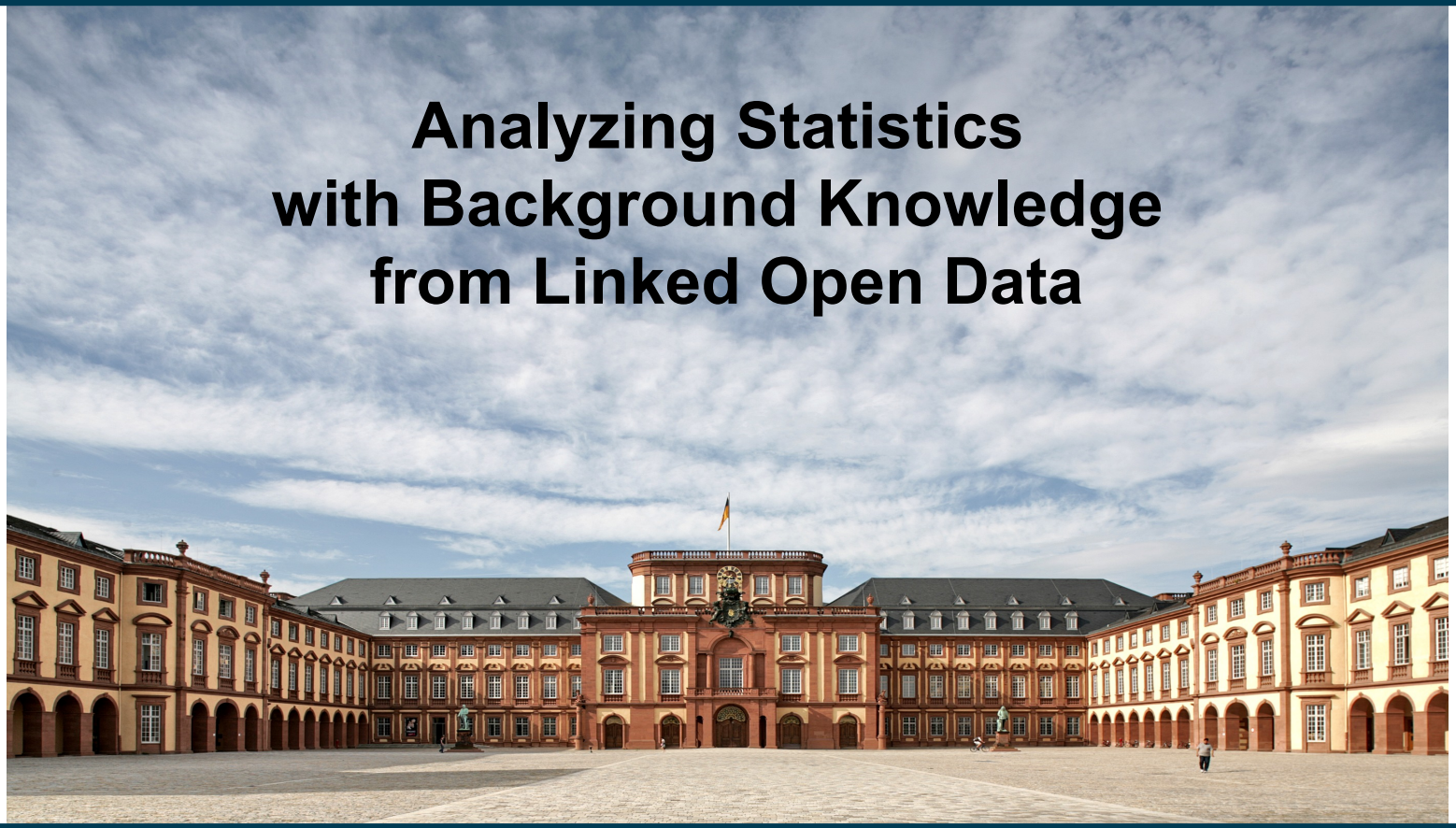
UNIVERSITÄT MANNHEIM

Analyzing Statistics with Background Knowledge from Linked Open Data



Petar Ristoski, **Heiko Paulheim**

Idea

- Background knowledge from LOD can help
  - finding explanations
  - creating more sophisticated visualizations
- Steps taken
  - linking the statistics datasets to LOD datasets
  - DBpedia, Eurostat, GADM, Linked Geo Data
  - Extracting features
- Finding correlations with unemployment rate
  - using only one target variable for demonstration purposes
  - works for arbitrary target variables

Linking to LOD Datasets

- Linking to DBpedia
  - using DBpedia Lookup
  - restricting results to Place and AdministrativeArea
  - select from many results by minimum edit distance

- Linking to Eurostat
  - using SPARQL to query for labels
  - querying for word 1-grams, 2-grams, … from original labels
  - selecting by minimum edit distance

Linking to LOD Datasets

- Linking to GADM
  - searching by name turned out to be error-prone
  - searching by coordinates (from DBpedia) is precise
    - but suffers from low recall
  - two-stage approach:
    - searching by coordinates
    - searching by average coordinates of all linked objects (in DBpedia)

- Total figures:
  - France regions: 27/27 DBpedia, 26/27 Eurostat, 27/27 GADM
  - France departments: 101/101 DBpedia, 101/101 GADM
  - Australia states: 8/9 DBpedia, 9/9 GADM
  - Australia SA3/SA4: no satisfying results, discarded

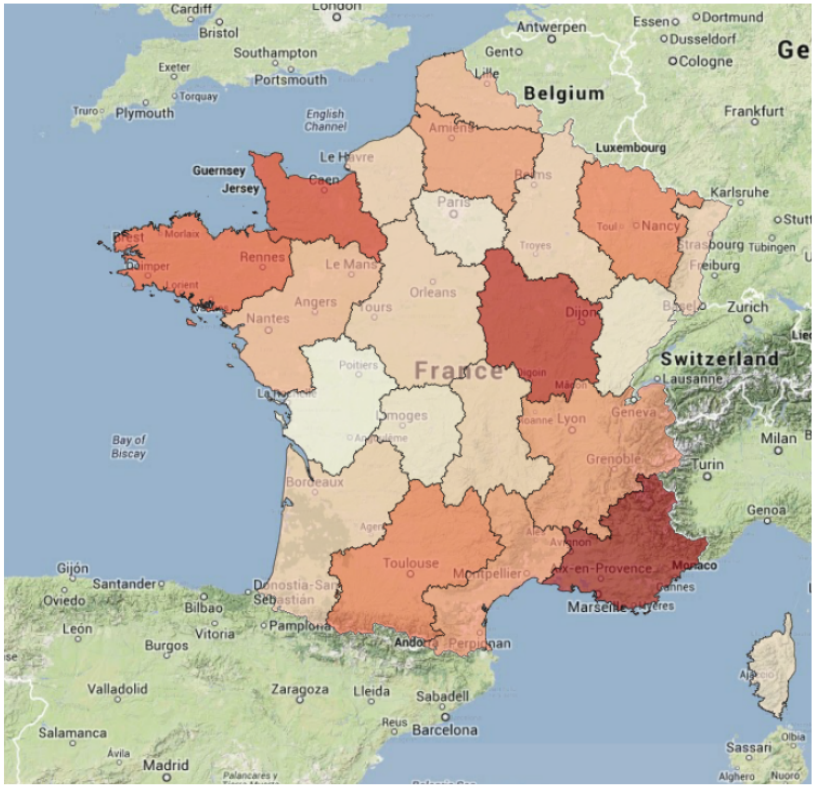
Feature Extraction

- Once the links have been created
  - get polygon shapes from GADM (for visualization)
  - get datatype properties from Eurostat/DBpedia
  - get direct types from DBpedia (incl. YAGO types)
  - get qualified relations from DBpedia

- Using information from Linked Geo Data
  - extract objects within GADM polygon, aggregate by type (e.g., region contains 125 police stations)
  - spatial queries only possible with rectangles
  - workaround: use minimum enclosing rectangle and filter afterwards

Visualization with GADM Polygons

- Polygons from GADM allow for visualization of unemployment on maps



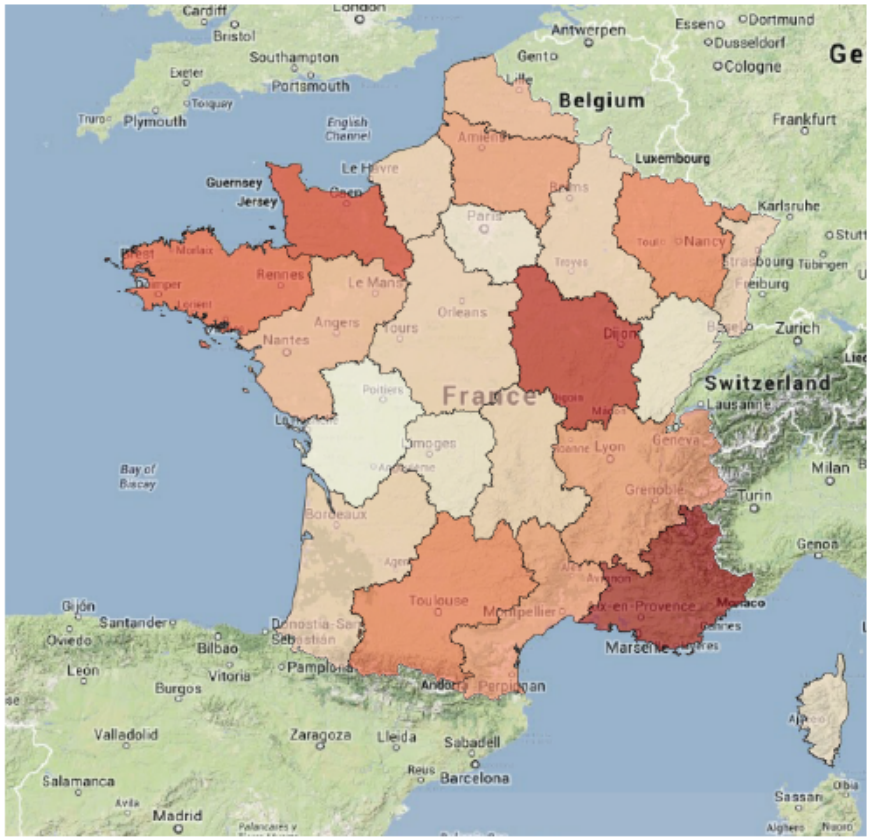
(a) Unemployment by region

Finding Correlations

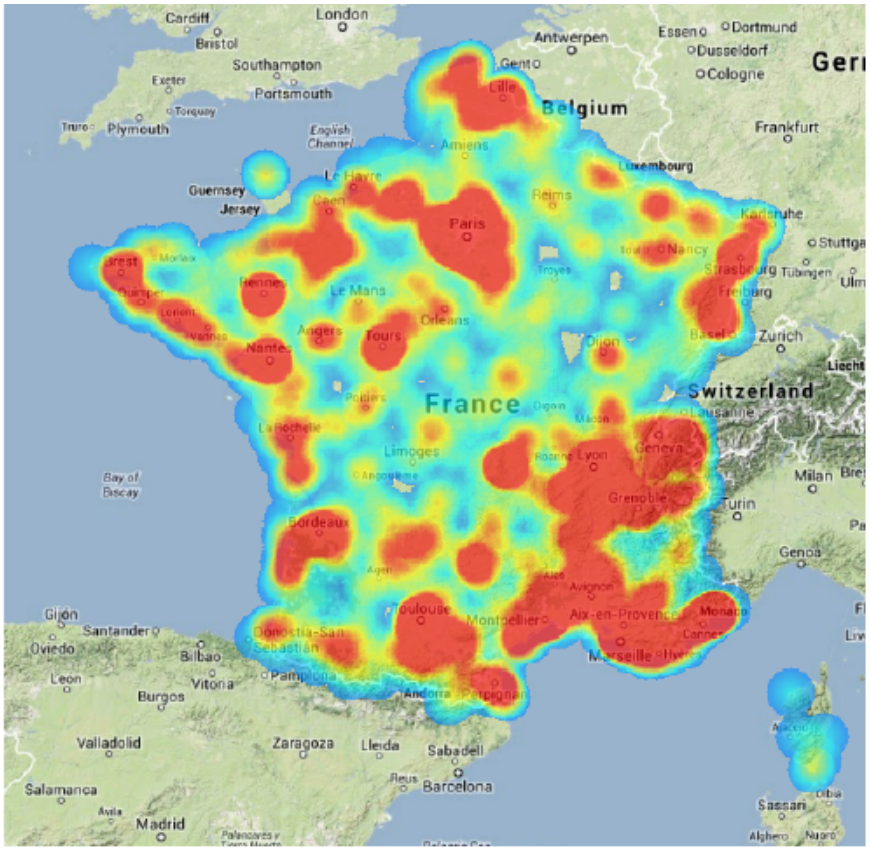
- Using extracted features to find interesting correlations
- Example correlation for unemployment in France:
  - African islands, Islands in the Indian Ocean, Outermost regions of the EU (positive)
  - GDP (negative)
  - Disposable income (negative)
  - Hospital beds/inhabitants (negative)
  - RnD spendings (negative)
  - Energy consumption (negative)
  - Population growth (positive)
  - Casualties in traffic accidents (negative)
  - Fast food restaurants (positive)
  - Police stations (positive)

Visualization Correlations

- e.g., unemployment rate ~ number of police stations



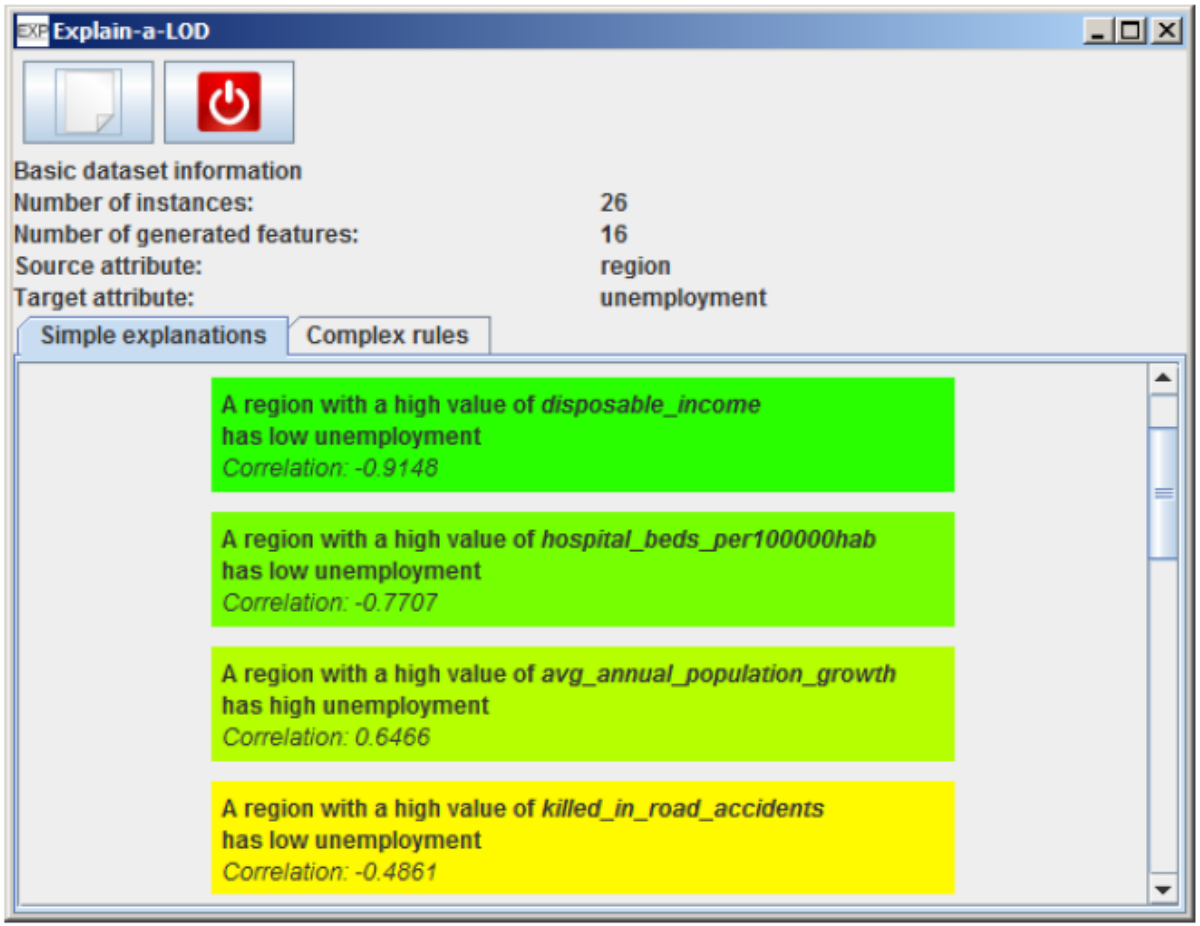
(a) Unemployment by region



(b) Heat map of police stations

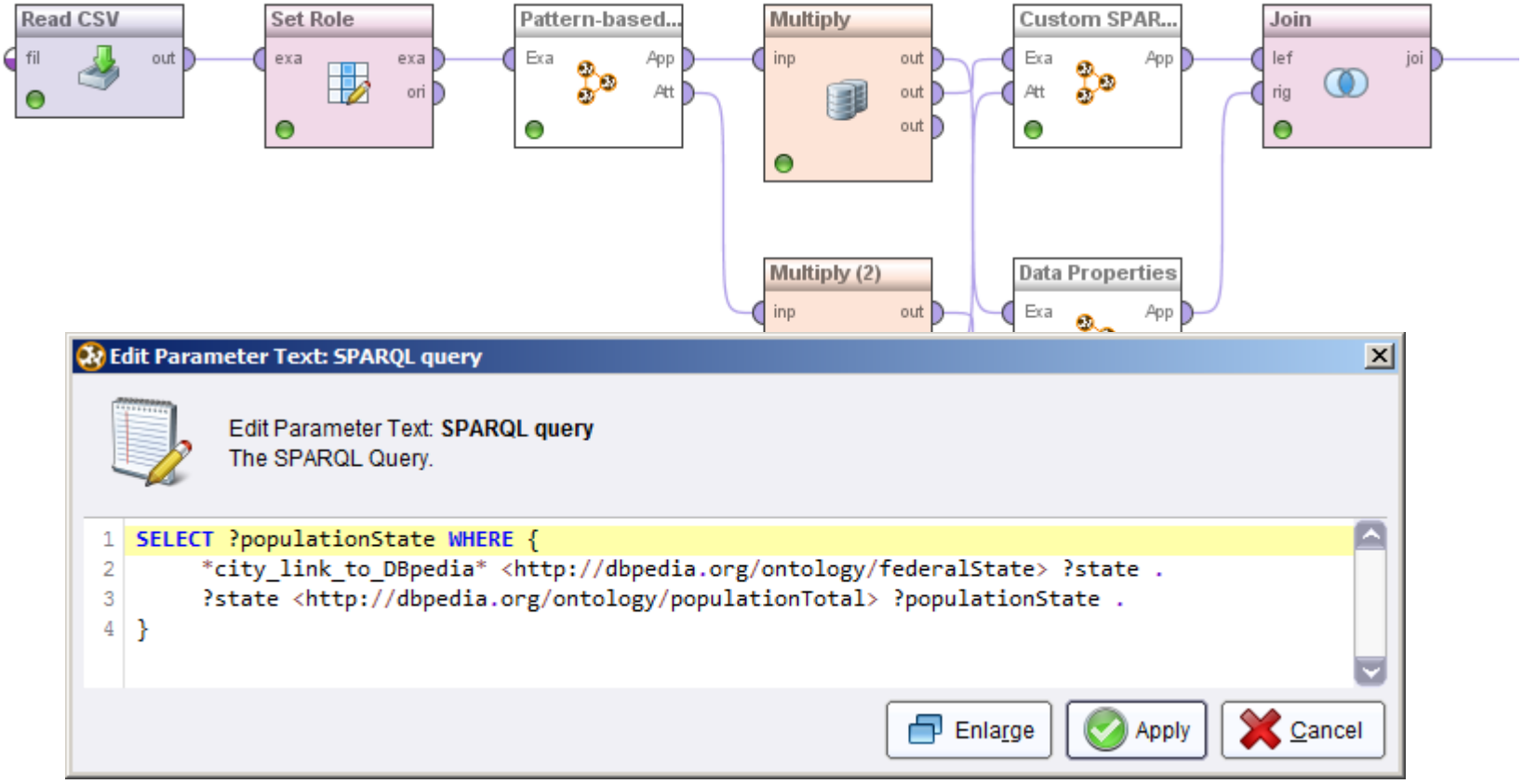
Tools

- FeGeLOD/Explain-a-LOD (ESWC 2012: best demo award)



Tools

- RapidMiner Linked Open Data Extension (2013)



Assets

- New modules for RapidMiner LOD extension
  - DBpedia Lookup linker
  - Label-based linker

- New links for DBpedia 3.9 release
  - DBpedia – GADM (39,000 links)

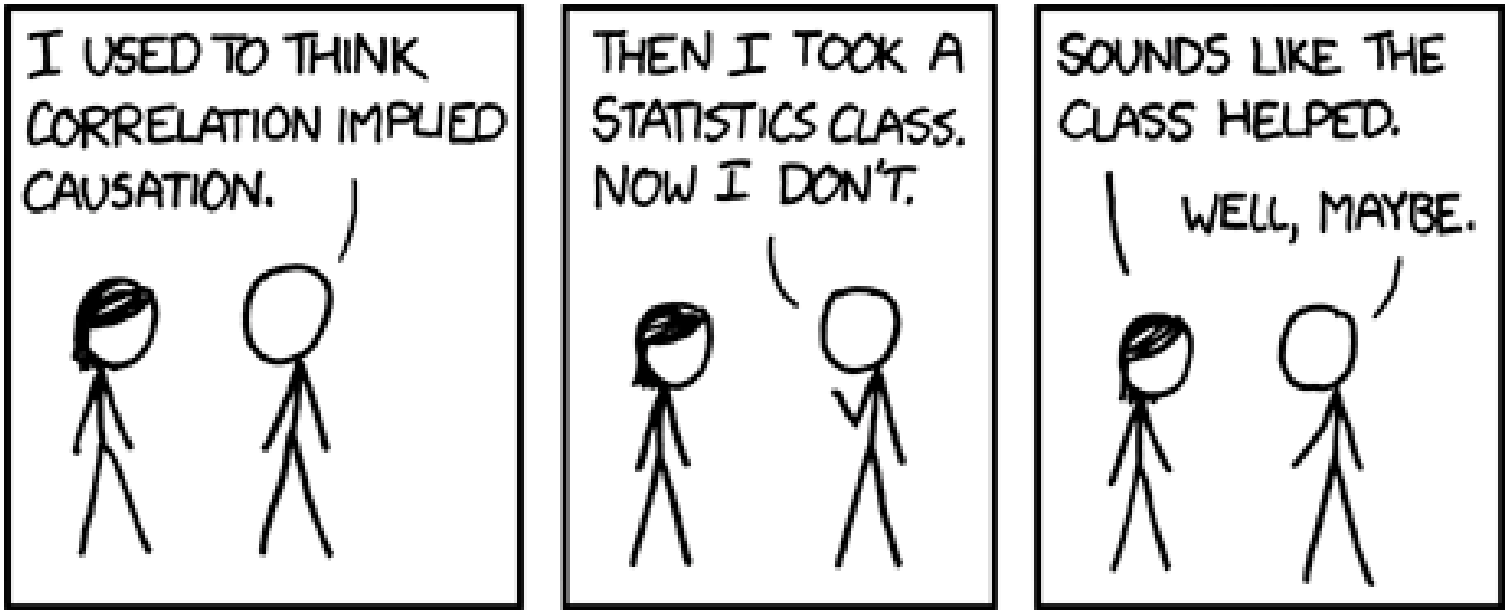
Conclusions & Lessons Learned

- Linked Open Data provides useful background knowledge
  - For finding explanations
  - For creating visualizations

- Some data sources are more suitable than others
  - official data sources (e.g. Eurostat) provide best results

Conclusions & Lessons Learned

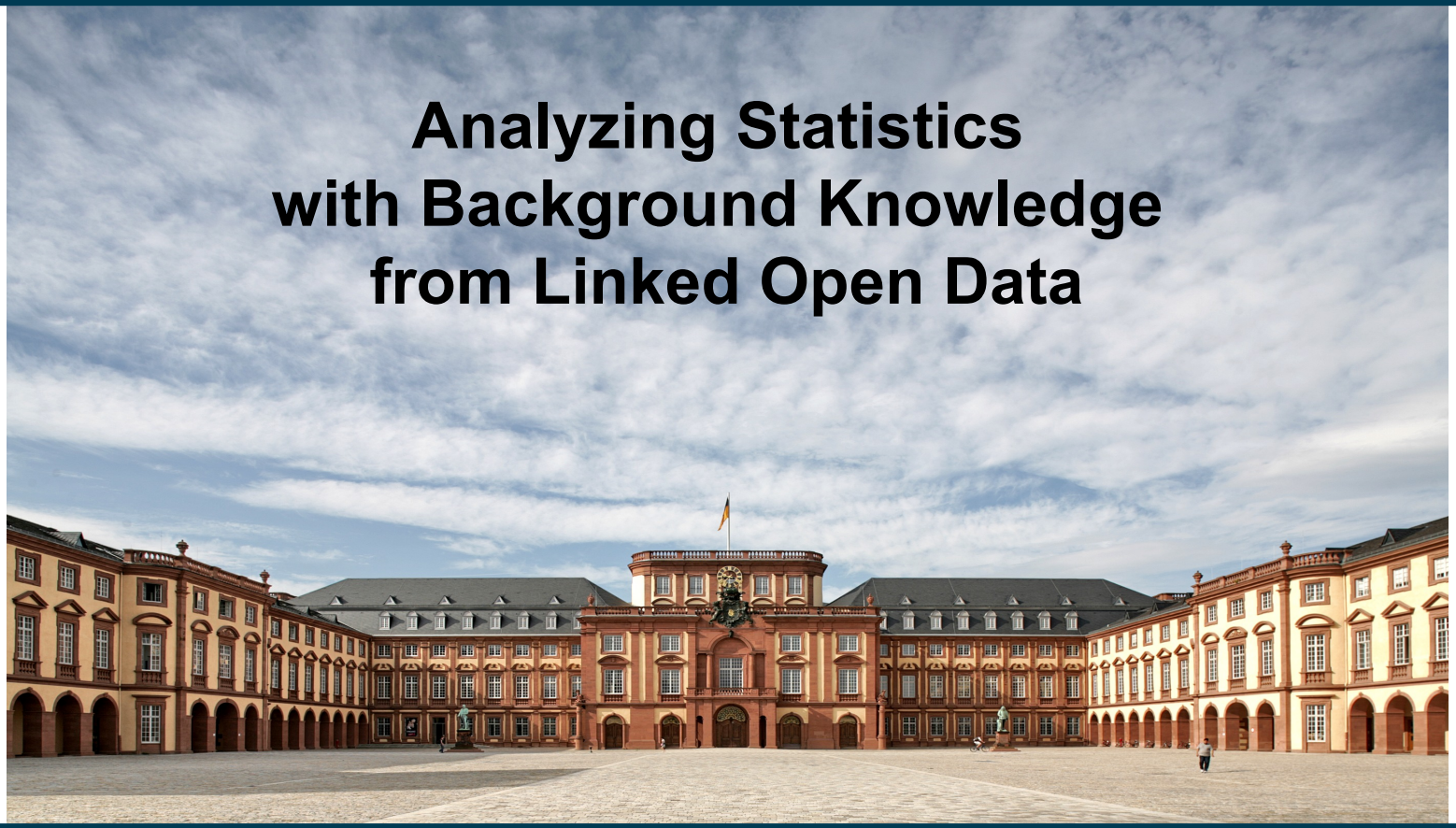
- Negative correlation: traffic accident casualties ~ unemployment
  - Fight unemployment by increasing traffic accidents?



http://xkcd.com/552/

UNIVERSITÄT MANNHEIM

Analyzing Statistics with Background Knowledge from Linked Open Data



Petar Ristoski, **Heiko Paulheim**