

Riding the Big IoT Data Wave

Complex Analytics for IoT Data Series



Themis Palpanas
Paris Descartes University



Telecom Paristech
Paris, April 2017



diNo 2

References

- papers
 - **ADS: The Adaptive Data Series Index.** VLDBJ 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.** VLDB 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>
- code and datasets
 - <http://www.mi.parisdescartes.fr/~themisp/isax2plus/>
- data series toolbox
 - <https://github.com/zoumpatianos/DSSStat>
- demo
 - <http://daslab.seas.harvard.edu/rinse/>

Acknowledgements

- Michele Linardi
- Anna Gogolou
- Botao Peng
- Karia Echihabi

Paris Descartes University

- Alessandro Camerra
- University of Trento*

- Stratos Idreos
 - Kostas Zoumpatianos
- Harvard University*

- Yin Lou
 - Johannes Gehrke
- Cornell University*

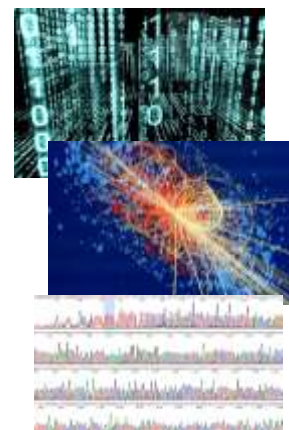
- Jin Shieh
 - Eamonn Keogh
- University of California at Riverside*



Themis Palpanas - Telecom Paristech, Apr 2017

Executive Summary

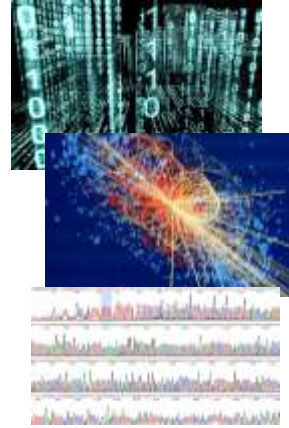
- data collected at unprecedented rates
- they enable data-driven scientific discovery
- lots of these data are sequences
 - takes **days-weeks** to analyze big sequence collections



Themis Palpanas - Telecom Paristech, Apr 2017

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



our work: **analyze big sequences** in **minutes/seconds**

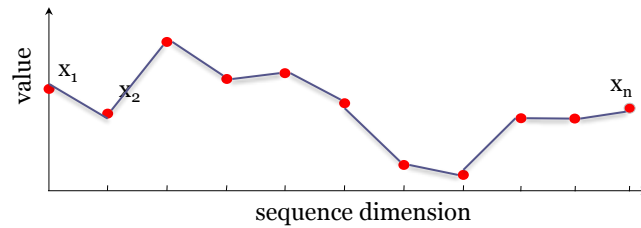
Themis Palpanas - Telecom Paristech, Apr 2017

Data series

Themis Palpanas - Telecom Paristech, Apr 2017

Data series

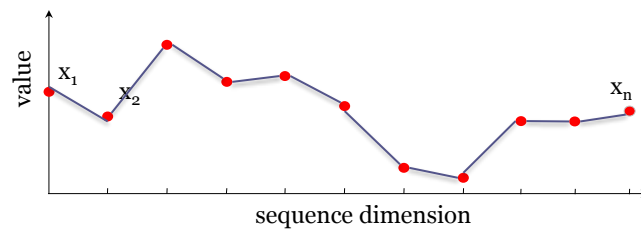
- Sequence of points ordered along some dimension



Themis Palpanas - Telecom Paristech, Apr 2017

Data series

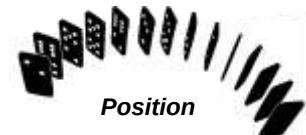
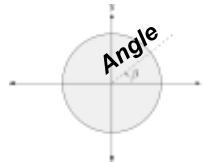
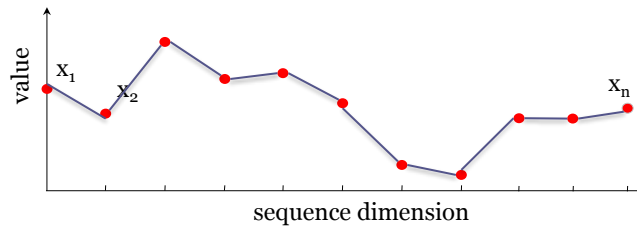
- Sequence of points ordered along some dimension



Themis Palpanas - Telecom Paristech, Apr 2017

Data series

- Sequence of points ordered along some dimension

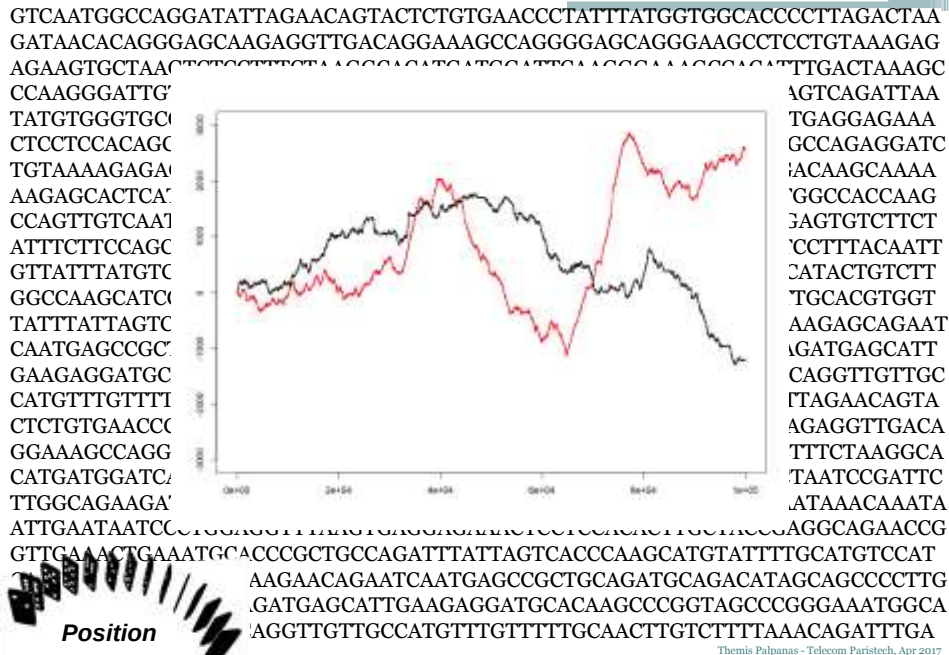
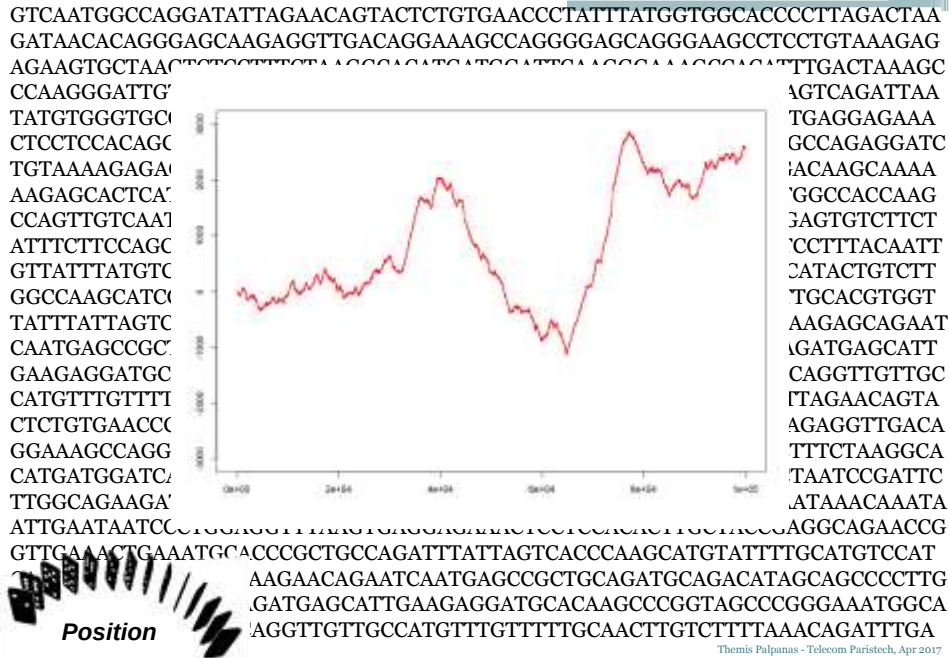


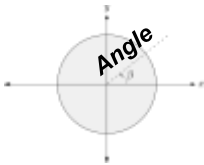
Thémis Palpanas - Telecom Paristech, Apr 2017

GTCAATGGCCAGGATATTAGAACAGTACTCTGTGAACCTATTTATGGTGGCACCCTTAGACTAA
GATAACACAGGGAGCAAGAGGTTGACAGGAAAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAG
AGAAGTGCTAAGTCTCCTTTCTAAGGCACATGATGGATTCAAGGGAAAAGCCACATTTGACTAAAGC
CCAAGGGATTGTTGCTTCTAATCCGATTTCTTGGCAGAAGATATTACAACTAAGAGTCAGATTAA
TATGTGGGTGCCAAAATAAAATAAAATAAATTGAATAATCCCTGGAGGTTTAAAGTGAGGAGAAA
CTCCTCCACAGCTTGCTACCGAGGCAGAACCGTTGAACTGAAATGCATCCGCCGCCAGAGGATC
TGTAAGAGAGAGGTTGTTACGAACTGGCAACTGCCAACCAAGTCCACCAATGGACAAGCAAAA
AAGAGCACTCATCTCATGCTCCCAAGGATCAACCTTCCAGAGTTTTCACTTAAAGTGCCACCAAG
CCAGTTGTCAATCCAGGGCTTTGACTGAAATCTAGGGCTTCATCCGCTACCTCAGAGTGTCTTCT
ATTTCTCCAGCCAGTGACAAATACAACAAACATCTGAGATGTTTTAGCTATAAATCCTTTACAATT
GTTATTTATGTCTTAACCTTTTGTATACCTGAAAAAGTAGGGGAAACAATAAGAACATACTGTCTT
GGCCAAGCATCCAAGGTTAAATGAGTTATGGAAATTCATTTGGGAGCCAAGACATTCACAGTGGT
TATTTATTAGTCACCCAAGCATGTATTTTGCATGTCCATCAGTTGTTCTTGGCCAAAAGAGCAGAAT
CAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTGCAAGGACAAGTCTGCAAGATGAGCATT
GAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCAGGCACTTACAAGAGCCAGGTTGTGTC
CATGTTTGTGTTTGGCAACTTGTCTATTTAAAGAGATTTGGGCAATGGCCAGGATATTAGAACAGTA
CTCTGTGAACCTATTTATGTTAGCAGCCCTTAGACTAAGATAACACAGGGAGCAAGAGGTTGACA
GGAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAGAGAAGTGCTAAGTCTCCTTTCTAAGGCA
CATGATGGATCAAGGGAAAGTCATTTGACTAAAGCCCAAGGGATTGTTGCTTCTAATCCGATTC
TTGGCAGAAGATATTGCAACTAAGAGTCAGATTAAATATGTGGGTGCCAAAATAAAATAAAATA
ATTGAATAATCCCTGGAGGTTTAAAGTGAGGAGAACTCCTCCACACTTGCTACCGAGGCAGAACCG
GTTGAACCTGAAATGCACCCGCTGCCAGATTTATAGTCACCCAAGCATGTATTTTGCATGTCCAT
AAGAACAGAATCAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTG
.GATGAGCATTGAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCA
AGGTTGTTGCCATGTTTGTGTTTGGCAACTTGCTTTTAAACAGATTGTA

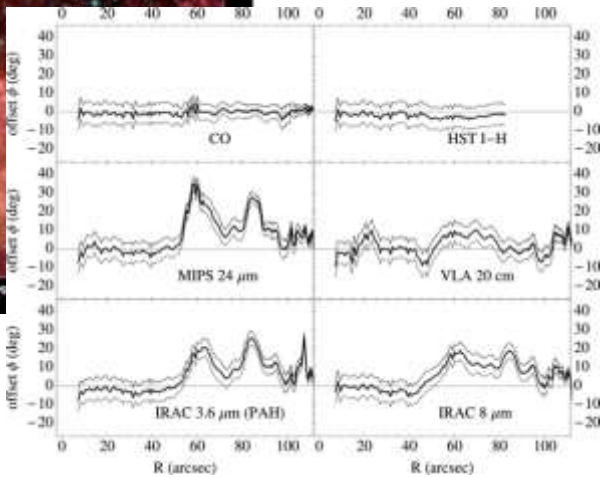
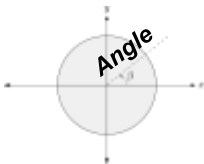
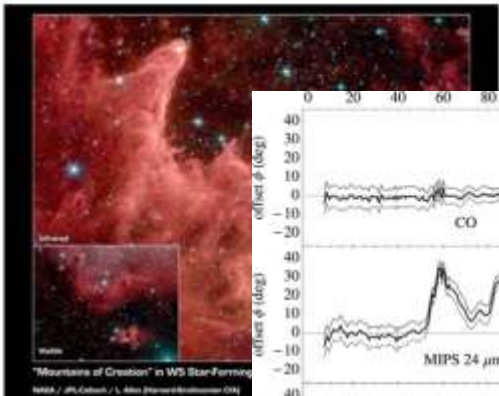


Thémis Palpanas - Telecom Paristech, Apr 2017





Themis Palpanas - Telecom Paristech, Apr 2017

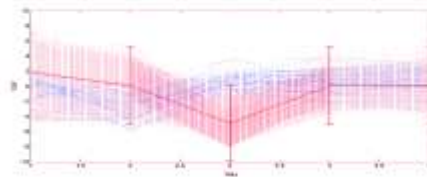


Schinnerer et al.

Themis Palpanas - Telecom Paristech, Apr 2017

Telecommunications

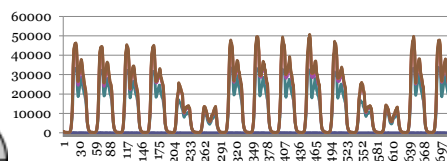
- analysis of **call activity** patterns
 - Telecom Italia



call activity for Easter Monday



clustermap of incoming calls time series



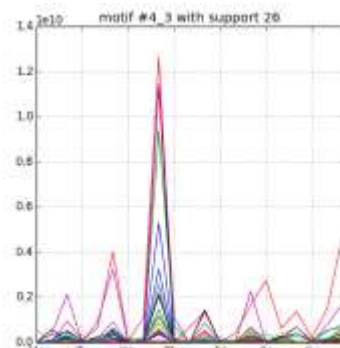
average number of calls for 5 smallest clusters



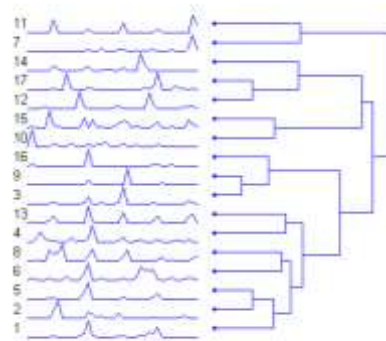
Themis Palpanas - Telecom Paristech, Apr 2017

Home Networks

- temporal **usage behavior analysis** of home networks
 - Portugal Telecom



(previously unknown) frequent behavior pattern



clustering based on user activity patterns



Themis Palpanas - Telecom Paristech, Apr 2017



Operation Health Monitoring



Themis Palpanas - Telecom Paristech, Apr 2017

32



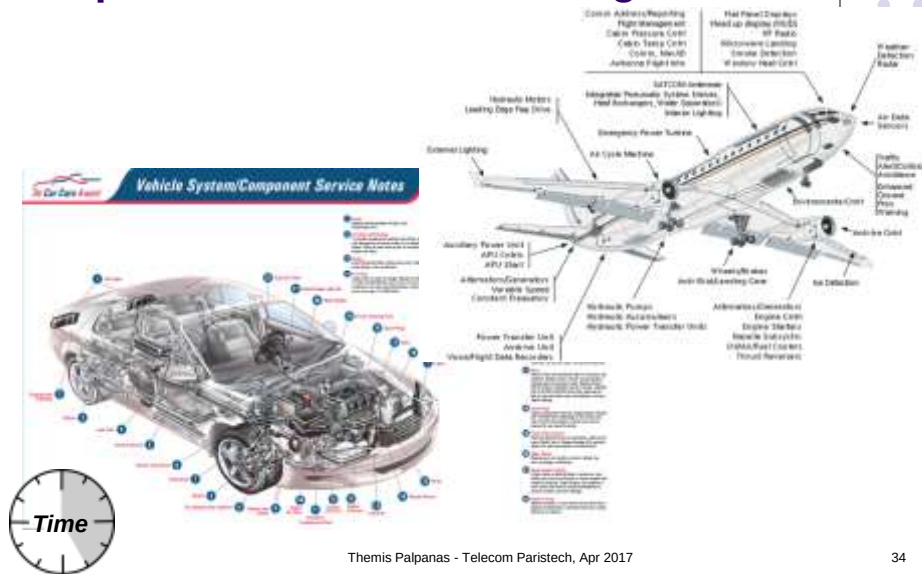
Operation Health Monitoring



Themis Palpanas - Telecom Paristech, Apr 2017

33

Operation Health Monitoring

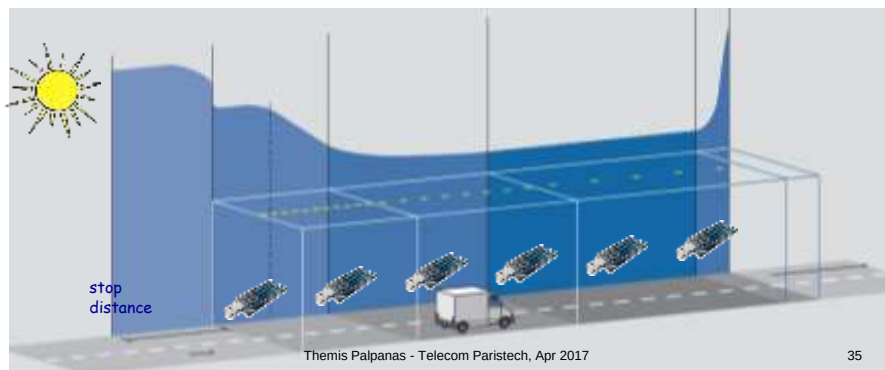


Themis Palpanas - Telecom Paristech, Apr 2017

34

Road Tunnel Monitoring and Control

- Lamp levels typically statically determined, ignoring environmental
 - **Overprovisioned** to meet the regulations
- Problems: waste energy and potential security hazard
- **Idea:** place wireless sensors along tunnel, adjust lamps to actual conditions
 - Eliminate overprovisioning, account for environmental variations



Themis Palpanas - Telecom Paristech, Apr 2017

35



dIN 37

Internet of Things: Data Acquisition

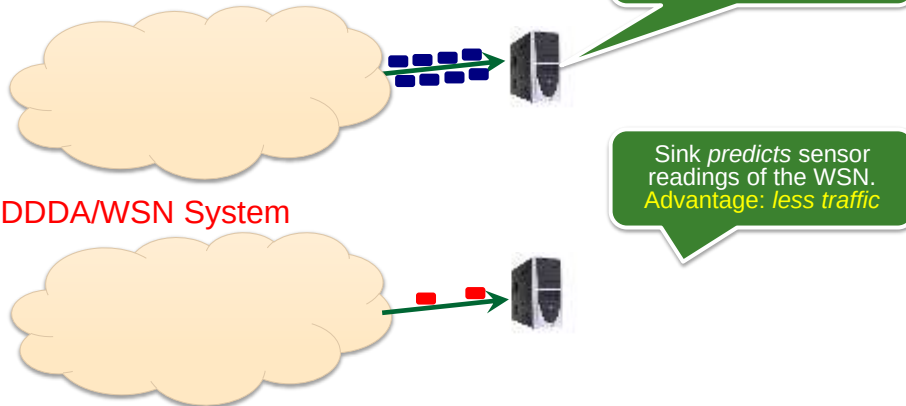
Usman Raza, Alessandro Camerra, Amy L. Murphy, Themis Palpanas, Gian Pietro Picco. **Practical Data Prediction for Real-World Wireless Sensor Networks**. TKDE 27(8), 2015

Usman Raza, Alessandro Camerra, Amy L. Murphy, Themis Palpanas, Gian Pietro Picco. **What Does Model-Driven Data Acquisition Really Achieve in Wireless Sensor Networks?**. PerCom 2012. **BEST PAPER**

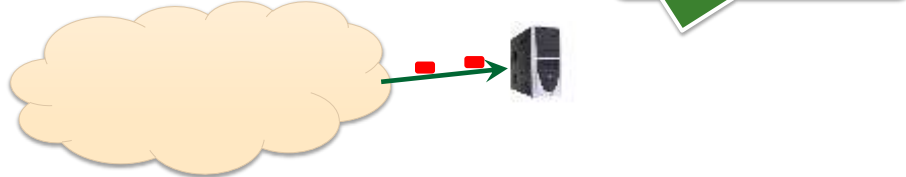
Themis Palpanas - Telecom Paristech, Apr 2017

Data-Driven Data Acquisition (DDDA)

Typical WSN System



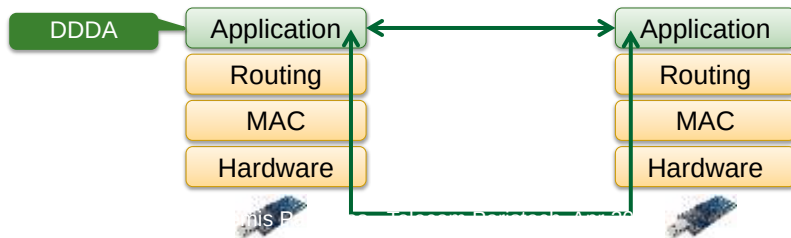
DDDAWSN System



38

What Does Data-Driven Data Acquisition *Really* Achieve?

- DDDA *well* studied in database community
 - Depending on scenario, *99% less traffic generated*
- Does 99% less traffic mean 99% more lifetime?
 - Current studies look only at application-layer
 - Full network stack influences lifetime



39

Derivative-Based Prediction (DBP)

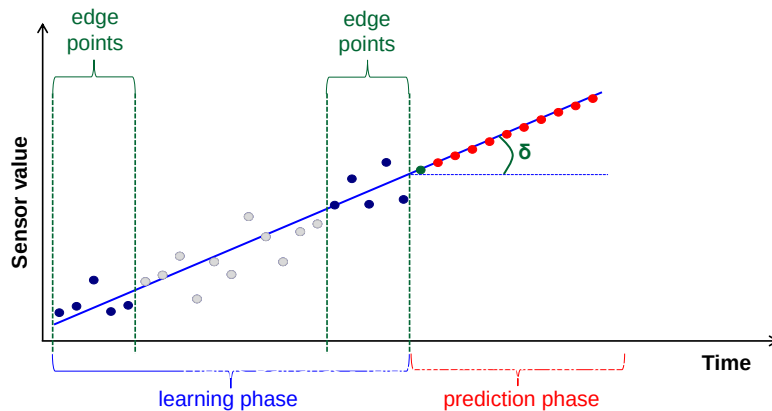
- DBP: a linear model
 - Easy to calculate the model
 - Easy to decide if the sensed data fit the model

Publications

TKDE '15

PERCOM '12

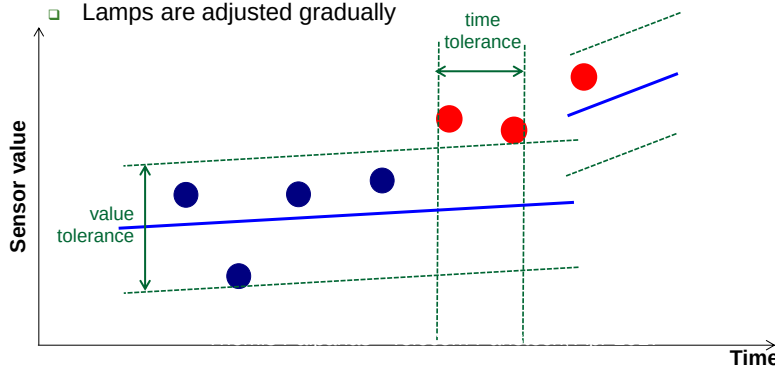
SPRINGER '12



40

Tolerance for Prediction Error

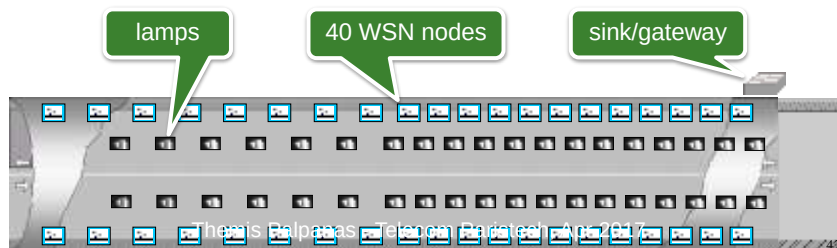
- Centralized control algorithm requires only *approximate* values
 - **Tolerances:** sensed values may temporarily exceed these values without requiring new model generation
- **Value tolerance** – maximum of a percentage and an absolute
 - Addresses inherent sensor error and variations of low light levels
- **Time tolerance** – in terms of sampling periods
 - Lamps are adjusted gradually



41

Testing the DBP Model

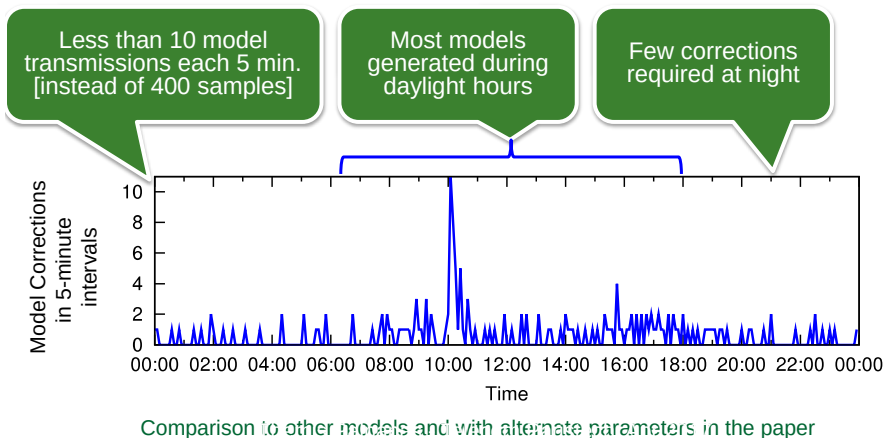
- Test scenario
 - 250 m long tunnel, 40 TMote equivalent nodes, 1 gateway
 - 47 days, samples every 30 seconds, **5.4 million measurements**
 - **DBP**: parameters established by lighting engineers



DBP Results:

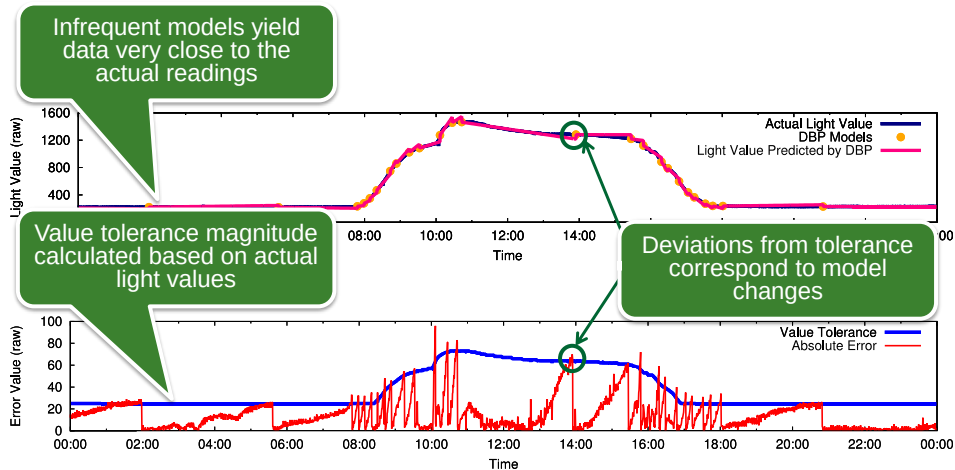
Excellent reduction in data traffic

- **99.1%** messages suppressed at the nodes



43

DBP Results: *Excellent data approximation*

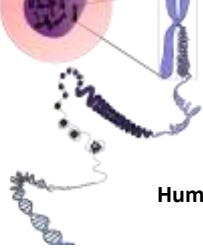


44

dIN 45

Internet of Things: Complex Data Analytics

Massive Data Series Collections

 NASA's Solar Observatory 1.5 TB per day Planned Large Synoptic Survey Telescope ~30 TB per night	 Human Genome project 130 TB
 Boeing jet 20 TB per hour	data center and services monitoring 2B data series 4M points/sec 

Themis Palpanas - Telecom Paristech, Apr 2017

46



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

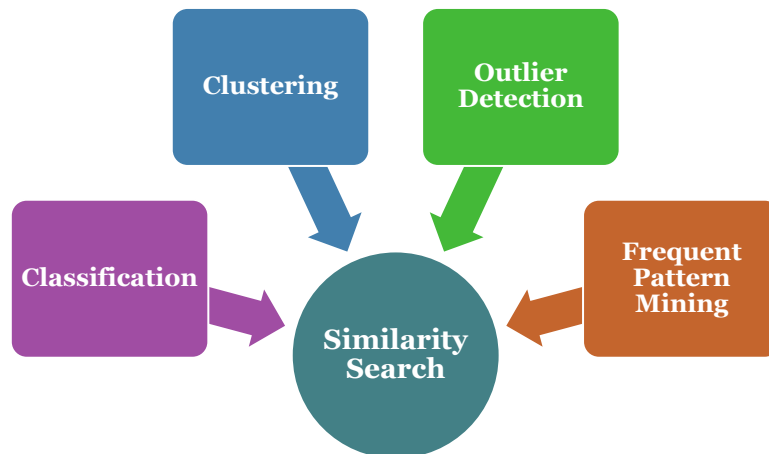
Themis Palpanas - Telecom Paristech, Apr 2017

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



Themis Palpanas - Telecom Paristech, Apr 2017

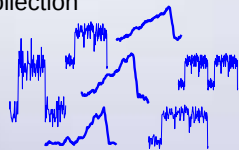
What do we want to do with them?
- complex analytics



Themis Palpanas - Telecom Paristech, Apr 2017

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

sequence
collection



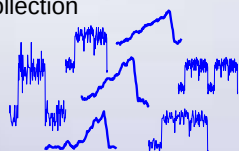
Themis Palpanas - Telecom Paristech, Apr 2017

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

query

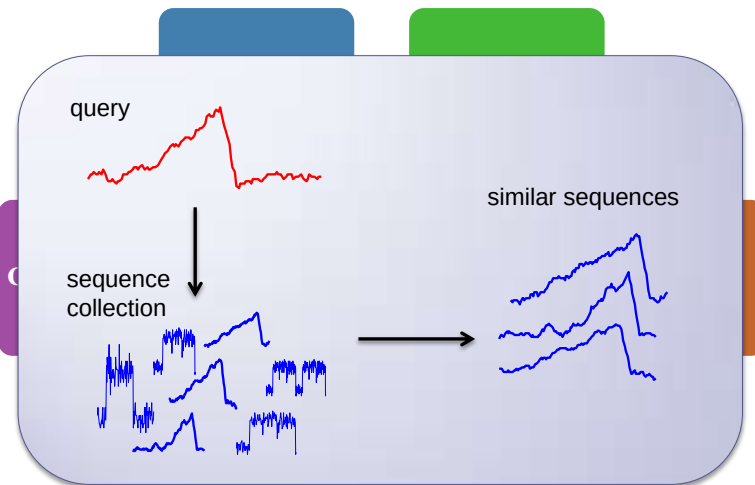


sequence
collection



Themis Palpanas - Telecom Paristech, Apr 2017

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



Themis Palpanas - Telecom Paristech, Apr 2017

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

Euclidean

$$D(X, Y) \equiv \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Themis Palpanas - Telecom Paristech, Apr 2017

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

Euclidean

$$D(X, Y) \equiv \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Dynamic Time
Warping (DTW)

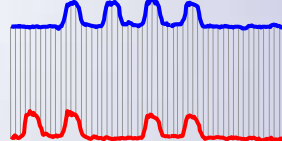
$$D_{dtw}(X, Y) = f(n, m)$$

$$f(i, j) = \|x_i - y_j\| + \min \begin{cases} f(i, j-1) \\ f(i-1, j) \\ f(i-1, j-1) \end{cases}$$

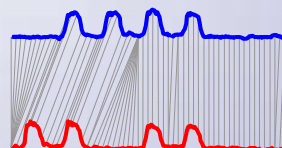
Themis Palpanas - Telecom Paristech, Apr 2017

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

Euclidean

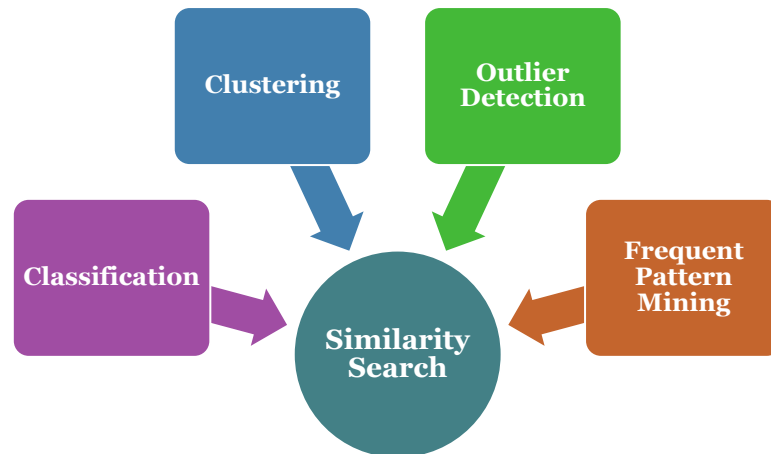


Dynamic Time
Warping (DTW)



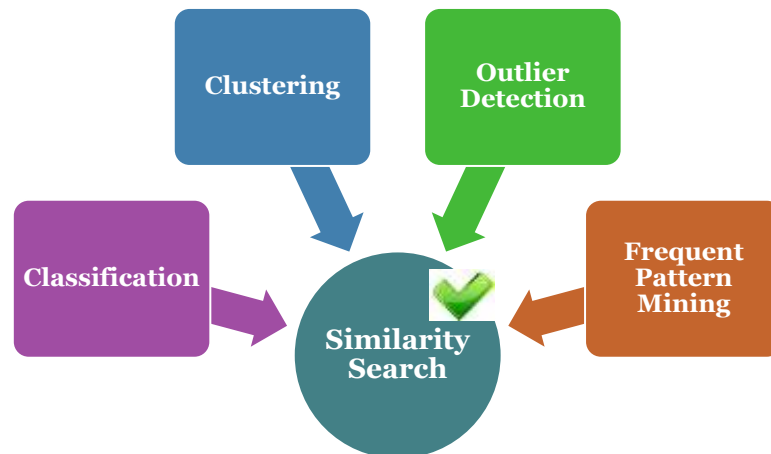
Themis Palpanas - Telecom Paristech, Apr 2017

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



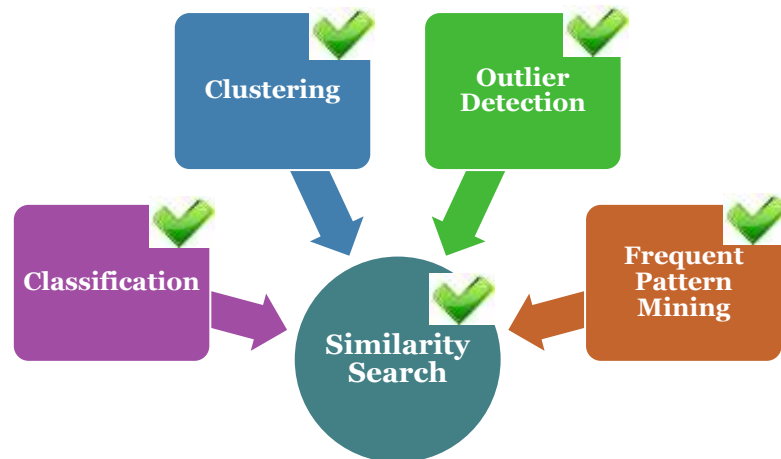
Themis Palpanas - Telecom Paristech, Apr 2017

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



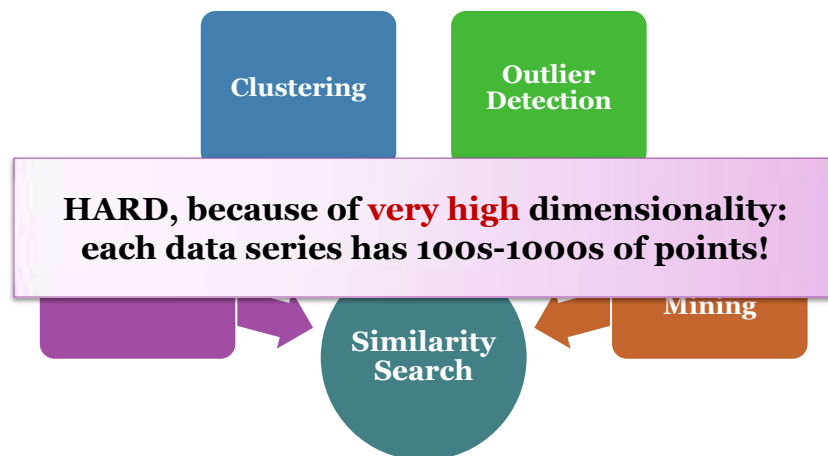
Themis Palpanas - Telecom Paristech, Apr 2017

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



Themis Palpanas - Telecom Paristech, Apr 2017

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



Themis Palpanas - Telecom Paristech, Apr 2017

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

Clustering

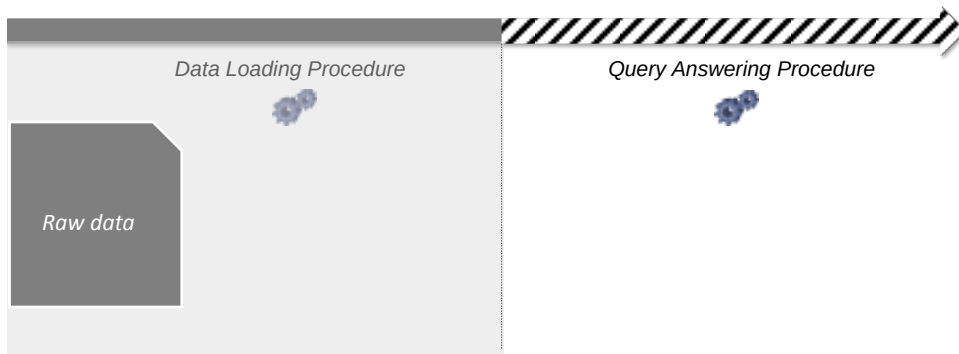
Outlier
Detection

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

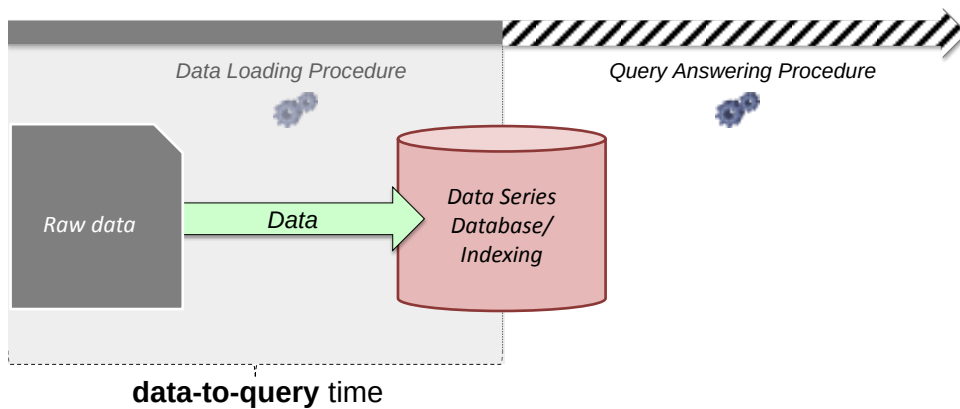
even HARDER, because of **very large** size:
millions to billions of data series (multi-TBs)!

Themis Palpanas - Telecom Paristech, Apr 2017

Query answering process



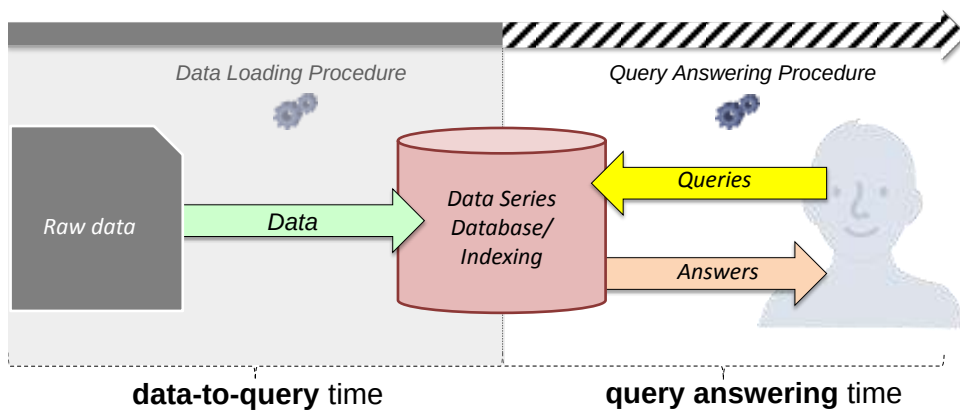
Query answering process



Themis Palpanas - Telecom Paristech, Apr 2017

62

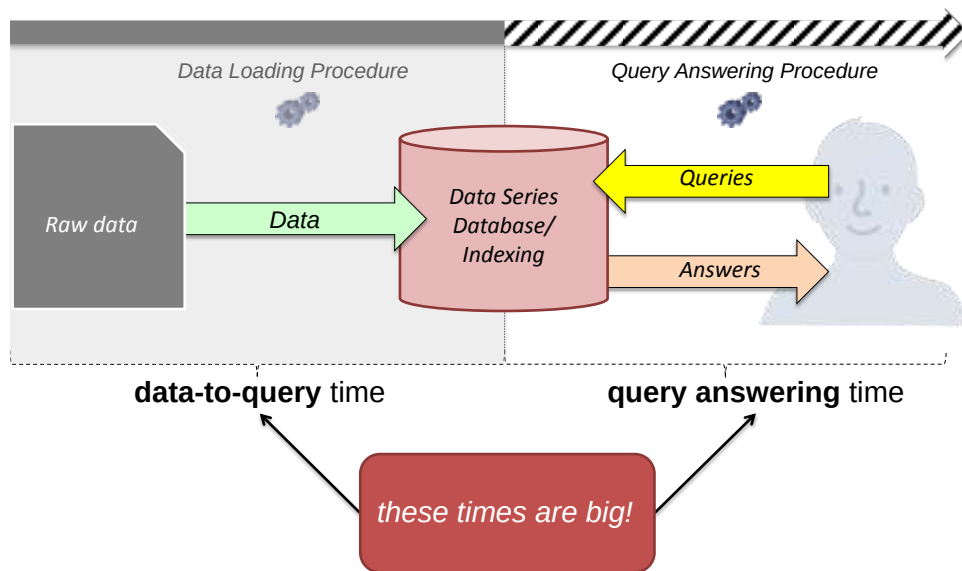
Query answering process



Themis Palpanas - Telecom Paristech, Apr 2017

63

Query answering process



Themis Palpanas - Telecom Paristech, Apr 2017

64

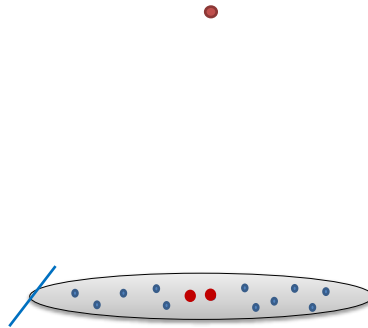
Similarity Search via Serial Scan



Themis Palpanas - Telecom Paristech, Apr 2017

65

Similarity Search via Serial Scan



Themis Palpanas - Telecom Paristech, Apr 2017

66

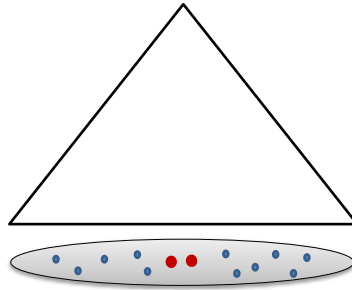
Similarity Search via Serial Scan



Themis Palpanas - Telecom Paristech, Apr 2017

67

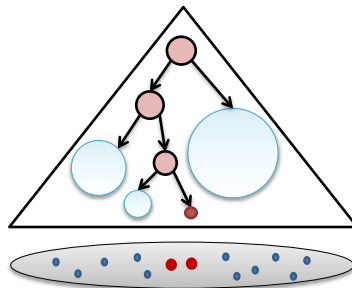
Similarity Search via Indexing



Themis Palpanas - Telecom Paristech, Apr 2017

68

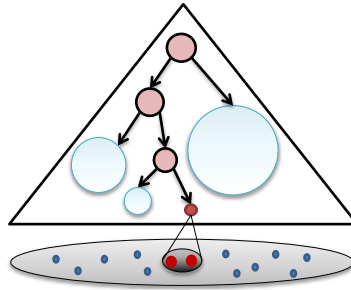
Similarity Search via Indexing



Themis Palpanas - Telecom Paristech, Apr 2017

69

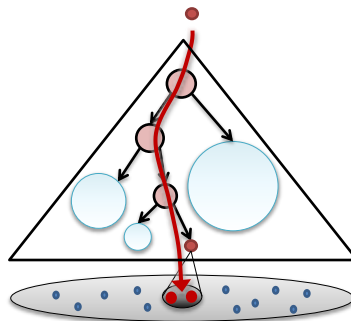
Similarity Search via Indexing



Themis Palpanas - Telecom Paristech, Apr 2017

70

Similarity Search via Indexing



Themis Palpanas - Telecom Paristech, Apr 2017

71

Traditional Approaches

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

Themis Palpanas - Telecom Paristech, Apr 2017

72

Traditional Approaches

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

serial scan takes 45 minutes/query



Query Answering

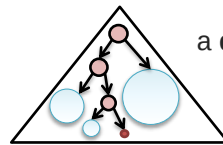
Themis Palpanas - Telecom Paristech, Apr 2017

73

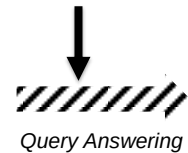
Traditional Approaches

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

serial scan takes 45 minutes/query



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



Themis Palpanas - Telecom Paristech, Apr 2017

74

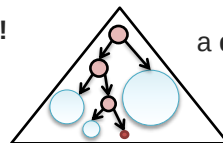
Traditional Approaches

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

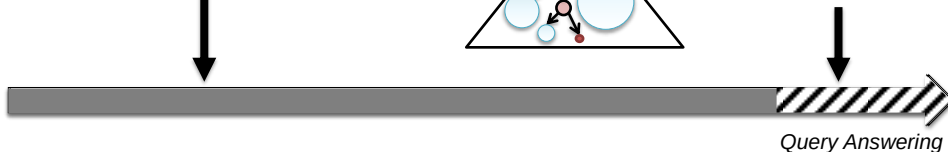
serial scan takes 45 minutes/query



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



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



Themis Palpanas - Telecom Paristech, Apr 2017

75

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



Themis Palpanas - Telecom Paristech, Apr 2017

76

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

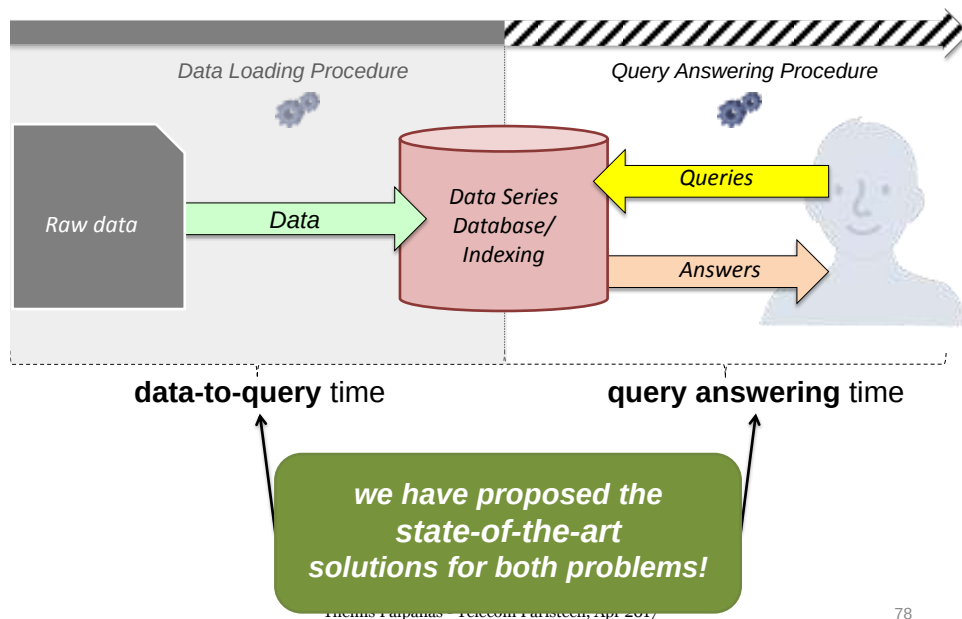


complex analytics in hours/days...

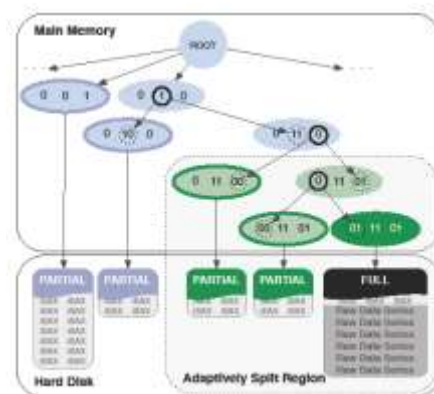
Themis Palpanas - Telecom Paristech, Apr 2017

77

Query answering process



Our Approach: ADS+



complex analytics in minutes/seconds!

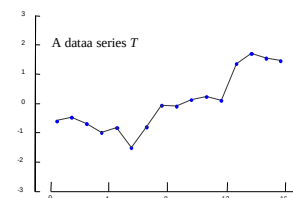
Outline

- background
 - SAX representation
 - iSAX representation
 - iSAX index
- proposed solution
 - bulk loading
 - splitting policy
 - adaptive solution
- experimental evaluation, case studies
- conclusions, future work, new challenges

Themis Palpanas - Telecom Paristech, Apr 2017

SAX Representation

- **Symbolic Aggregate approXimation (SAX)**
 - **(1)** Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)



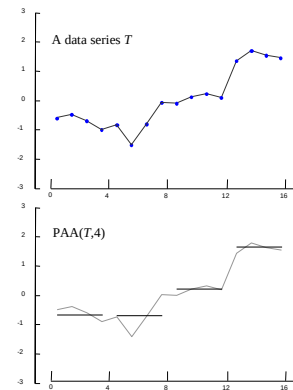
Themis Palpanas - Telecom Paristech, Apr 2017

SAX Representation

- **Symbolic Aggregate approXimation (SAX)**
 - (1) Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)
 - T typically normalized to $\mu = 0, \sigma = 1$

$$\text{PAA}(T, w) = \bar{T} = \bar{t}_1, \dots, \bar{t}_w$$

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

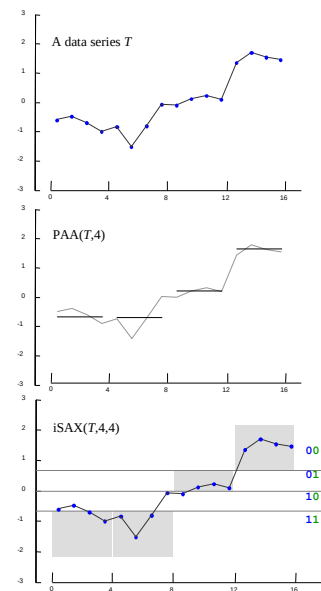


Themis Palpanas - Telecom Paristech, Apr 2017

SAX Representation

- **Symbolic Aggregate approXimation (SAX)**
 - (1) Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)
 - T typically normalized to $\mu = 0, \sigma = 1$
 - $\text{PAA}(T, w) = \bar{T} = \bar{t}_1, \dots, \bar{t}_w$

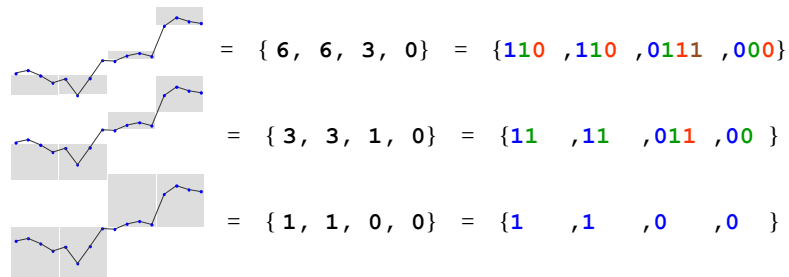
$$\text{where } \bar{t}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} T_j$$
 - (2) Discretize into a vector of symbols
 - Breakpoints map to small alphabet α of symbols



Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Representation

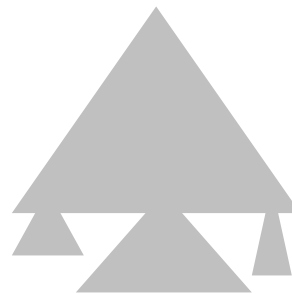
- iSAX offers a bit-aware, quantized, multi-resolution representation with variable granularity



Themis Palpanas - Telecom Paristech, Apr 2017

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 $\leq th$

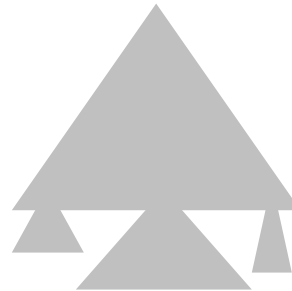


Themis Palpanas - Telecom Paristech, Apr 2017

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 $\leq th$

e.g., $th=4$, $w=4$, $b=1$

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$


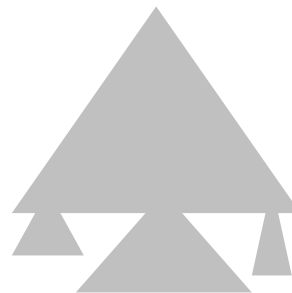
Themis Palpanas - Telecom Paristech, Apr 2017

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 $\leq th$

e.g., $th=4$, $w=4$, $b=1$

Insert: $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

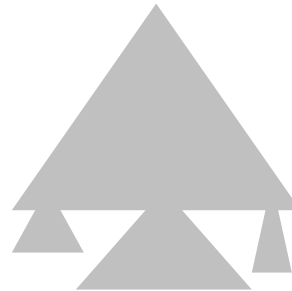
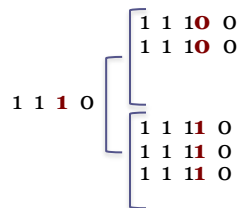


Themis Palpanas - Telecom Paristech, Apr 2017

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 $\leq th$

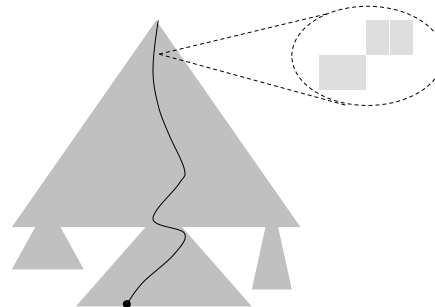
e.g., $th=4$, $w=4$, $b=1$



Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index

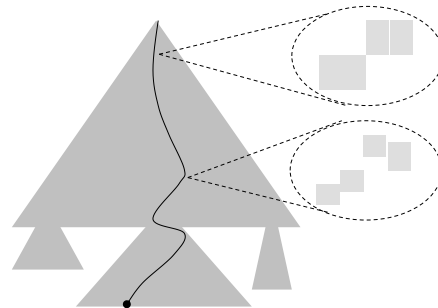
- non-balanced tree-based index with non-overlapping regions, and controlled fan-out rate
 - base cardinality b (optional), segments w , threshold th
 - hierarchically subdivides SAX space until num. entries $\leq th$



Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index

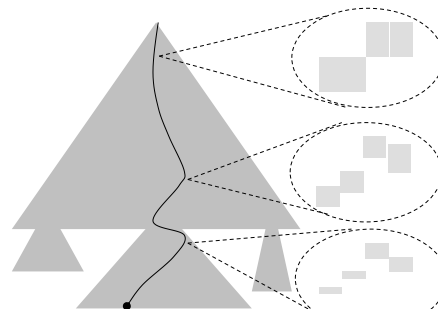
- non-balanced tree-based index with non-overlapping regions, and controlled fan-out rate
 - base cardinality b (optional), segments w , threshold th
 - hierarchically subdivides SAX space until num. entries $\leq th$



Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index

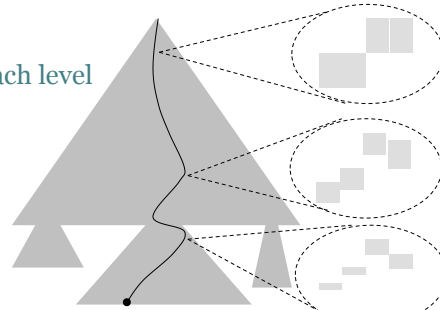
- non-balanced tree-based index with non-overlapping regions, and controlled fan-out rate
 - base cardinality b (optional), segments w , threshold th
 - hierarchically subdivides SAX space until num. entries $\leq th$



Themis Palpanas - Telecom Paristech, Apr 2017

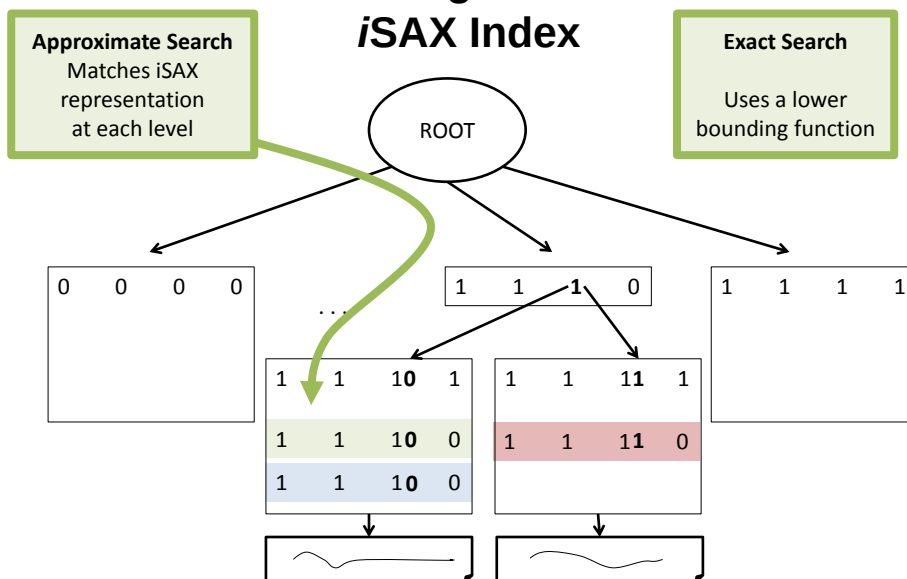
iSAX Index

- 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 - Telecom Paristech, Apr 2017

Background iSAX Index



Themis Palpanas - Telecom
Paristech, Apr 2017

97

iSAX Index: Shortcomings & Challenges

- this is a wonderful index!

Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index: Shortcomings & Challenges

- this is a wonderful index!
- ... but why does it take soooo long to build for huge datasets?

Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index: Shortcomings & Challenges

- this is a wonderful index!
- ... but why does it take soooo long to build for huge datasets?
- because iSAX implements
 - a **naive** node splitting policy
 - may lead to ineffective splits and additional disk I/O

Themis Palpanas - Telecom Paristech, Apr 2017

iSAX Index: Shortcomings & Challenges

- this is a wonderful index!
- ... but why does it take soooo long to build for huge datasets?
- because iSAX implements
 - a **naive** node splitting policy
 - may lead to ineffective splits and additional disk I/O
 - **no** bulk loading strategy
 - does not use available main memory to reduce disk I/O

Themis Palpanas - Telecom Paristech, Apr 2017

iSAX 2.0 Bulk Loading Algorithm

- design principles:
 - take advantage of available **main** memory
 - maximize **sequential** disk accesses

Themis Palpanas - Telecom Paristech, Apr 2017

iSAX 2.0 Bulk Loading Algorithm

- intuition for proposed solution:
 - for each leaf node, collect as many data series that belong to it as possible before materializing the leaf node
 - the raw values of data series in leaf nodes are written to disk

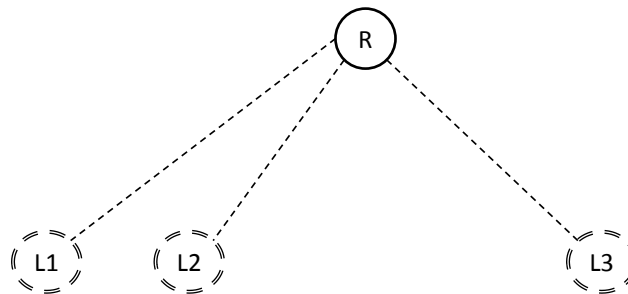
Themis Palpanas - Telecom Paristech, Apr 2017

iSAX 2.0

Bulk Loading Algorithm

- iterate between two phases (till all data series are indexed):
 - Phase 1
 - read data series and group them according to **first-level** nodes
 - use **all** available main memory
 - Phase 2
 - grow index by processing the subtree rooted at each one of the first-level nodes **one at-a-time**
 - flush leaf node contents to disk using **sequential** accesses

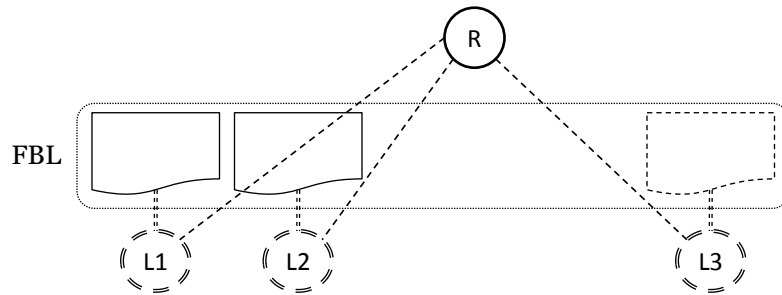
Themis Palpanas - Telecom Paristech, Apr 2017



main memory

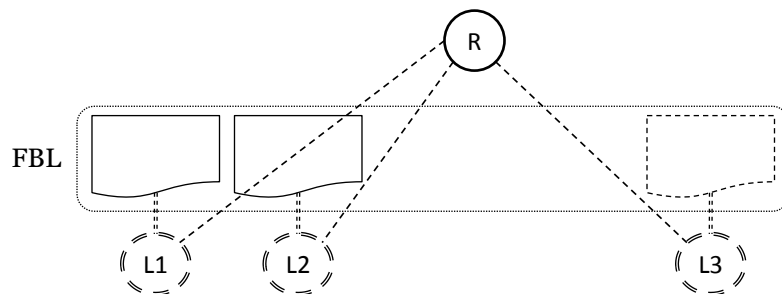
disk

Themis Palpanas - Telecom Paristech, Apr 2017



main memory
disk

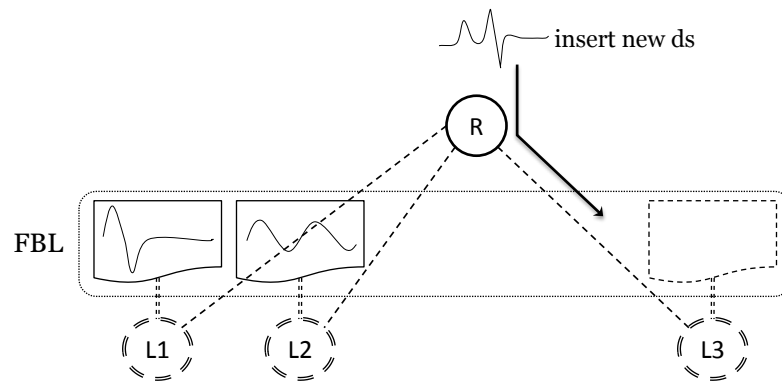
Themis Palpanas - Telecom Paristech, Apr 2017



main memory
disk

no limit in the size of FBLs!

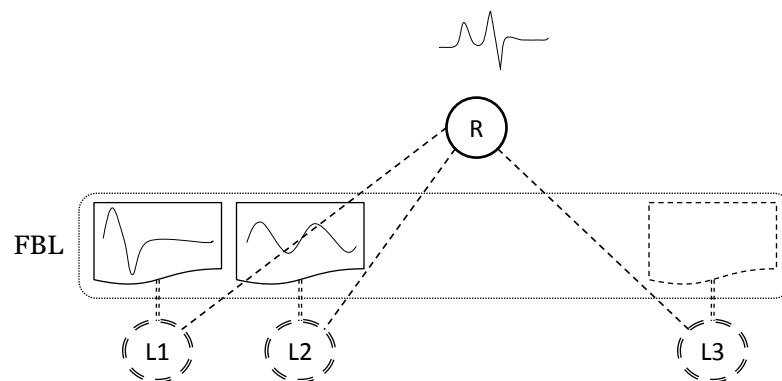
Themis Palpanas - Telecom Paristech, Apr 2017



main memory

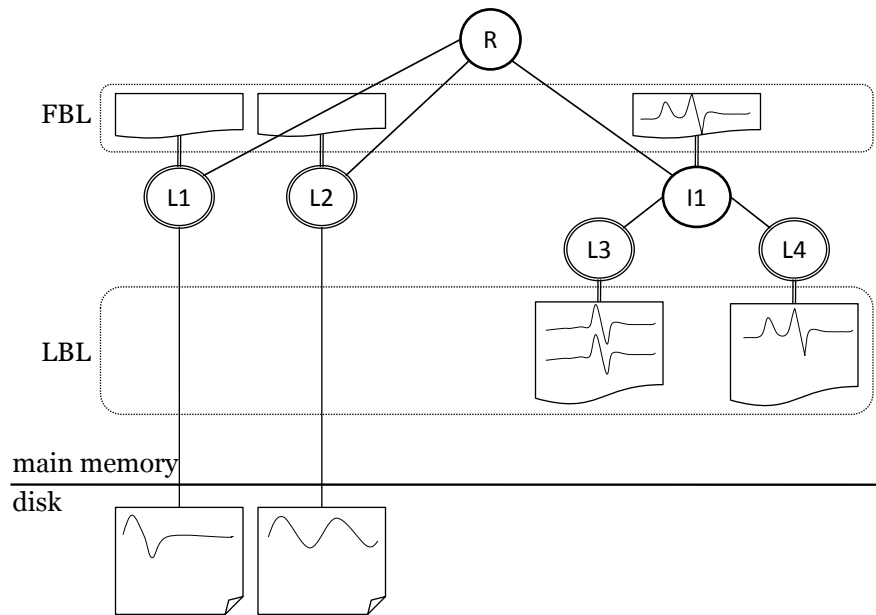
disk

Themis Palpanas - Telecom Paristech, Apr 2017

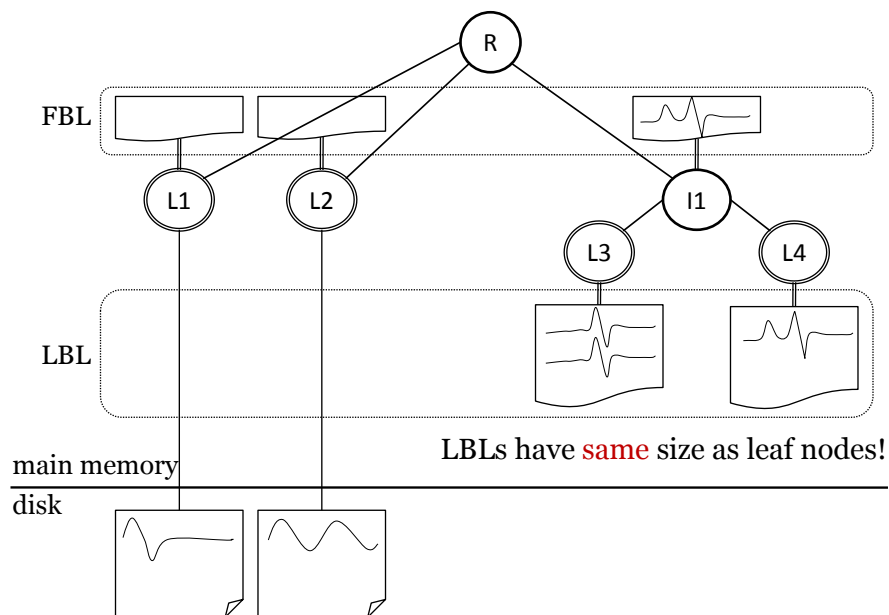


main memory
disk

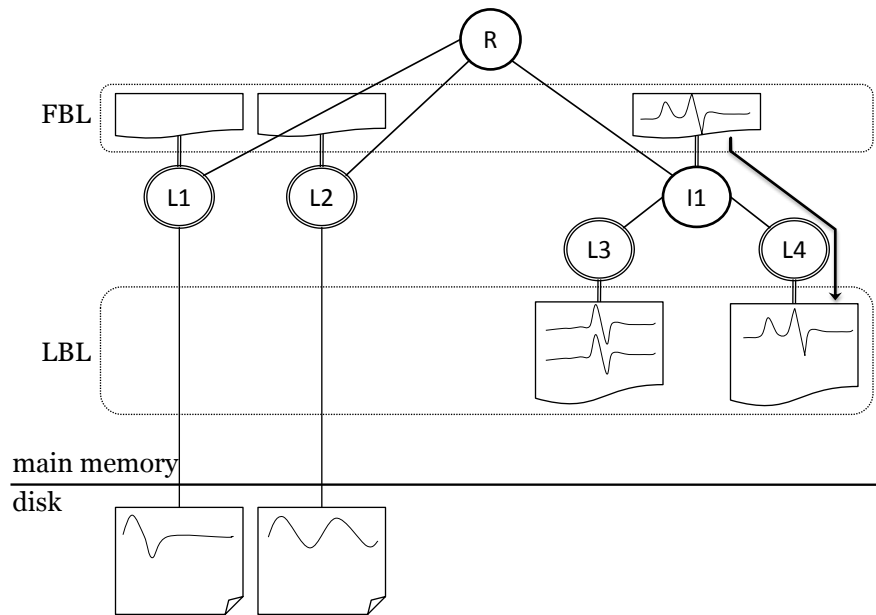
Themis Palpanas - Telecom Paristech, Apr 2017



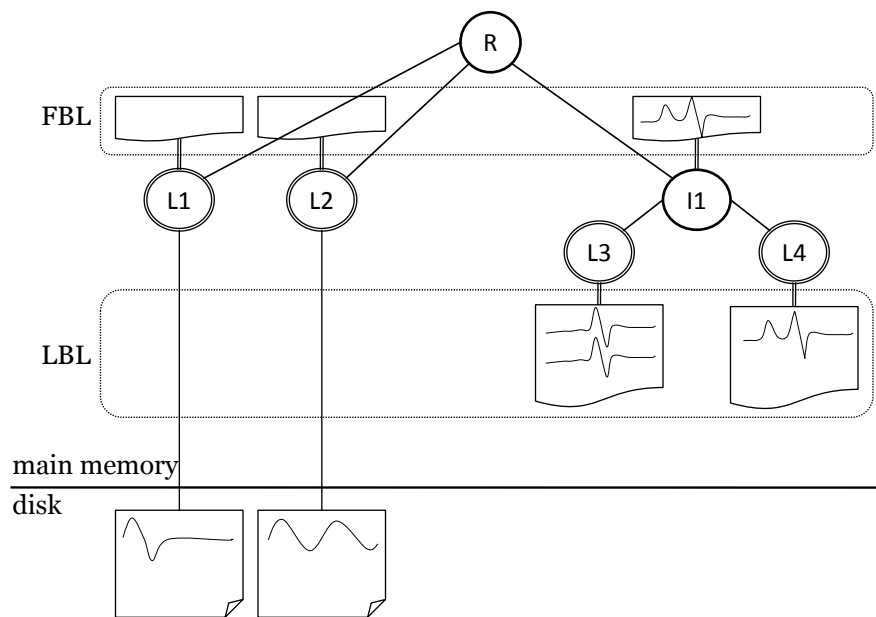
Themis Palpanas - Telecom Paristech, Apr 2017



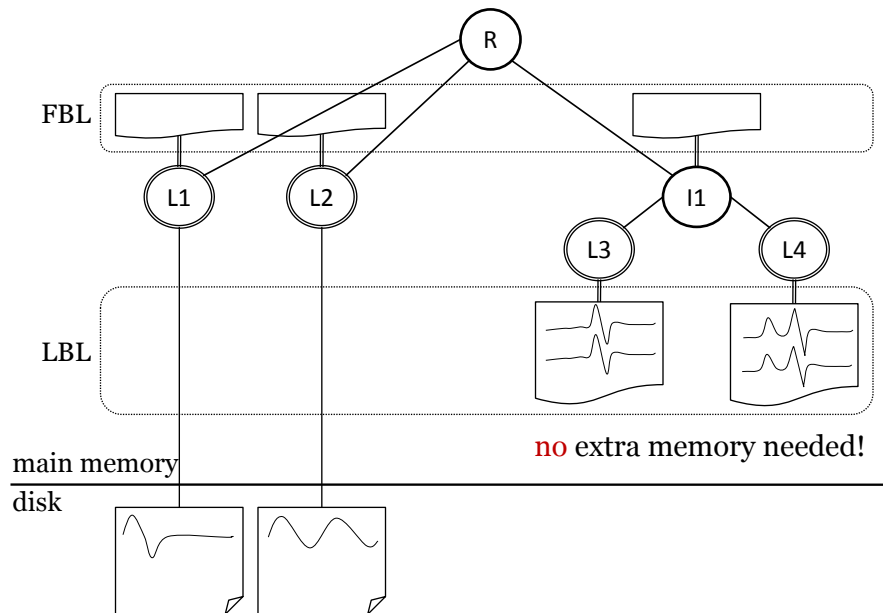
Themis Palpanas - Telecom Paristech, Apr 2017



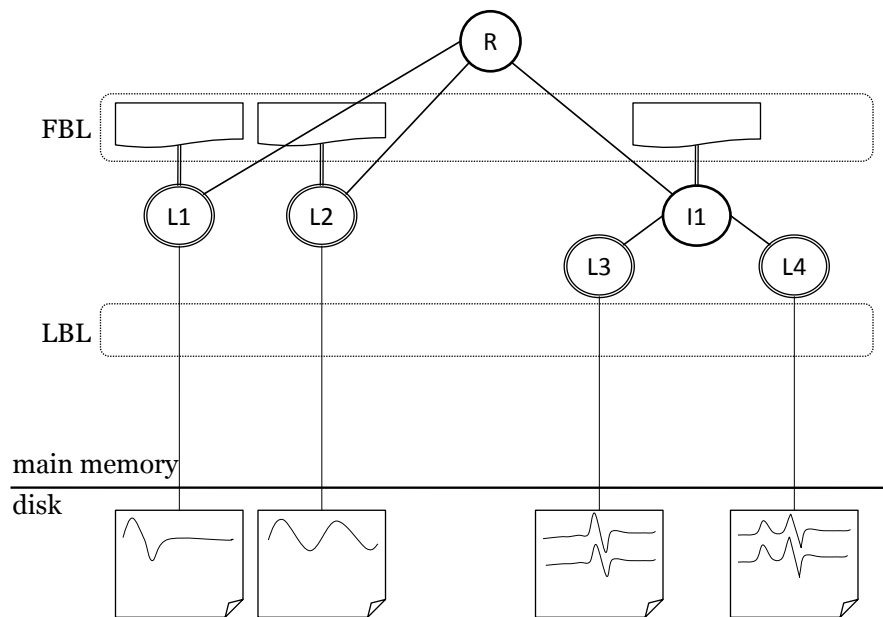
Themis Palpanas - Telecom Paristech, Apr 2017



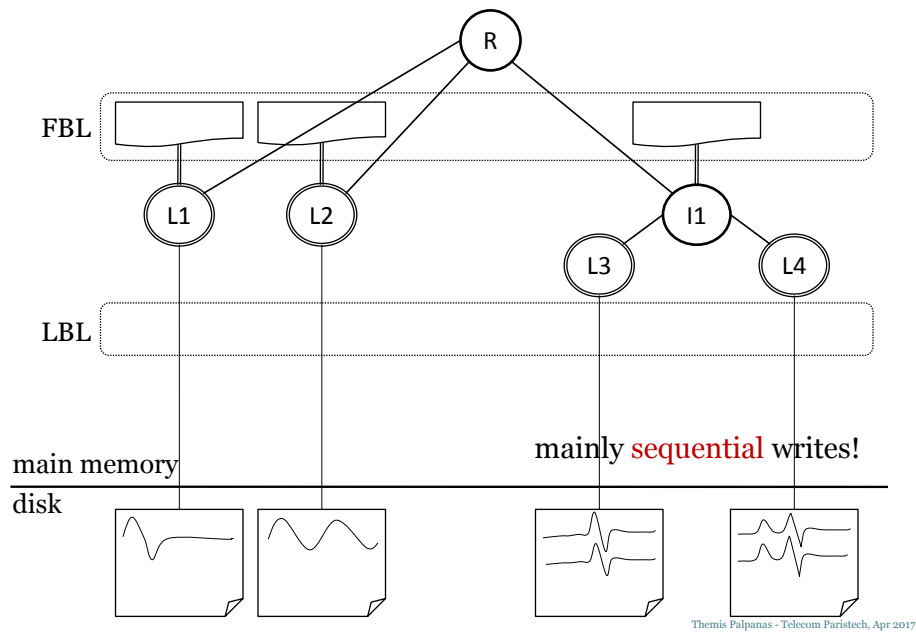
Themis Palpanas - Telecom Paristech, Apr 2017



Themis Palpanas - Telecom Paristech, Apr 2017

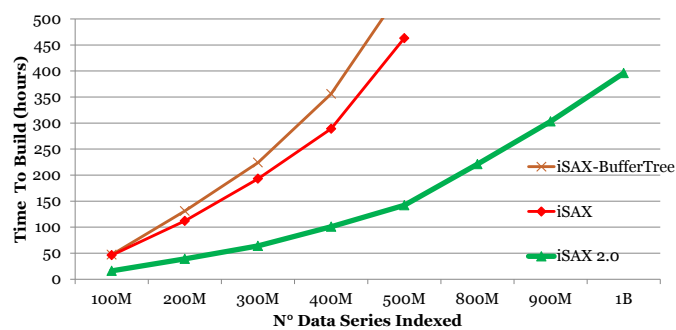


Themis Palpanas - Telecom Paristech, Apr 2017



Experimental Evaluation

Bulk Loading



- 1 Billion data series indexed in 16 days: 72% less time
- indexing time per data series: 0.001 sec

iSAX2+

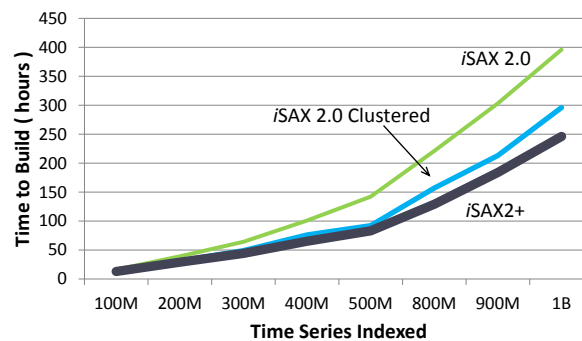
Publications

KAIS'14

- intuition for proposed solution:
 - iSAX grows fast at the beginning of bulk loading, its shape stabilizing well before the end of the process
 - several data series end up in leaf nodes that never need to split
 - implement lazy splitting:
 - move raw data to leaf node the first time
 - if leaf node splits, do not move raw data until the end of index building process

Themis Palpanas - Telecom Paristech, Apr 2017

Experimental Evaluation Bulk Loading



- **1 Billion** data series indexed in **10 days**: **82% less time** than iSAX
- indexing time per data series: **0.8 milliseconds**

Themis Palpanas - Telecom Paristech, Apr 2017

Drawback of iSAX2+

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

Themis Palpanas - Telecom Paristech, Apr 2017

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 - Telecom Paristech, Apr 2017

Adaptive Data Series Index: ADS+

Publications

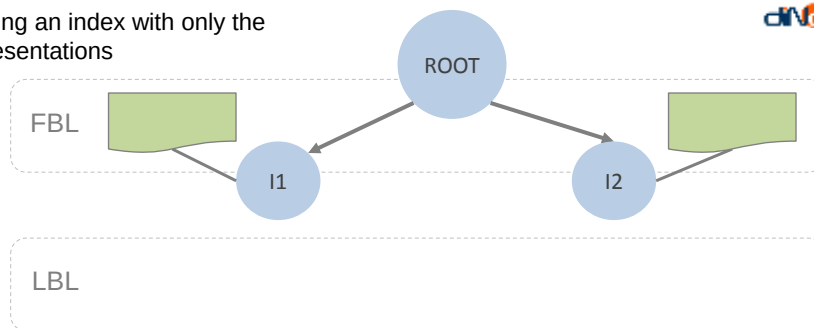
SIGMOD'14

VLDBJ'16

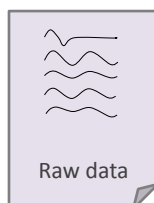
- intuition for proposed solution
 - build the iSAX index using the iSAX representations
 - just like iSAX2+
 - but start with a large leaf size
 - minimize initial cost
 - postpone leaf materialization to query time
 - only materialize (at query time) leaves needed by queries
 - parts that are queried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

Themis Palpanas - Telecom Paristech, Apr 2017

Start building an index with only the
iSAX representations

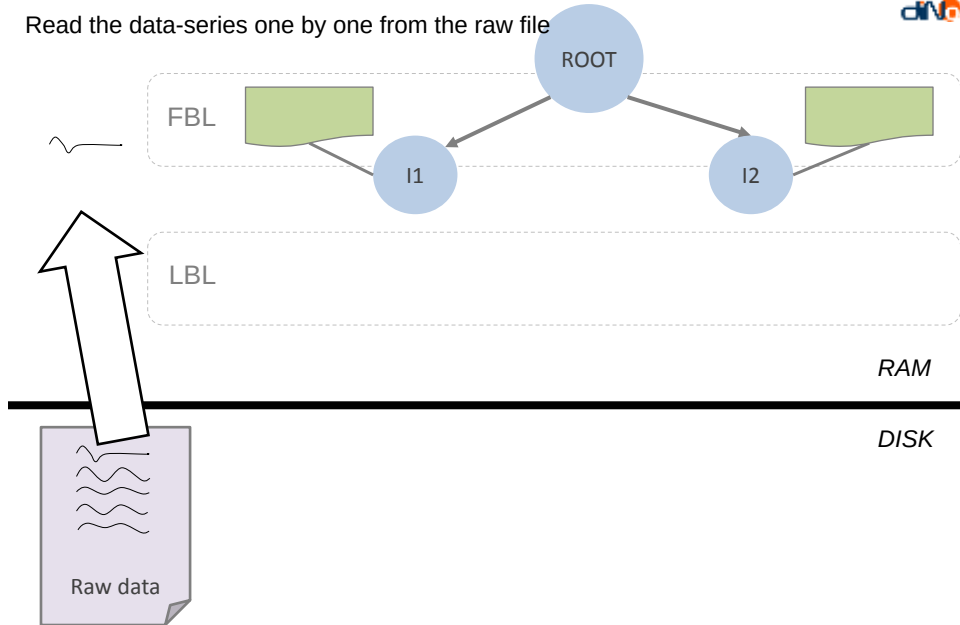


RAM



DISK

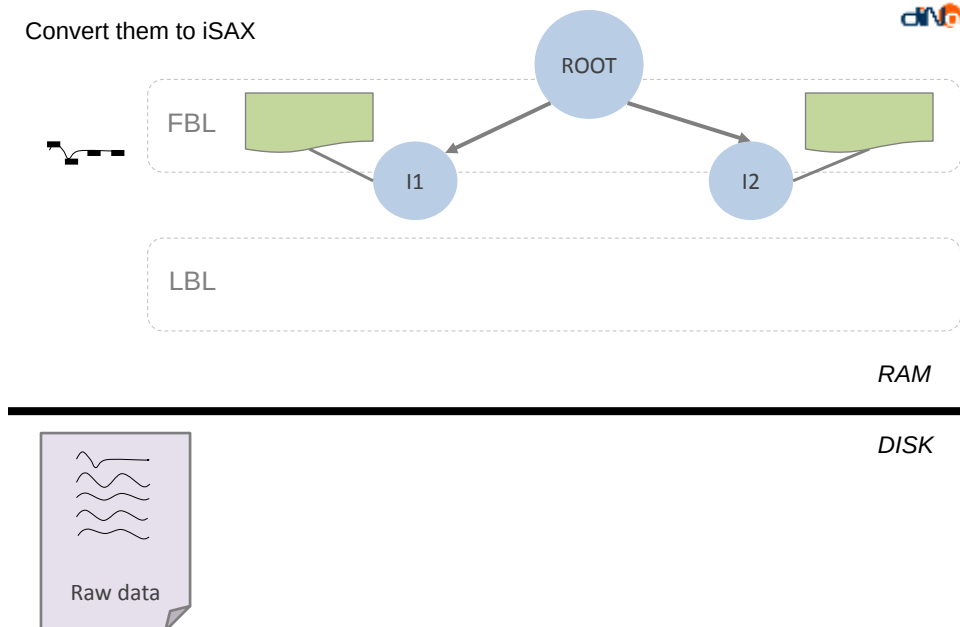
Read the data-series one by one from the raw file



Themis Palpanas - Telecom Paristech, Apr 2017

178

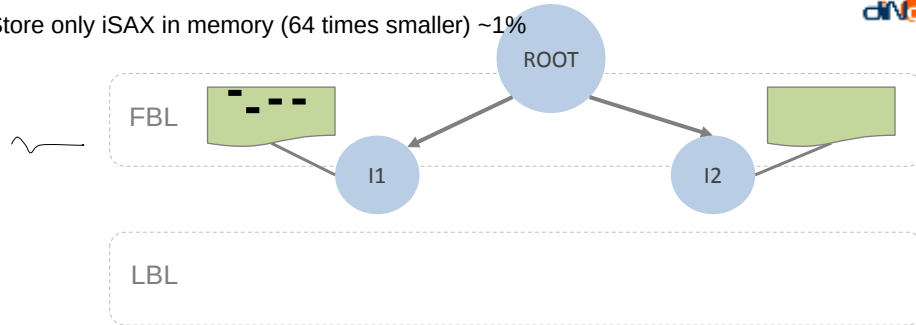
Convert them to iSAX



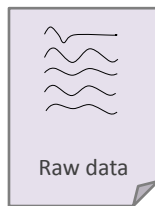
Themis Palpanas - Telecom Paristech, Apr 2017

179

Store only iSAX in memory (64 times smaller) ~1%



RAM

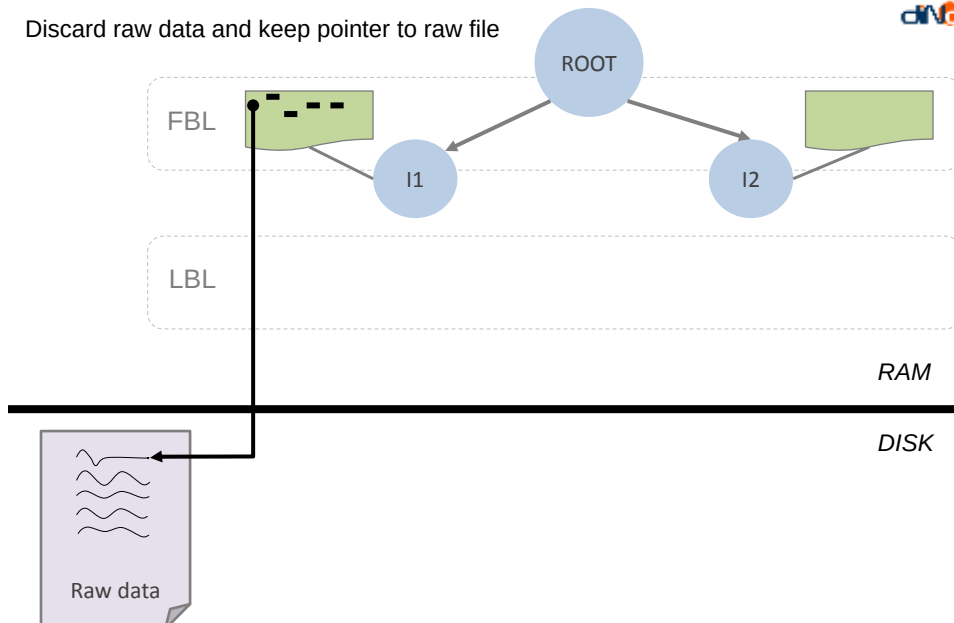


DISK

Themis Palpanas - Telecom Paristech, Apr 2017

180

Discard raw data and keep pointer to raw file



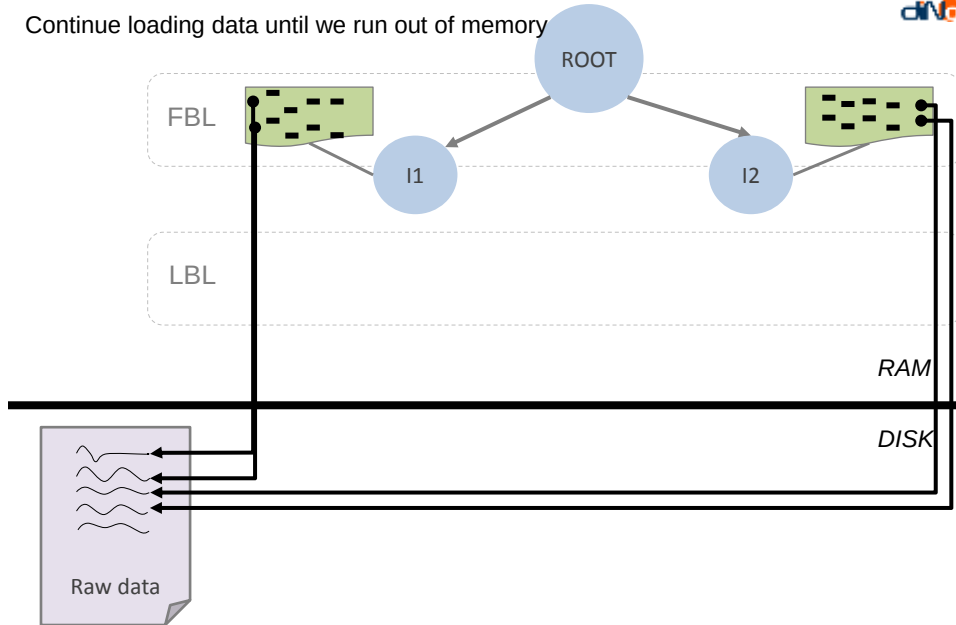
RAM

DISK

Themis Palpanas - Telecom Paristech, Apr 2017

181

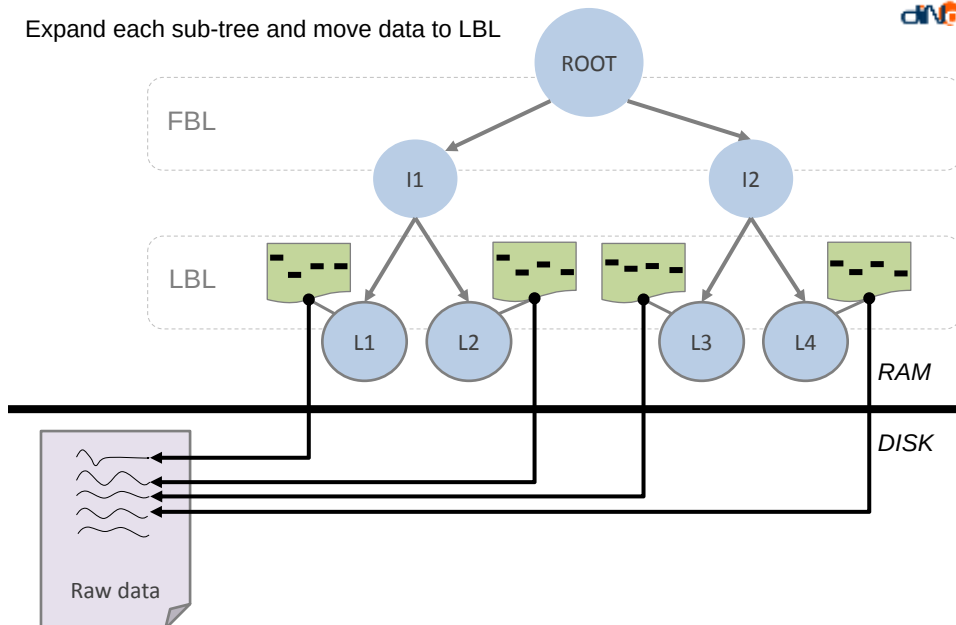
Continue loading data until we run out of memory



Themis Palpanas - Telecom Paristech, Apr 2017

182

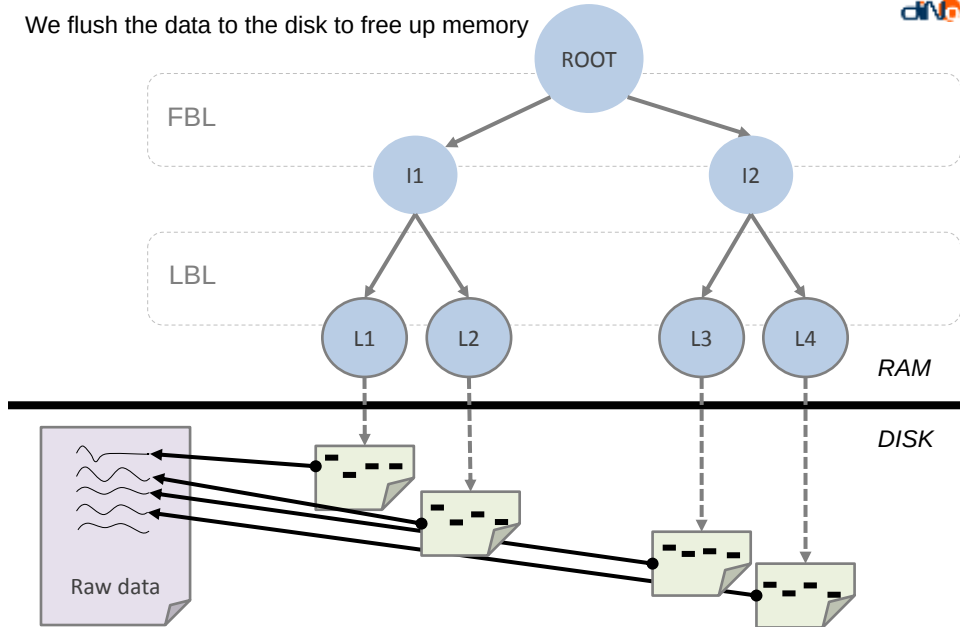
Expand each sub-tree and move data to LBL



Themis Palpanas - Telecom Paristech, Apr 2017

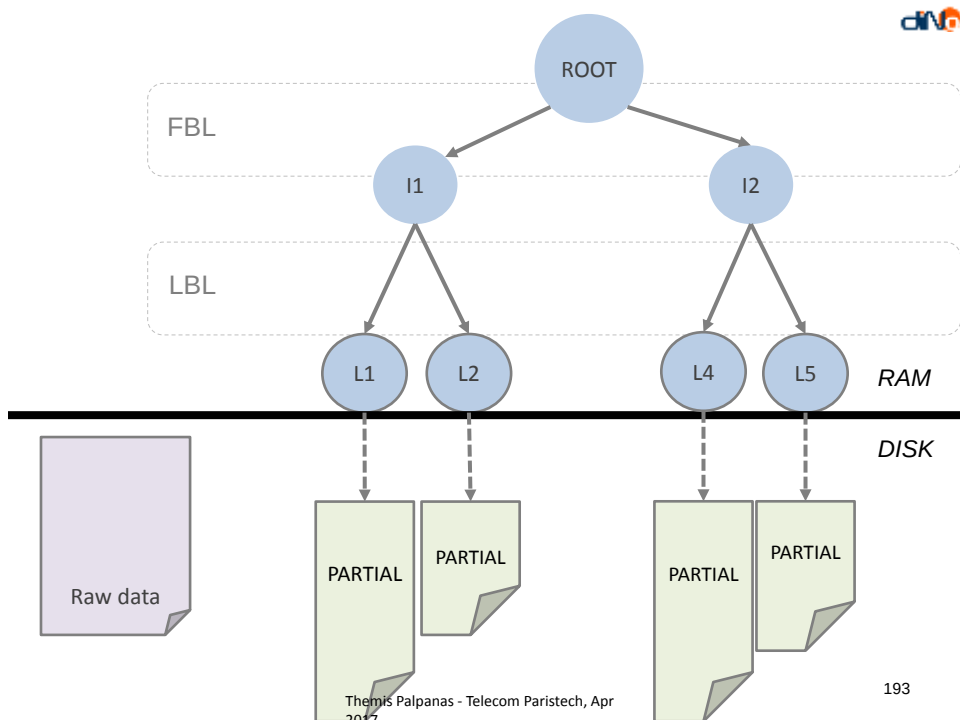
183

We flush the data to the disk to free up memory



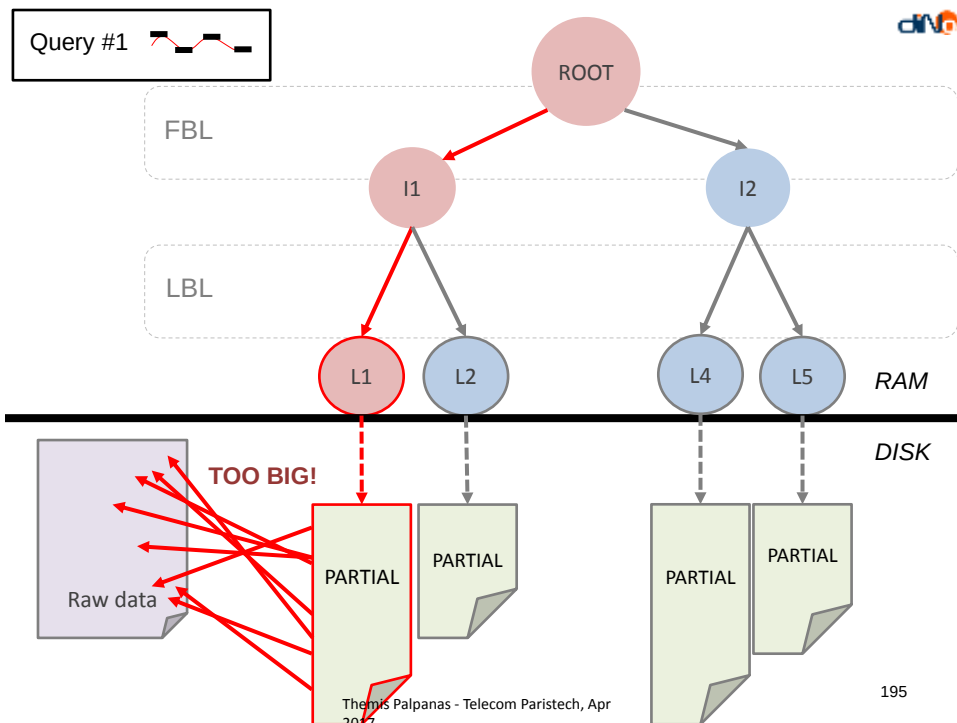
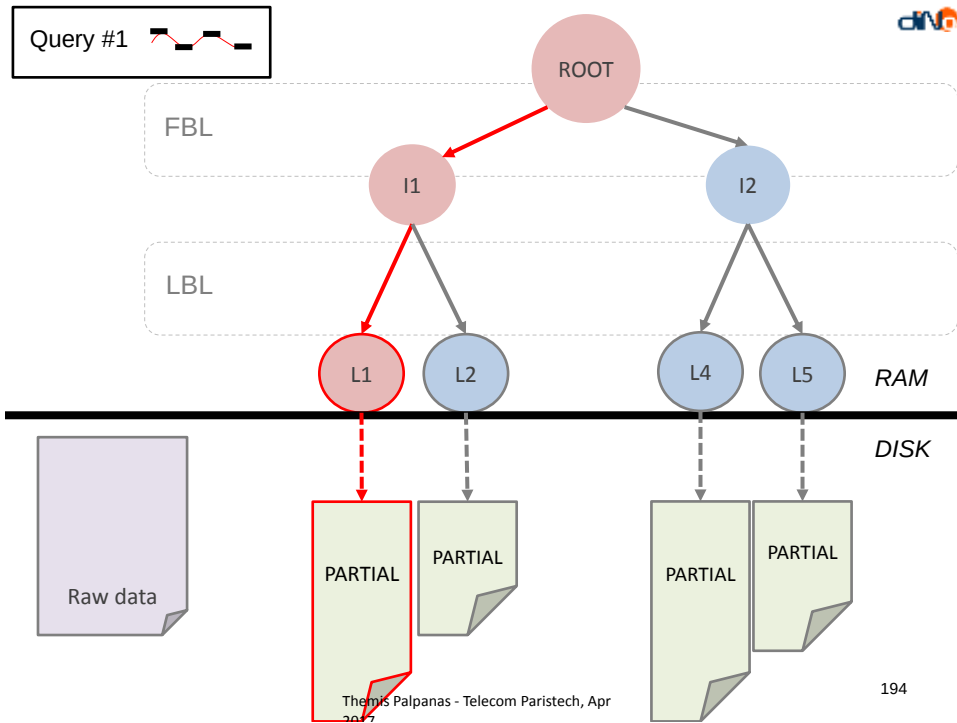
Themis Palpanas - Telecom Paristech, Apr 2017

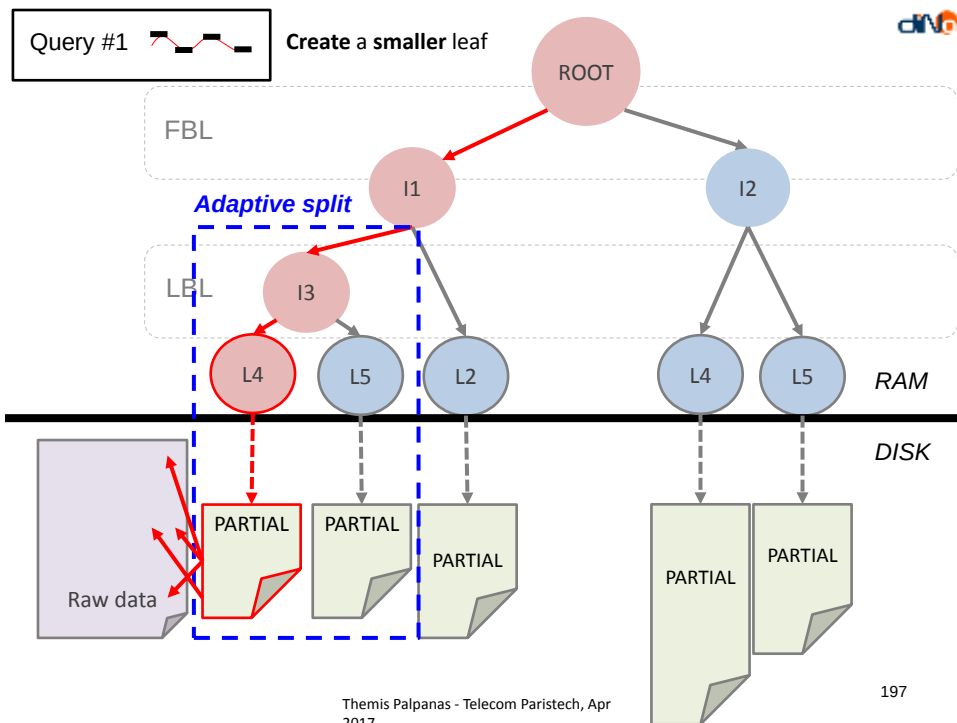
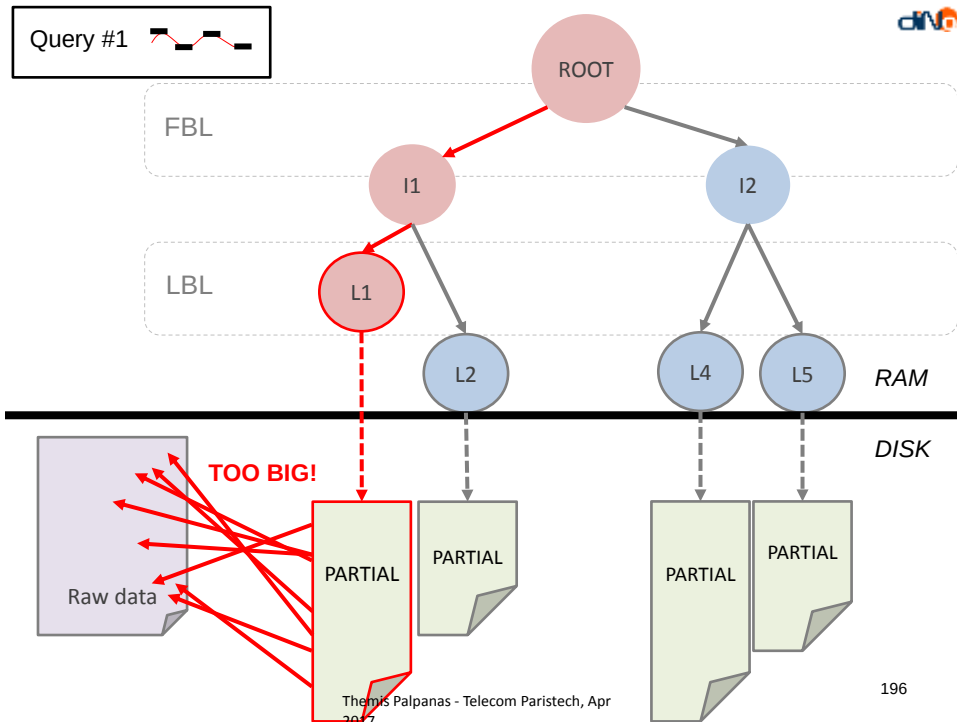
184

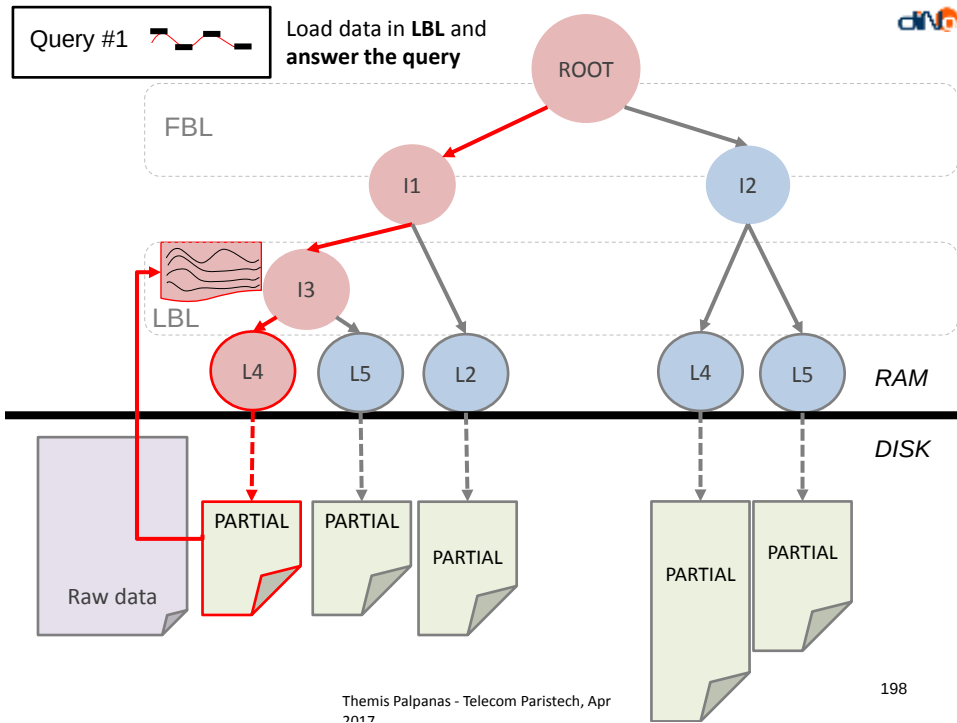


Themis Palpanas - Telecom Paristech, Apr 2017

193

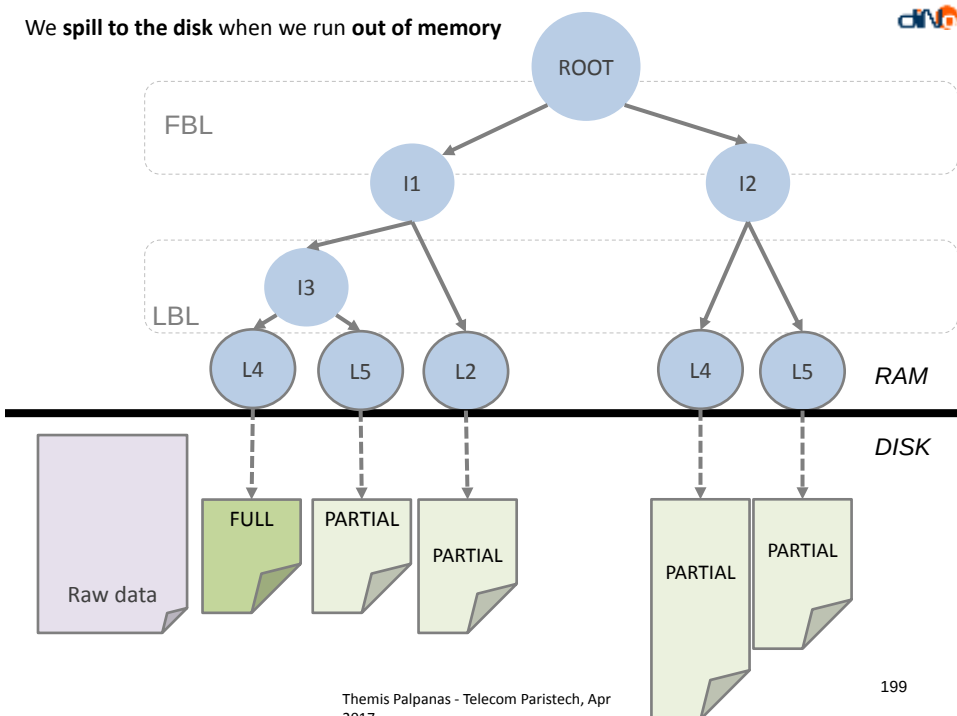






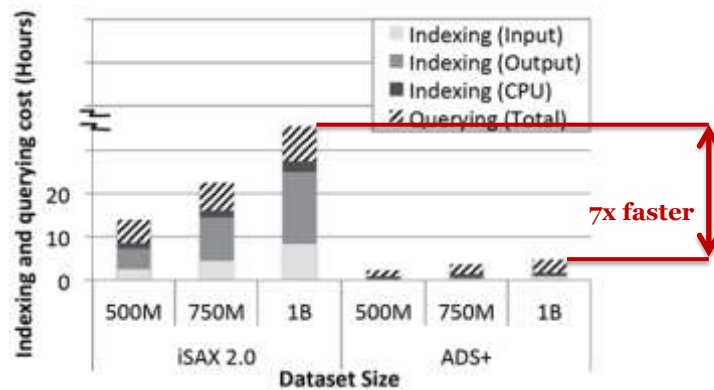
198

We **spill to the disk** when we run **out of memory**



199

Experimental Evaluation



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

Themis Palpanas - Telecom Paristech, Apr 2017

Comparison to *multi-dimensional indices*

measure data-to-query time
(just index 1 **billion** data-series)



1-3 orders of magnitude faster than multi-dimensional indexing methods

Themis Palpanas - Telecom Paristech, Apr 2017

201

Demo

- <http://www.mi.parisdescartes.fr/~themisp/rinse/>

Publications

VLDB'15



Themis Palpanas - Telecom Paristech, Apr 2017

Related Work

- significant amount of work in data series indexing
 - e.g., TS-Tree [Assent et al. '08], iSAX [Shieh & Keogh '08]
- **none** of these approaches
 - considered bulk loading
 - examined more than 1 Million data series
- several studies for index bulk loading
 - merge-based assume data is **pre-clustered** [Choubey et al. '99]
 - buffering-based work only for **balanced** indices [Arge et al. '02] [Van den Bercken & Seeger '01] [Soisalon-Soininen & Widmayer '03]
- Adaptive indexing/file reorganization for column stores
 - Database cracking [Idreos et al. '07], raw file cracking [Idreos et al. '11]

Themis Palpanas - Telecom Paristech, Apr 2017

Distribution/Parallelization/Cloud?

Themis Palpanas - Telecom Paristech, Apr 2017

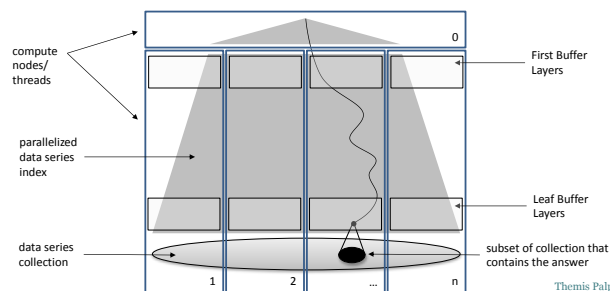
Distribution/Parallelization/Cloud?

- discussion so far assumed 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 - Telecom Paristech, Apr 2017

Distribution/Parallelization/Cloud?

- further scale-up and scale-out possible!
 - techniques inherently parallelizable
 - across cores, across machines
 - more involved solutions required when optimizing for energy
 - minimize total work



Conclusions

- proposed iSAX 2.0, iSAX 2.0 Clustered, iSAX2+, ADS+
 - indexing for very large data series collections
 - code and datasets: <http://www.mi.parisdescartes.fr/~themisp/isax2plus/>
 - current **state of the art**
- experimentally validated proposed approach
 - **first** published experiments with **1 Billion** data series

Conclusions

- proposed iSAX 2.0, iSAX 2.0 Clustered, iSAX2+, ADS+
 - indexing for very large data series collections
 - code and datasets: <http://www.mi.parisdescartes.fr/~themisp/isax2plus/>
 - current **state of the art**
- experimentally validated proposed approach
 - **first** published experiments with **1 Billion** data series
- case studies in diverse domains exhibit **usefulness** of approach
 - for the first time enable **pain-free** analysis of existing, vast collections of data series

Themis Palpanas - Telecom Paristech, Apr 2017

What Next?



new challenge: index and mine **10 billion** data series

Themis Palpanas - Telecom Paristech, Apr 2017

What Next?

- infrastructure monitoring
 - Facebook wants to manage 4B data series, 12M new values/sec



Themis Palpanas - Telecom Paristech, Apr 2017

What Next?



Themis Palpanas - Telecom Paristech, Apr 2017

What Next?



Themis Palpanas - Telecom Paristech, Apr 2017

What Next?



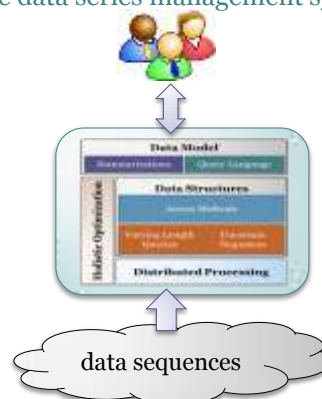
The Road Ahead

“enable practitioners and non-expert users to easily and efficiently manage and analyze massive data series collections”

Themis Palpanas - Telecom Paristech, Apr 2017

The Road Ahead

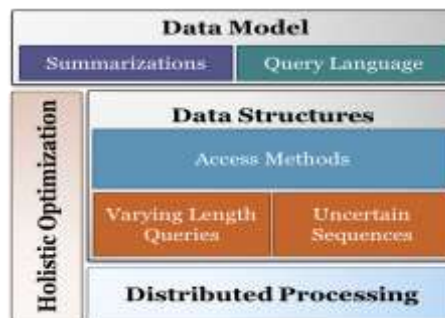
- Big Sequence Management System
 - general purpose data series management system



Themis Palpanas - Telecom Paristech, Apr 2017

The Road Ahead

- Big Sequence Management System



Themis Palpanas - Telecom Paristech, Apr 2017

collaborations!

Themis Palpanas - Telecom Paristech, Apr 2017



Current and Past Collaborations

- Infrastructure Monitoring
 - analysis and mining of hardware and software infrastructure for **health monitoring**
 - with *Facebook, EDF, Safran*
- Human Behavior Patterns
 - identification of different social groups, and analysis of their macro- and micro-**patterns of behavior**
 - with *IBM Research, Telecom Italia*
- Human Brain Activity
 - analysis of fully-detailed neurobiological data for explaining **brain functions**
 - with *ICM*
- Green Manufacturing
 - analysis and optimization of manufacturing processes for **energy savings**
 - with *SAP, Intel, Volvo, Infineon*
- eCrime
 - identification of **fraudulent activities** related to the telecommunication industry
 - with *Telecom Italia, Vodafone, Wind*
- World Sentiments and Opinions
 - analysis of **aggregate sentiment** for different social groups, **role of media** in public sentiment
 - with *Qatar Computing Research Institute, and Hewlett-Packard Labs*

Themis Palpanas - Telecom Paristech, Apr 2017

258

Data-Intensive and Knowledge-Oriented systems

