Aggregation Criteria

K. Gibert

Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, (IDEAI-UPC) Barcelona

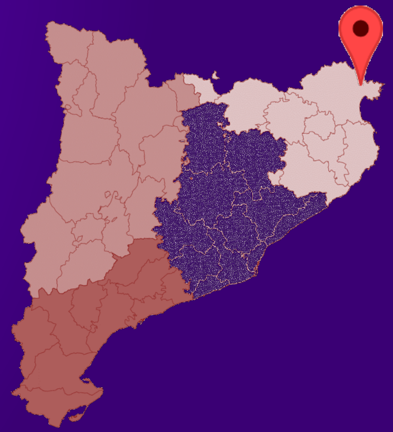
Dean of Illustrious Professional College of Informatics Engineering of Catalonia

President and founder of donesCOEINF

https://www.eio.upc.edu/en/homepages/karina,

karina.gibert@upc.edu, karina.gibert, @karinagibertk

Figuerenca



febrer de 2024

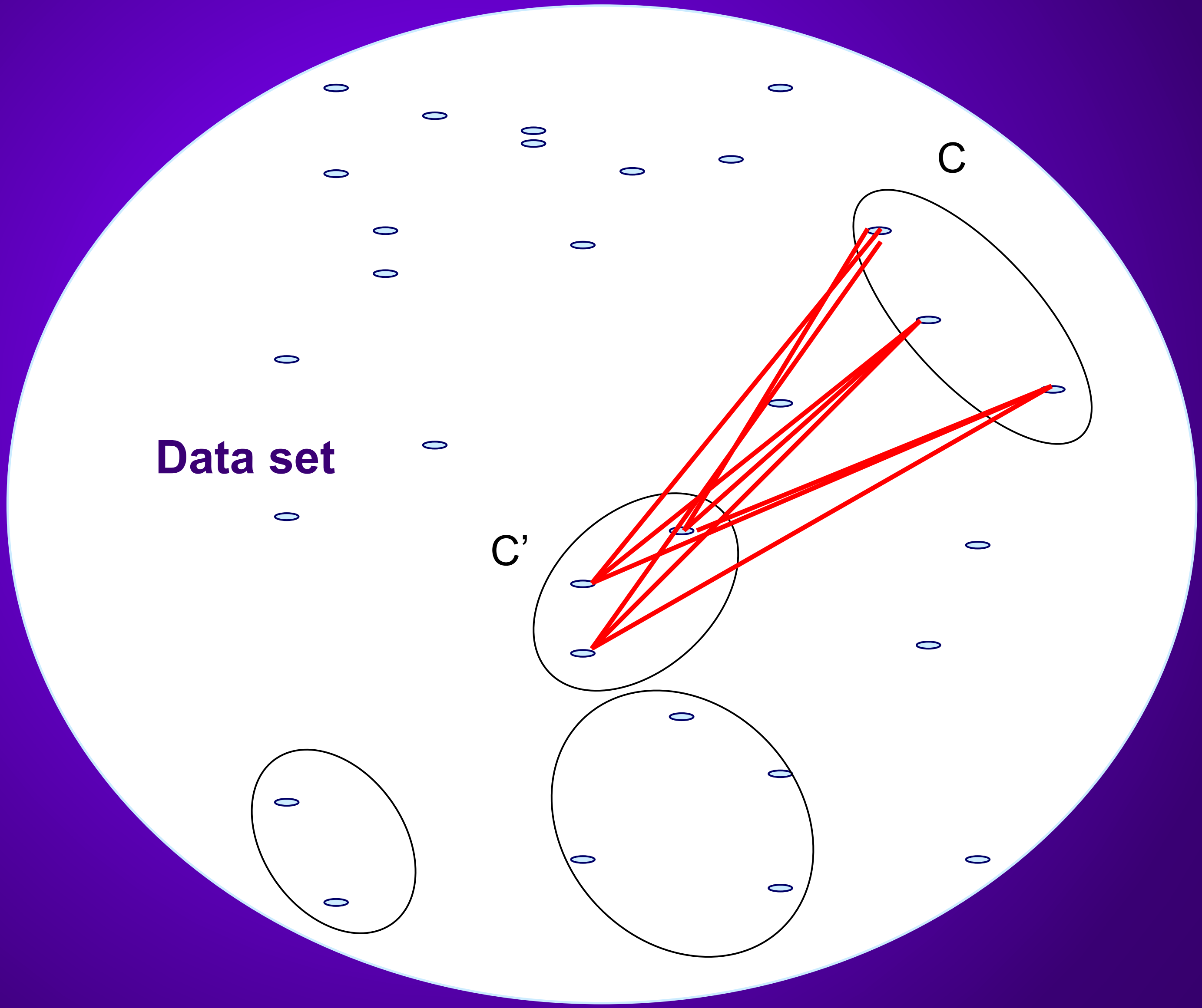
Ascendant hierarchical clustering

Hot points :

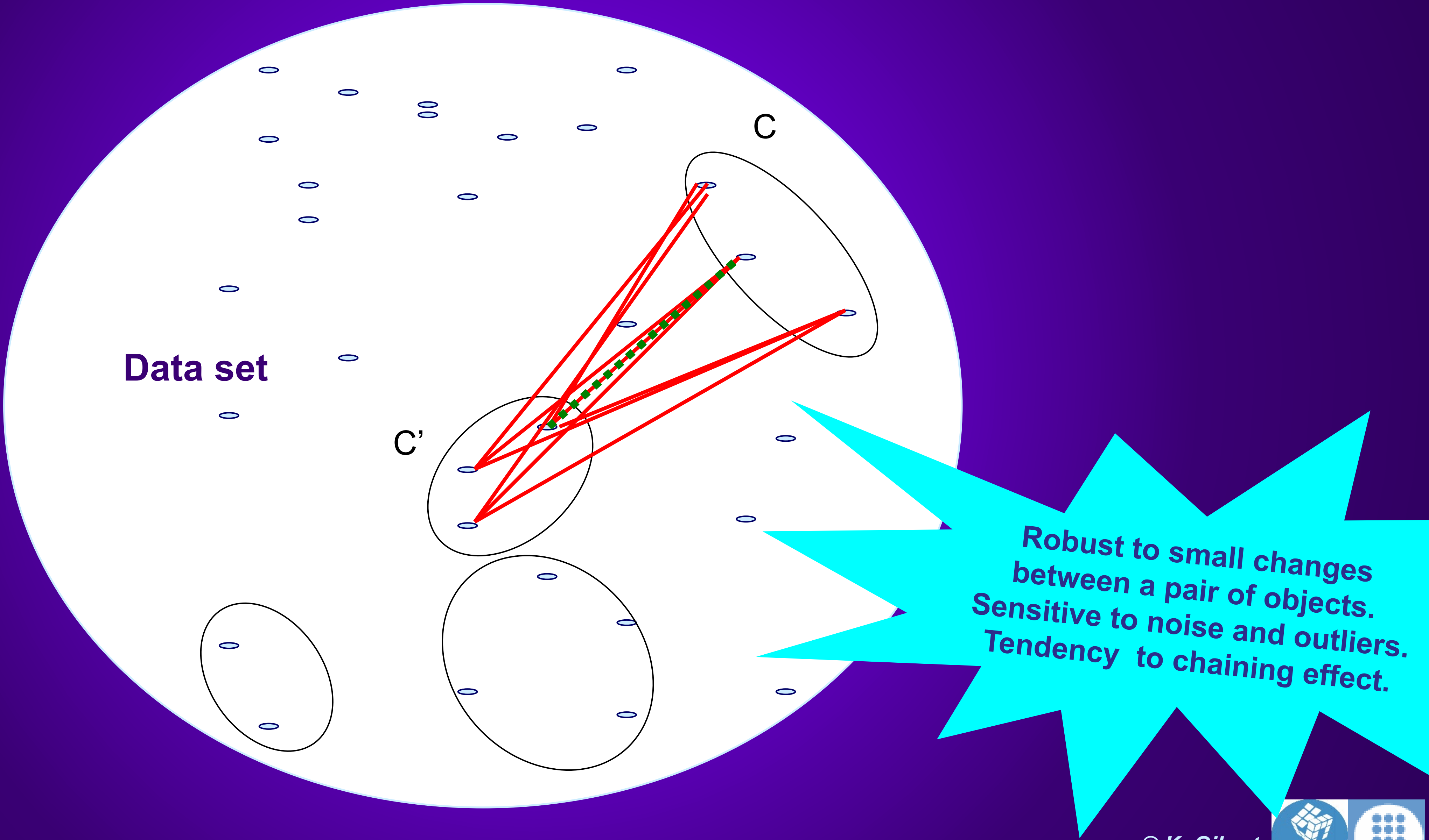
- Agregation criteria
  - Centroid criterion
  - Ward's criterion
- Distance between individuals
  - Only quantitative variables → Euclidean, Correlations
  - Only qualitative variables → χ^2 (Benzécri 80)
  - Heterogenous variables (compatibility measures) → Gower (71)
    Gowda i Diday (91)
    Gibert's Mixed (91)
    Ichino i Yaguchi (94)
    Ralambondrainy (95)
    Generalized Gibert's Mixed (13)

Aggregation Criteria

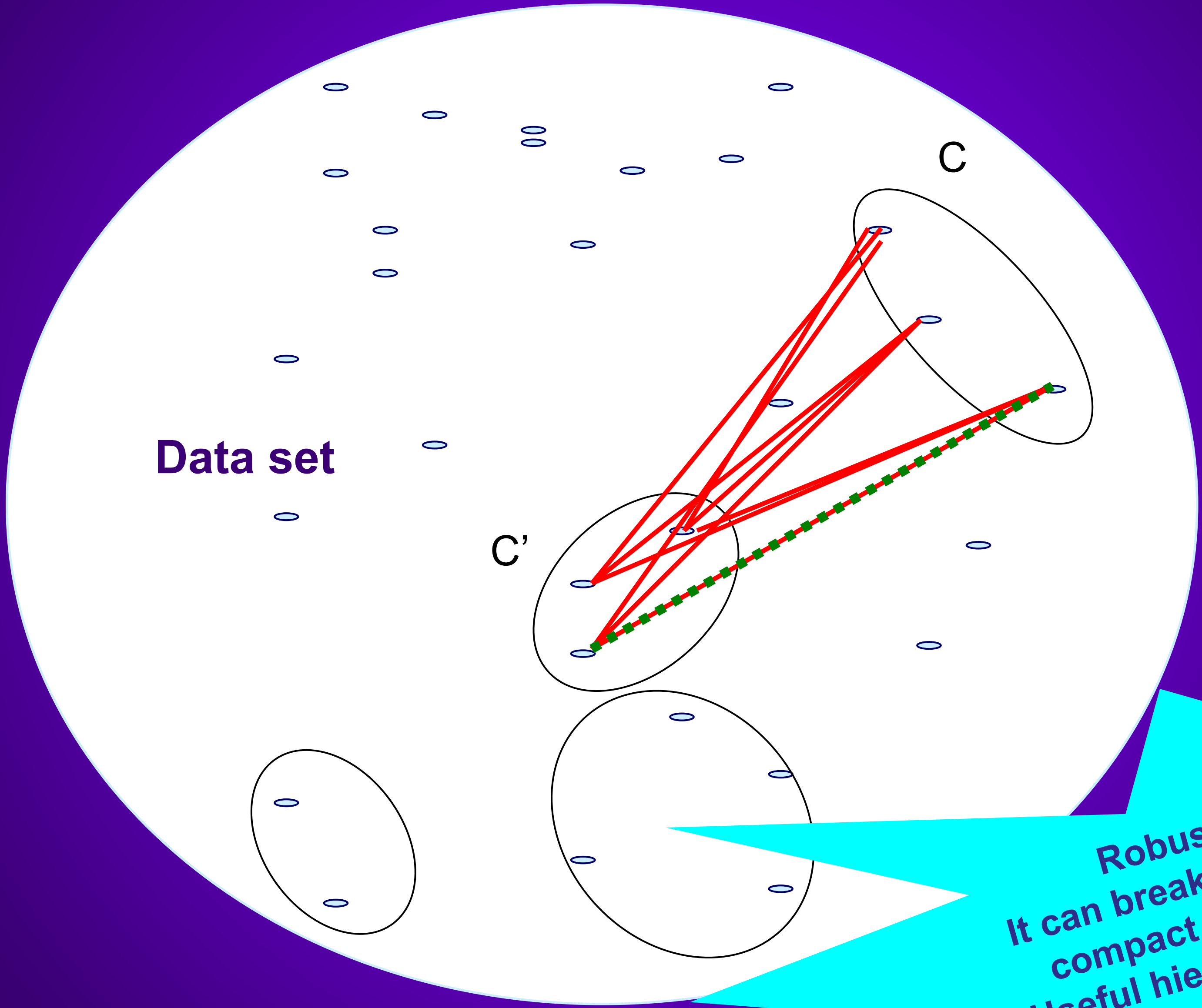
$d(C,C') = f(d(i,i')),\ i \in C,\ i' \in C'$



Single linkage (Sneath 1957) $d(C,C') = \min(d(i,i')),\ i \in C,\ i' \in C'$



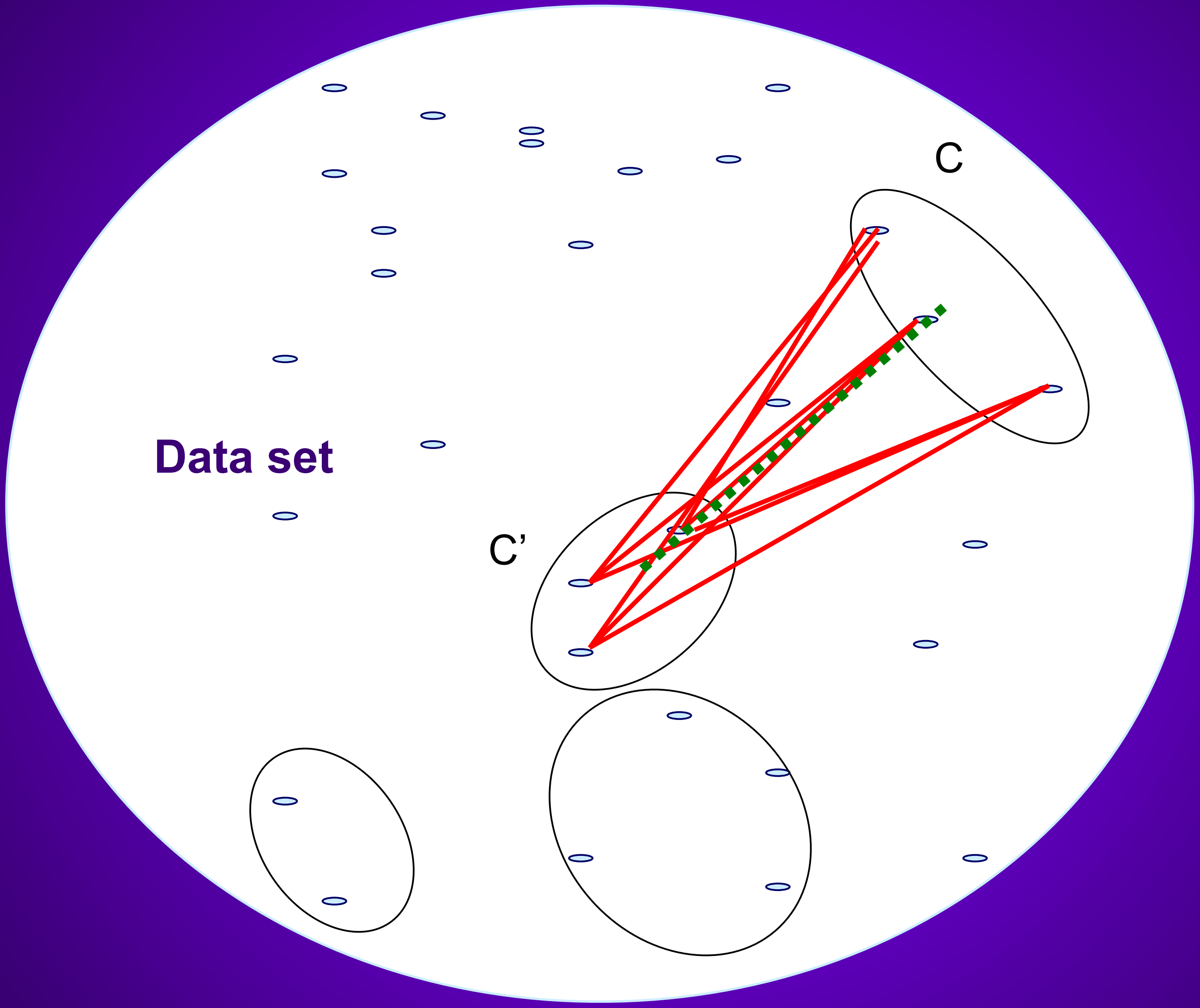
Complete linkage (Sorensen 1948) $d(C,C') = \max(d(i,i')), i \in C, i' \in C'$



Robust to outliers and noise.
It can break large clusters Conservative, compact clusters (Baeza-Yates 1992).
Useful hierarchies (Jain and Dubes 1998)

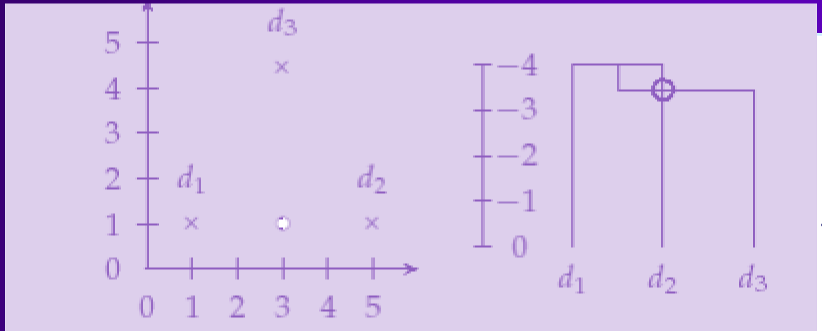
Average linkage

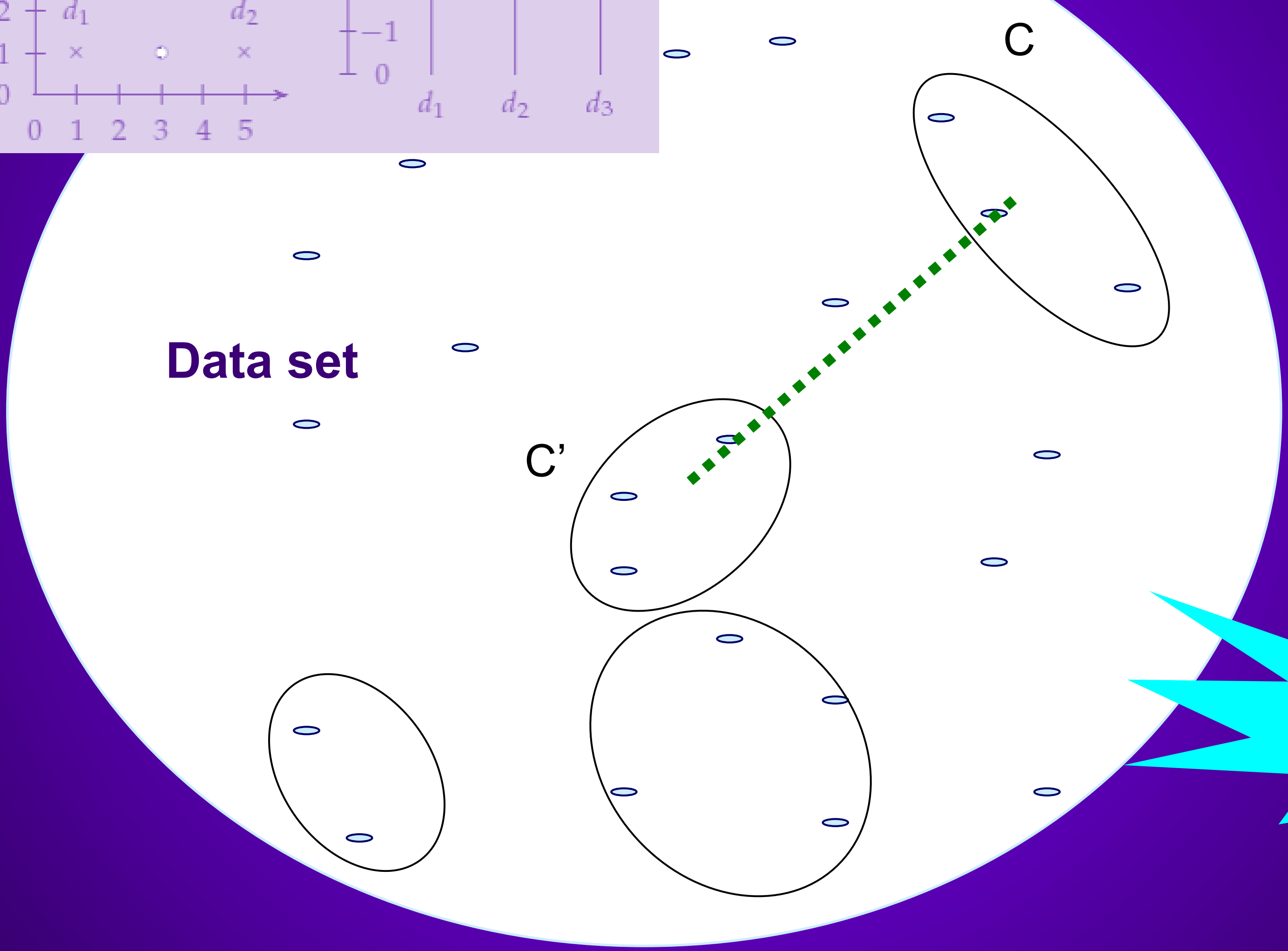
$d(C,C') = \text{mean}(d(i,i')),\ i \in C,\ i' \in C'$



Centroid

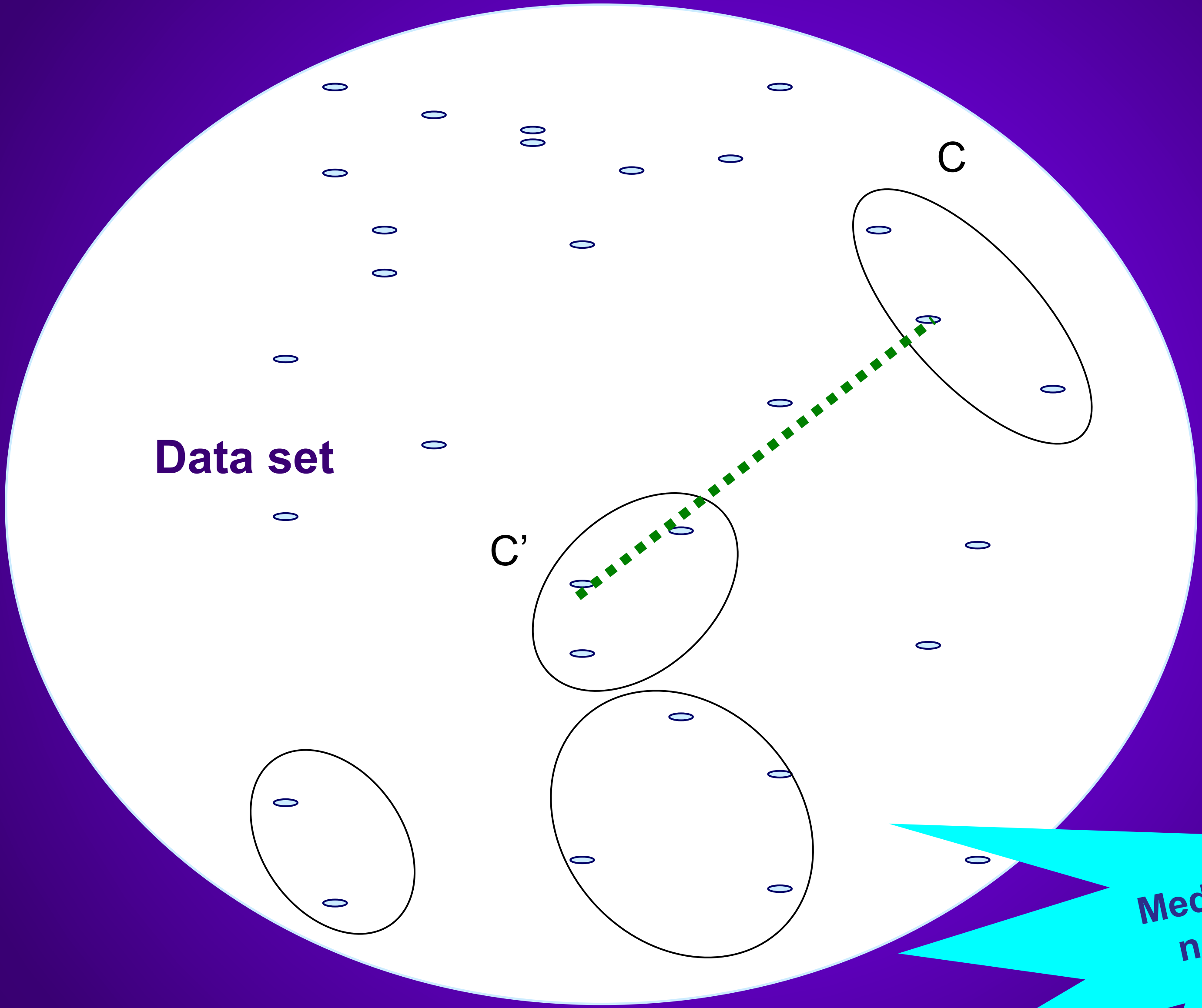
d(C,C’) = d(c,c’), c,c’ centroids





Produce inversions in the dendrogram

Median linkage (Gower 1967) d(C,C’) = d(c,c’)), c,c’ medoids



Median representativeness do not depend on cluster size

Ward's method (Ward 1963)

- Ascendant hierarchical method
- Group the two classes giving minimal inter-class inertia loss
- Inertia (physical concept)

$$M^2(\mathcal{I}/\bar{\imath}) = \sum_{i\in\mathcal{I}} m_i\, d^2(i,\bar{\imath})$$

- Huygens theorem

$$M^2(\mathcal{I}/\bar{\imath}) = Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) + Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}),$$

$$Mi^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) = \sum_{\mathcal{C}\in\mathcal{P}} M^2(\mathcal{C}/\bar{\imath}_{\mathcal{C}}) = \sum_{\mathcal{C}\in\mathcal{P}}\sum_{i\in\mathcal{C}} m_i\, d^2(i,\bar{\imath}_{\mathcal{C}})$$

$$Me^2_{\mathcal{P}}(\mathcal{I}/\bar{\imath}) = \sum_{\mathcal{C}\in\mathcal{P}} m_{\mathcal{C}}\, d^2(\bar{\imath}_{\mathcal{C}},\bar{\imath})$$

- Every aggregation increases intra-class inertia with

$$\Delta_\xi = Mi^2_{\mathcal{P}_{\xi+1}}(\mathcal{I}/\bar{\imath}) - Mi^2_{\mathcal{P}_\xi}(\mathcal{I}/\bar{\imath}) = m_{\mathcal{C}_e} d^2(\mathcal{C}_e,\bar{\imath}_{\mathcal{C}}) + m_{\mathcal{C}_d} d^2(\mathcal{C}_d,\bar{\imath}_{\mathcal{C}})$$

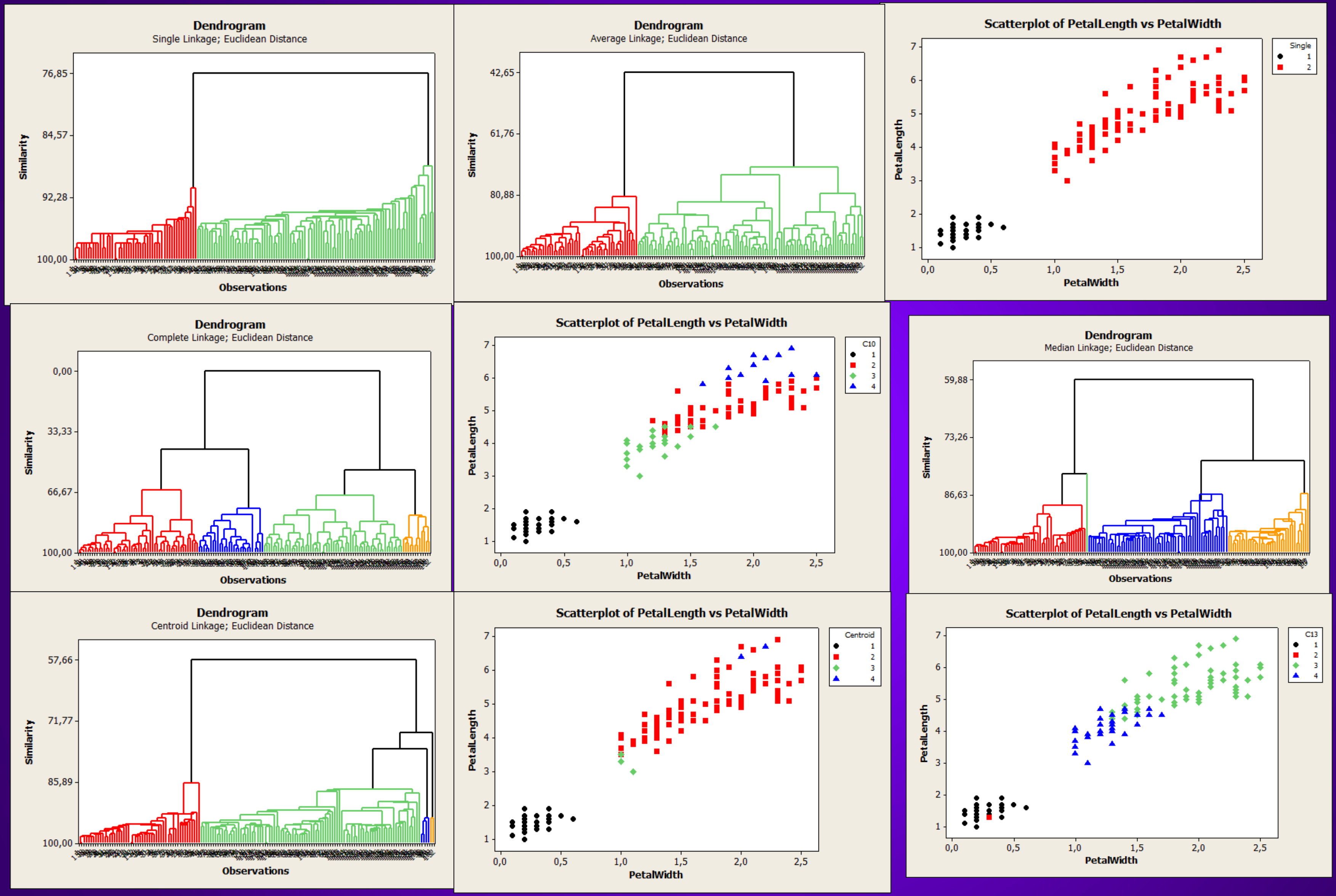
Minimize

- Inertia: related with quantity of information *(Thm Benzecri 1973)*
  - The more informative, more inertia
- Also with variability
  - The more variable, more inertia
- Quite popular

Interpretable classes

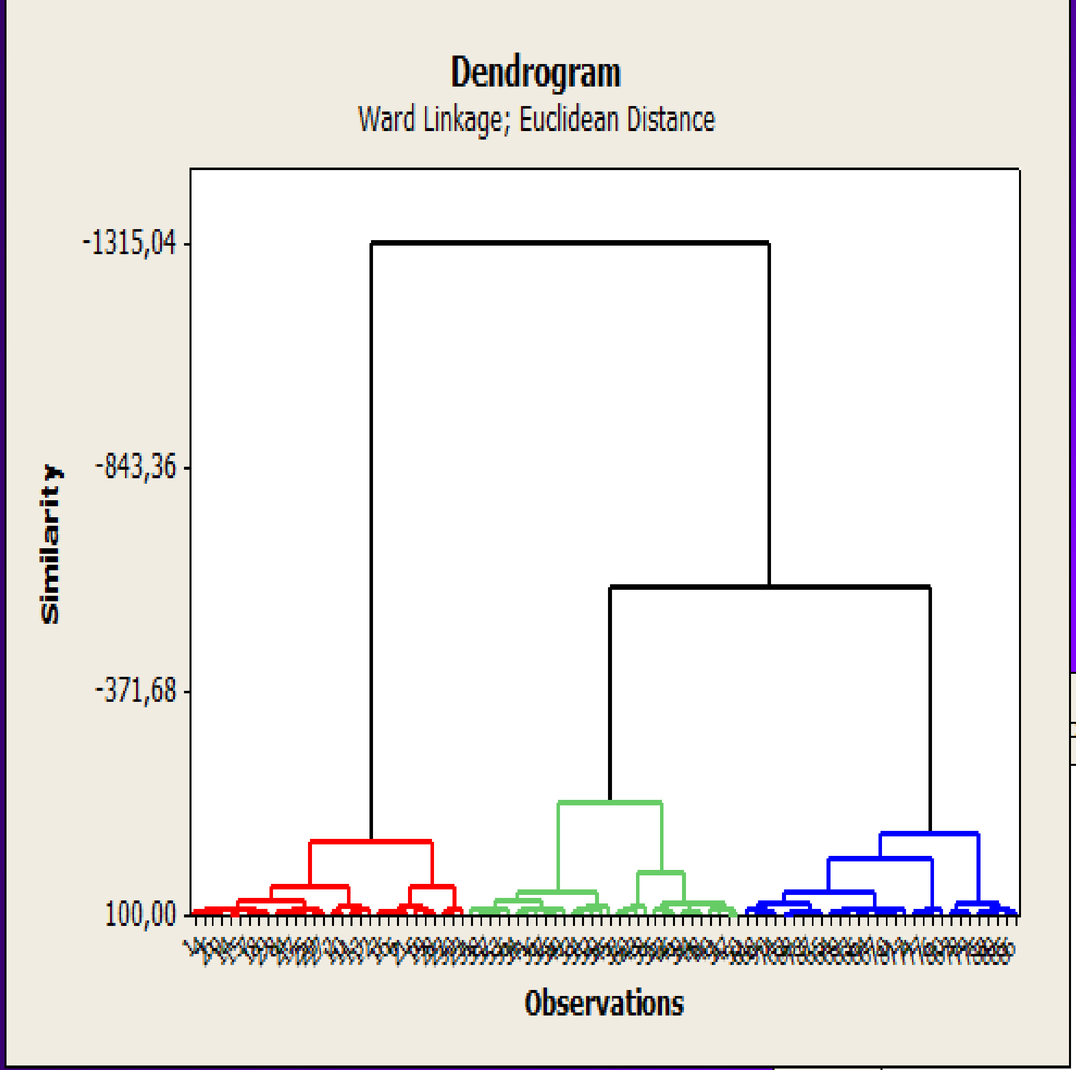
Can exagerate number of Exception-classes

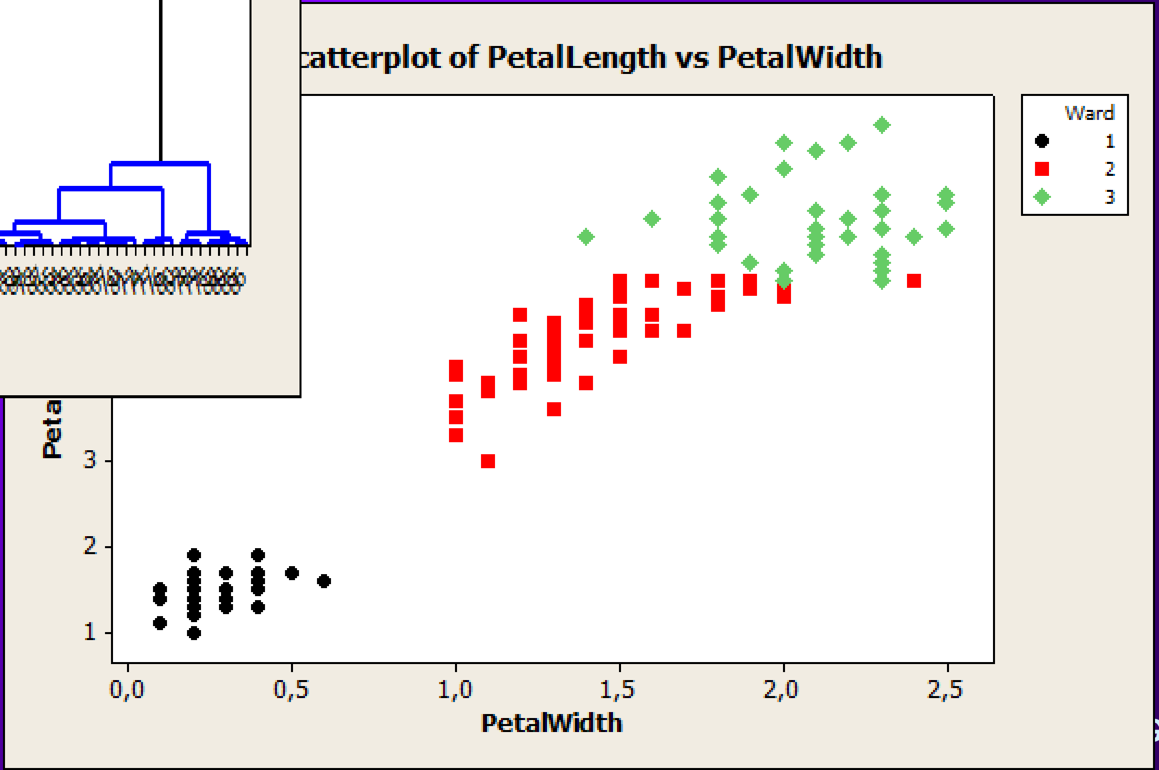
Iris Flowers dataset



Iris Flower Dataset

Ward's method





ibert

Karina Gibert

Dpt. Statistics and Operation Research

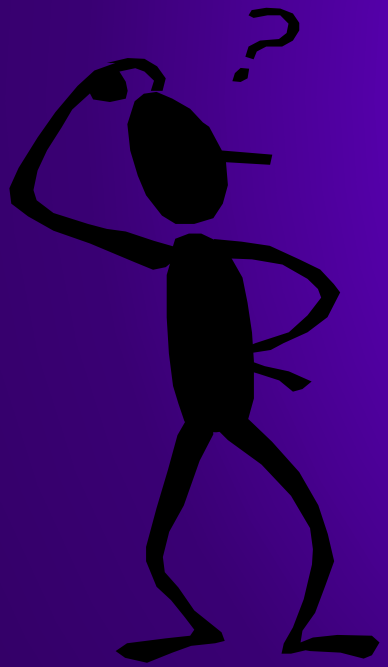
Knowledge Engineering and Machine Learning Research group at Intelligent Data Science and Artificial Intelligence Research Center

Research Institute of Sience and Technology of Sustainability

Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu

www.eio.upc.edu/en/homepages/karina



Are there any questions?...