Data collection and big data: opportunities and challenges

Bart van der Sloot

Senior Researcher
Tilburg Institute for Law, Technology, and Society (TILT)
Tilburg University, Netherlands

Overview

- (1) What is Big Data?
- (2) How can Big Data be applied for Digital Justice?
- (3) What are the potential risks and benefits?

(1) What is Big Data?

- Key sources:

- Datatilsynet: Big Data: privacy principles under pressure
- Podesta: Big Data: Seizing Opportunities, preserving values
- EDPS: Opinion on coherent enforcement of fundamental rights in the age of big data
- WP29: Statement on Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU
- Berlin Working Group: Privacy principles under pressure in the age of Big Data analytics
- Dutch Scientific Council for Government Policy: Exploring the Boundaries of Big Data

Defintion and delineation of Big Data

- The *Gartner Report* focusses on three matters when describing Big Data: increasing volume (amount of data), velocity (speed of data processing), and variety (range of data types and sources). This is also called the 3v model or 3v theory
- Authors have added new V's such as *Value* (Dijcks, 2012; Dumbill, 2013), *Variability* (Hopkins & Evelson, 2011; Tech America Foundation, 2012), *Veracity* (IBM, 2015) and *Virtual* (Zikopoulos et al 11; Akerkar et al 2015).

Defintion and delineation of Big Data

- The Article 29 Working Party: ‘Big Data is a term which refers to the enormous increase in access to and automated use of information. It refers to the gigantic amounts of digital data controlled by companies, authorities and other large organizations which are subjected to extensive analysis based on the use of algorithms. *Big Data may be used to identify general trends and correlations, but it can also be used such that it affects individuals directly.’*
- The *European Data Protection Supervisor*: ‘Big data means large amounts of different types of data produced at high speed from multiple sources, whose handling and analysis require new and more powerful processors and algorithms. Not all of these data are personal, but many players in the digital economy increasingly rely on the large scale collection of and trade in personal information. As well as benefits, these growing markets pose specific risks to individual's rights to privacy and to data protection.’

Defintion and delineation of Big Data

- The Estonian DPA describes Big Data as ‘collected and processed open datasets, which are defined by quantity, plurality of data formats and data origination and processing speed.’
- The Luxembourg DPA: ‘Big Data stems from the collection of large structured or unstructured datasets, the possible merger of such datasets as well as the analysis of these data through computer algorithms. It usually refers to datasets which cannot be stored, managed and analysed with average technical means due to their size. Personal data can also be a part of Big Data but Big Data usually extends beyond that, containing aggregated and anonymous data.’
- The Dutch DPA: ‘Big Data is all about collecting as much information as possible ; storing it in ever larger databases ; combining data that is collected for different purposes ; and applying algorithms to find correlations and unexpected new information.’
- The Slovenian DPA: ‘Big Data is a broad term for processing of large amounts of different types of data, including personal data, acquired from multiple sources in various formats. Big Data revolves around predictive analytics – acquiring new knowledge from large data sets which requires new and more powerful processing applications.’
- The UK DPA: ‘repurposing data; using algorithms to find correlations in datasets rather than constructing traditional queries; and bringing together data from a variety of sources, including structured and unstructured data.’
- The Swedish DPA argues that ‘the concept is used for situations where large amounts of data are gathered in order to be made available for different purposes, not always precisely determined in advance.’

Defintion and delineation of Big Data

- Umbrella term
- *Open Data:* Lots of Big Data initiatives are linked to Open Data. Open Data is the idea, as the name suggests, that (government) data should be public. Traditionally, it is linked to the strive for transparency in the public sector and for more control over government power by media and/or citizens. In particular, the Estonian DPA is very explicit about the relationship between Open Data and Big Data. Big Data is defined as ‘collected and processed open datasets, which are defined by quantity, plurality of data formats and data origination and processing speed’. The desk research also shows a clear link between the two concepts in some countries, such as Australia, France, Japan and the United Kingdom.

Defintion and delineation of Big Data

- *Re-Use:* Linked to Open Data is the idea of re-use of data. Yet there is one important difference. While Open Data traditionally concerned the transparency of and control on government power, there re-use of (government) data is specifically intended to promote the commercial exploitation of these data by businesses and private parties. The re-use of Public Sector Information is stimulated through the PSI Directive of the European Union. But more in general, re-use refers to the idea that data can be used for another purpose than for which they were originally collected. The Norwegian DPA, inter alia, has suggested the relationship between Big Data and the re-use of data. The Norwegians use the definition of the Working Group 29, 'but also add what in our opinion is the key aspect of Big Data, namely that it is about the compilation of data from several different sources. In other words, it is not just the volume in itself that is of interest, but the fact that secondary value is derived from the data through reuse and analysis.' The desk research also showed a link between the two concepts. In France, for example, Big Data is primarily seen as a phenomenon based on the re-use of data for new purposes and on the combination of different data and datasets. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Directive 2013/37/EU of the European Parliament and the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information.

Defintion and delineation of Big Data

- *Internet of things:* The term the Internet of Things refers to the idea that more and more things are connected to the Internet. This may include cars, lampposts, refrigerators, pants, or whatever object. This allows for the development of smart devices - for example, a refrigerator that records that the milk is out and automatically orders new. By providing all objects with a sensor, large quantities of data can be collected. Therefore, Big Data and the Internet of Things are often mentioned in the same breath. An example would be the DPA of the United Kingdom noting ‘that big data may involve not only data that has been consciously provided by data subjects, but also personal data that has been observed (e.g. from Internet of Things devices), derived from other data or inferred through analytics and profiling.’

Defintion and delineation of Big Data

- *Smart:* Because of the applications of the internet of things and the constantly communicating devices and computers, the development of smart products and services has spiralled. Examples of such developments are smart cities, smart devices and smart robots. The desk research indicates that in a number of countries, a link is made between such developments and Big Data systems, for example the United States and the United Kingdom. Also, the DPA from Luxembourg emphasizes the relationship with smart systems, such as smart metering. ‘At a national level, a system of smart metering for electricity and gas has been launched. The project is however still in a testing phase. - The CNPD has not issued any decisions, reports or opinions that are directly dealing with Big Data. The Commission has however issued an opinion in a related matter, namely with regard to the problematic raised by smart metering. In 2013, the CNPD issued an opinion on smart metering. The main argument of the opinion highlights the necessity to clearly define the purposes of the data processing as well as the retention periods of the data related to smart metering.’

Defintion and delineation of Big Data

- *Profiling:* A term that is often associated with Big Data and is sometimes included as part of the definition of Big Data is profiling. Because increasingly large data sets are collected and analysed, the conclusions and correlations are mostly formulated on a general or group level. This mainly involves statistical correlations, sometimes of a predictive nature. Germany is developing new laws on profiling and a number of DPAs emphasize the relationship of Big Data with profiling, such as the DPA of Netherlands, Slovenia, the UK and Belgium. The latter argues: 'The general data protection law applies, and we expect that de new data protection regulation will be able to provide a partial answer (profiling) to big data issues (legal interpretation of the EU legal framework).'

Defintion and delineation of Big Data

- *Algoritmes:* A term that recurs in very many definitions of Big Data is algorithms. This applies to the definition of Working Party 29, the EDPS and a number of DPAs such as that of Luxembourg, the Netherlands and the UK. A number of countries also have a special focus on algorithms. In Australia, a ‘Program Protocol’ applies to certain cases – a report may be issues in which the following elements are contained: a description of the data, a specification of each matchings algorithm, the expected risks and how they will be addressed, the means for checking the integrity and the security measures used.

Defintion and delineation of Big Data

- *Cloud Computing:* Cloud computing is also often associated with Big Data processes. In particular, in China and Israel, the two terms are often connected to each other. For example, the Chinese vice-premier stressed that the government wants to make better use of technologies like Big Data and cloud computing to support innovation; according to the prime minister mobile Internet, cloud computing, Big Data and the Internet of Things are integrated with production processes, and will thus be an important engine for economic growth. In Israel, the plan is for the army to have a cloud where all data are stored in 2015 - there is even talk of a "combat computing cloud", a data center that will make available different tools to forces on the ground. Also, some DPAs suggest a relationship between cloud computing and Big Data; the Slovenian DPA states, for example, that ‘new concepts and paradigms, such as cloud computing or big data should not lower or undermine the current levels of data protection as a fundamental human right.’

Use in practice of Big Data

- In the United States, more than $ 200 million was reserved for a research and development initiative for Big Data, to be spent by six federal government departments; the army invested the most in Big Data projects, namely $ 250 million; $ 160 million was invested in a smart cities initiative, investing in 25 collaborations focused on data usage.
- In the United Kingdom, £ 159 million was spent on high-quality computer and network infrastructure, there are £ 189 million in investments to support Big Data and to develop the data infrastructure of the UK and £ 10.7 million will be spent on a center for Big Data and space technologies. In addition, £ 42 million will be spent on the Alan Turing Institute for analysis and application of big data, £ 50 million for 'The Digital Catapult', where researchers and industry are brought together to come up with innovative products and lastly, the Minister of Universities and Science in February 2014 announced a new investment of £ 73 million in Big Data. This is used for bioinformatics, open data projects, research and the use of environmental data.
- In South-Africa, the government has invested 2 billion South-African Rand, approximately € 126.8 million, in the Square Kilometre Array (SKA) project. A project which revolves around very large data sets.
- In France, seven research projects related to Big Data were given € 11.5 million.
- In Germany, the Ministry of Education and Research invested € 10 million in Big Data research institutes and € 20 million in Big Data research; this ministry will also invest approximately € 6.4 million in the project Abida, a four-year interdisciplinary research project on the social and economic effects of large data sets.

Use in practice of Big Data

- What are the areas in which Big Data is already used?
  - Internet companies: advertisements
  - Health care sector: total genome analysis
  - Taxs authorities: risk profiles
  - Police: predictive policing
  - Intelligence services: terror prevention

(2) How can Big Data be applied for Digital Justice?

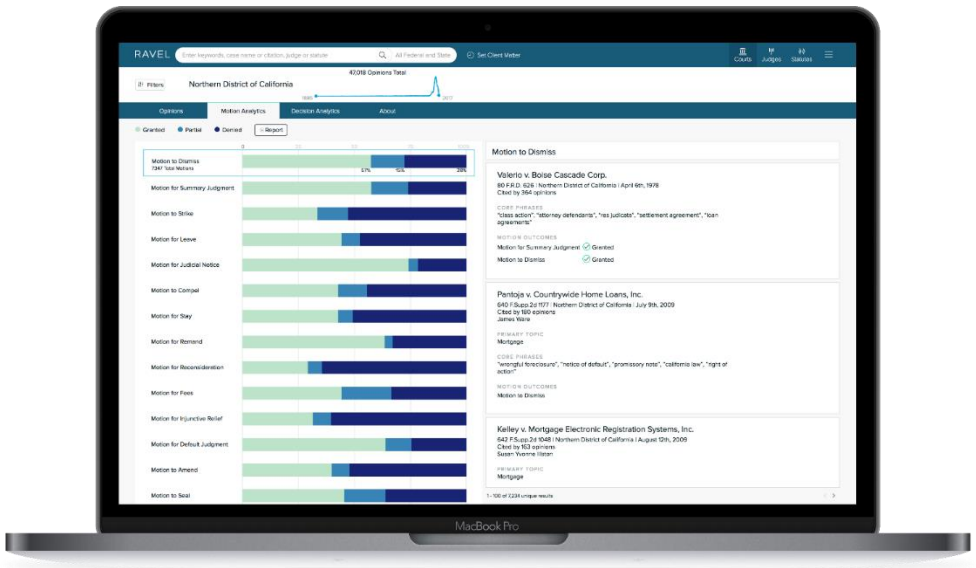
Current applications

- Relatively little experiments untill now

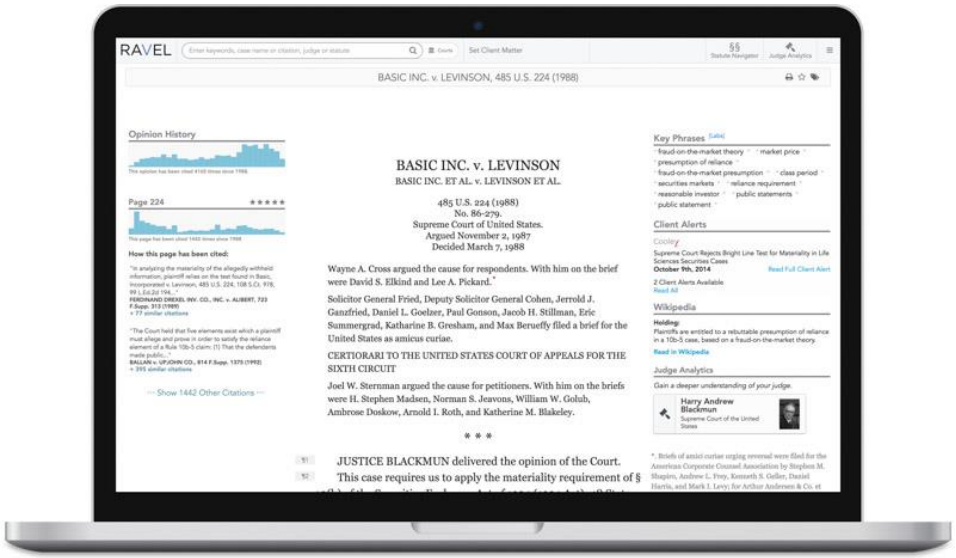
- But there are initiatives, such as:
  - ROSS, which is partly based on IBM’s Watson
  - Ravellaw
  - Various academics and institutions who experiment with it

ROSS

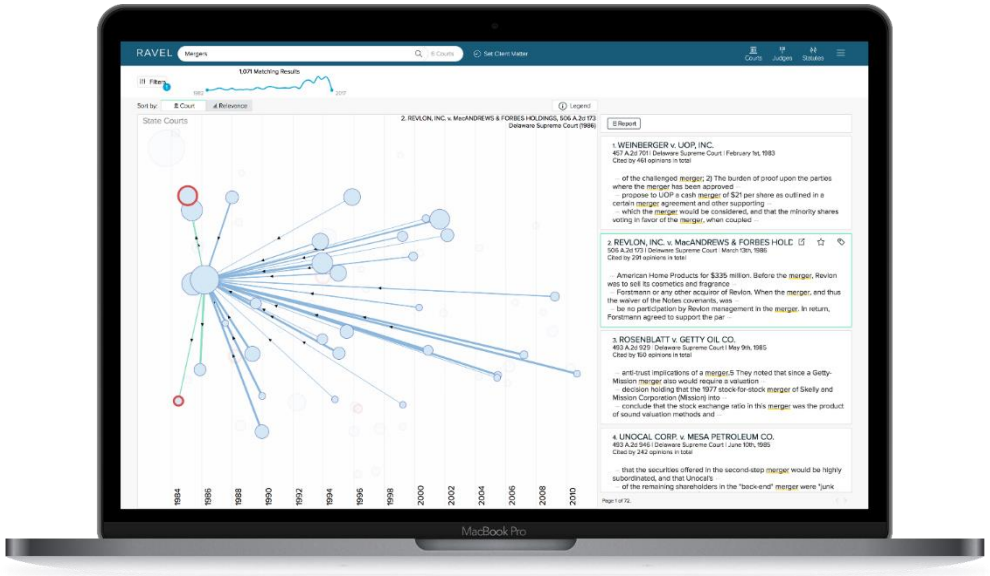
Ravellaw: Court Analytics



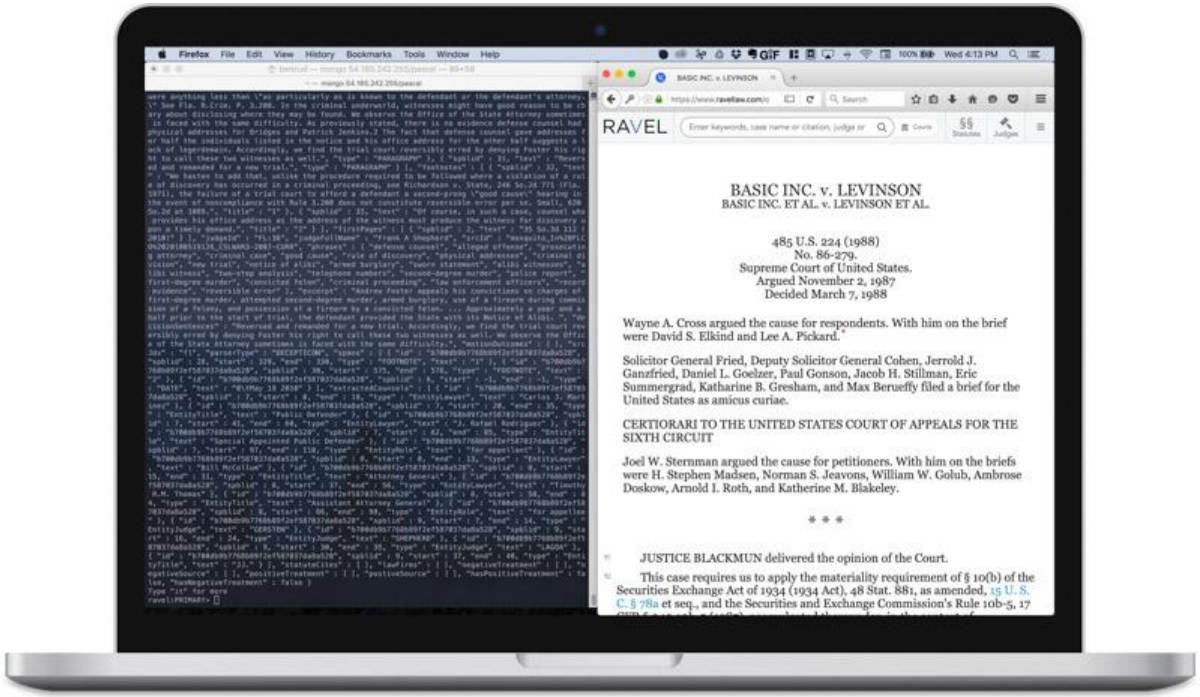
Ravellaw: Case Analytics



Research Tools



Ravellaw: Data Services



- https://vimeo.com/163718987

Law prof claims computer model predicts SCOTUS decisions with 70% accuracy

- Posted Jul 29, 2014 09:30 pm CDT By Debra Cassens Weiss
- Updated: A South Texas College of Law assistant professor who developed a Supreme Court fantasy league says he and two colleagues have developed a computer model that can predict decisions of the court and individual justices.
- Law professor Josh Blackman, writing at his blog, says his computer model, applied to cases since 1953, correctly identifies 69.7 percent of the court's affirmances and reversals, as well as 70.9 percent of the votes of individual justices. A paper at SSRN has details.
- The predictions are based on data that was available before the court's decision. Ninety variables are used, including the party of the appointing president, the court era in which the decision is written, the court and justice's ideological direction, and the agreement level of the court. The model compares predictions for each case to what actually happened, learning which variables work and which don't.
- The computer model categorizes cases using a Supreme Court database at Washington University created with a grant from the National Science Foundation, though Blackman and his colleagues will have to code cases for the upcoming term. Working on the project with Blackman are Michigan State University law professor Daniel Martin Katz and Michael James Bommarito II of Bommarito Consulting.
- "While other models have achieved comparable accuracy rates," Blackman writes, "they were only designed to work at a single point in time with a single set of nine justices. Our model has proven consistently accurate at predicting six decades of behavior of thirty justices appointed by 13 presidents."
- Blackman says he will be hosting a tournament where fantasy SCOTUS players compete against the algorithm. "What IBM's Watson did on Jeopardy, our model aims to do for the Supreme Court," he writes.
- Blackman introduced the project at his blog on April 1. The April Fool's joke, he tells the ABA Journal, was partly inspired by the actual project. Beginning the first Monday in October, Blackman says, a new website will allow users to pull down a pending Supreme Court case for a detailed prediction of what will happen.
- Asked if the database would have real-world applications, Blackman said the model could be used by lawyers weighing whether to settle or litigate a case. "If you have intelligence that's reliable about how the court will decide the case, you can make a more informed litigation decision," Blackman says.

This AI Can Accurately Predict the Outcome of Human Rights Trials

- Jordan Pearson Oct 24 2016, 2:00am
- A team of researchers from University College London (UCL) has devised an algorithm that can predict whether or not a human rights complaint is legitimate, with 79 percent accuracy. This technology, the researchers say, could automate the human rights pipeline by analyzing applications and prioritizing them for the court's human judges.
- "It's important to give priority to cases where there was likely a violation of a person's human rights," said Nikos Aletras, a UCL computer scientist and co-author of a paper describing the work published on Sunday in *PeerJ Computer Science*, in an interview.
- "The court has a huge queue of cases that have not been processed and it's quite easy to say if some of them have a high probability of violation, and others have a low probability of violation," added Vasileios Lampos, Aletras' colleague and also a co-author of the paper. "If a tool could discriminate between the classes and prioritize the cases with a high probability, then those people will get justice sooner."
- The approach used by the team is fairly simple, as far as the quickly advancing field of deep learning goes. They first trained a Natural Language Processing neural network on a database of court decisions, which contains the facts of the case, the circumstances surrounding it, the applicable laws, and details about the applicant such as country of origin. This way, the program "learned" which of these aspects is most likely to correlate with a particular ruling.

(3) What are the potential risks and benefits?

ROSS

ROSS: Impact Identified

Reduction in Research Time from Incorporating Use of ROSS

- **30.3%** over Boolean alone
- **22.3%** over Natural Language alone

Increase in Information Retrieval Quality Compared to Boolean and Natural Language Search

- **42.9%** more relevant authorities retrieved
- **30.3%** more results constituted relevant authorities
- **86.9%** higher Normalized Discounted Cumulative Gain

Estimated Business Impact & ROI

- **$8,466 - $13,067** annual revenue increase per attorney based on a 25% conversion of unbillable time to billable time
- **176.4% to 544.5%** resulting return on investment

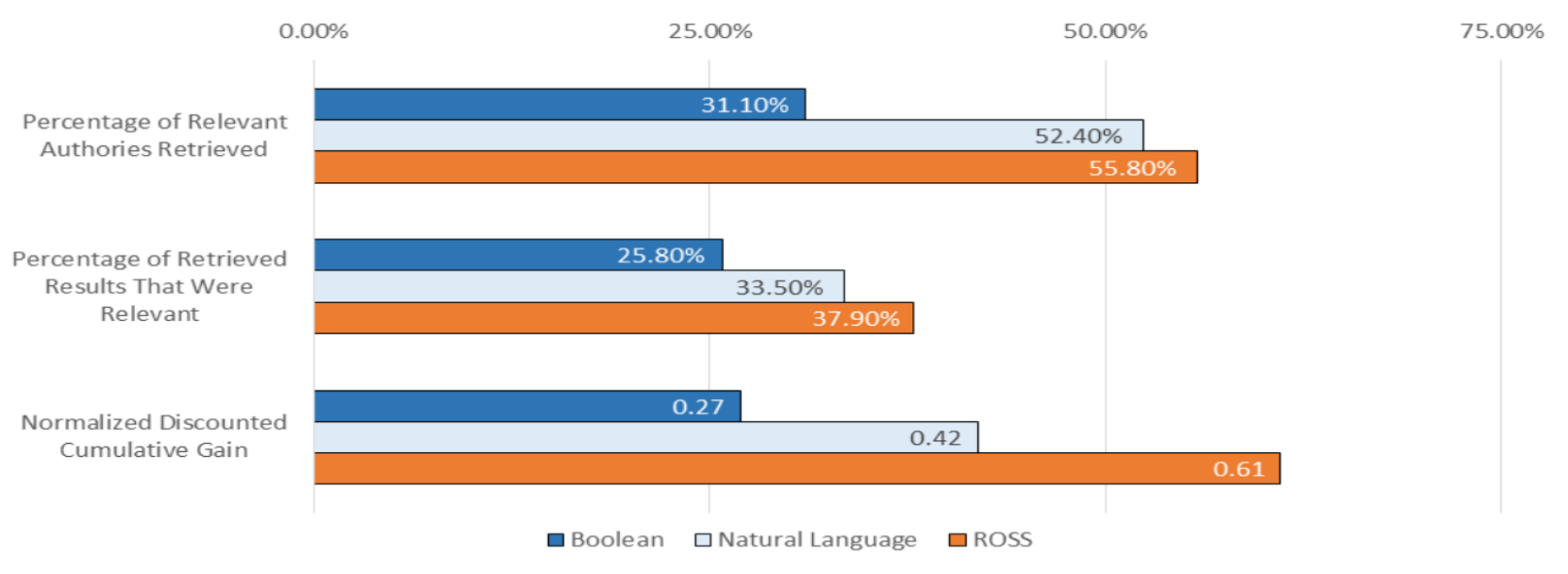
ROSS

Table 1: Key Assessment Factors Benchmarked

Category	Factor	Definition	Measurement
Information Retrieval Quality	Thoroughness	Portion of the total pool of existing relevant authorities that were retrieved	Percentage of the total set of relevant authorities that were retrieved
	Accuracy	Portion of total results retrieved that included relevant authorities	Percentage of the total results retrieved that represent relevant authorities
	Ranking Effectiveness	Relative placement of relevant authorities within the list of top results retrieved	Normalized Discounted Cumulative Gain
User Satisfaction	Ease of Use	Participant's satisfaction with the ease of use of the research approach.	Self-reported Likert Scale responses to standardized satisfaction questions
	Confidence	Participant's perceived confidence that he or she obtained a complete answer	Self-reported Likert Scale responses to standardized satisfaction questions
Research Efficiency	Time to Complete	Amount of time required for a participant to obtain a satisfactory answer the question.	Total time the researcher spent using the research approach to obtain an answer

ROSS

Figure 1: Information Retrieval Effectiveness of Legal Search Tools Based on Observed Queries



ROSS

Table 2: Average Agreement with Statements Describing User Experience with the Toolsets Employed

		Boolean	Natural Language	ROSS & Boolean	ROSS & Natural Language
Usability	I found the tool's user interface to be intuitive and easy to use.	3.3	4.0	5.0	5.0
	The search results returned by the tool were concise and primarily contained cases that were relevant to my legal questions.	4.0	3.5	5.0	5.0
	The search results returned by the tool did not include a large number of results that were NOT RELEVANT to my legal questions.	2.0	2.7	4.7	4.7
Confidence	It was easy to find all of the cases required to give a complete answer to my legal questions using the tool.	3.5	3.3	4.8	5.0
	I am confident that the tool returned all of the cases required to give a complete answer to my legal questions.	3.3	3.5	4.8	5.0
Scale – 1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly Agree					

Karin Aarde & Corien Prins

- The use of Big Data in the judiciary can have an affect on all phases in the procedure:
  - Submitting a claim
  - Access to files and documents
  - The hearing
  - The decision
  - Archiving

Dangers

Social and ethical dangers of Big Data

- **Power imbalance & Mathew effect:** Individuals, as a general rule, have limited power to influence how large corporations behave. Extensive use of Big Data analytics may increase the imbalance between large corporations on the one hand and the consumers on the other. It is the companies that collect personal data that extract the ever-growing value inherent in the analysis and processing of such information, and not the individuals who submit the information. Rather, the transaction may be to the consumer's disadvantage in the sense that it can ex- pose them to potential future vulnerabilities (for example, with regard to employment opportunities, bank loans, or health insurance options).
- **Data determinism and discrimination:** The “Big data-mindset” is based on the assumption that the more data you collect and have access to, the better, more reasoned and accurate decisions you will be able to make. But collection of more data may not necessarily entail more knowledge. More data may also result in more confusion and more false positives. Extensive use of automated decisions and prediction analyses may have adverse consequences for individuals. Algorithms are not neutral, but reflect choices, among others, about data, connections, inferences, interpretations, and thresholds for inclusion that advances a specific purpose. 32 Big Data may hence consolidate existing prejudices and stereotyping, as well as reinforce social exclusion and stratification. Use of correlation analysis may also yield completely incorrect results for individuals. Correlation is often mistaken for causality. If the analyses show that individuals who like X have an eighty per cent probability rating of being exposed to Y, it is impossible to conclude that this will occur in 100 per cent of the cases. Thus, discrimination on the basis of statistical analysis may become a privacy issue. A development where more and more decisions in society are based on use of algorithms may result in a ”Dictatorship of Data”, where we are no longer judged on the basis of our actual actions, but on the basis of what the data indicate will be our probable actions.

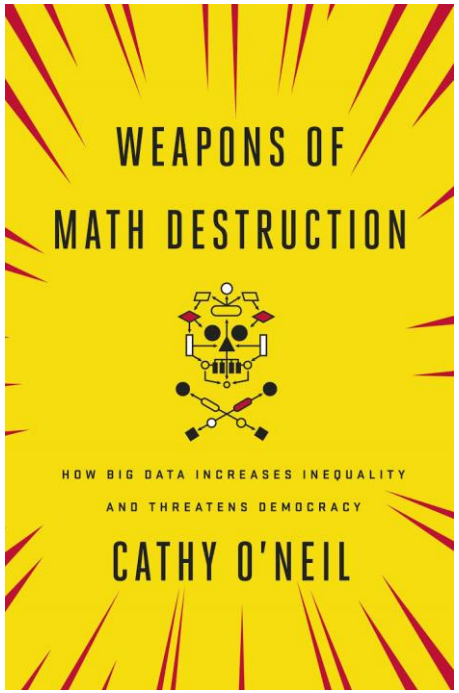
Social and ethical dangers of Big Data

- **The Chilling effect:** If there is a development where credit scores and insurance premiums are based solely or primarily on the information we leave behind in various contexts on the Internet and in other arenas in our daily life, this may be of consequence for the protection of privacy and how we behave. In ten years, our children may not be able to obtain insurance coverage because we disclosed in a social network that we are predisposed for a genetic disorder, for example. This may result in us exercising restraint when we participate in society at large, or that we actively adapt our behaviour – both online and elsewhere. We may fear that the tracks we leave behind in various contexts may have an impact on future decisions, such as the possibility of finding work, obtaining loans, insurance, etc. It may even deter users from seeking out alternative points of view online for fear of being identified, profiled or discovered. With regard to the authorities' use of Big Data, uncertainty concerning which data sources are used for collecting information and how they are utilised may threaten our confidence in the authorities. This in turn may have a negative impact on the very foundation for an open and healthy democracy. Poor protection of our privacy may weaken democracy as citizens limit their participation in open exchanges of viewpoints. In a worst case scenario, extensive use of Big Data may have a chilling effect on freedom of expression if the premises for such use are not revealed and cannot be independently verified.
- **Echo chambers:** Personalisation of the web, with customised media and news services based on the individual's web behaviour, will also have an impact on the framework conditions for public debates and exchanges of ideas – important premises for a healthy democracy. This is not primarily a privacy challenge, but constitutes a challenge for society at large. The danger associated with so-called ”echo chambers” or ”filter bubbles” is that the population will only be exposed to content which confirms their own attitudes and values. The exchange of ideas and viewpoints may be curbed when individuals are more rarely exposed to viewpoints different from their own.
- **Transparency paradox:** The citizen is becoming more and more transparent to the government, while the government is becoming more an more in-transparent to the citizen.

Principles that might be undermined

- Equality of arms
- Presumption of innocence
- Legibility of judgements
- Non-discrimination
- Fairness

Big Data: Usefull tools or Weapons of Math Destruction?



Questions:

- (1) How can Big Data be used in the field of Digital Justice to improve accessibility and transparancy for citizens?

- (2) How can Big Data be used in the field of Digital Justice to improve the efficiency and effectiveness of the procedure and organisation?

- (3) How can Big Data be used in the field of Digital Justice to improve court judgements in terms of quality and fairness?