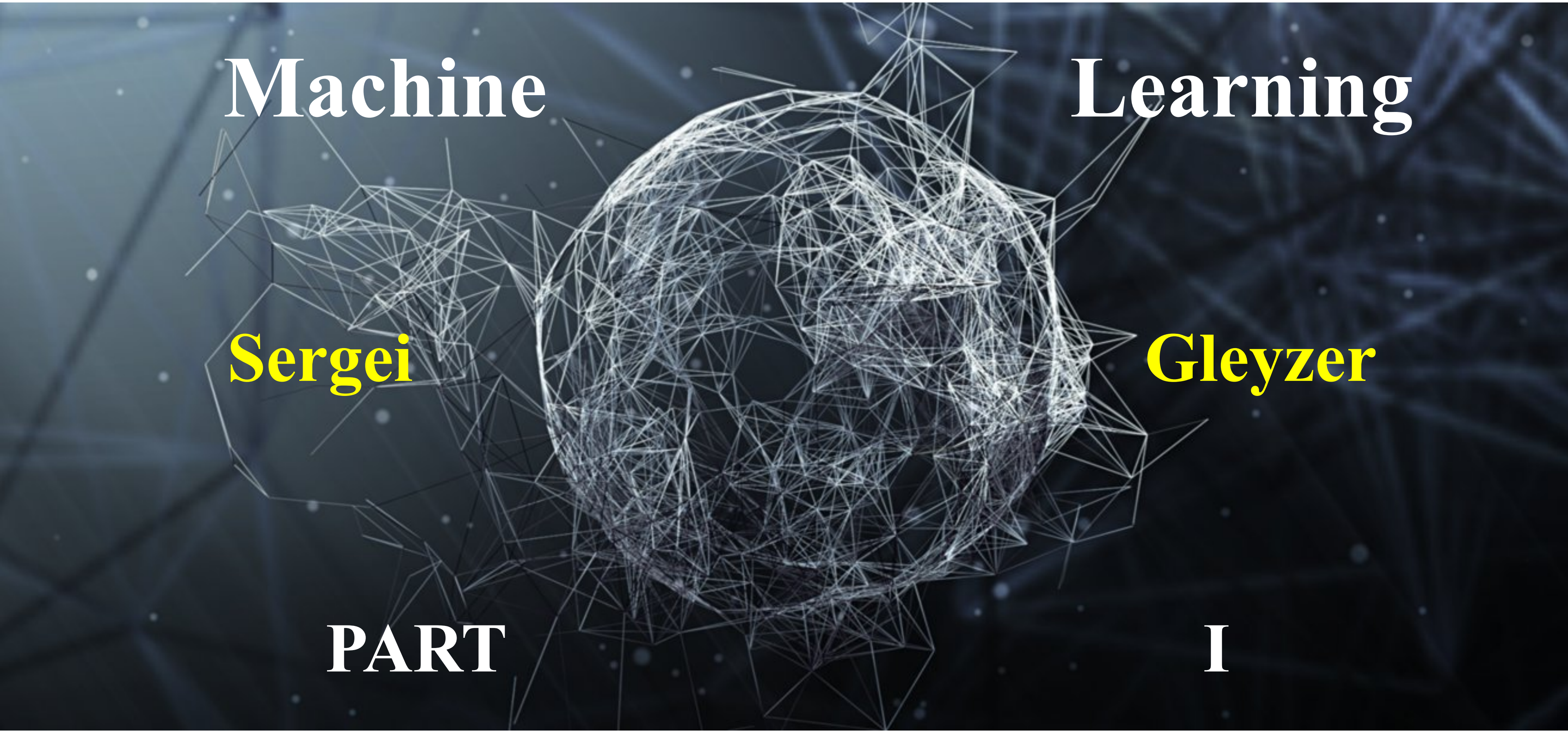


Machine Learning

Sergei Gleyzer

PART I

CERN Open Lab Summer Student Lecture

July 25, 2016

Outline

- **What is Machine Learning**
- **in Theory**
- **in Practice**

Machine Learning Basics

Machine Learning

What is Machine Learning?

- Study of algorithms that improve their performance **P** for a given task **T** with more experience **E**

Sample tasks: identifying faces, Higgs bosons

In Computer Science

Machine learning already preferred approach:

- Speech recognition, Natural language processing
- Computer vision, Robot control
- Medical outcomes analysis

Machine Learning field is growing fast

- Improved algorithms
- Increased data capture
- Software too complex to write by hand

A Little History

1950s First methods invented

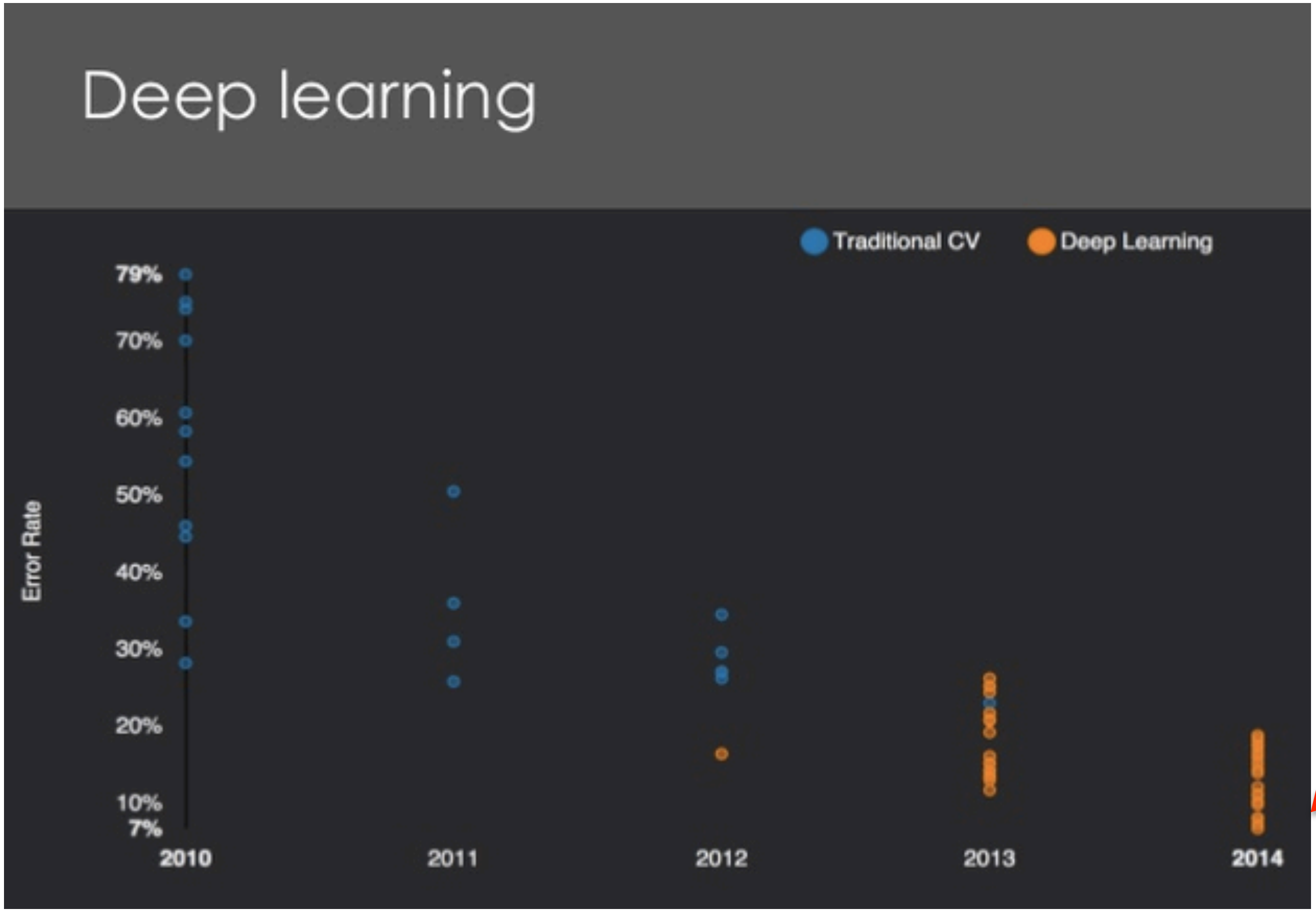
1960-80s Slow growth, focus on knowledge

1990s Computing power growth, new learning methods, data-centric focus

2000-10s Wide use of machine learning in all spheres of research and industry

2010s Improvement of learning, high parallelism, deep learning

Diving Deeper



Huge Progress

Machine Learning Theory

Machine Learning

General Approach:

- Given **training** data $T_D = \{y, \mathbf{x}\} = (y,x)_1 \ldots (y,x)_N$, **function space** $\{f\}$ and a **constraint** on these functions, teach a machine to learn the **mapping** $y = f(x)$

Machine Learning

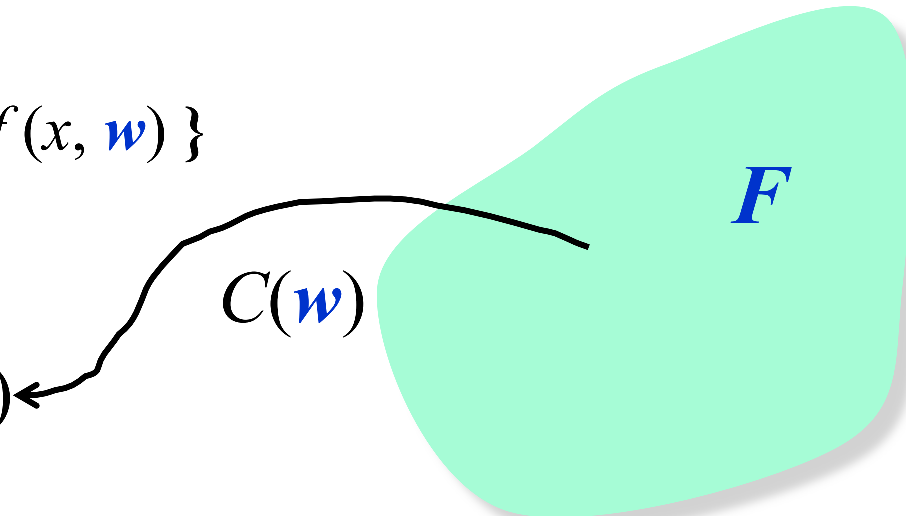
Choose

Function space $\quad F = \{ f(x, \mathbf{w}) \}$

Constraint $\quad C$

Loss function* $\quad L$

$f(x, \mathbf{w}^*)$



Method

Find $f(x)$ by minimizing the empirical risk $R(w)$

$$R[f_w] = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i, w))$$

subject to the constraint $C(\mathbf{w})$

*The loss function measures the cost of choosing badly

Machine Learning

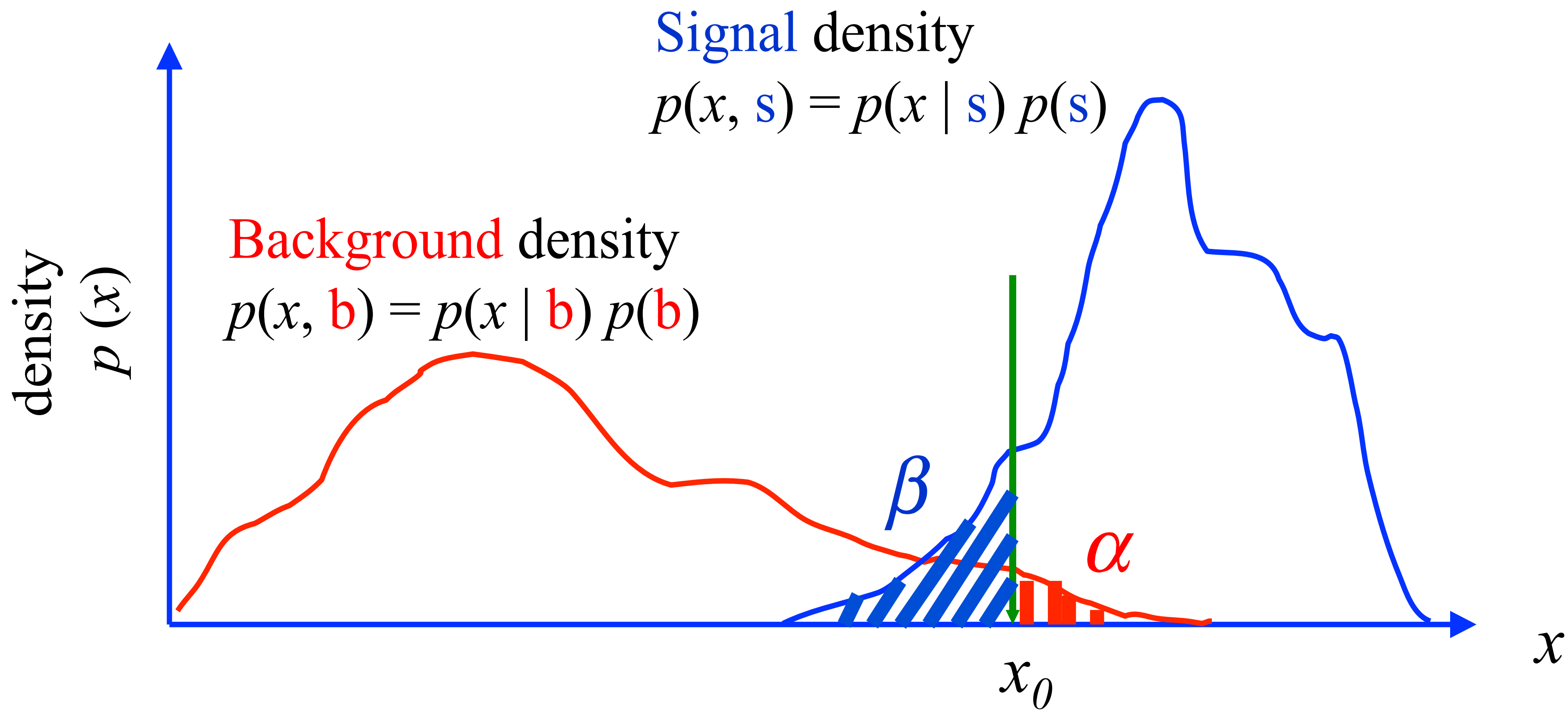
Many methods (e.g., neural networks, boosted decision trees, rule-based systems, random forests,…) use the quadratic loss

$$L(y, f(x, w)) = [y - f(x, w)]^2$$

and choose $f(x, w^*)$ by minimizing the ***constrained*** mean square empirical risk

$$R[f_w] = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i, w)]^2 + C(w)$$

Classification Theory

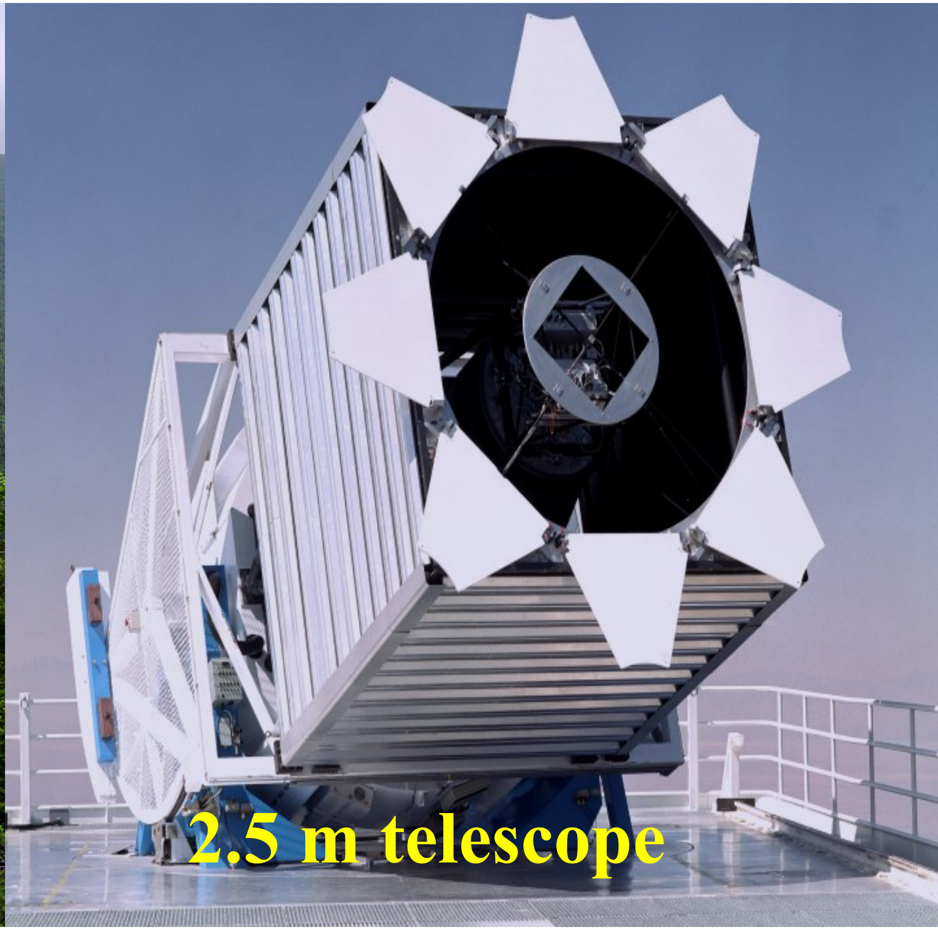


Optimality criterion: minimize the error rate, $\alpha + \beta$

Machine Learning in Practice

How Big is Big Data?

Sloan Digital Sky Survey



Collected more data in the first two weeks

than was collected in the history of astronomy

Large Synoptic Survey Telescope

3200 Megapixel camera

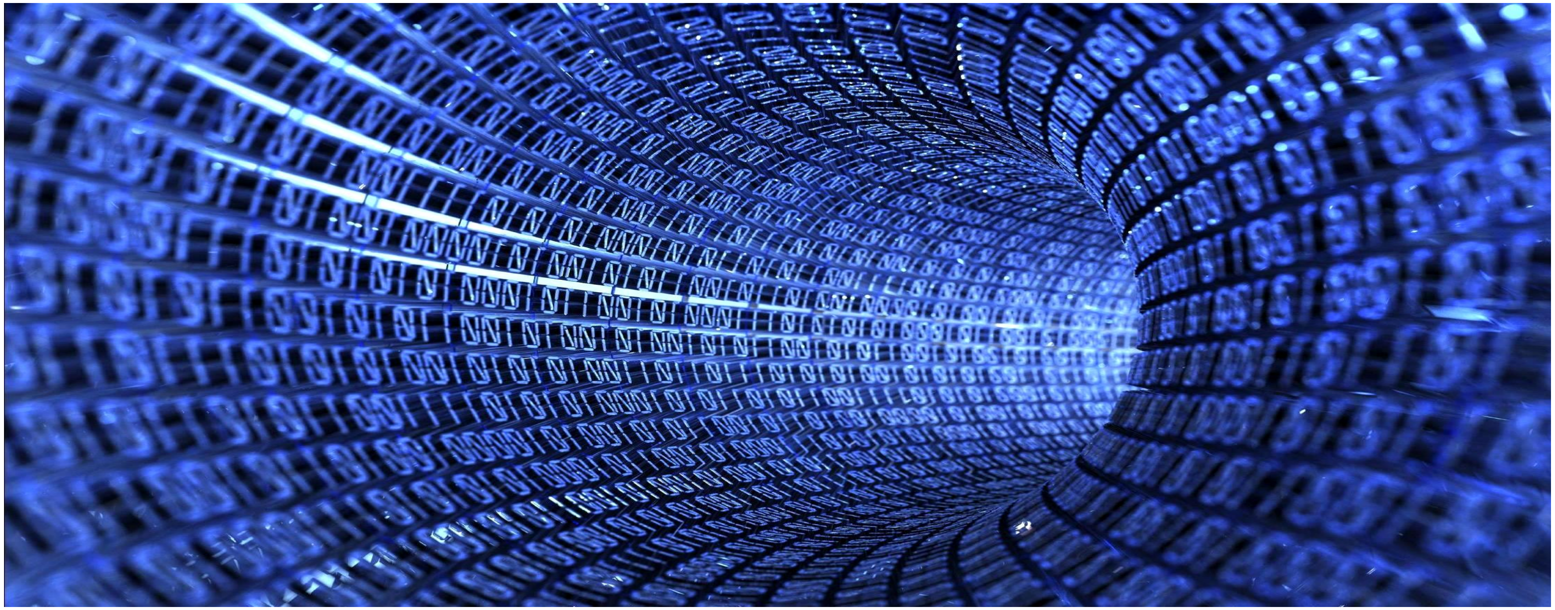
Will create a movie of the sky in different frequencies for ~10 years

Data-taking expected to begin in 2022

Large Synoptic Survey Telescope

Big Data

Project	Expected Data	Period
SDSS	100 Tb	2000 - 2015
LSST	100 000 Tb	2022 - 2032
LHC	15 000 000 Tb	2010 - 2035

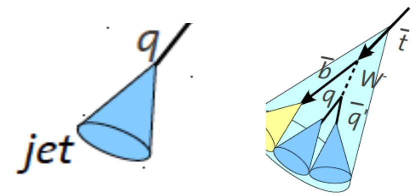


How do LHC Experiments Use Machine Learning?

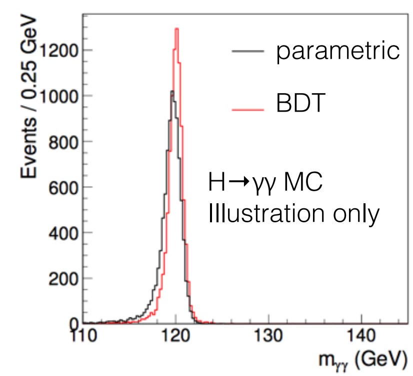
LHC Applications

Classification

- **Particle identification**
  – Is this particle a photon or a jet?



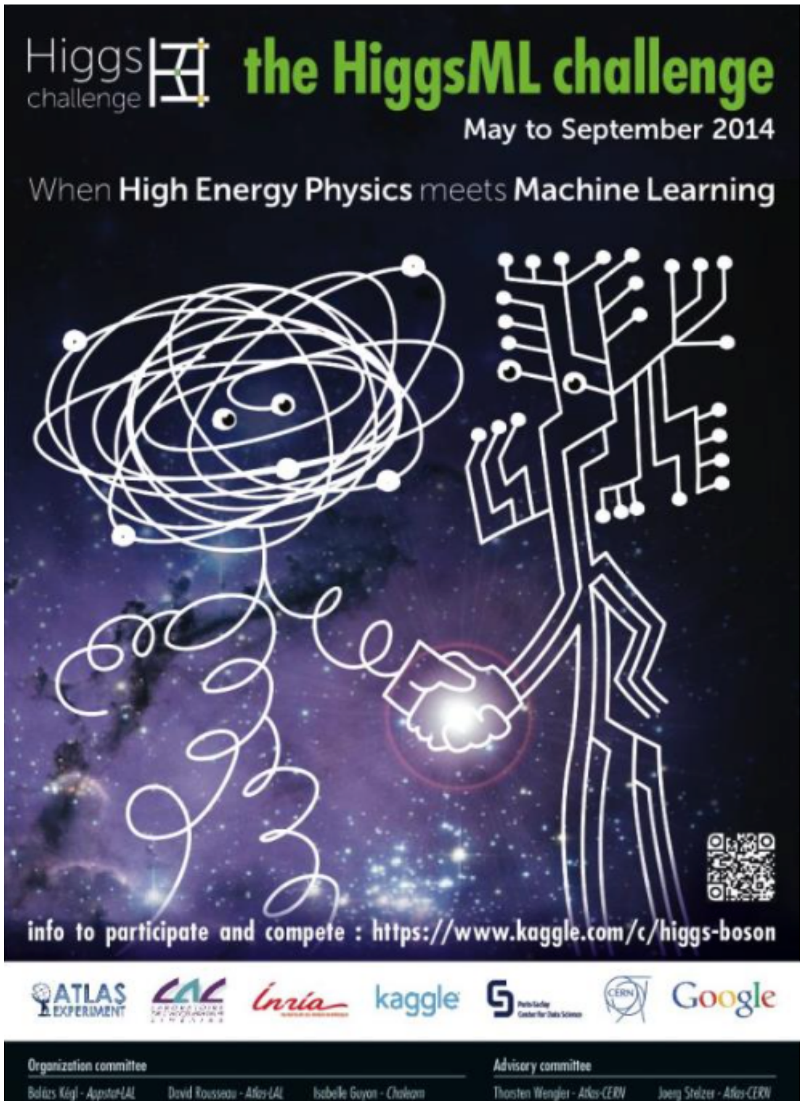
- **Advanced Pattern Recognition**
  – Clustering detector hits, jet sub-structure
- **Searches for new Physics**
  – Is this a Higgs/SUSY event or background



Function Estimation

- **Energy/Momentum estimation**
  – Better estimate using Machine Learning regression

Higgs Challenge



- Big success !
- 1785 teams (1942 people) have participated
- 6517 people have downloaded the data
- Most popular challenge on the Kaggle platform (until spring 2015)
- 35772 solutions uploaded
- 136 forum topics with 1100 posts

- Similar challenge by LHCb

Types of Learning

- **Supervised:**
  - All training data are **labeled**
- **Unsupervised**
  - All training data are **not labeled**
- **Semi-supervised**
  - Some training data is **not labeled**

Classification

Goal:

Achieve **lowest probability** of error on unseen cases $\{<x^{(i)}, y^{(i)}>\}$

Approach:

Inductively **learn** from labeled **examples** where classes are known

Classification

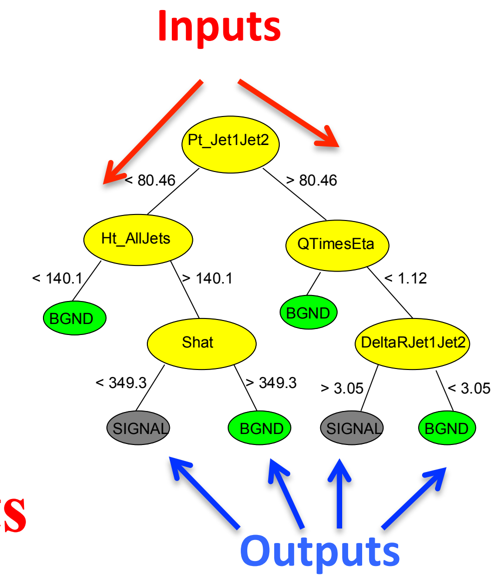
Distinguish f(x), g(x) using Training set of observations

{**inputs** , **outputs**}

Pass **observations** to a learning algorithm

neural network, decision tree

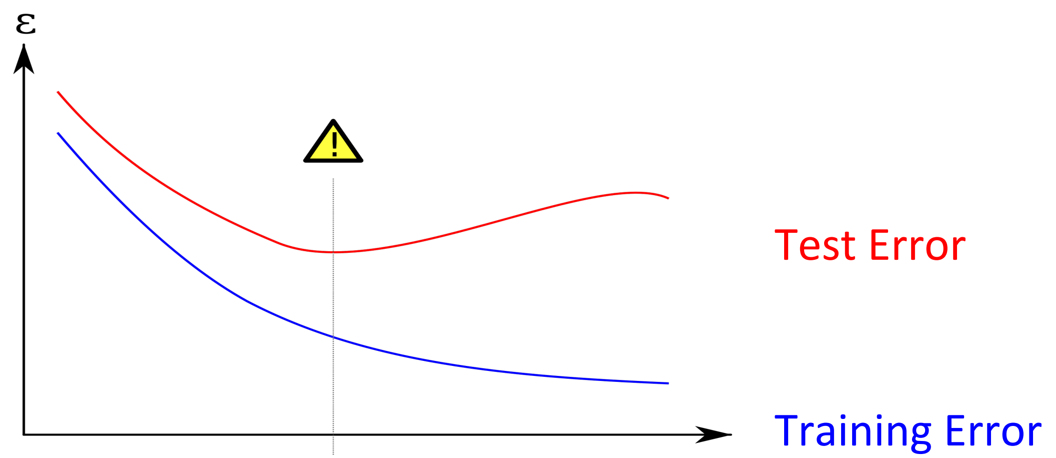
that produces **outputs** in response to **inputs**



Use another set of observations to evaluate

Training

- Split data into at least two sets
  - Keep training and testing sets separated

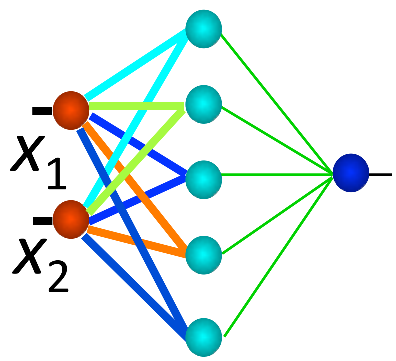


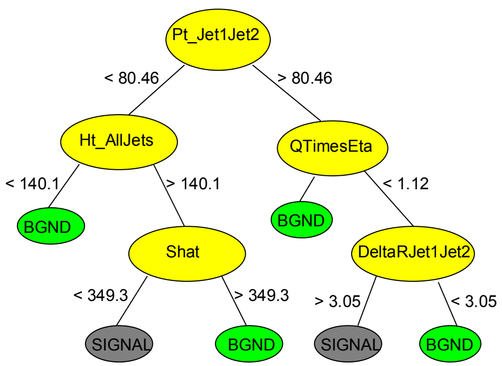
  - Monitor training and testing error rates
    - Watch out for overtraining

Popular Methods

Incomplete list of learning algorithms:

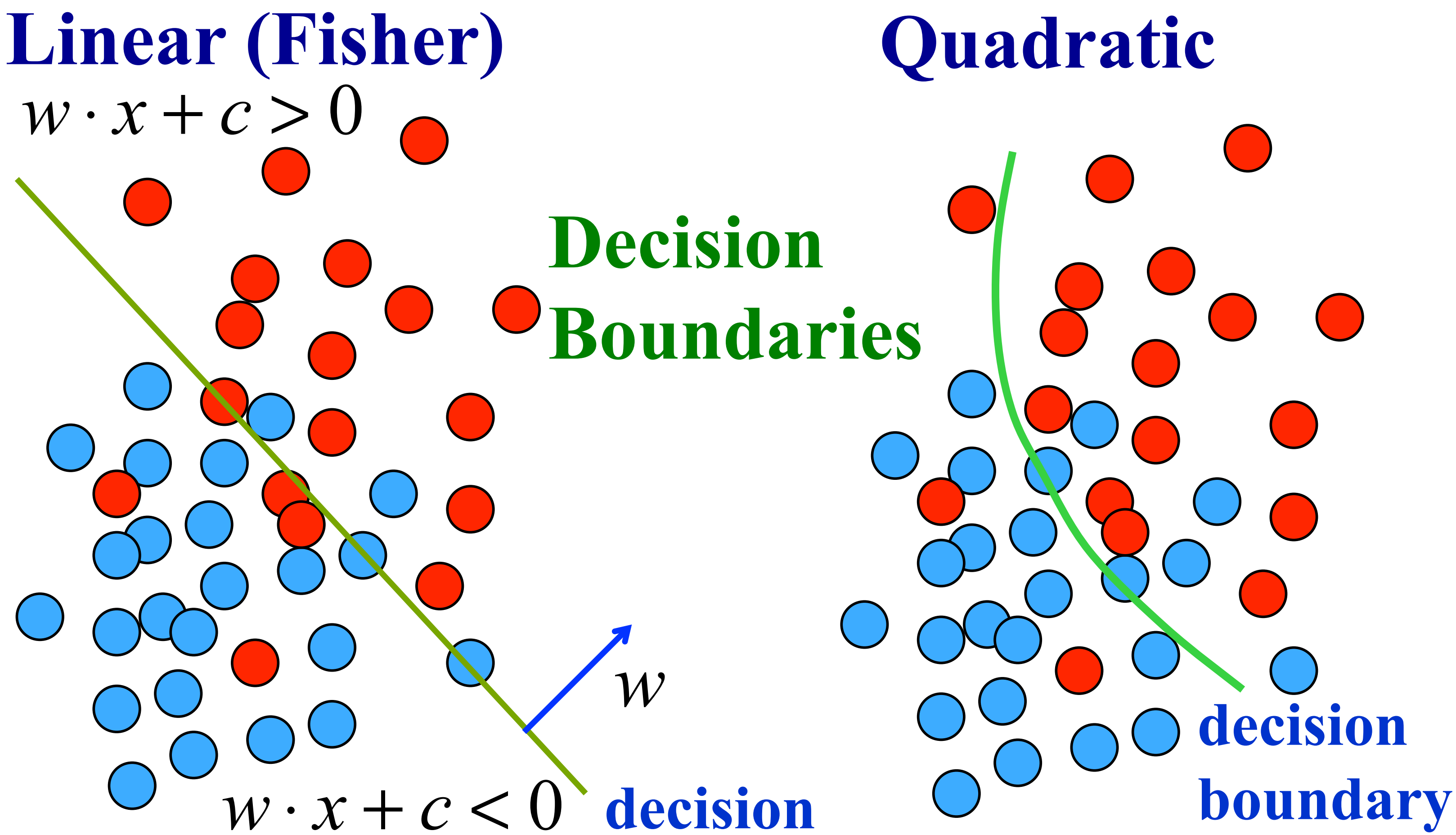
- Fisher (Linear) Discriminant
- Quadratic Discriminant
- Support Vector Machines
- Decision Trees
- Neural Networks
- Bayesian Neural Networks
- Genetic Algorithms
- Random Forest





CONSTRUCTING CLASSIFIERS

Linear and Quadratic



Decision Trees

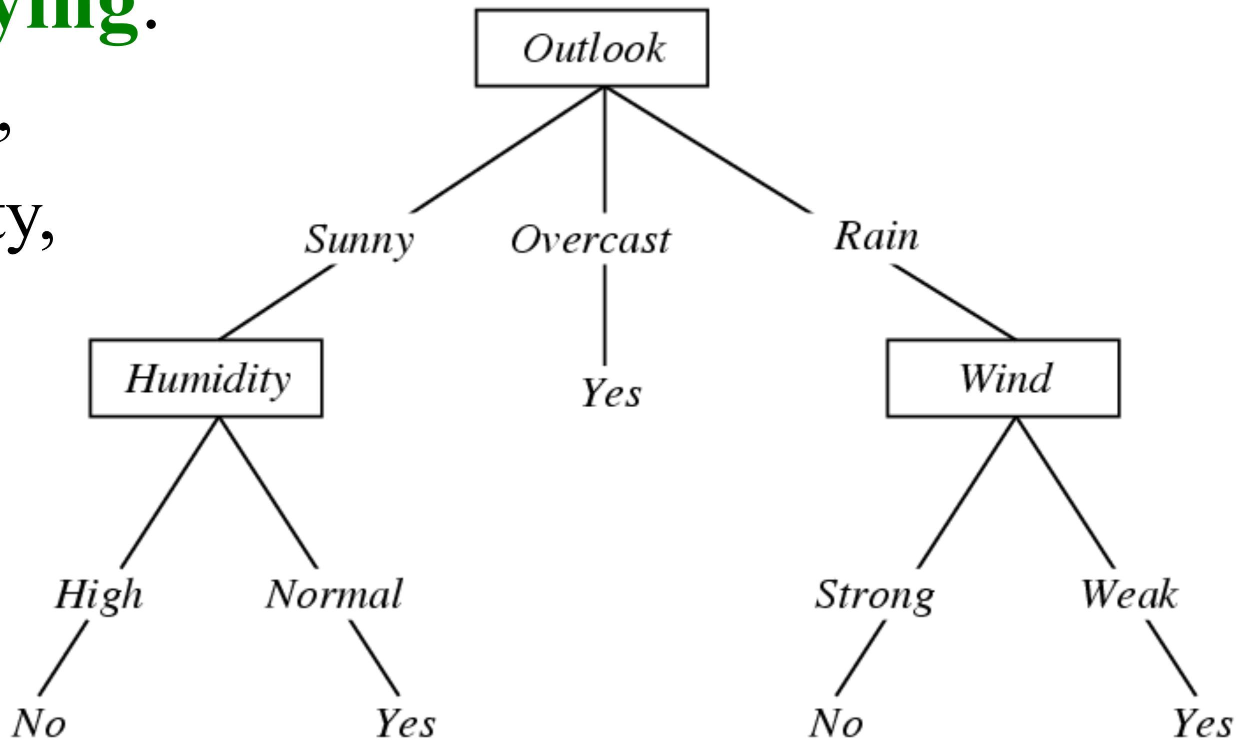
Decision Trees

- Decision trees is a simple Classifier
- Golf-Playing:

  f(outlook,

  humidity,

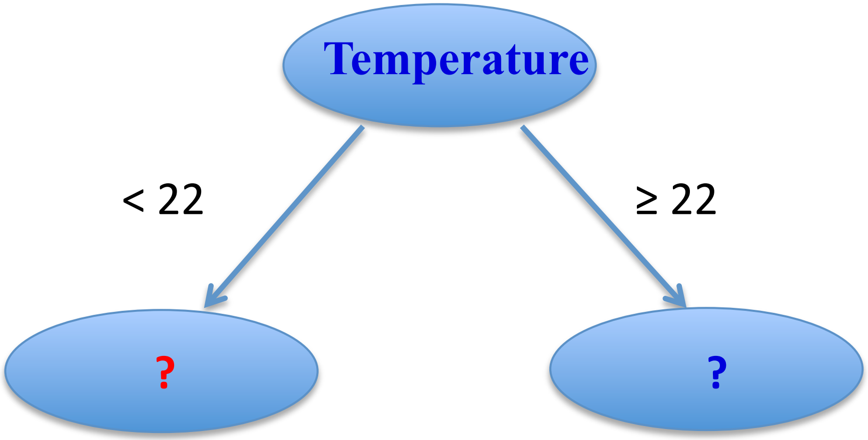
  wind)



Decision Trees

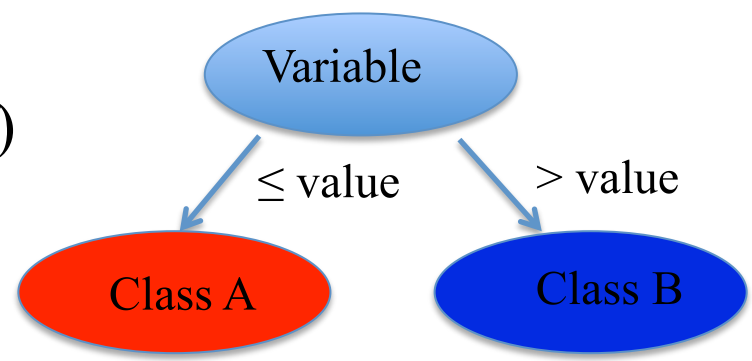
Building a tree:

- Scan along each variable
  - propose a **DECISION**
    - A splitting value that maximizes class separation (binary branching)



Decision Trees

- Choose a **decision** that leads to greatest separation between classes **A** and **B**
  - – Build regions of increasing purity $\dfrac{N_A}{N_A + N_B}$
  - – Stop when no further improvement from additional splitting
    - Reach terminal node (leaf)
    - Assign purity-based class



Proceed to Hands-On I (c5.0)

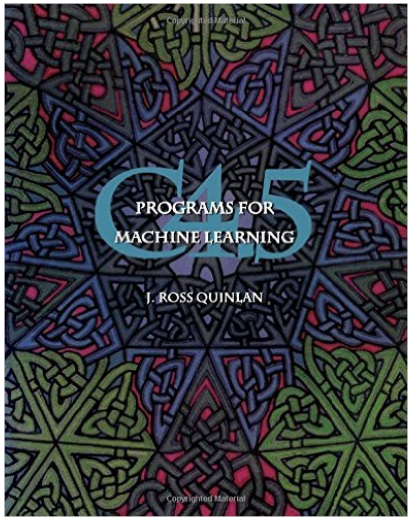
Part I: Decision Trees

Hands-On Part I

1. Login to CERNBox: http://cernbox.cern.ch
2. Open Swan: http://swan001.cern.ch
3. Open new terminal: new → terminal
4. Clone the code: git clone https://github.com/iml-wg/c50.git
5. Go to c50 directory: cd c50/

C5.0

- Classic **ML tool** for
  - **decision trees**
  - **rules**
  - **boosted classifiers**
- Written by **J.R. Quinlan**
  - Name: ID3 → C4.5 → C5.0
    - Use c5.0 to familiarize with decision tree based classifiers



Hands-On Part I

Examples: playing golf, breast-cancer

- Create your first classifiers
  - **Decision trees**
    - c5.0 –f golf
    - c5.0 –f breast-cancer

    Needed:

    .names file that includes the names of classes and variables, and variable types(continuous/discrete)

    .data file with values for each variable and class

Tutorial Part I

- Look at **Decision Tree** structure(s)
- Consider **accuracy** of predictions
  - Prediction errors
    - on training examples
    - on testing examples
  - Understand **confusion** matrix

```
  (a)   (b)    <-classified as
 ----  ----
  125     5    (a): class 2
    6    63    (b): class 4
```

Rules

Decision Rules:

- Deconstruct **Decision Tree**
- Set of **if** – **then** – **else** rules
  - – Example of "weak" learners (better than random guessing)
  - – Become a competitive classifier in an ensemble
    - RuleFit: Friedman, Popescu, 2005

Proceed to Tutorial (c5.0)

Part II: Rules

Tutorial Part II

Examples: playing golf, breast-cancer

- Create your first classifiers
  - **Rules**
    - c5.0 –r –f golf
    - c5.0 –r –f breast-cancer
  - Compare Rule(s) to Decision Tree(s)
    - Note: all decision trees are rules but not all rules are trees

Classifier Performance

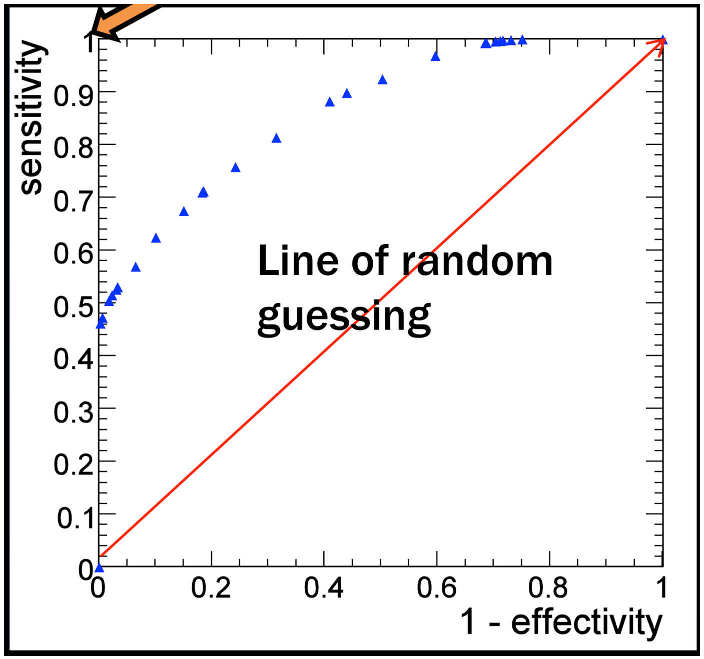
Classifier Performance

Receiver Operating Characteristic (ROC)

Commonly used metric

Shows the **relationship** between correctly classified positive cases (sensitivity) and incorrectly classified negative cases (1-effectivity)

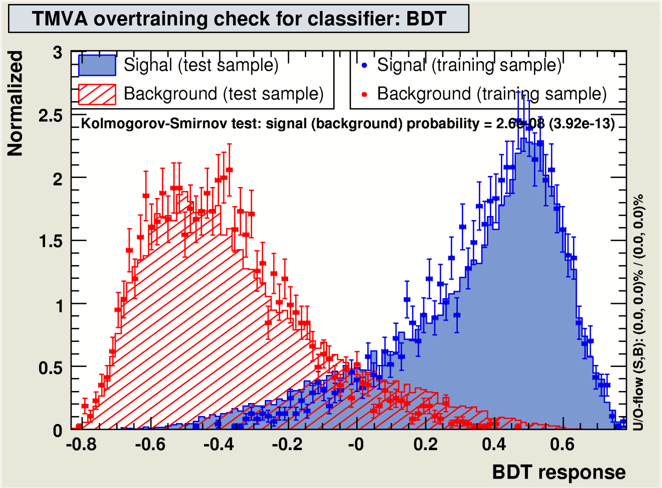
Perfect Classifier



Over-Training

Over-training or over-fitting sometimes occurs when too many parameters for data size

- **Diagnose with**
  - Divergent training -testing error slopes
  - Kolmogorov-Smirnov tests of classifier output



- **Treat with**
  - Reduce number of parameters
  - Prune decision trees

Pruning

Decision trees can become large and complex and risk over-fitting the data

Pruning removes less powerful or possibly noisy parts of the tree

- start from the leaves and work back up
- Pruned trees smaller in size, easier to interpret

ML Today

- **Large ensembles** of classifiers
- **Deep vs. shallow learning**
  - Neural networks with many more hidden layers
- **Combination** of semi/un-supervised learning with supervised learning

Summary

- **Machine Learning** is a very powerful field with an expanding number of applications in high energy physics
  - **Basic Methods**: Linear, Quadratic, Decision Trees, Decision Rules
  - **More methods** on Wednesday
  - Many methods available: good to experiment

Classifier Performance

Receiver Operating Characteristic (ROC)