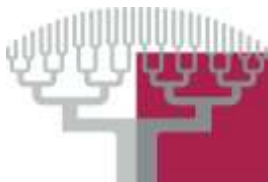


# Enabling Exploratory Analysis of Very Large Scientific Data



Themis Palpanas  
*Paris Descartes University*



## Short Bio

- current positions:
  - professor of computer science, Paris Descartes University, France
  - co-director of diNo group (14 members), France
- previous positions:
  - vice-chair of Data and Knowledge Engineering group, Trento RISE (90 members), Italy
  - director of dbTrento group (20 members), Italy
  - general chair for the international conference on Very Large Data Bases (2013), Italy
  - professor of computer science, University of Trento, Italy
  - IBM T.J. Watson Research Center, USA
  - University of California, Riverside, USA
- Visiting Researcher in:
  - Microsoft Research, USA
  - IBM Almaden Research Center, USA
  - National University of Singapore
- founding member of Event Processing Technical Society
- past member of IBM Academy for Event Processing
- recipient of the international **IBM SUR Award**, 2012 (for excellence in research)
- holds **3 best-paper awards**
- author of **9 US patents** (3 of them implemented in commercial products)
- PhD in Computer Science from University of Toronto, Canada

# Data Intensive and Knowledge Oriented Systems

## diNo Group

- faculty members
  - Salima Benbernou
  - Ioana Ileana
  - Mourad Ouziri
  - Themis Palpanas
  - Soror Sahri
- Postdocs
  - Mehdi Bentounsi
  - Christophe Rodriguez
- PhD students
  - Iheb BenAmor
  - Adel Benaissa
  - Isam Ellayddi
  - Xin Huang
  - Michele Linardi
  - Pavlos Paraskevopoulos
  - Kostas Zoumpatianos



# diNo Mission

## Data-driven Knowledge Discovery

**research areas**

- integration
- quality
- management
- analytics
- mining
- privacy
- security

**techniques**

- entity resolution
- adaptive indexing
- streaming
- algorithms
- distributed
- processing
- semantic
- engineering
- crowdsourcing

**applications**

- finance
- smart cities
- social sciences
- healthcare
- scientific data
- manufacturing
- service providers

## in Big Data

# Big Data

- large scale data
- streaming data
- heterogeneous data
- private data
- uncertain data

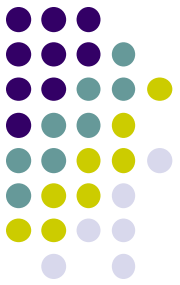
*confluence of all the above!*

# Our Work

- large scale data
- streaming data
- heterogeneous data
- private data
- uncertain data

## **funded by:**

European Commission, Facebook, IBM Research, Hewlett-Packard Labs, Telecom Italia, Autonomous Province of Trento



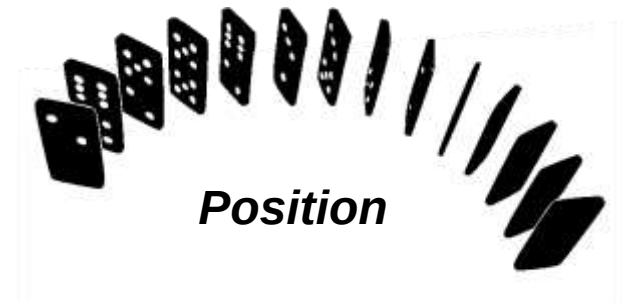
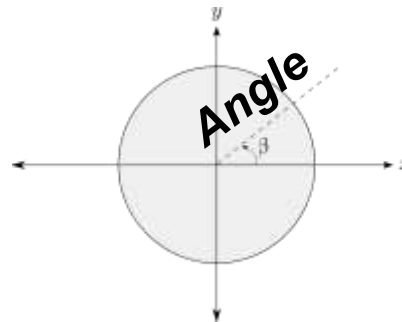
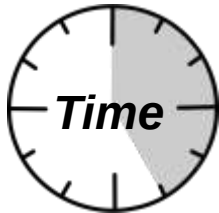
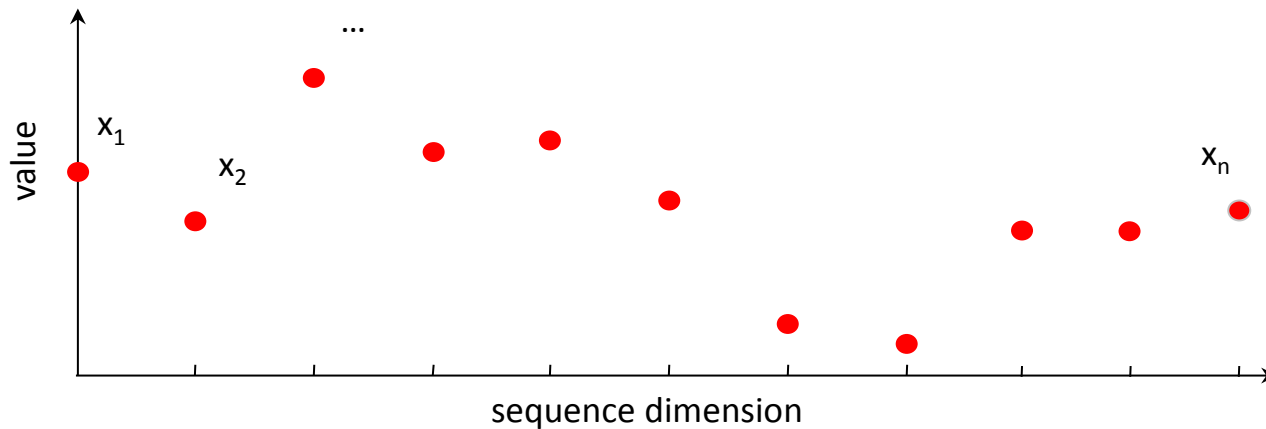
## *scalable indexing and mining of massive data series*

# References

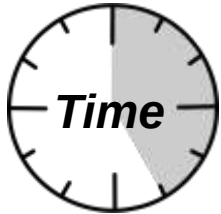
- papers
  - Themis Palpanas. **Data Series Management: The Road to Big Sequence Analytics**. ACM SIGMOD Record 44(2), 2015, Special Issue on Visionary Ideas in Data Management
    - <http://www.mi.parisdescartes.fr/~themisp/publications/sigrec15-bisemvision.pdf>
  - Kostas Zoumpatianos, Yin Lou, Themis Palpanas, Johannes Gehrke. **Query Workloads for Data-Series Indexes**. KDD 2015
    - <http://www.mi.parisdescartes.fr/~themisp/publications/kdd15-bends.pdf>
  - Kostas Zoumpatianos, Stratos Idreos, Themis Palpanas. **RINSE: Interactive Data Series Exploration**. VLDB 2015
    - <http://www.mi.parisdescartes.fr/~themisp/publications/vldb15-rinse.pdf>
  - K. Zoumpatianos, S. Idreos, T. Palpanas. **Indexing for Interactive Exploration of Big Data Series**. SIGMOD 2014
    - <http://www.mi.parisdescartes.fr/~themisp/publications/sigmod14-ads.pdf>
  - A. Camerra, J. Shieh, T. Palpanas, T. Rakthanmanon, E. Keogh. **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>
  - A. Camerra, T. Palpanas, J. Shieh, E. Keogh. **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/DStat>
- demo
  - <http://daslab.seas.harvard.edu/rinse/>

# Data Series

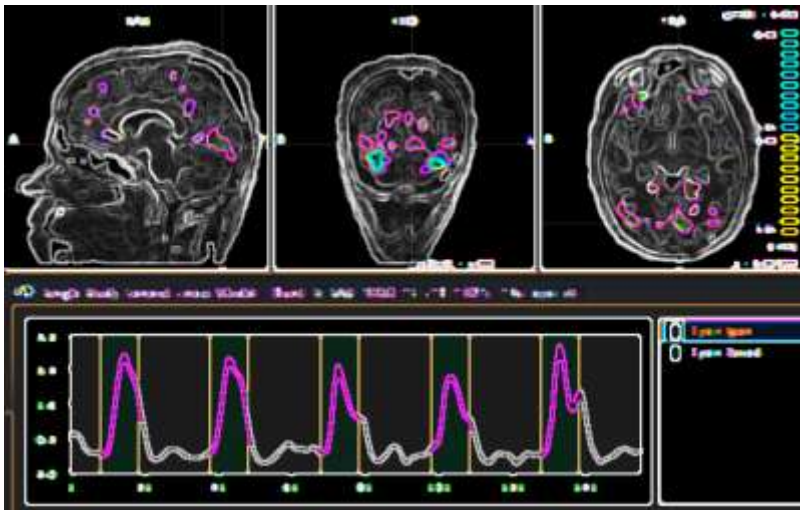
Sequences of points **ordered** over a **dimension**



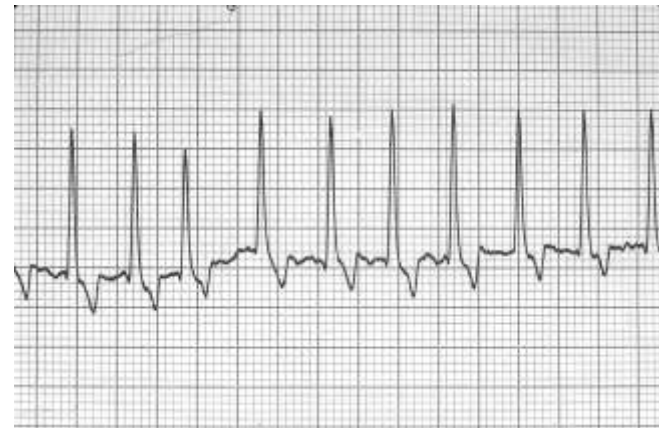
# Examples



**Financial Data**

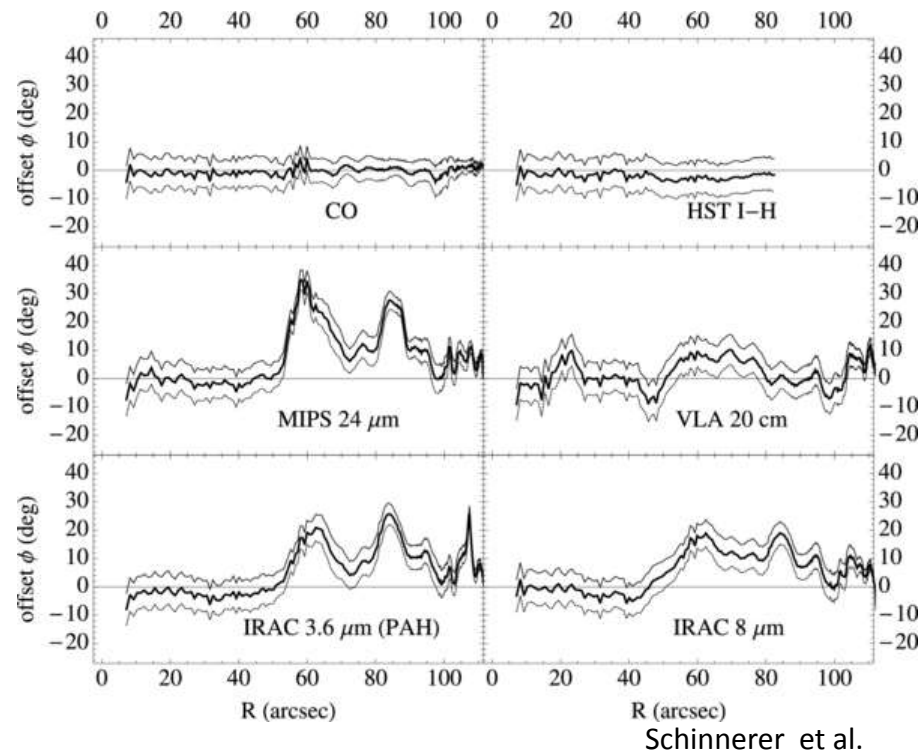
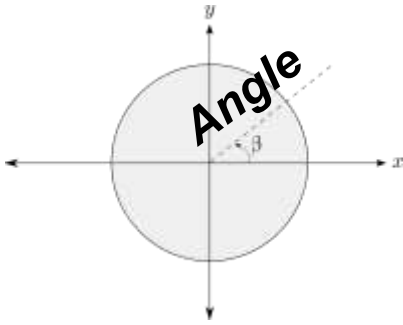


**Neuroscience Data (fMRI)**

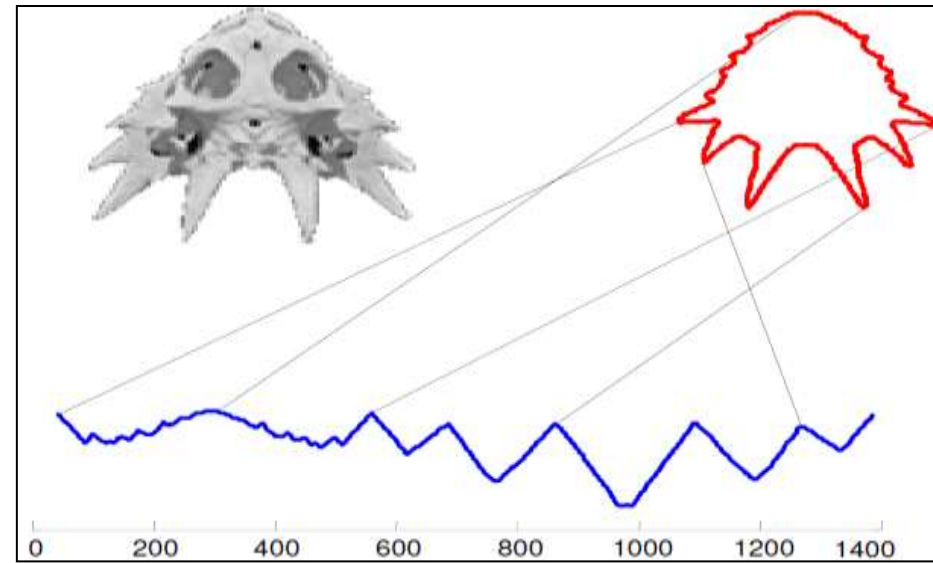


**Health Data (ECG)**

# Examples

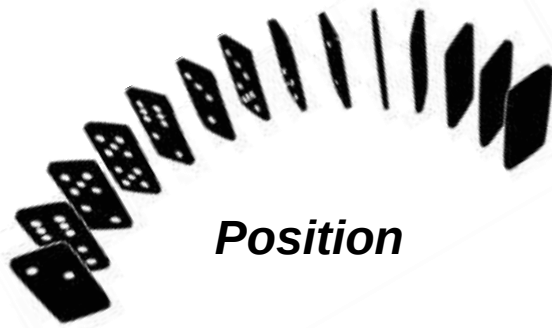


## Astronomical data



*By Prof. E. Keogh*

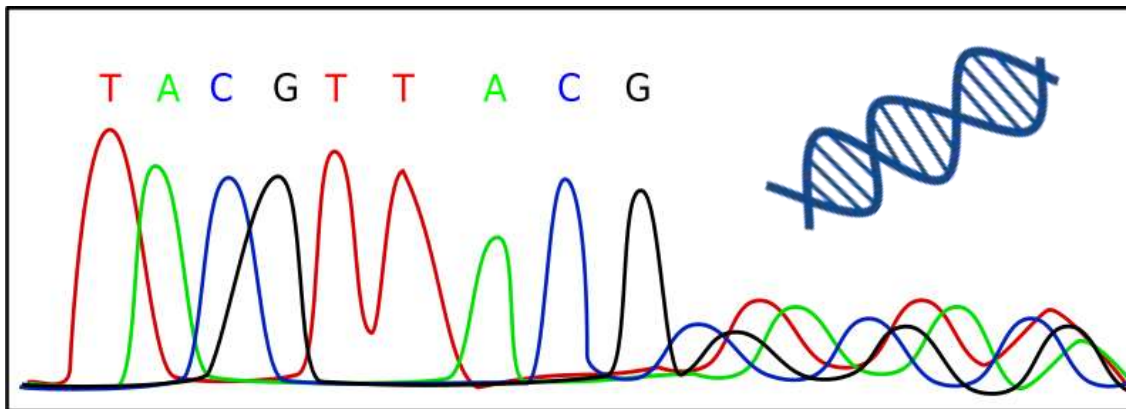
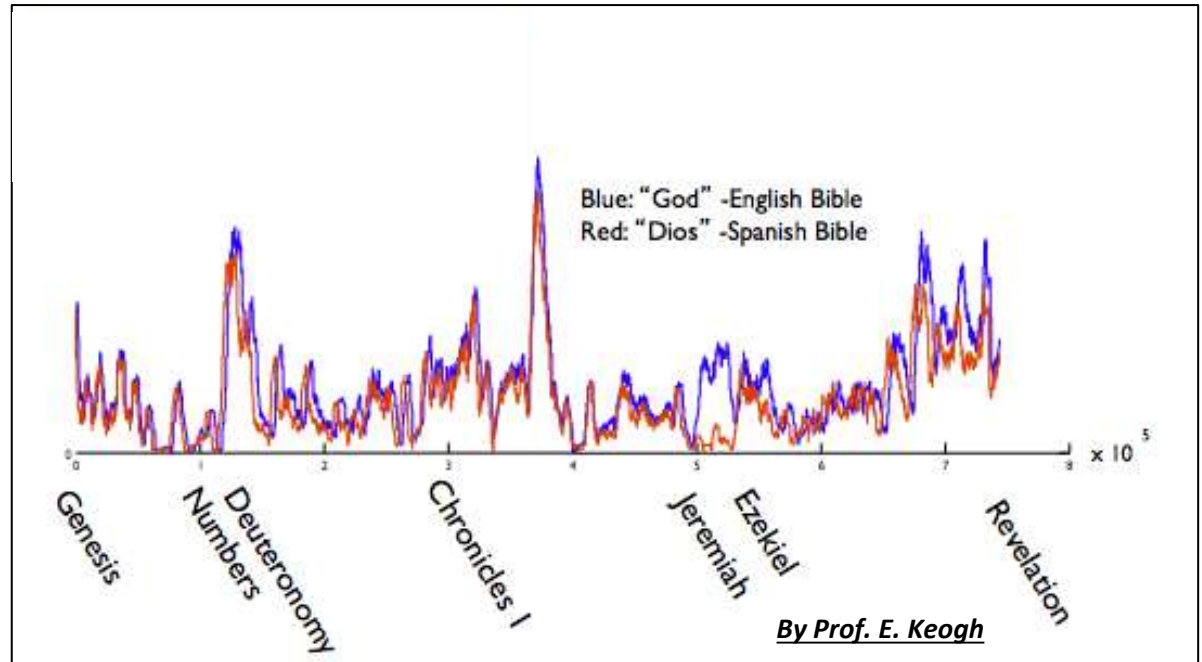
## Shapes



*Position*

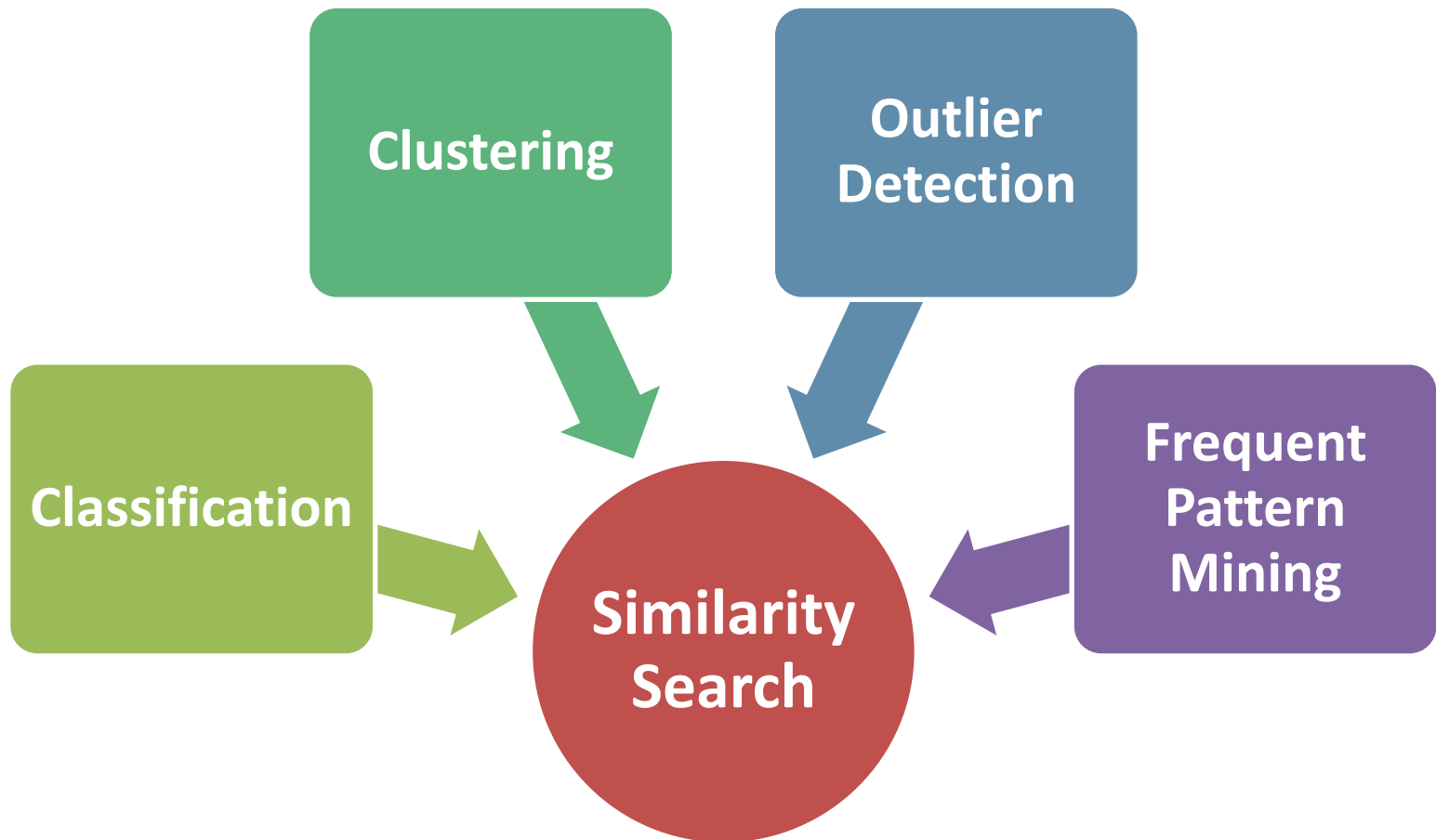
# Examples

**Text Data**

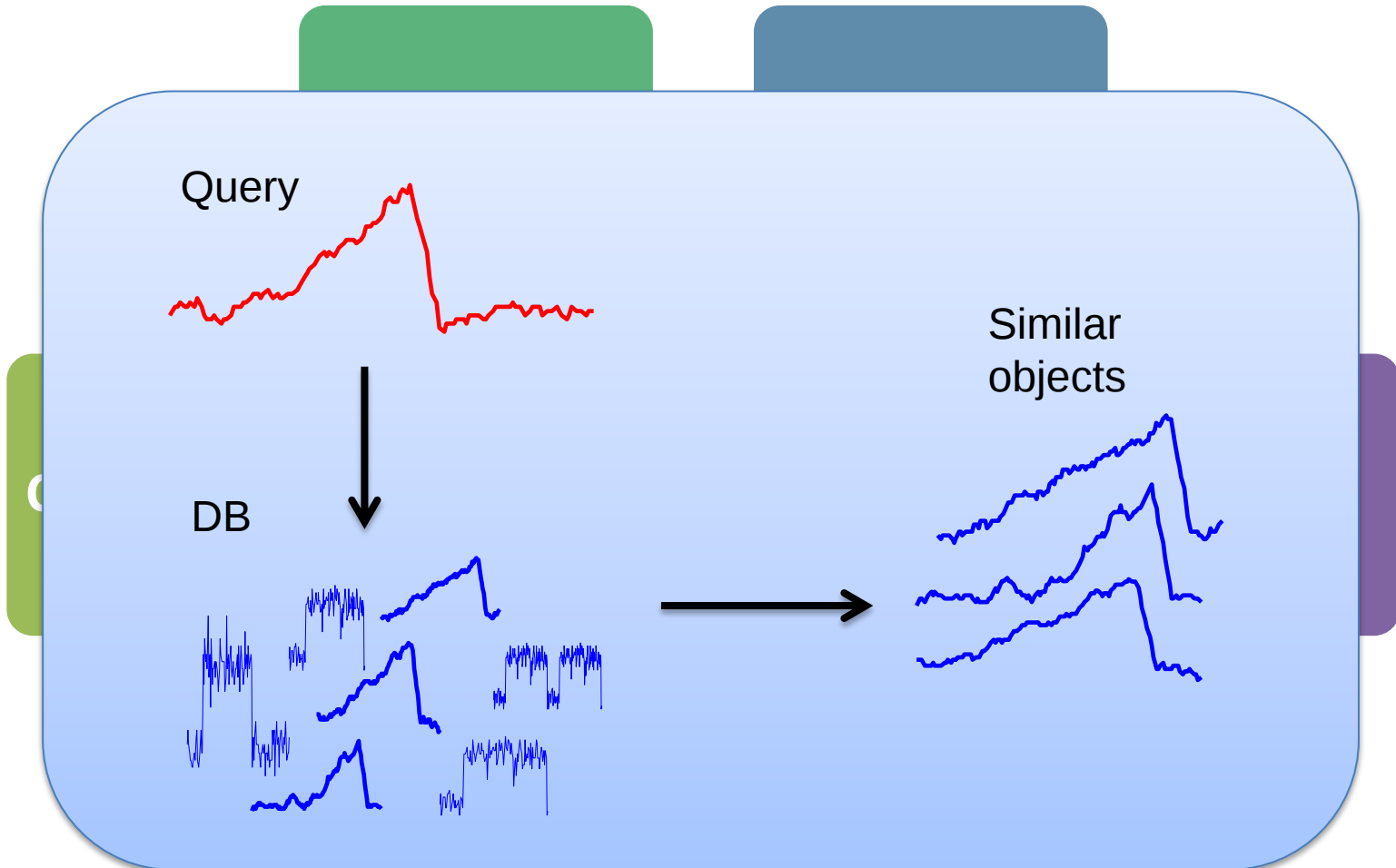


**Genome Data**

# Data Series Analytics



# Data Series Analytics



# Data Series Analytics

Query

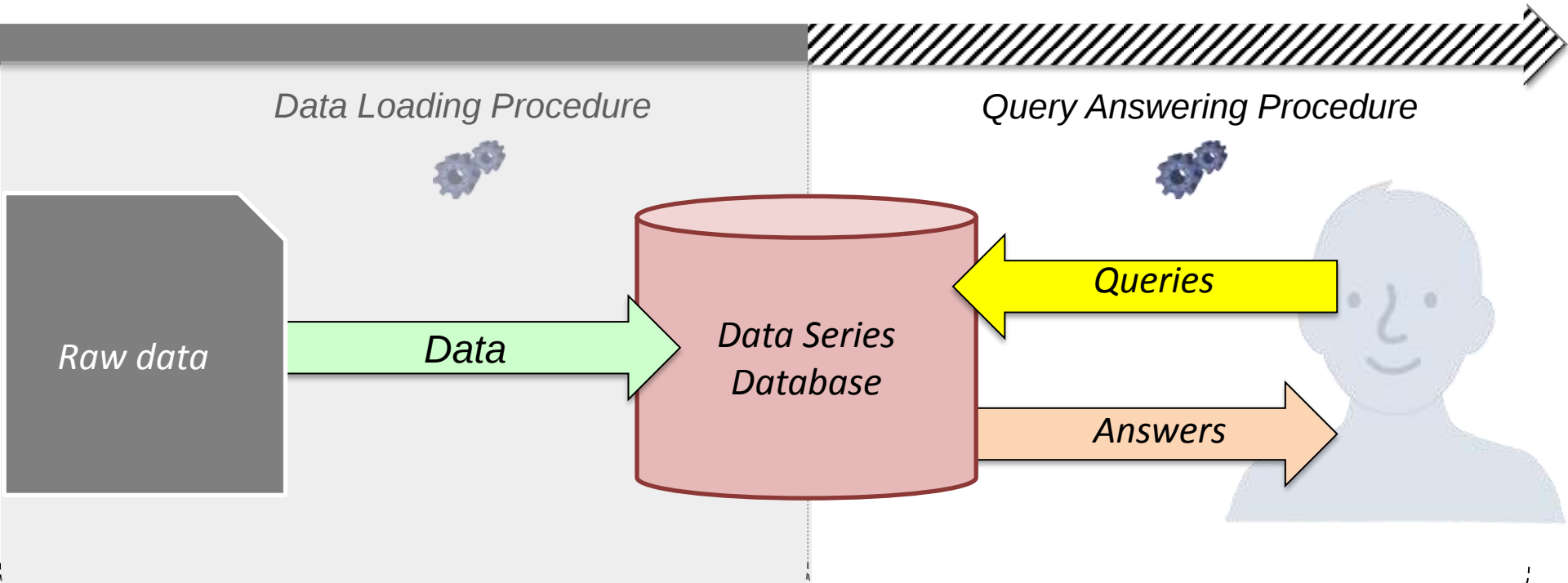


Similar  
objects



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

# Query answering process



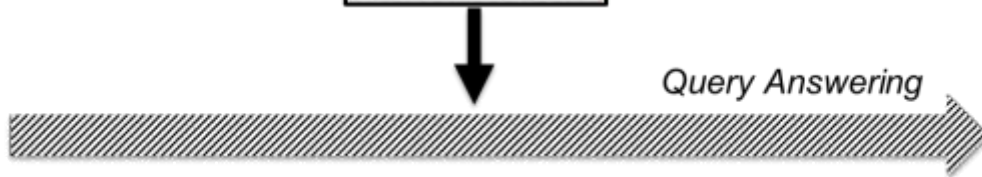
data-to-query time

query answering time

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

# Traditional Approaches

answer nearest neighbor queries on a 1TB dataset  
serial scan takes **45 minutes** per query!



**Bad news:** building the index takes **too long**

**27 hours** to index 1 Billion data series

optimized implementation in C, optimal parameters,  
modern hardware (Intel Xeon, 64GB RAM, RAID0 disks)



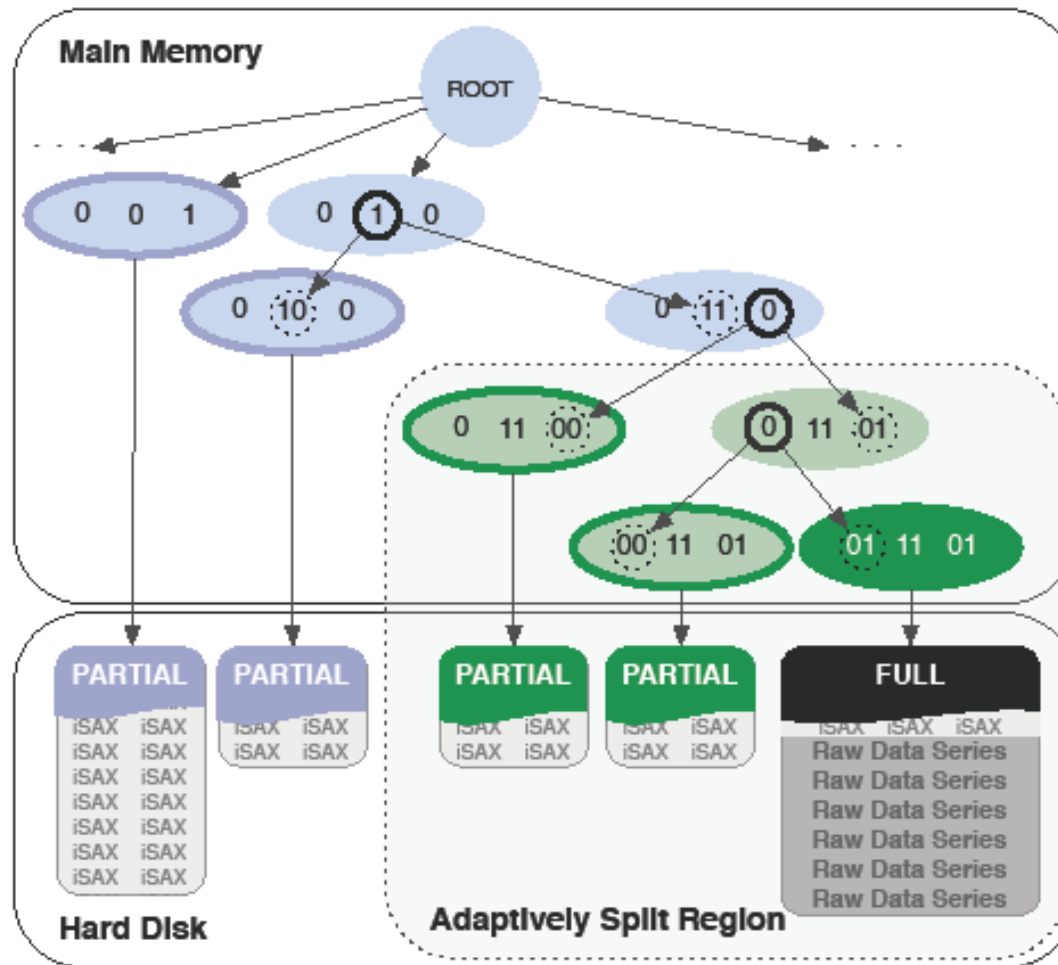
*Data Loading*

*Query Answering*



**complex analytics in hours/days...**

# Data Series Indexes: iSAX2+, ADS+

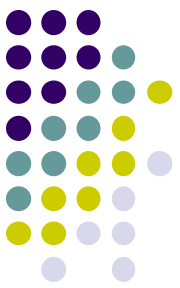


**complex analytics in minutes/seconds!**

# Conclusions

- work on techniques and algorithms for
  - online processing and complex analysis of streaming data
  - efficient identification of interesting patterns hidden in huge volumes of data

*open to collaborations!*



# Current and Past Collaborations

- Infrastructure Monitoring
  - analysis and mining of hardware and software infrastructure for **health monitoring**
  - with **Facebook**
- 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 **CIMeC**
- Green Manufacturing
  - analysis and optimization of manufacturing processes for **energy savings**
  - with **SAP, Intel, Volvo**
- 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 **Hewlett-Packard Labs**

thank you!

google: themis palpanas

