Introduction to Data Science

January 11, 2016

About this course

DATA 5000: Introduction to Data Science

Some highlights:

- Topics for data scientists
- R
- IBM Cognos Workspace, IBM SPSS Modeler, Watson Analytics
- VCL cloud
- Course projects

Evaluation

Course Project

- 10% Project proposal, due 25 January, 2016
- 10% Presentation outline, due 17 March, 2016
- 30% Presentation, last two classes 28 March and 4 April, 2016
- 50% Project paper, due April 11, 2016

Details will be discussed later today.

Contact information

Olga Baysal

Email: olga.baysal@carleton.ca

Office hours: By appointment or via Slack

Office: HP 5125D

Website: http://olgabaysal.com/teaching/winter16/data5000.html

Boyan Bejanov

Email: boyanbejanov@cmail.carleton.ca

Office hours: By appointment or via Slack

Office: none

Website: http://scs.carleton.ca/~boyanbejanov/data5000

What is Data Science?

Business efficiency: Wal-Mart

http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html

Business marketing: Target

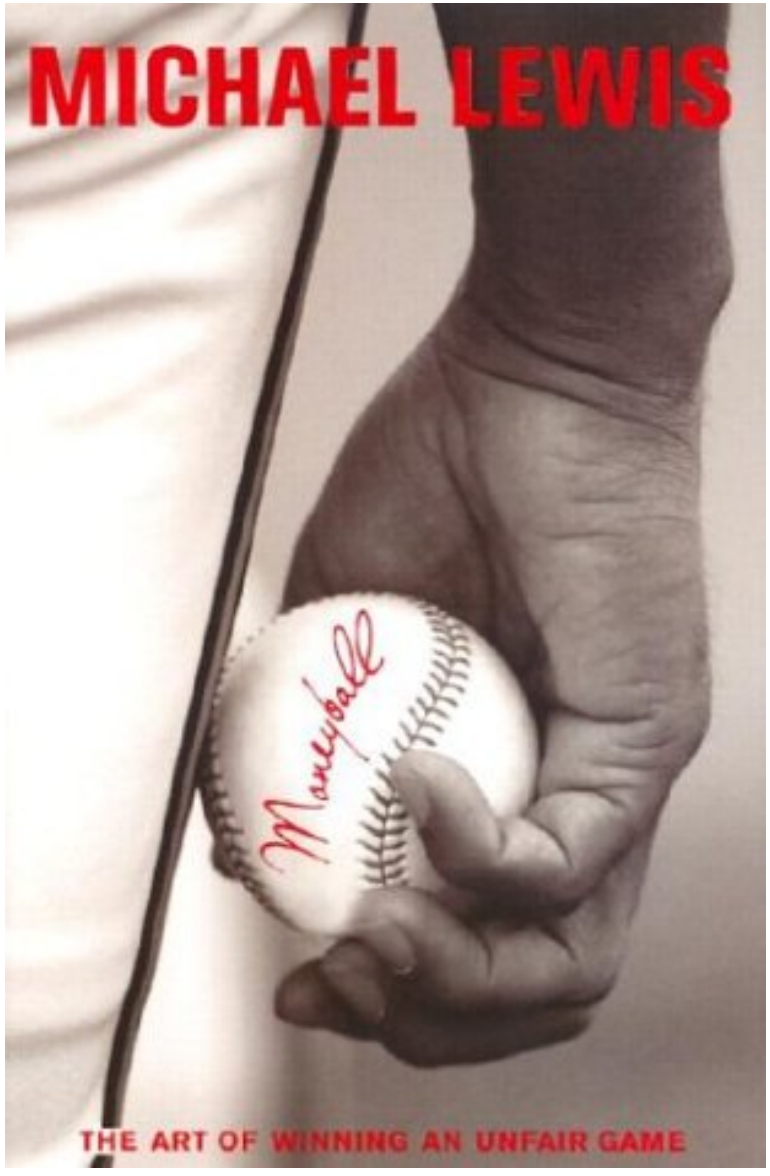
http://tinyurl.com/7jbntx3

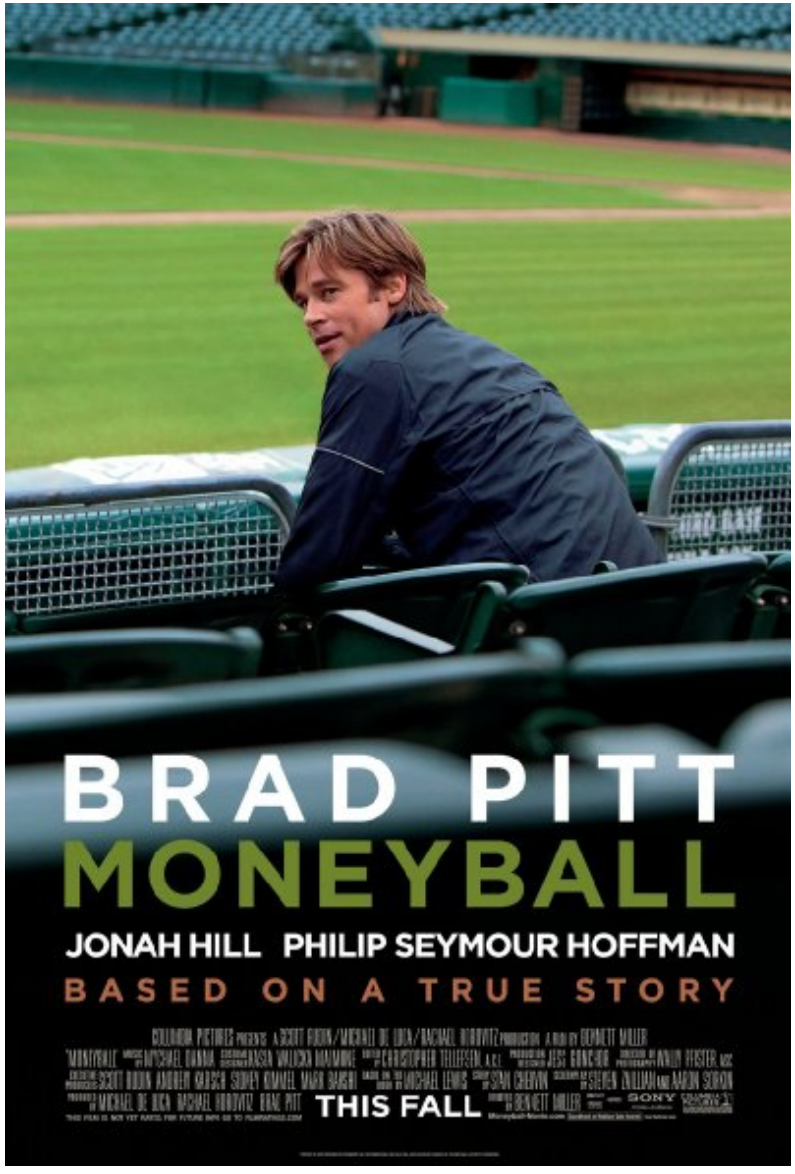
Recommendations: Netflix

- In October 2006 Netflix held a competition for the best algorithm to predict user ratings of movies.
- The winner must improve Netflix' own algorithm by at least 10%
- Award was given in September 2009.

`http://www2.research.att.com/~volinsky/netflix/bpc.html`

Sports analytics



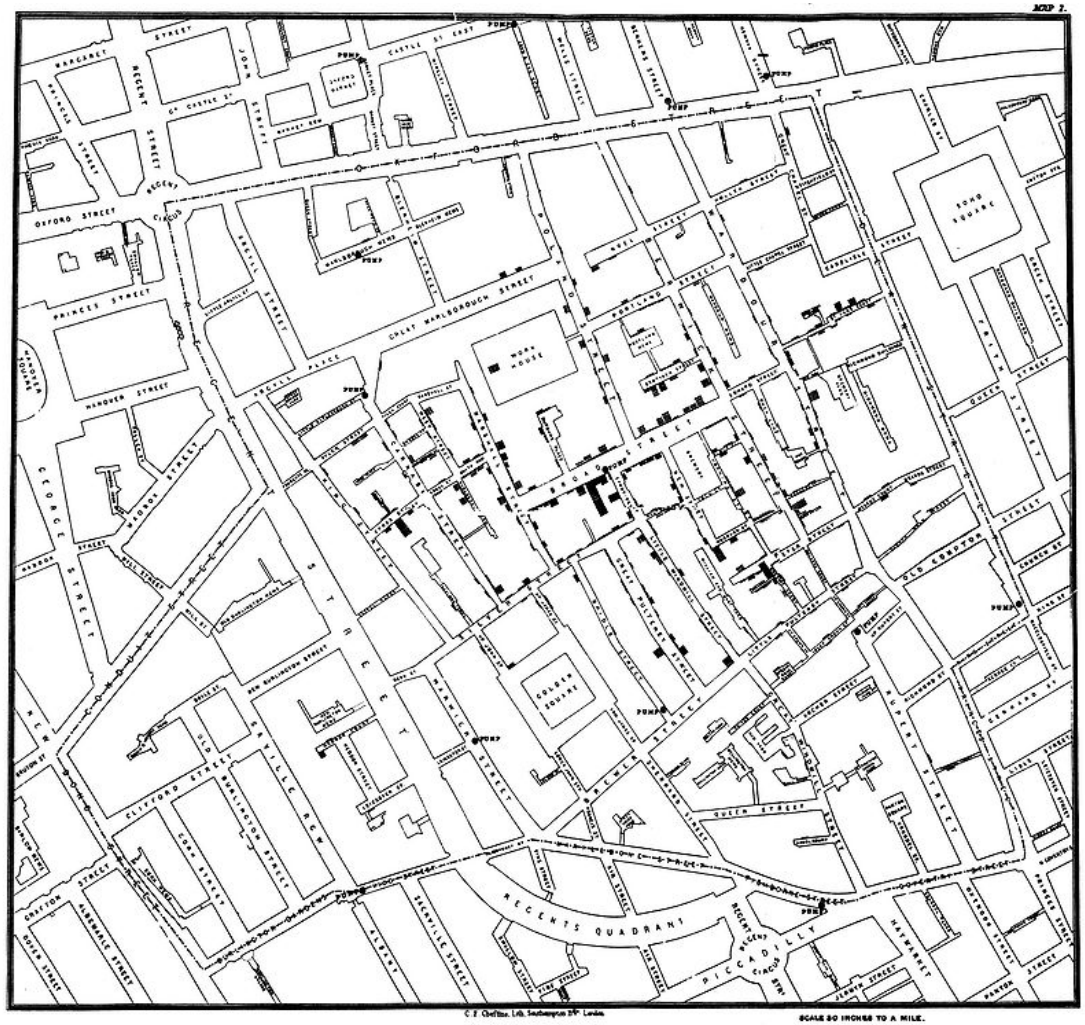


Many others

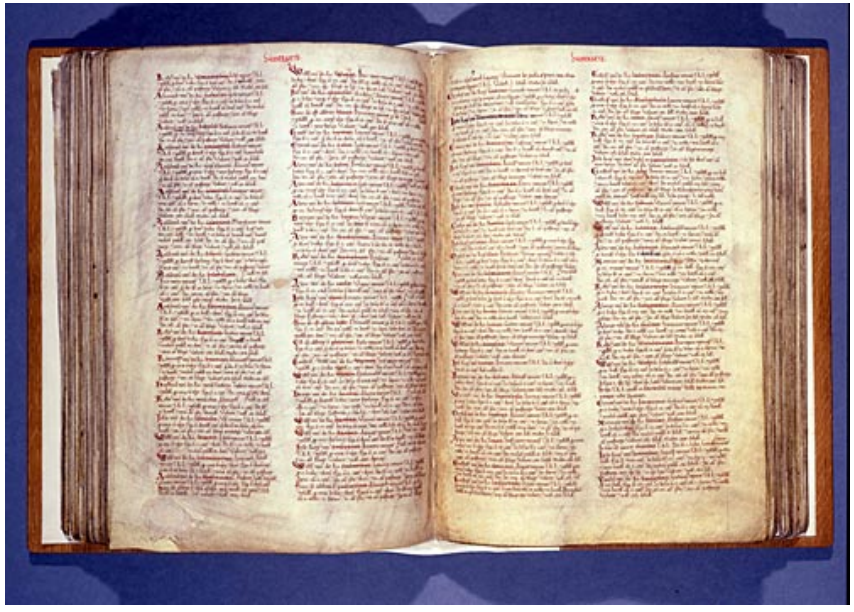
- Cities: `http://data.cityofchicago.org/`
- Physics: `http://particlefever.com/`
- Politics: `http://53eig.ht/1zPmuCD`
- Social networks
- Biology
- Medicine
- etc.

Cholera outbreak in London 1856

- Physician John Snow links the outbreak to a contaminated well by plotting number of cases on a map
- Started the science of epidemiology



The Winchester Roll of 1086



a.k.a. Domesday Book

- Commissioned in 1085 by William the Conqueror
- Record of the Great Survey of England
- Last used to settle dispute in court in the 1960s!

http://www.domesdaybook.co.uk/

Data in the 20-th century

What problems were solved?

- Engineering: design of machines
- Sciences: formulation of theories

How were problems solved?

- Empirically
- Theories
- Computation

Data in the 21-st century

How is today different?

- More data is available
- More data is digital
- More data is observed, rather than generated by a designed experiment

Data in the 21-st century

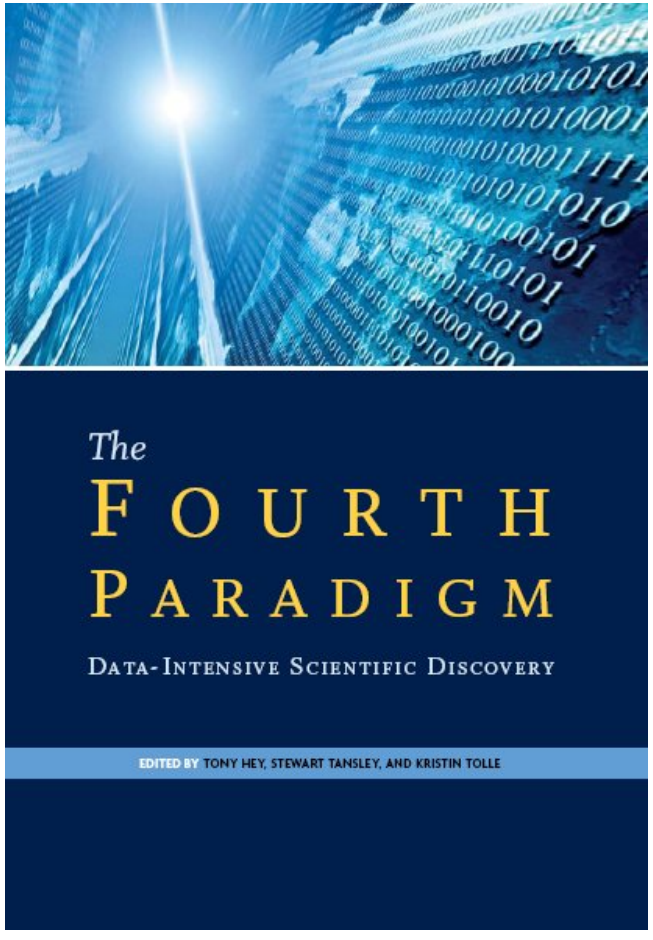
What problems are solved today?

- Spell checking
- Face recognition
- Sentiment analysis
- Optimal routing
- High-frequency trading algorithms
- just to name a few . . .

Data in the 21-st century

How are problems solved today?

- Empirically
- Theories
- Computation
- **Data exploration**



http://research.microsoft.com/en-us/collaboration/fourthparadigm/

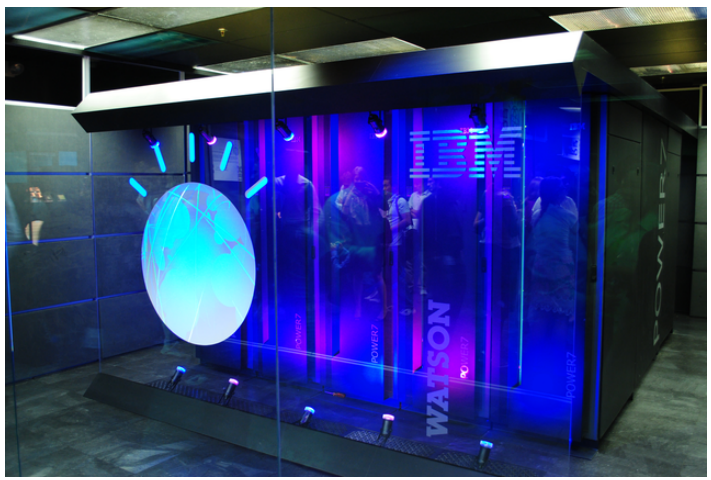
For example

Network security:

- 20-th century: based on rules and signatures
- 21-st century: data mining traffic logs, cf. `http://www.bro.org/`

Artificial Intelligence:

vs.



A good question

So, what is data science?

Who are the data scientists?

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

ANALYTICS

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Skills:

- make discoveries while swimming in data
- don't allow technical limitations to bog down solutions
- often fashion their own tools
- skilled in storytelling with data

Some data-driven companies:

- Google, Wal-Mart, Twitter, LinkedIn, Amazon

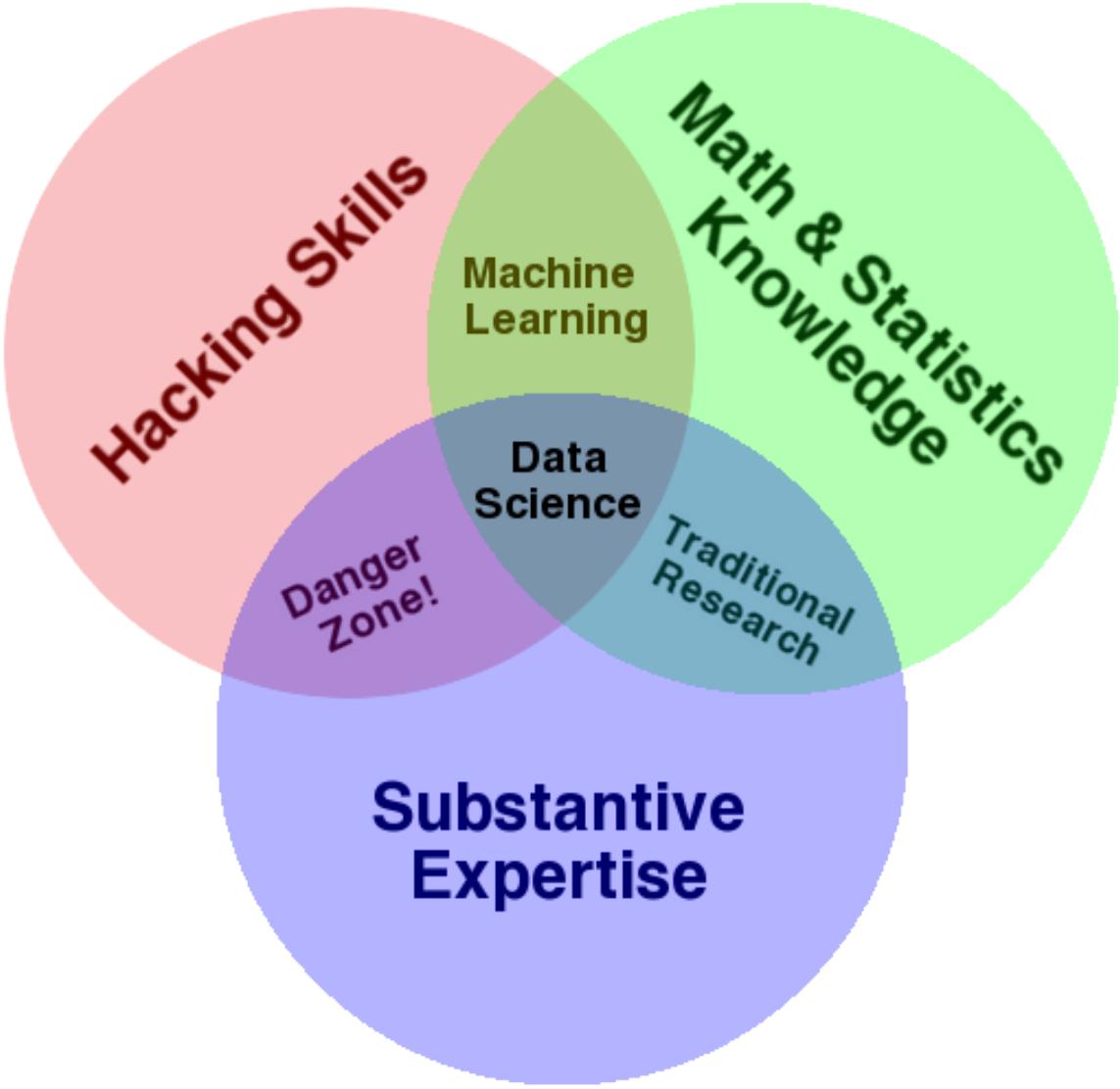
What data scientists do

- Ask a question
- Get relevant data
- Prepare data for analysis
  - outliers, missing values, incorrect values
- Explore data
  - understand the world as it is (was)
- Statistical model
  - estimate/train and validate model
  - predict what will (likely) happen
- Communicate results
  - tell a story
  - recommend

Data scientist skills

- Computer science
  - programming, hacking skills
- Statistics
  - probability, distributions, modelling
- Mathematics
  - linear algebra, calculus, optimization
- Domain expertise
  - storytelling, pose question, interpret result
- Communication
  - presentation, data visualization

Drew Conway's Venn diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Tentative course schedule

11 Jan	First class.
25 Jan	Project proposals due by end of day.
1 Feb	Cognos Workspace, TBC.
15 Feb	Reading week, no class
22 Feb	SPSS Modeler, TBC.
7 Mar	Watson Analytics, TBC.
	Presentation outlines due by March 17.
14, 21 Mar	Guest lectures.
28 Mar	Project presentations.
4 Apr	Project presentations, last class.
11 Apr	Project papers due.

Books

Note: These books are not required.

Books used for this course:

- *Doing Data Science*
  by Cathy O'Neil and Rachel Schutt
- *Data Mining And Business Analytics With R*
  by Johannes Ledolter
- *Data Science for Business*
  by Foster Provost and Tom Fawcett

Other good books:

- *An Introduction to Statistical Learning*
  by T. Hastie, R. Tibshirani et al.
- *The Elements of Statistical Learning*
  by T. Hastie, R. Tibshirani et al.

Projects

Teams of 2 - no individual projects, no larger groups. *No teams with all members from the same department!*

Email me your *team name* (optional), and *team members* by January 17, 2016 (before next class).

Project proposals are due January 25, 2016. Proposal should describe your *question*, the *dataset* and *an idea* of what you'll do with it. *Keep it short.*

Some project ideas and datasets are listed on the course website: `http://olgabaysal.com/teaching/winter16/data5000.html#datasets`.