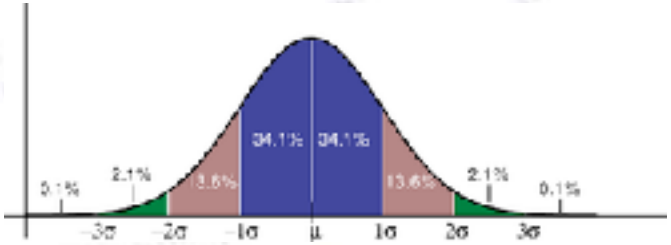
Applied Statistics

Multivariate analysis



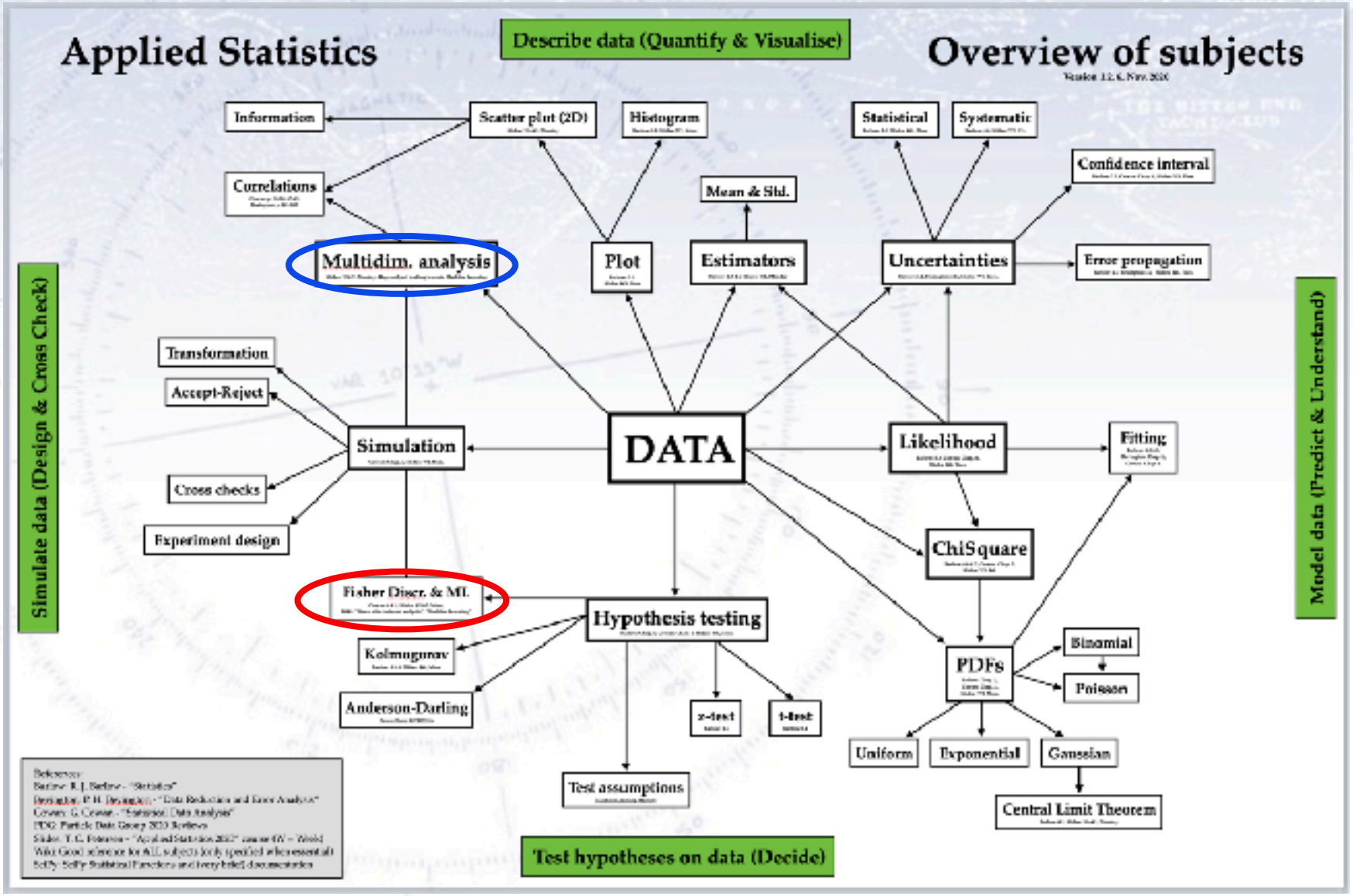



Troels C. Petersen (NBI)

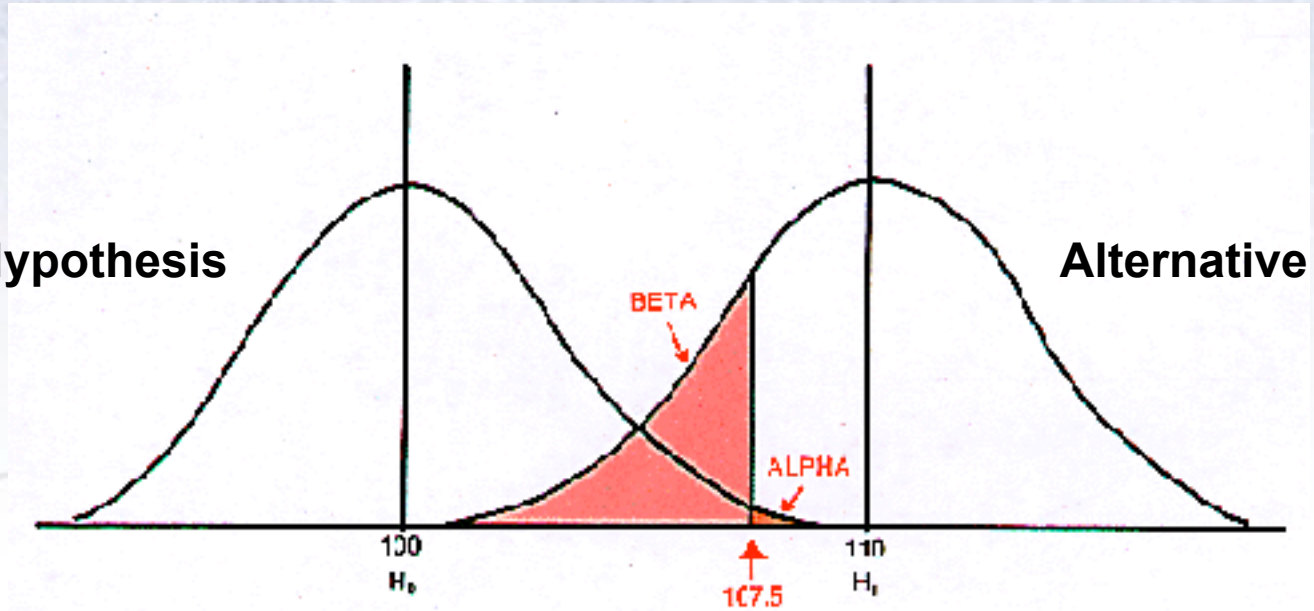


"Statistics is merely a quantisation of common sense"

MultiVariate Analysis & Fisher Discr.



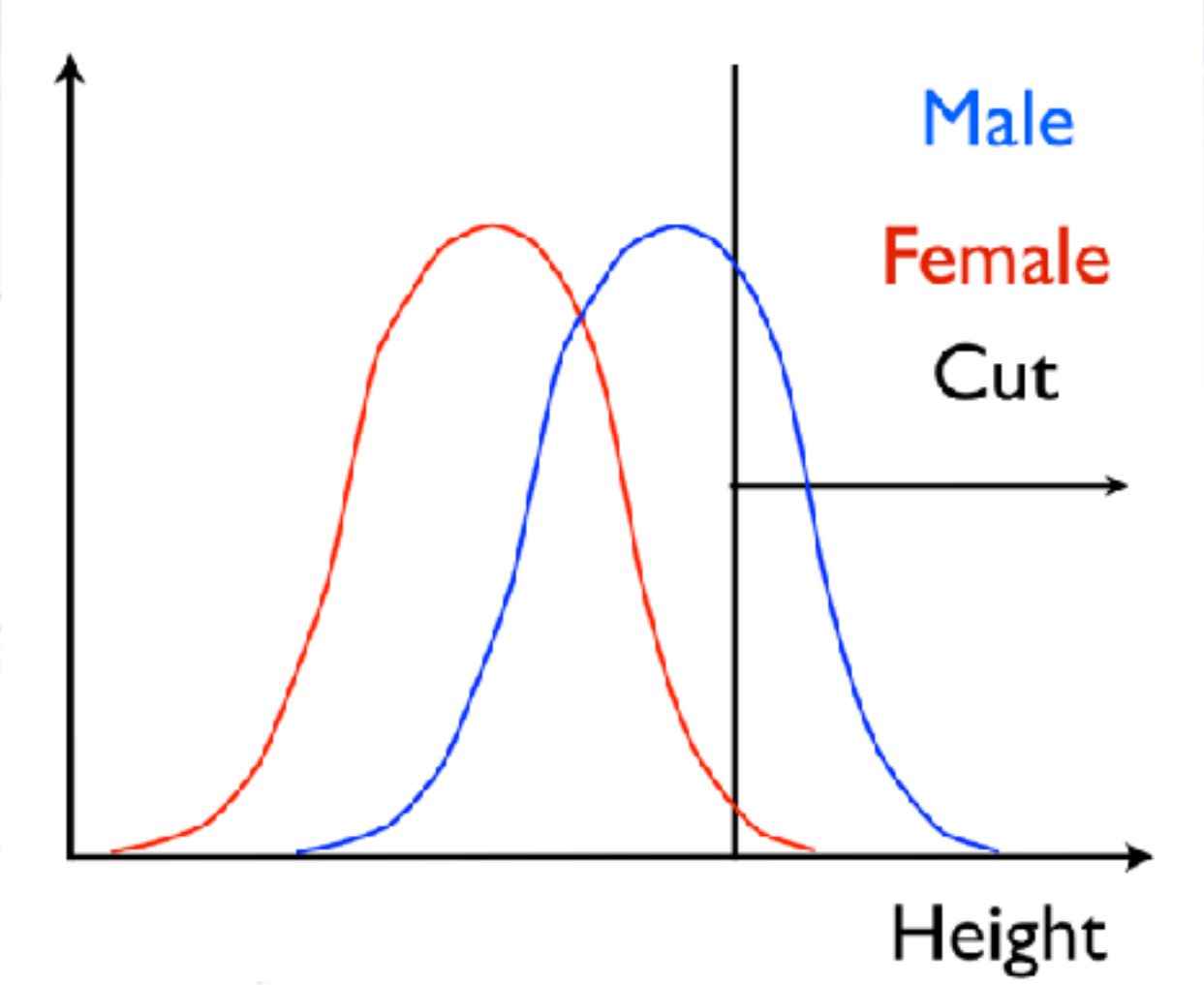
Separating hypothesis



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

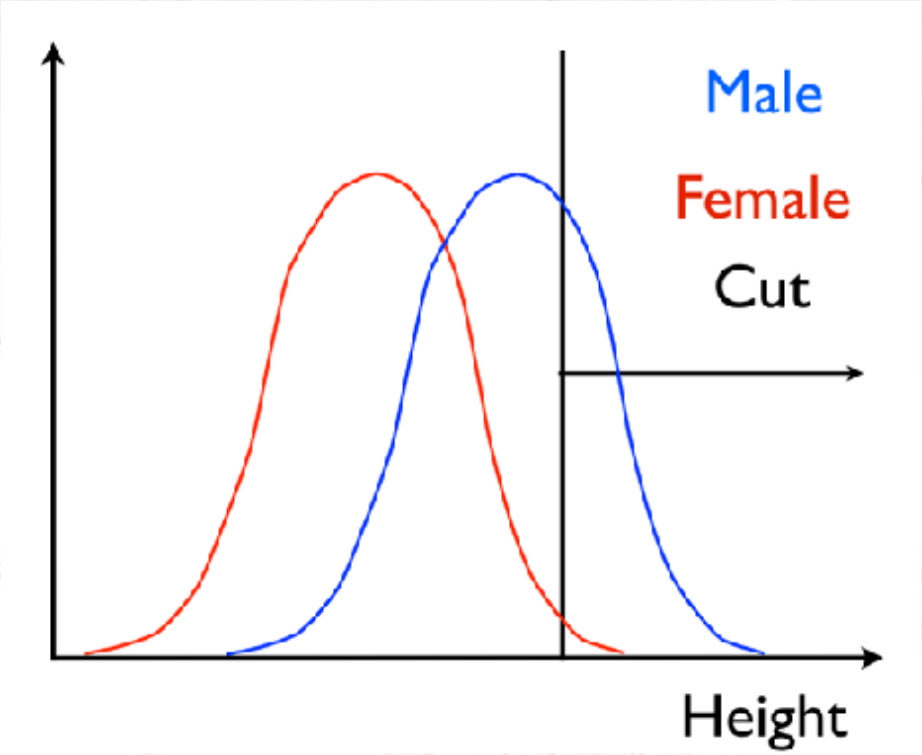
Simple Example

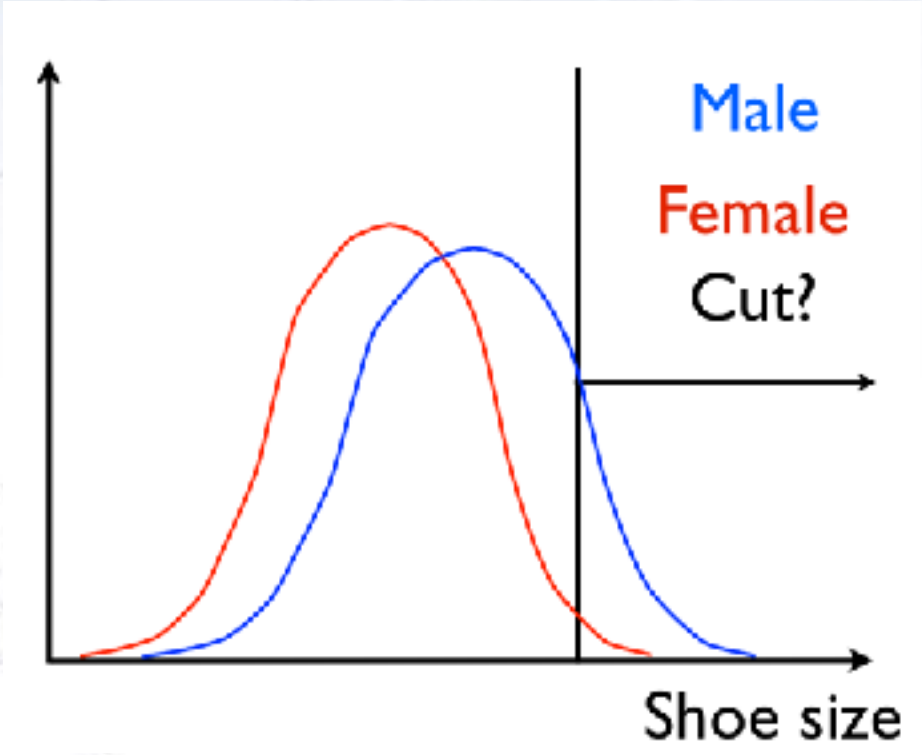
Problem: You want to figure out a method for getting sample that is 95% male!
Solution: Gather height data from 10000 people, Estimate cut with 95% purity!



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!

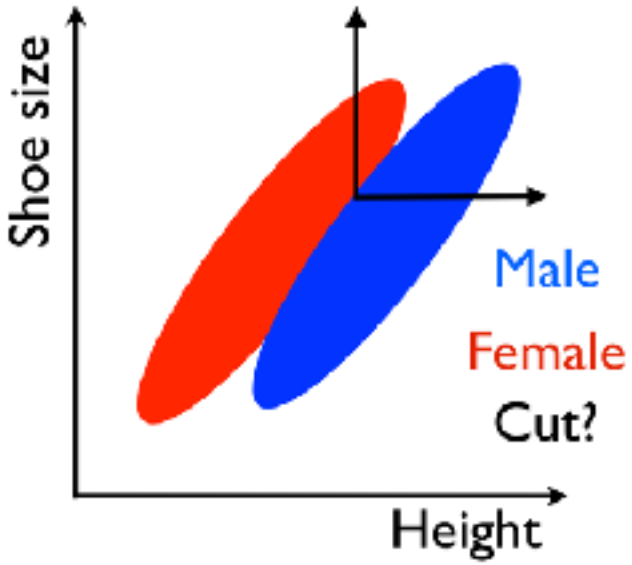
Simple Example

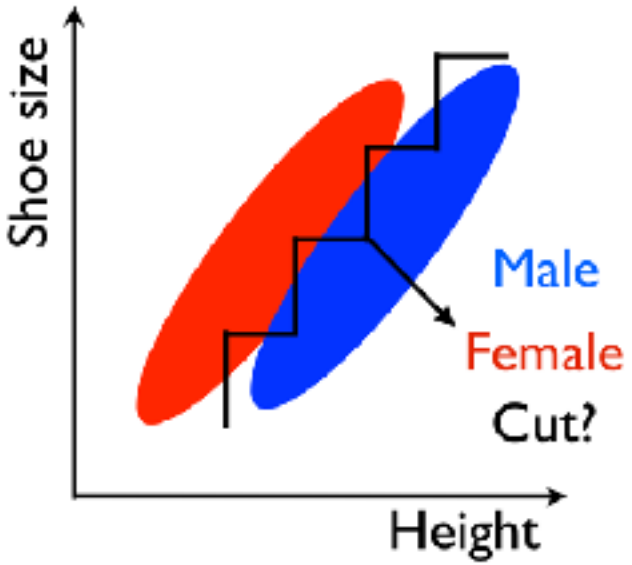
So we look if the data is correlated, and consider the options:

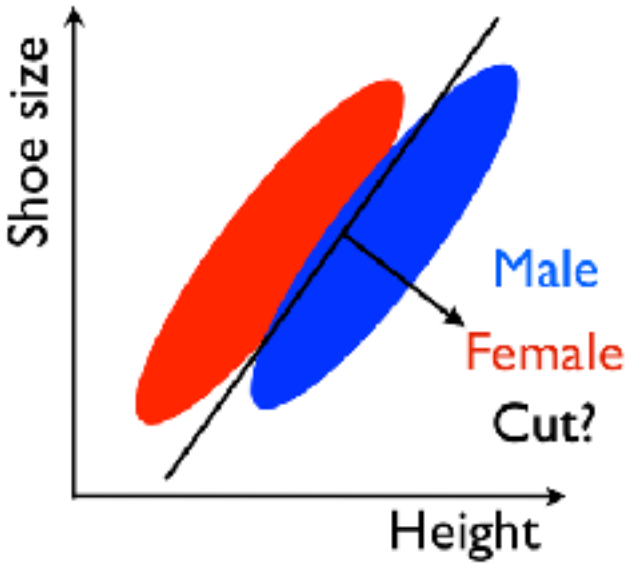
Cut on each var?
Poor efficiency!

Advanced cut?
Clumsy and hard to implement

Combine var?
Smart and promising







The latter is the Linear Discriminant Analysis (LDA) aka. Fisher discriminant!
It has the advantage of being simple and applicable in many dimensions easily!

Simple Example

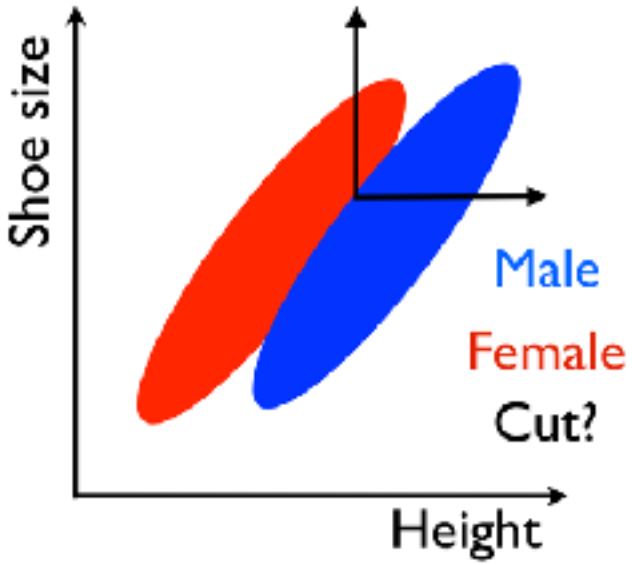
So we look if the data is correlated, and consider the options:

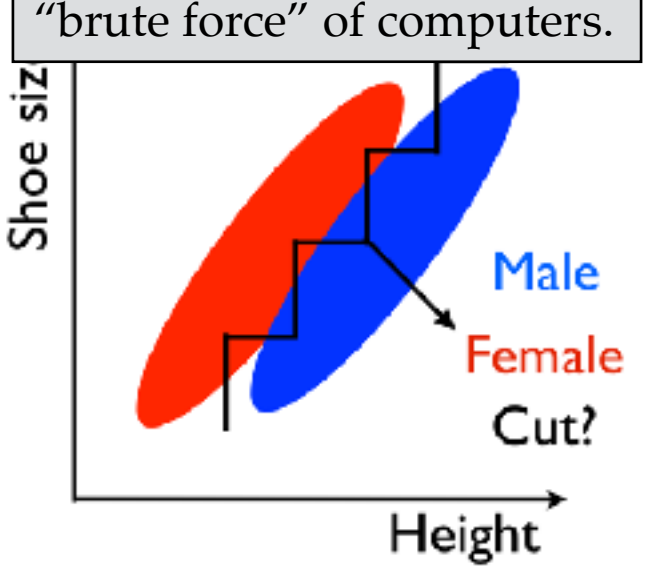
Cut on each var?
Poor efficiency!

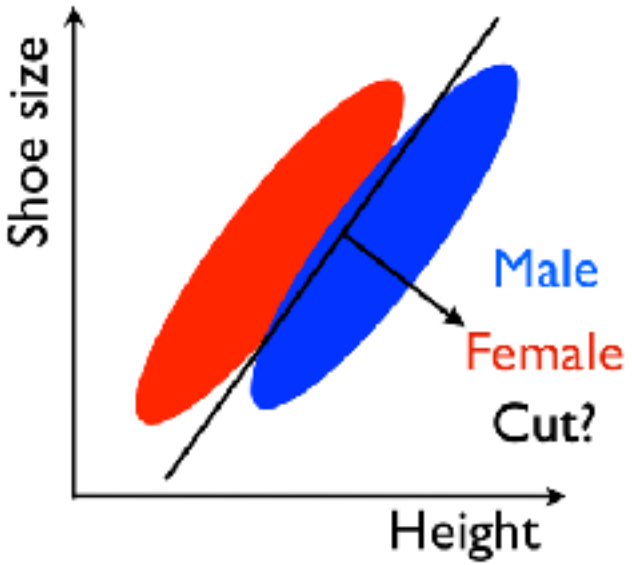
Advanced cut?

Interestingly, this is exactly the approach of tree-based learning in ML, through "brute force" of computers.

Combine var?
Smart and promising







The latter is the Linear Discriminant Analysis (LDA) aka. Fisher discriminant!
It has the advantage of being simple and applicable in many dimensions easily!

Separating Classes/Types

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

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

LDA / Fisher Discriminant

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

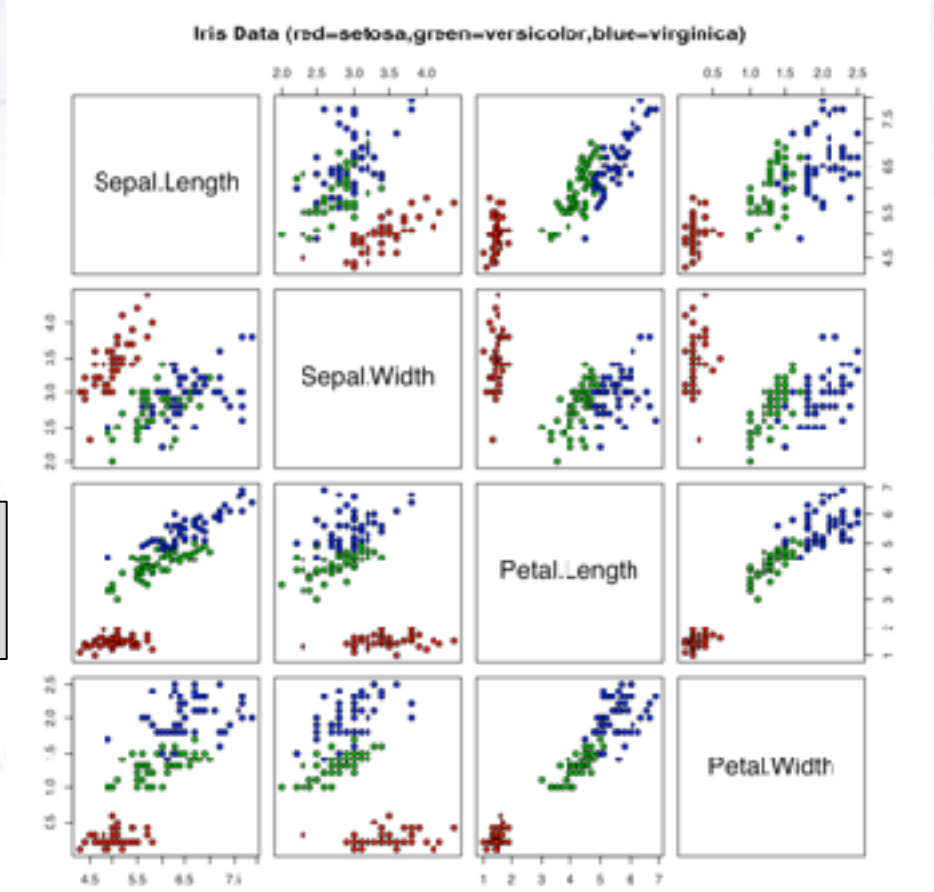
Q: How to combine the variables?
A: Use the Fisher Discriminant:

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

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.



LDA / 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!

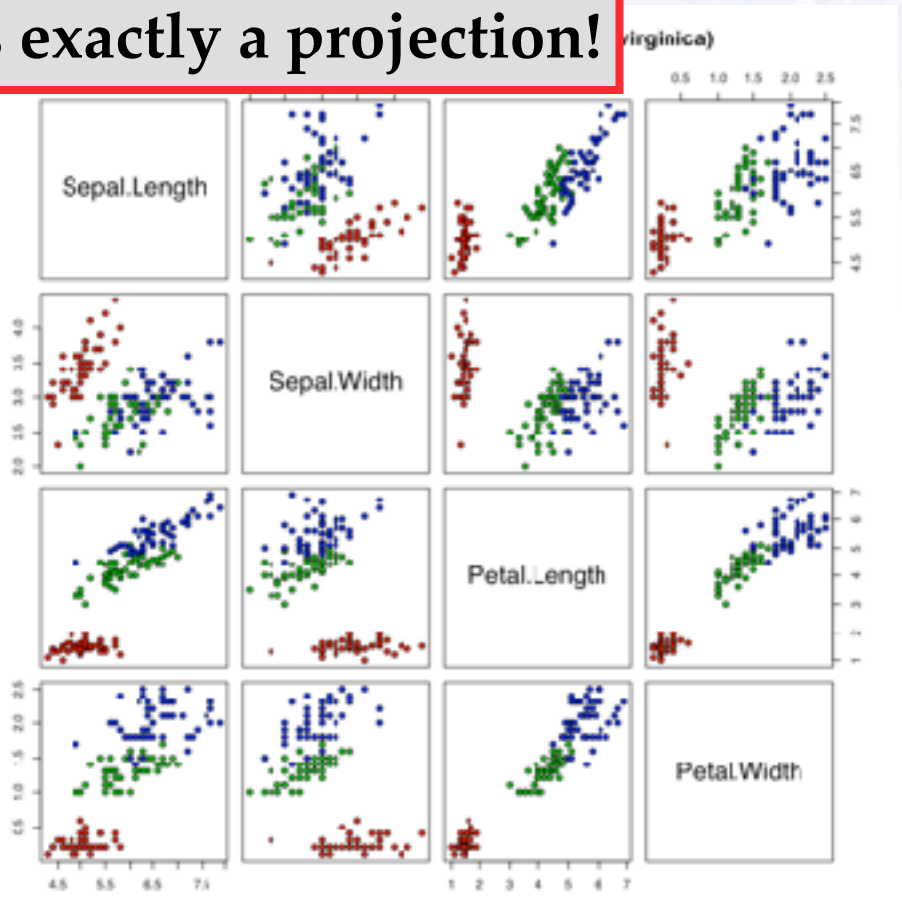
$$\mathcal{F} = w_0 + \vec{w} \cdot \vec{x}$$

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.



LDA / Fisher Discriminant

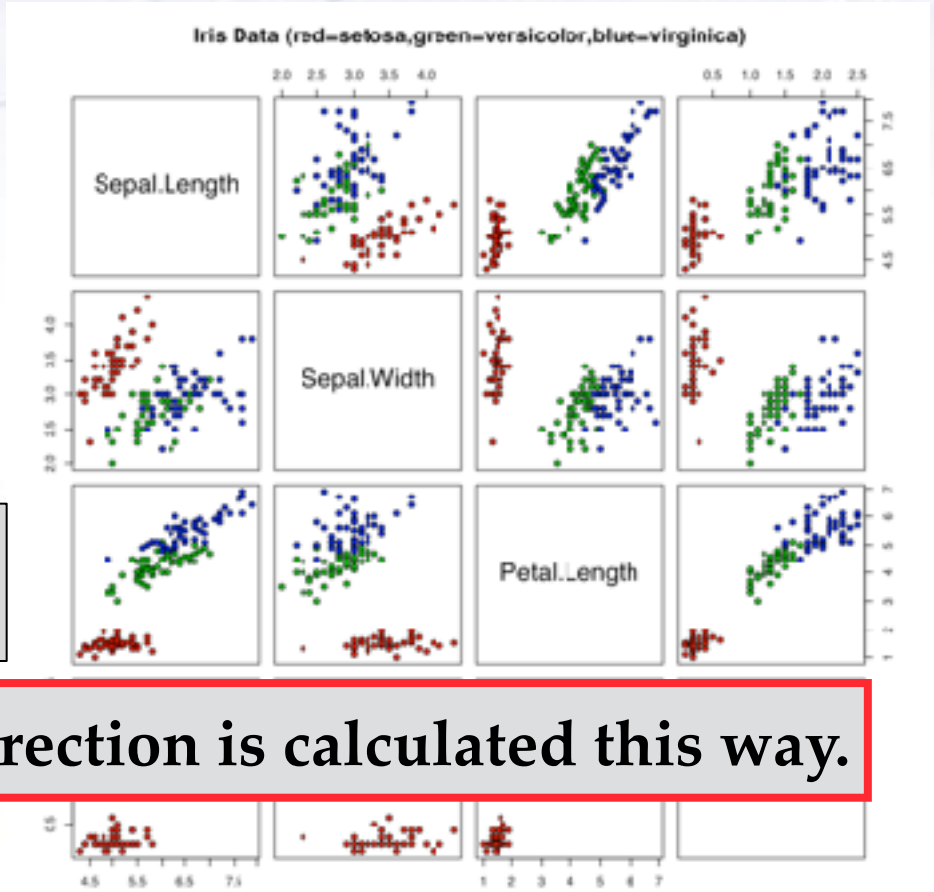
You want to separate two types/classes (A and B) of events using several measurements.

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

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

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

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



The optimal projection direction is calculated this way.

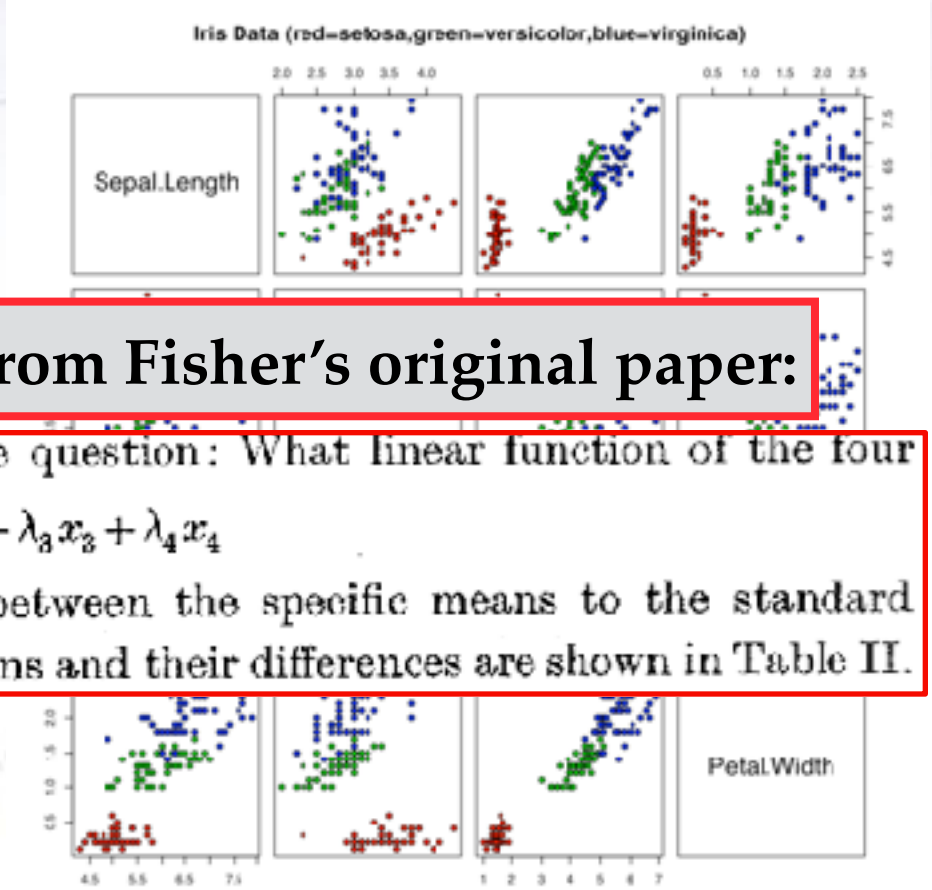
This incorporates the linear correlations into the separation capability.

LDA / 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:

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



Q: How to choose the values of w?

From Fisher's original paper:

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.

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

LDA / Fisher Discriminant

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

Q: How to combine

A: Use the Fisher D

This is just an offset (one can also scale)!

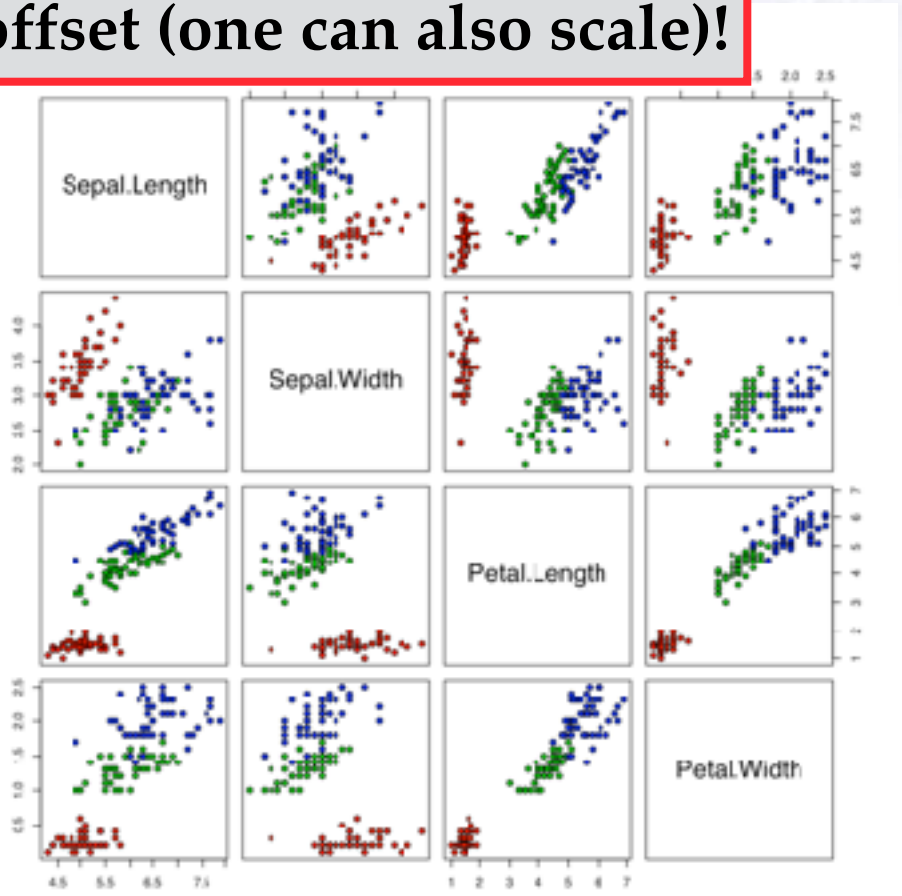
$$\mathcal{F} = w_0 + \vec{w} \cdot \vec{x}$$

Q: How to choose the values of w?

A: Inverting the covariance matrices:

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

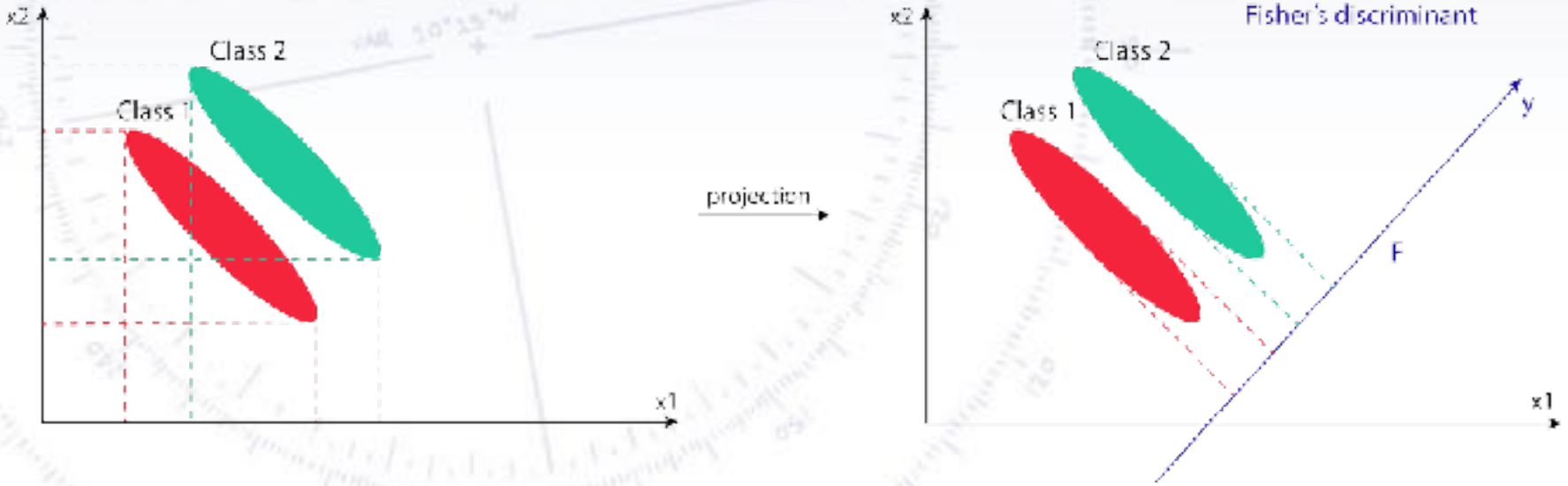
This can be calculated analytically, and incorporates the linear correlations into the separation capability.



LDA / Fisher Discriminant

<u>Executive summary:</u>

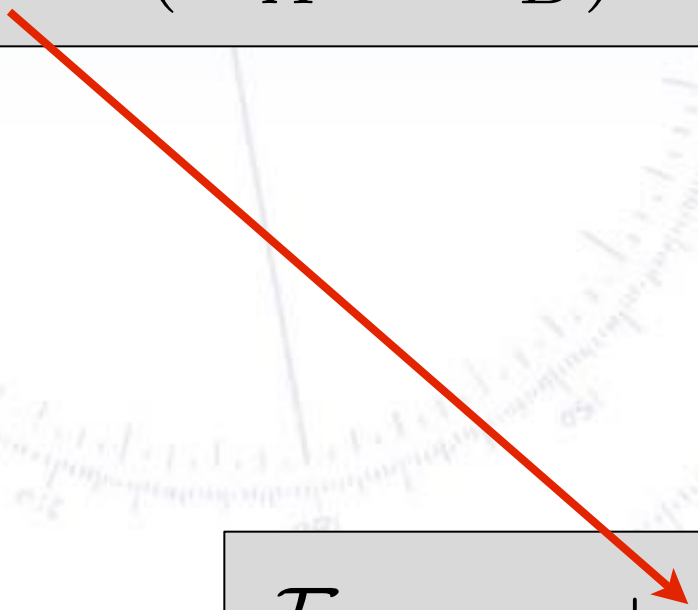
Fisher's Discriminant uses a linear combination of variables to give a single variable with the maximum possible separation (for linear combinations!).



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

LDA / Fisher Discriminant

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



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

LDA / 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.

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

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

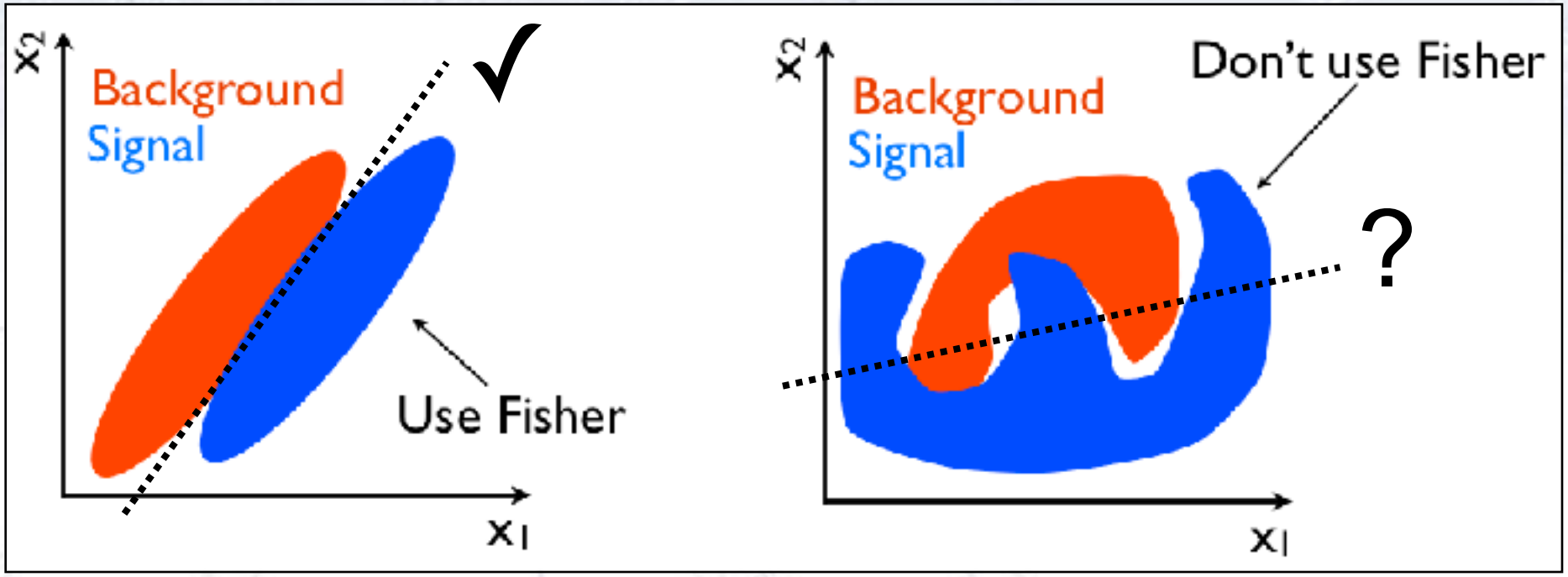
Given weights (w), you take your input variables (x) and combine them linearly as follows:

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

F is what you base your decision on.

Selecting signal in 2D

Now let us try in 2 dimensions (two 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**.

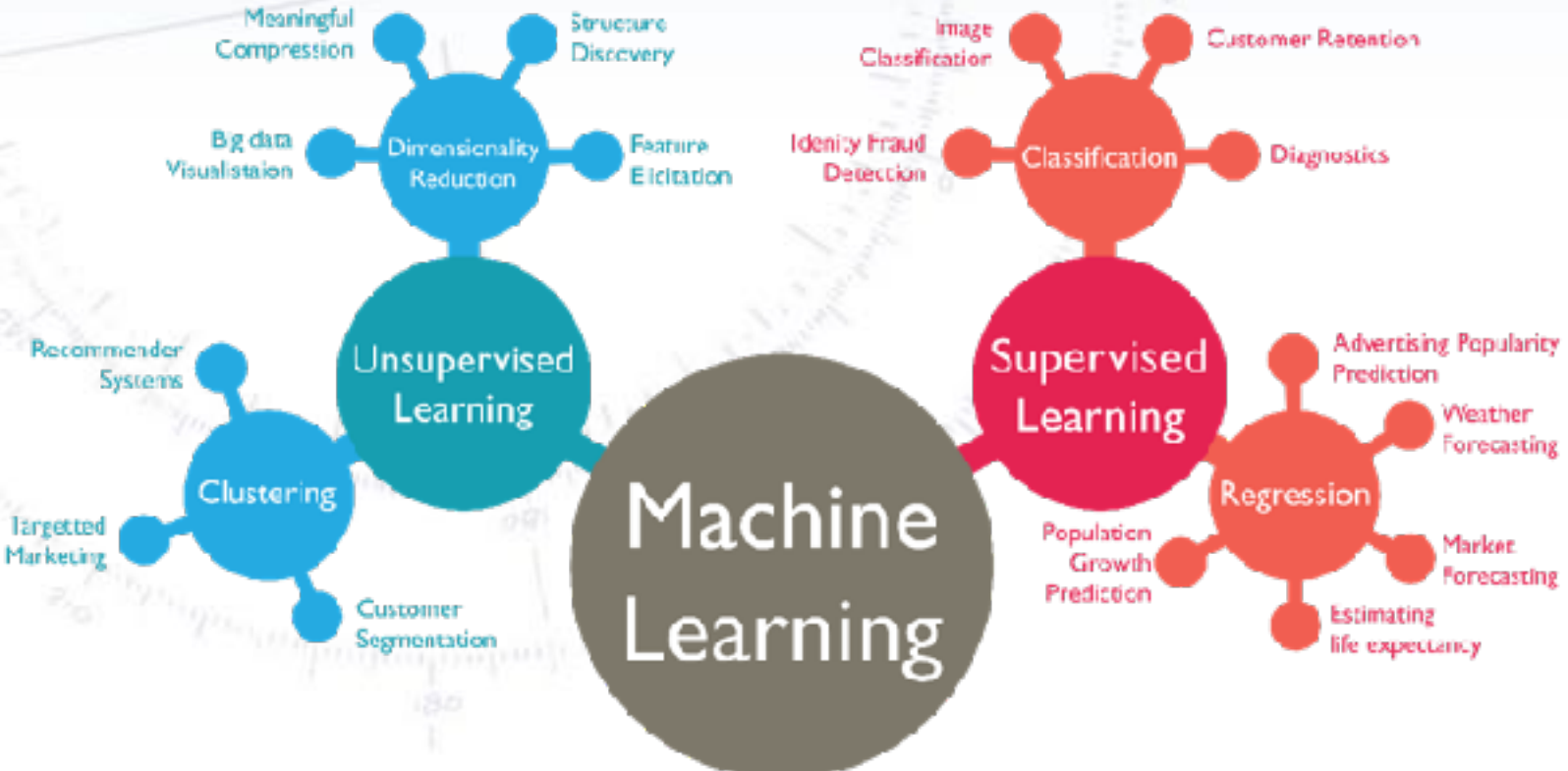
But this is **very rarely** the case, and therefore more "tough" methods are needed...

Relation to ML

Machine Learning

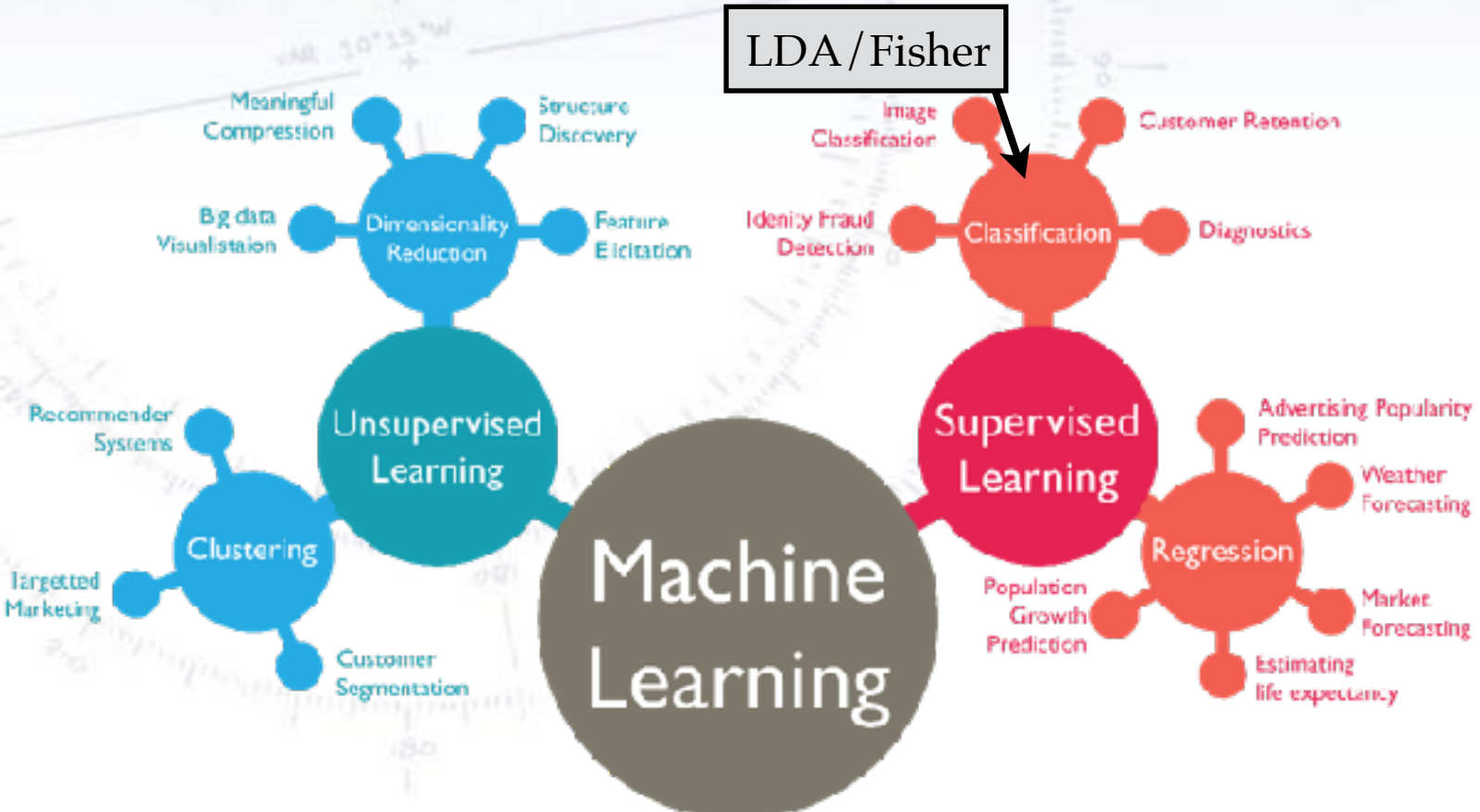
Unsupervised vs. Supervised
Classification vs. Regression

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).



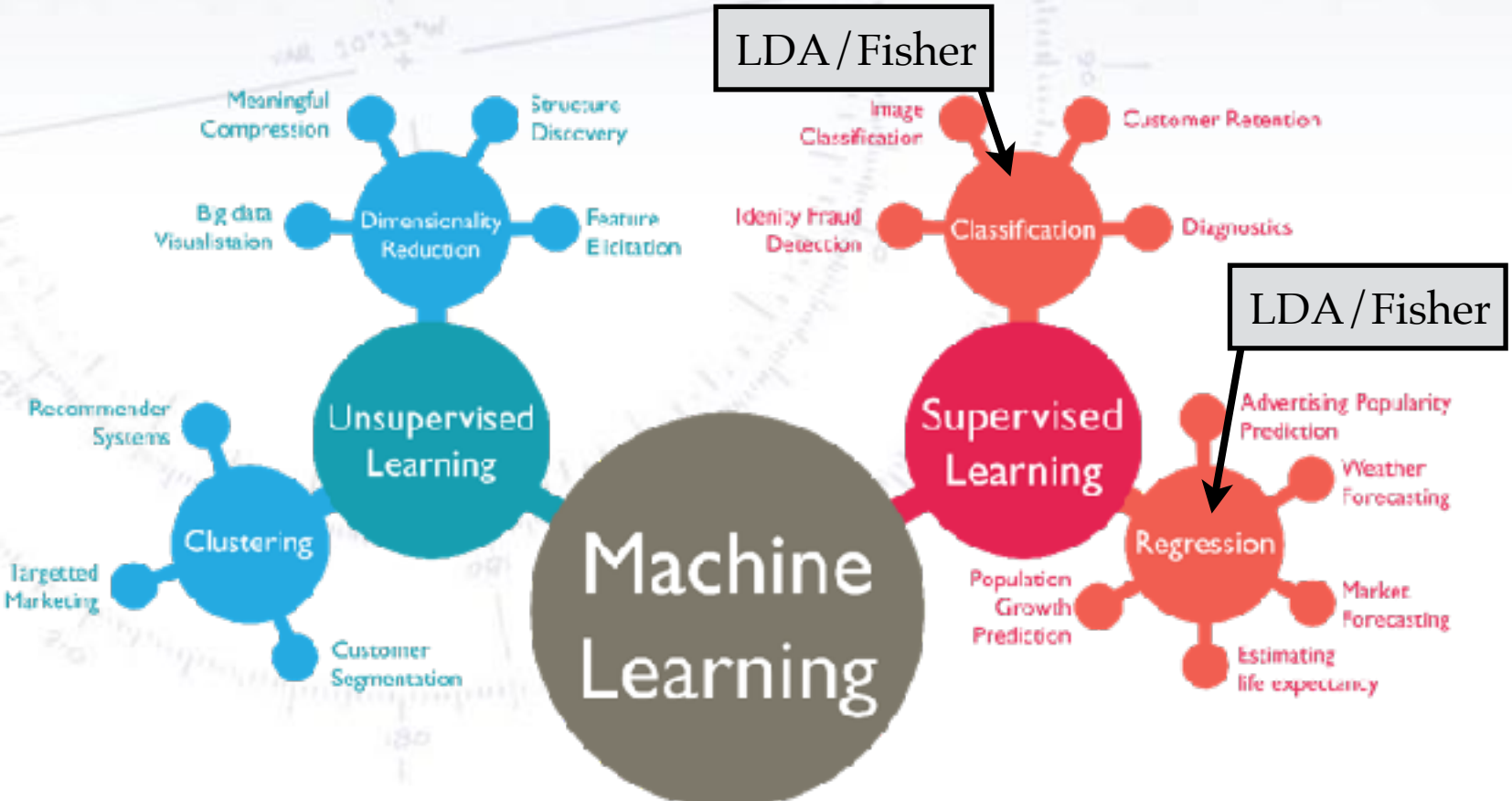
Unsupervised vs. Supervised
Classification vs. Regression

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).



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).



Relation to PCA

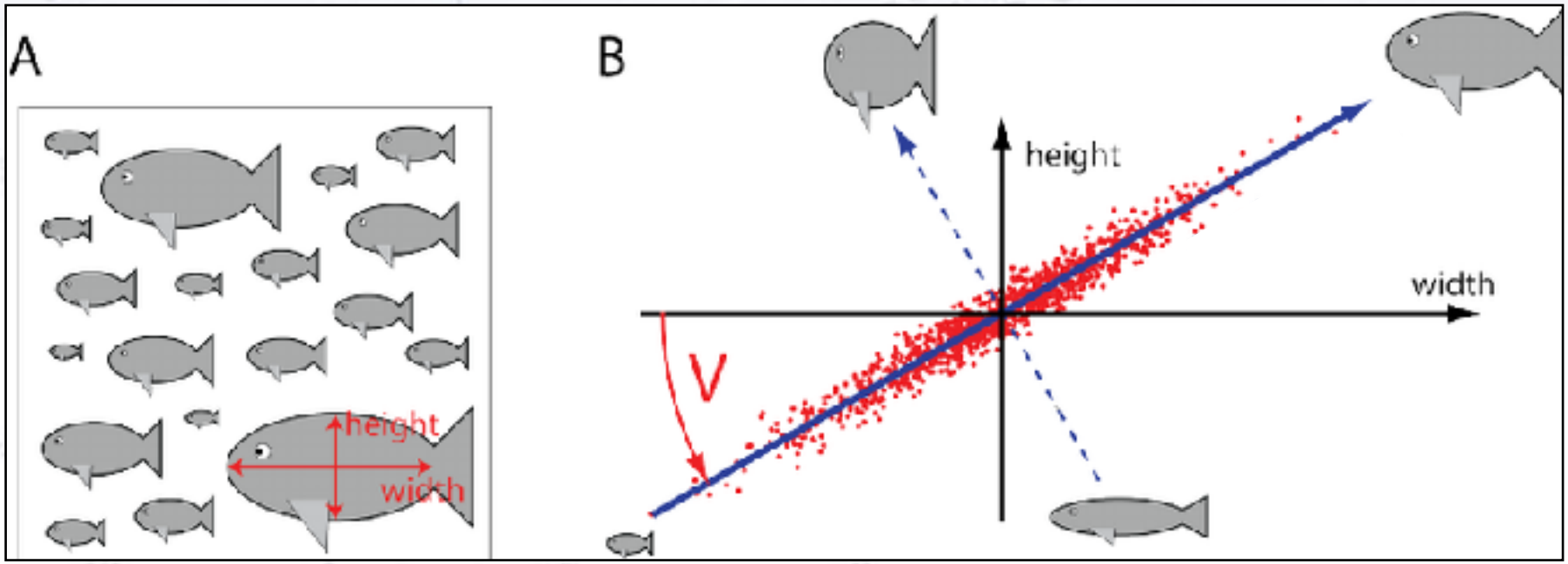
Principle Component Analysis

Principle Component Analysis (PCA)

PCA is a technique for **dimensionality reduction** of a dataset, accomplished through **linearly transforming** the data into a new coordinate system where the **maximum variation** in the data is spanned by fewer (usually two!) dimensions than the initial data.

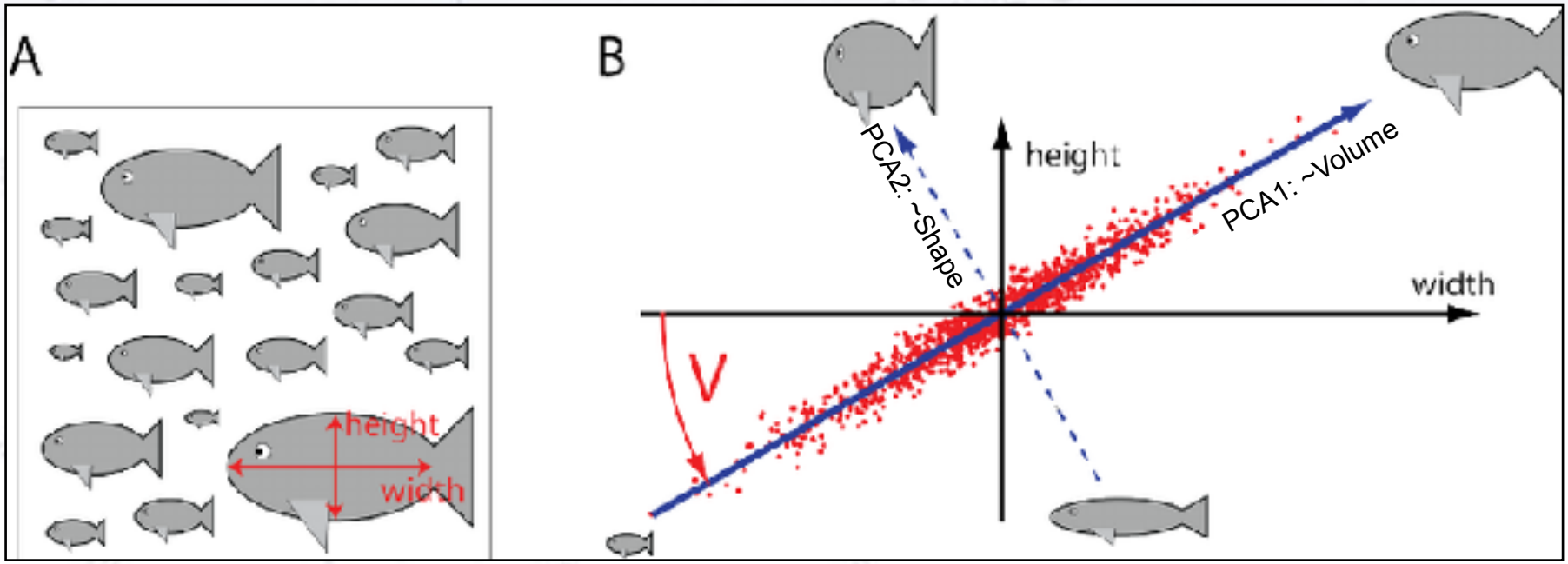
Principle Component Analysis (PCA)

PCA is a technique for **dimensionality reduction** of a dataset, accomplished through **linearly transforming** the data into a new coordinate system where the **maximum variation** in the data is spanned by fewer (usually two!) dimensions than the initial data.



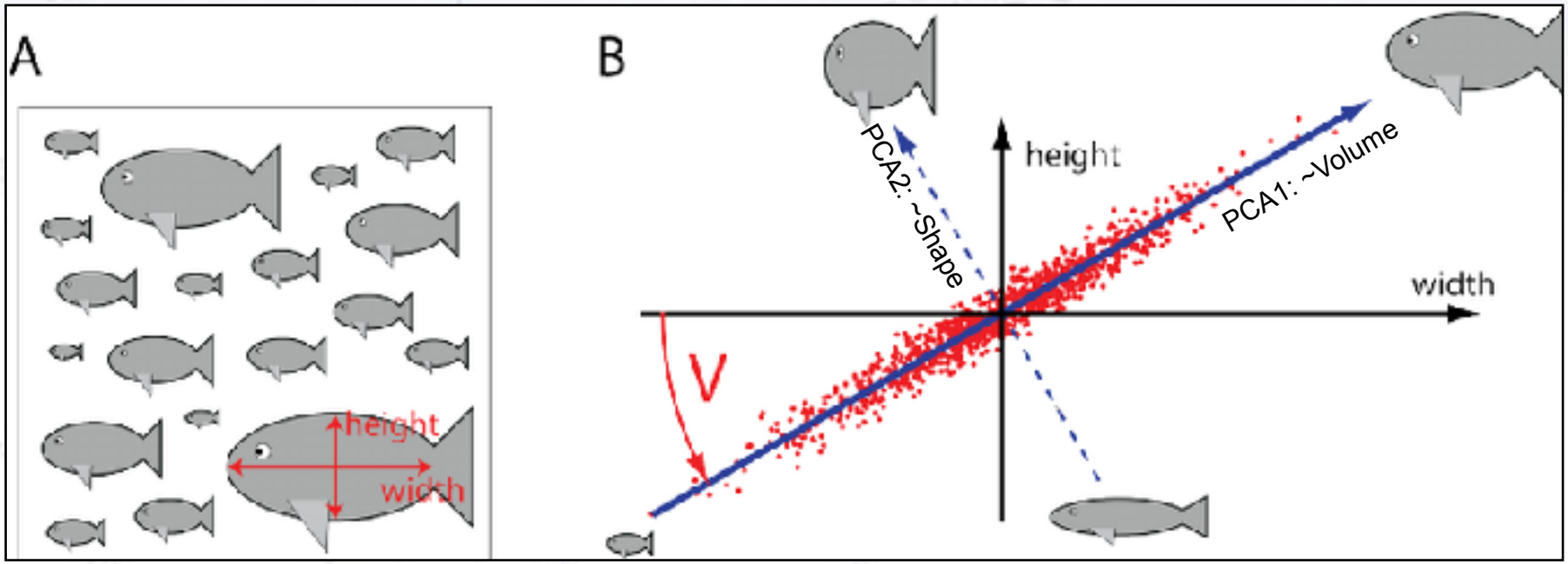
Principle Component Analysis (PCA)

PCA is a technique for **dimensionality reduction** of a dataset, accomplished through **linearly transforming** the data into a new coordinate system where the **maximum variation** in the data is spanned by fewer (usually two!) dimensions than the initial data.



Principle Component Analysis (PCA)

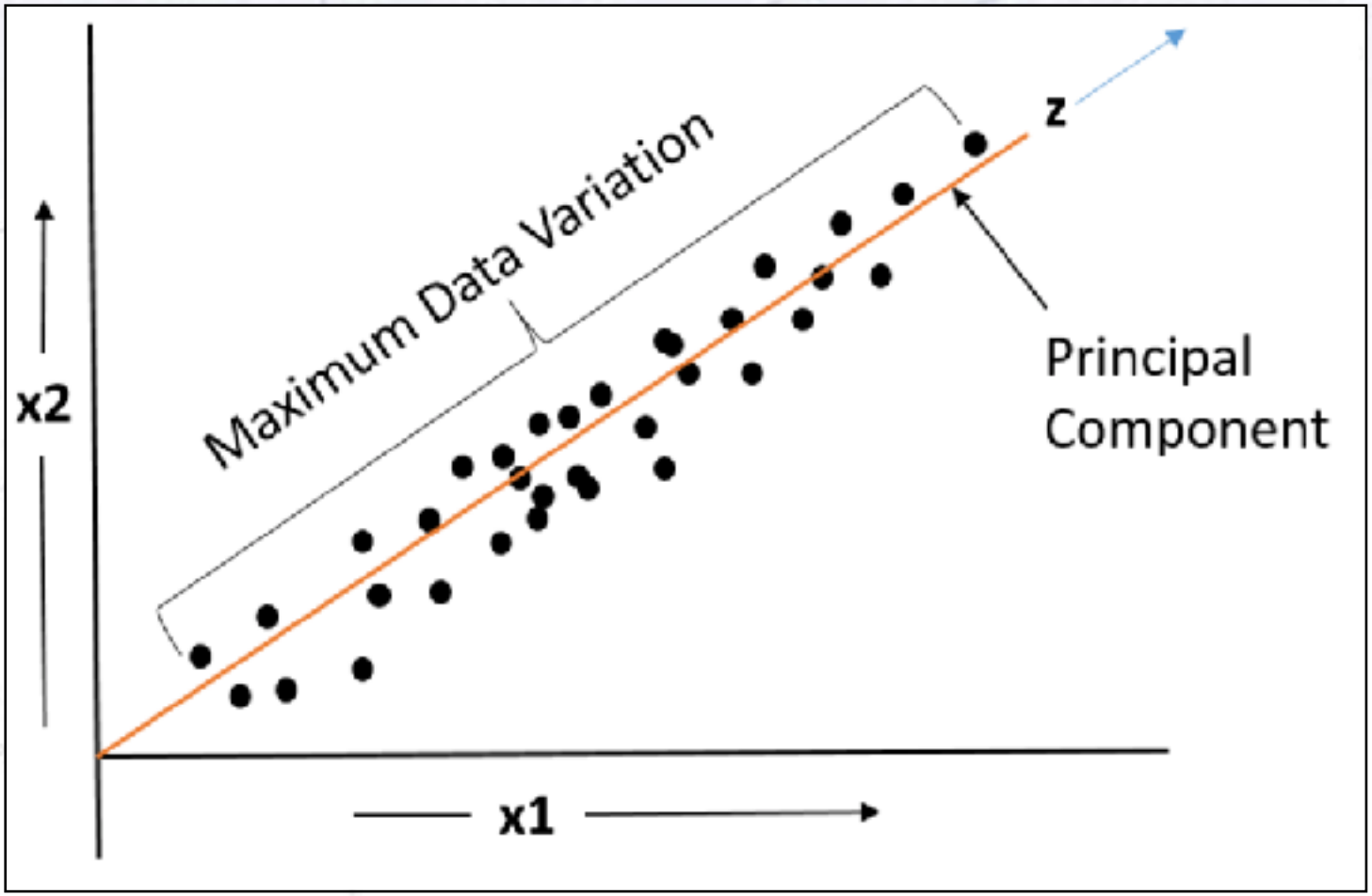
PCA is a technique for **dimensionality reduction** of a dataset, accomplished through **linearly transforming** the data into a new coordinate system where the **maximum variation** in the data is spanned by fewer (usually two!) dimensions than the initial data.



The PCA directions constitute an **orthonormal basis**. PCA is the process of computing the principal components and using them to perform a change of basis on the data, often **using only the first few principal components**.

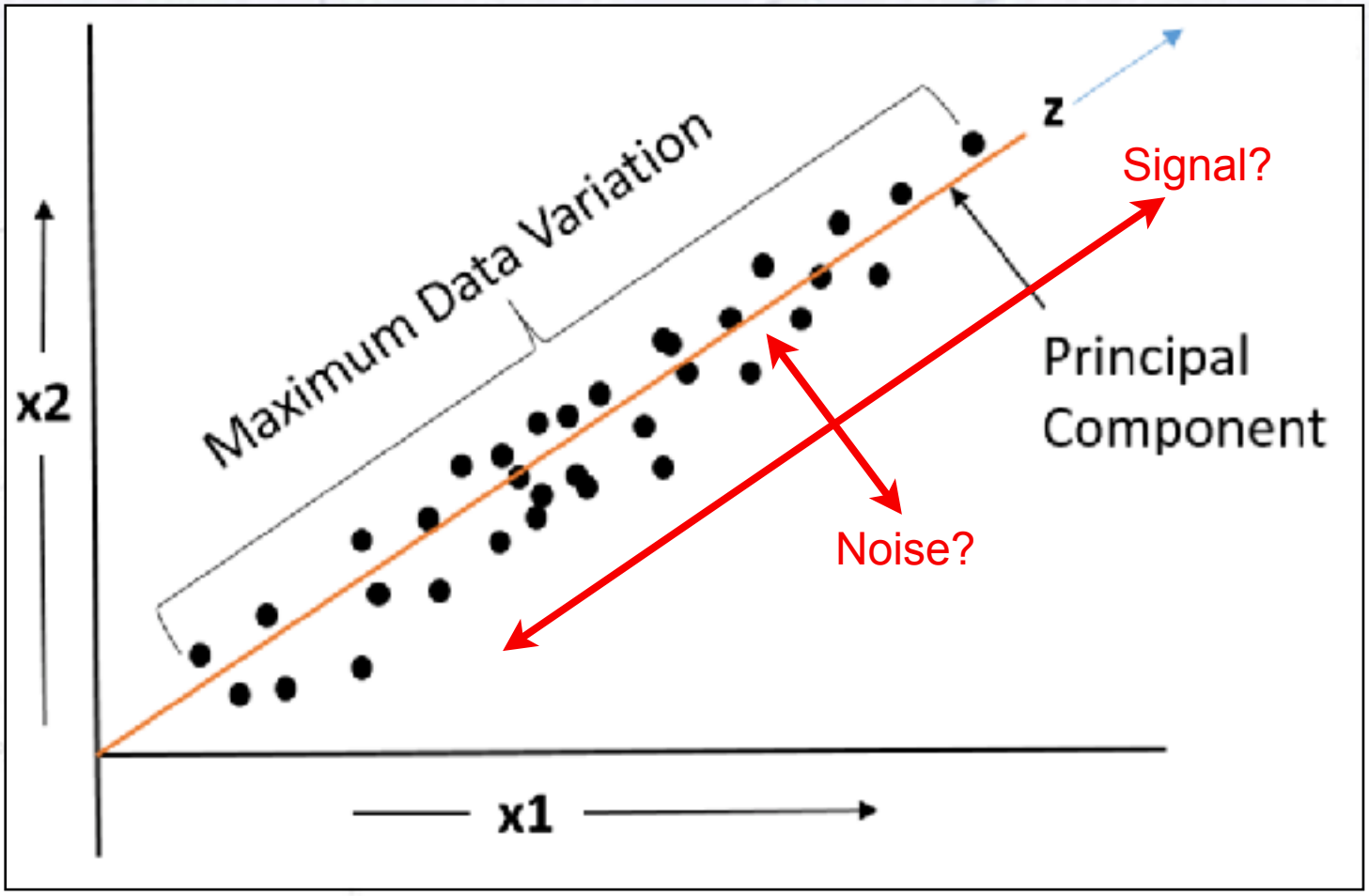
Principle Component Analysis (PCA)

Why consider the largest variation? Well, the "hope" is that any possible signal lies in the (principle) direction of maximal variance, while noise in the data lies along the "lesser" directions:



Principle Component Analysis (PCA)

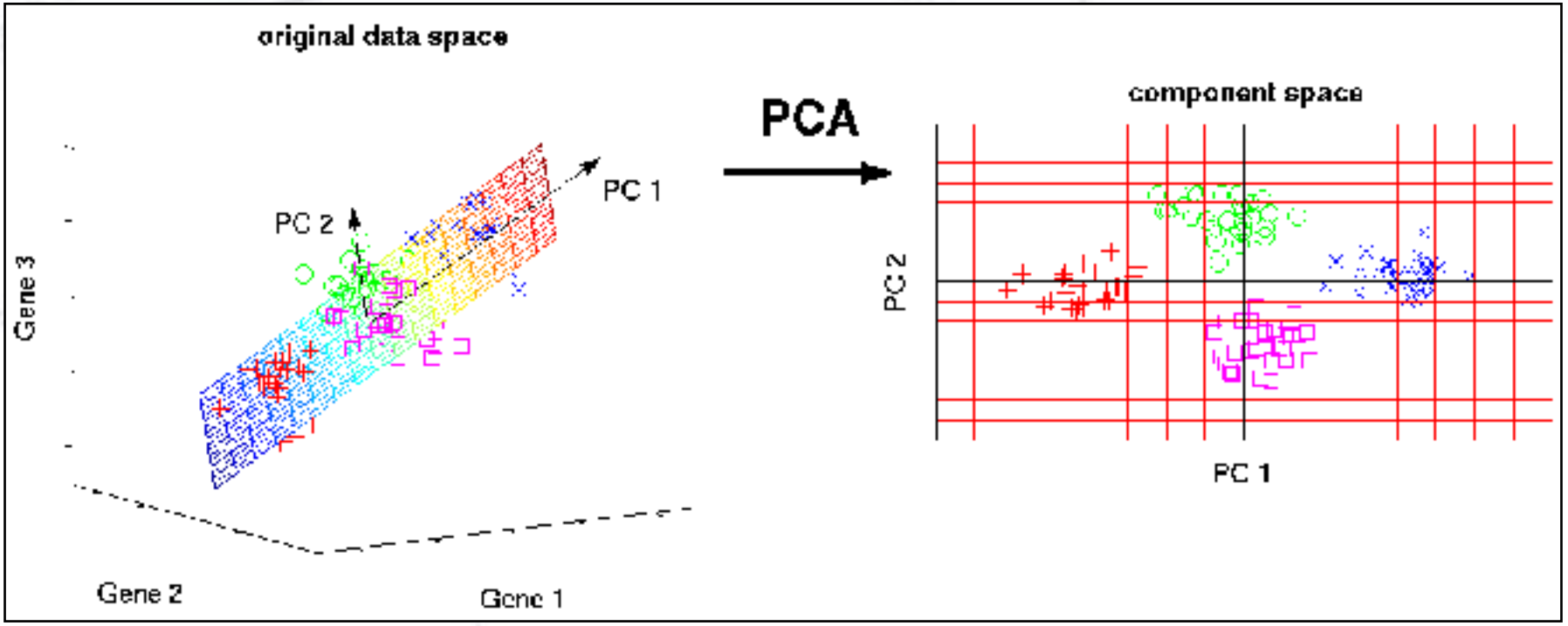
Why consider the largest variation? Well, the "hope" is that any possible signal lies in the (principle) direction of maximal variance, while noise in the data lies along the "lesser" directions:



Principle Component Analysis (PCA)

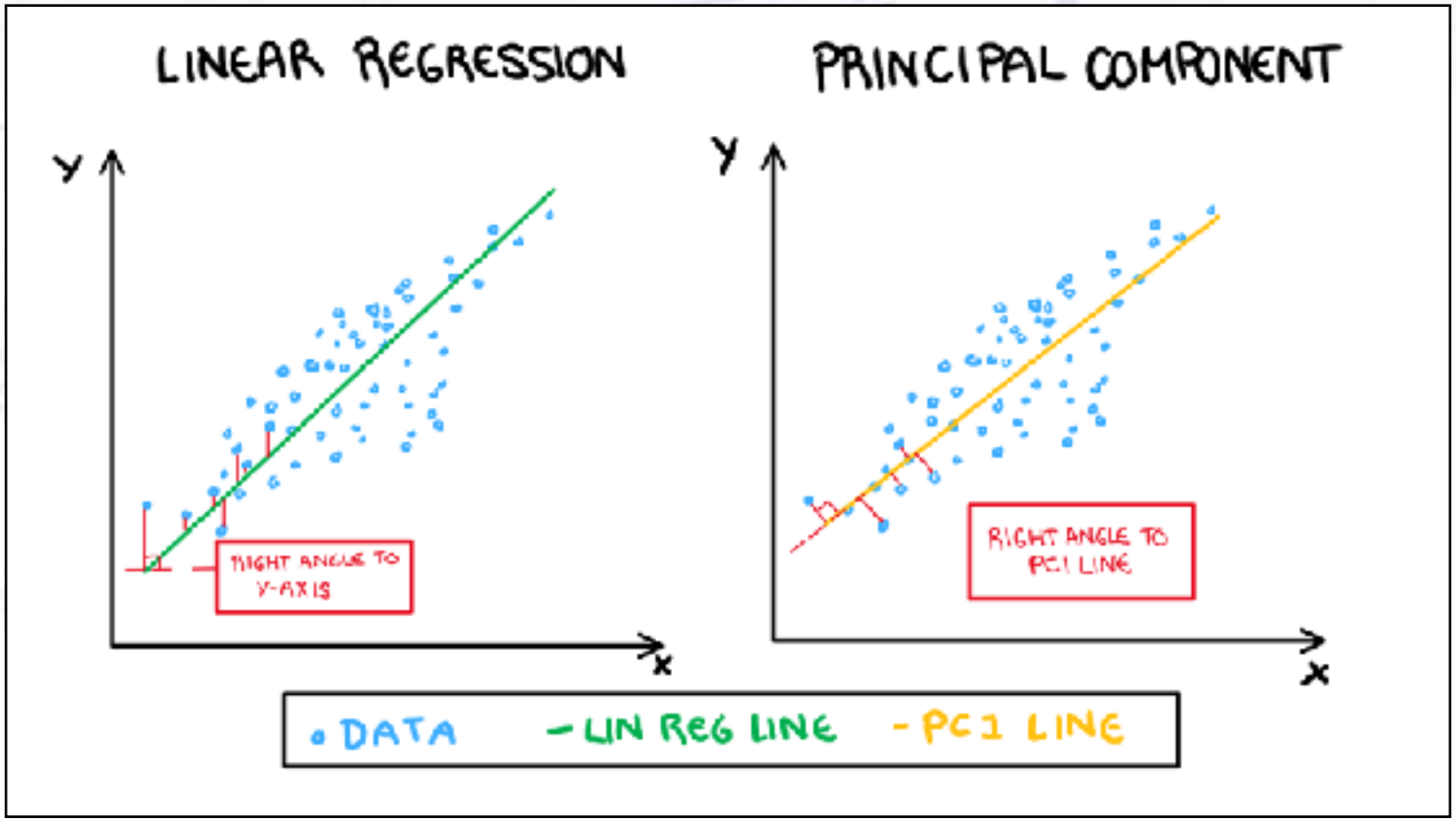
PCA is defined as an orthogonal linear transformation to a new coordinate system, where the i'th greatest variance by some scalar projection of the data comes to lie on the i'th coordinate (called the i'th principal component).

Shown below is an example of a PCA transformation from 3D to 2D, considering the two PCs with the greatest variance:



Principle Component Analysis (PCA)

The principal components of data (points in N-dimensional space) are a sequence of unit vectors, where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first i-1 vectors. Here, a best-fitting line is defined as one that minimises the average squared perpendicular distance from the points to the line.



Principle Component Analysis (PCA)

PCA is defined as an orthogonal linear transformation to a new coordinate system, where the i'th greatest variance by some scalar projection of the data comes to lie on the i'th coordinate (called the i'th principal component).

$$\vec{w}_1 = \arg\max_{||w||=1} \left[\sum_i (\vec{x}_i \cdot \vec{w})^2 \right]$$

Principle Component Analysis (PCA)

PCA is defined as an orthogonal linear transformation to a new coordinate system, where the i'th greatest variance by some scalar projection of the data comes to lie on the i'th coordinate (called the i'th principal component).

The PCA transformation is defined as a set of size q of p-dimensional vectors:

$$\vec{w}_{(k)} = (w_1, \ldots, w_p)_{(k)}$$

that map each data point to a new vector of principle component "scores", given by:

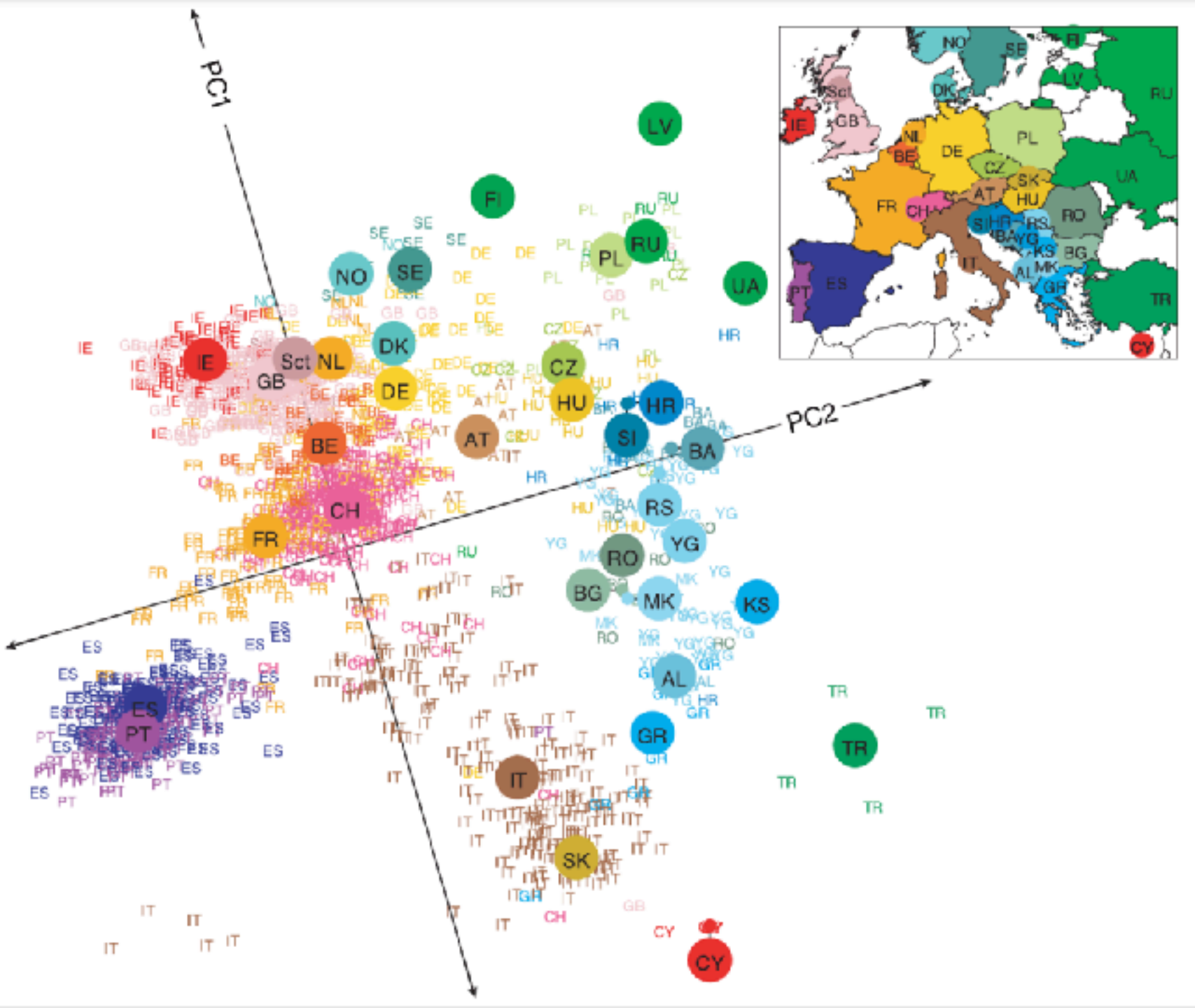
$$\vec{t}_{(i)} = (t_1, \ldots, w_q)_{(i)} = \vec{x}_{(i)} \cdot \vec{w}_{(k)}$$

in such a way that the individual variables of t considered over the data set successively inherit the maximum possible variance from the original data.

The PC are calculated as:

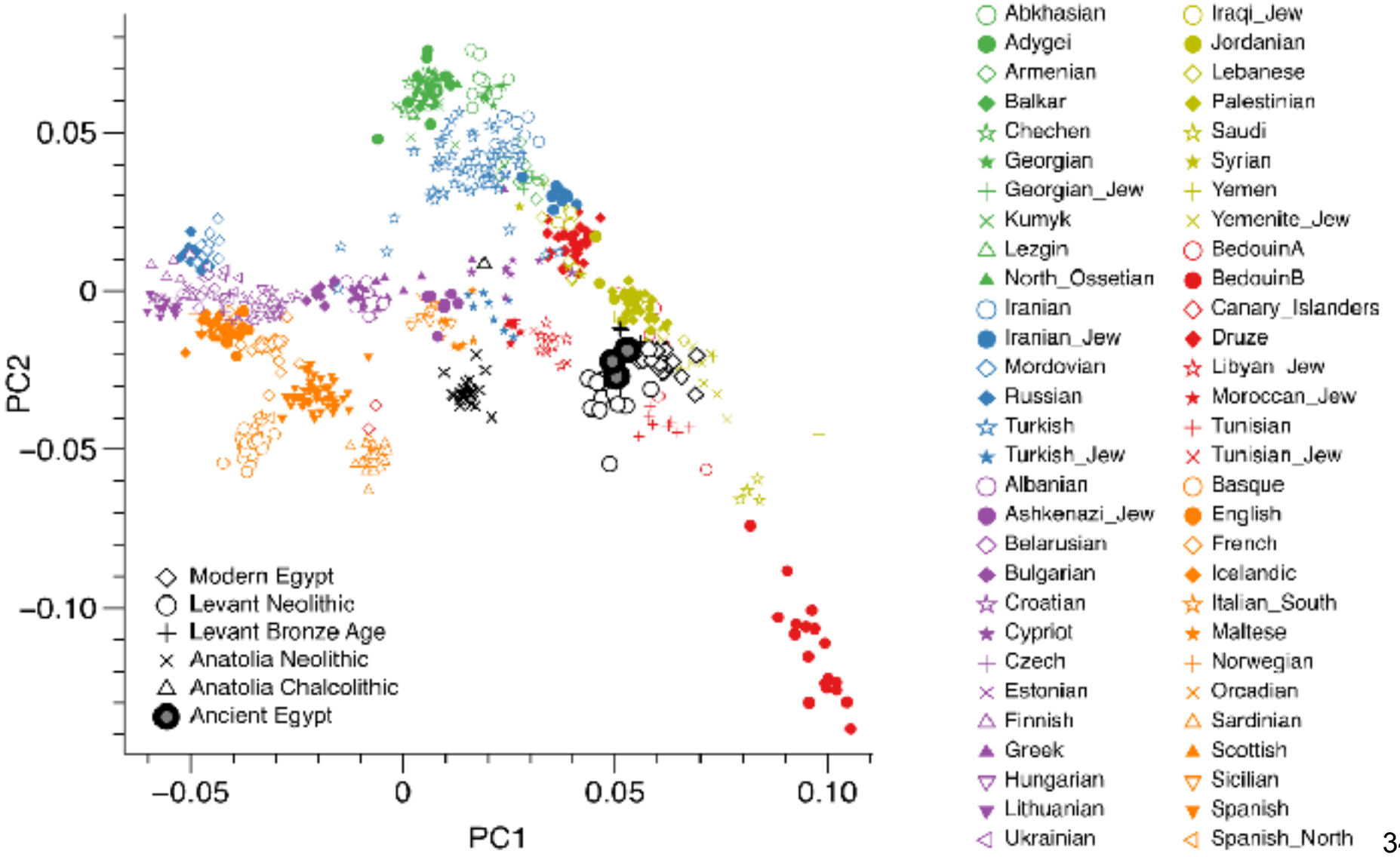
$$\vec{w}_1 = \arg\max_{||w||=1} \left[\sum_i (\vec{x}_i \cdot \vec{w})^2 \right]$$

PCA of European Genes



Principle Component Analysis (PCA)

PCA can be used to determine "proximity" in very high dimensional spaces:



PCA vs. LDA (Fisher)

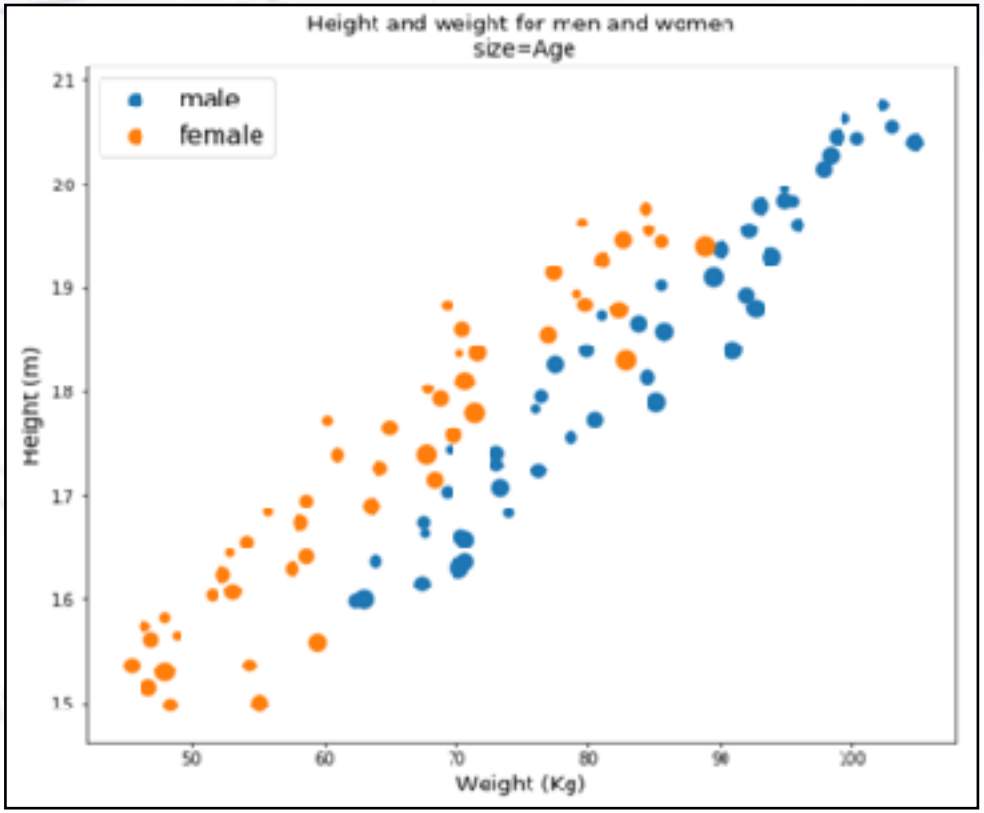
So what is the difference, and when to use one or the other method?

PCA is an unsupervised algorithm while **LDA is a supervised** algorithm. This means that **PCA finds directions of maximum variance** regardless of class labels while **LDA finds directions of maximum class separability**.

PCA vs. LDA (Fisher)

So what is the difference, and when to use one or the other method?

PCA is an unsupervised algorithm while **LDA is a supervised** algorithm. This means that **PCA finds directions of maximum variance** regardless of class labels while **LDA finds directions of maximum class separability**.



PCA vs. LDA (Fisher)

So what is the difference, and when to use one or the other method?

PCA is an unsupervised algorithm while **LDA is a supervised** algorithm. This means that **PCA finds directions of maximum variance** regardless of class labels while **LDA finds directions of maximum class separability**.

Conclusion:
Use LDA, if you have labelled (training) data and want good classification.

Use PCA, if you don't have data with labels, and want to find structures in data.

The dimensionality reduction of the PCA allows you to **get a visual inspection in 2D!**