Analyzing Big Data

Bodó, B.; van de Velde, B.

DOI
10.1007/978-3-030-16065-4_20

Publication date
2019

Document Version
Submitted manuscript

Published in
The Palgrave Handbook of Methods for Media Policy Research

Big Data & Data Science in information law and policy research

Balázs Bodó
Institute for Information Law, University of Amsterdam, bodo@uva.nl

Bob van de Velde
Informatics Institute, University of Amsterdam

Abstract

We present two projects that used big data, and data science methods to support law and policy research with empirical evidence on digital media production and consumption. The simple case concerns the automatic scraping of news media websites to gather data on what is being published by news organizations. The complex case is about Robin, a research infrastructure which allows volunteers to donate their web browsing data stream so the process of personalized communications online can be studied. We discuss the issues researchers need to consider during the planning, data collection, and analysis phases of big data based research. We conclude that despite the limitations, difficulties and well-justified critique, social scientists, legal scholars, and researchers working in the humanities need to develop individual skills, and institutional competencies in big data methods, because data science is quickly becoming to be an indispensable part of the methodological tool-set of these disciplines.

Introduction

Big Data, and increasingly Data Science, have become key buzzwords prevalent in many discussions about technology, business, media, policy and virtually all other fields. To separate true benefits from buzz, gains from graft, and learning from loss, this chapter discusses how to leverage the ‘big data’ / ‘data science’ toolkit(s) to benefit policy formulation and evaluation of media policy. The toolkit includes data collection tools such as scraping, simulation, online experiments, monitoring and crowdsourcing, but also data analysis tools such as automated detection of sentiment, subjectivity, speech types, topics, and other content *at scale*. The promise of data-based approaches are equally wide, ranging from epistemological to ethical to practical concerns. Both good and bad are represented in two how-to’s that center around the case of media-content production and consumption monitoring.

The domain in which we discuss big data based approaches is evidence based media policy. Media policy, especially in Europe is very much concerned with health of the public sphere and the media system which hosts the democratic debates within society. In the late 2010’s social media platforms grew to play a prominent role in the circulation of news, with many potential

concerns due to the weaponization of social media by malicious state actors, the prevalence of fake news, polarization of public debates, and the growth of filter bubbles due to personalization (Borgesius Zuiderveen et al. 2016; Bodó, Helberger, and de Vreese 2017). Consequently a strong need emerged from both policymakers and researchers to have a better understanding of the impact of algorithmic intermediaries on the information diet of citizens, on the societal circulation of news and other information, and in general of how the process of information personalization unfolds in the digital space.

This gap in the oversight and control capabilities of societies over algorithmic agents led to an algorithmic control crisis (Bodó et al 2017), prompting research into how data-based approaches can be developed to grasp the fragmented online media landscape. This contribution describes one such effort of setting up the data-collection, data-processing and data-analysis tools to monitor and study the personalized online information landscape. We'll spoil the take-away message: Big Data requires Big Planning.

So why engage in Big Data research? The answer is simple: the online media landscape is increasingly fragmented, diverse, personalized and opaque. Most information is received through websites, search, apps and other means. Asking people about media consumption may yield a plethora of websites or worse: aggregators such as google news or Facebook that do not tell you what content they actually saw. The times of doing content analysis on a small set of national newspapers and asking people which they read to know what they read are over. And not all data is easy to obtain simply by visiting the same websites. Services and content can be highly personalized, both in terms of advertisement as well as content. It is vital to formulate policy based on what is actually happening, but this is increasingly obscured. Big Data approaches can be used to create, collect, analyze and learn from actual behavior to support policy research.

What are big data and data science?

"Big Data", as a compound noun, has been around in 'the' (academic) literature for more than a decade. Early work focused mainly on *big* data in the sense of 'hard-to-manage'. Case in point is the 2009 paper by Adam Jacobs (2009) referring to 'big' not as hard to store (100 GB dataset can be stored in affordable commodity hardware), but as hard to analyze, as most prevalent databases cannot handle billions of rows of data. He remarks that such sizes can be attained even by single, relatively well-visited, websites with modest runtimes (i.e. months). Jeffrey Cohen et. al. (2009) still talk about a mostly technical approach, but argue specifically for the stronger integration (Magnetic) of heterogeneous (Agile) data that is used beyond operational purposes (Deep)[1]. In these contexts "Big Data" is a computing paradigm, a new way of analyzing data: not in the memory of a single (big) machine, but spread across multiple (small) machines by leveraging analysis techniques tailored towards large scale. Indeed, a whole industry of tools tailored towards such modes of big data analysis has developed (Chen & Zhang, 2014).

But Big Data is more than just a computing paradigm. Chris Anderson's infamous Wired article on "The end of Theory" (2008) pushes Big Data as a research paradigm. This paradigm is characterized by *inductive, data-driven, 'correlation is enough'* approaches enabled by the aforementioned large scale data and statistical tools. Enthusiasm also struck the business world

[1] This constitutes their "MAD" framework. Talk about tortured acronyms!

with perhaps over-optimistic reports by McKinsey (Manyika et al. 2011), the Harvard Business review (McAfee et al. 2012), and more nuanced books about the pro's and cons (Walker 2014; Mayer-Schönberger and Cukier 2013). Kitchin (2014) provides what is perhaps the most well-known definition of Big Data as three v's: volume that is too big handle, velocity because most of this data comes in in real-time, variety because most of this data does not conform to any explicit specification. Others have later added veracity, because such data is often not all reliable, variability, as data-inflow waxes and wanes rather than keeps up a continuous stream, *value,* as data is often near-worthless without analysis (see a summary in Gandomi & Haider, 2014). A key part of Kitchin's argument is that most of this Big Data -indeed the way most academics still talk about Big Data- is characterized as *trace data* generated as a by-product of day-to-day activities, e.g.: internet browsing, mobile phone use, records of transactions. Big Data approaches are then characterized as an *exploratory science* that looks for patterns in data generated without scientific design. Big Data is thus characterized as a method of science that leverages (trace) data generated at big scale for non-scientific purposes to computationally investigate scientifically interesting phenomena.

Data Science can be considered the "complement" of Big Data. It aims to bridge the resources and methodologies of Big Data to enable data-driven decision-making (Provost & Fawcett, 2013, Hazen et. al. 2014). To enable Big Data to aid in policy development, analysis and evaluation, social science must adopt the computational methods and investigative approaches required to tackle the size and nature of Big Datasets (Shah, Cappella & Neuman, 2015). The role of Data Science is that of bridging between the technical challenges and scientific needs. As such, data scientists require skills in the domain (here media policy research), programming (to extract, transform, load -ETL- big data into the right format to answer specific questions) and statistics (to analyze patterns). In this sense, Data Science is the 'glue' required to merge Big Data and social science.

There are some important lessons in this discussion. First, the technical roots of the Big Data move highlight the challenges in handling and analyzing this kind of data. Second, the use of Big Data specifically in the social science domain raises further challenges in translating research questions to scale-able analysis code and the results back to theoretically interpretable insights. Third, Big Data research has focused on post-hoc interpretations of data, with exploration as the central research posture. In a sense, Big Data in the literature is a way to let 'data speak for itself'.

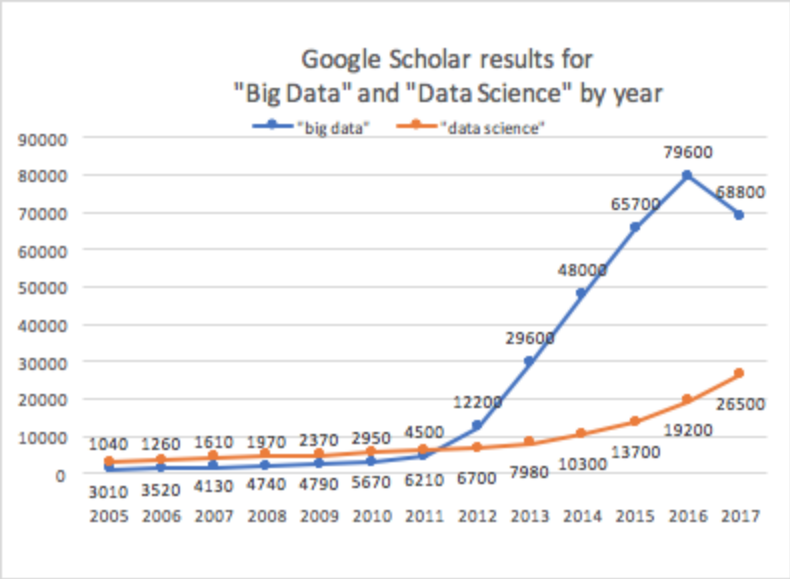

Fig 1: The growth of the number of scholarly works referencing Big Data and data Science[2]

The pros and cons of the Big Data approach

The promise of Big Data lies in the use of digital traces generated in unobtrusive ways. By virtue of scale and through the Data Science craft, such data should then provide a deeper insight into social processes. Because it is unobtrusive trace data, it should be externally valid (i.e. it works in the real world not just the lab). Because of its scale, even small effects can be captured. Because of its diversity, it should enable holistic interpretations of what is happening. In short, there is ample ground for epistemological exuberance. But there is no free lunch. There have been some notable critiques that any research looking to leverage Big Data should bear in mind. Here, we will distinguish three lines of critique: 1) Critiques dealing with Big Data research as a (novel) epistemology, 2) discussions of the ethical and legal issues related to the use of this type of data and 3) The more practical limitations faced by those who engage in Big Data research.

Conditions under which big data approach may make sense

There is an increasing number of questions in social sciences, which due to their particular structure could and sometimes do benefit from it. The two case studies that we later present were aimed at providing empirical evidence on the circulation of information in the public sphere to inform media policy. This domain is well positioned to be studied through big data methods for a number of reasons:

- Traditional methods (such as survey based methods to gather data on media exposure from a representative sample, media diary methods, etc.) have serious limitations. For example, recall bias, sample size, representativity are costly to address and overcome

[2] Drawn from a scholar.google.com searches of "Big Data" and separately "Data Science" using the 'from' and 'to' selector for each year (e.g. from 2012 to 2012, from 2013 to 2013) and relying on the total results indicator.

- Automated data collection is feasible as the behavior to observe is either takes place in the public (news production), or takes place via observable bottlenecks (news consumption on digital devices, social media, etc.)
- Digitization reduced the cost of digital automated data collection, as well as of analysis (content and metadata of published news can be collected automatically in the same step)
- A complete software ecosystem developed to facilitate data gathering (i.e. web scraping software packages for science)
- Complete market ecosystems emerged to collect, structure and sell data (data broker markets)

These developments radically reduced the costs of conducting media production and consumption related research, and enabled researchers to conduct large scale automated scraping and analysis of online resources, the observation of online behavior of large populations, and conduct online experiments

Critiques of big data on epistemological grounds

boyd & Crawford (2012) probably provide the most thorough epistemological critique of Big Data. First and foremost comes their challenge to the effect of the Big Data discourse on the concept of knowledge. Big Data is seen as a key part of the 'computational turn' in social science. Quoting, as humanities scholars do, Latour, boyd and Crawford raise concerns over the impact of Big Data as a tool on the conceptualization of social reality in science. They even go so far as to proclaim Big Data a new ontotheology in line with Berry (2011). In plain terms, they warn that the nature of Big Data as *trace data resulting as a by-product* means that reality can only be understood to the extent that it is captured in these by-products. The lack of designed measures not only biases attention in research but even reduces the space of possible inquiry. Second, boyd and Crawford attack the claim of objectivity and accuracy in Big Data research. Objectivity is -- in their eyes -- suspect because "claims to objectivity are necessarily made by subjects and are based on subjective observations and choices" (boyd & Crawford 2012:667). Another point is the ability to find any pattern imaginable provided the dataset is big enough (apophenia)[3]. The core here is that problems of quality in the dataset cannot be solved simply by scale alone and interpretation is at least as -- but quite possible more -- challenging when dealing with Big Data. Third, the sources of Big Data yield concerns over sample quality. Often used sources such as Facebook and Twitter API's hide the filtering mechanics employed by their retrieval algorithms. In addition, their users may not be representative of societies, thereby reducing external generalizability of findings based on platform-specific data[4]. This is compounded by the black-box

[3] Unfortunately, this argument is linked to a 2007 publication that is neither about Big Data nor about datasets that have enough directions to find any pattern imaginable. Perhaps the authors got a bit over-eager here.

[4] Political and communications science made heavy use of Twitter data in recent years because of its availability, leading to a whole body of literature on the impact of social media on democracy, public sphere, polarization, etc. In the wake of the fake news and weaponization of social media crises, Twitter (as well as Facebook) purged hundreds of thousands of profiles identified as bots, and fake accounts.

nature of these data sources with regard to the underlying quality of their storage and retrieval systems. Indeed, a 2014 paper by Lazer, Kenny, King and Vespignani (2014) showcases the failure of a Google-search based Big Data system that aimed to predict actual flu using time and geographically coded search terms. Finally, a data-driven concept drift may occur. The particular example of boyd and Crawford lies in social network analysis, where the analysis of social media ties has superseded self-reported friendship networks. They point out how there are strong conceptual differences between the two types of 'friendship ties', whereas a stronger tie in self-reported relations can express closeness, a strong behavioral tie regarding email ties to colleagues may not reflect the same affective closeness.

These points deserve careful attention in any case, but it is especially important to address them in studies which cannot rely on any other data source but those gathered by big data methods. The study of online communication is such a case, where the online production, dissemination, and consumption of, or exposure to digital information can only be meaningfully studied using such methods.

Ethical issues

The cataloguing and analysis of all the ethical, and legal issues around the collection and use of trace data is beyond the scope of this chapter. Bodó et al. (2017) provides a detailed overview of most issues, coupled with practical solutions. Here we only point at two issues we found most challenging to tackle.

First, the protection of (often highly sensitive) personal data, and the **privacy** of users is a central concern in Big Data studies. The European General Data Protection Regulation[5] (GDPR), formalizes in legal terms many ethical considerations: it requires that data subjects are informed about the data collection, it spells out the rights of data subjects vis-à-vis those who collect and process data, it defines the conditions and limits of data collection and analysis, among others (Bodó et al. 2017). While in the US different, more relaxed rules apply, the scope the GDPR is very wide, and applies to every research project which collects data from EU citizens. Researchers are advised to err on the side of caution, and strive towards GDPR compliance. Most big data projects will fall under the scope of GDPR, and even if not, the rules formulated there represent high ethical and legal data protection standards.

Second, Big Data research necessitates planning for the **afterlife of data**. Once collected, structured and analyzed, big data sets can be highly valuable, yet their reuse is often problematic. There are many issues to consider. The recent Cambridge Analytica scandal brought the commercial reuse of data collected for research purposes to the forefront (Cadwallad & Graham-Harrison 2018). But even for reasons of access and reproducibility, researchers should aim to set out clear guidelines regarding the (re-)use of gathered data for external researchers. In addition, both the platforms created to gather the data, as well as the gathered data should be open-source so that other institutions have the ability to replicate and expand on research. Yet, data protection

Lacking any follow up studies, we cannot even estimate how these accounts biased the results of virtually all studies from the last five years.

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. (L 119)

and Intellectual property rules may prevent the straightforward release of such datasets, as they may contain personal data, or the data may be protected by third party copyrights. Can a researcher share a collection of social media posts from regular consumers, or do they have a say in them? Can snippets be published as examples in papers? Can the data be shared to outside collaborators? These thorny questions need to be considered, and may require professional legal advice.

Practical limitations

The most important practical consideration relates to the **cost** of developing one's own data-generating research infrastructure, vs the cost of buying access to data collected by third parties. In part due to data-ownership, the use of existing / secondary data such as posts to social media platforms (Tweets, Facebook posts, Instagram pictures etc.) at scale will require 1) paid access through data-brokers (e.g. GNIP) or 2) Custom software build for research purposes. Either option provides some practical limitations: buying data can require substantial amounts of budget, while hiring the expertise to create and/or run software required for data collection can be costly and hard. Even the collection of primary data will require the use of specialized services (e.g. panel companies that collect tracking data) or the creation of custom collection, storage and analysis functionalities. The expenses and risks related to these approaches make Big Data research especially suitable for bigger, multi-research question projects rather than smaller projects of limited scope.

Finding the people is itself a hard task. A data scientist should be proficient in 1) Statistics (at scale), 2) Programming (including system administration) and 3) the domain (media policy theory). Data scientists that actually excel in all three fields are rare, and appropriately called 'unicorns'. Often, projects will be tempted to use a convenience mix existing faculty, but as we suggest in Case #2, there is no simple relay between a domain expert theoretician to a programmer who will then push some results to a statistician that finally returns an answer. Without technological and Big Data statistical knowledge, it is impossible to tell the feasible from the impossible, and easy to mistake clear skies for grey.

Getting started with Big Data

A Big Data project will generally entail (roughly) four stages: 1) Planning/Design, 2) Development, 3) Collection and 4) Analysis. Each of these stages contain a number of steps:

<u>Planning</u>

1. Planning your project
2. Getting a team together
3. Brainstorming questions & answers

<u>Development</u>

4. Choosing technology
5. Developing a pilot

Data collection

6. Moving into production
7. Maintaining the data collection infrastructure
8. Winding down
9. Data storage and the ever-after

Analysis

10. Doing analysis

In the following two case studies we demonstrate the issues related to the practical implementation of these steps. Case 1 is about setting up a system that tracks online news production. This system automatically scrapes the websites of online news media and stores the content for further analysis. Case 1 is the 'easy' case, as over the last decades a substantial amount of experience accumulated on web-scraping; the necessary software tools are rather sophisticated; and the data collection, storage and analysis procedures are standardized. The second case describes an effort to capture and analyze data on the use of personalized online services. As we describe, this case is 'hard', due to the complexity of the task, and the lack of similar efforts in the past.

Case 1: Scraping news sources (the 'easy' case)

In box

One of the key concerns of media policy is the 'diversity' of information which is produced by news organizations, and made available to the public. Many other questions about news production patterns, such as reliance on news wires, or reproduction of publications of other sources also require insights into what news content has been produced. Many such questions can be answered by a quantitative approach, through the automatic scraping and text analysis of online information sources. Newspapers are often available through aggregators (e.g. LexisNexis) or simply by subscribing to them. Online, sources may be more diverse, may update irregularly and may be unavailable from aggregators. This necessitates data collection using scraping technology. A scraper is essentially a script (small program) that retrieves content from a given webpage. This case is a 'generic' approach to scraping these web pages for later analysis.

Planning

To start automated collection of news or other content (scraping), you need to plan what, how and when to retrieve content. The main project decisions are simple: which sources do you wish to collect, how often (e.g. once per day) and for how long. The team can be quite small, depending greatly on the number of sources to be collected and the complexity of these sources. Student assistants or junior researchers trained in a scripting language should be able to create a webpage-specific scraper in about a week. Complicated websites that use javascript rendering

and anti-scraping techniques may require developers with particular experience in web technologies. With a small team, you can have a quick look at the websites to get a feeling for their complexity. Analysis depends on the researchers that want to work with the data, and should definitely include someone familiar with content analysis. Questions should be determined to know what the appropriate intervals of collection are (e.g. hourly, daily, monthly), which sources must be included, whether images or comment-sections should be included etcetera. It is often wise to include at least the publication time, author, url, body text and title of a page as these are the most often used features of webpages during analysis.

Development

For scraping purposes, it's good to know where the scrapers will run (e.g. on a laptop or PC, or a server somewhere). The main point is the reliable retrieval of the websites. Storage of webpages can generally be done on consumer harddisks. The resulting files of scraping can be stored in plaintext files, JSON-files, or in database solutions such as SQL, MongoDB or Elasticsearch. Depending on the size of the resulting collection, which is heavily contingent on runtime and collection interval, a database might be required for easy retrieval of results for automated content analysis. These transfer can be done post-hoc and does not necessarily be done before collection. In addition, the programming language used for scraping does not matter that much, although it is recommended to use a simple language that many people can work with (R or Python are currently favorites in the research field). A pilot should be run to test whether all necessary data of all sources is collected. Generally, the best way to proceed is to run for a week and try using the resulting data for some basic word-count over time analyses.

Data collection

After the retrieval scripts have been developed and tested, moving things to actual use (production) should be relatively trivial. The important thing is to regularly check whether the scripts have run and whether the data is correctly stored. When websites change, scrapers will often 'break' and stop collecting the right content. At these points, a new scraper must be developed. Setting up a warning system, such as an automated email on script failure, may help as an early-warning approach. After the data collection phase is over, it should simply be a matter of stopping the scripts from automatically running. It is generally a good practice to back-up the used scripts as well as logs of progress and the resulting data on a dedicated harddisk.

Data analysis

As scraping results in content data, content analysis methods of almost all kinds can be applied to the result. For example, quantitative tools can be used to produce time-series data by plotting the occurrence of words or phrases in documents over time. The results can also be analyzed through qualitative methods. With the appropriate software packages automatic sentiment detection, content classification, named entity recognition can be achieved, and the scarped documents can be automatically sub-setted based on various criteria, including, but not limited to the aforementioned ones.

Case 2: Tracking news consumption online (the ‘hard’ case)

In Box

The subject of the second case study is Robin. Robin is custom technology developed at the University of Amsterdam to study the process of personalized online communication, by collecting and analyzing how individual users interact with (potentially) personalized online services, such as digital advertising, e-commerce, social media, etc. Essentially no element in the process of personalized communications is observable from the outside, without the active support of one or both parties. Corporations do not share such information. Traditional methods, such as diary research methods cannot provide an accurate account of consumption, let alone exposure.

At the time of planning substantial legal, technical and methodological hurdles limited commercial data tracking services’ ability to provide insight into personalized communications. Consequently, the only possibility to study information personalization was via the design and development of a custom observation platform, which intercepts all online communications between users and the personalized services. In Bodó et al. (2017) we detailed the design process and the various legal, technological, scientific, organizational considerations and hurdles we faced before and during implementation. Here we provide an ex-post analysis, after the conclusion of the data collection period.

The most important roadblock of studying personalized online communications is that by nature such communication is private, and its external observability is very limited. Personalization is based on data collected by online service providers on the individuals, and personalized information is delivered to the individual to their (mobile) devices. The second case deals with a data-collection infrastructure to collect web-traffic of participants in a long-term study (see box).

Planning

Matching research objectives with methods

Studying online communications requires careful planning about 1) The technical process of data capture, storage and analysis, 2) Organizational factors, such as having partner organizations to arrange panels; the provision of technological infrastructure; and having adequate technical and theoretical expertise in the team, 3) The protocols to ensure that the data collection is legal and passes ethical scrutiny. Here, balances must be struck: retraining or hiring new staff were necessary, picking partners that are fully onboard with innovative research and weighing the richness of data with potential legal/ethical concerns. In our case, the planning phase produced a number of documents that guided different aspects of the implementation phase: the technical development, the legal compliance, the panel recruitment, the implementation of data security, access protocols, etc.

Getting a team together

The planning phase should reveal that the complexity of the task requires a team with very diverse, and wide reaching skill sets, the 'data scientists'. The long-term collection of live-stream data poses substantial technical, legal and organizational challenges, which all required the involvement of high level, specialized expertise:

- **Software development**. Specialized software development expertise was required to develop a number of different technologies: (1) a browser plugin to be installed and used by a general public; (2) back end data capture, processing and storage systems based on big data technology stacks; (3) systems of technical monitoring and maintenance to ensure the non-stop availability of the data capture over a year; and (4) custom data analytics technologies to extract structured data from the unstructured data stream. The nature of data collection required non-stop technical supervision necessitating the involvement of a professional technology service provider. While a number of companies offer software development services for academia, few specialize in the development of consumer facing (as opposed to scientist facing) software. Few (academic) companies provide big-data software that can operate without interruption or nuisance.
- **Hardware:** Big data often requires big hardware. The capturing of trace data requires excellent network connectivity and high levels of availability. The high volume of data requires specialized data storage and retrieval infrastructures. The highly sensitive nature of the captured data requires state of the art data security. In-house solutions require more upfront investment and technical staff, but cloud providers may be more expensive. A good estimate of required machines and storage space is a must.
- **Participant recruitment, panel management**: Many online tracking projects work with a convenience sample as they recruit data donors from internet volunteers. This is hardly an option for studies that want to make claims based on representative samples. This necessitates the involvement of professional panel companies, which must be willing to expose their panel to innovate research, but also have adequate size panels, and support capacity.
- **Legal:** Capturing large amounts of unfiltered internet traffic raises serious data protection and privacy issues, both legal and ethical. While the research was hosted by an institute for information law, specialized in data protection and privacy issues, the tasks and responsibilities related to data protection and privacy are not that of a legal researcher. Creating an adequate legal framework, including informed consent, requires significant investment.
- **Project management:** The successful implementation of such a large scale project required the coordination of a wide and complex web of partners, including the hardware infrastructure provider, the panel provider, designers, researchers, developers, legal compliance, etc. Such coordination requires resources, skills and expertise, which is not usually available in the domains of legal or communications research. Because big data projects are generally outside the comfort zone of most collaborators, project management must provide budgetary control, flexibility and persistence during the planning, implementation and analysis phases.

- **Research and analysis:** Finally, big data based research requires advanced data analytics skill from scientist. In effect, researchers will need to be recruited or trained in programming and computational methods. Bringing in external support has limited potential for a number of reasons. First, it is very difficult for universities to compete for data scientists with firms on the job market. Second, there is no quality data scientist without their own research interest / agenda, meaning that they will not have time, capacity, or willingness to support unskilled researchers in data wrangling tasks. Third, without at least some expertise in data science, researchers are unable to formulate big-data specific research questions, and analytical approaches.

Brainstorming questions & answers

Big data based policy, and socio-legal research constitutes a radical departure from the traditional methods of these disciplines which tend to rely on desk research or doctrinal research. One needs to understand both the qualities and the limitations of the data that can be acquired with big data methods in order to be able to formulate appropriate research questions. To illustrate this with an example, researchers who are unfamiliar with the fundamentals of web technologies, and the technical processed of personalization will struggle to operationalize, and understand the limits of the detection and interpretation of the traces of online personalization.

Development

Beyond a certain level of complexity, the development of the research software and hardware infrastructure is probably best done by an external contractor. The functional specification, however, needs to be produced by the researchers. Also, researchers must take responsibility for certain technological choices. Where, in which country will the collected data be stored and analyzed? -- this question has technical, legal, and cost implications. Which browsers, and operating systems need to be supported? -- this question will affect the sample bias. What is the process of technological onboarding? -- this question will affect the sample bias and data quality. An early specification of research goals is crucial to make the right decisions when implementation faces tough tradeoffs between development costs, technological functionalities, and research goal limitations. Without an in-depth understanding of the trade-offs intrinsic to the development process, such questions are hard to answer correctly, and suboptimal choices have long-term consequences as they define the data that is collected later.

Data collection

Moving into production: User onboarding

Assuming all the testing and planning has been completed in the previous phase, the biggest challenge of moving into production is the onboarding of research participants. In the case of Robin this meant a complex process with the following steps: (1) participants were selected from a survey panel based on whether they complied with the technical requirements, and were invited to participate in the research; (2) participants were informed about the nature of the research, and the extent of the data collection. We invited a graphic designer specialized designing informed consent forms to design the information and consent acquisition process; and

we pre-tested the informed consent procedure on a small sample; (3) users who consented to participate needed to pass through a complicated software installation process, and were assigned a unique authentication key.

All of these steps were necessary either for ethical, legal or technical reasons. However, at each step, the number of participants dropped significantly, and at the end of the process only 10% of the eligible panel members was turned into data-producing research participants.

Maintaining the data collection infrastructure

While the data collection phase seems to be less challenging than the planning and development phase, it is important to keep in mind that it still requires substantial amounts of resources, and of a different kind than development. Things that will pop-up during the project can include:

- Monitoring and quality assurance of the data collection technical process, with adequate intervention, error handling and backup capacities,
- Customer service and panel management, that deals with issues such as users' technical and other questions,
- Preparedness to detect and handle data breaches, hacks, and other potentially devastating legal/technical issues,
- Monitoring and detecting deterioration in the panel (churn) or in the data quality (due to changes in the nature of the data).

All these tasks need resources, skills, and manpower, which need to be planned for in advance.

Winding down. Data storage and the ever-after

At the end of the data collection period, participants need to be safely off-boarded. Installed software needs to be removed, access credentials revoked, etc. Live, data collecting systems need to be shut down, and data needs to be preserved in long term storage for access and reproducibility.

Analysis

In our case no live data was analyzed. Since the analysis of real-time flow-type data requires different amounts of resources and types competencies than the analysis of static (stock) data, we do not discuss that here. Dealing with stock-type big data has enough challenges in itself, such as:

- **Data structure**: The data Robin collected was raw web log data, as it passed between the users' browsers and the servers on the Internet. Extracting structured information from vast amounts of unstructured data was a huge challenge that we could only address by developing a custom data extraction toolset.
- **Data quantity**: the sheer amount data makes data search and extraction tasks slow and burdensome. In our case, *only the index* of the collected data made up more than 3 Tb of data, or more than 310 million records, while the actual collected data, including images and the full content of the tracked websites was by orders of magnitude higher than that.

Managing these challenges requires familiarity with the data, and adequate data wrangling and software development skills.

Conclusion

With the prevalence of digitization, even the most mundane soft sciences, like social sciences, humanities, or law face problems that hard sciences such as physics had been struggling with for the last several decades. Social science research is increasingly data and technology intensive. Scientific fields, like particle physics had to develop internal capacities to build highly complex instruments such as the Large Hadron Collider, and vast infrastructures that can analyze the huge amounts of data such instruments generate. Though the challenges social scientists face are by orders of magnitude smaller and more manageable than those in physics or genomics, the transformations that social science disciplines must undergo are no less dramatic. The step from doing research in excel to doing advanced data processing in R is huge, and it is not the only one most departments need to take. Social sciences need to build a diverse set of scientific skills and expertise which also include strong management capacities that are able to handle the complex development and operational tasks related to big-data ready research infrastructures. One can have all the data in the world, but without these skills there is no big data research.

Further Reading

- Bodó, B., Helberger, N., Irion, K., Borgesius Zuiderveen, F. J., Moller, J., van der Velde, B., … Vreese, C. H. de. (2017). Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents. Yale Journal of Law & Technology, 19, 133.
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.
- Shah, D. V, Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. The ANNALS of the American Academy of Political and Social Science, 659(1), 6–13. https://doi.org/10.1177/0002716215572084
- Borgesius, F. Z., Gray, J., & Eechoud, M. V. (2015). Open data, privacy, and fair information principles: Towards a balancing framework. Berkeley Tech. LJ, 30, 2073.
- Mitchell, R. (2018). Web Scraping with Python: Collecting More Data from the Modern Web. O'Reilly Media, Inc..

References

- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7), 16-07.

- Berry, D. (2011) 'The Computational Turn: Thinking About the Digital Humanities', Culture Machine. vol 12. [online] Available at:http://www.culturemachine.net/index.php/cm/article/view/440/470 (11 July 2011).
- Bodó, B., Helberger, N., Irion, K., Borgesius Zuiderveen, F. J., Moller, J., van der Velde, B., … Vreese, C. H. de. (2017). Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents. Yale Journal of Law & Technology, 19, 133.
- Bodó, B., Helberger, N., & de Vreese, C. H. (2017). Political micro-targeting: a Manchurian candidate or just a dark horse? *Internet Policy Review*, *6*(4).
- Borgesius, F. Z., Gray, J., & Eechoud, M. V. (2015). Open data, privacy, and fair information principles: Towards a balancing framework. Berkeley Tech. LJ, 30, 2073.
- Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., Vreese, C. H. de, & Helberger, N. (2016). Should we worry about filter bubbles? An interdisciplinary inquiry into self-selected and pre-selected personalised communication. *Internet Policy Review*, *5*(1).
- danah boyd & Kate Crawford (2012) CRITICAL QUESTIONS FOR BIG DATA, Information, Communication & Society, 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878
- Cadwalladr, C., & Graham-Harrison, E. (2018). How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. Retrieved on April, 10.
- Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). MAD skills. Proceedings of the VLDB Endowment, 2(2), 1481–1492. https://doi.org/10.14778/1687553.1687576Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, 154, 72–80. https://doi.org/https://doi.org/10.1016/j.ijpe.2014.04.018
- Jacobs, A. (2009). The pathologies of big data. Communications of the ACM, 52(8), 36. https://doi.org/10.1145/1536616.1536632
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.
- Mayer-Schönberger, V., & Cukier, K. (2014). Learning with big data: The future of education. Houghton Mifflin Harcourt.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.
- Shah, D. V, Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. The ANNALS of the American

Academy of Political and Social Science, 659(1), 6–13. https://doi.org/10.1177/0002716215572084

- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51–59. https://doi.org/10.1089/big.2013.1508
- John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think. Taylor & Francis.
- Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314–347. https://doi.org/https://doi.org/10.1016/j.ins.2014.01.015