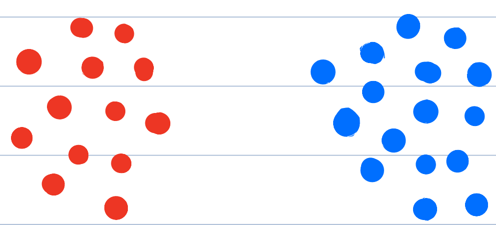
AMAT 583 Lec 19, 11/5/19

<u>Today</u>: Clustering, continued

<u>Review</u>

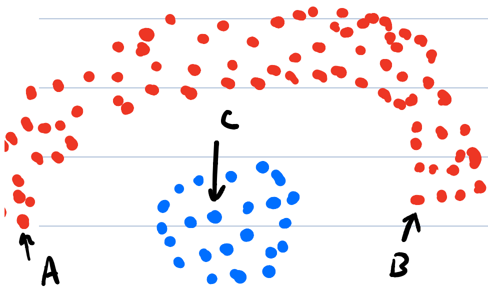
Rough, informal description of the clustering problem: Given a finite subset $X \subset \mathbb{R}^n$, break up X into subsets so that points in the same subset are close, and points within different subsets are far.

<u>Example</u> ($n=2$)



intuitively, there are two clusters here

Example



Again, intuition suggests that there are two clusters, but points A and B are further from each other than the are to C, even though it seems A and B should be clustered together, while C should be in a different cluster.

This suggests a defect with the rough definition of clustering given above. Indeed, devising a good defintion of clustering is problematic, and there is no universally agreed upon definition. Instead, There are many propasal for different definitions.

<u>A typical scientific setting</u>: Clustering breast cancers into subtypes.

<u>Motivation</u>: If we distinguish between different cancer subtypes, we can studying + treat the different subtypes separately (a divide + conquer approach).

For example, we may have $X = \{x^1, \ldots, x^{300}\} \subseteq \mathbb{R}^{24{,}000}$

- 300 breast cancer patients + heathy patients
- Each x^i represents a tissue sample from a patient.
- We consider the level of expression of 24,000 genes in each tissue sample.
- Letting $x^i = (x^i_1, x^i_2, \ldots, x^i_{24{,}000})$, x^i_j is the level of expression of gene j in tissue sample i.

Gene expression levels are measured using RNA sequencing.

Clusters in X should correspond to cancer subtypes.

Formal specification of the input and output of the clustering problem.

Set Theory language

For a set T, a set P of non-empty subsets of T is a partition if each element of T belongs to exactly one element of P.

A subpartition P a partition of a subset of T.
Thus every element of T belongs to at most one element of P.

<u>Example</u>: Let $T = \{1,2,3,4\}$.

$P = \{\{1,2\},\{3,4\}\}$ is a partition of P and also a subpartition.

$P = \{\{1,2\},\{4\}\}$ is a subpartition but not a partition.

<u>Input of the clustering problem:</u>

1) A finite set of points $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^n$ OR
2) more generally, X may be a finite metric space, represented as an $n \times n$ matrix D with $D_{i,j} = d(x_i, x_j)$ (where $d: X \times X \to [0,\infty)$ is the metric.

for example X could be a finite set of 3-D configurations of a protein, and d could be RMSD.

<u>Output of the clustering problem</u>

1) A subpartition P of X (usually a partition, actually)
2) A hierarchical partition or subpartition

A <u>clustering method</u> is a function mapping an input to an output

Definition: A hierarchical partition of a set X is a family of partitions of X $\{P_\alpha\}_{\alpha\in[0,\infty)}$ such that for any $\alpha \le \beta \in [0,\infty)$, if x any y belong to the same element of P_α, then they belong to the same element of P_β.

Definition: A hierarchical subpartition is defined the same way, except replacing the word "partition" with "subpartition."

Note: Sometimes, it will be more convenient to think of α as belonging to the non-negative integers $\mathbb{N}$ than to $[0,\infty)$.

Next, I want to show an example of a clustering method called single linkage clustering, a topologically flavored method.

For this, we will need to define graphs.

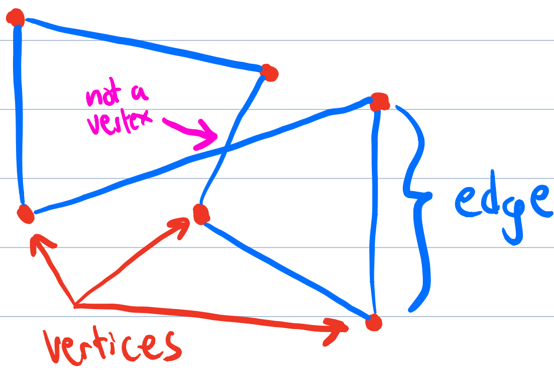
- Graphs are very important constructions in computer science, mathematics, and statistics

- **Note:** These graphs are not the same as the graphs of of functions you've seen since high school.

We distinguish between <u>directed</u> and <u>undirected</u> graphs.

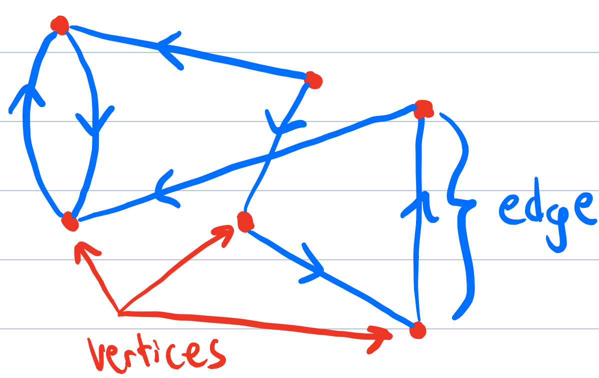
Intuitively, an undirected graph is a collection of points (vertices) with edges connecting them.

We can draw this in the plane like so



However, formally, we don't usually assume that the vertices live in $\mathbb{R}^2$.

A directed graph is a similar kind of object, but the edges are each assumed to have a <u>direction</u> from one vertex to another.



For now, we will be concerned only with undirected graphs, so we define these formally.

Definition: An <u>undirected graph</u> is a pair (V, E) where V is any set (called the vertex set) and E is a set of two element subsets of V (called the edge set.