A Subjective Introduction into Machine Learning

Georgiy Bobashev, Ph.D.

Contact: Georgiy Bobashev
Center for Data Science
RTI International
3040 Cornwallis Rd. P.O. Box 12194
Research Triangle Park, NC 27709
tel: (919) 541-6167
fax: (919) 541-6722
e-mail: bobashev@rti.org
web: www.rti.org/bobashev

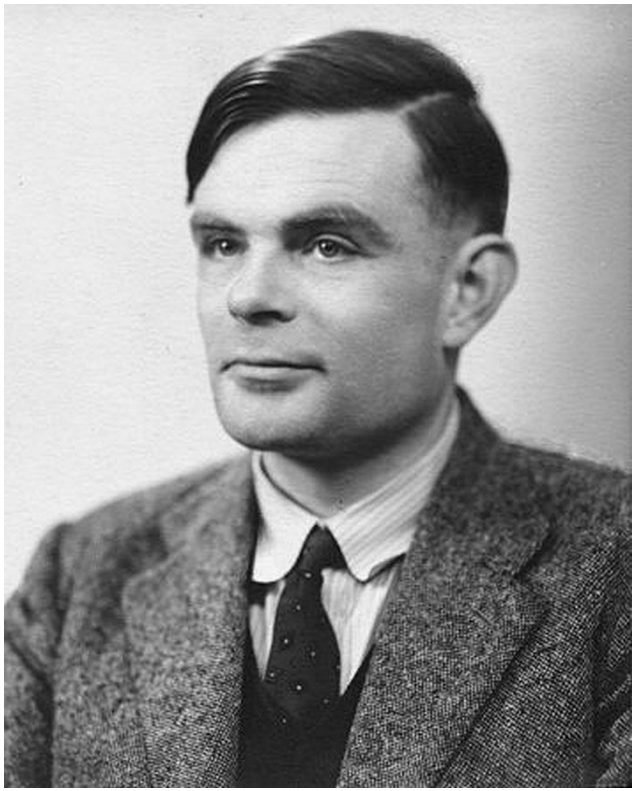
RTI International is a registered trademark and a trade name of Research Triangle Institute.

Objectives

- Provide an overview/demystification of ML
- Illustrate some of the main approaches
- Illustrate Deep Learning with Tensorflow

What is Machine Learning?

- From pattern recognition to computational learning theory in artificial intelligence
- Algorithms that can learn from and make predictions on data



Some Terminology (from Helen Shin)

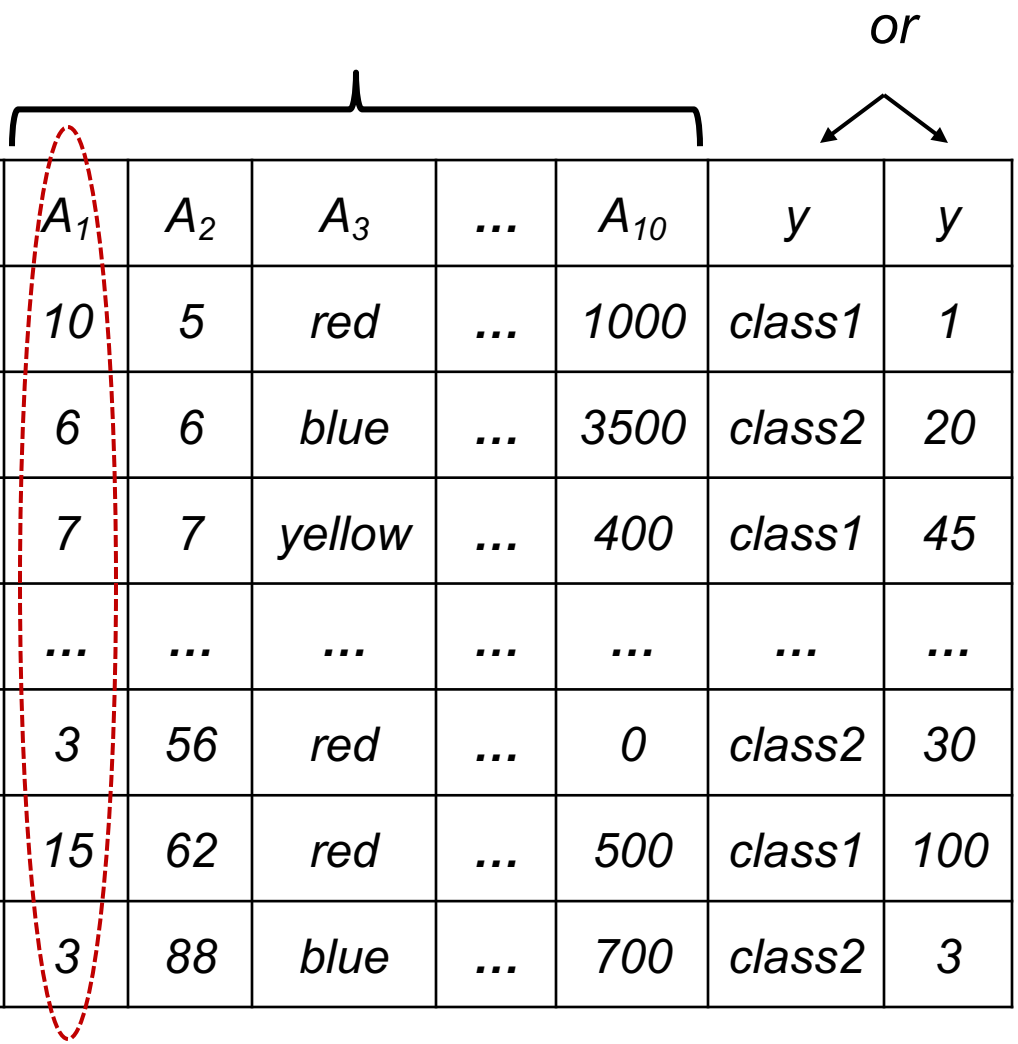
Data Set

20 records,
20 samples,
20 observations,
20 objects,
20 data points,
20 individuals,
20 experimental units,
etc.

	A_1	A_2	A_3	...	A_{10}	y	y
X_1	10	5	red	...	1000	class1	1
X_2	6	6	blue	...	3500	class2	20
X_3	7	7	yellow	...	400	class1	45
...	...	...	...	...	...	...	...
X_{18}	3	56	red	...	0	class2	30
X_{19}	15	62	red	...	500	class1	100
X_{20}	3	88	blue	...	700	class2	3

Some Terminology (continued)

A_j
attribute,
feature,
descriptor,
input variable,
predictor variable,
exogeneous variable,
etc.



	A_1	A_2	A_3	...	A_{10}	y	y
X_1	10	5	red	...	1000	class1	1
X_2	6	6	blue	...	3500	class2	20
X_3	7	7	yellow	...	400	class1	45
...	...	...	...	...	...	...	...
X_{18}	3	56	red	...	0	class2	30
X_{19}	15	62	red	...	500	class1	100
X_{20}	3	88	blue	...	700	class2	3

Some Terminology (continued)

X_i
input,
predictor,
etc.

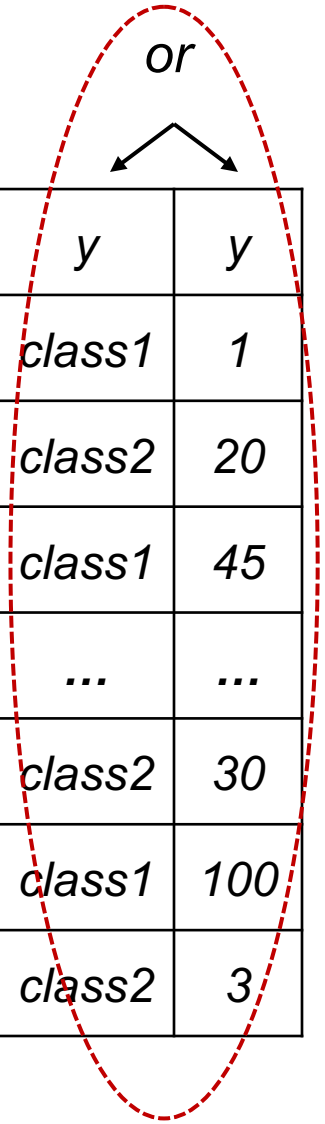
or

	A_1	A_2	A_3	...	A_{10}	y	y
X_1	10	5	red	...	1000	class1	1
X_2	6	6	blue	...	3500	class2	20
X_3	7	7	yellow	...	400	class1	45
...	...	...	...	...	...	...	...
X_{18}	3	56	red	...	0	class2	30
X_{19}	15	62	red	...	500	class1	100
X_{20}	3	88	blue	...	700	class2	3

Some Terminology (continued)

Y_i
output variable,
response,
target variable,
endogeneous variable,
label,
etc.

	A_1	A_2	A_3	...	A_{10}	y	y
X_1	10	5	red	...	1000	class1	1
X_2	6	6	blue	...	3500	class2	20
X_3	7	7	yellow	...	400	class1	45
...	...	...	...	...	...	...	...
X_{18}	3	56	red	...	0	class2	30
X_{19}	15	62	red	...	500	class1	100
X_{20}	3	88	blue	...	700	class2	3

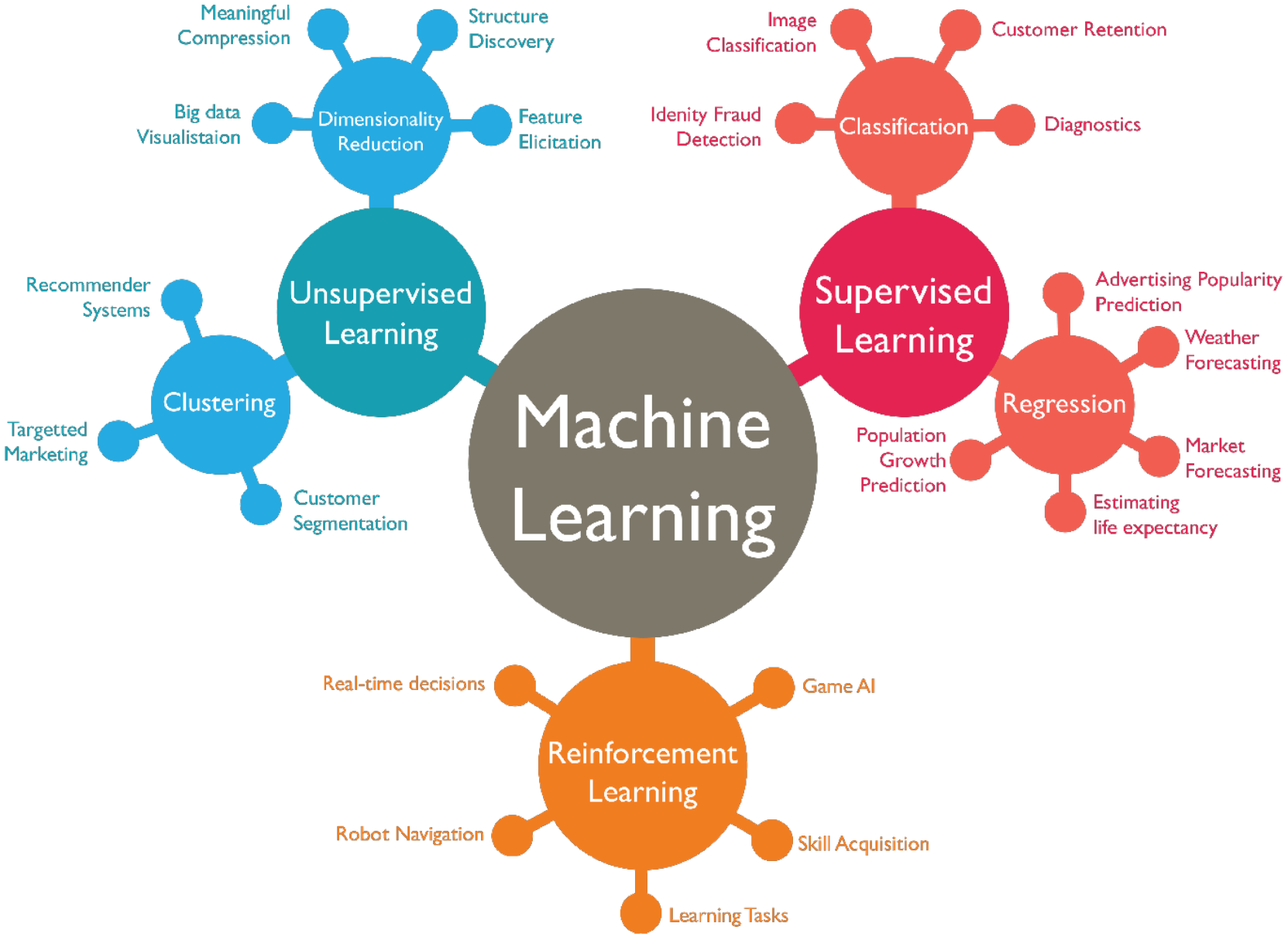


Difference Between Images and Surveys

- Image: rows and columns can be transposed (flipped) but not reshuffled
- Surveys: rows and columns can be reshuffled but not transposed

	A_1	A_2	A_3	...	A_{10}	y	y
X_1	10	5	red	...	1000	class1	1
X_2	6	6	blue	...	3500	class2	20
X_3	7	7	yellow	...	400	class1	45
...	...	...	...	...	...	...	...
X_{18}	3	56	red	...	0	class2	30
X_{19}	15	62	red	...	500	class1	100
X_{20}	3	88	blue	...	700	class2	3

Types of ML

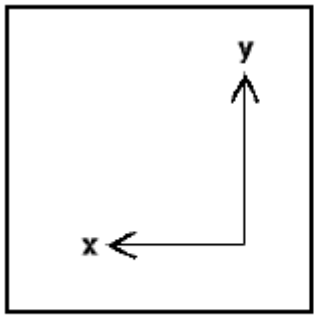

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Distance/Similarity Measures

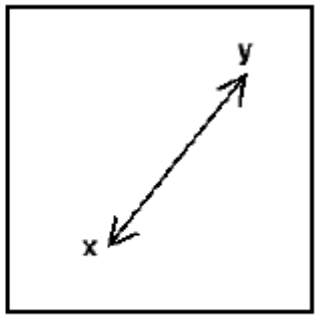
- Input: Two data points or sets of data, Output: Similarity (distance) measure between them
- Most common
  - Euclidean
  - Correlation (e.g., Mahalanobis)
  - Manhattan
  - Distribution-based
  - Jaccard (percent matches)
- More exotic
  - Graph-based
  - Rule-based

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}.$$



Manhattan



Euclidean

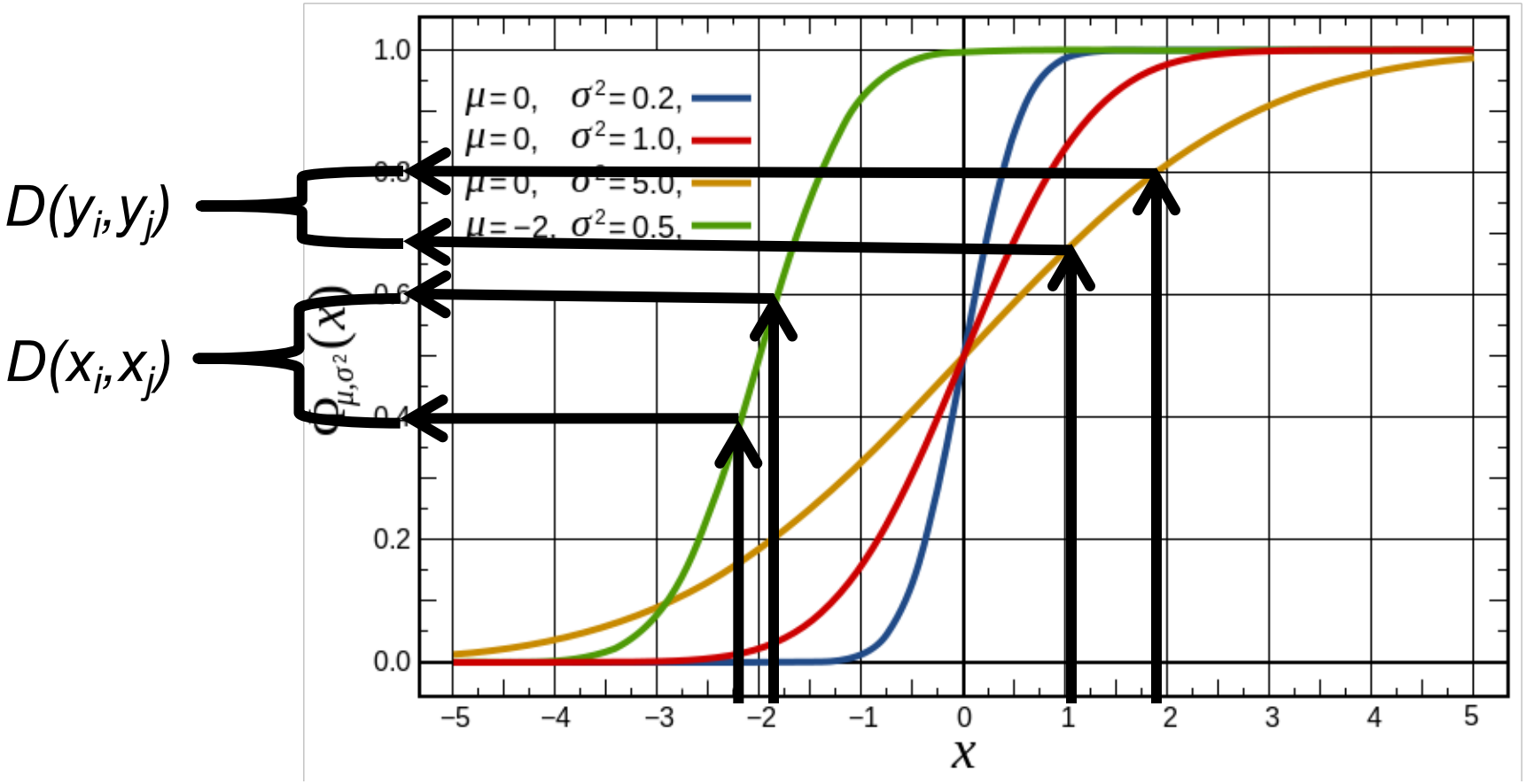
- Similarity between the observations and between the variables
- What is the difference between the variables and subjects?

Exploratory Analysis: *Similarity Measures*

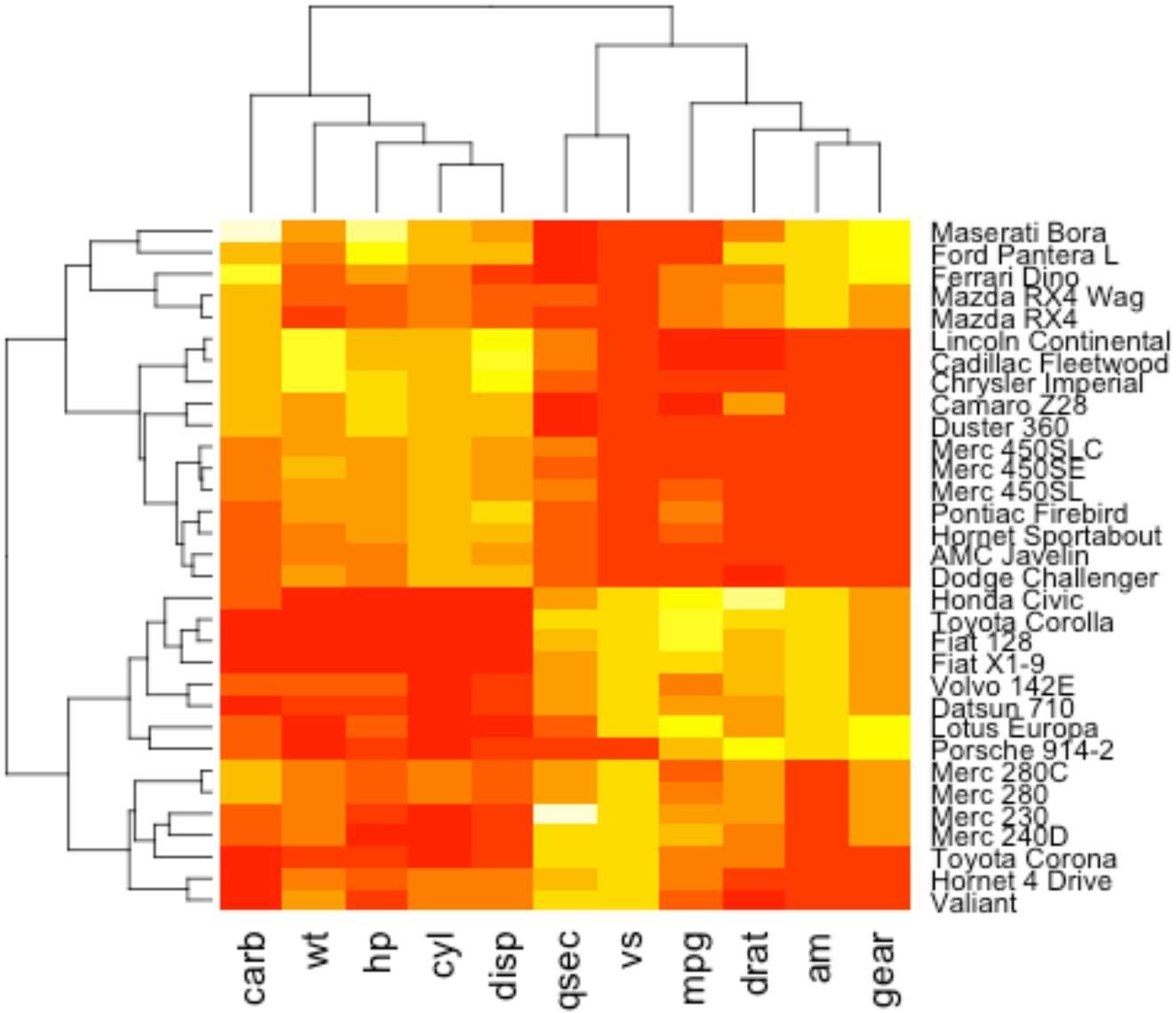
- Distribution-based:

$D(x_i,x_j)=|F(x_i)-F(x_j)|$

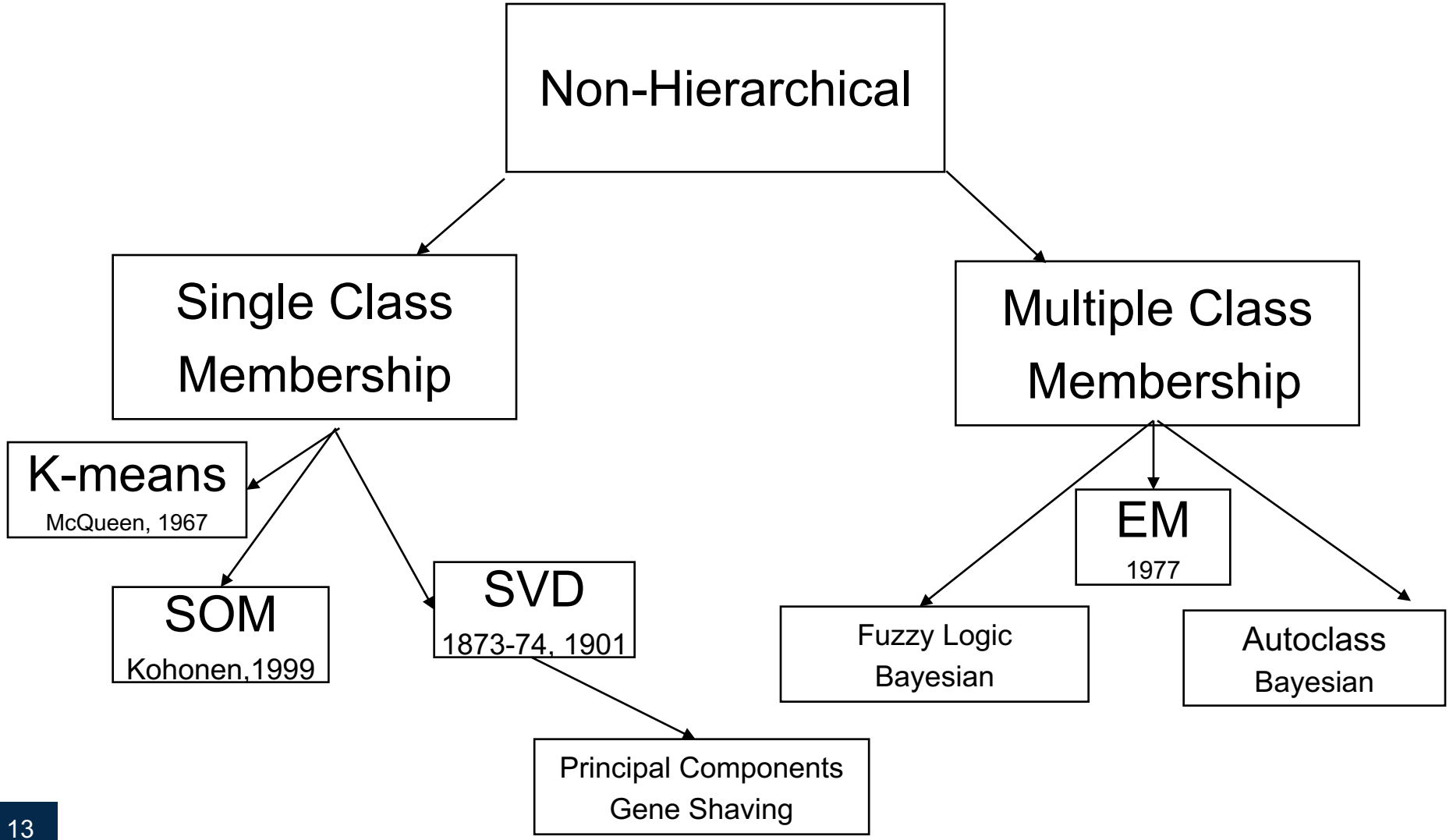
$D(i,j)=\Sigma|F(x_i)-F(x_i)|/n$



Unsupervised 2-way Hierarchical Clustering



Unsupervised Clustering
Pattern Identification: Non-hierarchical

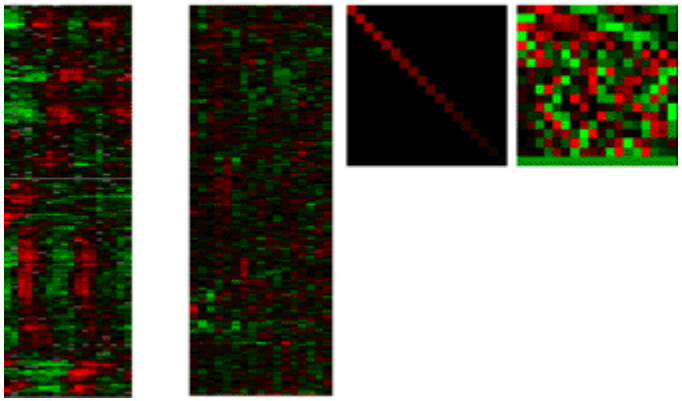


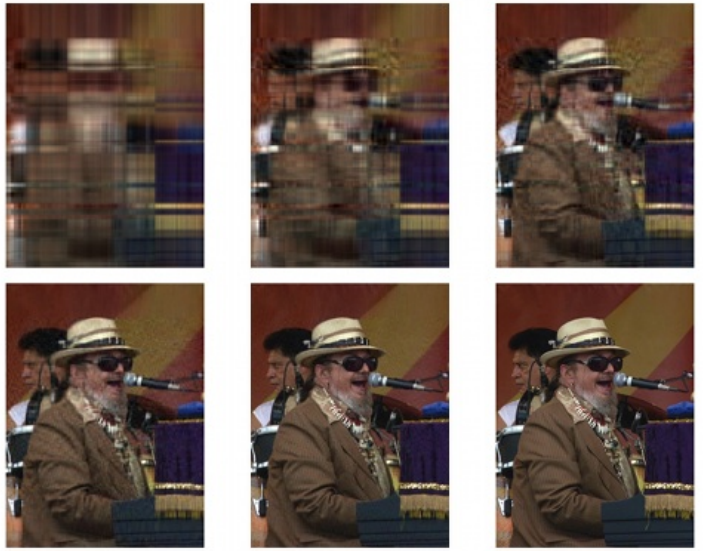
Singular Value Decomposition (Dimensionality Reduction)
Similar to Principal Components

Essentially it is the filtering of an image

1. Calculate the eigenvalues and eigenvectors and transform the data
2. Isolate a number of desired eigenvalues
3. Keeping the small number of eigenvectors decompose the matrix to obtain the smaller matrix

$$A = U \cdot W \cdot V^T$$

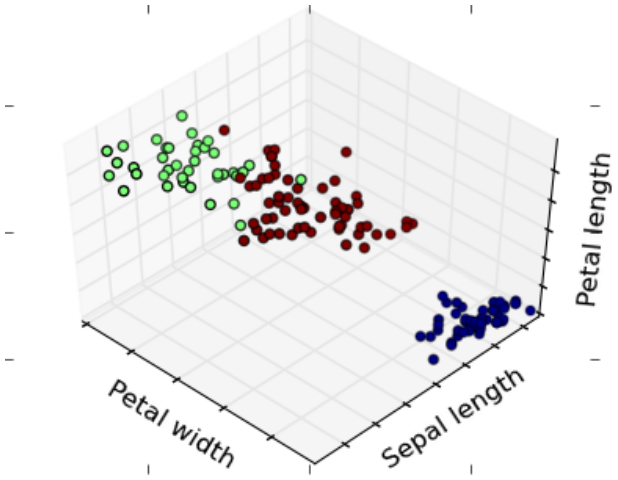


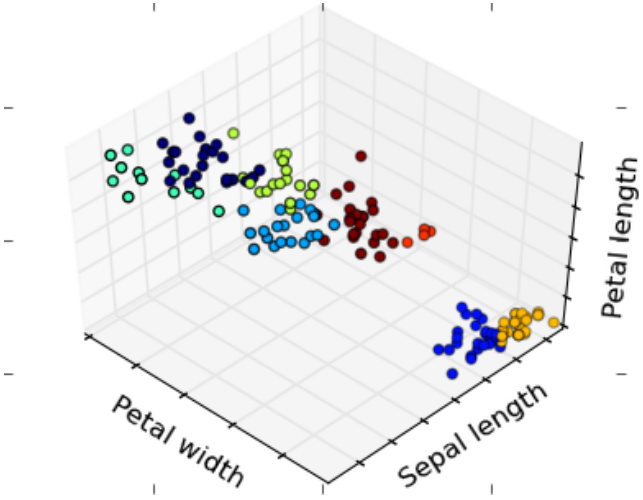


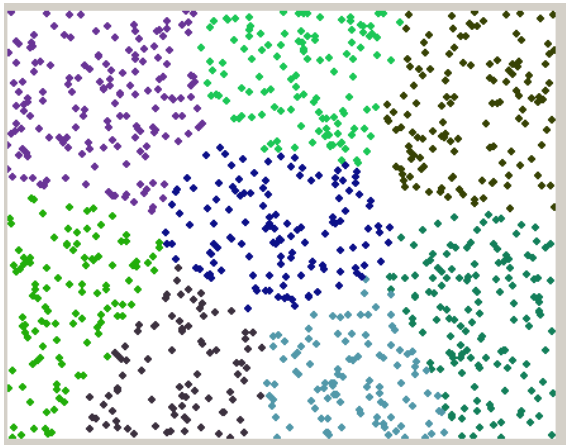
http://tulane.edu/sse/ccs/about/ccs_art_show_2011.cfm

K-Means

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

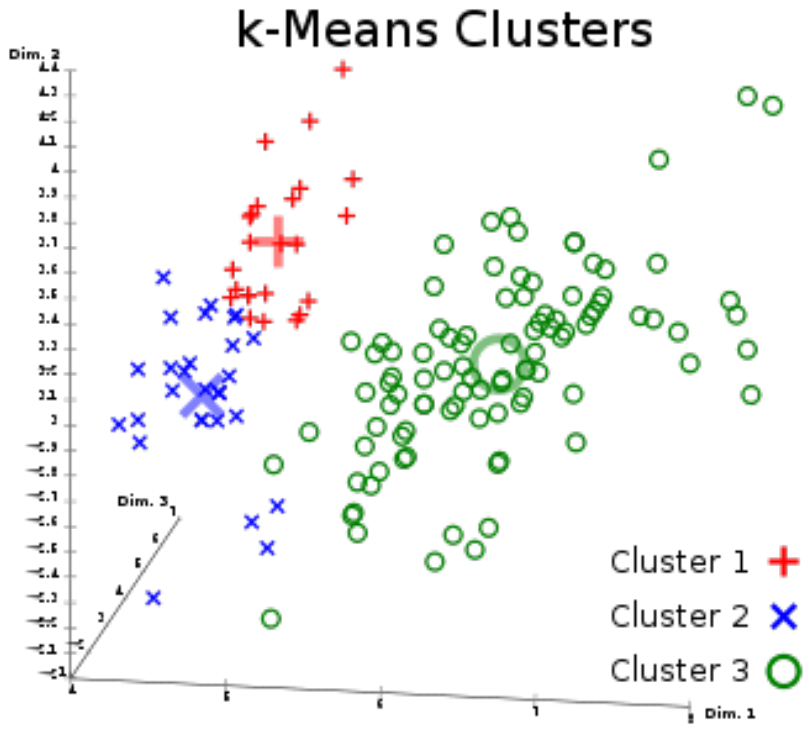


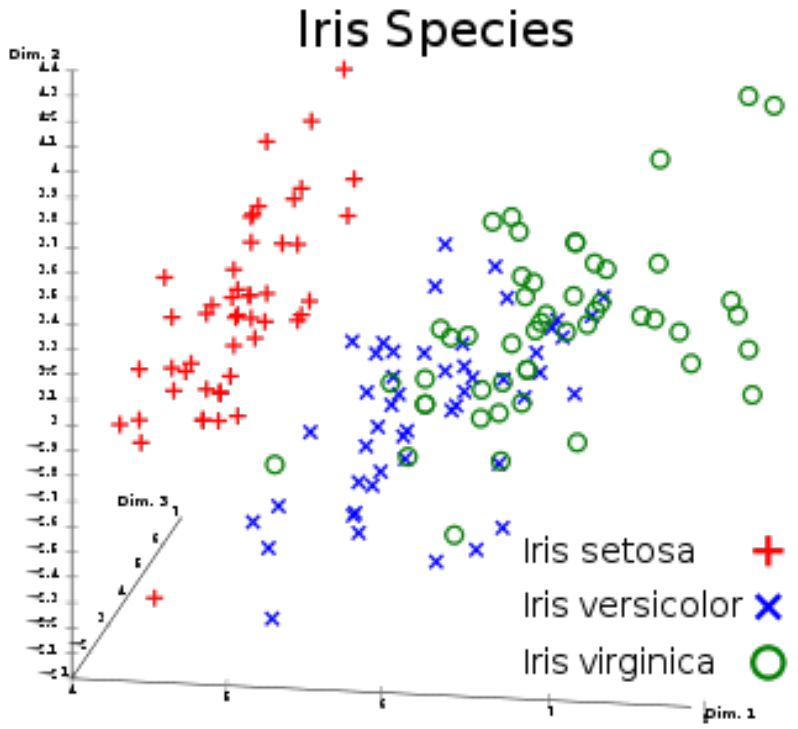




Linking Clustering to Outcomes

- A simple way is to link clusters to the outcomes is to first identify cluster membership and then use it in a regression or other type of model.
- Doesn't always work well. Outcomes could be very different, while clusters stay the same



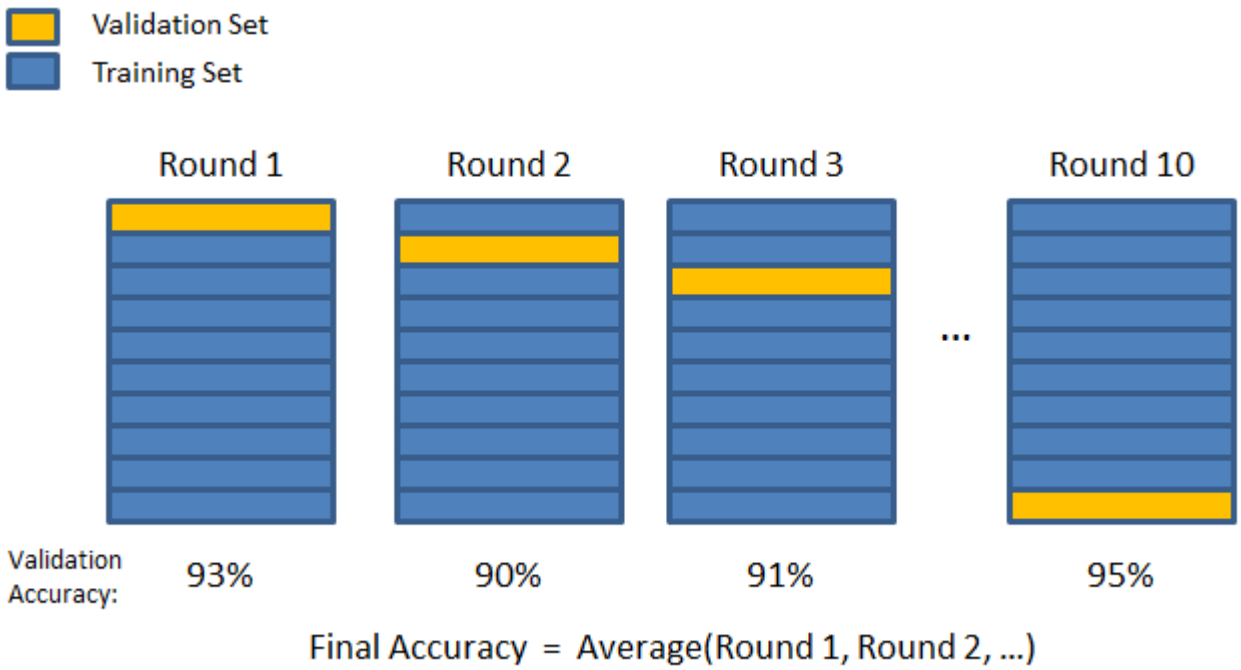


EM and K-means Clustering

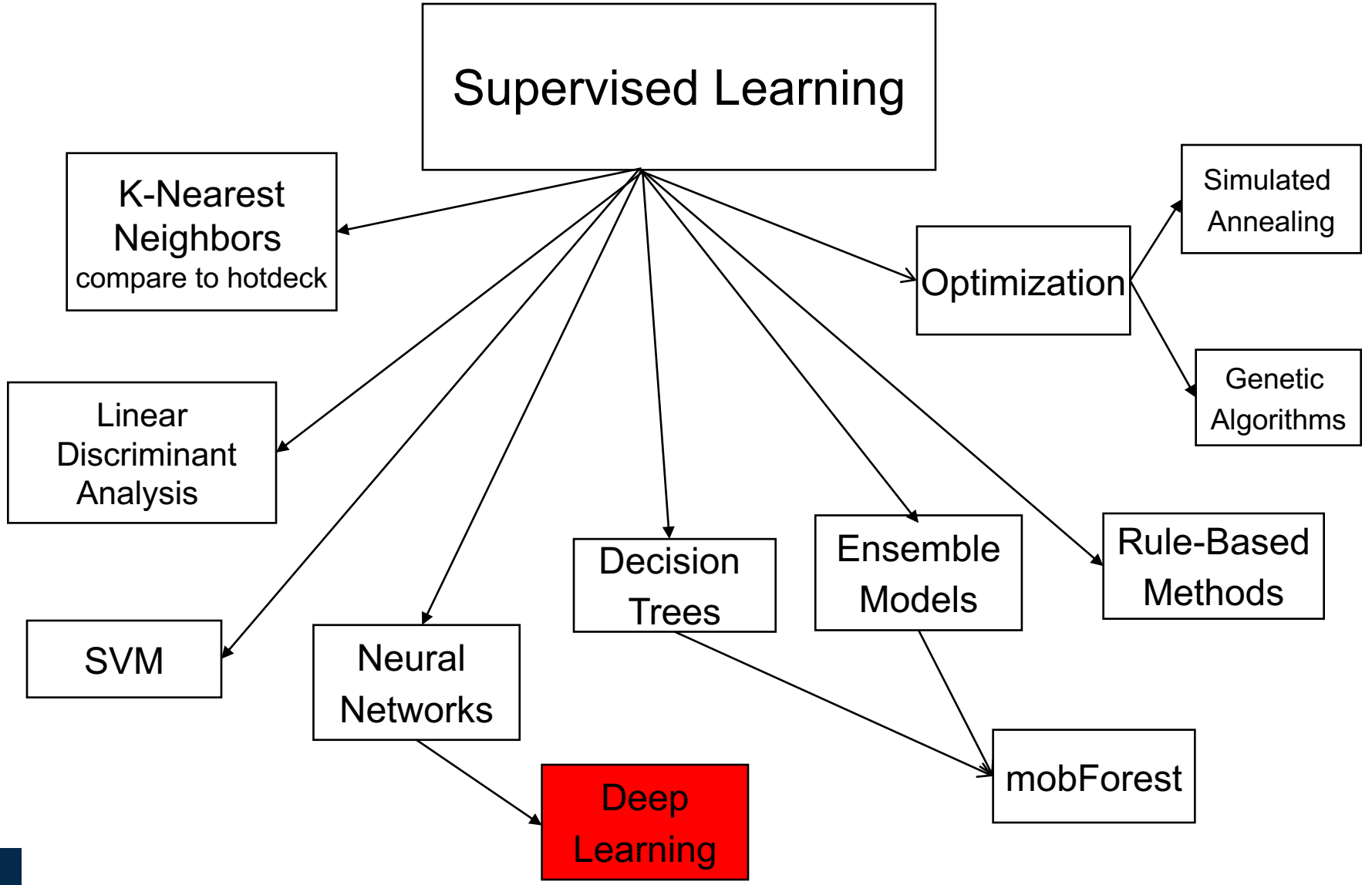
- K-Means
  - Hard clustering, an element belongs to only one cluster
  - Sensitive to outliers
- EM
  - Soft clustering, an element belongs to all clusters with different probabilities
  - Robust to outliers and can be used for numeric and nominal attributes
- EM Process
  1. Initiation. Use random distributions (in K-means random centers)
  2. M-step. For each point calculate probability being in each distribution (weight) (in K-means each point belonged to one of the centers)
  3. E-step. Based on the weights recalculate distribution parameters. (in K-means adjusted the centers)
  4. Continue alternating steps 2 and 3 until convergence

Training and Independent Test Datasets. At least cross-validation.

1. Training, Validation, Test Datasets
2. Training and Test Datasets
3. Cross-validation (e.g., k-fold)



Some Supervised Learning Methods



Continuous Outcomes

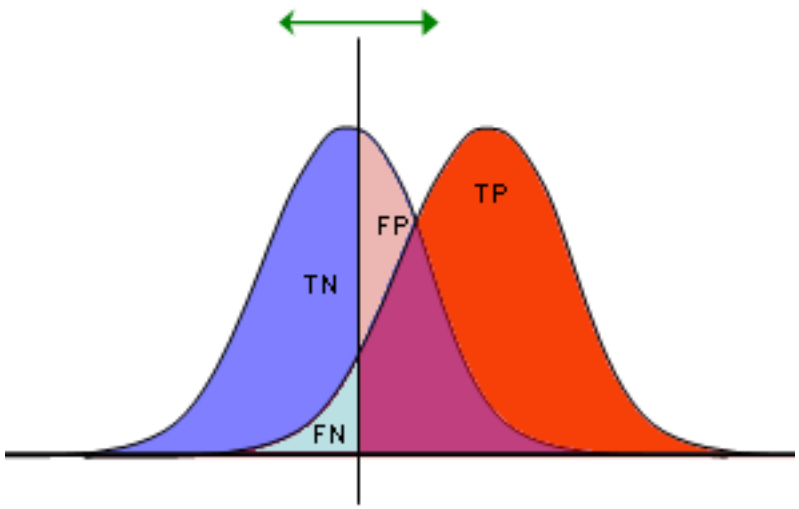
- Criterion: pseudo R^2

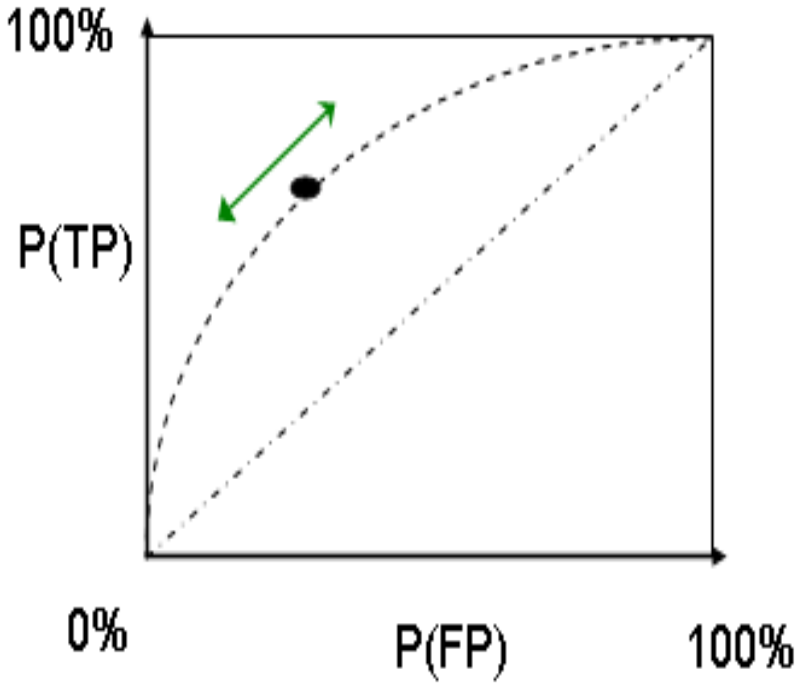
$$\text{Pseudo}R^2 = 1 - \frac{\text{mean square prediction error}}{\text{variance of the data}}$$

- Pseudo R^2 can be negative if the prediction performs worse than predicting the mean.

Prediction of Binary Outcomes

- **Receiver Operating Characteristic (ROC) curve**
  - **Area Under the Curve (AUC)**
  - **If AUC=0.5 then no association**



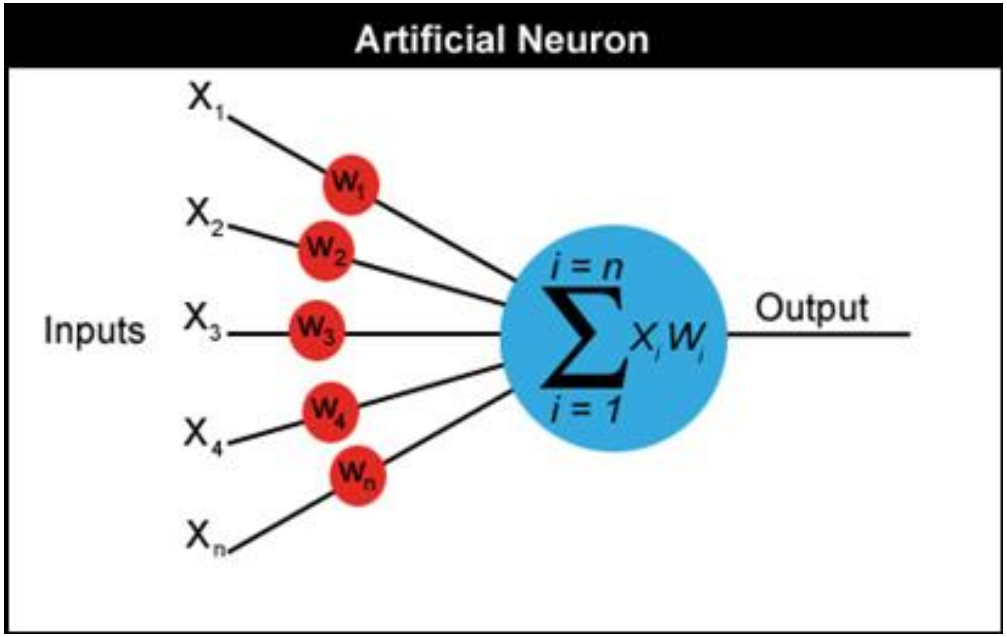


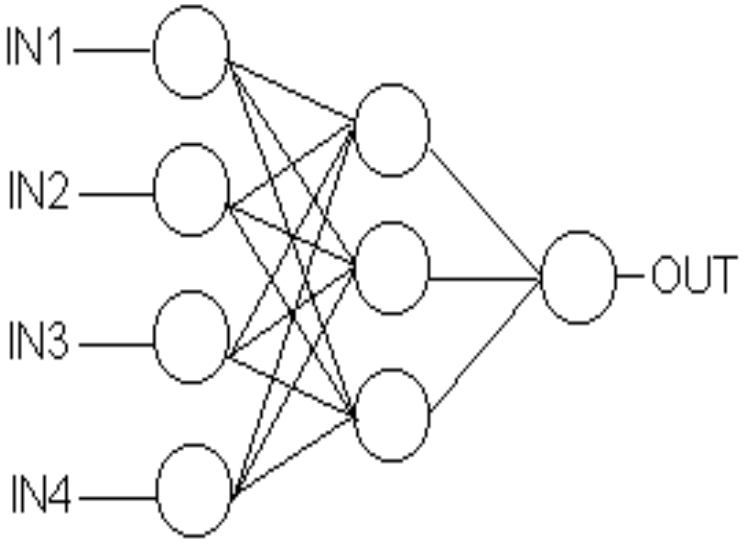
Neural Networks (ANN), Regression of regressions

Regression

$$Y = F(x) = w_0 + w_1 x_1 + \ldots + w_k x_k = \sum wx$$

$$Z = G(Y) = \beta_0 + \beta_1 Y_1 + \ldots + \beta_k Y_k = \sum \beta Y = \sum \beta F(x, w)$$

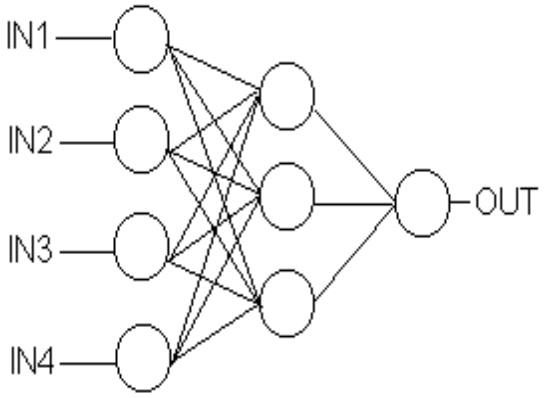




Neural Networks (ANN) (continued)

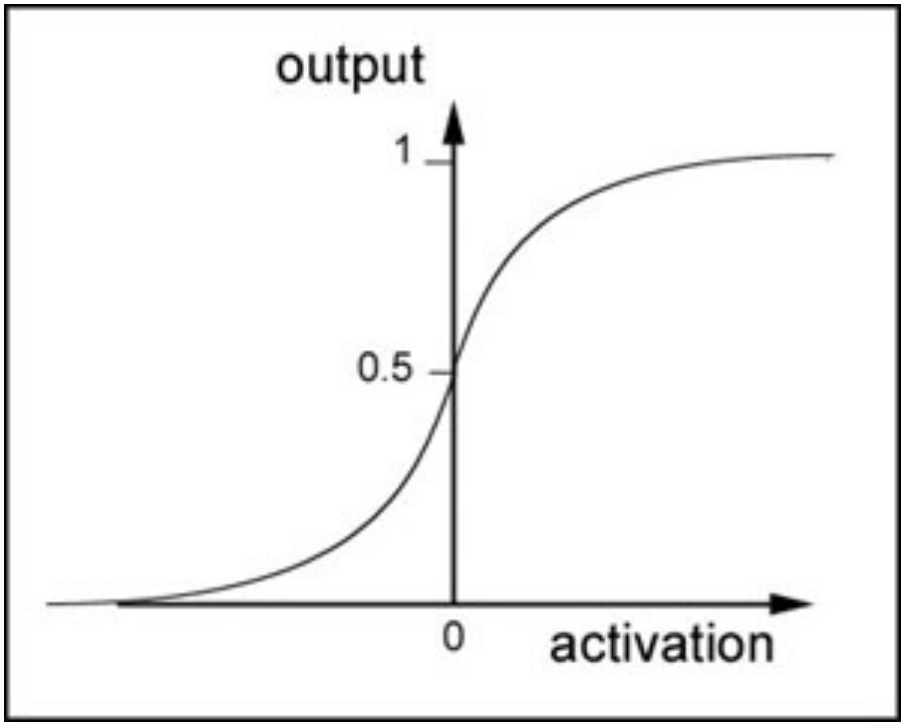
Logistic regression

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i}$$



$$P = \frac{1}{1 + e^{\beta_0 + \beta_1 Y_1 + \ldots + \beta_k Y_k}}$$

$$Y = \frac{1}{1 + e^{w_0 + w_1 x_1 + \ldots + w_k x_k}}$$



Neural Networks (ANN) (continued)

$$Z(x) = F(x, w)$$

- If F is smooth, then can calculate $\dfrac{\partial (Z - \hat{Z})^2}{\partial w}$

- Network training with error back propagation
- Find an initial set of weights *w.* Gradient descent
- Calculate the new value of $\hat{Z}$
- Evaluate the gradients
- Calculate a new set of weights *w*
- Repeat until converge

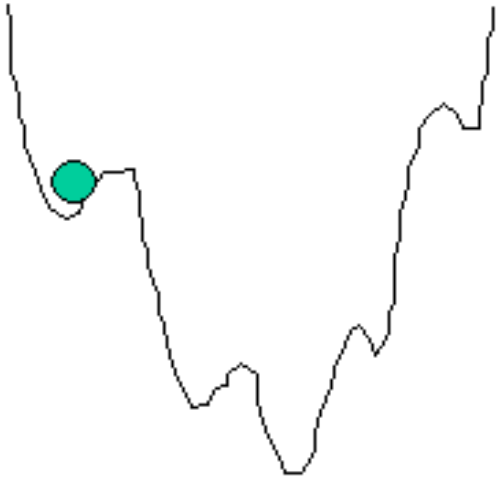
$$\frac{\partial (Z - \hat{Z})^2}{\partial w}$$

Neural Networks (ANN) (continued)

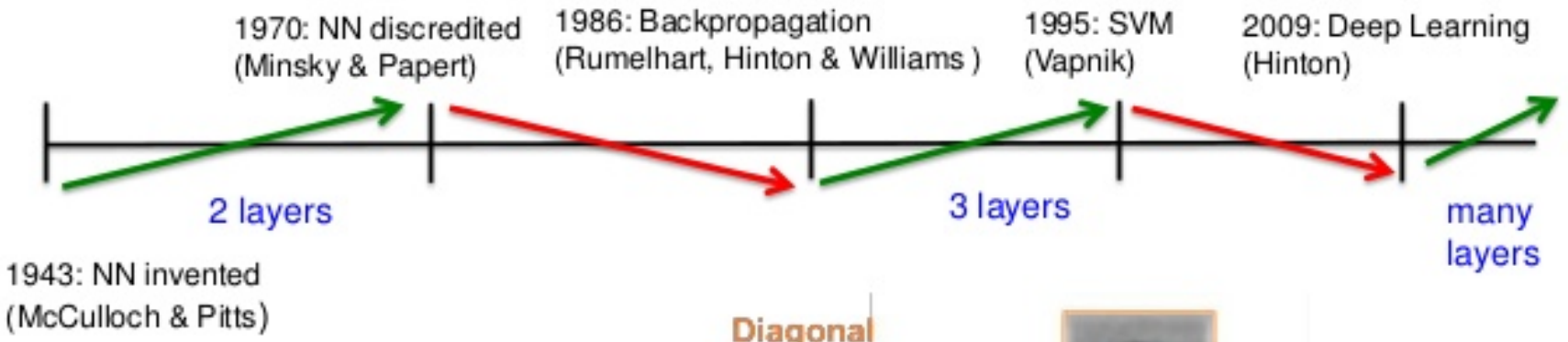
- The idea of “Energy.”
- Local minima

$$\frac{\partial (Z - \hat{Z})^2}{\partial w} = 0$$

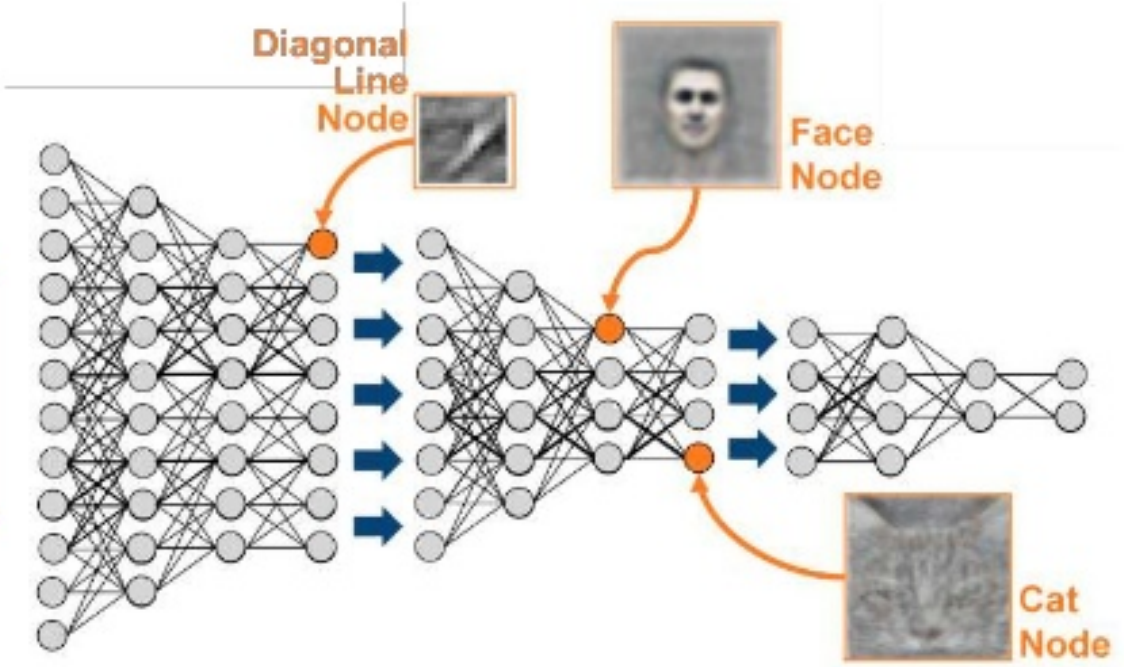
- Need to “Shake-up” the network to find the global minimum
- Simulated annealing (?)
- Closer to the global min, less shaking



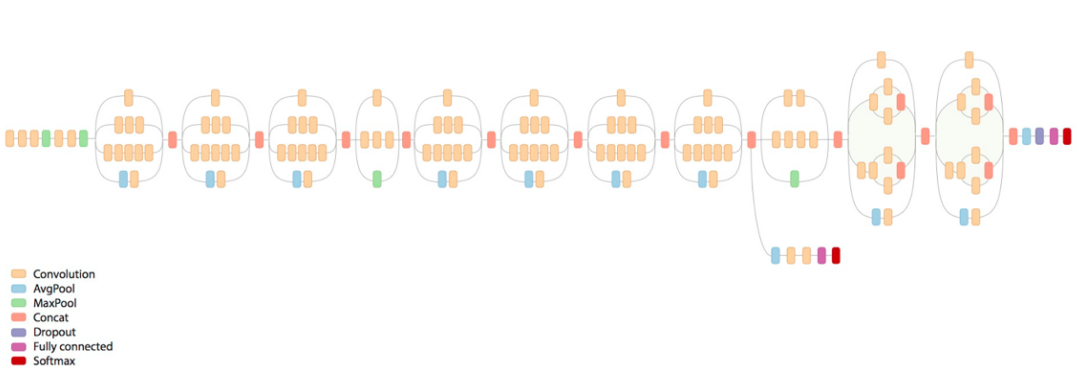
Shallow and Deep Learning



-Model Size: 10B parameters
-Used by: Yahoo!, Google, Microsoft, Baidu, IBM, Scyfer ☺



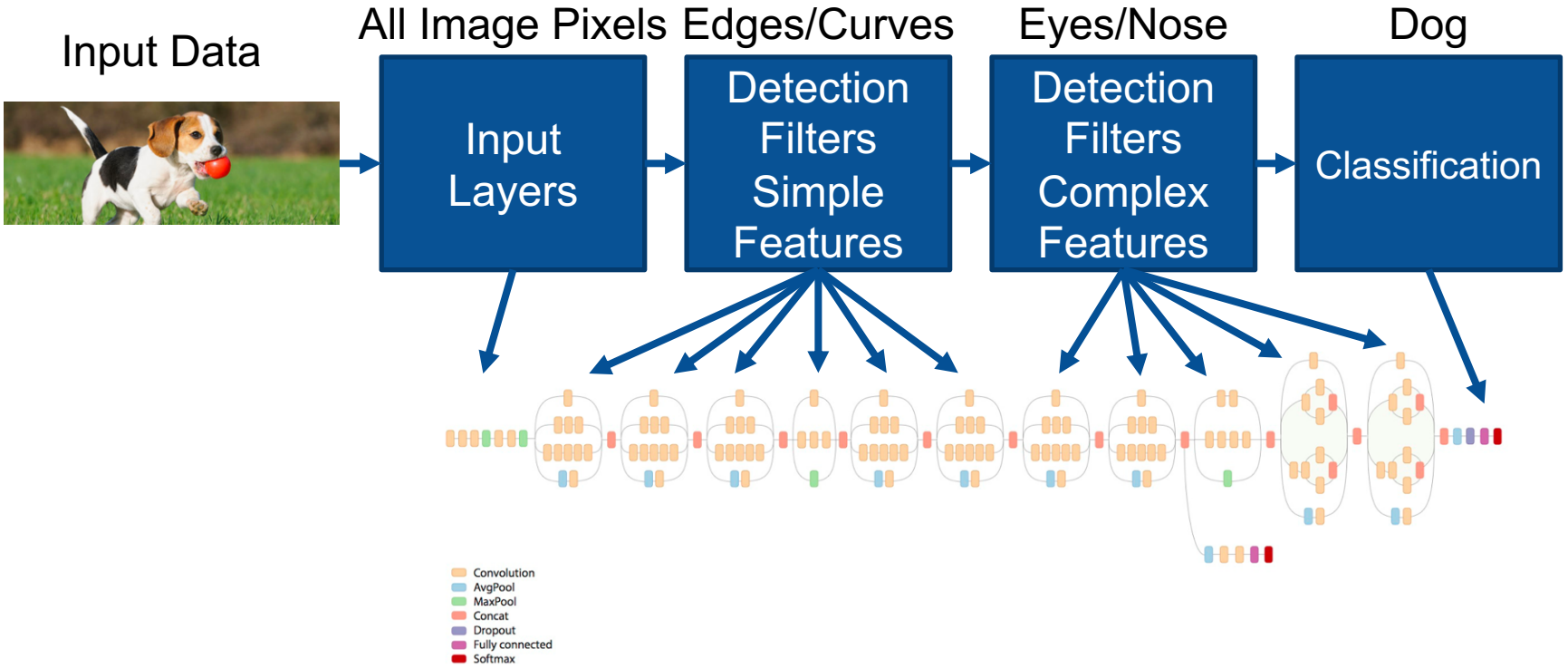
Object Recognition Using Neural Networks (Deep Learning)



DOG

CAT

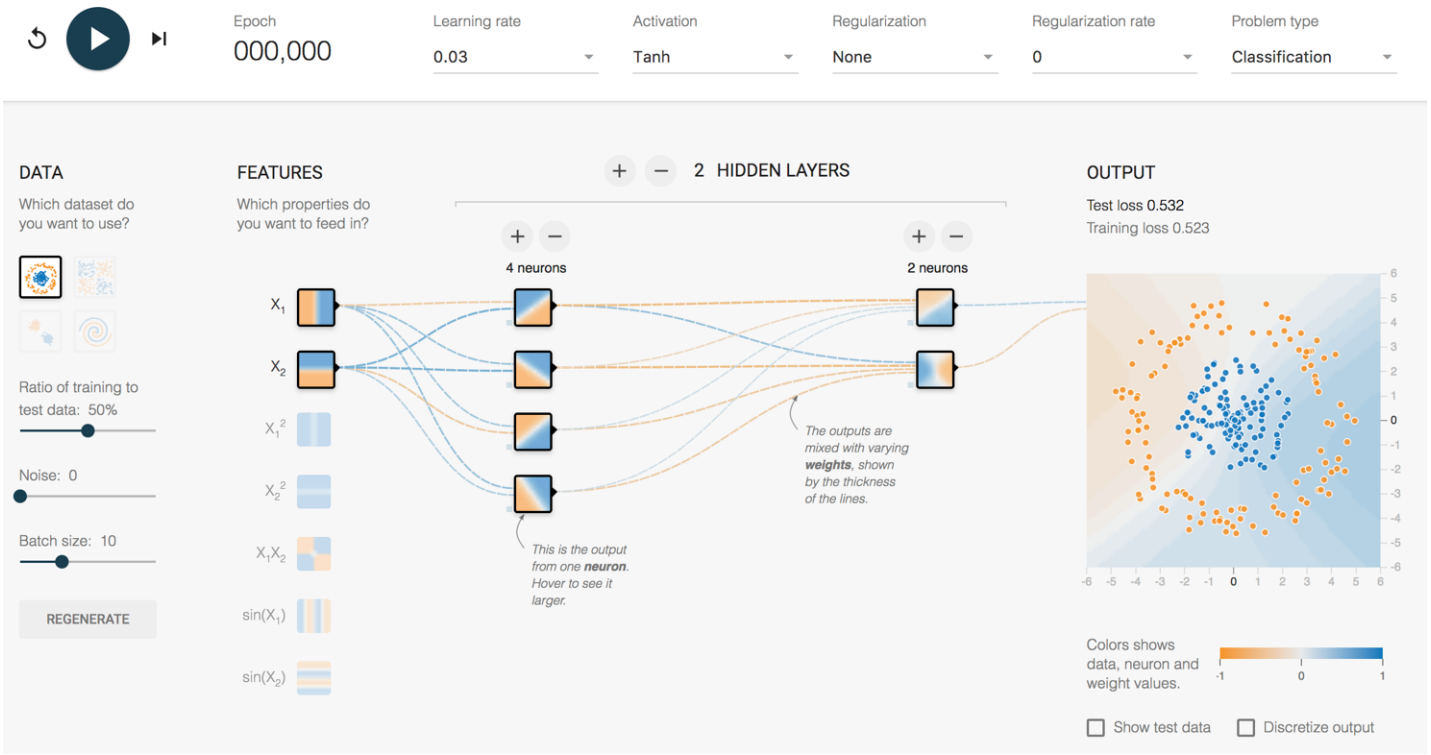
CNN Diagram for Image Analysis



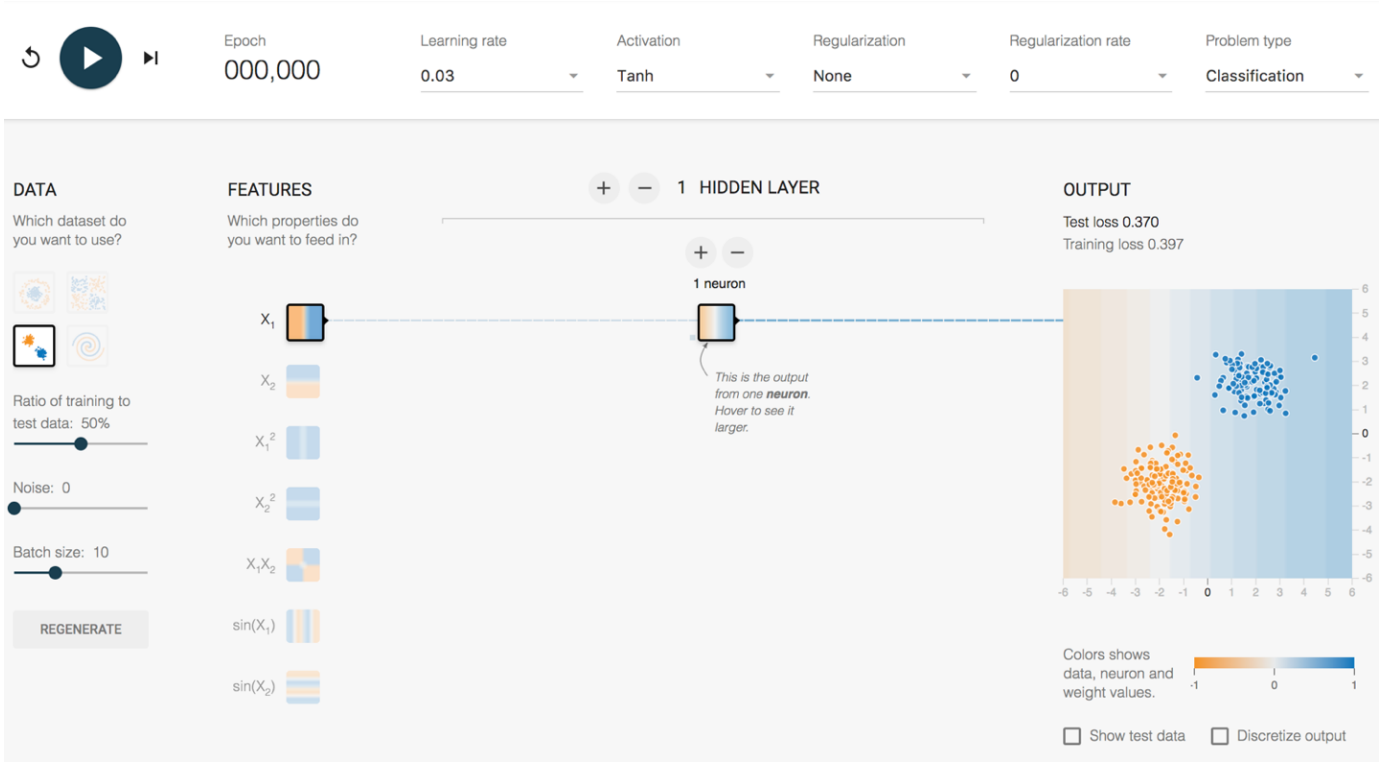
Inception v3
(Google CNN Architecture)

How does it work?

http://playground.tensorflow.org/



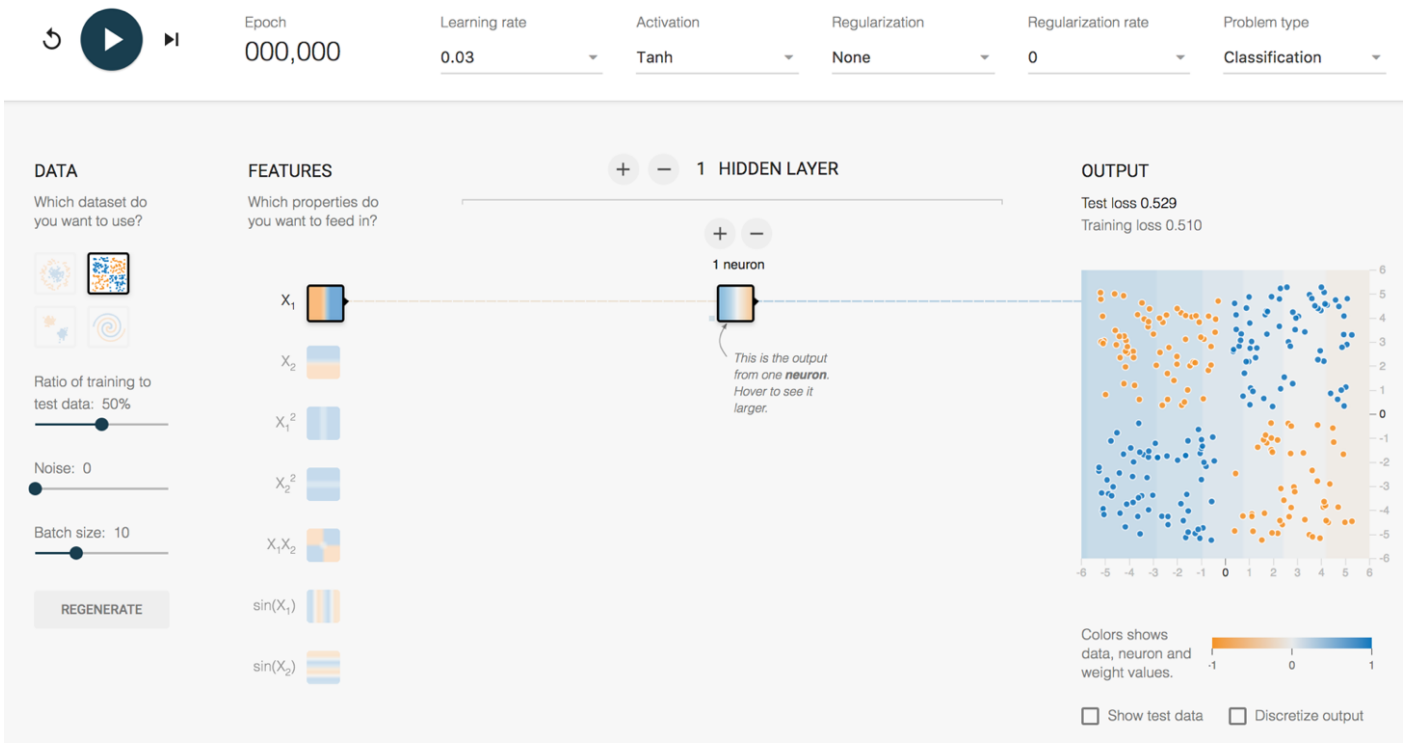
Playground Demo 1



2 Clumps 1 Feature 1 Hidden Layer

1 Neuron

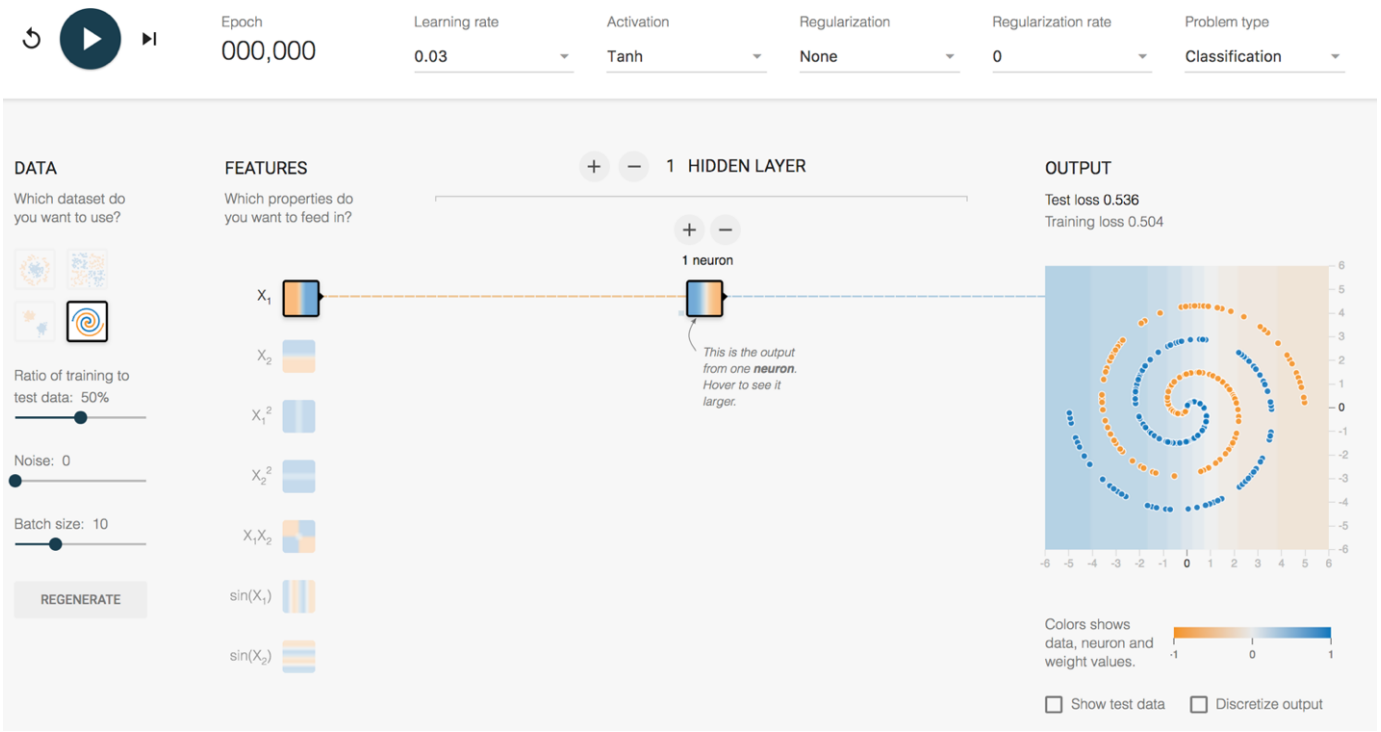
Playground Demo 2



4 Clumps 1 Feature 1 Hidden Layer

1 Neuron

Playground Demo 3

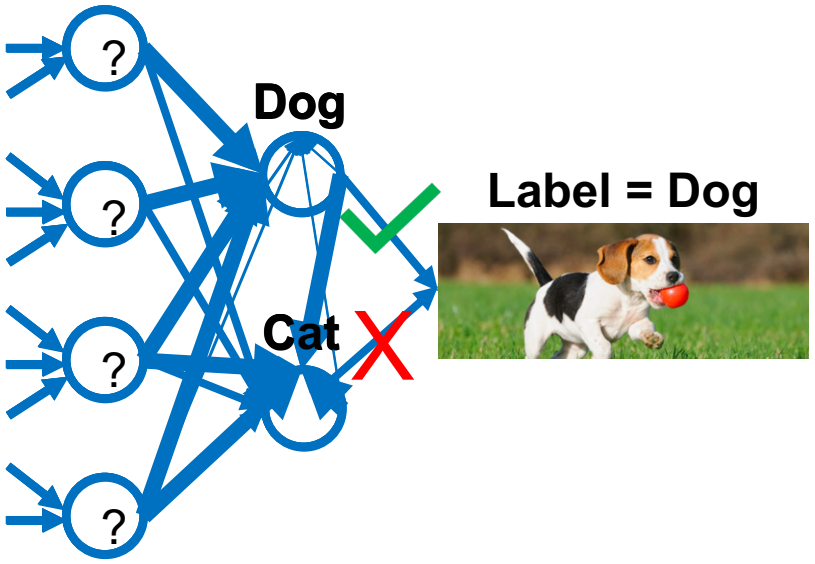


Spiral 1 Feature 1 Hidden Layer

1 Neuron

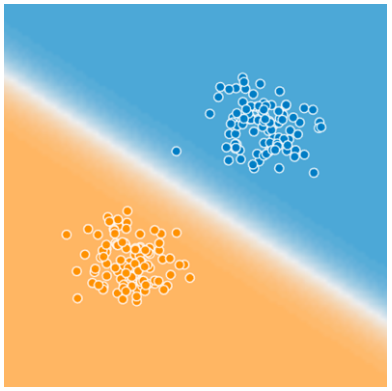
How does a CNN “Learn”?

- Same old backpropagation
- Initial network settings are random
- Test it using training data
- Adjust network by a certain amount (Learning Rate) to reduce the error
- Start at the end of the network and work backwards



Is that it?

Blue

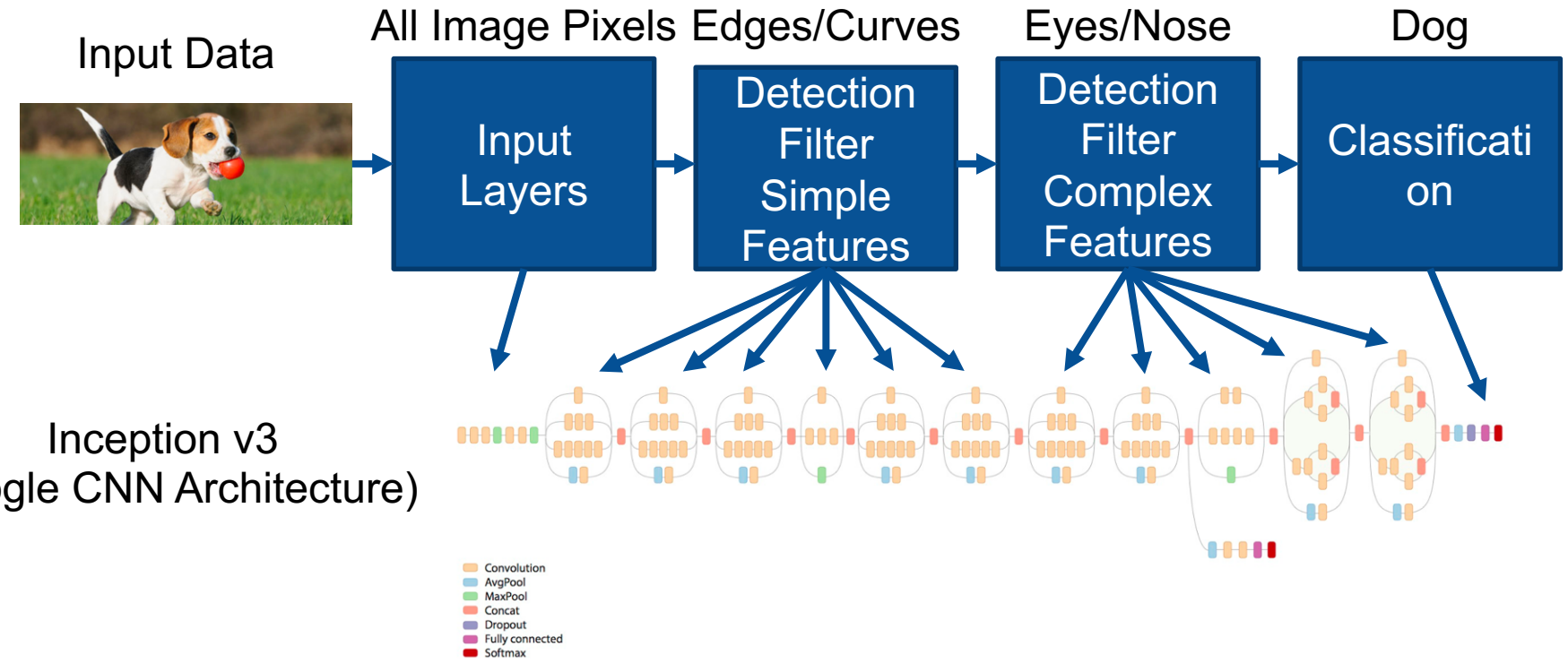


Orange

How to get to here?

Cat Typing on Laptop

How complex are these things?

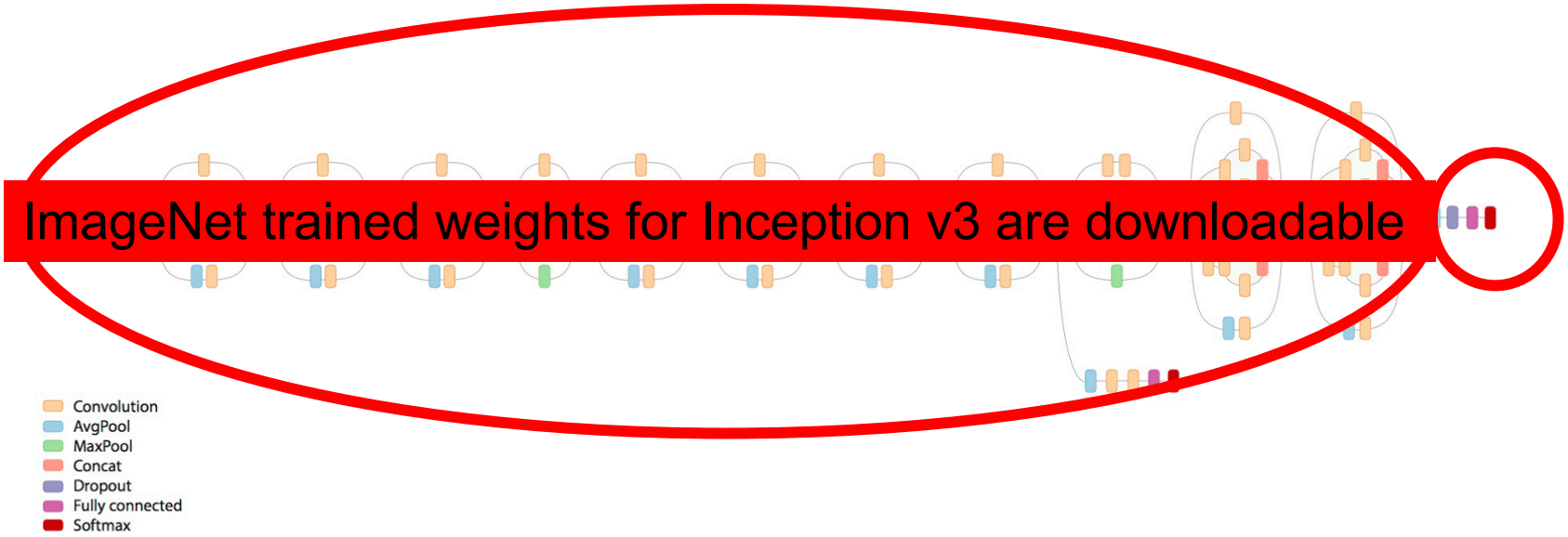


Inception v3
Google CNN Architecture)

Number of Parameters = 23,200,000

Transfer Learning is a Key Advance

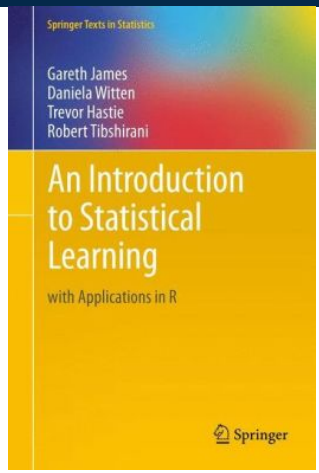
- Retraining the last layer is all that is needed for any human recognizable object



References

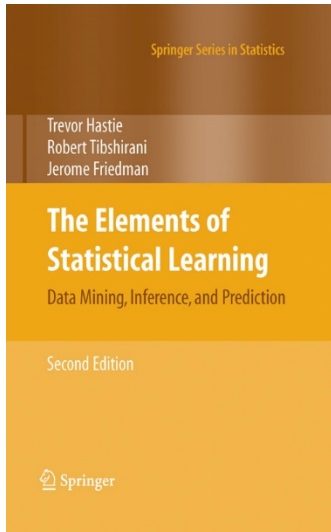
Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Second Ed.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.



- Available online FREE!
  - http://www-stat.stanford.edu/~tibs/ElemStatLearn/
  - http://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/

- Numerous ***public*** sources on Google
  1. R libraries. http://cran.r-project.org/web/views/MachineLearning.html
  2. Python libraries. www.*scikit-learn.org/*



Resources to Learn More

Entry Level Videos:

- Josh Gordon (Google) Machine Learning Tutorials https://www.youtube.com/watch?v=cKxRvEZd3Mw

- Siraj Raval - Various Artificial Intelligence Videos https://www.youtube.com/channel/UCWN3xxRkmTPmbKwht9FuE5A

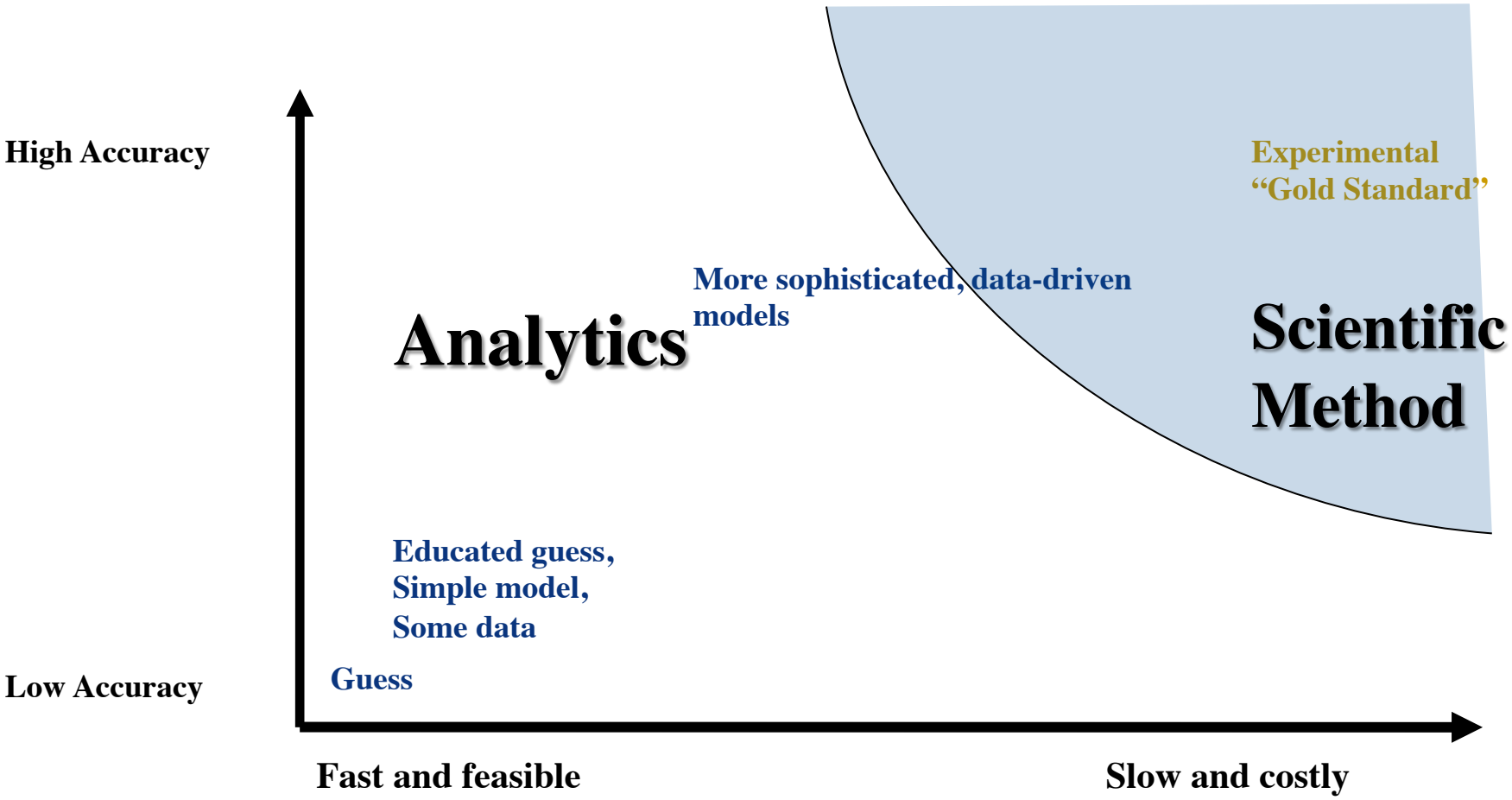
More Advanced

- Adrian Rosebrock (Lots of Python and OpenCV Materials) https://www.pyimagesearch.com/

Thanks you

- Questions?
- Contact info: bobashev@rti.org

Compromise of Precision and Speed



Philosophical Concept

Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.

John Tukey. The future of data analysis. *Annals of Mathematical Statistics*, 33(1), 1962, p. 13.