

Analytics and Knowledge Discovery from Big Data Series Collections



Themis Palpanas



Paris Descartes University
French University Institute



ISIP– Heraklion (Greece), May 2019



1

diNo 2

References 1

• papers

- **Generating Data Series Query Workloads.** VLDBJ 2018
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb18-benchmark.pdf>
- **Coconut: A Scalable Bottom-Up Approach for Building Data Series Indexes.** PVLDB 2018
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb18-coconut.pdf>
- **Comparing Similarity Perception in Time Series Visualizations.** VIS 2018
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vis2018.pdf>
- **Data Series Management: Fulfilling the Need for Big Sequence Analytics.** ICDM 2018
 - <http://www.mi.parisdescartes.fr/~themisp/publications/icde18-sms.pdf>
- **ULISSE: Ultra compact Index for Variable-Length Similarity SEarch in Data Series.** ICDE 2018
 - <http://www.mi.parisdescartes.fr/~themisp/publications/icde18-ulissee.pdf>
- **DPiSAX: Massively Distributed Partitioned iSAX.** ICDM 2017
 - <http://www.mi.parisdescartes.fr/~themisp/publications/icdm17-dpisax.pdf>
- **ADS: The Adaptive Data Series Index.** PVLDBJ 2016
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb16-ads.pdf>
- **Big Sequence Management: A Glimpse on the Past, the Present, and the Future.** LNCS, 2016
 - <http://www.mi.parisdescartes.fr/~themisp/publications/sofsem16-bisem.pdf>
- **Query Workloads for Data-Series Indexes.** KDD 2015
 - <http://www.mi.parisdescartes.fr/~themisp/publications/kdd15-bends.pdf>
- **RINSE: Interactive Data Series Exploration.** PVLDB 2015
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb15-rinse.pdf>
- **Indexing for Interactive Exploration of Big Data Series.** SIGMOD 2014
 - <http://www.mi.parisdescartes.fr/~themisp/publications/sigmod14-ads.pdf>
- **Beyond One Billion Time Series: Indexing and Mining Very Large Time Series Collections with iSAX2+.** KAIS 2014
 - <http://www.mi.parisdescartes.fr/~themisp/publications/kais14-isax2plus.pdf>
- **iSAX 2.0: Indexing and Mining One Billion Time Series.** ICDM 2010
 - <http://www.mi.parisdescartes.fr/~themisp/publications/icdm10-billiontimeseries.pdf>

Themis Palpanas - ISIP - May 2019

2

References 2

- papers
 - **Coconut Palm: Static and Streaming Data Series Exploration Now in your Palm.** SIGMOD 2019
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb19-lernaeanhydra.pdf>
 - **The Lernaean Hydra of Data Series Similarity Search: An Experimental Evaluation of the State of the Art.** PVLDB 2019
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb19-lernaeanhydra.pdf>
 - **Scalable, Variable-Length Similarity Search in Data Series: The ULISSE Approach.** PVLDB 2019
 - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb19-ulisse.pdf>
 - **Progressive Similarity Search on Time Series Data.** BigVis 2019
 - <http://helios.mi.parisdescartes.fr/~themisp/publications/bigvis19.pdf>
 - **PARIS: The Next Destination for Fast Data Series Indexing and Query Answering.** IEEE BigData 2018
 - <http://helios.mi.parisdescartes.fr/~themisp/publications/bigdata18.pdf>
 - **Massively Distributed Time Series Indexing and Querying.** TKDE 2018
 - <http://helios.mi.parisdescartes.fr/~themisp/publications/tkde18-dpisax.pdf>

Themis Palpanas - ISIP - May 2019

3

References

- code and datasets
 - **Coconut:** <https://github.com/kono925/coconut>
 - **ULISSE:** <http://helios.mi.parisdescartes.fr/~mlinardi/ULISSE.html>
 - **DPiSAX:** <https://github.com/DPiSAX/DPiSAX.github.io>
 - **ADS:** <https://github.com/zoumpatianos/ADS>
 - **iSAX2+:** <http://www.mi.parisdescartes.fr/~themisp/isax2plus/>
 - **PARIS:** <http://helios.mi.parisdescartes.fr/~themisp/paris/>
- data series toolbox
 - **DSStat:** <https://github.com/zoumpatianos/DSStat>
- demos
 - **Coconut Palm:** http://users.ics.forth.gr/~kondylak/Coconut_Palm/Coconut_Palm.html
 - **RINSE:** <http://daslab.seas.harvard.edu/rinse/>
- nestor project
 - <http://nestordb.com>

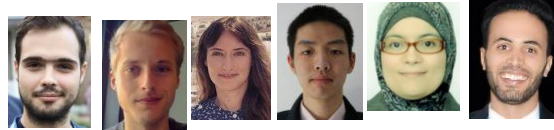
Themis Palpanas - ISIP - May 2019

4

Acknowledgements

Paris Descartes University

- Michele Linardi
- Anna Gogolou
- Botao Peng
- Karima Echihabi
- Paul Boniol
- Federico Roncallo



Harvard University

- Kostas Zoumpatianos
- Stratos Idreos
- Niv Dayan



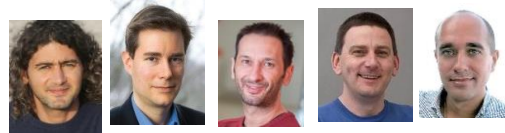
Cornell University/Microsoft

- Yin Lou
- Johannes Gehrke



Inria

- Anastasia Bezerianos
- Fanis Tsandilas
- Djamel-Edine Yagoubi
- Reza Akbarinia
- Florent Massegla



University of California at Riverside

- Jin Shieh
- Eammon Keogh

University of Crete/FORTH

- Haridimos Kondylakis
- Panagiota Fatourou

University of Trento

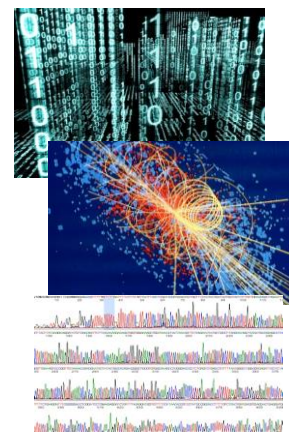
- Alessandro Camera

Themis Palpanas - ISIP - May 2019

5

Executive Summary

- data collected at unprecedented rates
- they enable data-driven scientific discovery
- lots of these data are sequences
 - takes **days-weeks** to analyze big sequence collections



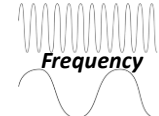
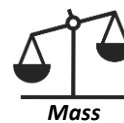
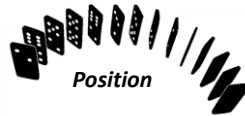
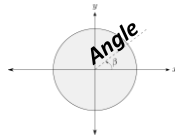
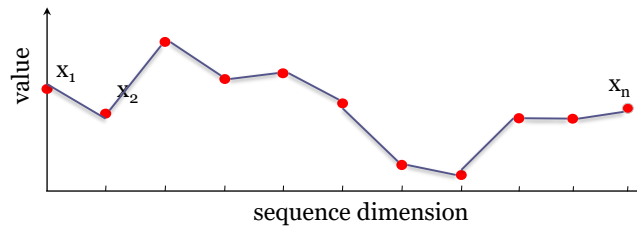
goal: **analyze big sequences** in **minutes/seconds**

Themis Palpanas - ISIP - May 2019

6

Data series

- Sequence of points ordered along some dimension

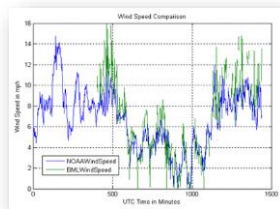


Themis Palpanas - ISIP - May 2019

7

Scientific Monitoring

- meteorology, oceanography, astronomy, finance, sociology, ...



Wind speed

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



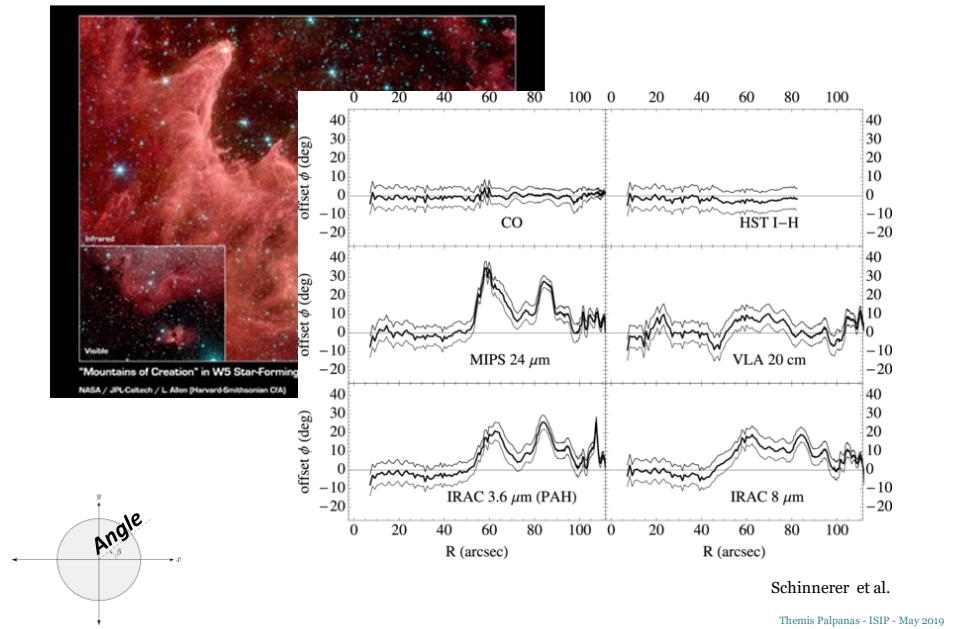
Historical stock quotes

http://money.cnn.com/2012/04/23/markets/walmart_stock/index.htm

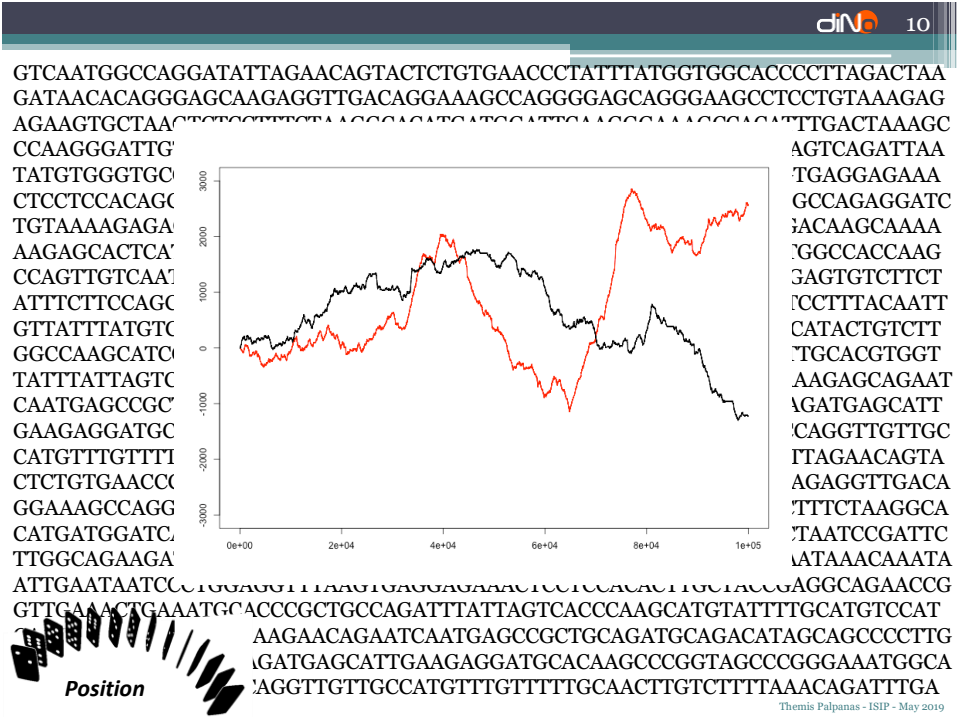


Themis Palpanas - ISIP - May 2019

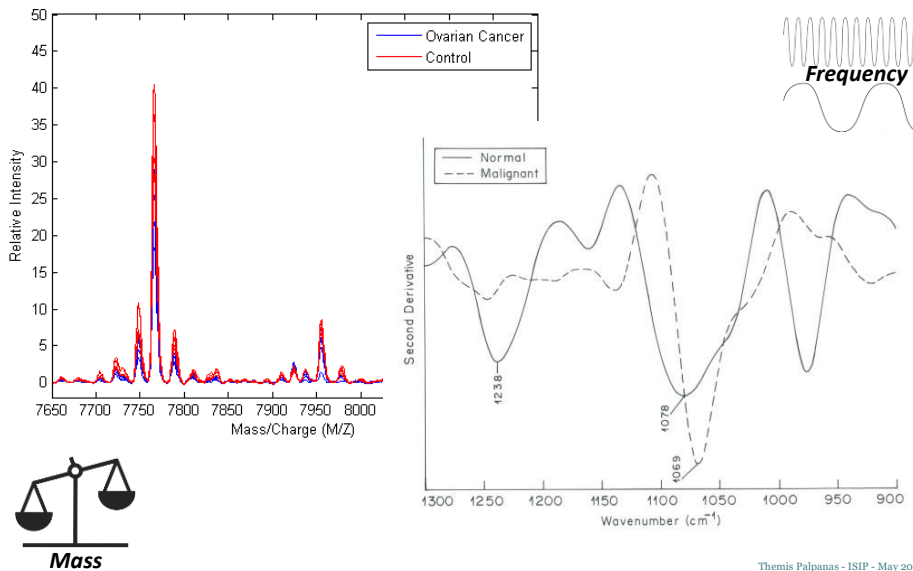
8



9



10



11

What do we want to do with them?
- simple query answering

select values
in time
interval

select values
in some
range

select some
data series

combinations
of those

12

What do we want to do with them?

- simple query answering

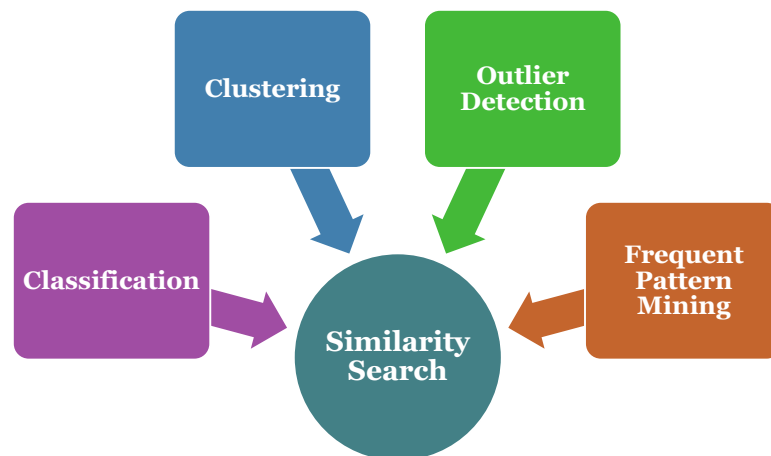
- a solved(?) problem
 - your favorite DBMS
 - ...
 - InfluxData
 - kx
 - Riak TS
 - OpenTSDB
 - TimescaleDB
 - KairosDB
 - Druid
 - ...

Themis Palpanas - ISIP - May 2019

13

What do we want to do with them?

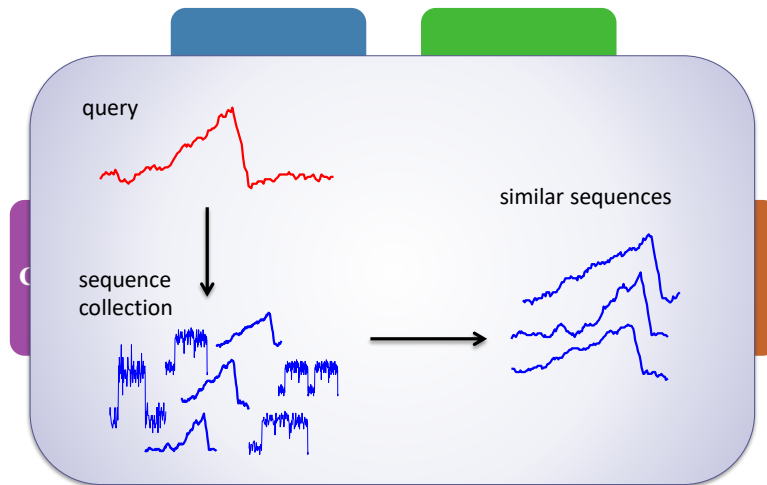
- complex analytics



Themis Palpanas - ISIP - May 2019

14

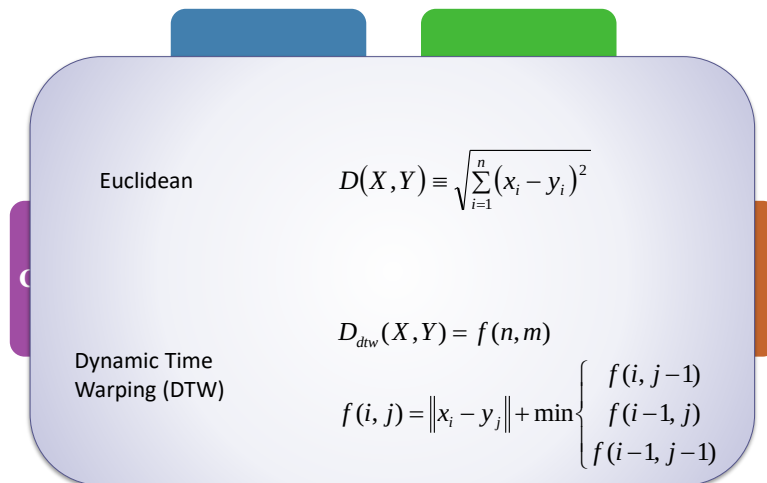
What do we want to do with them?
- complex analytics



Themis Palpanas - ISIP - May 2019

15

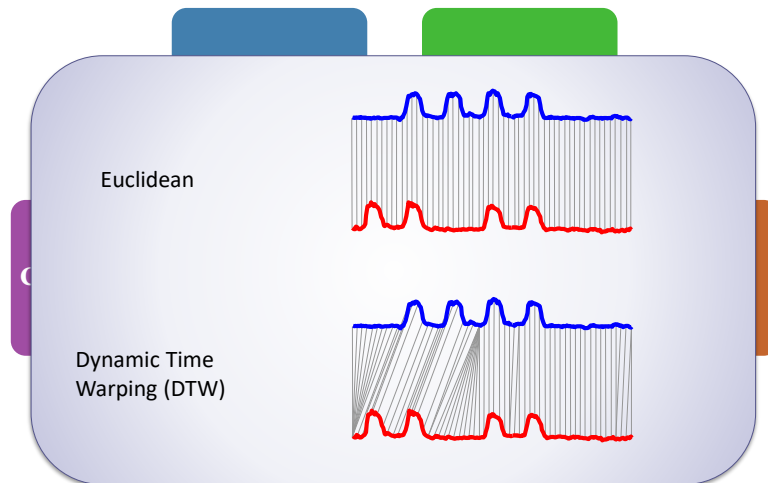
What do we want to do with them?
- complex analytics



Themis Palpanas - ISIP - May 2019

16

What do we want to do with them?
- complex analytics



Themis Palpanas - ISIP - May 2019

17

What do we want to do with them?
- complex analytics

Clustering

Outlier
Detection

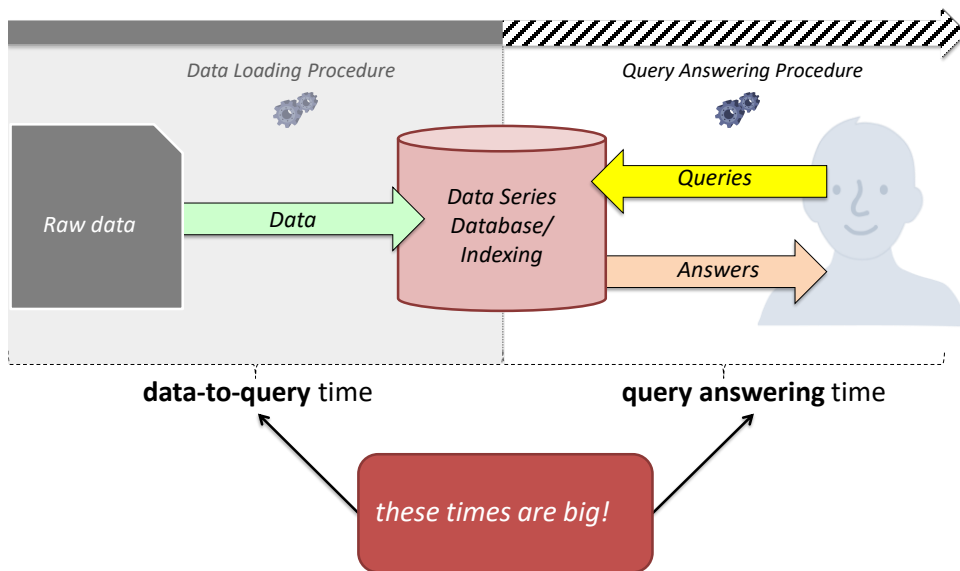
HARD, because of **very high dimensionality:
each data series has 100s-1000s of points!**

even HARDER, because of **very large size:
millions to billions of data series (multi-TBs)!**

Themis Palpanas - ISIP - May 2019

18

Query answering process

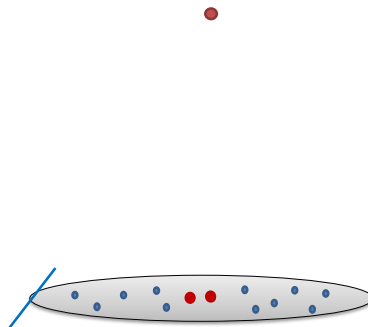


Themis Palpanas - ISIP - May 2019

19

19

Similarity Search via Serial Scan

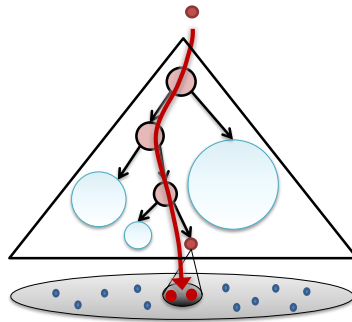


Themis Palpanas - ISIP - May 2019

20

20

Similarity Search via Indexing



Themis Palpanas - ISIP - May 2019

21

21

Traditional Approaches

answer **nearest neighbor queries** on a **1TB** dataset:

serial scan takes 45 minutes/query

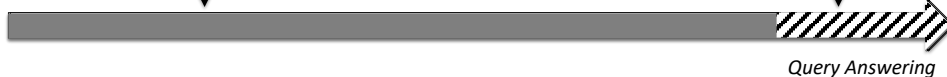


but building the index takes **too long!**

indexing a 1TB dataset takes days



a **data series index** can
reduce querying time



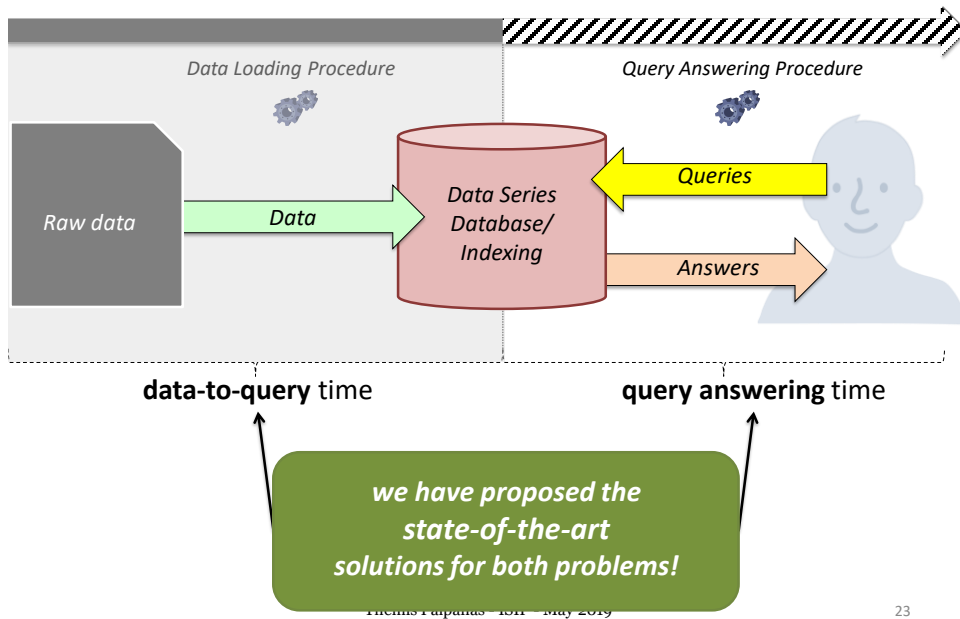
complex analytics in days...

Themis Palpanas - ISIP - May 2019

22

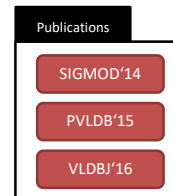
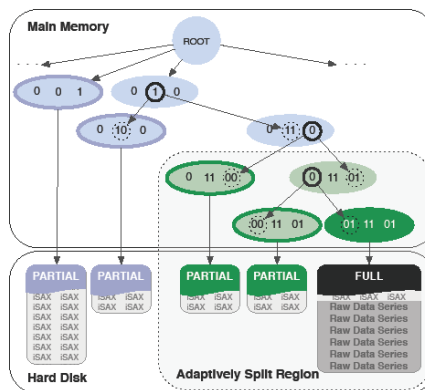
22

Query answering process



23

State of the Art Approach: ADS+



complex analytics in hours!

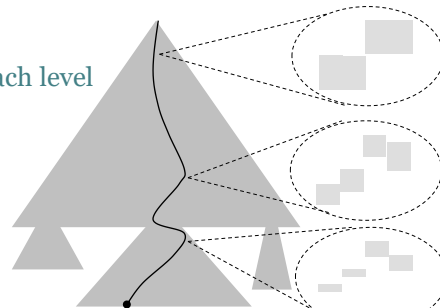
Themis Palpanas - ISIP - May 2019

24

24

iSAX Index

- non-balanced tree-based index with non-overlapping regions, and controlled fan-out rate
 - base cardinality b (optional), segments w , threshold th
 - hierarchically subdivides SAX space until num. entries $\leq th$
- Approximate Search
 - Match iSAX representation at each level
- Exact Search
 - Leverage approximate search
 - Prune search space
 - Lower bounding distance



Themis Palpanas - ISIP - May 2019

25

Drawback of iSAX2+

- cannot start answering queries until *entire* index is built!

Themis Palpanas - ISIP - May 2019

26

Adaptive Data Series Index: ADS+

- **novel paradigm** for building a data series index
 - do not build entire index and then answer queries
 - start answering queries by building the part of the index needed by those queries
- still guarantee **correct answers**

Themis Palpanas - ISIP - May 2019

27

Adaptive Data Series Index: ADS+

- intuition for proposed solution
 - build the iSAX index using the iSAX representations
 - just like iSAX2+
 - but start with a large leaf size
 - minimize initial cost
 - postpone leaf materialization to query time
 - only materialize (at query time) leaves needed by queries
 - parts that are queried more are refined more
 - use smaller leaf sizes (reduced leaf materialization and query answering costs)

Publications

SIGMOD'14

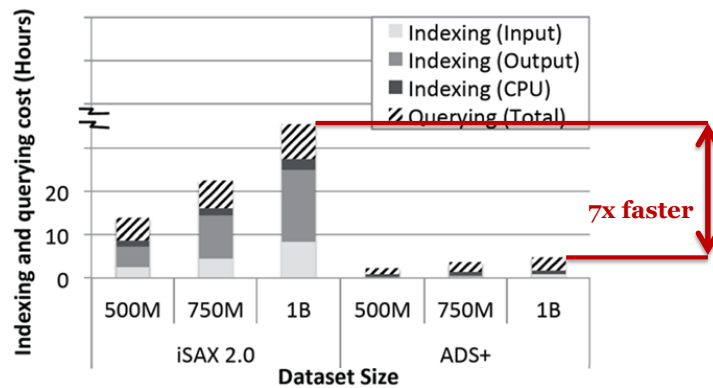
PVLDB'15

VLDBJ'16

Themis Palpanas - ISIP - May 2019

28

Experimental Evaluation



- iSAX 2.0 needs more than 35 hours to answer 100K approximate queries
- ADS+ answers 100K approximate queries in less than 5 hours

Themis Palpanas - ISIP - May 2019

29

Parallelization/Distribution?

- discussion so far assumed serial execution in a single core
 - focus on efficient resource utilization
 - squeeze the most out of a single core
 - produce scalable solutions at lowest possible cost
 - also suitable for analysts with no access to/expertise for clusters

Themis Palpanas - ISIP - May 2019

30

Need for Parallelization/Distribution

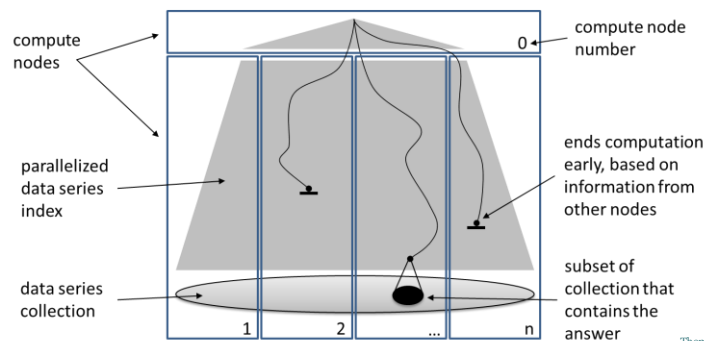
- take advantage of modern hardware!
 - Single Instruction Multiple Data (SIMD)
 - natural for data series operations
 - multi-tier CPU caches
 - design data structures aligned to cache lines
 - multi-core and multi-socket architectures
 - use parallelism inside each computation server
 - Graphics Processing Units (GPUs)
 - propose massively parallel techniques for GPUs
 - new storage solutions: SSDs, NVRAM
 - develop algorithms that take these new characteristics/tradeoffs into account
 - compute clusters
 - distribute operation over many machines

Themis Palpanas - ISIP - May 2019

31

Need for Parallelization/Distribution

- further scale-up and scale-out possible!
 - techniques inherently parallelizable
 - across cores, across machines



Themis Palpanas - ISIP - May 2019

32

Need for Parallelization/Distribution

Publications

HPCS'17

- further scale-up and scale-out possible!
 - techniques inherently parallelizable
 - across cores, across machines
- more involved solutions required when optimizing for energy
 - reducing execution time is relatively easy
 - minimizing total work (energy) is more challenging

Themis Palpanas - ISIP - May 2019

33

Need for Parallelization/Distribution

Publications

ICDM'17

TKDE'18

- **DPiSAX**: current solution for distributed processing (Spark)
 - balances work of different worker nodes

Themis Palpanas - ISIP - May 2019

34

Need for Parallelization/Distribution

Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)
 - balances work of different worker nodes
 - performs 2 orders of magnitude faster than centralized solution
- **ParIS**: current solution for modern hardware
 - completely masks out the CPU cost

Themis Palpanas - ISIP - May 2019

37

Need for Parallelization/Distribution

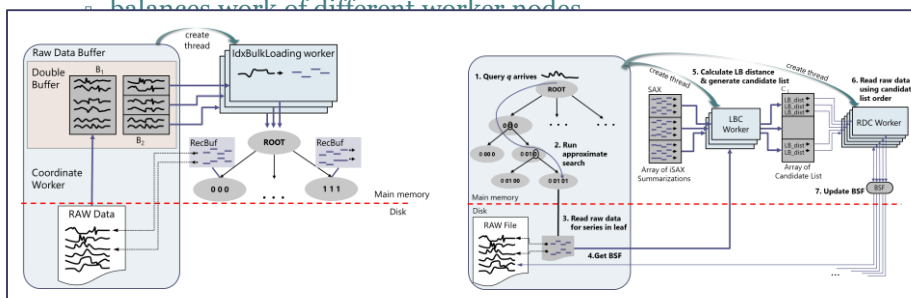
Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)
 - balances work of different worker nodes



Themis Palpanas - ISIP - May 2019

38

Need for Parallelization/Distribution

Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)
 - balances work of different worker nodes
 - performs 2 orders of magnitude faster than centralized solution
- **ParIS**: current solution for modern hardware
 - masks out the CPU cost
 - answers exact queries in the order of ms
 - 3 orders of magnitude faster than single-core solutions

Themis Palpanas - ISIP - May 2019

39

Need for Parallelization/Distribution

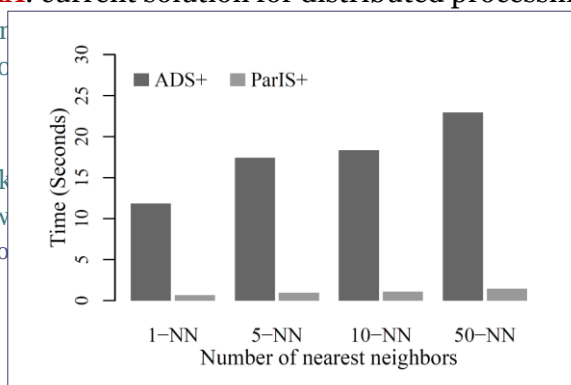
Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)
 - balanc
 - perform
- **ParIS**:
 - mask
 - answer
 - 3 o



Themis Palpanas - ISIP - May 2019

40

Need for Parallelization/Distribution

Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)

- balanc

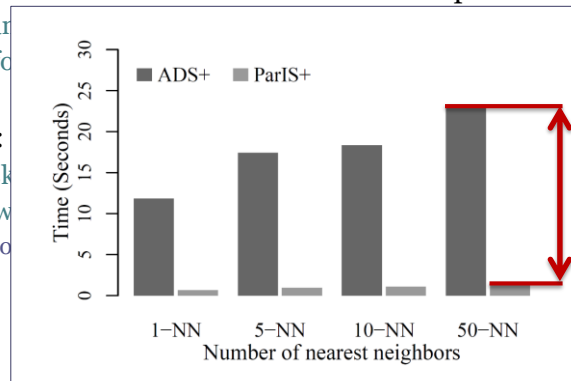
- perform

- **ParIS**:

- mask

- answer

- 30



d solution

18x faster

Themis Palpanas - ISIP - May 2019

41

Need for Parallelization/Distribution

Publications

ICDM'17

TKDE'18

BigData'18

- **DPiSAX**: current solution for distributed processing (Spark)

classifying 100K objects using a 100GB dataset goes down from **several days to **few hours**!**

- balanc

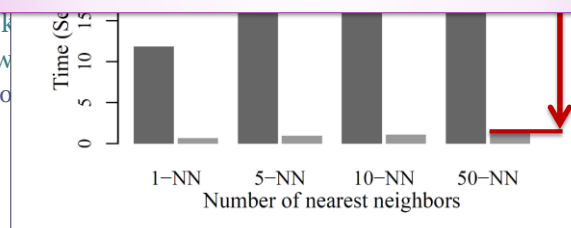
- perform

- **ParIS**:

- mask

- answer

- 30



18x faster

Themis Palpanas - ISIP - May 2019

42

Interactive Analytics?

- data series analytics is **computationally expensive**
 - very high inherent complexity
- may not always be possible to remove delays
 - but could try to hide them!

Themis Palpanas - ISIP - May 2019

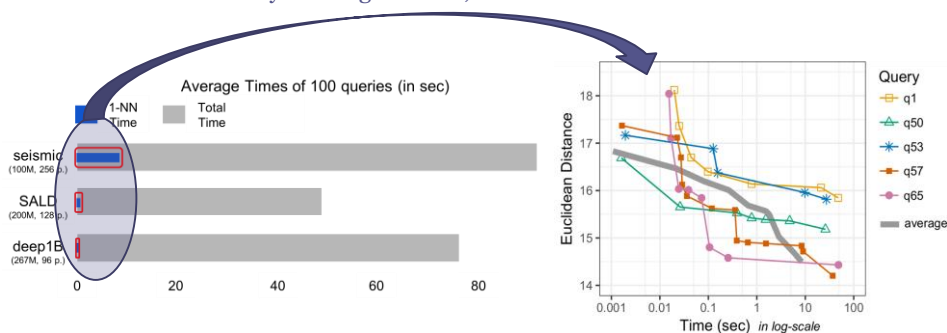
43

Need for Interactive Analytics

Publications

BigVis'19

- interaction with users offers **new opportunities**
 - **progressive answers**
 - produce intermediate results
 - iteratively converge to final, correct solution



Themis Palpanas - ISIP - May 2019

44

Need for Interactive Analytics

Publications

BigVis'19

- interaction with users offers **new opportunities**
 - **progressive** answers
 - produce intermediate results
 - iteratively converge to final, correct solution
 - provide bounds on the errors (of the intermediate results) along the way

Themis Palpanas - ISIP - May 2019

45

Need for Interactive Analytics

Publications

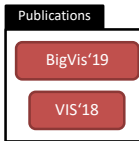
BigVis'19

- interaction with users offers **new opportunities**
 - **progressive** answers
 - produce intermediate results
 - iteratively converge to final, correct solution
 - provide bounds on the errors (of the intermediate results) along the way
 - **imprecise** queries
 - enable user to specify varying accuracy requirements for different parts of the same query

Themis Palpanas - ISIP - May 2019

46

Need for Interactive Analytics



- interaction with users offers **new opportunities**
 - **progressive** answers
 - produce intermediate results
 - iteratively converge to final, correct solution
 - provide bounds on the errors (of the intermediate results) along the way
 - **imprecise** queries
 - enable user to specify varying accuracy requirements for different parts of the same query
- several exciting **research problems** in intersection of visualization and data management
 - **frontend**: HCI/visualizations for querying/results display
 - **backend**: efficiently supporting these operations

Themis Palpanas - ISIP - May 2019

47

Benchmarking Data Series Indexes?

Themis Palpanas - ISIP - May 2019

48

Previous Studies

evaluate **performance** of **indexing methods** using **random queries**

- chosen from the data (with/without noise)



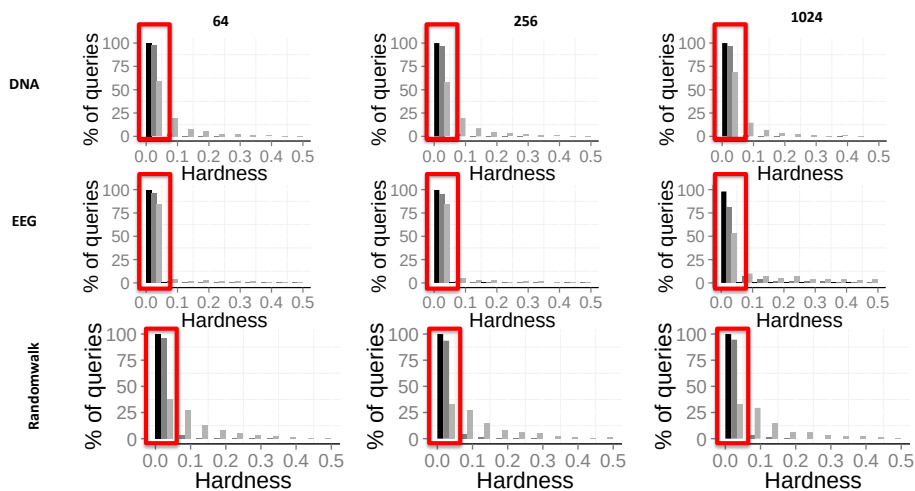
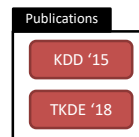
Themis Palpanas - ISIP - May 2019

49

49

Previous Workloads

Most previous workloads are *skewed* to *easy* queries

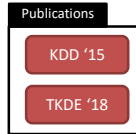


Themis Palpanas - ISIP - May 2019

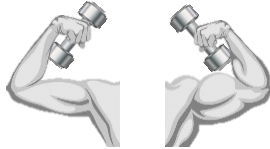
50

50

Benchmark Workloads



If all queries are **easy**
all indexes look **good**



If all queries are **hard**
all indexes look **bad**



need **methods** for **generating** queries of **varying hardness**



Themis Palpanas - ISIP - May 2019

51

51

dIN 52

Conclusions

- data series is a very **common** data type
 - across several different domains and applications
- complex data series analytics are **challenging**
 - have very high complexity
 - efficiency comes from data series management/indexing techniques
- current approaches used in practice are **ad-hoc**
 - waste of time and effort
 - suboptimal solutions
- need for **Sequence Management System**
 - optimize operations based on data/hardware characteristics
 - transparent to user
- several exciting **research opportunities**

Themis Palpanas - ISIP - May 2019

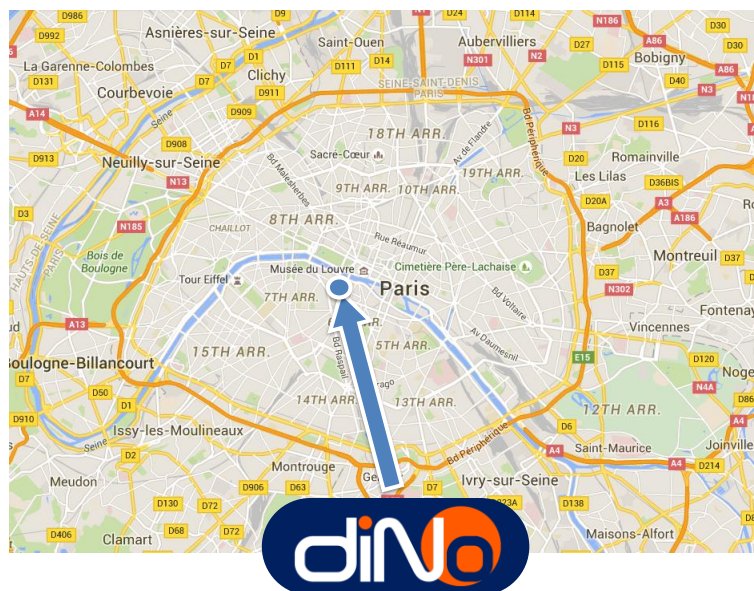
52

collaborations!

Themis Palpanas - ISIP - May 2019

53

Data-Intensive and Knowledge-Oriented systems



54

thank you!

google: Themis Palpanas

work supported by:



55

55