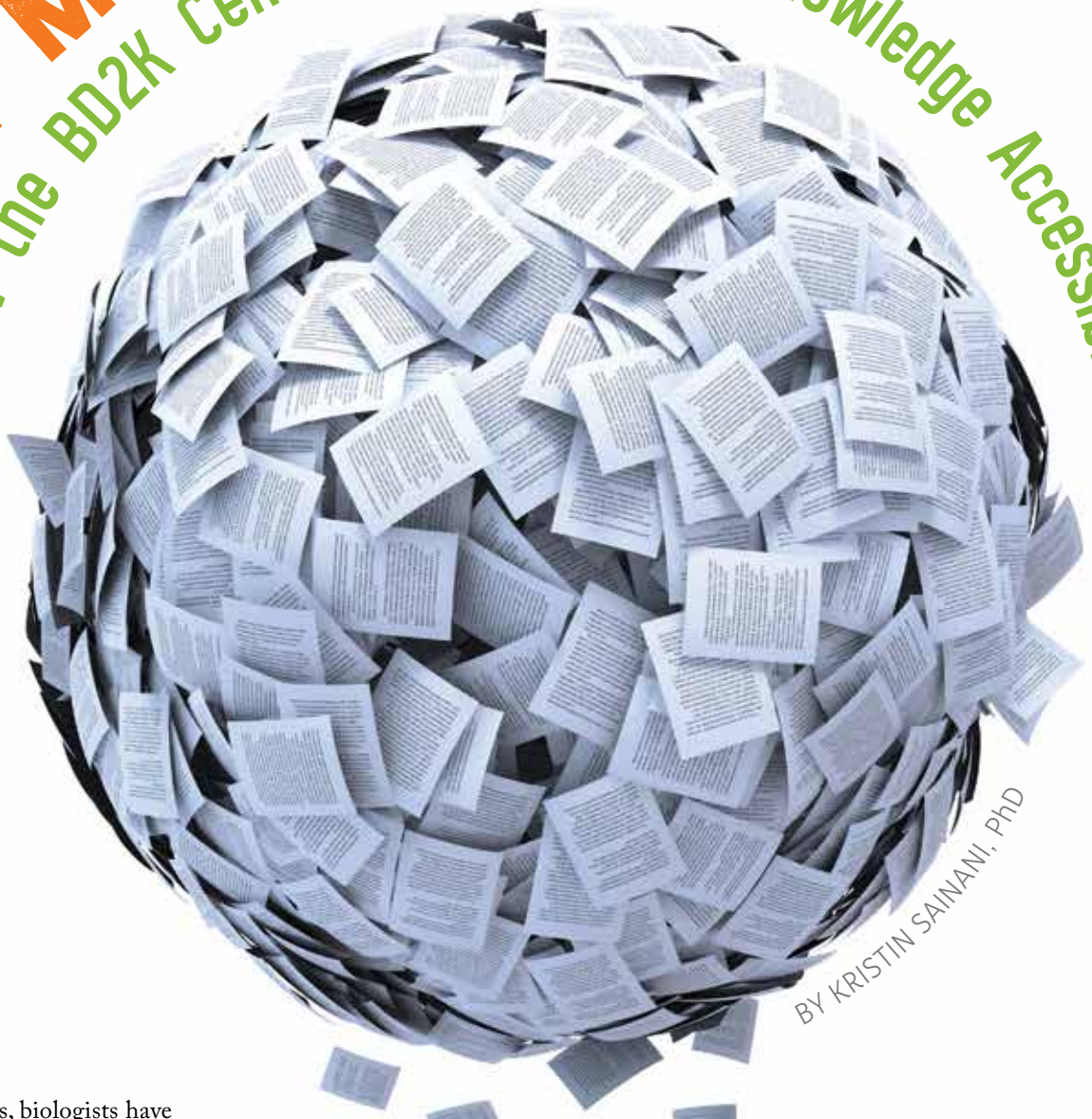
TEXT MINING:

How the BD2K Centers are Making Knowledge Accessible



BY KRISTIN SAINANI, PhD

Since the 1960s, biologists have manually curated data on 6,000 hereditary diseases for the OMIM database (Online Mendelian Inheritance in Man). The database is vital to doctors, who use it for differentially diagnosing genetic conditions; it's much faster and more accurate than asking Dr. Google. But human curators have only been adding about 50 records per month for years, lagging far behind the explosion of information on

gene-disease associations and gene variants currently available in the literature.

What if, instead, computers could curate the literature automatically? What if computers could also scan the millions of papers in PubMed and automatically discover biological networks or predict new uses for existing drugs? These are the many promises of text mining; and some are beginning to come true.

For example, a program called DeepDive, developed by **Christopher Ré**, **PhD**, assistant professor of computer science at Stanford, and data science lead at the **Mobilize Center**, can now quickly and accurately extract data from the text, figures, and tables of scientific papers. When applied to the paleontology literature as a test case, it extracted 100 times more facts from 10 times more papers than human curators, with an accuracy as good or better than that of humans. Unlike human curators, DeepDive doesn't get bored or tired, and it can re-read the entire literature anytime to grab new facts of interest.

In biomedicine, the need for high performance text-mining systems like DeepDive has never been more pressing. Most of the collective knowledge of biomedicine is trapped within published papers or buried within the medical notes and images found in electronic health records (EHRs). If researchers could teach computers to make sense of natural language and pictures, they could unlock this untapped knowledge. But teaching a computer to read is hard; and teaching a computer to read biomedical jargon is even harder. Scientists often write in complicated, convoluted prose; and doctors write in shorthand recognizable only to others in their specialty. Plus, many biological entities have ambiguous names; for example, there are genes named "cheap date", "onion ring", and "pray for elves."

Fortunately, there has been significant progress in biomedical text-mining in the past decade. "There have been a lot of new techniques discovered," says **KnowEnG Center** director **Jiawei Han**, **PhD**, professor of computer science at the University of Illinois, Urbana-Champaign. "A number of groups are moving text mining along to make it real time and high resolution." Specifically, the community has seen advances in two key text-mining tasks: recognizing entities (e.g., genes and drugs), and extracting relationships between entities (e.g., interactions between genes and drugs). Early systems relied on simple approaches such as matching words to dictionaries; making up simple rules (e.g., the word "kinase" denotes a protein); and assuming that two entities that co-occur in the same sentence are related. Later systems improved accuracy by incorporating machine-learning algorithms.

Now, BD2K researchers at the Mobilize, KnowEnG, and **bioCADDIE** Centers are taking text mining to the next level by leveraging recent advancements in machine learning, such as deep learning and active learning. They are also finding ways to address machine learning's biggest bottleneck: the need for large amounts of hand-annotated data to train the systems. Finally, with tools such as DeepDive, they are putting cutting-edge methods in the hands of users. "It will be exciting to explore how some of these BD2K tools can be combined to form a nice, practical text-mining pipeline," says **Jason Fries**, **PhD**, a postdoctoral fellow in Stanford's Mobilize Center.

"It will be exciting to explore how some of these BD2K tools can be combined to form a nice, practical text-mining pipeline," says Jason Fries, PhD, a postdoctoral fellow in Stanford's Mobilize Center.

Even BD2K Centers that are not focusing on text-mining methods per se are getting into the text-mining game by using existing text-mining tools for novel applications, such as cleaning up the metadata in data repositories, an effort that's happening at the **Center for Predictive Computational Phenotyping** (**CPCP**).

"What's exciting is that we're moving from just demonstrating that these methods can extract information with reasonably good accuracy to now figuring out ways that this information can be used," says **Mark Craven**, **PhD**,

director of the CPCP and professor of biostatistics, biomedical informatics, and computer science at the University of Wisconsin-Madison.

RELEVANT NIH INSTITUTES:

NHGRI, NIBIB, NLM, and all disease-focused institutes including NCI, NHLBI, NIDDK, NINDS, NIAID, and NIAMS

Mobilize: Deep Learning for Text Mining

To achieve state-of-the-art performance in text and image mining, researchers at the Mobilize Center are turning to deep learning. Deep learning models are larger and more complex than traditional machine-learning models, and are driving revolutions in computer vision and speech recognition; for example, they power Apple's digital assistant Siri. "A lot of development has gone into deep learning in the last few years. The community has achieved state-of-the-art and lowered the bar to use," says **Alex Ratner**, a doctoral student in Ré's laboratory at the Mobilize Center.

But there's a catch: Deep learning models need massive amounts of annotated training data from which to learn. "You can label a couple hundred examples and can get a simple model to work, but you can't get one of these deep models to work," Ratner says. "Intuitively it makes sense that a much more complex model—one that has tens of millions of parameters—would need commensurately more data."

It might take weeks or even months for a team of graduate students working around the clock to generate enough training data for one text-mining task. Ré's lab is getting around this problem by having computers label the data. The computer-generated training data are imperfect, but, surprisingly: "You can get really good performance even if you have lower quality labels," Ratner says. Their tools—DeepDive (http://deepdive.stanford.edu/) and Snorkel (http://snorkel.stanford.edu)—actually outperform tools that require hand-labeled training data (so-called "supervised" models).

DeepDive automatically annotates training examples with the help of existing knowledge bases, a trick known as "distant" supervision. For example, if an existing database tells us that p53 down-regulates CHK1, then DeepDive would label the sentence: "It was therefore of interest to determine whether p53 affects CHK1," as a positive example of a gene-gene interaction. "You 'lightly' label everything," explains **Emily Mallory**, a doctoral student in the lab of **Russ Altman, MD, PhD**, professor of bioengineering, genetics, medicine, and biomedical data science at Stanford. While the labels can be wrong, DeepDive compensates for these inaccuracies with the sheer volume of examples.

DeepDive has been used in applications as far-flung as automatically curating the paleobiology literature to catching sex traffickers by text mining internet ads. In a 2016 paper in *Bioinformatics*, Mallory used DeepDive to automatically extract gene-gene interactions from more than 100,000 full-text articles from *PLoS One*, *PLoS Biology*, and *PLoS Genetics* (see sidebar: Mining for Gene-Gene Interactions, page 26).

It took Mallory a few months to perfect her gene-gene relation extractor because DeepDive requires multiple rounds of iteration and refinement to optimize performance. DeepDive also requires considerable programming knowhow—beyond the skills of a typical biologist. So Ratner and others on Ré's team have developed Snorkel, a successor to DeepDive that is more streamlined, more user-friendly, and achieves better performance.

Snorkel uses even weaker supervision than DeepDive. "Weak supervision is where you say 'I want to use even noisier input streams,'" Fries explains. Rather than relying on just a single knowledge base, you can throw in anything that might contain even a very noisy signal—hundreds of weakly related knowledge bases; training data labeled by lay annotators; or simple, error-prone rules, such as "anytime two chemicals occur in the same sentence label this as a causal relationship"—and Snorkel is able to learn something

Rather than relying on just a single knowledge base, you can throw in anything that might contain even a very noisy signal ... and Snorkel is able to learn something.

about that signal. Snorkel users input labeling functions via a simple interface that requires only basic programming skills. A typical novice user can write 30 to 40 such labeling functions in hours to days.

The key is that the labeling functions will assign multiple—often conflicting—labels to the same bit of text. Snorkel automatically looks at the patterns of agreement and disagreement to learn which labeling functions are better than others. Labeling functions that mostly agree with other labeling functions are assumed reliable and given the most weight; labeling functions that mostly run counter to the consensus are considered unreliable and given the least weight. The computer then tallies the votes of the "good" and "bad" labeling functions for a given extraction and assigns it a probability—e.g., there is an 85 percent probability that this sentence contains a gene-gene interaction. By assigning probabilistic rather than yes/no labels, "you're actually formally acknowledging and modeling the fact that this is weak and inaccurate supervision, not the ground truth," Ratner says.

These training data are then fed to deep-learning

Mining for Gene-Gene Interactions

In 2016, graduate student Emily Mallory used DeepDive to extract gene-gene interactions from more than 100,000 full-text articles from ***PLoS One, PLoS Biology,*** and ***PLoS Genetics***. Mallory first extracted about 1.7 million sentences containing mentions of at least two genes. She labeled sentences as positive for a gene-gene interaction if the gene pair could be found in the BioGRID or ChEA databases and negative if it could be found in the Negatome database (which documents genes and proteins unlikely to interact). This generated a training set with more than 100,000 imperfectly labeled sentences.

Using these training data, DeepDive learned 724 sentence features useful for classification—for example, the presence of the verb "bind" between two genes. When this model was applied to the 1.6 million unlabeled sentences, it identified 3,356 unique gene pairs where the probability of a true interaction was greater than 90 percent. (To account for uncertainties, including in recognizing gene mentions, DeepDive returns the probability of a true gene-gene interaction rather than a yes/no answer.)

In evaluation against a database of curated protein interactions and manual curation, the system achieved an F1 score of 59 percent, which is on par with state-of-the-art relation extractors that use human-labeled training data. Mallory is now planning to apply the framework to mine gene-gene, gene-disease, and other relations from 500,000 full-text articles available in PubMed Central.

algorithms that can handle uncertainty in the training labels. These algorithms devise a classification model that can be applied to new data for entity tagging or relation extraction.

Snorkel has shown impressive performance. State-of-the art chemical entity taggers that rely on human-labeled data have achieved an accuracy of 88 percent, as quantified by the F1 score (a common accuracy metric in text mining). On the same task, Fries' team got an F1 score of 87 percent with Snorkel when all they fed it was a dictionary. "We did it completely automatically—we just gave it a dictionary. So, it's completely for free," Fries says. Adding some simple labeling functions improved performance. For a harder task—extracting causal chemical-disease relationships from PubMed abstracts—the top team in the 2015 BioCreAtIvE competition achieved an F1 score of 57 percent using 1000 human-labeled PubMed abstracts. Ratner's team built a Snorkel-based extractor that bested this mark without using any human-labeled data. They used 33 labeling functions applied to hundreds of thousands of unlabeled PubMed abstracts. "We can pour in unlabeled data, and we actually get scaling," Ratner says.

Mallory is now collaborating with the FDA to build a Snorkel-based tool for extracting gut microbiome relationships, such as drug-microbiome and chemical-microbiome interactions, from the biomedical literature. Fries and Ratner are working on Snorkel applications that extract information from the clinical notes of electronic health records. For example, Fries is collaborating with Stanford postdoctoral fellow **Allison Callahan, PhD**, to extract mentions of pain and other symptoms from 500,000 clinical notes for 3500 hip and knee replacement patients. When combined with structured data from the electronic health records, unstructured data from clinical notes may help doctors predict which patients' implants will fail, as well as generate early warnings when specific devices are causing problems (see page 21 for sidebar story about Callahan's work). Snorkel is also being used outside of biomedicine—for example, researchers at the Hoover Institution are using Snorkel to extract data from military combat notes to try to determine what factors cause militants to join or leave insurgencies.

Ré's lab is also building tools on top of Snorkel to extract data from images, figures, and tables. Ratner is collaborating with radiologists who study bone tumors to develop a Snorkel-based tool that can accurately classify images of bone lesions as cancerous or not. For images, users write labeling functions that consider visual features, such as edges, as well as text in titles and captions. Snorkel-based tools that read tables in the biomedical literature are also in development. These tools can help augment manual data curation efforts, such as for the GWAS Catalog (an online catalog of published genome-wide association studies). For example, a computer could extract results from every supplemental GWAS table in the published literature.

KnowEnG: From Phrases to Relations

Researchers at the KnowEnG Center are also exploiting weak and distant supervision to make state-of-the-art text-mining tools that require minimal labeling from domain experts. KnowEnG's director, Jiawei Han, has developed a suite of text-mining tools that work on everything from tweets, to the New York Times, to the scientific literature. In the past few years, Han's lab has been focusing on applications in biomedicine.

"Jiawei is a mainstream text-mining person. For him to enter bio-text mining is very exciting," says KnowEnG co-director **Saurabh Sinha, PhD**, professor of computer science at the University of Illinois, Urbana-Champaign. "The underlying tools that his lab has developed are making new functionalities that I care about happen."

Han's team first built tools to mine phrases out of text. Text-mining tools need to recognize that certain words go together—such as "congenital heart disease" or "Obama administration." "Extraction of phrases is critical towards information extraction because many concepts, entities, and relations are manifested in phrases," Han says. The tools are portable across domains and languages, and also

Text-mining tools need to recognize that certain words go together—such as "congenital heart disease" or "Obama administration."

require minimal or no hand labeling. "We worked out a very powerful method using either no training at all or weak training or distant training."

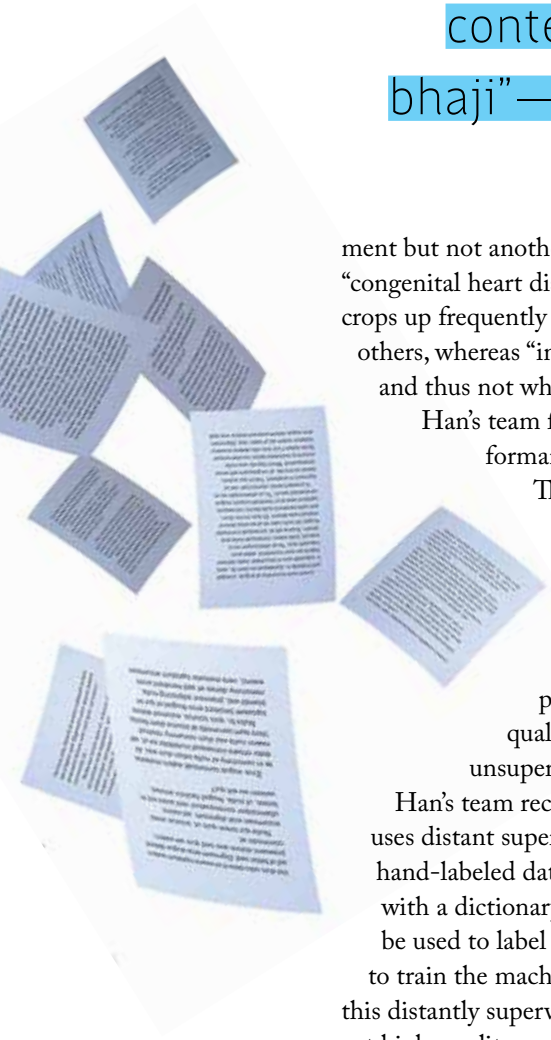
Han's team first built ToPMine in 2014, which is an unsupervised method and requires no training data. ToPMine identifies salient phrases using statistical clues, such as how frequently a given string of words appears in the corpus (popularity), how often the words appear together versus apart (concordance), and how often they appear in one document but not another (distinctiveness). For example, "congenital heart disease" is distinctive because it crops up frequently in some documents but rarely in others, whereas "important problem" is ubiquitous and thus not what Han calls a "quality phrase."

Han's team found they could improve performance by adding weak supervision. Their tool SegPhrase incorporates a machine-learning model that can be trained with a tiny amount of labeled data—just 300 labeled phrases for a 1 gigabyte corpus. The model generates better-quality phrases than completely unsupervised methods such as ToPMine.

Han's team recently built AutoPhrase, which uses distant supervision to obviate the need for hand-labeled data. Users provide AutoPhrase with a dictionary or knowledge base that can be used to label enough phrases in the corpus to train the machine-learning model. "We like this distantly supervised method because you can get high-quality results without experts," Han says. "It's also powerful because it works on many languages. It could also recognize Chinese phrases if we gave it a Chinese Wikipedia, for example."

Han's team has also developed an entity tagger called ClusType, which builds on their phrase-mining tools. ClusType first uses distant supervision to tag some entities in the corpus. Then it leverages the context clues around the labeled entities to tag additional entities. For example, based on Wikipedia, ClusType may tag ice cream as a food in: "The waiter served ice cream." When ClusType later comes across an unlabeled phrase in a similar context—for example, "The waiter served pav bhaji"—it is able to predict that pav bhaji is a food. Newly labeled entities give new context clues, and the whole cycle repeats until the corpus is adequately labeled. When applied to news stories, Yelp reviews, and tweets, ClusType yielded an average 37 percent improvement over the next best method for tagging entities. Han's team has extended this to CoType, which works in a similar manner but types both entities and relationships between entities simultaneously.

ClusType may tag ice cream as a food in: "The waiter served ice cream." When ClusType later comes across an unlabeled phrase in a similar context—for example, "The waiter served pav bhaji"—it is able to predict that pav bhaji is a food.



Han's suite of text-mining tools are publicly available (at https://github.com/KnowEnG) and are being built into the KnowEnG knowledge engine. The KnowEnG Center is also partnering with Heart BD2K to use the tools to solve specific biomedical problems (see sidebar: Ranking Proteins in Heart Disease, opposite).

Han's team is also working on an exciting new search tool that embeds entity recognition into the search. If you search in PubMed or Google Scholar, these search engines treat genes, proteins, metabolites, and drugs like any other words. But what if the search engine could recognize genes, proteins, metabolites, and drugs as biological entities? "That's a type of query/response interface to the literature that supports a much richer space of queries," Sinha says.

Working together with existing biomedical knowledge bases, Han's tools can tag entities in queries and papers. "His tools can recognize that there are different types of terms in there. They have built in the prior knowledge of what are genes, what are proteins, what are drugs, and so on," Sinha explains. Now, Han's team

Ranking Proteins in Heart Disease

The **Heart BD2K**'s director, **Peipei Ping**, **PhD**, asked Jiawei Han's team at KnowEnG to help them use the biomedical literature to comparatively rank 250 proteins known to be involved in heart disease. Ping is professor of physiology, medicine/cardiology, and bioinformatics at the University of California, Los Angeles. "The problem is there are millions of papers in cardiology. Nobody can read one million papers in a lifetime. But a computer can," Han says. "What if your computer could read those articles to give you a comparative summary?" Han's and Ping's labs collaborated to build a pipeline called Context-Aware Semantic Online Analytical Processing (caseOLAP), which incorporates SegPhrase.

Ping's team wanted to know which of the 250 proteins were most relevant for each of the six major types of heart disease—cerebrovascular accidents, cardiomyopathies, ischemic heart diseases, arrhythmias, valve disease, and congenital heart disease. They used phrase mining to group abstracts by disease and to discover the predominant proteins for each disease. CaseOLAP calculated a text-mining score for each disease-protein pair based on the quality of the mined phrases, how frequently a given protein appeared in the abstracts of a given disease, and how distinct a given protein was for one disease versus the other five.

"The thing that I found amazing was just how much information could be processed. This is something that a human being just cannot do," says **David Liem**, **MD**, **PhD**, a scientist at UCLA and the project's clinical study coordinator. When they examined the top-ranking proteins, they got some unexpected insights. "Some of the findings were no surprise. For example, we found a lot of inflammatory proteins and proteins involved in hemodynamic regulation," Liem says. "But what was a surprise to us was we found a lot of proteins that are involved in neurodegenerative diseases." This unforeseen link between heart disease and neurodegenrative disease has been confirmed in other recent studies, Liem notes.

The discoveries could help doctors predict new drug targets for heart disease, Ping says. Han and Ping plan to expand the project to explore the role of 8,000 additional proteins and also to rank protein-protein interactions for each type of heart disease. The tool could also be applied to electronic medical records to mine clinical notes. "This collaboration opens possibilities for many other projects," Ping says.

is working on exactly how to use this information to give the most reasonable ranking of papers. "The problem isn't solved yet, but we have an army of really smart graduate students working on this," Sinha says.

bioCADDIE: Customized Pipelines for Text Mining

The bioCADDIE Center is developing dataMED, a search engine for publicly available datasets that does for data what PubMed does for papers. Similar to Han's search tool, dataMED embeds entity recognition into the search. So, it's not suprising that text-mining expert **Hua Xu**, **PhD**, was tapped to lead development. Xu is a professor in the School of Biomedical Informatics at the University of Texas Health Science Center at Houston.

> "[T]here are millions of papers in cardiology. Nobody can read one million papers in a lifetime. But a computer can," Han says.

Xu's lab works on text-mining methods, software, and applications for clinical data. "I view it like a circle. You have a new proven method for NLP [natural language processing]; you implement that into software; and then you use the software to extract information for clinical studies," Xu says. "Then the clinical study actually suggests needs for new technology, and this feeds back to the methods development."

Like scientists at Mobilize and KnowEnG, Xu wants to reduce demands for costly annotated training data. Rather than turning to weak and distant supervision, however, Xu's lab is taking a different tack—an approach called interactive machine learning.

Interactive machine learning loops humans into the learning process. The idea is that if a person injects critical insights at the right time, this can improve efficiency and performance. Normally, human experts label random stretches of training data. But in active learning—a type of interactive machine learning—only the most informative examples are selected for labeling, Xu explains. At the beginning of active learning, a human expert annotates a small amount of randomly selected training data, which is fed into a machine-learning algorithm to build a model. The computer then attempts to classify the unlabeled data using the model—and passes back the ones for which it has the most trouble. The human labels these, and passes them back to the computer. This cycle repeats until the model achieves sufficient accuracy. This approach has the potential to significantly reduce training data for the same performance, Xu's team has shown.

Xu's lab has also built several machine learning–based tools for entity tagging and relation extraction that have taken first or second place in major text-mining challenges including the i2b2 NLP Challenge, SemEval (Semantic Evaluation) and BioCreAtIve (Critical Assessment of Information Extraction in Biology). Xu's team has made these tools available as part of dataMed.

Sometimes, text-mining tools need to be tuned to local data. For example, different hospitals and even different specialties within the same hospital may have idiosyncrasies in how they talk about diseases and traits. "For an institution that doesn't have a strong NLP team, it could be very challenging," Xu says. To address this issue, Xu's team built a user-friendly clinical text-mining system called CLAMP (Clinical Language Annotation, Modeling, and Processing Toolkit, http://clamp.uth.

edu/). With CLAMP, users drag and drop ready-made components—such as part-of-speech taggers, dictionaries, and machine-learning modules—to create a customized pipeline. "They can modify each component, including directly annotating local data and clicking a button to train a machine-learning module," Xu says. "With this interface, users who don't have much NLP experience can build high-performance NLP pipelines for their own tasks," Xu says. The tool is freely available to academics, and has been downloaded by about 50 institutions.

For example, a researcher can enter "HIV replication" in gene-search mode and get a ranked list of genes relevant to HIV replication. The system searches millions of abstracts in PubMed to find those relevant to the user's query, and then applies a named entity recognizer to identify mentions of genes or small molecules. "The challenge is trying to filter out false positives," says CPCP director Mark Craven. For example, the common English word "cat" could be mistaken for the abbreviation for the catalase gene, CAT. To help avoid errors, the named entity recognizer considers properties of the word, such as the presence of italics, as well as the context around it. "We look at lots of pieces of evidence like that to decide 'do I think this is a gene name, yes or no?'" Craven says. The software then ranks the genes based on how many query-specific abstracts mention the gene and how frequently the gene is mentioned in other abstracts.

With CLAMP, Xu says, "users who don't have much NLP experience can build high-performance NLP pipelines for their own tasks."

The same cutting-edge text-mining tools that are baked into CLAMP are also being used in dataMED, bioCADDIE's dataset search tool (https://datamed.org). The dataMED development team first indexed all the datasets in the data repositories, using NLP to mine free-text fields for mentions of genes, diseases, and chemical names. Then the team built a search tool that embeds recognized entities into the search. For example, if a user types in breast cancer and NFKappaB, the tool recognizes these as a disease and gene, respectively, and maps them to their standardized ontology concepts. Then it expands the search to include synonyms of these concepts (e.g., "tumor of breast" for "breast cancer").

dataMED has already indexed more than 63 data repositories that contain a total of 1.4 million biomedical datasets.

CPCP: Putting Text-Mining Tools to New Uses

Though text-mining methods are not a focus of CPCP, Center researchers have developed some novel ways to use text mining for different kinds of PubMed searches and to clean up metadata in data repositories. GADGET, developed by Craven, uses standard indexing and text-mining algorithms to search PubMed and return, for a given query, genes and metabolites rather than articles. Biologists at the University of Wisconsin-Madison, are already using the tool to accelerate their science. For example, one stem cell lab searches for genes that might help them steer stem cells to a given fate. Another lab is using the tool to help figure out the networks of host genes involved in HIV replication. "We found that we can get better network models by pulling in this evidence from the literature as identified by GADGET in addition to the genes that are coming directly from experiments," Craven says. GADGET is freely available here: http://gadget.biostat.wisc.edu/.

Another lead investigator of CPCP, **Colin Dewey, PhD**, associate professor of biostatistics and medical informatics at the University of Wisconsin-Madison, is using text mining to clean up the metadata of the Sequence Read Archive (SRA) data repository.

The SRA stores next-generation sequencing reads for 2.1 million samples from 90,000 worldwide studies. Scientists hope to mine these data for new insights; for example, by studying all available lung cancer RNA-seq data, scientists might be able to pinpoint gene expression patterns that characterize the disease. But combining data across different studies is difficult because the samples aren't labeled consistently. "The metadata are not standardized or normalized," Dewey says. "People just make up their own names for the attributes, as well as the

values of those attributes." Attributes (such as "cell line") and their values (such as "HeLa cells") vary widely due to misspellings, synonyms, abbreviations, and the use of natural language descriptions.

So, Dewey's team devised a novel computational pipeline (https://github.com/deweylab/metasra-pipeline) that automatically cleans up the metadata in the SRA.

Off-the-shelf text-mining tools yield an unacceptably high false-positive rate when applied to SRA metadata. "There are a lot of cases when there's an entity mentioned in the metadata that is not actually describing the sample of interest," Dewey says. For example, a standard named-entity recognizer will extract the word "breast" from "breast cancer" and infer that this is the anatomical source of the sample. But the sample may have been taken from blood rather than breast tissue.

Dewey's team built a system that's similar to a named entity recognizer but "with a bunch of heuristics added to remove the errors introduced by such systems," he says. The system builds a graph, starting with the attribute-value pair from the original metadata. The attribute and value are each mapped to terms from biomedical ontologies (represented as nodes on the graph). But Dewey's team then subjects the graph to a series of custom-made reasoning rules and operations. For example, one of these heuristics recognizes that "breast" should not be mapped to an anatomical location if the word "breast" in the metadata is part of a larger phrase (e.g., "breast cancer") that maps to an ontology term. Another rule tells the system that the abbreviations "F" and "f" indicate a female sample when they are paired with an attribute that maps to "sex." The system also extracts numerical values—such as age—from metadata. "That's a novel aspect of our system that will be helpful for doing aggregate analyses using those numerical values as covariates."

In initial tests on human samples assayed by RNA-seq experiments on the Illumina platform, the system achieved recall rates as good as standard named-entity recognizers (85 to 90 percent) but better false positive rates (precision of 90 to 95 percent). The MetaSRA database is available at http://deweylab.biostat.wisc.edu/metasra. Dewey's team plans to expand the database in the future.

Looking Back; Moving Forward

Nine years ago, *Biomedical Computation Review* (Summer 2008) published an article about text mining. One challenge identified then remains a major bottleneck today—data accessibility. Out of 14 million English-language abstracts in PubMed, only about a million are accessible for full-text mining, Mallory says; and when it comes to electronic health records, privacy issues complicate data access. For text mining to realize its full potential, researchers will have to make headway on this issue.

But there have been a lot of wins over the past nine years, thanks in part to work by the BD2K Centers. Text-mining tools have gotten more powerful and, importantly, more usable by doctors and biologists—as evidenced by user-friendly programs such as Snorkel, CLAMP, and GADGET. Text-mining tools are also being used in more real-world applications than ever before—from curating and scanning the literature to making it easier to find and pool publicly available datasets. It might also be possible to build a pipeline from the BD2K tools described here. For example, Fries says, Han's unsupervised learning tools are great for potentially discovering new patterns and building domain dictionaries, but they are also very noisy. Snorkel could be used to unify these dictionaries into a more robust extraction system. "The different BD2K tools being developed provide complementary ways to tackle the text mining problem," he says.

"The different BD2K tools being developed provide complementary ways to tackle the text mining problem," [Fries] says.

Nine years from now, perhaps computer curators will have replaced human curators for the OMIM database. As a result, within moments of a new paper hitting PubMed, new knowledge will be deposited in OMIM automatically—making it possible for doctors to instantly use that knowledge to help patients. □