15-388/688 - Practical Data Science: Introduction

J. Zico Kolter
Carnegie Mellon University

Fall 2016

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

Some possible definitions

Data science is the application of computational and statistical techniques to address or gain insight into some problem in the real world

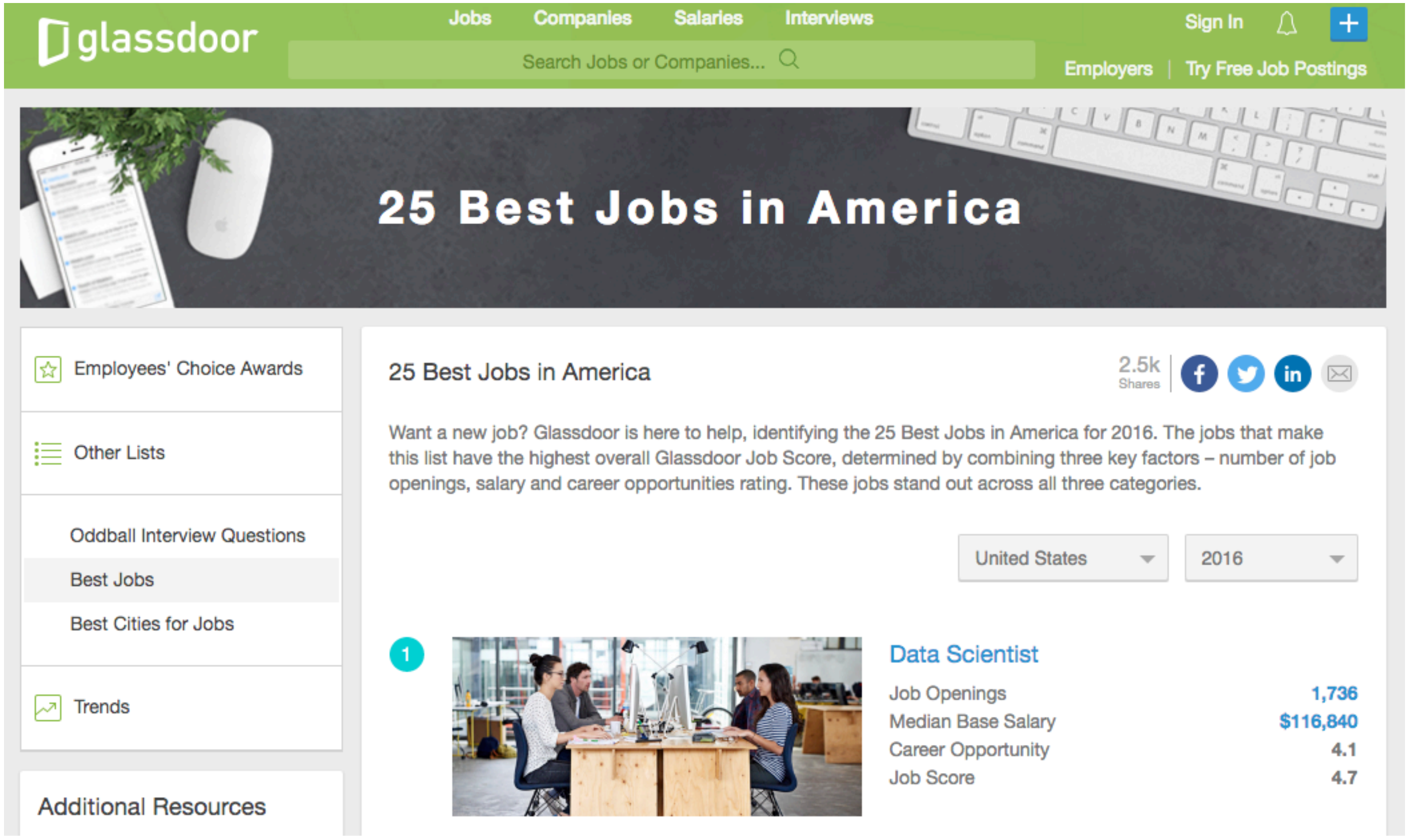
Some possible definitions

Data science is the application of **computational** and **statistical** techniques to address or gain insight into some problem in the **real world**

Some possible definitions

Data science = statistics +
data processing +
machine learning +
scientific inquiry +
visualization +
business analytics +
big data + …

Data science is the best job in America

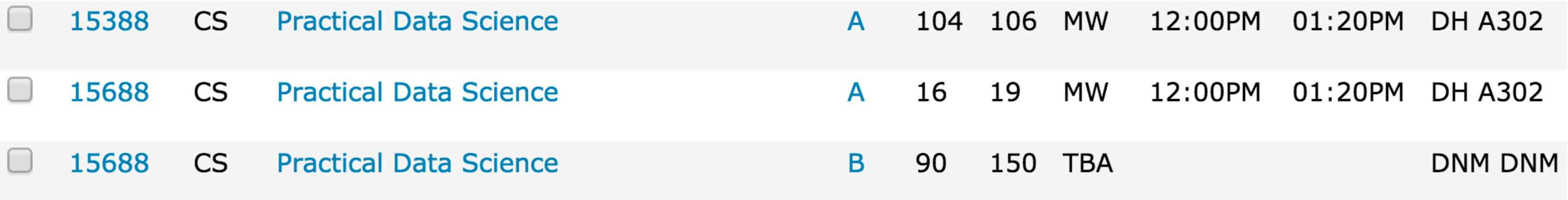


25 Best Jobs in America

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

1 Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7



□	15388	CS	Practical Data Science	A	104	106	MW	12:00PM	01:20PM	DH A302
□	15688	CS	Practical Data Science	A	16	19	MW	12:00PM	01:20PM	DH A302
□	15688	CS	Practical Data Science	B	90	150	TBA			DNM DNM

(+85 on waitlists)

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

Data science is not machine learning

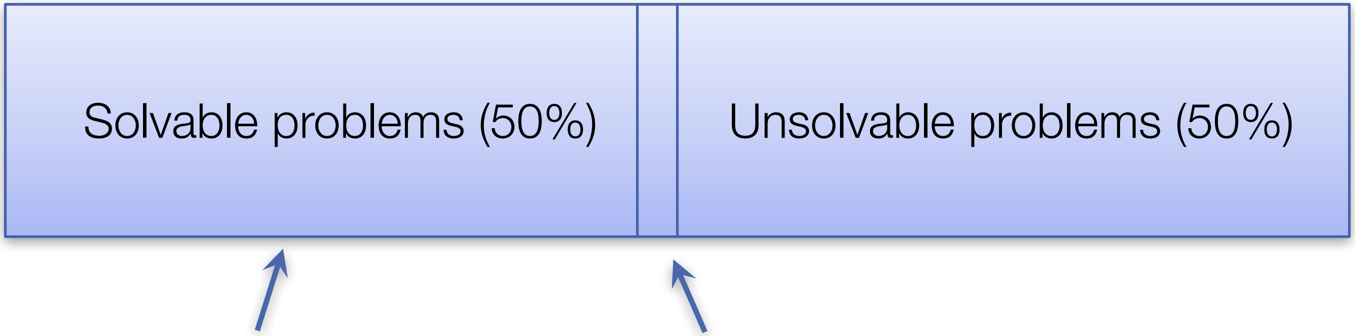
Machine learning involves computation and statistics, but has traditionally been not very concerned about answer *scientific questions*

Machine learning has a heavy focus on fancy algorithms…

... but sometimes the best way to solve a problem is just by visualizing the data, for instance

Data science is not machine learning

Data problems we would like to solve



Solvable problems (50%)

Unsolvable problems (50%)

Problems that can use “simple” machine learning (49%)

Problems that need, e.g. deep learning (1%)

Data science is not statistics

“Analyzing data computationally, to understand some phenomenon in the real world, you say? … that sounds an awful lot like statistics”

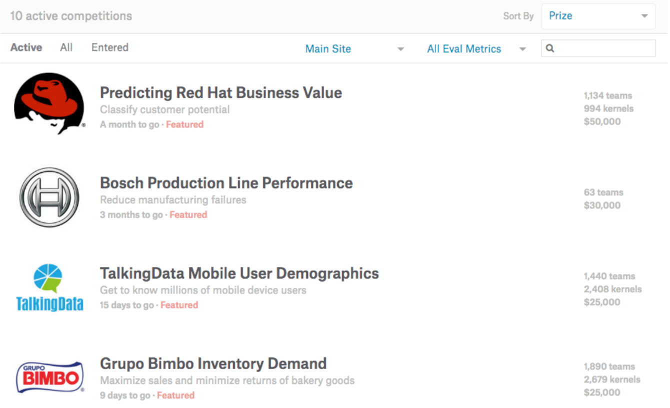
Statistics (at least the academic type) has evolved a lot more along the mathematical/theoretical frontier

Not many statistics courses have a lecture on e.g. web scraping, or a lot of data processing more generally

Plus, statisticians use R, while data scientists use Python ... clearly these are completely different fields

Data science is not data science competitions

kaggle™



Predicting Red Hat Business Value
Classify customer potential
A month to go · Featured
1,134 teams
994 kernels
$50,000

Bosch Production Line Performance
Reduce manufacturing failures
3 months to go · Featured
63 teams
$30,000

TalkingData Mobile User Demographics
Get to know millions of mobile device users
15 days to go · Featured
1,440 teams
2,408 kernels
$25,000

Grupo Bimbo Inventory Demand
Maximize sales and minimize returns of bakery goods
9 days to go · Featured
1,890 teams
2,679 kernels
$25,000

Data science competitions like Kaggle ask you to optimize a metric on a fixed data set

This may or may not ultimately solve the desired business/scientific problem

Data science is the iterative cycle of *designing* a concrete problem, building an algorithm to solve it (or determining that this is not possible), and evaluating what insights this provides for the real underlying question

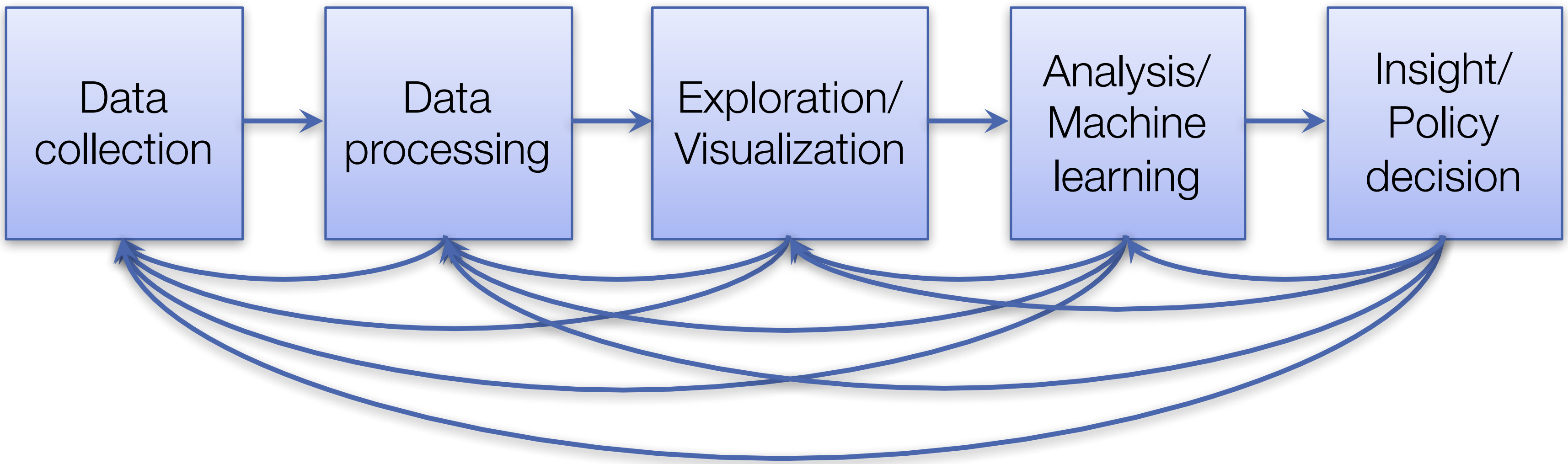
Data science is not big data

Sometimes, in order to truly understand and answer your question, you need massive amounts of data

But sometimes you don’t

Don’t create more work for yourself than you need to

Back to what data science is



Outline

What is data science?

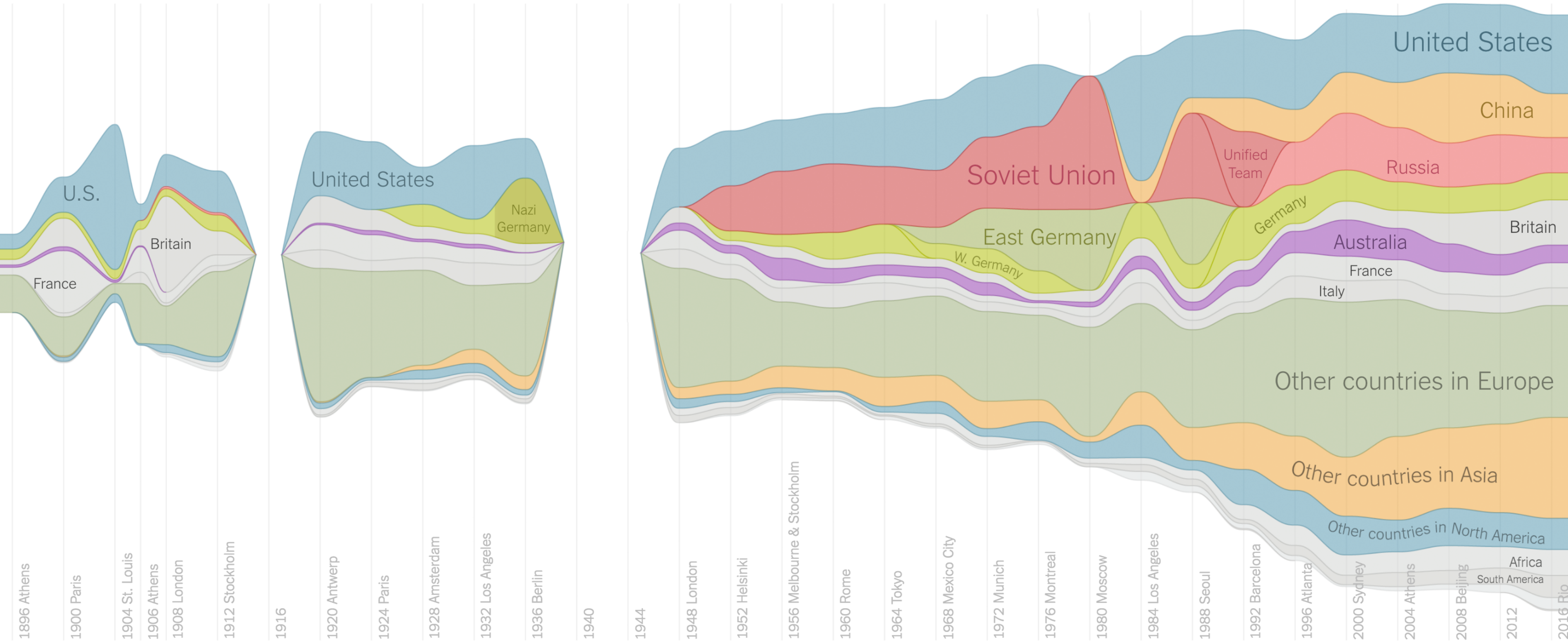
What is data science not?

(A few) data science examples

Course objectives and topics

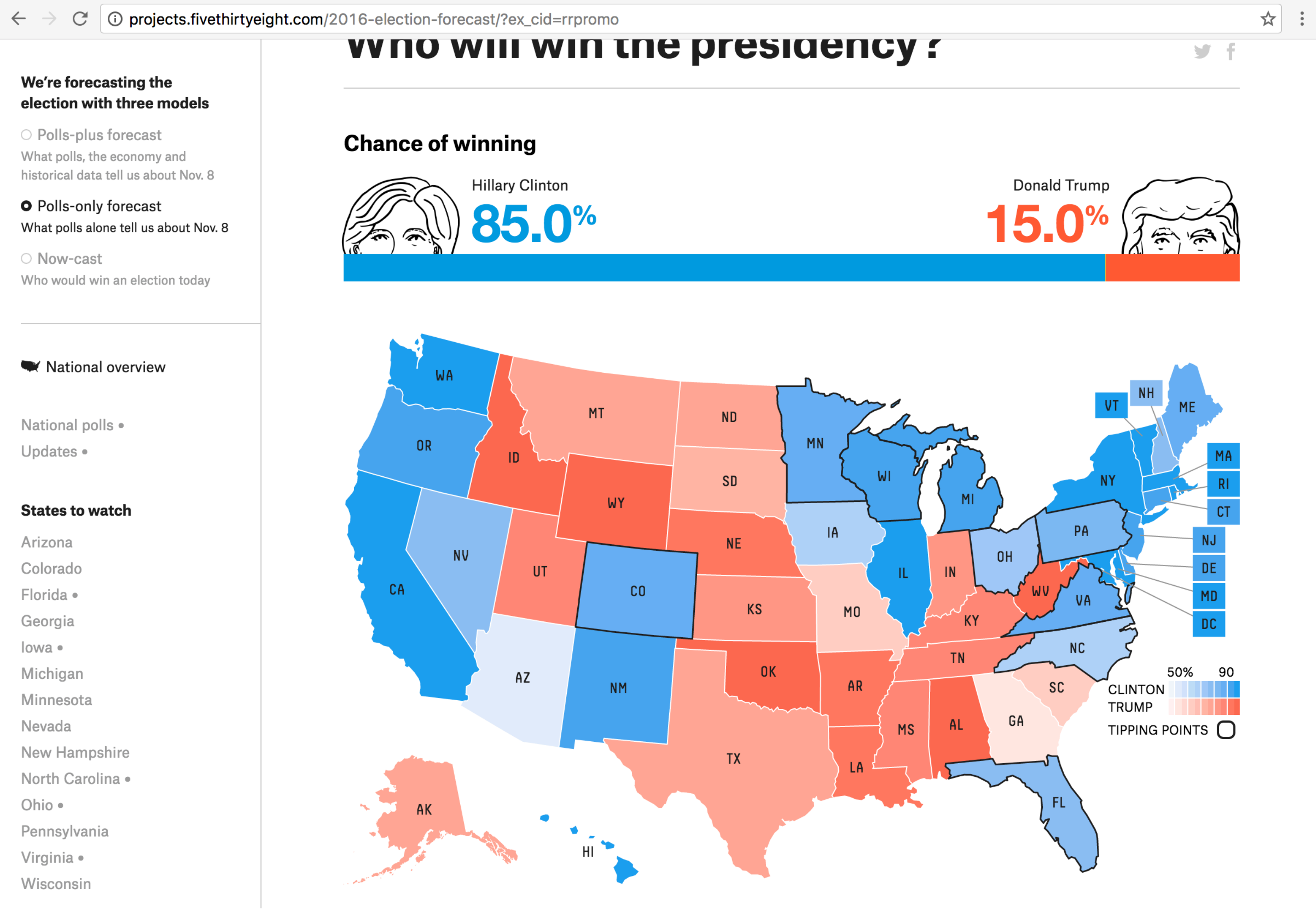
Course logistics

Olympic medals



http://www.nytimes.com/interactive/2016/08/08/sports/olympics/history-olympic-dominance-charts.html

FiveThirtyEight



projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo

We're forecasting the election with three models

○ Polls-plus forecast
What polls, the economy and historical data tell us about Nov. 8

● Polls-only forecast
What polls alone tell us about Nov. 8

○ Now-cast
Who would win an election today

National overview

National polls •
Updates •

States to watch

Arizona
Colorado
Florida •
Georgia
Iowa •
Michigan
Minnesota
Nevada
New Hampshire
North Carolina •
Ohio •
Pennsylvania
Virginia •
Wisconsin

Who will win the presidency?

Chance of winning

Hillary Clinton 85.0%

Donald Trump 15.0%

Poverty Mapping

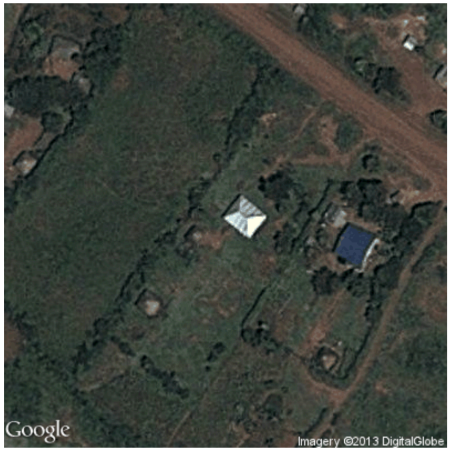


Figure 2: Example of metal roof in center of satellite image.

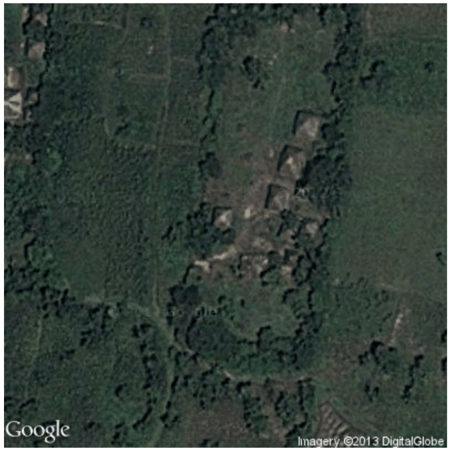
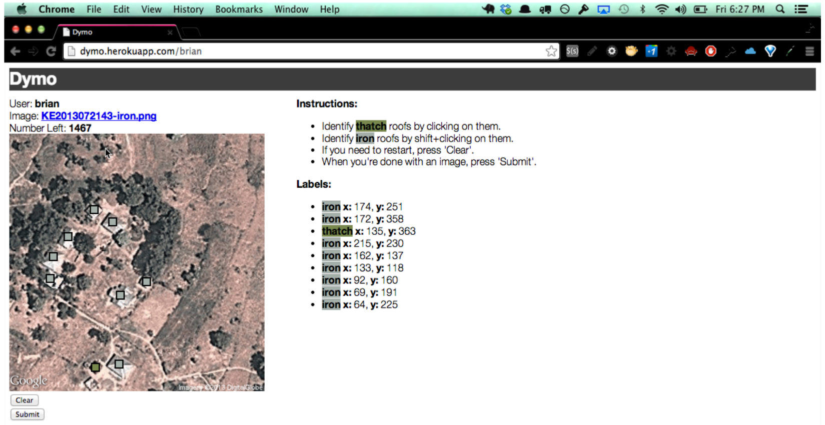


Figure 3: Example of thatched roof in center of satellite image.

Abelson, Varshney, and Sun. "Targeting Direct Cash Transfers to the Extremely Poor," 2012



Dymo

User: **brian**
Image: KE2013072143-iron.png
Number Left: **1467**

Instructions:

- Identify **thatch** roofs by clicking on them.
- Identify **iron** roofs by shift+clicking on them.
- If you need to restart, press 'Clear'.
- When you're done with an image, press 'Submit'.

Labels:

- **iron x:** 174, **y:** 251
- **iron x:** 172, **y:** 358
- **thatch x:** 135, **y:** 363
- **iron x:** 215, **y:** 230
- **iron x:** 162, **y:** 137
- **iron x:** 133, **y:** 118
- **iron x:** 92, **y:** 160
- **iron x:** 69, **y:** 191
- **iron x:** 64, **y:** 225

Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.

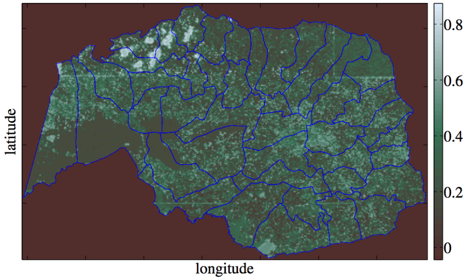


Figure 11: Heat map of proportion of roofs that are metal in the region of interest.

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

Learning objectives of this course

After taking this course, you should…

… understand the full data science pipeline, and be familiar with programming tools to accomplish the different portions

... be able to collect data from unstructured sources and store it using appropriate structure such as relational databases, graphs, matrices, etc

... know to explore and visualize your data

... be able to analyze your data rigorously using a variety of statistical and machine learning approaches

Topics covered (subject to change)

Data collection and management: relational data, matrices and vectors, graphs and networks, free text processing, geographical data

Statistical modeling and machine learning: linear and nonlinear classification and regression, regularization, data cleaning, hypothesis testing, kernel methods and SVMs, boosting, clustering, dimensionality reduction, recommender systems, deep learning, probabilistic models, scalable ML

Visualization: basic visualization and data exploration, data presentation and interactivity, high dimensional visualization

Philosophy: tools and deeper understand

Most of the techniques we will teach in this course have mature tools that you will likely use in practice

But, the philosophy of this course is that you will use these tools most effectively when you understand what is going on under the hood

This course will teach you some of the more common tools, but (especially in 15-688 problem sets), you will also need to implement some of the underlying methods

Example: we'll teach you how to run machine learning algorithms using scikit-learn library, but you'll also need to implement many of the algorithms yourself

Differences between 15-388/688 and XX

There are many courses that cover similar or related material (10-601, 10-701, 11-663, 05-839, 36-402, etc)

In general, this course puts a high emphasis on exploring and analyzing real (unprepared) data, managing the entire data science pipeline

Compared to other machine learning or statistics courses, there is relatively little theory, higher emphasis on implementation and use on practical data sets

Recommended background

The only formal prerequisite for this course is an intro to programming (if you have taken one at another university, this is fine)

We recommend that students have **experience with Python**, ideally some background in **probability and statistics, and linear algebra**

If you don’t have background in these areas, you may still sign up, but be aware that you will probably need to learn some of these items as the class goes on (we will be providing pointers to references)

General rule of thumb: If the homework seems hard, but you have ideas about how to proceed, you probably have the right level of background; if the homework seems hard and you have no idea how to proceed, this may be the wrong course

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

Instructors

Zico Kolter

Eric Wong

Dhivya Eswaran

Jonathan Dinu

Course materials and discussion

All course material (slides, lecture videos, assignments, any supplemental notes or documentation), is available on the course webpage

http://www.datasciencecourse.org

Slides posted one hour before class, videos up ~2-3 hours after

Course discussion will take place on the Piazza Forum (15-688)

http://www.piazza.com

You ***must*** sign up for the Piazza forum ***with your andrew email*** by the end of the first week of class

15-388 vs. 15-688

Two versions of the course: 15-388 (undergrad, 9 unit), 15-688 (graduate, 12 unit)

Courses are identical (same lectures, assignments, etc) except that 15-688 problem sets have an additional question per assignment, usually requiring that students implement some advanced technique

Undergraduates ***may take 15-688*** for 12 units, but please wait until enrollment shakes out (for now, just start doing the 15-688 questions on the homeworks)

Course waitlist and DNM section

We currently have many more students enrolled than available space

To allows in as many people as possible, we added Section B, a DNM (does not meet) section to 15-688, courses are identical except that lectures are online

Will I get off the waitlist?

15-388: Yes
15-688-A: Probably not
15-688-B: Yes

First time course

This course is being taught for the first time

We hope you will learn a lot, and have fun, and we are doing our best to develop content and assignments that teaches you what we think are the most important concepts in data science

But, please be aware that there *are* going to be glitches (typos in the slides, bugs in the homework or autograders)

What we do promise is that we'll be very responsive and issue notices of any major fixes on Piazza

Grading

Grading breakdown is posted on the web site (updated):

55% homework
15% tutorial
25% class project
5% class participation

Class participation is judged based upon participation in the Piazza forum (you will get full credit if you post at least one standard deviation below the mean number of posts/student)

Final grades are assigned on a curve (separate for 15-388 and 15-688 versions)

Homeworks

One homework assignment every two weeks: released on Wednesdays by midnight, due the Wednesday two weeks later at midnight

We may miss this deadline sometimes (we are sorry in advance, we will of course also extend the due date)

Work will be largely (solely?) about ***writing code*** to solve problems

Homeworks are are in the form of Jupyter notebooks, ***solutions autograded by autolab***

https://autolab.andrew.cmu.edu/
(not ~~https://autolab.cs.cmu.edu/~~ !)

Autograding

The meta-goal for this course is to have a *scalable* introduction to data science (~300 students enrolled and on the waitlist, 3 TAs)

We believe that the current best way to achieve scalability is through heavy use of autograding

But, it’s also not perfect, so the reality is that there are some components of the assignments that we don’t evaluate quantitatively

This presents an additional problem for data science, where part of the process is developing scientific conclusions from the data (this is what the class project is for)

Late days

Assignments are due at 11:59pm (midnight) on Wednesdays

You have ***6 late days*** to use over the course of the semester

Each assignment can use a maximum of ***3 late days*** (midnight Saturday)

You cannot use late days for final project submission

Tutorial

The best way to learn a subject is to teach it

In lieu of a midterm, students will design a mini-tutorial, in the form of a Jupyter notebook, on a subject of their choice (though we will also provide suggestions)

Your tutorial will be read by the instructors, but also by other students, and peer grading will factor in to your final grade on the tutorial

Class project

A major component of the class: goal is to take a real-world domain that you are interested in, and apply data science methodologies to gain insight into the domain

Work to be done in groups of 2-3 students

Final report will be a Jupyter Notebook working through the analysis of your data, including code and visual results

Also presented in a video presentation (in lieu of final)

Class projects *must* be focused on some real data problem (ideally one that you collect yourself), not an already-curated data set

Office ~~hours~~ lunches

TAs will hold “traditional” office hours, schedule to be posted on the course web page, but when you have questions, we much prefer that you post to the Piazza forum (lets the whole class see the question)

My own office hours have often been dominated by a few students who ask lots of questions about homeworks, and I would instead like the chance get to know everyone in the course

After class each day, I’ll go to lunch with a group of 4-5 students, to talk about the problems you’re interested in and how the course material can apply (***plus you get a free lunch***)

Signup page will be linked on piazza forum

Academic integrity and homeworks

You can discuss ideas and methodology for the homeworks or tutorial with other students in the course, but ***you must write your solutions completely independently***

We will be running automated code-checking tools to assess similar submissions or submissions that use code from other sources

You ***may*** use snippets of code from sources like Stack Overflow, as long as you cite these properly (put a comment above and below whatever portion of code is copied), but be reasonable

See CMU's academic integrity policy:
http://www.cmu.edu/academic-integrity/

Student well-being

CMU and courses like this one are stressful environments

In my experience, most academic integrity violations are the product of these environments and decisions made out of desperation

Please don’t let it get to this point (or potentially much worse)

Don’t sacrifice quality of life for this course: still make time to sleep, eat well, exercise

Up next

Next class: web scraping and data collection

First homework released on Wednesday, use it as a gauge to determine if the course is right for you