TOOLBOX

AI TAMES THE SCIENTIFIC LITERATURE

As artificially intelligent tools for literature and data exploration evolve, developers seek to automate how hypotheses are generated and validated.

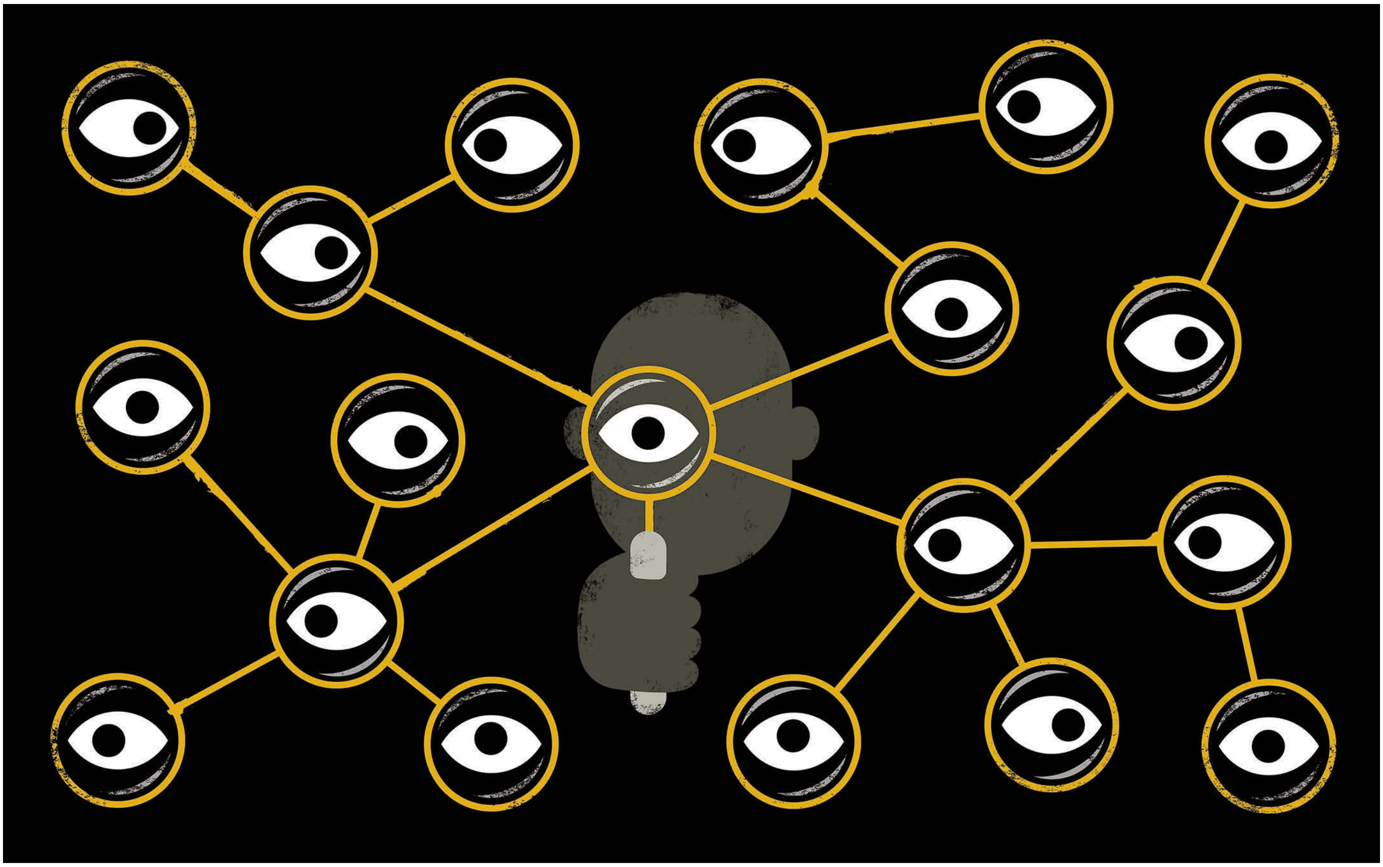

ILLUSTRATION BY THE PROJECT TWINS

BY ANDY EXTANCE

When computer scientist Christian Berger's team sought to get its project about self-driving vehicle algorithms on the road, it faced a daunting obstacle. The scientists, at the University of Gothenburg in Sweden, found an overwhelming number of papers on the topic — more than 10,000 — in a systematic literature review. Investigating them properly would have taken a year, Berger says.

Luckily, they had help: a literature-exploration tool powered by artificial intelligence (AI), called Iris.ai. Using a 300-to-500-word description of a researcher's problem, or the URL of an existing paper, the Berlin-based service returns a map of thousands of matching documents, visually grouped by topic. The results, Berger says, provide "a quick and nevertheless precise overview of what should be relevant to a certain research question".

Iris.ai is among a bevy of new AI-based search tools offering targeted navigation of the knowledge landscape. Such tools include the popular Semantic Scholar, developed by the Allen Institute for Artificial Intelligence in Seattle, Washington, and Microsoft Academic. Although each tool serves a specific niche, they all provide scientists with a different look at the scientific literature than do conventional tools such as PubMed and Google Scholar. Many are helping researchers to validate existing scientific hypotheses. And some, by revealing hidden connections between findings, can even suggest new hypotheses for guiding experiments.

Such tools provide "state-of-the-art information retrieval", says Giovanni Colavizza, a research data scientist at the Alan Turing Institute in London, who studies full-text analysis of scholarly publications. Whereas conventional tools act largely as citation indices, AI-based ones can offer a more penetrating view of the literature, Colavizza says.

That said, these tools are often expensive, and limited by the fraction of the scientific literature they search. "They are not meant to give you an exhaustive search," says Suzanne Fricke, an animal-health librarian at Washington State University in Pullman, who has written a resource review on Semantic Scholar (S. Fricke *J. Med. Lib. Assoc.* **106**, 145–147; ▶

▸ 2018). Some, for example, "are meant to get you quickly caught up on a topic, which is why they should be used in conjunction with other tools". Berger echoes this sentiment: "Blindly using any research engine doesn't answer every question automatically."

TEACHING SCIENCE TO MACHINES

AI-based 'speed-readers' are useful because the scientific literature is so vast. By one estimate, new papers are published worldwide at a rate of 1 million each year — that's one every 30 seconds. It is practically impossible for researchers to keep up, even in their own narrow disciplines. So, some seek to computationally tame the flood.

The algorithms powering such tools typically perform two functions — they extract scientific content and provide advanced services, such as filtering, ranking and grouping search results. Algorithms extracting scientific content often exploit natural language processing (NLP) techniques, which seek to interpret language as humans use it, Colavizza explains. Developers can use supervised machine learning, for example — which involves 'tagging' entities, such as a paper's authors and references, in training sets to teach algorithms to identify and extract them.

To provide more-advanced services, algorithms often construct 'knowledge graphs' that detail relationships between the extracted entities and show them to users. For example, the AI could suggest that a drug and a protein are related if they're mentioned in the same sentence. "The knowledge graph encodes this as an explicit relationship in a database, and not just in a sentence on a document, essentially making it machine readable," Colavizza says.

Iris.ai takes a different approach, Colavizza notes, grouping documents into topics defined by the words they use. Iris.ai trawls the Connecting Repositories collection, a searchable database of more than 134 million open-access papers, as well as journals to which the user's library provides access. The tool blends three algorithms to create 'document fingerprints' that reflect word-usage frequencies, which are then used to rank papers according to relevance, says Iris.ai chief technology officer Viktor Botev.

The result is a map of related papers, but eventually the company plans to supplement those results by identifying hypotheses explored in each paper as well. It is also developing a parallel, blockchain-based effort called Project Aiur, which seeks to use AI to check every aspect of a research paper against other scientific documents, thus validating hypotheses.

Colavizza says that tools such as Iris.ai — free for basic queries, but costing upwards of €20,000 (US$23,000) a year for premium access, which allows more-nuanced searches — can accelerate researchers' entry into new fields. "It facilitates initial exploration of the literature in a domain in which I'm marginally familiar," he says.

Experts seeking deeper insights into their own specialities might consider free AI-powered tools such as Microsoft Academic or Semantic Scholar, Colavizza suggests. Another similar option is Dimensions, a tool whose basic use is free but which costs to search and analyse grant and patent data, as well as to access data using the programmable Dimensions Search Language. (Dimensions is created by technology firm Digital Science, operated by the Holtzbrinck Publishing Group, which also has a majority share in *Nature*'s publisher.)

Semantic Scholar has a browser-based search bar that closely mimics engines such as Google. But it gives more information than Google Scholar to help experts to prioritize results, Colavizza says. That includes popularity metrics, topics such as data sets and methods, and the exact excerpt in which text is cited. "I was very surprised to find that they also capture indirect citations," Colavizza adds — such as when a method or idea is so well established that researchers don't refer to its origin.

Doug Raymond, Semantic Scholar's general manager, says that one million people use the service each month. Semantic Scholar uses NLP to extract information while simultaneously building connections to determine whether information is relevant and reputable, Raymond says. It can identify non-obvious connections, such as methodologies in computer science that are relevant to computational biology, he adds, and it can help to identify unsolved problems or important hypotheses to validate or disprove. Currently, Semantic Scholar incorporates more than 40 million documents from computer and biomedical science, and its corpus is growing, says Raymond. "Ultimately, we'd like to incorporate all academic knowledge."

For other tools, such as SourceData from the European Molecular Biology Organization (EMBO) in Heidelberg, Germany, experimental data are a more central concern. As chief editor of *Molecular Systems Biology*, an EMBO publication, Thomas Lemberger wants to make the data underlying figures easier to find and interrogate. SourceData therefore delves into figures and their captions to list biological objects involved in an experiment, such as small molecules, genes or organisms. It then allows researchers to query those relationships, identifying papers that address the question. For instance, searching, 'Does insulin affect glucose?' retrieves ten papers in which the "influence of insulin (molecule) on glucose (molecule) is measured".

> *"Ultimately, we'd like to incorporate all academic knowledge."*

SourceData is at an early stage, Lemberger says, having generated a knowledge graph comprising 20,000 experiments that were manually curated during the editing process for roughly 1,000 articles. The online tool is currently limited to querying this data set, but Lemberger and his colleagues are training machine-learning algorithms on it. The SourceData team is also working on a modified neuroscience-focused version of the tool with an interdisciplinary neuroscience consortium led by neurobiologist Matthew Larkum at Humboldt University in Berlin. Elsewhere, IBM Watson Health in Cambridge, Massachusetts, announced in August that it will combine its AI with genomics data from Springer Nature to help oncologists to define treatments. (*Nature*'s news team is editorially independent of its publisher.)

HYPOTHETICALLY USEFUL

Among those embarking on hypothesis generation are the roughly 20 customers of Euretos, based in Utrecht, the Netherlands. Arie Baak, who co-founded Euretos, explains that the company sells tools to industry and academia, mainly for biomarker and drug-target discovery and validation, for prices he did not disclose.

Euretos uses NLP to interpret research papers, but this is secondary to the 200-plus biomedical-data repositories it integrates. To understand them, the tool relies on the many 'ontologies' — that is, structured keyword lists — that life scientists have created to define and connect concepts in their subject areas.

Baak demonstrates by searching for a signalling protein called CXCL13. Above the resulting publication list are categories such as 'metabolites' or 'diseases'. The screen looks much like Google Scholar or PubMed at this stage, with an ordered list of results. But clicking on a category reveals extra dimensions. Selecting 'genes', for instance, pulls up a list of the genes associated with CXCL13, ranked by how many publications mention them; another click brings up diagrams illustrating connections between CXCL13 and other genes.

Researchers at the Leiden University Medical Centre (LUMC) in the Netherlands have shown that this approach can yield new hypotheses, identifying candidate diseases that existing drugs might treat. The team presented its results at the Semantic Web Applications and Tools for Health Care and Life Sciences meeting in Rome in December 2017. They have also used Euretos to identify gene-expression changes in a neurological disorder called spinocerebellar ataxia type 3 (L. Toonen *et al. Mol. Neurodegener.* **13**, 31; 2018).

So, should researchers worry that AI-based hypothesis generation could put them out of a job? Not according to Colavizza. Hypothesis generation is a "very challenging ambition", he says, and improvements initially will be incremental. The hypotheses suggested so far are therefore "mostly in the realm of the relatively unsurprising ones", Colavizza says.

That will probably change, of course. But surprising or not, computer-generated hypotheses must still be tested. And that requires human researchers. "One should never believe an auto-generated hypothesis first-hand without investigating the underlying evidence," warns LUMC researcher Kristina Hettne. "Even though these tools can assist in collecting the known evidence, experimental validation is a must." ■

Andy Extance *is a freelance writer based in Exeter, UK.*