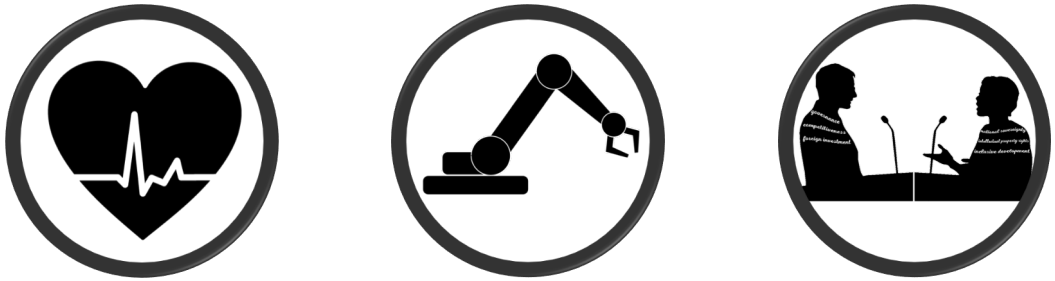
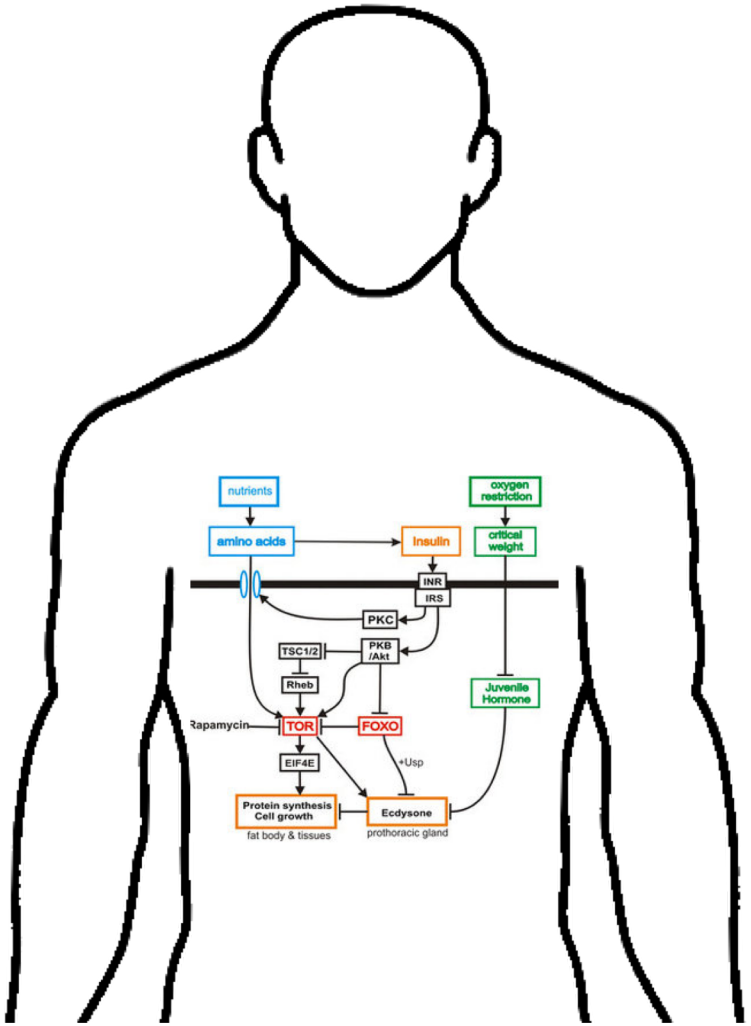


Solving World Problems

Learning One Example at a Time

Problem: Signaling Networks



To treat/cure disease, we need to understand the chains of protein interactions that determine how cells process signals from their environment

Problem: Robotic Interaction

To safely operate in distant/dangerous terrain, we need robotic teammates that can autonomously perceive, understand, interact with, and manipulate their environment

Problem: Winning on the Merits

To enact the best policies, we need to understand what makes for a strong argument

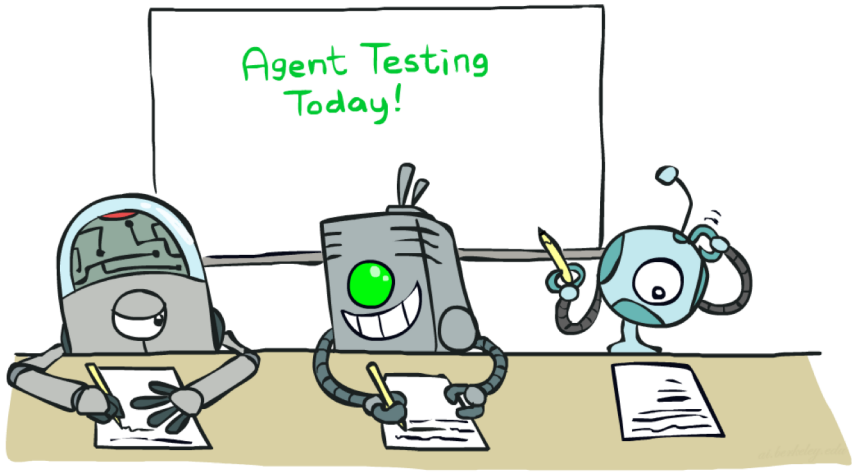
Solving World Problems

The common link in solving these and many other complex problems is the potential for using **Machine Learning**, which exploits…

- Big Data: examples from the world
- Cloud Computation: cheap, fast processing
- Algorithms: automatically uncovering value

Machine Learning

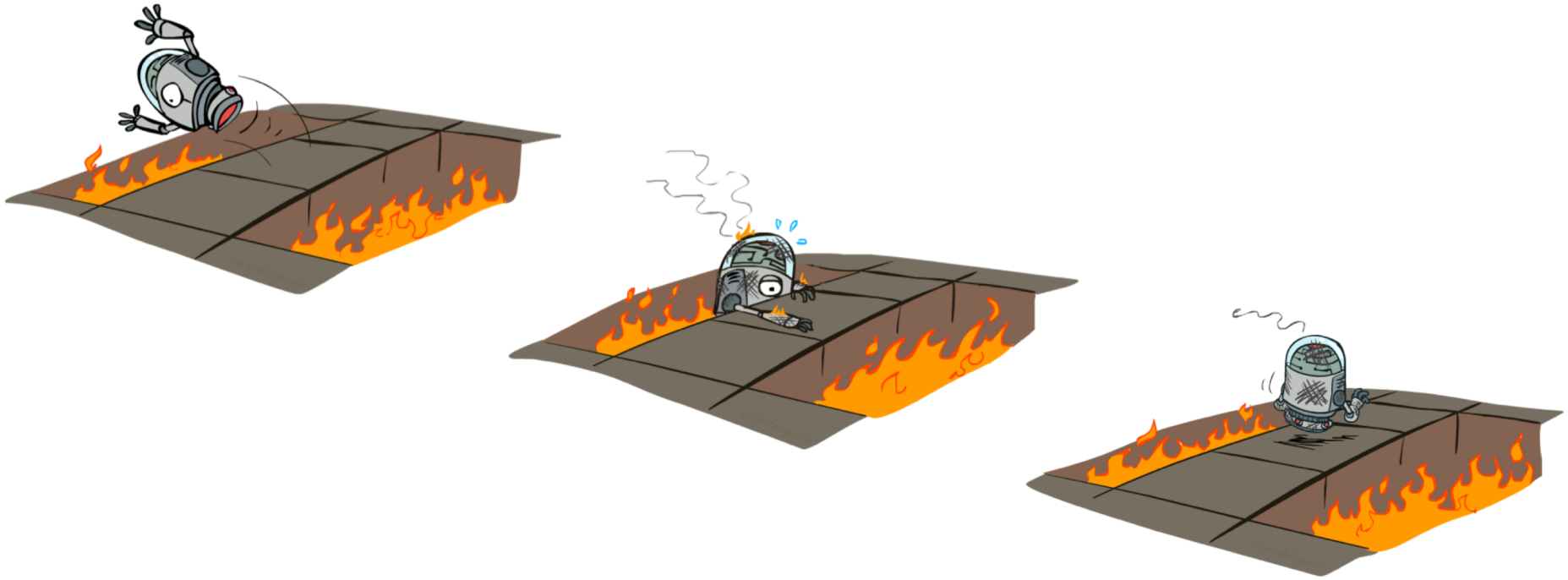
1. What is it?
   – Why you should care!
2. How does it work?
3. You + ML



http://ai.berkeley.edu

What is Machine Learning?

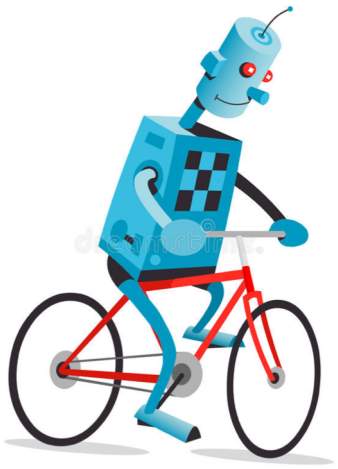
Computer programs that can improve performance with experience



But Wait…

Why Learn?

Many complex tasks are hard to describe, but easy to learn from experience



Why Now?

Data sources and powerful computing are increasingly cheap and plentiful

Natural Language Processing (NLP)

Modern NLP algorithms are typically based on statistical ML

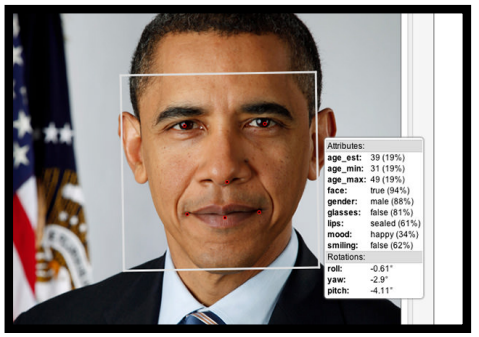
Applications

- Summarization
- Machine Translation
- Speech Processing
- Sentiment Analysis

...

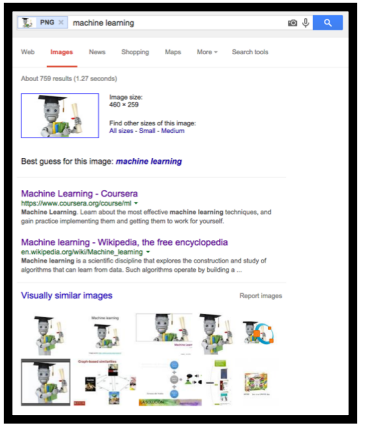
Computer Vision

Methods for acquiring, processing, analyzing, and understanding images

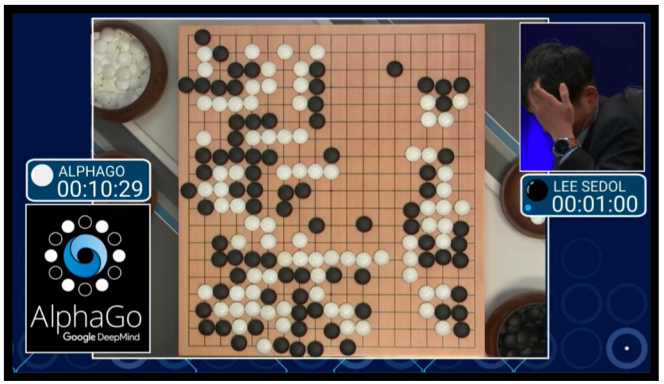


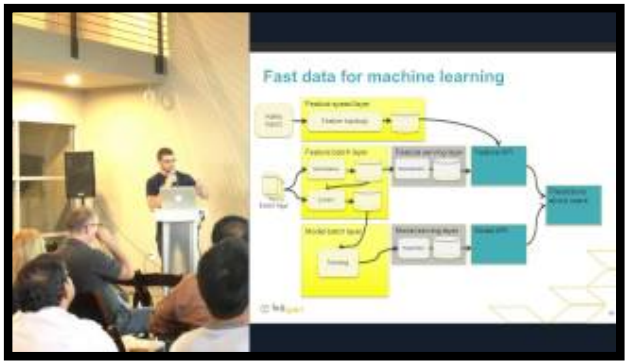
Applications

- Image search
- Facial recognition
- Object tracking
- Image restoration
- …

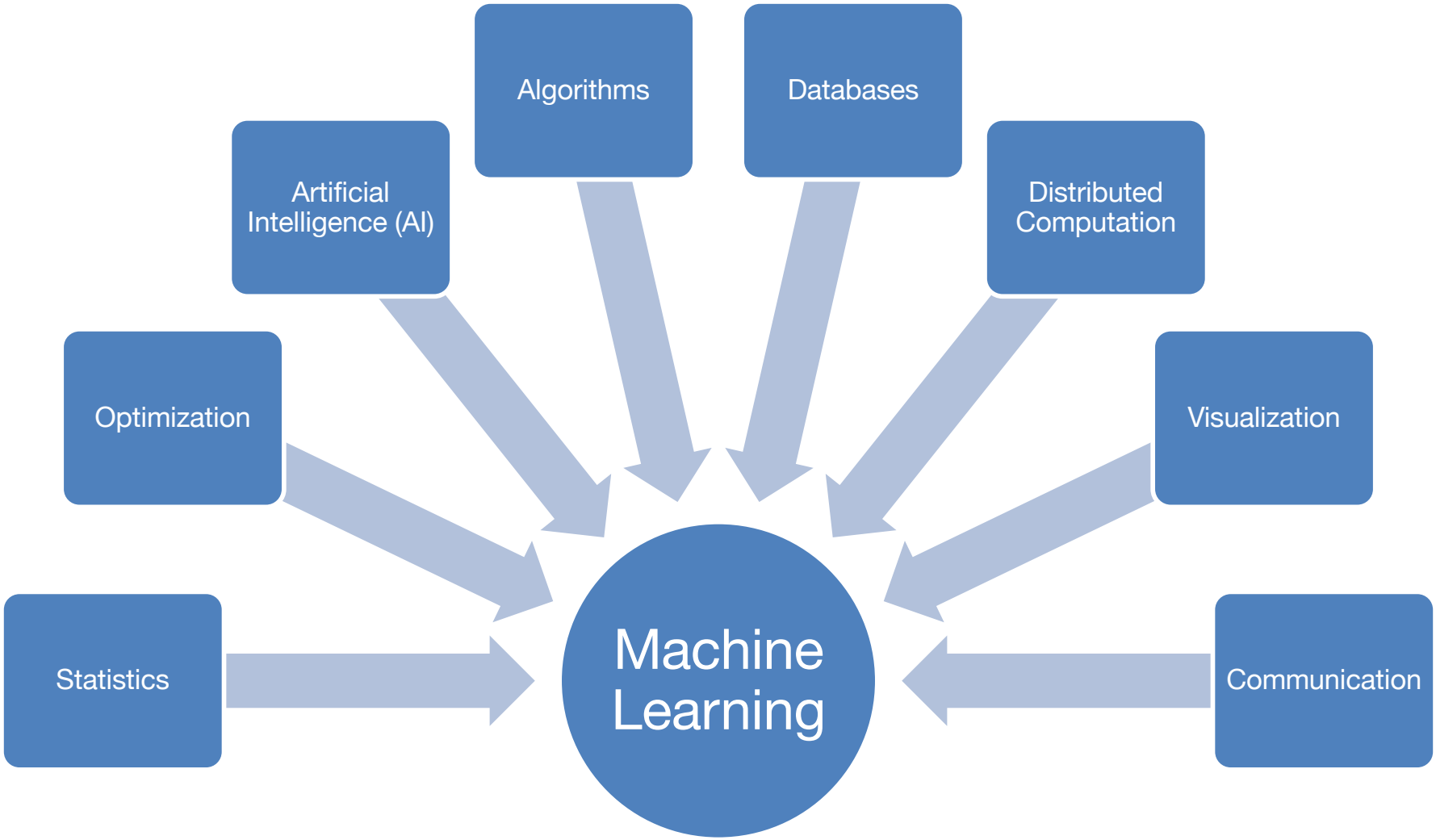


Games, Robotics, Medicine, Ads, …

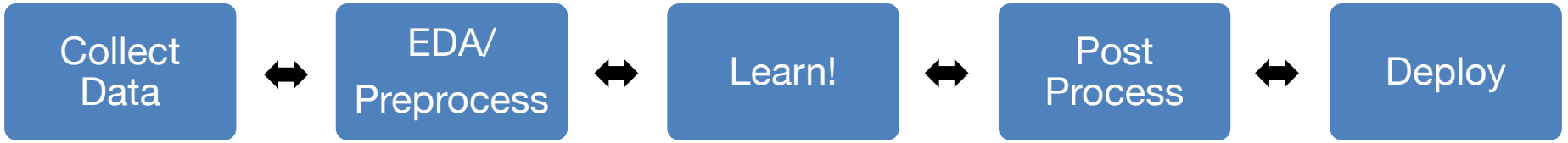




Fusing Disciplines



ML/DS Pipeline



API
Sensors
DBMS query

Association
Classification
Clustering
Outlier detection

Pipelining
Scaling

Integration
Normalization
Feature selection
Binning
Dimension red.

Evaluation
Interpretation
Visualization

Jobs!

Position	Salary
Data Scientist	$117,345
Software Engineer	$103,035

The New York Times

TECHNOLOGY

A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit

By CADE METZ APRIL 19, 2018

Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent

Nearly all big tech compa
intelligence project, and t
experts millions of dollar

查看简体中文版 | 查看繁體中

By CADE METZ

Typical A.I. specialists, including both Ph.D.s fresh out of school and people with less education and just a few years of experience, can be paid from $300,000 to $500,000 a year or more in salary and company stock, according to nine people who work for major tech companies or have entertained job offers from them. All of them requested anonymity because they did not want to damage their professional prospects.

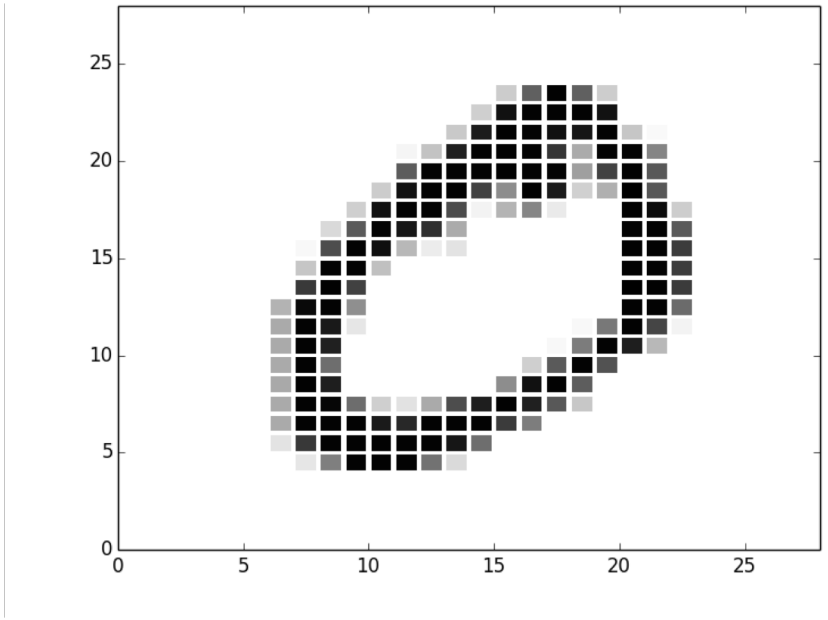
"Software Is Eating the World, but AI Is Going to Eat Software"
- Jensen Huang (CEO, NVIDIA)

*glassdoor.com, USA National Avg as of April 4, 2019

Machine Learning Tasks

- ***Supervised***
  - Given a dataset of input-output pairs, learn a function that maps future (novel) inputs to appropriate outputs
- ***Unsupervised***
  - Given a dataset and a hypothesis, find interesting patterns/parameters
- ***Reinforcement***
  - Learn an optional action ***policy*** over time; given an environment that provides states, affords actions, and provides feedback as numerical ***reward***, maximize the *expected* future reward

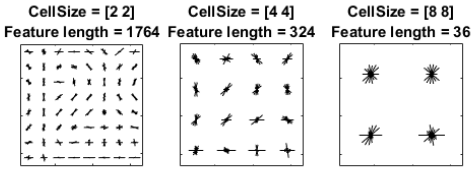
ML Terminology



example, instance
Unit of input

Composed of ***features***
(or ***attributes***)

- In this case, we could represent each digit via raw pixels: 28x28=784-pixel ***vector*** of greyscale values [0-255]
  - ***Dimensionality***: number of features per instance (|vector|)
- But other ***data representations*** are possible, and might be advantageous



- In general, the problem of ***feature selection*** is challenging

Instances/Features = Table

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Instance

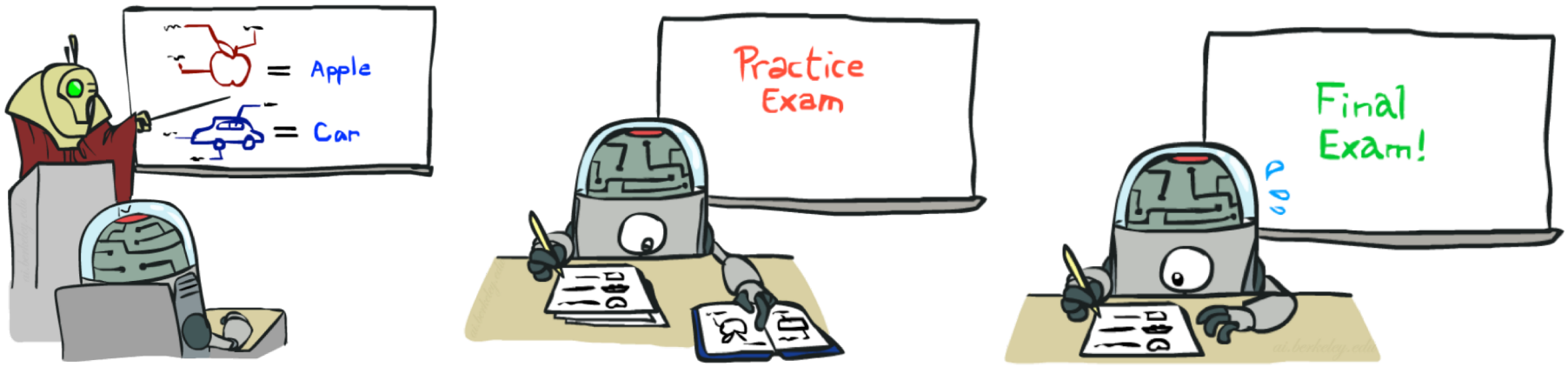
"Target" Feature

When trying to predict a particular feature given the others

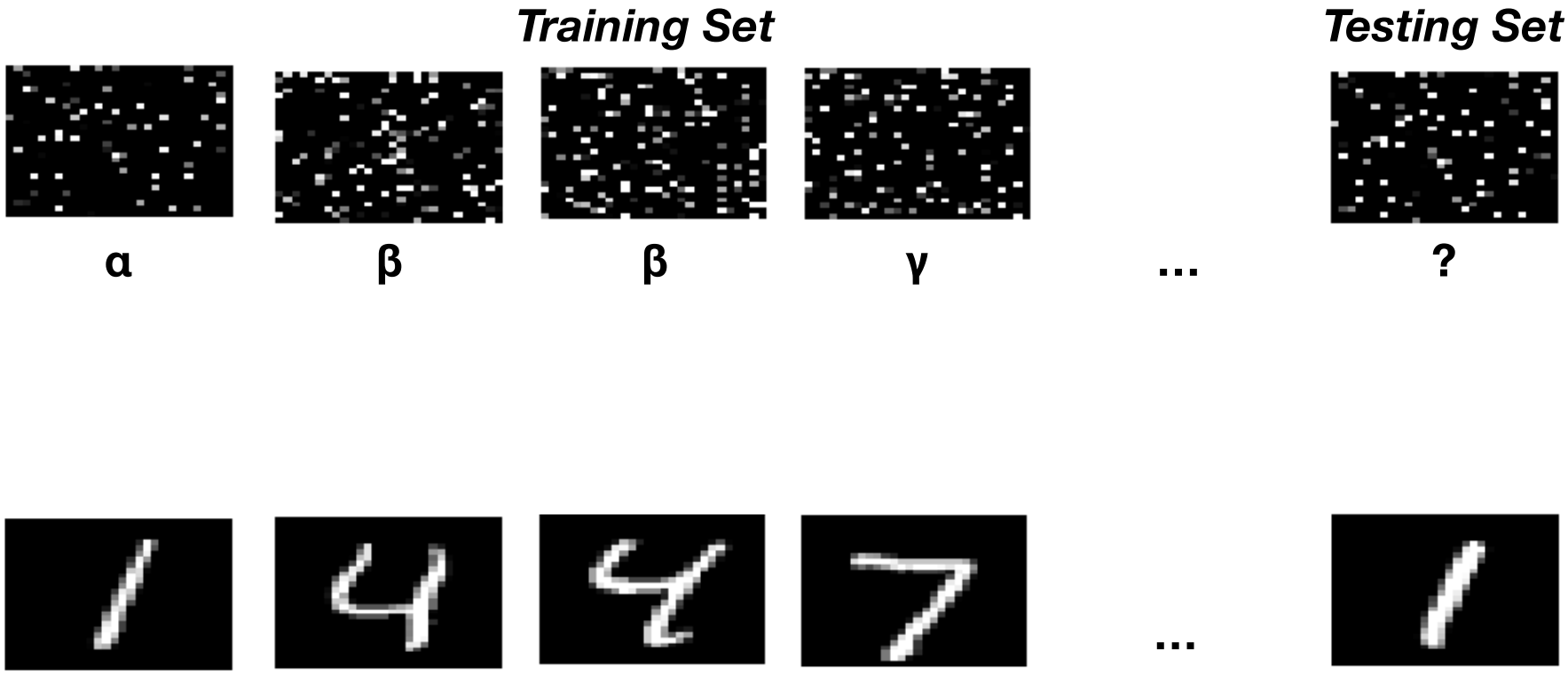
target*, *label*, *class*, *concept, dependent

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Supervised Learning



Supervised Learning in Action



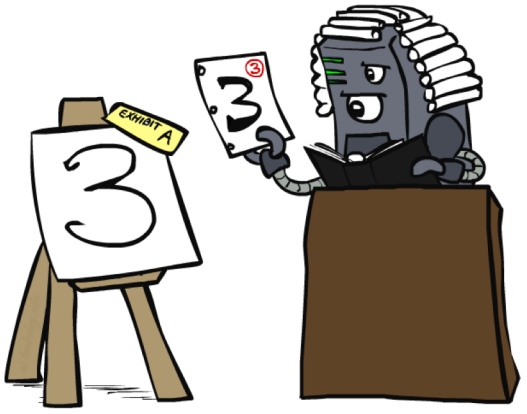
A Simple ML Technique: **kNN**

Training

- Store all examples

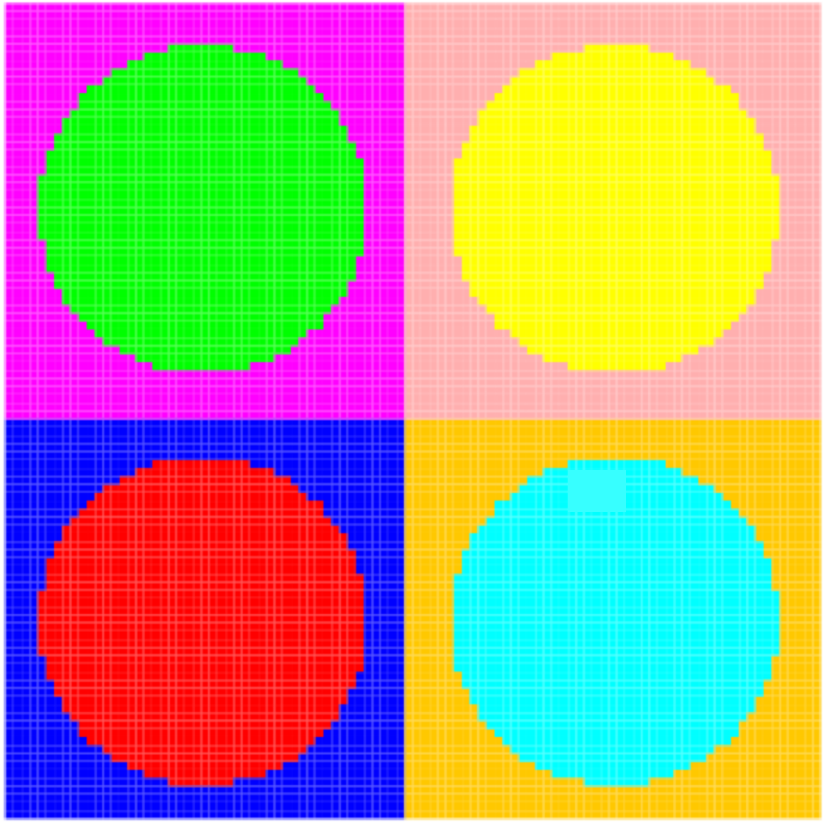
Testing

- Find the *k* nearest neighbors to input
- Vote on output

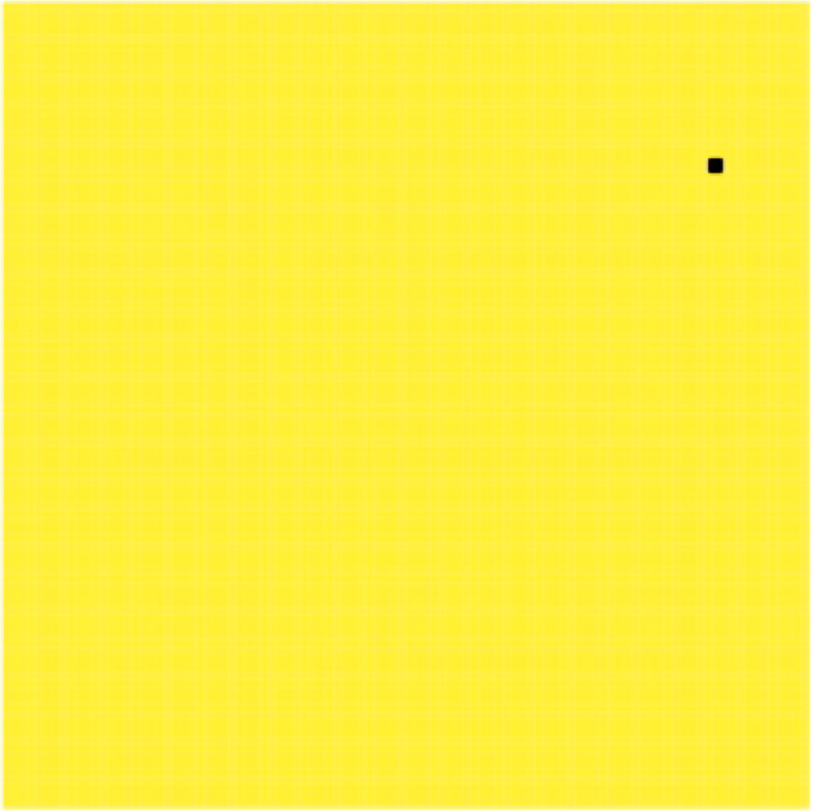


2D Multiclass Classification

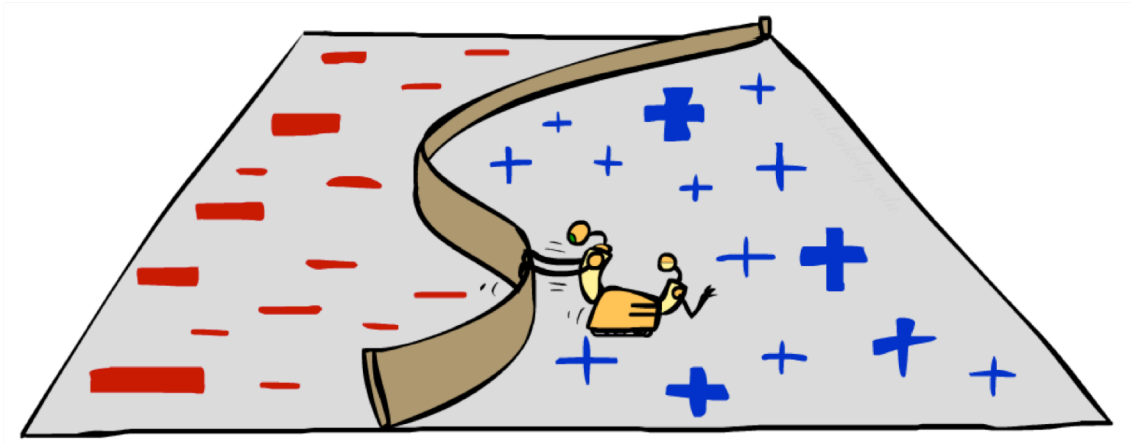
Boundary Tree

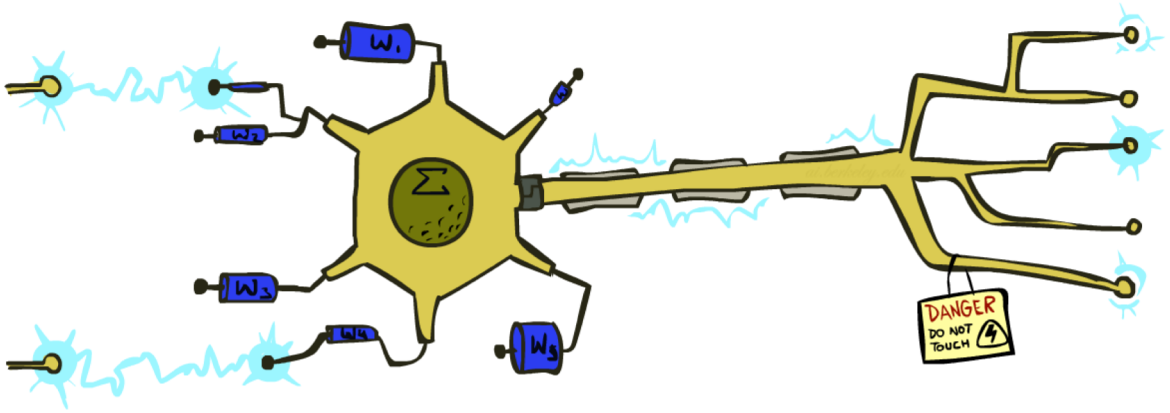


1-NN via Linear Scan



Many Approaches





Supervised Tasks (1)

Classification

- Discrete target
- Binary vs. multi-class



SepalLength	SepalWidth	PetalLength	PetalWidth	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa

Supervised Tasks (2)

Regression

- Continuous target

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
14	8	454	220	4354	9	70	1	chevrolet impala
14	8	440	215	4312	8.5	70	1	plymouth fury iii
14	8	455	225	4425	10	70	1	pontiac catalina
15	8	390	190	3850	8.5	70	1	amc ambassador dpl
15	8	383	170	3563	10	70	1	dodge challenger se
14	8	340	160	3609	8	70	1	plymouth 'cuda 340
15	8	400	150	3761	9.5	70	1	chevrolet monte carlo
14	8	455	225	3086	10	70	1	buick estate wagon (sw)
24	4	113	95	2372	15	70	3	toyota corona mark ii
22	6	198	95	2833	15.5	70	1	plymouth duster
18	6	199	97	2774	15.5	70	1	amc hornet
21	6	200	85	2587	16	70	1	ford maverick
27	4	97	88	2130	14.5	70	3	datsun pl510
26	4	97	46	1835	20.5	70	2	volkswagen 1131 deluxe sedan
25	4	110	87	2672	17.5	70	2	peugeot 504
24	4	107	90	2430	14.5	70	2	audi 100 ls
25	4	104	95	2375	17.5	70	2	saab 99e
26	4	121	113	2234	12.5	70	2	bmw 2002

ML in Python via `scikit-learn`

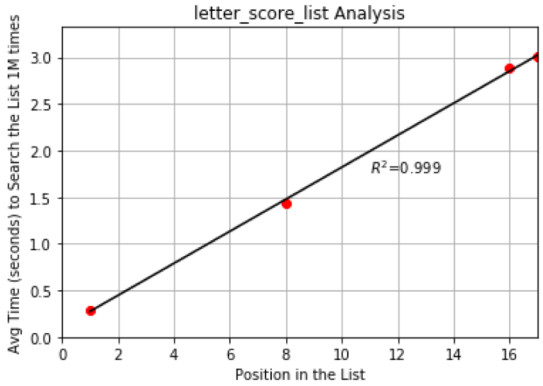
```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(model_x_values, list_times)

print(model.coef_)
print(model.intercept_)
print(model.score(model_x_values, list_times))

print(fit.predict([[x] for x in all_positions]))
```

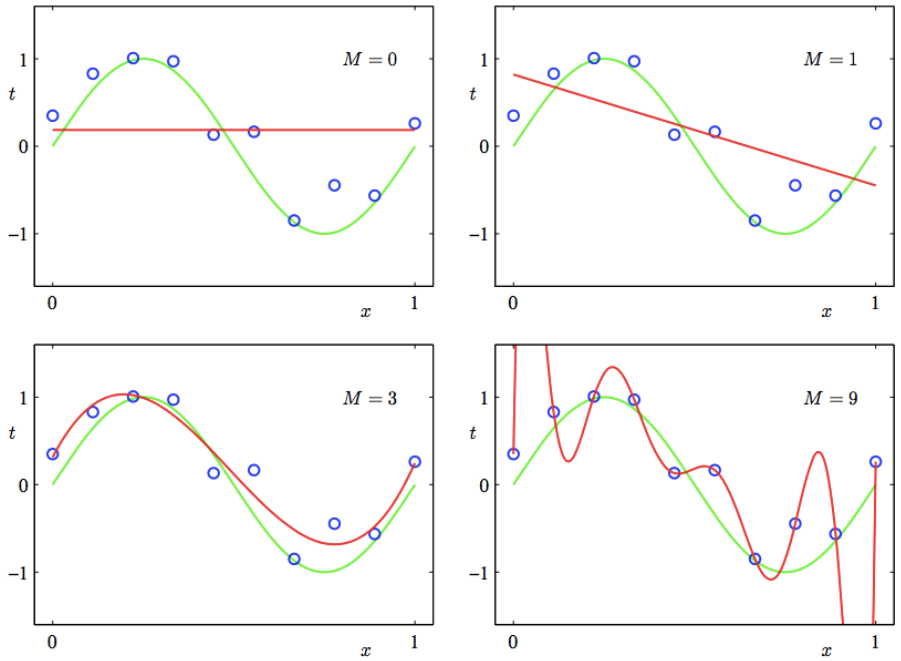
```
[0.17175444]
0.10182840236686364
0.9992863167421991
[0.27358284 0.44533728 0.61709172 0.78884615 0.96060059 1.13235503
 1.30410947 1.47586391 1.64761834 1.81937278 1.99112722 2.16288166
 2.33463609 2.50639053 2.67814497 2.84989941 3.02165385]
```



Under/Over-fitting

Underfitting: the model does not capture the important relationship(s)

Overfitting: the model describes noise instead of the underlying relationship



<u>Approaches</u>

- ***Regularization***
- Robust evaluation
  - Cross validation

Validation Set

- One approach in an ML-application pipeline is to use a ***validation*** dataset (could be a ***holdout*** from the training set)

- Each model is built using just training; the validation dataset is then used to compare performance and/or select model parameters

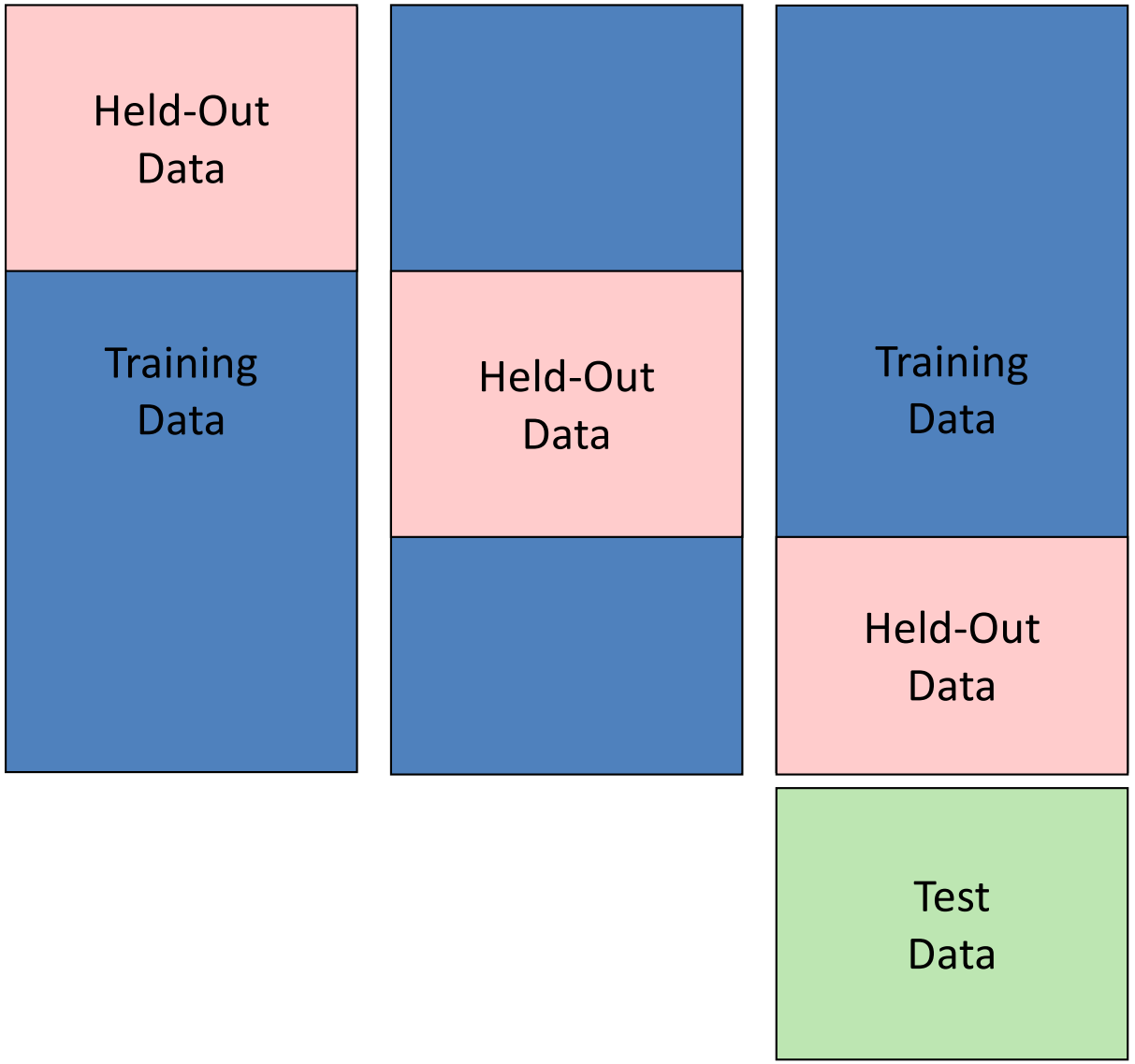
- But still, the final performance is only measured via an independent test set

More Training Data = Better

- In general, the greater the amount of training data, the better we expect the learning algorithm to perform
  - But we also want reasonable amounts of validation/testing data!

- So how do we not delude ourselves, achieve high performance, *and* a reasonable expectation of future performance?

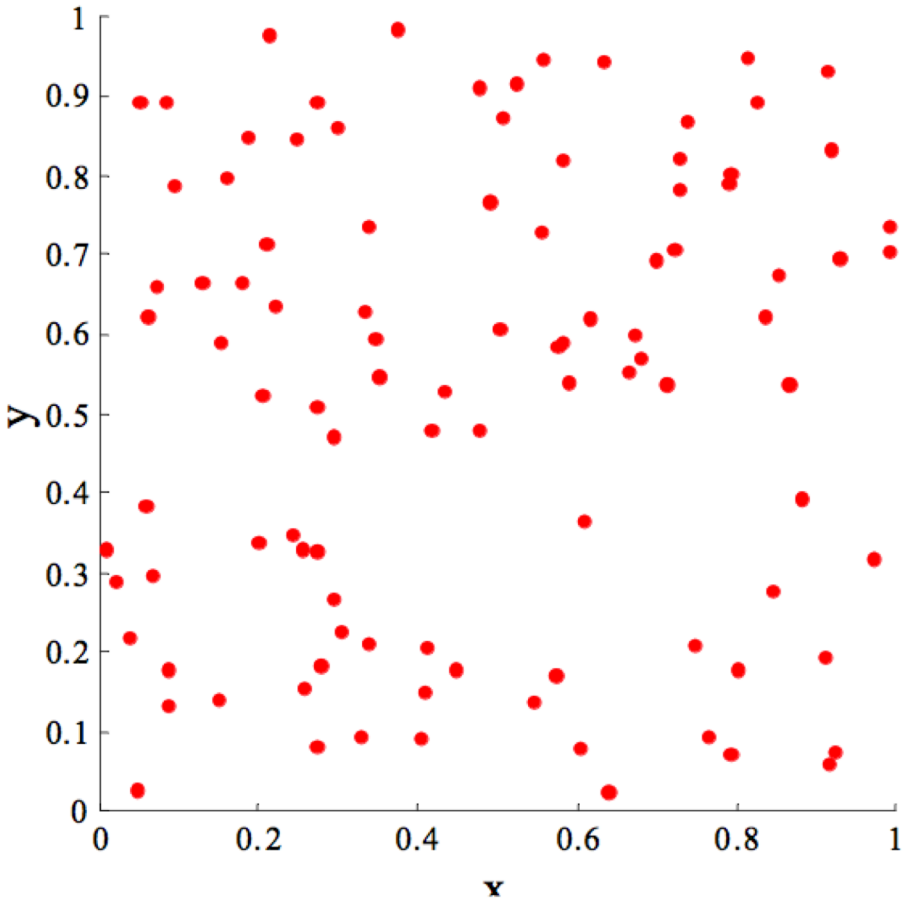
k-Fold Cross Validation



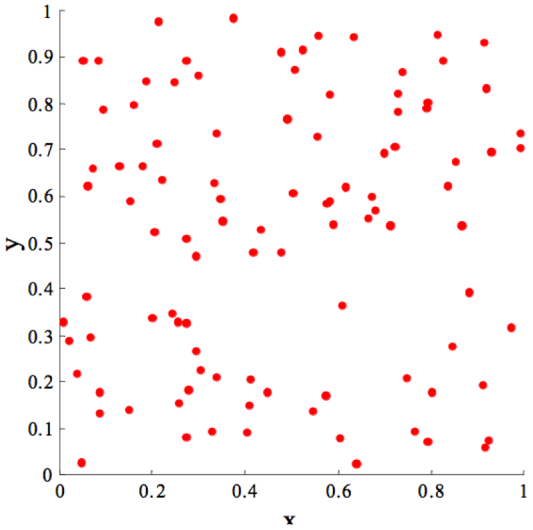
One (of many) Challenges

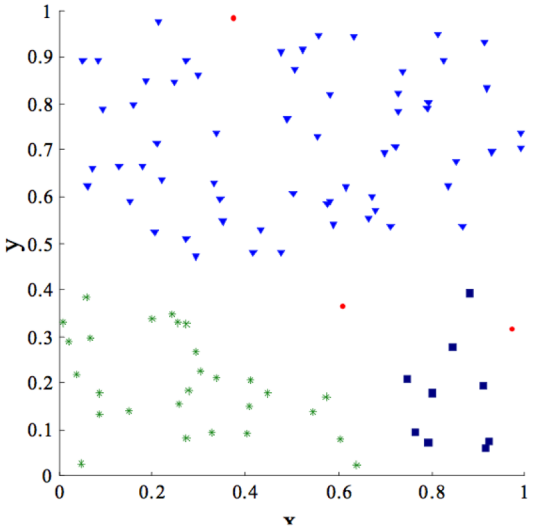
https://xkcd.com/1838/

What's a "Good" (Unsupervised) Answer?



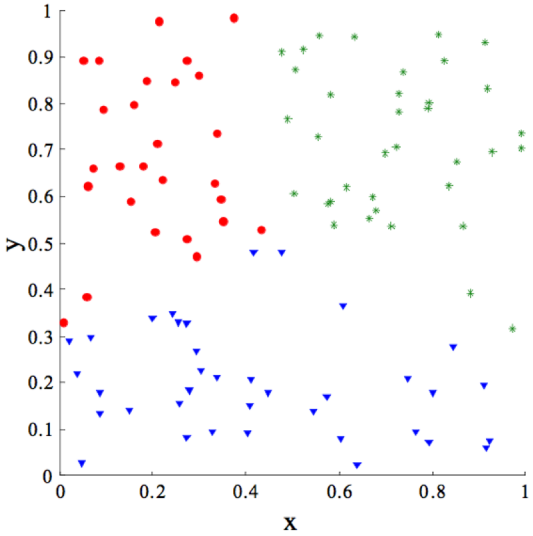
Did I Learn **Correctly**?

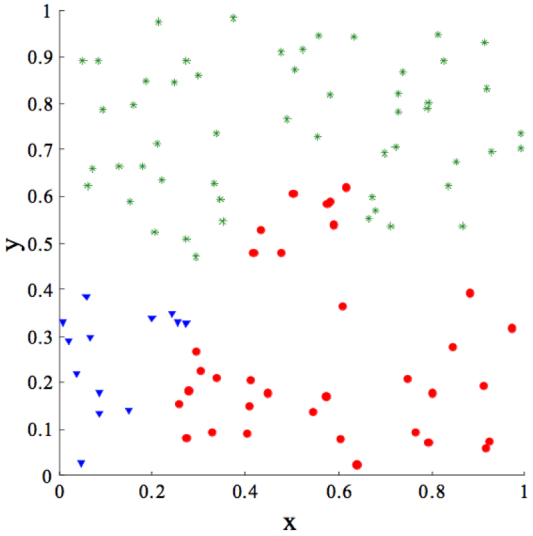




DBSCAN

K-means

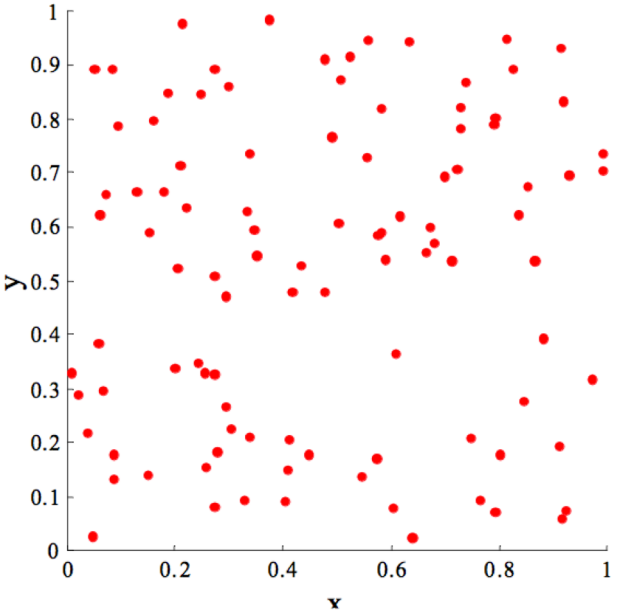




Complete Link

Did I Learn **Ethically** and **Safely**?

Random Points



- 👵🏽👴🏽 vs 👶🏿
- 🚌 vs 🚲 vs 🚶🏾‍♀️

Pop Quiz!

1. Given a dataset of past credit-card transactions (known to be fraudulent or not), build a system to identify future fraud

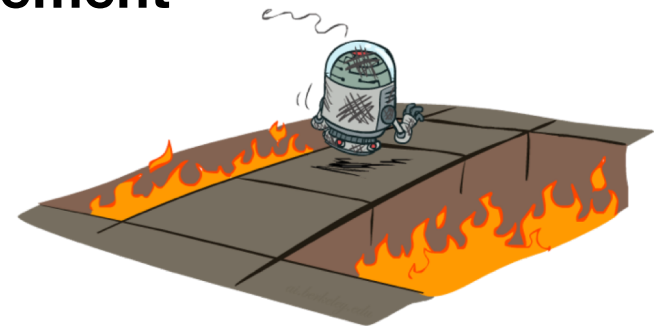
Supervised (classification)

2. If we assume incoming CS1 students are bi-modal, but normally distributed, find the average grades of the two groups

Unsupervised

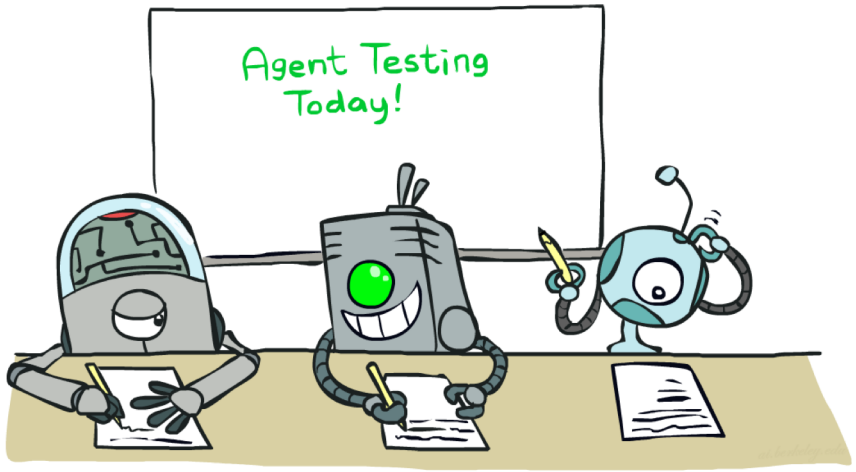
3. Build an Atari system that learns game-winning techniques via actually playing and adjusting actions based upon score changes

Reinforcement



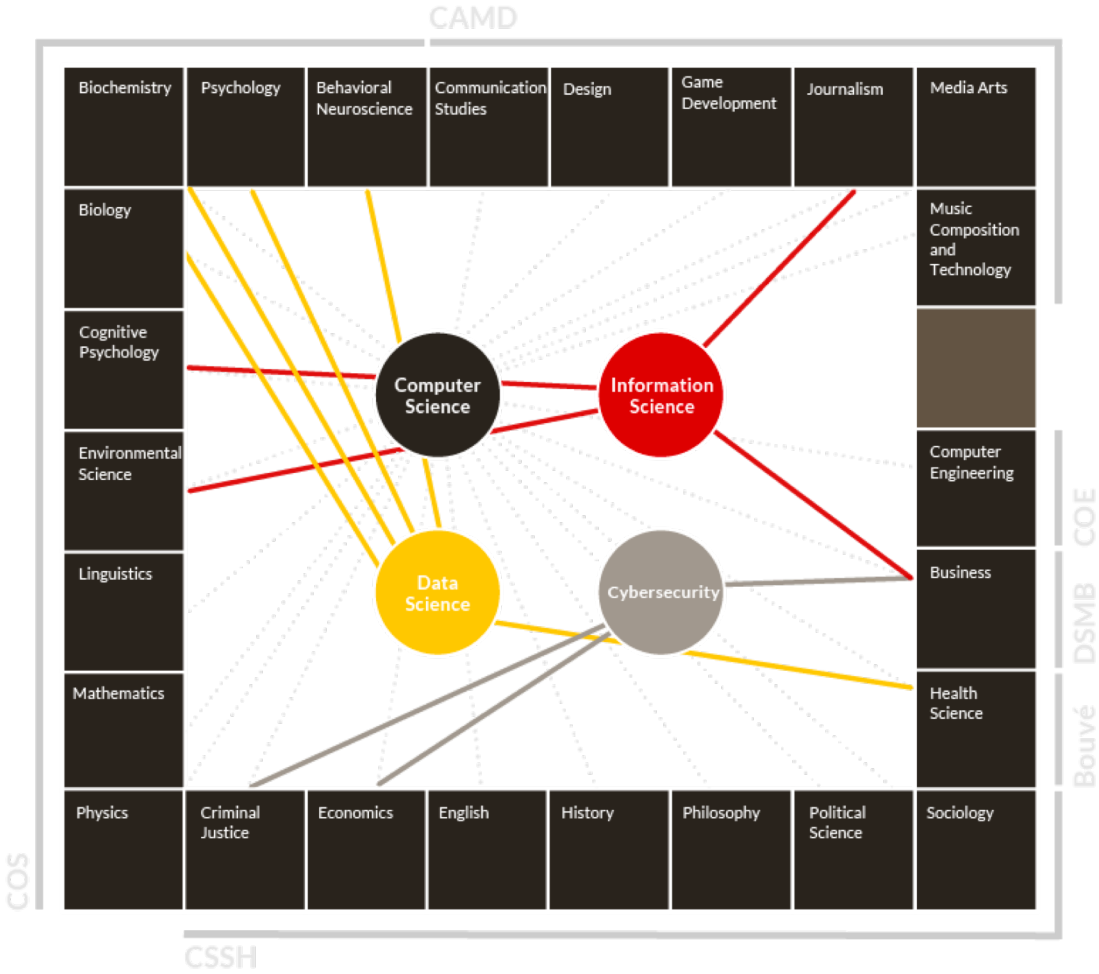
Opportunities for (Machine) Learning

1. Courses
2. Research



http://ai.berkeley.edu

Computing @ Northeastern

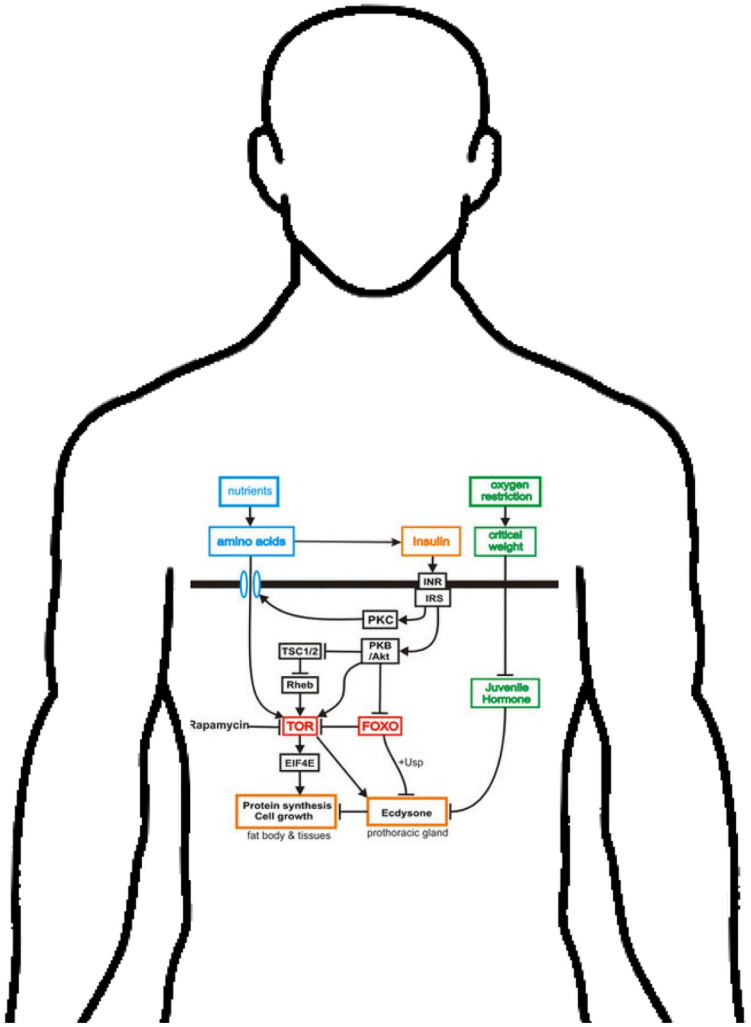


Some Offerings…

- Artificial Intelligence
- Supervised Machine Learning and Learning Theory
- Unsupervised Machine Learning and Data Mining
- Reinforcement Learning
- Natural Language Processing
- Advanced Machine Learning
- Information Presentation and Visualization
- Robotic Science and Systems
- Pattern Recognition and Computer Vision

…

Learning Signaling Networks



- Using NLP to aggregate and analyze scientific findings
- Given data, learn network structure

Learning Robotic Interaction

- Large-scale robotic grasp training
- Inter-planetary!

Learning to Win on the Merits

- Understanding content/style vs strength
- Making new arguments from past experience and data

Thank You :)

Questions?