Whirlwind Tour of Machine Learning (Pattern Recognition)

- Well established academic fields
- Participation from statisticians, computer scientists, mathematicians and engineers
- Many useful, tried-and-true techniques
- Resurgence
  - SVM, boosting
  - Deep learning

Common Statements

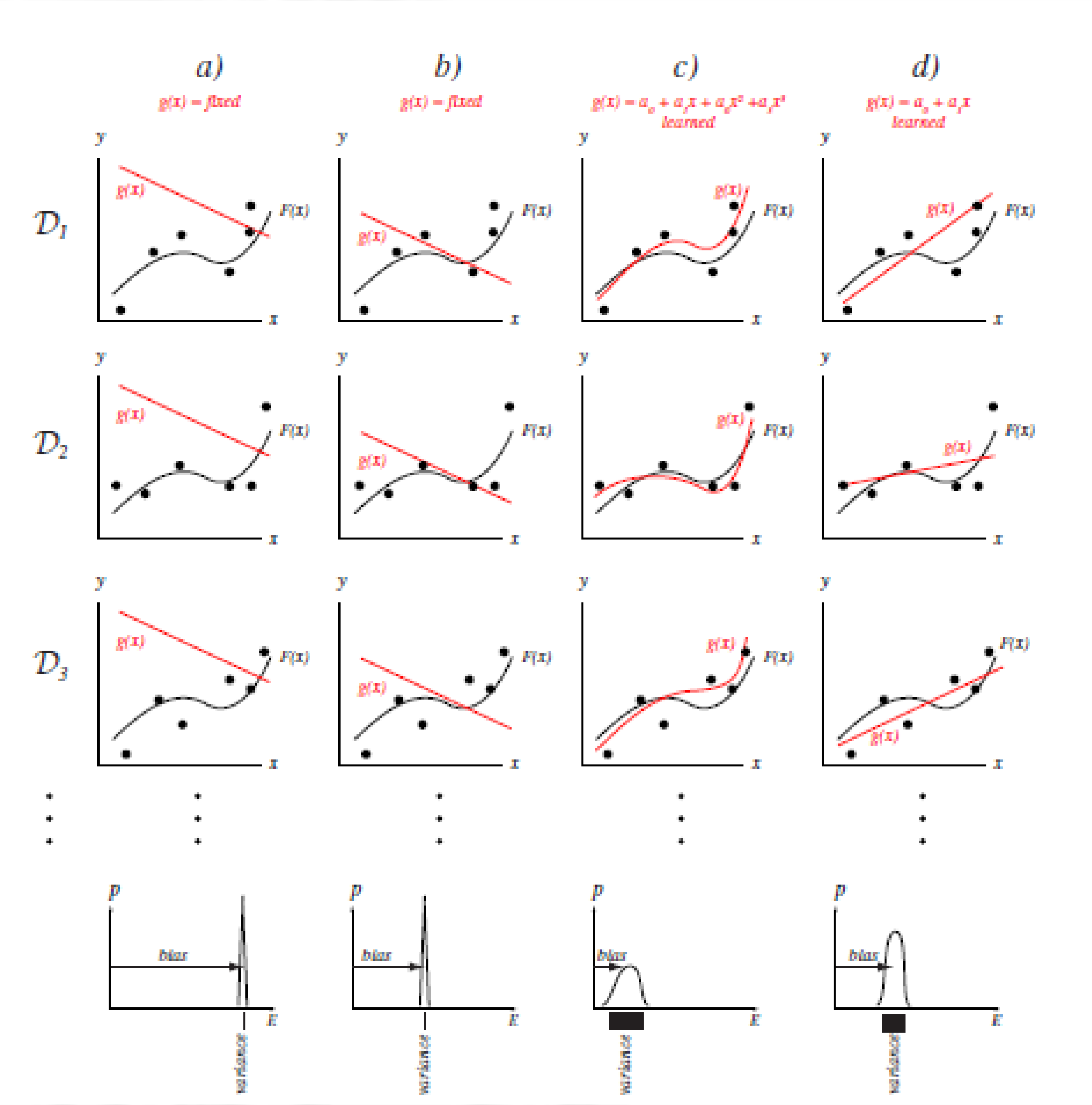
- Objects characterized by traits (features)
  - You credit score, income, debt in loan application
  - Your age, look, hobby in dating services
  - Website usage, traffic, geographical loc in intrusion detection
  - SIFT, wavelet features in image analysis
  - Finger print patterns (ridges, bifurcation, terminations, etc.) in matching
  - Location, school district, # of rooms, square footage in housing search
  - Color, options, engine size, # of doors in auto market

General Approach

- Supervised f(x) -> y
  - Learn with a teacher model
  - Classification (discrete) & regression (continuous)
  - Training + validation, testing, and deployment
- Unsupervised f(x) only
  - Learning without a teacher model
  - Abnormality finding, data mining, etc.
- RL
  - Sequence (e.g., temporal) data
  - No or sparse training data and feedback
  - Maximize reward functions

Supervised Regression

- Curve, surface fitting
- Right order



Universal Dilemma

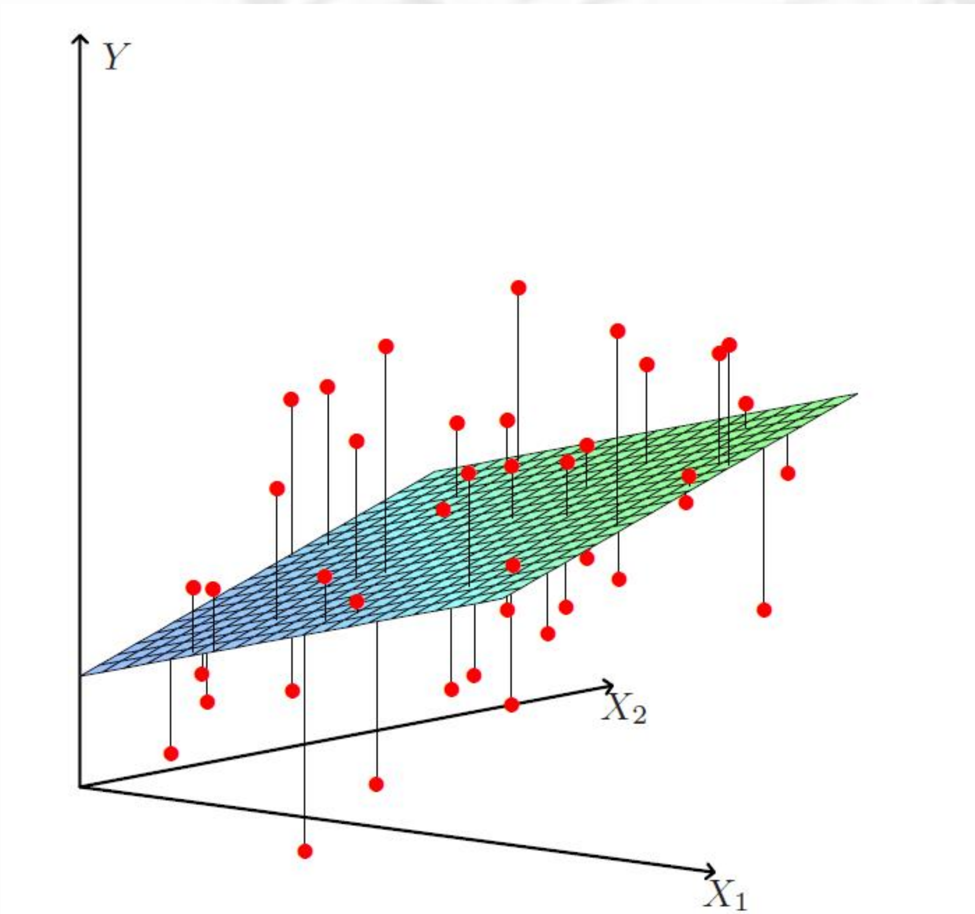
- Bias
  - Error
  - How fitting confirms to data
- Variance
  - How fitting confirms to each other

Universal Dilemma

- Small bias and large variance
  - Way too complicated models
  - Overfitting
- Large bias and small variance
  - Naïve model
  - Insufficient training data
- Under general assumptions there is a subtle trade off – you cannot optimize both!

Supervised classification

- Separation of different class samples



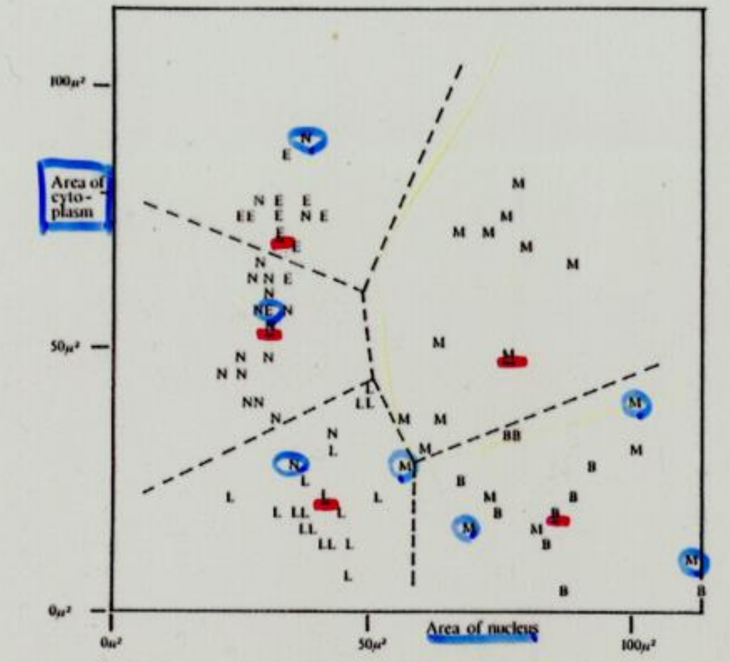


Figure 14-1. Scattergram of cytoplasm area versus nuclear area for five different common types of white blood cells. The letters denote the different classes, with the centroids underlined. The dashed lines show linear boundaries that best separate the classes. Several samples are misclassified. (Plotted from data in "Automated Leukocyte Recognition" by I.T. Young, Ph.D. thesis, MIT, Cambridge, Massachusetts, 1969.)

Various Types of Errors

$$\text{Precision} = \frac{tp}{tp + fp}$$

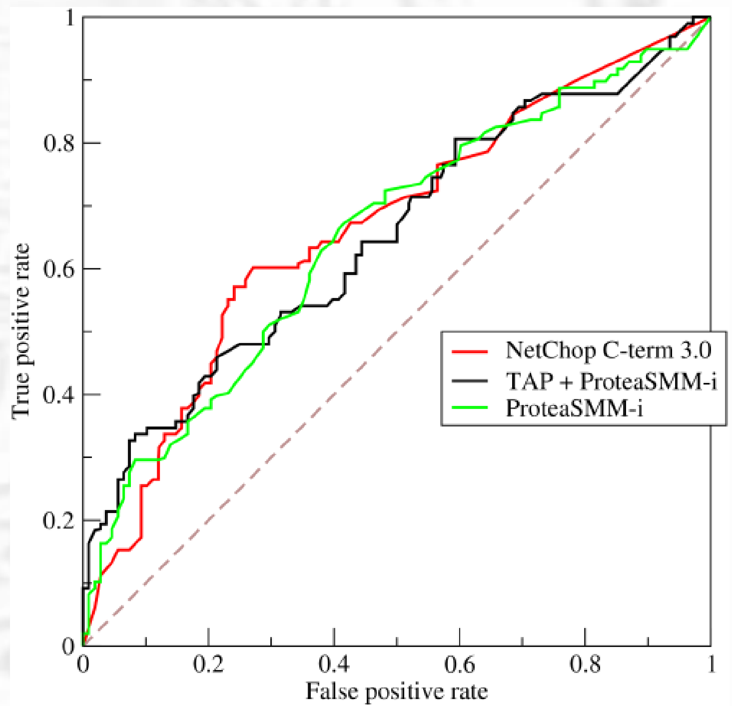
$$\text{Recall} = \frac{tp}{tp + fn}$$

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy

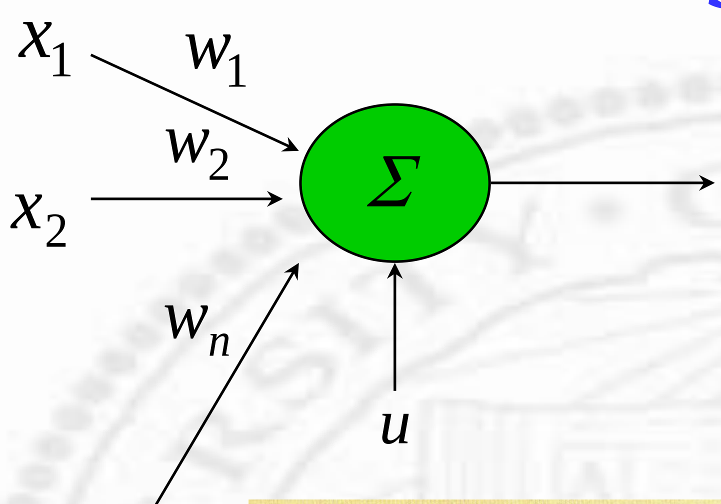
Precision vs. Recall

- A very common measure used in PR and MI community
- One goes up and the other HAS to go down
- A range of options (Receiver operating characteristic curves)
- Area under the curve

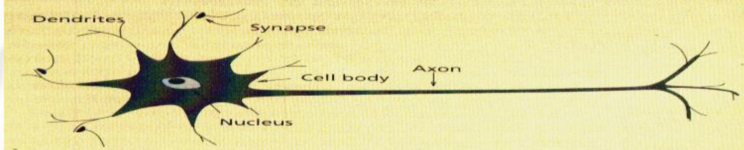
as a goodness measure



Simple Tool



x_1
w_1
w_2
x_2
Σ
w_n
u
x_n



Dendrites
Synapse
Cell body
Axon
Nucleus

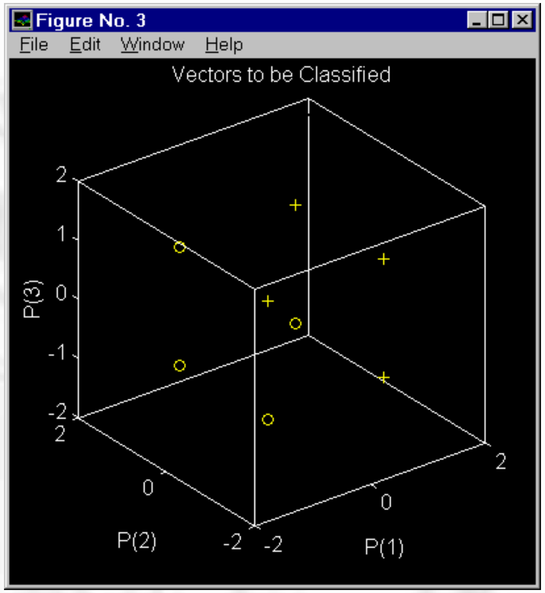
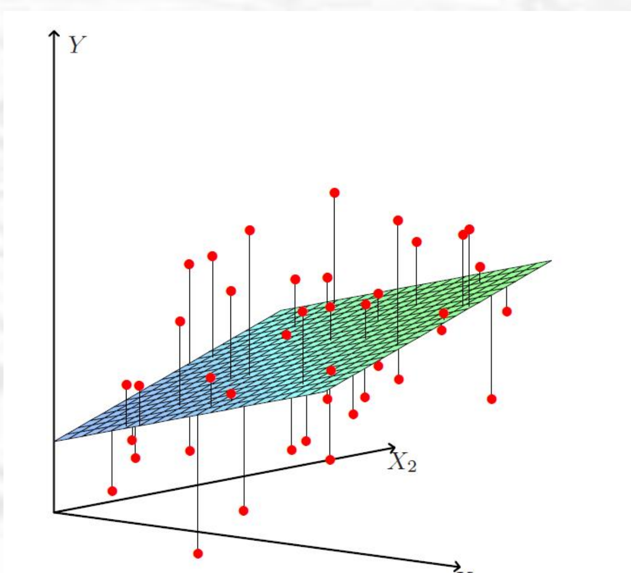


Figure No. 3
File Edit Window Help
Vectors to be Classified
P(3)
P(2)
P(1)



Y
X_2

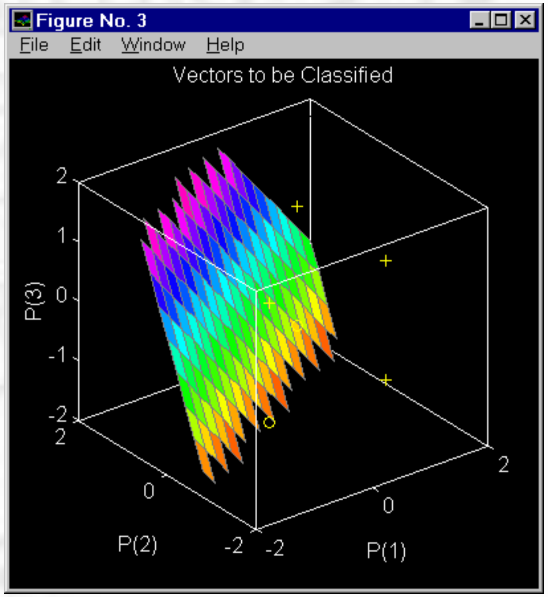
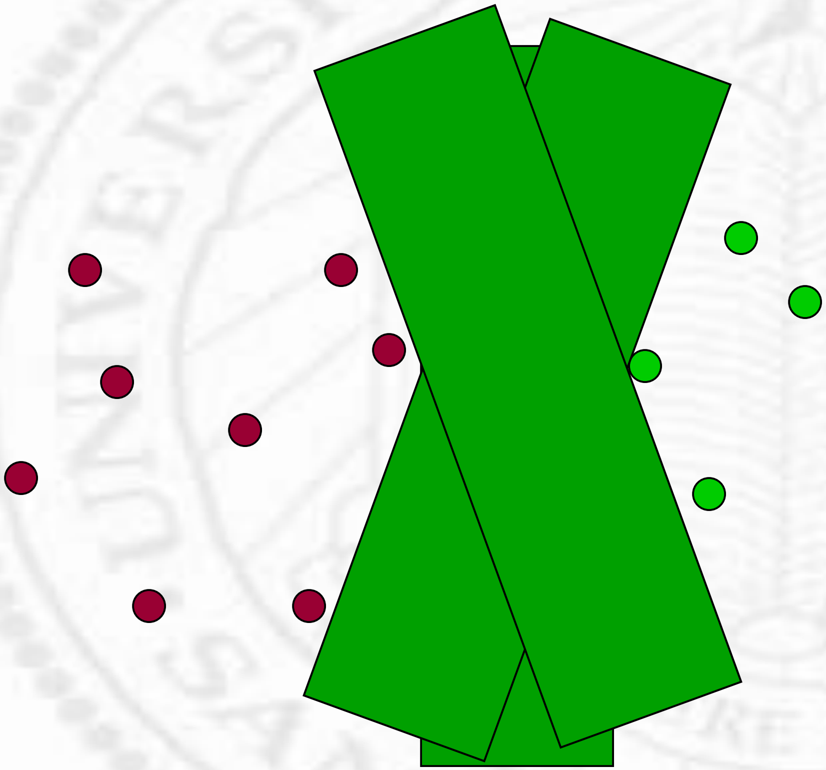
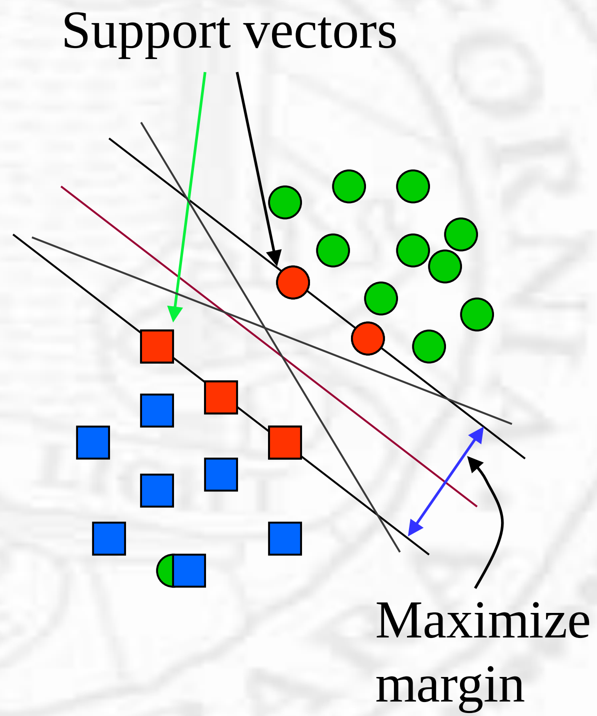


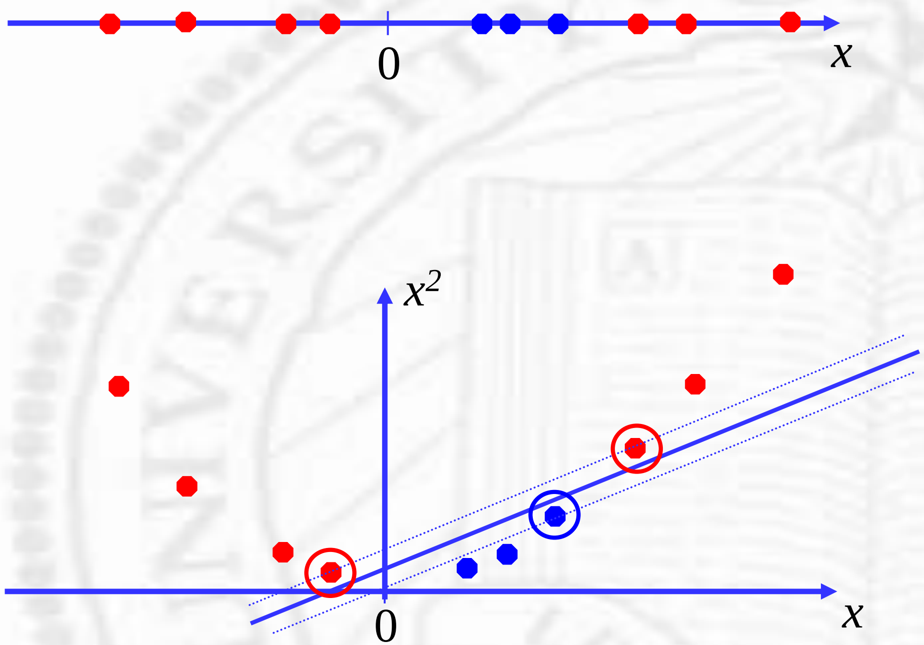
Figure No. 3
File Edit Window Help
Vectors to be Classified
P(3)
P(2)
P(1)

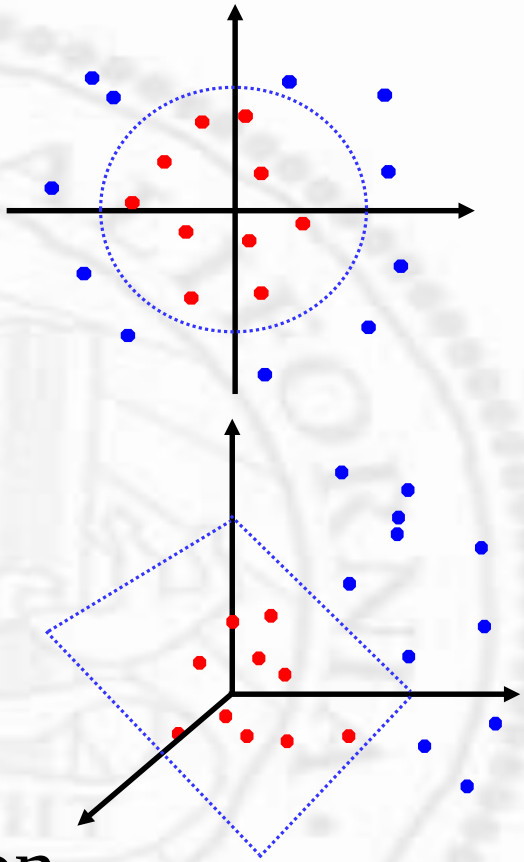
More Advanced Math Tools





“Massage” data

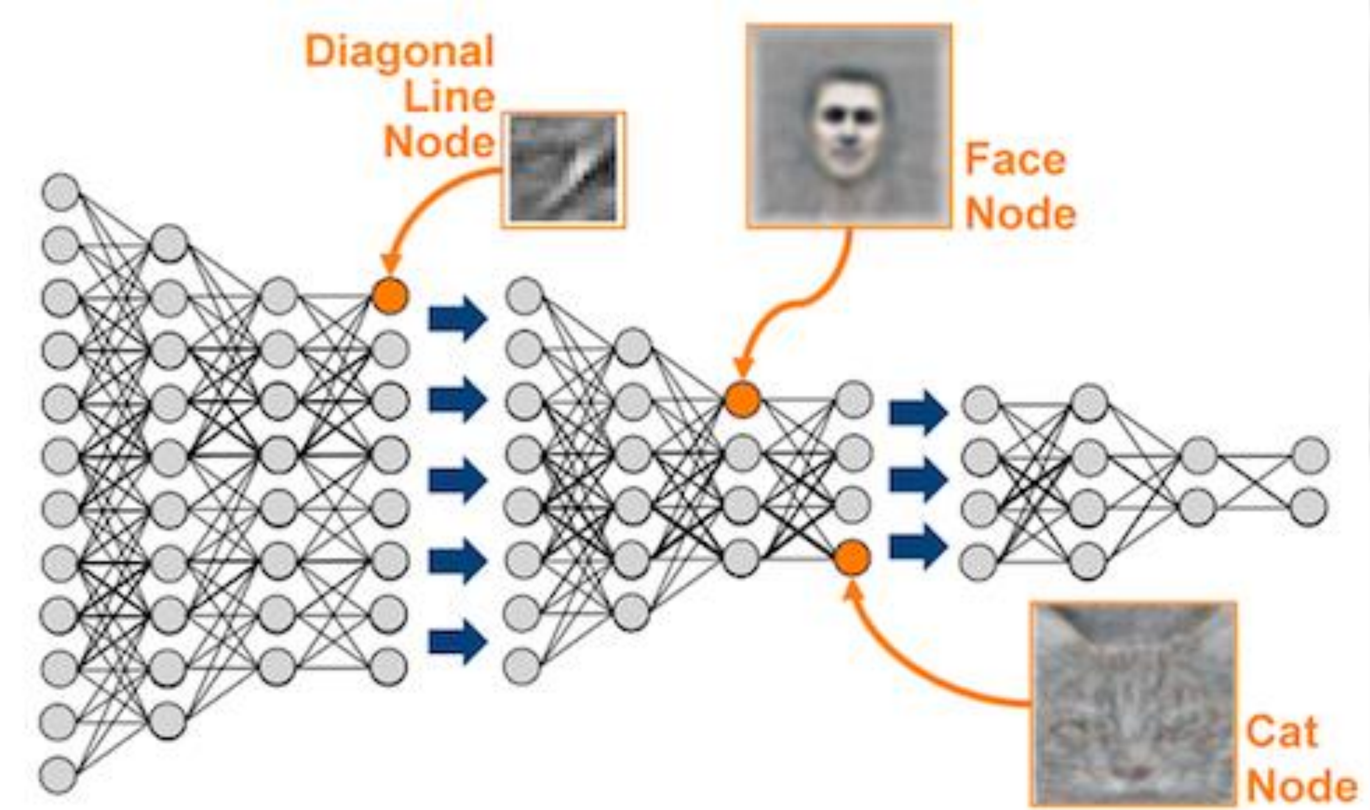




- Nonlinear data transformation
- Combined SVM + Data Massage (Kernel methods) is the most popular and robust methods (up to 5 years ago)

"Massage" Classifiers

- "Massage" classifiers
- Use more than one
- Use a hierarchy



General Training Procedures

- Data collection, shuffling, batching
- Repeat
  - Training with known f(x)-> y
  - Validation with non-overlapping f(x)-> y
  - Error drop, keep training results
  - Error not drop, discard training results
- Final accuracy figures (on non-overlapping testing data set)

Unsupervised Clustering