These Elsevier collaborations use machine learning to turn data into knowledge

Computer scientists are finding new ways to extract information from the literature, making data more accessible in bioinformatics, chemistry and clinical decision support

By Lilah Sturges December 14, 2017

Elsevier Connect

Scientists are finding new ways to extract knowledge from data. (Image © istock.com/jxfzsy)

If there's one thing scientists don't lack these days, it's information. Making sense of that

information is currently stored in formats that are designed to be read by humans, such as research articles, but cannot be easily leveraged computationally for other purposes.

Until recently, for scientific data to become readily accessible by other means, a human had to interpret documents and enter relevant data into a database by hand. And while humans do an excellent job of comprehending and synthesizing information, human effort doesn't scale up the way computing power does. With the number and density of data sources ever on the rise, Elsevier and its collaborators hope to change the paradigm to enable computers to manage as much of the job of synthesizing data as possible, allowing humans to do better science with the results.

AI for clinical decision support

Nowhere is the need for scientific data to be accessible more readily apparent than in a doctor's office. Dr. Richard Boyce , Associate Professor in the Department of Biomedical Informatics at the University of Pittsburgh , and his team are working with Elsevier to provide physicians with improved clinical decision support for prescribing medications.

Doctors are regularly faced with complex decisions in prescribing drugs and in weighing the risks and benefits of multiple interacting factors. Circumstances such as drug-drug interactions and pharmacogenetics (genetic factors that influence a patient's outcome from a medication) are often overlooked because clinicians don't have the information to guide them.

Prof: Richard Boyce, PhD

"This an area where evidence is coming from a variety of different sources," said Dr. Boyce, "and those sources tend to be very siloed and heterogeneous."

Those whose job it is to synthesize the evidence into information that is usable by clinicians struggle with the complexity of the problem. "They struggle to find the information," he added. "They struggle to assess it when they can finally locate it, and they often are not able to organize the synthesized information in a format that's useful for clinical application."

Part of the problem is that there are numerous sources of information about drug-drug interaction, such as scientific articles, product labeling, and new drug applications. "So in those various sources," Dr. Boyce said, "we're imagining that at the time that they publish these artifacts: what if they're able to explain what the main scientific claim is, and able to explicitly annotate the data that supports that claim and the method that supports the data?"

As for next steps:

> *"We've been building a framework so that for a small set of drugs, we will have explicitly annotated the claims, the data and the methods for scientific articles and product labeling, and so on. (These) are important to look at when you're thinking about drug interactions. …"*
> *"Then we're building a search portal which is going to be used in a study. We're going to engage people who work at companies where they synthesize this information and say, 'Hey, why don't you use your normal approach and also use our tool?' And we're going to look at questions like: Are they going to be able to do their job more quickly, more effectively, more accurately? Are they satisfied with the end result? We hope it will be a big improvement from the user's perspective, and it should make (the process) more effective. And on the flip side, that should be a great gain for search and retrieval."*

One of the primary hurdles Dr. Boyce faced was getting access from publishers to use journal articles in his research. Though drug labeling information can be downloaded for free, using full-text articles creates licensing issues. "We had to contact every single publisher and get their permission," he said, "and Elsevier was very generous about giving us blanket permission for the articles that we listed as needing to annotate. They're also okay with us publishing the annotations publicly, which is not true across all publishers."

Going forward, Dr. Boyce would like to include machine reading in his work. "We'd like to work on a task," he said, "where we try to supplement manual annotation with machine-extracted information, but we're not quite there yet. Unless the machine-read information is extremely accurate, sometimes the machines will actually confuse the person and they'll spend a lot more time trying to correct the machine than doing their job."

Using machine learning to predict the cellular function of proteins

The key to moving forward appears to lie in machine learning. "It should be said that machine learning has played a large part in bioinformatics for as long as that word 'bioinformatics' has been in common usage," said Dr. David Jones , Professor of Bioinformatics and head of the Bioinformatics Group in the Department of Computer Science at University College, London . "Of course, these methods have become hugely more powerful in recent years, particularly in the area of neural networks, where we can now apply 'deep learning' algorithms to biological data, but still the ideas have been around for a long time."

David Jones, PhD, Professor of Bioinformatics and head of the Bioinformatics Group in the Department

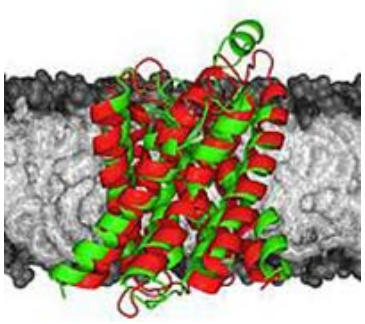
Dr. Jones's team uses machine learning to predict the cellular function of proteins based on information available in academic literature: He explained:

of Computer Science at University College, London.

> *"With so much gene sequence data, we can now infer subtle patterns of co-evolution between pairs of amino acids; that is, we can determine how a mutation in one position of a protein influences the likelihood of observing a mutation somewhere else. With advanced deep learning, this information can not only tell us a lot about the structure of the protein but also which other proteins it might interact with."*

Often, biological experiments such as gene expression profiles can uncover previously unknown proteins that might be interesting targets for research, but these targets tend not to be well-studied and so give little insight to researchers. "A key role in bioinformatics," said Dr. Jones, "is to help generate novel hypotheses for scientists to test in the lab. A scientist can use our software to generate a list of possible functions that we think a protein is likely to have, and can then go to the lab to validation those predictions."

"If we can develop better to tools to make so-called 'de novo' functional predictions for proteins like that," he said, "such as information on cellular localization or possible molecular interaction partners, this could have a lot of impact for drug discovery or modeling of disease pathways."



Transmembrane protein modeling from the Jones lab at University College London

One promising avenue in Dr. Jones's work is determining how much extra information on protein function his team can get from applying their algorithms not just to data in the public domain but also to data accessible through Elsevier's Pathway Studio, with its vast database of published biological literature. "Although that's still very much work in progress," he said, "we are nonetheless seeing encouraging results."

Although the technology has improved, said Dr. Jones:

> *"None of this would make any difference if we also didn't have orders of magnitude more data. In many respects, it has been the huge growth in the size of data sets that we have to analyze that has had the most impact. For example, in protein folding, we now commonly have thousands or even tens of thousands of protein sequences which are evolutionarily related to a protein we are trying to analyze."*

Extracting information about chemistry reactions from research

articles

"Everything we know about science is in a research publication," said Dr. Karin Verspoor . A Professor in the School of Computing and Information Systems at the University of Melbourne , Dr. Verspoor is building machine learning tools to extract information about chemical reactions from chemistry research articles:

Prof. Karin Verspoor, PhD

> *"It gets studied in a lab and then it gets written up as a scientific publication. And sometimes that information gets into a database, but not always. If we can get a machine to process the info and make it easer to find and structure so that other kinds of analyses can be done on that information, then that would be a huge resource."*

"We decided to focus our project around the context of reactions," she said, "and essentially try to work on the algorithms for extracting chemical interactions information and chemical reaction information from the literature."

She explained:

> *"There's a lot of information that is specifically chemical in nature in the chemistry literature that is about reactions, and a lot of that information is expressed in what I would call 'pseudo-natural language.' It's using English to describe a particular chemical process, but it's highly domain-specific and has a particular pattern to it. We're starting with information that's focused on reactions and also information that's expressed in tables. There's a lot of scientific content that's expressed in tables in a research publications, and from an NLP (natural language processing) perspective, that's particularly challenging information because tables don't have the linguistic clues around them to help decipher the language and semantics of the table."*

The collaboration between her lab and Elsevier is a natural one, Dr. Verspoor said: "My focus has been on information extraction from scientific literature. Elsevier publishes literature and (adds) value to that literature. One of the main things we're going to be doing is actually to leverage the internal expertise that already exists at Elsevier doing manual analysis of the literature to give us those examples that we need to train the machine learning model.

"There were teams at Elsevier already working on extraction from these publications and

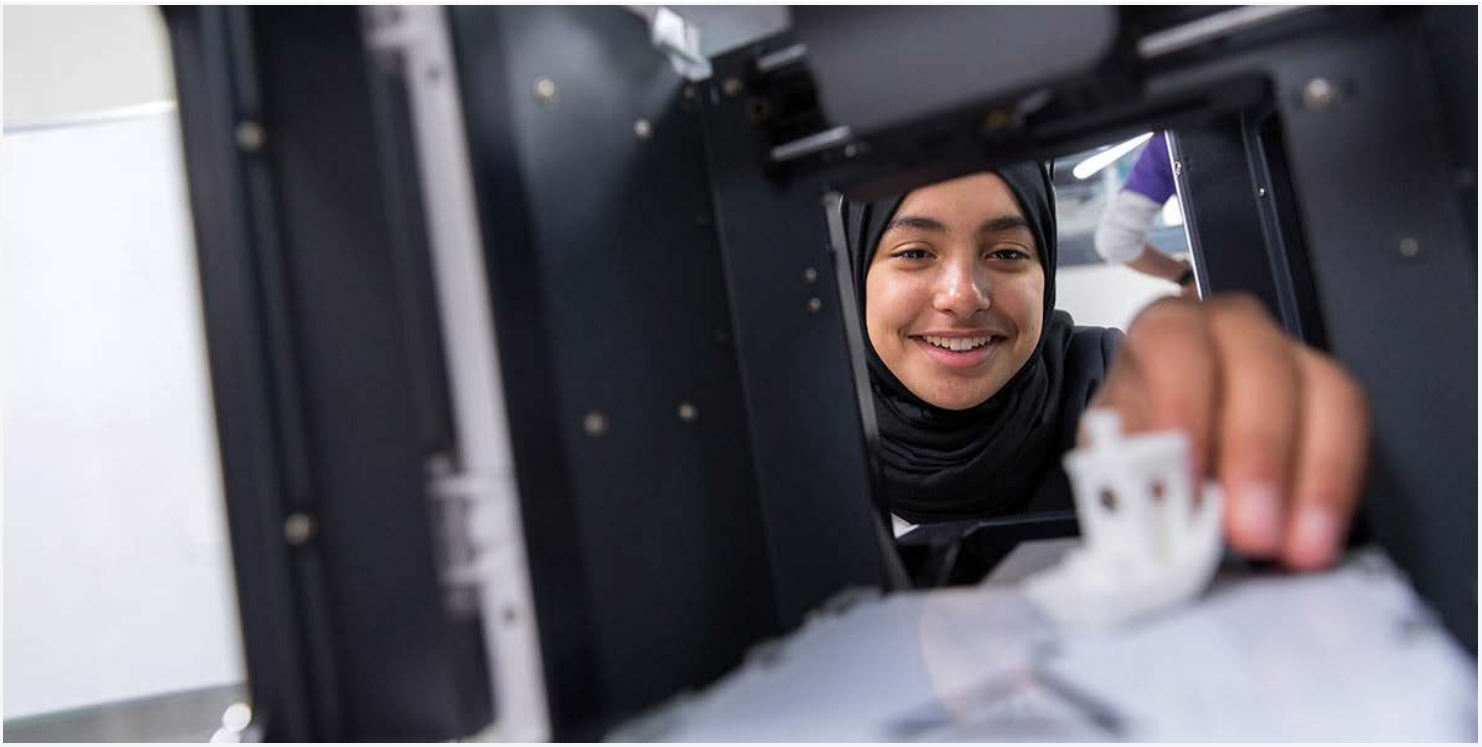
making them available in tools like Reaxys, and most of that process has been happening through a largely manual effort," she said. "Obviously that is a process that is extremely human-intensive and doesn't scale very well when we have a million new research publications coming out every year. It's just an incredible amount of information."

Elsevier, she explained, is also contributing to the broader research community with information extraction from the literature and text mining. For example, her group and Elsevier plan to make the annotated resources available more broadly to the scientific community. "So it won't be just me building methods for my group for this set of problems, but other people around the world who are interested in working on these tasks can contribute as well.

"It won't just help Elsevier," she said, "It'll actually help move the entire discipline forward."

Latest posts

How social science is driving a sustainable future (with special issue)



Makerspace creates tech opportunities for young people

Harvard and Elsevier explore collaborations in data science

In photos: Celebrating LGBTQ Pride at Elsevier — with new chapter in NYC

How data scientists are tackling hunger and social change

Tags

Data & Analytics | Technology | Trends | Innovation

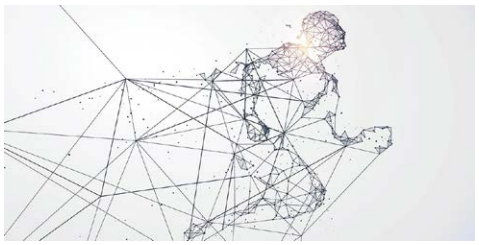
Contributors

Lilah Sturges

Lilah Sturges has written articles on topics ranging from Hamlet to the transgender experience. She has also authored two novels, as well as numerous short stories and comic books for which she has been nominated for both the Eisner and Ignatz awards. Prior to her career as a writer, she worked as a web developer and software architect. She has a passion for science and once considered becoming a physicist, though she ultimately graduated from the University of Texas at Austin with a degree in English.

Follow me on Facebook ↗ or LinkedIn ↗
.

Related stories



5 reasons data is a key ingredient for AI applications

By Paul Groth, PhD

How researchers are using NLP and machine learning to ease your information overload

Why today's tech jobs need creative minds

By Alison Bert, DMA

August 30, 2017

December 05, 2017

Comments↗

Elsevier Connect

By Lilah Sturges

October 16, 2017

Comments↗

Elsevier Connect

Comments↗

Elsevier Connect

Comments

Solutions

Researchers

About Elsevier

How can we help?