Model-Based Database Systems

Kasun S Perera[1], Martin Hahmann[1], Wolfgang Lehner[1], Christian Thomsen[2], Torben Bach Pedersen[2]

Technische Universiät Dresden[1] – Germany , Aalborg University[2] – Denmark

TECHNISCHE UNIVERSITÄT DRESDEN

Motivation

❖ WE NEED TO EXTRACT INFORMATION FROM DATA
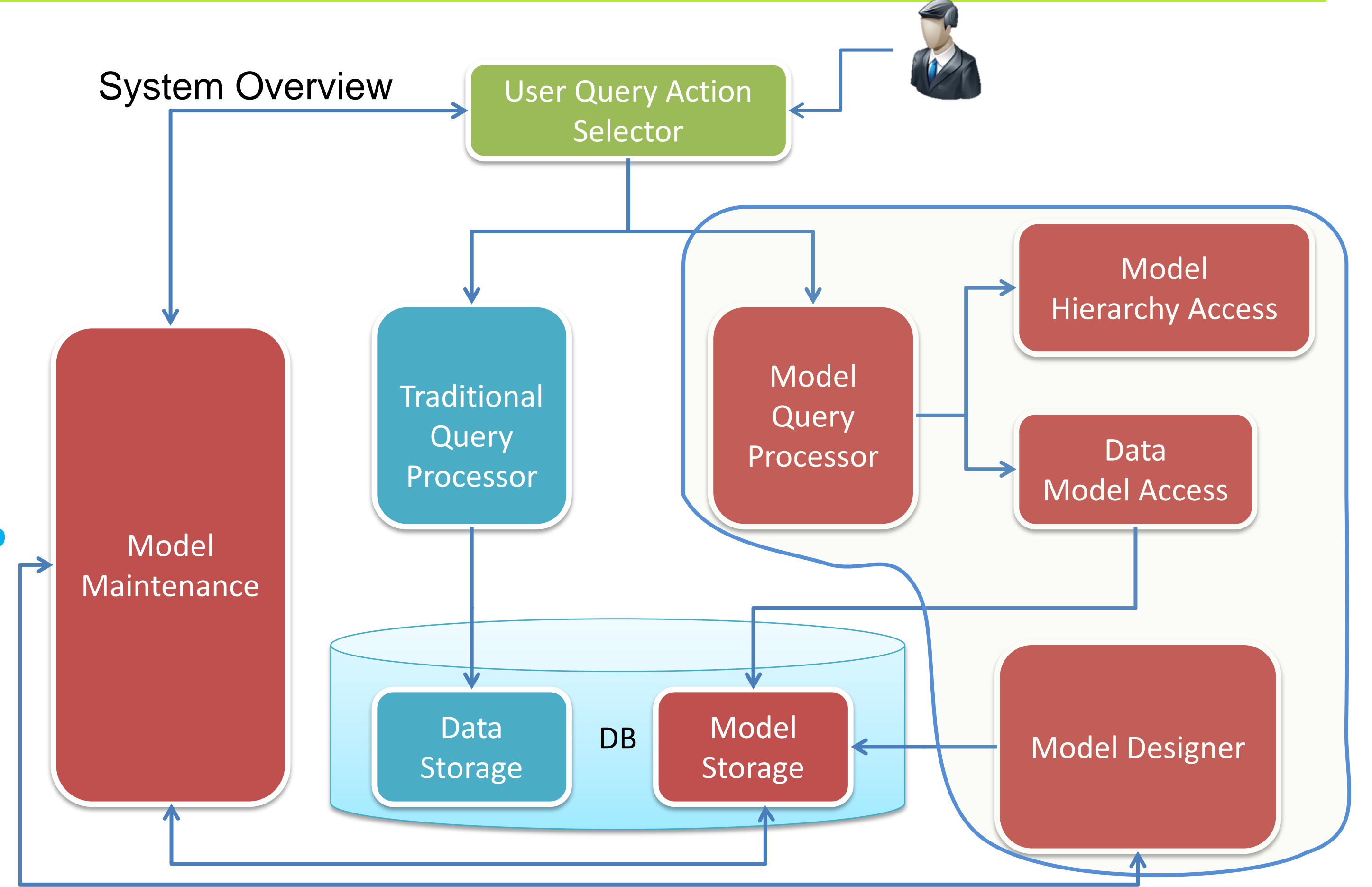
❖ PATTERNS ARE MORE IMPORTANT THAN DATA

❖ DATA IS GETTING BIGGER

❖ INFORMATION EXTRACTION IS GETTING HARDER

Model: A simplified description, a mathematical one, of a system or a process, to assist in calculations and predictions: (Oxford Dic)
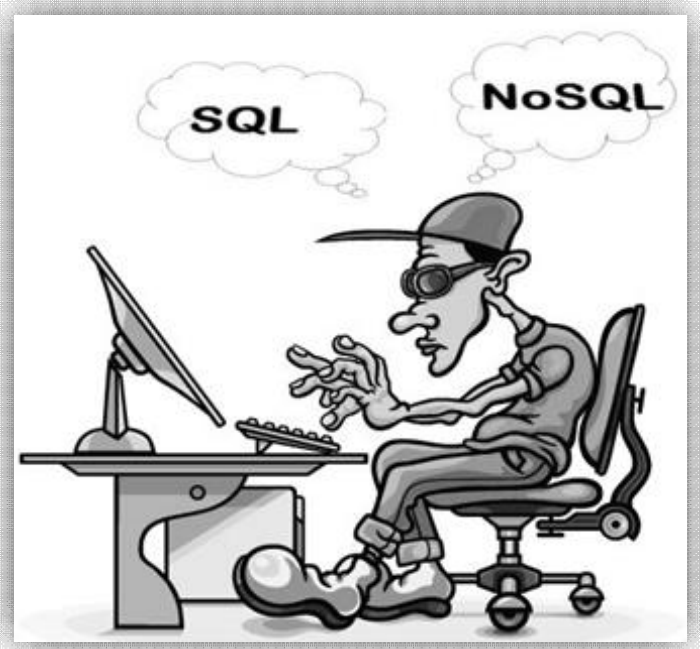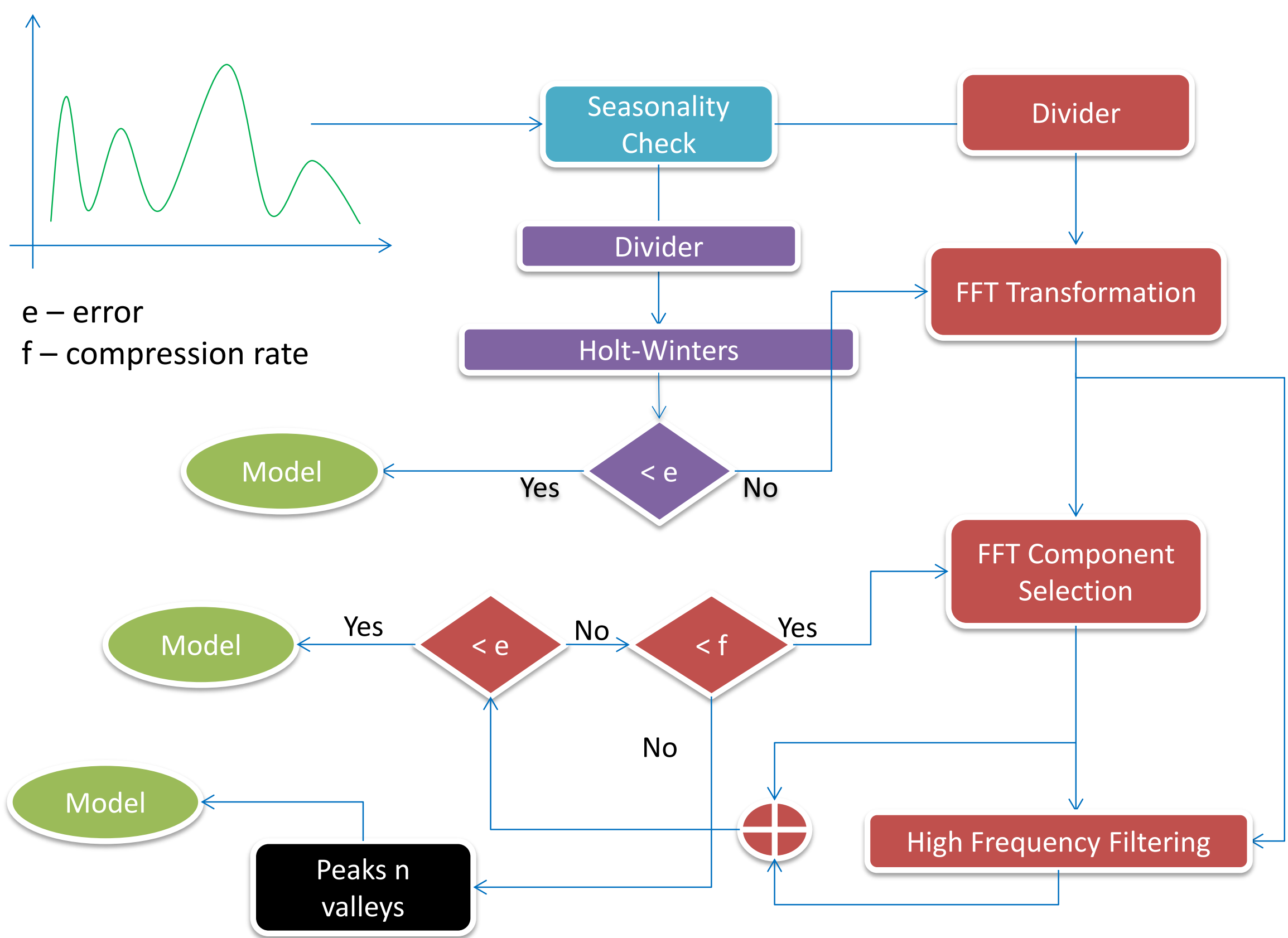
❖ RDBMS CAN NO LONGER SUPPORT COMPLEX QUERIES

❖ DIFFERENT TECHNIQUES TO DEAL WITH DIFFERENT PROBLEMS
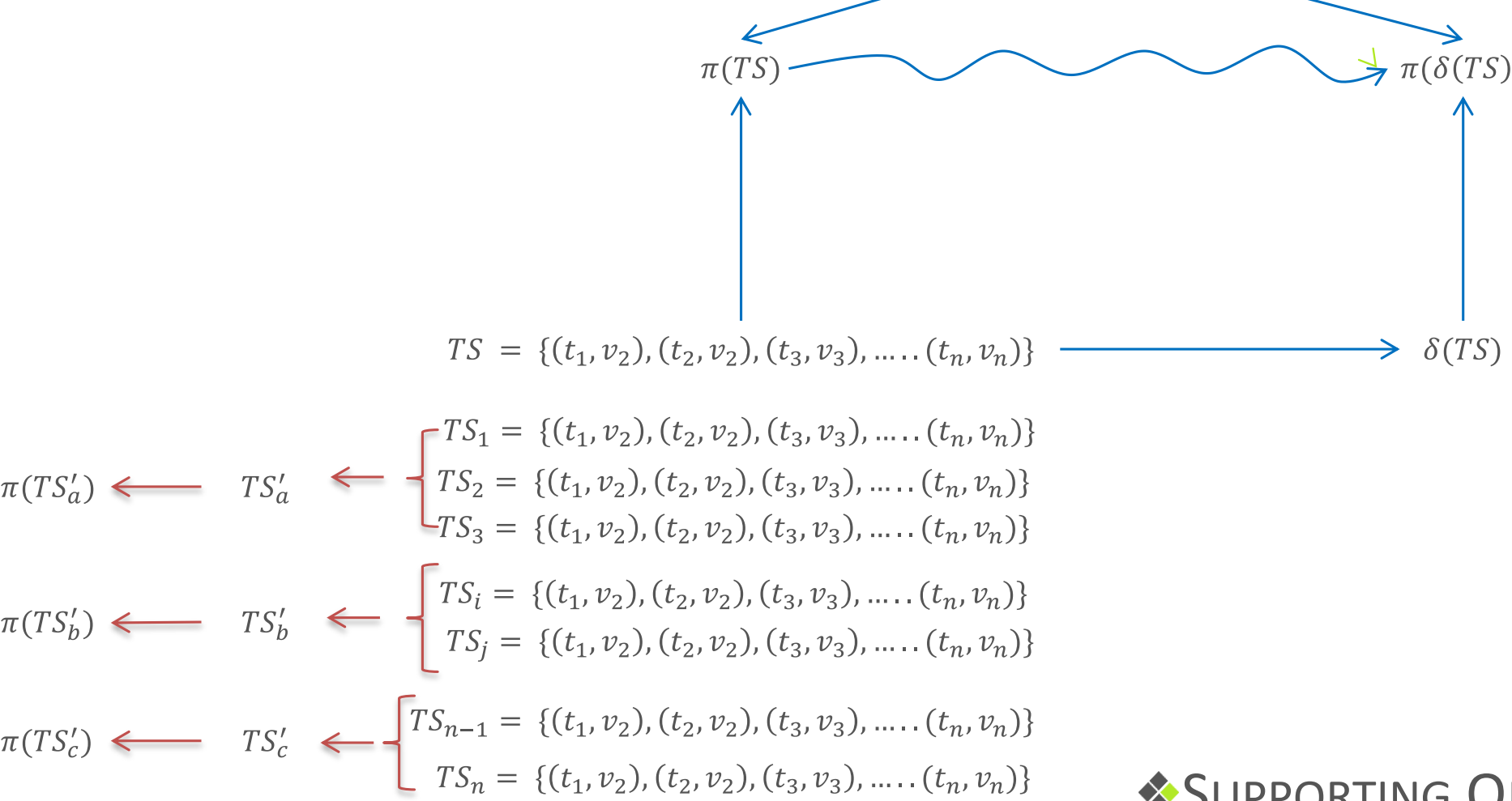
❖ EFFICIENT INFORMATION EXTRACTION WITH RELATIVE ACCURACY

System Overview

User Query Action Selector

Model Maintenance — Traditional Query Processor — Model Query Processor — Model Hierarchy Access — Data Model Access — Model Designer

DB — Data Storage — Model Storage

Phase 1 – Time Series Modeling and Querying

❖ HOLT WINTERS MODEL
- Widely used forecasting model
- Fit the model to data and extend the model
- Use the intermediate results to represent the full time series
- Works well with seasonal time series

❖ LANDMARK POINTS (PEAKS AND VALLEYS)
- Simple to implement
- Adjustable to compression rate, accuracy
- Querying is relatively easy

❖ FAST FOURIER TRANSFORMATION
- Sound mathematical background
- Adjustable compression rates
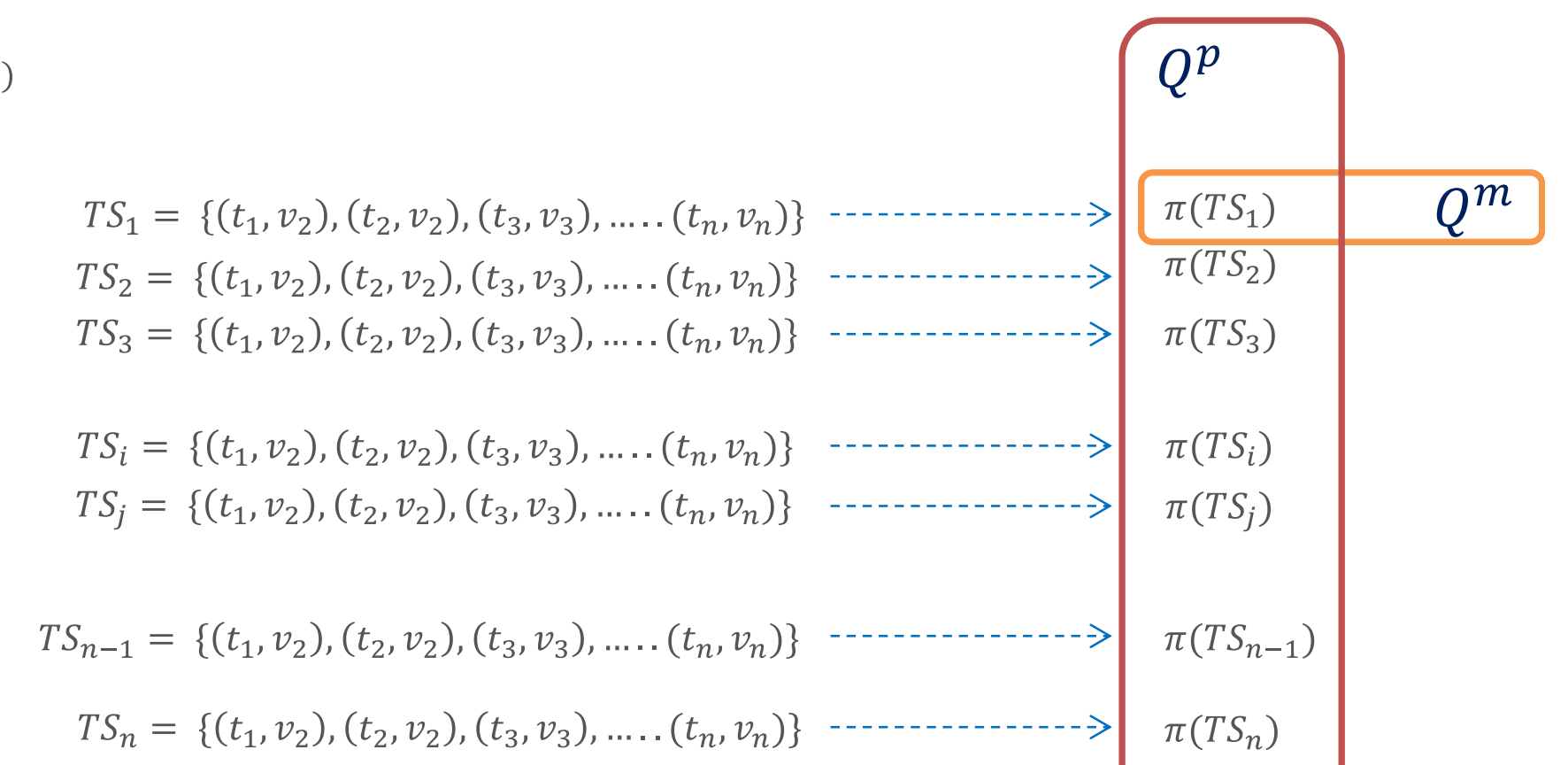- Works well on both seasonal and non-seasonal time series

Time Series	Model	Original Length	RMSE	MAPE	Compression Factor
NYISO	1 Week window HW	26280	8.61	10.77	5.25
NYISO	2 weeks window HW	26280	9.16	11.71	8.62
UK Energy Demand	1 week window HW	17520	2574.42	5.52	5.25
UK Energy Demand	2 weeks window HW	17520	2785.86	7.18	8.62
UK Internet Traffic	Peaks and Valleys	19887	71.63	0.68	2.08
Site1 Wind		157967	0.58	3.42	4.17
Site1 Power		157967	6.97	INF	2.61
Site2 Wind		157967	0.44	2.88	4.44
Site2 Power		157967	4.24	INF	2.67
Site1 Power	FFT – 5% Components	157967	9.55	INF(15.01)	20
Site1 Power	FFT – 10% Components	157967	5.79	INF(8.40)	10
Site1 Power	FFT – 25% Components	157967	2.86	INF(3.44)	4

e – error
f – compression rate

Seasonality Check — Divider — Holt-Winters — < e — Model — FFT Transformation — FFT Component Selection — High Frequency Filtering — Peaks n valleys — < f — Model

Building the Model Pool

Aggregated Query (Q)

$\pi(TS)$ ... $\pi(\delta(TS))$

$TS = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$ → $\delta(TS)$

$TS_1 = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$
$TS_2 = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$
$TS_3 = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$

$TS_i = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$
$TS_j = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$

$TS_{n-1} = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$
$TS_n = \{(t_1, v_2), (t_2, v_2), (t_3, v_3), \ldots (t_n, v_n)\}$

Querying the Models

Q^p
Q^m

$\pi(TS_1)$
$\pi(TS_2)$
$\pi(TS_3)$
$\pi(TS_i)$
$\pi(TS_j)$
$\pi(TS_{n-1})$
$\pi(TS_n)$
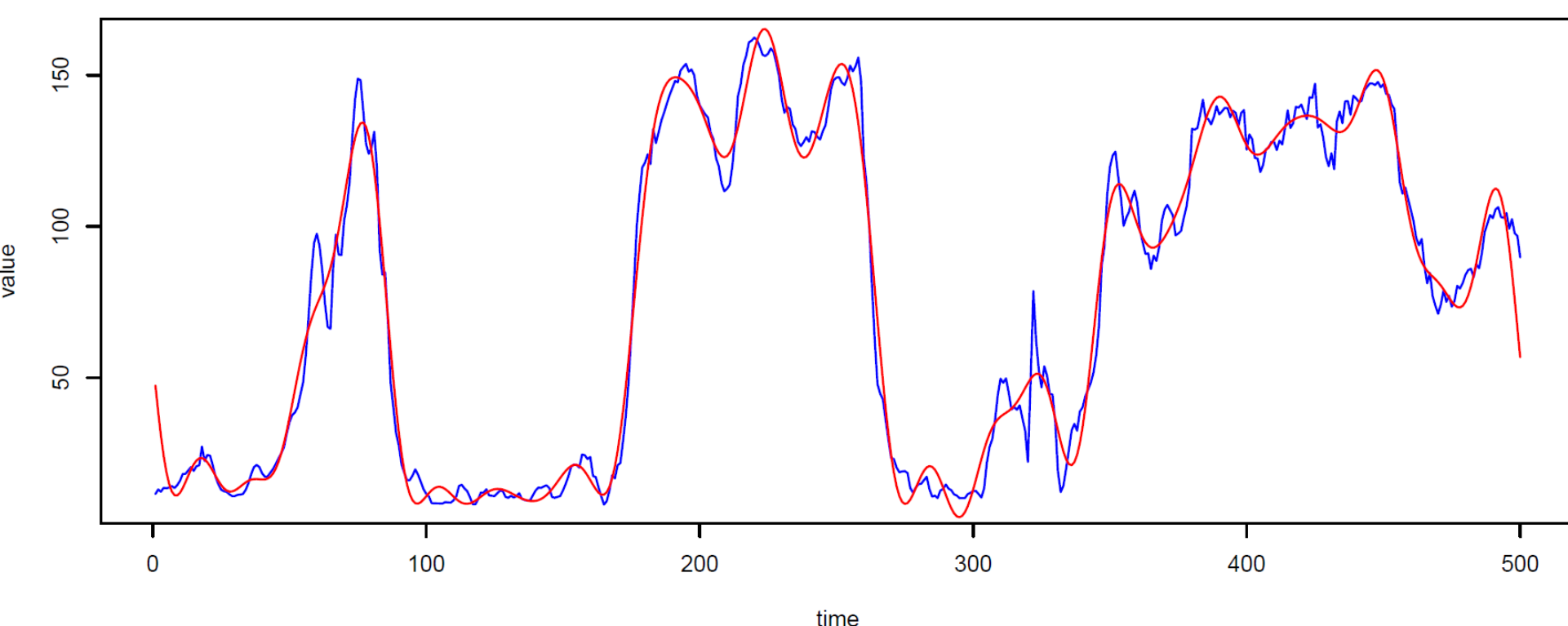
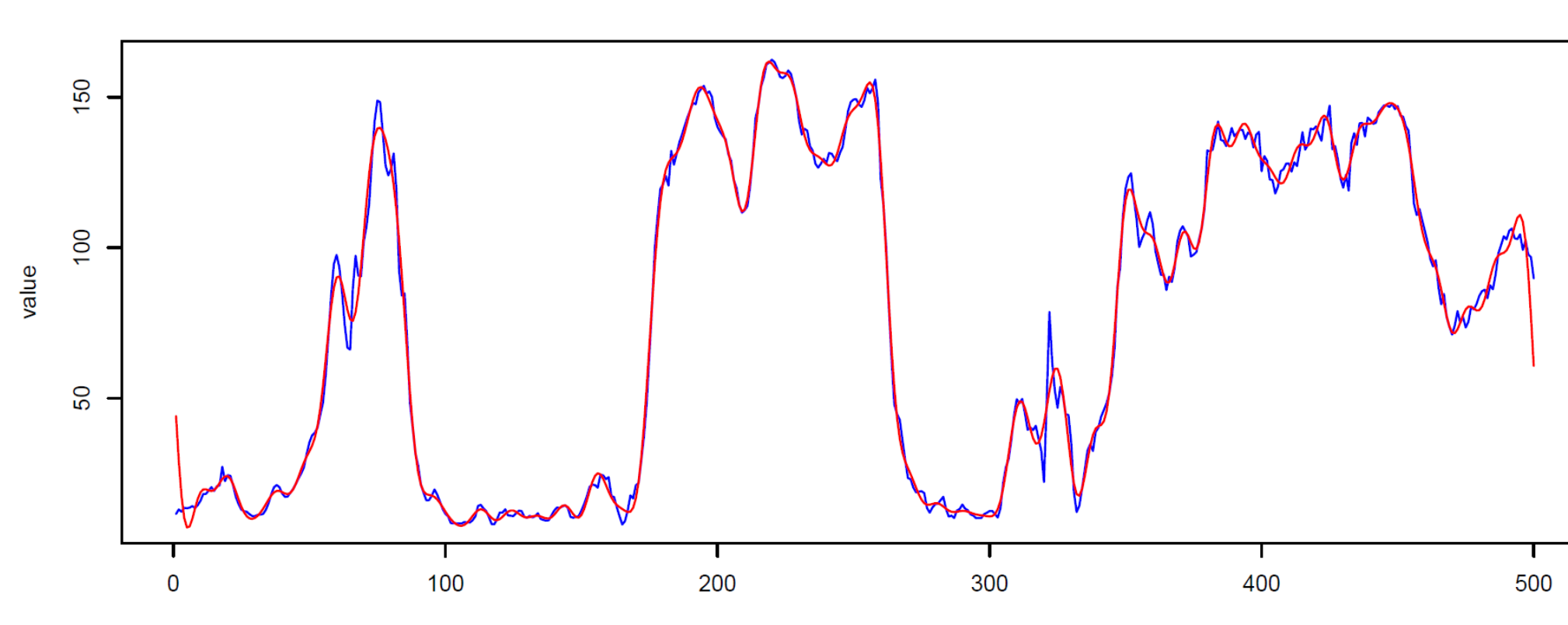❖ SUPPORTING QUERIES
- Simple Aggregate Queries
  - Time Domain, Spatial Domain, Simple Statistical Queries
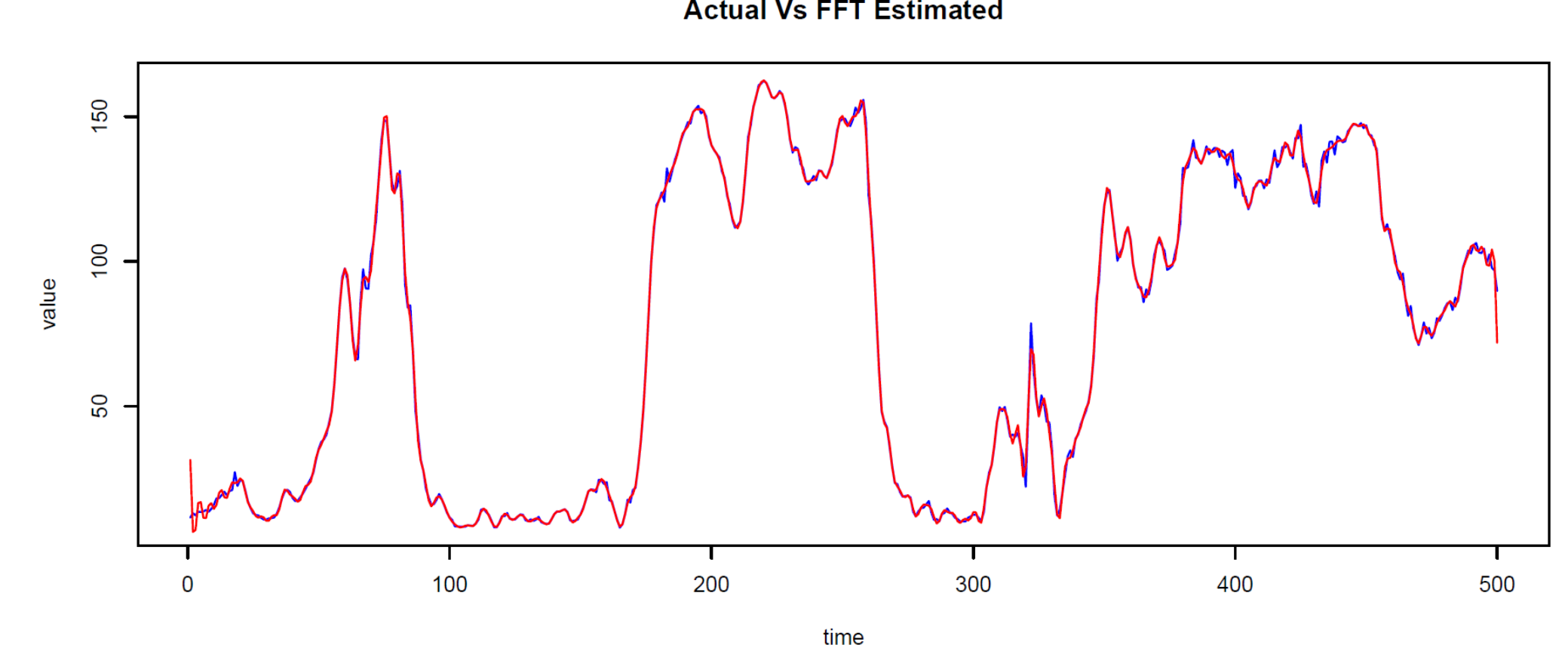- YtoD – Year to Date Queries
- Moving Average, Trend
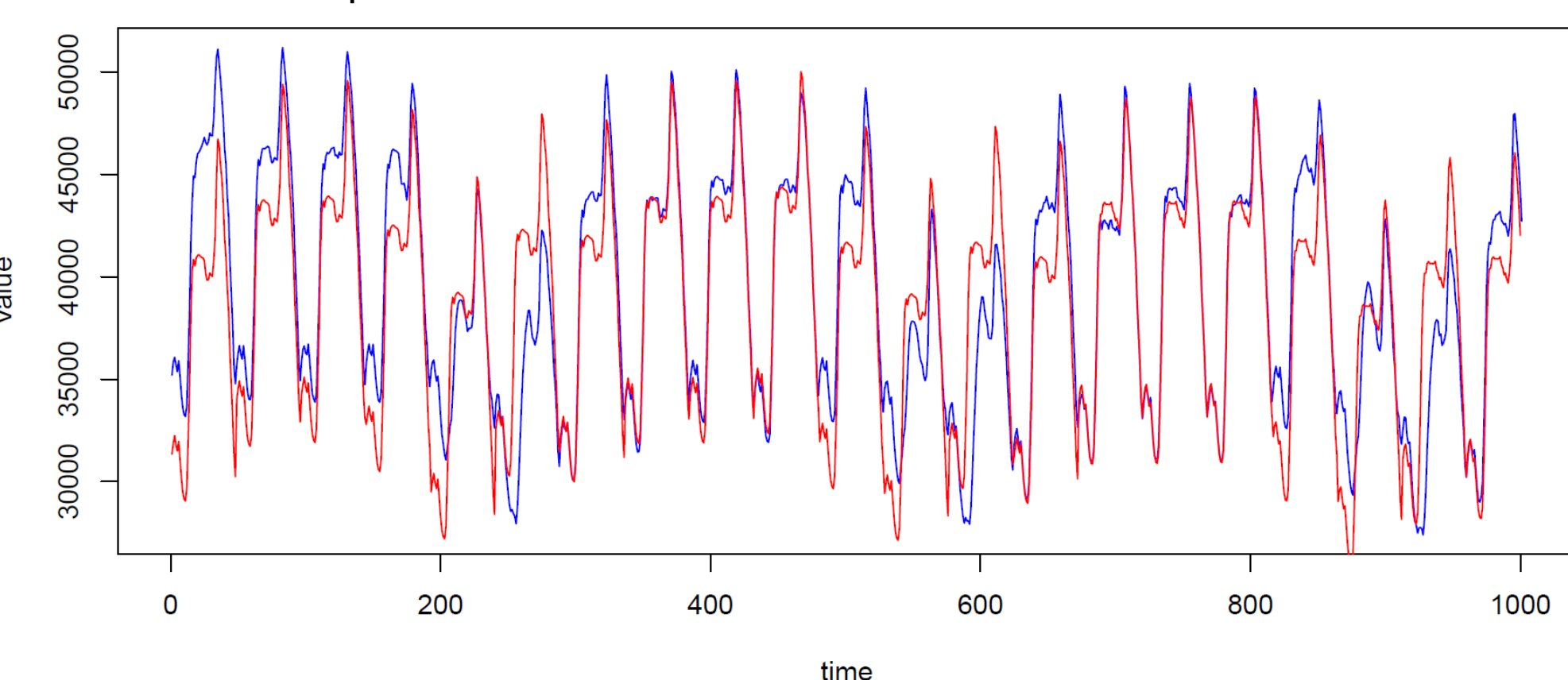
25 FTT Components 20 times compression
Actual Vs FFT Estimated

50 FTT Components 10 times compression
Actual Vs FFT Estimated

125 FTT Components 4 times compression
Actual Vs FFT Estimated

2 Weeks Holt Winters 8.6 times compression
Actual Vs HW Estimated

Peaks and Valleys 3.4 times compression
Actual Vs Peaks n Valleys
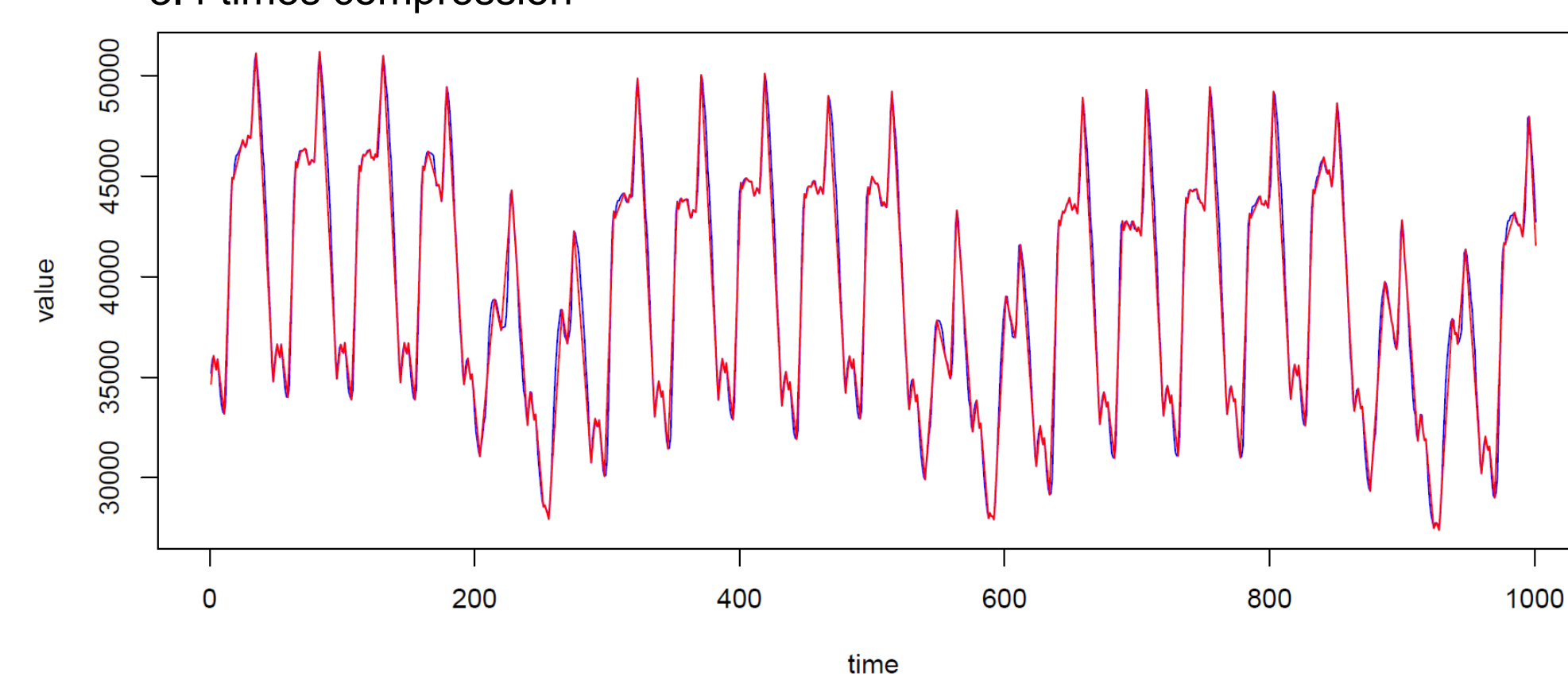
10 window Moving Average 50 FFT Components
MA Vs FFT

TECHNISCHE UNIVERSITÄT DRESDEN

http://wwwdb.inf.tu-dresden.de/research

Database Technology Group