Introduction to Data Exploration

Strings and Sequences

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Understand string, sequence, and time series data

Common Data Representations

Relational/Object Oriented data

Vector Space (spatial or high-dimensional) data

Strings, sequences, and time series data

Trees and graphs

Fuzzy and probabilistic data

Strings, sequences, time series

- A string or sequence, $S = (c_1, c_2, \ldots c_N)$, is a finite sequence of symbols. Here, N denotes the length of the string or sequence and c_i are from an alphabet of symbols.

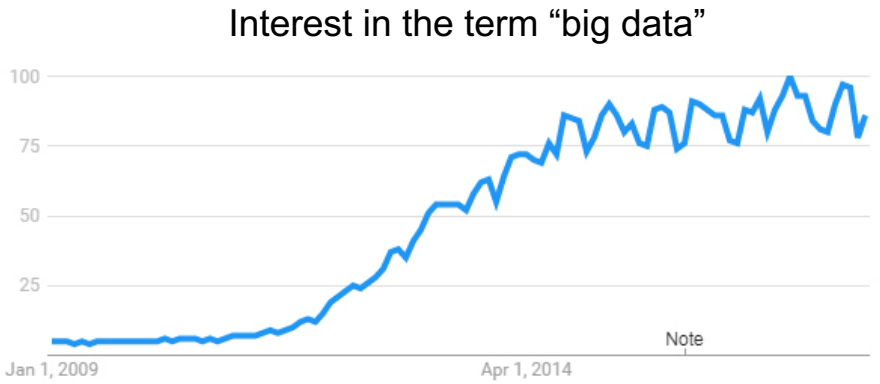
abcbbbaabbaabcbbbaaabbc

Strings, sequences, time series

- A *string or sequence*, $S = (c_1, c_2, \ldots c_N)$, *is a finite sequence of symbols*. Here, *N* denotes the length of the string or sequence and c_i are from an alphabet of symbols.

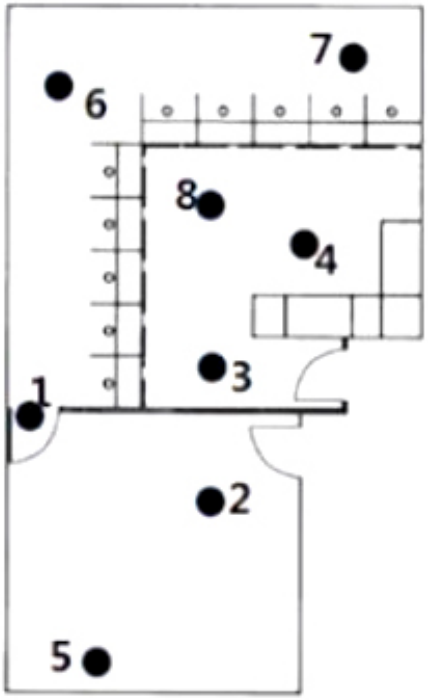
abcbbbaabbaabcbbbaaabbc

- A *time series*, $T = (d_1, d_2, \ldots, d_N)$, *is a finite sequence of data values*. Here, N denotes the length of the time series and $d_i \in R$

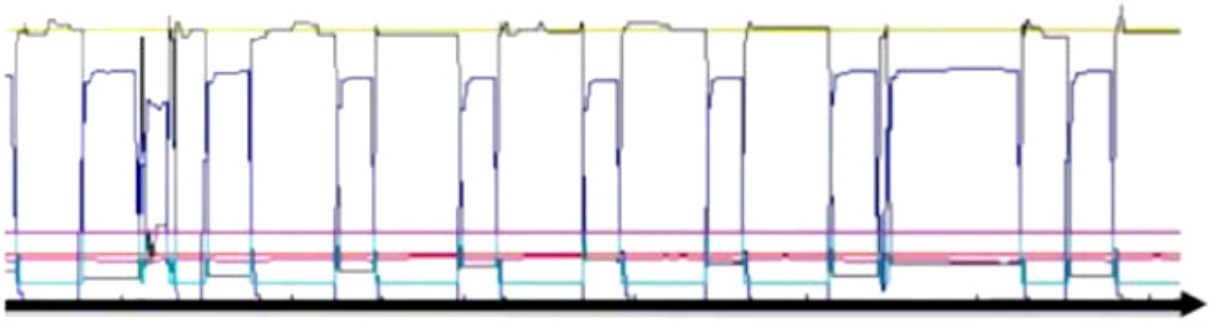


https://trends.google.com/trends/explore?date=2008-12-19%202018-01-19&q=big%20data

Multi-variate time series



Temperature recordings over a period of time



Time

Strings/sequence matching and search

Prefix search:

- Find all strings that start with “tab”
  - “table”; “tabular”; “tablet”;…

Subsequence search:

- Find all strings that contain the subsequence “ark”
  - “marketing”; “spark”; “quark”;…

Subsequence match:

- Find the longest matching subsequence between “plasticity” and “scholastic”
- Find the most frequently repeating 3 character subsequence
  - “abcbbbaabbaabcbbbaaabbc”

How similar are two strings?

- “table” vs. “cable”?
- “table” vs. “tale”?
- “table” vs. “tackle”?

Edit Distance

Edit distance between two sequences is the minimum number of edit operations needed to convert one sequence to the other:

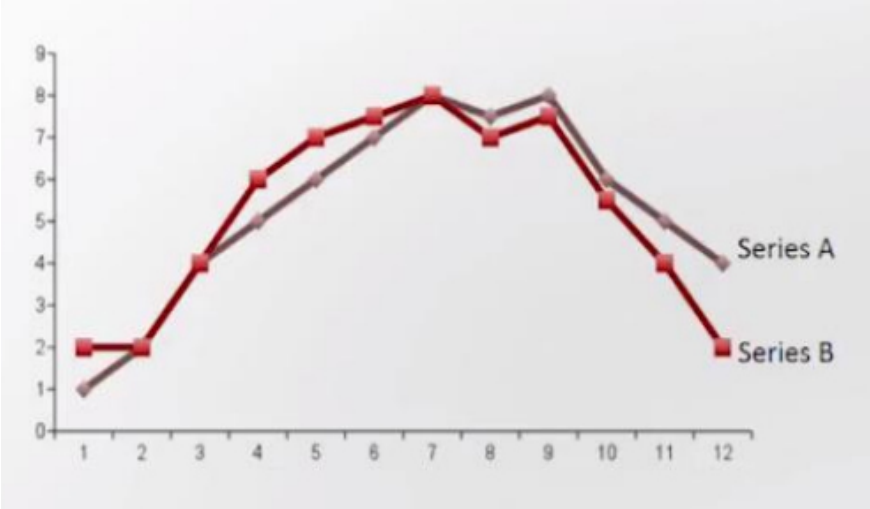
- "**table**" vs. "**cable**"
  - 1 replacement ("t" with "c")
- "**table**" vs. "**tale**"
  - 1 deletion ("b")
- "**table**" vs. "**tackle**"
  - 1 deletion ("b") and 2 insertions ("c" and "k")
  - 1 replacement ("b" with "c") and 1 insertion ("k")

Time Series Matching

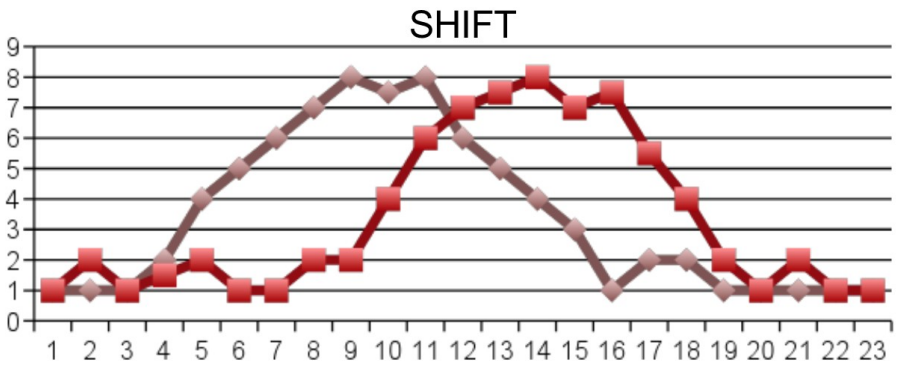
Synchronous/Non-Elastic Distance and Similarity Measures:

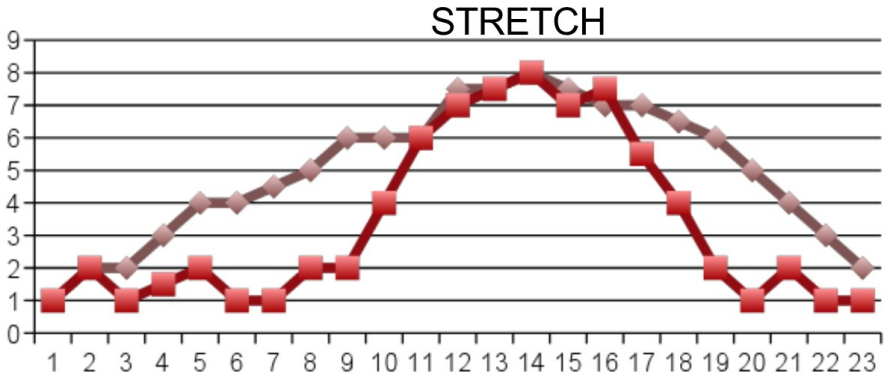
- Euclidean distance

$$\left(\sum_{i=1\ldots12} a_i^2 - b_i^2 \right)^{1/2}$$



Asynchrony in Time Series

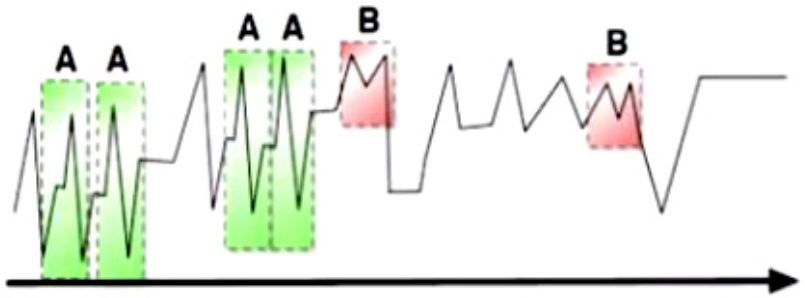


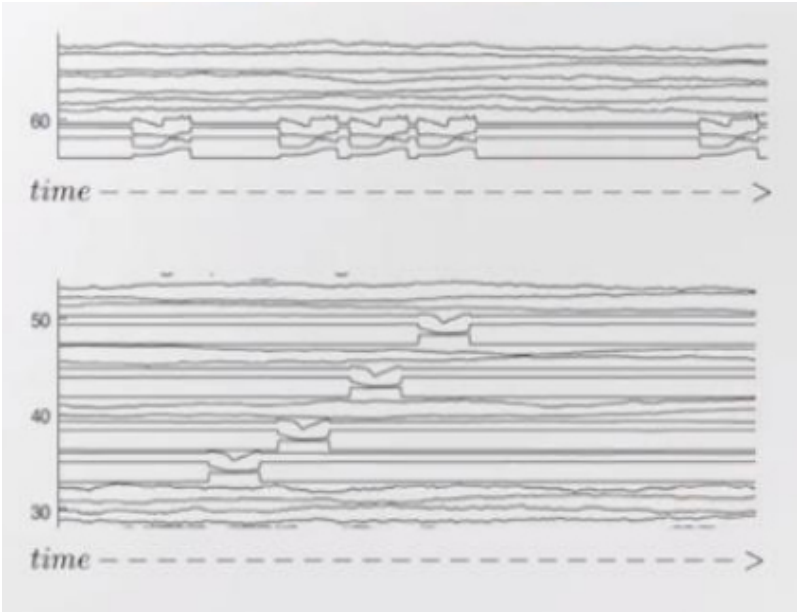


Asynchronous/Elastic Distance and Similarity Measures:

- Edit Distance, ED
- Dynamic Time Warping, DTW
- Feature-based Alignment, RMT

Motifs





- Frequently repeating patterns in time series
- Motifs can also occur in multi-variate time series