DATA ANALYTICS

CERN openlab

Stefano Alberto Russo
IT-DB-DBF
Stefano.alberto.russo@cern.ch

Purpose of the project

Why data analytics at CERN?

- Lot of monitoring data stored in the past years
  - this data contains a lot of information:
    - we want to understand how to extract it

This is the purpose of the Openlab data analytics project

…to obtain added value from not so actively used data

Topics:

- The R project for statistical computing
- The R Studio IDE
- Oracle R Enterprise (ORE)
- The Openlab working layout
- The CASTOR use case
  - Data sources and management
- Live demo
- Conclusions

About me:

- Involved in the CASTOR monitoring
  - Developed real time Cockpit
  - Set up Hadoop cluster to store raw log data on Hadoop (10+ TB)
- Now working on data analytics in Openlab
- Found limitations of NoSQL for advanced data analysis
- Started investigating other solutions:
  - home made stuff, R and Oracle R Enterprise

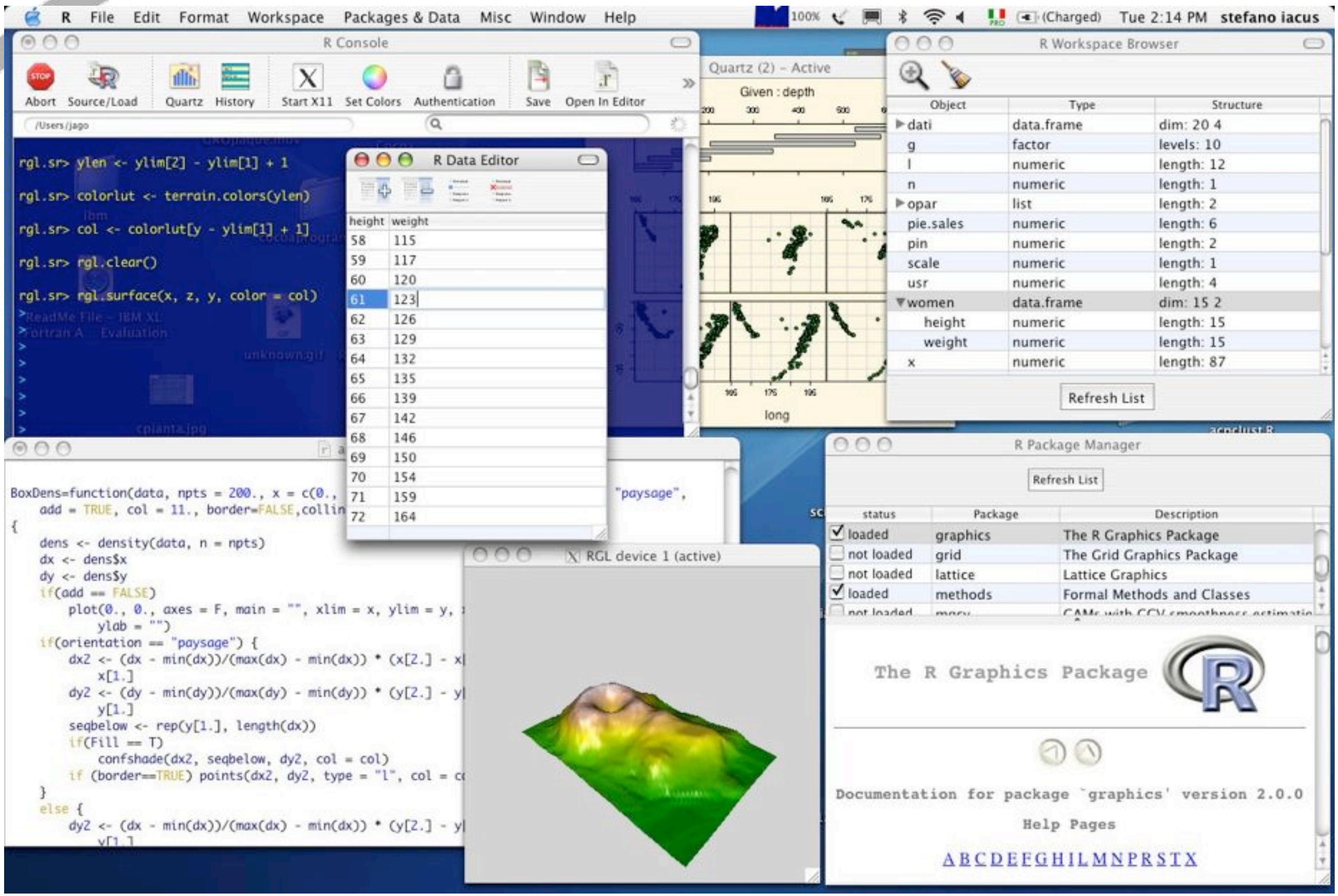
The R Project

The R Project for Statistical Computing

- Software environment for statistical computing and graphics
- Free and open source
- Standard and advanced statistical techniques
  - linear and nonlinear modeling
  - classical statistical tests
  - time-series analysis
  - classification, clustering
  - machine learning (neural networks, SVM, …)
- Highly extensible

…gives a meaning to your data

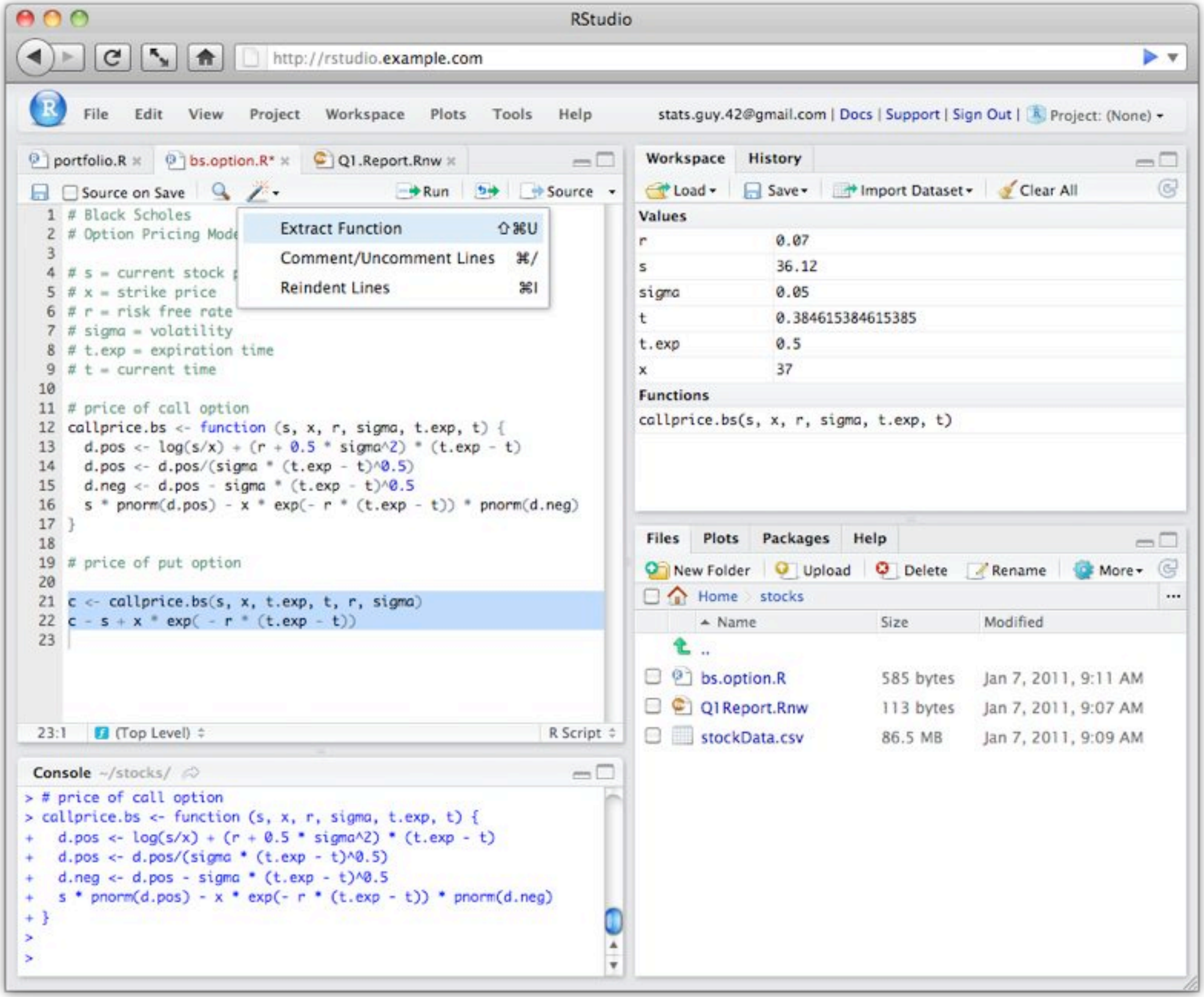
The R Project



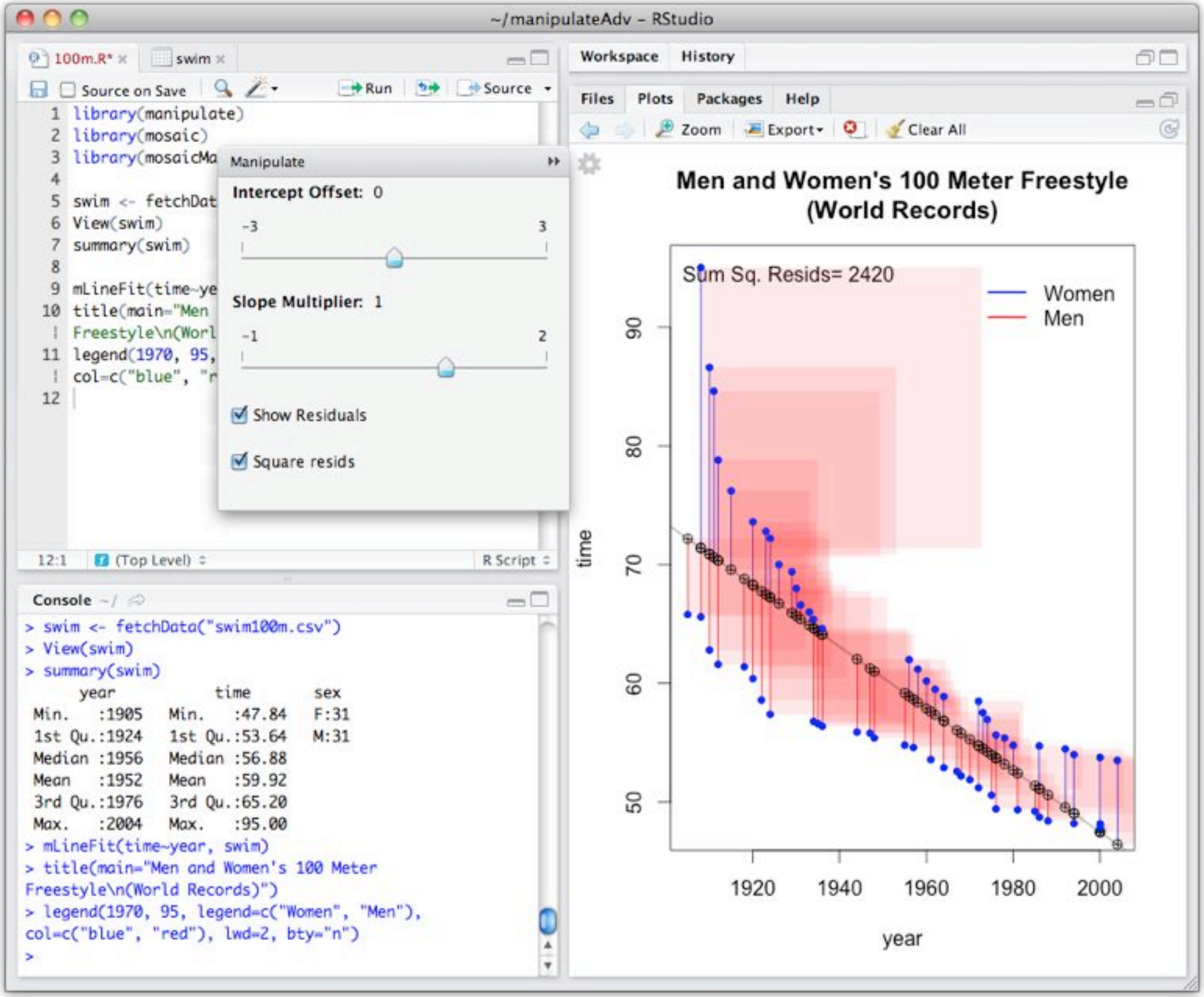
R Studio IDE: overview

- Free and open source **IDE** for R
- “Take control of your R code”
- Windows, Mac Linux
- Web (RStudio server)

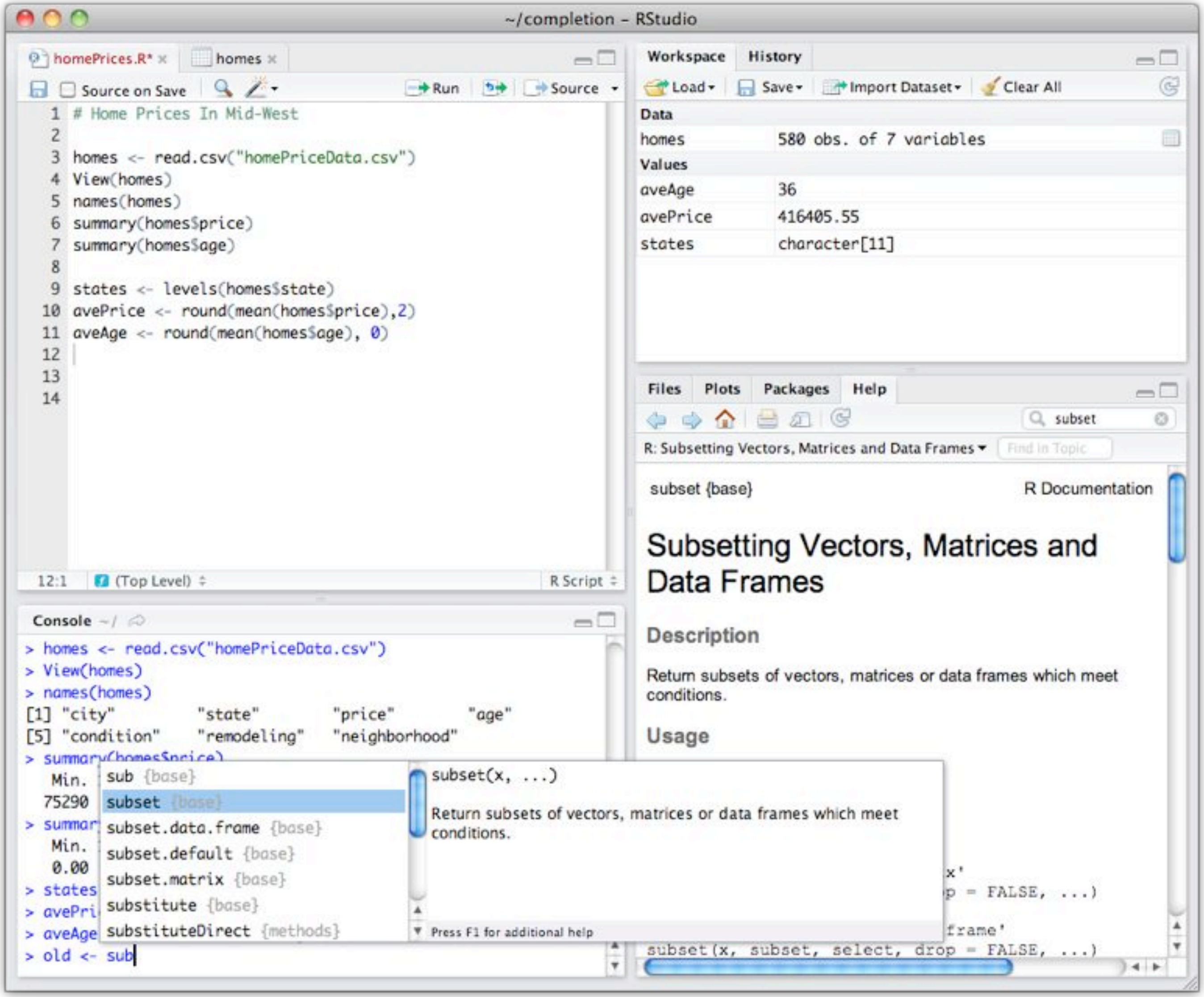
R Studio


RStudio
http://rstudio.example.com
File Edit View Project Workspace Plots Tools Help
stats.guy.42@gmail.com | Docs | Support | Sign Out | Project: (None)
portfolio.R bs.option.R* Q1.Report.Rnw
Source on Save
Run
Source
Extract Function ⇧⌘U
Comment/Uncomment Lines ⌘/
Reindent Lines ⌘I
# Black Scholes
# Option Pricing Mode
# s = current stock p
# x = strike price
# r = risk free rate
# sigma = volatility
# t.exp = expiration time
# t = current time
# price of call option
callprice.bs <- function (s, x, r, sigma, t.exp, t) {
d.pos <- log(s/x) + (r + 0.5 * sigma^2) * (t.exp - t)
d.pos <- d.pos/(sigma * (t.exp - t)^0.5)
d.neg <- d.pos - sigma * (t.exp - t)^0.5
s * pnorm(d.pos) - x * exp(- r * (t.exp - t)) * pnorm(d.neg)
}
# price of put option
c <- callprice.bs(s, x, t.exp, t, r, sigma)
c - s + x * exp(- r * (t.exp - t))
23:1 (Top Level) R Script
Workspace History
Load Save Import Dataset Clear All
Values
r 0.07
s 36.12
sigma 0.05
t 0.384615384615385
t.exp 0.5
x 37
Functions
callprice.bs(s, x, r, sigma, t.exp, t)
Files Plots Packages Help
New Folder Upload Delete Rename More
Home stocks
Name Size Modified
bs.option.R 585 bytes Jan 7, 2011, 9:11 AM
Q1Report.Rnw 113 bytes Jan 7, 2011, 9:07 AM
stockData.csv 86.5 MB Jan 7, 2011, 9:09 AM
Console ~/stocks/
> # price of call option
> callprice.bs <- function (s, x, r, sigma, t.exp, t) {
+ d.pos <- log(s/x) + (r + 0.5 * sigma^2) * (t.exp - t)
+ d.pos <- d.pos/(sigma * (t.exp - t)^0.5)
+ d.neg <- d.pos - sigma * (t.exp - t)^0.5
+ s * pnorm(d.pos) - x * exp(- r * (t.exp - t)) * pnorm(d.neg)
+ }

R Studio


~/manipulateAdv – RStudio
100m.R*
swim
Source on Save
Run
Source
library(manipulate)
library(mosaic)
library(mosaicMa
swim <- fetchDat
View(swim)
summary(swim)
mLineFit(time~ye
title(main="Men
Freestyle\n(Worl
legend(1970, 95,
col=c("blue", "r
Manipulate
Intercept Offset: 0
Slope Multiplier: 1
Show Residuals
Square resids
12:1
(Top Level)
R Script
Console ~/
> swim <- fetchData("swim100m.csv")
> View(swim)
> summary(swim)
year time sex
Min. :1905 Min. :47.84 F:31
1st Qu.:1924 1st Qu.:53.64 M:31
Median :1956 Median :56.88
Mean :1952 Mean :59.92
3rd Qu.:1976 3rd Qu.:65.20
Max. :2004 Max. :95.00
> mLineFit(time~year, swim)
> title(main="Men and Women's 100 Meter Freestyle\n(World Records)")
> legend(1970, 95, legend=c("Women", "Men"), col=c("blue", "red"), lwd=2, bty="n")
Workspace
History
Files
Plots
Packages
Help
Zoom
Export
Clear All
Men and Women's 100 Meter Freestyle
(World Records)
Sum Sq. Resids= 2420
Women
Men
time
year

R Studio


~/completion - RStudio
homePrices.R*
homes
Source on Save
Run
Source
# Home Prices In Mid-West
homes <- read.csv("homePriceData.csv")
View(homes)
names(homes)
summary(homes$price)
summary(homes$age)
states <- levels(homes$state)
avePrice <- round(mean(homes$price),2)
aveAge <- round(mean(homes$age), 0)
12:1
(Top Level)
R Script
Workspace
History
Load
Save
Import Dataset
Clear All
Data
homes
580 obs. of 7 variables
Values
aveAge
36
avePrice
416405.55
states
character[11]
Files
Plots
Packages
Help
subset
R: Subsetting Vectors, Matrices and Data Frames
Find in Topic
subset {base}
R Documentation
Subsetting Vectors, Matrices and Data Frames
Description
Return subsets of vectors, matrices or data frames which meet conditions.
Usage
Console ~/
> homes <- read.csv("homePriceData.csv")
> View(homes)
> names(homes)
[1] "city" "state" "price" "age"
[5] "condition" "remodeling" "neighborhood"
sub {base}
subset {base}
subset.data.frame {base}
subset.default {base}
subset.matrix {base}
substitute {base}
substituteDirect {methods}
subset(x, ...)
Return subsets of vectors, matrices or data frames which meet conditions.
Press F1 for additional help
> old <- sub

Oracle Database Overview

- Widely used
- At CERN since the first releases
- Scalable
- Many advanced features
  - Data consistency
  - Concurrent user support
  - PL/SQL
  - Parallel query system
  - Analytical functions
  - …

Oracle R Enterprise

Oracle **R** **E**nterprise (**ORE**) overview

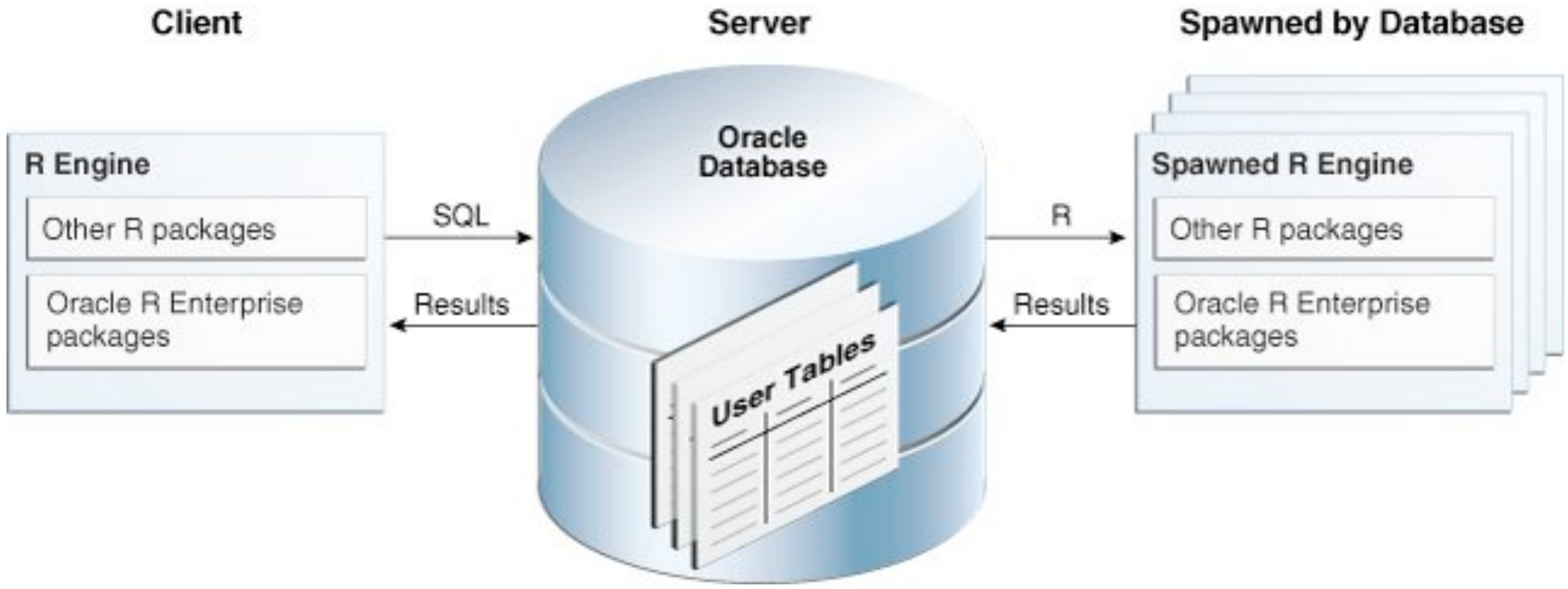
Leader in data management + Leader in data analysis

- R integrated in Oracle core product
  - R operates on data stored in the Oracle Database
- parallelism and scalability of the database
- automate data analysis

Features

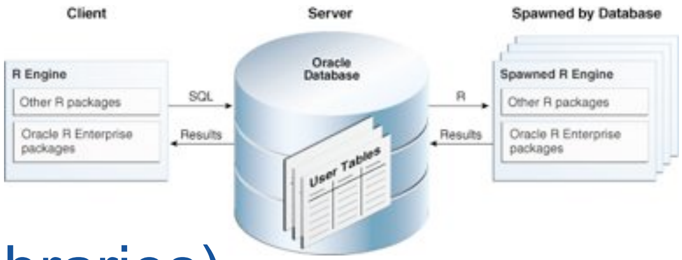
- **Transparency Layer:** allows R code and R packages run almost out of the box
  - bottom line: **intercepts R functions**
- Data management and organization in a DB much easier than CSV style
- R on a database approach already possible but:
  - ORE allows **in-database data analysis**
- Scalable and Big Data ready

Layout


Client
Server
Spawned by Database
R Engine
Other R packages
Oracle R Enterprise packages
SQL
Results
Oracle Database
User Tables
R
Results
Spawned R Engine
Other R packages
Oracle R Enterprise packages

Oracle R Enterprise

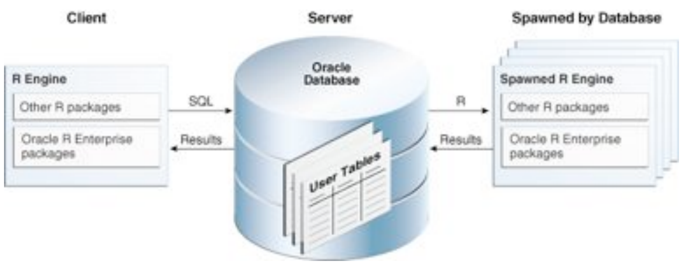
Client



- Collection of **R packages** (libraries)
  - Allows **connect to an Oracle Database** and to interact with it
- **Any R command** usable
- Provides a set of **highly optimized**, ORE specific **functions**
  - corr() -> ore.cor()
  - lm() -> ore.lm()
- Functions intercept data transforms, statistical functions, and Oracle R Enterprise-specific functions

Oracle R Enterprise

Server



- collection of **PL/SQL procedures and libraries**
  - augment Oracle Database with the capabilities required to support an Oracle R Enterprise client
  - **SQL query parallel** execution
- embedded R execution
  - Oracle Database spawns **R engines**: **data parallelism**
- Access to **tables**, **views**, and external tables in the database

Why ORE?

- Advanced data analysis requires **advanced statistical and machine learning models**

  …which are ready to use in R
- If you have both **many variables to evaluate** and **large amounts of data:**
  - Management is complex with CSV-style data store
  - NoSQL solutions cannot be used for correlation analysis
  - standard R in-memory approach can fail

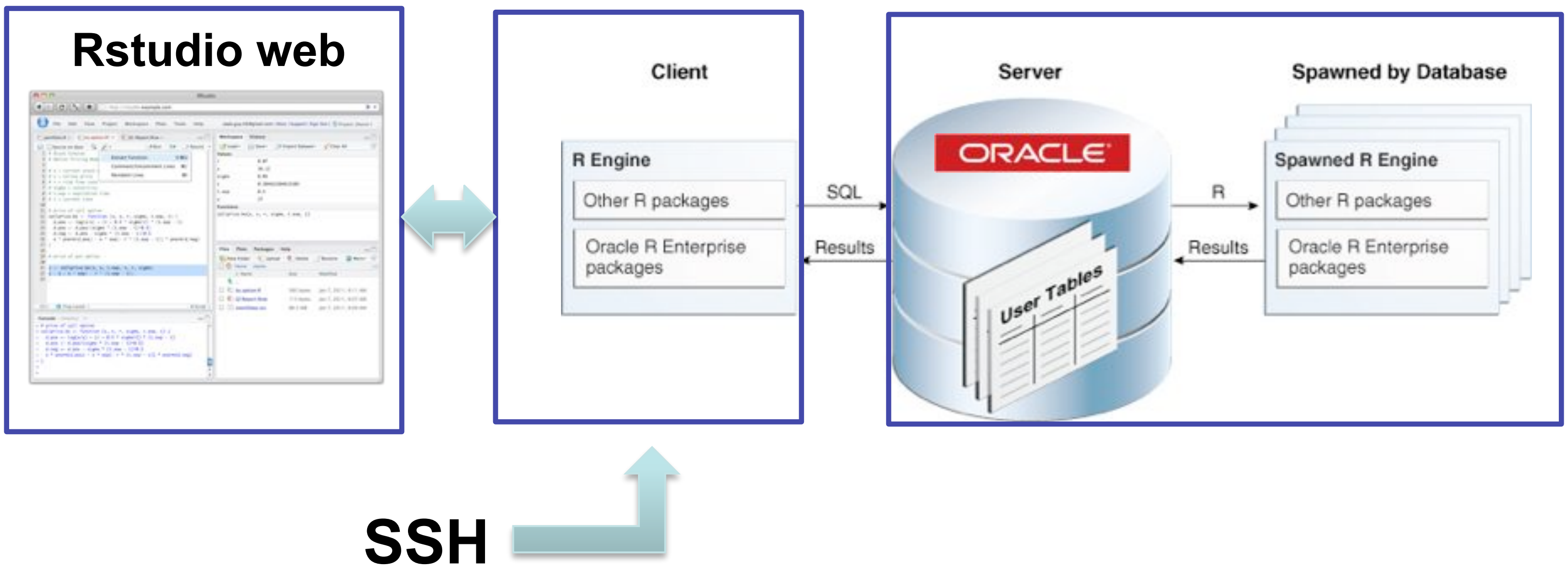
  (ORE let data be analysed inside the DB)

The Openlab working layout

Installation-free Workstation/Laptop

Virtual Machine

Server(s) with Oracle Database



Modular, Flexible, user-friendly

The CASTOR use case

A test case:

the CERN **A**dvanced **STOR**age Manager

CASTOR

CERN Advanced STORage manager

The CASTOR use case

CASTOR use case overview

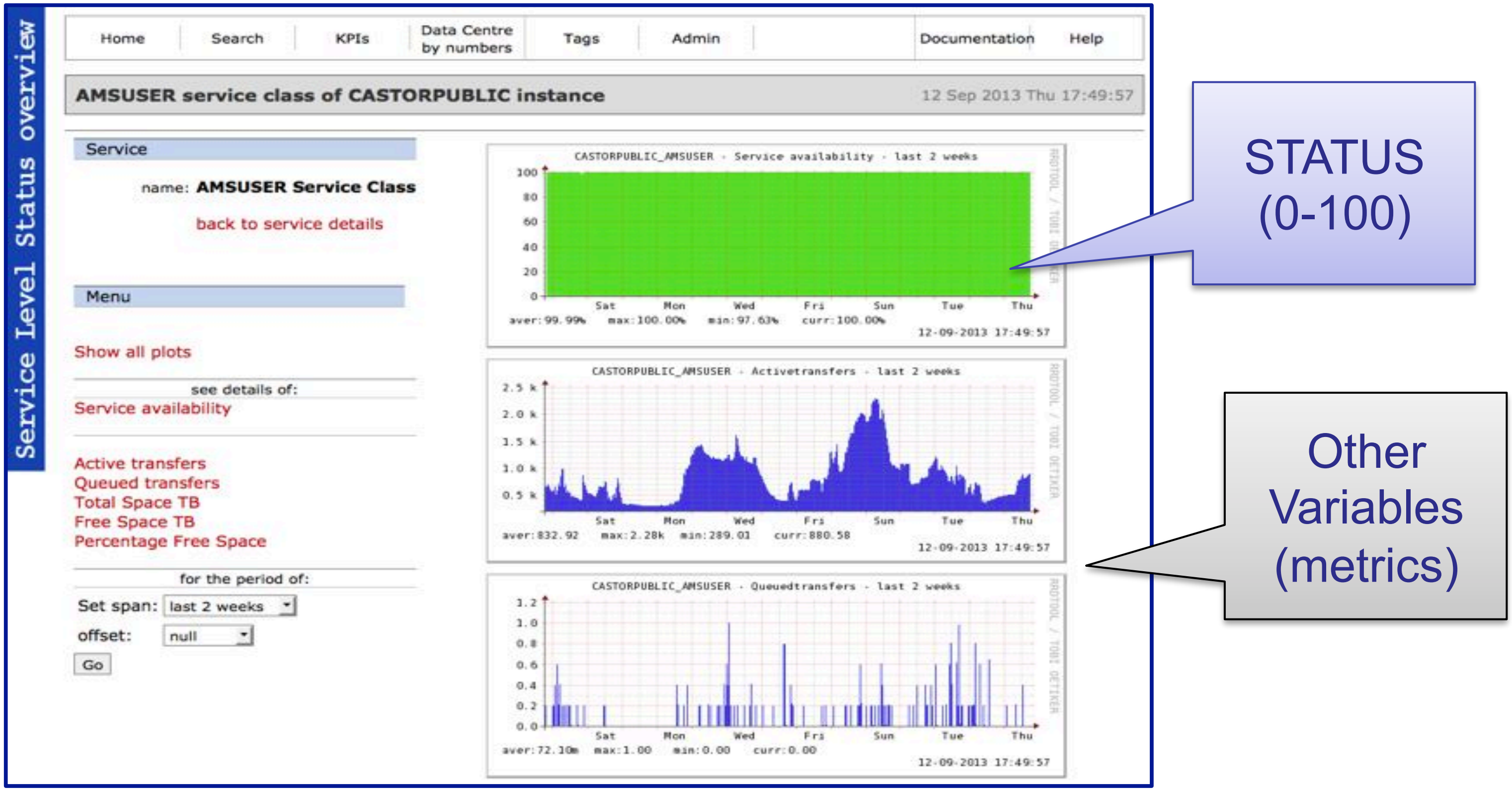
- Lot of (complex) **log data** recorded in the past years from **various systems**
  - (mainly, ***time series***)
- **CASTOR TEAM:** *Can we obtain useful information from it?*

Understanding if and how the **anomalies** in the system are correlated with the **adverse events** would have *great value*

– Performance
– Cause of errors
– Anomaly detection
– Predictions
– Early warning systems

The CASTOR use case

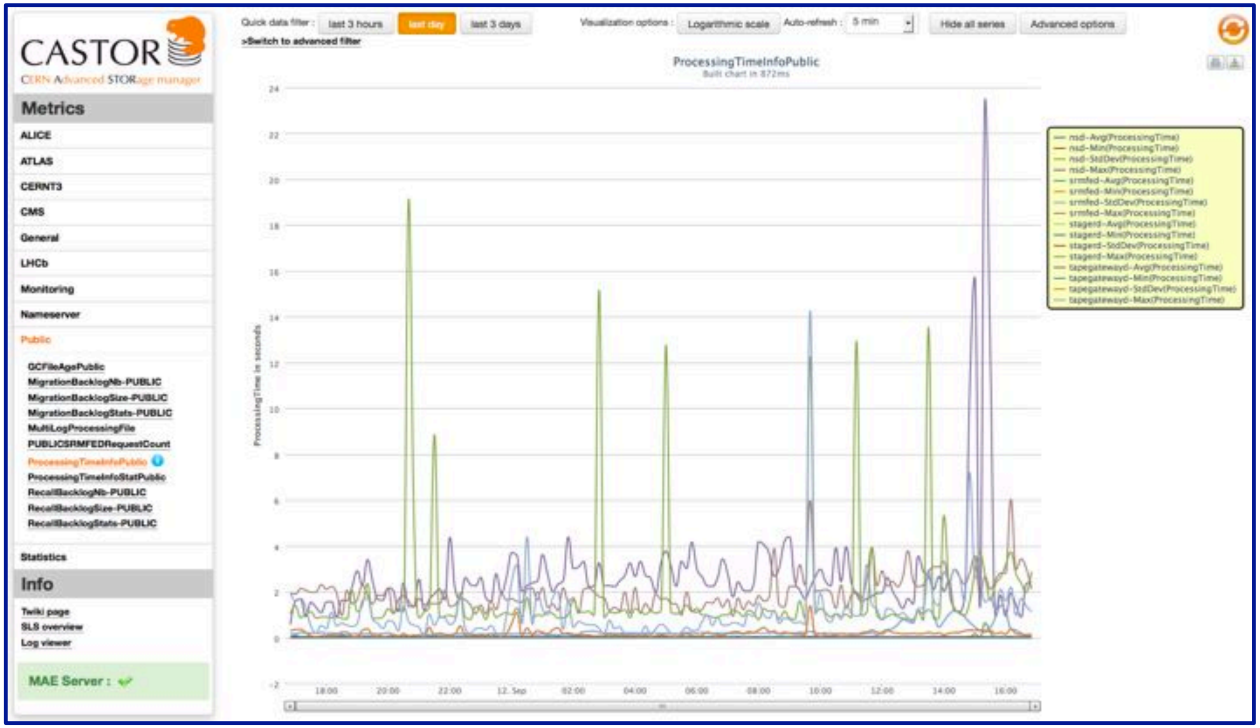
Data source: Service Level Status (SLS)



Data format: a new data point only if the status changes

The CASTOR use case

Data source: Castor Cockpit



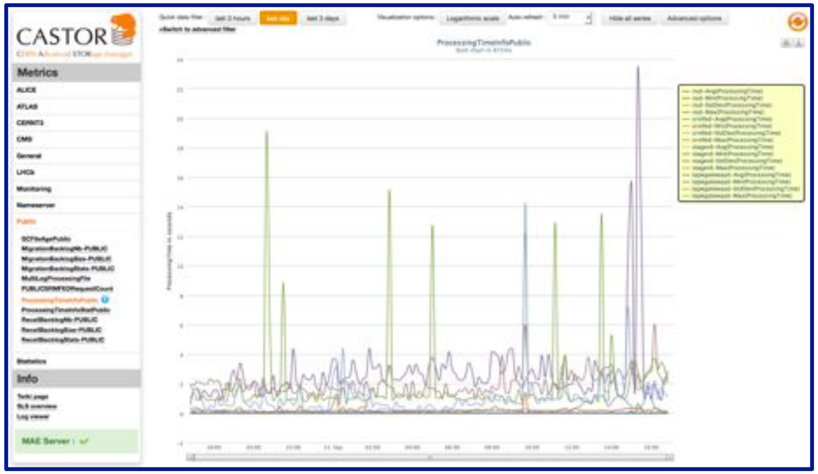
Data format: a new data point every x seconds,

- No sync between metrics
- Missing bins

The CASTOR use case

Data source: Hadoop via Castor Cockpit

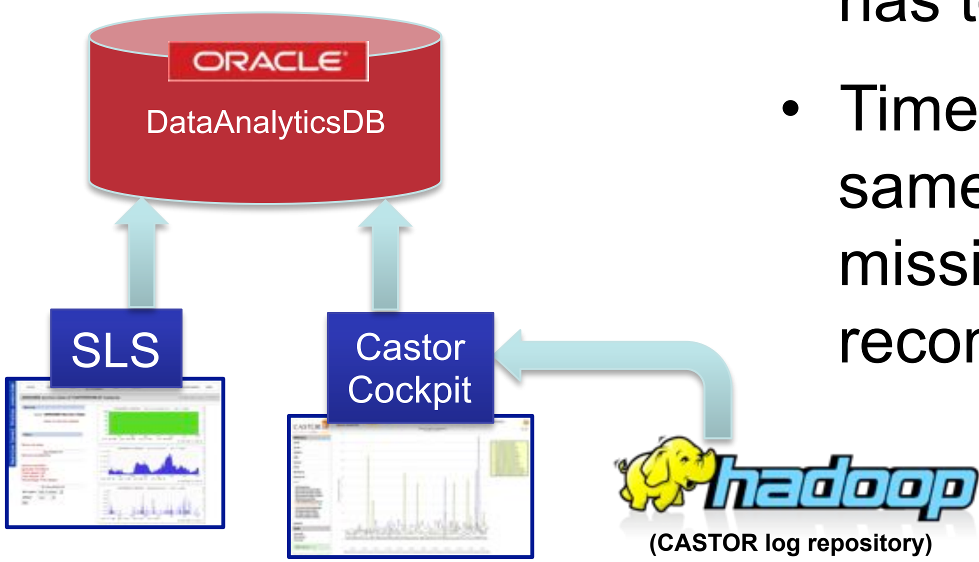
- The Castor Cockpit can load data from the Hadoop respository (Raw data, Terabytes)



CASTOR data mining: approach

APPROACH:

CENTRALIZE and STANDARDIZE data



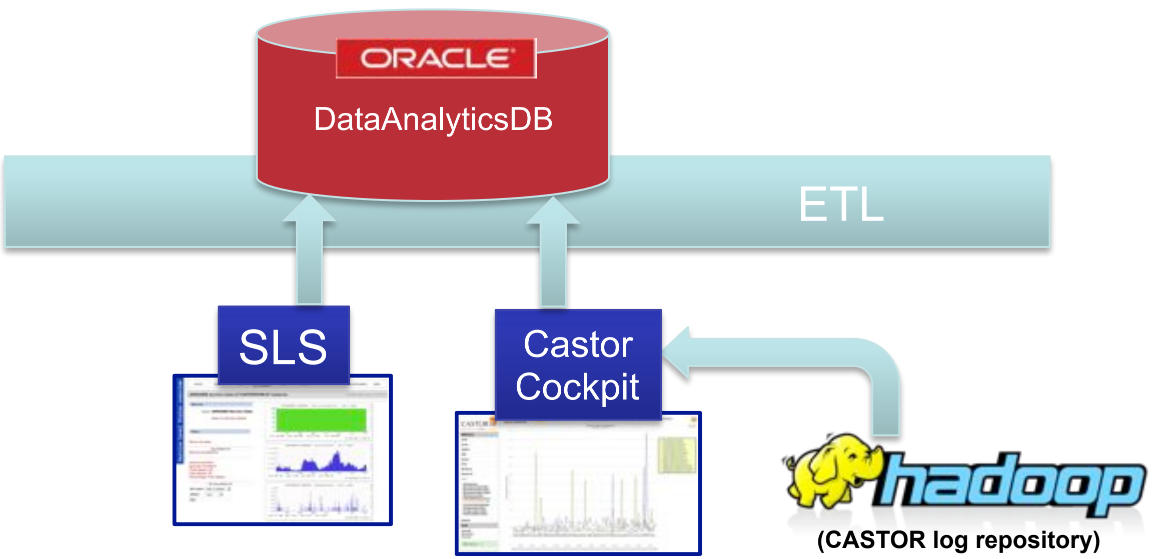
- Data from various sources has to be standardized
- Time series must have the same frequency and missing values has to be reconstructed.

CASTOR data mining: approach

APPROACH:

CENTRALIZE and STANDARDIZE data

Home made ETL process



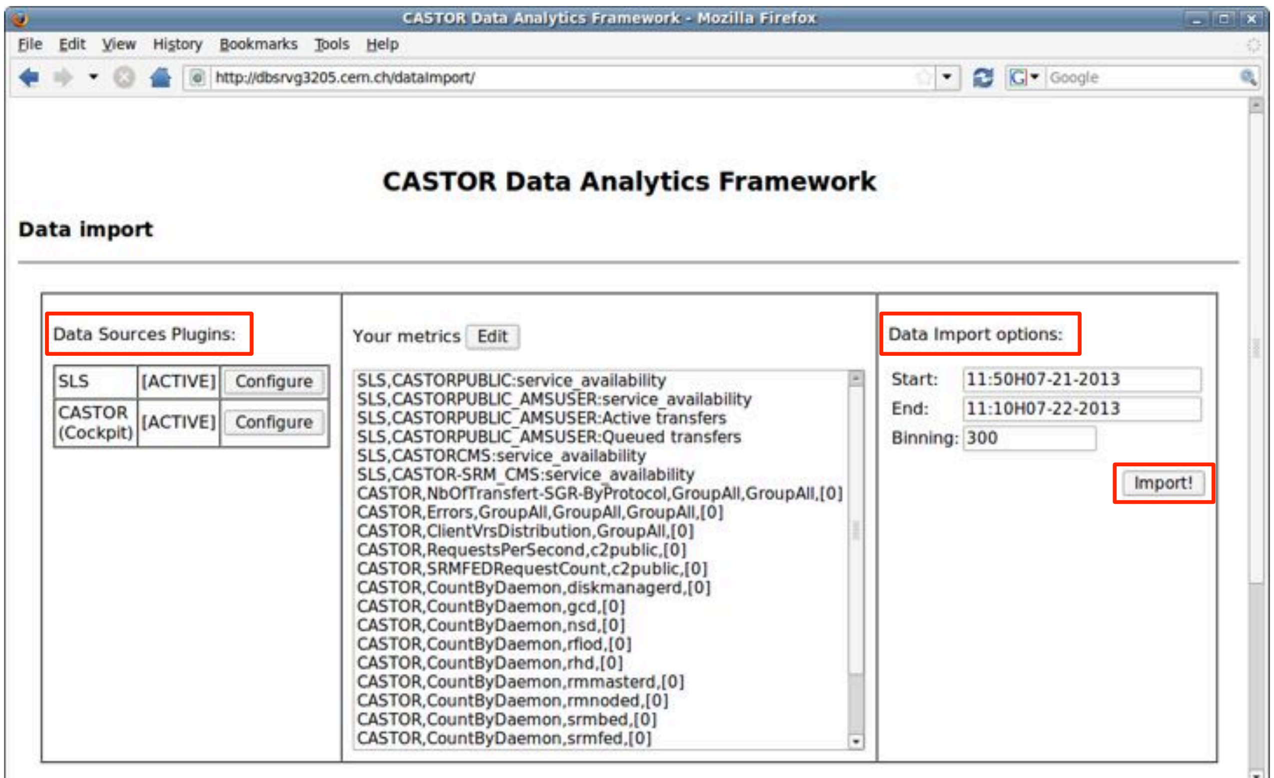
- Extract
- Transform
- Load

CASTOR data mining: approach

APPROACH:

CENTRALIZE and STANDARDIZE data

Home made ETL process



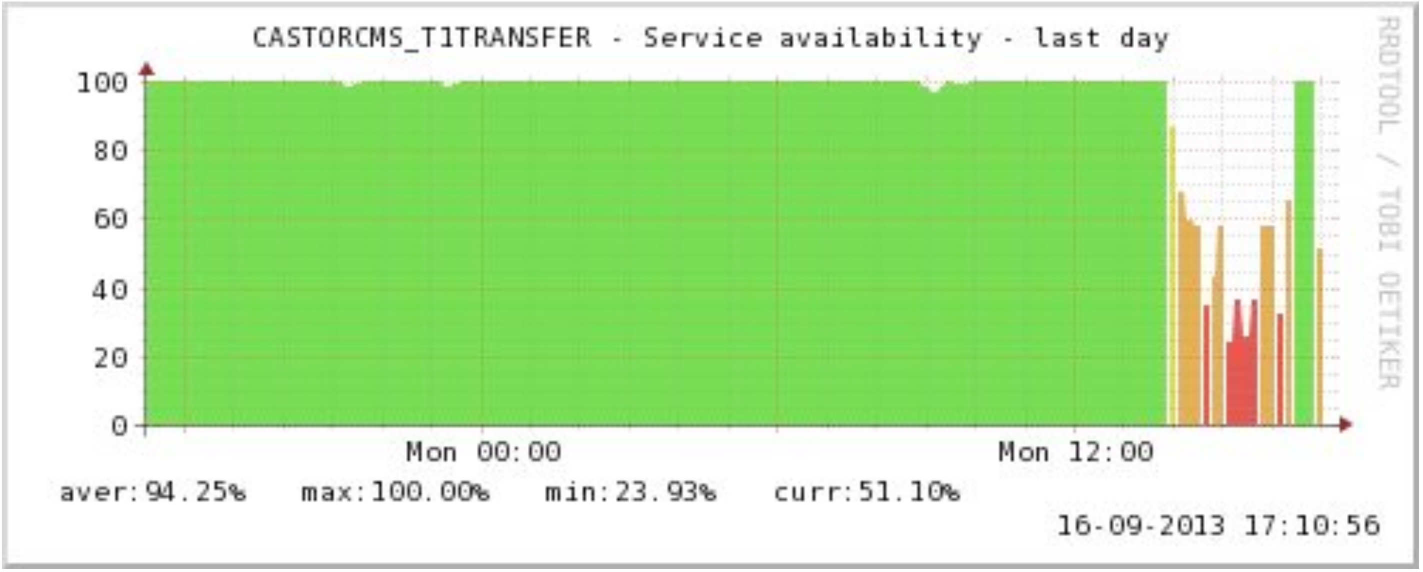
- Extract
- Transform
- Load

Python backend

The CASTOR use case

"Adverse Events"

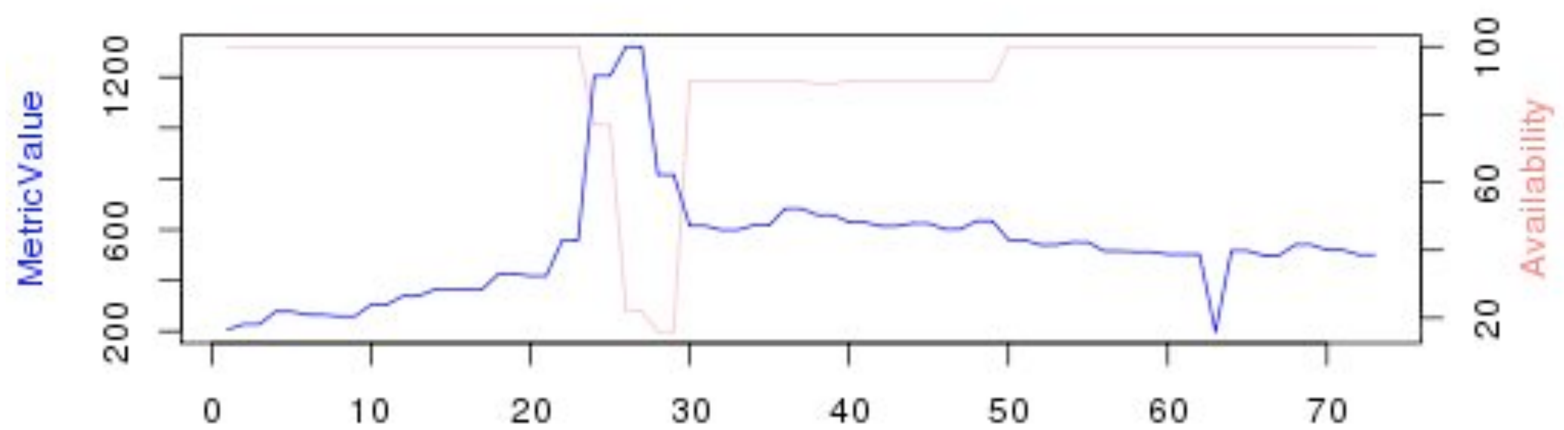
- An *adverse event* is a downtime/problem/failure in the CASTOR system
- **The SLS *service availability* metric identifies in first approximation the "adverse events":**



The CASTOR use case

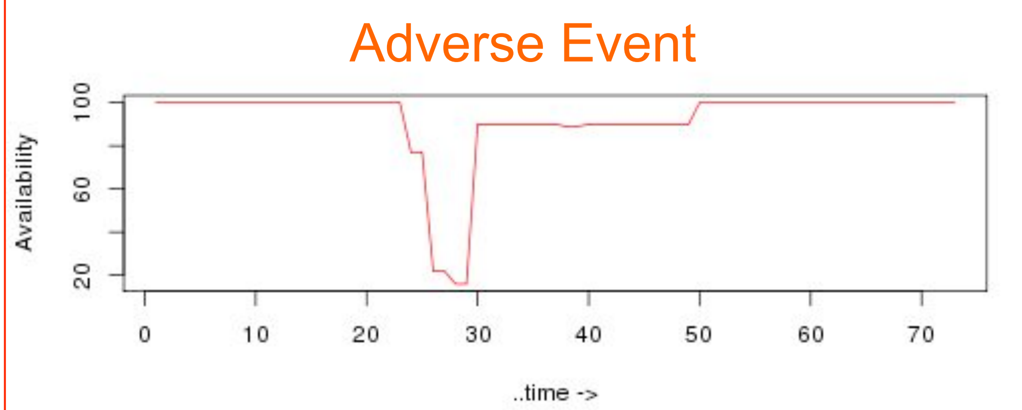
Correlating Anomalies

- **Easy to correlate:**

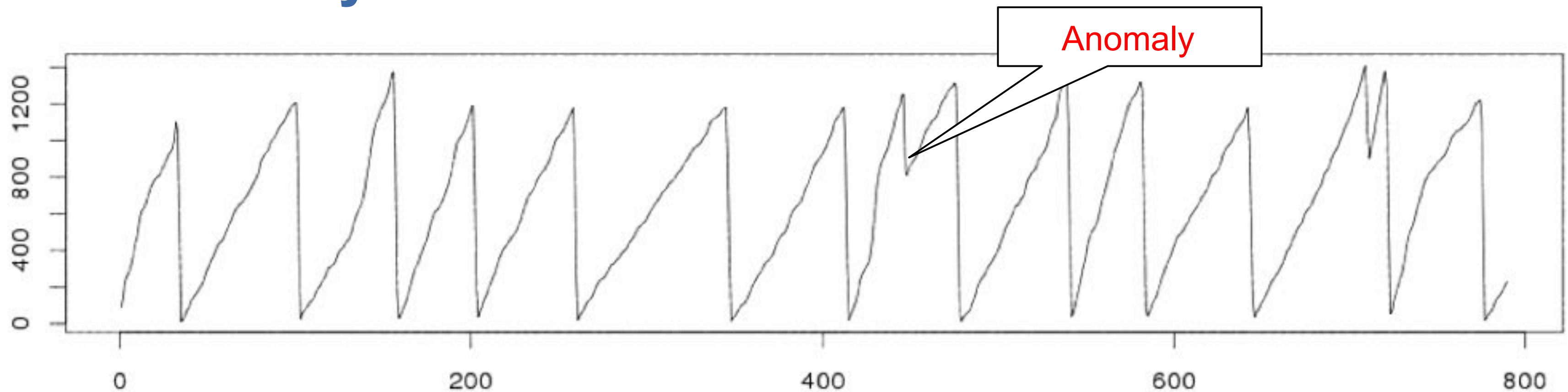


The CASTOR use case

Correlating Anomalies



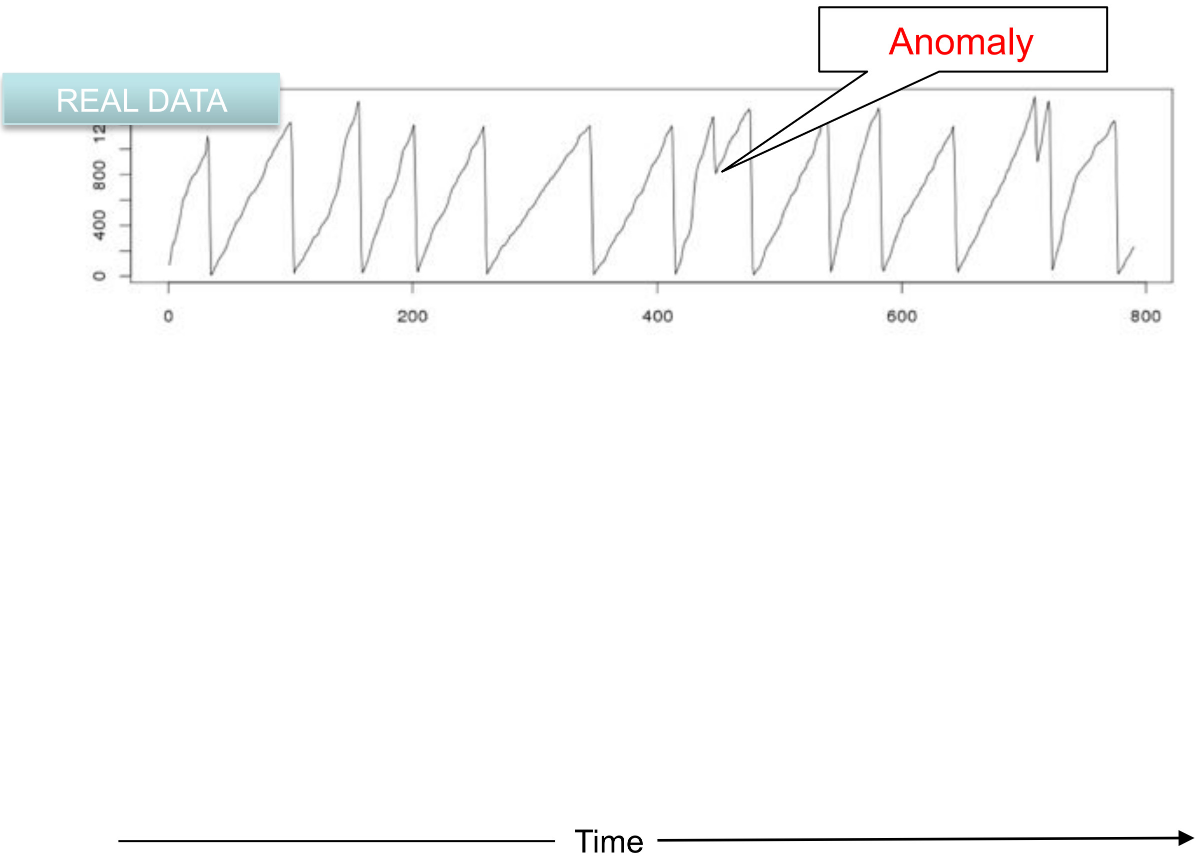
- **Not so easy:**



Example algorithm: data to be transferred to tape.

Queue data, then transfer to tape in one-go

Anomaly detection

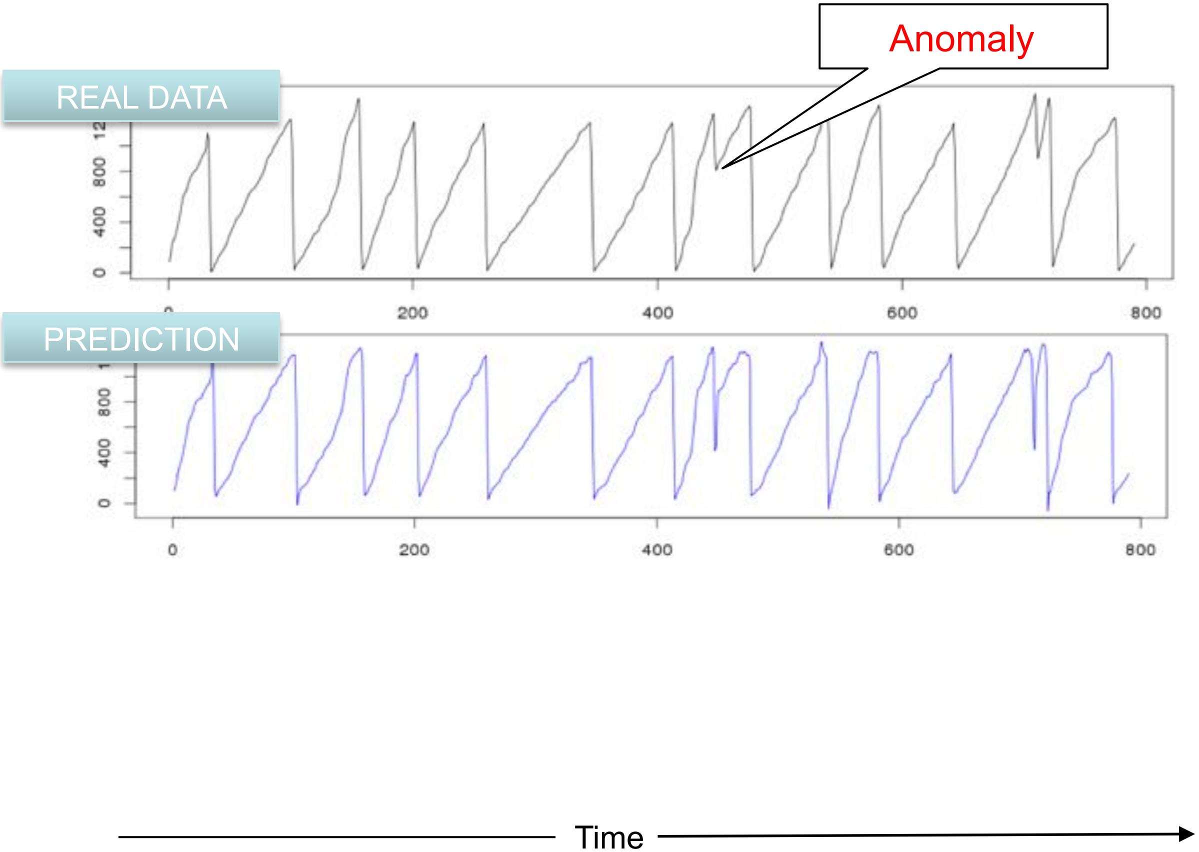


1) Build a **SVM** (*Neural Network like*) **model**
- self trained
- <u>no supervision</u>

Anomaly detection

1) Build a **SVM** (*Neural Network like*) **model**
   - self trained
   - no supervision

2) Predict



The CASTOR use case

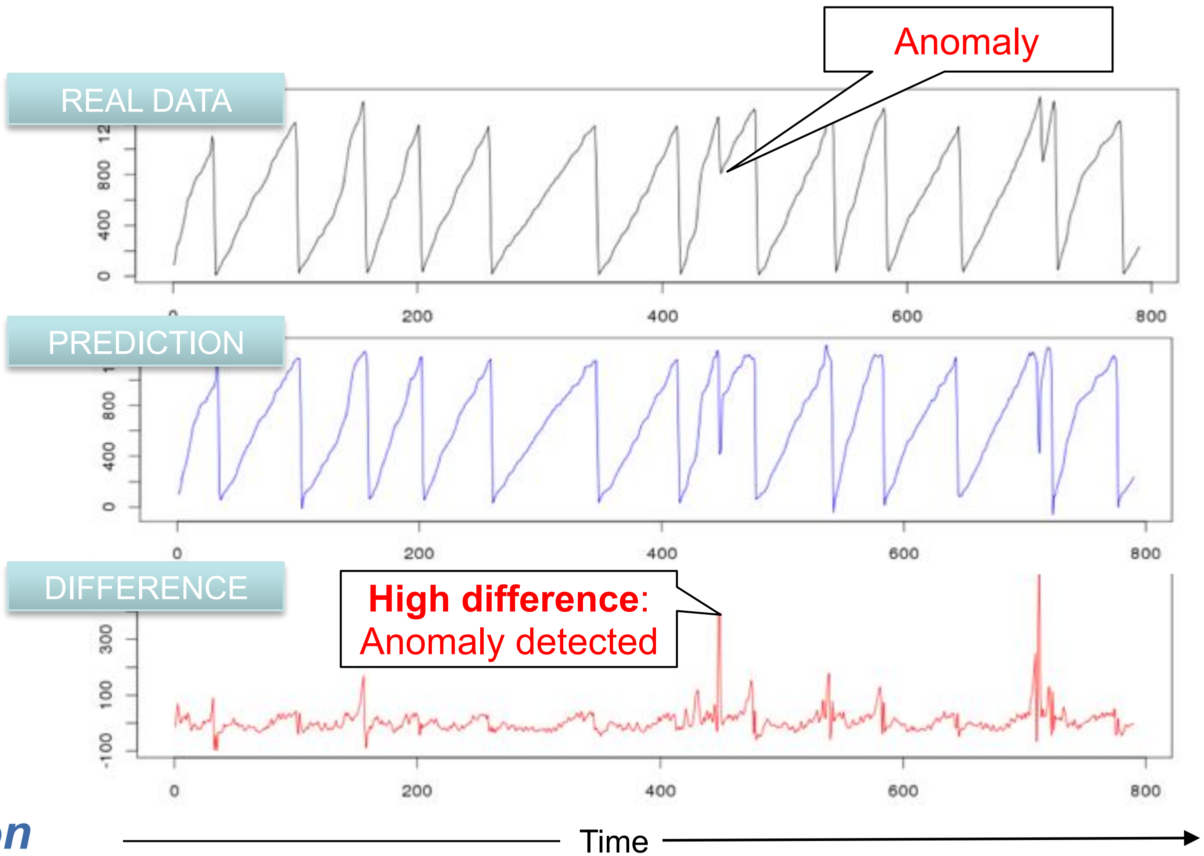
Anomaly detection

1) Build a **SVM** (*Neural Network like*) **model**
   - self trained
   - no supervision

2) Predict and compare:

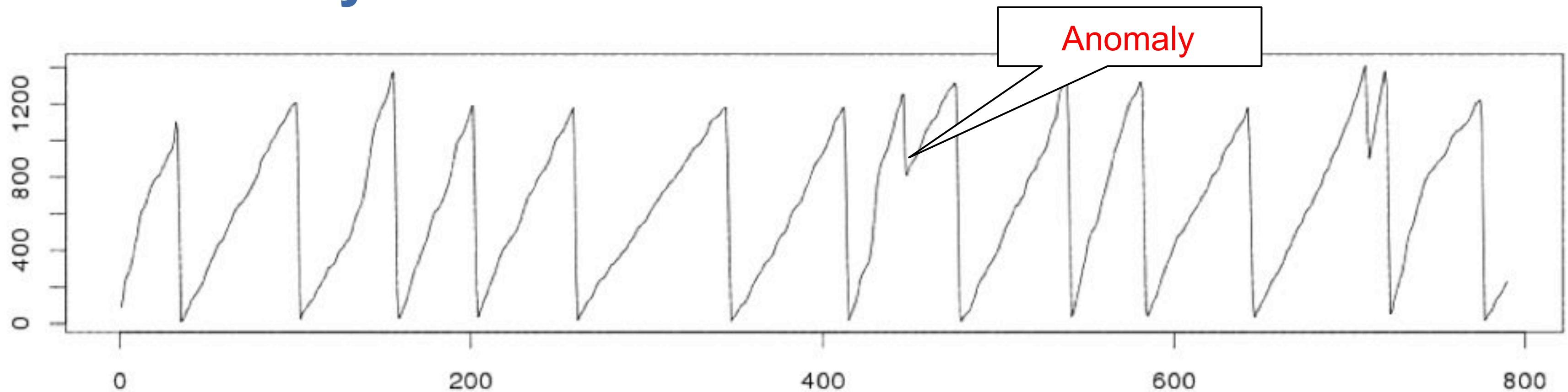
Real data Vs Prediction

- **Blindly recognize anomalies**
- ***No other information required*** *(i.e. thresholds)*



Correlating Anomalies

- **Not so easy:**

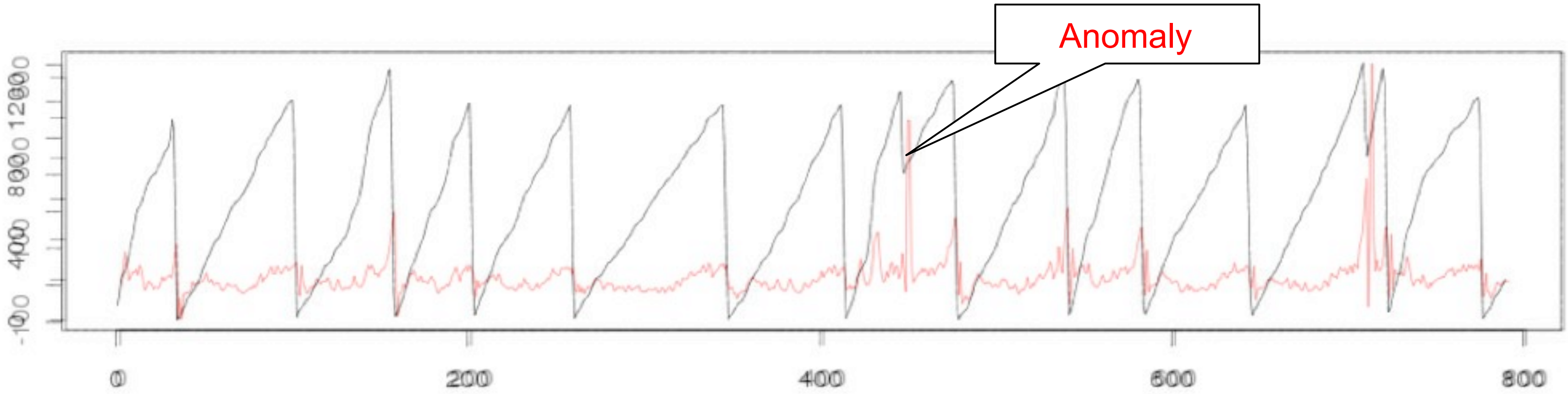


Example algorithm: data to be transferred to tape.
Queue data, then transfer to tape in one-go

The CASTOR use case

Correlating Anomalies

- **Becomes now…**

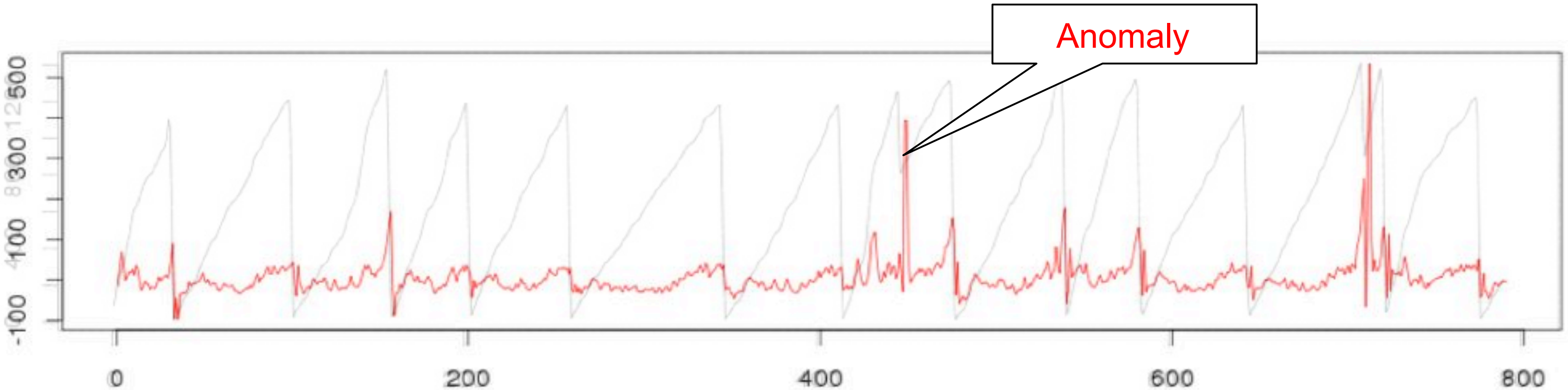


Example algorithm: data to be transferred to tape.
Queue data, then transfer to tape in one-go

The CASTOR use case

Correlating Anomalies

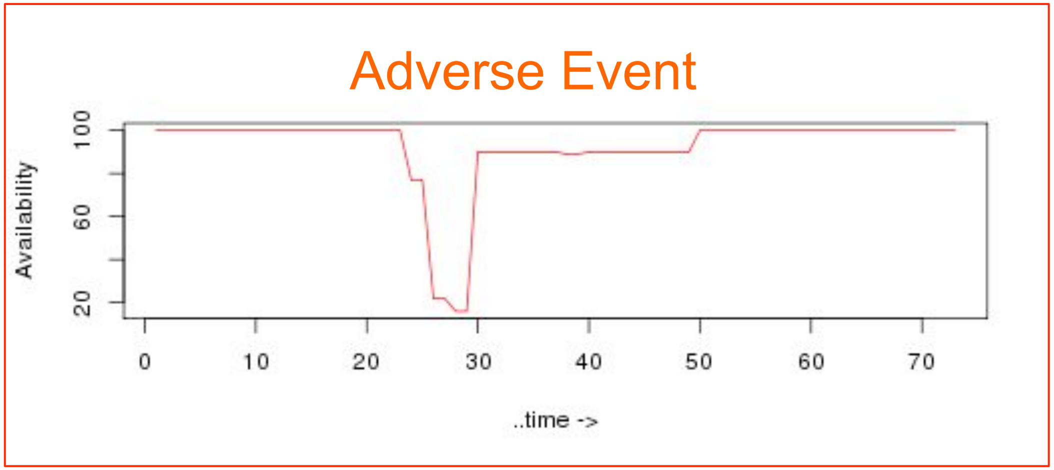
- **Becomes now…**



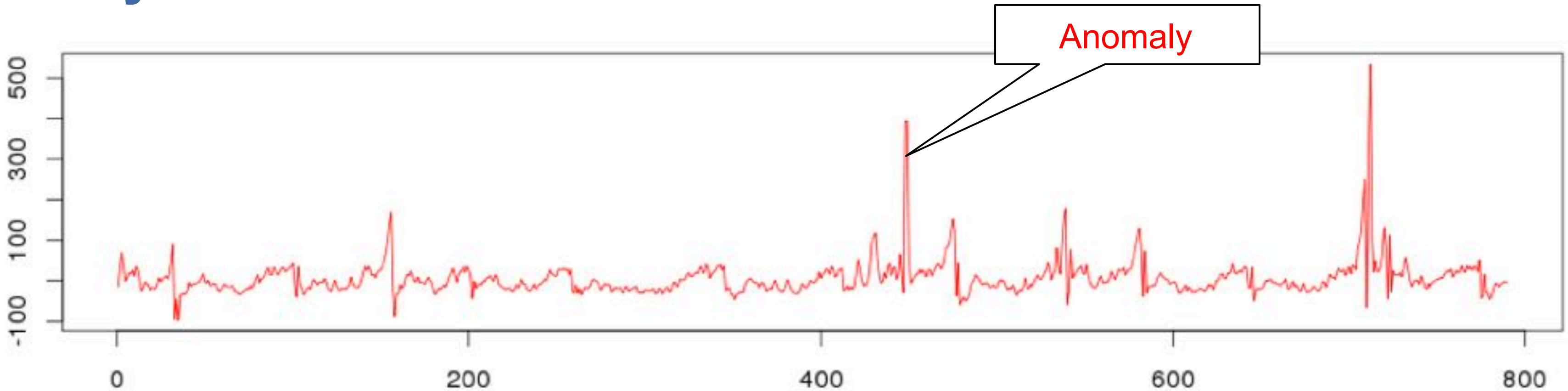
Example algorithm: data to be transferred to tape.
Queue data, then transfer to tape in one-go

The CASTOR use case

Correlating Anomalies



- **Way easier:**



Example algorithm: data to be transferred to tape.

Queue data, then transfer to tape in one-go

The CASTOR use case

Why ORE for the CASTOR use case?

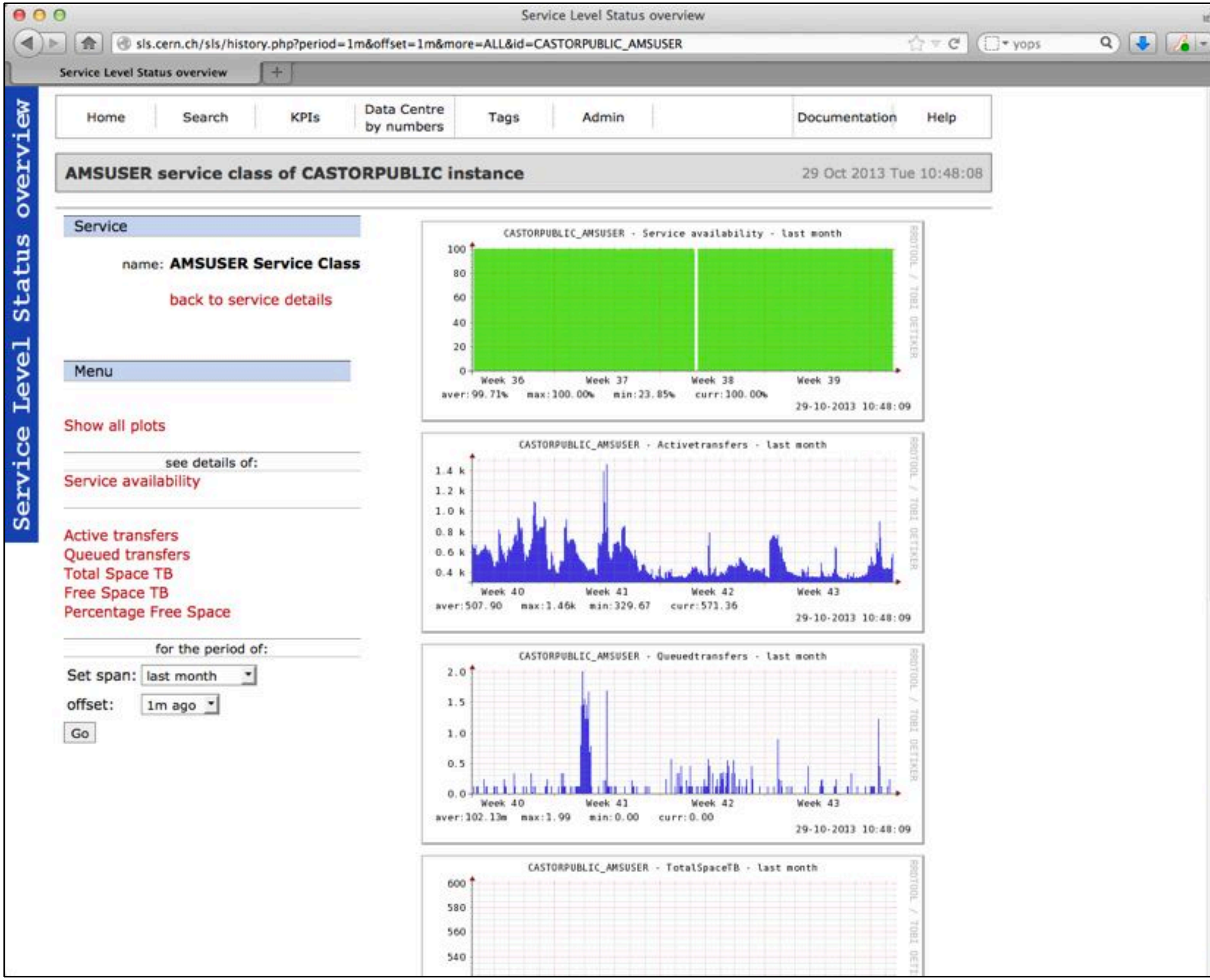
- Detecting anomalies requires advanced statistical and machine learning models: ready to use in R
- The CASTOR use case involves two challenges:
  - Many variables to evaluate (management is complex)
    - *SLS Status broken down by instance and service class*
    - *Several other SLS variables for every instance and service class*
    - *Cockpit data broken down by many other fields (i.e. daemon, message, hostname)*
  - Large amounts of data to process
    - Correlations cannot be computed over NoSQL solutions (Hadoop)
    - standard R in-memory approach can fail

In-database R (ORE) solves all these problems

DEMO SESSION

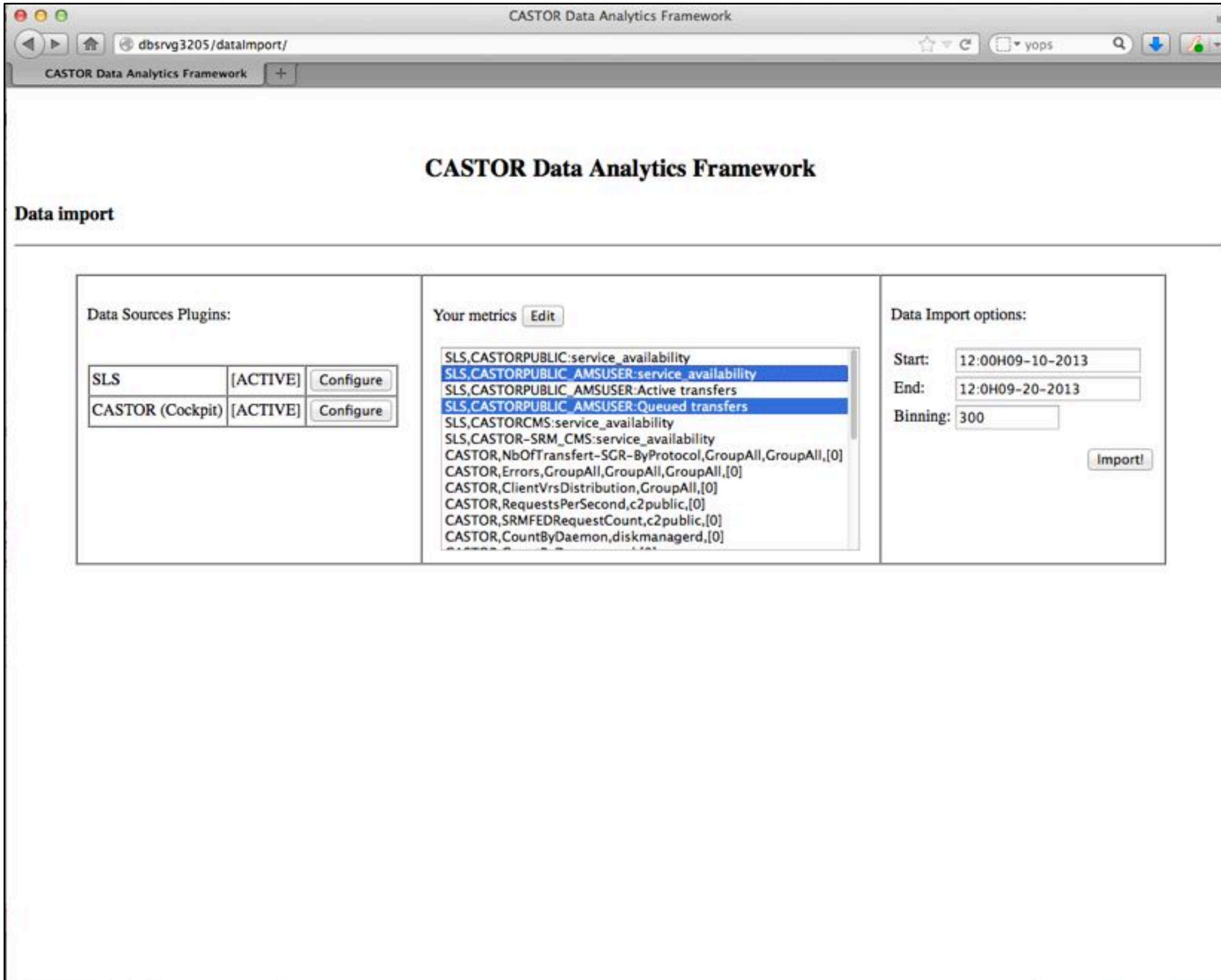
The CASTOR use case

DEMO SESSION



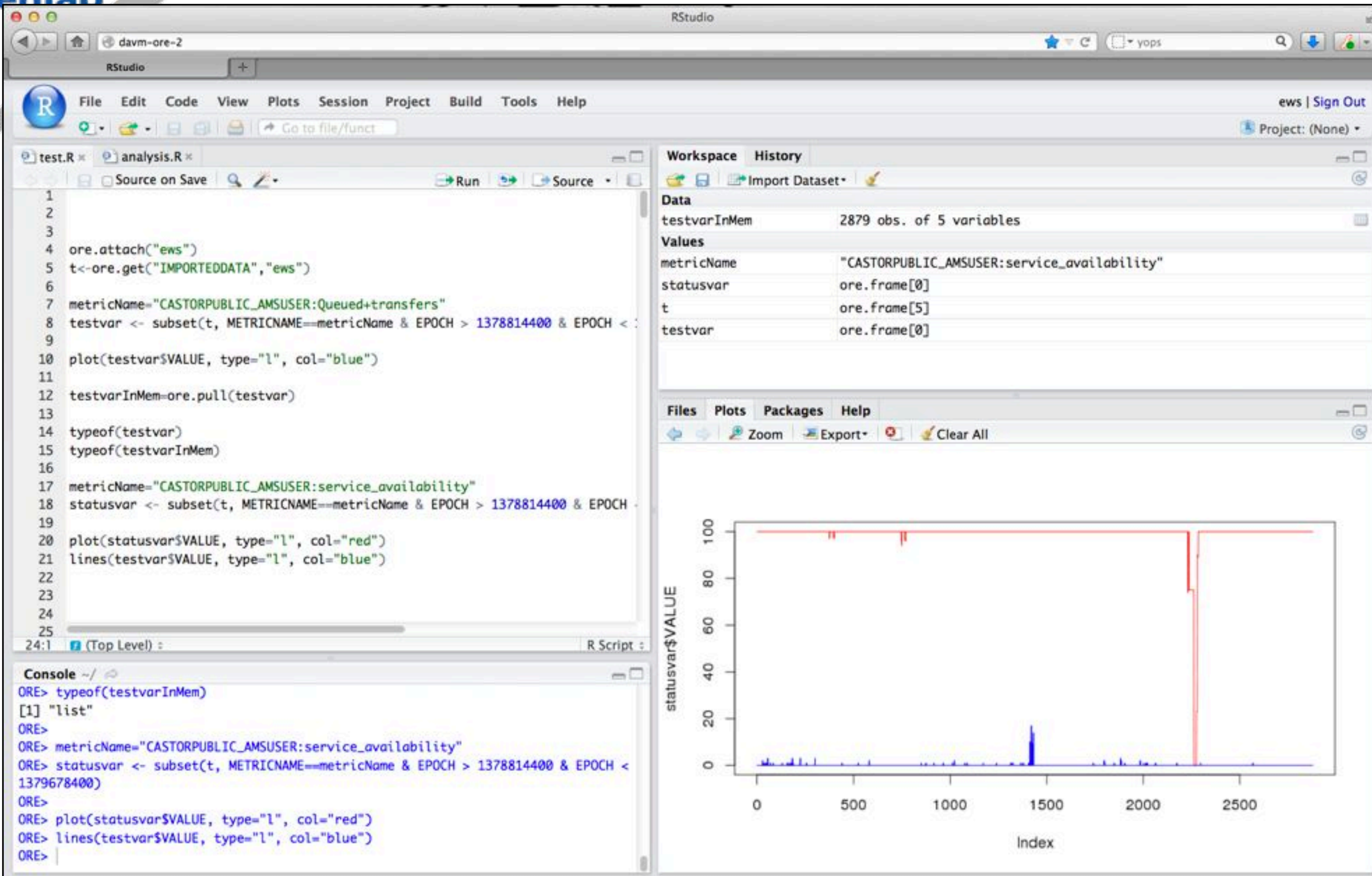
The CASTOR use case

DEMO SESSION



The CASTOR use case

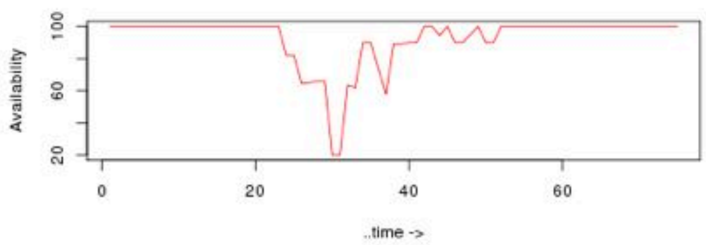
DEMO SESSION

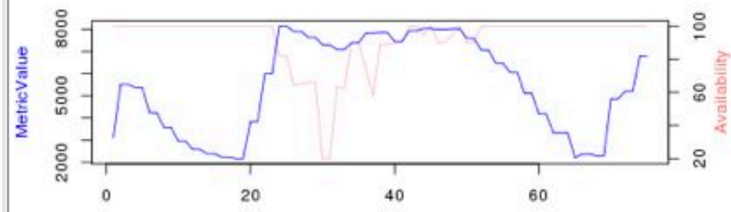


The CASTOR use case

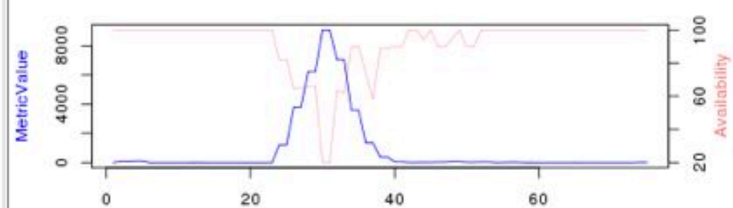
DEMO SESSION

CASTORPUBLIC_AMSUSER:service_availability - Thu, 21 Mar 2013 09:40:00 +0100





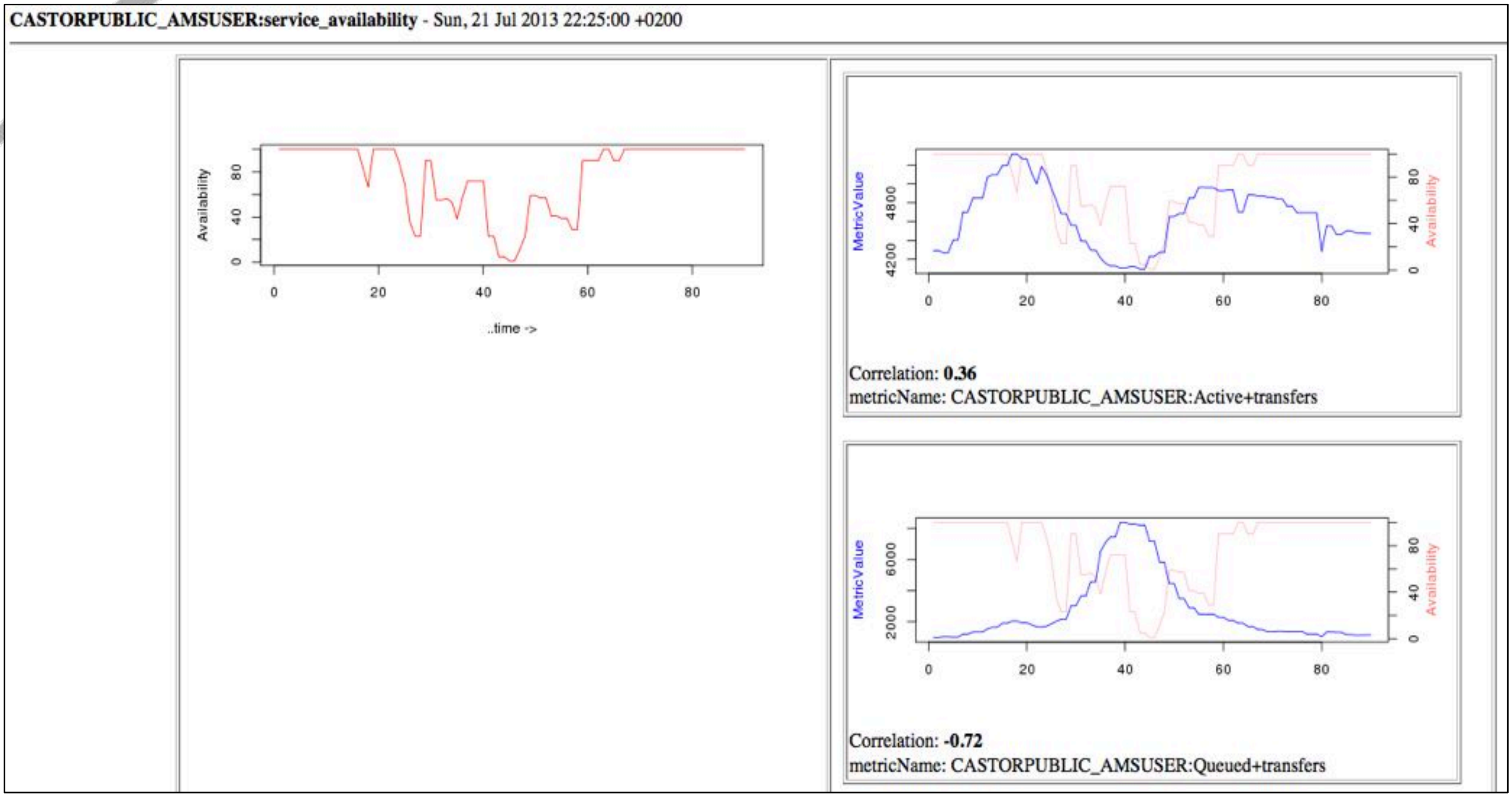
Correlation: **-0.47**
metricName: CASTORPUBLIC_AMSUSER:Active+transfers



Correlation: **-0.90**
metricName: CASTORPUBLIC_AMSUSER:Queued+transfers

The CASTOR use case

DEMO SESSION



Conclusions

Conclusions

- **R** is a **powerful** and interesting tool for data analysis
- **ORE brings R into a scalable DB engine (**solving problems of data management, analysis and scalability)
- We actually can obtain **information** and **added value** from not so actively used **data**
- Lesson learnt: The more you aggregate, the more you can give a meaning to your data:

3rd aggregation: correlate results

2nd aggregation: DB, ORE

1st aggregation: SLS / CASTOR Cockpit

RAW Data

Future plans

- Work with castor team to deliver daily reports of CASTOR analysis
- Extend to other use cases
- Promote the data analytics project
  - 20th November: CERN data analytics use cases workshop
  - February 2014: II Openlab Workshop on Big Data Analytics

Thanks to the CASTOR team and Oracle!

- **II Openlab Workshop on Big Data Analytics (February 2014)**
  - **Big Data Solutions**
  - **Big Data Analytics Technologies**
    - Advance Analytics
    - Information Discovery
    - In-Database Analytics
    - Logs Analysis
    - Advance Visualization
  - **Main Challenges**

M.Sc. in Soft-computing and Intelligence Information Systems in 2011 both at the University of Granada, Spain. He is a member of the "Soft Computing and Intelligent Information Systems" (SCI2S) research group and the "Distributed Computational Intelligence and Time Series" research lab (DICITS). In 2007 he joined the Beam department of the European Centre for Nuclear Research (CERN). A that time he was co-leading the Control Configuration Database and the database integration of the Front-End Software Architecture. Both projects of critical importance for improving the CERN's accelerators complex control systems. Currently, Manuel is part of the CERN IT

architects from Oracle Development. In this role he h the honor to work with companies like Amazon, Yaho Mastercard, Nielsen, Bank of America, Allianz, BN Paribas, Deutsche Bank, Turkcell, Garanti Bank, S Telekom and many more. He started as Consultant, th moved into Project Management, Presales and final Business Development.

End

Thank you for your attention!

- Questions..?