# Generating Location-Based News Leads for National Politics Reporting

**Nicholas Diakopoulos**
Northwestern University

**Madison Dong**
Northwestern University

**Leonard Bronner**
The Washington Post

**Jeremy Bowers**
The Washington Post

**Abstract**

Computational news discovery refers to the use of algorithms to orient editorial attention to potentially newsworthy events or information prior to publication. In this paper we describe the design, development, and initial evaluation of a computational news discovery tool, called Lead Locator, which is geared towards supplementing national politics reporting by suggesting potentially interesting locations to report on. Based on massive amounts of data from a national voter file, Lead Locator ranks counties based on statistical properties such as their extremity in the distribution of a variable of interest (e.g. voter turnout) as well as their political relevance in terms of shifts in voting patterns. It then presents an automatically generated tip sheet of potentially interesting locations that reporters can interactively browse and search to help inform their reporting ideas.

**Keywords**

algorithmic newsworthiness, computational news discovery, lead discovery, politics reporting, text generation

## 1 Introduction

Some of the earliest writing about computational journalism emphasizes the potential for computing to alert reporters to interesting trends, anomalies, or newsworthy entities in data [9]. The promise is that computing can increase the efficiency and scale at which new news stories can be identified [8]. Such editorial orientation algorithms have now been developed to support a variety of journalistic use cases, including social media monitoring, fact spotting, and document screening for investigations, among others [4]. Here we refer to such applications of computing as *computational news discovery* (CND), which we define as: the use of algorithms to orient editorial attention to potentially newsworthy events or information prior to publication. The focus of this paper is on describing the design and development of one such tool—Lead Locator—at The Washington Post.

Lead Locator is aimed at helping national politics reporters identify geographic areas that are of potential interest for reporting on the electorate in the 2020 U.S. elections. Initial interviews with reporters suggested a need for helping to find places and slices of the electorate where reporters might go to better understand the people in that place. To support this use-case Lead Locator uses data mining to analyzes a national voter file tracking every registered voter in the U.S. in order to rank counties based on their potential newsworthiness to reporters. We then use natural language generation to automatically produce tip sheets that are presented to reporters in an interactive interface that conveys why each county may be of interest to their reporting. This use case is a promising candidate for demonstrating the value of computational news discovery in journalism because of the sheer scale of data [18]. Leads which might otherwise be impossible to find could develop into unique and original contributions that differentiate coverage, and engineering effort can be amortized over an extended use period.

The central question motivating Lead Locator is: How can we orient politics reporters by suggesting interesting places to go to talk to voters? We were also interested in questions of how to apply data mining in ways that translate traditional notions of newsworthiness into algorithms. Additionally, we wanted to investigate how to present leads to reporters in ways which capture their attention and get them interested in following up on the lead. The premise is not that the system would automatically generate publishable news reports, but rather that it would function as a stimulus to help reporters ideate, hypothesize, and ultimately do additional reporting work to develop these auto-generated leads into fully fleshed out news reports. In the next section we outline in more detail the previous work on computational news discovery. Then, we describe the Lead Locator system in detail including the domain, data, algorithmic newsworthiness, tip sheet text generation, and user-interface design. We conclude the paper with a discussion of initial user feedback and plans for further evaluation and future work.

## 2 Computational News Discovery

A number of systems have demonstrated how computing can contribute to finding interesting news leads for journalists. For instance, the City Beat system was deployed as an ambient newsroom display that alerted journalists to potentially newsworthy events based on social media postings in New York City [16]. The system's evaluation identified several challenges, such as the difficulty of translating newsworthiness into algorithms, and the variability with which different newsrooms wanted to define it. More recently, Reuters developed the Tracer system to monitor Twitter, detect breaking news events such as floods and earthquakes, and present

*C+J '20, March 20–21, 2020, Boston, MA*
.

those events to journalists in an interactive interface [12]. The system sped up Reuters' news alerts in many cases, and demonstrated progress in computationally operationalizing newsworthiness factors such as topicality, scale of event, impact (e.g. human, physical, or financial), location, and rarity of the event. The work presented here further grapples with the challenge of implementing newsworthiness in code, in particular for our use-case of national politics reporting.

Other systems have been developed to monitor numerical data streams and trigger alerts when rules related to trends or outliers are matched [17]. The Marple system was able to detect anomalies, outliers, and trends in municipal data sets [11], and has graduated into a full-fledged lead discovery system renamed "Newsworthy" which operates using EuroStat data (see https://www.newsworthy.se/en/). The Local News Engine scans local data from courts, housing developments, and business licences to identify and alert journalists to the names of newsworthy people, places, or companies [14]. Computational news discovery has also been applied to fact checking workflows to help monitor and identify fact checkable claims for journalists [7]. Automated writing has been used in ways that make localized stories available to reporters, some of which are then used as leads inform additional local reporting [4]. Interactive data-driven systems have been deployed to help journalists identify newsworthy patterns in data [3], identify potentially interesting angles for stories based on user comments [13], and surface hypotheses for investigative journalists from large document sets [2, 5].

Given the various demonstrations of CND systems, some research has begun to examine their integration in journalism practice, their limitations and constraints, and their applicability to different types of reporting scenarios [4, 18] . In this work we consider a scenario that has received less attention: national politics reporting, and describe the design of the Lead Locator system to support this use-case.

## 3 Lead Locator: Design and Development

Lead Locator was iteratively designed by the Washington Post's Newsroom Engineering team in collaboration with reporters and editors in the newsroom. We applied a user-centered and agile methodology in order to improve the utility of the tool for practicing journalists. This involved an initial set of interviews with reporters and editors to better understand the use case for the tool, as well as additional user testing and interviews as the tool was developed to ensure the prototype was responsive to users' needs and interests. As designers, we additionally considered the previous work related to computational news discovery systems when defining user interface requirements as well as the overall expected workflow of the tool. In the following sections, we describe the design and development of the system in detail.

### The Domain: National Politics Reporting

In order to better understand the domain of national politics reporting, we interviewed five national politics reporters and editors at The Washington Post. We kept the initial interviews fairly open-ended by asking questions about challenges in reporting, approaches for sourcing stories, and use of data and other tools. The goal was to better understand the use case that our system should support, as well as to identify ideas for initial features that would enable reporters.

One approach to national politics reporting that was mentioned several times was a focus on voters. There was an interest in identifying small specific groups of people that are key to understanding political trends nationally (e.g. 20-somethings, African-Americans in urban areas, etc.). Reporters typically will develop seed ideas based on previous conversations with voters or other sources and then pitch that to their editor in an ongoing conversation. In some cases an idea will come down from an editor. The two core challenges our interviewees identified are (1) figuring out *where to go* and then, (2) *who to talk to* once you go there. In this work we focus on addressing the first challenge.

Reporters will sometimes use data to identify interesting locations, such as by looking at the vote margin between Democrats and Republicans, if a congressional district has flipped from one party to another, or by looking at specific data such as unemployment rates to see where there are above average areas. It can be hard to know where to go to find a "surprising" story. This is confounded by the idea that sometimes there's an interesting idea but the timing isn't right: reporters must also consider the question of "why now?". Reporters typically want to find a place where there's some local hook with respect to a broader idea. They're also interested in going to places that are "untouched" or that at least haven't received major recent coverage.

Once reporters and their editors reach consensus to move forward with a story idea reporters will oftentimes go to a location and then attend events in that location where the types of people they're looking to interview are likely to congregate (e.g. could be a rally, yoga class, coffee shop, post office, retail store, etc). They may also reach out to different organizations that might put them in touch with a certain type of person (e.g. a teacher's union).

### The Data

The Washington Post purchases a national voter file from L2 Political[1]. This voter file is a dataset containing voting history and demographic information for every registered voter in the U.S. (~190 million rows). It contains 650 columns, including primary and general election vote history going back until the 2000 election, demographic data such as gender, ethnicity, and age and information on party registration

[1] https://l2political.com/our-data/

and interests such as subscriptions to hunting, gardening, or cooking magazines. The data is originally generated from a number of different sources, including voter registration, vote history, public records, and other private databases. The Post only considers large aggregations of data.

Each piece of demographic information is either taken directly from the voter registration records (whether they are public depends on the state) or statistically modeled by L2. The fact that the data and data accuracy may vary from state to state is a significant challenge to using the data in a reliable way for journalism. L2 suggests a lower bound for the accuracy of some predicted columns, offering statements suggesting the accuracy "is 85% or better". However, L2 does not give any accuracy guarantee or transparency into their modeling process. These data limitations, and the variance even between states, means that it is best to use the voter file only for hypothesis generation. Voter file data is updated by L2 one to two times per month.

In addition to the voter file we use other data to define the political relevance or geographic context of a location. For instance we use the U.S. Election Atlas[2] for historical voting results tabulated by county for 2012, 2016, and 2018. We also use data from the Daily Kos[3] on open congressional seats. Finally, we use census data for population and we use estimates of whether a county is urban, suburban, or rural [4].

### Algorithmic Newsworthiness

In order to encourage usage an important design goal for Lead Locator was to align leads with established expectations for what journalists typically find newsworthy. There are a wide range of newsworthiness criteria that have been identified in the literature including factors such as exclusivity, conflict, surprise, reference to elites, magnitude and significance, proximity, audience expectations, and organizational agenda [10]. And while there have been some prior efforts at translating a few of these news values into algorithms [11, 12], challenging research remains to be done on how to computationally specify and configure a range of newsworthiness definitions [4, 18]. In this work we focus on operationalizing three conceptual newsworthiness factors that we apply to each county in our dataset: *novelty*, *political relevance*, and *magnitude*.

*Novelty.* Here we outline three perspectives on novelty which we think apply to this use case: contrast to central tendency, contrast to known relationships, and contrast to geographic context (we have only implemented the first one, leaving the latter two for future work).

- *Contrast to Central Tendency.* Our basic approach to capturing novelty is to contrast a variable to a measure of the central tendency of the distribution of that variable. For instance, we might take the turnout rate of a county and compare that to the average turnout rate for counties in a state. To formalize this we compute the signed z-score, which is the number of standard deviations a data point is above or below the mean of the distribution. In ranking counties we take the absolute value of the z-score, however we use the sign of the score in rendering the text of the tip sheet. This method can also be used to define distributions based on transformations of other variables. For instance, we might be interested in the difference of two variables, such as the male turnout rate and the female turnout rate, and then compute z-scores based on that difference.
- *Contrast to Known Relationships.* This perspective is meant to capture the degree of deviation of a data point from a known relationship such as a correlation. If, for instance, we consider that in general there is a correlation between the political engagement of voters and the total number of donations to campaigns in a location, it would be potentially newsworthy to find locations which buck this trend. In other words, locations which have high political engagement but low donations, or low political engagement and high donations.
- *Contrast to Geographic Context.* A location may be particularly different to its neighbors. In other words a variable for the location may contrast substantially with surrounding locations within some radius of interest. The goal is to try to identify locations that may be bucking some local geographic trend. In the future we are interested in applying Moran's I, a measure of spatial autocorrelation, to help identify geographic "interestingness" [6].

*Political Relevance.* Previous work has defined relevance in the context of newsworthiness as the "intensity of damage or benefit of an event" [1]. In this work we take political relevance to mean the importance of a location to the outcome of the 2020 election. In particular we try to operationalize this idea in four ways: (1) vote margin between the two dominant parties; (2) shift in that vote margin from 2012 to 2016, and 2016 to 2018; (3) whether the location flipped back and forth in its support of one party or another between 2012, 2016, and 2018; and (4) whether there is an open congressional seat overlapping the location in 2020.

*Magnitude.* The newsworthiness factor of magnitude captures the idea that stories tend to be more newsworthy when they involve or impact a large number of people [10]. Here we operationalize this idea straightforwardly by using county population as a measure of magnitude. In the future we are interested to explore the prevalence of a particular pattern in

[2]https://uselectionatlas.org/

[3]https://www.dailykos.com/stories/2018/2/21/1742660/-The-ultimate-Daily-Kos-Elections-guide-to-all-of-our-data-sets

[4]https://gist.github.com/gebelo/a3bae4b4bef43b3680392423c2fbb220

terms of the number of counties that exhibit it as another aspect of magnitude. So, for instance, if a county demonstrates some pattern of exceptional new registrations amongst Hispanic voters, it may be even more interesting if that pattern is present in 80% of counties within a state.

### Tip Sheet Generation

Counties are ranked by calculating and summing a set of weighted scores to arrive at a final score used to rank. Each tip sheet is generated based on a configuration file which defines the overall score composition and weights as well as the main measure (e.g. 2018 Turnout), the dimensions of interest (e.g. Democrat, Republican), transformations (e.g. difference between Democrat and Republican), filters (e.g. turnout of a particular ethnicity), and normalization (i.e. what variable should function as a denominator for another variable). The configuration file also contains metadata such as the definition of terms used, data provenance, and a list of geographic scopes to compute tip sheets (e.g. specific states). We opted for a heavily authored configuration so that tip sheets could be steered towards interesting patterns. In the future we may allow for interactive configuration so that reporters can dynamically define what they want to rise to the top [13].

Once scores for each county are calculated we use automated data-to-text generation to output text to populate the tip sheet [15]. The tip sheet consists of three structured sections which include at least one and possibly multiple bullet points that are generated based on the statistical salience of various sub-scores. A point has text generated if its associated sub-score is above a threshold (typically 1.0 for a numeric score, or "true" for a boolean score). In other words, points will not be generated unless they meet some level of "interestingness" as defined by the underlying score statistic.

Different templates are used to render the text for different scores. Currently we have templates for the main measure, for a gender or party contrast, for political relevance, and for location context (See Figure 1). The templates are relatively simple, but include conditional logic to insert appropriate comparisons (e.g. "higher", "lower", "the same") or adjectives of direction (e.g. "left", "right") or magnitude (e.g. "small", "moderate", or "large" increase). In the future we hope to generalize and expand the templates used to provide additional variability and complexity to the language generation.

### User-Interface Design

During our initial requirements gathering interviews, reporters expressed varying interests, skillsets, and workflows. For example, some had more experience with data, and some had greater interest in specific topics and populations. Primary concerns for the interface design were therefore displaying compelling information about each lead and creating a workflow that would be useful for a wide range of reporters. Doing so required turning dense information into a clear, skimmable format. In addition, because there are so many counties and metrics in the voter file, it was important to give the appropriate amount of context to pique reporter interest, and help explain why a lead might be newsworthy to pursue further. With all that in mind, the interface consists of two main areas: tip sheets, and the index.

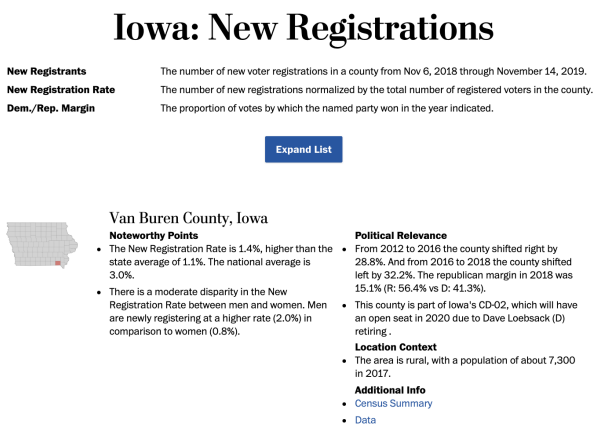


**Figure 1: Tip Sheet for New Registrations in Iowa.**

*Tip Sheets.* Tip sheets are titled with the geographical scope and metric (See Figure 1). Below that is a list of relevant data definitions to aid proper interpretation, followed by the ranked list of counties. Each item on the list contains a locator map on the left in order to provide quick visual context. To the right of that is the county name, along with four sections of bullet points. The first, "noteworthy points," contains points of general statistical interest. Other sections are shorter, and listed in a separate column to provide visual emphasis on them as a collective, and to increase readability. They consist of "political relevance," (i.e. how the county voted in recent elections or seats that will soon be up for election); "location context," which lists population and an urban, suburban, or rural categorization; and, lastly, "additional info," which has links to a more detailed Census Reporter profile, and to view the raw data, in case the user is interested in further research for a story.

Each tip sheet defaults to showing "top leads" only, i.e. counties that are determined to be above some minimum threshold score. However, user testing suggested that while some prefer a predetermined list, others want to browse *all* counties in a state. We therefore include a toggle button to show the full list of all counties in the geographic scope (still ordered according to the ranking metric of the tip sheet).

*Index Page* Reporters expressed two use cases for looking at the site: general inspiration for stories, and location specificity, i.e. having a place already in mind, such as a state, and wanting to be pointed somewhere even more specific. Because of that, the index page serves as a directory for quickly


**Figure 2: Lead Locator index page.**

browsing tip sheets. It contains a searchable list of all available tip sheets, ordered by recency, in the form of content previews that include their geographical scope, metric, and date of creation (See Fig. 2). A search option combined with the full list of tip sheets serves both use cases: If users want a general overview of recent updates, they can see recent tip sheets at the top and can casually browse. At the same time, users can search for a specific state or metric and will be shown relevant tip sheets only. The index also includes a link to an about page, with explanations on the methodology and how reporters should attribute the tool as a source. This was included to provide additional transparency into the data and process for creating tip sheets.

## 4 Evaluation

We iteratively developed the Lead Locator by periodically asking reporters and editors for feedback. In total we evaluated the tool with eight journalists, in sessions that ranged from 30 to 90 minutes. We asked users to "think aloud" as they looked at a tip sheet and try to explain what was interesting (or not) about each lead, what kinds of questions or hypotheses it prompted, and how they might follow up on a lead. We further probed our participants using a card sort exercise consisting of 13 of the variables available in the voter file, which we asked them to rank according to their interests and questions that might be spurred from each variable. This helped direct our development of new tip sheets around variables that were of interest. Overall this feedback informed our understanding of the need for various features (e.g. showing all counties), newsworthiness metrics (e.g. needing to better account for the magnitude of a place), or pieces of information on the tipsheet (e.g. population, urban / rural / suburban designation) as well as for the different approaches reporters had in terms of coming to the tool with more crystallized hypotheses versus having a more open mind.

As it is deployed more widely we are interested in evaluating Lead Locator in an ongoing fashion, including by using quantitative ratings of lead relevance and quality, perceived utility, and overall satisfaction, as well as with qualitative feedback on the experience of using the tool and whether the leads appear to be stimulating new and original reporting. We would also like to measure the number of tip-sheets accessed by reporters, the time spent browsing leads, and the number of story-assists the tool provides.

## References

[1] F. Badenschier and H. Wormer. 2011. Issue Selection in Science Journalism: Towards a Special Theory of News Values for Science News? In *The Sciences' Media Connection –Public Communication and its Repercussions*. Springer Netherlands, 59–85.

[2] M Brehmer, S Ingram, and J Stray. 2014. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *In IEEE TVCG* 20, 12 (2014), 2271–2280.

[3] Meredith Broussard. 2015. Artificial Intelligence for Investigative Reporting. *Digital Journalism* 3, 6 (May 2015), 814–831.

[4] Nicholas Diakopoulos. 2019. *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

[5] C Felix, A V Pandey, and E Bertini. 2015. RevEx: Visual Investigative Journalism with A Million Healthcare Reviews. In *Proc. Computation + Journalism Symposium*.

[6] T. Gao, J. Hullman, E. Adar, B. Hecht, and N. Diakopoulos. 2014. NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News . In *Proc. CHI*.

[7] Lucas Graves. 2018. *Understanding the Promise and Limits of Automated Fact-Checking*. Technical Report.

[8] James T Hamilton. 2016. *Democracy's Detectives*. Harvard University Press.

[9] J T Hamilton and F Turner. 2009. *Accountability through algorithm: Developing the field of computational journalism*. Report from the Center for Advanced Study in the Behavioral Sciences.

[10] Tony Harcup and Deirdre O'Neill. 2016. What is news? *Journalism Studies* 23, 1 (March 2016), 1–19.

[11] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. In *Computation + Journalism Symposium*.

[12] A. Nourbakhsh, Q. Li, X. Liu, and S. Shah. 2017. "Breaking" Disasters - Predicting and Characterizing the Global News Value of Natural and Man-made Disasters. In *Data Science + Journalism Workshop*.

[13] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. *Proc. CHI* (2016).

[14] William Perrin. 2017. Local News Engine: Can the machine help spot diamonds in the dust? In *Data Journalism Past, Present, Future*, J. Mair, R. Keeble, M. Lucero, and M. Moore (Eds.). Abramis.

[15] Ehud Reiter. 2007. An architecture for data-to-text systems. *Proc. Workshop on Natural Language Generation (ENLG)* (2007).

[16] R Schwartz, M Naaman, and R Teodoro. 2015. Editorial algorithms: Using social media to discover and report local news. In *Proc. ICWSM*.

[17] M Shearer, B Simon, and C Geiger. 2014. Datastringer: easy dataset monitoring for journalists. In *Computation + Journalism Symposium*.

[18] Jonathan Stray. 2019. Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism* 7, 8 (2019), 1076–1097.