# Event Classification

# Information Extract

The LHC experiments are expensive

~ $\$10^{10}$ (accelerator and experiments)

the competition is intense

(ATLAS vs. CMS) vs. Tevatron

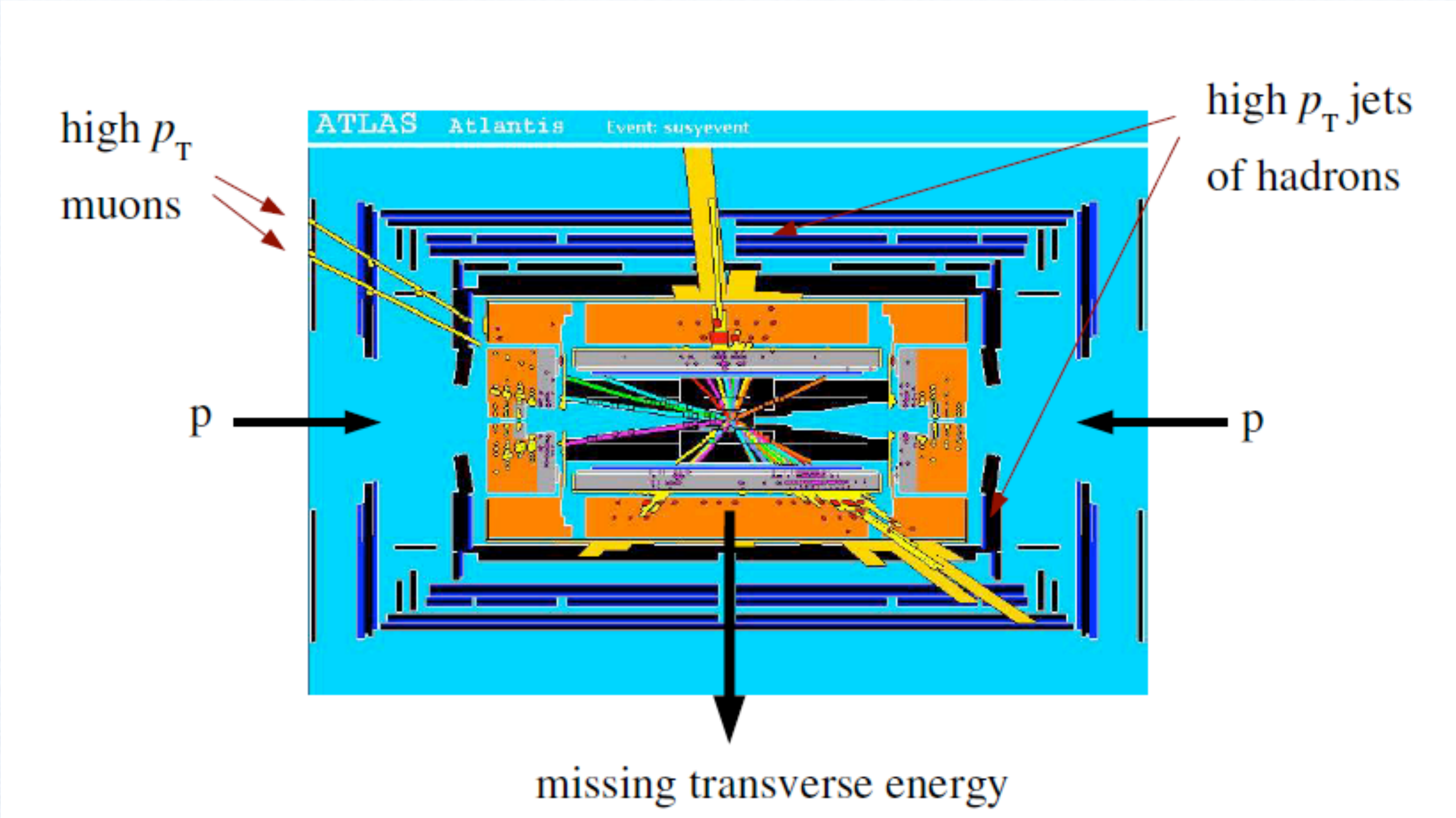
and the stakes are high:


4 sigma effect → (goat)

5 sigma effect → (medal)

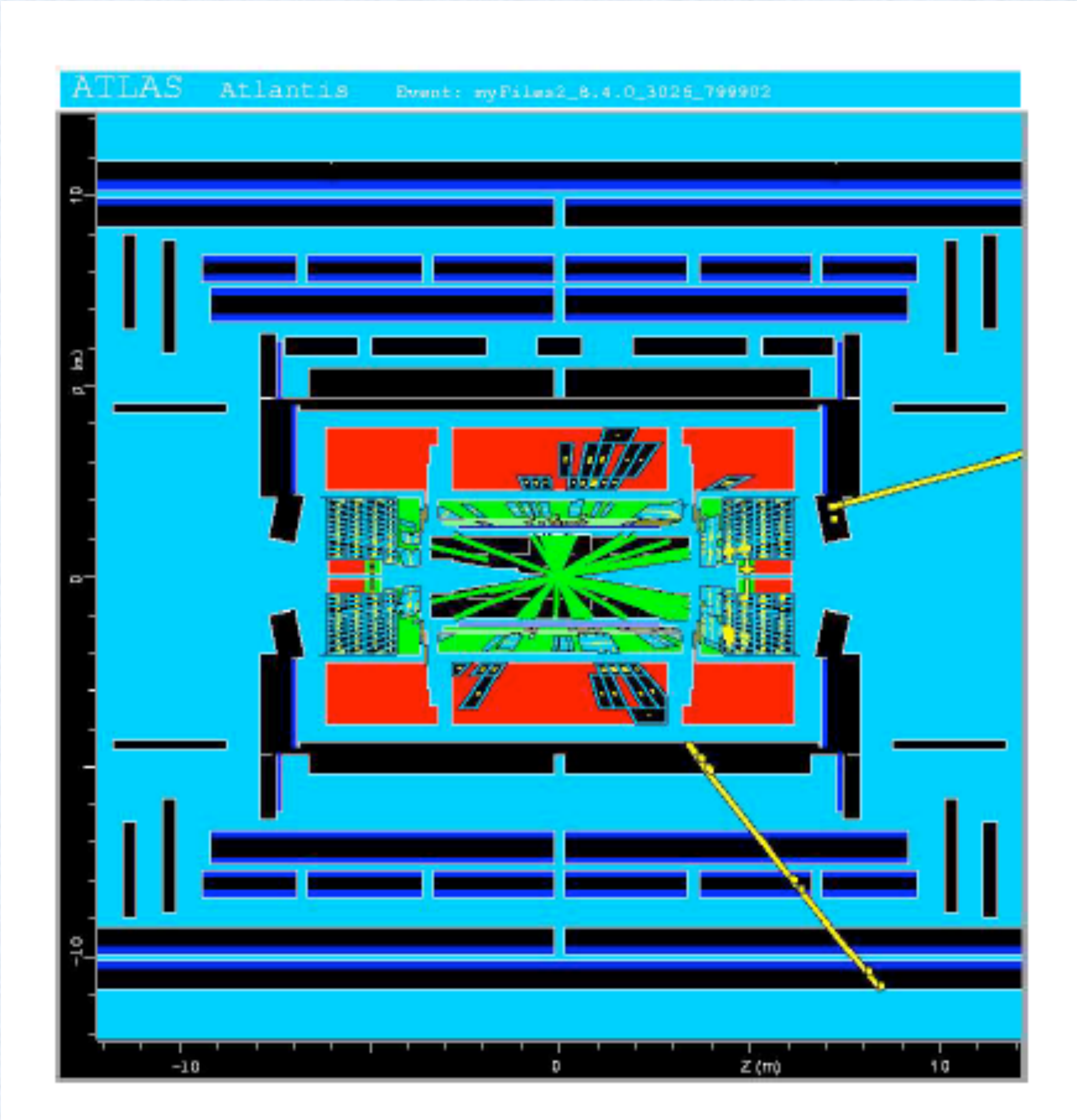
So there is a strong motivation to extract all possible information from the data.

# Signal Event: SUSY



# Background Event

- top quark event



# Basic Problem

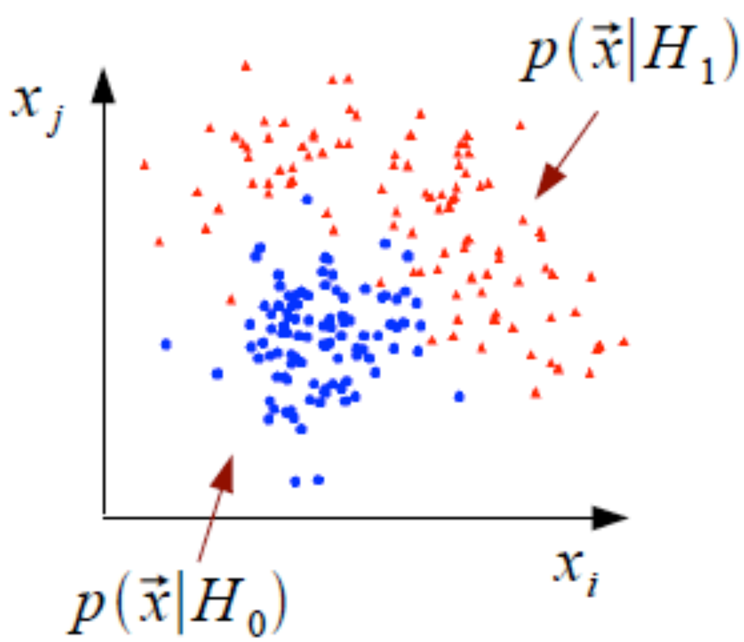
Suppose for each event we measure a set of numbers $\vec{x} = (x_1, \ldots, x_n)$

$x_1$ = jet $p_T$

$x_2$ = missing energy

$x_3$ = particle i.d. measure, ...

$\vec{x}$ follows some $n$-dimensional joint probability density, which depends on the type of event produced, i.e., was it $pp \to t\bar{t}$, $pp \to \tilde{g}\tilde{g}, \ldots$




E.g. hypotheses (class labels) $H_0$, $H_1$, ...

Often simply "signal", "background"

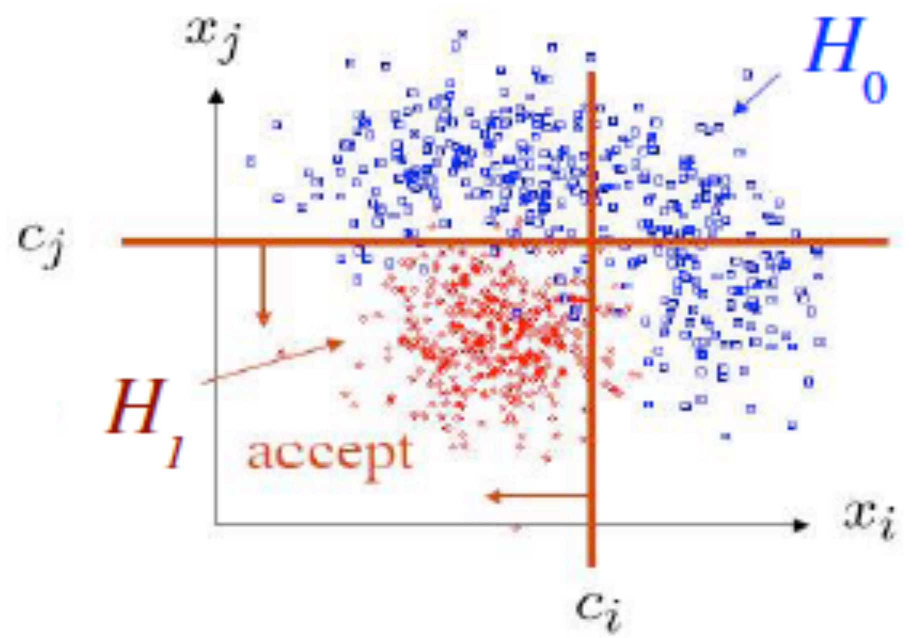
We want to separate (classify) the event types in a way that exploits the information carried in many variables.

# GOAL

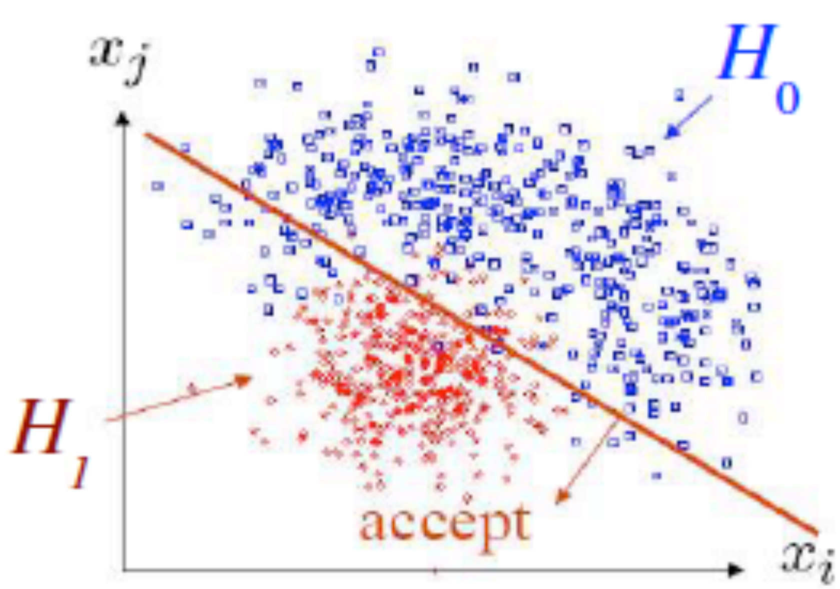
Maybe select events with "cuts":

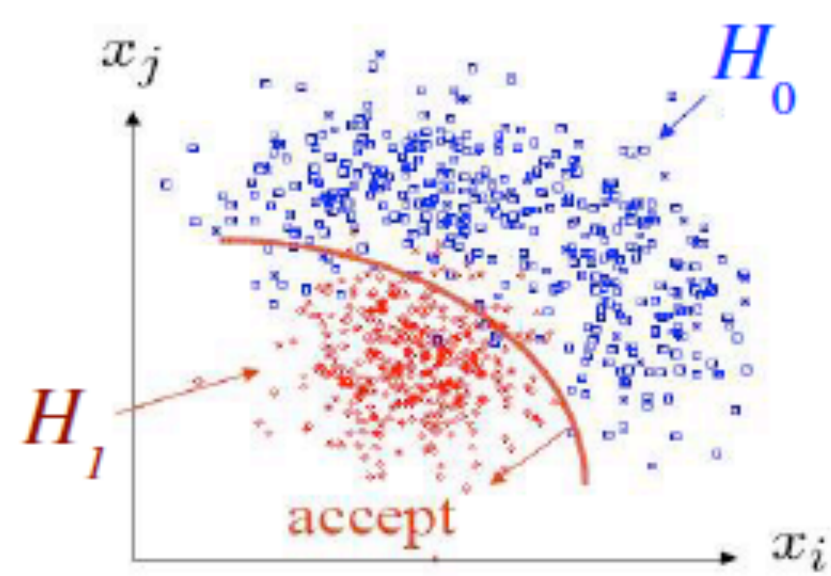
$x_i < c_i$

$x_j < c_j$



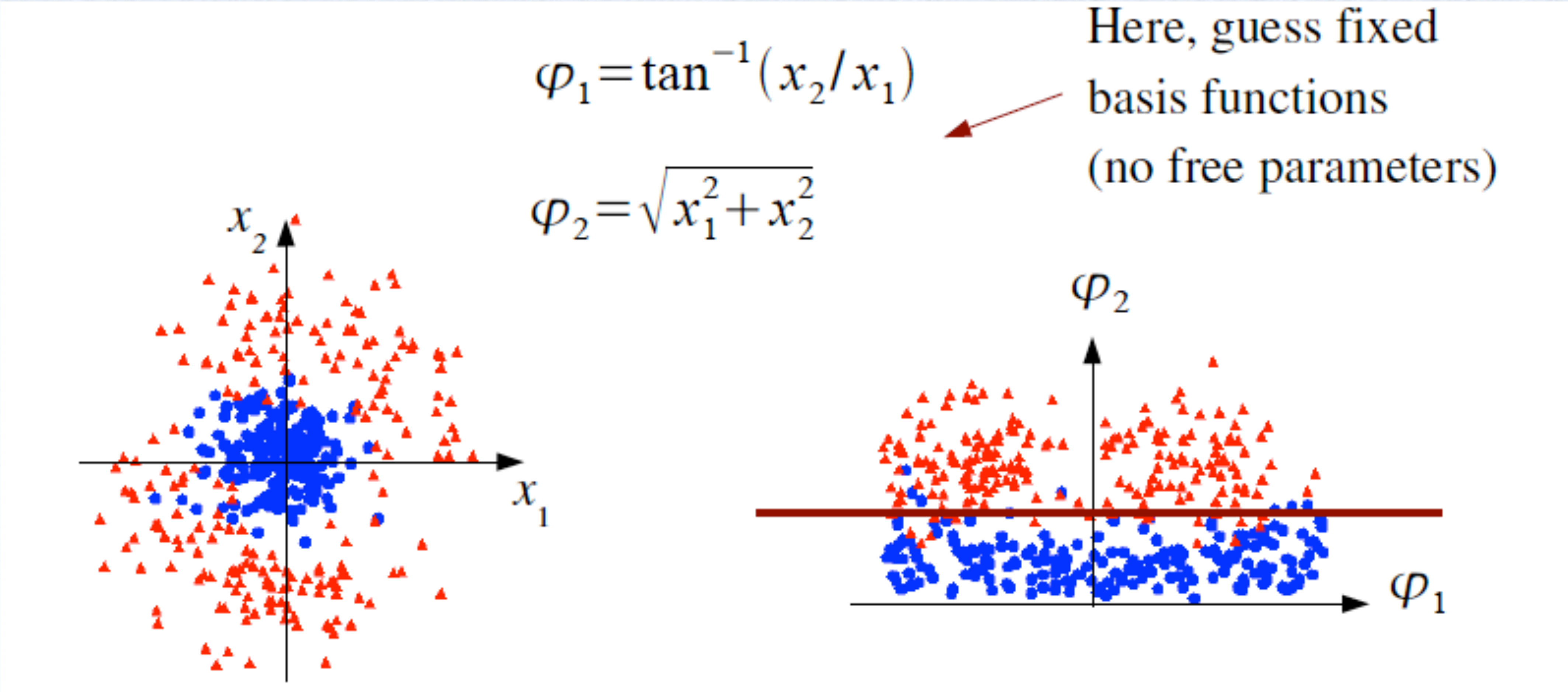

Or maybe use some other type of decision boundary:



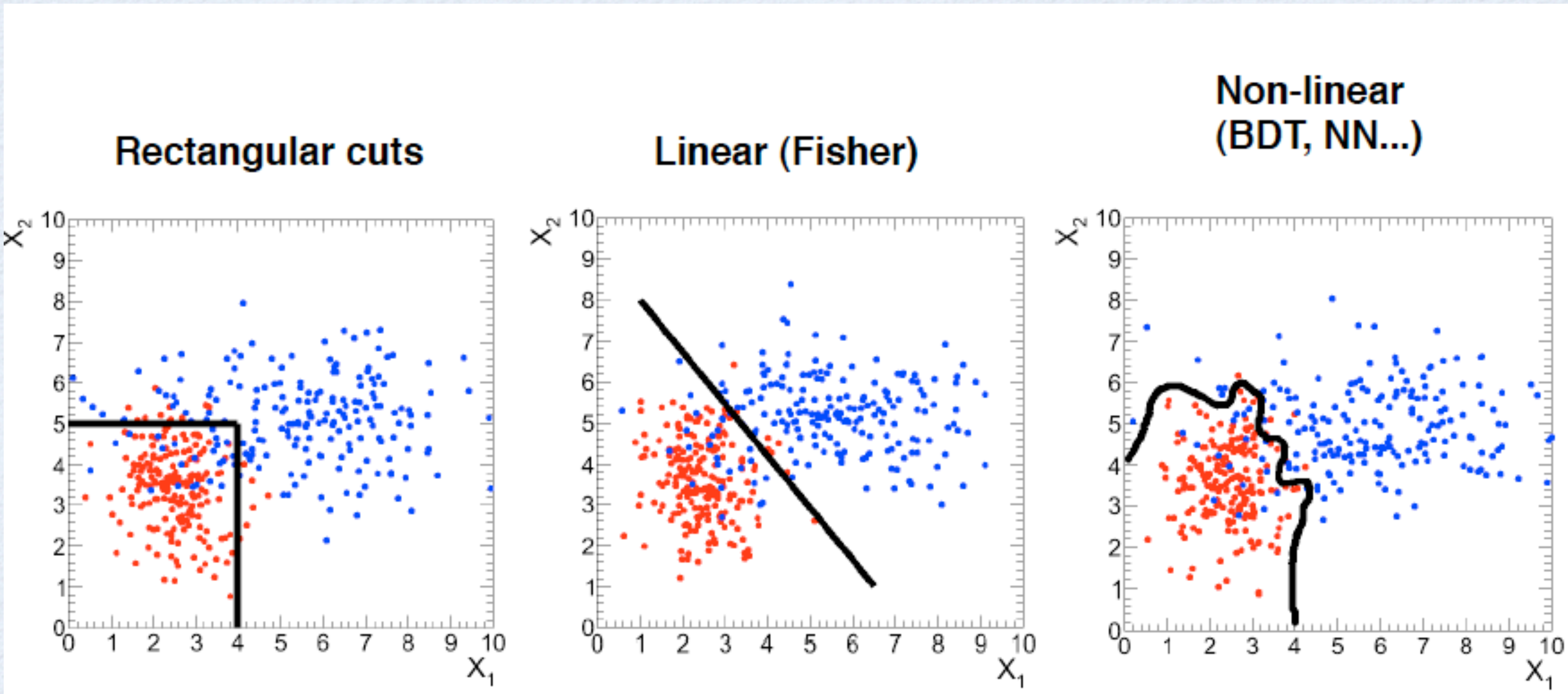

Goal of multivariate analysis is to do this in an "optimal" way.

- We can try to find a transformation so that the transformed "feature space" variables can be separated better by a linear boundary:



# Possible solutions

| optimized cuts | transform or find relation | optimized solution |
|---|---|---|

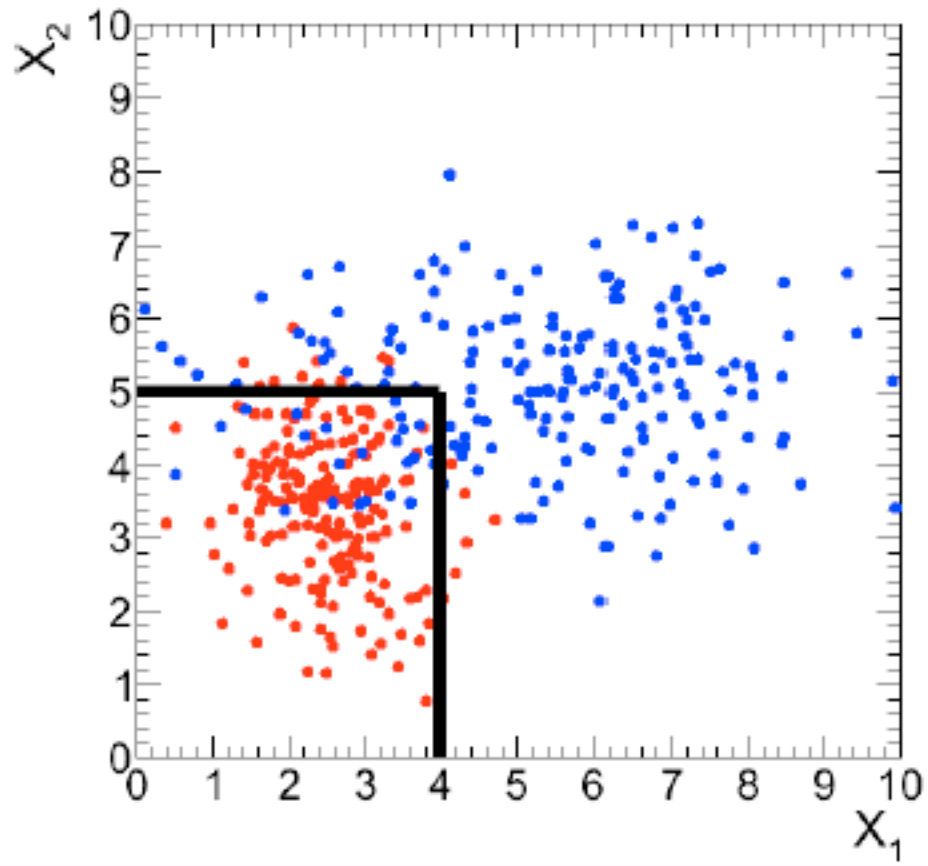


# Rectangular cuts

- Simplest multivariate method, very intuitive
- All HEP analyses are using rectangular cuts, not always completely optimized

**Rectangular cuts optimization :**
- Grid search, Monte-Carlo sampling
- Genetic algorithm
- Simulated annealing

**Characteristics :**
- Difficult to discriminate signal from background if too much correlations
- Optimization difficult to handle with high number of variables




Define the signal region :

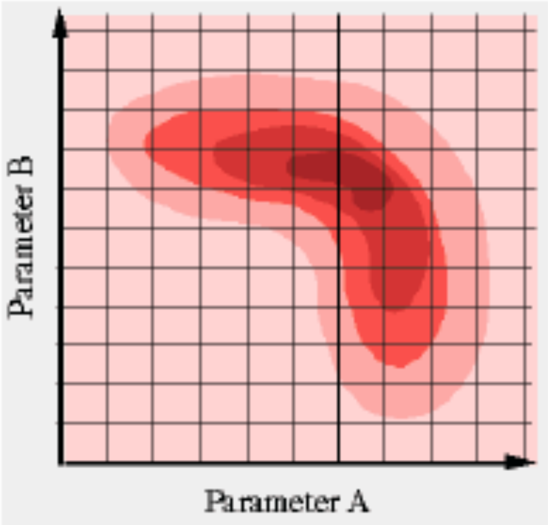
$a1 < x1 < a2$,

$b1 < x2 < b2$

...

# CHOOSING CUTS

## How to find the best set of cuts for a given criterion ?




### Grid search

- Try N points (usually very large) of the phase-space equally spaced in each dimensions

=> Impossible with high number of variables (too much CPU time)

### Monte-Carlo sampling

- Try N points randomly chosen in the phase space

=> Usually performs better, but still non optimal

Both are good global minimum finder but have poor accuracy

### Examples of criterion :

- Maximize the signal efficiency for a given background rejection
- Maximize the significance

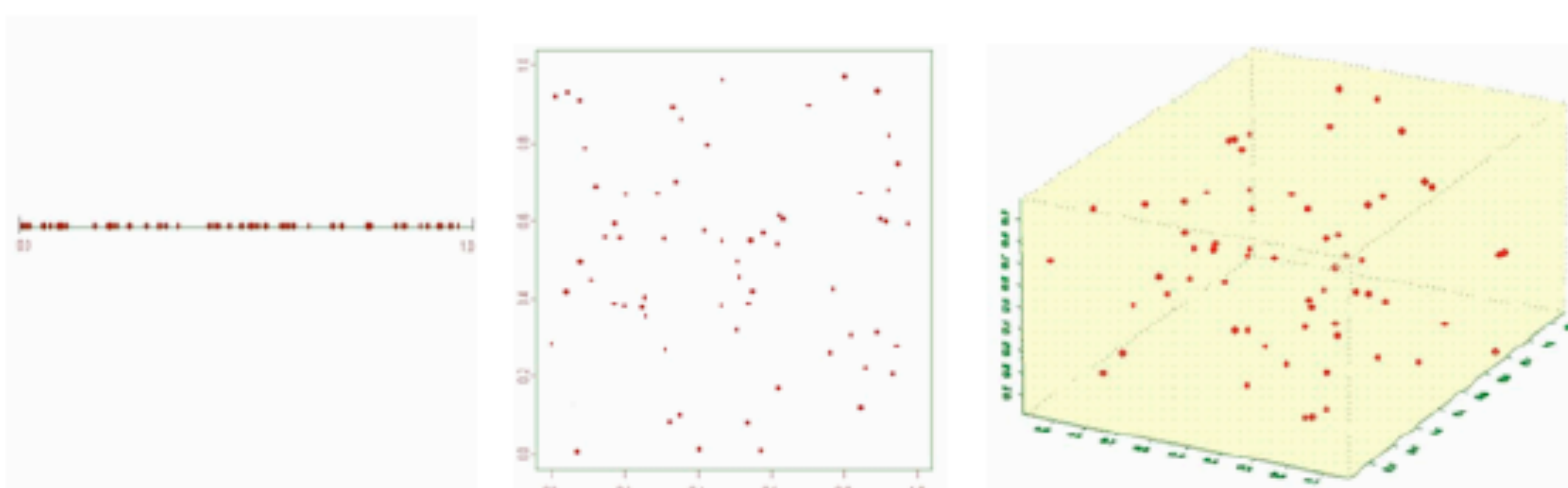
# DIMENSIONALITY

**Grid search and Monte-Carlo sampling suffer from the curse of dimensionality :**

- For one variables, trying 100 working points is easy

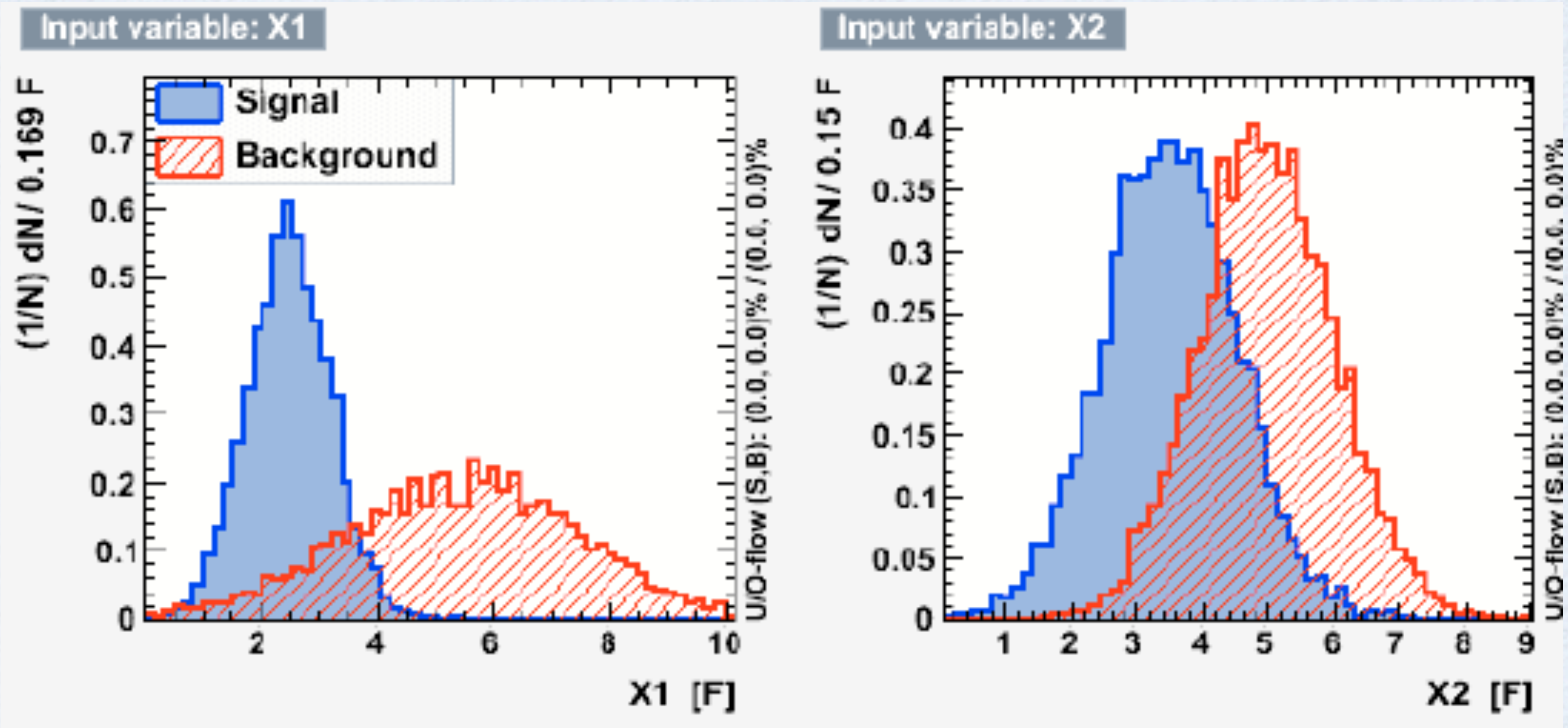
- For two variables, 100 working points will lead to not well covered phase-space because each points have more distance between them
- 100x100 points should be used

- Increasing number of variables will lead this algorithm to be impossible in practice



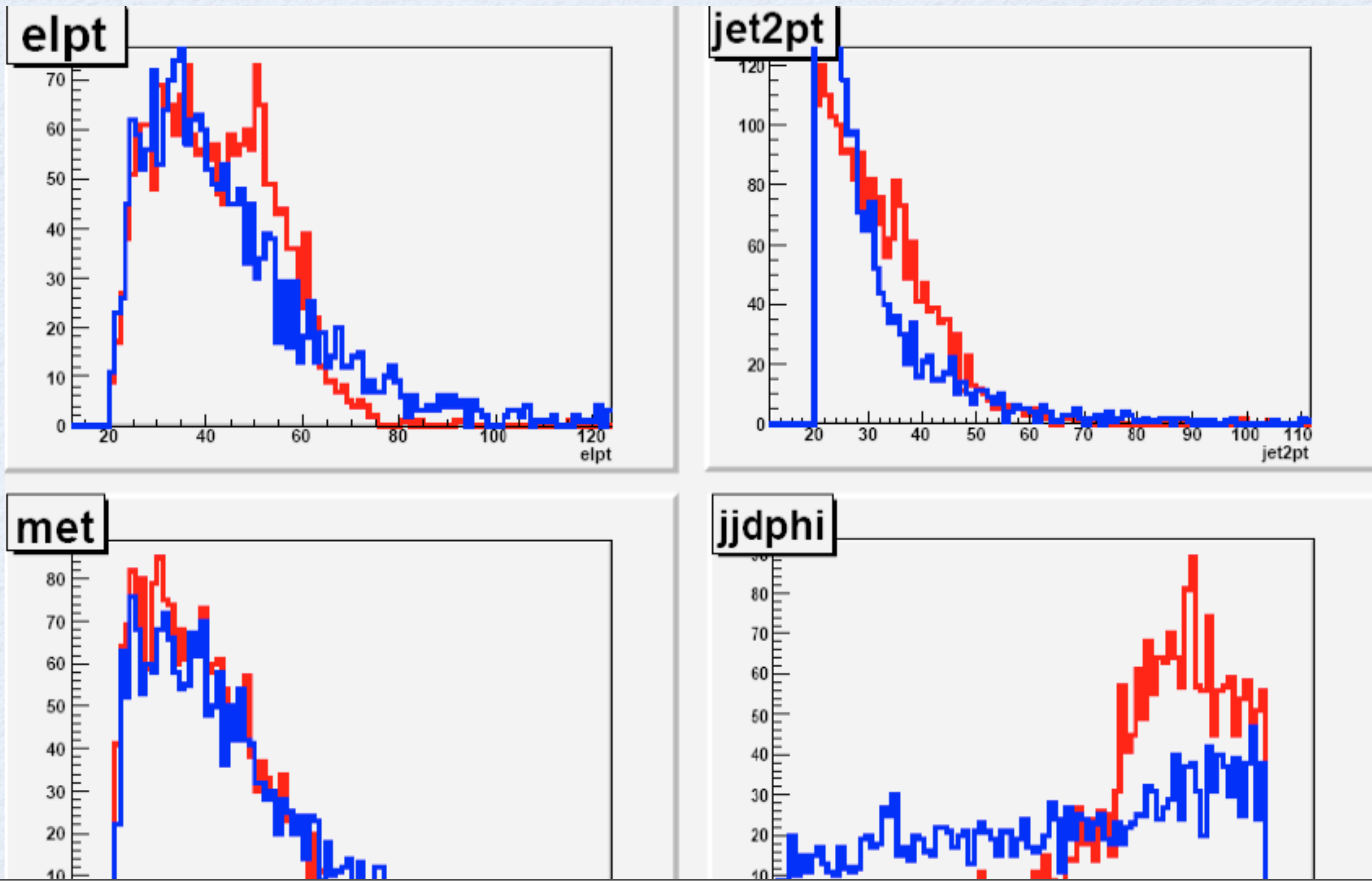
# Multiple Variables

Might be able to use `square cuts'
if your distributions look like



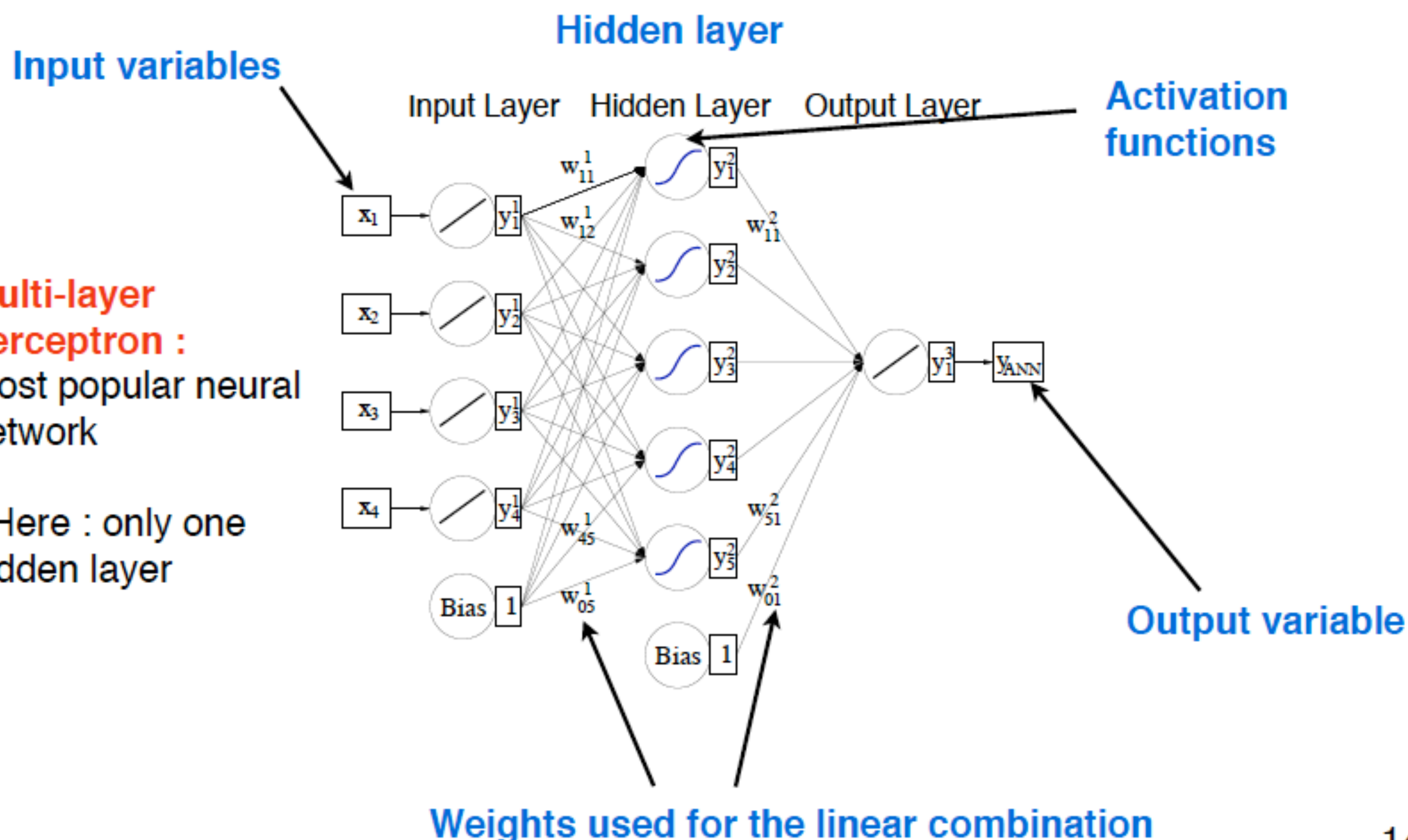
# Multivariate Methods

But what if?

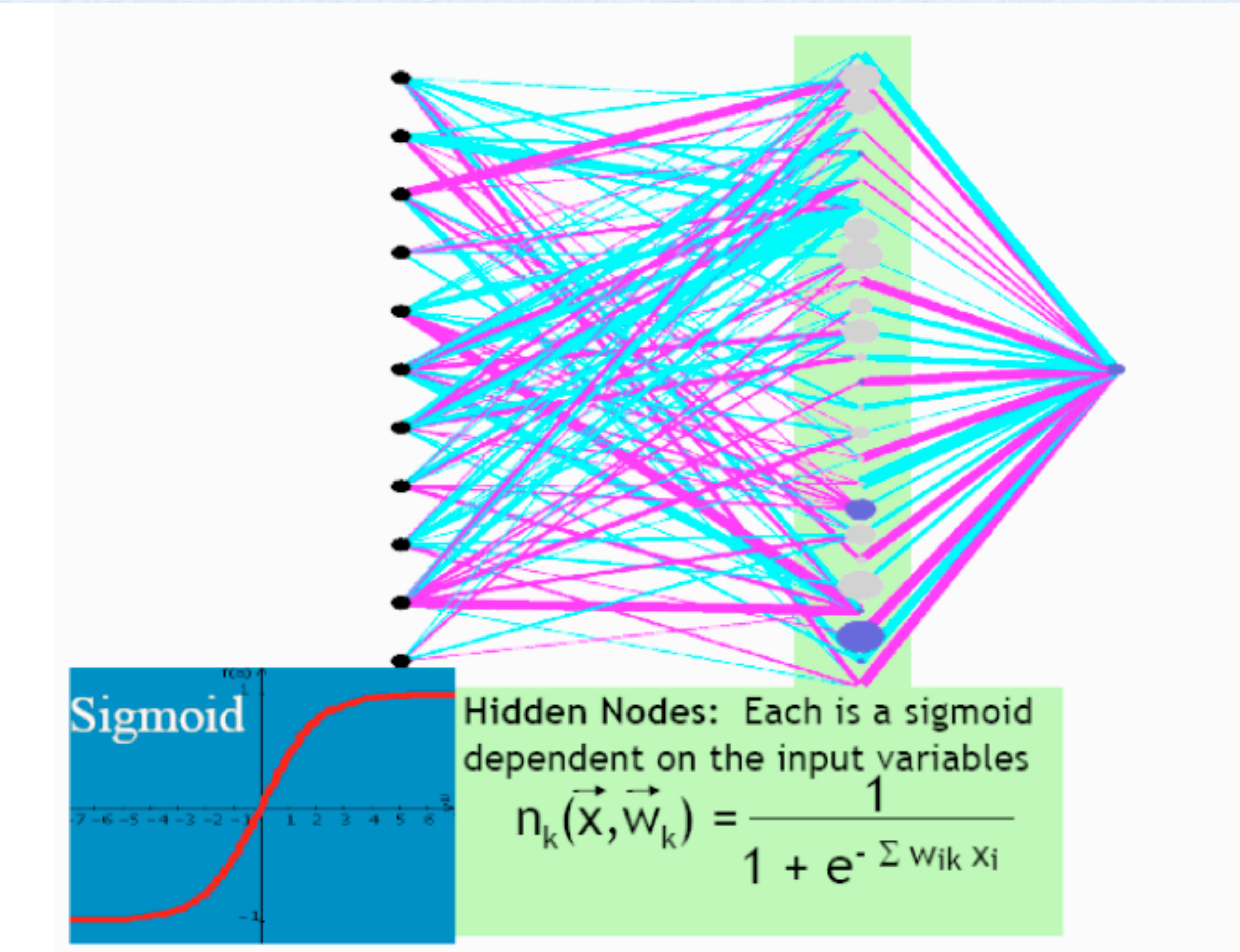


# Neural Networks

- **Multi-layer perceptron :** most popular neural network
- Here : only one hidden layer

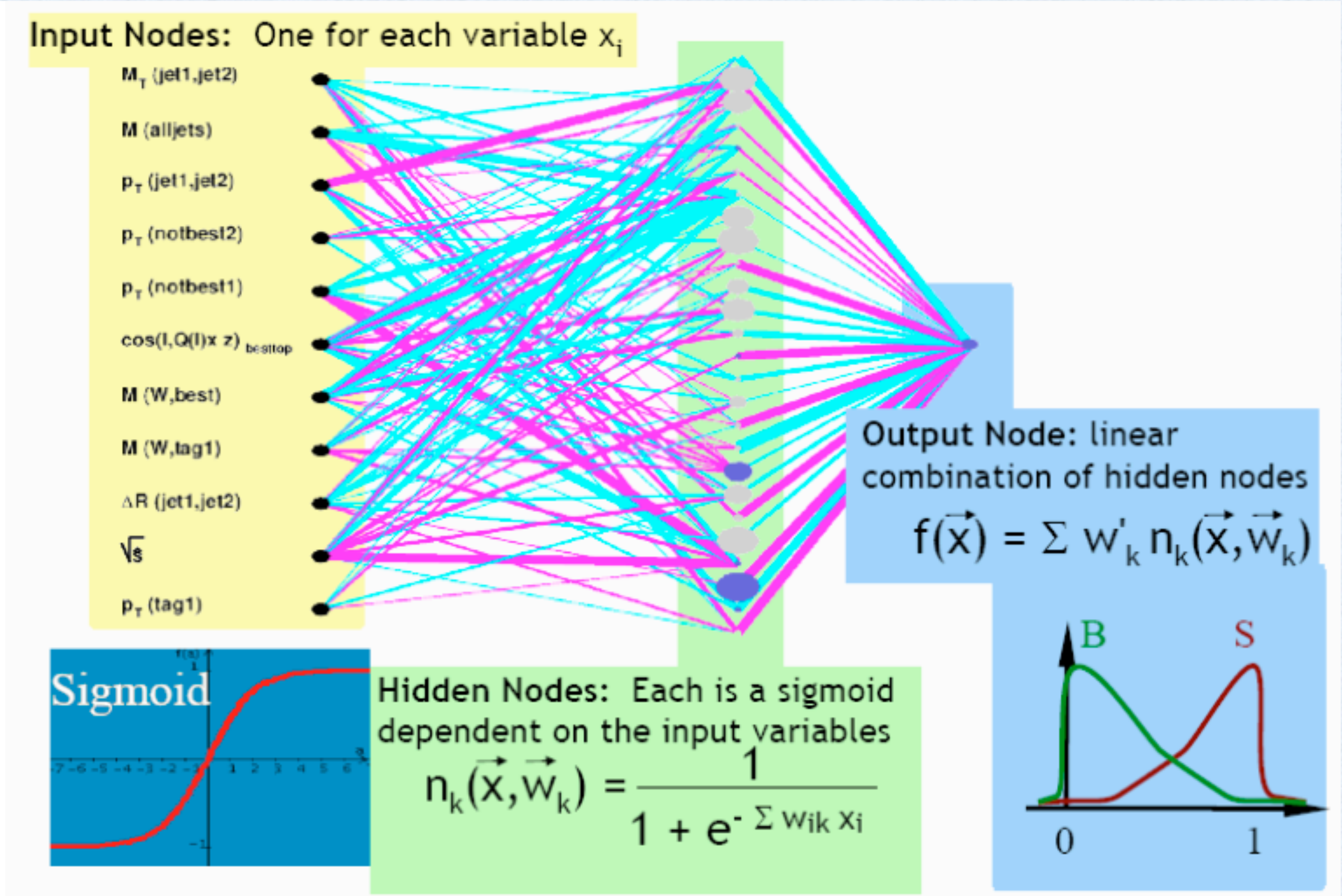


# At the node, transform




$$n_k(\vec{x},\vec{w}_k) = \frac{1}{1 + e^{-\Sigma w_{ik} x_i}}$$

# FROM MANY ...ONE



# Example

- ROOT tutorial file placed on blackboard.
- Download and run it - it will go download a file from the web and create a neural network