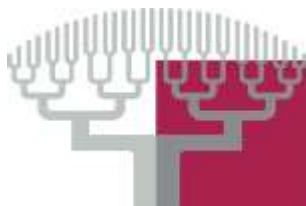


Data Series Management



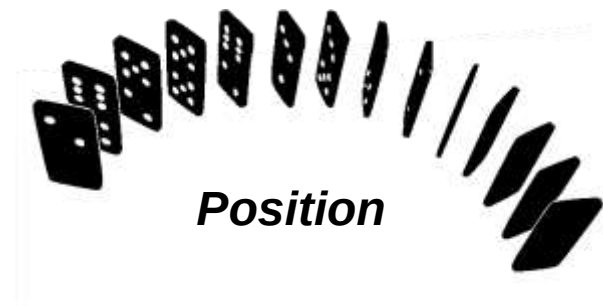
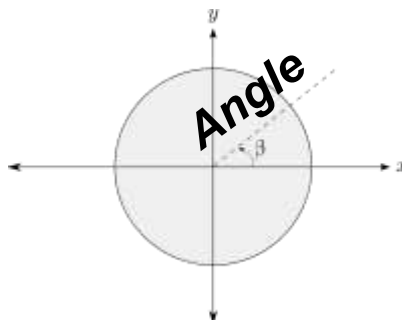
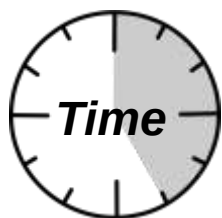
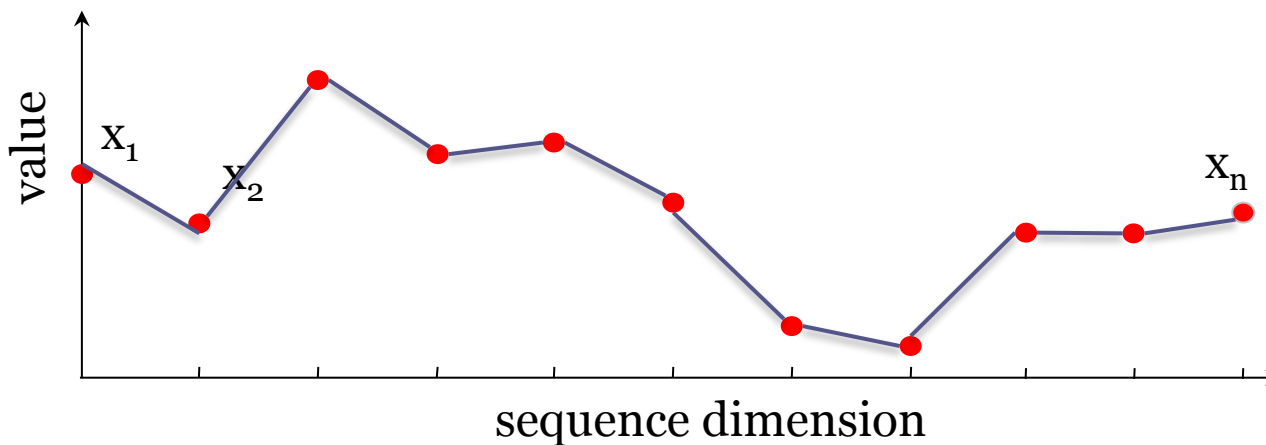
Themis Palpanas
University of Paris V - Paris Descartes

Interdisciplinary Workshop on Time Series Analysis
(CNRS/MASTODONS)
Paris, June 2016



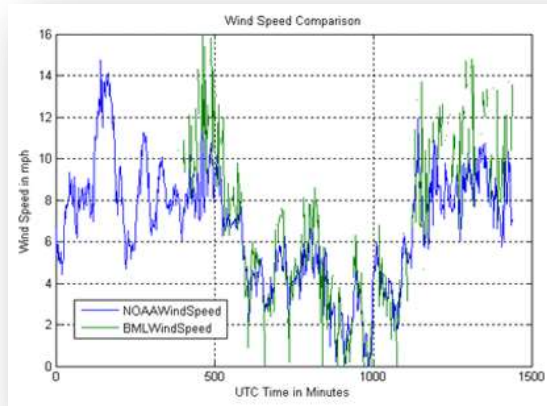
Data series

- Sequence of points ordered along some dimension



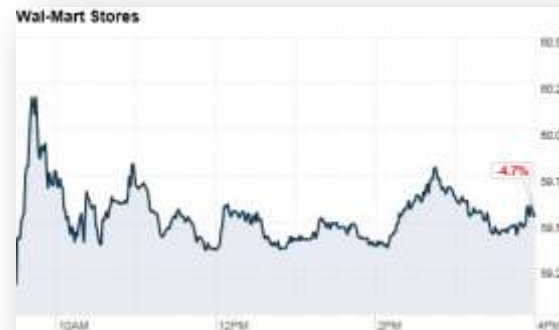
Scientific Monitoring

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



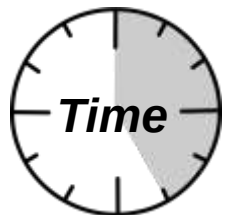
Wind speed

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



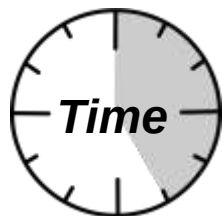
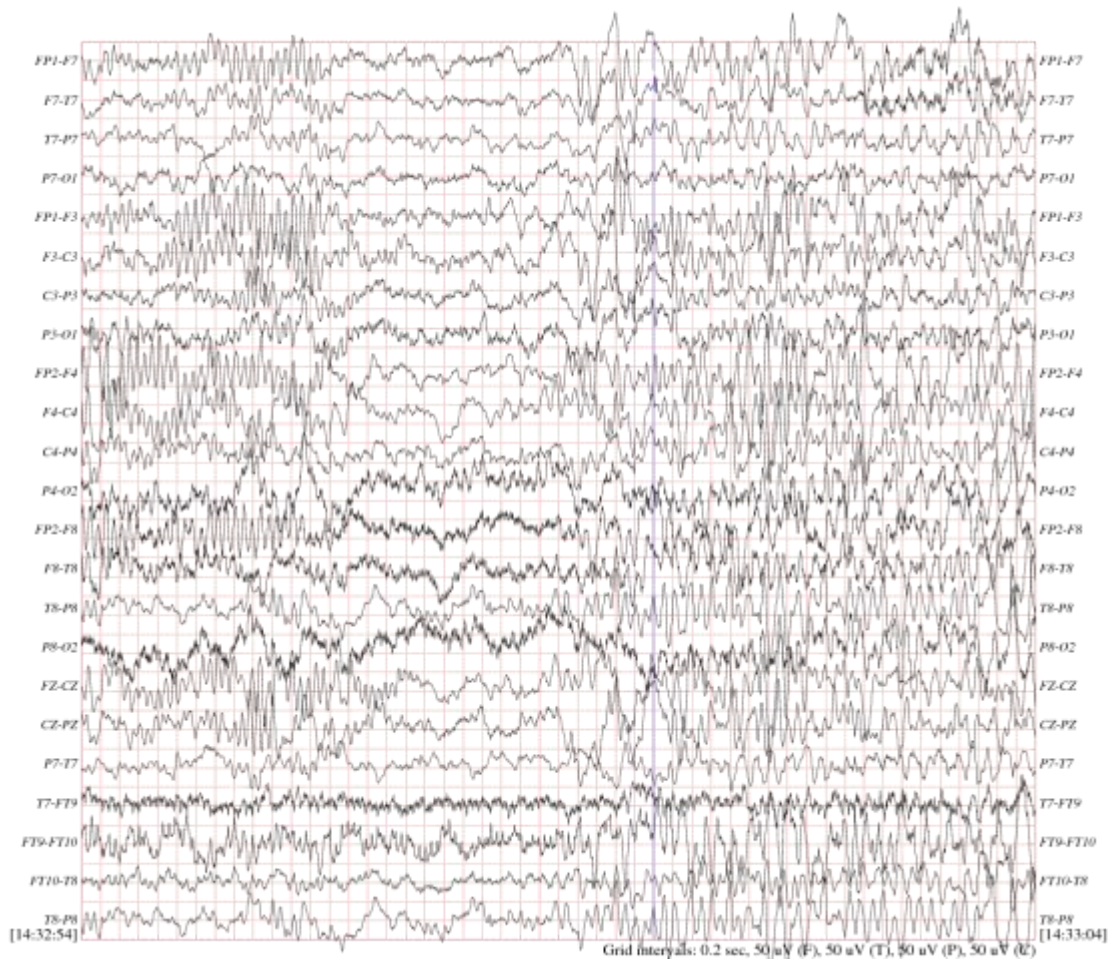
Historical stock quotes

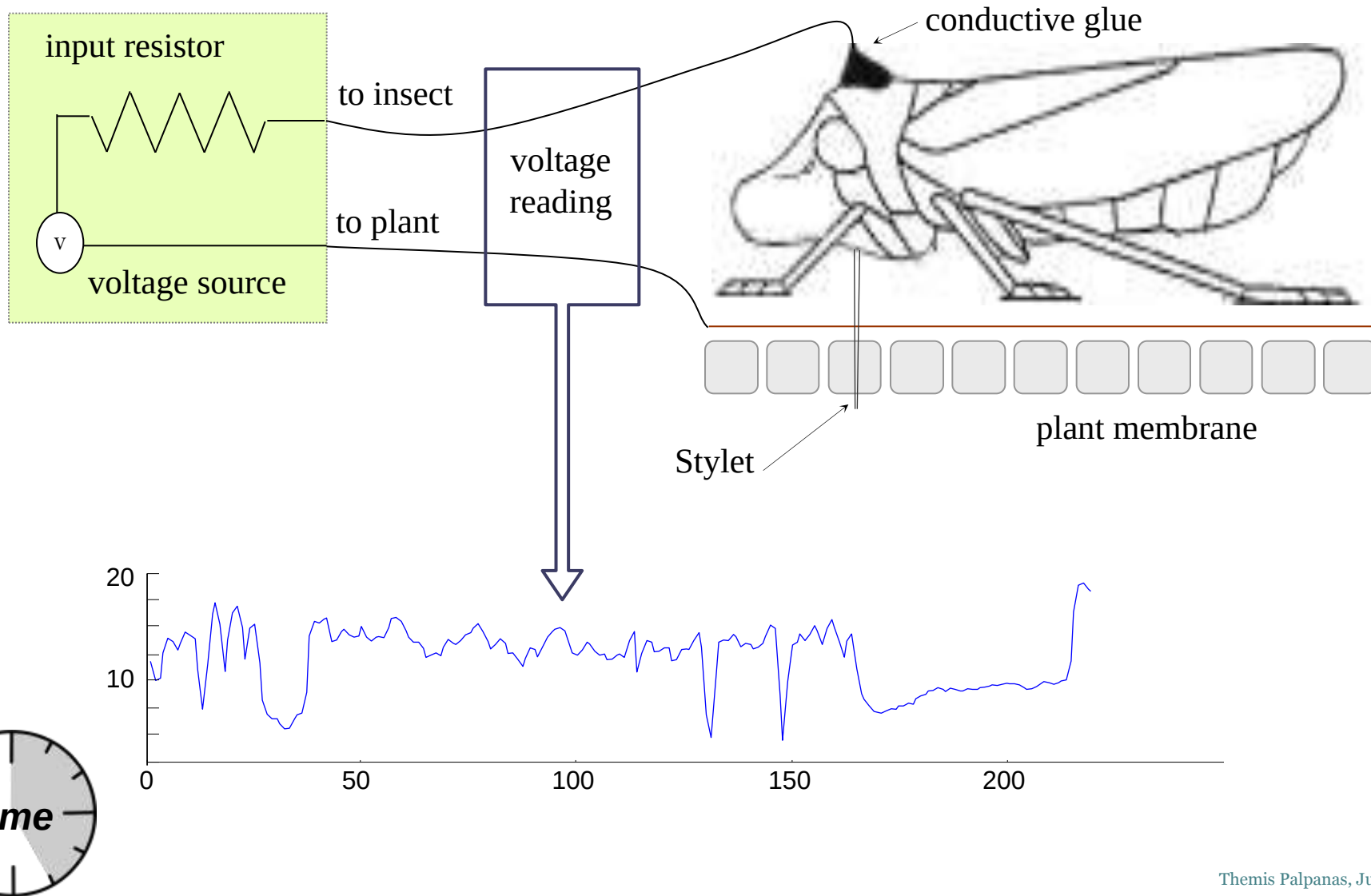
http://money.cnn.com/2012/04/23/markets/walmart_stock/index.htm

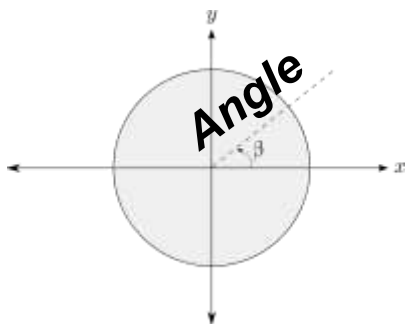
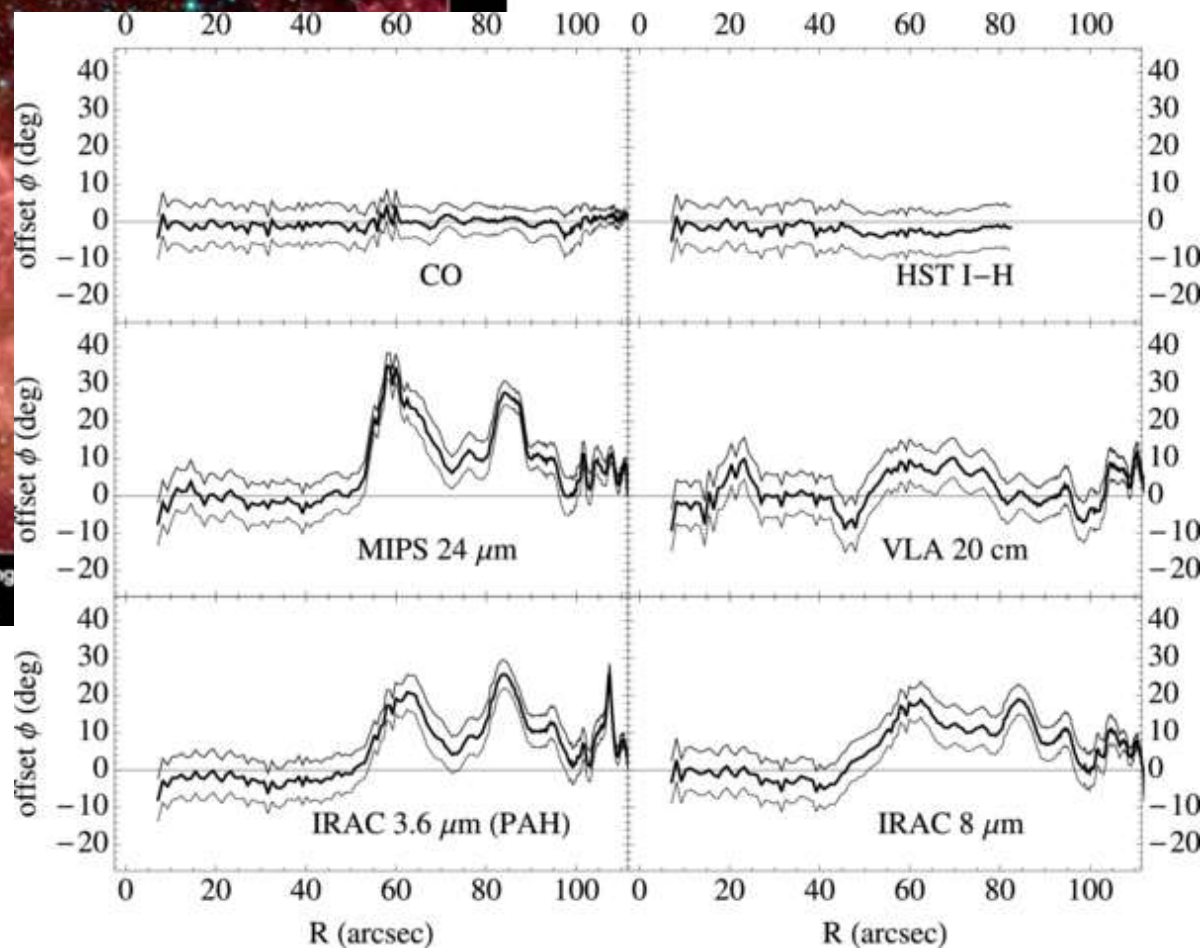
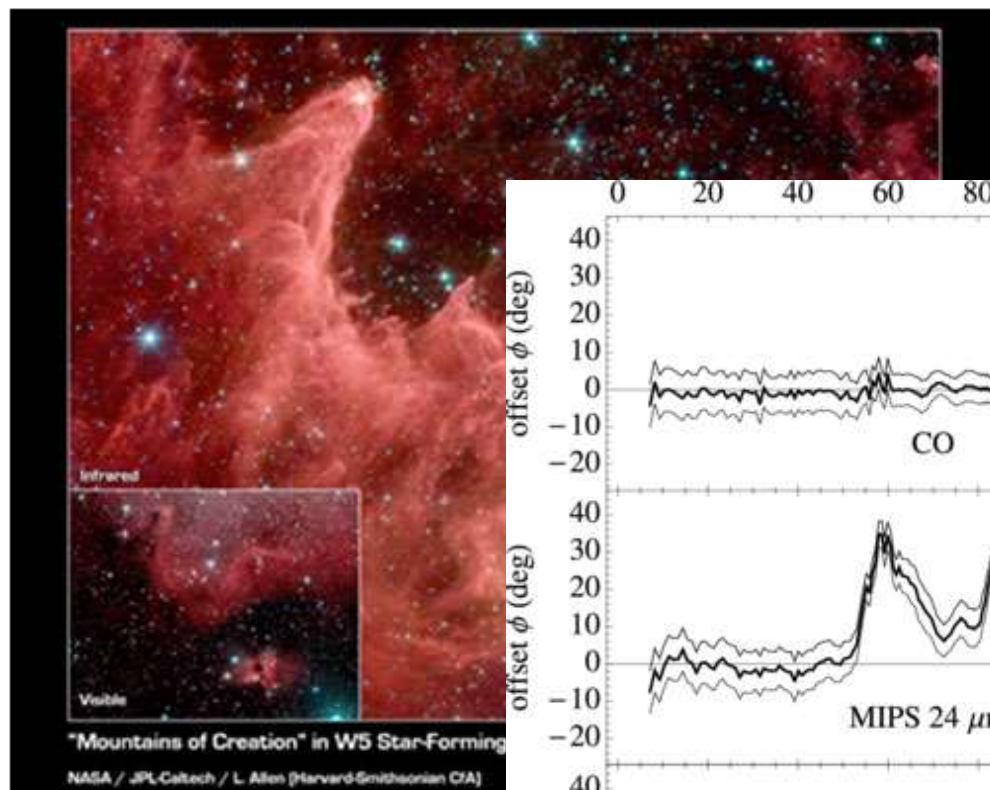




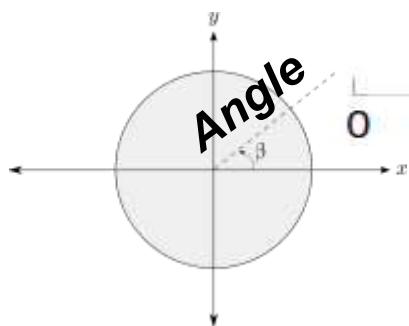
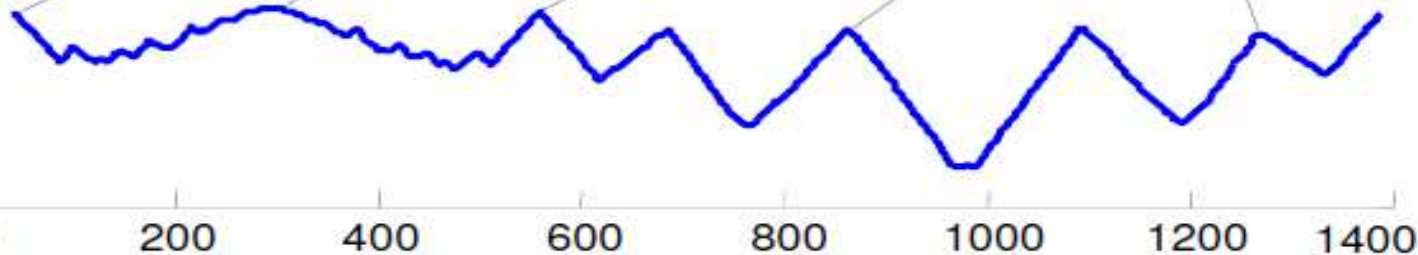
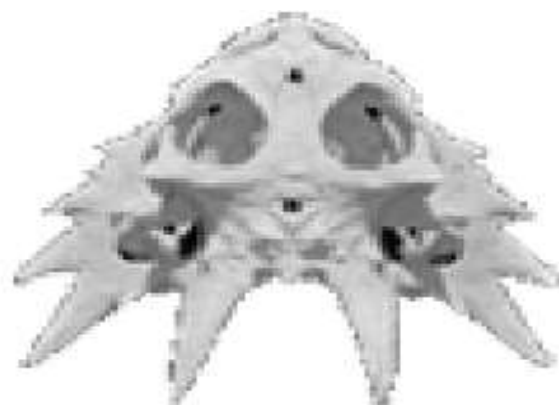
Neuroscience Data (EEG)

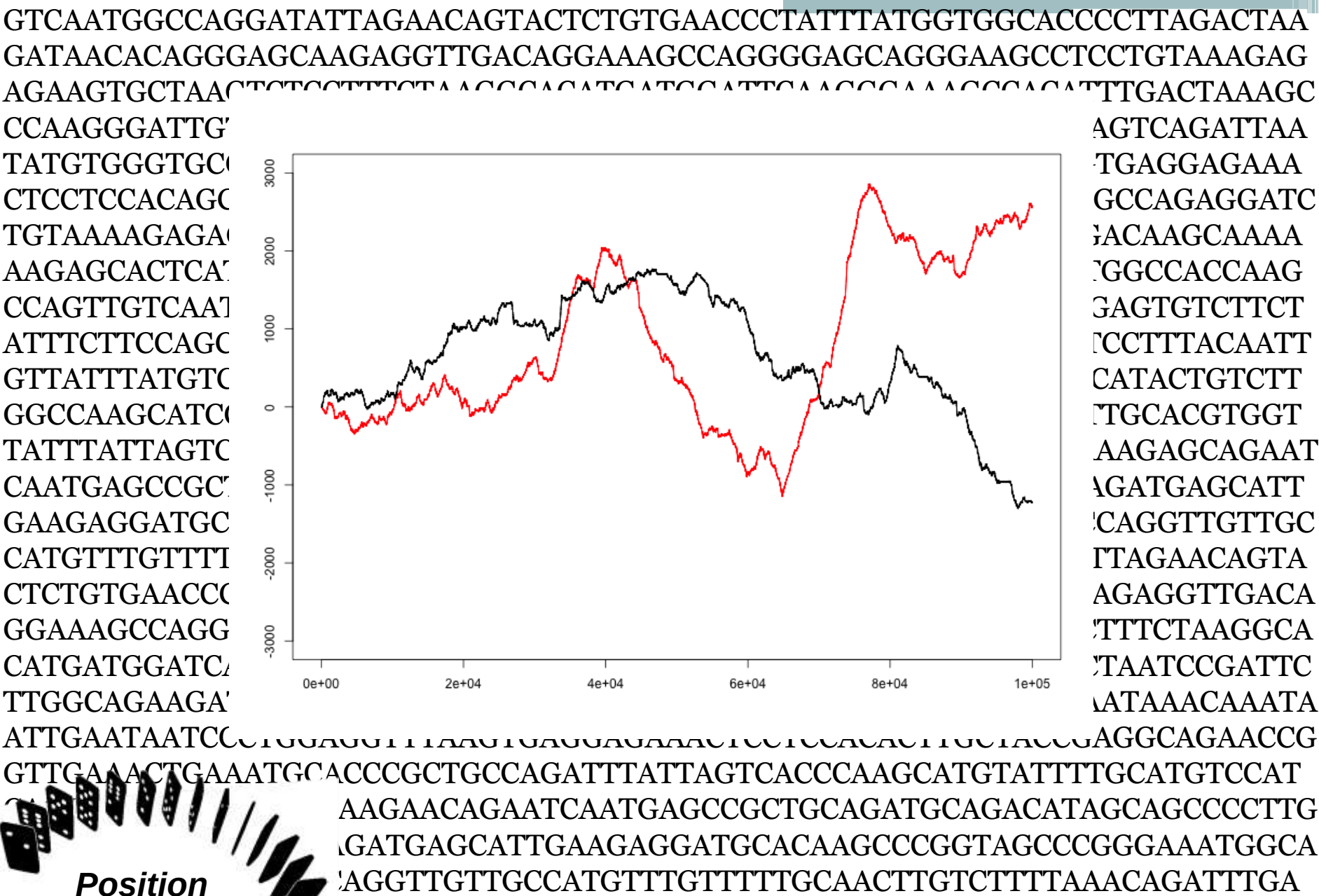




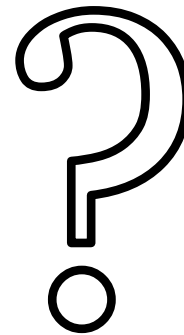
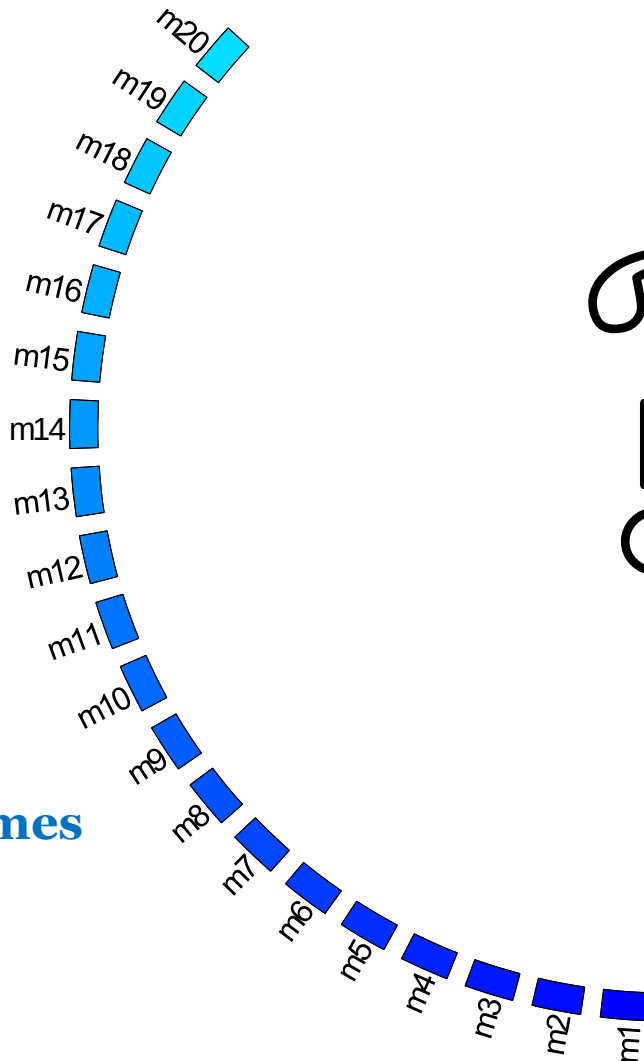


Schinnerer et al.

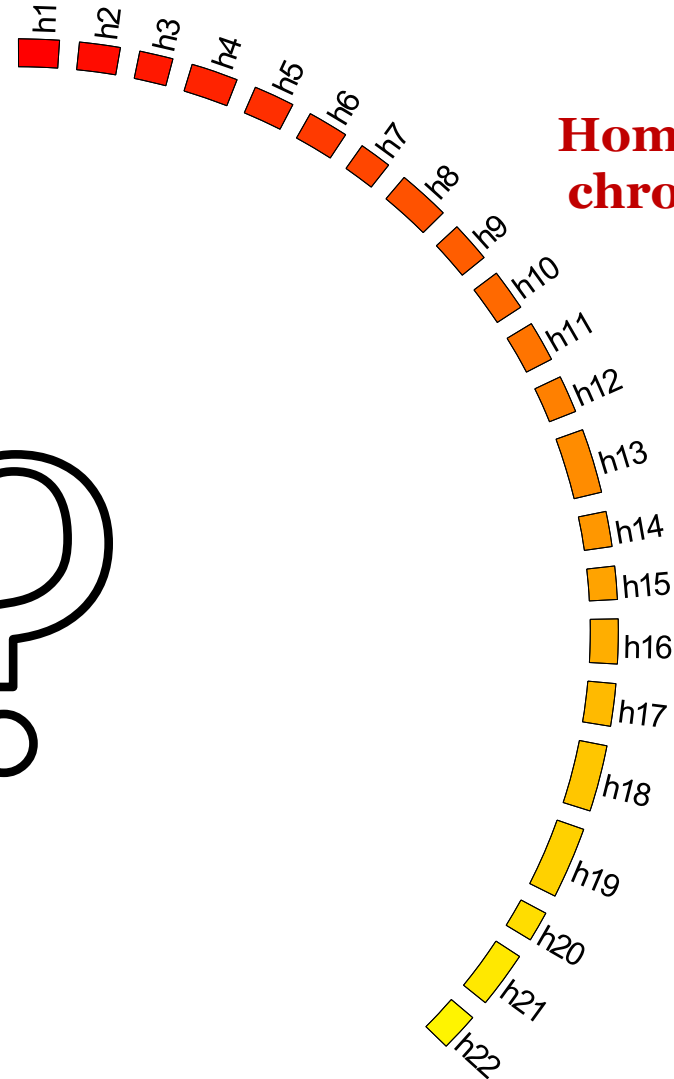




**Rhesus
Macaque
chromosomes**

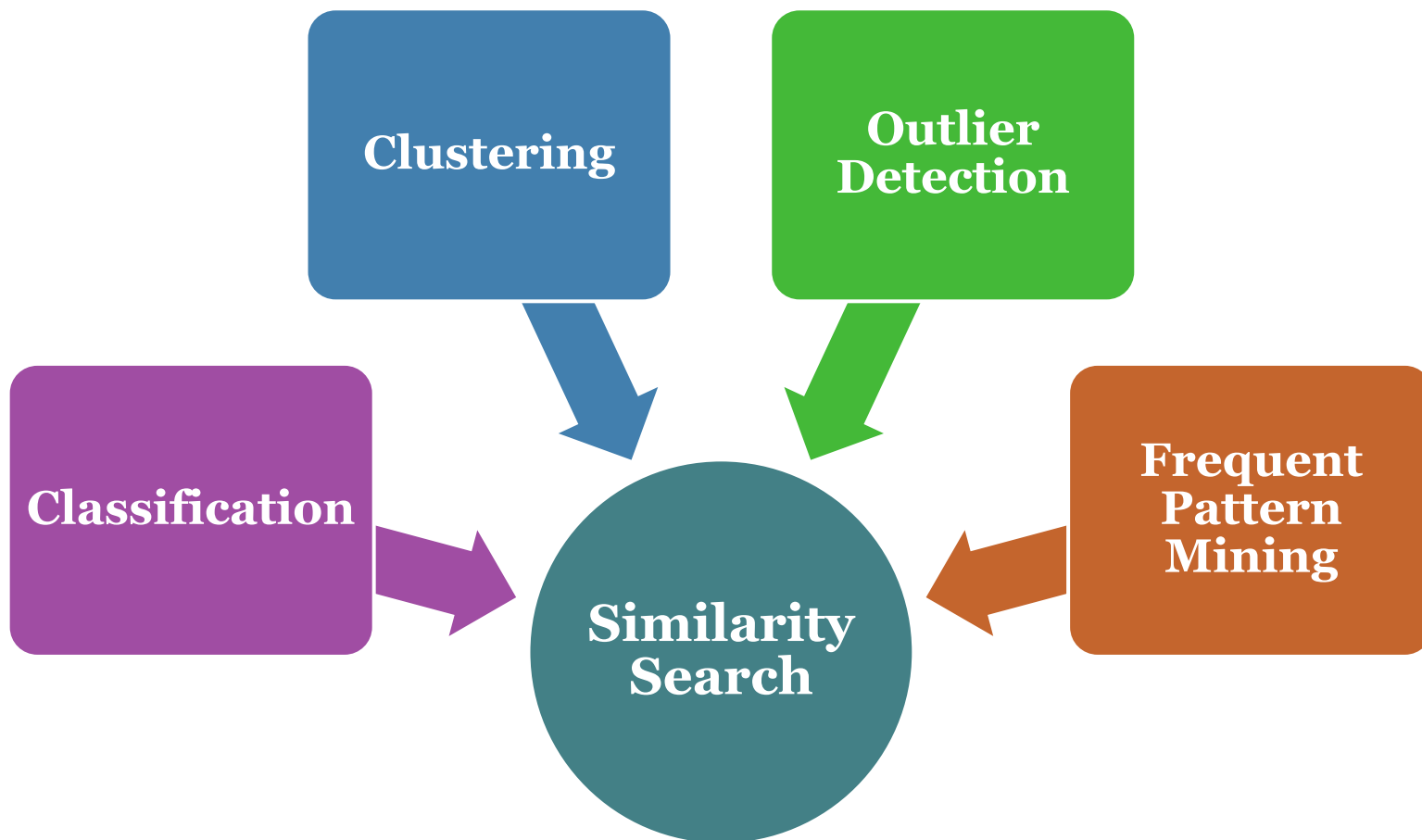


**Homo Sapiens
chromosomes**



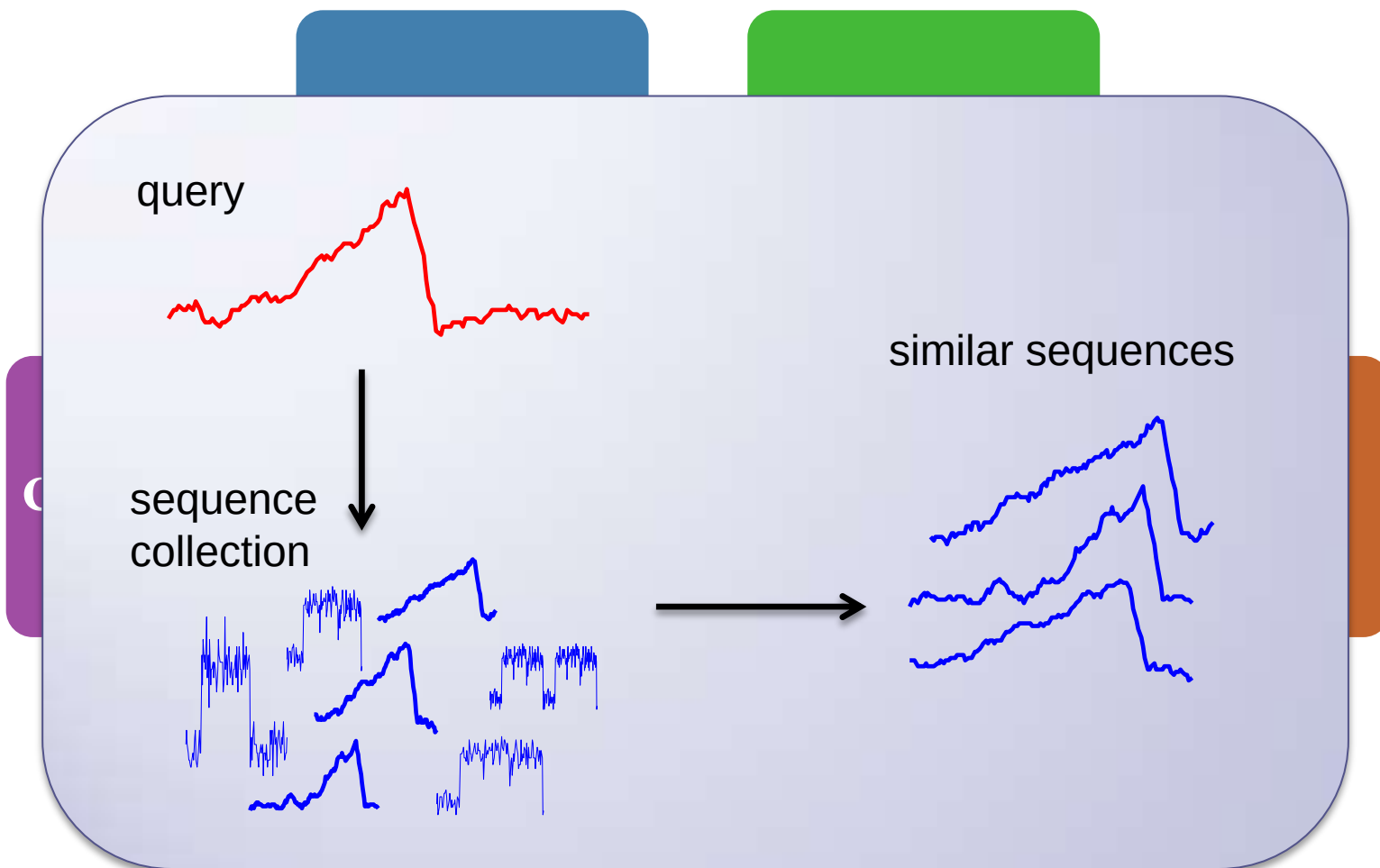
What do we want to do with them?

- complex analytics



What do we want to do with them?

- complex analytics



What do we want to do with them?

- complex analytics

Euclidean

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

Dynamic Time
Warping (DTW)

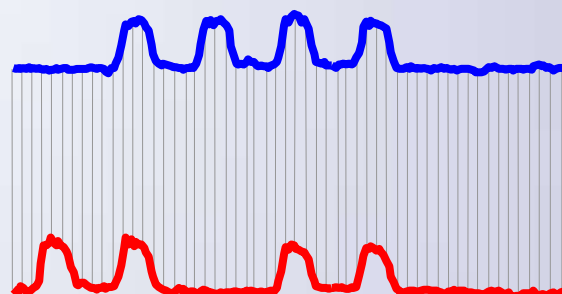
$$D_{dtw}(X, Y) = f(n, m)$$

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

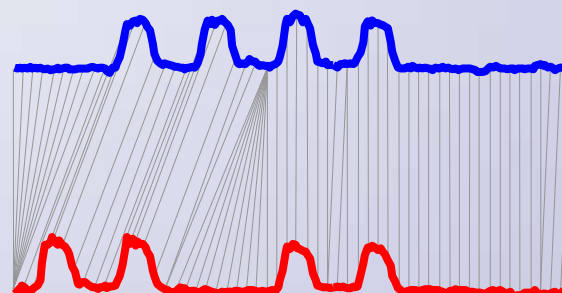
What do we want to do with them?

- complex analytics

Euclidean



Dynamic Time
Warping (DTW)



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)!**

Big Data Series



Large Synoptic
Survey Telescope
~30 TB / day

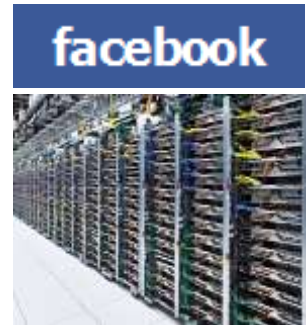


High-Throughput
Next Generation Sequencing
100 GB / day

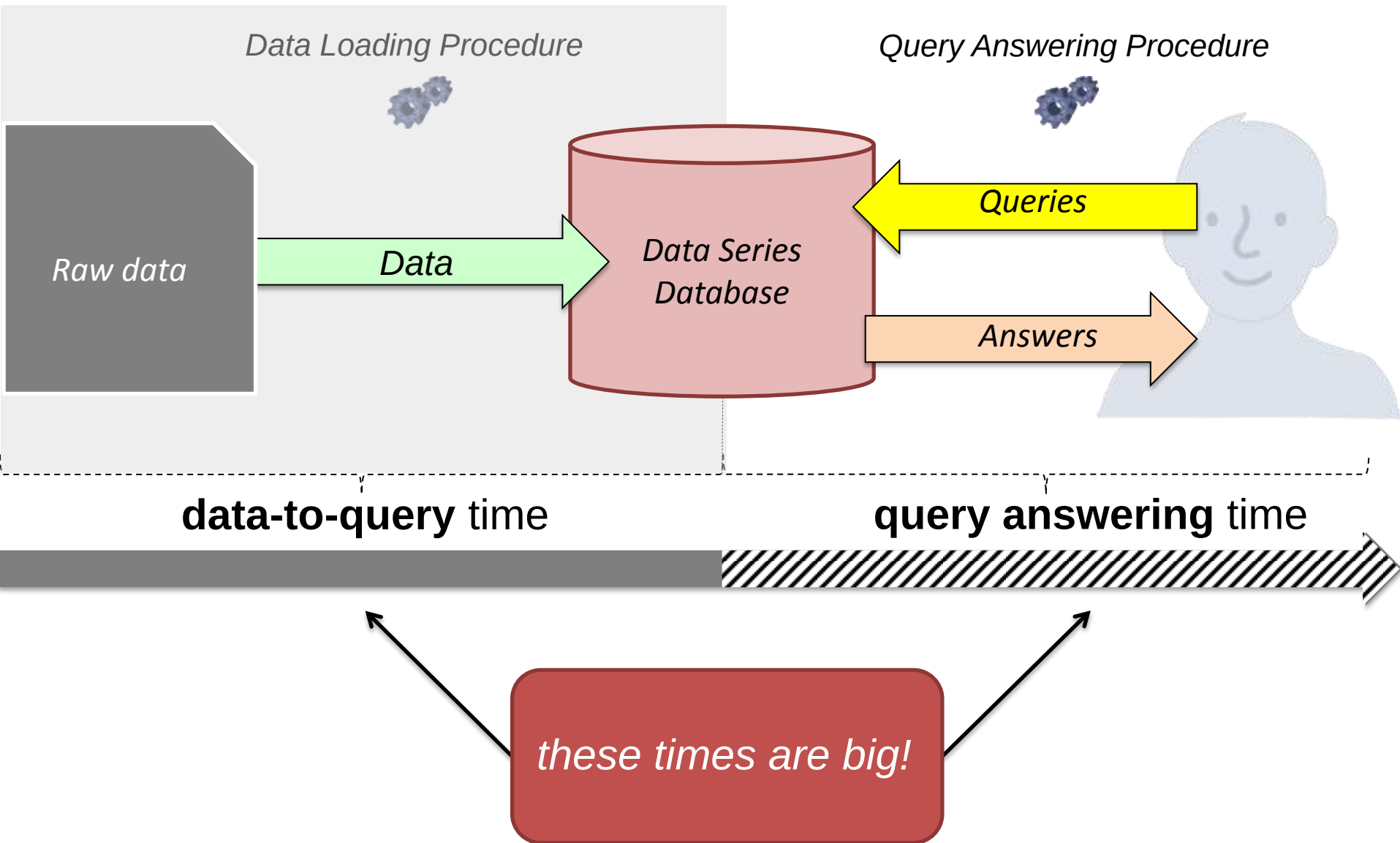


Boeing jet
20 TB / hour

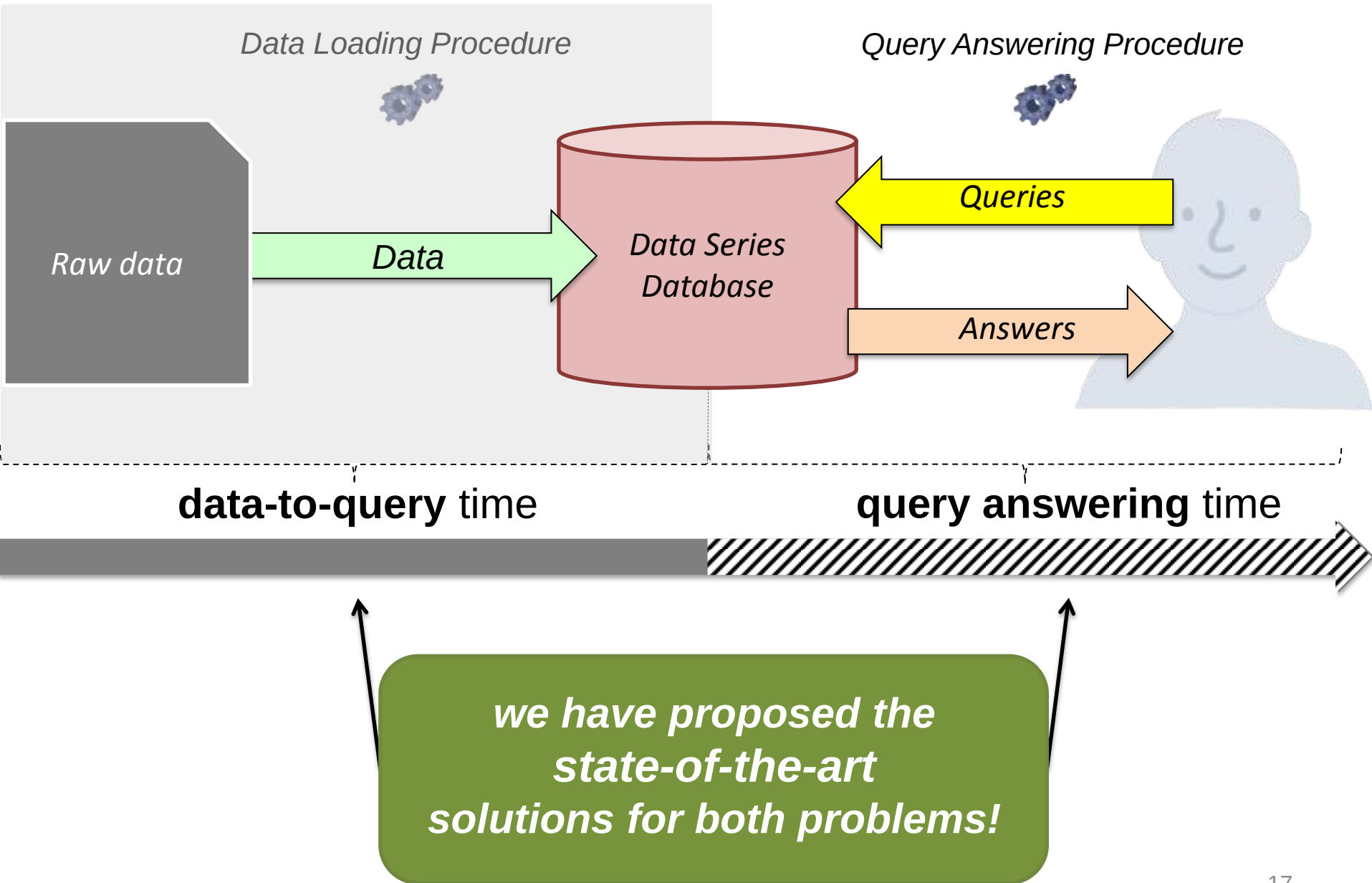
data center and
services monitoring
4B data series
12M points/sec



Query answering process



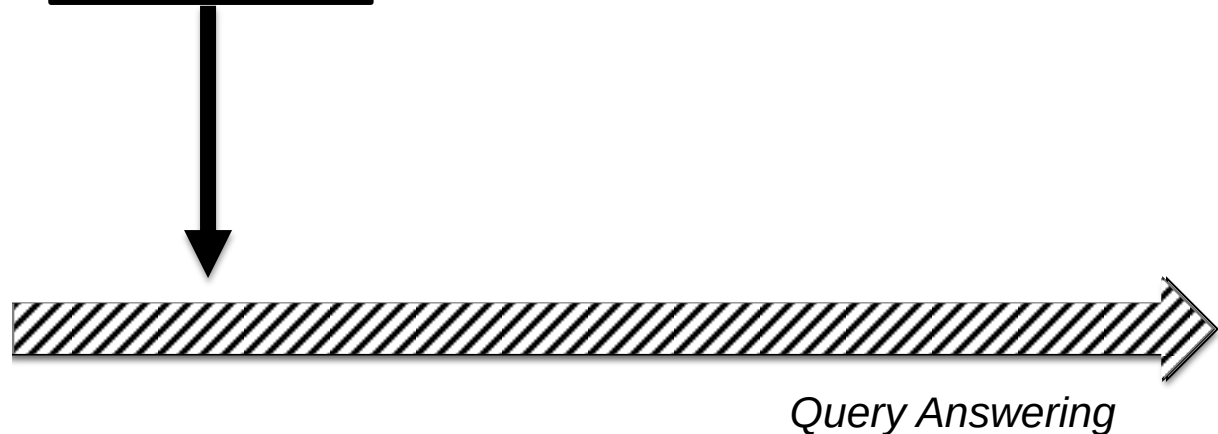
Query answering process



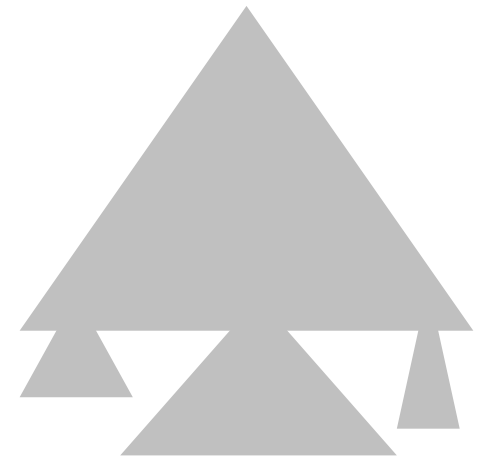
Ok, let's do it!

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

...serial scan takes **45 minutes** per query!



Good news: use **data series index**
for lowering **query answering time**

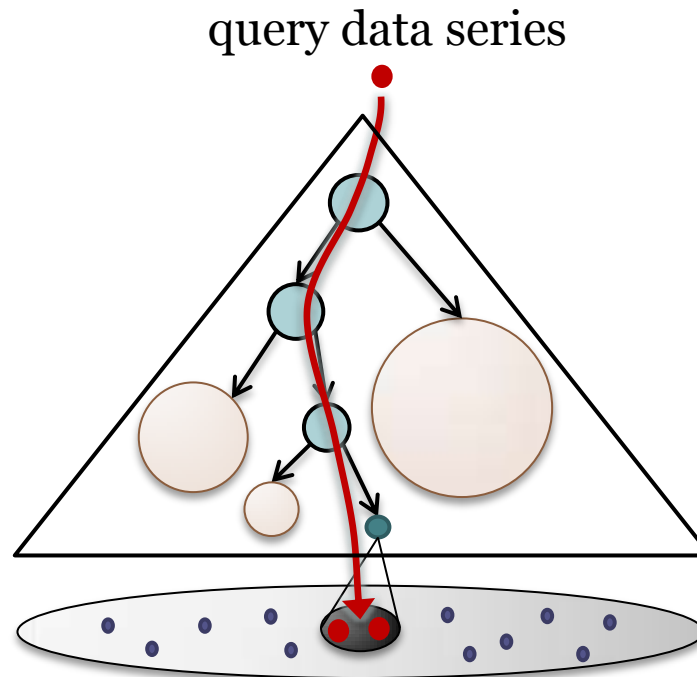


iSAX 2.0 for indexing and mining
data series collections [*Camerra et al.*
2008]



Query Answering

a few words on: Data Series Indexing



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

indexing 1 Billion data series
requires **days**

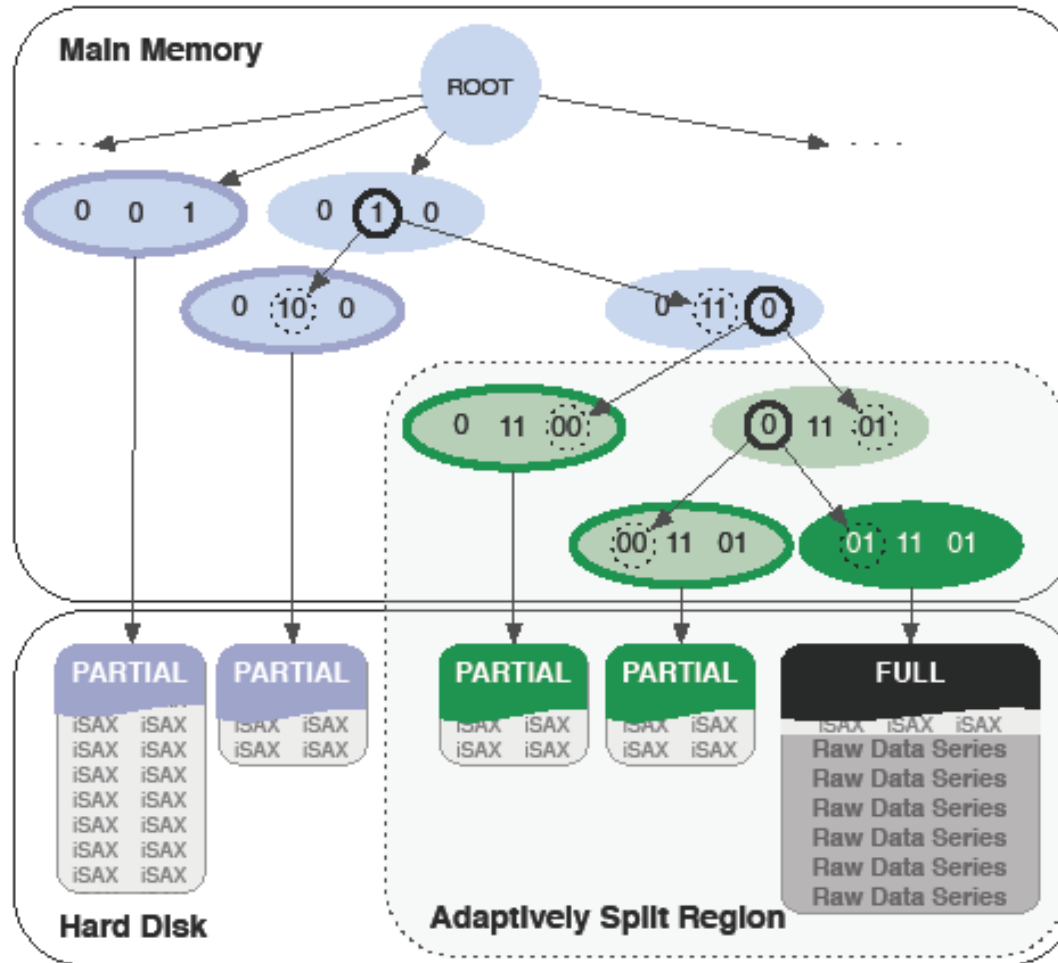


Data Loading

Query Answering

complex analytics in hours/days...

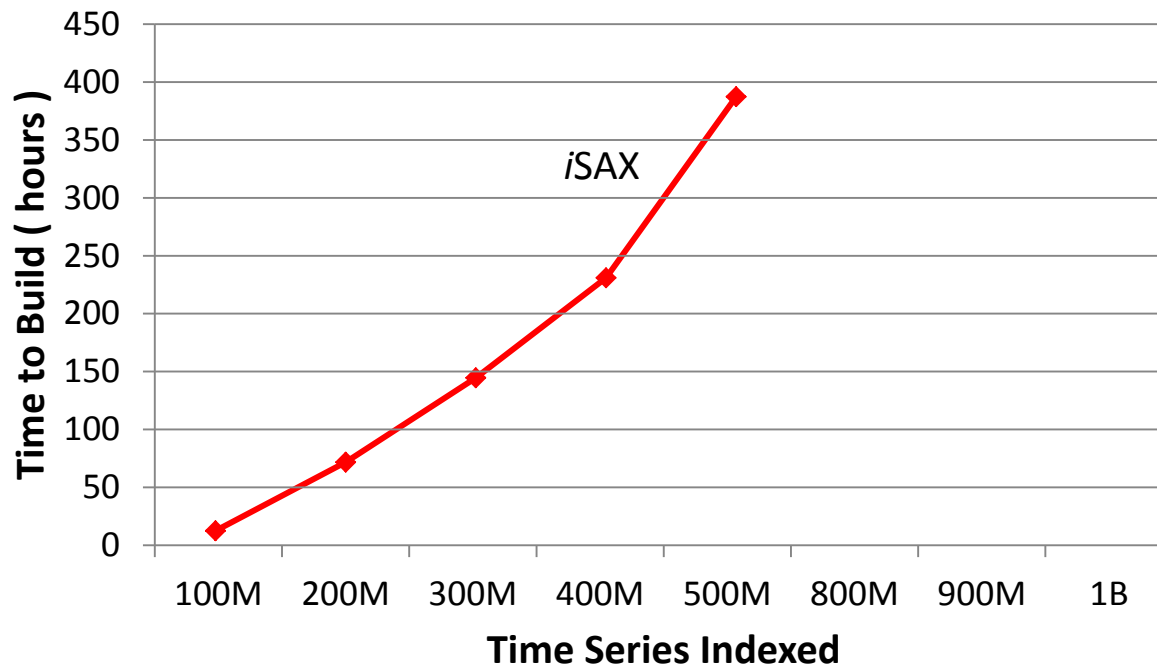
Data Series Indexes: iSAX2+, ADS+



complex analytics in minutes/seconds!

Experimental Evaluation

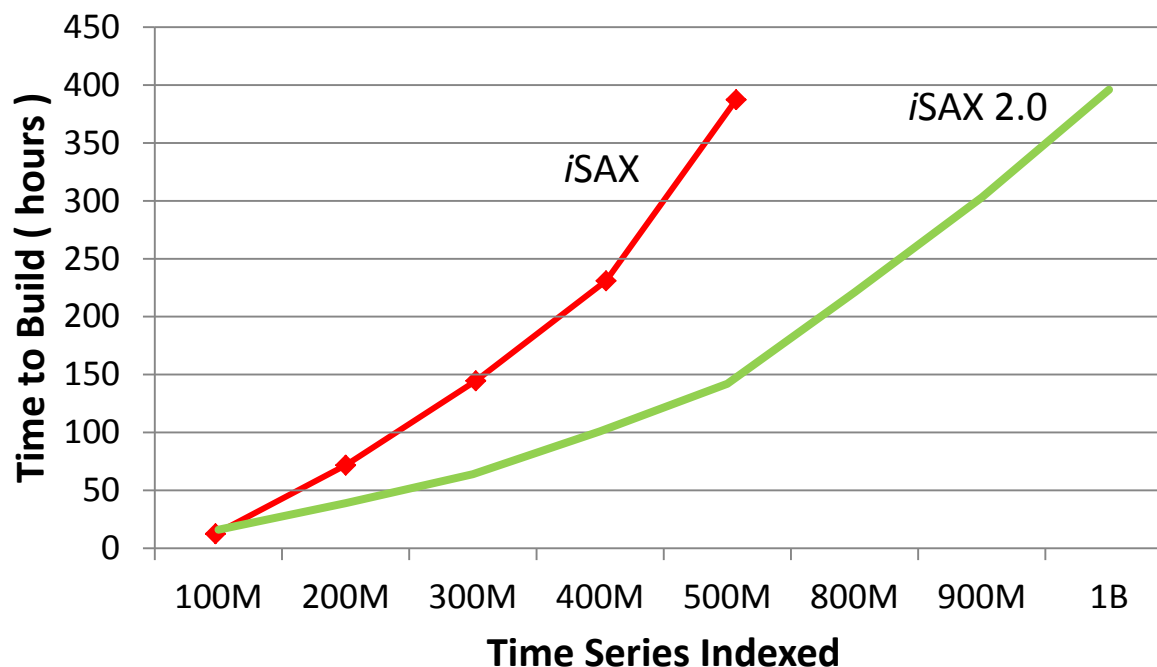
initial approach



- 1 Billion data series indexed in 60 days

Experimental Evaluation

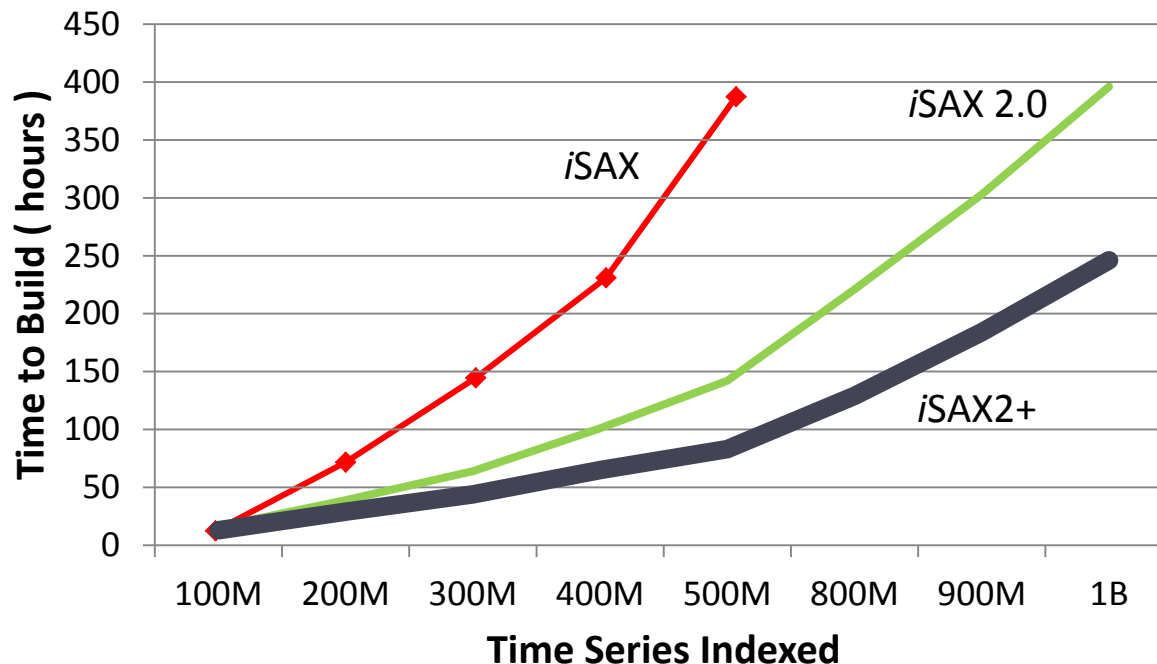
initial approach



- **1 Billion** data series indexed in **16 days**: **72% less time** than iSAX
- indexing time per data series: **1 millisecond**

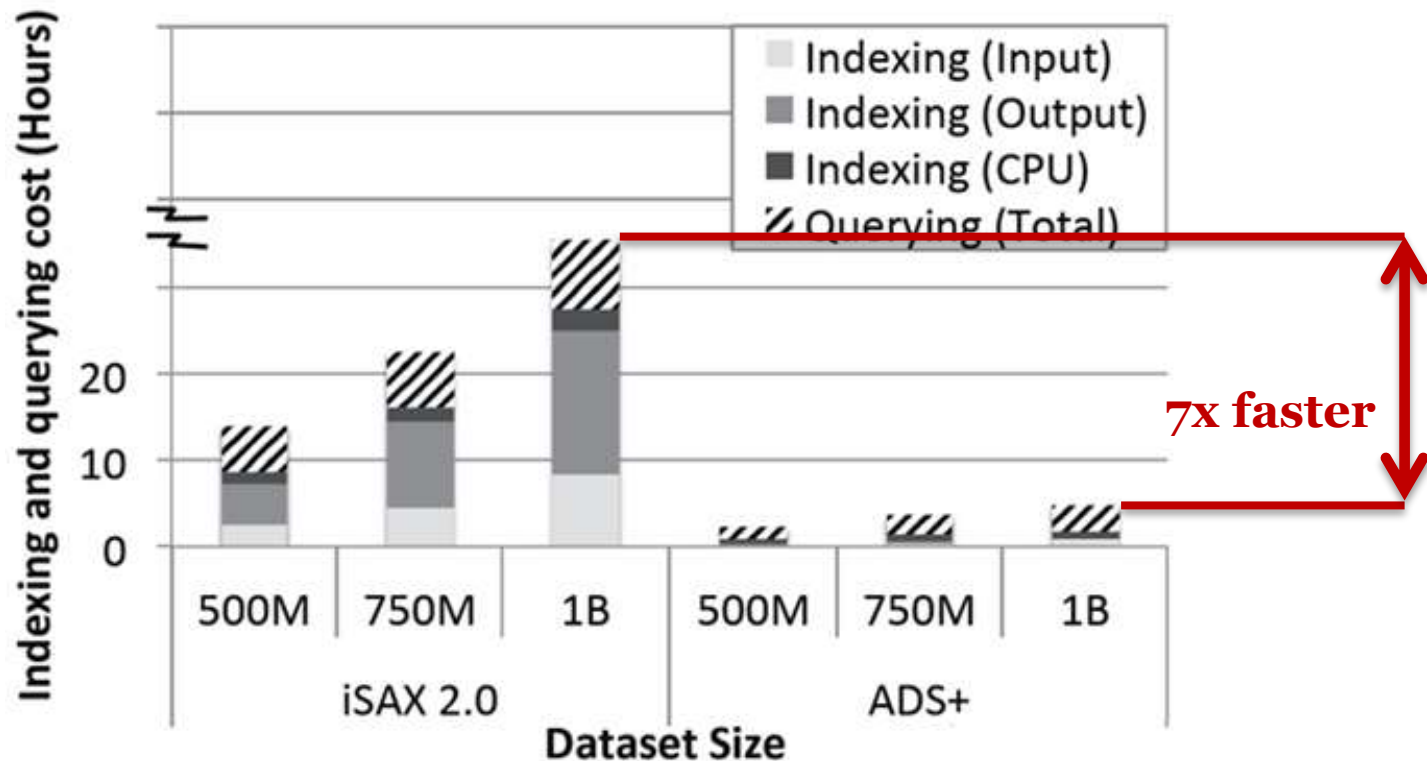
Experimental Evaluation

initial approach



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

Experimental Evaluation

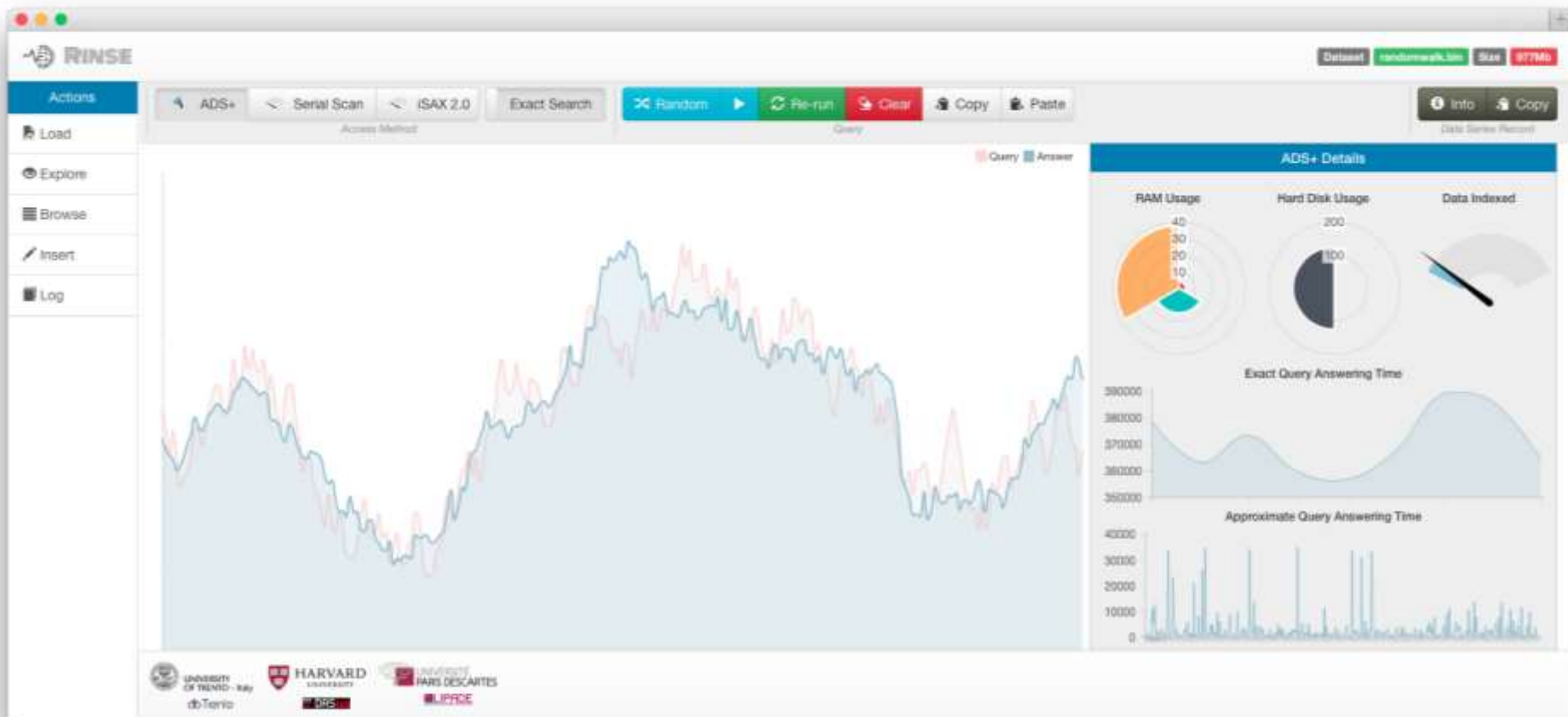


- iSAX 2.0 needs more than **35 hours** to answer 100K queries
- ADS+ answers 100K queries in less than **5 hours**

Demo

VLDB 2015

- <http://daslab.seas.harvard.edu/rinse/>



Conclusions

- developed 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**
 - **orders of magnitude** better performance
- experimentally validated our 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 very large data series collections

Time Series Analysis Workshop

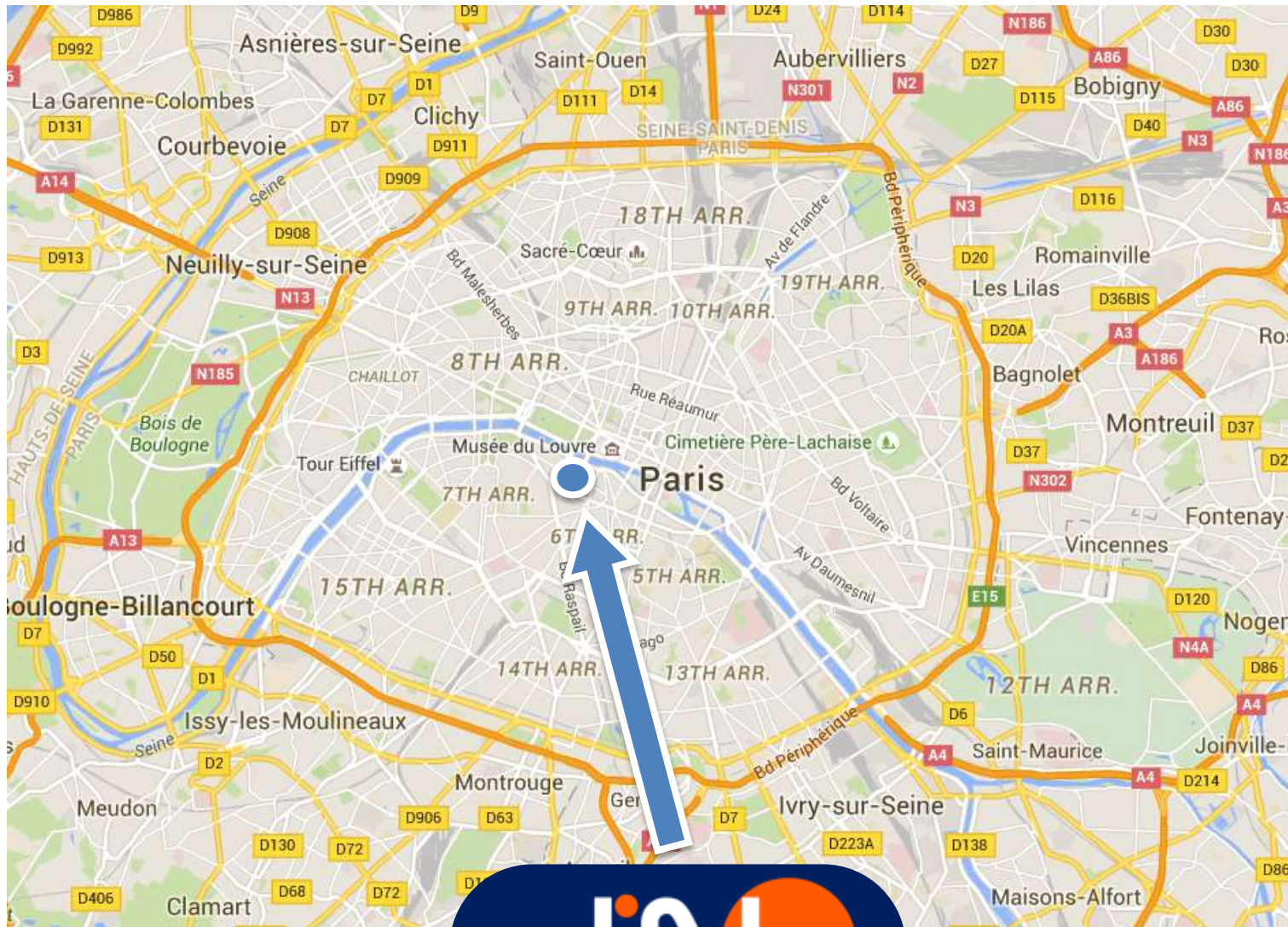
- interdisciplinary workshop on time series analysis
 - supported by CNRS / MASTODONS
- brings together different disciplines
 - astrophysics, neuroscience, engineering, ...

Time Series Analysis Workshop

- interdisciplinary workshop on time series analysis
 - supported by CNRS / MASTODONS
- brings together different disciplines
 - astrophysics, neuroscience, engineering, ...
- goals:
 - study characteristics/problems of large sequence collections
 - share experiences/expertise on data series analysis
 - develop techniques for efficient management and processing

collaborations!

Data-Intensive and Knowledge-Oriented systems



References

- papers

- **Big Sequence Management: A Glimpse on the Past, the Present, and the Future.** LNCS 9587, 2016
 - <http://www.mi.parisdescartes.fr/~themisp/publications/sofsem16-bisem.pdf>
- **Data Series Management: The Road to Big Sequence Analytics.** SIGREC 44(2), 2015
 - <http://www.mi.parisdescartes.fr/~themisp/publications/sigrec15-bisemvision.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/DStat>

- demo

- <http://daslab.seas.harvard.edu/rinse/>

thank you!

google: Themis Palpanas