Responsibility of algorithms and data

Serge Abiteboul
INRIA et ENS Cachan
Conseil national du numérique – Académie des Sciences

Joint work with **Julia Stoyanovich**, Drexel University

Organization

- Introduction
- The properties we would like to enforce
  - Fairness
  - Transparency
  - Accessibility by all
- How responsibility can be enforced
  - Education
  - Regulation
  - User unions
  - Technology

The situation

- A few companies concentrate most of the world's data and analytic power
  - They have the means to destroy business competition in large portions of the economy
- A few companies control all your personal data
  - They determine what information you are exposed to
  - They guide many of your decisions
  - They potentially infringe on your privacy and freedom.
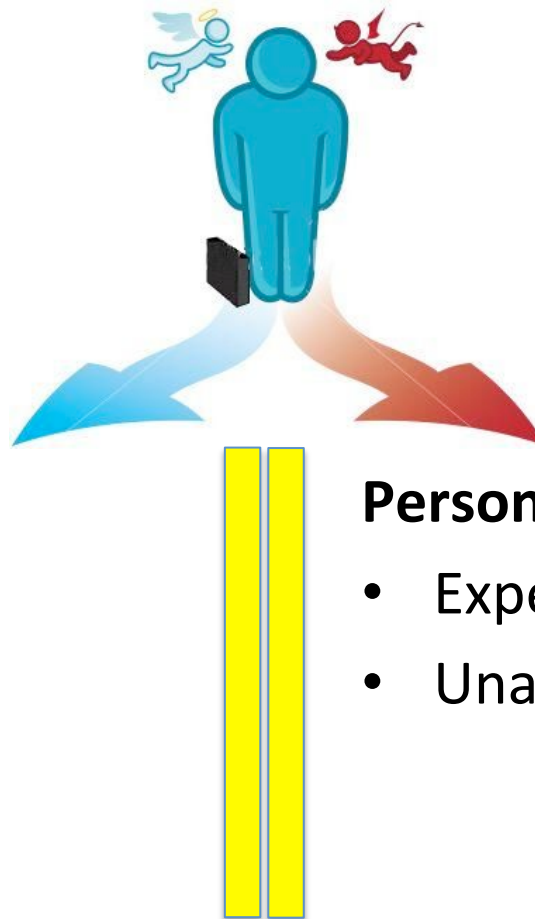- What should we do about that ?

Data, responsibly

Wrong ideas that we hear all the time

- Algorithms are of somewhat divine essence
  - "*The*" algorithm (Google), "*the*" computer (HAL)
  - No! There are many computers and infinitely many algorithms
  - They can be wrong
- They have unlimited power
  - No! They have limits
  - No! There are things they cannot do
- Algorithms are neutral
  - Algorithms are designed by humans and do what they are told to do
- They cannot explain what they do & cannot be challenged
  - Depends which algorithms
  - Some are very "rational"
  - One can verify what they do

Analysis of
Medical data of a person
Her genome
All her social data

Illustration:
Big data &
health

Personalized medicine

- Personalized care

- Predictive measures

Personalized insurance policies

- Expensive for people at risk

- Unaffordable for some

The same technology makes both possible

Responsibility of data analysis algorithms
and not of algorithms in general

We consider for instance

- Web page ranking in Google

- Match making in Meet

- Recommendation in Netflix

- Surveillance by intelligence services
  - Snowden
  - *Loi de Renseignement*

We will not consider

- Killer drones

- Stock market crash in 2010

- Program safety in airlines, nuclear plants…

- Program security in your telephone/computer/car…

- Computer-aided surgery

Similar huge power, so similar issues

The problem is of course not the technology, but what we do with it, what Antoinette calls "the big data ideology"

Because of its tremendous power, massive data analysis (aka "big data") must have responsibilities

Organization

- ✓ Introduction
- The properties we would like to enforce
  - Fairness
  - Transparency
  - Accessibility by all
- How responsibility can be enforced
  - Education
  - Regulation
  - User unions
  - Technology

Fairness – Lack of bias

- Origins of bias
  - data collection
    - E.g., a crime dataset in which some cities are under-represented
  - data analysis
    - E.g., a search engine that skews recommendations in favor of advertising customers
- This bias may even be illegal
  - Offer less advantageous financial products to members of minority groups (a practice known as steering)
- Analogy : analysis of scientific data
  - Should explain how data was obtained
  - Should explain which analysis was carried on it
  - Experiments should be reproducible

Data, responsibly

Fairness – Neutrality

- Such a tremendous power, must come with responsibilities
  - CNNum reports on Net and Platform neutrality
- Some general resources should be « neutral playing field »
  - An Internet provider who refuses to serve Youtube videos or give degraded service for them
  - An App Store who refuses some applications for various reasons or favor some service against another
- Limits the freedom of individuals
- Threatens fair business competition

Fairness – Diversity

- Relevance ranking (for recommendation) is typically based on popularity
  - Ignores less common information (in the tail) that constitutes in fact the overwhelming majority
  - Lack of diversity can lead to discrimination, exclusion.
- Examples
  - on-line dating platform like Match.com
  - a crowdsourcing marketplace like Amazon Mechanical Turk
  - or a funding platform like Kickstarter.

The rich gets richer & the poor gets poorer

Transparency

- Example: lack of transparency in Facebook data processing
  - In general, unreadable End-user license agreement
- Users want to control what is recorded about them, and how that information is used
- Transparency facilitates verification that a service performs as it should, as is promised
- Also allows a data provider to verify that data are well used as it has specified.

Equal accessibility to all

- Data and analysis means more and more concentrated → oligopolies
- Natural outcome of fair competition?
- Why this is not acceptable
  - Loss of freedom of choice for the user.
  - Discourage innovation
  - Eventually leads to an increase of the price of services

Data, responsibly

Organization

✓ Introduction
✓ The properties we would like to enforce
  – Fairness
  – Transparency
  – Accessibility by all
• How responsibility can be enforced
  – Education
  – Regulation
  – User unions
  – Technology

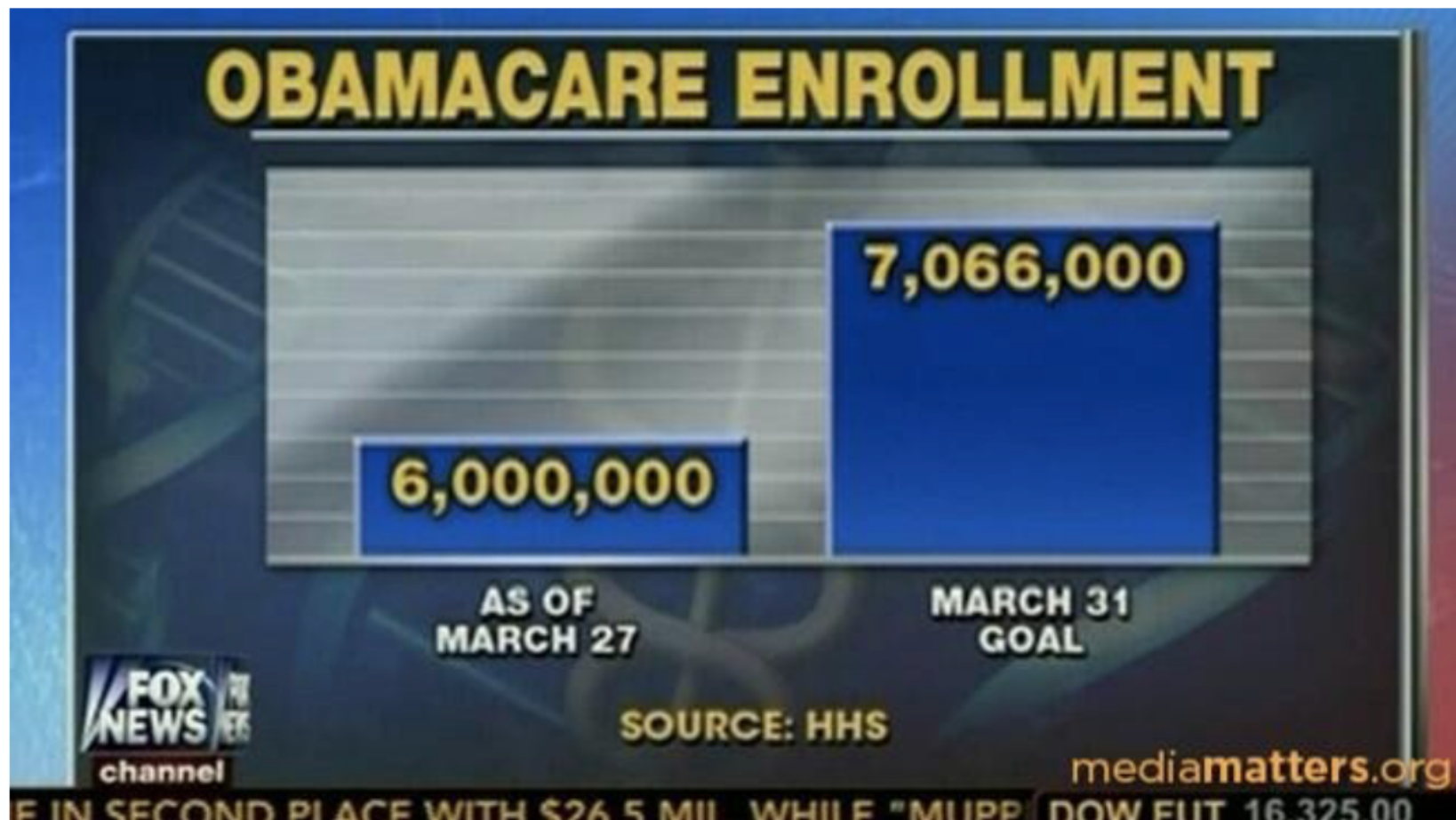# Education

Digital literacy issues

Everyone should learn basis in informatics and basis in data analysis

- To understand the digital world you are living in
- To decide your life in the digital world
- To participate in/contribute to the digital world

Report of CNNum on e-inclusion & education

Education (2)

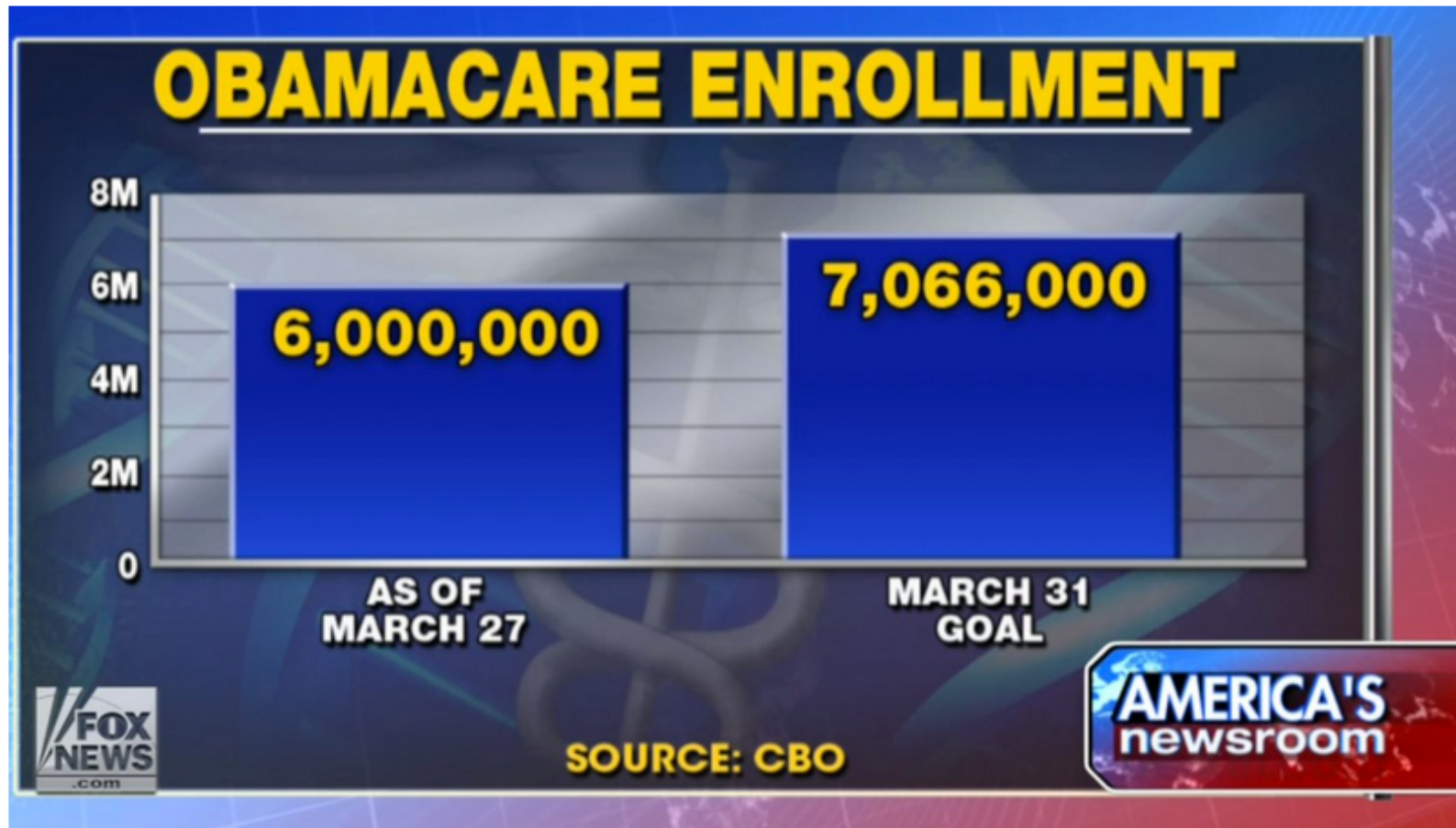http://www.businessinsider.com/fox-news-obamacare-chart-2014-3

Education (2')

Fox News

Government regulation

- This is a complex task
  - Big data technology is complex
  - The technology is changing rapidly
  - Governments lack the competence
  - And the issues are often international
- Governments should
  - Be aware of the problems of data analysis responsibility
  - Define principles and general guidelines
  - Encourage good practices
  - Fight against bad practices such as the building of oligopolies

User associations

- Users have enormous power
  - Example: The Instagram 2012 case
  - Instagram changed its privacy policy to give the new owner, FB, broad access to user data and photos for commercial use
  - forced to change it back under user pressure
- Users should get better organized to
  - discuss with business
  - specify best practices
  - prevent unfair and opaque practices

Technology

- Should provide proper tools
  - To collect data and analyze it responsibly
  - To verify that some analysis was performed responsibly
  - Easier if responsibility is taken into account as early as possible, *by design responsibility*
- To check the behavior of a program, one can
  - Analyze its code ≈ proof of mathematical theorems
  - Analyze its effect ≈ study of phenomena (such as climate or the human heart)
- Simpler in open setting : open data, open source
  - Useful but not sufficient: bug in the SSL library of Debian
    - Weak secrecy of keys for 2 years

Technology: Black box analysis

- More complicated but possible (sometimes)
- Ranking of search engines
  - Statistical analysis
- Sometimes hard/impossible to verify
  - Allow Intelligence black boxes at Internet providers
    - By law, should only look at metadata
    - How do you check?
  - Some data should deleted after a certain time
    - How do you check?

Technology: Machine learning

- Massive data analysis
  - Classical techniques don't scale
  - Machine learning does
- Amazing results
- But
  - Unclear scientific foundations
  - Difficult to explain specific results
  - Does not distinguish between correlation and causality

Conclusion

 to the digital world!

- The massive use of digital information has modified in depth all facets of our life : work, science, education, health, politics, etc.

- We will soon be living in a world surrounded by machines that
  - acquire knowledge for us
  - remember knowledge for us
  - reason for us
  - communicate with other machines at a level unthinkable before

- The use of massive data analysis is the newcomer in that context
  - The system can learn things at a level unthinkable before

Conclusion

- Data analysis has tremendous power
  - Great potential for developing science and innovation
  - Great risk of destabilizing economy and increasing inequalities
  - What will we do with it?

- Religion…science…machines
  - Knowledge used to be determined by religion
  - Knowledge was later on determined scientifically
  - Will knowledge be determined by machines?
  - Then this resembles religion when explanations are not available

Data, responsibly