Data Science and BigData: a Game-changer for Society, Science and Innovation

Roberto Trasarti

- Document for G7 Academy, March 2017, authored by Fabio Beltram: Scuola Normale, Pisa, **Fosca Giannotti**: Istituto Scienza e Tecnologie dell'Informazione, CNR, Pisa, Dino Pedreschi: Dipartimento di Informatica, Univ. Pisa, Pisa
- Document for G7 Academy, March 2018: "Realizing our Digital Future and shaping its impact on Knowledge, Industry, and WorkForce

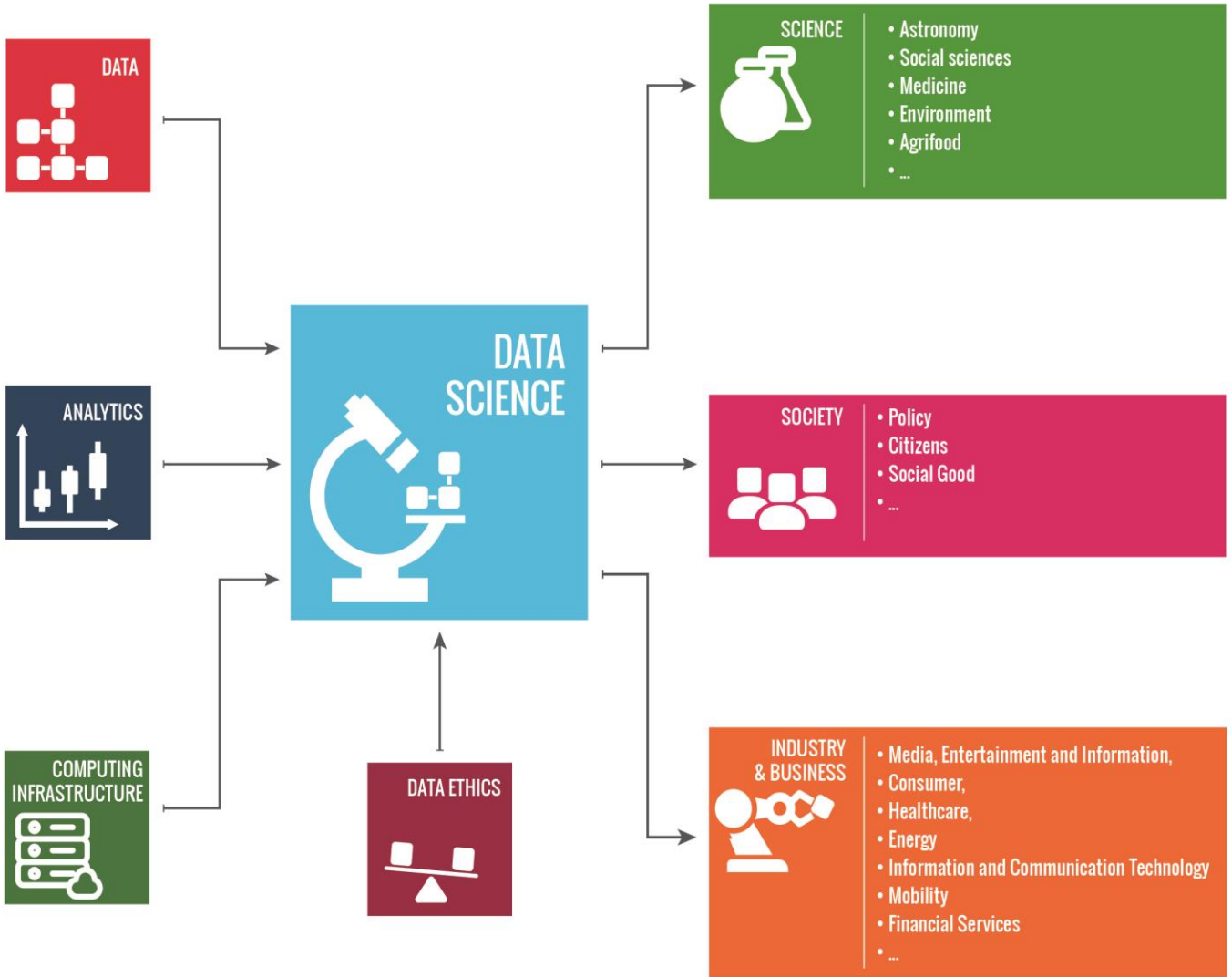
SoBigData

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

KDD Lab

What is data science?

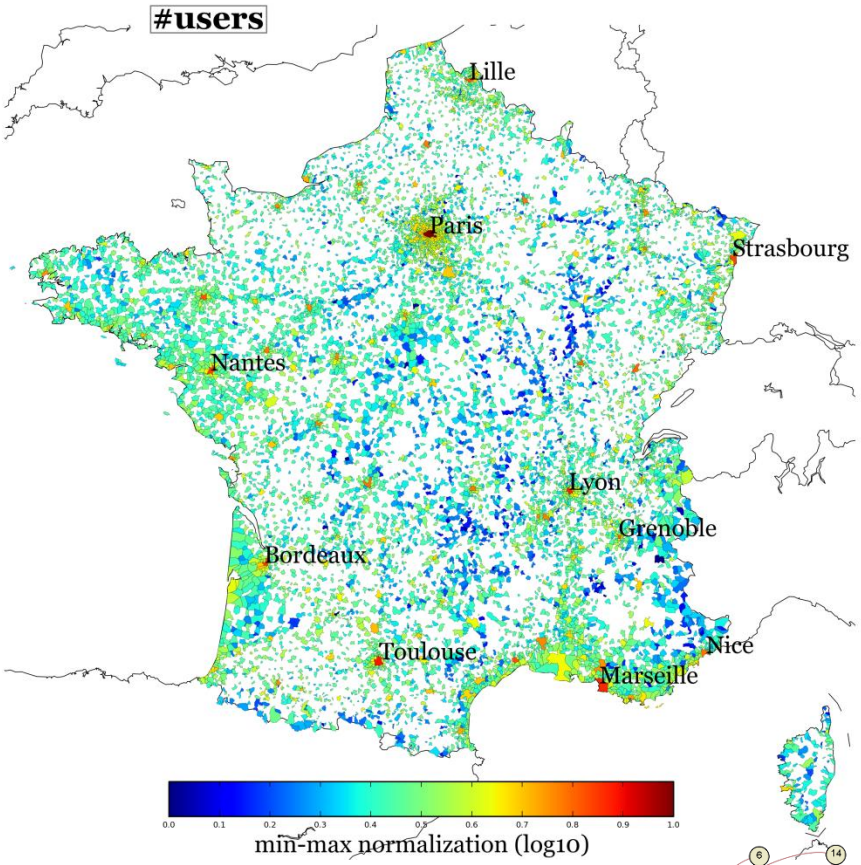
data availability, sophisticated analysis techniques, and scalable infrastructures brought what we call today “Data Science”

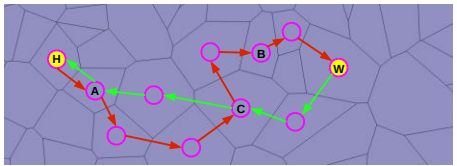


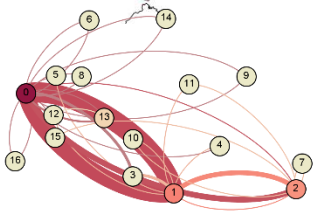
Data Science is a driver for disruptive innovation

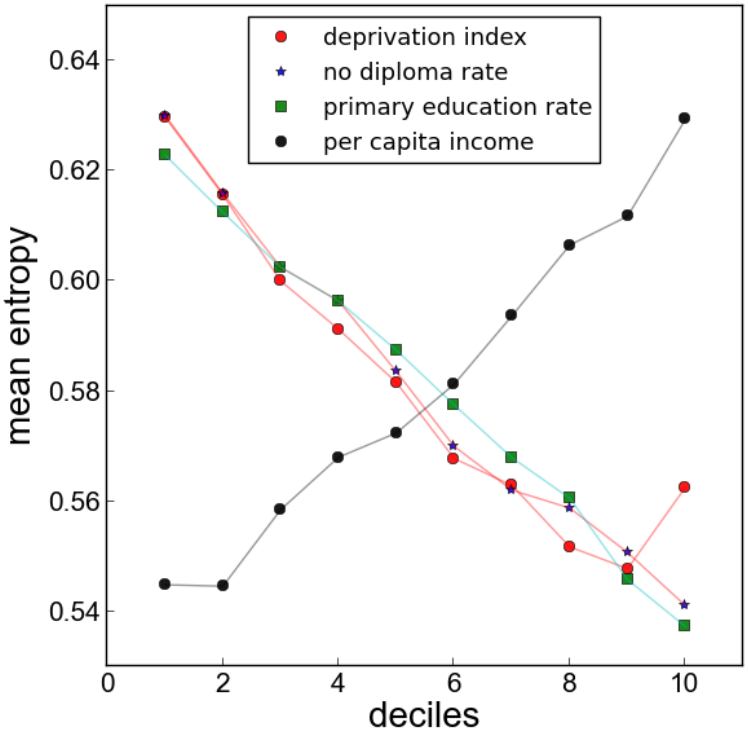
- Empowers citizens, scientists, communities, business and institutions with a **Digital Time Machine** to:
  - Explore the past and present to gain better self-knowledge
  - Explore plausible future to reason on the consequences of decision making

Mobility Diversity and Well-being



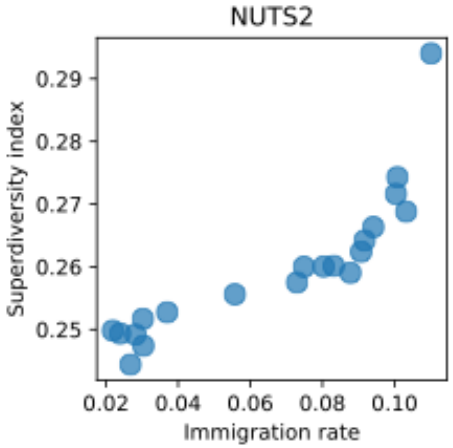


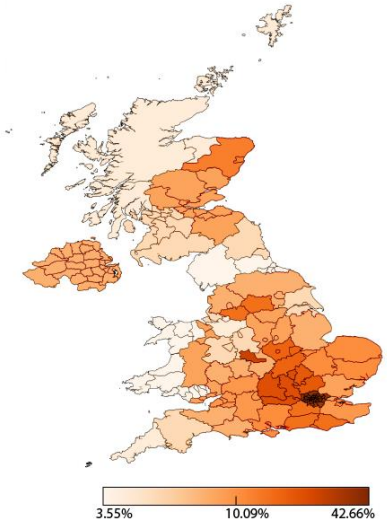




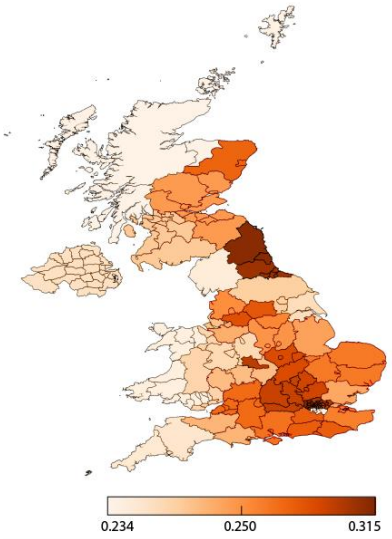
$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i \qquad p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j$$

Migration and Superdiversity on Twitter



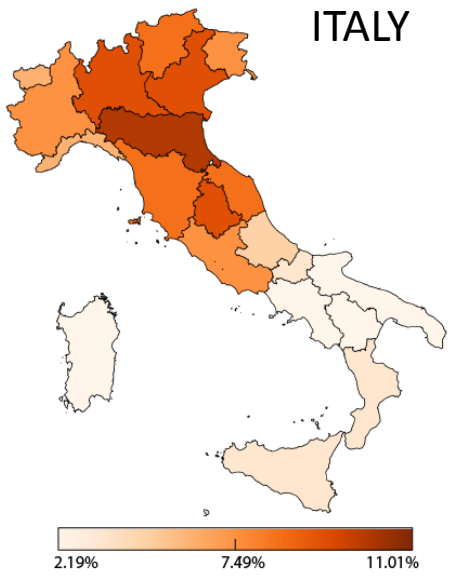


Immigration rate

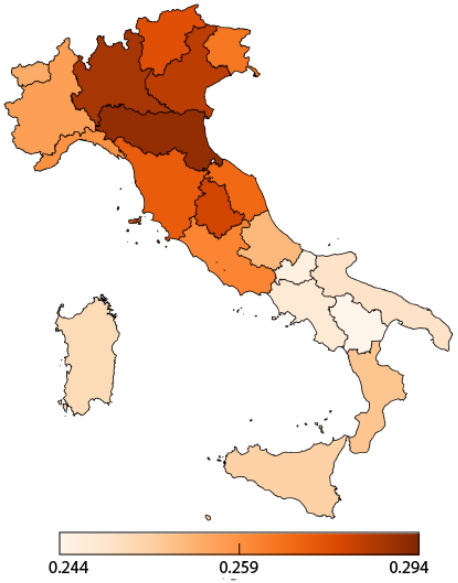


Superdiversity

ITALY

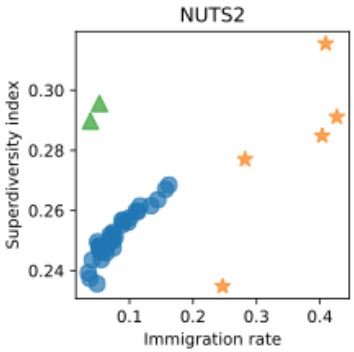


Immigration rate



Superdiversity

UK

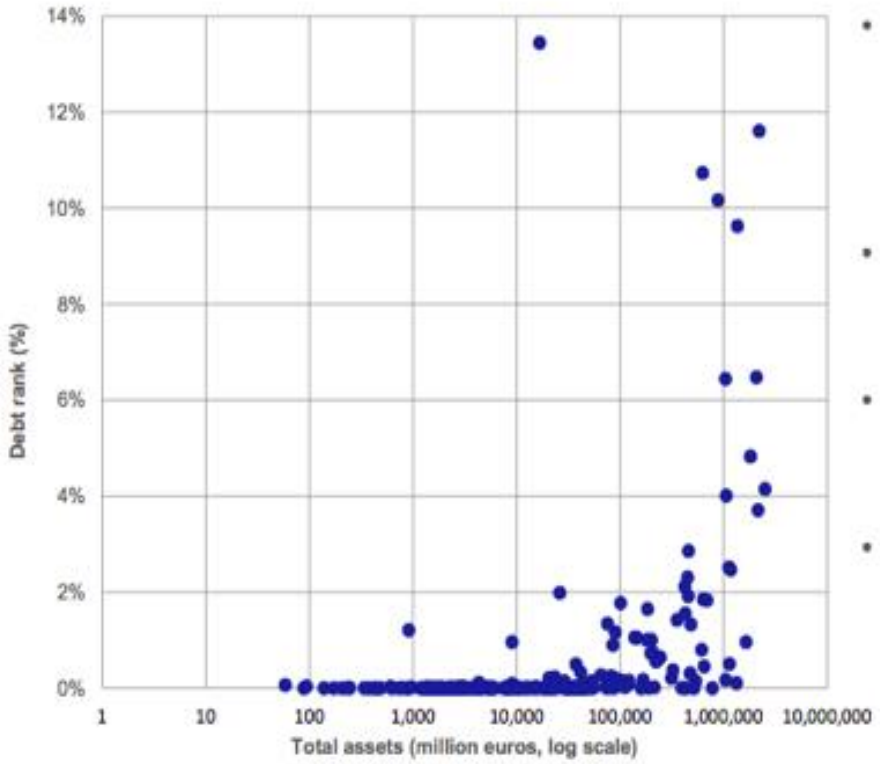


- **Immigration rates:** JRC D4R data challenge
- **Superdiversity:** distance between the emotional content of words in the standard language and on Twitter

Estimating propagation effect of financial distress

WS3: Indicator of marginal bank contagion risk

➢ **Effect of bank failure on euro interbank network (example Dec. 08)**



- Transmission not only through defaults but also proportional to Furfine exposures, relative losses and relative capitalisation of banks
- Contagion risk larger than found in traditional default simulations
- Largest banks have systemic effect (non-linear) but wide dispersion
- Helps, inter alia, to understand the systemic importance of individual banks and how it evolves over time

Simulation of the overall loss of equity (in % of total) among all banks active in TARGET2 caused by individual bank failures ("debt rank" methodology based on a further development of Battiston et al. (2012)) and bank size.
Source: di Iasio, Rainone, Rocco and Vacirca (2013).

TOWARDS REALIZING OUR DIGITAL FUTURE: PERILS OF BIGDATA

Risks

- To convince ourselves that “Privacy is dead”
  - the right to keep personal sphere private as much a person wants is the salt of democracy
- the model of GAFA latifundists. They are good and offer us useful services.
  - Maximiazing “like” and “followers amplifies polarization, is against diversity
- They are BAADDs (by the Economist in Jan. 2018 in **Taming the Titans)**: Big, Anti-competitive, Addictive and Destructive to Democracy

(How the power of data will drive EU Economy)
http://datalandscape.eu/sites/default/files/report EDM_D2.2_First_Report_on_Policy_Conclusions_20.04.2018.pdf

DEMOCRATIZING DATA: TOWARDS AN OPEN SPACE OF DATA SHARING

Towards a new deal on personal data

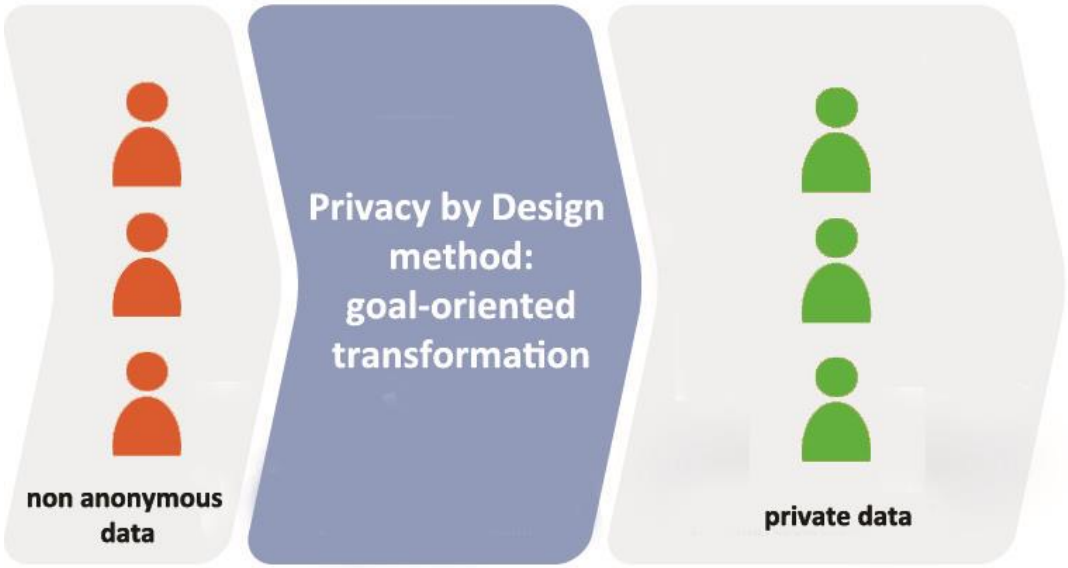
- **Full control of personal data / knowledge**
  - From informed consent to awareness, support for the management of own personal data and knowledge
- **Data liberation**
  - Right to withdraw personal data at any moment in full from any service provider
- **Oblivion**
  - Right to having personal data forgotten
- **Public good**
  - Right to have full access to the collective knowledge

The GDPR is a first step

- Introduces important novelties
  - New Obligations
  - New Rights

PRIVACY BY DESIGN

- Design data driven process that implement the **privacy-by-design & by-default** principle



- Consider privacy at every stage of their business
- Integrate privacy requirements “by design” into their business model.

Technologies

Big Data, Big Risks

- **Big data is algorithmic, therefore it cannot be biased!** And yet…
- All traditional evils of social discrimination, and many new ones, exhibit themselves in the big data ecosystem
- Because of its tremendous **power**, massive data analysis must be used **responsibly**
- Technology alone won't do: also need **policy**, **user involvement** and **education** efforts

The danger of black boxes

- The COMPAS score (Correctional Offender Management Profiling for Alternative Sanctions)
- A 137-questions questionnaire and a predictive model for “risk of crime recidivism.” The model is a proprietary secret of Northpointe, Inc.

- The data journalists at propublica.org have shown that the model has a strong ethnic bias
  - blacks who did not reoffend are classified as high risk twice as much as whites who did not reoffend
  - whites who did reoffend were classified as low risk twice as much as blacks who did reoffend.

The danger of black boxes

- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on …

The danger of black boxes

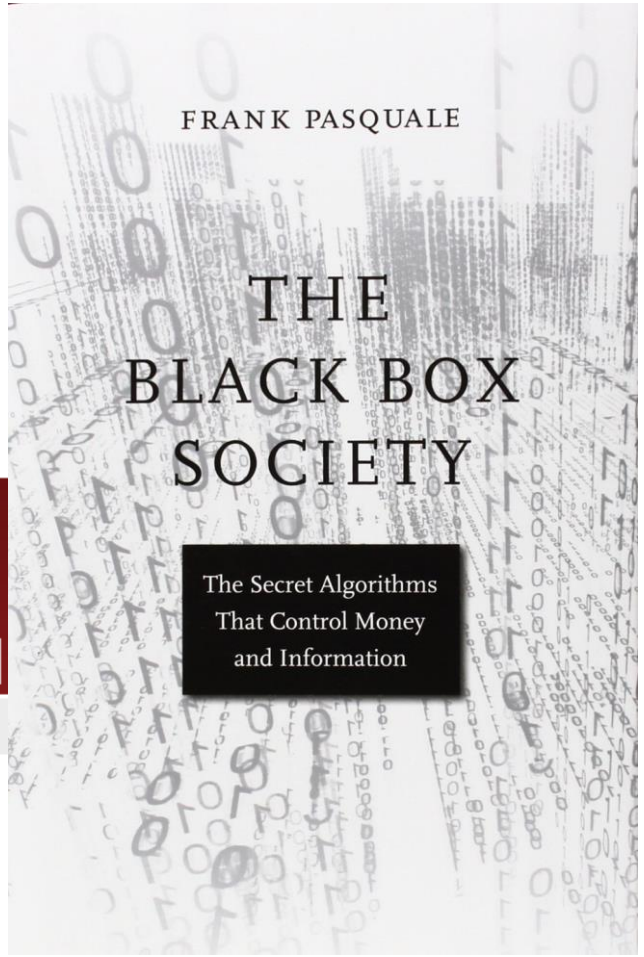
- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on … **the presence of snow in the background!**

(a) Husky classified as wolf

(b) Explanation

Transparent algorithms to build trust

- **Systems that recommend humans making a decision should explain why**
- Gartner says, “by 2018, half of business ethics violations will occur through improper use of Big Data analytics.”



nature International weekly journal of science

NATURE | EDITORIAL

More accountability for big-data algorithms

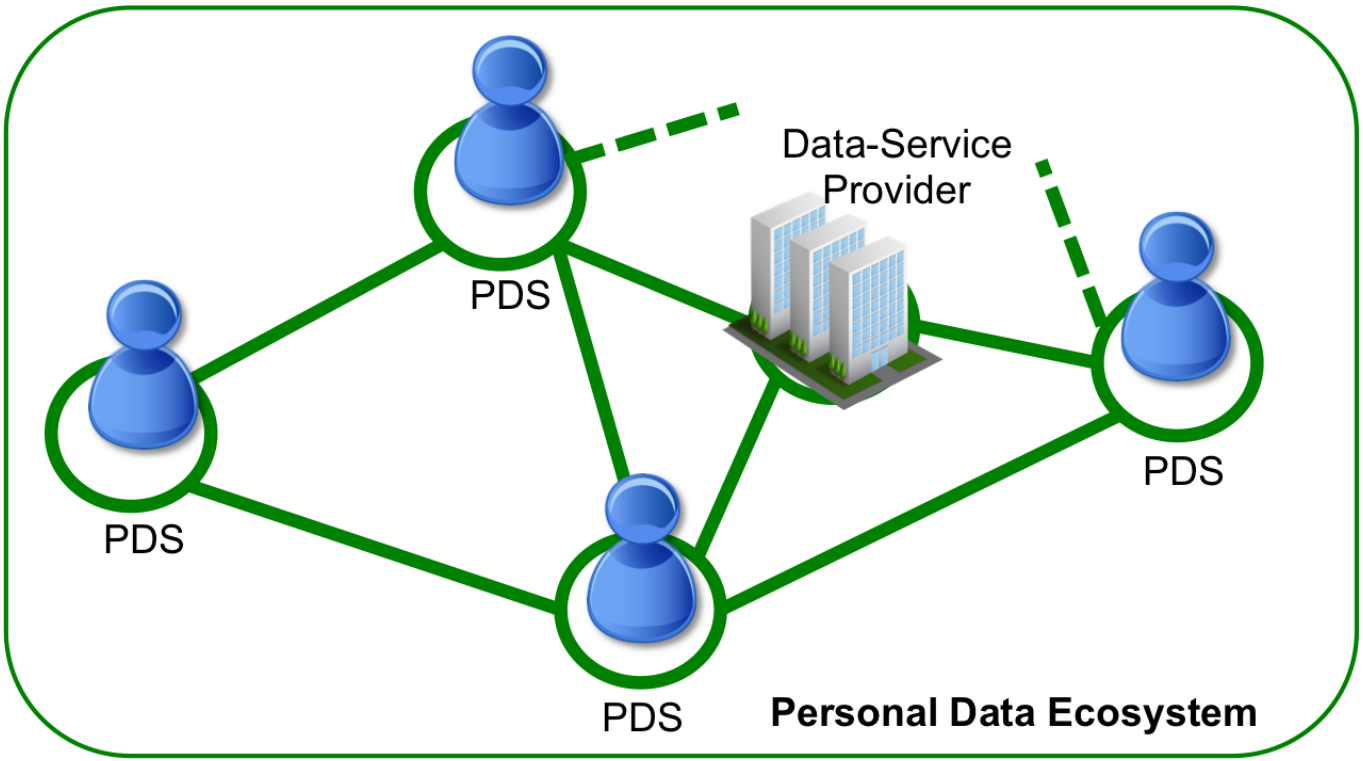
To avoid bias and improve transparency, algorithm designers must make data sources and profiles public.

21 September 2016

MORE COURAGE FROM EU IS NEEDED

A change of perspective: Personal Data Ecosystem

- Personal data collection and knowledge mining need to be balanced with *participation*, based on a much greater awareness of the value of own personal data for each one of us and the communities that we inhabit, at all scales.



ECONOMIC AND BUSINESS OUTLOOK

The EU Data market place

- **EU data market** :
  - in 2016 estimated at almost EUR 60 Billion
  - in 2020 could amount to more than EUR 106 Billion
- **Total number of data companies** in the EU
  - neared the threshold of 255,000 units in 2016,
  - and might rech 360,000 units in 2020.
- **The data economy**
  - represented almost 2% of the EU GDP in 2016.
  - the data economy impacting 4% on the total EU economy in 2020
- The EU data market employed
  - **6.1 million data workers** in 2016
  - 10.4 million in 2020 according to the High-Growth Scenario.

WORLD ECONOMIC FORUM

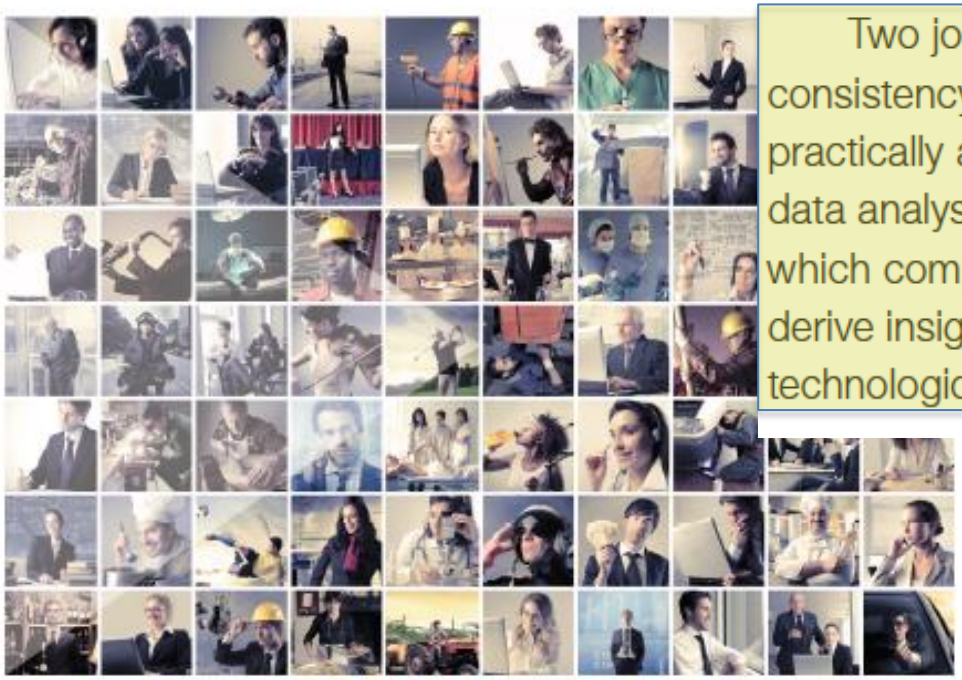
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Global Challenge Insight Report

The Future of Jobs

Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution

January 2016



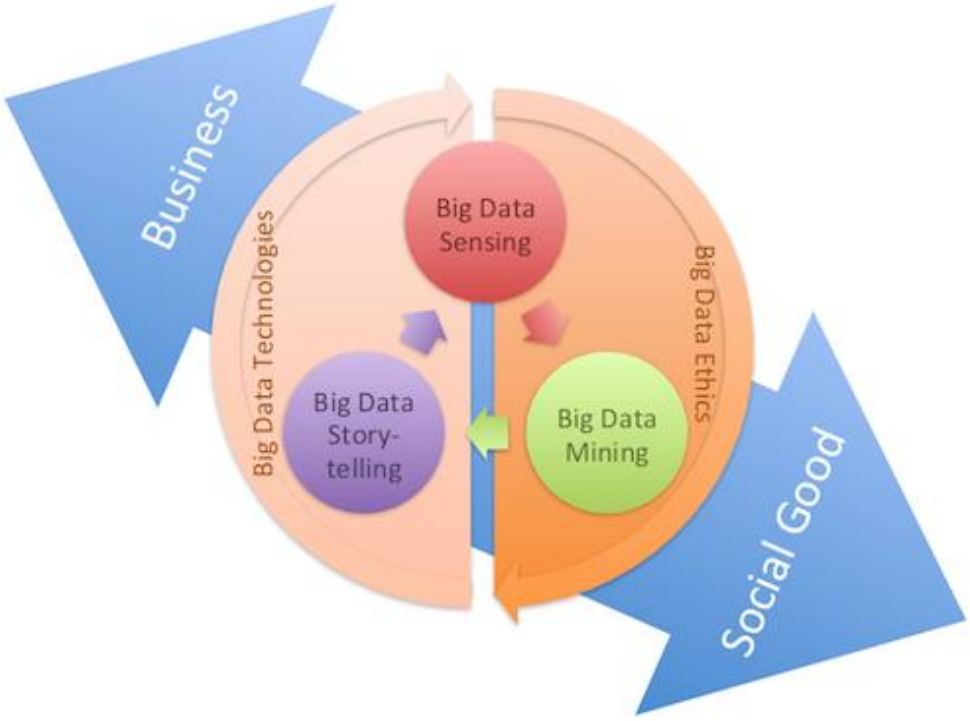
New and Emerging Roles

Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

Two job types stand out due to the frequency and consistency with which they were mentioned across practically all industries and geographies. The first are data analysts, as already frequently mentioned above, which companies expect will help them make sense and derive insights from the torrent of data generated by the technological disruptions referenced above. The second

http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

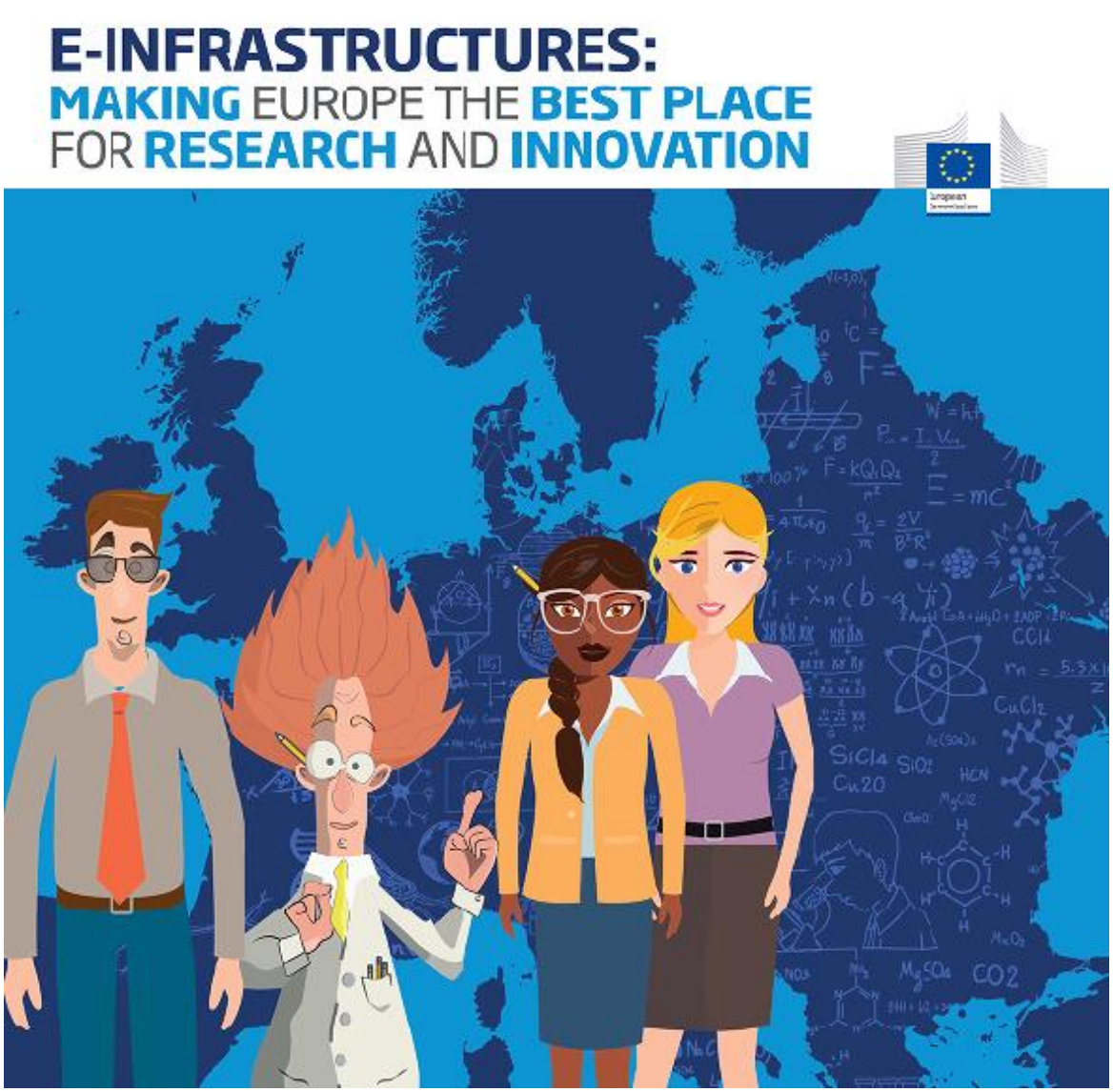
The novel data scientist



- *deep analytical talent* – people with technical skills in statistics and machine learning, for example, who are capable of analyzing large volumes of data to derive business insights;

- *data managers and analysts* who have the skills to be effective consumers of big data insights – i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions;

- *supporting technology personnel* who develop, implement, and maintain the hardware and software tools such as databases and analytic programs needed to make use of big data



- *Not a single centre but a network of excellences interconnected*
- *Connecting people, technologies and data.*

SoBigData Research Infrastructure

www.sobigdata.eu

H2020 excellent science research infrastructure

SCUOLA NORMALE SUPERIORE

IN SUPREMÆ DIGNITATIS 1343

SCUOLA SUPERIORE SANT'ANNA PISA

IMT SCUOLA ALTI STUDI LUCCA

Consiglio Nazionale delle Ricerche

Biblio

- The Big Data Value Strategic Research Innovation Agenda (SRIA) (2017) http://www.bdva.eu/
- Big Data Analytics: towards a European research Agenda ERCIM white paper on Big Data Analytics (2014)https://www.ercim.eu/news/387-ercim-white-paper-on-big-data-analytics