# 15-388/688 - Practical Data Science: Introduction

J. Zico Kolter
Carnegie Mellon University
Fall 2019

# 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

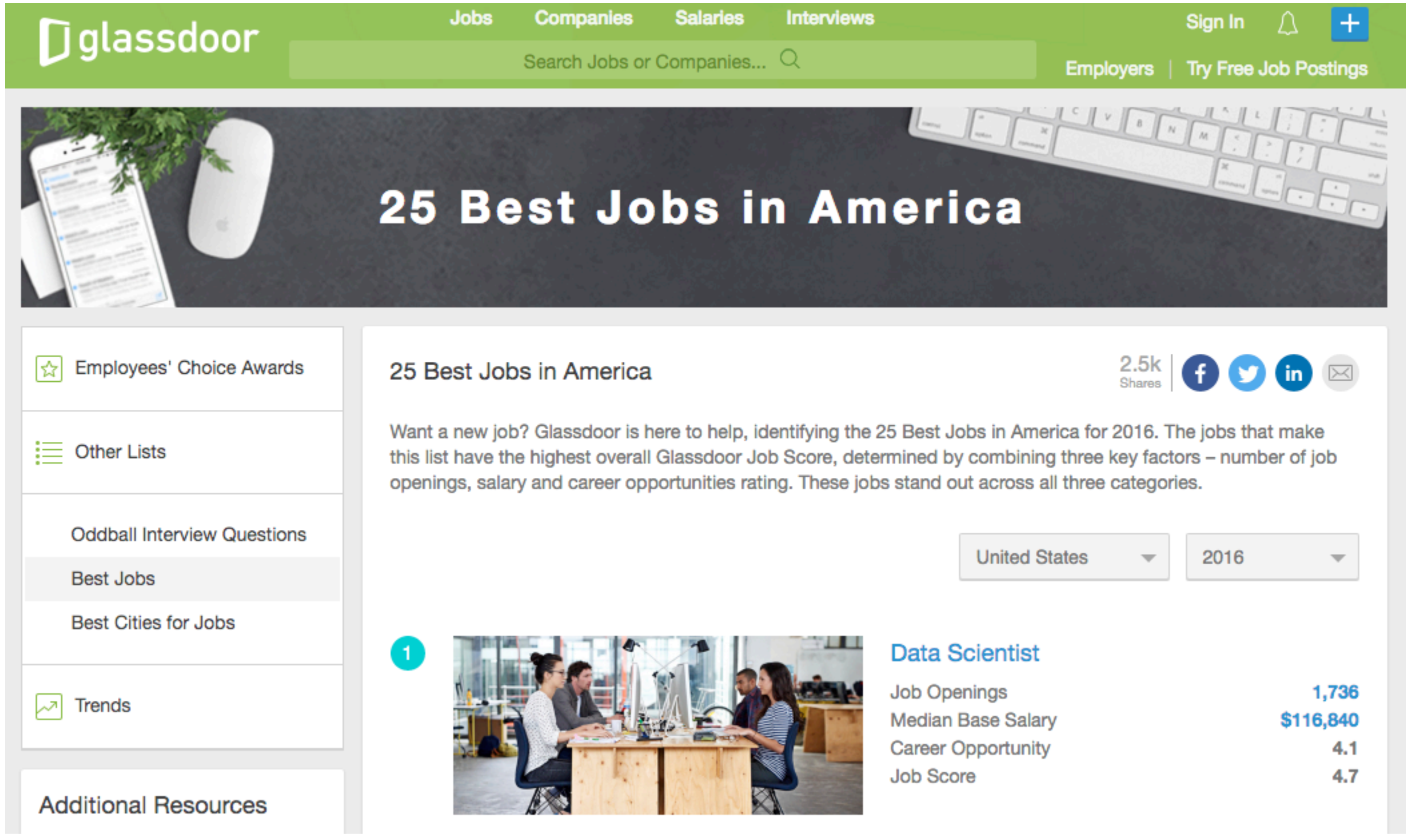
# Some possible definitions

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



# 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

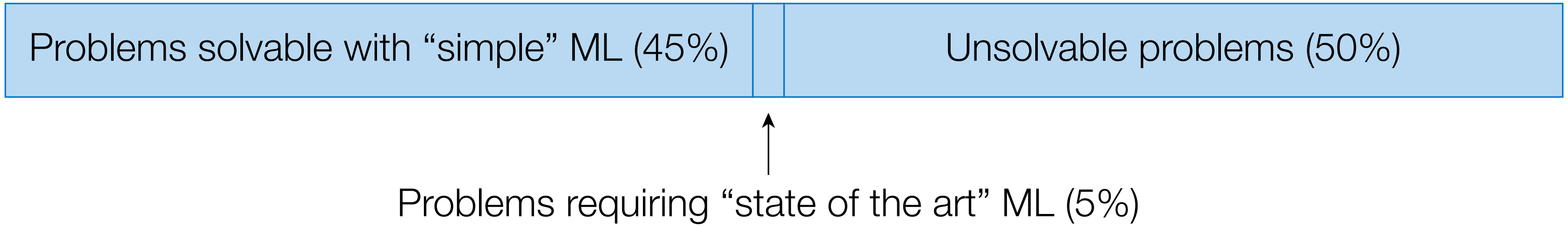
Machine learning involves computation and statistics, but has not (traditionally) been very concerned about answering *scientific questions*

Machine learning has a heavy focus on fancy algorithms…

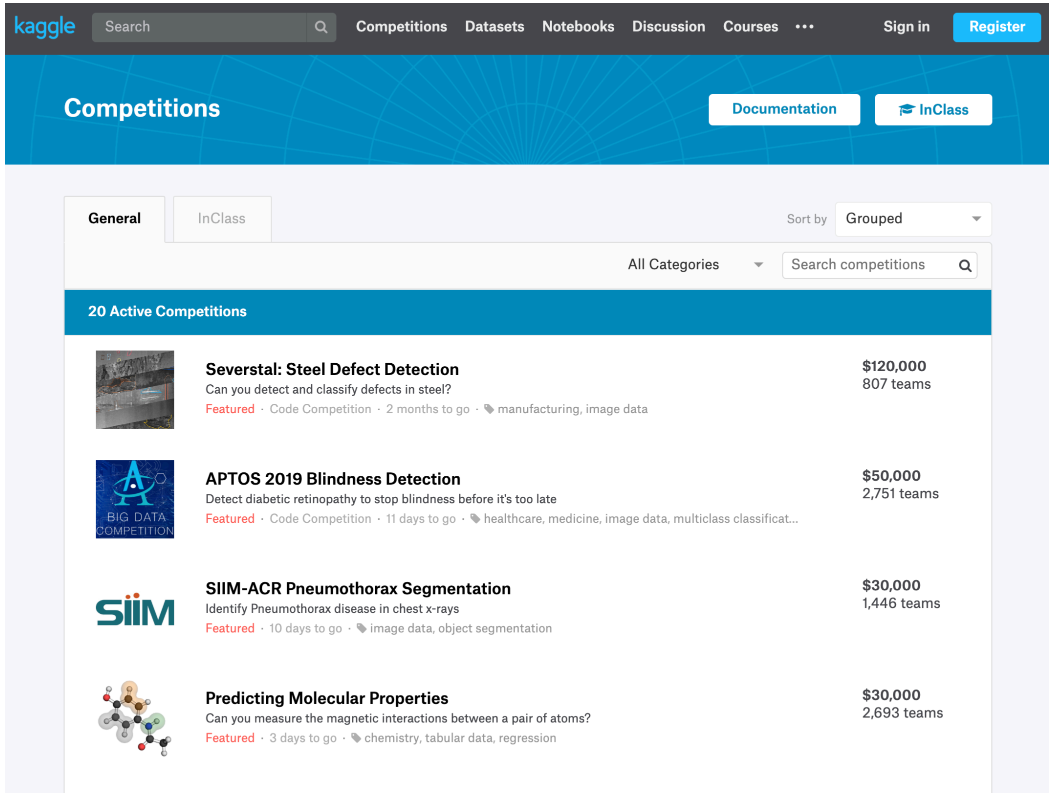
... but sometimes the best way to solve a problem is just by visualizing the data, for instance

# Data science is not machine learning

**Universe of machine learning problems**



# 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 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

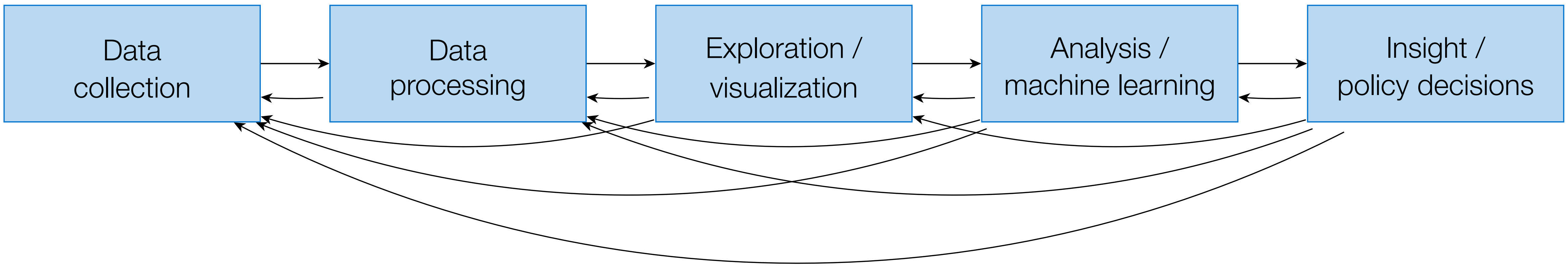
# Data science is not big data

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.

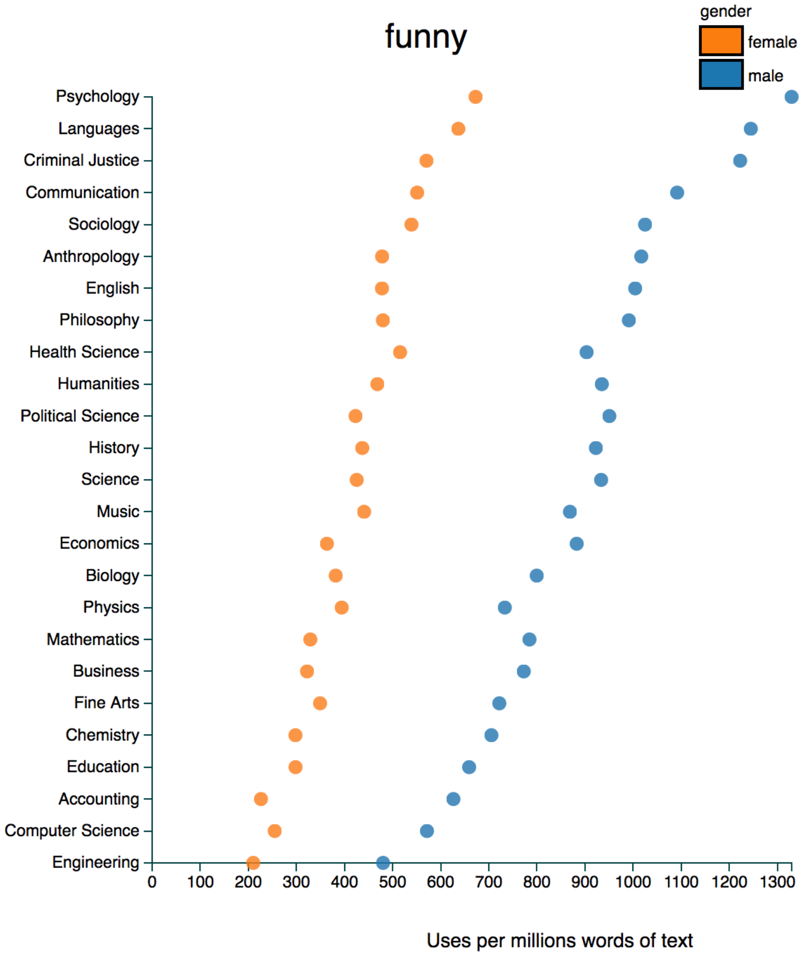
You can enter any other word (or two-word phrase) into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). You can also limit to just negative or positive reviews (based on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



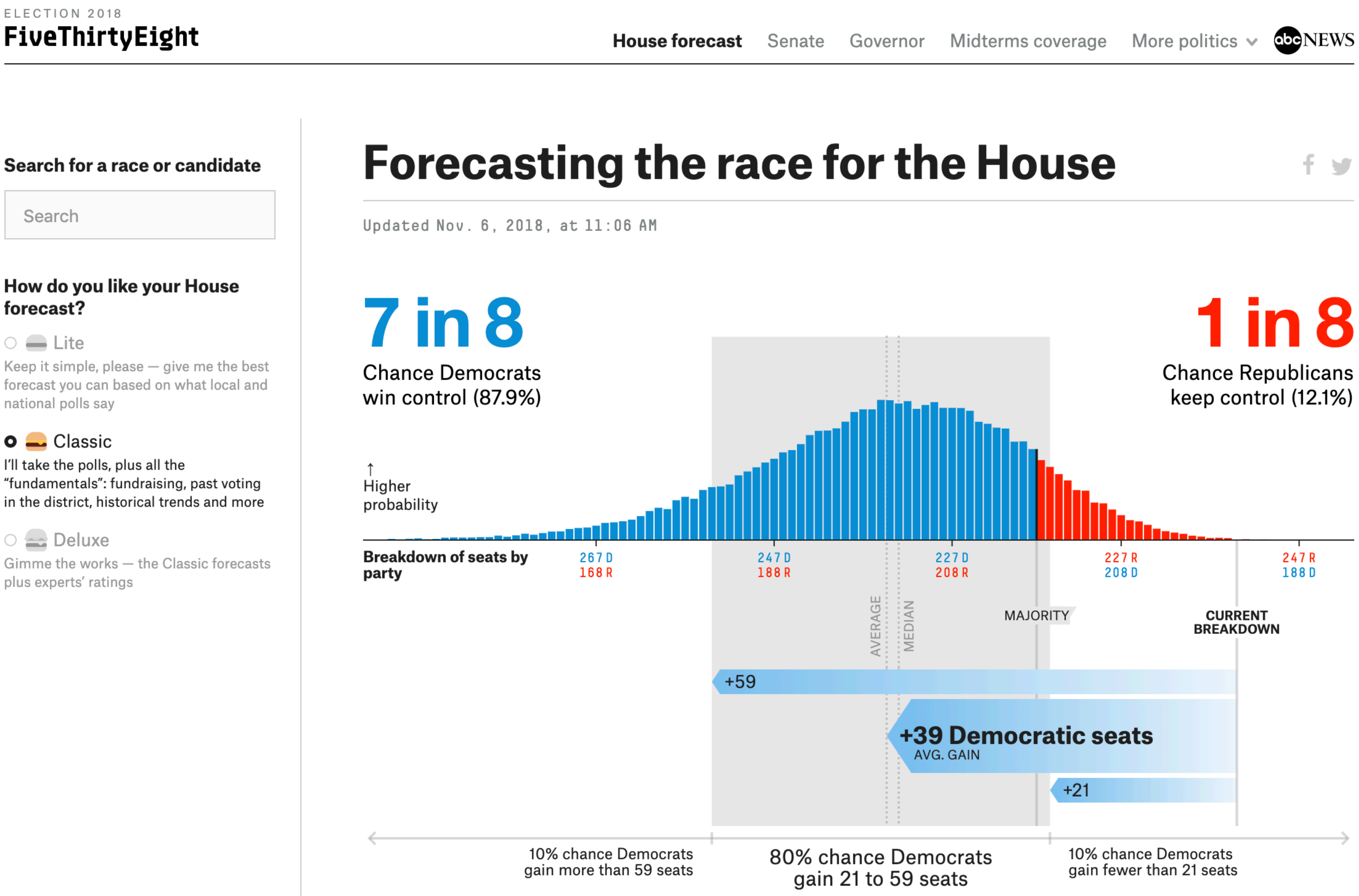

http://benschmidt.org/profGender/

# Obligatory quote

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

-John Tukey

# FiveThirtyEight




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

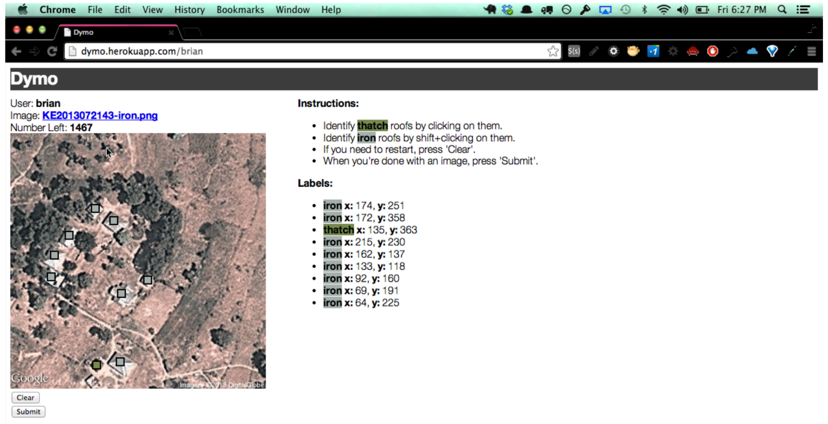



**Figure 6: Screen shot of application deployed for crowdsourced labeling of roofs in satellite images.**

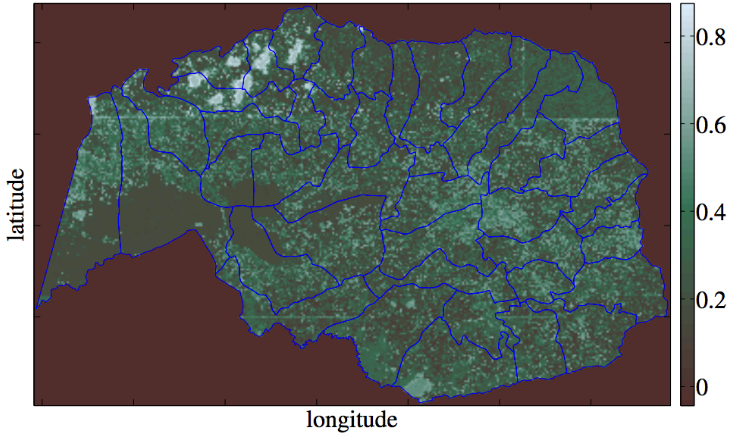



**Figure 11: Heat map of proportion of roofs that are metal in the region of interest.**

# 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)

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 materials and discussion

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

http://www.datasciencecourse.org

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

All forums/discussion and homeworks will be submitted online via Diderot, signup instructions on the course page, under “Assignments”

http://diderot.one

# 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

The reality is that by the first few weeks of the semester, there will be room in the course, even if you are in Section B

Will I get off the waitlist?

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

# Auditing?

Auditing is permitted, but *only* within the B section of 15-688 (i.e., non-auditors will have preference for in-class version of the course)

The requirements to pass an audit are to receive at least **50% of the points on 4 out of the 5 assignments** (out of the whole assignment, so both the two 388 questions and the one additional 688 question)

No tutorial or final project are required for audit

We discourage final projects consisting of some full-credit participants and some auditors, unless you have a very good reason

# Course videos

All lectures will be recorded, made available on the course website (a permanent link to all the videos will also be posted)

Attendance still required for the Section A students (more on this in a moment)

Videos are being made **publicly available** this semester, so be aware of this if you sit nearby the camera

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)

# Grading

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

50% homework
15% tutorial
25% class project
10% class participation

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

# Homeworks

One homework assignment every two weeks: released on Thursdays by midnight, due the Thursday two weeks later at midnight (though first homework is already released, **due 9/12**)

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 Diderot:** http://diderot.one

# 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

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

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

Each assignment can use a maximum of **2 late days** (midnight Saturday)

You cannot use late days for final project submission

# Class participation

For 15-388/688A (in-class sections), class attendance is required: class participation grade will come from **participating in in-class Diderot polls** (you don't need to submit the right answer, just an answer)

For 15-688B (online section), you will need to watch all the videos lectures (Panopto system tracks this), and **answer a short quiz, within one week** of the lecture

If you are in Section A and miss a class, you should watch the video and take the corresponding quiz; if you are in the B section and attend class (and answer poll), you don't need to watch the video or answer the quiz

Additional extra credit participation for *answering* student questions on Diderot

# 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

# Academic integrity and homeworks

All submitted content (code and prose for homeworks, tutorials, and and final project) must be your own original content

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

# 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 today, use it as a gauge (after a few of the next lectures) to determine if the course is right for you