Analytics and Decision Making with Big Data: Big Opportunities and Big Pitfalls

Sertac Karaman
Assistant Professor of Aeronautics and Astronautics
Laboratory for Information and Decision Systems
Institute for Data, Systems, and Society
Massachusetts Institute of Technology

What is in Big Data?

- **Define Practical Big Data with 4V's:**
  - Volume (Data Quantity)
  - Velocity (Data Rate)
  - Variety (Data Types)
  - Veracity (Messiness)

Why Do We Have Big Data Now?
Why Did Not This Exist Before?

- **The Key Enablers of Big Data**
  - Increase in (affordable) storage capacities
  - Increase in processing speed and (affordable) computation
  - Increase in the availability of data
    - Massive deployment of sensors, data acquisition (e.g., smart phone)

- As a result, we generate 2.5 quintilion bytes of data every day.
  (2.5 x 10^{18}, or roughly 2.5 million terabytes)

- It is estimated that 90% of all our data was generated in the last two years.

MIT AEROASTRO

How Much Data Do We Generate Really?

- **Every minute, we generate:**
  - More than 204 million email messages
  - Over 2 million Google search queries
  - 48 hours of new YouTube videos
  - 684,000 bits of content shared on Facebook
  - More than 100,000 tweets
  - $272,000 spent on e-commerce

Big Data: What is Good, What is Bad?

- **The Good:**
  - Big
  - Timely
  - Predictive (sometimes)
  - Cheap (usually)
- **The Bad:**
  - Unknown population representation
  - Issues of data quality
  - Privacy and confidentiality issues
  - Difficult to assess accuracy and uncertainty

Where are Some of the Classical Applications?

- Smarter health care:
  - 80% of data is unstructured and clinically relevant.
  - Data in multiple places: lab and imaging reports, physician notes, medical correspondence, claims.

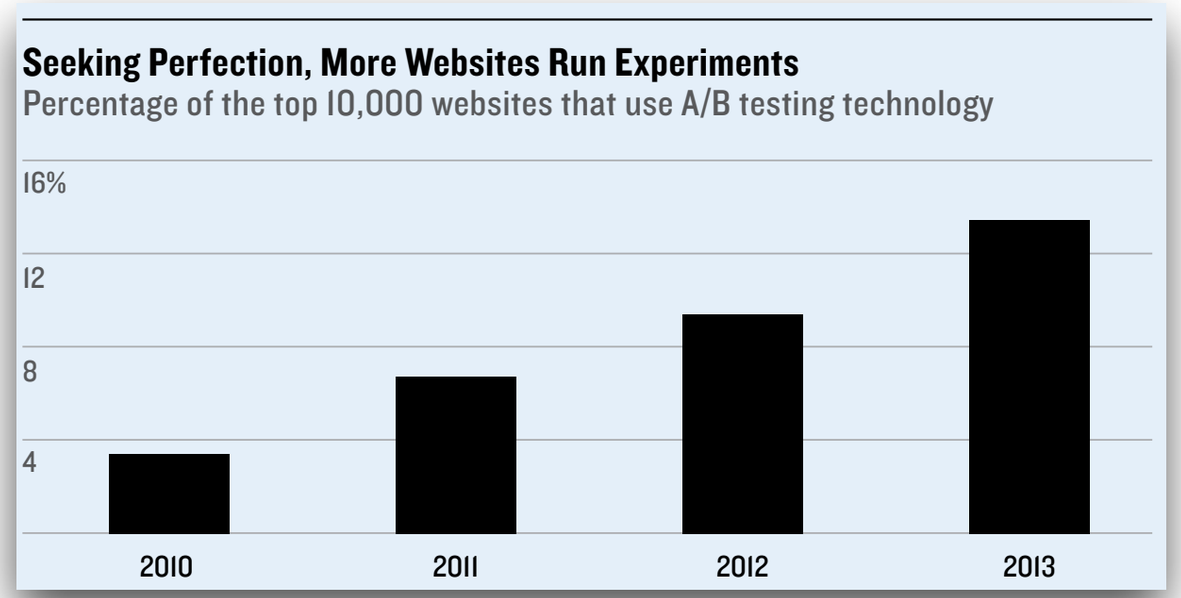
How Can We Utilize Big Data Now?
Experimentation for Better User Experience

- Optimizely runs “experiments” to decide how your website looks, how it is organized, and what goes on your website.
    - Slightly different versions of the website are provided the different users, and users’ interaction with the website is measured. — The data decides the winner.

Optimizely

- Google experimented with 41 different shades of blue to decide the current.

- Optimizely made many of the devisions for Obama’s online campaign.



MIT AEROASTRO

Scientific Method for Business Development

- Scientific method is based on **controlled experiments** that offer evidence for how nature really works.
- Big tech companies like, Facebook, Google, Microsoft, Amazon utilize experiments reap benefits/profits utilizing big data.

- However, there are number of pitfalls:

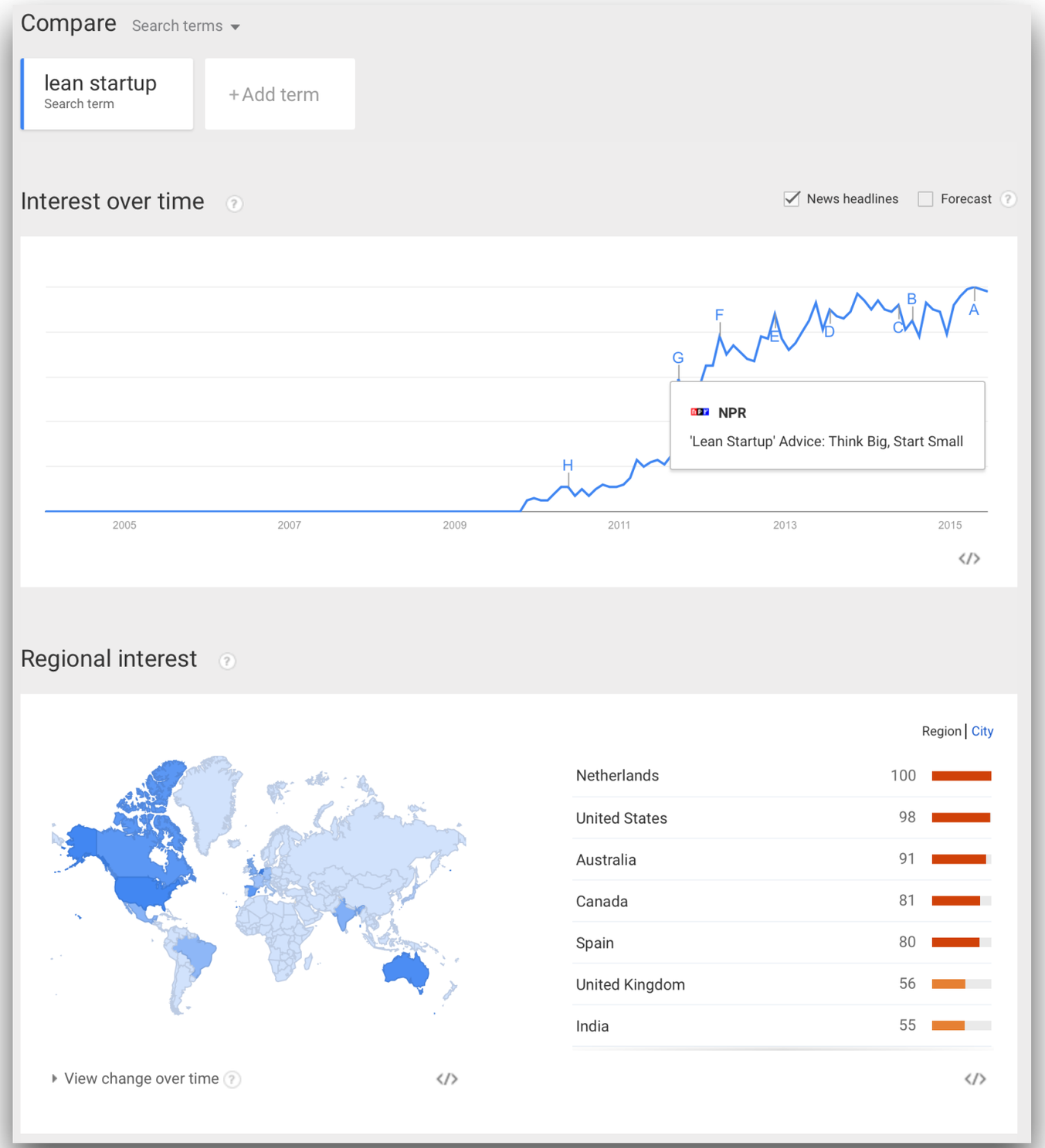
The New York Times

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

JUNE 29, 2014

Facebook revealed that it had altered the news feeds of over half a million users in its study.

Quickly Rule Out Bad Ideas



- Do you have an interesting idea?
- Quickly experiment with a few customers and observe the results, before you invest.

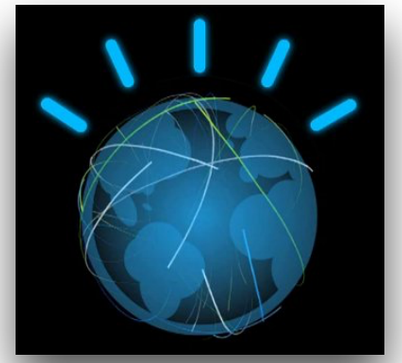
MIT AEROASTRO

Other Potential Pitfalls

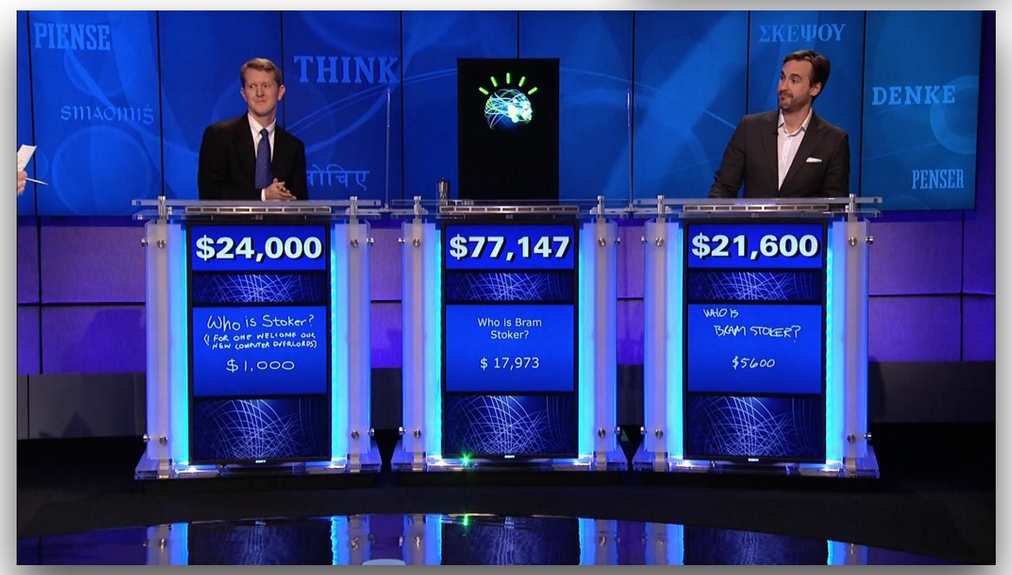
- **Over-reliance on data:**
  - Could not see the 2007-2009 financial crises, mainly due to confidence for positive data.

- Complex human phycology, human interactions create data:
  - Why “Charlie Bit Me” and “Gangnam Style” have been the most-viewed videos on YouTube?

Where is Watson Headed?



- Watson become famous by winning *Jeopardy!*



- IBM created the Watson Group, working solely on Watson, as a result of merging with ‘cognitive computing’ groups at IBM.
  - The Watson Group has its own operational officers (e.g., its own CTO), and it directly reports to IBM’s CEO.
  - Roughly 2,000 people work for the Watson Group
  - IBM hopes to generate $10 billion in revenues.
  - Currently, the group generates roughly $100 million in revenues.

- “You have to put this stuff into action and refine it. You need examples to work on”
  - Rob High, *Watson Group CTO.*

MIT AEROASTRO

What Can Watson Do Today?

- Focuses on the analysis of unstructured data:
  - Analysis of text, images, audio, and video.
- Applications of audio and video are also vast:
  - Audio data in call centers, meetings.
  - Video data in cameras and movies (e.g., consider a movie recommendation tool that actually watches the movies!)

- “It will be a Watson system that can hear, see, and talk.” - Rob High

- The Watson Group currently targets:
  - biomedical industry (e.g., cancer diagnosis)

What May Be Even More Powerful?
On Human-Computer Decision Making

- IBM's computer Deep Blue defeated Gary Kasparov in 1997. It was big!
- Even bigger in the chess world:
  - In 1995, two U.S. amateurs armed with three PC's defeated the best players/computers!!
  - It changed the field, many grandmasters started training with computers.

What May Be Even More Powerful?
On Human-Computer Decision Making

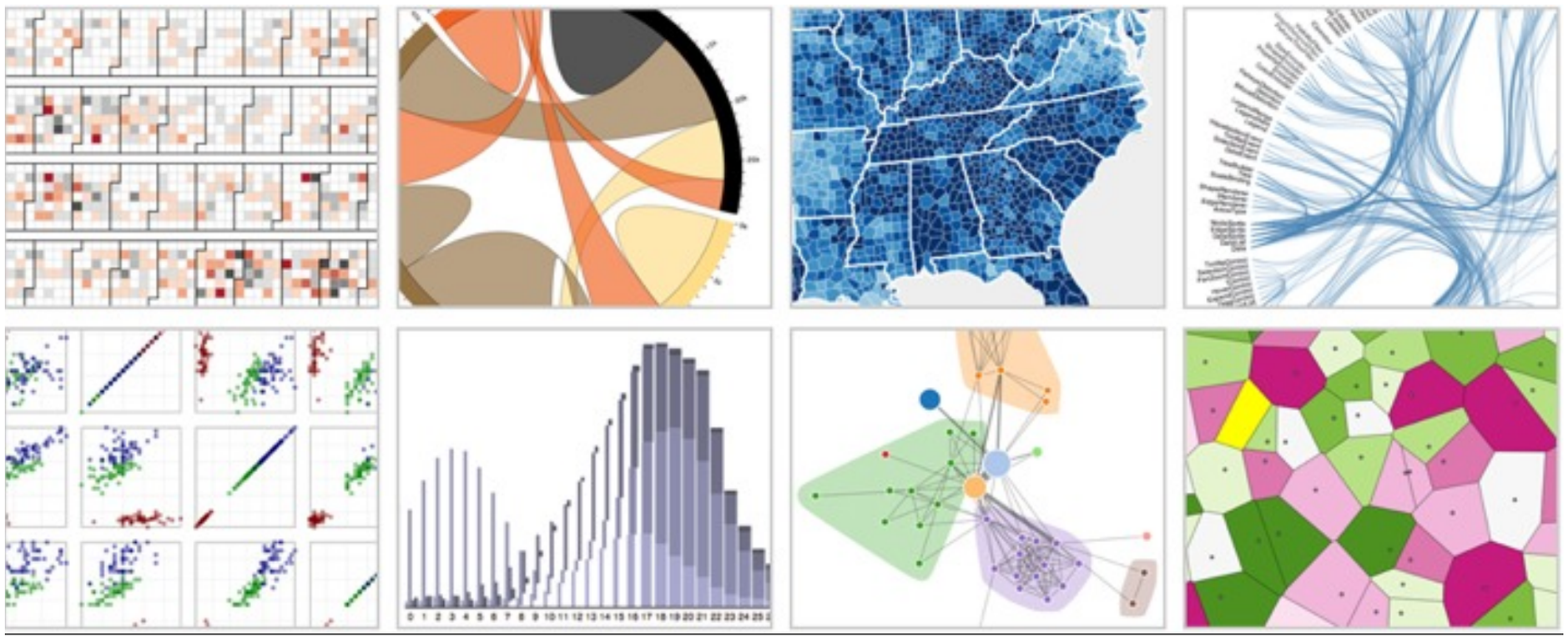
- Palantir is founded by ex-PayPal-engineers, who designed an automated system to catch fraud using big data.
- Their software caught 80% of the cases. The remaining 20% chased by people.
- At Palantir, they provide tools that fosters, what they call “human-computer symbiosis”.
- Palantir was valued at $15 billion in January of 2015.

Data Visualization

- Good “human-computer symbiosis” requires tools that go beyond optimization, statistics, and machine learning.

- Excellent **data visualization** and **interfaces** may be key tools for the next-generation managers, engineers, business professionals, …

- *We will look into rapid data visualization and rapid interfacing next!*



MIT AEROASTRO