Some ideas for formalizing clustering schemes

Gunnar Carlsson and Facundo Mémoli

memoli@math.stanford.edu

http://comptop.stanford.edu
http://math.stanford.edu/~memoli

NIPS 2009

Clustering

- Clustering plays a central role in Data Analysis. It can give useful information about the structure of the data.
- Not much known about theoretical properties of clustering methods. Which methods are **stable**?
- In practice, when dealing with large datasets, one is forced to subsample the data: clustering the whole dataset is infeasible. How do the answers based on two different subsamples compare? Can I guarantee that we obtain similar answers when these subsamples are similar ?
- I'll describe work we've done in the last 3 years [**CM08,CM09-um,CM-IFCS-09**].

Standard Clustering

In this context, given a finite metric space (X, d), a clustering method f returns a partition of X:

$$f(X, d) \in \mathcal{P}(X).$$

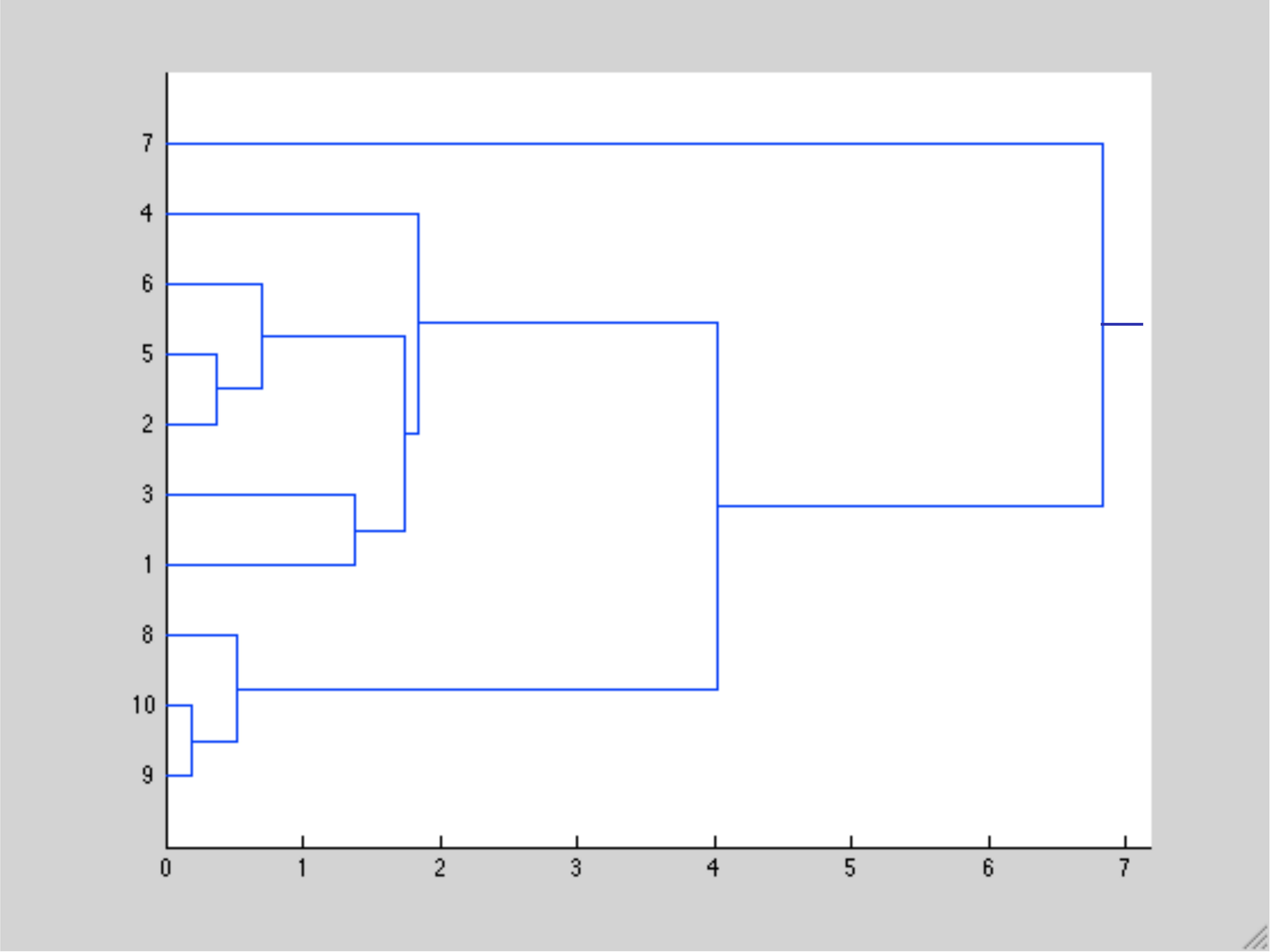
Hierarchical Clustering

Given a finite metric space (X, d), a clustering method f returns a nested family of partitions, or **dendrogram** (a.k.a. persistent set) of X:

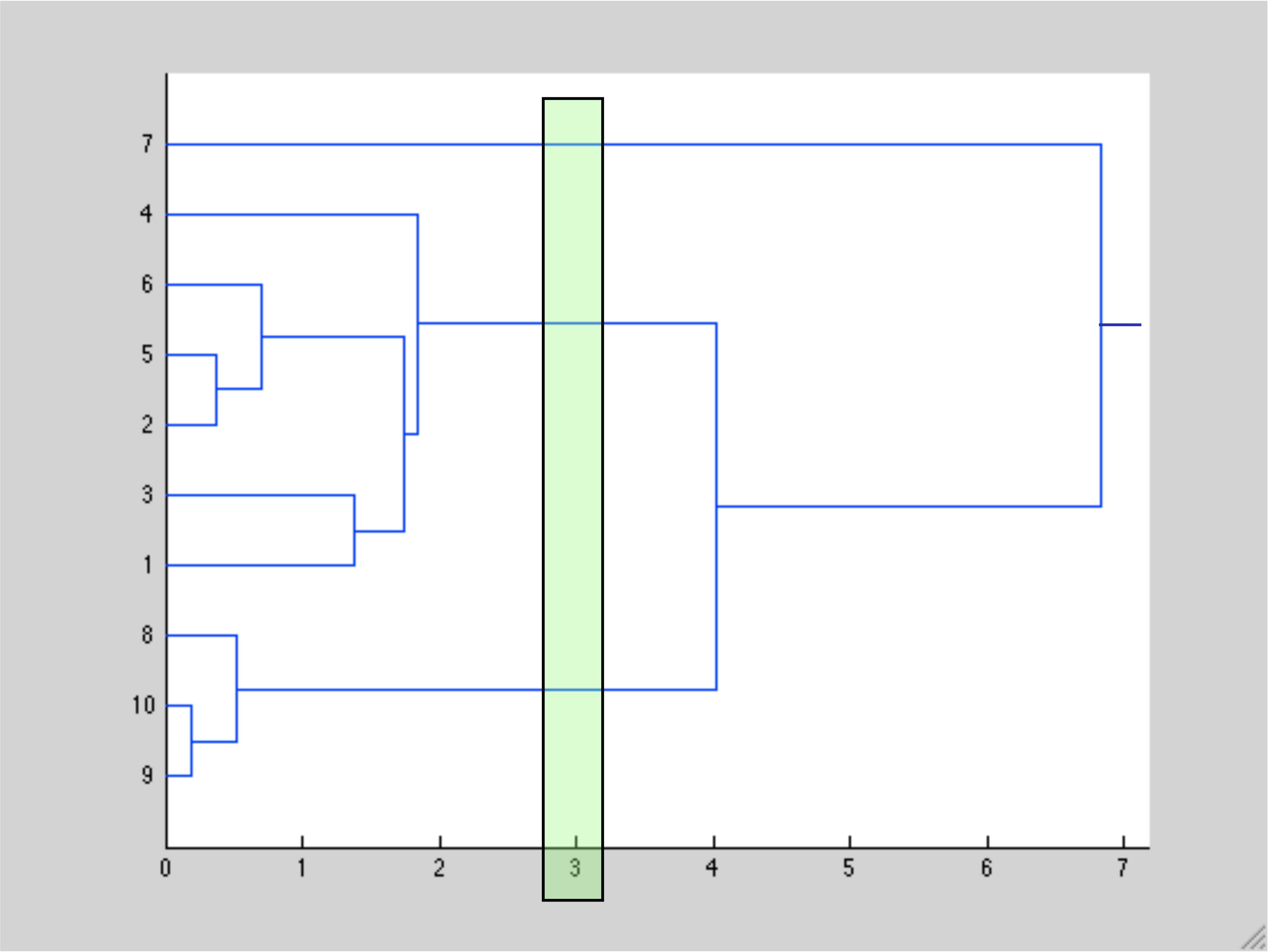
$$f(X, d) \in \mathcal{D}(X)$$

where $\mathcal{D}(X) = \{(X, \theta) | \, \theta : [0, \infty) \to \mathcal{P}(X)\}$ s.t.

1. $\theta(0) = \{\{x_1\}, \ldots, \{x_n\}\}$.
2. There exists t_0 s.t. $\theta(t)$ is the *single block partition* for all $t \geq t_0$.
3. If $r \leq s$ then $\theta(r)$ *refines* $\theta(s)$.
4. For all r there exists $\varepsilon > 0$ s.t. $\theta(r) = \theta(t)$ for $t \in [r, r + \varepsilon]$.


7
4
6
5
2
3
1
8
10
9
0
1
2
3
4
5
6
7

$$\theta(3) = \{\{7\}, \{4, 6, 5, 2, 3, 1\}, \{8, 9, 10\}\}$$



Standard Clustering: desirable properties

$$f(X, d) = \Gamma \in \mathcal{P}(X).$$

- **Scale Invariance**: For all $\alpha > 0$, $f(X, \alpha \cdot d) = \Gamma$.

- **Richness**: Fix finite set X. Require that for all $\Gamma \in \mathcal{P}(X)$, *there exists* d_Γ, metric on X s.t. $f(X, d_\Gamma) = \Gamma$.

- **Consistency**: Let $\Gamma = \{B_1, \ldots, B_\ell\}$. Let $\widehat{d}$ be any metric on X s.t.

  1. for all $x, x' \in B_\alpha$, $\widehat{d}(x, x') \leq d(x, x')$ and
  2. for all $x \in B_\alpha$, $x' \in B_{\alpha'}$, $\alpha \neq \alpha'$, $\widehat{d}(x, x') \geq d(x, x')$.

  Then, $f(X, \widehat{d}) = \Gamma$.

Kleinberg's Theorem: bad news

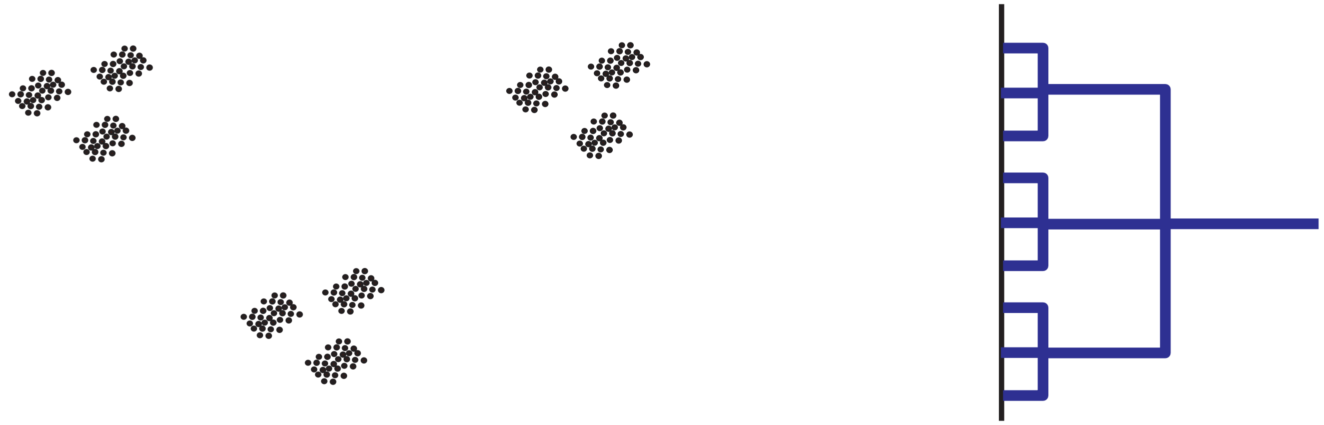
Theorem 1. *There is no standard clustering algorithm satisfying scale invariance, richness and consistency.*

Kleinberg’s Theorem: bad news

Theorem 1. *There is no standard clustering algorithm satisfying scale invariance, richness and consistency.*

Comments

- This is one more reason why one may feel that it is more sensible to look at hierarchical clustering.
- Sometimes datasets have multiscale structure, so standard clustering may not be applicable.
- So we now concentrate on hierarchical clustering methods. We wil prove a theorem in the spirit of Kleinberg’s but instead of non-existence, we’ll obtain *uniqueness*.



Hierarchical Clustering

We deal with *agglomerative* HC. For a finite metric space (X, d), its *separation* is

$$\operatorname{sep}(X, d) = \min_{x \neq x'} d(x, x').$$

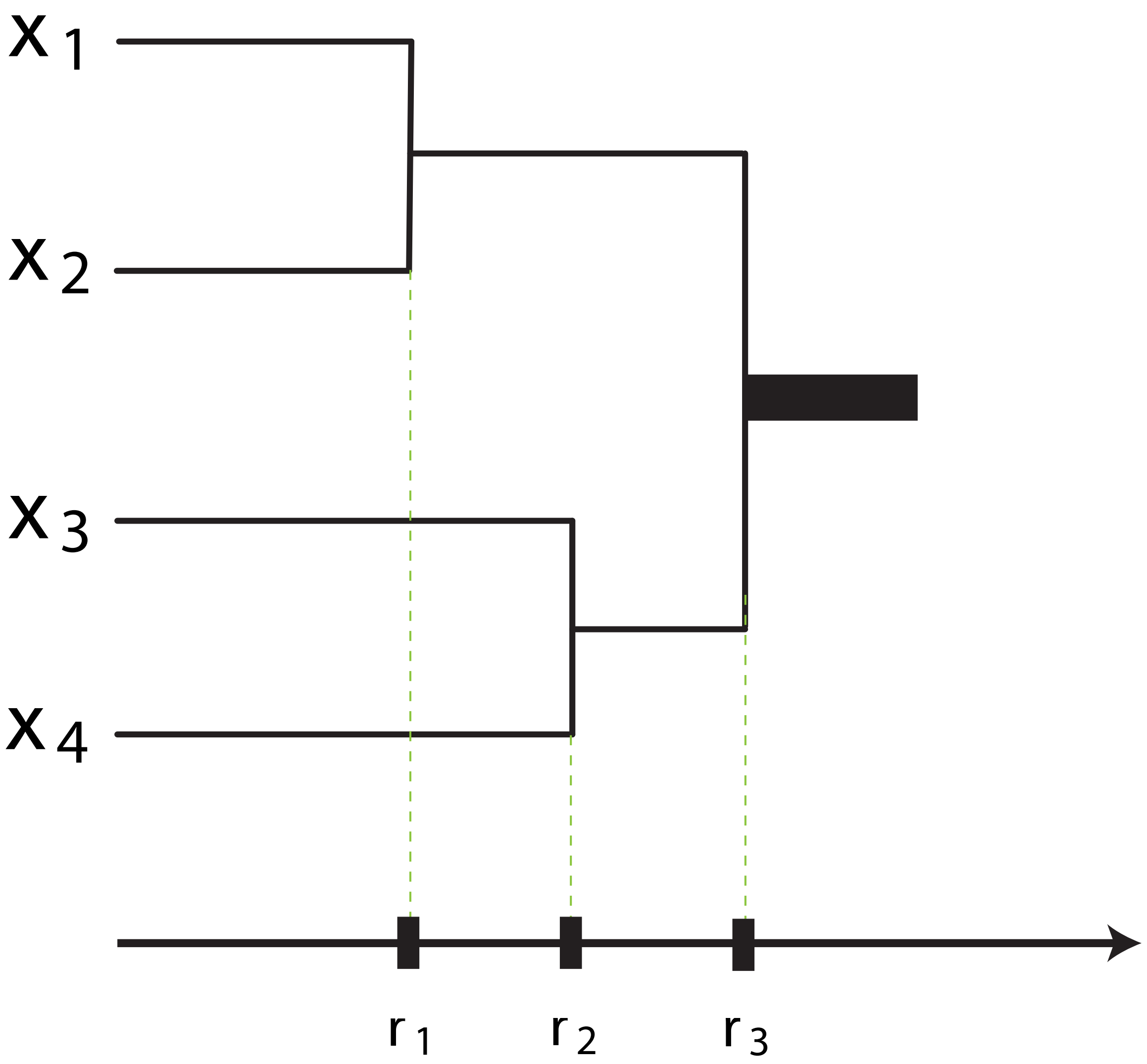
- The idea is to start with the partition of X into singletons and then begin agglomerating blocks according to some rule.

- Well known methods/rules are those given by **single, average** and **complete linkage**.

- Continue agglomerating until you are left with one single block.

- Record the values of the **linkage parameter** for which there are mergings and obtain a hierarchical decomposition of X, i.e. a dendrogram over X.

From Dendrograms to Ultrametrics



$$((u_\theta)) = \begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 0 & r_1 & r_3 & r_3 \\ x_2 & r_1 & 0 & r_3 & r_3 \\ x_3 & r_3 & r_3 & 0 & r_2 \\ x_4 & r_3 & r_3 & r_2 & 0 \end{array}$$

$$\theta \xrightarrow{\Psi} u_\theta$$

HC methods: reformulation in terms of ultrametrics

- An ultrametric u on a set X is a function $u : X \times X \to \mathbb{R}^+$ s.t.
  - $u(x, x') = 0$ if and only if $x = x'$.
  - $u(x, x') = u(x', x)$.
  - $\max(u(x, x'), u(x', x'')) \geqslant u(x, x'')$ for all $x, x', x'' \in X$.
- Let $\mathcal{U}(X)$ denote the collection of all ultrametrics on the set X.

HC methods: reformulation in terms of ultrametrics

- An ultrametric u on a set X is a function $u : X \times X \to \mathbb{R}^+$ s.t.
  - $u(x, x') = 0$ if and only if $x = x'$.
  - $u(x, x') = u(x', x)$.
  - $\max(u(x, x'), u(x', x'')) \geqslant u(x, x'')$ for all $x, x', x'' \in X$.
- Let $\mathcal{U}(X)$ denote the collection of all ultrametrics on the set X.
- It turns out that ultrametrics and dendrograms are **equivalent**.

Theorem. *For any given finite set X, there exists a bijection $\Psi : \mathcal{D}(X) \longrightarrow \mathcal{U}(X)$ such that*

$$x, x' \in B \in \theta(t) \Longleftrightarrow \Psi(\theta)(x, x') \leqslant t$$

for all dendrograms θ.

Hierarchical clustering: formulation

We represent dendrograms (= rooted trees) as *ultrametric* spaces: (X, u) is an ultrametric space if and only if for all $x, x', x'' \in X$,

$$\max(u(x,x'), u(x',x'')) \geq u(x,x'').$$

Let $\mathcal{X} = \sqcup_{n\geq 1}\mathcal{X}_n$ denote set of all finite metric spaces and $\mathcal{U} = \sqcup_{n\geq 1}\mathcal{U}_n$ all finite ultrametric spaces. Then, a hierarchical clustering method can be regarded as a map

$$T : \mathcal{X} \to \mathcal{U}$$

s.t. $\mathcal{X}_n \ni (X, d) \mapsto (X, u) \in \mathcal{U}_n$.

Hierarchical clustering: formulation

We represent dendrograms (= rooted trees) as *ultrametric* spaces: (X, u) is an ultrametric space if and only if for all $x, x', x'' \in X$,

$$\max(u(x, x'), u(x', x'')) \geq u(x, x'').$$

Let $\mathcal{X} = \sqcup_{n \geq 1} \mathcal{X}_n$ denote set of all finite metric spaces and $\mathcal{U} = \sqcup_{n \geq 1} \mathcal{U}_n$ all finite ultrametric spaces. Then, a hierarchical clustering method can be regarded as a map

$$T : \mathcal{X} \to \mathcal{U}$$

s.t. $\mathcal{X}_n \ni (X, d) \mapsto (X, u) \in \mathcal{U}_n$.

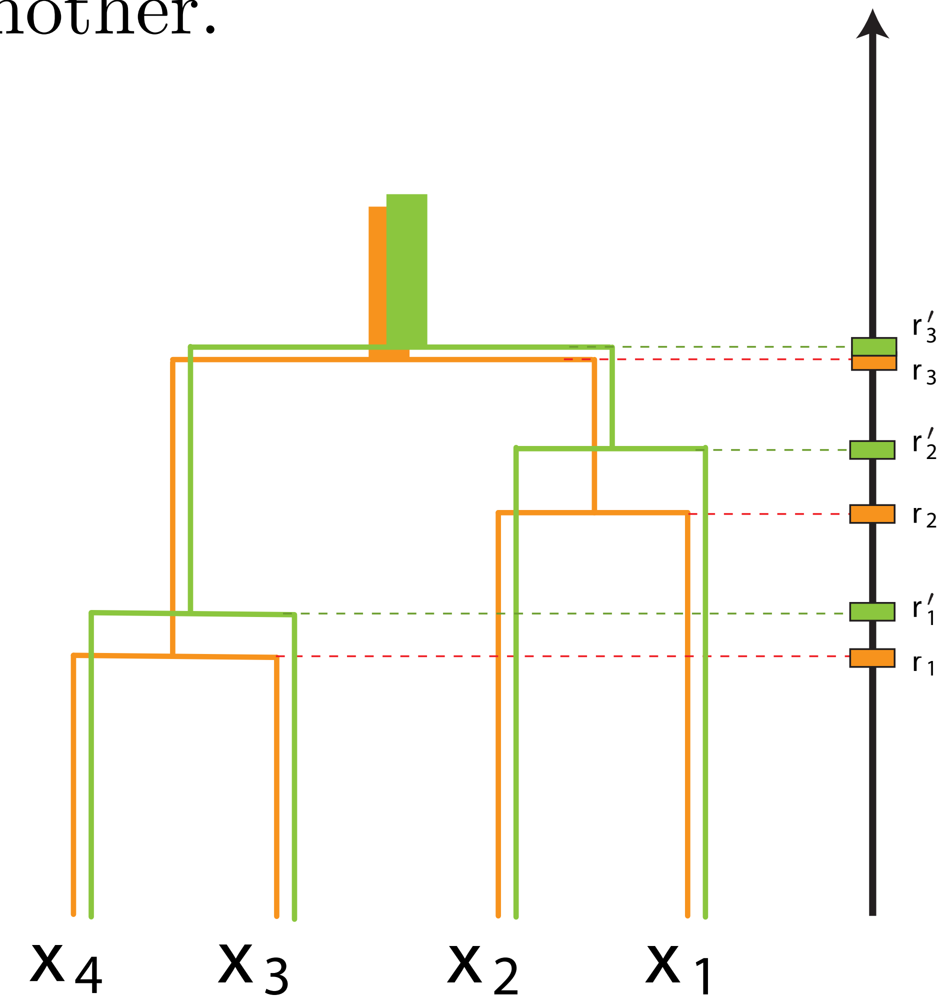
Remark. *The interpretation is that $u(x, x')$ measures the* ***effort*** *or* ***cost*** *of merging x and x' into the same cluster.*

Example: measuring distance between dendrograms

One of the consequences of the flexibility offered by the ultrametric representation of dendrograms is that one can now define some useful notions of **distance between dendrgrams**. Consider for example the case when α and β are two dendrograms over a given set X. Then, the condition that

$$\max_{x,x'} \left|\Psi(\alpha)(x,x') - \Psi(\beta)(x,x')\right| \leqslant \eta$$

translates into the fact that the points at which x and x' merge are within η of eachother.

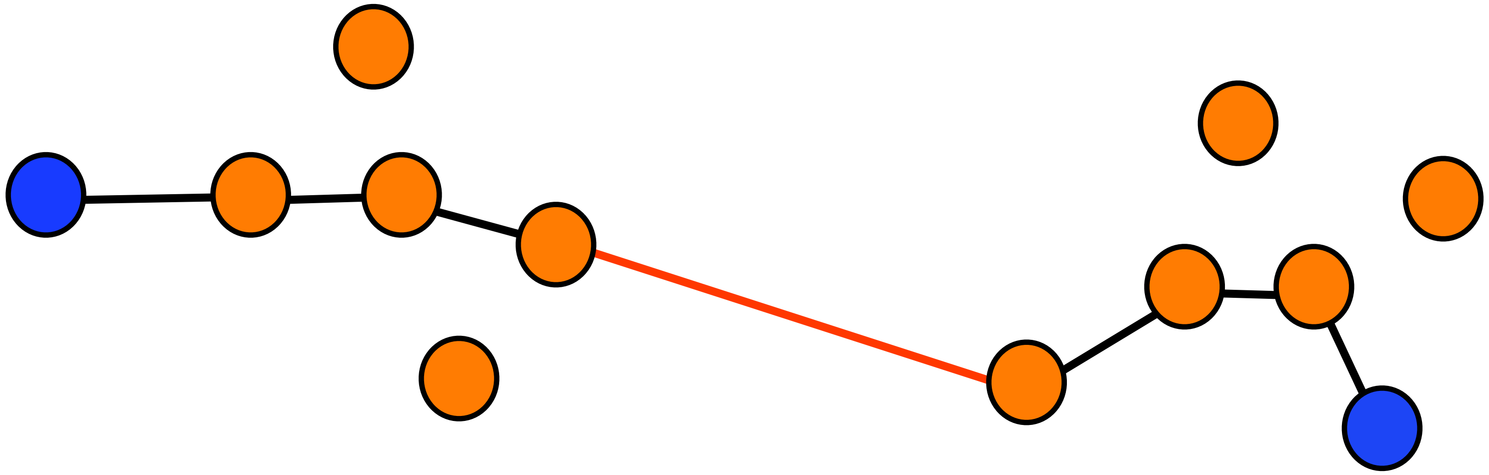


$$\max_i |r_i - r_i'| \leqslant \eta$$

Canonical construction

SL HC can be proved to be equivalent to the **maximal subdominant ultrametric**: $T^* : \mathcal{X} \to \mathcal{U}$ given by $T^*(X,d) = (X,u^*)$ where

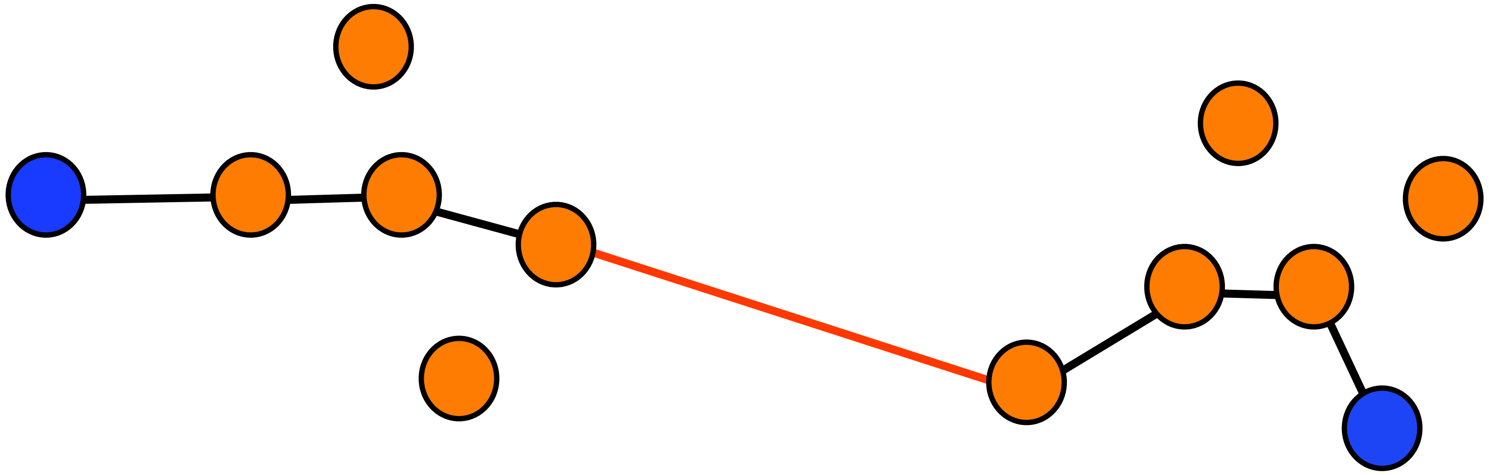
$$u^*(x,x') := \min\left\{ \max_{0\le i\le n-1} d(x_i,x_{i+1});\ x = x_0, x_1, \ldots, x_n = x' \right\}.$$



Canonical construction

SL HC can be proved to be equivalent to the **maximal subdominant ultrametric**: $T^* : \mathcal{X} \to \mathcal{U}$ given by $T^*(X,d) = (X,u^*)$ where

$$u^*(x,x') := \min \left\{ \max_{0 \leq i \leq n-1} d(x_i, x_{i+1});\; x = x_0, x_1, \dots, x_n = x' \right\}.$$



Indeed, one can prove that

Proposition. *Let (X,d) be any finite metric space and write $T^*(X,d) = (X,u^*)$. Then, the dendrogram $\Psi^{-1}(u^*)$ is equal to the one produced by SL HC applied to (X,d).*

A characterization theorem for SL, [CM08], [CM09-um]

Theorem 1. *Let T be a clustering method s.t.*

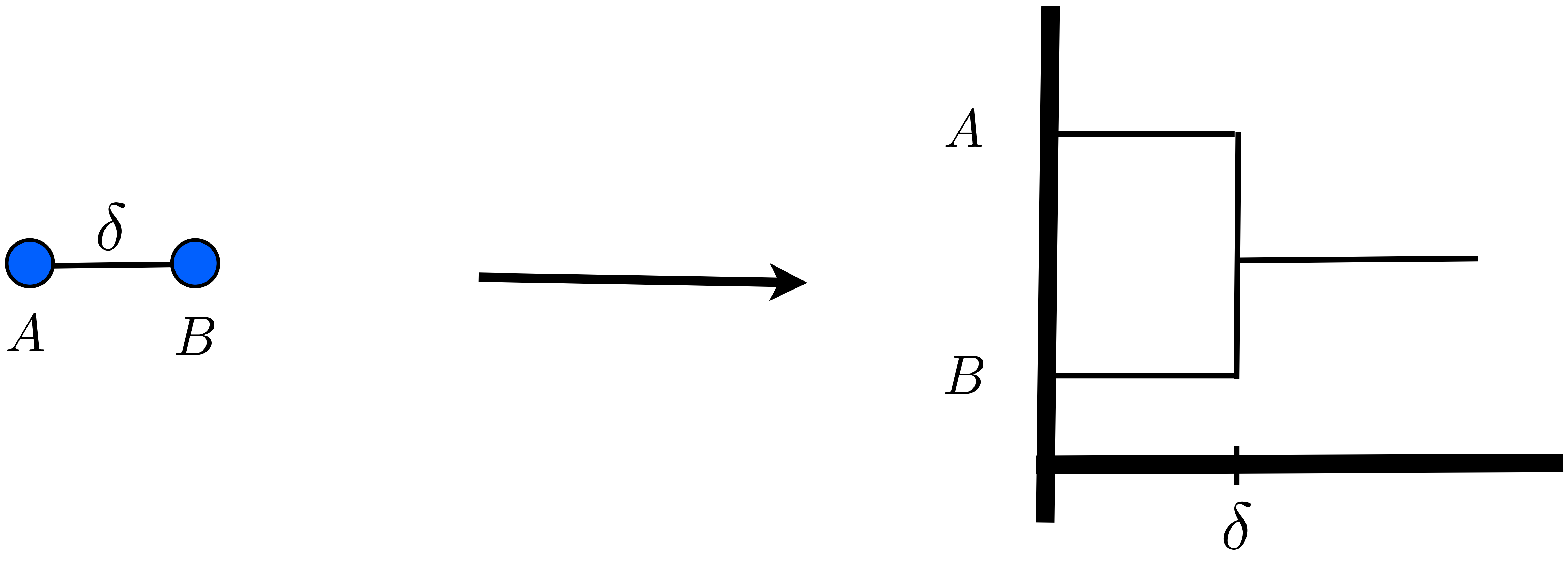
1. $T(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)) = (\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right))$ *for all* $\delta > 0$.
2. *For all* $X, Y \in \mathcal{X}$ *and* $\phi : X \to Y$ *s.t.* $d_X(x,x') \geq d_Y(\phi(x), \phi(x'))$,
$$u_X(x,x') \geq u_Y(\phi(x), \phi(x'))$$
for all $x, x' \in X$, *where* $T(X, d_X) = (X, u_X)$ *and* $T(Y, d_Y) = (Y, u_Y)$.
3. *For all* $(X, d) \in \mathcal{X}$,
$$u(x,x') \geq sep(X,d) \text{ for all } x \neq x' \in X$$
where $T(X,d) = (X,u)$.

Then $T = T^*$.

Interpretation of the conditions of the theorem

Condition I:

for all $\delta > 0$



Standard Clustering: desirable properties

$$f(X, d) = \Gamma \in \mathcal{P}(X).$$

- **Scale Invariance**: For all $\alpha > 0$, $f(X, \alpha \cdot d) = \Gamma$.

- **Richness**: Fix finite set X. Require that for all $\Gamma \in \mathcal{P}(X)$, *there exists* d_Γ, metric on X s.t. $f(X, d_\Gamma) = \Gamma$.

- **Consistency**: Let $\Gamma = \{B_1, \ldots, B_\ell\}$. Let $\widehat{d}$ be any metric on X s.t.

  1. for all $x, x' \in B_\alpha$, $\widehat{d}(x, x') \leq d(x, x')$ and
  2. for all $x \in B_\alpha$, $x' \in B_{\alpha'}$, $\alpha \neq \alpha'$, $\widehat{d}(x, x') \geq d(x, x')$.

  Then, $f(X, \widehat{d}) = \Gamma$.

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $d_X(x, x') \geqslant d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$. Then

$$u_X(x, x') \geqslant u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.$$

This means roughly that decreasing the distances has the effect of reducing the **cost** of merging points.

Cf. Kleinberg's *consistency* property.

$$\begin{tikzcd}
(X, d_X) \arrow[r, "T"] \arrow[d, "\phi"'] & (X, u_X) \arrow[d, "\phi"] \\
(Y, d_Y) \arrow[r, "T"] & (Y, u_Y)
\end{tikzcd} \tag{1}$$

(this would be called **functoriality**)

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $\boxed{d_X(x, x') \geqslant d_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.}$ Then

$$u_X(x, x') \geqslant u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.$$

This means roughly that decreasing the distances has the effect of reducing the **cost** of merging points.

Cf. Kleinberg's *consistency* property.

$$\begin{tikzcd}
(X, d_X) \arrow[r, "T"] \arrow[d, "\phi"'] & (X, u_X) \arrow[d, "\phi"] \\
(Y, d_Y) \arrow[r, "T"] & (Y, u_Y)
\end{tikzcd} \tag{1}$$

(this would be called **functoriality**)

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $d_X(x, x') \geqslant d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$. Then

$$\boxed{u_X(x, x') \geqslant u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.}$$

This means roughly that decreasing the distances has the effect of reducing the **cost** of merging points.

Cf. Kleinberg's *consistency* property.

$$\begin{tikzcd}
(X, d_X) \arrow[r, "T"] \arrow[d, "\phi"'] & (X, u_X) \arrow[d, "\phi"] \\
(Y, d_Y) \arrow[r, "T"] & (Y, u_Y)
\end{tikzcd} \tag{1}$$

(this would be called **functoriality**)

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $d_X(x, x') \geqslant d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$. Then

$$u_X(x, x') \geqslant u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.$$

This means roughly that decreasing the distances has the effect of reducing the **cost** of merging points.

Cf. Kleinberg's *consistency* property.

$$\begin{tikzcd}
(X, d_X) \arrow[r, "T"] \arrow[d, "\phi"'] & (X, u_X) \arrow[d, "\phi"] \\
(Y, d_Y) \arrow[r, "T"] & (Y, u_Y)
\end{tikzcd} \qquad (1)$$

(this would be called **functoriality**)

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $d_X(x,x') \geqslant d_Y(\phi(x),\phi(x'))$ for all $x, x' \in X$. Then

$$u_X(x,x') \geqslant u_Y(\phi(x),\phi(x')) \text{ for all } x, x' \in X.$$

This means roughly that decreasing the distances has the effect of reducing the **cost** of merging points.

Cf. Kleinberg's *consistency* property.

$$\begin{tikzcd}
(X, d_X) \arrow[r, "T"] \arrow[d, swap, "\phi"] & \boxed{(X, u_X)} \arrow[d, "\phi"] \\
(Y, d_Y) \arrow[r, "T"] & \boxed{(Y, u_Y)}
\end{tikzcd} \tag{1}$$

(this would be called **functoriality**)

Condition II

Let $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ s.t. $d_X(x, x') \geqslant d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$. Then

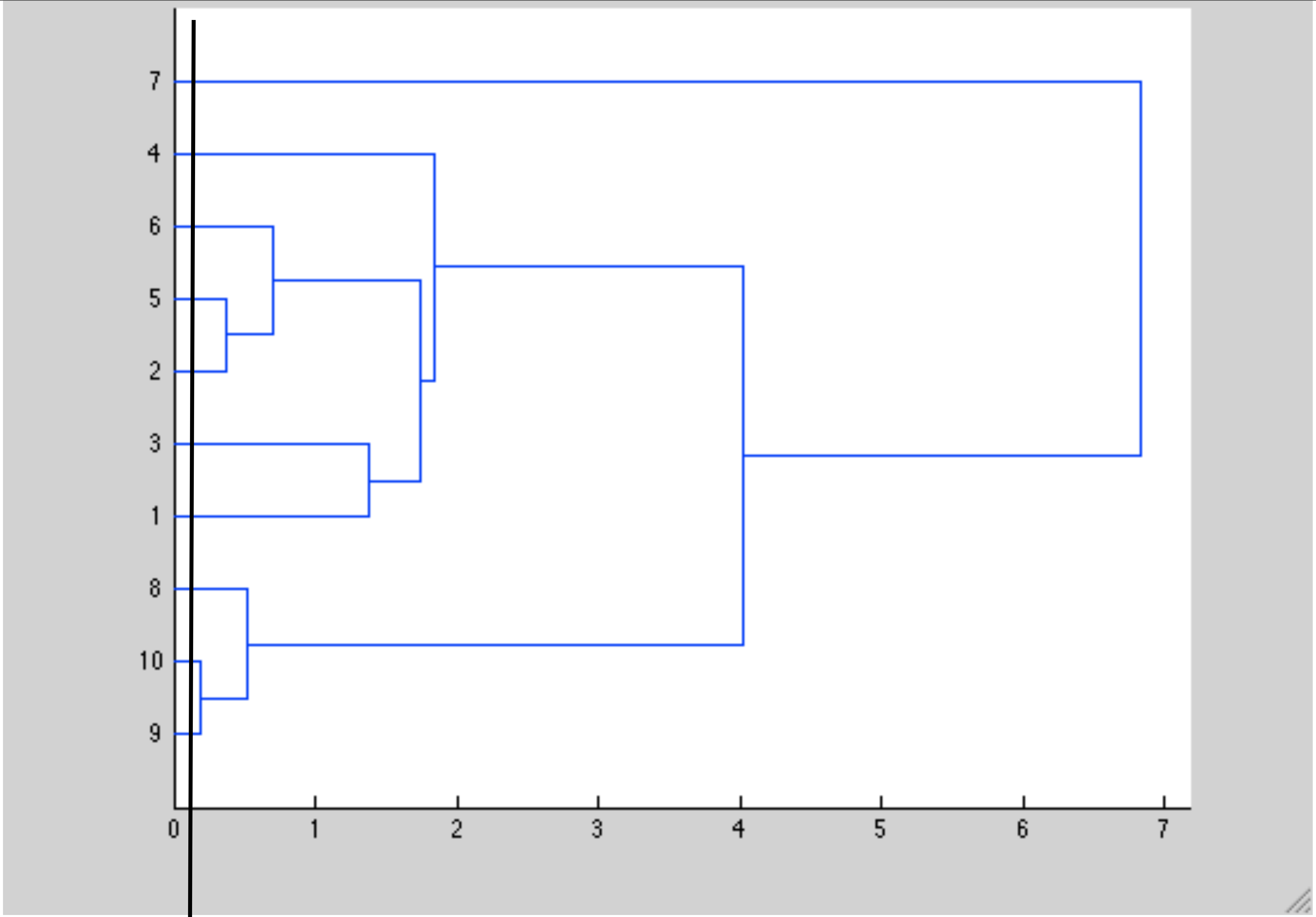
$$u_X(x, x') \geqslant u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.$$

This means roughly that decreasing (not reducing) the distances has the effect of reducing (not increasing) the **cost** of merging points.

Condition III

$$u(x, x') \geqslant \mathrm{sep}(X, d) \text{ for all } x, x' \in X.$$

This means roughly that the cost of merging to points has to be at least the *separation* of the space.



$\mathrm{sep}(X, d)$

A characterization theorem for SL, [CM08], [CM09-um]

Theorem 1. *Let T be a clustering method s.t.*

1. $T(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)) = (\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right))$ *for all* $\delta > 0$.
2. *For all* $X, Y \in \mathcal{X}$ *and* $\phi : X \to Y$ *s.t.* $d_X(x,x') \geq d_Y(\phi(x), \phi(x'))$,
$$u_X(x,x') \geq u_Y(\phi(x), \phi(x'))$$
for all $x, x' \in X$, *where* $T(X, d_X) = (X, u_X)$ *and* $T(Y, d_Y) = (Y, u_Y)$.
3. *For all* $(X, d) \in \mathcal{X}$,
$$u(x,x') \geq sep(X,d) \text{ for all } x \neq x' \in X$$
where $T(X,d) = (X,u)$.

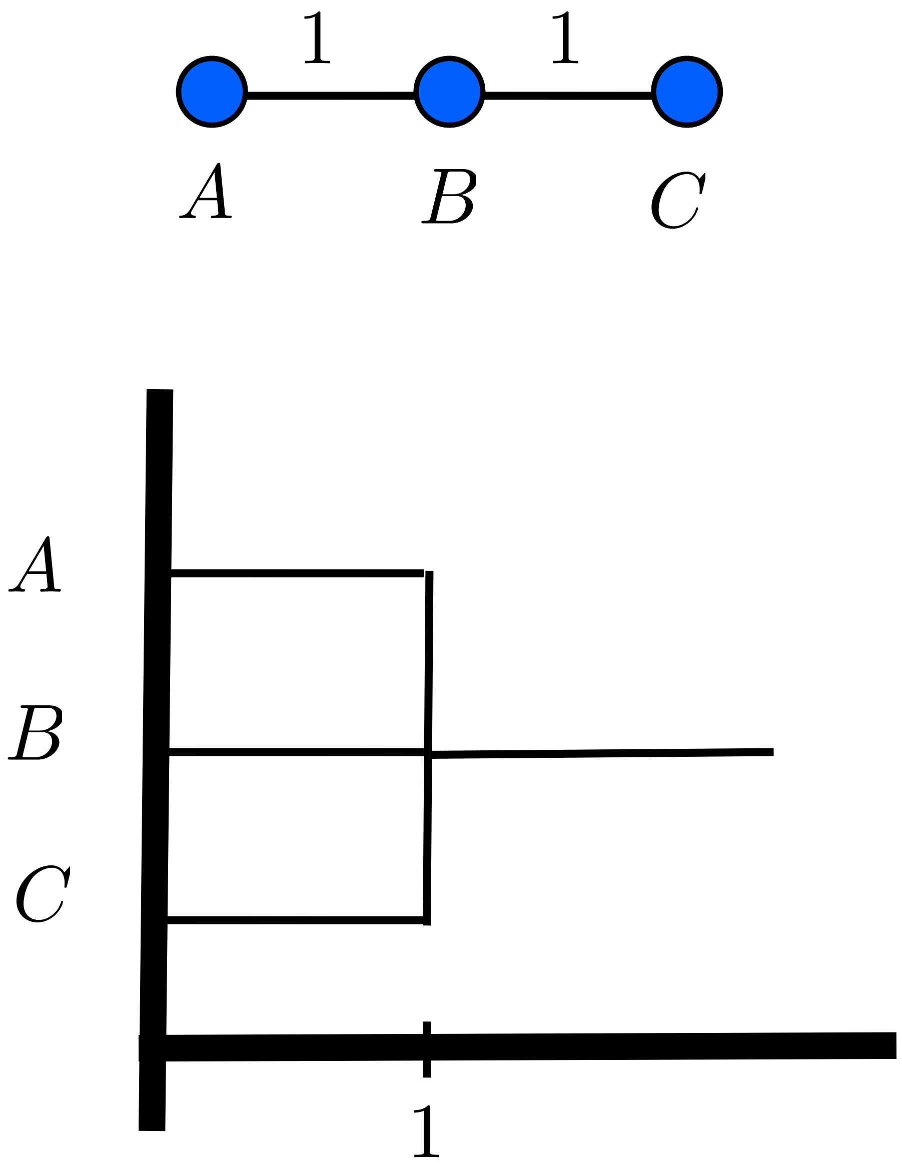
Then $T = T^*$.

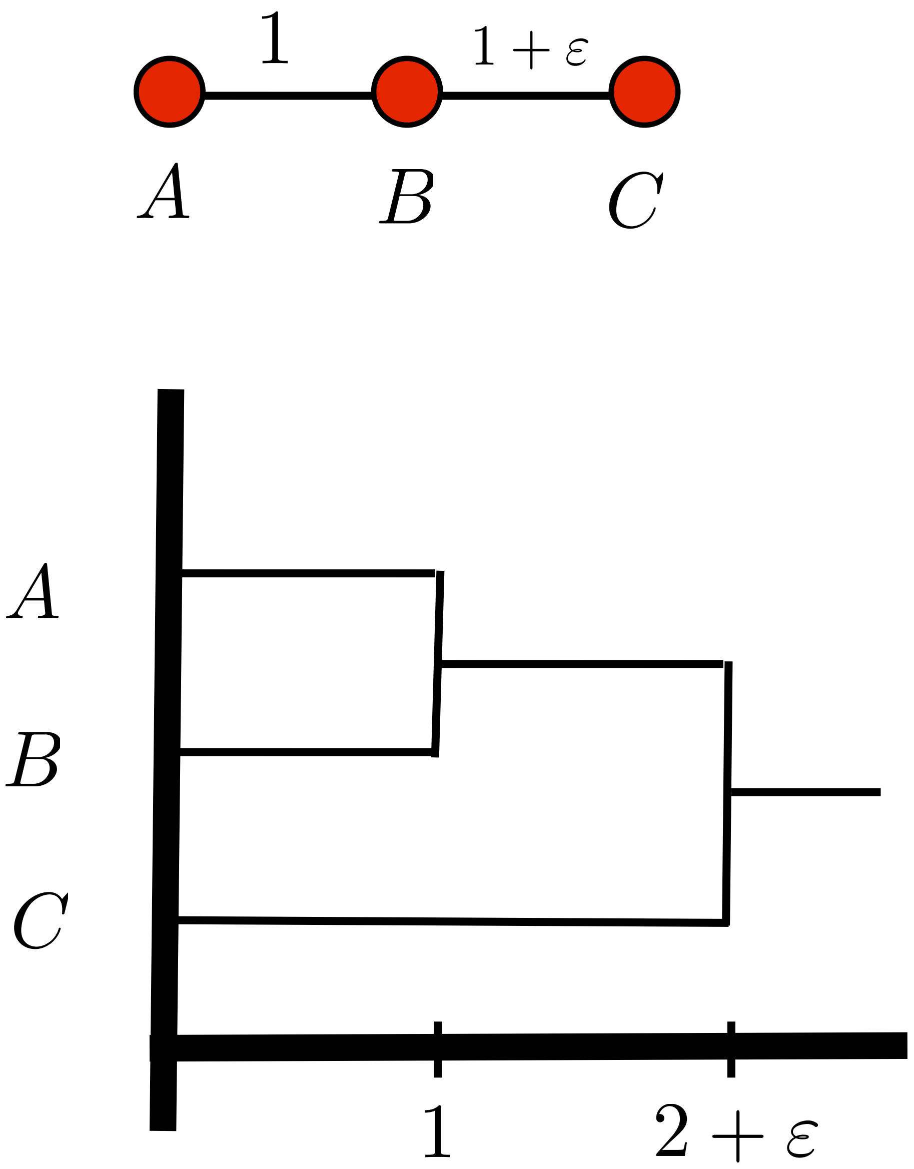
Two other aspects of our work

- Stability
- Convergence

Stability properties of HC methods

- CL and AL are not stable!!
- SL is stable.





Stability of SL HC, [CM08], [CM09-um]

Proposition 1. *For any finite metric spaces* (X, d_X) *and* (Y, d_Y)

$$d_{\mathcal{GH}}((X, d_X), (Y, d_Y)) \geq d_{\mathcal{GH}}(T^*(X, d_X), T^*(Y, d_Y)).$$

Moral: metrically similar subsets of my data will yield similar clustering results, when the clustering method is **SL**.

Stability of SL HC, [CM08], [CM09-um]

Proposition 1. *For any finite metric spaces* (X, d_X) *and* (Y, d_Y)

$$d_{\mathcal{GH}}((X, d_X), (Y, d_Y)) \geq d_{\mathcal{GH}}(T^*(X, d_X), T^*(Y, d_Y)).$$

Moral: metrically similar subsets of my data will yield similar clustering results, when the clustering method is **SL**.

Consequence: Convergence

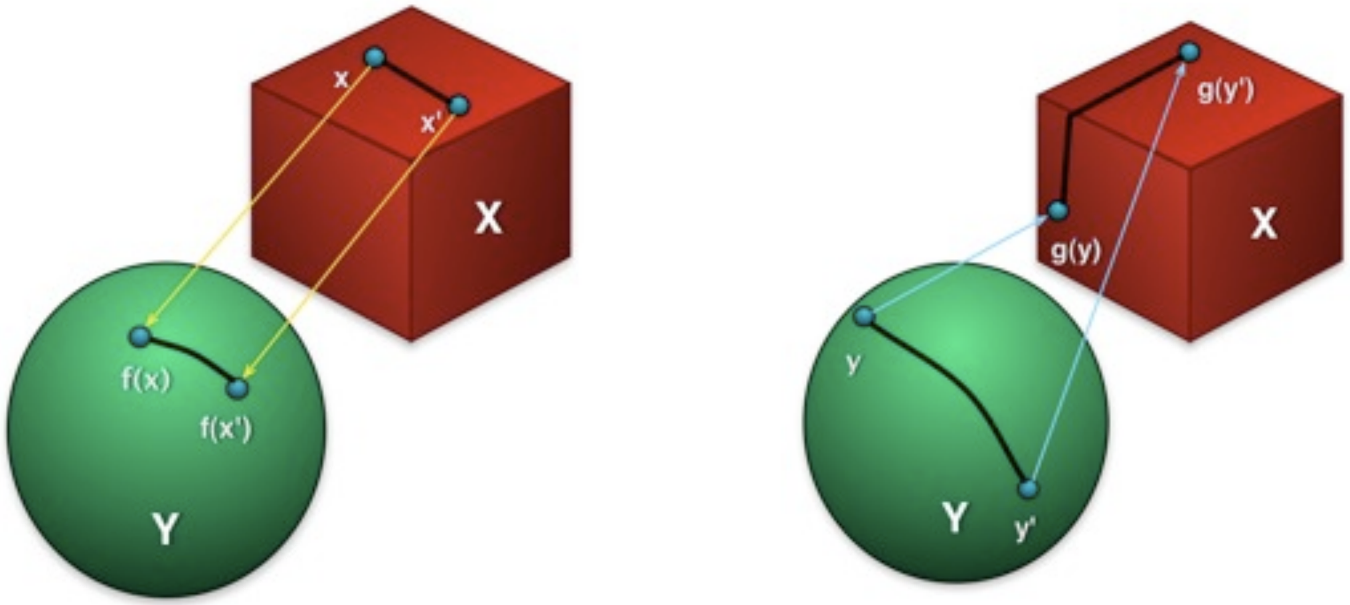
The Gromov-Hausdorff distance

- It is well studied and well understood notion of distance between metric spaces.
- It is insensitive to relabelling (actually to *isometries*)
- We view dendrogram as (ultra) metric spaces ⇒ we can use the GH distance to compare dendrograms.
- Roughly the definition is the following: $d_{\mathcal{GH}}\big(X,Y\big) \leqslant \eta$ if and only if there exist maps $f : X \to Y$ and $g : Y \to X$ with the property that

$$|d_X(x,x') - d_Y(f(x),f(x'))| \leqslant \eta \text{ for all } x,x' \in X$$

and

$$|d_Y(y,y') - d_X(g(y),g(y'))| \leqslant \eta \text{ for all } y,y' \in Y.$$



The Gromov-Hausdorff distance: dendrograms

In terms of dendrograms,

$$d_{\mathcal{GH}}\big(\Psi(\theta_X), \Psi(\theta_Y)\big) \leqslant \eta$$

means that there exist f and g s.t.

- two points x, x' fall in the same same block of $\theta_X(t)$ implies that $f(x)$ and $f(x')$ fall in the same block of $\theta_Y(t')$ for some $t' \in [t-\eta, t+\eta]$.
- two points y, y' fall in the same same block of $\theta_Y(t)$ implies that $g(y)$ and $g(y')$ fall in the same block of $\theta_X(t')$ for some $t' \in [t-\eta, t+\eta]$.

Another aspect of our work: convergence

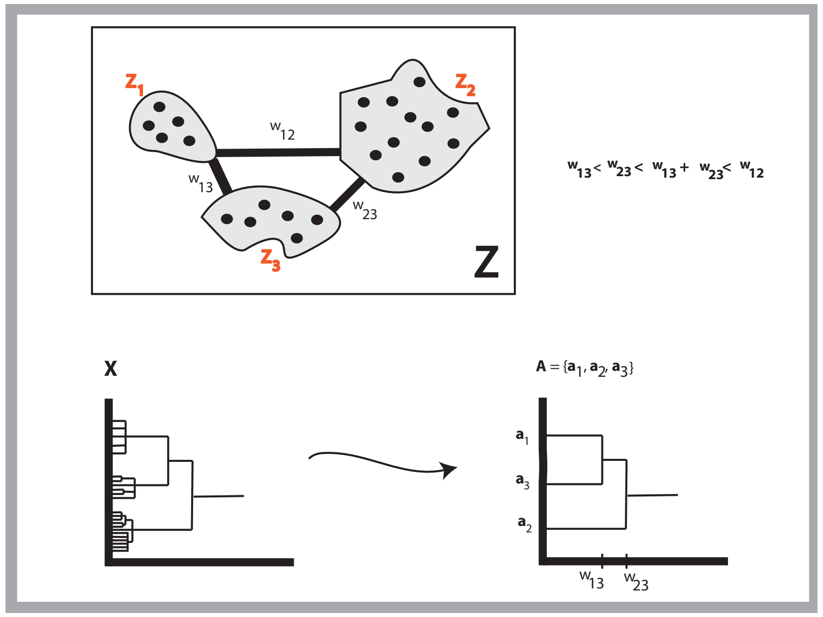
Say you are given finitely many random i.i.d. samples $X_n = \{x_1, x_2, \ldots, x_n\}$ from a metric space (Z, d_Z), where each x_i is distributed according to a probability measure μ **compactly supported** on Z. Then, compute θ_{X_n} the **SL** dendrogram of X_n.

The question is: what does θ_{X_n} converge to (if at all)?

We answer this question in our work and generalize a classical result by Hartigan regarding the properties of SL. Namely, we prove that

$$\mathbf{P}\left(\lim_n \theta_{X_n} = \theta_\mu\right) = 1$$

for some dendrogram θ_μ that captures the multiscale structure of $\text{supp}\,[\mu]$.



Another aspect of our work: convergence

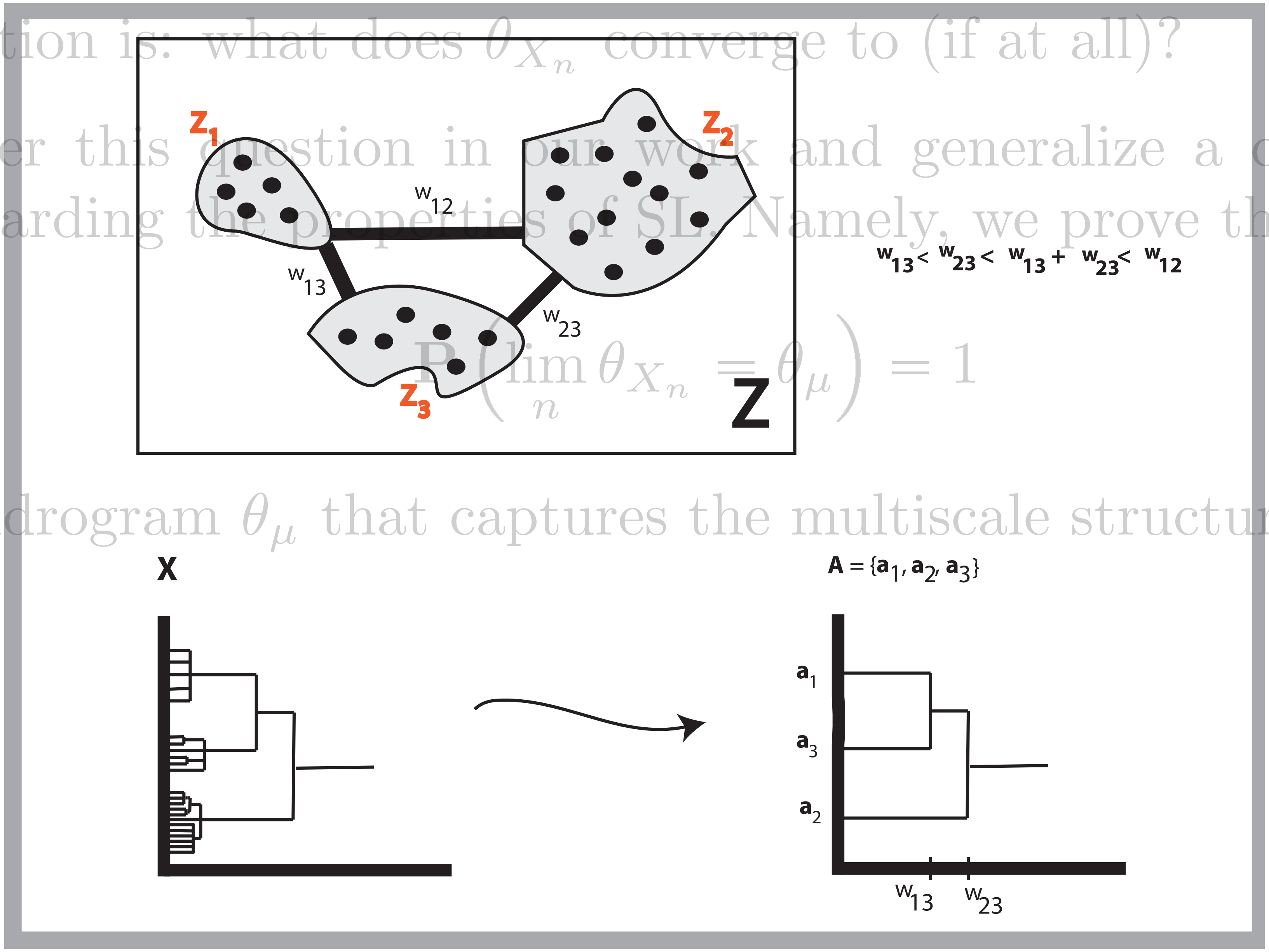
Say you are given finitely many random i.i.d. samples $X_n = \{x_1, x_2, \ldots, x_n\}$ from a metric space (Z, d_Z), where each x_i is distributed according to a probability measure μ **compactly supported** on Z. Then, compute θ_{X_n} the **SL** dendrogram of X_n.

The question is: what does θ_{X_n} converge to (if at all)?

We answer this question in our work and generalize a classical result by Hartigan regarding the properties of SL. Namely, we prove that

$$\mathbb{P}\left(\lim_n \theta_{X_n} = \theta_\mu\right) = 1$$

for some dendrogram θ_μ that captures the multiscale structure of supp $[\mu]$.



Discussion

- SL HC is stable and enjoys all nice properties but it is derided by practicioners because of its insensitivity to density: **chaining effect**.

- AL, CL do exhibit sensitivity to density, yet they are theoretically unsound

  – The standard version: because it is not well behaved under permutations.

  – The ”fixed” version: because it is unstable!

- As a solution we propose to look at **two-parameter clustering**: look at certain two-dimensional analogues of dendrograms [**CM-IFCS-09**].

- Another line of work: study different trade-offs in the properties required from standard clustering.

- The underlying concepts in our work are **functoriality** and **metric geometry**.

References

[**CM-IFCS-09**] Gunnar Carlsson and Facundo Mémoli. Multiparameter clustering methods. In *IFCS 2009*, 2009.

[**Kl02**] Jon M. Kleinberg. An impossibility theorem for clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 446–453. MIT Press, 2002.

[**CM08**] G. Carlsson and F. Mémoli. Persistent Clustering and a Theorem of J. Kleinberg. *ArXiv e-prints*, August 2008.

[**CM-SC**] Gunnar Carlsson and Facundo Mémoli. Classifying clustering schemes. Technical report, 2009. In preparation.

[**CM09-um**] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering algorithms. Technical report, 2009.

[**JD88**] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice Hall Advanced Reference Series. Prentice Hall Inc., Englewood Cliffs, NJ, 1988.

`memoli@math.stanford.edu`

`http://comptop.stanford.edu`
`http://math.stanford.edu/~memoli`