Text-mining Tools for Optimizing Community Database Curation Workflows in Neuroscience

Kyle H. Ambert

Department of Biomedical Informatics

Oregon Health & Science University

A thesis submitted for the degree of

Doctor of Philosophy

April 29th, 2013

School of Medicine
Oregon Health & Science University

Certificate of Approval

This is to certify that the PhD Dissertation of

Kyle H. Ambert

"Text-mining Tools for Optimizing Community Database Curation Workflows in Neuroscience"

Has been approved

Dissertation Advisor – Aaron M. Cohen

Elis Boudreau

Committee Member – Eilis Boudreau

Gully B

Committee Member – Gully Burns

Committee Member – Melissa Haendel

Committee Member – Brian Roark

Committee Member – Kemal Sonmez

Contents

List of Figures

List of Tables

Acknowledgements

When I was eight years old, my mom signed me up for a chemistry summer camp for kids, and, on my first day attending, I fell in love with science. I was fascinated by the reactions I could create just by combining two substances that, on their own, appeared relatively uninteresting. I suppose the camp instructors saw the spark of interest they had created in me, and fanned the flame of interest, as any good teacher would. In retrospect, I suppose they regretted this, as, left to my own devices, I decided to explore new and interesting chemical combinations that were not a part of the curriculum, but which would surely be of interest to a young dab hand, such as myself. On the playground, watching from a distance as the fire department put out the flames that were the byproduct of my explorations, I considered that not all reactions were meant for a wooden table in a poorly-ventilated elementary school science wing. The school, fortunately, was able to be saved, but I could not: that was the day I became a scientist.

Over the years, the specific focus of my interests shifted, but, looking back, each shift contributed in some way to completing this dissertation. Along the way, there have been several teachers and mentors who have helped me become the scientist I am today. There were several key people who have challenged me to become a better writer and communicator, Jan Wickes, my 7^{th} grade humanities teacher, was probably the first to show me that writing could be informative and entertaining at the same time, allowing me to write history essays from the perspective of a fictional Keeshond, somehow transported back to the 4^{th} century. Similarly, Bill Goslow, my high school AP U.S. History teacher, taught me to focus my writing, and to find joy in the craft itself. Scott Janes, my music teacher for six years, through middle school and high school, also taught me a great deal about communication, instilling in me the ability to comfortably stand in front of an audience and still be myself. Wayne Robertson, my manager at the Oregon State University Writing Center, taught me that writing conventions still apply to technical writing and scientific manuscripts, and that concise technical explanations that are accessible to laypeople are priceless. Finally, the folks at the OHSU StudentSpeak blog–Rob West, Mark Kemball, and Allison Fryer, in particular–for helping me re-discover creative writing while working on my PhD.

In terms of science, there are many people who have influenced me, but three in particular stand out. Bill Struthers, my teacher and advisor at Wheaton College, taught me to find joy in discovery. During my time working with him, he went out of his way to encourage scientific curiosity in myself and many others. The summer of my Junior year, he created a Summer Behavioral Neuroscience research program that allowed myself and two others to spend two months of our va-cation in Neuroscience education and research that was beyond the scope of the school's curriculum at the time. Bill was a big inspiration for me, and his obvious care for his students was what inspired me to pursue teaching and academic research. Once I began my graduate career in the Behavioral Neuroscience department at Oregon Health and Science University, Shannon McWeeney was an immediately inspirational instructor. At the time, she was charged with teaching the Behavioral Neuroscience cohorts basic statistics. She took a curriculum that is so typically taught as memorizing a series of rules and made it come alive as the deep, complex field that it is. Once I joined the Department of Biomedical Informatics, it was the perspective she gave me which drove my research: the layers of complexity around us are what make things interesting and worthy of our attention.

Once I joined the Department of Biomedical Informatics, I began to work with my eventual dissertation advisor, Aaron Cohen. In my life I've met people who have inspired me to be a better scientist, and people who challenge me to be a better person. I've met a handful of people with whom I feel I can talk about anything, and some who inevitably have an insight into a situation I haven't considered. Aaron is all of these. He taught me how to take an abstract idea, and make into an interesting research question. He taught me how to stop thinking about interesting research questions and actually write some research code. He taught me how to translate an abstract idea into something that can be interpreted by the Python compiler, and then to write up the results in a way that might even interest others. Most importantly, he taught me that a successful research career doesn't have to come at the expense of spending time with family.

There are two other people who have been pivotal in getting me to the point of writing up my dissertation. First, my Mom. Growing up, I benefitted from a Mom who was a natural teacher. Maybe I just never really thought about it too much, but, somehow, it wasn't until I was in 7^{th} grade that I realized that other kids my age didn't get homework over the summer. I would always ask my Mom questions about how the world worked. Anything–photosynthesis, the Bernoulli effect, the electoral college, iambic pentameter–whether she had an immediate answer for me, or didn't know the first thing, my Mom would either teach me about the subject, or show me how I could teach myself. It's interesting to me, in retrospect, how empowering it can be for a kid to hear, "you know...I don't really know about that. Why don't we go to the Library and find out", in response to a question. This willingness to embrace the unknown is something that continues to influence me as a researcher today. Finally, my fiancée, Kim. Kim has been an ever-inspiringing presence for me through the process of researching and writing this dissertation. I could never have finished this without her understanding, love, and encouragement. She let me try out a lot of the ideas written about here on her before they could be articulated as code, helping me figure out the places where I didn't yet understand what I was trying to do.

Second-to-finally, there are several people who deserve acknowledgment with regard to this work in particular. First, my committee, Aaron Cohen, Brian Roark, Eilis Boudreau, Gully Burns, Kemal Sonmez, and Melissa Haendel were incredibly insightful as to the potential importance of my research, and holes in logic that I hadn't noticed. Alan Ruttenberg, of Science Commons, had some early contributions to the direction of these studies, and helped out with the annotation standards used in my reading. The people at the Neuron Registry were also helpful for getting this project started, helping out with setting the annotation standards and allowing me extended chunks of time to discuss their curation work flow. Finally, my "science friends" Steven Bedrick, Ted Laderas, David Gibbs, Zephy McKanna, and Nima Behkami were all helpful as attentive people, always willing to hear out my new ideas, as were all of the National Library of Medicine Fellows. The Allen Brain Institute, Allison Anacker, and the Ryabinin lab at OHSU's Department of Behavioral Neuroscience were also helpful for letting me try out some of my machine learning ideas in a Neuroscience research setting. Lastly, to my brother, Jake, one of my best friends, the most interesting guy I know, and whose pride in me is worth so much: may the power of Vectron bring prosperity to your house.

I've learned a lot since I was an 8-year old aspiring chemist. For one, nothing caught on fire during this dissertation. So, with that, I will leave the reader to consider my work.

To my fiancée & the love of my life:
Thanks for letting me try to impress you with algorithms.

—Kyle.

Abstract

The emphasis of multilevel modeling techniques in the Neurosciences has led to an increased need for large-scale databases containing neuroscientific data. Despite this, such databases are not being populated at a rate commensurate with their demand amongst Computational Neuroscientists. The reasons for this are common to scientific database curation in general–limitation of resources. Much of Neuroscience's long tradition of research has been documented in computationally inaccessible formats, such as the *pdf*, making large-scale data extraction laborious and expensive. Here we present three sets of studies designed to construct automated tools for alleviating three bottlenecks in the workflow of a community-curated knowledge base of neuroscience-related information. *Virk*, the first of these tools, is designed specifically with under-developed knowledge bases in mind, using active learning to allow them to quickly bootstrap their development. *Flokka*, our second tool, is designed for prioritizing a set of potentially-relevant manuscripts, so that they can be examined in order of their likely relevance. *Finna*, our final tool, is designed to rank-order the composite paragraphs of a likely relevant document, in terms of the probability that the paragraph contains information that is of interest to the database. Each of our systems attained a level of performance indicating its potential usefulness in the real world. In addition, we present a data set consisting of 962 expert-curated neuroscience documents–to our knowledge, the first data set of its kind. Each document was annotated at the document level, in terms of their relevance for a neuron-related knowledge base, and at the sentence level, in terms of whether a particular sentence communicates information that would lead to the document being included in the knowledge base.

Chapter 1

Introduction

Like most domains in biological research, neuroscience has experienced a recent explosion in the volume of published information [260]. The history of neuroscience can arguably be traced back at least as far as the works of Camillo Golgi and Santiago Ramón y Cajal, in the early twentieth century. Since that time, Neuroscience has become increasingly fractionated into various sub-domains, incorporating elements of Molecular Biology, Genetics, Computer Science, and Cognitive Science, to name but a handful. Each of these domains has proven equally prolific, such that a simple Google Scholar search for "*neuro**" yields nearly a million and a half results. To say that any one scientist can or should have this volume of information available for immediate recall in his or her head is folly, and yet, in order to efficiently advance the field of research, this can seem exactly what would be required. How can we, as neuroscientists, be sure we're not repeating ourselves, investigating experimental hypotheses that have long-since been addressed? How can scientists efficiently synthesize the knowledge within a particular neuroscientific sub-domain in order to see where the gaps in our knowledge lie? Given the diversity of training background in the neuroscience community, how can we be sure we're not falling subject to communication errors, using differing terminology to refer to similar neuroanatomical concepts, and therefore losing opportunities to make new conceptual connections? These are the kinds of questions that neuroinformatics and text-mining attempts to address. Each of these questions has been posed in the past, and a variety of solutions have been devised. Several of the solutions that have shown to provide greatest benefit, and most potential for continued use, are derived from a subdomain of machine learning called text-mining.

In this chapter, we will review many of the important developments in text-mining research, as well as how they apply and can be applied to research the behavioral neurosciences. Aside from its importance to Neuroscience in-and-of themselves, the information reviewed here sets the context in which this dissertation will be cast. Neuroscience is an incredibly diverse field, having many data sharing, terminology integration, and anatomy-related problems that make this an interesting area in which to conduct a set of text-mining studies. As the importance of data sharing and data integration increases, the importance of automated solutions to solving many of the workflow issues that arise in the data curation for the neurosciences will continue to increase in importance as well. Before turning to our work, we will first review some of the issues associated with working with text in the neurosciences, highlighting some of the interesting problems, important achievements, and future directions that it will likely go.

1.1 The Importance of Terminologies & Data Integration to Neuroscience

Neuroscience is an incredibly diverse field, consisting of researchers from chemistry, neurobiology, cognitive science, and mathematics, to name a handful. Although united by a shared interest in the study of the brain, each field has its own way of communicating–the cognitive psychologist might refer to Brodmann area IV, while the behavioral neuroscientist might refer to the primary motor cortex. Researchers in the field are not typically confused by this diversity in language, but computers often are. To the non-informatician, this may not seem like much of a problem–after all, computers don't need to "understand" concepts, they just need to efficiently manipulate them in accordance with a user's instructions. Unfortunately, this is very much not the case. Although neuroinformatics is still a young field, the heterogeneity of terms in neuroscience is already an interesting problem being addressed in order to improve mathematical modeling, machine learning document classification systems, and information retrieval systems, with a particular focus on neuroanatomical terminologies. Terminologies can be helpful tools for facilitating communication between colleagues in related disciplines and sub-disciplines, and aid in data sharing. Ontologies are related, as they allow for the definition of hierarchical types of objects and abstract concepts in a way that is understandable to both machines and human readers. Here we will discuss two example systems: NeuroNames, and the NIFSTD & BIRNLex Ontologies.

1.1.1 NeuroNames: A Neuroanatomical Nomenclature

Co-created by Douglas Bowden and Richard Martin [43; 200], NeuroNames[1] was one of the first popular neuroanatomical terminologies in the field. At the time it was first published, there was an absence of machine-readable neuroanatomical terminologies, making even something as seemingly straightforward as finding articles pertaining to a particular neuroscience sub-discipline difficult [42]. In order to facilitate scholarly communication and information retrieval in the neurosciences, Bowden and colleagues set out to define a "comprehensive set of mutually exclusive primary structures that constitute the brain" [42]. NeuroNames consists of 15,000 neuroanatomical terms, spanning 2500 brain-related concepts, culled from textbooks, atlases, and research articles. [41] One of the most important contributions of the NeuroNames vocabulary is that it constitutes one of the first attempts to standardize neuroanatomical terms, by serving as a reference point for neuroscientists, and by providing a standardized set of terms that unites multiply-defined anatomical structures by combining the concept name and the author and year of publication of the publication in which the term appeared (e.g., Area 9 of Brodmann-1909).

1.1.2 Leveraging Neuroscience Ontologies & Vocabularies in New Resources

The Neuroscience Information Framework (NIF) has made significant contributions to fulfilling the need for standardized terminologies in the Neurosciences.

[1] http://braininfo.rprc.washington.edu/

Their standardized ontology (the NIFSTD) is an hierarchically-structured collection of neuroscience-related terminologies, including terms used for describing neuroscience data, methods, anatomy, and digital resources [47; 139]. The project is an extension of the Biomedical Informatics Research Network (BIRN) project [201], is formatted in the style of a semantic wiki, as the NeuroLex, [26; 176; 177], and is easily downloadable in *owl* file format[1], the standard format for describing ontologies (see Figure 1, for an example). The idea is that Neuroinformaticians developing their own resources will be inclined to fold the NIFSTD ontology into their own resources, rather than developing a new set of terms, as has so often been the case in the past. In fact, this movement has already begun to take hold. For example, [203], used the NIFSTD to connect entities in clinical descriptions of human disease to model systems, thus bridging phenotypes in animal models from behavioral research to descriptions of human pathological features.

On the surface, terminologies and ontologies may not seem like useful resources to bench neuroscientists, as they seem something far removed from their day-to-day research activities. However, they begin to address what has long been recognized as a difficult problem that is deeply integrated into the way neuroscientists think about the brain. Sometimes called the *neuron classification problem* [38], the question of what constitutes necessary and sufficient criteria for distinguishing one type of neuron from another, dates back to the foundation of Neuroscience itself, with Camillo Golgi and Santiago Ramón y Cajal (Clarke & Jacyna, 1987). Are histological differences sufficient for distinguishing one cell type from another, or should spatial location in the brain be a factor as well? Within a particular

[1] http://purl.org/nif/ontology/nif.owl

region of the brain (e.g., central nucleus of the amygdala), is directionality also important (e.g., lateral, ventral, etc.)? These are the questions that neuroinformaticians, in collaboration with molecular neuroanatomists, aim to address. The decisions that are made will facilitate how researchers interact with one another, both in terms of scholarly discourse (e.g., how we describe neuron-related findings), as well as in terms of how they share data with each another. As users, other neuroscientists will benefit from further development of these tools by being able to better collaborate with other researchers in related disciplines.

Although the NIF/NIFSTD has made great strides in addressing the problems of data and information integration in the neurosciences, they have not solved it entirely. The NIFSTD is quite useful for integrating together the information that has been released in disparate data sources, moving them to a common language of neuroscience. However, the NIF, and sources like it, are only as useful as the data sources they are able to obtain information from. Integrating one's resource with the NIF means that any changes made to the local version of the tool that has been integrated are immediately available to users of the NIF–it's useful because there are a variety of constantly-updating resources that it is able to obtain information from, not necessarily because it is creating information itself. Thus, the same data curation problems remain, but are not the consideration of the NIF *per se*. Instead, the resources that it is obtaining information of each need to have general-purpose tools available to them that can streamline their respective workflows, allowing them to efficiently update their resources as new information is discovered.

1.2 Information Retrieval in Neuroscience

Information Retrieval (IR) is a sub-discipline of computer science that is concerned with developing accurate algorithms for retrieving information from databases of documents or textual information [126]. In general, IR systems are designed to take users' search requests (queries), identify relevant data in a database, and return a ranked list of results that is ordered according to likelihood of relevance to the input query [126]. Such systems are quite common in today's information-heavy age, with common examples being Google search, PubMed, or Apple's Spotlight system, on the OSX operating system.

In the Biomedical sciences, IR is most-commonly associated with the National Library of Medicine's PubMed search engine[1], which queries against a database of over 21 million peer-reviewed scientific publications. In addition to joining query terms via standard boolean operators (e.g., AND, NOT, OR[2]), PubMed also utilizes a vector representation of the query to identify the most relevant related articles. [141] Although PubMed is one of the first resources many researchers will use when performing a literature search, it is not without its limitations.

Domain-specific IR systems can provide several advantages over general-purpose ones, such as PubMed. Although general-purpose biomedical IR solutions will often suffice, there are situations where neuroscientists can need specialized search tools [22]. For example, a researcher conducting a literature review on retrograde tracer studies could run a simple PubMed query *retrograde tracer*, and

[1] http://www.ncbi.nlm.nih.gov/pubmed/

[2] http://www.ncbi.nlm.nih.gov/books/NBK3827/

obtain approximately 2700 results (query performed July, 2012). The enumerated publications will be articles in which the term *retrograde* and the term *tracer* appeared in at least one of 64 data fields (e.g., Abstract, MeSH Term, Title; for a full and up-to-date list, see[1]. If the researcher is only interested in studies that actually used retrograde tracing as an experimental method, the results returned by PubMed are likely to contain many documents that are not of interest (e.g., 35 of the results obtained were review articles), and, in addition, are likely to not identify publications that would have been relevant (e.g., studies that used retrograde tracing, but did not include this exact term in the titles or abstracts). Because of this, the researcher performing the literature review will have to spend time manually going through the entire list of results to identify the publications that are genuine of interest, in addition to performing extra queries, to obtain articles that were not initially identified. The costs associated with performing these tasks are often prohibitive; thus, neuroinformaticians have constructed specialized search tools for the neuroscience literature base that can overcome this difficulty. Two of the major developments in Neuroscience Information Retrieval (NIR) solutions that have come about in the last five years are Textpresso for Neuroscience [218], and the platform developed by the Neuroscience Information Framework [105]. As these two systems have taken somewhat different approaches to addressing NIR, we'll discuss each in turn.

[1] http://www.nlm.nih.gov/bsd/mms/medlineelements.html

1.2.0.1 Textpresso for Neuroscience: A Combination Information Retrieval & Extraction System

Textpresso for Neuroscience is the Neuroscience-specific version of the popular Textpresso system, from Müller, Kenny, and Sternberg, of Howard Hughes Medical Institute and California Institute of Technology [217]. Textpresso is an IR system distinguished by two key components: the ability to perform full text searches, and the use of an ontology (see previous section), allowing for defining types of objects and abstract concepts in a way that is understandable by both machines and human readers. One can easily perform a search for *anchor cell* in a general-purpose search engine, but many of the documents returned may end up being, for example, about maritime justice systems. If we are truly interested in documents only refering to anchor cells in the biological sense, an ontology could be useful for informing the search system that anchor cells are a type of biological cell in the *C. elegans*, and are characterized by production of the signaling molecule LIN-3/EGF [128]. To allow full text searching, Textpresso uses the xpdf software[1] in combination with journal-specific templates, which allow them to extract the plain text from the PDF representation of a publication with some degree of accuracy. This approach contrasts with that taken by PubMed, which uses publisher-supplied metadata (e.g., keywords) for their database. Although this approach is limited somewhat by the hit-or-miss process of extracting text from a PDF, it does allow users to query against the entire document, which can be advantageous, particularly if users wish to query based on text that is likely to be found in figure captions [130]. Similarly advantageous is Textpresso's use of ontologies to facilitate accurate searching of the text. In the original Textpresso

[1] http://www.foolabs.com/xpdf/

paper [217], Müller and colleagues describe a variety of categories that were used to mark up their documents, enabling a variety of concepts to be included in a search query, including biological concepts, relationships, and descriptions. [218] For example, to search for brain areas in which the TRP channel TRPC1 is found, the user could specify to include TRPC1, and select the categories *brain area*, and *NIF (neural) stem cell types*.

To extend their approach to the Neurosciences, Müller and colleagues included publications from 18 Neuroscience journals that were selected in collaboration with the NIF [303]. As of the time of this writing, their system allows full text searching for over 100,000 neuroscience publications, and allows for the specification of several neuroscience-related term categories and sub-categories (Table 1). Textpresso for Neuroscience can be accessed either through the systems' main website[1], or through their webservice. In addition, it has been incorporated into the NIF [117].

The Textpresso for Neuroscience system can be used by research scientists outside of neuroinformatics to further their own work. Because the Textpresso system allows for full text searching of research publications, users can perform more specific queries that are targeted at text occurring throughout the document. If one is interested in retrieving documents based on information that is in figure captions (where experimental results are frequently described with greater concision), this would be possible with Textpresso, since the entire text is indexed, but it would only be possible for the open access publications that are indexed

[1] http://www.textpresso.org/neuroscience/

by PubMed. A major limitation of the system, however, is that its bibliography has not been updated since 2009 (*website accessed on July, 2012*). This highlights a shortcoming of many digital resources: it is typically more common for research scientists to receive grant funding for a project aiming to develop new methods for using or accessing digital resources than it is for one that will maintain said resource beyond its initial funding period. An incredibly useful tool, such as Textpresso for Neuroscience, is only as good as the data it indexes, and, since the number of neuroscience-related publications is always increasing, without ongoing support it can quickly become out of date. This highlights the need for developing general-purpose tools for maintaining the distributed data sources that are used to inform information retrieval tools such as this, as well as the one we review next, the Neuroscience Information Framework.

1.2.0.2 Information Retrieval Using the Neuroscience Information Framework

The Neuroscience Information Framework was created as a part of the National Institute of Health's Blueprint for Neuroscience Research, in 2004. [27; 105] A complete description of the NIF can be found in Chapter three of this volume. Briefly, the NIF distinguishes itself from more traditional document IR systems (e.g., PubMed) by providing a central framework with which existing online Neuroscience resources can be integrated. These resources aren't just limited to documents–they include expression data (e.g., as documented in BrainSpan[1]), atlases (e.g., as documented in the Allen Mouse Brain Atlas[2]), and imaging

[1] http://www.brainspan.org/

[2] http://www.brain-map.org/

databases (e.g., as documented in the Brede Database[1]). This diversification stems from the NIF's driving goal, to facilitate access to, and integration of, heterogenous neuroscience data, for the purpose of enabling new discoveries to be made, and new neuroinformatics tools to be developed [105].

Integrating dynamically-updating data from geographically-distributed resources can be something of a daunting task, since all data needs to be mapped from different views of the human brain into a common data model, but, if carried out properly, it provides significant advantages to users. The NIF currently offers three levels of data integration to neuroscientists who have information resources they would like to make available. The most in-depth of these levels allows contributors to integrate their data into the larger NIF data federation by submitting schema information and database views to the NIF mediator. They use a concept mapping tool to map the data to the tables, fields, and values in the NeuroLex ontology[2]. This allows resource providers to leave their data in its original format, maintaining its integrity, and leaving any necessary transformations to be made in the ontology mapping stage. This allows for updates to the content to be made available as they happen. From the perspective of the user, this deep-level integration means that queries performed on the NIF's main page will be run against a variety of neuroscience data resource simultaneously, with the results packaged in a way that's meaningful and easy to navigate. For example, running the query *Amygdala basolateral nucleus pyramidal neuron* on the NIF returns 189 literature results, and several results from the data federation–four brain regions,

[1] http://neuro.imm.dtu.dk/services/jerne/brede/
[2] http://neurolex.org

two genes, four grants, and two diseases (*query performed July, 2012*). If more than one of these resource categories were of interest to a user, and he or she weren't using the NIF, multiple queries would need to be performed on several external databases (e.g., BAMS, OMIM, and NIH RePORTER) using different query formats and terminologies, which would be time-consuming to perform, and would leave the scientist to do the integration of the retrieved results.

One use case for a resource like the NIF is that of data integration. Because the NIF takes care of mapping multiple heterogenous data resources back to a common data ontology, it is possible to query across multiple data types in a meaningful way. To return to the *Amygdala basolateral nucleus pyramidal neuron* query example, if a scientist were interested in doing a study involving this cell type, he or she could learn that four grants have been funded to NIH institutions on this topic, but that the most recent one ended in 2011. One would also find that, in the Online Mendelian Inheritance in Man (OMIM) database, it related to brain-derived neurotrophic factor (BDNF), obsessive-compulsive disorder, and congenital central hypoventilation syndrome. All of this information would be helpful to developing a new hypothesis or designing a study, and it is immediately available in one integrated resource.

A second use case relates more directly to text-mining experiments that might be conducted by or for behavioral neuroscientists. Behavioral assays, such as the elevated plus maze [243], conditioned place preference [78], or the adjusting-amount procedure [212], are the backbone of behavioral neuroscience. Such procedures are used as behavioral models of disease, and used, for example, to evaluate the efficacy of drugs for treating disease. If a scientist were conducting a literature review on the use of the adjusting-amount procedure in evaluating the effects of dopamine-2 receptor antagonists on impulsive choice, they could perform a query in PubMed, and manually sift through the many documents it would return. Carrying out the same task using the NIF, however, would allow the researcher to leverage the previously-described ontology, ensuring that the results returned are indeed relevant to both the behavioral procedure in question and the specific class of drugs. That is, the results would include instances of the procedure and drug themselves, rather than just the words themselves (i.e., *adjusting-amount procedure* as a method, rather than documents containing the words *adjusting-amount* and *procedure*). As it stands, this tool is useful enough, but the future possibilities for this type of information retrieval could greatly affect the way literature reviews are conducted in the behavioral sciences. For example, using a procedure similar that described in the CoCoMac classification experiment described in the following section, one could use the NIF to obtain documents in which certain behavioral procedures are known to have been used. These data could be used to create a document classifier that would then identify research publications in which the procedure was used, but which had not been identified by the NIF either because they were newly-published, or because of publisher error. Such tools, like the ones we will present in this dissertation, are necessary for maintaining the relevance of such knowledge bases.

1.3 Supervised Text Classification in the Neurosciences

The frequency and volume of newly-published scientific literature is quickly making the maintenance of publicly-available scientific databases unrealistic and costly. Assuming a newly published article is identified as potentially containing relevant information, database curators can spend up to 48 hours determining whether it should be included in their database, and manually extracting the relevant information from the full text document. Therefore, supervised document classification systems are an increasingly effective machine learning tool to promote efficiency for the many text-related tasks in biomedical science [69]. In such systems, a collection of documents are manually annotated with regards to some criteria–for example, include/exclude in a database, or relevant/irrelevant for a literature review, and are then used to train a classifier to make judgments on documents that have not yet been seen. Cohen and colleagues [65; 66; 303] have used such an approach to provide text-mining support tools to the Systematic Review community. In this work, the Medline records associated with documents are used as input features to a classifier that assigns each a relevance judgement for a number of systematic review topics. In a more biomedical application, they have also used text classification for using the text in the i2b2 challenge tasks for mining clinical discharge summaries to predict smoking status [64], obesity-related disease comorbidity status [8], and identification of biomedical concepts, assertions, and relations (e.g., type II diabetes, "disease is present," and "hypertension was controlled by hydrochlorothiazide," respectively) [9; 68].

In the neurosciences, document classification is manifest in the maintenance of databases documenting primary-source experimental data on, for example, neuroanatomical connectivity. Many of these databases have become invaluable resources for neuroscientists studying connectivity itself [36; 270], and a useful reference for behavioral neuroscientists in conducting lesion or micro-injection studies. Despite the frequency with which they are used, the information contained in such connectivity databases is often based on user-submitted connection information, and it may or may not be possible for the database owner to find enough time to verify the information for themselves, or to identify new information and update the database.

Gully Burns and colleagues' Scientific Knowledge Mine (SciKnowMine) project is an important development for behavioral researchers. [125; 237; 238] They recently showed how their document classification/biocuration pipeline can be used to help curation at the Mouse Genome Informatics group[48]. They take an all-in-one approach to solving the problem of applied text-mining, providing a system that stores documents, extracts text from PDFs, pre-processes data, maps the text to an ontology, and outputs the data to web services. They used this system at the MGI to perform automated document triage (identifying which documents in a large data set are irrelevant for some curation task). Burns and colleagues' unified system approach to text-mining is an important example of how machine learning experts and neuroinformaticians are beginning to recognize the importance of making their tools accessible and useful for performing common tasks in research scientists' work flows; such tools are motivating examples for the work in this dissertation. Similarly, the work of Lynette Hirschman, Gully Burns, and others [51; 53; 130; 233] has shown how text-mining can be used to optimize biocuration workflows in the molecular sciences. In particular, text-mining can be useful for the document triage task described above, wherein bio-entity identification and normalization (i.e., removing specific mentions of biological entities from text prior to classification) can be leveraged to develop a useful document classification system, or to suggest relations for annotation in a database. For example, in a recent study where we built a document classifier for identifying protein-protein interaction (PPI)-related information [9], we observed that replacing protein mentions in the text of documents with a normalized feature (e.g., changing "5-HT Receptor" to "PROTEIN_MENTION") led to improved classification performance. The reason for this is that in many biocuration classification procedures, it is more important that the classifier use the contextual features surrounding annotatable information than the specific entities themselves. In the case of neuroanatomical connection classification, this would be akin to relying more on features like *connects*, *afferent*, and *efferent*, rather than ones like *hippocampus*, *cortex*, and *striatum*. Similar to the PPI normalization case described above, the contextual features will allow the classifier to more-easily identify documents containing annotatable information regarding neuroanatomy that it has not previously seen.

1.3.0.3 Classification for the CoCoMac Database – An Example of Text-mining for the Neurosciences

Text classification experiments can be fairly complex, but as a rule of thumb, there are generally five elements to a text-classification pipeline:

1. *Text extraction:* free text is extracted from a PDF document (e.g., in [238]),

website, or some other input resource, and put in a format readable by the classification software. This could be a directory of *txt* files, an *xml* file, or a database.

2. *Pre-processing:* this step is important to get the extracted text into a regularized and predictable form [9]. In the above-mentioned PPI study, we found that an important feature of a document classifier for identifying papers containing PPI-related information was a step in which we removed all mentions of specific proteins. Classification systems make their judgements based on the characteristics of the input documents. Thus, if one's goal is to create a system for identifying documents containing a variety of PPIs, and not just those that were observed in the training data, removing specific PPI mentions forces the classifier to make its judgments based on other document characteristics, for example the sorts of sentence structures that often describe relation information between two proteins (e.g., *"our data demonstrate that PROTEIN interacts with PROTEIN"*). Other procedures frequently done during pre-processing is the removal of all punctuation in the text, and case-normalization.

3. *Tokenization:* In this step, the pre-processed documents are split into individual tokens, or features. A simple normalization procedure that is frequently used in text-mining experiments is simple unigram tokenization. This approach splits the document into a "bag of words", wherein each feature is a word, and no ordering is conserved. Other approaches will be based on bi- or tri-grams (individual pairs or trios of words, respectively), which retain some word ordering observed in the original document.

4. *Modeling:* The collection of tokens resulting from the tokenization step is next modeled for use by the classification algorithm. Binary feature modeling is a commonly-used modeling procedure in which the unique set of features observed in the entire training document collection is assigned a position in vector. Each document is then represented as a vector of the same length, in which each position contains either a zero or a one, corresponding to the absence or presence of that feature within the document in question.

5. *Classification:* The classification algorithm is given a set of (*vector, true class label*) pairs (during classifier training), or just document vectors (during classification), and using whatever classification procedure has been selected for the task, it will either learn the mathematical relationship between document feature vectors and their class labels (in training), or predict the class label of new documents (during classification). Many classification algorithms exist, but Support Vector Machines [142], and Naïve Bayes [205] are commonly-used procedures in text classification.

As a proof of concept for the application of text classification in the neurosciences, we developed a machine learning framework for automating the identification of sentences containing neuroanatomical connectivity information appropriated for incorporation into the CoCoMac online database of Macaque connectivity information[1]. The CoCoMac database was selected for several reasons. First, it contains a great deal of connectivity information indexed according to the PubMed Identifier (PMID) associated with the article from which the information was

[1] http://www.CoCoMac.org

obtained. Many online neuroscientific databases contain a combination of unpublished experimental data and peer-reviewed results, and since this proof-of-concept system is concerned with verifying the information that has been accepted into the scientific body of knowledge, it made sense to choose a database specifically focusing on the published literature. Second, the CoCoMac database has an intuitive, built-in URL search interface that makes it easy for an automated system to pull down information on an as-needed basis, rather than having one or more individuals spend time performing manual information retrieval. Third, CoCoMac's article curation process is rigorous and well-documented. Furthermore, the CoCoMac database has not been updated since 2005, due, according to its founder, to the fact that verifying the information contained in one article can take up to two days (Kotter, 2009; personal communication)–emphasizing the need for automated methods for streamlining the curation process, which we will present in this dissertation.

We created a classifier that, given a list of connections supposedly documented within an article, would identify the sentences in the article's abstract containing this information. We first obtained a complete list of PMID IDs contained in the CoCoMac database (approximately 600 IDs), and located an electronic version of the fulltext for each using PubMed, Google, and Google Scholar. Even though the present set of experiments was based on sentence-level classification judgments in the abstract, an important follow-up experiment is to expand our classification to Results sections in full text (see Chapters 2, 3, and 4), as well and therefore our studies included only those abstracts for which we could obtain the entire document (approximately 250). For this subset, we extracted the abstracts from their respective PDFs. In order to train a classifier to identify connectivity information at the sentence level, it was necessary for us to manually markup a subset of our abstracts using the Knowtator annotation plugin for the Protege ontology management system [227], identifying those sentences containing connectivity information, as well as any single- or multi-word strings that refer to a particular neuroanatomical concept. For this proof-of-concept we only annotated 60 articles in our data set, however this resulted in a dataset containing approximately 600 sentence/connectivity judgment pairs. We performed cross-validation on these data to develop a baseline support vector machine (SVM [285])-based classifier against which we compared the results of various feature selection and resampling experiments. For thoroughness, we compared the performance of our SVM-based systems to that of a non-SVM classifier, kIGNN, a mutual information-based k-nearest neighbor classifier that has been shown to be effective in identifying documents containing protein-protein interaction-related information [9]

The performance of our baseline system, according to the area under the receiver operating characteristic (AUC), is depicted in 1.1. For the AUC, random classification would equate to a value of 0.5. Although our baseline system performs better than random (0.63±0.05), an examination of the ratio of positive classes in light of previous research [62] led us to hypothesize that the over-abundance of negative class-sentences was leading to poor performance. To overcome this, we used a previously-described resampling method [62], in which we sampled (with replacement) from our existing dataset to create a new one, but increased the probability that a given sample would be from the positive class. Performance of this approach is depicted in Figure 3 for a range of probabilities for obtaining a positive class sample (1-5: $1x$ through $5x$ as likely). Importantly, since this is a resampling method, even though the $1x$ probability level is equivalent to our baseline system, this method results in a dataset five times as large as that of our baseline system. This is reflected in the fact that the AUC of the baseline and $1x$ system are roughly the same, but the 1x confidence intervals are much tighter.

We were interested in determining feature selection and feature generation methods that would lead to improved performance. Here, we examined the effects of neuroanatomical term normalization and neuroanatomical term-based distance feature generation on performance. Using the neuroanatomy markups obtained during our Knowtator annotation procedure, we replaced all recognized neuroanatomical features with a single common feature. To examine the effects of doing this on performance, we plotted the information gain associated with each

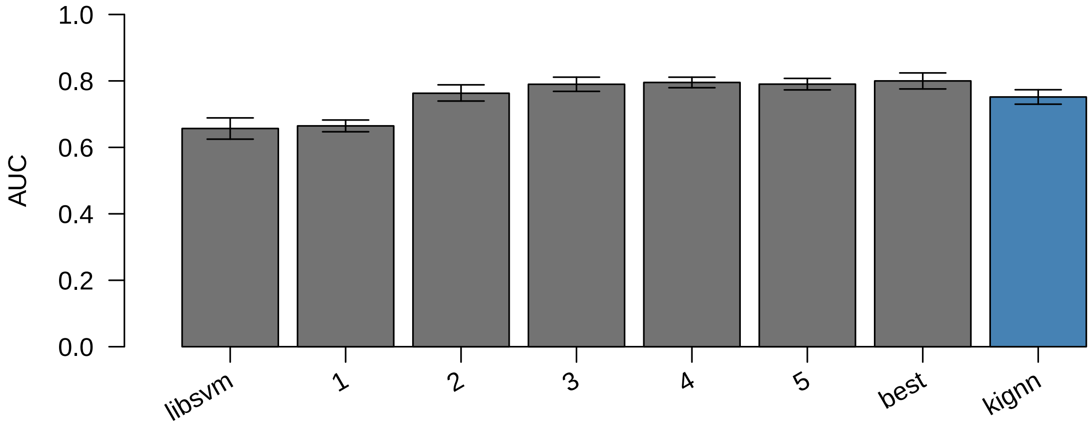

Figure 1.1: AUC (with 95% confidence intervals) comparisons of our baseline (libsvm) and various number of costs for misclassifying a positive sentence (1-5), with a previously-successful relationship extraction system (*k*IGNN).

feature for our normalized and non-normalized datasets (Figure 1.2 normalized: blue; non-normalized: black). As this figure makes clear, when all neuroanatomical terms are replaced with a common feature, the peak of the information gain is sharper and shifted to the left. This implies that many of the predicitive features in the non-normalized collection were neuroanatomical terms, and that performance would be improved by grouping all these into a single feature. In terms of qualitative implications, this would mean that one of the best ways our classification system was able to distinguish between sentences that were positive or negative for containing connectivity information was whether they contained neuroanatomical terms. Figure 1.3 depicts the distribution of the average distance between neuroanatomical terms within each sentence for the positive (black) and negative (red) classes. The results depicted in Figure 1.3 fit well with those depicted here–the peak of the distribution for the negative class is sharply centered around 0 (meaning that one or fewer neuroanatomical terms were contained in the sentence.). The positive class is also centered around 0, but it drops less gradually toward positive values. Based on these results, we hypothesized that normalizing our dataset for neuroanatomical terms, as well as including a feature describing the average distance between neuroanatomical terms in a given sentence, would improve performance of our classifier. This combination of features led to substantial improvement in our cross-validation studies (AUC: 0.81).

This proof-of-concept text classification experiment demonstrates the feasibility of developing a sentence-level neuroanatomical relationship classifier using a small number of annotated articles. We were able to achieve a level of performance that could be useful for performing actual classification tasks (i.e., AUC$\geqslant$0.80) by using a support vector machine classifier and cost-based resampling methods. In

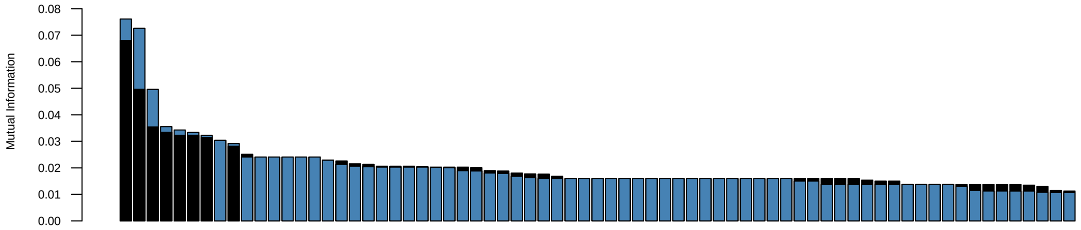

Figure 1.2: Feature information gain with (blue) and without (black) neuroanatomical term normalization, for the CoCoMac classification experiment.

practice, Neuroscientists could use a system such as this to extract a literature-based *connectome* for a particular model organism. In particular, this tool could be integrated with a system recently developed by French and colleagues [100; 101] to identify specific brain regions and pull down their gene expression-related information from the Allen Brain Atlas [180]. Integrating all this information could be used to create an integrated visual map of brain connections and their gene expression data that could be used, for example, to model spatial correlation of gene expressions in the brain.

Although the supervised classification approach to developing a knowledge base is a useful one, it is not without limitations. We will review these in the next section.

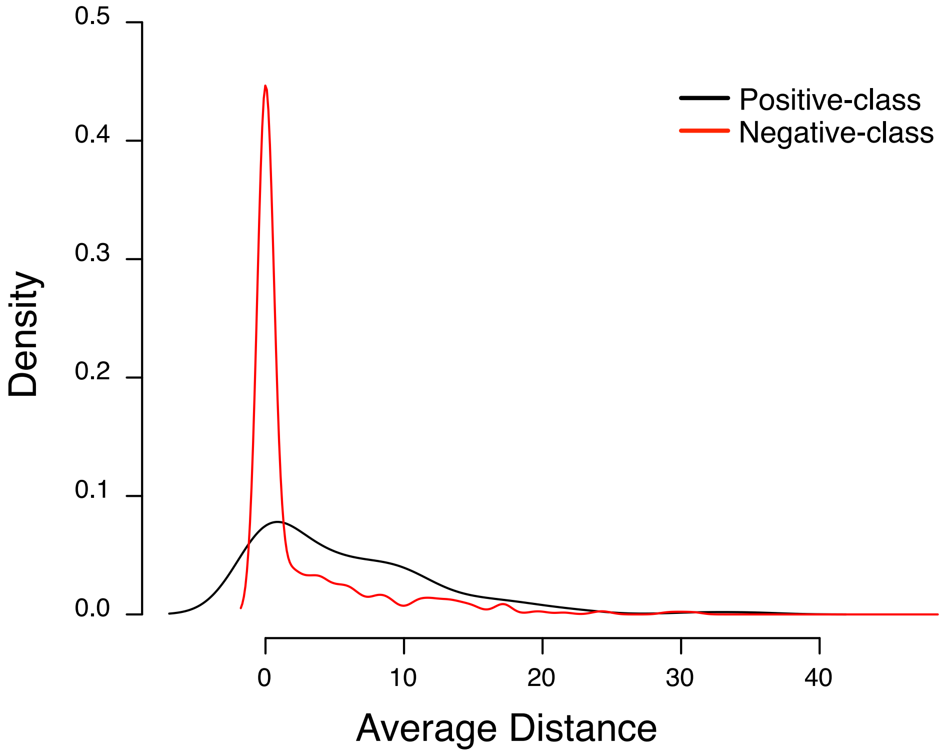

Figure 1.3: Distribution of average distance between neuroanatomical terms in the positive (black) and negative (red) classes, for the CoCoMac classification experiment.

1.3.0.4 Efficient Approaches to Classification: Knowledge Mining

One alternative to using machine learning for assisting manual database curation is that of automated mining from document databases. Because the financial and time costs associated with developing a large curated document collection is often prohibitive, researchers will sometimes perform automated association mining, in which textual features are extracted from a large collection of input documents and used to either further one's understanding of the relationships between the documents themselves, or to develop hypotheses that can be investigated on their own. Voytek and colleagues [290], for example, used co-occurrences of brain region mentions, cognitive functions, and brain-related diseases to demonstrate that known relationships can be extracted in an automated and scalable way by using clustering algorithms. Importantly, they were able to extend this approach to semi-automatically generate hypotheses regarding "holes" in the literature–associations between brain structure and function, or function and disease which are likely to exist, but lack support in the literature. For example, they discovered that the structure *striatum* and the term *migraine* were strongly related to the term *serotonin* (they co-occurred in nearly 3000 publications for each relationship), yet the *striatum* and *migraine* had only 16 shared publications themselves, indicating that this association may exist but be understudied.

French and colleagues [99] used knowledge mining to automatically map neuroanatomical identifiers found in a large volume of journal abstracts from the Journal of Comparative Neurology (JCN) to connect over 100,000 brain region mentions to 8,225 normalized brain region concepts in a database. In this work, they used an annotated collection of abstracts from JCN and other Neuroscience journals [97], expanding all abbreviations in the text, and manually identified the brain region mentions they contained. They also put together a dictionary of 7,145 brain regions having formal unique identifiers from the NeuroNames vocabulary [41], NIFSTD/BIRNLex [47], Brede Database [225], Brain Architecture Management System [40], and Allen Mouse Brain Reference Atlas [85]. In total, they used five different techniques to link the free-text neuroanatomical mentions to the compiled set of terms: exact string matching, bag of words, stemming, bag of stems (similar to gap-edit global string matching [271]), and the Lexical OWL Ontology Matcher, which allows for the specification of specific types of entities.

[111] Scientists interested in using these resources could incorporate their annotated data (freely available at[1] into a classification system like the ones described in the previous section.

1.4 A Case Study in Neuroinformatics Knowledge Base Maintenance: The Neuron Registry

In the introduction to his recent book on Systems Biology and Neuroscience, Neuroscientist Olaf Sporns proposed that “we cannot fully understand brain function unless we approach the brain on multiple scales.” [267] To approach our understanding of the brain on multiple scales, Computational Neuroscience has turned to multilevel mathematical modeling—a collection of techniques which allow for mathematically representing the layered complexity of the brain. Though effective, such models require large volumes of data for each layer they represent, which is most commonly obtained in some form of database.

Neuroscience is fractionated into many sub-disciplines, each of which, though concerned with respective questions of interest, is in some way motivated by extending our knowledge and understanding of how the brain works [161]. In response to this fractionation, the collective goal of *reverse-engineering the brain* [247] was recently set by Neuroscientists and Engineers. In practice, a reverse-engineered brain would be a mathematical representation of the brain as a system,

[1] http://www.chibi.ubc.ca/WhiteText

which would be useful for conducting *in silico* simulations that would increase our knowledge of brain operation in diseased and healthy states. Such knowledge would translate to better, more targeted treatment for neurological disorders, like Traumatic Brain Injury [131; 157], Alzheimer's Disease [215; 260], and the relationship between variation in an individual's brain structure and function [100; 266; 268]. This future for Neuroscience will be attained by leveraging the wealth of knowledge already obtained over its productive history, in the form of mathematical models and simulations of the brain. Such simulations have been, and will continue to be, useful for determining the relationship between brain structure, function, and treatment. A significant barrier to achieving this goal, however, is getting the extant Neuroscience-related knowledge into machine-accessible databases. Using a manual approach for such a task is unrealistic; it would require an excessive amount of financial and personnel resources that are unavailable. Thus, scalable machine learning methods must be created, for identifying and extracting Neuroscience-related knowledge from the published literature, and subsequently storing it in machine-readable databases for computational modeling.

1.4.1 Databases & Research Science in the Information Age

Databases (or, in the case of a database of qualitative scientific knowledge, *knowledge bases*, in this work) play an important role in the way modern scientific research is conducted. In general terms, they have two purposes[58]: sharing the results of scientific research in a meaningful way, and providing structured data that is useful for leading to previously unknown scientific knowledge. In either case, the integration of database technology in scientific research tends to be preceded by the development of methods that yield a high volume of data [264]. In Genomics, for example, an event that is often cited as important in leading to the prevalence of sequence databases is the development of rapid methods for determining the base sequences in DNA [253]. More than just for sharing sequences themselves, these databases have been important tools for generating new knowledge in Genomics and other fields alike. For example, comparing similarities between the mouse and human genome has led to discoveries of new gene regulatory elements [124], and genetic events that may have given rise to differences between the respective genomes [274]. Similarly, the availability of sequence databases for multiple organisms has enabled the phylogenetic mapping of organisms, allowing us to learn about the evolutionary history of species [228].

1.4.2 The Importance of Databases to Neuroscience Research

In partial contrast with the Genomics field, the integration of the neuroscientific knowledge base with database technology has been only somewhat preceeded by the increased use of high-data-volume methods, such as electroencephalography (EEG) and neuroimaging. Adoption of database technology has been partly motivated by its success in bioinformatics (e.g., the Allen Brain Atlas[179]), and by the need for meaningful communication between sub-disciplines in the field (e.g., the NeuroNames brain hierarchy [43]; see Chapter 3), as well as the desire for data-informed computational models of brain function.

In the late 1990s and early 2000s, discussion of databases within the realm of Neuroscience was largely concerned with the community's reluctance for sharing data [58; 159; 160]. At the time, the primary barrier to digitally representing Neuroscience knowledge was an apparent proprietary view, held by many investigators in the field, regarding their discoveries. This no longer seems to be a substantial barrier; there are several possible reasons for this. For one, a possible reason for this might be the obvious success of databases and data sharing in related fields, like molecular biology and genomics. Alternatively, the increasing emphasis of funding organizations, such as the National Institute of Health, on having a plan for making data available to researchers may be an important influence. Just as likely, the shift in thinking may be due to the influence of the International Neuroinformatics Coordinating Facility [1] and similar organizations on shaping the way research Neuroscientists think about the future of their field [94].

1.4.3 The Neuron Registry: A Community-Curated Knowledge Base for Neuroscience

The Neuron Registry (NR) is a community database being developed by the Neuron Registry Task Force (NRTF), under the International Neuroinformatics Coordinating Facility's (INCF) Program on Ontologies of Neural Structures (PONS). The stated purpose of the NRTF [2] is to put structure into place for a knowledge base of neuronal cell types, providing a formal means for describing

[1] http://www.incf.org/

[2] http://pons.neurocommons.org/page/Neuron_registry

and quantifying existing cell types from their properties, and populating it with information from the literature. Among the task force's goals with regard to the development of the NR database itself, three key responsibilities stand out:

1 To initially populate the registry.

2 To continuously curate the registry.

3 To establish a self-propagating long-term system for maintenance and updates.

1.4.3.1 Do we need another knowledge base?

At times, it can seem that the landscape of the internet is scattered with various scientific knowledge bases. Each likely has a niche amongst a certain group of research, and even the ones which are no longer updated were probably at one point quite useful for the users the developers were targeting. There are many barriers to maintaining such services–lack of publicity, inability to easily fit into the target user groups' research workflow, or possibly failure to consider a viable plan for their long-term maintenance. Database developers working in a quickly-changing field, such as Neuroscience, must be aware that future research is likely to bring major changes to the field which, if not they are not careful, could potentially render the database obsolete, by not fitting into their way of representing the data. Neuroimaging databases, for example, have been met with much this same issue on more than one occasion[284]. Facing obsolescence due to drastic changes in imaging technology, changes in the capabilities of the internet (e.g., semantic web technology), and shifting opinions on whether to share raw or pre-processed data, considerable effort has been made in that community to change the way their data are stored and shared.

1.4.3.2 A Clinically-relevant Use Case for the Neuron Registry

The importance of a well-structured database of neuronal attributes extends beyond bench research by providing a centralized computationally-accessible warehouse of neuron-related Neuroscience findings that can be used to further our understanding of the diseased brain. Clinicians[3], Psychiatrists, in particular[138], have expressed a desire for the modeling community to create models that are more obviously clinically-relevant, especially to the extent that they don't oversimplify the biology involved. Many Neuroscientists [17; 20; 31; 138; 247] argue that multilevel mathematical models of the brain will be a key tool for in increasing our understanding of how the brain works, and translating that knowledge into clinically-relevant findings. Such models are mathematical representations of the integrated layers of the brain, and are used to recreate computational simulations of various Neuroscientific findings, and subsequently generate hypotheses that can be tested in the lab. For example, Migliore and colleagues [209] recently used a computational modeling strategy to investigate the biological basis of hallucinations in Schizophrenia. An important characteristic symptom associated with Schizophrenia is hallucinations, a phenomenon which has been thought to arise from problems in hippocampal-mediated associative recall[182]. Much is known about the electrophysiological and morphological properties of neurons in the hippocampus, particularly in the cornu ammonis 1 (CA1) region[259], which is known to be important in memory encoding and recall. Using data obtained from Ascoli's[18] Neuromorpho project[1], Migliore and colleagues were able to illustrate how a change in context-dependent input to an ensemble of CA1 synapses would lead to activation of perceptions not relevant to one's immediate context (i.e., hallucinations). By drawing qualitative (e.g., receptor types) and quantitative information (e.g., specific parameter values) directly from the Neuroscience literature, their model was able to demonstrate one possible mechanism of a well known and poorly understood psychiatric symptom. Although similar modeling efforts have been previously undertaken by Computational Neuroscientists and Psychiatrists[112; 234; 265; 289], many of these models were created using a simplified representation of neurons and their properties. Although this approach can be useful in some cases, Migliore and colleagues were particularly interested in understanding the role of the individual neuron in generating the schizophrenic symptoms. Thus, particular care was taken to make sure that the neuron properties used in the model reflected those which had been experimentally verified. Their approach resulted in a model which was not only consistent with experimental and clinical findings, but was able to generate a hypothesis for inconsistent results regarding hippocampal activity in schizophrenic subjects, as measured by functional magnetic resonance imaging (fMRI), and to make a testable prediction for context-dependent associative learning in schizophrenics.

Such models are going to play an important role in moving bench and clinical Neuroscience forward, but are only made possible by leveraging empirical neuronal data. Although much of this data has already been collected and published over the years, much of it is only available in hard-copy, or *pdf* form, which is ineffi-

[1] http://neuromorpho.org/

cient to access. The data used to carry out the Migliore study was obtained from existing public resources for Computational Neuroscientists, such as the aforementioned NeuroMorpho project, which stores digital reconstructions of neurons, and ModelDB[129], an online resource for storing mathematical Neurosciences models[1]. These resources are useful for studies involving neuronal morphology and their electrophysiological properties, but in order to extend such simulations to incorporate circuits of realistically-depicted neurons, additional resources are needed.

1.4.3.3 The Neuron Registry as an Aid to Developing Neuroinformatics

The Neuron Registry will be an important computational modeling resource for making the jump from biologically-realistic models of individual neurons, to networks of biologically-realistic neurons–a leap that will be necessary to make, in order to create mechanistic models of complex behavior, such as addiction, or to move from modeling the mechanisms underlying particular disease symptoms, to those which lead to, or persist during, a disease state. Much is known about the properties of neurons, but, until recently, it has remained unclear how to best distinguish types of neurons from one another. Moreover, the question of what constitutes a neuron "type" has been rather *ad hoc*[39]. The consequence of this, for computational neuroscientists, is that models operating at the level of "neuron type" may or may not be using legitimate experimental findings to inform parameter value and model topology selection. Take the case of the pyra-

[1] http://senselab.med.yale.edu/modeldb/

midal cell, for example. The CA1 region of the hippocampus is populated with pyramidal cells, as is the CA3 region. Although pyramidal cells in both regions express glutamatergic NMDA receptors, however in each region the cellular transduction pathways regulating NMDA receptor expression differentially respond to Ca2+levels resulting in a down-regulation of NMDA receptors in CA3 have been shown to have a much less pronounced effect in CA1[114]. This example highlights both a potential problem for Computational Neuroscientists, as well as its solution–neurons should be defined in terms of their properties, and not solely by whether they are pyramidal cells, bipolar cells, etc. The Neuron Registry will provide a computationally-accessible framework for defining neurons in such a way, allowing modelers and bench researchers to have access to this information in a structured way that will not only be able to be incorporated into mathematical models, but will allow the research community to identify where there are gaps or contradictions in our current neuron-related knowledge.

The frequency and volume of newly-published scientific literature is quickly making the maintenance of publicly available scientific databases unrealistic and costly. In the Neurosciences, this problem is manifest in the maintenance of databases documenting primary-source experimental data on, for example, neuroanatomy or neuronal morphology. Many of these databases have become invaluable resources to bench and Computational Neuroscientists alike[35; 269]. The information contained in such databases is often based on user-submitted knowledge, and, despite the frequency with which they are used, it may or may not be possible for database owners to obtain sufficient resources to verify the information themselves. I propose that machine learning can provide a solution to this lack of resources. And, although efforts for maintaining our neuroscientific knowledge in databases have been underway for several years [58; 73; 74; 76; 163; 164], little work has been done in developing targeted Machine Learning and Text-mining methods for minimizing the human effort involved in their curation, making this a useful target of new research agendas. Document classification has been shown to be particularly useful in related fields. Supervised document classification is a machine learning technique in which a set of documents are manually labeled as positive and/or negative examples of some criterion of interest. An extension of this procedure, that we'll refer to as database submission classification, is related, in that it would, given a document and information that it is supposed to contain, automatically classify the submission as correct or incorrect.

Workflows for large-scale databases have many well-documented areas where machine learning methods can improve their efficiency [65; 66; 69]. Although the results found in related biological domains can often inform the application of document classification-like techniques to a new one, there are often domain-specific aspects that ought to be considered as well [9]. As such, it is important for both the machine learning and Neuroinformatics communities that such text mining methods be studied within the context of Neuroscience. For one, it is important for the Neurosciences, since it will allow text mining methods to be optimized specifically for performing within its feature space. For machine learning theorists, it will deepen our understanding of the relationship between the characteristics of a particular domain's textual feature space (i.e., the way experts discuss their field), and the relative performance of various text classification algorithms [9].

The main goal of the proposed studies is to design, build, and evaluate a system for automating certain aspects of the textual, neuron-related, community database curation workflow. The specific target-users of this system will be those involved in data curation at the Neuron Registry, however, the general framework used in these studies will be relevant to those involved in database curation for the Neurosciences in general. The Neuron Registry was chosen for several reasons, in addition to those already discussed [1.4.3.3]. First, biologically-sound multilevel models of neural circuitry will all necessarily leverage the information contained in a database of neuronal attributes. As such, the Neuron Registry stands to become an important source of modeling information in Neuroinformatics. Second, is the amount of already curated information contained in the Neuron Registry. The process of identifying and verifying new material for inclusion in the such a database is laborious, and so it is unsurprising that the Neuron Registry houses only a small amount of information to date. The small amount of data already contained means that it will greatly benefit from our efforts, in a way that an already established database might not. Despite this, preliminary studies indicate that it contains enough curated information to support a machine learning approach to identifying and prioritizing new documents for inclusion. Third, the Neuron Registry has adopted a collaborative approach to curation–an approach largely unstudied, in terms of how text-mining and machine learning can effectively aid its curation.

1.5 Key Contributions of this Dissertation

The results and methods used in the experiments described in this dissertation have made several key contributions to the fields of neuroinformatics, biocuration, and neuroscience.

1. The data set we developed to use in our experiments is unique, and can be used by neuroinformaticians and machine learning developers for their own experiments. This documents in this data set were obtained by constructing searches that a typical neurobiologist would perform to find articles of interest, and 962 of these were annotated at the document- and sentence level, in terms of their relevance to a knowledge base of interest, and is useful for developing information retrieval systems, supervised document classifiers, or information extraction systems. (Chapter 2.)

2. We have developed a general approach to bootstrapping the development of an under-represented knowledge base in the biomedical sciences. This technique will be useful to knowledge base developers who wish make efficient use of the time they have for identifying documents for inclusion in a knowledge base. Our approach, an adapted for of active learning, improves its accuracy as more information is given to it, over the course of knowledge base development, and will drastically decrease the amount of time spent reading irrelevant publications. (Chapter 2.)

3. We have developed a supervised document classification system useful for identifying publications that are relevant to a neuron-related knowledge base.

4. We have created several hand-curated regular expression lists for identifying NeuroLex and Methods-related terms in the text of publications.

5. We have developed a system for ranking paragraphs in a document, in terms of their likely relevance for containing information of interest for a knowledge base. Our general approach could be applied outside of the neurosciences, as the techniques used did not rely on neuroscience-specific information.

1.6 Thesis Overview

This thesis is comprised of three main chapters, each describing the development and evaluation of a text-mining-based tool designed to target a specific bottleneck in the curation workflow of a community knowledge base in the neurosciences:

1. *Virk* (Chapter 2): an active learning tool designed specifically with under-developed knowledge bases in mind, allowing them to quickly bootstrap their development.

2. *Flokka* (Chapter 3): an document classification tool that will rank-order documents, in terms of their likelihood of containing information that is relevant to some knowledge base.

3. *Finna* (Chapter 4): a paragraph ranking tool that, given a manuscript that is likely to contain relevant information for a knowledge base, will prioritize those paragraphs in terms of how likely they are to contain the information that would lead to being included in a knowledge base.

Taken together, these tools can enable the start-up and maintenance of knowledge bases in the Neurosciences, and constitute an important contribution to the fields of neuroscience, biocuration, and applied machine learning.

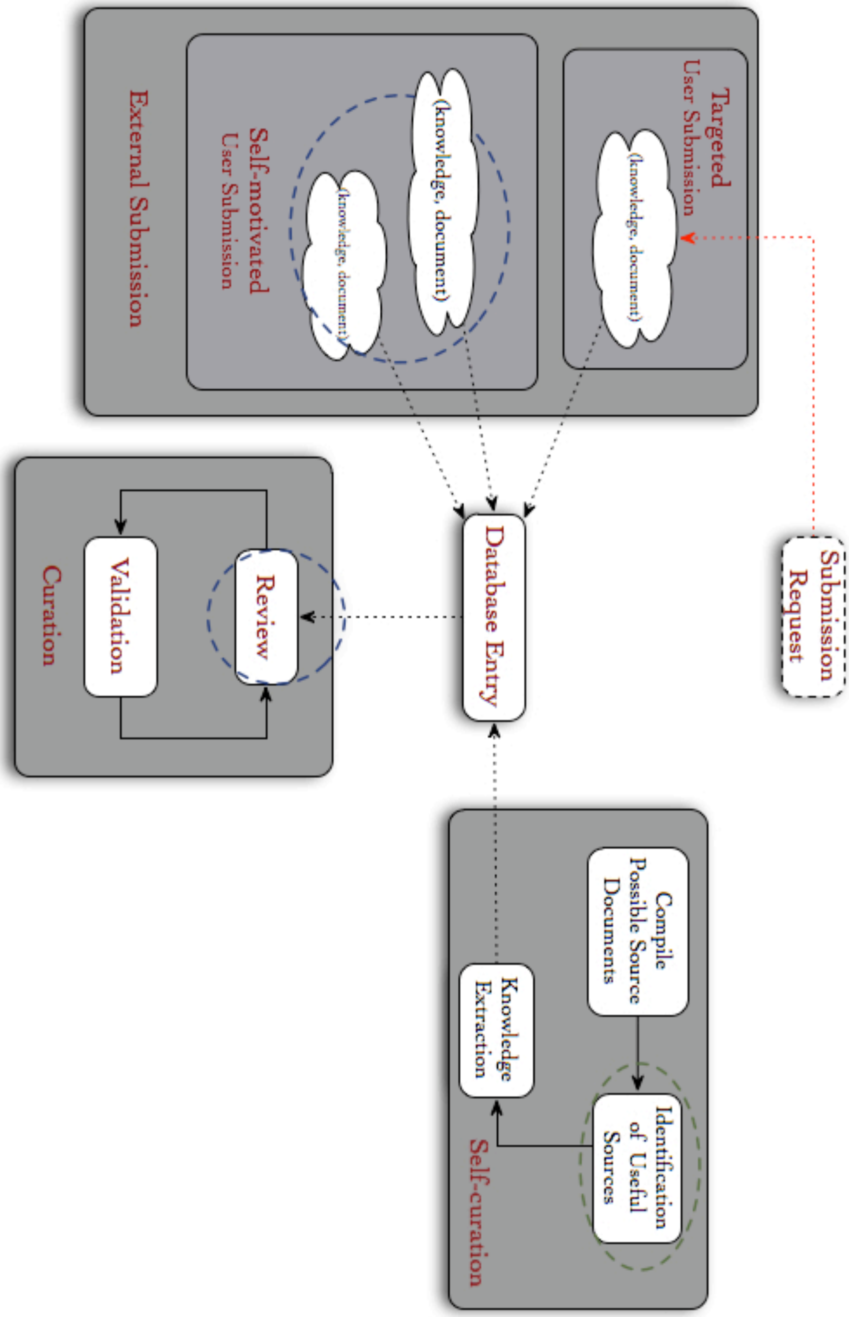

Figure 1.4: Diagram describing the submission/curation workflow at the Neuron Registry. Circled areas denoted points in the workflow which can give rise to bottle necks or inefficiencies. Blue: delay from lack of self-motivated user submissions. Green: bottleneck from needing to read through a collection of potential source documents, some of which will won't be used. Purple: bottleneck due to the process of identifying useful substring(s) in the source documents that haven't been triaged. Red: bottleneck due to validation of (knowledge, document) pairs, from either user submissions, or in-house identification.

Chapter 2

Virk: An Active Learning System for Bootstrapping New Curated Neuroinformatics Knowledge Bases

2.1 Introduction

In 2008, Howe and colleagues proposed that in the next five years the field of biocuration should create a mechanism by which community-based curation efforts can be facilitated[134]. This challenge has been taken on and applied in already successful online databases, such as FlyBase[1], the online repository of drosophila genetic information, and GrainGenes[2], a browser for the *triticeae* and

[1] http://flybase.org/

[2] http://wheat.pw.usda.gov/GG2/index.shtml

avena genomes, but is not so easily translated to new databases, which do not already have an active contributing community, or are able to support the professional staff to maintain it. As the neuroinformatics community begins to rely more and more on online resources containing structured information that can be used for large-scale mathematical modeling and simulation, the ability to efficiently create a new database and build it up to the point where it can be a useful resource to an active community will increase in importance, even as the amount of information that must be manually search through increases with exponentially-increasing rates of publication. It is because of this that the community must turn to automated techniques that have demonstrated their effectiveness in the field of machine learning. Such techniques are known as recommender systems [240].

The process of manual curation of data from published papers into underlying databases in an important (and mostly unacknowledged) bottleneck for developers of neuroinformatics systems. For example, the first version of the CoCoMac system (*Collations of Connectivity data on the Macaque brain*[1][272]) is a neuroinformatics database project concerned with inter-area connections in the cerebral cortex of the Macaque. It is a mature solution for a problem that was under consideration by national committees concerned, as far back as 1989 (L.W. Swanson, personal communication). CoCoMac currently contains roughly 2.0 $\times 10^4$ connection reports, reflecting the dedicated effort of a small curation team of the course of years of work. Due to the machine-readable nature of much of the data in their field, bioinformatics systems in molecular biology are usually larger by *several or-*

[1] http://cocomac.org

ders of magnitude. The Uniprot KB release for February 2013 contains 3.03 x 10^7 entries. Naturally, this disparity is due to many factors, including the levels of available resources for curation, the general utility of the data being housed, and the relative size of user communities. In any case, the rate-determining step for developing informatics systems of any size is the speed of curation, and so accelerating that process is an important central goal.

To mitigate the problems arising from having more information that needs to be annotated, but not necessarily a commensurate increase in available resources, biocurators should adopt automated approaches to identifying documents that contain information relevant to their particular knowledge base, such as active learning (AL) systems. An extensive review of AL methods is available from Burr Settles, [258] but, briefly, AL is a type of supervised machine learning (ML), in which a classification algorithm works collaboratively with an expert user to train a classifier as efficiently as possible (in terms of the expert's effort), by requesting gold-standard annotations for the data the AL system deems most informative. Such methods could be incredibly useful for neuroinformaticians starting up new knowledge bases, by helping them to efficiently create a publication recommendation system that would allow them to spend more time reviewing manuscripts that are likely to contain information relevant to their group's interests.

A substantial body of work has demonstrated the effectiveness of AL and recommender systems for efficiently developing a document classifier that has been incrementally trained on gold-standard data. Mohamed and colleagues, for example, used AL to develop a protein-protein interaction predictor [214], and Arens, in conjunction with a support vector machine (SVM) classifier, used AL for learning document ranking functions in a biomedical information retrieval task [16]. SVMs and AL have also been paired together for the identification of documents that are eligible for inclusion in a systematic review [292]. Here, Wallace and colleagues adapted previously-developed AL strategies for biomedical document classification, by taking into account the commonly-observed highly-skewed class distribution in such publications.

The Neuron Registry (NR) is a community-curated knowledge base under the direction of the Neuron Registry Task Force (NRTF[1]), a part of the International Neuroinformatics Coordinating Facility (INCF) Program on Ontologies of Neural Structures (PONS). The primary goal of the NRTF is to create the infrastructure for a machine-readable knowledge base of neuronal cell types, providing a formal means for describing and quantifying existing cell types from their properties, and populating it with information that has been extracted from the primary literature.

As a community-curated knowledge base, growth of the NR is contingent upon user submissions–the problem of adding new information to the system has been largely left to the people who use it. For knowledge bases that already have a strong user-base and an active community (e.g., Wikipedia), new submissions are frequently being made. This makes sense–Wikipedia is one of the more frequently-accessed web sites in the world; for a less-well-known resource, such as the NR (which has contributions from only 13 individuals, to date), some level of useful-

[1] http://pons.neurocommons.org/page/Neuron_registry

ness will need to be demonstrated before it becomes something researchers are regularly willing to submit new information [49]. Given the scope of information that is relevant to the NR, a great many more contributions will need to be made before it can be used as a reliable, machine-readable repository of our neuron-related knowledge. This is a common problem in informatics, in general, and neuroinformatics, in particular. New web-based resources are frequently created and made publicly available for use in others research. Initially, the creation and maintenance of such resources is often supported by the grant that lead to their starting up, but it is uncommon for funds to be available for the continued maintenance of a resource that hasnt already demonstrated meaningful contributions to the research community[10]. This can lead to the gradual decline of a resource, to the point where it is no longer a reliable, up-to-date snapshot of the community's knowledge. For example, this happened with the well-known CoCoMac database, which was unable to keep up with the pace of the published literature on Macaque connectivity beyond 2008, because of increasing rates of publication and limited resources (Kötter, personal communication; 2009). Thus, informaticians interested in creating accessible knowledge bases for the research community are left with a dilemma: how can they create a resource and deploy it with sufficient information useful to the community, without spending a great deal of time and money on curating the information they wish to include before the user community has been established? From the ML community, the answer to such a problem has been AL and recommender systems.

Although AL has been shown to be useful for identifying documents which will provide the most information to a supervised classification system, no one has yet used AL for simultaneously identifying new documents containing relevant information for a knowledge base while training a new document classifier for later use in updating the knowledge base. While the creation of a data set for training document classifiers is useful to the biocuration community (for a review, see [130]), it should also be possible to use AL to streamline the process of identifying initial documents that contain information that is of interest to under-developed knowledge bases which don't contain sufficient data to train an accurate classifier. There are two main hurdles to achieving this goal. First, to do this, a system will need to simultaneously identify documents that are likely to contain information of interest while identifying documents on which it cannot reliably make a judgment. Second, the existing methods for evaluating AL systems have been designed to work with a large, fixed corpus of data already annotated, which is not available for our purposes–no annotated full text corpus of neuron-related documents has been made available to the public. What's more, existing evaluation metrics primarily focus on the accuracy of the classifier being built and the rate at which it was able to achieve peak performance [258]. While these aspects of the proposed system are important, we are also interested in the point at which the maximum number of relevant documents are identified at the minimum amount of annotation effort–this trade-off does not exist in typical AL applications.

Here, we present and evaluate *Virk*, an AL system that is able to rapidly bootstrap knowledge base development. Over the course of the study we dramatically increase the coverage of the NR, which will make it possible for the knowledge base to be a more useful resource neuroscientists, and create a unique, publicly-available gold-standard document collection for the neurosciences that will be of great use to neuroinformaticians in the future. We describe a gold-standard-trained recommender system that was used to efficiently develop the NR, contributing important, expert-curated knowledge to the neuroinformatics community, and demonstrating for the first time that, with minimal effort spent on tuning a classification algorithm workflow, an AL system can provide meaningful contributions to the biocuration workflow. Importantly, our system is designed to specifically address to bottlenecks in the NR curation work-flow (Figure 1.4). As a tool primarily designed for bootstrapping the start-up of new knowledge bases, the Virk system will be helpful in the "Self-Curation" node depicted in Figure 1.4. New knowledge base developers can have a significant volume of publication they must review, but may have no prior information that can help them efficiently prioritize the order in which they do their reading. Virk will help biocurators by interactively re-ordering a list of publications in terms of their likely relevance, updating its judgements after receiving feedback from the curation staff.

2.2 Methods

2.2.1 Collecting a Full Text Neuron-related Document Set

Data collection proceeded in two main stages: journal selection and article selection. First, we determined which neuroscience-related journals to use to build our corpus. Our primary goal was to build a document collection that adequately represented the diversity of the neuroscience literature, so that our classifier would be exposed to articles about neuroimaging, computational neuroscience, and behavior, in addition to documents containing information on neuron-related experiments. At the same time, we wanted to obtain a sufficient number of documents containing information relevant to the NR. Thus, we downloaded the all entries in *vertebrate neuron* category on NeuroLex[1], an online, community-curated neuroscience lexicon, and found which journals most often include these terms within the MEDLINE records of the research articles they publish. We wanted to be certain that the articles we eventually included in our corpus had complete MEDLINE records and were representative of the sorts of terminology used in newly-published research, so we decided to limit our document selection to those published during the year 2010.

PubMed[2] queries were constructed for each selected NeuroLex term, each taking the form "NEURON", where NEURON corresponded to an entry in the NeuroLex (e.g., "*Amygdala basolateral nuclear complex pyramidal neuron*"). We rank-ordered the journals, in terms of their frequency of using NeuroLex terms in 2010. The top eight journals are listed in Table 2.1. We limited our selected journals to eight, because the ninth journal was the *Proceeds of the National Academy of Science*, which would have doubled the size of our corpus and would have also diluted the concentration of annotatable information by adding many non-neuroscience articles (e.g., geology or astronomy). Column n_{2010} in Table 2.1 shows the number of publications associated with that particular journal in 2010. The complete MEDLINE records for all 5932 articles included in Table 2.1 were downloaded and stored in a mongoDB database[3]. We chose a document-oriented database for this work because they allow us to efficiently represent MEDLINE

[1] http://www.neurolex.org

[2] http://http://www.ncbi.nlm.nih.gov/pubmed/

[3] www.mongodb.org

Journal	n_{2010}
The Journal of Neuroscience	1457
Brain Research	1267
Neuroscience Letters	1056
Neuroscience	910
The Journal of Physiology	495
The Journal of Comparative Neurology	276
Hearing Research	256
Journal of Neurophysiology	215
Total	5932

Table 2.1: Table of PubMed results by the top nine-matching journals in the data set, based on performing queries with the NeuroLex terms. n_{2010} denotes the number of 2010 publications associated with that journal.

records, which often have fields that differ from document to document. Since our University library had subscriptions to the eight journals of interest, we attempted to download the full text for each. Due to time constraints, we only attempted to download each of the 5932 full text documents once–if an error occurred, we simply skipped that document. In all, we were able to successfully download 3336 of the documents we were interested in. Since the approximately 2500 articles we were unable to download were distributed amongst all the journals of interest, and 3336 was likely to be an adequately-sized data set, no further attempt to retrieve these articles was made.

The 3336 documents were associated with their respective MEDLINE records in our database, and distributed into one of four document pools, according to Figure 2.1.

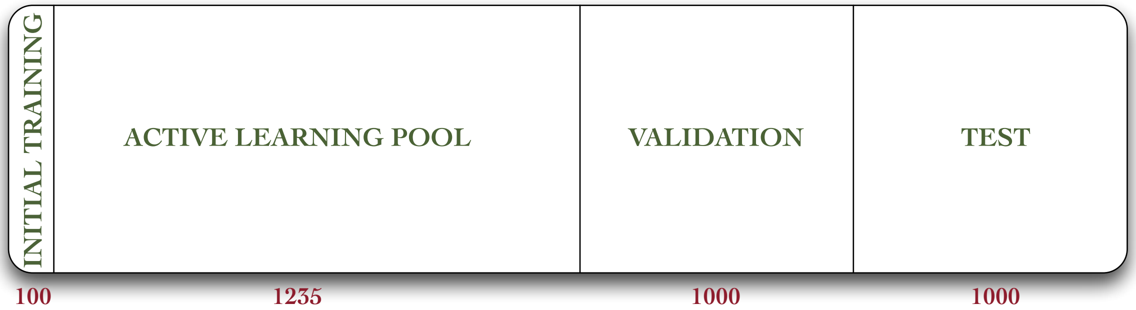

Figure 2.1: Diagrammatic representation of the distribution of documents in our corpus. Documents were randomly allocated to either the initial training, active learning pool, validation, or test collections.

2.2.2 Procedure for Annotating the Documents

One of the goals of this study was to build up a document classification data set for identifying publications that are likely to contain information that is relevant to the Neuron Registry (NR). The NR is a collection of neuron-related information, in the form of rows of (neuron type, relation, value) tuples (e.g., *CA1 pyramidal cell*, *located in*, *CA1 stratum oriens*), and an associated reference (e.g., a PubMed identifier). Previous work has shown that having well-defined annotation schema and criteria is important for building up a consistent document collection for machine learning. Thus, in collaboration with Giorgio Ascoli at the Neuron Registry, we developed an annotation schema where an article was marked *excluded* unless it could meet the following inclusion criteria:

1. The document appears in a peer-reviewed scientific journal.

2. The document is a primary source of the information in question (i.e., a primary, citable communication of the information in question (and not, for example, a review article).

Accepted Neuron Registry Neuron Relations	
Expresses protein	Does not express protein
Has molecule	Does not have molecule
Makes contact to	Does not make contact to
Receives contact from	Does not receive contact from
Located in	Not located in
Has current	Has firing pattern
Has part	Lacks part
Has orientation	Generates
Has mRNA transcript	Expresses gene
Lacks quality	Has quality
Has size	Has shape

Table 2.2: The list of accepted Neuron Registry neuron relation values used for creating the annotated document collection. The list was generated by examining entries already annotated into the neuron registry, and examining neuroscience publications.

3. The document contains all parts of the (Neuron Type, Relation, Value, Publication ID) tuple.

4. The Neuron Type and Relation identified in the document in question are found in the accepted set of values:

   (a) The Neuron Type must either map to one of the types listed on neurolex.org[1], or, if it's not included, a strong case must be able to made for needing to include it.

   (b) The relation must be an accepted NR relation type (see Table 2.2).

5. The (Neuron Type, Relation, Value, Publication ID) tuple must not already be included in the NR.

During the process described below, documents were annotated for inclusion in the NR, and, if it was determined the document ought to be included, the specific

[1] http://neurolex.org/wiki/Neuron_Types_With_Definitions

text which led to this judgment was extracted and the information was uploaded to the NR website.

2.2.3 Developing & Training the Baseline Classifier

100 randomly-selected documents were annotated for the Initial Training collection (Figure 2.1). We used these documents to conduct a set of classification experiments that would help us determine various aspects of the active learning classifier. We were interested in creating a classifier that could perform reasonably well without using many enriched-engineered features (e.g., named entity recognition, or part of speech tagging). We made an *a priori* decision to use a support vector machine classifier with linear kernel[90], using default parameter settings, as we have used this for a baseline classifier in previous work[8; 9; 63; 66; 68]. Moreover, part of the motivation for the present experiments was to create a document corpus that could be used for supervised document classification experiments down the road. If a more complex classifier using engineered features were used here, it's possible that the selections it makes could affect the results of those future experiments. Thus, to focus on the AL process, we aimed to use the simplest possible classification pipeline. We decided there were three aspects of our classification system we could optimize for the baseline classifier: input features, feature normalization, and modeling type. The input features we investigated were all combinations of MeSH terms, abstract unigrams, and title unigrams (as obtained from the associated MEDLINE record), and full text unigrams (as obtained from the above-described pdf-to-text extraction procedure). We considered two feature vector normalization techniques–one in which the fea-

tures from different sections of the document are simply combined into a single larger vector of greater dimension, and another which applies L2 normalization to the vector components. Finally, we considered two feature modeling methods: binary, and mutual information-based. In binary feature modeling, each document in the training collection is represented as an n-length vector of 0's and 1's, where n corresponds to the unique set of features in the training collection, and a given index in each document's vector maps to the same feature. When applying the training model to unseen data, then, only the features that were observed in the training collection are used to make the classification judgment–any previously unseen features are ignored. Mutual Information-based feature modeling is similar, however, the role of each index of the vector in this method corresponds to the mutual information of that feature for the classification problem at hand, where we define mutual information, or information gain, of feature j in all documents as

$$IG_{x_{.j}} = H(x_{.j}) - H(x_{.j}|y_i), \tag{2.1}$$

where y_i corresponds to the true class label y of document i, and $H(x_{.j})$ is the entropy of feature j in the collection of documents, defined as

$$H(x_{.j}) = -\sum p(x_{.j}) \log p(x_{.j}) \tag{2.2}$$

The results from our baseline classifier experiments are shown in Table 2.3. As is shown there, the two top-performing systems, in terms of the AUC observed on 5x2 cross-validation using the 100 manually-annotated documents from the Initial Training collection (Figure 2.1), used features from the title, abstract, and MeSH terms from each paper, and binary feature modeling using either

Feature Type				Normalizing		Default	
Text	Title	Abstract	MeSH	IG	Binary	IG	Binary
X	X	X	X	0.737	0.711	0.678	0.680
X	X		X	0.713	0.798	0.705	0.797
X			X	0.596	0.689	0.564	0.670
X	X	X		0.755	0.707	0.684	0.668
X		X		0.755	0.707	0.664	0.668
X	X			0.702	0.787	0.689	0.774
X				0.621	0.707	0.567	0.666
	X	X	X	0.739	0.807	0.746	0.808
	X		X	0.706	0.662	0.775	0.659
	X	X		0.804	0.650	0.745	0.607
	X			0.637	0.687	0.651	0.659
		X	X	0.675	0.797	0.802	0.801
			X	0.638	0.627	0.699	0.635
		X		0.751	0.640	0.726	0.600

Table 2.3: Summary of AUC results observed in the baseline classifier cross-validation (five repetitions of two-way; 5x2) experiments. An *X* in any of the four left columns indicates the inclusion of textual features from that portion of the document. The four right-hand columns show the AUC observed when using either the Normalizing Feature Combiner (columns five and six), or the Default Feature Combiner (columns seven and eight), along with either the Information Gain-based (IG) or binary modelers. The AUCs highlighted in red denote the two top-scoring system configurations. The top-performing baseline systems were both obtained from systems using unigrams derived from documents' title, abstract, and MeSH terms.

the default or normalizing feature combiners. The default mode for combining features involved simply combining all features into a binary vector, while the normalization approach involves also normalizing the length of the vector to 1. Since the difference between the two top-performing systems was small (0.001), we opted to use the simpler of the two–the one using default feature combining–for our active learning experiments.

2.2.4 The Active Learning Procedure

We had two simultaneous goals with our active learning system: to learn more about the efficiency of our ML approach in the domain of neuron-related documents, and to identify the greatest number of annotatable documents for the NR as possible, in the least amount of time, and with the fewest-possible total documents examined. In a typical AL text-mining experiment, ML scientists will use a corpus of documents that has already been annotated for a particular task (e.g., the Reuters Corpus, as in [37; 89; 186; 206; 255; 277]). This is because such studies are concerned with coming to a greater understanding of what classification approaches, modeling techniques, and input feature types lead to best classifier performance in an AL framework. Here, as is often the case with an under-developed or new knowledge base, we have no gold standard available to us. If such a corpus were available, we would simply train a document classifier using the data available, and use the classifier to identify newly-published documents that contain information relevant to the NR. Although it would be possible to create a classifier using the little data already in the NR, it results in a classifier trained to identify documents containing only a small set of neuron-related concepts (e.g., only documents containing, for example, the word *purkinje cell*). This is because, when only a small amount of data is available for a very broad and terminologically-rich field, such as neuroscience, there is not enough information available to create a general representation of the sorts of documents that one is interested in. Often times, the small amount of data available may be about a small set of concepts for which there has been a great deal of research (e.g., *purkinje cells*), making mentions of specific neuroanatomical features highly-predictive terms for a classifier built to identify documents containing information that is similar to those already included in the knowledge base. Ideally, a classifier would make its judgments based on more general concepts, such as methods that are often used in NR-relevant publications (e.g., patch-clamp), or observation-related words associated with those types of methods (e.g., current), but, to learn these types of associations, the classifier would need to be presented with many more examples of relevant and irrelevant documents than were available to the NR, (or, indeed, than often are available to many new knowledge bases). Here we create a method for using AL to bootstrap the development of a knowledge base while simultaneously training a document classifier. To our knowledge, this is the first-published method accomplishing this task.

A work-flow diagram of our annotation procedure can be seen in Figure 2.2. We trained our baseline classifier on the annotated Initial Training sample, and classified all 1235 of the documents in the Active Learning pool. We rank-ordered these judgments in terms of confidence, where a confidence of 1.0 is a document that our system is highly confident is one containing annotatable information, to 0.0, which is one the system is least confident the document contains annotatable information (or, most confident that it does not contain such information). There are a variety of active learning sample selection methods that have been used in previous work (for a review, see [258]). We chose this approach for its simplicity and efficiency. From the rank-ordered list, we identified 30 documents–the top 20 highest confidencethat were most likely to contain annotatable information, and the 10 that the classifier was least certain about (i.e., the 10 nearest to a confidence value of 0.5). These numbers of documents were selected because we thought they would provide the right balance of possible granularity in our performance metrics, while still being small enough that we could detect changes in classifier performance. These articles were then read in full, and annotated as being positive or negative for containing information relevant to the NR (the terms positive-class/relevant and negative-class/not relevant are used interchangeably in this manuscript). For those which were found to contain annotatable information, the relevant data was manually extracted and immediately uploaded to the NR. The annotated 10 uncertain documents were then added to the documents from the Initial Training sample (giving 110 annotated documents), the model was re-trained, and the remaining documents in the Active Learning pool were re-classified. The whole process was repeated for 20 iterations. We chose to only include the 10 uncertain documents, rather than using all 30 annotated at each iteration, in the data the model was trained on because we hypothesized that, while adding all 30 documents would likely help boost the model's performance, in terms of AUC, it wouldn't necessarily help us pick the most useful documents for classifier training, and bias the classifier toward positive prediction (see Discussion section).

This procedure highlights the dual purpose of this set of experiments: the 10 uncertain documents are akin to those that would be added at each iteration of a typical active learning experiment, while the 20 most likely relevant documents are identified at each iteration of our procedure so that we are more likely to identify documents containing actual annotatable information at each iteration.

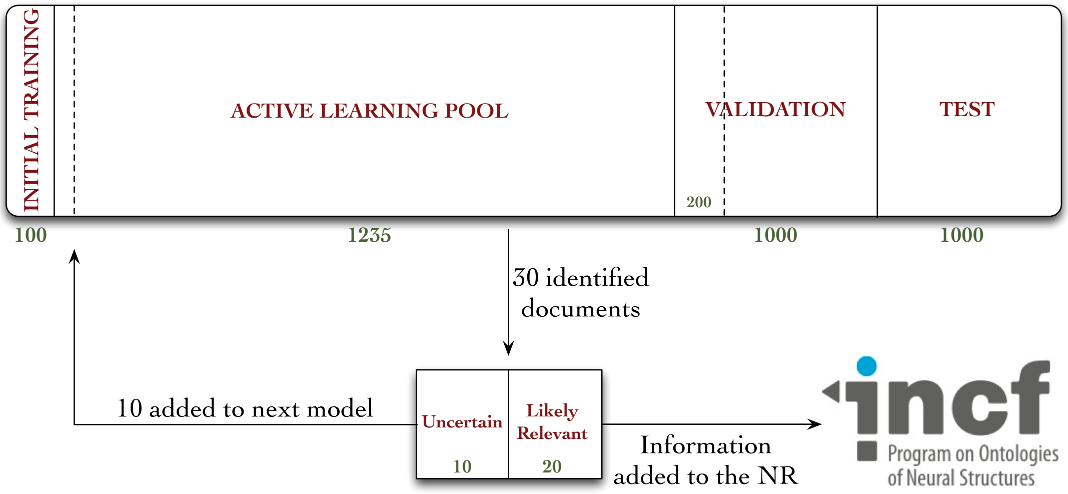

Figure 2.2: Work-flow diagram for annotation in the active learning experiment. A total of 962 documents were annotated during these experiments–670 during the active learning procedure, 92 during the random validation experiments, and an additional 200 for the hold-out test collection, used to evaluate the Virk system against the random system.

2.2.5 Evaluating the Dual Purpose Approach to Using Active Learning

We evaluated our general approach in terms of the change in classifier performance over active learning iterations, and the change in the ratio of the ML-predicted relevant to truly relevant documents. The former relates to the performance metrics often used in other AL and text classification studies–change in the area under the ROC curve (AUC) over learning iterations. We chose AUC as a performance metric because we were primarily interested in our system producing accurate rank-orderings, rather than completely accurate predictions–that is, it was more important to us that the top 20 documents that were predicted to contain annotatable information actually contained such information, than whether or not the SVM actually predicted them as belonging to the positive class. The later performance metric has to do with our goal of developing an AL system that is able to bootstrap knowledge base development by identifying publications that are likely to be relevant. We would expect that, if our system is able to accomplish this task, the number of truly relevant publications that it identifies would increase during the initial stages of training, level out for a time, and then begin to decline again, once the relevant documents in the AL pool collection become more rare.

In order to determine whether the classifier would be better off being trained by all 30 manually-annotated documents at each iteration, rather than just the 10 uncertain documents, we ran a set of experiments comparing the two possible approaches. To do this, we annotated an additional 200 randomly-selected documents from the hold-out validation pool. We trained a classifier using the documents that were selected up to each iteration in the original run of AL experiments, classifying the 200 validation documents using one of two methods–either a model which was trained using only the 10 uncertain documents identified at each iteration, or using a model trained on all 30 documents (the 10 uncertain documents, and the 20 predicted relevant ones).

Finally, we wished to compare the performance of our system against a randomly-performing system not using an AL document selection mechanism. To do this, we started a new active learning system from the same 100 documents selected to initialize the Virk system, and then randomly selected 10 documents from the active learning document pool over 19 sampling iterations. If a selected document had not previously been annotated, the document was annotated, and, if it was a positive-class document, it was used to add one or more annotations to the NR. To simulate the process of random sampling over 190 iterations (10 rounds of 19 iterations each), we randomly assigned each of these documents to one iteration in each round. At each iteration of both systems, the systems were trained on their annotated training data and evaluated against 200 randomly selected documents from the previously-described 200 documents that were annotated from the hold-out validation collection.

2.3 Results

100 randomly-selected documents were annotated for the first iteration of training, of which 8 were manually determined to contain NR-relevant information. Based on this, we inferred an 8% positive sample inclusion rate (4.4%-14.8%, 90% CI, based on the binomial distribution) in the larger population of potentially-included documents.

Over 20 iterations of AL, a total of 670 full text documents were annotated, over the course of four months, for containing information to include in the NR. Of those, 159 were identified for inclusion, with the remaining 511 being excluded.

Thus, after twenty iteration we observed a positive inclusion rate of approximately 24%–well outside the originally-projected inclusion rate of between 4.4% and 14.8%. This, of course, is to be expected from a system designed, in part, to identify positive-class documents. Figure 2.3 depicts the progress of our annotation and active learning procedure. From this figure, it is clear that there are

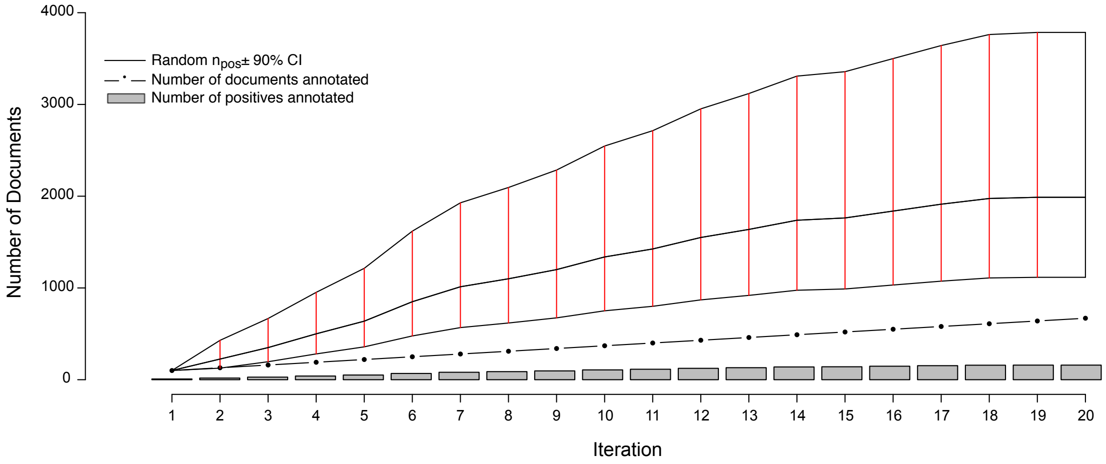

Figure 2.3: Performance statistics for the active learning system over iterations of document curation. The grey bars show the cumulating number of positively-annotated documents, while the black dotted line indicate the total number of documents annotated at a given iteration (increasing by 30 at each iteration after the first). The solid black line intersected with solid red lines indicates the estimated number of randomly-selected documents (±90% CI) that, at any iteration, would need to be annotated in order to obtain the same number of positive documents identified by Virk by that iteration. After three rounds of annotation, the average number of documents that would have to be read for a given number of positives is statistically significantly greater than that needing to be annotated with the Virk system.

different perspectives from which one can assess performance of our system–such as the rate at which positives are identified by Virk, and the savings conferred by Virk over the random system. One consequence of the system identifying a finite number of relevant documents from a fixed pool is that, as the number of positive-class documents in the pool depletes, the classification task becomes more difficult as the positive documents become more rare. To account for this, we evaluated our system in terms of an adjusted positive inclusion rate, defined as

$$AIR = \frac{\frac{n_{pos20}}{20}}{\frac{\hat{n}_{posRemaining}}{n_{Remaining}}}, \tag{2.3}$$

where n_{pos20} is the number of positive-class documents identified in a round of classification, and $\hat{n}_{posRemaining}$ is the number of positives estimated to remain in the active learning pool, based on the initially-estimated positive prevalence rate, and $n_{Remaining}$ is the number of documents remaining in the AL pool. This metric will adjust the fraction of positive-class documents found during one iteration of AL by the number of positive-class documents that are estimated to remain in the AL pool, thus accounting for the change in difficulty of the task at each iteration.

In order to evaluate the effect of only using the 10 uncertain documents to train our classifier (as opposed to all 30 annotated), we ran an experiment to compare the relative performance (in terms of AUC) of a system classifying a hold-out validation set of 200 documents using either a model trained on just the uncertain documents at each iteration, versus one trained on all annotated documents at each iteration. Of these 200 documents, 28 were found to contain relevant information for the NR, while the remaining 172 did not. This resulted in a positive-prevalence rate of 14%, which is within the 90% confidence interval of the original estimated positive-prevalence rate conducted at the outset of the study. The results of this experiment are shown in Figure 2.4. In terms of AUC, the system trained using the all annotated documents consistently out-performs the one trained using only the 10 uncertain documents at each iteration, though both systems begin to converge to similar values after 20 iterations. This implies that, despite the fact that the classifier is getting trained on a corpus of documents with a class-distribution different from that of the larger population, the extra information contained in the additional documents improved the ranking performance of the classifier. Importantly, however, this does not reflect the impact that this change in training samples might have on our ability to identify the most informative documents for subsequent training, based on the most uncertain predictions of the classifier. With the lowest-confidence sample selection method that we used here, including the 20 most confident documents in the training would have raised the proportion of positive samples in the training data. This likely would have biased the confidence estimates upward, leading to the system being trained with more negative documents and fewer positives. We will return to this issue in the Discussion section.

Prior to beginning this study, there were 235 entries in the NR, derived from 16 different journals, and submitted by 13 different authors. The majority of these submissions were added by the NR development team over the course of one year (between April 2010, and April 2011). Over the course of four months of annotation using Virk, an additional 257 annotations were added to the NR (more than the number of entries that it included prior to our work). This expanded the NR coverage of NeuroLex[1] neuron types from 16% to 55%. Using the class prevalences derived from our initial sample, using a random-selection approach, one would need to review between 160%-570% more documents than our approach

[1] http://neurolex.org

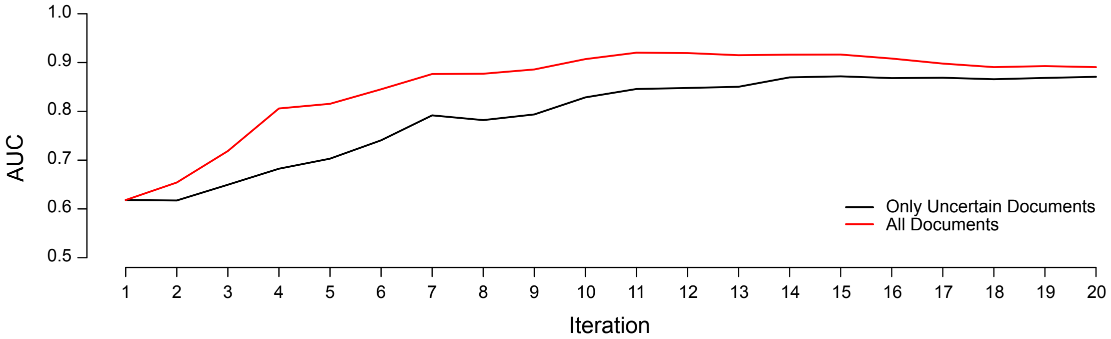

Figure 2.4: Relative performance of a system trained using only the uncertain documents at each iteration of active learning (black), versus all documents annotated up to that point (red), in terms of AUC. At Iteration 1, both systems are trained using the same data (the 100 initially-annotated documents), and thus score the same. After that, the system trained using all the data consistently outperforms the one using only the uncertain data, though both begin to converge to similar values after 20 iterations.

required (between 1116-3785 total documents).

To examine the validity of our selection approach and performance metrics, we compared the Virk system to that of a system using the data selected for the first iteration of Virk, and 10 randomly-selected documents at each of 19 iterations (if the document had not already been annotated during the Virk process, it was annotated and added to the NR, if necessary), using the system to classify the 200 hold-out validation documents every iteration. The random process was repeated 10 times, so that we could calculate confidence intervals.

The results of the random validation experiment are shown in Figures 2.5 and 2.6. To compare the Virk and Random Validation systems, we used area under the receiver-operator curve (AUC) obtained from training on the data available at a particular iteration, and classifying the hold-out validation set of 200 documents that were randomly selected from the Validation partition of the data set. As can be seen in Figure 2.5, for the first five iteration, the Virk system performs worse than random, although the performance differences are not substantially significant. By the sixth iteration, performance of the Virk system greatly exceeds that of the Random Validation system. Peak performance by Virk (AUC $\approx$ 0.87) appears to occur around iteration 14, while peak performance of Random Validation levels out by iteration twelve (AUC $\approx$ 0.72). Figure 2.6 compares the number of positive-class documents identified by the Virk and Random Validation systems. After 20 iterations, Virk identified 159 documents containing information relevant to the NR, whereas the Random Validation system only identified an average of 36 documents. After 4 iterations, our system exceeded the average number of positive-class documents identified using 20 iterations of random sampling. On average, the Random Validation system was able to identify approximately 1.5 positive-class documents per iteration (compared to approximately 8.0, by Virk). Thus, one would have to complete 106 iterations of annotation by random sampling to achieve what our approach was able to do in 20–a greater than 500% difference in work savings. These results demonstrate that Virk is able to quickly out-perform the standard document identification approaches used by biocurators today–our approach was able to identify significantly more relevant documents than the standard approach, and was able to do so with significantly less annotation effort.

While performance metrics such as AUC and number of positive documents identified are important to assessing a classification system, in the case of active learning, they do not necessarily tell the whole story. Besides being able to establish

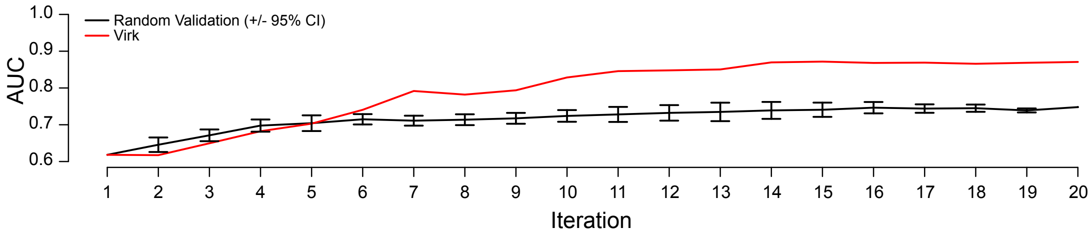

Figure 2.5: Performance evaluation comparing AUC of the Virk (red line) and Random Validation (black line) systems over iterations of active learning. The Random Validation system AUC was averaged over 10 random samplings, so that standard error could be calculated (bars). After six iternation, the Virk system outperform the 95% confidence interval for the random validation system.

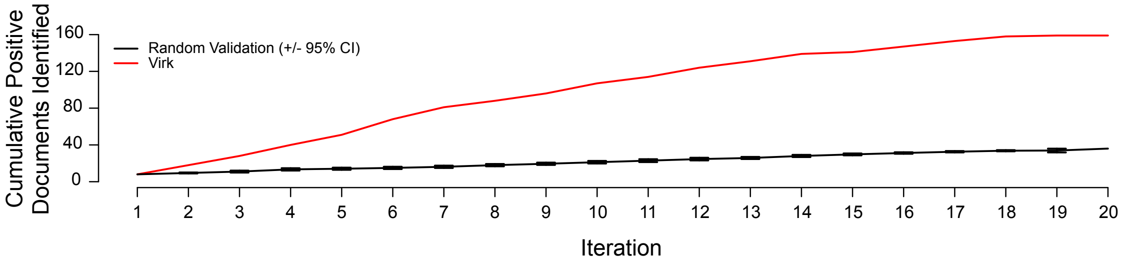

Figure 2.6: Number of positive-class documents identified over 20 iterations by Virk (red line) and Random Validation (black line, ± 95% confidence interval). After 20 iterations, the random validation system identified the number of positives found after only three iterations of the Virk system.

that our system can make accurate classifications, we also wanted to understand the trade-off annotators must make when deciding whether to use the system developed at any particular point or make additional annotations. To address this, we developed a metric called *goodness:work*, which quantifies the level of benefit obtained from accurate classifications made by the system relative to the amount of work that has been done in developing it up to that point. Biocurators could use a measure such as this to make informed judgments about when to stop data curation to train an active learning system, or to make informed decisions about how many documents would need to be annotated for future related systems. We define the goodness:work measure at iteration i as

$$(g:w)_i = \frac{n_{positivesidentified}/n_{totalpositivesestimated}}{n_{annotateddocument}/n_{documentsinALpool}} \tag{2.4}$$

goodness:work over iterations is depicted in Figure 2.7. goodness:work steadily increases, until it is maximized around iteration 7, where it remains stable for eight more iterations before beginning to drop around iteration 15, likely because, at that point, the number of positive documents remaining in the AL Pool has decreased enough that they are more difficult to find.

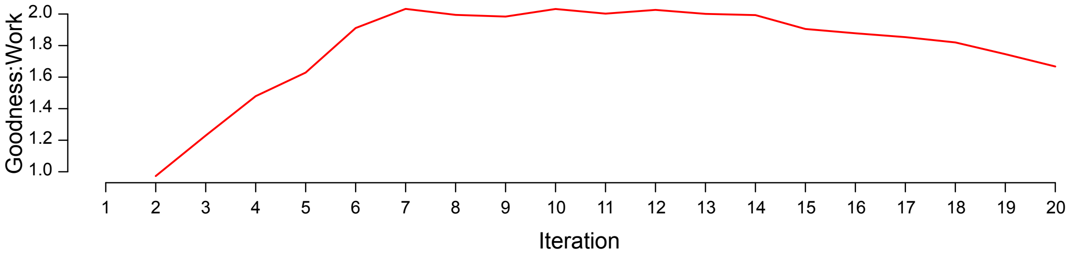

Figure 2.7: goodness:work ratio over iterations of active learning. No data exists at the first iteration because no active learning has yet taken place. Between iterations 2 and 7, the goodness:work ratio increases, being approximately level until iteration 15, where it begins to decline.

To better understand the contribution of different features to performance across iterations, we created ranked lists of the highest information gain features at each of the twenty classification tasks. Information gain was calculated according to

$$I(X|Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{2.5}$$

where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and

$p(y)$ are the marginal probability distribution functions of X and Y, respectively. The top twenty information gain-scoring features at each iteration of the active learning procedure are shown in Figure 2.8, where the word in the first position is the one that was the most informative for classification on that particular round, and its color corresponds to which part of the document meta data it came from (either title (black), abstract (red), or MeSH term (blue)). As the figure shows, the predictive features used at the first iteration of AL are a mix of cell-related terms (e.g., *ganglion*), relations (e.g., *expressed*), and stop words (e.g., *at*, or *than*). After the first iteration, stop words (defined here as terms unrelated to biological research, such as *at*, and *in*) are hardly used, and more cell-related terminology begins to show up. By the final iteration, a stable selection of features has been identified, coming from the title, abstract, and MeSH terms. *Patch-clamp*, for example, is a method used to study ion channels in the cell, and could be used to collect data for a variety of types of NR submissions. The presence of *ganglion* (likely, from dorsal root ganglion cell) and *purkinje* (from purkinje cell) are not surprising either–both of these cell types have been extensively studied in the literature.

2.4 Discussion

The creation and maintenance of machine-readable public repositories of scientific information is an increasingly important and interesting area of informatics research. In this work, we demonstrate the utility of AL systems for aiding in the community-curated database biocuration workflow. Using a simple approach, incorporating binary feature modeling of documents title, abstract, and MeSH terms, and an SVM classifier, we developed a system that shows how AL can be used by biocurators to quickly add annotated resources to an under-developed database while simultaneously training an ML classifier for later large-scale use. For completeness, here, we assumed that no previously-curated data was available for the initial training data. Its important to note that, for many knowledge bases, this would not be the case–some small amount of seed training data may already be available to biocurators, making the start-up costs of using our approach quite small.

Our research was able to make several important contributions to the NR, as well as to the neuroinformatics community at-large by improving several aspects of the NR knowledge base–we increased the number of entries by a factor of 90%, we increased its coverage of terms from the NeuroLex ontology, and we created an expert-curated neuroscience document collection that has been manually according to whether each contains information that is relevant to the NR (i.e., it contains curatable information about neuron-related phenomena). According to our calculations, based on annotation rates observed in this study, we were able to make approximately 1-2 years-worth of annotation contributions to the NR in only four months. The first two of these contributions will help make the NR into a resource that can be used by multi-level modelers in neuroinformatics (e.g., researchers interested in integrating information across multiple levels of the brain), as well as researchers wishing to find cited information on various aspects of the central nervous system at the neuron level. By improving the coverage and depth of the NR, we have helped increase the visibility of the NR, and therefore increase the likelihood that it will obtain new users and experts who will be willing to add their own contributions to the database.

The 962-document manually-curated collection that was created to run our experiments is also an important and usefulcontribution. As the neuroscience literature base continues to grow, methods for information retrieval and information extraction will continue to increase in importance, to neuroscientists and neuroinformaticians alike. Many such resources are trained on expertly-curated document collections that can be expensive to obtain. Here, we created a large collection of documents for training supervised algorithms. Our document collection has been annotated at both the document (i.e., relevant v. irrelevant) and sentence level (the sentence(s) containing the information that led to a *relevant* judgment), so, in addition to being useful for training document classification systems, it could be used for training structure classification algorithms for information extraction. No such resource was previously available for neuroinformatics.

One limitation of the presently-described experiments is that only one curator (KHA) was used to assign inclusion/exclusion judgments to documents in the seed training collection, active learning pool, and evaluation collection. Although the curator has a graduate-level education in the neurosciences, and the annotation criteria used was developed in collaboration with a professional biocurators and the NR staff, it would have been preferable to use a team of curators to assign labels, so that inter-annotator agreement statistics could have been calculated. Despite this limitation, all the annotations submitted over the course of this study have been reviewed and accepted by the NR curation staff. Along similar lines, in order to simplify the full text document acquisition step of our experiments, we limited our training data collection to only articles published in 2010, in nine top neuroscience journals. We made this choice in order to ensure that adequate metadata would be available in MEDLINE, and to maximize the number of articles that we would encounter that would be relevant to the NR, while still being true to the diversity of the neuroscience literature base as a whole. Although this wasnt an especially limiting assumption, future work could extend that presented here by taking a larger, more chronologically diverse slice of neuroscience publications. This would enable future work to look at the role of concept drift (e.g., [93]) in the performance of AL and recommender systems for biocuration.

Similar to others [8; 66; 67; 165; 166; 279; 292], we used AUC and ranking to evaluate the system and prioritize the literature for annotation work. Another potential area for future research lies in optimizing the number of uncertain documents used to train the classifier at each iteration (here, 10 documents) and the number of predicted-relevant documents used for annotation and evaluation (here, 20 documents). Although our values were methodically selected, according to the characteristics of the task we wished to perform, a more principled approach would be to derive these values based on the characteristics of the corpus from which relevant documents are being drawn, in order to both optimize the rate of training as well as the rate of accumulation of annotated articles. For example, in corpora where relevant documents are more prevalent, it might make more sense to increase the number of top-ranked documents drawn at each iteration for annotation into the knowledge base.

The results of our experiment comparing a system trained with only the uncertain documents to one trained on all the available documents are intriguing. While the system trained on all the available annotated documents consistently outperformed the system trained only on the uncertain documents in terms of AUC, this comparison is incomplete. An important part of our system is the method used to select the most informative documents for manual annotation and addition to the training set in the next iteration. In our system, we used the simple approach of choosing the 10 documents for which the classifier had the most uncertainty - the documents with the lowest confidence in their predictions. This enabled us to keep the prevalence of the training set approximately equal to that of the document pool. If we had included all of the annotated documents in each round of training (both the uncertain and confident documents), the training set would become gradually more and more skewed towards the positive documents and therefore our simple approach of selecting the documents with the most uncertainty would also be subject to this bias. It is unclear what the impact would be on the performance of our system in this situation. More sophisticated means of choosing the most informative documents could avoid this problem and allow training on all of the annotated documents at each stage without risk of biasing the classifier. However, these methods tend to be much more computationally and algorithmically complex than the simple method that was effective here[91; 258; 277].

Another possibility is to start off training on only the annotated uncertain documents and after some number of iterations switch to including all of the annotated documents. An avenue for future investigation will be to examine a priori methods for identifying the point at which this switch should be madein our studies, based on Figure 2.4, it appeared that this point occurred somewhere between iterations 10 and 14, but this may have been influenced by some aspects of our experiments (e.g., class distribution, or the number of documents used from training at each iteration). Finally, although our system is adept at expanding an online knowledge base (one bottleneck in the workflow of a community-curated database), it does nothing to address other inefficiencies, such as identifying likely erroneous submissions, recognizing newly-published articles that contain information of interest, or identifying where in an article the annotatable information could be found. Each of these, however, should be points of focus for future work.

2.5 Conclusion

In the present study we have demonstrated an approach to bootstrapping the development of knowledge bases with active learning. Using a simple support vector machine classification system, Virk was able to efficiently aid document discovery for biocuration of the Neuron Registry. Over the four months our system was used, we were able to increase the coverage of the Neuron Registry, more than doubling the number of MEDLINE-indexed references included in the knowledge base–an addition which would have required between 1-2 years by standard approaches used today. In addition to contributing to the content of the Neuron Registry, our approach resulted in a highly-attuned document classifier which will be used in future studies to identify newly-published relevant documents. The 962 annotated document collection created to conduct the experiments we have presented is also an important contribution. High-quality sentence-level annotations are not commonly available in the general biomedical informatics field and rare in neuroinformatics. To our knowledge, ours is the only one of its kind, and, to encourage additional work on text classification in the neurosciences, we have made it available as supplementary data.

Title Abstract MeSH

Rank (IG)	1	2	3	4	5	6	7	8	9	10
1	ganglion	expressed	cells	cells	cells	cells	cells	cells	cells	cells
2	since	retinal	purkinje	purkinje	ganglion	ganglion	humans	humans	ganglion	ganglion
3	have	mouse	purkinje	purkinje	retina	humans	patch-clamp	ganglion	humans	humans
4	bipolar	cells	purkinje	ganglion	humans	retina	ganglion	animals	patch-clamp	granule
5	subsequently	ganglion	located	purkinje	purkinje	animals	animals	patch-clamp	granule	ganglia,
6	retinal	located	expressed	expressed	purkinje	amacrine	currents	retina	currents	animals
7	muller	throughout	ganglion	animals	animals	purkinje	potentials/drug	granule	ganglia,	patch-clamp
8	peptide	putative	provide	humans	amacrine	purkinje	retina	currents	animals	currents
9	than	retinal	encoding	target	cell	retinal	patch-clamp	ganglia,	patch-clamp	ganglion
10	mouse	mice,	putative	retina	retinal	patch-clamp	amacrine	postsynaptic	techniques	patch-clamp
11	gene	proteins/genetics/*metabolism	amino	amacrine	purkinje	cells	granule	amacrine	retina	amacrine
12	their	cell	retina	encoding	layer	granule	purkinje	patch-clamp	amacrine	retina
13	expressed	expression	retinal	putative	cells	cell	purkinje	purkinje	purkinje	root
14	specific	purkinje	cell	cells/metabolism	retina	patch-clamp	ganglia,	techniques	whole-cell	cells
15	receptors	purkinje	acid	cells	cerebellar	ganglia,	cells	postsynaptic	cells	purkinje
16	identified	purkinje	animals	provide	ganglia,	layer	postsynaptic	antagonists/pharmacology	ganglion	drg
17	expression	types	cells	layer	ganglion	retinal	postsynaptic	cells	antagonists/pharmacology	techniques
18	at	mouse	glutamate	cell	identified	currents	antagonists/pharmacology	expression	overall	rat
19	via	target	glia	after	encoding	purkinje	techniques	potentials/drug	purkinje	inner
20	adult	since	specific	expression	rabbit	adult	potassium	inner	cell	in

Rank (IG)	11	12	13	14	15	16	17	18	19	20
1	cells	ganglion	ganglion	ganglion	ganglion	ganglion	ganglion	ganglion	ganglion	ganglion
2	humans	humans	humans	humans	humans	currents	currents	currents	currents	currents
3	ganglion	cells	cells	cells	cells	humans	humans	root	root	root
4	animals	animals	animals	currents	currents	animals	root	humans	patch-clamp	patch-clamp
5	granule	granule	granule	animals	animals	cells	patch-clamp	patch-clamp	cells	cells
6	patch-clamp	ganglia,	patch-clamp	ganglia,	ganglia,	ganglia,	animals	cells	humans	humans
7	ganglia,	patch-clamp	ganglia,	root	root	root	granule	ganglia,	techniques	techniques
8	in	in	currents	granule	granule	patch-clamp	cells	postsynaptic	postsynaptic	ganglia,
9	currents	root	patch-clamp	ganglion	patch-clamp	granule	ganglia,	techniques	ganglia,	postsynaptic
10	ganglion	ganglion	root	patch-clamp	ganglion	postsynaptic	postsynaptic	animals	ganglion	ganglion
11	patch-clamp	patch-clamp	in	pyramidal	patch-clamp	patch-clamp	techniques	ganglion	animals	animals
12	rat	rat	cell	patch-clamp	pyramidal	techniques	ganglion	granule	granule	granule
13	cells	drg	ganglion	cells	postsynaptic	ganglion	slices	drg	potentials/drug	slices
14	amacrine	currents	cells	postsynaptic	cells	slices	pyramidal	slices	slices	drg
15	retina	the	drg	techniques	techniques	pyramidal	patch-clamp	potentials/drug	drg	neurons
16	the	of	postsynaptic	drg	drg	drg	drg	pyramidal	patch-clamp	cell
17	of	amacrine	amacrine	amacrine	hair	potentials/drug	potentials/drug	neurons	cell	purkinje
18	root	retina	retina	retina	retina	pyramidal	neurons	patch-clamp	neurons	pyramidal
19	purkinje	cells	the	hair	amacrine	postsynaptic	pyramidal	pyramidal	purkinje	potentials/drug
20	techniques	cell	of	pyramidal	hair	retina	rat	purkinje	pyramidal	patch-clamp

Figure 2.8: The top 20 rank-ordered features, in terms of information gain, over iterations of active learning. The color of the text denotes which section of the document's associated MEDLINE record the term came from: either title (black), abstract (red), or MeSH (blue). Certain terms, such as ganglion are found across many iterations, though its position in the rank-ordered list changes, while others, such as "via" were less informative, appearing only in the first iteration.

Chapter 3

Flokka: A Document Prioritization System for Curating Databases in the Neuroscience

3.1 Introduction

Developing effective automated methods for streamlining biocuration workflows is essential to the continued advancement of Neuroinformatics. Although databases documenting the current state of our brain-related knowledge are currently under development, they contain only a small fraction of the information comprising the current state of our brain-related knowledge. What's more, as Neuroinformatics as a field continues to rely on multi-level modeling as a strategy for understanding the complexities of the brain, computationally-accessible databases of

Neuroscience information extracted from the primary literature will become more and more important. In order to make these databases into effective resources for mathematical models and Neuroscience researchers alike, we need to develop strategies for stream-lining biocuration work-flows that rely on text-mining.

Although using text-mining for document prioritization is not new, there are several complications associated with using such a system for curating a database with information derived from Neuroscience manuscripts. First, the language used to discuss neuroanatomy is quite heterogeneous. Neuroscience is made up of scientists from a variety of backgrounds (e.g., Biology, Chemistry, Physics, Computer Science), each of which has its own writing conventions and ways of discussing neuroanatomy. Second, Neuroscience research is done in a variety of species–the manner by which the different parts of the brain have been parsed into structures and substructures frequently differs between these species (e.g., birds have a region known as Area X, which is thought to be a homologue of the basal ganglia, in mammals [86]).

The overall goal of this set of studies was to identify a combination of classification algorithms and feature generation methods that can lead to good performance in a document ranking system for prioritizing manuscripts in terms of their relevance for the Neuron Registry (NR), a community-curated database of neuron-related information. *Flokka*, our system, addresses a specific problem in the NR curation work-flow that is also manifest in other community-curated knowledge bases. Figure 1.4 depicts the NR submission/curation work-flow. We have developed the Flokka system to help both external and internal submissions, as well as the curation process itself. First, this tool will be able to determine the likelihood that a publication submitted as a reference to a new entry contains information of relevance to the NR, which could be used as a warning system for flagging erroneous submission. Second, this tool can be used by internal and external sources alike, to identify newly-published manuscripts that will likely contain information of interest to the NR. Finally, this system can be used to prioritize documents, in terms of their likelihood for containing relevant information, which can be applied to internal development of the NR, when curators need to read through a large set of publications in a short amount of times.

3.2 Methods

3.2.1 Obtaining the Document Collection

For the set of experiments described here, we used the annotated document corpus described in [12] (see Chapter 2). For a full description of the data collection methods, we refer the reader to that chapter. Briefly, the documents used in this study were selected from the articles published in the 8 neuroscience journals (see Table 2.1) in 2010 using an active learning procedure. Documents in this study were assessed for their relevance to the Neuron Registry[1] (NR), and assigned either a positive-class or negative-class label, according to whether they contained information that could be included in the NR. 762 documents were randomly assigned to either the training collection (572 documents: 118 positive-class, 454 negative-class), for developing and comparing various system configurations, or

[1] http://incfnrci.appspot.com/

the hold-out testing collection (190 documents: 43 positive-class, 147 negative-class), for verifying the performance of our chosen systems on previously-unseen data.

3.2.2 Evaluating the Classifiers

In order to identify the best set of classifiers, we evaluated several classification algorithms with five repetitions of 2-way cross-validation, in terms of area under the curve (AUC), and with a common set of features–unigrams constructed from each document's title, abstract, and MeSH terms. We investigated a range of parameterizations for a variety of classification algorithms and implementations, including support vector machines (SVM; [142]), k-nearest neightbors [2; 9], centroid classification [116], logistic regression [299], adaboost with decision stumps [221], naïve bayes [152; 153], decision trees [235], random forests [184], star [96], and winnow [188].

Our previous work [10; 62; 63; 66; 70] has shown that SVMs are particularly effective for classifying biomedical text, however we have never directly examined the differences between different implementations of the SVM algorithm on the same data. Thus, here, we also compared performance two implementations of the linear support vector machine algorithm: SVMlight, and Weka. In addition, since many of the classifiers have parameters which need to be optimized (e.g., polynomial svm, and random forests), we investigated a range of parameter values using cross-validation. In an ideal situation, we would have had a separate hold-out data set for parameter optimization, however, since the data set used in these experiments is not large, we opted to just used the cross-validation set for parameter optimization. For classifiers where a broad range of parameter values showed equal, or near equal, performance, we selected a set of parameters settings lying in the middle of the equivalent range. The classifiers with different parameter values, and the ranges we investigated, are shown in Table 3.1.

Classifier	Parameter(s)	Range(s)	Selected Value(s)
SVM (Linear Kernel)	C	$1\ x\ 10^{-9} \to 1\ x\ 10^{6}$	1.0
SVM (RBF Kernel)	γ	$1\ x\ 10^{-4} \to 10$	0.02
SVM (Sigmoid Kernel)	S	$1\ x\ 10^{-3} \to 5$	$1\ x\ 10^{-3}$
SVM (Polynomial Kernel)	d, S	1-5, $1\ x\ 10^{-5} \to 0.09$	4, $7\ x\ 10^{-4}$
Random Forest	n_{trees}, $n_{features}$	$1 - 100$, $1 - 512$	74, 32
Adaboost DS	$n_{iterations}$	$1 - 700$	300

Table 3.1: Classifier parameter settings investigated in our study. Many of the algorithms we used did not have parameters needing to be optimized, and so were excluded from this list. For those here, we selected a logical range of possible values based on our previous experience with using these classifiers.

We ran our set of cross-validation experiments so that we could choose a group of classifiers with which to evaluate against a variety of features, in a later experiment. To do this, we had several criteria: performance (in terms of AUC), diversity (so that we could evaluate how different types of classifiers perform against different types of input features), and training and classification speed.

3.2.3 Feature Generation Methods

Once we selected a group of classifiers using the baseline set of input features, we investigated more complex features derived from neuroscience-related terminology. Since many of these approaches resulted in quantitative features–as opposed to the previous experiments, which used binary features–we had to incorporate a method for representing these continuously-valued features in the language of binary feature vectors, which are used by our classifiers. We will describe these methods, the All Feature Value Sum Normalizing Modeler (AVSM) and the Recursive Partitioning Summing Modeler (RPSM) following our description of the quantitative features we investigated.

3.2.3.1 Methods for Generating Continuously-Valued Features

3.2.3.1.1 NeuroNames: Quantifying Neuroanatomical Term Mentions

To determine the contribution of the presence of neuroanatomical-related terminology to whether a document contained information relevant to the NR, we created a list of regular expressions from the NeuroNames terminology [42; 43; 200], and incorporated two features based on these into our experiments–the total number of NeuroName mentions in the full text and MEDLINE records of the input data, and the total number of unique NeuroName mentions (e.g., if the term *Amygdala* was mentioned 20 times in a document, this we result in one unique count and 20 total counts).

3.2.3.1.2 NeuroLex: Quantifying Terms from a Neuroscience Ontology The vertebrate neuron branch of the community-curated ontology NeuroLex[1] were used to create regular expressions for identifying the presence of each term within the full text of the training documents. Initial investigation showed that the only term which was directly identifiable in the training corpus was *retinal ganglion cell*, which occurred multiple times in both positive- and negative-class documents. This is not entirely surprising–Neuroscience is a het-

[1] http://neurolex.org/wiki/Vertebrate_Neuron_overview

erogeneous research research discipline that is composed of investigators from a variety of academic backgrounds, and each has its own particular way of referring to neuroanatomical entities and writing conventions. For example, *amygdala basolateral nucleus stellate neuron*, could be referred to in a publication as *stellate neuron in the basolateral nucleus of the amygdala*, which would not be matched by our approach. Because manually creating a list of regular expressions covering all the possible ways each of the 247 neuroanatomical entities in the NeuroLex, we instead hand-selected substrings from the NeuroLex entries which would likely be found in whatever way someone used to refer to the concept. For example, the *amygdala basolateral nucleus stellate neuron* became *stellate neuron*, and the *retinal ganglion cell* became ganglion cell.

3.2.3.1.3 Methods: Quantifying Analytical Technique Mentions Regular expressions were manually constructed that would enable us to identify methods-related terms mentioned within the text of articles' title, abstract, MeSH, and figure captions. To determine which methods should be represented, the full text of 10 positive-class documents from our training collection were examined for the mention of methods-related terms, and regular expressions were constructed for each that enabled fuzzy matching (e.g., the presence or absence of hyphens, etc.). To expand this set, the terms were searched for on the MeSH Heading Term browser[1], and relevant methods-related terms from higher up on the tree were added to our list, where those occurred (e.g., after search for the term *Immunohistochemistry*, we examined terms higher up in its hierarchy, from the *Immunologic*

[1] http://www.nlm.nih.gov/mesh/MBrowser.html

Techniques category). Additional category subsections were also examined based on the author's (KHA) expertise in neuroscience-related methods. This approach led us to add terms from the following MeSH categories:

1. Analytical Chemistry Techniques
2. Cytological Techniques
3. Genetic Techniques
4. Immunologic Techniques
5. Microscopy
6. Molecular Probe Techniques

Finally, we investigated whether there was useful information to be found in the average and minimum word-based distances between these types of quantitative feature terms. Our thinking was that, since a valid entry for the NR includes an entity, a value, and a relationship, in the full text of documents, these will frequently be expressed as sentences describing some finding, and the method by which it was discovered. Thus, we investigated the mean and minimum document-level distances between methods-related terms and NeuroNames terms, and methods-related terms and NeuroLex terms. For example, a full text sentence like "*Neurons in the Amygdala were investigated using standard histological methods*", would have a 5-word methods/NeuroNames distance, and a 8-word methods/NeuroLex distance.

3.2.3.2 Full Text Features

Since full text data was available to us, we wanted to evaluate whether using features derived from the full text document contributed performance gains over-and-above those obtained from using the MEDLINE record alone. To use the full text in our classification workflow, we needed to convert the *.pdf* representation of the manuscripts to something plain text. For a full description of our approach, see [12] (Chapter 2). Briefly, we processed the previously-obtained pdf records using the *pdftotext* function in the xpdf library[1], and cleaned up the resultant plain text with a manually-constructed set of regular expressions.

In using our full text features, we wanted to investigate two hypotheses: that using text-based and quantitative features derived from the full text of a manuscript (as opposed to just the MEDLINE record) builds a more accurate classifier, and that only using features derived from full text paragraphs that are likely to contain important information leads to improved performance over-and-above using the entire set o full text features. To investigate our first hypothesis, we conducted a set of classification studies using a combination of feature types identified as potentially useful in our previous experiments. For that latter, we hypothesized that the text in table and figure captions of manuscripts would be more likely to improve classification than that elsewhere in the manuscript. Using our training data set, we examined which parts of the manuscript the annotatable information was found. Table 3.2 summarizes these results. Here, the collection of annotated information obtained from [12] (see Chapter 2) was manually verified by cross-referencing all annotations with the original *.pdf* of the respective manuscripts.

[1] http://www.foolabs.com/xpdf/

Most annotatable information in the training collection was obtained from publication abstracts–something already used in the classification systems not using full text information. The next-most-common locations were the Results subsection and figure captions–neither of which are used in MEDLINE record-only classification systems. Although more annotatable information was found in the results subsection, the information density of a figure caption is likely to be higher–that is, the ratio of useful text is likely to be higher here, which could potentially lead to improved classification performance. Thus, we created a regular expression for extracting figure captions from our pre-processed plain text manuscripts, and used the extracted text for a set of experiments mirroring the full text ones described above.

Subsection	Count
Abstract	58
Results	40
Figure Captions	34
Title	11
Methods	5

Table 3.2: Table describing the locations of annotatable information in the full text. This table was manually-extracted from the annotation data obtained in [12] (see Chapter 2), and the full text locations were individually verified by cross-referencing with the original *.pdf* representation of the manuscript. As we hypothesized, many annotations could be identified in the abstracts of the documents we reviewed, with the Results sections and figure captions having many annotations as well.

3.2.3.3 Quantitative Feature Modeling Approaches

The above-described quantitative features were modeled as set of features for a binary feature vector using the AVSM and RPSM techniques.

3.2.4 Train/Classify Experiments

We selected a group of classifiers and feature generation methods for further evaluation against our hold-out testing data set. Classifiers were selected based on performance and diversity of approach, while combinations of feature generation methods were selected solely based on performance both in the presence or absence of full text information (i.e., Full text + MEDLINE record versus MEDLINE record alone). Based on our classifier experiments, we selected five classifiers to evaluate against the hold-out testing data: Linear SVM ($C = 1.0$), Naïve Bayes, kIGNN, Random Forest ($n_{trees} = 74$, $n_{features} = 32$), and AdaBoost with Decision Stumps ($n_{iterations} = 300$). We also selected seven groupings of feature generation methods–five for when full text is available, and two for when it is not (Table 3.3). We evaluated these classifiers, with these feature configurations, by training the model on the entire training collection, and classifying the hold-out testing collection. We selected an overall best-performing system from all classifier/feature generation method combinations based on system performance (AUC), selecting one for the case where full text is unavailable, and one for the case that it is available.

3.2.4.1 Determining the Effect of Training Data Set Size

To determine the contribution of training set size to performance–and to see whether further performance gains could be achieved, we did a series of train/classify experiments in which n-percent (where n=10-100, by 10s) of the training data was randomly sampled for creating a classification model to use against the entirety of the hold-out testing data collection. At each percent, we conducted 30

Configuration	Feature Generation Methods
1	Captions, NeuroLex, Abstract, Title, MeSH
2	NeuroLex, Abstract, Title, MeSH
3	NeuroLex, Abstract, Title (Trigrams), MeSH
4	Captions, NeuroLex, Abstract, Title (Trigrams), MeSH
5	Full text (Bigrams), NeuroLex, Abstract, Title (Trigrams), MeSH
6	NeuroLex (MEDLINE only), Abstract, Title, MeSH
7	Abstract, Title, MeSH

Table 3.3: Table of our feature generation method sets for the train/classify experiments, using the hold-out testing data for classification. Feature sets for full text availability are shown in black, while those for use against the MEDLINE record only are in red. From these configurations, we selected two best-performing systems–one for when full text is available, and one for use with MEDLINE records alone.

train/classify experiments and averaged the AUCs, in addition to computing 95% confidence intervals for our experiment.

3.3 Results

Initial classification experiments were done using five repetitions of 2-way cross-validation on the hold-out training data, using a variety of classifiers built on models constructed from a baseline set of features (word-based unigrams derived from document titles, abstracts, and MeSH terms). Since many of these classifiers had parameters that needed to be optimized, we first examined the range of parameter values that lead to good performance for each classifier, comparing each classifier parametrization using AUC on five repetitions of 2-way cross-validation using the above-described set of baseline features. In addition, since there were a variety of implementations for the SVM algorithm, we also examined whether there were performance differences between the SVM$_{light}$ and Weka implementations of the linear-kernel SVM classifier.

3.3.1 Selecting Classifier Algorithms

3.3.1.1 SVM$_{light}$ v. Weka SMO

To determine which implementation of the SVM classification algorithm should be used in our studies, we analyzed the performance of the SVM$_{light}$ and Weka SMO [149; 232] implementations across a range of C-parameter values. For our C-parameter values, we selected and handful of values in the upper and lower extremes (e.g., 1.0 x 10^{-8}, 1.0 x 10^{7}), as well as a consecutive range of values in the area of typical C-parameter settings. The results of this study are depicted in Figure 3.1. As can be seen there, SVM$_{light}$ consistently out-performed the Weka implementation. Both implementations showed a step-like increase in AUC at C-parameter values of 0.1, with the Weka implementation showing the largest gains at this point. We selected C=1.0 for SVM systems studied from here on out, as that value is in the middle of the range of best-performing configurations, and will therefore be likely to generalize to other data sets.

3.3.1.2 SVM$_{light}$ Radial Basis Function, Sigmoid, & Polynomial Kernels

We examined a range of parameter values for the radial basis function (RBF) and sigmoid kernels next. Besides the C-parameter, which we fixed at 1.0, the RBF kernel has a γ parameter. We examined a range of values between 0.001 - 0.01, 0.01 - 0.03, 0.03 - 0.1, and 0.1 - 1.0. The results of this study are shown in Figure 3.2. As this figure shows, there is a range of acceptable parameter values, with

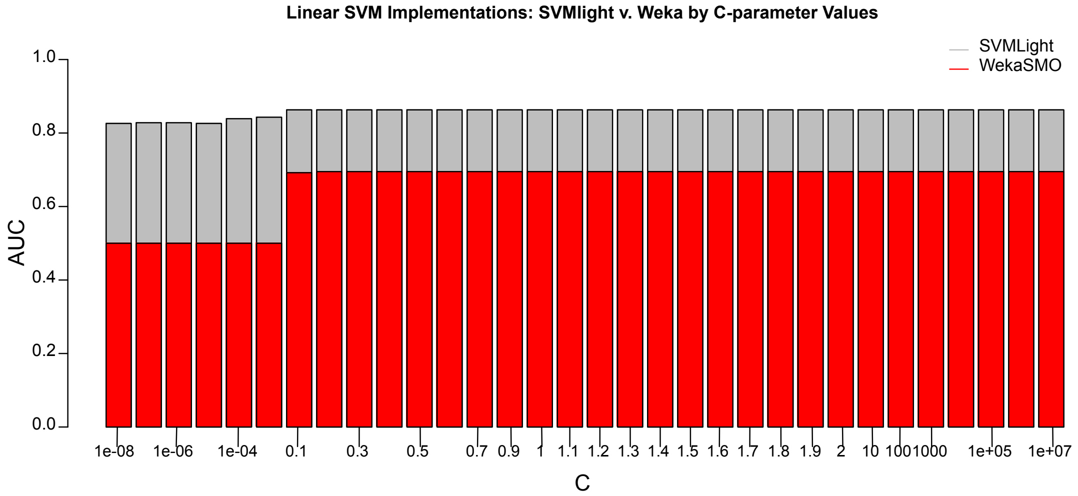

Figure 3.1: Performance (AUC) of the SVM$_{light}$ and Weka implementations of the linear-kernel SVM classification algorithms. SVM$_{light}$ consistently out-performed the Weka implementation. Both implementations showed a step-like increase in AUC at C-parameter values of 0.1.

0.02 being in the middle of that range, having an AUC of 0.862, showing a slight performance gain over the linear kernel. We also examined a range of S-parameter values (0.01 - 0.1, 0.1 - 1, and 1 - 5) for the SVM$_{light}$ sigmoid kernel. The results of this experiment are shown in Figure 3.3. As can be seen there, the best-performing sigmoid kernel parametrization is with $S = 0.01$, achieving an AUC of 0.839, a negligible improvement over the linear SVM. Finally, we examined $2^{nd}-$, $3^{rd}-$, $4^{th}-$, $5^{th}-$degree polynomial kernels across a range of $S-$parameter values (0.01 - 0.1, 0.1 - 1, and 1 - 5). The results for this experiment can be seen in Figure 3.4. Interestingly, each degree kernel had nearly the same maximum AUC (0.855; the $2^{nd}-$degree kernel achieved the negligibly-different 0.854), but they differed in the frequency with which their parameter tunings approximated peak performance. The $2^{nd}-$degree kernel, for example, tended to score near this value across the entire parameter range investigated, whereas the $5^{th}-$degree kernel had only a single parameter setting that lead to this performance. Figure 3.5 shows another way of looking at these data. Here, we plotted the distribution of AUC scores across $S-$parameter values by each kernel degree. Here, the $2^{rd}-$degree kernel had the highest peak around the maximum AUC, meaning that, in the absence of any prior knowledge about a specific $S-$parameter value, would would be most likely to achieve peak performance with a value randomly selected in the range investigated here with that kernel.

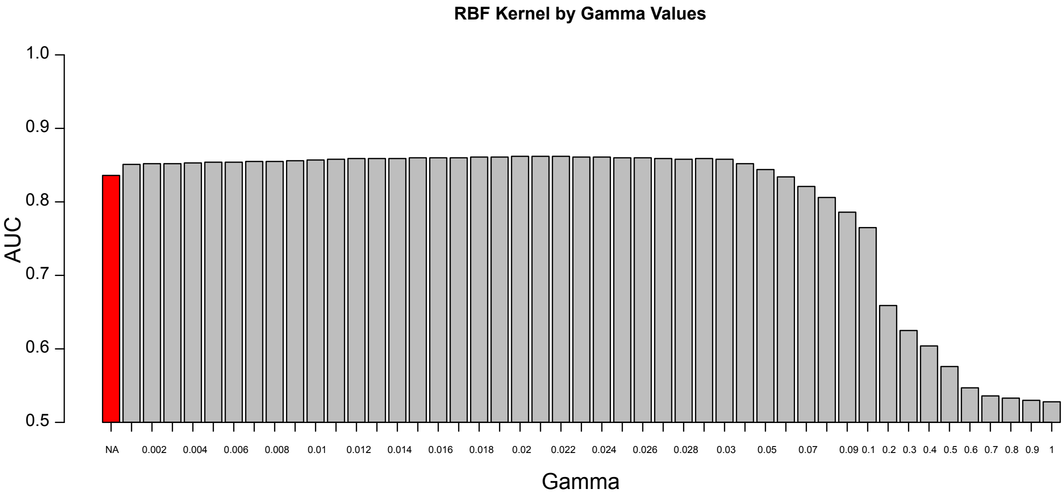

Figure 3.2: Performance (AUC) of the SVM_{light} radial basis function kernel across γ parameter values. For reference, the best-performing linear kernel SVM is shown in red. The RBF kernel tended to outperform the linear kernel, doing so over a broad range of γ parameter values, but eventually degrading at more extreme settings.

3.3.1.3 Adaboost with Decision Stumps & Random Forests

We examined two classification algorithms with parameters to optimize from the Weka machine learning library [120]–Adaboost with Decision Stumps (AdaboostDS)[102; 103], and Random Forests (RF) [44]. The only parameter that has to be opti-

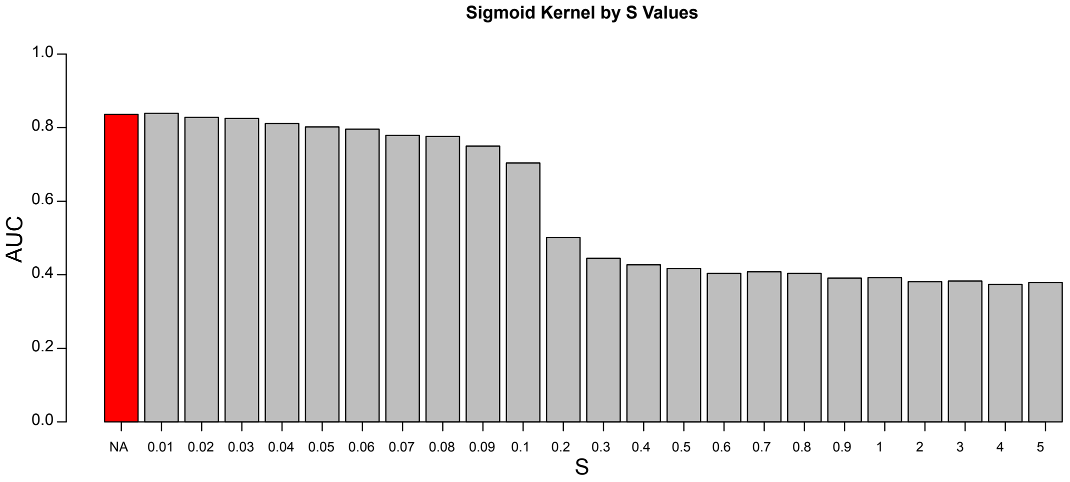

Figure 3.3: Performance (AUC) of the SVM$_{light}$ sigmoid kernel across S parameter values. For reference, the best-performing linear kernel SVM is shown in red. The sigmoid kernel only out-performed the linear kernel at one setting, but never did so with a significant difference in performance.

mized in AdaboostDS is the number of iterations the algorithm has to run. Here, we investigated a range of 1, 5, and 10, 10 - 100, and 100 - 700, shown in Figure 3.6. Here, we observed very poor performance with a single iteration, as would be expected, which quickly jumps to 0.790 at 10 iterations. After this, although performance gains are still seen by increasing the number of iterations, the gains are modest with each increase, but eventually peak at 0.837 at 600 iterations. We optimized the number of features and number of trees used by the RF algorithm. The results of these experiments are shown in Figure 3.7. There were several parametrizations for the number of trees that lead to acceptable performance here. As one would expect, the performance of the system increases with the more features that are added–with a single feature (solid red line), AUC is almost uniformly lower than other configurations, and, as the number of features allowed is increased, performance improves. To select a configuration of the RF

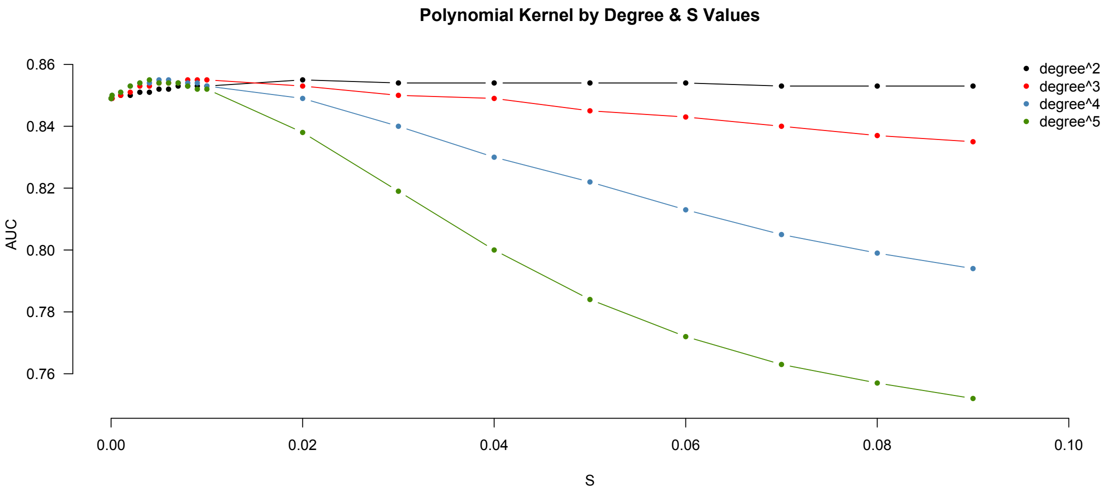

Figure 3.4: Performance (AUC) of the SVM$_{light}$ $2^{nd}-$, $3^{rd}-$, $4^{th}-$, $5^{th}-$degree polynomial kernels across a range of $S-$parameter values. As the degree in the polynomial kernel is increased, peak performance increases slightly, but does so at the expense of the range of values at which the performance can be obtained. The second-degree kernel, for example, had a broad range of peak performance values, whereas the fifth-degree kernel had a very narrow range.

algorithm for our later experiments, we aimed for selecting a number of trees that had a broad range of acceptable number of feature parameter values leading to good performance. For this feature, however, since we hadn't yet conducted our feature generation experiments, we didn't want to limit the number of features given to the algorithm. Thus, since 16, 32, and 64 features all seemed to perform in the same range, we selected 32 features, with 74 trees.

3.3.1.4 Comparison of Optimized Classifiers

We next compared the performance of the above-described optimized classifiers along with that of several other classifiers that did not have parameters needing to be optimized. A list of the classification algorithms investigated and their

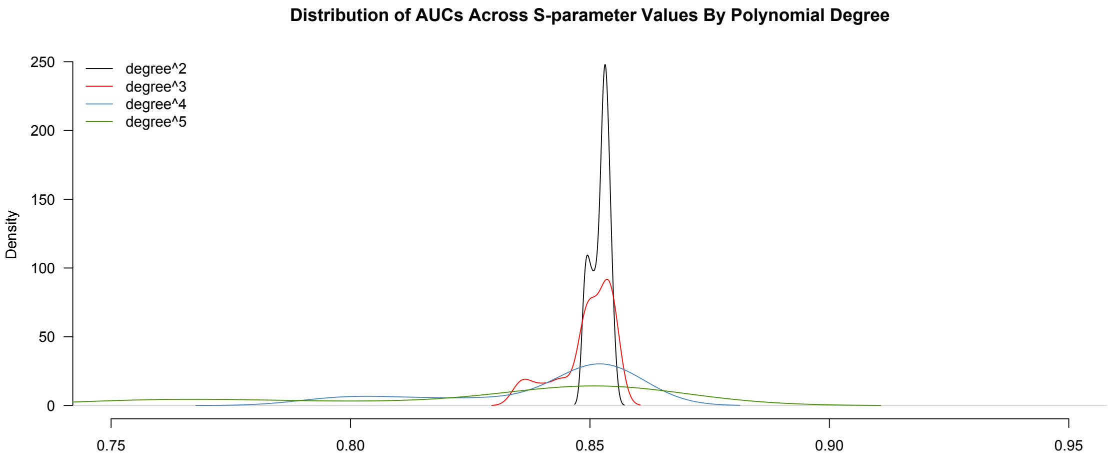

Figure 3.5: Distribution of AUCs across $S-$parameter values for each of the SVM$_{light}$ $2^{nd}-$, $3^{rd}-$, $4^{th}-$, $5^{th}-$degree polynomial kernels. The narrow range of peak performance for the fifth-degree kernel is evidenced here by the broad density of performance values.

origins follows:

1. k* (Star): a classification algorithm that makes judgements according to the characteristics of the documents positioned around some to-be-classified document within the feature space, using an entropy-based distance metric.[60]

2. IBK: the Weka implementation of the k-nearest neighbors algorithm. In the experiments described here, we used the default settings[1]: 1 nearest neighbor used in voting, votes unweighted by sample similarity.[2]

3. One Best Feature (OBF): makes predictions based on the single-most informative feature found in the training set, according to mutual information.

4. Complement Class Naïve Bayes (CNB) [239]

[1] http://www.cs.tufts.edu/ ablumer/weka/doc/weka.classifiers.IBk.html

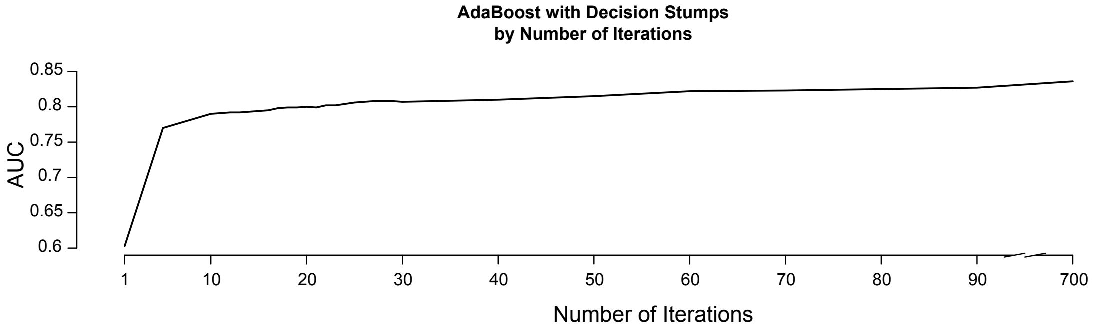

Figure 3.6: Performance (AUC) of the Adaboost with Decision Stumps classification algorithm, by number of iterations. Iteration values between 100 and 700 were all investigated, but showed very little variability. For the purposes of this figure, Iterations 100-600 have been withheld. Although there were many parameter settings at which the Adaboost system performed well, it took many hours for these experiments to run.

5. J48, Multiboost J48: The Weka implementation of C4.5 decision trees. For the experiments here, we used the default parameter settings–0.25 confidence threshold for pruning, a minimum of 2 instances per leaf, and three folds used for reducing error pruning.[236]

6. Winnow: an algorithm similar to the perceptron, that is known to perform well in high-dimensional, sparse data.[187; 188]

7. SMO (see above: 3.3.1.1). Weka's implementation of Support Vector Machines is based on Platt's sequential minimal optimization algorithm. [149; 232]

8. Auto Centroid: a classification algorithm that optimizes the number of per-class sub-clusters by optimizing for the F1 performance metric.

9. CF Centroid [116]

10. Logistic Regression [174]

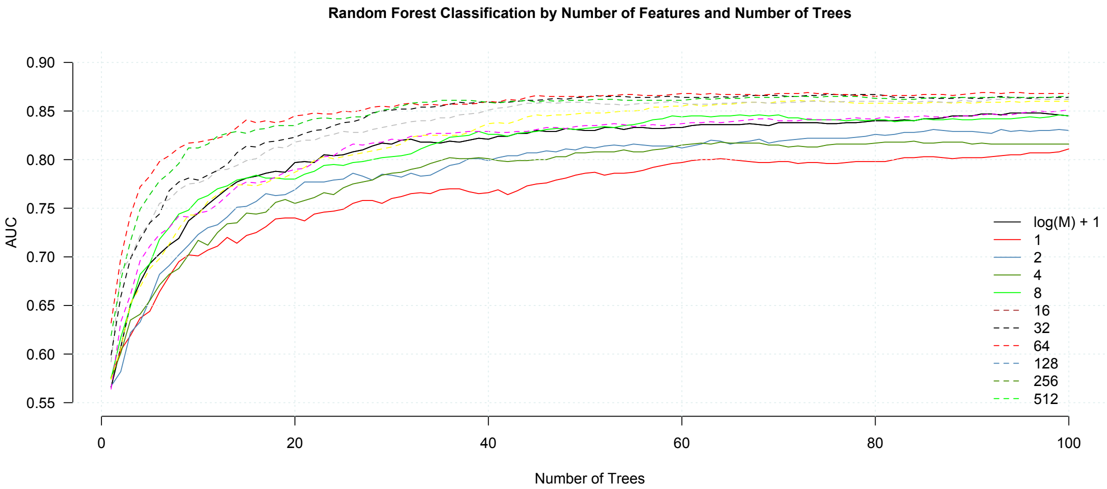

Figure 3.7: Performance (AUC) of Weka's implementation of the Random Forest classification algorithm. Data are separately-plotted according to the number of features used by the algorithm, over the number of trees. In the Weka implementation, the default number of features is set to $\log(M) + 1$, where M is the number of unique features identified, which is plotted here as a black solid line. There were many decently-performing parameter configurations for this classifier, with performance tending to level out around 40 trees.

11. Random Forest (see above: 3.3.1.3)

12. Adaboost with Decision Stumps (see above: 3.3.1.3)

13. Naïve Bayes [144]

14. Centroid [95]

15. kIGNN [9]

16. SVM (sigmoid kernel; see above: 3.3.1.2) [143]

17. SVM (2-degree polynomial kernel; see above: 3.3.1.2) [143]

18. SVM (3-degree polynomial kernel; see above: 3.3.1.2) [143]

19. SVM (RBF kernel; see above: 3.3.1.2) [143]

20. SVM (linear kernel; see above: 3.3.1.1) [143]

We compared the above-described algorithms by five-repetitions of two-way cross-validation on the training set, selecting five algorithms to use moving forward, for our feature generation experiments. We made our selections primarily based on performance (AUC), but also diversity in approach. For example, the top-five performing classification algorithms were all variants of the SVM$_{light}$ support vector machine implementation. Although these might be the best-performing systems on the baseline feature set, it is possible that they would all respond similarly to newly-introduced feature types. Thus, we selected five classification algorithms: SVM$_{light}$ (linear kernel), kIGNN, AdaBoost, and Random Forest. We selected these classifiers based on their diversity, performance, and speed.

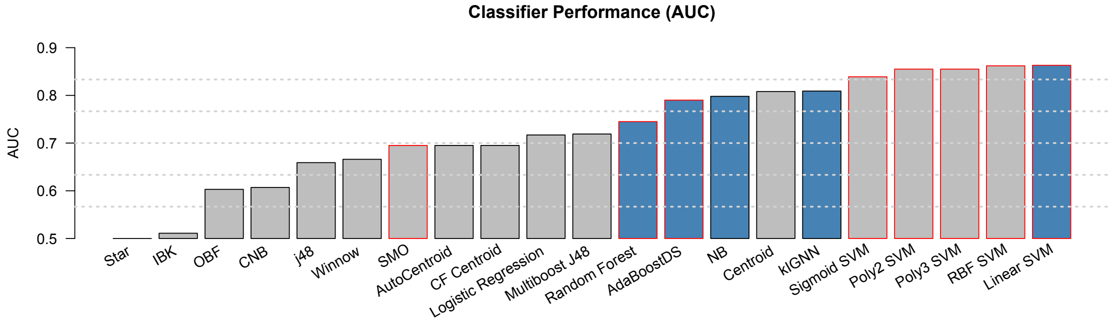

Figure 3.8: Performance (AUC) of the classification algorithms investigated in this study, using 5-repetitions of 2-way cross-validation on the hold-out training data. Bars outlined in red show the best-performance observed during parameter optimization, and bars shaded blue were selected for use in the feature generation experiments.

3.3.2 Feature Generation Methods Selection

To extend the feature types beyond that included in the baseline system configurations, we investigated a variety of binary and continuously-valued features. Many of these features had to do with various ways of quantifying the presence or absence or neuroscience- and experimental design-related terminologies. For completeness, we also investigated the utility of n-gram representations of documents' title, abstract, full text, and pre-processed full text, for identifying figure captions. We'll first review the results for our n-gram and full text experiments, before turning to the results of the quantitative feature experiments. For all experiments, we used the linear-kernel SVM$_{light}$ classifier set to default parameter settings, before examining the synergistic effects of combining the selected and optimized classifiers with the selected feature generation methods; methods were selected based on performance (AUC) in five-repetitions of two-way cross-validation.

3.3.2.1 n-gram Experiment results

We examined the efficacy of a classifier using unigram, bi-gram (with unigrams), or tri-gram (with bi-grams and unigrams) representations of different parts of the training documents. Full text features were obtained from the plain text output of *pdftotext*, following cleaning up with the manually-constructed regular expressions described in the Methods section. Caption-based features were obtained by pre-processing the full text with a manually-constructed regular expression for identifying figure and table captions. The results of these experiments are summarized in Table 3.4. Here, it can be seen that adding bi-grams and tri-grams had mixed effects across feature types, but, in general, did not make a great difference. Adding bi-grams and tri-grams to manuscript title unigrams led to a small, but not insignificant, performance gain over using unigrams alone, while abstracts were best-represented as unigrams. Full text bi-grams out performed other representations of full text, included all those for caption pre-processing.

Feature	Unigram	Bi-gram	Tri-gram
Title	0.736	0.741	0.743
Abstract	0.836	0.831	0.824
Full text	0.831	0.839	0.834
Captions	0.805	0.805	0.795

Table 3.4: Summary table of *n*-gram feature experiment results (in terms of AUC). The best-performing configuration for each feature type is printed in red. Representing title features as tri-grams gave a significant performance boost, but uni-grams were best for Abstract and Caption features.

3.3.2.2 Evaluating the Contribution of Full text & Captions

We next investigated the contribution of adding full text to other feature generation methods. Although there would likely be many situations in which our classification system would need to be used on documents for which full text data is not available (i.e., just the MEDLINE record would need to be used), we hypothesized that, in those situations where full text could be obtained, it might add some important information over and above that provided by other feature configurations. Thus, we ran a set of classification experiments in which full text or captions were added to titles, abstracts, and MeSH, in a step-wise fashion. The primary purpose of these experiment was to determine whether, on those occasions where full text is available, it is better to use it as-is, or to pre-filter it for text found in figure captions. Rather than simply compare the two configurations directly, we opted to also examine the unique contributions these feature sources make to the classification model as the informative MEDLINE-based features are added into the mix. We limited the present experiments to unigrams derived from the respective sources. The results of these experiments are shown in Figure 3.9. We observed that full text-based unigrams out-performed captions alone, but, as MEDLINE-based features begin to be added in to the mix, the caption-filtered full text systems began to out-perform the full text-based systems. Once titles, abstracts, and MeSH terms are in the mix, the caption-based system is one point better than full text. Comparing either group of systems to those not utilizing full text at all, both perform better, once abstracts are added as a feature source.

3.3.2.3 Evaluating the Contribution of Continuously-Valued Features

We investigated a collection of derived features we collectively refer to as *continuously-valued* features, to contrast them with the binary features that have already been described. For each of the features categories, we examined the distribution of values by class (Include v. Exclude), prior to examining their possible contribution to a classification system. Thus, we present the distribution results first, before turning to the classification results.

3.3.2.3.1 NeuroNames We examined the distribution of absolute NeuroNames term counts and unique NeuroName term counts in the MEDLINE records and full text of documents by their classes. The distributions of total NeuroNames counts for each class are shown in Figure 3.10. For both classes, the peak of the distribution is around 0–meaning no, or very few, NeuroNames terms were identified in both classes, implying that this feature alone may not be useful in distinguishing the two classes of documents. The distributions of unique NeuroNames counts are shown in Figure 3.11. As with the total count distributions, the peaks of these distributions center around 0, with a second, less extreme peak around 1. For completeness, we examined the efficacy of these to quantitative approaches combined with other features in a classifier system, although it does not appear that there is much information here for the classifier to find.

3.3.2.3.2 Methods-related Terms We examined the distribution of total methods-related term count and unique methods-related term counts, using our manually-constructed term list, in the MEDLINE records and full text of documents by their classes. The distributions for total methods-related term counts are shown in Figure 3.12. Here, there appears to be similar distribution peaks between 0 and 5 counts, with the Include-class documents tending to have closer to 5 counts than the Exclude-class ones. In contrast, Figure 3.13 shows much more similar distributions between the two classes, implying that it will be difficult for a classifier to identify documents of interest based on this feature alone.

3.3.2.3.3 NeuroLex We examine the distribution of total number and unique number of NeuroLex term occurrences in the training documents by class. As was discussed in the methods section (see 3.2.3.1), these terms were quantified by matching the MEDLINE and full text features against a set of manually-constructed regular expressions. In the figures showing these data, the blue lines correspond to the values the recursive partitioning modeler used to partition the continuous values into a set of binary values. We've included these here, because the results of our final optimized classifier and feature generation methods cross-validation experiments (see 3.3.3) indicated that the NeuroLex features were helpful to include in the final systems configurations to consider. The distributions of total NeuroLex counts by document class are shown in Figure 3.14. As can be seen in this figure, there are some small differences between the two distributions–the Exclude-class documents' peak is centered around 0, with a base that extends to near 10, meaning that they tend to have very few NeuroLex term mentions, while the Include-class documents have a much broader peak, which is nearer 10 at its apex. This implies that this feature type might be useful to a classifier for distinguishing the two classes, but, it could be argued, there may be better ways of looking at the data. The cuts made by the recursive partitioning modeler, for example, tend to be between 1 and 10, and are all clustered together, and are in places where the distributions don't have a great deal of differences. This means that, between repetitions of cross-validation, the recursive partitioner didn't have many different places to make helpful cuts in the data–it always made the cuts where the distributions didn't seem to have many differences. Thus, although it may have found some helpful information here, the small differences in the distributions imply that there may be a moderate amount of information here that would be useful to a classifier.

Figure 3.15 shows the distribution of the unique number of NeuroLex occurances identified in the training set documents by class label. As in Figure 3.14, the cuts made by the recursive partitioning modeler across repetitions of cross-validation are shown in blue. The Exclude-class distribution (black) is multi-modal, having peaks near 0, 1, and 2, with the highest being near 0. In contrast, the Include-class documents have a broad distribution with the peak near zero. The recursive partitioning modeler found a variety of places to make informative cuts in the values, each of which appears to be located at a place where the two distributions differ from one another. For example, the cut located near 0.75 is between two of the Exclude class peaks, and where there is very little density in the Include-class data. These results lend face validity to using NeuroLex term counts (total and unique) in a classification system going forward.

3.3.2.3.4 The Contribution of Term Distances The final set of feature types we studied had to do with term distances–the per-document mean and minimum distance between a method-related term and a NeuroNames term. Our original intention was to also create these features for the distances between methods-related terms and NeuroLex terms, however, our initial experiments for calculating distances between methods and NeuroNames terms yielded very little, by way of benefit to the classifier. What's more, calculating these features throughout the full text of each document was so time-consuming (i.e., several hours to complete a cross-validation run) that we decided these features would not likely be useful in practice. The results of our experiments on using the distances between methods-related terms and NeuroNames terms are shown in Figures 3.16 and 3.17. For both figures, the distribution of distances are similar between the document classes. The peaks on the far right of each graph are created by the default value of $1\ x\ 10^{19}$ used by the generator on documents where both methods and NeuroNames terms could not be found.

3.3.3 Selection of the Best Classifier System Configuration

Based on the results of the above experiments, and our desire to have a classification systems that can work quickly both in the presence and absence of full text availability (non-full text configurations are shown in red below), we picked seven sets of feature generation methods to evaluate in five repetitions of two-way cross-validation with the best-selected classification algorithms.:

1. Caption, NeuroLex, Abstract, Title, MeSH
2. NeuroLex, Abstract, Title, MeSH
3. NeuroLex, Abstract, Title (Tri-grams), MeSH
4. Captions NeuroLex, Abstract, Title (Tri-grams), MeSH
5. Full text (Bi-grams), NeuroLex, Abstract, Title (Tri-grams), MeSH
6. NeuroLex (MEDLINE only), Abstract, Title, MeSH
7. Abstract, Title, MeSH

The results of these experiments are shown in Figures 3.18 and 3.19. Looking at Figure 3.18, we can see that, overall, SVM tended to perform quite well (AUC $\geq$ 0.85). Even in the absence of full text and quantitative features (configuration 7), SVM achieves an AUC of 0.863. Looking across classifiers within feature configurations, SVM is only out-performed in configuration 2, by RF ($\delta = 0.002$) and configuration 5, by AdaBoost ($\delta = 0.02$). Figure 3.19 shows the same data, but reorganized to group the performance of the different classifiers together for each configuration. Here, we see that the feature configuration leading to the best overall performance when full text is available, is configureation 2, which lead to AUCs of 0.879 and 0.877 in RF and SVM, respectively. In the absence of full text, configuration 7 appears to be the best, leading to AUCs of 0.863 and 0.861 in SVM and RF, respectively. Interestingly, the addition of full text affected the classifiers differently. If we compare configurations 2 and 7, each of which consists of Abstract, Title, MeSH, and NeuroLex (but either run on full text, in configuration 2, or the MEDLINE record, in configuration 7), we see that AdaBoost gained 0.029 from the addition of features derived from the full text, while kIGNN lost 0.006.

3.3.4 Evaluation With the Hold-out Testing Data

To evaluate our system configurations, we trained classifiers on the entire training data set, and evaluated the hold-out testing set. Our goal was to identify a best-performing classification system for use when full text is available, and one for use when it is not. The results of this experiment are shown in Figure 3.20, and reorganized according to system configuration in Figure 3.20. Classifier and feature configurations were almost uniformly improved over the respective cross-validation performances, with the SVM systems showing some of the best improvement. One classifier (Adaboost, configuration 1) ran for 24 hours without finishing, so those data are missing here.

The overall best systems used SVM classifiers, with feature configuration 3 (C3; full text available), and SVM with feature configuration 6 (C6; full text unavailable). RF configuration 6 was a close second to SVM, coming in at 0.01 lower, implying that it may still be worth considering for future use. However, the SVM classifiers were drastically faster than the RF classifiers (finishing in seconds, rather than minutes), so they would likely be more desirable for deployment in the real world anyway. Thus, based on these results, we selected SVM configurations 3 and 6 for use in future real-world deployment.

3.3.4.1 Train/Classify Learning Curve

To make sure that the performance of our selected systems was not held back by insufficient training data, we examined learning curves for each system, down sampling the training set at increasing percentages (from 10 - 100), and classifying the hold-out testing set with the resultant model each time. The results of this experiment is shown in Figure 3.22. As would be expected, the confidence intervals narrow as the sampling percentage increases, and the C3 system performs consistently better than C6, and both systems approximate their cross-validation results at 50% sampling. Neither system appears to have plateaued in performance by 100%, implying that further performance gains could be obtained from adding more data to the training collection.

3.4 Discussion

Our recent work [12] (see Chapter 2) has put us in a unique position to study the complications associated with developing a document classification system for identifying publications containing information of interest for Neuroinformatics databases, such as the NR. Here, we have examined a variety of possible classification algorithms and feature generation methods that have begun to give us a good idea of what sorts of systems will be successful in this setting.

The studies described in this series of experiments were designed to identify a classification algorithm that could be used for real-world biocuration problems in Neuroinformatics. Based on this, we have succeeded: our classifiers are able to accurately rank a set of documents, in terms of their likelihood of containing information of interest to a database (here, the NR), and are able to do so very quickly (on the order of seconds). Both best-performing systems used features derived from our manually-created set of search terms based on the neuron branch of the NeuroLex, an open source Neuroscience ontology–something that has not, to our knowledge, been previously investigated as a source of feature types for this type of classification problem.

Interestingly, between our cross-validation and train/classify experiments, the best-performing system changed. This implies that, perhaps, our results are somewhat subject to the amount of training data available for these experiments. Since the train/classify experiment were conducted using more data, it is likely that the results observed here are truer to those that would be observed on a larger data set, but, to be sure, further experiments should be carried out when more data is available. This conjecture is further supported by the results of our learning curve experiments (Figure 3.22). Although the learning curves depicted in these experiments indicate that our results have begun to stabilize, it is clear that they have not completely done so. There were two other classification algorithms that are likely to warrant consideration in the future–RF, and AdaBoost.

Various configurations of the RF classifier tended to perform similarly to the best SVM configurations, but only exceeded the performance of the best SVMs with smaller amounts of training data. Possibly, the RF algorithm is more robust to smaller data sets, whereas SVM may get mislead in these situations. Similarly, AdaBoost tended to perform amongst the best classifiers. However, there are two limitations to this algorithm that would likely prevent it from being used in a real-world setting. First, it's performance was very much linked to the number of iterations allowed for the algorithm. Although the parameter settings used here were selected because of they appeared to lie in the middle of a stable parameter range, sensitivity to this parameter value may be triggered by using different data or with smaller data sets. Second, the Adaboost classifier was extremely slow. In fact, one of the train/classify configurations was allowed to run on a dual core MacBook Pro for 24 hours and never completed. Even if it had out-performed SVM, running for this long severely limits its deployment in the real world. In contrast, the linear-kernel SVM$_{light}$ ran very quickly and had only one parameter to optimize that had low variability across parameter values. Even if SVM was the third- or fourth-best-performing classifier, these characteristics would make it a very reasonable choice for deployment in the real world.

We investigated a variety of features here, two of which were based on hand-curated term lists. Although methods-related distance-based continuous-valued features have been shown to be effective in similar tasks (e.g., the BioCreative Protein-protein Interaction Document Task [9]), they did not here. Possibly, the curation task the classifier created in Ambert, 2011 was more *methods-centric* than the one here: the experimental detection of protein-protein interaction is limited to a handful of methodologies, whereas the scope of the NR is much more broad, incorporating a variety of bench methods. One interesting finding here is that, although features derived from the MEDLINE record (i.e., Abstract, Title, MeSH) tended to lead to decently-performing classifiers, the addition of full text often improved the best-performing systems. Although the information density of the text in MEDLINE-related features is much higher, which, in general, can lead to good performance for general-purpose document classification systems, it is possible that the specific findings that are curated in the NR can only be reasonably expressed in the full text, such as figure captions, or the results sections. Both of our best-performing systems incorporated continuous-valued features derived from the NeuroLex Ontology of Neuroscience. Despite the well-documented text-mining problems associated with the heterogeneity of language in the Neurosciences, it is interesting that the use of an expert-created adaptation of an open ontology, here, lead to improved performance. Although further experiments are necessary, we believe this demonstrates the utility of expert-curated term sets in classification problems that are very narrow in scope, such as is common to the field of biocuration.

Although the best-performing classifiers in the present set of experiments appear to be accurate enough for use in the real world, further investigation is necessary to verify that this is the case. In particular, an implementation study that examines how they can be best integrated into a biocuration workflow are important. Even though our systems achieved very high AUCs, it is important that they have an interface that is easily adopted by biocurators, and requires minimal alteration of their workflow. Another limitation of this study is that the data it uses was manually annotated by one Neuroscientist [12] (see Chapter 2). To better generalize to the field of Neuroscience in general, it would be beneficial to incorporate data that has been annotated by other Neuroscientists into this data set.

3.5 Conclusion

In this study we have demonstrated that an automated system can be used to identify documents containing information of interest for databases being developed by the Neuroinformatics community. By training on a manually-curated Neuroscience document set, we showed that two configurations of an SVM-based classification system can be used to identify documents of interest, both in the cases where full text articles are available, as well as those where they are not. In the presence of full text, we were able to achieve an AUC of 0.925, while in the absence we achieved an AUC of 0.917–both scores indicate these systems could be realistically used by Neuroinformaticians and Biocurators. Future work will need to focus on further quantifying the contribution of training data availability to the observed scores (i.e., can scores be improved further with additional data), and the actual implementation of a system such as ours into a biocuration workflow.

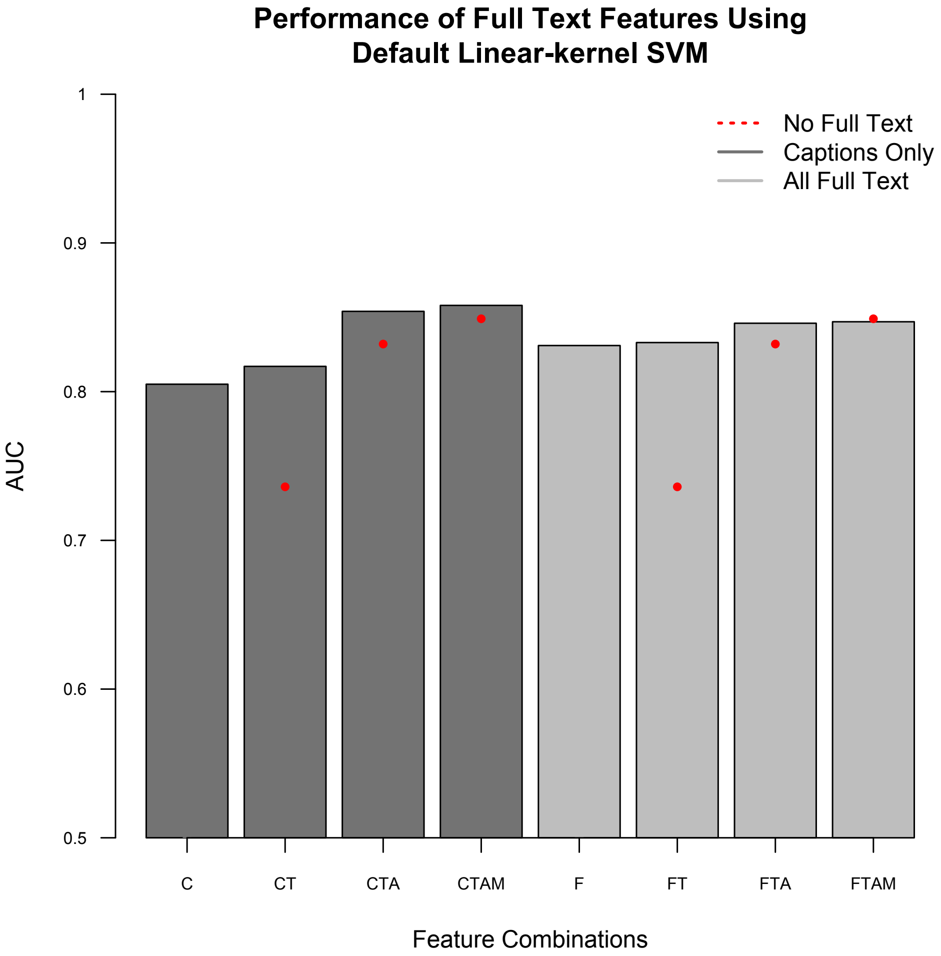

Figure 3.9: Performance of various feature combinations with two full text-type features. Combinations involving caption-derived features are shown in dark grey, while combinations involving the full text extracted from the plain text representation of the manuscripts are shown in light grey, and the corresponding feature combinations in the absence of either type of full text are depicted as a red dot. The addition of full text features tended to lead to performance gains, especially in the case of filtering the full text for figure captions. *Abbreviations: C=Captions, F=Full text, T=Title, A=Abstract, M=MeSH.*

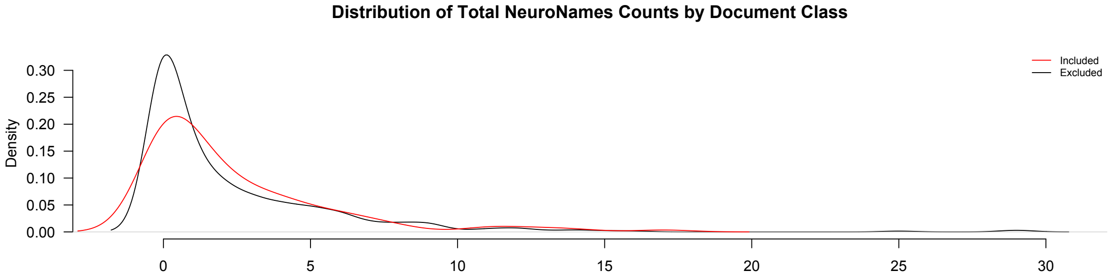

Figure 3.10: Distribution of total NeuroNames counts by Include (red) and Exclude (black) class label. The distribution of these terms did not differ greatly between classes.

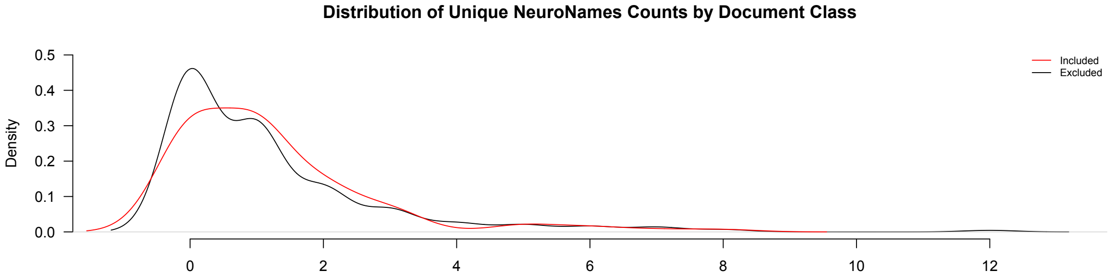

Figure 3.11: Distribution of unique NeuroNames counts by Include (red) and Exclude (black) class label. The distribution of these terms did not differ greatly between classes.

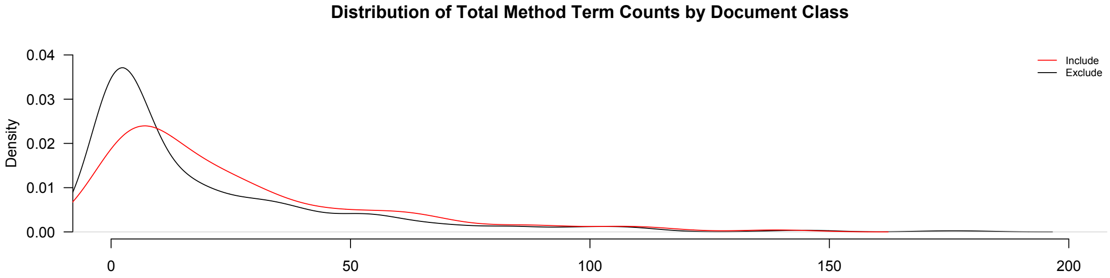

Figure 3.12: Distribution of total methods-related term counts by Include (red) and Exclude (black) class label. The distribution of these terms did not differ greatly between classes.

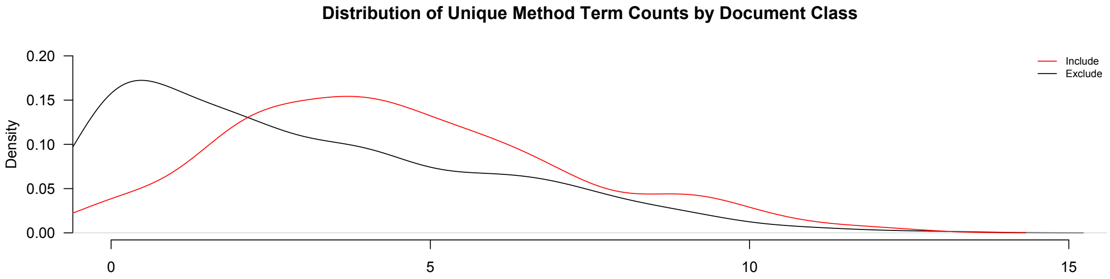

Figure 3.13: Distribution of unique methods-related term counts by Include (red) and Exclude (black) class label. The distribution of these terms did not differ enough to indicate this feature's usefulness for classification.

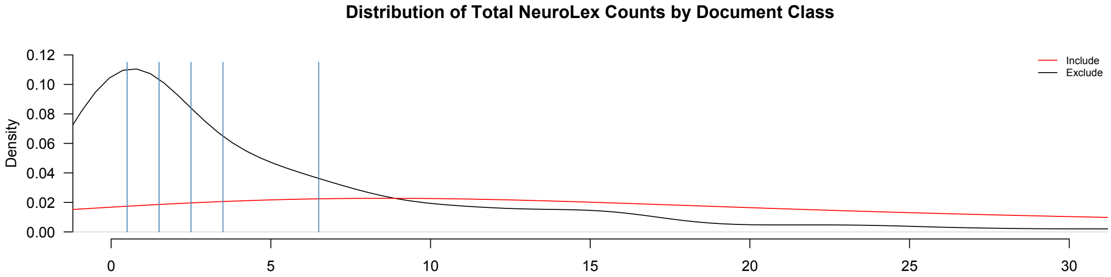

Figure 3.14: Distribution of total NeuroLex term counts by Include (red) and Exclude (black) class label. The blue lines denote places the recursive partitioning modeler made cuts during cross-validation.

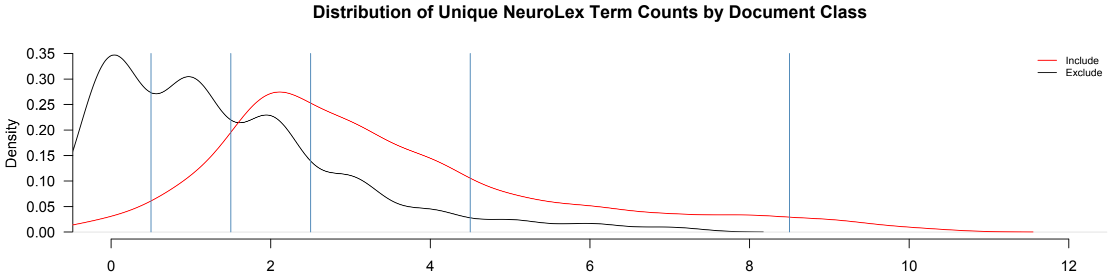

Figure 3.15: Distribution of unique NeuroLex term counts by Include (red) and Exclude (black) class label. The blue lines denote places the recursive partitioning modeler made cuts during cross-validation.

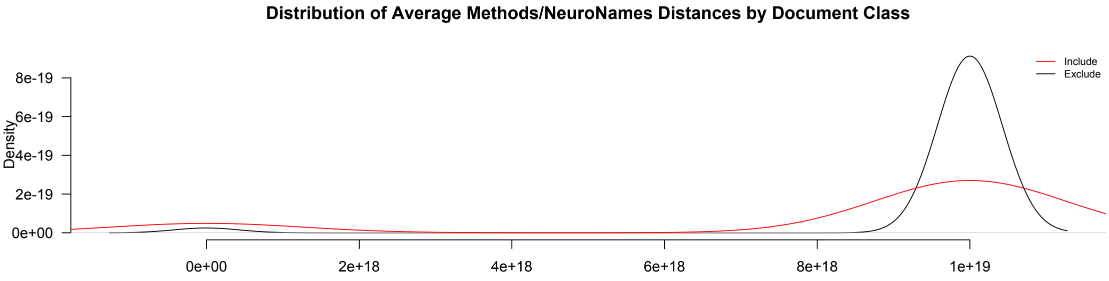

Figure 3.16: Distribution of the per-document average distances between a methods-related term and a NeuroNames term by document class. 1 x 10^{19} denotes the default value used by the feature generator when either a methods-related term or a NeuroNames term were not found in a document.

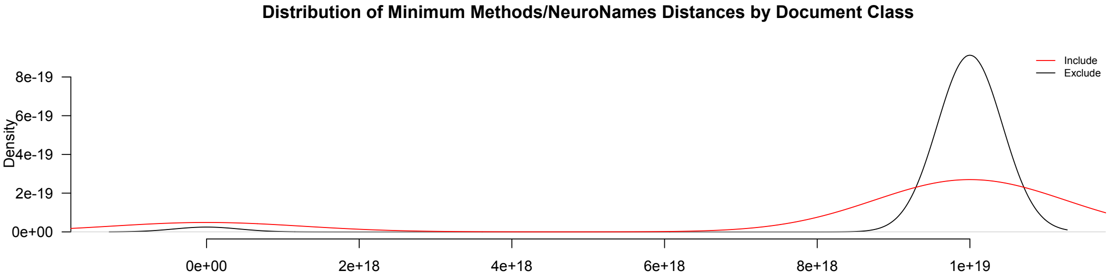

Figure 3.17: Distribution of the per-document minimum distances between a methods-related term and a NeuroNames term by document class. 1 x 10^{19} denotes the default value used by the feature generator when either a methods-related term or a NeuroNames term were not found in a document.

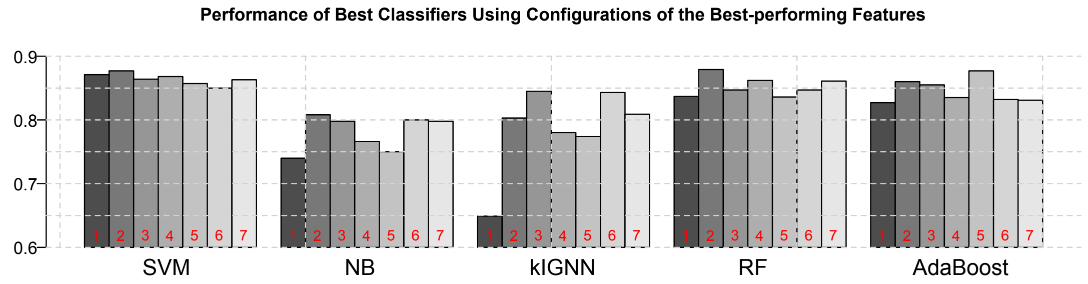

Figure 3.18: Performance of the best-performing classifiers with configurations of the best-performing feature types, during five repetitions of two-way cross-validation on the training data. SVM tended to perform well for all feature configurations, while *k*IGNN was more variable.

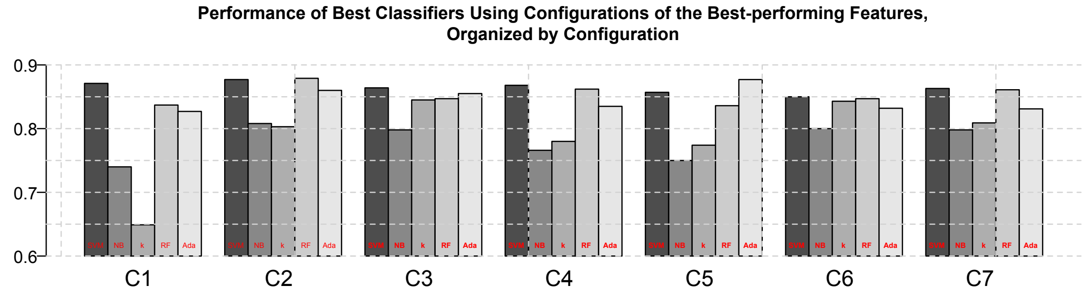

Figure 3.19: Performance of the best-performing classifiers with configurations of the best-performing feature types, during five repetitions of two-way cross-validation on the training data, organized by configuration. *Abbreviations: C=Configuration, k=kIGNN, Ada=AdaBoost.* Configurations 2, 3, and 6 tended to perform well for all classifiers. Classifiers appear in this order: SVM, NB, K, RF, Ada.

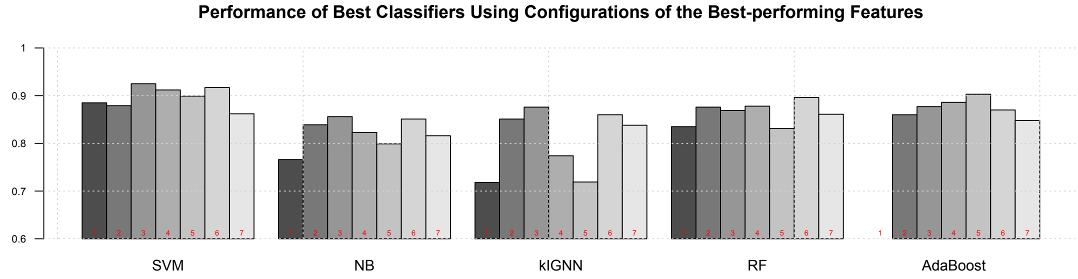

Figure 3.20: Performance of the best-performing classifiers with configurations of the best-performing feature types, after training on the training data, and classifying on the hold-out testing collection.

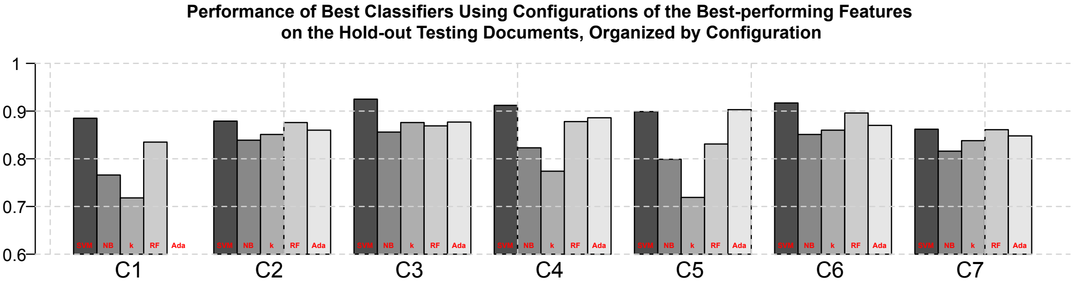

Figure 3.21: Performance of the best-performing classifiers with configurations of the best-performing feature types, after training on the training data, and classifying on the hold-out testing collection, organized by configuration. *Abbreviations: C=Configuration, k=kIGNN, Ada=AdaBoost.* Configurations 2, 3, and 6 tended to perform well for all classifiers. Classifiers appear in this order: SVM, NB, K, RF, Ada.

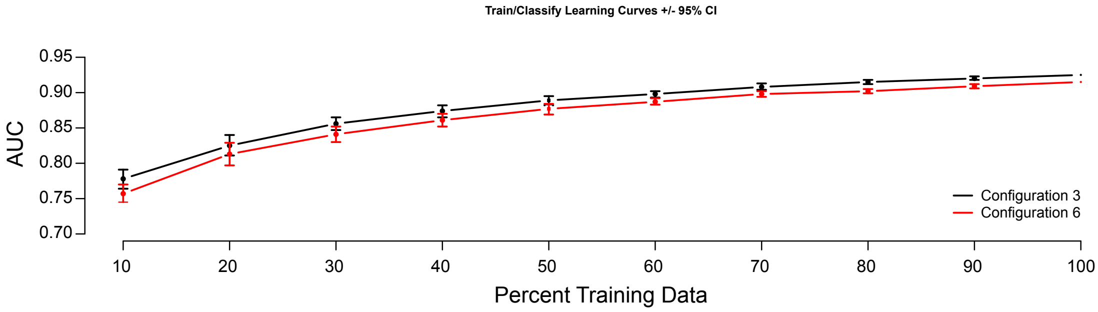

Figure 3.22: AUC Learning curve of the C3 (black) and C6 (red) systems. Training data was sampled 30 times at the percentages on the x-axis, to construct 95% confidence intervals. Both systems responded similarly to increasing data, with configuration 3 tending to be better than configuration 6.

Chapter 4

Finna: A Paragraph Prioritization System for Biocuration in the Neurosciences

4.1 Introduction

The manual creation of discipline-specific knowledge bases is an expensive and time-consuming process, requiring the effort of experts over an extended period of time. Although some general-purpose tools have been developed for streamlining the workflows of biocuration tasks [49; 53; 130; 146; 233; 237; 244; 296; 297], and some work has been done on developing task-specific solutions using text-mining (particularly for the curation of systematic reviews [13; 65; 65; 66; 66; 67; 292; 303]). Many of these approaches have focused on classifying documents in terms of the likely relevance for the curation task at hand. This, however, only solves one part of the problem–given a likely relevant document, curators must still read it, looking for the information of interest.

Here, we describe *Finna*, a system that, given a likely-relevant document, will re-order its composite paragraphs in terms if the likelihood that they contain the relevant information. This addresses an important bottle-neck in the curation workflow depicted in Figure 1.4. As a part of the curation process, NR curators will review any submissions to the knowledge base that they deem necessitate review. Once an entry has been added, the publication associated with the information is, by definition, a positive-class manuscript, even though the submission may have been made in error. The previously-described system (Chapter 3) can be used to determine the likelihood that such an error has been made, but, if further review is still necessary, it may not be necessary for the curator to read the entire manuscript in question to find and understand the new submission. Using the system we describe here, curators may be able to spend less time reviewing new submissions, leaving them more time to identify new sources.

4.2 Methods

To better address the typical workflow of biocurators at the NR, we extended our document-level classifier to rank-order the paragraphs of positive-class documents according to likelihood of containing information that is relevant to curators.

4.2.1 Constructing the Paragraph Document Collection

The gold-standard positive-class documents described in 3.2.1 were randomly-assigned to either a new training set (128 documents), or a hold-out testing set

(33 documents), and were broken up into their composite paragraphs. Because the true structure of the *.pdf*-versions of the documents were somewhat altered, following the previously-described *pdftotext* and regular expression text extraction procedure, we inferred paragraphs based on new-line-separation. Because the original annotation of the data used here was done at the sentence level, rather than at the document, we were able to identify which paragraph or paragraphs in each document contained the information that lead to a positive-class assignment. Thus, each paragraph was assigned either a positive- or negative-class label according to whether it contained annotated information that lead to a positive-class annotation for the document in which it was found. After this procedure, the training collection had 9983 paragraphs (158 positive-class, 9825 negative-class), and the testing collection had 2026 paragraphs (35 positive-class, and 1991 negative-class).

4.2.2 Paragraph Classifier System Design

Rather than run a full series of classifier experiments for this extension of our previously-described system, we used the best-performing classification algorithm identified in the document classification train/classify experiments 3.2.4.

4.2.2.1 Feature Methods

Since many of the types of document-level features used in those experiments (e.g., document title, abstract, and MeSH terms) can't be used in the present task, we investigated the NeuroLex quantitative feature described in our document-level classification experiments 3.2.3.1, as well as using word-based n-grams ($n = 1-5$). In addition, because previous work (3.2) showed that annotated information tended to occur in similar places between documents, we hypothesized that a feature describing where in the document a particular paragraph is located would be helpful for classification. Thus, we created a continuous-valued feature indicating the absolute paragraph order number within a document. Since recursive partitioning tended to be the most useful continuous-valued feature modeling technique in the previous study, we used it here as well.

4.2.2.2 System Evaluation

Since this the classification task described in this study is fundamentally a ranking task, we used AUC as our primary performance metric. System configurations were done by performing five repetitions of 2-way cross-validation on the training data set. The feature configurations that showed some usefulness were then used in a train/classify experiment, in which a model is trained using the training data set, and classified using the hold-out test data set.

As a secondary performance metric for our train/classify experiments, we examined the median number of paragraphs that would need to be read for each publication in the hold-out classification set, in order to identify a paragraph containing information that is relevant to the neuron registry. We re-examined each the best-performing system in the train/classify experiments from this perspective. For the purposes of interpreting this metric, we compared this value to the median number of paragraphs a reviewer would have to read, if they started from the first paragraph, as published in the *pdf* version of the manuscript, and stopped once they reached the sentence(s) leading to the manuscript being included in the NR.

4.3 Results

Our initial classification experiment was done using five repetitions of two-way cross-validation on the training data collection. We examined two sets of system configurations–n-grams, where n=1-5, with paragraph locations, and paragraph locations alone. The results of this experiment are shown in Figure 4.1. From this figure, it is clear that increasing the number of n in the n-grams improves performance, which asymptotically approaches its peak by n = 5. Adding paragraph locations give a small, but not insignificant, performance boost, with maximum performance being achieved at 5-grams with paragraph location information. We ran the same system configurations training on the entire training collection, and evaluating the resultant models against the hold-out testing collection. The results of these experiments are shown in Figure 4.2. The results observed here mirror those seen in the cross-validation experiments–a steady increase in performance with increasing size of n-grams, and a small performance boost obtained by adding paragraph location information. Interstingly, if we compare the

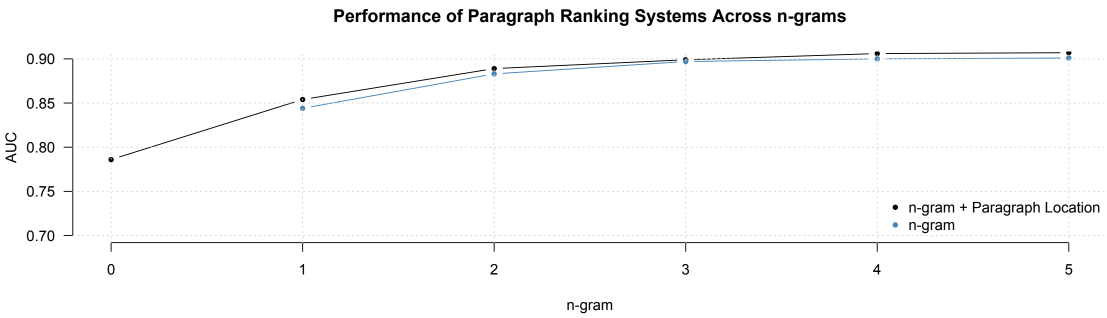

Figure 4.1: Performance of two configurations of the Finna system across increasing n-grams during five repetitions of two-way cross-validation. The n-gram with paragraph location system was better than n-grams alone, but there was not a great difference.

performance increases between 4-gram and 5-gram in the cross-validation and train/classify experiments, it appears that AUC has begun to level off in the train/classify experiments, implying that investigating larger n-grams, or more training data, may not show dramatic improvements in performance. Based on these observations, we chose 4-grams with paragraph location information as our best-performing system, since it performed negligibly worse than the 5-gram system (0.906 v. 0.907), and was considerably faster. Finally, we looked at the median number of paragraphs in the best-performing system (4-grams, and paragraph locations) that would have to be read in each document, in order to find the paragraph containing the positive-class sentence, if they were read in order of the system-generated rankings. We compared this value to the standard approach to document review–starting from the first paragraph, and reading until the information leading to a positive annotation has been identified. Based on looking at the distribution of paragraph rankings and paragraph locations (the standard) (see Figure 4.3), the proper measure of central tendency for these metrics appears

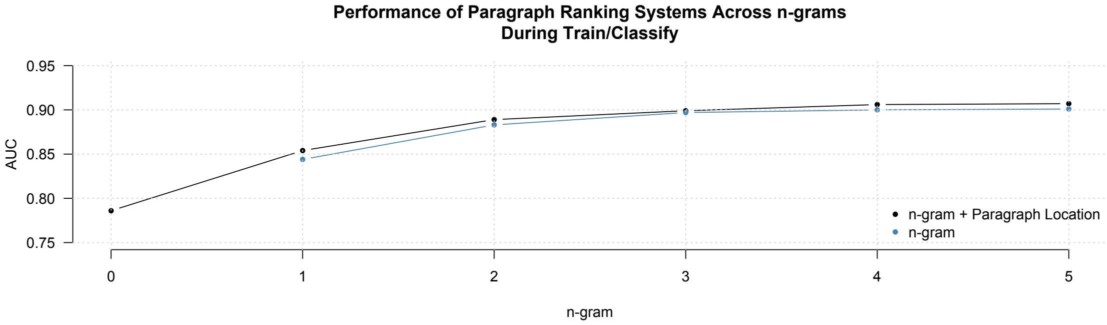

Figure 4.2: Performance of two configurations of the Finna system across increasing n-grams after training on the entire training data set, and classifying the hold-out testing set. The n-gram with paragraph location system was better than n-grams alone, but there was not a great difference.

to be the median. For the *Finna* approach, a median of 2 paragraphs in each document have to be read by annotators, whereas in the standard approach, a median of 6 paragraphs have to be read. The shape of these distributions highlights another interesting difference.

The *Finna* system tended to perform very well on the majority of documents (1-4 paragraphs needing to be read), while the standard approach was more variable ($sd_{Finna} = 9.4$ v. $sd_{Standard} = 15.4$), resulting in many documents requiring over 20 paragraphs to be read. Although *Finna* yielded an average 7.4 paragraphs savings in reading, over the standard approach, there were a few documents that were outliers, in system performance. To delve into our results further, we examined Finna's performance in terms of the document locations of the annotated information (abstract, results, methods, & figure captions). Table 4.1 summarizes the results of this analysis. Overall, the *Finna* system tended to out-perform the standard approach, especially resulting in large reading savings for annotations

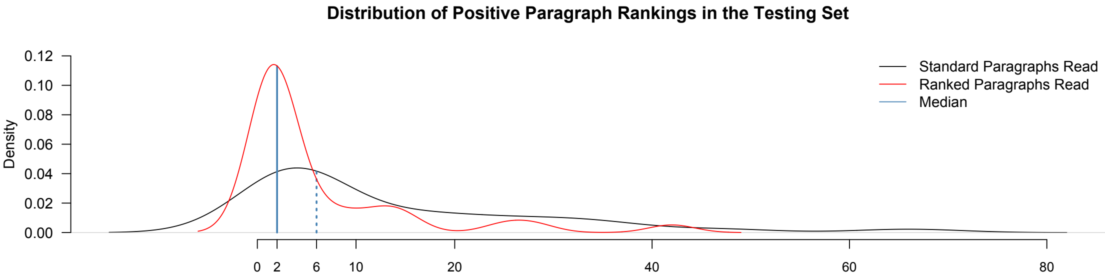

Figure 4.3: Distribution of positive-class paragraph rankings in the hold-out testing data set, after ranking based on a model trained on the training collection. The distribution of the number of paragraphs an annotator would have to read using the standard approach is shown in black, while the number of paragraphs he or she would have to read is shown in red. The median number for each distribution is shown in blue (dotted for the standard approach, and solid for the Finna system).

Subsection	Count	Finna	Standard	Reading Savings (Finna)
Abstract	16	1.1	3.8	2.7
Title	3	5.0	2.0	-3.0
Results	7	11.1	28.7	17.6
Figure Captions	7	13.6	26.0	12.4
Methods	0	NA	NA	NA
Summary	33	2	6	7.4

Table 4.1: Table summarizing the results of the *Finna paragraphs to read* analysis, broken down by the mean results for the document section in which the annotated information was found. The *Finna* system tended to out-perform the standard approach, unless the annotated information was found in the Title section. In the Summary row, text appearing in black is a sum for that column, text in red is the overall median performance for that system, and text in blue is the mean of that column. Data for annotations found in the Methods section are not present here, as, in the hold-out testing document collection, no annotatable information was found in the Methods section.

found in the results section and figure captions. Annotations found in the document title, however, tended to lead to a small loss in paragraphs needing to be read.

4.4 Discussion

The present set of experiments describes a classification system that can be used for rank-ordering the paragraphs within a manuscript, in terms of their likelihood for containing information of interest to a Neuroscience biocuration task. By training off of a subset of expert annotated documents known to contain information that is relevant to the NR, [12] (see Chapter 2) we were able to create a system that performed well, both in terms of AUC (0.906), and a new metric, median number of paragraphs to read in a set of documents (median = 2).

We observed some variation in system performance, depending on where the annotatable information was found in the original document. In general, *Finna* lead to a savings in the amount of reading needing to be done by annotators, as compared with the standard approach, except in the case of information found in document titles. Although the standard approach is a logical standard of comparison, in terms of evaluating the practical utility of a paragraph re-ranking system, for understanding the performance from a machine learning perspective, this may not be the best metric. For example, using our paragraph parsing method, the title is always the first paragraph of the document, so any small changes in paragraph rank by our system can result in apparent performance degradation. Results sections, conversely, tend to occur toward the end of a manuscript, putting paragraph re-rankers at an advantage, from the perspective of evaluating in terms of the number of paragraphs reviewers have to read. Here, our system lead to a mean savings of 17.6 paragraphs, for documents with annotatable information in the results section. This is not to say that median paragraphs to read is a useless metric. Rather, systems such as *Finna*, that need to be optimized in terms of machine learning performance and practical performance alike should be evaluated using more than one metric. Here, we've created a system that performs well using standard machine learning metrics for document ranking (AUC), as well as in terms of a new metric designed to measure performance for how the system will actually be used. Future research should focus on ways to further synthesize these two classes of performance metric, in a way that could be useful for comparing systems that are being deployed to perform a specific task.

Based on our results, this system is ready for evaluation as a component of real-world biocuration workflow, likely working in consort with *Flokka*, our previously-described document classification system[11]. In this situation, biocurators would use the *Flokka* system to identify publications that are highly-likely to contain information of relevance to a particular biocuration task, and the system would automatically re-order the paragraphs of each predicted positive-class publication in terms of their likelihood for having the information that lead to being assigned a positive-class label. In order to be truly useful to a team of biocurators, such a system we need to be evaluated within the context of their workflow–it would need to be integrated in the least-obtrusive way, ideally allowing the curators to focus on adding expert-annotated data to their databases, rather than on using the system.

One limitation of our system is that the data it was trained on was generated by a single Neuroscientist. Although the curator had graduate-level training in the Neurosciences, it is not possible at this time to evaluate the extent to which his annotations are generalizable to the larger population of Neuroscience-related publications. To do this, a data set annotated by multiple experts would be needed, so that inter-annotator agreement statistics could be computed. Since the present set of experiments only used paragraph-level information for classification, one possible extension of the *Finna* system would be to incorporate document-level information into the classification workflow. For example, it may be possible to use the MEDLINE-derived MeSH terms to adjust the prior probability of certain paragraphs containing important information: the occurrence of certain NeuroLex entries [11; 26] in the MeSH terms and certain paragraphs may imply that paragraph is more likely to succinctly communicate the main finding(s) of the manuscript. Beyond this, another way to improve future systems would be to model MEDLINE-related elements of the document (e.g., abstract and title) separately, removing them from the document. This would likely improve *Finna*'s handling of the Title paragraphs.

4.5 Conclusion

In this study we have demonstrated that an automated system can be used to rank-order a publication's paragraphs, in terms of their likelihood for containing information that is of interest to a biocuration task in the Neurosciences. By training on a manually-curated Neuroscience document set that has been pre-filtered to only include documents that do, in fact, contain information of interest, we showed that a simple configuration of an SVM-based classification system can be used to identify paragraphs of interest. We were able to achieve an AUC of 0.906 with our selected system, which was able to complete the classification task on the order of several seconds. Based on its level of performance and its speed, this system could be integrated into a real-world biocuration workflow, and used to streamline the process of curating a knowledge base.

Chapter 5

Discussion

In this thesis we have presented a set of text-mining tools for making the process of biocuration in the Neurosciences more efficient. More specifically, we've focused our work on addressing several work-flow bottlenecks at the Neuron Registry (Figure 1.4). Although our studies focused on one specific neuroscience database more efficient, our approach is generalizable to any community-curated database that has a workflow similar to the Neuron Registry (NR). As community-based curation increases in popularity, in will become more important for the bottlenecks in such a framework to have automated solutions. For example, if the NR were receiving 100 submissions per week, that would require the curation staff to manually verify the content in 100 documents per week, which is not likely to be realistic for a relatively small group. By using the *Flokka* + *Finna* system, the curation staff may be able to automatically weed out erroneous submissions, while efficiently reading only the likely important paragraphs in predicted-positive manuscripts.

One of the most important contributions of this thesis is our general-purpose approach to bootstrapping the startup of computationally-accessible knowledge bases 2. In order to train the types of classifiers used in Chapters 3 and 4, one needs to have access to sufficient training data to build a classifier that can accurately carry out the task at hand. This is not a problem specific to the Neurosciences, or even text-mining, rather, it is one which pervades the entire field of supervised machine learning. Using our approach, we were able to create a gold-standard data set for the training and evaluation of the classifiers needed to help the NR, all while adding entries to the relatively small database. We did this in up to $\frac{1}{6}$ the time standard approaches to data set development would have taken, more than doubling the number of entries in the NR during that time.

Although there is face validity to the tools we created being useful to real-world biocuration tasks, the extent to which they can fit into existing biocuration workflows has yet to be evaluated. An observation in favor of them being successful is that creating the data set described in Chapter 2 was itself a real-world biocuration task. That said, it is likely that every team charged with creating computationally-accessible knowledge bases will have their own particular workflow, some of which will have a more natural integration of our tools than others. As a first pass at examining the utility of our tools outside of the biocuration tasks performed here, we conducted a mock biocuration task that is a subset of the types of tasks performed at the NR. To do this, in collaboration with the Allen Brain Institute, we created a classification system specifically designed to identify documents that are likely to contain information about neuronal gene expression, and deployed our system to an OHSU web-server, using a simple XML-RPC interface. We publicly released the code for using this system onto github[1], in addition to including it in the Appendix of this dissertation (see Chapter 6). We created a classification model based on performance (AUC) during five repetitions of 2-way cross-validation, using a modified version of the data set described in Chapter 3 (modified such that documents associated with gene expression information in the NR were positive-class, and all others were negative-class). The cross-validation results of this study are presented in Figure 5.1. Based on the studies reported in Chapter 3, we expected some configuration of SVM to perform best on this task, however, the overall best-performing system used the *k*IGNN algorithm, title, abstract, and MeSH, obtained from MEDLINE records, along with a quantitative feature representing the number of Allen Brain Institute gene mentions [34; 85] in the full text of the documents. These results imply that, while SVM may be the overall best-performing classifier for the tasks investigated in this thesis, more specific curation tasks may require differently-configured systems. This will be an avenue for future investigation.

Although future studies will need to confirm this, implementing this gene expression classifier system as an XML-RPC service required minimal effort on the part of the author, and one of the curators at the NR (*personal communication*). The interface can take either a single PubMed Identifier (pmid), or a batch of pmids, using the full text-based system (*k*IGNN) if full text can be easily found, and the MEDLINE record-only system (SVM) if it cannot. Since the NR typically request a pmid be associated with a new submission, it would be straightforward for their server to simply pass the pmid through our web service, notifying the

[1] https://github.com/ambertk/NeuronRegistryCorpus.git

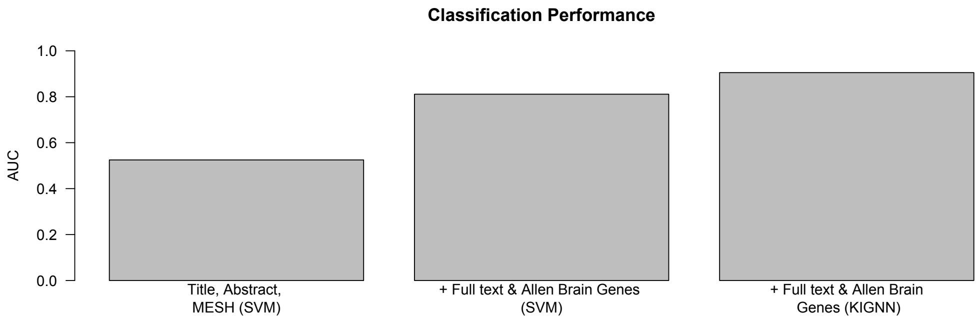

Figure 5.1: Performance of three classification systems for identifying documents gene expression-related information. kIGNN performed best for this problem, in contrast to those described in Chapter 3.

curators if a new submission has been made that is associated with a confidence value below some threshold they can determine.

Another avenue for future investigations lies in examining the change in Neuroscience terminologies over time. Concept drift [93] is a well-documented phenomenon that has been investigated primarily in large corpuses of news reports (e.g., the Reuters corpus), but has received very little attention in the context of the Neurosciences. Neuroscience is a rapdily-changing field [260]; it is made up of a variety of disciplines, each having their own way of communicating things, and each having their own rates of concept change over time. As such, a multi-discipline knowledge base, such as the NR, is likely to be affected by concept drift in some interesting ways that will benefit from study.

5.1 Conclusion

Here, we have presented a suite of text-mining tools that are useful for biocuration tasks in general, and streamlining curation workflows in the Neurosciences, in particular. They can be used in conjunction with one another, or individually–each solving an important biocuration task in its own right. *Virk*, an active learning system for bootstrapping the curation of knowledge bases, was able to more than double the number of entries in the Neuron Registry, while simultaneously generating a gold-standard data set for the creation and evaluation of neuroscience document classifiers, in $\frac{1}{6}$ the time of standard approaches. *Flokka*, a document classification system using Support Vector Machines and information derived from the NeuroLex open-scource ontology of Neurosciences, achieved an AUC of 0.925, showing great promise for its deployment in the real world. *Finna*, a paragraph ranking system was a first approach at solving a relatively new problem in biocuration, achieved an AUC of 0.90, and, based on our results, would require biocurators to read a median of 2 paragraphs for the NR curation task, compared with 6 paragraphs, using the standard reading order. Taken together, our results show that effective text-mining solutions for the neurosciences require the effective use of the open-source resources available to neuroscientists today, and a deep understanding of how neuroscientists communicate in publications.

Chapter 6

Appendix 1: Neuron Registry Data Set

In addition to releasing the data set generated over the course of our active learning experiments in Chapter 2, we have included our data set here. The data are presented in two parts. First, we present the document-level judgements, where documents are denoted by their PubMed ID, an *INCLUDE* judgement indicates it was determined to contain relevant information for the Neuron Registry, and an *EXCLUDE* judgement indicates that it did not. Second, we present the sentence-level annotations for the positive class documents. Documents are again designated according to their PubMed ID, and the respective *SUPPORT* column entries indicate the sentence span containing information leading to an *INCLUDE* judgement. Importantly, because some documents had more than one sentence leading to an *INCLUDE* judgement, some documents have multiple rows.

6.1 Document-level Annotations

JUDGEMENT	PMID
EXCLUDE	19481585
EXCLUDE	19576959
EXCLUDE	19699275
EXCLUDE	19748564
EXCLUDE	19770192
EXCLUDE	19772906
EXCLUDE	19778592
EXCLUDE	19786083
EXCLUDE	19788920
INCLUDE	19796672
EXCLUDE	19804817
INCLUDE	19815003
INCLUDE	19815055
EXCLUDE	19818832
EXCLUDE	19818833
EXCLUDE	19818840
INCLUDE	19819309
EXCLUDE	19822542
EXCLUDE	19825395
INCLUDE	19833105
INCLUDE	19833108
EXCLUDE	19835848
EXCLUDE	19836361
EXCLUDE	19836362
INCLUDE	19837134
INCLUDE	19837136
EXCLUDE	19837137
INCLUDE	19837138
EXCLUDE	19837139
EXCLUDE	19840840
EXCLUDE	19846621
EXCLUDE	19850105
EXCLUDE	19850107
INCLUDE	19850111
EXCLUDE	19853029
INCLUDE	19853587

EXCLUDE	19853643
EXCLUDE	19854241
EXCLUDE	19854242
EXCLUDE	19854243
EXCLUDE	19854244
EXCLUDE	19857467
INCLUDE	19857553
EXCLUDE	19857554
EXCLUDE	19857562
EXCLUDE	19874866
EXCLUDE	19874869
EXCLUDE	19874873
EXCLUDE	19878705
EXCLUDE	19878711
INCLUDE	19879331
EXCLUDE	19879335
EXCLUDE	19879860
EXCLUDE	19879921
EXCLUDE	19879926
EXCLUDE	19879927
EXCLUDE	19883739
EXCLUDE	19884315
EXCLUDE	19889845
EXCLUDE	19889850
EXCLUDE	19892005
EXCLUDE	19895868
EXCLUDE	19895870
EXCLUDE	19895872
EXCLUDE	19896521
INCLUDE	19897018
EXCLUDE	19900959
INCLUDE	19903514
EXCLUDE	19906874
INCLUDE	19906875
EXCLUDE	19906876
EXCLUDE	19906877
EXCLUDE	19906878
EXCLUDE	19906880
INCLUDE	19906884
EXCLUDE	19906885
INCLUDE	19906886

EXCLUDE	19909731
INCLUDE	19909790
EXCLUDE	19909792
INCLUDE	19909793
EXCLUDE	19909794
EXCLUDE	19909796
EXCLUDE	19909797
EXCLUDE	19913606
EXCLUDE	19914223
EXCLUDE	19914335
EXCLUDE	19914337
EXCLUDE	19914353
EXCLUDE	19917566
EXCLUDE	19917569
EXCLUDE	19917570
EXCLUDE	19922772
INCLUDE	19923250
EXCLUDE	19925847
INCLUDE	19925852
INCLUDE	19925855
INCLUDE	19931229
EXCLUDE	19931230
EXCLUDE	19932692
EXCLUDE	19932735
EXCLUDE	19932740
EXCLUDE	19933754
EXCLUDE	19939956
EXCLUDE	19939958
EXCLUDE	19941837
EXCLUDE	19944141
EXCLUDE	19944736
EXCLUDE	19944742
EXCLUDE	19945444
EXCLUDE	19945511
INCLUDE	19948656
EXCLUDE	19948659
EXCLUDE	19948660
EXCLUDE	19958811
EXCLUDE	19958814
EXCLUDE	19958815
INCLUDE	19958820

EXCLUDE	19961902
INCLUDE	19961903
INCLUDE	19961904
EXCLUDE	19961905
INCLUDE	19961906
EXCLUDE	19961908
EXCLUDE	19962370
INCLUDE	19962427
EXCLUDE	19962428
EXCLUDE	19962429
EXCLUDE	19963038
EXCLUDE	19963040
EXCLUDE	19963054
EXCLUDE	19968968
EXCLUDE	19969043
EXCLUDE	20004651
INCLUDE	20004700
EXCLUDE	20004702
EXCLUDE	20004709
EXCLUDE	20004713
EXCLUDE	20005920
EXCLUDE	20005922
INCLUDE	20005924
EXCLUDE	20006674
EXCLUDE	20006677
EXCLUDE	20006679
INCLUDE	20006972
EXCLUDE	20006975
INCLUDE	20018227
EXCLUDE	20018232
INCLUDE	20018832
INCLUDE	20018836
EXCLUDE	20025854
EXCLUDE	20025939
EXCLUDE	20026089
INCLUDE	20026091
EXCLUDE	20026180
EXCLUDE	20026181
EXCLUDE	20026246
EXCLUDE	20026247
EXCLUDE	20026250

EXCLUDE	20026251
EXCLUDE	20026266
EXCLUDE	20026315
EXCLUDE	20026383
EXCLUDE	20032230
INCLUDE	20032232
EXCLUDE	20032240
EXCLUDE	20032241
EXCLUDE	20034478
EXCLUDE	20034544
EXCLUDE	20034545
EXCLUDE	20035829
EXCLUDE	20036314
EXCLUDE	20036315
EXCLUDE	20036714
EXCLUDE	20036717
EXCLUDE	20036723
EXCLUDE	20038443
EXCLUDE	20038444
EXCLUDE	20040367
INCLUDE	20042702
INCLUDE	20043887
EXCLUDE	20043974
EXCLUDE	20043976
EXCLUDE	20045038
EXCLUDE	20045451
EXCLUDE	20045719
EXCLUDE	20045894
EXCLUDE	20045897
INCLUDE	20045899
EXCLUDE	20045901
EXCLUDE	20051233
INCLUDE	20053845
EXCLUDE	20053849
EXCLUDE	20056127
INCLUDE	20056129
EXCLUDE	20056130
EXCLUDE	20056131
INCLUDE	20056135
EXCLUDE	20056139
EXCLUDE	20059989

EXCLUDE	20060032
INCLUDE	20060034
INCLUDE	20060035
EXCLUDE	20060436
INCLUDE	20060438
INCLUDE	20060460
INCLUDE	20060884
INCLUDE	20064491
EXCLUDE	20064853
EXCLUDE	20064855
EXCLUDE	20071623
EXCLUDE	20071625
EXCLUDE	20071632
INCLUDE	20074625
INCLUDE	20074632
EXCLUDE	20079337
EXCLUDE	20079346
EXCLUDE	20079403
EXCLUDE	20079808
INCLUDE	20080147
EXCLUDE	20080149
EXCLUDE	20080150
EXCLUDE	20080152
EXCLUDE	20089816
EXCLUDE	20093174
EXCLUDE	20096330
EXCLUDE	20096331
EXCLUDE	20096333
INCLUDE	20096335
EXCLUDE	20096669
EXCLUDE	20096750
EXCLUDE	20096752
EXCLUDE	20097264
EXCLUDE	20097279
EXCLUDE	20100739
EXCLUDE	20100741
EXCLUDE	20105451
EXCLUDE	20105453
EXCLUDE	20107118
INCLUDE	20107127
EXCLUDE	20107128

EXCLUDE	20107129
INCLUDE	20107132
EXCLUDE	20109529
INCLUDE	20114035
EXCLUDE	20116415
EXCLUDE	20116419
EXCLUDE	20117173
EXCLUDE	20117174
EXCLUDE	20122901
EXCLUDE	20123000
EXCLUDE	20123002
EXCLUDE	20123120
EXCLUDE	20123783
EXCLUDE	20123784
EXCLUDE	20130038
EXCLUDE	20130040
EXCLUDE	20130041
INCLUDE	20130043
EXCLUDE	20132864
EXCLUDE	20132866
INCLUDE	20132867
EXCLUDE	20132869
INCLUDE	20138028
EXCLUDE	20138119
EXCLUDE	20138122
EXCLUDE	20138127
EXCLUDE	20138851
EXCLUDE	20138970
EXCLUDE	20138974
EXCLUDE	20141765
EXCLUDE	20142269
EXCLUDE	20142271
EXCLUDE	20142274
EXCLUDE	20142275
EXCLUDE	20144689
EXCLUDE	20144691
INCLUDE	20144697
EXCLUDE	20144699
EXCLUDE	20147417
EXCLUDE	20149783
INCLUDE	20149841

EXCLUDE	20149842
EXCLUDE	20152880
EXCLUDE	20152881
INCLUDE	20152884
EXCLUDE	20152885
EXCLUDE	20153298
EXCLUDE	20153299
EXCLUDE	20153403
EXCLUDE	20153737
EXCLUDE	20153738
EXCLUDE	20153807
EXCLUDE	20156524
EXCLUDE	20164390
EXCLUDE	20167206
INCLUDE	20167256
EXCLUDE	20167258
EXCLUDE	20167259
INCLUDE	20167261
EXCLUDE	20176001
EXCLUDE	20176082
EXCLUDE	20178832
EXCLUDE	20178834
EXCLUDE	20181735
INCLUDE	20184943
INCLUDE	20184948
EXCLUDE	20184949
EXCLUDE	20188142
INCLUDE	20188149
EXCLUDE	20188152
EXCLUDE	20188154
EXCLUDE	20193738
EXCLUDE	20193741
EXCLUDE	20193743
EXCLUDE	20193750
EXCLUDE	20194124
EXCLUDE	20194125
EXCLUDE	20194127
EXCLUDE	20194131
EXCLUDE	20194133
EXCLUDE	20197064
INCLUDE	20206235

EXCLUDE	20211154
EXCLUDE	20211609
EXCLUDE	20211610
INCLUDE	20211697
EXCLUDE	20211698
INCLUDE	20211700
EXCLUDE	20211704
INCLUDE	20211975
EXCLUDE	20211983
EXCLUDE	20219631
EXCLUDE	20219633
EXCLUDE	20219634
EXCLUDE	20219648
EXCLUDE	20220081
EXCLUDE	20223226
INCLUDE	20223280
INCLUDE	20223282
EXCLUDE	20226230
INCLUDE	20226232
INCLUDE	20226768
INCLUDE	20227462
EXCLUDE	20227464
EXCLUDE	20231147
EXCLUDE	20298684
EXCLUDE	20298749
EXCLUDE	20298759
EXCLUDE	20298762
EXCLUDE	20303337
EXCLUDE	20303338
EXCLUDE	20304030
INCLUDE	20307510
EXCLUDE	20307636
EXCLUDE	20308253
EXCLUDE	20332016
EXCLUDE	20338151
EXCLUDE	20338225
EXCLUDE	20338226
INCLUDE	20346391
INCLUDE	20347011
INCLUDE	20347939
EXCLUDE	20347945

EXCLUDE	20350587
EXCLUDE	20350589
EXCLUDE	20351041
INCLUDE	20351049
EXCLUDE	20353762
EXCLUDE	20359525
EXCLUDE	20359526
EXCLUDE	20360027
INCLUDE	20362644
EXCLUDE	20371267
EXCLUDE	20371268
EXCLUDE	20371269
EXCLUDE	20371277
EXCLUDE	20371377
EXCLUDE	20375140
EXCLUDE	20380824
EXCLUDE	20381470
INCLUDE	20381473
EXCLUDE	20381586
EXCLUDE	20381587
EXCLUDE	20381588
EXCLUDE	20381589
EXCLUDE	20382134
EXCLUDE	20382135
EXCLUDE	20382206
EXCLUDE	20385204
EXCLUDE	20385205
EXCLUDE	20388498
INCLUDE	20394799
EXCLUDE	20394801
INCLUDE	20394802
EXCLUDE	20398736
EXCLUDE	20398738
EXCLUDE	20399252
EXCLUDE	20399253
EXCLUDE	20399256
EXCLUDE	20403413
EXCLUDE	20406669
INCLUDE	20416359
EXCLUDE	20417250
INCLUDE	20417252

EXCLUDE	20417256
EXCLUDE	20417694
INCLUDE	20420813
EXCLUDE	20421283
EXCLUDE	20421284
EXCLUDE	20423699
INCLUDE	20430080
INCLUDE	20430082
INCLUDE	20430087
INCLUDE	20430089
EXCLUDE	20430090
INCLUDE	20433897
EXCLUDE	20433898
INCLUDE	20433901
EXCLUDE	20433905
EXCLUDE	20434522
EXCLUDE	20434528
EXCLUDE	20435099
EXCLUDE	20436038
EXCLUDE	20436041
EXCLUDE	20436042
INCLUDE	20438721
INCLUDE	20438805
EXCLUDE	20438808
INCLUDE	20438810
INCLUDE	20438814
EXCLUDE	20438823
EXCLUDE	20438824
EXCLUDE	20442265
EXCLUDE	20451504
EXCLUDE	20451507
INCLUDE	20451586
EXCLUDE	20452401
EXCLUDE	20452406
EXCLUDE	20457221
EXCLUDE	20457223
INCLUDE	20457226
EXCLUDE	20457238
INCLUDE	20460115
INCLUDE	20466037
EXCLUDE	20470764

EXCLUDE	20470865
EXCLUDE	20470866
INCLUDE	20470874
INCLUDE	20471377
EXCLUDE	20471378
EXCLUDE	20472034
INCLUDE	20472035
EXCLUDE	20478276
EXCLUDE	20478278
EXCLUDE	20478356
INCLUDE	20478357
EXCLUDE	20478359
EXCLUDE	20478361
EXCLUDE	20478367
EXCLUDE	20478368
INCLUDE	20488168
EXCLUDE	20493176
EXCLUDE	20493235
EXCLUDE	20493242
EXCLUDE	20493932
EXCLUDE	20498227
EXCLUDE	20498228
EXCLUDE	20498233
INCLUDE	20498234
EXCLUDE	20501327
EXCLUDE	20510339
INCLUDE	20510892
EXCLUDE	20513366
EXCLUDE	20515664
EXCLUDE	20516339
EXCLUDE	20516340
EXCLUDE	20516342
EXCLUDE	20516344
EXCLUDE	20516349
EXCLUDE	20519317
EXCLUDE	20519318
INCLUDE	20519320
EXCLUDE	20537989
EXCLUDE	20538047
EXCLUDE	20538048
EXCLUDE	20540934

EXCLUDE	20540989
EXCLUDE	20542089
EXCLUDE	20542092
INCLUDE	20542093
EXCLUDE	20542094
EXCLUDE	20546705
EXCLUDE	20546711
EXCLUDE	20547144
EXCLUDE	20547679
EXCLUDE	20547682
EXCLUDE	20553803
EXCLUDE	20553817
EXCLUDE	20558148
EXCLUDE	20561510
EXCLUDE	20561561
INCLUDE	20561573
INCLUDE	20561574
EXCLUDE	20566385
EXCLUDE	20570602
EXCLUDE	20570603
EXCLUDE	20570607
EXCLUDE	20570714
EXCLUDE	20570717
EXCLUDE	20573572
EXCLUDE	20580637
EXCLUDE	20580659
EXCLUDE	20580660
EXCLUDE	20580772
EXCLUDE	20580784
EXCLUDE	20580788
INCLUDE	20580801
EXCLUDE	20594945
EXCLUDE	20595020
EXCLUDE	20595050
EXCLUDE	20595051
EXCLUDE	20599476
EXCLUDE	20599478
EXCLUDE	20599586
EXCLUDE	20599592
EXCLUDE	20599821
EXCLUDE	20599833

EXCLUDE	20599835
EXCLUDE	20599836
EXCLUDE	20600590
INCLUDE	20600592
EXCLUDE	20600595
EXCLUDE	20600600
EXCLUDE	20600601
EXCLUDE	20600607
EXCLUDE	20600609
EXCLUDE	20600612
EXCLUDE	20600618
EXCLUDE	20600620
EXCLUDE	20600623
EXCLUDE	20600639
EXCLUDE	20600640
EXCLUDE	20600644
EXCLUDE	20600647
EXCLUDE	20600648
EXCLUDE	20600649
EXCLUDE	20600654
EXCLUDE	20600655
EXCLUDE	20600656
INCLUDE	20600657
INCLUDE	20600667
INCLUDE	20600669
EXCLUDE	20600675
EXCLUDE	20600738
INCLUDE	20600740
EXCLUDE	20603183
INCLUDE	20603186
EXCLUDE	20603189
EXCLUDE	20603191
EXCLUDE	20603193
EXCLUDE	20603331
EXCLUDE	20603333
EXCLUDE	20603337
INCLUDE	20603338
EXCLUDE	20609381
EXCLUDE	20610036
EXCLUDE	20619318
INCLUDE	20620193

EXCLUDE	20620197
EXCLUDE	20620199
EXCLUDE	20621161
EXCLUDE	20624375
EXCLUDE	20624377
INCLUDE	20624793
EXCLUDE	20624794
EXCLUDE	20630476
EXCLUDE	20630477
EXCLUDE	20633613
EXCLUDE	20634182
EXCLUDE	20634183
EXCLUDE	20634184
EXCLUDE	20637742
EXCLUDE	20637833
INCLUDE	20637834
EXCLUDE	20638442
EXCLUDE	20638447
EXCLUDE	20638463
INCLUDE	20638464
EXCLUDE	20638959
EXCLUDE	20639000
EXCLUDE	20639005
EXCLUDE	20643194
EXCLUDE	20643767
EXCLUDE	20643768
EXCLUDE	20643777
EXCLUDE	20650306
EXCLUDE	20650307
EXCLUDE	20654589
EXCLUDE	20654597
EXCLUDE	20654699
EXCLUDE	20654702
EXCLUDE	20655300
EXCLUDE	20655362
EXCLUDE	20659540
EXCLUDE	20660565
EXCLUDE	20660566
EXCLUDE	20660568
EXCLUDE	20673789
EXCLUDE	20674554

INCLUDE	20674557
EXCLUDE	20674683
INCLUDE	20674684
INCLUDE	20674686
INCLUDE	20674687
EXCLUDE	20675813
INCLUDE	20678546
INCLUDE	20678549
EXCLUDE	20678553
EXCLUDE	20678554
EXCLUDE	20678556
EXCLUDE	20679351
INCLUDE	20679355
EXCLUDE	20685230
EXCLUDE	20685231
INCLUDE	20685388
EXCLUDE	20685612
EXCLUDE	20688135
INCLUDE	20691167
EXCLUDE	20691766
INCLUDE	20691767
EXCLUDE	20693292
EXCLUDE	20693294
EXCLUDE	20696148
EXCLUDE	20696211
EXCLUDE	20696214
EXCLUDE	20696229
EXCLUDE	20705061
EXCLUDE	20705119
INCLUDE	20707989
EXCLUDE	20708656
EXCLUDE	20709036
EXCLUDE	20709146
EXCLUDE	20709148
EXCLUDE	20709149
EXCLUDE	20709151
INCLUDE	20709153
EXCLUDE	20710039
EXCLUDE	20713023
INCLUDE	20713027
EXCLUDE	20723582

INCLUDE	20724365
EXCLUDE	20727397
EXCLUDE	20727939
EXCLUDE	20727945
INCLUDE	20727947
INCLUDE	20727948
EXCLUDE	20727949
EXCLUDE	20728435
EXCLUDE	20728507
EXCLUDE	20728508
EXCLUDE	20732309
EXCLUDE	20732394
EXCLUDE	20735999
EXCLUDE	20736419
INCLUDE	20736420
EXCLUDE	20800646
INCLUDE	20800648
EXCLUDE	20800661
INCLUDE	20800662
EXCLUDE	20800664
EXCLUDE	20804821
INCLUDE	20807519
EXCLUDE	20807792
INCLUDE	20807794
EXCLUDE	20810359
EXCLUDE	20813167
INCLUDE	20813177
EXCLUDE	20813178
EXCLUDE	20816724
EXCLUDE	20816763
EXCLUDE	20816918
EXCLUDE	20816923
EXCLUDE	20816926
EXCLUDE	20817076
INCLUDE	20817079
INCLUDE	20819943
EXCLUDE	20819944
EXCLUDE	20828545
EXCLUDE	20833228
EXCLUDE	20833230
EXCLUDE	20837104

EXCLUDE	20837105
INCLUDE	20837107
EXCLUDE	20837108
EXCLUDE	20837109
EXCLUDE	20837640
EXCLUDE	20837644
EXCLUDE	20840842
EXCLUDE	20843834
INCLUDE	20846512
EXCLUDE	20849913
EXCLUDE	20849921
EXCLUDE	20849931
INCLUDE	20850419
EXCLUDE	20851161
EXCLUDE	20851162
EXCLUDE	20851169
INCLUDE	20851170
EXCLUDE	20854877
EXCLUDE	20854880
INCLUDE	20854882
EXCLUDE	20855438
EXCLUDE	20858463
INCLUDE	20858468
EXCLUDE	20868728
EXCLUDE	20868730
EXCLUDE	20868734
EXCLUDE	20869350
EXCLUDE	20870010
EXCLUDE	20870012
EXCLUDE	20870014
EXCLUDE	20875798
EXCLUDE	20875840
EXCLUDE	20875843
EXCLUDE	20876203
EXCLUDE	20883673
EXCLUDE	20884323
INCLUDE	20884331
EXCLUDE	20884333
EXCLUDE	20888891
EXCLUDE	20889487
EXCLUDE	20889488

EXCLUDE	20921195
EXCLUDE	20921201
EXCLUDE	20921202
EXCLUDE	20923697
EXCLUDE	20933576
EXCLUDE	20933580
EXCLUDE	20933583
INCLUDE	20937710
EXCLUDE	20937711
INCLUDE	20950672
EXCLUDE	20950673
EXCLUDE	20951682
EXCLUDE	20951774
EXCLUDE	20952374
EXCLUDE	20952375
EXCLUDE	20955770
INCLUDE	20961999
EXCLUDE	20962000
EXCLUDE	20962003
EXCLUDE	20965158
EXCLUDE	21037311
EXCLUDE	21037312
INCLUDE	21041525
EXCLUDE	21078598
EXCLUDE	21078602
EXCLUDE	21109037
EXCLUDE	21123203
EXCLUDE	21123204
EXCLUDE	21173085

6.2 Sentence-level Annotations

PMID	SUPPORT
19796672	As for their immunohistochemical localization, the AQP-2 protein is expressed on the basal side of the basal cells of the SV, and proteins of AQP-3 and V2-R are expressed on the apical side of the basal cells.

19815003 In both the normal and epileptic hippocampus, aromatase was detected in numerous CA1-CA3 pyramidal neurons, in granule cells of the dentate gyrus and in interneurons that co-expressed the calcium-binding proteins calbindin, calretinin or parvalbumin.

19815055 In the present study, the protein expression and the electrophysiological characteristics of HCN channels were investigated in nodose ganglion (NG) afferent neurons (A-fiber and C-fiber neurons) from sham and streptozotocin (STZ)-induced diabetic rats.

19815055 Protein expression of HCN channel isoforms in nodose ganglia from sham and diabetic rats, measured by Western blot (A-D). Data are meanSE, n8 rats in each group.

19819309 A typical GABAergic AC had a soma larger than 10 m in diameter with multiple long processes (left, Fig. 1A). The cell was positive to the anti-GABA antibody (right, Fig. 1A).

19833105 The SCG10 positive cells were scattered throughout the neuronal layer in tissue from all ages analyzed but appeared to be more numerous in young animals.

19833108 In the cerebellum, some labeled cells were observed in the deep cerebellar nuclei and in the Purkinje cell layer.

19833108 LGI1 expression in the hippocampal formation was restricted to the pyramidal and granular layers, whereas scattered labeling was observed outside these areas (Fig. 2C).

19833108 Weak LGI4 labeling was observed in the pyramidal and granular layers of the hipoccampal formation, whereas scattered putative interneurons showed strong labeling (Fig. 6B).

19837134 As is known, disinhibition in the CA1-region can cause a depolarization of pyramidal cells.

19837136 Furthermore, as revealed by costaining with an antibody against ChAT, the dendrites of these cells showed co-stratification with the processes of starburst amacrine cells (Fig. 1D).

19837136 In retinal vertical sections, the dendrites of these cells were found to bistratify in the inner plexiform layer (IPL) (Left panel in Fig. 1D).

19837136 The cell somas were located in both the inner nuclear layer (INL) and GCL (Left panels of Fig. 2A, B).

19837136 The cell somas were located in the ganglion cell layer (GCL), and all of them were ganglion cells because each of them displayed an axon, as indicated by an arrow in Fig. 1C (Also see left panel in Fig. 1D).

19837136 These EYFP-positive cells were found to be starburst amacrine cells, the only cholinergic cells in the retina.

19837138 Twenty-four hour NMDA (10 M) exposure produced marked neurodegeneration (350% of control cultures) in the CA1 pyramidal cell region that was significantly reduced by co-exposure to ifenprodil or DL-2-Amino-5-phosphonopentanoic acid (APV).

19850111 The neurons inside the dorsal root ganglia of the lumbar area express Cdh7 and Cdh20 weakly, but not Cdh19 (drg in Fig. 3S, U).

19853587 To determine the functions of the other family members, 4.1G and 4.1B, we observed their expression patterns in developing stereocilia in mice inner ear hair cells. 4.1G is expressed in the basal tapers of the stereocilia bundle in early postnatal stages. 4.1B was specifically and constantly expressed in the stereocilia tips during postnatal development. Additionally, we found that 4.1B is ablated in the hair cells of both myosin XV and whirlin mutant mice at all stages in hair cell development.

19857553 No difference was detected in CB mRNA and protein levels between aged and adult rats (P ¿ 0.05).

19879331 CFP-expressing cells in retinas with optic nerve transection.

19897018 In this report we use BK channels in frog (Rana pipiens) hair cells to monitor dynamic changes in intracellular Ca2 concentration during transient influxes of Ca2, showing that BK current magnitude and delay to onset are correlated with the rate and duration of Ca2 entry through Ca2 channels.

19903514 All the OHC nuclei (red) were normal/intact, but prestin staining in some OHCs was lower compared to the others.

19906875 LTP Based on the preceding data, we deduce that HFS activates the NMDA receptors on spinal dorsal horn neurons, which leads to endogenous ROS generation, and then ROS leads tofEPSP slope (% of control) 0 20 40 60 80 Time (min) 100 120.

19906884 BCs then release glutamate onto downstream ganglion cells (GCs) and amacrine cells (Acs).

19906886 The pre-Bo - tC region was functionally identified in the VRC, where I and E neurons were found, by the typical tachypneic response produced by DLH microinjection into the left VRC.

19909790 The chemoanatomical organization of the visual sector of the cat's thalamic reticular nucleus (TRN)–that is at the dorsal lateral geniculate nucleus (dLGN) and at the pulvinar nucleus (Pul)–was investigated with two novel cytoarchitectonic markers.

19909790 The labeled neurons showed typically fusiform morphology with dendrites orienting in the plane of TRN.

19909793 Bright-field microautoradiographs from Cresyl Violet stained brain sections showing BDNF mRNA labeled cells (black grains) in the layers II of cerebral cortex (CTX), in the CA3 pyramidal cell layer (CA3) and in the dentate gyrus (DG) of the hippocampal formation.

19923250 Potassium current inhibited by 5 mM 4-Aminopyridine treatment.

19925852 The VCA was negative, as was the DCN (Fig. 3d), including the granule cell layer.

19925855 We found that excitatory postsynaptic currents (EPSCs) of pyramidal neurons were rapidly depressed by 0.1 Hz stimulation in acutely prepared slices from rats at 11-12 postnatal days, while this phenomena disappeared in slices from young adolescent rats (23-24 postnatal days).

19931229 The present study investigated the anatomical distribution of cannabinoid-1 receptor (CB1r) in the LC and its association with mu-opioid receptor (MOR).

19948656 We show that the speed of ramp-like mechanical stimulation determines the dynamics of mechanically activated current responses and hence the type of DRG neuron most likely to be activated.

19958820 Immunohistochemical study revealed a dense network of amylin-immunoreactive (irAMY) cell processes in the superficial dorsal horn of the mice.

19961903 Histological analysis method in pyramidal layer of hippocampal CA3 region was performed following under procedures (Sapolsky et al., 1985).

19961903 In addition, we examined that serum Gc increased by restraint stress aggravated kainic acid (KA)-induced neuronal death in hippocampal CA3 region.

19961904 In parallel, we observed a strong upregulation of prodynorphin mRNA in the spinal cord after CCI, with no changes in the expression of proenkephalin or pronociceptin.

19961906 Application of group I mGluR agonist (RS)-3,5-dihydroxyphenylglycine (DHPG) reversibly suppressed spontaneous inhibitory postsynaptic currents (IPSCs).

19962427 Furthermore, the differential expression of estradiol receptors in the dorsal and ventral MePD did not lead to distinct spine number in these subregions when circulating ovarian steroids peak in proestrus.

20004700 SB significantly reduced threshold shift, central auditory function damage, and cochlear function deficits, suggesting that SB may protect auditory function in NIHL and that the active constituent may be a flavonoid, baicalein.

20004700 This study examined the effects of baicalin, baicalein, and Scutellaria baicalensis (SB) extract against NIHL in a mouse model.

20005924 Acute spinal cord slices from 6 to 10 day old mice were used to record EPSCs evoked in visually identified superficial DH neurons by dorsal root primary afferent stimulation.

20006972 In capsaicinsensitive DRG neurons from wild-type mice, acid (¿pH 5.0) evoked [Ca2]i increases, but not in DRG neurons from transient receptor potential V1 (TRPV1) (/) mice.

20018227 Retrograde tracing using a patch loaded with Fast blue (FB) was applied to all four chambers of the rat heart and labeled cardiac spinal afferents were characterized by using three neurochemical markers.

20018832 Group I metabotropic glutamate receptors (mGluRs) activate median preoptic nucleus (MnPO) neurons and induce an inward current.

20018836 A: application of nipecotic acid (1 mM), a nonselective GABA transporter inhibitor, significantly increased the amplitude of Itonic.

20026091 The D1 dopamine receptor (D1R) antibody-labeled small cells resembling medium spiny neurons (black arrows).

20032232 Phosphoinositide-3-kinase (PI3K) localizes to the olfactory cilia and can be activated by odorants. A: Western blot analysis of the catalytic p110 subunit using a pan-specific PI3K antibody (p110pan) and antibodies against p110 and p110 in rat spleen extract, deciliated olfactory receptor neuron (ORN) membranes, and olfactory cilia-enriched membranes.

20042702 Figure 4B shows that nearly all chopper units (5/7) of our dataset had very different ISI statistics when only their phase preference was maintained. This in turns means that most units showed significant mode-locking behavior.

20043887 These data reveal that the basal dendritic trees of cells in A1 continue to grow for a much longer period, and attain almost double the number of spines, as compared with those in V1.

20045899 These two populations can also be discriminated by the presence of pro-inflammatory peptides and by the expression of neurotrophin receptors; IB4neurones have high levels of neuropeptides such as calcitonin gene-related peptide and substance P, and express receptors for nerve growth factor, whereas IB4+ neurones are neuropeptide poor and express receptors for glial cell line-derived neurotrophic factor.

20053845 Effects of thalamic neuromodulators on FS cells in barrel cortex.

20056129 After facial nerve axotomy, TLR2 mRNA was significantly upregulated in the facial motor nucleus and co-immunofluorescence localized TLR2 to CD68+ microglia, but not GFAP+ astrocytes.

20056135 Purkinje cells of the cerebellum are irSST.

20060034 Quantification of zif268- and Homer1a-labeled CA1 neurons.

20060035 Properties of synaptic transmission from the reticular formation dorsal to the facial nucleus to trigeminal motoneurons during early postnatal development in rats.

20060438 Functional and in situ hybridization evidence that preganglionic sympathetic vasoconstrictor neurons express ghrelin receptors.

20060460 Gentamicin is ototoxic to all hair cells in the fish lateral line system.

20060884 IK,n was blocked by the KCNQchannel blockers, linopirdine (100 lM) and XE991 (10 lM), but was insensitive to both IK,f blocker, tetraethylammonium (TEA), and IK,s blocker, 4-aminopyridine (4-AP).

20064491 At PND 8, Ucn 1-ir was present in pIIIu of all mice examined and had increased 3.2 fold from levels at PND 4. Another 2.4 fold increase was observed at PND 12, with mature cell counts leveling off at PND 16, as similar values were observed in the late juvenile/early adolescent mice at PND 24.

20064491 Brightfield immunohistochemical staining for Ucn 1 and CART showed that Ucn 1immunoreactivity (ir) was absent at PND 1, while CART-ir was already apparent in pIIIu at birth, a finding indicating that although the pIIIu neurons have already migrated to their adult position, Ucn 1 expression is triggered in them at later postnatal stages. Ucn 1-ir gradually increased with age, approaching adult levels at PND 16.

20064491 CART-positive cells were strongly labeled in pIII starting at PND 1.

20064491 In contrast, CART is present in pIIIu and other brain regions at both ages.

20064491 Representative sagittal sections showing postnatal development of CART-ir in pIII at high magnification. CART-positive cells in pIII at PND 1 (A), PND 4 (B), PND 8 (C), PND 12 (D), PND 16 (E), and PND 24 (F).

20064491 Representative sagittal sections showing postnatal development of Ucn 1-ir in pIIIu at high magnification. Ucn 1-positive cells in the pIIIu at PND 1 (A), PND 4 (B), PND 8 (C), PND 12 (D), PND 16 (E), and PND 24 (F).

20074625 CGRP positive neurons treated with riluzole showed a significant increase in neurons with complex branching (42.3.01.2%, Fig. 1C-E) compared with those with no outgrowths (23.41.3%) and to complex outgrowths in vehicle treated cultures (25.72.0%, Fig. 1A, B). Quantitative analysis of the longest neurite per cell (Fig. 5) showed that riluzole significantly increased neurite length in CGRP positive neurons (310.622.3 m) compared to vehicle treated cultures (192.536.5 m).

20074625 Examples of CGRP positive neurons in vehicle treated (A, B) and riluzole treated (C, D) cultures. Neurons are stained with III Tubulin (A, C) to identify soma and neurites and are co-stained with CGRP (C, D) to identify the specific DRG subpopulation. (E) shows quantitative analysis of neurite branch pattern in the CGRP positive neurons.

20074625 Examples of IB4 positive neurons in vehicle treated (A, B) and riluzole treated (C, D) cultures. Neurons are stained with III Tubulin (A, C) to identify soma and neurites and are co-stained with IB4 (C, D) to identify the specific DRG subpopulation. (E) shows quantitative analysis of neurite branch pattern in the IB4 positive neurons.

20074625 This study explored the effects of exogenous administration of 0.1 M riluzole on the neurite growth of specific subpopulations of adult rat dorsal root ganglion (DRG) neurons in vitro. Neuronal branching and neurite length were measured in calcitonin gene related peptide (CGRP), Griffonia simplicifolia Isolectin B4 (IB4), N52 and parvalbumin positive neuronal subpopulations. Riluzole was found to enhance neurite branching in both CGRP and IB4 positive neurons compared to vehicle treated cultures. However, neurite length was only significantly increased in CGRP positive neurons in riluzole treated cultures.

20074632 The purpose of the present study was to examine the effects of E2 on gentamicin-induced apoptotic cell death in outer hair cells. The basal turn organ of Corti explants from p3 or p4 rats were maintained in a tissue culture and exposed to 100 lM gentamicin for 48 h. The effects of E2 on gentamicin-induced outer hair cell loss, JNK activation, and staining for terminal deoxynucleotidyl transferase-mediated biotinylated UTP nick-end labeling (TUNEL) were examined. E2 significantly decreased gentamicin-induced outer hair cell loss in a dose-dependent manner.

20080147 IL-1RI immunoreactivity was detected in some neurons (particularly, CA1 pyramidal cells; Fig. 6C1) as well as astrocytes.

20096335 The results presented here show that the GABAA receptor population which is involved in tonic GABA mediated inhibition of cerebellar granule cells is relatively insensitive to 1,5benzodiazepine as it is to "classical" 1,4-benzodiazepines.

20107127 Paraoxon enhanced the frequency and amplitude of spontaneous excitatory postsynaptic currents (sEPSCs). A and B: representative sEPSC recordings obtained from dentate granule cells before (control) and after application of 3 M paraoxon (PXN).

20107132 Responses to only the four basic taste stimuli were included in the cluster analysis for comparison with those of our previous single unit investigations in the geniculate ganglion (Breza et al. 2006, 2007; Lundy and Contreras 1999). Analysis of agglomeration by way of the screen plot (data not shown) indicated that an abrupt upward deflection occurs around 0.3, separating the neurons into five groups, as indicated by the solid vertical line in the cluster analysis.

20114035 Bath application of flunitrazepam (500 nM) altered the properties of spontaneous GABAergic inhibitory postsynaptic currents (sIPSCs) in whole-cell recordings from rat layer II/III pyramidal neurons (Fig. 2).

20130043 A CSD analysis reveals sinks and sources of Cl ions in different layers. Sources of Cl were revealed in the alveus/str. oriens and upper granule cell layer while sinks were located in the stratum oriens and lacunosum-moleculare and the hilus.

20132867 A slight decrease of current amplitudes was induced by ischemia.

20138028 Newly delaminated ganglion mother cells still expressed VDUP1 (yellow arrowheads).

20144697 Role for ionic fluxes on cell death and apoptotic volume decrease in cultured cerebellar granule neurons.

20149841 In rats treated as neonates with anesthesia-only, Type II neurons demonstrated increased spontaneous and UBD-evoked activity following adult intravesical zymosan treatment whereas Type I neurons demonstrated decreased spontaneous and UBD-evoked activity relative to controls.

20152884 The pre-B-tzinger complex (pre-B-tC), a subregion of the ventrolateral medulla involved in respiratory rhythm generation, contains intrinsically bursting pacemaker neurons.

20167256 Compression was found to shorten the apical, but not basal, dendrites of underlying layer III and V cortical pyramidal neurons and reduced dendritic spines on the entire dendritic arbor immediately.

20167261 Phenytoin concurrently increased background inhibition (Ibg) but decreased background excitation (Ebg).

20184943 Cellular proliferation in the subgranular zone of the adult male Syrian hamster.

20184948 Immunohistochemistry results showed that the Cx30.3 protein was clearly present in the ganglion cells of the SG (Fig. 3a, arrows).

20188149 Properties of GABAergic inputs and glutamate receptors of YFP and YFP CRc.

20206235 Cochleograms showing degree of outer hair cell (OHC, dashed line) and inner hair cell (IHC, solid line) loss as function of percent distance from the apex of the noise exposed cochlea (126 dB, 100 Hz narrowband noise centered at 12 kHz, 2 h) in nine rats allowed to survive for 10 wk (B1, D3, E, F, C1, D1, C3, A4, B4).

20211697 Here we study the expression and localization of BKCa channels and CGRP in the rat trigeminal ganglion (TG) and the trigeminal nucleus caudalis (TNC) as these structures are involved in migraine pain.

20211700 Effects of D1R activation on membrane potential and resting conductances of NAc MSNs. Representative traces from two MSNs showing membrane depolarization (A) and inward current (B) elicited by a 10-min bath application of the D1R agonist, SKF-38393 (30 M).

20211975 BCT depolarisation evoked short-latency, AMPA/kainate receptor-mediated EPSCs in connected GCL neurons.

20211975 To address this, I have made paired recordings from BC terminals (BCTs) and neurons in the ganglion cell layer (GCL) in goldfish retinal slices.

20223280 Expression of R1 in the rat retina.

20223282 TOOTH PULP INFLAMMATION INCREASES BRAIN-DERIVED NEUROTROPHIC FACTOR EXPRESSION IN RODENT TRIGEMINAL GANGLION NEURONS.

20226232 Despite the lack of interaction between GABA and glutamate, blocking GABAA receptors significantly accelerated the onset of the Purkinje cell ”ischemic” depolarization (ID), as assessed with current-clamp recordings from Purkinje cells or field potential recordings in the dendritic field of the Purkinje cells.

20226768 Dopamine D5 receptor immunoreactivity is differentially distributed in GABAergic interneurons and pyramidal cells in the rat medial prefrontal cortex.

20227462 In addition, the induction of Hrd1 was concentrated on the granular cell layer and the pyramidal layer, following the distribution character as CA3 ¿ CA2 ¿ CA1 was evaluation of OD value of Hrd1 immunolabelling (F = 26.163, p ¡ 0.001).

20307510 In agreement with the results from CA1 cells, we observed increased GR-IR in the nuclear compartment after glucocorticoid treatment visualized with H300 (Fig. 5A, p ¡ 0.01), but not with M20 (Fig. 5B, p = 0.62).

20346391 Up-regulation of CCR2 receptor protein in the injured DRG.

20347011 Transient forebrain ischemia induced by bilateral common carotid artery occlusion (BCCAO) for 20 min increases cell proliferation in the dentate gyrus (DG) of adult mice.

20347939 The stimulation of a dorsal root with rectangular pulses of 0.5 ms at 0.1 Hz evoked monosynaptic (MSR) and polysynaptic reflex (PSR) potentials in the segmental ventral root.

20351049 Using extracellular recording and voltage-sensitive dye imaging in rat and mouse Purkinje cells, we show that both simple and complex spikes are generated in the proximal axon, 15-20 m from the soma.

20362644 Kainate-induced delayed onset of excitotoxicity with functional loss unrelated to the extent of neuronal damage in the in vitro spinal cord.

20362644 Motoneurons were counted as large ventral horn cells immunopositive for SMI 32 in laminae VIII and IX.

20381473 Photographic representation of the pyramidal neurons of BLA stained with rapid Golgi method.

20394799 Cerebellar granule cells were prepared from 8-day-old rats and cultured as described previously [11], with minor modifications.

20394799 Dose-dependent effects of five compounds on histone H3 acetylation levels in cerebellar granule cells.

20394802 Immunohistochemical labelling obtained with mAChR (M1R-M5R) antibodies on DRG sections, Cy3 conjugated secondary antisera.

20416359 Electrical stimulation of the dorsal root induced a reproducible eEPSC in most of the SG neurons recorded (92%; n = 65), 69.2% of them were monosynaptic or mono plus polysynaptic (Fig. 1).

20417252 Using in situ hybridization and film autoradiography, an obvious 2 mRNA signal in the TH-defined LC (Fig. 1A) was revealed by the two antisense probes designed to hybridize to different locations of human GABAA receptor 2 subunit mRNA.

20420813 Modulation of NMDA and AMPA-mediated synaptic transmission by CB1 receptors in frontal cortical pyramidal cells.

20430080 Cochleograms revealed no gross destruction of hair cells in the non-diabetic groups or the Diabetes-NAC group; however, a significant number of outer hair cells (OHCs) were lost in the Diabetes-Saline group.

20430082 Immunopositive layer V pyramidal cells are observed in all areas of auditory cortex and are consistently the most intensely reactive cells.

20430087 The expression of purinergic receptors (P2X) on rat vestibular ganglion neurons (VGNs) was examined using whole-cell patch-clamp recordings.

20430089 In contrast, salicylate had no effect on the spontaneous or evoked firing of cartwheel cells indicating that salicylate's suppressive effects are specific to fusiform cells.

20433897 Co-localization of cyclin B1- and CDK4-immunoreactivities in cerebellar Purkinje cells labeled with calbindin.

20433901 Muscarine induced firing of BCs.

20438721 In the chicken retina, protocadherin-19 was expressed as early as embryonic day 5 and was localized in the ganglion cell layer, inner plexiform layer, and optic nerve layer.

20438805 Our results indicate that the co-cultures of OECs and SGCs can be successfully established and that both OECs and OEC-CM promote SGCs survival in vitro.

20438805 SGCs survival was most enhanced when co-cultured with OECs. Both Olfactory bulb (OB) and OECs were proved to express BMP-4 and NCAM while BMPR-1A and a7 integrin were also detected in cochlea and SGCs.

20438810 5HT1A receptor density was increased by 23% in the CA1 region of the hippocampus of adult rats treated with 100 g/kg HU210 for 4 days compared to vehicle treated controls. The same treatment increased mRNA expression by 27% and by 14% in the CA1 region and dentate gyrus of the hippocampus.

20438814 Thus, in the presence of CNP the threshold for LTP induction was shifted to higher stimulus frequencies, a modulation that showed layer-specific differences in area CA1. Effects of CNP were prevented by the NPR-B antagonist HS-142-1.

20451586 Most retrogradely labeled cells were located in the ipsilateral medial nucleus of the trapezoid body (MNTB) and contralateral anteroventral cochlear nucleus.

20457226 Recording of evoked fEPSPs was performed by extracellular glass microelectrode (0.6 -1.0 M resistance) using Axopatch-1D amplifier (Molecular Devices, Axon Instruments, Inc., CA, USA) from the hippocampal CA1 pyramidal neurons at the apical dendritic layers.

20460115 In these motoneurons, EPSCs and GABAergic IPSCs were blocked by the application of CNQX, AP-5 and bicuculline.

20466037 A moderate concentration of NE (10 M) and the 1 receptor agonist phenylephrine (10 M) depolarized and increased spontaneous or current injection-evoked spiking in GCs. By contrast, low NE concentrations (0.1-1.0 M) or the 2 receptor agonist clonidine (Clon, 10 M) hyperpolarized and decreased the discharge of GCs.

20470874 Histological analysis of cochleas showed that hair cell lesions are most severe in Sod1-/- Cdh23ahl/ ahl mice followed closely by Sod1-/- Cdh23ahl/ahl mice and much smaller in Sod1-/- Cdh23-/- and Sod1-/- Cdh23-/- mice.

20471377 In addition, PSD reduced the Ih amplitude and the rebound excitability of CA1 pyramidal neurons.

20472035 We observed immunoreactivity for CGRP, CLR and RAMP1, in the human trigeminal ganglion: 49% of the neurons expressed CGRP, 37% CLR and 36% RAMP1.

20478357 In the current study, we explored whether chronic salicylate exposure could induce apoptosis in outer hair cells (OHCs) and spiral ganglion neurons (SGNs) of the cochlea.

20488168 It was shown that the potentiation effect that low concentration Zn2+ (10 -M) exerted on the amplitude of the current mediated by Ca2+permeable AMPA receptors was more remarkable in the presence of moderate concentration of CTZ (20 -M).

20498234 As the GABAergic system is critical for retinal development, we then performed in vivo gramicidin perforated-patch whole-cell recording to characterize the developmental change of GABAergic action in RGCs.

20510892 A sizable population of large spiny neurons in the amygdala and their axons are intensely eYFP+ (Fig. 5, Supplemental material Fig. SM5).

20510892 Again, virtually all of the labeled cells had the morphology of projection neurons; they were either granule cells in dentate gyrus (DG) or pyramidal cells in Ammon's horn fields (CA1, CA2 and CA3) and subiculum (Fig. 1B, Supplementary material Figs. SM1-SM5).

20510892 Substantial numbers of pyramidal neurons in all neocortical areas were eYFP+; we did not observe labeled neurons with a non-pyramidal morphology.

20519320 Interestingly 100 m picrotoxin or BIC potentiated the MSR, depressed the DRP, and produced a long lasting motoneurone after-discharge. Furosemide, a selective antagonist of extrasynaptic GABAA receptors, affects receptor subtypes with 4/6 subunits, and in a similar way to higher concentrations of PTX or BIC, also potentiated the MSR but did not affect the DRP, suggesting the presence of 4/6 GABAA receptors at motoneurones.

20542093 Note the tonic irregular firing of NA-LC neurons during episodes of sniffing (A) and grooming (B, C), characterized by sustained theta waves on the cortical EEG.

20561573 Aminoglycosides are known to enter hair cells via apical endocytosis or permeation of the mechanotransduction channels on the apical surface of hair cells, and presumably from endolymph in vivo.

20561574 We investigated the contribution of systemic inhibition on spike timing in SBCs by iontophoretic application of glycine- and GABA-receptor antagonists (strychnine, bicuculline). Discharge rate increased in one-third of the units during antagonist application, which was accompanied by a deterioration of phase-coupling accuracy in half of those units.

20580801 The cochlear implant electrode array is located in the inner region and the SG neurons lie in the outer region.

20600592 NMDA exposure produced a significant increase in PI uptake in the pyramidal cell layer of the CA1 in METH-naive tissue.

20600657 We found peaks in spike cross-correlograms indicating correlated activity on both fast (peak width 1-50 ms) and slow (peak width¿50 ms) time scales, only in pairs with convergent glomerular projections.

20600667 The improvement of auditory function by FA was paralleled by a significant reduction in oxidative stress, apoptosis and increase in hair cell viability in the organ of Corti.

20600669 Perfusion of slices with SR101 (1 M) for 10 min induced long-term potentiation of intrinsic neuronal excitability (LTP-IE) and a long-lasting increase in evoked EPSCs (eEPSCs) in CA1 pyramidal neurons in hippocampal slices.

20600740 In conclusion, amplitude-modulated chronic electrical stimulation with a high pulse rate does not affect survival, morphology and functionality of spiral ganglion cells with the exception of eABR latencies.

20603186 However, TBI significantly decreased the number of Purkinje neurons (P ¡ 0.05), whereas treadmill exercise significantly alleviated reduction of Purkinje neurons by the TBI (Fig. 1B arrows, P ¡ 0.05).

20603338 The experiments described above with photolytic release of L-glutamate indicate that the facilitation of the mGluR1 current seen with AMPA receptor antagonists at the PF-Purkinje cell synapse is a postsynaptic phenomenon, involving cross-talk between AMPARs and the mGluR signalling pathway.

20620193 Pacemaker currents in mouse locus coeruleus neurons.

20624793 To examine the temporal tuning of the DSGCs, the cells were stimulated with either a grating drifted over the receptive-field centre at a range of velocities or with a light spot flickered at different temporal frequencies.

20637834 Moreover, spontaneous excitatory postsynaptic currents (sEPSCs) were increased by isotonic increases in [Na+]o in the parvocellular neurons. Bath application AMPA receptor antagonist CNQX or non-selective glutamate antagonist kynurenic acid almost completely blocked the sEPSCs.

20638464 A two way ANOVA (strain, age) showed that hair cell number was not significantly affected by strain (F(1,22) - 2.87, p ¿ .05) but was significantly affected by age (F(2,22) - 23.98, p ¡ .0001), with a significant interaction between strain and age (F(2,22) - 9.045, p ¡ .01).

20674557 Expression of ORC3 and ORC5 in cerebellar granule cells differentiating in culture.

20674684 Following SE, IL-18 immunoreactivity was increased in CA1-3 pyramidal cells as well as dentate granule cells.

20674686 Here, we use a different approach to identify and quantify the subpopulations of SPN that contain the mRNA for pituitary adenylate cyclase activating polypeptide (PACAP) or enkephalin.

20674687 Hypoxia (2% O2 for 24 h) also promoted death of DRG neurons, and was further enhanced when mechanical strain and hypoxia were combined.

20674687 Mechanical injury (20% tensile strain) led to significant neuronal cell death (assessed by ethidium homodimer-1 labelling), which was proportional to strain duration (5 min, 1 h, 6 h or 18 h).

20678546 Direction-selective ganglion cells (DSGCs) respond with robust spiking to image motion in a particular direction. Previously, two main types of DSGCs have been described in rabbit retina: the ON-OFF DSGCs respond to both increases and decreases in illumination, whereas the ON DSGCs respond only to increases in illumination.

20678546 Two types of ON direction-selective ganglion cells in rabbit retina.

20678549 The GABAergic projection from the basal forebrain ends selectively on interneurons, specifically on type 1 periglomerular cells and granule cells, and is likely to control the activity of the olfactory bulb via disinhibition of principal cells.

20679355 For example, cerebellar granule neurons cultured for 16 days undergo reproducible inactivating inward sodium current and non-inactivating outward potassium current upon repeated voltage clamped cycles of 0 mV depolarization (Fig. 5A and Suppl. Table 5).

20685388 Inner hair cells compress and rectify the signal.

20691167 Purkinje cell numbers in the female rat cerebellum

20691767 Immature GC were sparsely distributed in the sub granular zone of the DG or the inner third of the granular layer.

20707989 Microstimulation of the granule cell layer of both transverse or sagittal slices evoked a local membrane depolarization restricted to a radial wedge, but these radial responses did not activate measurable molecular layer beams in transverse slices

20709153 Effect of ceramide on cochlear hair cells. Representative microphotographs of hair cells cultured with 10, 100, or 200 M ceramide (without gentamicin) (upper, phalloidin staining). Quantitative analysis of hair cell loss in explants treated with ceramide (without gentamicin) for 48 h (lower). Ceramide itself induced hair cell loss at 150 and 200 M (*one-way ANOVA and Bonferroni test: p ¡ 0.05).

20713027 Within the CA3 the CnB1 and CnB2 isoforms (Sham Figs. 1 and 3) appear to be predominantly in the stratum pyramidale with little expression within the dendritic and axonal layers of the stratum radiatum and stratum oriens.

20724365 The cerebellar cortex is crucial for sensorimotor integration. Sensorimotor inputs converge on cerebellar Purkinje cells via two afferent pathways: the climbing fibre pathway triggering complex spikes, and the mossy fibre-parallel fibre pathway, modulating the simple spike activities of Purkinje cells.

20724365 We show that most Purkinje cells in ipsilateral crus 1 and crus 2 of awake mice respond to whisker stimulation with complex spike and/or simple spike responses.

20727947 Population based quantification of dendrites: evidence for the lack of microtubule-associate protein 2a,b in Purkinje cell spiny dendrites.

20727948 Identification of PTEN in the differentiating HCs of inner ear.

20736420 We found that a direct microinjection of AAV vectors into the vagal nodose ganglia in vivo leads to selective, effective and long-lasting transduction of the vast majority of primary sensory vagal neurons without transduction of parasympathetic efferent neurons.

20800648 DRGs co-cultured with mechanically injured ASTs from C3-deficient mice also showed improved neurite outgrowth.

20800662 In adult rat nodose ganglion neurons, application of 1 M THC caused a significant inhibition of 5-HT3 receptors, extent of which correlated with the density of 5-HTinduced currents, indicating that the observed THC effects occur in mammalian neurons.

20807519 These results indicate that cannabinoid inhibition of nociceptive reflexes produced by WIN-2 and THC may result from inhibition of dorsal horn neurons through a KOR-dependent mechanism.

20807794 Confocal imaging of sEAAT2B labelling in retina A, sEAAT2B labelling (red) with and without background Acridine Orange (AO) (green) nuclear stain in a retinal section. sEAAT2B labels both regular and displaced Off-bipolar cells located in the INL and ONL, respectively. B, single scanning imaging from a double-labelled flat-mounted retina, in which SV2 and sEAAT2B are separately located at photoreceptor terminals and the postsynaptic dendrites, respectively, in the distal OPL.

20807794 DHKA causes a large enhancement in the light-offset current.

20807794 DHKA enhances the light-offset response.

20807794 DHKA-elicited currents in rodand cone-dominated bipolar cells in dark-adapted retinal slices at holding voltage -60 mV, E Cl = -60 mV Aa, DHKA elicits an inward current, which can be blocked by CNQX, in cone-dominated Off-bipolar cells (n = 15).

20807794 Off-bipolar cell synapses (Fig. 4Ab), only EAAT2B uptake is blocked

20807794 The contribution of sEAAT2A to light responses in rodand cone-dominated bipolar cells A, cone-dominated Off-bipolar cells display a transient inward current at the offset of a 2 s light stimulus.

20813177 Two clear populations were identified consistent with: principal neurons which are involved in detecting interaural intensity differences (IIDs) and efferent neurons of the lateral olivocochlear (LOC) system which project to the cochlea.

20817079 The effects of cocaine on BrdU labeling in the SGZ: the effects of chronic cocaine exposure on the number of BrdU+ cells in the SGZ (A), at 1, 3 and 5 days post-labeling (A), and according to spatial distribution along the dorso-ventral axis (B). Chronic cocaine treatment results in a significantly greater number of BrdU+ cells in the SGZ and the total number of BrdU+ cells significantly decreased as the post-labeling time increased from 1 day to 3 days and 1 day to 5 days postlabeling/cocaine abstinence (A).

20819943 Light increases the gap junctional coupling of retinal ganglion cells.

20837107 Postnatal development enhances the effects of cholinergic inputs on recruitment threshold and firing rate of rat oculomotor nucleus motoneurons.

20846512 In 1-month-old Bax-deficient (Bax-/-) mice, distinct subsets of DRG neurons that were immunopositive for TrkA, CGRP, TRPV1 or TrkC, were all increased in number and exhibited cell atrophy compared to wild type DRG neurons.

20850419	In the CA3 region and dentate gyrus the range of variation in mRNA expression was significantly reduced gradually.
20851170	5-Lipoxygenase in mouse cerebellar Purkinje cells.
20854882	The results revealed an agerelated decrease in macular axo-spinous synapses that was not reversed by CR that occurred in the absence of changes in the size of synapses or spines.
20858468	Results showed that application of rat recombinant TNF-(rrTNF) into the cultured normal adult rat DRG neurons increased the immunoreactive (IR) of Nav1.3 localized mainly around the cell membrane and pre-treatment with PDTC blocked the change dosedependently.
20884331	We analyzed the long-term consequences of asphyxial cardiac arrest for hippocampal cell proliferation in rats to evaluate if the ischaemia-induced degenerated CA1 region may be repopulated by endogenous (stem) cells. Analysis of BrdU-incorporation demonstrated an increase at 7, 21 as well as 90 days after global ischaemia in the hippocampal CA1 pyramidal cell layer.
20937710	Vagal sensory neurons are situated in the nodose (placode derived) and jugular ganglion (neural crest derived).
20950672	S1P significantly increased the rate of AMPA-mEPSCs recorded from CA3 pyramidal neurons, without affecting their amplitude (P0.01, Kolmogorov-Simirnov test) (Fig. 1A).
20961999	The intrinsic membrane and firing properties of the pyramidal neurons were not changed by the lesion.
21041525	Western blot data showed that P2X3 receptors were significantly upregulated in doral root ganglion (DRG) of CHF rats whereas VR1 receptors were significantly downregulated.

Bibliography

[1] A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Retchsteiner, K. Verspoor, Z. Wang, and L. Rocha. Uncovering protein-protein interactions in the bibliome. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop, Volume ISBN 84-933255-6*, volume 2, pages 247–255. Citeseer, 2007.

[2] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991. 88, 102

[3] SH Ahmed, M. Graupner, and B. Gutkin. Computational approaches to the neurobiology of drug addiction. *Pharmacopsychiatry*, 42(1):S144–S152, 2009. 40

[4] H. Akil, M.E. Martone, and D.C. Van Essen. Challenges and Opportunities in Mining Neuroscience Data. *Science*, 331(6018):708, 2011.

[5] S.P. Akula, R.N. Miriyala, H. Thota, A.A. Rao, and S. Gedela. Techniques for integrating-omics data. *Bioinformation*, 3(6):284, 2009.

[6] R.B. Altman, C.M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, et al. Text mining for biology-the way forward: opinions from leading scientists. *Genome biology*, 9(Suppl 2):S7, 2008.

[7] K.H. Ambert. A system for identifying neuroanatomical connection information in free text. In *Oregon Health and Science University Student Research Forum*, 2010.

[8] K.H. Ambert and A.M. Cohen. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *Journal of the American Medical Informatics Association*, 16(4):590, 2009. 23, 61, 80

[9] KH Ambert and AM Cohen. k-Information Gain Scaled Nearest Neighbors: A Novel Approach to Classifying Protein-Protein Interactions in Free-Text. *IEEE Transaction on Computational Biology and Bioinformatics*, 2011. 23, 25, 26, 29, 44, 61, 88, 104, 116

[10] KH Ambert and AM Cohen. Text-mining and neuroscience. *International review of neurobiology*, 103:109, 2012. 54, 88

[11] K.H. Ambert, A.M. Cohen, G.A.P.C. Burns, Eilis A. Boudreau, and Kemal Sonmez. *Flokka*: A document prioritization system for curating databases in the neuroscience. *In preparation*, 2013. 135

[12] K.H. Ambert, A.M. Cohen, G.A.P.C. Burns, Eilis A. Boudreau, and Kemal Sonmez. *Virk*: An active learning system for boostrapping new curated neuroinformatics knowledge bases. *In preparation*, 2013. xii, 87, 93, 94, 114, 118, 133

[13] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas. Supporting systematic reviews using text mining. *Social Science Computer Review*, 27(4):509–523, 2009. 126

[14] JR Anderson, BW Jones, JH Yang, MV Shaw, CB Watt, P. Koshevoy, J. Spaltenstein, E. Jurrus, V. Kannan, et al. An ultrastructural framework for neural circuitry mapping. *PLoS Computational Biology*, 2010.

[15] R. Arens. Learning SVM Ranking Function from User Feedback Using Document Metadata and Active Learning in the Biomedical Domain. 2010.

[16] R. Arens. Learning svm ranking function from user feedback using document metadata and active learning in the biomedical domain. *Preference Learning*, page 363, 2010. 53

[17] G.A. Ascoli. From data to knowledge. *Neuroinformatics*, 1(2):145–147, 2003. 40

[18] G.A. Ascoli. Mobilizing the base of neuroscience data: the case of neuronal morphologies. *Nature Reviews Neuroscience*, 7(4):318–324, 2006. 41

[19] G.A. Ascoli. Successes and rewards in sharing digital reconstructions of neuronal morphology. *Neuroinformatics*, 5(3):154–160, 2007.

[20] G.A. Ascoli. The coming of age of the hippocampome. *Neuroinformatics*, pages 1–3, 2010. 40

[21] G.A. Ascoli. The coming of age of the hippocampome. *Neuroinformatics*, 8(1):1–3, 2010.

[22] G.A. Ascoli. Twenty questions for neuroscience metadata. *Neuroinformatics*, pages 1–3, 2012. 15

[23] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[24] N. Ashish, J.L. Ambite, M. Muslea, and J.A. Turner. Neuroscience Data Integration through Mediation: An (F) BIRN Case Study. *Frontiers in Neuroinformatics*, 4, 2010.

[25] N.J. Bahr and A.M. Cohen. Discovering synergistic qualities of published authors to enhance translational research. In *AMIA Annual Symposium Proceedings*, volume 2008, page 31. American Medical Informatics Association, 2008.

[26] A.E. Bandrowski. Biological resource catalog: Nif and neurolex. 2011. 13, 135

[27] R.W. Baughman, R. Farkas, M. Guzman, and M.F. Huerta. The national institutes of health blueprint for neuroscience research. *The Journal of neuroscience*, 26(41):10329–10331, 2006. 19

[28] A. Bernard, S.A. Sorensen, and E.S. Lein. Shifting the paradigm: new approaches for characterizing and classifying neurons. *Current opinion in neurobiology*, 19(5):530–536, 2009.

[29] D. Berrar, N. Sato, and A. Schuster. Quo vadis, artificial intelligence? 2010.

[30] G. Bezgin, A.T. Reid, D. Schubert, and R. K
"otter. Matching spatial with ontological brain regions using java tools for visualization, database access, and integrated data analysis. *Neuroinformatics*, 7(1):7–22, 2009.

[31] J.G. Bjaalie. Understanding the brain through neuroinformatics. *Frontiers in neuroscience*, 2(1):19, 2008. 40

[32] J. Bjorne, F. Ginter, J. Heimonen, S. Pyysalo, and T. Salakoski. Learning to extract biological event and relation graphs. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA'09)*, 2009.

[33] F.E. Bloom, J.H. Morrison, and W.G. Young. Neuroinformatics: A new tool for studying the brain. *Journal of affective disorders*, 92(1):133–138, 2006.

[34] J.W. Bohland, H. Bokil, C.B. Allen, and P.P. Mitra. The brain atlas concordance problem: Quantitative comparison of anatomical parcellations. 2009. 139

[35] J.W. Bohland, C. Wu, H. Barbas, H. Bokil, M. Bota, H.C. Breiter, H.T. Cline, J.C. Doyle, P.J. Freed, R.J. Greenspan, et al. A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. *PLoS Computational Biology*, 5(3), 2009. 43

[36] J.W. Bohland, C. Wu, H. Barbas, H. Bokil, M. Bota, H.C. Breiter, H.T. Cline, J.C. Doyle, P.J. Freed, R.J. Greenspan, et al. A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. *Arxiv preprint arXiv:0901.4598*, 2009. 24

[37] A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *The Journal of Machine Learning Research*, 6:1579–1619, 2005. 64

[38] M. Bota and L.W. Swanson. The neuron classification problem. *Brain research reviews*, 56(1):79–88, 2007. 13

[39] M. Bota and L.W. Swanson. The neuron classification problem. *Brain research reviews*, 56(1):79–88, 2007. 42

[40] M. Bota and L.W. Swanson. Bams neuroanatomical ontology: design and implementation. *Frontiers in neuroinformatics*, 2, 2008. 34

[41] D.M. Bowden, M. Dubach, and J. Park. Creating neuroscience ontologies. *METHODS IN MOLECULAR BIOLOGY*, 401:67, 2007. 12, 34

[42] D.M. Bowden and M.F. Dubach. Neuronames 2002. *Neuroinformatics*, 1(1):43–59, 2003. 12, 90

[43] D.M. Bowden and R.F. Martin. NeuroNames brain hierarchy. *Neuroimage*, 2(1):63–83, 1995. 12, 37, 90

[44] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 99

[45] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262, 2000.

[46] W. Bug, C. Gustafson, A. Shahar, S. Gefen, Y. Fan, L. Bertrand, and J. Nissanov. Brain spatial normalization. *Methods in molecular biology (Clifton, NJ)*, 401:211, 2007.

[47] W.J. Bug, G.A. Ascoli, J.S. Grethe, A. Gupta, C. Fennema-Notestine, A.R. Laird, S.D. Larson, D. Rubin, G.M. Shepherd, J.A. Turner, et al. The nifstd and birnlex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6(3):175–194, 2008. 13, 34

[48] C.J. Bult, J.T. Eppig, J.A. Kadin, J.E. Richardson, J.A. Blake, et al. The mouse genome database (mgd): mouse biology and model systems. *Nucleic acids research*, 36(suppl 1):D724–D728, 2008. 24

[49] S. Burge, T.K. Attwood, A. Bateman, T.Z. Berardini, M. Cherry, C. O'Donovan, et al. Biocurators and biocuration: surveying the 21st century challenges. *Database: the journal of biological databases and curation*, 2012, 2012. 54, 126

[50] G. Burns, D. Feng, and E. Hovy. Intelligent approaches to mining the primary research literature: techniques, systems, and examples. *Computational Intelligence in Medical Informatics*, pages 17–50, 2008.

[51] G. Burns, D. Feng, and E. Hovy. Intelligent approaches to mining the primary research literature: techniques, systems, and examples. *Computational Intelligence in Medical Informatics*, pages 17–50, 2008. 25

[52] G.A.P.C. Burns and W.C. Cheng. Tools for knowledge acquisition within the neuroscholar system and their application to anatomical tract-tracing data. *Journal of Biomedical Discovery and Collaboration*, 1(1):10, 2006.

[53] G.A.P.C. Burns, M. Krallinger, K. Cohen, C. Wu, and L. Hirschman. Studying biocuration workflows. 2009. 25, 126

[54] G.A.P.C. Burns and M.P. Young. Analysis of the connectional organization of neural systems associated with the hippocampus in rats. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1393):55–70, 2000.

[55] L. Cao and P.S. Yu. Behavior informatics: An informatics perspective for behavior studies. *IEEE Intelligent Informatics Bulletin*, 2009.

[56] K.H. Cheung, E. Lim, M. Samwald, H. Chen, L. Marenco, M.E. Holford, T.M. Morse, P. Mutalik, G.M. Shepherd, and P.L. Miller. Approaches to neuroscience data integration. *Briefings in bioinformatics*, 10(4):345–353, 2009.

[57] J.H. Chiang, H.H. Liu, and Y.T. Huang. Condensing biomedical journal texts through paragraph ranking. *Bioinformatics*, 2011.

[58] M. Chicurel. Databasing the brain. *Nature*, 406(6798):822–825, 2000. 36, 38, 44

[59] E. Clarke. Jacyna, ls (1987). *Nineteenth-century origins of neuroscientific concepts.*

[60] John G Cleary, Leonard E Trigg, et al. K*: An instance-based learner using an entropic distance measure. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 108–114. MORGAN KAUFMANN PUBLISHERS, INC., 1995. 102

[61] A.M. Cohen. An effective general purpose approach for automated biomedical document classification. American Medical Informatics Association, 2006.

[62] A.M. Cohen. An effective general purpose approach for automated biomedical document classification. In *AMIA Annual Symposium Proceedings*, volume 2006, page 161. American Medical Informatics Association, 2006. 29, 88

[63] A.M. Cohen. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of the American Medical Informatics Association*, 15(1):32, 2008. 61, 88

[64] A.M. Cohen. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of the American Medical Informatics Association*, 15(1):32–35, 2008. 23

[65] A.M. Cohen, C.E. Adams, J.M. Davis, C. Yu, P.S. Yu, W. Meng, L. Duggan, M. McDonagh, and N.R. Smalheiser. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 376–380. ACM, 2010. 23, 44, 126

[66] A.M. Cohen, K. Ambert, and M. McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association*, 16(5):690–704, 2009. 23, 44, 61, 80, 88, 126

[67] A.M. Cohen, K. Ambert, and M. McDonagh. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA Annual Symposium Proceedings*, volume 2010, page 121. American Medical Informatics Association, 2010. 80, 126

[68] A.M. Cohen, K. Ambert, J. Yang, R. Felder, R. Sproat, B. Roark, K. Hollingshead, and K. Baker. Ohsu/portland vamc team participation in the 2010 i2b2/va challenge tasks. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2*, 2010. 23, 61

[69] A.M. Cohen and W.R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57, 2005. 23, 44

[70] A.M. Cohen and W.R. Hersh. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *Journal of Biomedical Discovery and Collaboration*, 1(1):4, 2006. 88

[71] A.M. Cohen, W.R. Hersh, K. Peterson, and P.Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206, 2006.

[72] C. Crasto, L. Marenco, P. Miller, and G. Shepherd. Olfactory receptor database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic acids research*, 30(1):354–360, 2002.

[73] C.J. Crasto, L.N. Marenco, M. Migliore, B. Mao, P.M. Nadkarni, P. Miller, and G.M. Shepherd. Text mining neuroscience journal articles to populate neuroscience databases. *Neuroinformatics*, 1(3):215–237, 2003. 44

[74] C.J. Crasto, P. Masiar, and P.L. Miller. Neuroextract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *Journal of the American Medical Informatics Association*, 14(3):355, 2007. 44

[75] C.J. Crasto, P. Masiar, and P.L. Miller. Neuroextract: facilitating neuroscience-oriented retrieval from broadly-focused bioscience databases using text-based query mediation. *Journal of the American Medical Informatics Association*, 14(3):355, 2007.

[76] C.J. Crasto and G.M. Shepherd. Managing knowledge in neuroscience. *Methods in Molecular Biology*, 401:3, 2007. 44

[77] S.M. Crook and F.W. Howell. Xml for data representation and model specification in neuroscience. *Neuroinformatics*, page 53, 2007.

[78] C.L. Cunningham, C.M. Gremel, and P.A. Groblewski. Drug-induced conditioned place preference and aversion in mice. *Nature protocols*, 1(4):1662–1670, 2006. 21

[79] A.P. Davis, C.G. Murphy, C.A. Saraceni-Richards, M.C. Rosenstein, T.C. Wiegers, and C.J. Mattingly. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Research*, 37(suppl 1):D786, 2009.

[80] A.P. Davison, T.M. Morse, M. Migliore, G.M. Shepherd, and M.L. Hines. Semi-automated population of an online database of neuronal models (modeldb) with citation information, using pubmed for validation. *Neuroinformatics*, 2(3):327–332, 2004.

[81] F. Denis. PAC learning from positive statistical queries. In *Algorithmic Learning Theory*, pages 112–126. Springer, 1998.

[82] D. Derom, R.A. Schmidt, I. McLeod, and B.A. Hewitt. Using metaneva for structuring, managing and retrieving animal data in the cognitive neurosciences. 2010.

[83] T.G. Dietterich and G. Bakiri. Error-correcting output codes: A general method for improving multiclass inductive learning programs. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, pages 572–577. Citeseer, 1991.

[84] T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Arxiv preprint cs/9501101*, 1995.

[85] H.W. Dong. *The Allen reference atlas: A digital color brain atlas of the C57Bl/6J male mouse.* John Wiley & Sons Inc, 2008. 34, 139

[86] Allison J Doupe, David J Perkel, Anton Reiner, and Edward A Stern. Birdbrains could teach basal ganglia research a new song. *Trends in neurosciences*, 28(7):353–363, 2005. 86

[87] W. Duch. Computational models of dementia and neurological problems. *METHODS IN MOLECULAR BIOLOGY*, 401:305, 2007.

[88] R. El-Yaniv and M. Nisenson. Optimal single-class classification strategies. *Advances in Neural Information Processing Systems*, 19:377, 2007.

[89] S. Ertekin, J. Huang, and C.L. Giles. Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 823–824. ACM, 2007. 64

[90] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 61

[91] R.L. Figueroa, Q. Zeng-Treitler, L.H. Ngo, S. Goryachev, and E.P. Wiechmann. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, 19(5):809–816, 2012. 82

[92] K. Fissell. Workflow-based approaches to neuroimaging analysis. *Methods in molecular biology (Clifton, NJ)*, 4:6, 2007.

[93] G. Forman. Tackling concept drift by temporal inductive transfer. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 252–259. ACM, 2006. 80, 140

[94] L. Forsberg and P. Roland. 1st incf workshop on neuroimaging database integration. 2008. 38

[95] Eibe Frank and Remco Bouckaert. Naive bayes for text classification with unbalanced classes. *Knowledge Discovery in Databases: PKDD 2006*, pages 503–510, 2006. 104

[96] Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H Witten, and Len Trigg. Weka. *Data Mining and Knowledge Discovery Handbook*, pages 1305–1314, 2005. 88

[97] L. French, S. Lane, L. Xu, and P. Pavlidis. Automated recognition of brain region mentions in neuroscience literature. *Frontiers in Neuroinformatics*, 2009. 34

[98] L. French and P. Pavlidis. Informatics in neuroscience. *Briefings in bioinformatics*, 2007.

[99] L. French and P. Pavlidis. Using text mining to link journal articles to neuroanatomical databases. *The Journal of Comparative Neurology*, 2011. 34

[100] L. French, P. Pavlidis, and O. Sporns. Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain. *PLoS Computational Biology*, 7(1):795–799, 2011. 32, 36

[101] L.H. French. Bioinformatics for neuroanatomical connectivity. 2012. 32

[102] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 124–133. MORGAN KAUFMANN PUBLISHERS, INC., 1999. 99

[103] Yoav Freund, Robert Schapire, and N Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999. 99

[104] D. Gardner, H. Akil, G.A. Ascoli, D.M. Bowden, W. Bug, D.E. Donohue, D.H. Goldberg, B. Grafstein, J.S. Grethe, A. Gupta, et al. The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160, 2008.

[105] D. Gardner, H. Akil, G.A. Ascoli, D.M. Bowden, W. Bug, D.E. Donohue, D.H. Goldberg, B. Grafstein, J.S. Grethe, A. Gupta, et al. The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160, 2008. 16, 19, 20

[106] D. Gardner, D.H. Goldberg, B. Grafstein, A. Robert, and E.P. Gardner. Terminology for neuroscience data discovery: multi-tree syntax and investigator-derived semantics. *Neuroinformatics*, 6(3):161–174, 2008.

[107] D. Gardner and G.M. Shepherd. A gateway to the future of neuroinformatics. *Neuroinformatics*, 2(3):271–274, 2004.

[108] D. Gardner, A.W. Toga, G.A. Ascoli, J.T. Beatty, J.F. Brinkley, A.M. Dale, P.T. Fox, E.P. Gardner, J.S. George, N. Goddard, et al. Towards effective and rewarding data sharing. *Neuroinformatics*, 1(3):289–295, 2003.

[109] A. Gelbukh, N.O. Kang, and S.Y. Han. Combining sources of evidence for recognition of relevant passages in texts. *Advanced Distributed Systems*, pages 283–290, 2005.

[110] D.H. Geschwind and G. Konopka. Neuroscience in the era of functional genomics and systems biology. *Nature*, 461(7266):908–915, 2009.

[111] A. Ghazvinian, N.F. Noy, and M.A. Musen. Creating mappings for ontologies in biomedicine: Simple methods work. In *AMIA Annual Symposium Proceedings*, volume 2009, page 198. American Medical Informatics Association, 2009. 35

[112] CD Gore, M. Bányai, PJ Gray, V. Diwadkar, and P. Érdi. Pathological Effects of Cortical Architecture on Working Memory in Schizophrenia. *Pharmacopsychiatry*, 43(S 01):S92–S97, 2010. 41

[113] O.L. Griffith, S.B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M.C. Sleumer, M. Bilenky, M. Haeussler, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic acids research*, 36(Database issue):D107, 2008.

[114] A.A. Grishin, C.E. Gee, U. Gerber, and P. Benquet. Differential calcium-dependent modulation of NMDA currents in CA1 and CA3 hippocampal pyramidal cells. *The Journal of neuroscience*, 24(2):350, 2004. 43

[115] M. Gruenberger, R. Alberts, D. Smedley, M. Swertz, P. Schofield, et al. Towards the integration of mouse databases-definition and implementation of solutions to two use-cases in mouse functional genomics. *BMC Research Notes*, 3(1):16, 2010.

[116] Hu Guan, Jingyu Zhou, and Minyi Guo. A class-feature-centroid classifier for text categorization. In *Proceedings of the 18th international conference on World wide web*, pages 201–210. ACM, 2009. 88, 103

[117] A. Gupta, W. Bug, L. Marenco, X. Qian, C. Condit, A. Rangarajan, H.M. Müller, P.L. Miller, B. Sanders, J.S. Grethe, et al. Federated access to heterogeneous information resources in the neuroscience information framework (nif). *Neuroinformatics*, 6(3):205–217, 2008. 18

[118] M.A. Haendel, N.A. Vasilevsky, and J.A. Wirz. Dealing with data: A case study on information and data management literacy. *PLoS Biology*, 10(5):e1001339, 2012.

[119] P. Hagmann, L. Cammoun, X. Gigandet, S. Gerhard, P. Ellen Grant, V. Wedeen, R. Meuli, J.P. Thiran, C.J. Honey, and O. Sporns. Mr connectomics: Principles and challenges. *Journal of Neuroscience Methods*, 2010.

[120] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. 99

[121] D.J. Hamilton, G.M. Shepherd, M.E. Martone, and G.A. Ascoli. An ontological approach to describing neurons and their relationships. *Frontiers in Neuroinformatics*, 6, 2012.

[122] L. Han, J. van Hemert, and R. Baldock. Automatically Identifying and Annotating Mouse Embryo Gene Expression Patterns. *Bioinformatics*, 2011.

[123] M. Hanke, Y.O. Halchenko, J.V. Haxby, and S. Pollmann. Statistical learning analysis in neuroscience: aiming for transparency. 2010.

[124] R.C. Hardison, J. Oeltjen, and W. Miller. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome research*, 7(10):959, 1997. 37

[125] K.G. Helmer, J.L. Ambite, J. Ames, R. Ananthakrishnan, G. Burns, A.L. Chervenak, I. Foster, L. Liming, D. Keator, F. Macciardi, et al. Enabling collaborative research using the biomedical informatics research network (birn). *Journal of the American Medical Informatics Association*, 18(4):416–422, 2011. 24

[126] W.R. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Verlag, 2009. 15

[127] CC Hilgetag, GA Burns, MA O'Neill, JW Scannell, and MP Young. Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1393):91, 2000.

[128] R.J. Hill, P.W. Sternberg, et al. The gene lin-3 encodes an inductive signal for vulval development in c. elegans. *Nature*, 358(6386):470, 1992. 17

[129] M.L. Hines, T. Morse, M. Migliore, N.T. Carnevale, and G.M. Shepherd. ModelDB: a database to support computational neuroscience. *Journal of Computational Neuroscience*, 17(1):7–11, 2004. 42

[130] L. Hirschman, G.A.P.C. Burns, M. Krallinger, C. Arighi, K.B. Cohen, A. Valencia, C.H. Wu, A. Chatr-Aryamontri, K.G. Dowell, E. Huala, et al. Text mining for the biocuration workflow. *Database: The Journal of Biological Databases and Curation*, 2012, 2012. 17, 25, 55, 126

[131] A. Hobbs. Mapping variation in brain structure and function: Implications for rehabilitation. *The Journal of Head Trauma Rehabilitation*, 14(6):616, 1999. 36

[132] W.T. Hole and S. Srinivasan. Adding neuronames to the umls metathesaurus. *Neuroinformatics*, 1(1):61–63, 2003.

[133] B. Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470, 2003.

[134] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S. St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008. 50

[135] M.Y. Hsiao, D.Y. Chen, and J.H. Chen. Constructing human brain-function association models from fmri literature. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 1188–1191. IEEE, 2007.

[136] http://www.ncbi.nlm.nih.gov/books/NBK3827/. Pubmed help.

[137] M.F. Huerta, Y. Liu, and D.L. Glanzman. A view of the digital landscape for neuroscience at nih. *Neuroinformatics*, 4(2):131–137, 2006.

[138] Q.J.M. Huys, M. Moutoussis, and J. Williams. Are computational models of any use to psychiatry? *Neural Networks*, 2011. 40

[139] F.T. Imam, S.D. Larson, J.S. Grethe, A. Gupta, A. Bandrowski, and M.E. Martone. Nifstd and neurolex: A comprehensive neuroscience ontology development based on multiple biomedical ontologies and community involvement. 13

[140] S.J. JB, R. Inoue, and J. Prasad. SQUINT–SVM for Identification of Relevant Sections in Web Pages for Web Search.

[141] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006. 15

[142] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142, 1998. 27, 88

[143] Thorsten Joachims. Making large scale svm learning practical. 1999. 104, 105

[144] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995. 104

[145] S.H. Joshi, J.D. Van Horn, and A.W. Toga. Interactive exploration of neuroanatomical meta-spaces. 2009.

[146] N. Karamanis, I. Lewin, R. Seal, R. Drysdale, and E. Briscoe. Integrating natural language processing with flybase curation. In *Pac Symp Biocomput*, volume 12, pages 245–56, 2007. 126

[147] S Karkar, S Faisan, L Thoraval, and JR Foucher. A multi-level parcellation approach for brain functional connectivity analysis. In *31st Annual International Conference of the IEE EMBS*, 2009.

[148] N. Kasabov, V. Jain, L. Benuskova, P. Gottgtroy, and F. Joseph. Integration of Brain-Gene Ontology and Simulation Systems for Learning, Modelling and Discovery. *Computational Intelligence in Medical Informatics*, pages 221–234, 2008.

[149] S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649, 2001. 97, 103

[150] DN Kennedy. Making connections in the connectome era. *Neuroinformatics*, 2010.

[151] D.N. Kennedy. The benefits of preparing data for sharing even when you don't. *Neuroinformatics*, pages 1–2, 2012.

[152] Ashraf Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. *AI 2004: Advances in Artificial Intelligence*, pages 235–252, 2005. 88

[153] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. Some effective techniques for naive bayes text classification. *Knowledge and Data Engineering, IEEE Transactions on*, 18(11):1457–1466, 2006. 88

[154] J. Kinoshita and T. Clark. Alzforum. *Methods in molecular biology (Clifton, NJ)*, 401:365, 2007.

[155] SA Knock, AR McIntosh, O. Sporns, R. K
”otter, P. Hagmann, and VK Jirsa. The effects of physiologically plausible connectivity structure on local and global dynamics in large scale brain models. *Journal of Neuroscience Methods*, 2009.

[156] H. Knublauch, R. Fergerson, N. Noy, and M. Musen. The protégé owl plugin: An open development environment for semantic web applications. *The Semantic Web–ISWC 2004*, pages 229–243, 2004.

[157] F.H. Kobeissy, S. Sadasivan, M.W. Oli, G. Robinson, S.F. Larner, Z. Zhang, R.L. Hayes, and K.K.W. Wang. Neuroproteomics and systems biology-based discovery of protein biomarkers for traumatic brain injury and clinical validation. *PROTEOMICS-Clinical Applications*, 2(10-11):1467–1483, 2008. 36

[158] A. Koike, Y. Niwa, and T. Takagi. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236, 2005.

[159] SH Koslow. Should the neuroscience community make a paradigm shift to sharing primary data? *Nature neuroscience*, 3(9):863, 2000. 38

[160] S.H. Koslow. Sharing primary data: a threat or asset to discovery? *Nature Reviews Neuroscience*, 3(4):311–313, 2002. 38

[161] S.H. Koslow and M.F. Huerta. *Neuroinformatics: an overview of the human brain project*. Lawrence Erlbaum, 1997. 35

[162] R.N. Kostoff. Expanded information retrieval using full-text searching. *Journal of Information Science*, 36(1):104, 2010.

[163] R. Kotter. Neuroscience databases: tools for exploring brain structure–function relationships. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1111, 2001. 44

[164] R. Kotter, J. Maier, W. Margas, K. Zilles, A. Schleicher, and A. Bozkurt. Databasing receptor distributions in the brain. *Methods in molecular biology (Clifton, NJ)*, 401:267, 2007. 44

[165] A. Kouznetsov, S. Matwin, D. Inkpen, A. Razavi, O. Frunza, M. Sehatkar, L. Seaward, and P. O’Blenis. Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. *Advances in Artificial Intelligence*, pages 224–228, 2009. 80

[166] M. Krallinger, F. Leitner, and A. Valencia. The biocreative ii. 5 challenge overview. *Proceedings of the BioCreative II. 5 Workshop 2009 on Digital Annotations*, 19, 2009. 80

[167] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology*, 9(Suppl 2):S1, 2008.

[168] M.A. Krogel and T. Scheffer. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Machine Learning*, 57(1):61–81, 2004.

[169] L.I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.

[170] A.E. Kurylas, T. Rohlfing, S. Krofczik, A. Jenett, and U. Homberg. Standardized atlas of the brain of the desert locust, schistocerca gregaria. *Cell and tissue research*, 333(1):125–145, 2008.

[171] H. Lam, L. Marenco, T. Clark, Y. Gao, J. Kinoshita, G. Shepherd, P. Miller, E. Wu, G. Wong, N. Liu, et al. AlzPharm: integration of neurodegeneration data using RDF. *BMC bioinformatics*, 8(Suppl 3):S4, 2007.

[172] H.Y.K. Lam, L. Marenco, T. Clark, Y. Gao, J. Kinoshita, G. Shepherd, P. Miller, E. Wu, G. Wong, N. Liu, et al. Semantic Web Meets e-Neuroscience: an RDF use case. In *Proceedings of International Workshop on Semantic e-Science, ASWC*, pages 158–70. Citeseer, 2006.

[173] H.Y.K. Lam, L. Marenco, G.M. Shepherd, P.L. Miller, and K.H. Cheung. Using web ontology language to integrate heterogeneous databases in the neurosciences. In *AMIA Annual Symposium Proceedings*, volume 2006, page 464. American Medical Informatics Association, 2006.

[174] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine Learning*, 59(1):161–205, 2005. 103

[175] L.S. Larkey and W.B. Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297. ACM, 1996.

[176] S. Larson, F. Iman, R. Bakker, L. Pham, and M. Martone. A multi-scale parts list for the brain: community-based ontology curation for neuroinformatics with neurolex. org. *Neuroinformatics*, 2010. 13

[177] S.D. Larson and M.E. Martone. Ontologies for neuroscience: What are they and what are they good for? *Frontiers in neuroscience*, 3(1):60, 2009. 13

[178] TB Leergaard, CC Hilgetag, and O. Sporns. Research topic mapping the connectome: Multi-level analysis of brain connectivity. *Front. Neuroinform*, 6:14, 2012.

[179] E.S. Lein, M.J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A.F. Boe, M.S. Boguski, K.S. Brockway, E.J. Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2006. 37

[180] E.S. Lein, M.J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A.F. Boe, M.S. Boguski, K.S. Brockway, E.J. Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2006. 32

[181] F. Leitner, S.A. Mardis, M. Krallinger, G. Cesareni, L.A. Hirschman, and A. Valencia. An overview of BioCreative II. 5. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 385–399, 2010.

[182] M.F. Lenzenweger. Schizophrenia: refining the phenotype, resolving endophenotypes. *Behaviour research and therapy*, 37(3):281–295, 1999. 40

[183] F. Letouzey, F. Denis, and R. Gilleron. Learning from positive and unlabeled examples. In *Algorithmic Learning Theory*, pages 71–85. Springer, 2009.

[184] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. 88

[185] A.M.L. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk, and J. Del-Favero. Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome biology*, 12(6):R57, 2011.

[186] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 591–597. JOHN WILEY & SONS LTD, 1997. 64

[187] Nicholas Littlestone. Mistake bounds and logarithmic linear-threshold learning algorithms. 1990. 103

[188] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988. 88, 103

[189] N. Liu. Brain mapping with high-resolution fmri technology. *METHODS IN MOLECULAR BIOLOGY*, 401:195, 2007.

[190] Y. Livneh and A. Mizrahi. A time for atlases and atlases for time. 2009.

[191] B. lomiej Wilkowski. Knowledge discovery in neuroinformatics.

[192] B. lomiej Wilkowski. Semantic approaches for knowledge discovery and retrieval in biomedicine.

[193] A. Losonczy, J.K. Makara, and J.C. Magee. Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature*, 452(7186):436–441, 2008.

[194] W.W. Lytton and M. Stewart. Data mining through simulation. *METHODS IN MOLECULAR BIOLOGY*, 401:155, 2007.

[195] A. MacKenzie-Graham, J. Boline, and A.W. Toga. Brain atlases and neuroanatomic imaging. *METHODS IN MOLECULAR BIOLOGY*, 401:183, 2007.

[196] A. Maedche and S. Staab. Ontology learning for the semantic web. *Intelligent Systems, IEEE*, 16(2):72–79, 2005.

[197] L. Marenco, P. Nadkarni, M. Martone, and A. Gupta. Interoperability across neuroscience databases. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, 401:23, 2007.

[198] L. Marenco, R. Wang, and P. Nadkarni. Automated database mediation using ontological metadata mappings. *Journal of the American Medical Informatics Association*, 16(5):723–737, 2009.

[199] L.N. Marenco, Wang R., G.M.C Shepherd, and P.L. Miller. The nif disco framework: Facilitating automated integration of neuroscience content on the web. *Neuroinformatics*, 2010.

[200] R.F. Martin, J. Dubach, and D. Bowden. Neuronames: human/macaque neuroanatomical nomenclature. In *Proceedings, 14th Annual Symposium on Computer Applications in Medical Care*, pages 1018–9, 1990. 12, 90

[201] M.E. Martone, A. Gupta, and M.H. Ellisman. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nature neuroscience*, 7(5):467–472, 2004. 13

[202] E. Mavritsaki, D. Heinke, H. Allen, G. Deco, and G.W. Humphreys. Bridging the gap between physiology and behavior: Evidence from the sSoTS model of human visual attention. *Psychological review*, 118(1):3, 2011.

[203] S.M. Maynard, C.J. Mungall, S.E. Lewis, and M.E. Martone. A knowledge based approach to matching human neurodegenerative disease and associated animal models. *Neuroscience*, 230, 2010. 13

[204] R. Mazumder, D.A. Natale, J.A.E. Julio, L.S. Yeh, and C.H. Wu. Community annotation in biology. *Biology direct*, 5(1):12, 2010.

[205] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48, 1998. 27

[206] A. McCallum and K. Nigam. Employing em in pool-based active learning for text classification. In *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 350–358, 1998. 64

[207] D.P. McCloskey, M.E. Kress, S.P. Imberman, I. Kushnir, and S. Briffa-Mirabella. From market baskets to mole rats: using data mining techniques to analyze rfid data describing laboratory animal behavior. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 301–306. ACM, 2011.

[208] FA Middleton, C. Rosenow, A. Vailaya, A. Kuchinsky, MT Pato, and CN Pato. Integrating genetic, functional genomic, and bioinformatics data in a systems biology approach to complex diseases: Application to schizophrenia. *METHODS IN MOLECULAR BIOLOGY*, 401:337, 2007.

[209] M. Migliore, I. De Blasi, D. Tegolo, and R. Migliore. A modeling study suggesting how a reduction in the context-dependent input on CA1 pyramidal neurons could generate schizophrenic behavior. *Neural Networks*, 2011. 40

[210] TC Minter. Single-class classification. In *Symposium on Machine Processing of Remotely Sensed Data, pp. 2A12–2A15*, 1975.

[211] Y. Mishchencko, J.T. Vogelstein, and L. Paninski. A bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *Annals of Applied Statistics*, 2010.

[212] S.H. Mitchell and A.J. Rosenthal. Effects of multiple delayed rewards on delay discounting in an adjusting amount procedure. *Behavioural processes*, 64(3):273–286, 2003. 21

[213] T. Mohamed, J. Carbonell, and M. Ganapathiraju. Active learning for human protein-protein interaction prediction. *BMC bioinformatics*, 11(Suppl 1):S57, 2010.

[214] T.P. Mohamed, J.G. Carbonell, and M.K. Ganapathiraju. Active learning for human protein-protein interaction prediction. *BMC bioinformatics*, 11(Suppl 1):S57, 2010. 52

[215] T.M. Morse, P.G. Mutalik, K.H. Cheung, P.L. Miller, and G.M. Shepherd. P1-141: Modeling the “dendritic hypothesis” of Alzheimer pathogenesis: A neuroinformatics approach for elucidating part of the amyloid beta cascade. 2008. 36

[216] A.A. Moustafa and M.A. Gluck. Computational cognitive models of prefrontal-striatal-hippocampal interactions in Parkinson’s disease and schizophrenia. *Neural Networks*, 2011.

[217] H.M. Müller, E.E. Kenny, and P.W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11), 2004. 17, 18

[218] H.M. Müller, A. Rangarajan, T.K. Teal, and P.W. Sternberg. Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics*, 6(3):195–204, 2008. 16, 18

[219] C.J. Mungall, D. Anderson, A. Bandrowski, B. Canada, A. Chatyr, K.C. Aryamontri, P.M. Conn, K. Dolinski, M. Ellisman, J. Eppig, et al. An ontology* based approach to linking model organisms and resources to human diseases.

[220] P. Nadkarni and L. Marenco. Database architectures for neuroscience applications. *METHODS IN MOLECULAR BIOLOGY*, 401:37, 2007.

[221] Tyrone Nicholas. Using adaboost and decision stumps to identify spam e-mail, 2003. 88

[222] F.Å. Nielsen. Lost in localization: A solution with neuroinformatics 2.0? *Neuroimage*, 2009.

[223] F.Å. Nielsen, D. Balslev, and L.K. Hansen. Mining the posterior cingulate: segregation between memory and pain components. *Neuroimage*, 27(3):520–532, 2005.

[224] FA Nielsen, MS Christensen, KH Madsen, TE Lund, and LK Hansen. fmri neuroinformatics. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):112–119, 2006.

[225] F.Å. Nielsen, L.K. Hansen, and D. Balslev. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics*, 2(4):369–379, 2004. 34

[226] F.Å. Nielsen, L.K. Hansen, and D. Balslev. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics*, 2(4):369–379, 2004.

[227] P.V. Ogren. Knowtator: A Protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations*, pages 273–275. Association for Computational Linguistics, 2006. 29

[228] R.D.M. Page and M.A. Charleston. Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution*, 13(9):356–359, 1998. 37

[229] M.S. Palmer, D.A. Dahl, R.J. Schiffman, L. Hirschman, M. Linebarger, and J. Dowding. Recovering implicit information. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 10–19. Association for Computational Linguistics, 1986.

[230] G. Paxinos and C. Watson. *The Rat Brain in Stereotaxic Coordinates: Hard Cover Edition*. Academic press, 2007.

[231] A.R. Pico, T. Kelder, M.P. Van Iersel, K. Hanspers, B.R. Conklin, and C. Evelo. WikiPathways: pathway editing for the people. *PLoS biology*, 6(7), 2008.

[232] John C Platt. 12 fast training of support vector machines using sequential minimal optimization. 1999. 97, 103

[233] S. Pokkunuri, C. Ramakrishnan, E. Riloff, E. Hovy, and G.A.P.C. Burns. The role of information extraction in the design of a document triage application for biocuration. In *Proceedings of BioNLP 2011 Workshop*, pages 46–55. Association for Computational Linguistics, 2011. 25, 126

[234] Z. Qi, GW Miller, and EO Voit. Computational Modeling of Synaptic Neurotransmission as a Tool for Assessing Dopamine Hypotheses of Schizophrenia. *Pharmacopsychiatry*, 43(S 01):S50–S60, 2010. 41

[235] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986. 88

[236] John Ross Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993. 103

[237] C. Ramakrishnan, W.A. Baumgartner Jr, J.A. Blake, G.A.P.C. Burns, K.B. Cohen, H. Drabkin, J. Eppig, E. Hovy, C.N. Hsu, L.E. Hunter, et al. Building the Scientific Knowledge Mine (SciKnowMine1): a community-driven framework for text mining tools in direct service to biocuration. *New Challenges For NLP Frameworks Programme*, page 9. 24, 126

[238] C. Ramakrishnan, A. Patnia, E. Hovy, G.A.P.C. Burns, RH Ramirez-Gonzalez, R. Bonnal, M. Caccamo, D. MacLean, RW Grosse-Kunstleve, TC Terwilliger, et al. Layout-aware text extraction from full-text pdf of scientific articles. *Source Code for Biology and Medicine*, 7(1):7, 2012. 24, 25

[239] Jason D Rennie, Lawrence Shih, Jaime Teevan, David Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 616, 2003. 102

[240] P. Resnick and H.R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997. 51

[241] F. Rinaldi, G. Schneider, and S. Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*, 2012.

[242] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, L. Hunter, et al. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pac Symp Biocomput*, volume 2000, pages 515–524, 2000.

[243] RJ Rodgers and A. Dalvi. Anxiety, defence and the elevated plus-maze. *Neuroscience & Biobehavioral Reviews*, 21(6):801–810, 1997. 21

[244] R. Rodriguez-Esteban, I. Iossifov, and A. Rzhetsky. Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput Biol*, 2(9):e118, 2006. 126

[245] F. Roli and G. Giacinto. Design of multiple classifier systems. *Series in Machine Perception and Artificial Intelligence*, 47:199–226, 2002.

[246] G.D. Rosen, E.J. Chesler, K.F. Manly, and R.W. Williams. An informatics approach to systems neurogenetics. *METHODS IN MOLECULAR BIOLOGY*, 401:287, 2007.

[247] B. Roysam, W. Shain, and G.A. Ascoli. The central role of neuroinformatics in the national academy of engineering's grandest challenge: Reverse engineer the brain. *Neuroinformatics*, 7(1):1–5, 2009. 35, 40

[248] T. Russ, C. Ramakrishnan, E. Hovy, M. Bota, and G. Burns. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. *BMC bioinformatics*, 12(1):351, 2011.

[249] A. Ruttenberg, J.A. Rees, M. Samwald, and M.S. Marshall. Life sciences on the semantic web: the neurocommons and beyond. *Briefings in bioinformatics*, 10(2):193, 2009.

[250] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P.A. Duboué, W. Weng, W.J. Wilbur, et al. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53, 2004.

[251] A. Rzhetsky, D. Wajngurt, N. Park, and T. Zheng. Probing genetic overlap among complex human phenotypes. *Proceedings of the National Academy of Sciences*, 104(28):11694, 2007.

[252] M. Samwald, H. Chen, A. Ruttenberg, E. Lim, L. Marenco, P. Miller, G. Shepherd, and K.H. Cheung. Semantic senselab: Implementing the vision of the semantic web in neuroscience. *Artificial Intelligence in Medicine*, 48(1):21–28, 2010.

[253] F. Sanger and A.R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase* 1. *Journal of Molecular Biology*, 94(3):441–446, 1975. 37

[254] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 839–846. Citeseer, 2000.

[255] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 839–846. Citeseer, 2000. 64

[256] B. Scholkopf, K. Tsuda, and J.P. Vert. *Kernel methods in computational biology*. The MIT Press, 2004.

[257] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191, 2005.

[258] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010. 52, 55, 66, 82

[259] G.M. Shepherd. *The synaptic organization of the brain*. Oxford University Press, USA, 2004. 40

[260] G.M. Shepherd, J.S. Mirsky, M.D. Healy, M.S. Singer, E. Skoufos, M.S. Hines, P.M. Nadkarni, and P.L. Miller. The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends in Neurosciences*, 21(11):460–468, 1998. 9, 36, 140

[261] M.S. Simpson and D. Demner-Fushman. Biomedical text mining: A survey of recent progress. *Mining Text Data*, pages 465–517, 2012.

[262] T.I.A.N. Simulators. Model structure analysis in neuron. *Neuroinformatics*, page 91, 2007.

[263] N.R. Smalheiser. Informatics and hypothesis-driven research. *EMBO reports*, 3(8):702, 2002.

[264] T.F. Smith. The history of the genetic sequence databases. *Genomics*, 6(4):701, 1990. 37

[265] K.M. Spencer. The functional consequences of cortical circuit abnormalities on gamma oscillations in schizophrenia: insights from computational modeling. *Frontiers in Human Neuroscience*, 3, 2009. 41

[266] O. Sporns. Graph theory methods for the analysis of neural connectivity patterns. *Neuroscience databases: A practical guide*, pages 171–186, 2002. 36

[267] O. Sporns. *Networks of the Brain*. The MIT Press, 2010. 35

[268] O. Sporns, D.R. Chialvo, M. Kaiser, and C.C. Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, 2004. 36

[269] O. Sporns, G. Tononi, and G.M. Edelman. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, 10(2):127, 2000. 43

[270] O. Sporns, G. Tononi, and GM Edelman. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, 10(2):127–141, 2000. 24

[271] PR Srinivas, S.H. Wei, N. Cristianini, EG Jones, and FA Gorin. Comparison of vector space model methodologies to reconcile cross-species neuroanatomical concepts. *Neuroinformatics*, 3(2):115–131, 2005. 34

[272] K.E. Stephan, L. Kamper, A. Bozkurt, G.A.P.C. Burns, M.P. Young, and R. Kötter. Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1159, 2001. 51

[273] L.L. Sun and X.Z. Wang. A survey on active learning strategy. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 1, pages 161–166. IEEE, 2010.

[274] Y. Suzuki, R. Yamashita, M. Shirota, Y. Sakakibara, J. Chiba, J. Mizushima-Sugano, K. Nakai, and S. Sugano. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome research*, 14(9):1711, 2004. 37

[275] L.W. Swanson. *Brain maps*. Academic Press, 2004.

[276] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[277] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002. 64, 82

[278] S. Usui. Visiome: neuroinformatics research in vision project. *Neural Networks*, 16(9):1293–1300, 2003.

[279] Ö. Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009. 80

[280] G. Valentini and F. Masulli. Ensembles of learning machines. *Neural Nets*, pages 3–20, 2002.

[281] K.R.A. Van Dijk, T. Hedden, A. Venkataraman, K.C. Evans, S.W. Lazar, and R.L. Buckner. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *Journal of neurophysiology*, 103(1):297, 2010.

[282] J.D. Van Horn and C.A. Ball. Domain-specific data sharing in neuroscience: What do we have to learn from each other? *Neuroinformatics*, 6(2):117–121, 2008.

[283] J.D. Van Horn, S.T. Grafton, D. Rockmore, and M.S. Gazzaniga. Sharing neuroimaging studies of human cognition. *Nature Neuroscience*, 7(5):473–481, 2004.

[284] J.D. Van Horn and A.W. Toga. Is it time to re-prioritize neuroimaging databases and digital repositories? *Neuroimage*, 47(4):1720–1734, 2009. 39

[285] V.N. Vapnik. *The nature of statistical learning theory*. springer, 2000. 29

[286] A. Vercelli. Brain maps and connectivity representation. *Neuroinformatics*, 4(4):319–320, 2006.

[287] P. Villoslada and J.R. Oksenberg. Neuroinformatics in clinical practice: are computers going to help neurological patients and their physicians? *Future Neurology*, 1(2):159–170, 2006.

[288] A. Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

[289] TP Vogels and LF Abbott. Gating deficits in model networks: a path to schizophrenia? *Pharmacopsychiatry*, 40(1):73, 2007. 41

[290] J.B. Voytek, B. Voytek, B. Voytek, and M.B.G. Hall. Automated cognome construction and semi-automated hypothesis generation. 33

[291] B.C. Wallace, T.A. Trikalinos, J. Lau, C. Brodley, and C.H. Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55, 2010.

[292] B.C. Wallace, T.A. Trikalinos, J. Lau, C. Brodley, and C.H. Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55, 2010. 53, 80, 126

[293] C.A. Walsh. Neuroscience in the post-genome era: an overview. *TRENDS in Neurosciences*, 24(7):363–364, 2001.

[294] X. Wan and P. Pavlidis. Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics*, 5(3):161–175, 2007.

[295] Y.W. Webster, R.C. Gudivada, E.R. Dow, J. Koehler, and M. Palakal. A hybrid method to discover and rank cross-disciplinary associations. In *2009 IEEE International Conference on Bioinformatics and Biomedicine*, pages 362–365. IEEE, 2009.

[296] T.C. Wiegers. Developing a Text Mining Prototype for the Comparative Toxicogenomics Database Biocuration Process. 2009. 126

[297] T.C. Wiegers, A.P. Davis, K.B. Cohen, L. Hirschman, and C.J. Mattingly. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database(CTD). *BMC bioinformatics*, 10(1):326, 2009. 126

[298] G.L. Winsor, R. Lo, S.J.H. Sui, K.S.E. Ung, S. Huang, D. Cheng, W.K.H. Ching, R.E.W. Hancock, and F.S.L. Brinkman. Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic acids research*, 33(Database Issue):D338, 2005.

[299] Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations. 1999. 88

[300] L. Wolf, C. Goldberg, N. Manor, R. Sharan, and E. Ruppin. Gene expression in the rodent brain is associated with its regional connectivity. *PLoS computational biology*, 7(5):e1002040, 2011.

[301] S.T.C. Wong and S.H. Koslow. Human brain program research progress in bioinformatics/neuroinformatics. *Journal of the American Medical Informatics Association*, 8(1):103, 2001.

[302] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *Pac Symp Biocomput*, volume 1, pages 408–19, 2001.

[303] J.J. Yang, A.M. Cohen, and M.S. McDonagh. SYRIAC: The SYstematic Review Information Automated Collection System A Data Warehouse for Facilitating Automated Biomedical Text Classification. In *AMIA Annual Symposium Proceedings*, volume 2008, page 825. American Medical Informatics Association, 2008. 18, 23, 126

[304] L. Young, D. Vismer, M.J. McAuliffe, S.W. Tu, L. Tennakoon, A.K. Das, V. Astakhov, A. Gupta, and S. Jeffrey. Ontology driven data integration for autism research. In *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS*, 2009.

[305] H. Yu. Svmc: Single-class classification with support vector machines. In *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 18, pages 567–574. Citeseer, 2003.

[306] H. Yu, C.X. Zhai, and J. Han. Text classification from positive and unlabeled documents. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 232–239. ACM, 2003.

[307] S. Yu, L.C. Tranchevent, B. De Moor, and Y. Moreau. Gene prioritization and clustering by multi-view text mining. *BMC bioinformatics*, 11(1):28, 2010.