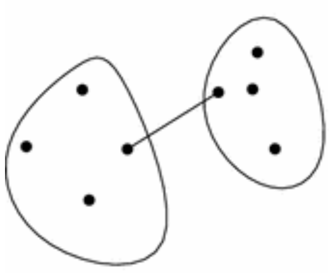
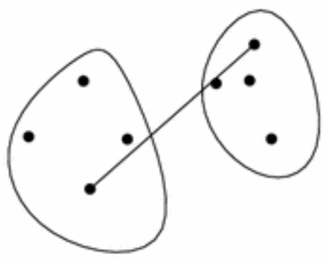
Strategies to Define a Good Linkage Distance

Let D_{x_i,x_j} the distance (e.g., euclidean) between any two elements of X.

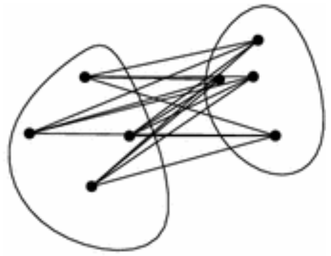
We need to define a subset distance Δ_{X_i,X_j} between any two subsets.



Single Linkage. $\Delta_{X_i,X_j} = min_{x_i \in X_i, x_j \in X_j} D_{x_i,x_j}$



Complete Linkage. $\Delta_{X_i,X_j} = max_{x_i \in X_i, x_j \in X_j} D_{x_i,x_j}$

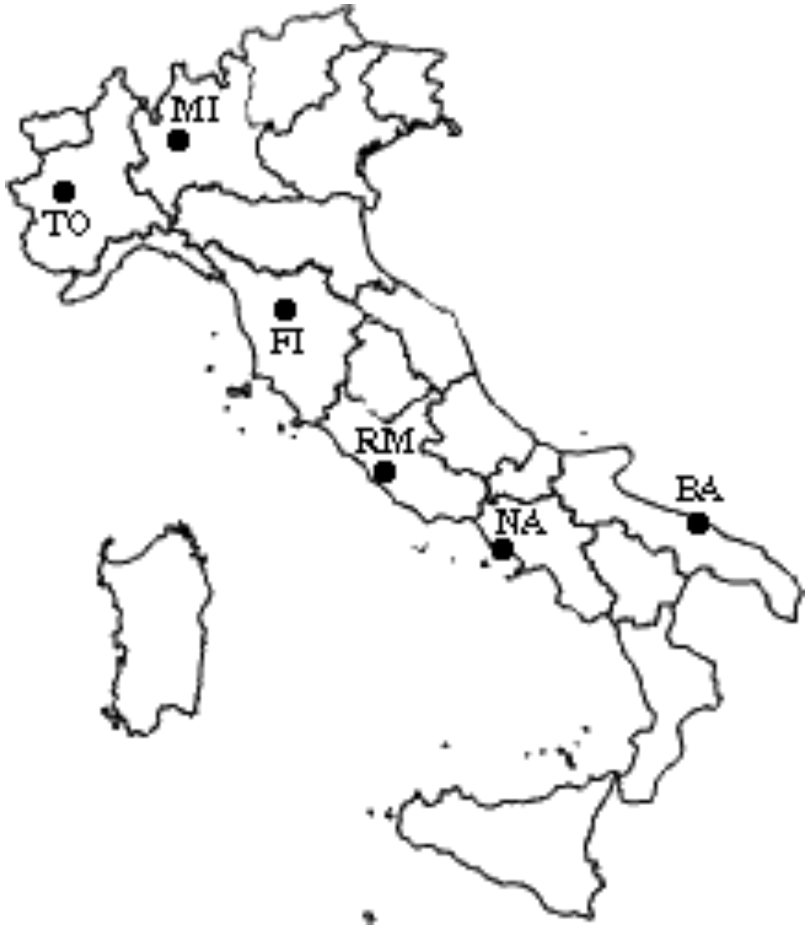


Group Average Linkage. $\Delta_{X_i,X_j} = \dfrac{1}{|X_i||X_j|} \sum_{x_i \in X_i} \sum_{x_j \in X_j} D_{x_i,x_j}$

Nielsen, F. (2016). Hierarchical clustering. In Introduction to HPC with MPI for Data Science (pp. 195-211). Springer, Cham.

Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

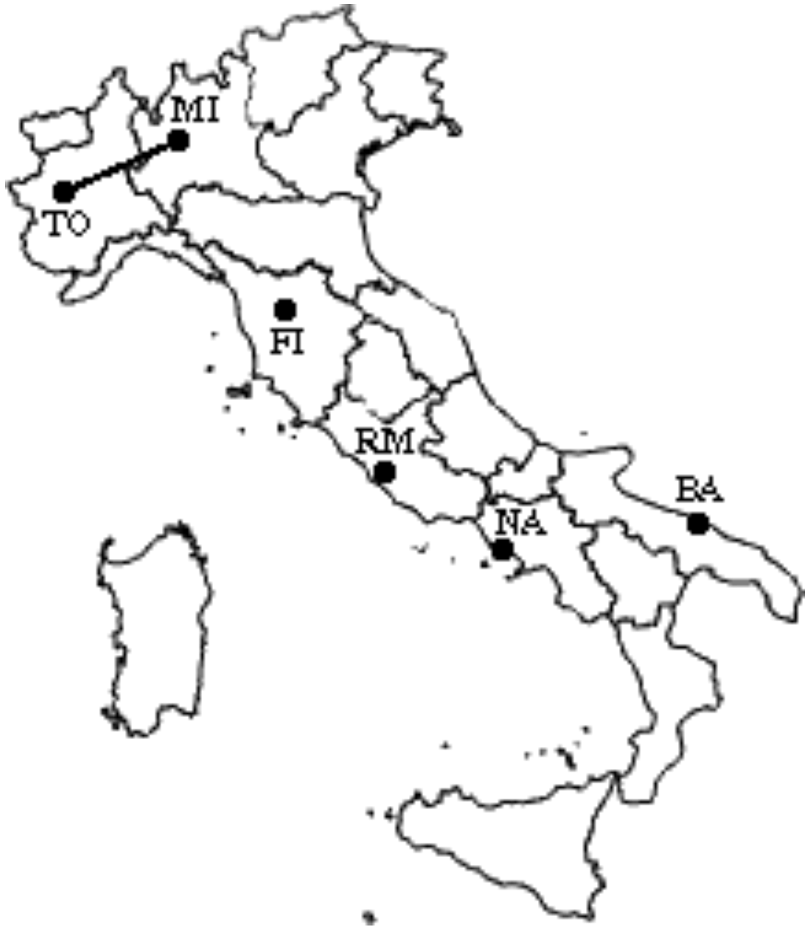
	BA	FI	MI	NA	RM	TO
BA	-	662	877	255	412	996
FI	662	-	295	468	268	400
MI	877	295	-	754	564	138
NA	255	468	754	-	219	869
RM	412	268	564	219	-	669
TO	996	400	138	869	669	-



JADS Jheronimus Academy of Data Science

Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

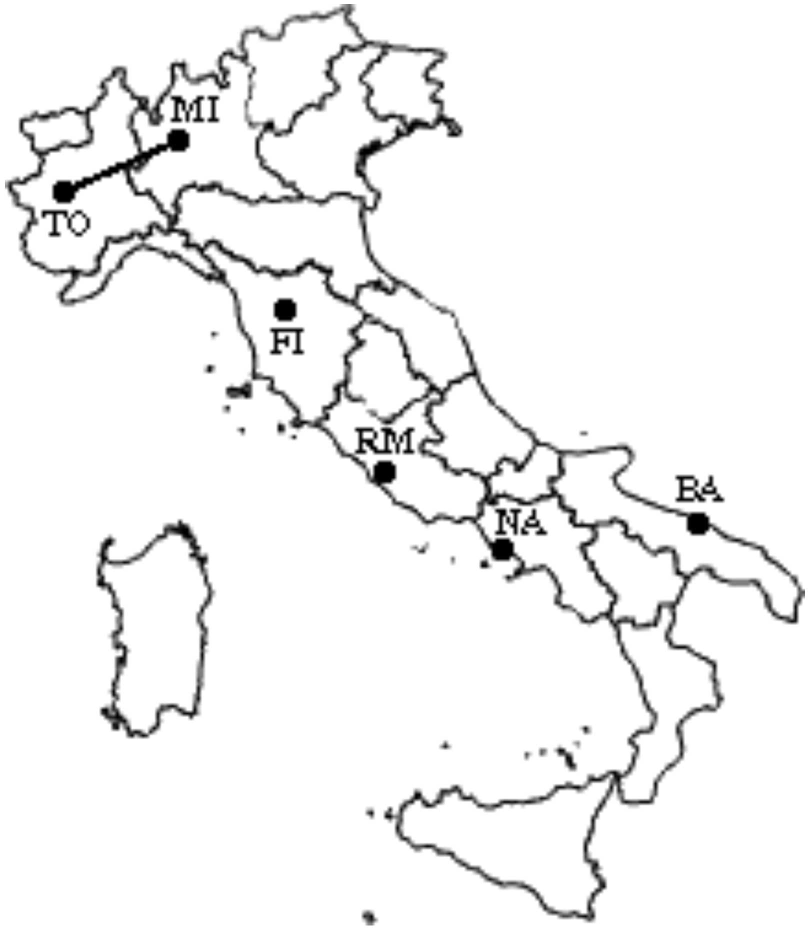
	BA	FI	MI	NA	RM	TO
BA	-	662	877	255	412	996
FI	662	-	295	468	268	400
MI	877	295	-	754	564	138
NA	255	468	754	-	219	869
RM	412	268	564	219	-	669
TO	996	400	138	869	669	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

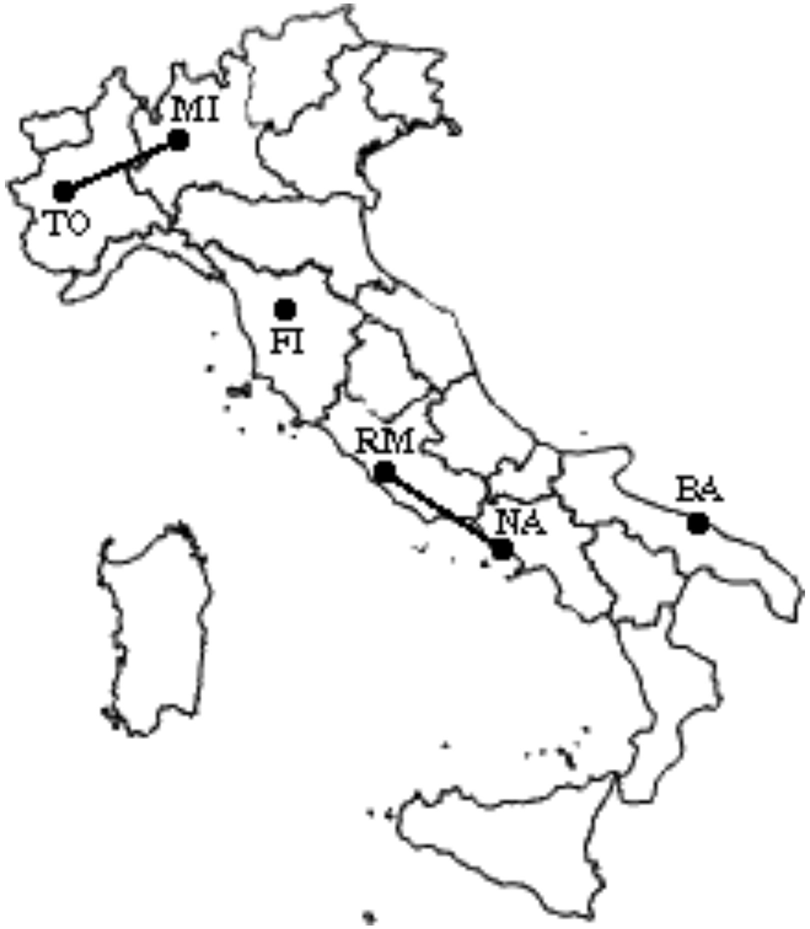
	BA	FI	MI	NA	RM	TO
BA	-	662	877	255	412	996
FI	662	-	295	468	268	400
MI	877	295	-	754	564	138
NA	255	468	754	-	219	869
RM	412	268	564	219	-	669
TO	996	400	138	869	669	-

	BA	FI	MI/TO	NA	RM
BA	-	662	{877, 996}	255	412
FI	662	-	{295, 400}	468	268
MI/TO	{877, 996}	{295, 400}	-	{754, 869}	{564, 669}
NA	255	468	{754, 869}	-	219
RM	412	268	{564, 669}	219	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

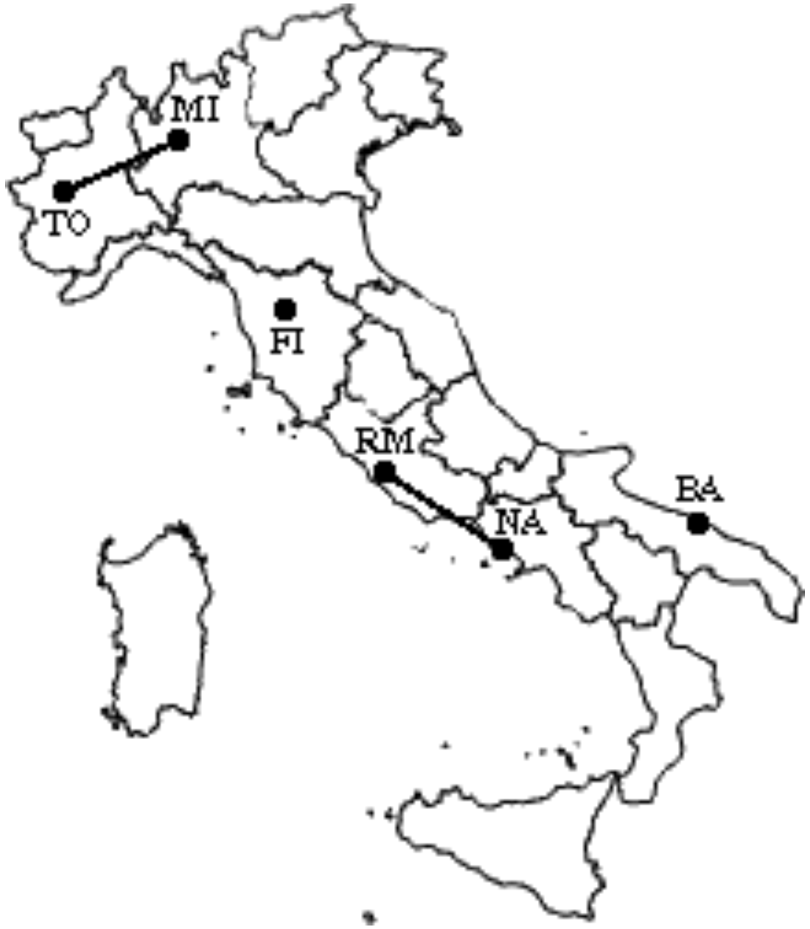
	BA	FI	MI/TO	NA	RM
BA	-	662	{877, 996}	255	412
FI	662	-	{295, 400}	468	268
MI/TO	{877, 996}	{295, 400}	-	{754, 869}	{564, 669}
NA	255	468	{754, 869}	-	219
RM	412	268	{564, 669}	219	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

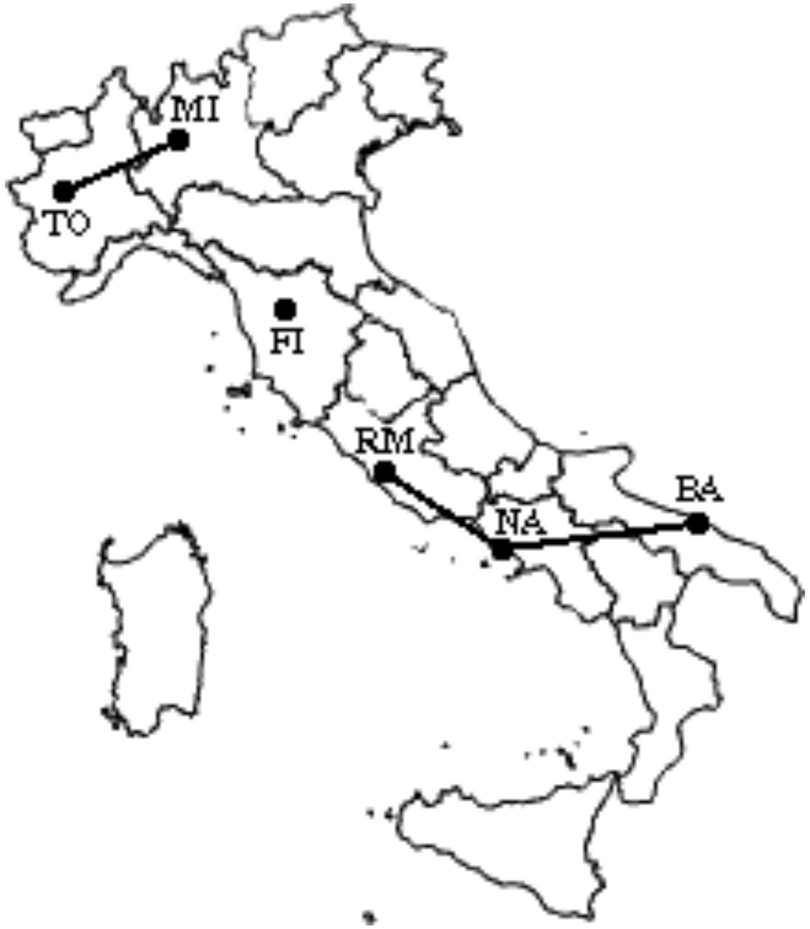
	BA	FI	MI/TO	NA	RM
BA	-	662	{877, 996}	255	412
FI	662	-	{295, 400}	468	268
MI/TO	{877, 996}	{295, 400}	-	{754, 869}	{564, 669}
NA	255	468	{754, 869}	-	219
RM	412	268	{564, 669}	219	-

	BA	FI	MI/TO	NA/RM
BA	-	662	{877, 996}	{255, 412}
FI	662	-	{295, 400}	{268, 468}
MI/TO	{877, 996}	{295, 400}	-	{564, 669, 754, 869}
NA/RM	{255, 412}	{268, 468}	{564, 669, 754, 869}	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

	BA	FI	MI/TO	NA/RM
BA	-	662	{877, 996}	{255, 412}
FI	662	-	{295, 400}	{268, 468}
MI/TO	{877, 996}	{295, 400}	-	{564, 669, 754, 869}
NA/RM	{255, 412}	{268, 468}	{564, 669, 754, 869}	-

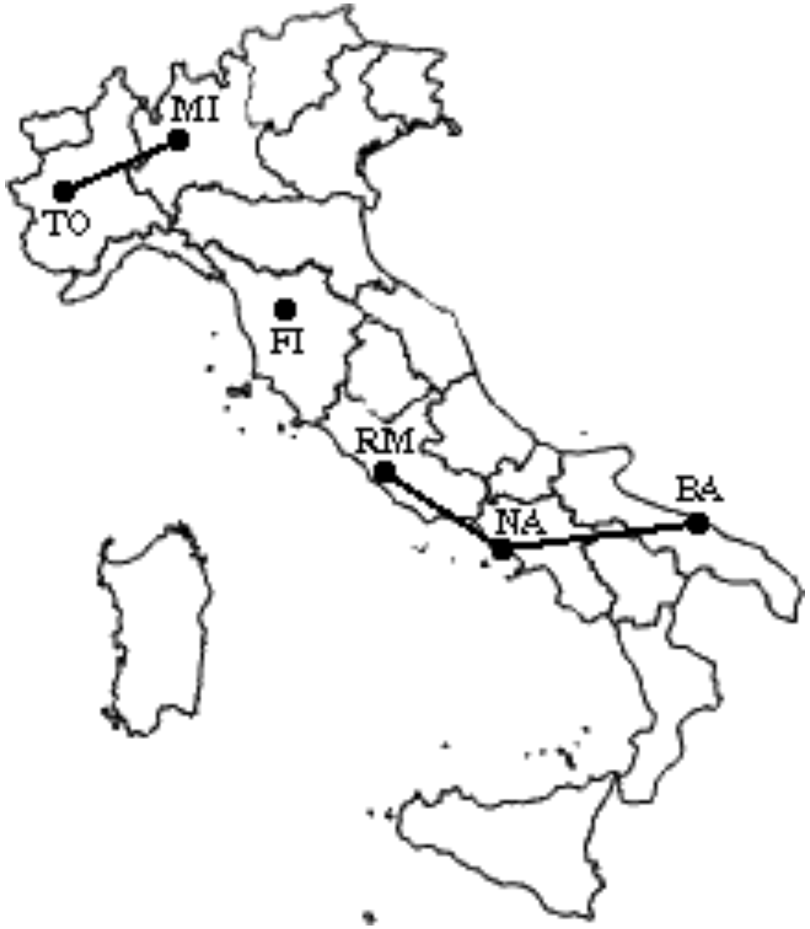


JADS Jheronimus Academy of Data Science

Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

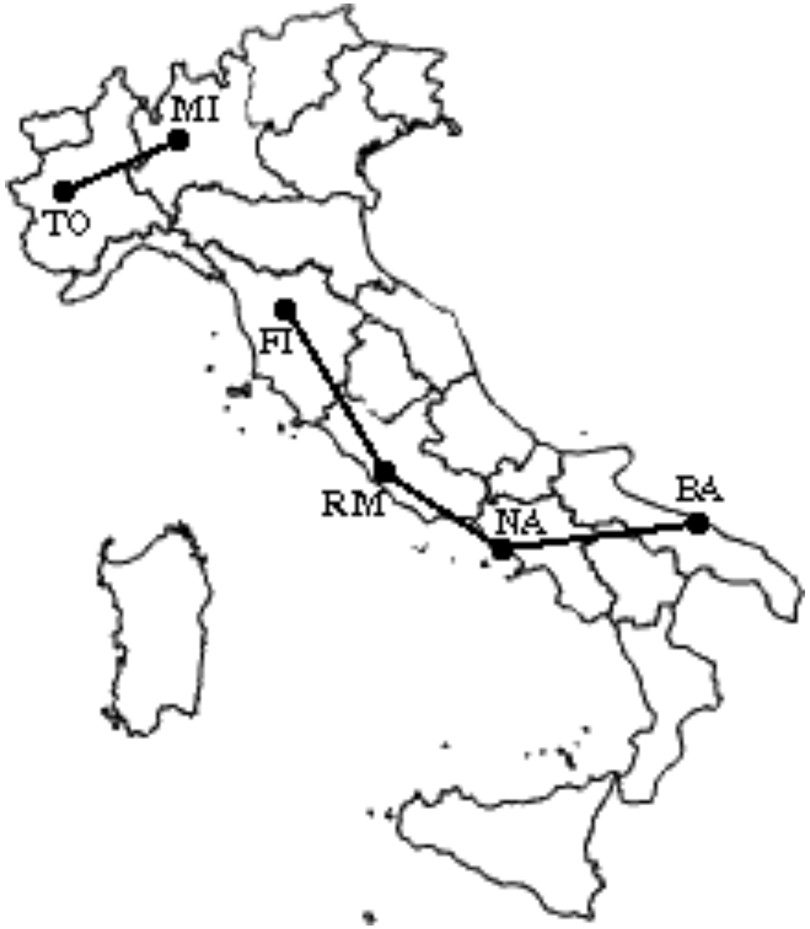
	BA	FI	MI/TO	NA/RM
BA	-	662	{877, 996}	{255, 412}
FI	662	-	{295, 400}	{268, 468}
MI/TO	{877, 996}	{295, 400}	-	{564, 669, 754, 869}
NA/RM	{255, 412}	{268, 468}	{564, 669, 754, 869}	-

	FI	MI/TO	BA/NA/RM
FI	-	{295, 400}	{268, 468, 662}
MI/TO	{295, 400}	-	{564, 669, 754, 869, 877, 996}
BA/NA/RM	{268, 468, 662}	{564, 669, 754, 869, 877, 996}	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

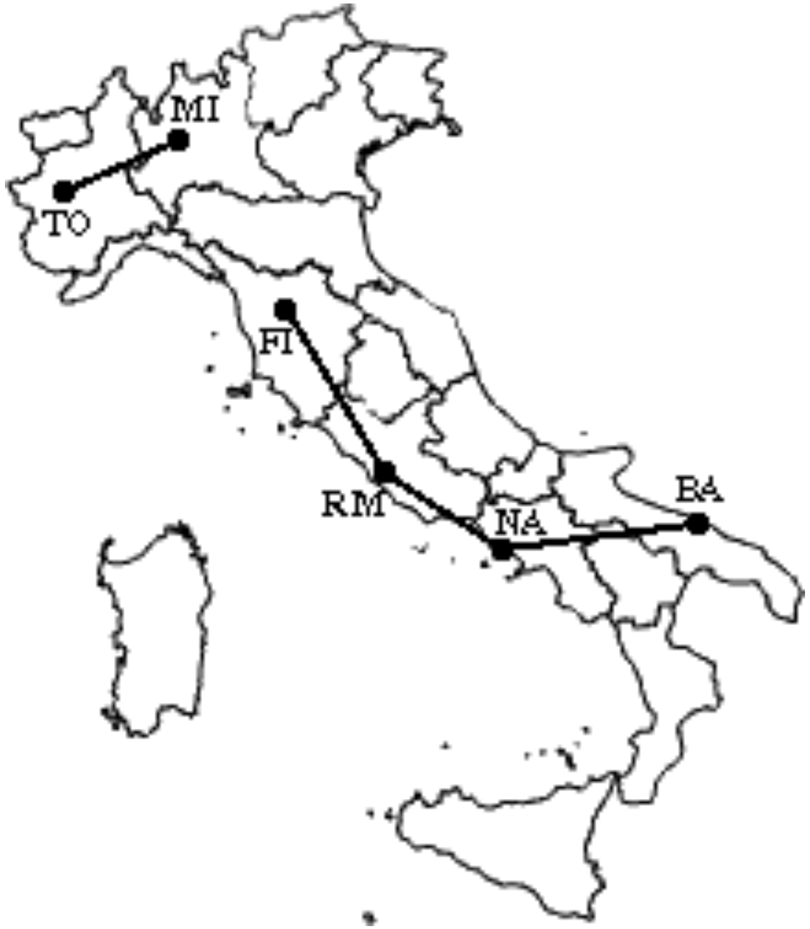
	FI	MI/TO	BA/NA/RM
FI	-	{295, 400}	{268, 468, 662}
MI/TO	{295, 400}	-	{564, 669, 754, 869, 877, 996}
BA/NA/RM	{268, 468, 662}	{564, 669, 754, 869, 877, 996}	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

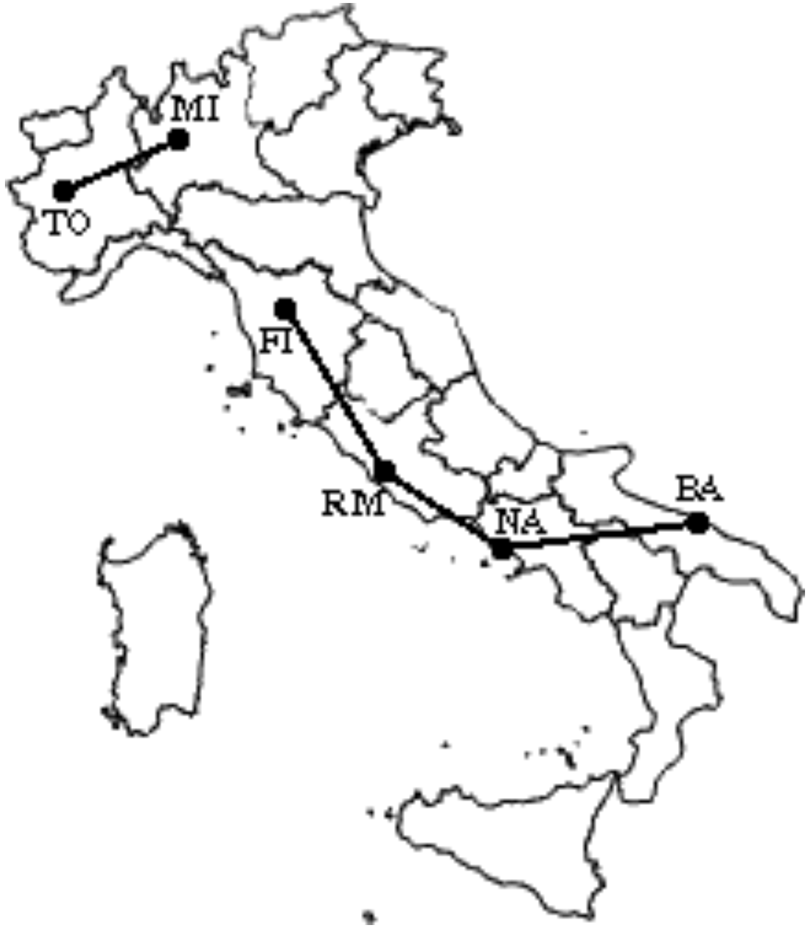
	FI	MI/TO	BA/NA/RM
FI	-	{295, 400}	{268, 468, 662}
MI/TO	{295, 400}	-	{564, 669, 754, 869, 877, 996}
BA/NA/RM	{268, 468, 662}	{564, 669, 754, 869, 877, 996}	-

	MI/TO	BA/FI/NA/RM
MI/TO	-	{295, 400, 564, 669, 754, 869, 877, 996}
BA/FI/NA/RM	{295, 400, 564, 669, 754, 869, 877, 996}	-



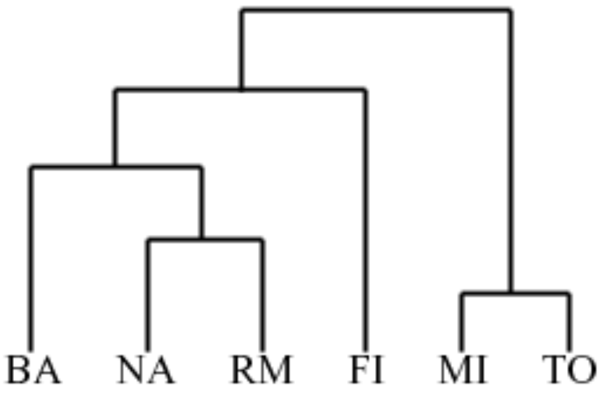
Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

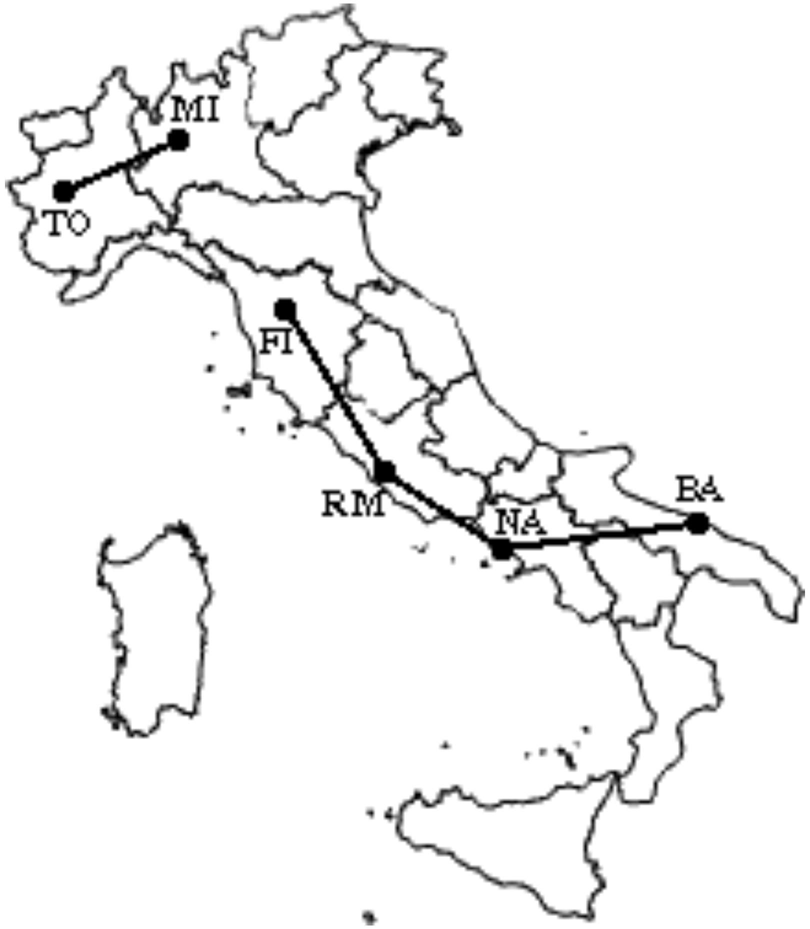
	MI/TO	BA/FI/NA/RM
MI/TO	-	{295, 400, 564, 669, 754, 869, 877, 996}
BA/FI/NA/RM	{295, 400, 564, 669, 754, 869, 877, 996}	-



Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

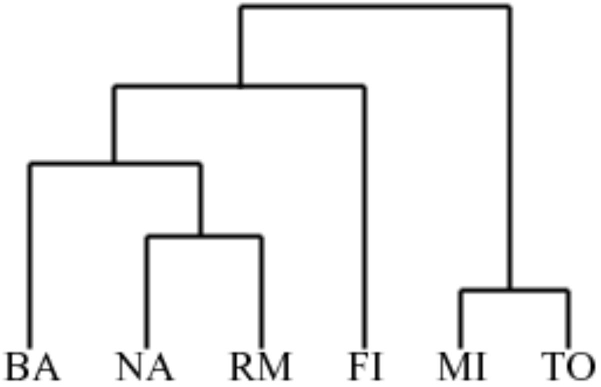
	MI/TO	BA/FI/NA/RM
MI/TO	-	{295, 400, 564, 669, 754, 869, 877, 996}
BA/FI/NA/RM	{295, 400, 564, 669, 754, 869, 877, 996}	-

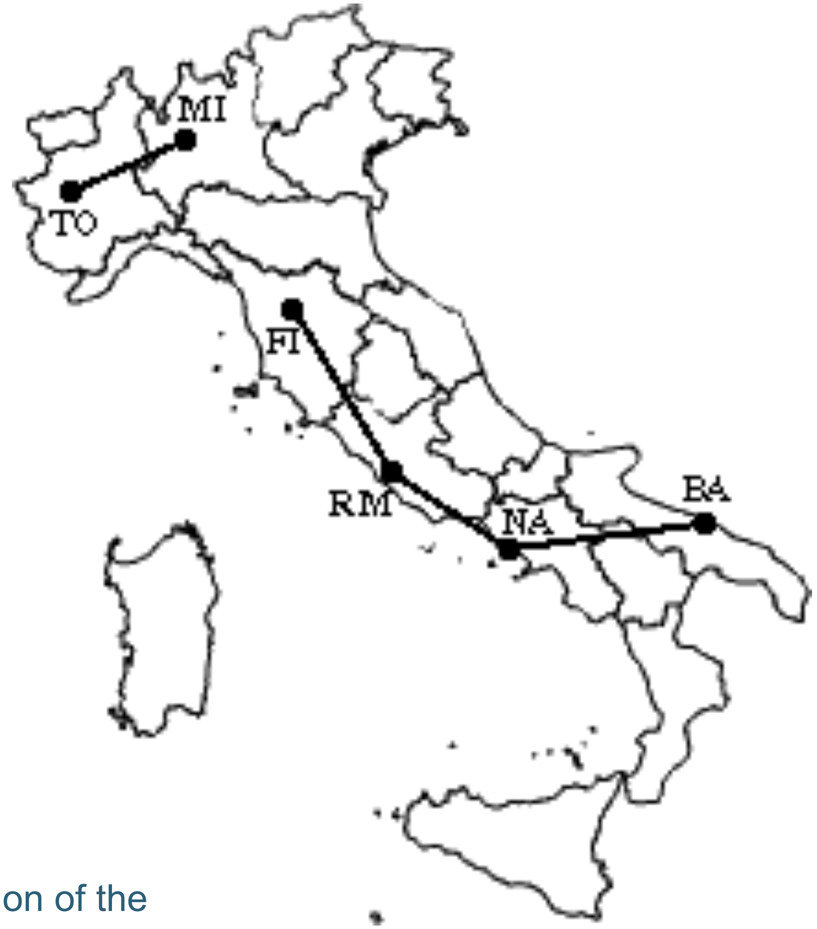




Hierarchical Clustering
Step by Step - Single Linkage (Distanza Minima)

	MI/TO	BA/FI/NA/RM
MI/TO	-	{295, 400, 564, 669, 754, 869, 877, 996}
BA/FI/NA/RM	{295, 400, 564, 669, 754, 869, 877, 996}	-



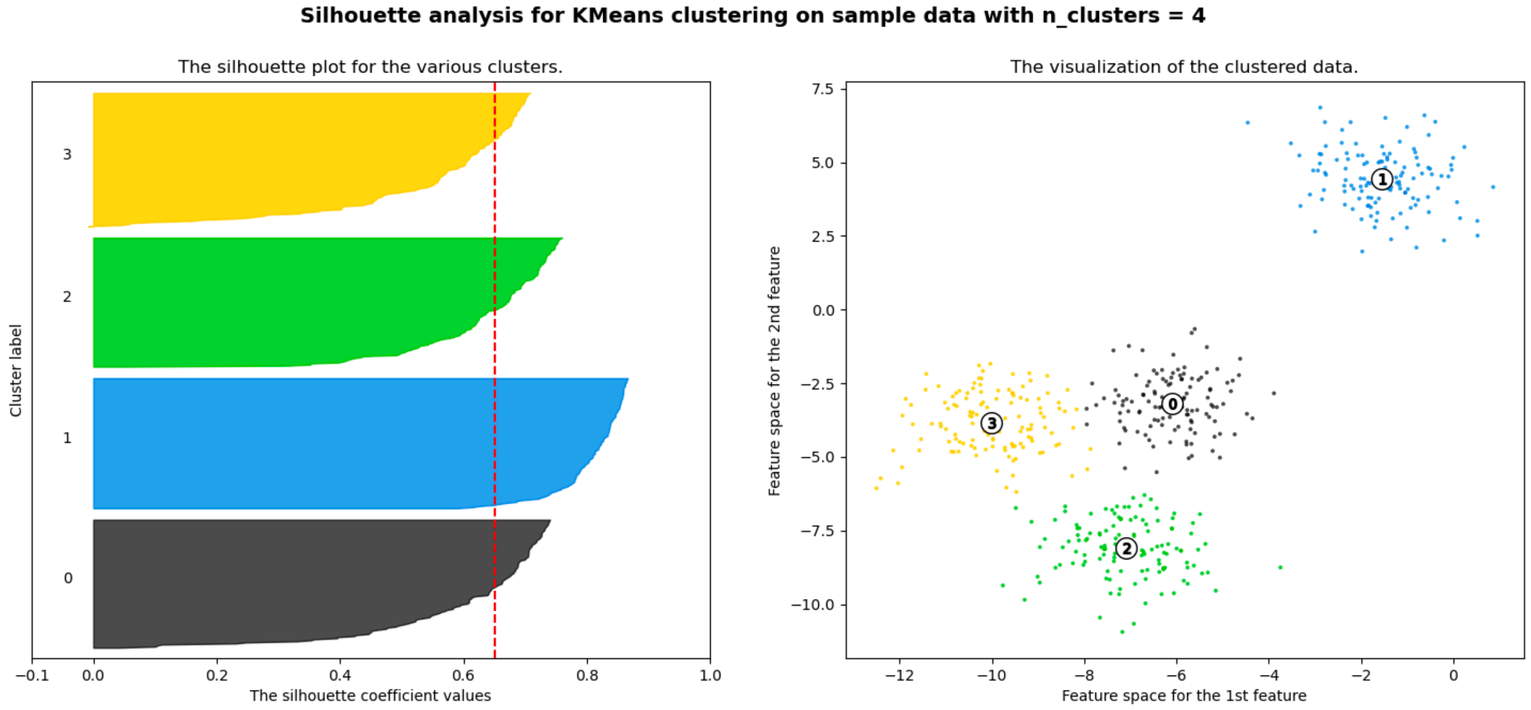


The main weakness of clustering methods is that they **do not scale well**.
Clustering methods require calculating the distance between all pairs of data points in each iteration of the algorithm. This can become computationally expensive as the number of data points grows larger.
Dendrograms cannot tell you how many clusters you should have. The number of clusters to use is typically determined through a separate process, such as a clustering validity index or domain expertise.

K-Means - How to select k?

Silhouette Analysis

Silhouette coefficients near +1 indicate that a sample is far away from the neighboring clusters.
A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.
Negative values indicate that those samples might have been assigned to the wrong cluster.



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Jheronimus Academy of Data Science