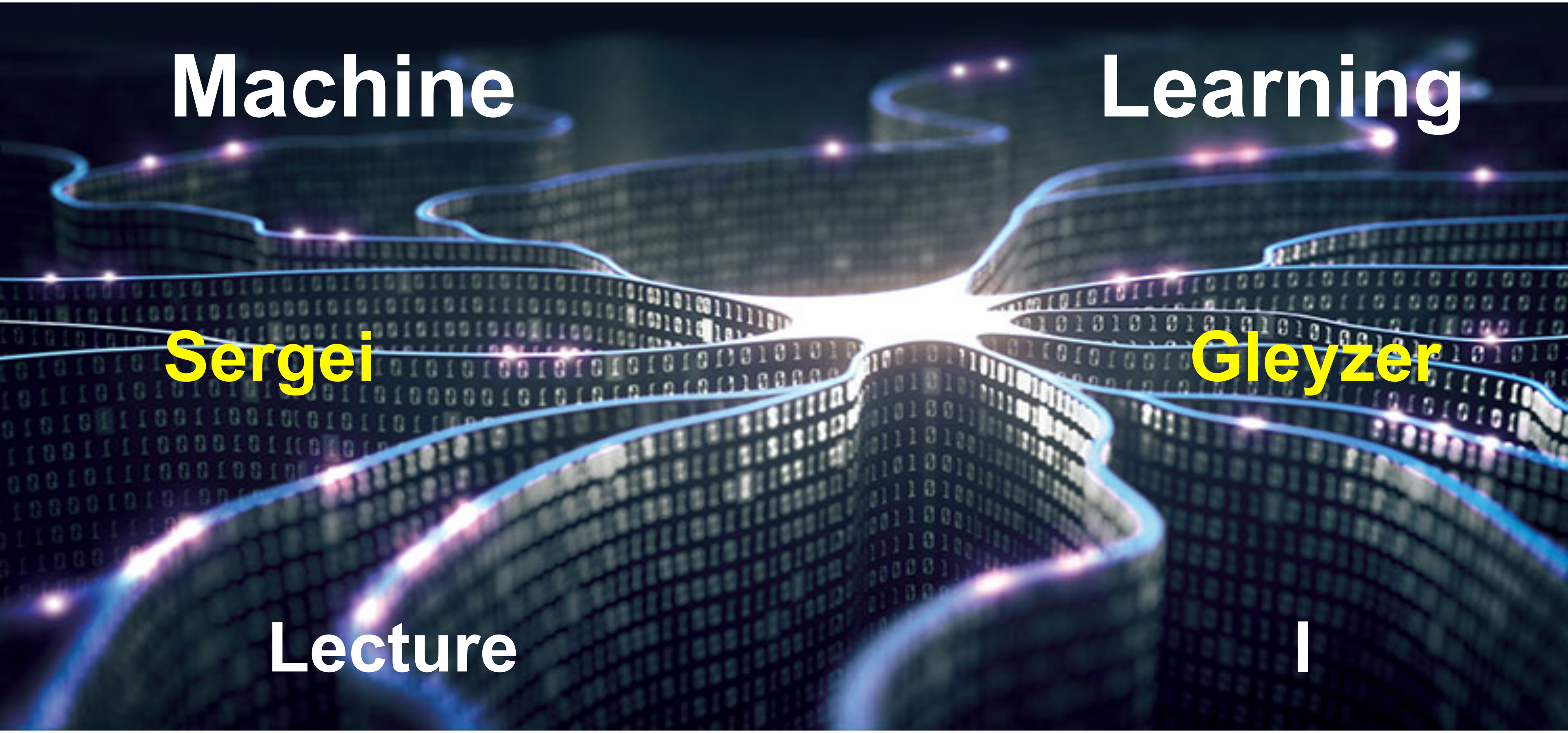
UF UNIVERSITY of FLORIDA



Machine Learning

Sergei Gleyzer

Lecture I

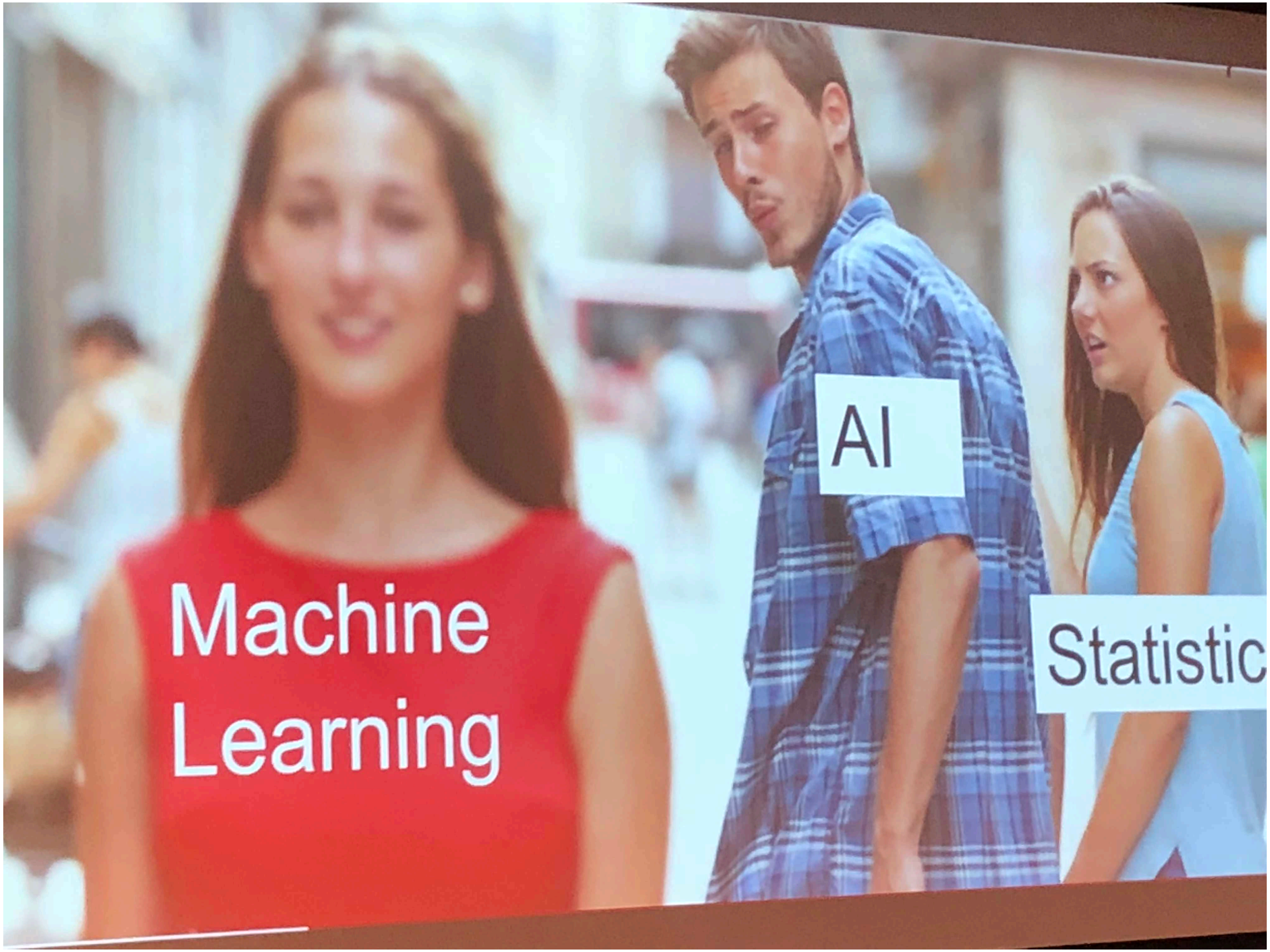
UPRM Lectures

April 24, 2019

Today's Outline

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

Machine Learning Basics


Machine
Learning
AI
Statistic

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

Already the preferred approach to:

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

Growing fast

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

Machine Learning

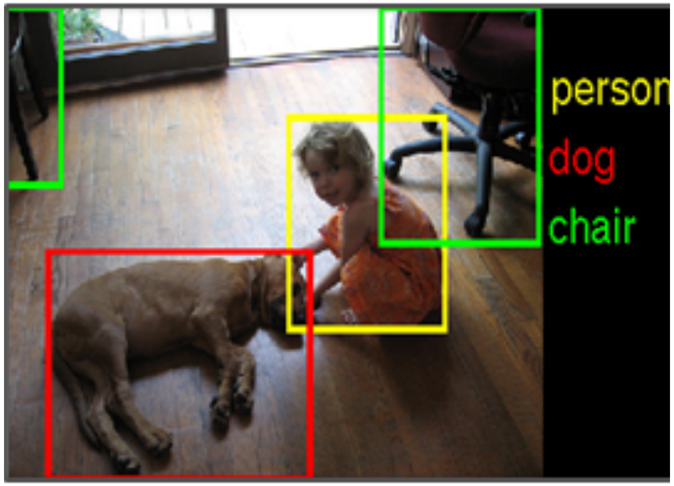
General Approach:

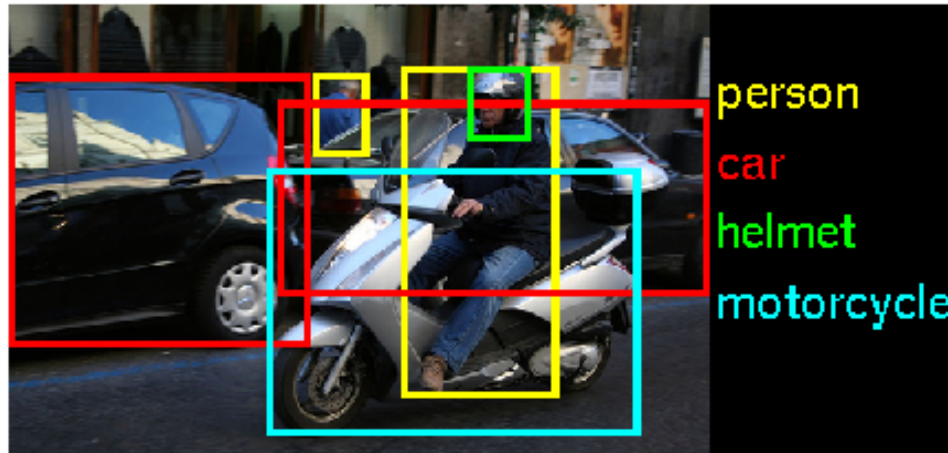
Given **training** data $T_D = \{y, \mathbf{x}\} = (y,x)_1 \ldots (y,x)_N$,

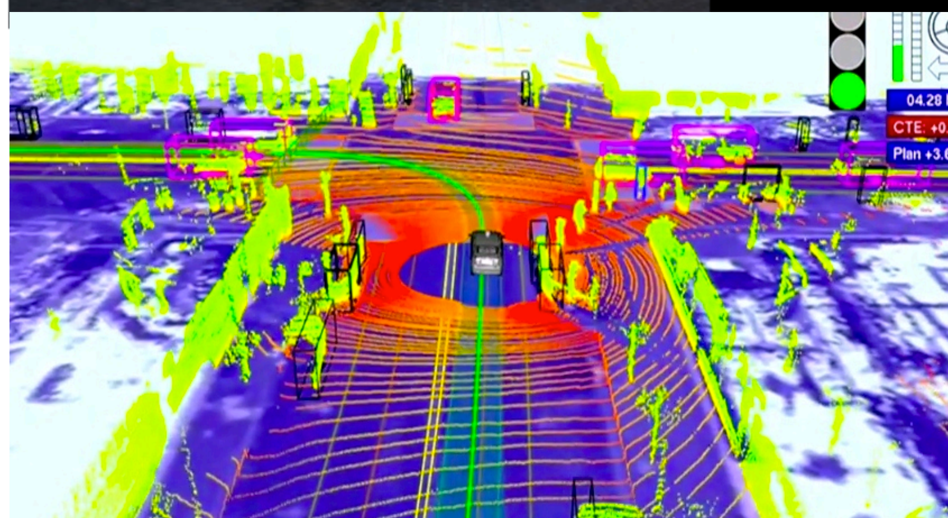
function space $\{f\}$ and a
constraint on these functions

Teach a machine to learn the **mapping** $y = f(x)$

Examples







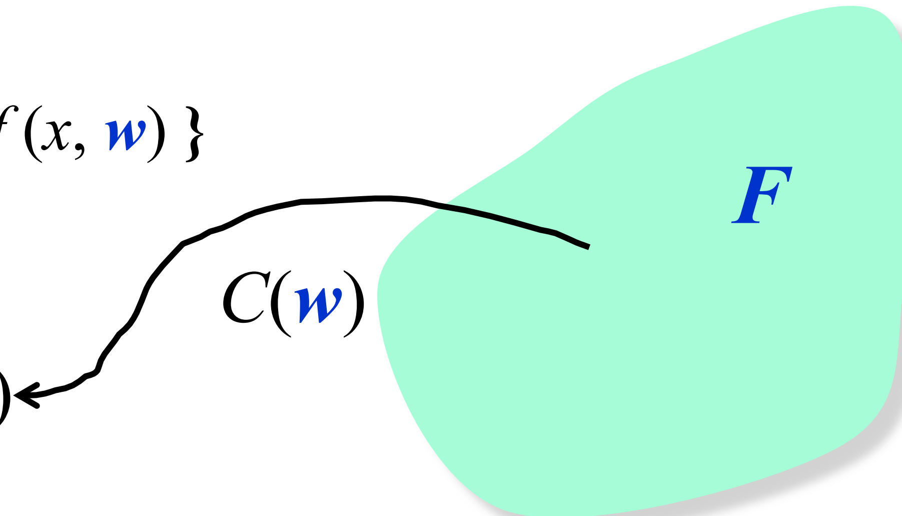
Machine Learning

Choose

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

Constraint $\boldsymbol{C}$

Loss function* $\boldsymbol{L}$



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

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 $C(\boldsymbol{w})$

*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

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

History

1950s: First methods invented

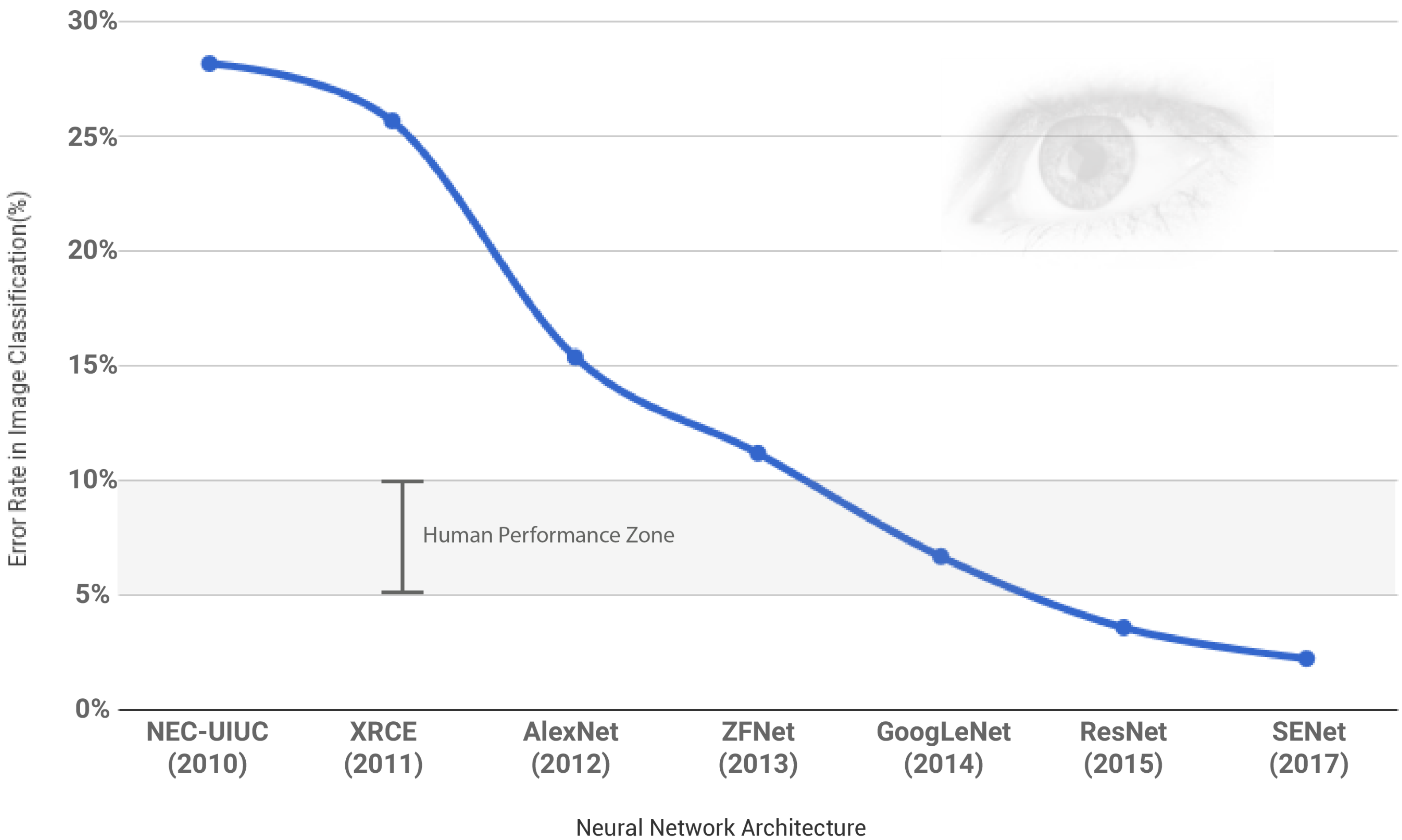
1960-80s: Focus on knowledge

1990s: Computing power, new learning methods, data-centrism

2000-10s: Wider use research and industry

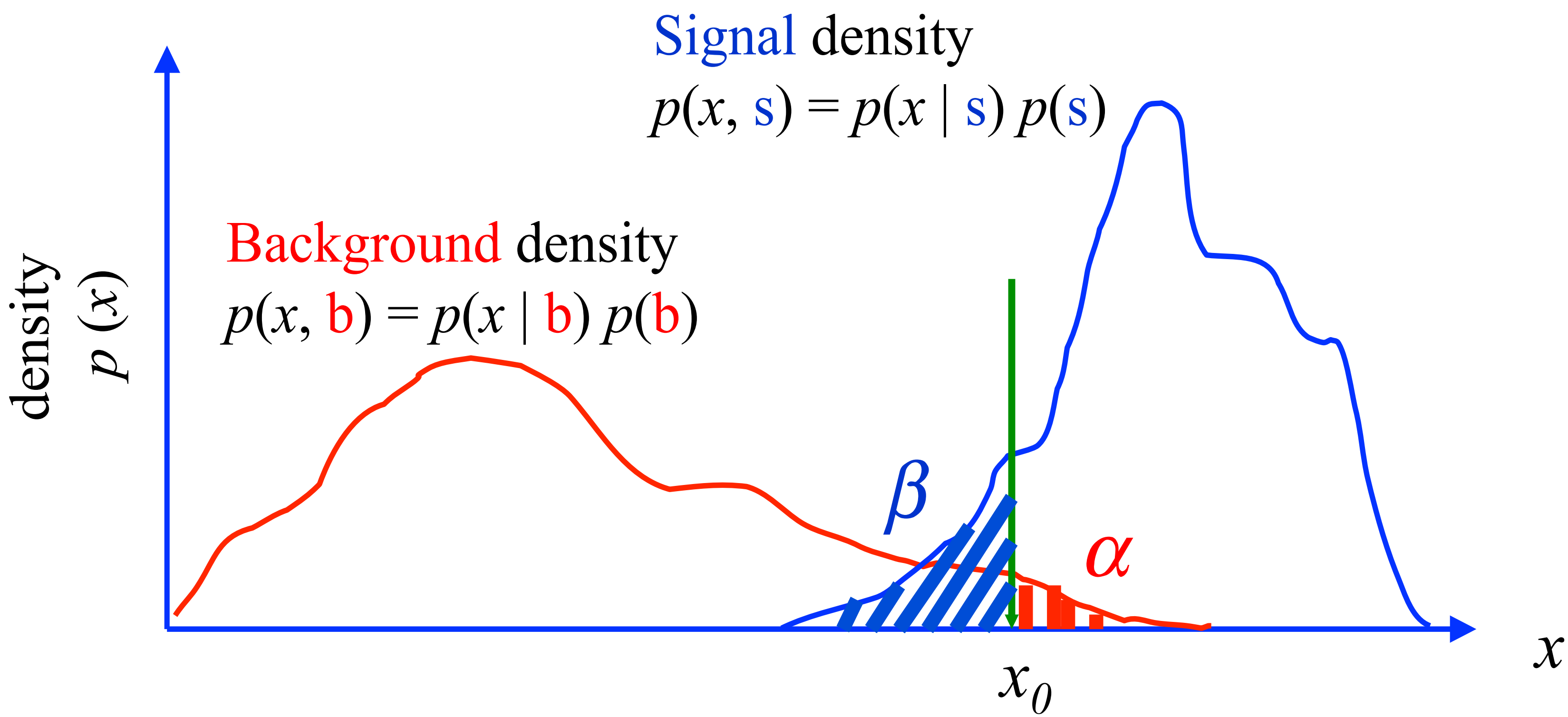
2010s: Learning improvement, dedicated hardware, deep learning

Diving Deeper



Classification Theory

Classification Theory



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

Classification Theory

The total loss L arising from classification errors is given by

$$L = L_b \int H(f)\, p(x,b)\, dx + L_s \int [1 - H(f)]\, p(x,s)\, dx$$

Cost of background misclassification

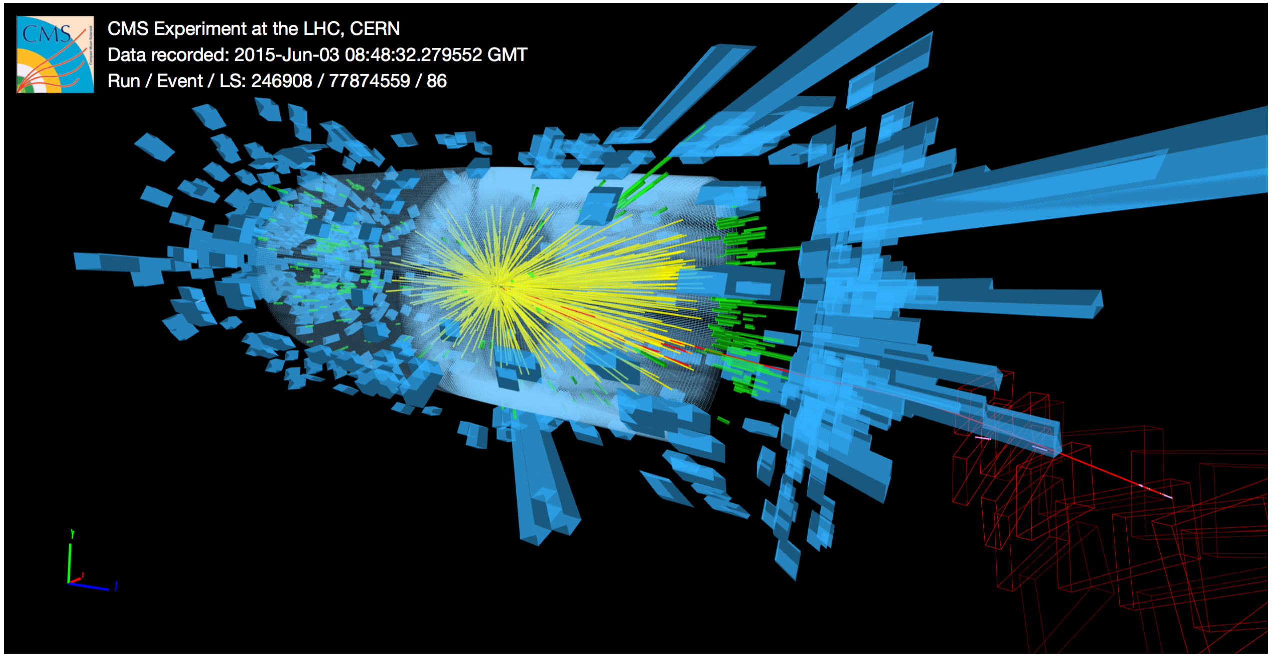
Cost of signal misclassification

where $f(x) = 0$ defines a decision boundary
such that $f(x) > 0$ defines the acceptance region

$H(f)$ is the Heaviside step function:

$H(f) = 1$ if f > 0, 0 otherwise

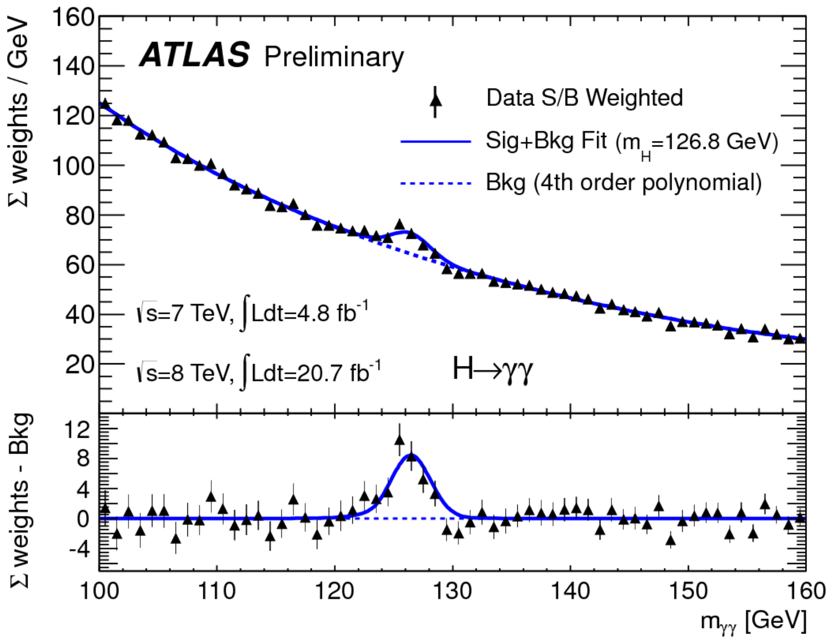
Classification in Practice

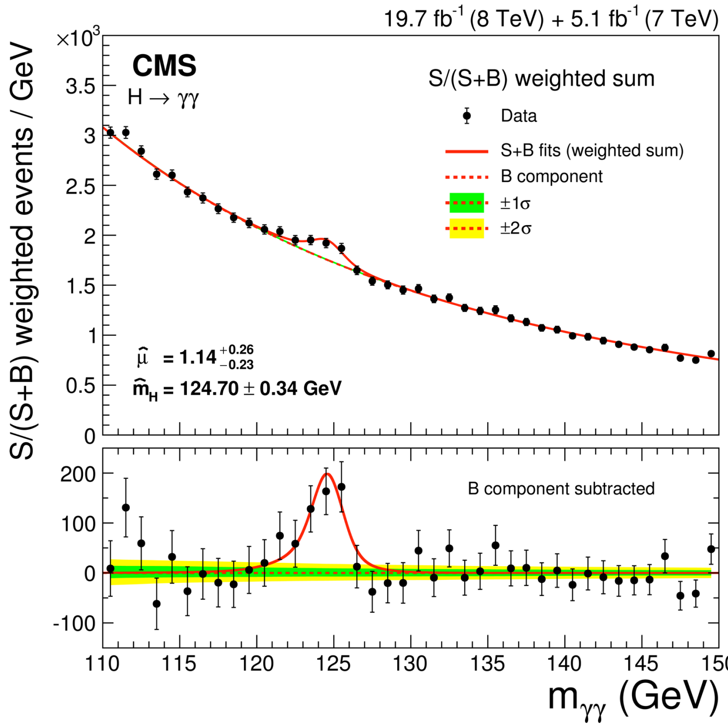


In Particle Physics

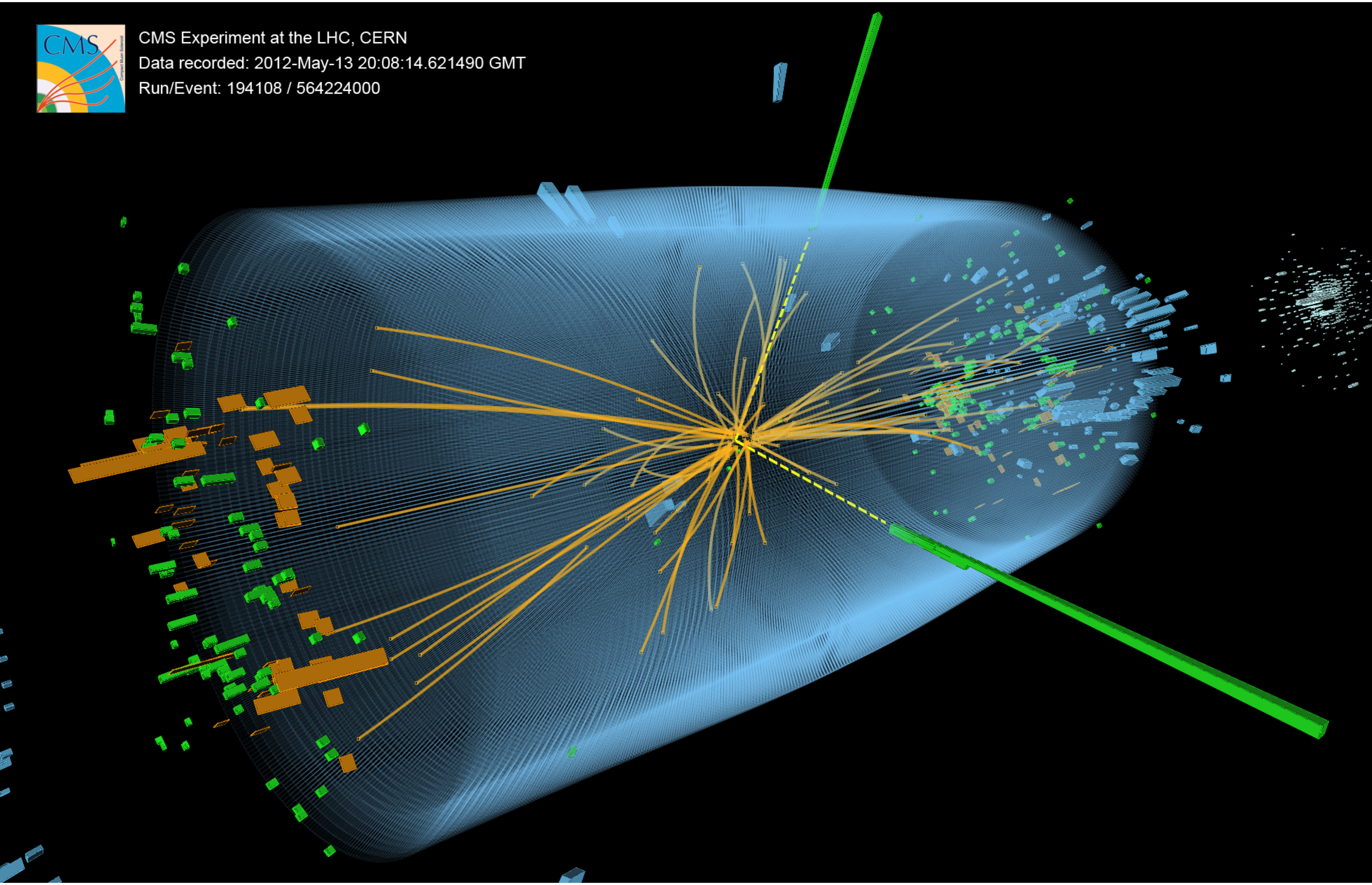
Higgs Boson Discovery

Higgs to di-photons

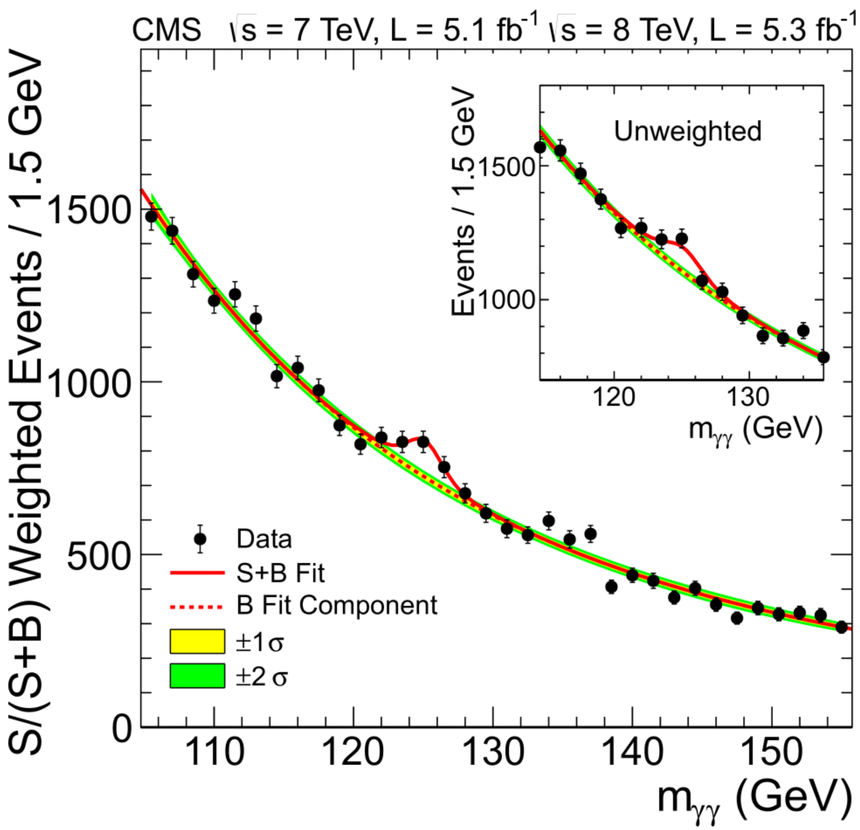




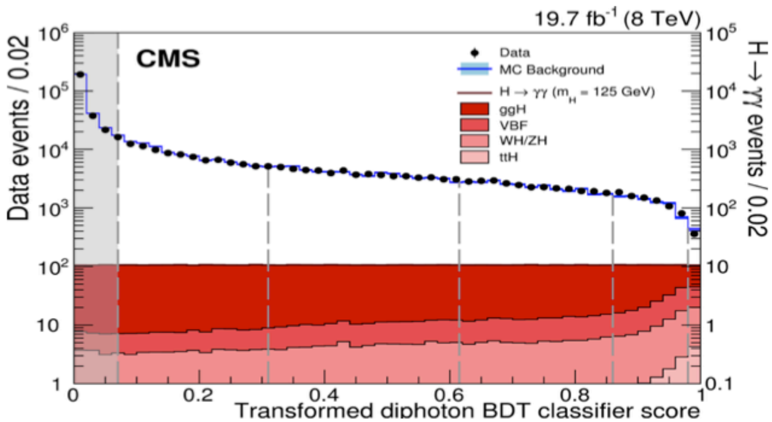
ATLAS **CMS**


CMS
CMS Experiment at the LHC, CERN
Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000

in Higgs Discovery



- Identification of particles
- Identification of interactions
- Energy regression
- Event selection

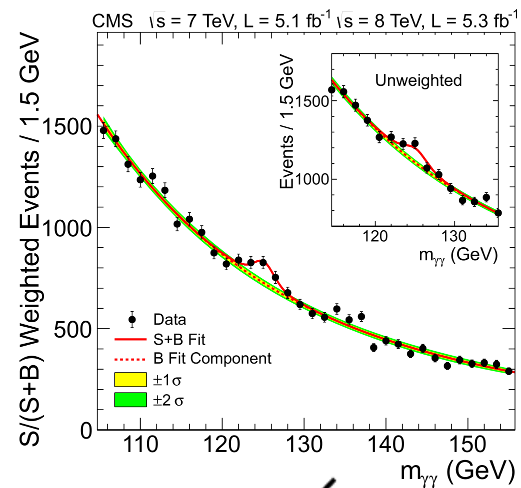


Improvement in analysis from all four areas

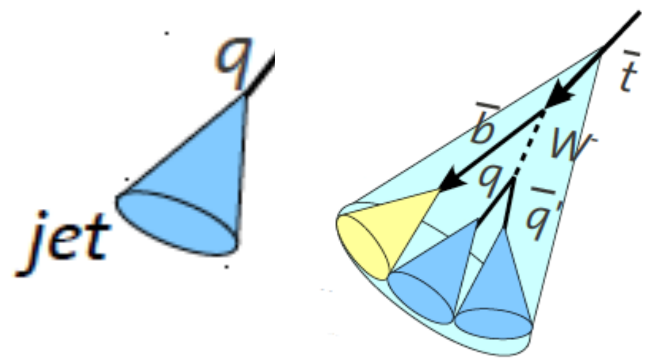
ML in HEP Today

Machine learning already at forefront of what we do:

- Physics object **identification**
- Event type **classification**
- Object properties **regression**



Expanding quickly

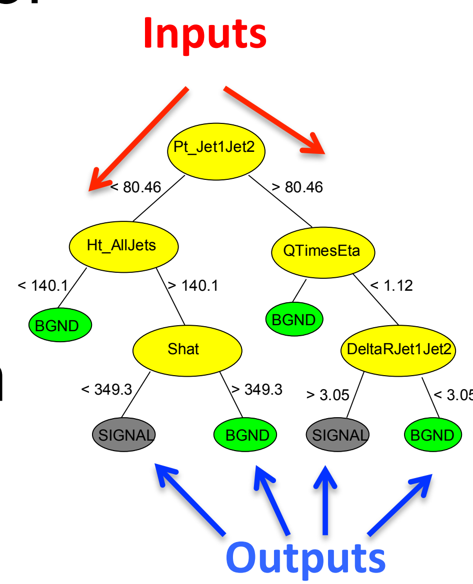


CONSTRUCTING CLASSIFIERS

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

Learning Types

Supervised Learning

Labeled examples with known classes

Examples: cats/dogs

Unsupervised Learning

Un-labeled data

Examples: clustering, anomaly detection

Classification

Primary Goal:

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

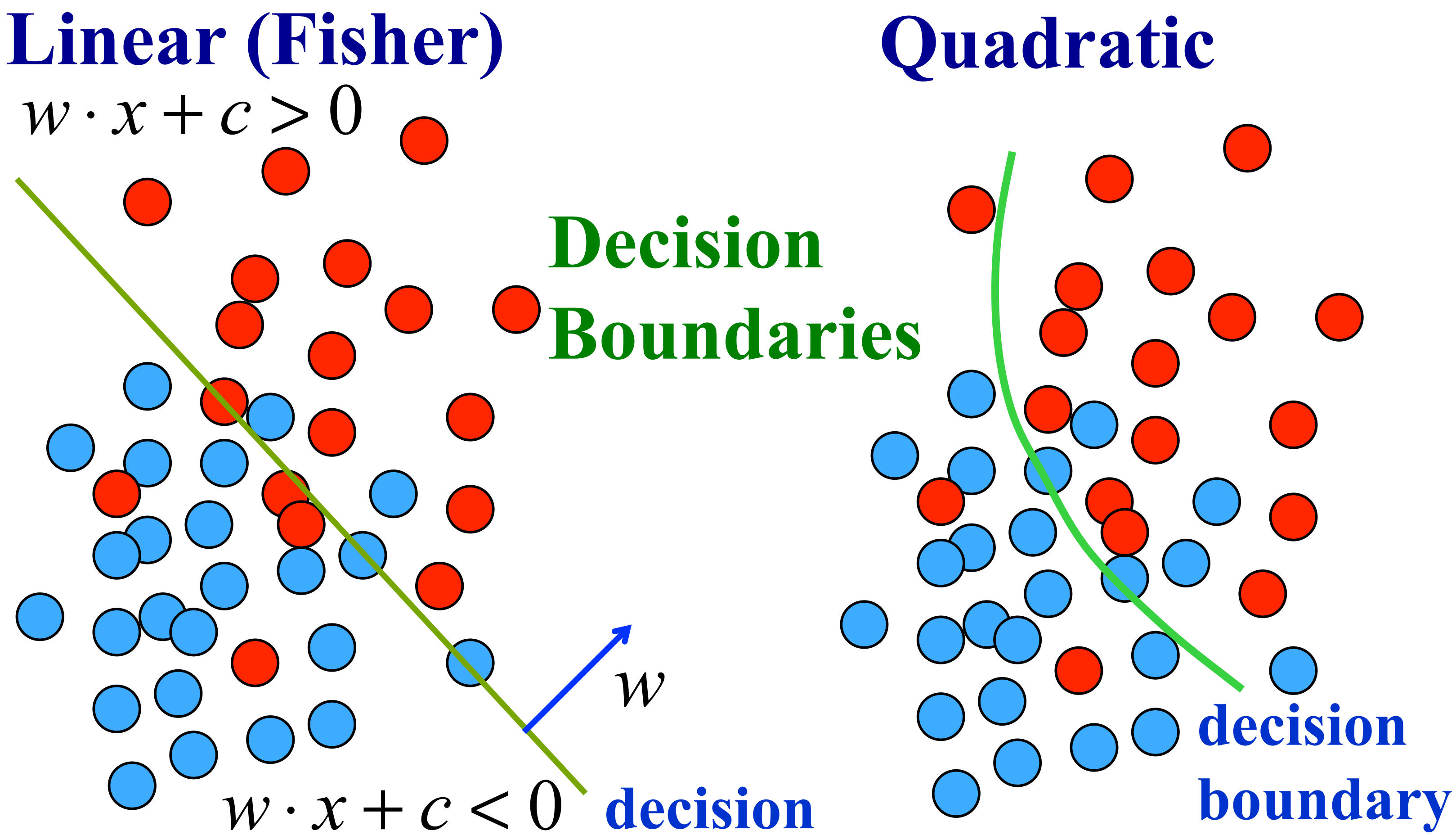
Approach:

Inductively learn from labeled examples where classes are known

ML Algorithms

- Fisher, Quadratic
- Naïve Bayes (Likelihood)
- Kernel Density Estimation
- Random Grid Search
- Rule ensembles
- Boosted decision trees
- Random forests
- Support vector machines
- Genetic algorithms
- Deep learning neural networks

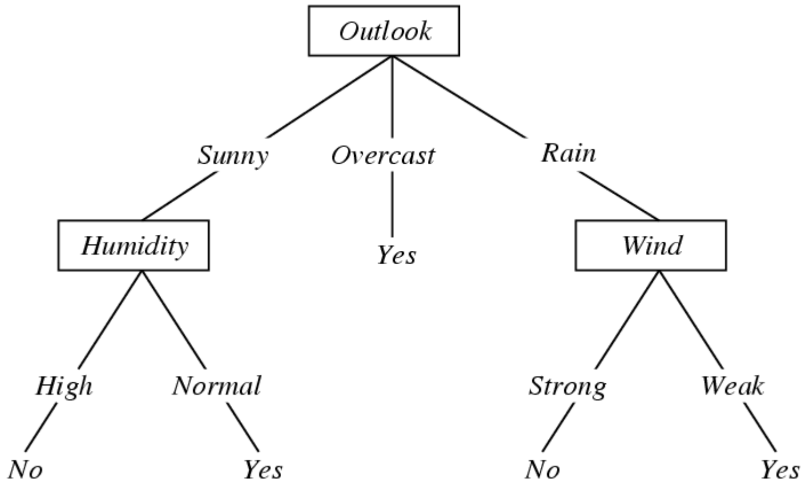
Linear and Quadratic



Binary Decision Trees

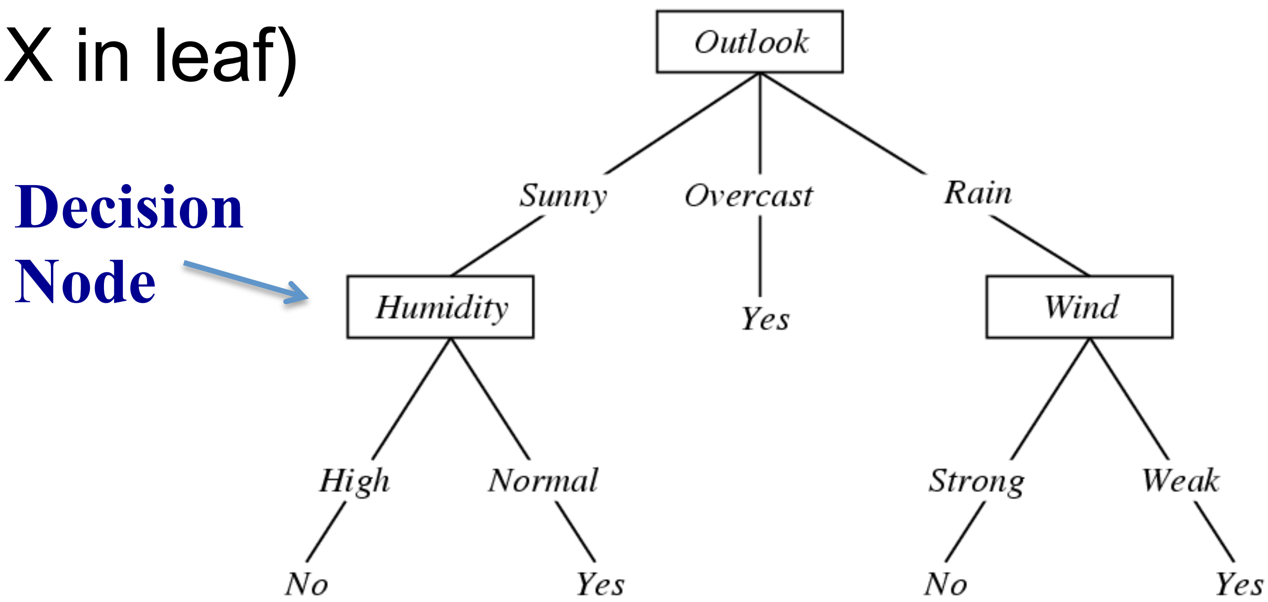
Decision Trees

- **Decision trees** are recursively constructed **multidimensional histograms**
  - Each leaf associated to the value (**class**) of f(x) to be approximated
  - Golf-Playing Tree: f(outlook, humidity, wind, temperature)



Decision Trees

- Each **internal** node: test one attribute X_i
- Each **branch:** selects one value for X_i
- Each **leaf** node: predict Y
  - Or P(Y|X in leaf)



Decision Tree Learning

- Set of possible instances: **X**
  - **instance** is a feature vector

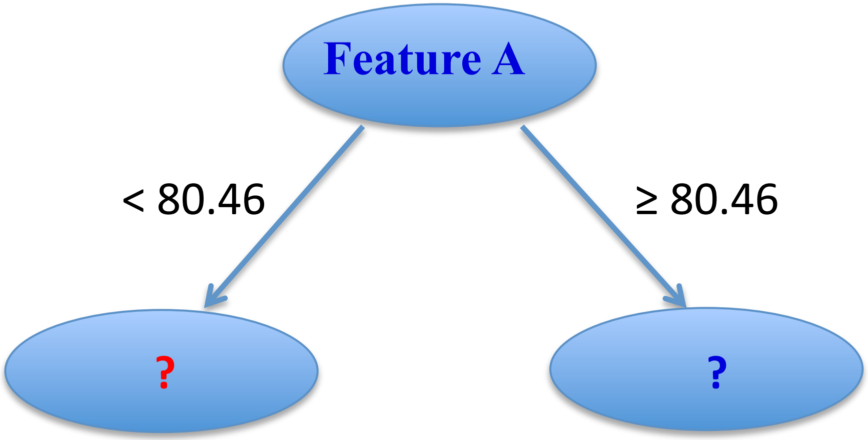
    e.g. < Humidity = High, Wind = weak,
    Outlook = rain, Temp = hot >

- Unknown target function f: **X** → **Y**
  - **Y** is discrete valued (class)

Decision Trees

Building a tree:

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

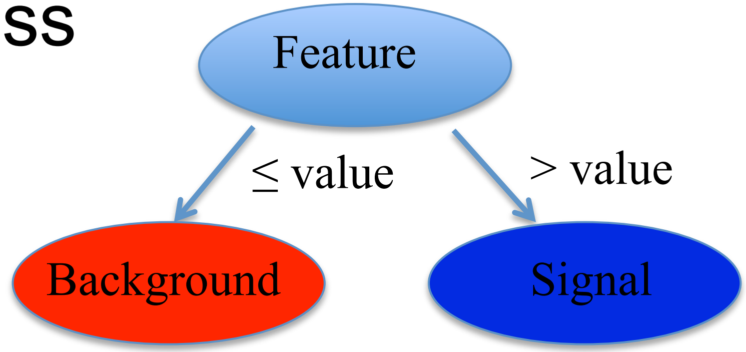


Decision Trees

Choose **split** that leads to greatest separation among classes

- Based on the information gained
  - Build regions of increasing purity
  - Stop when no improvement from branching
  - Reach terminal node (leaf) and assign purity-based class

$$\frac{N_{signal}}{N_{signal} + N_{background}}$$



Separation Gain

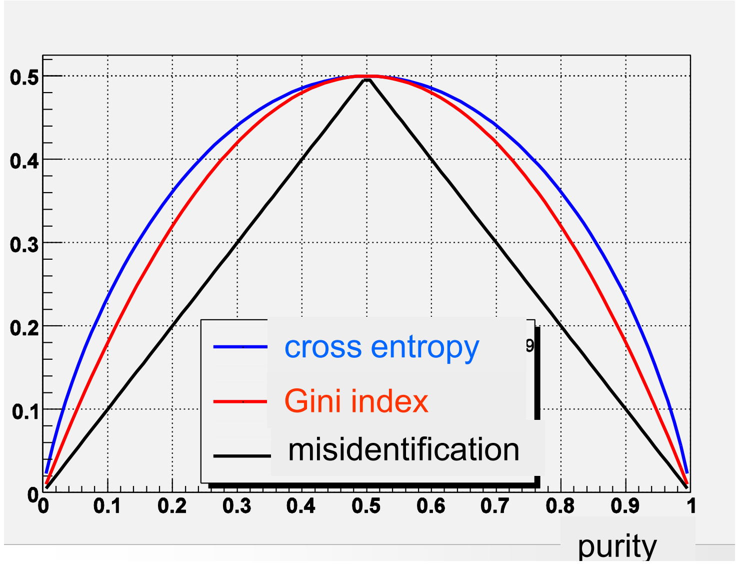
Popular Separation Gain Measures

- Cross-Entropy

  $-p \ln p + (1-p) \ln(1-p)$

- Gini Index

  $p\,(1-p)$



Want to lower entropy from split

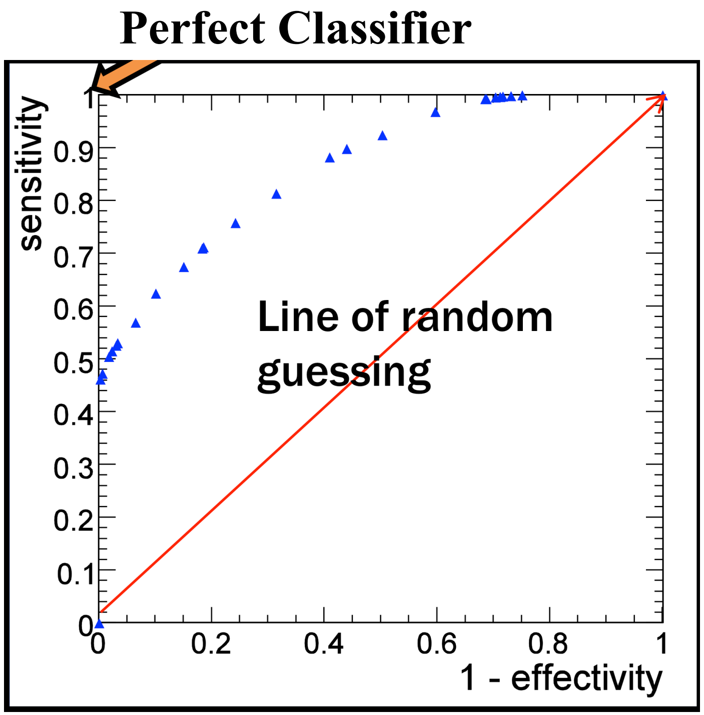
Pruning

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

- **Pruning:** remove parts of tree that are less powerful or noisy
  - start from the leaves and work back up
  - pruned trees smaller in size, easier to interpret

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)

Summary

- **Machine learning focuses on algorithms capable of learning**
  - Basic methods: linear, quadratic
  - Tree-based methods:
    - Decision Trees, Rules
  - Ensembles:
    - Boosted Decision Trees, Random Forests
  - Next lecture: Deep learning algorithms