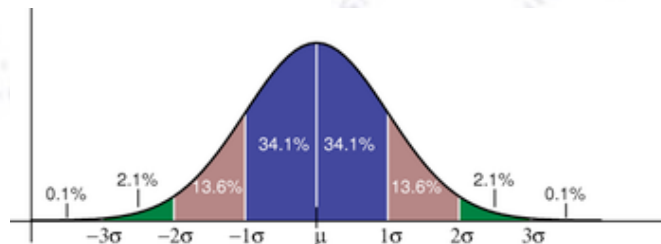


Machine Learning

An introduction



Troels C. Petersen (NBI)

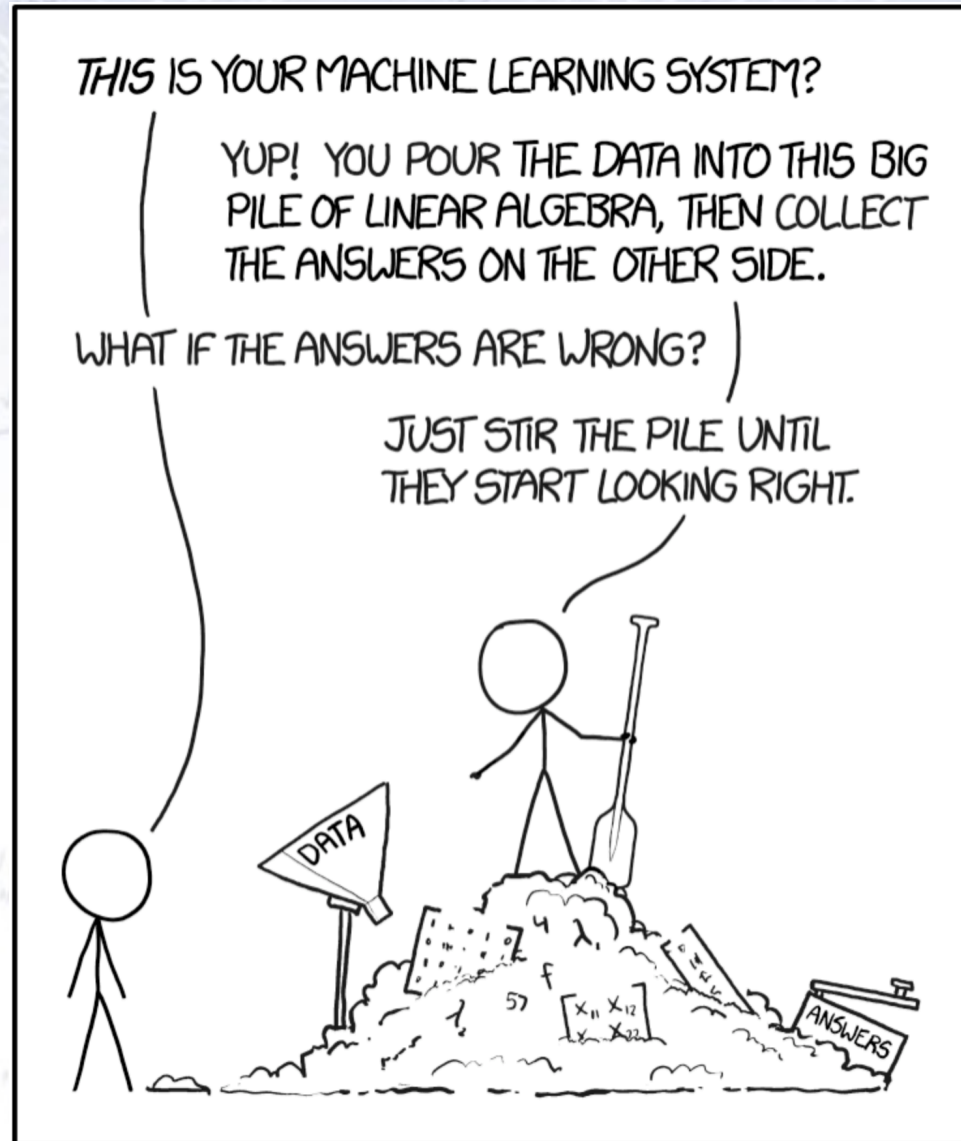


"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"



What is ML?

What is Machine Learning?



What is Machine Learning?

While there is no formal definition, an early attempt is the following intuition:

“Machine learning programs can perform tasks without being explicitly programmed to do so.”

[Arthur Samuel, US computer pioneer 1901-1990]

What is Machine Learning?

While there is no formal definition, an early attempt is the following intuition:

“Machine learning programs can perform tasks without being explicitly programmed to do so.”

[Arthur Samuel, US computer pioneer 1901-1990]

An attempt at a more formal definition is:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E ."

[T. Mitchell, "Machine Learning" 1997]

What is Machine Learning?

While there is no formal definition, an early attempt is the following intuition:

“Machine learning programs can perform tasks without being explicitly programmed to do so.”

[Arthur Samuel, US computer pioneer 1901-1990]

An attempt at a more formal definition is:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E ."

[T. Mitchell, "Machine Learning" 1997]

Under all circumstances, ML allows the analysis and understanding of data, that is complex in terms of both size, dimensionality, quality, and relations [TP].



Two main ingredients

The Universal Approximation Theorem

Theorem 5.1.1 (Universal Approximation Theorem) ¹⁰ Let σ be a non-constant, bounded, and monotone-increasing continuous function. Let I_{m_0} denote the m_0 -dimensional unit hypercube $[0, 1]^{m_0}$. The space of continuous functions on I_{m_0} is denoted as $C(I_{m_0})$. Then given any function $f \in C(I_{m_0})$ and $\epsilon > 0$ there exists a set of real constants a_i, b_i and w_{ij} , where $i = 1, \dots, m_1$ and $j = 1, \dots, m_0$ such that we may define

$$F(x_1, \dots, x_{m_0}) = \sum_{i=1}^{m_1} a_i \sigma \left(\sum_{j=1}^{m_0} w_{ij} x_j + b_i \right) \quad (5.6)$$

as an approximate realization of the function f ; that is,

$$|F(x_1, \dots, x_{m_0}) - f(x_1, \dots, x_{m_0})| < \epsilon \quad (5.7)$$

for all x_1, x_2, \dots, x_{m_0} that lie in the input space.

Universal Approx. Theorems

One main ingredient behind ML are **Universal Approximation Theorems (UAT)**.

These imply that Neural Networks can approximate a very wide variety of functions given simple function constraints and enough degrees of freedom.

This typically entails a large amount of weights, for which the UATs give no recipe on how to find - only that such a construction is possible.

Universal Approx. Theorems

One main ingredient behind ML are **Universal Approximation Theorems (UAT)**.

These imply that Neural Networks can approximate a very wide variety of functions given simple function constraints and enough degrees of freedom.

This typically entails a large amount of weights, for which the UATs give no recipe on how to find - only that such a construction is possible.

Part of this course is learning how to find these!

Decision Trees and K-Nearest Neighbour algorithms are also capable of “universal approximation” (i.e. have forms of UATs).

A UAT has also been worked out for Graph Neural Networks... in 2020!

Universal Approx. Theorems

Regarding UATs, as far as learning is concerned, whether the class is really universal or not is not overly important:

If one assumes that there is no noise in the training set, then there will still be infinitely many functions that pass through all training points and not all of them will have the same error on an unseen point (i.e. the test set).

Universal Approx. Theorems

Regarding UATs, as far as learning is concerned, whether the class is really universal or not is not overly important:

If one assumes that there is no noise in the training set, then there will still be infinitely many functions that pass through all training points and not all of them will have the same error on an unseen point (i.e. the test set).

Thus, one can ask for what sort of functions the approximation applies.

All differentiable functions? Typically, NNs are restricted to this class.

All continuous functions? All measurable functions? All computable functions?

As it turns out, the real deal is characterising that class of functions that can be approximated.

Universal Approx. Theorems

Regarding UATs, as far as learning is concerned, whether the class is really universal or not is not overly important:

If one assumes that there is no noise in the training set, then there will still be infinitely many functions that pass through all training points and not all of them will have the same error on an unseen point (i.e. the test set).

Thus, one can ask for what sort of functions the approximation applies.

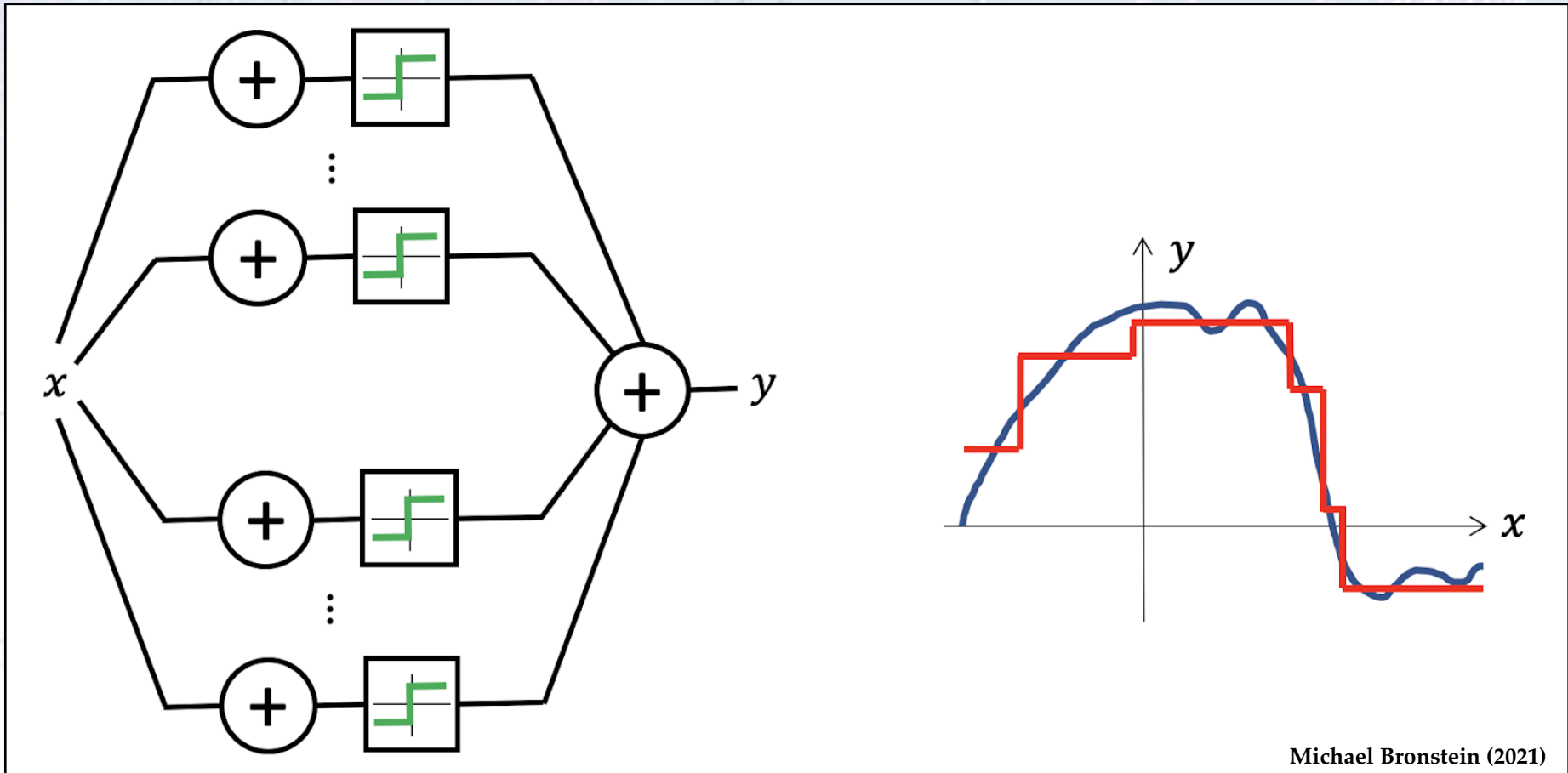
All differentiable functions? Typically, NNs are restricted to this class.

All continuous functions? All measurable functions? All computable functions?

As it turns out, the real deal is characterising that class of functions that can be approximated.

However, we don't really care about that - we simply assume, that with enough liberty/complexity, the functions can approximate what we want.

Universal Approx. Theorems



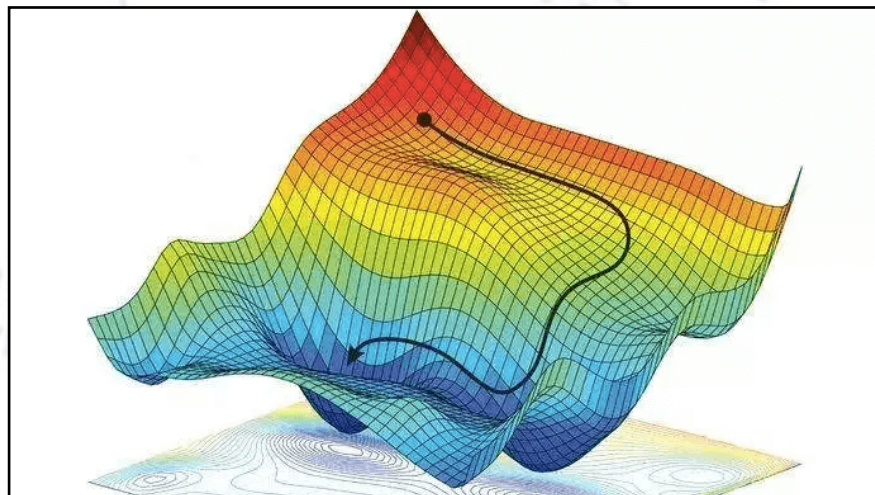
However, we don't really care about that - we simply assume, that with enough liberty/complexity, the functions can approximate what we want.

Stochastic Gradient Descent

The way to obtain the parameters / weights of ML algorithms, is generally by **Stochastic Gradient Descent**.

This “back propagation” algorithm works by computing the gradient of the loss function (to be optimised) with respect to each weight using the chain rule.

One thus computes the gradient one layer at a time, iterating backwards from the last layer (avoiding redundancies). See Goodfellow et al. for details.



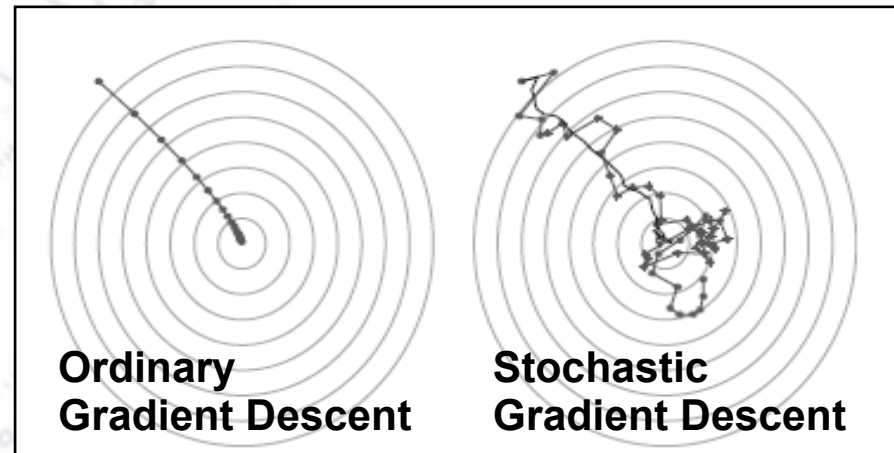
Stochastic Gradient Descent

The way to obtain the parameters / weights of ML algorithms, is generally by **Stochastic Gradient Descent**.

This “back propagation” algorithm works by computing the gradient of the loss function (to be optimised) with respect to each weight using the chain rule.

One thus computes the gradient one layer at a time, iterating backwards from the last layer (avoiding redundancies). See Goodfellow et al. for details.

The gradient descent is made stochastic (and fast) by only considering a fraction (called a “batch”) of the data, when calculating the step in the search for optimal parameters for the algorithm. This allow for stochastic jumping, that avoids local (false) minima.



Ingredients for ML

So now we know that at least in principle:

- a solution exists (Universal Approximation Theorem) and
- that it can be found (Stochastic Gradient Descent).

But this does not in reality make us capable of getting ML results.

We (at least) also need:

- actual functions/ algorithms for making approximations
- knowledge about how to tell them what to learn
- a scheme for how to use the data we have available

Ingredients for ML

So now we know that at least in principle:

- a solution exists (Universal Approximation Theorem) and
- that it can be found (Stochastic Gradient Descent).

But this does not in reality make us capable of getting ML results.

We (at least) also need:

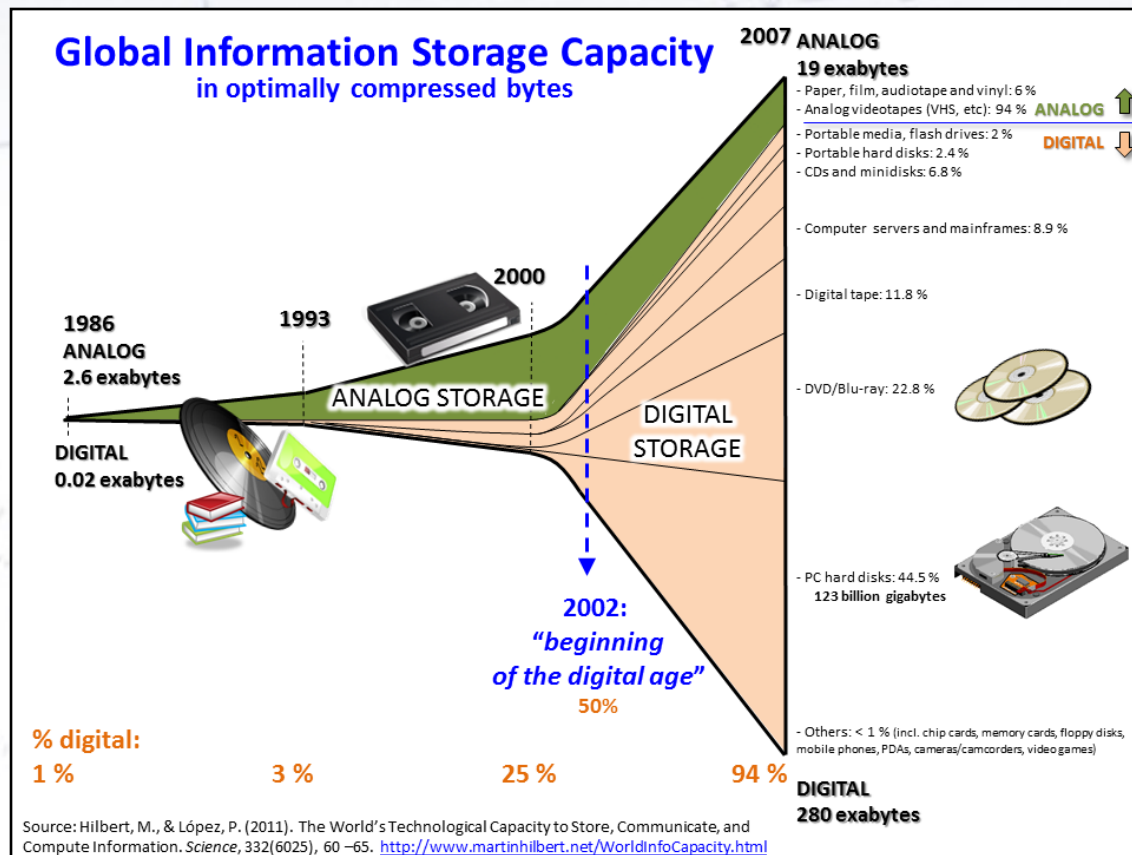
- actual functions/ algorithms for making approximations
Boosted Decision Trees (BDTs) & Neural Networks (NNs)
- knowledge about how to tell them what to learn
Loss functions (and now to minimise these)
- a scheme for how to use the data we have available
Training, validation, and testing samples & Cross Validation



Why ML?

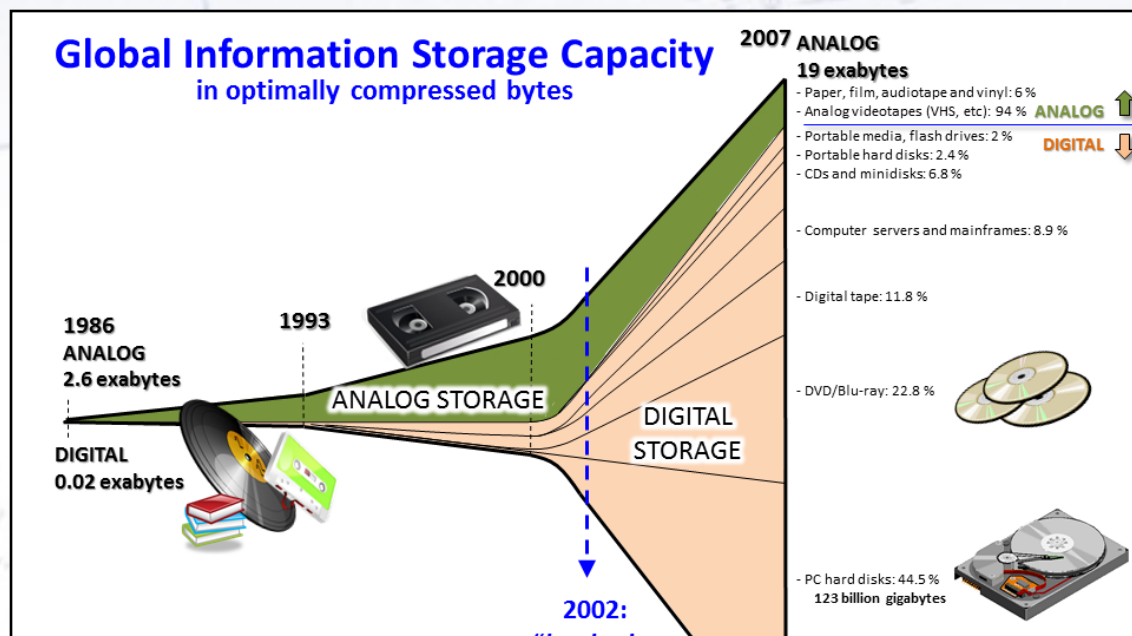
Why Machine Learning?

Part of the “rising” of Machine Learning has been the explosion in data volume, and the easy access to mine it (i.e. internet-of-things), but also the growth in data storage and processing capabilities.



Why Machine Learning?

Part of the “rising” of Machine Learning has been the explosion in data volume, and the easy access to mine it (i.e. internet-of-things), but also the growth in data storage and processing capabilities.



In a digital world, both academia and business has an advantage in understanding their (growing) data volumes. Machine Learning is a powerful tool to do exactly that!

Dimensionality and Complexity

Humans are good at seeing/understanding data in few dimensions!

However, as dimensionality grows, complexity grows exponentially (“curse of dimensionality”), and humans are generally not geared for such challenges.

	Low dim.	High dim.
Linear	Humans: Computers:	Humans: Computers:
Non-linear	Humans: Computers:	Humans: Computers:

Computers, on the other hand, are OK with high dimensionality, albeit the growth of the challenge, but have a harder time facing non-linear issues.

However, through smart algorithms, computers have learned to deal with it all!

That is essentially what Machine Learning has enabled!

Dimensionality and Complexity

Humans are good at seeing/understanding data in few dimensions!

However, as dimensionality grows, complexity grows exponentially (“curse of dimensionality”), and humans are generally not geared for such challenges.

	Low dim.	High dim.
Linear	Humans: ✓ Computers: ✓	Humans: ÷ Computers: ✓
Non-linear	Humans: Computers:	Humans: Computers:

Computers, on the other hand, are OK with high dimensionality, albeit the growth of the challenge, but have a harder time facing non-linear issues.

However, through smart algorithms, computers have learned to deal with it all!

That is essentially what Machine Learning has enabled!

Dimensionality and Complexity

Humans are good at seeing/understanding data in few dimensions!

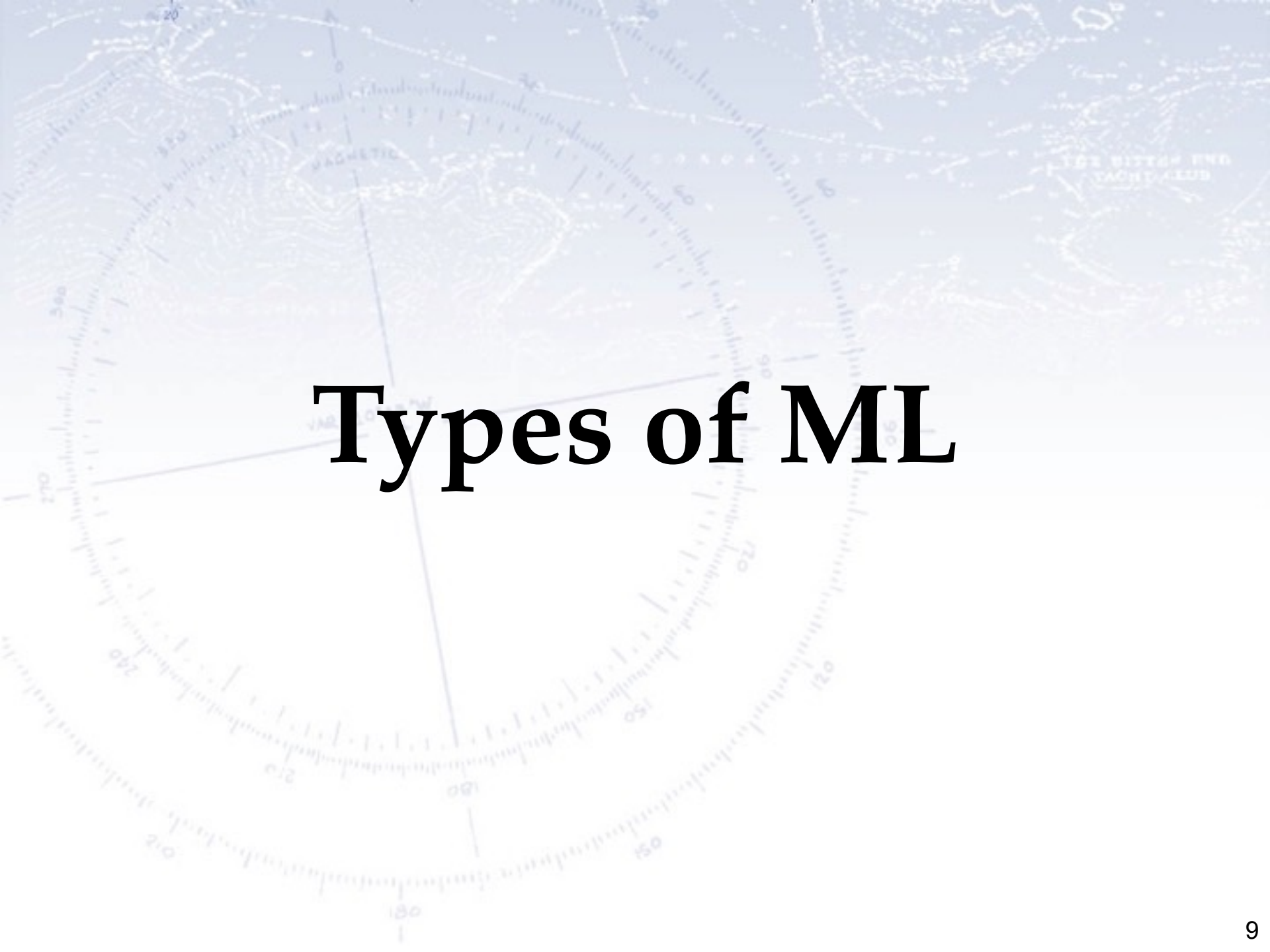
However, as dimensionality grows, complexity grows exponentially (“curse of dimensionality”), and humans are generally not geared for such challenges.

	Low dim.	High dim.
Linear	Humans: ✓ Computers: ✓	Humans: ÷ Computers: ✓
Non-linear	Humans: ✓ Computers: (✓)	Humans: ÷ Computers: (✓)

Computers, on the other hand, are OK with high dimensionality, albeit the growth of the challenge, but have a harder time facing non-linear issues.

However, through smart algorithms, computers have learned to deal with it all!

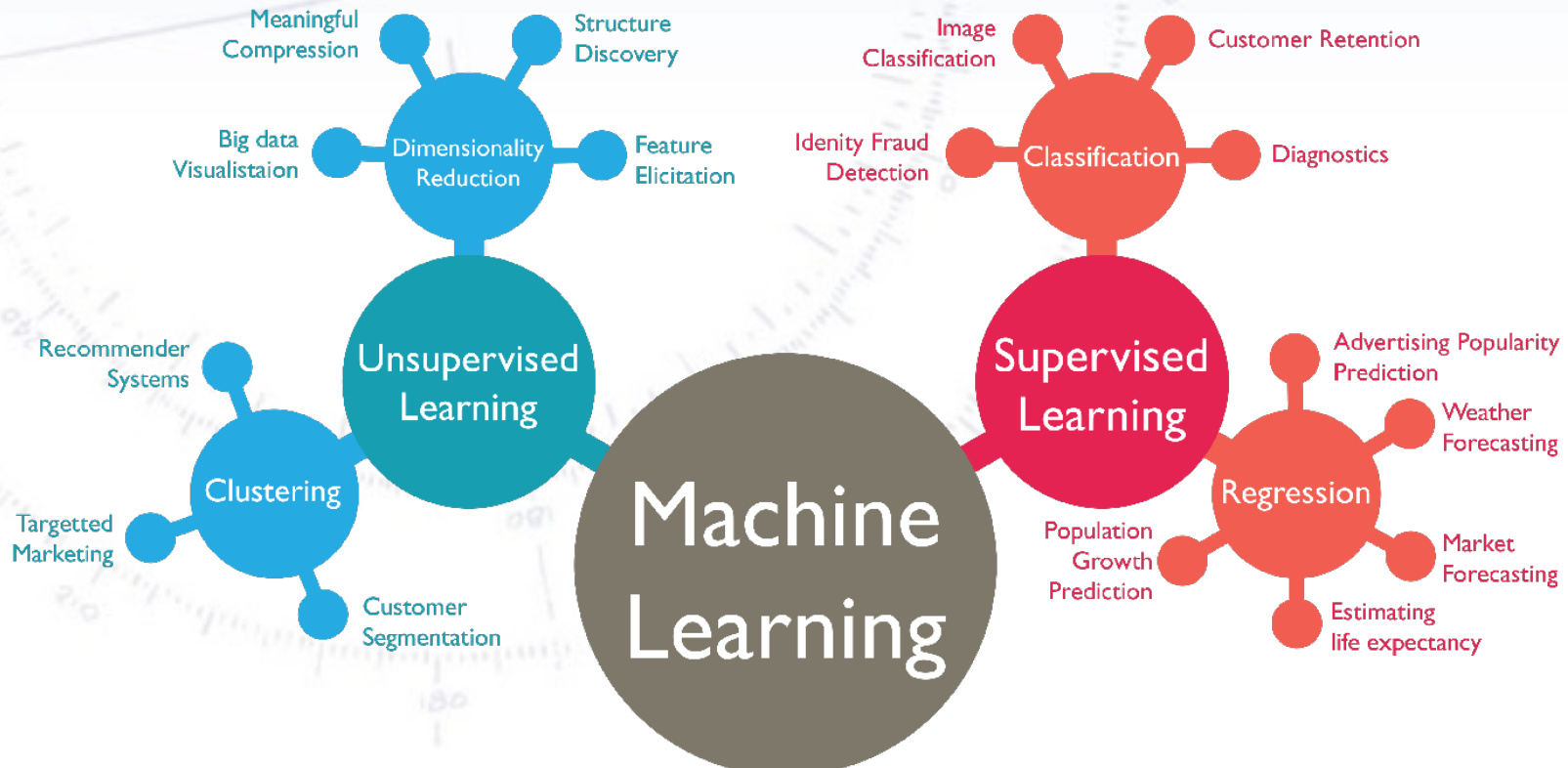
That is essentially what Machine Learning has enabled!



Types of ML

Unsupervised vs. Supervised Classification vs. Regression

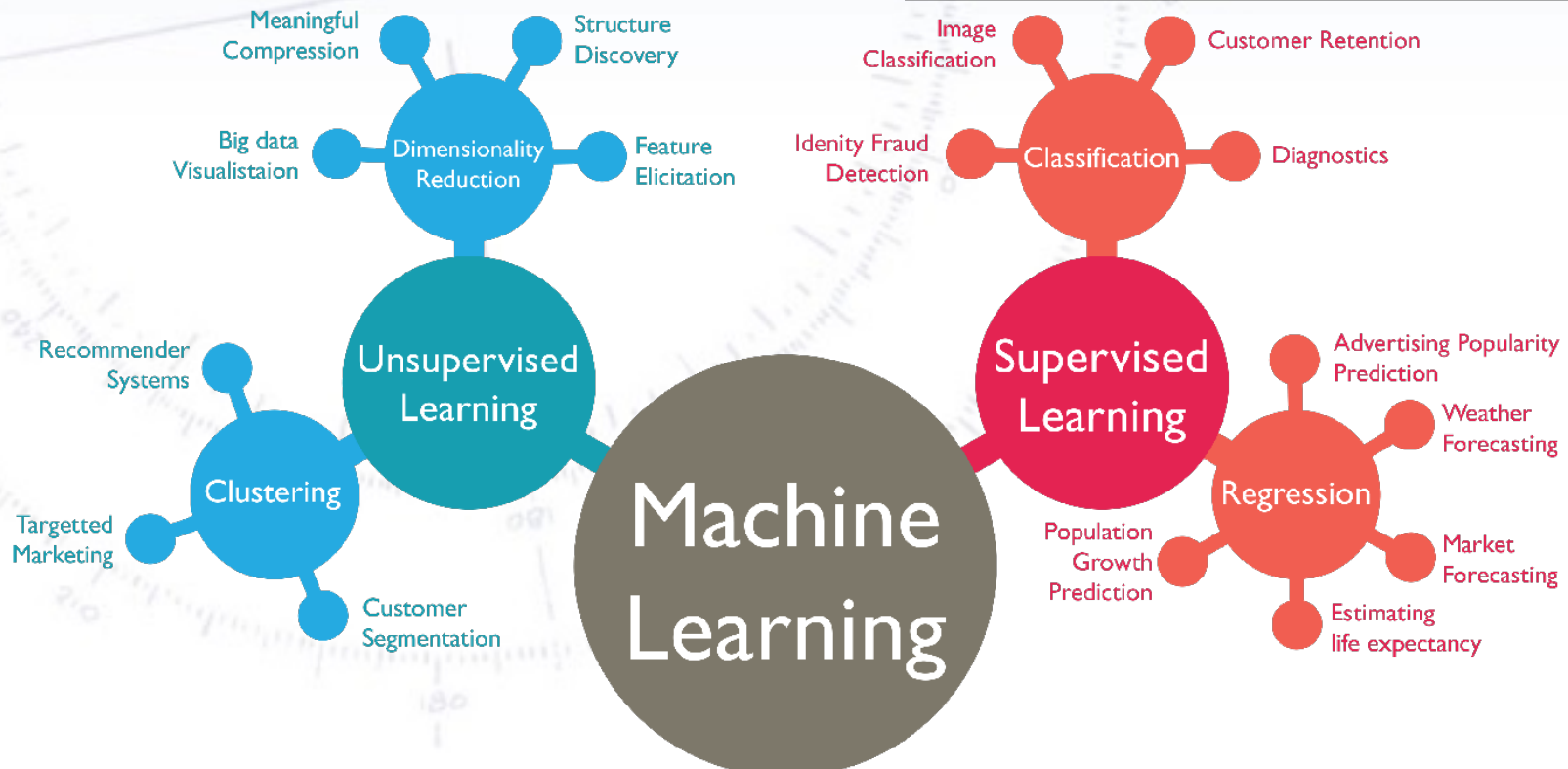
Machine Learning can be supervised (you have correctly labelled examples) or unsupervised (you don't)... [or reinforced]. Following this, one can be using ML to either classify (is it A or B?) or for regression (estimate of X).



Unsupervised vs. Supervised Classification vs. Regression

Machine Learning can be supervised (you have correctly labelled examples) or unsupervised (you don't)... [or reinforced]. Following this, one can be using ML to either classify (is it A or B?) or for regression (estimate of X).

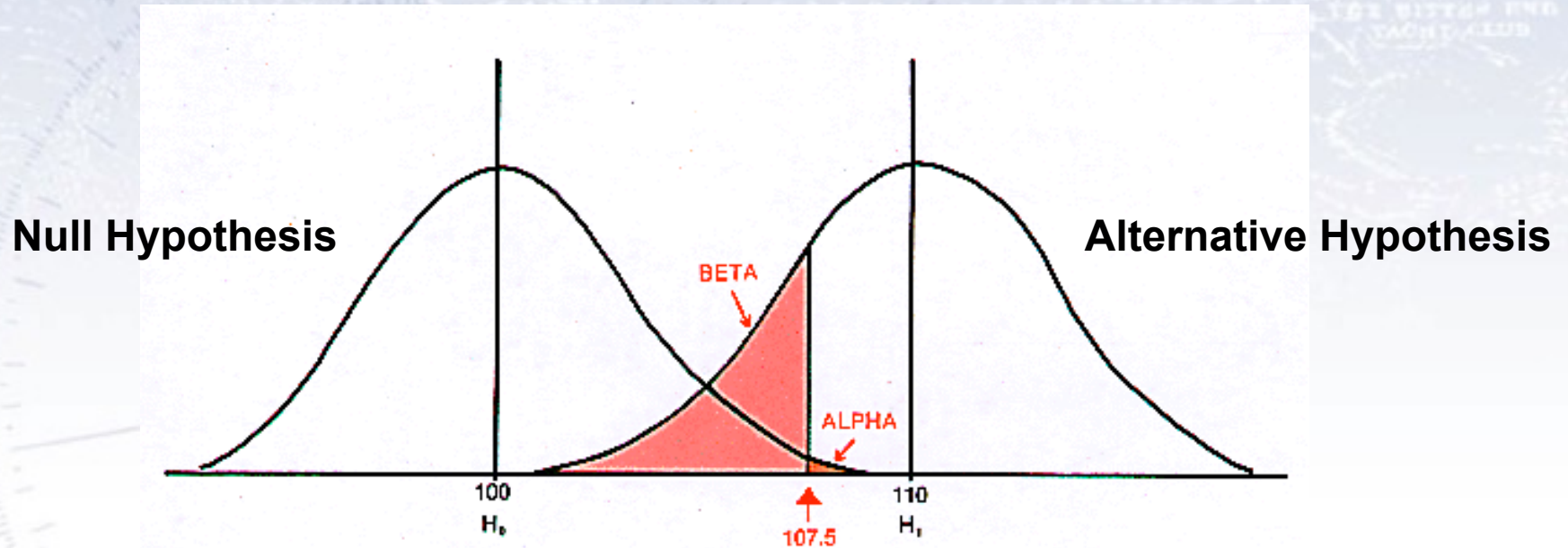
We will be mostly on this side!





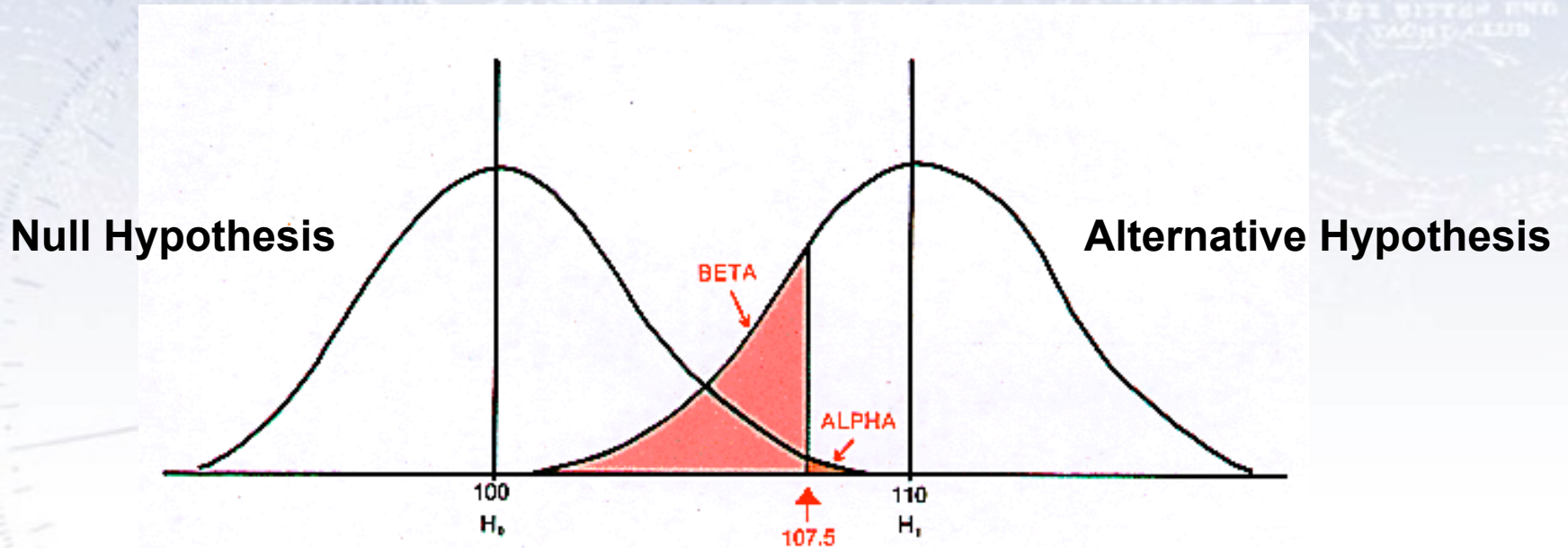
Target of ML

Classification



		REALITY	
		Null is True	Null is False
STATISTICAL DECISION:	Do Not Reject Null	$1 - \alpha$ Correct	β Type II error
	Reject Null	α Type I error	$1 - \beta$ Correct

Classification



Machine Learning typically enables a better separation between hypothesis

DECISION:

Reject Null

α Type I error	$1 - \beta$ Correct
--------------------------	------------------------

Hypothesis testing

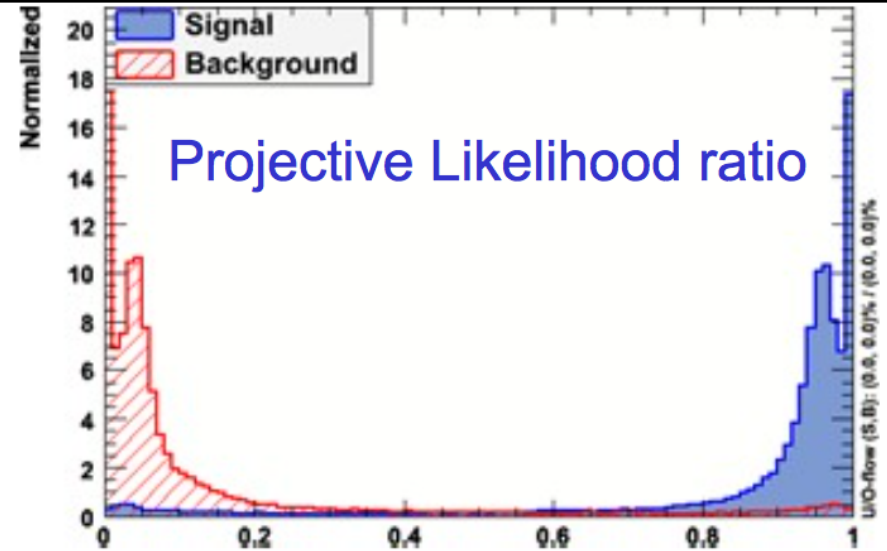
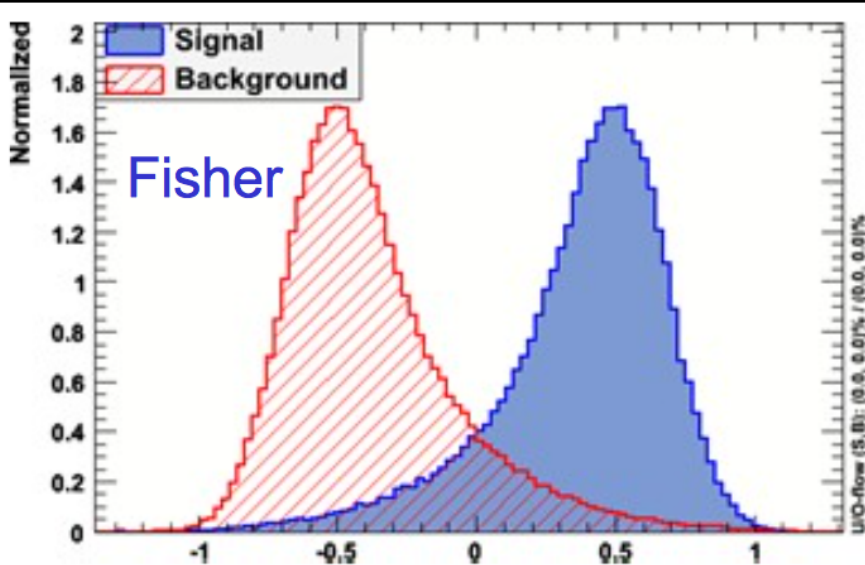
Hypothesis testing is like a criminal trial. The basic “null” hypothesis is **Innocent** (called H_0) and this is the hypothesis we want to test, compared to an “alternative” hypothesis, **Guilty** (called H_1).

Innocence is initially assumed, and this hypothesis is only rejected, if enough evidence proves otherwise, i.e. that the probability of innocence is very small (“beyond reasonable doubt”). This is summarised in a **Contingency Table**:

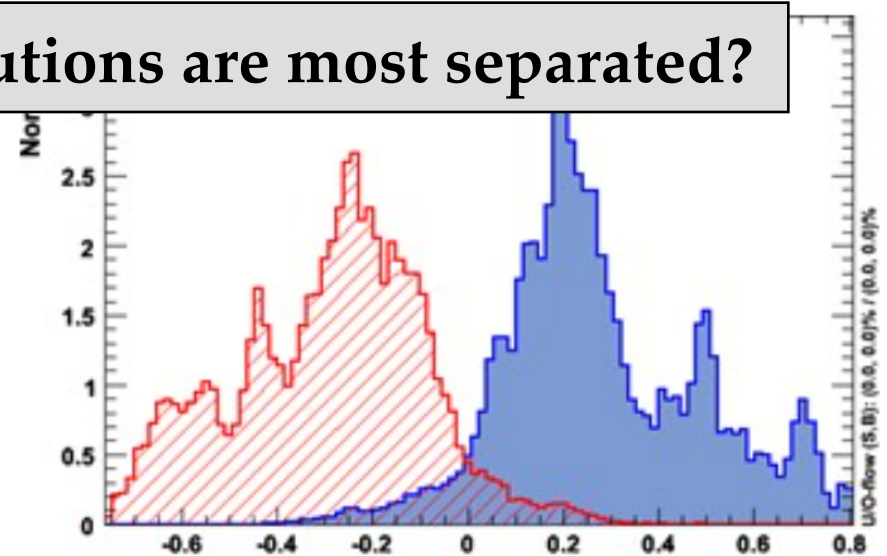
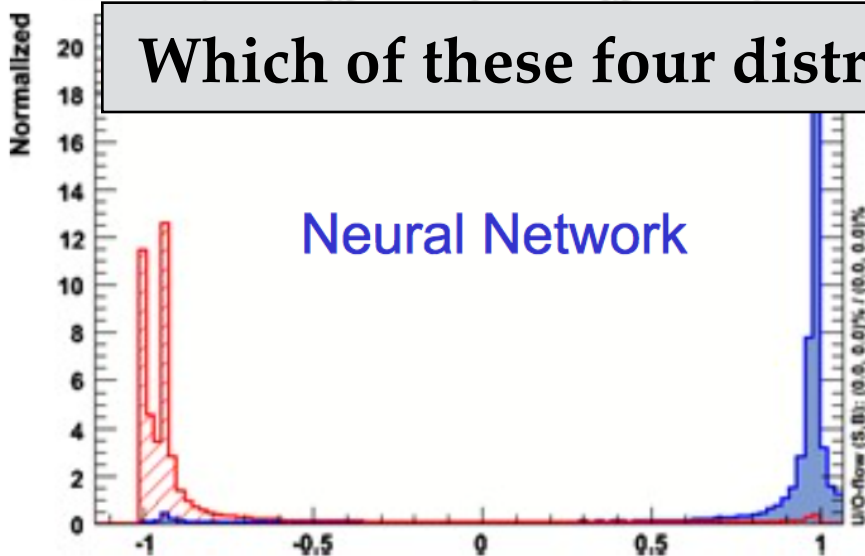
	Truly innocent (H_0 is true)	Truly guilty (H_1 is true)
Acquittal (Accept H_0)	Right decision True Positive (TP)	Wrong decision False Negative (FN)
Conviction (Reject H_0)	Wrong decision False Positive (FP)	Right decision True Negative (TN)

The rate of FP and FN are correlated, and one can only choose one of these!

Measuring separation

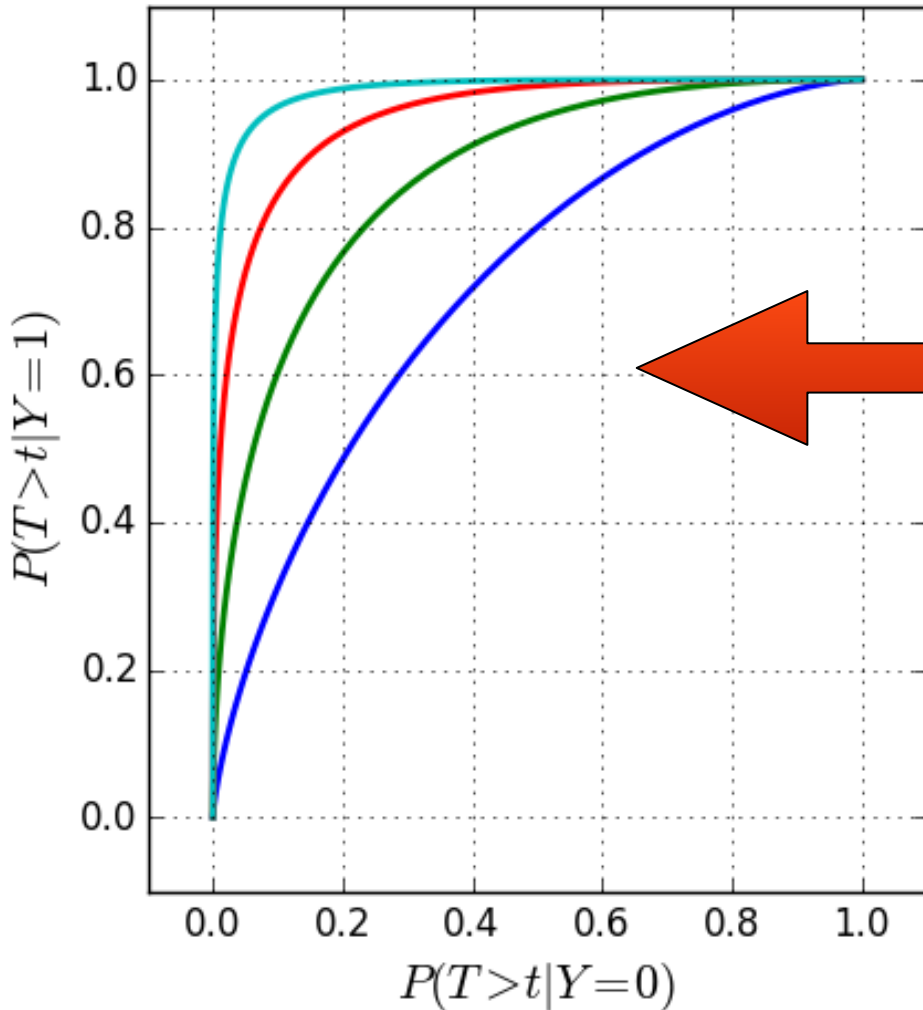


Which of these four distributions are most separated?

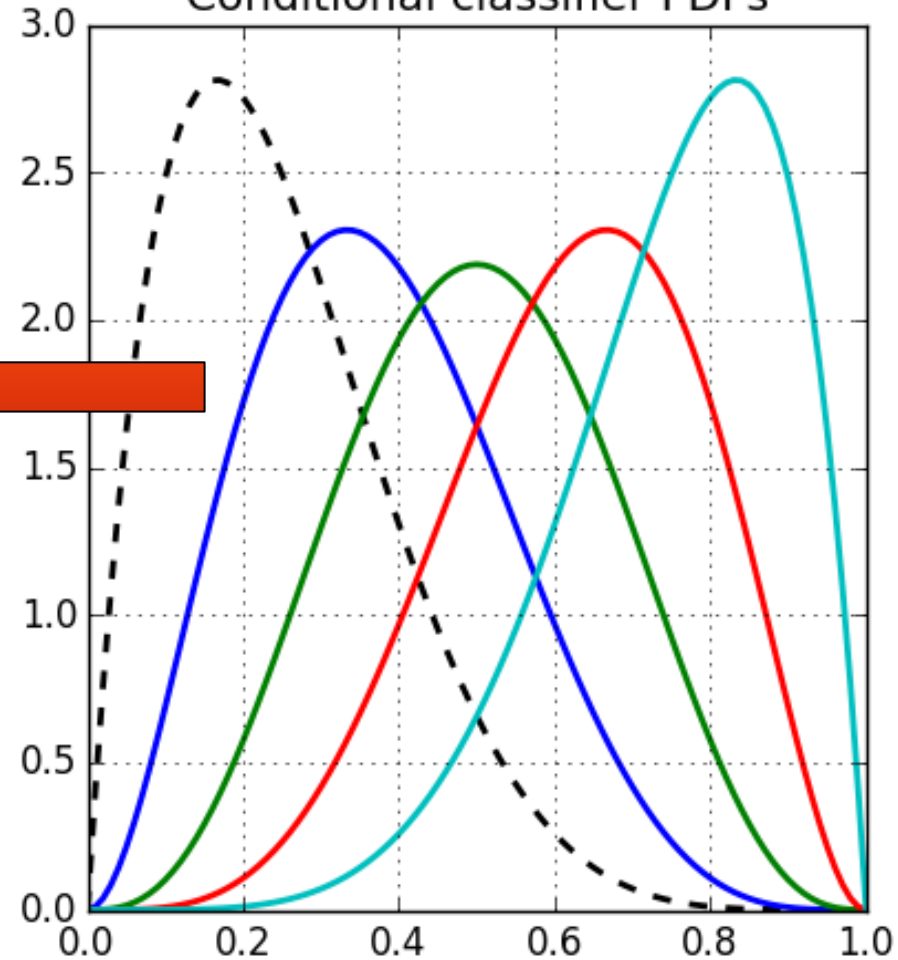


Simple case

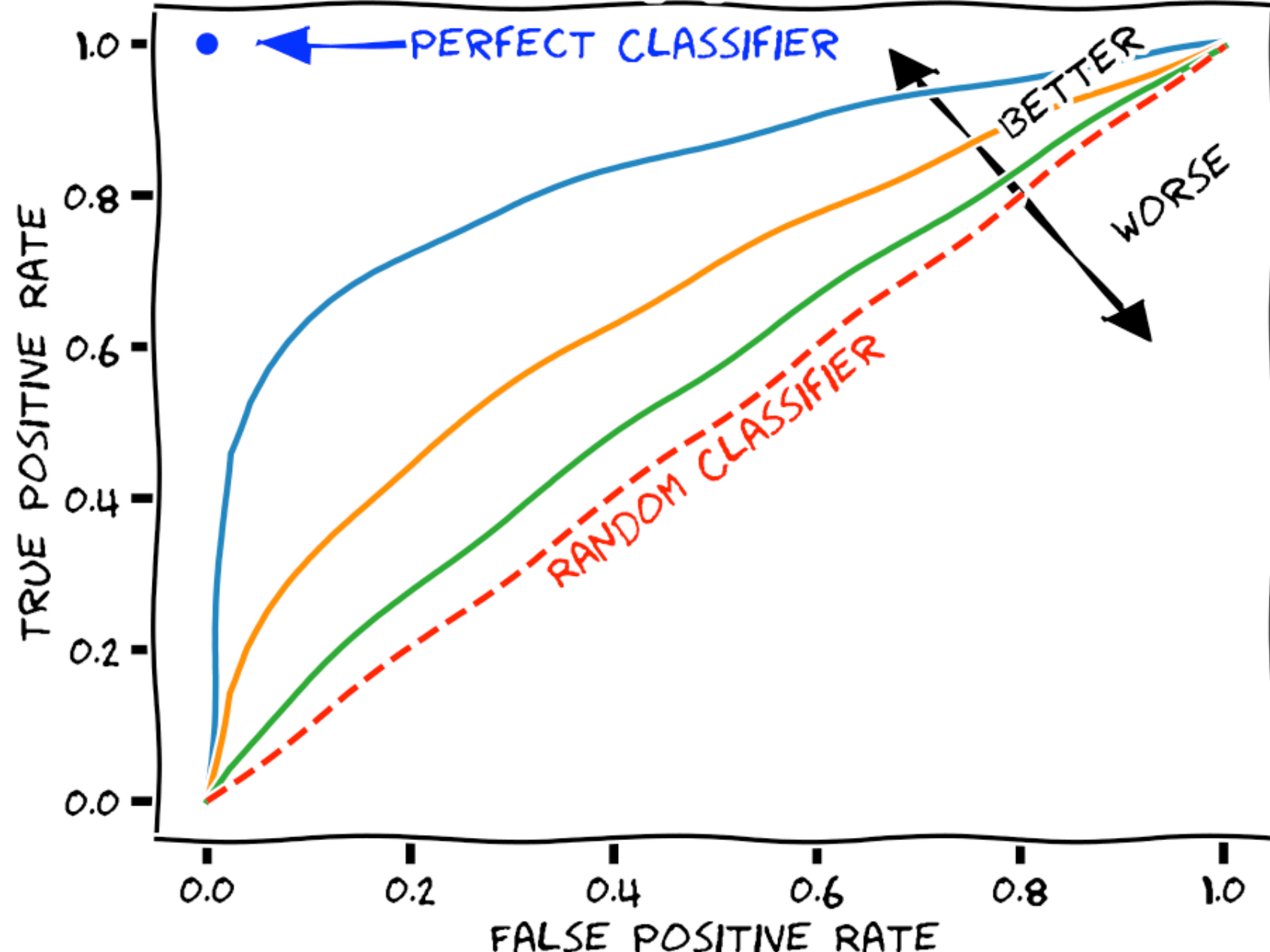
ROC curves



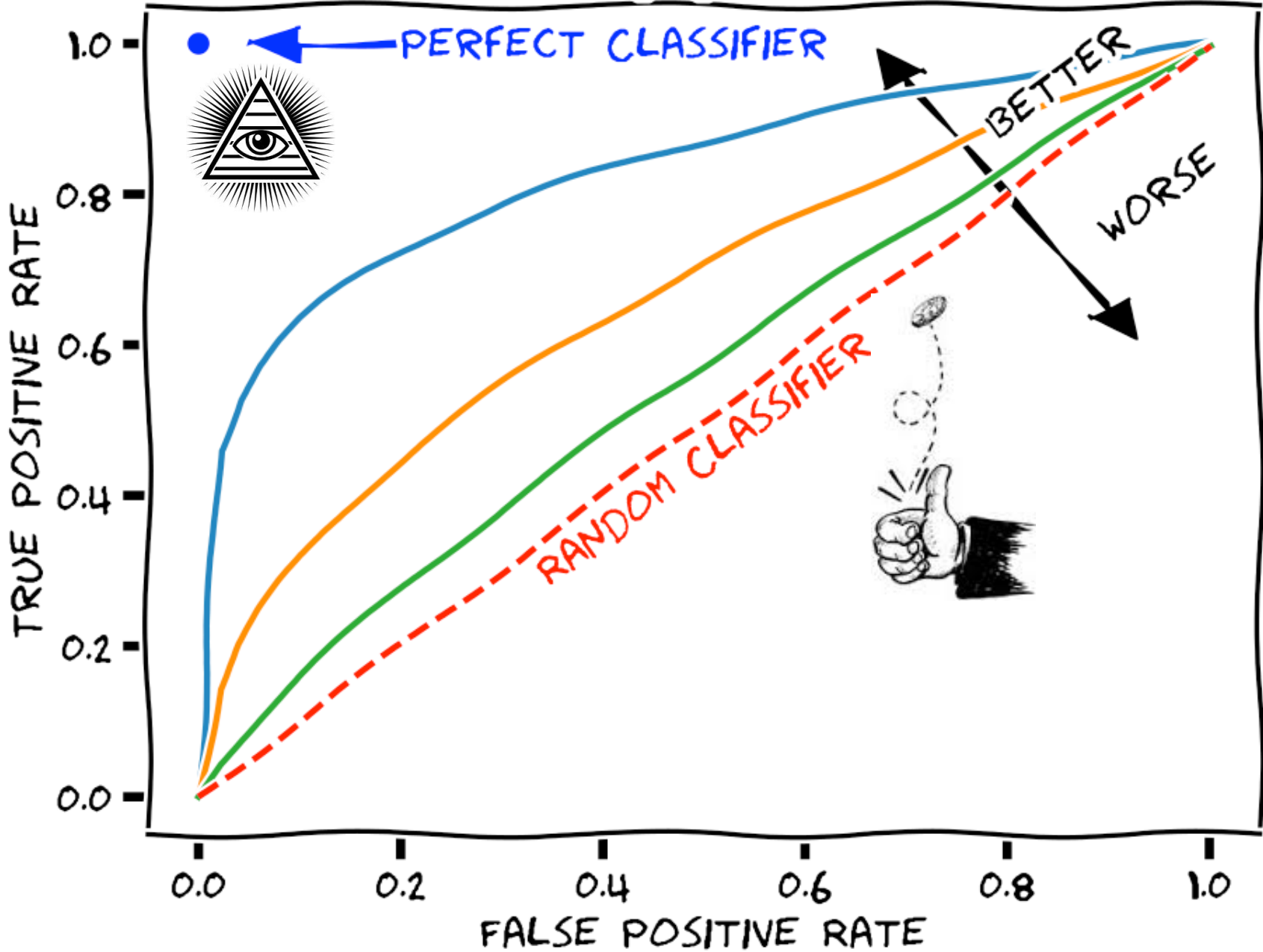
Conditional classifier PDFs



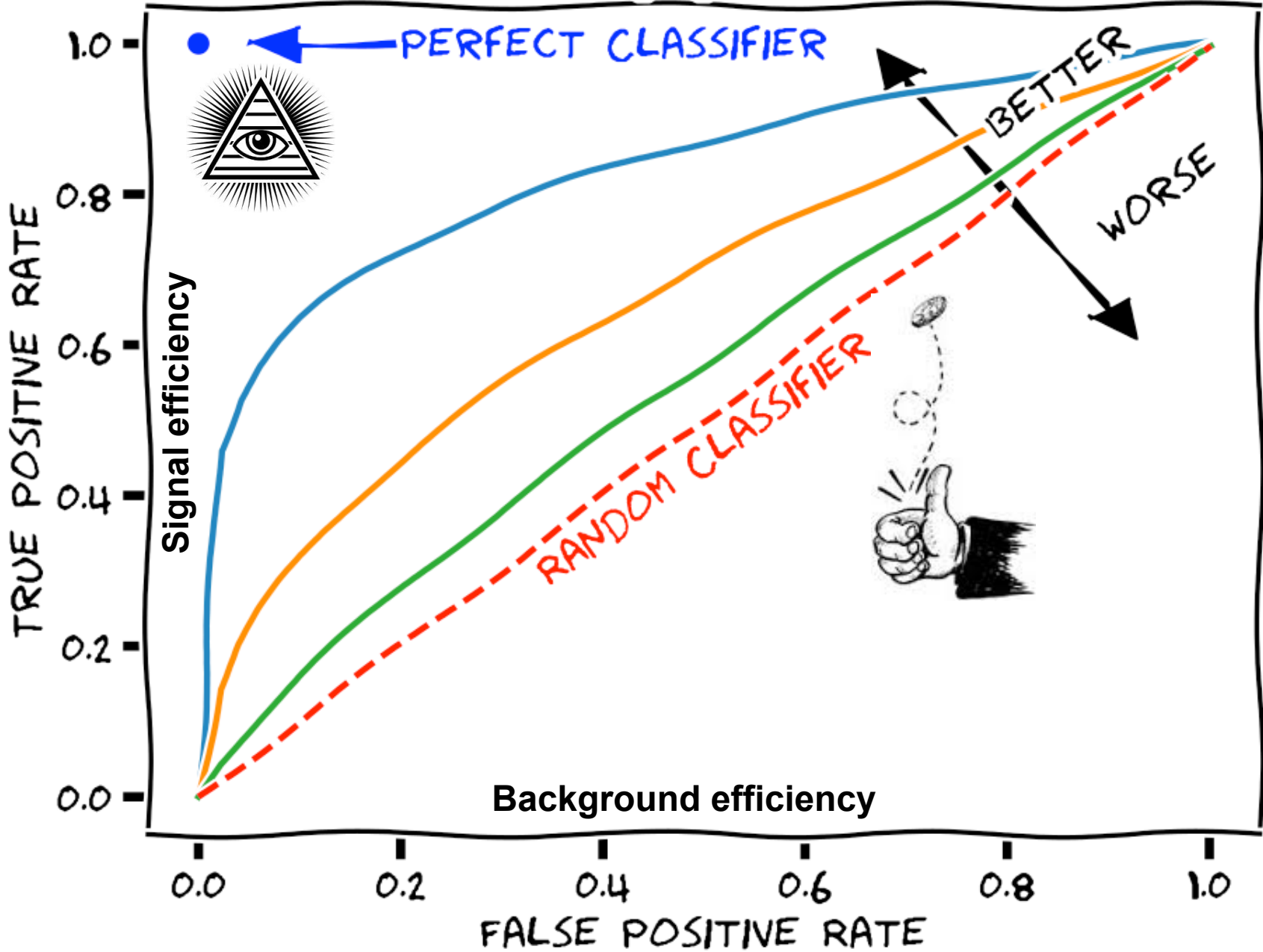
ROC CURVE



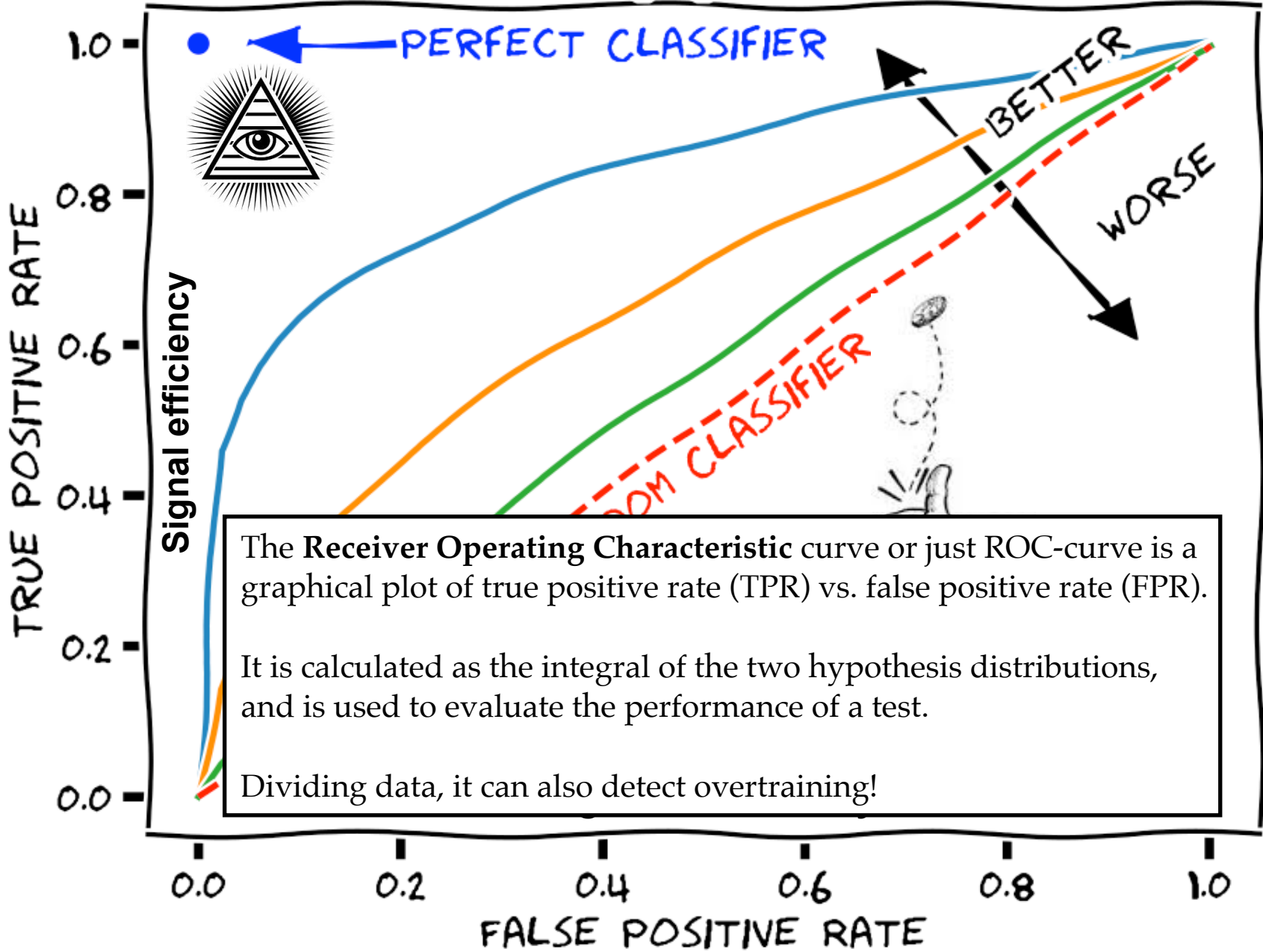
ROC CURVE



ROC CURVE



ROC CURVE



The **Receiver Operating Characteristic** curve or just ROC-curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR).
It is calculated as the integral of the two hypothesis distributions, and is used to evaluate the performance of a test.
Dividing data, it can also detect overtraining!

Which metric to use?

There are a ton of metrics in hypothesis testing, see below. However, those in the boxes below are the most central ones.

One metric - not mentioned here - is the Area Under the Curve (AUC), which is simply an integral of the ROC curve (thus 1 is perfect score). This is sometimes used to optimise performance (loss), but not great!

		True condition			
		Condition positive	Condition negative		
Total population				Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	
				F ₁ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	

Matthew's Correlation Coefficient

Given a Contingency Table:

	Got well	Remained ill
Medicin	28	5
No Medicin	19	9

One of the commonly used measures of separation the MCC, which (in this case) is the Pearson ρ , and related to the ChiSquare:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Read more at:

https://en.wikipedia.org/wiki/Phi_coefficient

However, when optimising an algorithm and giving continuous scores in the range $]0,1[$, there are other things to consider (see talk on Loss Functions).

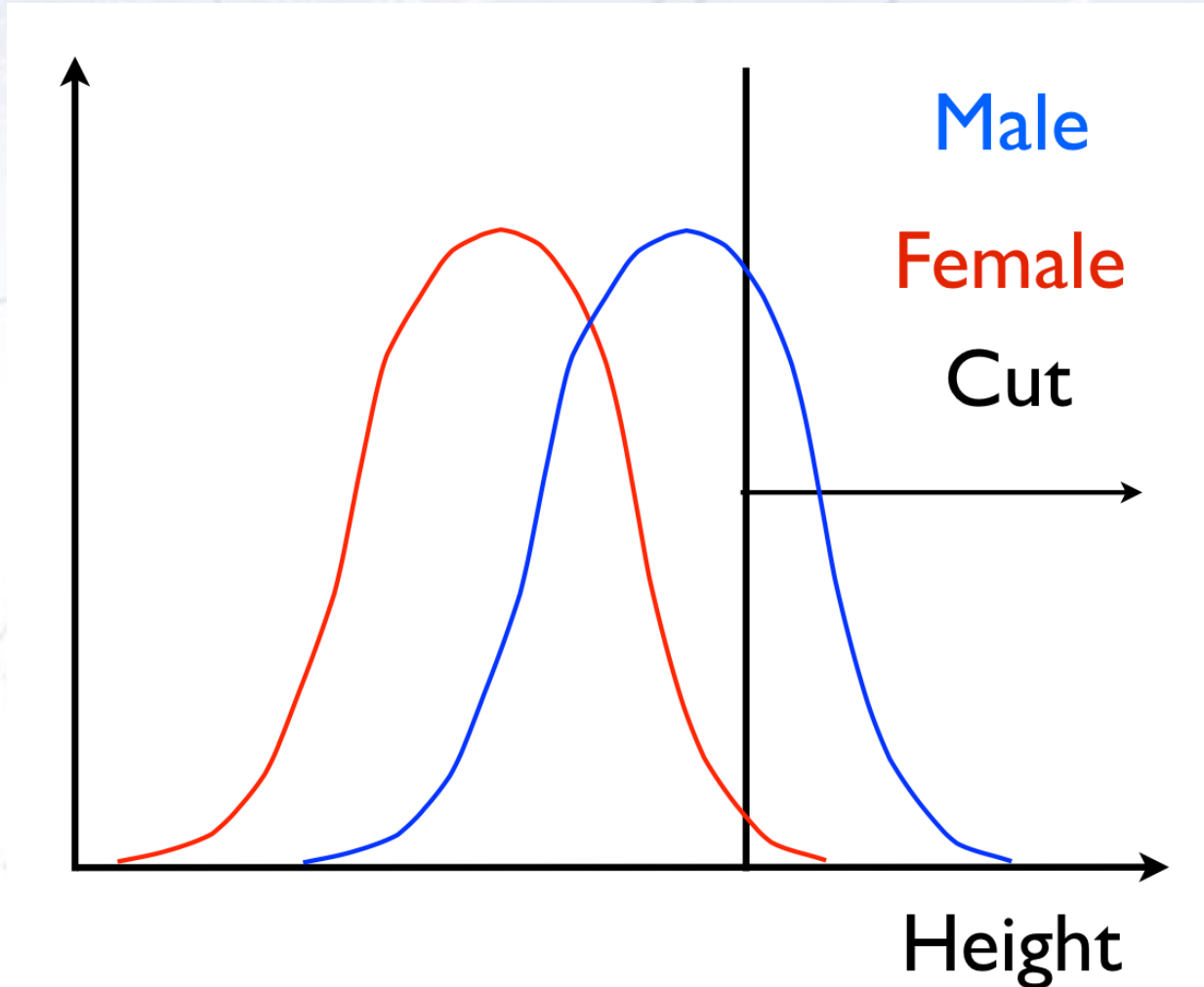


The linear analysis case

Simple Example

Problem: You want to figure out a method for getting sample that is mostly male!

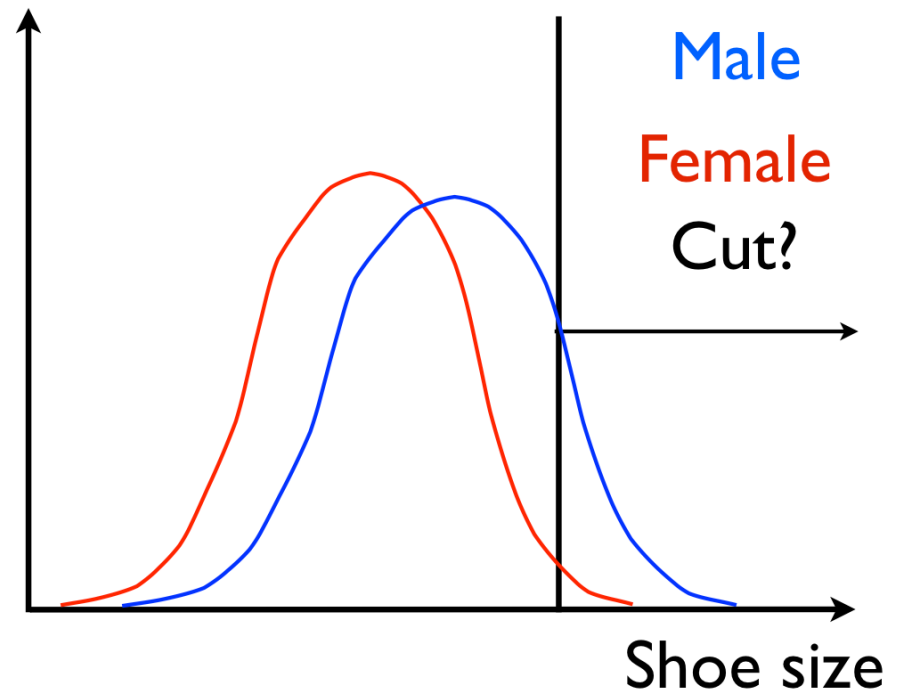
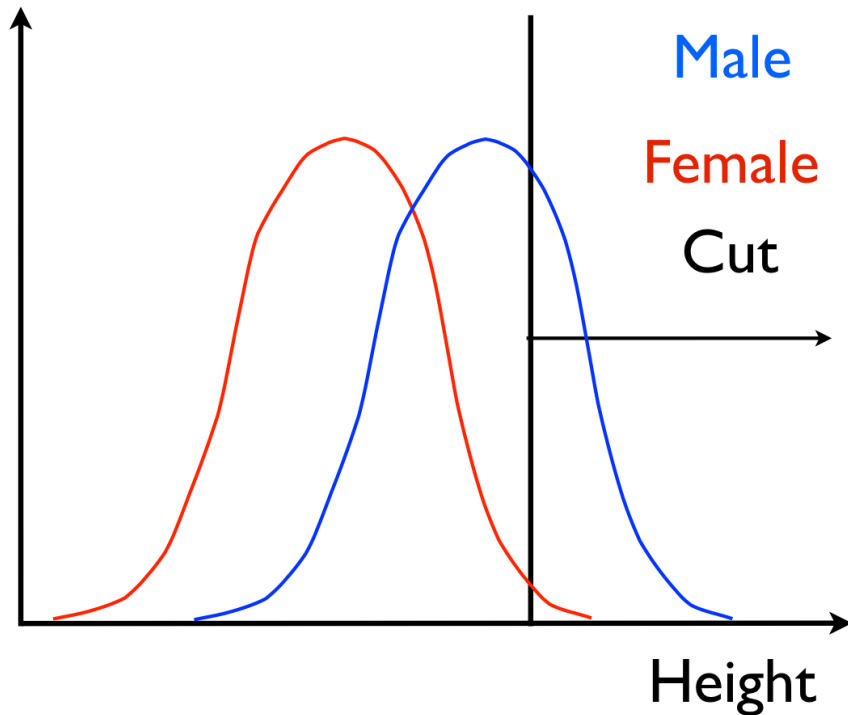
Solution: Gather height data from 10000 people, Estimate cut with 95% purity!



Simple Example

Additional data: The data you find also contains shoe size!

How to use this? Well, it is more information, but should you cut on it?



The question is, what is the best way to use this (possibly correlated) information!

Simple Example

So we look if the data is correlated, and consider the options:

Cut on each var?

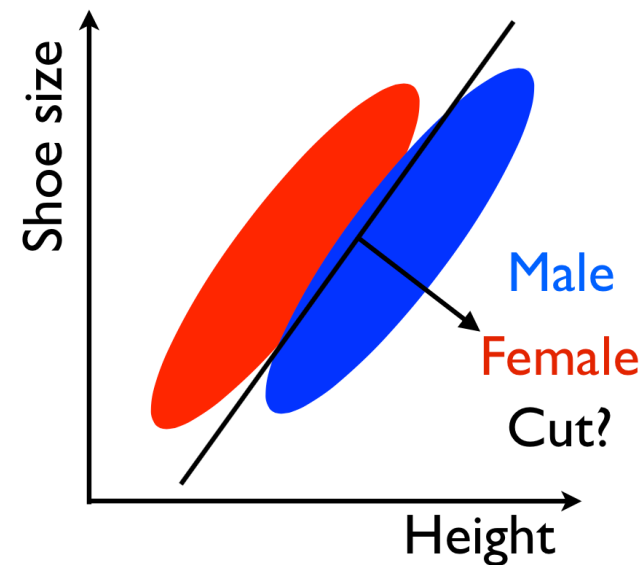
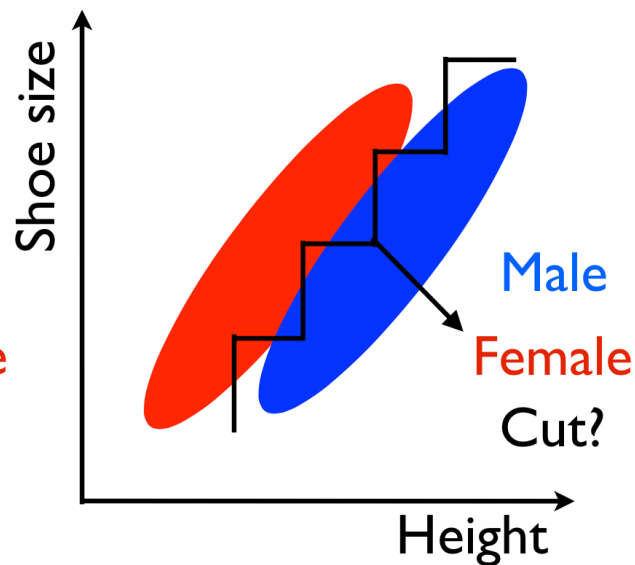
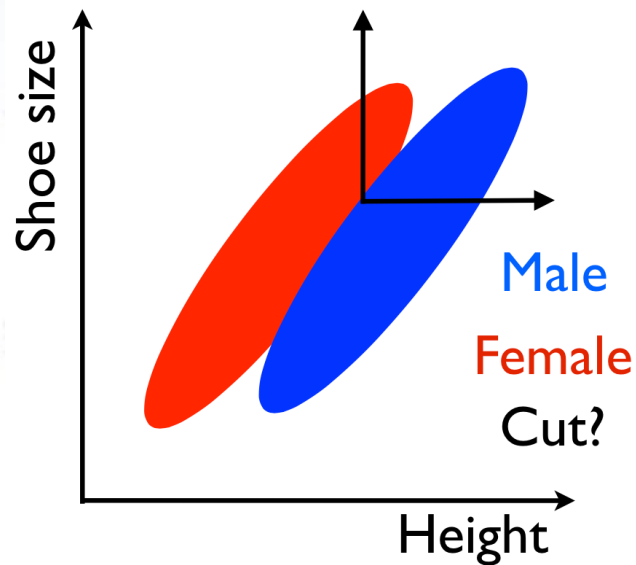
Poor efficiency!

Advanced cut?

**Clumsy and
hard to implement**

Combine var?

**Smart and
promising**



The latter approach is the Fisher discriminant!

It has the advantage of being simple and applicable in many dimensions easily!

Simple Example

So we look if the data is correlated, and consider the options:

Cut on each var?

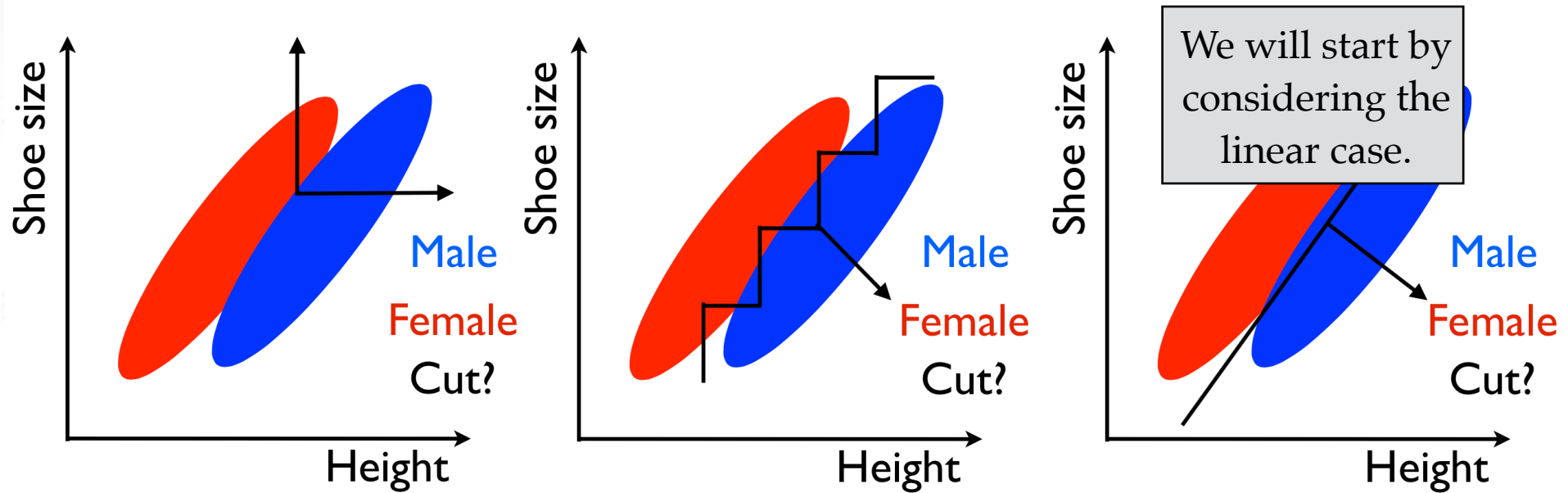
Poor efficiency!

Advanced cut?

Clumsy and
hard to implement

Combine var?

Smart and
promising



The latter approach is the Fisher discriminant!

It has the advantage of being simple and applicable in many dimensions easily!

Simple Example

So we look if the data is correlated, and consider the options:

Cut on each var?

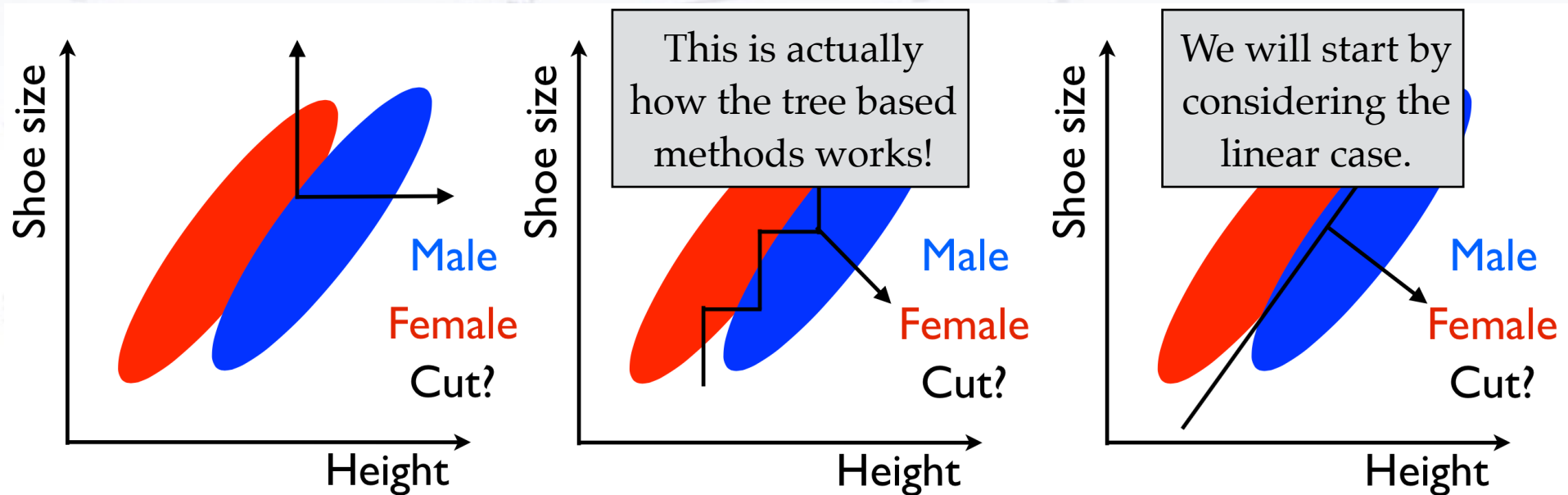
Poor efficiency!

Advanced cut?

Clumsy and
hard to implement

Combine var?

Smart and
promising



The latter approach is the Fisher discriminant!




It has the advantage of being simple and applicable in many dimensions easily!

Separating data

Fisher's friend, Anderson, came home from picking Irises in the Gaspé peninsula...

180 MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

Table I

<i>Iris setosa</i>				<i>Iris versicolor</i>				<i>Iris virginica</i>			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
											
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3

Fisher's Linear Discriminant

You want to separate two types/classes (A and B) of events using several measurements.

Q: How to combine the variables?

A: Use the Fisher Discriminant:

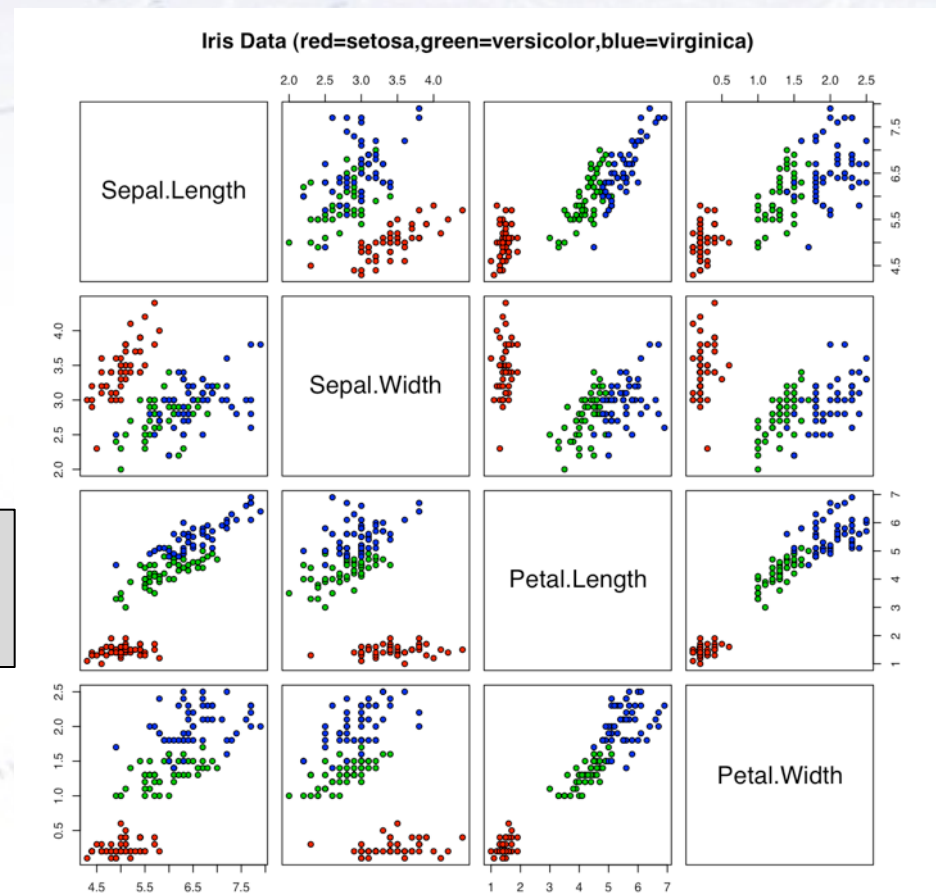
$$\mathcal{F} = w_0 + \vec{w} \cdot \vec{x}$$

Q: How to choose the values of w ?

A: Inverting the covariance matrices:

$$\vec{w} = (\Sigma_A + \Sigma_B)^{-1} (\vec{\mu}_A - \vec{\mu}_B)$$

This can be calculated analytically, and incorporates the linear correlations into the separation capability.



Fisher's Linear Discriminant

You want to separate two types/classes (A and B) of events using several measurements.

Q: How to combine the variables?

A: Use the Fisher Discriminant:

measurements are given. We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II.

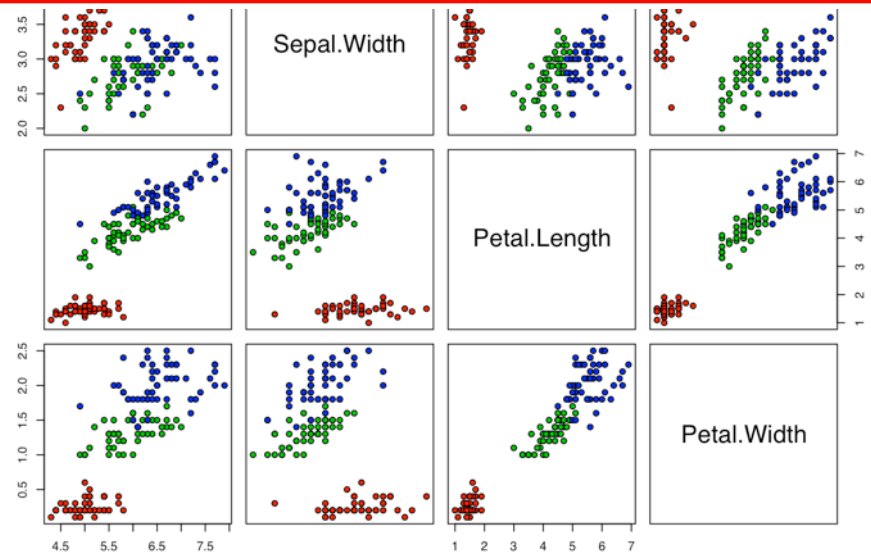
Q: How to choose the values of w ?

A: Inverting the covariance matrices:

$$\vec{w} = (\Sigma_A + \Sigma_B)^{-1} (\vec{\mu}_A - \vec{\mu}_B)$$

This can be calculated analytically, and incorporates the linear correlations into the separation capability.

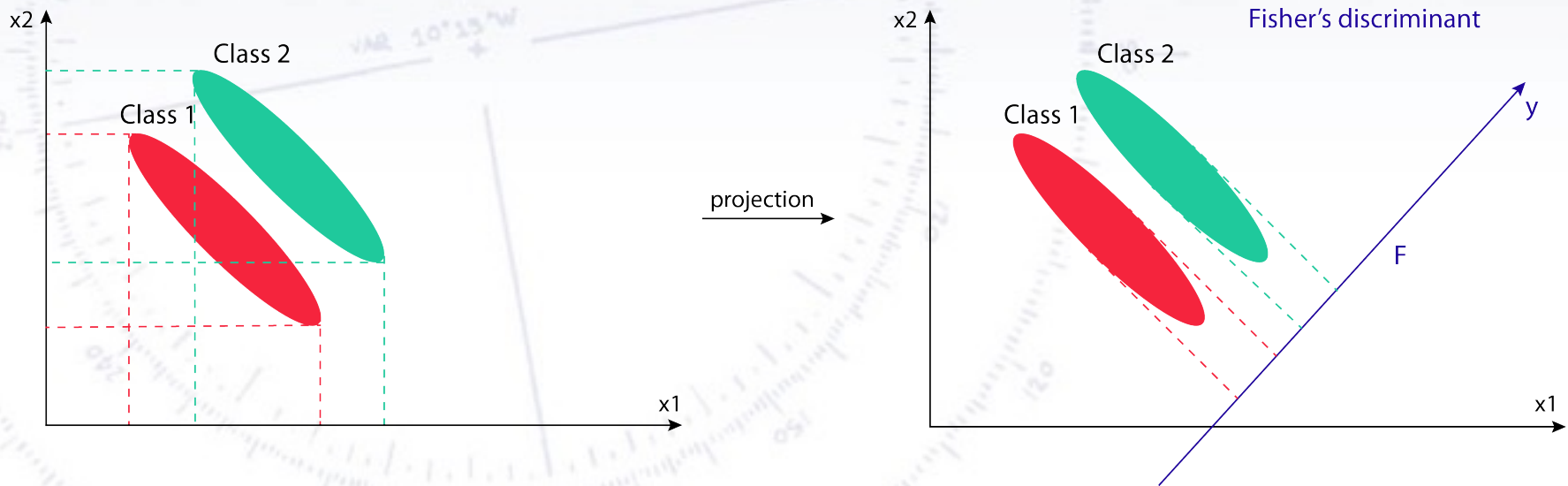
Iris Data (red=setosa,green=versicolor,blue=virginica)



Fisher's Linear Discriminant

Executive summary:

Fisher's Discriminant uses a linear combination of variables to give a single variable with the maximum possible separation (for linear combinations!).



It is for all practical purposes a projection (in a Euclidian space)!

Fisher's Linear Discriminant

The details of the formula are outlined below:

You have two samples, A and B, that you want to separate.

For each input variable (x), you calculate the mean (μ), and form a vector of these.

$$\vec{w} = (\Sigma_A + \Sigma_B)^{-1} (\vec{\mu}_A - \vec{\mu}_B)$$

Using the input variables (x), you calculate the covariance matrix (Σ) for each species (A/B), add these and invert.

Given weights (w), you take your input variables (x) and combine them linearly as follows:

$$\mathcal{F} = w_0 + \vec{w} \cdot \vec{x}$$

F is what you base your decision on.

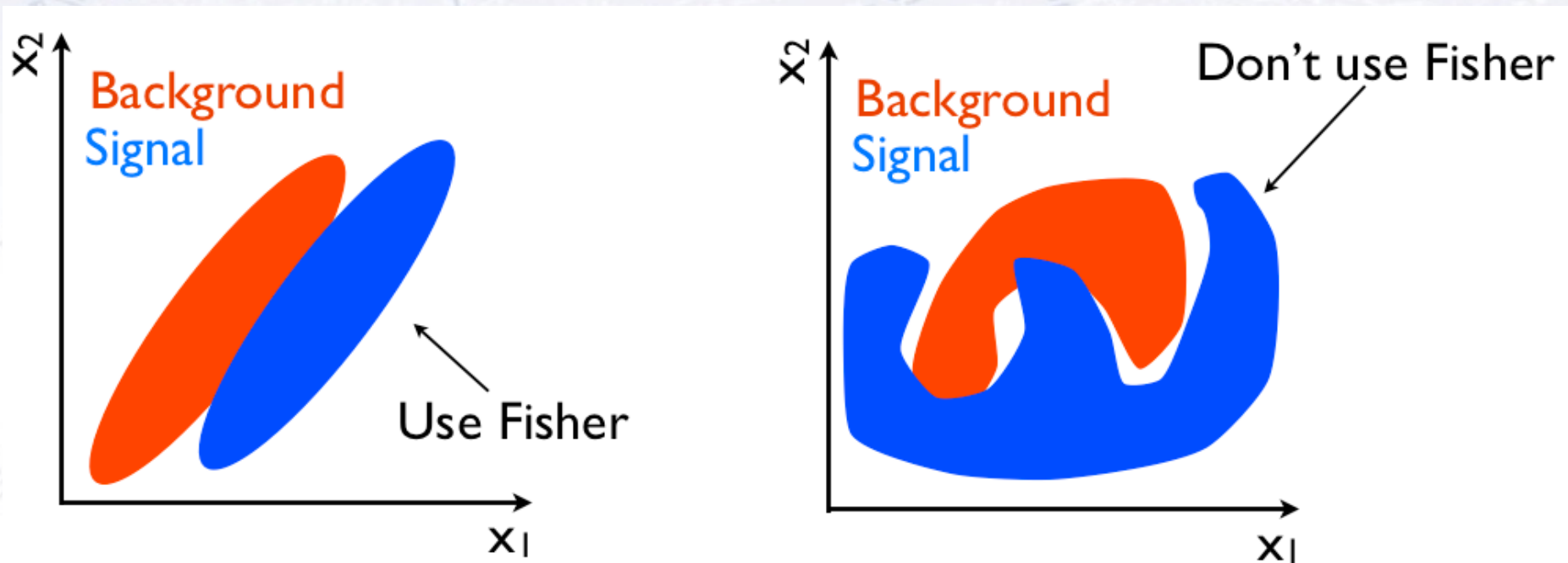


The background is a faded nautical chart. It features several curved lines representing magnetic isogons, labeled with values such as 0, 30, 60, 90, 120, 150, 180, 210, 240, 270, and 300. A specific location is marked with a cross and labeled '10°15'W'. The word 'MAGNETIC' is visible on the chart. In the upper right corner, there is a label '102 BITTER END TACHT/FLUG'. The overall image has a light blue and white color scheme.

The non-linear case

Non-linear cases

While the Fisher Discriminant uses all separations and **linear correlations**, it does not perform optimally, when there are **non-linear correlations** present:



If the PDFs of signal and background are known, then one can use a likelihood. But this is **very rarely** the case, and hence one should move on to the Fisher. However, if correlations are non-linear, more “tough” methods are needed...

(Boosted) Decision Trees

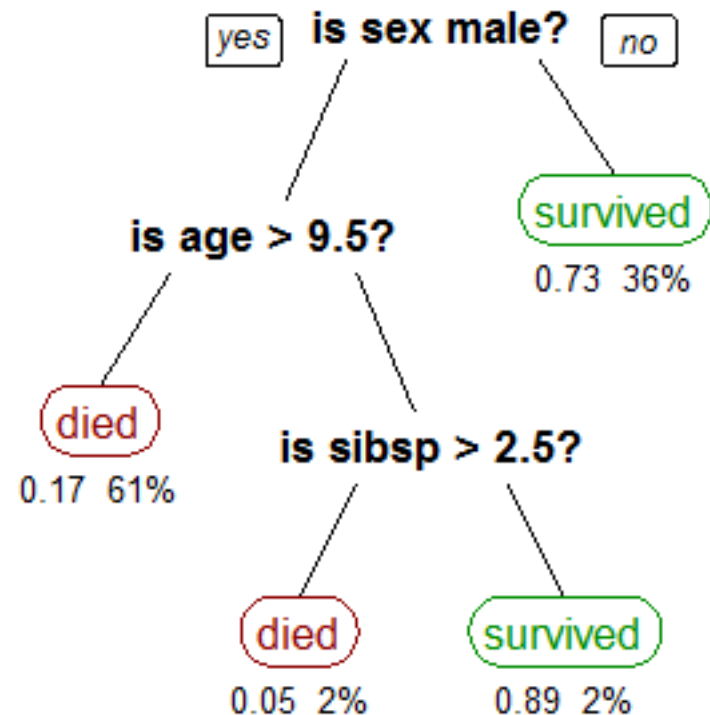
Can become very complex.

Good for discrete problems.

“Good for all problems!!!”

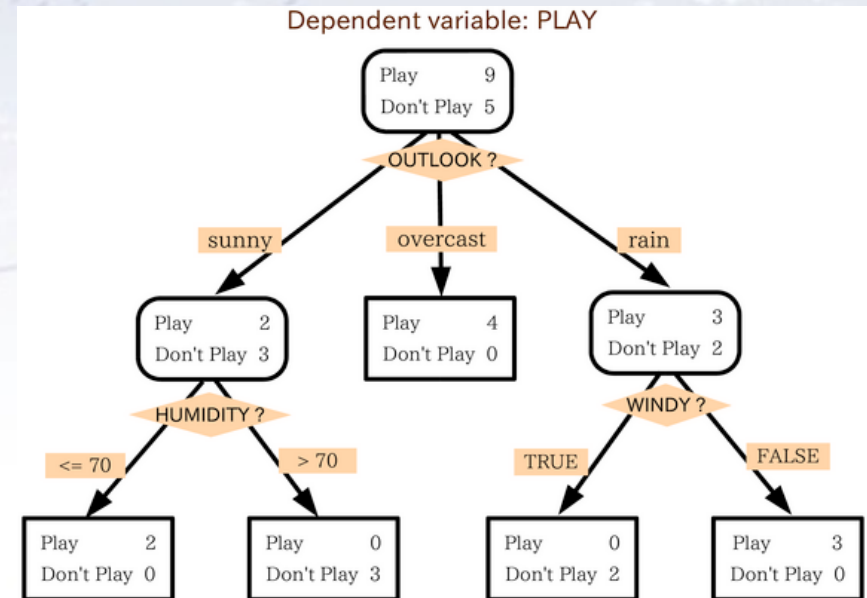
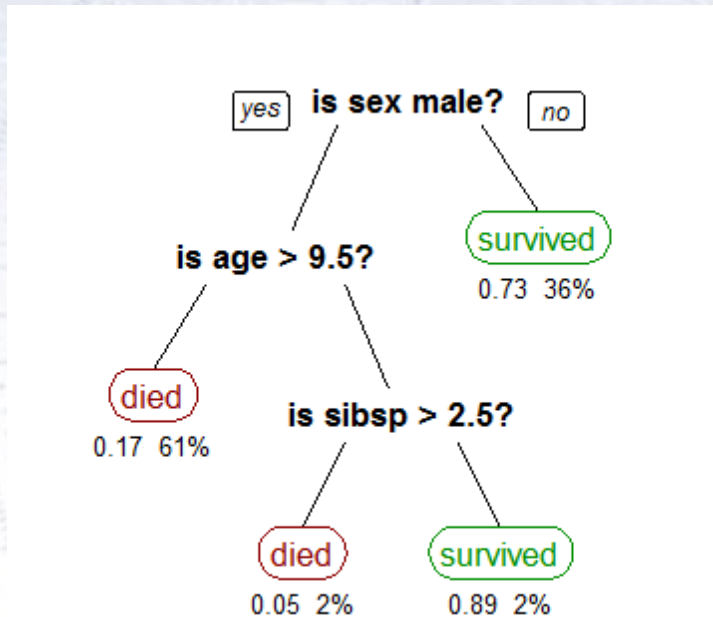
Not always highest efficiency, though...

Boosting adds to separation.



* Example decision tree on a simple algorithm for predicting survival of Titanic!

Boosted Decision Trees (BDT)



*Decision tree learning uses a **decision tree** as a **predictive model** which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in **statistics**, **data mining** and **machine learning**.*

[Wikipedia, Introduction to Decision Tree Learning]

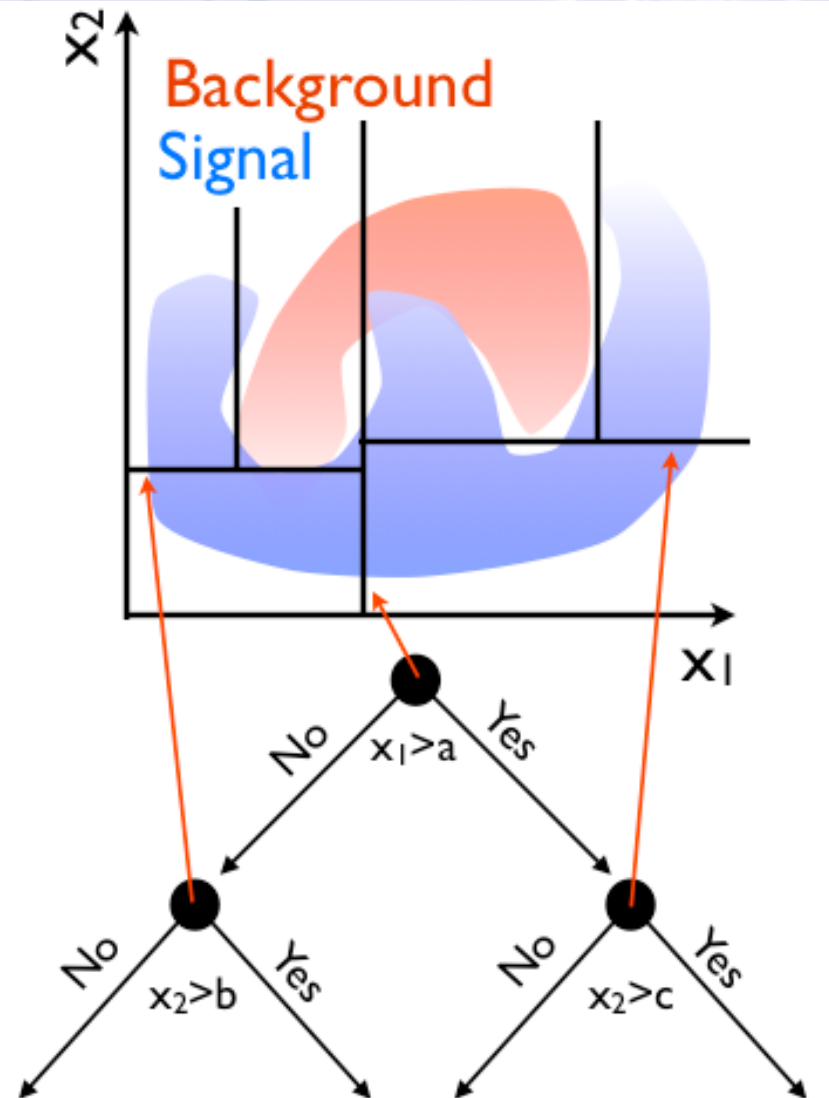
Boosted Decision Trees

A decision tree divides the parameter space, starting with the maximal separation. In the end each part has a probability of being signal or background.

- Works in 95+% of all problems!
- Fully uses non-linear correlations.

But BDTs require a lot of data for training, and is sensitive to overtraining.

Overtraining can be reduced by limiting the number of nodes and number of trees.



Boosting...

There is no reason, why you can not have more trees. Each tree is a simple classifier, but many can be combined!

To avoid N identical trees, one assigns a higher weight to events that are hard to classify, i.e. boosting:

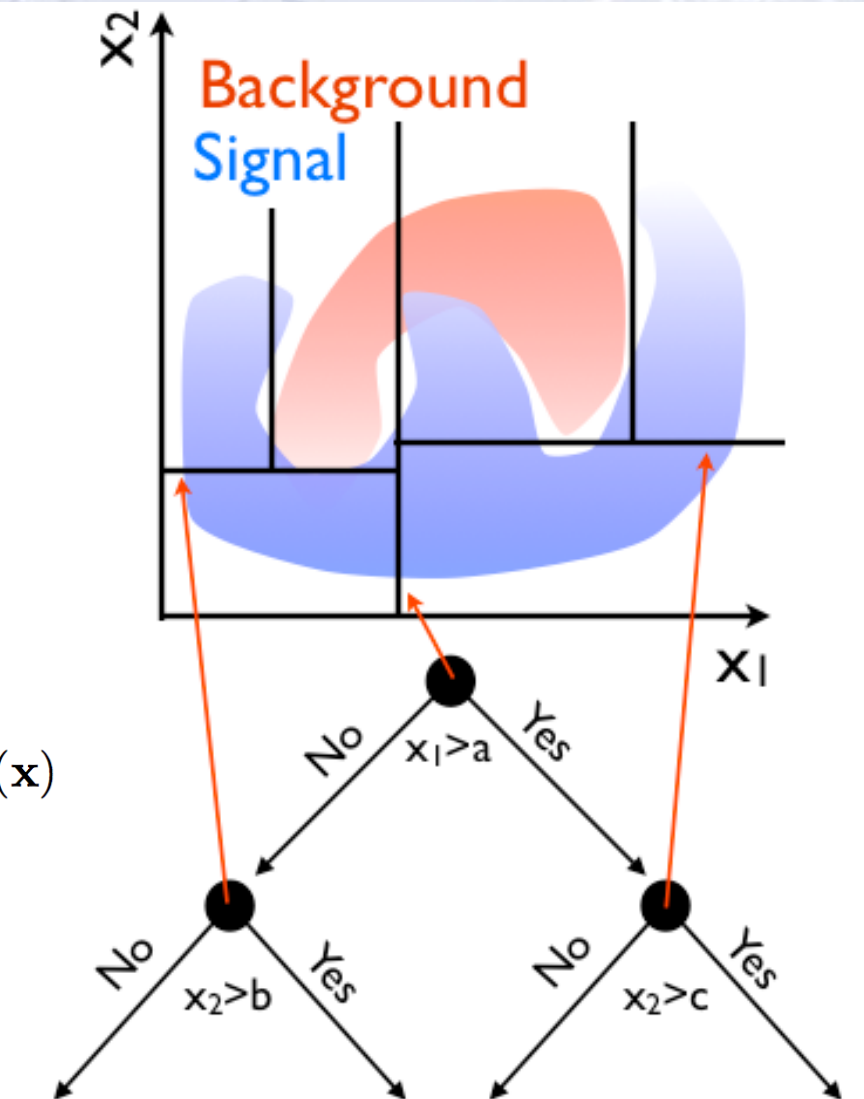
First classifier

Boost weight

$$\alpha = \frac{1 - \text{err}}{\text{err}}$$
$$y_{\text{Boost}}(\mathbf{x}) = \frac{1}{N_{\text{collection}}} \cdot \sum_i^{N_{\text{collection}}} \ln(\alpha_i) \cdot h_i(\mathbf{x})$$

Parameters in event N

Individual tree



Neural Networks

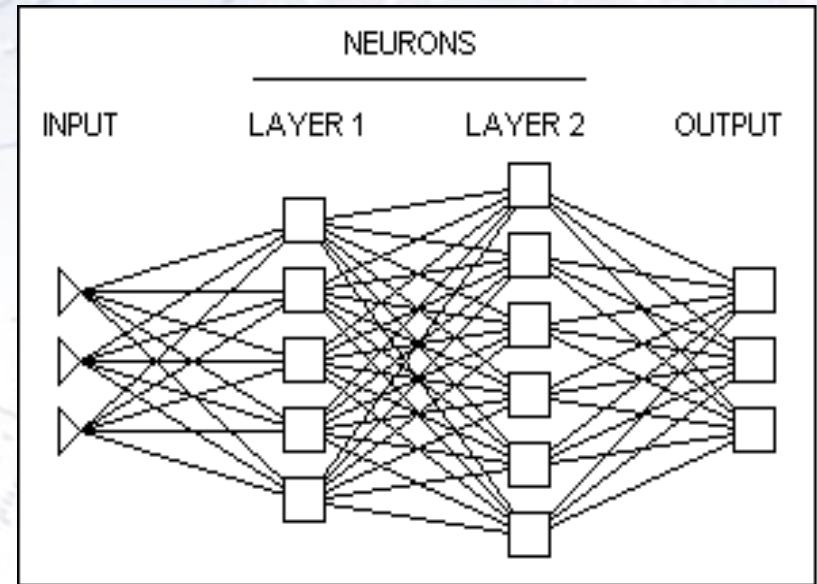
Can become very complex.

Good for continuous problems.

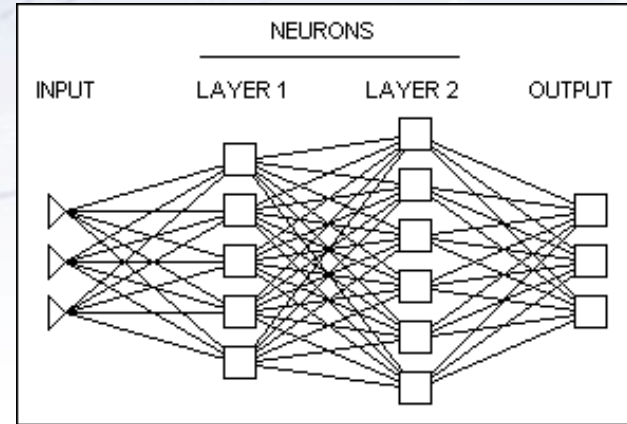
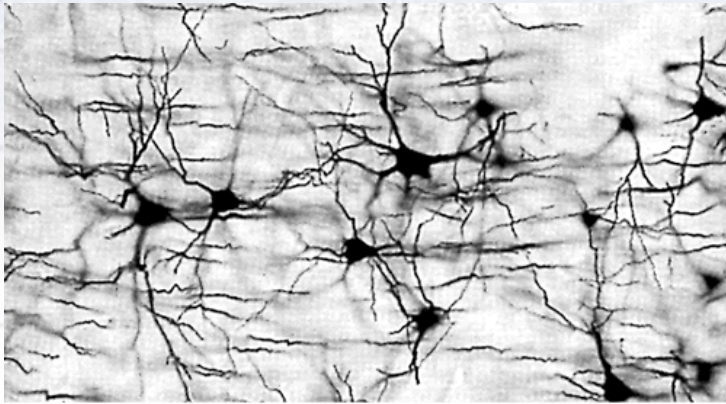
Sometimes hard to train!

Very versatile approach that can also be applied to images, text, etc.

Easily produces multiple outputs.



Neural Networks (NN)



*In machine learning and related fields, artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems (in particular the brain) which is capable of **machine learning** as well as **pattern recognition**.*

*Neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including **computer vision** and **speech recognition**.*

[Wikipedia, Introduction to Artificial Neural Network]

Neural Networks

Neural Networks combine the input variables using a “activation” function $s(x)$ to assign, if the variable indicates signal or background.

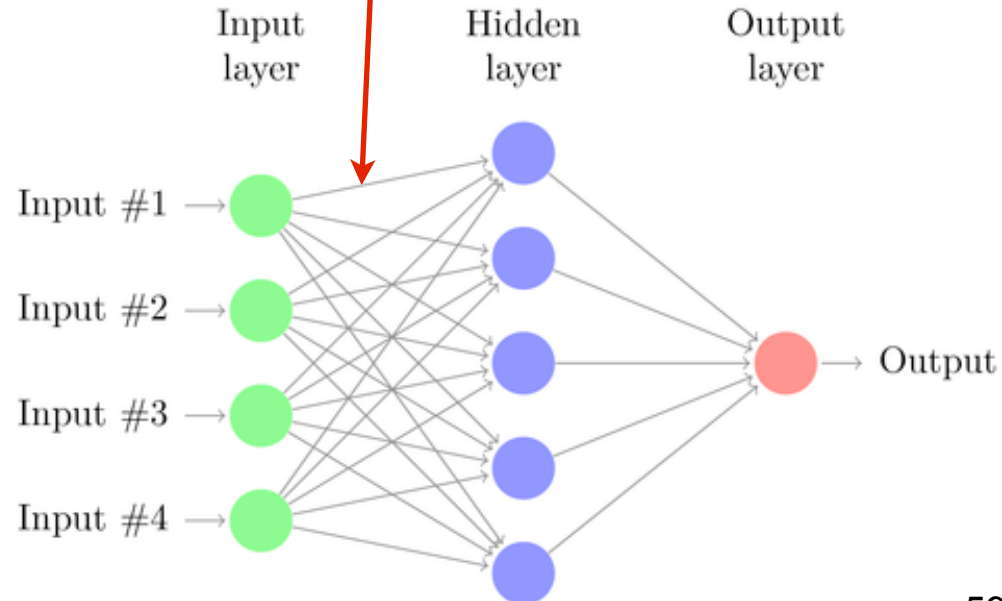
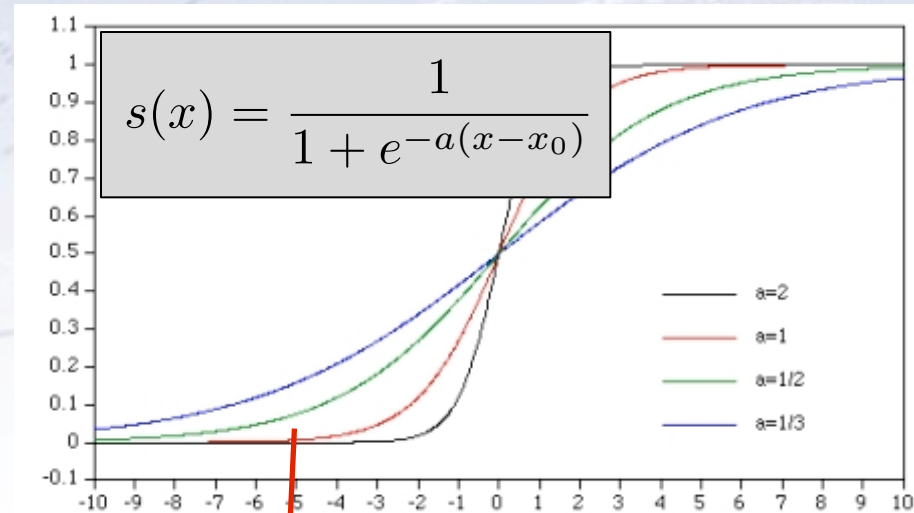
The simplest is a single layer perceptron:

$$t(x) = s\left(a_0 + \sum a_i x_i\right)$$

This can be generalised to a multilayer perceptron:

$$t(x) = s\left(a_i + \sum a_i h_i(x)\right)$$
$$h_i(x) = s\left(w_{i0} + \sum w_{ij} x_j\right)$$

Activation function can be any sigmoid function.

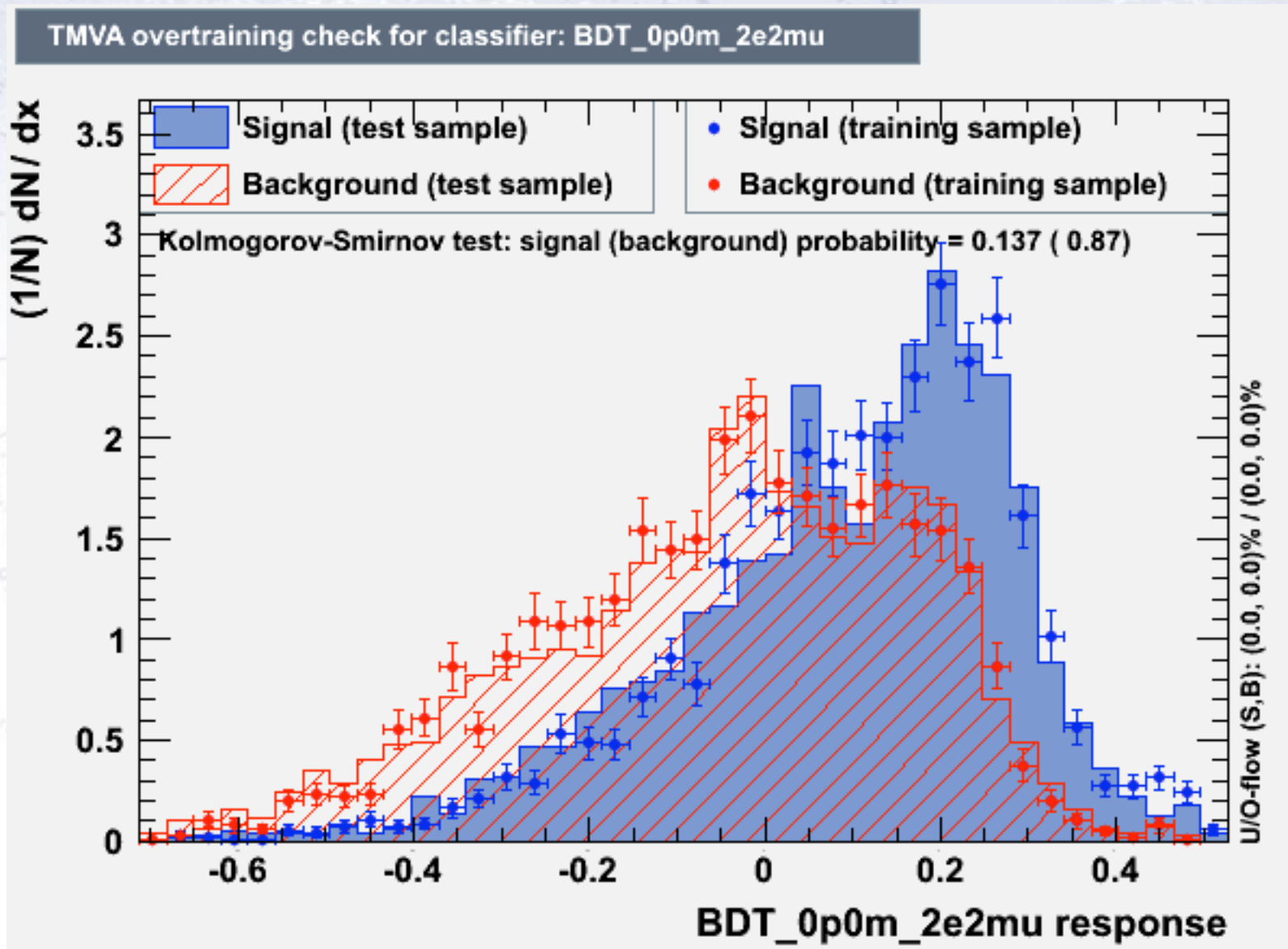


The background features a nautical chart of the Bitter End Yacht Club area. A magnetic compass rose is overlaid on the chart, showing magnetic variation. The text 'MAGNETIC' is visible on the compass rose, along with 'VAR 10°15' W'. The chart includes depth soundings and the name 'THE BITTER END YACHT CLUB'.

Training & Over-training

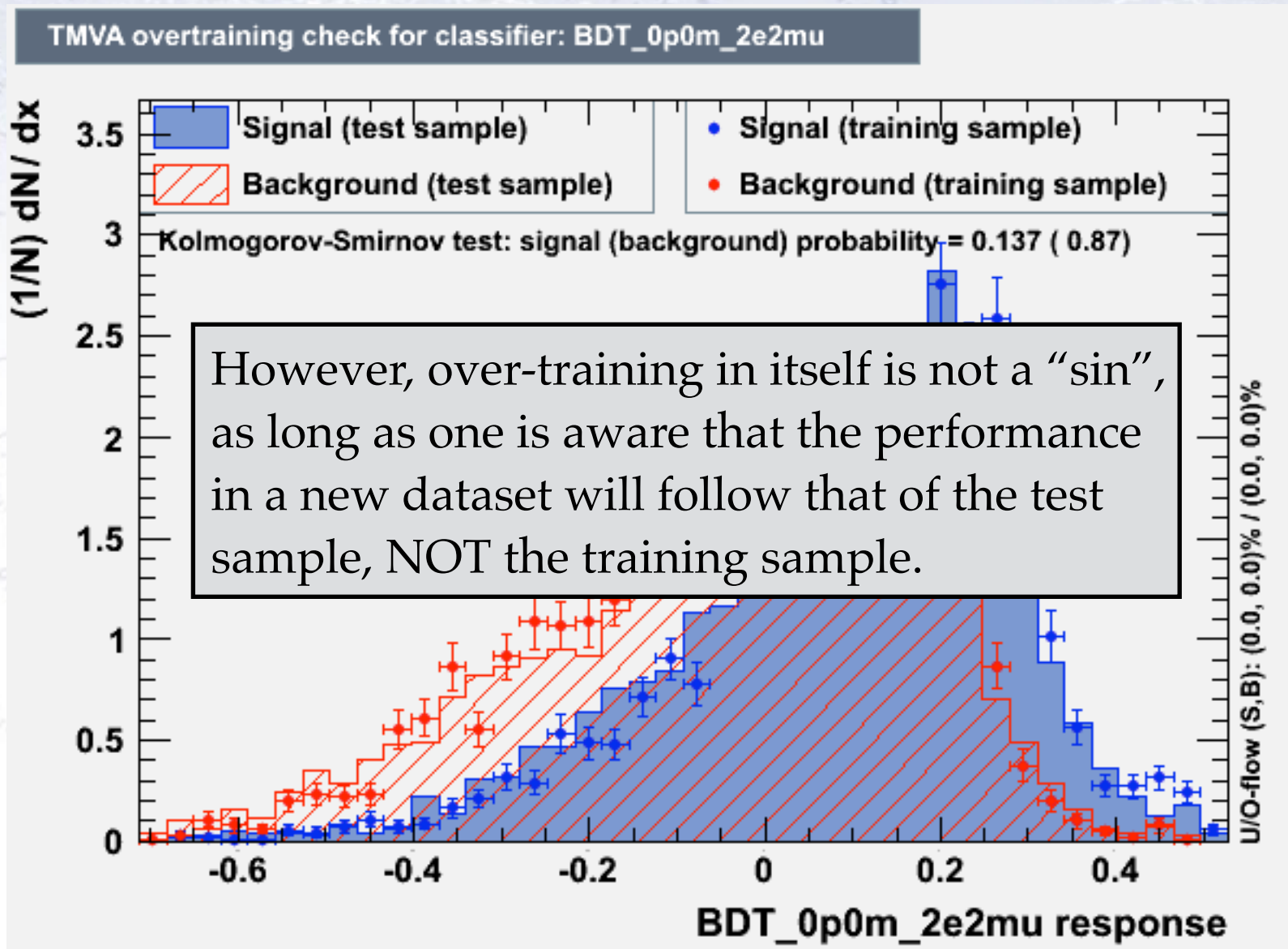
Test for simple over-training

In order to test for overtraining, half the sample is used for training, the other for testing:



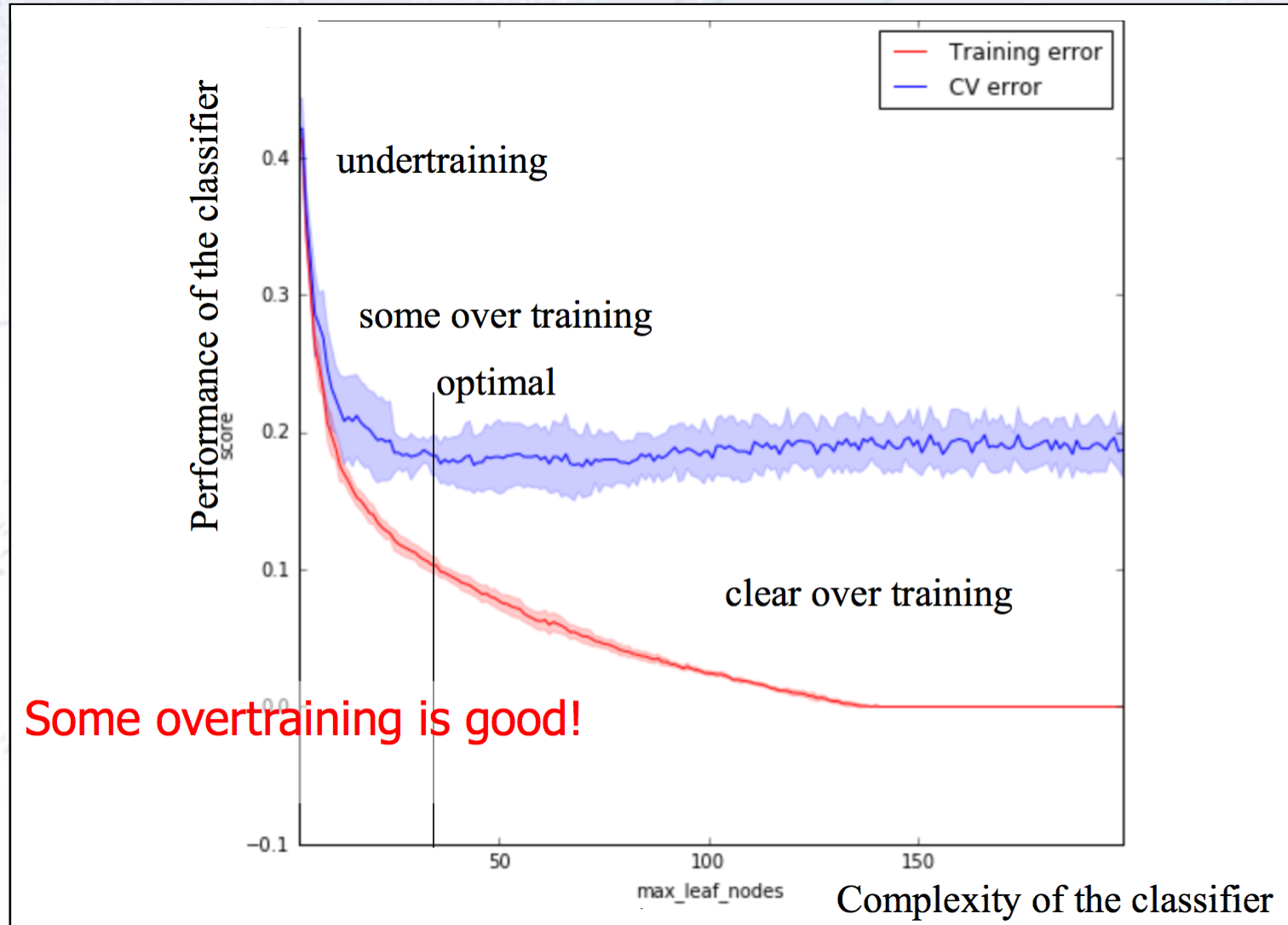
Test for simple over-training

In order to test for overtraining, half the sample is used for training, the other for testing:



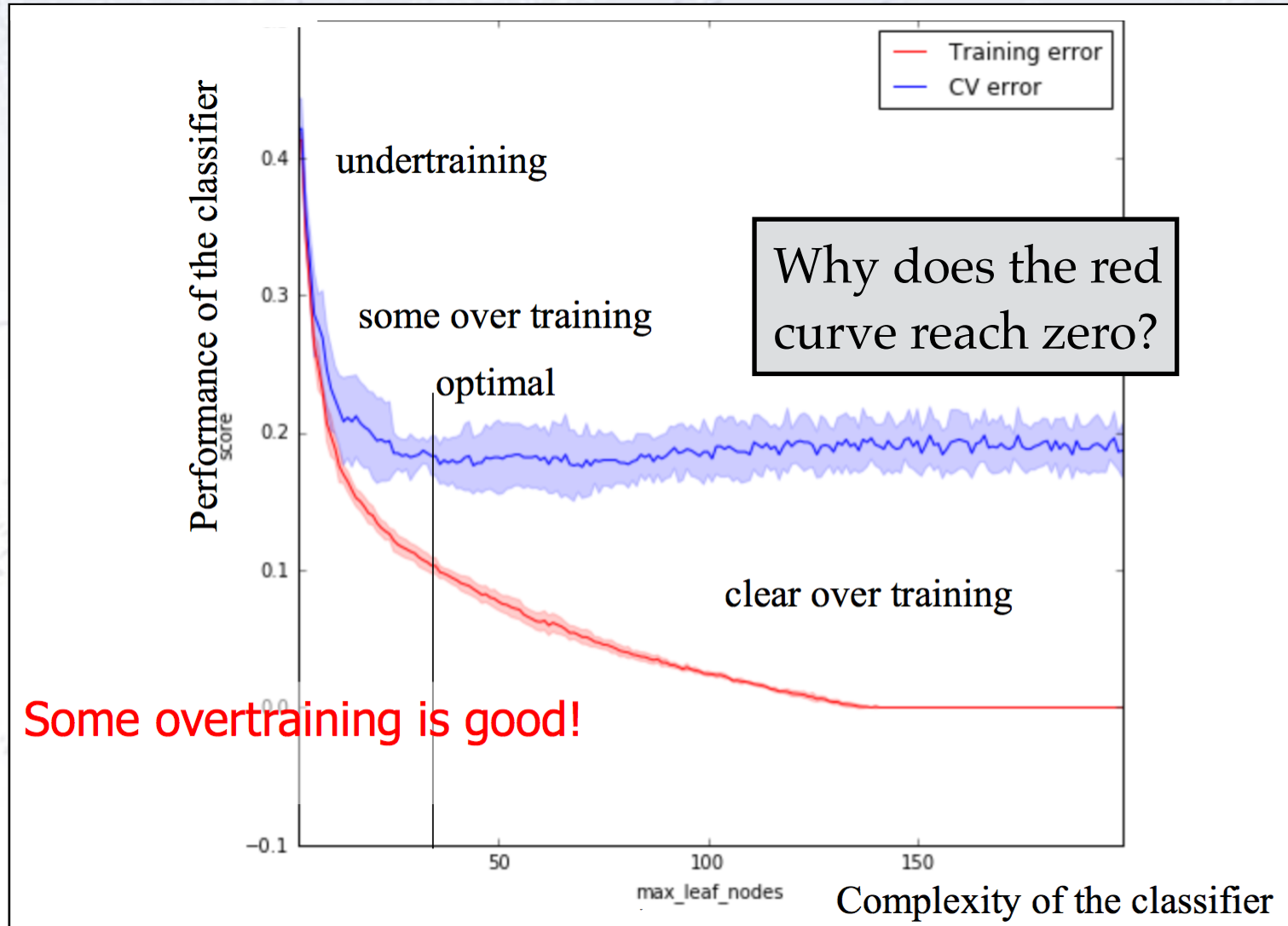
Real overtraining

The “real” limit of overtraining, is when the (Cross) Validation (CV) error starts to grow!



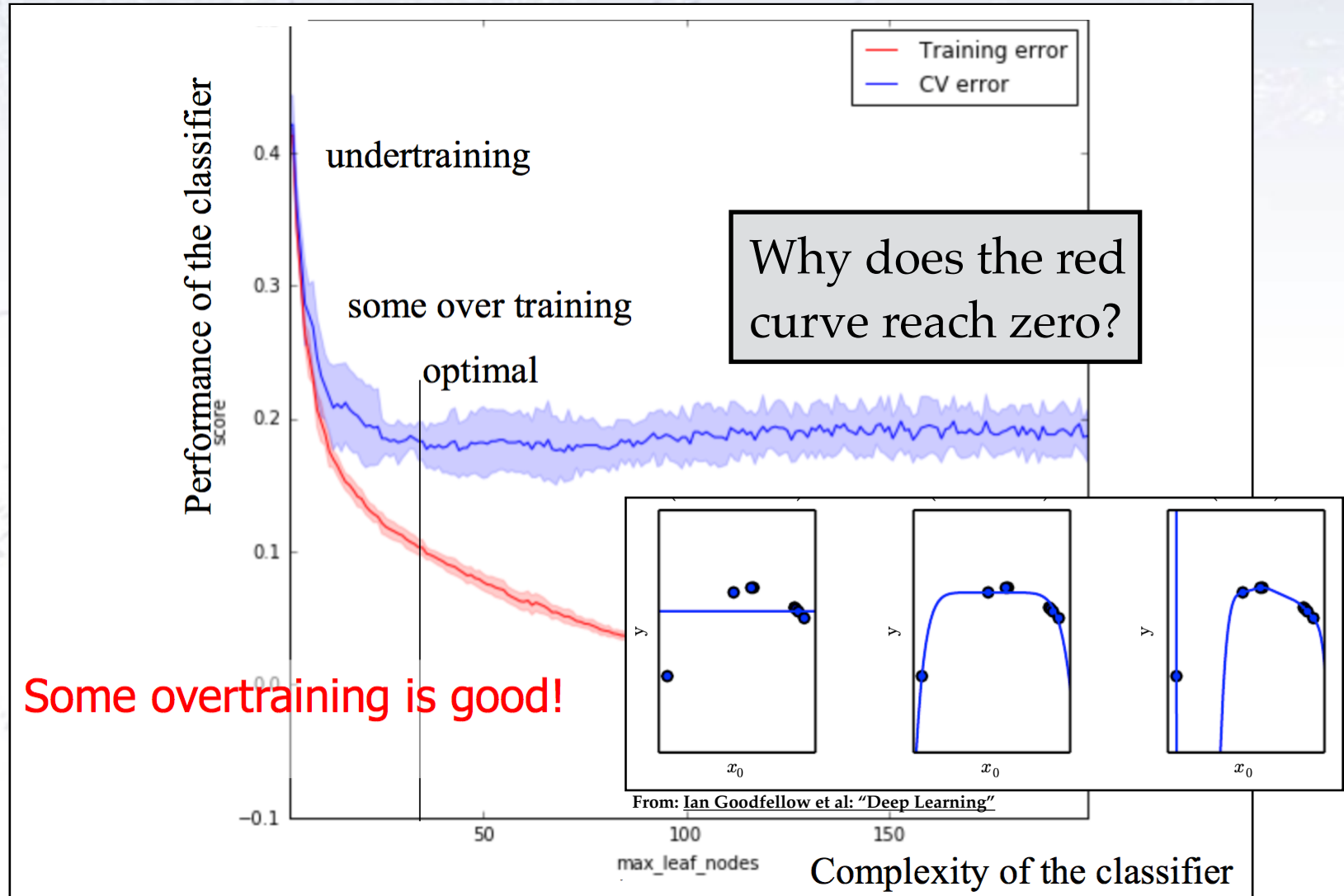
Real overtraining

The “real” limit of overtraining, is when the (Cross) Validation (CV) error starts to grow!



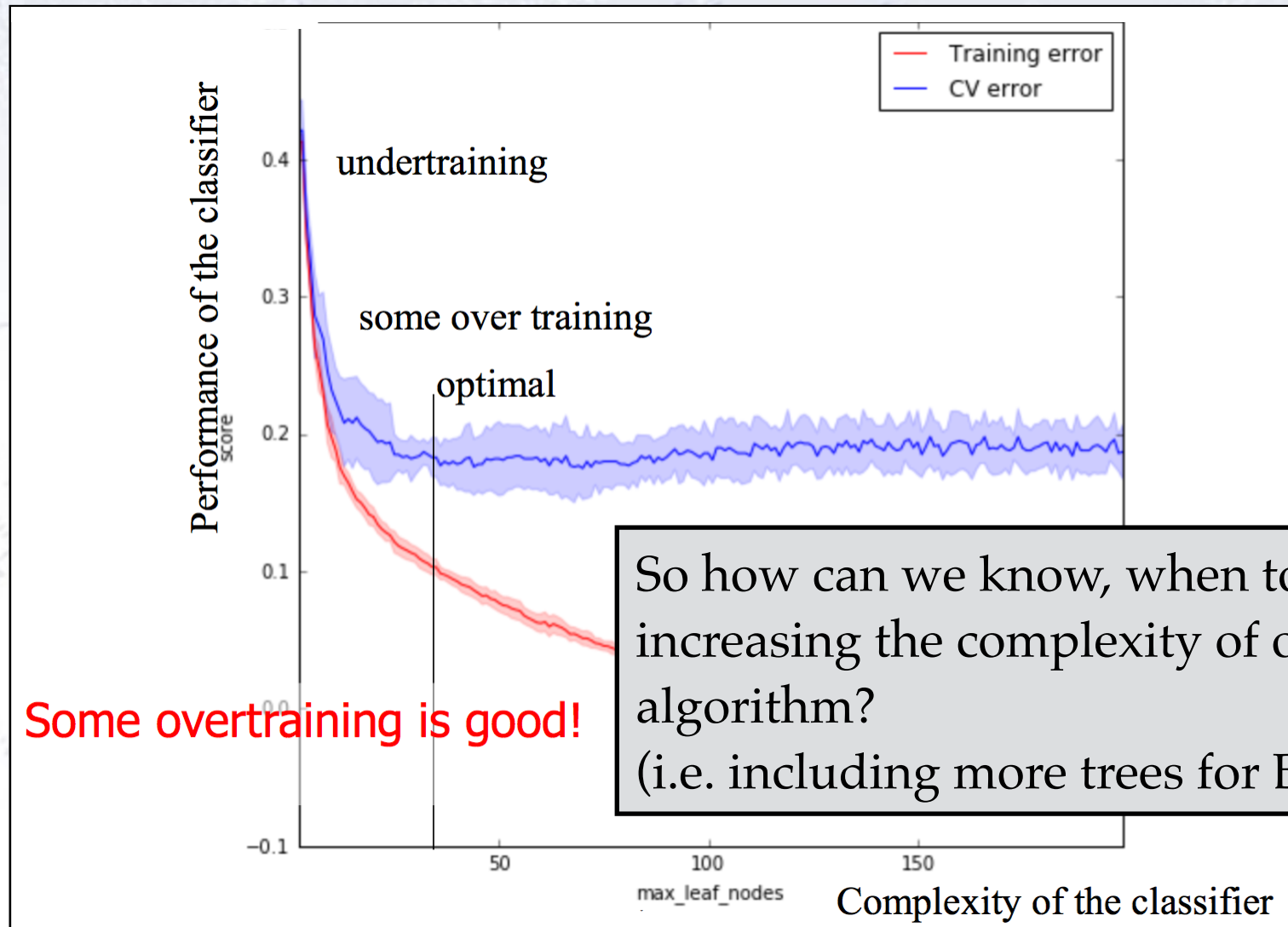
Real overtraining

The “real” limit of overtraining, is when the (Cross) Validation (CV) error starts to grow!



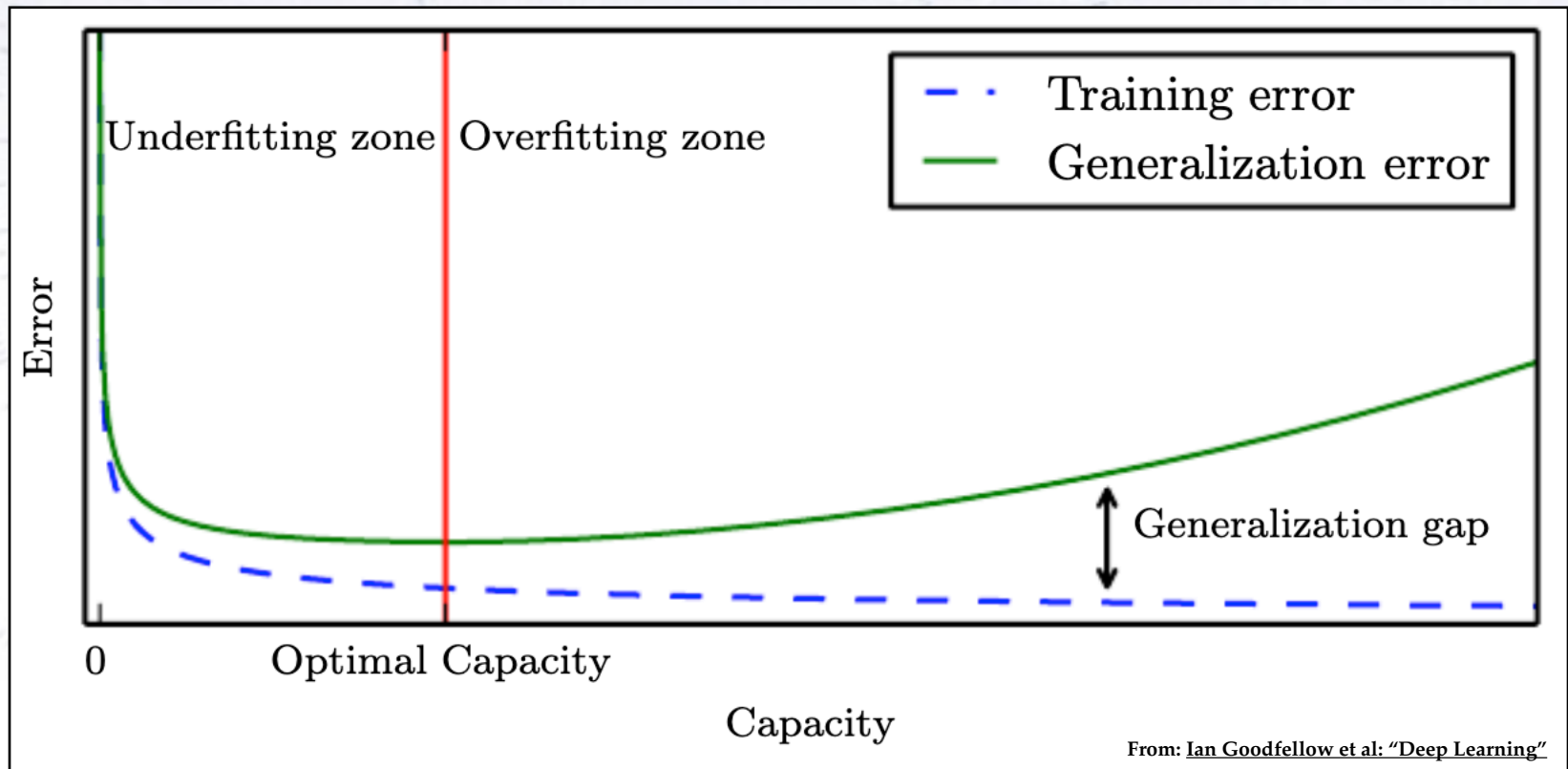
Real overtraining

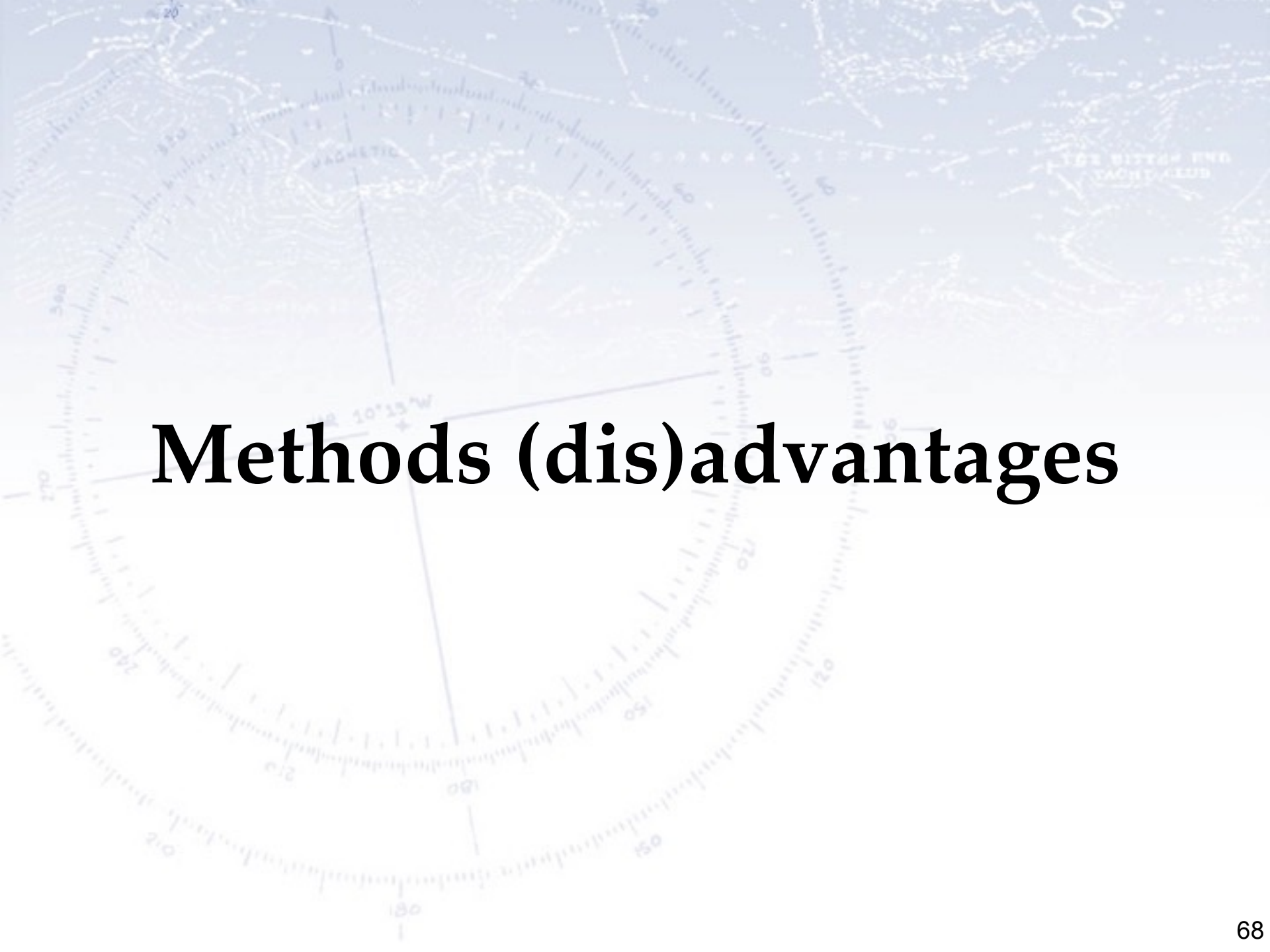
The “real” limit of overtraining, is when the (Cross) Validation (CV) error starts to grow!



Real overtraining

The “real” limit of overtraining, is when the (Cross) Validation (CV) error starts to grow!





Methods (dis)advantages

Method's (dis-)advantages

Another comparison is done in Elements of Statistical Learning II (ESL II), where linear methods are not included.

As can be seen, Neural Networks are “difficult” in almost all respects, but performant.

For trees, the case is almost the opposite.

However, I don't agree with the evaluation of the predictive power of trees.

At least not for normal structured data.

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of “mixed” type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

Method's (dis-)advantages

Another comparison is done in Elements of Statistical Learning II (ESL II), where linear methods are not included.

As can be seen, Neural Networks are “difficult” in almost all respects, but performant.

For trees, the case is almost the opposite.

However, I don't agree with the evaluation of the predictive power of trees.

At least not for normal structured data.

For tabular data, I disagree!

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of “mixed” type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

Method's (dis-)advantages

Another comparison is done in Elements of Statistical Learning II (ESL II), where linear methods are not included.

As can be seen, Neural Networks are “difficult” in almost all respects, but performant.

For trees, the case is almost the opposite.

However, I don't agree with the evaluation of the predictive power of trees.

At least not for normal structured data.

For tabular data, I disagree!

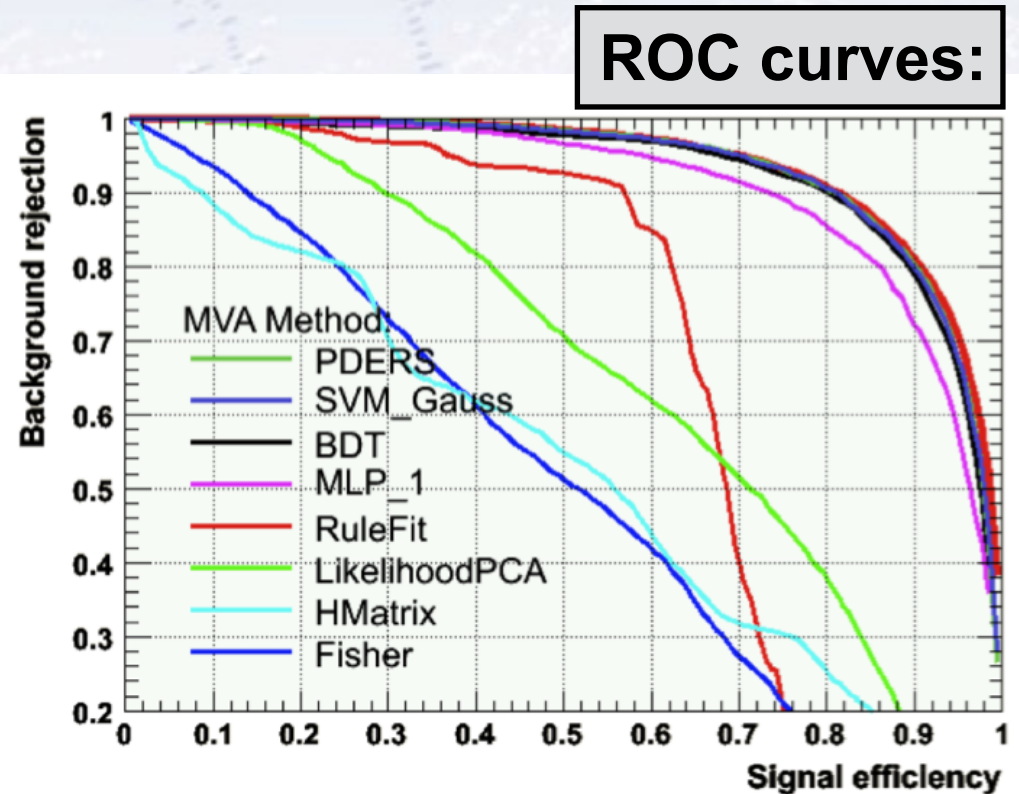
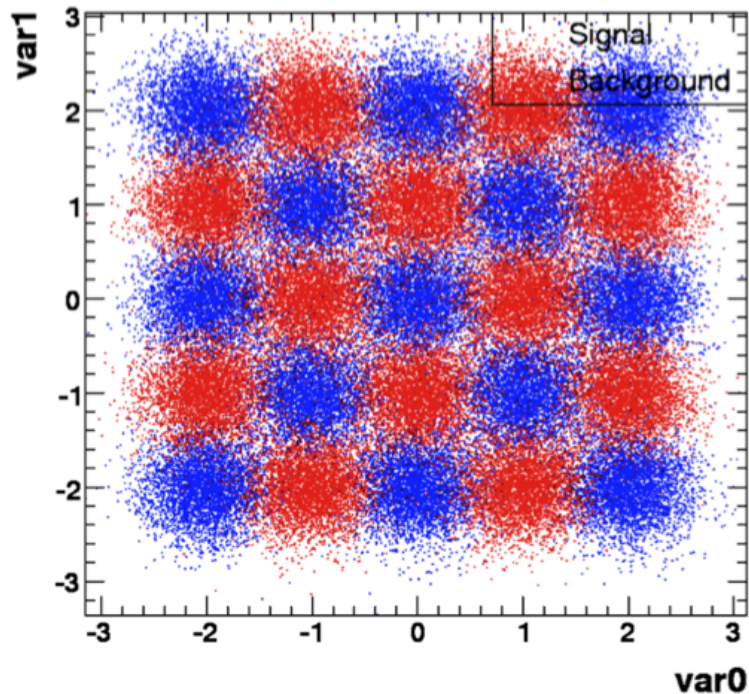
Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of “mixed” type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

...and others do too [<https://arxiv.org/abs/2110.01889>]

From ESL II, Chapter 10.7

Performance comparison

Left figure shows the distribution of signal and background used for test.
Right figure shows the resulting separation using various MVA methods.



The theoretical limit is known from the Neyman-Pearson lemma using the (known/correct) PDFs in a likelihood.

In all fairness, this is a case that is great for the BDT...

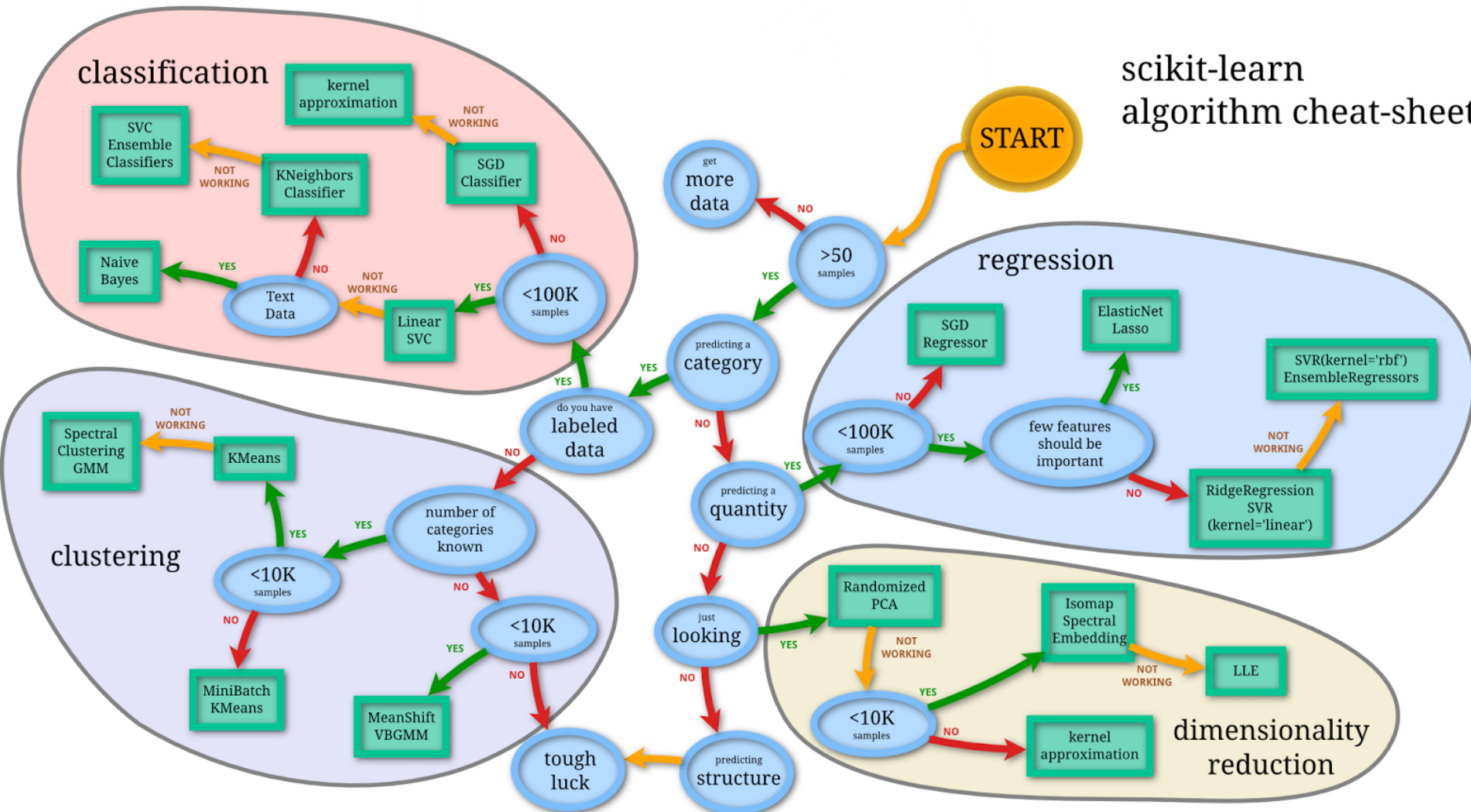


How to choose method?

Which method to use?

There is no good / simple answer to this, though people have tried, e.g.:

scikit-learn
algorithm cheat-sheet



Which method to use?

There is no good / simple answer to this, though people have tried, e.g.:

scikit-learn
algorithm cheat-sheet

