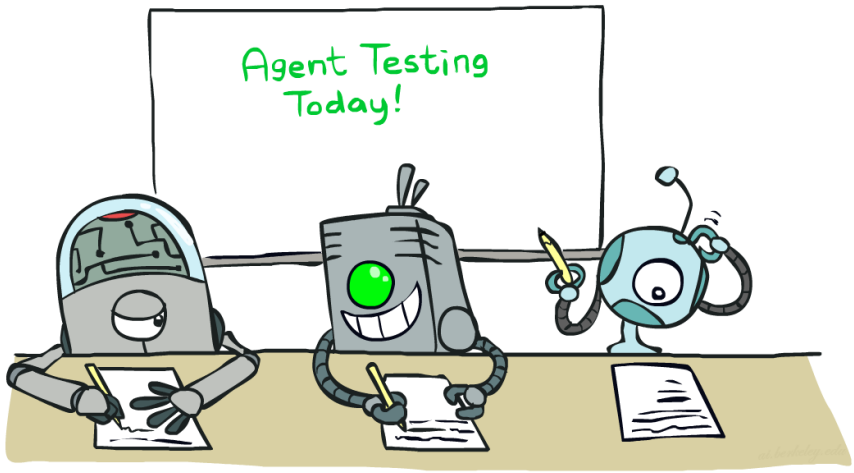
Introduction to Machine Learning

Lecture 7

How can we develop systems that learn from examples?

Agenda

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



What is Machine Learning (ML)?

The study/construction of algorithms that can learn from data

The study of algorithms that improve their performance **P** at some task **T** with experience **E**

– Tom Mitchell (CMU)

Fusion of algorithms, artificial intelligence, statistics, optimization theory, visualization, …

Natural Language Processing (NLP)

Modern NLP algorithms are typically based on statistical ML

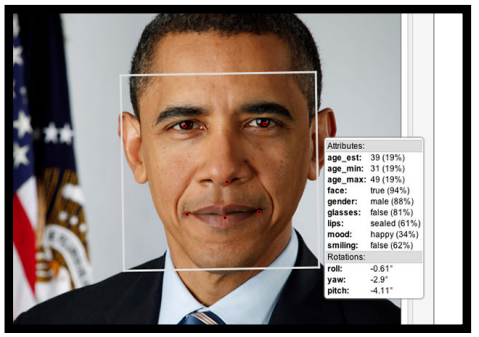
Applications

- Summarization
- Machine Translation
- Speech Processing
- Sentiment Analysis

…

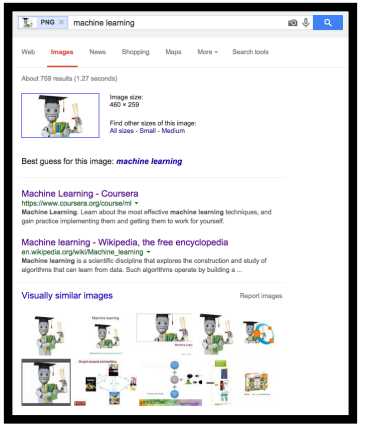
Computer Vision

Methods for acquiring, processing, analyzing, and understanding images

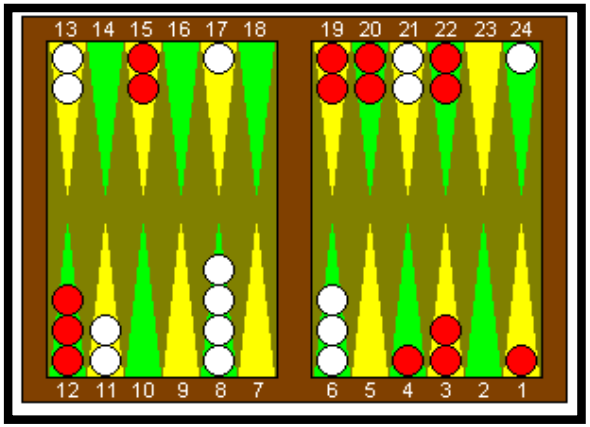


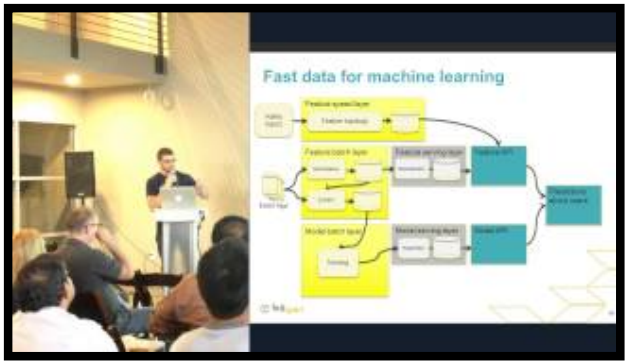
Applications

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



Games, Robotics, Medicine, Ads, …





Machine Learning is in Demand!

Position	Salary*
Data Scientist	**$113,436**
Machine Learning Engineer	**$114,826**
Software Engineer	$95,195

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."
– Josh Blumenstock (UW)

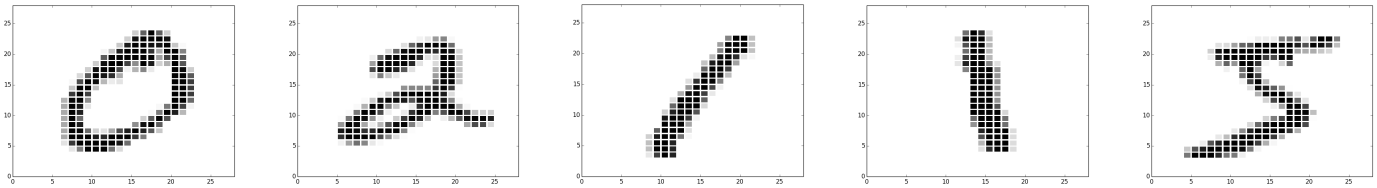
"Data Scientist = statistician + programmer + coach + storyteller + artist"
– Shlomo Aragmon (Ill. Inst. of Tech)

*glassdoor.com, National Avg as of March 19, 2017

Key Terminology

Let's consider a task [that we will revisit in greater detail]: handwritten digit recognition

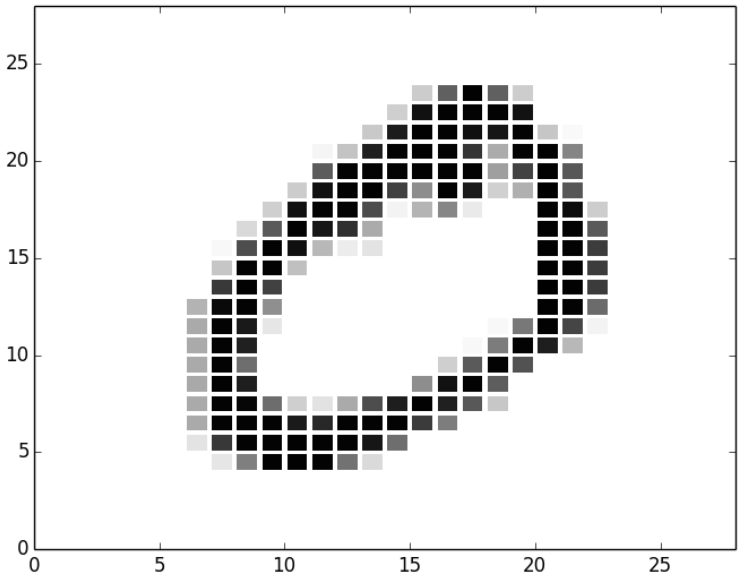
Given as input…



Have the computer correctly identify…

0 2 1 1 5

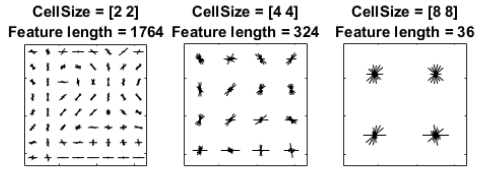
Instances and Features



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

Spot the Vocabulary!

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

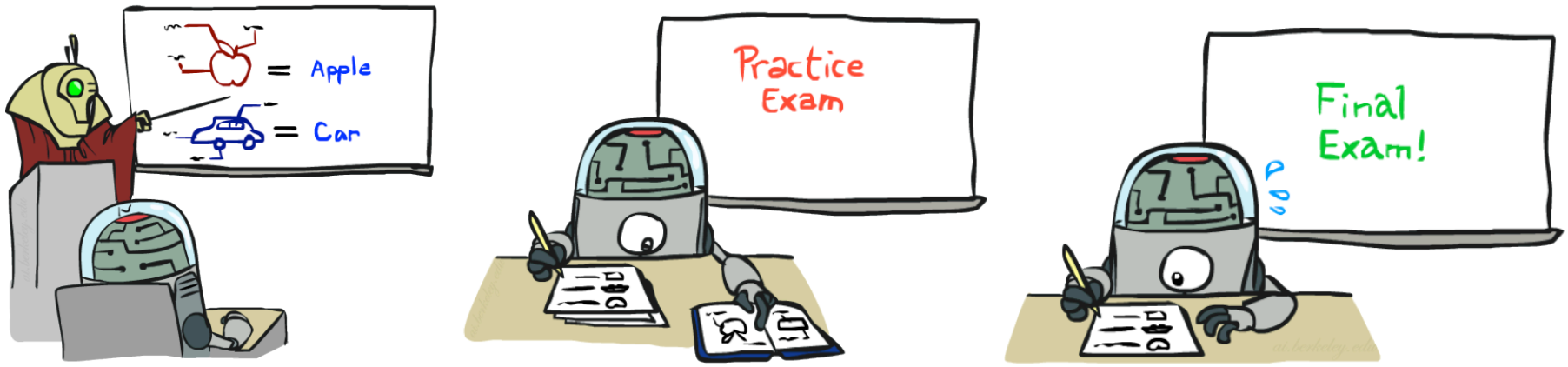
target*, *label*, *class*, *concept, dependent

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

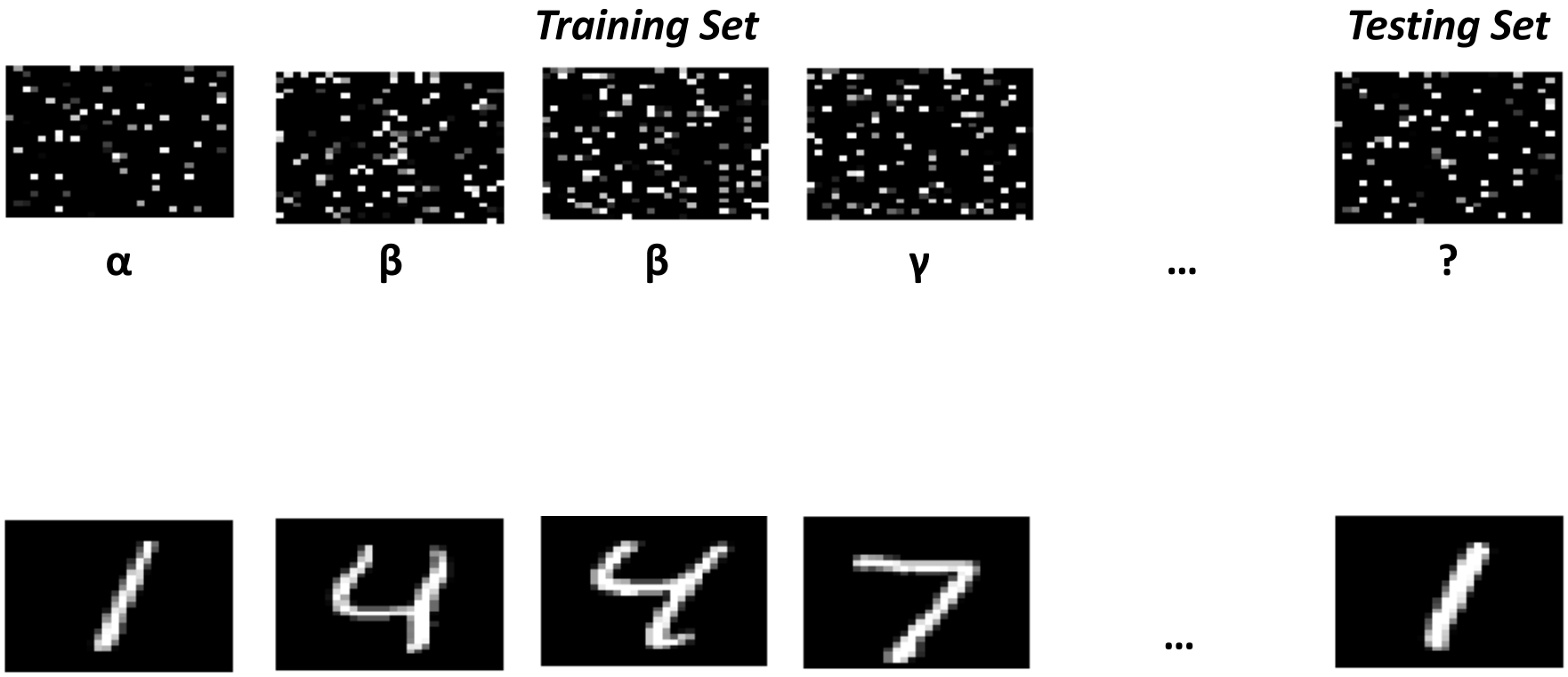
Machine Learning Tasks

- ***Supervised***
  - Given a ***training set*** and a target variable, ***generalize***; measured over a ***testing set***
- ***Unsupervised***
  - Given a dataset, find "interesting" patterns; potentially no "right" answer
- ***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)



Goal: ***generalization***

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

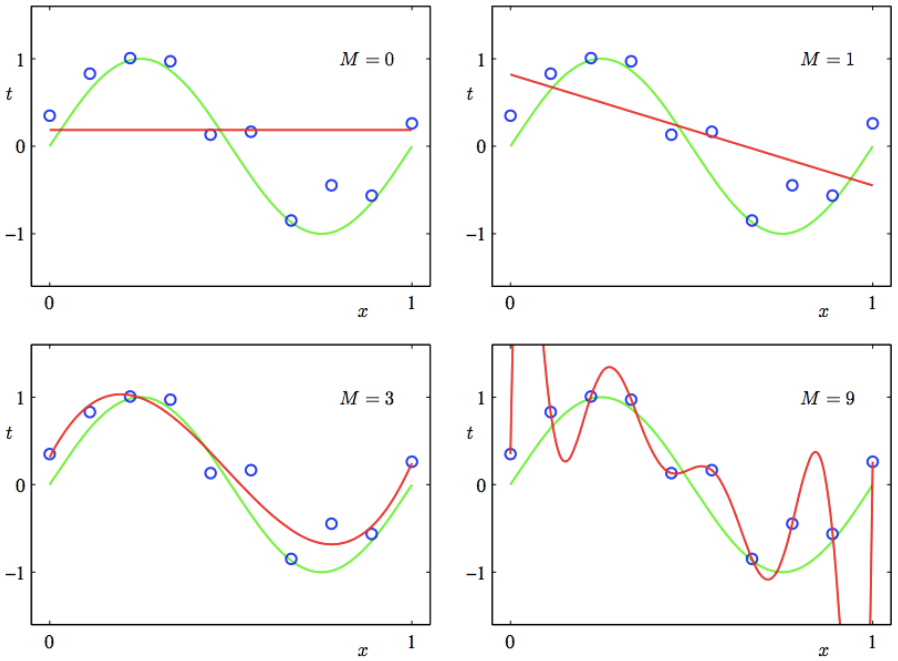
- Continuous target

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

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!

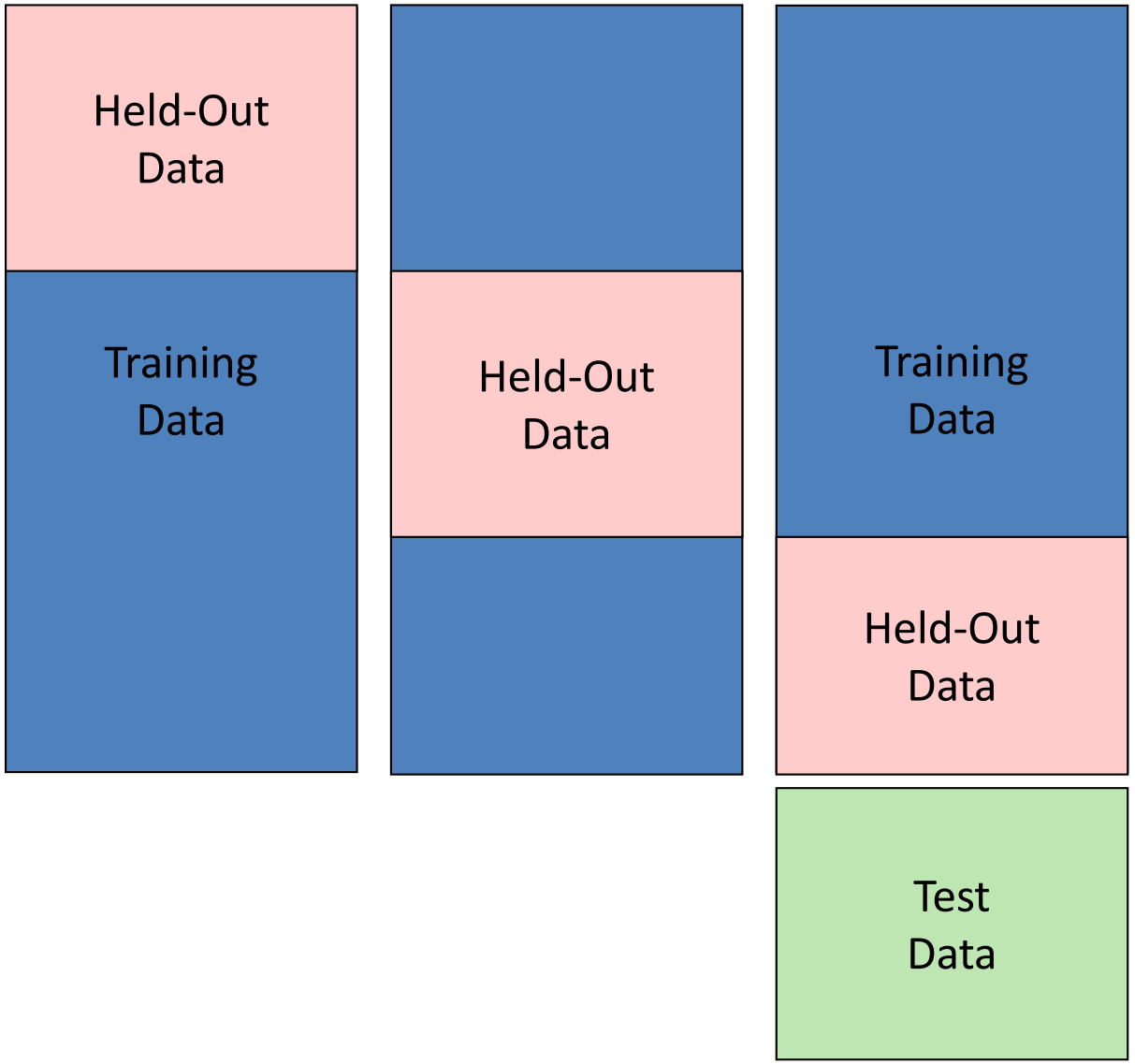
- So how do we not delude ourselves, achieve high performance, *and* a reasonable expectation of future performance?

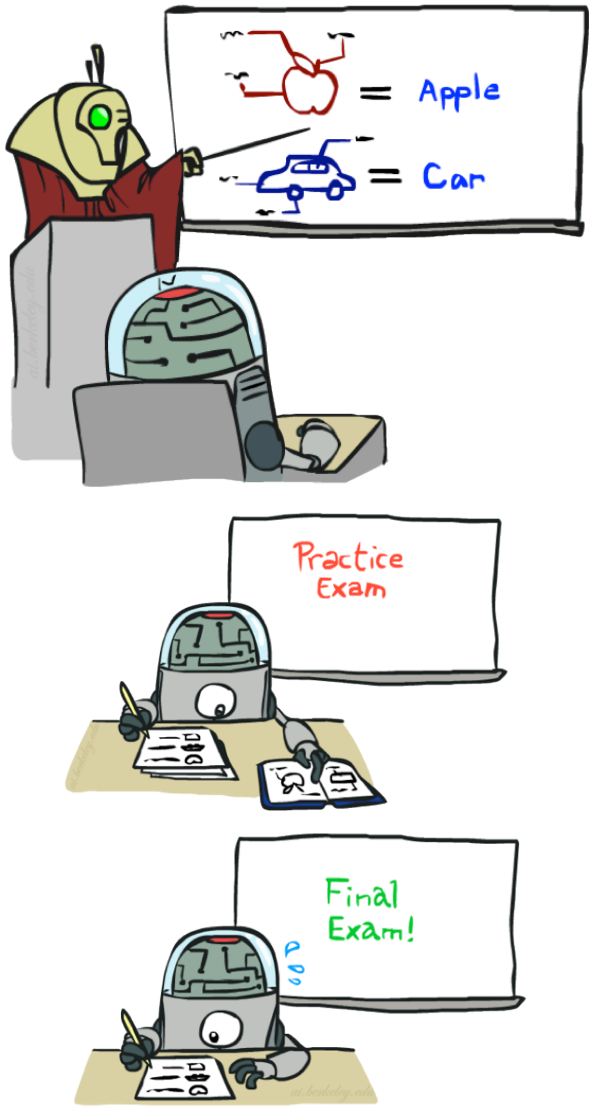
k-Fold Cross-Validation

- Basic approach
  - Divide the data into k randomly selected partitions (typically 10)
  - For each, use the fold as test data, the remainder as training data (i.e. repeat the train/test process k times)
  - Average results

- To control for unfortunate outcomes in random selection, consider repeating (e.g. 10 x 10-fold cross validation = 100 train/test)
  - Expensive!

k-Fold Cross Validation Visualized





Common Algorithms

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

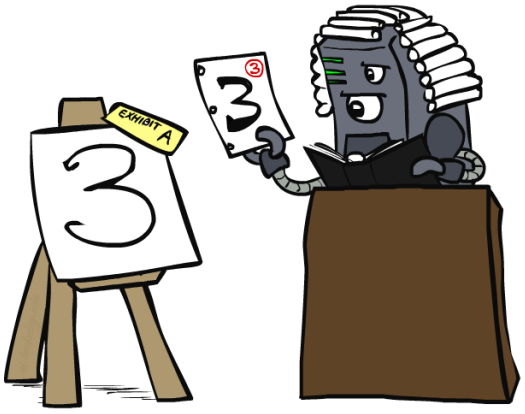
kNN

Training

- Store all examples

Testing

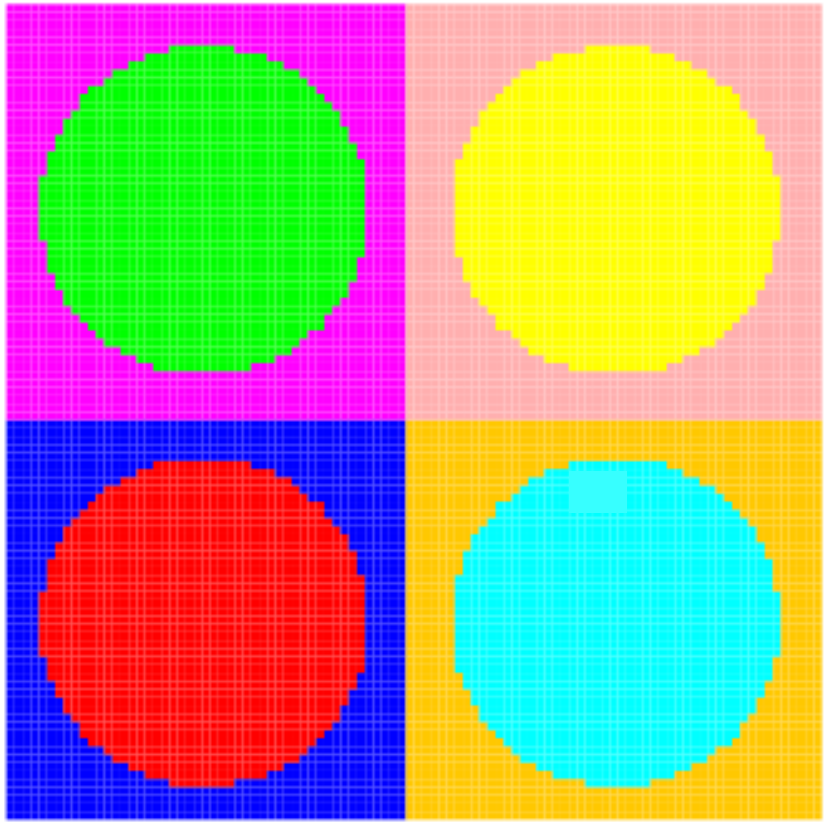
- Find the nearest *k* neighbors to target
  - Via distance function
- Vote on class



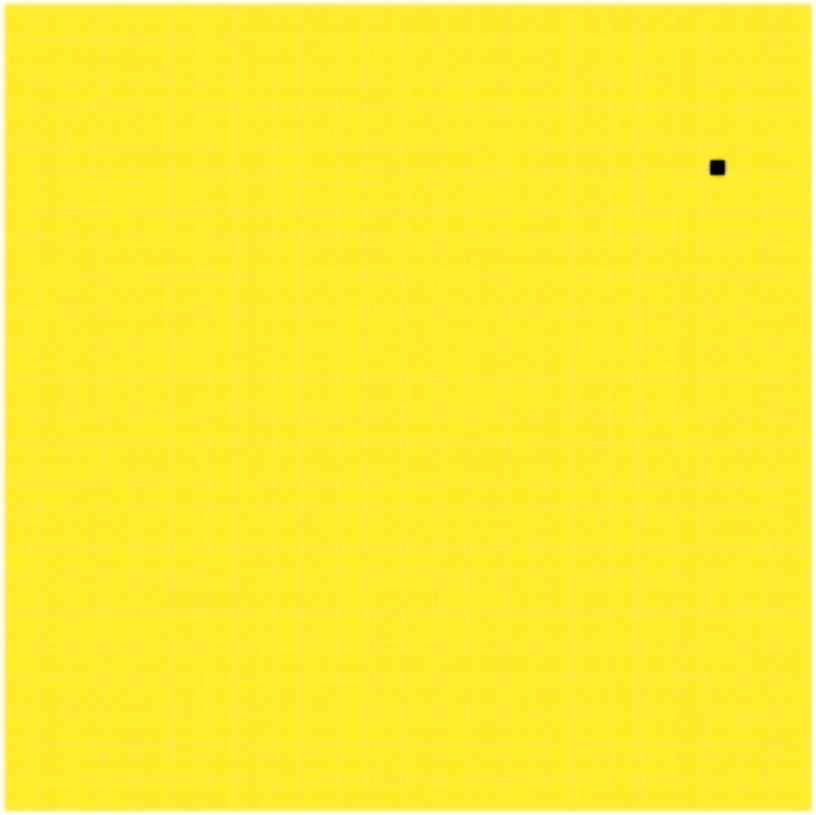
Non-parametric algorithm (i.e. grows with |examples|!)

2D Multiclass Classification

Boundary Tree



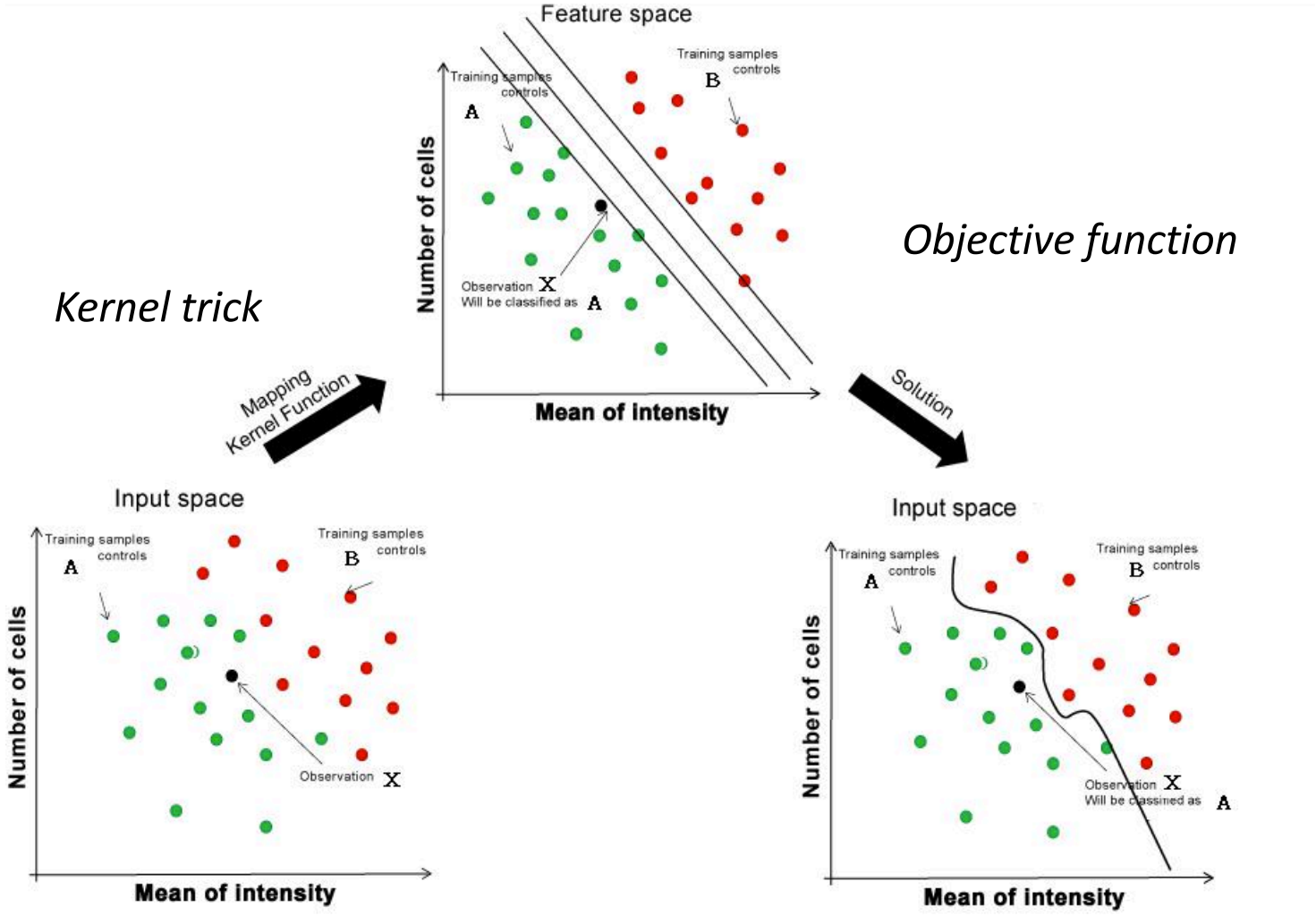
1-NN via Linear Scan



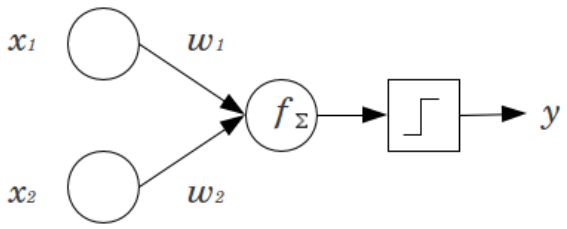
Decision Trees/Forests

Explicit knowledge Representation, vs. ***implicit***

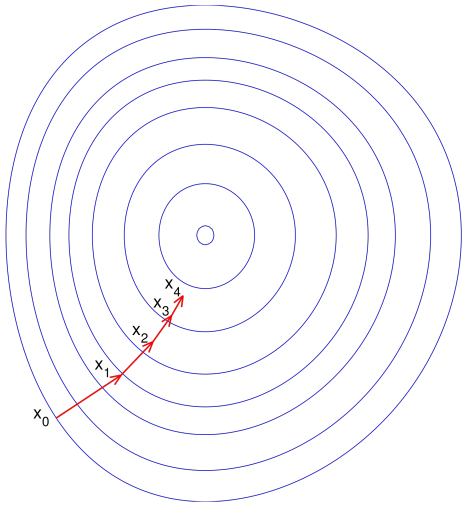
Support Vector Machine (SVM)



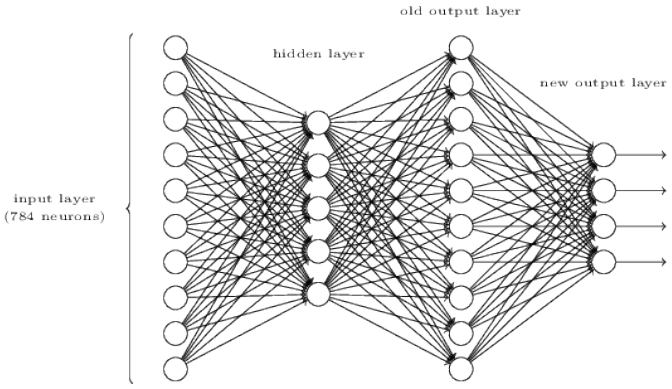
Artificial Neural Networks (ANN)



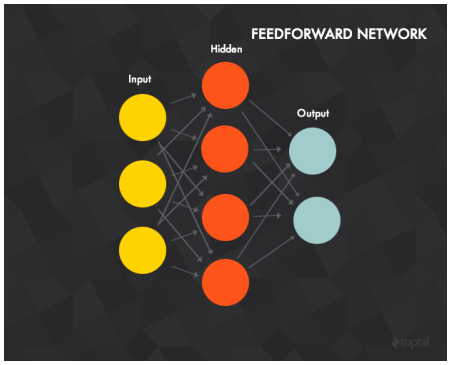
Perceptron
Linear classifier



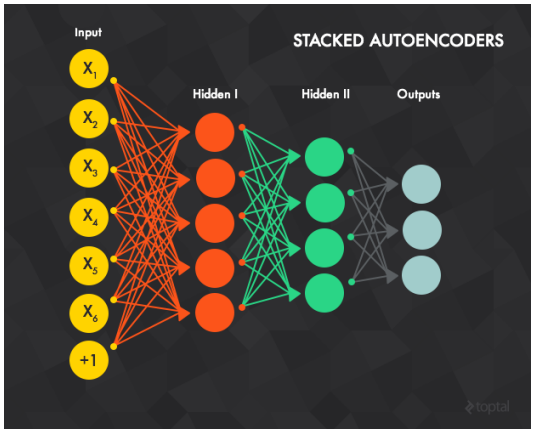
Gradient descent



Backpropagation



Feedforward vs.
Recurrant



Deep Architectures
Vanishing Gradient

Unsupervised Learning

No right answer, find "interesting" structure or patterns in the data

Tasks

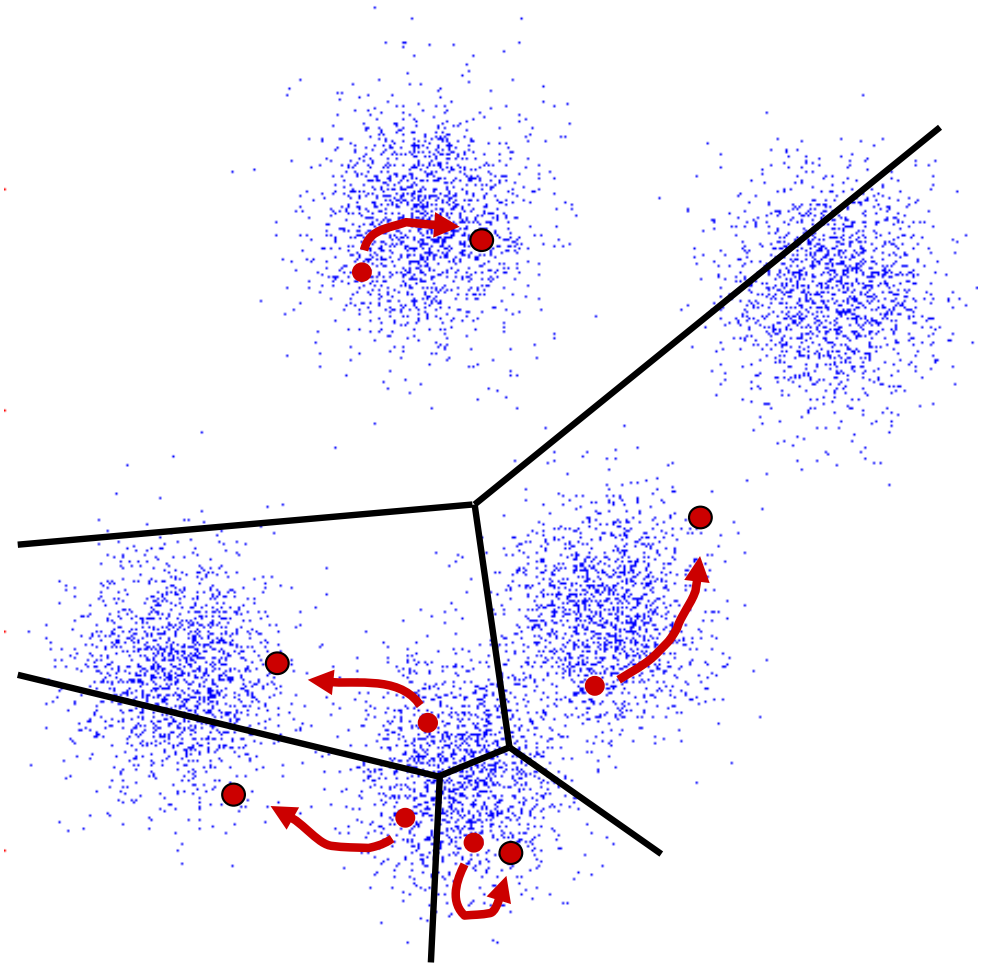
- Clustering
- Dimensionality reduction
- Density estimation
- Discovering graph structure
- Matrix completion

Common Algorithms

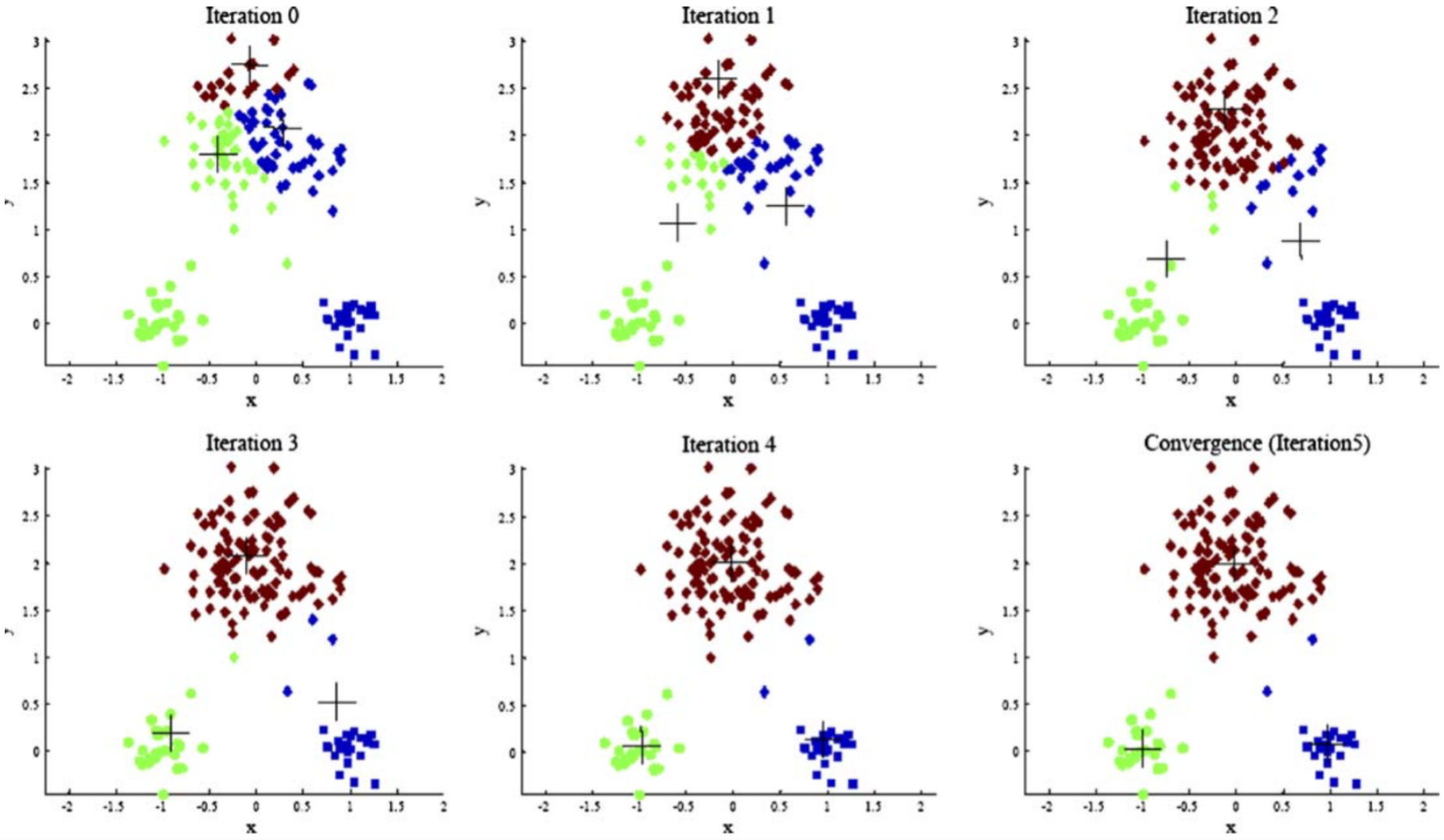
- k-Means Clustering
- Collaborative Filtering
- Principle Component Analysis (PCA)
- Expectation Maximization (EM)
- Artificial Neural Networks (e.g. RBM)

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

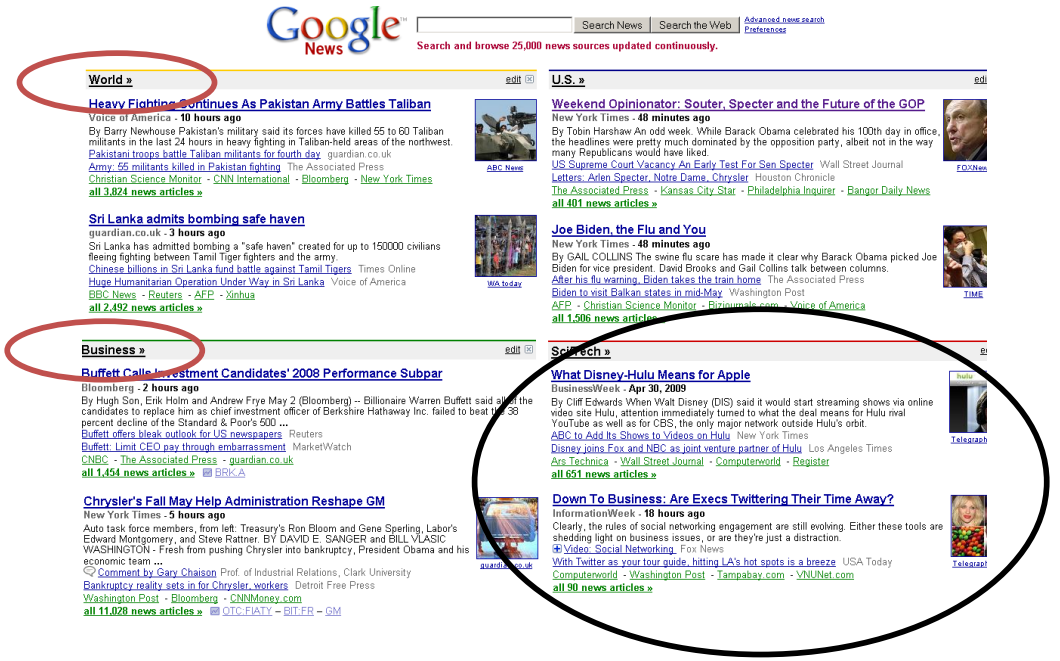


k-Means Clustering (2)



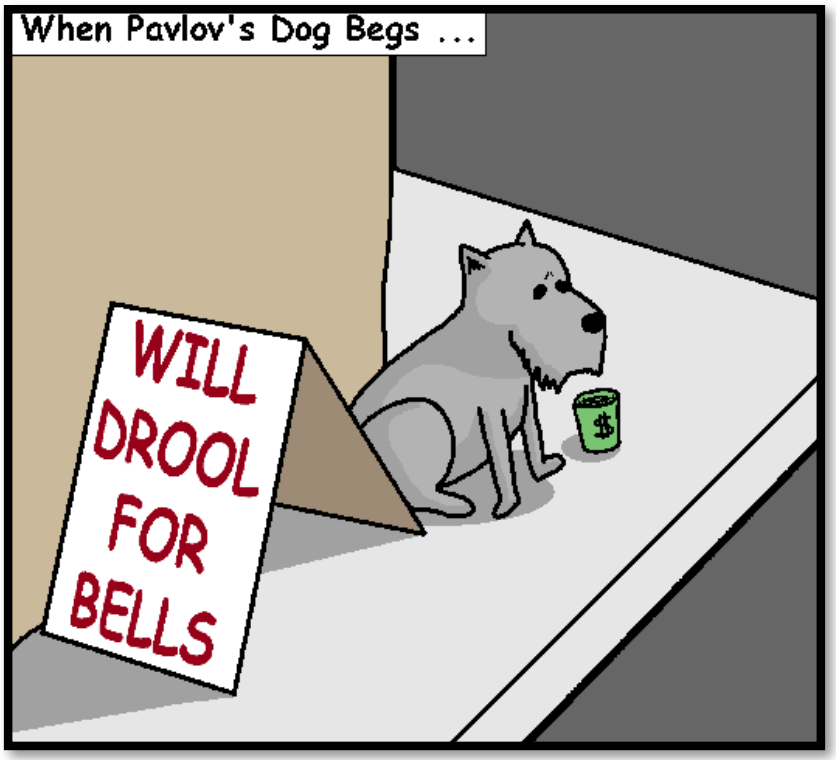
Example: Google News

- Group articles
  - Unsupervised
- Group labels
  - Supervised



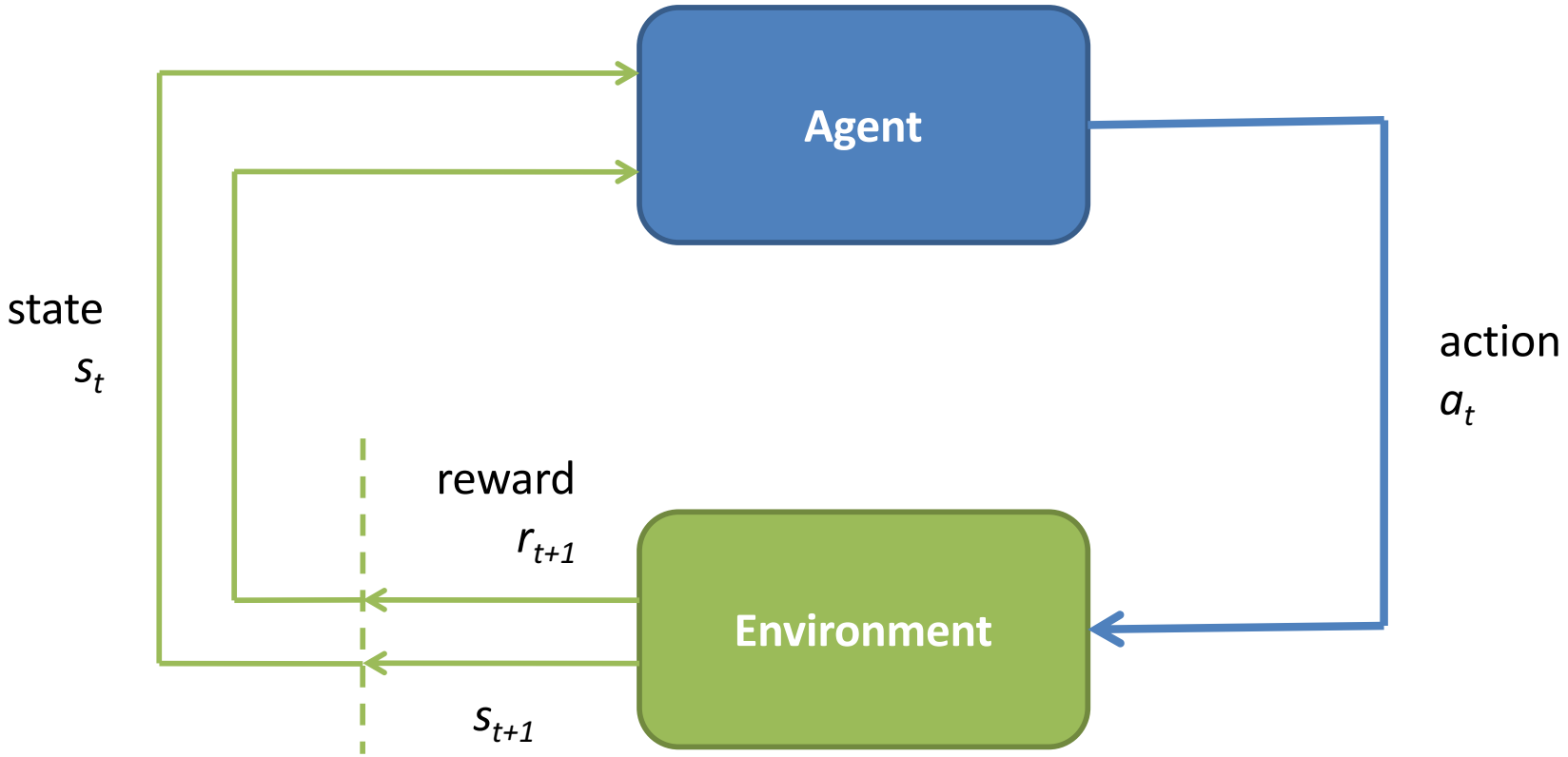
Reinforcement Learning (RL)

Choose actions to maximize future reward



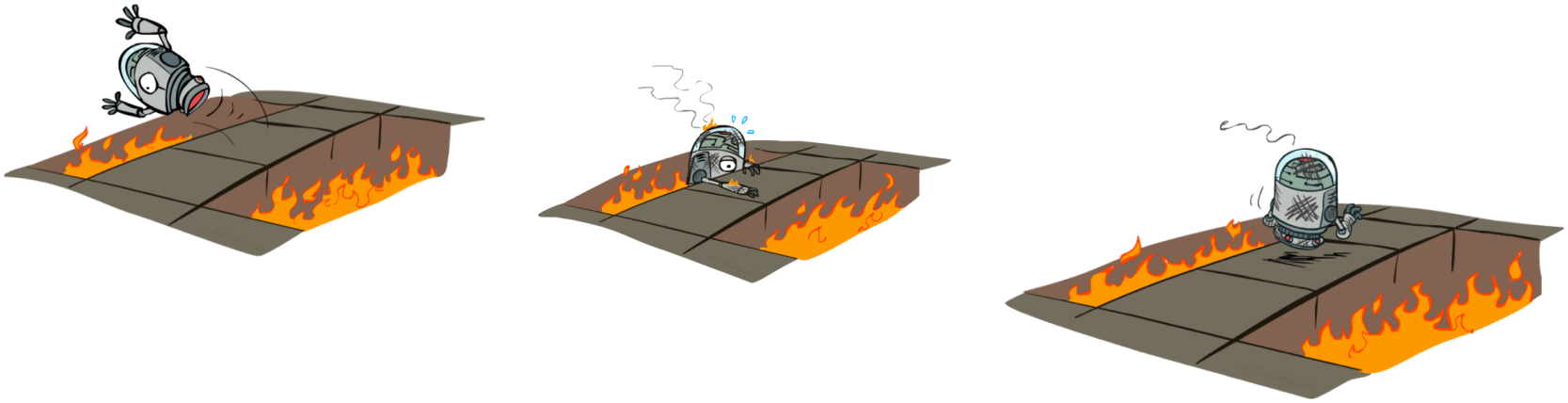
The RL Cycle

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



Temporal Difference (TD) Learning

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

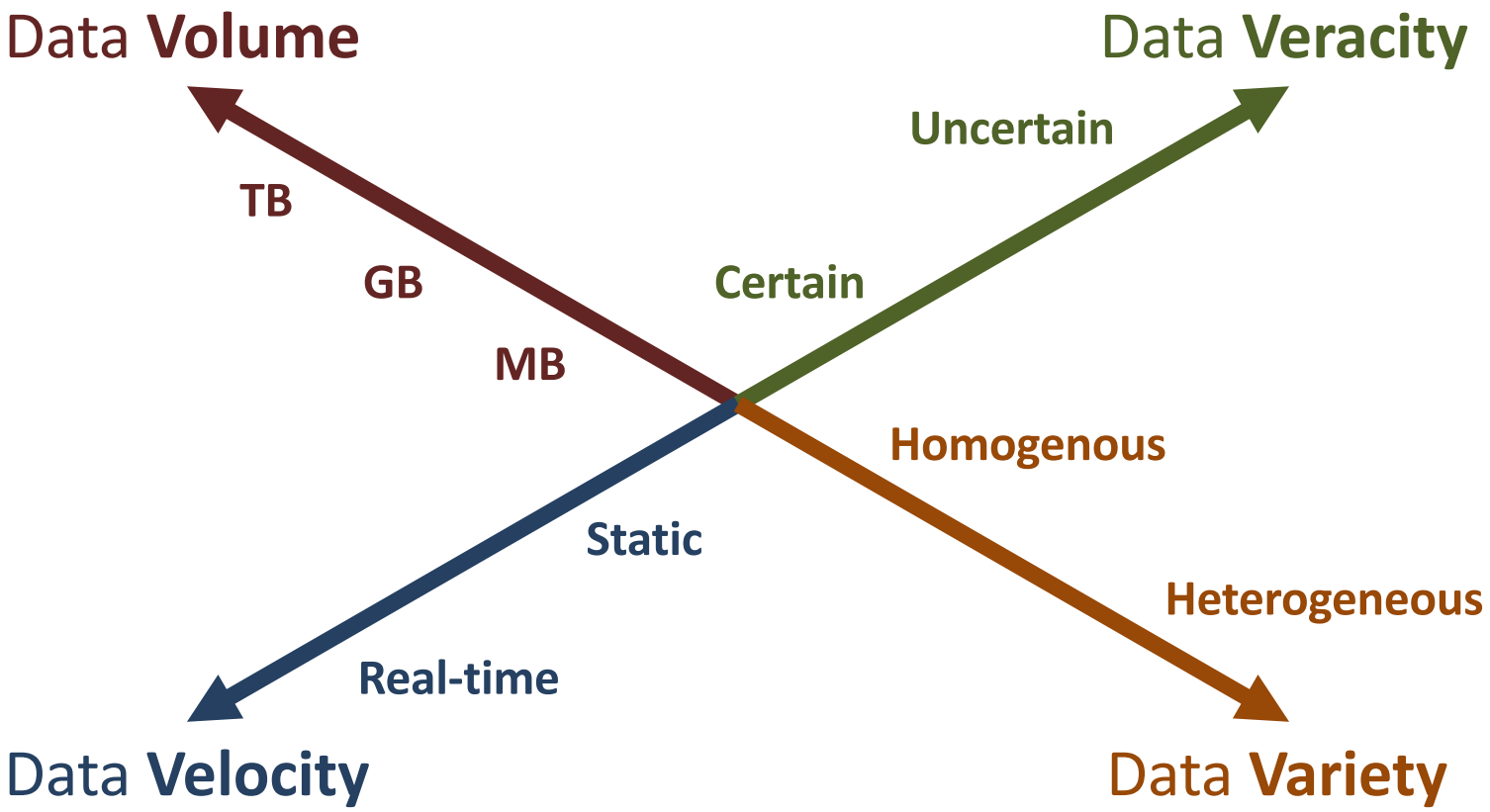


- Evidence that some neurons (dopamine) operate similarly
- Lead to world-class play via TD-Gammon (neural network trained via TD-learning)

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

Machine Learning Applications

1. Collect the data
2. Preprocess the data
3. Analyze the input data
   - Model selection
4. Train, evaluate
5. Deployment

Collecting Data

- Public data sets
  - RSS feeds
- Application Programming Interface (API)
- Generate via sensors/logs

Preprocessing

- Converting formats
  - Binning
  - Mapping
  - Cleaning

Data Analysis

- Identifying incorrect/outlier/missing data
- Use domain knowledge & simple statistical/visual results
  - Model selection
  - Feature selection/production
- Understand under/over-representation

Train, Evaluate

- Methods for meta-parameter selection (e.g. *k* in kNN)
  - Cross validation
- Iteration is likely, might consider multiple models if algorithmic assumptions to not match application/data

Application Deployment

- Automate the data collection/processing pipeline
- May have to re-iterate given…
  - Real-world data
  - Performance constraints
  - Changes in application requirements

Summary

- Machine Learning is the study of algorithms that can learn from data
- Datasets are typically represented as a set of *n* instances/examples, each composed of k-dimensional feature vectors
- Machine Learning tasks include supervised (classification, regression), unsupervised, and reinforcement
- In the search for generalization over training data, supervised algorithms are seeking an ideal tradeoff between under/over fitting
- Machine Learning applications involve an iterative process of data collection/preprocessing/analysis, training/evaluation, and eventual deployment