

What is Machine Learning?

Nate Derbinsky

Associate Teaching Professor

Director of Teaching Faculty



My Path to CCIS @ Northeastern

bitX solutions 1998-2009 **BitX Solutions, Inc.** Founder & President
• {.gov .edu .org .com} x {desktop web mobile}

 2002-2006 **NC State.** BS Computer Science
• TA, DBMS



2006-2012 **U of Michigan.** MS/PhD Comp Sci and Eng
• TA, AI+DBMS



2012-2014 **Disney Research.** Postdoctoral Associate
• Machine Learning, Optimization, Robotics

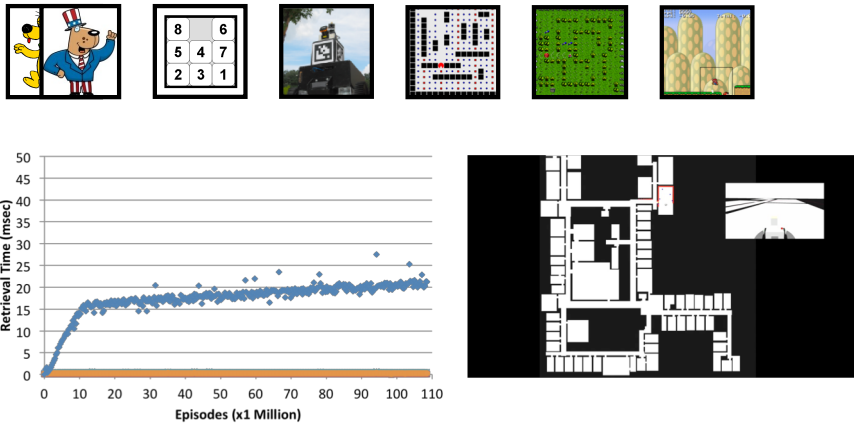


2014-2017 **Wentworth.** Assistant Professor
• 3-3, Research/Service Learning

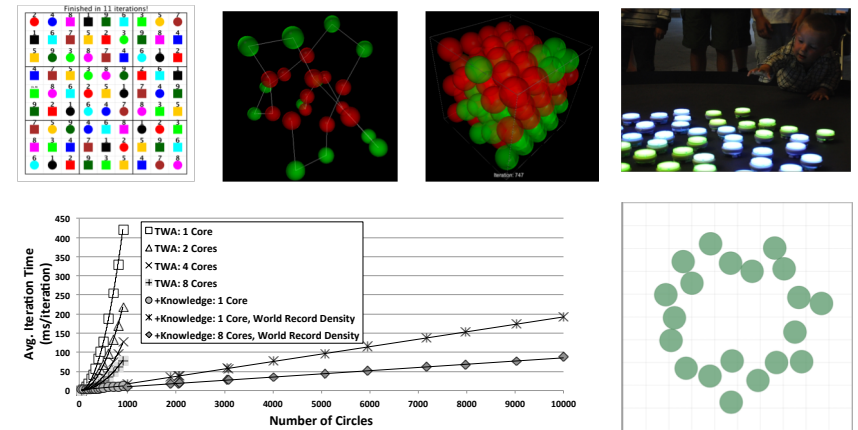


Research Interests

Cognitive Systems



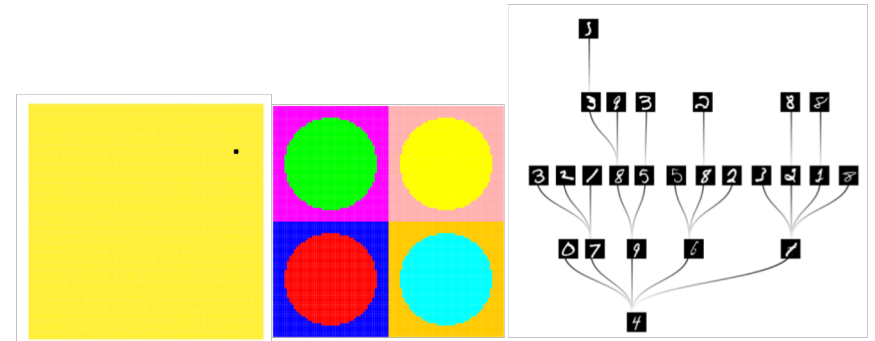
Scalable Optimization



AI Applications/Education



Online ML



What is Machine Learning?

October 26, 2018

Teaching

K-12/ICT-D



UG/Grad

- CS1/2
 - OOP, Foundations
- Databases, Web
- AI, Machine Learning
- HTMAA
 - RPi, Arduino



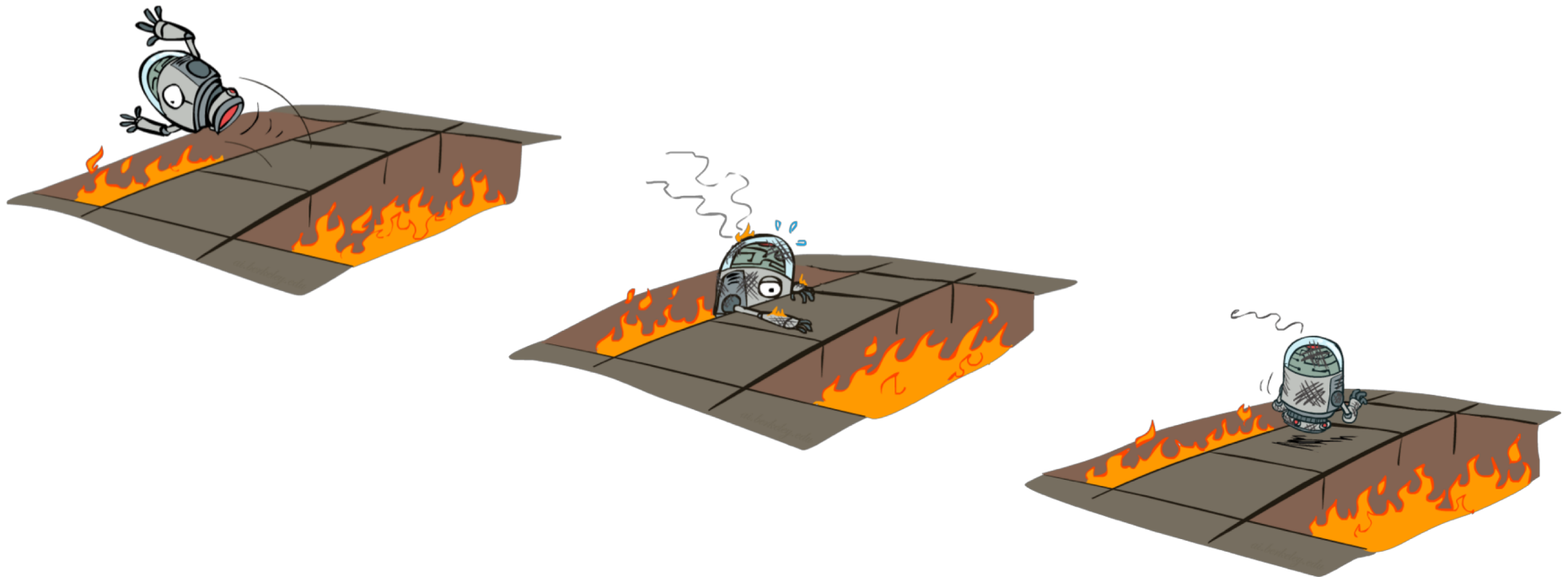
Agenda

- What is Machine Learning?
- Key Terminology/Tasks
- Challenges/Issues



What is Machine Learning?

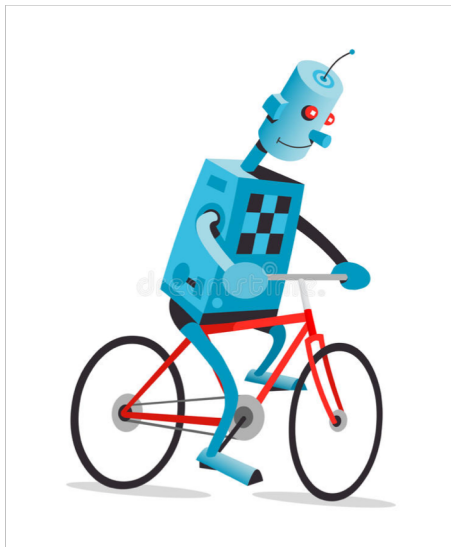
Computer programs that can improve performance with experience



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

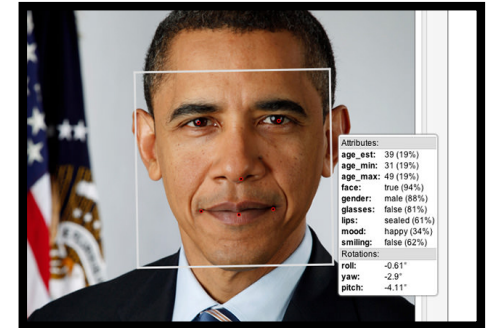
- Summarization
- Machine Translation
- Speech Processing
- Sentiment Analysis

...



Computer Vision

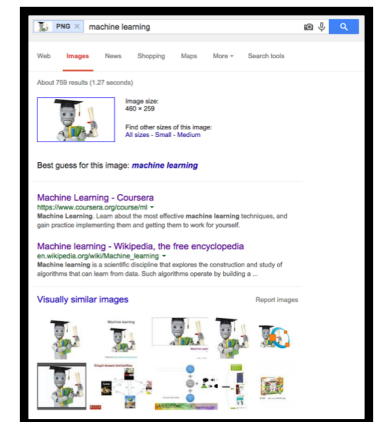
Methods for acquiring, processing, analyzing, and understanding images



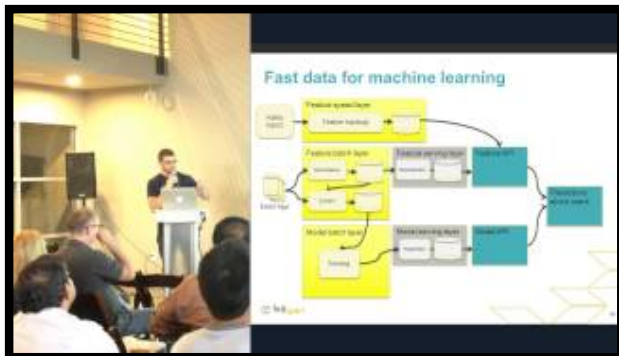
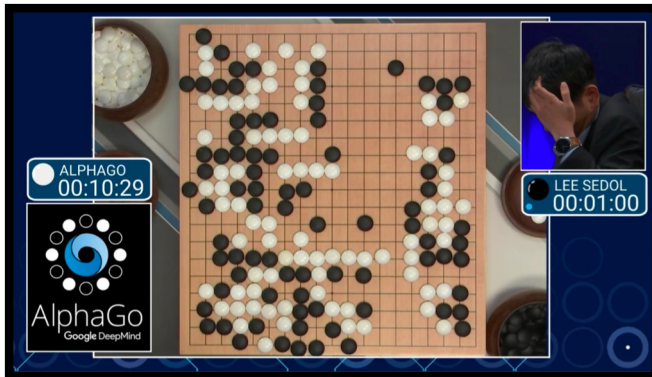
Applications

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

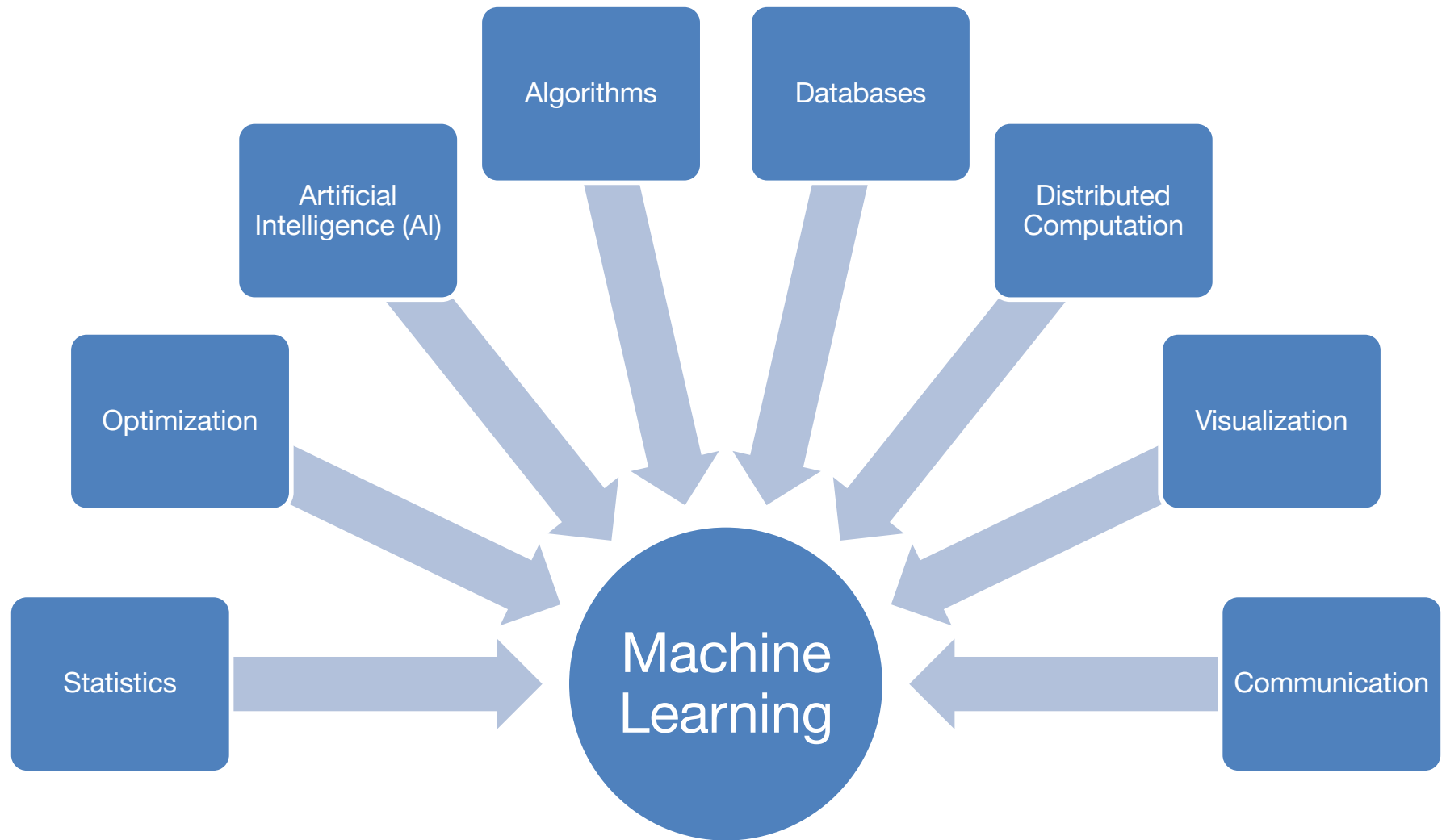
...



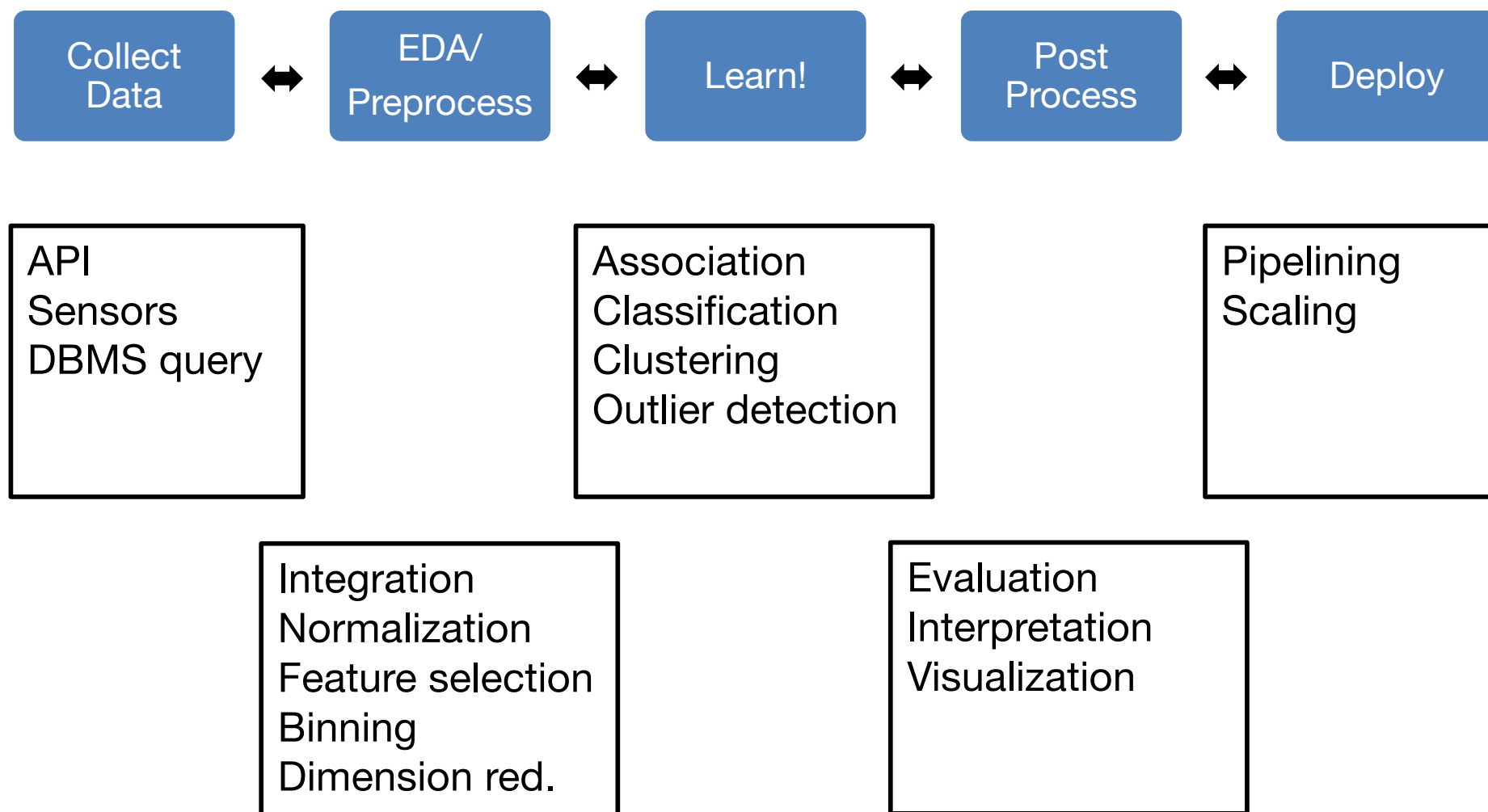
Games, Robotics, Medicine, Ads, ...



Fusing Disciplines



Machine Learning Pipeline



Jobs!

Position	Salary
Data Scientist	\$139,840
Software Engineer	\$115,462



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

Nearly all big tech companies are competing for artificial intelligence project, and the experts millions of dollars

[查看简体中文版](#) | [查看繁体中文版](#)

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.

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

– Josh Blumenstock (UW)

“Data Scientist = statistician + programmer + coach + storyteller + artist”

– Shlomo Aragmon (Ill. Inst. of Tech)

“Software Is Eating the World, but AI Is Going to Eat Software”

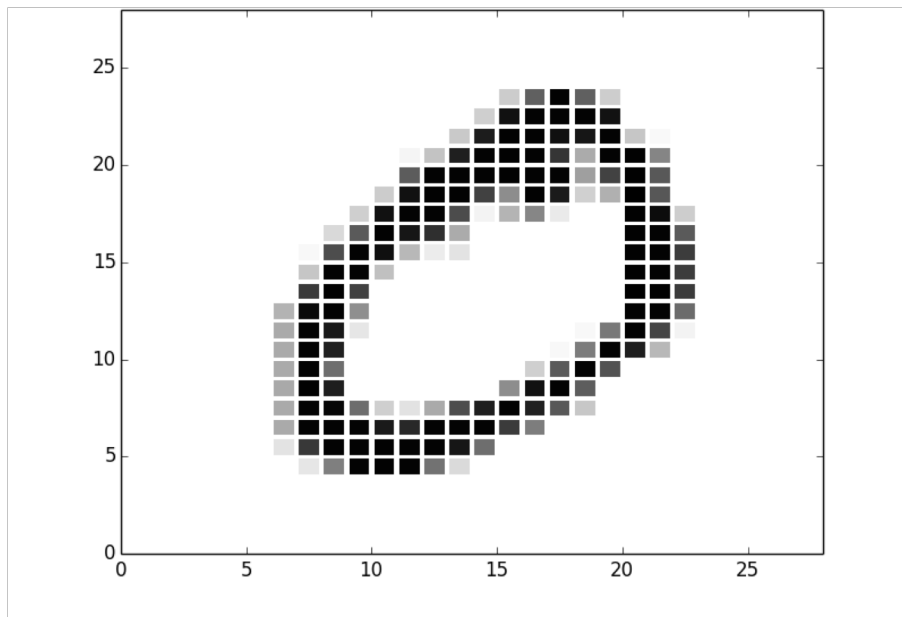
– Jensen Huang (CEO, NVIDIA)

*glassdoor.com, National Avg as of October 20, 2018



What is Machine Learning?

Common Terminology

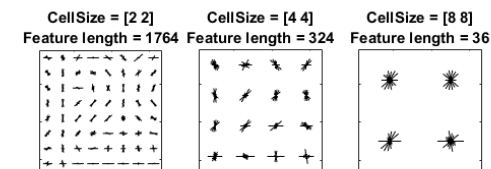


example, instance

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

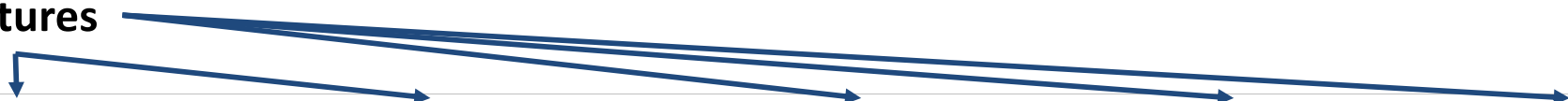


- In general, the problem of ***feature selection*** is challenging



Instances/Features = Table

Features



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

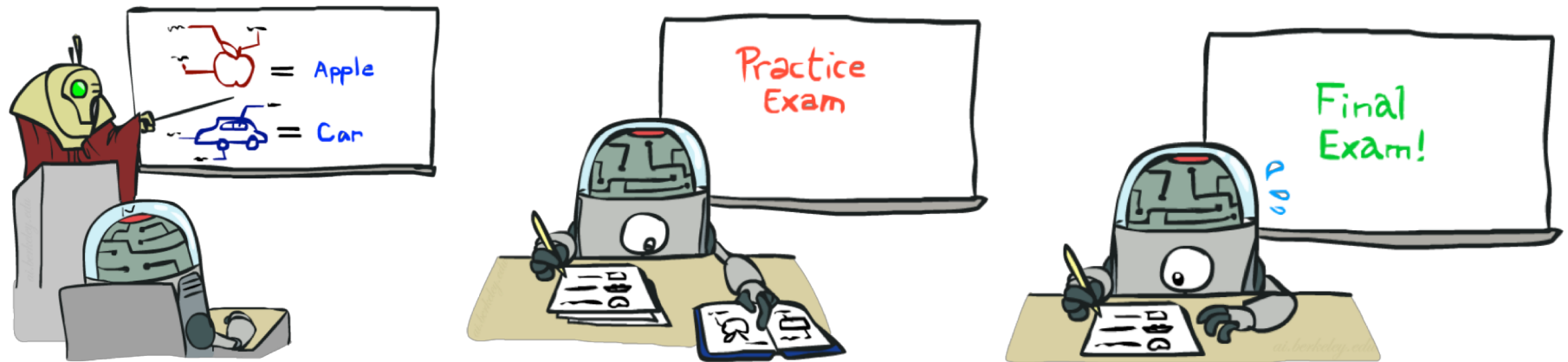


Machine Learning Tasks

- ***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



Supervised Learning (1)



Supervised Learning (2)

Training Set



α



β



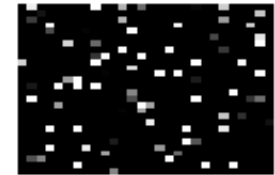
β



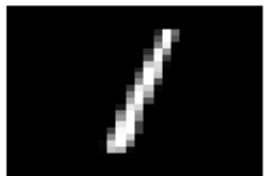
γ

...

Testing Set



?



...



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	datsum 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



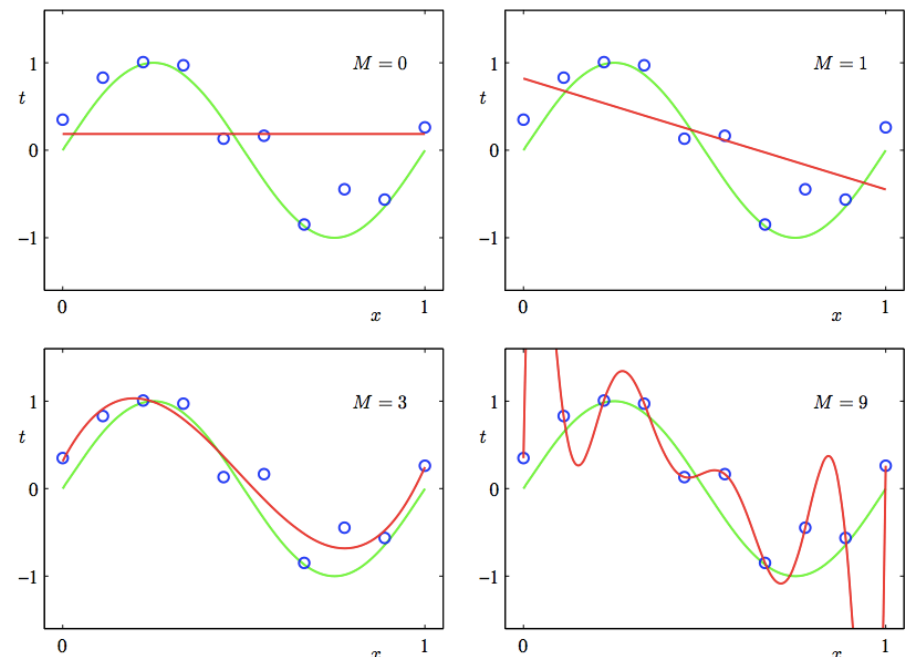
Under/Over-fitting

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

Overfitting: the model describes noise instead of the underlying relationship

Approaches

- **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

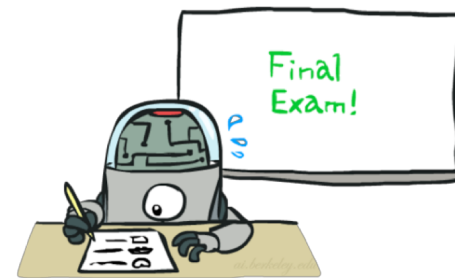
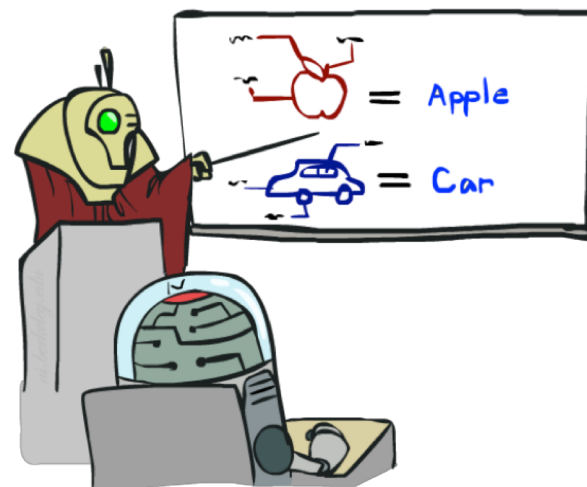
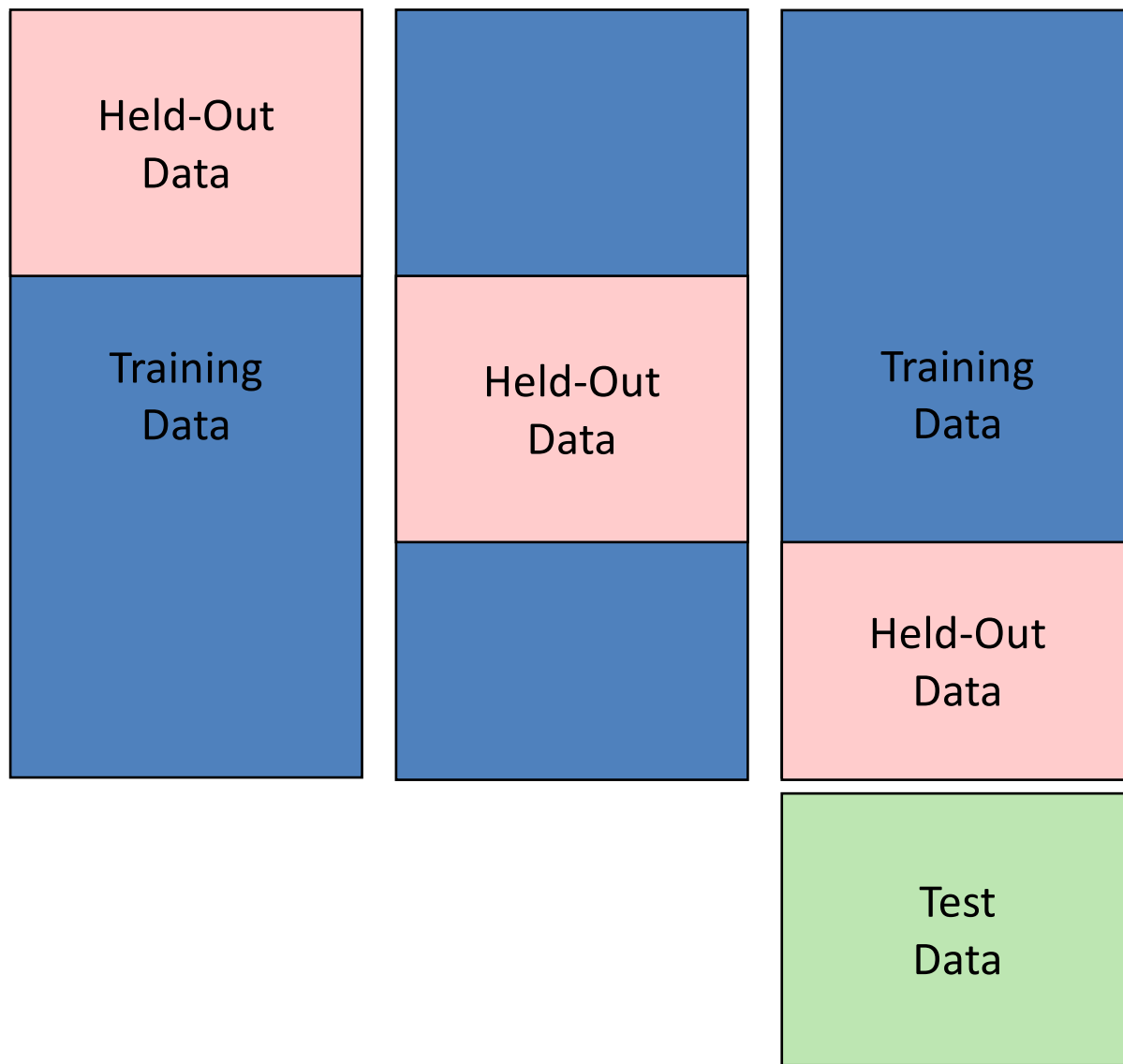


More Training Data = Better

- 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



Common Algorithms

- Instance-based
 - **Nearest Neighbor (kNN)**
- Tree-based
 - ID3, C4.5, Random Forests
- Optimization-based
 - **Linear regression**, logistic regression, support vector machines (SVM)
- Probabilistic
 - Naïve Bayes, HMM
- Artificial Neural Networks
 - Backpropagation
 - Deep learning



kNN

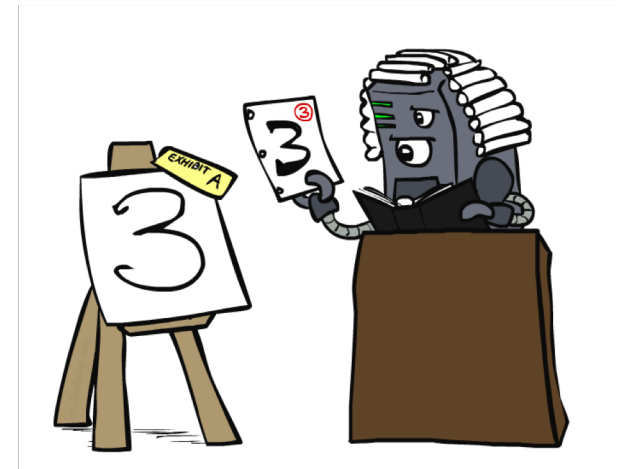
Training

- Store all examples

Testing

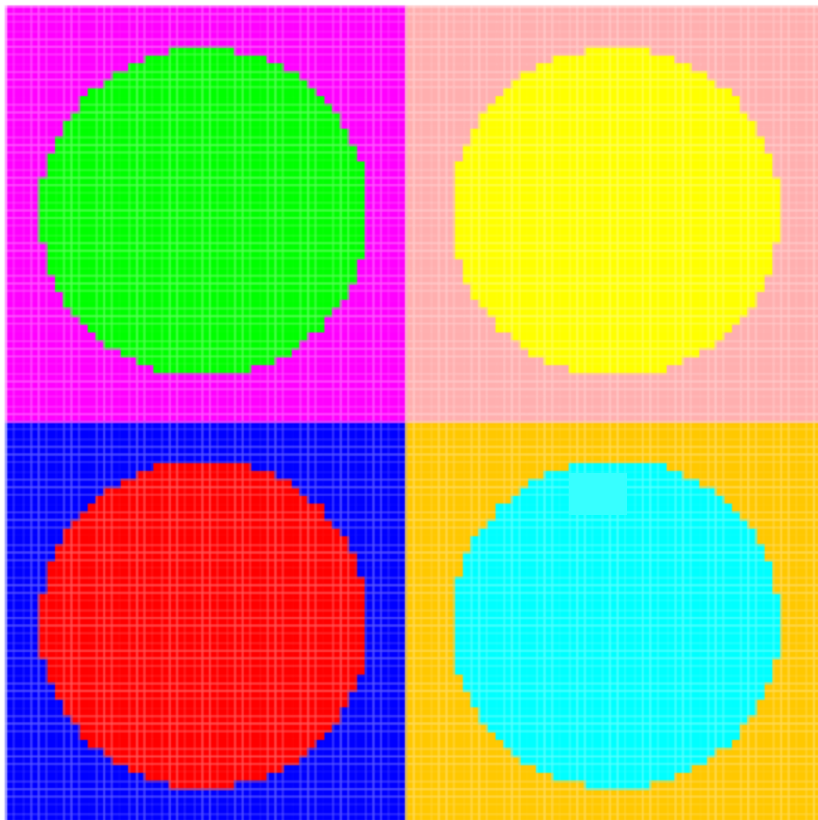
- Find the nearest k neighbors to target
 - Via distance function
- Vote on class

Non-parametric algorithm
(i.e. grows with |examples|!)

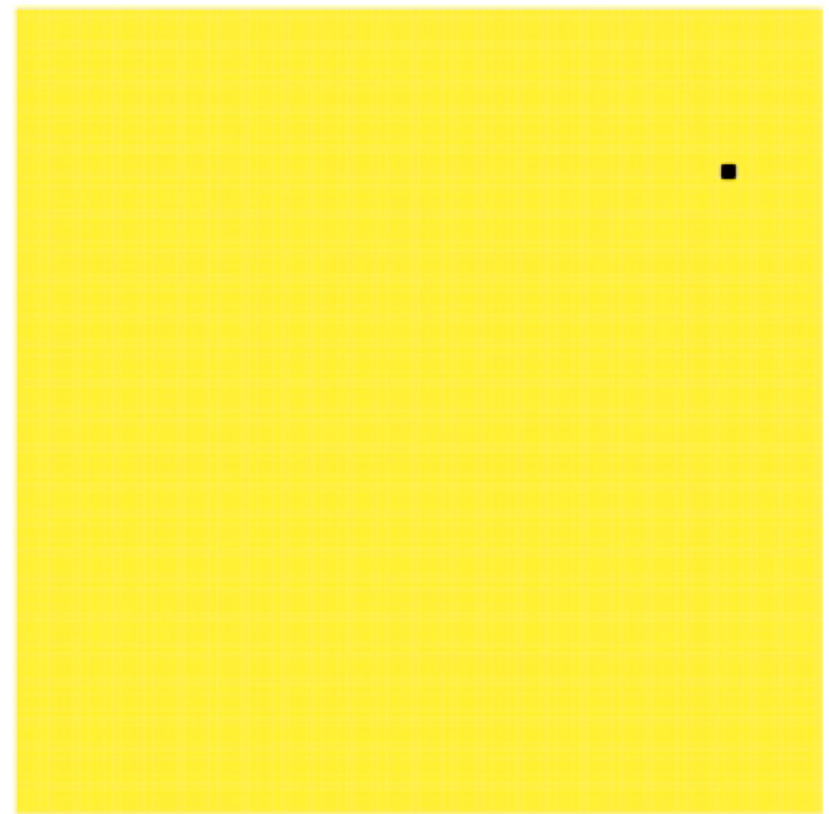


2D Multiclass Classification

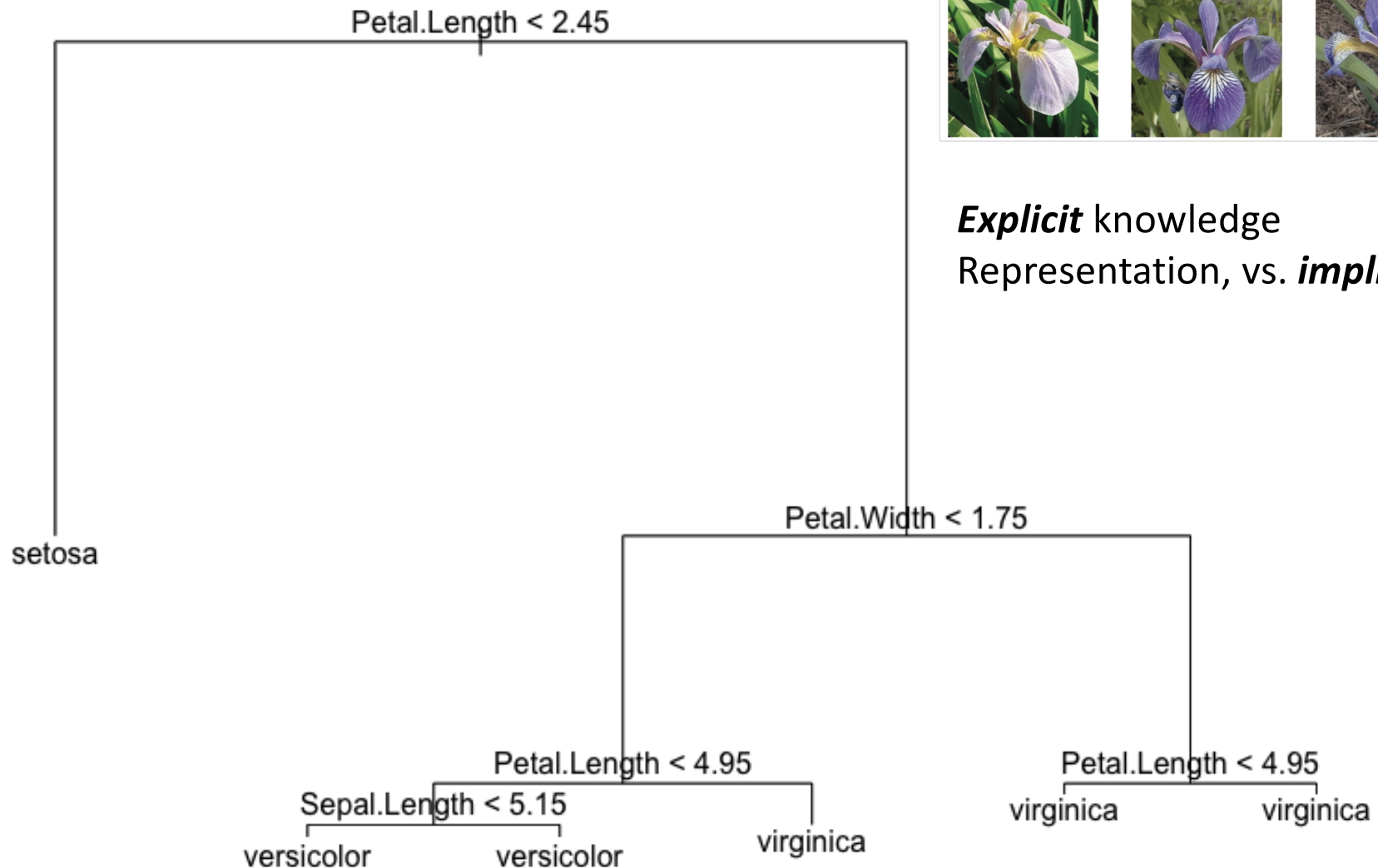
Boundary Tree



1-NN via Linear Scan



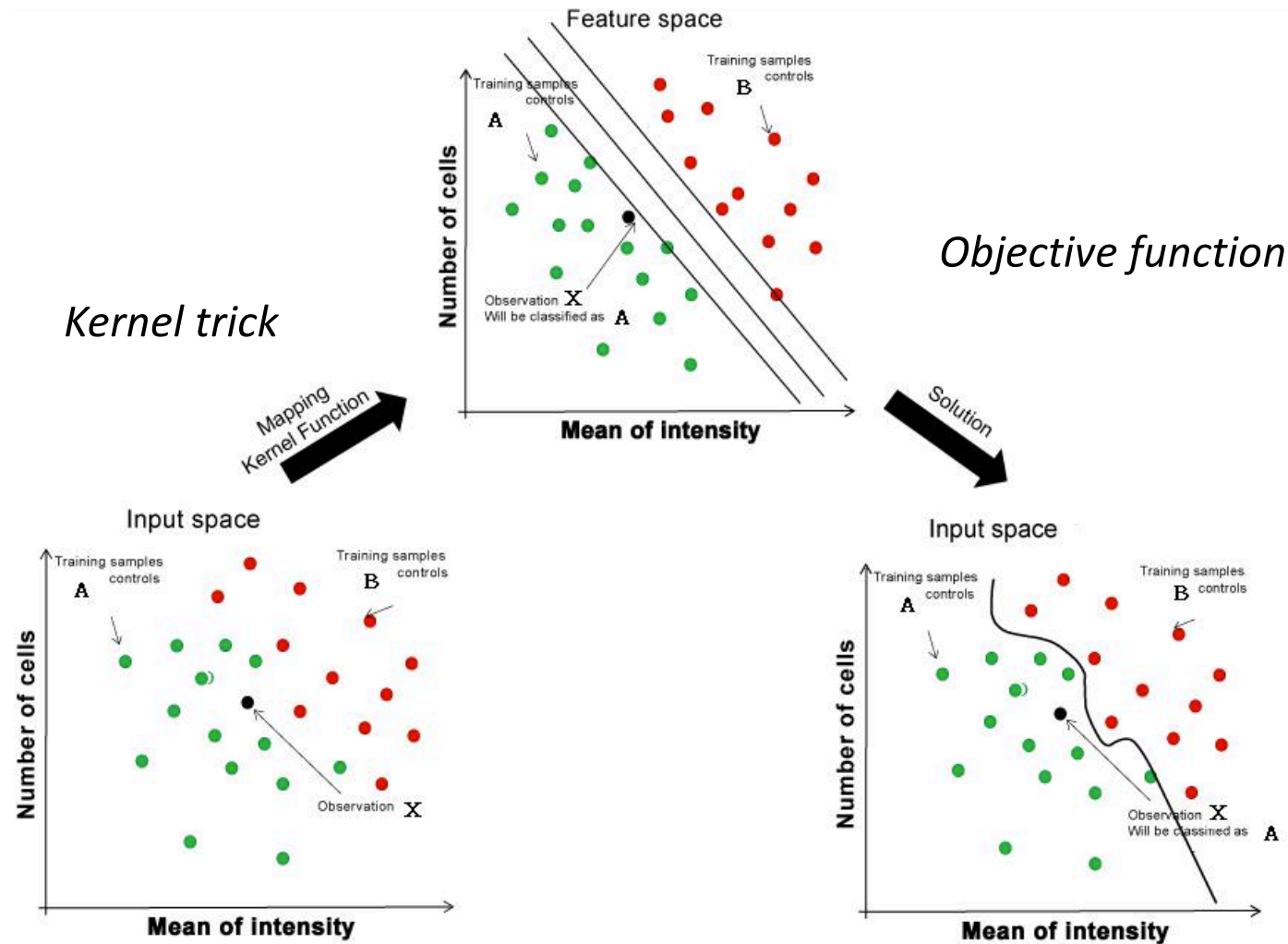
Decision Trees/Forests



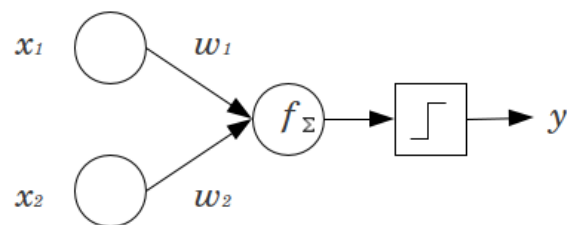
Explicit knowledge
Representation, vs. **implicit**



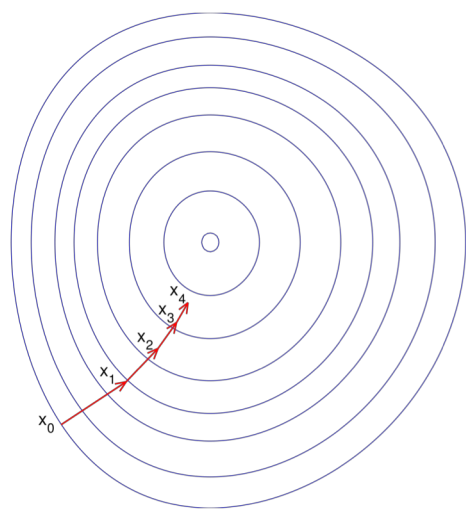
Support Vector Machine (SVM)



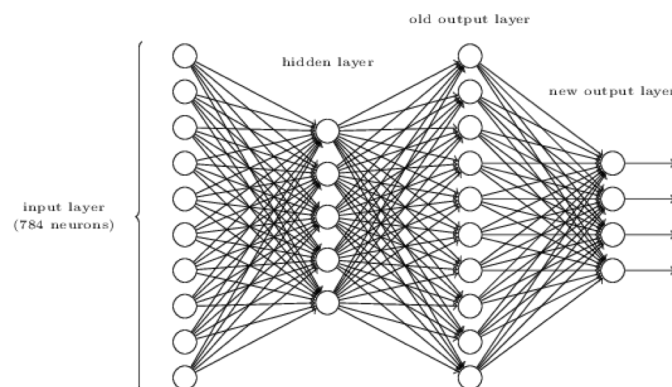
Artificial Neural Networks (ANN)



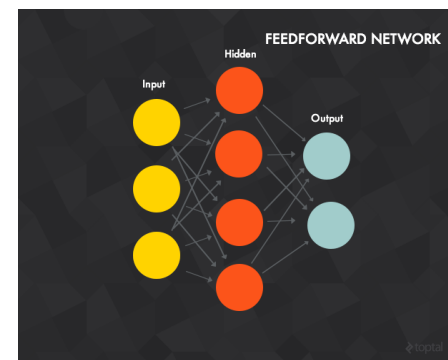
Perceptron
Linear classifier



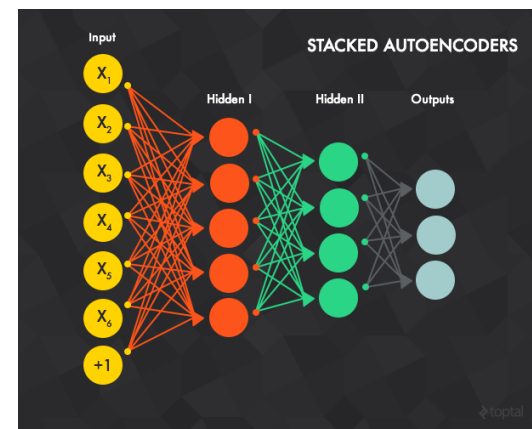
Gradient descent



Backpropagation



Feedforward vs.
Recurrent



Deep Architectures
Vanishing Gradient

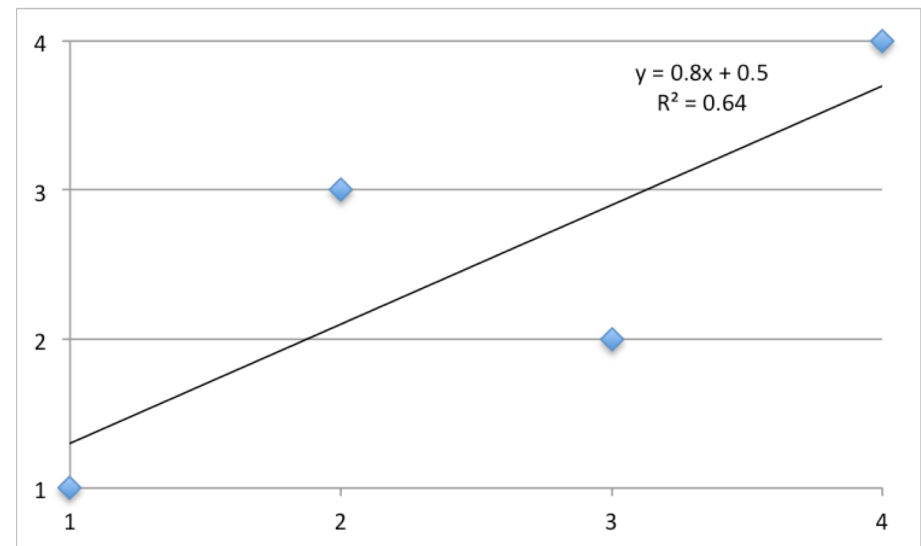


Example: Linear Regression

Input

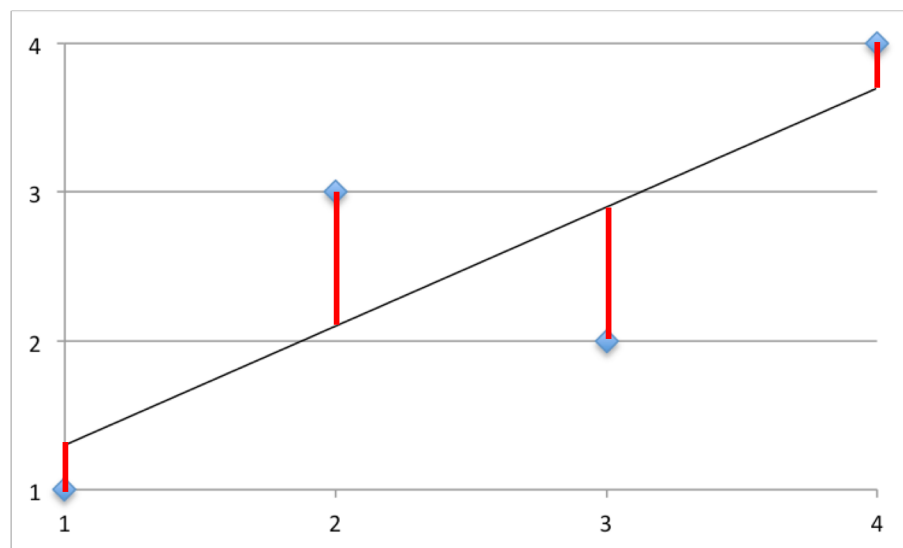
x	y
1	1
2	3
3	2
4	4

Output



Linear Regression as Optimization

- Why this line?
 - Minimizer **error**
- In 2D, the algorithm tries to find a slope and intercept that yields the smallest sum of the square of the error (SSE)



$$\arg \min_{m,b} \sum_{i=1}^N e_i^2 = (y_i - (mx_i + b))^2$$

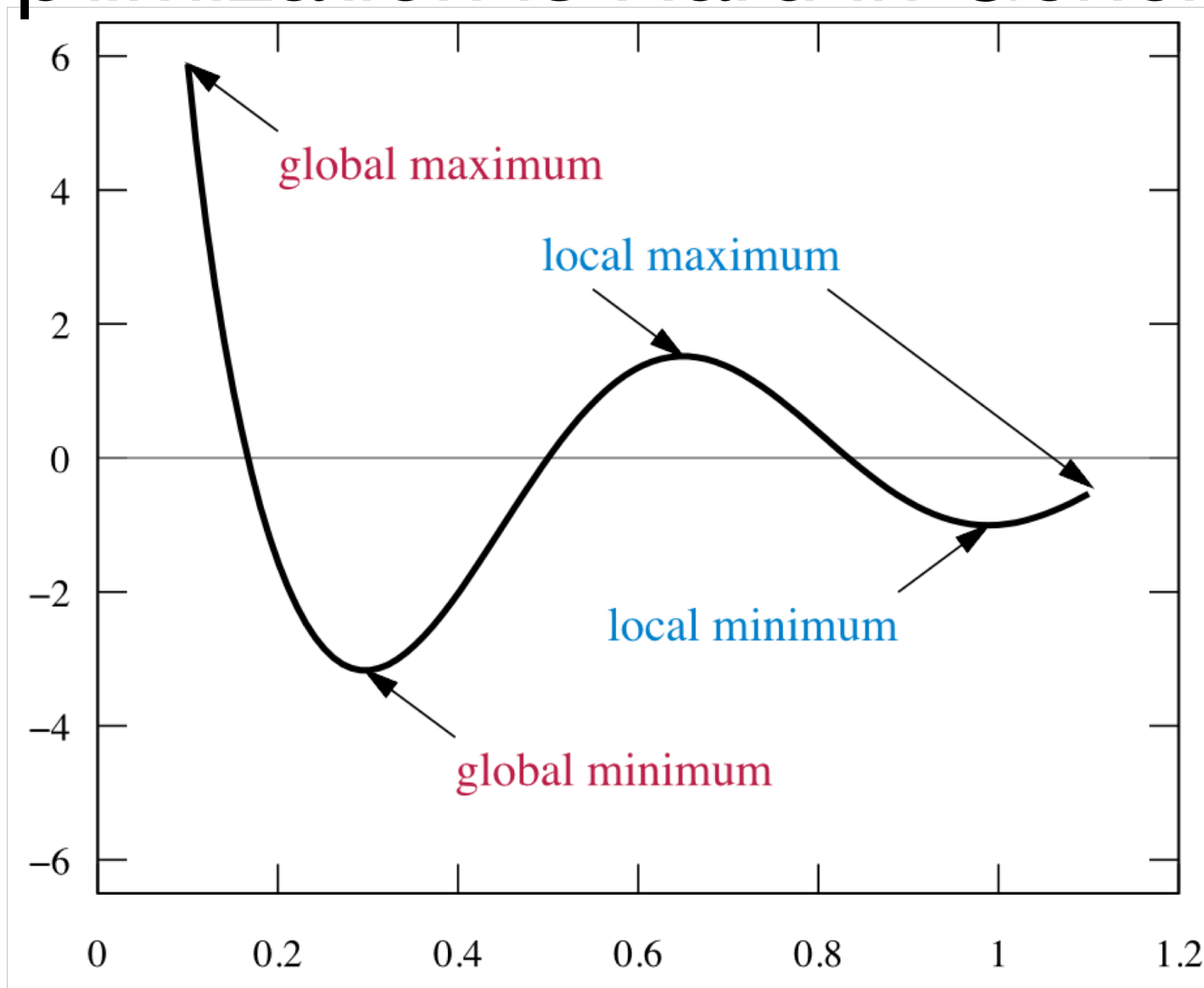


Machine Learning via Optimization

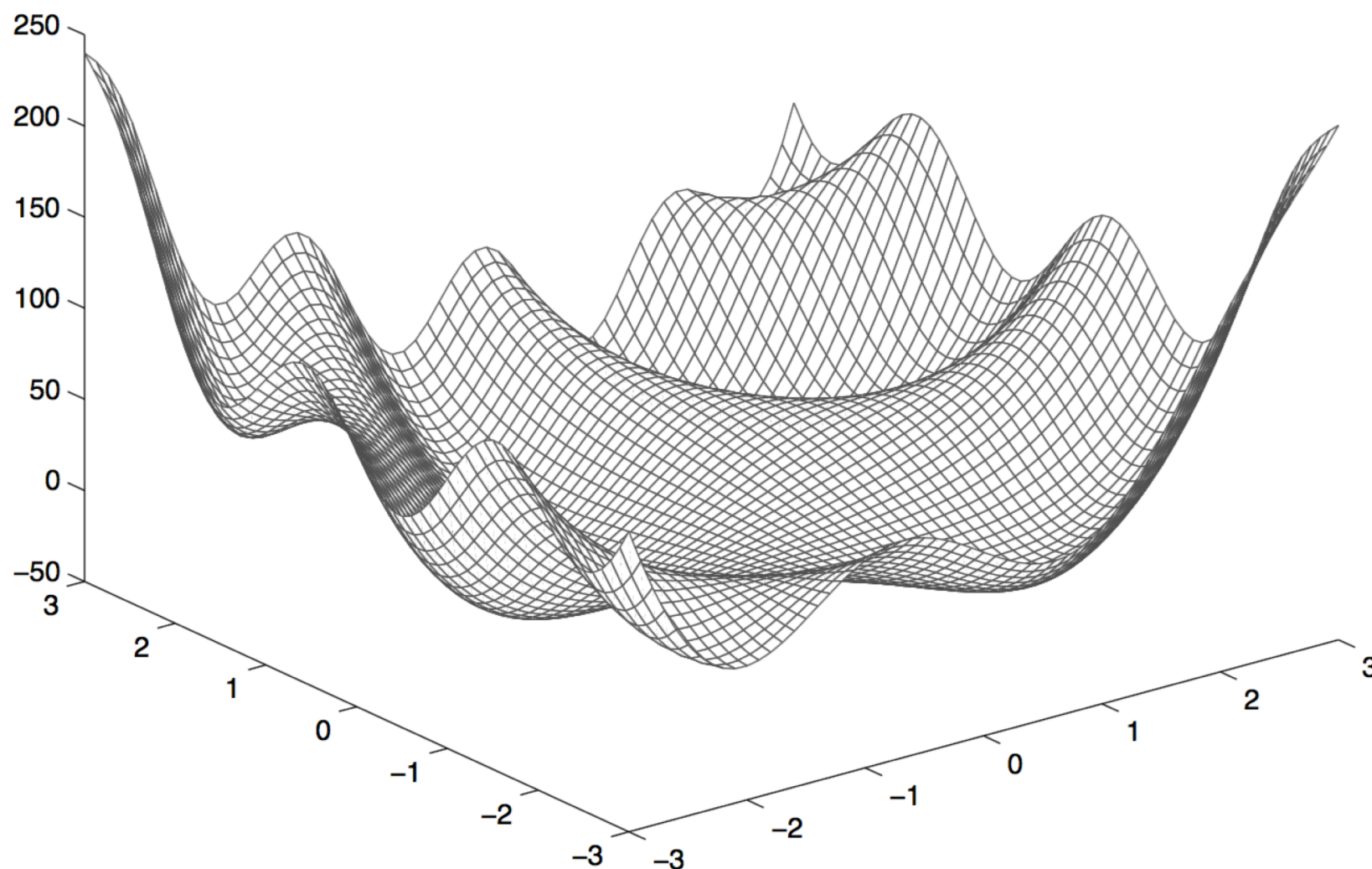
1. Define an error function
2. Find model parameters that minimize the error function given the data
 - Sometimes closed-form solution (e.g. linear)
 - Sometimes [iterative] solution [with guarantees] (e.g. convex)
 - Most of the time will require approximation
 - Iteration (limited by number, delta)
 - Softening constraints
 - Post-processing
 - ...



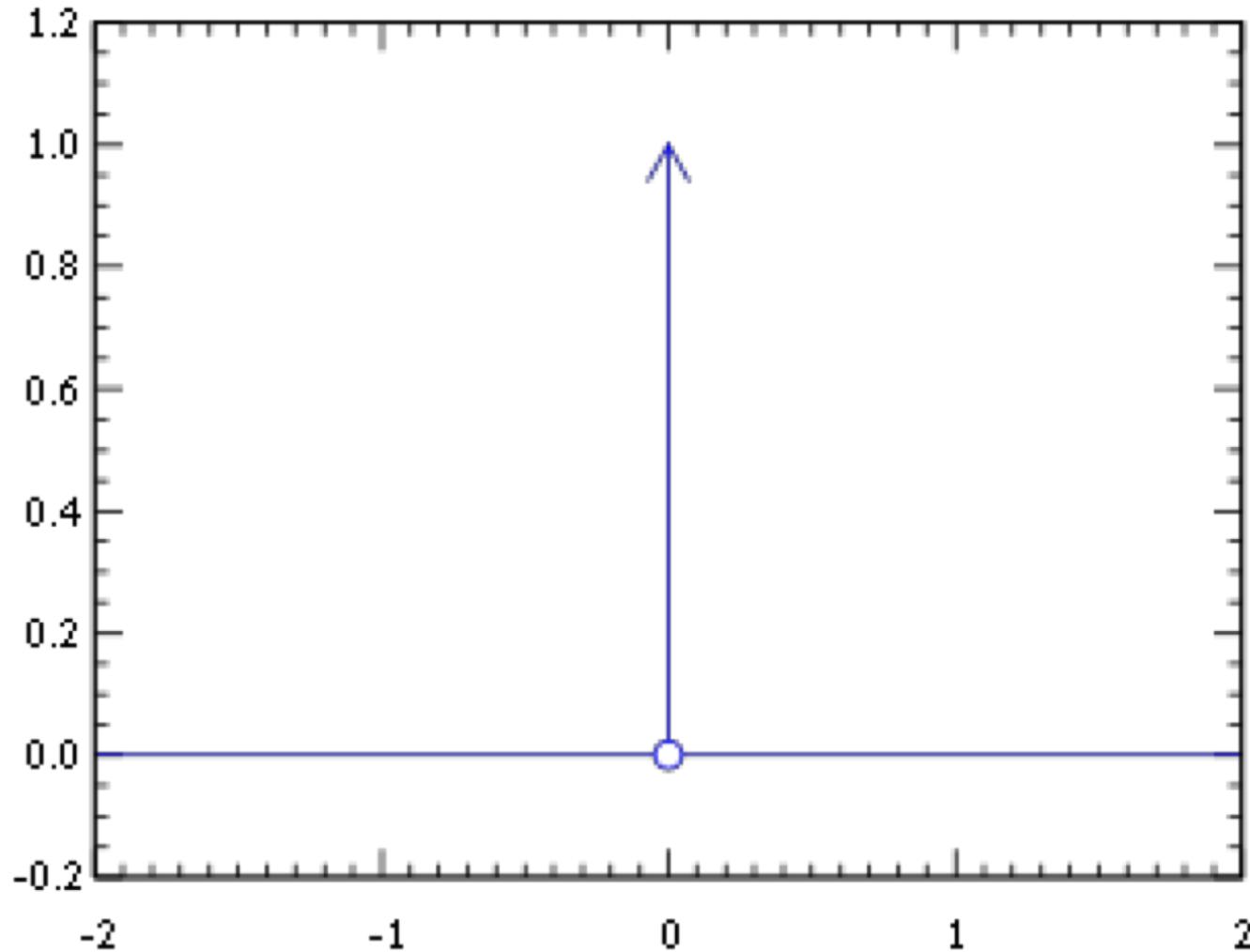
Optimization is Hard in General



Consider Many [Cursed] Dimensions



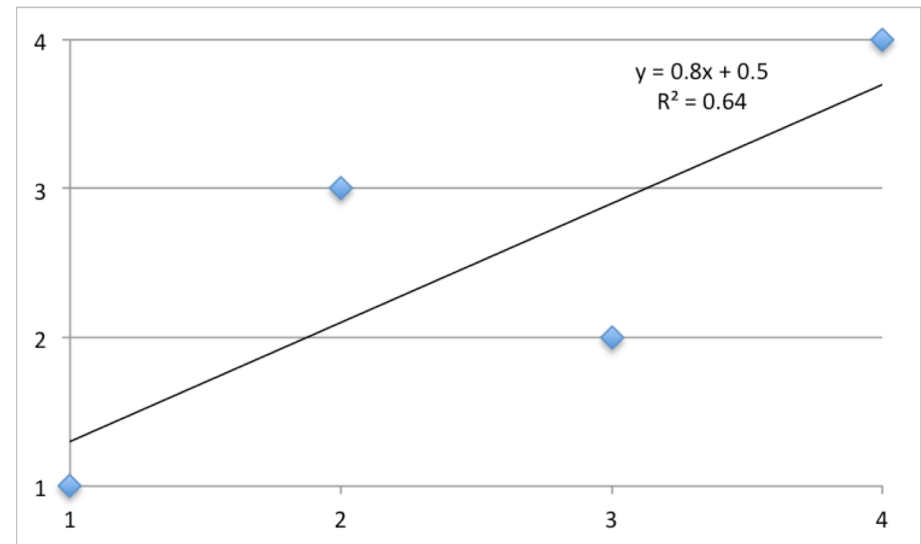
Consider Discontinuities



Linear Regression

Recipe

1. Define error function
2. Find parameter values that minimize error given the data

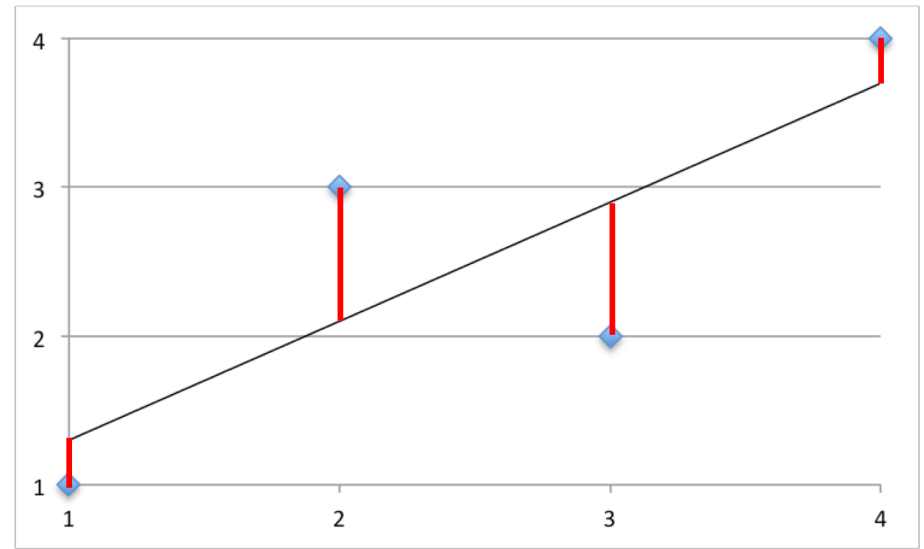


Error Function

Sum of Squared Error (SSE)

aka Residual Sum of Squares (RSS)

$$\text{SSE}_{\text{line}} = \sum_{i=1}^N (y_i - f(x_i))^2 = (y_i - (mx_i + b))^2$$



Algebra (1)

$$\begin{aligned}\text{SSE}_{\text{line}} &= \sum_{i=1}^N (y_i - (mx_i + b))^2 \\ &= \sum_{i=1}^N y_i^2 - 2y_i(mx_i + b) + (mx_i + b)^2 \\ &= \sum_{i=1}^N y_i^2 - 2mx_iy_i - 2by_i + m^2x_i^2 + 2mbx_i + b^2\end{aligned}$$



Algebra (2)

$$\sum_{i=1}^N a_i = N\bar{a}$$

so...

$$\begin{aligned}\text{SSE}_{\text{line}} &= \sum_{i=1}^N y_i^2 - 2mx_iy_i - 2by_i + m^2x_i^2 + 2mbx_i + b^2 \\ &= N\bar{y}^2 - 2Nm\bar{x}\bar{y} - 2Nb\bar{y} + Nm^2\bar{x}^2 + 2Nmb\bar{x} + Nb^2\end{aligned}$$



Recall: Critical Points

- For a differentiable function of several variables, a critical point is a value in its domain where all partial derivatives are zero
- So to find the point at which error is minimized, we take partial derivatives of the error function w.r.t. the parameters, set these equal to 0, solve



Calculus (1)

$$\text{SSE}_{\text{line}} = N\overline{y^2} - 2Nm\overline{xy} - 2Nb\overline{y} + Nm^2\overline{x^2} + 2Nmb\overline{x} + Nb^2$$

$$\frac{\partial \text{SSE}_{\text{line}}}{\partial m} = -2N\overline{xy} + 2Nm\overline{x^2} + 2Nb\overline{x} = 0$$

$$\frac{\partial \text{SSE}_{\text{line}}}{\partial b} = -2N\overline{y} + 2Nm\overline{x} + 2Nb = 0$$



Algebra (3)

$$\frac{\partial \text{SSE}_{\text{line}}}{\partial b} = -2N\bar{y} + 2Nm\bar{x} + 2Nb = 0$$

$$0 = -2N\bar{y} + 2Nm\bar{x} + 2Nb$$

$$0 = -\bar{y} + m\bar{x} + b$$

$$\bar{y} = m\bar{x} + b \quad (\bar{x}, \bar{y})$$



Algebra (4)

$$\frac{\partial \text{SSE}_{\text{line}}}{\partial m} = -2N\overline{xy} + 2Nm\overline{x^2} + 2Nb\overline{x} = 0$$

$$0 = -2N\overline{xy} + 2Nm\overline{x^2} + 2Nb\overline{x}$$

$$0 = -\overline{xy} + m\overline{x^2} + b\overline{x}$$

$$\overline{xy} = m\overline{x^2} + b\overline{x}$$

$$\frac{\overline{xy}}{\overline{x}} = m\frac{\overline{x^2}}{\overline{x}} + b$$

$$\left(\frac{\overline{x^2}}{\overline{x}}, \frac{\overline{xy}}{\overline{x}}\right)$$



And Finally...

$$(\bar{x}, \bar{y})$$

$$\left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$$

$$m = \frac{\frac{\overline{xy}}{\bar{x}} - \bar{y}}{\frac{\overline{x^2}}{\bar{x}} - \bar{x}}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$\bar{y} = m\bar{x} + b$$

$$b = \bar{y} - m\bar{x}$$



The Multi-Dimensional Case

We begin with an analogous representation

$$Y = XB + e$$

where...

- **Y** is $N \times 1$
- **X** is $N \times (k+1)$; extra 1 to multiply intercept
- **B** is $(k+1) \times 1$; first intercept, then coefficients
- **e** is $N \times 1$



k-Dimensional Linear Regression

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1k} \\ 1 & x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & & & & & \\ \cdot & \cdot & & & & & \\ \cdot & \cdot & & & & & \\ 1 & x_{N1} & x_{N2} & \cdot & \cdot & \cdot & x_{Nk} \end{bmatrix} \begin{bmatrix} b \\ m_1 \\ m_2 \\ \cdot \\ \cdot \\ \cdot \\ m_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{bmatrix}$$



Step 1: Error Function

- We will use the same error method as last time, which is SSE (i.e. square the difference between Y and XB)

$$\begin{aligned} \text{SSE} &= e^T e \\ &= (Y - XB)^T (Y - XB) \end{aligned}$$



Matrix Algebra (1)

$$\begin{aligned}\text{SSE} &= (Y - XB)^\top (Y - XB) \\ &= (Y^\top - B^\top X^\top)(Y - XB) \\ &= Y^\top Y - Y^\top XB - B^\top X^\top Y + B^\top X^\top XB \\ &= Y^\top Y - 2Y^\top XB + B^\top X^\top XB\end{aligned}$$



Matrix Calculus (1)

$$\text{SSE} = Y^{\top}Y - 2Y^{\top}XB + B^{\top}X^{\top}XB$$

$$\frac{\partial \text{SSE}}{\partial B} = -2X^{\top}Y + 2X^{\top}XB$$



Matrix Algebra (2)

$$0 = -2X^T Y + 2X^T X B$$

$$-2X^T X B = -2X^T Y$$

$$X^T X B = X^T Y$$

$$B = (X^T X)^{-1} X^T Y$$



Unsupervised Learning

Find structure or patterns in data

Tasks

- Clustering
- Dimensionality reduction
- Density estimation
- Discovering graph structure
- Matrix completion
- ...



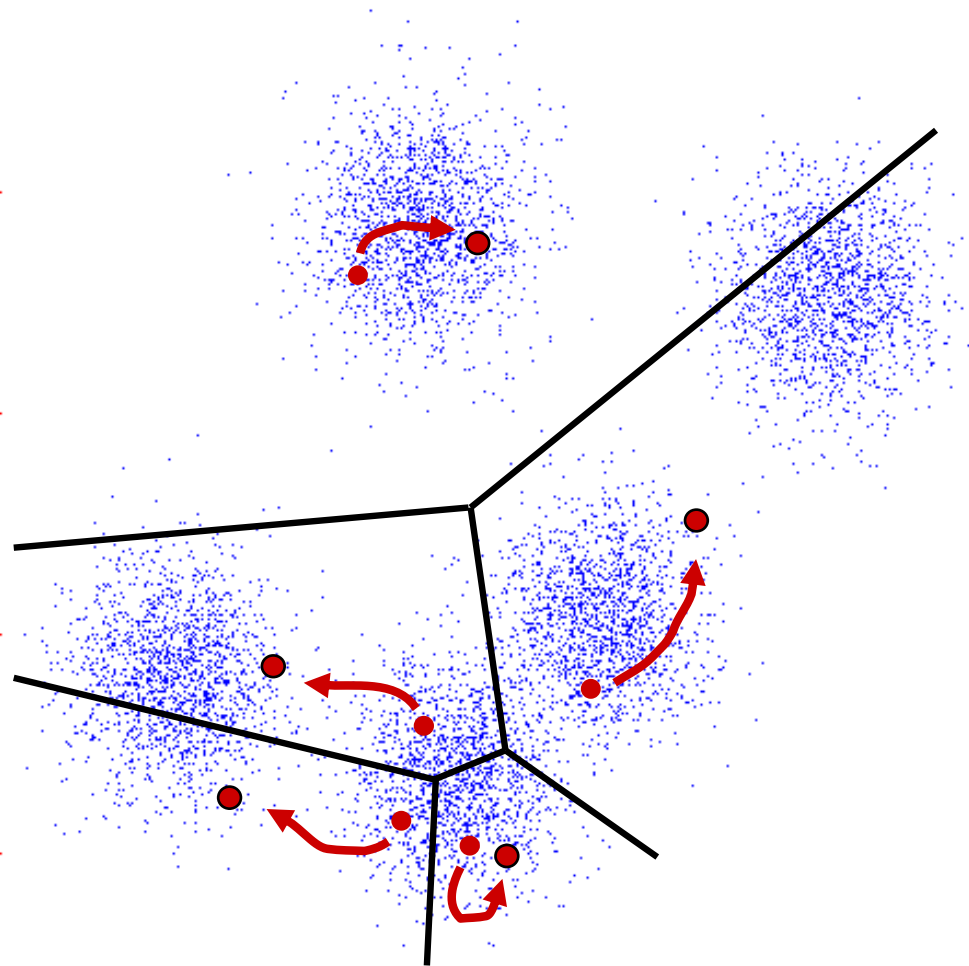
Common Algorithms

- k-Means Clustering
- Collaborative Filtering
- Principle Component Analysis (PCA)

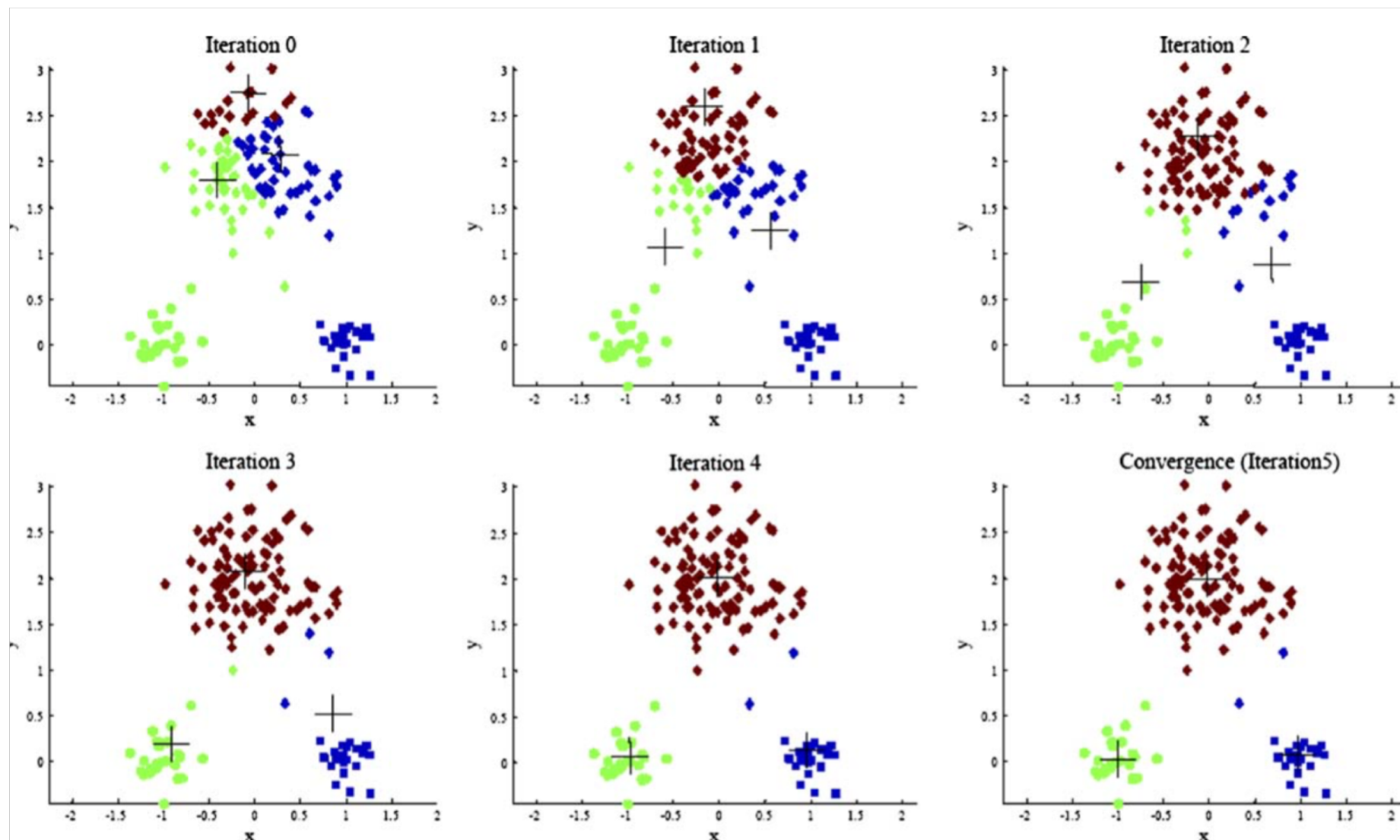


k -Means Clustering (1)

- Pick K random points as cluster centers (means)
- Alternate:
 - Assign data instances to closest mean
 - Assign each mean to the average of its assigned points
- Stop when no points' assignments change



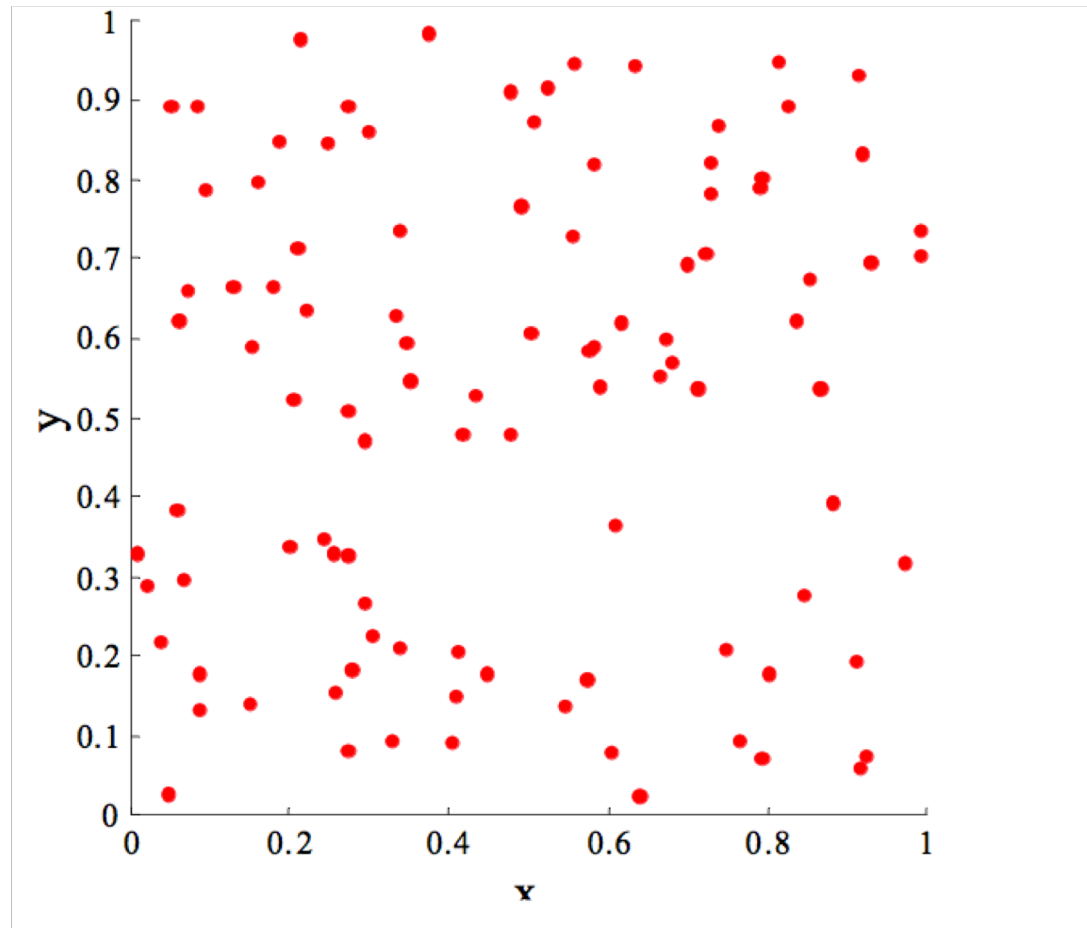
k -Means Clustering (2)



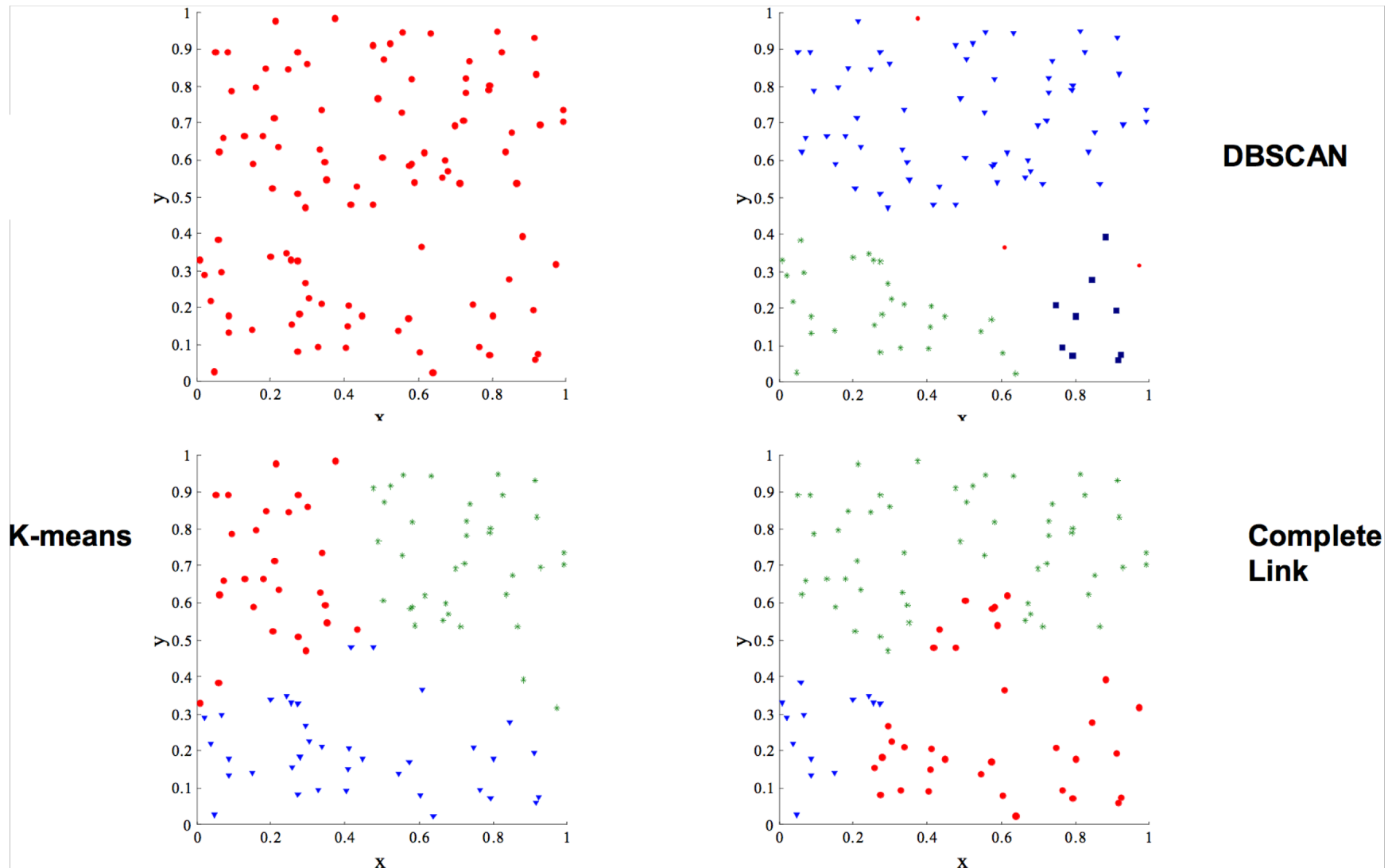
ML ala XKCD



What Makes for a “Good” Clustering?



Did I Cluster *Well*?



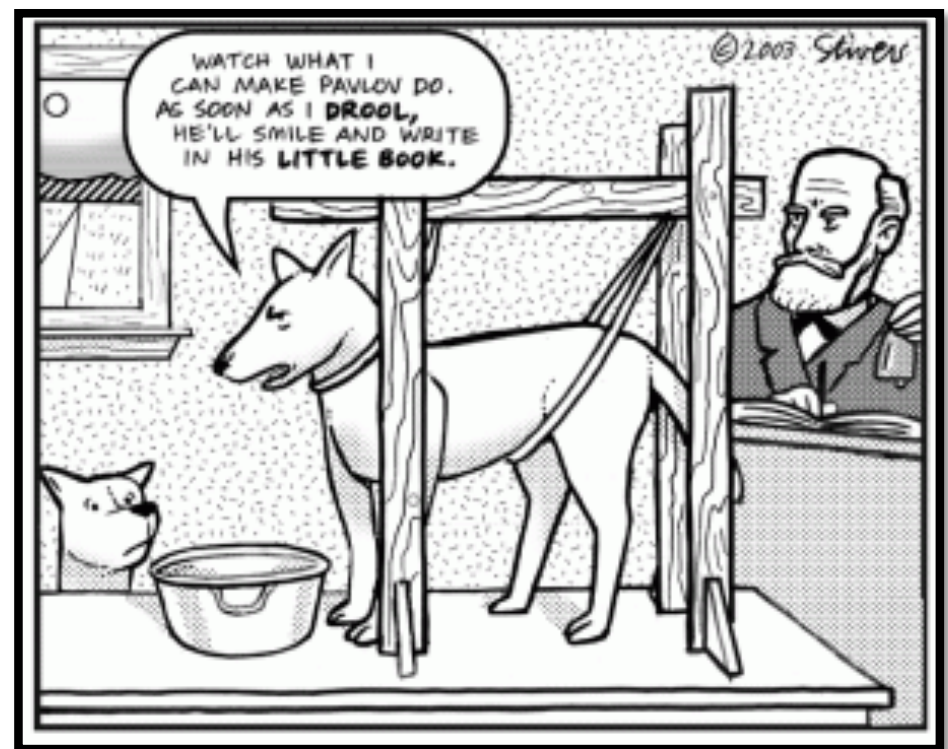
Key Evaluation Questions

1. Does non-random structure actually exist in the data?
2. What is the correct number of clusters?
3. How well do the results of a cluster analysis fit the data?
4. How well do the results of a cluster analysis adhere to externally known results?
5. Given two clusterings – which is better?



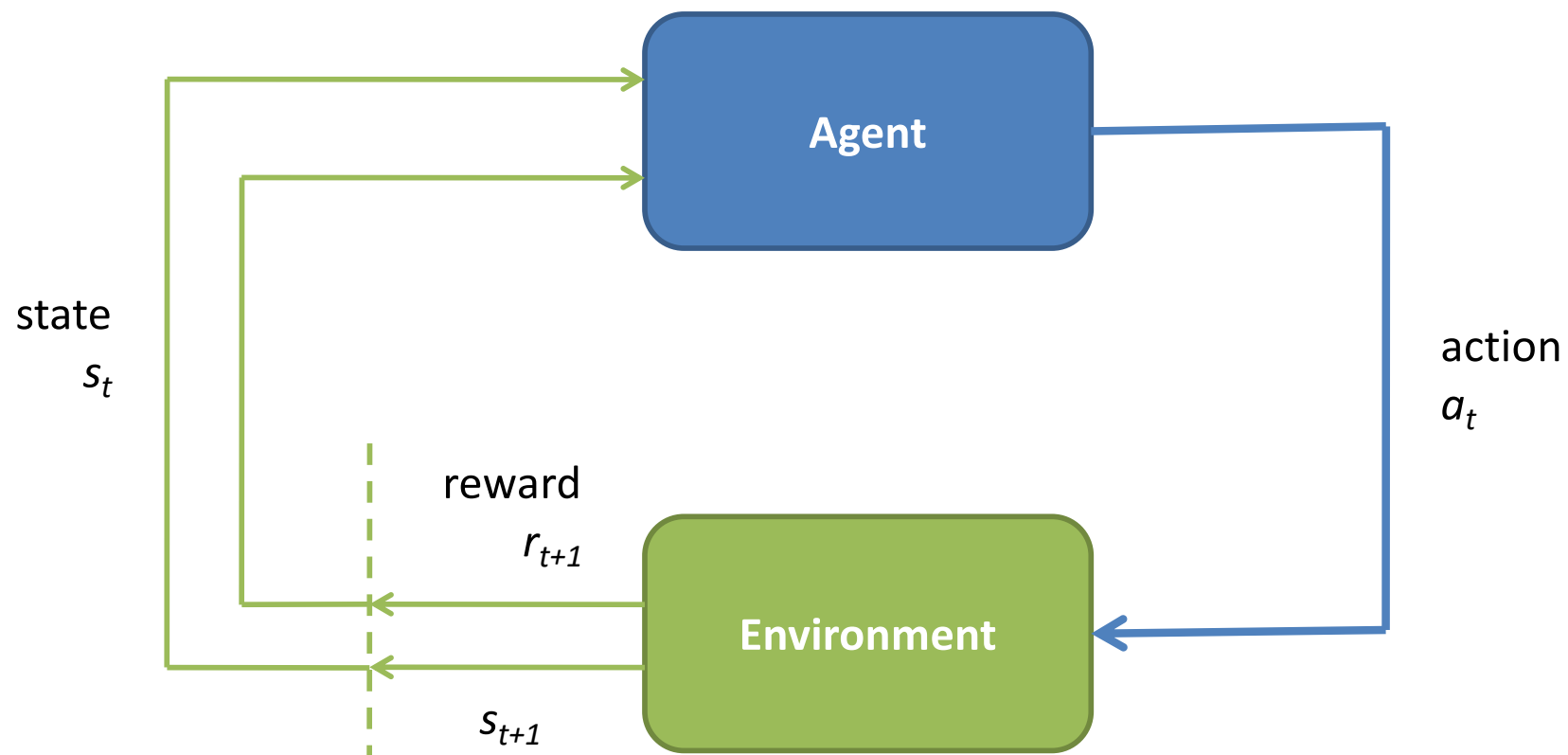
Reinforcement Learning (RL)

Choose actions to maximize future reward



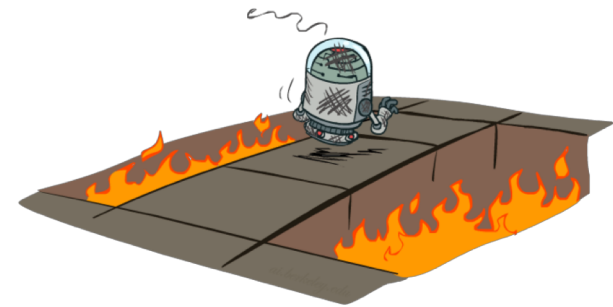
The RL Cycle

Issues. credit assignment, exploration vs. exploitation, reward function, ...



Temporal Difference (TD) Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

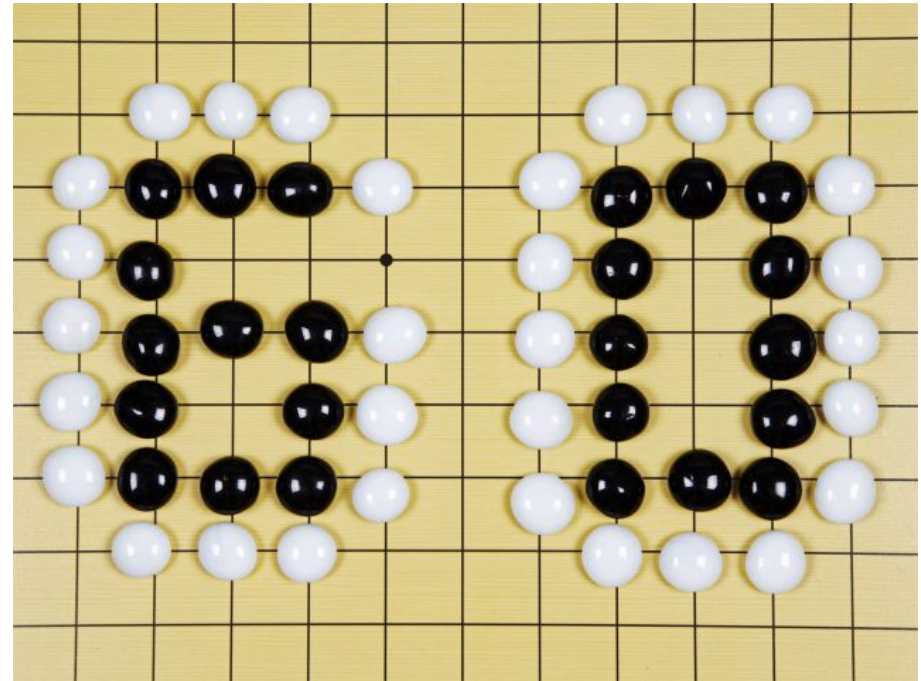


- Evidence that some neurons (dopamine) operate similarly
- Lead to world-class play via TD-Gammon (neural network trained via TD-learning)

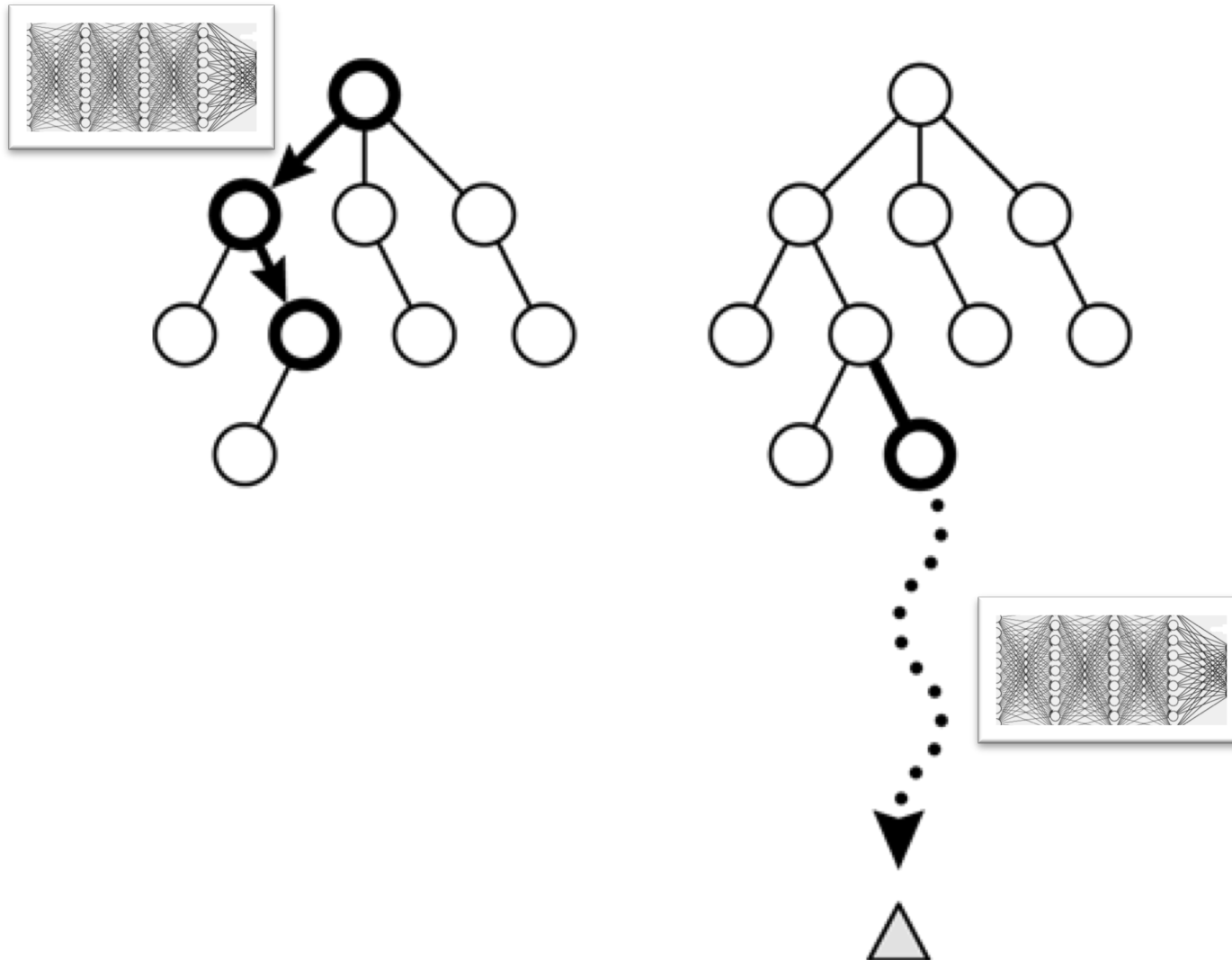


AI + ML = AlphaGo

- Until recently, AI was not competitive at champion level
 - 2015: beat Fan Hui, European champion (2-dan; 5-0)
 - 2016: beat Lee Sedol, one of the best players in the world (9-dan; 4-1)
 - 2017: beat Ke Jie, #1 in the world (9-dan; 3-0)
- MCTS + ANNs for **policy** (what to do) and **evaluation** (how good is a board state)



MCTS + 2xANNs



Checkup

1. Build an Atari system that learns game-winning techniques via actually playing and adjusting actions based upon score changes
 2. Given a dataset of past credit-card transactions (known to be fraudulent or not), build a system to identify future fraud
 3. If we assume incoming CS1 students are bi-modal, but normally distributed, find the average grades of the two groups
1. RL
 2. Supervised
 - Classification
 3. Unsupervised
 - Parameter estimation

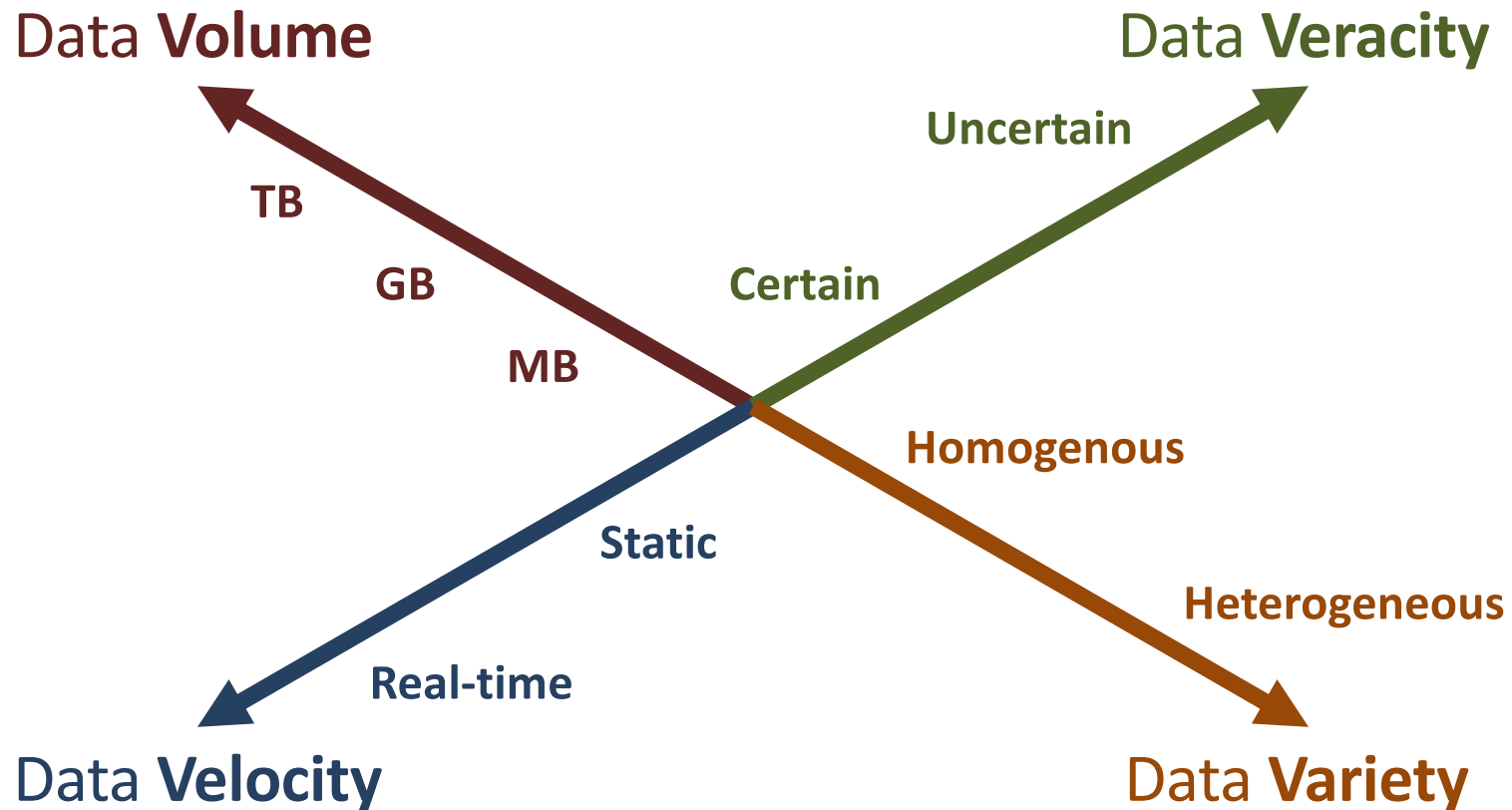


Issues/Challenges

- Big Data
- Curse of Dimensionality
- No Free Lunch



Big Data – The Four V's



Parametric algorithm: model does not grow with data size



The Curse of Dimensionality

“Various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.” – Wikipedia

- Memory requirement increases
- Required sampling increases
- Distance functions become less useful

...



No Free Lunch

- There is no universally best model – a set of assumptions that works well in one domain may work poorly in another
- We need many different models, and algorithms that have different speed-accuracy-complexity tradeoffs



Thank You :)

Questions?

Nate Derbinsky

Associate Teaching Professor

Director of Teaching Faculty

WVH 208B

n.derbinsky@northeastern.edu

<https://derbinsky.info>

