11-2019

What do you mean? Research in the Age of Machines

Arthur J. Boston
Murray State University, aboston@murraystate.edu

Recommended Citation

Arthur "A.J." Boston

What do you mean?

Research in the Age of Machines

"What Do You Mean?" was an undeniable bop of its era in which Justin Bieber explores the ambiguities of romantic communication. (I pinky promise this will soon make sense for scholarly communication librarians interested in artificial intelligence [AI].) When the single hit airwaves in 2015, there was a meta-debate over what Bieber meant to add to public discourse with lyrics like "What do you mean? Oh, oh, when you nod your head yes, but you wanna say no."[1] It is unlikely Bieber had consent culture in mind,[2] but the failure of his songwriting team to take into account that some audiences might interpret it that way was ironic, considering the song is all about interpreting signals.

Like pop music, innovation often inspires unforeseen takes. Consider the Internet, an infrastructure built for a faster means of communication. Or Spandex, a fabric developed for freer movement of the body. For one generation, the Internet and Spandex were the fruits of a war effort. For another generation, they mean Instagramming in athleisure.[3] Imagine some early ARPANET boss rallying his staff around that as a goal—you can't even.

Recently, University of California-San Francisco researchers trained a machine-learning algorithm to decode words and phrases from speech signals in the brain, which could lead to neuroprosthetics capable of restoring speech systems for people who have lost communication abilities.[4]

For Facebook, a major investor in this research, their interpretation of this tech-nology is a future brain-computer interface that would allow users to navigate between screens and type up posts, free of effort from hands or voice. Such an interface would minimize the frictions necessary for consumers to feed their data into Facebook's highly profitable algorithms. What do users mean? Facebook wants to know.

The Facebook tech blog wrote that technology is not "inevitable, and it is never neutral—it's always situated within a specific social and historical context."[5] One context worth remembering is the social media company's history with data handling, such as when Cambridge Analytica received data on 87 million Facebook users that could then be rendered through more than 100 data models to "target" and "predict the behavior of like-minded people."[6] (And to be fair to Cambridge Analytica, that's basically the Facebook business model.)

Data mining and machine-learning are a great boon for political campaigns and corporate marketing wings that thrive on the ability to uncover hidden connections in consumer behavior, in order to influence it. Such practices are problematic, but they are no less effective for that fact. In the classic fashion of late-capitalism, efforts that could do good for humankind using these ad-

Arthur "A.J." Boston is scholarly communication librarian at Murray State University Libraries, email: aboston@murraystate.edu

vances are often stymied if they run counter to an overall profit maximization narrative. Research articles, for instance, are routinely placed behind paywalls, consequently leaving underfunded scholars, the public at large, and even machines, unable to build meaning or create new connections between knowledge resources.

Wait—machines?

What does research mean, according to a machine?

Carl Malamud, a longtime crusader for open information, recently "teamed up with Indian researchers to build a gigantic store of text and images" equivalent in size to the Web of Science core collection. The goal for this electronic database is not for researchers to find and read individual articles, but for computer software to crawl the "world's scientific literature to pull out insights without actually reading the text."[7] At present, whether the vision Malamud proposes will ultimately jibe with copyright is an open question.

While it is unclear to what extent publishers will bully progress under the banner of copyright, the potential for knowledge advances made possible when machines access the scholarly corpus are being realized in other areas. Machine learning-generated word maps have become "established tools" for data scientists to uncover semantic relationship between huge swaths of literature.[8] Paper Digest and Scholarcy hope to assist overwhelmed readers with article summaries and key takeaways. Google Scholar, Semantic Scholar, and Meta (a Chan Zuckerburg joint) are each machine-learning programs built to aid article discovery for readers. And editors have at their disposal "quantitative tools that complement the[ir] qualitative expertise" to help "estimate the future impact" of manuscripts under review.[9]

Literature citation sentiment is also a fascinating area of growth for machine-learning advancement. Take CiTO, which is a Citation Typing Ontology that gives scholars a vocabulary to "capture their cita-

tion intent" whenever they cite a study.[10] This idea was recently built upon with the "Annotation Platform for Citation Typing at Scale," which enables authors to *rapidly* classify their in-text citations "according to purpose and influence."[11] Just earlier this year, Scite.ai unveiled a machine-learning tool that *automatically* detects whether an article's citing papers were written in support or contradiction of the cited article claims. If we take these developments together—the existence of citation ontologies and platforms for authors to encode them—we can begin to consider how a machine-learning tool (like Scite.ai) might evolve if fed rich, human-generated citation sentiment data. The implications are startling.

What does it all mean, for libraries?

If (or when) citation counts become nuanced reflections of sentiment from citing papers, we have to consider what might be downstream effects on literature discovery, library purchase and subscription decisions, research funding decisions, journal editorial decisions and subsequent author writing choices, teaching, and so on. There are any number of potential effects, but the first hypothetical for librarians to decide is whether we will be active partners in shaping the outcome or not. If we're in, there's work to be done, both in technical and critical terms.

When MIT Libraries Director Chris Bourg gave a talk, saying it was past time that digital libraries were taken to the next level with AI and machine-learning, she urged that our use of these tools support our missions and values.[12] As Thomas Padilla writes, there are values-based implications to consider as our born-digital collections come to be "treated as data rather than simple surrogates of physical objects."[13] Research labs might build an automated thinking solution today, and we might begin to use it tomorrow, but without understanding possible complications, we accrue what Jonathan Zittrain calls: "intellectual debt." We can pay off these debts by establishing a clearer understand-

ing over time. For progress to occur, a dab of intellectual debt might be necessary here and there. When we continually fail to pay these debts off, interests accrue.

Most "machine-learning models cannot offer reasons for their ongoing judgements," says Zittrain, and misfires can be "triggered intentionally by someone who knows just what kind of data to feed into that system,"[14] or even triggered unintentionally by someone who does not realize that a data set was suboptimal to begin with. Either way, garbage in, garbage out, as the adage goes. The failure of humans to recognize what constitutes garbage, or "bad" data, can "unintentionally reify human behavior," writes Charlie Harper in a paper introducing librarians to issues that "raise deep questions about the future role of [machine-learning] in society."[15]

"Garbage in, garbage out" is among these issues, such as when a facial recognition program poorly recognizes darker-skinned women relative to its recognition of lighter-skinned men as a result of biased or incomplete training data. Other examples Harper discusses are the privacy issues when AI uncovers otherwise hidden personal traits, or the challenges deepfakes pose toward our sense of reality.

What will the librarians mean to communicate?

As a scholarly communication librarian, the areas of machine-learning enhancement I've been closely following are those that aid in the publishing and research cycle, such as Scite.ai and Scholarcy. While I am eager to share this new class of tools with the students and faculty members on my campus, I'm also thinking about the attendant intellectual debt.

To illustrate, consider SCIgen, an algorithm that generates spoof computer science articles full of random nonsense. It was a lesson well-learned for the editors who were later informed that they had accepted some of these spoofs into their conference proceedings. Knowing that SCIgen has already

been used in this mostly prankish way, it is a fair assumption that at some point, more malevolently intentioned entities will use something like SCIgen to generate false or misleading information, but otherwise logically written articles, perhaps in support of medicines still under trial or in contradiction of particular sciences prone to political ire, like climate change. Flood enough journal submission portals with these, and some number of spoofs will invariably get published.

And so, when I discuss the benefits of an AI-powered research tool with a local researcher, it should be my response to also discuss hypothetical threats. Threats like discovering papers, once plugged into Scite.ai, appear to be overwhelmingly supported or contradicted by the citing literature. Perhaps there is scientific consensus, or maybe it's the case that the literature has been flooded with intentional spoofs.

Likewise, if I introduce journal editors to AI-enabled editorial tools, it will be incumbent on me to warn of the chance that past (and present) publication biases could possibly creep into the underpinning algorithms. Some manuscript types, like null result studies, currently don't have a probable chance to help build impact for a journal. If a tool that an editor has invested in recommends not publishing such studies, the editor might feel pressure to follow that guidance, which would be a net negative for the state of science. These are just two hypothetical threats that I can imagine, to say nothing of those that I cannot.

"Answers without theory, found and deployed in different areas," Zittrain wrote, "can complicate one another in unpredictable ways."[16] And this is really the point: for librarians to have a theory to accompany these new solutions before putting them into practice, to have our values firmly in mind before we incorporate new technology into libraries and the research process, and to critically face the obvious and unforeseen complications to come.

As librarians introduce these shiny new things on our campuses, it is imperative to strive toward developing value-laden theories about

them beforehand, to know what it is that we mean to communicate.

As a famous social media company once blogged: "Technology is never neutral."[17] And neither should be the sentiment with which we discuss it.

Notes

1. Justin Bieber, lyrics to "What Do You Mean?" *Genius,* 2015, https://genius.com /Justin-bieber-what-do-you-mean-lyrics (accessed August 20, 2019).

2. Elizabeth Denton, "Here's Why Justin Bieber's "What Do You Mean" Lyrics Are Sparking Debate About Consent," Seventeen.com, https://www.seventeen.com /celebrity/music/news/a33634/heres-why-some-people-think-justin-biebers-what-do-you-mean-promotes-date-rape/ (accessed August 20, 2019).

3. Jia Tolentino, "Athleisure, barre and kale: the tyranny of the ideal woman," *The Guardian,* https://www.theguardian.com /news/2019/aug/02/athleisure-barre-kale -tyranny-ideal-woman-labour (accessed August 20, 2019).

4. David A. Moses, Matthew K. Leonard, Joseph G. Makin, and Edward F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature* 10, 3096 (July 2019): 1–14, https://doi.org/10.1038/s41467-019-10994-4 (accessed August 20, 2019).

5. Tech@facebook, "Imagining a new interface: Hands-free communication without saying a word," Tech@facebook, https://tech. fb.com/imagining-a-new-interface-hands -free-communication-without-saying-a-word / (accessed August 20, 2019).

6. Cecilia Kang and Sheera Frenkel, "Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users," *New York Times,* April 2018, https:// www.nytimes.com/2018/04/04/technology /mark-zuckerberg-testify-congress.html (accessed August 20, 2019).

7. Priyanka Pulla, "The plan to mine the world's research papers," *Nature* 571 (July 2019): 316–18, https://www.nature.com /articles/d41586-019-02142-1 (accessed August 20, 2019).

8. Olexandr Isayev, "Text mining facilitates materials discovery," *Nature* 571 (July 2019): 42–43, https://www.nature.com /articles/d41586-019-01978-x (accessed August 20, 2019).

9. Meta, "Enabling editors through machine learning," Medium, https://medium. com/@meta_6493/enabling-editors-through -machine-learning-81b528b496ce (accessed August 20, 2019).

10. SPAR Ontologies, "About SPAR," http://www.sparontologies.net/about (accessed August 20, 2019).

11. David Pride, Jozef Harag, and Petr Knoth, "ACT: An Annotation Platform for Citation Typing at Scale," in: JCDL 2019 - ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES 2019 (Pride, David and Knoth, Petr eds.), Jun 2-6 2019, Urbana-Champaign, Illinois, http://oro.open. ac.uk/60670/ (accessed August 20, 2019).

12. Chris Bourg, "What happens to libraries and librarians when machines can read all the books?" *Feral Librarian,* https:// chrisbourg.wordpress.com/2017/03/16/what -happens-to-libraries-and-librarians-when -machines-can-read-all-the-books/ (accessed August 20, 2019).

13. Thomas Padilla, "Collections as data: Implications for enclosure," *College & Research Libraries News* [Online], 79.6 (2018): 296.https://crln.acrl.org/index.php/crlnews /article/view/17003/18751 (accessed August 20, 2019).

14. Jonathan Zittrain, "The Hidden Costs of Automated Thinking," *The New Yorker,* https://www.newyorker.com/tech /annals-of-technology/the-hidden-costs-of -automated-thinking (accessed August 20, 2019).

15. Charlie Harper, *Code4Lib Journal:* 41 (August 2018): "Machine Learning and the Library or: How I Learned to Stop Worrying and Love My Robot Overlords," https:// journal.code4lib.org/articles/13671 (accessed August 20, 2019).

16. Zittrain, "The Hidden Costs of Automated Thinking."

17. Tech@facebook, "Imagining a new interface." ✂