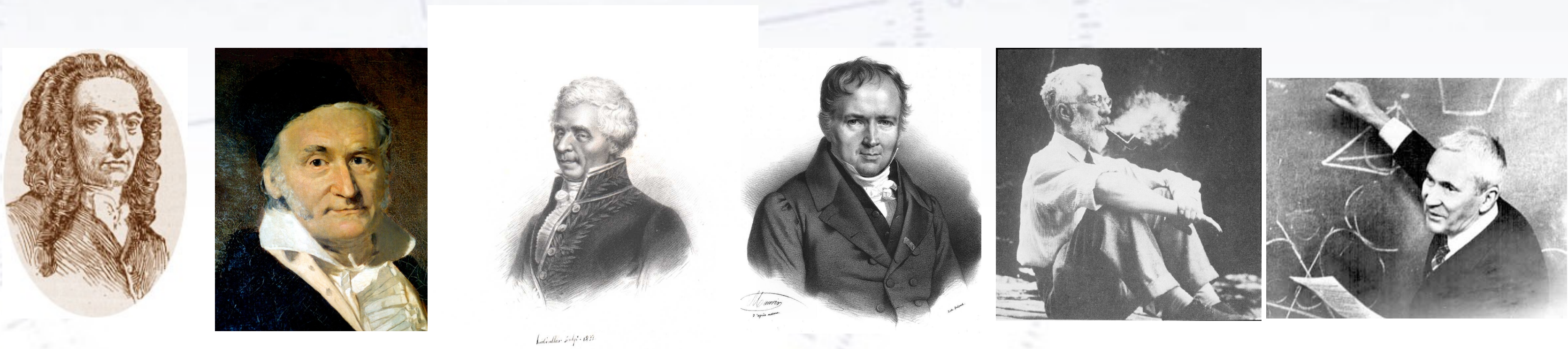
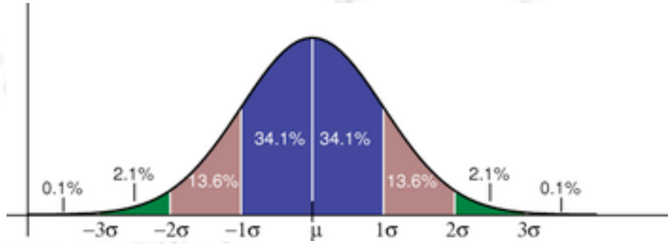
Applied Statistics

Multivariate analysis

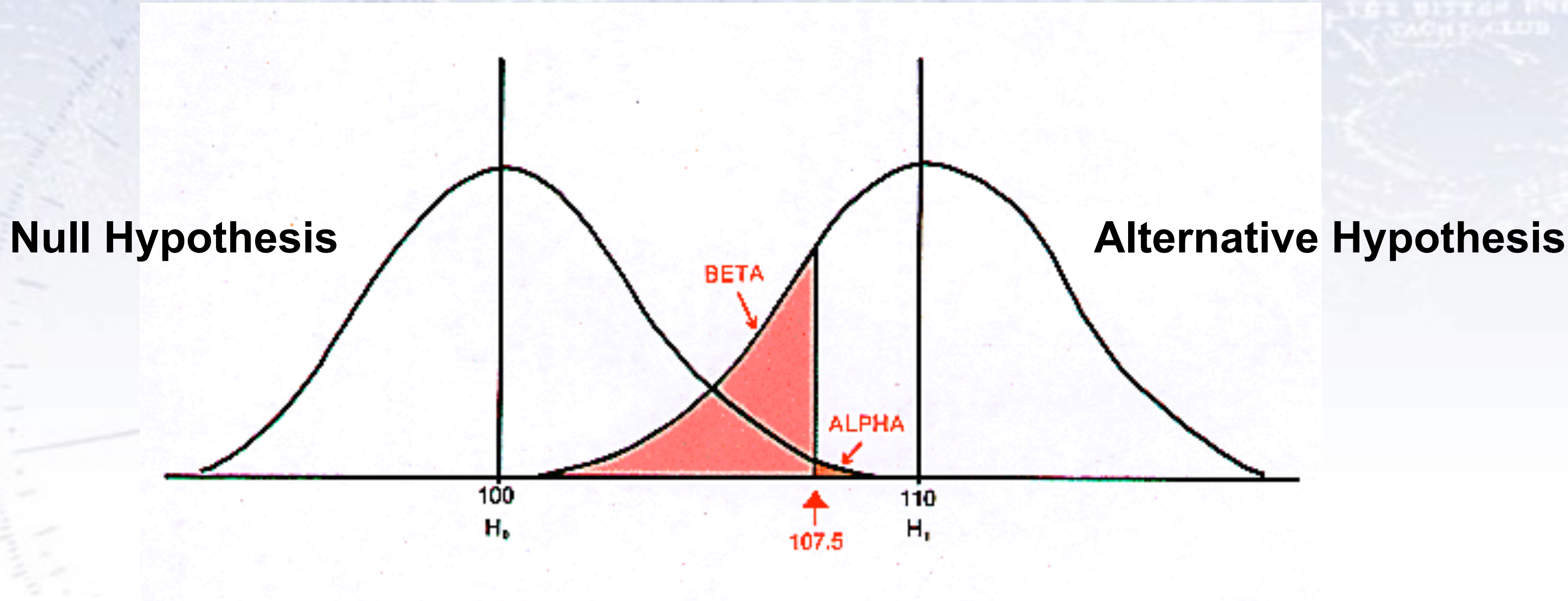


Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

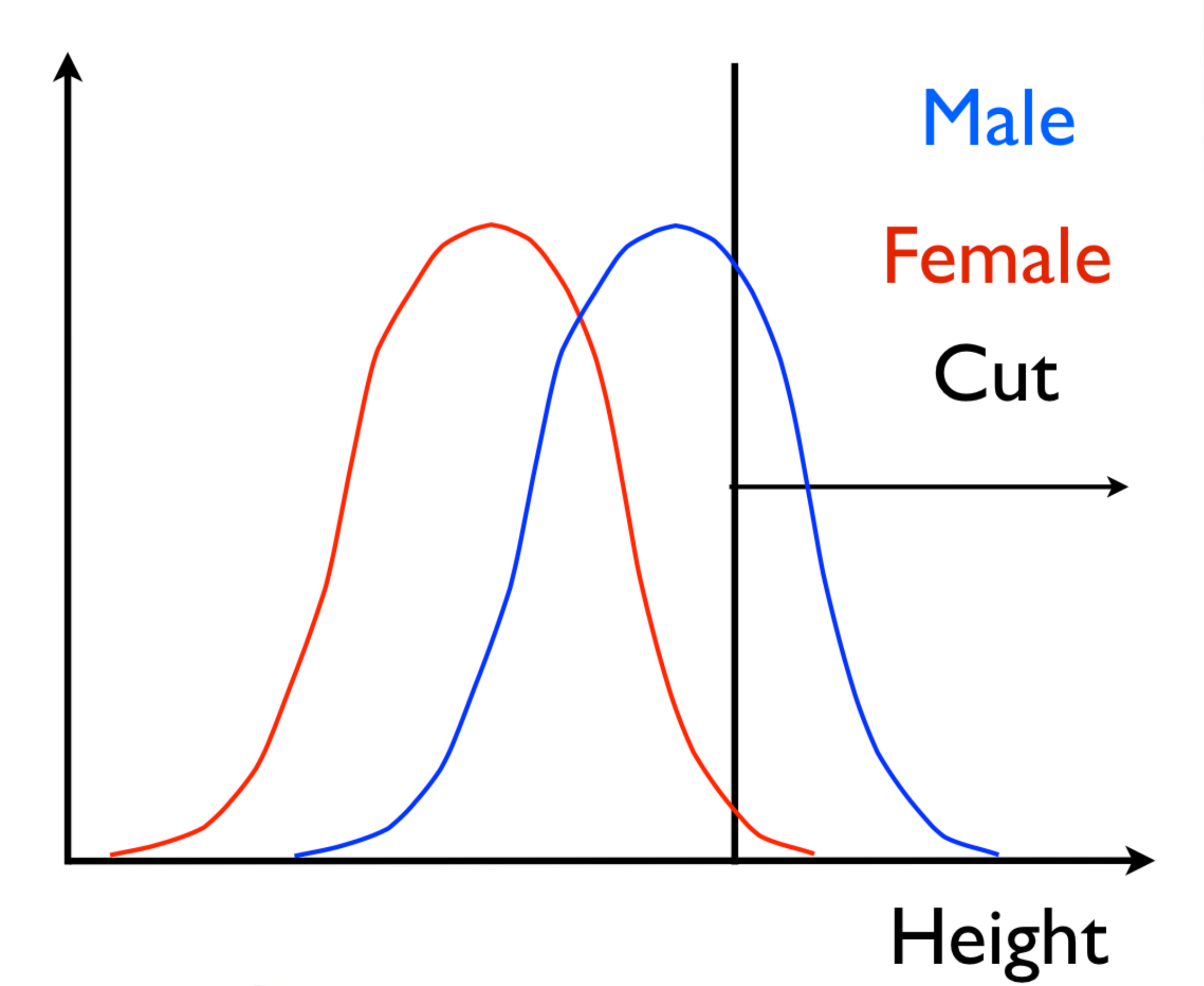
Separating hypothesis



		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

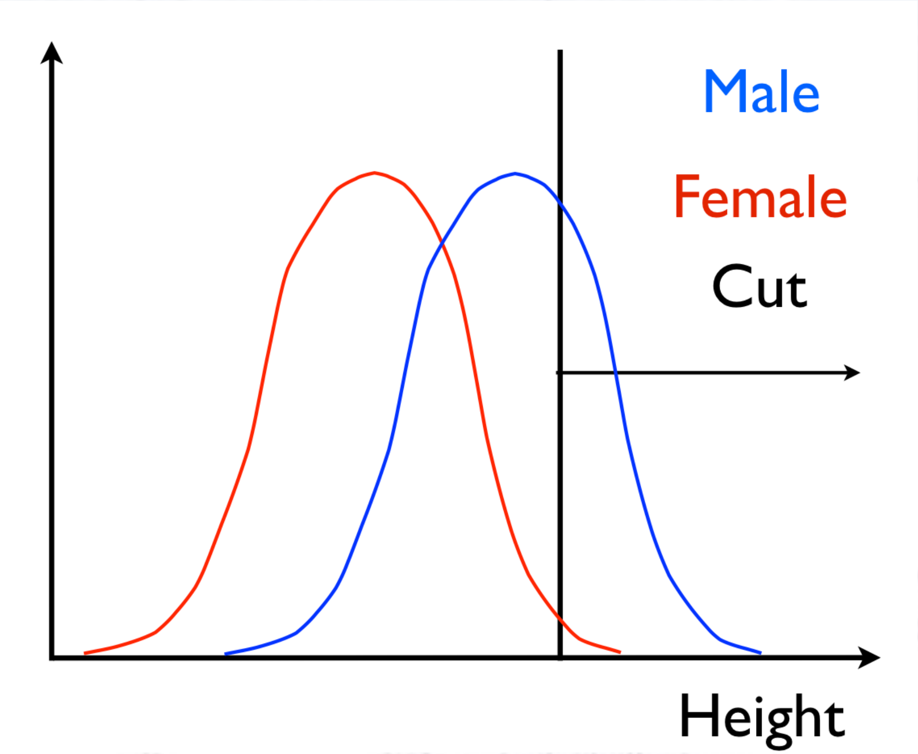
Simple Example

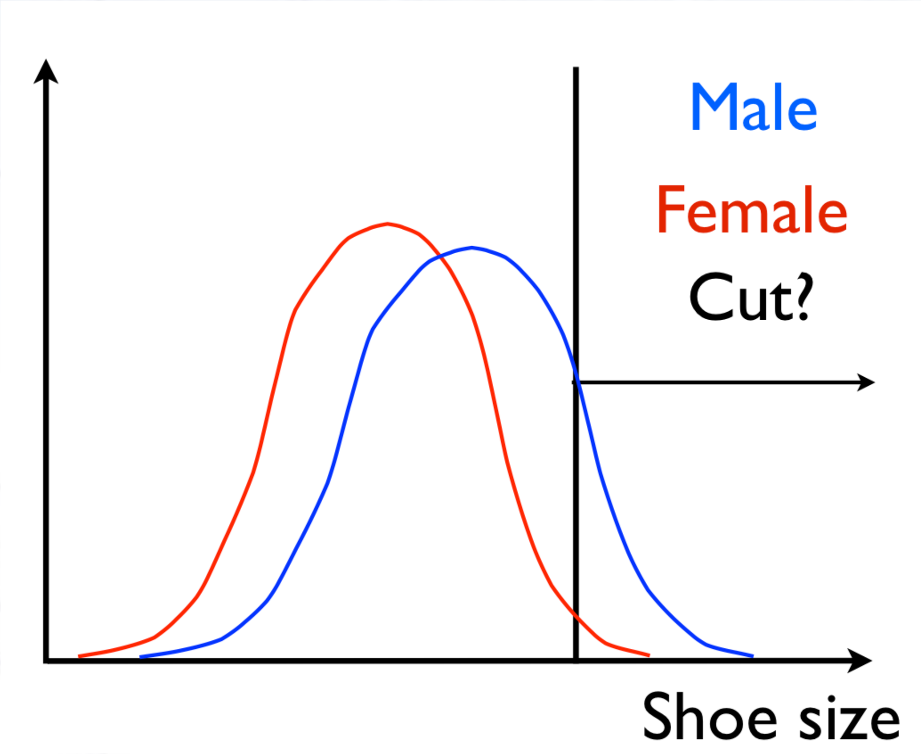
Problem: You want to figure out a method for getting sample that is 95% male!
Solution: Gather height data from 10000 people, Estimate cut with 95% purity!



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!

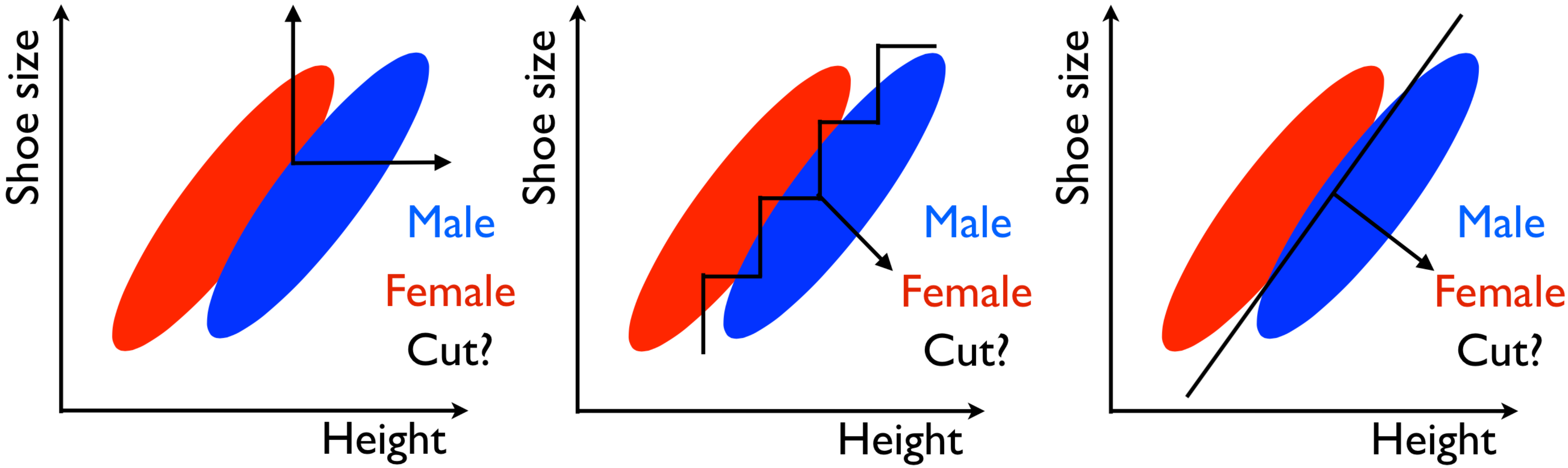
Simple Example

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!

Separating data

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

180 MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
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 Discriminant

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

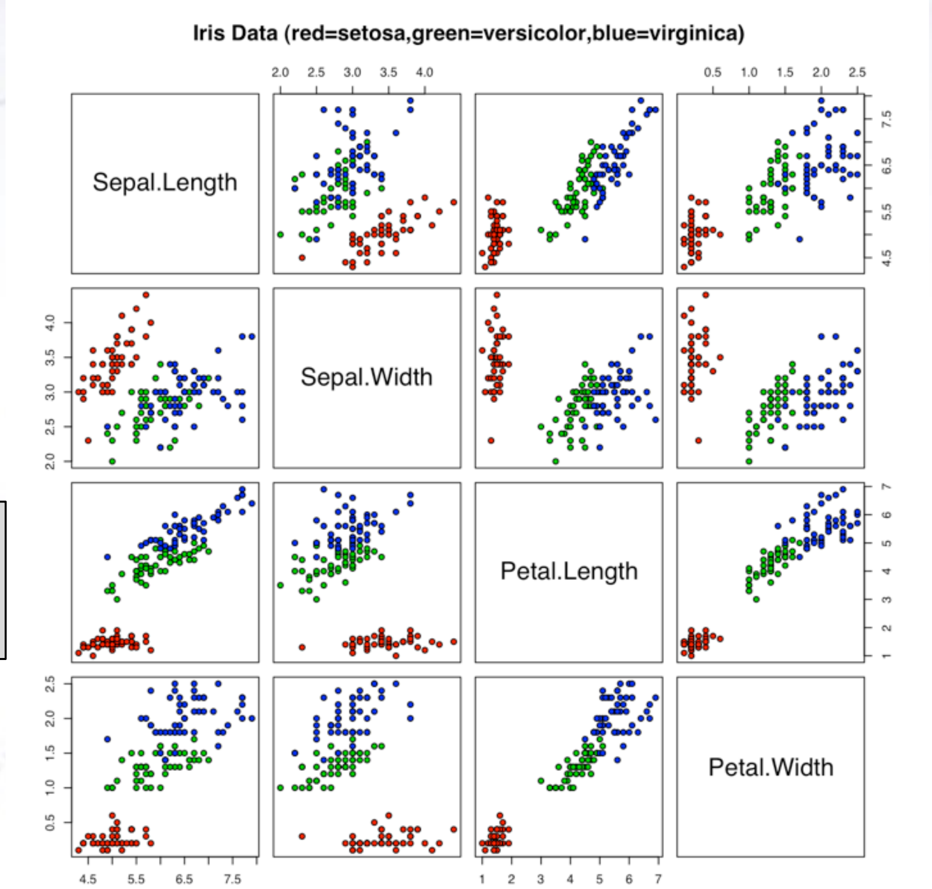
Q: How to combine the variables?
A: Use the Fisher Discriminant:

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

Q: How to choose the values of w?
A: Inverting the covariance matrices:

$$\boxed{\vec{w} = (\mathbf{\Sigma}_A + \mathbf{\Sigma}_B)^{-1} \;\; (\vec{\mu}_A - \vec{\mu}_B)}$$

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



Fisher Discriminant

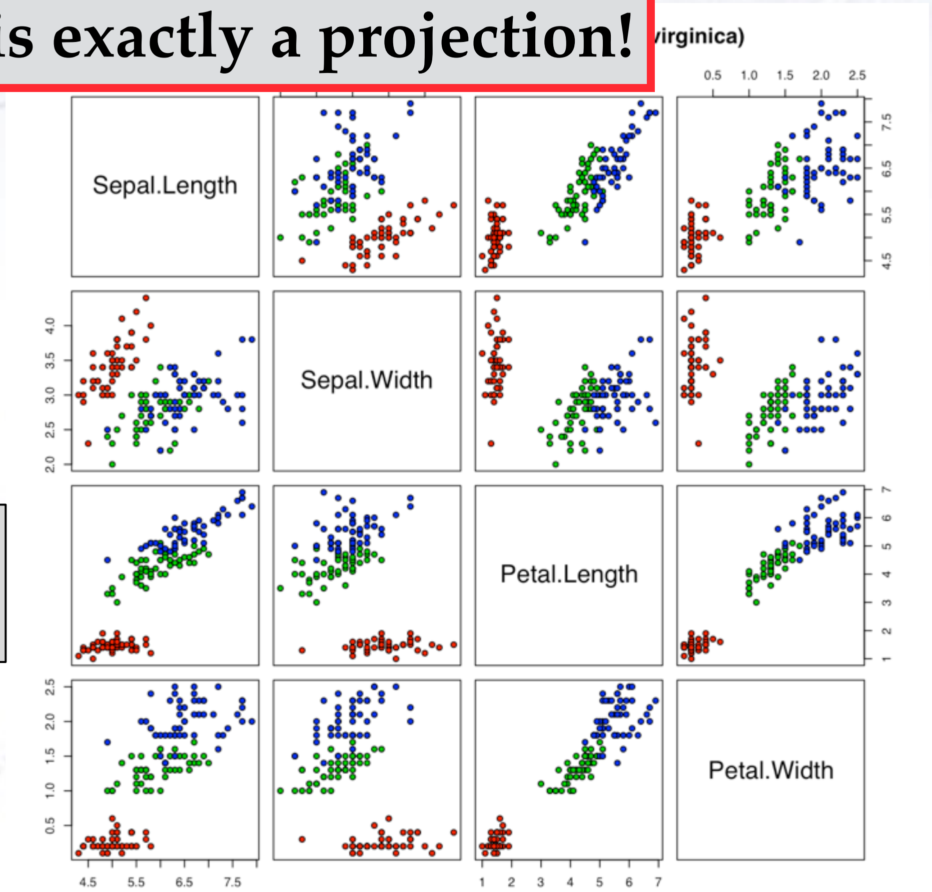
You want to separate two types/classes (A and B) of events using several measurements.

Q: How to combine the variable

A: Use the Fisher Discriminant:

This is exactly a projection!

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



Q: How to choose the values of w?

A: Inverting the covariance matrices:

$$\vec{w} = (\mathbf{\Sigma}_A + \mathbf{\Sigma}_B)^{-1} \; (\vec{\mu}_A - \vec{\mu}_B)$$

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

Fisher Discriminant

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

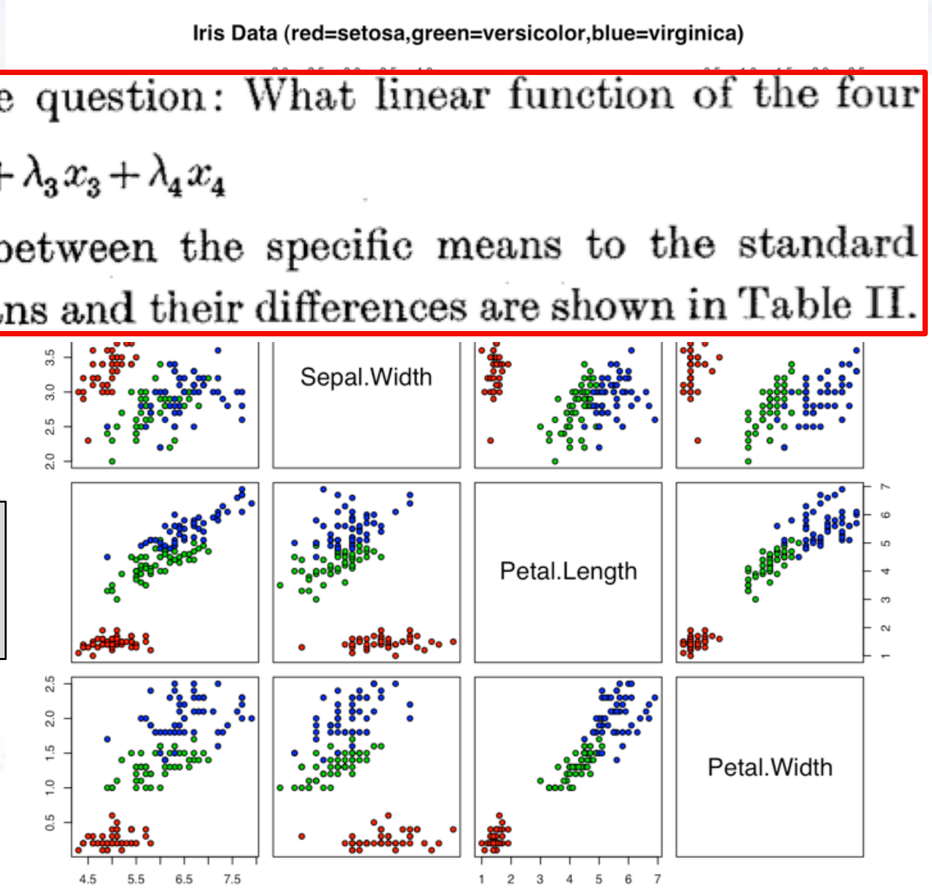
Q: How to combine the variables?

A: Use the Fisher Discriminant:

ments 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?

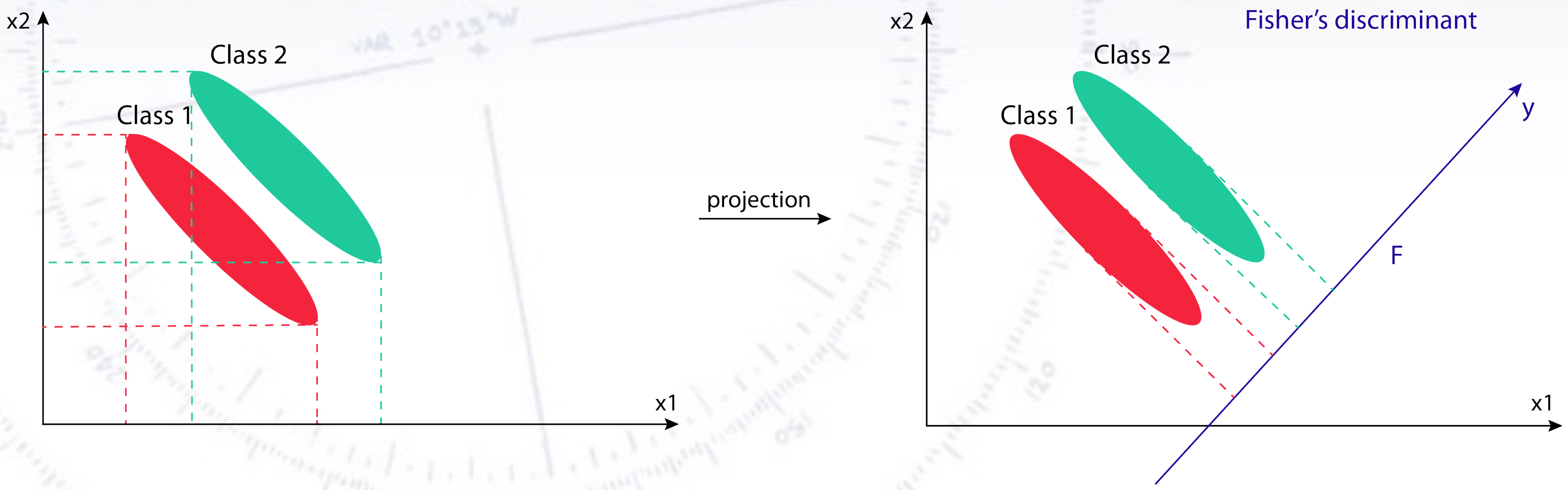
A: Inverting the covariance matrices:

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

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

Fisher Discriminant

<u>Executive summary:</u>
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 Discriminant

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

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

Fisher 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.

$$\vec{w} = (\mathbf{\Sigma}_A + \mathbf{\Sigma}_B)^{-1} \; (\vec{\mu}_A - \vec{\mu}_B)$$

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

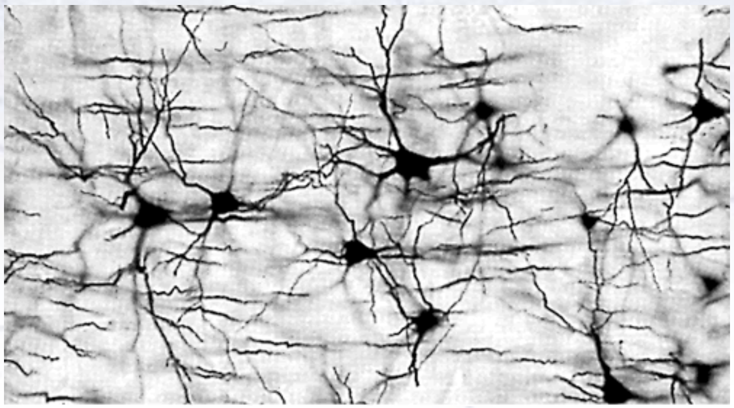
Given weights (w), you take your input variables (x) and combine them linearly as follows:

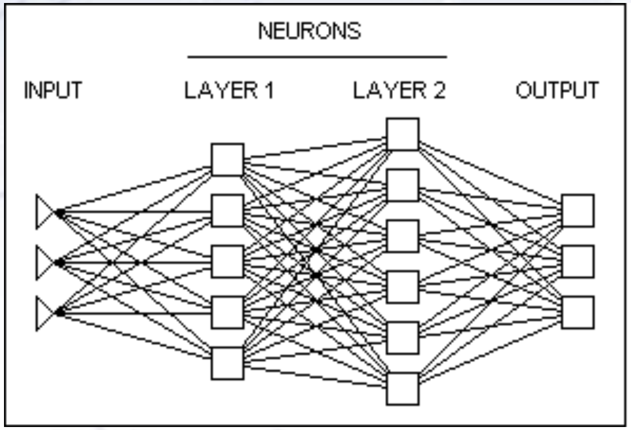
$$\mathcal{F} = w_0 + \vec{w} \cdot \vec{x}$$

F is what you base your decision on.

Data Mining

Seeing patterns in data and using it!

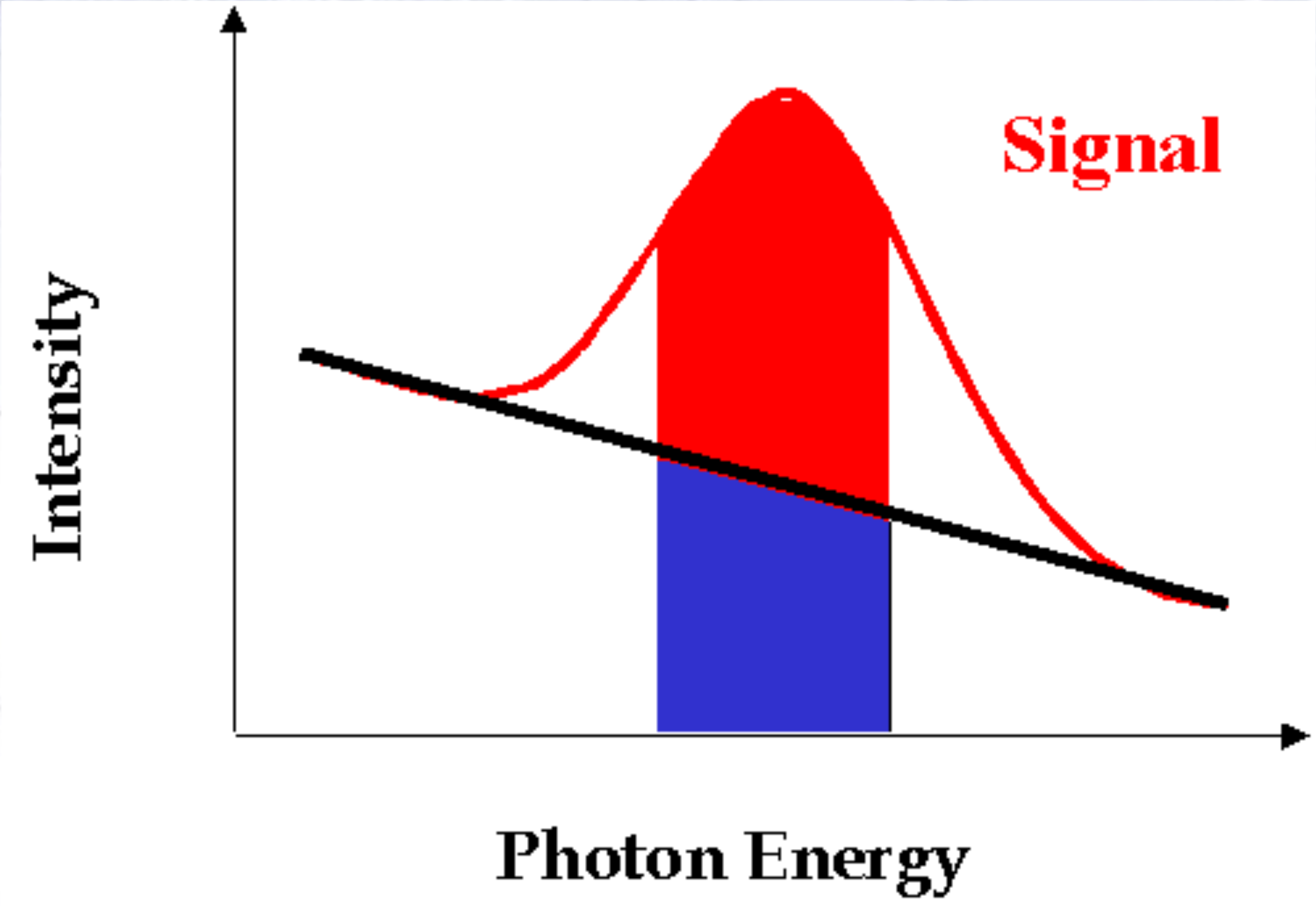




Data mining** is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and **scientific discovery.

[Wikipedia, Introduction to Data Mining]

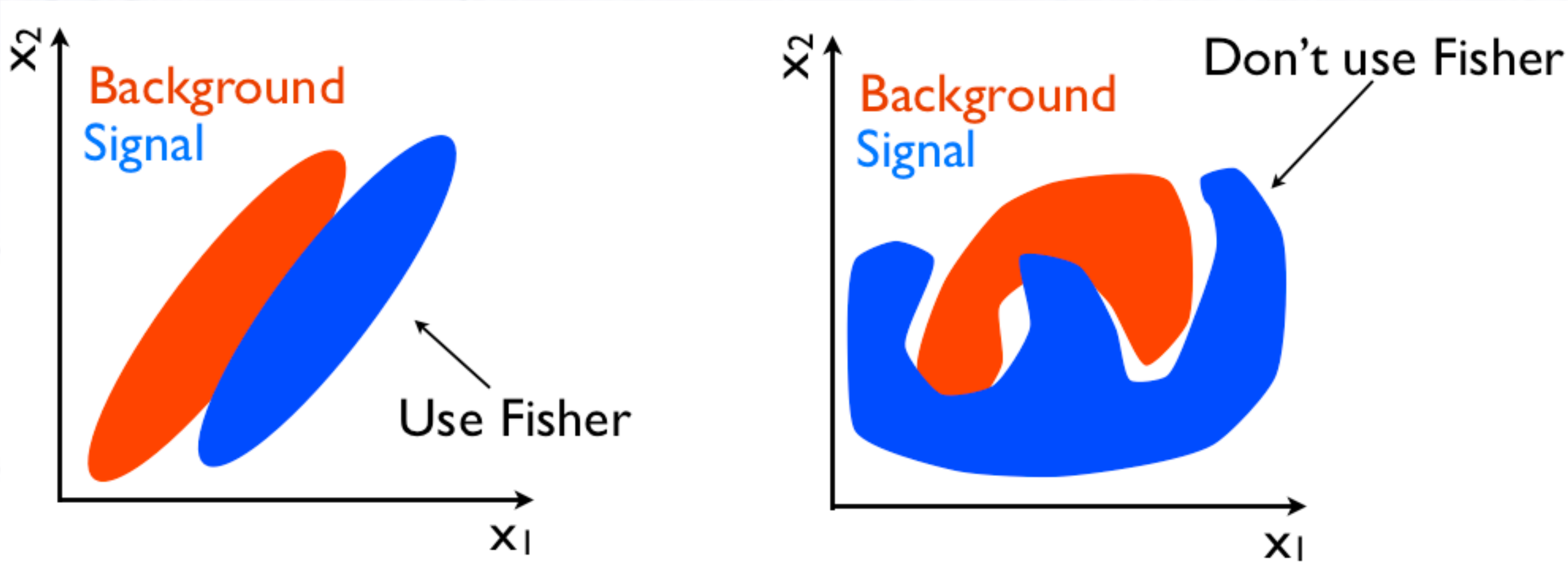
Cuts



Classical case (signal peak on background)…

Cuts – in 2 dimensions

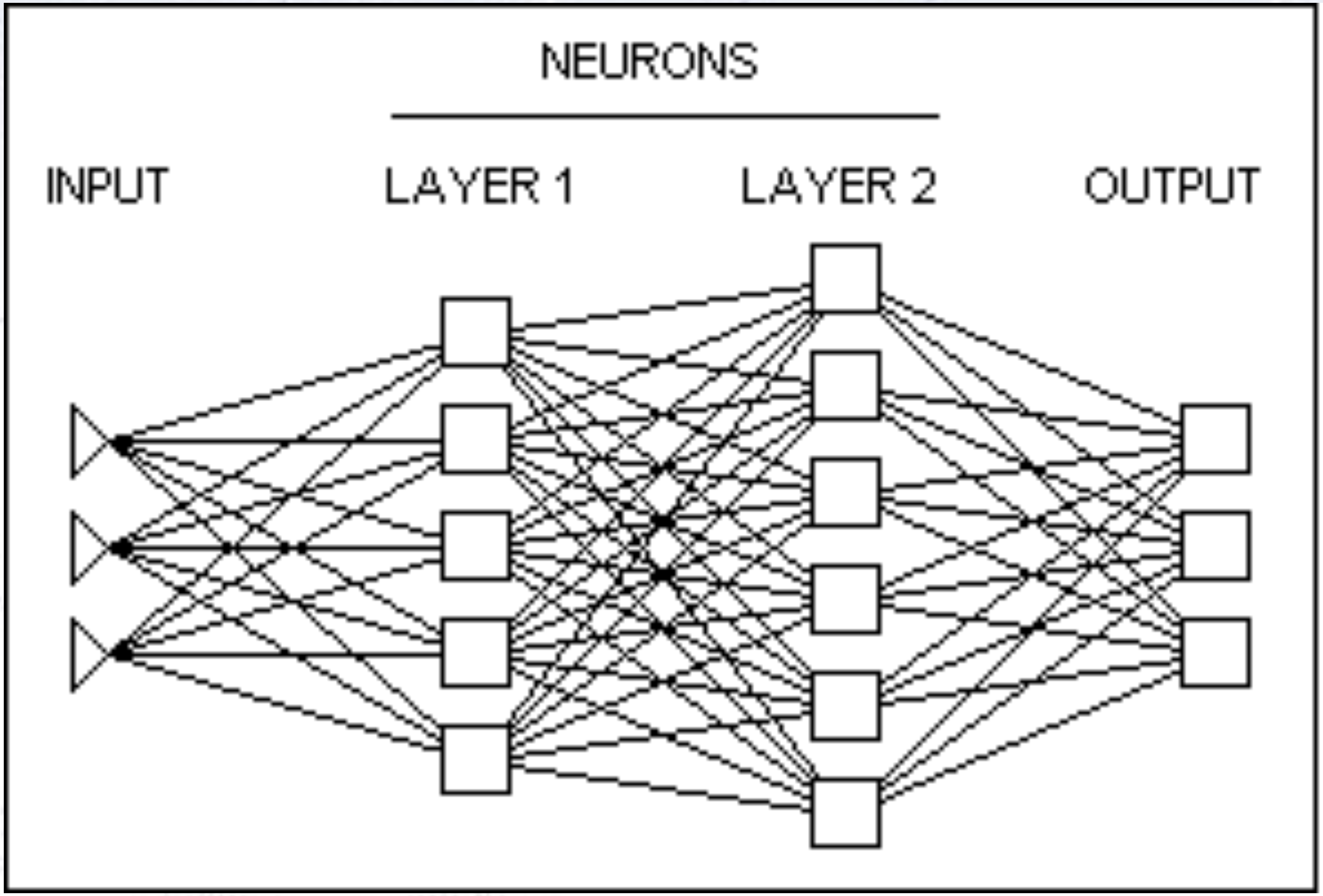
Now let's try in 2 dimensions (2 cases):



Not as simple as the 1 dimensional case!

Correlations now has to be taken into account.

Neural Networks

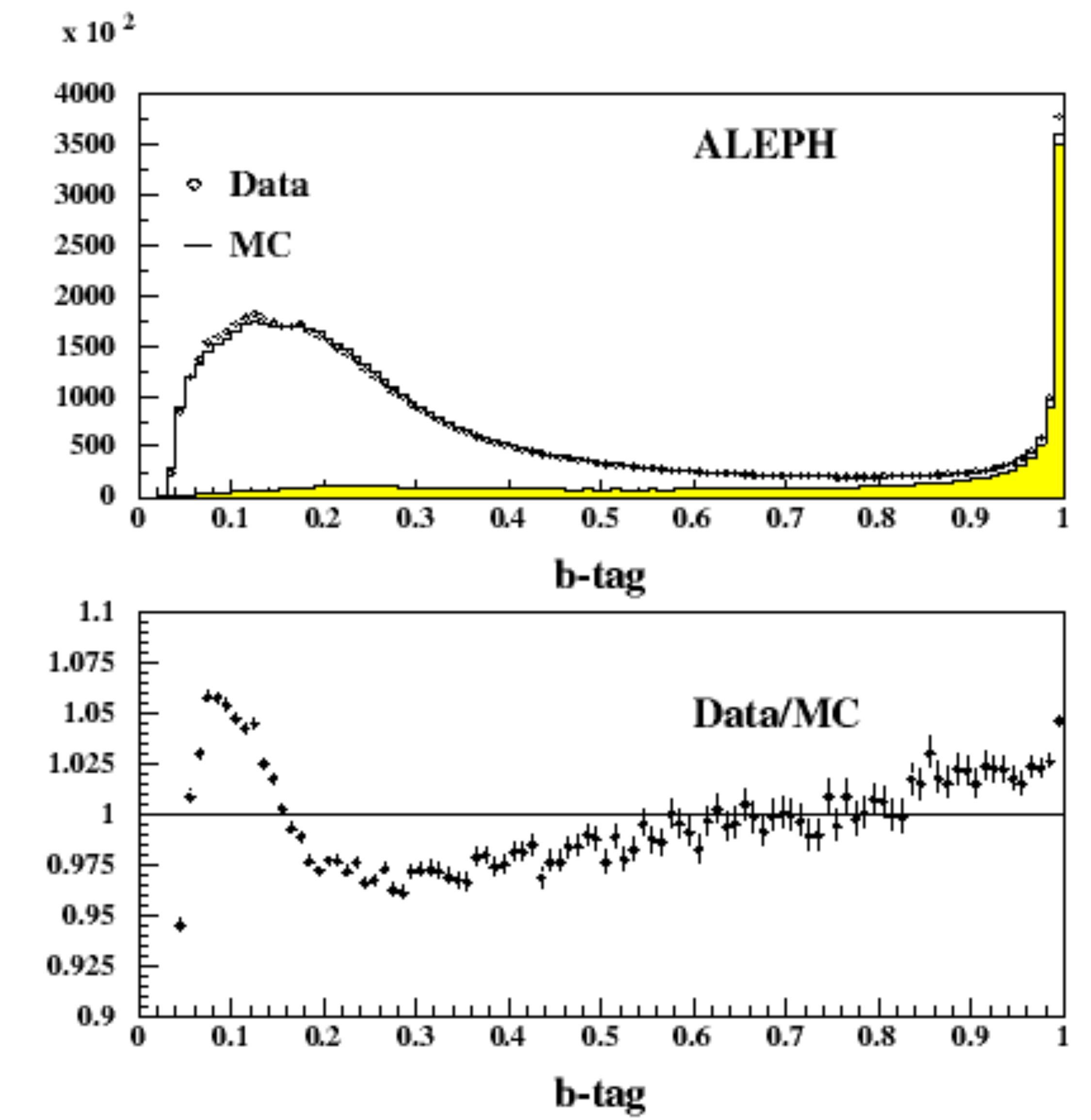


Neural Networks

An example from CERN is the ALEPH collaboration at LEP.

Used to determine if a jet is from a b-quark or not.

Very large statistics, and of very great importance.



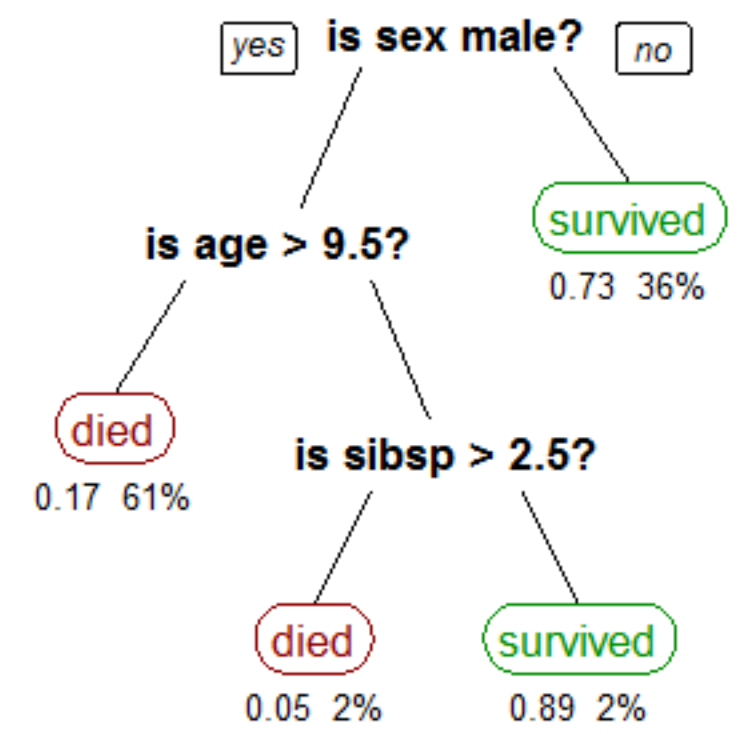
Boosted Decision Trees

Can become very complex.

Good for discrete problems.

Not always as efficient.

Boosting adds to separation.



Advanced MVA…
…more on Monday