#NICAR15: Lessons from computational social sciences

Hanna Wallach (@hannawallach) is a researcher at Microsoft Research NYC and computer science professor at UMass Amherst, with a specialization in machine learning. During her postdoc, she realized that she wanted to use machine learning to understand how groups of people interact. Over the past few years, her work has been in this field of "computational social sciences" — a weird mishmash of computer scientists and sociologists who normally don't work together.

First, Hanna tells us about a 24-hour project where she and Columbia journalism professor Nick Lemann analyzed previously-classified government documents. Using machine learning, they looked at snippets of text to try and determine what might get redacted. Similarly, machine learning could be used to analyze a large corpus of emails, such as the Jeb Bush email dump.

You'd think that there are a lot of machine learning researchers looking at email data, but that's not true at all. "People don't write emails for journalists and researchers to analyze," says Hanna. This data can be really messy — there's no standard format. For example, quoted text poses a huge problem, as there are so many strategies for representing it. There are not many publicly-available email data sets, either, so there's little incentive for researchers to analyze it.

Much of Hanna's research is in the area of statistical topic modeling, which tries to answer the question: what is a set of documents about? These topics can be really specific and assign probabilities based on different keywords. Moreover, this technique can show the specific topic breakdown of a specific document, e.g. this email is 20% A, 50% B, and 30% C. This topical analysis can serve as a first stepping stone for surfacing stories and delving into these documents.

While there's been a lot of quantitative work looking at government at the federal level, there hasn't been much work on the local level. Using sunshine laws, Hanna requested internal email archives from every county in North Carolina. For those who didn't respond, she divided them into two groups: the first was simply asked again, while the second was asked with a list of counties that had complied. The latter had a much higher response rate, showing that counties are looking to their geographic peers for guidance on how to comply with these requests. In the end, Hanna was able to put together one of the largest email data sets gathered for any academic research project.

On the topic of interdisciplinary research, Hanna says it's important to keep the pace quick — that's what computer scientists are used to. As a result, it's essential to sit down with collaborators and explicitly listen to what research questions they're specifically interested in. Sometimes, there are great off-the-shelf machine learning tools that already exist and you don't need to bring in a researcher to collaborate with. One common task is classification (Naive Bayes), and there are many libraries in most popular programming languages that can do this for you. That being said, there isn't much incentive for the people working on these projects to create something really easy for end users.

From the point of view of getting help with using specific software, looking things up online forums are adequate. But if you want to learn about the underlying algorithms, universities are your best friends — you can often consult with students who are interested the field and professors (when they're free). Look at course syllabi for more resources, or email professors for their materials. Hanna recommends Kevin Murphy's textbook on machine learning (warning: math!) but it has a great table of contents that breaks down the field into helpful subtopics.

"It's very easy to take some data, take some machine learning techniques, get some output, and publish it," says Hanna. Every machine learning technique has implicit assumptions, and with the increased availability of these black-box solutions, it's easy to forget that. "Nothing that comes out of any of these algorithms is truth." It's important to contextualize your output in light of those assumptions.