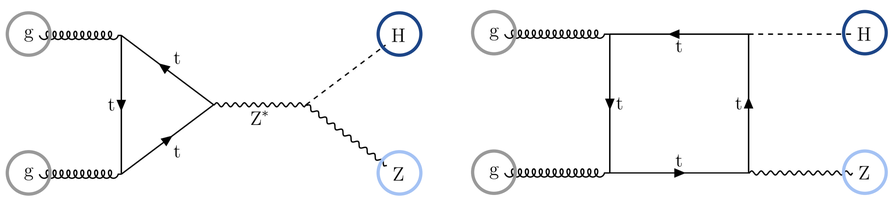
Binning High-Dimensional Classifier Output for HEP Analyses through a Clustering Algorithm

Svenja Diekmann, Niclas Eich & Martin Erdmann - RWTH Aachen University

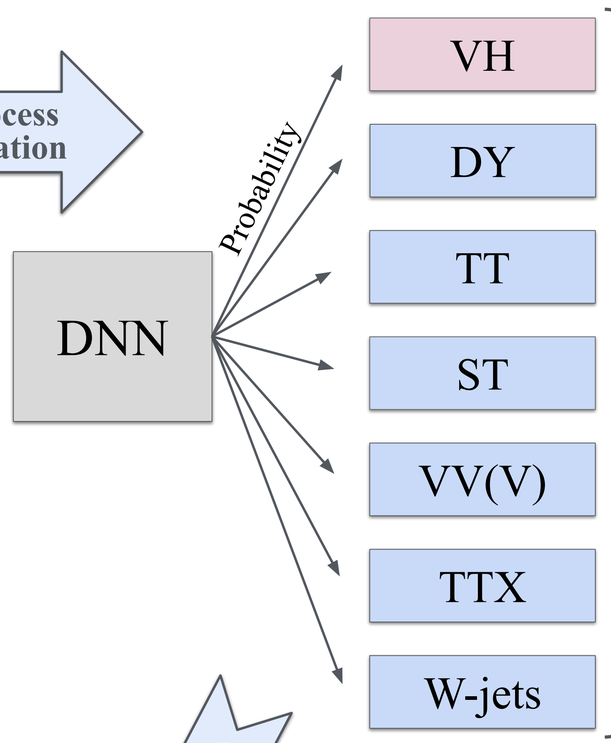
Motivation & Context

- search for VH
- specialised analysis for
  gg →ZH production
- yet unmeasured
- small cross section:
  - need for sensitivity enhancement

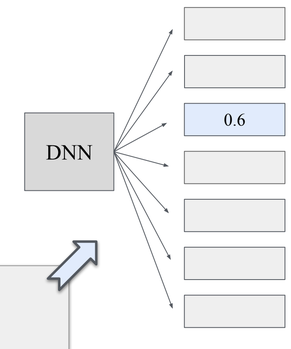


Process Assignment by DNN

Multiprocess Classification



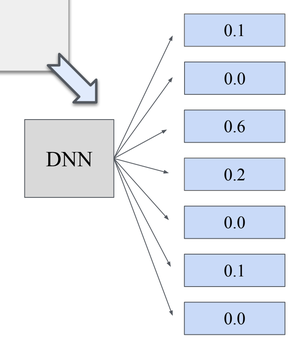
Standard Approach:
- projection to 1 dimension
- loss of information



High dimensional output:

$E_1 = (p_{VH}, p_{DY}, p_{TT}, \ldots)$
$E_2 = (p_{VH}, p_{DY}, p_{TT}, \ldots)$
$E_3 = (p_{VH}, p_{DY}, p_{TT}, \ldots)$
…

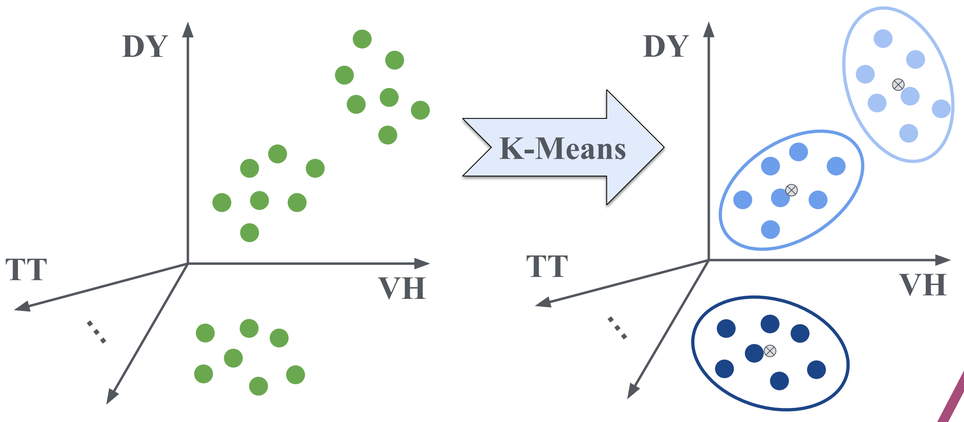
New Approach:
- keep **all** dimensions
- **no** loss of information



New Approach

High-Dimensional Clustering

K-Means Clustering Algorithm:
- random initialisation of **k** cluster centers
- eventwise assignment to closest cluster
- iterative update of cluster centers
  by **mean** of assigned events

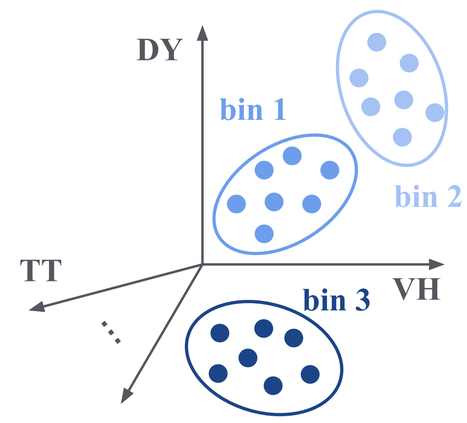


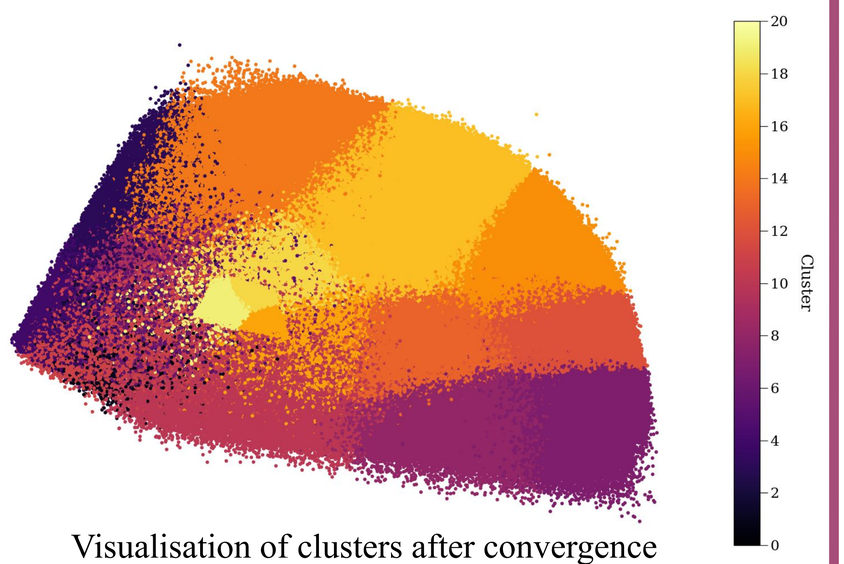
Convergence

Agnostic binning algorithm in high-dimensional space

Binning from Clusters

Identify clusters as bins for likelihood fit





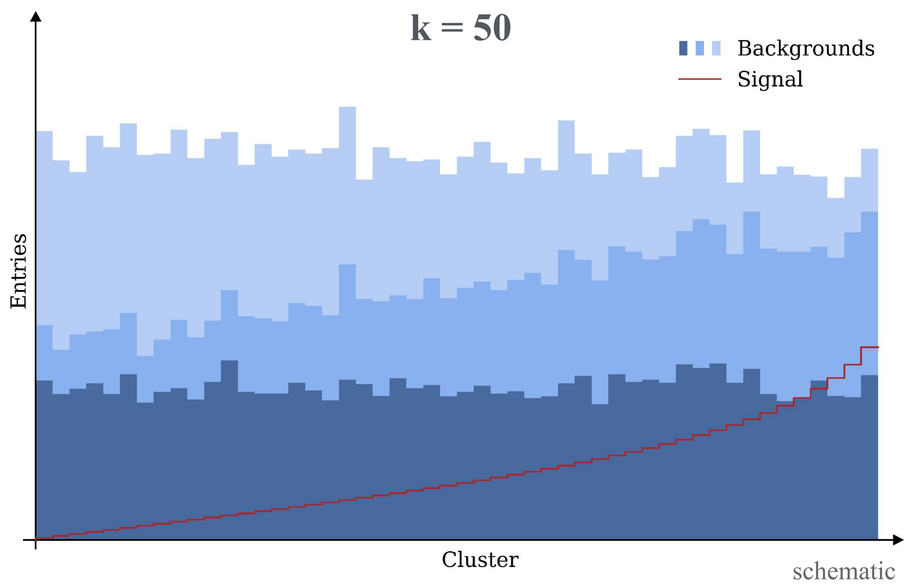
Visualisation of clusters after convergence by **P**rincipal **C**omponent **A**nalysis in 2D

k = 20

Physical Interpretation

Resulting Histograms

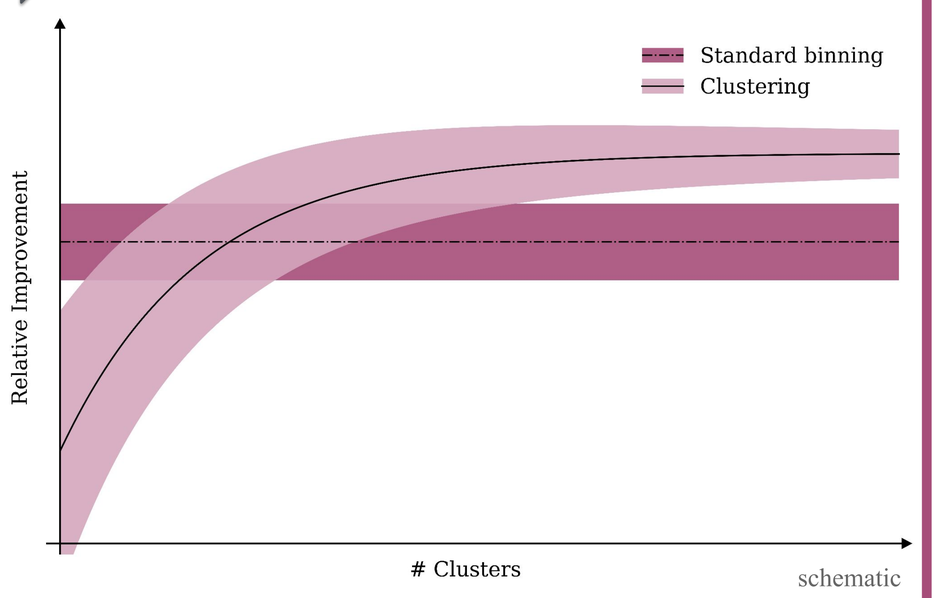
- clustering creates high signal and high background regions
- cluster index has no physical meaning
  - sorted by the total signal contribution (red line)



Inference

Sensitivity Enhancement

Sensitivity increase for high numbers of clusters:
- improvement compared to the standard method including statistical uncertainties



ACAT 2022 - svenja.diekmann@cern.ch

Federal Ministry of Education and Research
DFG Deutsche Forschungsgemeinschaft
Physics Institute III
RWTH AACHEN UNIVERSITY