International Journal of Communication 12(2018), 473–488 1932–8036/20180005

100 Billion Data Rows per Second: Media Analytics in the Early 21st Century

LEV MANOVICH
City University of New York, USA
Cultural Analytics Lab, USA

This article describes the newest stage in the development of modern technological media. I call this stage "media analytics." It follows the previous stages of massive reproduction (1500–), broadcasting (1920–), the use of computers for media creation workflows (1981–), the Web as global content creation and distribution network (1993–), and social media platforms (2004–), to name just a few such stages. Unlike other stages, the new stage is not focused on new mechanisms for creation, publishing, or distribution of media, although it also affects these operations. Instead, this new stage is about automatic computational *analysis* of the content of all online digital media, personal online behaviors and communication, and automatic actions based on this analysis.

Keywords: machine learning, social media, culture industry, user-generated content, data science

> Culture today is infecting everything with sameness. Film, radio, and magazines form a system. . . . Interested parties like to explain culture industry in technological terms. Its millions of participants, they argue, demand reproduction processes that inevitably lead to the use of standard processes to meet the same needs at countless locations. . . . In reality, the cycle of manipulation and retroactive need is unifying the system ever more tightly. (Horkheimer & Adorno, 1944/2002, pp. 94–95)

> Scuba is Facebook's fast slice-and-dice data store. It stores thousands of tables in about 100 terabytes in memory. It ingests millions of new rows per second and deletes just as many. Throughput peaks around 100 queries per second, scanning 100 billion rows per second, with most response times under 1 second. (Wiener & Bronson, 2014, para. 9)

> Our data is literally a *big deal*. Measuring every second of engagement on every single page on most every major website in the globe means a scientifically defined insane amount of data. (Chartbeat, 2015, para. 2)

Lev Manovich: Manovich.lev@gmail.com
Date submitted: 2016–08–08

The history of technological media can be imagined as a series of many overlapping stages. At each stage, new technologies and new practices for creating, storing, distributing, and using content become prominent. But these practices do not replace each other in a linear fashion. Instead, the older ones continue to coexist along with the new ones. For example, consider mass reproduction of print (1500–), broadcasting (1920–), use of personal computers for media creation (1981–), the Web as a publishing and distribution platform (1993–), and social networks and media sharing sites (2004–), to name just a few of these practices. All of them are active today, although over long periods of time, the earlier practices may become less important or get transformed in significant ways.

This article aims to describe the newest stage in the development of modern technological media. I call this stage "media analytics." Unlike other stages, it is not at its core about creation, publishing, or distribution, although it also affects these operations. The core of this new stage is automatic computational *analysis* of the content of all media available online as well as online personal and group behaviors and communication. Because the industry does not have a single term to refer to these practices, I will go ahead and coin a name for them. Let us call them *media analytics*.

The motivations and uses of media analytics are multiple, but they all are related to the scale of digital culture in the early 21st century. This scale is the volume of digital content—the Web has 14 billion Web pages, 2 billion photos are shared daily, media streaming service Spotify has 30 million songs, and so on. It is also the numbers of people who share, interact with, or purchase this content. As of early 2017, there were 2.5 billion active social network users and 3.7 billion Internet users, and these numbers continue to grow. Therefore, to say that media analytics and the rise of the "big data" paradigm are related is an understatement: In fact, Google and Facebook developed the next generation of technologies to store, retrieve, and analyze big data that are now also used in other fields because of their volumes of media and interaction records.

Media Analytics Examples

The companies that sell cultural goods and services via websites or apps (e.g., Amazon, Apple, Spotify, and Netflix), organize and make searchable information and knowledge (Google, Baidu, and Yandex), provide recommendations (Yelp, TripAdvisor), and enable social communication, information sharing (Facebook, QQ, WeChat, WhatsApp, Twitter, etc.), and media sharing (Instagram, Pinterest, YouTube, and iQiyi) all rely on computational analysis of massive media data sets and data streams. These data include the following:

- Traces of users' online behavior (i.e., digital footprints): visiting websites, following links, sharing posts and "liking," viewing and clicking on ads
- Traces of physical behavior: geographical location, date and time when a user posts to social networks, location of a user computer connected to the Internet
- Media content created by companies: songs, video, books, and movies
- Media content created by users of social networks: posts, conversations, images, and video

I am using the term *data sets* to refer to static or "historical" data organized in databases before automatic analysis. The term *historical* in industrial data analytics applications means everything that is more than a few seconds, or sometimes even fractions of a second, in the past. *Data streams* refers to the data that arrive in real time and are analyzed continuously using platforms such as Spark Streaming and Storm (Asay, 2015). In both cases, collected data are also stored using platforms such as Cassandra, HBase, and MongoDB. So far, digital humanities and computational social sciences have only been analyzing historical static data sets; meanwhile, the industry has been increasingly using real-time analysis of data streams that are larger and require special platforms mentioned earlier.

Let us consider one example of the computational analysis of media content and the use of this analysis. Spotify analyzes many characteristics of each music track in its collection of more than 30 million tracks. These characteristics, or "features," are also made available to external developers via the Spotify API method "Get Audio Features for a Track" (Spotify, 2016). The current specification for this method lists 13 features. Many of them are built on top of more low-level features extracted by algorithms from the track audio file. These features are "acousticness," danceability, duration in milliseconds, energy, "instrumentalness," key, "liveness," loudness, mode, "speechiness," tempo, time signature, and valence.

Spotify and other music streaming services use such extracted features to automatically create custom playlists for users starting with a song, album, artist, or genre. You can start with a single song, and the app's algorithms select and stream songs that are close to it in a feature space. The advantage of this method is that the new songs do not have to belong to the same album or artist—they only need to share some musical features with the previous songs.

There are numerous other examples of media analytics. For example, to make its search service possible, Google continuously analyzes full content and markup of billions of Web pages. It looks at *every* page on the Web that its spiders can reach—its text, layout, fonts used, images, and so on, using over 200 signals in total (Google, 2016a). E-mail spam detection relies on analysis of texts of numerous e-mails. Amazon analyzes purchases of millions of its customers to recommend books. Netflix analyzes choices of millions of subscribers to recommend films and TV shows. It also analyzes information on all its offerings to create more than 70,000 genre categories (Madrigal, 2014). Contextual advertising systems such as AdSense analyze content of Web pages and automatically select the relevant ads to show. Video game companies capture gaming actions of millions of players and use this to optimize game design. Facebook's algorithm analyzes all updates by all friends of every user to automatically select which ones to show in user feed (if you are using default "Top Stories" option; Dredge, 2014). And it does this for all posts of their 1.6 billion users. (According to the estimates by Mikael Huss [2014], in 2014, Facebook was processing 600 TB of new data per day.) Other examples of the use of media analytics in the industry include automatic translation (Google, Skype) and recommendations for people to follow or add to your friends list (Twitter, Facebook). Using the voice interface in Google Search, Google Voice Transcriptions (Beaufays, 2015), Microsoft Cortana, or Siri also depends on computational analysis of millions of hours of previous voice interactions.

The development of algorithms and software that make this data collection, analysis, and subsequent actions possible is carried out by researchers in a number of academic fields, including

machine learning, computer vision, music information retrieval, computational linguistics, natural language processing, data mining, and other areas of computer science. Many of these fields started to develop in the 1950s, with the key concept of "information retrieval" appearing around 1950. The newest term is *data science*, which became popular after 2010. It refers to professionals who know contemporary algorithms and methods for data analysis (described today by overlapping terms of *machine learning*, *data mining*, and *AI*) as well as classical statistics, and can implement gathering, analysis, reporting, and storage of big data using current technologies, such as platforms I referenced earlier.

People outside the industry may be surprised to learn that many key parts of media analytics technologies are open-sourced. To speed up the progress of research, most top companies regularly share many parts of their code. For example, on November 9, 2015, Google open-sourced TensorFlow, its data and media analysis system that powers many of its services (Pichai, 2015). Other companies, such as Facebook and Microsoft, also open-sourced their software systems for organizing massive data sets (Cassandra and Hive are two popular systems from Facebook, and they are now used by numerous commercial and nonprofit organizations). The reverse is also true: The data from community mapping project Openstreetmap.org (with more than 2 million members) is used by many commercial companies, including Microsoft and Craigslist, in their applications (Sawers, 2014). The most popular programming languages used for media analytics research today are open source R and Python.

If we want to date the establishment of the practices of the massive analysis of content and interaction data across the culture industry, we may pick up 1995 as the starting date (early Web search engines) and 2010 (when Facebook reached 500 million users) as the date these practices fully matured. Today, media analytics is taken for granted, with every large company offering social networking or selling media goods online, doing this daily and increasingly in real time. The same analysis is performed by hundreds of companies that offer social media dashboards—Web tools for monitoring and analyzing user activity and posting content—and perform custom analysis for numerous clients, both profit and nonprofit. (Their customers include private and public universities.)

The Two Parts of Media Analytics

Media analytics is the new stage of media technology that impacts everyday *cultural* experiences of significant percentages of populations in dozens of countries who use the Internet and computing devices. One part of media analytics—the practices of gathering and algorithmic analysis of user interaction data (i.e., digital traces)—has received significant attention. However, most discussions of these practices are focused on political and social issues such as privacy, surveillance, access rights, discrimination, fairness, biases, and so on, as opposed to history and theory of technological media.

The second part of media analytics—the practices of algorithmic analysis of all types of online media content by the industry—has received less attention in comparison. However, only if we consider the two parts of media analytics together—analysis of user interaction data and analysis of media content—does the magnitude of the shift that gradually took place between 1995 and 2010 become fully apparent. Although articles in popular media have discussed details of computational analysis of cultural content and data in some cases, such as Google Search, Netflix's recommendation system, or the 2008

Obama election campaign, they have not explained that media analytics is now used throughout the culture industry (Segal, 2011; Vanderbilt, 2013).

Media analytics practices and technologies are employed in most platforms and services where people share, purchase, and interact with cultural products and with each other. They are used by companies to automatically select *what* will be shown on these platforms to each user, and *how* and *when*, including updates from friends and recommended content. Perhaps most important, they are built into many apps and Web services used not only by companies and nonprofits but also by millions of individuals who now participate in the culture industry not only as consumers but also as content and opinion creators—George Ritzer and Nathan Jurgenson (2010) called such combination of consumption and production "prosumer capitalism." For example, Google Analytics for websites and blogs, and analytics dashboards provided by Facebook, Twitter, and other major social networks are used by millions to fine-tune their content and posting strategies.

Both parts of media analytics are historically new. At the time when Max Horkheimer and Theodor Adorno (1944/2002) were writing their book, interpersonal and group interactions were not part of the culture industry. But today, they have now also become "industrialized"—influenced in part by algorithms deciding what content, updates, and information from people in your networks to show you. These interactions are also industrialized in a different sense—interfaces and tools of social networks and messaging apps are designed with input from UI (user interaction) scientists and designers who test endless possibilities to ensure that every UI element, such as buttons and menus, is optimized and engineered to achieve maximum results.

The second part of media analytics—computational analysis of media content—is also very recent in terms of its use by the culture industry. The first computer technologies that could retrieve computer-encoded text in response to a query were introduced in the 1940s. In a conference held in in 1948, "Holmstrom described a 'machine called the Univac' capable of searching for text references associated with a subject code. The code and text were stored on a magnetic steel tape" (Sanderson & Croft, 2012, paras. 13–14). Calvin Mooers coined the term *information retrieval* in his master's thesis at MIT and published his definition of the term in 1950. According to this definition, information retrieval is "finding information whose location or very existence is a priori unknown" (quoted in Garfield, 1997, para. 2). While the earliest systems only used subject and author codes, in the late 1950s, IBM computer scientist Hans Peter Luhn introduced full-text processing that I identify as the real start of media analytics.

In the 1980s, the first search engines applied information retrieval technology to the files on the Internet. After the World Wide Web started to grow, new search engines for websites were created. The first well-known engine that searched text of websites was WebCrawler, released in 1994. In the second part of the 1990s, many search engines, including Yahoo!, Magellan, Lycos, Infoseek, Excite, and AltaVista, continued analysis of Web text. And in the 2000s, massive analysis of other types of online media, including images, video, and songs, also started. For example, in early 2016, an image search service by TinEye indexed over 14 billion Web images ("How Many Images," 2016). Music streaming services such as Spotify and Deezer analyze characteristics of millions of songs and use this analysis for recommendation. By early 2017, Spotify analyzed 30 million songs and automatically generated 2 billion

playlists. YouTube analyzes content of posted videos to see if a new video matches an already existing item in the database of millions of copyrighted videos (Google, 2016b).

Automation of Media Analysis

If we look at the cultural analytics stage of media history in terms of automation, it follows the earlier stage when software tools and computers were adapted for *authoring* individual media products (see Manovich, 2013a, for a detailed discussion of this history and its cultural effects.) The important moments in this history are introductions of the Quantel Paintbox for video effects (1981), Microsoft Word for writing (1983), Amiga for video editing (1985), PageMaker for desktop publishing (1985), Illustrator for vector drawing (1987), and Photoshop for image editing (1990). These software tools made possible faster workflows, exchanging and sharing of projects' digital files and assets, creation of modular content (e.g., layers in Photoshop), and the ability to easily change parts of the created content in the future. Later, these tools were joined by other technologies that enable computational media authoring, such as render farms and media workflow management.

The tools of media analytics are different—they automate the analysis of (1) billions of pieces of media content available online, and (2) data from trillions of interactions between users and software services and apps. For example, Google analyzes content of images on the Web, and when you enter a search term, the system shows all or only some images depending on your selection in the Safe Search option. And if this is desired, they also make possible automatic actions based on this analysis—for example, automatic ads placement.

So what is now being automated is no longer a creation of individual media items, but a presentation of all Web content and retrieval of relevant content. This includes selection and filtering (what to show), content placement (behavioral advertising), and discovery (search, recommendations). Another growing application is "how to show"—for example, popular news portal Mashable that has 8.5 million followers on Twitter (as of March 2017) automatically adjusts the placement of content pieces based on real-time analysis of users' interactions with this content. Another application of media analytics is "what to create"—for example, in 2015, *The New York Times* writers started to use an in-house application that recommends topics to cover (for other examples, see LeCompte, 2015; Podolny, 2015).

Just as the adoption of computers for media authoring gradually democratized this process, the development of concepts, techniques, software, and hardware for media analytics also democratizes its use. Today, every creator of Web content has free tools that until recently were only available to big advertising agencies or marketers. Every person who runs a blog site or posts content on her or his social media networks can now act as a media company, studying the data about clicks, reshares, and likes, paying to promote any post, and systematically planning what and where she or he shares. All popular media sharing and networking platforms, from Facebook, YouTube, and Twitter to Academia.edu, show people detailed graphs and statistics on interaction with network users with their content.

As another example, consider MailChimp, the popular service for sending and tracking mass e-mails. When I use MailChimp to send an e-mail to my small mailing list (MailChimp is currently free for up

to 2,000 e-mail addresses and 12,000 e-mails per month), I use their Send Time Optimization option. It analyzes data from my previous e-mail campaigns and "determines the best sending time for the subscribers you're sending to, and distributes it at the optimal time" (MailChimp, 2017, para. 5). To create my posts for Facebook and Twitter, I use the Buffer app, which also calculates the best time for me to post to each network. If I want to promote my Facebook page or Twitter posts, I can use the free advertising features that can create a custom audience for my campaign by selecting users on their networks based on hundreds of settings, including country, age, gender, interests, and behaviors. While category-based market segmentation was already used earlier in marketing and advertising, Twitter also allows you to "target users who are similar to the people who already follow you" of any of the accounts you specify (Twitter, 2017, para. 7). In this new situation, I no longer have to start with explicit categories or terms—instead, I can let Twitter's media analytics build a custom audience for me.

In the case of Web giants such as Google and Facebook, their technical and talent resources for data analysis and access to the data about the use of their services by hundreds of millions of people daily give them significant advantages. These resources allow these companies to analyze user interactions and act on them in ways that are *quantitatively* different from an individual user or a business using Google Analytics or Facebook analytics on their own accounts, or using any of the social media dashboards—but *qualitatively*, in terms of concepts and most of the technologies, it is exactly the same. One key difference between giants such as Google, Facebook, Baidu, and eBay, and smaller companies is that the former have top scientists developing their machine learning systems (i.e., the modern form of AI) that analyze and make decisions based on billions of data points captured in near real time. Another difference is the fact that Google and Facebook dominate online search and advertising in many countries, and therefore, they have a disproportional effect on discovery of new content and information by hundreds of millions of people.

So media analytics is big, and it is used throughout the culture industry. But still, why do I call it a "stage," as opposed to just one among other "trends" of the contemporary culture industry? Because in some industries, media analytics is used to algorithmically process and act on *every* cultural artifact. For example, digital music services that use media analytics accounted for 70% of music revenues in the United States in 2014 (Richter, 2014). Media analytics is also used to analyze and act on *every* user interaction on platforms used by the majority of younger people in dozens of countries (e.g., Facebook, Baidu, Tumblr, Instagram, etc.). It is the new logic of how media works internally and how it functions in society. In short, it is crucial both practically and theoretically. Any future discussion of media theory or communication has to start with this situation.

(Of course, I am not saying that nothing else has happened after 1993 in media technologies. I can list many other important developments, such as the move from hierarchical organization of information to search, the rise of social media, the integration of geolocation information, mobile computing, the integration of cameras and Web browsing into phones, and the switch to supervised machine learning across media analytics applications and other areas of data analysis after 2010.)

Companies that are key players in "big media" data processing are only 10 to 15 years old—Google, Baidu, VK, Amazon, eBay, Facebook, Instagram, etc. They developed in a Web era, as opposed to the older 20th-century cultural industry players such as movie studios or book publishers. These older

players were, and continue to be, the producers of "professional" content. The newer players act as interfaces between people and this professional content, as well as "user-generated content." The older players are gradually moving toward adoption of analytics, but key decisions (e.g., publishing a book) are still made by individuals following their instincts. In contrast, new players from the beginning built their business on computational media analytics.

What they analyze and optimize is primarily distribution, marketing, advertising, discovery, and recommendations, that is, the part of the culture industry where customers find, purchase, and "use" cultural products. However, the same computational paradigms are also implemented by social network companies. From this perspective, the users of these networks become "products" to each other. For example, Amazon algorithms analyze data about what goods people look at and what they purchase and use this analysis to provide personal recommendations to each of its users. In parallel, Facebook algorithms analyze what people do on Facebook to select what content appears in each person's News Feed (Luckerson, 2015). (According to the current default setting, Facebook will show you only some of these posts, which it calls "Top Stories," automatically selected by its algorithms. This setting can be changed by going to the News Feed tab and selecting "Most Recent" instead of "Top Stories.")

Although the word *algorithms* and the term *algorithmic culture* are convenient because they seem to nicely sum up the concepts of automatic analysis and decision making, they can be also misleading—and that is why I use *analytics* instead. The most frequently used technology today for big data analysis and prediction is machine learning, and it is quite different from our common understanding of an algorithm as a finite sequence of steps executed to accomplish some task. Some machine learning applications are "interpretable," but many, if not the majority, are not. The process of creating a computer system often leads to a "black box," which has good practical performance but is not interpretable, that is, we do not know how it generates results (Annany et al., 2015; Mencar, 2013). For these reasons, I prefer to avoid using the terms *algorithms* and *algorithmic* when referring to the real-world systems deployed by companies to analyze data, make predictions, or execute automatic actions based on this analysis. My preferred term is *software*, which is more general—it does not assume that the system uses traditional algorithms, nor that these algorithms are interpretable (Manovich, 2013b).

Media analytics is the key aspect of "materiality" of media today. Fifteen years ago, this concept may have been used in discussions of computer hardware, programming languages, databases, network protocols, and media authoring, publishing, and sharing software (Manovich, 2013a). Today, media materiality is also about big data storage and processing technologies such as Hadoop and Storm, paradigms such as supervised machine learning and deep learning, and the popular machine learning algorithms such as k-means, decision trees, support vector machines, and kNN. Materiality is Facebook "scanning 100 billion rows per second" (Wiener & Bronson, 2014, para. 9) and Google processing 100+ TB of data per day (estimated in Huss, 2014). Materiality is also Google automatically creating "multiple [predictive] models for every person based on the time of the day" (Woodie, 2015, para. 18).

Automation of Media Actions

So far, our discussion has focused on automatic analysis of media content and user interactions with the content. I now want to discuss another novel aspect of media culture today that is enabled by media analytics: automation of "media actions" based on the results of previous and/or real-time analysis. These actions can be divided into two types: (1) automatic actions partly controlled by explicit user's inputs or chosen settings and (2) automatic actions not controlled by explicit user inputs.

Examples of *automatic actions partly controlled by explicit user's inputs or chosen settings* include search results returned in response to a text search query, image search results produced in response to the user choosing an image type to find, and music tracks recommended by a music streaming service in response to the user's initial selection of a musician or tracks. For example, Google image search options currently have a choice of face, photo, clip art, line drawing, or animation, and full color, a dominant color, and black and white. Examples of settings that can be changed by users are ads chosen by the system to show in response to the user's ad preferences, and types of images shown in response to "safe search" settings.

These users' inputs and settings are combined with the results of content and interactions analysis to determine the actions taken by the software. The choice of actions may combine previous data from the particular user and data for all other users—such as purchasing history of all Amazon customers. Other information can be also used to determine actions. For instance, real-time algorithmic actions that involve thousands of ads determine which ads will be shown be on the user's page at a given moment.

Automatic actions not controlled by explicit user inputs depend on the analysis of user interaction activity but do not require the user to choose anything explicitly. In other words, a user "votes" with all his or her previous actions. The automatic filtering in Google e-mail into "Important" and "Everything" is a good example of this type of action. Most of the automatic actions we encounter in our interactions with Web services and apps today can be partly controlled by us via settings; however, not every user is willing to spend time to understand and change the default settings for every service (e.g., https://www.facebook.com/settings).

We also divide automatic actions into two types, depending on whether they are arrived at in a deterministic or nondeterministic way.

Deterministic actions are produced by computation that always generates the same outputs given the same inputs.

Nondeterministic actions are produced by computation that may generate many different outputs given the same inputs. Today, most algorithmic decision making that uses big data relies on probability theory, statistics, and machine learning. This includes automatic decision making in Web services and apps of the culture industry. For example, a recommendation system may generate different results every time by adding a random parameter to vary results. But even when a computational system uses

deterministic methods, it can still generate different actions every time if the data used as input have changed—as typically is the case with constantly evolving Web or social networks.

The overall result is another new condition of media—what we are shown and recommended every time is not completely determined by us or by system designers. This shift from strictly deterministic technologies and practices of the culture industry in the 20th century to nondeterministic technologies in the first decade of the 21st century is another important aspect of the new stage of media culture. What was strictly the realm of experimental arts—use of indeterminacy by John Cage, or stochastic processes by Iannis Xenakis to create or perform compositions now—in a way, has been adopted by the culture industry as a method to deal with the new massive scale of available content. But of course, the goal now is rather different—not to create a possibly uncomfortable and shocking aesthetic experience, but to expose a person to more of the existing content that fits with the person's existing taste, as manifested in her or his previous choice. However, we should keep in mind that industry recommendation systems can be also used to expand one's taste and knowledge, if one gradually keeps moving further from his or her initial selections—and certainly Web hyperlinking structure, Wikipedia, open-access publications, and all kinds of other Web content can be also used to do this.

In addition to the examples of automatic actions based on media analytics I already mentioned, there are many other types of such actions that also make contemporary media different from the past. For example, the data on users' interaction with a Web service, app, or device are also often used to make automatic design adjustments in the Web service, app, or device. The data also are used to create more cognitive automation, allowing the system to "anticipate" what users many need at any given location and time and deliver the information best tailored to this location, moment, user profile, and type of activity. The term *context-aware* is often used to describe computer systems that can react to a location, time, identity, and activity (Dey & Abowd, 1999). The Google Now assistant is one example of such context-aware computing.

Twentieth-century industrial and software designers and advertisers used user testing, focus groups, and other techniques to test new products and to refine them. But in the media analytics stage, a service or a product can automatically adjust its behavior for each individual user based on his or her interaction history as well as analysis of interactions of every other user with the service or product. Following the model popularized by Google, every Web and app user has become a better tester of many constantly changing systems that learn from every interaction.

Media Analytics and Cultural Analytics

Many of the cultural effects—as opposed to economic, social, and political—of the new computational organization of media culture have not been yet systematically studied empirically by either industry or academic researchers. For example, we know now many things about the language of conservative and liberal Twitter users in the United States or the political polarization on the same platform (Association for Psychological Science, 2015; Kaplan, 2015). But we do not know anything about the differences in types of content shared on Instagram in thousands of cities worldwide, or the evolution in topics of hundreds of millions of blogs over the past 10 years. The industry does extract some of this

information and uses it in search and recommendation services, but it does not publish this information. We should also keep in mind that the industry is typically interested in the analysis of the current trends in relation to certain content and user activities (e.g., all social media mentions of a particular brand), as opposed to historical or large-scale cross-cultural analysis that is of interest to academics.

However, one thing is clear to me. The same data analysis methods that are used in the culture industry can be also used to quantitatively research and theorize cultural effects of media analytics. In 2005, when industrial media analytics was just emerging, I introduced a term, *cultural analytics*, to refer to the use of computational methods to explore massive samples of contemporary digital media to ask questions relevant to media studies and humanities. Over the past 10 years, researchers in computer science, computational social science, and digital humanities have published tens of thousands of interesting studies that apply these methods to the analysis of literature, music, art, historical newspaper content, and social networks, including Facebook, Twitter, Flickr, and Instagram. (For an overview, see Manovich, 2015.) However, computational analysis of large volumes of media content—such as images, video, and sound, as opposed to users' online activities—has not yet become the norm in media and communication studies. To motivate such future research and also give it a name, we can coin the term *computational media studies*.

In their 1944 book *Dialectic of Enlightenment*, Horkheimer and Adorno (1944/2002) introduced a term *culture industry*. The book was written in Los Angeles when the Hollywood studio system was in its "classical"—that is, most integrated—period. There were eight major film conglomerates, and five of them (Fox, Paramount, RKO, Warner Brothers, and Loew's) had their own production studios, distribution divisions, theater chains, directors, and actors. According to some film theorists, the films produced by these studios during this period also had a very consistent style and narrative construction (Bordwell, Thompson, & Staiger, 1985). Regardless of whether Horkheimer and Adorno already fully formed their ideas before arriving in Los Angeles as emigrants from Germany, the tone of the book and its statements, such as the famous quote, "Culture today is infecting everything with sameness," seem to fit particularly well with the Hollywood classical era (Horkheimer & Adorno, 1944/2002, p. 94).

How does the new "computational base" (i.e., media analytics) affect both the products that the culture industry creates and what consumers get to see and choose? For example, do computational recommendation systems used today by Amazon, YouTube, Netflix, Spotify, Google Play, and other companies help people choose apps, books, videos, movies, or songs more widely (i.e., long tail effect), or do they, on the contrary, guide them toward "top lists"? What about systems used by Twitter and Facebook to recommend to us whom to follow and which groups to join? (For an example of the industry publication that presents details of its recommendation system, see Gupta & Singh, 2013; for an example of the analysis of the effects of an industry recommendation system on media consumption, see Zhou, Khemmarat, & Gao, 2010.)

Or consider the interfaces and tools of popular media capture and sharing apps, such as Instagram, with its standard set of filters and adjustment controls appearing in certain order on the user's phone. Does this lead to homogenization of image styles, with the same few filters dominating the rest? These questions about diversity versus homogeneity can now be studied quantitatively using large-scale

cultural data from the Web and computational methods for data analysis. For example, in our Cultural Analytics Laboratory (http://lab.culturalanalytics.info), we compared the use of Instagram filters in 2.3 million photos shared in 13 global cities and found remarkable consistency between the cities (Hochman & Manovich, 2013). The relative frequencies of different filters were similar across the cities, and their popularity was almost perfectly correlated with the order of their appearance in the Instagram app interface.

Digitization of historical cultural media also makes it possible to analyze the diversity versus homogeneity dimension of culture historically. A group of researchers published an article titled "Measuring the Evolution of Contemporary Western Popular Music" (Serrà, Corral, Boguñá, Haro, & Arcos, 2012), in which they applied computational methods to a data set of 464,411 distinct music recordings for the 1955–2010 period. Recently, many researchers from computer and information sciences also have been studying the aesthetic preferences and dynamics of attention in social networks. As an example of such an article, consider "An Image Is Worth More Than a Thousand Favorites" (Schifanella, Redi, & Aiello, 2015). The article presents "analysis of ordinary people's aesthetics perception of Web images" using 9 million Flickr images with Creative Commons licenses. Reviewing the large body of quantitative research that uses large data, the authors stated,

> The dynamics of attention in social media tend to obey power laws. Attention concentrates on a relatively small number of popular items and neglecting the vast majority of content produced by the crowd. Although popularity can be an indication of the perceived value of an item within its community, previous research has hinted to the fact that popularity is distinct from intrinsic quality. As a result, content with low visibility but high quality lurks in the tail of the popularity distribution. This phenomenon can be particularly evident in the case of photo-sharing communities, where valuable photographers who are not highly engaged in online social interactions contribute with high-quality pictures that remain unseen. (para. 1)

The authors proposed an algorithm that can find "unpopular" images (i.e., images that have been seen by only a small proportion of users) that are equal in aesthetic quality to the popular images. Implementing such algorithm would allow more creators to find audiences for their works. Such research exemplifies the potential of computational media studies to go beyond generating descriptions and "critique" of cultural situations by offering constructive solutions that can change these situations.

Although the use of large-scale computational media analysis of content and interaction data from hundreds of millions of users gives top companies such as Google, Facebook, Instagram, Amazon, and Netflix lots of power, we have to remember that they are not simply the new iterations of tightly integrated Hollywood conglomerates from the 1940s. If the 20th-century culture industry was creating, distributing, and marketing content (movies, books, songs, and TV programs), the newer cultural industry of our own time (i.e., the companies such as those just listed) is focusing on organizing, presenting, and recommending content created by others. Media analytics—analyzing media content and people's online interactions—is done to support this goal and to support advertising on these platforms, which is often their main source of income. (In other words, these companies in most cases are not content creators themselves.) These "others" include both professional producers and hundreds of millions of ordinary

casual users, as well as millions of people who are situated on many points in between these extremes. The examples are social media mini-celebrities; people who work freelance or have studios such as fitness and yoga instructors, hair stylists, or interior decorators; owners of small shops; creators of anime music videos; 35 million artists who share their works on Deviantart.com; 28 million academics who have accounts on Academia.edu; and so on.

And the content itself is also qualitatively different from what was produced at the time when Horkheimer and Adorno wrote their book (1940s): It is not only songs, films, books, and TV shows, but also our individual posts, messages, images, videos shared on Twitter, Facebook, Vine, Instagram, YouTube, and Vimeo, academic papers, code, and so on. If all content published by the entire culture industry in the 1940s in the United States probably was under a few million items per year, today all content shared on social networks adds up to many billions of items every day. "Surfacing" the variability of this content so we can understand and interpret it can only be done using computational methods. Until recently, these methods have been used only by computer scientists—but, just as the new fields of digital humanities, digital history, and digital art history have now started to apply them in their own fields, it is only a matter of time before media studies will start doing the same.

References

Ananny, M., Karahalios, K., Sandvig, C., & Wilson, C. (2015). *Auditing algorithms from the outside: Methods and implications*. Retrieved from https://auditingalgorithms.wordpress.com/rationale/

Asay, M. (2015, March 23). Beyond Hadoop: The streaming future of big data. *InfoWorld*. Retrieved from http://www.infoworld.com/article/2900504/big-data/beyond-hadoop-streaming-future-of-big-data.html

Association for Psychological Science. (2015, August 27). Political polarization on Twitter depends on the issue. *ScienceDaily*. Retrieved from http://www.sciencedaily.com/releases/2015/08/150827083423.htm

Beaufays, F. (2015, August 11). The neural networks behind Google Voice transcription [Blog post]. Retrieved from https://research.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html

Bordwell, D., Thompson, K., & Staiger, J. (1985). *The classical Hollywood cinema: Film style and mode of production to 1960*. New York, NY: Columbia University Press.

Chartbeat. (2015). *About*. Retrieved from https://chartbeat.com/about/

Dey, A. K., & Abowd, G. D. (1999). *Towards a better understanding of context and context-awareness*. Retrieved from ftp://ftp.cc.gatech.edu/pub/gvu/tr/1999/99-22.pdf

Dredge, S. (2014, June 30). How does Facebook decide what to show in my news feed? *The Guardian*. Retrieved from https://www.theguardian.com/technology/2014/jun/30/facebook-news-feed-filters-emotion-study

Garfield, E. (1997). A tribute to Calvin N. Mooers, a pioneer of information retrieval. *The Scientist, 11*(6), 9. Retrieved from http://www.garfield.library.upenn.edu/commentaries/tsv11(06)p09y19970317.pdf

Google. (2016a). *How search works*. Retrieved from https://www.google.com/search/howsearchworks

Google. (2016b). *How Content ID works* [Video file]. Retrieved from https://support.google.com/youtube/answer/2797370?hl=en

Gupta, A. & Singh, K. (2013). Location based personalized restaurant recommendation system for mobile environments. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*. Sri Jayachamarajendra College of Engineering (SJCE), Mysore, India. doi:10.1109/ICACCI.2013.6637223

Hochman, N., & Manovich, L. (2013). Zooming into an Instagram city: Reading the local through social media. *First Monday*, *18*(7). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4711/3698

Horkheimer, M., & Adorno, T. W. (2002). *Dialectic of enlightenment* (E. Jephcott, Trans.) Stanford, CA: Stanford University Press. (Original work published 1944)

How many images are in TinEye's search index? (2016, February 21). *TinEye*. Retrieved from https://www.tineye.com/faq#count

Huss, M. (2014, June 24). Data size estimates [Blog post]. Retrieved from https://followthedata.wordpress.com/2014/06/24/data-size-estimates/

Kaplan, K. (2015, September 18). Your Twitter feed says more about your political views than you think, study says. *Los Angeles Times*. Retrieved from http://www.latimes.com/science/la-sci-sn-twitter-political-conservative-republicans-20150917-story.html

LeCompte, C. (2015, September 1). Automation in the newsroom. *Nieman Reports*. Retrieved from http://niemanreports.org/articles/automation-in-the-newsroom/

Luckerson, V. (2015, July 9). Here's how Facebook's news feed actually works. *TIME*. Retrieved from http://time.com/3950525/facebook-news-feed-algorithm/

Madrigal, A. C. (2014, January 2). How Netflix reverse engineered Hollywood. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

MailChimp. (2017, October 5). Use send time optimization. Retrieved from https://kb.mailchimp.com/delivery/deliverability-research/use-send-time-optimization

Manovich, L. (2013a). *Software takes command*. London, UK: Bloomsbury.

Manovich, L. (2013b, December 16). The algorithms of our lives. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/The-Algorithms-of-Our-Lives-/143557/

Manovich, L. (2015). *The science of culture? Social computing, digital humanities, and cultural analytics*. Retrieved from http://manovich.net/index.php/projects/cultural-analytics-social-computing

Mencar, C. (2013, July 2). What do you mean by "interpretability" in models? *ResearchGate*. Retrieved from https://www.researchgate.net/post/What_do_you_mean_by_interpretability_in_models

Pichai, S. (2015, November 9). TensorFlow: Smarter machine learning, for everyone [Blog post]. Retrieved from https://googleblog.blogspot.com/2015/11/tensorflow-smarter-machine-learning-for.html

Podolny, S. (2015, March 7). If an algorithm wrote this, how would you even know? *The New York Times*. Retrieved from http://www.nytimes.com/2015/03/08/opinion/sunday/if-an-algorithm-wrote-this-how-would-you-even-know.html

Richter, F. (2014, September 30). Digital accounts for nearly 70% of U.S. music revenues. *Statista*. Retrieved from https://www.statista.com/chart/2773/digital-music-in-the-united-states/

Ritzer, G., & Jurgenson, N. (2010). Production, consumption, prosumption: The nature of capitalism in the age of the digital "prosumer." *Journal of Consumer Culture*, *10*(1), 13–36. doi:10.1177/1469540509354673

Sanderson, M., & Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE, 100*, 1444–1451. Retrieved from http://ciir-publications.cs.umass.edu/getpdf.php?id=1066

Sawers, P. (2014, February 28). The rise of OpenStreetMap: A quest to conquer Google's mapping empire. *The Next Web*. Retrieved from http://thenextweb.com/insider/2014/02/28/openstreetmap/

Schifanella, R., Redi, M., & Aiello, L. (2015). An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr pictures. *arXiv*. Retrieved from http://arxiv.org/pdf/1505.03358.pdf

Segal, D. (2011, February 12). The dirty little secrets of search. *The New York Times*. Retrieved from http://www.nytimes.com/2011/02/13/business/13search.html?_r=2

Serrà, J., Corral, Á., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the evolution of contemporary Western popular music. *Scientific Reports, 2*. doi:10.1038/srep00521

Spotify. (2016). Get audio features for a track. Retrieved from https://developer.spotify.com/web-api/get-audio-features/

Twitter. (2017). Follower targeting on Twitter. Retrieved from https://business.twitter.com/en/targeting/follower.html

Vanderbilt, T. (2013, August 7). The science behind the Netflix algorithms that decide what you'll watch next. *Wired*. Retrieved from http://www.wired.com/2013/08/qq_netflix-algorithm/

Wiener, J., & Bronson, N. (2014, October 22). Facebook's top open data problems. *Facebook*. Retrieved from https://research.fb.com/facebook-s-top-open-data-problems/

Woodie, A. (2015, February 9). The rise of predictive modeling factories. *Datanami*. Retrieved from https://www.datanami.com/2015/02/09/rise-predictive-modeling-factories/

Zhou, R., Khemmarat, S., & Gao, L. (2010). The impact of YouTube recommendation system on video views. *Proceedings of the 2010 ACM Internet Measurement Conference*. Melbourne, Australia. Retrieved from http://conferences.sigcomm.org/imc/2010/papers/p404.pdf