DIGITAL
JOURNALISM

Digital Journalism

ISSN: 2167-0811 (Print) 2167-082X (Online) Journal homepage: http://www.tandfonline.com/loi/rdij20

Clarifying Journalism's Quantitative Turn

Mark Coddington

To cite this article: Mark Coddington (2015) Clarifying Journalism's Quantitative Turn, Digital Journalism, 3:3, 331-348, DOI: 10.1080/21670811.2014.976400

To link to this article: http://dx.doi.org/10.1080/21670811.2014.976400

Published online: 07 Nov 2014.

CLARIFYING JOURNALISM'S QUANTITATIVE TURN

A typology for evaluating data journalism, computational journalism, and computer-assisted reporting

Mark Coddington

As quantitative forms have become more prevalent in professional journalism, it has become increasingly important to distinguish between them and examine their roles in contemporary journalistic practice. This study defines and compares three quantitative forms of journalism—computer-assisted reporting, data journalism, and computational journalism—examining the points of overlap and divergence among their journalistic values and practices. After setting the three forms against the cultural backdrop of the convergence between the open-source movement and professional journalistic norms, the study introduces a four-part typology to evaluate their epistemological and professional dimensions. In it, the three forms are classified according to their orientation toward professional expertise or networked participation, transparency or opacity, big data or targeted sampling, and a vision of an active or passive public. These three quantitative journalistic forms are ultimately characterized as related but distinct approaches to integrating the values of open-source culture and social science with those of professional journalism, each with its own flaws but also its own distinct contribution to democratically robust journalistic practice.

KEYWORDS big data; computational journalism; computer-assisted reporting; data journalism; expertise; open source; professional journalism; public; transparency

Introduction

Professional journalism has historically been built around two elements—textual and visual. Numbers have long had a role in journalism as well, but American journalists have consistently downplayed their importance in making up their professional skillset, leading to a notorious difficulty in presenting numerical data accurately and responsibly (Maier 2002). A notable exception has been the professional subfield of computer-assisted reporting (CAR), which has focused on journalistically analyzing quantitative data for at least 40 years. Over the past several years, this data-driven strain of journalism has become more prominent within the profession as it has converged with the increasingly ubiquitous digitization of information both personal and

Digital Journalism, 2015
Vol. 3, No. 3, 331–348, http://dx.doi.org/10.1080/21670811.2014.976400

public. As more information has become ones and zeroes at its most elemental level, more journalism has involved gathering, analyzing, and computing that information as quantitative data as well. Journalism appears to be taking, as Petre (2013) puts it, "a quantitative turn."

This wave of quantitatively oriented journalism has deep democratic roots; various forms of it are tied to open government advocacy (Parasie and Dagiral 2013) and the public-service tradition of investigative journalism (Cox 2000). It has great potential to broaden journalism's ability to make democratic institutions more responsive and legible to the public, but even within this sub-area of journalism, views of the public and the journalistic process are broadly disparate. Where the CAR of the 1990s was generally a single, unified concept for both professionals and scholars, the area has splintered into a set of ambiguously related practices variously termed by researchers computational journalism (Flew et al. 2012; Karlsen and Stavelin 2014), programmer-journalism (Parasie and Dagiral 2013), open-source journalism (Lewis and Usher 2013), or data journalism (Appelgren and Nygren 2014; Fink and Schudson 2014; Gynnild 2014), among others.

The journalists engaged in these practices seem particularly unconcerned with classifying their work *vis-à-vis* professional journalism, a sentiment most famously expressed in a short blog post by developer Adrian Holovaty (2009) that answered the question "Is data journalism?" with "Who cares?" This has resulted in several of the aforementioned terms being thrown together within professional discourse as synonyms. For researchers, however, these definitional questions are fundamental to analyzing these practices as sites of professional and cultural meaning, without which it is difficult for a coherent body of scholarship to be built. Indeed, the nascent scholarship in the area is often characterized by initial attempts to define these forms of journalism, each of which has largely been well-conceived and conceptually useful. But taken collectively, they have produced a cacophony of overlapping and indistinct definitions that forms a shaky foundation for deeper research into these practices. As these data-driven forms of journalism move closer to the center of professional journalistic practice, it is imperative that scholars do not treat them as simple synonyms but think carefully about the significant differences between the forms they take and their implications for changing journalistic practice as a whole.

Building on the work of Parasie and Dagiral (2013), Gynnild (2014), and Stavelin (2014) to delineate differences between these practices, this study is an attempt to develop a typology for analyzing forms within this quantitative area of journalism. It examines three professional practices—CAR, data journalism, and computational journalism—along four professional and epistemological dimensions. The analysis will begin with a brief discussion of the cultural background against which these practices are operating, then proceed with an introduction to the three practices, and finally an evaluation of each practice against each of the four dimensions.

Open-source Culture

These new forms of journalistic practice are emerging within an increasing interaction between programmers and journalists, as more programmers have begun to move into professional newsrooms and professional journalists have become increasingly drawn to programming's technical capabilities and cultural norms, which have been heavily influenced by the open-source movement.

The term "open source" as a technological principle was born in the late 1990s as a more palatable and widely accessible offshoot of the free software movement. Both movements focused on the ability to freely access, modify, and redistribute software as a manifestation of the universal right to access to information and knowledge (Coleman 2013; Kelty 2008). While open-source is intrinsically oriented not toward journalism but toward software, Lewis and Usher (2013) explained its application to journalism through four principles: transparency, iteration, tinkering, and participation. Each of those principles arises from the process of collaboratively building and sharing software, the practice at the core of the open-source software movement. And as Lewis and Usher explained, each is gradually becoming more prevalent within professional journalistic culture as a small subset of more computing-oriented journalists are drawn to the open-source ideals of creativity, experimentation, and liberation of information. In this way, the principles of open source have been an important common ground for bringing together "hacks" (journalists) and "hackers" (technologists).

Data-driven Journalism Practices

The three journalistic practices examined here are not mutually exclusive. Since they have very similar professional and epistemological roots, they will inevitably overlap, in some cases significantly. Actual cases of these practices will often display characteristics of more than one of these categories, as well as the marks of open-source principles. Key institutions have been involved in the perpetuation of more than one of these practices; for example, the National Institute for Computer-Assisted Reporting (NICAR) was the central organization in computer-assisted reporting during the 1990s and is now a central organization in connecting and training those who practice data journalism (Fink and Anderson 2014). In addition, many of the journalists who engage in these practices themselves tend to emphasize their continuity; data journalists generally characterize themselves as following in the same tradition as CAR. But there are significant differences between these forms of practice, and the following is an attempt to pull them apart and clarify them conceptually. This paper relies heavily on research into these practices within the United States and Scandinavia, since those have been the most thoroughly studied geographical settings for this work. It thus broadly describes the forms as they are generally practiced in those environments, though national and local variations certainly exist, both within these areas and outside them.

Computer-assisted Reporting

Though the use of computers in journalism dates back to the 1950s (Cox 2000), the *de facto* godfather of CAR is Philip Meyer, who outlined a new form called precision journalism in a book of the same name (Meyer 1973). Precision journalism was modeled after social science, using empirical methods (particularly surveys and content analysis) and statistical analysis to achieve more definitive answers to journalistic questions. It was not until the late 1980s and early 1990s that precision journalism, since recast as

CAR, began to make significant inroads into newsrooms, led by several high-profile, Pulitzer Prize-winning stories that became an important vehicle for professional validation (Houston 1996).

CAR became closely tied to investigative reporting, often being seen as an auxiliary tool to aid in long-term, public-affairs journalism projects (Cox 2000; Gynnild 2014; Parasie and Dagiral 2013). Though CAR journalists often fought against the perception that their practices were only for time-consuming investigative story packages—an association that may ultimately have limited CAR's adoption within professional journalism (Gynnild 2014), they also encouraged it at times, characterizing it as, in the words of one CAR pioneer, "the new investigative journalism" (Jaspin 1993). The term CAR has fallen out of favor since the early 2000s as its technology has broadly diffused throughout newsrooms; Meyer himself called in 1999 for the moniker to be retired, describing it as an "embarrassing reminder that we are entering the 21st century as the only profession in which computer users feel the need to call attention to ourselves" (Meyer 1999, 4). Meyer's call ultimately went unheeded, as CAR continues to be practiced in journalism, though it appears to be invoked more often as a historical mode of quantitative journalism than a contemporary practice. A comparison between CAR and data journalism or computational journalism, as this paper undertakes, is thus a characterization more of change in practice over time than a comparison of contemporaneous practices.

While CAR had its roots in social science-based statistical methods, it came to embody two sets of practices: the data gathering and statistical analysis descended from Meyer's precision journalism, and more general computer-based information-gathering skills such as online and archival research and even email interviews (Miller 1998; Yarnall et al. 2008). The more general information-gathering skills have become so elemental a part of journalistic work that they can no longer be considered, in Powers' (2011) terms, "technologically specific work," though the statistical- and data-oriented forms of CAR remain such because of their relative lack of diffusion. This is the form of CAR that this paper refers to with the term, and the one that serves as the foundation for the modern approaches of data journalism and computational journalism (Gynnild 2014).

Data Journalism

Sometimes referred to as data-driven journalism, data journalism seems to have taken up the mantle of CAR in contemporary professional journalism. Though it is less preferred by scholars, data journalism appears to be the term of choice in the news industry for journalism based on data analysis and the presentation of such analysis (though note the ambivalence toward the term found by Appelgren and Nygren 2014). Professional definitions have tended to be broad, characterizing data journalism as essentially any activity that deals with data in conjunction with journalistic reporting and editing or toward journalistic ends, as in Stray's (2011) definition of data journalism as "obtaining, reporting on, curating and publishing data in the public interest." Several others have defined data journalism in terms of its convergence between several disparate fields and practices, characterizing it as a hybrid form that encompasses statistical analysis, computer science, visualization and web design, and reporting (Bell 2012; Bradshaw 2010; Thibodeaux 2011). Data journalism has also been closely associated with the use and proliferation of open data and open-source tools to analyze and display that data (Gynnild 2014), though open data is not necessarily or exclusively a part of its domain of practice (Parasie and Dagiral 2013).

Data journalism has been ascendant since the late 2000s, before which time most data analysis within newsrooms had either been in the form of CAR or in news organizations that dealt largely in specialist financial information (Bell 2012). Though it is not a central element of professional journalistic work, it has made significant inroads into the news industry, with heavy demand throughout the profession despite a relatively small number of dedicated data journalists and relative rarity outside of the most resource-rich news organizations (Fink and Anderson 2014; Howard 2014). Young and Hermida (2014) argue that a new professional class of data journalists is beginning to form, though they have often appropriated computational methods to fit dominant professional practices. One particularly celebrated example of data journalism was *The Guardian*'s 2009–10 project reporting on the expense claims of Members of the United Kingdom's Parliament, in which the newspaper published 460,000 pages of expense reports online and asked their readers to sort through them and flag questionable claims. The project resulted in investigative reports and data visualizations led many Members of Parliament to re-examine and re-pay some of their claims. This project exemplifies the data journalism model in its focus on opening data to the public and its use of public input to drive data analysis, visualization, and reporting (Gray, Bounegru, and Chambers 2012).

While data journalism is often used within the context of investigative projects such as *The Guardian*'s, it is much more loosely coupled with investigative journalism than was CAR. Some scholars and professionals have emphasized the continuity between CAR and data journalism (e.g., Gordon 2013; Gray, Bounegru, and Chambers 2012), but data journalism's decoupling with investigative journalism and integration into broader journalistic practices marks a significant break between CAR and data journalism (Gray, Bounegru, and Chambers 2012; Marshall 2011; Minkoff 2010). Other distinctions include data journalism's emphasis on visualization as a core practice through a close connection between visualization design and journalistic values (Gordon 2013; Weber and Rall 2013), and an epistemological break in which data journalism views readers as co-constructors of truths and moral claims (Parasie and Dagiral 2013).

Computational Journalism

Computational journalism has at times been used by scholars to include CAR and data journalism, conflating the previous two forms; indeed, the most common definition of computational journalism seems to encompass both CAR and data journalism: "the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism" (Hamilton and Turner 2009, 2). But by defining them so broadly, this definition does not allow much room to draw significant distinctions between each of the three practices. Instead, following Diakopoulos (2011), I define computational journalism here as a strand of technologically oriented journalism centered on the application of computing and computational thinking to the practices of information gathering, sense-making, and information presentation, rather than the journalistic use of data or social science methods more generally.

Stavelin (2014) helpfully emphasizes the application of computational tools and methods in the service of journalistic aims in his definition of computational journalism, though he notes that it goes beyond a particular set of tools to a set of processes built on a particular mode of thought known as computational thinking. This form of thinking—developed as a concept by Wing (2006, 2008) but with roots in mid-twentieth-century computer science (National Research Council of the National Academies 2010)—is built around abstraction and automation. Abstraction, the ability to break down information or problems beyond their immediate material context, is the central element of computational thinking. It is a cognitive process, rather than a practice necessarily done by computer; computing, then, is simply the automation of abstracted information and processes (Wing 2008). These automation processes often take the form of algorithms, which are occasionally considered a third element of computational thinking (Flew et al. 2012). Algorithms are the abstraction of a step-by-step procedure taking an input and producing an output to accomplish a defined outcome (Diakopoulos 2014a; Wing 2008). Algorithms can prioritize, classify, and filter information, and can be involved in journalism at several stages, including distribution—as in search results and audience metrics—determining topics to cover, or even writing stories themselves (Anderson 2013a; Carlson 2014).

Even with this narrowed definition of computational journalism, there are a variety of types of projects that might fit under its umbrella. Diakopoulos (2014a) describes the use of algorithmic processes in reporting on other algorithms, such as ProPublica's recreation of the algorithm used to send personalized campaign emails in the 2012 US presidential campaign or *The Wall Street Journal*'s use of simulated user profiles to determine the algorithms governing price discrimination in online commerce. More directly, computational processes can produce the news content itself, as in Narrative Science's use of structured data to produce automated financial and sports articles (Bell 2012). A more widely applicable example of computational journalism may be DocumentCloud, founded by ProPublica and *New York Times* journalists in 2008, which hosts user-submitted and user-annotated documents and provides computational tools to process them, such as optical character recognition (Cohen et al. 2011). Though they involve different stages of the journalistic process and different levels of human involvement, each of these examples involve the core elements of computational journalism—practices or services built around computational tools in the service of journalistic ends.

Like data journalism, computational journalism has been characterized as a descendant of CAR. Gynnild (2014) identifies Philip Meyer as an ahead-of-his-time pioneer of computational thinking's application to journalism, and others have said that journalists have been practicing computational journalism (or computational thinking) for decades without labeling it as such (e.g., Linch 2010). Indeed, CAR is built in part around the use of simple computational processes—most commonly database analysis—to sort information. However, as Diakopoulos (2011) notes, computational journalism goes beyond CAR in its focus on the processing capabilities of computing, particularly aggregating, automating, and abstracting information. Likewise, there is also significant commonality between data journalism and computational journalism in the use of computational tools and collaborative processes to analyze and present data. But, as Stray (2011) points out, not all data journalism is computational; computational journalism works primarily through abstraction of information to produce computable models, while data journalism works primarily through analysis of data sets to produce data-oriented stories (Stavelin 2014). The three practices, then, are distinct quantitatively oriented journalistic forms: CAR is rooted in social science methods and the deliberate style and public-affairs orientation of investigative journalism, data journalism is characterized by its participatory openness and cross-field hybridity, and computational journalism is focused on the application of the processes of abstraction and automation to information.

Typology

Having outlined each of the concepts, we now turn toward the effort to classify and differentiate them. The typology (visualized in Figure 1) that follows examines four dimensions: two of them are professional—professional expertise versus networked information and transparency versus opacity. One, big data versus targeted sampling, is epistemological, and the final one has a professional/moral dimension—the vision of an active versus passive public. The dimensions of this typology are ideal types (Weber 1947), generalized forms not meant to capture the details of a particular case, but instead intended to serve as ideal forms against which individual cases and genres might be compared. As such, the classification of each form of journalism into this typology necessarily involves broad generalizations. Because of the overlap in practice among CAR, data journalism, and computational journalism, the typology is not meant to be a definitive placement of these genres regarding each type, but rather an initial guide used to evaluate any computational or data-oriented project, tool, or organization.

This typology was developed through a close reading of about 90 texts on CAR, data journalism, and computational journalism, within both academic and professional discourse. The professional discourse on the subject consisted largely of articles in journalism reviews such as the *Columbia Journalism Review* and *Nieman Journalism Lab*, textbooks, and other blog posts and articles by data and computational journalism professionals, gathered through several years of personal collection, augmented by archive searches of journalism publications and snowball-style data gathering through hyperlinks and citations. After analysis of a subset of these texts, an initial typology was developed; the analytical corpus was expanded and the typology revised and refined after academic feedback.

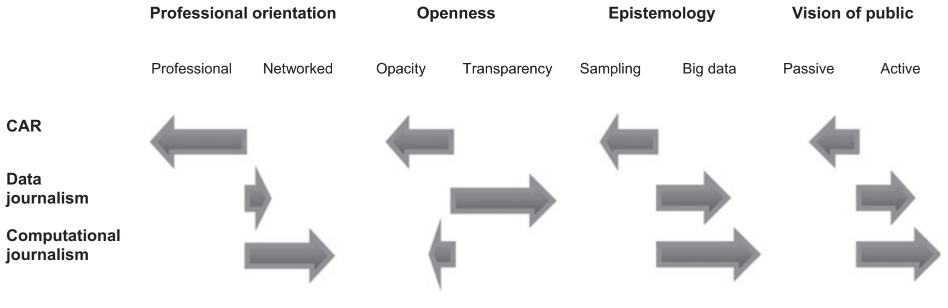

FIGURE 1
A visualized typology of data-driven journalism forms

Professional Expertise Versus Networked Information

The first dimension of the typology is an orientation toward openness and broad participation on one end and professional expertise and limited participation on the other. Expressed organizationally, it is the difference between a production process limited to professionals within institutional organizations and one open to a networked, loosely joined group consisting of both professionals and non-professionals. More acutely, it addresses the practices' relationship to the norms and practices of traditional professional journalism, particularly the degree to which they are subordinated to the specialized knowledge and institutionalized routines of traditional reporting. This tension between broad participation and professional control has been a defining one in twenty-first-century journalism (Lewis 2012), though it is magnified at the intersection between computing and journalism, as distributed participation is a fundamental element of open-source practice (Lewis and Usher 2013).

Throughout its history, CAR has been continually subordinated to professional norms and framed as a way to enhance professional expertise. Books and articles defending CAR and explaining its practices are filled with admonitions that CAR does not replace or threaten traditional reporting, depicting it as simply a new tool in the service of existing practices, rather than a new way of seeing news or information. "Nothing can replace good, old-fashioned reporting, but CAR is an additional tool," said one investigative reporter in a typical statement (Garrison 1996, 116). Data is similarly seen within CAR as entirely secondary to human-oriented aspects of a story—that is, the ones that must be gathered through traditional, "shoe-leather" practices of interviewing and direct observation (e.g., Houston 1996; Jaspin 1993; Miller 1998). This subjugation to the methods of professional methods is also evident in CAR's close ties to investigative journalism, which sits at the core of journalists' professional identity (Ettema and Glasser 1998). In CAR, claims from data are precipitated by leads based on reporting and are subjected to journalistic practices such as cross-checking and interviewing that are drawn from and subordinated to investigative journalism (Parasie 2014). The effect is that, as Philip Meyer put it, CAR is "the same old journalism but with better tools" (Miller 1988, 36).

This placement of CAR strictly in the service of professional norms and practices puts it squarely within the professionalist, "high modern" paradigm prevalent in journalism during the era in which it developed (Flew et al. 2012). In this model, data held no value of their own except to produce stories, and "the computer-assisted reporter was still primarily a journalist rather than a technologist; the underlying goal was to produce a better story" (Lewis and Usher 2013, 605; Parasie and Dagiral 2013). This foregrounding of story has continually pulled CAR back into the realm of the investigative reporting-oriented professional practices, like interviewing and examination of documents, around which journalists are most keen to build their professional expertise and identity (Coddington 2014). As Meyer (2002) argues, CAR has also existed in tension with the professional journalistic norms into which it is embedded. In particular, CAR's emphasis on the analysis of data collected according to social scientific principles is a real challenge to traditional journalism, which tends to defer to the expertise of official sources and the authority of anecdotal example and personal experience. Still, by and large, CAR is a form of data processing that is subordinated almost completely under the principles of professional journalism.

Data journalism retains CAR's emphasis on subordinating data to the professional journalistic value of narrative and the "story." Just as in CAR, data journalism discourse foregrounds telling the story over using data, though it is looser in its connection to traditional journalistic practices in producing those narratives (Fink and Anderson 2014; Stavelin 2014). As Howard (2014, 5) asserts in his definition of the practice: "data journalism is telling stories with numbers, or finding stories in them." In data journalism, however, the expertise needed to determine the story has spread beyond the strictly professional realm of CAR. This storytelling work no longer requires interviews and other professional journalistic practices, but only examination of the data. This opens up the expertise of using data to tell stories to anyone capable of accurately drawing meaning from that data, professional journalist or not. Data journalist and researcher Liliana Bounegru explains this shift aptly:

> By enabling anyone to drill down into data sources and find information that is relevant to them, as well as to verify assertions and challenge commonly received assumptions, data journalism effectively represents the mass democratisation of resources, tools, techniques and methodologies that were previously used by specialists—whether investigative reporters, social scientists, statisticians, analysts or other experts. (Gray, Bounegru, and Chambers 2012, "Data Journalism in Perspective")

In practical terms, this openness toward non-professional involvement leads many data journalism projects to involve opening data sets to the audience and developing tools for them to explore or personalize them (Parasie and Dagiral 2013), as well as crowdsourcing the data and analysis stemming from it—inverting the normal computational mode of using software to compute human data by instead providing data to humans to process (Appelgren and Nygren 2014; Stavelin 2014, 44). Data journalism retains an emphasis on editorial selection and professional news judgment in analyzing and presenting data (Stray 2010), but it does so while also building around a recognition that expertise in analyzing and drawing meaning from that data often exists outside of the profession, among the audience. Though data journalists see their work as fundamentally sense-making as other professional journalists do, they have opened up that sense-making process to be a collective one, bringing the citizen alongside the professional (Parasie and Dagiral 2013).

A distributed information production process is even more central to computational journalism than to data journalism. Much like open-source journalism, computational journalism and computational thinking are at their core collaborative processes. Computational thinking is fundamentally a group phenomenon rather than an individual one (National Research Council of the National Academies 2010), and computational journalism is oriented around the belief that human expertise is located in crowds, rather than small, closely guarded enclaves. Computational journalism is an effort to harness that expertise, taking advantage of emerging sets of tools that allow for broad, many-to-many collaboration (Cohen et al. 2011; Flew et al. 2012). Computational journalism can shed the emphasis on narrative and storytelling that tends to draw CAR and data journalism back toward professional journalistic practices and news judgment, as it tends to be more focused on producing a tangible product or platform than a narrative (Diakopoulos 2013; Stavelin 2014), though Diakopoulos (2011, 2014b) also details forms of computational journalism that are oriented toward finding and telling stories.

There are limits to the distributed structure and practices of computational journalism. It is much more reliant on technical expertise—most notably, advanced programming skills—that while not limited to a particular profession (and certainly not journalism) can nonetheless be quite difficult to acquire. Additionally, Karlsen and Stavelin (2014) found that even those highly specialized technical skills can still be subordinated to journalistic ones when computational journalism is practiced within a traditional newsroom. Still, of the three forms examined here, computational journalism is least wedded to professional journalistic norms and practices and most essentially distributed and networked in its practice.

Transparency Versus Opacity

Transparency has been an ascendant journalistic value over the past decade, one characterized as a crucial element to establishing credibility with an increasingly mistrustful public (Karlsson 2010; Plaisance 2007). Though professional journalists have long advocated for open information for themselves, they have been much less willing to open up the process by which they produce news to the public. Lewis and Usher (2013) describe transparency as a key element of the open-source movement, though journalists have been slow to pick the value up because of their concerns about its threat to their professional autonomy. Karlsson (2010) classifies two distinct strains of journalistic transparency: disclosure transparency, or openness about how news is produced, and participatory transparency, or the ability of those outside the profession to participate in the journalistic process. I will focus here on disclosure transparency, which Karlsson notes was technically achievable in the pre-digital media system, but was largely barred by a closed professional culture.

CAR is grounded in the same modernist professional journalistic culture that has typically resisted efforts to make its professional practices transparent to the public. Though it has exhibited a stronger inclination toward disclosure transparency than that culture through the transparency of social science methods (Meyer 2002), some traces of that opacity are evident. In CAR, as Taylor (2009) notes, the data are meant to be invisible within a story—something to be included, but downplayed so as not to detract from the story's core human elements. The advice given in Miller's (1998) CAR textbook comports with this description: choose carefully what numbers to include, and only lead with those numbers when they are particularly compelling. Otherwise, the data should recede into the background. As for the process by which those numbers are gathered, Miller advocates transparency, but equally emphasizes how it should be limited: "Explain, when necessary and relevant, how you gathered your information. But don't go overboard" (Miller 1998, 225). In the CAR paradigm, neither methods nor the data itself are the story, so both should be set in the background to the extent that they infringe on journalists' professional abilities to filter data and find meaning in it for the audience.

By contrast, transparency of both process and product are a core element of data journalism. Some of that transparency has come to data journalism by way of the open-source philosophy, as Gynnild (2014) characterized the use of open-source tools and open data as a defining element of data journalism. Unlike in CAR, publishing the data alongside articles based on it is so fundamental to the practice of data journalism that it is described as something that "goes without saying" (Gray, Bounegru, and Chambers 2012, "Engaging People Around Your Data"). *The Guardian*'s Simon Rogers (Stray 2010) described this data publication as the primary difference between how his paper approaches data now and how it did so a decade prior, during the heyday of CAR. In Rogers' account, the shift has been a response to demand driven by the access to unfiltered information elsewhere on the internet. Online audiences, Rogers said, "want the interpretation and the analysis from people, but they also want the veracity of seeing the real thing, without having it aggregated or put together. They just want to see the raw data" (Stray 2010). In data journalism, displaying this kind of transparency does not undermine the story the journalist is trying to convey; it simply adds to it.

The role of transparency is much less settled within computational journalism, thanks to particular obstacles endemic to computational work. Algorithmic transparency, Diakopoulos (2014b) argues, is much more difficult than data transparency, as it involves additional labor costs for both creating and making sense of an algorithm for public consumption. Likewise, Stavelin (2014) contends that software is opaque by nature, and thus any transparency in computational journalism is chiefly borrowed from professional journalistic values, rather than coming from within its own native framework. Computational journalism does, however, have its own normative well from which to draw an orientation toward transparency—namely the influence of the open-source software movement (Lewis and Usher 2013; Parasie and Dagiral 2013), whose commitment to transparency far outstrips that of professional journalism. To the degree, then, that computational journalists adhere to the ideals of that movement, they may be able to overcome the barriers to disclosure that exist within the work they do. Diakopoulos (2014b) offers a promising model to incorporate transparency into the journalistic use of algorithms, though he acknowledges the tensions inherent in such an adaptation of journalistic norms.

Targeted Sampling Versus Big Data

The third dimension of the typology is epistemological, having to do with how data is gathered and analyzed in order to generate conclusions and knowledge. On one pole is data gathered through targeted means such as sampling, with conclusions reached through inference and causality placed at a premium. This is generally the epistemological approach of classic social science. On the other pole is a focus on large data sets or collections of information that are obtained through attempts at capturing the totality of a phenomenon, with an emphasis on exploratory analysis and simple correlation rather than causation. This roughly corresponds to the epistemology of the "big data" movement, which Mayer-Schönberger and Cukier (2013) have helpfully set in stark contrast to that of traditional social science. From a big-data perspective, simple correlation and exploratory rather than hypothesis-driven analysis are often sufficient because the size of the database overcomes any analytical shortcomings with it (Bollier 2010; Mayer-Schönberger and Cukier 2013).

CAR is located toward the targeted sampling pole of this dimension. Meyer's (2002) precision journalism philosophy out of which CAR grew is not just deeply rooted in the practice of social science; it *is* social science, simply translated for journalists. CAR remained rooted in that social science mindset with an emphasis on hypothesis testing and survey research, especially early on. As CAR grew, however, more of its projects demonstrated an openness to more complete data and less statistically rigorous analysis. As we will see, data journalism and computational journalism are in part responses to a dramatic rise in information scale, and CAR was also a response to certain forms of information abundance (Parasie and Dagiral 2013). But it typically dealt with a somewhat smaller scale of data, and it often—but not always—used sampling and statistical analysis as a method to produce intelligibility for large data sets.

Data journalists often emphasize the exponential increase in the amount of data being collected and the size of individual data sets as a key element of what is new about their practice (Gray, Bounegru, and Chambers 2012; Howard 2014; Rogers 2011). When the primary task shifts from finding and collecting data to processing it, the analysis of that data accordingly shifts from being driven by hypotheses that spurred the gathering of that data to a more inductive and exploratory approach. Tellingly, while Appelgren and Nygren (2014) described the data journalists they studied as being tied to Meyer's methods, none of those journalists mentioned social scientific methods, instead emphasizing the size of the data sets they dealt with. Rogers (2011) also ties this change to the increasing speed of data journalism, noting that the old form of the practice often involved weeks of in-depth data analysis, while the new form prioritizes producing analyses as quickly as possible. Both the scale of the data and the pace of the work, then, push data journalism toward a more exploratory, big-data form of analysis.

Computational journalism is similarly oriented to a big-data epistemology, largely because it is responding, just as data journalism is, to a shift toward increasing information abundance (Flew et al. 2012). The speed issue faced by data journalism is much less present here, but the foregrounding of computational methods encourages a particular inclination toward use of unaltered large data sets. Such computational methods allow extremely large data sets to be handled in full, thus eliminating the need for sampling. Parasie and Dagiral (2013) explain that the programmer-journalists in their study "do not consider statistics as a major tool because, in their opinion, data do not hide anything if they are granular and complete" (863). They eschew sampling because they do not believe such procedures can produce new knowledge from data. Instead, that intelligibility comes from the ability to access complete data through skilled use of computational tools.

Vision of the Public: Active Versus Passive

The conception of the public has been a central element of modern professional journalism; its invocation has been a foundation for journalistic claims of authority. Journalism has historically seen the public as a unitary, rational, and fixed body, but the online environment has deeply complicated this vision, at once revealing the public as fragmented and creating the potential for a more interactive and participatory public (Anderson 2013b). This vision of a fragmented and participatory public creates tension with journalists' professional norms of autonomy and authority, leading journalists to continue to resist seeing the public as a productive and interactive part of the journalistic process (Lewis 2012). Each of these three forms acts on a vision of a more active public than does traditional professional journalism, though the degree of that public's activity and generative value varies widely.

The public is crucial to the work of CAR, but in a much less active way than in data journalism. As a form closely tied to investigative journalism, CAR relies on the public to supply the moral outrage that it works to produce. As Ettema and Glasser (1998) argue, the normative aim of investigative journalism is to highlight violations of the moral order, as determined by the public for which those journalists write, and their response to the story. In this way, "*every* investigation must be understood as a call to the conscience of a community" (Ettema and Glasser 1998, 187)—a test of the public's consensus on community values. The public's role is to respond to the perceived moral outrage in a way that upholds their community values and condemns violators. This role for the public typically does not involve analyzing or contributing to the data themselves; CAR aims to use the truths in public data to set the public agenda, rather than giving the public an active role in determining its own meaning from data (Parasie and Dagiral 2013).

Like CAR, data journalism is also built around informing the public about critical issues, but the public is involved to a greater degree and to different ends. The goal of data journalism is to allow the public to analyze and draw understanding from data themselves, with the data journalist's role being to access and present the data on the public's behalf. This has a substantial influence on the process of data journalism itself, which is oriented around creating utility for the user. In developing data journalism products—often data visualization or Web applications—their usefulness to the audience is a prime consideration (Gray, Bounegru, and Chambers 2012; Stray 2010). Consider the contrast with CAR, whose primary measure of "impact" is in influence not on the public itself, but on institutions or officials through public outrage. In data journalism, the public plays a much more direct role, as the goal is more simply to provide a useful way for the public to enhance its own understanding of, and draw its own meaning from, public issues. On the other hand, as Fink and Anderson (2014) note, while many data journalists profess a devotion to serving an active public, their conception of that public is still primarily rationalized and anonymized through online metrics, rather than as a personal or reciprocal participant in the journalistic process.

Computational journalism also views its public as a collection of rational, participatory users who are capable of producing understanding from data themselves. The members of the public, in this view, expect to interact with the information they encounter, and the goal of computational journalism is to provide them with the tools they need to perform their own filtering and abstraction with it (Flew et al. 2012; Hamilton and Turner 2009). As Gynnild (2014) notes, the computational view of the audience as autonomous and creative enough to perform their own searches of data—allowed by computational tools—is part of what enables the publication of data in itself to be considered journalistic. This overall view of an interactive, autonomous public that expects to be engaged with data is very similar to that of data journalism; if anything, computational journalism's envisioned public is even more empowered in its ability to do its own computational thinking on the data it can access.

Conclusion

This typology is only an initial attempt to classify more systematically these data-driven journalistic practices. These dimensions are hardly the only ones differentiating them, and this area of journalism remains unsettled, so new dimensions and forms of practice may emerge over the next several years. Still, this typology indicates a significant gap between the professional and epistemological orientations of CAR, on the one hand, and both data journalism and computational journalism, on the other (see Figure 1). This divide has its origins in the cultural background from which each has approached journalism: CAR arose out of an effort to marry social science with modern professional journalism, and especially investigative journalism. Data journalism and computational journalism, on the other hand, have arisen from the intersection of professional journalism with open-source culture. Each represents a different amalgam between those two social realms, but the fact that those combinations are being made from very similar raw cultural materials gives them much more in common with each other than with CAR.

As it stands, data journalism is the closest we have to the melding of professional journalism and both open-source and computational principles, as advocated by Lewis and Usher (2013) and others. Data journalists' statements that narrative, storytelling, and traditional reporting are still important parts of good data journalism (e.g., Gray, Bounegru, and Chambers 2012) are attempts to closely link themselves to the dominant professional view of journalism. By reiterating the importance of traditional journalistic work, they help to ensure that their own work is taken seriously by professional journalism—that they are seen as continuing its practices, rather than harming them (Powers 2011).

Like data journalism, computational journalism is a blend between professional journalism and open-source culture, though through its tighter connection to programming it moves closer to the influence of open-source culture than does data journalism. Computational journalism thus inherits a strong emphasis on open and networked workflows but also remains more materially and technically oriented than data journalism. The bridge to professional journalism and to CAR is a bit further here than with data journalism: the concept of computational thinking, of abstracting data when approaching complex tasks or objects of news down to granular, discrete elements, does not appear to have a precedent or analog in pre-computer-age journalism. Gynnild (2014) does, however, identify Meyer as an ahead-of-his-time pioneer of computational thinking's application to journalism. Though computational journalism differs significantly from CAR in many of its emphases and animating principles, its emphasis on abstracting journalistic inquiries to large-scale and quantifiable forms, and using computational methods to filter and analyze large bodies of information, can be traced to CAR's influence.

Despite its generalized nature and the fluidity of the practices it covers, this typology offers a useful orienting framework for future research into these emergent forms. It highlights several under-researched dimensions that may be especially fruitful for gaining a fuller understanding of data-driven journalistic practices and their relationship to both professional journalism and the public. First, scholars would do well to focus more closely on the epistemological elements of each of the forms—the ways in which their constructions of facts and knowledge compare and contrast to each other and to other professional journalistic practices. This is one of data-driven journalism's starkest points of divergence from the modern professional journalistic mindset, and further work that fleshes out the epistemological roots of these practices, such as Parasie (2014) in this issue, would be most helpful in outlining its contours.

Second, research should delve deeper into the shifting position of data-driven journalism in relationship to the larger field of professional journalism. On this point, I echo Anderson's (2013a) call to approach these forms of journalism from an institutional or field perspective, examining the social and cultural power struggles within this emerging field and in relation to adjacent fields such as traditional journalism or computer science. As this field grows and coheres, its autonomy from and flows of influence and capital between adjacent fields may be crucial in shaping broader journalistic practice.

Finally, beyond general statements about commitment to openness and participation, the relationship between these journalistic forms and the public has received little scrutiny. Research should more fully examine these journalists' vision of the public and their relationship to it, including their audiences' reception of their work. We have little knowledge of whether data journalists' openness to the public is being substantively reciprocated, or the epistemological and attitudinal frameworks in which audiences are consuming and evaluating the journalism they produce. Research from such an audience-centric perspective could extend our currently one-dimensional understanding of data-driven journalism and the public. To the extent that a quantitative turn is indeed occurring within journalism, it becomes particularly important to examine the ways such a turn changes its alignment with both the profession's traditional values and practices as well as the public.

ACKNOWLEDGEMENTS

The author thanks Steve Reese and Seth Lewis, as well as the anonymous reviewers for *Digital Journalism*, for their feedback on earlier drafts of this paper.

REFERENCES

Anderson, C. W. 2013a. "Towards a Sociology of Computational and Algorithmic Journalism." *New Media and Society* 15: 1005–1021.

Anderson, C. W. 2013b. *Rebuilding the News: Metropolitan Journalism in the Digital Age*. Philadelphia, PA: Temple University Press.

Appelgren, Ester, and Gunnar Nygren. 2014. "Data Journalism in Sweden: Introducing New Methods and Genres of Journalism into 'Old' Organizations." *Digital Journalism* 2: 394–405. doi:10.1080/21670811.2014.884344.

Bell, Emily. 2012. "Journalism by Numbers." *Columbia Journalism Review*. September 5. http://www.cjr.org/cover_story/journalism_by_numbers.php?page=all.

Bollier, David. 2010. *The Promise and Peril of Big Data*. Washington, DC: The Aspen Institute.

Bradshaw, Paul. 2010. "How to Be a Data Journalist." *The Guardian*, October 1. http://www.theguardian.com/news/datablog/2010/oct/01/data-journalism-how-to-guide.

Carlson, Matt. 2014. "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority." *Digital Journalism*. doi:10.1080/21670811.2014.976412.

Coddington, Mark. 2014. "Defending Judgment and Context in 'Original Reporting': Journalists' Construction of Newswork in a Networked Age." *Journalism* 15: 678–695.

Cohen, Sarah, Chengkai Li, Jun Yang, and Cong Yu. 2011. "Computational Journalism: A Call to Arms to Database Researchers." Paper presented at the 5th Biennial Conference on Innovative Data Systems Research (CIDR '11), Asilomar, CA, January 9–12. http://ranger.uta.edu/~cli/pubs/2011/cjdb-cidr11-clyy-nov10.pdf.

Coleman, E. Gabriella. 2013. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton, NJ: Princeton University Press.

Cox, Melisma. 2000. "The Development of Computer-assisted Reporting." Paper presented at AEJMC 2000, Phoenix, AZ, August 9–12. http://com.miami.edu/car/cox00.pdf.

Diakopoulos, Nicholas. 2011. "A Functional Roadmap for Innovation in Computational Journalism." April 22. http://www.nickdiakopoulos.com/2011/04/22/a-functional-roadmap-for-innovation-in-computational-journalism/.

Diakopoulos, Nicholas. 2013. "Finding Tools Vs. Making Tools: Discovering Common Ground between Computer Science and Journalism." *Nieman Journalism Lab*, February 14. http://www.niemanlab.org/2013/02/finding-tools-vs-making-tools-discovering-common-ground-between-computer-science-and-journalism/.

Diakopoulos, Nicholas. 2014a. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures." *Digital Journalism*. doi:10.1080/21670811.2014.976411.

Diakopoulos, Nicholas. 2014b. *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. New York: Tow Center for Digital Journalism.

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

Fink, Katherine, and C. W. Anderson. 2014. "Data Journalism in the United States: Beyond the 'Usual Suspects'." *Journalism Studies*. doi:10.1080/1461670X.2014.939852.

Fink, Katherine, and Michael Schudson. 2014. "The Rise of Contextual Journalism, 1950s–2000s." *Journalism* 15: 3–20.

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2012. "The Promise of Computational Journalism." *Journalism Practice* 6: 157–171.

Garrison, Bruce. 1996. "Tools Daily Newspapers Use in Computer-Assisted Reporting." *Newspaper Research Journal* 17 (1): 113–126.

Gordon, Rich. 2013. "Want to Build a Data Journalism Team? You'll Need These Three People." *Northwestern University Knight Lab*, June 28. http://knightlab.northwestern.edu/2013/06/28/want-to-build-a-data-journalism-team-youll-need-these-three-people/.

Gray, Jonathan, Liliana Bounegru, and Lucy Chambers, eds. 2012. *The Data Journalism Handbook*. http://datajournalismhandbook.org/1.0/en/index.html.

Gynnild, Astrid. 2014. "Journalism Innovation Leads to Innovation Journalism: The Impact of Computational Exploration on Changing Mindsets." *Journalism* 15: 713–730. doi:10.1177/1464884913486393.

Hamilton, James T., and Fred Turner. 2009. "Accountability through Algorithm: Developing the Field of Computational Journalism." Report given at the Behavioral Sciences Summer Workshop, Stanford, CA, July 27–31. http://www.stanford.edu/~fturner/Hamilton%20Turner%20Acc%20by%20Alg%20Final.pdf.

Hermida, Alfred, and Mary-Lynn Young. 2014. "From Mr. and Mrs. Outlier to Central Tendencies: Computational Journalism and Crime Reporting at the *Los Angeles Times*." *Digital Journalism*. doi:10.1080/21670811.2014.976409.

Holovaty, Adrian. 2009. "The Definitive, Two-Part Answer to 'is Data Journalism?'" May 21. http://www.holovaty.com/writing/data-is-journalism/.

Houston, Brant. 1996. *Computer-assisted Reporting: A Practical Guide*. New York: St. Martin's.

Howard, Alexander Benjamin. 2014. *The Art and Science of Data-driven Journalism*. New York: Tow Center for Digital Journalism. http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf.

Jaspin, Elliot. 1993. "The New Investigative Journalism: Exploring Public Records by Computer." In *Demystifying Media Technology*, edited by John V. Pavlik and Everette E. Dennis, 142–149. Mountain View, CA: Mayfield.

Karlsen, Joakim, and Eirik Stavelin. 2014. "Computational Journalism in Norwegian Newsrooms." *Journalism Practice* 8: 34–48. doi:10.1080/17512786.2013.813190.

Karlsson, Michael. 2010. "Rituals of Transparency: Evaluating Online News Outlets' Uses of Transparency Rituals in the United States, United Kingdom and Sweden." *Journalism Studies* 11: 535–545.

Kelty, Christopher M. 2008. *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke University Press.

Lewis, Seth C. 2012. "The Tension between Professional Control and Open Participation: Journalism and Its Boundaries." *Information, Communication and Society* 15: 836–866. doi:10.1080/1369118X.2012.674150.

Lewis, Seth C., and Nikki Usher. 2013. "Open Source and Journalism: Toward New Frameworks for Imagining News Innovation." *Media, Culture and Society* 35: 602–619. doi:10.1177/0163443713485494.

Linch, Greg. 2010. "Why Computational Thinking Should be the Core of the New Journalism Mindset." *Publish2*, April 30. http://blog.publish2.com/2010/04/30/computational-thinking-new-journalism-mindset/.

Maier, Scott R. 2002. "Numbers in the News: A Mathematics Audit of a Daily Newspaper." *Journalism Studies* 3: 507–519.

Marshall, Sarah. 2011. "10 Things Every Journalist Should Know about Data." *News: Rewired*, April 26. http://www.newsrewired.com/2011/04/26/10-things-every-journalist-should-know-about-data/.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Eamon Dolan/Houghton Mifflin Harcourt.

Meyer, Philip. 1973. *Precision Journalism*. Bloomington, IL: Indiana University Press.

Meyer, Philip. 1999. "The Future of CAR: Declare Victory and Get out!." In *When Nerds and Words Collide: Reflections on the Development of Computer Assisted Reporting*, edited by Nora Paul, 4–5. St. Petersburg, FL: Poynter Institute.

Meyer, Philip. 2002. *Precision Journalism*. 4th ed. Lanham, MD: Rowman and Littlefield.

Miller, Tim. 1988. "The Data-base Revolution." *Columbia Journalism Review* 27 (3): 35–38.

Miller, Lisa C. 1998. *Power Journalism: Computer-assisted Reporting*. Fort Worth, TX: Harcourt Brace and Co.

Minkoff, Michelle. 2010. "Bringing Data Journalism into Curricula." March 24. http://michelleminkoff.com/2010/03/24/bringing-data-journalism-into-curricula/.

National Research Council of the National Academies. 2010. *Report of a Workshop on the Scope and Nature of Computational Thinking*. Washington, DC: National Academies Press.

Parasie, Sylvain. 2014. "Data-driven Revelation? Epistemological Tensions in Investigative Journalism in the Age of 'Big Data'." *Digital Journalism*. doi:10.1080/21670811.2014.976408.

Parasie, Sylvain, and Eric Dagiral. 2013. "Data-driven Journalism and the Public Good: 'Computer-Assisted Reporters' and 'Programmer-Journalists' in Chicago." *New Media and Society* 15: 853–871.

Petre, Caitlin. 2013. "A Quantitative Turn in Journalism?" *Tow Center for Digital Journalism*, October 30. http://towcenter.org/blog/a-quantitative-turn-in-journalism/.

Plaisance, Patrick L. 2007. "Transparency: An Assessment of the Kantian Roots of a Key Element in Media Ethics Practice." *Journal of Mass Media Ethics* 22: 187–207.

Powers, Matthew. 2011. "'In Forms that are Familiar and Yet-to-Be Invented': American Journalism and the Discourse of Technologically Specific Work." *Journal of Communication Inquiry* 36: 24–43.

Rogers, Simon. 2011. "Data Journalism at the Guardian: What is It and How Do We Do It?" *The Guardian*, July 28. http://www.theguardian.com/news/datablog/2011/jul/28/data-journalism.

Stavelin, Eirik. 2014. *Computational Journalism: When Journalism Meets Programming*. PhD diss., Norway: University of Bergen.

Stray, Jonathan. 2010. "How the Guardian is Pioneering Data Journalism with Free Tools." *Nieman Journalism Lab*, August 5. http://www.niemanlab.org/2010/08/how-the-guardian-is-pioneering-data-journalism-with-free-tools/.

Stray, Jonathan. 2011. "A Computational Journalism Reading List." January 31. http://jonathanstray.com/a-computational-journalism-reading-list.

Taylor, Megan. 2009. "How Computer-Assisted Reporters Evolved into Programmer/Journalists." *PBS MediaShift*, August 7. http://www.pbs.org/mediashift/2009/08/how-computer-assisted-reporters-evolved-into-programmerjournalists219.

Thibodeaux, Troy. 2011. "5 Tips for Getting Started in Data Journalism." *Poynter*, October 6. http://www.poynter.org/how-tos/digital-strategies/147734/5-tips-for-getting-started-in-data-journalism/.

Weber, Max. 1947. *The Theory of Social and Economic Organization*. Translated and edited by Talcott Parsons. New York: Free Press.

Weber, Wibke, and Hannes Rall. 2013. "'We Are Journalists': Production Practices, Attitudes and a Case Study of the New York times Newsroom." In *Interaktive Infografiken*, edited by Wibke Weber, Miguel Burmester, and Ralph Tille, 161–172. Wiesbaden, Germany: Springer Vieweg.

Wing, Jeannette M. 2006. "Computational Thinking." *Communications of the ACM* 49 (3): 33–35.

Wing, Jeannette M. 2008. "Computational Thinking and Thinking about Computing." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 366: 3717–3725.

Yarnall, Louise, J. T. Johnson, Luke Rinne, and Michael Andrew Ranney. 2008. "How Post-Secondary Journalism Educators Teach Advanced CAR Data Analysis Skills in the Digital Age." *Journalism and Mass Communication Educator* 63: 146–164.

Mark Coddington, School of Journalism, University of Texas at Austin, USA. E-mail: markcoddington@gmail.com. Web: http://markcoddington.com