Introduction to Machine Learning

CS 534: Machine Learning

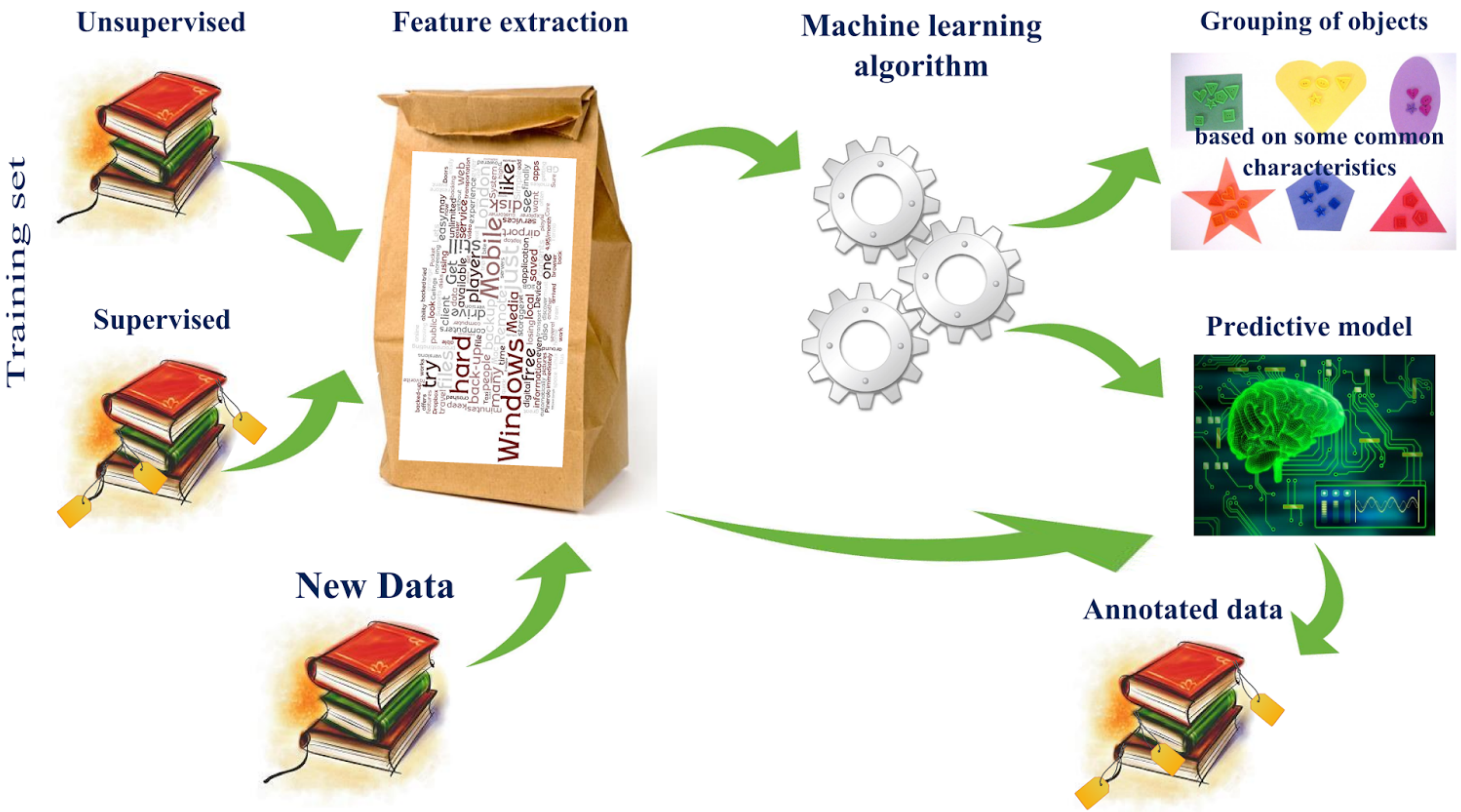
Slides adapted from Lee Cooper, David Sontag, Luke Zettlemoyer, Pedro Domingos, and Carlos Guestrin

Why Machine Learning?

“We are drowning in information and starving for knowledge.” — John Naisbitt

- Big data era
- Use algorithms to discover new relationships, scale tasks, and perform decision making under uncertainty

Machine learning workflow



http://nkonst.com/machine-learning-explained-simple-words/

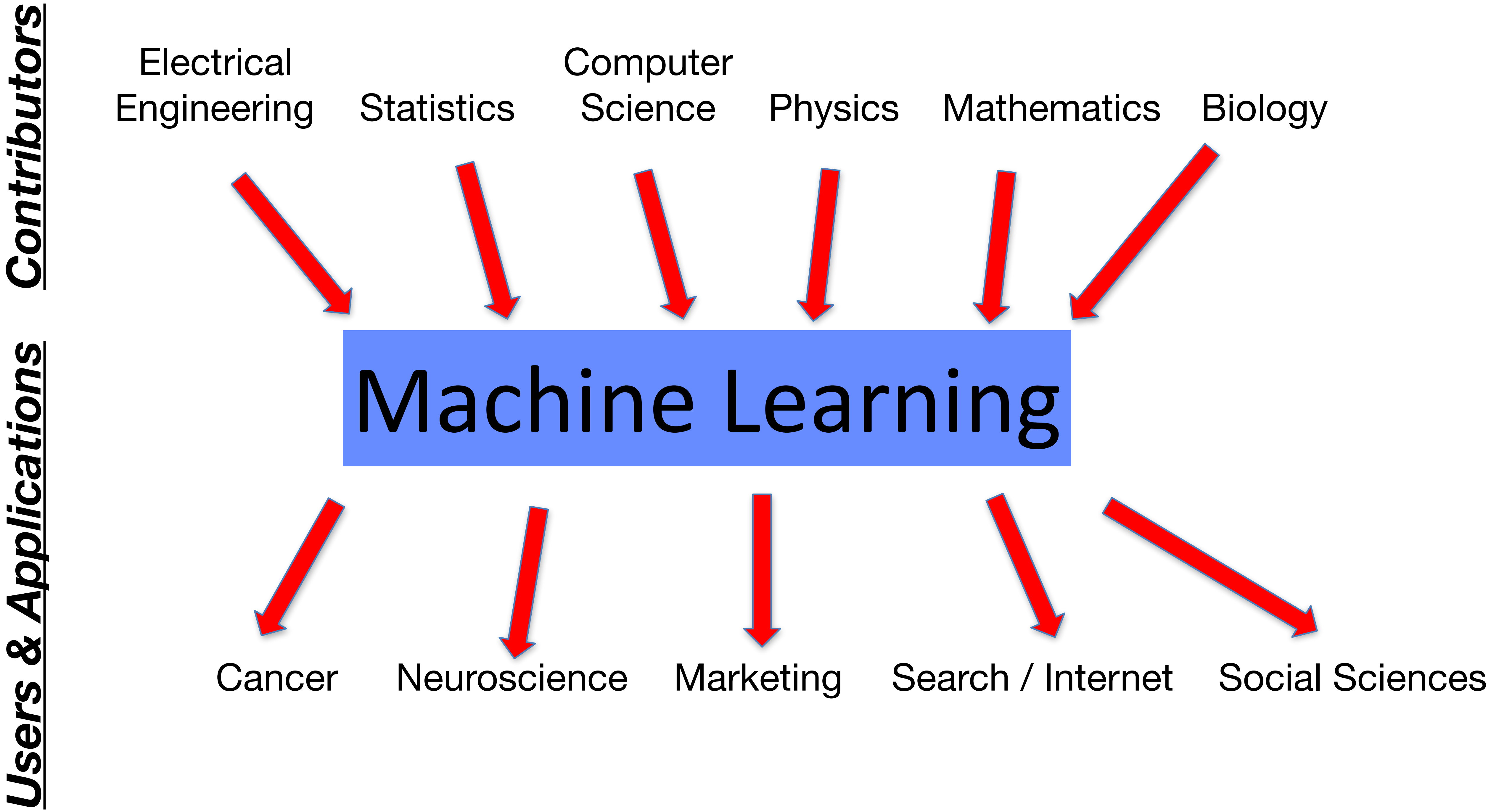
ML, DM, or AI?

- Machine Learning (ML) - The study, design and development of algorithms that endow learning capabilities

- Data Mining (DM) – Using ML and statistical techniques to learn something new from data

- Artificial Intelligence (AI) – Broad study of developing intelligent agents (think Turing Test)

Diverse Community



Journals & Conferences

- Journals
  - Transactions on Pattern Analysis and Machine Intelligence (IEEE)
  - Journal of Machine Learning Research (ACM)
  - Machine Learning (Springer)
  - Foundations and Trends in Machine Learning (ACM)

- Conferences
  - ICML – International Conference on Machine Learning
  - NIPS – Neural Information Processing Systems
  - UAI – Uncertainty in Artificial Intelligence
  - CVPR – Computer Vision Pattern Recognition

Complete list at Microsoft Research Rankings

All About Benchmarking

- Abundant data available to compare algorithms
- Required for publication
- Makes ML more of a science
- Still difficult to make fair comparisons
  - What parameters to use
  - Difficult to sweep parameter space

Tools for Machine Learning

- Python

  Pros: free, fast, many algorithms available

  Cons: can be slow

Code examples will be provided in Python

- R

  Pros: free, standard in bioinformatics & statistics, great vector graphics

  Cons: extremely slow, poorly documented, bad language conventions

- Matlab

  Pros: fast, large user community & codebase, well documented

  Cons: not free

Course Logistics

Course Website

http://joyceho.github.io/cs534-s17/index.html

- Lectures
- Assignments
- Example code (when applicable)

About Me (Joyce Ho)

- Undergraduate / MEng from MIT
- PhD from University of Texas at Austin
- Research interests:
  - Data Mining / Machine Learning
  - Healthcare Informatics
- More information: http://joyceho.github.io

Contact Information

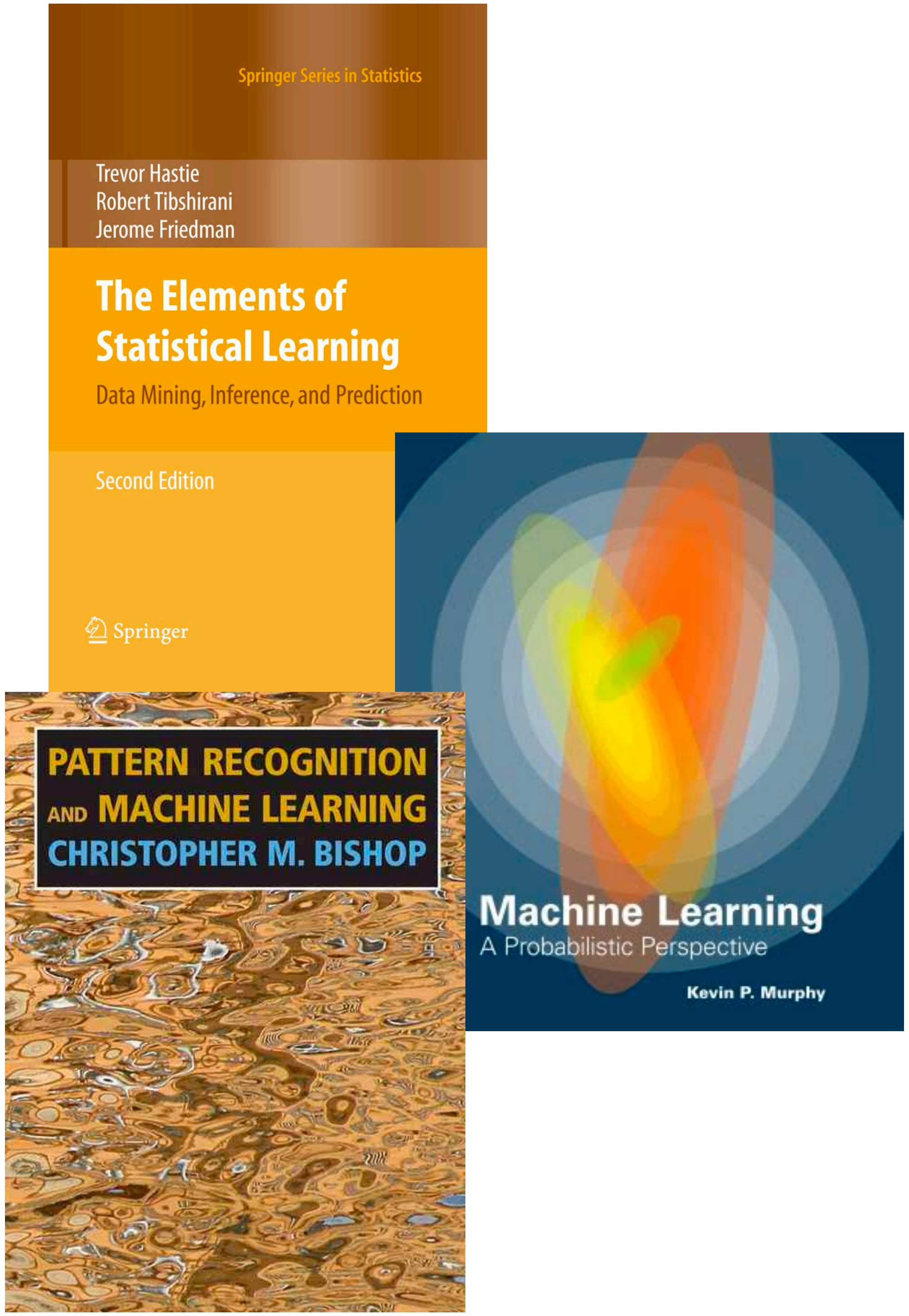
- Email: joyce.c.ho@emory.edu
- Office Hours @ MSC W414
  - M 1:00 pm - 3:30 pm
  - W 9:30 am - 12:00 pm

Communication

- Piazza: http://piazza.com/emory/spring2017/cs534
  - Announcements
  - Questions + Discussions
  - Assignment Clarifications + Slide Corrections
- Office Hours
- By Appointment!

Course Textbook

- Elements of Statistical Learning
  - PDF available online
- Machine Learning: a Probabilistic Perspective, by Kevin Murphy (Optional)
- Pattern Recognition and Machine Learning, by Christopher Bishop (Optional)


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
Machine Learning
A Probabilistic Perspective
Kevin P. Murphy

Evaluation

- 4-5 assignments (40%)
    - Both theory & programming
- Midterm (15%)
- Project (40%)
- Participation (5%)

Collaboration Policy

- Try the assignments on your own first
- Discuss with others if necessary
- Write-up solutions on your own
- List the people you collaborated with

Project

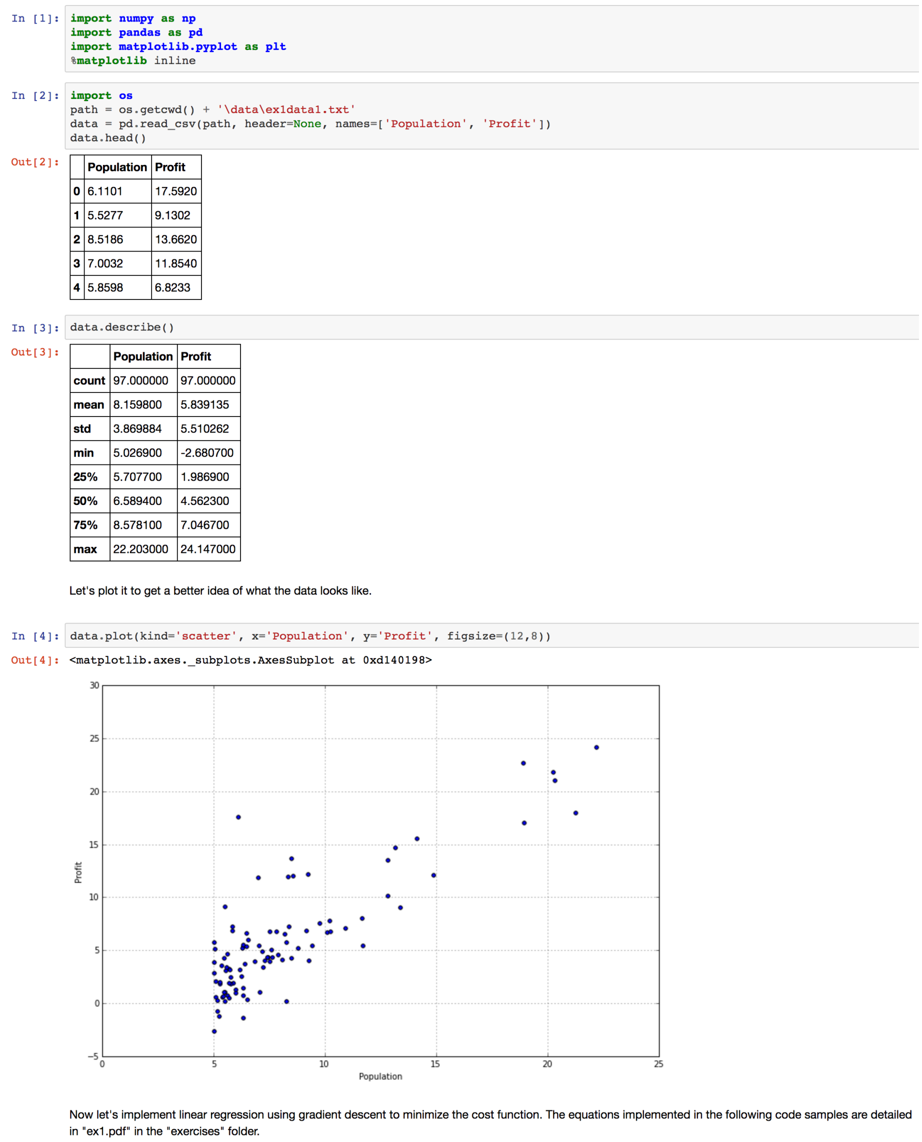
- Work in groups of 1-2
- Emphasis on public data sets (e.g., Kaggle competitions, MovieLens, KDD Cup, etc.)
- Project proposal due by spring break for feedback
- Goal is to either develop a new algorithm or try multiple algorithms to achieve good performance

Prerequisites

- REQUIRED:
  - Probability theory
  - Linear algebra
- STRONGLY RECOMMENDED:
  - Statistics
  - Good programming skills

Jupyter [IPython] Notebooks

- Designed to make lectures more ‘interactive’ — in-class activity to learn together
- Provide a means for more exercises to better understand the material
- Learn some Python along the way



```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
```

```
In [2]: import os
        path = os.getcwd() + '\data\ex1data1.txt'
        data = pd.read_csv(path, header=None, names=['Population', 'Profit'])
        data.head()
```

Out[2]:

	Population	Profit
0	6.1101	17.5920
1	5.5277	9.1302
2	8.5186	13.6620
3	7.0032	11.8540
4	5.8598	6.8233

```
In [3]: data.describe()
```

Out[3]:

	Population	Profit
count	97.000000	97.000000
mean	8.159800	5.839135
std	3.869884	5.510262
min	5.026900	-2.680700
25%	5.707700	1.986900
50%	6.589400	4.562300
75%	8.578100	7.046700
max	22.203000	24.147000

Let's plot it to get a better idea of what the data looks like.

```
In [4]: data.plot(kind='scatter', x='Population', y='Profit', figsize=(12,8))
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0xd140198>
```

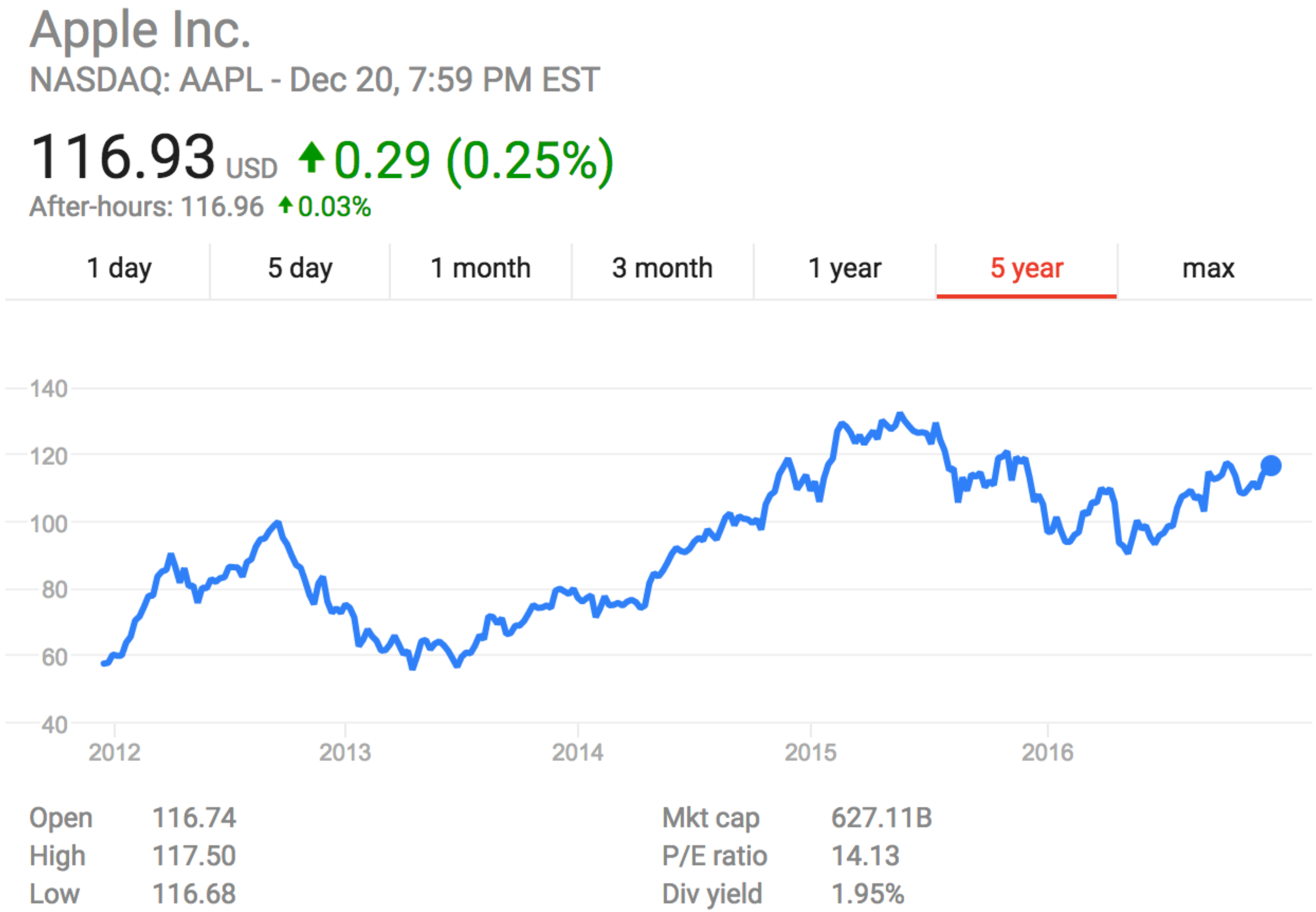
Now let's implement linear regression using gradient descent to minimize the cost function. The equations implemented in the following code samples are detailed in "ex1.pdf" in the "exercises" folder.

Preview of Topics

Linear Regression

- How can we build linear models to predict continuous-valued outcomes?
- How can we analyze these models to understand the importance of features?

Linear Regression: Stock Market



Apple Inc.

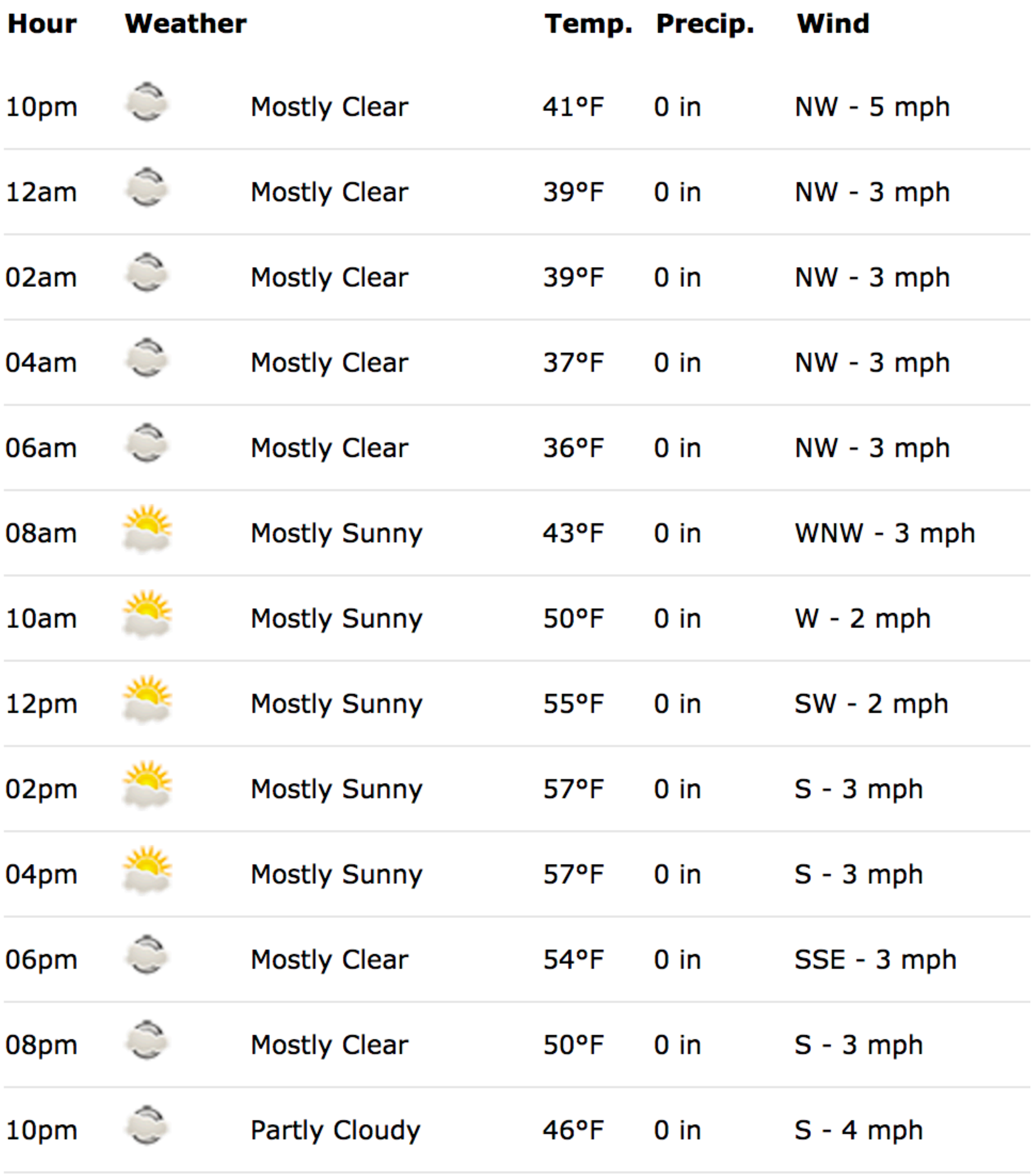
NASDAQ: AAPL - Dec 20, 7:59 PM EST

116.93 USD ↑0.29 (0.25%)

After-hours: 116.96 ↑0.03%

Open	116.74	Mkt cap	627.11B
High	117.50	P/E ratio	14.13
Low	116.68	Div yield	1.95%

Linear Regression: Weather

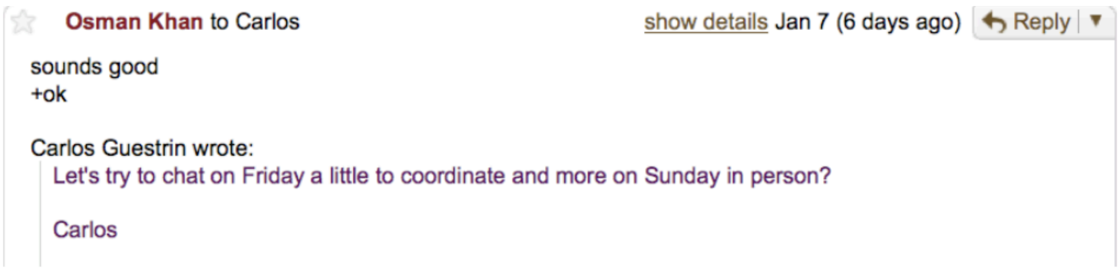


Hour	Weather	Temp.	Precip.	Wind
10pm	Mostly Clear	41°F	0 in	NW - 5 mph
12am	Mostly Clear	39°F	0 in	NW - 3 mph
02am	Mostly Clear	39°F	0 in	NW - 3 mph
04am	Mostly Clear	37°F	0 in	NW - 3 mph
06am	Mostly Clear	36°F	0 in	NW - 3 mph
08am	Mostly Sunny	43°F	0 in	WNW - 3 mph
10am	Mostly Sunny	50°F	0 in	W - 2 mph
12pm	Mostly Sunny	55°F	0 in	SW - 2 mph
02pm	Mostly Sunny	57°F	0 in	S - 3 mph
04pm	Mostly Sunny	57°F	0 in	S - 3 mph
06pm	Mostly Clear	54°F	0 in	SSE - 3 mph
08pm	Mostly Clear	50°F	0 in	S - 3 mph
10pm	Partly Cloudy	46°F	0 in	S - 4 mph

Linear Classifiers

- What if the response variable is categorical or discrete?
- What are strategies for selecting the best features and dealing with noise?

Linear Classifiers: Spam Filtering

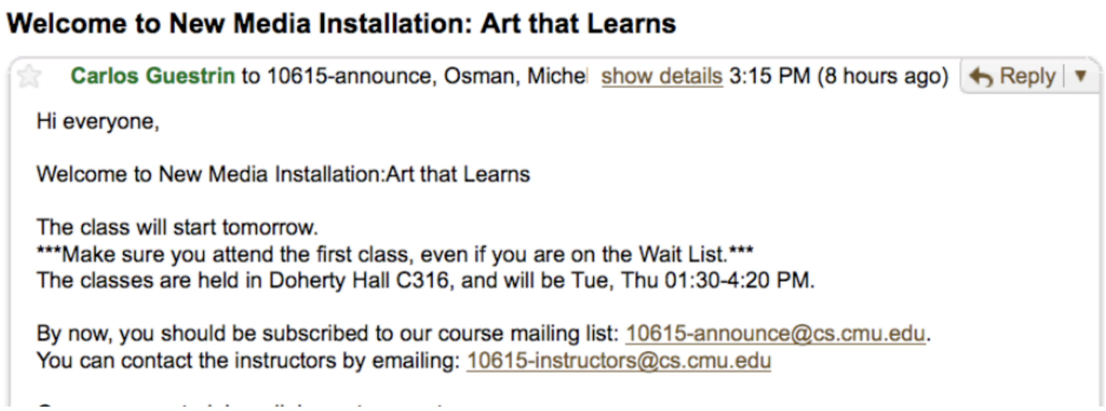


Osman Khan to Carlos — show details Jan 7 (6 days ago) — Reply

sounds good
+ok

Carlos Guestrin wrote:

> Let's try to chat on Friday a little to coordinate and more on Sunday in person?
>
> Carlos



Welcome to New Media Installation: Art that Learns

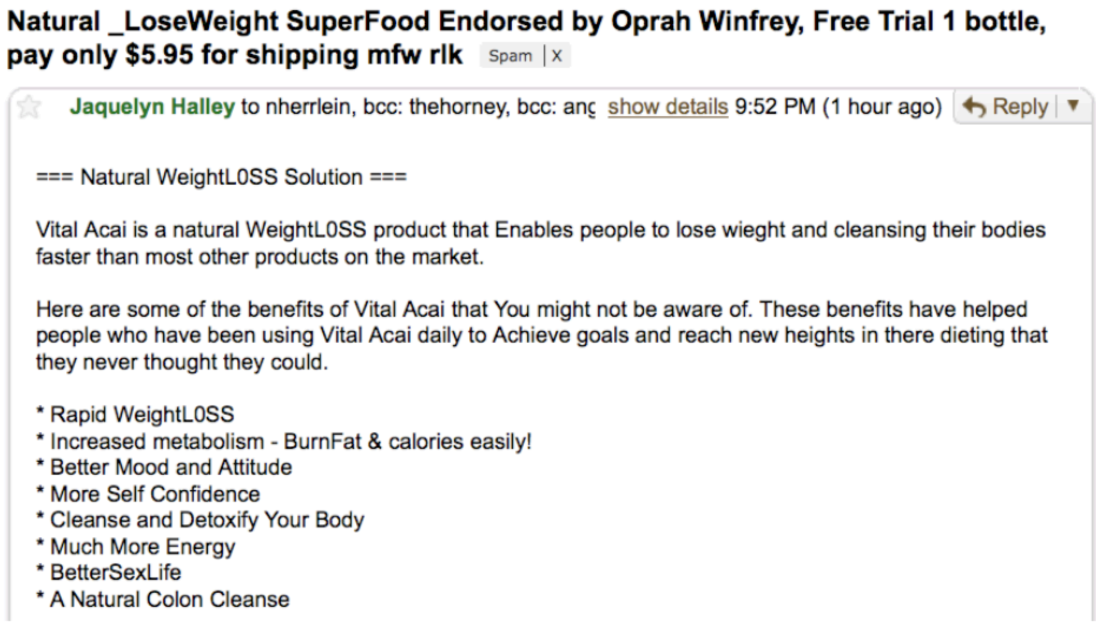
Carlos Guestrin to 10615-announce, Osman, Michel — show details 3:15 PM (8 hours ago) — Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang — show details 9:52 PM (1 hour ago) — Reply

=== Natural WeightL0SS Solution ===

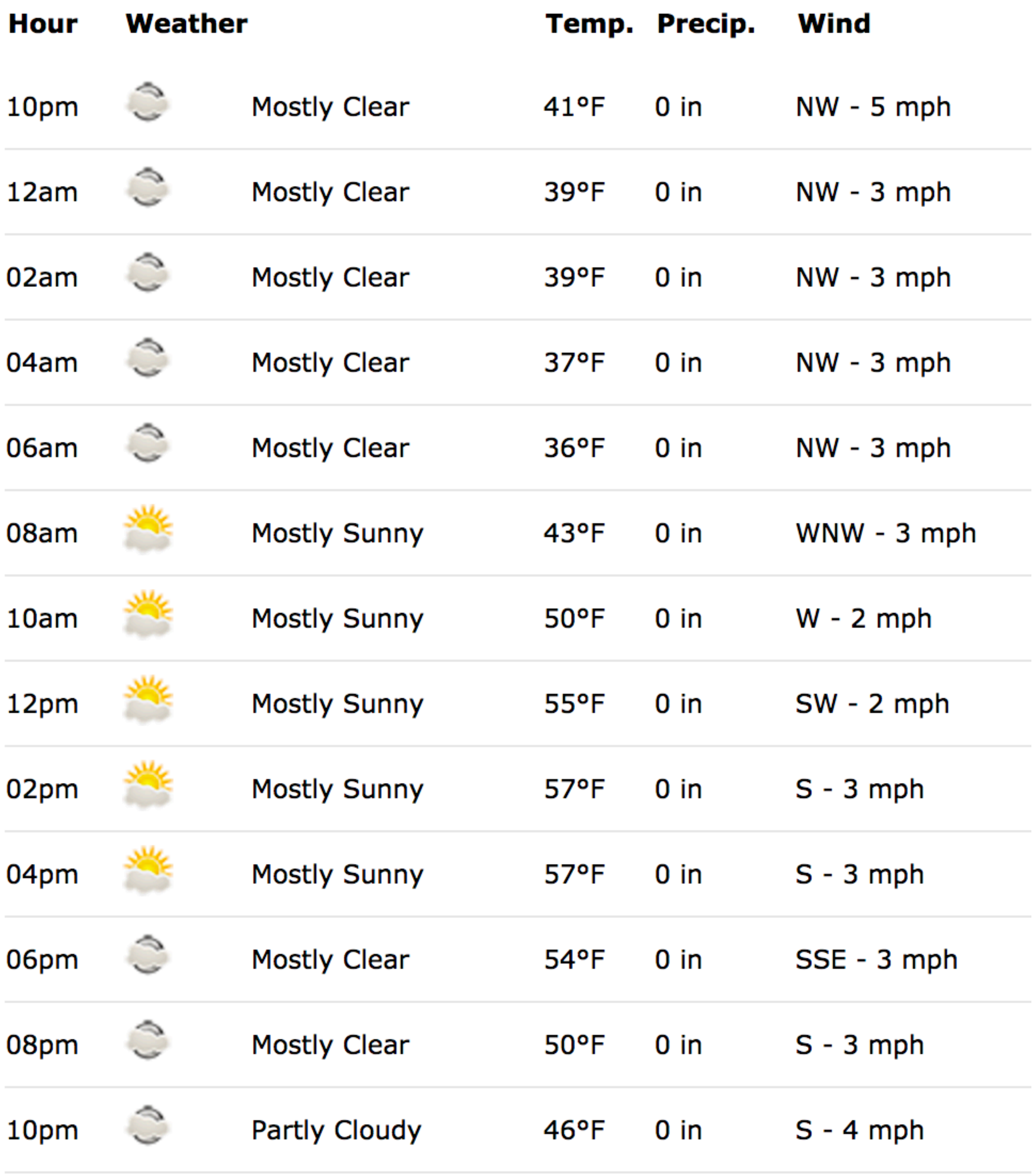
Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

spam
vs
not spam

Linear Classifiers: Weather Prediction



Hour	Weather		Temp.	Precip.	Wind
10pm	🌙	Mostly Clear	41°F	0 in	NW - 5 mph
12am	🌙	Mostly Clear	39°F	0 in	NW - 3 mph
02am	🌙	Mostly Clear	39°F	0 in	NW - 3 mph
04am	🌙	Mostly Clear	37°F	0 in	NW - 3 mph
06am	🌙	Mostly Clear	36°F	0 in	NW - 3 mph
08am	🌤	Mostly Sunny	43°F	0 in	WNW - 3 mph
10am	🌤	Mostly Sunny	50°F	0 in	W - 2 mph
12pm	🌤	Mostly Sunny	55°F	0 in	SW - 2 mph
02pm	🌤	Mostly Sunny	57°F	0 in	S - 3 mph
04pm	🌤	Mostly Sunny	57°F	0 in	S - 3 mph
06pm	🌙	Mostly Clear	54°F	0 in	SSE - 3 mph
08pm	🌙	Mostly Clear	50°F	0 in	S - 3 mph
10pm	🌙	Partly Cloudy	46°F	0 in	S - 4 mph

Learning Theory

- How can we gauge the accuracy of a hypothesis on unseen data?

- How do we quantify our ability to ***generalize*** as a function of the amount of training data and the hypothesis space?

- How do we find the best hypothesis?

Occam's Razor Principle

- William of Occam: Monk living in the 14th century
- Principle of parsimony:
  "One should not increase, beyond what is necessary, the number of entities required to explain anything"
- When many solutions are available for a given problem, we should select simplest one

But what does simplest really mean?

Samy Bengio

Validation

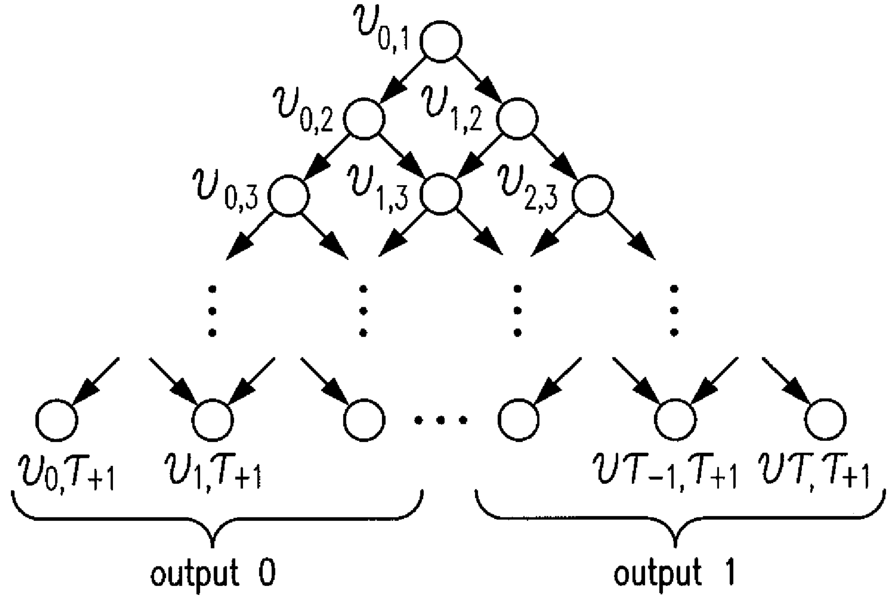
- How do we objectively measure performance of ML algorithms?

- How do we select the best algorithms and parameter values?

- Probably the most important topic — most abused and neglected

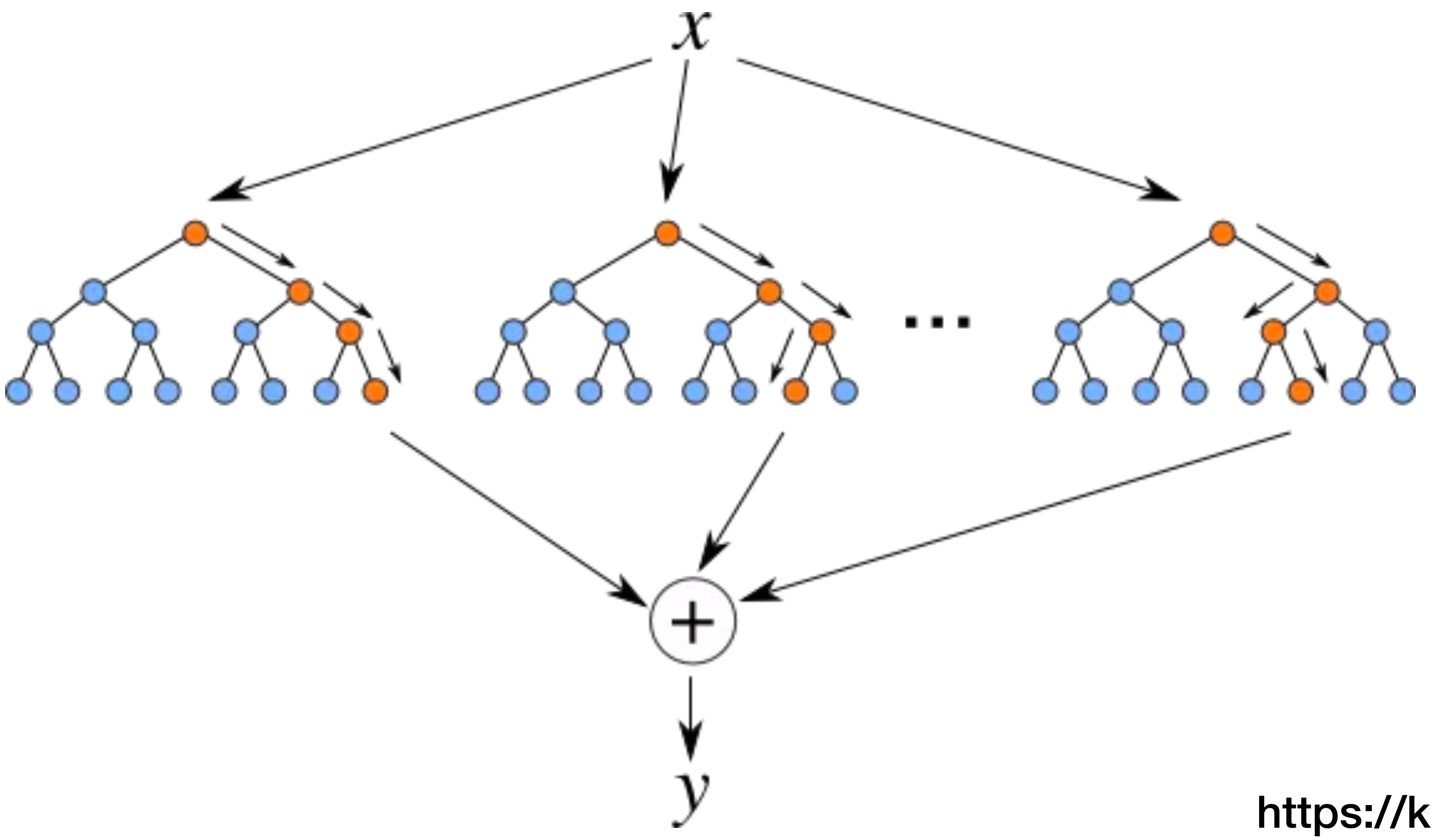
Boosting, Trees, & Additive Models

- Can we achieve good performance by combining many primitive models?
- Can we build a strong learning from many weak learners?



Ensembles & Random Forests

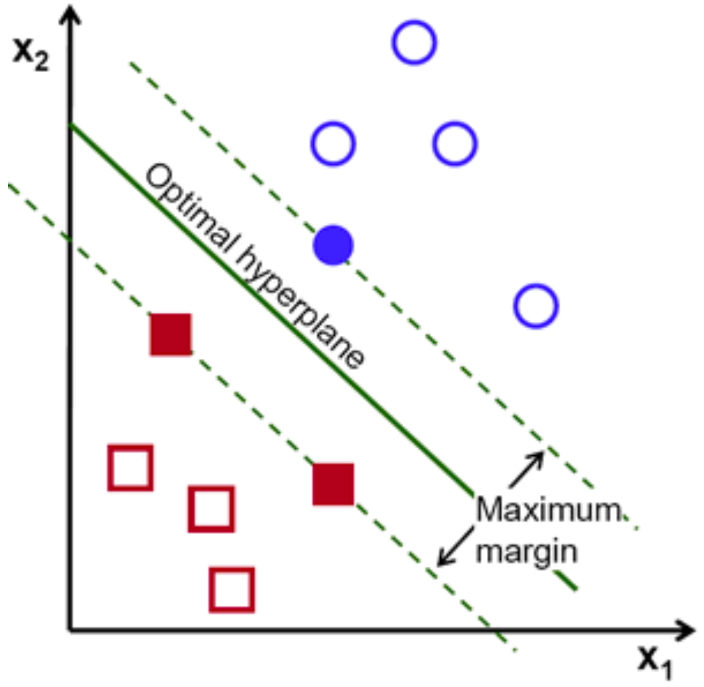
- One of the most popular methods around
- Combining trees in a special way to get powerful classifier



https://kgpdag.wordpress.com/

Support Vector Machines

- Widely used classifier
- How can we implement non-linear classifiers by automatically transforming the data?

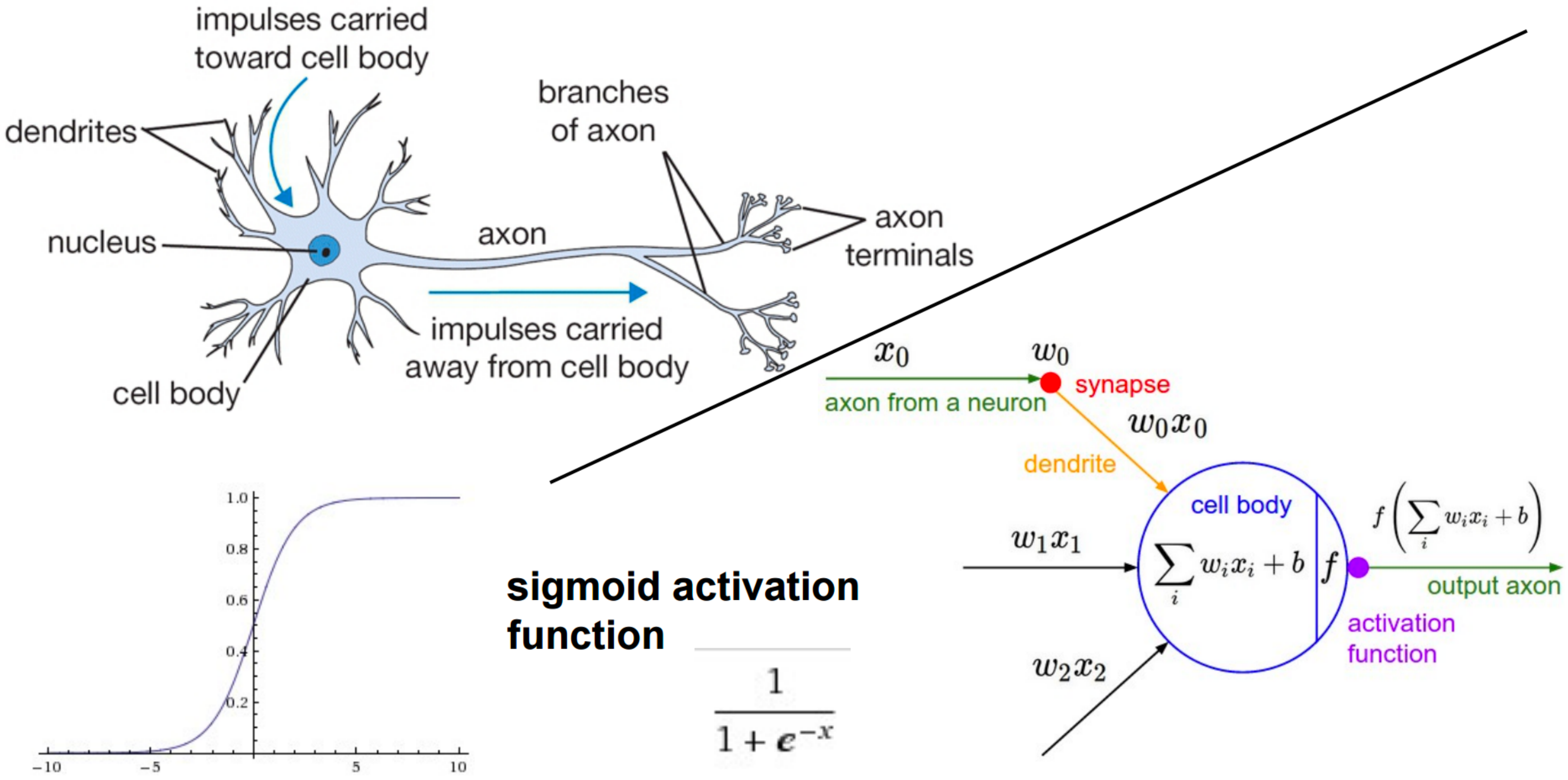


http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

Neural Networks

- Can we use biology as an inspiration for classification?

- Connect virtual neurons in “natural” architectures and train to make decisions



sigmoid activation function

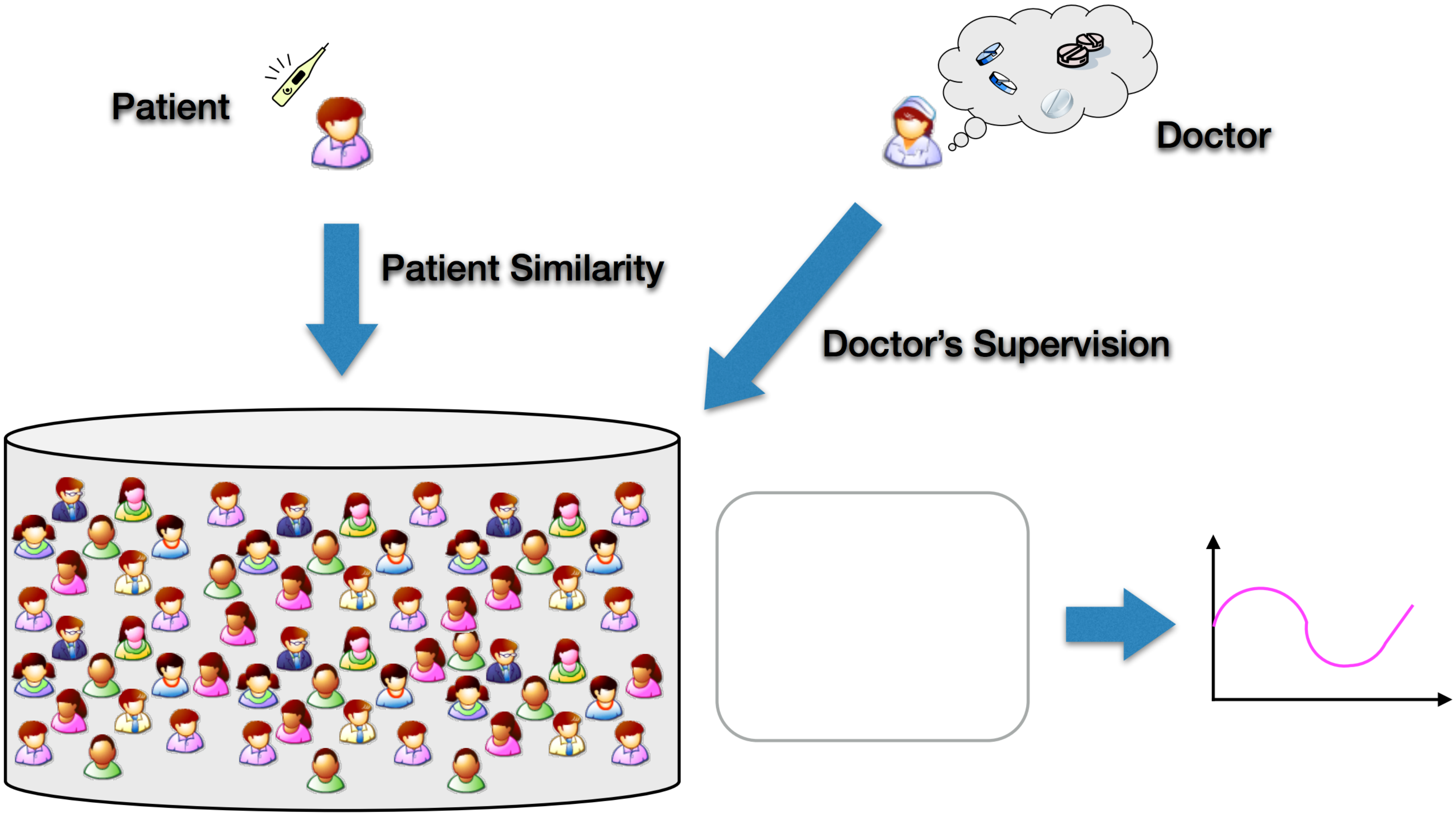
$$\frac{1}{1+e^{-x}}$$

http://vision.stanford.edu/teaching/cs231n/slides/lecture5.pdf

Prototype Methods

- What if we don’t have an underlying model of how the data is shaped?
- Can we use the relationships between data points to develop good classifiers?

Prototype Methods: kNN



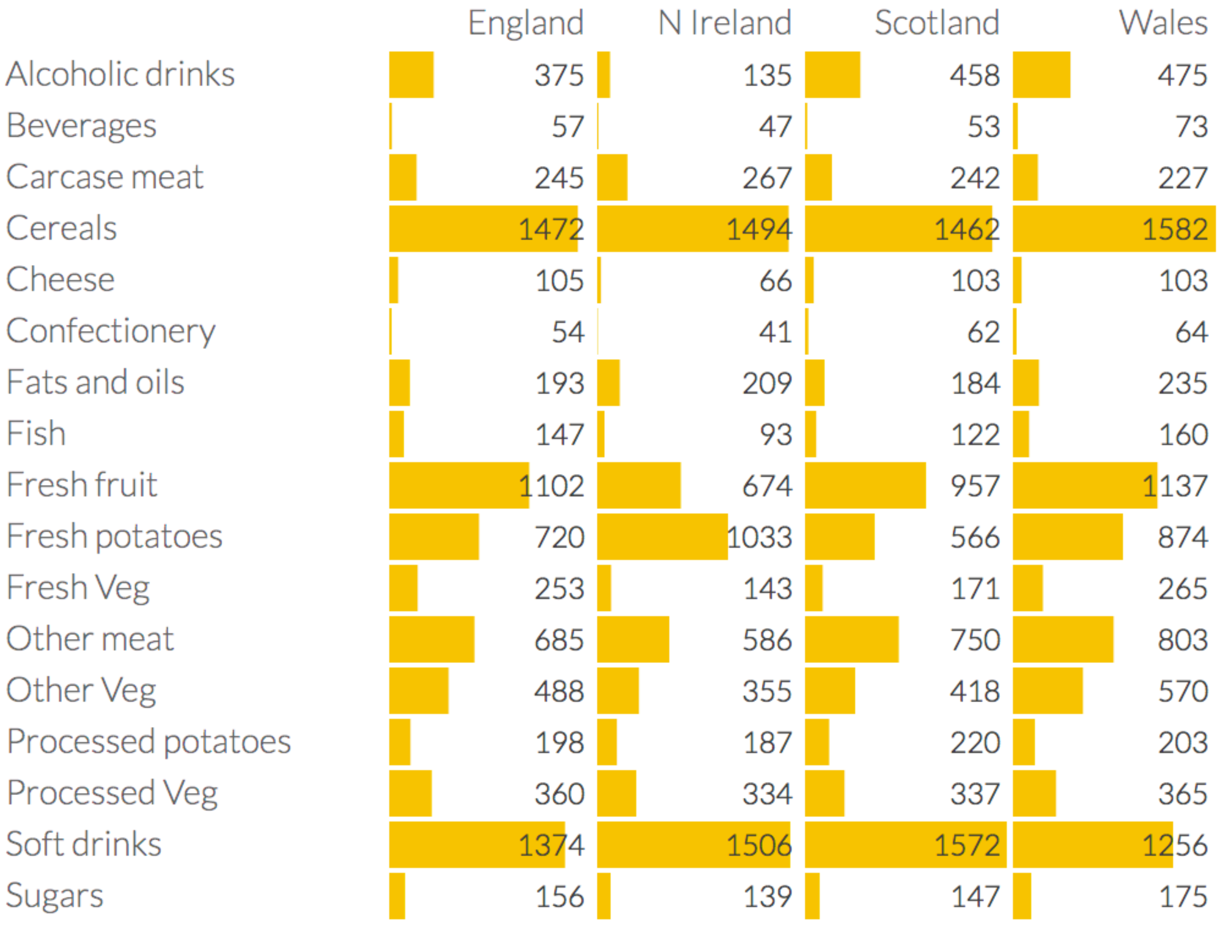
http://engr.uconn.edu/~fwang/tutorials/CIKM14_Tutorial.pdf

Unsupervised Learning

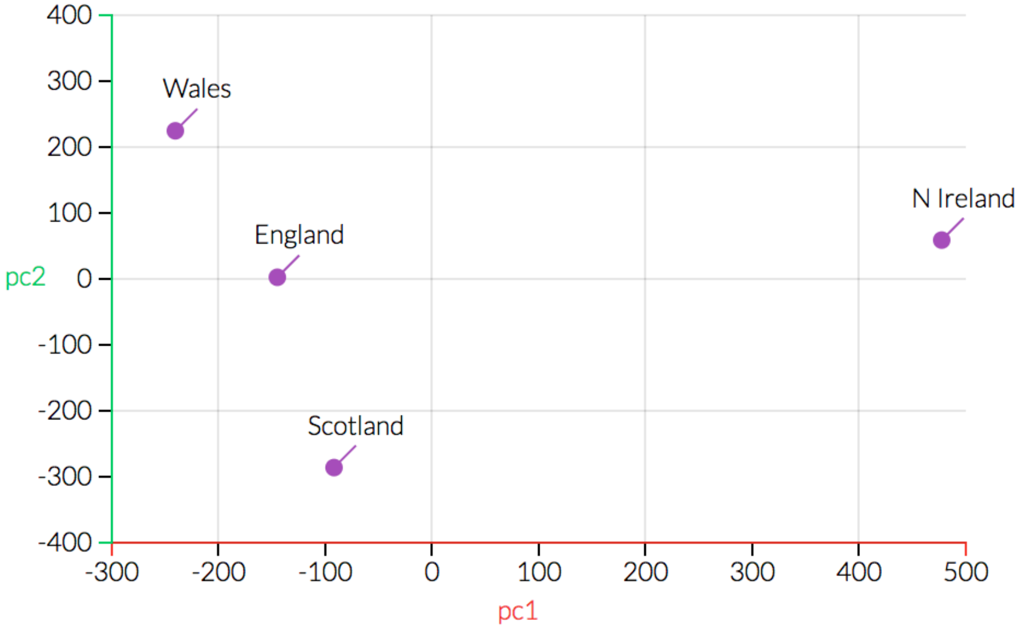
- What happens when we don’t know the outcome or have classes?

- How to explore data to look for structure and patterns?

Unsupervised Learning: Visualization

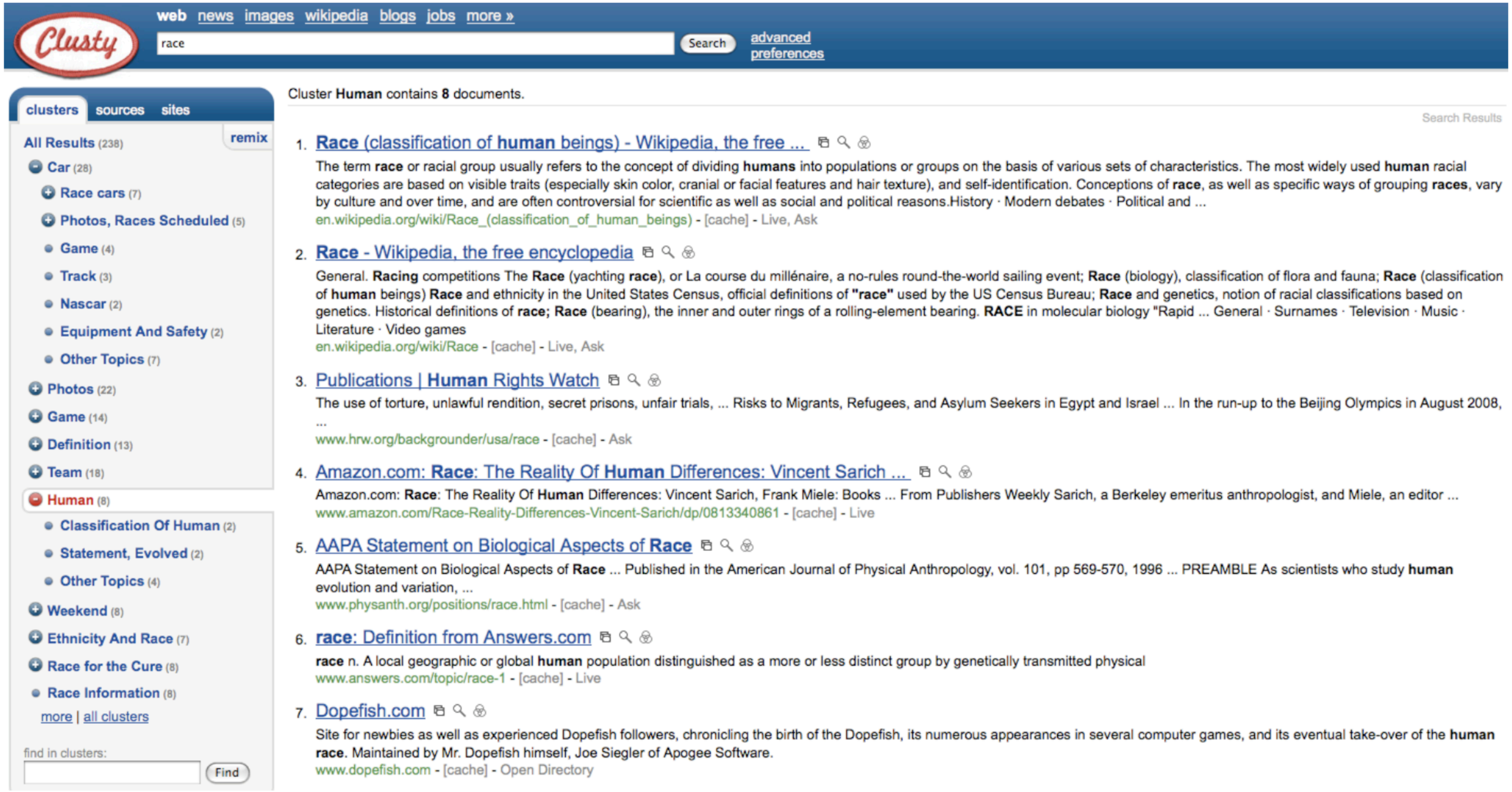


	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



http://setosa.io/ev/principal-component-analysis/

Unsupervised Learning: Clustering



- All Results (238)
  - Car (28)
    - Race cars (7)
    - Photos, Races Scheduled (5)
    - Game (4)
    - Track (3)
    - Nascar (2)
    - Equipment And Safety (2)
    - Other Topics (7)
  - Photos (22)
  - Game (14)
  - Definition (13)
  - Team (18)
  - Human (8)
    - Classification Of Human (2)
    - Statement, Evolved (2)
    - Other Topics (4)
  - Weekend (8)
  - Ethnicity And Race (7)
  - Race for the Cure (8)
  - Race Information (8)
  - more | all clusters

Cluster **Human** contains **8** documents.

Search Results

1. **Race** (classification of **human** beings) - Wikipedia, the free ...
   The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **races**, vary by culture and over time, and are often controversial for scientific as well as social and political reasons.History · Modern debates · Political and ...
   en.wikipedia.org/wiki/Race_(classification_of_human_beings) - [cache] - Live, Ask

2. **Race** - Wikipedia, the free encyclopedia
   General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of **human** beings) **Race** and ethnicity in the United States Census, official definitions of **"race"** used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games
   en.wikipedia.org/wiki/Race - [cache] - Live, Ask

3. Publications | **Human** Rights Watch
   The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
   www.hrw.org/backgrounder/usa/race - [cache] - Ask

4. Amazon.com: **Race**: The Reality Of **Human** Differences: Vincent Sarich ...
   Amazon.com: **Race**: The Reality Of **Human** Differences: Vincent Sarich, Frank Miele: Books ... From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
   www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 - [cache] - Live

5. AAPA Statement on Biological Aspects of **Race**
   AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study **human** evolution and variation, ...
   www.physanth.org/positions/race.html - [cache] - Ask

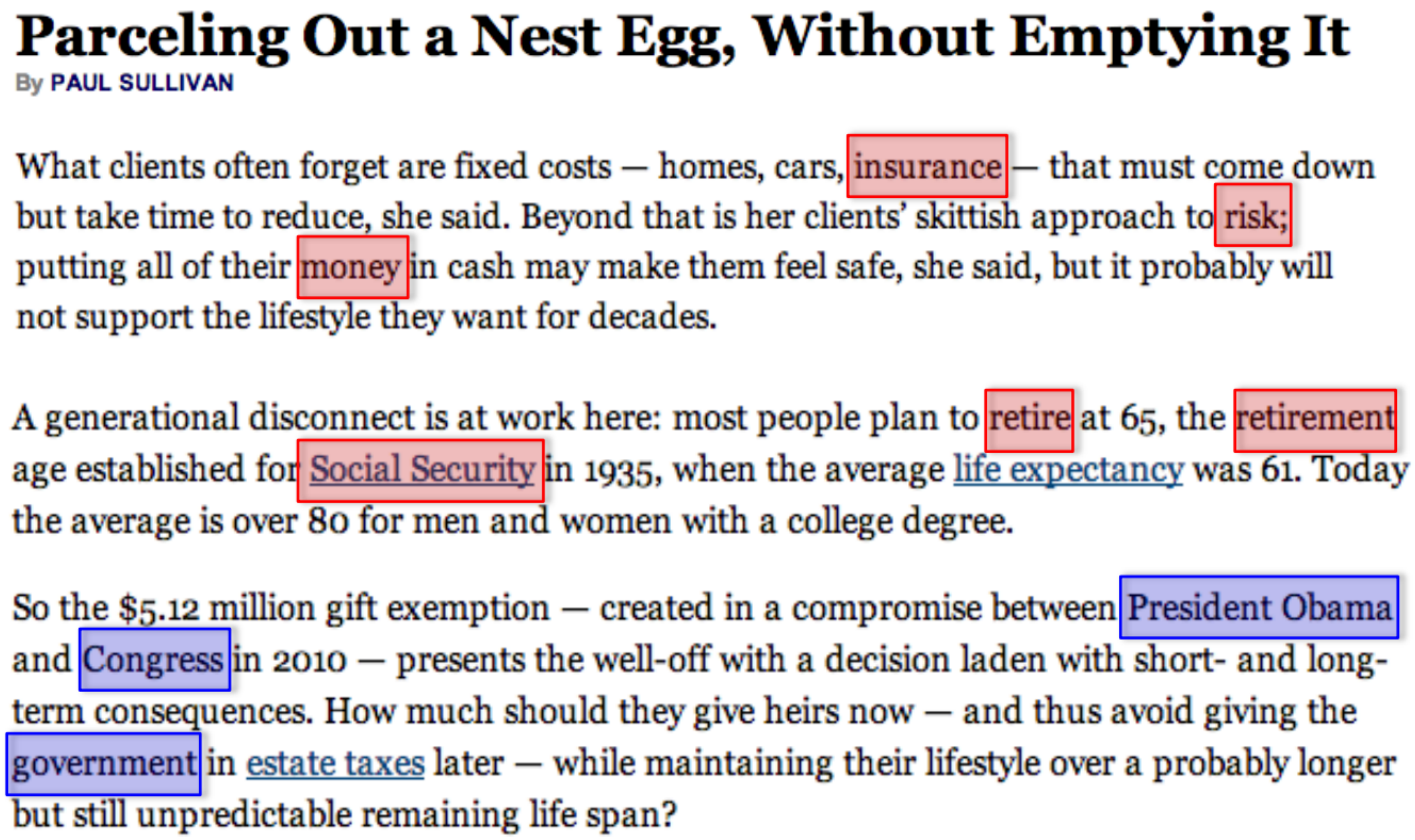
6. **race**: Definition from Answers.com
   race n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically transmitted physical
   www.answers.com/topic/race-1 - [cache] - Live

7. Dopefish.com
   Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
   www.dopefish.com - [cache] - Open Directory

Unsupervised Learning: Topic Models

Personal Finance: (money, 0.15), (retire, 0.10), (risk, 0.03) …

Politics: (President Obama, 0.10), (congress, 0.08), (government, 0.07), …



Parceling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

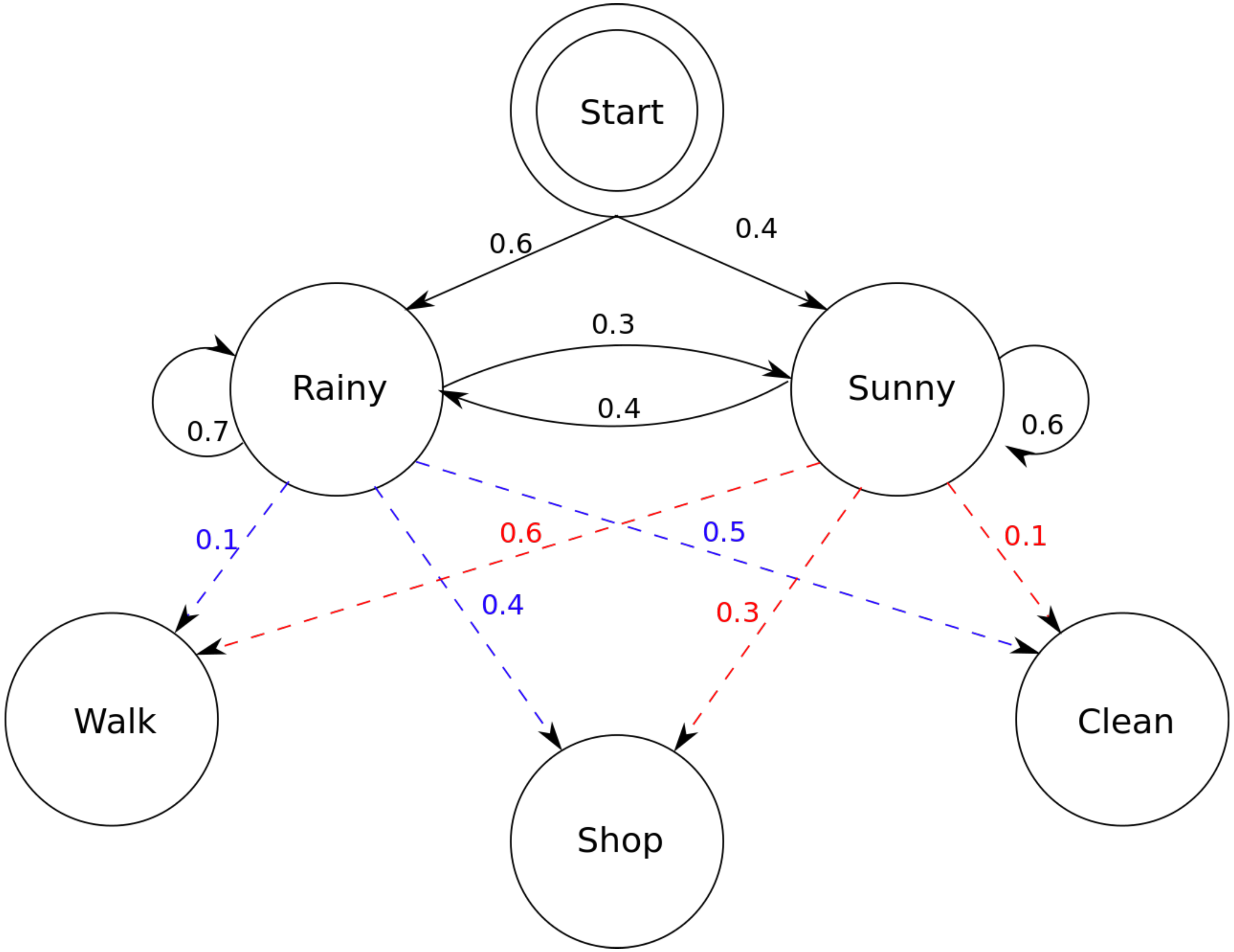
What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients' skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between President Obama and Congress in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government in estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?

Graphical Models

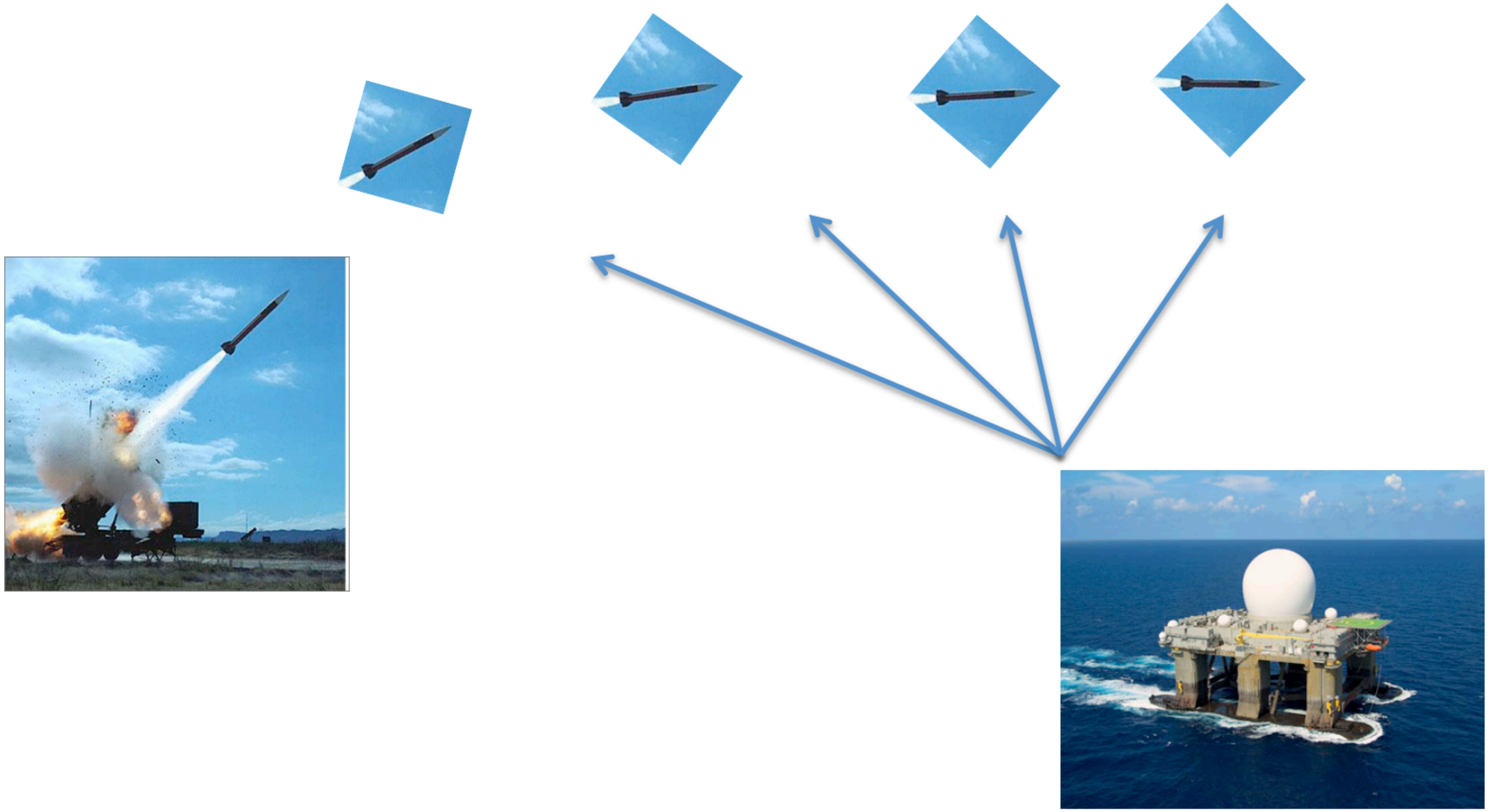
- Marriage between graph theory and probability theory
- Great for modeling sequential data (e.g., time series, speech processing)



https://en.wikipedia.org/wiki/Hidden_Markov_model

Graphical Models: Tracking

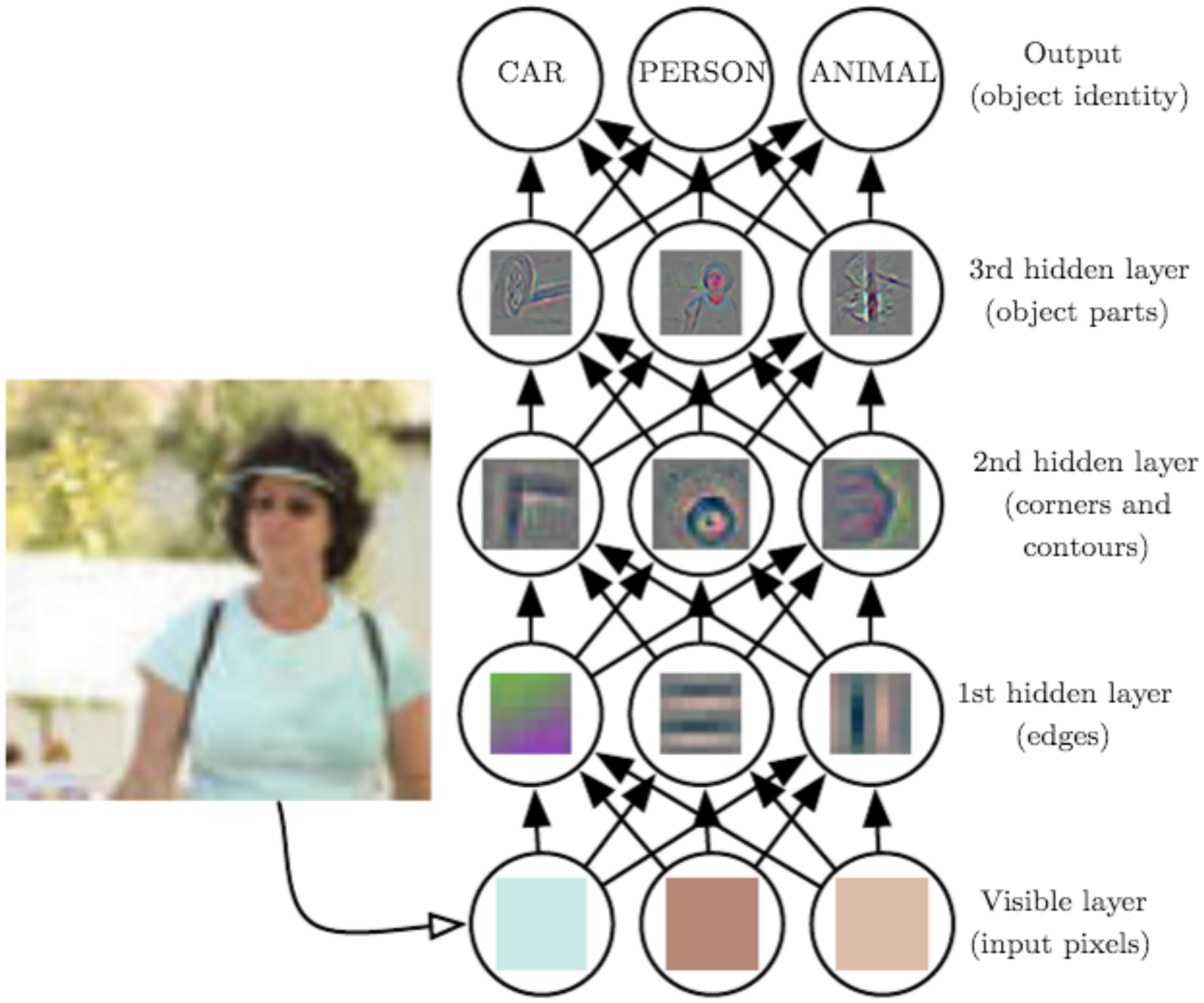
Observe noisy measurements of missile location



Where is the missile now? Where will it be in 1 minute?

Deep Learning

- Form of representation learning
- Aimed at learning feature hierarchies
- Features from higher levels of the hierarchy are formed by lower level features
- Each hidden layer allows for more complex features of input



http://www.deeplearningbook.org/contents/intro.html

Recommendation Systems

How to build a system that provides or suggests items to the end users?

iPython Setup

Jupyter [iPython] Notebook

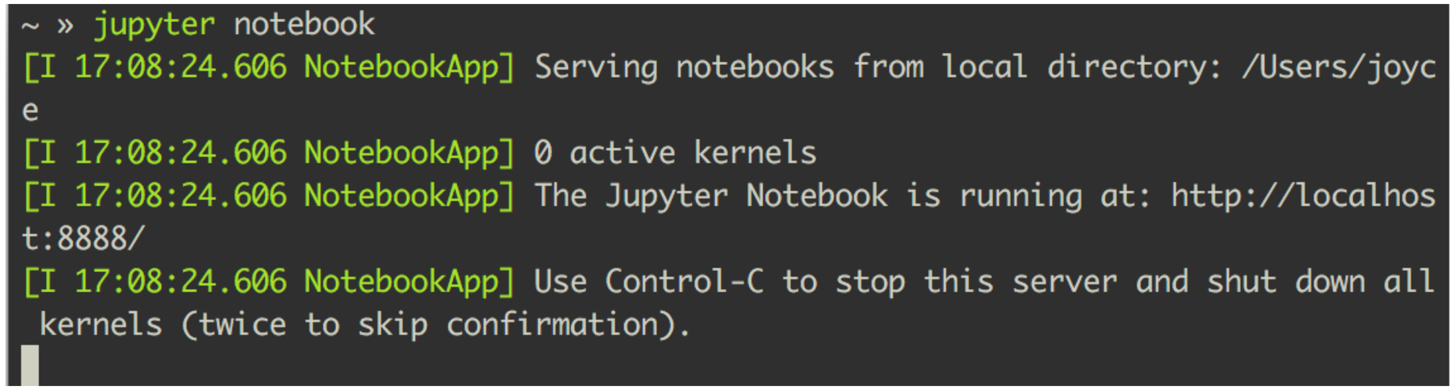
- Interactive computational environment which save output in a nice notebook format
  - Combine code execution, rich text, math, plots, and may other things
  - Supports markdown, LaTeX, HTML, etc
  - Popular in Data Science and can be easily shared with others
- More information: http://jupyter.org/

Jupyter Notebook Setup

- Suggestion: Anaconda, an open source data science platform powered by Python: https://www.continuum.io/downloads
  - Contains Windows, OS X, Linux installations
  - Use either Python 2.7 or Python 3.5
    (Note: class will use 2.7 syntax primarily)

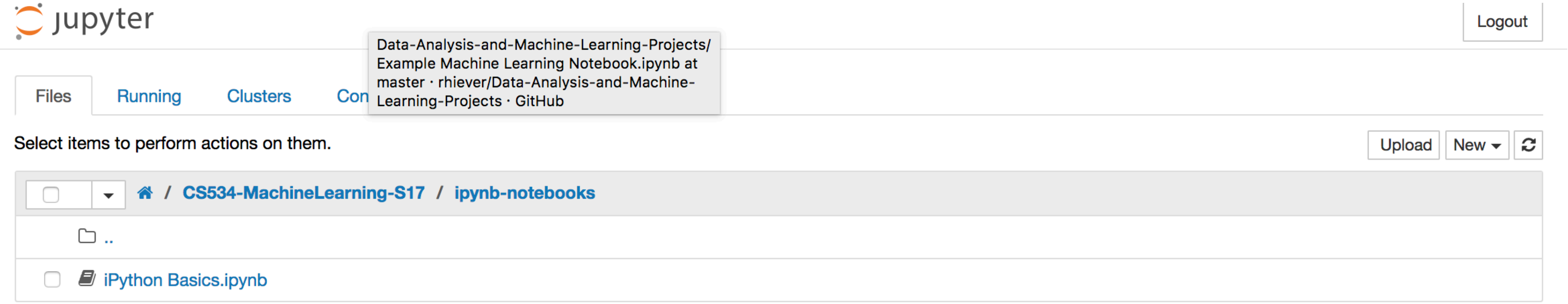
Jupyter Notebook Setup (2)

- Once installed, to start, just open a terminal and run jupyter notebook



```
~ » jupyter notebook
[I 17:08:24.606 NotebookApp] Serving notebooks from local directory: /Users/joyc
e
[I 17:08:24.606 NotebookApp] 0 active kernels
[I 17:08:24.606 NotebookApp] The Jupyter Notebook is running at: http://localhos
t:8888/
[I 17:08:24.606 NotebookApp] Use Control-C to stop this server and shut down all
 kernels (twice to skip confirmation).
```

- A browser should open with jupyter running and you can import the .ipynb notebook from today's activity



Data-Analysis-and-Machine-Learning-Projects/Example Machine Learning Notebook.ipynb at master · rhiever/Data-Analysis-and-Machine-Learning-Projects · GitHub

Select items to perform actions on them.

/ CS534-MachineLearning-S17 / ipynb-notebooks

..

□ iPython Basics.ipynb