DATA-DRIVEN REVELATION? Epistemological tensions in investigative journalism in the age of 'big data'

Sylvain Parasie

DATA-DRIVEN REVELATION?
Epistemological tensions in investigative journalism in the age of 'big data'

Sylvain Parasie

As an increasing number of reporters see databases and algorithms as appropriate means of doing investigation, journalism has been challenged in recent years by the following question: to what extent would the processing of huge datasets allow journalists to produce new types of revelations that rely less on normative assumptions? Drawing on the analysis of a particular investigation by the San Francisco-based Center for Investigative Reporting*, this article points out the existence of epistemological tensions in the making of journalistic revelations that involve the processing of vast amounts of data. First, I show that the design of data-processing artifacts can match the traditional epistemology of journalistic investigation, but only with great efforts and resources from the organization. Second, I point out that the use of these artifacts by journalists follows two opposite paths to produce the revelation: a "hypothesis-driven" path and a "data-driven" path. Such findings contribute to a better understanding of how news organizations produce justified beliefs, as data-processing artifacts become major components of the newsroom's environment.*

KEYWORDS
big data; computational journalism; data journalism; epistemologies of journalism; investigative reporting

Introduction

In the last decade, computer databases and algorithms have made their way into news organizations (Gray et al. 2012; Lewis 2011), where they are used in particular as tools supporting journalistic investigation. With the rise of so-called "data journalism" in the United States as well as in Europe, a growing number of journalists and programmers see data-processing tools as appropriate means of uncovering officials' wrongdoings, social inequities or environmental issues (Cohen et al. 2011; Parasie and Dagiral 2013). Established news organizations (such as *The New York Times* or *The Guardian*) as well as nonprofit organizations (such as ProPublica) and less formal groups of investigative journalists have produced revelations based on data-processing techniques.

Such initiatives cannot be isolated from a wider phenomenon often labeled "big data." This popular expression refers to the processing of massive quantities of

information—government records, genetic sequences, traces left by internet users, etc.—in various domains such as scientific research, public policies or business. One of the promises of "big data" is that the statistical processing of huge datasets could facilitate revelations about nature or society without relying on any theoretical or normative assumptions (Anderson 2008). As the authors of a recent book suggest, "big data may offer a fresh look and new insights precisely because it is unencumbered by the conventional thinking and inherent biases implicit in the theories of a specific field" (Mayer-Schönberger and Cukier 2013, 71).

Journalists may be receptive to this renewed promise of objectivity. Since the late 19th century, the distinction between facts and values has indeed been a strong occupational norm for North American journalists, who have largely emphasized the ideal of a reporter gathering "facts" in a detached, unbiased and impersonal manner (Schudson 1978, 2001). And if we consider investigative journalists especially, a realist conception of the truth is often said to be prevalent among them: in their opinion, there is only one true and complete statement of what happened; and what happened is independent of their investigation. But scholars have shown that, in practice, investigative journalists base their investigation on a strong interdependence between facts and values. The facts they collect in the investigative process are inherently value-loaded and, conversely, the values they rely on hold some facts (Ettema and Glasser 1998). This is why we have to ask: to what extent does the processing of vast amounts of data affect how journalists produce knowledge in the process of an investigation? Does it modify the way they distinguish between facts and values in the making of a revelation?

Studying the "epistemologies" of knowledge producers is a classical approach in the sociology of knowledge, and more broadly in the area of science and technology studies (Knorr-Cetina 1999). It implies analyzing how actors make knowledge claims that they collectively find acceptable, and not evaluating whether those claims are valid or not. In the area of journalism studies, James Ettema and Theodore Glasser (1998, 2005) have adopted this perspective fruitfully to study investigative journalists, and other scholars extended this questioning to other types of journalism (Godler and Reich 2013). Since the goal is not to evaluate whether journalists' knowledge claims are valid or not, but merely to examine what journalists consider to be acceptable claims, this has been a significant departure from many sociological works that have been skeptical about journalistic claims to objectivity (Tuchman 1972; Hallin 2005).

My approach differs in two respects from the usual research on journalistic epistemologies. First, I take into account, in greater depth than do previous works, how journalists rely on the material environment of their organization to decide whether their knowledge claims are justified or not. For journalists, databases and algorithms are not black boxes providing unquestionable results, and we need to examine the material basis on which they collectively hold a specific output as being justified. Second, following Bruno Latour and Steve Woolgar's account of how science is made in the laboratory (Latour and Woolgar 1979), I disregard innovators' public accounts of how technology affects the epistemology of reporting, and focus rather on the often tortuous history of how justified beliefs are collectively produced in relation to artifacts.

Drawing on the analysis of a particular investigation performed by a San Francisco-based news organization, this article shows the existence of epistemological tensions in the making of journalistic revelations that involve the processing of vast

amounts of data. Given the empirical limitation of this case, my intention is not to identify emergent trends regarding the use of data-processing artifacts in newsrooms in the U.S. and elsewhere. My aim is rather to point out the existence of epistemological tensions affecting investigative projects that involve that kind of artifact. These tensions are of two types. First, I show that the design of data-processing artifacts can fit the traditional epistemology of journalistic investigation, but only as the result of a long and costly process. Second, I point out that the use of these artifacts by journalists follows two opposite paths to produce the revelation.

Methods

In April 2011 the *Center for Investigative Reporting*, based in the San Francisco Bay area, revealed that the State of California had failed to enforce earthquake safety standards in public schools. I studied that particular investigation for two main reasons. First, because a team composed of journalists with heterogeneous backgrounds performed it: experienced investigative reporters with few skills in data processing, journalists with a background in "computer-assisted reporting," and programmer-journalists connected with the "data science" community that is very active in the area. Since the team had no shared culture concerning the design and use of data-processing artifacts, this project provided a good opportunity to examine the epistemological tensions in the collective making of a revelation in relation to such artifacts. The second reason is that this 19-month investigation consisted of a combination of quantitative and qualitative research methods. Journalists not only designed databases, performed statistical analyses, and built an interactive map; they also collected interviews and documentation. This afforded us the opportunity to identify precisely how data processing gave rise to epistemological tensions, in relation to the use of more traditional methods.

The fieldwork was carried out in collaboration with Eric Dagiral in the San Francisco Bay area between mid-August and mid-September 2012. We conducted five in-depth interviews with journalists involved in the project. Our questions aimed at reconstructing the history of this investigation and of how the team collectively dealt with databases and algorithms over time. Analyzing the rise of certainties and doubts within the team and how they evolved over the 19-month period was a major concern in this study. I also analyzed several versions of the databases designed by the team, examining how they were structured over time and processed to design outputs (maps, tables, lists) used in the investigation.

The Problematic Epistemologies of a "Data-Driven" Investigation

With the growth and spread of data-processing artifacts, several institutions have come to question the grounds of their knowledge production. Scientific institutions, notably, have been challenged by the idea that the statistical processing of huge datasets might allow them to produce new types of revelations about nature or society that rely less on theoretical or normative assumptions. Scholars in the study of science and technology have shown the existence of epistemological tensions within some scientific areas, between "hypothesis-driven" and "data-driven" perspectives (Strasser 2011). In the biomedical sciences, for instance, Keating and Cambrosio (2012) have

shown that the field of cancer research features conflict between bio-informaticians, who support a data-driven view, and biostatisticians, who support a hypothesis-driven view.

Although data-processing artifacts are obviously not as commonplace in journalism as they are in science, databases and algorithms have been used in journalistic investigation since the late 1960s within the North American tradition of "computer-assisted reporting" (Cox 2000), and have been increasingly regular tools for news organizations since the 1990s (Garrison 1998). Nevertheless, as the "data-driven approach" conflicts with the established epistemologies of investigative reporting and computer-assisted reporting, it may be problematic for investigative reporters to adopt it.

The Epistemology of Investigative Journalism

Ettema and Glasser (1998, 2005) offer us a compelling portrait of the epistemology of U.S. investigative journalism in the late 1990s. They showed that this epistemology has three main elements. Firstly, investigative reporters rely on an "externalist" approach to the truth, assuming that "there is only one true and complete description of the way the world is" (Ettema and Glasser 1998, 134). They strongly believe they can and must find out what really happened. Secondly, investigative journalists collect several accounts of what happened from different sources—through interviews and documentation—and then produce a new account that they believe is more authoritative. What is really important to them is the correspondence not between reality and their account, but rather between the various accounts they have collected. According to Ettema and Glasser, investigative reporters build their new authoritative account on the assumption that the reality must be coherent, determinate, and non-contradictory (137). Thirdly, the facts and the story emerge simultaneously in the investigation. While on the one hand the reporter collects documents and interviews on the basis of an initial story that identifies the issues, on the other hand the collected facts limit his or her choice of a story.

Such an epistemology is far removed from a model where the facts are first collected with few assumptions, and are then analyzed and composed into a story. Identifying leads in the data, formulating hypotheses, and collecting facts on the basis of an already-structured story: all of this might cause some epistemological problems for journalists as they consider the adoption of a "data-driven approach" in investigative reporting.

The Epistemology of Computer-Assisted Reporting

Dealing with data-processing artifacts in order to investigate is far from being a recent phenomenon in the U.S. Since the late 1960s, the "computer-assisted reporting" tradition has not only fostered the adoption of these artifacts in North American newsrooms, but also developed a coherent framework of epistemological standards to make databases more regular aspects of investigation (Parasie and Dagiral 2013). Such standards have been explored and disseminated by the National Institute for Computer-Assisted Reporting (NICAR) and the various handbooks published on the topic (Garrison 1998; Houston et al. 2002).

One major standard is that data has no journalistic value in and of itself: database-oriented operations are viewed as valuable only when they are subordinated

to a story idea (Garrison 1998, 281). Accordingly, the reporter has to process the data on the basis of assumptions regarding the issue concerned, the actors involved, and their responsibilities. Another standard is that journalistic norms remain entirely valid in such operations: checking data for accuracy, cross-comparing the accounts from various sources, etc. Moreover, this framework does not attribute any significant value to the processing of huge and complete datasets; only the handling of samples is important (Parasie and Dagiral 2013). It thus follows that the epistemological framework conveyed by the computer-assisted reporting tradition conflicts with the "data-driven" approach to journalistic revelations.

But there may be a huge gap between how journalists view epistemology in their public accounts and the actual epistemological grounds of their investigation—as in science (Latour and Woolgar 1979). I therefore chose to document how journalists collectively produce a revelation based on the processing of vast amounts of data.

The Challenge of Adjusting the Artifacts to Established Epistemologies

On April 7, 2011, the San Francisco-based Center for Investigative Reporting (CIR) revealed systemic breakdowns in the way the State of California enforced seismic safety standards in the construction of public school buildings. For nineteen months a dedicated team within this non-profit news organization relied on a combination of various methods to produce this revelation:

> Tonight, a 19-month investigation by the Center for Investigative Reporting finds the state is failing to enforce earthquake safety standards in Californian public schools. It uncovers faulty constructions as well as a troubling lack of oversight by those in charge of keeping our children safe. (Transcript from a KQED television special, April 15, 2011)

During the investigation, the team designed databases and algorithms that matched the established epistemologies of journalistic investigation. Accordingly, it rejected the idea that the artifacts could encapsulate a hidden truth. Thus, the adjustment that had been so problematical for a long time was finally made, but it was demanding in terms of the resources made available by the organization.

Revelation

A series of articles entitled "On Shaky Ground" were published on the CIR website in April 2011. They made the following assertions:

1. State regulators have routinely failed to enforce California's landmark earthquake safety law for public schools, by allowing children and teachers to occupy buildings with potential safety hazards reported during construction.
2. State regulators have approved for jobs most of the inspectors accused of falsification and absence.
3. The state has made it virtually impossible for school districts to access a fund set aside for urgent seismic repairs.

4. Lobbyists and private interests have largely captured the regulation of school seismic safety.

Those articles put the blame primarily on the Division of the State Architects, the agency in charge of enforcing the regulation, and claimed that children in California public schools were put at risk in case of an earthquake. Moreover, an interactive map was released on the website, allowing any user to check whether a particular school was reviewed as safe or not, and its proximity to seismic zones.

[Insert Figure 1 about here]

As Corey Johnson—the journalist who led the investigation—told us, "The thought from experts and the larger public was that all schools were complying. So it took a lot of them by surprise when we reported that they did not." And this revelation was taken seriously, not only by parents concerned about the safety of their children, but also by governments and other media organizations. This led to a nationwide scandal as local and national media spread the news and commented on it.[1] The majority leader of the California State Legislature said, "It is unacceptable to allow children to use facilities that are unsafe," and the state legislature initiated a public inquiry and introduced a bill intended to enhance the seismic safety of public schools.[2] The investigative team consequently received several professional awards, including a finalist honor for the Pulitzer Prize in 2012 and top distinction from the Investigative Reporters and Editors association in the same year.

Origins of a "Big Story"

In September 2009, the reporter Corey Johnson was asked to write a story on the seismic safety of Californian schools for the 20th anniversary of the Loma Prieta earthquake.[3] At the time, the general opinion was that the Field Act was correctly enforced. Passed in 1933, this Act had created an agency, the Division of the State Architect (DSA), whose mission is to approve school building projects and to regularly inspect the construction while underway.

As he started looking for information about the seismic safety of schools in California, Corey first interviewed a couple of experts. In order to obtain a list of unsafe schools, he submitted a data request to the DSA. The agency sent him back a spreadsheet indicating that more than 9,000 schools in California did not comply with safety standards. At that point, Corey realized he was onto a "big story":

> Once I got that, that's when I knew that there was a big story here because the law was so strict that it said not one single school can violate this law—that 9,000 in a list that appear to violate the law. So how does that happen? (Johnson, interview, August 31, 2012)

The spreadsheet enabled him to persuade his chief editors that he should have more time and resources to investigate deeper. This file appears to have been what he calls an "initial guide" in his investigation because it offered two important features. First, it gave quantified insight on the regulatory failure. Second, it offered information about

the potentially unsafe schools, notably their names and addresses, which allowed him to believe that the investigation was "do-able." This initial view on the data in the investigative process has many similarities with the way in which investigative reporters assess the value of a tip (Ettema and Glasser 1998).

At this stage, the CIR chief editors strongly advised the team to design a comprehensive database of schools located in seismic hazard zones. The goal was then both to produce an artifact allowing reporters to evaluate and prove the massive regulatory breakdown, and to enable parents to check the safety of their children's school through an interactive map. This database was originally designed from three different datasets from government sources: one from the DSA regarding the safety status of every public school in California; another from the California Geological Survey—the agency in charge of the identification and the mapping of geological hazards in California—regarding the location of every seismic zone in the state; and another from the California Department of Education, which provided information about the location of every school building in the state. Throughout the investigation, this database was fed with several other datasets from the same and other institutions.

This database was designed essentially for gathering and collating, in the same artifact, various accounts held by different institutions. Since the accounts of the DSA were seriously questionable in a context of a massive regulatory crisis, the team saw the inclusion of other accounts from other agencies in the same database as a good means of producing a new authoritative account of the issue. Reporters could then collect information about the location of a particular school to point out that the fact of a school not having been inspected, or having been reviewed as unsafe because of its location near the fault—thus revealing a major failure of the regulatory agency. As noted above, this approach is largely in line with the established epistemology of investigative journalism.

"A Massive and Complicated Topic"

Once the journalists involved were convinced a "big story" was at stake, the investigation followed two different but connected paths: Corey Johnson collected interviews and documents to account for this massive regulatory crisis, while Agustin Armendariz, a "data analyst," was in charge of designing the database. However, as Agustin started merging the various datasets in October 2009, a host of difficulties arose, essentially due to the emergence of a gap between the epistemological framework that the journalists involved had relied on, and the features of the investigation.

The first difficulty stemmed from the extreme messiness of the data from state agencies. Schools were often labeled under different names in the DSA datasets and in those from the Department of Education. As there are tens of thousands of public schools in California, the merging of such messy datasets seemed almost impossible:

> It was a hard task. I mean, there was the messiness of the information. There was me trying to figure out how to convey to people what I think we can and can't do in a way that they would trust. (...) The big challenge of that process is that it was very much us getting to know each other and dealing with a massive, massive and complicated topic. (Armendariz, interview, September 12, 2012)

The inaccuracy of the data was another source of concern for Agustin and the journalists involved. Because state regulators seemed to have experienced major organizational issues, it appeared risky for the team to take it for granted that the regulators' records were accurate and factual. Moreover, as the datasets from the Department of Education were full of misspelled school names and poorly located school buildings, locating each school was very difficult. Because he used algorithms affecting the distance of each school from seismic zones, Agustin felt he could not guarantee the accuracy of the schools' location on the map:

> This is not survey data. I don't know how these things were projected, and I reprojected them when I stitched them together. I don't know how accurate that school point is. It's not the footprint of the school, and I don't even know where that building falls on campus. I can't make a 50-foot measurement like that. (Armendariz, interview, September 12, 2012)

The messiness and inaccuracy of the data also raised an ethical concern. As Agustin told us, it was very important for the CIR to avoid making wrong claims about the safety of a school. Falsely claiming that a school is unsafe for children might provoke unfounded reactions from people and ruin the reputation of the news organization:

> No information is ever clean. No data is ever perfect. I'm willing to accept that. But given that limitation, it's really important and really necessary and really hard to figure out what we can responsibly say with this information. (Armendariz, interview, September 12, 2012)

In the first six months of the investigation the most serious difficulties concerned the schools' location and the evaluation of their safety. For the journalists involved in the investigation, the geological data seemed the only element they could take for granted. Unlike government data, the geological information was produced by scientists. But in February 2010, Corey got a tip from a former geologist who used to work for the regulatory agency. According to this source, the geological map of seismic zones in California had changed substantially over the last decades, and not for scientific reasons. After checking state archives, Corey found out that the map had indeed been largely redrawn—under the pressure of private interests. Even geological data appeared to be political and, accordingly, largely questionable.

Designing a Database is a "Reporting Process"

A gap was opening up between the established epistemologies and the design of such a massive and comprehensive database. Moreover, it became obvious for the team that another big issue was the growing rift between the two fronts of the investigation. Whereas Corey was collecting enormous amounts of information that consisted mostly of qualitative data, Agustin was focused only on technology. The team found it very difficult to connect what Agustin was seeing on his computer and the evidence Corey was collecting in the field.

The team consequently adjusted its way of considering the data-processing artifacts in the investigative process: designing a database, or building a web application, had to be viewed as a "reporting process." Instead of seeing the database in itself as encapsulating a truth, the journalists considered it rather as a fragile construction that needed ongoing adjustments to get it to correspond to the "real-life evidence" collected on the ground. Kendall Taggart, who was hired in August 2010 as an intern and then as a "data reporter," described to us how challenging it had been to match the database to the qualitative information collected on the ground:

> Two weeks before it went up, one of the schools that we'd written about that was in Corey's story was, in the state database it was called like high school number 2. And then in real life it was called Southeast Middle. So when you went to the website for Southeast Middle it showed no problems and we'd just spent months writing a print story about some of the problems at that school and Corey saw that. You can see why from a data perspective someone who's not embedded and doesn't know that high school number 2 is Southeast Middle is going to miss it. It does happen a lot if you don't figure out ways to make sure that everything, including building an app, is a reporting process. (Taggart, interview, September 6, 2012)

Until October 2010, the journalists continuously adjusted the database to the "real-life evidence" that was collected on the ground. Such a shift entailed a substantial increase in the human resources dedicated to the investigation—starting with two journalists, the team ended up with 11 journalists. It also resulted in the definition of a new job assignment as "data reporters" were hired to check systematically a large part of all the data stored in the database.

The making of algorithms was also viewed as having to stick to the epistemological standards of an investigation. It implied the integration of ethical considerations into the algorithms, concerning what journalists considered as their professional duty towards their audience. As noted above, the team was particularly concerned about the faulty location of schools and the unfounded reactions this might cause. Agustin therefore decided to integrate a "buffer" into the algorithm, to increase the size of the zone taken into account by it:

> What I did is that once you stitch together a map of the seismic hazards in California, and you put the schools in proximity to those, can I lay a buffer, a half-mile buffer, around that point at which they say the school is at and see if there are features, you know, hazards that fall within it. (Armendariz, interview, September 12, 2012)

Throughout the process, the data-processing artifacts involved were adjusted to the established standards of an investigation. The idea that the database could encapsulate a hidden truth had not been envisioned at all in the investigative process. On the contrary, the team reacted by sticking more firmly to the established epistemological standards.

Emergence of Confidence in the Data

Adjusting the artifacts to the established epistemologies was a demanding process for the organization. The total cost has been estimated at $550,000, most of that in staff expenses (Doctor, 2011). But it ended up producing collective confidence in the data. In late September 2010, after almost a full year of collective work on the data, Agustin gave Corey a list of more than a hundred schools that had failed to comply with safety standards. And Corey found a good match between the data and the qualitative evidence collected on the ground:

> As Corey went out, down this target list I gave him, he wasn't able to disprove in real life any of the things that I was finding in the technology. So as he came back and as I understood that he was seeing the same things I was seeing, you know, the same things I was seeing in the data, he was seeing in the documents, he was seeing at the campuses, he was seeing in interviews. (Armendariz, interview, September 12, 2012)

From then on the team's confidence in the data never wavered. The team finally managed to produce a collective confidence in the artifacts, reducing the tensions that had arisen from the will to comply with the established epistemologies. This successful adjustment resulted from a long and costly process in which the building of a shared epistemic culture within the team was crucial.

Producing Justified Beliefs with Data-Processing Artifacts

The process of a journalistic investigation fundamentally consists in the collective production of justified beliefs about the world. Accordingly, news organizations, like scientific institutions (Knorr-Cetina 1999), rely on epistemic cultures that frame the valid ways of justifying a knowledge claim. In the last decade, new connections between the journalism community and the computer worlds have partly renewed these epistemic cultures in the U.S. (Parasie and Dagiral 2013). This has led to the emergence of another tension regarding the way data-processing artifacts support the making of a journalistic revelation. More specifically, the "On Shaky Ground" investigation provides evidence of two opposite ways of justifying beliefs with data-processing artifacts: a "hypothesis-driven approach" and a "data-driven approach."

Testing Hypotheses

Among the assertions at the core of the revelation, the following two have been inherently based on the use of data-processing artifacts:

1. State regulators have routinely failed to enforce California's landmark earthquake safety law for public schools.
2. Regulators have approved for jobs most of the inspectors accused of falsification and absence.

Throughout the investigation, the team's dominant approach was to formulate hypotheses first and subsequently to perform statistical processing in order to confirm or

discard these hypotheses. In this hypothesis-driven approach, statistical sampling appears to be a major tool—much more important than the completeness of the data.

One key operation was to evaluate the size of the regulatory crisis. As mentioned above, the lead reporter, Corey Johnson, began his investigation with a total of almost 9,000 schools that did not comply with safety standards. The hypothesis was therefore that many public schools failed to comply with legal standards. But the team quickly realized this number could not be taken for granted, given the poor quality of regulators' records. In particular, many of the school names were misspelled, were assigned to the wrong district, or did not even correspond to existing schools. In order to find a conservative number of unsafe schools, Agustin extracted a random sample of 370 schools so that the team's reporters could manually check which ones actually corresponded to an existing school. They found that 30 percent of the schools listed by the DSA could not be matched to official schools. They concluded that 6 schools out of 10 in California had at least one uncertified building project.

The claim that state regulators were guilty of supporting inefficient and/or dishonest inspectors relied on the same approach. From interviews, the team found that some inspectors did not show up during the buildings' construction, and that others agreed to certify projects that obviously could not comply with legal standards. The hypothesis was that state regulators had failed to control inspectors and may even have encouraged them to behave badly. In order to prove that claim, Agustin designed a second database concerning the inspectors' evaluation. The team obtained 17,000 inspector-rating forms corresponding to nearly 1,800 inspectors over 30 years, and the reporters entered the information into this database. Agustin extracted a list of 300 inspectors who had received poor ratings, and found that 66 percent of them were approved for additional jobs. Here again, making a hypothesis and performing statistical sampling appeared to be a decisive way to produce a justified belief.

The database of school seismic safety was thus occasionally used in the investigation to point out some "examples" of unsafe schools on which reporters could then investigate further. The processing of the data relied here on explicit hypotheses made by the journalist. The application developer involved in the project recalls many instances when reporters came to Agustin for such "examples":

> The reporters did come back to Agustin multiple times and be like, 'I need an example of X, can you find one?' And he did a lot of that. Once the data was in a structure, he was great for being able to find specific examples. Because we would, for example, hear that, 'Whether these schools that are on the AB300 list of school buildings that have the most structural problems, can you find me one of those that are really close to a fault?' And so that was the kind of thing that the dataset was great for. So he could then give her ten examples and she'd go check it out. (Michael Corey, interview, August 24, 2012)

Although prevalent in the "On Shaky Ground" project, this hypothesis-driven approach has been challenged by another approach.

Letting the Data Speak

The collective design of a comprehensive database of every public school in California also gave reporters the opportunity to distance themselves from a strictly hypothesis-driven view. In October 2010, after almost a year of checking the accuracy of the data, Agustin extracted a list of schools from the cross-tabulation of each school's safety status and its proximity to the fault. Here again, a hypothesis was made, but less to prove a precise claim than to organize the joint collection of facts. Kendall recalls how she used the "hit list" of unsafe schools to collect new evidence on the ground:

> (We were) using the data to identify what schools we thought were the worst, what schools clearly had safety problems, and really drilling down on what was happening there by talking to the structural engineers and the inspectors, all the people involved in that project. (Taggart, interview, September 6, 2012)

As a result, the database was used in the process as a means of disseminating the team's resources more efficiently. It allowed the reporters to collect more facts on the ground, and made it easier for them to cross-compare single cases of unsafe schools.

The reporters also relied on the database to gain more information from sources. As Corey recalls it, he showed the map of schools' seismic safety to geological experts in order to elicit additional information from them:

> Once we had a chance to talk to another earthquake engineer (…). She knew which areas were somewhat hot zones for geological fault activity. So by seeing our map, she was able to help us to better understand threats—threats to some of the schools in a way that we just didn't have the technical understanding and knowledge of. So in that way, the mapping has helped because other people were able to put other information on top of those maps and really bring the issue to life for us so we could help the public understand why this is a big deal. (Johnson, interview, August 31, 2012)

More critically, the database was used to monitor the inconsistent accounts of state regulators. As mentioned above, the reporters strongly suspected the Division of State Architects of manipulating their records. In order to check their claims about the crisis, the team decided to regularly add in their general database the updated datasets from state regulators. This regular update allowed the team to hold regulators more accountable for the situation:

> Often times for fact-checking the state's own claims about how many schools still had problems, having our own data that we'd gotten from them, every two months we were re-upping it, made it possible to know what kinds of changes they were making and whether or not they made sense. (Taggart, interview, September 6, 2012)

Although they relied mostly on a hypothesis-driven approach, other aspects of the investigation suggest that the data-processing artifacts were also used as means of organizing the collection of facts by reporters and obtaining supplementary and even unintentional accounts from sources. This approach strongly differs from the previous

one. Instead of formulating strong hypotheses that are subsequently proven or discarded on the basis of data processing, reporters instead expect from data processing the identification of new and unexpected stories.

Two Paths to Produce Justified Beliefs

Throughout the investigation, the production of justified beliefs with data-processing artifacts followed two opposite paths. From an analytical perspective, it would be impossible to assert that one path was more valid than the other; each of them provided specific norms to produce shared beliefs, relying on specific epistemological grounds.

[Insert Table 1 about here]

The first path has its roots in the computer-assisted reporting tradition in the U.S. (Parasie and Dagiral 2013). According to this model, a journalist becomes interested in a dataset only because he or she has some leads regarding a particular issue. From this perspective, data-processing artifacts allow reporters to confirm or invalidate a lead they may have found in interviews or documentation. Statistical sampling is viewed here as an appropriate tool to prove or disprove the lead. The completeness of the data as well as the possibility for the reporter to access granular data (e.g., the safety status of a school) is not a major concern. What is much more important here is the possibility to grasp social groups (e.g., inspectors) through data analysis, and to identify unfair situations or wrongdoings.

The second path has its roots in the new connections that have developed between the computer worlds and journalism since the mid-2000s (Lewis and Usher 2013). It values highly the completeness of the data as well as the possibility to access granular data. Taking the lead as a starting point for the data processing does not appear here to be compulsory for the reporter. The completeness of the data allows him or her to explore the data and be open to a new and unexpected story.

Within the team, the epistemological tension between the two paths did not give rise to a major controversy. The two paths were made compatible, but their coexistence fed the discussion about the possibility to explore the data by relying on lighter hypotheses. The journalists connected to computer innovators of the Bay Area, in particular, supported this claim against the rest of the team. As a “news applications developer” involved in the “data science community,” Michael Corey embodied that position:

> It’s funny sometimes because reporters or editors, when we tell them about a project, one of the first things they’ll say is, ‘Okay, well, what’s the lead?’ And we’re like, ‘Well, there’s not really a lead. That’s not really the point.’ And sort of the idea of data not as an end into itself, because I think it’s wrong. But there’s just not having to do the traditional news lead or there’s kind of a gotcha moment or there’s a problem or highlighting. It’s like, ‘No, we want to use the data as a vehicle to tell a story.’ (Corey, interview, August 24, 2012)

During the “On Shaky Ground” investigation, the team experienced two opposite ways of justifying beliefs with data-processing artifacts. On the one hand, this resulted in an epistemological tension between reporters with different backgrounds; on the other, these opposite ways were made compatible in the investigative process, as the same artifacts paradoxically helped to reduce the potential conflicts and to better organize collective work within the organization.

Conclusion

This article aims at contributing to the analysis of how technology affects the epistemologies of journalism. Because of the specificity of the case examined, the study does not bring to light a coherent epistemological framework that could apply to every investigative project involving a relation to data-processing artifacts, conducted in the U.S. or abroad. Instead, it points out the existence of epistemological tensions in the making of journalistic revelations through the processing of huge quantities of data. Because the “On Shaky Ground” project is the result of a collective organization consisting of journalists with different epistemological backgrounds, the tensions made visible here may be more broadly shared across other newsrooms.

This paper makes contributions to two streams of scholarship. The first deals with the future of investigative reporting and the role of technology therein. Noting the decline of investigative journalism in the U.S. in recent years, several scholars have emphasized the impact of the financial crisis and online technologies (Schudson 2010, Siles and Boczkowski 2012). Other scholars have nevertheless shown great expectations about how technology—especially data-processing artifacts—could facilitate investigation by lowering its cost (Cohen et al. 2011). The present study provides ambivalent arguments in this debate. On the one hand, it shows that the adjustment of artifacts to the established epistemologies of investigation has been a long and costly process. Mobilizing such resources may appear particularly problematic as most news organizations experience economic difficulties. But on the other hand, the analysis shows how one organization has succeeded in building the elements of a shared epistemic culture—which may reduce the cost of future projects. It thus suggests that data-processing artifacts can be used to enhance the collective organization of an investigation.

The second contribution of this paper concerns the study of journalistic knowledge. The notion of news as a form of knowledge is a well-established tradition in sociology, but analyzing the knowledge claims held by reporters is still rare in research (Ettema and Glasser 1998, Godler and Reich 2013), and usually poorly connected to technological matters. The present study suggests that it is crucial to study how journalists make such claims in relation to artifacts. Because of new connections with the computer worlds, news organizations experience alternative ways of producing justified beliefs from data (Parasie and Dagiral 2013).

Two limitations in this study should be addressed by further research. First, the organization that conducted the “On Shaky Ground” project is strongly committed to the established epistemologies of investigative reporting. Further research should study the making of investigative projects by organizations that are less respectful of established epistemologies. It could eventually point out distinct material and moral processes whereby they collectively make justified beliefs in relation to artifacts. Second, the case

analyzed here is U.S.-specific. Because news organizations based in other countries regularly conduct comparable projects, it seems crucial to understand whether they encounter similar tensions. The limited diffusion of a “computer-assisted reporting” tradition outside the U.S., however, may profoundly shape the way news organizations globally deal with the processing of vast amounts of data.

ACKNOWLEDGEMENTS

This article has greatly benefited from close readings by Eric Dagiral. I also thank the special issue editor, Seth Lewis, and two anonymous reviewers for their most helpful suggestions.

FUNDING

The French Ministry of Culture and Communication supported this work.

NOTES

1. This series reached 7 million people in three days (Rosenthal 2011).
2. Introduced in Marsh 2012, the bill was finally dropped in September 2012, officially for financial reasons.
3. This earthquake caused the death of 63 people in the San Francisco bay area on October 17th, 1989.

REFERENCES

Anderson, Chris. 2008. “The End of Theory: the Data Deluge Makes the Scientific Method Obsolete.” *Wired*, June 23. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Cohen, Sarah, James T. Hamilton and Fred Turner. 2011. “Computational Journalism.” *Communications of the ACM* 54 (10): 66–71.

Cox, Melisma. 2000. “The Development of Computer-Assisted Reporting.” Paper presented to the newspaper division, association for education in journalism and mass communication, southeast colloquium, University of North Carolina, Chapel Hill, NC, March 17–18.

Doctor, Ken. 2011. “The Newsonomics of a Single Investigative Story.” *Nieman Journalism Lab*, April 21. http://www.niemanlab.org/2011/04/the-newsonomics-of-a-single-investigative-story/

Ettema, James S. and Theodore L. Glasser. 1998. *Custodians of Conscience. Investigative Journalism and Public Virtue*. New York: Columbia University Press.

Ettema, James S. and Theodore L. Glasser. 2005. “On the Epistemology of Investigative Journalism.” In *Journalism: The Democratic Craft*, edited by Stuart G. Adam and Roy P. Clark, 126-140. New York: Oxford University Press.

Garrison, Bruce. 1998. *Computer-Assisted Reporting*. Mahwah, NJ: Lawrence Erlbaum Associates.

Godler, Yigal and Zvi Reich. 2013. “How Journalists Think about Facts. Theorizing the Social Conditions behind Epistemological Beliefs.” *Journalism Studies* 14 (1): 94–112.

Gray, Jonathan, Liliana Bounegru and Lucy Chambers. 2012. *The Data Journalism Handbook*. Sebastopol, CA: O’Reilly Media.

Hallin, Daniel C. 2005. *We Keep America on Top of the World: Television Journalism and the Public Sphere*. London: Routledge.

Houston, Brant, Len Bruzzese and Steve Weinberg. 2002. *The Investigative Reporter’s Handbook: A Guide to Documents, Databases and Techniques*. New York: Bedford/St. Martin’s Press.

Keating, Peter and Alberto Cambrosio. 2012. “Too Many Numbers: Microarrays in Clinical Cancer Research.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 37–51.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Latour, Bruno and Steve Woolgar. 1979. *Laboratory Life. The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Lewis, Seth C. 2011. “Journalism Innovation and Participation: An Analysis of the Knight News Challenge.” *International Journal of Communication* 5: 1623–1648.

Lewis, Seth C. and Nikki Usher. 2013. Open source and journalism: Toward new frameworks for imagining news innovation. *Media, Culture & Society*, 35(5), 602–619.

Mayer-Schönberger, Viktor and Kenneth Cukier. 2013. *Big Data: A Revolution that Will Transform How we Live, Work and Think*. Boston and New York: Houghton Mifflin Harcourt.

Parasie, Sylvain and Éric Dagiral. 2013. “Data-driven Journalism and the Public Good. Computer-Assisted Reporters and Programmer-Journalists in Chicago.” *New Media and Society* 15 (6): 853–871.

Rosenthal, Robert J. 2011. “Reinventing Journalism. An Unexpected Personal Journey from Journalist to Publisher.” *California Watch*, October 4. http://californiawatch.org/project/reinventing-journalism

Schudson, Michael. 1978. *Discovering the News. A Social History of American Newspapers*. New York: Basic Books.

Schudson, Michael. 2001. “The Objectivity Norm in American Journalism.” *Journalism* 2 (2): 149–170.

Schudson, Michael. 2010. “News in crisis in the United States: Panic—and Beyond.” In *The Changing Business of Journalism and its Implications for Democracy*, edited by David A.L. Levy and Rasmus K. Nielsen, 95–106. Oxford: Reuters Institute for the Study of Journalism.

Siles, Ignacio and Pablo J. Boczkowski. 2012. “Making Sense of the Newspaper Crisis: A Critical Assessment of Existing Research and an Agenda for Future Work.” *New Media and Society* 14 (8): 1375–1394.

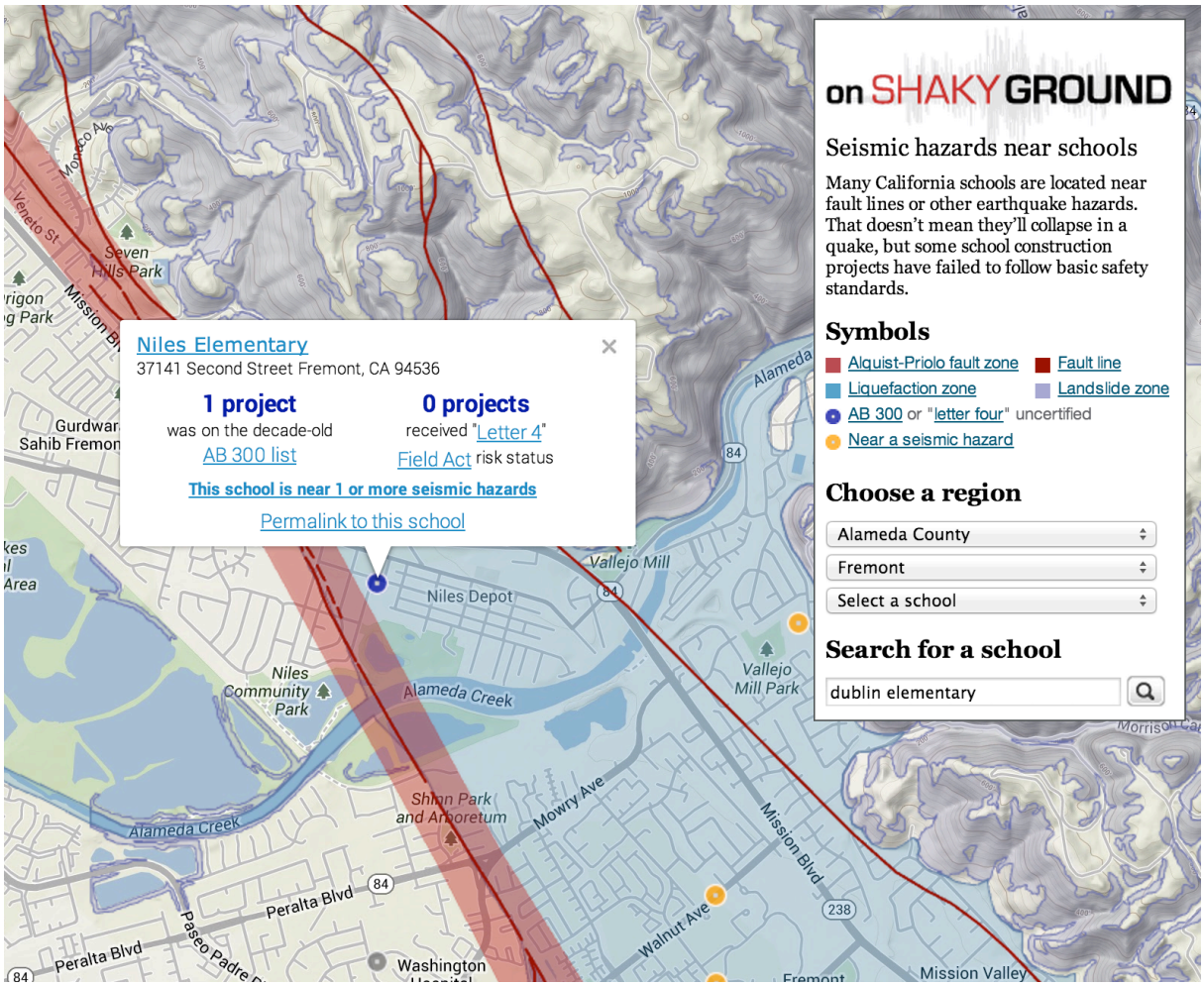
Strasser, Bruno. 2012. “Data-Driven Sciences: From Wonder Cabinets to Electronic Databases.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 85–87.

Tuchman, Gaye. 2012. “Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity.” *American Journal of Sociology* 77 (4): 660–679.

Author contact details

Sylvain Parasie, University of Paris-Est Marne-la-Vallée, LATTS, France.
sylvain.parasie@univ-paris-est.fr

Figure 1. Interactive map of seismic hazards near Californian schools



Note: The picture shows the Niles Elementary School, in Fremont, California. This school is located near two seismic hazards and has one building project that is listed on a 2002 inventory of school buildings with potentially dangerous seismic hazards.

Table 1. Two paths to produce justified beliefs

	The hypothesis-driven path	**The data-driven path**
Database's most valued feature	Its availability for sampling	Its completeness
Starting assumptions	A strong hypothesis	A light hypothesis
What makes the data interesting for reporters	It allows reporters to confirm or invalidate a lead	It allows journalists to explore the information, and eventually to find new leads
Level of investigation	Aggregates and social groups	Granular or incident-level
Connection to the collection of qualitative information	Qualitative information helps to formulate the hypothesis, and to illustrate the related claim	The processing of the data helps to collect supplementary information
Ways of making governments accountable	By identifying unfair situations or faulty actions from data sampling	By identifying inconsistencies in the tracking of government actions over time