MACHINE LEARNING

BRETT WUJEK, SAS INSTITUTE INC.

AGENDA

MACHINE LEARNING

- Background
- Use cases in healthcare, insurance, retail and banking
- Examples:
  - Unsupervised Learning –Principle Component Analysis
  - Supervised Learning – Support Vector Machines
  - Semi-supervised Learning – Deep Learning
  - Supervised Learning – Ensemble Models
- Resources

MACHINE LEARNING BACKGROUND

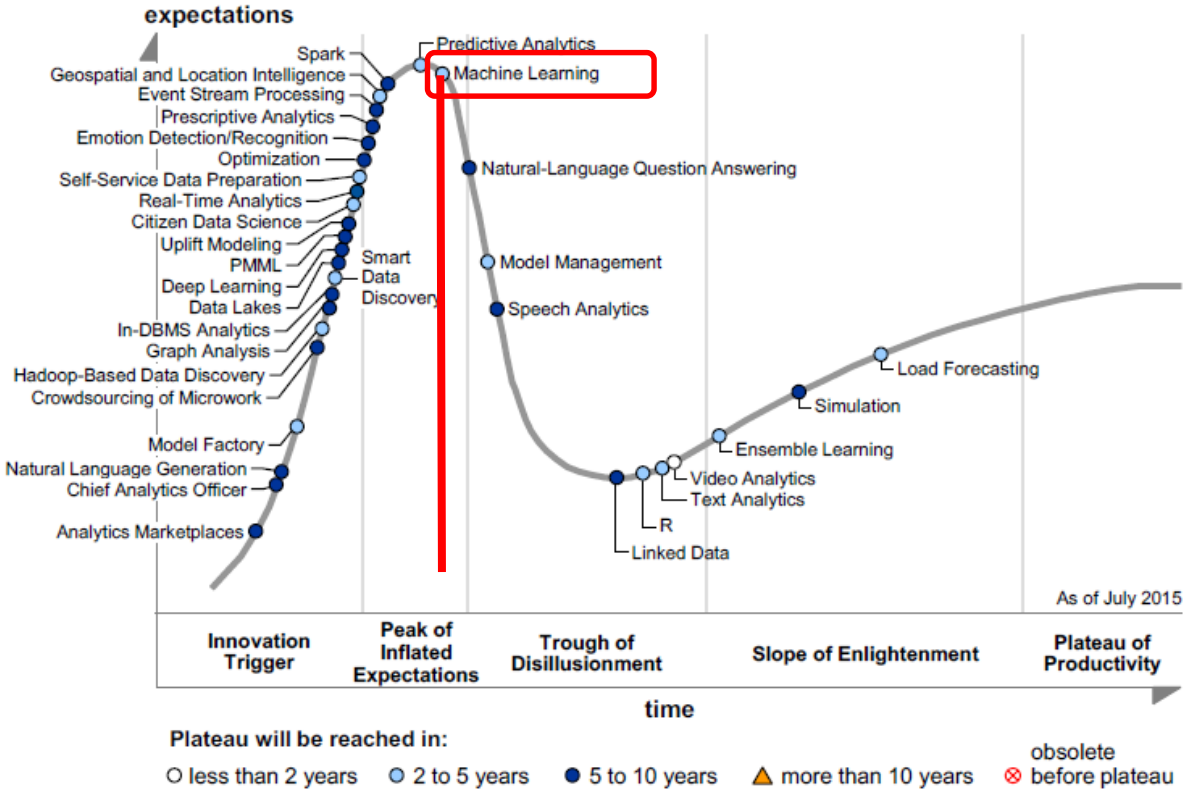
WHAT IS MACHINE LEARNING?

Wikipedia:
*"Machine learning is a scientific discipline that deals with the construction and study of algorithms that can **learn from data**. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, **rather than following only explicitly programmed instructions**."*

SAS:
*Machine learning is a branch of artificial intelligence that **automates** the building of systems that learn from data, identify patterns, and make decisions – with **minimal human intervention**.*

MACHINE LEARNING

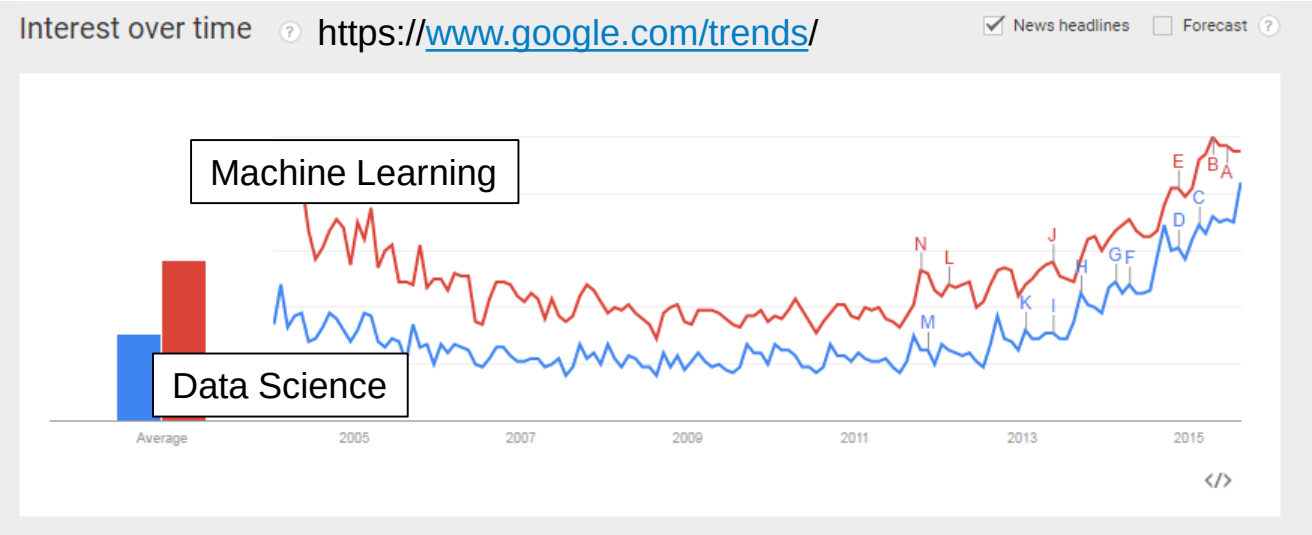
GARTNER HYPE CYCLE (ADVANCED ANALYTICS)



Source: Gartner (July 2015)

MACHINE LEARNING | WHY NOW?

- Powerful computing resources have become available
- Data is a commodity: more and different types of data have become available
- Rise of the “data scientist”



WHY WOULD YOU USE MACHINE LEARNING?

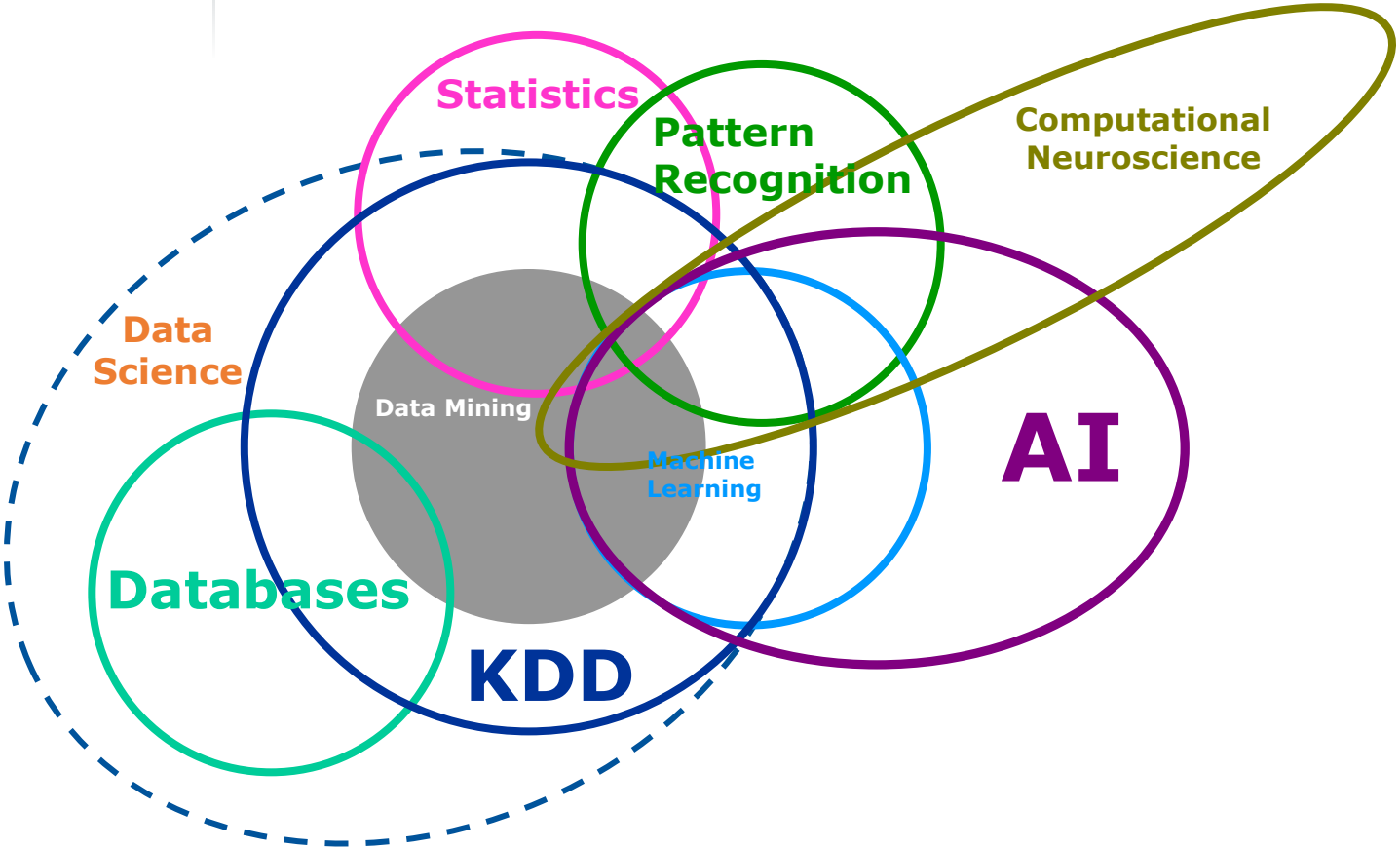
- Machine learning is often used in situations where the **predictive accuracy of a model is more important than the interpretability of a model**.

- Common applications of machine learning include:
  - Pattern recognition
  - Anomaly detection
  - Medical diagnosis
  - Document classification

- Machine learning shares many approaches with statistical modeling, data mining, data science, and other related fields.

MACHINE LEARNING BACKGROUND

MULTIDISCIPLINARY NATURE OF BIG DATA ANALYSIS



MACHINE LEARNING | EVERYDAY USE CASES

- Internet search
- Digital adds
- Recommenders
- Image recognition – tagging photos, etc.
- Fraud/risk
- Upselling/cross-selling
- Churn
- Human resource management

EXAMPLES OF MACHINE LEARNING IN HEALTHCARE, BANKING, INSURANCE AND OTHER INDUSTRIES

TARGETED TREATMENT

- ~232,000 people in the US diagnosed with breast cancer in 2015
- ~40,000 will die from breast cancer in 2015

	Men	Women
Incidence (new cases)	1.3 per 100,000	124.3 per 100,000
Mortality (deaths)	0.3 per 100,000	21.5 per 100,000

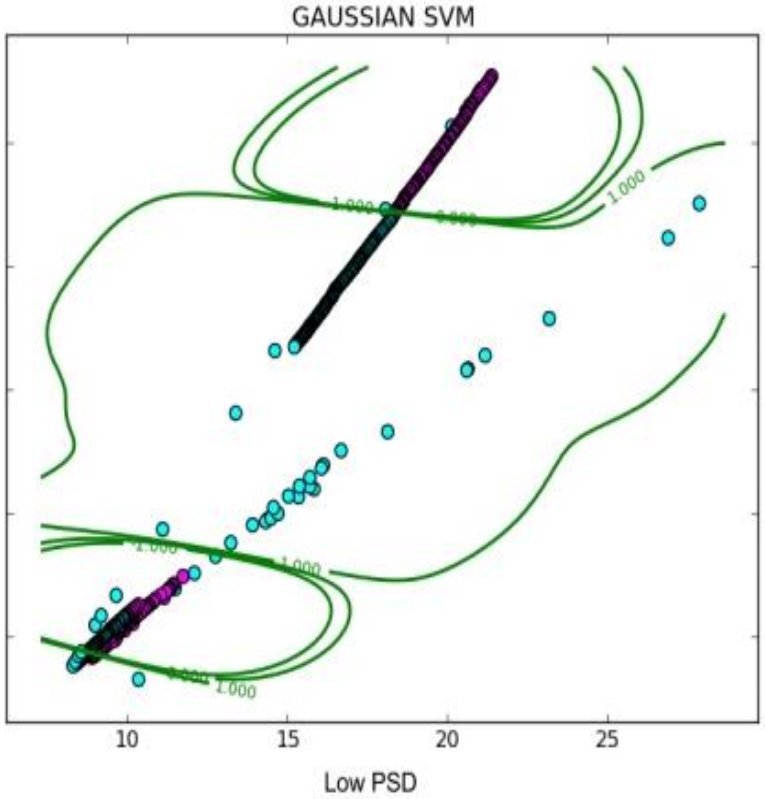
- 60 different drugs approved by FDA
- Tamoxifen
  - Harsh side effects (including increased risk of uterine cancer)
  - Success rate of 80%

80% effective in 100% of patients? → Big Data + Compute Resources + Machine Learning → 100% effective in 80% of patients (0% effective in rest)

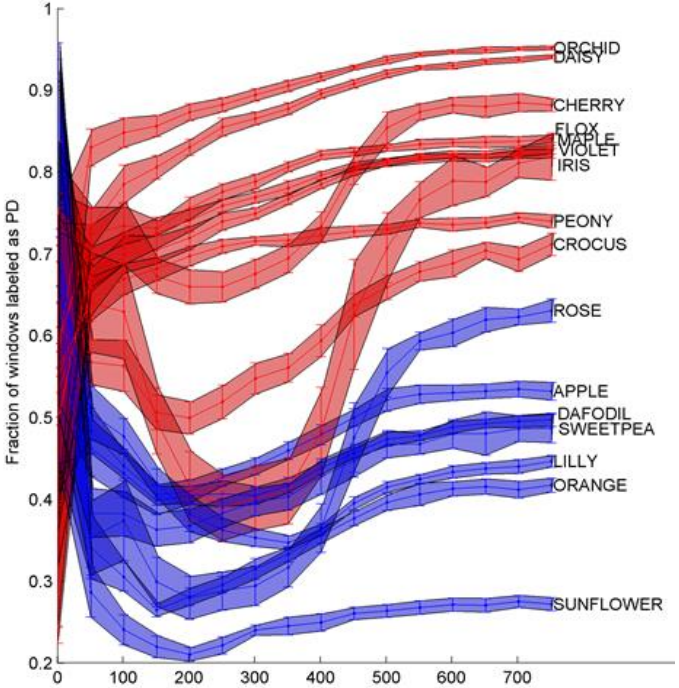
Big Data, Data Mining, and Machine Learning, Jared Dean

EARLY DETECTION AND MONITORING

- $10,000 Parkinson's Data Challenge
- 6000 hours of data from 16 subjects (9 PD, 7 Control)
- Data collected from mobile phones
  - Audio
  - Accelerometry
  - Compass
  - Ambient
  - Proximity
  - Battery level
  - GPS



Source: Michael Wang



Source: Andrea Mannini

MEDICATION ADHERENCE

- 32 million Americans use three or more medicines daily

> ***Drugs don't work in patients who don't take them***
> C. Everett Koop, MD

- 75% of adults are non-adherent in one or more ways
- The economic impact of non-adherence is estimated to cost $100 billion annually

AiCure uses mobile technology and facial recognition to determine if the right person is taking a given drug at the right time.

Face recognition
Medication identification
Ingestion

HIPAA-compliant network

BIG DATA IN DRUG DEVELOPMENT

Berg, a Boston-area startup, studies healthy tissues to understand the body's molecular and cellular natural defenses – and what leads to a disease's pathogenesis.

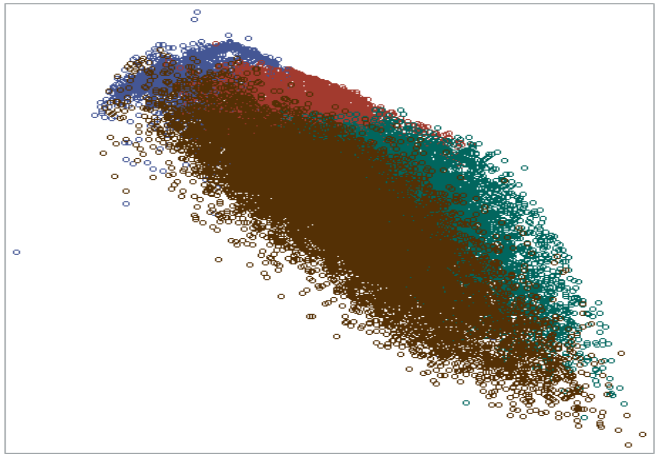
It's using concepts of machine learning and big data to scope out potential drug compounds – ones that could have more broad-ranging benefits, pivoting away from today's trend toward targeted therapies.

MedCityNews

MACHINE LEARNING IN INSURANCE

AUTOMOBILE INSURANCE

- Telematics driving behavior captured from a mobile app
- 46 Variables: Speed, acceleration, deceleration, jerk, left turns, right turns, …
- Goal: Stratify observations into low to high risk groups
- Cluster analysis and principle component analysis is used



Cluster One
- 7% of the trips
- Avg. trip – 11 minutes
- Slow, Short Trips

Cluster Two
- 18% of the trips
- Avg. trip – 16 min
- Less Aggressive

Population
- Average trip – 18 min
- 50% of time spent under 20 MPH

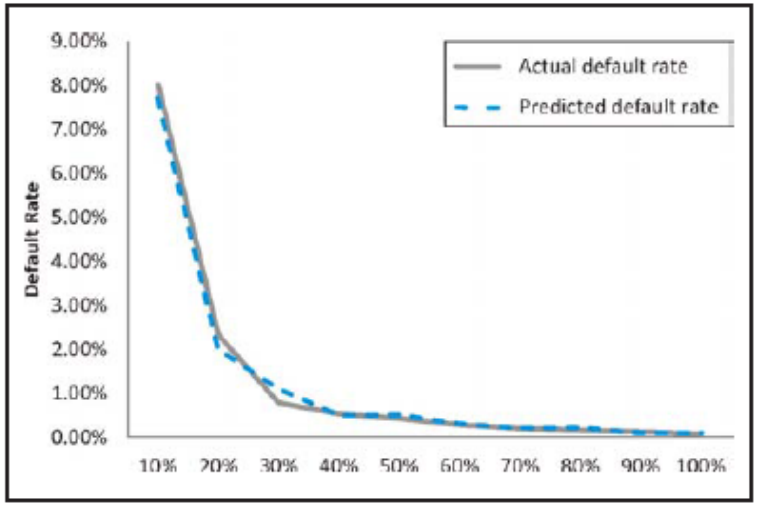
Cluster Three
- 25% of the trips
- Avg. trip – 28 min
- Fast, less aggressive, long Trips

Cluster Four
- 46% of the trips
- Avg. trip – 17 min
- More Aggressive

MACHINE LEARNING IN BANKING

FRAUD RISK PROTECTION MODEL USING LOGISTIC REGRESSION

- The model assigns a ‘probability score’ of default (PF) to each merchant for a possible fraud risk.
- PF score warns the management in advance of probable future losses on merchant accounts.
- Banks rank order merchants based on their PF score, and focus on the relatively riskier set of merchants.



The model can capture 62 percent frauds in the first decile

MACHINE LEARNING FOR RECOMMENDATION

Netflix Data Set

	movie 1	movie 2	movie 3	movie 4	…
User A	3	2	?	5	…
User B	?	5	4	3	..
User C	1	?	?	1	…
User D	4	1	5	2	…
User E	1	3	?	?	…
.	.	.	.	.	.

- 480K users
- 18K movies
- 100M ratings 1-5 (99%ratings missing)

Goal:

$ 1 M prize for 10% reduction in RMSE over Cinematch

BelKor's Pragmatic Chaos

Declared winners on 9/21/2009

Used ensemble of models, an important ingredient being Low-rank factorization

MACHINE LEARNING IN RETAIL

TARGETED MARKETING

- Another form of recommender
- Only send to those who need incentive

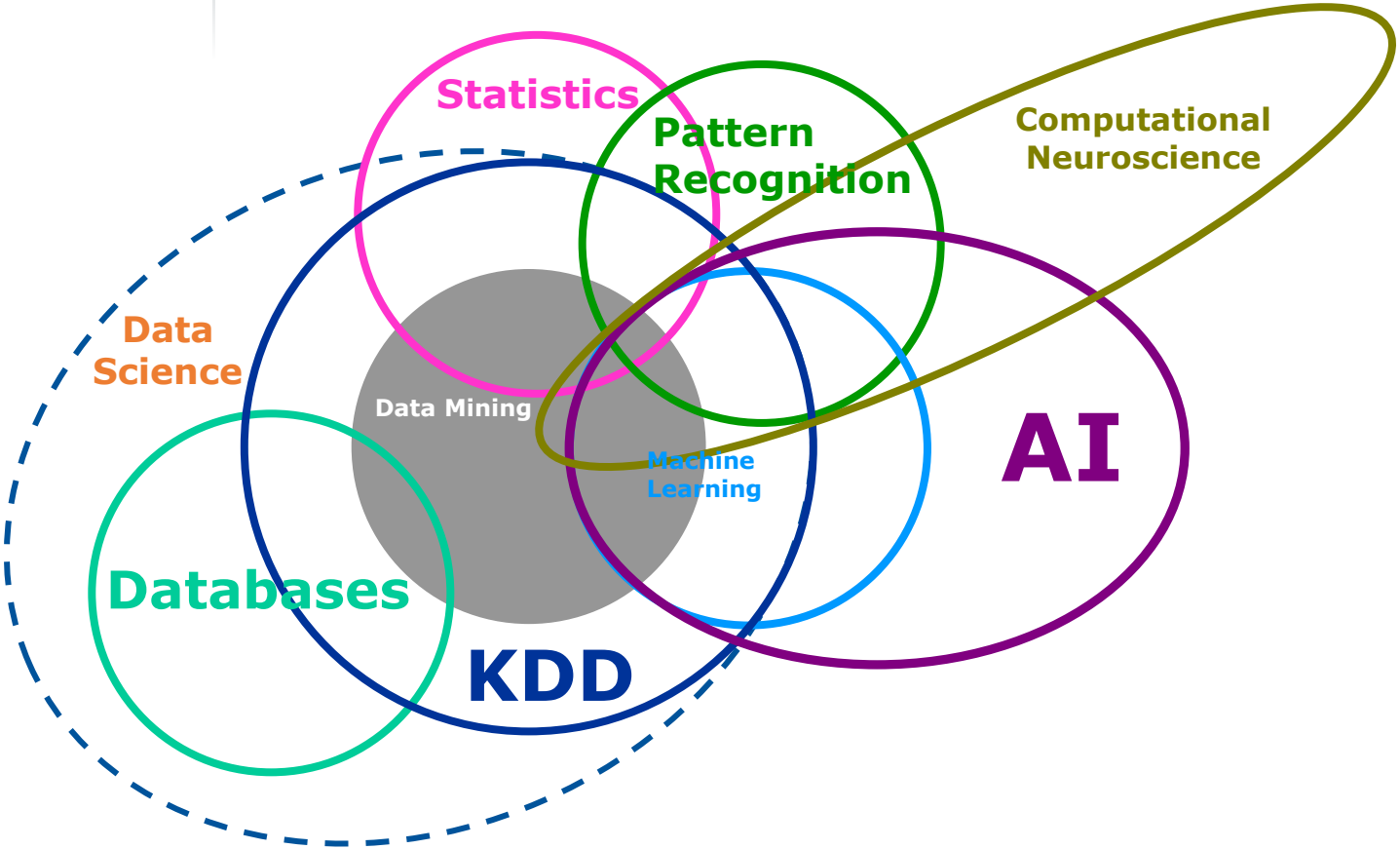
MACHINE LEARNING: THE ALGORITHMS

TERMINOLOGY

Machine Learning Term	Multidisciplinary Synonyms
Case, instance, example	Observation, record, row, data point
Feature, input	Independent variable, variable, column
Label	Dependent variable, target
Class	Categorical target variable level
Train	Fit
Score	Predict

MACHINE LEARNING BACKGROUND

MULTIDISCIPLINARY NATURE OF BIG DATA ANALYSIS



Machine Learning

Data Mining

SUPERVISED LEARNING	UNSUPERVISED LEARNING	SEMI-SUPERVISED LEARNING

– Regression
  – LASSO regression
  – Logistic regression
  – Ridge regression
– Decision tree
  – Gradient boosting
  – Random forests
– Neural networks
– SVM
– Naive Bayes
– Neighbors
– Gaussian processes
	– A priori rules
	– Clustering
	  – *k*-means clustering
	  – Mean shift clustering
	  – Spectral clustering
	– Kernel density estimation
	– Nonnegative matrix factorization
	– PCA
	  – Kernel PCA
	  – Sparse PCA
	– Singular value decomposition
	– SOM
		– Prediction and classification*
		– Clustering*
		– EM
		– FSVM
		– Manifold regularization
		– Autoencoders
		  – Multilayer perceptron
		  – Restricted Boltzmann machines

Know target	**Don't know target**	**Sometimes know target**

TRANSDUCTION

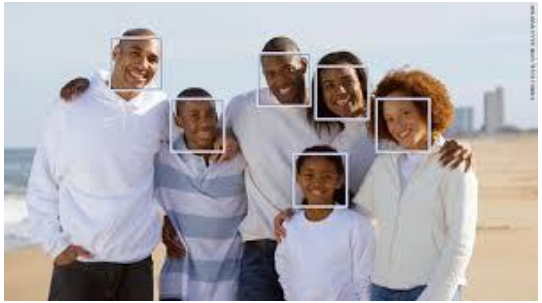
REINFORCEMENT LEARNING

DEVELOPMENTAL LEARNING

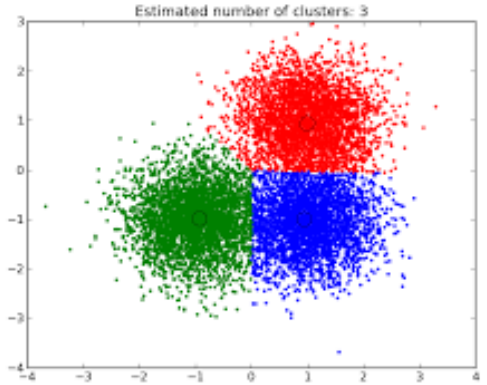
*In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering.

PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

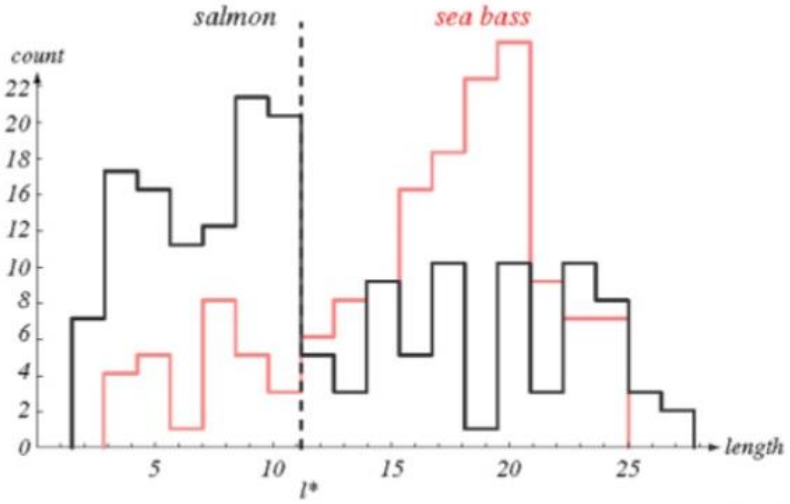
PCA IS USED IN COUNTLESS MACHINE LEARNING APPLICATIONS



- Fraud detection
- Word and character recognition
- Speech recognition
- Email spam detection
- Texture classification
- Face Recognition



SEA BASS OR SALMON?

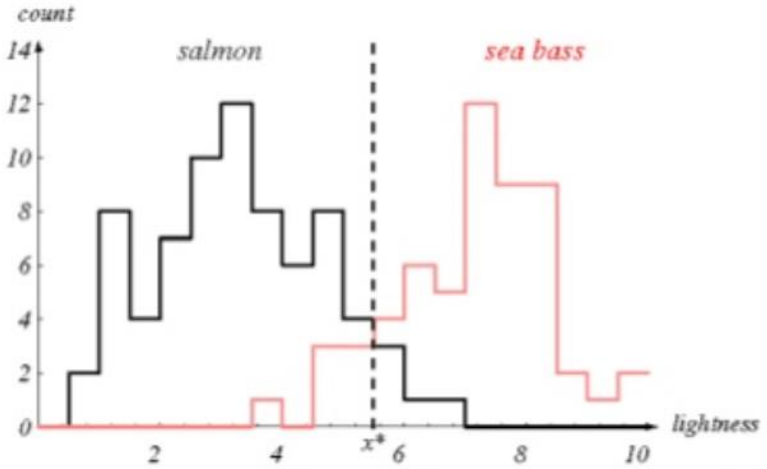


Can sea bass and salmon be distinguished based on length?

Length is not a good classifier!

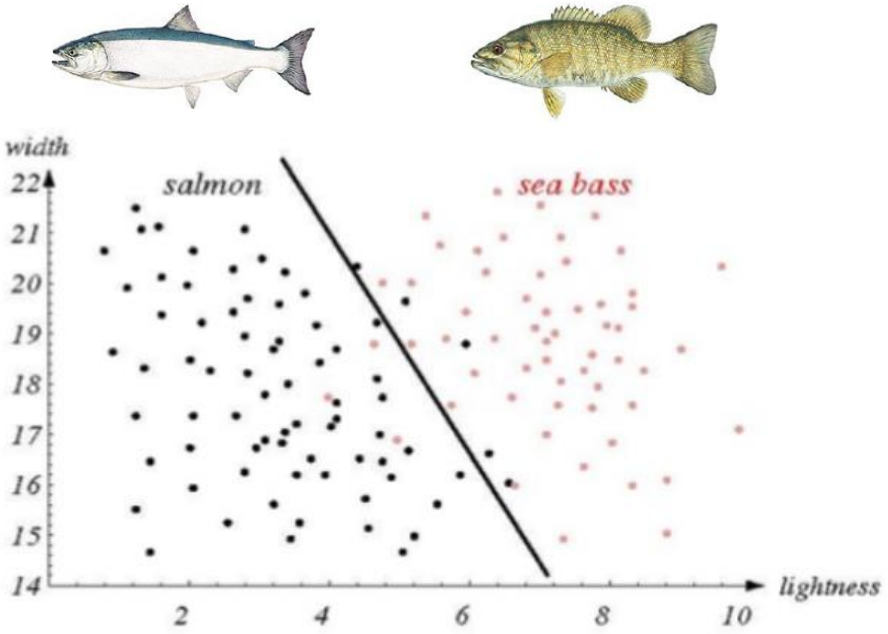
SEA BASS OR SALMON?

Can sea bass and salmon be distinguished based on lightness?



Lightness provides a better separation

SEA BASS OR SALMON?



Width and lightness used jointly

Almost perfect separation!

TYPICAL DATA MINING PROBLEM: 100s OR 1000s OF FEATURES

Key questions:

- Should we use all available features in the analysis?
- What happens when there is feature redundancy or correlation?

When the number of features is large, they are often correlated. Hence, there is redundancy in the data.

WHAT IS PRINCIPAL COMPONENT ANALYSIS?

Principal component analysis (PCA) converts a set of possibly correlated features into a set of linearly uncorrelated features called principal components.

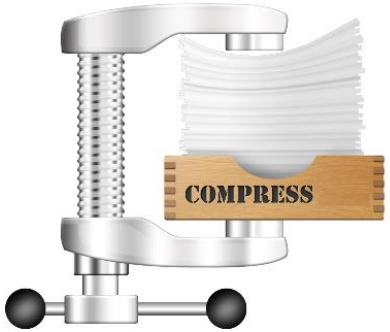
Principal components are:

- Linear combinations of the original variables
- The most meaningful basis to re-express a data set

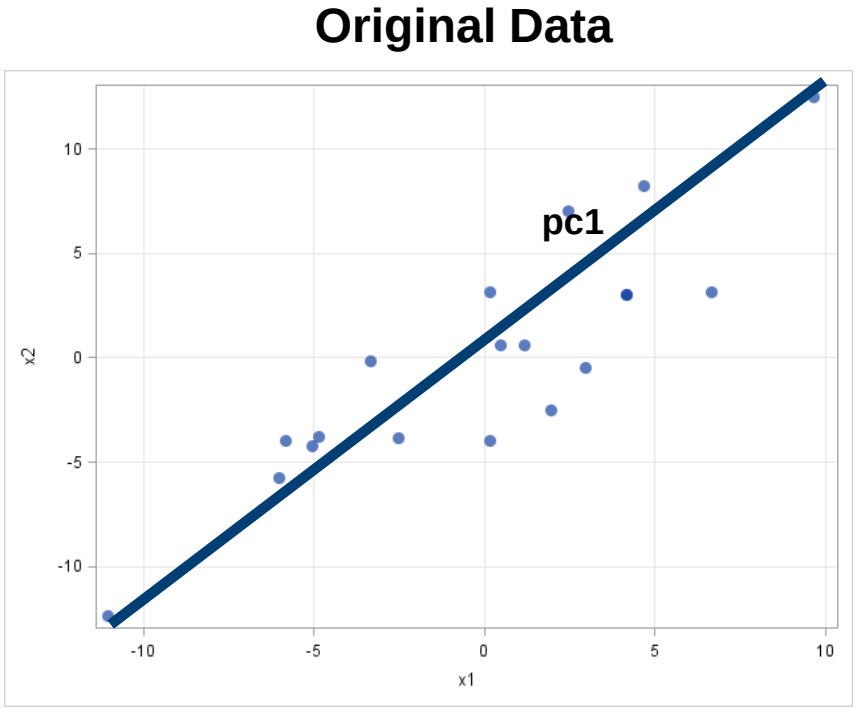
WHAT IS PRINCIPAL COMPONENT ANALYSIS?

Reduce the dimensionality:

- Reduce memory and disk space needed to store the data
- Reveal hidden, simplified structures
- Solve issues of multicollinearity
- Visualize higher-dimensional data
- Detect outliers



TWO-DIMENSIONAL EXAMPLE

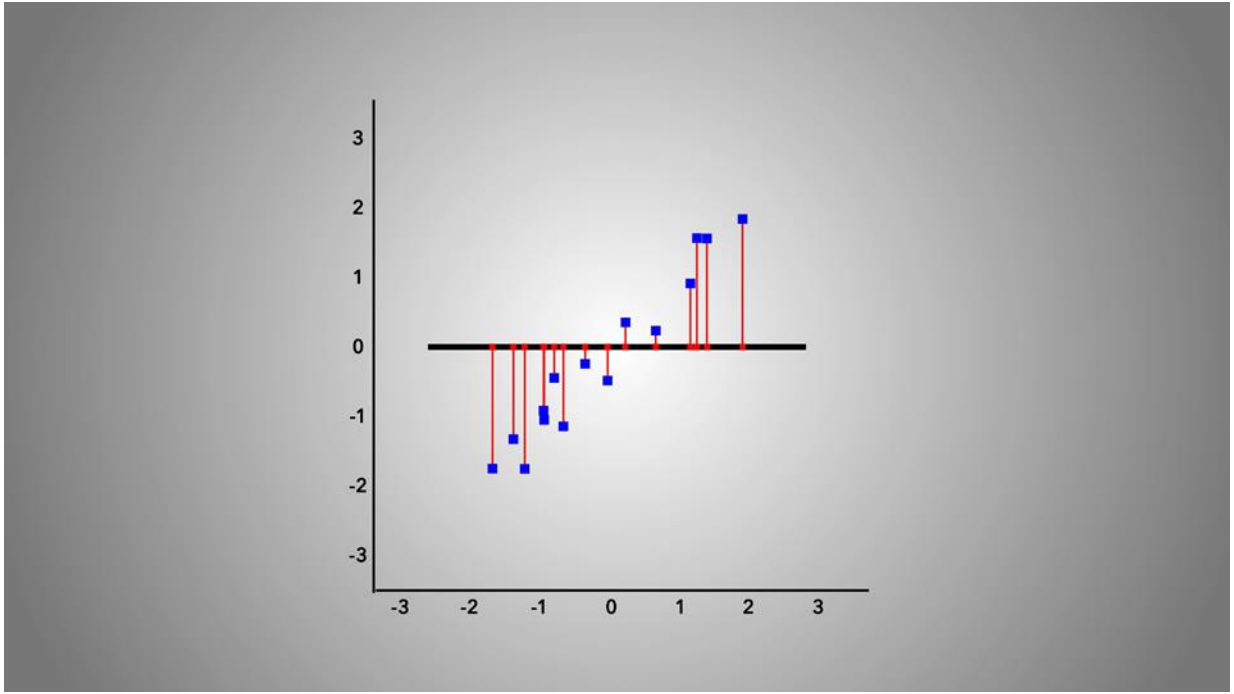


The first principle component is found according to two criteria:

- The variation of values along the principal component direction should be maximal
- The reconstruction error should be minimal if we reconstruct the original variables.

PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

TWO-DIMENSIONAL EXAMPLE

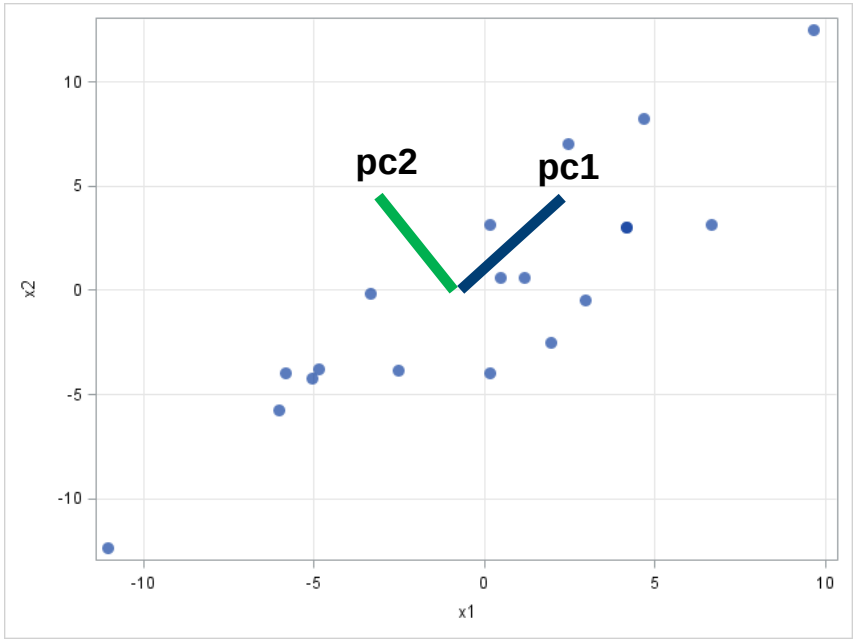


PCA.mp4

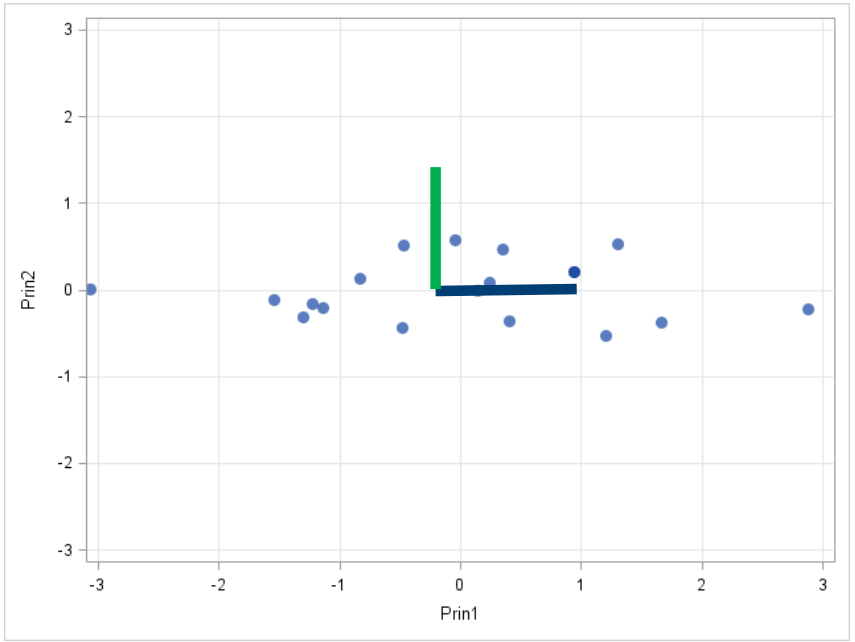
PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

TWO-DIMENSIONAL EXAMPLE

Original Data

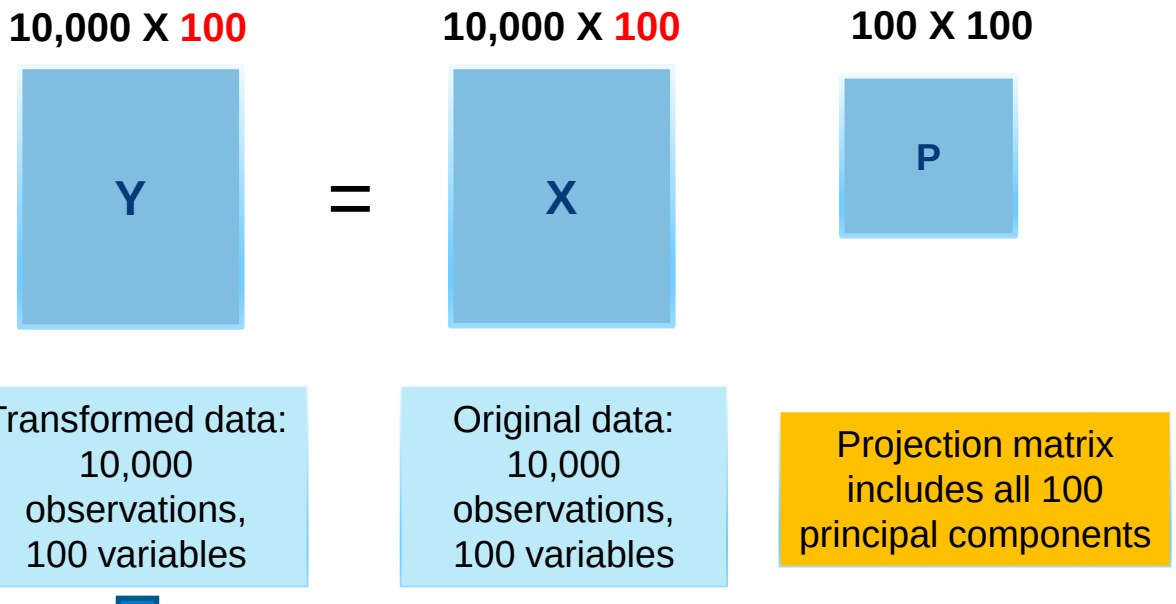


PCA Output



PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

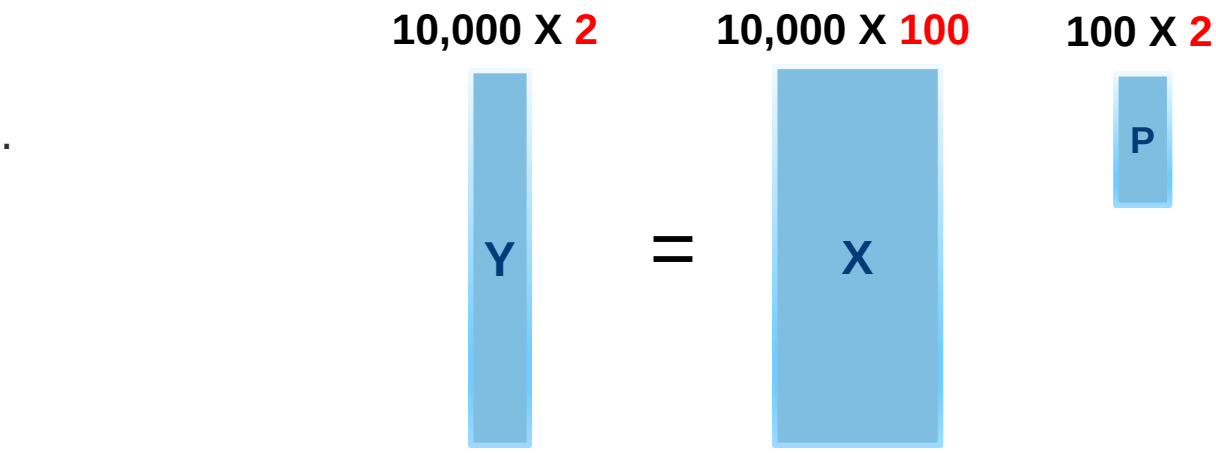
NO DIMENSION REDUCTION: 100 → 100



PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

DIMENSION REDUCTION: 100 → 2

Choose the first two principal components that have the highest variance!

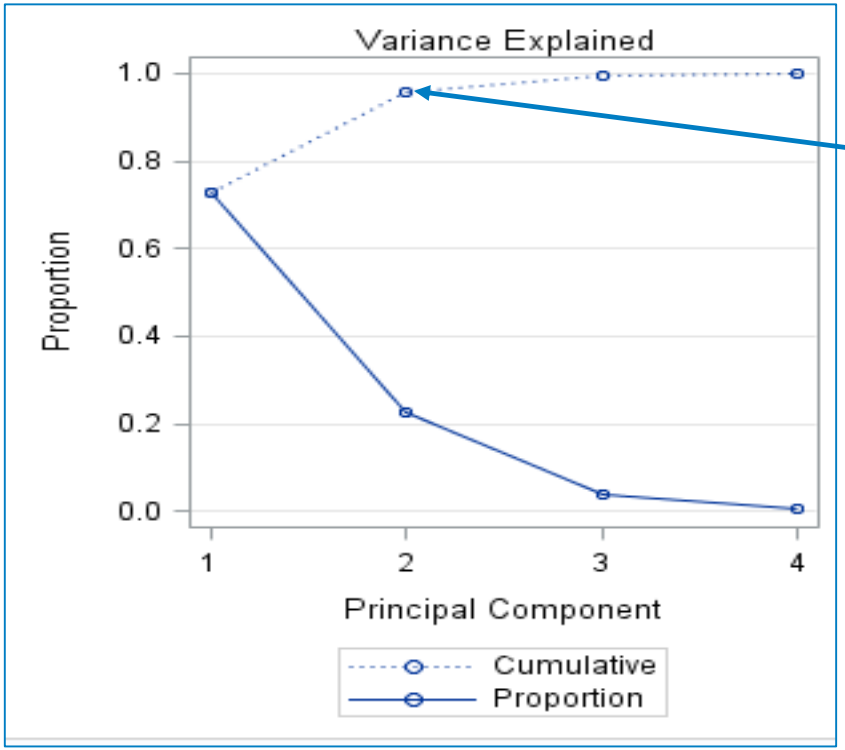


PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

HOW TO CHOOSE THE NUMBER OF PRINCIPAL COMPONENTS?

```
proc princomp  data=SASHELP.IRIS;
    var  SepalLength SepalWidth
         PetalLength PetalWidth;
run;
```

Eigenvectors			
		Prin1	Prin2
SepalLength	Sepal Length (mm)	0.521066	0.377418
SepalWidth	Sepal Width (mm)	-.269347	0.923296
PetalLength	Petal Length (mm)	0.580413	0.024492
PetalWidth	Petal Width (mm)	0.564857	0.066942



96% of the variance in the data is explained by the **first two** principal components

FACE RECOGNITION EXAMPLE

Training set consists of 100 images

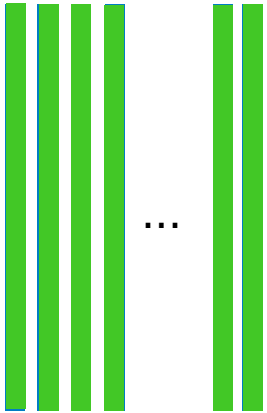
For each image

Face image

50x50

Face vector

2500x1



100 face vectors

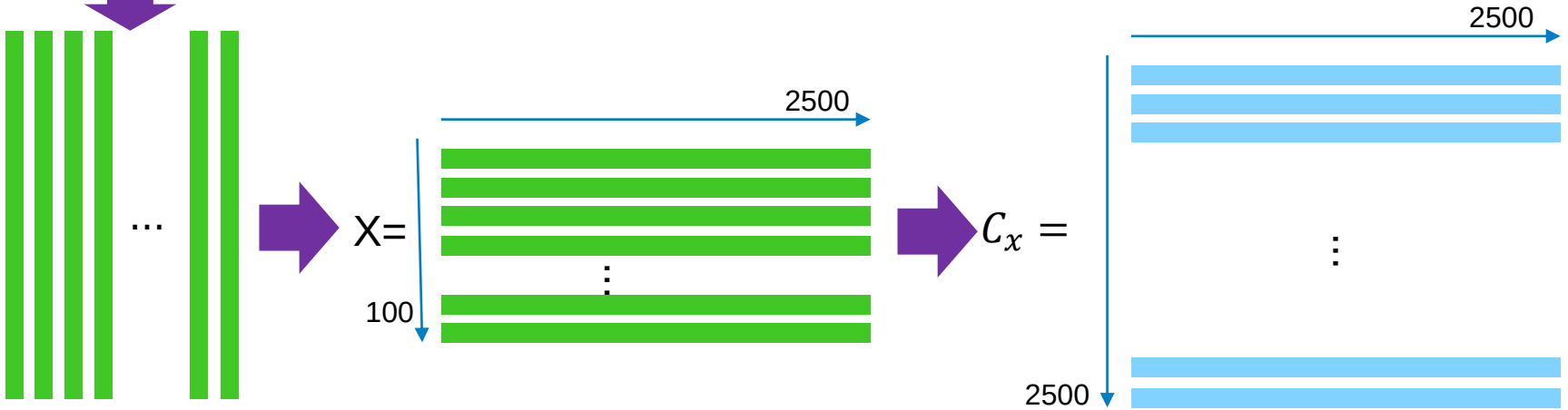
Calculate average face vector

Subtract average face vector from each face vector

FACE RECOGNITION EXAMPLE

Training set consists of 100 images

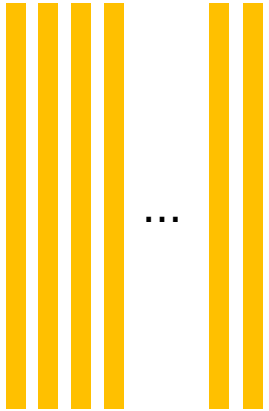
Calculate principal components of X, which are eigenvectors of $C_x = X^t X$



100 face vectors

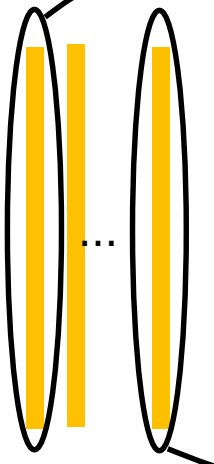
FACE RECOGNITION EXAMPLE

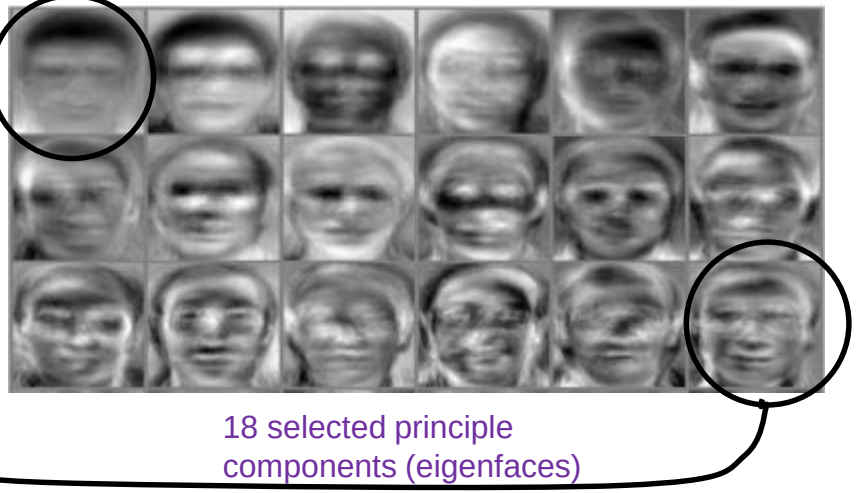
Training set consists of 100 images



2500 principle components, each 2500 x 1 dimensional

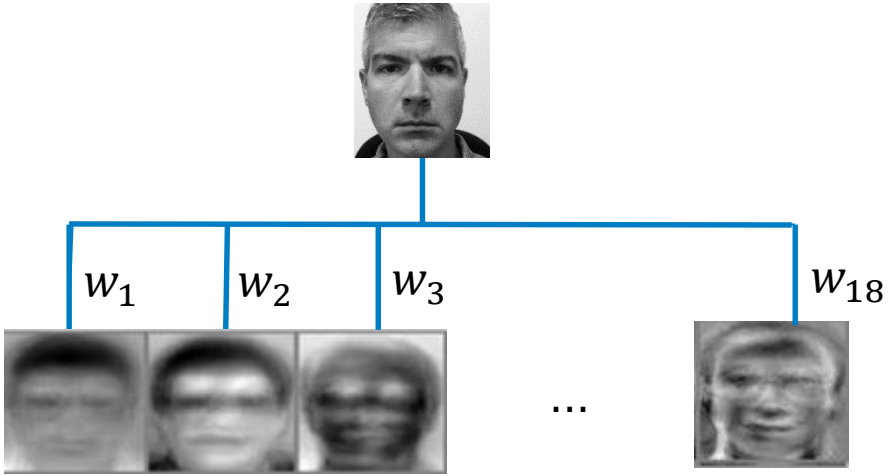
Select the 18 principle components that have the largest variation





18 selected principle components (eigenfaces)

FACE RECOGNITION EXAMPLE



Represent each image in the training set as a linear combination of eigenfaces

$$\begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_{18} \end{bmatrix} \quad \cdots \quad \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_{18} \end{bmatrix}$$

Weight vector for image 1

Weight vector for image 100

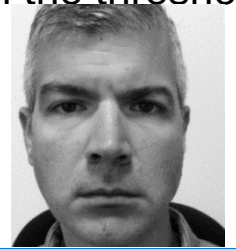
PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

RECOGNIZE UNKNOWN FACE

?

?

1. Convert the unknown ir
2. Standardize the face ve
3. Represent this face vec ation of 18 eigenfaces
4. Calculate the distance b veight vector and the weight vector for each image in the tr
5. If the smallest distance is larger than the threshold distance → unknown person
6. If the smallest distance is less than the threshold distance, then the face is recognized as →

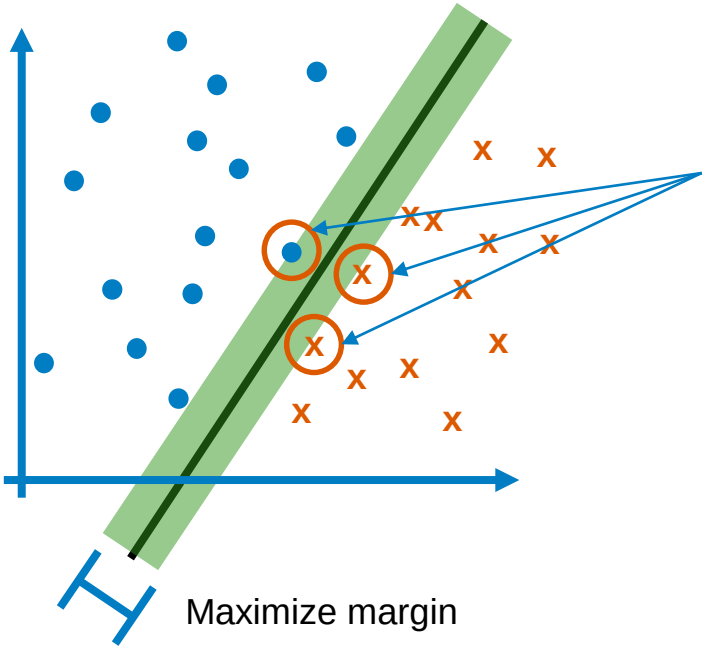


EXAMPLE 2: SUPERVISED LEARNING

SUPPORT VECTOR MACHINES

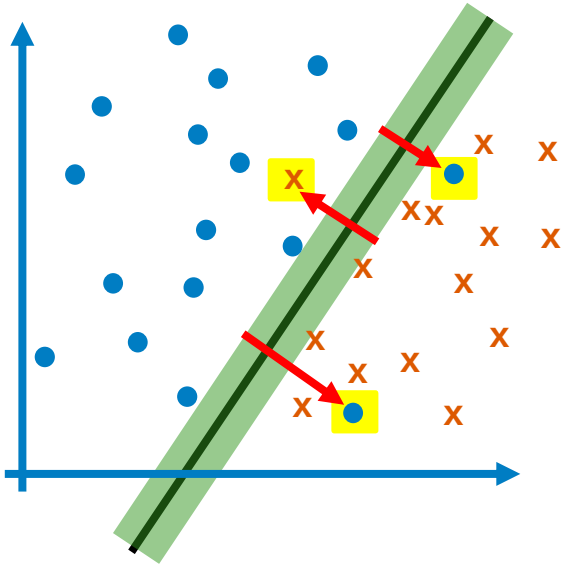
BINARY CLASSIFICATION MODELS

Construct a hyperplane that maximizes margin between two classes



Support vectors: points that lie on (define) the margin

INTRODUCING A PENALTY

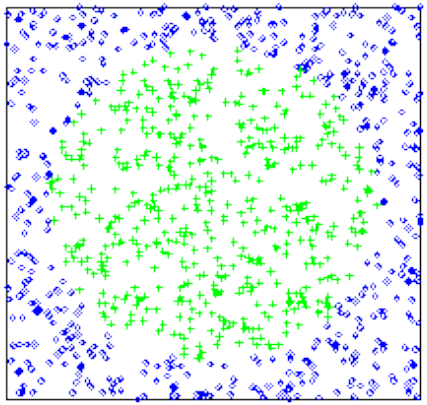


Introduce a penalty C based on distance from misclassified points to their side of the margin

Tradeoff:

- Wide margin = more training points misclassified but generalizes better to future data
- Narrow margin = fits the training points better but might be overfit to the training data

SEPARATION DIFFICULTIES



Circle-in-the-Square data set

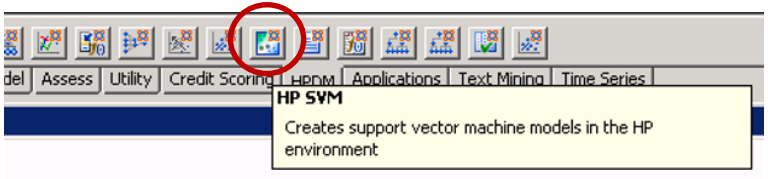
2 classes: In, Out

http://techlab.bu.edu/classer/data_sets

KEY TUNING PARAMETERS

HP SVM node in SAS Enterprise Miner

- Distributes margin optimization operations distributed
- Exploits all CPUs using concurrent threads



$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i,$$

Penalty C (default 1)

Linear

Polynomial (2 or 3 degrees)

$$K(x_i, x_j) = (ax_i^T x_j + r)^d$$

Radial basis function

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$$

Sigmoid function

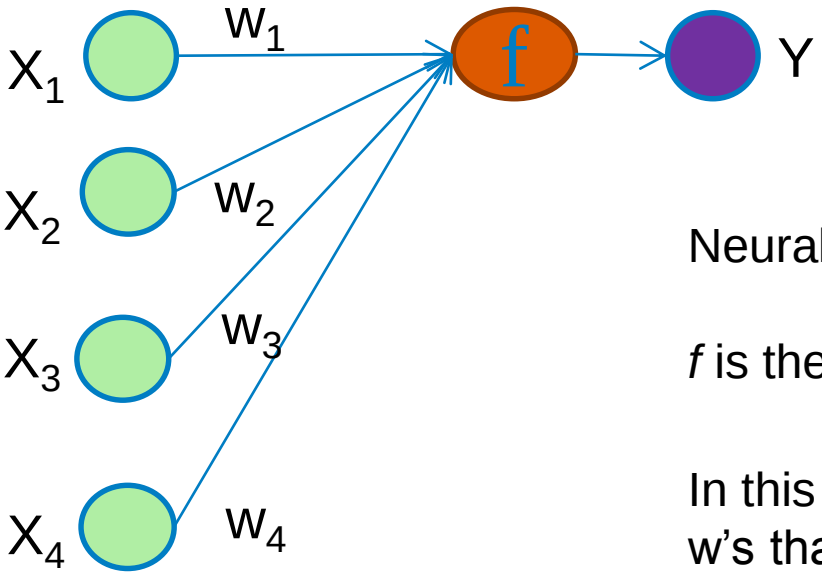
$$K(x_i, x_j) = \tanh(ax_i^T x_j + r)$$

Kernel and kernel parameters

EXAMPLE 3: SEMI-SUPERVISED LEARNING

NEURAL NETWORKS & DEEP LEARNING

NEURAL NETWORKS



Neural network **compute node**

f is the so-called **activation function**

In this example there are four weights w's that need to be determined

NEURAL NETWORKS

The prediction forumla for a NN is given by

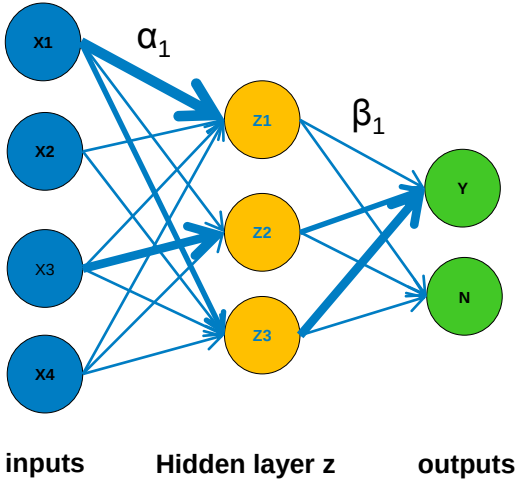
$$\mathrm{P(Y\,|X)} = g(T_Y)$$
$$T_Y = \beta_{0Y} + \beta_Y^T Z$$
$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X)$$

The functions g and σ are defined as

$$g(T_Y) = \frac{e^{T_Y}}{e^{T_N} + e^{T_Y}} \quad , \sigma(x) = \frac{1}{1+e^{-x}}$$

In case of a binary classifier $P(N|X) = 1 - P(Y|X)$

The model weights α and β have to be estimated from the data



ESTIMATING THE WEIGHTS

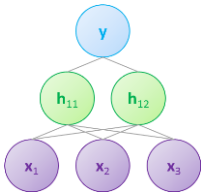
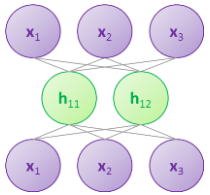
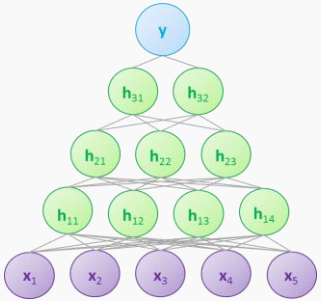
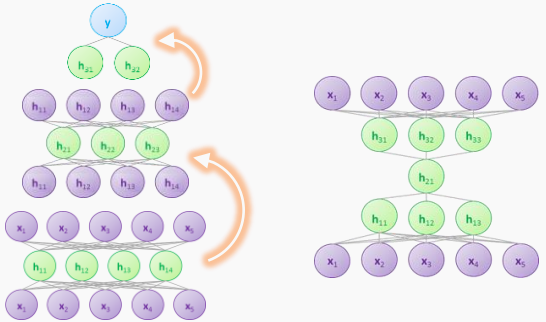
Back propagation algorithm

- Randomly choose small values for all w_i' s
- For each data point (observation)
  1. Calculate the neural net prediction
  2. Calculate the error E (for example: E = (actual – prediction)^2)
  3. Adjust weights w according to:

$$w^{new}{}_i = wi + \Delta wi$$

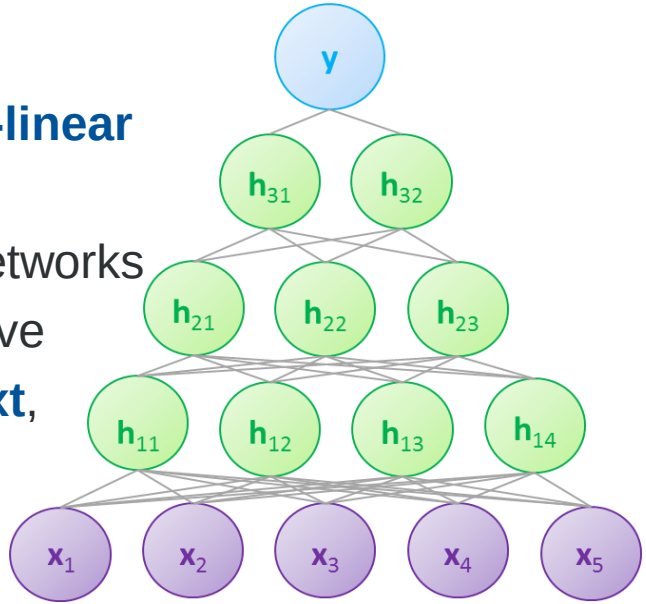
$$\Delta wi = -\alpha \frac{\partial E}{\partial w_i}$$

  4. Stop if error E is small enough.

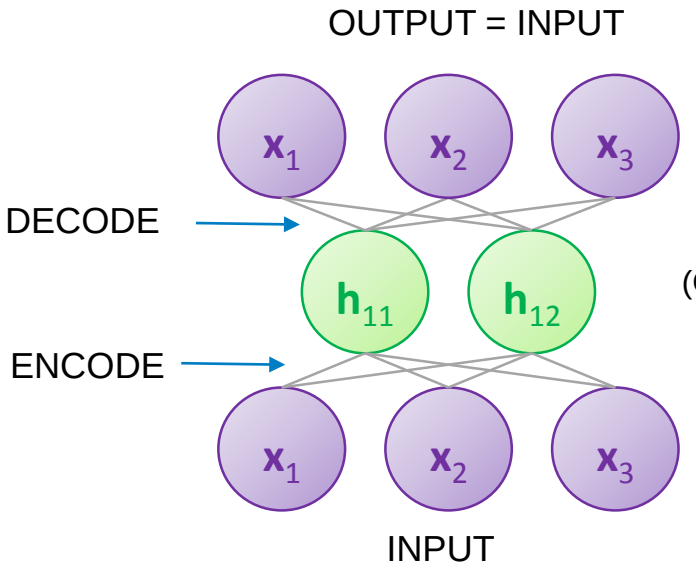
	Supervised	Unsupervised
Single-layer network	 Multilayer perceptron (MLP)	 Autoencoder (with noise injection: denoising autoencoder)
Deep Network	 Deep belief network or Deep MLP	 Stacked autoencoder (with noise injection: stacked denoising autoencoder)

WHAT IS DEEP LEARNING?

- **Neural networks** with **many layers** and different types of …
  - Activation functions.
  - Network architectures.
  - Sophisticated optimization routines.
- **Each layer represents an optimally weighted, non-linear combination of the inputs**.
- **Extremely accurate** predictions using deep neural networks
- The most common applications of deep learning involve **pattern recognition** in unstructured data, such as **text**, **photos**, **videos** and **sound**.



Unsupervised neural networks that use inputs to predict the inputs

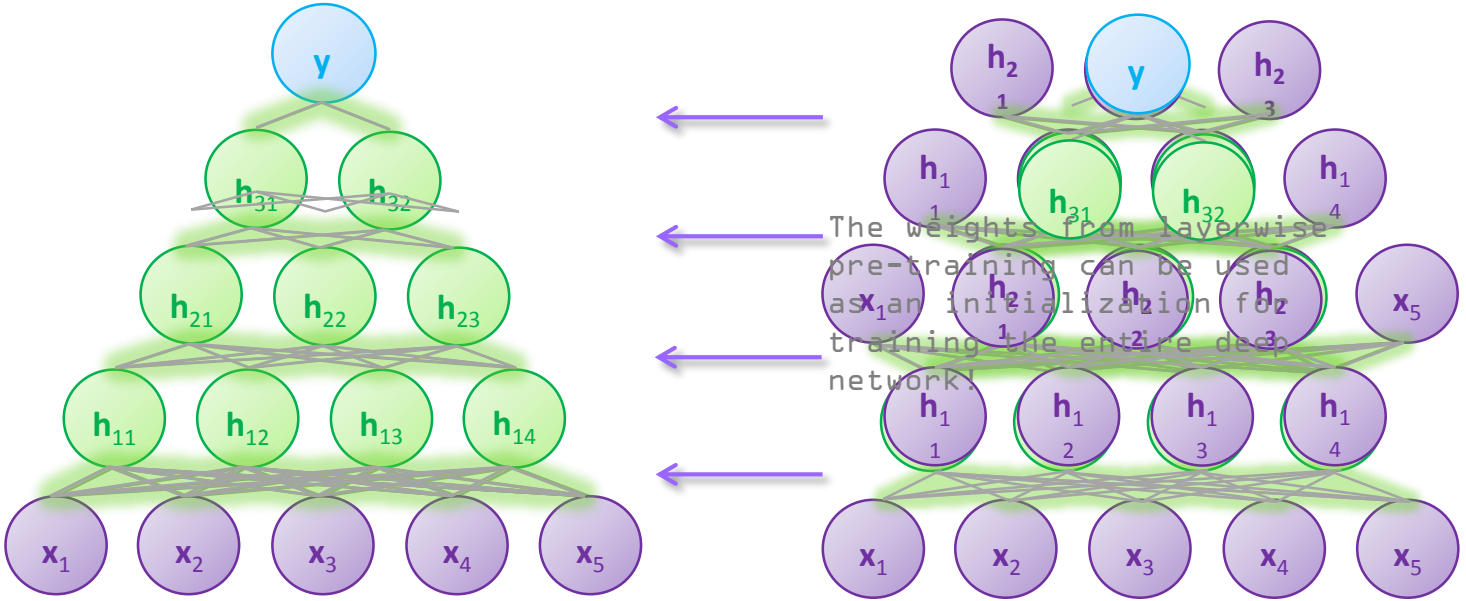


(Often more hidden layers with many nodes)

Note: Linear activation function corresponds with 2 dimensional principle components analysis

http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf

DEEP LEARNING

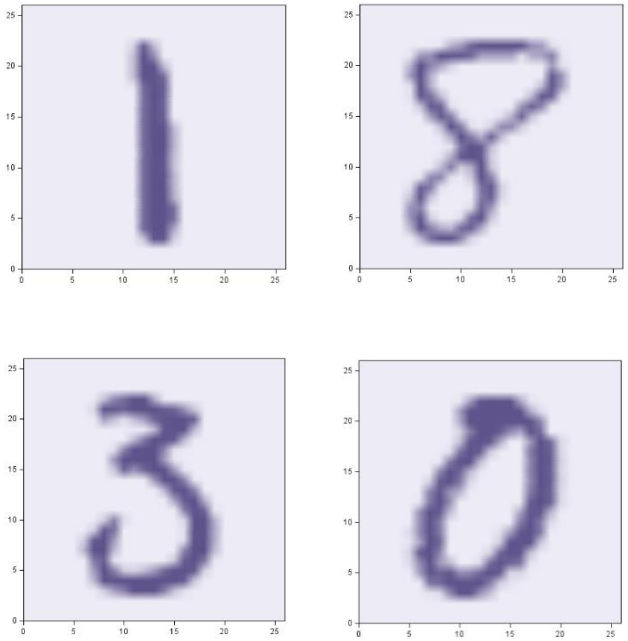


The weights from layerwise pre-training can be used as an initialization for training the entire deep network!

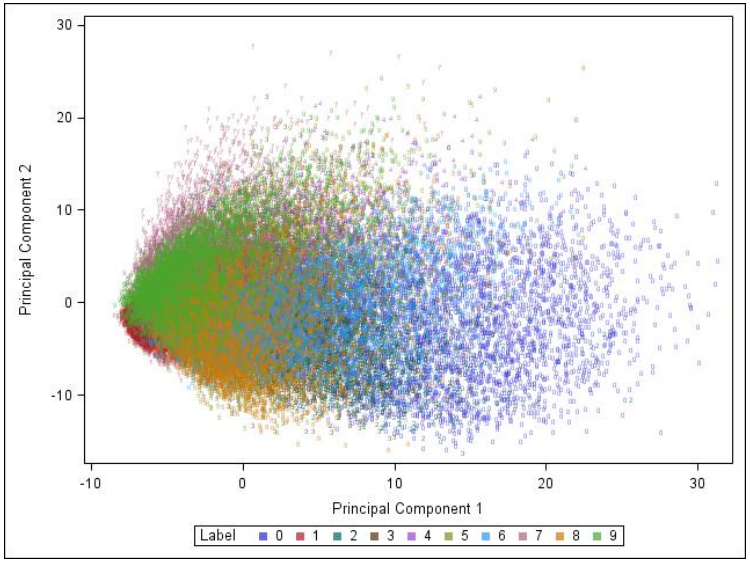
Many separate, unsupervised, single hidden-layer networks are used to initialize a larger unsupervised network in a layerwise fashion

DIGIT RECOGNITION - CLASSIC MNIST TRAINING DATA

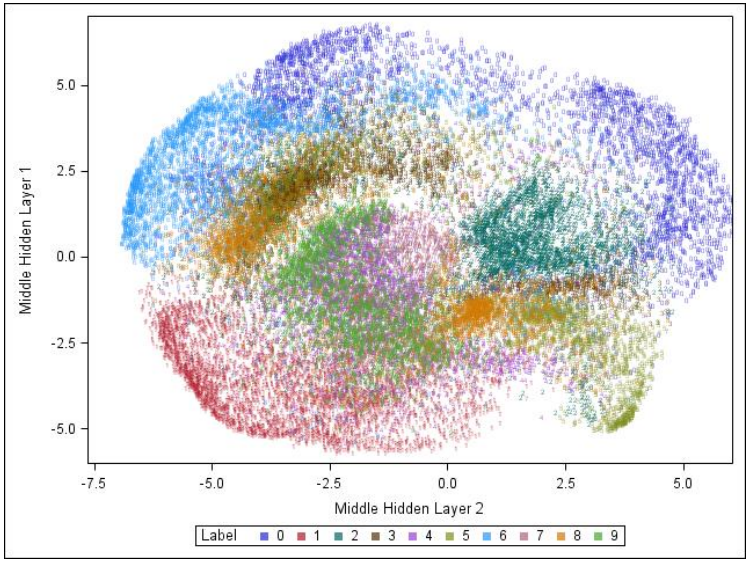
- 784 features form a 28x28 digital grid
- Greyscale features range from 0 to 255
- 60,000 labeled training images
  (785 variables, including 1 nominal target)
- 10,000 unlabeled test images
  (784 input variables)



First Two Principal Components



Output of middle hidden layer from 400-300-100-2-100-300-400 Stacked Denoising Autoencoder



MEDICARE PROVIDERS DATA

Medicare Data:

- Billing information
- Number of severity of procedures billed
- Whether the provider is a university hospital
- …

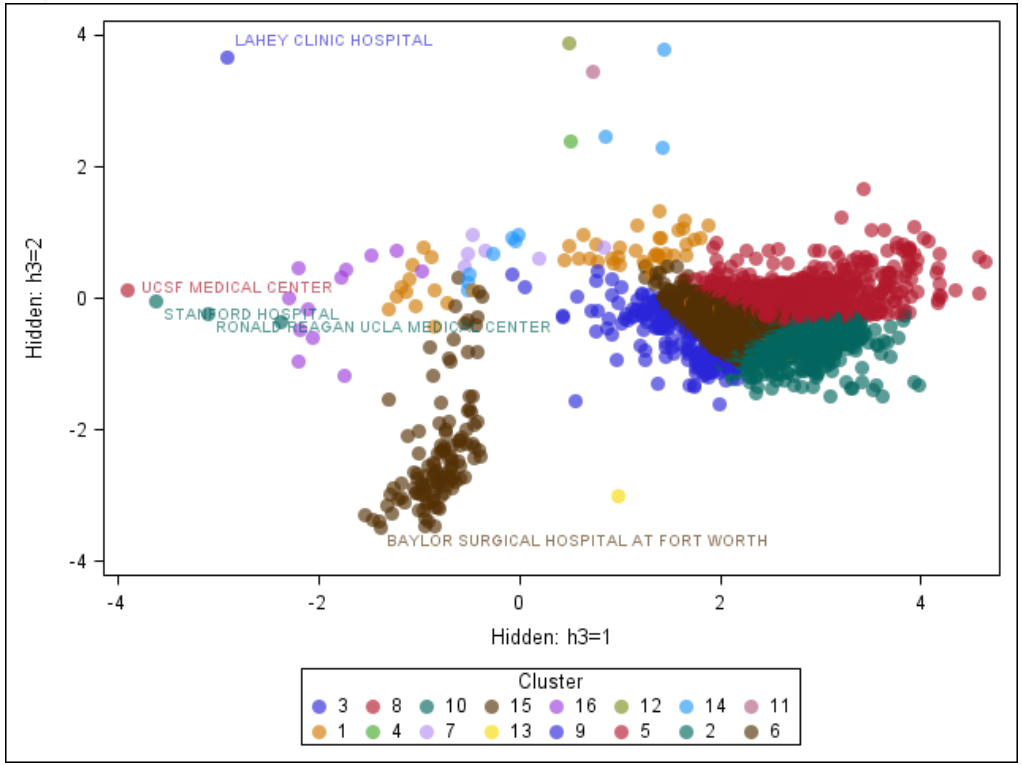
Goals:

- Which Medicare providers are similar to each other?
- Which Medicare providers are outliers?

Methodology to analyze the data:

- Create 16 k-means clusters
- Train simple **denoising autoencoder with 5 hidden layers**
- Score training data with trained neural network
- **Keep 2-dimensional middle layer** as new feature space
- Detect outliers as points from the origin of the 2-dimenasional space

VISUALIZATION AFTER DIMENSION REDUCTION



EXAMPLE 4: SUPERVISED LEARNING

ENSEMBLE MODELS

ENSEMBLE MODELING
FOR MACHINE LEARNING

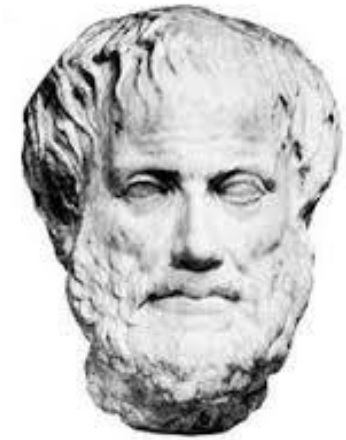
CONCEPT

http://www.englishchamberchoir.com/

Investment Portfolio

Blended Scotch

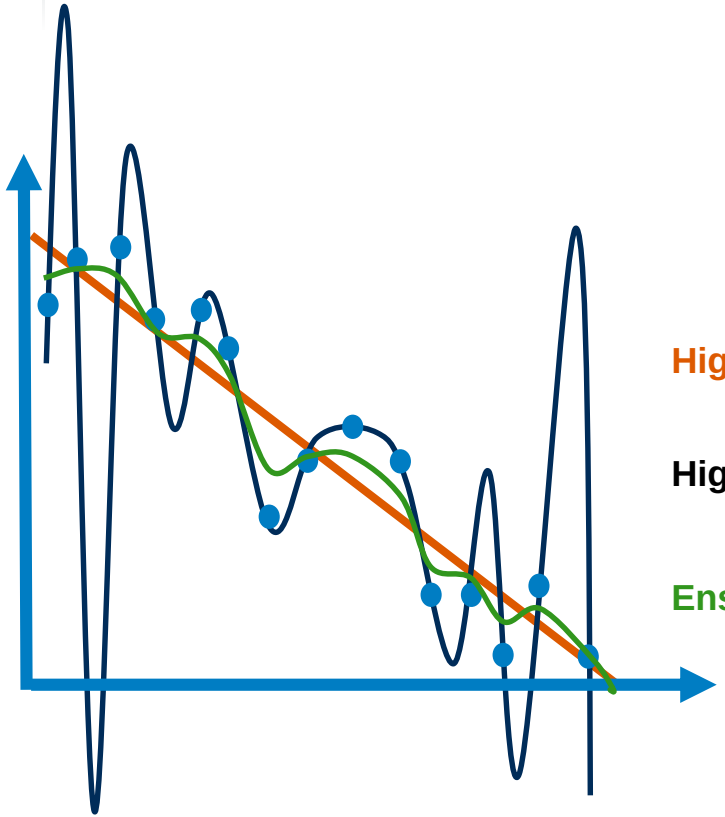
CONCEPT



"Wisdom of the crowd" – Aristotle ('Politics')

- collective wisdom of many is likely more accurate than any one

CONCEPT



High bias model (underfit)

+

High variance model (overfit)

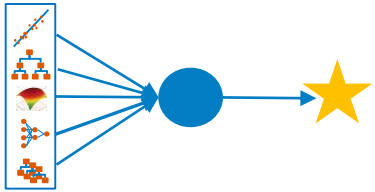
↓

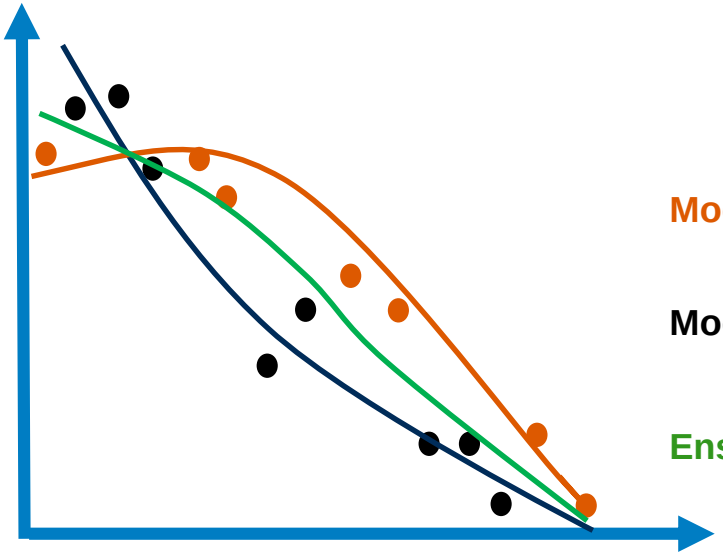
Ensemble (average)

ENSEMBLE MODELING
FOR MACHINE LEARNING

CONCEPT

- Combine strengths
- Compensate for weaknesses
- Generalize for future data





Model with Sample #1

+

Model with Sample #2

Ensemble (average)

COMMON USAGE

- Predictive modeling competitions - Teams very often join forces at later stages

KDD
bringing together the data mining, data science and analytics community since 1989

kaggle

NETFLIX

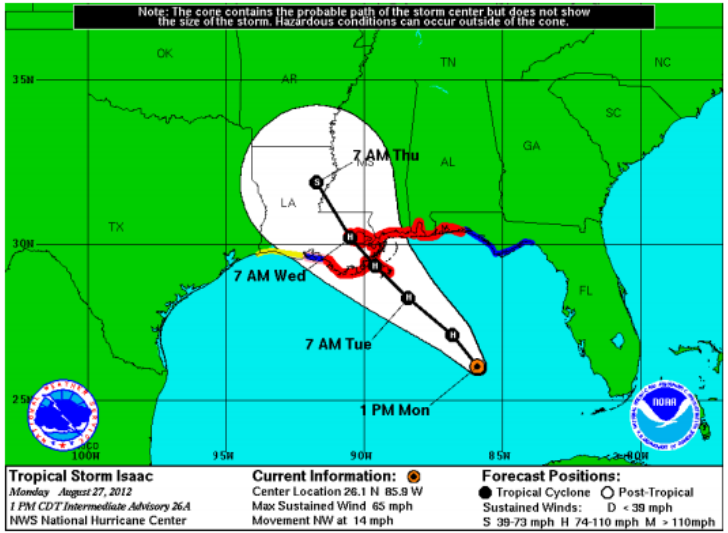
"Arguably, the Netflix Prize's most convincing lesson is that a disparity of approaches drawn from a diverse crowd is more effective than a smaller number of more powerful techniques" – Wired magazine

http://www.netflixprize.com/

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

COMMON USAGE

- Forecasting
  - Election forecasting (polls from various markets and demographics)
  - Business forecasting (market factors)
  - Weather forecasting



http://www.nhc.noaa.gov/

COMMON USAGE

- Meteorology ensemble models for forecasting

European Ensemble Mean

European Deterministic (Single)

http://www.wral.com/fishel-weekend-snow-prediction-is-an-outlier/14396482/

ASSIGNING OUTPUT

- Interval Targets – Average value
- Categorical Targets – Average of posterior probabilities OR majority voting

Model	Probability of YES	Probability of NO	
Model 1	0.6	0.4	Yes
Model 2	0.7	0.3	Yes
Model 3	0.4	0.6	No
Model 4	0.35	0.65	No
Model 5	0.3	0.7	No
	0.47	0.53 NO	

Posterior probabilities:
Yes = 0.4, No = 0.6

https://communities.sas.com/thread/78171

- Spread of ensemble member results gives indication of reliability

APPROACHES

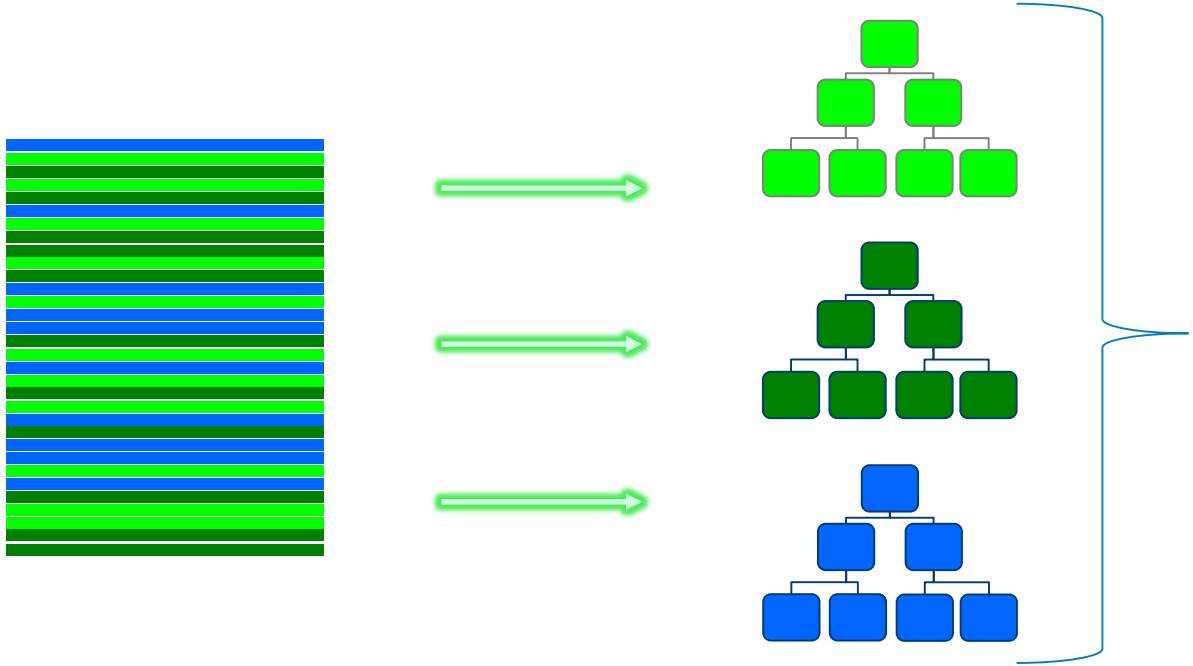
Ways to incorporate multiple models:

- One algorithm, different data samples
- One algorithm, different configuration options
- Different algorithms
- Expert knowledge

APPROACHES

One algorithm, different data samples

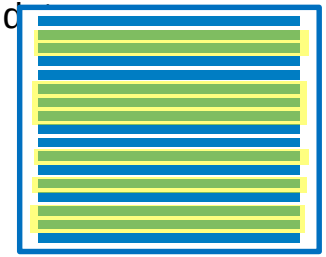
Bagging (Bootstrap Aggregating) – parallel training and combining of base learners

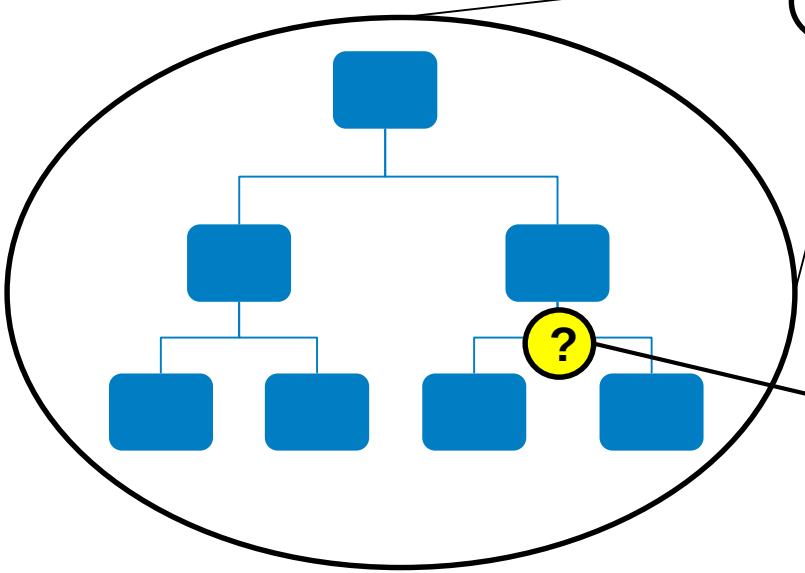


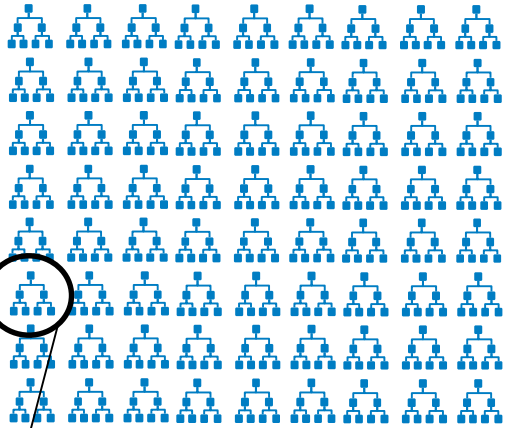
RANDOM FORESTS

- Ensemble of decision trees
- Score new observations by majority vote or average

Each tree is trained from a sample of the full d



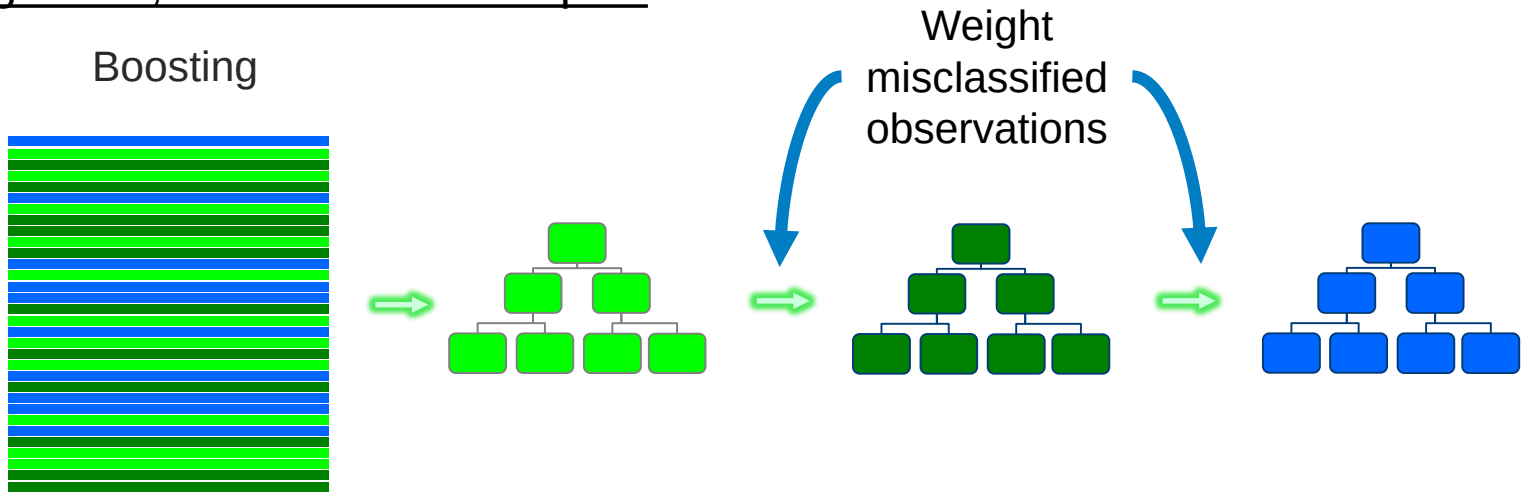




Variable candidates for splitting rule are random subset of all variables (bias reduction)

One algorithm, different data samples

Boosting



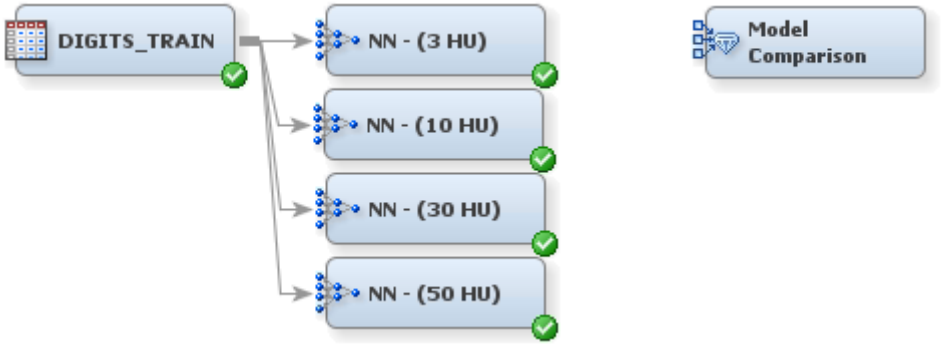
Gradient Boosting

Each iteration trains a model to predict the residuals of the previous model (ie, model the error)

Boosting degrades with noisy data – weights may not be appropriate from one run to the next

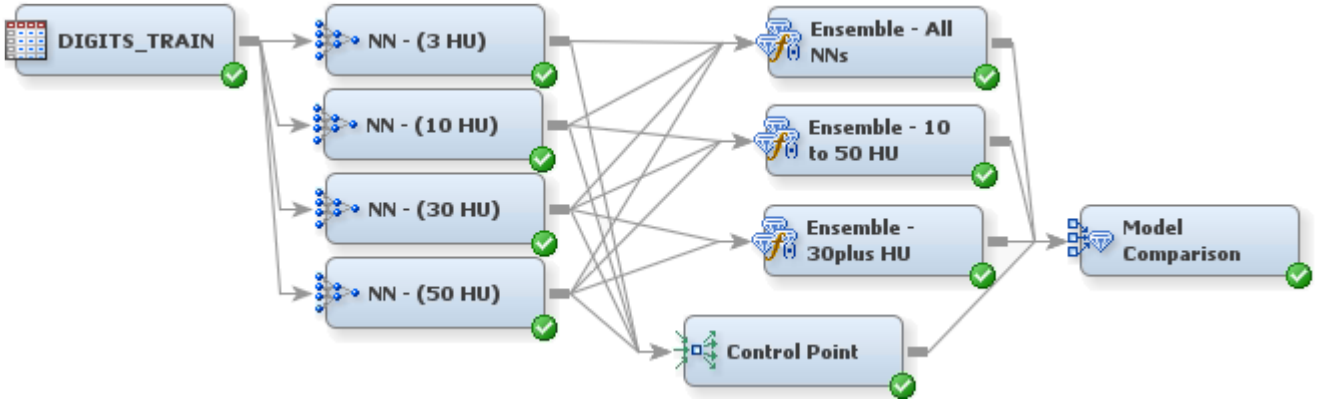
MODIFIED ALGORITHM OPTIONS

One algorithm, different configuration options



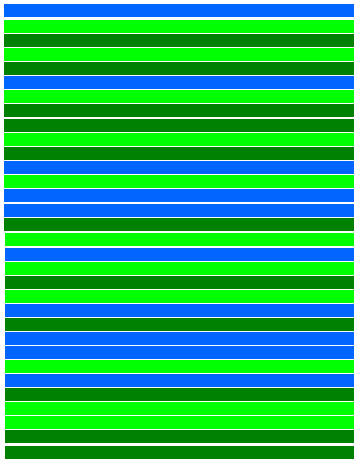
MODIFIED ALGORITHM OPTIONS

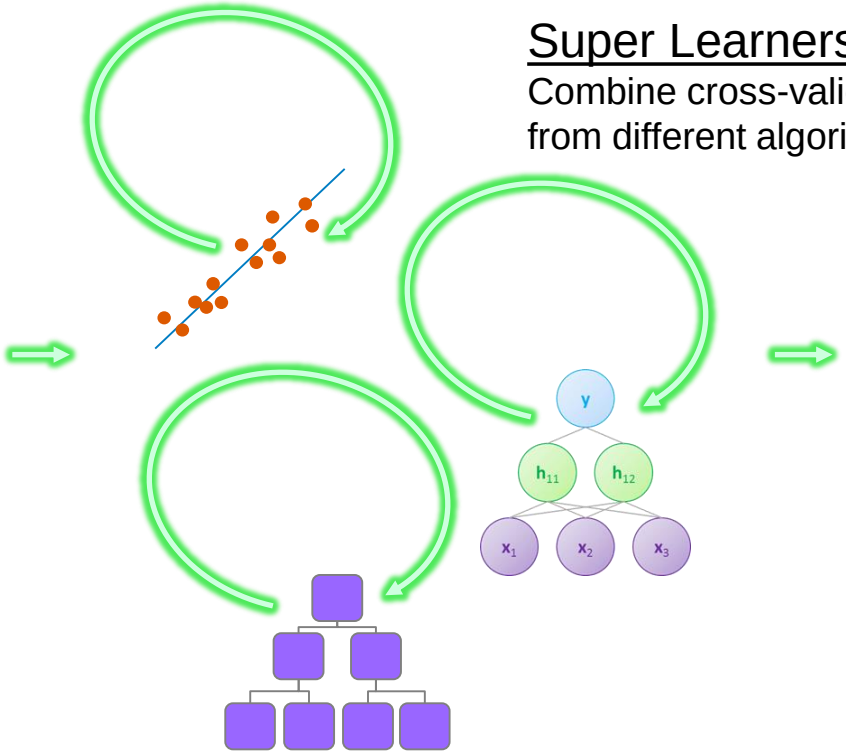
One algorithm, different configuration options



MULTIPLE ALGORITHMS

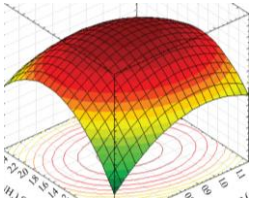
Different algorithms





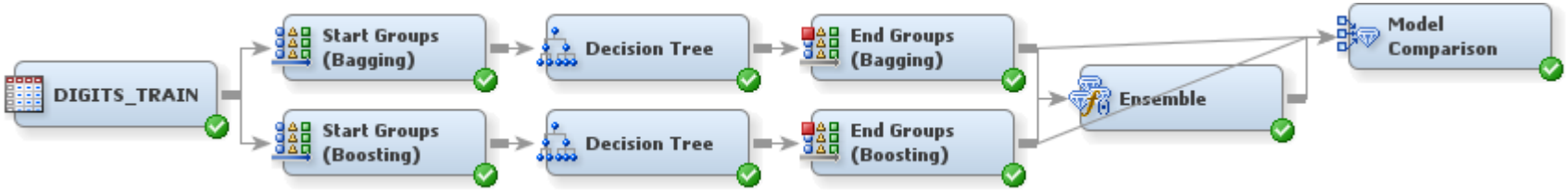
Super Learners

Combine cross-validated results of models from different algorithms

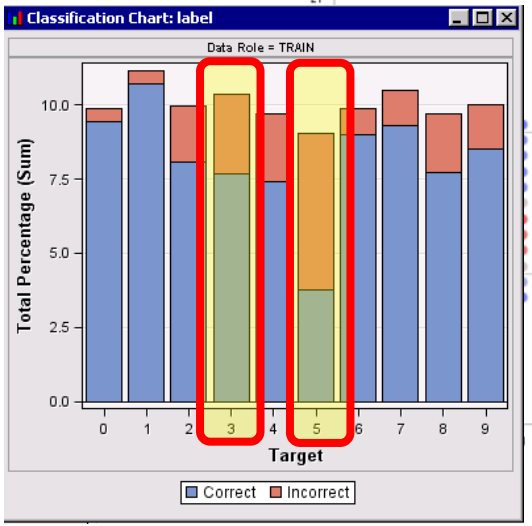


ENSEMBLE MODELING
FOR MACHINE LEARNING

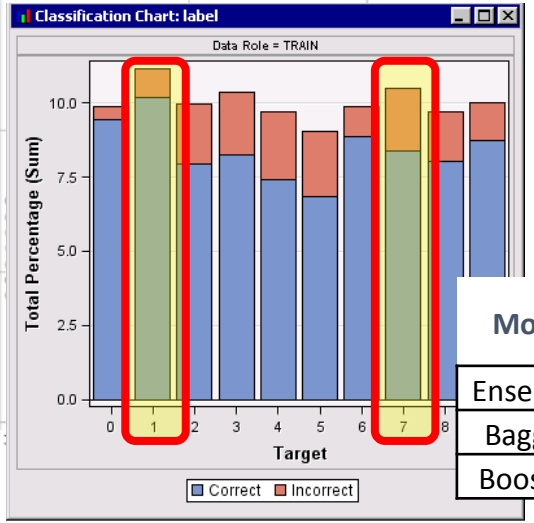
IMPROVED FIT STATISTICS



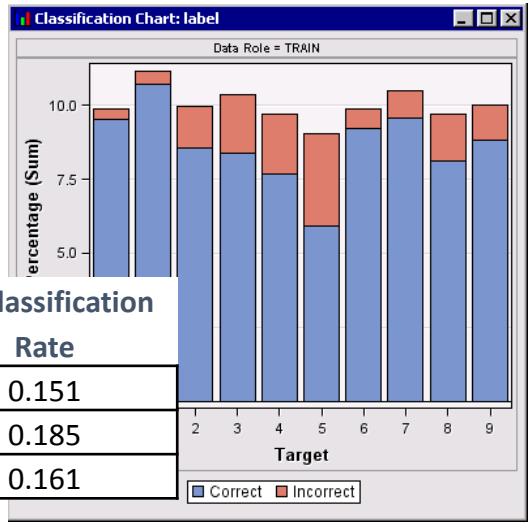
Bagging



Boosting



Ensemble



Model	Misclassification Rate
Ensemble	0.151
Bagging	0.185
Boosting	0.161

MACHINE LEARNING RESOURCES

SAS RESOURCES

An Introduction to Machine Learning

http://blogs.sas.com/content/sascom/2015/08/11/an-introduction-to-machine-learning/

SAS Data Mining Community

https://communities.sas.com/data-mining

SAS University Edition (FREE!)

http://www.sas.com/en_us/software/university-edition.html

Products Page

http://www.sas.com/en_us/insights/analytics/machine-learning.html

"Overview of Machine Learning with SAS Enterprise Miner"

http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf
http://support.sas.com/rnd/papers/sasgf14/313_2014.zip

GitHub

https://github.com/sassoftware/enlighten-apply
https://github.com/sassoftware/enlighten-deep

MISCELLANEOUS RESOURCES

- **"Big Data, Data Mining, and Machine Learning"**
  http://www.sas.com/store/prodBK_66081_en.html
- **Cloudera data science study materials**
  http://www.cloudera.com/content/dev-center/en/home/developer-admin-resources/new-to-data-science.html
  http://cloudera.com/content/cloudera/en/training/certification/ccp-ds/essentials/prep.html
- **Kaggle data mining competitions**
  http://www.kaggle.com/
- **Python machine learning packages**
  OpenCV: http://opencv.org/
  Pandas: http://pandas.pydata.org/
  Scikit-Learn: http://scikit-learn.org/stable/
  Theano: http://deeplearning.net/software/theano/
- **R machine learning task view**
  http://cran.r-project.org/web/views/MachineLearning.html
- **Quora: list of data mining and machine learning papers**
  http://www.quora.com/Data-Mining/What-are-the-must-read-papers-on-data-mining-and-machine-learning

THANK YOU