Parallel Data Series Indexing and Similarity Search on Modern Hardware

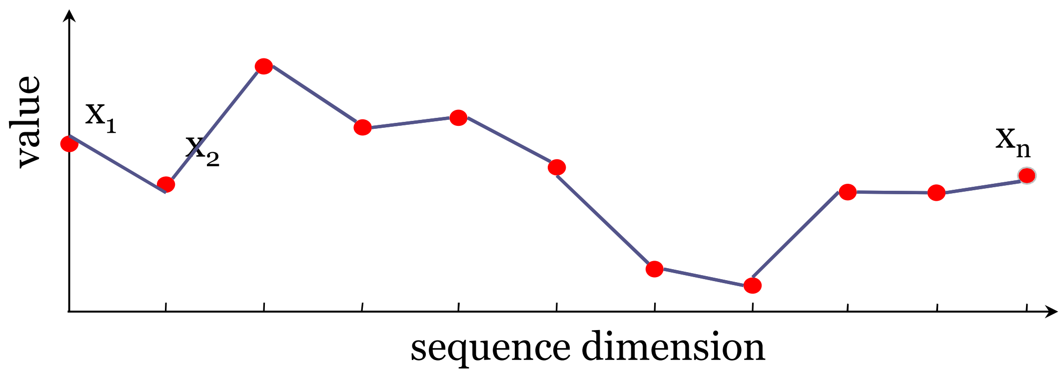
Panagiota Fatourou, Professor

University of Crete and FORTH
Laboratoire d'Informatique Paris Descartes, Université de Paris Cité

Joint work with:
Botao Peng, Chinese Academy of Sciences and **Themis Palpanas**, Université de Paris Cité

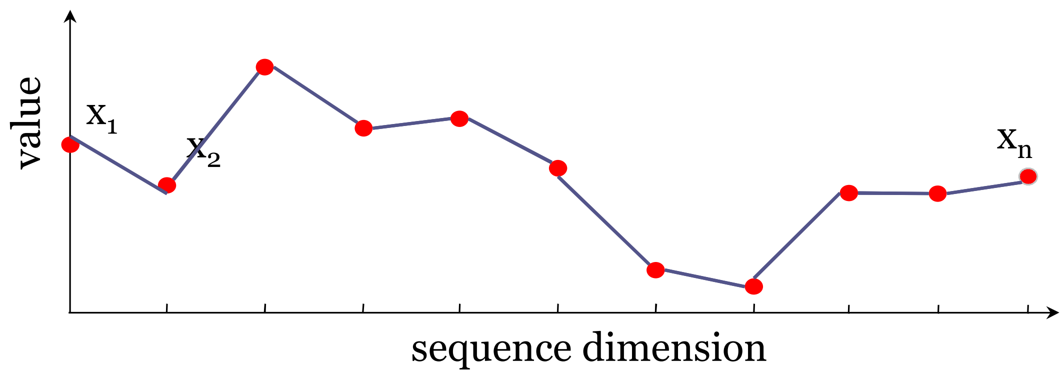
Data series

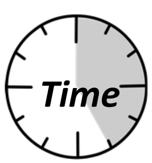
- Sequence of points ordered along some dimension



Data series

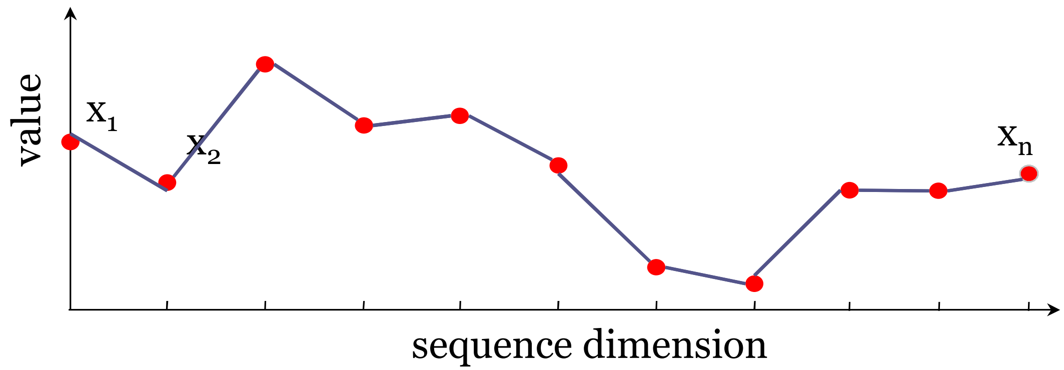
- Sequence of points ordered along some dimension

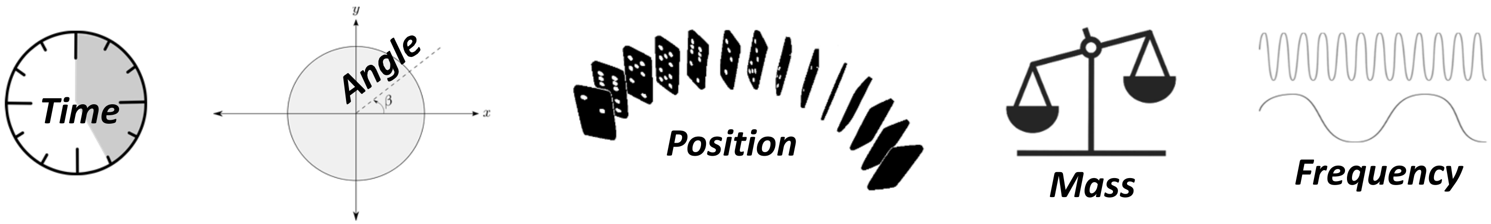




Data series

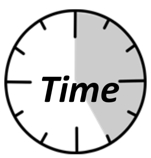
- Sequence of points ordered along some dimension

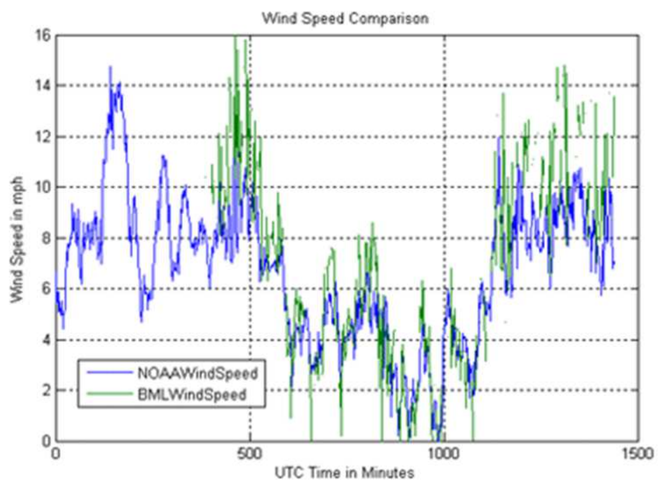




Scientific Monitoring

meteorology, oceanography, volcanology, seismology, astronomy, finance, sociology, etc.





Wind speed

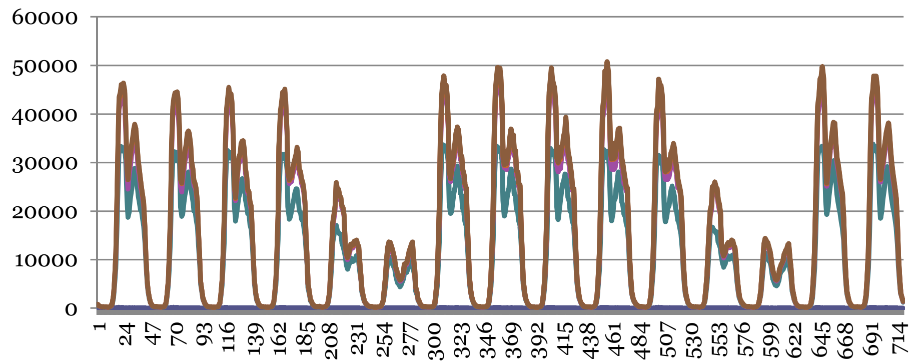
From ocean observing node project, http://bml.ucdavis.edu/boon/wind.html

Volcanic Activity Indicators

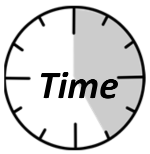
From British Geological Survey
https://www.bgs.ac.uk/geology-projects/volcanoes/

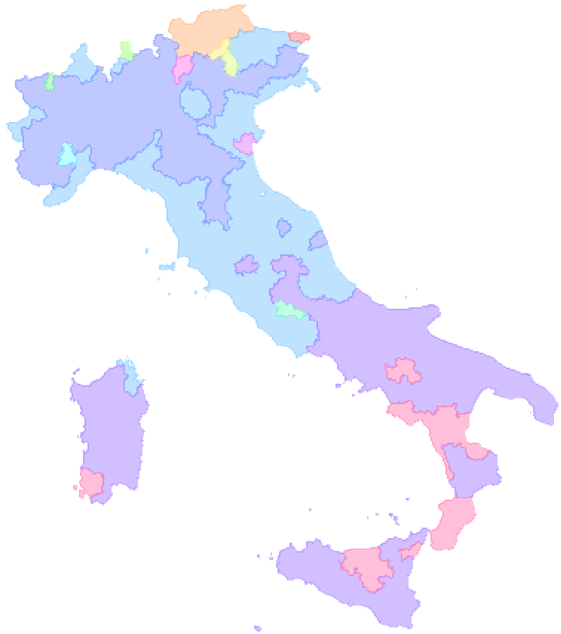
Telecommunications

- analysis of call activity patterns, Telecom Italia



average number of calls

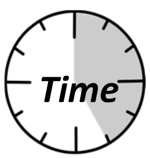


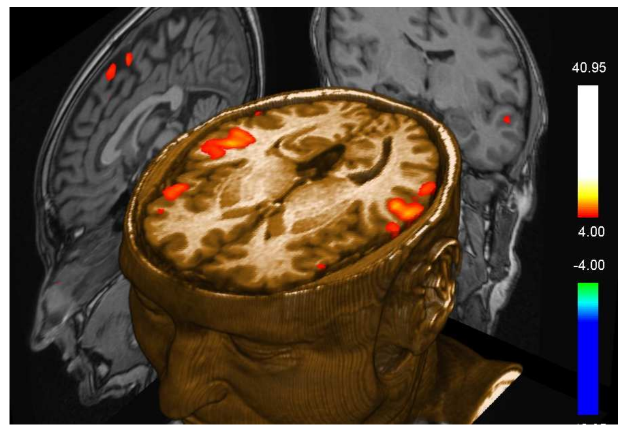


clustermap of incoming calls time series

Neuroscience

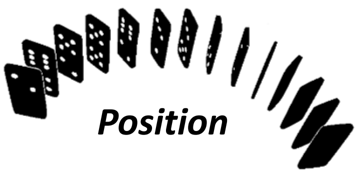
- functional Resonance Magnetic Imaging (fMRI) data
  - primary experimental tool of neuroscientists
  - reveal how different parts of brain respond to stimuli

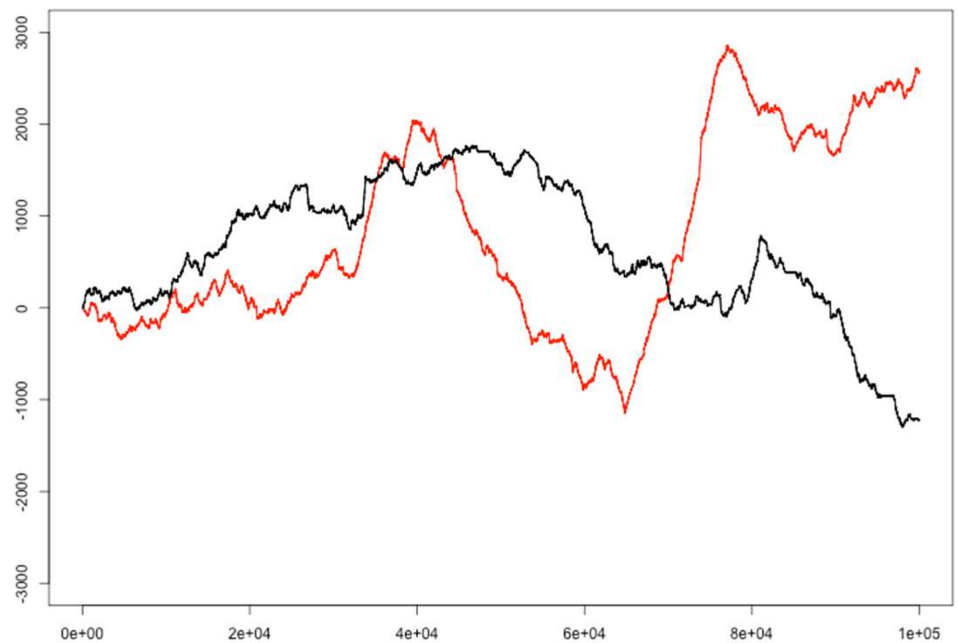




Biology

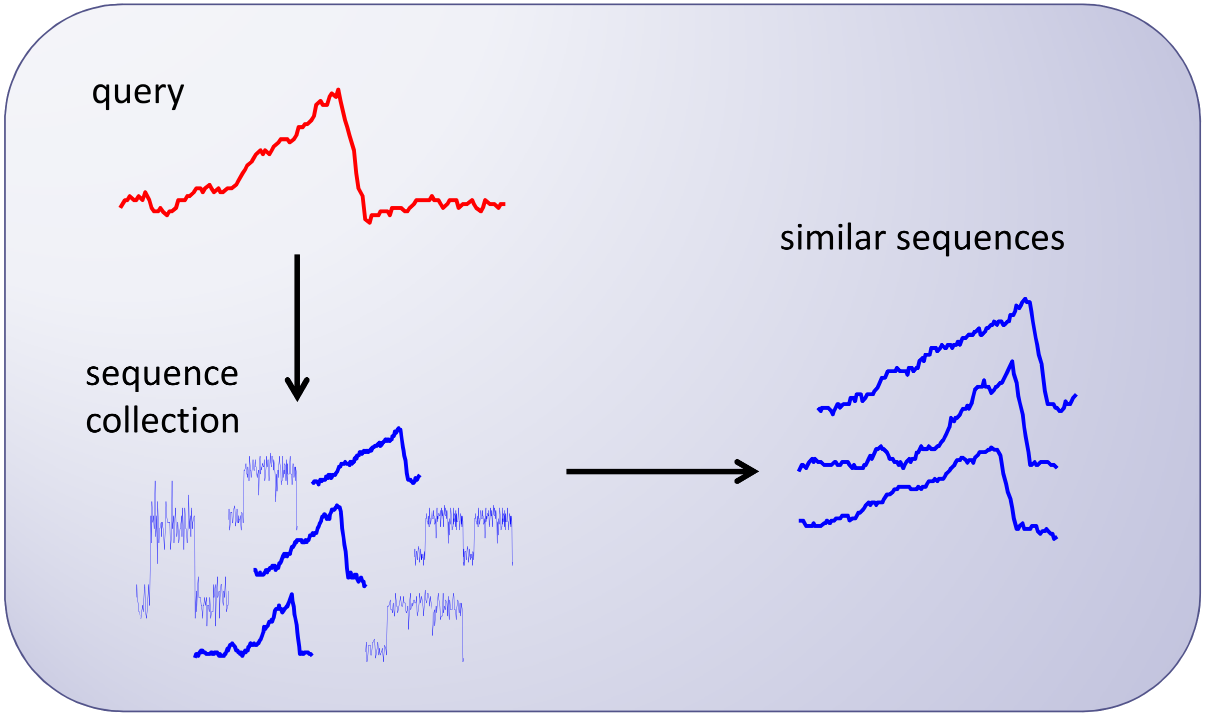
GTCAATGGCCAGGATATTAGAACAGTACTCTGTGAACCCTATTTATGGTGGCACCCCTTAG
ACTAAGA GGGAAGCCTCC
TGTAAAG AGGGAAAGCCA
CATTTGA AAGATATTACA
AACTAAG TGAATAATCCC
TGGAGGT ACCGGTTGAA
ACTGAAA AACTGGCAACT
GCCAACC CTCCCAAGGAT
CAACCTT AGGGCTTTGG
ACTGAAA AGCCAGTGACA
AATACAA TTTATGTCTTA
ACTTTTG TGGCCAAGCAT
CCAAGGT GTGGTTATTTA
TTAGTCA AGAGCAGAATC
AATGAGC GCAAGATGAG
CATTGAA CAAGAGCCCA
GGTTGTT AATGGCCAGG
ATATTAGAACAGTACTCTGTGAACCCTATTTATGGTAGCACCCCTTAGACTAAGATAACACA
GGGAGCAAGAGGTT



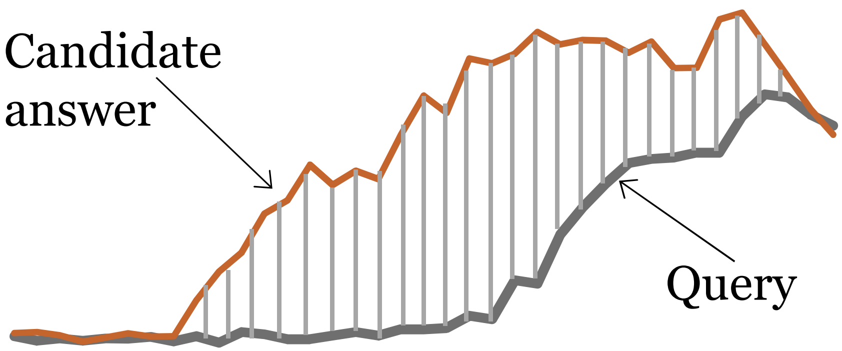


What do we want to do with data series?

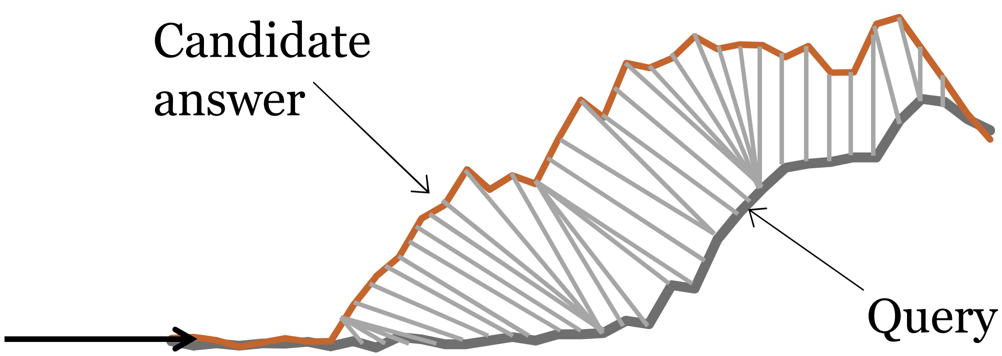
What do we want to do with data series?



What do we want to do with data series? Complex analytics



Euclidean Distance



Dynamic Time Warping

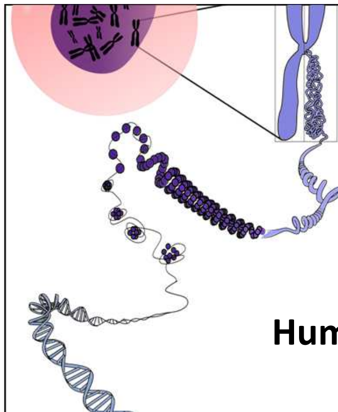
Challenge - Massive data Series collections

NASA's Solar Observatory

1.5 TB per day

Large Synoptic Survey Telescope

~30 TB per night



Human Genome project

130 TB

passenger aircrafts

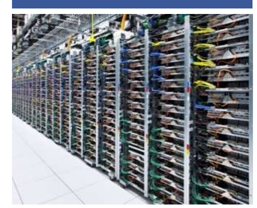
20 TB per hour

data center and services monitoring

2B data series

4M points/sec

facebook



HARD, because of very high dimensionality: each data series has several hundreds to several thousands of points!

Contributions

- **ParIS+**, a disk-based concurrent data series index for modern hardware
  - Multi-threaded design, SIMD instructions
  - Similarity search up to 15x faster (10sec on 100GB dataset)

  IEEE Trans. on Knowledge & Data Engineering 2021, IEEE Big Data 2018

- **MESSI**, an in-memory data series index for modern hardware
  - Lower synchronization index design
  - Novel algorithms for query processing
  - Similarity search at interactive speeds (50msec on 100GB dataset)

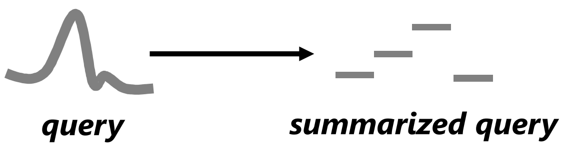
  VLDB Journal 2021, IEEE International Conference on Data Engineering (ICDE) 2020

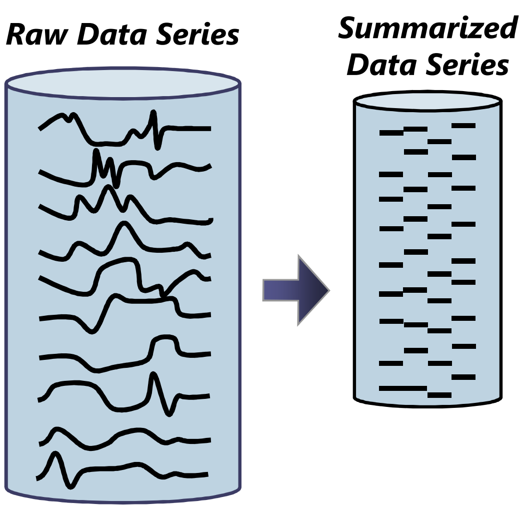
- **SING**, a data series similarity search accelerated by GPUs
  - CPU-GPU collaborative framework
  - Expands scalability of exact similarity search (~30msec on 100GB dataset)

  IEEE International Conference on Data Engineering (ICDE) 2020

State-of-the-Art: ADS, the Adaptive Data Series Index

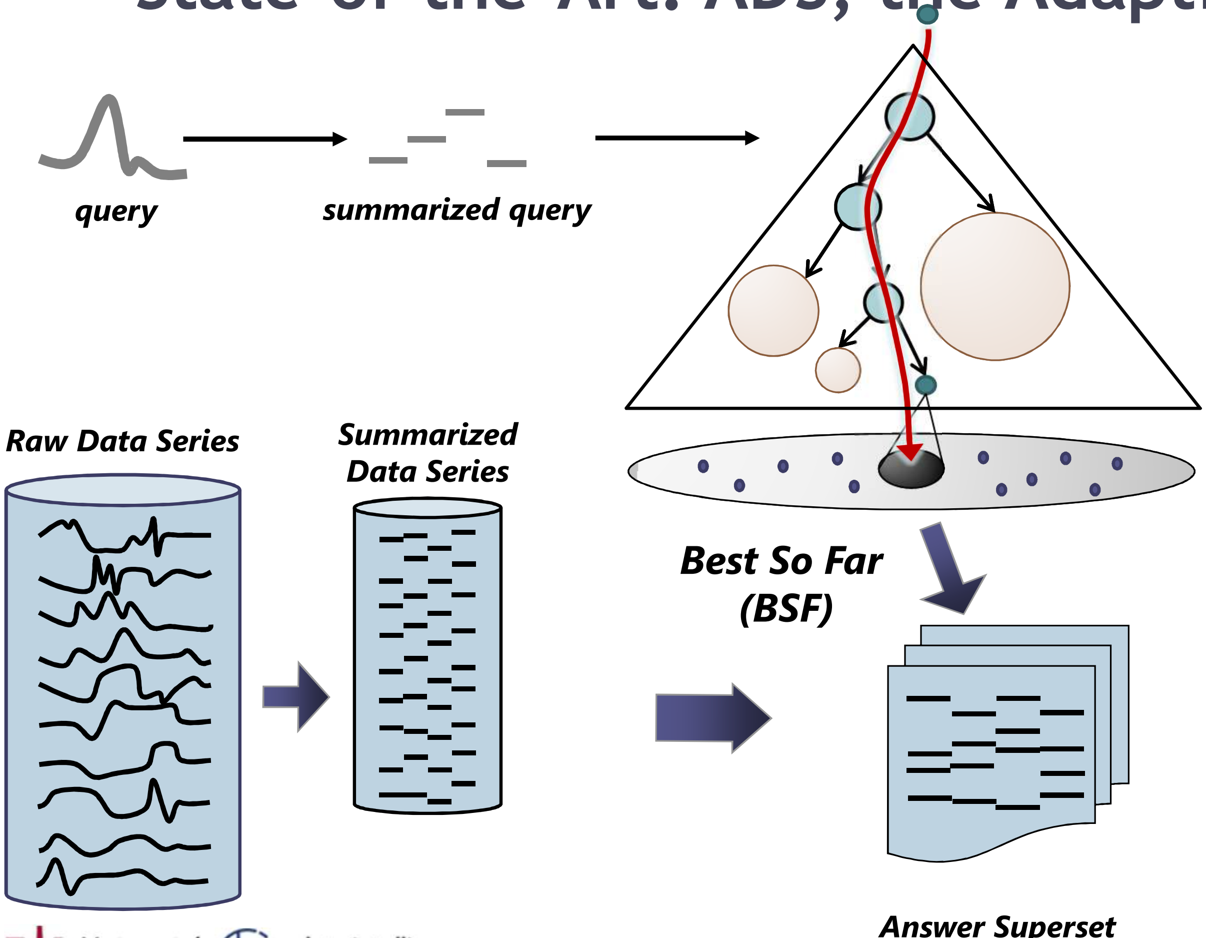
[Zoumpatianos et al., VLDBJ'16]





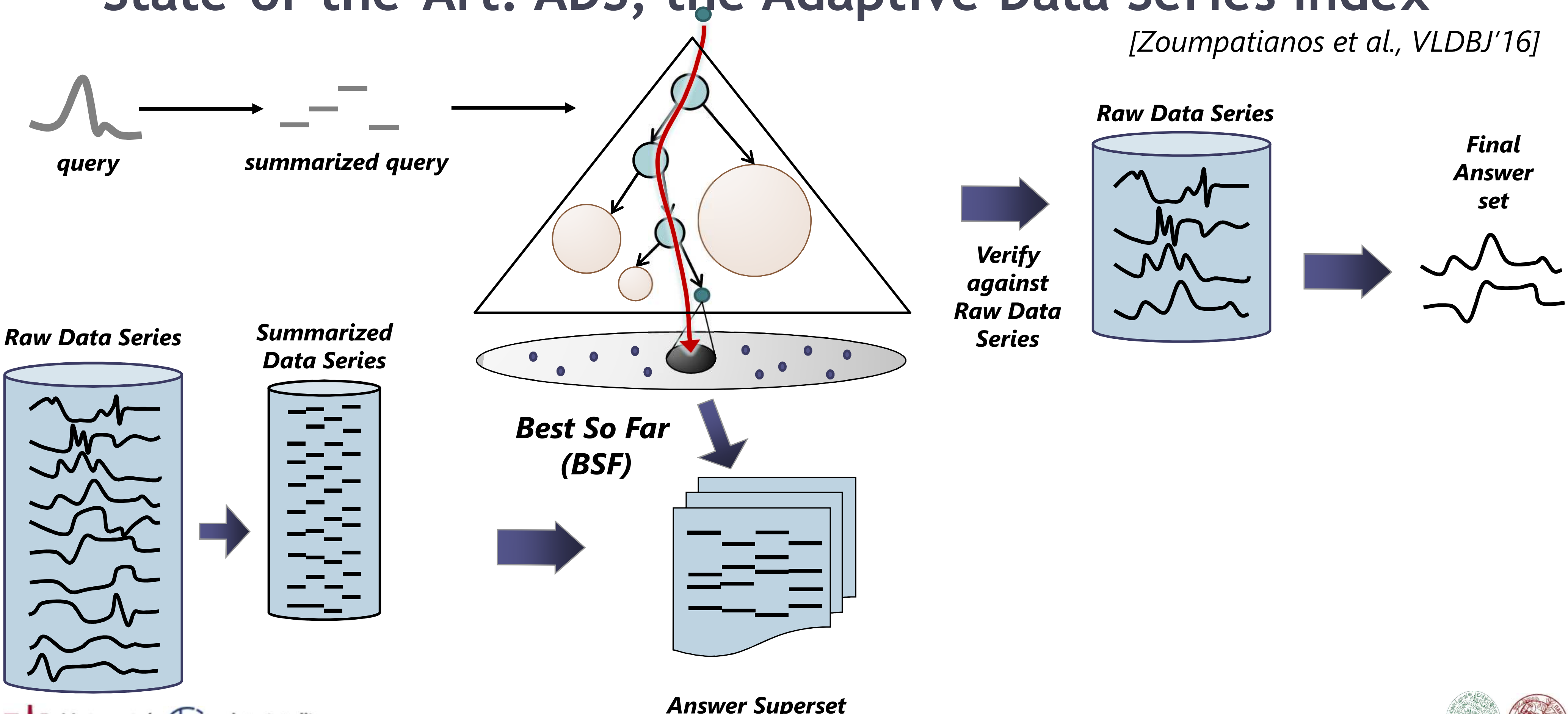
State-of-the-Art: ADS, the Adaptive Data Series Index

[Zoumpatianos et al., VLDBJ'16]



State-of-the-Art: ADS, the Adaptive Data Series Index

[Zoumpatianos et al., VLDBJ'16]



Symbolic Aggregate approXimation (SAX)

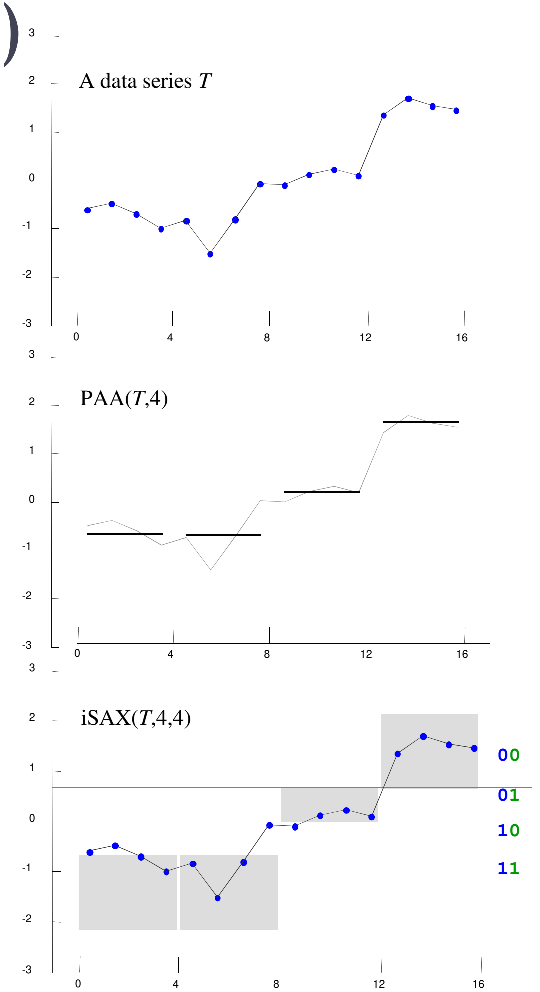
(1) Represent data series T of length $\boldsymbol{n}$ with $\boldsymbol{w}$ segments using Piecewise Aggregate Approximation (PAA)

- T typically normalized to $\mu = 0, \sigma = 1$
- PAA(T,w) = $\bar{T} = \bar{t}_1, \ldots, \bar{t}_w$

  where $\bar{t}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} T_j$

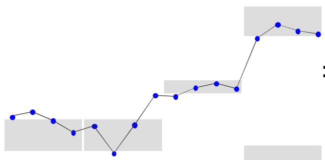
(2) Discretize into a vector of symbols

- Breakpoints map to small alphabet $\boldsymbol{a}$ of symbols

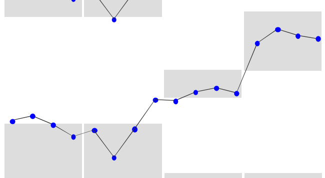


iSAX Representation

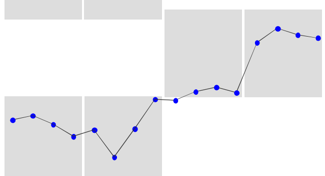
- iSAX offers a bit-aware, quantized, multi-resolution representation with variable granularity



= { 6, 6, 3, 0} = {110 ,110 ,0111 ,000}



= { 3, 3, 1, 0} = {11 ,11 ,011 ,00 }



= { 1, 1, 0, 0} = {1 ,1 ,0 ,0 }

Lower Bound distance Calculation: Calculate distance between the iSAX summary of a data series and the query PAA

Real Distance Calculation: Calculate real Euclidean distance between the query and a data series

Lower-Bound Property of iSAX summaries: If the lower bound distance between a query Q and the data series DS is higher than a value v, then the real distance between Q and DS is also higher than v.