Data Analytics using KNIME open source tool

Team: Olga Bode, Eric Dörheit, Dorothea Poljak

Supervisor: Dr. Marcela Charfuelan

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/

Overview

1. Handling of big files
2. Extract, Transform, Load (ETL) with Knime
3. Approaches to detect anomalies

Handling of big files

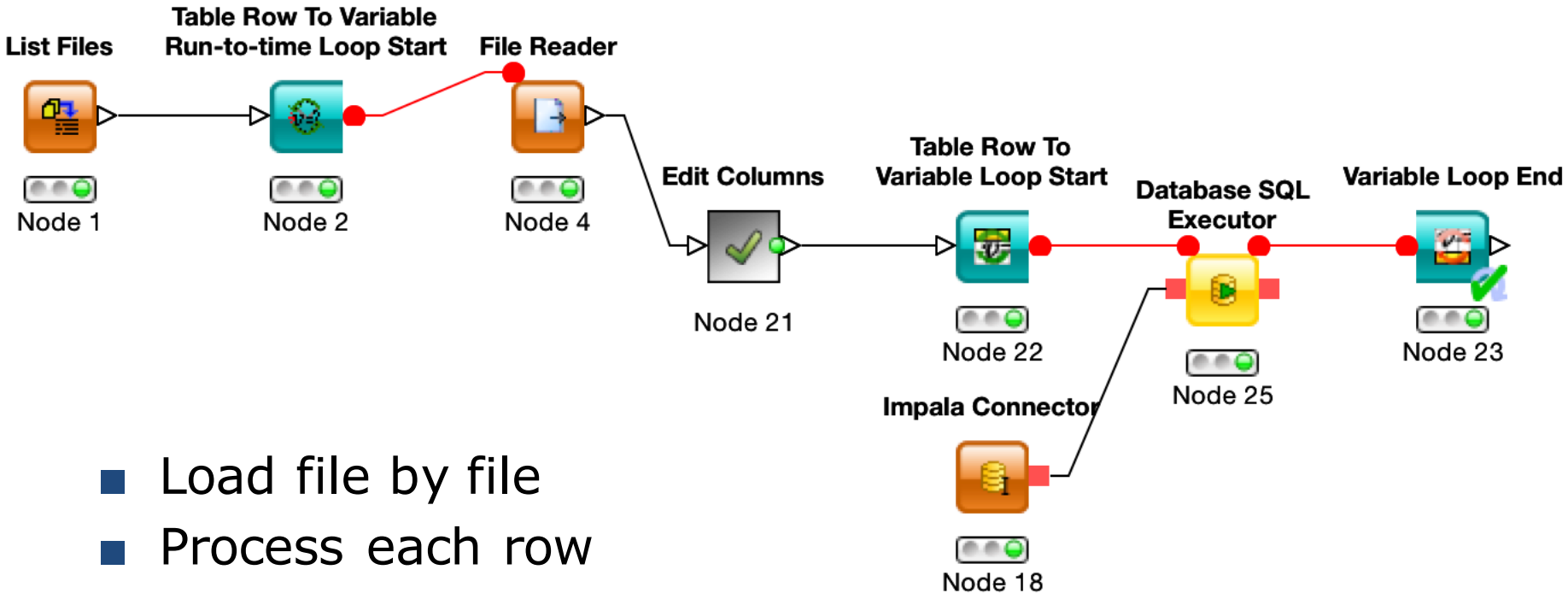
- Dataset does not fit into memory
- File (~6GB) not yet imported in Database, HDFS etc.

→ Split file into smaller files

→ Iterate over file list and process each file

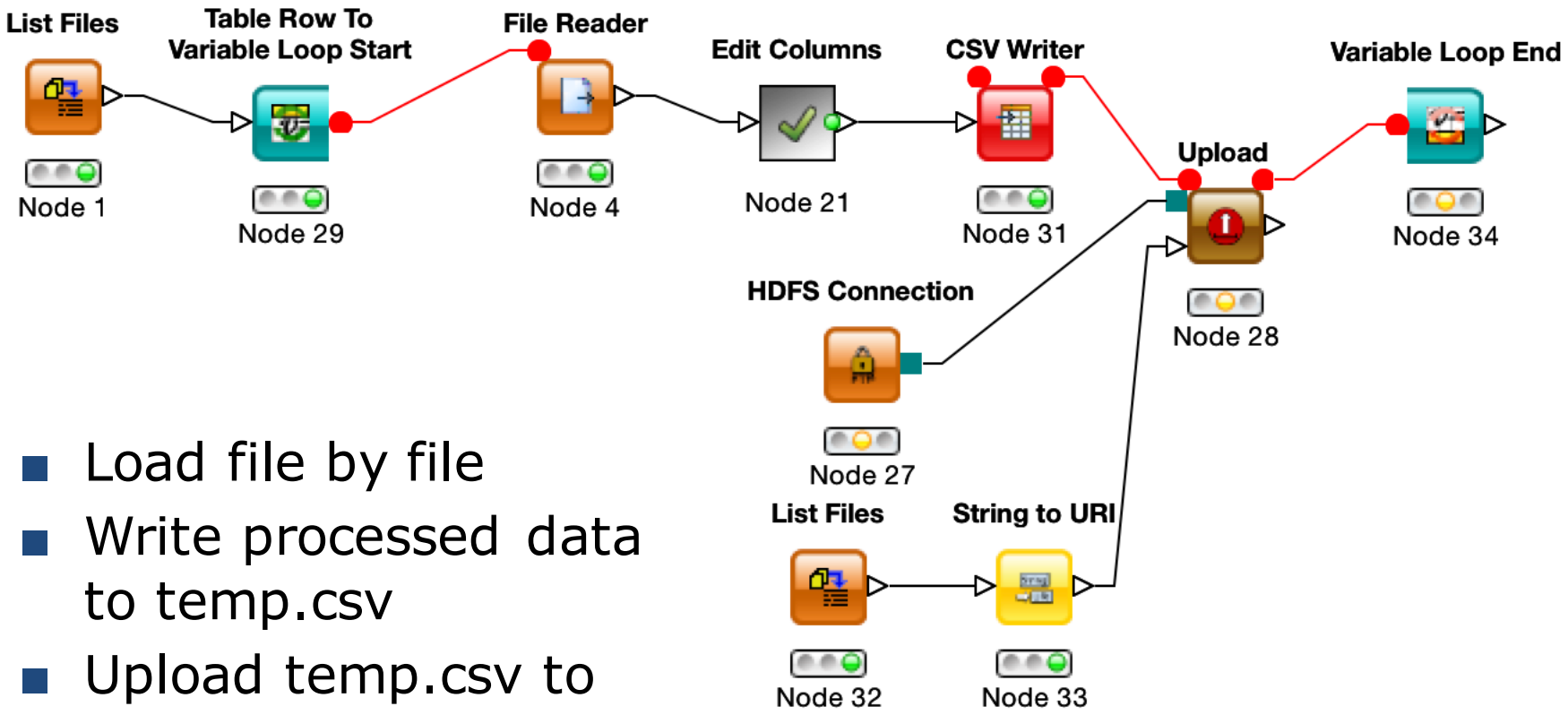
→ Import small files to Impala, Database etc.

Handling of big files



- Load file by file
- Process each row
- Insert row into Impala table

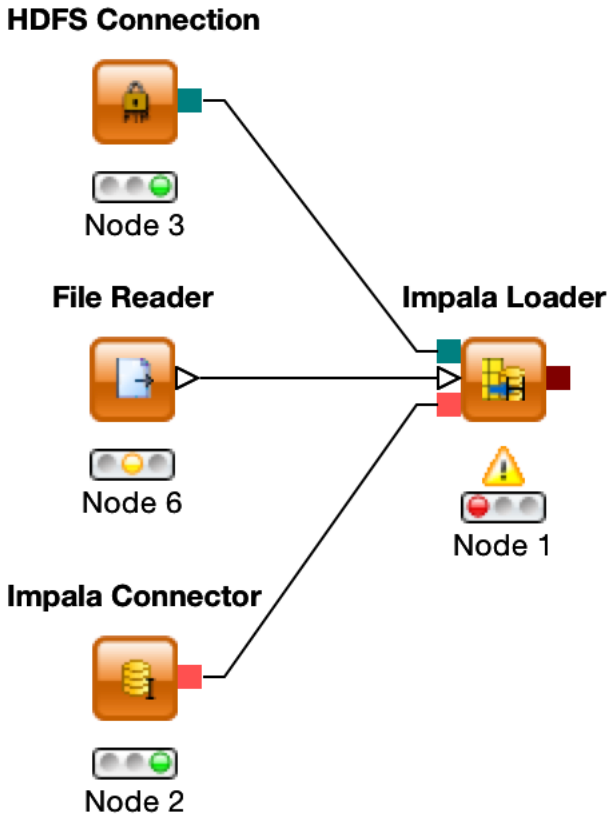
Handling of big files



- Load file by file
- Write processed data to temp.csv
- Upload temp.csv to HDFS

Handling of big files

- Write file temporarily to HDFS
- Import file into Impala
- Columns of input file must not be named as SQL commands!



Dataset

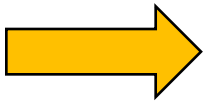
ts	index	mf01	mf02	mf03	pc13	pc14	pc15	pc25	pc26	pc27	res	bm05	bm06	bm07	bm08	bm09	bm10	pp01	pp02
pp03	pp04	pp05	pp06	pp07	pp08	pp09	pp10	pp11	pp12	pp13	pp14	pp15	pp16	pp17	pp18	pp19	pp20	pp21	pp31
pp32	pp33	pp34	pp35	pp36	pc01	pc02	pc03	pc04	pc05	pc06	pc19	pc20	pc21	pc22	pc232012-02-22T17:22:36.4814244+00:0				
2772001	13781	14917	8645	0071	0186	24508	0000	0000	0000	0000	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
2012-02-22T17:22:36.4914240+00:00					2772002	13783	14918	8645	0070	0189	24506	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5014147+00:00					2772003	13784	14921	8645	0070	0187	24507	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5114943+00:00					2772004	13784	14923	8645	0071	0190	24506	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5214255+00:00					2772005	13787	14924	8644	0069	0192	24507	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5314238+00:00					2772006	13789	14928	8641	0071	0191	24510	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5414222+00:00					2772007	13789	14931	8639	0076	0186	24514	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5514250+00:00					2772008	13791	14933	8637	0075	0186	24516	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5614259+00:00					2772009	13796	14937	8637	0072	0191	24512	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5714332+00:00					2772010	13797	14938	8638	0075	0192	24510	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5814361+00:00					2772011	13795	14939	8638	0076	0194	24509	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.5914254+00:00					2772012	13797	14942	8638	0076	0192	24510	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.6014296+00:00					2772013	13803	14946	8638	0076	0192	24512	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.6114298+00:00					2772014	13803	14949	8638	0078	0194	24508	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.6214307+00:00					2772015	13800	14952	8635	0078	0197	24506	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2012-02-22T17:22:36.6314163+00:00					2772016	13803	14955	8638	0078	0194	24507	0000	0000	0000	0000	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0

Challenge:

1. filter out the columns that do not change

2. process the first column, ts, to split the value in various columns containing: year, month, day, hour, minute, second

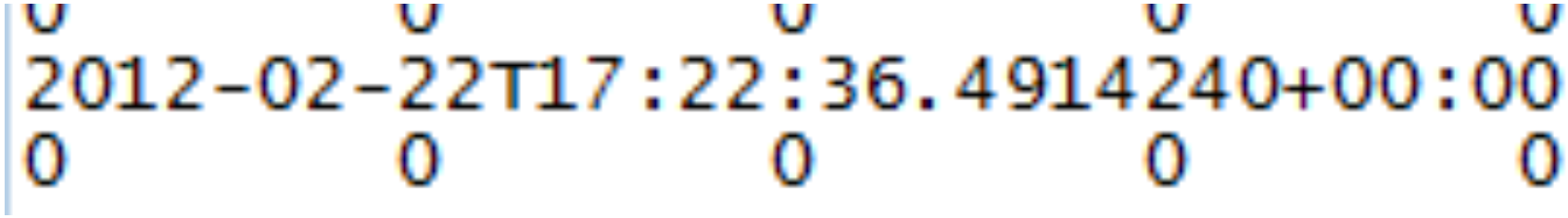
3. process the data, numerical columns, not the binary ones, and generate averages per second, save the data in another file



Data pre-processing

Timestamp structure

ts:

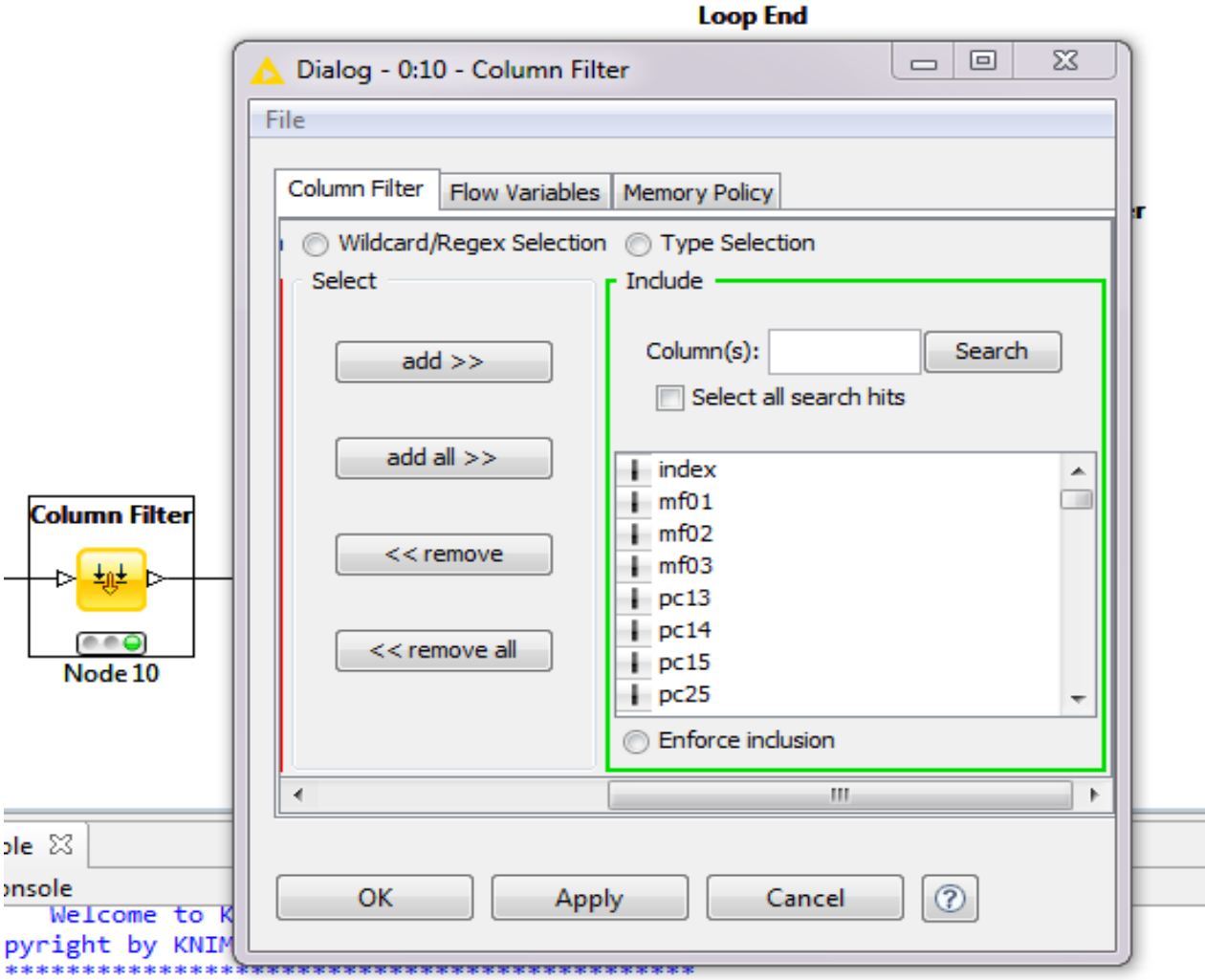


Column Filter

Challenge:

1. filter out the columns that do not change

-> **Column Filter.**

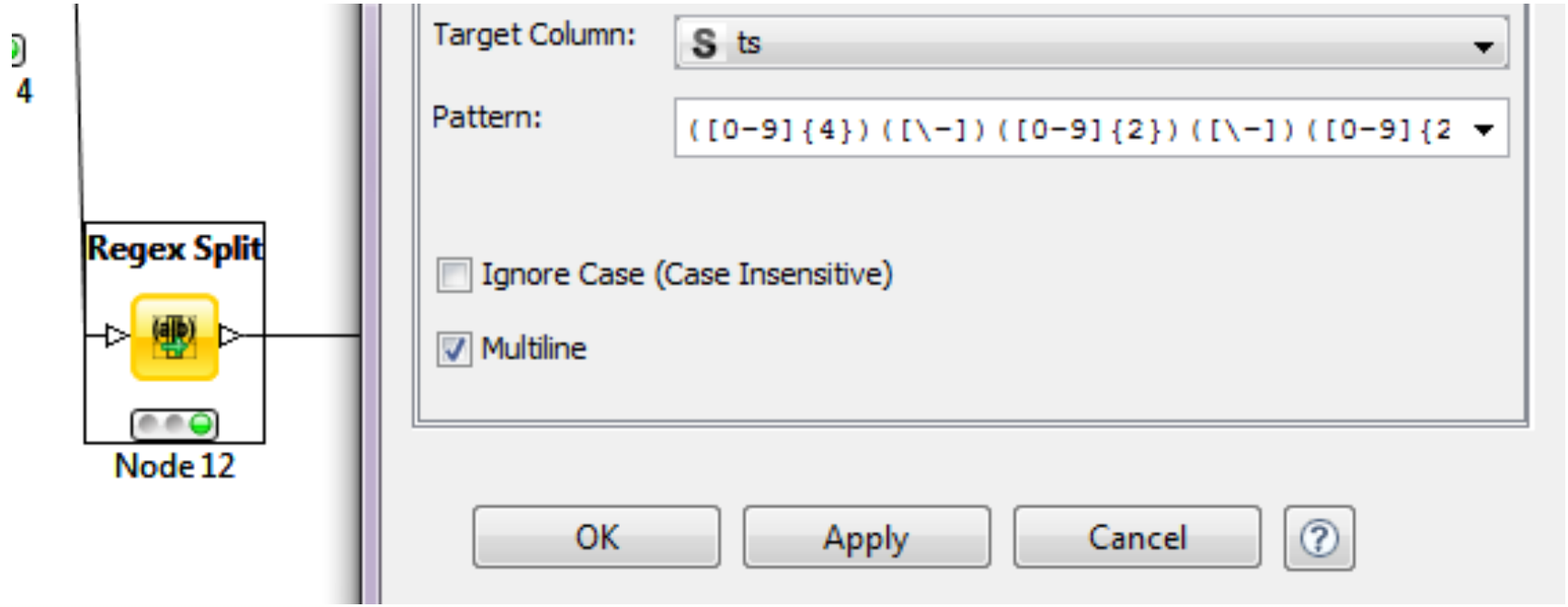


Split Timestamp

Challenge:

2. process the first column, ts, to split the value in various columns containing: year, month, day, hour, minute, second

-> Regex Split

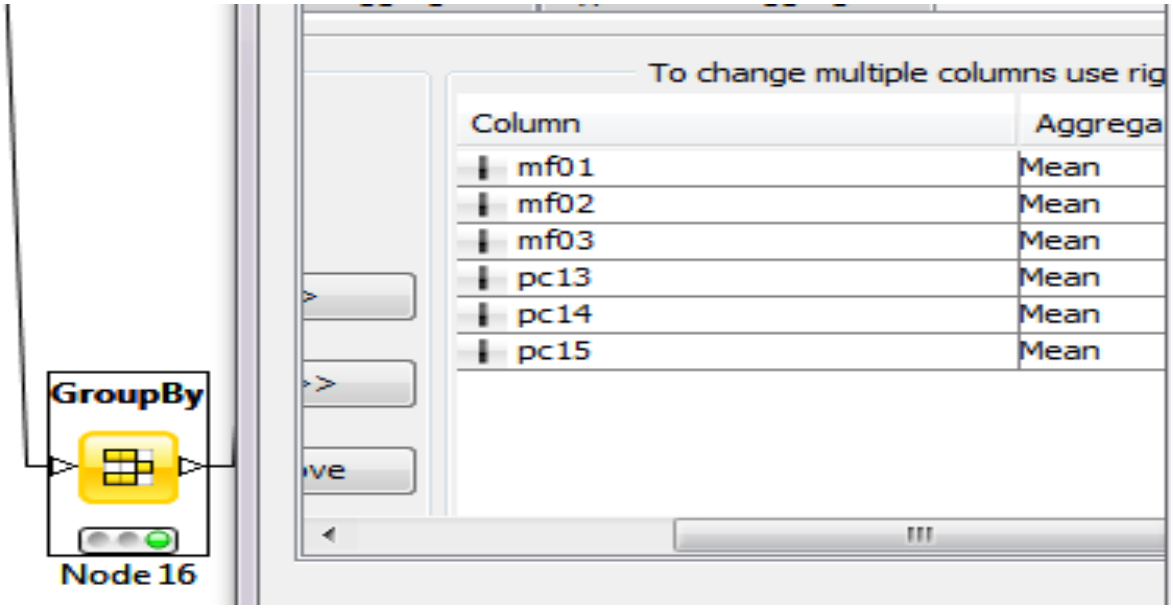


Group and compute average

Challenge:

3. process the data, numerical columns, not the binary ones, and generate averages per second, save the data in another file

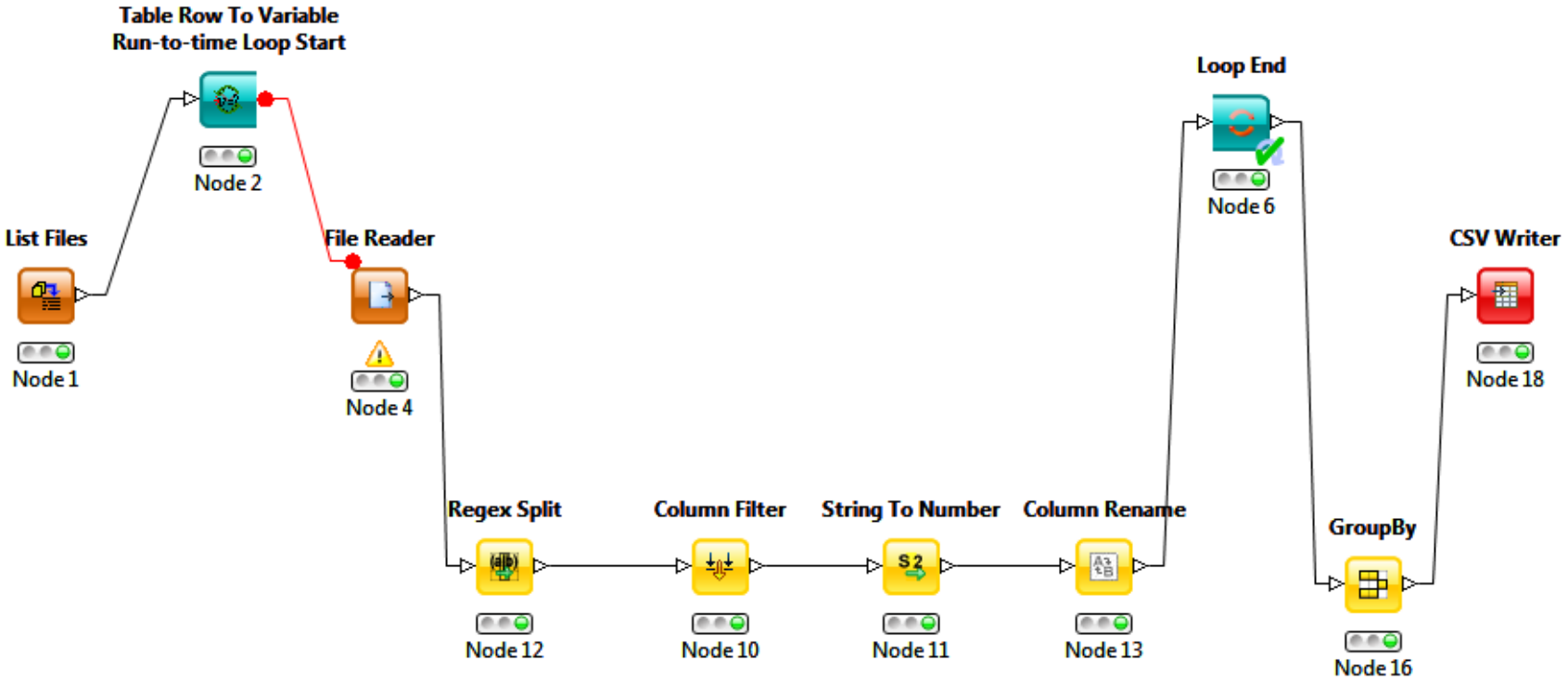
-> GroupBy and Aggregation



New table:

year	month	day	hour	min	sec	Mean(...	Mean(...	Mean(...	Mean(p...	Mean(p...	Mean(p...
2012	2	22	17	22	36	13,800.212	14,957.365	8,633.25	75.231	188.596	21,647.135

Workflow:



What are anomalies?

Anomalies:

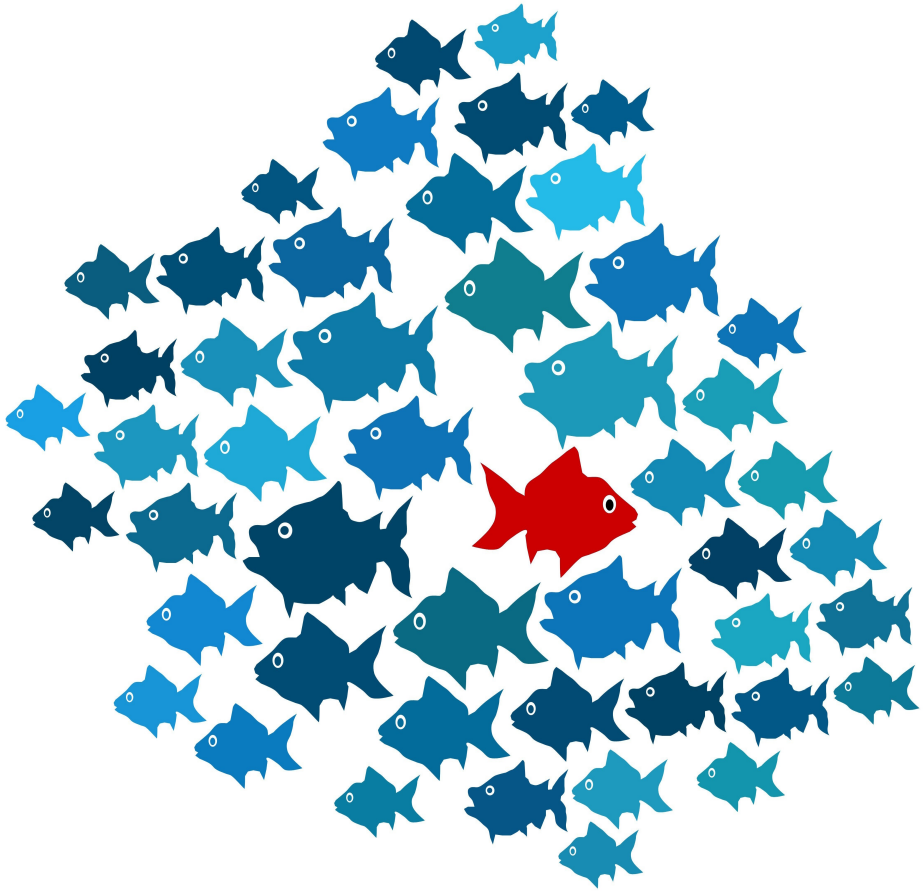
patterns in data that do not conform to a well defined notion of normal behavior.

Interestingness of anomaly:

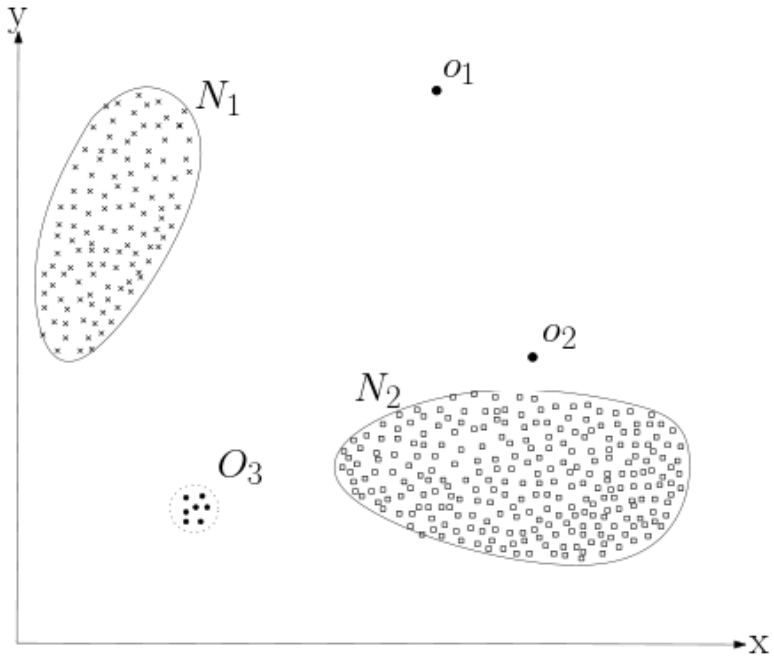
- breakdown of a system
- intrusion

Anomaly detection approach:

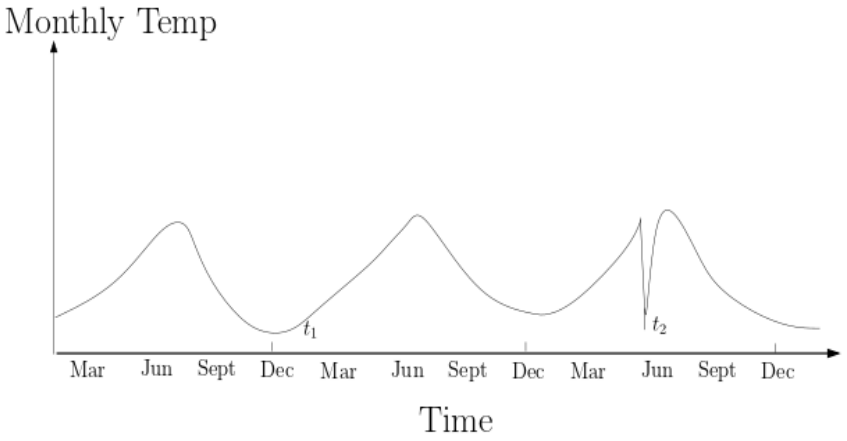
to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an anomaly.



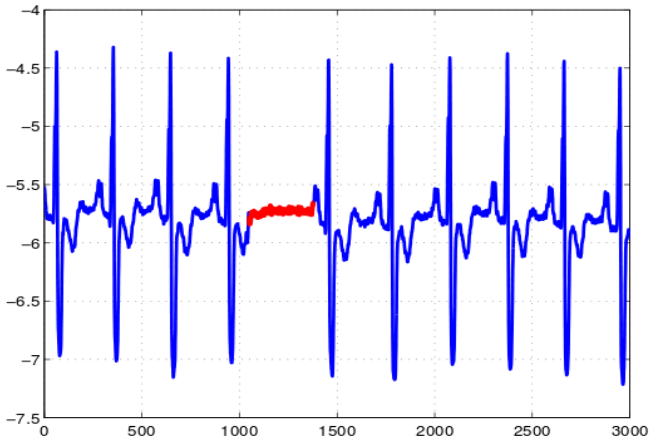
Type of anomaly



Pic 1. Point Anomalies in a 2-dimensional data set

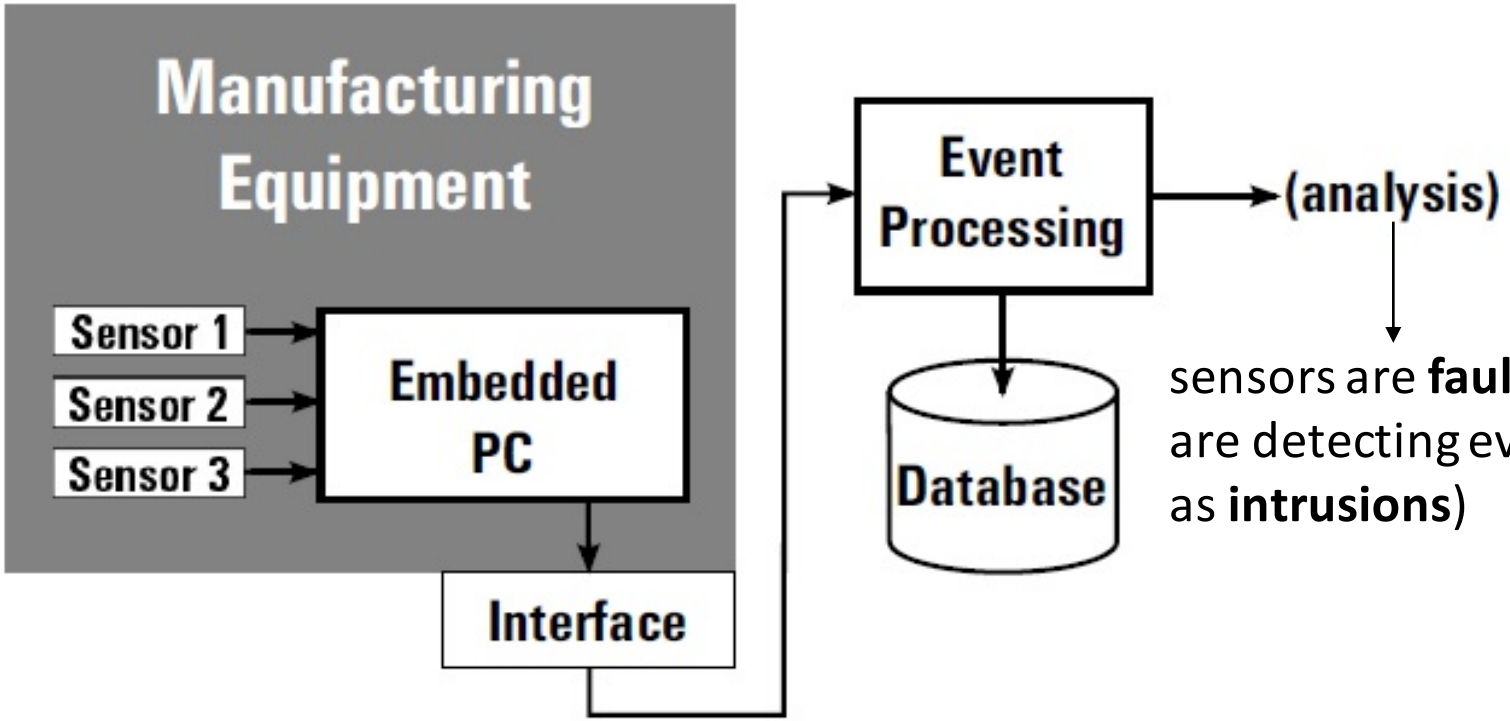


Pic 2. Contextual anomaly t2
in a temperature time series.



Pic 3. Collective anomaly in an human electrocardiogram output

Sensor Networks: streaming mode

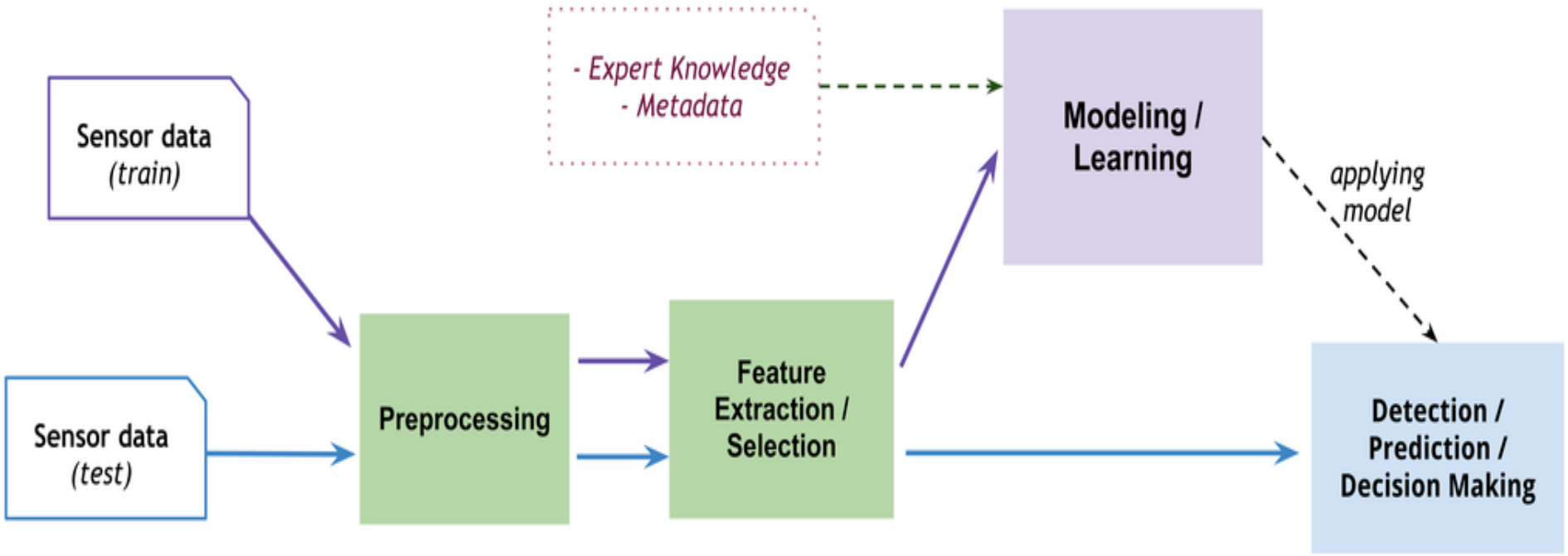


sensors are **faulty**, or they are detecting events (such as **intrusions**)

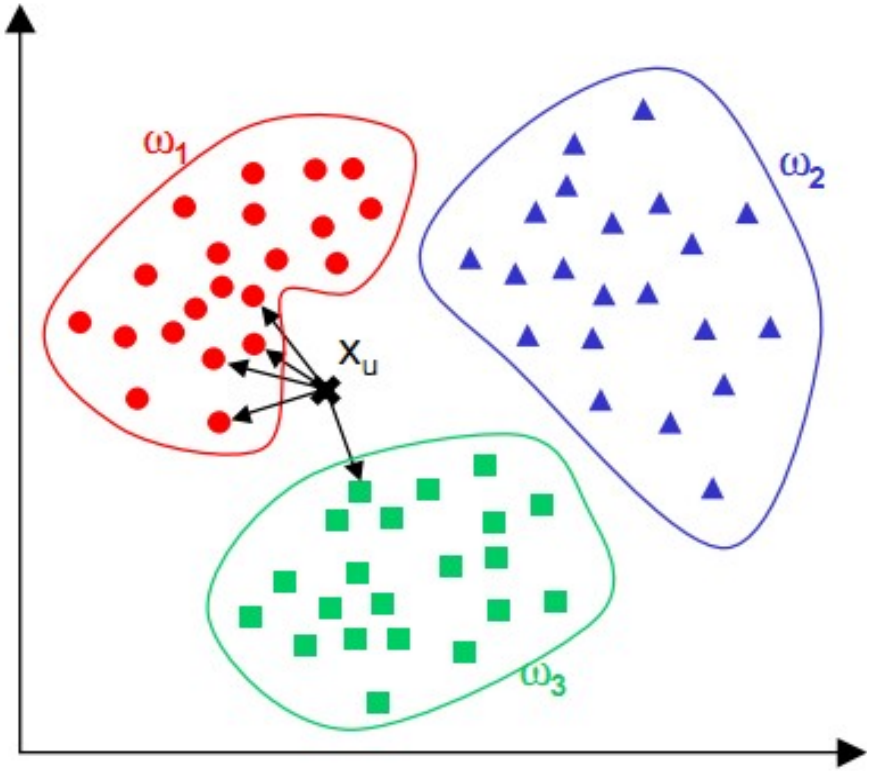
Anomaly Detection Techniques

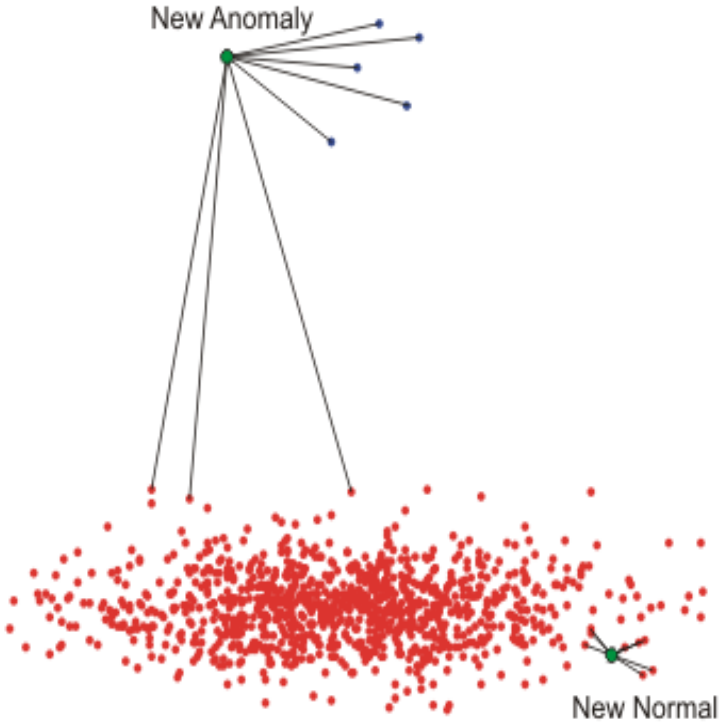
- Classification
- Nearest Neighbor
- Clustering

Classification

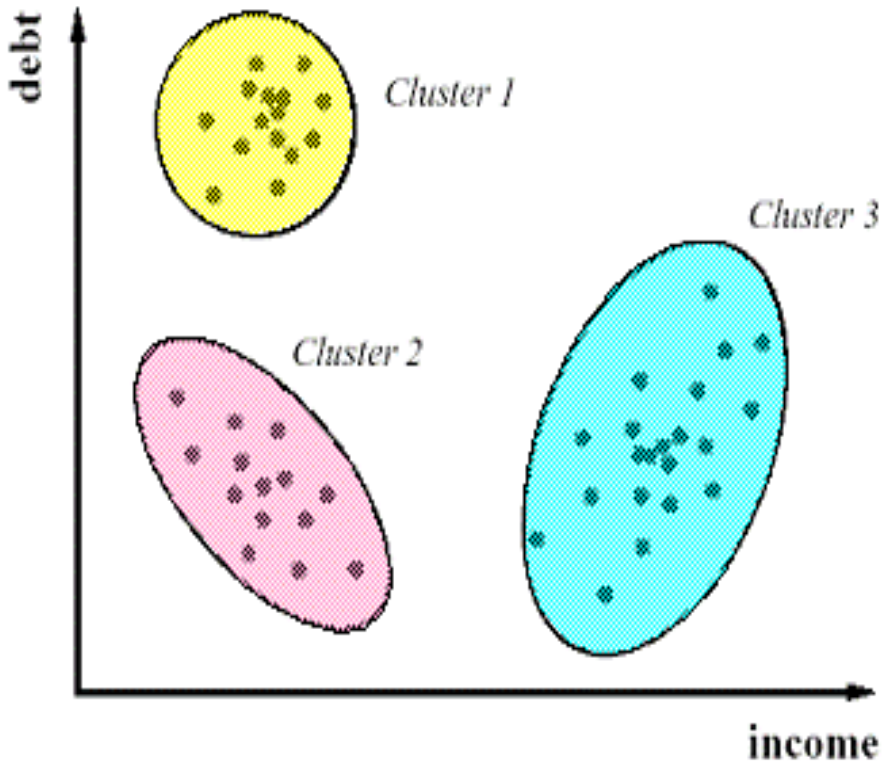


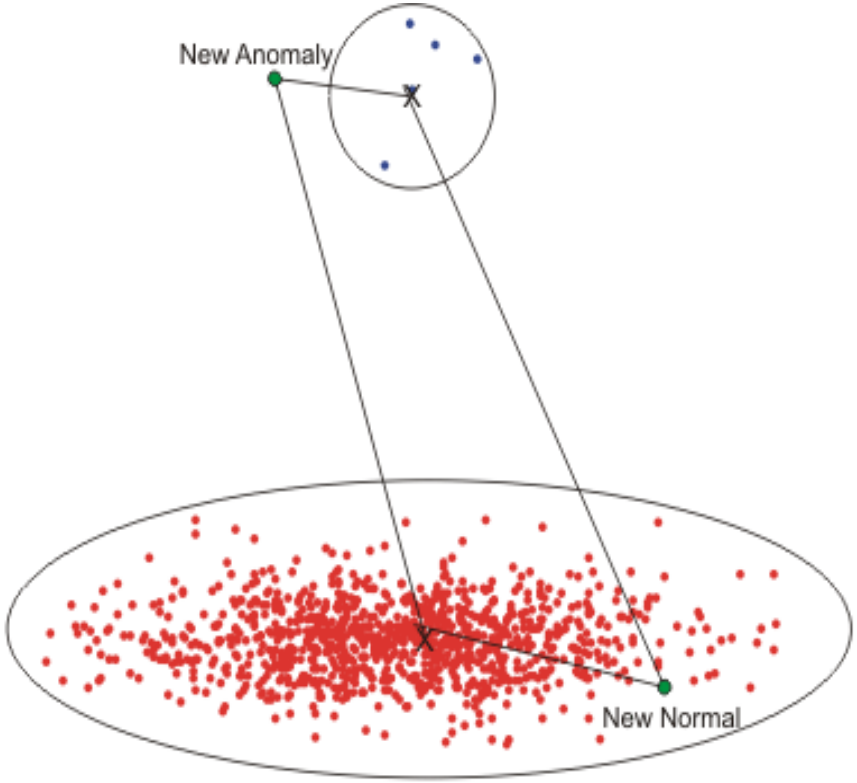
K-nearest Neighbor (KNN)





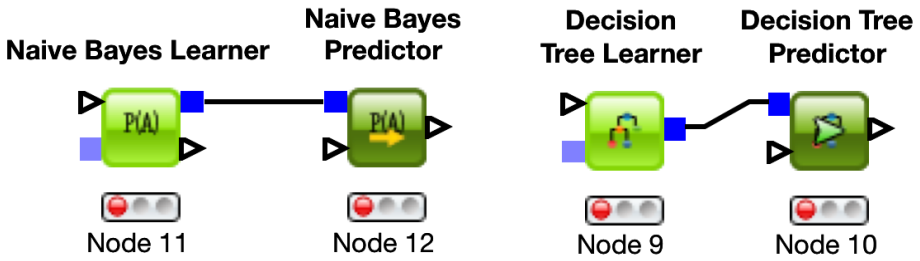
Clustering


debt
Cluster 1
Cluster 2
Cluster 3
income


New Anomaly
New Normal

Next Steps

- Apply methods of anomaly detection on DEBS Grand Challenge dataset with Knime:
  - Classification



  - K-nearest Neighbor

  - Clustering