# Applied Statistics

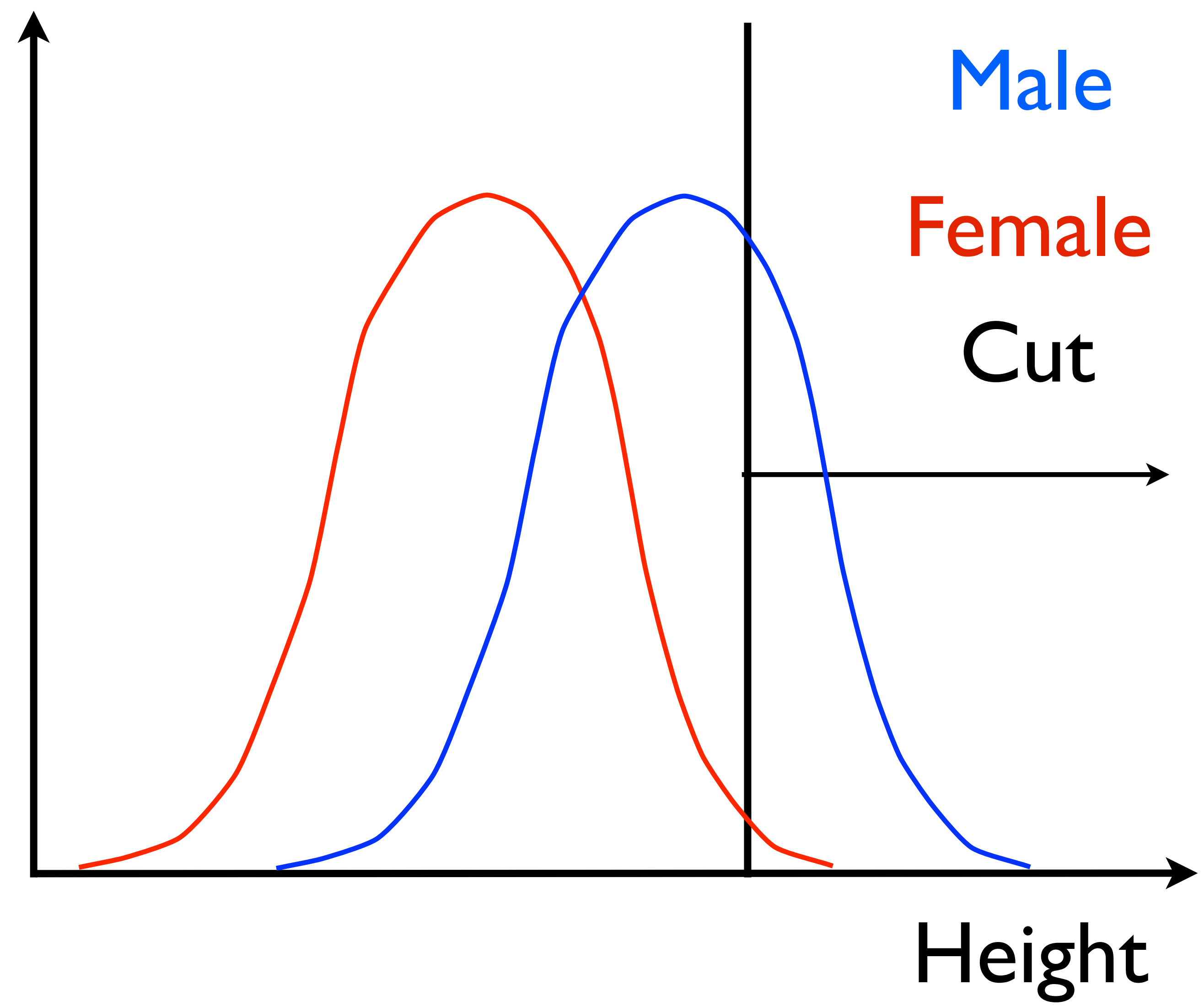
Week 6 - Multivariate Analysis

# This week

- Monday:
  - Intro to Multivariate analysis - Fisher discriminant/Iris data
- Tuesday:
  - Working on project two
- Friday:
  - A peek into more involved Multivariate techniques / machine learning

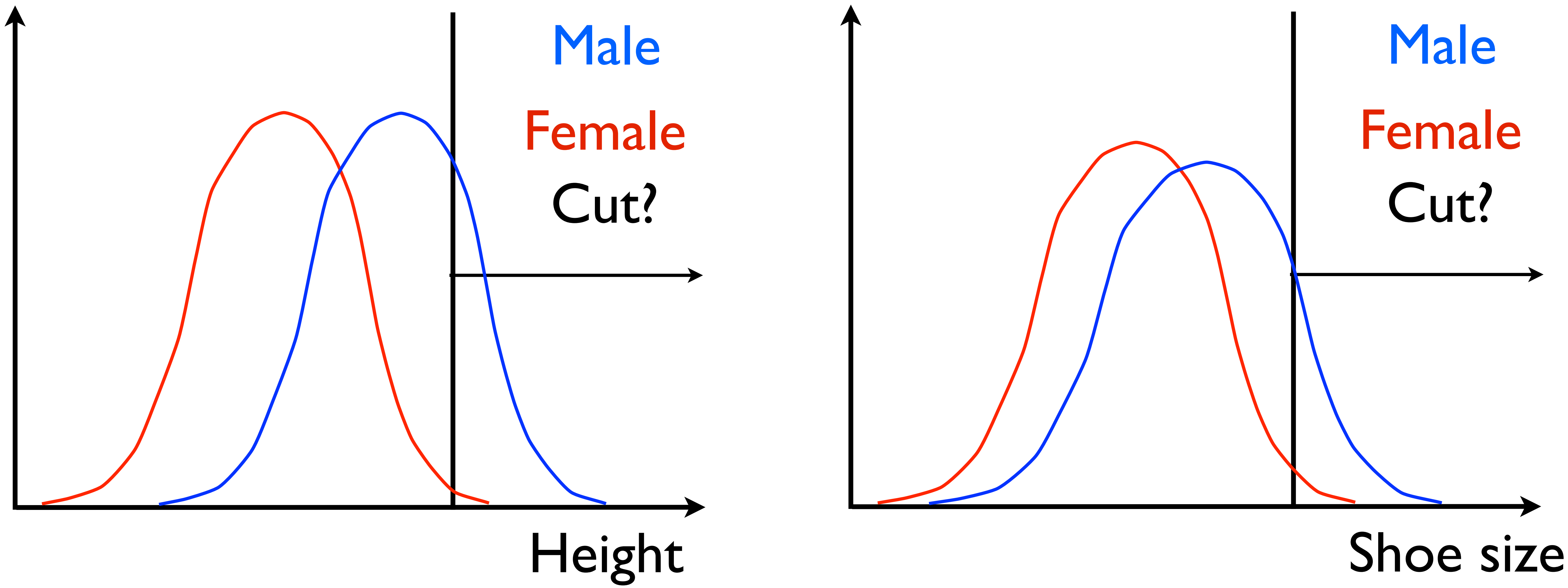
# Introduction - A simple example

- You want to figure out a method s.t. you are 95% sure a person is male:
- Easy: Gather height data from 10000 people, Estimate cut with 95% purity



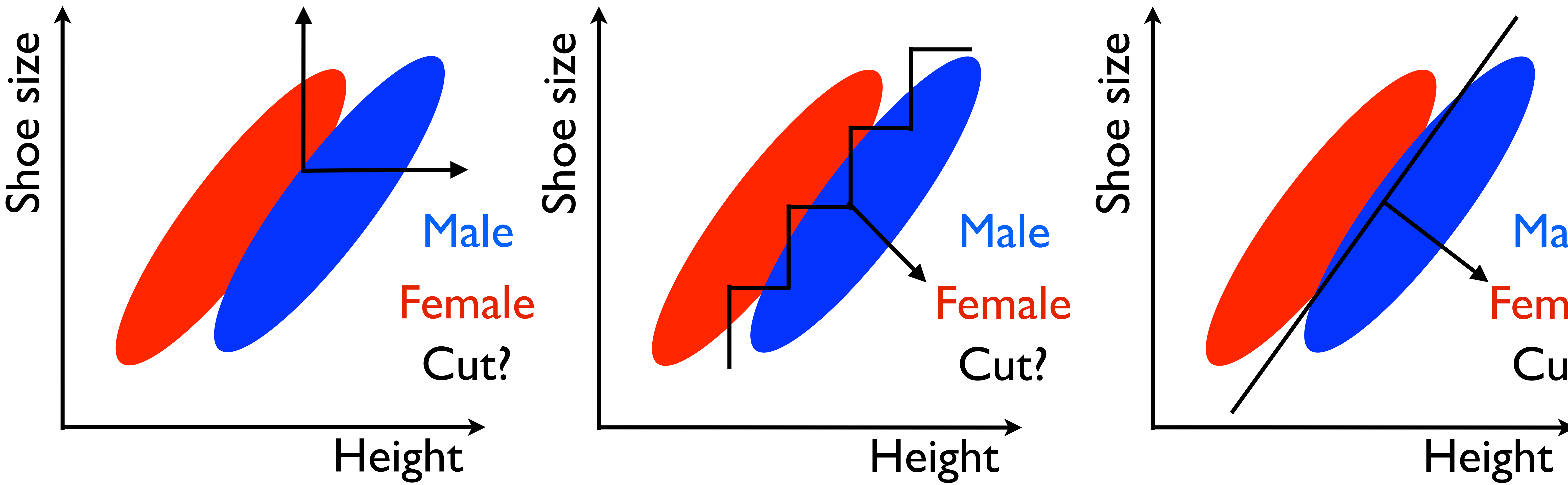
# Introduction - A simple example

- You want to figure out a method s.t. you are 95% sure a person is male:
- Your friend now gives you shoe size data as well:
  - More information: Better separation
  - Cut on both observables?



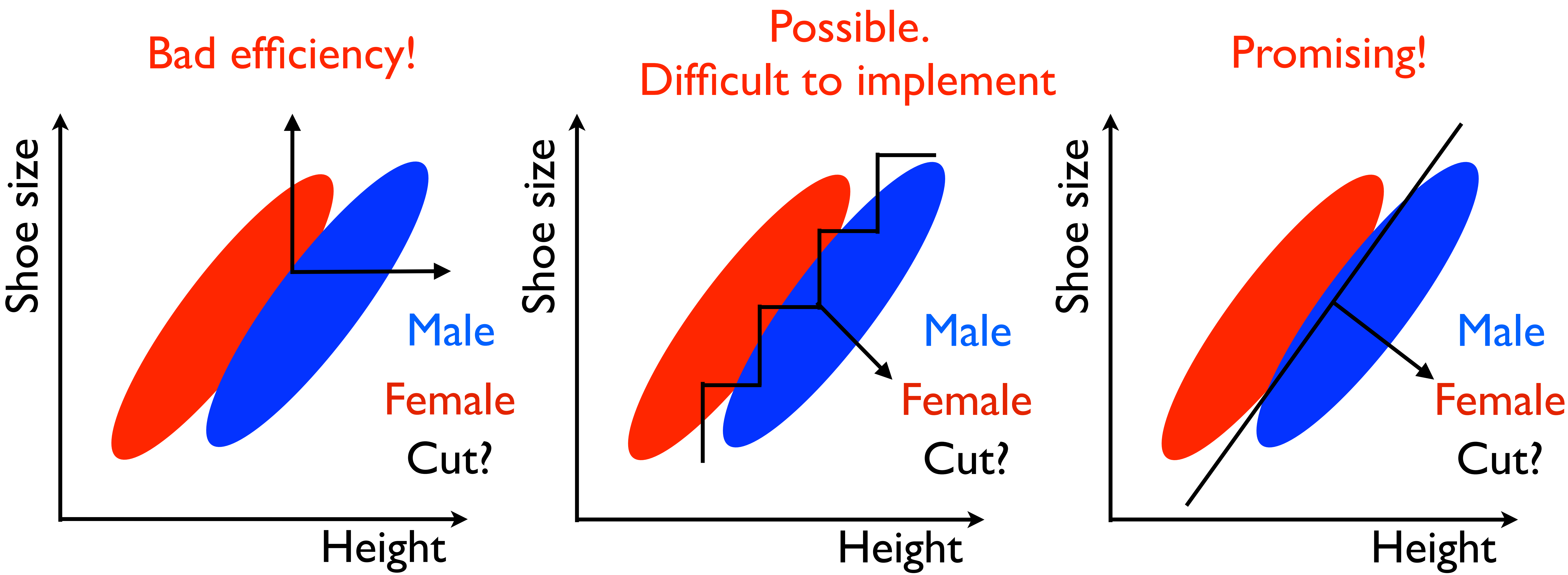
# Introduction - A simple example

- You want to figure out a method s.t. you are 95% sure a person is male:
- Your friend now gives you shoe size data as well:
  - Look at observable correlations



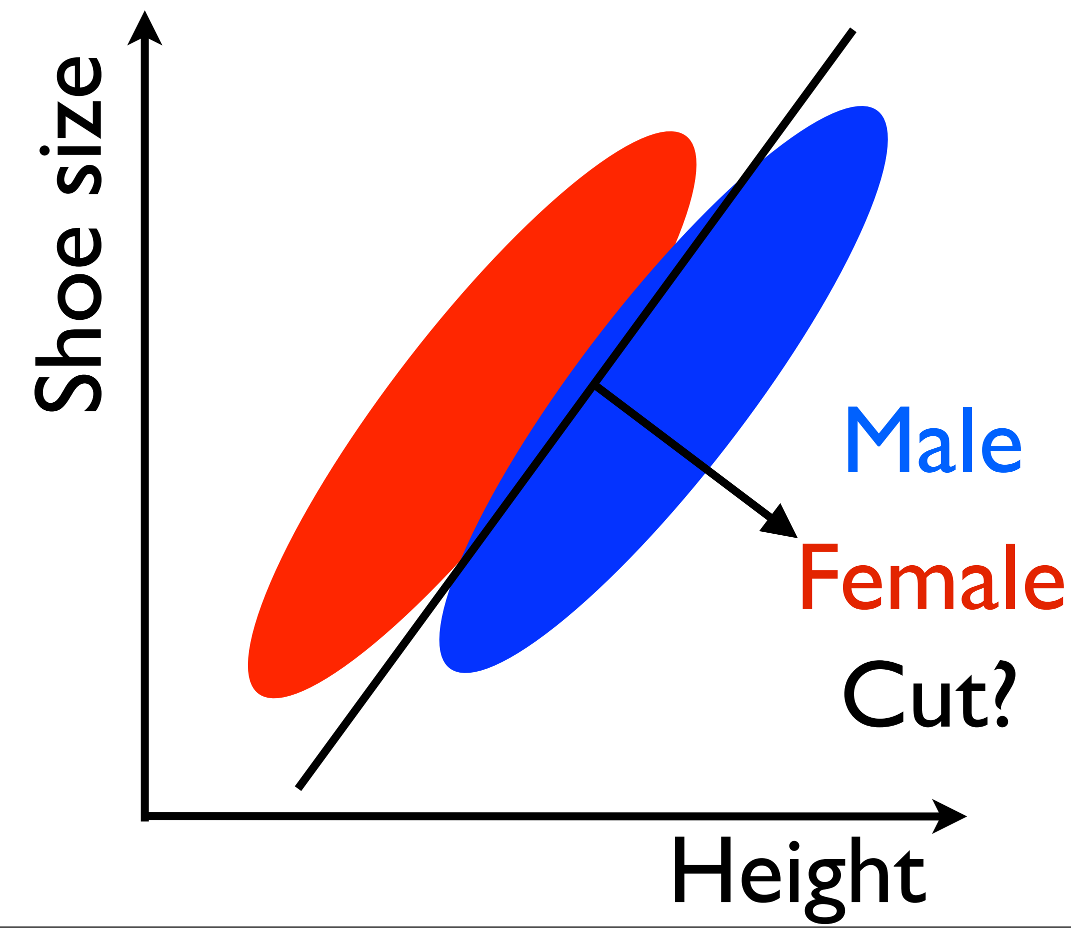
# Introduction - A simple example

- You want to figure out a method s.t. you are 95% sure a person is male:
- Your friend now gives you shoe size data as well:



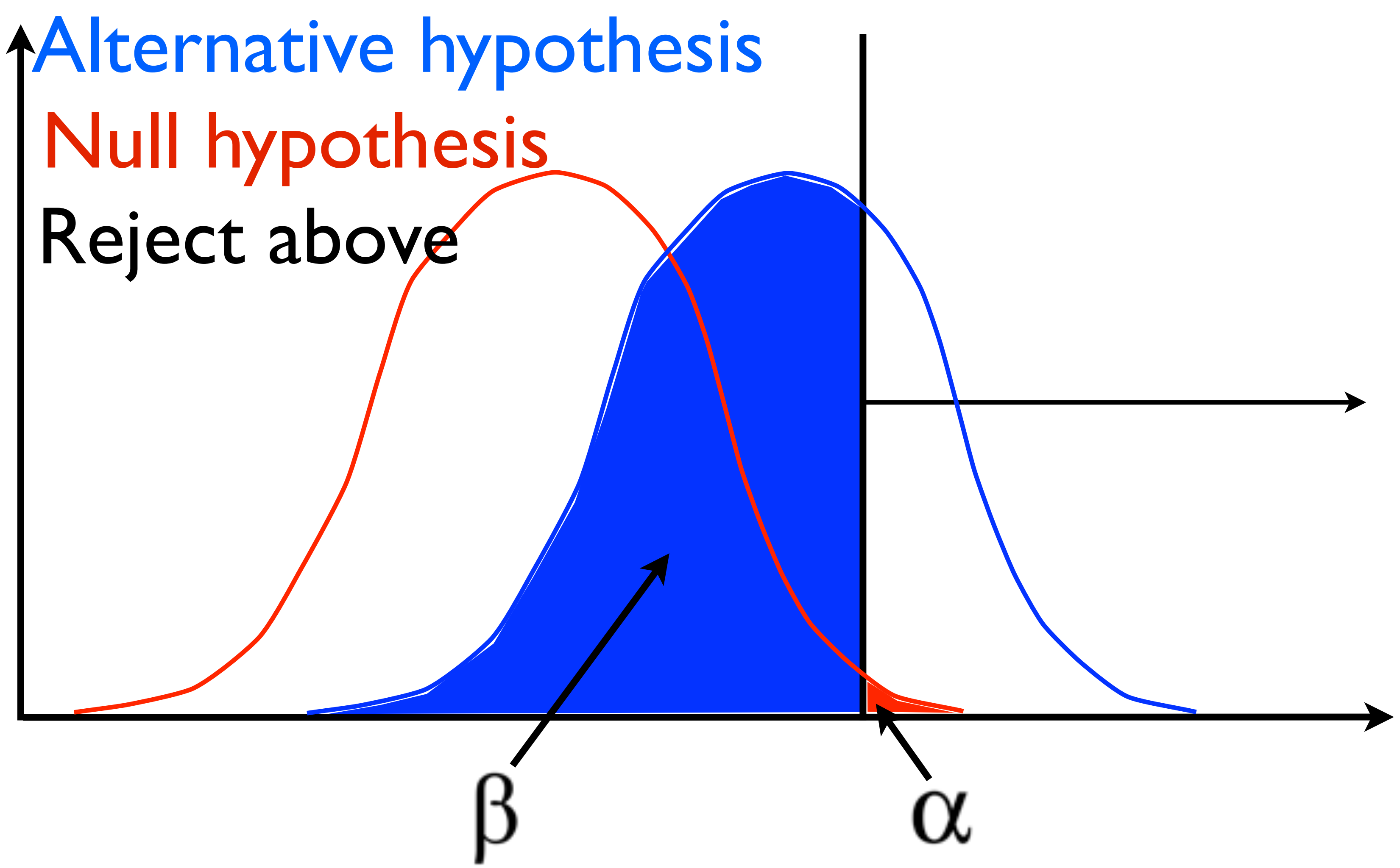
# Introduction - A simple example

- You want to figure out a method s.t. you are 95% sure a person is male:
- Add weight, etc. to your data.
  - Not intuitive how to define cut in N dimensions.
- Moreover, you are going to run out of statistics <u>very</u> fast if you try to populate an N dimensional histogram.
  - 10 bins in 9 dimensions will require $>10^9$ entries...
- <u>Introduce the Fisher discriminant</u>
- But first a short word on separation



# Separation of data

- Separation a Null and and Alternative Hypothesis
  - In general: Null is what you are ‘testing’. Alternative is what you are comparing to.
  - Example below is trying to keep as much Null as possible.

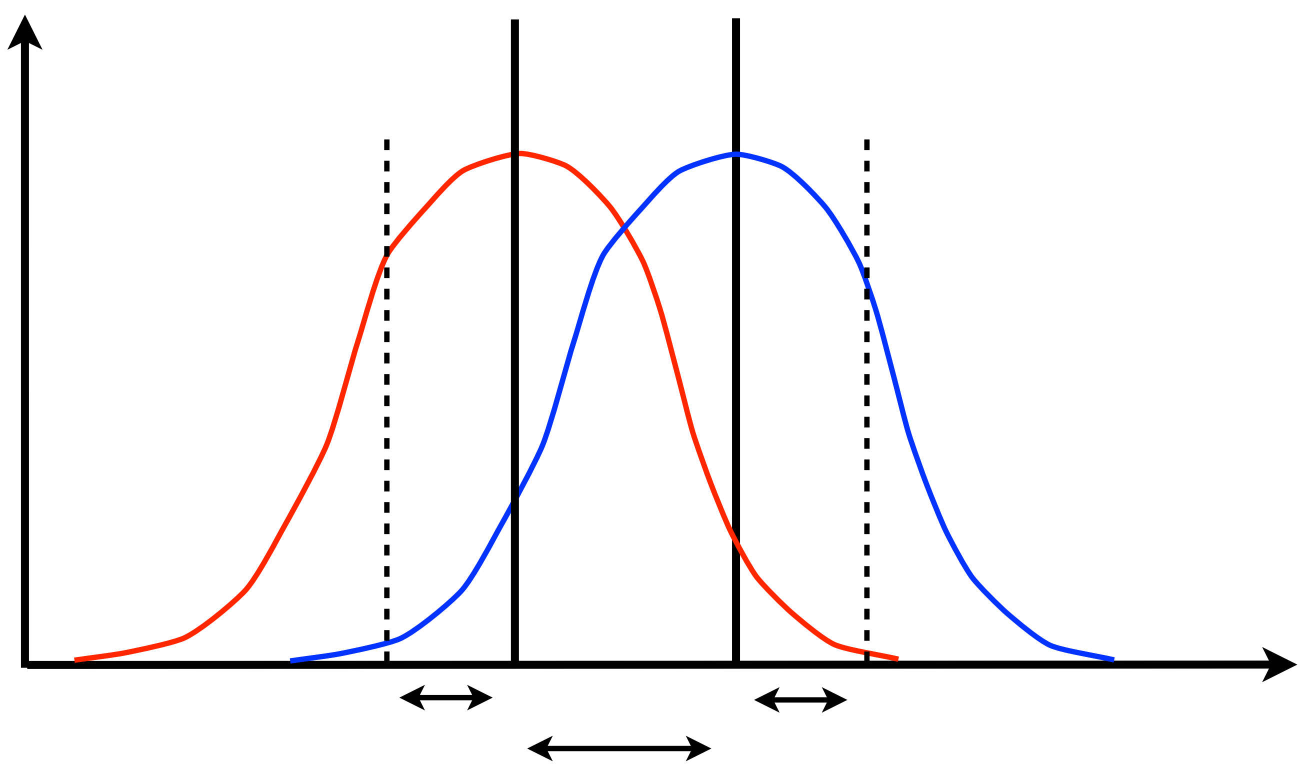



Reality

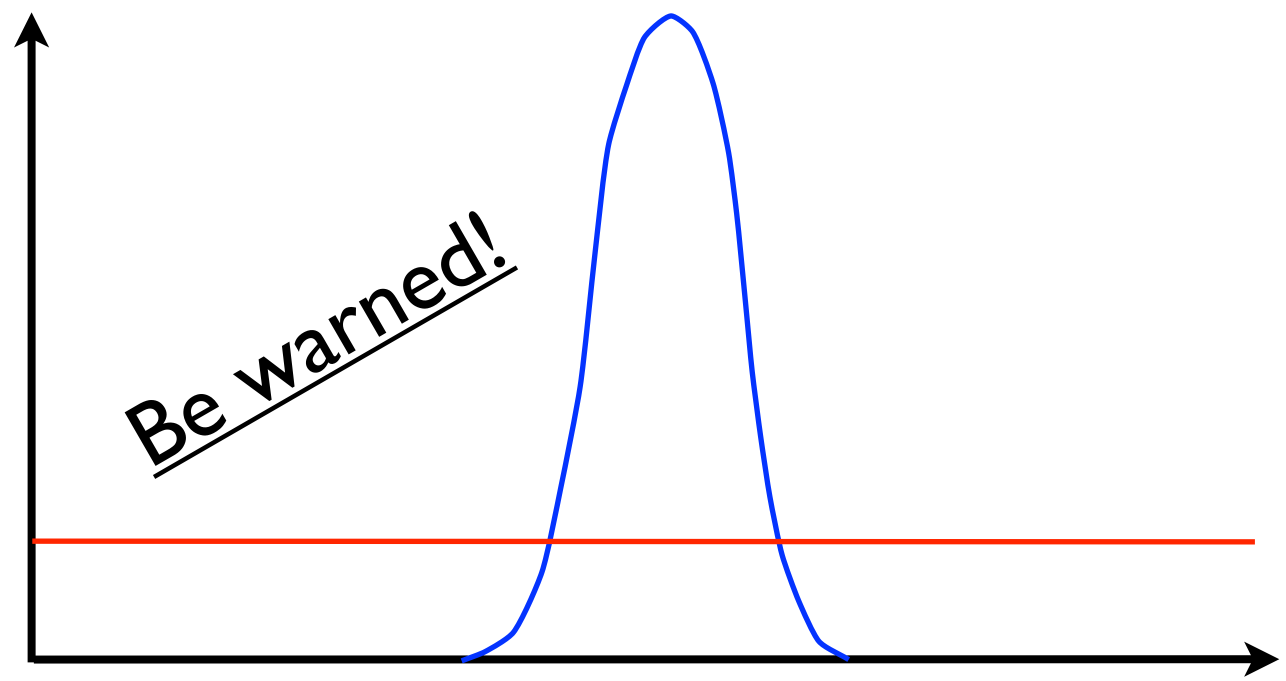
| Statistical decision | Null is true | Null is false |
| --- | --- | --- |
| Do not reject Null | 1-α Correct | β Type II error |
| Reject Null | α Type I error | 1-β Correct |

# Separation of data

- Separation a Null and and Alternative Hypothesis
- As always, there does not exist a general measure of separation.
- Today, we will use: $$\frac{(\tau_{Null} - \tau_{Alt})^2}{\Sigma^2_{Null} + \Sigma^2_{Alt}}$$



Intuitive: Distance between means in units of widths



# Multivariate Analysis

- Given a vector of data **x**, construct a test statistic to distinguish two hypotheses: $H_{Null}$ and $H_{Alt}$:

- Optimal: The Likelihood Ratio.

Discriminating function

$$\text{Test statistic} \longrightarrow t(\mathbf{x}) = \frac{f(\mathbf{x}|H_{Null})}{f(\mathbf{x}|H_{Alt})}$$

Vector of measurements $x = (x_1, \ldots, x_N)$

The value of the test statistic for a series of measurement can then be compared to what you know from simulation, or data of known outcome. This should tell you the probability that Null can describe your observations.

- Most real life cases, f can not be determined analytically.

- Estimated from simulation or data of known type ($H_{Null}/H_{Alt}$)

- Can in general be N dimensional

- Instead of populating histograms with many dimensions. Create an effective form of f and optimize under some criteria: Multivariate Analysis.

# The Fisher Discriminant

- Idea: Construct t as a linear combination of **x**:

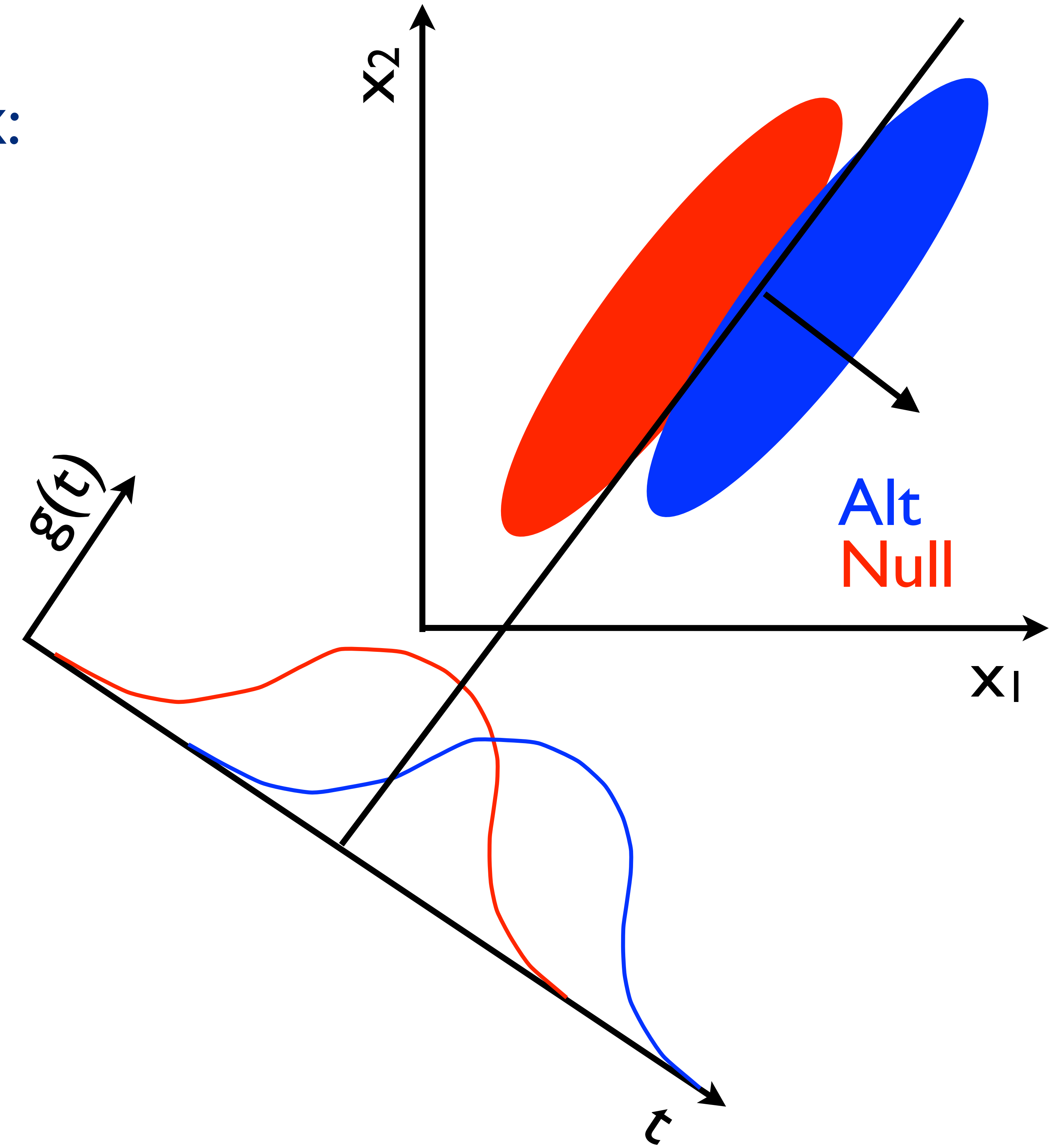
$$t(\mathbf{x}) = \sum_{i=1}^{N} a_i x_i = a^T x$$

- We just need to figure out what a is.

- Goal is to construct them s.t. we have the maximal separation between the functions:

$$g(t|H_{Null}) \qquad g(t|H_{Alt})$$

- This requires a bit of Linear Algebra



# The Fisher Discriminant

- Start by calculating mean and covariance for Null and Alt:

$$(\mu_k)_i = \int x_i f(x|H_k) dx_1 \dots dx_N$$

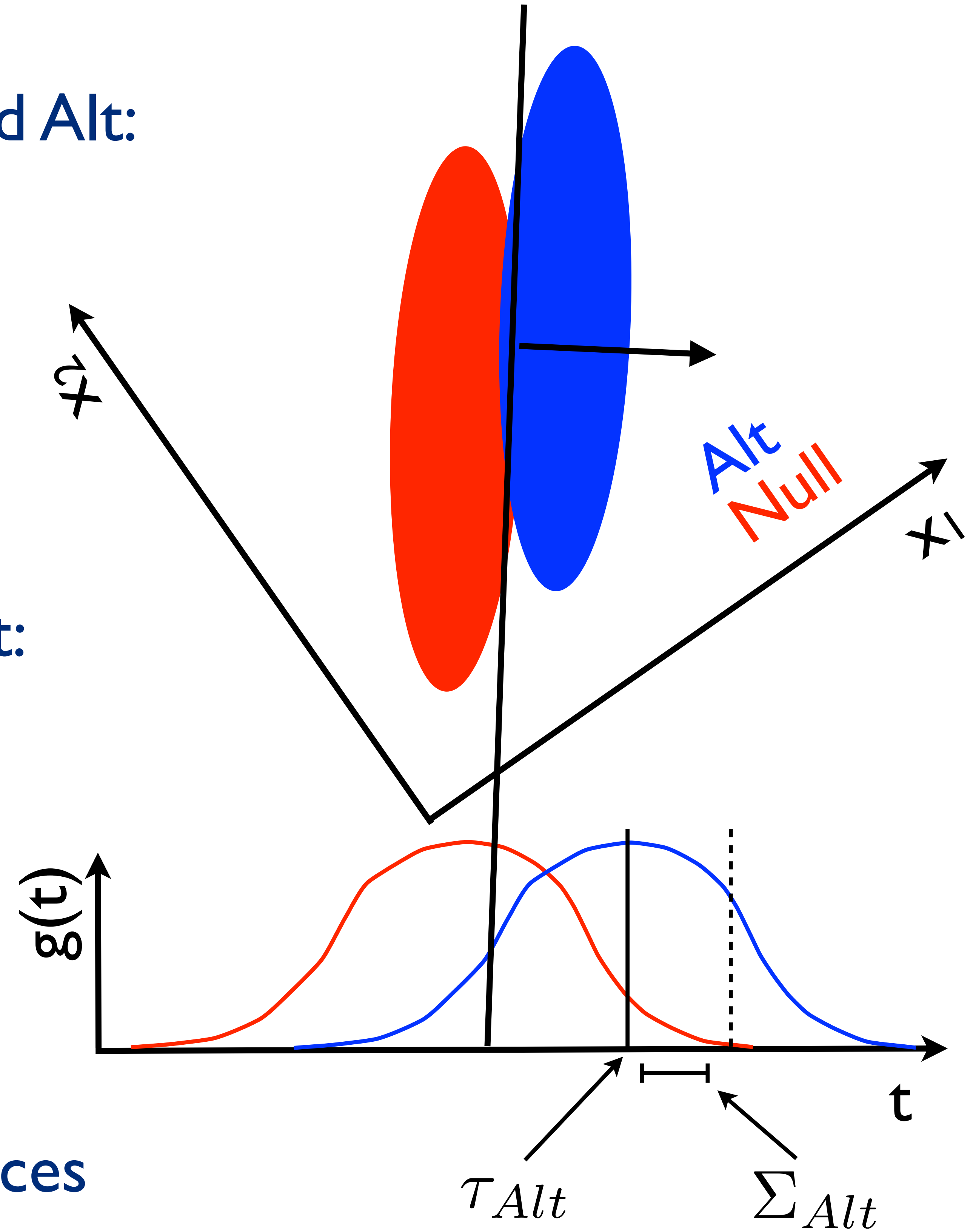
$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(x|H_k) dx_1 \dots dx_N$$

- Similarly, we can calculate the mean and variance of t:

$$\tau_k = \int t g(t|H_k) dt = a^T \mu_k$$

$$\Sigma_k^2 = \int (t-\tau_k)^2 g(t|H_k) dt = a^T V_k a$$

  - k's indicate Null or Alt, i's/j's are measurement indices



# The Fisher Discriminant

- Find a s.t. the following separation is optimized:

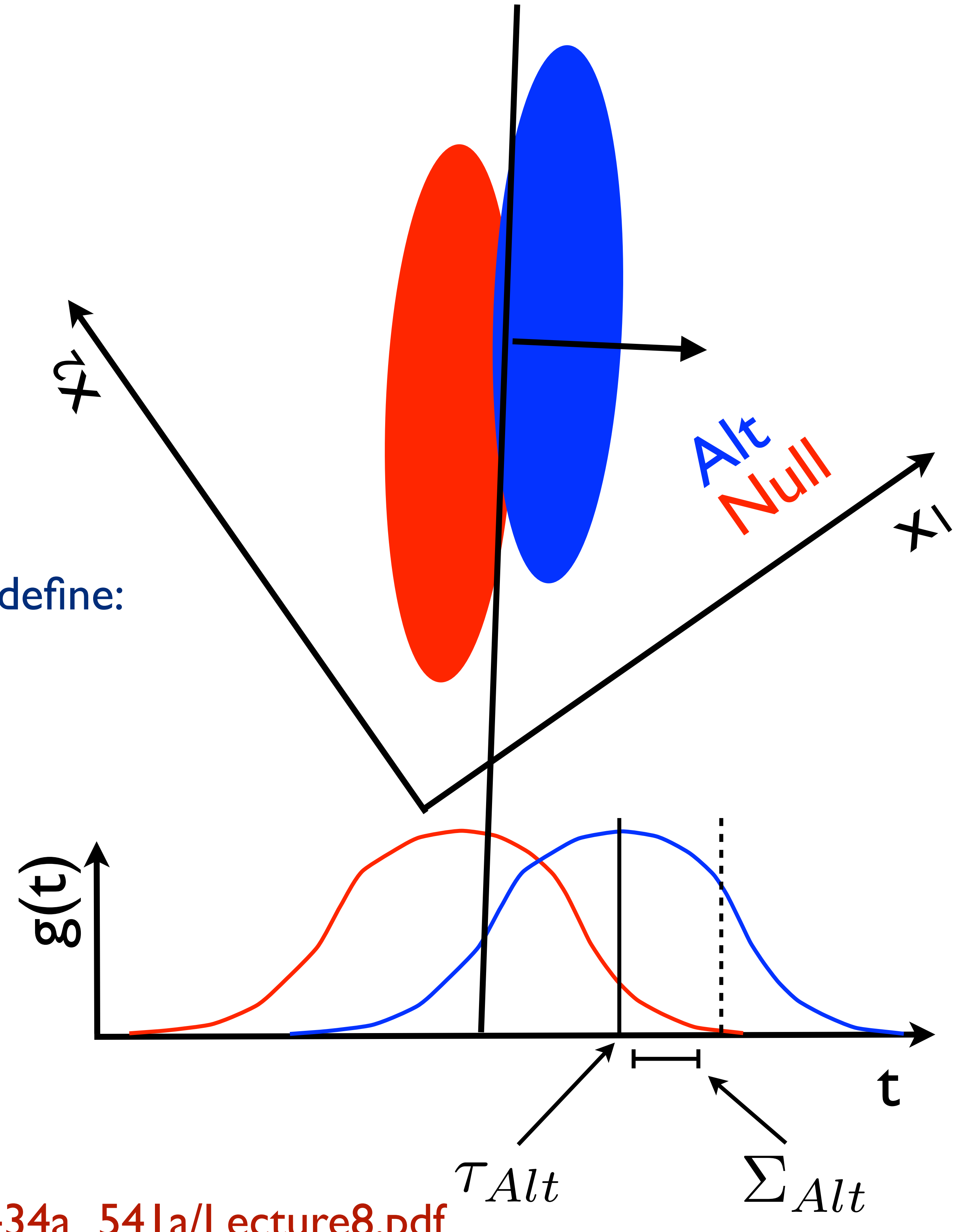
$$J(a) = \frac{(\tau_{Null} - \tau_{Alt})^2}{\Sigma_{Null}^2 + \Sigma_{Alt}^2}$$

- Derive and set equal to zero:
  - Go trough rather lengthy derivation (see link)... Turns out it is useful to define:

$$W_{ij} = (V_{Null} + V_{Alt})_{ij}$$

- Resulting in a should be proportional to:

$$a \propto W^{-1}(\mu_{Null} - \mu_{Alt})$$



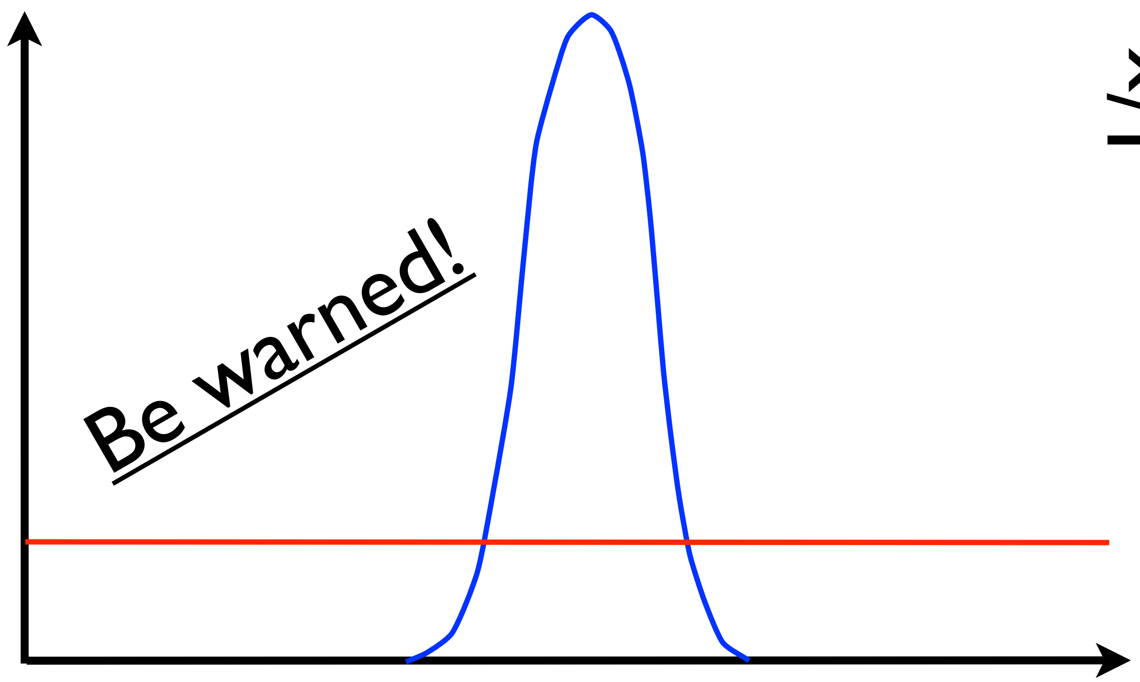

Formal derivation can be found here: http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture8.pdf

# The Fisher Discriminant - Summary

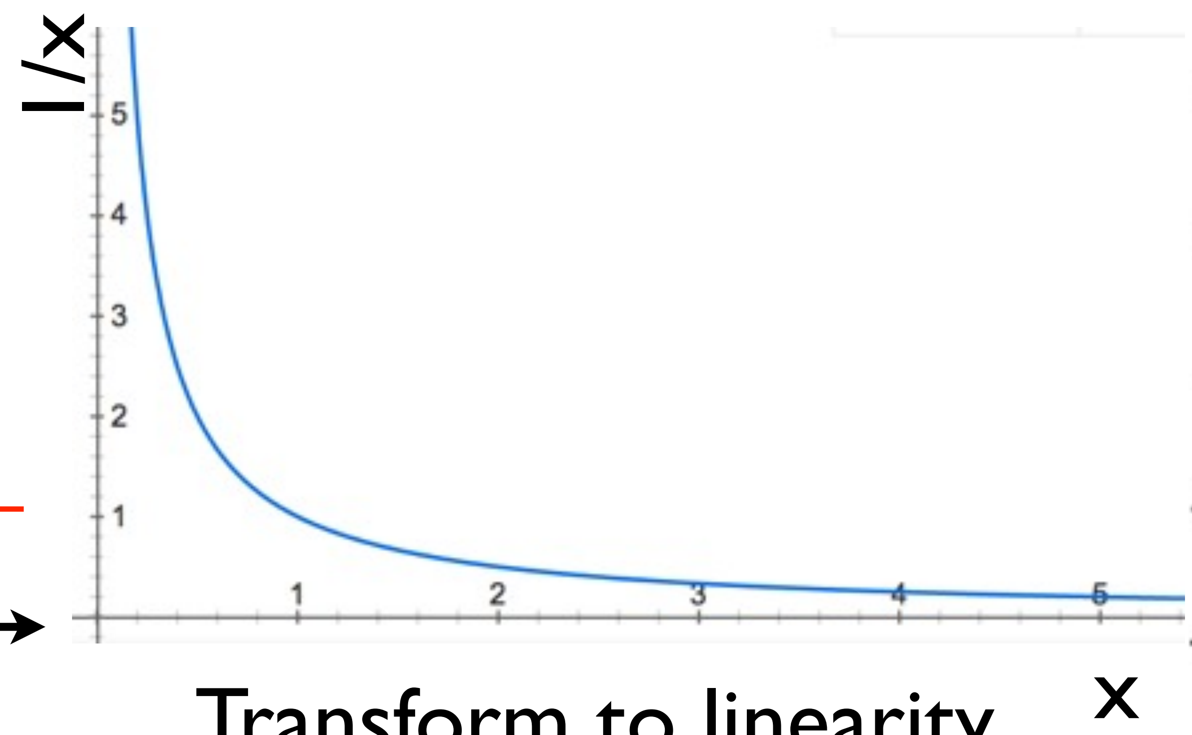
- In order to construct the linear discriminant:
- Calculate means from 'training' data sample: $(\mu_k)_i = \int x_i f(x|H_k) dx_1 \ldots dx_N$
- Calculate covariance matrices: $(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(x|H_k) dx_1 \ldots dx_N$
- Sum and invert covariance matrices: $W_{ij} = (V_{Null} + V_{Alt})_{ij}$
- a can now be calculated from this: $a \propto W^{-1}(\mu_{Null} - \mu_{Alt})$
- Finally the Fisher discriminant can be calculated: $t(\mathbf{x}) = \sum_{i=1}^{N} a_i x_i = a^T x$

# The Fisher Discriminant - Summary

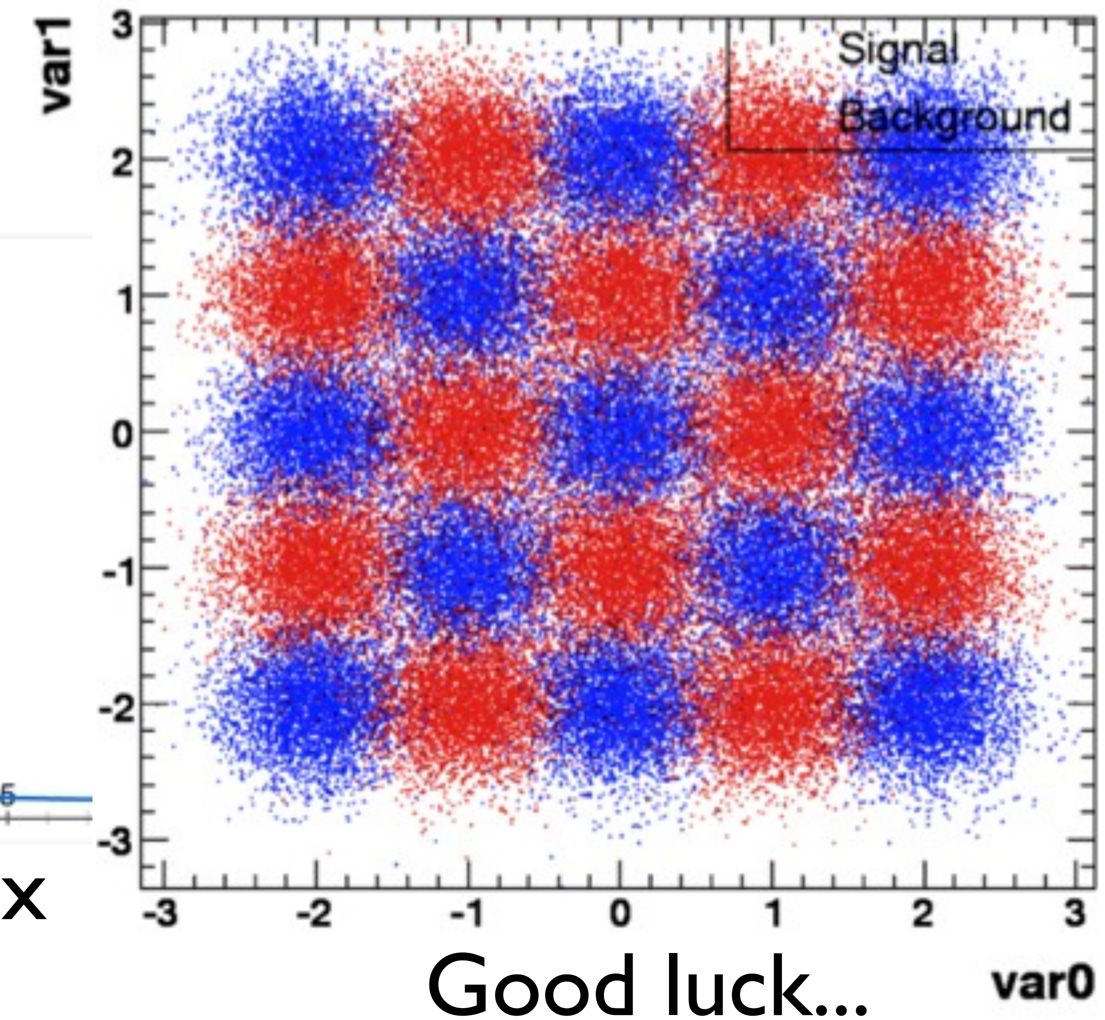
- Having constructed the ‘optimal’ observable, it is always worthwhile looking back and see what was assumed.




Define variables properly




Transform to linearity



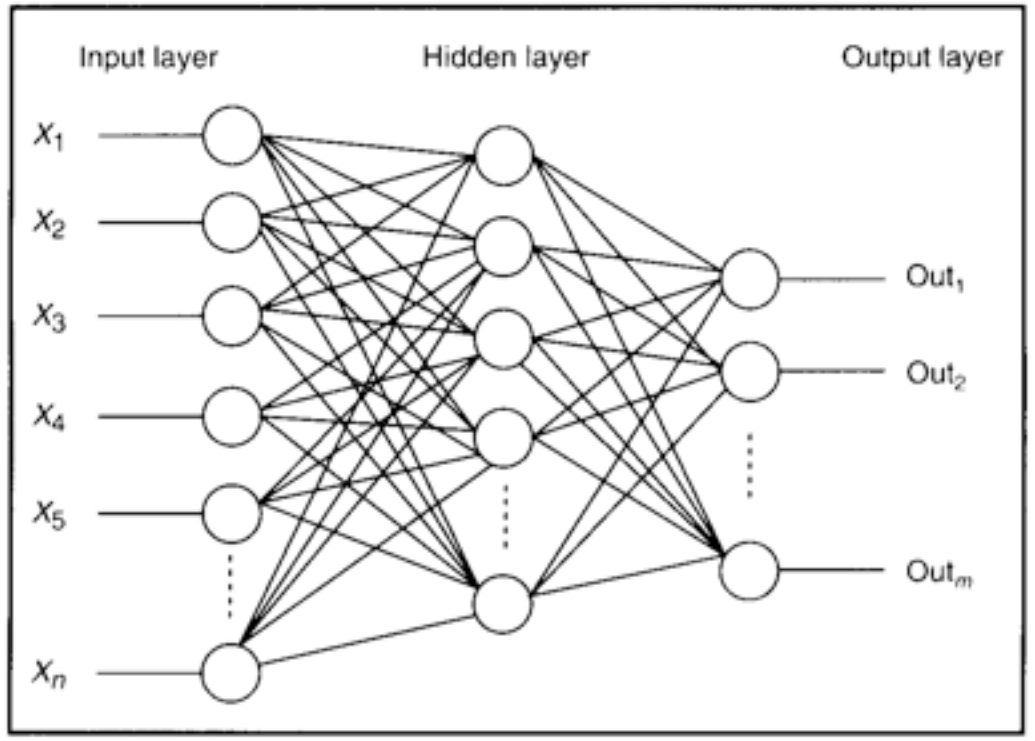

Good luck...

- Needs linearly correlated left/right separated data.

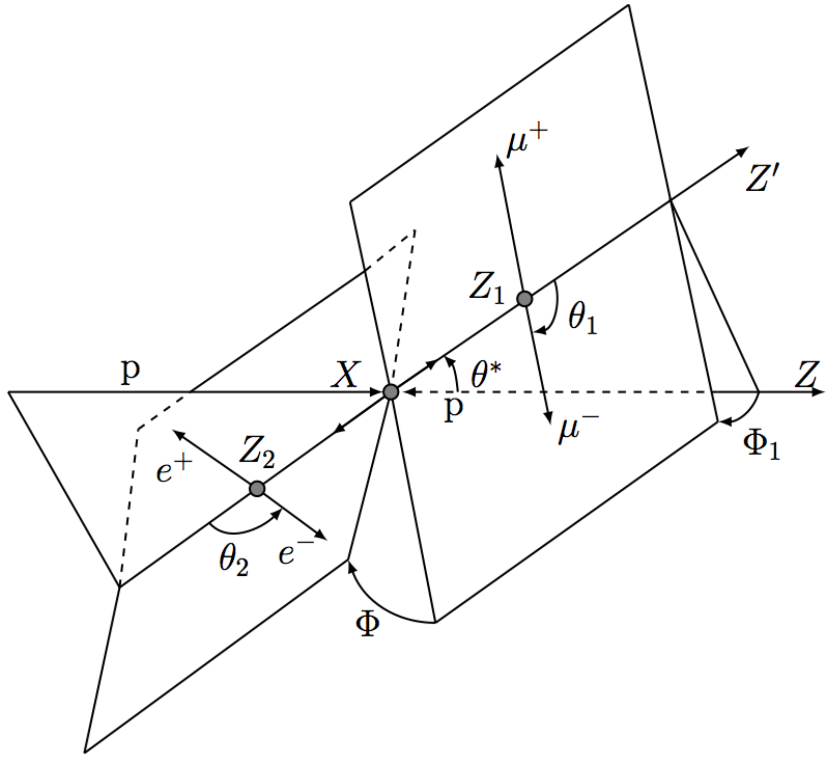
- That being said, the (relative) simplicity of the method makes it useful in a large variety of real life experiments.

# MVA summary

- In a perfect world, it would be possible to calculate the Likelihood ratio between any Null and Alternative hypothesis.
- Ratio can be tried to be estimated by histograms. I.e. using simulation or data, gathered before the experiment.
- As the number of factors that influence the likelihood grow, this becomes more and more unreasonable. Required statistics grow as res$^{\text{Dimension}}$.

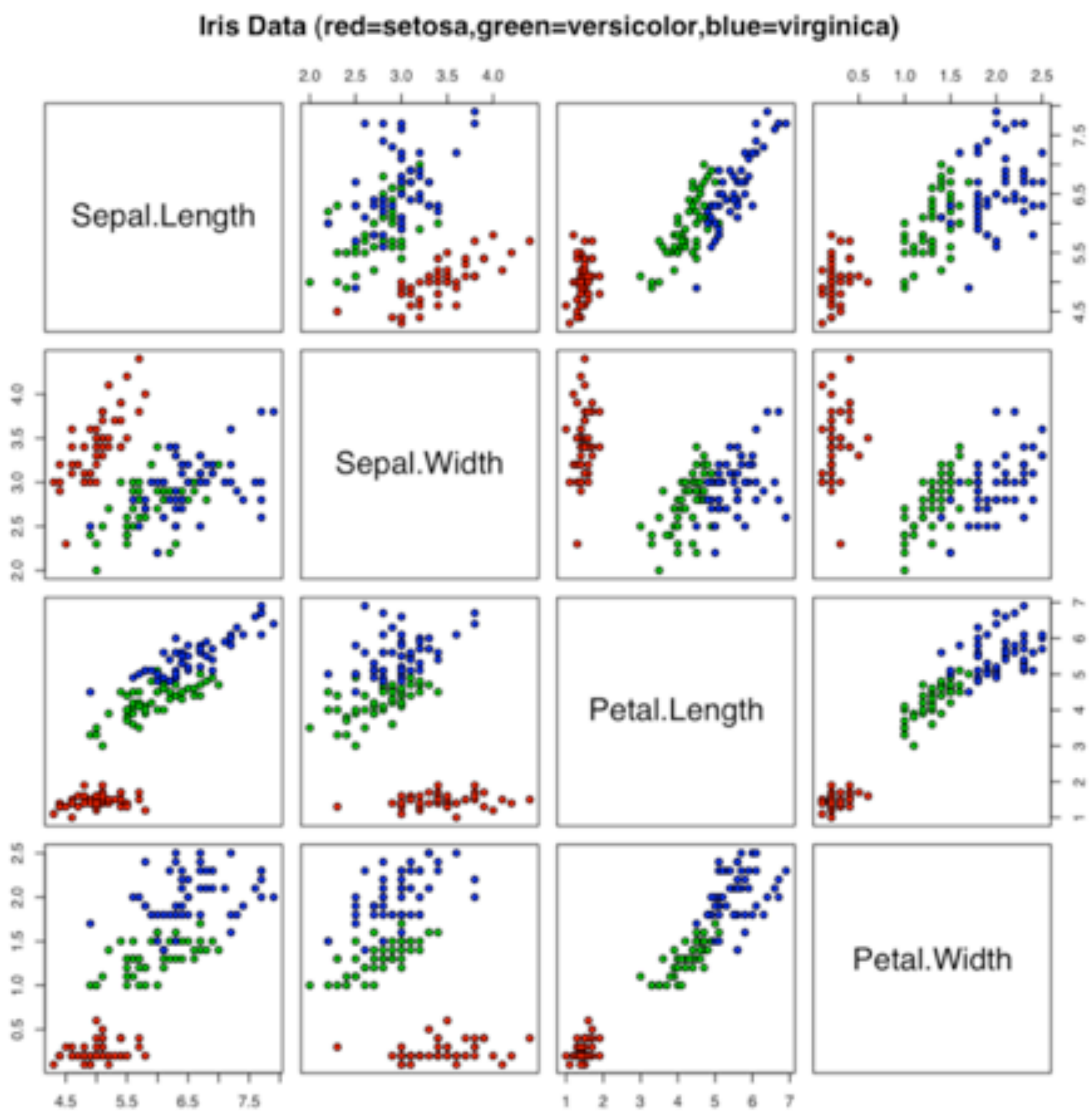






# The Fisher Discriminant

- Todays exercise:
  - Construct discriminant for data set that originally inspired Fisher.
  - Data: Measurements of Irises picked by Fisher's friend Anderson on the Gaspe peninsula...




Iris Virginica

Iris Versicolor

Iris Setosa

# Multivariate Analysis

- Next time:
  - Elaborate on separation measures: More applicable tools for real life analysis
  - Introduce tools (superficially) that can handle more complex examples.
  - Risk factors contributing to heart disease, based on data collected in South Africa.