

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

The New York Times > x
www.nytimes.com/2004/11/14/business/yourmoney/14wal.html

The New York Times **Business** [Great Getaways - FREE luxury travel deals](#)

NYTimes: [Home](#) - [Site Index](#) - [Archive](#) - [Help](#) Welcome, - [Member Center](#) - [Log Out](#)

Go to a Section Quotes: Site Search:

[NYTimes.com](#) > [Business](#) > [Your Money](#)

What Wal-Mart Knows About Customers' Habits

By **CONSTANCE L. HAYS**
Published: November 14, 2004

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at [Wal-Mart Stores](#) decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.



Illustration by The New York Times

ARTICLE TOOLS

- [E-Mail This Article](#)
- [Printer-Friendly Format](#)
- [Most E-Mailed Articles](#)
- [Reprints & Permissions](#)

Advertisement



AMERICAN EXPRESS

Take Charge.

- Pay **NO ANNUAL FEE** for the first year.
- Earn a **Welcome Bonus of 30,000 points.**

Built for Business Owners™

Let's Talk →

*Conditions apply

NYTIMES.COM ACCESS
Now at select **hotels and airport lounges** around the world.

[CHECK OUT OUR EXPANDING LIST >](#)

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



Business marketing: Target

The image is a screenshot of a web browser displaying a Forbes article. The browser's address bar shows the URL: www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant. The Forbes logo is visible in the top left, and navigation links for 'New Posts', 'Most Popular', 'Lists', 'Video', and '2 Free Issues of Forbes' are in the top right. The article is by Kashmir Hill, a Forbes Staff member, with a 'FOLLOW' button. The article is categorized as 'TECH' and was published on 2/16/2012 at 11:02AM, with 2,623,696 views. The main headline is 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did'. Below the headline are options to 'Share', 'Comment Now', and 'Follow Comments'. The article text begins with 'Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy.' To the right of the text is a large red Target bullseye logo with the word 'TARGET' in red below it. Further right is a black advertisement for Symantec with the text 'INTELLIGENT SECURITY SOLUTIONS from Symantec' and a 'Learn more' link. A Symantec logo is also visible in the top right of the article area.

<http://tinyurl.com/7jbntx3>



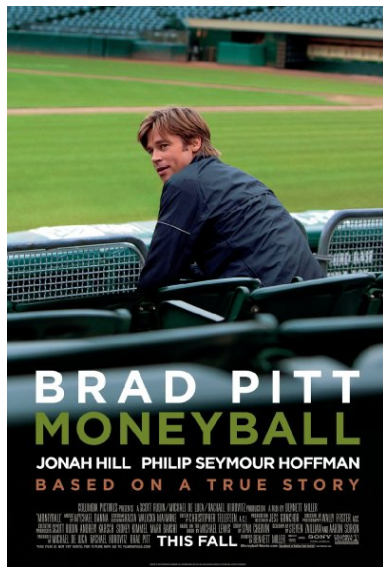
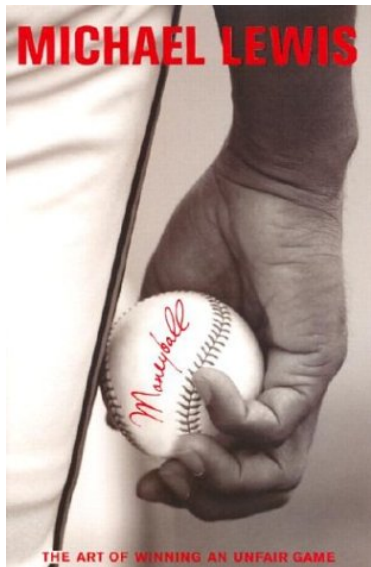
Recommendations: Netflix

- 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



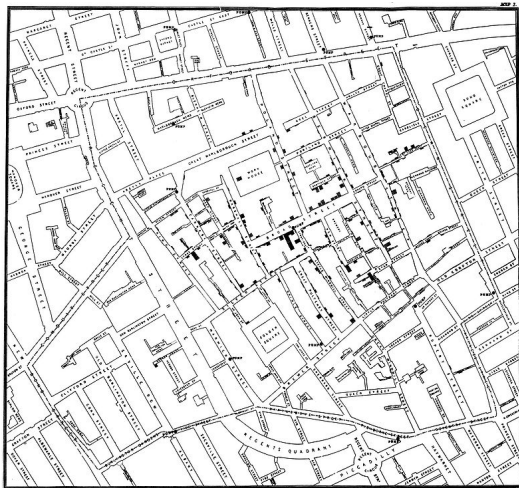
Many others

- **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?

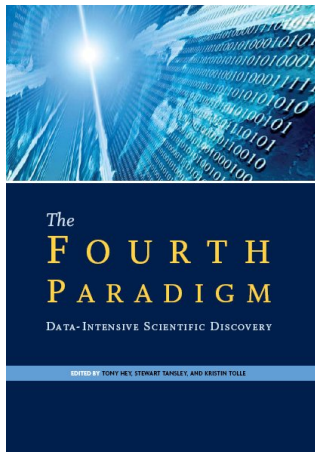
- Spell checking
- Face recognition
- Sentiment analysis
- Optimal routing
- High-frequency trading algorithms
- just to name a few . . .



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:

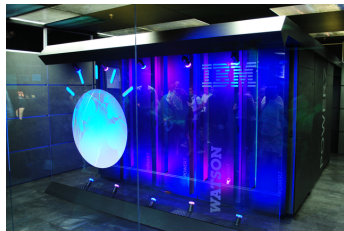
- 20-th century: based on rules and signatures
- 21-st century: data mining traffic logs, cf.

<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

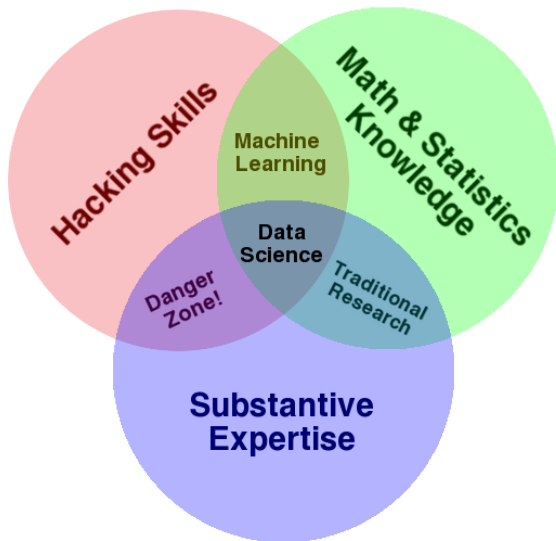


Data scientist skills

- 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



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>.

