Measuring distance and similarity of data objects

Many different data

- documents (webpages, books)
- records of users
- graphs
- images
- videos
- Strings (DNA sequences)
- Timeseries
- **How do we compare them?**

Data Representation

dataset X as a collection of objects

write x, y, z, ... for objects in X

at this point no assumption about the representation of objects in X

x can be

- real-valued vectors
- binary vectors
- sets
- time series
- images

Distance function

want to define function

$$d : X \times X \to \mathbb{R}$$

what properties should d have?

Distance functions

$d(x, y) \geq 0$	non negativity
$d(x, y) = 0 \text{ iff } x = y$	isolation
$d(x, y) = d(y, x)$	symmetry
$d(x, y) \leq d(x, z) + d(z, y)$	triangle inequality

Metric distance functions and metric spaces

a distance function that satisfies all properties

non-negativity,

isolation,

symmetry, and

triangle inequality

is called a metric

a data space equipped with a metric function is called metric space

Distance and similarity functions

distance function $d : X \times X \to \mathbb{R}$

large for dissimilar objects

similarity function $s : X \times X \to \mathbb{R}$

large for similar objects

often similarity **s** is between 0 and 1

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

$$s(x, y) \propto e^{-d(x,y)}$$

Distance functions for real-valued vectors

- L_p norms or Minkowski distance:

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- p = 1, L_1, Manhattan (or city block) or Hamming distance:

$$L_1(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i| \right)$$

Distance functions for real-valued vectors

- L_p norms or Minkowski distance:

$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- p = 2, L_2, Euclidean distance:

$$L_2(x,y) = \left(\sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2}$$

Data structures

data matrix

$$\begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}$$

distance matrix

$$\begin{pmatrix} 0 & \dots & & & \\ d(2,1) & 0 & \dots & & \\ \vdots & & & & \\ d(n,1) & d(n,2) & \dots & d(n,n-1) & 0 \end{pmatrix}$$

Similarity functions for real-valued vectors

- Dot product or cosine similarity

$$\cos(x, y) = \frac{x \cdot y}{||x||\,||y||}$$

- Can we construct a distance function out of this?
- When use the one and when the other?

Distance functions for 0/1 data

x	0	1	0	0	1	0	0	1	0
y	1	0	0	0	0	1	0	1	1

$$L_1(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i| \right)$$

How good is Hamming distance for 0-1 vectors?

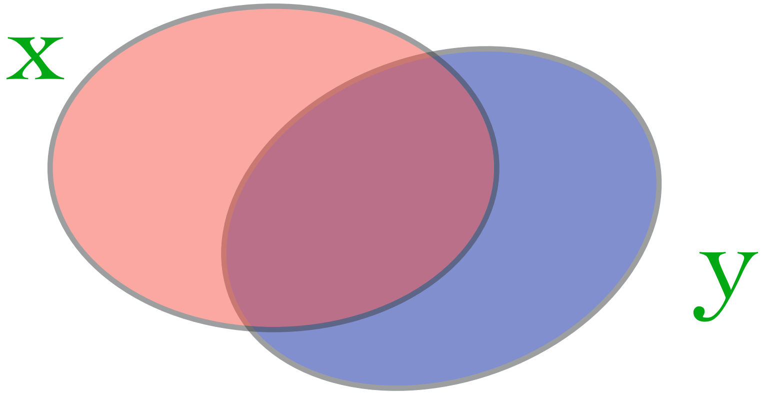
- **Drawback**

- Documents represented as sets (of words)
- Two cases
  - Two **very large** documents -- almost identical -- but for 5 terms
  - Two **very small** documents, with 5 terms each, disjoint

Distance functions for binary vectors or **sets**

- **Jaccard** similarity between binary vectors x and y (Range?)

$$\mathrm{JSim}(x,y) = \frac{|x \cap y|}{|x \cup y|}$$

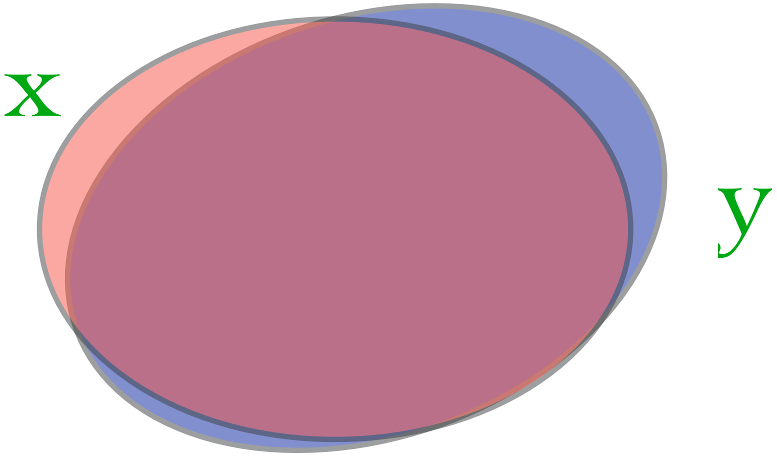


- **Jaccard** distance (Range?):

$$\mathrm{JDist}(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$$

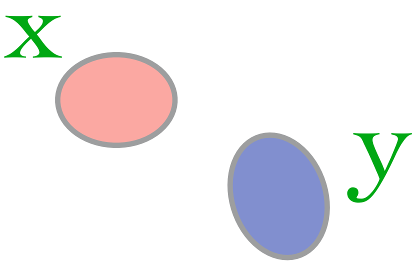
The previous example

- Case 1 (very large almost identical documents)



$J(x, y)$ almost 1

- Case 2 (small disjoint documents)



$J(x, y) = 0$

Distance functions between strings

strings x and y of equal length

modification of the Hamming distance

add 1 for all positions that are different

x	c	g	t	a	a	c	g
y	g	a	t	t	a	c	a

string Hamming distance = 4

drawbacks?

Distance functions between strings

1. strings must have equal length
2. what about

x	a	g	a	t	t	a	c
y	g	a	t	t	a	c	a

string Hamming distance = 6

String edit distance

consider two strings x and y

try to change one to another

only single-character edits are allowed

- insert character
- delete character
- substitute character

edit distance is the minimum number of such operations

not necessary to have equal length!

String edit distance

example

x   a g a t t a c

remove a

g a t t a c

add a

y   g a t t a c a

string edit distance = 2

String edit distance

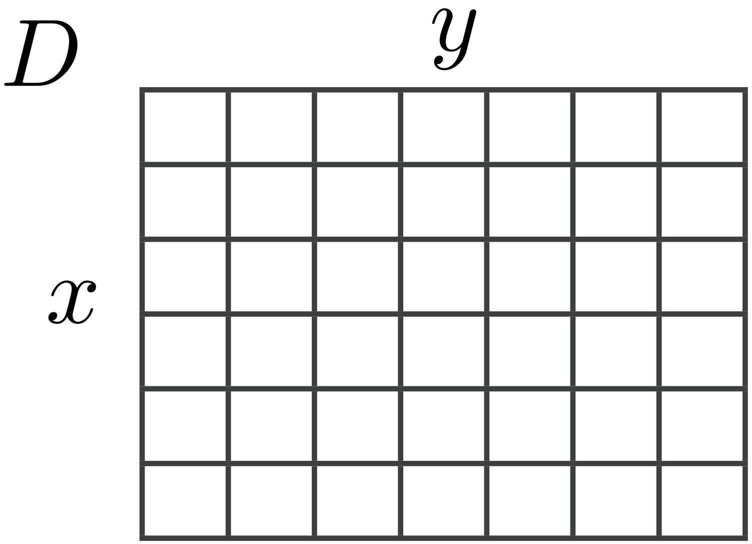
consider two strings x and y of lengths n and m, respectively

how can I compute the string edit distance between x and y?

how expensive is this computation?

Computing the edit distance

- Dynamic programming
- Form nxm distance matrix D (x of length n, y of length m)



- D(i,j) is the optimal distance between strings x[1..i] and y[1..j]

Computing the edit distance

- How to compute D(i,j)?
- Either
  - match the last two characters (substitution)
  - match by deleting the last char in one string
  - match by deleting the last character in the other string

Computing edit distance

$$\begin{aligned} D(i,j) = \min\{&D(i-1,j) + \mathrm{del}(X[i]), \\ &D(i,j-1) + \mathrm{ins}(Y[j]), \\ &D(i-1,j-1) + \mathrm{sub}(X[i],Y[j])\} \end{aligned}$$

- Running time? Metric?

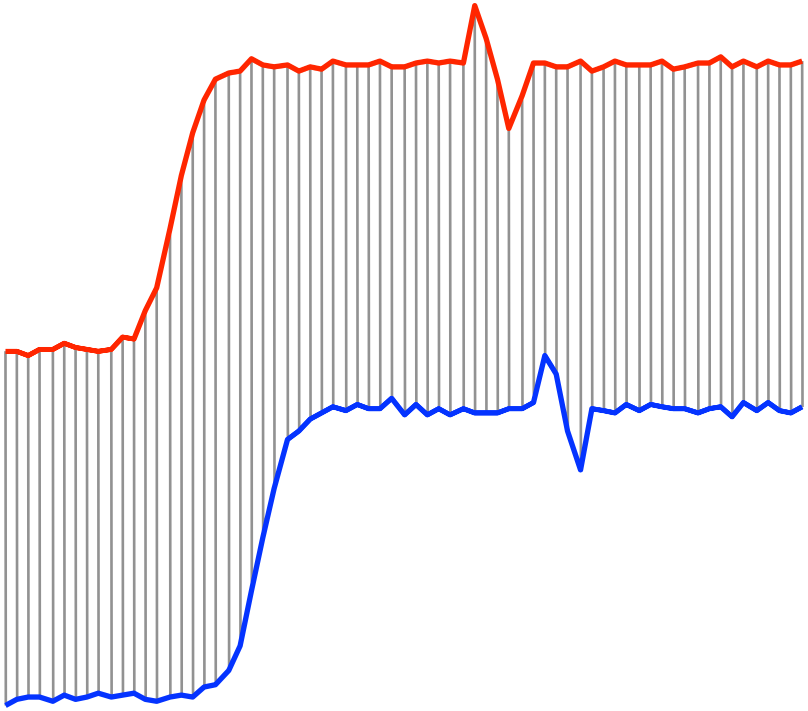
Distance function between time series

- time series can be seen as vectors
- apply existing distance metrics
- L-norms

- what can go wrong?

Distance functions between time series

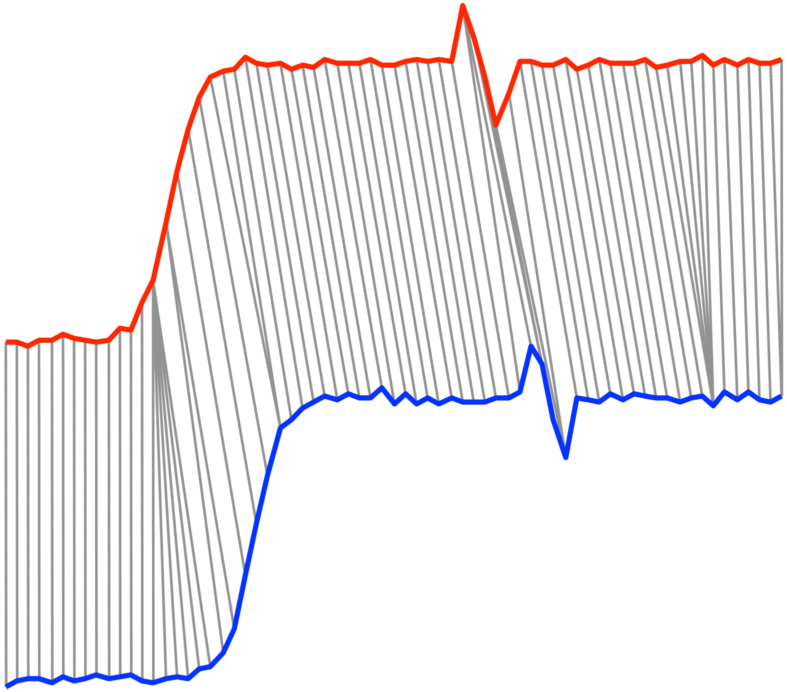
- Euclidean distance between time series



figures from Eamonn Keogh www.cs.ucr.edu/~eamonn/DTW_myths.ppt

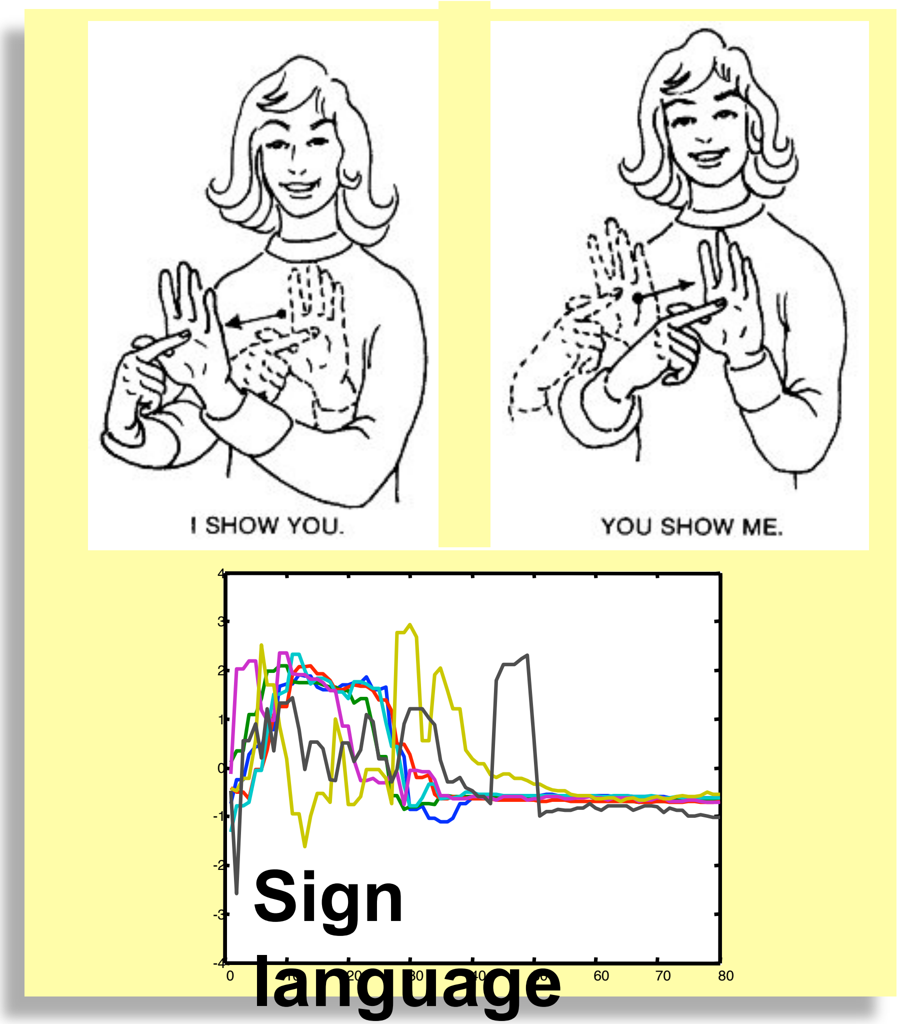
Dynamic time warping

- Alleviate the problems with Euclidean distance



figures from Eamonn Keogh www.cs.ucr.edu/~eamonn/DTW_myths.ppt

Dynamic time warping



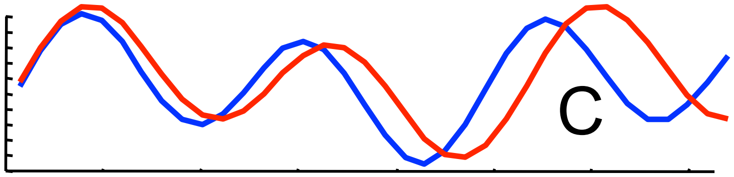
Sign language

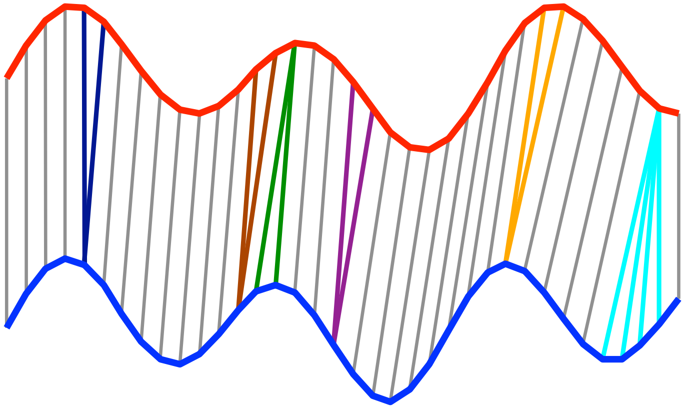
- Quite useful in practice

figures from Eamonn Keogh www.cs.ucr.edu/~eamonn/DTW_myths.ppt

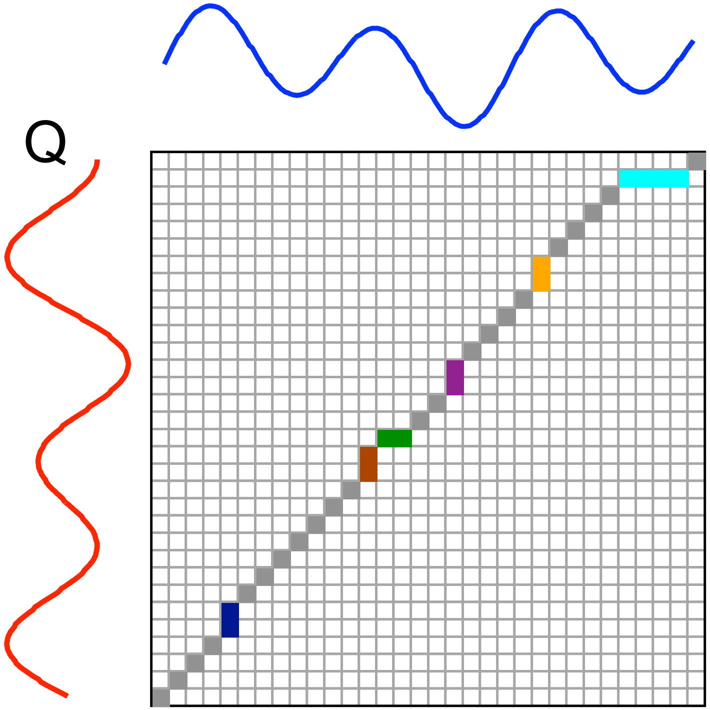
Dynamic time warping

- how to compute it?





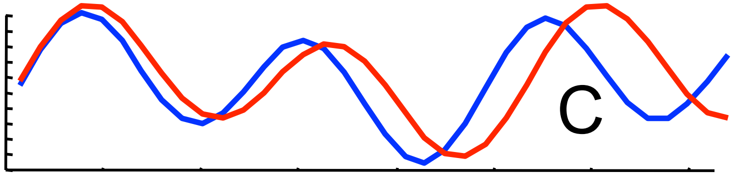
- Dynamic programming

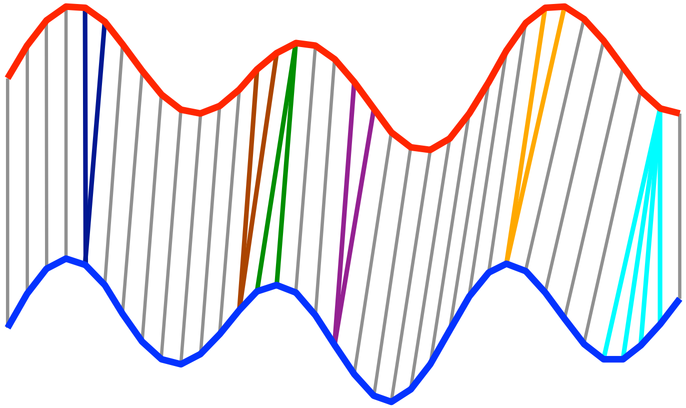


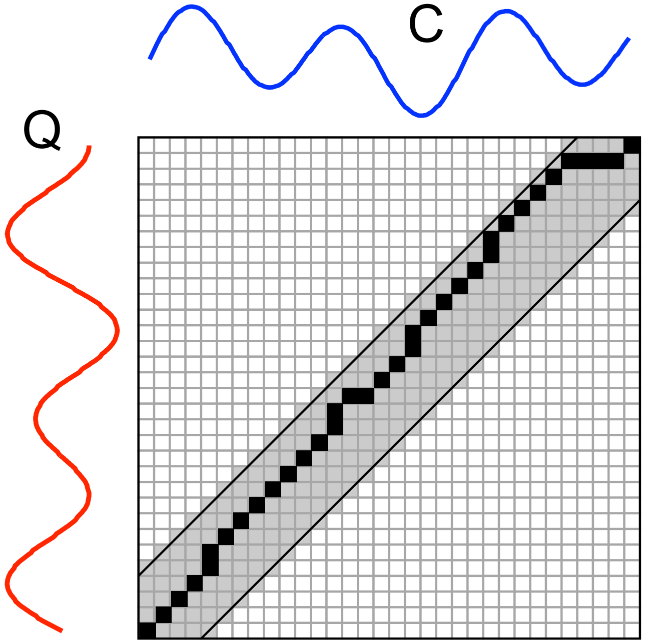
figures from Eamonn Keogh www.cs.ucr.edu/~eamonn/DTW_myths.ppt

Dynamic time warping

- constraints for more efficient computation







figures from Eamonn Keogh www.cs.ucr.edu/~eamonn/DTW_myths.ppt