Summer semester / 2016

universitätbonn

Outlier Detection on Financial RDF Data

Mentor: Christiane Engels

Group: Zuhair Almhithawi, Nayef Roqaya, Berivan Ekmez

Final Presentation

Semantic Web

Data Mining
Open Data

Outline

1) Motivation
2) Project overview
3) Requirements
4) System architecture
5) Results
6) Demo

Motivation

- Open Government and Data Transparency initiatives => increasing number of datasets
- automatically analyze data sets.

One aspect of analyzing data sets is finding unusual values, i.e. outliers

Project overview

Outlier Detection on Financial RDF Data

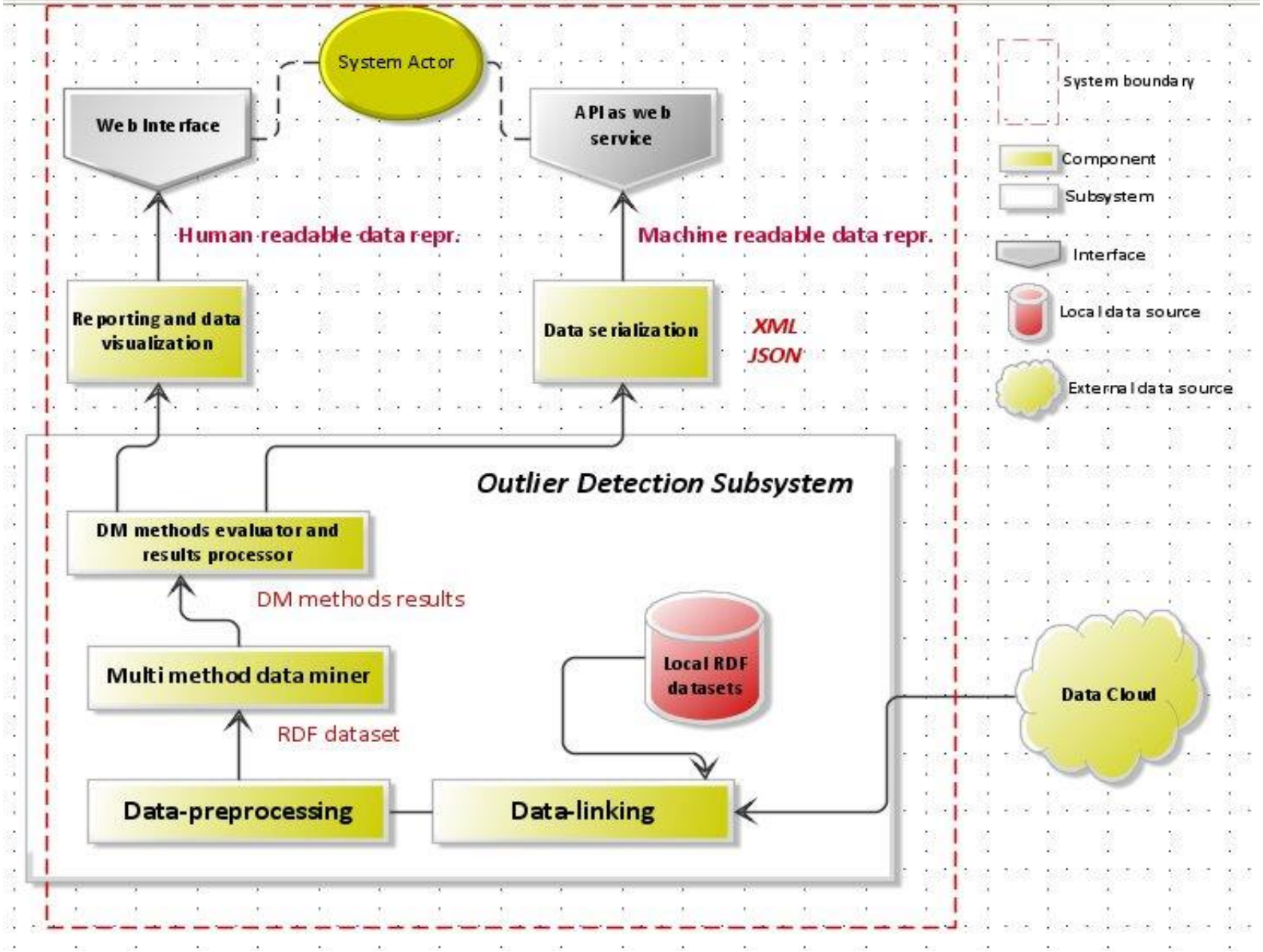
Objectives:

- ✓ Provide more insight into financial data by finding unusual values, i.e. outliers or anomalies:
  - ➢ irregular behavior (corruption, fraud, . . .)
  - ➢ regions of special interest that e.g. require more subsidies
- ✓ Compare results of different outlier detection methods.

Requirements

- User can import RDF dataset(s):
  - Data Cube vocabulary
- Link local datasets to external dataset cloud:
  - DBPedia
- Apply and compare different outlier detection methods:
  - K-means
  - Chauvenet's Criterion
- Visualize the results:
  - Google Maps

System architecture



System architecture – data-linking module

Enrichment:

Why?

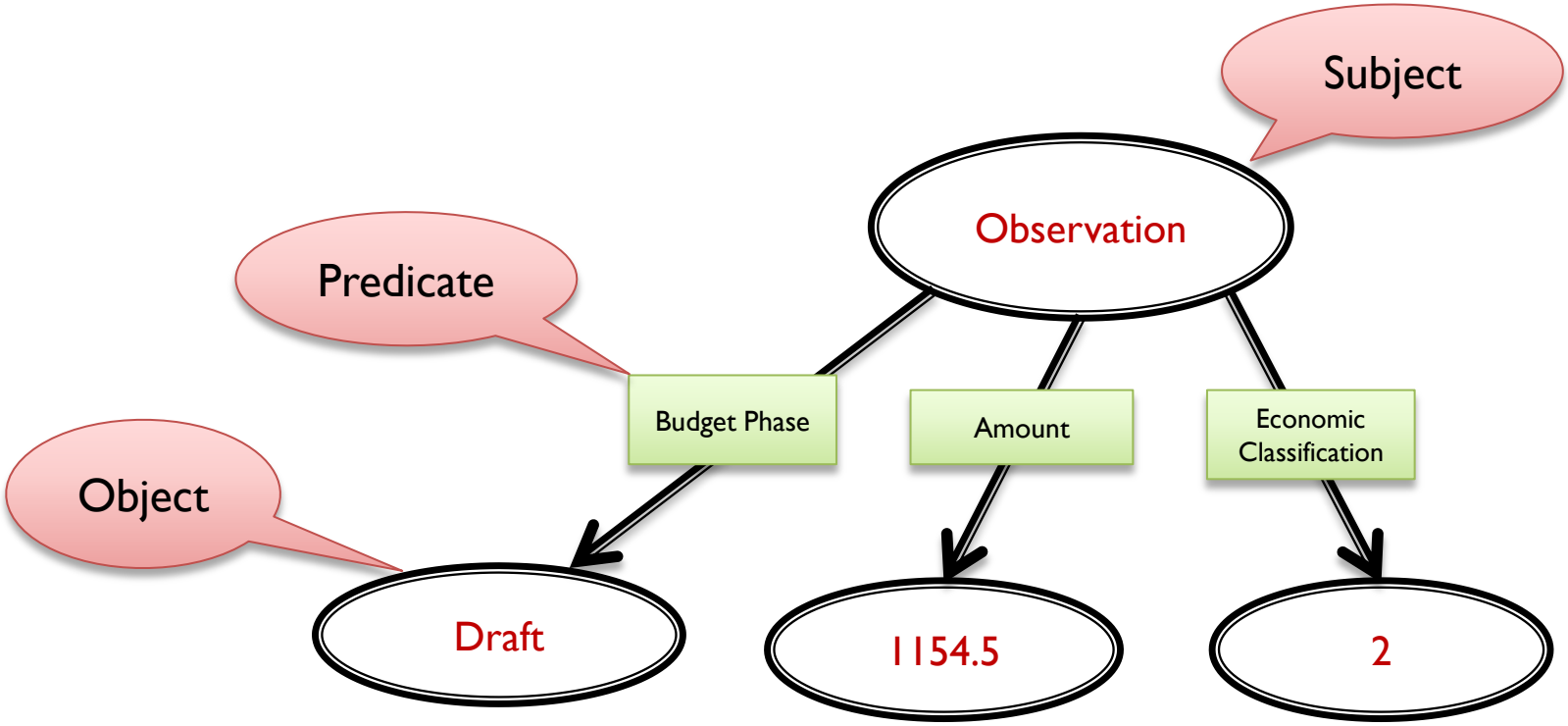
- ✓ More accurate subpopulation outcomes.

How?

- ❖ DBPedia as an external data repository.
- ❖ SPARQL queries performed on Fuseki Jena server.

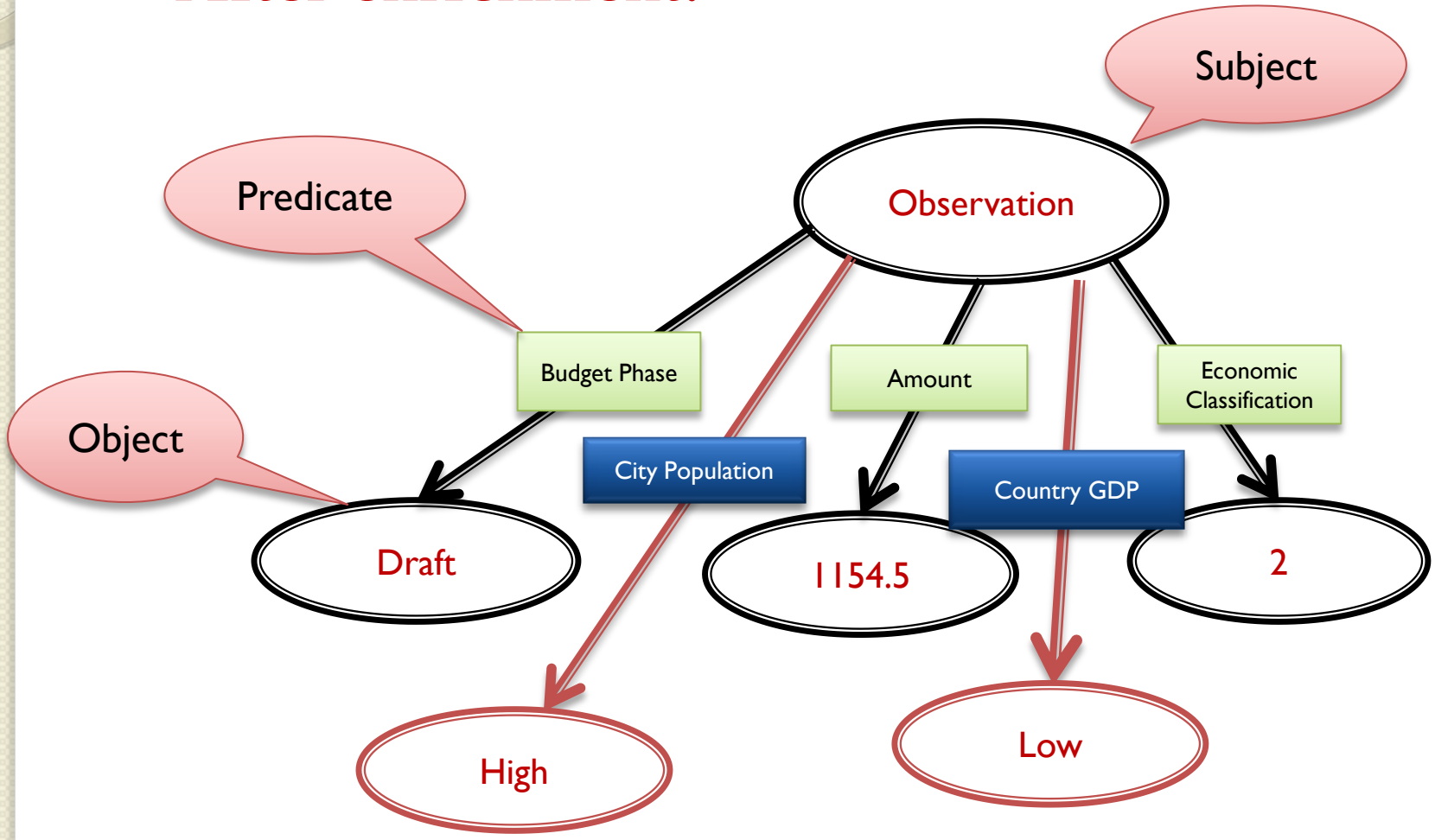
System architecture – data-linking module

- Before enrichment:



System architecture – data-linking module

- After enrichment:



System architecture – data pre-processing module

- Sub population :

Why?

- Run outlier detection on original dataset !!?

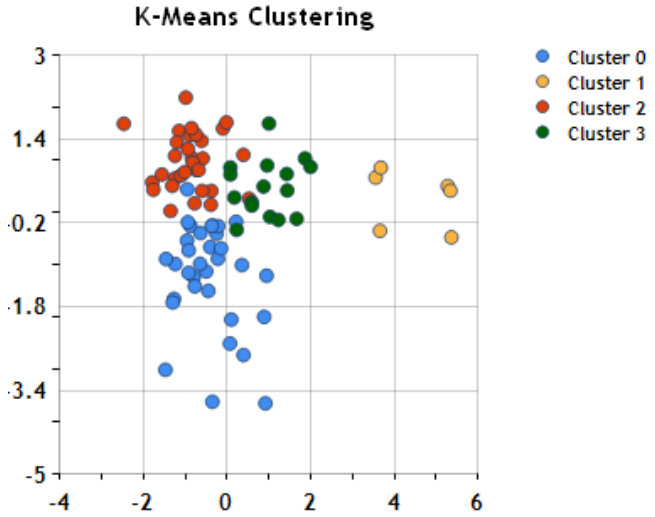
- Separation of original dataset into fragments gives more precise outlier results.

- Pruning when: No reducing or low KLV or number of instances is low

System architecture – outlier detection module

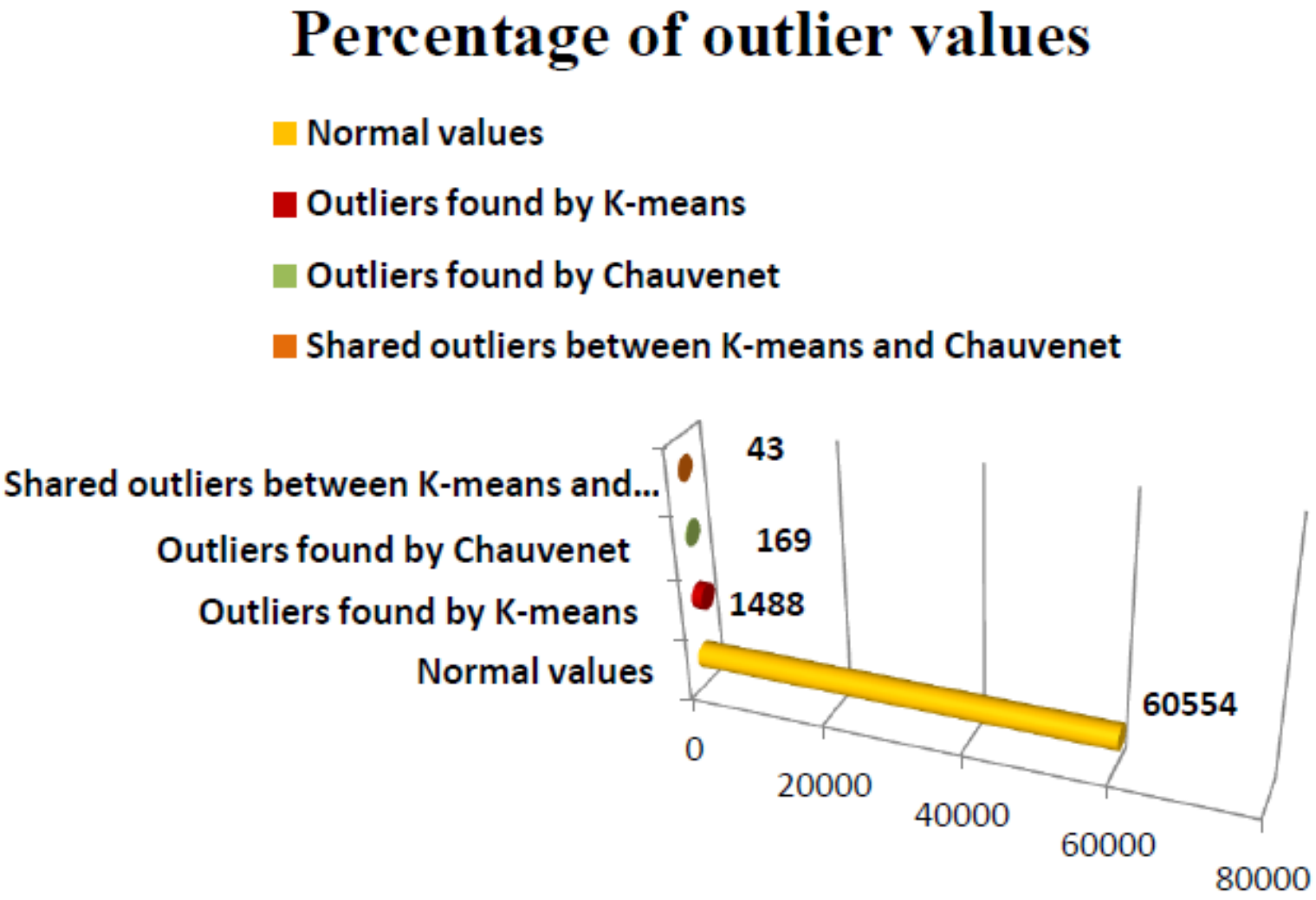
- Applied outlier detection methods:

K- Means

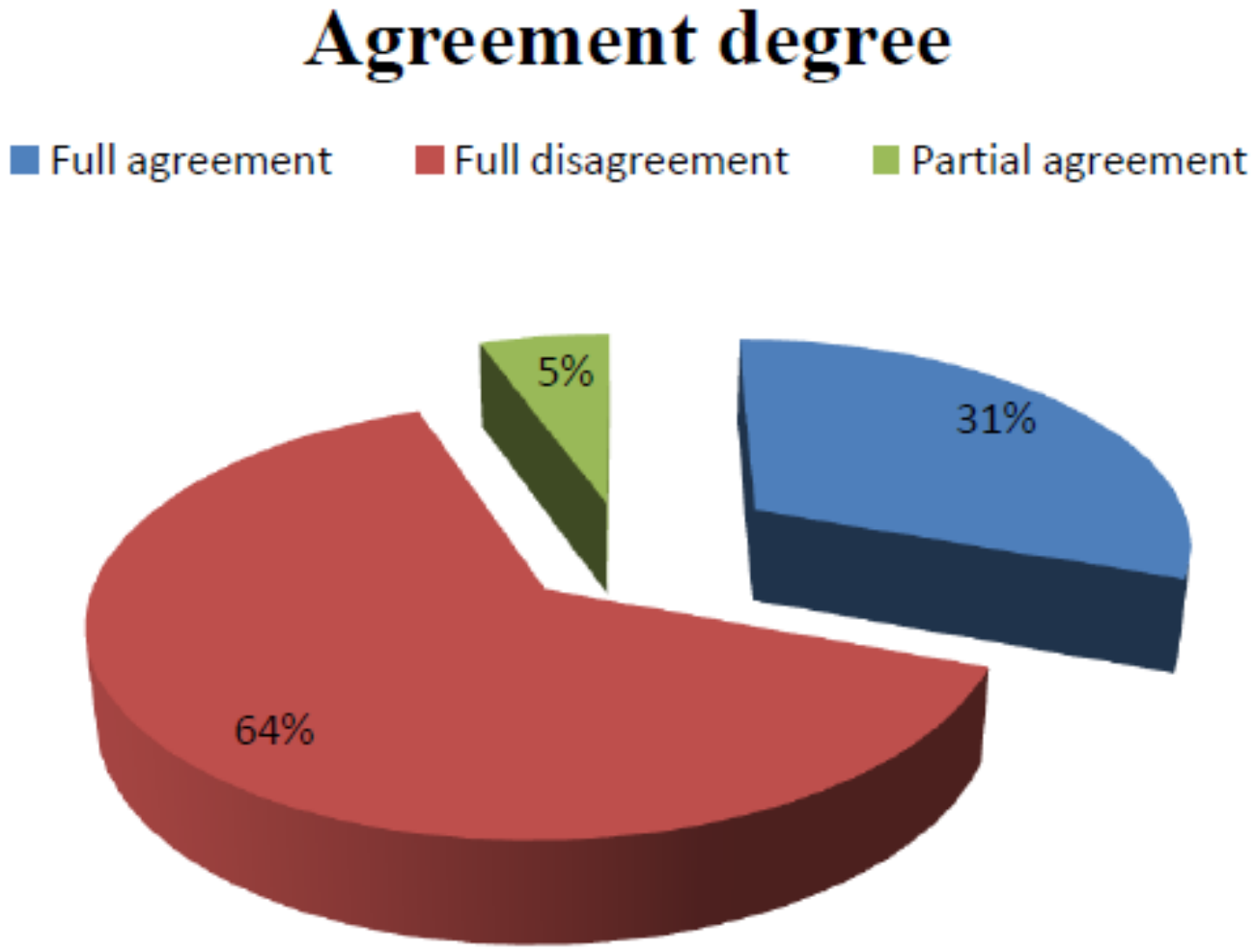


Chauvenet's Criterion

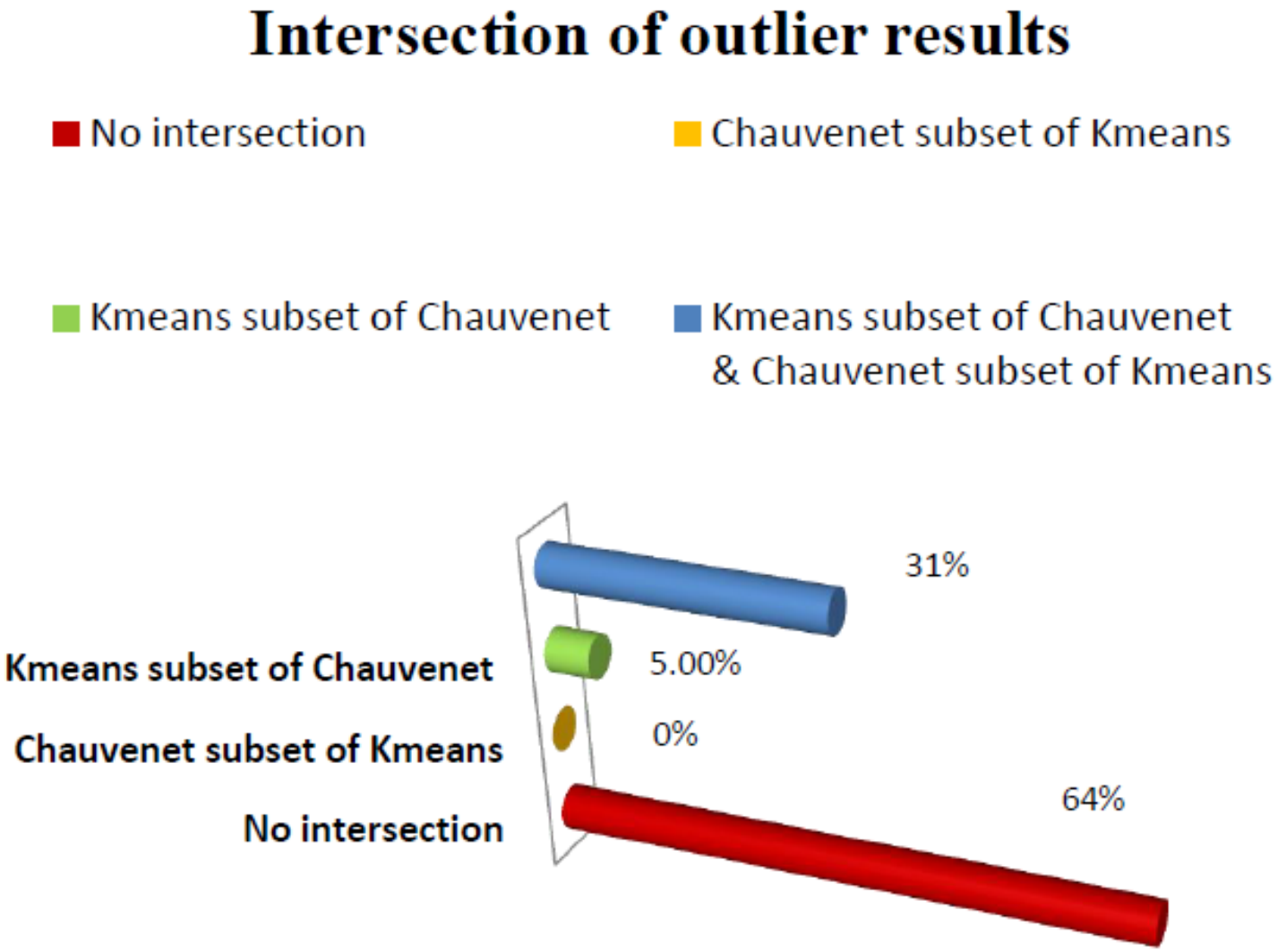
Results


Percentage of outlier values
Normal values
Outliers found by K-means
Outliers found by Chauvenet
Shared outliers between K-means and Chauvenet
Shared outliers between K-means and...
Outliers found by Chauvenet
Outliers found by K-means
Normal values
43
169
1488
60554
0
20000
40000
60000
80000

Results


Agreement degree
Full agreement
Full disagreement
Partial agreement
5%
31%
64%

Results


Intersection of outlier results
No intersection
Chauvenet subset of Kmeans
Kmeans subset of Chauvenet
Kmeans subset of Chauvenet & Chauvenet subset of Kmeans
31%
5.00%
0%
64%
Kmeans subset of Chauvenet
Chauvenet subset of Kmeans
No intersection

Demo