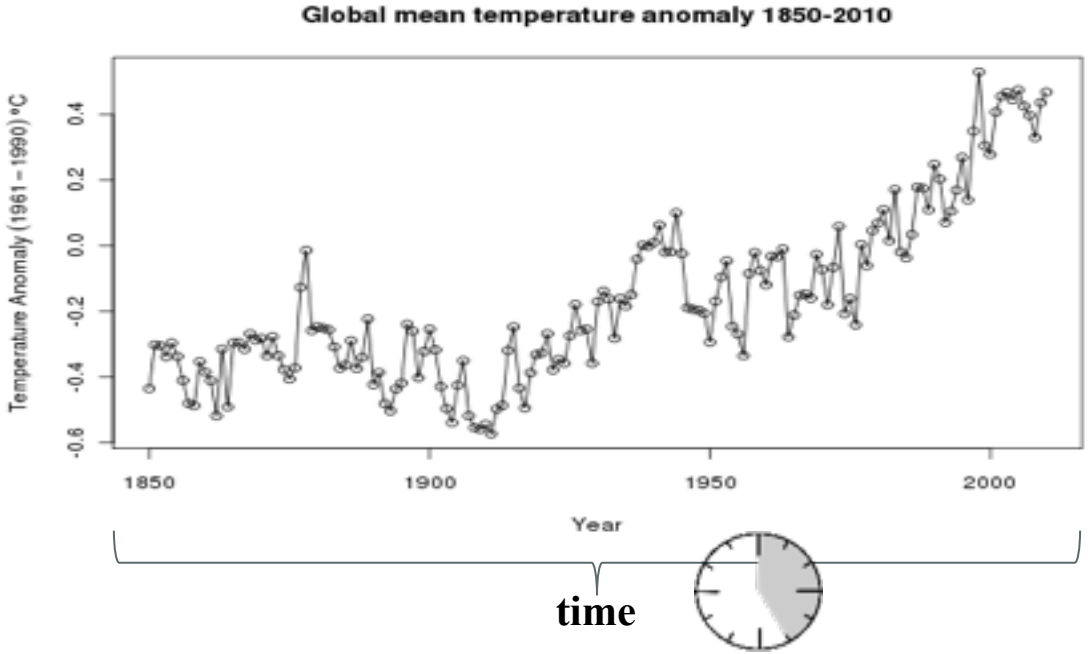
Interactive Visual Exploration of Large Data Series

Anna Gogolou

Time series

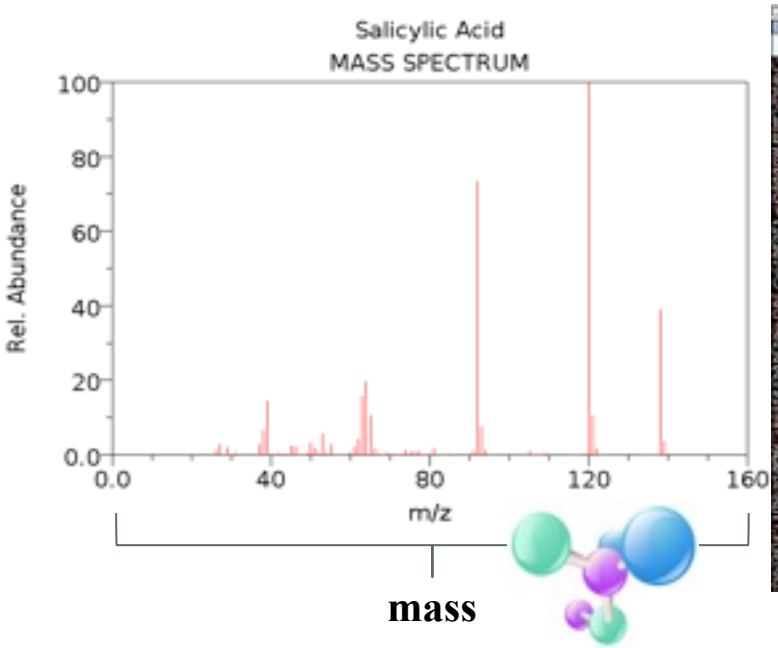
What are time series?



Data sequences ordered along the dimension of time

Data series

Observations, measurements ordered along a dimension (time, angle, mass, position, etc.)

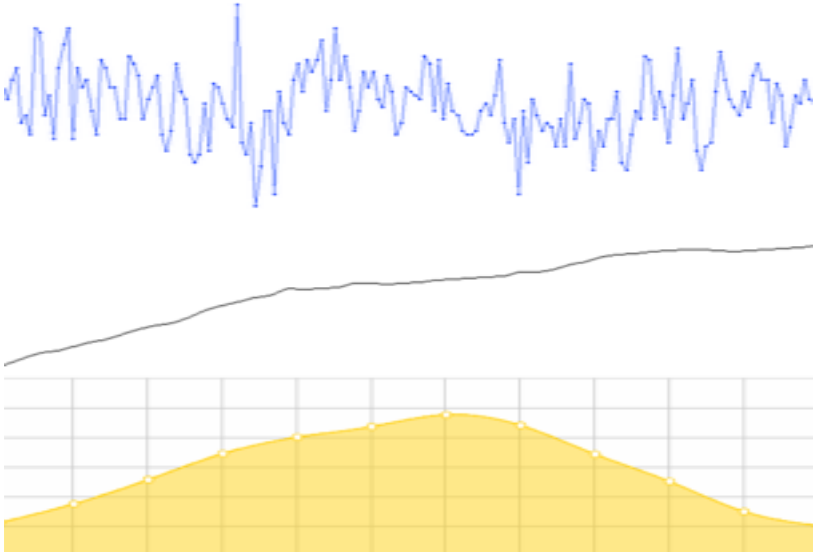


Why do we care about data series?

They occur in many scientific, medical, business, and social domains:

astronomy, biology, neuroscience, smart cities, nuclear power plants, finance, politics, …

- ➔ Our systolic/diastolic blood pressure
- ➔ Donald Trump’s popularity rating
- ➔ The annual sunshine in Paris



Outline

Motivation

Our goal

Existing data series visualization tools

Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

Motivation

How do we manage, explore and analyze large data series (order of terabytes)?

- More and more data series are produced every day:
  - 1 hour of ElectroCardioGraphy (ECG) data: 1 Gigabyte
  - Typical Weblog: 5 Gigabytes per week
  - EDF Database: ~100TB in total for all reactors (58), all sensors (10000 sensors per reactor), over all years and growing …

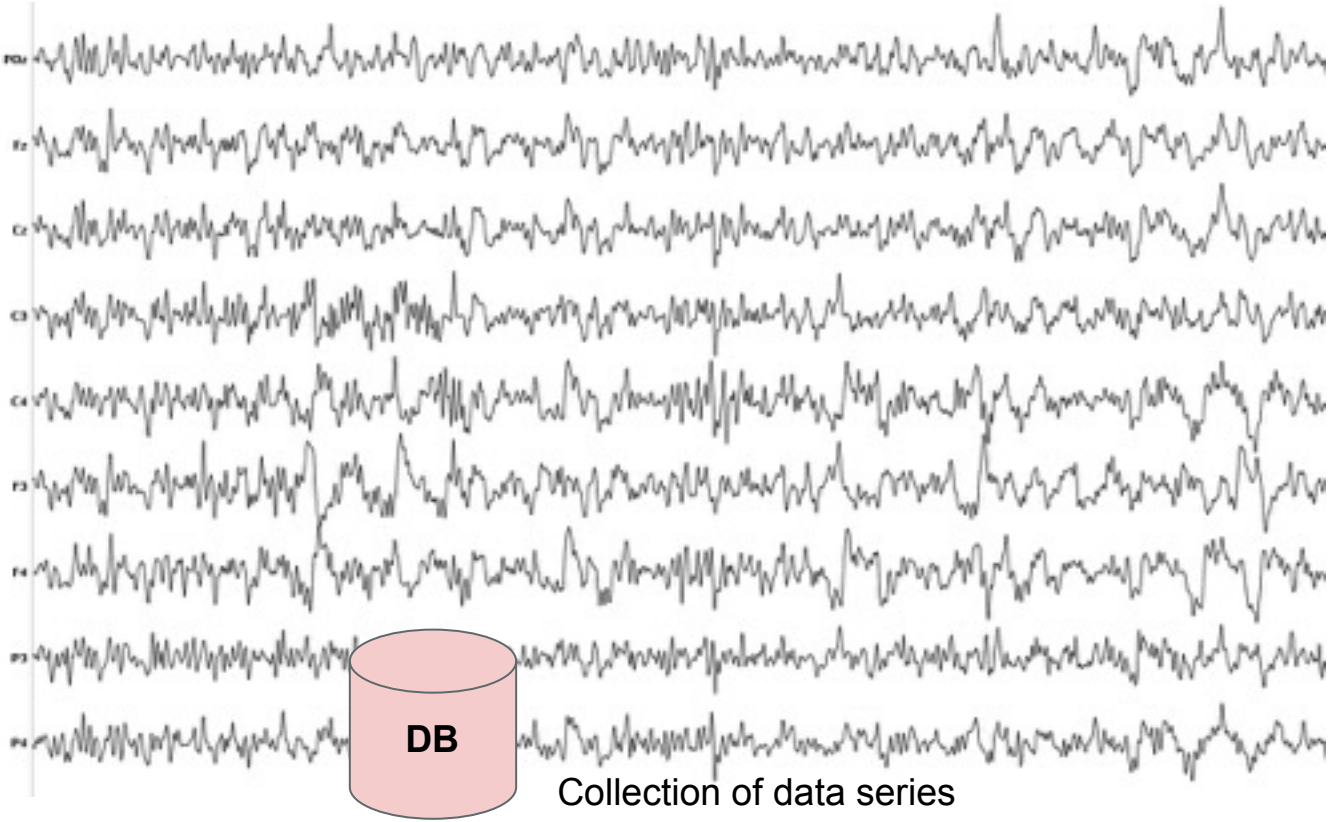
Data series analysis

Analysis Tasks:

- Pattern Matching
- Frequent Pattern Mining
- Clustering
- Classification
- Outlier Detection
- ...

Pattern matching

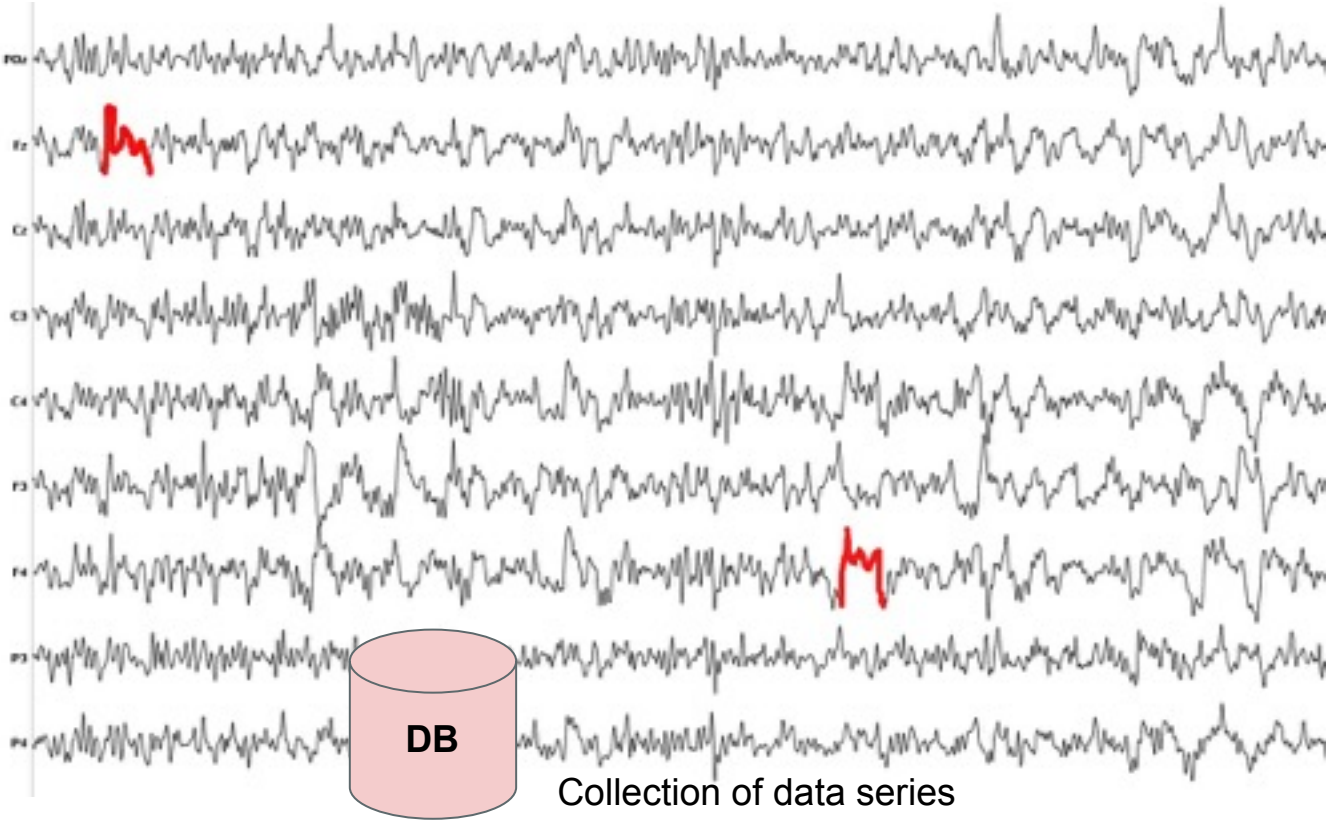
User-defined pattern



Collection of data series

Pattern matching

User-defined pattern



Collection of data series

Outline

Motivation

Our goal

Existing data series visualization tools

Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

Our goal

Interactive visual exploration of large data series

Interdisciplinary topic in the areas of:

- Human-Computer Interaction (HCI) & Information Visualization
  - Understand what analysts do with their data
  - Support interactive exploration of large datasets
- Databases
  - Analyze efficiently large data series
  - Get answers quickly

Outline

Motivation

Our goal

Existing data series visualization tools

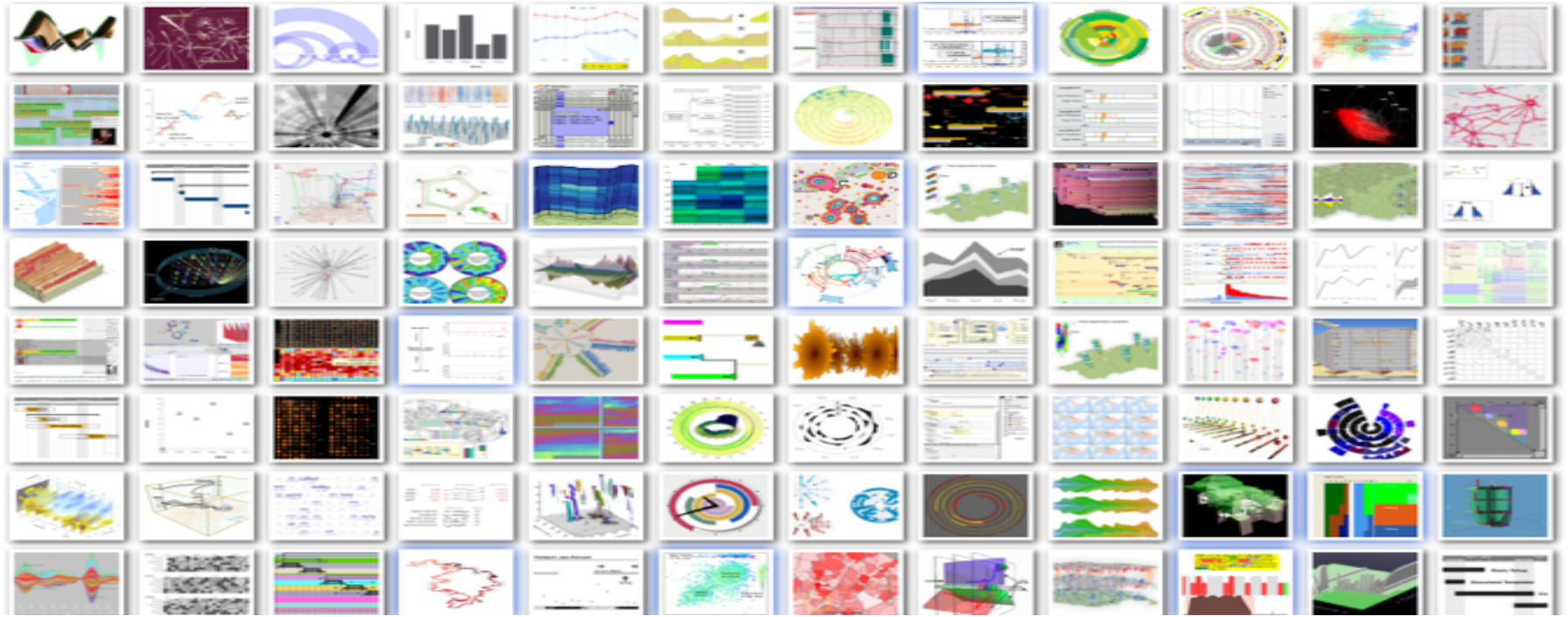
Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

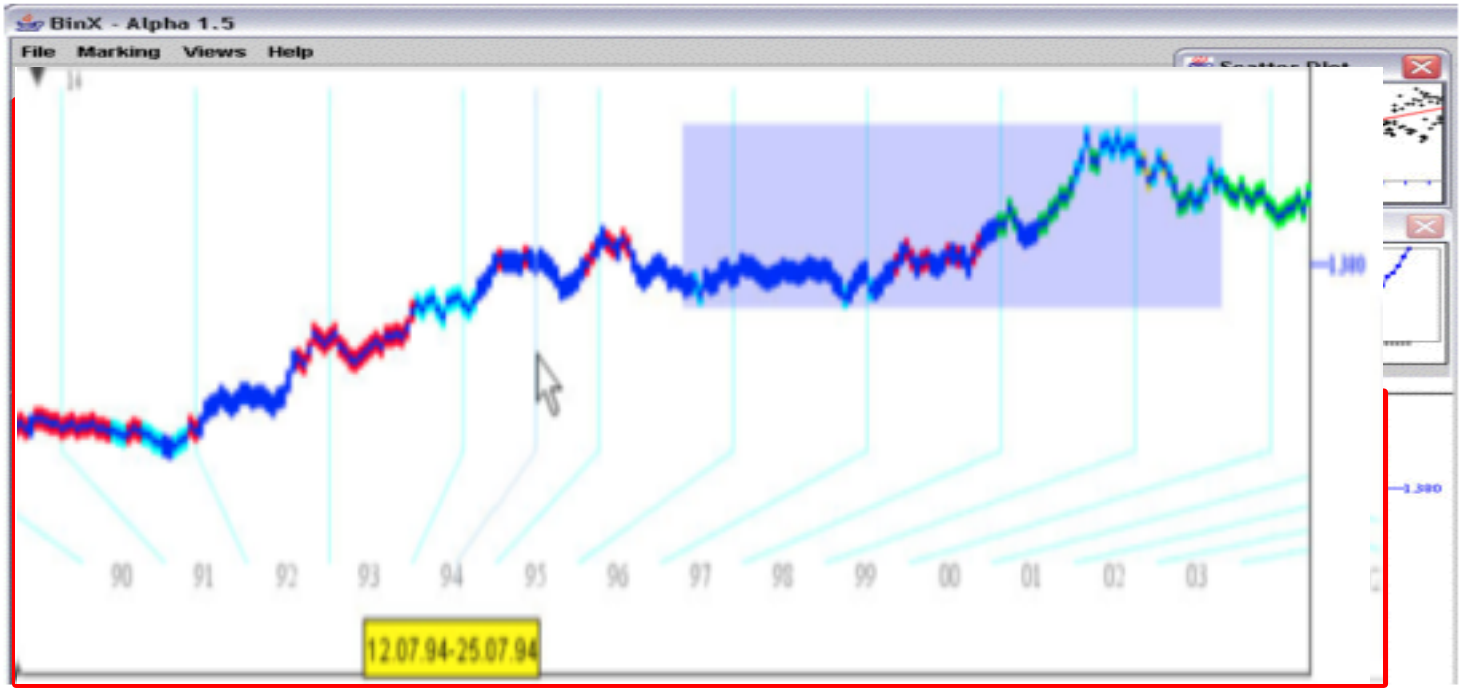
Existing data series visualization tools



From http://survey.timeviz.net/

BinX

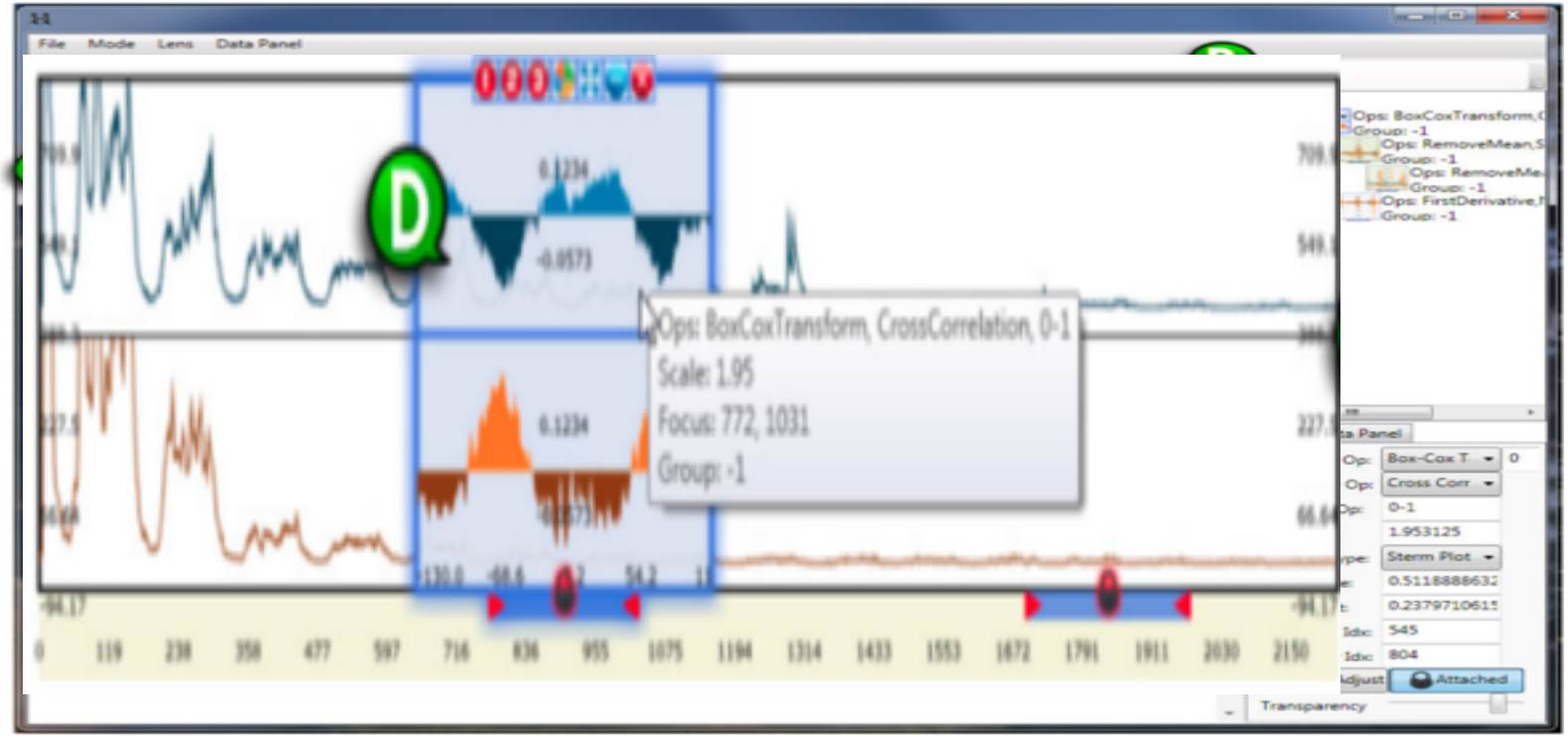
High level of aggregation (Overview)



Lior Berry, Tamara Munzner, "Binx: Dynamic exploration of time series datasets across aggregation levels", *Information Visualization 2004. INFOVIS 2004. IEEE Symposium on pages p2-p2. IEEE*, vol. 5, 2004. [doi>10.1109/INFVIS.2004.11]

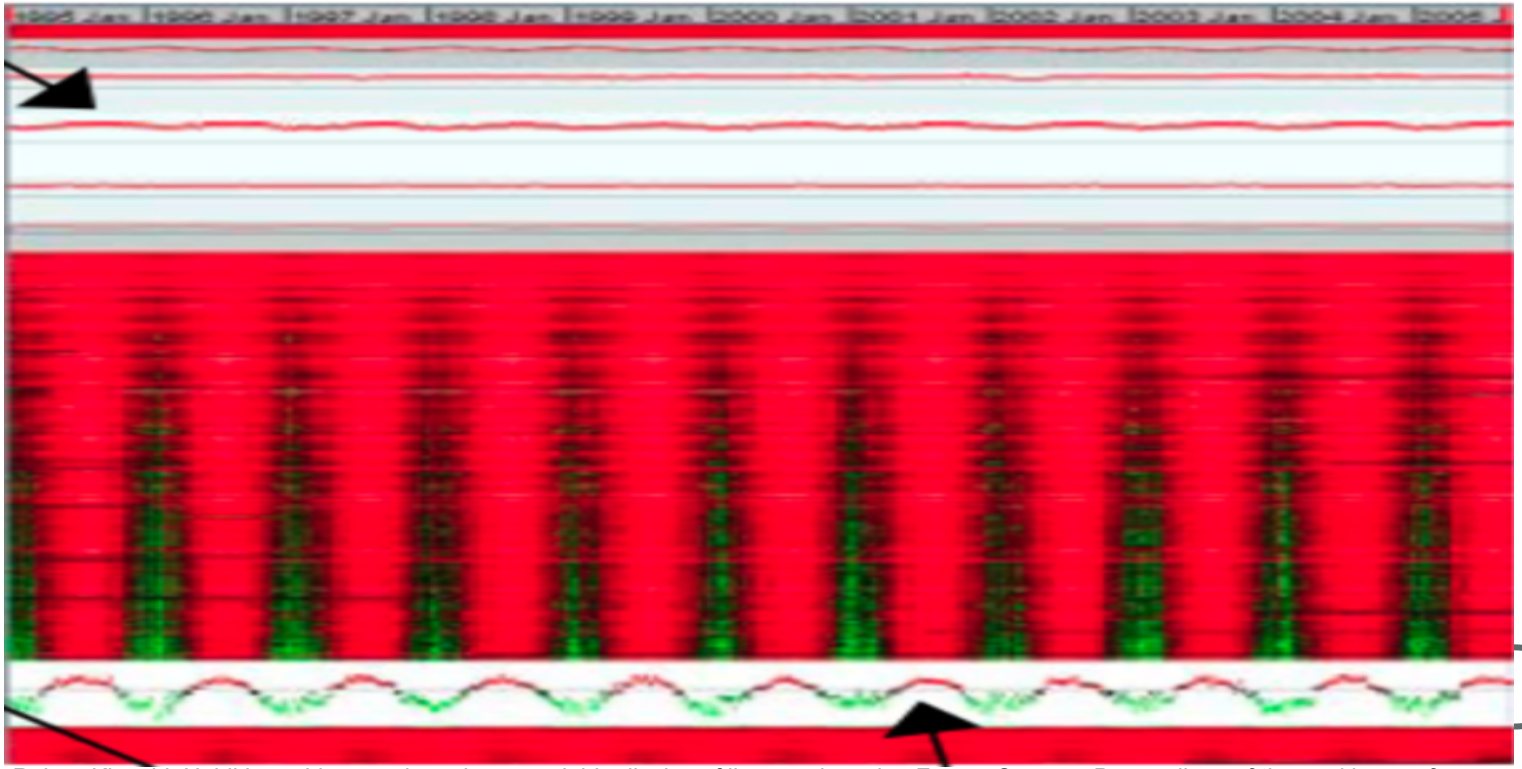
ChronoLenses

Data series chart panel



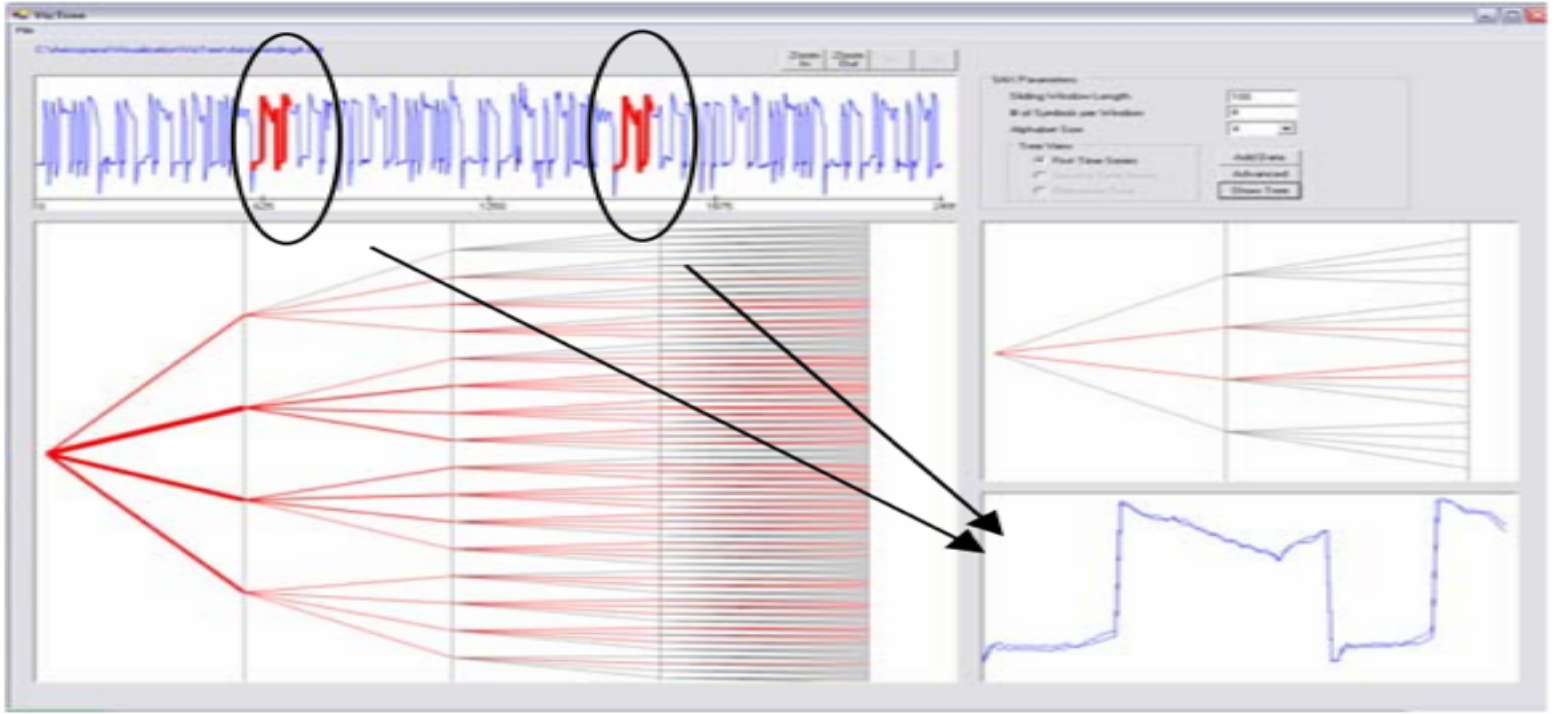
Jian Zhao, Fanny Chevalier, Emmanuel Pietriga, Ravin Balakrishnan, Exploratory Analysis of Time-Series with ChronoLenses, *IEEE Transactions on Visualization and Computer Graphics*, v.17 n.12, p.2422-2431, December 2011 [doi>10.1109/TVCG.2011.195]

Line graph explorer



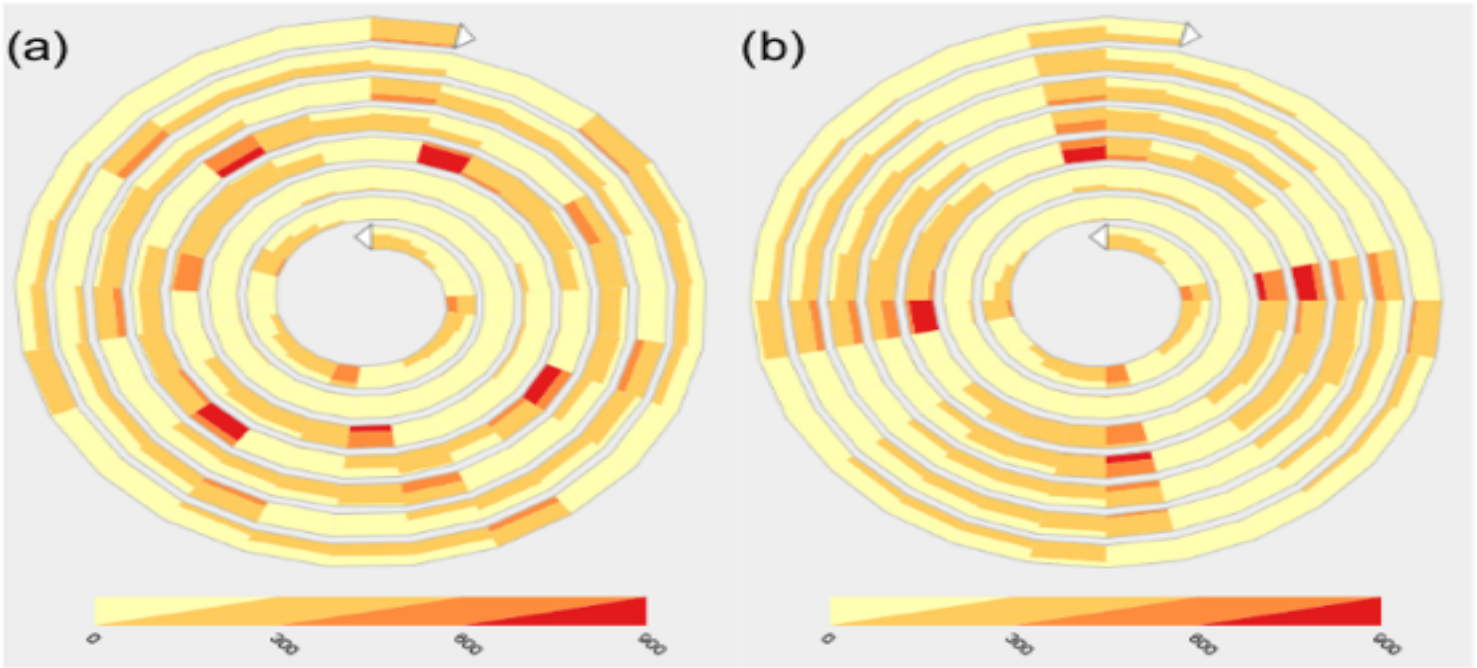
Robert Kincaid, Heidi Lam, Line graph explorer: scalable display of line graphs using Focus+Context, *Proceedings of the working conference on Advanced visual interfaces*, May 23-26, 2006, Venezia, Italy [doi>10.1145/1133265.1133348]

Viztree



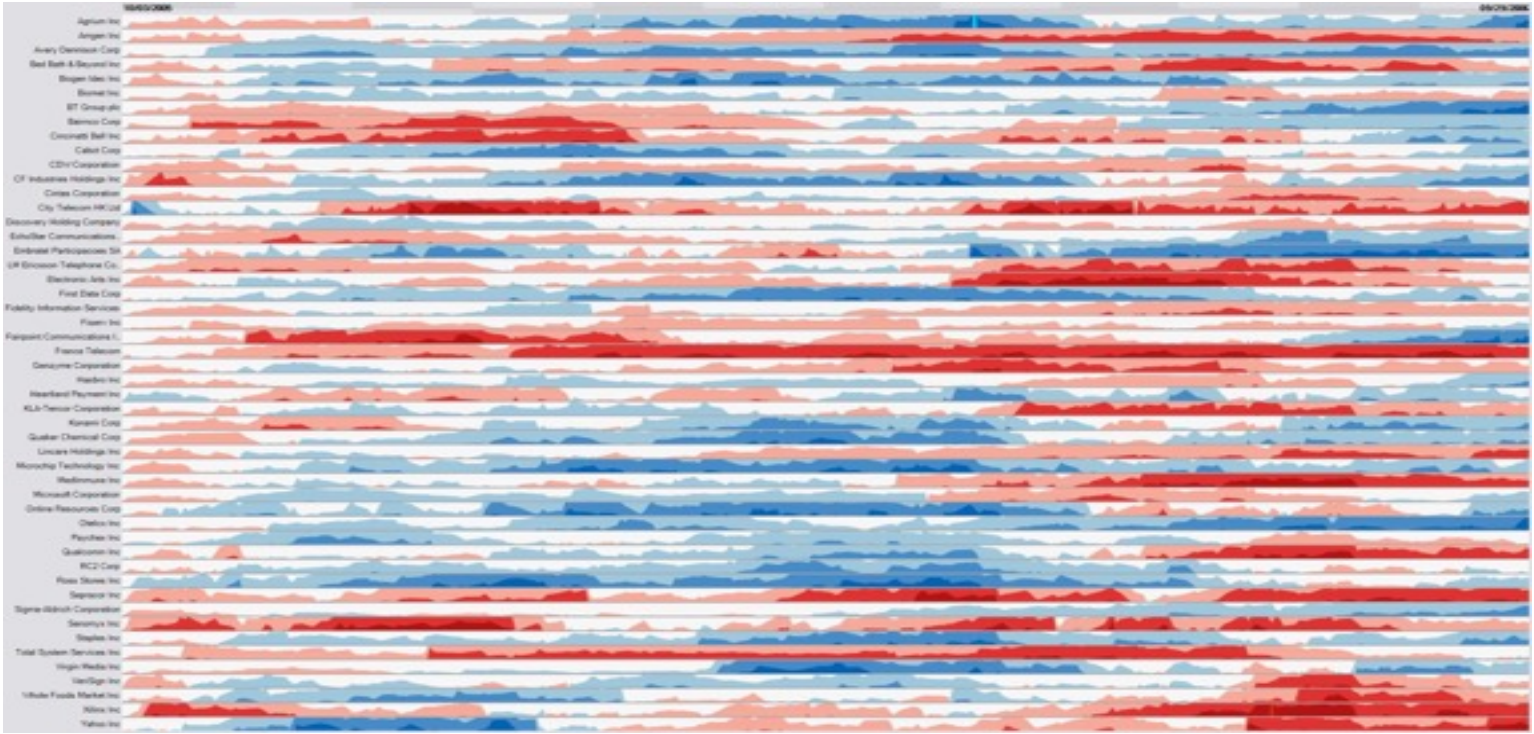
Jessica Lin, Eamonn Keogh, Stefano Lonardi, Visualizing and discovering non-trivial patterns in large time series databases, *Information Visualization*, v.4 n.2, p.61-82, July 2005 [doi>10.1057/palgrave.ivs.9500089]

Spirals



Christian Tominski, Heidrun Schumann, Enhanced interactive spiral display, *Proceedings of the annual SIGRAD conference*, pp.53–56 (2008)

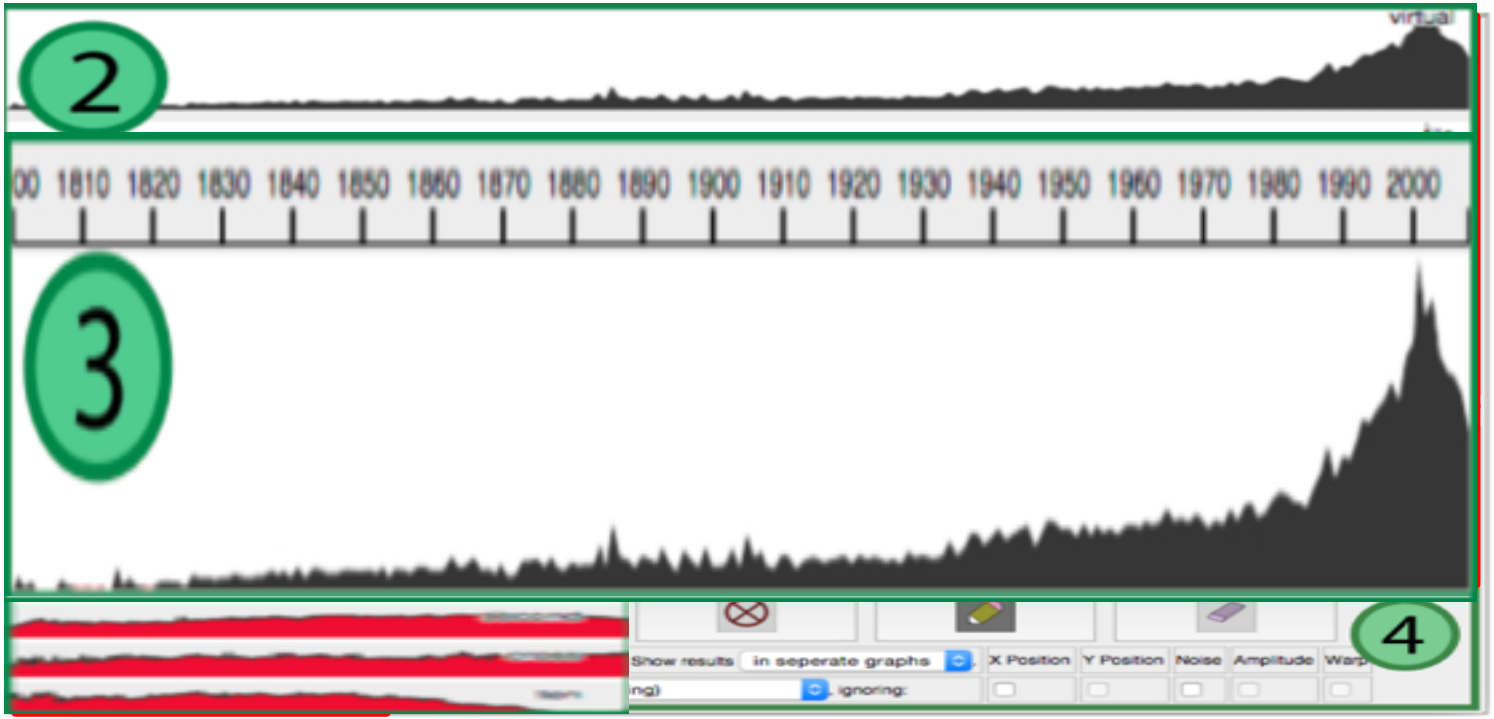
Horizon charts



Jeffrey Heer, Nicholas Kong, Maneesh Agrawala, Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 04-09, 2009, Boston, MA, USA [doi>10.1145/1518701.1518897]

The semantics of sketch

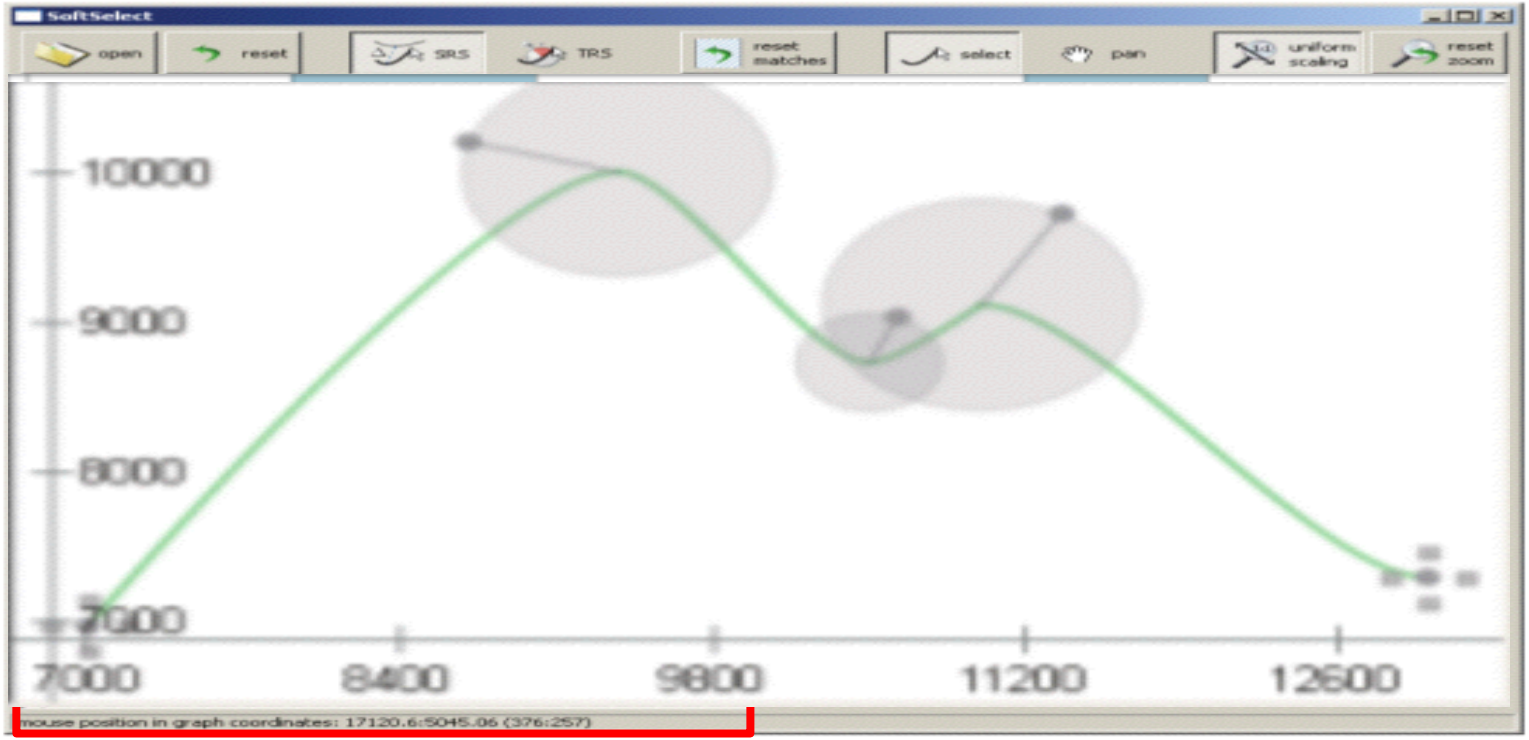
The matching panel



Michael Correll, Michael Gleicher, The Semantics of Sketch: A Visual Query System For Times Series Data, *Proceedings of the 2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, October 2016

SoftSelect

The graph-display and main-interaction area



Christian Holz, Steven Feiner, Relaxed selection techniques for querying time-series graphs, *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, October 04-07, 2009, Victoria, BC, Canada [doi>10.1145/1622176.1622217]

Outline

Motivation

Our goal

Existing data series visualization tools

Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

What about handling billions of data series?

None of these tools support scalability to terabytes of data:

How do we visualize them?

- Limited number of pixels

Visual scalability

- Limited human cognitive resources

Interactive response time scalability

How do we interact with them?

- Days to answer a single query

Limit: < 100 ms

J. Nielsen, Response times: The 3 important limits, https://www.nngroup.com/articles/response-times-3-important-limits/, January 1, 1993

Outline

Motivation

Our goal

Existing data series visualization tools

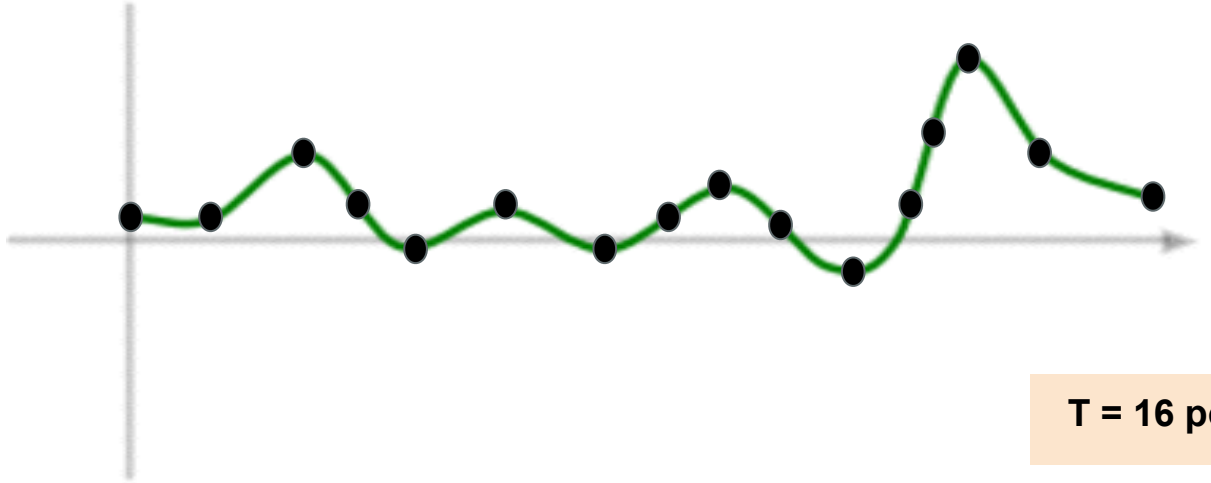
Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

Data series summarization

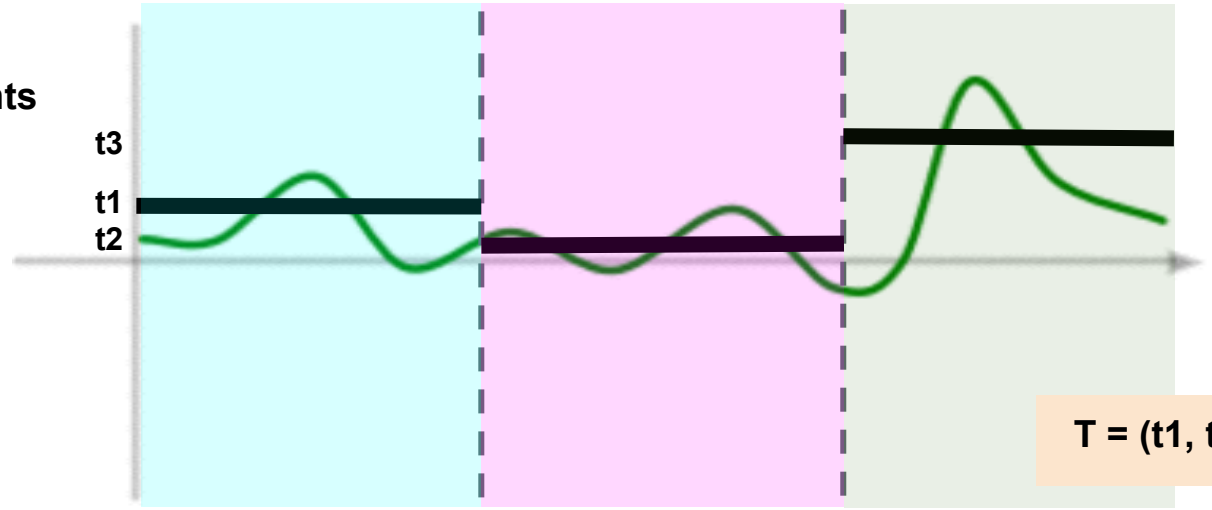


T = 16 points

A raw data series T

Data series summarization

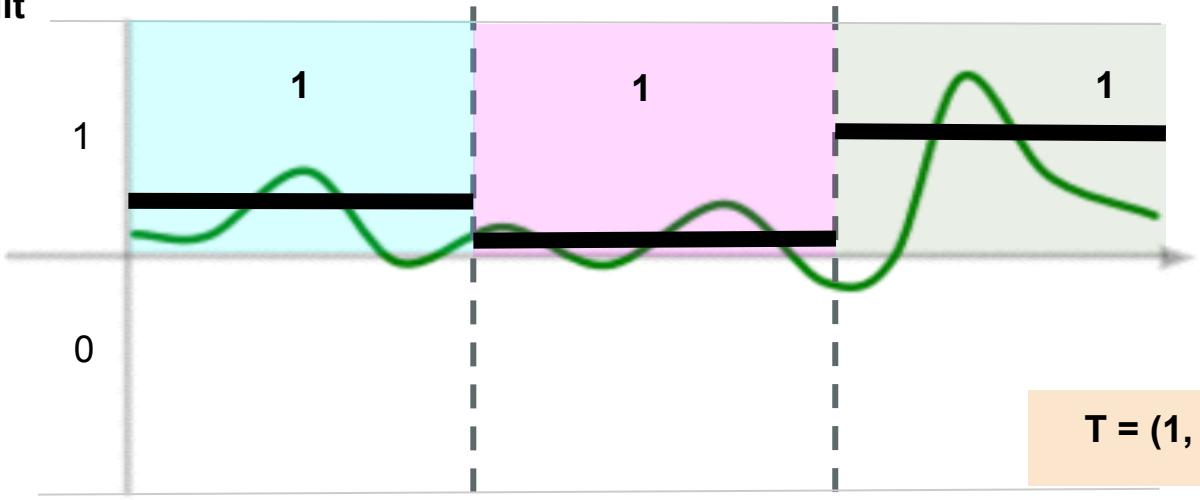
Let's split X-axis into 3 equal segments



Piecewise Aggregate Approximation (PAA)

Data series summarization

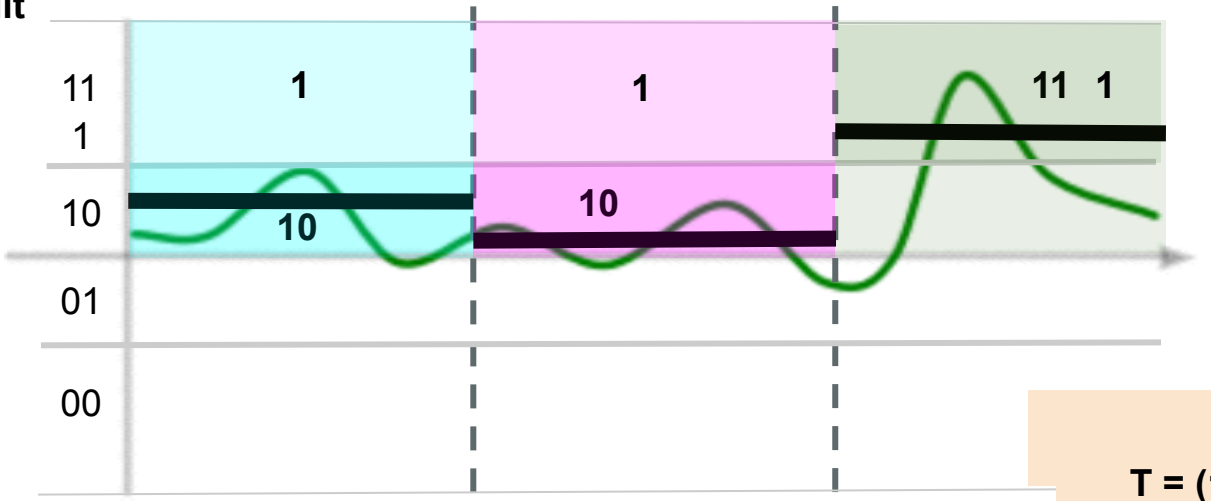
Let's further split Y-axis



Symbolic Aggregate approXimation (SAX)

Data series summarization

Let's further split Y-axis



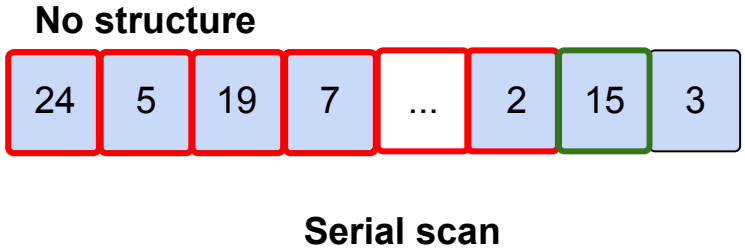
Symbolic Aggregate approXimation (SAX)

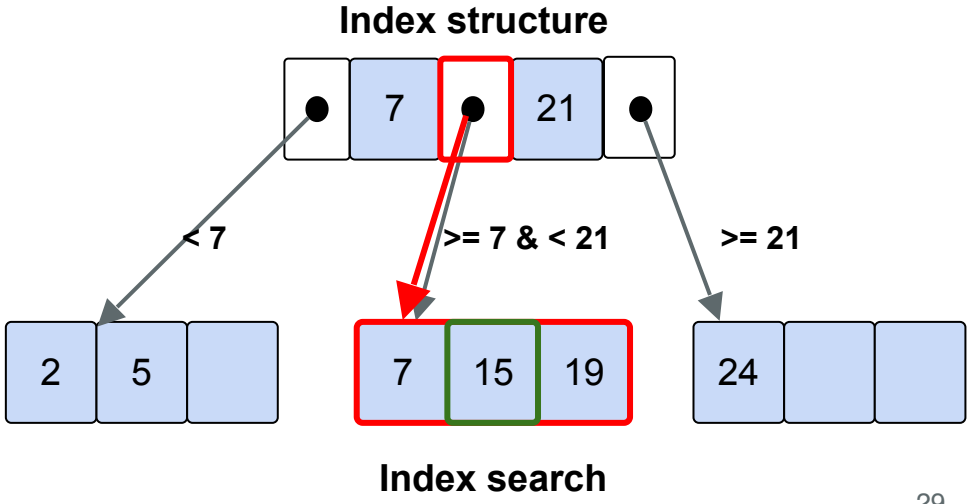
Indexing

How can queries be answered faster?

Using index structures (usually trees), which improve the speed of database operations on database tables (insert, select, delete, ...).

How do indexes work?





Data series indexes

Data series indexes follow a similar tree structure.

They are built on top of data series summarizations.

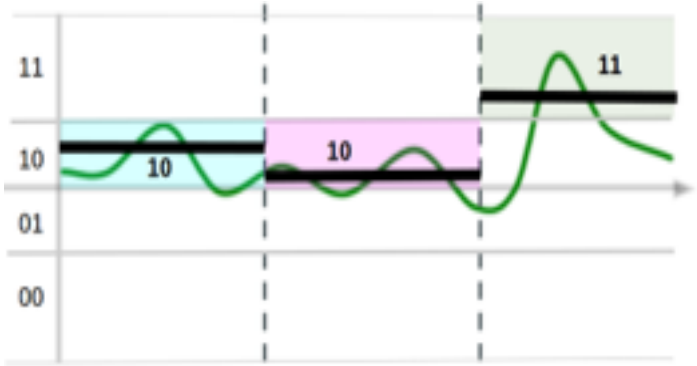
ADS+[1] is the state-of-the-art data series index. It is based on:

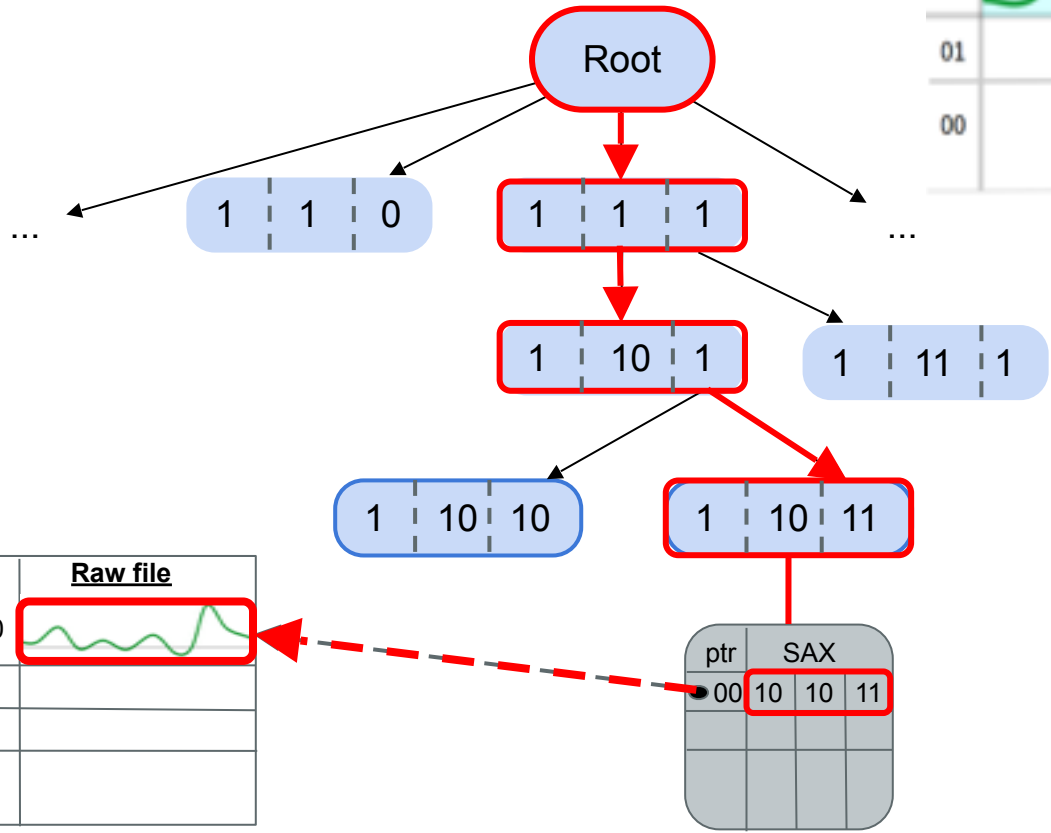
SAX representation

Euclidean Distance (ED) between SAX representations

[1]Zoumpatianos, K., Idreos, S., Palpanas, T., ADS: the adaptive data series index, *The VLDB Journal* (2016) 25: 843. [doi>10.1007/s00778-016-0442-5]

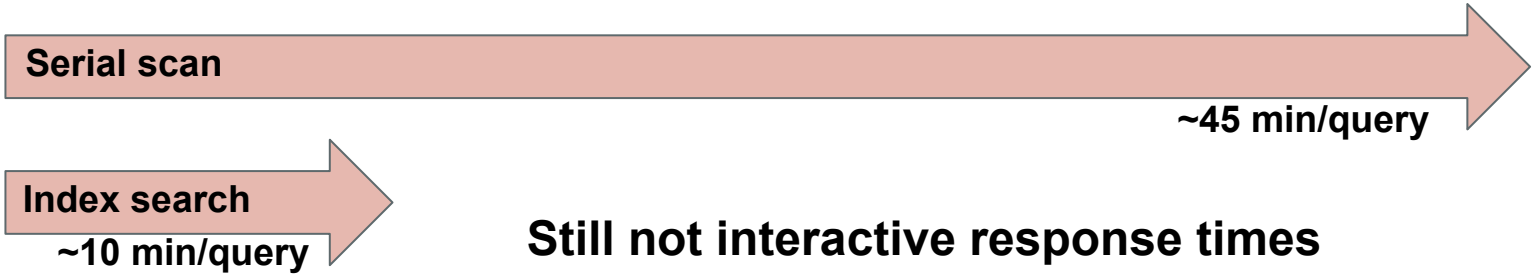
ADS+ index


11
10
01
00
10
10
11


Root
...
1 1 0
1 1 1
...
1 10 1
1 11 1
1 10 10
1 10 11
Raw file
00
ptr
SAX
00 10 10 11

Response times

Finding the most similar pattern on 1 **billion** data series:



Still not interactive response times
(not in the order of milliseconds)

Outline

Motivation

Our goal

Existing data series visualization tools

Scalability problems

State-of-the-art in data series management

Approximate vs exact search

Progressive visualizations

Approximate vs exact search

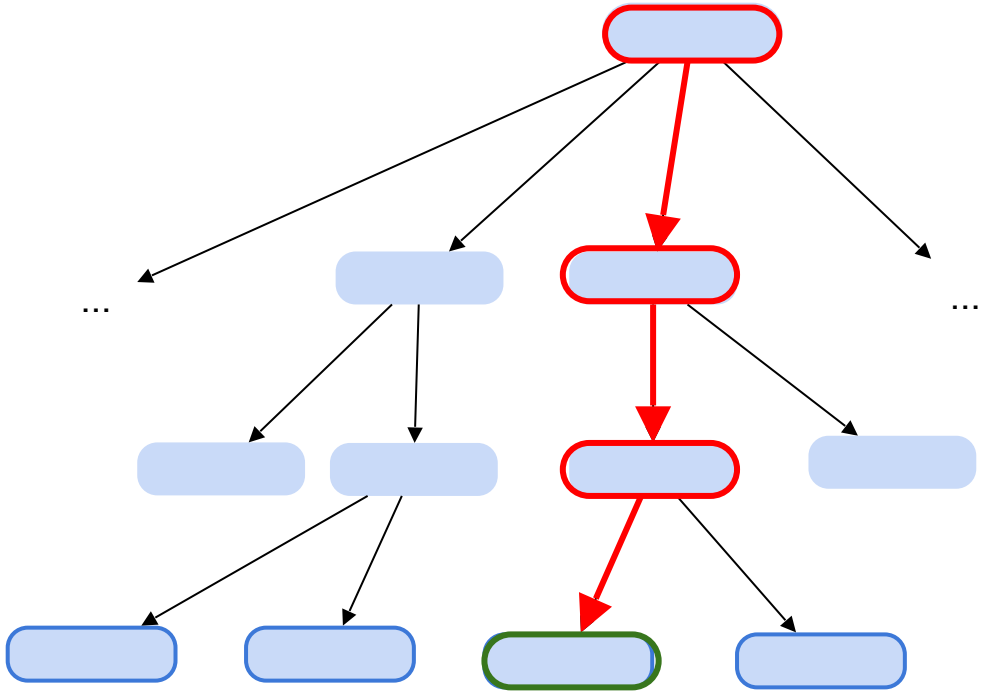
Approximate search:

- Returns answer not guaranteed to be the best one
  Associated with an approximation error
- Very fast (order of milliseconds)

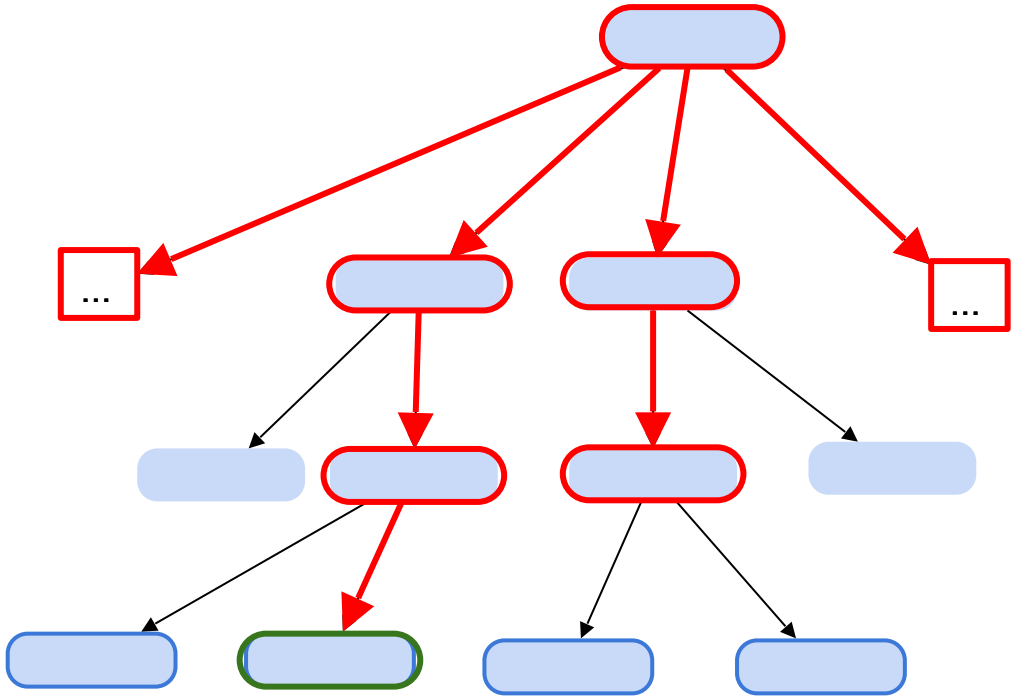
Exact search:

- Always returns exact correct answer
- Slow (order of minutes)

Approximate search



Exact search



Outline

Motivation

Our goal

Existing data series visualization tools

Scalability problems

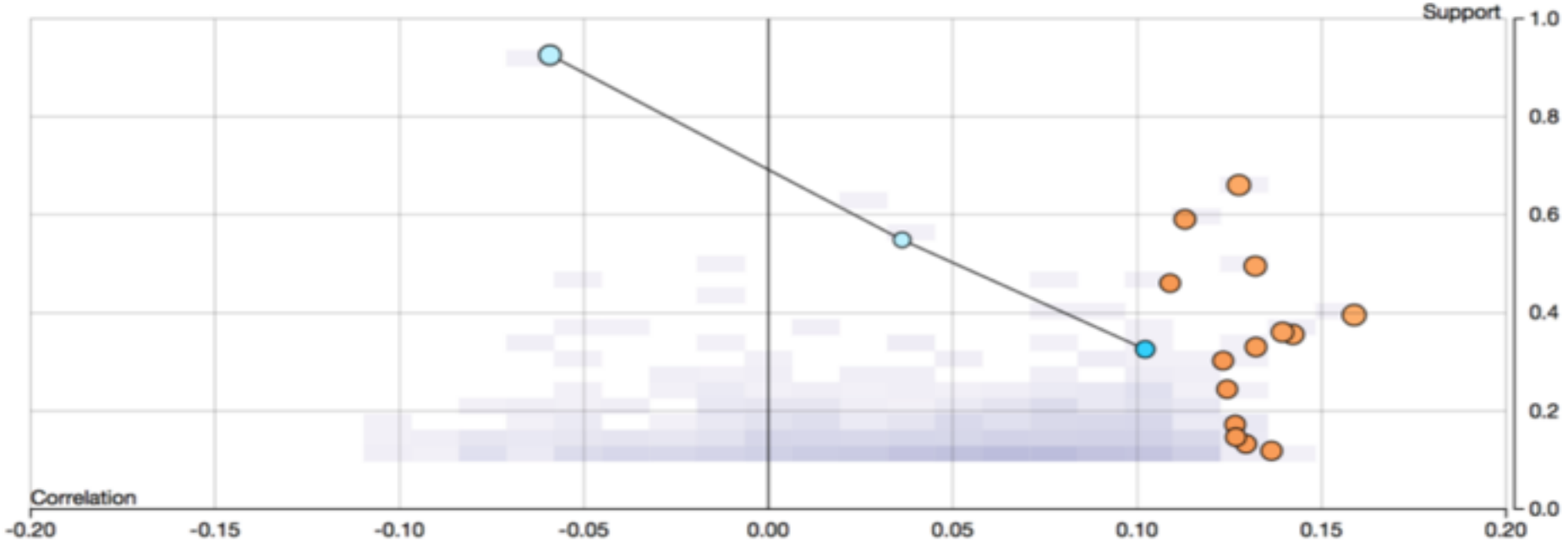
State-of-the-art in data series management

Approximate vs exact search

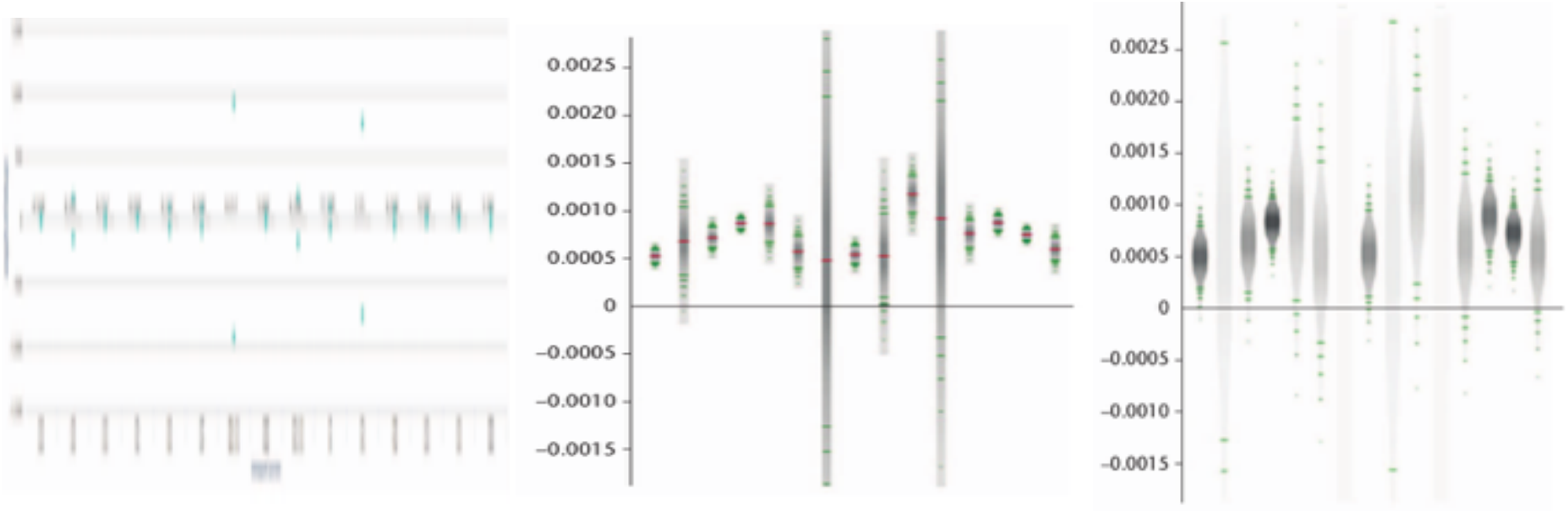
Progressive visualizations

Progressive visualizations

How can users progressively move from approximate to exact results?



Visualizing approximation error



Danyel Fisher, Steven M. Drucker, Arnd Christian Koenig, Exploratory Visualization Involving Incremental, Approximate Database Queries and Uncertainty, *Journal IEEE Computer Graphics and Applications*, Volume 32, Issue 4, July 2012, Pages 55-62, [doi>10.1109/MCG.2012.48]

Challenges

Provide quickly partial results

Visualize approximation error

Support iteration and refinement over approximate & progressive results

Easily focus on subspaces of interest

Conclusions

Open problems:

- Scalability to terabytes of data
  - Current visualization tools cannot handle large data series collections
  - Need techniques that:
    - effectively visualize large volumes of data
    - have interactive response times (< 100ms)
- Support for iteration and refinement of approximate & progressive results
  - Currently no such support for data series analytics
  - Need techniques that:
    - visually inform users on progress of task
    - convey information on accuracy of current results