Big Data, privacy and ethics: current trends and future challenges

Sébastien Gambs
Université du Québec à Montréal (UQAM)

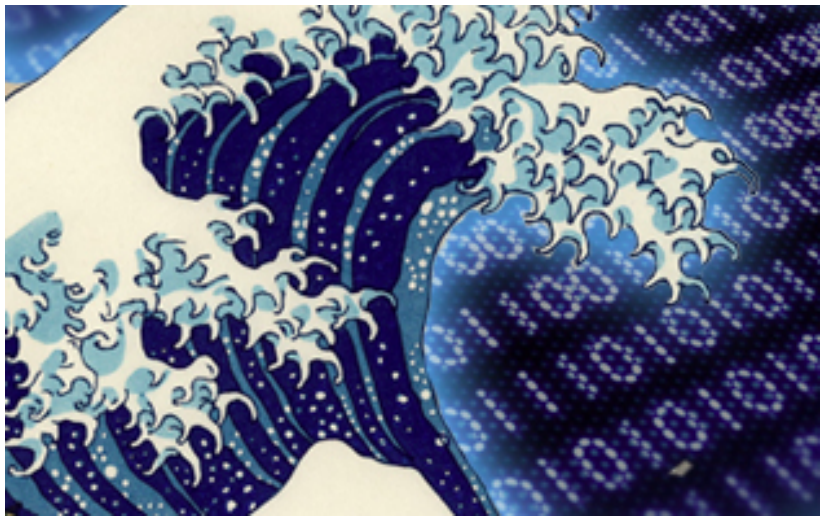
gambs.sebastien@uqam.ca

24 April 2017

Introduction

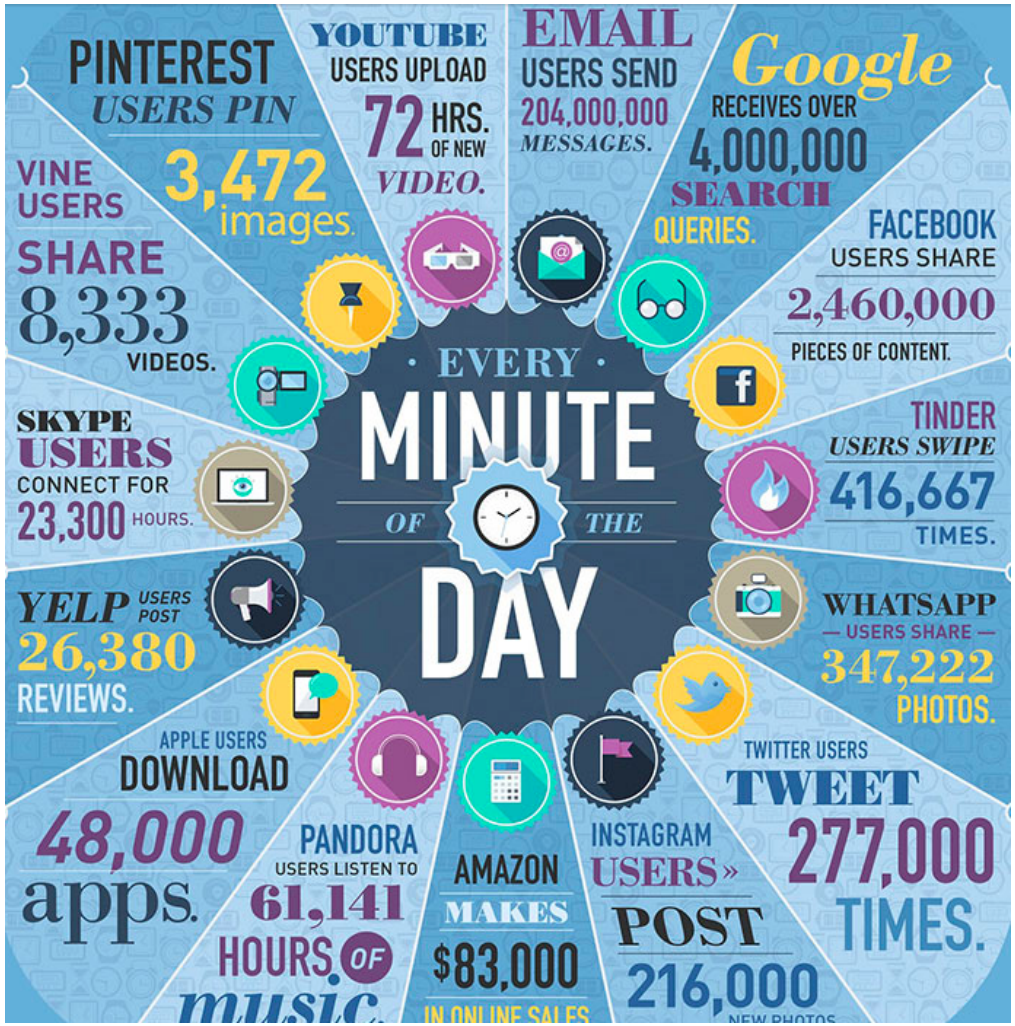
Big Data

- Broadly refers to the massive increase of the amount and diversity of data collected and available.

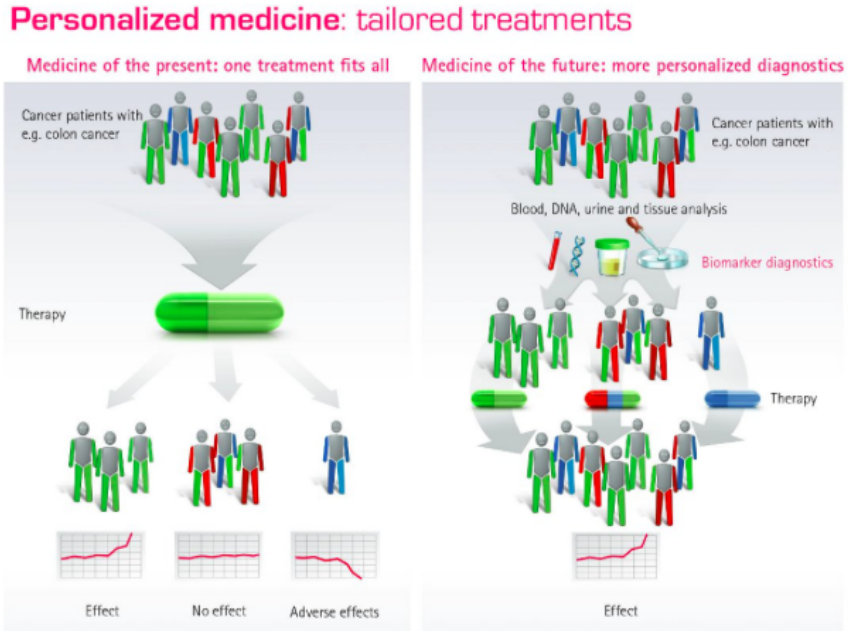


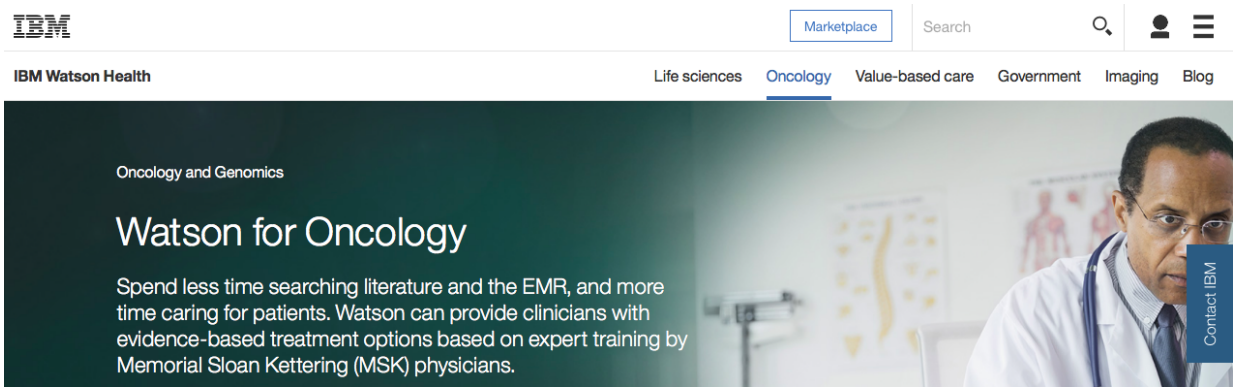
- Technical characterization: often define in terms of the five "V"' (*Volume*, *Variety*, *Velocity*, *Variability* and *Veracity*).
- Main promise of Big Data: offer the possibility to realize inferences with an unprecedented level of accuracy and details.

A glimpse at Big personal Data



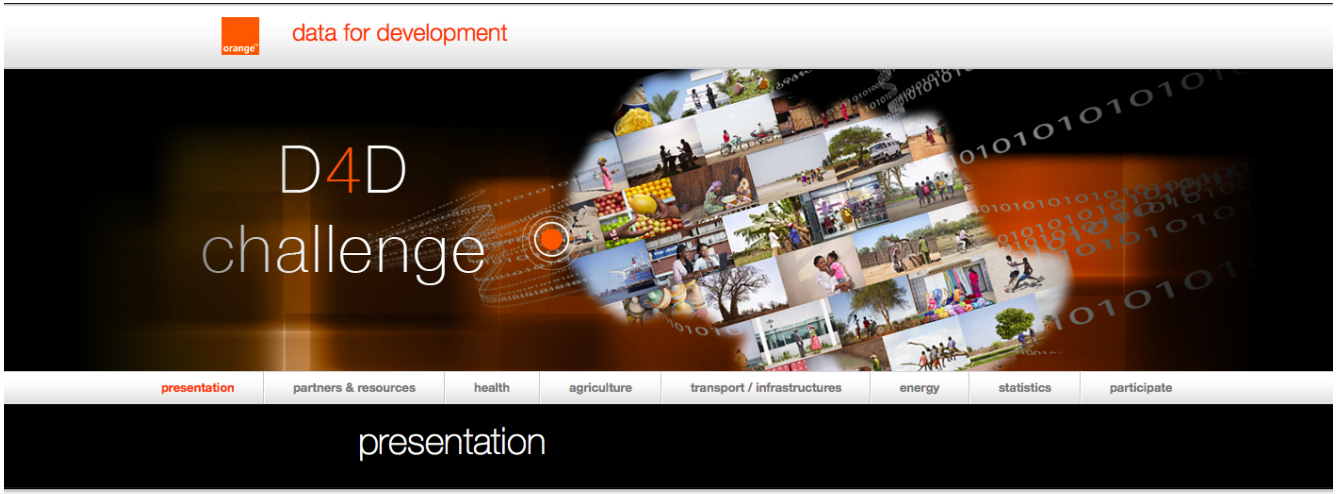
Personalized medicine - IBM Watson advisor for cancer





Large-scale mobility analytics

Objective: publication of the mobility traces of users issued from phone usage (Call Details Records).

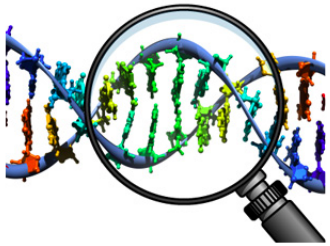


One year after the 'Data For Development Côte d'Ivoire Challenge', Sonatel and the Orange Group are now launching a 'D4D Challenge' in Senegal. For this purpose, they are making available to the scientific community, in the exclusive context of the D4D Competition, a series of statistical databases and samples extracted from the mobile network management signals.

Fundamental question: how to anonymize the data before publishing it to limit the privacy risks for the users whose mobility is recorded in the data?

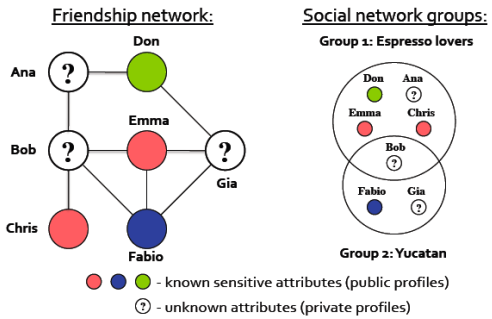
Other types of data with strong inference potential

1. Genomic/medical data.



- Possible risks : inference on genetic diseases or tendency to develop particular health problems, leakage of information about ethnic origin and genomics of relatives, genetic discrimination, . . .

2. Social data.



- Possible risks : reconstruction of the social graph, inferences on political opinions, religion, sexual orientations, hobbies, . . .

Factor 1: augmentation of the easiness of recording our life

- Recent technological developments increase the capacity to record the real and the virtual world.
- Examples :

Factor 2: open data movement

- Consequence : release of important amount of data.
- Originally, this data was mainly public information but...
- there is more and more pressure for institutions to open dataset composed of personal information. Example :

The National Pupil Database is not open data

Published on **10 July 2012**
Countries: **United Kingdom**

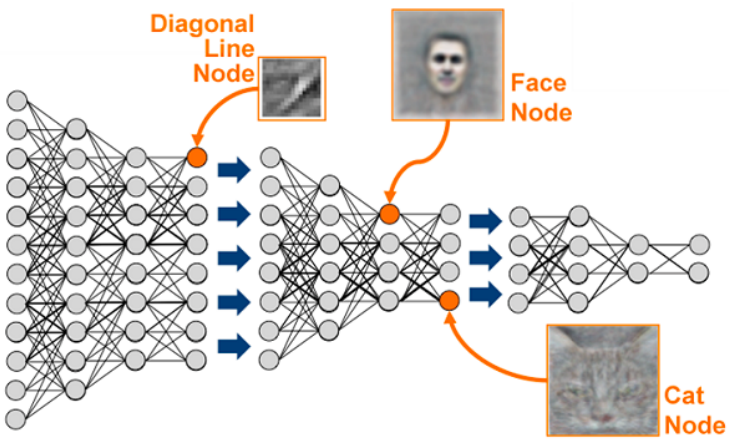
Share

This weekend, the Department for Education sponsored an "appathon", allowing attendees access to the National Pupil Database (which holds information like exam results, special education needs, truancy records and eligibility for free school meals on every child at every state school in the country) and inviting people to build "apps".

The database contains over 400 variables and the records of around 600,000 children. With so many variables, it is a relatively simple task to identify individual children who in any way stand out from the crowd, e.g. those who've performed unusually well in rare subjects. The kind of information the database holds is extremely sensitive and children may have gone out of their way to conceal it from their classmates. Make no mistake - this is intensely personal stuff, not "open data", and any suggestion otherwise betrays a fundamental misunderstanding of both categories. Accordingly, additional safeguards of process and content must be applied.

Factor 3: deep learning revolution

- Revolution: "quantum leap" in the prediction accuracy in many domains.
- Recent success: automatic generation of textual description from a picture, victory of AlphaGo against a professional go player in 2016.



- Possible due to algorithmic advances in machine learning through the Deep Learning approach combined with the increase in computational power and the amount of data available.

Privacy

Privacy

- Privacy is one of the fundamental right of individuals:
  - Universal Declaration of the Human Rights at the assembly of the United Nations (Article 12), 1948.
  - European General Data Protection Regulation (GDPR), voted in 2016 will become effective in 2018.
- One of the main challenge of the "Information Society".

- Risk: collect and use of digital traces and personal data for fraudulent purposes.
- Examples: targeted spam, identity theft, profiling, (unfair) discrimination (to be discussed later).

Impact of Big Data on privacy

1. Magnification of the privacy risks due to the increase in volume and diversity of the personal data collected and the computational power to process them.
2. Often data collected about individuals are "re-used" for a different purpose without asking their consent.
3. The inferences that are possible with Big Data are much more fine-grained and precise than before.
4. Massive release of data without taking into account the privacy aspect $\Rightarrow$ major privacy breach
   - Once a data is disclosed, it is there forever.
5. Ethics of inference : what are the inferences that are acceptable for the society and which ones are not?

Example of sensitive inference: predictive policing

THE VERGE TWEET SHARE

Struggling to reduce its high murder rate, the city of Chicago has become an incubator for experimental policing techniques. Community policing, stop and frisk, "interruption" tactics — the city has tried many strategies. Perhaps most controversial and promising has been the city's futuristic "heat list" — an algorithm-generated list identifying people most likely to be involved in a shooting.

The hope was that the list would allow police to provide social services to people in danger, while also preventing likely shooters from picking up a gun. But a new report from the RAND Corporation shows nothing of the sort has happened. Instead, it indicates that the list is, at best, not even as effective as a most wanted list. At worst, it unnecessarily targets people for police attention, creating a new form of profiling.

IT UNNECESSARILY TARGETS PEOPLE FOR POLICE ATTENTION

Funded through a $2 million grant from the National Institute of Justice, the list's algorithm identifies people by looking not only at arrests, but also whether someone is socially connected with a known shooter or shooting victim. The program also has a kind of pre-crime feature in which police visit people on the list before any crime has been committed.

Privacy enhancing technologies

Privacy Enhancing Technologies (*PETs*) : ensemble of techniques for protecting the privacy of an individual and offer him a better control on his personal data.
Example of PET : homomorphic encryption (see Caroline Fontaine talk's tomorrow).

Two fundamental principles behind the PETs :

- Data minimization : only the information necessary for completing a particular purpose should be collected/revealed.
- Data sovereignty : enable a user to keep the control on his personal data and how they are collected and disseminated.

Personally identifiable information

- Personally identifiable information : ensemble of information that can be used to uniquely identified an individual.
- Examples : first and last name, social security number, place and date of birth, physical and email address, phone number, credit card number, biometric data (such as fingerprint and DNA), ...
- Sensitive because they identify uniquely an individual and can be used to easily cross-referenced databases.
- Main limits of the definition :
  - does not take into account some attributes or patterns in the data that can seem innocuous individually but can identified an individual when combined together (quasi-identifiers).
  - does not take into account the inference potential of the data considered.

Pseudonymization is not an alternative to anonymization

Replacing the name of a person by a pseudonym $\not\Rightarrow$ preservation of the privacy of this individual

A Face Is Exposed for AOL Searcher No. 4417749

The New York Times — August 8, 2006

What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com
4417749	jcpenny	2006-04-28	14:16:05		
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenney.com
4417749	tortus and turtles	2006-04-29	13:12:47		
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26		
4417749	fingers going numb	2006-05-02	17:35:47		
4417749	dances by laura	2006-05-02	17:59:32		
4417749	dances by lori	2006-05-02	17:59:57		
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13		
4417749	single dances in atlanta	2006-05-02	18:01:50		
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com
4417749	tyroid	2006-05-06	16:55:34		
4417749	thyroid	2006-05-06	16:55:44		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:14:52		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:17		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:43		

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going though a divorce. I thought there was a place called 'Dances by Lori,' for singles."

"I wanted to find out what my house was worth."

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

(Extract from an article from the New York Times, 6 August 2006)

SUICA's privacy leak (July 2013)

SUICA scandal and the privacy in the era of Big Data

Recently, East Japan Railways Company (JR), oneof the biggest companies in Japan apologized for the sales of the data of its consumers.

JR manages a system of electronic money called SUICA, which started from an electronic ticket for JR's trains but now is one of the strongest electronic money widely used all over Japan.

Recently, JR announced that JR and its partner company would sell the marketing report based on the data of SUICA users on (1) which station they get on and get off, (2) when they used the train, (3) how old they are, and (4) their gender. For this purpose, JR gives the anonymized data of SUICA users to its partner company. At first, JR got no consent from the users. There was no opt-in nor opt-out for the sale of their data. When the announcement was made, there are many people opposing the usage, claiming that the sales of user data are an invasion of users' privacy.

From the viewpoint of the Act on the Protection of Personal Information, which aims to protect personal information, JR may make an argument that it is not violating the Act. The act generally prohibits the transfer of personal information without obtaining the consent. But it is understood that the personal information defined does not include the anonymized information.

Legal requirements to evaluate anonymization methods

General Data Protection Regulation (Article 16):

> *"To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."*

- Consequence : evaluation of risk of de-anonymization should take into account the ressources needed to conduct the re-identification and should be done on a regular basis (*risk-based approach*).
- The French law "for a Digital Republic" (October 2016) also recognized the right for the French data protection authority (the CNIL) to certify anonymization processes.

Inference attack

- Inference attack : the adversary takes as input a published dataset (and possibly some background knowledge) and tries to infer some personal information regarding individuals contained in the dataset.

- Main challenge : to be able to give some privacy guarantees even against an adversary having some auxiliary knowledge.
- We may not even be able to model this *a priori* knowledge.
- Remark: maybe my data is private today but it may not be so in the future due to the public release of some other data.

Example : inference attacks on location data

Joint work with Marc-Olivier Killijian (LAAS-CNRS) and Miguel Núñez del Prado (Universidad del Pacifico).

Main objective : quantify the privacy risks of location data.

Types of attacks:

1. Identification of important places, called *Point of Interests* (POI), characterizing the interests of an individual.
   - Example: home, place of work, gymnasium, political headquarters, medical center, ...
2. Prediction of the movement patterns of an individual, such as his past, present and future locations.
3. Linking the records of the same individual contained in the same dataset or in different datasets (either anonymized or under different pseudonyms).

De-anonymization attack

- De-anonymization attack : the adversary takes as input a sanitized dataset and some background knowledge and tries to infer the identities of the individuals contained in the dataset.
- Specific form of inference attack.
- Example : Sweeney's original de-anonymization attack via linking.

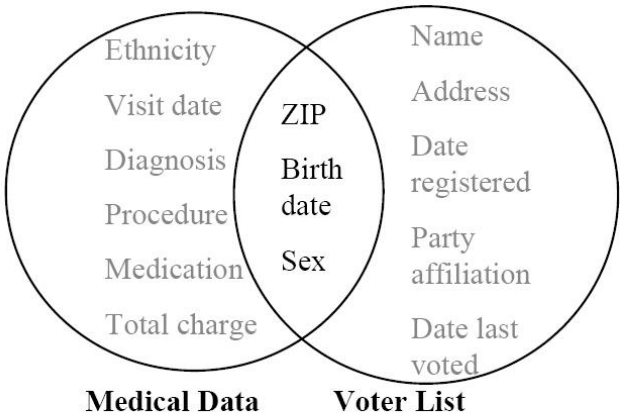
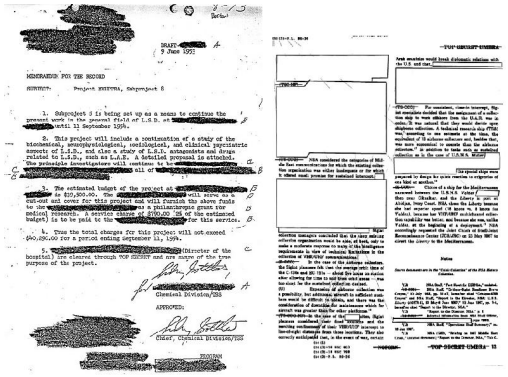


Figure 1 Linking to re-identify data

- The *re-identification risk* measures the success probability of this attack.

Sanitization

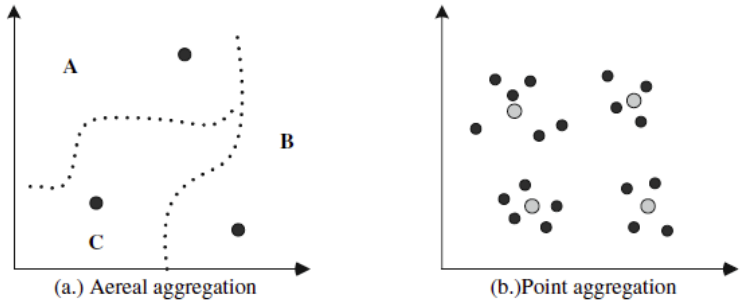
Sanitization : *process increasing the uncertainty in the data in order to preserve privacy*.
$\Rightarrow$ Inherent trade-off between the desired level of privacy and the utility of the sanitized data.
Typical application : public release of data (offline or online context).



Examples drawn from the "sanitization" entry on Wikipedia

Classical sanitization mechanisms

- Perturbation : addition of noise to the true value.
- Aggregation : merge several data into a single one.
- Generalization : loss of granularity of information.



- Deletion : erasure of the information related to a particular attribute.
- Remark : the absence of information can sometimes lead to a privacy breach (*e.g.* : removing the information on the disease of a patient record only if he has a sexual disease).
- Introduction of fake data : addition of artificial records in a database to "hide" the true data.

Fundamental ingredients for sanitization

1. Privacy model : what does it mean for released data to be respectful of privacy?
2. Sanitization algorithm : how to modify the data to reach the property defined by the privacy model?
3. Utility measure : how to quantify the utility of the resulting data?

k-anonymity (Sweeney 02)

- Privacy guarantee : in each group of the sanitized dataset, each invidividual will be identical to a least $k-1$ others.
- Reach by a combination of generalization and suppression.
- Example of use : sanitization of medical data.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

- Main challenge : extracting useful knowledge while preserving the confidentiality of individual sensitive data.

Intersection attack

- Question : suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

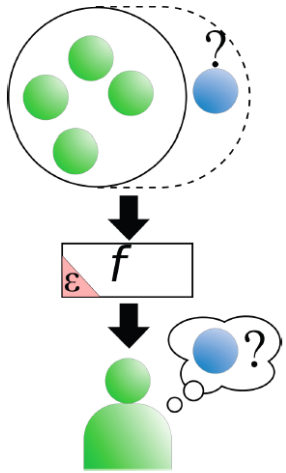
(b)

The key property: composition

- A "good" privacy model should provide some guarantees about the total leak of information revealed by two (or more) sanitized datasets.
- More precisely, if the first release reveals b_1 of information and the second release b_2 bits of information, the total amount of information leaked should not be more than $O(b_1 + b_2)$ bits.
- Remark : most of the existing privacy models do not have any composition property with the exception of differential privacy (Dwork 06).

Differential privacy: principle (Dwork 06)

- Privacy notion developed within the community of private data analysis that has gained a widespread adoption.

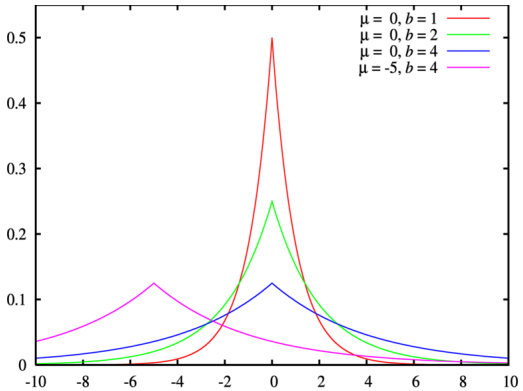


- Basically ensures that whether or not an item is in the profile of an individual does not influence too much the output.
- Give strong privacy guarantees that hold independently of the auxiliary knowledge of the adversary and compose well.

Implementing differential privacy

Possible techniques to implement differential privacy :

- Addition of noise to the output of an algorithm (ex: Laplacian mechanism).



- Perturbation of the input given to the algorithm.
- Randomization of the behaviour of the algorithm.
- Creation of a synthetic database or a data structure "summarizing and aggregating the data".
- Sampling mechanisms.

Fire and Ice Japanese competition on anonymization and re-identification attacks (2015, 2016 and 2017)

- Objective : evaluate empirically the efficiency of anonymization methods and re-identification attacks.

- Similar in spirit to other competitions in machine learning or security.
- See talks of Hiroaki Kikuchi (Meiji University) and Hiroshi Nakagawa (University of Tokyo) in privacy WG session for more details.

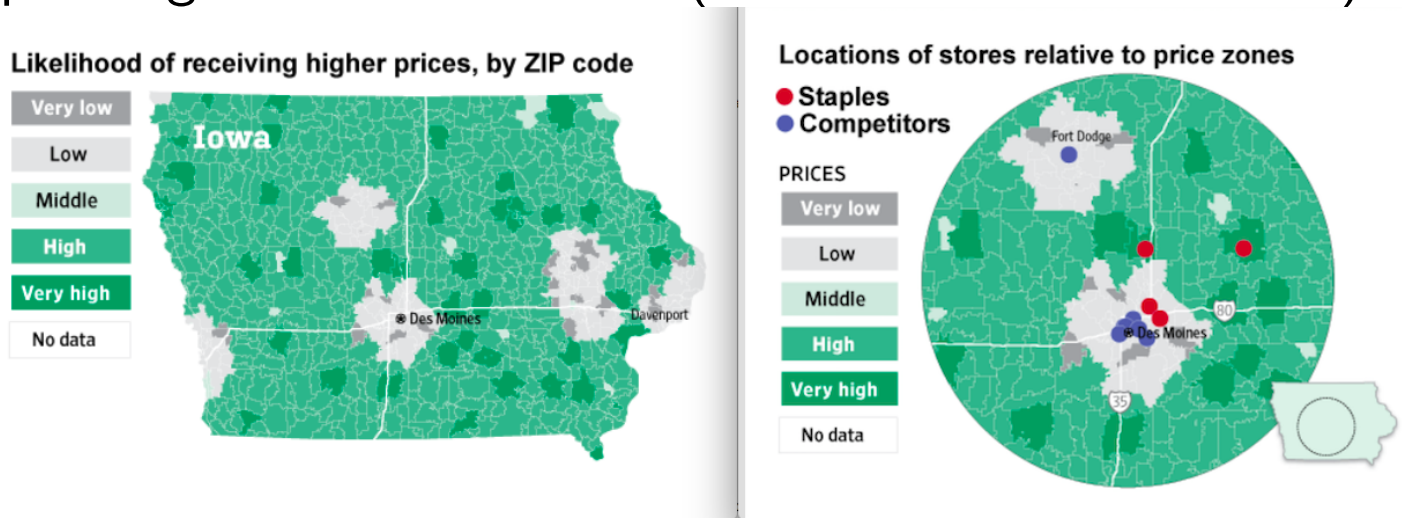
Next steps

- To broaden the impact and the outreach to the privacy community, we have submitted a proposal (accepted) to held an international competition on sanitization mechanisms and inference attacks in the annual Privacy Enhancing Technologies Symposium (PETS).
- Schedule :
  - This year: workshop at PETS for preparing the competition (definition of privacy and utility metrics, choice of the dataset, setting of the competition, ...).
  - Next year: international competition + workshop at PETS to report on the outcomes and the best algorithms for sanitization and inference.
- Parallel event : Shonan meeting on "Anonymization methods and inference attacks: theory and practice" (March 2018).

Transparency, accountability and fairness

The fuzzy border between personalization and discrimination

- Example : price personalization of the Staples website depending of the localization (Wall Street Journal 2012).



- Possible discriminations : poor credit, high insurance rate, refusal to employment or access to schools, denial to some function.

Right to fairness and transparency (GDPR, Article 70)

In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, [...] and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

Possible origin of the bias

1. Problem in the data collection due to some error or the fact that the data is inherently biased.
   - Examples : mistake in the profile of the user, dataset reflects discriminatory decision against a particular population.
2. Inaccuracy due to the learning algorithm.
   - Example : the algorithm is very accurate, except for 1% of the individuals.

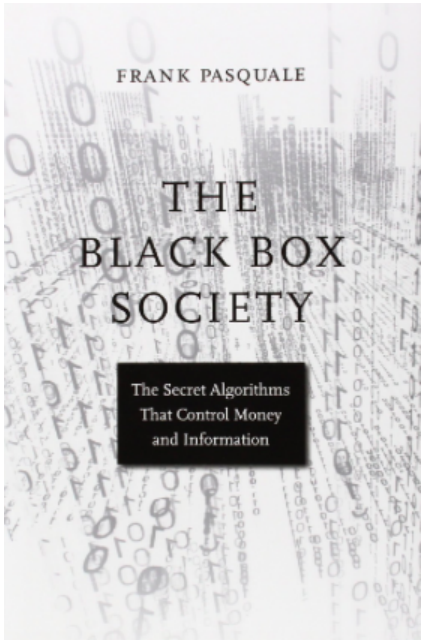
"I'd like to meet the algorithm that thought we'd be a good match."

Opacity of machine learning algorithms

- Machine learning has a central role in most of the personalized systems.
- Opacity: difficulty of understanding and explaining their decision due to their complex design.
- Example: the classifier outputted by a deep learning algorithm is typically composed of many layers of neural networks.
- Risk of "algorithmic dictatorship" (Rouvroy): loss of control of individuals on their digital lives due to automated decision if there is no remediation procedure.

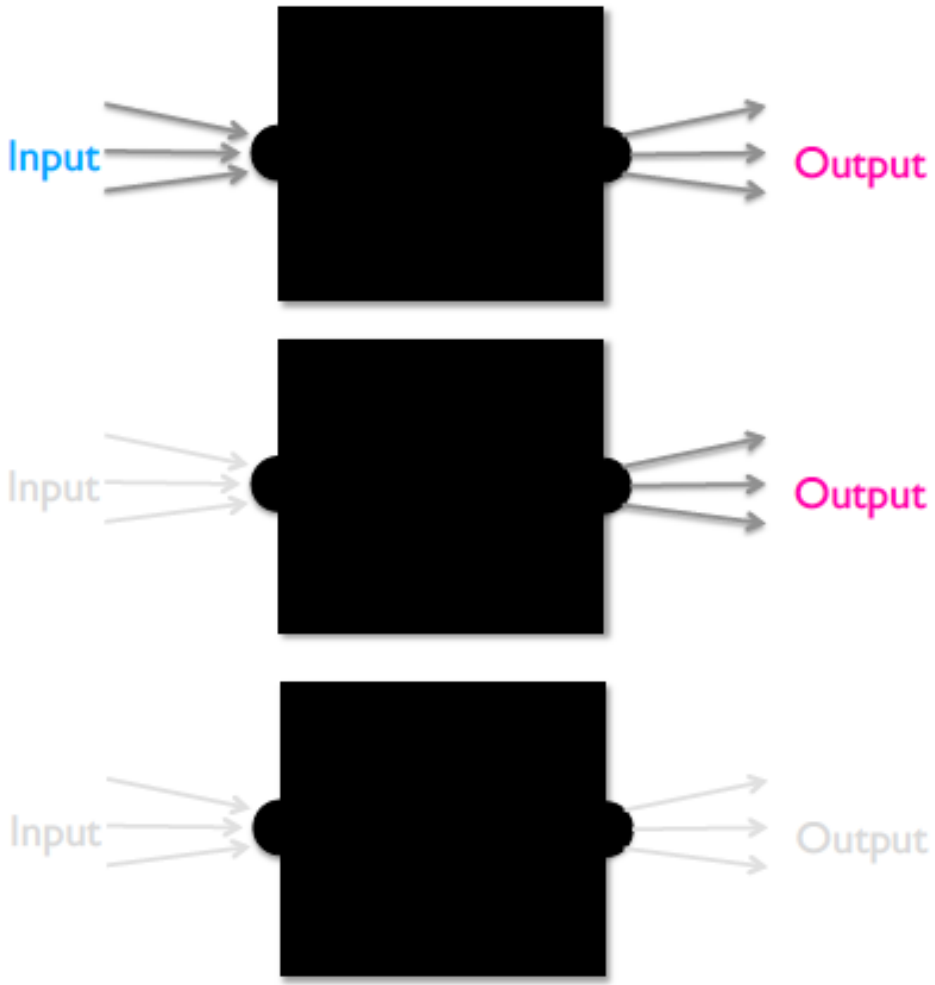
Transparency as a first step

- Asymmetry of information: strong difference between what the system knows about a person and what the person knows about the system.



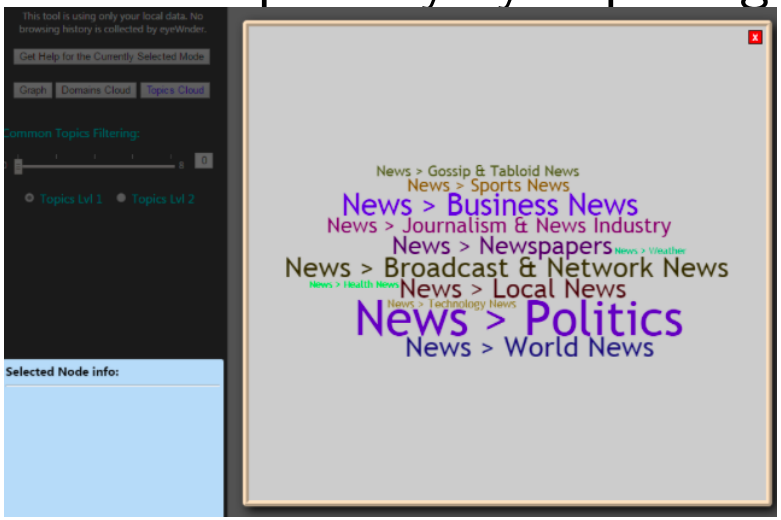
- Lack of transparency leads to lack of trust.
- Strong need to improve the transparency of information systems.

Possible cases for analyzing the black-box (Diakopoulos 16)



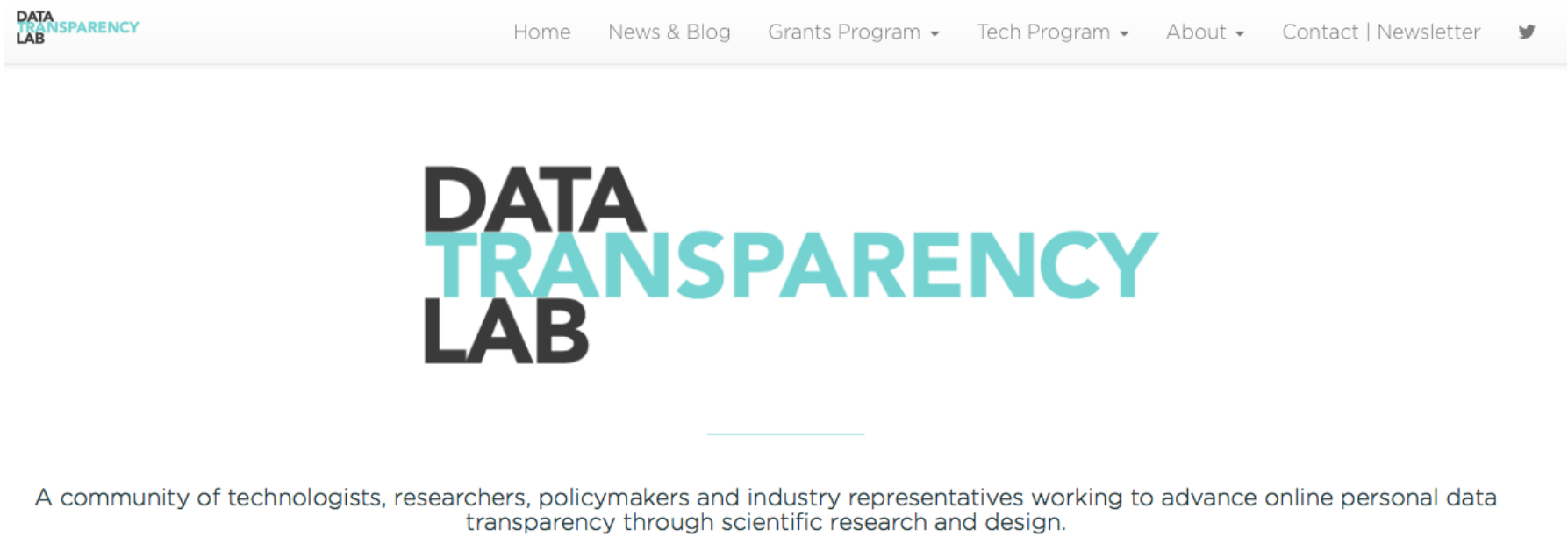
Possible approaches to transparency

1. Regulatory approaches to force companies to let users examine and correct the information collected about them.
2. Methods to increase transparency by "opening the black-box".



- Tools to reach transparency by design.
- Examples: publication of the source code, use of an interpretable model in machine learning.

Example of community effort to increase transparency



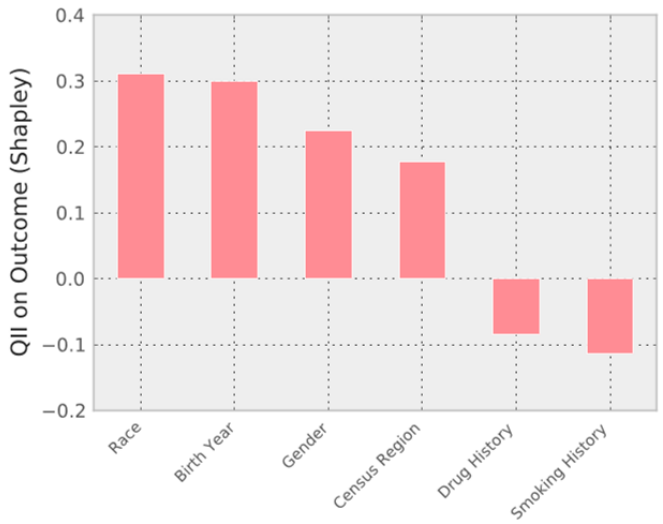
Towards algorithmic accountability

- Caveat: transparency does not necessarily means interpretability or accountability.
- Example: the code of an application could be public but too complex to be comprehend by a human.
- Strong need for the development of tools that can analyze and certify the code of the program.
- Objective: verify that the execution of the program match the intended behaviour or the ethical values that are expected from it.
- Strong link with the notion of *loyalty* (does the system behave as it promises).

Measuring discrimination

- Example: measurement of quantitative input influence.

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

- Challenge: possibility of *indirect discrimination* in which the discriminatory attribute is inferred through other attributes.
- Example: even if the ethnicity is not asked from the user, in some countries it strongly correlates with the ZIP code.

Defining discrimination and fairness

- Disparate impact: criterion in the US law to measure inequality of treatment.
- Inequal treatment occurs if (% of the minority group hired)/(% of the majority group hired) > 0.8
- Group fairness: the statistics of the decisions targeting a particular group are approximately the same than the overall population.
- Individual fairness: two individuals whose profiles are similar (with the exception of the protected attributes) should receive a similar outcome.
- Difficulty: some studies have shown than some of these metrics are incomparable.

Enhancing fairness

- Ultimate objective: being able to increase fairness while not impact too much accuracy
- Examples of possible approaches:
  - Sample the input data to remove its original bias,
  - Change the design of the algorithm so that it becomes discrimination-aware by design,
  - Adapt the output produced by the algorithm (*e.g.*, the classifier) to reduce discrimination.
- Active subject of research but still in its infancy, much remains to be done.

Conclusion

Conclusion (1/2)

- Observation 1 : the capacity to record and store personal data as increased rapidly these last years.
- Observation 2 : "Big Data" will result in more and more being available $\Rightarrow$ increase of inference possibilities.
- Observation 3 : the "Open data" movement will lead to the release of a huge amount of dataset $\Rightarrow$ worsen the privacy impact of Big Data (observation 2).
- The advent of Big Data magnifies the privacy risks that were already existing but also raises new ethical issues.
- Main challenge : balance the social and economical benefits of Big Data with the protection of privacy and fundamental rights of individuals.

Conclusion (2/2)

- Strong need for research and scientific cooperation in Big Data :
  - for determining how to address privacy in this context,
  - for the design of new protection and sanitization mechanisms
  - as well as for inference attacks for assessing the privacy level they provide.
  - for finding solutions for addressing the transparency, fairness and accountability issues
- Overall objective: being able to reap the benefits of Big Data by not only to protecting the privacy of individuals but also making sure that they remain in control of their digital lives.

This is the end

Thanks for your attention
Questions?

Right to object to automated decision (GDPR, Article 70)

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes "profiling" that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, [...].