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

Forthcoming in *International Journal of Communication*, special issue "Digital Traces in Context."

Lev Manovich | manovich.net
Professor, Ph.D. Program in Computer Science,
The Graduate Center, City University of New York.
Director, Software Studies Initiative | softwarestudies.com

Analytics dashboard in offices of Vogue UK. Screenshot from video "Alexa Chung Uncovers Fashion Industry Secrets - Full Series One | Future of Fashion | British Vogue, "published on YouTube.com on October 27, 2015, http://www.youtube.com/watch?v=Bi2nc_xxnvs.

"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 is ever more tightly." Theodor Adorno and Max Horkheimer, "The Culture Industry: Enlightenment as Mass Deception," in *Dialectic of Enlightenment*, 1944, http://web.stanford.edu/dept/DLCL/files/pdf/adorno_culture_industry.pdf.

Facebook 2015 stats: "Photo uploads total 300 million per day"; "968 million people log onto Facebook daily"; "50% of 18-24 year-olds go on Facebook when they wake up." Source: "The Top 20 Valuable Facebook Statistics," October 2015, https://zephoria.com/top-15-valuable-facebook-statistics/.

"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." *Facebook Top Open Data Problems*, 2014, https://research.facebook.com/blog/1522692927972019/facebook-s-top-open-data-problems/.

"Being able to iterate quickly on thousands of models will require being able to train and score models simultaneously. This approach allows Cisco (an H20 customer) to run 60,000 propensity to buy models every three months, or to allow Google to not only have a model for every individual, but to have *multiple models for every person based on the time of the day*." Alex Woodie, "The Rise of Predictive Modeling Factories," February 9, 2015, http://www.datanami.com/2015/02/09/rise-predictive-modeling-factories/.

"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." About page from chartbeat.com, social media monitoring and optimization platform, https://chartbeat.com/about/, accessed 11/25/2015.

In the early 21st century, culture industry was significantly reshaped by "big data" paradigm - but as of now, only some elements of this shift have been described by journalists and academics.

(I am using the term *culture industry* as opposed to "digital culture" or "digital media" because today *all* culture industries create digital products that are disseminated digitally online. This includes games, movies, music, TV shows, e-books, online advertising, apps, etc. So I don't think we need to add word "digital" anymore when we are talking about culture.)

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

- traces of users' online behavior: browsing pages, following links, sharing posts and "liking," selecting content items to play, view or read, clicking on ads;
- traces of physical activity: places and times when users post to social networks, online gameplay actions;
- media content created by companies – songs, video, books, movies;
- media content created by users of social networks – posts, conversations, images, video

Similarly, human-computer interaction – for example using voice interface in Google Search, Google Voice Transcriptions, Microsoft Cortana, or Siri – also depend on computational analysis of millions of hours of previous voice interactions.

(Note about terminology: I use the term "data sets" to refer to static or "historical" data organized in databases prior to automatic analysis. The term "historical" in industrial data analytics applications mean 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 arrives in real time and is analyzed continuously using platforms such as Spark Streaming and Storm. In both cases, collected data is also stored using platforms such as Cassandra, HBase, and

MongoDB. So far, digital humanities and computational social sciences have only been analyzing historical static datasets; meanwhile industry has been increasingly using real-time analysis of data streams that are larger and require special platforms mentioned above.)

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 its spiders can reach - its text, layout, fonts used, images and so on, using over 200 signals in total. (Web search was the first massive instantiation of media analytics.) To be able to recommend music, the streaming services such as Spotify and Deezer analyze characteristics of millions of songs. For example, Echonest that powers many online music services used its algorithms to analyze 36,774,820 songs by 3,230,888 artists. Email spam detection relies on analysis of texts of numerous emails. Amazon analyzes purchases of millions of its customers to recommend books. Netflix analyzes choices of millions of subscribes to recommend films and TV show. It also analyzes information on all its offerings to create over 70,000 genre categories. 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 games design. YouTube scans posted videos to see if a new video matches some item in the database of millions of copyrighted videos. Facebook algorithm analyzes all updates by every friends of every user to automatically select which ones to show in user feed (if you are using default "Top Stories" option). And it does this for all posts of their 1.6 billion users. (According to the estimates, in 2014 Facebook was processing 600 TB of new data per day.) Other examples of use media analytics in the industry include automatic translation (Google, Skype) and recommendations for people to follow or add to your friends list (Twitter, Facebook).

The development of algorithms and software systems that make this data collection, analysis and subsequent actions possible is carried out by researchers in a number of academic fields including data science, machine learning, data mining, computer vision, music information retrieval, computational linguistics, natural language processing, and computer science in general. Most of these fields started to develop already in the 1950s, with the key concept of "information retrieval" introduced in 1950. The newest term is "data science" that became popular after 2010. It refers to professionals who know contemporary algorithms and methods for data analysis (described by overlapping umbrella terms of "data mining," "machine learning," 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 above.

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. Other companies such as Facebook and Microsoft also open-sourced their software systems for organizing massive datasets (Cassandra and Hive are two popular systems from Facebook and they are now used by numerous commercial and non-profit organizations.) The reverse is also true: the data from community mapping project openstreetmap.org (with over two million members) is used by many commercial companies including Microsoft and craigslist in their applications. The most programming language used for media analytics research today is free R that is constantly being researchers from universities, labs, and companies.

If we want to date the establishment of the practices of the massive *analysis of content and interaction data* across 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 taking 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 also perform custom analysis for numerous clients, both profit and non-profit. (Their customers include private and public universities.)

Computational analysis of content of all cultural products being created and user interactions with this content and with each other is the new stage in the development of modern technological media. It follows the previous stages of massive reproduction (1500-), broadcasting (1920-), automation of media authoring using computers (1981-), use of web for creation of content and distribution (1993-), to name just a few. Since the industry does not have a single term to refer to all practices that characterize this new stage, we can go ahead and coin a name for it. Let's call it **media analytics**.

To the best of my knowledge, this novel aspect of contemporary media culture has not yet been discussed systematically in the academy. And while articles in popular media have covered computational analysis of cultural content and data

in some particular cases, such Google Search, Netflix recommendation system, or 2008 Obama election campaign, they have not explained that media analytics is now used throughout culture industry.

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 Internet and computing devices. (For figures about use of Internet and social media in the USA by different demographics groups, see latest Pew Research Center Internet & Tech reports.)

To be fair, we should note that one part of media analytics – the practices of *gathering and algorithmic analysis of user interaction data* - received significant attention. However, almost all discussions of this have been only in relation to political and social issues such as privacy, surveillance, access rights, discrimination, fairness, biases, etc., as opposed to *history and theory of technological media*.

In contrast, the second key part - the practices of *algorithmic analysis of all types of online media content by the industry* - received very little attention. One likely reason for this absence is that many journalists and academics in social sciences and media studies are interested mainly in social and political effects and uses of media, as opposed to "technical details" beneath its surface. While media analytics technologies and concepts are widely discussed in computer and data sciences in business publications, in conferences and trade shows, in leading science journals, and being taught to millions of students worldwide in computer science and data science classes, they are not discussed in either popular press or by academics outside of technology and science fields.

This lack of systematic knowledge on the part of many academics and journalists who write about digital cultures about the details of the computational processes that drive web services, apps, desktop applications, video games, search, image detection, voice recognition, recommendation systems, behavioral advertising, and so on, as well as contemporary software engineering and the field of data science in general often prevents them, in my view, from seeing the full picture. (Understanding of many of these details does require knowledge of computer science, and today very few people in academic humanists, social sciences or journalism have this background.) This is the reason of why many academics and journalists recently adapted the single term "algorithm" (or "algorithmic") to refer to the sum total of many very different computational processes and data infrastructures where algorithms is only one of many parts.

In particular, people often use this term to refer to systems that use *supervised machine learning* and therefore *are not algorithmic* in the accepted meaning of this concept. As Ian Bogust correctly noted in *The Atlantic* (01/15/2015), "Concepts like "algorithm" have become sloppy shorthands, slang terms for the act of mistaking multipart complex systems for simple, singular ones." (In the same article Bogust incorrectly describes me as somebody who focuses on *algorithms*, while in reality I have been advocating the study of *software*, the term I use to refer to such "multipart complex systems." See my book Software Takes Command, 2013). For example, while many presentations at innovative conferences on Governing Algorithms in 2013 and Algorithms and Accountability in 2015 organized by NYU Law Institute made interesting and important arguments, some of the presentations used the term "algorithms" too broadly.

Only if we consider the two parts of media analytics together - analysis of user interaction data and analysis of cultural content – the magnitude of the shift that took place between 1995 and 2010 becomes fully apparent. This is why I am proposing that we should think of *media analytics* as the new condition of *culture industry* and also as a new stage in *media history*. Because its use is now so central to industry as a whole, and because it affects all cultural activities mediated by the web and the apps, we need to start thinking beyond any particular instances.

To reiterate this point: the algorithmic analysis of "cultural data" and algorithmic decision-making is not only at work in a few most visible areas such as Google Search and Facebook News. 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, how, and when* will be shown on these platforms to each user, including updates from their friends and recommended content. And perhaps most importantly, they are built into many apps and web services used not only by *companies and non-profits* but also by *millions of individuals* who now participate in culture industry not only as consumers but also as content and opinion creators. (George Ritzer and Nathan Jurgenson call 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 Adorno and Horkheimer were writing their book, *interpersonal and group interactions* were not part of 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 assure that every UI element such as buttons and menus is optimized and engineered to achieve maximum results.

The computational analysis part is also very recent in terms of its use by culture industry. The idea and first computer technologies that could perform retrieve the computer-encoded text in response to a quarry were already introduced in 1940s. In the 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 and Croft, The History of Information Retrieval Research). Calvin Mooers coined the term "information retrieval" in his Master Thesis at MIT paper and published his definition of the term in 1950 ("finding information whose location or very existence is a priori unknown," quoted in see Eugene Garfield, A Tribute To Calvin N. Mooers, A Pioneer Of Information Retrieval, 1977). 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, first search engines applied information retrieval technology to the files on the internet. After the World Wide started to grow, new search engines for the websites were created. The first well-known engine that searched texts of web sites was 1994 Web Crawler. In second part of the 1990s, many search engines including Yahoo!, Magellan, Lycos, Infoseek, Excite, AltaVista continued analysis of web text. And in 2000s, the massive analysis of other types of online media including images, video and songs also started. For example, in early 2016 image search service by TinEye indexed over 14 billion web images (https://www.tineye.com/faq#count, retrieved 2/21/2016).

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 my Software Takes Command for 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 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 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 are now being automated are no longer creation of individual media items but *presentation* of all web content and *retrieval* of relevant content. This includes *selection and filtering* (what to show), *promotion* (advertising of content), and *discovery* (search, recommendations). Another growing application is "how to show" – for example, popular news portal Mashable that currently has 6.73 followers on Twitter (https://twitter.com/mashable, 02/21/2016) automatically adjusts the placement of content pieces based on real time analysis of users' interactions with this content. Yet another application is "what to create" – for example, in 2015 New York Time writers started to use in-house application that recommends topics to cover (for other examples, see Shelley Podolny, If an Algorithm Wrote This, How Would You Even Know? 03/08/2015, and Celeste Lecompte, Automation in the Newsroom, 09/01/2015).

Just as the adoption of computers for media authoring gradually democratized this process, the development of concepts, techniques, software, and hardware (i.e., computer clusters) for media analytics also democratizes its use. Now 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/his social media networks can now act as a media company, studying the data about clicks, re-shares, and likes, and paying to promote any post, and systematically planning what and where she/he shares. All popular media sharing and networking platforms, from Facebook, YouTube and Twitter to acamedia.edu show people detailed graphs and statistics on

interaction with network users with her content. As another example, consider MailChimp, the popular service for sending and tracking mass emails. When I use MailChimp to send an email to my small mailing list (MailChimp is currently free for up to 2,000 email addresses and 12,000 emails per month), I use their Send Time Optimization option. It analyzes data from my previous email campaigns and "determines the best sending time for the subscribers you're sending to, and distributes it at the optimal time." To create my posts for Facebook and Twitter, I use Buffer app that 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 custom audience for my campaign by selecting users on their networks based on hundreds of settings including country, age, gender, interests, behaviors. While category-based market segmentation was already used earlier in marketing and advertising, Twitter also allow you to "reach users with interests similar to followers" of any of the accounts you specify. 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 people daily gives them significant advantages. It allows 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, 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 culture industry. But still, why do I call it a "stage" as opposed to just one among other "trends" of 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 U.S in 9/2014). Media analytics is also used to analyze and act on

every user interaction on platforms used by majority of younger people in dozens of countries (i.e., Facebook, Baidu, Tumblr, Instagram, etc.). It's 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, media theory or communication has to start with this situation.

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

The companies that are key players in "big media" data processing are all only 10-15 years – 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 towards adoption of analytics, but key decisions (for example, publishing a particular 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, i.e. the part of culture industry where customers find, purchase, and "use" cultural products. However, the same computational paradigms are also implemented by social network services. 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 appear in each person News Feed. (According to the current default setting, Facebook will show you only some of these posts it calls "Top Stories" automatically selected by its algorithms. This setting can be changed by going to News Feed tab and selecting "Most Recent" instead of "Top Stories.")

(Although the word "algorithms" or the term "algorithmic culture" are convenient because they seems to nicely sum up the concepts of automatic

analysis and decision making, they can be also misleading – and that's 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 of machine learning applications are "interpretable," but many, if not majority, are not – the process of creating a computer system leads to a "black box" which has good practical performance but is not interpretable, i.e. we don't know how it generates results. For a useful discussion, see What do you mean by interpretability in models on researchgate.com. For a description of the research to "audit algorithms" see Auditing Algorithms From the Outside**.** For these reasons, I think that it's better 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 data 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. See the section "Can we analyze the code of software programs?" in my 2013 article in *Chronicle of Higher Education*.)

Media analytics is the key aspects of "materiality" of media today. In other words: *Materiality* now is not only about hardware, or databases, or media authoring, publishing and sharing software as it was in early 2000s (see again Software Takes Command). Today, it is also about big data storage and processing technologies such as Hadoop and Storm, paradigms such as supervised machine learning, the particular data analysis trends such as "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" and Google processing 100+ TB of data per day (2014 estimate). Materiality is also Google automatically creating "multiple [predictive] models for every person based on the time of the day."

Let me summarize and systematize the previous discussion:

Media analytics refers to two types of practices: 1) automatic analysis of media content and user interactions with the content, and 2) automatic actions based on the results of this analysis.

1. The *analysis* part is always fully automated. The results of analysis can be used to drive actions, but this is not required. The action part is also fully automated

and it be generated in response to user inputs, or without them. Google search offers an example of the system where the actions depend on previous analysis and user inputs. Google continuously indexes all web pages including dynamically generated content and content of apps it can access. This is analysis part. When a user enters input into the search interface using text, image or voice, Google systems return the results drawn from index. This is the action in response to user input.

Use of social media monitoring tools today is an example of analysis typically not connected to automatic actions. I can use Buffer, Hootsuite, Sprout Social, Piwik , and dozens of other free or paid tools to analyze user engagement with my own websites and social media accounts, or social media activity in general related to any topic, in many languages, and across dozens of global social networks. After I discover some patterns that I want to change, I may adjust my strategy of posting to Twitter, Facebook or Instagram, but these adjustments would not happen automatically.

The *analysis* practices can be divided into three types:

1.1. *Analysis of media content*. Examples include content of web pages and apps analyzed by search engines; analysis of photos and their metadata to detect faces or enable categorization by places and content performed by photo apps and photo sharing services; YouTube analysis of newly shared video to compare them to its database of copyrighted video and detect copies.

1.2. *Analysis of user interactions with content*. Examples of interactions include choosing particular items in the list of search results, "liking" content on Facebook, Twitter, or Instagram, retweeting, clicking on online ads, and viewing, or reading, watching, or listening to content items in multiple media (papers on academia.edu, products on Amazon, music tracks on Spotify, etc.).

1.3. *Analysis of user's interactions with other users* of a given service. For example, on Facebook I can start following a particular user; add this user (with her/his permission) to my friends list; write a message; and also "poke," "report," and "block." All these behaviors are recorded and analyzed by Facebook and used in some of its systems that drive certain automatic actions such as deciding which new items to show to each user.

2. The *action* practices can be divided into two types:

2.1. Automatic *actions partly controlled by explicit user's particular inputs or chosen settings.* Examples of inputs: search results produced in response to a text search query; filtered image search results produced in response user choosing type of image (Flickr currently offers search by key color, "minimalist," "patterns," or image orientation); similar music tracks chosen by a music streaming service in response to user's initial selection of a musician or tracks. Examples of using settings that can be changed by users: ads chosen by the system to show in response to user's ad preferences; types of image shown in response to "safe search" settings.

Users inputs and settings are combined with the results of content and interactions analysis to determine the actions. The interactions may combine previous interaction 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 auctions that involve thousands of ads determine which ads will be shown be on the user's page at a particular moment.

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

Finally, we also *divide automatic actions into two types*, depending on whether they are arrived at in deterministic or non-deterministic way:

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

2. *Non-deterministic actions*. These actions use 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 culture industry. (For example, a recommendation system may generate different results every time because it may use randomness to vary results).

Note that even if deterministic system is used in web service or app, it can still generate different actions every time if the data used as input has changed – as it typically the case with constantly evolving web or social network service content.

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 culture industry in the 20th century to non-deterministic technologies in the first decade of the 21st century is another important aspect of media analytics. What was strictly the realm of experimental arts – use of indeterminacy by John Cage, or stochastic processes by Iannis Xenakis to create and/or perform compositions has now, in a way, has been adopted by culture industry as a way to deal with the new massive scale of available content. But of course, the goal and the method now is rather different – not to create possibly uncomfortable and shocking aesthetic experience but to expose a person to more of existing content fits with a person existing taste, as manifested in her/his previous choice. However, we should keep in mind that industry recommendation system can be also used to expand your taste and knowledge, if you gradually keep moving further from your initial selections - and certainly web hyperlinking structure, Wikipedia, open access publications and all other kinds of web content can be used to do this.

One thing I should add to my outline above is another important use of collected interaction data that also makes the new media analytics stage different. The data on users' interaction with the web service, an app or a device is also often used to make *automatic design adjustments* of this web service, app or a device. It is also used to create more *cognitive automation*, allowing the system to "anticipate" what uses need at any given location and time, and deliver the information best tailed 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 location, time, identity, and activity. Google Now assistant is a good example of such context-aware computing.

The automatic changes in order of menus shown by a system to a particular user based on this user her/his interaction history is an example of design automation that uses interaction data. Of course, 20th century industrial and software designers and advertisers also 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 this user interaction history as well as the analysis of every other user of this service or products. Following the model popularized by Google, every web and app user had become a better tester of many constantly changing systems that learn from every interaction.

At this point you, the reader, may get impatient and wonder when I will deliver what critics and media theorists are supposed to deliver when they talk about contemporary life and in particular use of technologies: a critique of what I am describing. Where is the word "critical" in my text? Why I am not invoking "capitalism," "commodity," "fetishism," or "resistance"? Why I am not talking about "hidden agenda" and "biases" of data technologies, or "end of privacy"? Where is my moralistic judgment?

None of this is coming. Why? Because, in contrast to what media critics like to tell you, I believe that computing and data analysis technologies are *neutral*. They don't come with some built-in social and economic ideologies and effects, and they are hardly the tools of capitalism, profit making, or oppression. Exactly the same analytics algorithms (linear regression, k-means cluster analysis, Principal Component Analysis, and so on) or massive data processing technologies (Cassandra, MongoDB, etc.) are used to analyze people's behavior in social networks, to look for cure for cancer, to look for potential terrorists, to select ads that appear in your YouTube video, to study the human microbiome, to motivate people to live healthy lifestyles, to get more people to vote for a particular candidate during presidential elections (think of use of analytics in Obama 2008 and 2012 campaigns), to suggest to New York Times editors which stories they should publish, to generate automatic news layouts on Buzzfeed, etc. Media analytics benefits not only big companies but also many millions of small business, freelances and non-profits. The same algorithms and data gathering, storage and analysis technologies are used by companies and government agencies in USA, UK, Russia, Brazil, China, and dozens of other countries for thousands of different applications. They are used to control and to liberate, to create new knowledge and to limit what we know, to help find love and to encourage us to consume more, to spy on us and to help us escape surveillance, to organize protests and to track them. In other words, their use is so varied that any claim that they are "tools of capitalism" is simply ungrounded (unless you also want to also claim that arithmetic, calculus, rhetoric, electricity, space flight, and every other human technology ever invented are all tools of capitalism.)

This does not mean that the adoption of large-scale data processing and analysis across culture industry does not significantly change it. Nor does it mean that it is now any less of an "industry," in the sense of having distinct forms of organization and standardization (such as are "likes," "favorites," line graphs showing numbers of people engaging with your content, or maps showing the countries where these people are located.) On the contrary – some of marketing and advertising techniques, the ways companies engage customers online and also cultural products are new, and they are in the last few years all came to rely on big scale media analytics.

Many of the *cultural* (as opposed to economic, social, and political) effects of these developments have not been yet systematically studied empirically by either industry or academic researchers. For example, we know now many things about the language by conservative and liberal Twitter users in the U.S. or political polarization on the same platform. But we don't know anything about the differences in types of content shared on Instagram in thousands of cities worldwide, or the evolution in cultural topics in hundreds of millions of blogs over last ten years. The industry does extract some of this information and uses it in their search and recommendation services, but they don't publish this information itself. We should also keep in mind that industry is typically interested in the analysis of the current trends in relation to particular content and user activities (for example, 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 culture industry to select and standardize content and communication can be also used to quantitatively research and theorize cultural effects of media analytics. (In our lab we have been using such methods to analyze visual content such as millions of Instagram images, but not yet large interaction data). But such analysis will gradually emerge, and we already can give it a name: *computational media studies*.

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 cultural datasets including user-generated content in humanities context. Since then, researchers published lots 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, The Science of Culture? Social Computing, Digital Humanities, and Cultural Analytics, 2015.) However, since computational analysis of content or user interactions data has not yet been used in media and communication studies, the term *computational media studies* can be useful to motivate this research.

The term "culture industry" that is used in the title of this text was introduced by Adorno and Horkheimer in their 1944 book *Dialectic of Enlightenment*. The book was written in Los Angeles when Hollywood studio system was in its "classical," i.e. most integrated period. There were eight major film conglomerates; five of them (Fox, Paramount, RKO, Warner Brothers, and Loew's) had their production studios, distribution divisions, theatre chains, and their own 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 (see David Bordwell, Janet Staiger, Kristin Thompson, *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*, published in 1985.) Regardless of whether Adorno and Horkheimer already fully formed their ideas before arriving to Los Angeles as emigrants from Germany, the tone of the book and its particular statements such as famous "culture today is infecting everything with sameness" seem to fit particularly well to Hollywood classical era.

How does the new "computational base" (i.e., "media analytics") affect both the products culture industry creates, and what consumers get to see and choose? For example, do computational recommendation systems used today by Amazon, YouTube, Netflix, Spotify, Apple iTunes Radio, Google Play and others help people chose apps, books, videos, movies, or songs more widely (i.e., long tail effect), or do they, on the contrary, guide them towards "top lists"? What about recommendation systems used by Twitter and Facebook to recommend to us who to follow and which groups to join? 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 particular order on your phone. Does this lead to homogenization of image styles, with the same few filters dominating over the rest (currently 24 in total)?

These questions - such as diversity vs. homogeneity - can now be studied quantitatively using large-scale cultural data from the web and modern computational methods for data analysis. For example, in my lab we compared the use of Instagram filters in 2.3 million photos shared in 13 global cities, and found remarkable consistency between the cities. Digitization of historical

cultural content also makes it possible to analyze this question historically. In 2012 a group of researchers published a paper titled Measuring the Evolution of Contemporary Western Popular Music where they applied computational methods to the dataset of 464,411 distinct music recordings for 1955 - 2010 period. Recently, many researchers from computer and information sciences have also been studying the aesthetic preferences and dynamics of attention in social networks. As an example of such paper, consider 2015 papers An Image is Worth More than a Thousand Favorites from the scientists at Yahoo Research Labs in Barcelona. The paper presents "analysis of ordinary people's aesthetics perception of web images" using nine million Flickr images with Creative Commons licenses. Reviewing the large body of quantitative research that uses large data, the authors state:

> 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.

The authors propose an algorithm that can find "unpopular" images (i.e. images that have been seen by only small proportion of users) that equal in aesthetic quality to the popular images. Implementing such algorithm would allow more creators to find audiences for their works. Such research exemplifies 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, TV

programs), the newer cultural industry of our own time (i.e., the companies such as the ones listed above) is focusing on *organizing, presenting, and recommending content created by others as well as capturing and analyzing people interactions with this content.* (In other words, these companies are not content creators themselves.) These "others" include both professional producers and hundreds of millions of ordinary casual users, as well 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 studies such fitness and yoga instructors, hair stylists or interior decorators; 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 Adorno and Max Horkheimer wrote their book (1940s): it is not only songs, films, books and TV shows but also our individual posts, messages, images and video shared on Twitter, Facebook, Vine, Instagram, YouTube, Vimeo, academic papers, code, etc. If content published by all culture industry in the 1940s in the U.S. probably was under a few million items per year, today content shared on social networks constitutes contains many billions of items every day. "Surfacing" the variability of this content so we can understand and interpret it can only be done using the computational methods. Until recently, these methods have been only used by computer scientists - but, just as the new fields 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.

[Fall 2015 – Spring 2016]