3

AUTOMATED CONTENT PRODUCTION

The Swedes love their soccer. So much so that in mid-2016 the Swedish local media company Östgöta Media decided to launch a new site called "Klackspark" to cover every local soccer game in the eastern province of Östergötland. "It's like the national sport of Sweden and you play it even if you're not that good at it," Nils Olauson, the publisher of Klackspark, told me. "I played in Division 6, one of the lower leagues, maybe until I was thirty-four just because it's fun and there's still a lot of prestige in the game. You want to be the best in your neighborhood, you want to be the best in your part of the town, and you get your rivals even in that kind of low league." The sport has a high cultural significance. And when practically every neighborhood has its own team, spanning six divisions of local men's soccer and four divisions of local women's soccer, not to mention the major national and international leagues, that's a lot of games. To reach that breadth of coverage, Klackspark strategically employs automated software writing algorithms, which take structured data about each local game and automatically write and publish a short, roughly one-hundred-word factual summary of what happened in the game. It's not too fancy really. Any given story might recount who scored the goals in addition to the history and league standing of the teams that played. But the automation provides a foundational *breadth* to the coverage—anyone looking for the quick facts about a local match can find that story on the site.

Olauson described how Klackspark's automated stories are orchestrated with fourteen sports reporters who then add a bit of "spice," layering on details and human interest to the stories on the site. The reporters receive alerts when the software detects something newsworthy or unique in a lower-league game. "When a girl had ten goals in the same game, we had one of our reporters call her up and talk to her. He wrote an article about it, and that article was one of the most read pieces on Klackspark that week." The automation is serving a dual purpose: writing straight factual stories that are directly published and, as

detailed in Chapter 2, alerting human reporters to what could be a juicy story if only they did some additional reporting, got some quotes, and fleshed it out. Straight automation provides breadth of coverage, and automation plus professional reporters adds depth to coverage. The automation does the routine work, and the reporters get to focus on more interesting stories. No jobs lost either, at least not yet.

This type of hybrid scenario was the norm for the news organizations I spoke to about their use of content automation. Newsrooms see content automation as being largely complementary to journalists' work.[1] Yes, there are instances in content production where there is complete automation, and if you squint, you might even say there is artificial intelligence operating in narrow targeted areas. But the state of the art is still far from autonomously operating in the unbounded environment of the world and from doing the contextualized interpretation and nuanced communication required of journalists.

This chapter focuses on the content creation phase of editorial production. As various content creation tasks are delegated to automation and algorithms, new opportunities emerge for reinventing editorial processes and practices. I first examine the technical capabilities and potential of automated content production algorithms. Then I detail how algorithms enable faster, larger scale, more accurate, and more personalized journalism, which creates new business opportunities for news organizations. At the same time, this chapter also points out that automated content production has some very real limitations, such as data contingencies and difficulties matching human flexibility and quality in reporting on a dynamic world. This leads to the next topic, a consideration of how people and automation will work together, both in the design and operation of these systems, as well as in how this collaboration will impact the evolution of human tasks and roles. This chapter concludes by exploring what might be next for automated content production.

How Automated Text Writing Works

The basic premise of automated text production is to take structured data, as one might find in a database or spreadsheet, and have an algorithm translate that data into written text. This process is referred to as "natural language generation" (NLG). At the simpler end of NLG are rule-based techniques that work like "Mad Libs"—that is, there are prewritten templates with gaps where numbers are dynamically inserted from a dataset according to manually crafted rules (see Figure 3.1). More advanced template approaches are imbued with rule-sets that

incorporate linguistic knowledge and facilitate more sophisticated and dynamic text production. Such techniques can conjugate verbs in different tenses or decline nouns to make them grammatical. Sophisticated templates can be blended in with simpler ones. For instance, a simple template could be used for a headline so that it's more attention getting, but then a dynamic template could drive more heavily descriptive parts of the story.[2] Many rule-based NLG systems are built according to a standard model that includes three distinct stages: document planning, microplanning, and document realization.[3]

The document planning stage consists of determining and selecting *what* to communicate and then *how* to structure that information in paragraphs and sentences. Deciding what to communicate is impacted by what the reader is interested in, what the writer is trying to accomplish (for example, explain or persuade), and by constraints such as the available space and data. Document structure reflects the editorial priority and importance of information as well as higher-level discourse objectives such as telling a story versus explaining a timeline. In the domain of weather a document plan might prioritize the salience or ordering of information related to warnings, winds, visibility, and temperature, among other factors.[4] In election result articles it may take into account the "interestingness" of particular candidates or municipalities, as well as whether a win or loss is a statistical aberration.[5] Document planning works to enumerate all the things the data *could* say and then prioritizes those facts according to newsworthiness criteria.[6] More sophisticated systems may identify "angles" for stories in order to help structure the narrative based on rare events or domain-specific metrics.[7] Depending on what the data indicates, example angles for a sports story might be "back-and-forth horserace," "heroic individual performance," "strong team effort," or "came out of a slump."[8]

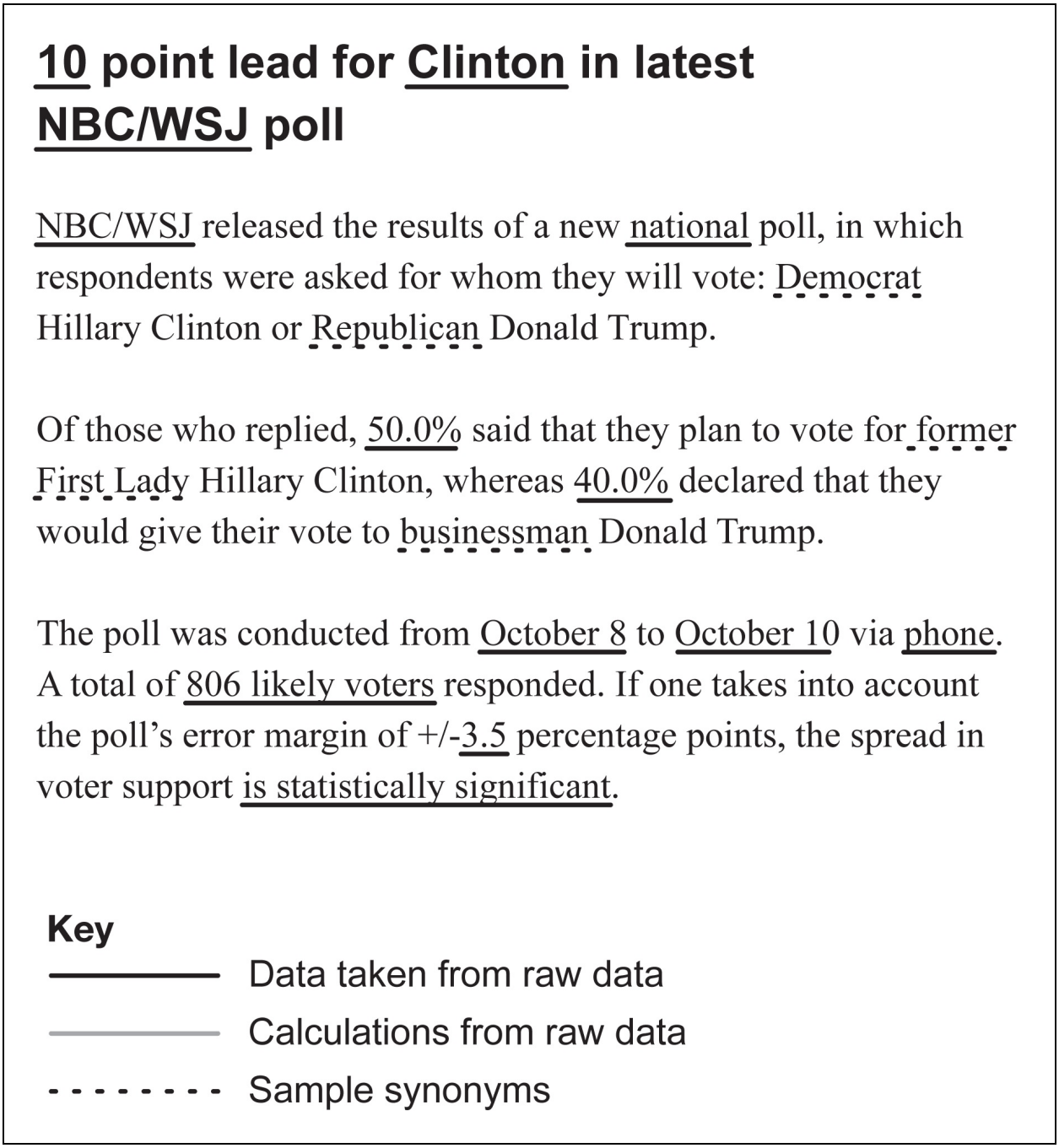

Figure 3.1. An excerpt of an automatically generated article on PollyVote.com reporting on the results of a US election poll from 2016. Underline style indicates different types of dynamic text. *Source:* Andreas Graefe, "Computational Campaign Coverage, *Columbia Journalism Review,* July 5, 2017 (used with permission of the author).

The next stage, microplanning, consists of making word and syntax choices at the level of sentences and phrases. This phase of text generation is important because it impacts the variability and complexity of the language output, as well as how publication or genre-specific style guides, tones, or reading levels are

produced. Microplanning entails nuanced choices, such as deciding among different templates for conveying the same information or among referring expressions, which specify different ways of mentioning the same person or company in an article. The first time a player is mentioned in a soccer article, their whole name might be used, the second time just the last name might suffice, and if there is a third time, it might be more interesting to use an attribute of the person such as "the 25 year-old."[9] Additional variability can be injected into output texts by integrating synonym selections from word ontologies, which define structured relationships between different words (see an example in figure 3.1). Careful word choice can steer the text away from monotony by blocking the use of certain verbs or phrases if they've already been included in a text. In the case of rule-based systems microplanning decisions reflect editorial choices deliberately coded into the algorithm.

Finally, the realization stage of NLG walks through a linguistic specification for the planned document to generate the actual text. This stage satisfies grammatical constraints, such as making a subject and verb agree, declining an adjective, making sure a noun is pluralized correctly, or rendering a question with a question mark. The realization stage is the most robust and well-studied aspect of NLG, with mature toolkits such as SimpleNLG being used by news organizations to generate text.[10] Many of the decisions at the level of text realization relate to grammar and so there is less potential for journalists to imbue editorial values in the algorithms driving this stage.

An alternative to the standard approach to NLG involves data-driven statistical techniques that learn patterns of language use from large corpora of examples. For instance, machine learning can automatically classify which points of data to include in an American football recap article. This contributes to the document planning stage by considering the context of data points and how that affects inclusion decisions.[11] Such techniques require a high degree of engineering to produce output in closely constrained scenarios. Moreover, rich datasets of content and data need to be available. In one case, a wind forecast statistical NLG system required a parallel corpus of manually written wind forecast texts as well as the aligned source data for each text.[12] News organizations with large corpora of articles could potentially utilize those texts as inputs to train statistical NLG systems.

Purely statistical approaches may, however, come up against barriers to adoption because they can introduce unpredictable errors. In the shorter term, a fruitful path forward appears to be the marriage of template- and statistically

based techniques. In 2013 Thomson-Reuters demonstrated a research system that could extract and cluster templates from a corpus of examples, creating a template database.[13] Then the system could generate a new text by iteratively selecting the best template given the available data for each successive sentence. The text production quality was competitive with the original texts, and had the additional benefit of introducing more variability into the output texts, which addressed a weakness of purely template-based approaches.

Beyond Automated Writing

Automated content production is not limited to writing texts, and can work with different inputs besides structured data, including the full range of unstructured texts, photos, and videos proliferating online. To be useful, however, these media must often first be converted into more structured data. Algorithms extract semantics, tags, or annotations that are then used to structure and guide content production. For text this involves natural language understanding (NLU) and for images or video this involves computer vision to detect or classify visual properties of interest. For instance, Wibbitz, a system that semi-automates the production of videos, uses computer vision to identify faces in input photos and videos so it can appropriately frame and crop the visual output.

Another type of content automation is summarization. Given the inundation of social media content, there's a lot of potential value to using algorithms to crunch things down into output summaries that people can skim. Algorithms can summarize events as diverse as the Facebook initial public offering, the British Petroleum oil spill in 2010, or a World Cup match by curating sets of representative media, such as tweets or photos.[14] Summarization can also generate a headline to share on social media, a compact presentation for a news browsing app, or a set of important take-aways from the story.[15] Summarization approaches can be either extractive or abstractive.[16] Extractive summarization corresponds to the task of finding the most representative sentences or text fragments from a document or set of documents, whereas abstractive summarization can synthesize entirely new sentences and words that didn't exist in the original text. Summarization algorithms embed a range of meaningful editorial decisions, such as prioritizing information from inputs and then selecting informative, readable, and diverse visual content. While summarization technology is advancing, it's sometimes hard to tell whether one summary is much better than another. Wibbitz uses summarization algorithms to reduce an input text into a set of points to be illustrated in video, but as Neville Mehta from

Law Street Media explained, “Sometimes you get something that’s shorter but it’s not doing the original piece justice.… There is quite a bit of hand editing that goes into it just to maintain the level of quality that we want.”

Video generation is considerably more challenging than text because it entails automatically tailoring the style, cropping, motion, and cuts of the visual, while also considering aspects such as the timing between the visual and textual overlays to the video. A system like Wibbitz must curate visual material from a database (such as Getty Images) and then edit that material together coherently. Early research systems capable of editing and synthesizing video stories began to emerge towards the late 2000s,[17] but only in the last few years has the technology been commercialized by the likes of Wibbitz and its competitor, Wochit. Wibbitz is further advancing the field by integrating machine learning that can learn editorial importance from user interactions with the tool. As human editors tune rough cuts, the system learns which content is most interesting to include for other videos on similar topics.

Another output medium for automated content production is data visualization. Visualizations are increasingly used in the news media for storytelling and involve mapping data to visual representations such as charts, maps, timelines, or networks to help convey narratives and engage an audience.[18] Data visualizations are also quite multimodal and often require incorporating a healthy dose of text to aid interpretation.[19] Weaving text and graphics together raises new challenges about which medium to use to convey which types of information, how to refer back and forth between the visualization and the text, and decide which aspects are most important to visualize for the overall story.[20] Automatically generating captions or other descriptive text goes hand in hand with automating data visualization.[21] For instance, systems can automatically produce annotated stock visualizations or annotated maps to augment news articles (see Figure 3.2).[22] The editorial decisions encoded into these algorithms are not only in selecting what to show, but also how to show it, and what to make most visually salient through labeling, highlighting, or additional annotation. The Associated Press (AP) produces automatically generated graphics on topics such as the Olympics, finance, and box office numbers, which are then distributed on the wire. Other publishers such as *Der Spiegel* and Reuters are beginning to experiment with automated data visualizations that can augment or even anchor articles.[23]

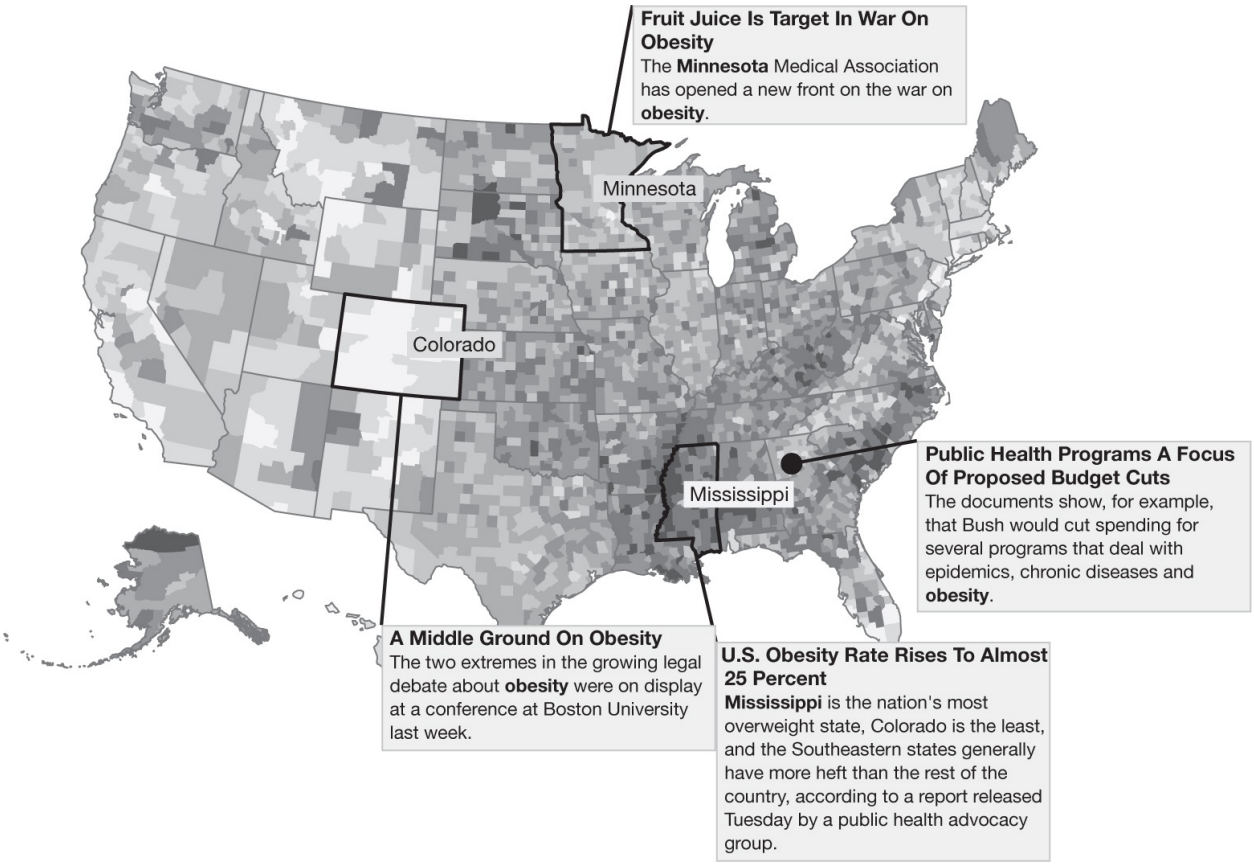

Figure 3.2. A map visualization of obesity rates in the United States, including annotations of various areas of interest. The map was automatically generated based on an input article on obesity and diabetes. The NewsViews tool was originally described in T. Gao, J. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, "NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News," *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (New York: ACM, 2014).

Machine learning is opening up fascinating new possibilities for automated content, including the wholesale synthesis of new images, videos, or texts on the basis of training data. Research prototypes have generated video using other videos or audio, as well as synthesized images based on photographic corpora, voices using audio of potentially important sources such as politicians, and text to mimic user-generated comments.[24] One prototype demonstrated a system that can take ordinary video footage of a person's face from a webcam, understand the facial expressions made, and then map those onto another video's face in real time. The demo is provocative: videos show the face of Putin or Trump mimicking the same facial movements as an actor's.[25] Faces can now be swapped from one body to another, creating what are popularly known as "deepfakes."[26] Other systems take in an audio clip and synthesize photo-realistic interview video that synchronizes the audio to the mouth shape of the speaker

based on training footage.[27] Photos can also be synthesized entirely from data. A machine-learning model trained on a database of 30,000 high resolution photographs of celebrities can output photos of faces for people who don't exist.[28] The photos have the look and feel of celebrity headshots and are striking in their quality (see Figure 3.3). In terms of text, neural networks can generate Yelp reviews that are rated just as "useful" as reviews from real people—essentially going unnoticed.[29]

Figure 3.3. Automatically synthesized faces of people who do not exist created using a neural network model trained on celebrity photos.

Source: T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *presented at the International Conference on Learning Representations*, Vancouver, Canada, 2018, https://arxiv.org/abs/1710.10196, licensed via Creative Commons Attribution-NonCommercial 4.0 International, https://creativecommons.org/licenses/by-nc/4.0/.

Algorithms that synthesize entirely new images, videos, and texts challenge the veracity of visual media and the authenticity of written media. The creation of fake videos showing sources saying something they did not could wreak havoc on the use of visual documentary evidence in journalism.[30] Photoshop has been undermining trust in visual media for years, but these new technologies create a whole new potential for scale and personalization. The output of these systems is rapidly advancing in believability, though to a trained eye there may be subtle signs that the videos aren't quite right, such as the flicker of an imperfectly synthesized mouth or a glazed look in the eyes. As the synthesis

technology moves ahead quickly, forensics technology that can identify synthesized facial imagery or synthesized texts is also being developed.[31] Even if synthesized content might sometimes fool the human eye, the forensic algorithm's statistical eye will know it's faked. Still, an arms race is taking shape in which journalists will need to be equipped with sophisticated forensics tools in order to evaluate the authenticity of potentially synthesized media.

Automated Content in Use

The news industry is still actively exploring different domains and use-cases where there is the most strategic potential for content automation. Automated text writing has been used for more than two decades to produce weather forecasts.[32] In finance and markets, Bloomberg has been automatically generating written news alerts for its terminal for more than a decade, and Reuters currently uses automation to facilitate the writing of market recap reports on a daily basis. In 2011, Forbes rolled out automated corporate earnings previews and reports using technology from Narrative Science, and in 2014, the AP began publishing automated earnings reports as well. *Le Monde* deployed automated writing to help report French election results in 2015, and in 2017 in Finland the public broadcaster Yle as well as the newspaper *Helsingin Sanomat* and academic project Valtteri reported the results of the Finnish municipal elections using automation.[33] In the run up to the 2016 US election, automation was used by election forecasting site PollyVote to render written stories about the nearly daily election polls that were coming out.[34]

A mid-2017 survey of fourteen news agencies found that eleven were already using automation or were actively developing it, mostly in the domains of finance and sports.[35] Various publishers have pursued use-cases in specific sports such as soccer, ice hockey, and baseball.[36] The *Washington Post* has used the technology to cover events ranging from the 2016 Olympics to local football games, as well as to cover the 2016 US elections.[37] ProPublica and the *LA Times* have dabbled in generating automatic content in domains such as education, crime, and earthquake reporting.[38] MittMedia in Sweden writes articles about local real-estate transactions automatically.[39] Beyond articles, the *Washington Post* and other publications have explored automated headline writing,[40] and McClatchy has used the technique to convey information from a database as part of a larger investigative story.[41] Some outlets, including the UK Press Association in collaboration with Urbs Media, are exploring the coverage of topics such as road conditions, public safety, crime, and health using open

municipal data.

Automation's Advantages

Over time innovators are expanding the scope of content genres and domains where automation offers a benefit. In some cases automation enables a new activity that people simply would not be capable of, whereas in other cases it offloads repetitive tasks from journalists or allows new scenarios of coverage that while possible for people to perform would be cost prohibitive. Automation affords one or more of several advantages in content creation by enabling speed, scale, accuracy, and personalization.[42]

Being the first to break a news story is often a competitive strategy for news organizations. They see speed as a way to drive traffic and enhance authority, despite there not necessarily being the same demand or urgency for speed from the audience.[43] But in some scenarios, and for some end users of news information, speed really is *the* defining factor. At Reuters as well as Bloomberg, which both sell specialized information terminals to stock traders, automation parses text documents, such as earnings releases, and almost instantaneously generates and publishes a headline to the terminal interface that reflects whether the company beat or missed earnings expectations. It's simple stuff written to specific standardized text templates, such as "IBM missed expectations by 0.12: 3.02 versus 3.14." But when new financial information can move a stock, being the first to have access to information can mean profit—and so speed is essential for traders. Finance appears to be the sole domain where full automation—complete autonomy across the entire production pipeline—has gained traction. To interject a human in the loop would diminish the value of the information to traders by slowing it down. Even a non-zero margin of error is outweighed by the sheer demand for speed. Automation for speed in this case is about providing a service that simply would not be possible to provide using human labor.

The domain of finance itself is not necessarily the defining factor in making speed critical, however. The importance of speed is more a reflection of the information needs of particular end users. Take for instance the AP's implementation of quarterly earnings reports described in the introduction to this book. Unlike the headlines pushed to the Reuters and Bloomberg terminals, AP's reports are not published within milliseconds of a corporate earnings release, but instead may be pushed to the wire anywhere from five to ten minutes or even an hour after an earnings release. Thereafter, they may be published by local AP-affiliated newspapers for the sake of local retail investors or individuals who are

interested in a particular company. In determining whether speed will be an advantage in a particular automated content scenario, designers should look to how end users will actually use the information. For instance, the *LA Times'* Quake Bot, which reports on earthquakes automatically, generates and publishes story stubs very quickly because this can serve to alert readers to potential risks or safety issues for themselves or loved ones who may be within the earthquake zone. For both prescheduled events (such as earnings releases) and unscheduled ones (such as earthquakes) automation for speed depends on the needs of the audience for the information.[44]

The potential for scale, of increasing coverage and breadth, is another driving factor in the adoption of automated content. A computer algorithm can churn out hundreds or thousands of variations of a story quite easily once set up, parameterized, and fed with a stream of data. Automated content use cases are emerging in areas such as weather, sports, and elections, where the cost of developing an automated system can be amortized over a high volume of output with a long tail that might not otherwise attract dedicated attention from journalists. For instance, instead of covering only 15 percent of congressional races as it had in the past, in the 2016 US elections the *Washington Post* was able to cover elections in all states, including 435 house races, 34 senate seats, and 12 gubernatorial races. That's a dramatic expansion of coverage. And although the *Post* could have reached that scale of coverage using human reporters in prior elections, the costs would have been prohibitive. Automation offers the possibility to create content for ever smaller audiences, relieving the pressure to make editorial decisions that balance labor availability with newsworthiness judgments about things like the magnitude of the event.[45] What constitutes "newsworthy" changes when it's cheap and easy to cover basically everything.[46]

Increased breadth of coverage can allow for more widespread access to information which can in turn impact other societal processes. For instance, one study showed that the earnings reports rolled out by the AP have affected stock-trading behavior. The research examined 2,433 companies on the US exchanges that didn't have any financial earnings coverage by AP prior to the introduction of automation in October 2014.[47] By the end of the sample period, in October 2015, about two-thirds of the companies tracked had received some automated coverage and exhibited higher liquidity and trading volume in comparison to those that didn't receive automated coverage. The relative increase in volume was about 11 percent.

Scale can be considered across different dimensions, including categorical

units such as sports teams (for instance, Klackspark's soccer coverage), geographic units (such as the *Washington Post'*s elections coverage), or temporal units (such as periodic polls in politics). The idea of scaling over time has an additional benefit: rather than widening coverage over a conceptual or spatial domain, it allows for consistent and uniform coverage of the same type of events over time. The *LA Times'* Quake Bot, which automatically writes stories about earthquakes using data provided by the United States Geological Survey (USGS), demonstrates this idea.[48] Earthquakes are always happening in California, and while there was audience demand to know when they happened, the *LA Times'* coverage was inconsistent. No reporter really wants to have to write a repetitive pro forma story about a local, relatively minor earthquake. Automation allowed the paper to offload the tedium of writing those basic articles while increasing the consistency of coverage, since basic rules about the magnitude of quakes that warranted a story were baked into code. Ben Welsh, the head of the data team at the *LA Times,* remarked that the reporter who had been tasked to write those stories in the past was now able to focus on more important investigations, such as what buildings in the city were unsafe for the next big earthquake. Scaling over time can contribute to time savings and potentially to labor reallocation that can buttress more substantive coverage of a topic.

Accuracy of text can be another advantage of automated content production. As Andreas Graefe writes in the *Guide to Automated Journalism,* "Algorithms do not get tired or distracted, and—assuming that they are programed correctly and the underlying data are accurate—they do not make simple mistakes like misspellings, calculation errors, or overlooking facts." The remaining errors are usually attributable to issues with the underlying data.[49] Lisa Gibbs, the business editor at the AP who helped with the roll-out of the automated earnings reports, concurred: "The error rate is lower than it was with human reporters.… Robots do not make typos and they do not make math errors." And if an error is detected in the output, the software needs to be fixed only once and then it can go back to operating accurately across its thousands of outputs. So once the systems are set up and debugged, they're consistent in their application of a procedure. But that doesn't always translate into more accuracy per se. Yes, automated content production can mitigate the clerical errors related to typos and math mistakes, but different types of errors and accuracy issues can crop up. The different accuracy profile of automated content in turn creates new demands on editors to recognize not the errors of sloppy fingers, but of missing data, unaccounted for

context, or other algorithmic limitations.[50]

Speed, scale, and accuracy are all possible reasons for adopting automated content, but there's another affordance of the technology that is still nascent: personalization. A variety of avenues have been explored for news personalization, such as adaptive navigation, dynamic recommendations, geo-targeted editions, or just reordering of content on a page to match an individual's interests.[51] The idea of news personalization has been around almost as long as the World Wide Web,[52] and many modern reading interfaces, from apps, to newsfeeds and homepages, routinely make use of these strategies.

Here I want to focus on personalization at the level of the content itself. This might, for instance, entail automatically rewriting the words of an article to appeal best to an individual's interests or to the characteristics of a particular publication's audience. For instance, financial content can be written quite differently for a general purpose audience consuming the AP's earnings reports as compared to the expert audience reading a niche financial news site. Personalization in this context means any adaptation of a communication to a user. More specifically, content personalization is "an automated change to a set of facts that appear in an article's content based on properties of the reader."[53] Adaptations occur with respect to some model of the user, which could include age, gender, education level, location, political affiliation, topical interests, or other attributes of the individual. When a user sets his or her own model, this is often referred to as "customization," whereas when the model is implicitly determined (such as by observing behavior), this is referred to as "personalization."[54] Oftentimes there is some combination of customization and personalization whereby media is adapted somewhat through personalization and then the user can make further adjustments to customize it. Research on baseball stories that can be interactively customized by end users to focus on particular players or parts of the game showed that such adaptive articles were rated as more informative, interesting, and pleasant to read than nonadaptive articles.[55] It's unclear whether such results hold for implicitly personalized articles, but if they do, this could improve the user experience of news consumption.

Localization is a more narrow type of personalization reflecting adaptations based specifically on geography or language. For instance, localized weather reports are produced for thousands of locations throughout the world.[56] In their pilot, Urbs Media localized articles for the thirty-three boroughs of London using open data to write an article about the diabetes incidence in each borough. In a project together with the U.K. Press Association called "RADAR"

(Reporters and Data and Robots), Urbs is expanding this approach to provide localized reports for a range of topics including health, education, crime, transportation, housing, and the environment.[57] A staff of four writers can produce fifteen story templates each week, each of which can in turn generate 250 localized versions.[58] Similarly, Hoodline, a local outlet in San Francisco, is experimenting with automatically localized stories about new or stand-out neighborhood restaurants using Yelp data.[59] Meanwhile, the Norwegian News Agency is considering localizing its automated soccer articles by creating different versions and angles on a story depending on which team a locale typically cheers for. So a home team loss might be framed in a softer way than when that same team crushes a crosstown rival. Another aspect of coverage localization relates to adapting content to a desired language of consumption. This has advantages for global organizations that publish content in multiple languages around the world.

Many of these applications of personalized content generation are in a nascent or developing stage. There are, however, already examples of personalized or customized content integrated into larger data-driven articles from the *New York Times*, the *Washington Post*, and *Vox*.[60] For example, a *Post* article entitled "America's Great Housing Divide: Are You a Winner or Loser?" maps and explains how real-estate prices have shifted over the last decade, showing the dip and uneven recovery after the 2008 financial crisis on a zip-code by zip-code basis. The article highlights several nationally interesting locations such as the San Francisco Bay Area, but it also presents a map and some text showing the locality of the viewer of the article. Using the location information modern web browsers infer based on a user's internet protocol (IP) address, the article is able to adapt the presentation of the content to the nearest metropolitan area and zip code. When I accessed the page in Washington, DC, it read, "In 2015, a single-family home there was worth $637,165 on average, about 79 percent more than in 2004. It's a densely populated, mostly black area. The home values are typically higher than most of the homes in the Washington-Arlington-Alexandria, DC-VA- area." The text is straightforward but is adapted to me as someone who at the time was living in the LeDroit Park neighborhood of Washington, DC, making it more contextually relevant.

The Business Case

"For us it's not about saving money—it's about doing stuff we couldn't otherwise do," underscored Ben Welsh about the possibilities for automation in

the *LA Times* newsroom. Whether it's speeding up production, increasing breadth, enhancing accuracy, or enabling new types of personalization, there is great potential for applying automated content production in journalism. But as editorial possibilities are explored, the business incentives and value propositions also need to be worked out. Can automated content help with some of the sustainability concerns that face newsrooms with shrinking budgets?

Some organizations are already seeing the business impacts of automated content in terms of competitiveness based on speed or breadth. Law Street Media is a niche publisher of law and policy topics for a millennial audience. It doesn't have a big—or really any—video production staff, but it has been able to use the Wibbitz semi-automated production tool to adapt content for a visuals-seeking audience and thus to expand its reach. Wibbitz, the toolmaker, makes money through revenue sharing. "Instead of having … 400 text articles written every day and only having a video on twenty of them, they [publishers] can now have video on 90 percent of them," Zohar Dayan, the founder of Wibbitz told me. Expanding the breadth of video content creates more advertising inventory, and in some cases Wibbitz gets a cut of the revenue that the advertising creates. Value accrues both to the creator of the automation tool and to the publisher.

Automated content can also create business value by reducing or shifting labor costs. Sören Karlsson, the CEO of United Robots, estimated that their soccer recap-writing software has saved one of its clients about $150,000 in a year. Some freelancers were no longer needed to call arenas to find out how each game turned out, for instance. This is more of a shift of labor than anything else since the effort (and cost) needed to collect the data has been subtracted away from each local newsroom and centralized by a new business entity that itself employs people to gather the data. There can be some cost savings by media organizations looking to shift labor onto perhaps more efficient and specialized technology provider companies.

Key business metrics can also be favorably impacted by automated content. The *Washington Post* found that automatically produced articles can drive a substantial number of clicks. The paper published about 500 automatically written articles during the 2016 US elections, which in total generated more than 500,000 page views.[61] In China, Xinhuazhiyun, a joint venture between Alibaba and Xinhua News Agency, created an automated soccer video highlight system and deployed it during the World Cup in 2018. The 37,581 clips produced, including things like goal highlights and coach reactions, generated more than 120 million views in total. Klackspark's automated articles are also generating

new traffic for Östgöta Media: the site had about 30,000 unique visitors per month in early 2017, and in an indication of a higher level of engagement, visitors now read more articles per visit on Klackspark. These numbers have made it fairly easy to attract sponsorship for the site from organizations that want to speak to soccer-loving audiences. Klackspark is harnessing the attention and value the automation provides to move toward a subscription model, where access to the site is packaged with access to one of Östgöta Media's local media properties. It's essentially seen as a value-add to the local news package, along the lines of: "Subscribe to our local paper and get access to this great soccer-news resource as part of the deal." Automated articles can also help drive subscription conversion directly. MittMedia's automatically produced real estate articles have driven hundreds of new subscriptions in the first few months of operation.[62]

Automation can, in some cases, also enhance the visibility of content on search engines such as Google. To the extent that search engine ranking criteria are known, these factors can be baked into how content is automatically produced so that it is ranked more highly by the search engine. This provides more visibility and traffic to the content and, theoretically at least, more advertising revenue as a result. For instance, Google ranks editorial content more highly when a picture is included. Realizing this, the German automated content provider Retresco automatically selects an image or picture to go along with the written football recap articles it generates.

Automated content also creates new opportunities for products. In the future perhaps we will see native content (that is, content sponsored by an organization with a stake in the topic) mass-localized for different markets using automation. But as with any new technical advance, designers should ask if an innovation comes into tension with desired values. Automation for scale appears to assume a "more is better" standard for news content. In the case of financial news at the AP, that has helped liquidity around stocks. But is more content *always* better—for society, or for business? For that matter, is more speed or personalization always better? Practitioners will need to weigh the benefits and trade-offs as they balance business and journalistic commitments.

Barriers to Adoption

Innovation in automated content continues to push the boundary of what can be accomplished by taking advantage of speed, scale, accuracy, and personalization. And while there will be a growing number of use cases where these benefits

make sense, either financially or competitively, a range of drawbacks will prevent a fully automated newsroom of the future. The limitations of content automation include a heavy reliance on data, the difficulty of moving beyond the frontier of its initial design, cognitive disadvantages such as a lack of ability to interpret and explain, and bounded writing quality. Some of these limitations, such as the dependency on data, are inherent to algorithmic production processes, whereas other issues may eventually succumb to the forward march of technical progress as artificial intelligence advances.

Data, Data, Data

The availability of data is perhaps the most central limiting factor for automated content production. Whether numerical data, textual or visual media corpora, or knowledge bases, automated content is all about the data that's available to drive the process. Datafication—the process of creating data from observations of the world—becomes a stricture that holds back more widespread use of automation simply because aspects of the world that aren't digitized and represented as data cannot be algorithmically manipulated into content. The quality, breadth, and richness of available data all impact whether the automated content turns out compelling or bland.[63] Data also become a competitive differentiator: exclusive data mean exclusive content.[64] Automated content runs the risk of becoming homogenous and undifferentiated if every news organization simply relies on access to the same data feeds or open data sources. In the content landscape, news organizations may see new competitors (or collaborators) in organizations that already have deep and unique databases and can cheaply transform that data into content. Despite a reluctance among some news organizations,[65] the acquisition of quality numerical data streams and knowledge bases is relatively new territory where news organizations will need to invest to remain competitive.

Investment in data production could involve everything from sensor journalism—the deployment of cheap sensors to gather data—to creating new techniques for digitizing documents acquired during investigations.[66] Public records requests are a method many data journalists use to acquire data for their investigations; however it still remains to be seen whether automated content can be reliably built around this form of data acquisition. *Streams* of data are often the most compelling and valuable for automated content since they facilitate scale and amortization of development costs over time. But engineering a stream of data requires the entire chain from acquisition to editing and quality control to

be systematized, and perhaps automated as well. Data quality is essential but is still mostly reliant on iterative human attention and care.[67] News organizations able to innovate processes that can pipeline data acquired via public records requests or other sources will have a competitive advantage in providing unique automated content. The use of more sophisticated structured data and knowledge bases is another promising avenue for advancing the capabilities of automation.[68] The deliberate and editorially oriented design and population of event and knowledge abstractions through structured reporting processes allow for the meticulous datafication of particular events from the world. These data can then be used to drive more sophisticated narrative generation.

The textual outputs of an automated writing system are influenced by the editorial decisions about the data they are fed. There's a certain "algorithmic objectivity" or even "epistemic purity" that has been attributed to automated content—an adherence to a consistent factual rendition that confers a halo of authority.[69] But the apparent authority of automated content belies the messy, complex reality of datafication, which contorts the beautiful complexity of the world into a structured data scheme that invariably excludes nuance and context.[70] The surface realization of content may be autonomous or nearly autonomous, but editorial decisions driven by people or organizations with their own values, concerns, and priorities suffuse the data coursing through the system. The ways data are chosen, evaluated, cleaned, standardized, and validated all entail editorial decision-making that still largely lies outside the realm of automated writing, and in some cases demands closer ethical consideration.[71] For instance, standard linguistic resources that an automated writing system might rely on can embed societal biases (such as race or gender bias) that then refract through the system.[72]

Bias can emerge in automated content in several different ways. The bias of a data source relates to a host of issues that can arise due to the diversity of intents and methods different actors may use in the datafication process. The limitations of datasets must be clearly understood before journalists employ automation. Corporations and governments that create datasets are not creating them for the sake of objective journalism. Why were they created, and how are they intended to be used? Data can be easily pushed past its limits or intended uses, leading to validity issues.

Another issue is coverage bias. If automation has a strong dependency on the availability of data, then there may be a tendency to automatically cover topics or domains only if data are available. And even if data are generally available, a

missing row or gap could prevent coverage of a specific event. We already see this type of bias in terms of the initial use cases for automation: weather, sports, and finance all being data-rich domains. Other topics may not receive automated coverage if data are not available, leading to coverage bias. “Just because there are a lot of data in some domains, it doesn’t mean that that is actually something that is interesting or important to the society. We need to watch out so that you don’t let the data control what you would cover,” explained Karlsson from United Robots. Data quality also needs to be considered, since data journalists have been known to prefer datasets that are easily readable and error-free.[73] As a result automation editors may want to consciously consider how data availability and quality influence coverage using automation.

The data that feed automated content production are also subject to little-discussed security concerns that could affect quality. It’s not unreasonable to imagine bad actors manipulating or hacking data streams to inject malicious data that would result in misleading content being generated. As a simple example, if a hacker were able to make the data delivered to the AP for the earnings per share of a company appear to beat estimates instead of missing them, this could have an impact on investors’ trading decisions and allow the hacker to profit. The AP does have checks in place to catch wild anomalies in the data—such as a stock dropping by 90 percent—as well as means to alert an editor, but more subtle and undetected manipulations could still be possible. Additional research is needed to understand the security vulnerabilities of automated content systems.

Data are the primary and in many cases the only information source for automated content. As a result, datafication puts important limits on the range of content that can be feasibly automated. Data *must* be available to produce automated content, but sometimes information is locked away in nondigitized documents, difficult-to-index digitized documents such as handwritten forms,[74] or even more problematically, in human heads. If a beat, story, or topic relies on information tied up in human brains (that is, not recorded in some digital or digitizable medium), then interviews are needed to draw out that information. The current generation of automated journalism is most applicable when the story rests on information either directly in the data, or derivable from the data, but not outside of that data. Unless the input data captures all the nuances of a situation, which is highly unlikely, there will necessarily be context loss in the automated content. New ways to overcome context loss and lack of nuance inherent to datafication are needed in order to expand the range of utility for

automated content.

In cases where necessary information is coming directly from people, and when it may be "sensitive, complex, uncertain, and susceptible to misunderstanding, requiring intimacy, trust, assessment of commitment, and detection of lies,"[75] the interjection of reporters—people tasked with acquiring that information—will be essential. Handling sensitive leaks and meetings in dark alleys with shady sources will be the purview of human journalists for some time to come. Interviewing is relational. Even just getting access may depend on building a rapport with a particular source over time. Moreover, an interview may require not just blindly recording responses but instead demand adversarial push-back on falsehoods, follow-ups to clarify facts, or reframing unanswered questions to press for responses. For all of these reasons, automation may be deeply challenged in terms of its capacity to engage in meaningful journalistic interviews.[76]

But there are some hints for how automated content could be integrated with human question asking, such as when the AP has reporters do follow-up interviews to get a quote to augment some automated earnings stories. This pattern of collaboration could even be systematized by inserting gaps in templates where information acquired from human collaborators would be expected. For instance, a template might include an "insert quote from soccer coach here" flag and send a task request to a reporter to acquire a quote before publishing a story. In such mixed initiative interfaces an algorithm might prompt the reporter for certain inputs it deems necessary.[77] In one experiment potential sources were targeted on Twitter and asked a question about the waiting time in an airport queue (42 percent of users even responded).[78] But the gap between asking simple questions like this and the more substantive questions required of human reporting is vast. Meaningful research awaits to be done on developing automation that can ask meaningful questions and receive meaningful responses. Automation must learn to datify the world.

The Frontier of Design and Capability

Automated content production systems are designed and engineered to fit the use cases for which they're deployed. Rule-based automation is particularly brittle, often functioning reliably only within the boundaries of where its designers had thought through the problem enough to write down rules and program the logic. In certain domains such as sports, finance, weather, elections, or even local coverage of scheduled meetings, there is a fair bit of routine in journalism. Such

events are often scheduled in advance and unfold according to a set of expectations that are not particularly surprising. Sports have their own well-defined rules and boundaries. It is in routine events like these where an automated algorithm, itself a highly structured routine, will be most useful. But, almost by definition, routine events in which something nonroutine happens could be extremely newsworthy—like say, stadium rafters collapsing on a soccer match. Unlike automated systems, human journalists are adept at adapting and improvising in cases like these.[79]

Automation fundamentally lacks the flexibility to operate beyond the frontier of its own design. Automated systems don't know what they don't know. They lack a meta-thinking ability to see holes or gaps in their data or knowledge. This makes it difficult to cope with novelty in the world—a severe weakness for the domain of news information given that novelty is quite often newsworthy. When the Norwegian News Agency implemented its soccer writing algorithm, there was a lot of hand-wringing about what to do in outlier situations. At the end of the day the decision was to simply not create an automated story for events that had some kind of exceptional, out-of-bounds aspect to them. But that kind of recognition itself requires human meta-thinking; human oversight must be built into the system.

The fragility of automated content production algorithms also leads to reliability issues. Yes, once it's set up, an algorithm will run over and over again in the same way. It is reliable in the sense that it is consistent. But if the world changes even just a little bit, such as a data source changing its format, this could lead to an error. Reliability in the sense of being dependable is an issue for adoption of automated systems because it impacts trust.[80] This introduces limits on how far news organizations are willing to push automation: "The more sophisticated you try to get with it, the more you increase the chance of error," said Lisa Gibbs from the AP.[81] In high-risk, high-reward use cases such as finance, which are built around a need for speed, organizations such as Reuters have double and triple redundancy on their automation, including in some cases a human fallback.

The resilience of people allows them to accommodate error conditions and make corrections on the fly. Automated systems, by contrast, continue blindly powering forward until an engineer hits the kill switch, debugs the machine to fix the error, and then restarts the process. As a result, automated content needs editors who can monitor and check the reliability of the process on an ongoing basis. Oftentimes the errors that crop up with automation are data-related,

though sometimes they can also be introduced by algorithms that aren't quite up to the task.[82] Ben Welsh at the *LA Times* described an error that their earthquake reporting system made. It published a report based on USGS data, but then ten minutes later the USGS sent a correction saying it was a ghost reading caused by aftershocks in Japan. The *LA Times* updated the post and then wrote a second story about ghost earthquakes that are sometimes incorrectly reflected in the USGS data. Errors of this ilk keep cropping up, though: in June, 2017 the Quake Bot relayed another bogus report from faulty USGS data, this time because a date-time bug at the USGS sent out a faulty alert.[83] These instances point out the weakness of relying on a single, untriangulated stream of data, albeit one that is often reliable. Similar data-driven errors have also arisen at the AP, including a prominent error made in July 2015, when one of its earnings reports erroneously indicated that Netflix's share price was down, when instead it was up. The reason for the error was traced to the fact that the company had undergone a 7–1 stock split, but the algorithm didn't understand what a stock split was.[84] Thankfully, outlets such as the *LA Times* and AP have human oversight and corrections policies that help to mitigate or remediate such algorithmic errors.

Errors in data aside, algorithms have gotten reasonably proficient at automated content creation within the bounds of their design. But they also still have difficulty with more advanced cognitive tasks. Beyond simply describing *what* happened, algorithms struggle with explaining *why* things happened. They have difficulty interpreting information in new ways because they lack context and common sense. In the nomenclature of the classic six Ws of journalism—who, what, where, when, why, and how—automation is practicable with *who, what, where,* and *when,* particularly when given the right data and knowledge bases, but it still struggles with the *why* and *how,* which demand higher-level interpretation and causal reasoning abilities.

Generating explanations for why something is happening in society is daunting, even for people. While data-driven methods do exist for causal inference, it remains to be seen whether they rise to the level of reliability publishers would require in order to automatically publish such inferences. It's more likely that any explanation produced by an automated system would be further vetted or augmented by people. Explaining a complex social phenomena may demand data, context, or social understanding that is simply inaccessible to an automated content production system. One of the critiques of mass localization has been that if statistics aren't contextualized by reporters with local knowledge, articles can miss the bigger social reality.[85] Entire strands of

journalism, for instance those pertaining to cultural interpretation, may also be difficult for automation to address.[86] The lack of commonsense reasoning prevents systems from making inferences that would be easy to make for a human reporter, which further curtails their usage outside of the narrow domains where they've been engineered with the requisite knowledge.

One specific type of knowledge and reasoning that automated content still lacks is the legal kind. As a result, algorithms have the potential to violate media laws without realizing it. Consider for a moment the possibility of algorithmic defamation, defined as "a false statement of fact that exposes a person to hatred, ridicule or contempt, lowers him in the esteem of his peers, causes him to be shunned, or injures him in his business or trade."[87] In order for a statement to be considered defamation in the United States, it must be false but be perceived as fact and must harm the reputation of an individual or entity.[88] The United States has some of the most permissive free speech laws in the world, and for a defamation suit to hold up in court against a public figure, the statement must have been made with "knowledge of its falsity or reckless disregard for its truth or falsity"—the "actual malice" standard.[89] Presumably this could be proven if a programmer or news organization deliberately acted with malice to create an automated content algorithm that would spew defamatory statements. But there's a lower standard for defamatory speech against private individuals. If a news organization publishes an automatically produced libelous statement against a private individual, it could be liable if it is shown that the organization was negligent, which might include failing to properly clean data or fact check automated outputs.[90] The potential for algorithmic missteps that lead to legal liability is yet another reason to have human editors in the loop, or at the very least to engineer systems away from being able to make public statements that could hurt the reputation of private individuals.

Writing Quality

Despite being able to output perfectly readable and understandable texts, the quality of automated writing still has a way to go before it can reflect human-like standards of nuance and variability, not to mention complex uses of language such as metaphor and humor.[91] Variability is a key concern, especially for content domains such as sports, where it's likely for a single user to consume more than one piece of content at a time. Repetition could lead to boredom. "You really need to be able to express the game outcome in many different ways. Otherwise you will see that this is automated text and you will find that this is

repeating itself," explained Karlsson from United Robots. In domains such as finance there may be advantages to keeping the language straightforward and more robot-like—verbal parsimony and repetition can straightforwardly communicate essential facts to users—whereas in domains such as sports users may want more entertaining, emotional, or lively text.

Text quality of course varies greatly with the method used to automate the writing. Template-based methods have the highest quality because they encode the fluency of the human writer of the template. Yet templates also reduce text variability, increasing repetition and the potential to bore readers. Studies have begun to examine the perception of automatically written texts and found that readers can't always differentiate automatic from human-written texts but that there can be substantive differences in readability, credibility, and perception of quality.[92] An early study found that human-written articles were more "pleasant to read" than their automated counterparts.[93] A more recent study of more than 1,100 people in Germany found that computer-written articles were scored as less readable, but more credible than their human-written counterparts.[94] The difference in readability was substantial: the mean rating for the human-written articles was 34 percent higher than for computer-written articles. But the absolute differences for credibility were quite small, only about 7 percent. Another study of 300 European news readers found that message credibility of sports (but not finance) articles was about 6 percent higher for automated content.[95] In Finland, a study of the Valtteri system compared automatically generated municipal election articles to journalist-written articles and found that the automated articles were perceived to be of lower credibility, likability, quality, and relevance.[96] Feedback from the 152 study participants suggested that the automated articles were boring, repetitive, and monotonous, as well as that there were grammatical mistakes and issues with the writing style. Unlike previous studies which evaluated texts generated using templated sentences, Valtteri used phrase-level templates, phrase aggregation, referring expression generation, and automated document planning based on the detection and ranking of newsworthiness of facts to include in the story. In other words, it's a more complex NLG engine. But this complexity appears to introduce the potential for grammatical errors that may undermine the credibility of texts. Moreover, the phrase-level templates used didn't provide enough variability to make the texts interesting to users.

We can contrast template-driven automated content of perhaps passable (if not entirely compelling) quality to even lower-quality statistical NLG

approaches. Let's look at an abstractive summarization system trained on *New York Times* articles using a machine-learning technique.[97] Researchers asked human readers to rate the readability of the summaries produced in an evaluation of the system. Here is one example of the output of the model that produced summaries with the highest readability of those tested, run on an input article about a Formula One car race:

> Button was denied his 100th race for McLaren. The ERS prevented him from making it to the start-line. Button was his team mate in the 11 races in Bahrain. He quizzed after Nico Rosberg accused Lewis Hamilton of pulling off such a manoeuvre in China.

Though somewhat recognizable as topical text, the summary has awkward verbs, such as "quizzed," which do not fit the context, it doesn't explain acronyms, and it has fragments of nonsensical text.

The fluency of textual output produced by simple templates (see Figure 3.1), as well as being more straightforward to understand in terms of concrete rules, helps explain its adoption by news agencies in lieu of statistical methods.[98] Publishing garbled text consisting of ungrammatical nonsense would cast a long shadow on the credibility of a news outlet. If more advanced machine-learning approaches are to be integrated into content production, additional technical advances are needed.

Automated Content Is People!

In 2016 the White House predicted that as the result of advances in artificial intelligence, "Many occupations are likely to change as some of their associated tasks become automatable."[99] So what does automation mean for human tasks in news production? Even in scenarios where content is produced entirely autonomously with no human in the loop,[100] there is still a lot of human influence refracted through the system. Automation is designed, maintained, and supervised by people. If people are inserted at key points in the production process, they can oftentimes compensate for the shortcomings of automation described in the last section. And so determining how best to infuse or blend human intelligence into the hybrid process is a key question. "The value of these automated systems is the degree to which you can imbue them with editorial acumen," explained Jeremy Gilbert at the *Washington Post*. Here I consider how that editorial acumen is expressed via the design, development, and operation of automated systems, as well as what this means for how human roles may change.

Design and Development

The design and development process for automated content production systems is rife with opportunities for baking in editorial decision-making according to the editorial values and domain knowledge of whoever is involved in that design process, including reporters, editors, computational linguists, data scientists, software engineers, product managers, and perhaps even end users. Values become embedded into systems through everything from how constructs are defined and operationalized to how knowledge bases and data are structured, collected, or acquired, how content is annotated, how templates or fragments of text are written and translated for various output languages, and how general knowledge about the genre or domain comes to be encoded so that it can be effectively utilized by an algorithm.

One of the key editorial decisions that automated content systems must make is how they define what information is included or excluded. Whether that depends on some notion of "newness," "importance," "relevance," "significance," or "unexpectedness," data scientists will need to carefully define and measure that construct in order to use it in the algorithm. Decisions about these types of editorial criteria will have implications for how people perceive the content. "Some critics of the Quake Bot think it writes too many stories about smaller earthquakes," Ben Welsh explained. This was the result of a decision by the city editor, who simply declared that the bot should cover everything that's a magnitude 3.0 and higher earthquake in the Los Angeles area. These types of human decisions get enshrined in code and repeated over and over by the automation, so it's important to think them through carefully in the design process.[101] Otherwise different or even conflicting notions of newsworthiness may become unconsciously built into how algorithms come to "think" about such inclusion and exclusion decisions. It's often necessary for people involved in the design of these systems to be able to make their implicit judgments explicit, and with clear rationale.

People also heavily influence automated content production in terms of the creation and provision of data. For unstructured data or for systems relying on machine learning, human annotation of data is often an important step. Structured annotations are a key enabler of algorithmic production. For instance, automated video creation systems need a corpus of photos that is reliably annotated so that when a person is referred to in the script, a photo of the correct person can be shown. In early systems for producing automated video such annotations were supplied by people using controlled vocabularies of

concepts.[102] Newer systems, such as those of Wochit and Wibbitz, rely on vast corpora of content that have been tagged by professional photo agencies. In cases where yet more classification algorithms are used to automatically tag a photo or video with concepts, those algorithms have to be trained on data that was originally judged by people as containing those concepts. Automated systems are built atop layers and layers of human judgment reflected in concept annotations.

Domain-specific and general knowledge reflecting an understanding of the genre, narrative style, choice of words, or expectations for tone also need to be encoded for automated systems. For specific content areas, such as politics, sports, or finance, experts need to abstract and encode their domain knowledge and editorial thinking for use by the machine. For instance, semantic enrichment of valence about a particular domain might allow an automated system to write that a trend for a particular data series "improved" instead of just "increased." The meaning of words may also depend on constantly evolving knowledge of a domain: for example, in a story about election polls, the meaning of terms that are not well defined, such as "lead," "trend," or "momentum," could depend on context.[103] Retresco keeps former editors on staff so that different genres of text can be matched in terms of choices in grammar and word choice. Tom Meagher at the Marshall Project explained how the database they designed to support the automation in their "Next to Die" project was influenced by partners who had spent dozens of years covering capital punishment in their own states and as a result knew what factors to quantify for the automation. Wibbitz has an in-house editorial team of video editors and ex-journalists who work with research and development groups to incorporate their knowledge into the technology. "That's one of the most important parts of the company," according to its founder, Zohar Dayan. A key enabler of the next generation of automated content systems will be better user interfaces that will allow experts to impart their journalistic knowledge and domain expertise to the system.

Collaboration in Operation

Once an automated content production system has been designed and developed, it moves into an operational phase. In this stage as well, people are in constant contact and collaboration with the automation as they update it, augment it, edit it, validate it, and otherwise maintain and supervise its overall functioning. There are some tasks humans can't do, there are other tasks that automation can't do, and so a collaboration is the most obvious path forward. This raises important

questions about the nature of that human-computer interaction and collaboration, such as how to ensure common ground and understanding between a human user and the automated system, how to monitor the automation to ensure its reliability or to override it, how contributions by people and computers can be seamlessly interleaved, and when and how trust is developed when systems are fairly reliable but not 100 percent reliable.[104] As people interact with these systems, the nature of their skills, tasks, roles, and jobs will necessarily evolve, most likely to privilege abstract thinking, creativity, and problem-solving.[105]

Given the preponderance of template-driven approaches to automated content, writing is one of the areas where people will need to evolve their craft. Template writers need to approach a story with an understanding of what the available data *could* say—in essence to imagine how the data could give rise to different angles and stories and delineate the logic that would drive those variations. Related is the ability to transform the available data and organize it more suitably or advantageously to the way it will be used. Gary Rogers of Urbs London, who creates templates to tell stories using open data in the United Kingdom, explained it this way: "When you're writing a template for a story, you're not writing the story—you're writing the potential for every eventuality of the story." The combinatorics of the data need to be understood in relation to what would be interesting to convey in a story, something that is closely related to the computational thinking concept of parameterization described in Chapter 1. People will also need to be on hand to adapt or reparameterize story templates in view of data contingencies or shifts in data availability for different contexts, such as a national soccer match with lots of rich data and a local soccer match with a relative paucity of data. Available interfaces for template writing don't yet support writers in seeing the multitude of possibilities in a given dataset, or in parameterizing a story in terms of the available data.

But user interfaces are, in general, evolving to better support the template-writing task. Automated Insights, whose technology drives the AP earnings reports, markets a tool called "Wordsmith," while Arria, which is used by Urbs Media, has a tool called "Arria Studio." These are essentially user-interface innovations wrapped around an automated content production algorithm. You could think of it as an alien word processor that lets the author write fragments of text controlled by data-driven if-then-else rules (see Figure 3.4). Automating the document planning and microplanning stages of NLG is complex and demands a fair amount of both engineering and domain knowledge. These interfaces offload the most complex of the document planning and

microplanning decisions to a human writer who authors a meta-document of rules and text fragments. Updating and maintaining a portfolio of templates this way, however, requires a fair bit of new editorial work.

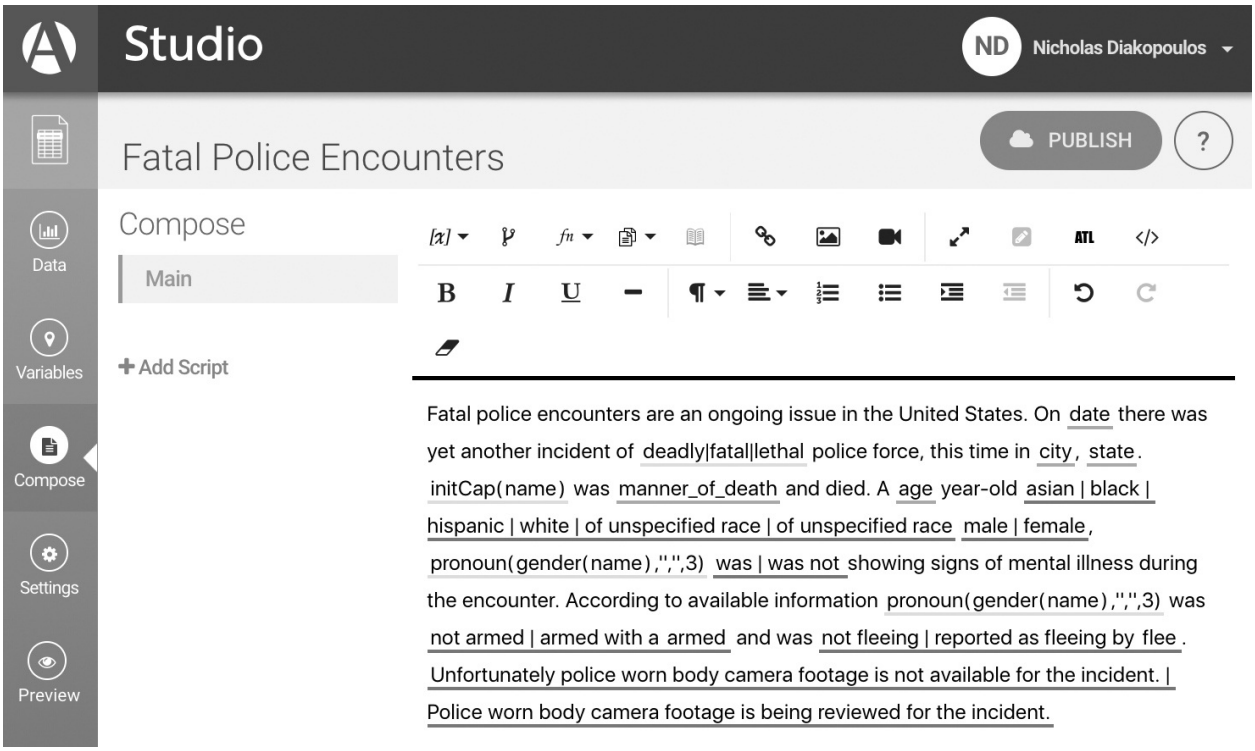

Figure 3.4. The Arria Studio user interface showing a story drafted using data from the *Washington Post* fatal police shootings database. Note the top row of buttons in the toolbar, used for inserting variables, conditional statements, synonyms, or other computational functions to transform the text. An alternative view provides direct access to the templating language. At the left are tabs that allow the user to explore the data, to choose among variables for the purposes of writing, and to preview the template's output with different rows of test data. *Source:* Arria Studio (https://app.studio.arria.com/).

If there's any area where there will be steady growth in employment, it's in the various shades of ongoing maintenance work that automated content systems require. People need to be involved in remediating errors that crop up in automated outputs, for instance. This may very well involve some exciting and creative detective work in debugging the system. But then there is the more prosaic maintenance: making sure the data streams are updated when the soccer leagues announce changes to teams, editing databases or spreadsheets to reflect new corporate details, keeping track of new or updated open data sources, or tweaking any of the myriad rule-sets for structure, genre, or style that have been baked into the design of the system. Lisa Gibbs at the AP told me that their

introduction of automated stories has led to new tasks for her colleagues: “One of the responsibilities of an editor on our breaking news desk now is essentially automation maintenance,” which involves tracking when companies move headquarters, split their stock, and so on. She estimates this takes about two hours per month of staff time, so it’s still a small slice of effort, but you could imagine that growing as more and more automation comes online. When the *LA Times’* Quake Bot missed a noticeable earthquake out in the middle of the Pacific Ocean, the geographic bounding box of the bot had to be adjusted so that its filter stretched farther out into the ocean. Updates like this will be needed as automated content is put through its paces and exposed to some of the corner or edge cases that designers maybe didn’t capture in their initial rules. From a business perspective, automated content probably won’t be competitive if it’s static for too long. People will need to be involved in continuously assessing and reassessing how it could be improved.

Supervision and management of automated systems are also increasingly occupying humans in the newsroom. As systems are developed, automation editors are needed to assess the quality of the content until it is good enough to publish, ideally foreseeing and rectifying potential errors in the design process. Regarding the development of soccer article-writing software, Helen Vogt from the Norwegian News Agency said, “I think we had at least 70 or 80 versions before we were ready to go live. And they [the sports editors] had a lot of fun testing it and tweaking it.” In the production of its Opportunity Gap project, ProPublica’s Scott Klein noted that editing 52,000 automatically written stories became a new challenge: “Editing one narrative does not mean you’ve edited them all.”[106] Sensible edits in one case might not entirely fit in another. Automated checks can contribute to quality assurance, but the provision of high-quality text still requires that representative samples of outputs be manually assessed.[107] At the RADAR project about one in ten stories is still manually checked to ensure that the raw data and story flow are accurately reflected in outputs. Much as methods for “distant reading” have helped digital humanities scholars understand corpora of texts via computational methods and visualizations,[108] building tools to support the notion of “distant editing” could allow editors to work on a much larger scale of texts.

There’s a distinct quality control or editing function that people must contribute during development of automation. People must also be available to make and post corrections in case the automation doesn’t behave as expected once it’s put into operation. When the *Washington Post* ran its automated

elections articles for the 2016 US elections, people were tasked with monitoring some of the stories using tools like virtual private networks (VPNs) to simulate loading the stories from different locations in order to see if the articles were adapting as expected. Tool makers have noted that a barrier to the adoption of personalized content is the effort needed to edit and check consistency of the many potential outputs.[109] As newsrooms expand their use of automation, people will need to keep an eye on the big picture—to know when to deploy, decommission, or redevelop a system over time.

Shifting Roles

The collaborations that emerge between journalists and automated content systems may lead to deskilling (loss of skills in the workforce), upskilling (an increase in skills to meet new demands from sophisticated tools), reskilling (retraining), or otherwise shift the role of human actors as they adjust to a new algorithmic presence.[110] Maintenance work related to updating datasets and knowledge bases, or annotating media, may not ultimately be that sexy, cognitively demanding, or high skill. This could have the effect of creating more entry-level jobs for which there would be more labor supply and lower wages. At the same time, automation can offload and substitute some of the tedium of, say, determining data-driven facts for a story. This may create more time for high-skilled journalists to do what they're trained for, namely reporting, finding and speaking to new sources, and writing in creative or compelling ways.[111] The need for high-skilled reporters is reinforced by the shortcomings of the current state-of-the-art technology, which cannot get at information where there are no data or answer questions such as why and how an event fits into the social fabric of society. The design, development, configuration, and supervision of automation may entail the need for a different kind of high-skilled individual as well—one who has mastered computational thinking, is highly creative, and understands the limits of the technology.

The reskilling of the existing journalistic workforce will involve learning tasks such as how to write templates to feed the automation. It will also include some light coding or at least familiarity with coding. "You have to be able to say here's a question a reporter would ask, or a framing of a story, or an angle and how do I encode that in computer code, or how could computer code ask and answer that question on its own, and return an answer that can then be republished," Welsh told me. Writing templates demands domain knowledge and a fluency in writing for that domain, as well as an essential understanding of the

possibilities of what the technology affords. At the *Washington Post*, Jeremy Gilbert observed that more experienced editors are often better at writing the templates because they already have experience helping to shape the writing of reporters and so they "have a better sense of structurally what is or should be." A structured journalism pilot project asked ten reporters to populate event and knowledge data structures, and found that the key skill that enabled some of those reporters to adapt quickly was a "general comfort with abstraction."[112] But this is skill enhancement, not replacement; reporting skills are just as important as ever. On the "Next to Die" project at the Marshall Project Tom Meagher explained that for structured journalism "the core task is keeping the database up to date and adding every new change to a case," but that "in some ways the reporting is almost exactly the same." In other words, reporting skills are still necessary, but the output of those reporting skills is to a structured database that then feeds the automated output.

These shifts are already to some extent reflected in new roles, both at the deskilled and the upskilled ends of the spectrum. Wibbitz claims that the results of its automatic process are a "rough cut" that's about 80 percent of the way there. People manage the last 20 percent, which involves fine tuning the story, swapping images based on relevance, or adjusting the transitions or caption timing. The people who do this don't have to be highly skilled video editors—one media organization I spoke to referred to the person tasked with editing the videos all day as "junior staff," an intern. On the other hand, the sports reporter who worked on the design of the first soccer story-writing software at the Norwegian News Agency decided to take on a new role as "news developer," and he now looks at tools, workflows, and opportunities where automation can provide value to the organization. Reuters has an automation bureau chief who is similarly tasked with both thinking through how to manage the existing automation in use and with envisioning new possibilities. And at the AP, a new position was created in 2014 for an automation editor who engages in much of the supervision, management, and strategy around automation use by the agency. These roles need editorial thinking as much as they need a capacity to understand data and the capabilities of state-of-the-art content automation technology. In other words they could be appropriate for a reskilled journalist or a reskilled data scientist. The implications of labor shifts will be borne out over time. Future research will need to observe how automated content production affects the composition of deskilling, reskilling, and upskilling and how that in turn impacts issues such as worker autonomy and skills education.[113]

Whither Automated Content?

Automated content production is still in a nascent state. It's beginning to gain traction as different organizations recognize the potential for speed and scale, but personalization is still largely on the horizon. There remains untapped communicative potential for combining text, video, and visualization in compelling ways. As limitations on data provision and automation capabilities give way to advances in engineering, what should we expect for automated content moving forward? Where should organizations look for opportunity as the technology advances?

Genre Expansion

One of the criticisms of automated journalism is that it's less "automated journalism" than "automated description."[114] A typology of journalistic forms is useful for seeing where else automation could expand. One such typology draws on rhetorical theory and offers five types of journalistic output: description, narration, exposition, argumentation, and criticism.[115] Description is essential in all of these modes and is well-suited to automation, given that the state of the world can be rendered factually based on collected data. For aspects of description that need to be more phenomenological, human collaborators can step in to add "flavor" such as quotes or other human-interest observations. Narration builds on description but offers more structure in presenting an event, perhaps even making it come across as more of a story by establishing connections between facts, events, and characters. This is what companies such as Narrative Science strive for: to be smart about document planning so that description starts to look more like narrative. The current state of the art in content automation gives us description and narrative.

Exposition starts to be more challenging. Exposition seeks to explain, interpret, predict, or advise and is familiar from formats such as op-ed columns or news analyses. This type of composition is harder for automation to render. Notable research prototypes of heavily engineered systems have, however, created expository essays or documentaries.[116] As explanatory AI becomes more sophisticated and knowledge structures are engineered to incorporate causality, we may see production-level automated content integrating aspects of explanation. Still, some aspects of exposition, such as commentary flavored by individual opinion, will remain outside the purview of automation, or will extrude human commentary that is parameterized and amplified through the machinery of the system. User-interfaces could allow for an opinionated

configuration of the content production pipeline. Personal interpretations of the definition of newsworthiness, for instance, as well as the word choices, templates, and tone used could then infuse what would be a more overtly subjective exposition.

Argumentation goes beyond exposition by adding the purpose of persuasion, of trying to sway a reader's attitudes and ultimately behaviors. Provided that the argument is coming from a human author and could be encoded in a template through human-driven document planning and that there is data that supports the argument, there are no real barriers to producing automated content that reflects that argument. Automated argumentation is an area that is ready to colonize, though perhaps less so by traditional journalists than by other strategic communicators.

A strong human collaborator enables automation to also participate in the composition of criticism. Criticism relies on the exercise of personal judgment and taste in appraising some newsworthy object or event. This is precisely what sophisticated sports commentators do when they critique the performance of a team or athlete. One example in this direction is a project from *Der Spiegel* that automatically produces tactical data visualizations, or maps of how a particular soccer team has performed.[117] But the maps are practically unpublishable on their own. A squad of sports reporters interpret the maps and add their own knowledge and critique of the game as they view the automatically produced visualization through their own critical lens. The automated visualization becomes a helpful tool for grounding a particular form of criticism.

Topical Opportunities

There is certainly room for automation to explore opportunities in exposition, argumentation, and criticism, as well as to deepen and expand descriptive and narrative capabilities. But there are also opportunities to apply existing techniques to new topics. We've already seen automation gaining traction in areas such as sports, finance, weather, and elections where data are available. Where else might there be demand for automated content if only data were available? And where is there enough of a potential for scale in terms of either raw scale of interest in publication of an event by different publishers or for scale over time because an event is routine or repetitive?

To examine these questions I gathered data from Event Registry, a content aggregator that as of mid-2017 had slurped up more than 180 million news articles from more than 30,000 news publishers worldwide. You might have

guessed that Event Registry is about *events*. The database clusters news articles that are about the same event, not unlike how Google News groups articles. Event Registry then enriches these events by automatically tagging them with categories—high-level groupings such as “Health” or “Science” but also much more fine-grained ones such as “Climate Change.” I collected all of the events from the Event Registry database for June 2017—more than 20,000—that referred to events in the United States and had at least some content in English. From these I tabulated the dominant categories of content based on number of events to see which categories show promise for scale over time. I also looked at the average number of articles per event to see which categories might benefit from scaling coverage over a range of different audiences.

Table 3.1 lists the top categories of content according to the number of events observed. Categories such as “Law Enforcement,” “Crime and Justice,” “Murder,” “Sex Offenses,” and “Fire and Security” all clocked hundreds to thousands of events over the course of a month, with the average number of articles per event being in the range of fifteen to twenty. The dominance of these categories is perhaps unsurprising given the old journalism adage, “If it bleeds, it leads.” It is also no surprise that categories relating to sports hit a sweet spot across the board—a great variety of sports-related topics received both routine and fairly widespread coverage. These include American favorites such as “Baseball,” “Professional Basketball,” and “Ice Hockey” as well as new classics such as “Competition Shooting” and “Auto Racing.” “Martial Arts” was another standout: there were about forty-six articles on average for each of the events, suggesting an opportunity for breadth. One surprise in the data was the apparent extent of media produced about video game events, which includes e-sports, video game tournaments, and fantasy sports. However, upon closer inspection, it appears Event Registry miscategorized some regular sporting events as “Video Game” events making it difficult to say how prevalent coverage of video game events really was.

Besides crime, sports, and potentially games there are several other categories of content that could take advantage of the potential for scale. Service-oriented topics related to transportation such as “Aviation” are promising, with 118 events and about 21 articles per event, as is “Health” with 178 events and 16 articles per event. Policy-oriented topics also have a potential for scale: “Immigration” had 109 events with an average of 19 articles per event, “Environment” had 103 events also with 19 articles on average, “Economic” issues had 79 events with an average of 24 articles, and “Campaigns & Elections” had 110 events with 21

articles on average. There is a range of openly available data related to economic issues, as well as for immigration and the environment, suggesting that these topics could perhaps be covered using automated content techniques. In general, however, cataloging the available data resources for each of these topics would be necessary to identify where the best opportunities lie.

Table 3.1. Categories of news in Event Registry with more than 100 distinct events in June 2017.*

Category	Event Count	Mean Article Count	Median Article Count
Society / Law / Law Enforcement	1,440	17	8
Reference / Education / Colleges & Universities	767	15	8
Business / Business_Services / Fire & Security	657	14	8
Games / Video_Games / Recreation	564	21	10
Games / Video_Games / Browser_Based	495	30	11
Society / Religion & Spirituality / Christianity	447	15	8
Society / Issues / Crime & Justice	389	18	8
Sports / Football / American	381	15	10
Society / Law / Services	355	26	9
Society / Law / Legal Information	235	30	9
Business / Industrial Goods & Services / Industrial Supply	221	13	7
Sports / Basketball / Professional	194	31	11
Recreation / Guns / Competition Shooting	190	31	8
Health / Public Health & Safety / Emergency Services	188	12	8
Society / Issues / Health	178	16	8
Shopping / Sports / Baseball	178	25	9
Home / Family / Parenting	176	14	10
Sports / Hockey / Ice Hockey	174	16	10
Arts / Performing Arts / Theater	160	24	8
Science / Biology / Flora & Fauna	155	18	8
Sports / Baseball / Youth	154	15	8
Society / Issues / Business	152	41	10
Society / Crime / Murder	148	17	8
Games / Video_Games / Downloads	140	25	10
Society / Issues / Transportation	139	15	7
Computers / Software / Accounting	129	20	9
Society / Crime / Sex Offenses	128	46	7
Society / People / Generations & Age Groups	120	14	7.5
Society / Issues / Warfare & Conflict	120	20	10
Business / Transportation & Logistics / Aviation	118	21	8

Computers / Internet / On the Web	117	25	11
Science / Technology / Energy	114	13	8
Sports / Motorsports / Auto Racing	114	19	8
Reference / Education / K through 12	113	10	8
Society / Politics / Campaigns & Elections	110	21	8
Society / Issues / Immigration	109	19	10
Arts / Television / Programs	108	16	8
Business / Energy / Utilities	108	18	8
Society / Issues / Environment	103	19	9
Recreation / Pets / Dogs	103	21	9

* Each category is shown with the mean and median number of English articles published for each event.

Of course, there are limitations to Event Registry data in terms of coverage and the extent of representation of media. Also, the numbers noted above deal only with the supply side of content, and the strategic deployment of automated content should also consider reader demand. But these data do provide hints about what categories or topics might be promising for automation. It's not an entirely unreasonable assumption that the coverage of these topics is in response to demand that's been expressed in some way by media consumers. The numbers suggest that there may still be opportunities for finding underexplored niches where data availability, technical capability, editorial interest, and some form of need for scale—either over time or for breadth—all intersect.

Automated content production offers a host of new opportunities to increase the speed, scale, accuracy, and personalization of news information, some of which are already creating revenue and competitive advantages for news organizations. At the same time, human roles and tasks are evolving to accommodate the design, maintenance, and supervision of automated content systems. Far from a panacea for news production, there are a host of limitations that go along with content automation. A dependency on data, as well as bounds on flexibility, interpretive ability, and writing quality place real constraints on how widely automated news production can spread. Still, some domains of content may yet be colonized by automated news content, as journalists find where end-user demand intersects with technical capabilities and constraints. Another milieu where these intersections are increasingly being explored is social media, where automated content takes on a social face in the form of newsbots.