Society | DOI:10.1145/2742484

Keith Kirkpatrick

# Putting the Data Science into Journalism

*News organizations increasingly use techniques like data mining, Web scraping, and data visualization to uncover information that would be impossible to identify and present manually.*

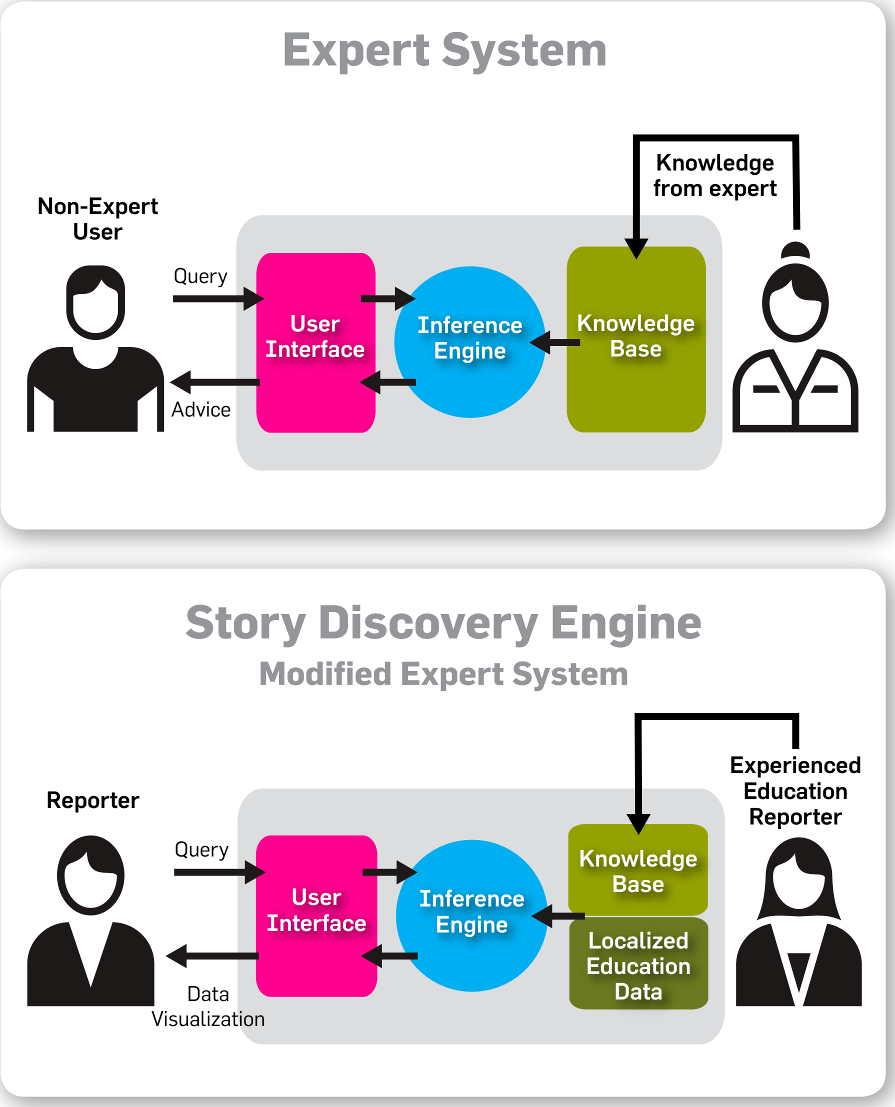
THE KEY ATTRIBUTES journalists must have—the ability to separate fact from opinion, a willingness to find and develop strong sources, and the curiosity to ask probing, intelligent questions—are still relevant in today's 140-character-or-less, ADHD-esque society. Yet increasingly, journalists dealing with technical topics often found in science or technology are turning to tools that were once solely the province of data analysts and computer scientists.

Data mining, Web scraping, classifying unstructured data types, and creating complex data visualizations are being utilized by news organizations to uncover data that would be impossible to compile manually, such as searching all Web retailers to find the average price of a particular item, given the vast number of potential sites to visit and the limited amount of time usually afforded to reporters on deadline. Additionally, the tools can be used to dig up or present data in a way that helps journalists generate story ideas, as well as presenting complex information to readers in ways they have not seen it presented before.

"There's this whole realization that if news organizations are to attract an audience, it's not going to be by spewing out the stuff that everyone else is spewing out," says David Herzog, a professor at the University of Missouri and academic advisor of the National Institute of Computer-Assisted Reporting (NICAR), part of Investigative Reporters and Editors (IRE), a non-profit organization dedicated to improving the quality of investigative reporting. "It is about giving the audience information that is unique, in-depth, that allows them to explore the data, and also engage with the audience," he adds.

One of the most interesting examples of programming technology being used to augment the reporting process comes from reporter and Temple University journalism professor Meredith Broussard, who was working on a series of stories for *The Atlantic* that focused on poor student test performance in the School District of Philadelphia.

As part of her investigative series, Broussard was trying to track down the number of textbooks in use in each public school in Philadelphia, and to



**A depiction of a traditional expert system compared with the Story Discovery Engine. Graphic by Meredith Broussard and Marcus McCarthy.**

IMAGE BY ANDRIJ BORYS ASSOCIATES, BASED ON GRAPHIC BY MEREDITH BROUSSARD & MARCUS MCCARTHY

see if the number matched the number of students enrolled. If there were not enough textbooks to go around, Broussard believed that could explain why so many students were performing poorly on tests directly tied to the curricula taught from those textbooks (a few main educational publishers design and grade the tests, and publish the books that students use to prepare for the tests).

"I wanted to do a story on whether there were enough textbooks in the Philadelphia schools," based on the number of students enrolled, Broussard explains. However, "it turns out that it is a really complicated calculation, and there wasn't software available to do that calculation for me."

To support Broussard's reporting, she built a data-analysis tool to crunch the complex data sets being analyzed for the article. To understand whether there really was one set of textbooks for each student, she needed to match the hundreds or thousands of student names associated with each school with book inventory lists (which could include various versions of each book and multiple titles within a set), and then repeat that process across the hundreds of schools within the district.

"The machine supercharges the reporter," Broussard says, noting that humans are limited, in terms of time and processing power, in a way that machines are not. Broussard observes that one reporter generally can analyze a single data set, whereas a team of a reporter, analyst, and editor usually can handle two or three large datasets. "But I needed to crunch at least 15 datasets for the story I wanted to do, and there was no way to do it using existing technology, so I had to create new technology."

In the end, Broussard's analysis yielded some shocking results. The average Philadelphia public school had only 27% of the books in the district's recommended curriculum, and at least 10 schools had no books at all, according to their own records. Some other schools had books that were extremely out of date.

The software tool used by Broussard was actually a prototype for the Story Discovery Engine, a tool that reporters can use to accelerate the process of finding investigative story ideas. It takes a traditional expert model of assessing information (applying a reporter's insight to a known knowledge base) and abstracts out high-level rules, then implements these rules in the form of database logic. When localized data sets are added, it then creates visualizations of the output that can show the reporter data anomalies or new avenues for exploration.

"You set up parameters, and then the system will spit out a visualization that will allow the reporter to come up with story ideas," Broussard explains.

She is quick to note that no technology, not even the Story Discovery Engine, is designed to or could replace a human reporter's instincts on what makes a good story. Says Broussard: "There's no such thing as a machine that comes up with story ideas, and spits them out. You don't actually want that, either, because computers are insufficiently creative. A human investigative journalist can look at the facts, identify what's wrong with a situation, uncover the truth, and write a story that places the facts in context. A computer can't. Investigative journalism plays an important role in preserving democracy; we need watchdog journalists to hold institutions accountable, to shine a light on misdeeds. The Story Discovery Engine allows investigative journalists to quickly uncover many stories on a topic as a way of getting broader and deeper coverage of important civic issues."

"A human investigative journalist can look at the facts, identify what's wrong with the situation, uncover the truth, and write a story that places the facts in context. A computer can't."

Some of the traditional news sources that have embraced this new world of data visualization include *The New York Times*, *The Los Angeles Times*, *The Chicago Tribune*, and *The Washington Post*, each of which has dedicated a significant level of resources and staffing to augment traditional reporters, often creating entire teams of analysts, designers, and computer programmers to ensure the data uncovered during the course of researching a story can be presented in a way that is interactive and meaningful to readers, many of whom now expect to be able to interact with that story on a personal level.

One such example is the Dollars for Doctors project, a project conducted by ProPublica, a self-described "independent, non-profit newsroom that produces investigative journalism in the public interest." The Dollars for Doctors project compiled the payments made to physicians by drug companies, and created an interactive tool that allows users to see exactly which payments were made to their doctor. The data itself was compiled from disclosures from 17 pharmaceutical companies, and then assembled into a single, comprehensive database.

Broussard says data visualizations and interactive tools are appealing to today's news consumers because they involve the readers in the story and localize events occurring in the news.

"I don't know where you live, so I can't draft a story [specifically] for you and show you what's happening in your neighborhood," Broussard says. "What I can do, however, is create a news app that lets you search your own address and find what's relevant to your neighborhood. Readers feel more engaged when they can put themselves into a story and find out what it means for their own lives."

The use of data to generate stories or provide context or color within them isn't new; reporters traditionally have had to dig through mountains of content to find statistics or data that helped drive or support a story. However, a key challenge faced by many reporters—particularly in the scientific field—is quickly pulling together stories on deadline while trying to analyze a wide range of technical papers and reports.

That is where Science Surveyor, a project launched in May 2014 and funded by the David and Helen Gurley Brown Institute for Media Innovation at Columbia University, seeks to assist science journalists and improve reporting on deadline by using natural language processing algorithms to review and analyze scientific research papers, and then provide a more detailed picture of each paper's findings in terms of its views, funding source, and other key attributes that likely will help frame the paper and its results.

"It's meant to provide, on deadline, a more nuanced view [of] where a particular paper sits, is it part of a growing consensus or not, and how is it funded," explains Mark Hansen, director of the David and Helen Gurley Brown Institute for Media Innovation. Hansen says the tool is designed to sift through and analyze research papers, and then creates visualizations summarizing the papers' timeliness, sources of funding or support, and which are part of a consensus view and which are outliers in their findings. Then, reporters are able to integrate "a more nuanced view of the topic or a particular paper," Hansen says, noting that providing context for specific research findings is key to ensuring better science reporting.

Clearly, creating data visualizations, developing programs to scrape Web sites, or even creating a new application to mine specific databases are tasks that require programming skills, and not everyone is jumping on board.

**Broussard says data visualizations and interactive tools are appealing to today's news consumers because they involve the readers in the story and localize events occurring in the news.**

"You still have division in newsrooms, where some people don't want to have anything to do with data," Herzog says. "Then you have people in the middle, with spreadsheet skills or some analysis, for enterprise stories. Then you'll have people who are really adept at learning new technologies, and who are pretty fearless at jumping in and trying to solve problems, and how to create a visualization."

It is the latter group of people that some of the top graduate journalism programs are seeking to attract. One such example is the Lede Program, a post-baccalaureate certification program jointly offered by Columbia University's Graduate School of Journalism and Department of Computer Science. The Lede Program is designed to offer journalism students a background in data, code, and algorithms, each of which are becoming increasingly crucial to research and data-centric journalism.

"A good story is a good story, whether it has data or not," Hansen says. However, with noted journalism schools at the University of Missouri and Columbia University actively teaching data- and computer science-based skills, "my guess is that over time, the term 'data journalist' will disappear and it will just become journalism," Hansen says.

**Further Reading**

*Bainbridge, L.*
Ironies of automation. *New Technology and Human Error*, J. Rasmussen, K. Duncan, J. Leplat (Eds.). Wiley, Chichester, U.K., 1987, 271–283.

The National Institute for Computer-Assisted Reporting
http://ire.org/nicar/

*Broussard, M.*
Artificial Intelligence for Public Affairs Reporting, *Computational Journalism*, http://bit.ly/1DGxiGL

The Data Journalism Handbook
http://datajournalismhandbook.org/1.0/en/

*Laroia, A.*
Web scraping: Reliably and efficiently pull data from pages that don't expects it, http://bit.ly/1LNaCKk

**Keith Kirkpatrick** is principal of 4K Research & Consulting, LLC, based in Lynbrook, NY.

Milestones

# ACM Bestows A.M. Turing, Infosys Awards

**At press time, ACM announced the names of this year's recipients of two of its most prominent awards.**

**The A.M. Turing Award, ACM's most prestigious technical award, is given for major contributions of lasting importance to computing.**

**This year's ACM A.M. Turing Award is being presented to Michael Stonebraker, an adjunct professor of computer science at the Massachusetts Institute of Technology (MIT), for "fundamental contributions to the concepts and practices underlying modern database systems."**

**The ACM-Infosys Foundation Award in the Computing Sciences recognizes personal contributions by young scientists and system developers to a contemporary innovation that, through its depth, fundamental impact, and broad implications, exemplifies the greatest achievements in the discipline.**

**This year's ACM-Infosys Foundation Award is being presented to Dan Boneh, professor of computer science and electrical engineering at Stanford University and leader of the applied cryptography group there, for "the groundbreaking development of pairing-based cryptography and its application in identity-based cryptography."**

**Both awards will be presented at the ACM Awards Banquet in San Francisco in June.**

***Communications* will provide in-depth coverage of both award recipients in upcoming issues, beginning with interviews with Stonebraker in the June issue.**

***—Lawrence M. Fisher***