Summary

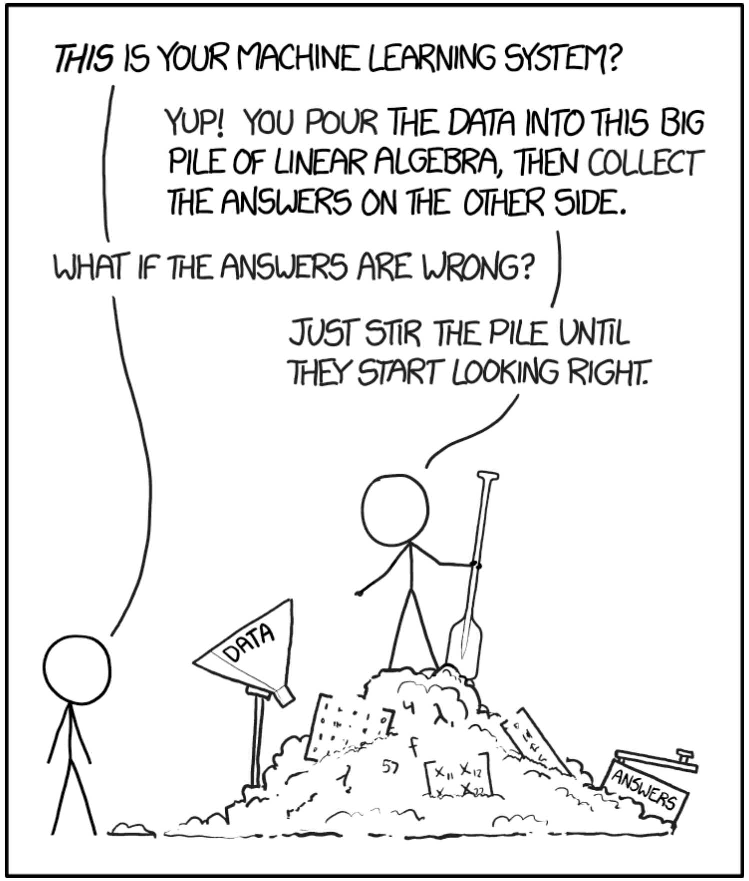
Pascal Fua
IC-CVLab

EPFL

What is Machine Learning?

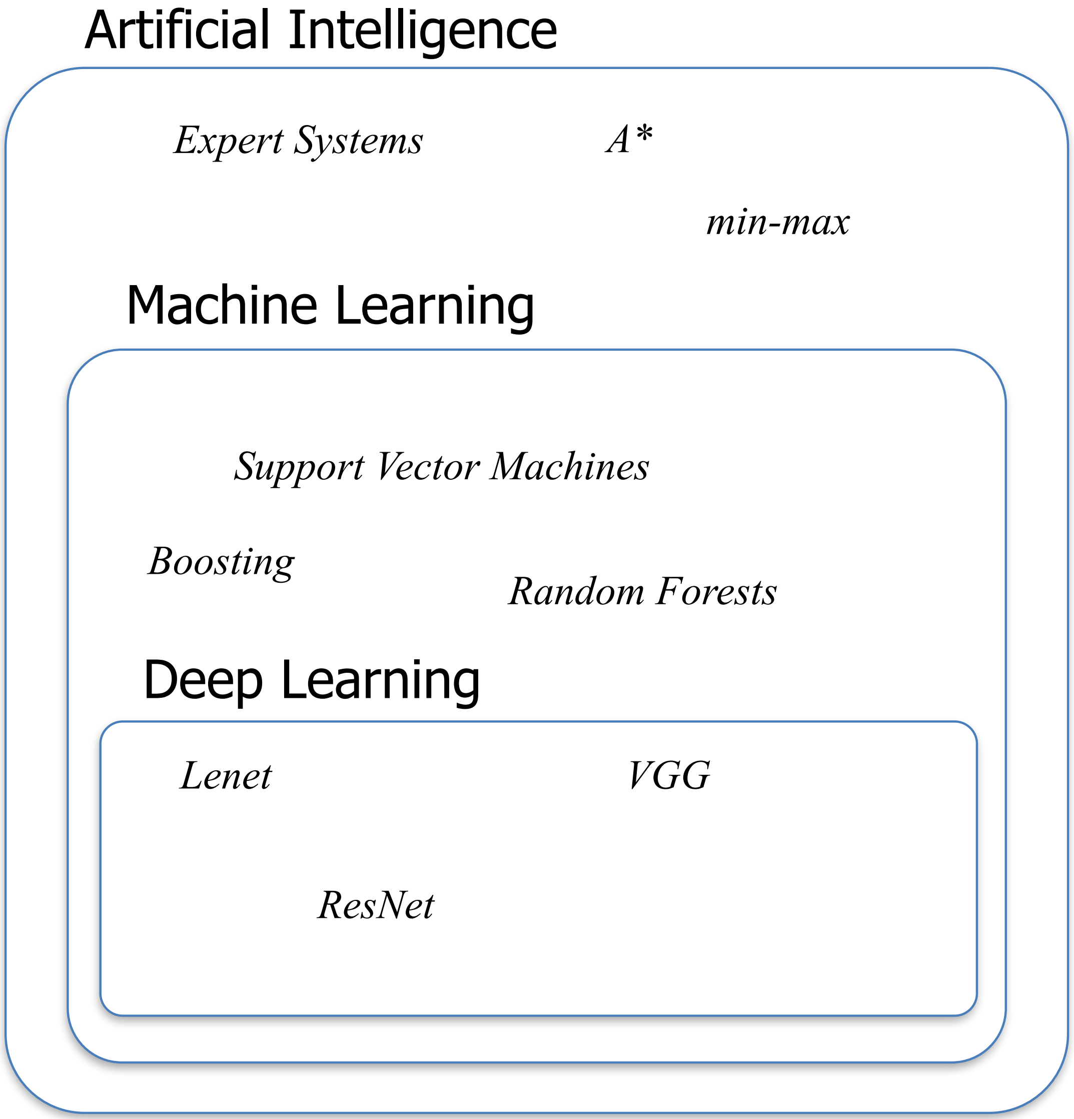
Machine learning algorithms:

- seek to provide knowledge to computers through data, observations, and interaction with the world,
- can make predictions given new observations,
- improve when given access to large amounts of training data.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/

Artificial Intelligence



Artificial Intelligence

Expert Systems

*A**

min-max

Machine Learning

Support Vector Machines

Boosting

Random Forests

Deep Learning

Lenet

VGG

ResNet

1997: Deep blue beats chess world champion

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
PAGE 462
NATURE.COM/NATURE

2017: AlphaGo beats go world champion

Self-Driving Cars

1985

DARPA ALV

2007

DARPA Urban Challenge

2014

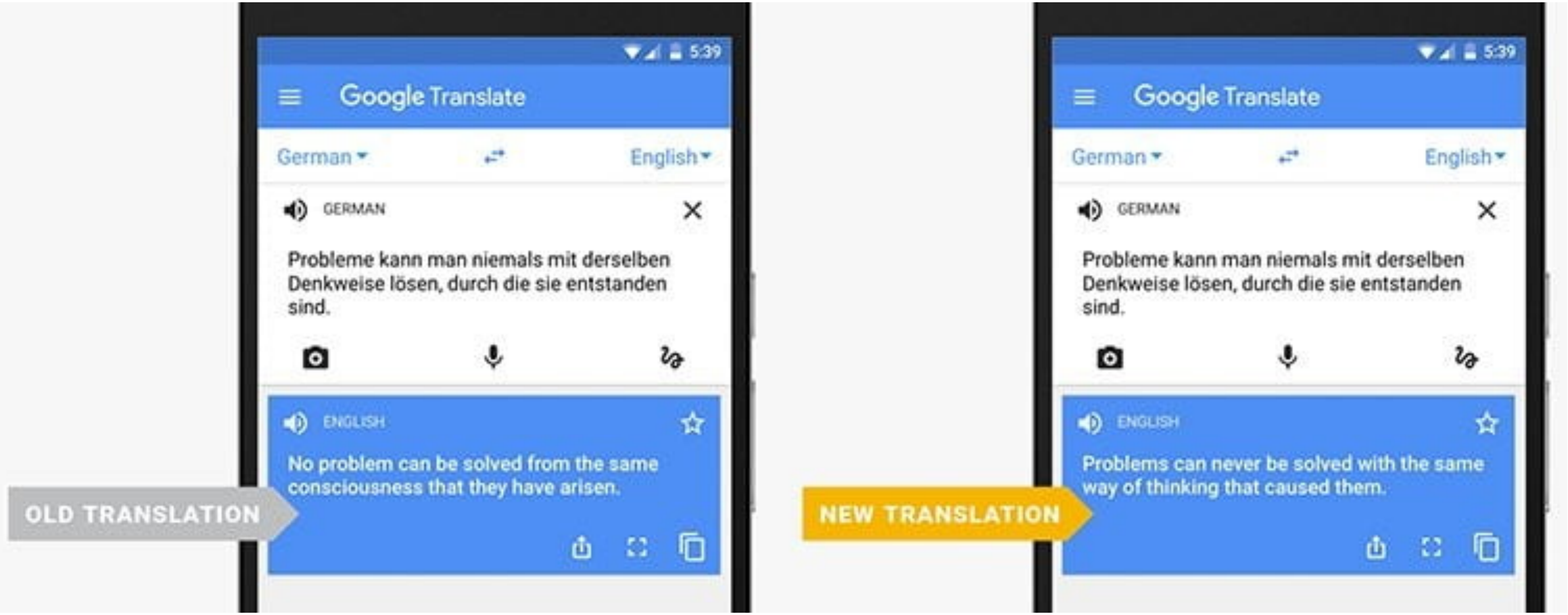
Google Cars

2016

Tesla

- More computing power.
- Better sensors.
- Detailed maps of the environment.
- Machine learning

Machine Translation



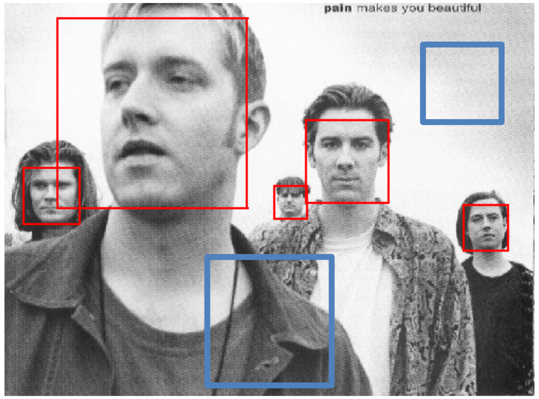
Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind.

No problem can be solved from the same consciousness that they have arisen.

Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind.

Problems can never be solved with the same way of thinking that caused them.

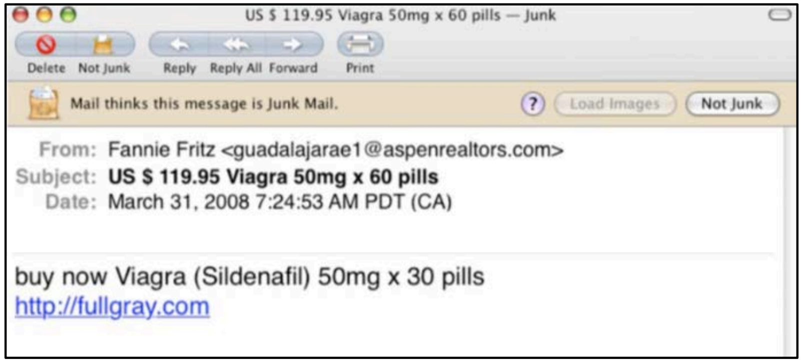
Classification vs Regression



□ Face

□ Not Face

Spam or not

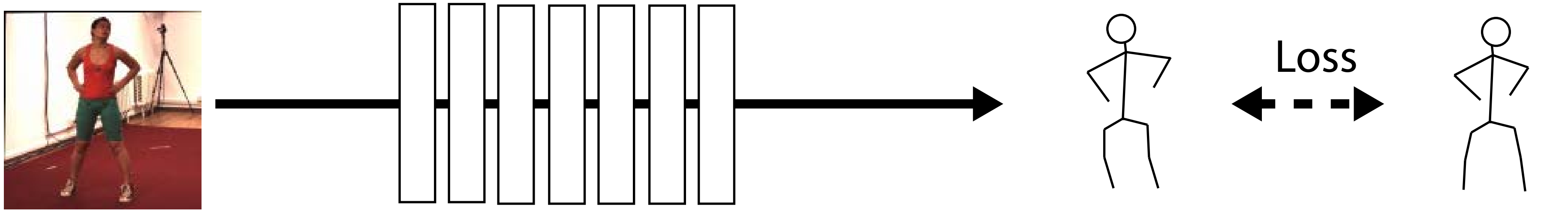


Mail thinks this message is Junk Mail.

From: Fannie Fritz <guadalajarae1@aspenrealtors.com>
Subject: **US $ 119.95 Viagra 50mg x 60 pills**
Date: March 31, 2008 7:24:53 AM PDT (CA)

buy now Viagra (Sildenafil) 50mg x 30 pills
http://fullgray.com

- Classification algorithms seek to estimate a mapping function y from the input vector **x** to discrete or categorical output variables.



Loss

Input image Deep CNN Output pose GT pose

- Regression algorithms seek to estimate the mapping function y from the input vector **x** to **numerical or continuous** output variables.

Supervised Classification

Minimize:

Predicted label for sample n

$$E(\mathbf{w}) = \sum_{n=1}^{N} L(y(\mathbf{x}_n; \mathbf{w}), t_n)$$

True label for sample n

- **x:** Feature vector
- **w:**Model parameters
- t: Label
- y: Predictor
- L: Loss Function
- E: Error Function

—> ML is an optimization problem

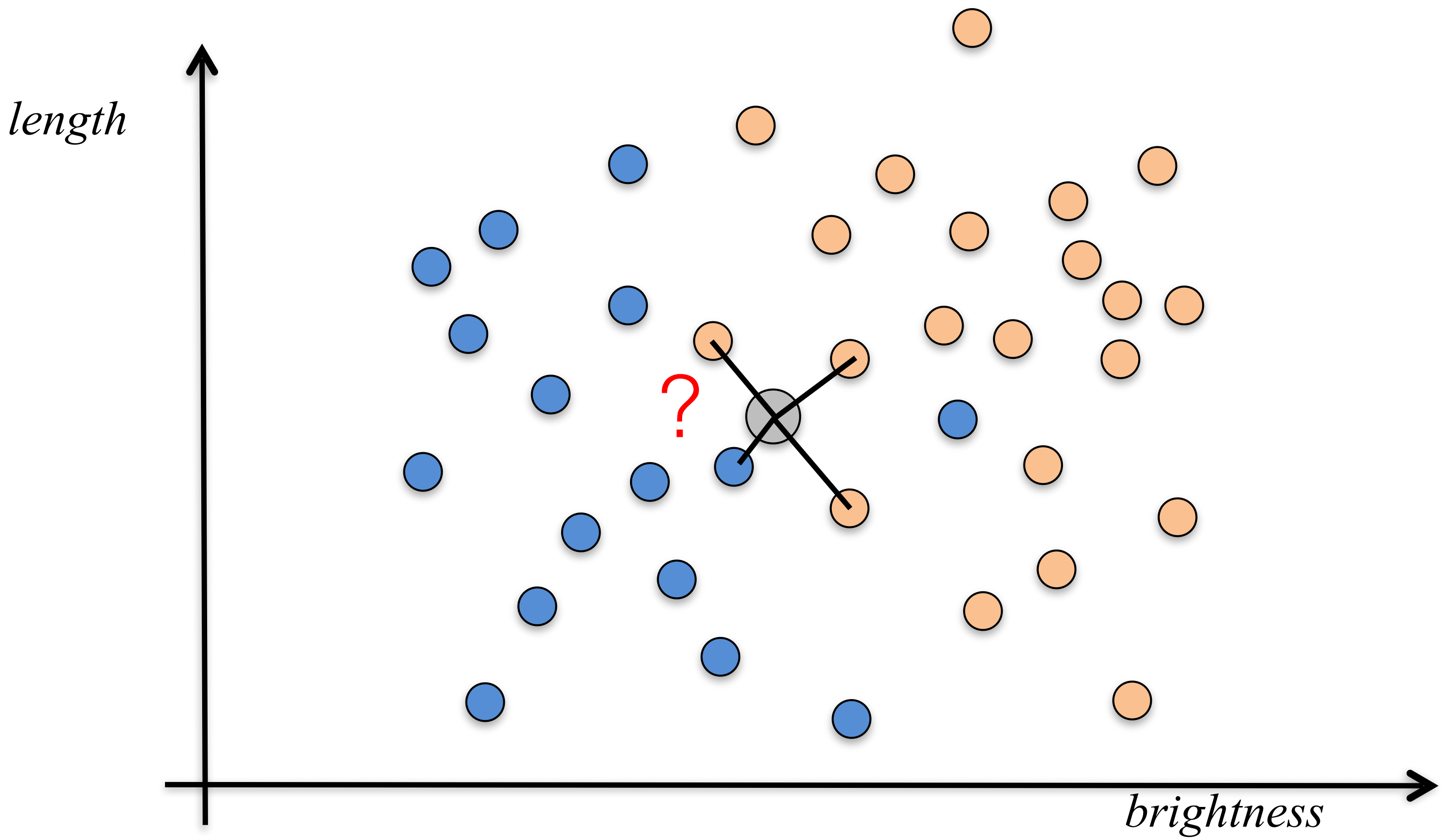
Classification Techniques

- Nearest-Neighbors and K-Means
- Logistic Regression
- Boosting
- Support Vector Machines
- Decision Trees and Forests
- Multilayer Perceptrons
- Convolutional Neural Networks
- Tranformers

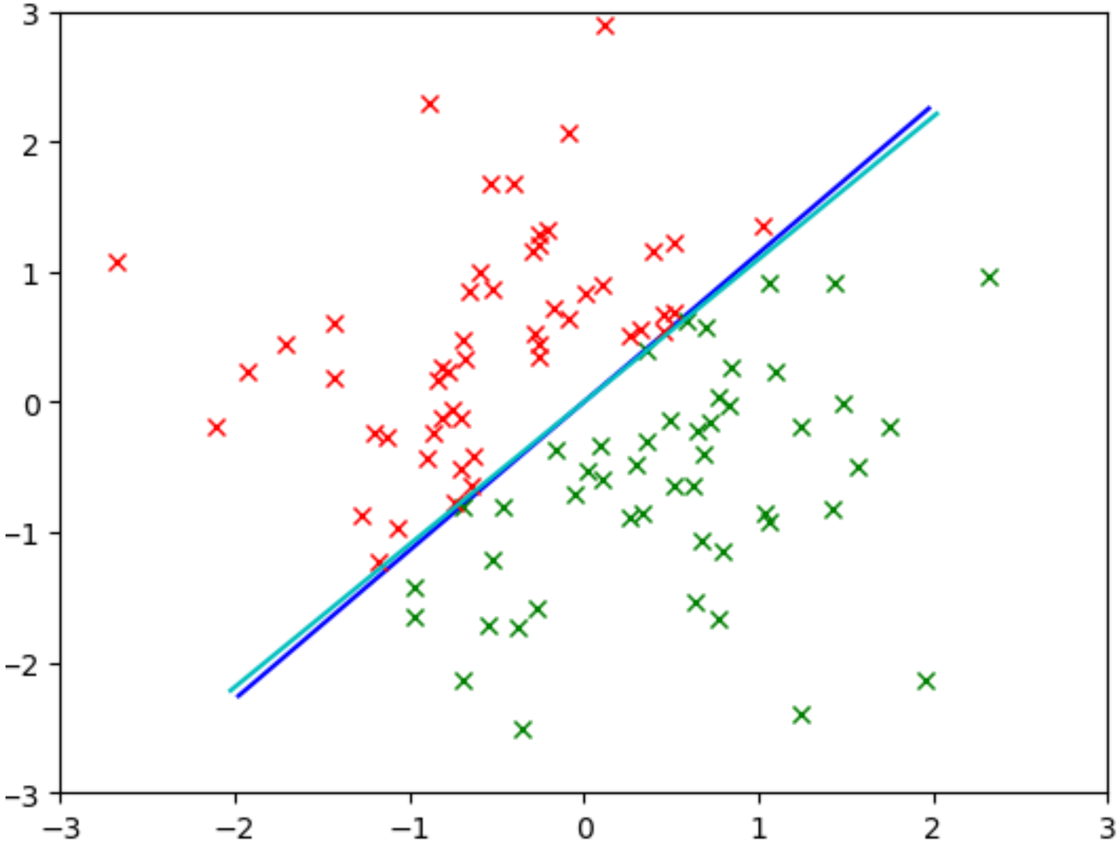
K-Nearest-Neighbor Classifier

Improved algorithm:

- Given a new **x** to be classified, find its k nearest neighbors in the training set.
- Classify the point according to the majority of labels of its nearest neighbors.



Perceptron

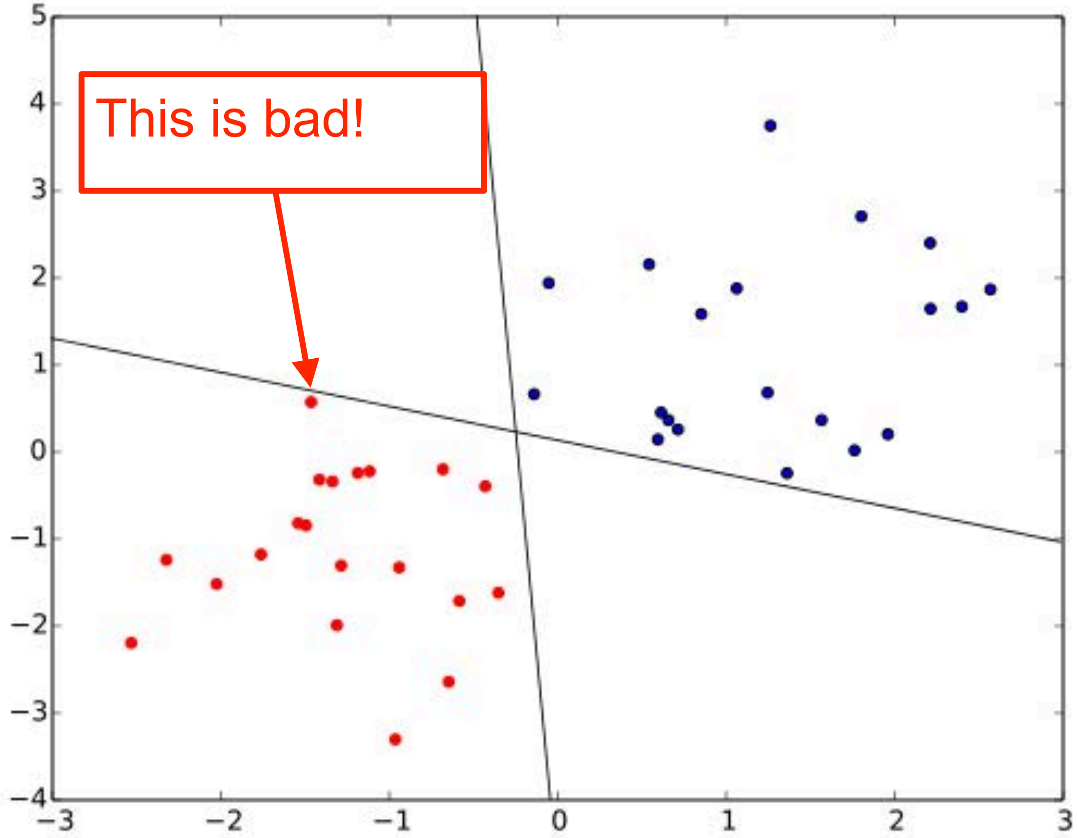


$$y(\mathbf{x}; \mathbf{w}, w_0) = \begin{cases} 1 & \text{if} \quad \mathbf{w}^T\mathbf{x} + w_0 \geq 0 \, , \\ -1 & \text{otherwise.} \end{cases}$$

Given the training set $\{(x_n, t_n)_{1 \leq n \leq N}\}$, choose a $\mathbf{w}$ that minimizes

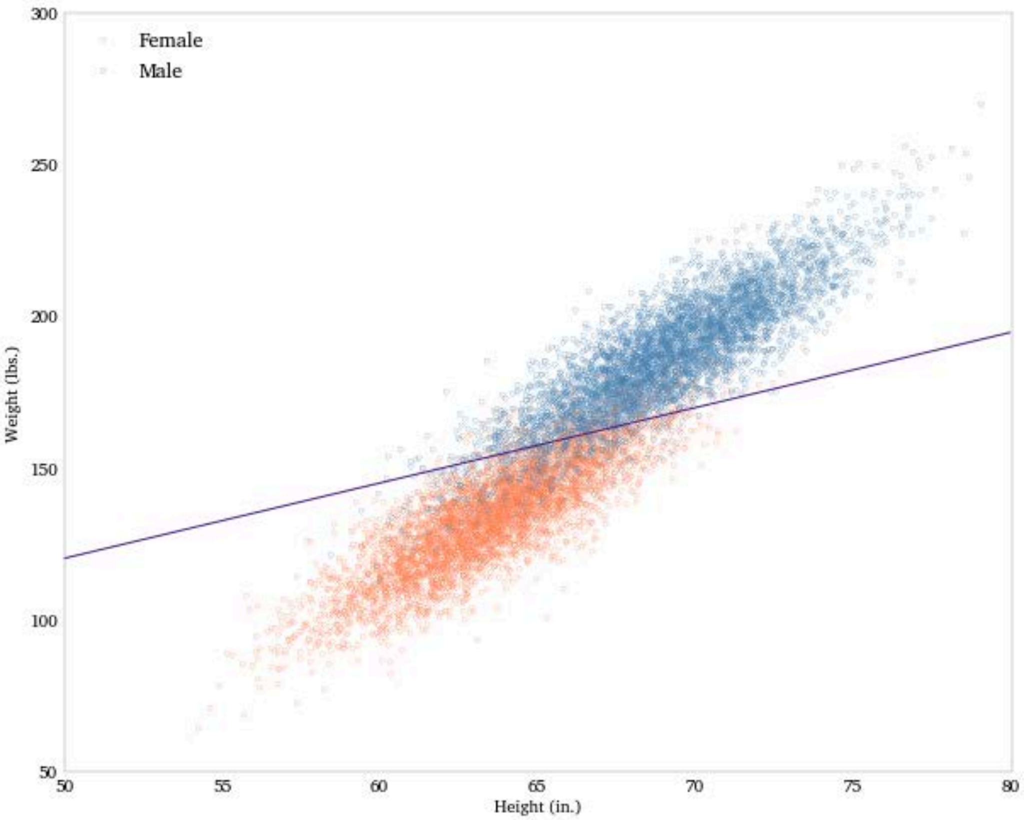
$$E(\mathbf{w}, w_0) = -\sum_{n=1}^{N} (\mathbf{w}^T\mathbf{x}_n + w_0)t_n$$

The Problem with the Perceptron

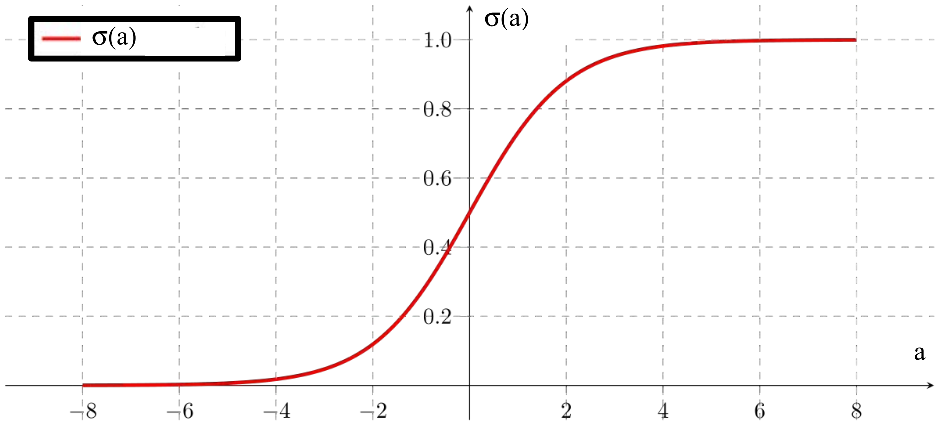


- Two different solutions among infinitely many.
- The perceptron has no way to favor one over the other.

Logistic Regression



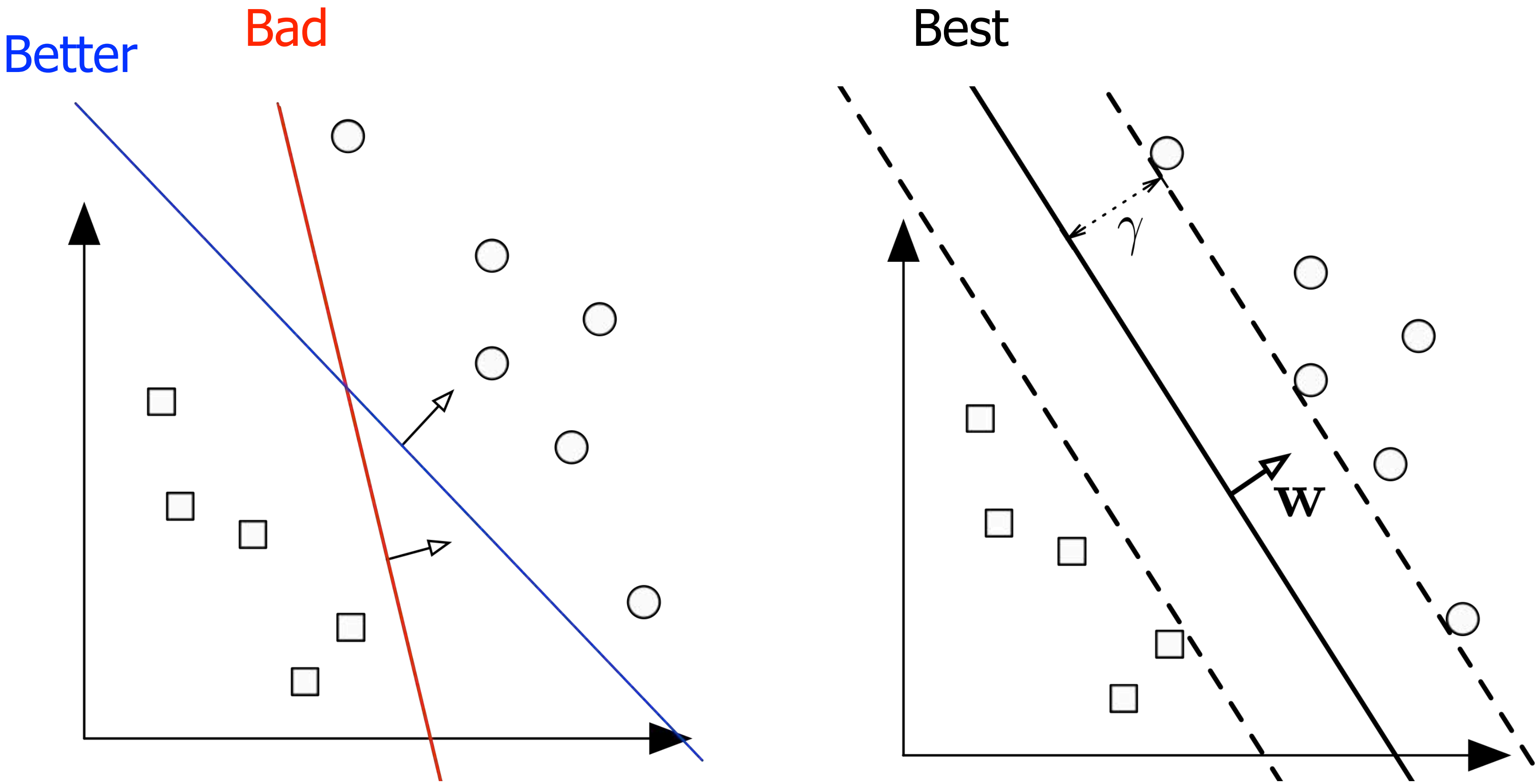
$$y(\mathbf{x}; \mathbf{w}, w_0) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$$
$$\approx p(t = 1, \mathbf{x})$$



Given the training set $\{(x_n, t_n)_{1 \leq n \leq N}\}$, choose a $\mathbf{w}$ that minimizes

$$E(\mathbf{w}, w_0) = -\sum_n \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \approx -\ln(p(\mathbf{t}|\mathbf{w}, w_0)) \ .$$

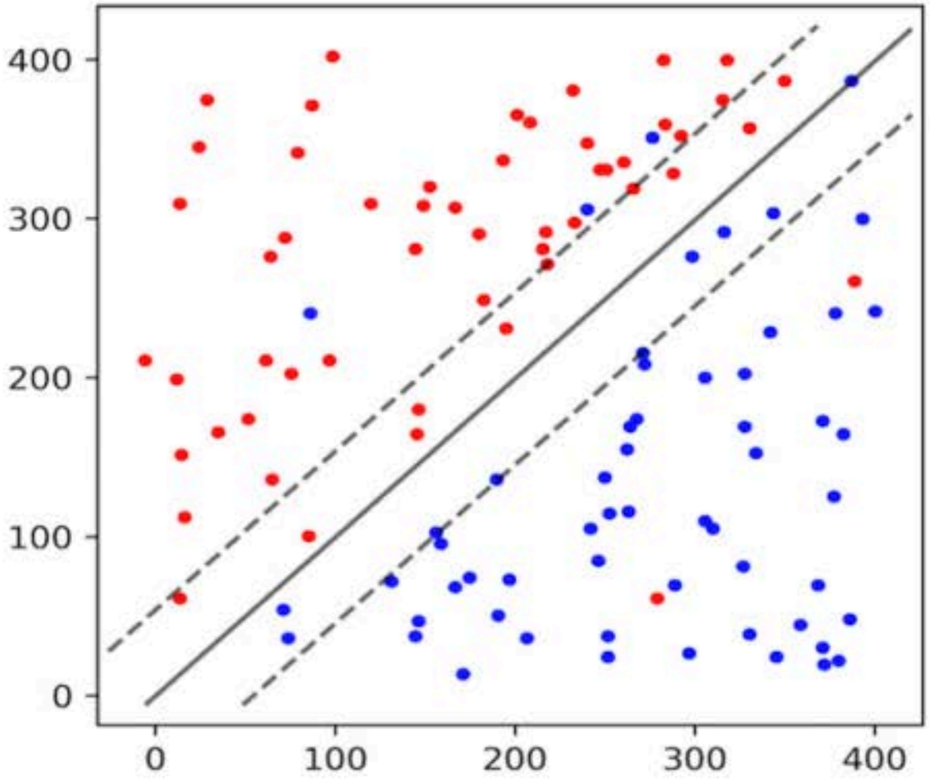
Maximizing the Margin



- The larger the margin, the better!
- In the presence of outliers, the logistic regression does not guarantee a large one.

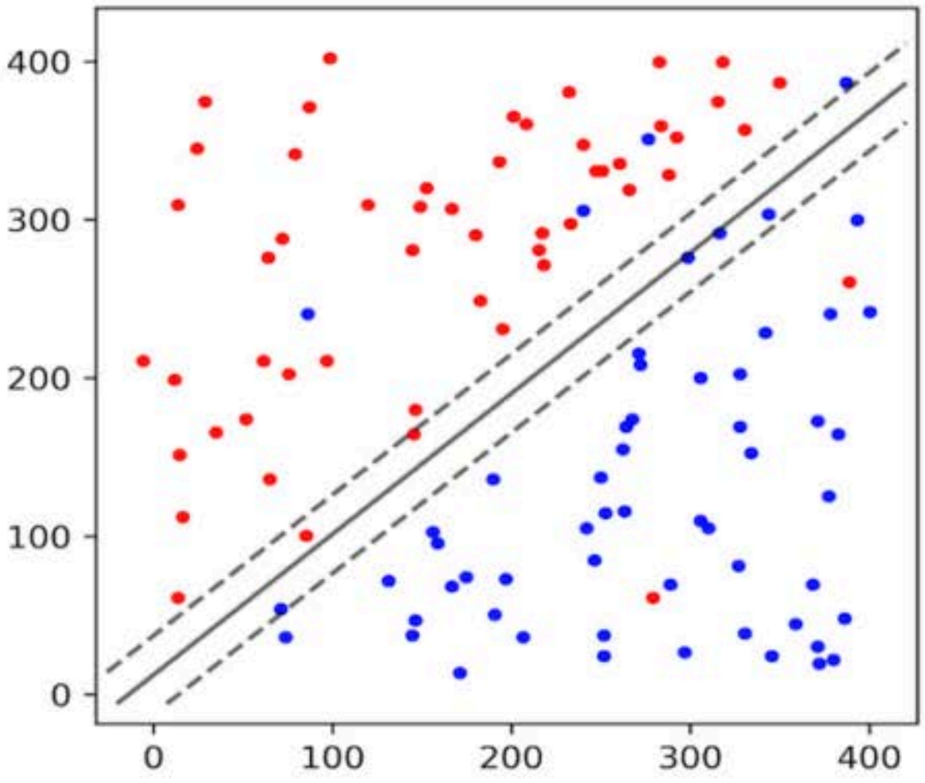
How do we maximize it?

Max Margin Classifier



C=1:

- Large margin.
- Many training samples misclassified.



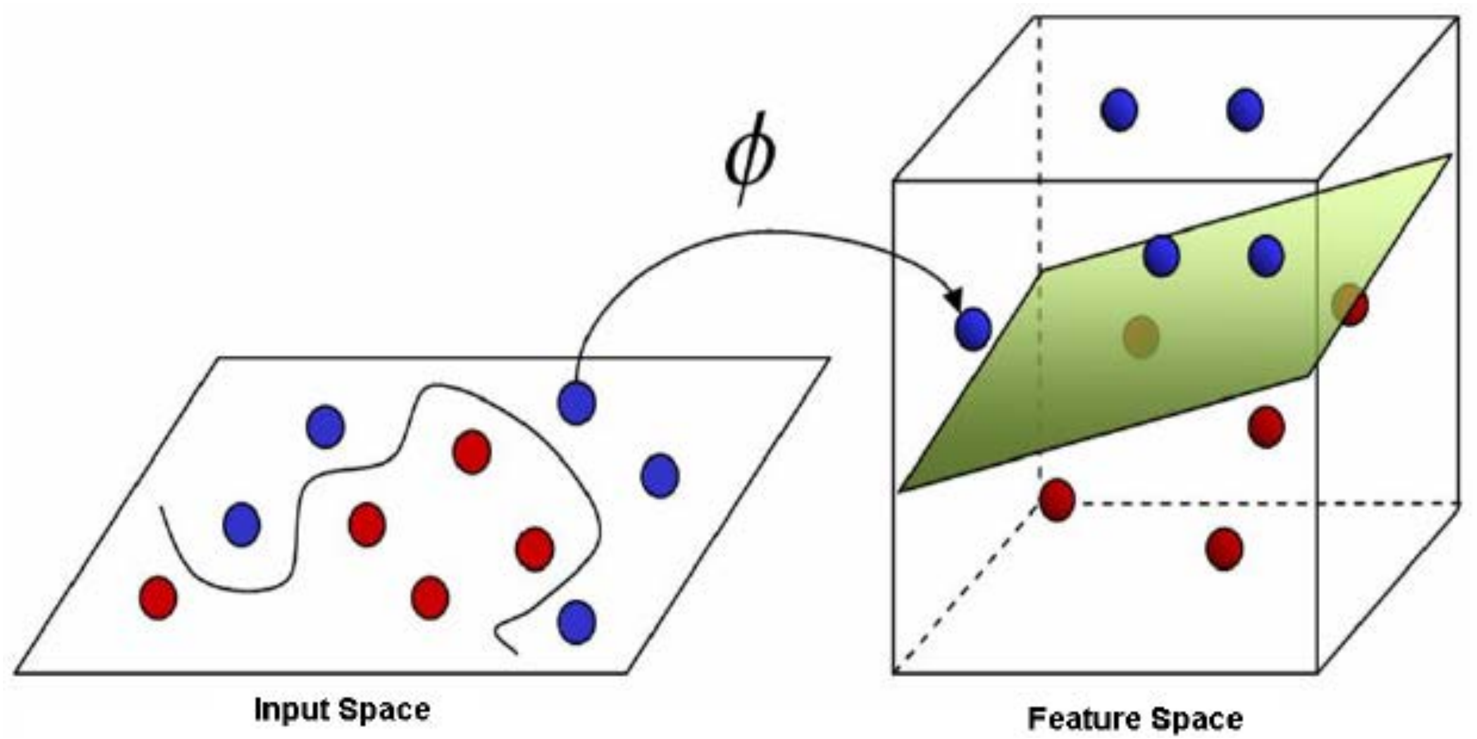
C=100:

- Small margin.
- Few training samples misclassified.

Which is best?

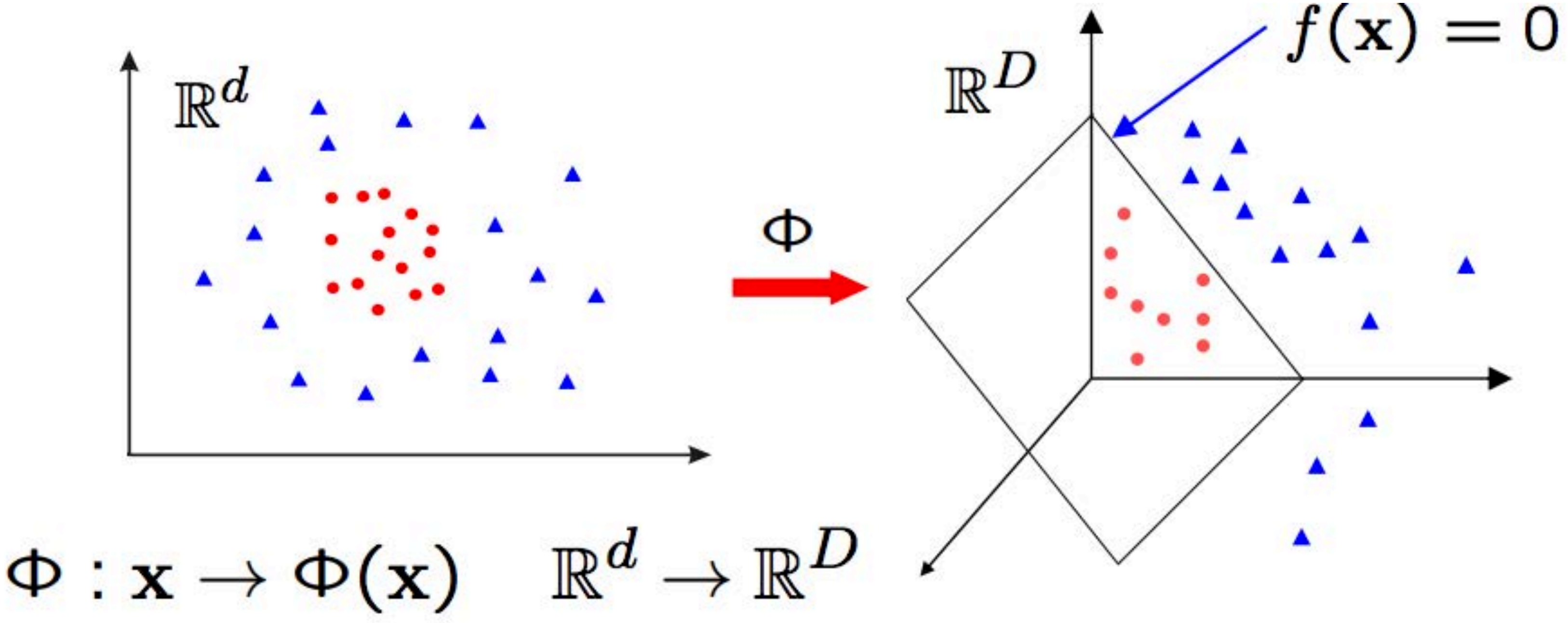
- It depends.
- Must use cross-validation, as we did for k-Means.

Non Linearly Separable Data



- Map to a higher dimensional space in which it is.
- Use an ensemble of classifiers.
- Use a deep network.

Classification in Feature Space


$\mathbb{R}^d$
Φ
$\mathbb{R}^D$
$f(\mathbf{x}) = 0$
$\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x}) \quad \mathbb{R}^d \rightarrow \mathbb{R}^D$

- Map from R^d to R^D
- Learn a linear classifier in R^D

$$\mathrm{y}\left(\mathbf{x}\right) = \sigma\left(\mathbf{w}^T \phi(\mathbf{x}) + w_0\right)$$

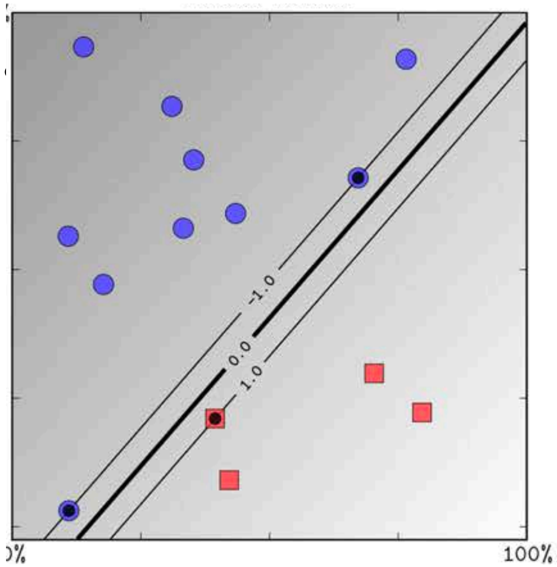
$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$$

Polynomial SVMs

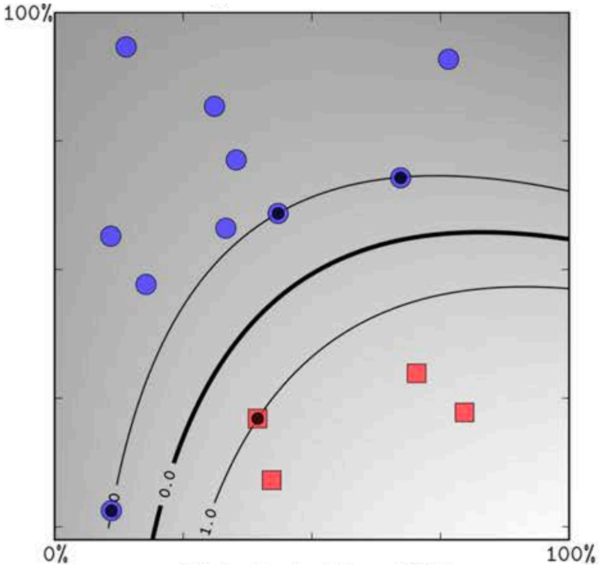
$$\mathbf{w}^* = min_{(\mathbf{w},\{\xi_\mathbf{n}\})}\frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n,$$

$$\text{subject to } \forall n, \quad t_n \cdot (\tilde{\mathbf{w}} \cdot \phi(\mathbf{x}_n)) \geq 1 - \xi_n \text{ and } \xi_n \geq 0.$$

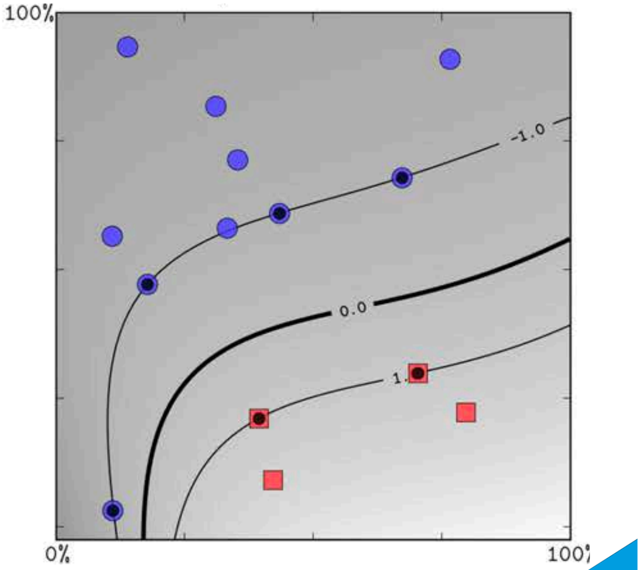
- C is constant that controls how costly constraint violations are.



M =1

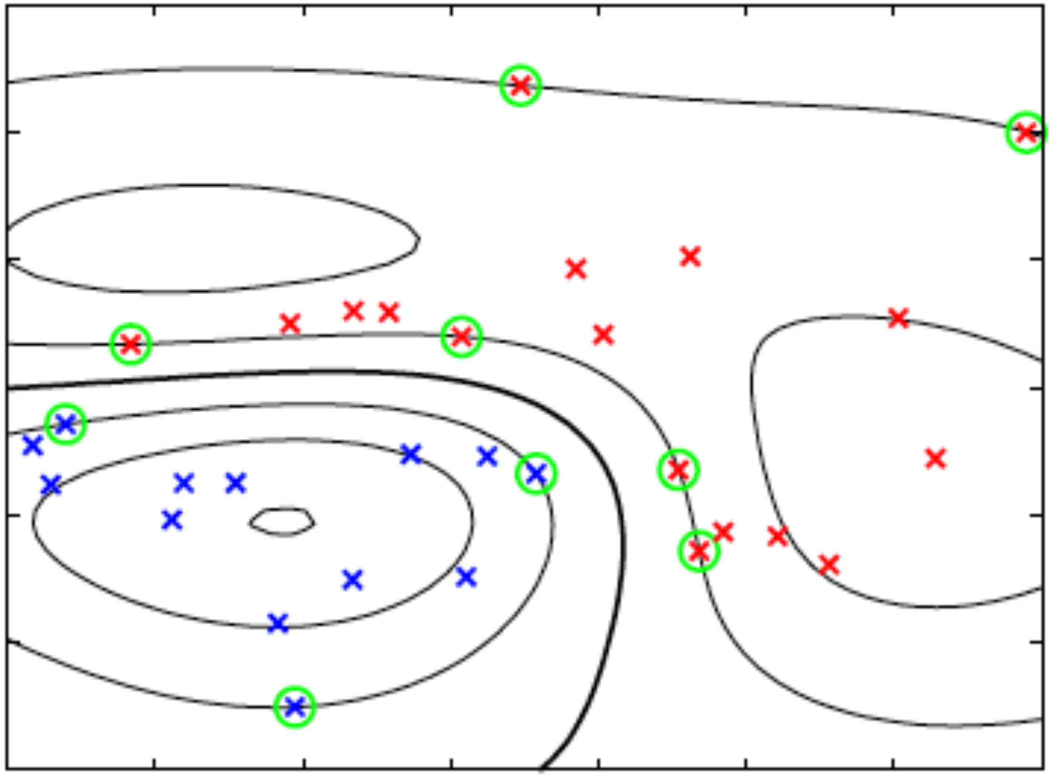


M = 2



M = 5

Kernel SVMs



$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \; .$$

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \; ,$$

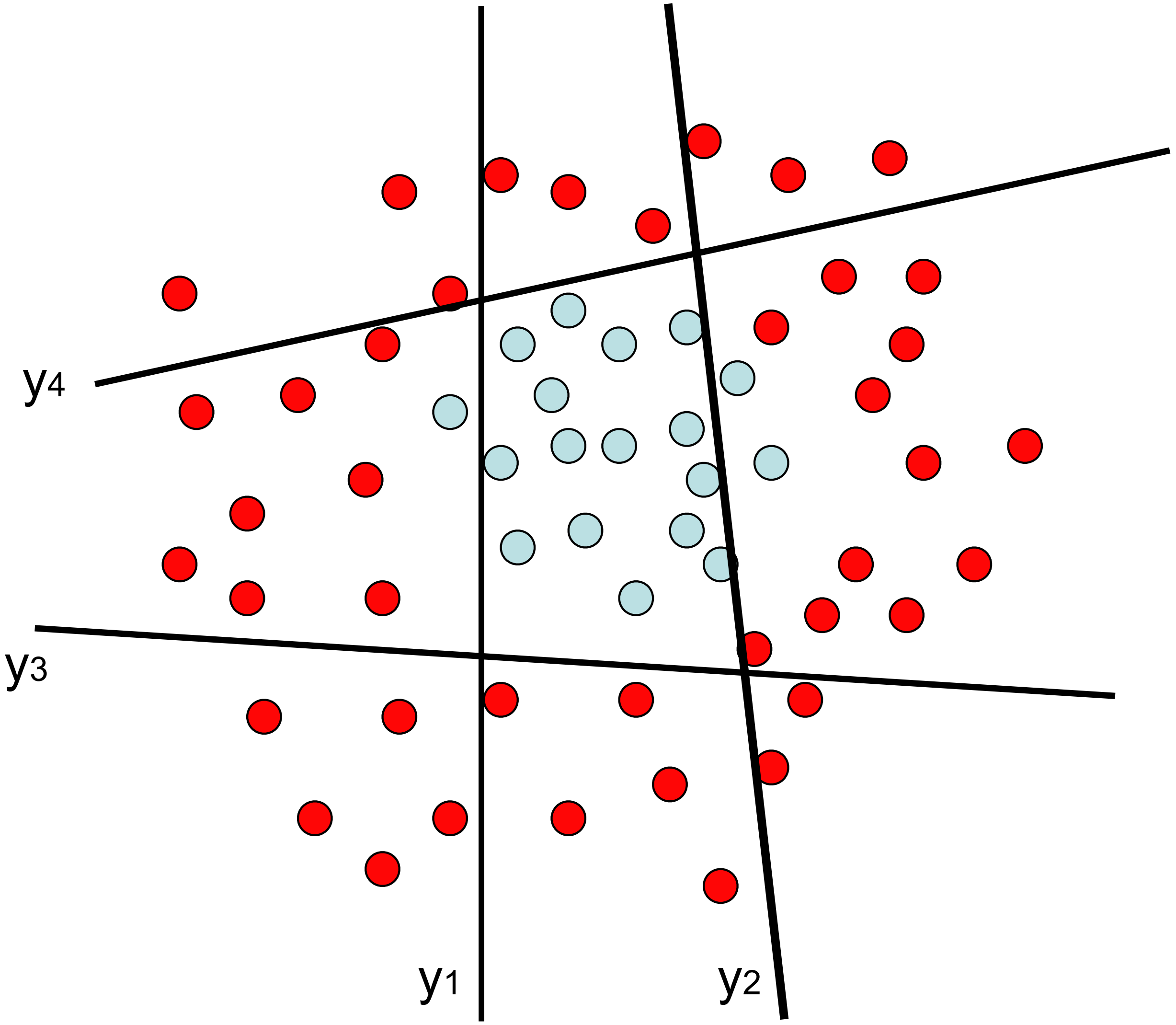
$$= \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \; ,$$

with $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') \; .$

- Only for a subset of the data points is a_n is non zero.
- The corresponding x_n are the support vectors and satisfy $t_n y(x_n)=1$.
- They are the only ones that need to be considered as test time.

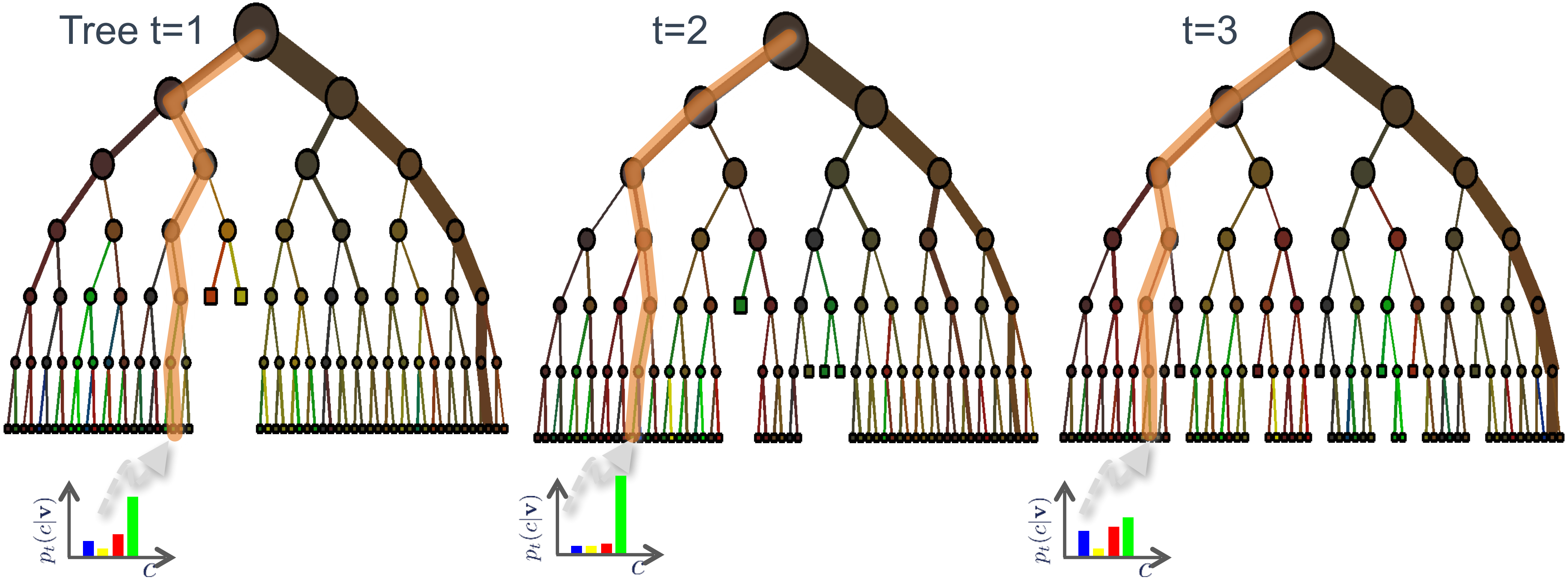
—> That is what makes SVMs practical!

AdaBoost

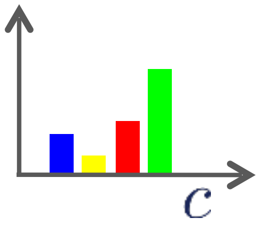


$$y(\mathbf{x}) = \alpha_1 y_1(\mathbf{x}) + \alpha_2 y_2(\mathbf{x}) + \alpha_3 y_3(\mathbf{x}) + \alpha_4 y_4(\mathbf{x})$$

Decision Forests

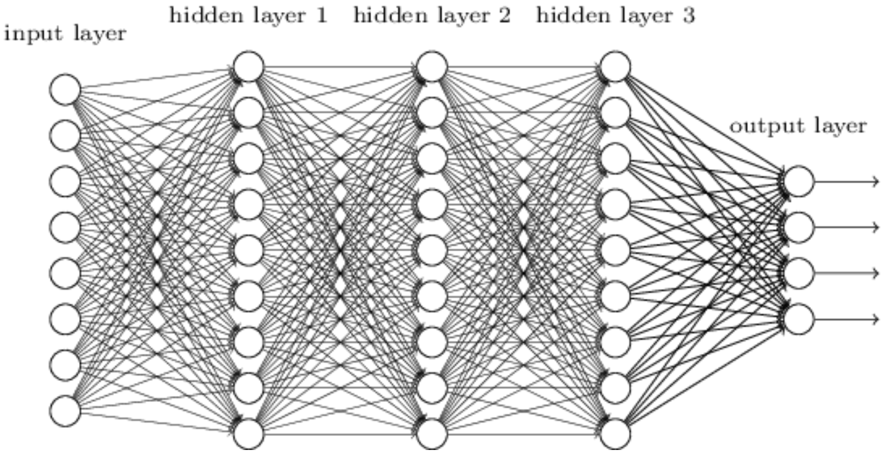


$$L(c,\mathbf{v}) = \frac{1}{T}\sum_t -\log(p_t(c|\mathbf{v}))$$

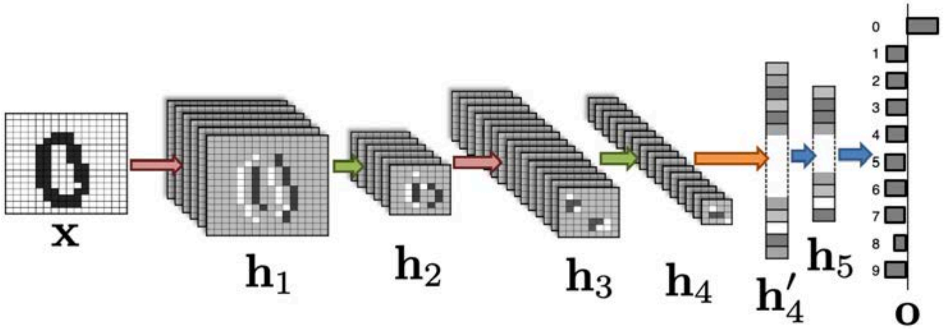


Simple and flexible approach.

Neural Networks



Fully connected



Convolutional

- Some of the most powerful current techniques around when enough training data is available.
- Convolutional Neural Nets are particularly well adapted for image processing.

Regression Techniques

- Linear Regression
- Polynomial Regression
- Neural Networks

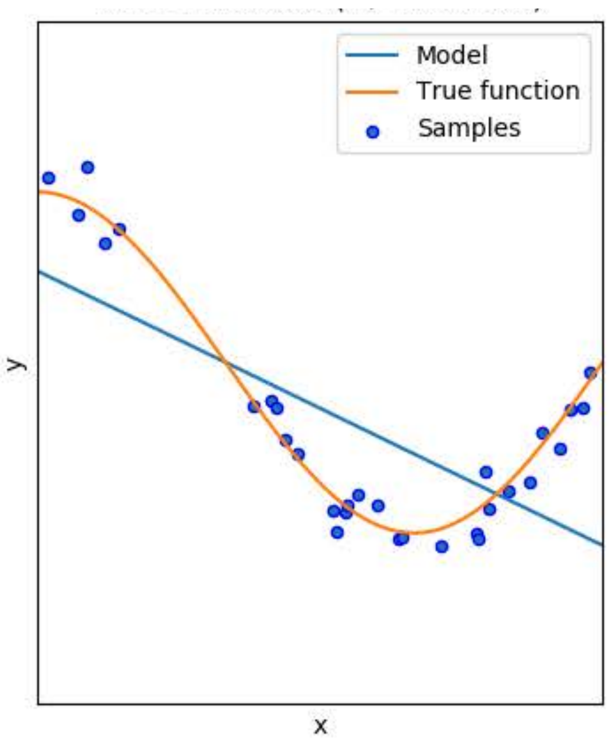
Supervised Regression

$$\text{Minimize } E(\mathbf{w}) = \sum_{n=1}^{N} L(y(\mathbf{x}_n; \mathbf{w}), t_n)$$

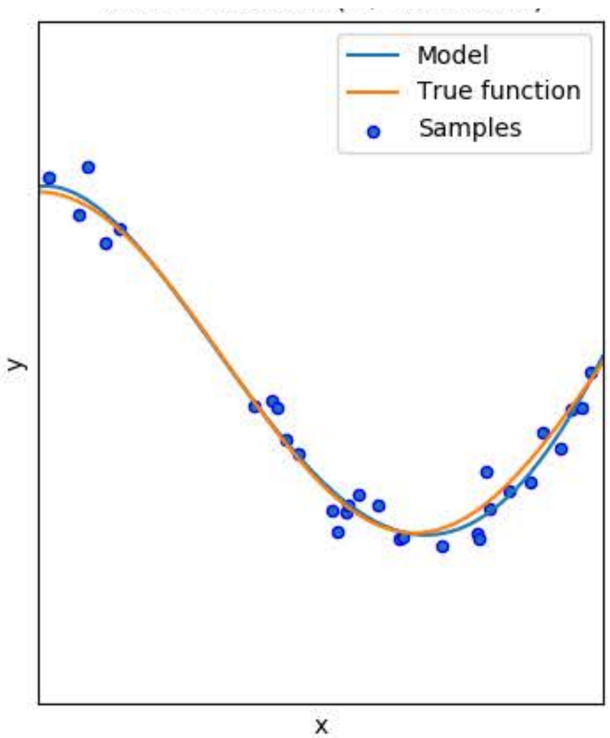
- **x:** Feature vector
- **w:**Model parameters
- t: Label
- y: Predictor
- L: Loss Function
- E: Error Function

Same as for classification, except for the fact that the t_n now denotes continuous values!

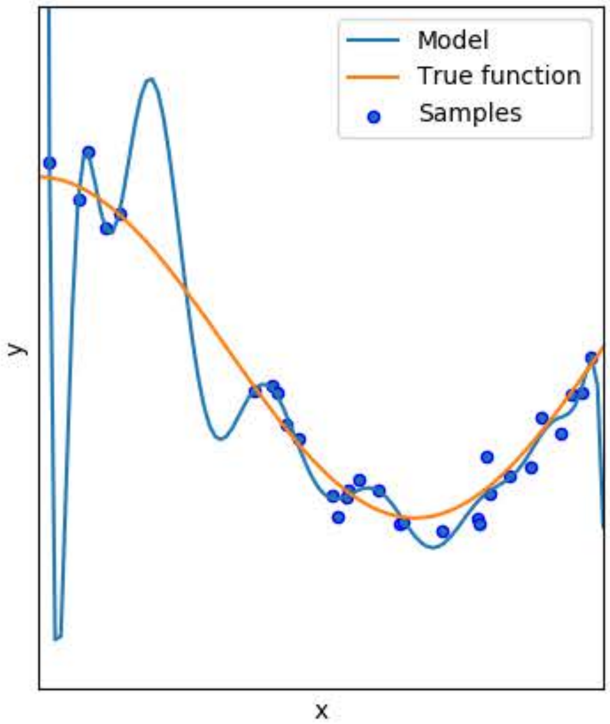
Linear and Non-Linear Regression



Order 1



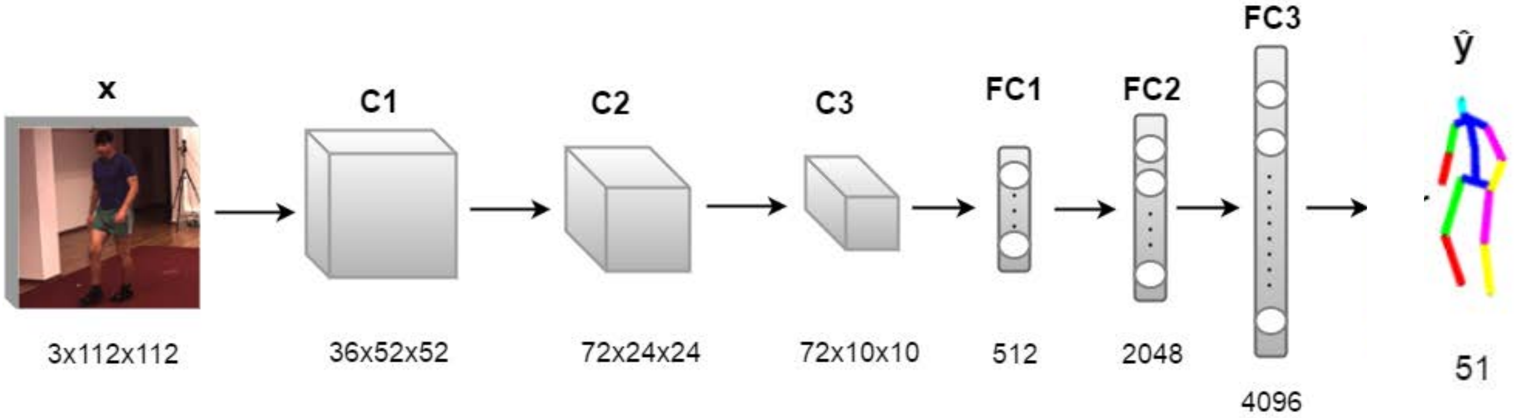
Order 4



Order 15

The trick is to find the best compromise between simplicity and goodness of fit.

Deep Networks



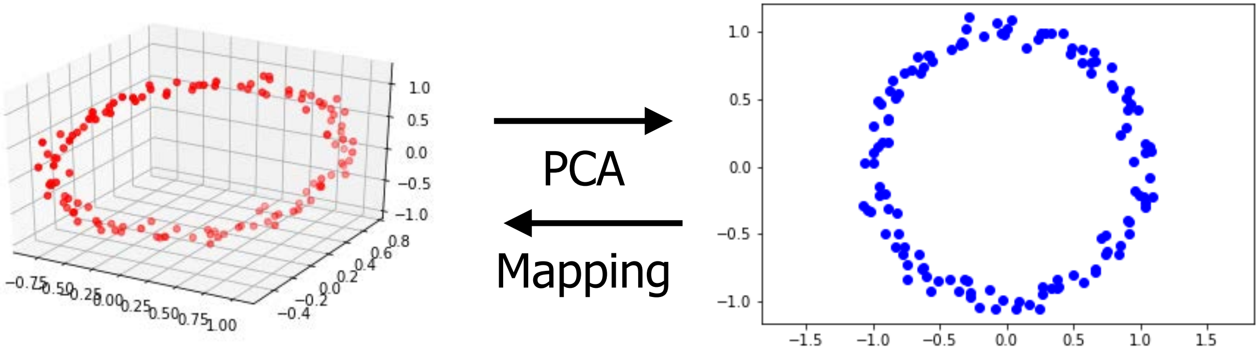
Input: $\mathbf{I}$

Output: $\{\mathbf{y}_j\}_{1 \leq j \leq J}$

Dimensionality Reduction Techniques

- PCA
- LDA
- Autoencoders

PCA



PCA

Mapping

$\mathbf{x}$

$\mathbf{y} = \mathbf{W}^{\mathbf{T}}(\mathbf{x} - \bar{\mathbf{x}})$

$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{W}\mathbf{y}$

- This mapping incurs some loss of information.
- However, the corresponding rectangular matrix **W** is the orthogonal matrix that minimizes the reconstruction error

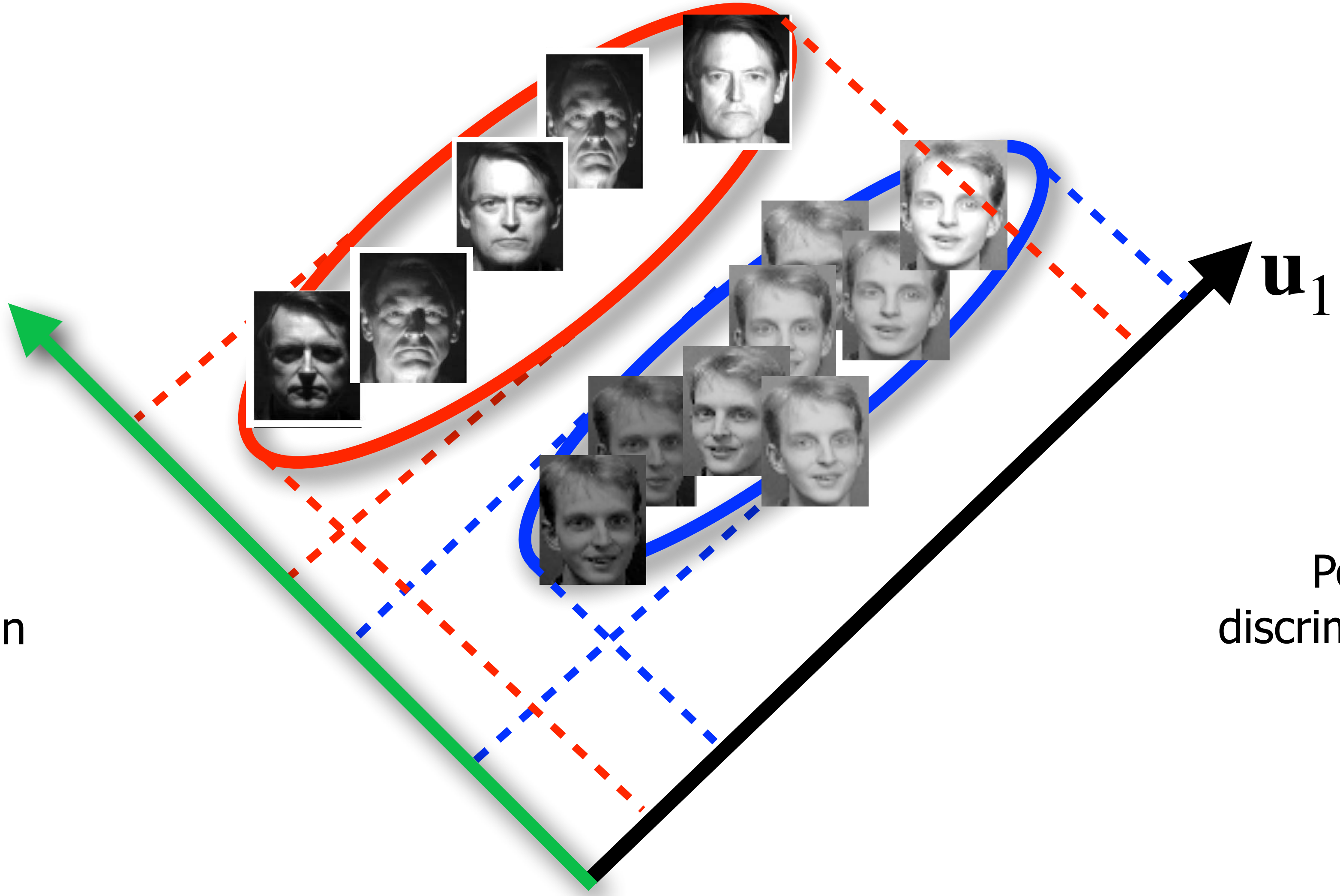
$$e = \|\hat{\mathbf{x}} - \mathbf{x}\|^2$$

where

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{W}\mathbf{y} = \bar{\mathbf{x}} + \mathbf{W}\mathbf{W}^T(\mathbf{x} - \bar{\mathbf{x}})$$

LDA

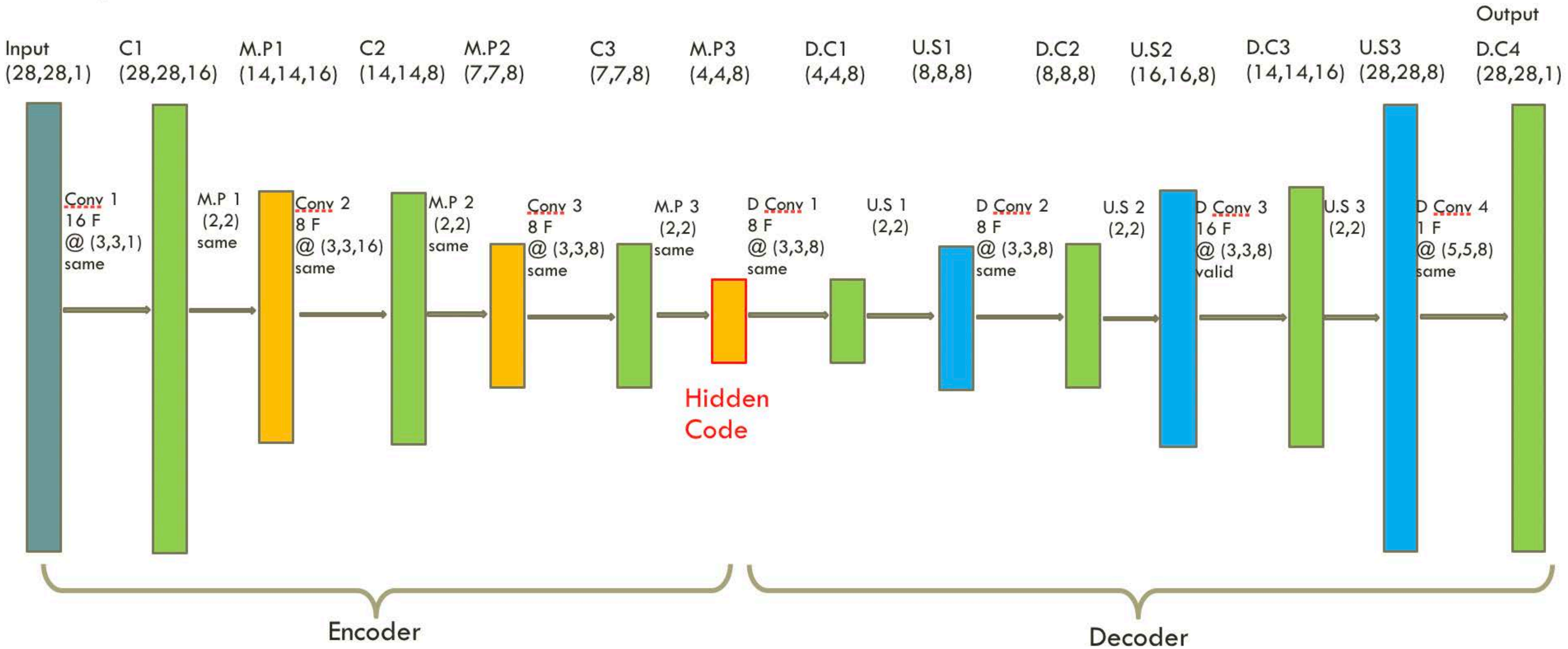
Maximise between class variance and minimize within class variance.

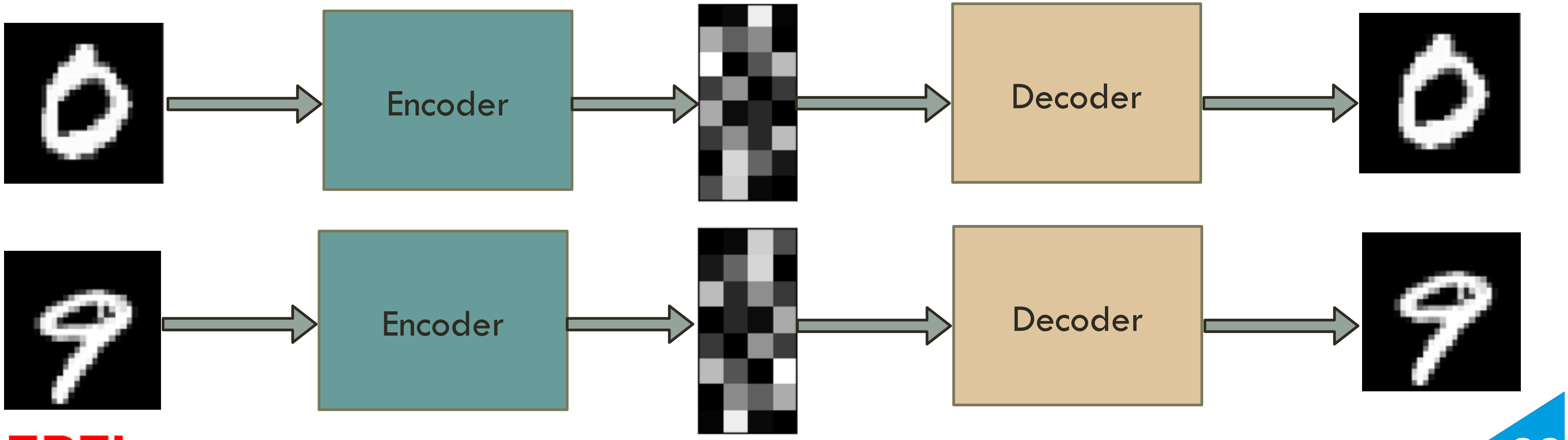


Good discrimination

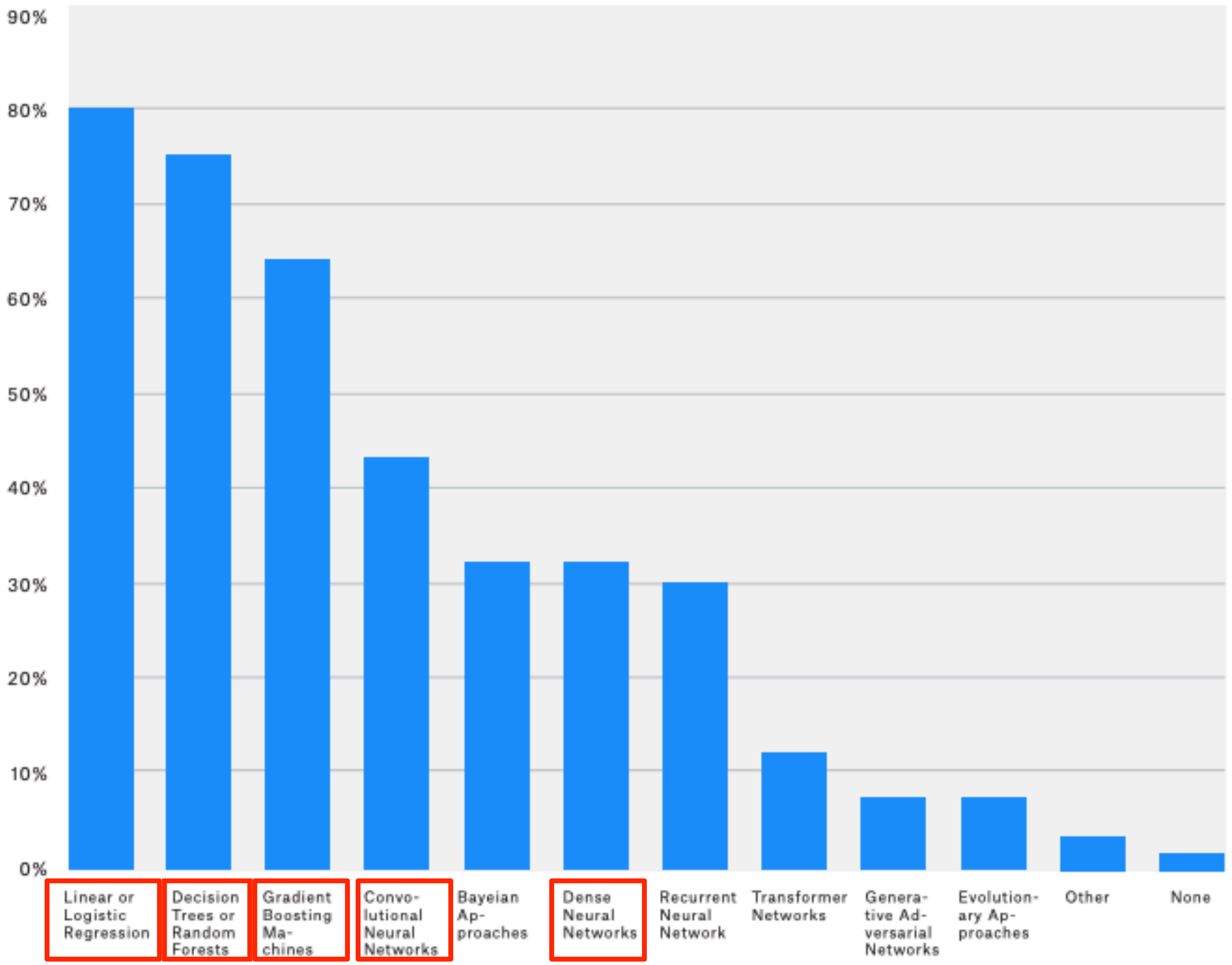
Poor discrimination

Autoencoders



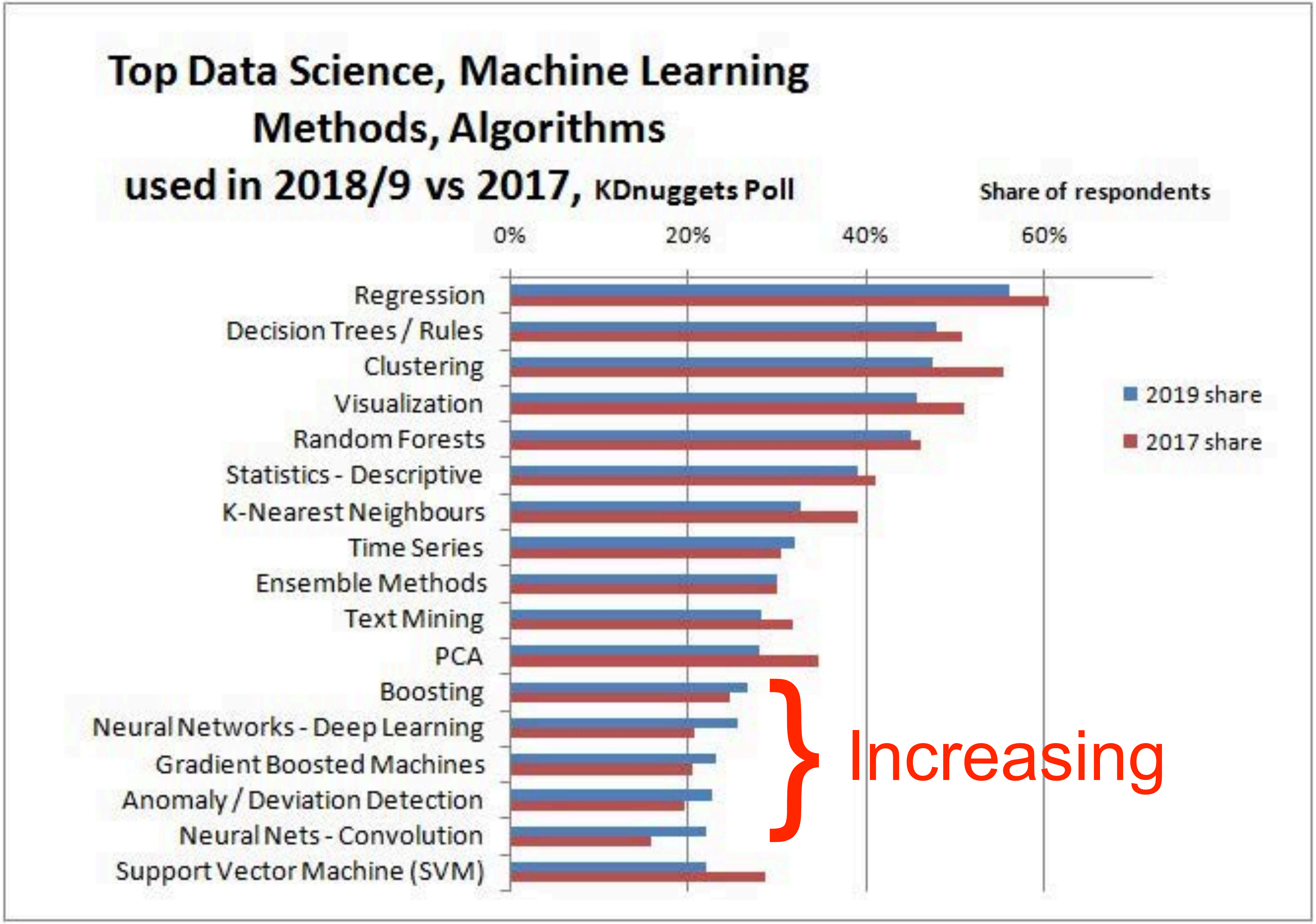


Kaggle Survey (2019)



What data science methods do you use at work?

Trends


Top Data Science, Machine Learning Methods, Algorithms used in 2018/9 vs 2017, KDnuggets Poll
Share of respondents
0%
20%
40%
60%
Regression
Decision Trees / Rules
Clustering
Visualization
Random Forests
Statistics - Descriptive
K-Nearest Neighbours
Time Series
Ensemble Methods
Text Mining
PCA
Boosting
Neural Networks - Deep Learning
Gradient Boosted Machines
Anomaly / Deviation Detection
Neural Nets - Convolution
Support Vector Machine (SVM)
2019 share
2017 share
Increasing

Logistic Regression on a Massive Scale

Ad Click Prediction at Google:

Methods such as regularized logistic regression are a natural fit for this problem setting. It is necessary to make predictions many billions of times per day and to quickly update the model as new clicks and non-clicks are observed.

—> The simpler methods are not going away and will probably co-exist with the more sophisticated ones.

McMahan et al., KDD '13

In Conclusion

Rule of thumb:

- Small training set: Use GPs.
- Medium training sets: Use boosted trees.
- Large training set: Use neural nets.

As for all such rules, there are exceptions. Real-time requirements define important ones.

Exam

- On June 27th.
- Possibility of extra-mural exam if you **cannot** come.
- 2.0 hours.
- 1 two-sided **hand-written** A4 page of notes.
- Questions on **non-indented** slides on webpage.

See you then.