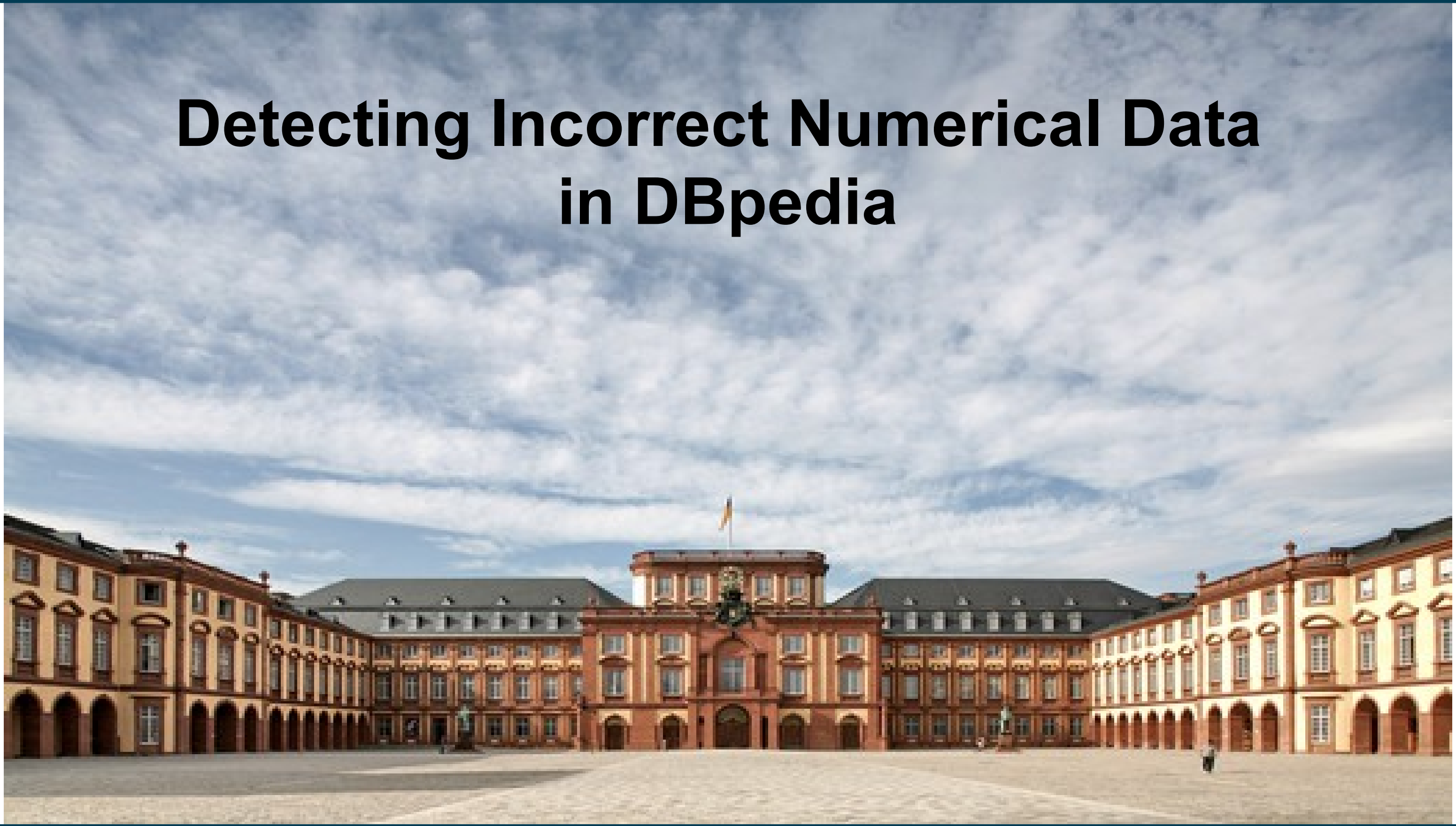
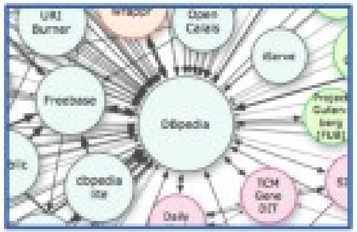
UNIVERSITÄT
MANNHEIM

Detecting Incorrect Numerical Data in DBpedia



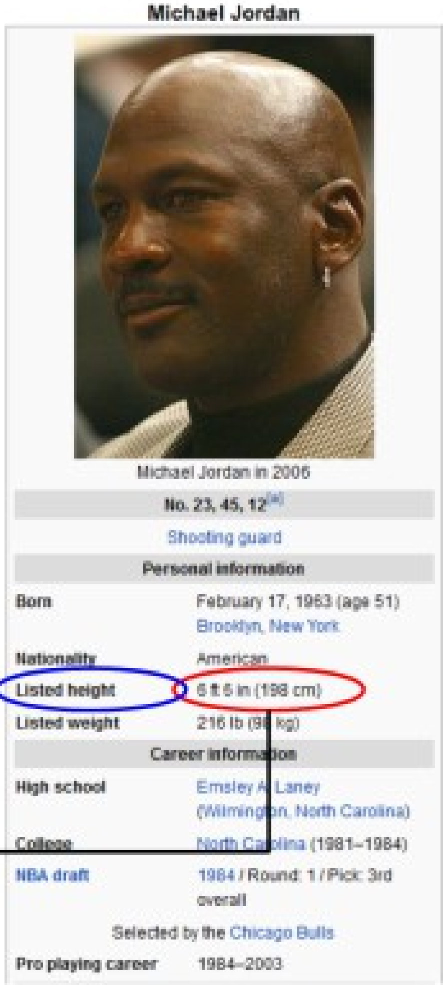
Dominik Wienand, **Heiko Paulheim**

Motivation



- DBpedia
  - extracts data from infoboxes in Wikipedia
  - based on crowd-sourced mappings to an ontology
- Example
  - Wikipedia page on Michael Jordan

```
dbpedia:Michael_Jordan
  dbpedia-owl:height
  "1.981200"^^xsd:double .
```



Michael Jordan

Michael Jordan in 2006

No. 23, 45, 12[a]	
Shooting guard	
Personal information	
Born	February 17, 1963 (age 51) Brooklyn, New York
Nationality	American
Listed height	6 ft 6 in (198 cm)
Listed weight	216 lb (98 kg)
Career information	
High school	Emsley A. Laney (Wilmington, North Carolina)
College	North Carolina (1981–1984)
NBA draft	1984 / Round: 1 / Pick: 3rd overall
Selected by the Chicago Bulls	
Pro playing career	1984–2003

Motivation

- Challenge
  - Wikipedia is made for humans, not machines
  - Input format in Wikipedia is not constrained

- The following are all valid representations of the same height value (and perfectly understandable by humans)
  - 6 ft 6 in, 6ft 6in, 6'6'', 6'6”, 6´6´´, ...
  - 1.98m, 1,98m, 1m 98, 1m 98cm, 198cm, 198 cm, ...
  - 6 ft 6 in (198 cm), 6ft 6in (1.98m), 6'6'' (1.98 m), ...
  - 6 ft 6 in[1], 6 ft 6 in [citation needed], ...
  - ...

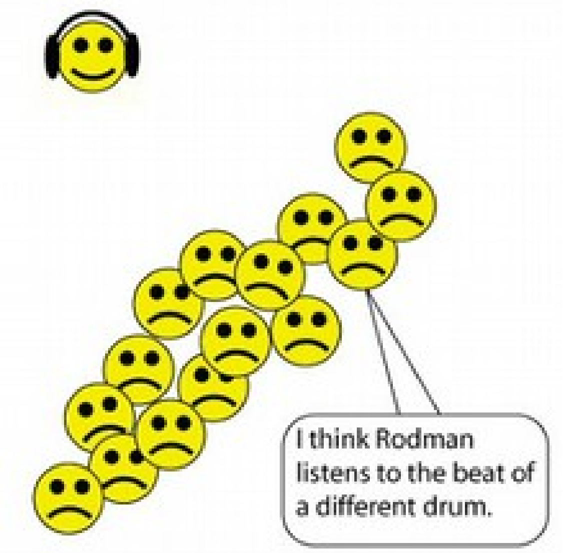
Motivation

- Challenge
  - We're (hopefully) slowly stepping out of the research labs
    - e.g., applications in Emergency Management, Finance, …
  - so we need reliable information
    - i.e., DBpedia has to be able to deal with all of those variants
- But
  - it is hard to cover each and every case
  - if the case is rare, we may not even know it
- Idea
  - A posteriori plausibility checking
  - Find values that are likely to be wrongly extracted

Idea

- Use outlier detection to find unlikely values
  - e.g., extremely large or small values

- Outlier Detection
  - "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs." (Grubbs,1969)
  - Outliers are not necessarily wrong!



Approach

- Basic approach
  - use all values of a numerical property (e.g., height) as a population
  - find outliers in that population

- Outlier detection approaches used
  - Median Absolute Deviation (Dispersion)
  - Interquartile Range
  - Kernel Density Estimation
  - Kernel Density Estimation iterative
    - i.e., remove found outliers and repeat

Median Absolute Deviation (MAD)

- MAD is the median deviation from the median of a sample, i.e.

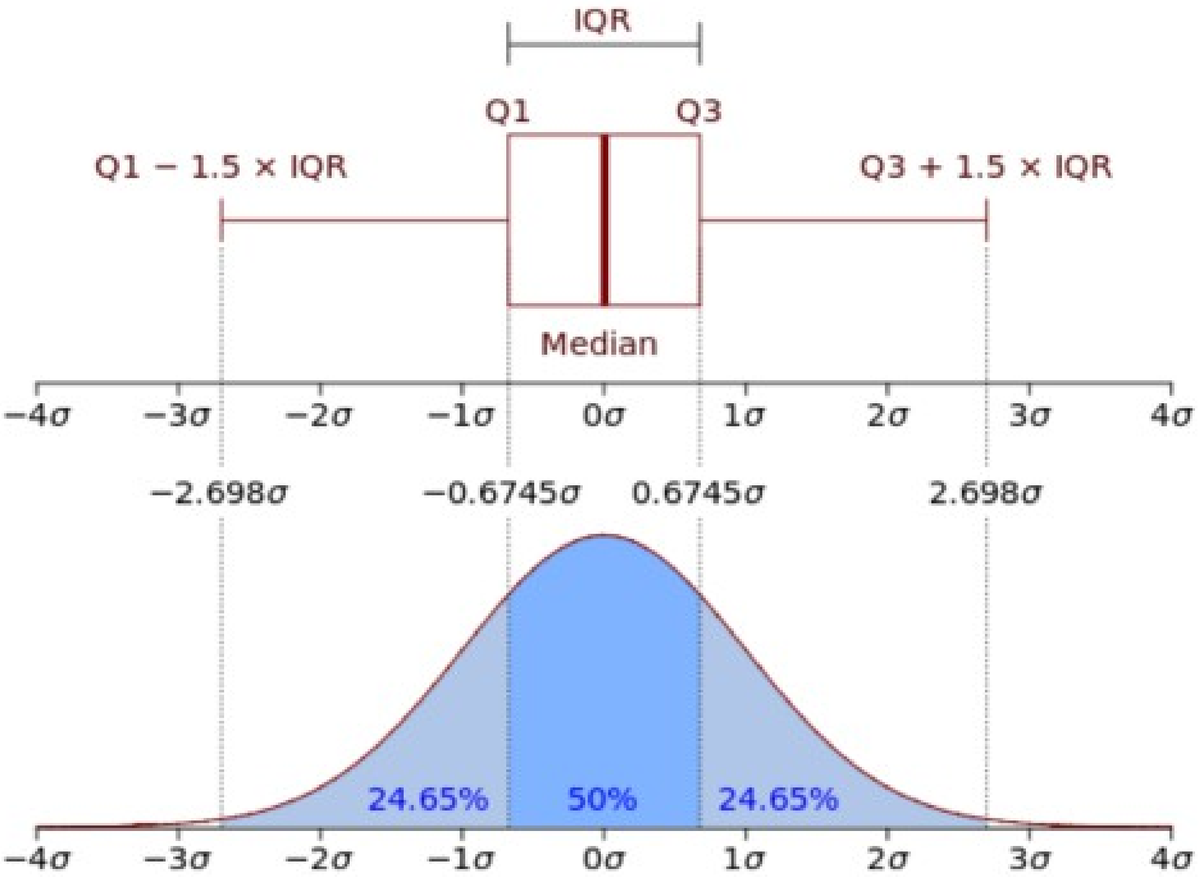
$$MAD := median_i |X_i - median_j(X_j)|$$

- MAD can be used for outlier detection
  - all values that are k*MAD away from the median are considered to be outliers
  - e.g., k=3

Carl Friedrich Gauss

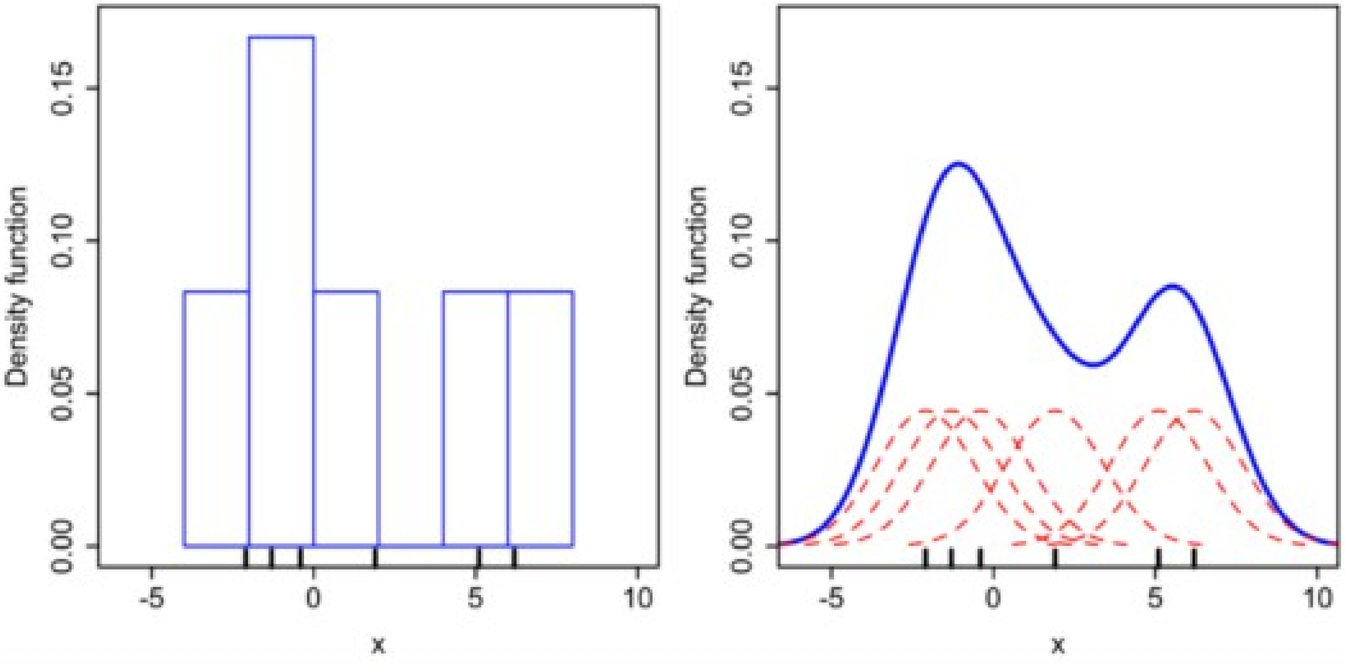
Interquartile Range

- Data is divided into four quartiles



Kernel Density Estimation

- Data populations is approximated as a sum of kernel functions
  - e.g., Gaussian normal distributions
  - function computes probability for “outlierness” of a value
  - faster approximation by Fast Fourier Transformation



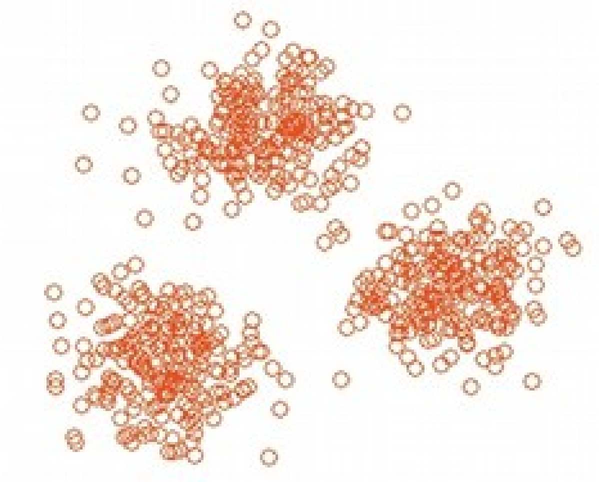
Approach

- Observation
  - some properties are used on a variety of different things
- Example: `dbpedia-owl:height`
  - persons, vehicles
- Example: `dbpedia-owl:population`
  - villages, cities, countries, continents
- Finding outliers in those mixed sets might be hard
  - refined approach: preprocess data
  - divide into subpopulations

Approach

- Preprocessing A: single type
  - split by single type
  - one data population per type (in the DBpedia ontology)
  - only the most specific type is used

- Preprocessing B: cluster by type vectors
  - each instance represented by vector of types
  - cluster instances with similar type vectors
  - EM algorithm (Weka)



Evaluation

- Two-fold evaluation
  - Pre study on three attributes (height, population, elevation)
  - Most promising approaches tested on random sample from DBpedia

- Evaluation strategy
  - a posteriori evaluation
  - each identified outlier is checked
  - bulk checking possible due to obvious clusters/patterns in outliers

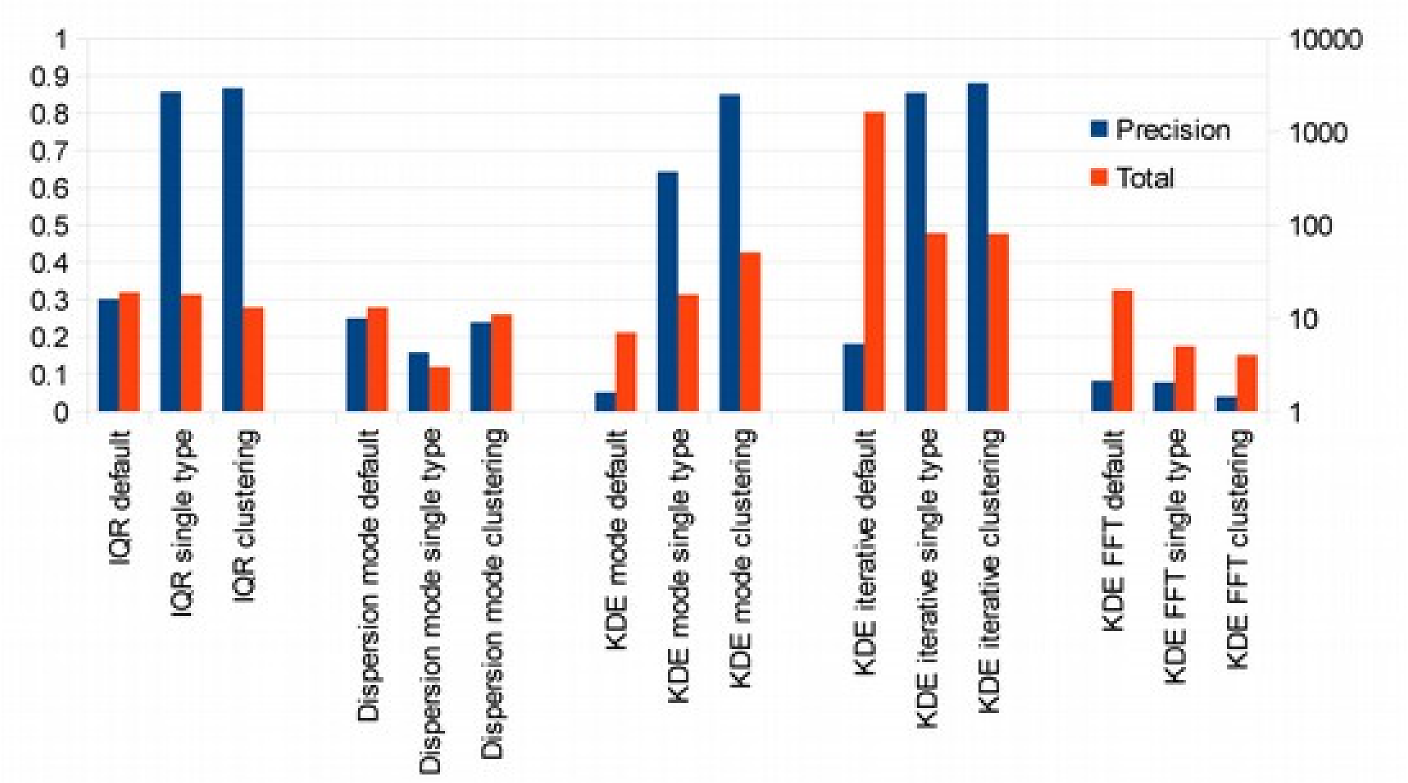
Evaluation: Pre-Study

- Sample sizes:
  - height: 52,522
  - population: 237,700
  - elevation: 206,977

- Different distributions
  - e.g., height: approximate normal distribution
  - e.g., population: power law distribution

Evaluation: Pre-Study

- Grouping and clustering improves the precision
- IQR and KDE iterative are best



Evaluation: Random Sample

- Findings from Pre-Study:
  - IQR and KDE iterative work well
  - clustering is too slow (>24h)
  - KDE FFT has poor precision

- Building a random sample
  - select 50 random resources
  - get all their datatype properties
  - retrieve all triples that use those properties
  - remove those properties that have <50% or <100 numbers as objects
- Resulting sample:
  - 12,054,727 triples

Evaluation: Random Sample

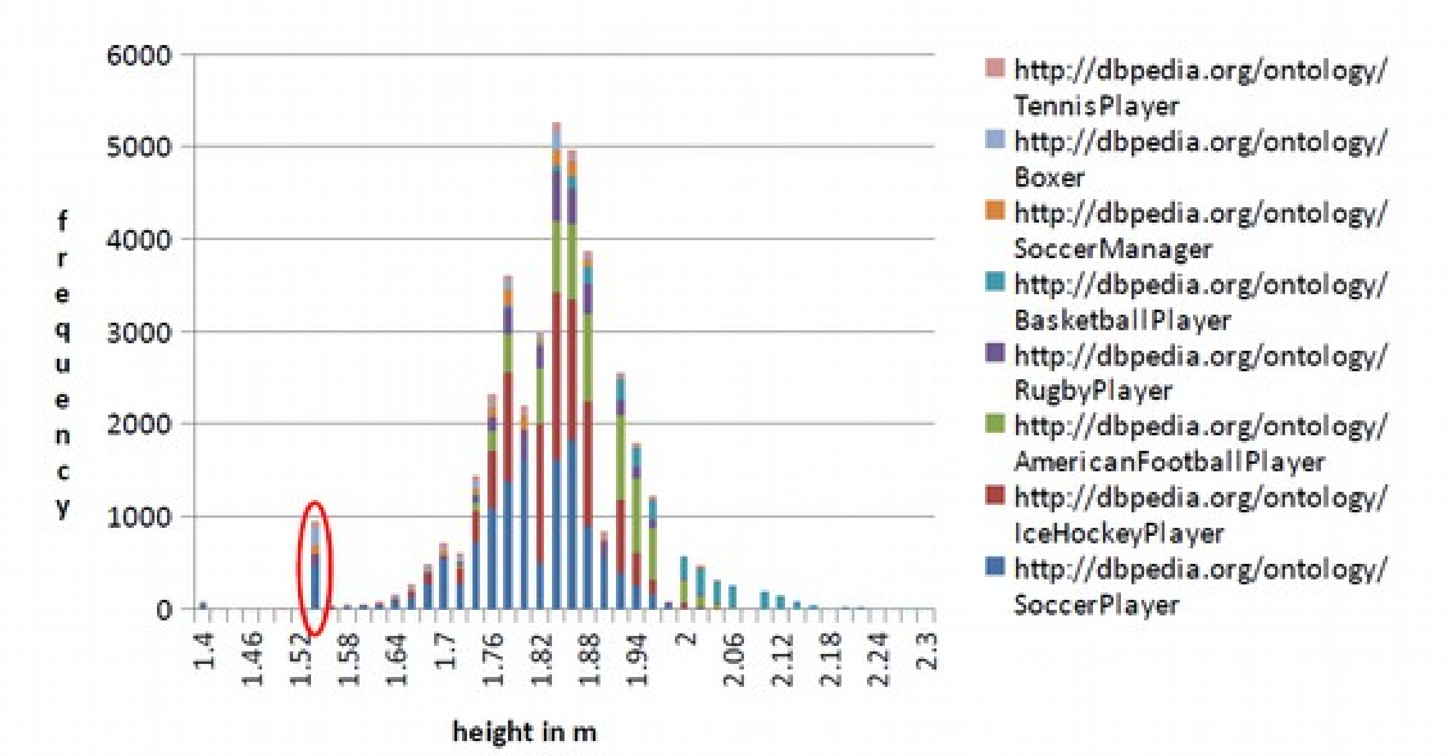
- Tried best performing configuration in pre-study
  - further explored parameter settings from there
  - IQR provides best trade-off between runtime and quality

- Best configuration for IQR:
  - 1,703 values marked as outliers
  - manually checked
    - shortcuts, e.g., all three digit ZIP codes for places in the US
  - Precision of outlier detection: 81%
    - i.e., roughly 1 in 1,000 values is wrong

Systematic Errors Found

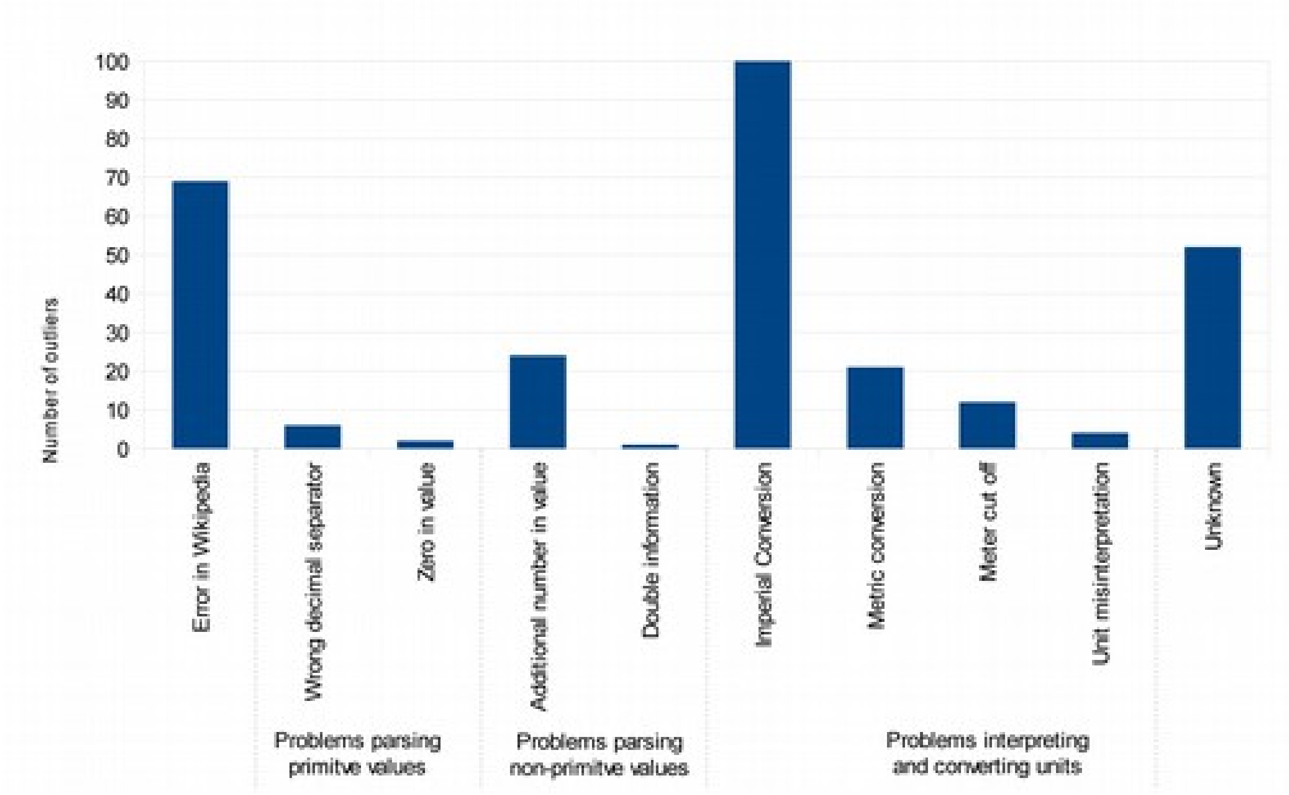
- The footprint of an error in the extraction framework code
  - here: imperial measures after 5' are truncated

    → observation: suspiciously many height values of 1.524m (=5')



Systematic Errors Found

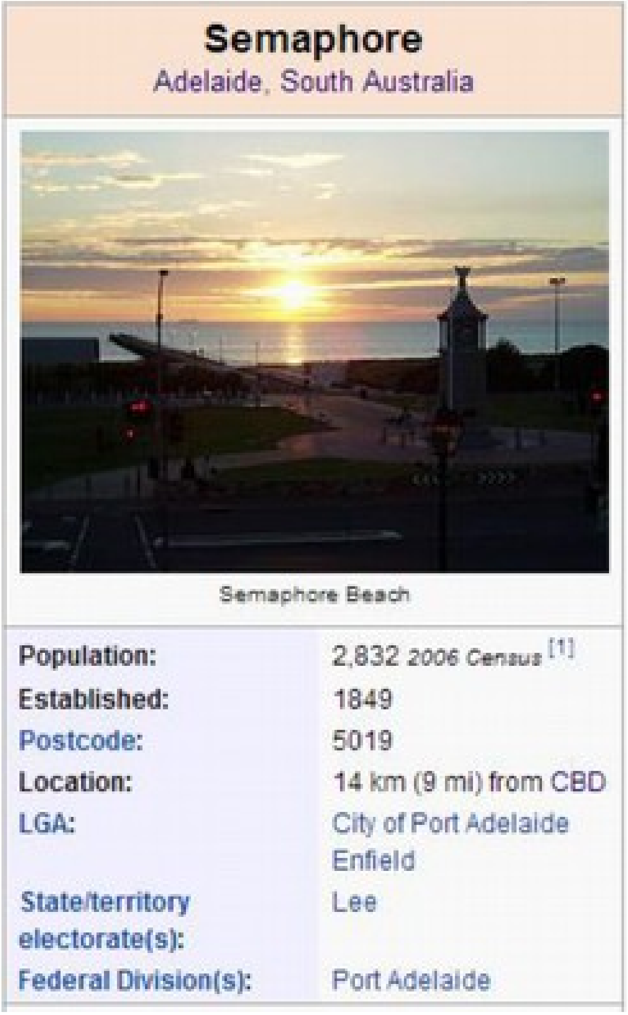
- Imperial conversion is the most severe problem
  - causes almost 90% of all outliers in our sample



Systematic Errors Found

- Additional numbers cause problems
- Example: village of *Semaphore*
  - population: 28,322,006

    (all of Australia: 23,379,555!)
  - a clear outlier among villages



Semaphore

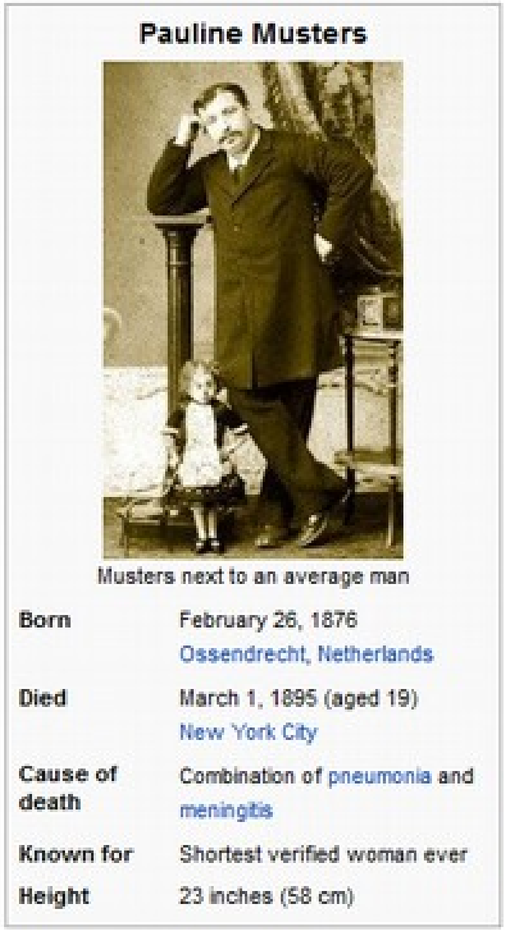
Adelaide, South Australia

Semaphore Beach

Population:	2,832 *2006 Census* [1]
Established:	1849
Postcode:	5019
Location:	14 km (9 mi) from CBD
LGA:	City of Port Adelaide Enfield
State/territory electorate(s):	Lee
Federal Division(s):	Port Adelaide

Limitations

- Telling natural outliers from errors
  - hard without additional evidence
- e.g., an adult person 58cm high
- e.g., a 7.4m high vehicle



Pauline Musters

Musters next to an average man

Born	February 26, 1876 Ossendrecht, Netherlands
Died	March 1, 1895 (aged 19) New York City
Cause of death	Combination of pneumonia and meningitis
Known for	Shortest verified woman ever
Height	23 inches (58 cm)

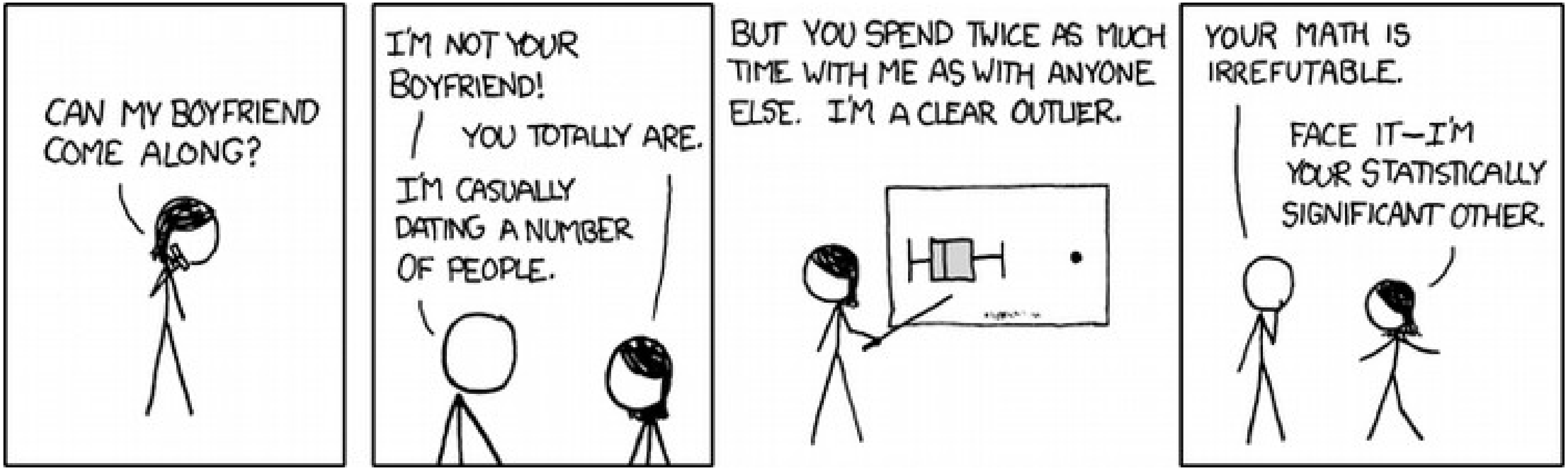
Beyond DBpedia

- So far, this work has been carried out only on DBpedia
- But can be transferred to any LOD dataset
- Particularly useful for crowdsourced/heuristic approaches
  - Information extraction from text (e.g., NELL, ReVerb)
  - Automatic fact completion
  - Datasets heuristically integrated from diverse sources

Ongoing Work

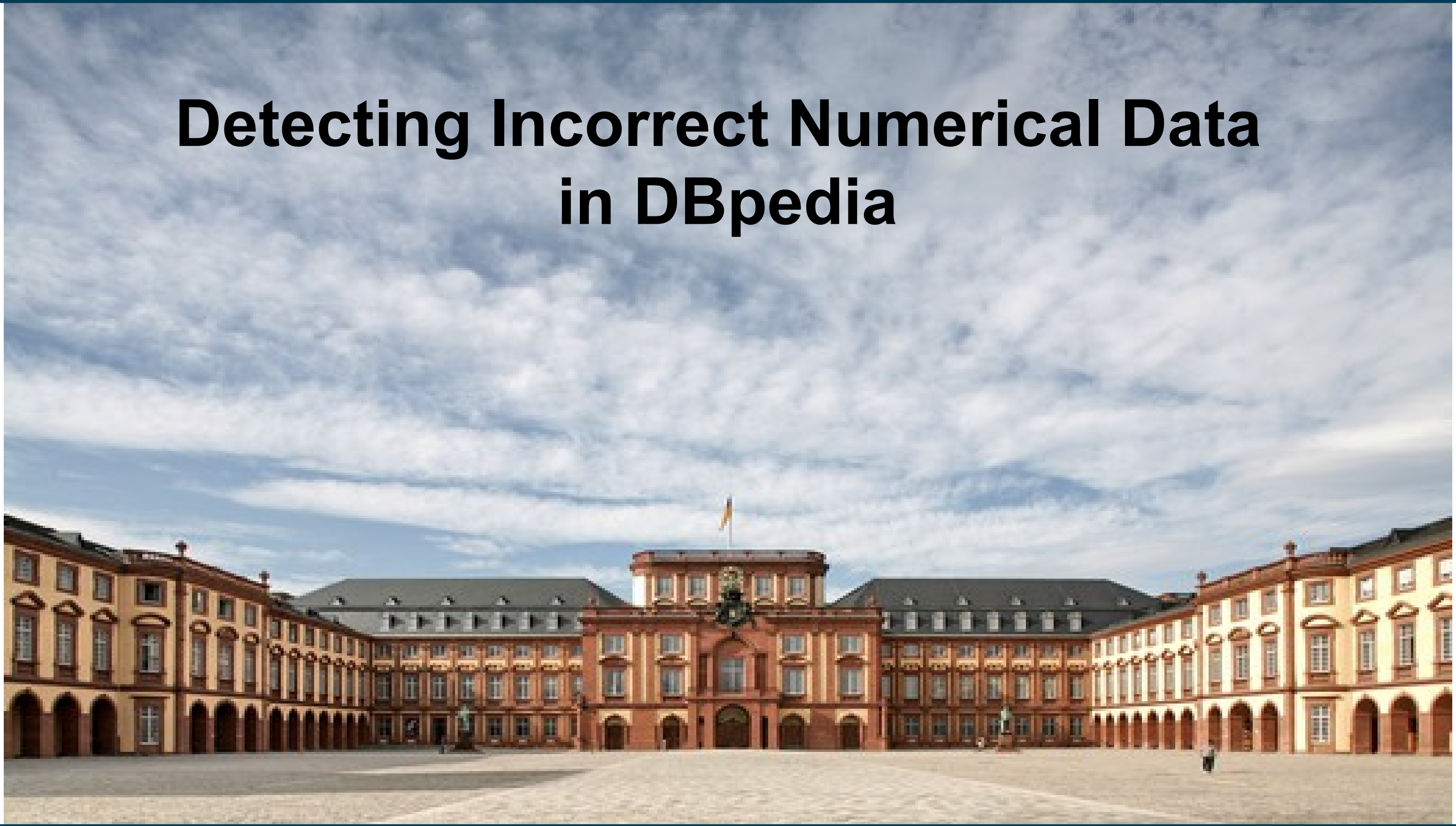
- Handling natural outliers
  - cross checking with other sources
  - for DBpedia: other language editions
- Preprocessing techniques
  - e.g., dynamically building a tree of meaningful subpopulations
- Pinpointing errors
  - text pattern induction on outliers found
  - e.g., `[0-9.,]* ([0-9]{4})` (years in parentheses cause problems)
  - could also help identifying natural outliers

Questions?


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

http://xkcd.com/539/

UNIVERSITÄT
MANNHEIM



Detecting Incorrect Numerical Data in DBpedia

Dominik Wienand, **Heiko Paulheim**