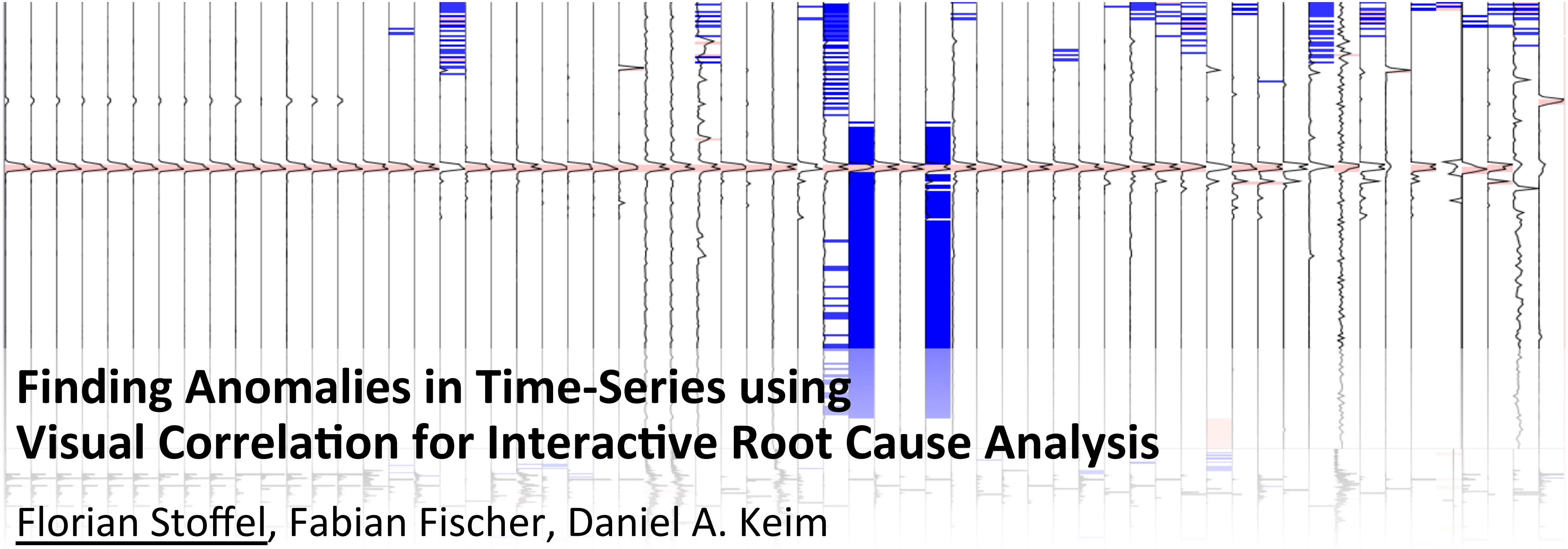
Data Analysis
and Visualization



Finding Anomalies in Time-Series using Visual Correlation for Interactive Root Cause Analysis

Florian Stoffel, Fabian Fischer, Daniel A. Keim

Data Analysis and Visualization Group | University of Konstanz

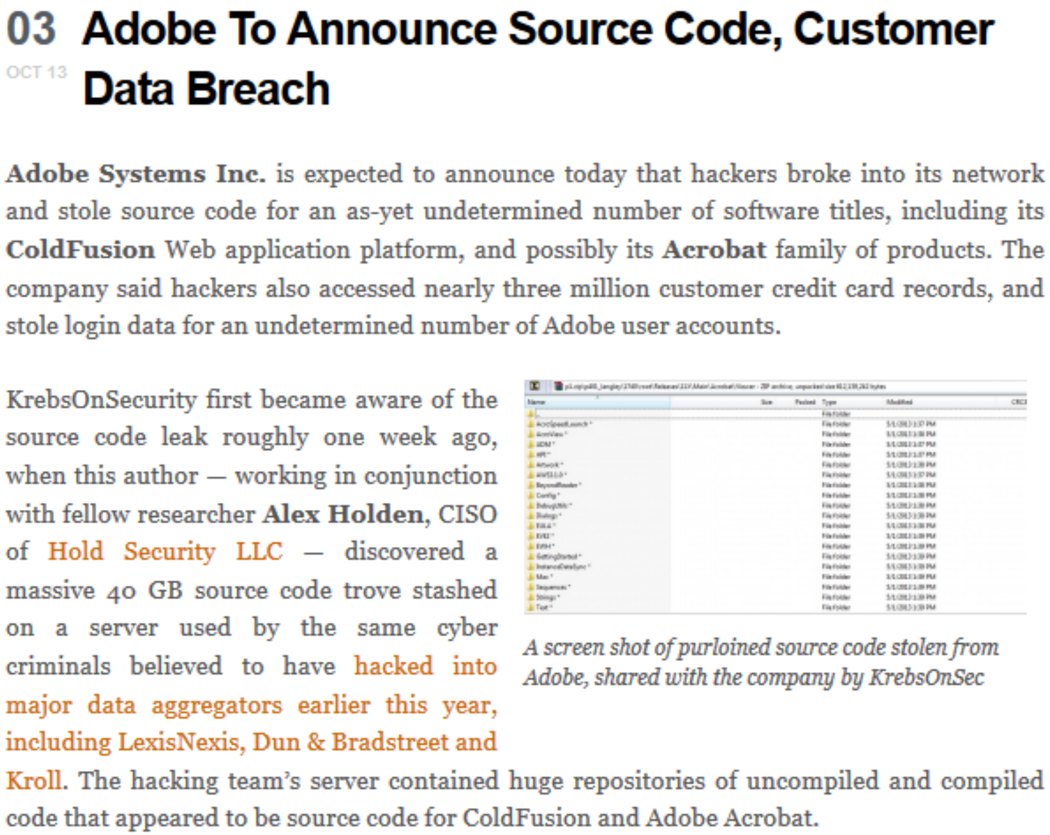
Motivation

- Computer networks are very important for the daily life and today‘s economy.
- Keeping them *secure* and available is crucial.
- Many different threats exist[1]:
    - ➢Cyber Crime
    - ➢Hacktivism
    - ➢Cyber Warfare
    - ➢Cyber Espionage

Source: *PEER1 Hosting/Jeff Johnston*

[1] according to http://hackmageddon.com/

Motivation II



03 OCT 13

Adobe To Announce Source Code, Customer Data Breach

Adobe Systems Inc. is expected to announce today that hackers broke into its network and stole source code for an as-yet undetermined number of software titles, including its **ColdFusion** Web application platform, and possibly its **Acrobat** family of products. The company said hackers also accessed nearly three million customer credit card records, and stole login data for an undetermined number of Adobe user accounts.

KrebsOnSecurity first became aware of the source code leak roughly one week ago, when this author — working in conjunction with fellow researcher **Alex Holden**, CISO of Hold Security LLC — discovered a massive 40 GB source code trove stashed on a server used by the same cyber criminals believed to have hacked into major data aggregators earlier this year, including LexisNexis, Dun & Bradstreet and Kroll. The hacking team's server contained huge repositories of uncompiled and compiled code that appeared to be source code for ColdFusion and Adobe Acrobat.

A screen shot of purloined source code stolen from Adobe, shared with the company by KrebsOnSec

Source: *krebsonsecurity.com*

Adobe Data Leak

- 40GB of source code.
- 2.9 million credit card and password records.

Wanted

- Ability to detect such incidents.
- If an incident is discovered, what services are involved?

Motivation III

Source: *www.jolyon.co.uk*

Approach:

- Incidents like security breaches and data theft should be detectable in network traffic data (**anomalies**).

Challenges:

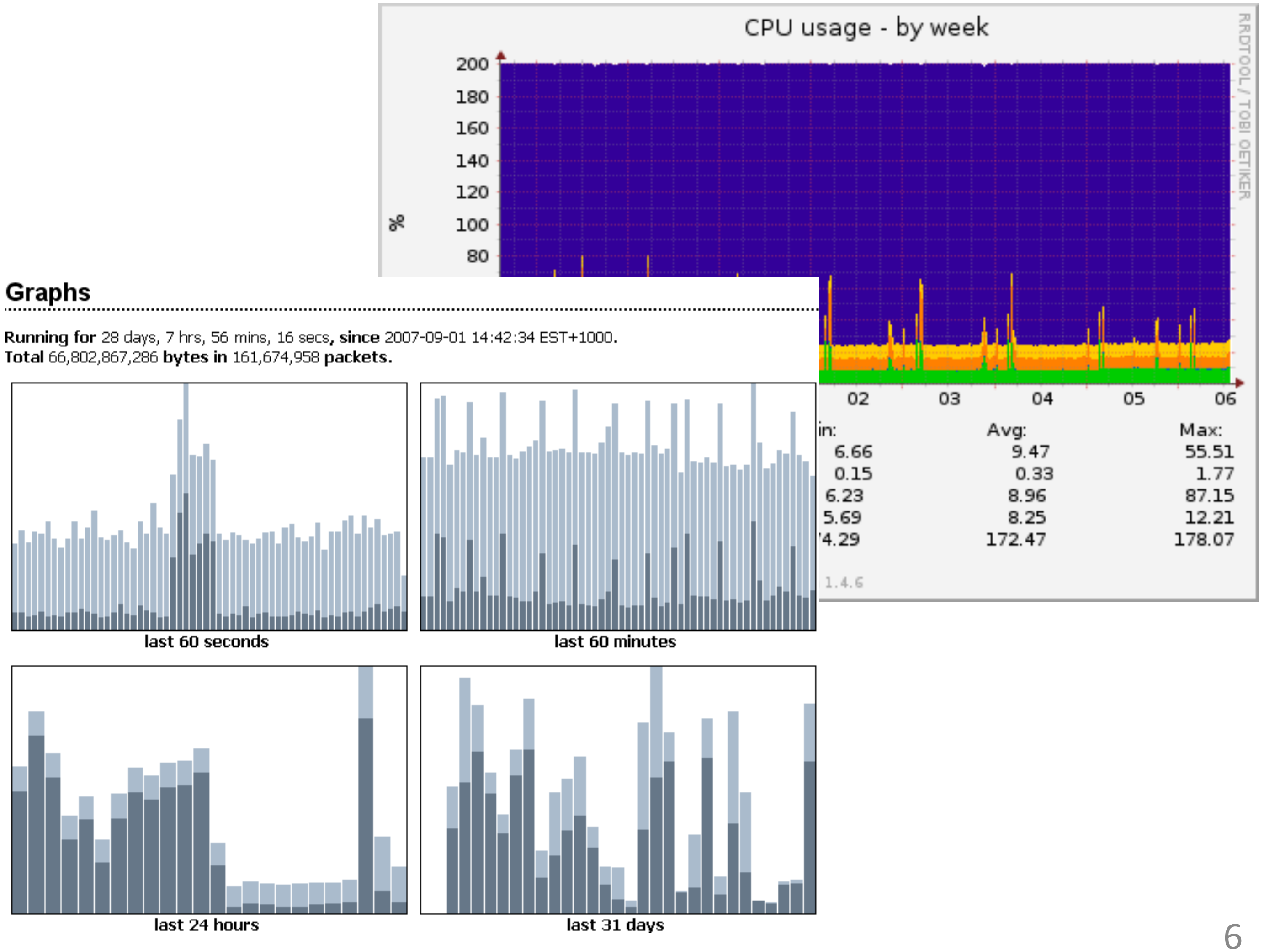
- Large amount of data.
- Define and search for anomalies.
- Scalable visualization.

Outline

- **Related Work**
  Network Monitoring
- **Data**
  Gathering, Processing, Modeling
- **Visualization**
  Design, Properties
- **System**
  Architecture, Usage
- **Conclusion**
  Summary, Future Work

Related Work – Network Monitoring

- Infrastructure and service monitoring.
- Network graphing.
- Manually defined thresholds trigger alarms.
- Usually no modelling or prediction.



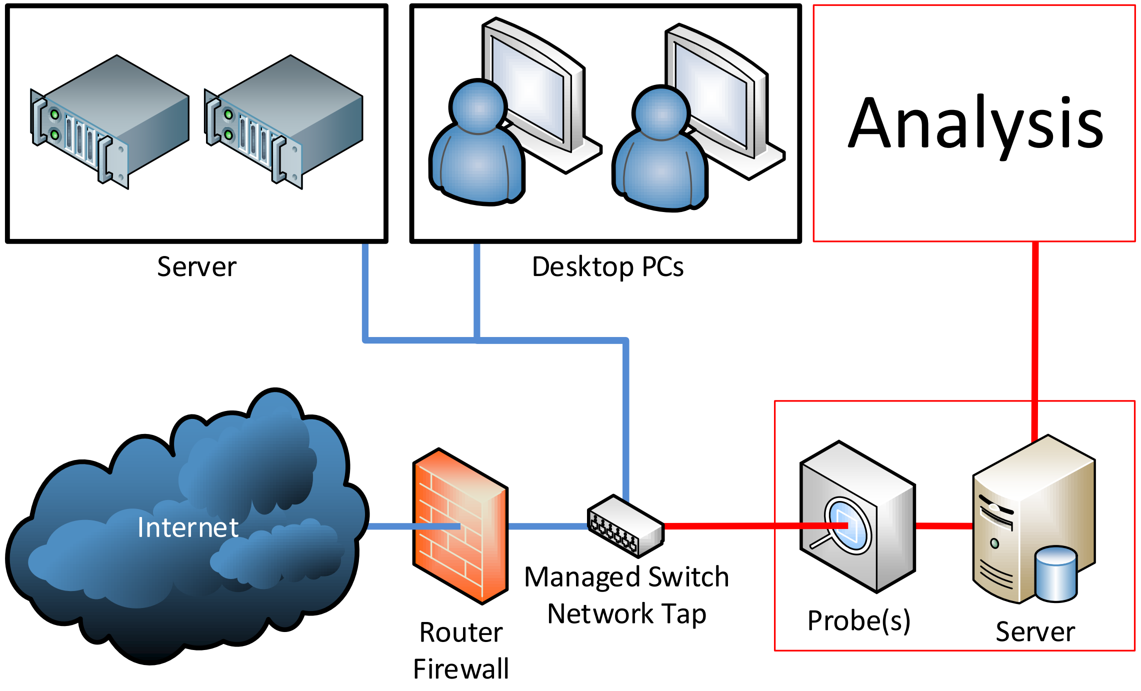
Graphs

Running for 28 days, 7 hrs, 56 mins, 16 secs, since 2007-09-01 14:42:34 EST+1000.
Total 66,802,867,286 bytes in 161,674,958 packets.

Data

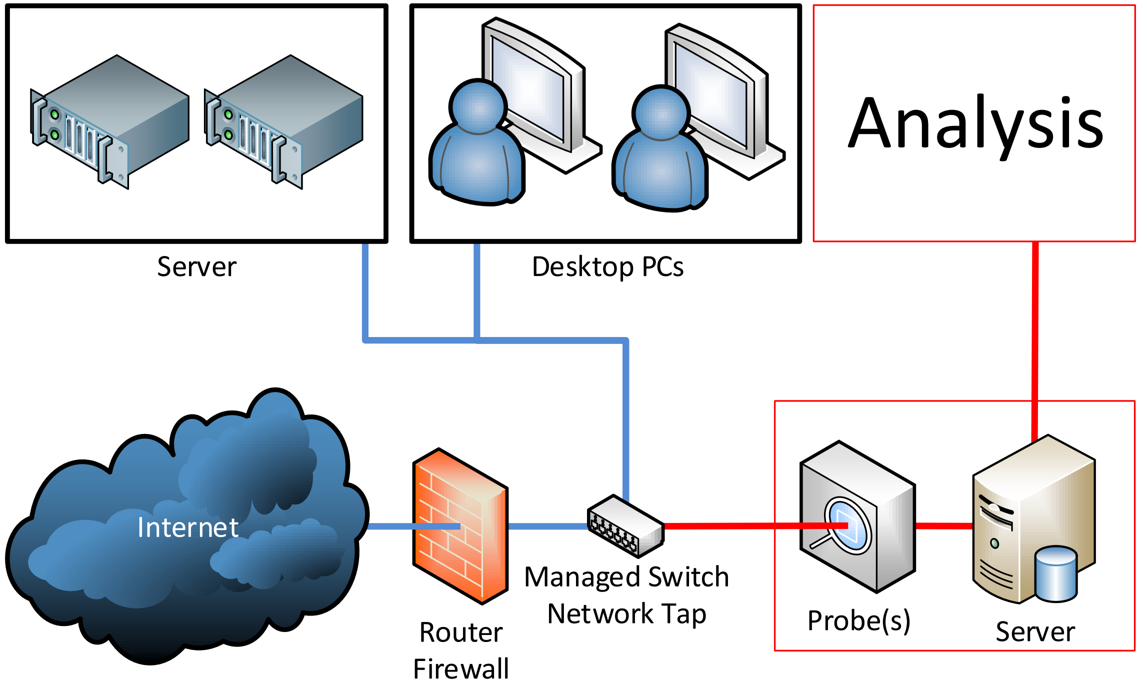
Data

- Data is gathered via *probes* analyzing network flows.
- A probe contains a number of different *descriptors*.
- Descriptors describe a specific property of a protocol, application, …
- If a descriptor matches, a numeric counter is incremented.



Data II

- Network with 30 workstations and 20 regular users.
- 11 month of data.
- Counter transmission interval: 5 minutes.
- Self developed, high performance time-series store.



Data III - Processing

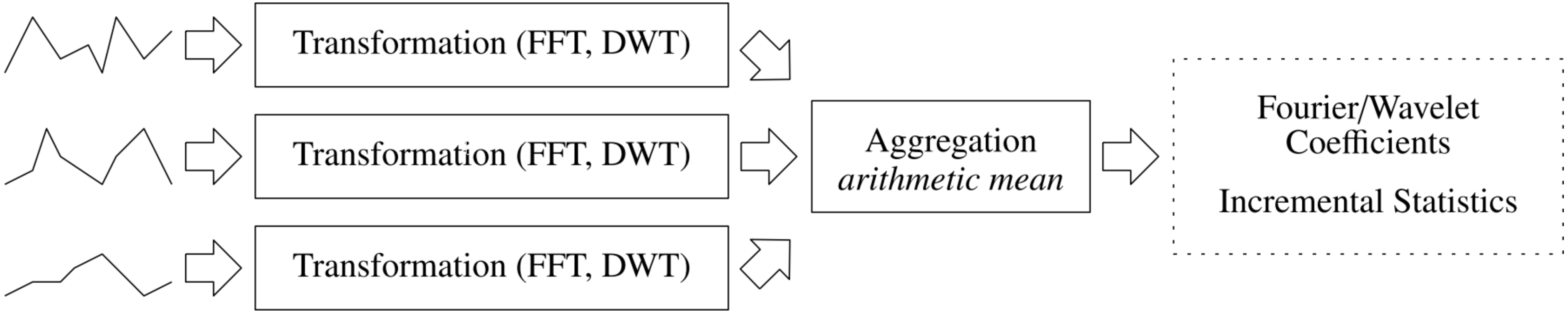
- Data restoration
  - ➢Linear interpolation of missing values.
- Re-sampling with a fixed interval.
  - ➢Continuous time-series.
- Incremental statistics computation.
  - ➢Min, Max, Mean, Variance per day
- Storage with indices over time and time-series.

Leads to: consistent data set with no missing values.

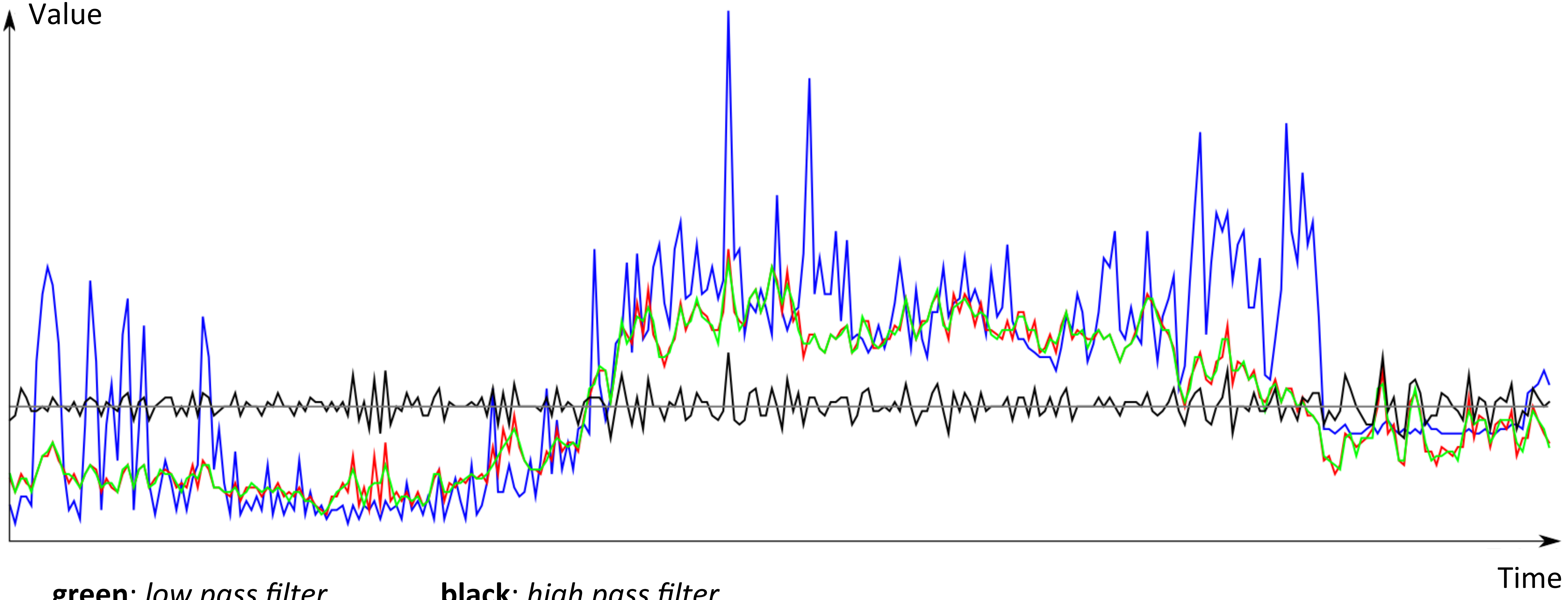
Data IV – Model

Both, Fourier and Wavelet Transformation can be used to decompose time-series data.

- **Fourier Transformation**: frequency domain, but no locality in time.
- **Wavelet Transformation**: locality in time, efficient.



Data IV – Model



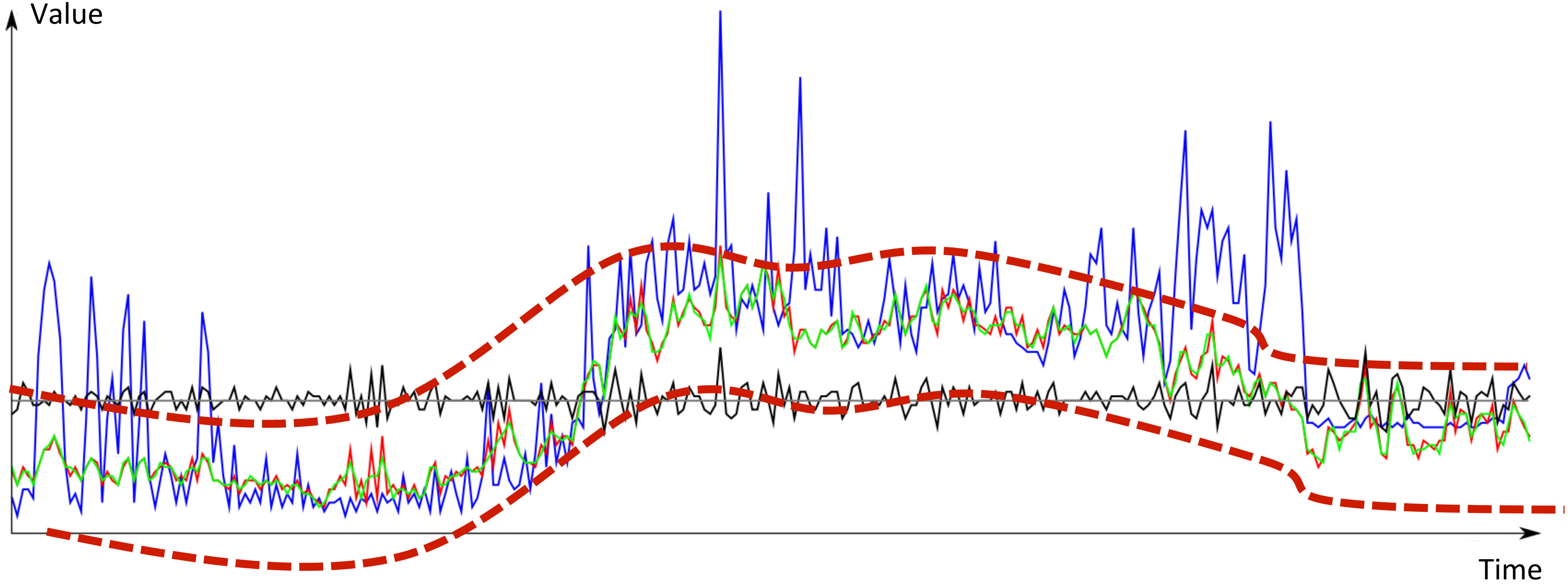
green: *low pass filter*

red: *raw model*

black: *high pass filter*

blue: *example time-series*

Data IV – Model & Anomalies



Define a time-span as ***anomalous***, when it lies outside some borders.

Visualization

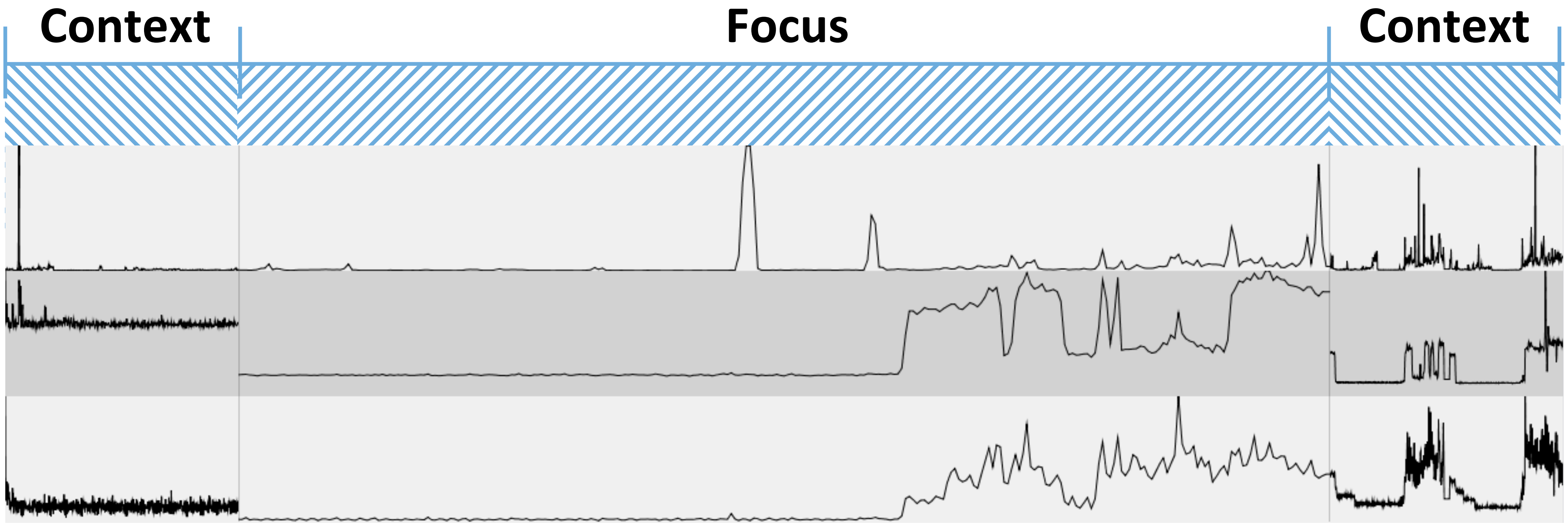
Visualization – Line Chart

- Line charts are widely used for not only time-series displays.
- Easy to interpret, easy to compare, well-known to analysts.

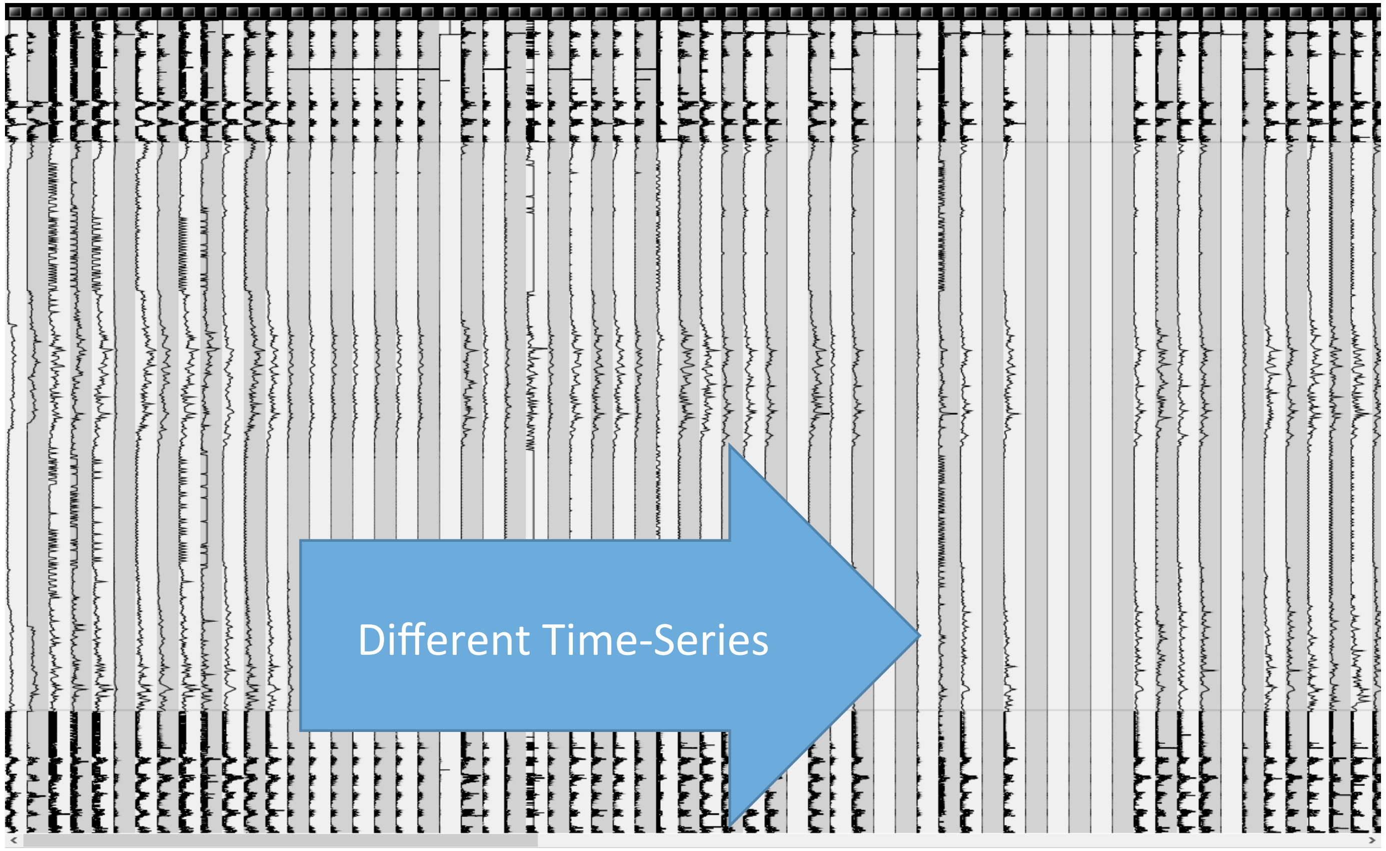
Our Modifications:

- Focus plus Context.
- Rotated view to foster visual, pre-attentive correlation.
- Space efficient representation on a standard PC screen.
- Scalable to any screen size.

Visualization II – Focus + Context



Visualization III – Rotated Line Chart

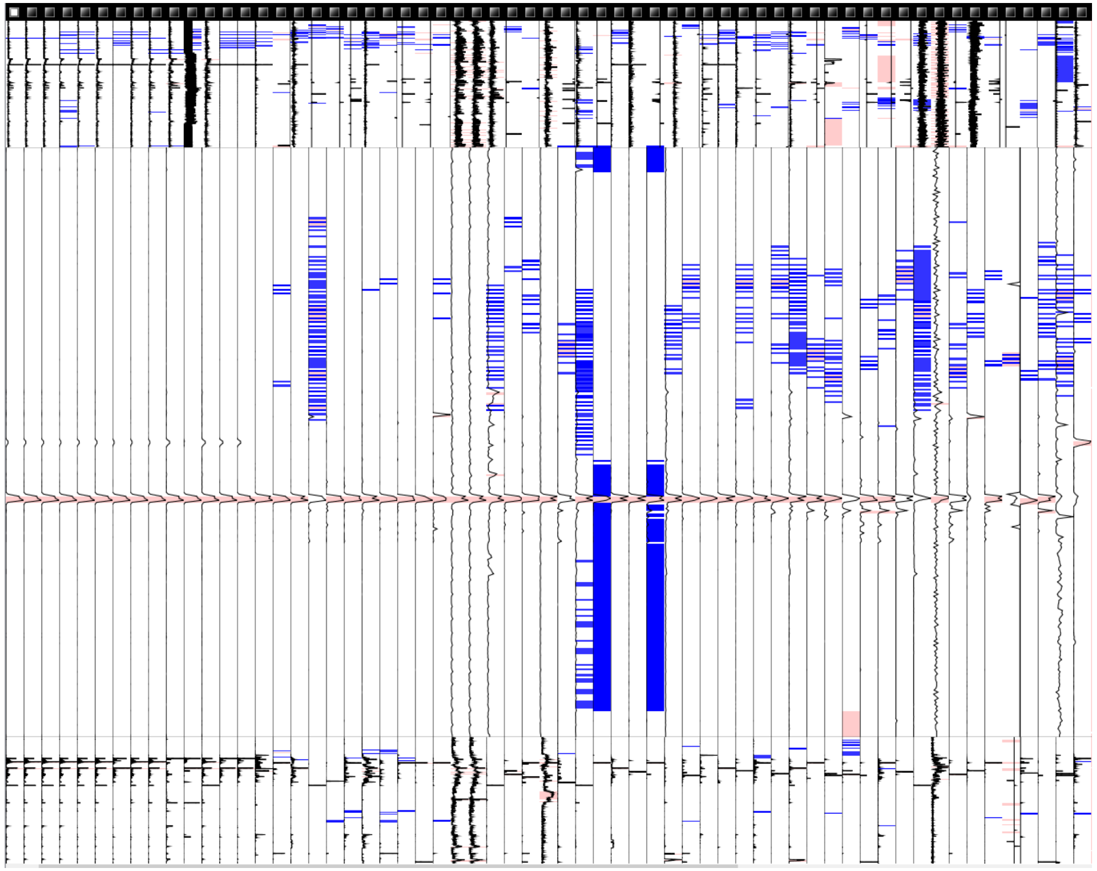


Idea:

- Weaken the intent to go along the time-axis and follow a series.
- Attempt to guide the analyst to comparison based on time-spans.

Visualization IV – Line Chart and Anomalies

- Background of the line charts used for anomaly display.
- Red: value above model and threshold
- Blue: value below model and threshold



Visualization V

- Visual Options

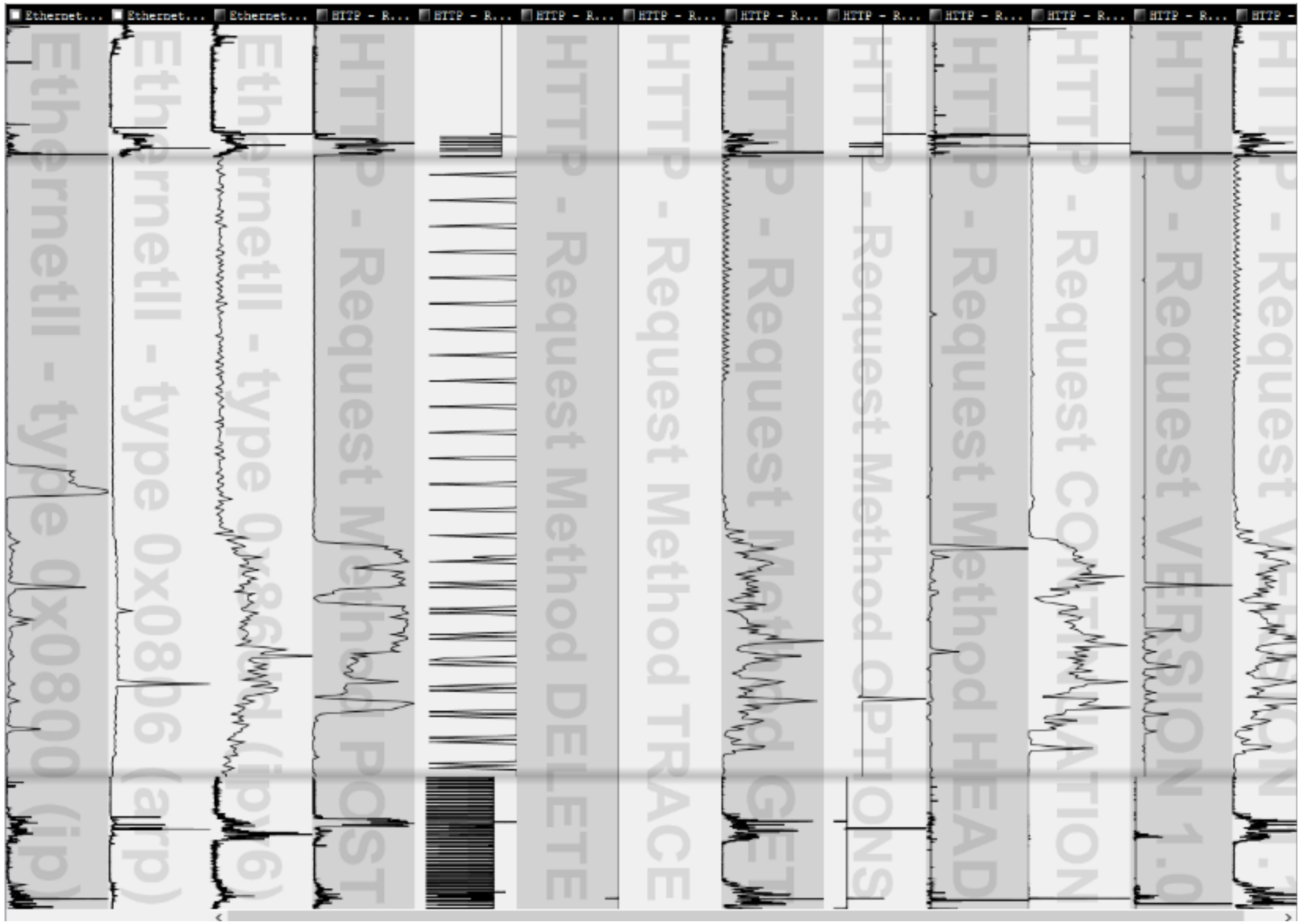
  width, height, colors, focus/ context area size, pinning, in-place comparison, scaling, ordering, …

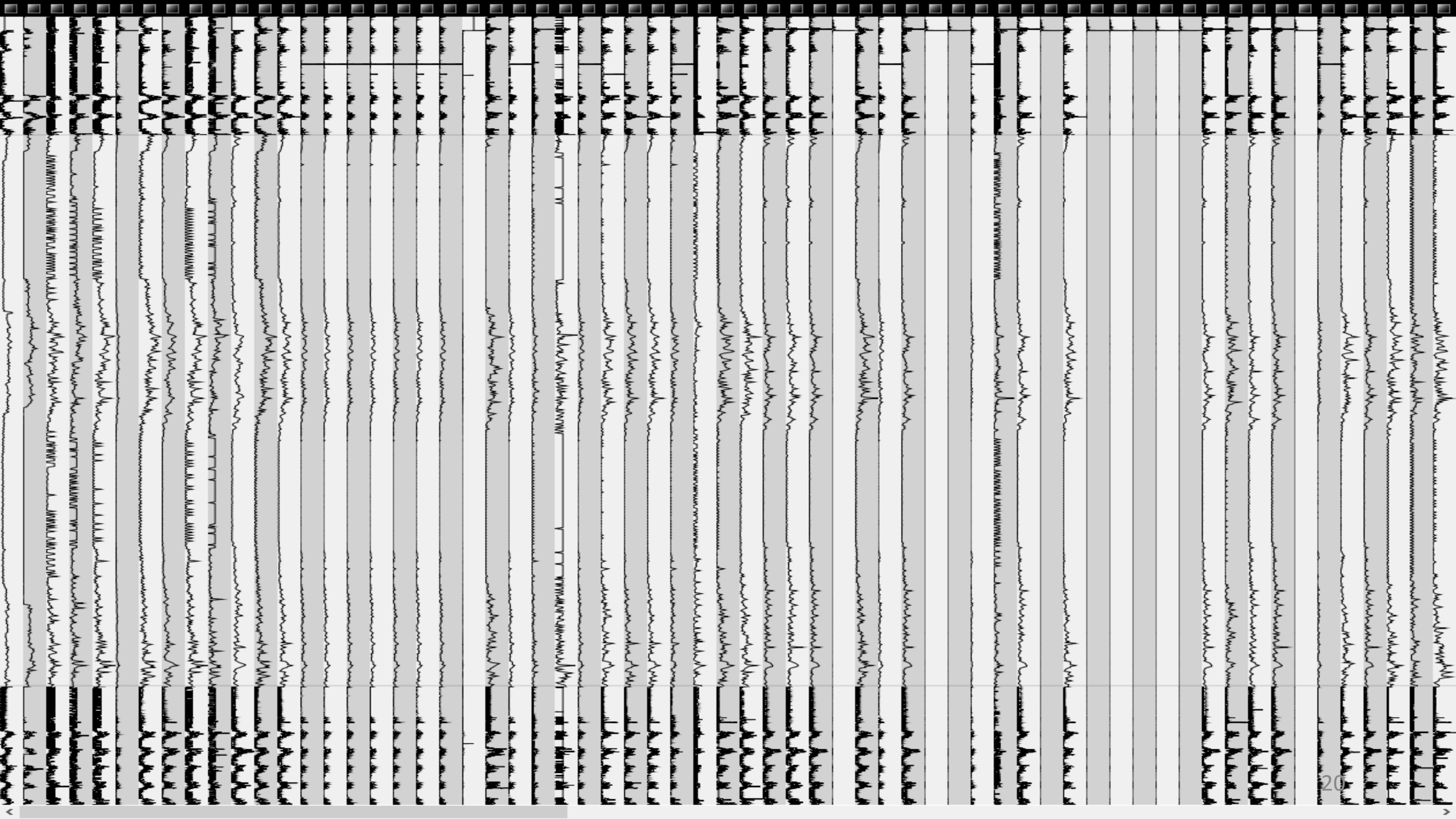
- Interaction

  browsing via mouse-wheel, zooming and querying based on mouse selection

- Synchronization

  location and changes in time visual options


Ethernet...
Ethernet...
Ethernet...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP - R...
HTTP -
Ethernetll - type 0x0800 (ip)
Ethernetll - type 0x0806 (arp)
Ethernetll - type 0x86dd (ipv6)
HTTP - Request Method POST
HTTP - Request Method DELETE
HTTP - Request Method TRACE
HTTP - Request Method GET
HTTP - Request Method OPTIONS
HTTP - Request Method HEAD
HTTP - Request CONTINUATION
HTTP - Request VERSION 1.0
HTTP - Request VERSION 1.1



System

System

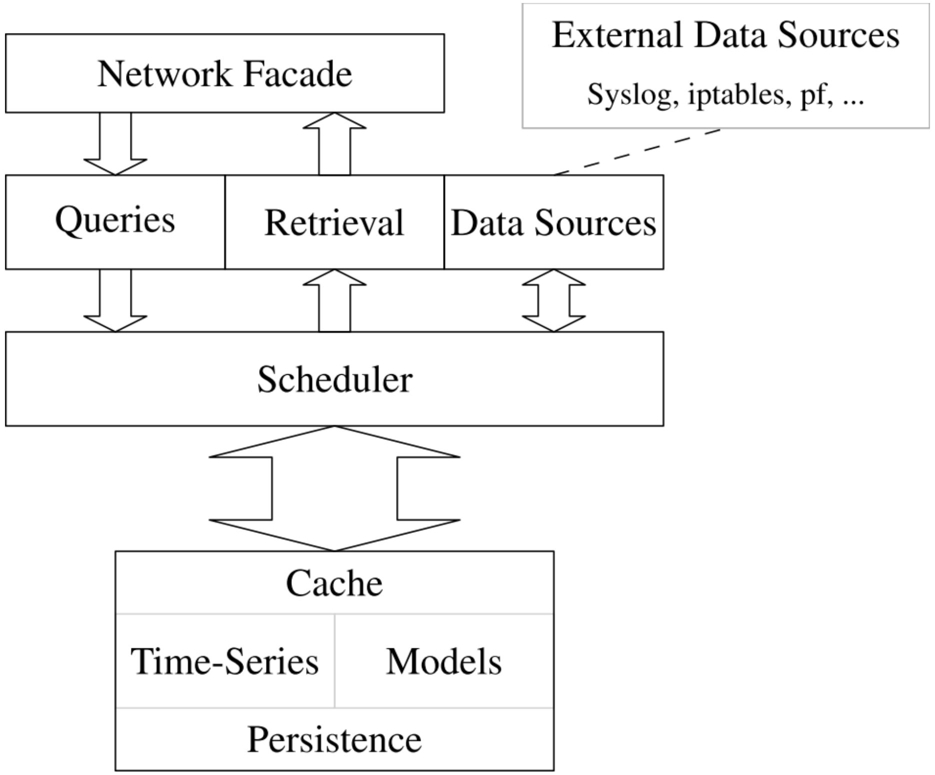
Two main components:

- Server
  - Time-series storage.
  - Analytics backend.
- Client
  - Time-series and visualization framework.
  - Multi-screen and multi-window support.
  - Access to the analytics backend of the server.

Third component:

- Network traffic analysis framework (probes).

System II



External Data Sources

Syslog, iptables, pf, …

Network Facade

Queries

Retrieval

Data Sources

Scheduler

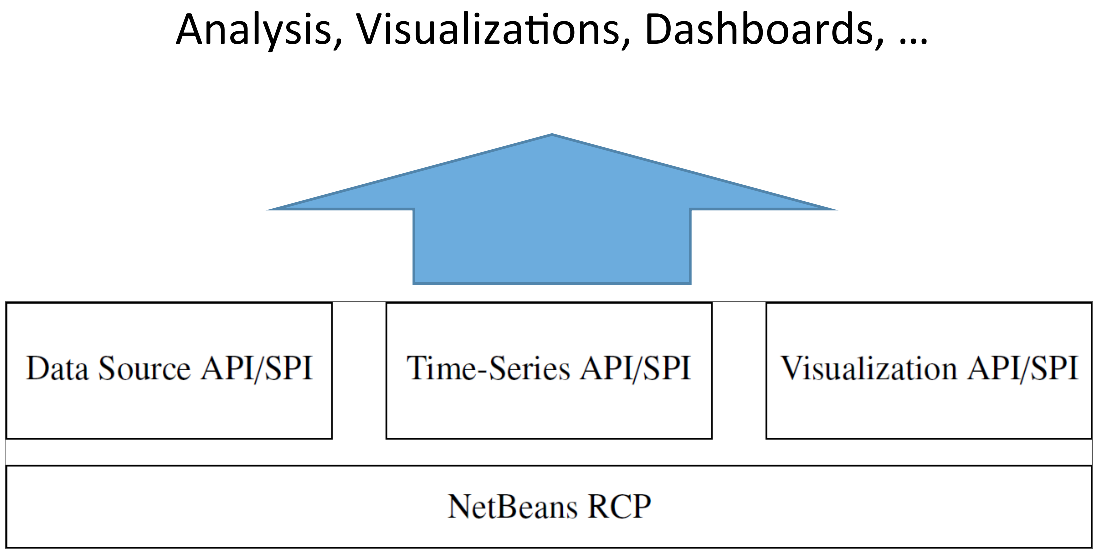
Cache

Time-Series

Models

Persistence

Server Components



Analysis, Visualizations, Dashboards, …

Data Source API/SPI

Time-Series API/SPI

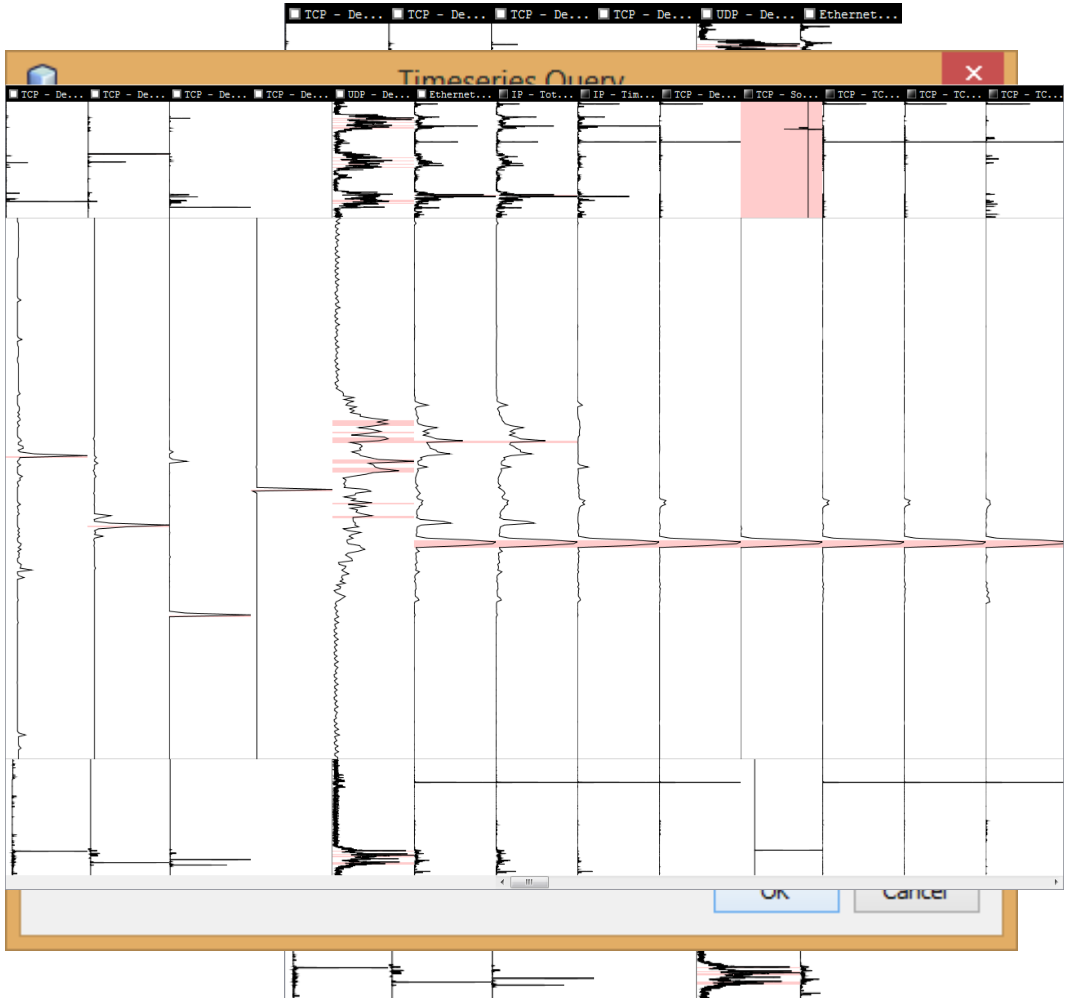
Visualization API/SPI

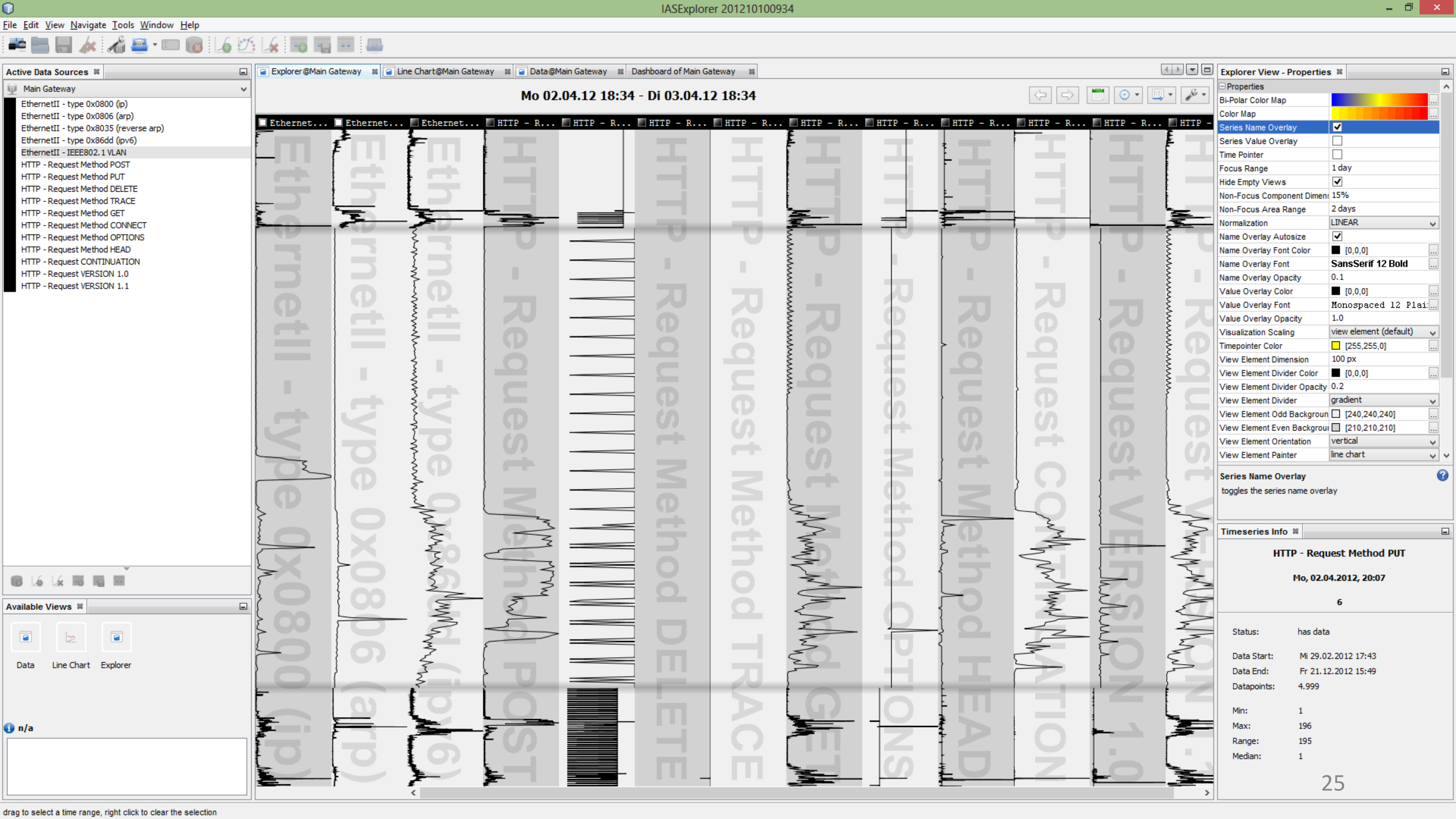
NetBeans RCP

Client Components

System III – Root Cause Analysis

1. Browse through time-series.
2. Find interesting time spans.
3. Formulate a similarity query.
4. Inspect the query results.
5. *(go at 1 if necessary)*
6. Start more detailed analysis in other tools, if necessary.




IASExplorer 201210100934
File Edit View Navigate Tools Window Help
Active Data Sources
Main Gateway
Explorer@Main Gateway
Line Chart@Main Gateway
Data@Main Gateway
Dashboard of Main Gateway
Mo 02.04.12 18:34 - Di 03.04.12 18:34
Ethernet...
HTTP - R...
EthernetII - type 0x0800 (ip)
EthernetII - type 0x0806 (arp)
EthernetII - type 0x86dd (ipv6)
HTTP - Request Method POST
HTTP - Request Method DELETE
HTTP - Request Method TRACE
HTTP - Request Method GET
HTTP - Request Method OPTIONS
HTTP - Request Method HEAD
HTTP - Request CONTINUATION
HTTP - Request VERSION 1.0
HTTP - Request VERSION 1.1
Available Views
Data
Line Chart
Explorer
n/a
Explorer View - Properties
Timeseries Info

EthernetII - type 0x0800 (ip)
EthernetII - type 0x0806 (arp)
EthernetII - type 0x8035 (reverse arp)
EthernetII - type 0x86dd (ipv6)
EthernetII - IEEE802.1 VLAN
HTTP - Request Method POST
HTTP - Request Method PUT
HTTP - Request Method DELETE
HTTP - Request Method TRACE
HTTP - Request Method GET
HTTP - Request Method CONNECT
HTTP - Request Method OPTIONS
HTTP - Request Method HEAD
HTTP - Request CONTINUATION
HTTP - Request VERSION 1.0
HTTP - Request VERSION 1.1

Properties

Bi-Polar Color Map	
Color Map	
Series Name Overlay	☑
Series Value Overlay	□
Time Pointer	□
Focus Range	1 day
Hide Empty Views	☑
Non-Focus Component Dimen:	15%
Non-Focus Area Range	2 days
Normalization	LINEAR
Name Overlay Autosize	☑
Name Overlay Font Color	[0,0,0]
Name Overlay Font	SansSerif 12 Bold
Name Overlay Opacity	0.1
Value Overlay Color	[0,0,0]
Value Overlay Font	Monospaced 12 Plai:
Value Overlay Opacity	1.0
Visualization Scaling	view element (default)
Timepointer Color	[255,255,0]
View Element Dimension	100 px
View Element Divider Color	[0,0,0]
View Element Divider Opacity	0.2
View Element Divider	gradient
View Element Odd Backgroun	[240,240,240]
View Element Even Backgrou	[210,210,210]
View Element Orientation	vertical
View Element Painter	line chart

Series Name Overlay

toggles the series name overlay

HTTP - Request Method PUT

Mo, 02.04.2012, 20:07

6

Status:	has data
Data Start:	Mi 29.02.2012 17:43
Data End:	Fr 21.12.2012 15:49
Datapoints:	4.999
Min:	1
Max:	196
Range:	195
Median:	1

drag to select a time range, right click to clear the selection

Conclusion

Framework for time-series processing, containing:

- Time-series modelling.
- Scalable time-series analysis.
- Scalable time-series visualization.

Future Work:

- Evaluate the line charts (rotation).
- Show the usefulness in a real world use-case with interesting data.
  - WIP: DNA Sequence Alignment Browser/Exploration.
- More Automatic methods.

Thank you very
much for your attention!

Questions?

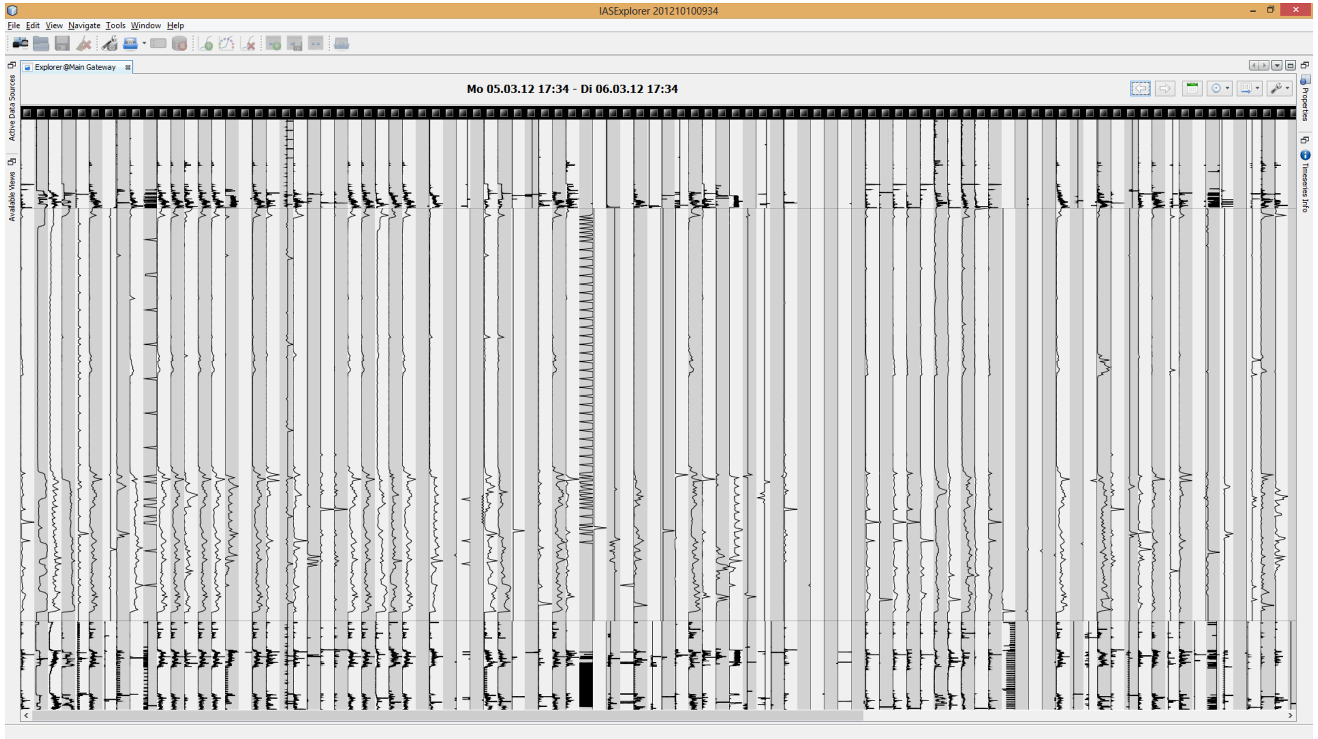
For more information
about our work
please contact

Florian Stoffel

Tel. +49 7531 88-3070
florian.stoffel@uni-konstanz.de

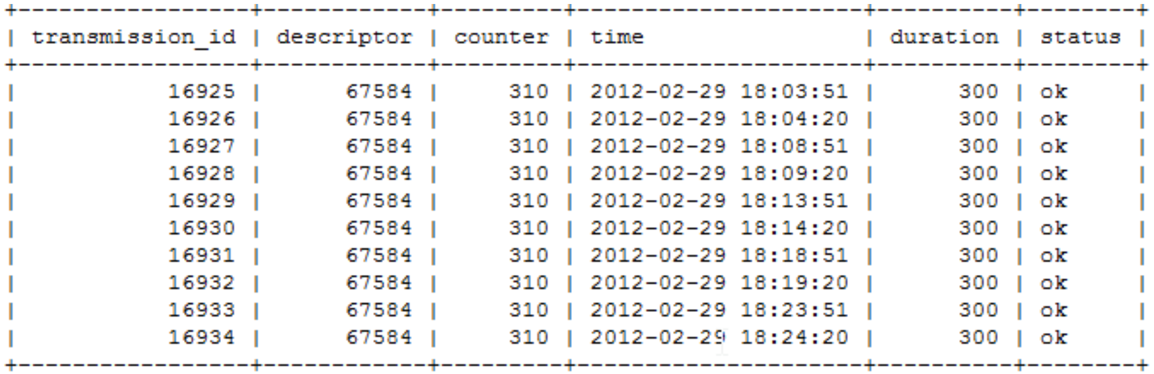
@florianstoffel

http://vis.uni-konstanz.de/



Dataset Details

- *transmission id*
  - ➢id of the data transmission
- *descriptor id*
  - ➢id of the descriptor
- *Counter*
  - ➢counter reading of the descriptor
- *transmission time*
  - ➢time of insertion
- Duration
  - ➢data gathering interval
- Status
  - ➢transmission status



```
+-----------------+------------+---------+---------------------+----------+--------+
| transmission_id | descriptor | counter | time                | duration | status |
+-----------------+------------+---------+---------------------+----------+--------+
|           16925 |      67584 |     310 | 2012-02-29 18:03:51 |      300 | ok     |
|           16926 |      67584 |     310 | 2012-02-29 18:04:20 |      300 | ok     |
|           16927 |      67584 |     310 | 2012-02-29 18:08:51 |      300 | ok     |
|           16928 |      67584 |     310 | 2012-02-29 18:09:20 |      300 | ok     |
|           16929 |      67584 |     310 | 2012-02-29 18:13:51 |      300 | ok     |
|           16930 |      67584 |     310 | 2012-02-29 18:14:20 |      300 | ok     |
|           16931 |      67584 |     310 | 2012-02-29 18:18:51 |      300 | ok     |
|           16932 |      67584 |     310 | 2012-02-29 18:19:20 |      300 | ok     |
|           16933 |      67584 |     310 | 2012-02-29 18:23:51 |      300 | ok     |
|           16934 |      67584 |     310 | 2012-02-29 18:24:20 |      300 | ok     |
+-----------------+------------+---------+---------------------+----------+--------+
```