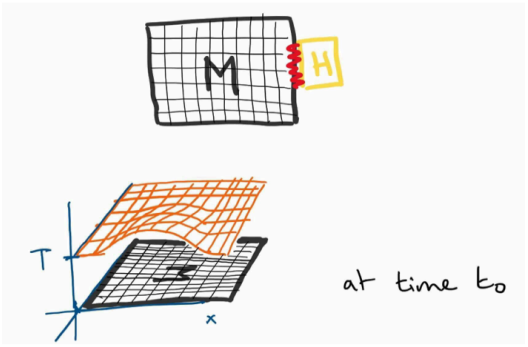
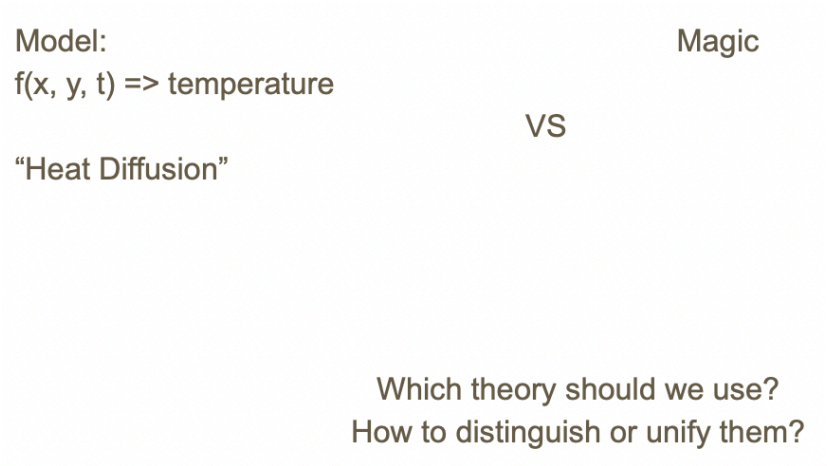
CAS CS 506
Lec 03

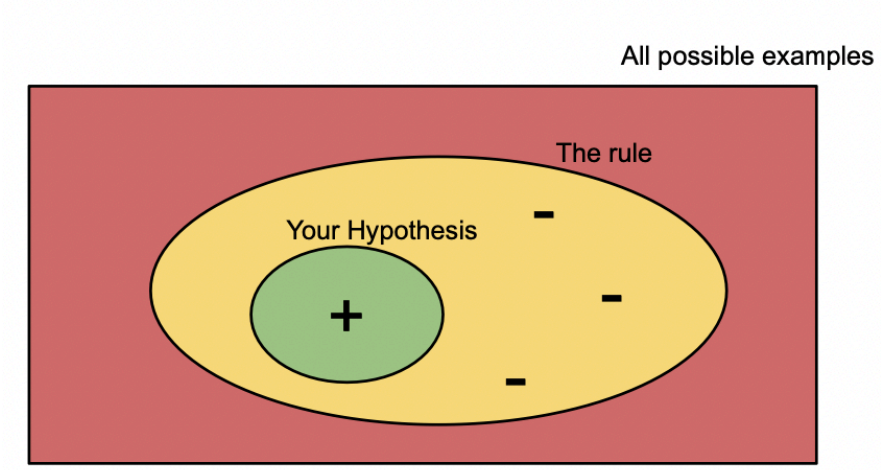
Intro to DS

1. Data Science
    a. Collection of methods and tools that allow for extracting knowledge from data
    b. Cross-disciplinary:
        i. Math
        ii. Statistics
        iii. Computer Science
        iv. Domain Expertise
    c. Know what you don't know!
2. Knowledge = Testable Predictions
    a. 

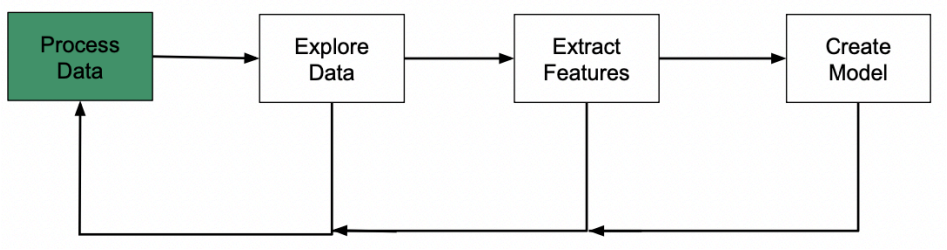
    b. 

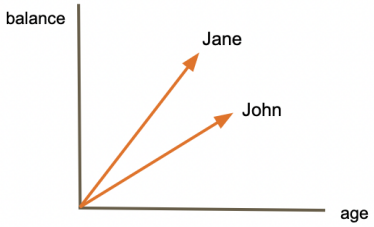
    c. 

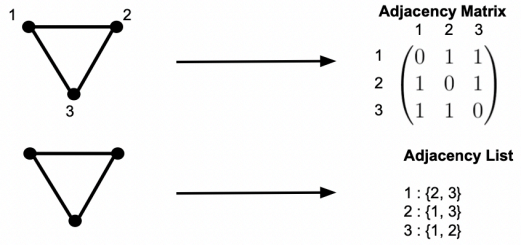
    d. If you can equally well explain every outcome, how can you have a definitive/deterministic anticipation of events?
    e. If you're equally good at explaining every outcome, you have zero knowledge

3. Confirmation Bias
    a. In a class just like this one, imagine playing the following game…
        i. I announce "(2, 4, 6) follows the rule".
        ii. Here are the examples submitted by one of the participants:
        (2,4,3) → No
        (6,8,10) → Yes
        (1,3,5) → Yes
        iii. After which, they proceed to write down their hypothesized rule. Would you have wanted to try more examples? If so, which and for what reason?
            1. Try example (7,8,9), (-5, -3, -1), and (-4, -3, -2)
    b. Challenges of data science
        i. Not all examples contribute similar amounts of information
        ii. A set of examples may not always be representative of the underlying rule
        iii. There may be infinitely many rules that match the examples provided
    c. Both positive and negative examples can falsify a hypothesis
        i. Positive Examples → Examples that would output True
        ii. Negative Examples → Examples that would output False
        iii. Try positive and negative examples equally (negative examples are equally important)
    d. Tendency to choose positive ones over negative ones
    e. 

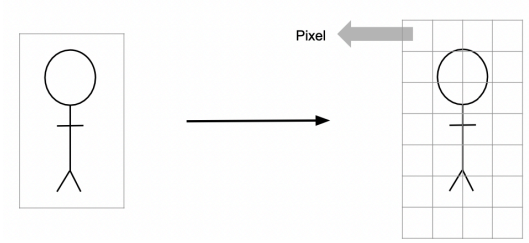
    f. The rule was (a < b < c)
    g. If we only tried positive examples of either (x, x+2, x+4) or (x, 2x, 3x) you would only get confirmation
4. Data Science Workflow (simplified)
    a. First ask what and who the model is used / intended for
        i. Is it just the general trend that is important of the exact predictions that are important?
        ii. Is this a problem that needs predictive tools to solve?

b. 

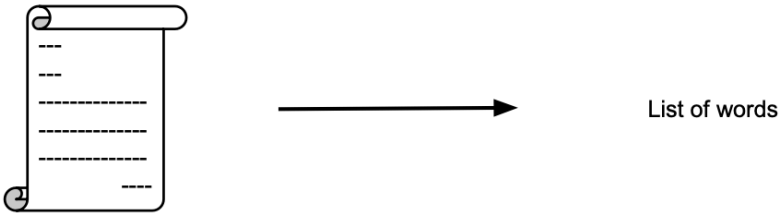
5. Data Processing
    a. What data should and shouldn't be used for the task?
    b. What to do with missing data?
    c. What to do with inconsistent data?
    d. What assumptions are you making with the transformations of the data?
6. Exploratory Data Analysis
    a. Describe, contextualize, and visualize the data
    b. What might be related to what you're trying to predict?
    c. Are there imbalances in the data?
7. Feature Extraction
    a. Are the features provided by the dataset the best features to use for the task?
    b. What other features can be extracted?
    c. Should existing features be transformed?
8. Finding the Right Model
    a. The success of this step depends entirely on the work done in previous steps - remember: garbage in, garbage out!
    b. Is your model easy to explain?
    c. When your model fails, can you explain why?
9. Types of Data
    a. Records
        i. m - dimensional points / vectors
        ii. Example: (name, age, balance) → ("John", 20, 100)
        iii. 

    b. Graphs
        i. Nodes connected by edges
        ii. 

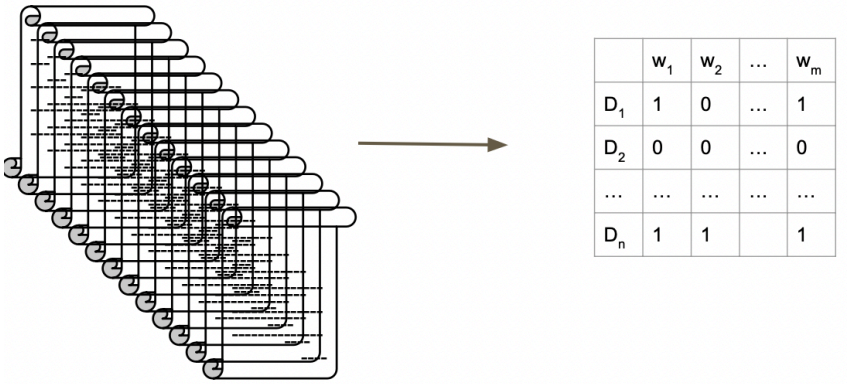
    c. Images

i. 

d. Text

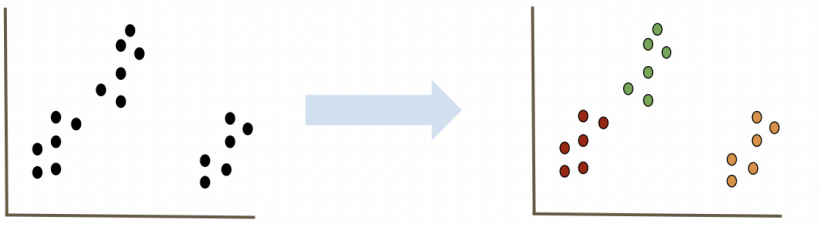
i. 

e. Corpus of Documents

i. 

	w_1	w_2	…	w_m
D_1	1	0	…	1
D_2	0	0	…	0
…	…	…	…	…
D_n	1	1		1

10. Types of Learning
    a. Unsupervised Learning
        i. Goal: Find an interesting structure in the data

ii. 

iii. What are some linear algebraic properties of the matrix of data? What does that tell me about the data?

iv. Goals:

1. Better understand / describe the data
    a. Data exploration / visualization step
    b. Find anomalies
    c. Recommender Systems (similiar users might be recommended the same things, emails similar to those marked as spam could be spam, etc.)
2. Extract Features
3. Fill in gaps in data

a. Data preprocessing step

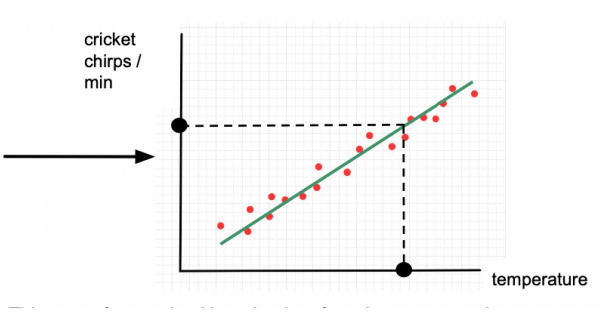
4. Make learning algorithms faster

a. Get rid of noise

b. Supervised Leanring

i.

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



ii.

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1