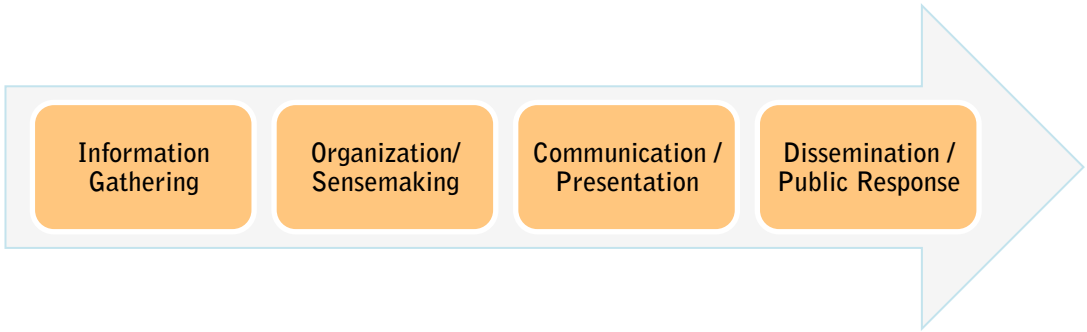
Nicholas Diakopoulos, Ph.D.
School of Communication and Information, Rutgers University
Original Version January, 2010; Updated April 2011.

A Functional Roadmap for Innovation in Computational Journalism

Overview

Journalism in all of its senses spans a spectrum of meaning ranging from social *purpose* (e.g. watchdogging), to professionalized *practice* (e.g. ethics and professional standards), to the functional *processes* that journalists employ. Innovation in journalism can happen within or across this hierarchy of meanings, but in this paper, in particular, I will explore the role that computing can play in the *process aspects* of journalism. My intent is to lay a foundation of computational thinking for journalistic processes upon which updated journalistic practices and reinvigorated journalistic purposes can be built.

From a process perspective, *Computational Journalism* is the application of computing and computational thinking to the activities of journalism including *information gathering*, *organization and sensemaking*, *communication and presentation*, and *dissemination and public response* to news information, all while upholding core values of journalism such as accuracy and verifiability. It is inclusive of CAR (Computer-Assisted Reporting) but distinctive in its focus on the *processing* capabilities (e.g. aggregating, relating, correlating, abstracting) of the computer in comparison to mundane aspects of storage or access. The field draws on technical sub-fields of computer science including information retrieval, artificial intelligence, content analysis, visualization, personalization, and recommender systems as well as aspects of social computing and information science.



While Computational Journalism is unlikely to ever replace journalists with computers it does promise a future where the goals of human journalists are greatly enabled and augmented through computing. Moreover, its pursuit may also inform developments in Computer Science, by, for example, driving research in visual analytics and visualization, time-critical information processing, trustworthy computing, and user interfaces.

In the remainder of this paper I will discuss opportunities for innovation along the lines of the process aspects of journalism identified above. My goal is to stimulate new research

and applications on these processes in the context of journalism and explore the challenges and opportunities in this space.

Information Gathering

The adoption of cheap and ubiquitous devices with photo and video capability has already had a substantial impact on how stories are reported, both in the mainstream media and through citizen journalism. While sensing hardware has gotten cheaper and more pervasive, social networking systems (e.g. Facebook) and social awareness streams (e.g. Twitter) have explicitly connected the *what* of sensing with the *who* is sensing or reporting.

The process of information gathering and reporting largely hinges on finding and verifying sources of information. Some of the best (and most difficult) journalism hinges on cultivating relationships over time with a personal network of sources. What's different about the sources that are available from social networks is that, although they are by and large public, they may not be *familiar* to the journalist. Finding the desired sources while characterizing the expertise and veracity of those sources represents a barrier to fully realizing the journalistic value from these networks.

There are at least four aspects of information gathering from social networks and awareness streams that can be enhanced computationally: (1) source expertise finding, (2) source characterization (e.g. historical biases), (3) cross-referencing and independence of breaking eye-witness reports, and (4) originating source of information determination. For instance, a computational process could automatically compute the sentiment (i.e. pro / con) of a source with respect to a range of topics or issues based on their history of Twitter messages. Such rankings could then be used to inform journalists about the background of a potential interviewee. Or, consider a breaking news scenario where a journalist is attempting to cross-reference messages for validity. Algorithms can be developed to estimate the independence of those sources or to trace information back to a likely originating source. These are just a few examples of the potential areas for technical innovation in the area of information gathering.

Organization and Sensemaking

With a growth in information gathering capabilities comes the difficulty of organizing and making sense of all of that information by journalists. This is a process where computers have already had a significant impact, namely though Computer Assisted Reporting (CAR). CAR tools are usually generic in the sense that they are widely applicable to different stories, though many tools are designed for specific data types such as geographic, temporal, or network.

While many CAR tools succeed in enabling journalists to organize their information there is still considerable room for improvement in the area of sensemaking. In particular, computational perception and content analysis enable computers to convert signals about the world (including everything from sensor values to Twitter messages) into semantically and contextually laden symbols (e.g. names of people or places) or

aggregate and derivative values (e.g. the sentiment or emotion of a message, the novelty or unusualness of a message with respect to an event).

Together with interactive and visual ways of presenting these computed, "semantic" facets of information there is a huge potential space for innovation in journalism tools. Some of this innovation is happening in other domains that draw on a similar process of sensemaking, such as intelligence analysis. These tools can be evaluated to better understand how they do or do not work in the context of journalism, and, in general, computational tools developed to enable sensemaking will need rigorous attention to the evaluation of their utility in real situations. Finally, sensemaking tools not only have potential for helping journalists but also for helping "readers" make sense of growing online repositories of newsworthy content and data.

Communication and Presentation

Once a story is organized and been made sense of the next process entails communicating and presenting it in a relevant and interesting way. And while I won't argue that every story demands it, there will be some stories that benefit from computationally infused *presentations* of content. A journalist might use computation in such a story by making models or data interactive in a way that informs the user moreso than reading a static story.

User interfaces need to innovate more generic paradigms to compellingly communicate complex stories via models, data, simulation, and games. For instance, recent research info playable data graphics has looked into how to add game elements such as goals, scores, and advancement into how users interact with online visualized data. Other types of newsgames explore editorial simulation or decision making processes. One thing to consider as we invent these new experiences is how journalistic norms and values play out in interactive media. There are certain notions of interactive rhetoric and literacy that need to be taken into account when training computational journalists.

As governmental data becomes emancipated from closed databases (as is the current executive order in the U.S.) the opportunities for telling stories through models, data, simulation, and games will only grow. There is a range of potential new (and not yet invented) storytelling forms that combines both elements of interactivity and computing with games, data, and news content. This will be an area ripe for alternative methods of communicating complex information in engaging and interactive formats.

Dissemination and Public Response

From a business perspective, one of the most disruptive shifts in journalism has been the process of digitization and dissemination of content online. This transition took content that was once constrained by a fixed medium and brought the variable costs of publishing space close to zero. The implication of this shift is that there is much more content out there and, practically speaking, many more ways to compete for attention for content. With unlimited space come the issues of *information overload* and *scale.*

Computation can improve the process of dissemination by addressing information overload and scale issues through, for instance, personalization and content adaptation

systems as well as recommender systems. Many of the methods developed will also be applicable to monetization strategies since the fundamental scale issue revolves around matching a paucity of attention with the right content in order to drive higher advertising revenue.

Another implication of unlimited publishing space is that instead of being constrained to a narrow “letters to the editor” page, public response can instead expand to whatever the community needs dictate. In managing the process of interaction with the public response, journalists are encountering this scale problem in terms of interacting with and moderating users’ content in online commenting systems.

In particular there is a lot that computation can offer to improve online commenting systems, both from the perspective of a journalist dealing with moderation as well as for users of the commenting system. Content analysis, such as natural language processing, computational linguistics, and standard information retrieval techniques can help with both the scale as well as the *quality* of the discourse by introducing new ways for filtering and organizing comments. For instance, content analysis could be used to rank comments by (1) relevance to the story, (2) subjectivity or objectivity, or (3) degree of politeness. This could aid the process of journalists interacting with readers as well as readers interacting with readers by making it easier to find high quality contributions.

Looking Ahead

Technology is rapidly changing the landscape of how news information is gathered, made sense of, communicated, and disseminated. To pave the way to the future, journalism schools need to train more computationally literate journalists who develop a deep understanding of notions of abstraction, modeling, parameterization, aggregation, scaleability, and programming. And while industry grapples with the culture clash between engineers and journalists as well as the classic innovator’s dilemma, there will be plenty of opportunities for the new computational journalists to reinvent the way news information is gathered, organized, presented, and disseminated.