Introduction to Data Science
Week 1, Lecture 1

Jeff Hammerbacher
January 17, 2012
Lecture Outline

- 1. Course Content
- 2. Course Logistics
- 3. Course Motivations
- 4. Doing Data Science
- 5. Homework
1. Course Content
Course Content

Target Setting

- Fast-growing organization looking to build a “data team”
- Data-driven decisions
- Data-intensive products and services
Course Content

Not the Target Setting

▪ Mature organization with well-defined problem
▪ Academic or industrial research
▪ Statistical consulting
Course Content

What We'll Cover

▪ Data Preparation
▪ Data Presentation
▪ Data Products
▪ Observation
▪ Experimentation
▪ Final Project
Course Content

What We Won’t Cover

▪ Probability and Statistics
▪ Machine Learning
▪ Data Mining
▪ Knowledge Discovery in Databases
▪ Text Mining/NLP
▪ Artificial Intelligence
▪ Big Data
▪ NoSQL
2. Course Logistics
Course Logistics

- Course Website: http://datascienc.es
- Instructors: Jeff Hammerbacher, Mike Franklin
- Course Times: 12:30 pm - 2:00 pm, Tuesday and Thursday (240 Bechtel)
- Office Hours: TBD
- Coursekit
  - URL: http://coursekit.com/app#course/194-16.hammerbacher
  - Invite code: JWGXU6
Course Logistics

Prerequisites

- Data manipulation: Python, R, databases
- Web programming: HTML, JavaScript, CSS
- Probability and Statistics
3. Course Motivations
Course Motivations

Personal
Course Motivations

Personal
Course Motivations

Personal
Course Motivations

Personal
Course Motivations

Personal

“Information Platforms and the Rise of the Data Scientist”
Course Motivations

Putting Data to Work

1935: “The Design of Experiments”
Course Motivations
Putting Data to Work

1939: “Quality Control”
Course Motivations
Putting Data to Work

1955: “Artificial Intelligence”
Course Motivations
Putting Data to Work

1958: “A Business Intelligence System”
Course Motivations
Putting Data to Work

1962: “Data Analysis”
Course Motivations
Putting Data to Work

1977: “Exploratory Data Analysis”
Course Motivations
Putting Data to Work

1989: “Business Intelligence”
Course Motivations
Putting Data to Work

1995: TDWI

THE DATA WAREHOUSING INSTITUTE
Course Motivations
Putting Data to Work

1996: “From Data Mining to Knowledge Discovery in Databases”
Course Motivations
Putting Data to Work

1997: “Machine Learning”

Thursday, January 19, 12
Course Motivations
The Emergence of Data Science

1994: “Managing Gigabytes”
Course Motivations
The Emergence of Data Science

1996: Google
Course Motivations
The Emergence of Data Science

Course Motivations
The Emergence of Data Science

2007: “The Case for DISC”
Course Motivations
The Emergence of Data Science

Course Motivations

The Emergence of Data Science

2009: “The Unreasonable Effectiveness of Data”
Course Motivations
The Emergence of Data Science

2007: “Competing on Analytics”
Course Motivations
The Emergence of Data Science

2007: “Super Crunchers”

“Groundbreaking... Not only is it fun to read, it just may change the way you think.”
—STEVEN D. LEVITT, coauthor of Freakonomics

SUPER CRUNCHERS
IAN Ayres
Course Motivations
The Emergence of Data Science

2007: “The Coming Exaflood”

WSJ.
Course Motivations
The Emergence of Data Science

2008: “The End of Science”
Course Motivations

The Emergence of Data Science

2010: “The Data Deluge”
Course Motivations
The Emergence of Data Science

2010: “What is Data Science?”
Course Motivations
The Emergence of Data Science

2011: “Data scientist: The hot new gig in tech”
4. Doing Data Science
Doing Data Science

My Process Model

• 1. Identify problem
• 2. Instrument data sources
• 3. Collect data
• 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
• 5. Build model
• 6. Evaluate model
• 7. Communicate results
Doing Data Science
Peter Huber

- 1. Inspection
- 2. Error checking
- 3. Modification
- 4. Comparison
- 5. Modeling and model fitting
- 6. Simulation
- 7. What-if analyses
- 8. Interpretation
- 9. Presentation of conclusions
Doing Data Science

Ben Fry

- 1. Acquire
- 2. Parse
- 3. Filter
- 4. Mine
- 5. Represent
- 6. Refine
- 7. Interact
Doing Data Science

Dataists

- 1. Obtain
- 2. Scrub
- 3. Explore
- 4. Model
- 5. Interpret
Doing Data Science
Colin Mallows

- 0. Identify data to collect and its relevance to your problem
- 1. Statistical specification of the problem
- 2. Method selection
- 3. Analysis of method
- 4. Interpret results for non-statisticians
Doing Data Science
Jim Gray

- 1. Capture
- 2. Curate
- 3. Communicate
Doing Data Science
Ted Johnson

- 1. Assemble an accurate and relevant data set
- 2. Choose the appropriate algorithm
5. Homework
Homework

- Email me
  - Name
  - Year
  - Major
  - Relevant course work
  - Relevant work experience
  - Why this course will be useful to you
  - Your dream data set
- Sign up for Coursekit
  
  Post a link to the most interesting example of Data Science that you can find