



Making Artificial Intelligence Work for Investigative Journalism

Jonathan Stray

To cite this article: Jonathan Stray (2019) Making Artificial Intelligence Work for Investigative Journalism, Digital Journalism, 7:8, 1076-1097, DOI: [10.1080/21670811.2019.1630289](https://doi.org/10.1080/21670811.2019.1630289)

To link to this article: <https://doi.org/10.1080/21670811.2019.1630289>



Published online: 02 Jul 2019.



Submit your article to this journal [↗](#)



Article views: 6018



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 37 View citing articles [↗](#)



Making Artificial Intelligence Work for Investigative Journalism

Jonathan Stray 

Graduate School of Journalism, Columbia University, New York, NY, USA

ABSTRACT

Many have envisioned the use of AI methods to find hidden patterns of public interest in large volumes of data, greatly reducing the cost of investigative journalism. But so far only a few investigative stories have utilized AI methods, in relatively narrow ways. This paper surveys what has been accomplished in investigative reporting using AI techniques, why it has been difficult to apply more advanced methods, and what sorts of investigative journalism problems might be solved by AI in the near term. Journalism problems are often unique to a particular story, which means that training data is not readily available and the cost of complex models cannot be amortized over multiple projects. Much of the data relevant to a story is not publicly accessible but in the hands of governments and private entities, often requiring collection, negotiation, or purchase. Journalistic inference requires very high accuracy, or extensive manual checking, to avoid the risk of libel. The factors that make some set of facts “newsworthy” are deeply sociopolitical and therefore difficult to encode computationally. The biggest near-term potential for AI in investigative journalism lies in data preparation tasks, such as data extraction from diverse documents and probabilistic cross-database record linkage.

KEYWORDS

Artificial intelligence; investigative journalism; computational journalism; algorithmic news; machine learning; natural-language processing; data cleaning

Introduction

Investigative journalism may be one of the most effective ways to discourage corruption and reveal wrongdoing across society, and recent analyses suggest that it is one of the most cost-effective as well (Hamilton 2016). If machine intelligence were applied to investigative journalism, it might monitor global feeds for important news, find socially relevant patterns among diverse data sets, and maybe even write up the resulting stories (Hansen et al. 2017; Marconi and Siegman 2017). However, AI is not currently widely used in investigative journalism, despite its supposed promise.

This paper contributes to the study of journalism and automation by unpacking, investigating, and re-developing the common assumption that AI can increase the power and reach of investigative journalists. Typically, the idea is that AI will reduce the cost of investigative journalism production by replacing certain types of tedious or

expensive human labor with cheap computation. While this is a real possibility, there are several key roadblocks, and the most significant potential applications of AI in investigative work may not look much like previous speculation.

We start by discussing the handful of uses of AI techniques in investigative work to date. An analysis of these cases and others suggests some fundamental reasons why it is hard to successfully apply AI in an investigative journalism context. However, there are some very interesting possibilities which seem like they could be tackled in the next few years with a program of applied research. Data ingestion and cleanup, which are often glossed over in AI research, consume a great deal of journalists' time and are good candidates for automation.

The intersection of investigative journalism and artificial intelligence is a small part of the intersection between computation and journalism generally.

The classic textbook *The Elements of Journalism* says that investigative journalism “puts emphasis on the role of the press as activist, reformer, and exposé” (Kovach and Rosenstiel 2014, 169). Hamilton’s (2016) study of the economics of the practice says that “investigative reporting involves original work, about substantive issues, that someone wants to keep secret” (10). Contemporary investigative reporting also frequently involves working with large volumes of public records and data, which is a natural opening for automation.

Artificial intelligence is an active field of computer science research, and also a wide set of engineering practices (Russell, Norvig, and Davis 2010). Some branches of AI, such as algorithms for playing chess, do not have obvious applications in journalism. Here, we are especially interested in those methods which might “lower the costs of discovering watchdog stories” (Hamilton 2016). In practice, these methods will be “narrow AI” and not “general AI” (Broussard 2018).

These definitions usefully exclude other types of journalism automation. For example, breaking news is not typically investigative reporting because there is no time for in-depth research, while journalistic data visualization is not typically an application of artificial intelligence because such work does not employ the computational methods developed by AI researchers. This paper also does not consider AI methods built into widely applicable tools. For example, email spam filtering and automated grammar checkers are used in every industry. Instead, *our focus is the application of AI theory and methods to problems that are unique to investigative reporting, or at least unsolved elsewhere.*

I start by reviewing related work, and collecting the various hopes that have been expressed for AI in journalism. Then I survey stories where AI methods were used successfully. These stories are not as numerous or as sophisticated as the hopes, so I examine why investigative journalism is a hard problem for AI, including detailed examples of unsolved problems. There are a variety of technical, legal, political, and philosophical challenges to build better AI for investigative journalism. One key interdisciplinary question is the algorithmic description of what counts as news—that is, what should we design our story-finding AI to find? Despite fundamental challenges, there remains great promise for AI in investigative journalism. I end by suggesting several ways that near-future AI could be applied productively, by helping with data preparation and cleaning.

Investigative AI in the Context of Journalism Automation

The subject of this paper is AI applied to investigative reporting, that is, story *production* as opposed to story distribution or promotion. AI methods are now commonly applied to the other areas of news work but are still relatively rare in story production. Conversely, automation is increasingly common in investigative work, but most of this would not be considered AI.

Many news organizations use machine learning techniques to solve a variety of business problems, including predicting the popularity or “virality” of stories in order to decide what to promote, modeling user behavior to increase subscriptions and minimize churn, and so on (Stone 2014; Prakash 2017). Machine learning-based news personalization systems are widely used by news publishers such as *The New York Times* (Spangher 2015) and news aggregation apps such as *Google News* (Das 2007).

Conversely, news articles are widely used as test data sets in AI research for problems such as named entity recognition, topic modeling (Newman et al. 2006), recommendation, and summarization (Paulus, Xiong, and Socher 2018). These techniques are relevant to investigative journalism tasks, but the AI models created in this line of research are trained on the *output* of journalists. This is unlikely to yield good results for journalism applications as the source material used in reporting is substantially more diverse and messy than most NLP training sets (Stray 2016a).

Automated story production techniques have come into wide use in the last few years, with major newsrooms, including the AP, Reuters, and Forbes producing thousands of stories a month based on structured data feeds of corporate earnings and sports scores (LeCompte 2015; Marconi and Siegman 2017). The process is akin to filling out a form, with some conditional elements to select from a finite set of sentences based on data values (e.g., “the home team emerged victorious” vs. “it was a sorry loss for the home team.”) While automated story production fundamentally challenges conceptions of the roles of humans and machines in journalism (Lewis, Guzman, and Schmidt 2019), automating the writing of investigative stories seems as if it would require artificial general intelligence, so we should not expect it soon.

The “computational journalism” literature gets closest to discussing the use of AI in story production. This relatively new term has been used in a variety of ways (Coddington 2015), including the use of computational techniques to find stories in data, and conversely the journalistic investigation of the properties of algorithms used by government and industry (Diakopoulos 2016). Both might be accelerated by AI. The 2011 definition of Cohen, Hamilton, and Turner is most relevant here:

Stories will emerge from stacks of financial disclosure forms, court records, legislative hearings, officials’ calendars or meeting notes, and regulators’ email messages that no one today has time or money to mine. With a suite of reporting tools, a journalist will be able to scan, transcribe, analyze, and visualize the patterns in these documents. (Cohen, Hamilton, and Turner 2011)

Computational methods are today routine in journalism, if unevenly distributed (Berret and Phillips 2016). There are now widely used journalism-specific tools for analyzing unstructured documents (such as DocumentCloud (Mor and Reich 2018), Overview (Brehmer et al. 2014; Stray 2016a), Tadam (Plattner, Orel, and Steiner 2016), and Tabula (Aristarán et al. 2013)) and data wrangling (such as CSVkit (Christopher et al.

2018), Open Refine (openrefine.org n.d.), and Dedupe.io (DataMade 2016)). Pioneering organizations such as the International Consortium of Investigative Journalists (ICIJ) and the Organized Crime and Corruption Reporting Project (OCCRP) are fusing diverse data sets in graph databases to facilitate network analysis (Cabra 2016; Stray 2017).

For the most part, current computational journalism efforts would not be considered “artificial intelligence” in the sense that they do not use AI methods. Admittedly, this distinction can be fuzzy. One practical question—and a key economic consideration—is whether or not these applications require the services of a developer trained in contemporary AI technology. So far, that has rarely been the case.

Hamilton and others have suggested cost-effectiveness as a core rationale for applying AI (Hamilton 2016; Cohen, Hamilton, and Turner 2011) and we explore this consideration below. But this does not specify what AI should be doing. Broussard articulates a remarkable role for AI in investigative journalism: to analyze data for differences between what is and what ought to be (Broussard 2015), an idea to which we will return.

The Assumption of AI Advantage

We start from the idea that AI will prove transformative for investigative journalism, which is widely held in both industry and academia. This is the core assumption that this paper explores. A report from Columbia Journalism School concludes that “AI tools can help journalists tell new kinds of stories that were previously too resource-impractical or technically out of reach” (Hansen et al. 2017). An Associated Press report says AI will “empower the creation of entirely new types of journalism” (Marconi and Siegman 2017).

In such discussions, AI is typically described as being able to “identify a pattern” (Hansen et al. 2017), “uncover social problems” (Broussard 2015), “tell the stories hidden in the data” (Holmes 2016), or otherwise illuminate previously unknown connections. This is exciting, but vague. Without the grounding of story case studies, it will be difficult to define the function of such pattern detection systems.

The other major claim is that AI will speed up investigative work. An example comes from the ICIJ:

The ICIJ didn’t utilize any AI technology during the investigative process [on the Panama Papers], but Matthew Caruana Galizia, the organization’s web applications developer, wishes they did.

“We were dealing with a vast amount of documents, and ICIJ just didn’t have the resources to investigate them all,” Galizia said. “But by using artificial intelligence, we would have been able to make that process much faster for all the journalists involved and end up with the same result.” (Marconi and Siegman 2017)

What type of “AI” is useful here, and which part of the investigative workflow will it accelerate, or what new types of stories will be possible? There are few concrete examples. In part, these questions are difficult to answer because there are surprisingly few descriptions of data-driven investigative journalism processes, that is, what investigative journalists actually do with data in the course of their work. Although journalists often discuss the methods they used to complete an individual story,

systematic summaries of investigative data practice are rare. There are detailed process descriptions of document mining in (Stray 2016a) and network analysis in (Stray 2017).

What Investigative AI Looks Like Now

The current uses of AI in investigative journalism are modest. To an AI researcher they may even seem trivial. Even so, these examples are important lessons in what journalists actually do, and may point the way to more ambitious applications.

Previous successful uses of AI in journalism fall into a few broad categories: document classification, language analysis, data cleaning, lead generation, and breaking news detection. Not all of these examples are investigative, but all have potential investigative applications. The stories produced in these ways might not have been possible without AI techniques, typically because they would have required too much manual labor. The seven stories and one system discussed in this section include many of the examples discussed within the data journalism community.

There are perhaps another dozen instances that might also be considered AI used for investigative reporting in (Stray 2016a; Stray 2017). The most comprehensive work on journalism automation lists about two dozen examples (Diakopoulos 2019) including most of those discussed here. This same small set of examples is repeatedly discussed at data journalism conferences such as NICAR (Shorey et al. 2018). The high overlap between sources suggests that there are a relatively small number of cases in total; no doubt there are others, but certainly not an order of magnitude more. In other words, AI methods are not yet commonly used for investigative reporting, and I will explore the reasons why below.

Document Classification

The most common use of machine learning in investigative journalism so far is supervised document classification. For the story “License to Betray” the Atlanta Journal Constitution scraped over 100,000 doctor disciplinary records from every state, looking for instances where doctors who had sexually abused patients were allowed to continue to practice (Teegardin et al. 2016). Logistic regression reduced the likely cases to 6000 documents, which they then read and coded manually (Stray 2016a).

The Los Angeles Times story “LAPD underreported serious assaults, skewing crime stats for 8 years” (Poston, Rubin, and Pesce 2015) was based on comparing the narrative descriptions of more than 400,000 incident reports with the category assigned by police, e.g. “aggravated assault.” They found that there was a systematic misclassification of assaults as less serious than they actually were, according to LAPD’s own definitions. Fortunately the reporters had manually reviewed one year’s worth of data for a previous story, providing a training set of over 20,000 incidents (Stray 2016a).

Language Analysis

Some stories have relied on NLP techniques such as topic modeling, clustering, word embeddings, sentiment analysis, etc.

In their 2014 story “The Echo Chamber,” Reuters reporters showed how a small group of elite lawyers have argued most of the cases before the US Supreme Court (Biskupic, Roberts, and Shiffman 2014). The reporters also broke down the number of accepted cases by type, for example, whether filed by a business, individual, or government agency. They accomplished this mostly by hiring 20 freelancers to read 10,300 petitions over a period of three months, but were able to gain some additional information through LDA topic modeling (Stray 2016a).

For the 2013 story “DHHS downplayed food stamp issues” (Dukes 2013), a WRAL-TV reporter used Overview (Brehmer et al. 2014) to automatically cluster 4500 pages of state government emails obtained through a Freedom of Information Act request. One large cluster corresponded to messages posted to an inter-county government email list. The reporter manually read this cluster and found messages showing the government officials knew that a web browser compatibility problem was causing delays, ultimately affecting 70,000 people (Dukes 2014).

Sentiment analysis has been used by journalists on social media data as a proxy for public opinion, but investigative journalism use is rare. For the Washington Post story “Whistleblowers say USAID’s IG removed critical details from public reports” (Higham and Rich 2014) reporters compared 12 draft reports with their final versions. Using sentiment analysis, they found that more than 400 negative references were removed before publication (Stray 2016a).

Monitoring for Breaking News

The advent of global public social media such as Twitter seems to offer enormous opportunity to find previously obscure news. It has taken some time and effort to successfully exploit this data stream for journalism. While monitoring for breaking news is not usually an investigative journalism application, this is one of the only production examples of more general “story finding” AI.

The Reuters Tracer system continuously ingests Twitter data, filters out spam and tweets which are not about events, then clusters tweets by event and ranks them for review by journalists (Liu et al. 2016; Stray 2016b). The system employs a number of trained models for tweet classification, clustering, and newsworthiness ranking, as shown in Figure 1. Out of a sample of 31 news events, in 27 cases Tracer found a corresponding tweet cluster faster than Reuters journalists were able to issue an alert using traditional reporting methods, often by a half hour or more.

Lead Generation

Several authors have noted that AI could be especially useful for journalistic lead generation, generating lines of inquiry rather than definitive conclusions (Hansen et al. 2017; Diakopoulos 2019; Shorey et al. 2018). Human involvement also avoids the potential accuracy and relevance problems of directly publishing automated output.

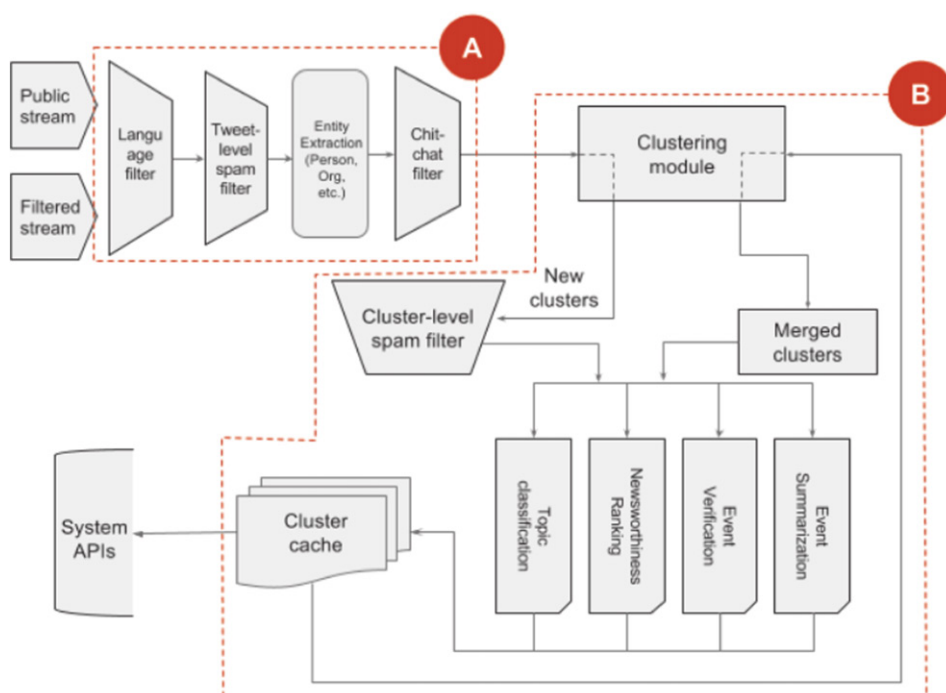
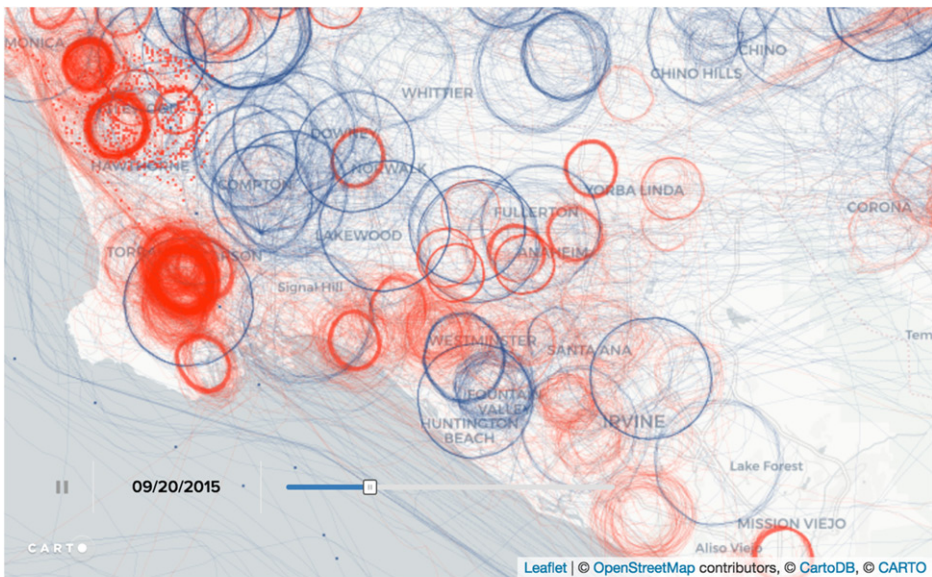


Figure 1. Machine learning system architecture for Reuters Tracer: (A) tweet processing module; (B) event detection module. Many of the stages in this diagram involve custom built and trained models. From Liu et al. (2016).

The *New York Times'* *Who The Hill* was designed to recognize the faces of US members of congress (Shorey et al. 2018) in images uploaded by readers. It was more of a curiosity than a serious reporting tool, but it contributed to at least one story when congresswoman Claudia Tenney was identified in a photograph taken at a fundraising party (Vogel and Shorey 2018).

Buzzfeed's story on US government surveillance planes is one of the most intriguing uses of machine learning in journalism. Reporters knew from previous stories that law enforcement would sometimes circle over major cities to take surveillance footage or capture cell phone signals (Aldhous and Seife 2016) (Figure 2). By training on the flight paths of known law enforcement planes, using features describing the flight path bounding box and flight speed and direction histograms, BuzzFeed was able to identify many planes for further investigation (Aldhous 2017).

This seems like a great success for machine learning on an important story. However, machine learning was not really necessary. In independent work, Bastien identified most of the same planes by ranking flights according to a simple metric designed to detect circling: the percentage of points on the flight path within 10 miles of the centroid (Bastien 2017). As several practitioners have pointed out, there are often simpler alternatives to machine learning (Shorey et al. 2018). In any case, the hard work of this story is not finding the planes, but the subsequent time-consuming investigation of who owns them, what they are doing, and whether it is legal and ethical.



■ FBI ■ DHS

PETER ALDHOUS / BUZZFEED NEWS

Figure 2. US government surveillance plane flight paths identified by machine learning (Aldhous and Seife 2016).

Why AI for Investigative Journalism Is Hard

So far, AI methods in investigative journalism have produced useful but modest results. There are perhaps a perhaps a few dozen examples to date, and there is a wide gap between these cases and the ambitious visions quoted above. AI has barely touched investigative journalism, let alone transformed it.

This failure could be an issue of technology diffusion or inadequate investment. Or perhaps the problem is simply difficult. There are a number of domain-specific issues that make it challenging to apply AI techniques to investigative journalism. Some of these appear to be fundamental, and unlikely to be solved quickly.

Data Availability

“Public data,” meaning data that is legally required to be accessible to citizens, is not necessarily publicly available. Often it must be requested, negotiated, scraped, or purchased. Certain national corporate registries, especially in tax havens such as Cyprus and Hong Kong, require company registration records to be purchased one at a time, meaning that it is impractically expensive to acquire the complete “public” data set. Surprisingly often, public records are not even digital. Fully a third of the document sets in a recent survey of document-driven investigative journalism projects arrived on physical paper (Stray 2016a).

The bad news is that journalists spend an enormous amount of time gathering data from a variety of sources. The good news is that, armed with suitable metadata

and integrations with existing data provider search systems, an AI assistant could propose scraping or purchasing the records required to answer a query, or help file and track public records requests as MuckRock already does (MuckRock 2018). Human sources will remain out of reach of automated methods for the foreseeable future; despite recent dramatic advances in conversational systems, such as Google's Duplex AI which can make simple phone calls to book services (Leviathan and Matias 2018), it will be a long time before machines can talk to people in an investigative context.

Journalism AI research is also hindered by the lack of standard training data sets. While "general" AI efforts such as question answering and document summarization may prove themselves valuable to journalists, investigative reporters also face some unique and uniquely complex data tasks. Creating specialized training and evaluation data repositories may be an important first step in producing AI research that leads to useful journalism applications. ProPublica's Free The Files project (ProPublica 2012), discussed below, is one labeled investigative journalism corpus that could be packaged and promoted as a research data set.

Unique Stories

When AI is used in a commercial setting there is typically an ongoing business problem to be solved. Transactional data such as clicks and purchases arrives in continuous streams, and a model trained on this data can be used as long as the underlying stream is stable.

By contrast, many data-driven investigative stories are never repeated. The Atlanta Journal-Constitution's model for finding documents describing sexual abuse by doctors will never be useful again, because there is not another backlog of 100,000 reports to examine. In such cases the cost of building a custom AI solution cannot be amortized over multiple stories.

I am aware of only one set of experiments on the time/cost/accuracy of machine learning vs. human information extraction in a journalism context, which suggests that the break-even point is on the order of a few hundred documents (Giorgi 2015). For document classification tasks, the domain of legal e-discovery is perhaps most similar to investigative journalism, and one vendor addresses the problem of fixed costs by noting that machine learning (called "predictive coding" in this domain) "has been successfully leveraged in cases with only a few hundred or thousand documents" (Robinson 2018). A survey of machine-assisted document-driven investigative journalism projects gives a median document set size of 4000 documents (Stray 2016a). This lower limit of hundreds to thousands of documents suggests that many document sets in journalism are simply too small to benefit from AI methods.

Challenging Problems

As part of an investigation of Donald Trump's business deals, students at Columbia Journalism School examined New York City real estate public records pertaining to several Trump properties. These records are available from the city's ACRIS database, covering a variety of contracts including mortgages and liens between dozens of

MODIFICATION AGREEMENT

THIS MODIFICATION AGREEMENT (this "Agreement"), dated as of the 23rd day of June, 2006, is made by BAYROCK/SAPIR ORGANIZATION LLC (formerly known as Bayrock/Zar Spring LLC), a Delaware limited liability company ("Borrower"), having its principal office c/o Bayrock Group L.L.C., Trump Tower, 725 Fifth Avenue, 24th Floor, New York, New York 10022, to FORTRESS CREDIT OPPORTUNITIES I LP, a Delaware limited partnership, having an address at 1345 Avenue of the Americas, 46th Floor, New York, New York 10105, as Agent on behalf of the Lenders set forth in the Loan Agreement (as hereinafter defined) (together with its successors and assigns, "Agent").

WITNESSETH:

WHEREAS, Borrower is the owner of the real property commonly known as 246 Spring Street located in the City of New York, County of New York and State of New York, such ownership interest being comprised of a fee simple interest in the Property described in Exhibit A attached hereto and made a part hereof (the "Property");

WHEREAS, Agent, on behalf of the Lenders, is the present owner and holder of the promissory note described on Schedule 1 attached hereto and made a part hereof (the "Existing Note"), which Existing Note evidences an indebtedness of Borrower to Agent, on behalf of the Lenders, in the outstanding principal amount of \$77,127,169.49;

Figure 3. The beginning of a document describing a modification to one of the loans used to finance the Trump Soho hotel (New York City ACRIIS document 2006083000784001).

parties over more than a decade. **Figure 3** is an example document concerning the Trump Soho hotel (later renamed The Dominick). The investigative questions are

- Who are the parties that currently own the building?
- What was the equity and outstanding debt of each party at each point in time?
- Who did they owe these debts to?

This can only be determined by painstakingly reading and reconstructing the series of documents filed for this property, which range from standardized forms to complex natural-language contracts. **Figure 4** shows part of a hand-built spreadsheet of the transactions contained in these documents, used by the reporters to answer these questions. The task is deterministic in the sense that there is a definite answer, though that answer may involve descriptions of financial relationships that fall outside of the simple categories in the above questions.

This is a multi-document comprehension problem that is well beyond the current state of the art of AI. Progress is likely to be slow: training data is scarce, expensive to produce, and unlikely to generalize. There are perhaps ten thousand New York City real estate developments of this size and complexity (New York City Department of Finance 2017), each one would require a dozen or so hours to generate a spreadsheet like the one above, and even a complete data set would not be large enough for current deep learning approaches. By comparison, the data sets used for much simpler "question answering" or "reading comprehension" AI research are orders of magnitude larger. The Stanford Question Answering Dataset includes 130,000 examples (Rajpurkar, Jia, and Liang 2018) and the CNN/Daily Mail training data set is over a million examples (Chen, Bolton, and Manning 2016).

Recorded / Filed	Document Type	Page s	Party1	Party2	Doc Amount
5/3/06 16:44	MORTGAGE AND CONSOLIDATION	14	BAYROCK/SAPIR ORGANIZATION LLC	FORTRESS CREDIT OPPORTUNITIES I LP	74,298,931.00
5/3/06 16:44	UCC3 AMENDMENT	14	BAYROCK/SAPIR ORGANIZATION LLC	FORTRESS CREDIT OPPORTUNITIES I LP	-
5/3/06 16:44	UCC3 AMENDMENT	13	BAYROCK/SAPIR ORGANIZATION LLC	FORTRESS CREDIT OPPORTUNITIES I LP	-
6/13/06 15:39	MISCELLANEOUS	1	246 SPRING STREET, LLC		-
8/23/06 11:43	MISCELLANEOUS	29	BAYROCK/SAPIR ORGANIZATION LLC		-
9/5/06 13:57	AGREEMENT	10	BAYROCK/SAPIR ORGANIZATION LLC	FORTRESS CREDIT OPPORTUNITIES I LP	9,884,807.00
5/3/07 11:00	ZONING LOT DESCRIPTION	12	BAYROCK/SAPIR ORGANIZATION LLC		-
9/24/07 14:19	MORTGAGE	33	BAYROCK/SPAPIR ORGANIZATIONLLC,	ISTAR FINANCIAL	87,000,000.00
9/24/07 14:19	ASSIGNMENT OF LEASES AND RENTS	13	BAYROCK/SPAPIR ORGANIZATIONLLC,725	ISTAR FINANCIAL	-
9/24/07 14:19	ASSIGNMENT, MORTGAGE	9	BAYROCK/SPAPIR ORGANIZATIONLLC,725	ISTAR FINANCIAL	-
9/24/07 14:19	UCC3 TERMINATION	16	BAYROCK/ZAR SPRING LLC C/O BAYROCK GROUP L.L.C.	FORTRESS CREDIT CORP.	-
9/24/07 14:19	MORTGAGE	31	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	28,237,515.00
9/24/07 14:19	ASSIGNMENT OF LEASES AND RENTS	14	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	-
9/24/07 14:19	INITIAL UCC1	10	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	-
9/24/07 14:19	MORTGAGE	31	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	9,762,485.00
9/24/07 14:19	ASSIGNMENT OF LEASES AND RENTS	13	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	-
9/24/07 14:19	INITIAL UCC1	10	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	-
9/24/07 14:19	INITIAL UCC1	10	BAYROCK/SAPIR ORGANIZATION LLC,	ISTAR FINANCIAL INC,	-
9/28/07 16:58	ADDITIONAL MORTGAGE TAX	1	BAYROCK/ SPAPIR ORGANIZATION LLC		2,816,263.00

Figure 4. Excerpt of the hand-built chronological list of New York City real estate public records concerning the Trump Soho hotel. Color coding indicates documents on the same date (Giannina Segnini/Columbia Journalism School).

The Need for Accuracy

Imagine a news organization which uses AI to examine public records to find suspected money laundering. Inaccurately suggesting that someone is involved in criminal activity is not only a serious violation of journalistic ethics, but it can also lead to an expensive libel lawsuit—even if the other 99% of inferences are correct. Although this has yet to be tested in court, it seems likely that US publishers will be liable for algorithmic errors: “news organizations should be concerned about liability for libelous automated journalism content affecting private plaintiffs, who can recover by proving the negligence on the part of the news organization” (Lewis, Sanders, and Carmody 2019).

It is unlikely that any AI system used in investigative journalism will reach 100% accuracy, in part because of the usual sources of error in AI systems (variance, generalization error, etc.) but more fundamentally because the available data is usually ambiguous. For example, there is no algorithm to determine whether the same name in two different databases actually refers to the same person or not. This requires more data, for example, the person’s address, but even then errors are possible: there could have been two people with the same name living at the same address at different times, or “Jr.” and “Sr.” suffices could be missing, or it could simply be an error in the data. Only manual research—perhaps a phone call to the landlord—can ultimately resolve such questions.

Thus, AI-generated results cannot be directly published if an incorrect result might injure someone’s reputation. This is not an issue when AI is used to rank items for human follow-up, as in BuzzFeed’s identification of potential surveillance flights. But if algorithmic results are to be published, they may first need to be individually checked by hand, in which case the computational advantages of scale and speed may be lost.

Despite errors, assaults drop

The Los Angeles Police Department misclassified an estimated 14,000 serious assaults from 2005 to 2012. Even with the errors factored in, serious assaults and violent crime showed a decline.

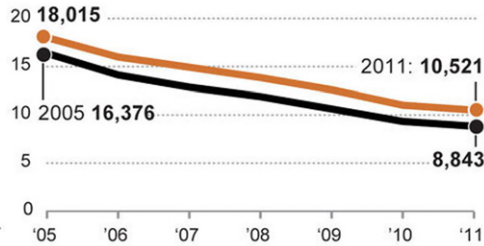
Figures in thousands

Adjusted aggravated assaults

Official aggravated assaults

Note: 2012 figures were excluded from the chart because they did not capture a full year of crime reporting.

Sources: Los Angeles Police Department; Times analysis.
Graphics reporting by Ben Poston



@latimesgraphics

Figure 5. LA Times' analysis of crime reports, showing that 14,000 assaults were recorded as less serious than narrative incident descriptions show. To compensate for the 24% error rate of their classifier, the reporters adjusted the totals to produce conservative estimates (Postin and Pesce 2015).

An alternative is to publish only aggregate information, which can sometimes be corrected for algorithmic uncertainty—though quantification of that uncertainty usually requires human review. For the LAPD crime misclassification story, the reporters reviewed a random sample of 2400 machine-labeled incidents and discovered that the classifier error rate was a hefty 24%. Rather than attempting to improve the classifier, they simply adjusted their yearly totals of misclassified crimes to produce the conservative estimates shown in Figure 5 (Postin and Pesce 2015).

Cost-Effectiveness

If there is an underlying thread to the problems so far, it is the issue of cost-effectiveness. AI may be able to help journalists find and produce stories that would otherwise be impossible; if it does not, it must help journalists do their work faster and cheaper. As Cohen Hamilton, and Turner (2011) point out, talking to human sources is often just as efficient as data analysis.

Journalists often collect records to address a specific question, which, when answered, marks the end of the analysis and the beginning of the story. This suggests a strict limit on the time and money invested in any document or data; it must be more effective or newsworthy than the alternative path of asking whistle-blowers or partisan insiders for the material. (Cohen, Hamilton, and Turner 2011)

The issue of time is multiplied by the relative market rates of different types of work. According to US salary data from job site Glassdoor, the average “reporter” salary is around \$50,000 while the average “artificial intelligence engineer” is closer to \$150,000. This constrains the amount of time that can be shifted from reporting to engineering, if automation is to increase efficiency in terms of stories per dollar. It also means that AI talent developed in the newsroom is in danger of leaving for better paid jobs.

Moreover, most of the data analyses performed in contemporary journalism can be done with a spreadsheet. AI will only be cost-effective for a subset of stories where:

- Data is a substantial and important source of information for the story.
- There is a data subtask which is at least partially automatable.
- Straightforward (non-AI) computational techniques are not sufficient.
- It would be faster and/or cheaper to apply an AI method instead of doing it manually.
- There is no good alternative, such as talking to a domain expert or inside source.

Today this is a rather small set of stories. Yet AI may still have enormous impact, if it can truly help find stories that a human alone would miss. Even one story can have an outsized impact. Money laundering investigations by the Sarajevo-based OCCRP have uncovered over 5 billion dollars stolen from public funds in Eastern European countries (OCCRP 2018). If an AI can marshal evidence that a human reporter missed, a single story might benefit thousands or millions of people.

What Is News?

Perhaps, the most complex challenge in AI-assisted investigative story production is the technical systematization of the concept of “news.” The description of “news values” by sociologist Stuart Hall seems as fresh today as when it was written in 1973:

News values are one of the most opaque structures of meaning in modern society. All ‘true journalists’ are supposed to possess it; few can or are willing to identify and define it. Journalists speak of ‘the news’ as if events selected themselves. Further, they speak as if which is the ‘most significant’ news story, or which ‘news angles’ are most salient are divinely inspired. Yet of the millions of events which occur every day in the world, only a tiny portion ever become visible as ‘potential news stories,’ and of this proportion, only a small fraction are actually produced as the day’s news in the news media. We appear to be dealing, then, with a ‘deep structure’ whose function as a selective device is un-transparent even to those who professionally most know how to operate it. (Hall 1973)

Of course, “news values” are not completely opaque even if they are hard for journalists to articulate, and decades of research have attempted to learn from journalists and their stories what counts as news. A recent review (Harcup and O’Neill 2017) suggests over a dozen criteria, such as *the power elite*, *conflict*, *surprise*, *magnitude*, *shareability*, *bad news*, and *celebrity*. Investigative journalism may or may not encode the same set of values as news generally, but it certainly uses *some* set of values to decide what is worth reporting. Embedding these values into code—teaching a computer to identify the fact patterns that constitute a “story”—is poorly explored. It is a major technical, political, and ethical challenge.

One approach is to design algorithmic definitions of newsworthiness from first principles. The Los Angeles Times’ earthquake reporting bot used data from the USGS Earthquake Notification Service to “automatically generate short reports on earthquakes above the ‘newsworthy’ threshold of a 3.0 magnitude” (LeCompte 2015). Others have used more sophisticated methods to interpret incoming data. The Marple system monitors Swedish crime data for potential stories and flags anomalous data

points. It models the average number of crimes per month from historical data, but there must still be a newsworthiness threshold. In this case, the researchers chose a statistical significance threshold (p value) of 0.0001, meaning that only data points with less than a 1-in-10,000 chance of being generated by the historical model will be flagged. The researchers found that this struck a good balance between missing “obvious” peaks and overwhelming the reporters with notifications—an approach they described as “ad-hoc” (Magnusson, Finnäs, and Wallentin 2016).

How should one decide on these thresholds? How to elicit them from reporters? Newsroom automation practitioners frequently describe this as a problem. For the Associated Press’ automated story production efforts,

Translating even the simplest data means converting the loose guidelines a human reporter might follow into concrete rules a computer can follow. For example, a human reporter might have a general idea of when a company’s performance was very different from analyst expectations, based on their knowledge of the industry. But for the algorithm, the AP had to specify exact ranges for which the spread between actual earnings and expectations is considered large or small. (LeCompte 2015)

The development of the Reuters Tracer system encountered similar obstacles:

Newsroom standards are rarely formal enough to turn into code. ... ‘The interesting exercise when you start moving to machines is you have to start codifying this,’ says Chua. ‘Much like trying to program ethics for self-driving cars, it’s an exercise in turning implicit judgments into clear instructions.’ (Stray 2016b)

Instead of trying to come up with explicit rules for newsworthiness, some researchers and practitioners have used human journalists’ output as training data. To train their Tracer system to decide whether an event is newsworthy, Reuters engineers created a set of 300 clusters of tweets around specific events, 63 of which were identified as newsworthy by journalists. And to evaluate the recall of the system, they collected all major news events over a period of one week as reported by Reuters, AP, and CNN (Liu et al. 2016).

Asking journalists to label training data or evaluate automated output avoids the problem of articulating explicit rules for newsworthiness. It also replicates any biases in existing reporting. For example, it is well established that crime reporting is biased. Metropolitan newspapers in the US report somewhere between 30% and 70% of the homicides in their city. Generally, a crime is more likely to be covered if the victim is young, female, white, and/or rich, or if the killing is particularly gruesome or involves multiple victims or sex. This produces a distorted picture of crime in the public imagination. Also, the focus on individual incidents as opposed to trends may explain why the majority of Americans believe that violent crime is increasing when it has been decreasing for decades in most cities (Stray 2012).

The codification of newsworthiness provides a unique opportunity to reflect on what investigative journalists cover and what they should cover. Rather than simply replicate what newsrooms do now, journalism AI researchers could entirely re-imagine reporting. However, this re-imagining will run into constraints. One team discovered that finding breaking news on social media requires monitoring active accounts with large audiences, and this means attending to “men in the media” even if one might wish to highlight other, less heard voices (Thurman et al. 2016). Investigative AI

designers will be forced to think deeply about both the goals and the practicalities of story detection.

Near-Term Promise for Investigative AI: Data Wrangling

So far, we have looked at the challenges to the grand visions of AI in investigative journalism. There are also problems that could be solved with near-future AI techniques. The biggest opportunity is speeding up data preparation and cleaning.

Data “wrangling” and cleaning makes up a large fraction of the time spent most data projects, with surveys showing numbers between 30% and 80%, yet it is not a particularly well-studied research topic (Kandel et al. 2011; Furche et al. 2016; Press 2016). The problem is particularly acute for investigative journalism because of the huge variety of different source document and data formats, even for the same type of data. The next sections give two real-world examples of data preparation tasks that AI could help with: data extraction from documents and cross-database record linkage.

Data Extraction from Heterogeneous Documents

While every TV station in the United States must disclose political ad sales, there is no requirement on format or standardization. This leads to a dizzying array of different form types, nearly as many as there are TV stations, three of which are reproduced in Figure 6. Standard OCR cannot cope with all these layouts, the need for high accuracy, and the required standardization and merging. The problem is so resistant to automation that for the 2012 election, ProPublica enlisted its readers in a crowd-sourced data transcription effort called Free The Files, which eventually manually entered data for about 17,000 of the 43,000 disclosure documents they obtained (ProPublica 2012).

There are deep learning methods to extract structured data from richly formatted documents (Wu et al. 2018) which could, in principle, be applied to document sets relevant to journalism. Services such as Amazon’s Textract are starting to offer similar capabilities commercially. It is not clear how AI trained on one document domain—such as corporate ownership records, financial disclosures, or court filings—would generalize to the others, which makes this a challenging research problem.

Record Linkage

Fusing databases has long been a basic investigative journalism technique. One early example is a 1985 story in which reporters cross-referenced a list of school bus drivers with a list of felons to find a disturbing amount of overlap. The resulting story led to policy changes, and 65 drivers had their licenses revoked (DeFleur 2013).

Record linkage is the process of determining that two records refer to the same entity, typically a person or company. When a database must be linked to itself—as in the case of identifying unique donors in campaign finance data—this is also known as deduplication. Because names are not unique, record linkage depends on the existence of other fields such as addresses, but even then it is often not possible to match

Contract Agreement Between: **WLWT**
1700 Young Street
Cincinnati, OH 45202
(513)412-5000
www.wlwt.com

Print Date 08/09/12 Page 1 of 7

CONTRACT

Contract / Revision 940867 /		Alt Order #
Product GENERAL		
Contract Dates 10/08/12 - 11/06/12		Estimate # 2452
Advertiser Mandel/R/Senator		Original Date / Revision 08/09/12 / 08/09/12
Billing Cycle EOM/EOC	Billing Calendar Broadcast	Cash/Trade Cash
Station WLWT	Account Executive Bob Sommerkamp	Sales Office Cincinnati
Special Handling		
Demographic Adults 25-54		
IDB#	Advertiser Code	Product Code
Agency Ref	Advertiser Ref	

And: Strategic Media Placement OH
7669 Stagers Loop
Delaware, OH 43015

*Line	Ch	Start Date	End Date	Description	Start/End Time	Days	Length	Spots/Week	Rate	Type	Spots	Amount
N 1	WLWT	10/08/12	11/06/12	5-6a news	5-6a		:30			NM	20	\$2,500.00
Class of Time - Immediately Pre-emptible without notice												
		Start Date	End Date	Weekdays				Spots/Week	Rate			
		Week: 10/08/12	10/14/12	MTWTF--				5	\$125.00			
		Week: 10/15/12	10/21/12	MTWTF--				5	\$125.00			
		Week: 10/22/12	10/28/12	MTWTF--				5	\$125.00			
		Week: 10/29/12	11/04/12	MTWTF--				5	\$125.00			
		Week: 11/05/12	11/11/12	MT-----				2	\$125.00			

--- CONTRACT COMMENT ---
*****BROADCAST INFORMATION***** PARAGRAPHS 49 AND 50 OF THE UNITED STATES FEDERAL COMMUNICATIONS COMMISSION'S REPORT AND ORDER NO. 07-217 PROVIDES THAT BROADCAST STATIONS' ADVERTISING CONTRACTS WILL NOT DISCRIMINATE ON THE BASIS OF RACE OR ETHNICITY, AND MUST CONTAIN NONDISCRIMINATION CLAUSES. COMPLIANCE WITH THIS ORDER, RACE COMMUNICATIONS, INC., IS INCLUDING ANY SUBSIDIARY OR DIVISION OF RACE DOES NOT DISCRIMINATE IN ANY BROADCAST ADVERTISING CONTRACT ON THE BASIS OF RACE OR ETHNICITY AND EVALUATES, NEGOTIATES AND COMPLETES ITS BROADCAST ADVERTISING CONTRACTS WITHOUT REGARD TO RACE OR ETHNICITY.

--- REMARKS ---
REVISED ORDER. SOME PORTION OF SPOTS CONVERTED TO 60S
TOTAL ADVERTISING UNCHANGED. PLS CNF

LT	Ln	Day	Time	Program	Len	Rate	Starts	Spots/Week	# of Weeks	Total Spots	Total Cost	Daypart
	*7	TH	7A-8A	60 MIN C2897	:30	\$250.00	10/25-10/25	1	1	1	\$250.00	
		Sales Remark: V3 - V3 Comment Changed, Spots/week Changed										
	*8	F	7A-8A	60 MIN C2897	:30	\$250.00	10/26-10/26	1	1	1	\$250.00	
		Sales Remark: V3 - V3 Comment Changed, Spots/week Changed										
	*13	TH	1230P-1P	THE	:30	\$60.00	10/25-10/25	0	0	0	\$0.00	
		Sales Remark: V3 1X CK V3 More than 2 codes changed										
	*14	F	1P-130P	AMF P3M-VIDEOS	:30	\$500.00	10/26-10/26	1	1	1	\$500.00	
		Sales Remark: V3 CHANGED TO 60 V3 More than 2 codes changed										
	*15	F	1P-130P	AMF P3M-VIDEOS	:30	\$200.00	10/29-10/29	0	0	0	\$0.00	
		Sales Remark: V3 1X CK V3 Comment Changed, Spots/week Changed										

Contract # **321083**

Schedule Dates 10/08/12-10/12/12
Advertiser PDP McCrory/Gov/NC (10811)
Agency Smart Media Group (1345)
Product Political - State Candidate (1071)
Brand CIRC14 GOV (134465)
Salesperson Telerep/Philadelphia, Philadelphia (1057)
Sales Office Telerep/Philadelphia
Buyer Name HANSEN,LINDSAY
Phone/Fax
CPE MCCO/ORDA/CIRC14
Account Types National/Political
Billing Type Standard
Comments MCCO/ORDA FOR GOVERNOR
RECEIVED FOR NC GOVERNOR

Date Entered 07/06/12
Last Modified 10/04/12
Entered By JH Hopton
CO-OP No
Headline # 06221195
Demo
Order Type Normal
Package Deal
Commission % 15.00
Commission \$ 72,117.50
Net Total \$40,332.50
Sales Tax

Charlotte (WSOC)
By Broadcast/Spots
Oct. 2012 \$4 \$47,450.00
Grand Total \$4 \$47,450.00

Smart Media Group
814 King Street
Suite 400
Alexandria, VA 22314

Line	Line Type / Break Type (Ref #)	Dates	Sec	Length	Run Times	SPW	Mo	Tu	We	Th	Fr	Sa	Su	Spots	Rate	Total	Status	Comments	Entered
1.1	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.2	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.3	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.4	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.5	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.6	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.7	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.8	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.9	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.10	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.11	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.12	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.13	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.14	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.15	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.16	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.17	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.18	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.19	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12
1.20	Normal Line / Spot	10/12/12-10/12/12	5	:30	5A-6A (EST)	1								1	\$375.00	\$375.00	Charter (WSOC)	DR PHIL	10/12

CONFIRMATION CONTRACT

Accepted-Agency/Advertiser: _____ Date: _____
Accepted-Station: _____ Date: _____
Comments: _____

WSOC-TV does not accept advertising contracts that impermissibly discriminate on the basis of race or ethnicity. This non-discrimination provision is a condition of each advertising contract with this station/affiliate and is written.

Figure 6. Three different political ad buy disclosures from the 2012 US election, from ProPublica's Free The Files (ProPublica 2012).

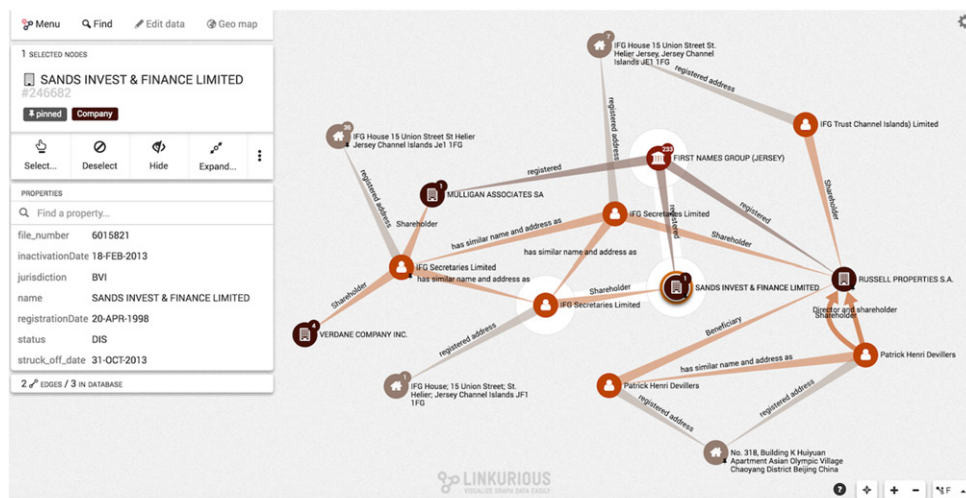


Figure 7. A subset of the Panama papers structured data, in a graph created by a reporter. Note the duplicate records from different databases, and also machine-generated linkages such as “has similar name and address as” (reproduced from Cabra 2016).

names with 100% certainty. Probabilistic record deduplication is already in use in journalism (DataMade 2016).

Recently, there have been a series of increasingly ambitious data fusion projects carried out by organizations such as ICIJ and OCCRP. The prototypical example is the Panama Papers. The structured data sets (which comprised only one part of the total leak) were loaded into the Neo4j graph database, then entities with similar names and addresses were given a “soft linkage” by adding edges, as shown in Figure 7. Journalists reported on the data by graphically exploring the networks around specific people and companies of interest (Stray 2017).

Automated linkage judgments must still be validated manually before publishing, because a crucial link which forms the basis of a public accusation of wrongdoing cannot rest on the vagaries of a particular model. Previous work proposes a hybrid model where the computer links records automatically and shows merged entities to the user, which can be expanded as needed to evaluate the underlying linkages (Stray 2017). One possible system, including an un-merged graph data store, is shown in Figure 8.

Conclusion

This paper has unpacked the idea that AI can be useful in investigative journalism, proceeding in three parts: case studies demonstrating how AI has been used to date, an analysis of the challenges that have prevented wider use, and a proposed near-term research focus on data wrangling, which is immediately useful and sidesteps many of the greatest difficulties.

Previous discussions of AI in investigative journalism have often dodged the details of exactly how it would be used. Generally, authors have imagined that AI could be used for

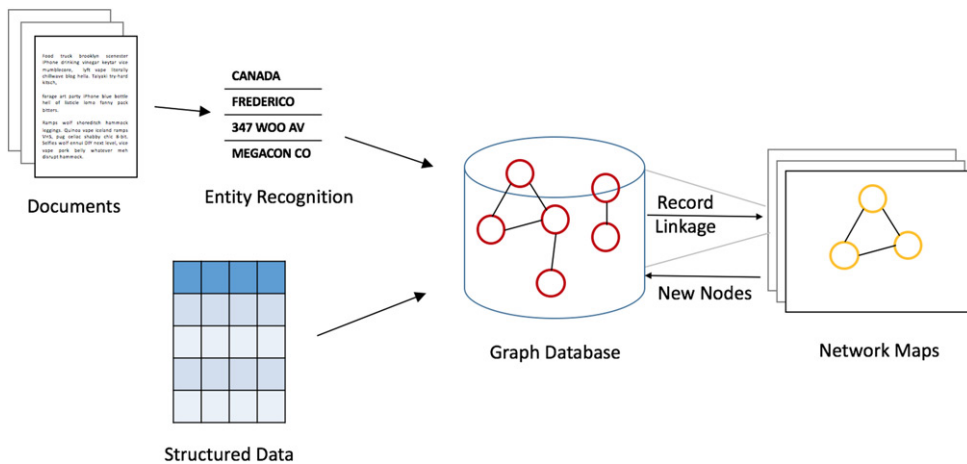


Figure 8. A proposed system for network-based investigative journalism, including AI-assisted record linkage. From Stray (2017).

“pattern recognition” within large data sets of public interest, greatly reducing the human effort required to produce investigative stories. There are several reasons why this will be difficult:

- Data access is a continual problem. Even public records frequently need to be requested, scraped, purchased, or negotiated.
- Investigative projects are often unique or “one off,” so the development costs of AI models cannot be amortized across stories.
- Investigative journalism problems are often well beyond the state of the art of current methods, such as complex multi-document summarization tasks.
- Professional ethics and libel law necessitate essentially perfect accuracy in any published inferences. This usually means AI output requires human checking.

Human labor is always an option, so these can all be understood as issues of cost-effectiveness. AI engineers are also substantially more expensive than reporters, which further constrains the economics. Even for tasks such as document classification, where existing AI methods are effective, problem set-up costs currently favor manual work for smaller data sets, perhaps up to a few thousand documents.

AI could also help find stories that humans would miss, either because it would be too expensive to have reporters read all of the relevant data or because the required pattern recognition is a cognitive task better suited to machines. Unfortunately, specifying which sorts of fact patterns constitute a “story” is an extremely challenging problem. It is difficult to translate notions of “newsworthiness” into code. The alternative is to have machine systems learn newsworthiness from human examples, but this will replicate any existing biases in coverage. Should investigative story-finding algorithms come into wide use, we should expect that they will be subjects of social and political controversy, as news recommendation algorithms already are.

There is at least one area where AI methods are likely to benefit investigative journalism in the near term: data cleaning and “wrangling.” This work typically consumes a substantial fraction of the time required to produce a data-driven investigative story, yet the required operations tend to be simpler and less open-ended than other investigative tasks. The primary source records that journalists must rely on are maddeningly diverse and messy, which makes data extraction and probabilistic record linkage promising targets for AI automation.

Beyond that, if the above challenges should prove surmountable, we are faced with the question of what, ultimately, investigative journalism AI should be used to accomplish. At the highest level of abstraction, “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)” (Broussard 2015). In principle, AI could be used to evaluate both the *is* and the *ought*. This is an enormously complex task. Determining what *is* is the core, hard task of investigative reporting. What *should* be is an even more complex question. Law and policy only capture part of what is right, people disagree for good reasons, and the answers are unavoidably political. Ultimately, investigative journalism AI requires not just vast quantities of public records and deep contextual understanding, but opinions on right and wrong.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Jonathan Stray  <http://orcid.org/0000-0003-4467-1239>

References

- Aldhous, Peter. 2017. “How BuzzFeed News Revealed Hidden Spy Planes in US Airspace.” *Columbia Journalism Review*. Accessed August 7 2017. https://www.buzzfeed.com/peteraldhous/hidden-spy-planes?utm_term=.oyox5J8g6#.nik4pQLYZ.
- Aldhous, Peter, and Charles Seife. 2016. “Spies in the Skies.” *Buzzfeed*. Accessed April 6 2016. https://www.buzzfeed.com/peteraldhous/spies-in-the-skies?utm_term=.ae7VDwyKL#.rhyKAN7bD.
- Aristarán, Manuel, Mike Tigas, Jeremy B. Merrill, and Jason Das. 2013. “Tabula: Extract Tables from PDFs.” 2013. <http://tabula.technology/>.
- Bastien, Laurent. 2017. “Finding Every Government Surveillance Flight.” 2017. https://github.com/laurentbastien/spyplanes/blob/master/laurent_bastien_DATA_SAMPLE.pdf.
- Berret, Charles, and Cheryl Phillips. 2016. “Teaching Data and Computational Journalism.” https://journalism.columbia.edu/system/files/content/teaching_data_and_computational_journalism.pdf.
- Biskupic, Joan, Janet Roberts, and John Shiffman. 2014. “The Echo Chamber.” *Reuters*. Accessed December 8 2014. <https://www.reuters.com/investigates/special-report/scotus/>.
- Brehmer, Matthew, Stephen Ingram, Jonathan Stray, and Tamara Munzner. 2014. “Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists.” *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 2271–2280.
- Broussard, Meredith. 2015. “Artificial Intelligence for Investigative Reporting.” *Digital Journalism* 3 (6): 814–831.

- Broussard, Meredith. 2018. *Artificial Unintelligence : How Computers Misunderstand the World*. Cambridge: MIT Press.
- Cabra, Mar. 2016. "How the ICIJ Used Neo4j to Unravel the Panama Papers." *Neo4j Blog*. 2016. <https://neo4j.com/blog/icij-neo4j-unravel-panama-papers/>.
- Chen, Danqi, Jason Bolton, and Christopher D. Manning. 2016. "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task." arXiv preprint arXiv:1606.02858.
- Christopher Groskopf 2018. "Csvkit 1.0.3." 2018. <https://csvkit.readthedocs.io/en/1.0.3/>.
- Coddington, Mark. 2015. "Clarifying Journalism's Quantitative Turn." *Digital Journalism* 3 (3): 331–348.
- Cohen, Sarah, James T. Hamilton, and Fred Turner. 2011. "Computational Journalism." *Communications of the ACM* 54 (10): 66.
- Das, A. S. M. Datar, A. Garg, and S. Rajaram. (2007, May). "Google news personalization: scalable online collaborative filtering" In Proceedings of the 16th international conference on World Wide Web, 271–280. ACM.
- DataMade. 2016. "Introducing Dedupe.io." 2016. <https://datamade.us/blog/introducing-dedupeio>.
- DeFleur, Margaret H. 2013. *Computer-Assisted Investigative Reporting: Development and Methodology*. Abingdon: Routledge.
- Diakopoulos, Nicholas. 2016. "Accountability in Algorithmic Decision Making." *Communications of the ACM* 59 (2): 56–62.
- Diakopoulos, Nicholas. 2019. *Automating the News*. Cambridge, MA: Harvard University Press.
- Dukes, Tyler. 2013. "Records: DHHS Downplayed Food Stamp Issues." *WRAL*. Accessed December 9 2013. <http://www.wral.com/records-dhhs-downplayed-food-stamp-glitches/13173174/>.
- Dukes, Tyler. 2014. "Human-Assisted Reporting Gets the Story." Source: An OpenNews Project. 2014. <https://source.opennews.org/articles/human-assisted-reporting/>.
- Furche, Tim, Georg Gottlob, Leonid Libkin, Giorgio Orsi, and Norman W. Paton. 2016. "Data Wrangling for Big Data: Challenges and Opportunities." Paper presented at Proceedings of the 19th International Conference on Extending Database Technology (EDBT), Bordeaux, France.
- Giorgi, Ariana. 2015. "An Analysis of Methods for Information Retrieval." 2015. https://github.com/arianagiorgi/masters-proj/blob/master/Giorgi_MP2015.pdf.
- Hall, Stuart. 1973. "The Determinations of News Photographs." In *The Manufacture of News: Social Problems, Deviance and the Mass Media*, edited by Stanley Cohen and Jock Young London, 226–247. Constable.
- Hamilton, James. 2016. *Democracy's Detectives: The Economics of Investigative Journalism*, Cambridge, MA: Harvard University Press.
- Hansen, Mark, Meritxell Roca-Sales, Jon Keegan, and George King. 2017. "Artificial Intelligence: Practice and Implications for Journalism." Tow Center for Digital Journalism, Columbia Journalism School, New York. <https://towcenter.org/research/artificial-intelligence-practice-and-implications-for-journalism/>.
- Harcup, Tony, and Deirdre O'Neill. 2017. "What Is News?" *Journalism Studies* 18 (12): 1470–1488.
- Higham, Scott, and Steven Rich. 2014. "Whistleblowers Say USAID's IG Removed Critical Details from Public Reports." *The Washington Post*. Accessed October 22 2014. https://www.washingtonpost.com/investigations/whistleblowers-say-usaids-ig-removed-critical-details-from-public-reports/2014/10/22/68fbc1a0-4031-11e4-b03f-de718edeb92f_story.html?utm_term=.df962e16a5d3.
- Holmes, Jonathan. 2016. "AI Is Already Making Inroads into Journalism but Could It Win a Pulitzer?" *The Guardian*. Accessed April 3 2016. <https://www.theguardian.com/media/2016/apr/03/artificial-intelligence-robot-reporter-pulitzer-prize>.
- Kandel, Sean, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank Van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. "Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data." *Information Visualization* 10 (4): 271–288.
- Kovach, Bill, and Tom Rosenstiel. 2014. *The Elements of Journalism*. 3rd ed. New York, NY: Three Rivers Press.

- LeCompte, Celeste. 2015. "Automation in the Newsroom." *Nieman Reports*. Accessed September 2015. <http://niemanreports.org/articles/automation-in-the-newsroom/>.
- Leviathan, Yaniv, and Yossi Matias. 2018. "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone." *Google AI Blog*. 2018. <https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html>.
- Lewis, Seth C., Amy Kristin Sanders, and Casey Carmody. 2019. "Libel by Algorithm? Automated Journalism and the Threat of Legal Liability." *Journalism & Mass Communication Quarterly*. 96 (1): 60–81
- Lewis, Seth C., Andrea L. Guzman, and Thomas R. Schmidt. 2019. "Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News." *Digital Journalism* 1–19.
- Liu, Xiaomo, Ramdev Wudali, Robert Martin, John Duprey, Arun Vachher, William Keenan, Sameena Shah, et al. 2016. "Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter." Paper presented at Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16, 207–216. New York, NY: ACM Press.
- Magnusson, Måns, Jens Finnäs, and Leonard Wallentin. 2016. "Finding the News Lead in the Data Haystack: Automated Local Data Journalism Using Crime Data." Paper presented at *Computation + Journalism Symposium*. Stanford University, Palo Alto, CA [http://journalism.stanford.edu/cj2016/files/Finding the news lead in the data haystack.pdf](http://journalism.stanford.edu/cj2016/files/Finding_the_news_lead_in_the_data_haystack.pdf)
- Marconi, Francesco, and Alex Siegman. 2017. "The Future of Augmented Journalism: A Guide for Newsrooms in the Age of Smart Machines." New York, NY: Associated Press.
- Mor, Niv, and Zvi Reich. 2018. "From "Trust Me" To "Show Me" Journalism: Can Document Cloud Help to Restore the Deteriorating Credibility of News?" *Journalism Practice*, 12 (9). 1091–1108.
- MuckRock. 2018. "About MuckRock." 2018. <https://www.muckrock.com/about/>.
- New York City Department of Finance. 2017. "Annual Property Tax Report." <http://www1.nyc.gov/site/finance/taxes/property-reports/property-reports-annual-property-tax.page>.
- Newman, David, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. "Analyzing Entities and Topics in News Articles Using Statistical Topic Models." In *International conference on intelligence and security informatics*, 93–104. Berlin, Heidelberg: Springer.
- OCCRP. 2018. "About Us." Organized Crime and Corruption Reporting Project. 2018. <https://www.occrp.org/en/about-us>.
- openrefine.org. n.d. "OpenRefine." Accessed April 8 2018. <http://openrefine.org/>.
- Paulus, Romain, Caiming Xiong, and Richard Socher. 2018. "A Deep Reinforced Model for Abstractive Summarization." Paper presented at Sixth International Conference on Learning Representations, Vancouver, Canada. <https://arxiv.org/pdf/1705.04304.pdf>.
- Plattner, Titus, Didier Orel, and Olivier Steiner. 2016. "Flexible Data Scraping, Multi-Language Indexing, Entity Extraction and Taxonomies: Tadam, a Swiss Tool to Deal with Huge Amounts of Unstructured Data." Paper presented at *Computation + Journalism Symposium*, Palo Alto, CA: Stanford University.
- Postin, Ben, and Anthony Pesce. 2015. "How We Reported This Story." *Los Angeles Times*. 2015. <http://www.latimes.com/local/cityhall/la-me-crime-stats-side-20151015-story.html>.
- Poston, Ben, Joel Rubin, and Anthony Pesce. 2015. "LAPD Underreported Serious Assaults, Skewing Crime Stats for 8 Years." *Los Angeles Times*. Accessed October 15 2015. <http://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html>.
- Prakash, Shailesh. 2017. "Journalism and Technology: Big Data, Personalization, Automation." Paper presented at *Computation + Journalism Symposium*. Evanston: Northwestern University. <https://www.youtube.com/watch?v=PqMvx089AQ4>.
- Press, Gil. 2016. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says." *Forbes*. 2016. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#38052d456f63>.
- ProPublica. 2012. "Free the Files." ProPublica. 2012. <https://projects.propublica.org/free-the-files/>.
- Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. "Know What You Don't Know: Unanswerable Questions for SQuAD." arXiv preprint arXiv:1806.03822.

- Robinson, Eric. 2018. "The Ultimate Predictive Coding Handbook." *The Ediscovery Blog*. 2018. <http://www.theeddiscoveryblog.com/2018/01/03/the-ultimate-handbook-for-mastering-predictive-coding/>.
- Russell, Stuart Jonathan, Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence : A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.
- Shorey, Rachel, Anthony Pesce, Chase Davis, and Peter Aldhous. 2018. "Getting Started with Machine Learning for Reporting." IRE NICAR. Chicago. 2018. <http://paldhous.github.io/NICAR/2018/machine-learning.html>.
- Spangher, Alexander. 2015. "Building the Next New York Times Recommendation Engine." *The New York Times*. Accessed August 11 2015. <https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/>.
- Stone, Martha L. 2014. "Big Data for Media." [https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Big Data For Media_0.pdf](https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Big%20Data%20For%20Media_0.pdf).
- Stray, Jonathan. 2012. "Beyond the Crime Scene: We Need New and Better Models for Crime Reporting » Nieman Journalism Lab." *Nieman Journalism Lab*. <http://www.niemanlab.org/2012/06/new-and-better-models-for-crime-reporting/>.
- Stray, Jonathan. 2016a. "What Do Journalists Do with Documents? Field Notes for Natural Language Processing Researchers." Paper presented at Computation + Journalism Symposium. Palo Alto, CA: Stanford University. [https://journalism.stanford.edu/cj2016/files/What do journalists do with documents.pdf](https://journalism.stanford.edu/cj2016/files/What%20do%20journalists%20do%20with%20documents.pdf).
- Stray, Jonathan. 2016b. "The Age of the Cyborg." *Columbia Journalism Review*. Accessed September 2016.
- Stray, Jonathan. 2017. "Network Analysis in Journalism: Practices and Possibilities." Paper presented at Data Science + Journalism Workshop. Halifax: ACM SIGKDD. https://drive.google.com/file/d/0B8CcT_0LwJ8QMzFjTWxLSFVkJVTg/view.
- Teegardin, Carrie, Danny Robbins, Jeff Ernsthansen, and Ariel Hart. 2016. "License to Betray." *Atlanta-Journal Constitution*. Accessed July 5 2016. http://doctors.ajc.com/doctors_sex_abuse/.
- Thurman, Neil, Steve Schifferes, Richard Fletcher, Nic Newman, Stephen Hunt, and Aljosha Karim Schapals. 2016. "Giving Computers a Nose for News: Exploring the Limits of Story Detection and Verification." *Digital Journalism* 4 (7): 838–848.
- Vogel, Kenneth P., and Rachel Shorey. 2018. "Trump Groups Raised Millions, Then Paid It Out to Loyalists and a Trump Hotel – The New York Times." *The New York Times*. Accessed January 24 2018. <https://www.nytimes.com/2018/01/24/us/politics/pro-trump-fundraising-trump-hotel.html>.
- Wu, Sen, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas, Philip Levis Schapals, and Christopher Ré. 2018. "Fonduer: Knowledge base construction from richly formatted data" In *Proceedings of the 2018 International Conference on Management of Data*, Houston, 1301–1316. ACM.