Data Series Management
The Next Challenge

Themis Palpanas
Paris Descartes University

Data Science DK - DABAI
Copenhagen, March 2017

References

- papers
  - **ADS: The Adaptive Data Series Index**. VLDBJ 2016
    - http://www.mi.parisdescartes.fr/~themisp/publications/vldbj16-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
  - ***i*SAX 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/DSStat
- 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





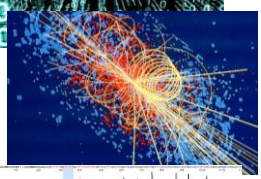




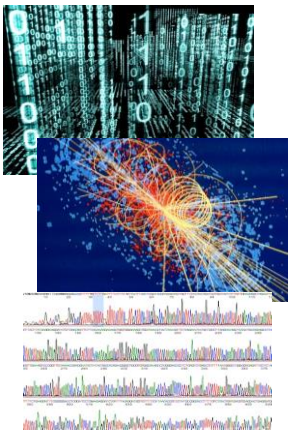


Executive Summary

- 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


Executive Summary



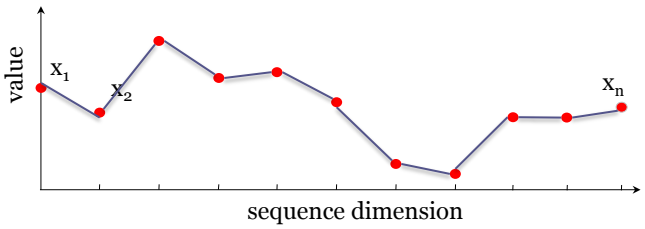
- data collected at unprecedented rates
- they enable data-driven scientific discovery
- lots of these data are sequences
  - takes **days-weeks** to analyze big sequence collections

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

Data series

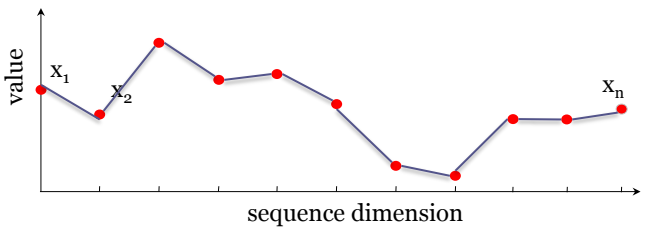
Data series

- Sequence of points ordered along some dimension



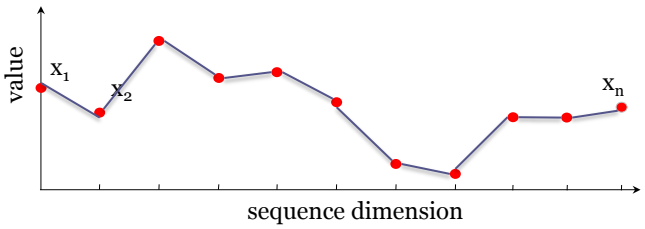
Data series

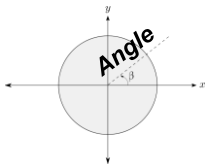
- Sequence of points ordered along some dimension

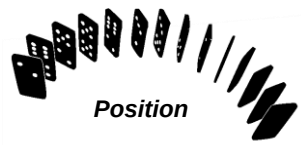


Data series

- Sequence of points ordered along some dimension

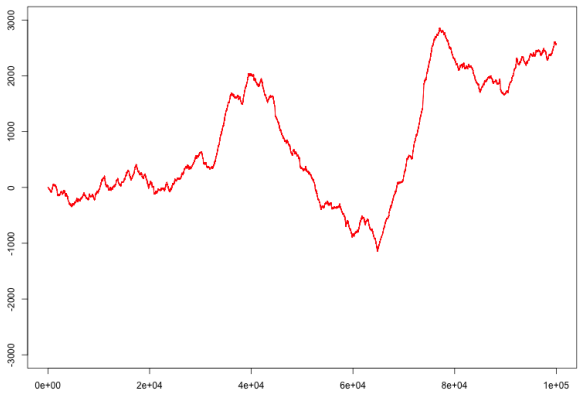




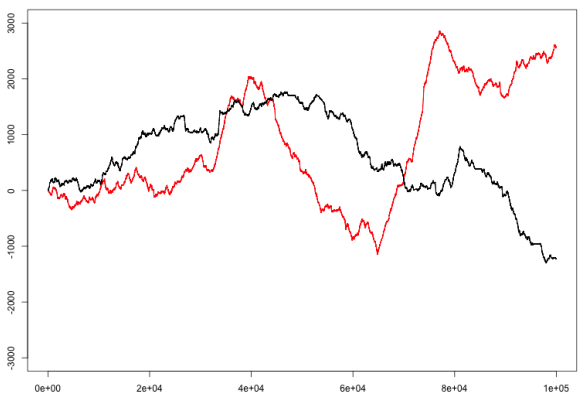


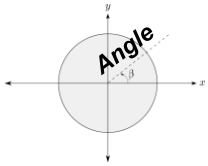
GTCAATGGCCAGGATATTAGAACAGTACTCTGTGAACCCTATTTATGGTGGCACCCCTTAGACTAA
GATAACACAGGGAGCAAGAGGTTGACAGGAAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAG
AGAAGTGCTAAGTCTCCTTTCTAAGGCACATGATGGATTCAAGGGAAAGCCACATTTGACTAAAGC
CCAAGGGATTGTTGCTTCTAATCCGATTTCTTGGCAGAAGATATTACAAACTAAGAGTCAGATTAA
TATGTGGGTGCCAAAATAAATAAACAAATAATTGAATAATCCCTGGAGGTTTAAGTGAGGAGAAA
CTCCTCCACAGCTTGCTACCGAGGCAGAACCGGTTGAAACTGAAATGCATCCGCCGCCAGAGGATC
TGTAAAAGAGAGGTTGTTACGAAACTGGCAACTGCCAACCAAAGTCCACCAATGGACAAGCAAAA
AAGAGCACTCATCTCATGCTCCCAAGGATCAACCTTCCCAGAGTTTTCACTTAAGTGGCCACCAAG
CCAGTTGTCAATCCAGGGCTTTGGACTGAAATCTAGGGCTTCATCCGCTACCTCAGAGTGTCTTCT
ATTTCTTCCAGCCAGTGACAAATACAACAAACATCTGAGATGTTTTAGCTATAAATCCTTTACAATT
GTTATTTATGTCTTAACTTTTGTTATACCTGGAAAAGTAGGGGAAACAATAAGAACATACTGTCTT
GGCCAAGCATCCAAGGTTAAATGAGTTATGGAAATTCATTTGGGAGCCAAGACATTGCACGTGGT
TATTTATTAGTCACCCAAGCATGTATTTTGCATGTCCATCAGTTGTTCTTGGCCAAAAGAGCAGAAT
CAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTGCAGGGACAAGTCTGCAAGATGAGCATT
GAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCAGGCACTTACAAGAGCCCAGGTTGTTGC
CATGTTTGTTTTTGCAACTTGTCTATTTAAAGAGATTTGGGCAATGGCCAGGATATTAGAACAGTA
CTCTGTGAACCCTATTTATGGTAGCACCCCTTAGACTAAGATAACACAGGGAGCAAGAGGTTGACA
GGAAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAGAGAAGTGCTAAGTCTCCTTTCTAAGGCA
CATGATGGATCAAGGGAAAGTCACATTTGACTAAAGCCCAAGGGATTGTTGCTTCTAATCCGATTC
TTGGCAGAAGATATTGCAAACTAAGAGTCAGATTAATATGTGGGTGCCAAAATAAATAAACAAATA
ATTGAATAATCCCTGGAGGTTTAAGTGAGGAGAAACTCCTCCACACTTGCTACCGAGGCAGAACCG
GTTGAAACTGAAATGCACCCGCTGCCAGATTTATTAGTCACCCAAGCATGTATTTTGCATGTCCAT
AAGAACAGAATCAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTG
GATGAGCATTGAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCA
AGGTTGTTGCCATGTTTGTTTTTGCAACTTGTCTTTTAAACAGATTTGA

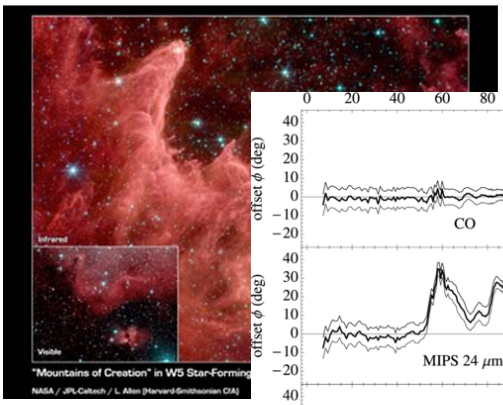
GTCAATGGCCAGGATATTAGAACAGTACTCTGTGAACCCTATTTATGGTGGCACCCCTTAGACTAA
GATAACACAGGGAGCAAGAGGTTGACAGGAAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAG
AGAAGTGCTAA TTGACTAAAGC
CCAAGGGATTG AGTCAGATTAA
TATGTGGGTGC TGAGGAGAAA
CTCCTCCACAGC GCCAGAGGATC
TGTAAAAGAGA ACAAGCAAAA
AAGAGCACTCA GGCCACCAAG
CCAGTTGTCAA AGTGTCTTCT
ATTTCTTCCAGC CCTTTACAATT
GTTATTTATGTC CATACTGTCTT
GGCCAAGCATC TGCACGTGGT
TATTTATTAGTC AAGAGCAGAAT
CAATGAGCCGC AGATGAGCATT
GAAGAGGATGC CAGGTTGTTGC
CATGTTTGTTTT TTAGAACAGTA
CTCTGTGAACC AGAGGTTGACA
GGAAAGCCAGG TTTCTAAGGCA
CATGATGGATC TAATCCGATTC
TTGGCAGAAGA ATAAACAAATA
ATTGAATAATCC AGGCAGAACCG
GTTGAAACTGAAATGCACCCGCTGCCAGATTTATTAGTCACCCAAGCATGTATTTTGCATGTCCAT
AAGAACAGAATCAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTG
GATGAGCATTGAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCA
AGGTTGTTGCCATGTTTGTTTTTGCAACTTGTCTTTTAAACAGATTTGA

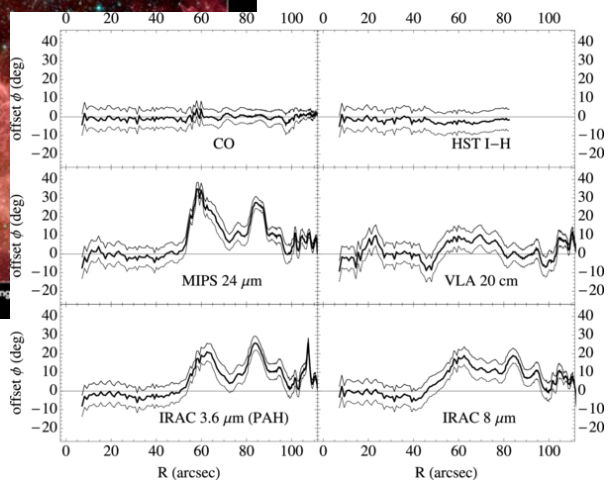


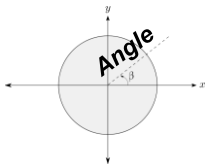
GTCAATGGCCAGGATATTAGAACAGTACTCTGTGAACCCTATTTATGGTGGCACCCCTTAGACTAA
GATAACACAGGGAGCAAGAGGTTGACAGGAAAGCCAGGGGAGCAGGGAAGCCTCCTGTAAAGAG
AGAAGTGCTAA TTGACTAAAGC
CCAAGGGATTG AGTCAGATTAA
TATGTGGGTGC TGAGGAGAAA
CTCCTCCACAGC GCCAGAGGATC
TGTAAAAGAGA ACAAGCAAAA
AAGAGCACTCA GGCCACCAAG
CCAGTTGTCAA AGTGTCTTCT
ATTTCTTCCAGC CCTTTACAATT
GTTATTTATGTC CATACTGTCTT
GGCCAAGCATC TGCACGTGGT
TATTTATTAGTC AAGAGCAGAAT
CAATGAGCCGC AGATGAGCATT
GAAGAGGATGC CAGGTTGTTGC
CATGTTTGTTTT TTAGAACAGTA
CTCTGTGAACC AGAGGTTGACA
GGAAAGCCAGG TTTCTAAGGCA
CATGATGGATC TAATCCGATTC
TTGGCAGAAGA ATAAACAAATA
ATTGAATAATCC AGGCAGAACCG
GTTGAAACTGAAATGCACCCGCTGCCAGATTTATTAGTCACCCAAGCATGTATTTTGCATGTCCAT
AAGAACAGAATCAATGAGCCGCTGCAGATGCAGACATAGCAGCCCCTTG
GATGAGCATTGAAGAGGATGCACAAGCCCGGTAGCCCGGGAAATGGCA
AGGTTGTTGCCATGTTTGTTTTTGCAACTTGTCTTTTAAACAGATTTGA







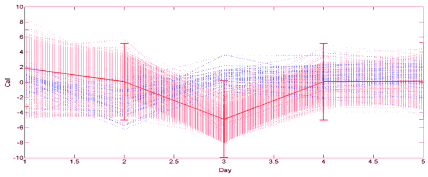




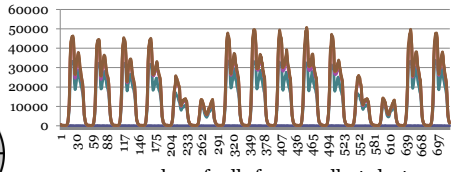
Schinnerer et al.

Telecommunications

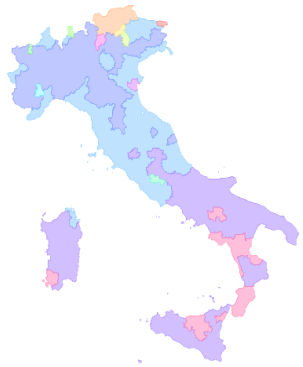
- analysis of call activity patterns
  - Telecom Italia



call activity for Easter Monday



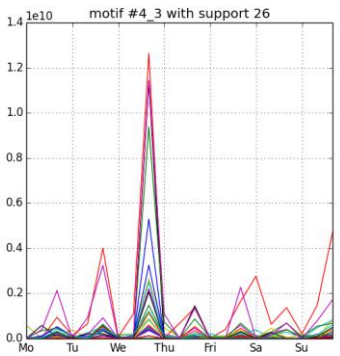
average number of calls for 5 smallest clusters



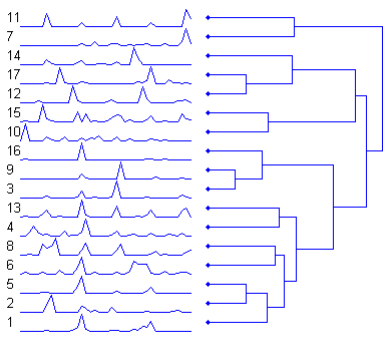
clustermap of incoming calls time series

Home Networks

- temporal usage behavior analysis of home networks
  - Portugal Telecom



(previously unknown) frequent behavior pattern

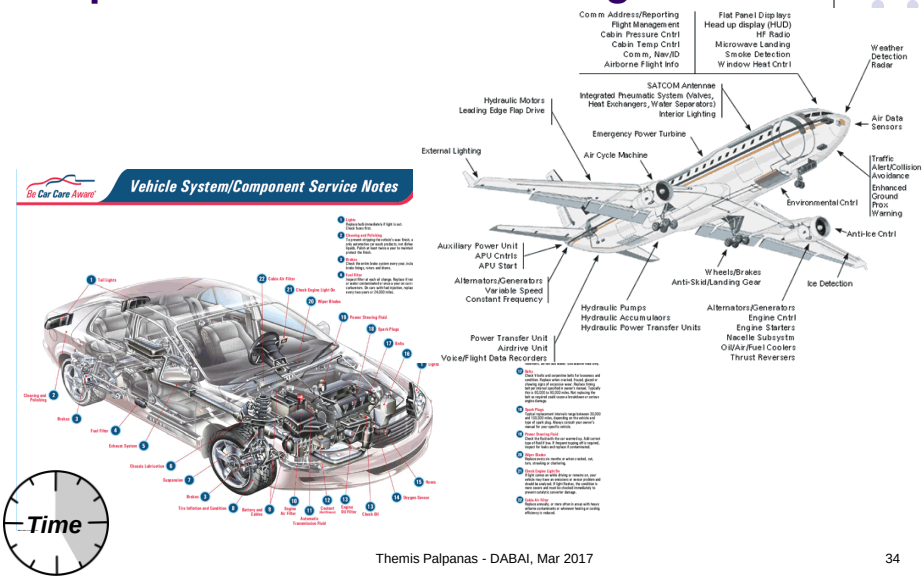


clustering based on user activity patterns

Operation Health Monitoring

Operation Health Monitoring

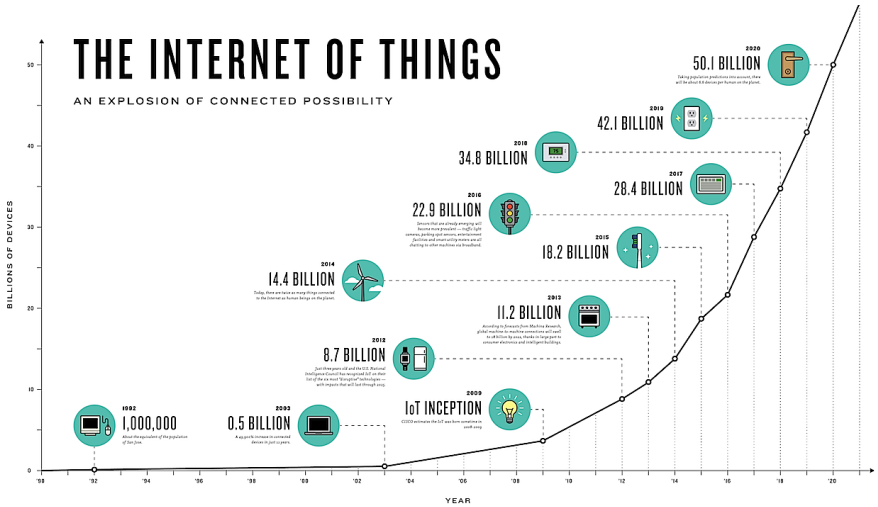
Operation Health Monitoring



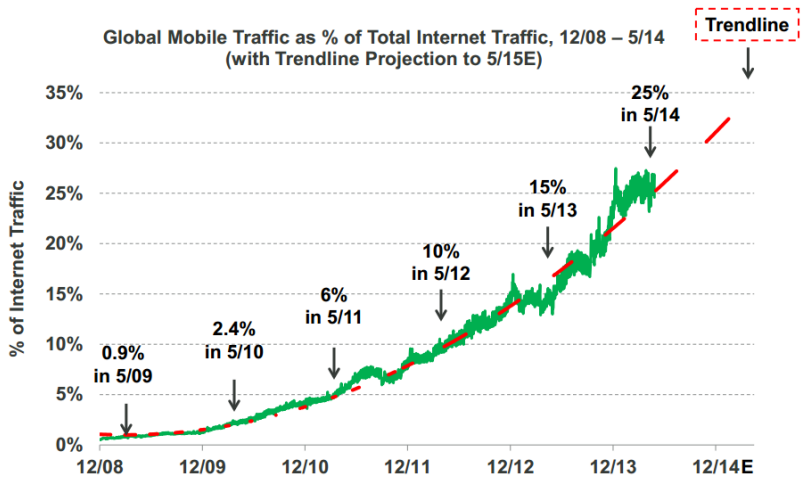

INTERNET of
THINGS


THE INTERNET of
THINGS

Internet of Things: Number of Connected Devices



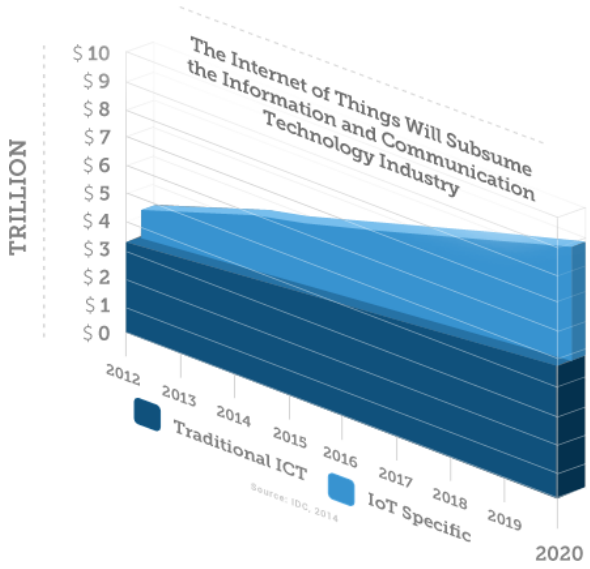
Internet of Things: Percentage of Network Traffic



@KPCB Source: StatCounter Global Stats, 5/14. Note that PC-based Internet data bolstered by streaming.

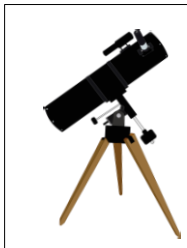
159

Internet of Things: ICT Market Share



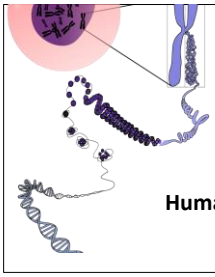
40

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

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

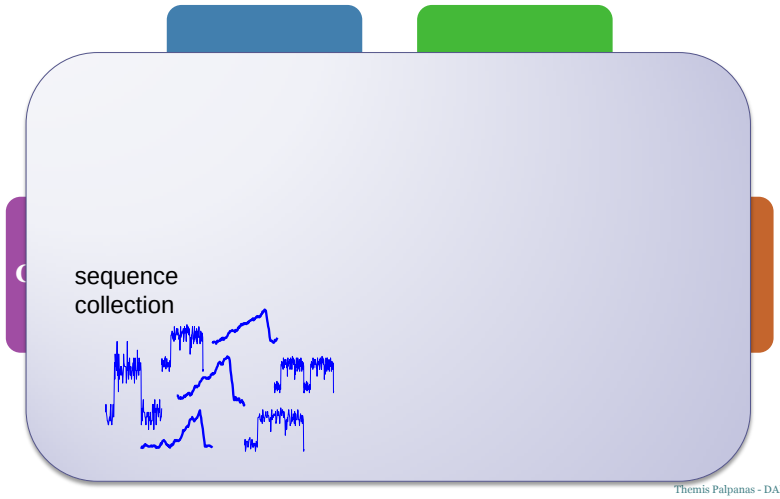
- select values in time interval
- select values in some range
- select some data series
- combinations of those

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

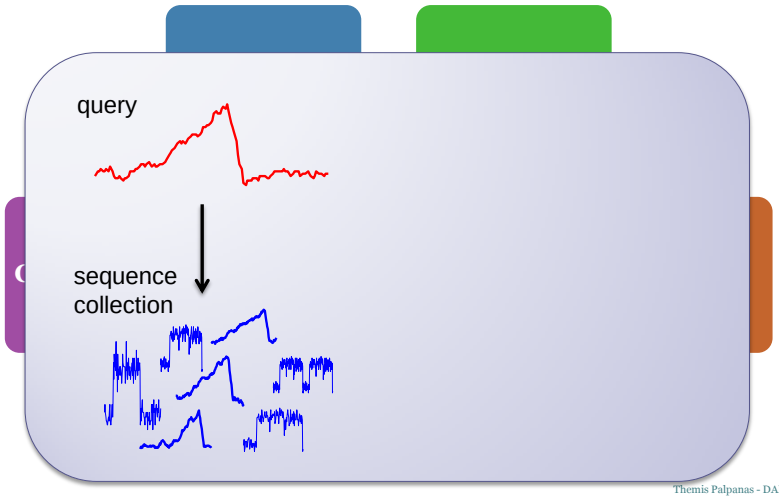
- Clustering
- Outlier Detection
- Classification
- Frequent Pattern Mining

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

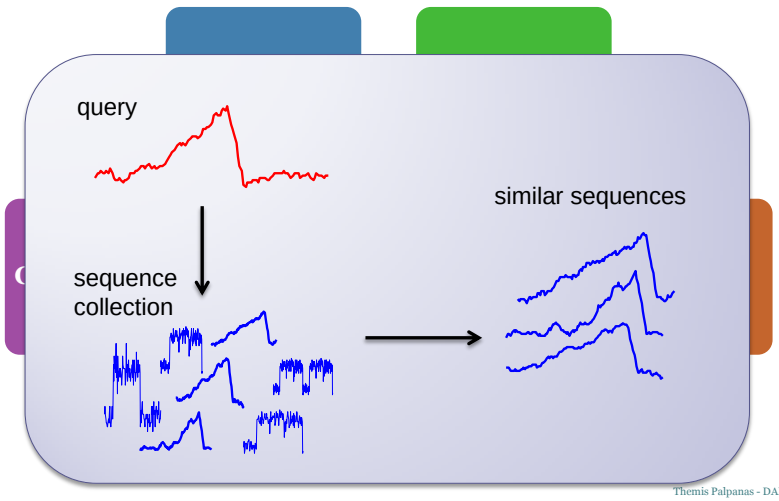
What do we want to do with them? - complex analytics



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



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



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

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

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

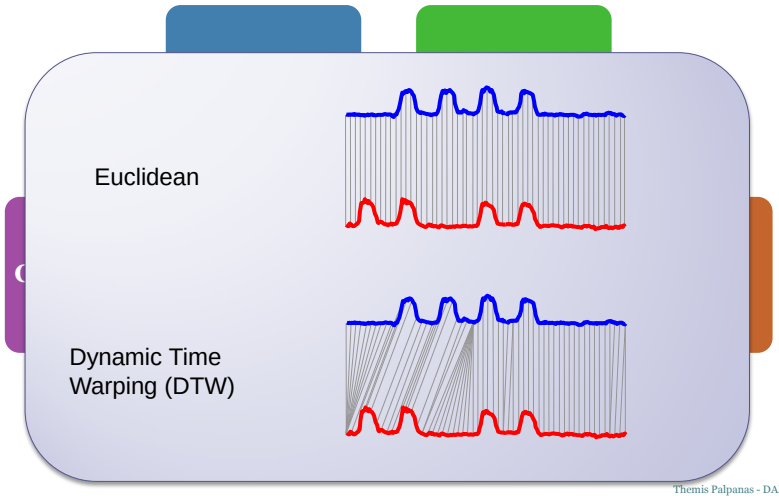
Euclidean $D(X,Y) \equiv \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Dynamic Time Warping (DTW)

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

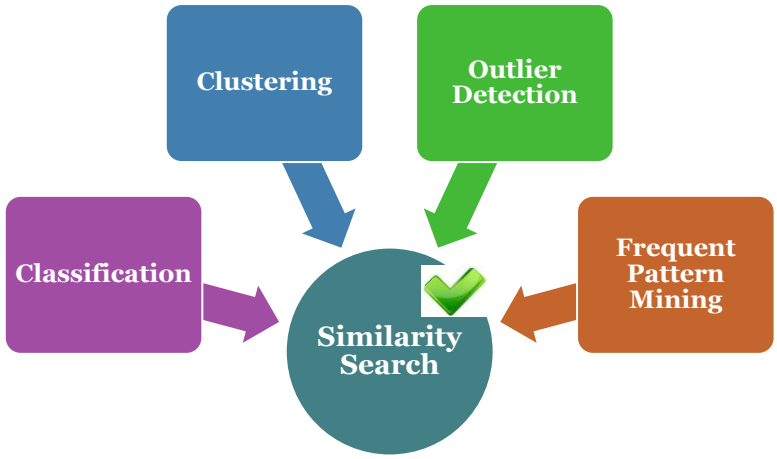
$$f(i,j) = \|x_i - y_j\| + \min\begin{cases} f(i,j-1) \\ f(i-1,j) \\ f(i-1,j-1) \end{cases}$$

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

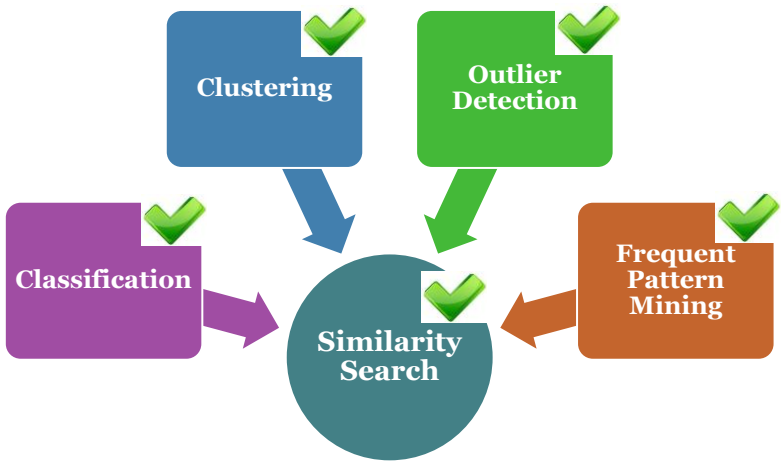


What do we want to do with them? - complex analytics

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

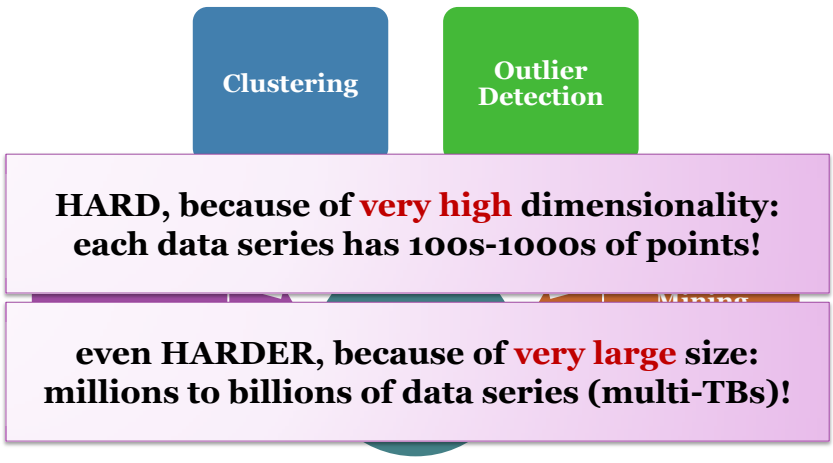


What do we want to do with them? - complex analytics

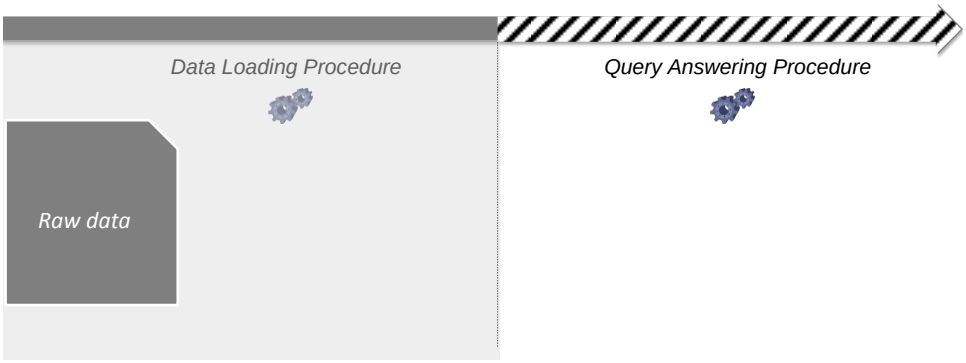


What do we want to do with them?
- complex analytics

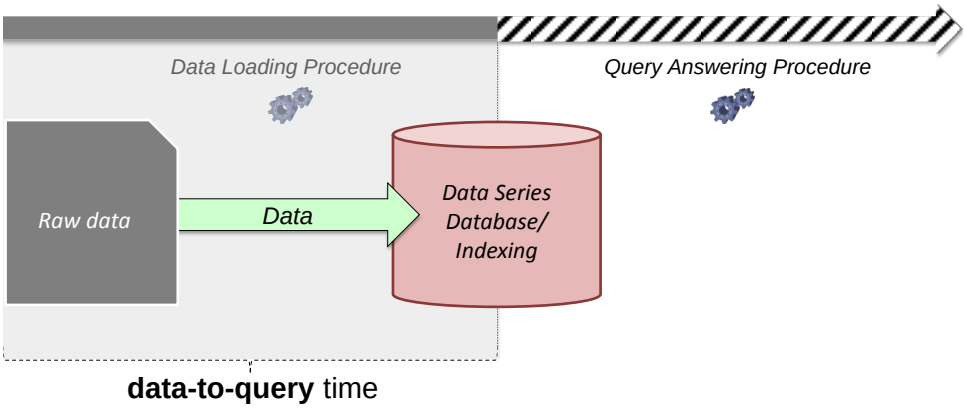
What do we want to do with them?
- complex analytics



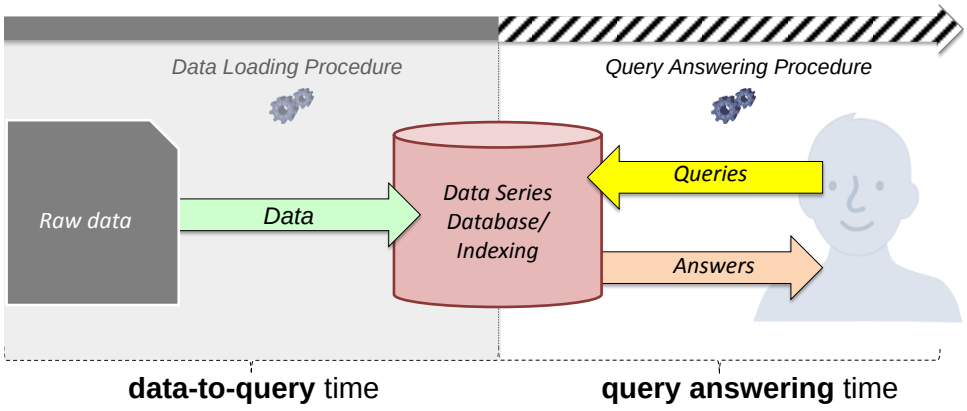
Query answering process



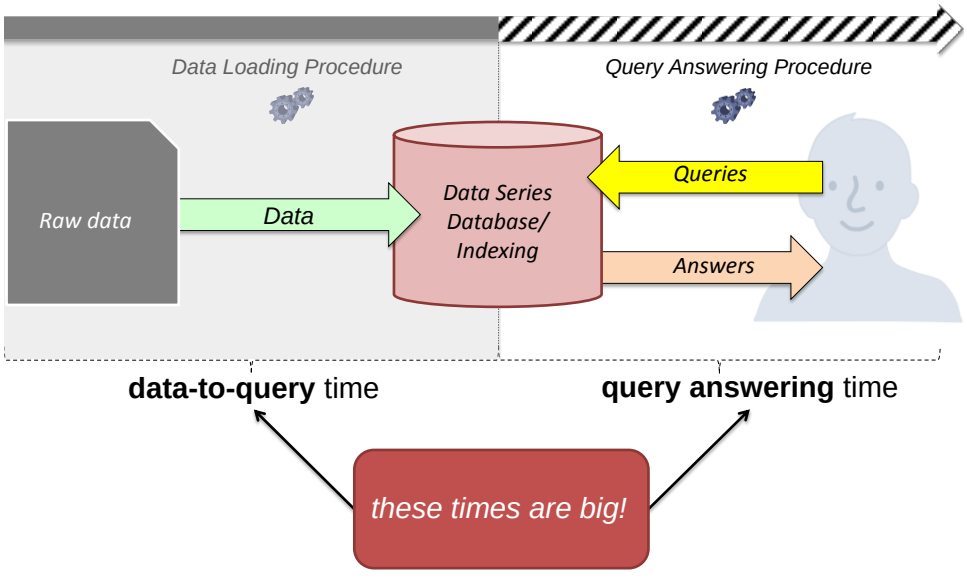
Query answering process



Query answering process

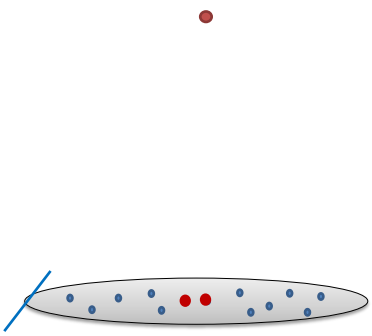


Query answering process



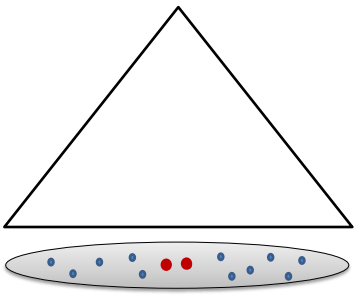
Similarity Search via
Serial Scan

Similarity Search via
Serial Scan

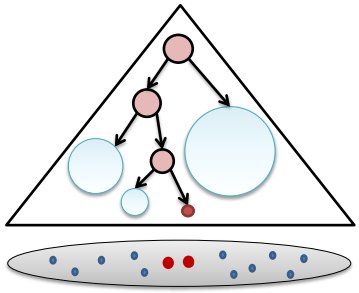


Similarity Search via
Serial Scan

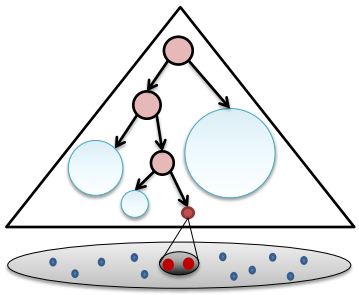
Similarity Search via
Indexing



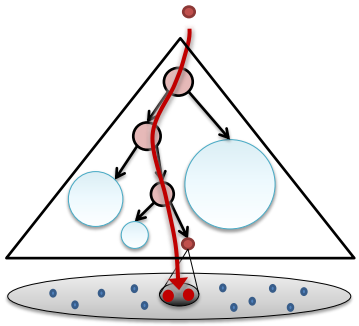
Similarity Search via Indexing



Similarity Search via Indexing



Similarity Search via
Indexing

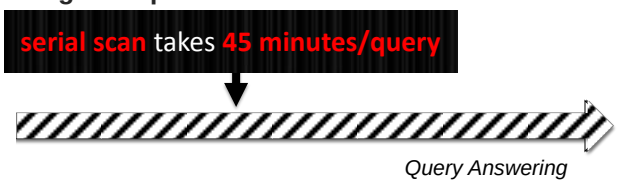


Traditional Approaches

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

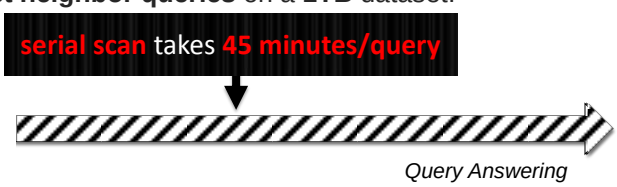
Traditional Approaches

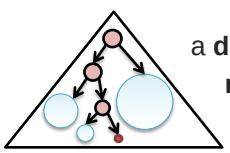
answer **nearest neighbor queries** on a **1TB** dataset:



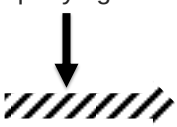
Traditional Approaches

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





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



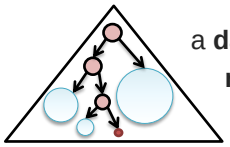
Traditional Approaches

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

serial scan takes **45 minutes/query**

Query Answering

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



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

Query Answering

Traditional Approaches

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

serial scan takes **45 minutes/query**

Query Answering

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

indexing a 1TB dataset takes **days**

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

Query Answering

Traditional Approaches

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

serial scan takes **45 minutes/query**

Query Answering

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

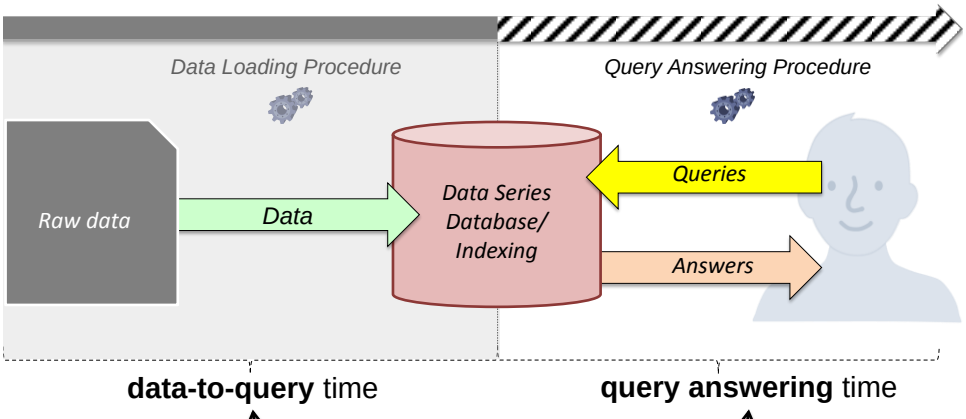
indexing a 1TB dataset takes **days**

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

Query Answering

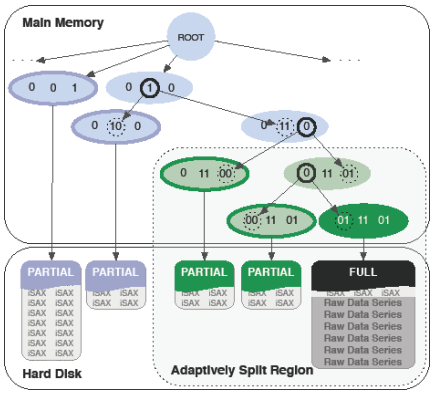
complex analytics in hours/days…

Query answering process



we have proposed the state-of-the-art solutions for both problems!

Our Approach: ADS+



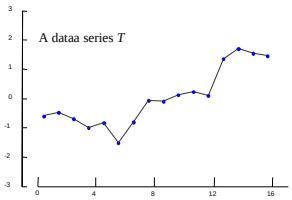
complex analytics in minutes/seconds!

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

SAX Representation

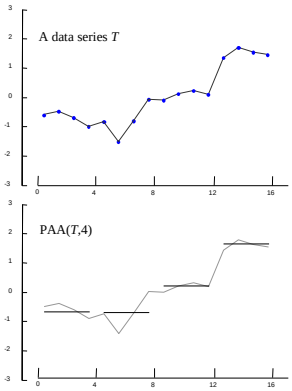
- **S**ymbolic **A**ggregate appro**X**imation (SAX)
  - **(1)** Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)



SAX Representation

- **S**ymbolic **A**ggregate appro**X**imation (SAX)
  - **(1)** Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)
    - T typically normalized to $\mu = 0$, $\sigma = 1$
    - $\text{PAA}(T,w) = \overline{T} = \bar{t}_1, \ldots, \bar{t}_w$

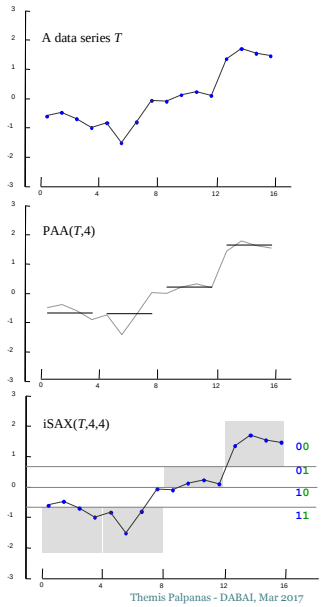
      where $\bar{t}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} T_j$



SAX Representation

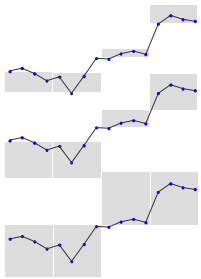
- **S**ymbolic **A**ggregate appro**X**imation (SAX)
  - **(1)** Represent data series T of length n with w segments using Piecewise Aggregate Approximation (PAA)
    - T typically normalized to $\mu = 0$, $\sigma = 1$
    - PAA(T,w) $= \overline{T} = \bar{t}_1, \ldots, \bar{t}_w$

      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 a of symbols



*i*SAX Representation

- *i*SAX offers a bit-aware, quantized, multi-resolution representation with variable granularity



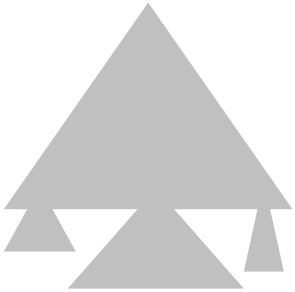
= { 6, 6, 3, 0} = {110 ,110 ,0111 ,000}

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

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

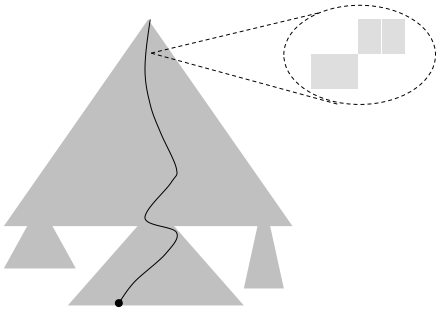
*i*SAX Index

- 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 ≤ ***th***



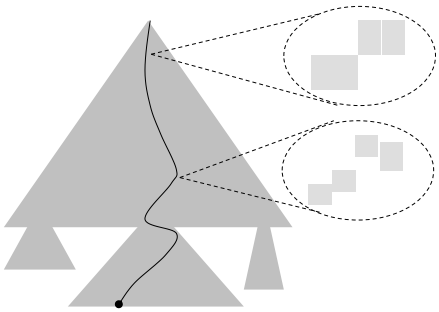
*i*SAX Index

- 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 ≤ ***th***



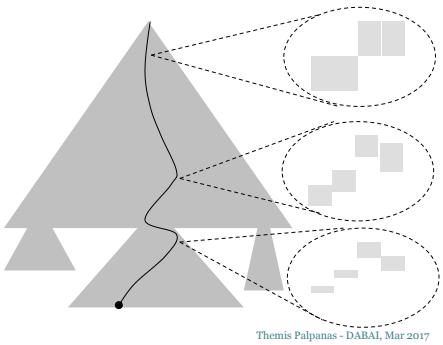
*i*SAX Index

- 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 ≤ ***th***



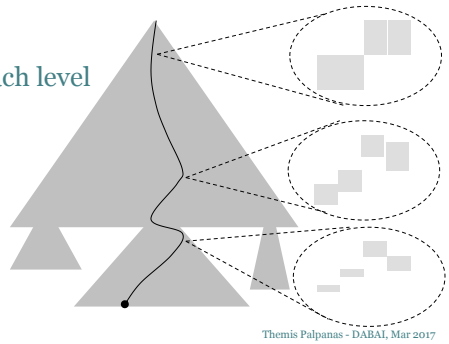
*i*SAX Index

- 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 ≤ ***th***



*i*SAX 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 ≤ ***th***
- Approximate Search
  - Match *i*SAX representation at each level
- Exact Search
  - Leverage approximate search
  - Prune search space
    - Lower bounding distance



iSAX 2.0
Bulk Loading Algorithm

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

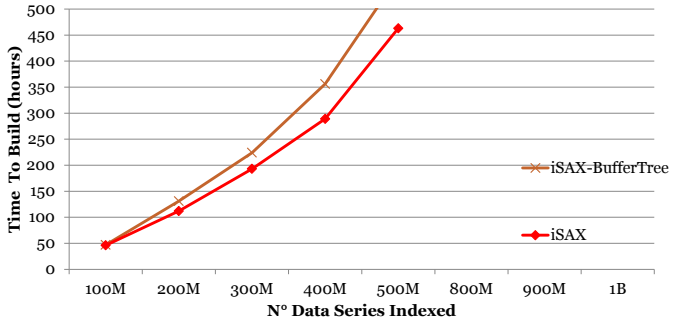
iSAX 2.0
Bulk Loading Algorithm

- 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

iSAX 2.0
Bulk Loading Algorithm

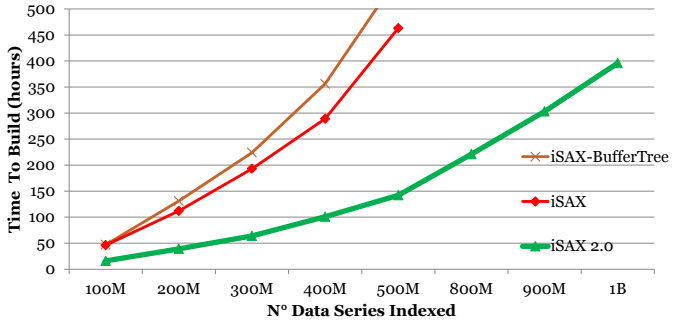
- iterate between two phases (till all data series are indexed):
  - Phase 1
    - read data series and group them according to first-level nodes
    - use all available main memory
  - Phase 2
    - grow index by processing the subtree rooted at each one of the first-level nodes one at-a-time
    -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

Experimental Evaluation Bulk Loading



- previous techniques take too long
  - 20 days for 500 Million data series
  - estimated about 2 months for 1 Billion data series

Experimental Evaluation Bulk Loading

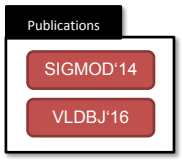


- 1 Billion data series indexed in 16 days: 72% less time
- indexing time per data series: 0.001 sec

Adaptive Data Series Index: ADS+

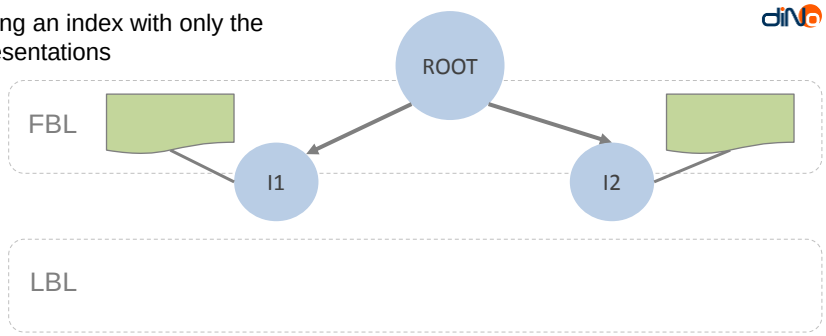
- novel paradigm for building a data series index
  - do not build entire index and then answer queries
  - start answering queries by building the part of the index needed by those queries
- still guarantee correct answers

Adaptive Data Series Index: ADS+

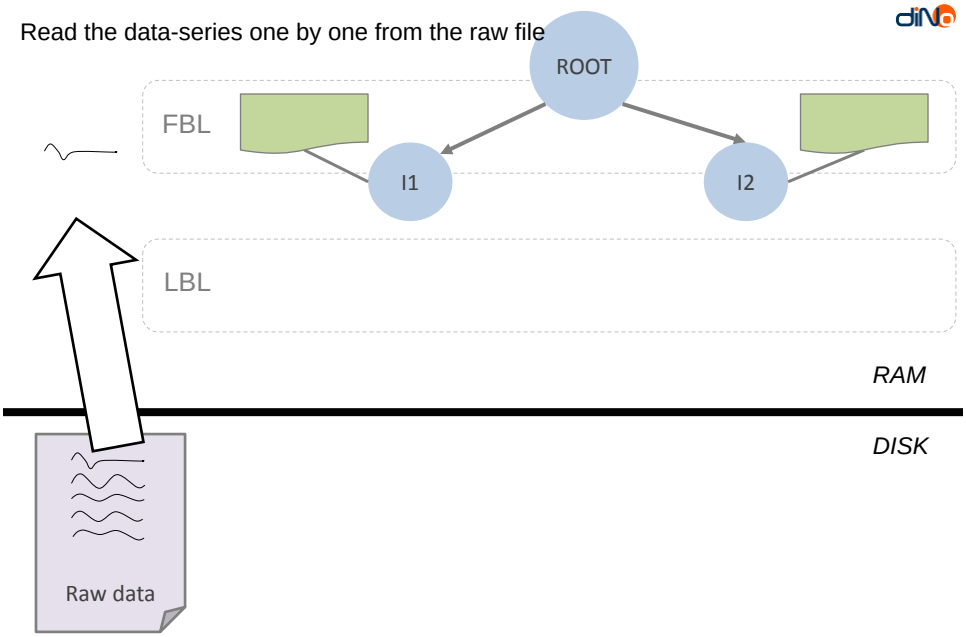


- intuition for proposed solution
  - build the iSAX index using the iSAX representations
    - just like iSAX2+
  - but start with a large leaf size
    - minimize initial cost
  - postpone leaf materialization to query time
    - only materialize (at query time) leaves needed by queries
  - parts that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

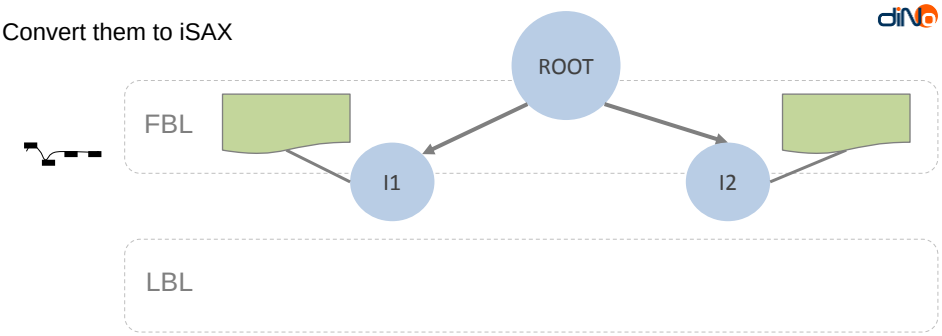
Start building an index with only the iSAX representations



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

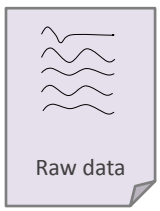


Convert them to iSAX

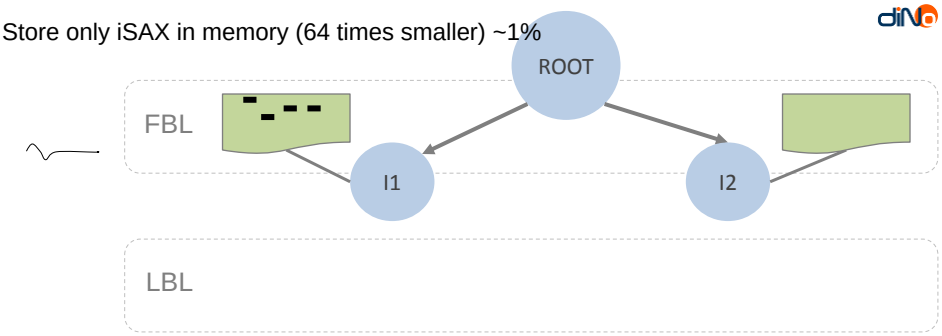


RAM

DISK

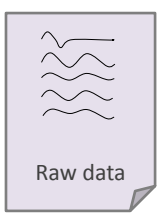


Store only iSAX in memory (64 times smaller) ~1%

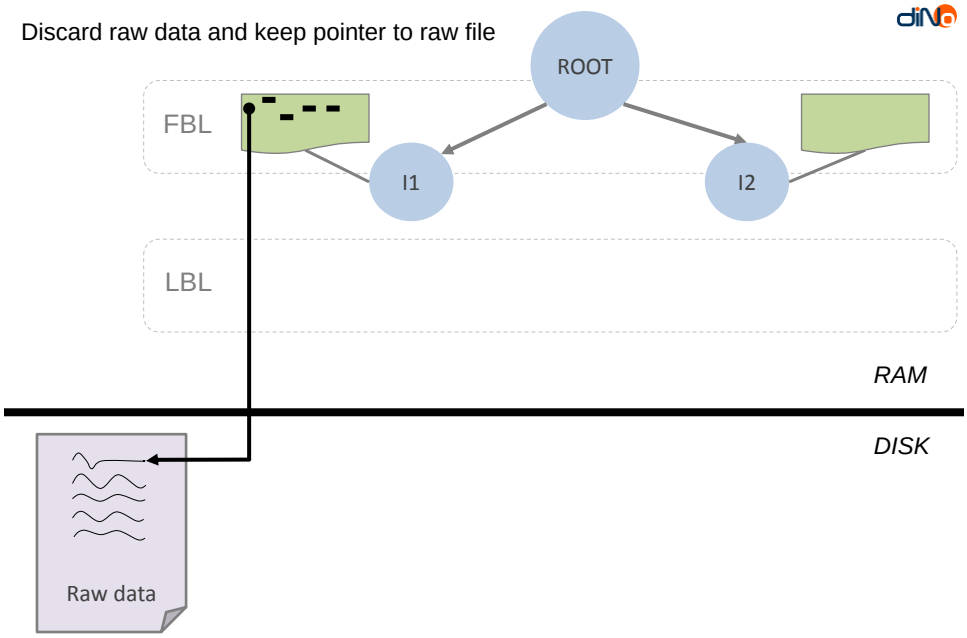


RAM

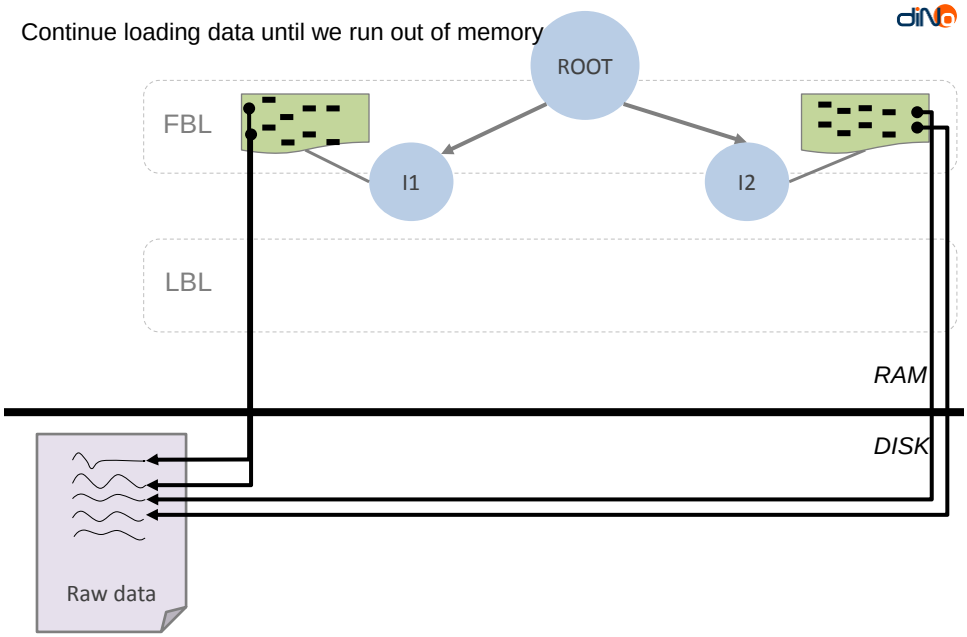
DISK



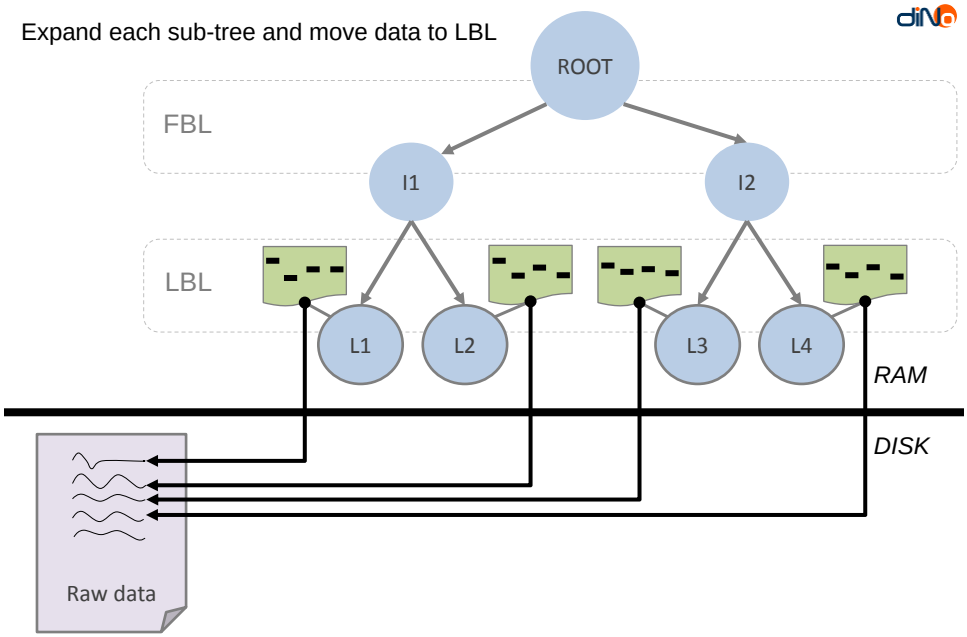
Discard raw data and keep pointer to raw file



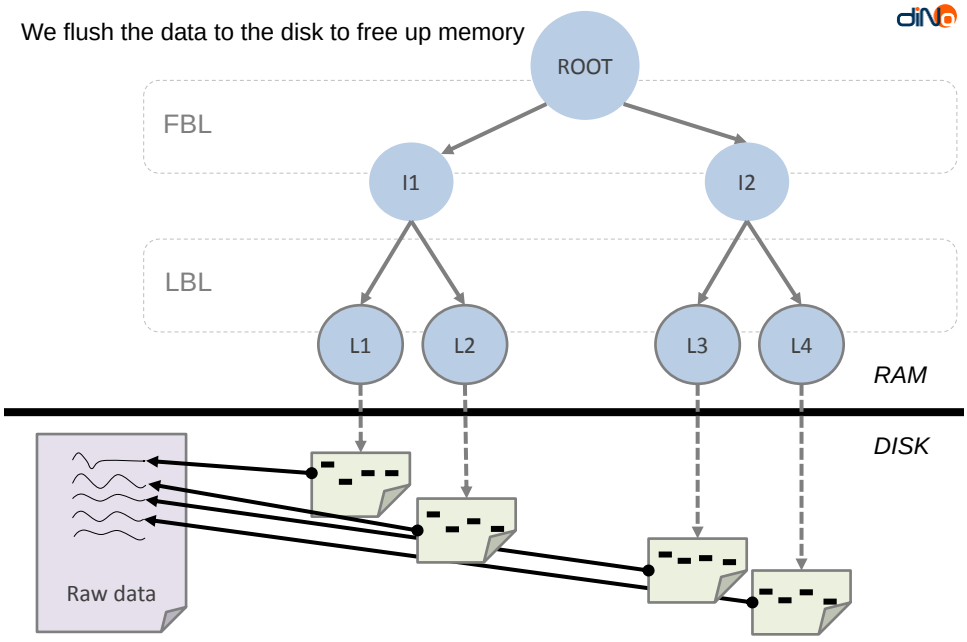
Continue loading data until we run out of memory

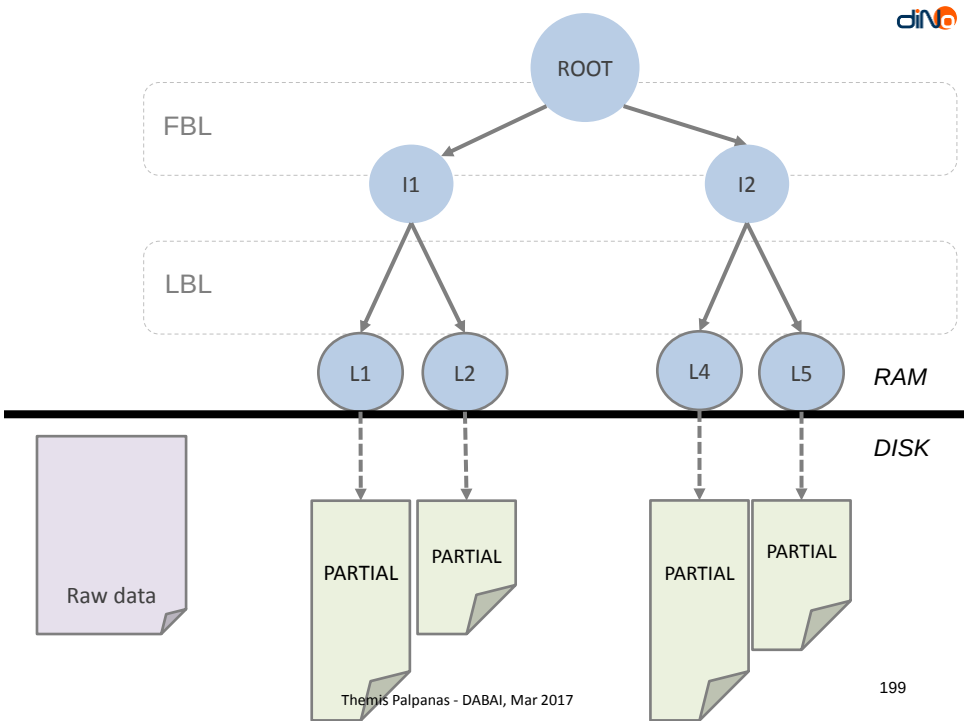


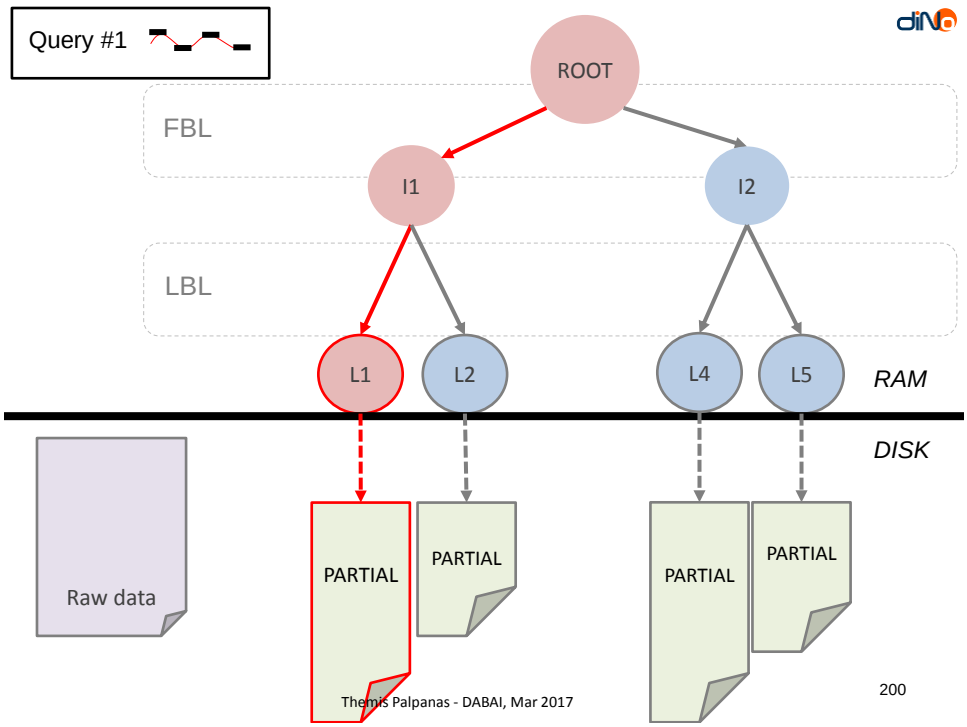
Expand each sub-tree and move data to LBL

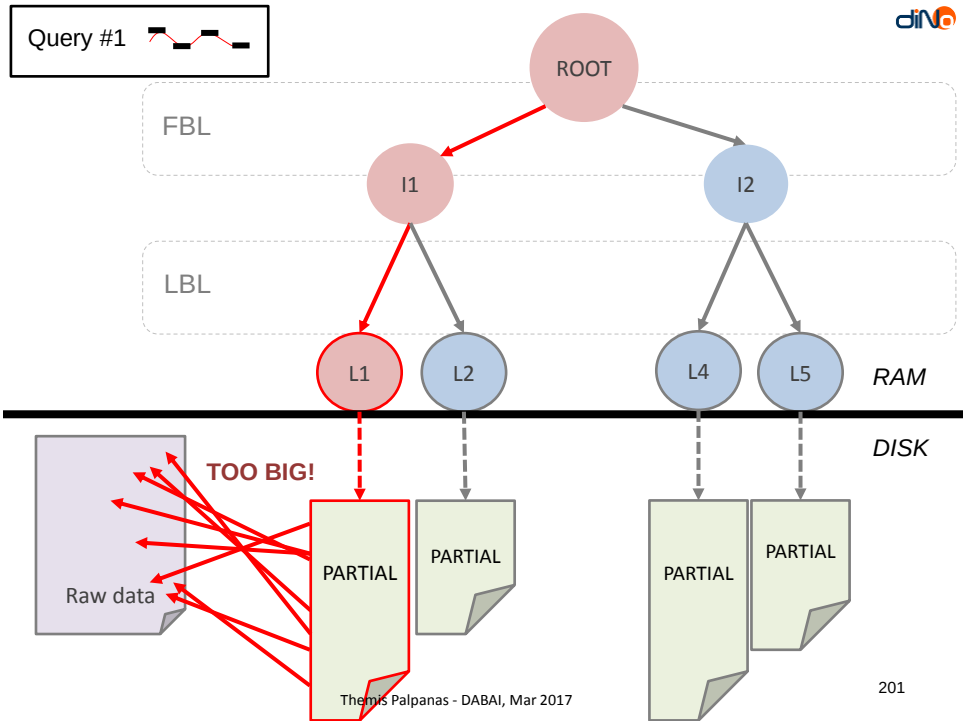


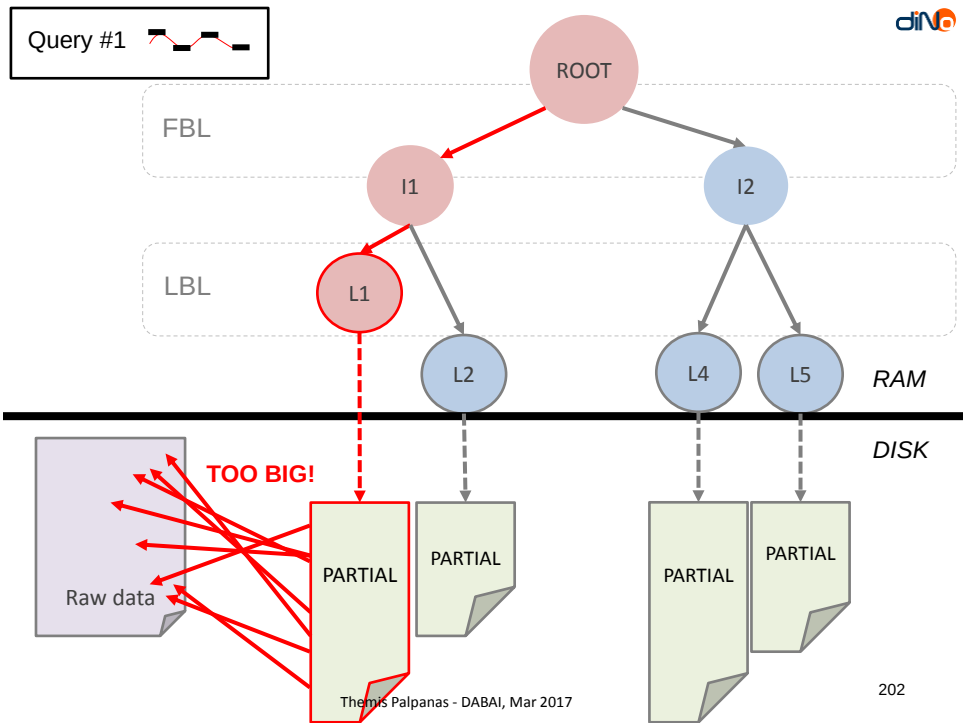
We flush the data to the disk to free up memory

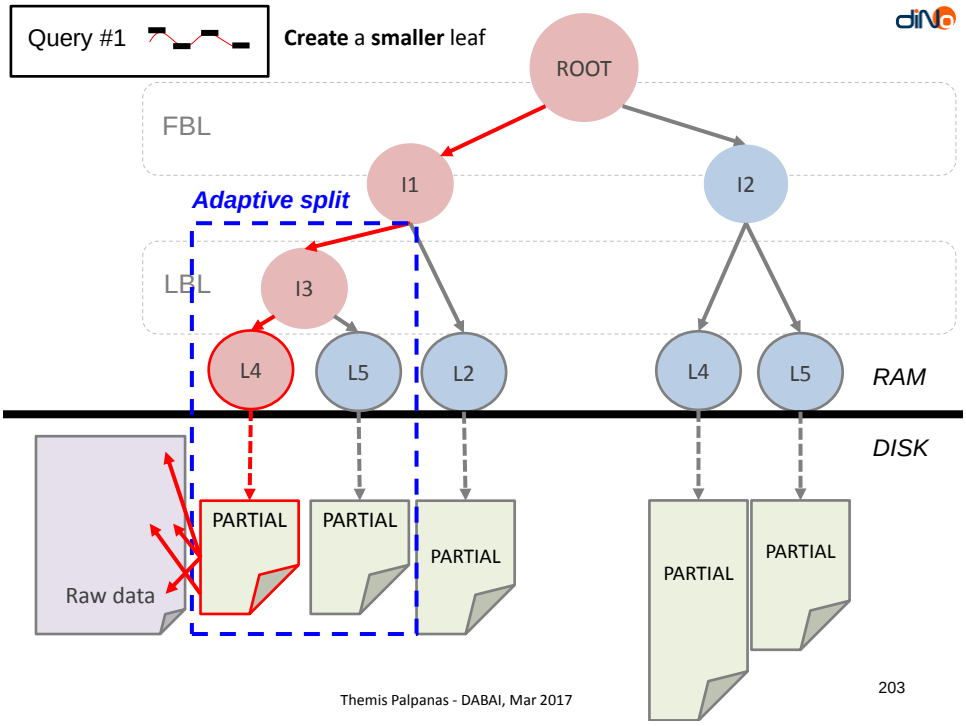


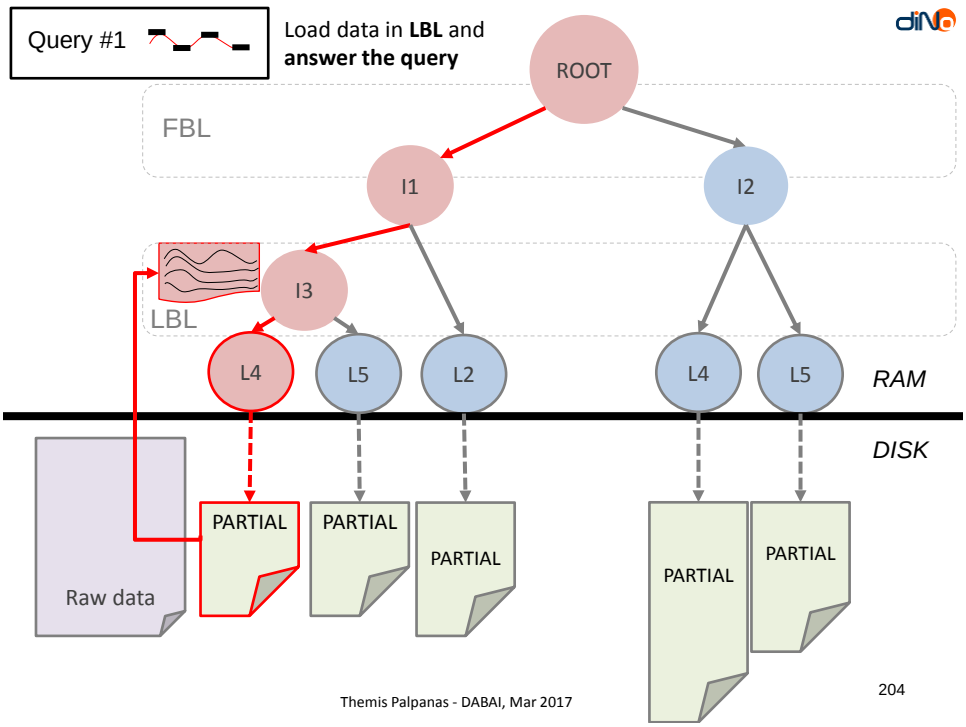

ROOT
FBL
I1
I2
LBL
L1
L2
L4
L5
RAM
DISK
Raw data
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Themis Palpanas - DABAI, Mar 2017
199


Query #1
ROOT
FBL
I1
I2
LBL
L1
L2
L4
L5
RAM
DISK
Raw data
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Themis Palpanas - DABAI, Mar 2017
200

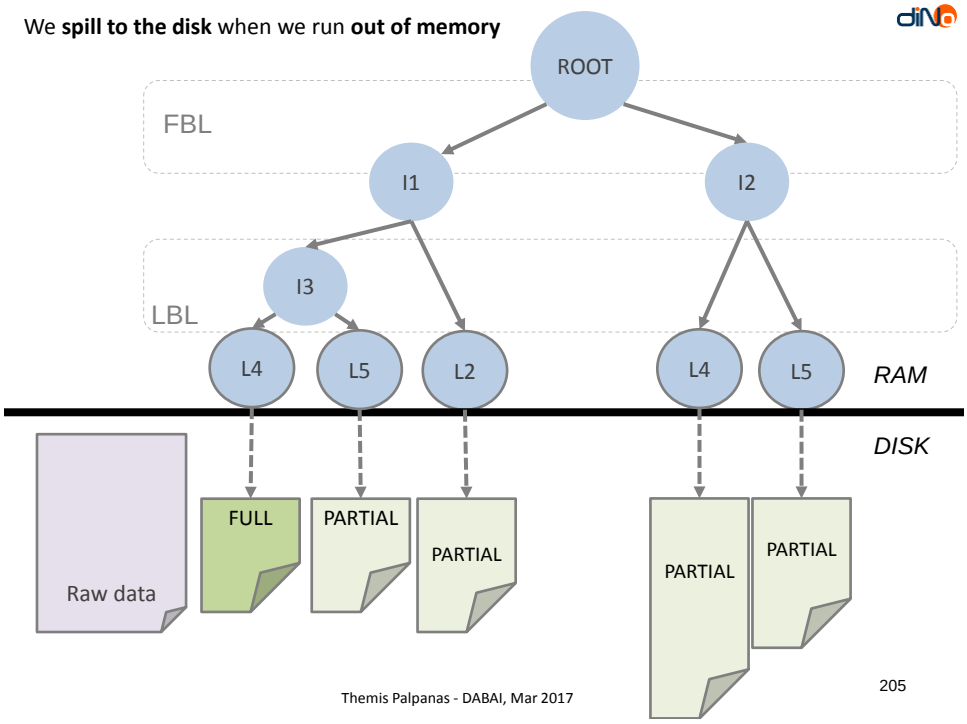

Query #1
ROOT
FBL
I1
I2
LBL
L1
L2
L4
L5
RAM
DISK
TOO BIG!
Raw data
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Themis Palpanas - DABAI, Mar 2017
201


Query #1
ROOT
FBL
I1
I2
LBL
L1
L2
L4
L5
RAM
DISK
TOO BIG!
Raw data
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Themis Palpanas - DABAI, Mar 2017
202


Query #1
Create a smaller leaf
ROOT
FBL
I1
I2
Adaptive split
LBL
I3
L4
L5
L2
L4
L5
RAM
DISK
PARTIAL
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Raw data
Themis Palpanas - DABAI, Mar 2017
203


Query #1
Load data in LBL and answer the query
ROOT
FBL
I1
I2
I3
LBL
L4
L5
L2
L4
L5
RAM
DISK
PARTIAL
PARTIAL
PARTIAL
PARTIAL
PARTIAL
Raw data
Themis Palpanas - DABAI, Mar 2017
204

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

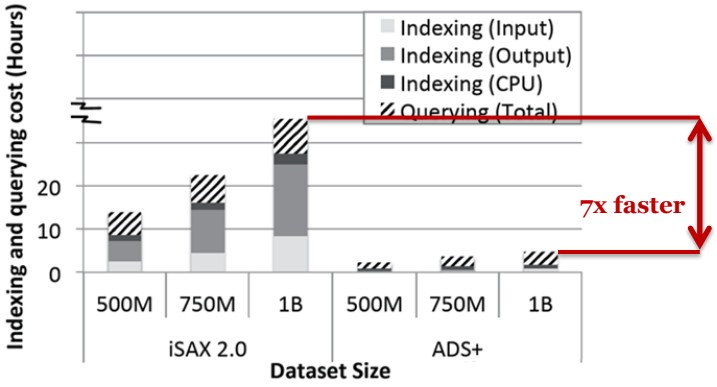


Themis Palpanas - DABAI, Mar 2017

205

206

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 - DABAI, Mar 2017

Comparison to *multi-dimensional indices*

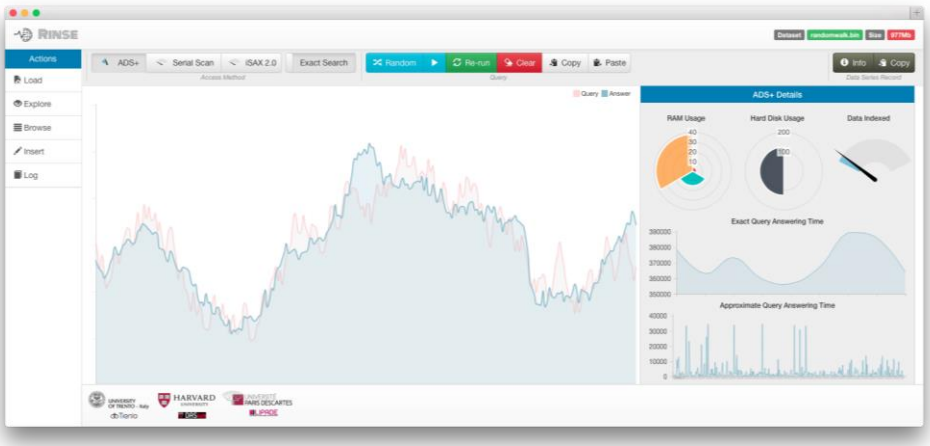
measure data-to-query time
(just index 1 **billion** data-series)

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

Demo

Publications
VLDB'15

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



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]*

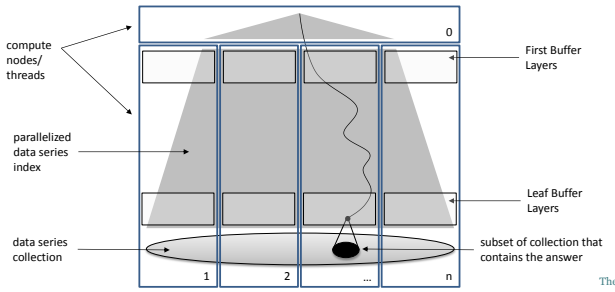
Distribution/Parallelization/Cloud?

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

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

What Next?

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

What Next?

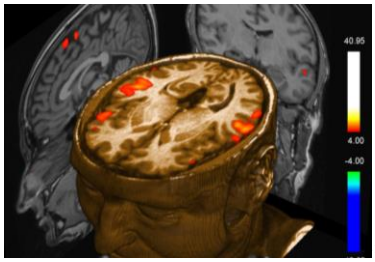
- functional Resonance Magnetic Imaging (fMRI) data
  - primary experimental tool of neuroscientists
  - reveal how different parts of brain respond to stimuli

What Next?

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

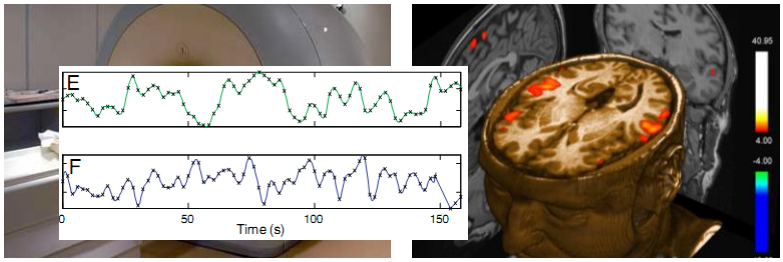
What Next?

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



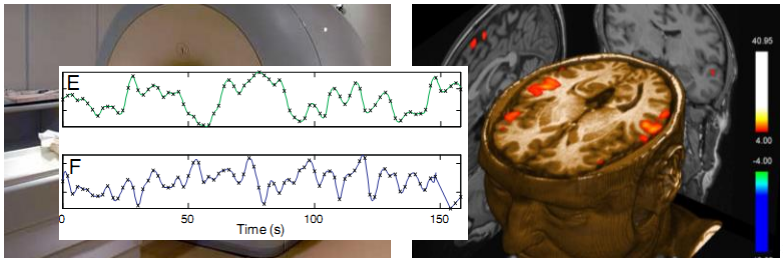
What Next?

- functional Resonance Magnetic Imaging (fMRI) data
  - primary experimental tool of neuroscientists
  - reveal how different parts of brain respond to stimuli
  - single experiment (1 subject, 1 test) produces
    - 60,000 data series of length 3,000: 12 GB



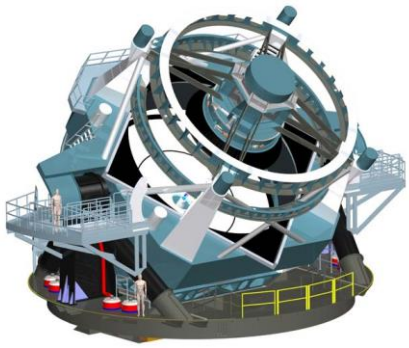
What Next?

- ADHD-200 Global Competition
  - classification task: detect Attention Deficit Hyperactivity Disorder
    - 776 subjects: 9 TB
      - equivalent to: 4.5 billion non-overlapping data series of size 256
      - equivalent to: 1100 billion overlapping data series of size 256



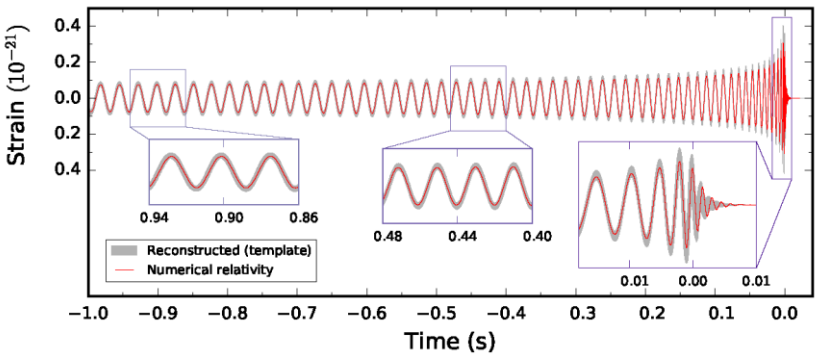
What Next?

- astronomy
  - LSST Telescope will be observing the entire sky every 3 nights: monitor 37B sky objects, generate 15TB/night



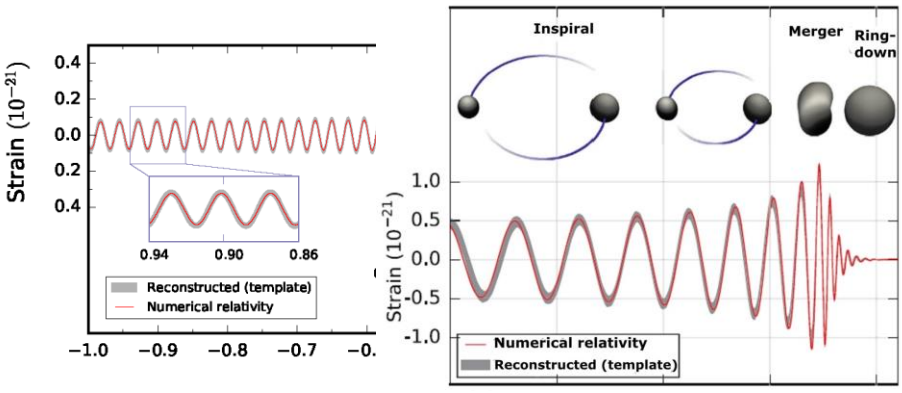
What Next?

- astronomy
  - gravitational wave detection



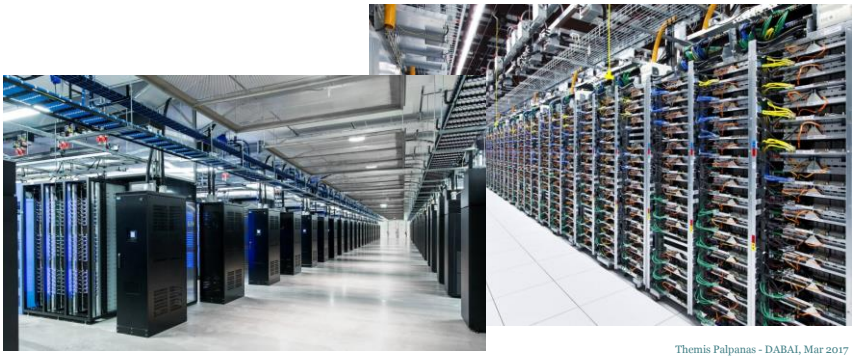
What Next?

- astronomy
  - gravitational wave detection



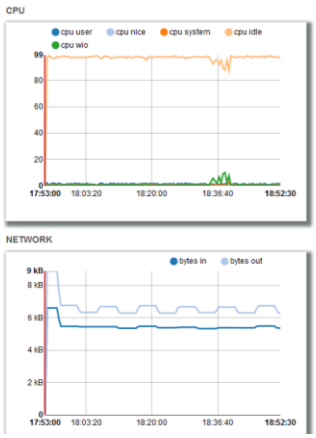
What Next?

- infrastructure monitoring



What Next?

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

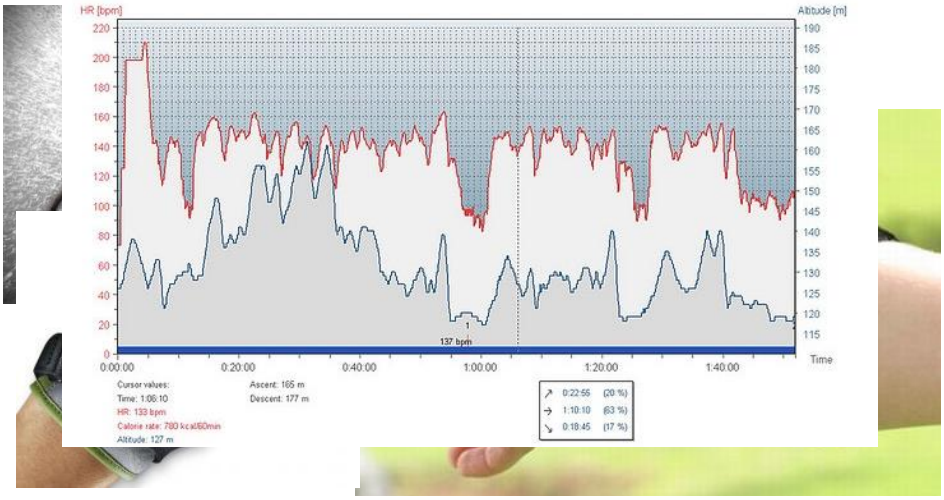


DABAI, Mar 2017

What Next?

Themis Palpanas - DABAI, Mar 2017

What Next?

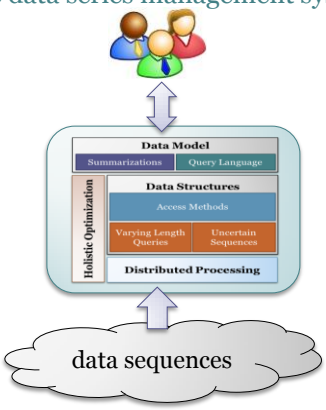


The Road Ahead

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

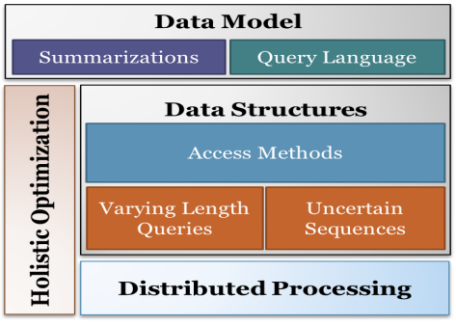
The Road Ahead

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



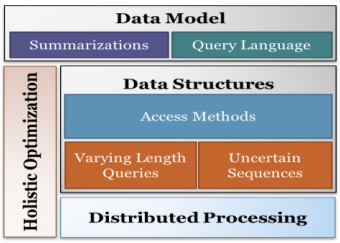
The Road Ahead

- Big Sequence Management System



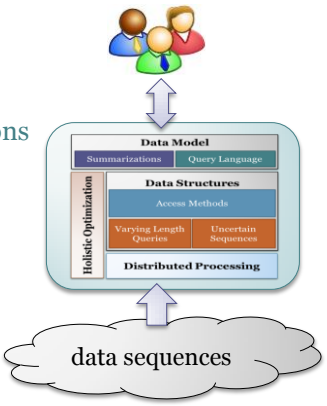
The Road Ahead

- Big Sequence Management System
  - physical and logical data independence
  - rich, declarative query language
  - query optimization
  - inherent scalability
  - support for streaming data series
  - support for uncertain data series



The Road Ahead
Crossroads with DABAI

- Big Sequence Management System
  - boost exploitation of big data (series)
  - enable development of 3rd party applications
    - for government
      - health, smart cities, …
    - for industry
      - wind farms, manufacturing, …
  - empower non-expert users
  - diffuse benefits to society at large



collaborations!

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*

Data-Intensive and Knowledge-Oriented systems

thank you!

google: Themis Palpanas