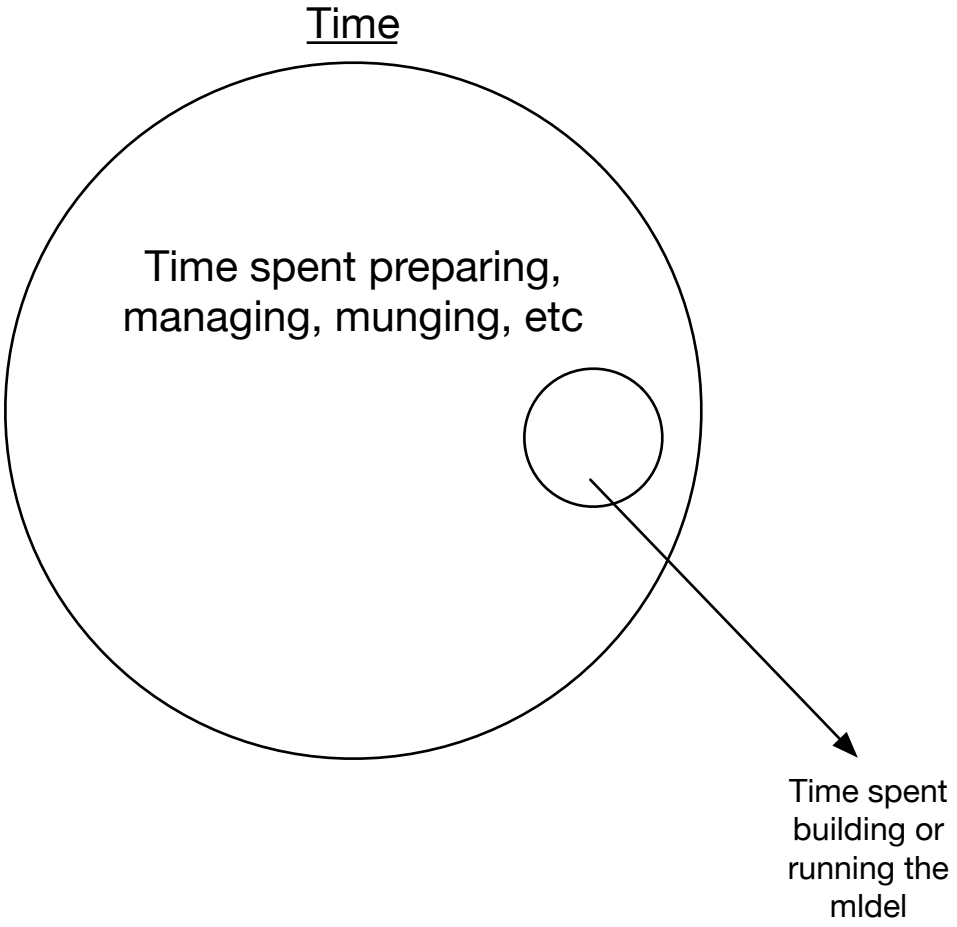
6.S080 Lecture 1 -- 9/4/2019

What is this course about?

The dirty parts of data science -- Tim gave a bunch of examples of interesting data science results; in this class we're going to teach you what it takes to get to those results in a careful, principled, way. Often you hear people say that 95% of the time getting a data science result.



In this class we're going to teach you all the stuff that happens in the 95%, and a little bit of what happens in that tiny circle, because other classes teach you about that tiny circle.

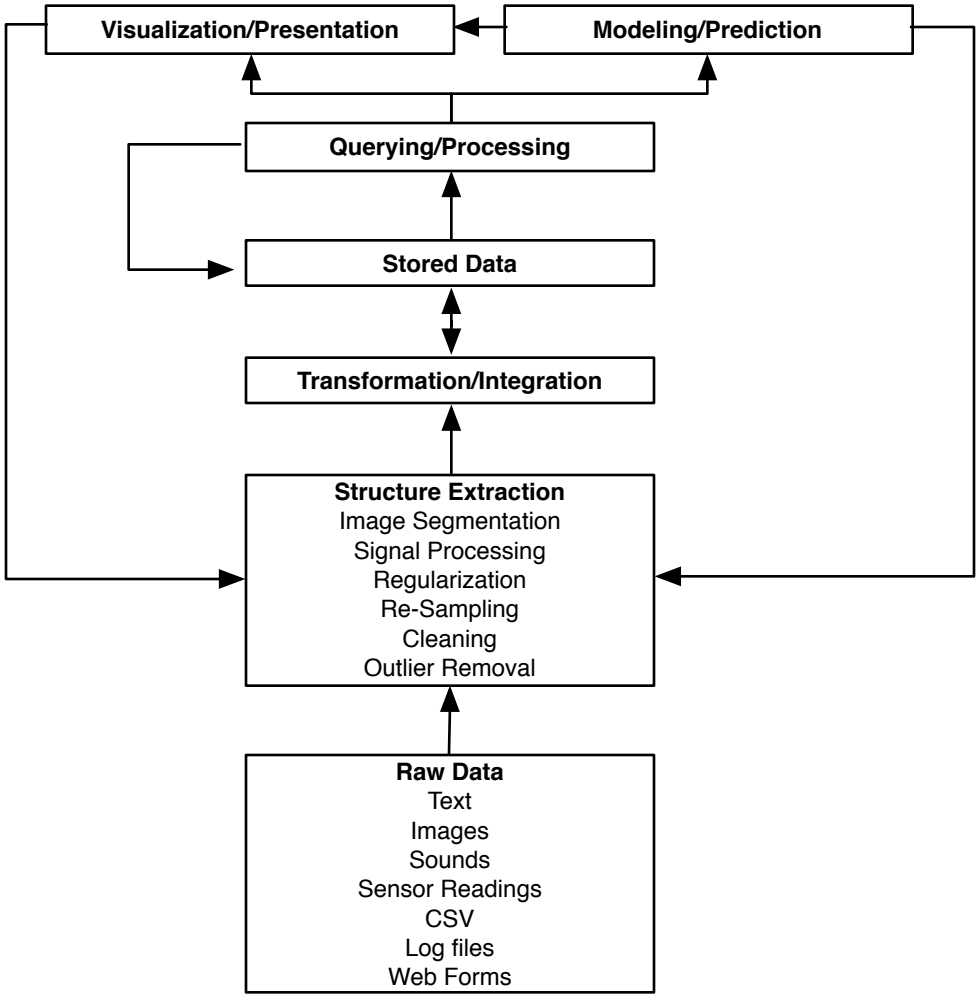
Why don't other classes teach this stuff?

Traditionally viewed as easy or obvious, but there has a been a huge proliferation of techniques and tools in the past decade or so, and we think it's useful for you to learn it for real.

A survey of systems, tools, algorithms, tips, tricks, etc to help make sense out of data.
Key concepts:

- data representation ("modeling")
- extracting structure ("wrangling")
- getting data sets to conform with each other, and align properly -- "integration and clenaing"
- different computational frameworks
    - with differing performance and scalability characteristics
    - and designed with different types of data in mind
- basic stats and machine learning
    - at scale
- visualization techniques
    - when data is large, needs to be interacted with, e
- how to know when you're making mistakes

6.S080 Pipeline



One Size Doesn't Fit All

Recent years have seen a proliferation of tools for processing data

SQL-based systems
SparkSQL
Athena
Redshift
Snowflake

NewSQL Systems
Text databases
Time series databases
Graph databases
Map Reduce/Hadoop/Spark

Text/ Semistructured databases (e.g., Mongo, XML databases)

Visualization Systems
d3
ggplot2
vege

...

Programming Languages
Python
Excel
SAS
SPSS
R
Matlab
Julia

Computing infrastructures, Mobile platforms, …

Course Organization and Structure

Two lectures per week, 1 in 3 will be a "hands on" in class start on a lab.

Lectures will involve readings, not as in-depth as a graduate CS class but some will be fairly technical.

We will have several lectures at the end that are "free" for special topics and for you to come to class to talk about your final projects.

Bulk of grades are for labs and projects.

50% project
40% lab
10% participation

No exams

We will start labs in class, you will need to have a laptop with course environment set up on it. Lab 0, due next Wednesday, is designed to ensure your laptop is properly set up.

Labs:

Labs will focus on practical, hands-on skills.

- working with data in pandas / sql
- data wrangling
- data cleaning and integration
- end to end data modeling and prediction
- data visualization
- scalable data processing tools (spark)
- stats and statistical errors

Typically labs will be available on Monday and due on Wednesday.

Labs can be done individually or in pairs. 5 "late days"

Project:

Should be done in teams of 3. Deliverables include a proposal, a mid-term report, and a final report. Mid-term report will involve a presentation to staff, and final report will involve a
Several options. One: design a system for processing data in some new way. Examples:
- build a tool to mine sentiment about a particular hashtag on twitter
- build a system that automatically attempts to select the best features for some class of ML problem

Two: pick a dataset of interest, and then do some analysis on it. This should be an end

to end analysis that includes prepping, loading, querying, and analyzing the data using tools we study.

Start thinking of some data now, and what you want to do with it.

There's lots of data out there:

- Election data (donations, fundrasing)
- Sports data, at surprising granularity (e.g., individual shots / passes)
- Education data (admissions, jobs, costs, loans)
- Medical data (medicare, billing, etc.)
- Federal funding (defense, nsf, etc)
- Real estate (transactions, property prices, restaurants, etc)

Example projects:

- Study whether basketball players actually have "hot streaks"
- Analyze demographics of admissions in public / private colleges
- Look for examples of "upcoding" in medicare data