Artificial Intelligence for Public Affairs Reporting

Meredith Broussard
Temple University
Annenberg Hall, 2020 N. 13th St.
Philadelphia, PA 19122
(215) 204-8358
merbroussard@temple.edu

ABSTRACT
In this paper, I describe the software system developed to support an investigative story series about textbooks missing from Philadelphia public schools. The software model, which is derived from an expert system, can be used as a template for news organizations that seek new tools for investigative reporting.

Categories and Subject Descriptors
I.2.5 [**Computing Methodologies**]: Programming Languages and Software – *Expert system tools and techniques.*

General Terms
Algorithms, Design, Experimentation, Human Factors, Theory.

Keywords
Expert systems, journalism, artificial intelligence.

1. INTRODUCTION
Computational methods offer promise for newsrooms seeking to amplify their investigative reporting capacity. This paper describes the software created for an investigative story series, the most recent installment of which is a story called "Why Schools Can't Win at Standardized Tests" that ran in *The Atlantic* in July 2014. The software is derived from an expert system. This model, dubbed the Story Discovery Engine, provides a template for news organizations that seek to leverage algorithms to accelerate the process of discovering investigative ideas on public affairs beats such as education, transportation, or campaign finance.

2. BACKGROUND
Journalism innovation theorists have suggested that tremendous possibilities exist in analyzing data to find investigative ideas. [1]–[4] Hamilton and Turner [5] write that the future of watchdog journalism may be found in using algorithms for accountability: "The best algorithms will essentially be public interest data mining. They will generate leads, hunches, and anomalies to investigate. It will remain for reporters and others interested in government performance to take the next step of tracking down the story behind the data pattern."

Accountability through algorithm can mean reverse-engineering an algorithm to discover how a company used an algorithm to influence the public [6]–[8] or it can mean designing an algorithm that is used to hold decision-makers accountable. I employ the latter meaning.

Tracking down a story in data requires specialized technical skills (to do the data-crunching) as well as journalistic expertise (to refine the story idea and craft appropriate prose). These skills until recently have tended to be segregated into different job categories and experience levels. A novice reporter might have sufficient technical skills to use pivot tables in a spreadsheet, for example, but might not have sufficient job experience to know that pivot table analysis could be applied to monthly government data releases on a particular beat. The promise of computational journalism is that such walls would be broken down through collaboration and training [9]. A successful computational journalism project might thus be described as one that uses computational thinking to bridge a knowledge gap.

This knowledge gap between the experienced and the novice reporter involves two types of knowledge: formal and informal. Formal knowledge includes rules of a system, as in knowing the rules of English grammar. An experienced education reporter has formal knowledge of his state's laws and policies around education. Informal knowledge includes domain expertise and rules of thumb based on experience. Informal knowledge for an experienced investigative reporter might include a rule of thumb like this: "If you have a natural disaster like Hurricane Katrina, and there is a big pool of money for hurricane relief, some of those funds will be misused; after a natural disaster, always follow up and find out where things went wrong with the government funds, and you'll find a story."

To come up with ideas the way an experienced reporter would, the novice reporter needs the informal knowledge that the experienced reporter has about where to find stories *plus* some of the formal knowledge about policies. In 2011, I found myself staring into exactly this type of knowledge gap. I was an experienced reporter, but my beat had been health, culture, and lifestyle topics. I wanted to investigate a question in education: do Philadelphia public school children have enough books to learn the material on the state-mandated standardized tests? I had data, I had methods, but I didn't have contacts. I wanted to talk to parents, teachers, and students at the city's best schools, and the city's worst schools, and see if there was a difference in the students' access to books. To do that, I needed to figure out which were the best schools, and which were the worst schools; I also needed to find people to talk to at each. There were more than 200 schools. The task was daunting.

Educational data is abundant, but the specific analysis I wanted had not been done before. It also involved numerous interdependencies and micro-judgments. To investigate the story I wanted to write, I turned to data journalism.

Data journalism is the practice of finding stories in numbers, and using numbers to tell stories [10]. It is an evolving practice [1] that may also be called data-driven journalism or computational journalism. Public affairs reporting is particularly suited to data journalism because government data is abundant and is often free.

I decided to develop a software system derived from an expert system. Public affairs reporting is ideally suited to expert system analysis because both disciplines rely on interpreting rules. An expert system works by drawing rule-based inferences from a knowledge base. A reporter on a public affairs beat such as campaign finance, education, or transportation must be familiar with a dizzying array of laws and policies at the federal and state

level. These laws and policies are articulated in text-based rules that are easily available online. The government uses data to track the success of its programs, and that data is frequently published online. Other data sets are available to reporters and citizens under the Freedom of Information Act of 1966.

I designed a software system that would allow me to juxtapose multiple data sets; calculate the answer to my investigative question; uncover additional stories; and identify potential sources.

3. THE STORY DISCOVERY ENGINE

3.1 Overview

Most people would be surprised at the idea that a public school wouldn't have enough books. After all, Pennsylvania law specifically says that the state provides books. In Philadelphia, however, students and parents regularly complain of textbook shortages.

Access to books is particularly critical because a school today is labeled a success or failure based on students' performance on high-stakes tests. The tests are highly specific and are aligned with state educational standards. The tests are also aligned with the textbooks sold by the three educational publishers that dominate the educational publishing market. These same publishers design and grade the standardized tests. It therefore stands to reason that if students don't have the right textbooks, they won't be able to do well on the tests even if they want to.

Answering the question of whether a single school has enough books is complex because each student in each grade studies at least 4 subjects every year. Asking if there are enough books in an entire school district is a massive task. With 200+ schools, the School District of Philadelphia is the eighth largest school district in the country. Many of the schools have high student turnover because students switch schools as they navigate the child welfare or juvenile justice systems [11]. The Children and Youth Division of the Philadelphia Department of Human Services serves an estimated 20,000 children and their families each year [12]. The District currently has 131,262 students in grades pre-K through 12, 87.3% of whom are economically disadvantaged. This is a significant issue because even if parents at each school fundraised for new books, there still might not be enough money.

My first step was creating a software system that would do some of the necessary investigative "thinking" for me. Clearly articulated rules in the real world can be translated easily into computer logic rules. Applying logical rules to a set of data allows the creation of a computational "intelligence" that a reporter can use to uncover social problems. Embedding formal and informal knowledge into the software would allow me (or any other reporter) to use the software as a reporting tool to refine story ideas and more efficiently find sources.

I called this reporting tool the Story Discovery Engine. It is based on the idea that it is possible to take some of the experienced reporter's knowledge, abstract out the high-level rules or heuristics, and implement these rules in the form of database logic. The data about the real world is fed in, the logical rules are applied, and the system presents the reporter with a visual representation of a specific site within the system. This visual representation triggers the reporter's creative process and allows him to quickly discover an idea for an investigative story.

3.2 Why Public Affairs Reporting

An investigation often arises when a reporter perceives a difference between *what is* (the observed reality) and *what should be* (as articulated in law or policy).

A high-impact investigative story looks at a situation where what is differs from what should be, and explains why. The reader can then use the narrative to create or enact a path to remedy the situation.

Public affairs reporting is thus ideally suited to the Story Discovery Engine model because *what should be* can be expressed in logical rules derived from laws and policies. The software compares this to *what is* (in this case education data), and shows how well (or how poorly) the rules are working in practice. The visual representation in the Story Discovery Engine allows the reporter to quickly spot anomalies that are worthy of further investigation. Most anomalies can be turned into news stories.

3.3 Equality

Though laws vary from state to state, in general American public school guarantees each student equal access to education. Regarding textbooks, Pennsylvania law states: "The board of school directors of each school district shall purchase all necessary furniture, equipment, textbooks, school supplies, and other appliances for the use of the public schools… and furnish the same free of cost for use in the schools of the district."[13]

This concept can be expressed mathematically in order to form a working definition of "enough books." A school with equal access to education is a school where:

number_of_books = number_of_students

This is not a straightforward calculation, however. Each student is enrolled in at least four core subjects, and a school day has about eight periods. I also wanted to create a way to resolve any issues I discovered: if there was a shortage of books at a school, I wanted the tool to calculate the amount of money it would take to resolve the shortfall. In this way, the tool would not just identify a problem; it would also allow the viewer to see a possible solution.

This calculation required accurate pricing data. At first, I planned to use an API to pull book prices. However, because each school district negotiates purchasing agreements with textbook wholesalers that allow for volume discounts, taking pricing data from a third party like Amazon or AbeBooks was not feasible. I used pricing data from the School District of Philadelphia's vendor agreements. Other data sets included a list of books at each school; a list of curricula in use at the District; the number of students enrolled in each grade at each school; five years of PSSA standardized test data for each grade at each school; and the necessary materials in each curriculum.

I created an object model that would account for all of the relationships between entities. For the sake of brevity, that object model is omitted from this paper.

3.4 Expert Systems

The Story Discovery Engine software is derived from a class of artificial intelligence programs called knowledge-based expert systems. Benfer [14] offers an excellent definition:

> Expert systems are computer programs that perform sophisticated tasks once thought possible only for human experts. If performance were the sole criterion for labeling a program an expert system, however, many decision support systems, statistical analyses, and spreadsheet programs could be called expert systems. Instead, the term "expert system" is generally reserved for systems that achieve expert-level performance, using artificial intelligence programming techniques such as symbolic representation, inference, and heuristic search (Buchanan, 1985). Knowledge-based systems can be distinguished from other branches of artificial intelligence research by their emphasis on domain-specific knowledge, rather than more general problem-solving strategies. Because their strength derives from such domain-specific knowledge rather than more general problem-solving strategies (Feigenbaum, 1977), expert systems are often called "knowledge-based."

Benfer argues that expert systems can provide an important mechanism for prompting new social science thinking, and expert system developers can learn from social scientists' rigorous methods of data collection and validation. He was the first to deploy an expert system in journalism:

> MUckraker, an expert system under development by New Directions in News and the Investigative Reporters and Editors Association at Missouri University, is a program to advise investigative reporters on how to approach people for interviews, how to prepare for those interviews, and how to examine a wide range of public documents in the conduct of an investigation. This program is designed to act much as an expert investigative reporter might, advising the user on strategies to try when sources are reluctant to be interviewed, pointing out documents that might be relevant to the investigation, and advising the user on how to organize his or her work. [14, p. 4]

Under the expert system model Benfer describes, the expert system would deliver to the reporter "advice" about whether the quantity of books in a school would be the appropriate basis for a story. The innovation in the Story Discovery Engine is that instead of advice, the expert system delivers an interactive data visualization. The data visualization is specifically designed to answer the most common questions a reporter might ask in order to assess whether a story might be found at a particular school.

It is significant that Benfer used social science methods in crafting an expert system for journalism. Social science thinking is at the heart of what today we call data journalism. Meyer [15] pioneered the application of social science methods to journalism in his 1967 Pulitzer Prize-winning story about race riots in Detroit; those methods were later codified in *Precision Journalism: A Reporter's Introduction to Social Science Methods*. Precision journalism methods informed computer-assisted reporting, which flourished in the 1980s with the advent of desktop computers in the newsroom. Today's online data journalists are incubated and organized by the Investigative Reporters and Editor's Association through the National Institute for Computer-Assisted Reporting (NICAR), which offers the Phil Meyer Reporting Award for a data-driven project each year.

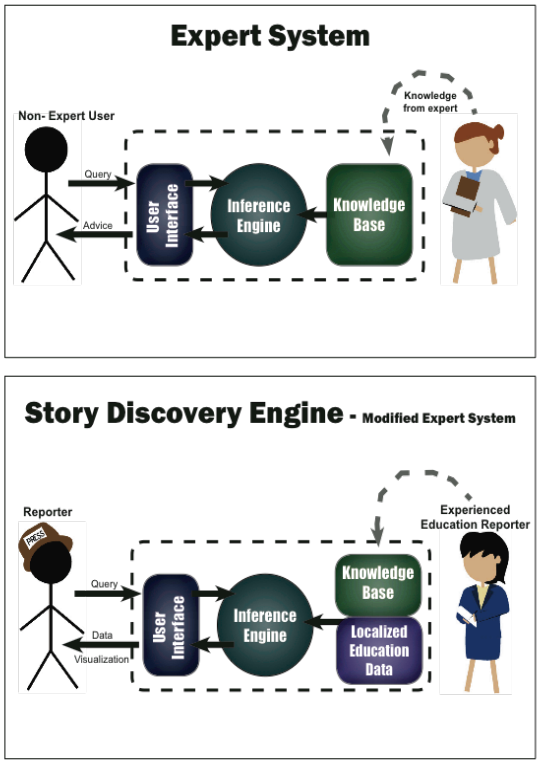

Figure 1. An expert system and the Story Discovery Engine

The Story Discovery Engine substitutes data visualization for the advice offered in the traditional expert system model. The reporter thus becomes part of the computational system: the reporter's interpretation of the simple data visualization is the final step in determining what (if any) story could result.

The notion that computer-based quantitative methods should augment humans, not replace them, is one of the principles of automated text analysis put forth by Grimmer and Stewart [16] in their analysis of possible pitfalls in automated content analysis. In recent years, communication scholars have frequently used the human workers who participate in Amazon's Mechanical Turk in order to code content in large data sets. In the Story Discovery Engine model, the reporter is a similarly essential part of the system.

Using the vast "computational" resources of the human brain, the reporter takes only moments to look at the data revealed by the system, leverage formal and informal knowledge, and make a judgment about the likelihood of a story. It would require vast amounts of computing power to get the computer to draw the same conclusions; it could take years to tease out all of the subtleties of human news judgment and implement them computationally. The human brain thus becomes an efficient part of the story-generating process, aided and augmented by the computational system.

It is worth mentioning at this point the relationship between software tools and reporters' productivity. Several web-based tools have been developed to help journalists be more efficient at

their investigative tasks. Tabula, for example, turns PDFs into text. One of the most consistent points of conflict between reporters and officials is the way that the officials provide information. Entire books have been written about the nuances of negotiating for access to public records [17], [18]. A successful tool for investigative journalism allows reporters to surmount common difficulties that interfere with reporting. Likewise, several data visualization tools have become popular to use on structured data. Putting census data into a data visualization tool like Tableau, which displays maps and bubble charts and other forms, allows the reporter to see patterns that would otherwise be invisible.

A small but growing subset of journalists is comfortable using data to enhance their abilities to investigate stories. However, those reporters are limited to using the number of data sets that they, or their newsroom team, can manage. Analyzing one data set is usually enough for a story. Analyzing two or three data sets and turning them into a story package requires a team that includes a programmer, designer, writer, and editor. [19]–[21] The Story Discovery Engine provides a way to use 15+ data sets simultaneously.

4. BACKGROUND FOR THE STORY

4.1 The Development Process

The Story Discovery Engine prototype launched online as a project called "Stacked Up." It has two parts: it is both a reporting tool and a presentation system for the stories I wrote using the reporting tool. The presentation system provides the user with a set of investigative stories and some explanatory text about the project. The reporting tool is a set of dynamic data visualizations that allowed me to write the investigative stories. The statistics and data that supported each story were original, derived from the data analysis resulting from the algorithm that forms the backbone of the project. The story in *The Atlantic* may be seen online at http://tinyurl.com/qhmnq4k and the project may be seen at http://www.stackedup.org.

In the reporting tool view, the reporter sees a page representing a single school. The page shows different types of data, organized so that specific types of investigative questions can be easily answered. Some such questions include:

How many students are in each grade in this school?

Where is the school located in the city?

How does this school's test results compare to the rest of the district?

Do there seem to be enough books for the students enrolled?

The system design anticipates the data points that a reporter needs to write a data-rich story and presents them in a centralized, easy-to-navigate format. The reporter leverages their domain expertise, clicks around to adjust some what-ifs to prompt the creative process, and comes up with a story idea. Because the story idea is targeted, it immediately becomes easier to identify appropriate sources.

The key is that the software doesn't try to solve a problem faced by all journalists on every beat. It tries to solve a specific problem on a specific beat, and in the process creates a way to solve other problems on that same beat. The Story Discovery Engine prototype was created and applied to education data, but the model can easily be applied to other beats as well.

Without going into excessive technical detail, the algorithm looked something like this:

Core_subjects = math, reading, social studies, science

School_curriculum = a curriculum package published by a major educational publisher (e.g., "Everyday Math")

Necessary_material = the minimum books or workbooks necessary to teach the school's curriculum package. This often means 2 items: a textbook and workbook.

```
For each school in School_District
  For each grade in school
    For each Core_subject
      For each Necessary_material in School_curriculum
        If
          NumberOf(students_in_grade) =
          NumberOf(necessary_material)
        Then
          Enough_materials = yes
        Else
          Enough_materials = no
```

Once the prototype existed, I looked at the data analysis and interviewed people to validate the findings. I developed hypotheses, reported them out, revised the hypotheses, and considered story formats as part of a long and intricate process. As predicted, the data revealed multiple potential stories about how books were "stacked up" in Philadelphia city schools.

The project took three developers six months to implement. While journalists might consider this a long development cycle, software developers might consider this a reasonable time frame. Moreover, now that the model is built, it can be replicated for any other school district in the US. Inputting the data and running the same analysis for a different district could happen in a matter of days, not months.

5. Findings and Directions for Future Research

I theorized that the Story Discovery Engine model could accelerate the production of ideas and stories on a public affairs beat. I prototyped the software and used it to report on a specific beat. The public reaction to the project validates the theory and suggests the Story Discovery Engine model as a valid option for creating impactful news.

Among the project's findings:

- Only a handful of Philadelphia schools seem to have enough books and learning materials to adequately teach students under the district's academic guidelines.

- At least 10 schools appear to have no books at all, others seem to have books that are wildly out of date, and some seem to have only the books that fit the curriculum guidelines established by a chief academic officer who left the district years ago.

- Despite investing in custom software to track its textbook inventory, the District did not require any of its employees to use the software.

- The District spent $111 million on textbooks between 2008 and 2013. Its inventory showed more than a million books. Nobody knew where they were; boxes and boxes of books lay unused and un-catalogued in the basement at District headquarters.

- The District published a recommended core curriculum, but didn't know if any of its schools were using it. There was no

systematic way to determine if struggling schools had the books and resources they needed for student success.

These findings, once published, were shared extensively on social media and prompted a number of changes at the School District of Philadelphia. Outcomes in subsequent weeks included:

- One highly paid administrator was found to be responsible for a number of textbook tracking failures. That administrator retired immediately. Another administrator involved in the failures left several months later.

- An internal investigation revealed that several school principals were buying textbooks from sales representatives with whom they had personal relationships instead of buying the textbooks recommended by the central administration. Some of the reps were former school principals. This practice was eliminated and cost savings were achieved [22].

- The School District of Philadelphia closed 24 schools at the end of the 2012-2013 school year, displacing approximately 4,000 students. Originally, the District planned to send all the books from the closing schools to the schools that were slated to receive the students. Instead, the District collected all of the books from the closing schools at a central location. An attempt was made to organize the books and reallocate them judiciously.

- An audit was performed so that the central administration was made aware of the curriculum officially in use at each school.

- Several local news organizations picked up the investigative stories and re-published them on their own websites, amplifying the audience for the stories.

This modest impact suggests that the reporting could be duplicated in other large cities like Philadelphia, all of which struggle with similar logistical issues around public education resources.

The Story Discovery Engine model also solves a logistical issue specific to newsrooms. A newsroom depends on highly specialized labor. The writers are good at writing, the editors are good at editing, the web producers are good at using the content management system, and the programmers are good at writing code. It makes sense to have the programmers write the code that teases out the facts the reporters need to write stories. Getting the reporters to write code is less practical. However, few newsrooms have the staff that would be required to write high-level code [20], [21], [23]. Despite the enthusiasm for data journalism, the logistics of performing data journalism have proved formidable for many news organizations.

Creating a Story Discovery Engine for a metropolitan area, then opening it up to the public, allows more people to leverage the code to write stories. The engine could also be implemented by a foundation and opened up to the public; the local press could use it to write stories without having to fund the development or hire and manage a software staff.

6. References

[1] E. Appelgren and G. Nygren, "DATA JOURNALISM IN SWEDEN: Introducing new methods and genres of journalism into 'old' organizations," *Digit. Journal.*, pp. 1–12, Feb. 2014.

[2] M. Dick, "INTERACTIVE INFOGRAPHICS AND NEWS VALUES," *Digit. Journal.*, pp. 1–17, Sep. 2013.

[3] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "RESEARCH METHODS IN THE AGE OF DIGITAL JOURNALISM: Massive-scale automated analysis of news-content—topics, style and gender," *Digit. Journal.*, vol. 1, no. 1, pp. 102–116, Feb. 2013.

[4] J. V. Pavlik, "INNOVATION AND THE FUTURE OF JOURNALISM," *Digit. Journal.*, vol. 1, no. 2, pp. 181–193, Jun. 2013.

[5] J. T. Hamilton and F. Turner, "Accountability Through Algorithm: Developing the Field of Computational Journalism," Stanford University, Center For Advanced Study in the Behavioral Sciences Summer Workshop, Jul. 2009.

[6] N. Diakopoulos, "Rage Against the Algorithms," *The Atlantic*, 03-Oct-2013.

[7] N. Diakopoulos, "Algorithmic Accountability Reporting: On the Investigation of Black Boxes." Tow Center for Digital Journalism at Columbia University, Feb-2014.

[8] L. Sweeney, "Discrimination in online ad delivery," *Commun. ACM*, vol. 56, no. 5, p. 44, May 2013.

[9] T. Flew, C. Spurgeon, A. Daniel, and A. Swift, "THE PROMISE OF COMPUTATIONAL JOURNALISM," *Journal. Pract.*, vol. 6, no. 2, pp. 157–171, Apr. 2012.

[10] A. B. Howard, "The Art & Science of Data-Driven Journalism," Columbia University, Tow Center for Digital Journalism, May 2014.

[11] Department of Human Services, City of Philadelphia, "2011 Annual Report," Annual Report, 2012.

[12] Department of Human Services, City of Philadelphia, "Children and Youth Division Home Page." 2014.

[13] *Pennsylvania Public School Code of 1949*, vol. 14. 1949.

[14] R. A. Benfer, *Expert systems*. Newbury Park, Calif: Sage Publications, 1991.

[15] P. Meyer, *Precision journalism: a reporter's introduction to social science methods*, 4th ed. Lanham, Md: Rowman & Littlefield Publishers, 2002.

[16] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Polit. Anal.*, vol. 21, no. 3, pp. 267–297, Jul. 2013.

[17] D. Cuillier, *The art of access: strategies for acquiring public records*. Washington, DC: CQ Press, 2011.

[18] D. Marburger, *Access with attitude: an advocate's guide to freedom of information in Ohio*. Athens: Ohio University Press, 2011.

[19] D. Domingo, "Interactivity in the daily routines of online newsrooms: dealing with an uncomfortable myth," *J. Comput.-Mediat. Commun.*, vol. 13, no. 3, pp. 680–704, Apr. 2008.

[20] S. Parasie and E. Dagiral, "Data-driven journalism and the public good: 'Computer-assisted-reporters' and 'programmer-journalists' in Chicago," *New Media Soc.*, vol. 15, no. 6, pp. 853–871, Nov. 2012.

[21] C. Royal, "The Journalist as Programmer: A Case Study of The New York Times Interactive News Technology Department," presented at the International Symposium in Online Journalism, The University of Texas at Austin, 2010.

[22] J. Diaz, "Personal communication," Nov-2013.

[23] R. W. McChesney, "FAREWELL TO JOURNALISM?: Time for a rethinking," *Journal. Pract.*, vol. 6, no. 5–6, pp. 614–626, Oct. 2012.