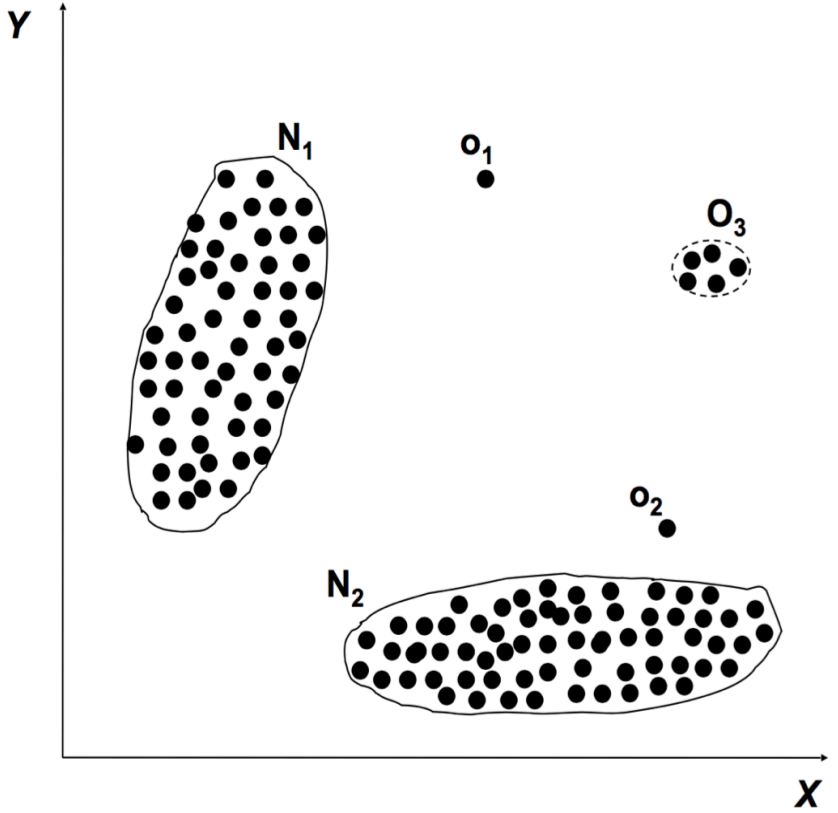
Anomaly: setting the stage

HSF RAMP, LAL-Orsay, 3rd May 2016

Anomaly : point level

- Also called outlier detection
- Two approaches:
  - Give the full data, ask the algorithm to cluster and find the lone entries : o1, o2, O3

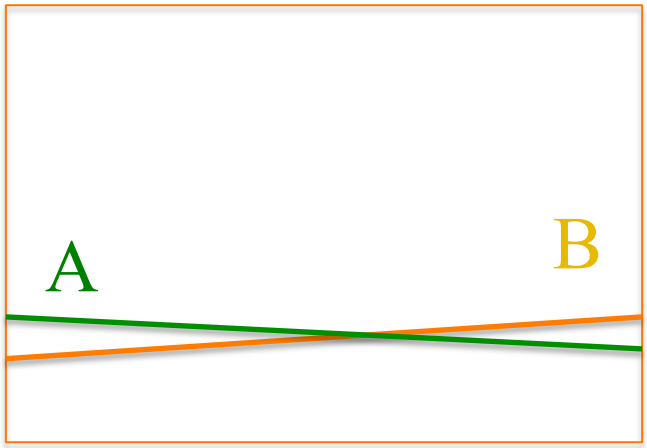


  - We have a training "normal" data set with N1 and N2. Algorithm should then spot o1,o2, O3 as "abnormal" i.e. "unlike N1 and N2" (no a priori model for outliers)
- Application : detector malfunction, sites malfunction, or even new physics discovery

Anomaly : population level

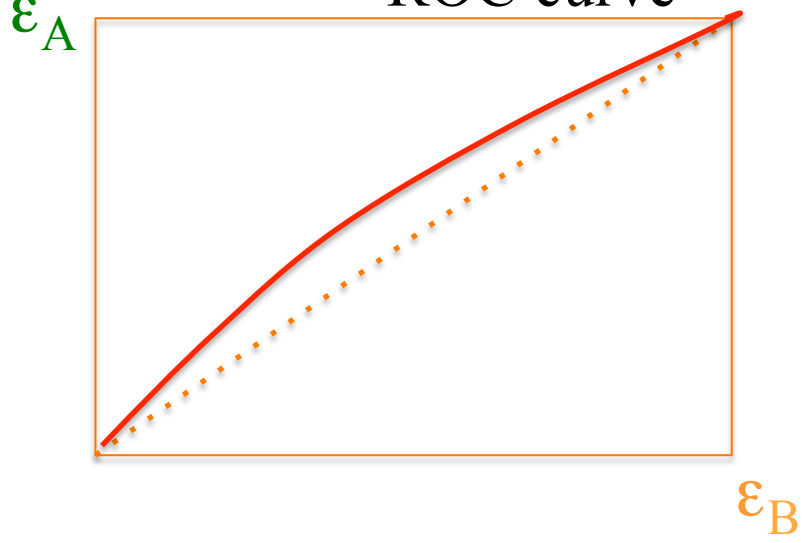
- Also called collective anomalies
- Suppose you have two independent samples A and B, *supposedly* statistically identical. E.g. A and B could be:
  - MC prod 1, MC prod 2
  - MC generator 1, MC generator 2
  - Derivation V12, Derivation V13
  - G4 Release 20.X.Y, release 20.X.Z
  - Production at CERN, production at BNL
  - Data of yesterday, Data of today
- How to verify that A and B are indeed identical ?
- Standard approach : overlay histograms of many carefully chosen variables, check for differences (e.g. KS test)
- ML approach : ~~ask an artificial scientist~~, train your favorite classifier to distinguish A from B, histogram the score, check the difference (e.g. AUC or KS test)
  - ➔only one distribution to check

Small non-local difference



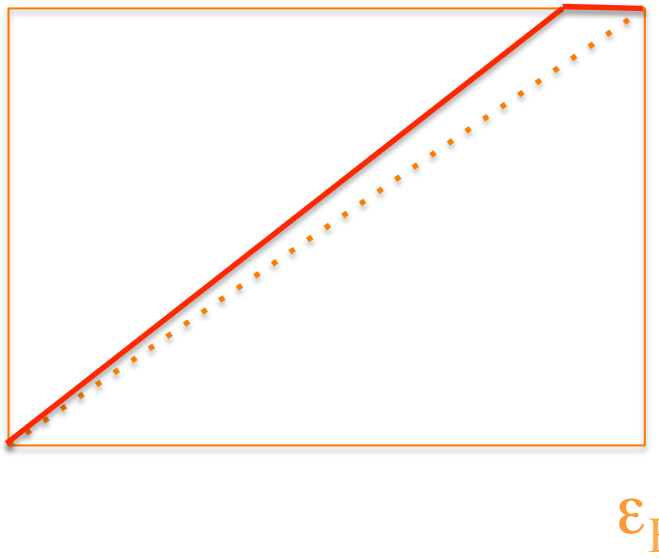
score

ROC curve



Local big difference (e.g. non overlapping distribution, hole)

score

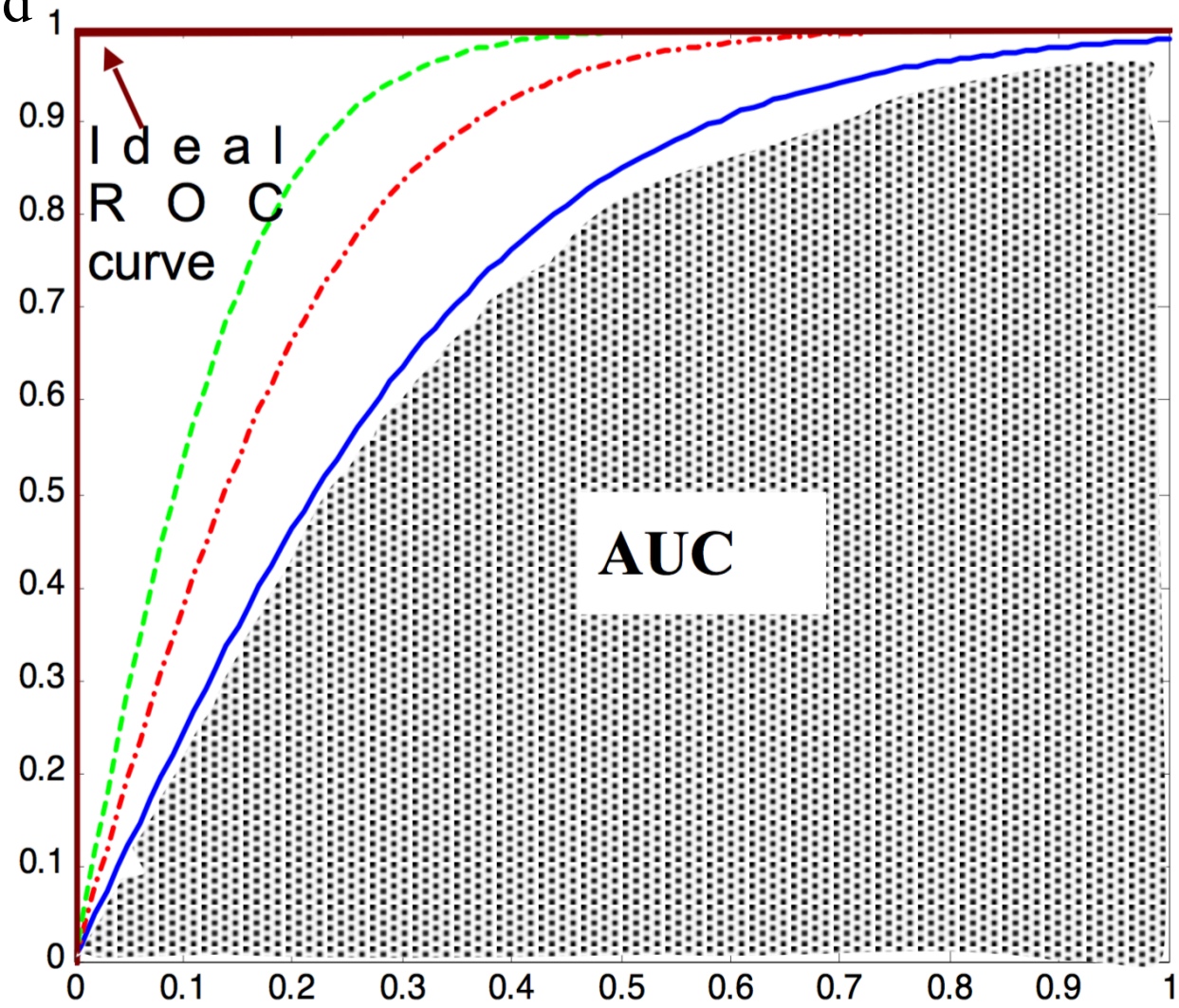


Today

- Dataset built from the Higgs Machine Learning challenge dataset (on CERN Open Data Portal)
  - Lepton, and tau hadron 3 momentum, MET : PRImary variables
  - DERived variables (computed from the above) from Htautau analysis
  - Jet variables dropped
- ➔reference dataset
- "Skewed" dataset built from the above, introducing small and big distortions (what has been done will be revealed in the course of the afternoon)
- Training dataset mixes the two populations, with isSkewed label to distinguish between them
- Classifier (your choice) trained on training dataset
- Goal is to statistically separate entries from the Skewed dataset from entries from the reference dataset
- AUC (Aread Under the Curve) of the ROC will be the figure of merit

AUC

Skewed



Reference