

Scientific Data Analysis: Course Introduction

Jarek Szlichta

<http://data.science.uoit.ca/>

Background

- Taught **Big Data Analytics** (CSCI 4030)
 - in winter
- High demand
 - Students seemed very happy 😊
- Now **Scientific Data Analysis**
- **New Data Science Specialization in CS!**
 - Formally approved last year
 - <http://www.science.uoit.ca/undergraduate/programs-and-information-for-prospective-students/computer-science/data-science-specialization.php>

Who Should Take the Course?

- Non-CS majors or early CS majors
- Not afraid of numbers
- Not afraid of computer tools
- Took equivalent of one programming class

What Does “Big Data” Mean?

(1) Collecting large amounts of data

- Via computers, sensors, people, events ...

(2) Doing something with it

- Make decisions, confirm hypotheses, gain insights, predict future ...

- **“Data Science” = Going from (1) to (2)**

Is Big Data a Fad?

Was computer programming a fad?

- **Ability to collect data will only increase**
- **Ability to analyze data will only improve**

Facebook Country..

1. China (1.339 billion)
2. India (1.218 billion)
3. Facebook (1 billion)

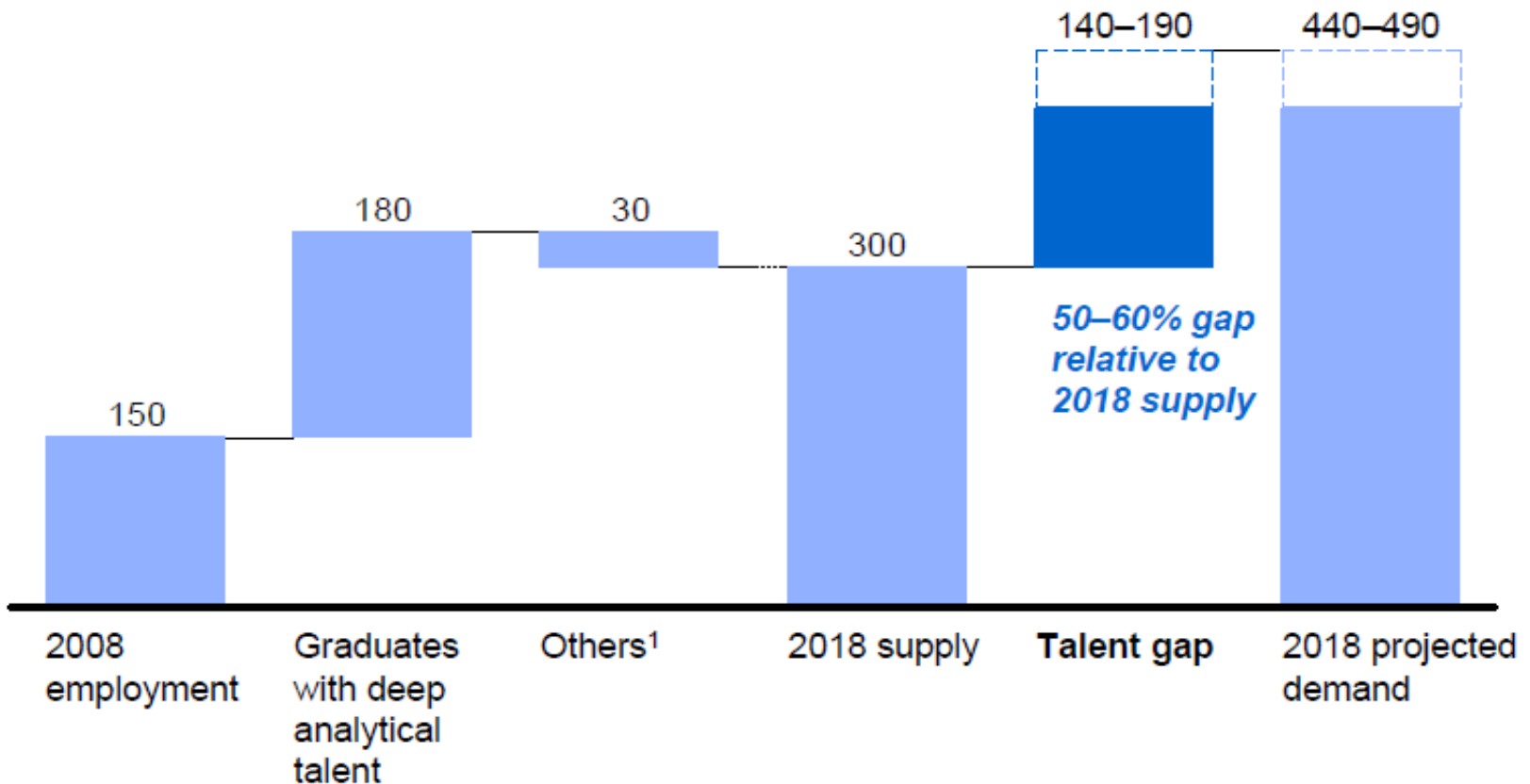


Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



¹ Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Rest of Today

- Introduction to good stuff
 - Utilities and discoveries
- Introduction to bad stuff
 - Pitfalls and privacy
- What the course will cover
- Class logistics

Utilities




How do you want that data?

Traffic



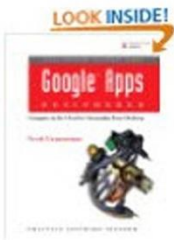
- (1) Collect Data
- (2) Do something with it

Recommendation Systems




Recommended for You


Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



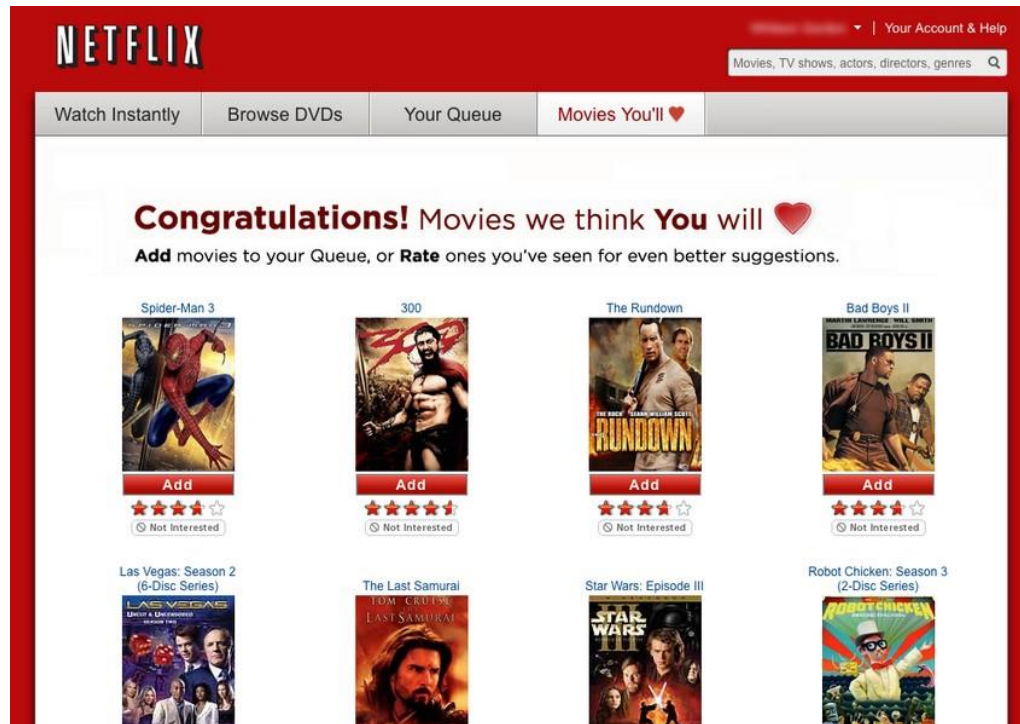
[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

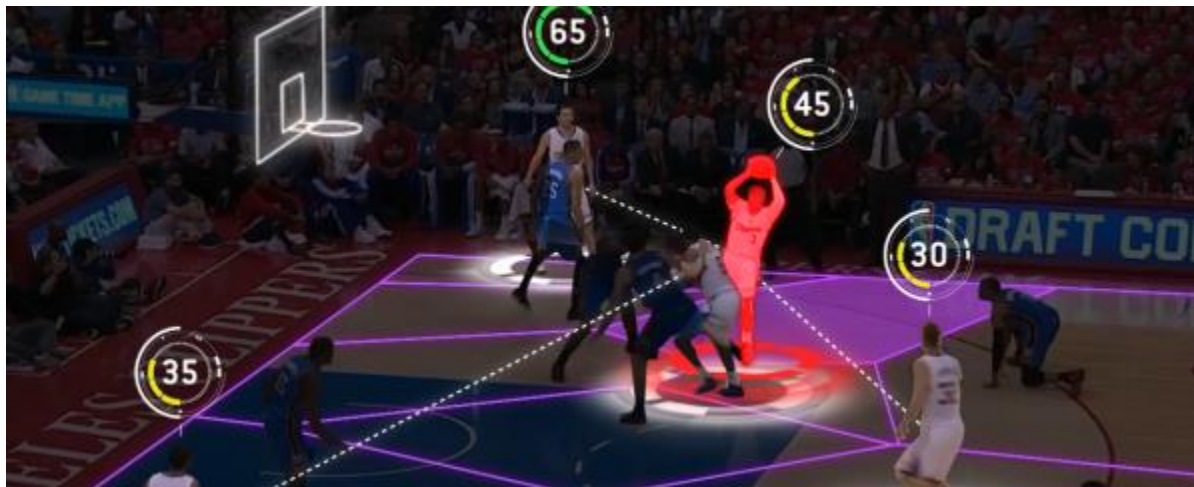
- (1) Collect Data
- (2) Do something with it

Recommendation Systems



- (1) Collect Data
- (2) Do something with it

Sports



- (1) Collect Data**
- (2) Do something with it**

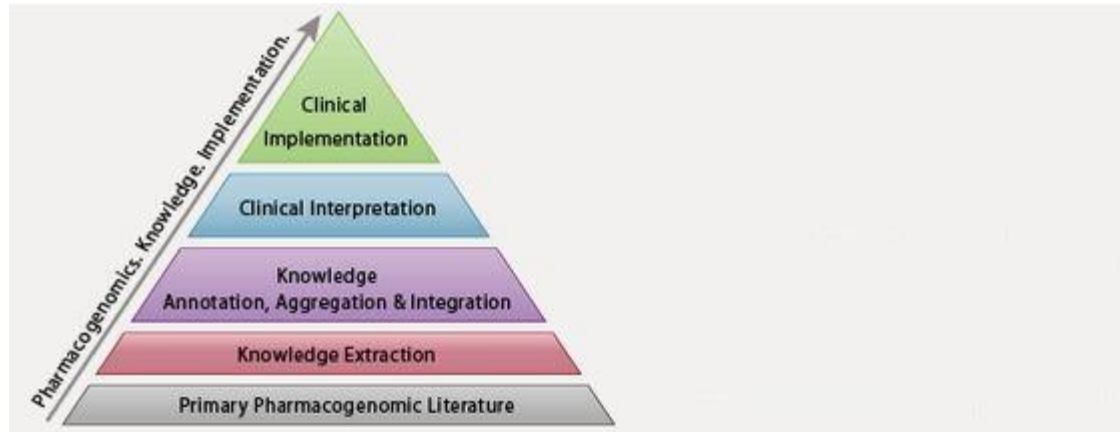
Advertising

<http://www.fastcompany.com/3036425/afacebook-users-challenge-to-facebook-heres-allmy-data-now-give-me-ads-i-like>

(1) Collect Data

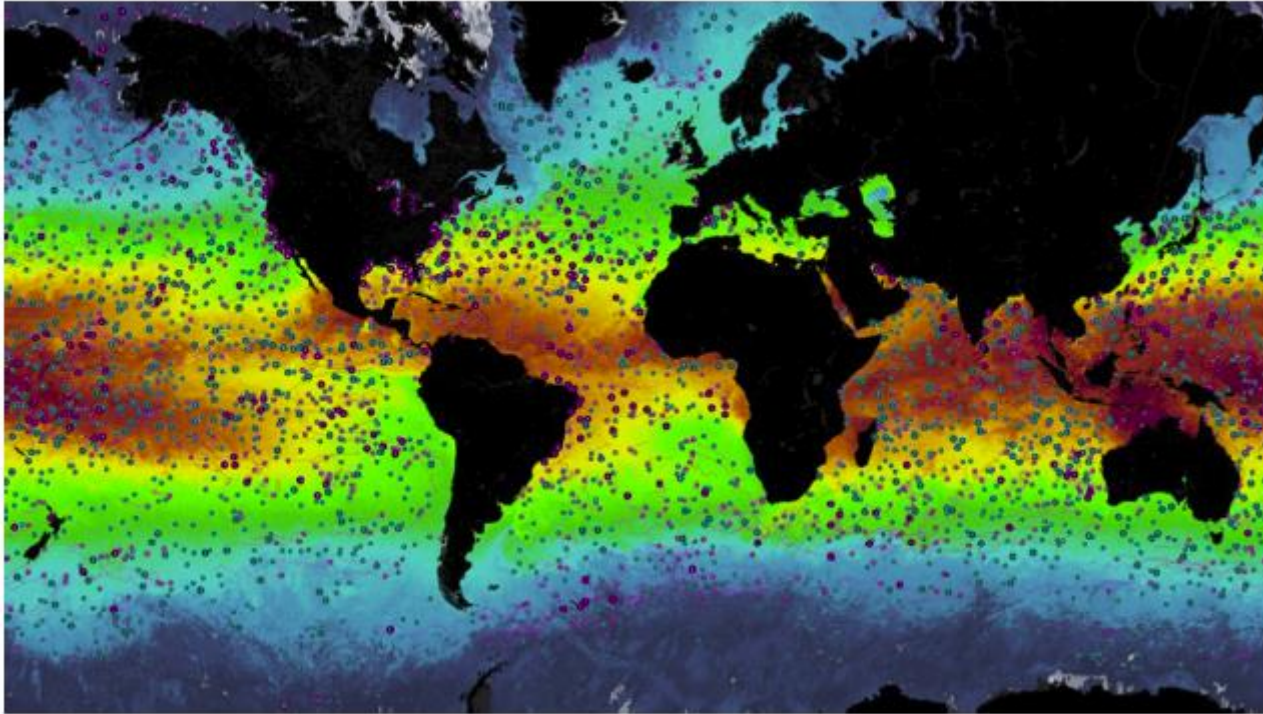
(2) Do something with it

Gene-Drug Relationships



PharmGKB collects, curates, and disseminates knowledge about the impact of human genetic variation on drug responses.

Ocean Health



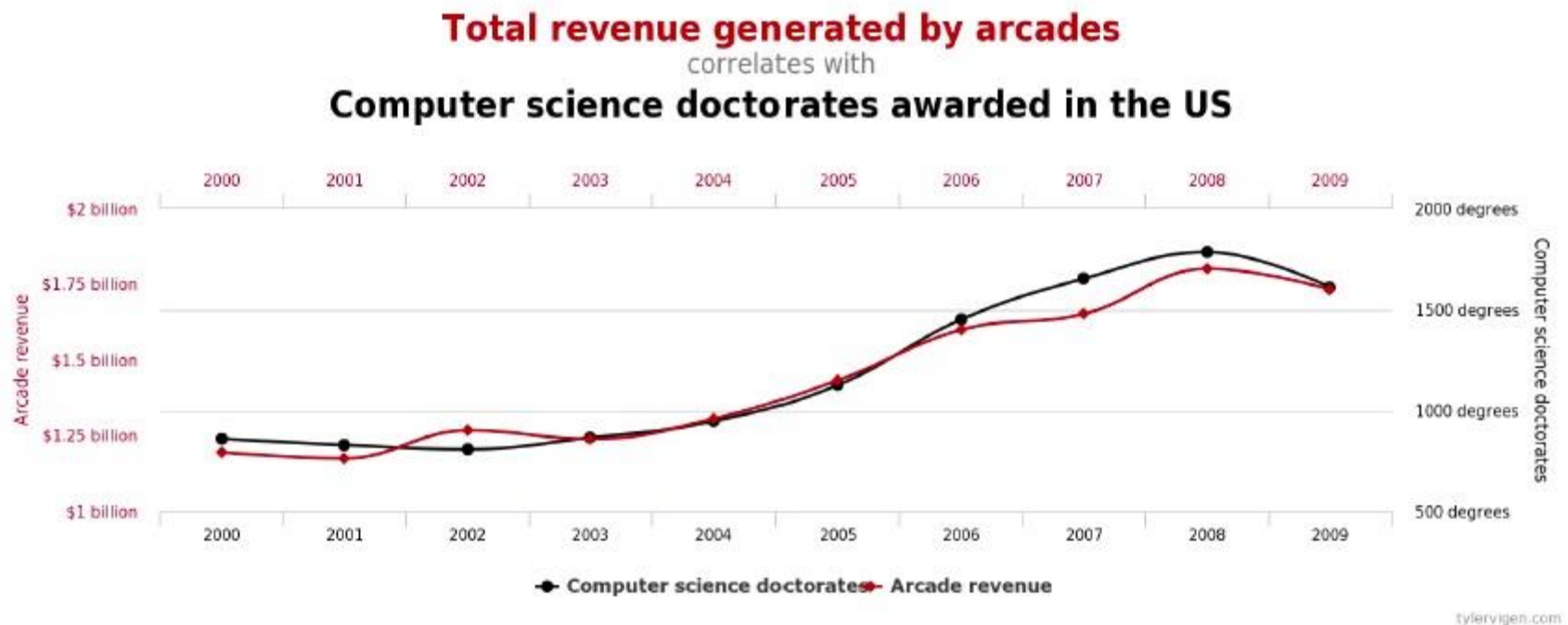
- 44,000 sensors, over 2 billion measurements
- Physical, chemical, biological

Pitfalls

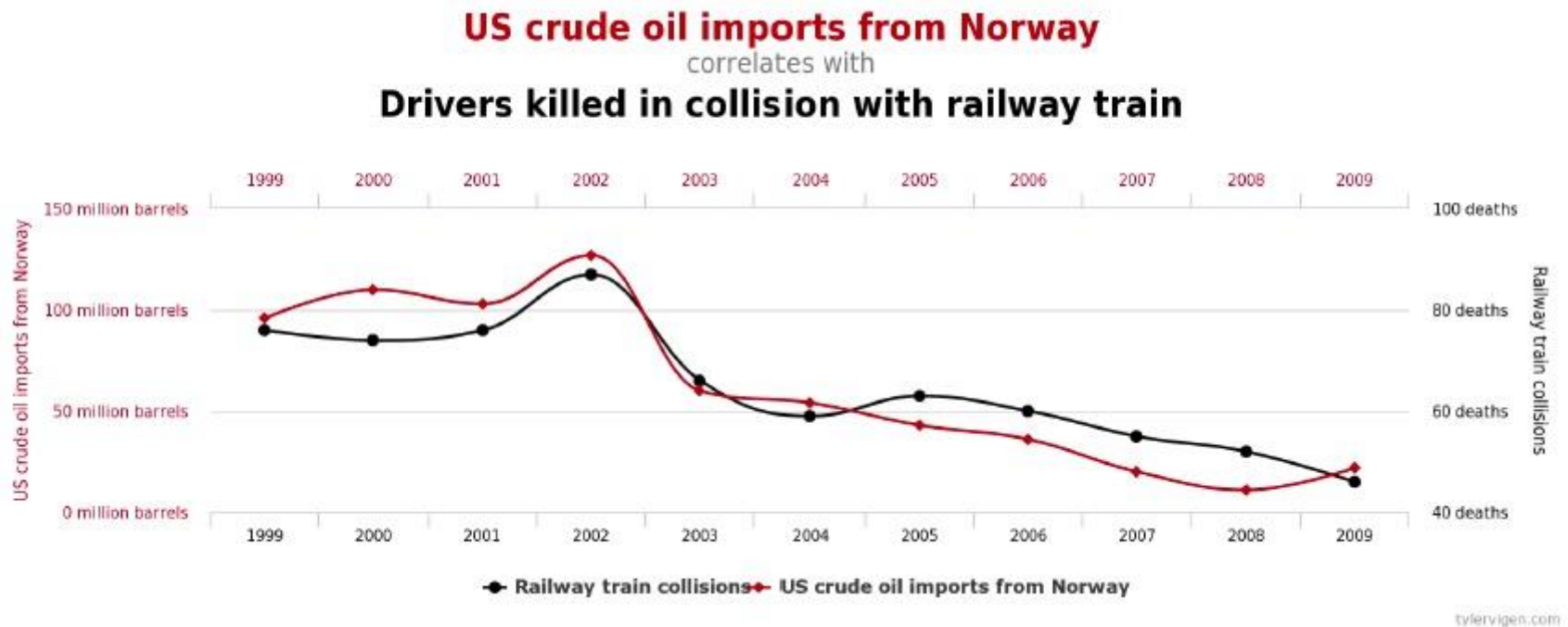
Pitfall: Causation



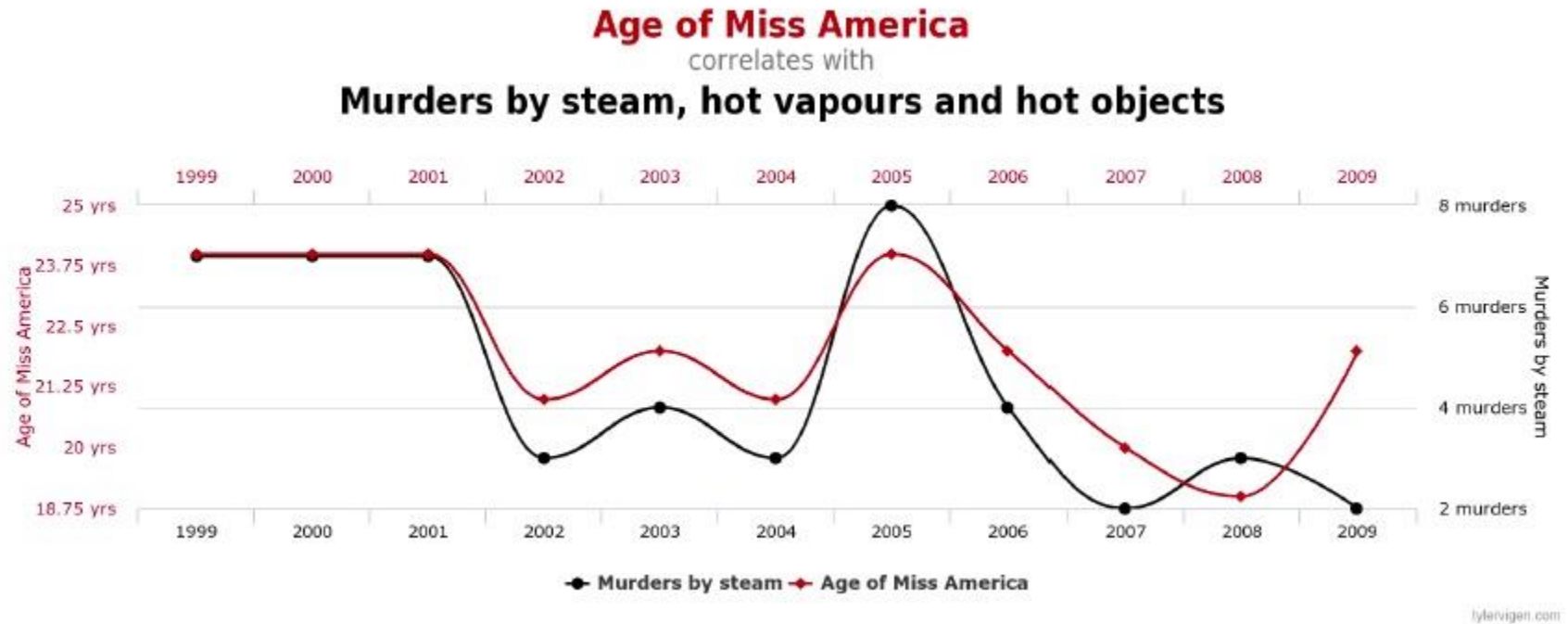
Pitfall: Correlation



Pitfall: Correlation



Pitfall: Correlation



Privacy

- Publicized cases of improper collection of individual data
- But individual data can also bleed legally through surprising channels



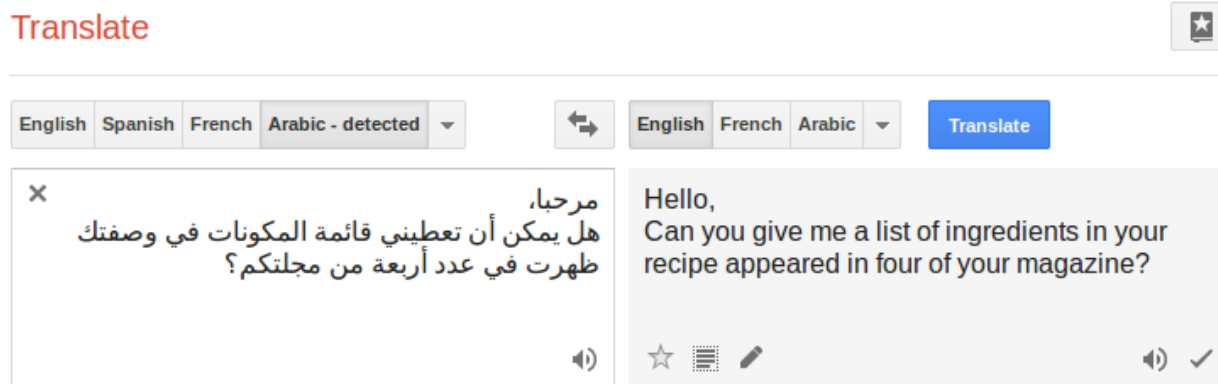
Google Flu Trends

- Geographic Flu Tracking based on user search terms
- Now it does not work
 - “Big Data Hubris”



Google Translate

- Automatically learn language translation from examples on the web.
 - Anyone see a problem?



Football Game Predictions

- Saturday: receive email from “Prescient Polly” predicting results of four Sunday football games. She’s right.
- Same thing the following weekend.
- And two more weekends.
- Fifth Saturday: Polly offers to place bets for you on Sunday games, for a fee.
 - **Should you do it?**

Football Game Predictions

- How many contacts does Polly need on week one for 100 potential betters on week five?
 - **65,536 x 100 (around 6.5 million)**

Enough Negativity

- Discoveries outweigh pitfalls
- Balance will only improve

What We will Cover

- **Data analysis techniques**
 - Basic data operations, data mining, machine learning (regression, classification, clustering)
- **Tools for data management & analysis**
 - Data visualization, SQL, Python, R, Weka, Google Spreadsheets
- **Anomaly detection, sampling and statistical significance**
- **Plus: guest speakers (IBM, Huawei, Dapasoftware, UofT), case studies, pitfalls, privacy issues**

About the Course

CSCI2000 Course Staff

■ TAs:

■ We have great TAs!

- Spencer Bryson (spencer.bryson@uoit.net)
- Lachlan Johnson (lachlan.johnston@uoit.net)

■ Office hours:

- **Me:** Thursdays 4pm-5pm, UA 4020
- See course website for TA office hours (TBD)

Course Logistics

- **Course website:**

- <http://data.science.uoit.ca>

- Lecture slides (at least 30min before the lecture)

- **Book: YES**

- Data Mining Concepts and Techniques: Jiawei Han and Micheline Kamber

- **Readings:** Recommended

- (posted online)

Logistics: Communication

- **Blackboard:**

- <https://uoit.blackboard.com>
 - Labs and Project / Midterm / Final / ..

- **Slack**

- **For e-mailing us, always use:**

- jaroslaw.szlichta@uoit.ca

- **We will post course announcements to the website (make sure you check it regularly)**

Auditors are welcome to sit-in & audit the class

Work for the Course

- **Tutorials & Project: 30% (10% + 20%)**
 - 10% for 10 tutorials (1% each)
 - 20% (midterm report + final report)
 - Tutorials will start in the week of 18th of September
- **Midterm: 20%**
- **Participation & Presentation: 10%**
- **Final exam: 40%**
- **It's going to be fun and hard work. 😊**
- **Data Science graduate positions open!**

Equipment



Introductory Reading List

■ **Overviews:**

- 5 Ways Big Data Will Change the World
 - <http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz4GCbfTw8O>
- Reinventing Society in the Wake of Big Data
 - <https://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data>

■ **Success Stories**

- The Promise of Big Data
 - <https://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/>
- Big Data: The Management Revolution
 - <https://hbr.org/2012/10/big-data-the-management-revolution/ar>

Introductory Reading List

■ Skepticism

- Eight (No Nine!) Problems With Big Data
 - http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=1
- Big Data: Are We Making a Big Mistake?
 - <https://next.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0>

■ History

- A Very Short History of Big Data
 - <http://www.forbes.com/forbes/welcome/>