BIG DATA

Rasmus Pagh

OpenITU, IT University of Copenhagen
January 23, 2015

Licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.
This images of book and magazine covers and the copyright for them are most likely owned either by the artist who created the covers or the publishers. It is believed that the use of low-resolution images qualifies as fair use.
What is BIG DATA?
What is BIG DATA
What is the fuss about BIG DATA?
# Big Data-based businesses

<table>
<thead>
<tr>
<th>Google (search engine)</th>
<th>Data</th>
<th>Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>The web, clickstreams, ...</td>
<td>Advertising</td>
<td></td>
</tr>
<tr>
<td>Amazon</td>
<td>Products, reviews, sales...</td>
<td>Retail</td>
</tr>
<tr>
<td>Facebook</td>
<td>User profiles, interaction, ...</td>
<td>Advertising</td>
</tr>
</tbody>
</table>
# Big Data-based businesses

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Business</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google</strong> (search engine)</td>
<td>The web, clickstreams, …</td>
<td>Advertising</td>
<td>What web pages should I visit?</td>
</tr>
<tr>
<td><strong>Amazon</strong></td>
<td>Products, reviews, sales…</td>
<td>Retail</td>
<td>What products should I buy?</td>
</tr>
<tr>
<td><strong>Facebook</strong></td>
<td>User profiles, interaction, …</td>
<td>Advertising</td>
<td>What things of interest are going on in my network?</td>
</tr>
</tbody>
</table>
Big Data-based businesses

<table>
<thead>
<tr>
<th>Data</th>
<th>Business</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google (search engine)</td>
<td>Advertising</td>
<td>What web pages should I visit?</td>
</tr>
<tr>
<td>Amazon</td>
<td>Retail</td>
<td>What products should I buy?</td>
</tr>
<tr>
<td>Facebook</td>
<td></td>
<td>What things of interest are going on in my network?</td>
</tr>
</tbody>
</table>

In common: Using data to enable better decisions.
Big Data on a smaller scale

Photo by Joe Mabel.
Rough definition

Big Data systems:

• Rely a lot of data (that will often fit on a laptop).
  Combines “all relevant data” from many sources, possibly real-time data.

• Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.
  Serves for “what?” or “how much?”, rather than “why?”
Rough definition

Big Data systems:

• Rely a lot of data (that will often fit on a laptop).

  Not: Playback of videos on Youtube and Netflix.

• Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.

  Served for “what?” or “how much?”; rather than “why?”
Big Data systems:

- Rely a lot of data (that will often fit on a laptop).

Not: Playback of videos on Youtube and Netflix.

- Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.

Not: Computing the strength of a bridge.
Rough definition

**Big Data systems:**

- Rely a lot of data (that will often fit on a laptop).
  
  Combines “all relevant data” from many sources, possibly real-time data.

- Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.
  
  Stares for “what?” or “how much?”, rather than “why?”
Big Data systems:

- Rely a lot of data (that will often fit on a laptop).
  - Often combine “all relevant data” from many sources, possibly real-time data.
- Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.
  - Settles for “what?” or “how much?”, rather than “why?”
**Big Data systems:**

- Rely a lot of data (that will often fit on a laptop).
  - Often combine “all relevant data” from many sources, possibly real-time data.
- Attempt to answer questions where it would be hard (or labour-intensive) for humans to give the answer.
  - Settles for “what?” or “how much?”, rather than “why?”
Software as rules

1. Lamp doesn't work
   - Lamp plugged in? [No] Plug in lamp
   - Lamp plugged in? [Yes] Bulb burned out?
     - Bulb burned out? [Yes] Replace bulb
     - Bulb burned out? [No] Buy new lamp
What is the rule that will allow us to recognize this animal?

Software as rules?
What is the rule that will allow us to recognize this animal?

Software as rules?
The three Vs of big data

- Volume
- Velocity
- Variety
The three Vs of big data

My research area

• Volume
• Velocity
• Variety
The three Vs of big data

- Volume
- Velocity
- Variety

My research area

Scalable Similarity Search
The three **Ps** of big data

- **Technology**
  - My research area
    - Volume
    - Velocity
    - Variety

- **Data sources**

- **Use of analysis**
The three **Ps** of big data

- Possibilities
- Pitfalls
- Privacy concerns

**Technology**
- My research area
  - Volume
  - Velocity
  - Variety

**Data sources**

**Use of analysis**
The three **Ps** of big data

- Possibilities
- Pitfalls
- Privacy concerns
Pitfall 1:

Unfairness

• **Issue**: If data reflects historical biases, these biases may be preserved.
Pitfall 1:

Unfairness

- **Issue**: If data reflects historical biases, these biases may be preserved.

“big data technologies can cause societal harms [...] such as discrimination against individuals and groups. This discrimination can be the inadvertent outcome of the way big data technologies are [...] used.”
Pitfall 2:

Opaqueness

• **Issue**: Many machine learning methods are essentially “black boxes”. A human cannot comprehend the meaning of the models they construct.
Pitfall 2: Opaqueness

- **Issue**: Many machine learning methods are essentially “black boxes”. A human cannot comprehend the meaning of the models they construct.

“some transparency is appropriate when a consumer’s experience is being altered based on their personal information, particularly in situations where companies offer differential pricing”
Issue: If it turns out that your Big Data-based decisions discriminate, can you claim that you were in good faith?
Pitfall 3: Uncertain accountability

- **Issue**: If it turns out that your Big Data-based decisions discriminate, can you claim that you were in good faith?

  “big data will require monitoring and transparency [...] Algorithmists would take a wow of impartiality and confidentiality, much as accountants [...] do now.”
Pitfall 4: Spurious discovery

Honey producing bee colonies (US) inversely correlates with Juvenile arrests for possession of marijuana (US)
Pitfall 4:

Spurious discovery
Pitfall 4:

Spurious discovery
• **Issue**: What does it even mean that certain data about me cannot be used?
The last P:

Privacy concerns

- **Issue:** What does it even mean that certain data about me cannot be used?
Many technologists are of the view that de-identification of data as a means of protecting individual privacy is, at best, a limited proposition.

**Issue**: What does it even *mean* that certain data about me cannot be used?
Can we do it?
Can we do it?

Distinguishing cause from effect using observational data: methods and benchmarks

Joris M. Mooij*
Institute for Informatics, University of Amsterdam
Postbox 94.323, 1090 GH Amsterdam, The Netherlands

Jonas Peters
Seminar for Statistics, ETH Zürich
Rämistrasse 101, 8092 Zürich, Switzerland

Dominik Janzing
Max Planck Institute for Spemannstraße 38, 72076

Jakob Zscheischler
Max Planck Institute for Biosciences
Hans-Knöll-Straße 10, 07745

Bernhard Schölkopf

Preserving Statistical Validity in Adaptive Data Analysis

Cynthia Dwork* Vitaly Feldman† Moritz Hardt‡ Toniann Pitassi§
Omer Reingold¶ Aaron Roth‖

November 12, 2014

Abstract

A great deal of effort has been devoted to reducing the risk of spurious scientific discoveries from the use of sophisticated validation techniques, to deep statistical methods for controlling the false discovery rate in multiple hypothesis testing. However, there is a fundamental disco
Take-home message

Big Data presents many Possibilities, but attention should be paid to Pitfalls, e.g.:

- Unfairness
- Opaqueness
- Uncertain accountability
- Spurious discovery

Privacy regulations need to be phrased in a balanced and technically meaningful way.
Thank you!

Slides available at www.itu.dk/people/pagh/papers/
Some pointers

• Cathy O’Neil: mathbabe.org/2014/08/25/gilian-tett-gets-it-very-wrong-on-racial-profiling/

• Moritz Hardt: medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

• Workshops on Differential Privacy; Fairness, Accountability, and Transparency in Machine Learning: simons.berkeley.edu/workshops/bigdata2013-4; www.fatml.org