✲ Clustering

- Given a set of data points, try to understand their structure
- Find similarities ⟶ Group similar data objects into clusters
- It is an unsupervised learning algorithm (No predefined class)
- **Good clustering:** high intra-class similarity, low inter-class similarity
- **Evaluation:** • Various measure of intra/inter cluster similarity • Manual inspection • Benchmarking on existing labels
- **Similarity** expressed in terms of distance function (different for interval-scaled, boolean, categorical, vector var.)
- Given set of points on a given space, a distance function $d(x,y)$ maps x and y to a real number, and satisfies:
  - $d(x,y) \geq 0$
  - $d(x,y) = 0$ if and only if $x = y$
  - $d(x,y) = d(y,x)$
  - $d(x,y) \leq d(x,z) + d(z,y)$

» Euclidean Distance

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

» Manhattan Distance

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sum_{i=1}^{n} |x_i - y_i|$$

» Jaccard Distance

$d(x,y) = 1 - J(x,y)$: Jaccard Distance

$J(x,y) = \frac{|x \cap y|}{|x \cup y|}$: Jaccard Similarity

→ **Example:**

	A	B	C	D	E
	0	0	1	0	0
	0	0	1	0	1

↳ Jaccard distance = 0.5 = $1 - \frac{1}{2}$
first example contains only C while second contains C and E

	A	B	C	D	E	F
	0	0	1	0	0	2
	0	0	1	0	1	1

↳ Jaccard distance is $1 - \frac{1}{3} = \frac{2}{3}$
Same: C (1)
Different: D and E (2)

⚠ In the second example if E was 20, 1; Jaccard Score would be the same ($\frac{2}{3}$) ⟶ underestimate in numerical variables (F)

» Hamming Distance

$d(x,y) = \frac{p - m}{p}$ (m: # of matching components, p: # of variables)

→ **Example:**

	A	B	C	D	E
	0	0	1	0	0
	0	0	1	0	1

→ Hamming distance = $\frac{5-4}{5} = 1/5$

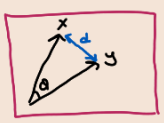
	A	B	C	D	E	F
	0	0	1	0	0	2
	0	0	1	0	1	1

→ Hamm. dist = $\frac{6-4}{6} = 2/6$

» Cosine Similarity

$$s(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

$d(x,y) = 1 - s(x,y)$



→ angular cosine distance:

$d(x,y) = c \times \frac{\arccos(s(x,y))}{\pi}$

$c = 1$ negative and positive values; $c = 2$ just positive values

» Edit Distance

$d(x,y) = |x| + |y| - 2|LCS|$ (LCS: longest common subsequence)

* distance between string x and y is the smallest # of insertion/deletion of single characters that will transform x→y

→ **Example:**

$x = a\ b\ c\ d\ e$
$y = a\ c\ f\ d\ e\ g$
{delete b, insert f, insert g}

Edit distance is 3
$|x| = 5$, $|y| = 6$, $|LCD| = 4$ which is acde

Some tools allow users to define their own distance function. But other provide only the usual distance function (like Euclidean, Manhatten...) → when custom function not an option we can try to transform the data

» Normalization

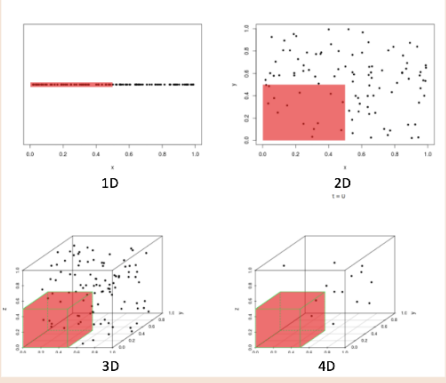
- Different attributes are measured on different scales → Need to normalization

Range Normalization: $x_i' = \frac{x_i - min_i x_i}{max_i x_i - min_i x_i}$ Score Normalization: $x_i' = \frac{x_i - \mu}{\sigma}$

- Other Normalization Approach: RobustScaler, Normalizer, MaxAbs Scaler.

» Curse of Dimensionality (High-dim data)

- As the # of dimensions in a dataset increases, distance measures become increasingly meaningless



- Possible Solutions → Dimensionality reduction
  → Subspace clustering