ACAT05

May 22 - 27, 2005

DESY, Zeuthen, Germany

The use of Clustering Techniques for the Classification of High Energy Physics Data

Mostafa MJAHED

Ecole Royale de l'Air, Mathematics and Systems Dept.

Marrakech, Morocco

The use of Clustering Techniques for the Classification of High Energy Physics Data

- Production of jets in e^+e^-
- Methodology
- The use of Clustering Techniques for the Classification of physics processes in e^+e^-
- Conclusion

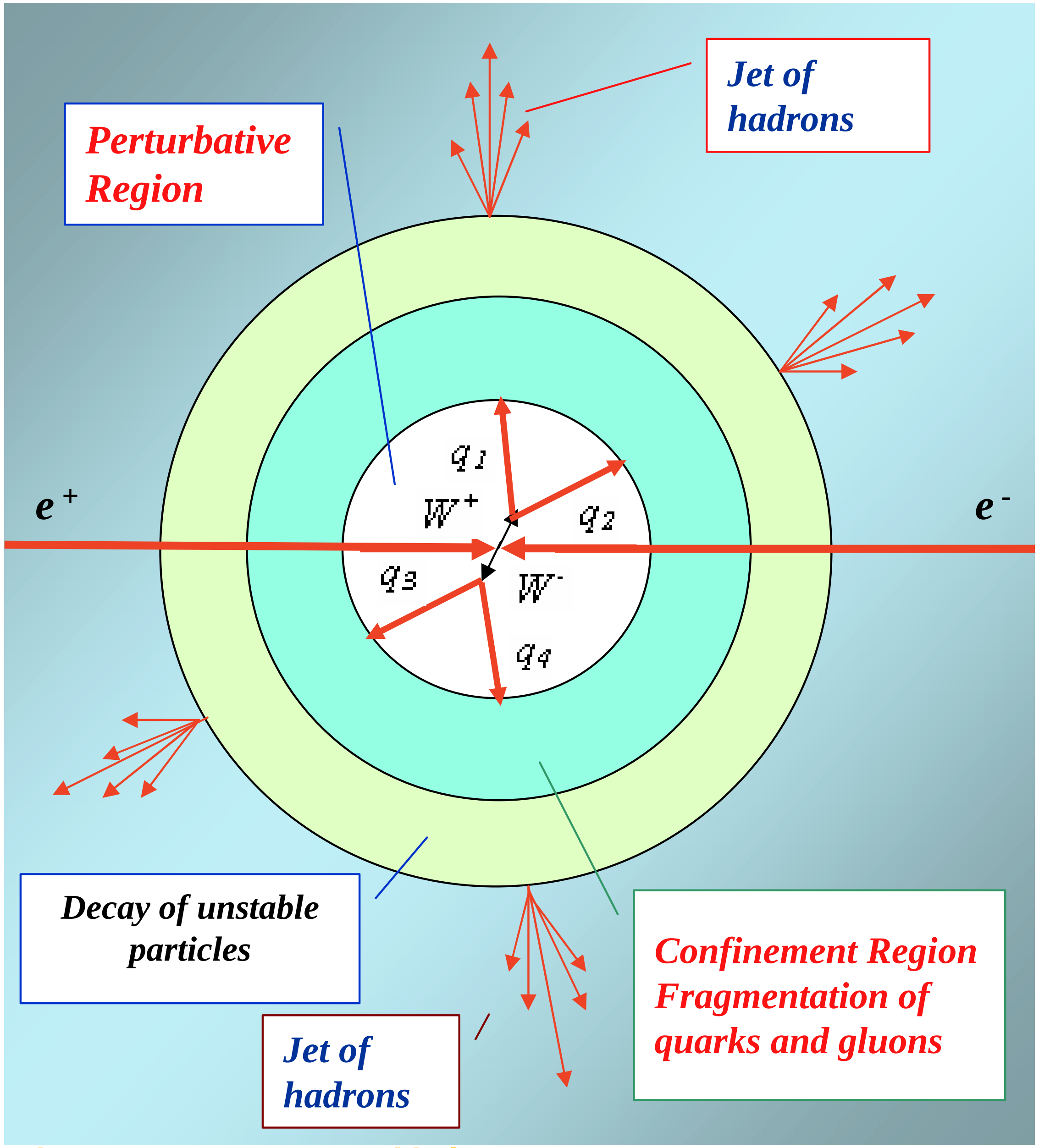
Production of jets in e^+e^-

- *Annihilation* $e^+e^- \rightarrow W^+W^-, ZZ, ZH$ *(H:Higgs) (LEP2 and beyond)*

- ***Decay of produced bosons:***
  $\gamma^*/Z^0 \rightarrow q\bar{q}$, $W^+ \rightarrow q_1 q_2, W^- \rightarrow q_3 q_4$
  $H^0 \rightarrow q\bar{q}$ …)

- ***Fragmentation of quarks and gluons and production of unstable particles***

- ***Decay of unstable particles to observed hadrons***



Production of jets in e^+e^-

***LEP2 and beyond:** observation of processes with dominants jets topologies:*

- ***Production of pairs W^+W^- :***

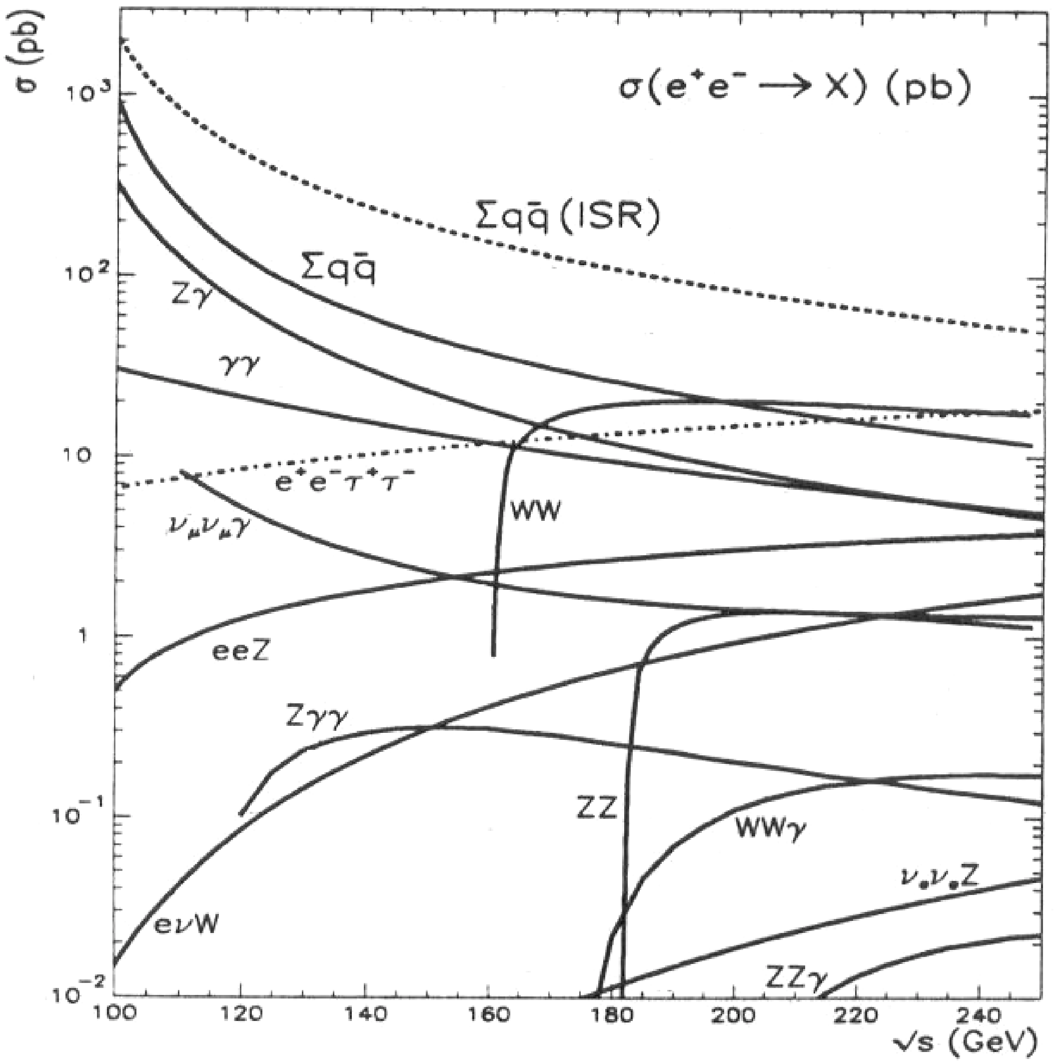
  $e^+e^- \rightarrow W^+W^- \rightarrow qql\nu_l\,,\ qqqq$

- ***Emergence of new particles as the Higgs Boson:***

  $e^+e^- \rightarrow ZH \rightarrow q\bar{q}b\bar{b}, \nu\bar{\nu}b\bar{b}\ (\tau^+\tau^- q\bar{q},\ q\bar{q}\tau^+\tau^-)$

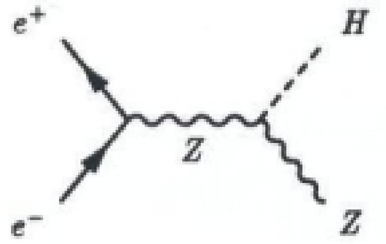
- ***Production of new processes:***

  $e^+e^- \rightarrow ZZ \rightarrow qq\, l\nu_l\,,\ qqqq,...$

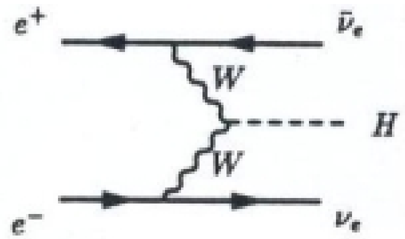


Higgs boson Production

Higgs-strahlung:
$e^+ e^- \to ZH$

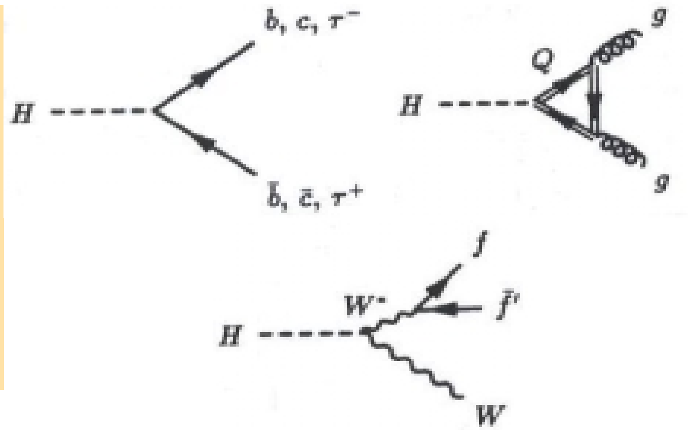


Fusion WW

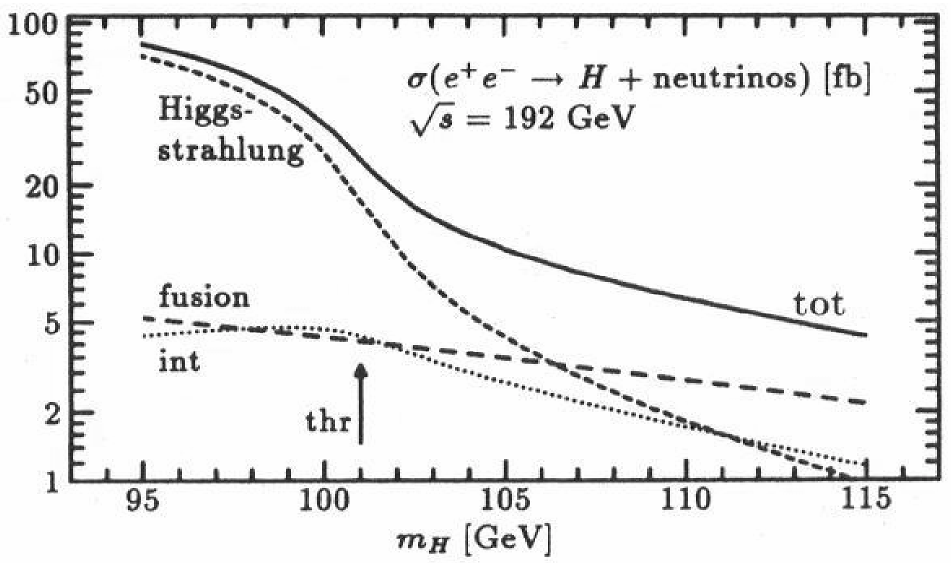


Decay Modes:

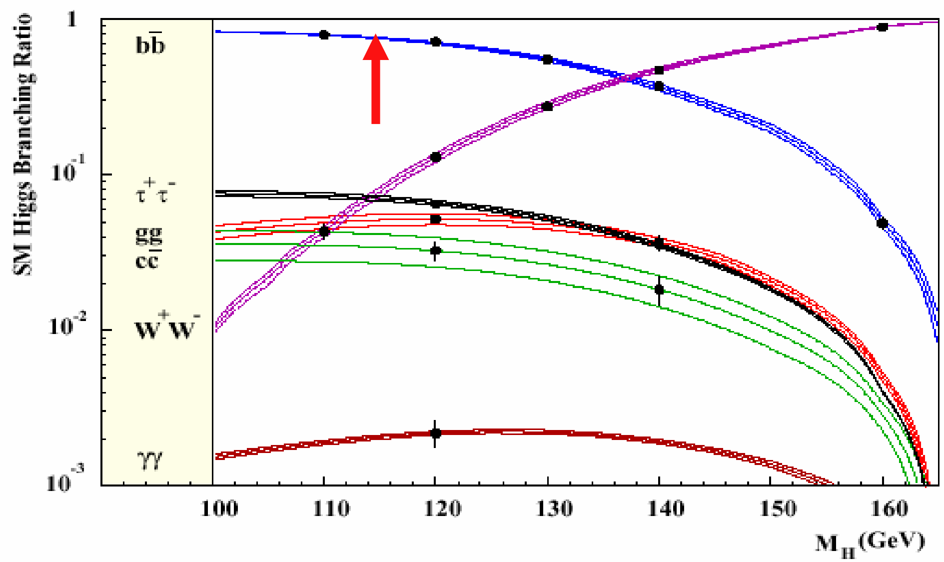
- *decay into quarks: $H \to bb$ and $H \to cc$*
- *leptonic decay $H \to \tau^+ \tau^-$*
- *gluonic decay $H \to g\,g$*
- *decay into virtual W boson pair: $H \to W^+W^-$*



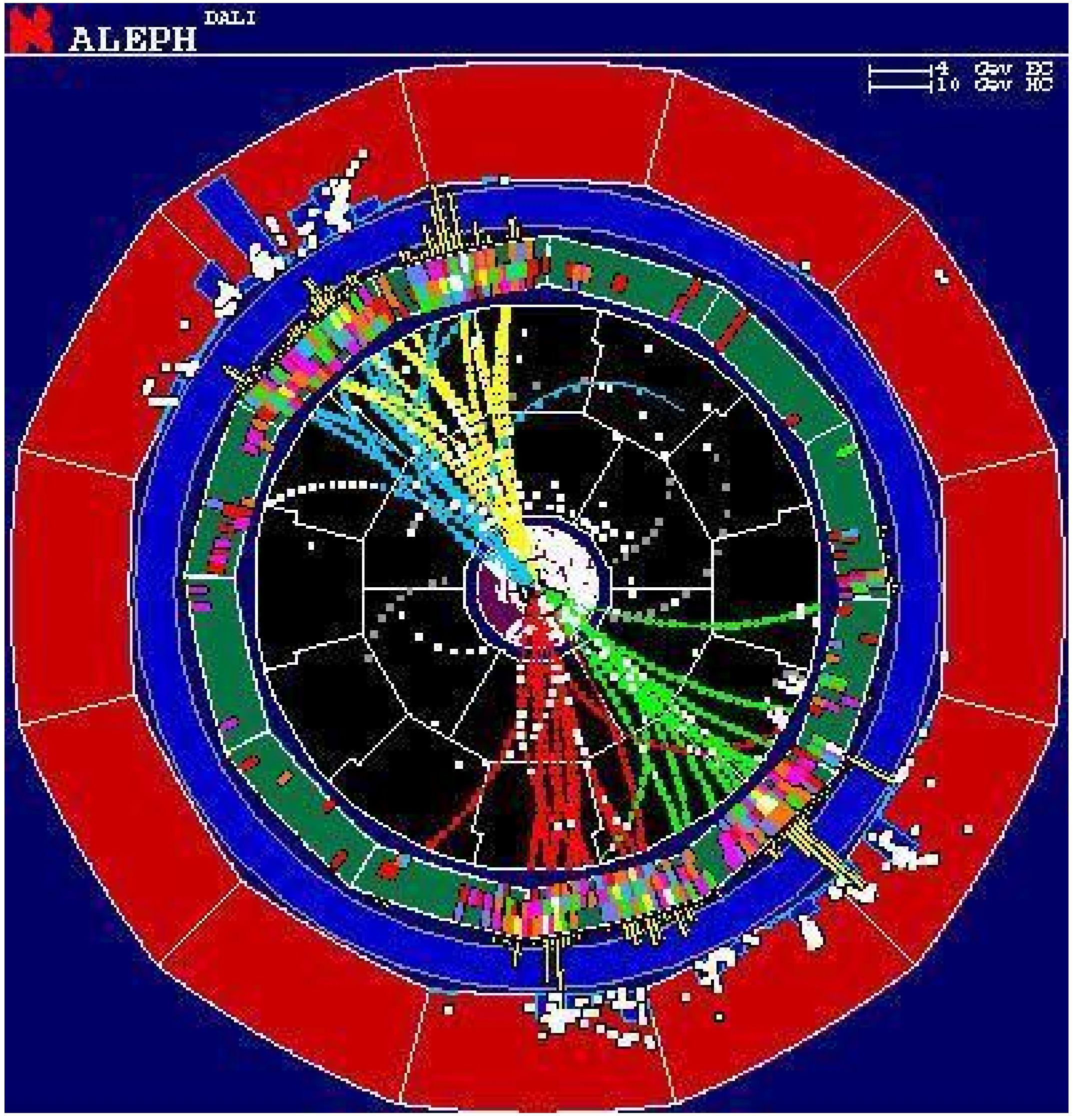
- ***Cross Section***



- ***Branching Ratio***



Production of jets in e^+e^-



- *HZ ALEPH candidate*
  $e^+ e^- \rightarrow H\,Z \rightarrow q\bar{q}b\bar{b}$

Jets analysis in e^+e^-

- *Analysis of W bosons pairs and research of new particles as the Higgs boson.*

- *Measure of the masse of W*
- *Measure of the Triple Gauge Coupling (TGC); coupling between 3 bosons*

Prediction of limits concerning the mass of the Higgs boson

- *These analyses are subjected to the identification of the different processes, with dominant jets topologies with a very high efficiency*

- **Need to use Pattern Recognition methods**

Pattern Recognition

$f: \quad X \rightarrow Y$

$x_i \in X \rightarrow y_j \in Y$

$$X(x_{ij}) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \rightarrow Y(y_j) = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_k \end{bmatrix}$$

- **Characterisation** *of events: research and selection of p variables or attributes*
- **Interpretation:** *definition of k classes*
- **Learning: association** $(x_i \rightarrow y_j) \Rightarrow f$
- **Decision** $(x_i \rightarrow y_j)$ *using f for any* x_i

Pattern Recognition Methods

- **Statistical Methods**
  - ***Principal Components Analysis PCA***
  - ***Decision Trees***
  - ***Discriminant Analysis …***
  - ***Clustering (Hierarchical, K-means, …)***

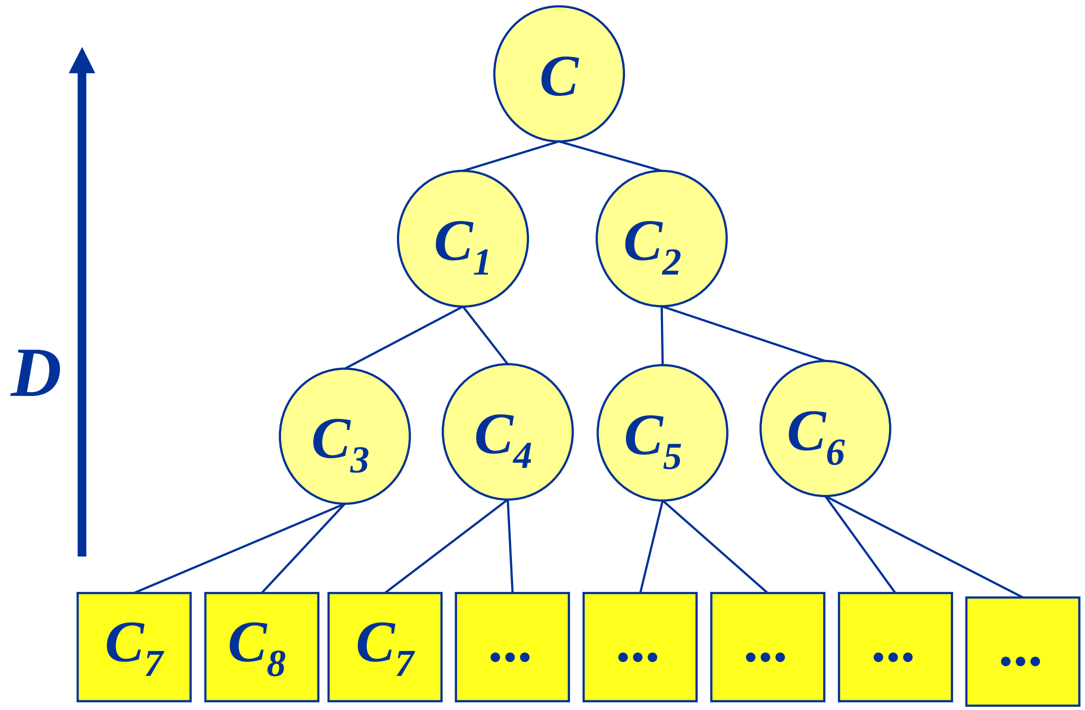
- **Connectionist Methods**
  - ***Neural Networks***
  - ***Genetic Algorithms …***

- **Other Methods**
  - ***Fuzzy Logic, Wavelets ...***

Hierarchical Clustering Technique

$$X(x_{ij}) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$D(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}$$



- *1. The distances between all the pairs of events x_i and x_j are computed*
- *2. Choice of the two most distant events: $C \rightarrow (C_1, C_2)$*
- *3. Assignation of all xi to the closer class C_1 or C_2*
- *4. Repeat the steps 2 and 3 for $C_1 \rightarrow (C_3, C_4)$ and $C_2 \rightarrow (C_5, C_6)$*
- *5. Repeat the step 4 for $C_i \rightarrow (C_j, C_k)$*

K-Means Clustering Technique

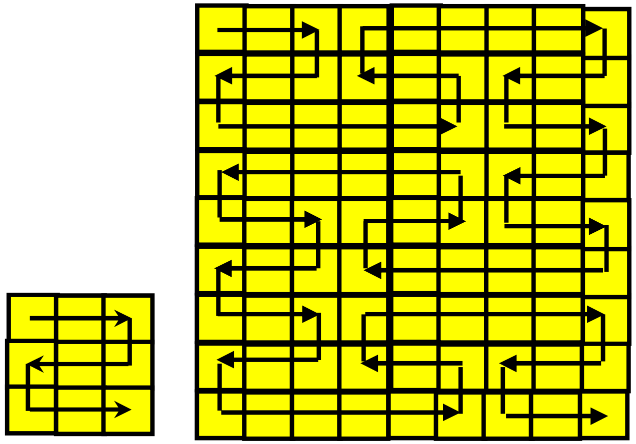
Given K, the K-means algorithm is implemented in 4 steps:

- *Partition events into K non empty subsets*
- *Compute seed points as the centroids (mean point) of the cluster*
- *Assign each event to the cluster with the nearest seed point*
- *Go back to step 2, stop when no more new assignment*

Parameters:

- *Choice of distances*
- *Supervised or unsupervised Learning*

Clustering by a Peano Scanning Technique



Example of an analytical Peano square-filling curve

- *Decomposition of data into p-dimensional unit hyper-cube*
  $$I_p = [0, 1] \times [0, 1] \times \ldots \times [0, 1]$$
- *Construction of a space filling curve* $F_p\,(t): I_1 \rightarrow I_p$
- *Compute the position of X (data) on the SFC, i.e.,* $t = \psi(x)$
- *Find the set K of nearest neighbours of t in the transformed learning set T*
- *Classify the test sample to the nearest class in set K*

Efficiency and Purity of a Pattern Recognition Method

- **Validation**

Test events	Classification	
	C_1	C_2
$C_1 : N_1$	N_{11}	N_{12}
$C_2 : N_2$	N_{21}	N_{22}
Total	M_1	M_2

- *Efficiency of classification for events of class C_i*

$$E_i = \frac{N_{ii}}{N_i}$$

- *Purity of classification for events of class C_i*

$$P_i = \frac{N_{ii}}{M_i}$$

Application

4 jets

$e^+ e^- \to HZ \to b\bar{b}q\bar{q}$

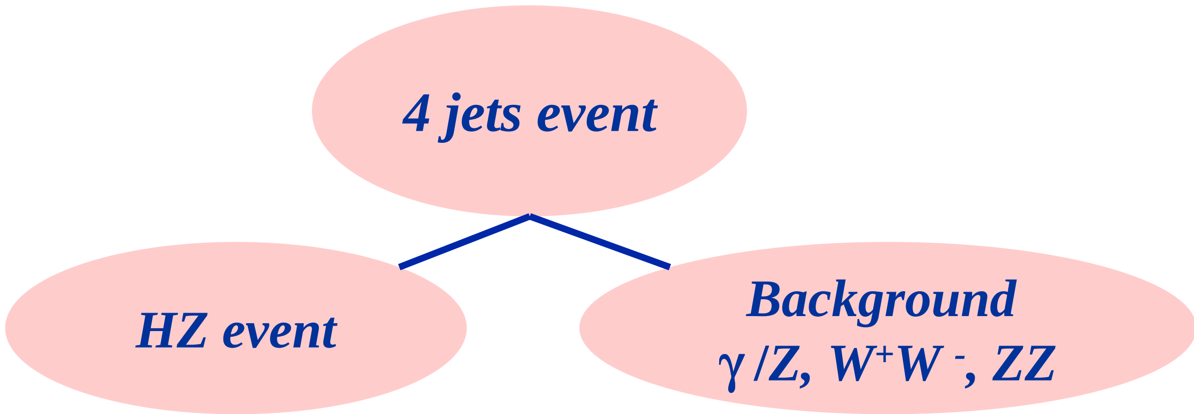
$e^+ e^- \to ZZ \to qqqq$

$e^+ e^- \to W^+W^- \to qqqq$

$e^+ e^- \to \gamma/Z \to q\bar{q}q\bar{q}$

- *Characterisation of the Higgs boson in the 4 jets channel,*
  $e^+e^- \to ZH \to q\bar{q}b\bar{b}$ *, by clustering techniques*

Characterisation of the Higgs boson in 4 jets channel $e^+e^- \to ZH \to q\bar{q}b\bar{b}$ by the use of clustering techniques



- *Events generated by the LUND MC (JETSET 7.4 and PYTHIA 5.7) at $\sqrt{s}$ = 300 GeV, in the 4 jets channel*
- *$e^+ e^- \to HZ \to q\bar{q}b\bar{b}$ (signal: Higgs boson events), M_H = 125 GeV/c^2*
- *$e^+e^- \to W^+W^- \to qqqq$, $e^+ e^- \to Z/\gamma \to q\bar{q}gg$, $q\bar{q}q\bar{q}$, $e^+ e^- \to ZZ \to q\bar{q}q\bar{q}$ (Background events)*

- ***Research of discriminating variables:** variables characterizing the presence of b quarks*

Variables

- ***Thrust***

$$T = max\sum_{i=1}^{N}(\vec{p}_i.\hat{n}) = max\sum_{i=1}^{N}|\vec{p}_{i//}|$$

- ***Sphericity S***

$$S = min\ S(\hat{n}) \qquad S(\hat{n}) = \frac{3}{2}\frac{\sum_{i=1}^{N}\vec{p}_{iT/\hat{n}}^{\,2}}{\sum_{i=1}^{N}\vec{p}_i^{\,2}}$$

- ***Boosted Aplanarity: BAP***

$$BAP = \frac{3}{2}min\frac{\sum_{i=1}^{N}|\vec{p}_{iT\,out}|^2}{\sum_{i=1}^{N}\vec{p}_i^{\,2}}$$

- ***Max3 (M_{jet}), Max3 (E_{jet}):***
the 3th value of the jet masses and jet energies in each event

- ***Bed: Event broadening***

Bed = Min B_{hemi}

$$B_{hemi} = \frac{\sum_{i=1}^{n_t}|p_{iT}|}{\sum_{i=1}^{n_t}|p_i|}$$

- ***Mincos: Min ($\cos\theta_{ij} + \cos\theta_{kl}$):***

The minimal sum of cosines by using all the permutations ijkl.

- ***Max (M_{jet}), Max (E_{jet}):***
the maximal value of the jet masses and jet energies in each event

- ***M_{min}, E_{min} :***
the 4th value of the jet masses and jet energies in each event

- ***Rapidity-impulsion weighted Moments M_{nm} :***

$$M_{nm} = \sum_{i\in Jet}\eta_i^{\,n}\,.\,p_{iT}^{\,m}$$

η_i rapidity:

$$\eta_i = \frac{1}{2}.Log\left(\frac{E_i + p_{i//}}{E_i - p_{i//}}\right)$$

Discriminating Power of variables

- *Test Function F_j*

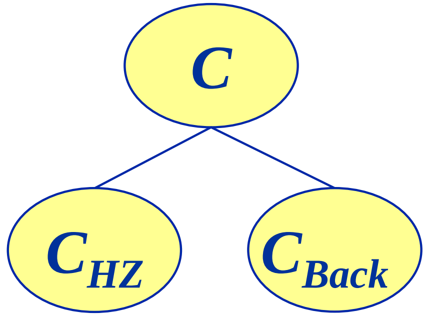
$$F_j = \frac{(n-k)}{(k-1)}\frac{B_j}{W_j} \qquad j=1, \ldots, 17.$$

- *B_j, W_j: Between and Within-classes Variance Matrix for variable j.*
- *n total number of events (signal+ background),*
- *k number of classes (2)*

- *The discriminating power of each variable V_j is proportional to the values of F_j (j=1, …, 17).*

Variable	Pouvoir discriminant: F			
	HZ/WW	$HZ/q\bar{q}q\bar{q}\ldots$	HZ/ZZ	HZ/All
T	0.042	0.092	0.005	0.085
Bed	0.021	0.213	0.056	0.132
S	0.066	0.084	0.032	0.054
Mincos	0.132	0.137	0.057	0.212
BAP	0.124	0.145	0.018	0.017
Max (E_{jet})	0.141	0.116	0.088	0.112
Max (M_{jet})	0.082	0.134	0.115	0.113
Max3 (E_{jet})	0.115	0.081	0.054	0.101
Max3 (M_{jet})	0.031	0.095	0.059	0.082
E_{min}	0.024	0.212	0.053	0.121
M_{min}	0.018	0.151	0.043	0.094
M_{11}	0.045	0.012	0.085	0.081
M_{21}	0.041	0.011	0.048	0.035
M_{31}	0.039	0.018	0.069	0.068
M_{41}	0.048	0.016	0.071	0.051
M_{51}	0.051	0.012	0.082	0.032
M_{61}	0.052	0.014	0.021	0.029

Hierarchical Clustering Classification



- ***The most separating distance $D_{HZ/Back}$ between the classes C_{HZ} and C_{Back} is searched and the corresponding cut $D_{HZ/Back}{}^{*}$ is computed.***
- ***The classification of a test event x_0 is then obtained according to the algorithm:***

$$\text{if } D_{HZ/Back}(x_o) \geq D_{HZ/Back}^{*} \text{ then } x_o \in C_{HZ} \text{ else } x_o \in C_{Back}$$

- $D_{HZ/Back} = 0.01\ Mincos + 0.32\ M_{axE} + 0.11\ M_{ax3E} + 0.52 E_{min} + 0.36\ BAP + 0.87\ Bed + 0.41\ M_{11} + 0.38\ M_{31}$
- $D_{HZ/Back}^{*} = 2.51$

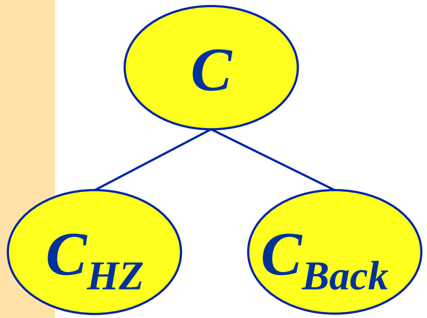
- **Classification of test events**

Test events		Hierarchical clustering		
		C_{HZ}	C_{Back}	
C_{HZ} : 1000		601	399	
C_{Back}	$C_{\gamma Z}$:1000	403	1791	597
	C_{ZZ} :1000	405		595
	C_{WW} :1000	401		599

K-Means Clustering Classification

For K=2, the K-means algorithm is implemented in 4 steps:

- *Partition events into 2 non empty subsets*
- *Compute seed points as the centroids (mean point) of the cluster*
- *Assign each event to the cluster with the nearest seed point*
- *Go back to step 2, stop when no more new assignment*



- **Classification of test events**

Test events		K-Means clustering: C_{HZ}	K-Means clustering: C_{Back}	
C_{HZ} : 1000		591	409	
C_{Back}	C_{YZ} :1000	411	1764	589
	C_{ZZ} :1000	415		585
	C_{WW} :1000	410		590

Peano space filling curve Clustering Classification

- *By using the training sample:*

  $X = (x_i\,(M_{11}, M_{21}, M_{31}, M_{41}, M_{51}, M_{61}, T, S, BAP, Bed, Mincos, M_{axE}, M_{axM}, M_{ax3E}, M_{ax3M}, E_{min}, M_{min}), i=1,\ldots, N=4000)$ *and the known class labels:* C_{HZ}, C_{back}*, an approximate Peano space filling curve is obtained, allowing to transform the 17-dimensional space into unit interval.*

- Classification of test events

Test events		Peano space filling curve clustering: C_{HZ}	Peano space filling curve clustering: C_{Back}	
C_{HZ} : 1000		581	419	
C_{Back}	$C_{\gamma Z}$:1000	430	1707	570
	C_{ZZ} :1000	437		563
	C_{WW} :1000	426		574

Comparison

- **Comparison between the 3 clustering methods**

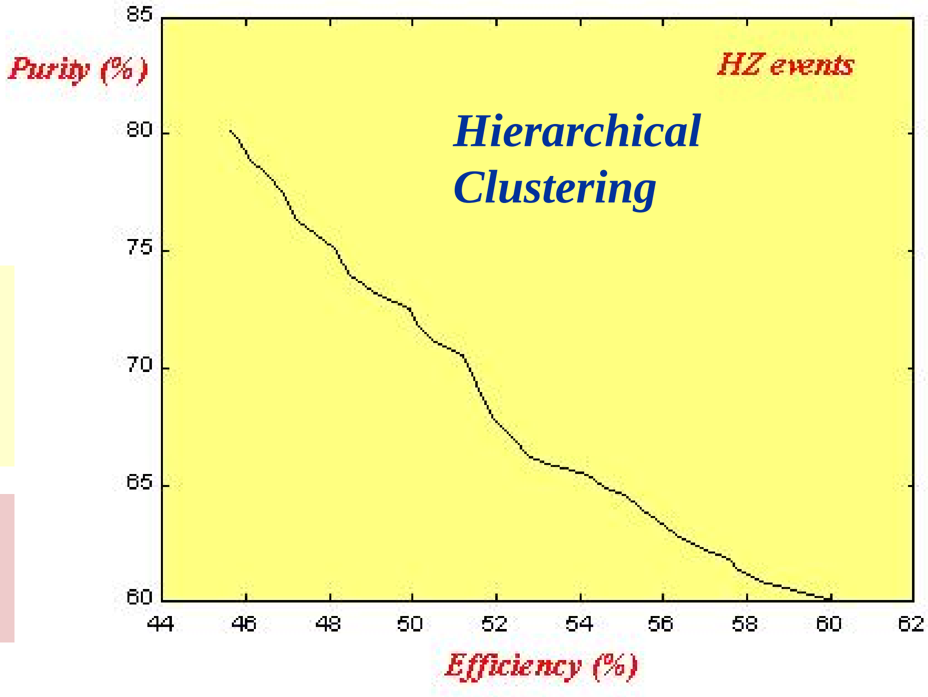
Method	*Efficiency (%)*		*Purity (%)*	
	C_{HZ}	C_{Back}	C_{HZ}	C_{Back}
Hierarchical Clustering	60.1	59.7	59.8	59.9
K-means Clustering	59.1	58.8	58.9	58.9
Peano scanning	58.1	56.9	57.4	57.6

- **Purity of classification vs cut's values D^* in hierarchical clustering**

$$D_{HZ/Back} = 0.01\ Mincos + 0.32\ M_{axE} + 0.11\ M_{ax3E} + .52E_{min} + 0.36\ BAP + 0.87\ Bed + 0.41\ M_{11} + 0.38\ M_{31} \qquad D_{HZ/Back}^* = 2.51$$

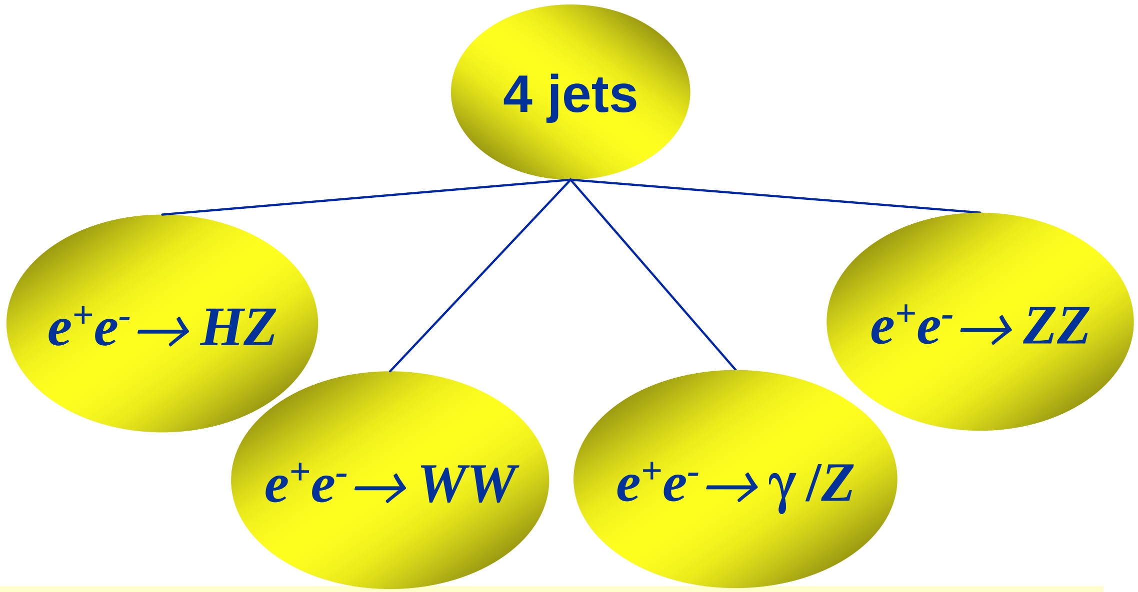
$D_{HZ/Back}^* = [1.65, 1.7, 1.75, …, 2.51, …, 2.65,$

$\Rightarrow$ *Purity(%) = [50, 51, 52, …,80]*



Conclusion

- **Variables**



- ***Characterisation of Higgs Boson events:***

  The most discriminating variables are: Mincos, M_{axE}, M_{ax3E}, E_{min}, BAP, Bed.

  They show the importance of information allowing to separate between b quark and udsc-quarks (separation between HZ events and background: $H \rightarrow bb$).

- ***Other variables as E_{min}, M_{min}, BAP, Bed, Mincos, may be used to identify events emerging from the background (i.e. $e^+e^- \rightarrow Z/\gamma \rightarrow 4$ jets).***

- ***Discrimination (γ/Z) / WW / ZZ:***

  using dijets properties: charge, broadness, presence of b quarks ...

Conclusion (cont)

Methods

- *Importance of Pattern Recognition Methods*
- *The improvement of an any identification is subjected to the multiplication of multidimensional effect offered by PR methods and the discriminating power of the proposed variable.*

- ***The hierarchical clustering method is more efficient than the other clustering techniques: its performances are in average 1 to 3 % higher than those obtained with the two other methods.***
- *Other cut's values $D^{*}_{HZ/Back}$ give other efficiencies and purities: We can reach values of purity permitting to identify the HZ events more efficiently*

- *Clustering techniques: comparative to other statistical methods : Discriminant Analysis, Decision trees,...*
- *Clustering techniques: less effective than neural networks and non linear discriminant analysis methods*