universitätbonn

Anomaly Detection in Numerical Financial Data

Christiane Engels

Enterprise Information Systems Group,
University of Bonn

07.04.2016

1. Motivation
2. Preliminaries & Related Work
3. Approach
4. Evaluation
5. Conclusion and Future Work

Motivation

- Open Government and Data Transparency initiatives
- Governments are forced by law to open (part of) their data by Freedom of Information laws in many countries
- Number of open data sets is constantly increasing

$\rightarrow$ There is a need to automatically analyze these data sets

Analyzing financial data

- One aspect of analyzing financial data is finding *unusual* values, i.e. *outliers* or *anomalies*.
- These may indicate:
  - errors in the data
  - irregular behavior (corruption, fraud, . . .)
  - regions of special interest that e.g. require more subsidies or a better handling of those

Outlier/Anomaly Detection

- "Anomaly detection refers to the problem of finding patterns in data that do not conform to the expected normal behavior."
  - Chandola et al. (2009)

Outlier/Anomaly Detection

- "Anomaly detection refers to the problem of finding patterns in data that do not conform to the expected normal behavior."
  - Chandola et al. (2009)
- General idea:

  Define a model representing normal behavior and declare any observation in the data that does not fit the model as an anomaly.

Outlier/Anomaly Detection

- Different approaches
  - Classification-based approaches
  - Statistical approaches
  - Clustering-based approaches
  - Density-based approaches
- Different categories
  - Supervised, semi-supervised, unsupervised
  - Depending on type of required labeled training data
- Different outputs
  - Binary label or outlier score

Anomaly Detection on Financial Data

- Large data sets
- Special structure
  - Amount of money spent/received is the target attribute for anomaly detection
  - Contextual attributes classifying the amount of money spent/received

Example: EU Funds

State	Region	Fund	Objective	EUAmount
Germany	Bavaria	EAFRD	Research & Innovation	2,000,000
Germany	Bavaria	EAFRD	Environment Protection	516,088,338
Germany	Bavaria	EAFRD	Social Inclusion	146,000,000
Germany	Bavaria	ESF	Social Inclusion	78,600,000
...	...	...	...	...

Example: EU Funds

universitätbonn

State	Region	Fund	Objective	EUAmount
Germany	Bavaria	EAFRD	Research & Innovation	2,000,000
Germany	Bavaria	EAFRD	Environment Protection	516,088,338
Germany	Bavaria	EAFRD	Social Inclusion	146,000,000
Germany	Bavaria	ESF	Social Inclusion	78,600,000
Germany	Saxony	EAFRD	Research & Innovation	7,872,000
Germany	Saxony	EAFRD	Environment Protection	126,370,270
Germany	Saxony	EAFRD	Social Inclusion	364,342,018
Germany	Saxony	ESF	Social Inclusion	206124996
France	French Guiana	EAFRD	Research & Innovation	10,070,000
France	French Guiana	EAFRD	Environment Protection	3,014,540
France	French Guiana	EAFRD	Social Inclusion	46,440,000
France	French Guiana	ESF	Social Inclusion	34,529,070
France	Reunion	EAFRD	Research & Innovation	51,470,000
France	Reunion	EAFRD	Environment Protection	99,825,000
France	Reunion	EAFRD	Social Inclusion	48,400,000
France	Reunion	ESF	Social Inclusion	100,786,000
...	...	...	...	...

Example: EU Funds

universitätbonn

State	Region	Fund	Objective	EUAmount
Germany	Bavaria	EAFRD	Research & Innovation	2,000,000
Germany	Bavaria	EAFRD	Environment Protection	516,088,338
Germany	Bavaria	EAFRD	Social Inclusion	146,000,000
Germany	Bavaria	ESF	Social Inclusion	78,600,000
Germany	Saxony	EAFRD	Research & Innovation	7,872,000
Germany	Saxony	EAFRD	Environment Protection	126,370,270
Germany	Saxony	EAFRD	Social Inclusion	364,342,018
Germany	Saxony	ESF	Social Inclusion	206124996
France	French Guiana	EAFRD	Research & Innovation	10,070,000
France	French Guiana	EAFRD	Environment Protection	3,014,540
France	French Guiana	EAFRD	Social Inclusion	46,440,000
France	French Guiana	ESF	Social Inclusion	34,529,070
France	Reunion	EAFRD	Research & Innovation	51,470,000
France	Reunion	EAFRD	Environment Protection	99,825,000
France	Reunion	EAFRD	Social Inclusion	48,400,000
France	Reunion	ESF	Social Inclusion	100,786,000
...	...	...	...	...

Example: EU Funds

universitätbonn

State	Region	Fund	Objective	EUAmount
Germany	Bavaria	EAFRD	Research & Innovation	2,000,000
Germany	Bavaria	EAFRD	Environment Protection	516,088,338
Germany	Bavaria	EAFRD	Social Inclusion	146,000,000
Germany	Bavaria	ESF	Social Inclusion	78,600,000
Germany	Saxony	EAFRD	Research & Innovation	7,872,000
Germany	Saxony	EAFRD	Environment Protection	126,370,270
Germany	Saxony	EAFRD	Social Inclusion	364,342,018
Germany	Saxony	ESF	Social Inclusion	206124996
France	French Guiana	EAFRD	Research & Innovation	10,070,000
France	French Guiana	EAFRD	Environment Protection	3,014,540
France	French Guiana	EAFRD	Social Inclusion	46,440,000
France	French Guiana	ESF	Social Inclusion	34,529,070
France	Reunion	EAFRD	Research & Innovation	51,470,000
France	Reunion	EAFRD	Environment Protection	99,825,000
France	Reunion	EAFRD	Social Inclusion	48,400,000
France	Reunion	ESF	Social Inclusion	100,786,000
...	...	...	...	...

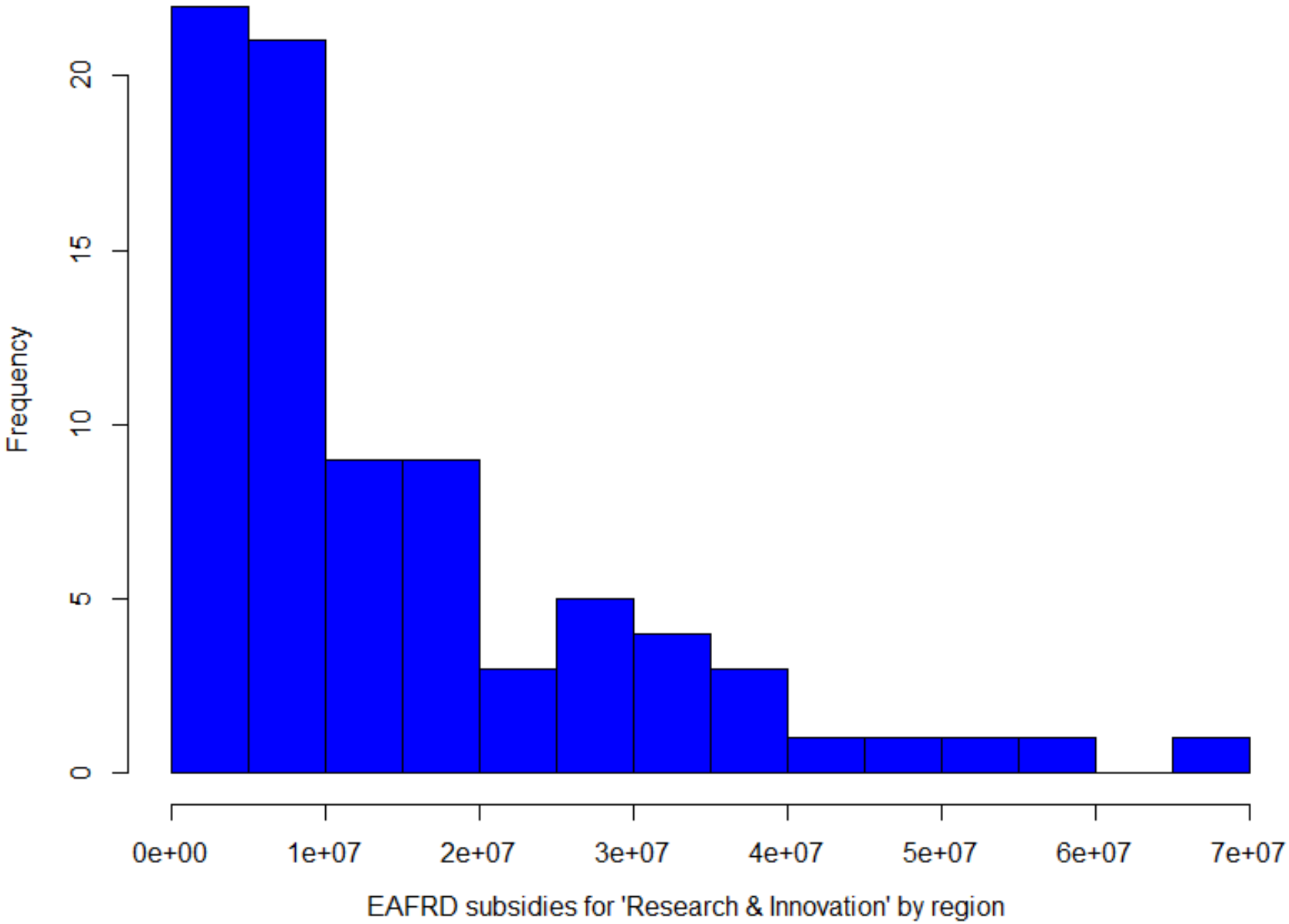
→ Approach: Apply outlier detection to appropriate subgroups of the data set comprising comparable items.

How to define appropriate subgroups?

- Fix a specific value (or range of values) for one or more dimensions
- This corresponds to slicing the data cube in the OBEU RDF data model.
- E.g.: Select a specific fund ("EAFRD") and a specific objective ("Research & Innovation") to compare the amount of EU subsidies received per region.

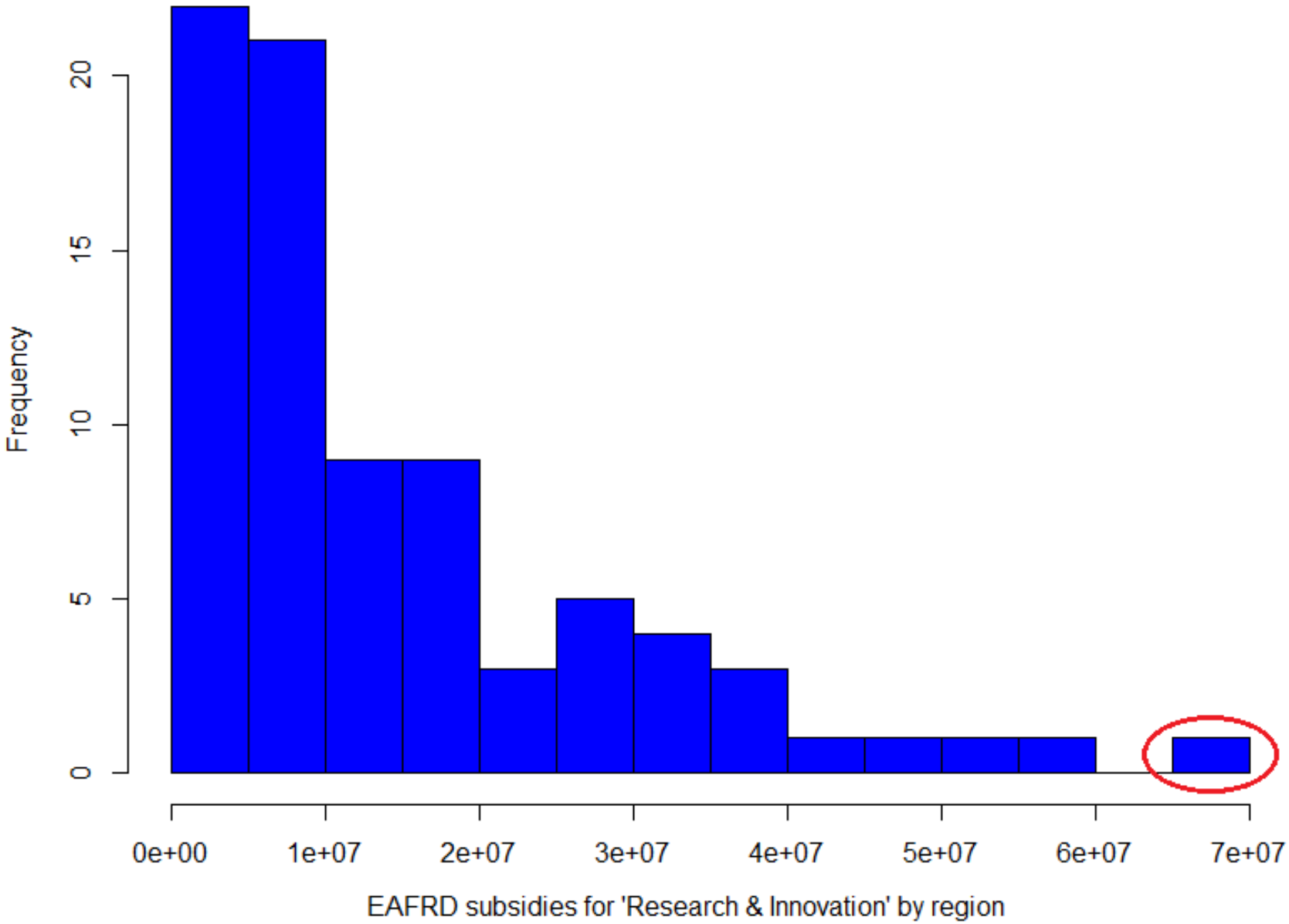
Example: EU Funds

universitätbonn



Example: EU Funds

universitätbonn



How to define appropriate subgroups?

- Slicing the data cube is not enough
- The items are still not comparable

→ Solution: Enrich the data set with additional features and apply a refined, more advanced grouping strategy.

How to define appropriate subgroups?

- Slicing the data cube is not enough
- The items are still not comparable

→ Solution: Enrich the data set with additional features and apply a refined, more advanced grouping strategy.

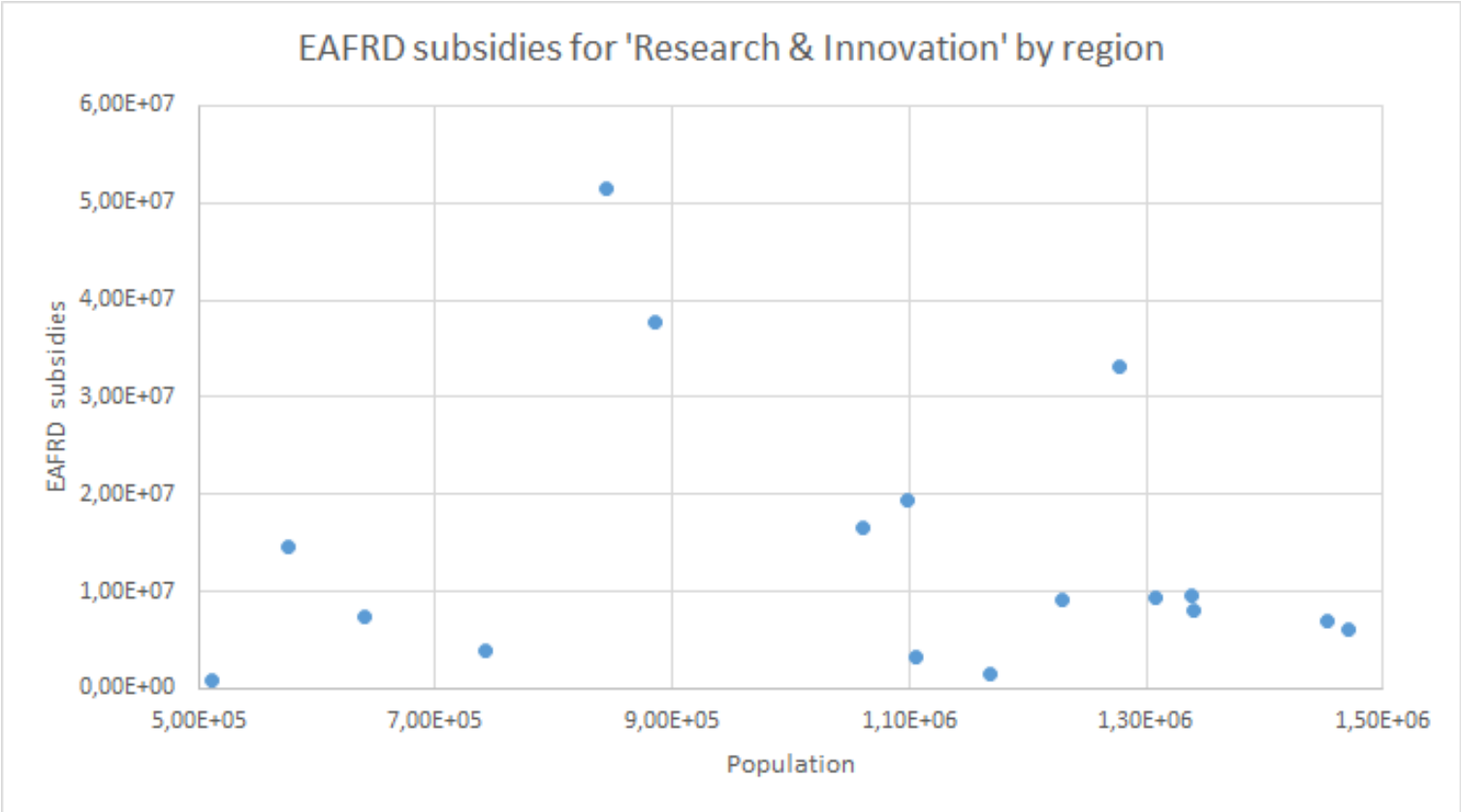
  - Use semantic layer for data enrichment
  - Use a subpopulation lattice as grouping strategy

Data Enrichment

- Fetch additional features from Linked Open Data sets
- Potential features:
  - Demographical features (population, social status, ...)
  - Geospatial features (area, location, ...)
  - Economical features (gdp, unemployment rate, economic sectors, ...)
- Potential data sets:
  - DBpedia, Eurostat, Geonames, LinkedGeoData, ...
- Follow links in the data set to fetch the features
- Possible in an automated way, using e.g. RapidMiner's Linked Open Data extension

Example: EU Funds

(Regions with 500.000 to 1.500.000 inhabitants)



universitätbonn

Subpopulation Lattice

– Fleischhacker et al. (2014)

- Applied to detect errors in numerical DBpedia properties
- Based on a paper by Melo et al. (2014)

Subpopulation Lattice

– Fleischhacker et al. (2014)

- Applied to detect errors in numerical DBpedia properties
- Based on a paper by Melo et al. (2014)
- Generate subpopulations using constraints
  - Define a property value constraint
    (i.e. fixing a value/value range for a property)
  - Limit the set of instances to those satisfying the constraint

Subpopulation Lattice

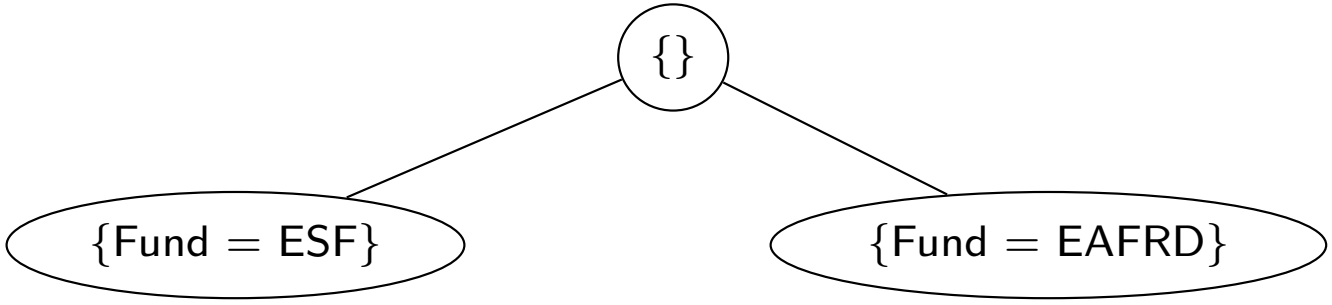
– Fleischhacker et al. (2014)

- Applied to detect errors in numerical DBpedia properties
- Based on a paper by Melo et al. (2014)
- Generate subpopulations using constraints
  - Define a property value constraint
    (i.e. fixing a value/value range for a property)
  - Limit the set of instances to those satisfying the constraint
- Apply outlier detection only to promising subpopulations
- Use a lattice to organize the search for promising subpopulations
- Each node is assigned a set of constraints that define the corresponding subpopulation

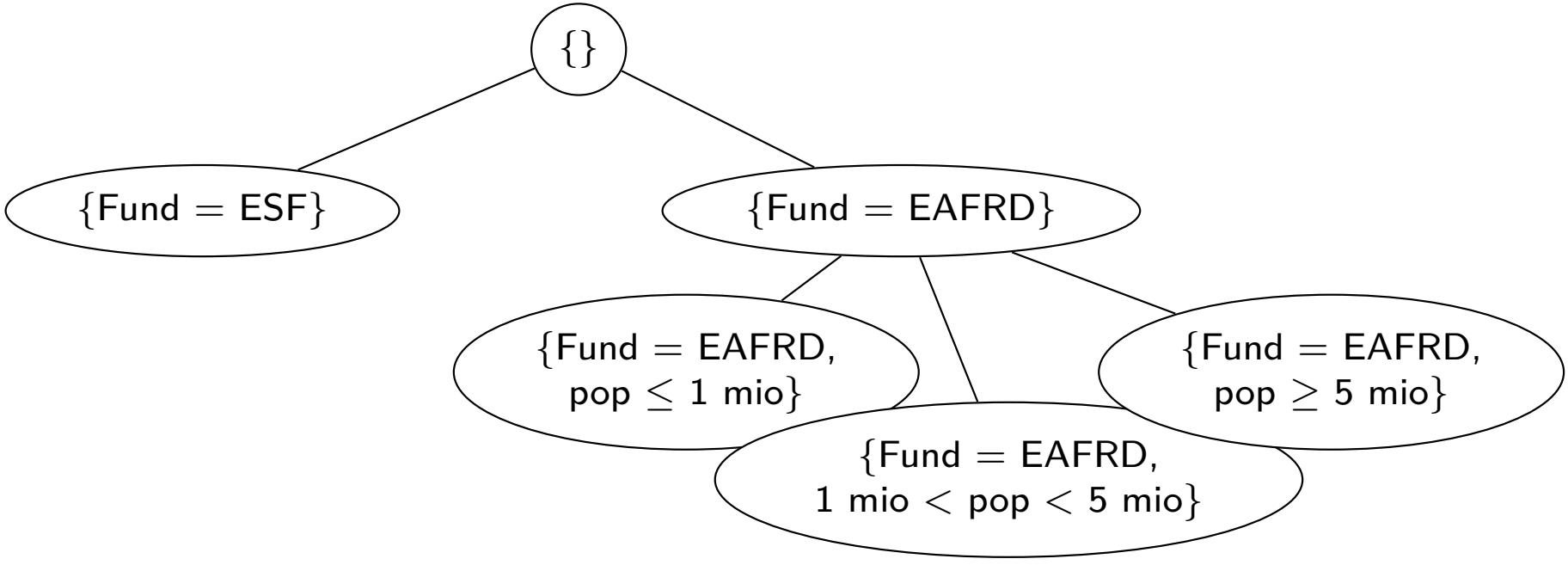
Subpopulation Lattice – Example

universitätbonn

Subpopulation Lattice – Example

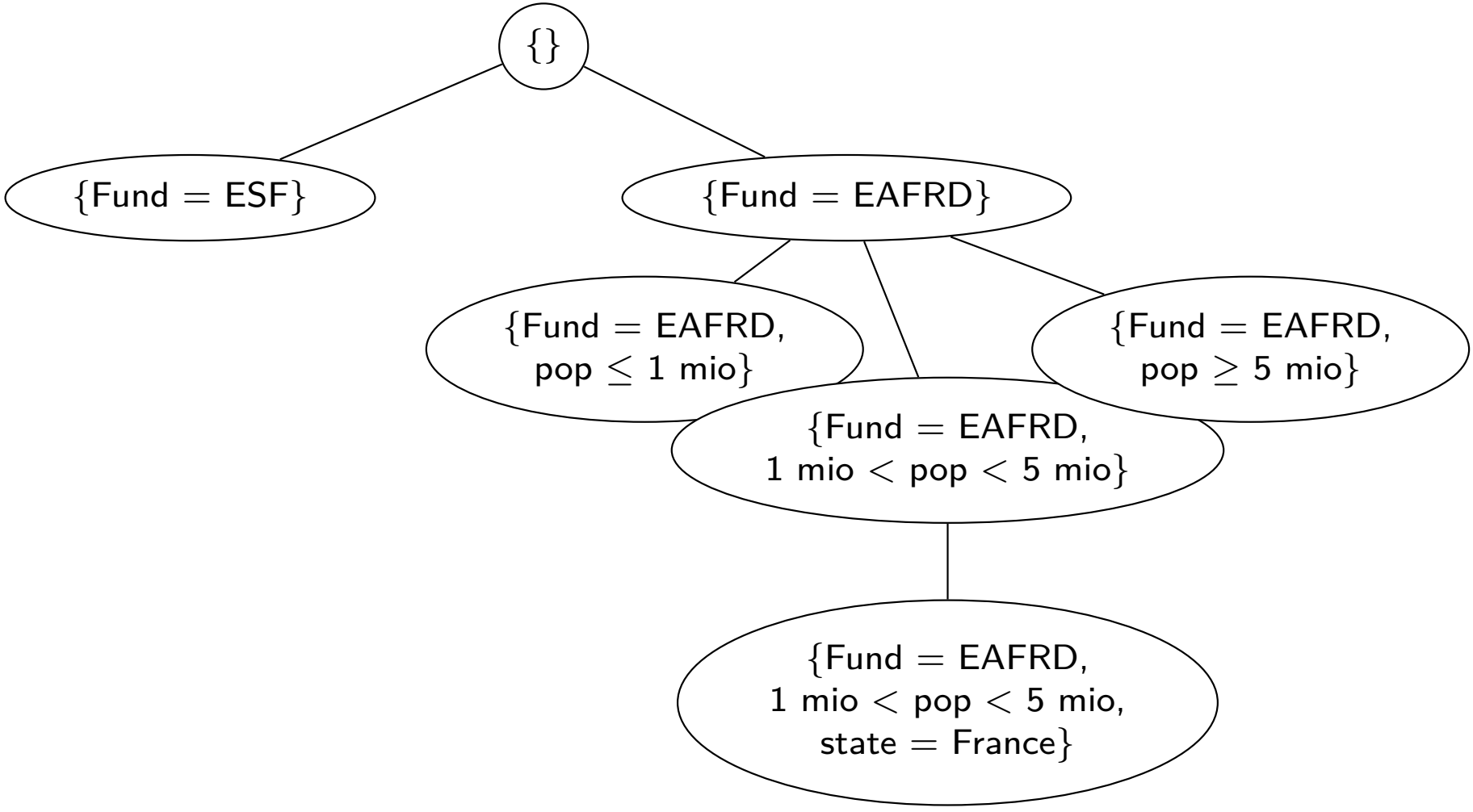


Subpopulation Lattice – Example



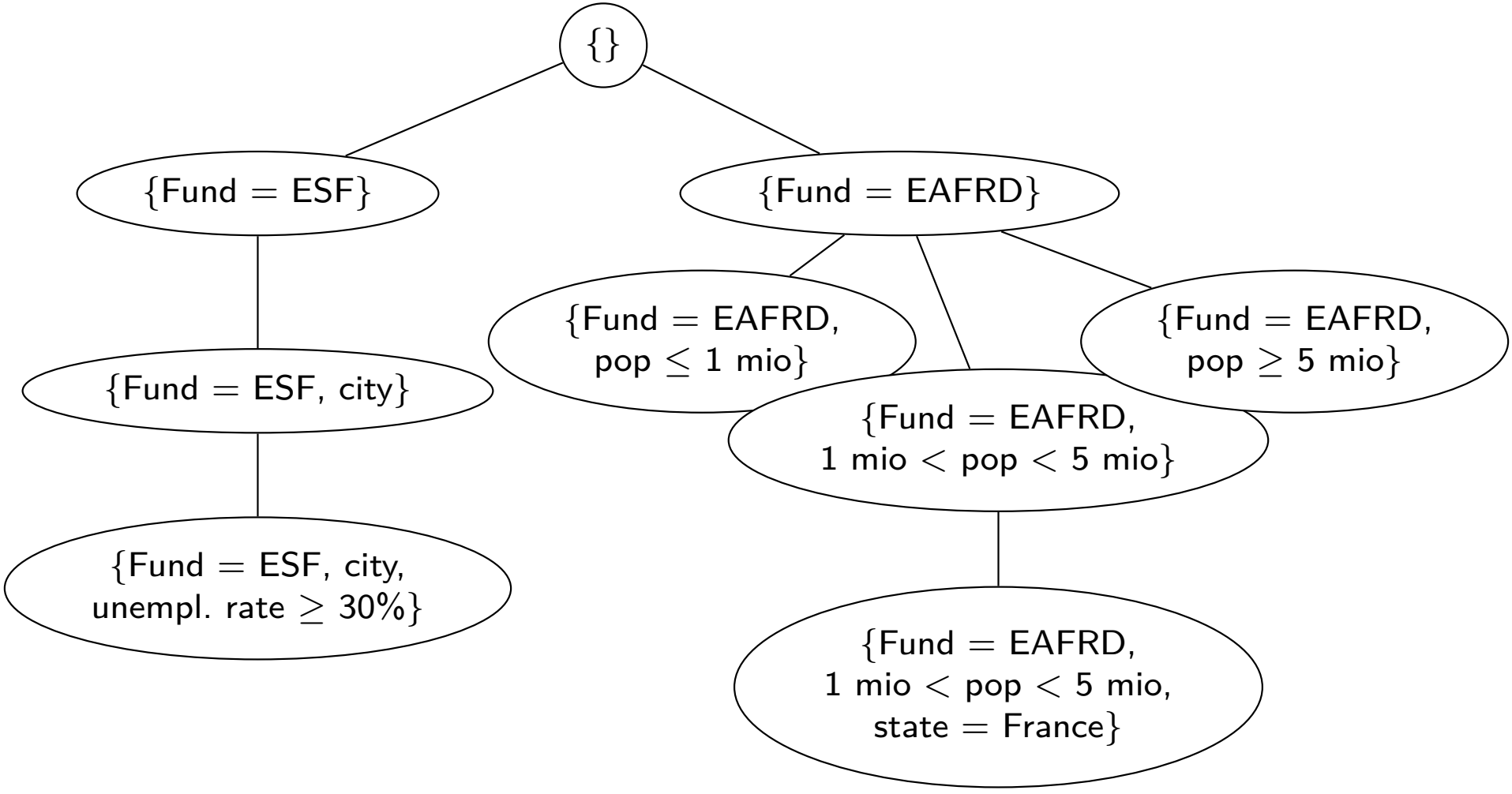
Subpopulation Lattice – Example

universitätbonn



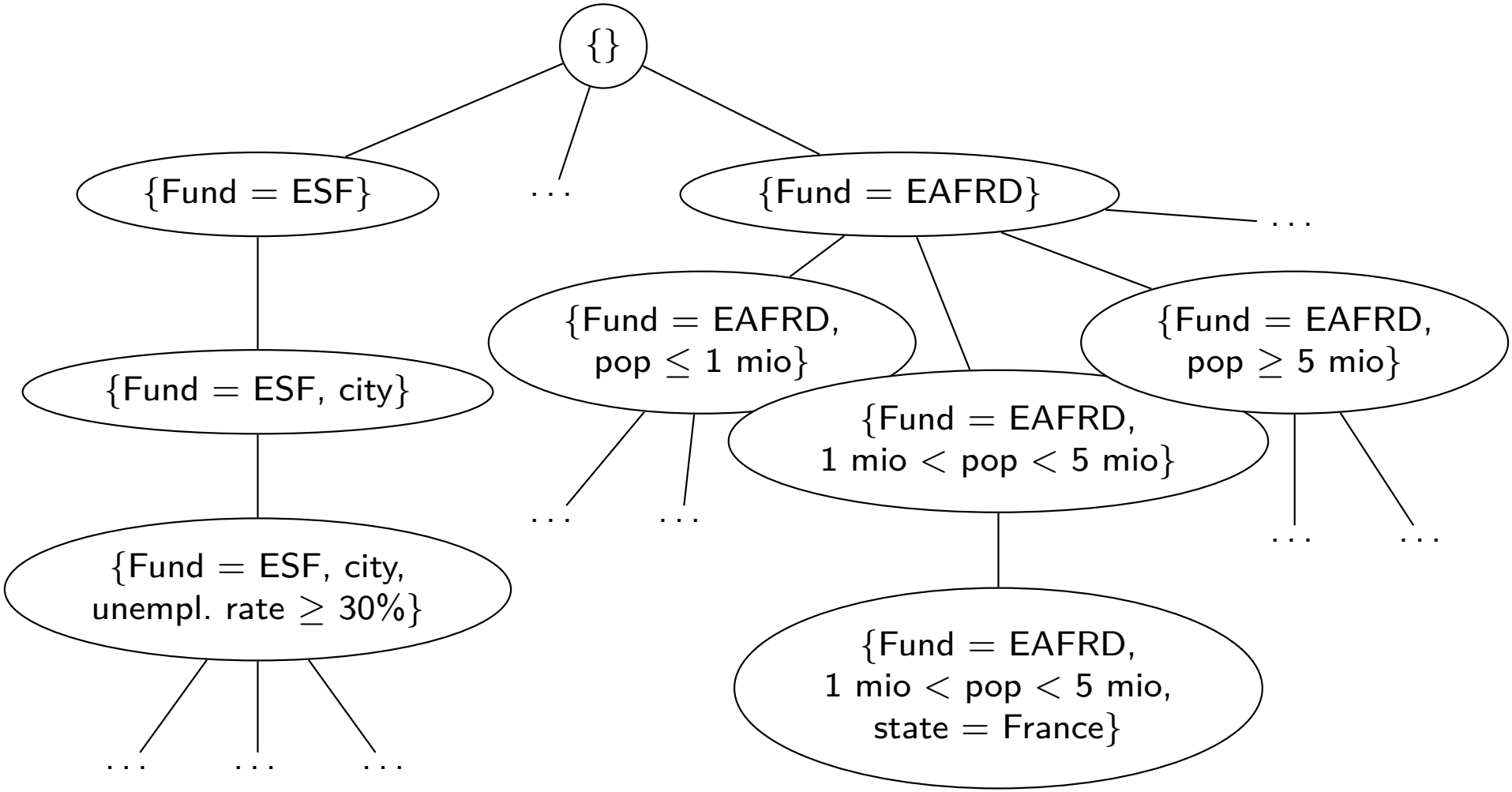
Subpopulation Lattice – Example

universitätbonn



Subpopulation Lattice – Example

universitätbonn



Subpopulation Lattice

- Extend lattice in a breadth-first-manner
- Maintain histogram of target value at each node
- Prune nodes
  - if a subpopulation is too specific
  - if the subpopulation is not further reduced
  - if the histogram does not change
- Note that the subpopulations are not disjoint, even if restricted to the leaves!

Subpopulation Lattice

- Apply outlier detection to all nodes of the lattice
- This results in an outlier score together with the corresponding constraints defining the subpopulations for each instance, i.e. amount of money spent/received
- Combine the outlier scores using an appropriate weighting scheme to a global score

Extract explanations

- From the subpopulation constraints we can extract explanations why a specific data point is considered as outlier
- Ensure interpretability of resulting scores
- Example:
  "Among regions with 500.000 to 1.500.000 inhabitants, Reunion and Umbria are receiving unusual amounts of EAFRD subsidies for the objective 'Research and Innovation'."

Approach (Summary)

- Enrich data set with additional features
- Compute subpopulation lattice grouping comparable items together
- Apply outlier detection on the lattice
- Combine different outlier scores using an appropriate weighting scheme
- Present results to the user together with explanations extracted from the corresponding subpopulation constraints

Experiments

- Plan to use RapidMiner
  - Software platform supporting all steps of the data mining process
  - Implementations of established data mining algorithms, including outlier various detection methods
  - LOD extension to deal with RDF data and fetch additional features
  - Process is split into separate *operators* (easy extendable, components can be replaced with alternative operators)
  - Possiblitiy to implement own operators
- Evaluation will be done on real world open government budget and spending data (OBEU project)

List of Experiments

- Which outlier score/approach provides the best results on the data?
- Which additional features/data sets to include?
- Compare different weighting schemes

Summary and Future Work

- Contribution
  - Combine two approaches (LOD enrichment, subpopulation lattice)
  - Apply to a new use case (financial data)
- Next steps
  - Run experiments on OBEU data
  - Evaluate approach

Comments?

Questions?

References

1. Chandola, Varun and Banerjee, Arindam and Kumar, Vipin: "Anomaly detection: A survey", ACM computing surveys (CSUR), 2009.
2. Fleischhacker, Daniel and Paulheim, Heiko and Bryl, Volha and Völker, Johanna and Bizer, Christian: "Detecting errors in numerical linked data using cross-checked outlier detection", The Semantic Web–ISWC, 2014.
3. Melo, André and Theobald, Martin and Völker, Johanna: "Correlation-based refinement of rules with numerical attributes", The Twenty-Seventh International Flairs Conference, 2014.
4. Paulheim, Heiko and Ristoski, Petar and Mitichkin, Evgeny and Bizer, Christian: "Data mining with background knowledge from the web", RapidMiner World, 2014.