



# Density-based clustering: (Hyper-)Graphs & Percolation

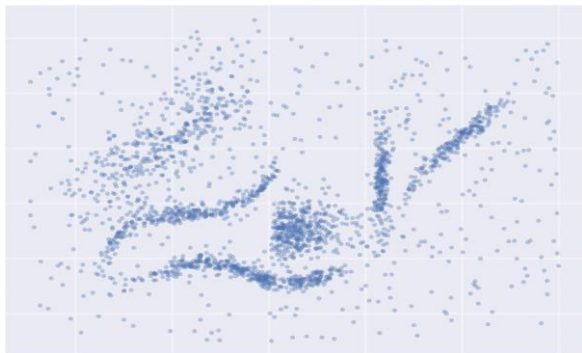
Louis Hauseux, Konstantin Avrachenkov & Josiane Zerubia  
*INRIA d'Université Côte d'Azur*

# OVERVIEW

- I – Introduction: Clustering**
- II – The model: Statistician point of view**
- III – Classical algorithms: Single-Linkage**
- IV – The heart phenomenon: Percolation**
- V – Benefits of Hypergraphs**
- VI – Experiments – Olive oil dataset**

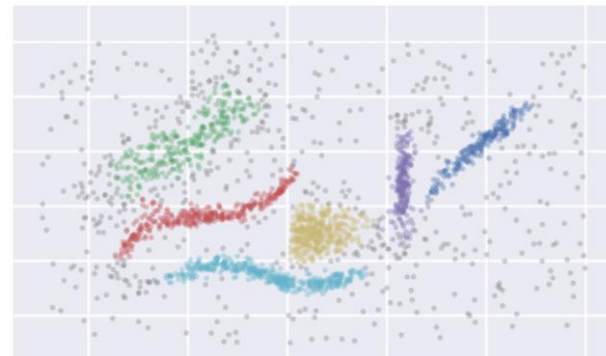
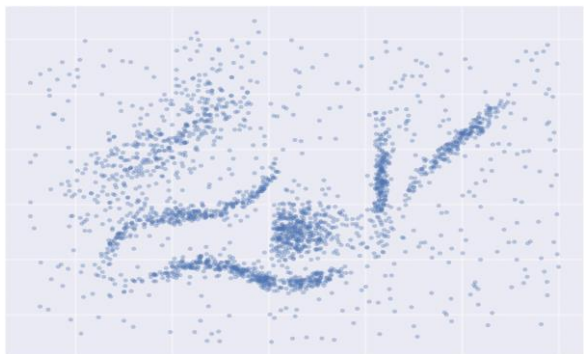
# I – Introduction

## Clustering



# I – Introduction

## Clustering



Hierarchical Density-  
Based SCAN Algorithm \*

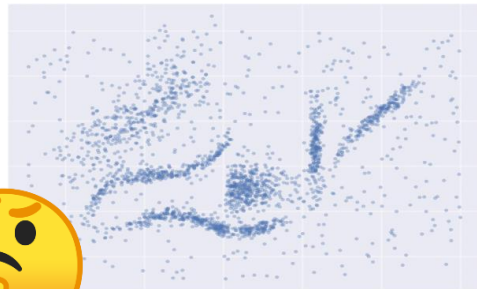
\* McInnes and Healy “Accelerated hierarchical density based clustering” (2017)

## II – The model

### Our approach: Statistician point of view

What is the underlying structure (= the clusters) ?

→ The density  $f$  of point-generation



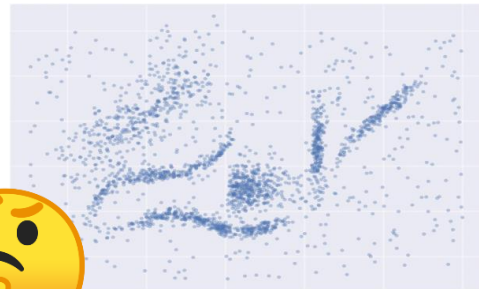
Suppose that all points are plotted IID w.r.t. to  $f$

## II – The model

### Our approach: Statistician point of view

What is the underlying structure (= the clusters) ?

→ The density  $f$  of point-generation



Goal: identify the « **High-Density Clusters** »\*

\*

J. Hartigan. *Clustering Algorithms* (1975)

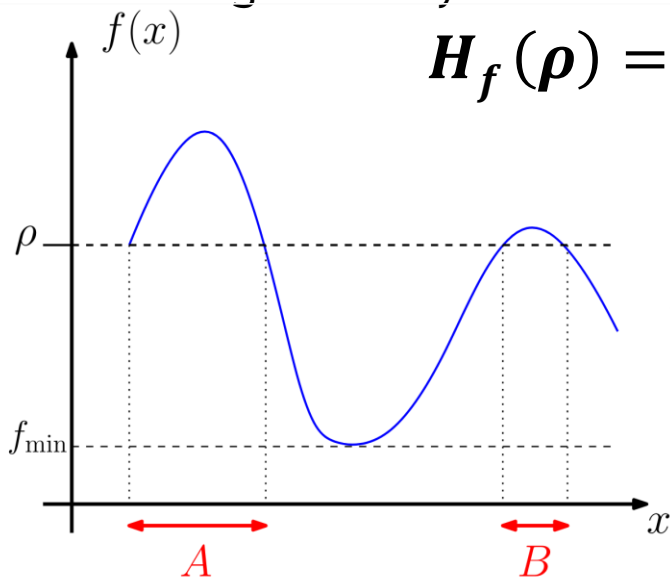
# II – The model

## High density clusters

High-Density Clusters  $H_f(\rho)$  of level  $\rho$  : the connected components of:

$$H_f(\rho) = \{x \in \mathbb{R}^d \mid f(x) \geq \rho\} \subseteq \mathbb{R}^d$$

Important !



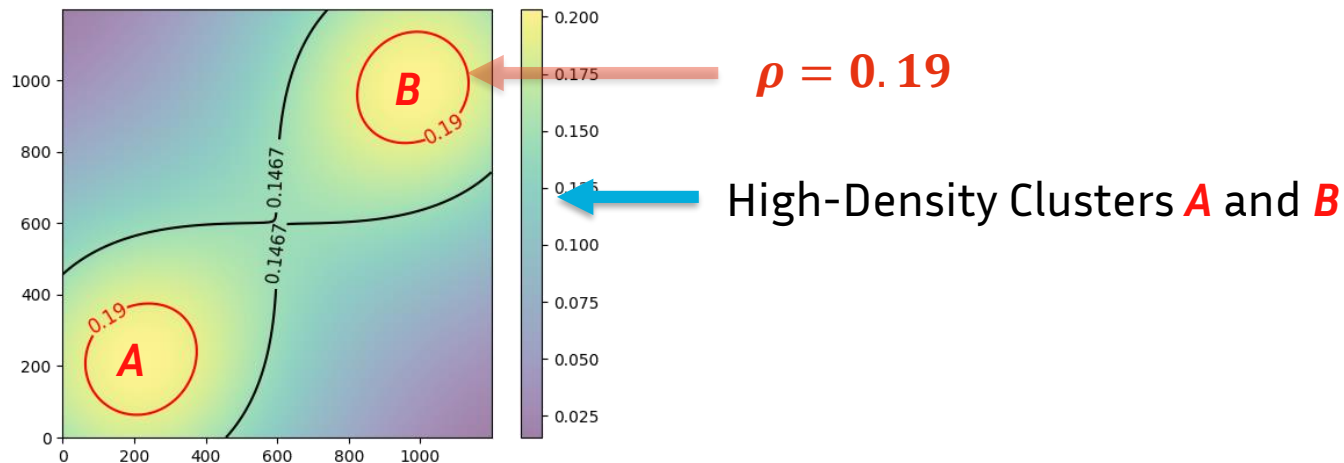
High-Density Clusters  $A$  and  $B$

# II – The model

## High density clusters

High-Density Clusters  $H_f(\rho)$  of level  $\rho$  : the connected components of:

$$H_f(\rho) = \{x \in \mathbb{R}^d \mid f(x) \geq \rho\} \subseteq \mathbb{R}^d$$





## II – The model... and the problem

First solution: Computing an estimator  $\hat{f}$  of  $f$

$$H_f(\rho) \approx H_{\hat{f}}(\rho)$$



Problem: discretizing space is exp. in  $\dim(\mathbb{R}^d)$

## II – The model... and the problem



**Problem: Computing  $\hat{f}$  is costly in  $\mathbb{R}^d$  with large  $d$**



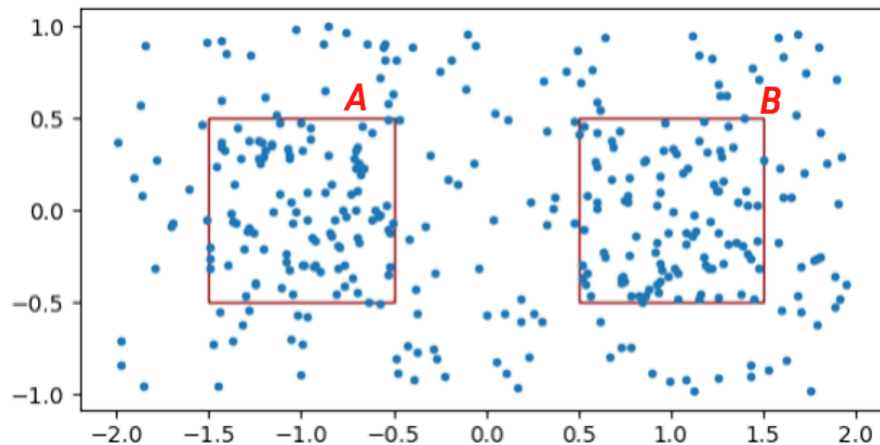
Typically: Geometric Graphs



**Construct Graphs on the data**

**➔ Statistical Analysis on Networks**

# III – Classical algorithms



Old but efficient:

- Single-Linkage  
= Geometric Graphs

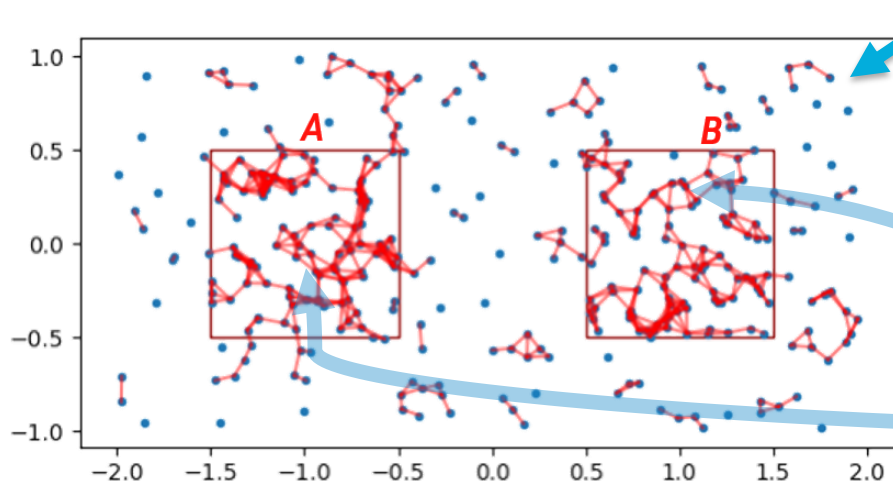
Recent and state-of-the-art:

- (H)DBSCAN
- Persistable\*

\* A. Rolle and L. Scoccola: “Stable and consistent density-based clustering” (2023)

# III – Single-Linkage ~ Geometric Graphs

Geometric graph with radius  
 $r \approx 0.14$

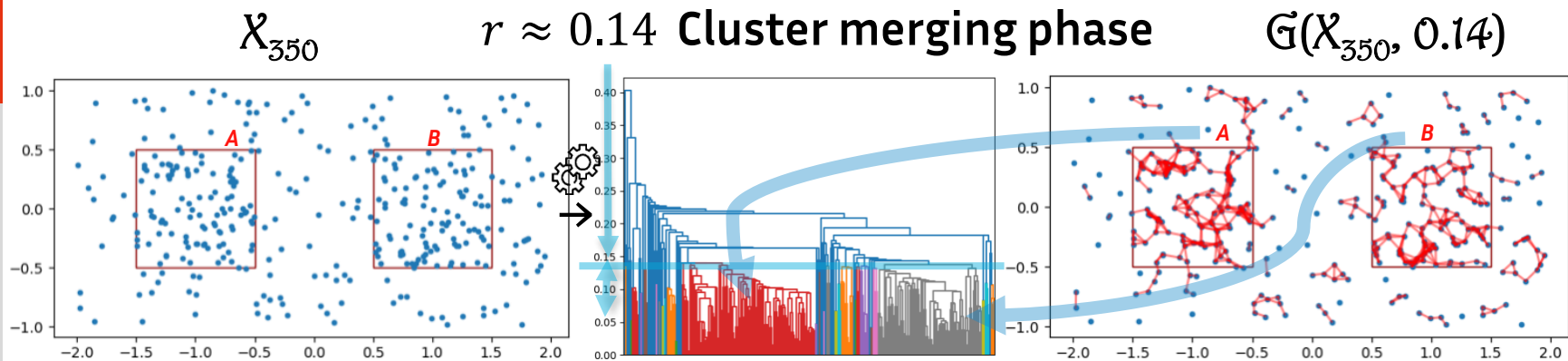


We almost recover the  
High-Density Cluster

**A**  
and **B**

# III – Single-Linkage ~ Geometric Graphs

Phase transition \*



\*

G. Parisi. *Statistical Field Theory* (1988)

## IV – Why does it work?



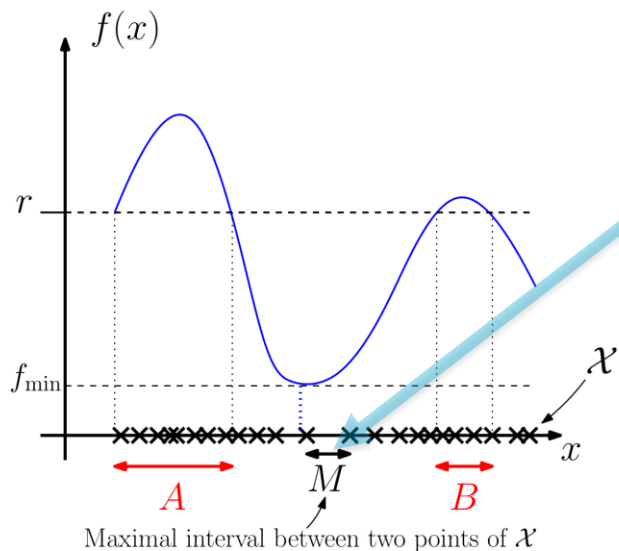
**Geometric Graphs ~ Empirical high-density clusters of  
1-Nearest Neighbor density estimator:**

$$\hat{f}_{1\text{-NN}}(y) = \frac{1}{R_y^d} \quad \text{where } R_y = \min_{x \in \mathcal{X}_n} \|x - y\|$$

High-density clusters  $\longleftrightarrow$  Connected components of  $\mathbb{G}$

## IV – Why does it work in $\mathbb{R}$ ?

➔ Single-Linkage is **consistent** in  $\mathbb{R}$  !!!



The « cut » between the two clusters, containing **A** or **B**

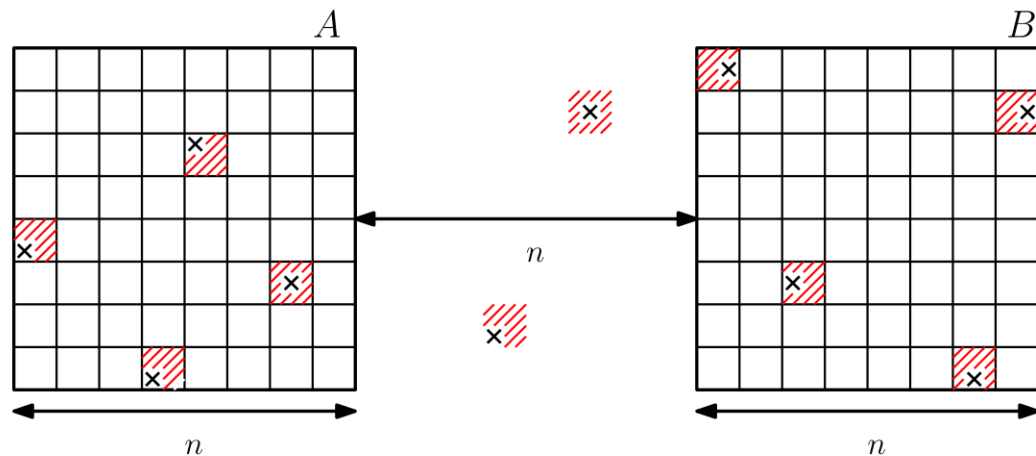
➔ But Single-Linkage is not **consistent** in  $\mathbb{R}^d$  for  $d \geq 2$  ...



## IV – Why does it *not* work in $\mathbb{R}^d$ ? ( $d \geq 2$ )

→ Single-Linkage is not consistent in  $\mathbb{R}^d$  for  $d \geq 2$  ... !!!

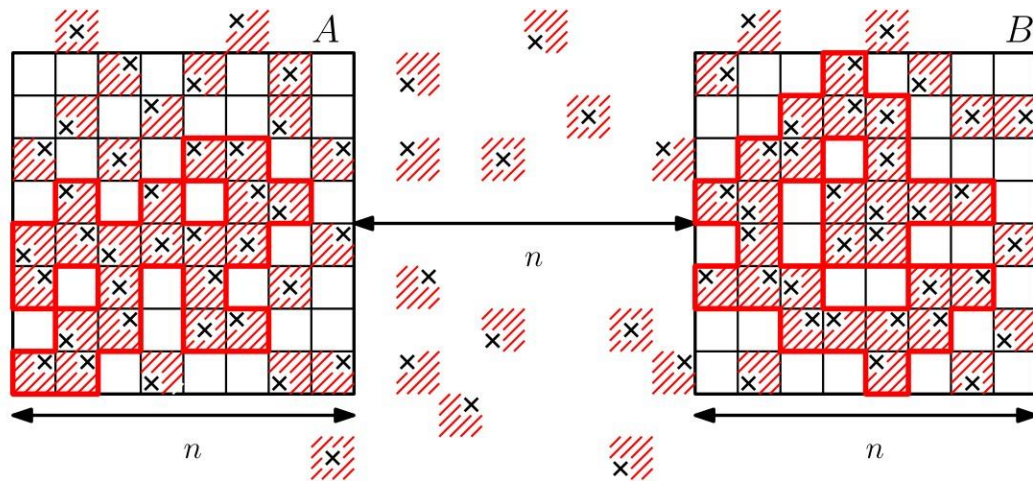
Discrete Site-Percolation on  $\mathbb{Z}^2$





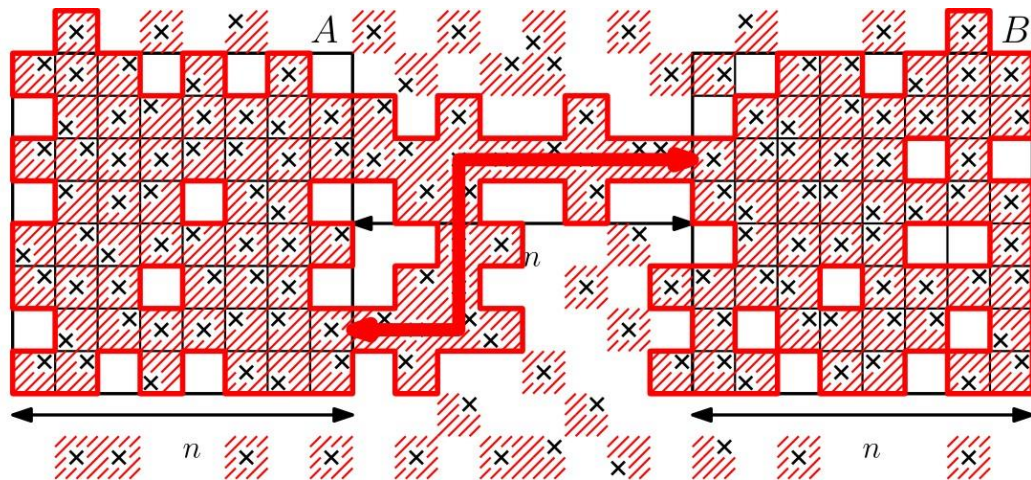
# IV – Why does it *not* work in $\mathbb{R}^d$ ? ( $d \geq 2$ )

Percolation occurs on  $A$  and  $B$



## IV – Why does it *not* work in $\mathbb{R}^d$ ? ( $d \geq 2$ )

Percolation occurs everywhere...



➔ The two clusters merge before having recovered **A** and **B** entirely...

## IV – The Percolation \*

Model : the  $x_1, \dots, x_n$  IID plotted uniformly on  $[0 ; 1]^2$ .

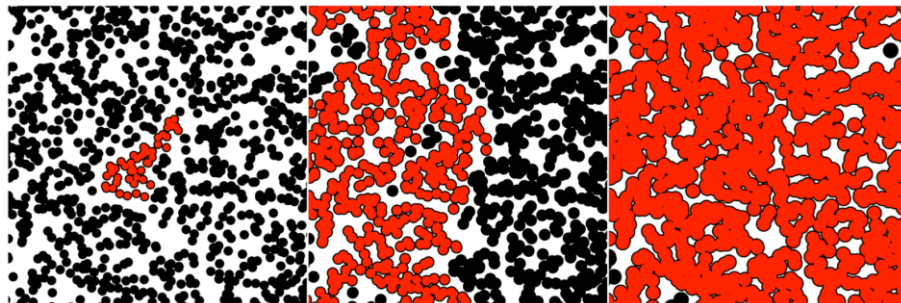
→ Poisson Point Process  $\lambda \leftarrow 1$

**Percolation** = giant component

→ **Very fast phenomenon**

Two sub-cases : there exists a critical value  $r_c \approx 1.2$ :

- If  $r < r_c$ , no great component (of size  $\Theta(\log(n))$ )
- If  $r > r_c$ , **one giant component** (of size  $\Theta(n)$ )



**Geometric Graphs / greatest component**  
 $r < r_c$        $r = r_c$        $r > r_c$

\*

M. Penrose. *Random Geometric Graphs* (2003)

## IV – Why does it *not* work in $\mathbb{R}^d$ ? ( $d \geq 2$ )

- Single-Linkage **works** in  $\mathbb{R}^1$  ... because there is **no** percolation
- S.-L. **does not work** perfectly in  $\mathbb{R}^d$  ... because **there is** percolation



Defeat on this battle... but war is not lost!

## IV – Percolation: a conclusion

- Single-Linkage **works** in  $\mathbb{R}^1$  ... because there is **no** percolation
- S.-L. **does not work** perfectly in  $\mathbb{R}^d$  ... because **there is** percolation

Compare performance of clustering algo.

Studying the percolation phenomenon:

- A **fraction** of high-density clusters is recoverable...
- Which fraction? Define the **Percolation rate** \*, \*\*




\*

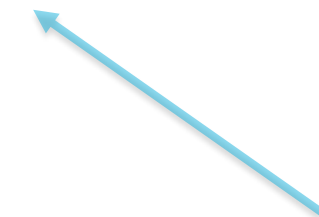
LH + KA + JZ, "Graph Based Approach for Galaxy Filament Extraction". Complex Networks (2023)

\*\*

Work to be submitted to SIAM Journal on Mathematics of Data Science

# V – Hypergraphs... for $K$ -high-density clusters

HDC of  $1\text{-NN}$   Graph Connected Components



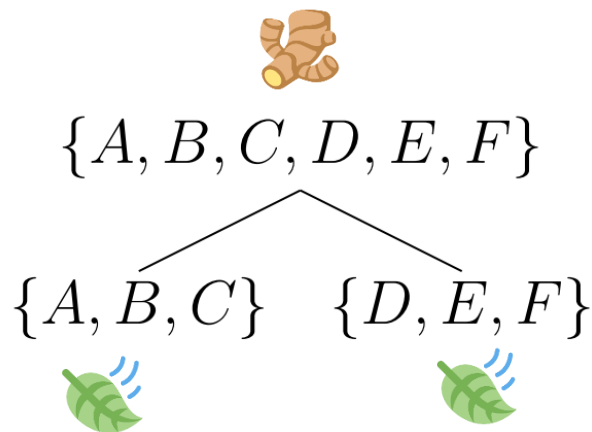
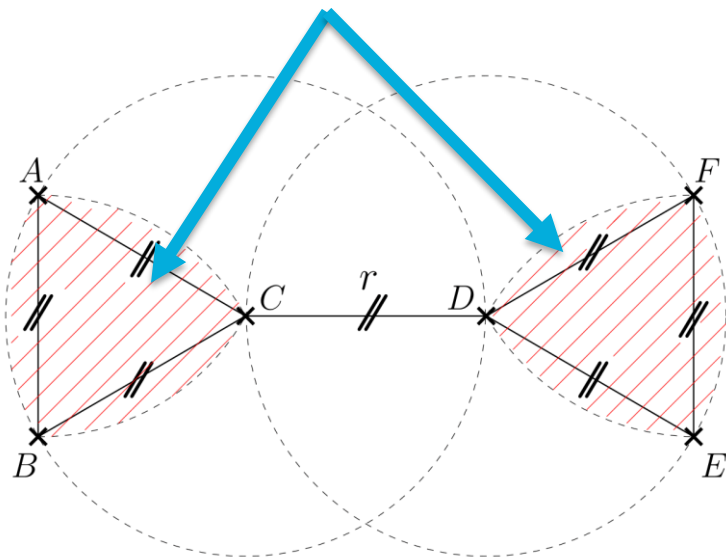
We would like  $K\text{-NN}$  to gain in robustness

# V – Hypergraphs... for $K$ -high-density clusters

➔ Weakness of RSL (or DBSCAN)

The 3-high-density clusters of level  $r$

Dendrogram of 3-NN density estimator



# $V - K$ -high-density clusters $\sim$ Hypergraphs

HDC of  $1\text{-}NN$   $\longleftrightarrow$  Graph Connected Components



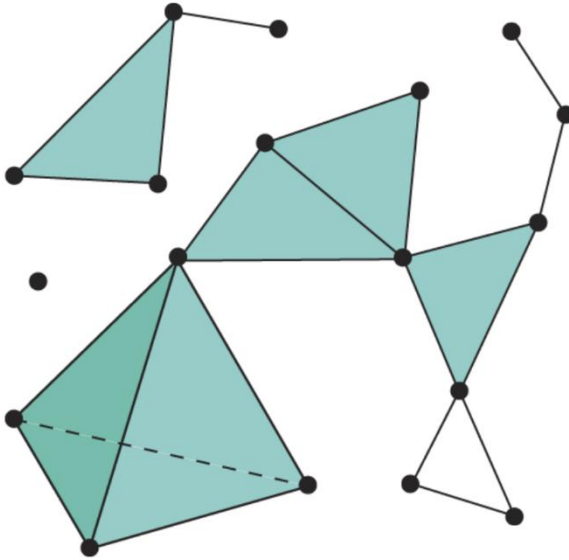
HDC of  $K\text{-}NN$   $\longleftrightarrow$  Hypergraph Connect. Comp. \*

\*

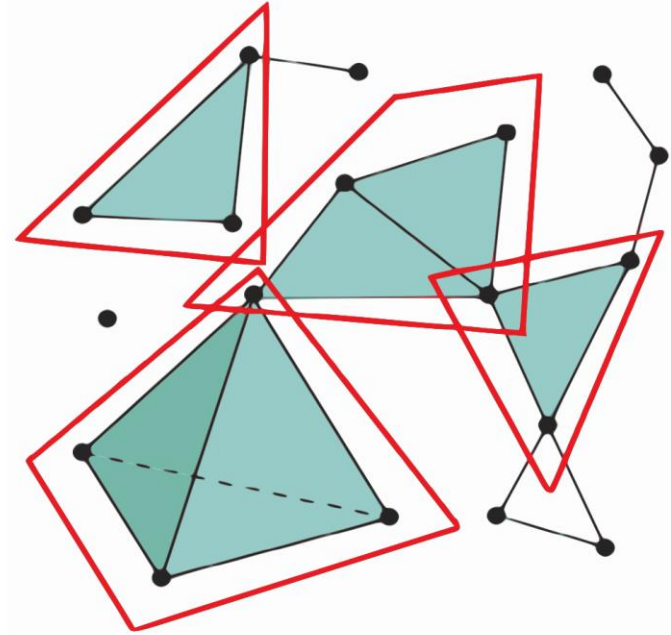
Work submitted (and accepted) to EUSIPCO



# V – Hypergraphs, Polyhedra



A hypergraph



**Polyhedra** on a hypergraph

# VI – Experiments – Olive Italian Oil Dataset \*

- \* M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia: **“Classification of olive oils from their fatty acid composition” (1983)**
- \*\* S. Scaldelai, L. Mاتيoli and M. Kleina: (2022)  
**“Multiclusterkde: A new algorithm for clustering based on multivariate kernel density estimation”**
- \*\*\* A. Rolle and L. Scoccola: **“Stable and consistent density-based clustering” (2023)**



Macro-area	Region
South	1 – North Apulia
	2 – Calabria
	3 – South Apulia
	4 – Sicilia
Sardinia	5 – Inland Sardinia
	6 – Coast Sardinia
Centro-settentrionale	7 – East Liguria
	8 – West Liguria
	9 – Umbria

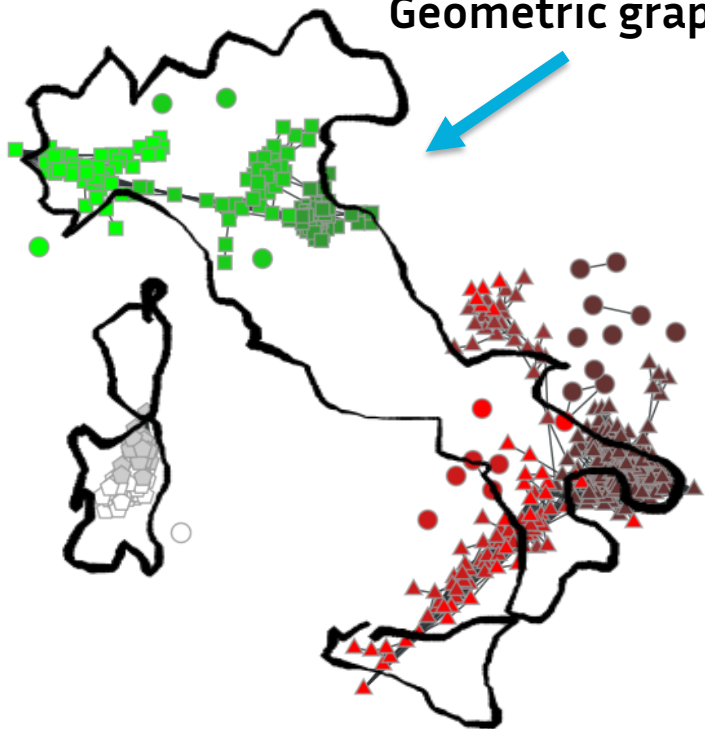
**572 samples of oil composition  
(vectors in  $\mathbb{R}^8$ )**

Fatty acids: Palmitic, Palmitoleic,  
Stearic, Oleic, Linoleic, Linolenic,  
Arachidic, Eicosenoic

**Can we recover the  
geographic clusters given  
only the fatty acids ?**

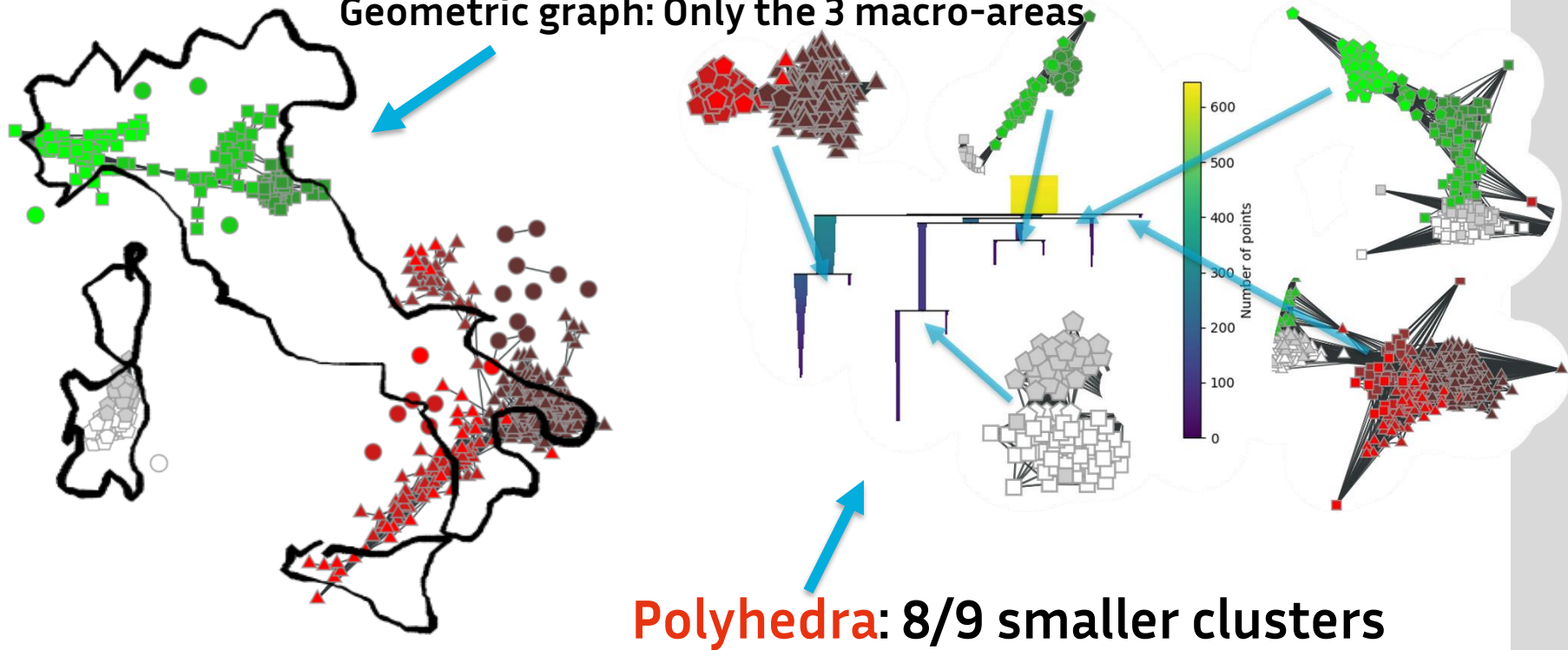
# VI – Experiments – Olive Italian Oil Dataset

Geometric graph: Only the 3 macro-areas



# VII – Experiments – Olive Italian Oil Dataset

Geometric graph: Only the 3 macro-areas



# VII – Experiments – Olive Italian Oil Dataset



Confusion Matrix: Persistable vs Polyhedra

Pers./Ours	1	2	3	4	5	6	7	8	9	Miss.
N. Apulia	12/14	0/0	0/0	-	0/0	0/0	0/0	0/0	0/0	13/11
Calabria	0/0	7/28	1/1	-	0/0	0/0	0/0	0/0	0/0	48/27
S. Apulia	0/0	0/0	100/167	-	0/0	0/0	0/0	0/0	0/0	106/39
Sicilia	3/5	0/1	0/2	-	0/0	0/0	0/0	0/0	0/0	33/28
Inl. Sard.	0/0	0/0	0/0	-	51/52	0/0	0/0	0/0	0/0	14/13
Coast S.	0/0	0/0	0/0	-	0/2	19/27	0/0	0/0	0/0	14/4
E. Ligur.	0/0	0/0	0/0	-	0/0	0/0	14/20	1/3	0/0	35/27
W. Ligur.	0/0	0/0	0/0	-	0/0	0/0	0/0	29/41	0/0	21/9
Umbria	0/0	0/0	0/0	-	0/0	0/0	0/6	0/0	42/25	9/20

# Bibliography

- John A. Hartigan. *Clustering Algorithms* (1975). *John Wiley & Sons*.
- M. Forina, C. Armanino, S. Lanteri, and E. Tiscornia: « Classification of olive oils from their fatty acid composition » (1983). *IUFoST Symposium*.
- M. Penrose. *Random Geometric Graphs* (2003). *Oxford Studies in Probability*.
- R. Stoica, Vicent J. Martinez, Jorge Mateu & Enn Saar. « Detection of cosmic filaments using the Candy model » (2005). *Astronomy & Astrophysics*.
- K. Chaudhuri and S. Dasgupta. « Rates of convergence for the cluster tree » (2010). *NIPS*.
- L. McInnes and J. Healy: « Accelerated hierarchical density based clustering » (2017). *ICDMW*.
- J.-D. Boissonnat, F. Chazal and M. Yvinec. *Geometric and Topological Inference* (2018). *Cambridge University Press*.
- S. Scaldelai, L. Matioli and M. Kleina: « Multiclusterkde: A new algorithm for clustering based on multivariate kernel density estimation » (2022). *J. Appl. Stat.*
- L. Hauseux & K. Avrachenkov & J. Zerubia. « Graph Based Approach for Galaxy Filament Extraction » (2023). *Intern. Conf. Of Complex Networks, Menton and HAL*.
- A. Rolle and L. Scoccola: « Stable and consistent density-based clustering » (2023). *Arxiv*.