15-388/688 - Practical Data Science: Introduction

Pat Virtue
Carnegie Mellon University
Spring 2022

Slide credits: CMU AI, Zico Kolter

Quick logistics

Slides will be available, including inked version

Lecture recording will be available

Quick logistics

Asking questions in main Zoom room

- Use the hand-raise icon
- Prof. can’t really see the chat

Answering questions in main Zoom room

- Just unmute and answer
- Use the hand-raise icon

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 + …

Lists of Best Jobs

Ranker

vote on >> entertainment // music //

1. National Geographic Traveler (4,249 / 839)
2. Secret Agent (2,545 / 1,076)
3. Private Island Caretaker (3,510 / 1,242)
4. Car Tester (2,676 / 1,177)

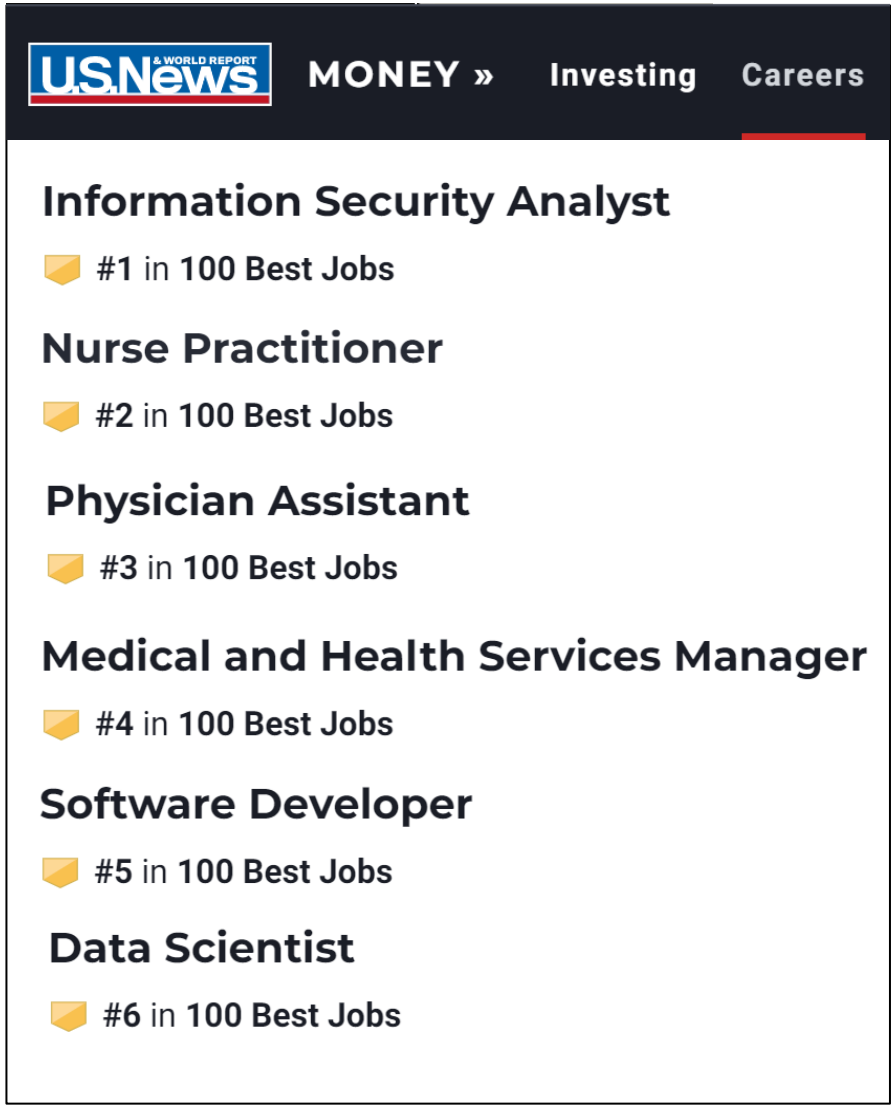
Coolest Jobs In The World!

1. Professional Sleeper

2. Netflix Tagger
3. Movie Critic
4. Water Slide Tester

https://www.ranker.com/crowdranked-list/best-jobs-in-the-world
https://www.panelplace.com/blogs/top-8-coolest-jobs-world

Lists of Best Jobs



https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm
https://money.usnews.com/careers/best-jobs/rankings/the-100-best-jobs

Outline

What is data science?

What is data science not?

(A few) data science examples

Course objectives and topics

Course logistics

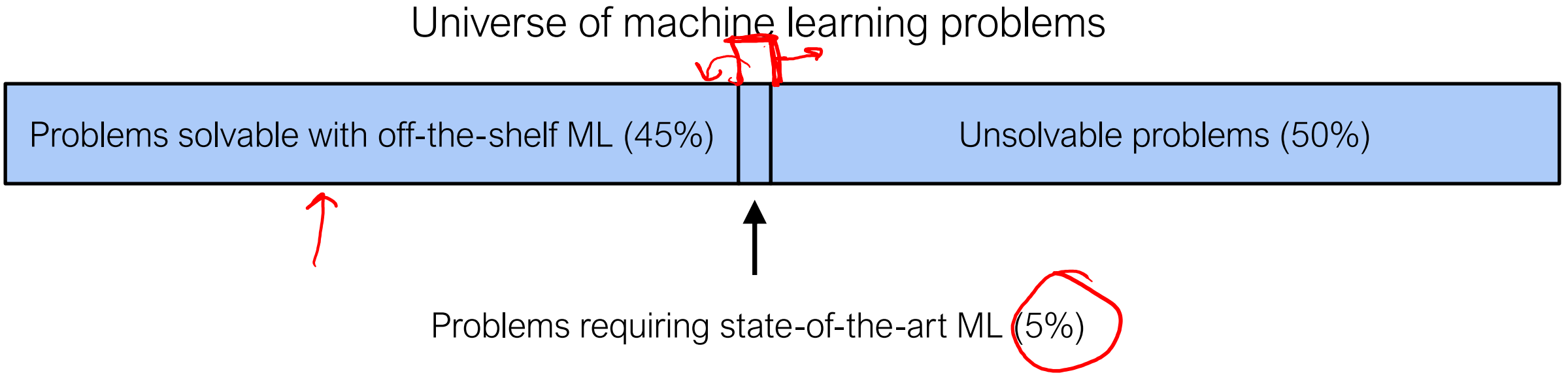
Data science is not machine learning

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

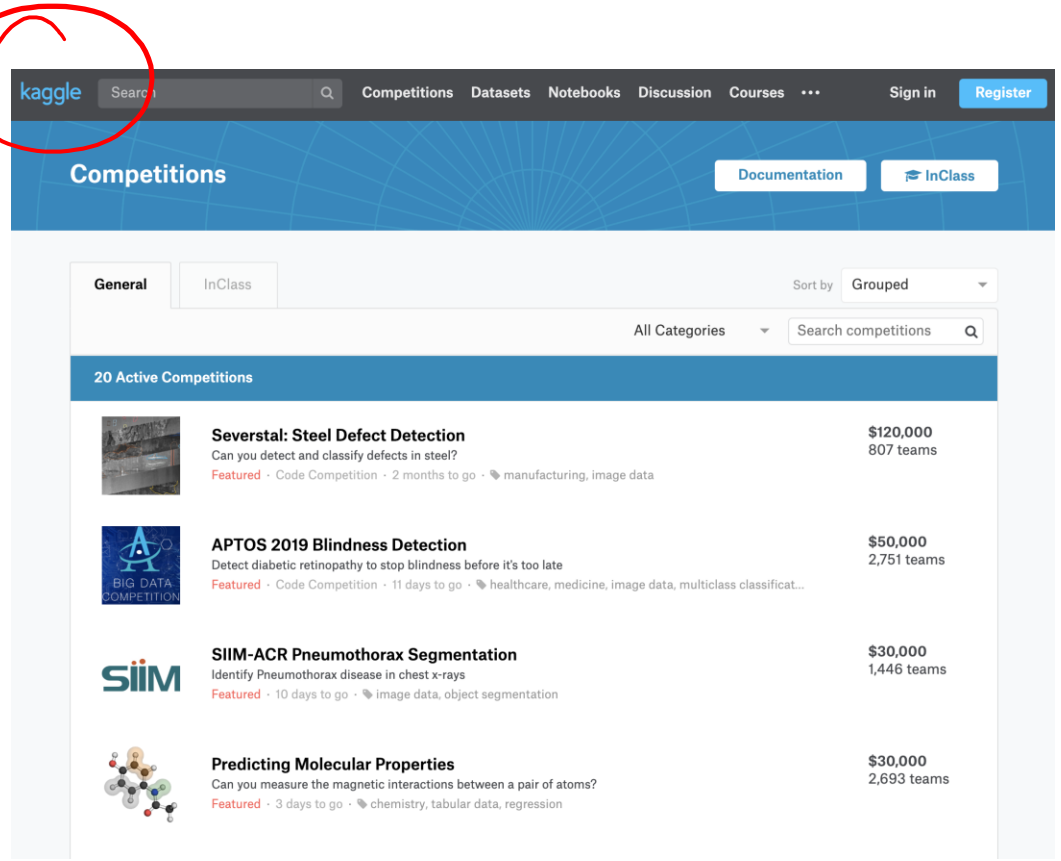
Machine learning has a heavy focus on fancy algorithms…

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

Data science is not machine learning



Data science is not machine learning competitions



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 machine learning competitions

"I was wondering if you could do that when it matters" -- *Braveheart*

https://youtu.be/Po3HbErxC-c

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

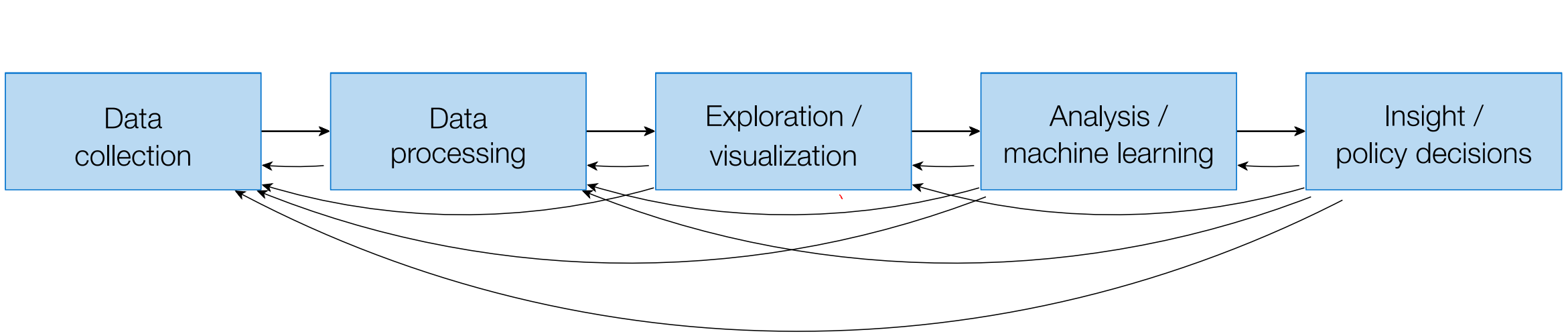
Data science is not big data

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

Gendered language in professor reviews

Gendered Language in Teacher Reviews

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

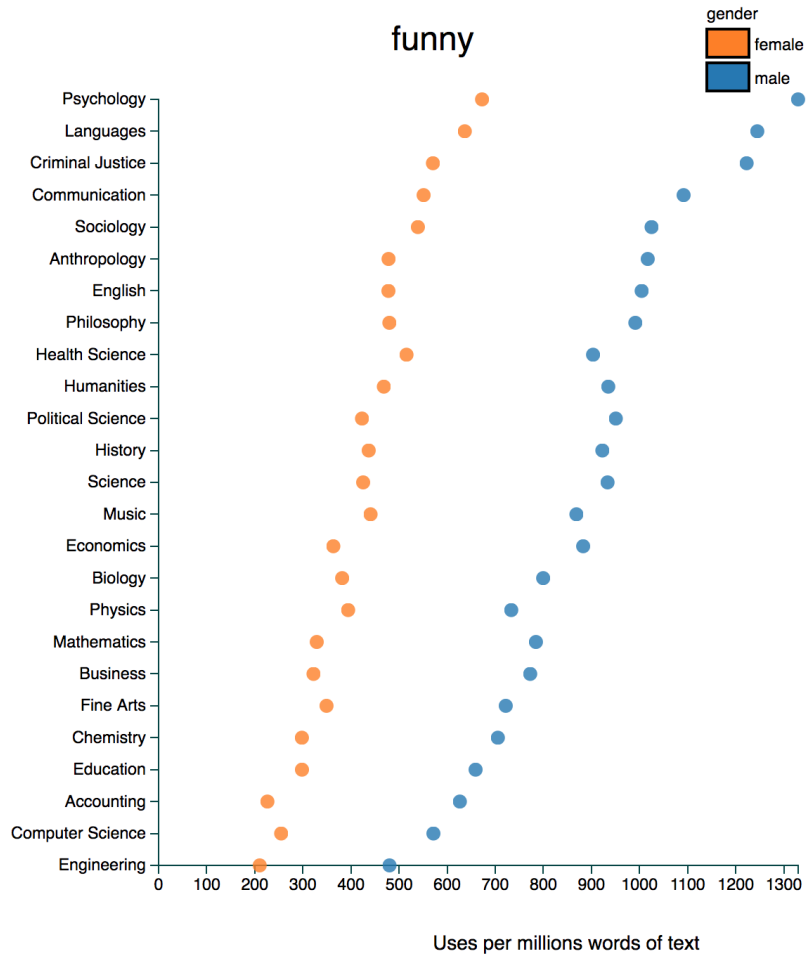
You can enter any other word (or two-word phrase) into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). You can also limit to just negative or positive reviews (based on the numeric ratings on the site). For some more background, see here.

Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

Search term(s) (case-insensitive):
use commas to aggregate multiple terms

funny

All ratings | Only positive | Only negative



http://benschmidt.org/profGender/

Obligatory quote

The greatest value of a picture is when it forces us to notice what we never expected to see.

-John Tukey

FiveThirtyEight

ELECTION 2018
FiveThirtyEight

House forecast | Senate | Governor | Midterms coverage | More politics | abcNEWS

Search for a race or candidate

Search

How do you like your House forecast?

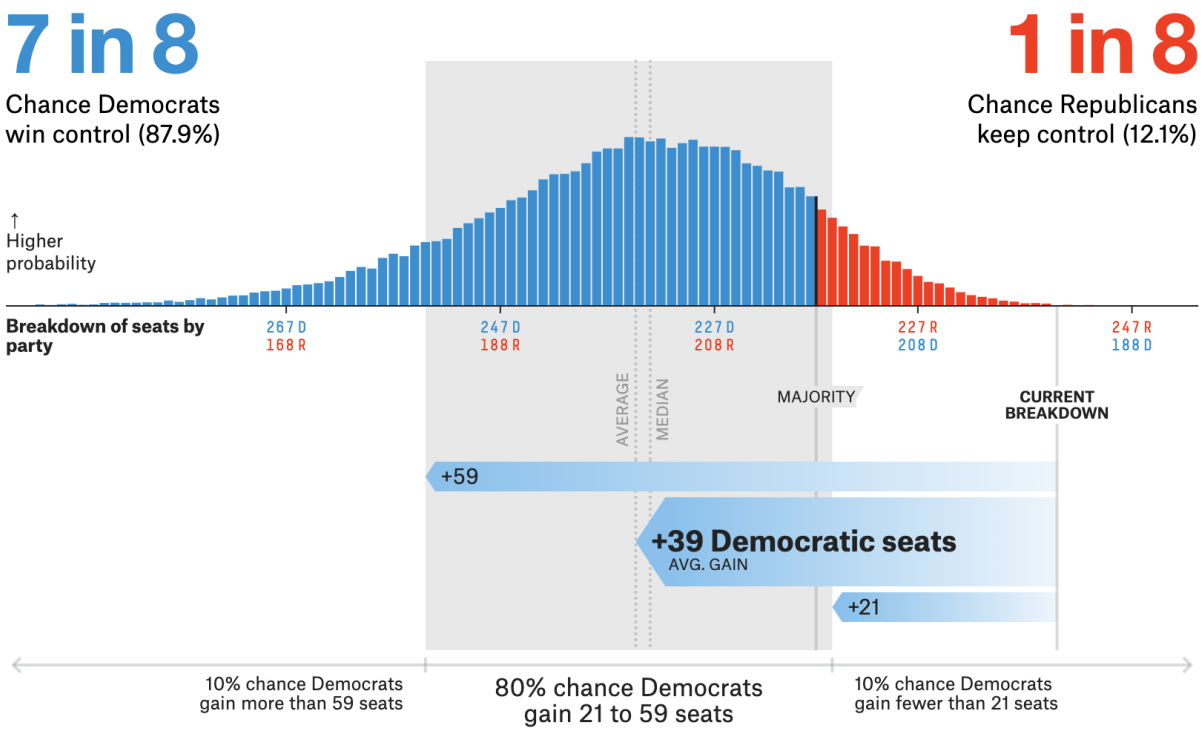
Lite
Keep it simple, please — give me the best forecast you can based on what local and national polls say

Classic
I'll take the polls, plus all the "fundamentals": fundraising, past voting in the district, historical trends and more

Deluxe
Gimme the works — the Classic forecasts plus experts' ratings

Forecasting the race for the House

Updated Nov. 6, 2018, at 11:06 AM



https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/

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

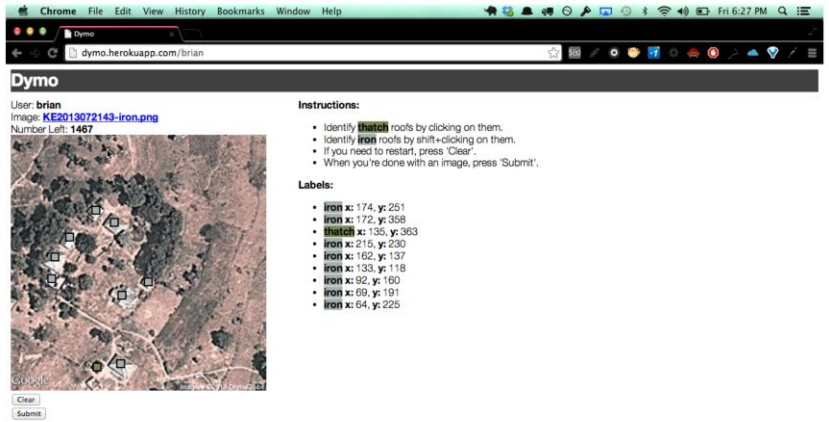


Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.

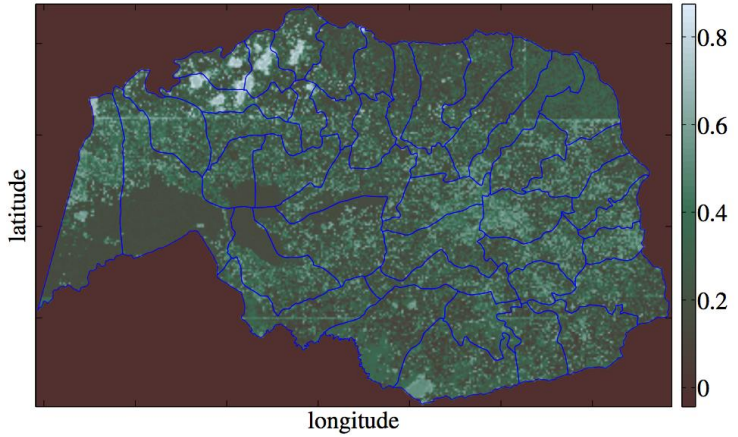


Figure 11: Heat map of proportion of roofs that are metal in the region of interest.

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

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 some 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)

15-112

We strongly 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

Course Team

Instructor

Pat Virtue

Students!

Office Hours

Course materials and discussion

All course material (slides, notes, lecture ~~videos~~, assignments) is available on the course webpage:

↳ Canvas

http://www.datasciencecourse.org

Slides posted before class, videos up ~2-3 hours after, notes posted asynchronously

Piazza for all communication

1 Public
2. ↳
3. ↳
4 Private

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

If we can get logistics to work out, 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)

388 → 688

388 ← 688

Course videos

All lectures will be recorded, made available on the course website and via Canvas (this is all that Canvas is used for for the course)

Students from any section may opt to view the class recordings instead of attending class (but of course, you won't be able to ask questions then)

Note that even if you ask a question in class, the video likely will not pick up your voice (I need to repeat questions after they are asked)

Auditing, Pass/Fail, and Waitlists

No auditing

Pass/fail up to you and your academic/program advisor

Waitlist: most students on waitlist won’t get in

Grading

Grading breakdown:

- 40% homework
- 20% tutorial
- 30% class project
- 10% class participation

A 90% ↑

B 80 ↓

70

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

Homeworks

One homework assignment every two weeks

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

Homeworks are are in the form of Jupyter notebooks (accessible via Colab, if desired), **solutions autograded via a new system we are developing, more info with first HW release**

Autograding

The meta-goal for this course is to have a *scalable* introduction to data science

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

This presents 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)

Note: tutorial and class project will be graded manually (by myself)

Late days

You have **5 late days** to use over the course of the semester

Each assignment can use a maximum of **2 late days**

You cannot use late days for tutorial or final project submission

Class participation

Participation points for answering polls in Piazza during lecture

Peer Instruction

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 Colab 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 Colab 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

Collaboration on homeworks

All submitted content (code and prose for homeworks, tutorials, and final project) should be your own content, written yourself

However, you *may* (in fact are encouraged to) discuss the homework with others in the class and on the discussion *including posting code*

- This creates some room for undue copying, but please obey the reasonable person principle: discuss as you see fit, but don't simply share answers

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 again, be reasonable

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