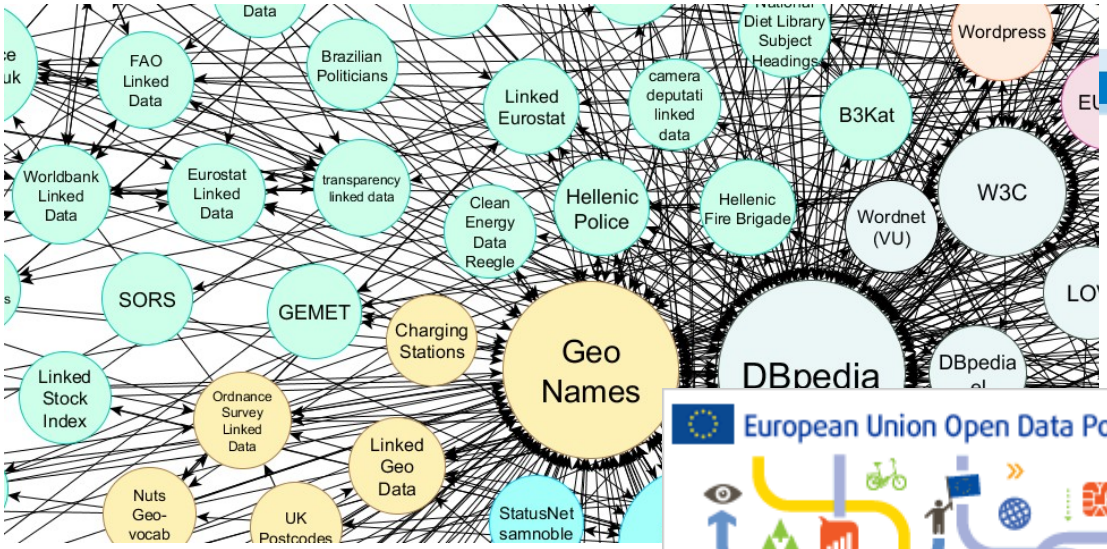
UNIVERSITÄT MANNHEIM

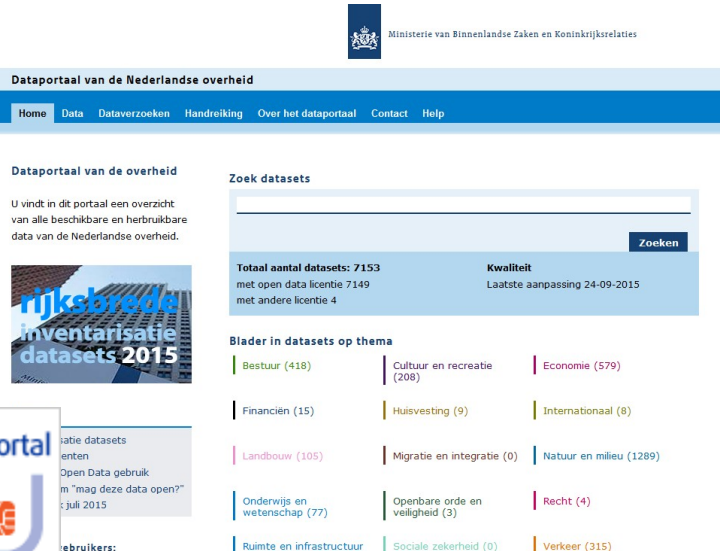
Linked Open Data enhanced Knowledge Discovery

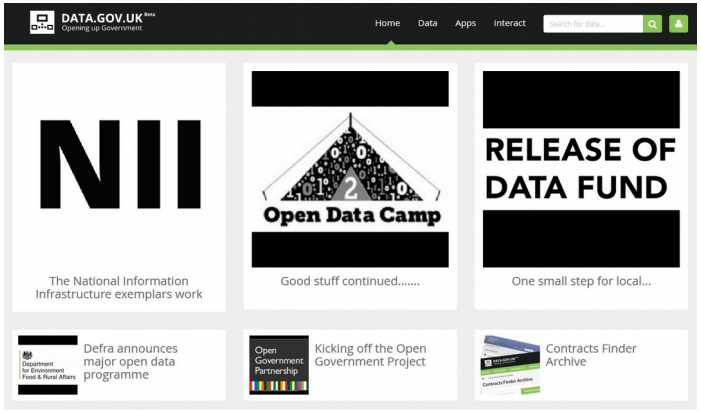
Introducing the RapidMiner Linked Open Data Extension

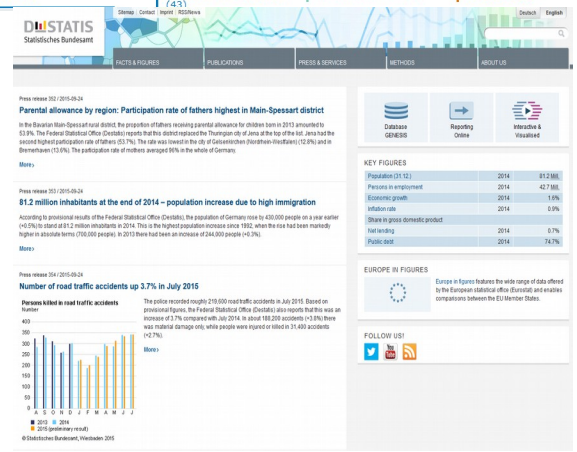
Heiko Paulheim

The Web is Full of Data...



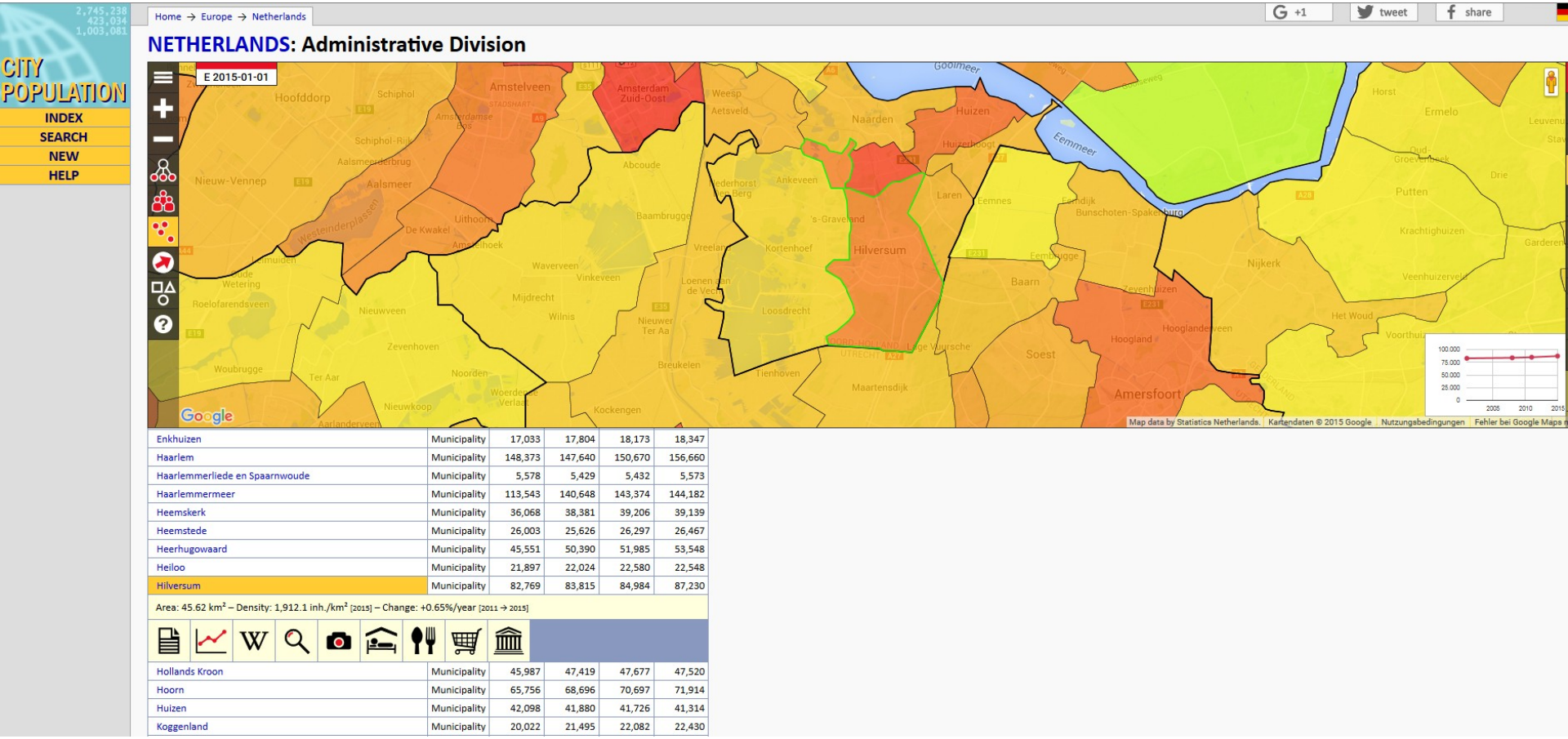






Motivating Example

- Understanding population changes in the Netherlands

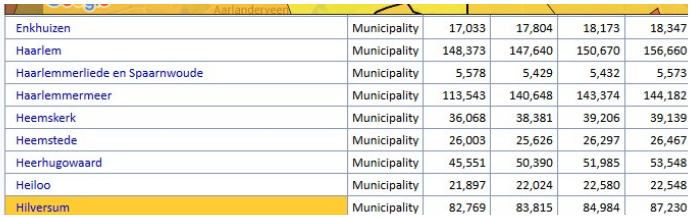


Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230
Area: 45.62 km² – Density: 1,912.1 inh./km² [2015] – Change: +0.65%/year [2011 → 2015]					
Hollands Kroon	Municipality	45,987	47,419	47,677	47,520
Hoorn	Municipality	65,756	68,696	70,697	71,914
Huizen	Municipality	42,098	41,880	41,726	41,314
Koggenland	Municipality	20,022	21,495	22,082	22,430

Motivating Example

- Understanding population changes in the Netherlands
- What we can see in the data
  - population changes by municipality are very diverse
  - ranging from -12% to +53% over the last 15 years

- What we cannot see from the data
  - How do growing regions differ from shrinking ones?
  - Which factors drive people's movements?



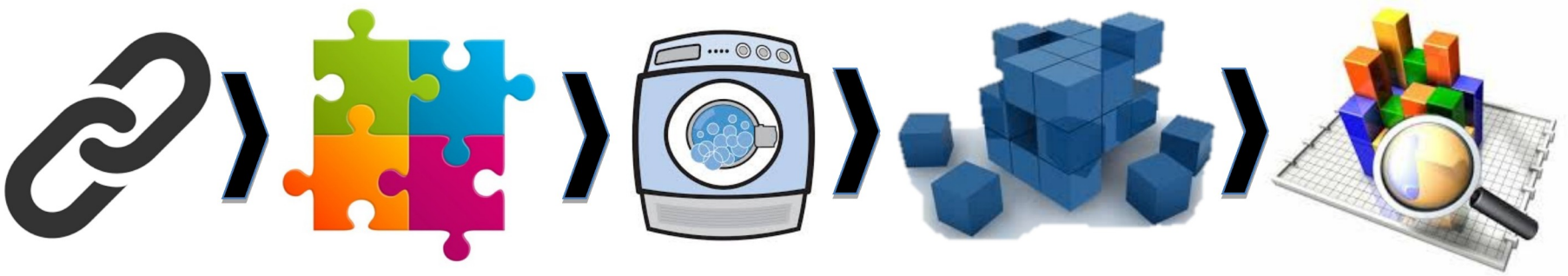
Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230

- As very often, we need more knowledge...

Motivating Example

- Proposed approach:
  - link data at hand to LOD Cloud
  - harvest additional information about regions
  - look for interesting patterns

Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230



link **combine** **cleanse** **transform** **analyze**

RapidMiner Linked Open Data Extension

Introducing RapidMiner:

- An open source platform for data mining and predictive analytics
- Processes are designed by wiring operators in a GUI
  (no programming)
- Operators for data loading, transformation, modeling, visualization, …
- Scalable, distributed, parallel processing in a cloud environment
- 200,000 active users

- Developers can write their own *extensions*

RapidMiner Linked Open Data Extension

- The extension adds operators for
  - accessing local and remote (Linked and non Linked) data
  - linking local to remote data
  - combining data from various sources
  - automatically following links to other datasets

- Data analysts can use it without knowing RDF, SPARQL, etc.

Example Use Case

- Understanding population changes in the Netherlands

- RapidMiner workflow:
  - Import original table
  - Link municipalities to DBpedia
    - alternative: link provinces to Eurostat
  - Build enriched table
  - Analyze the results

Enkhuizen	Municipality	17,033	17,804	18,173	18,347
Haarlem	Municipality	148,373	147,640	150,670	156,660
Haarlemmerliede en Spaarnwoude	Municipality	5,578	5,429	5,432	5,573
Haarlemmermeer	Municipality	113,543	140,648	143,374	144,182
Heemskerk	Municipality	36,068	38,381	39,206	39,139
Heemstede	Municipality	26,003	25,626	26,297	26,467
Heerhugowaard	Municipality	45,551	50,390	51,985	53,548
Heiloo	Municipality	21,897	22,024	22,580	22,548
Hilversum	Municipality	82,769	83,815	84,984	87,230

Example Findings

- Growing regions: Flevoland, Utrecht, North/South Holland
- Shrinking regions: Limburg, Groningen, Friesland

- Provincal capitals are growing
- Growth in regions with high population
- Growth in regions with high income
  - but also: growth in regions with high unemployment

Example Findings

- Negative correlation between growth and elevation?!

Behind the Scenes: RapidMiner LOD Extension

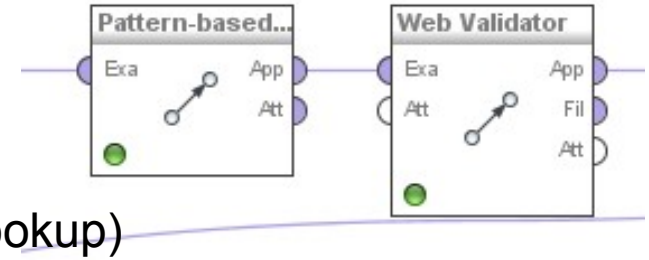
- RapidMiner uses a tabular data model

Input Table | Link | Additional Attributes

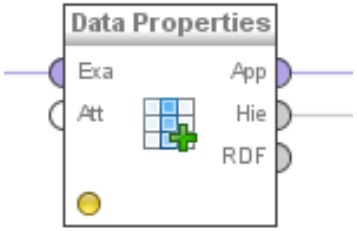
Row No.	cylinders	displacement	horsepower	weight	accelaeration	model	origin	mpg	car	car_uri	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	http://dbped...	ht
1	8	307	130	3504	12	70	1	18	chevrolet che	http://dbpedi	0	0	0	1	1	0	0	0
2	8	350	165	3693	11.500	70	1	15	buick skylark	http://dbpedi	1	0	0	1	1	0	0	0
3	8	318	150	3436	11	70	1	18	plymouth sa	http://dbpedi	1	0	0	1	1	0	0	0
4	8	304	150	3433	12	70	1	16	amc rebel	http://dbpedi	1	0	0	1	1	0	0	0
5	8	302	140	3449	10.500	70	1	17	ford torino	http://dbpedi	1	0	0	1	0	0	0	0
6	8	429	198	4341	10	70	1	15	ford galaxie	http://dbpedi	1	0	0	1	1	0	0	0
7	8	454	220	4354	9	70	1	14	chevrolet imp	http://dbpedi	1	0	0	1	1	0	0	0
8	8	440	215	4312	8.500	70	1	14	plymouth fur	http://dbpedi	1	0	0	1	1	0	0	0
9	8	455	225	4425	10	70	1	14	pontiac catal	http://dbpedi	1	1	0	1	1	0	0	0
10	8	390	190	3850	8.500	70	1	15	amc ambas	http://dbpedi	1	0	0	1	1	0	0	0
11	8	383	170	3563	10	70	1	15	dodge challe	http://dbpedi	1	0	0	1	1	0	0	0
12	8	340	160	3609	8	70	1	14	plymouth 'cu	http://dbpedi	1	0	0	1	1	0	0	0
13	8	400	150	3761	9.500	70	1	15	chevrolet mc	http://dbpedi	1	0	0	1	1	0	0	0
14	8	455	225	3086	10	70	1	14	buick estate	http://dbpedi	1	0	0	1	0	0	0	0
15	4	113	95	2372	15	70	3	24	toyota coron	http://dbpedi	1	0	0	1	1	0	0	0
16	6	198	95	2833	15.500	70	1	22	plymouth du	http://dbpedi	1	0	0	1	0	0	0	0
17	6	199	97	2774	15.500	70	1	18	amc hornet	http://dbpedi	1	0	0	1	1	0	0	0
18	6	200	85	2587	16	70	1	21	ford maveric	http://dbpedi	1	0	0	1	0	0	0	0
19	4	97	88	2130	14.500	70	3	27	datsun	http://dbpedi	0	0	0	1	0	0	0	0
20	4	97	46	1835	20.500	70	2	26	volkswagen	http://dbpedi	1	0	0	1	0	1	0	0
21	4	110	87	2672	17.500	70	2	25	peugeot 504	http://dbpedi	0	0	0	0	0	0	0	0
22	4	107	90	2430	14.500	70	2	24	audi 100	http://dbpedi	0	0	0	0	0	0	0	0
23	4	104	95	2375	17.500	70	2	25	saab 99	http://dbpedi	0	0	0	0	0	0	0	0
24	4	121	113	2234	12.500	70	2	26	bmw 2002	http://dbpedi	0	0	0	0	0	0	0	0
25	6	199	90	2648	15	70	1	21	amc gremlin	http://dbpedi	1	0	0	1	0	0	0	0
26	8	360	215	4615	14	70	1	10	ford f250	http://dbpedi	1	0	0	1	0	0	0	0
27	8	307	200	4376	15	70	1	10	chevy	http://dbpedi	1	0	0	1	1	0	0	0
28	8	318	210	4382	13.500	70	1	11	dodge d	http://dbpedi	0	0	0	1	0	0	0	0

Behind the Scenes: RapidMiner LOD Extension

- Linking local data to LOD Sources
  - based on URI patterns
  - based on text search
  - using specialized services (e.g., DBpedia Lookup)



- Following links
  - e.g., automatically follow all owl:sameAs links to other datasets to a certain depth



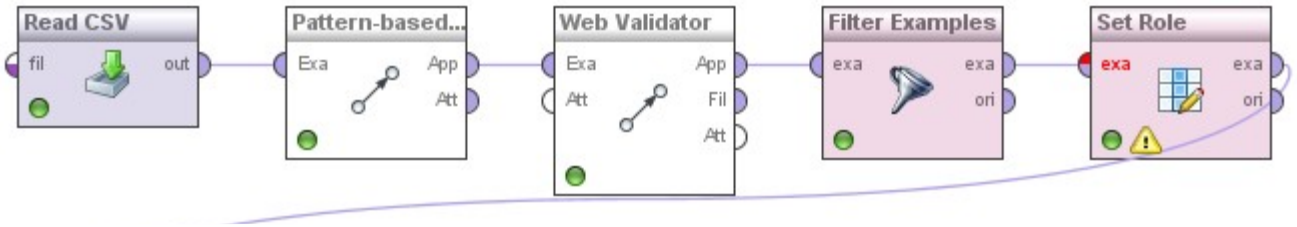
- Harvesting attributes
  - e.g., add all numeric attributes found
  - built-in support for aggregations

Behind the Scenes: RapidMiner LOD Extension

- Matching and fusion
  - e.g., many sources contain “population” as an attribute
  - automatic identification of similar attributes
  - automatic fusion using different policies

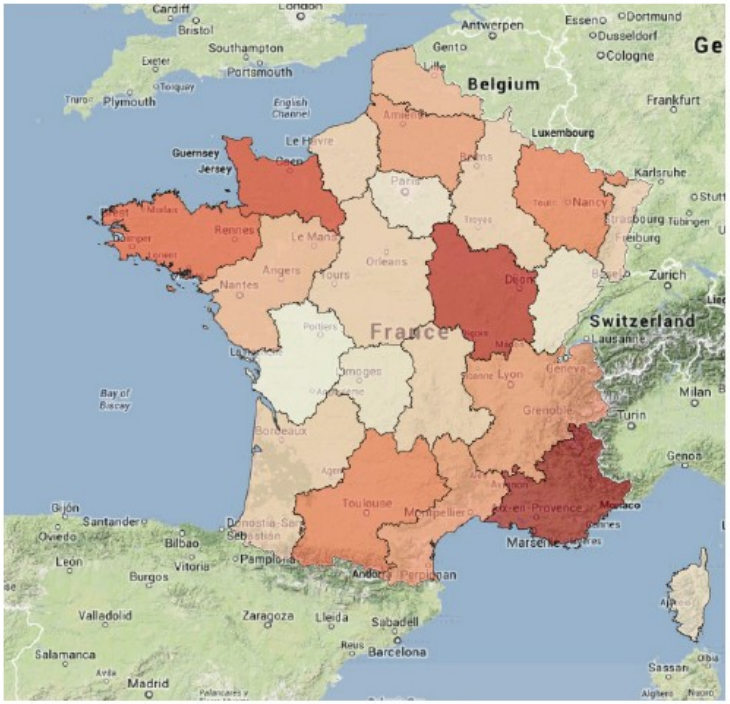
- Attribute set filtering
  - exploiting schema information
  - more effective in finding redundant attributes

Full RapidMiner Workflow for the Example

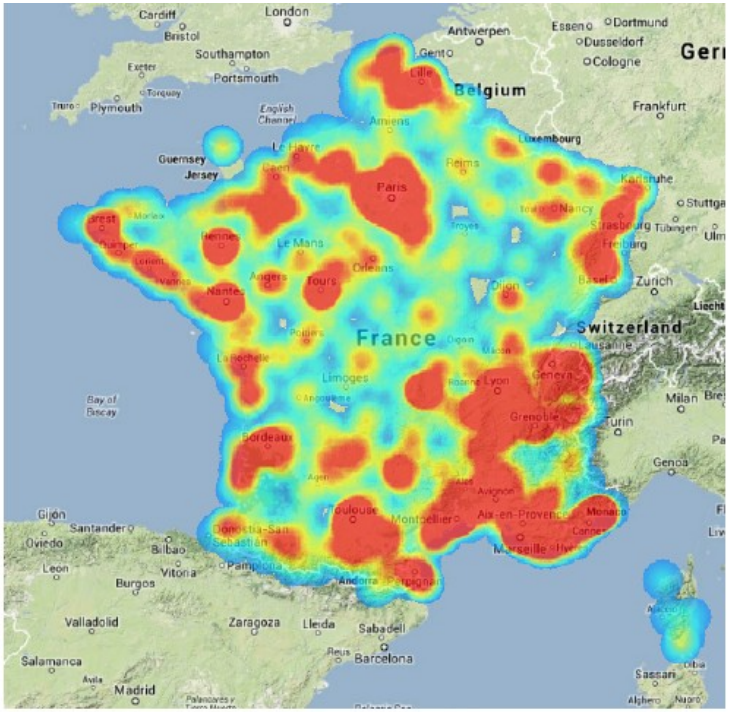


Other Examples

- Analyzing unemployment in France (SemStats'13)
  - using background knowledge from DBpedia, Eurostat, Linked Geo Data
  - exploiting links from DBpedia to GADM for visualization



(a) Unemployment by region



(b) Heat map of police stations

Other Examples

- Example correlations 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)

Other Examples

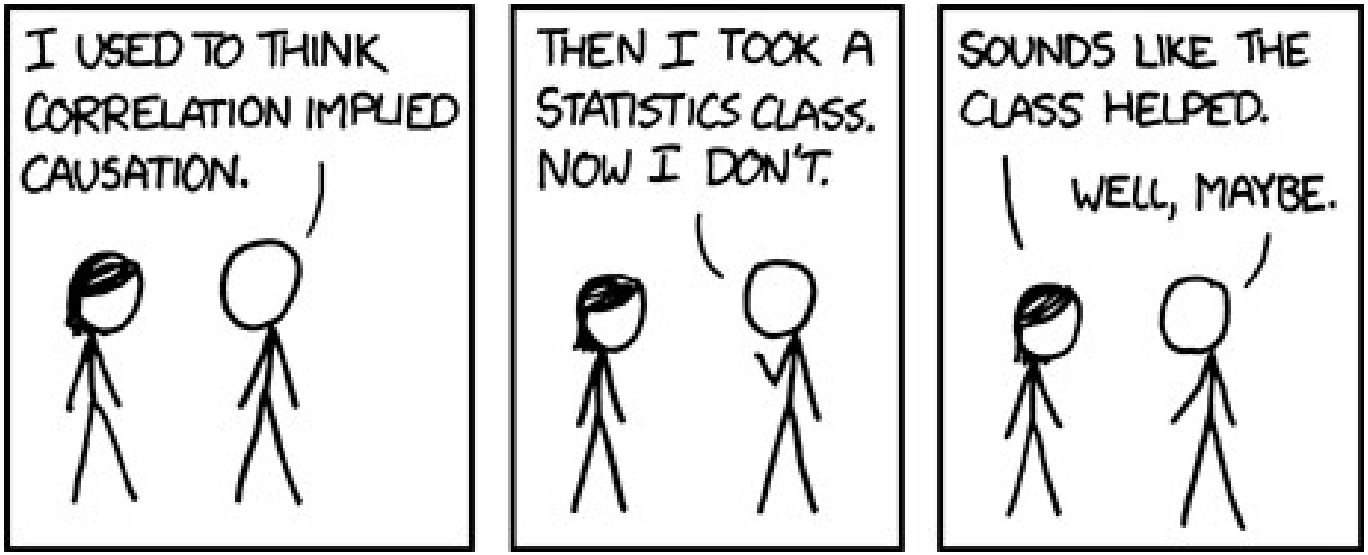
- Data Set: Suicide rates by country
  - http://www.washingtonpost.com/wp-srv/world/suiciderate.html
- Findings for suicide rates
  - Democracies have lower suicide rates than other forms of government
  - High HDI → low suicide rate
  - High population density → high suicide rate
  - By geography:
    - At the sea → low
    - In the mountains → high
  - High Gini index → low suicide rate
    - High Gini index ↔ unequal distribution of wealth
  - High usage of nuclear power → high suicide rates

Other Examples

- Data set: Durex worldwise survey on sexual activity
  - http://chartsbin.com/view/uya
- Findings:
  - By geography:
    - High in Europe, low in Asia
    - Low in Island states
  - By language:
    - English speaking: low
    - French speaking: high
  - Low average age → high activity
  - High GDP per capita → low activity
  - High unemployment rate → high activity
  - High number of ISP providers → low activity

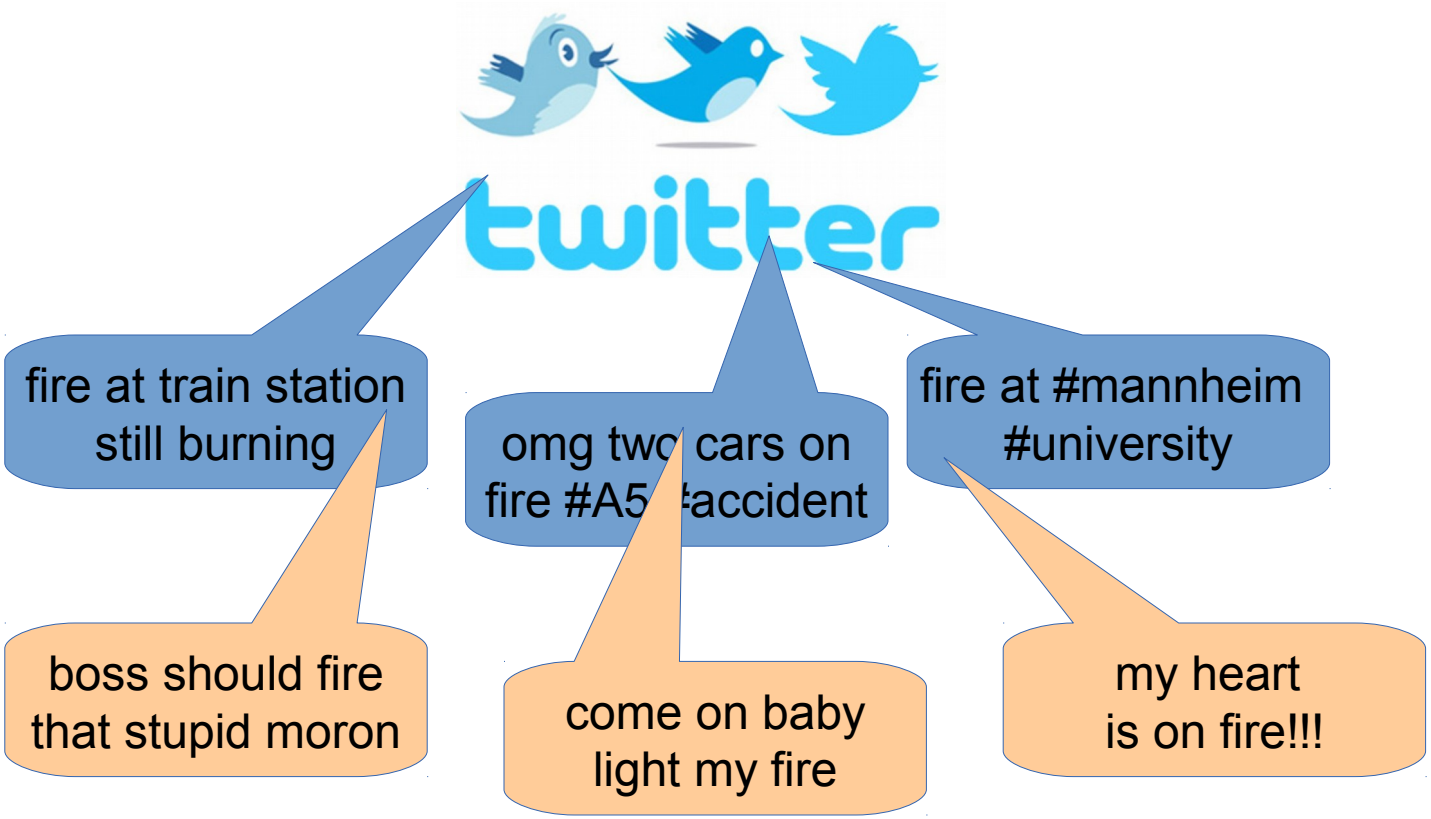
Caveat

- We have only been analyzing *correlations* here.



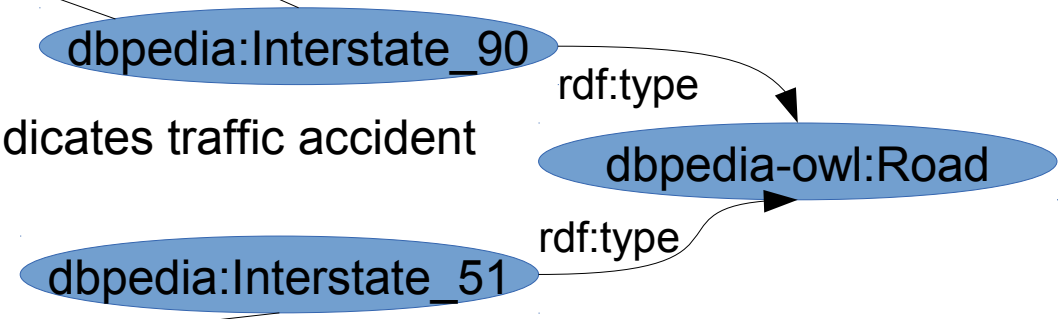
Other Use Cases

- Incident detection from Twitter



Other Use Cases

- Example set:
  - "Again crash on I90"
  - "Accident on I90"



- Model:
  - dbpedia-owl:Road → indicates traffic accident
- Applying the model:
  - "Two cars collided on I51" → indicates traffic accident
- Using LOD+RapidMiner
  - automatically learns a model
  - avoids overfitting

Other Use Cases

- Building Semantic Recommeder Systems (ESWC'14)
- Combines two extensions:
  - Linked Open Data extension
  - Recommender system extension
- Use data about books
  for content-based recommender
  - best system (out of 24)
    on two out of three tasks
  - used data from DBpedia
    and RDF Book Mashup

What is Special about Hilversum?

- Compare Hilversum to other Cities in the Netherlands
  - find distinctive features

- Finding the needles in the haystack of statements about Hilversum in DBpedia

What is Special about Hilversum?

- Compare Hilversum to other Cities in the Netherlands
  - find distinctive features

Top Facts for **Hilversum** compared to the entities of the class **Cities In The Netherlands**:

Show 10 entries Search:

Expand	Statements	Rules
		Entities of type **CitiesInTheNetherlands** usually have **isPartOf**, but **Hilversum** doesn't have!
		Entities of type **CitiesInTheNetherlands** usually don't have **hometown**, but **Hilversum** has!
		Entities of type **CitiesInTheNetherlands** usually are not of type **OlympicModernPentathlonVenues**, but **Hilversum** is!
		Entities of type **CitiesInTheNetherlands** usually don't have **recordedIn**, but **Hilversum** has!
		Entities of type **CitiesInTheNetherlands** usually are not of type **OlympicEquestrianVenues**, but **Hilversum** is!
		Entities of type **CitiesInTheNetherlands** usually don't have **headquarter**, but **Hilversum** has!
		Entities of type **CitiesInTheNetherlands** usually don't have **location**, but **Hilversum** has!
		Entities of type **CitiesInTheNetherlands** usually don't have **ground**, but **Hilversum** has!
		Entities of type **CitiesInTheNetherlands** usually are not of type **PopulatedPlacesInNorthHolland**, but **Hilversum** is!
		Entities of type **CitiesInTheNetherlands** usually are not of type **1928SummerOlympicVenues**, but **Hilversum** is!

Showing 1 to 10 of 10 entries Previous 1 Next

What is Special about Hilversum?

- Compare Hilversum to other Cities in the Netherlands
  - find distinctive features

- TopFacts application
  - Demonstration at ISWC 2015
  - Combines Linked Open Data with attribute-wise outlier detection [see Paulheim/Meusel, Machine Learning 100(2-3), 2015]

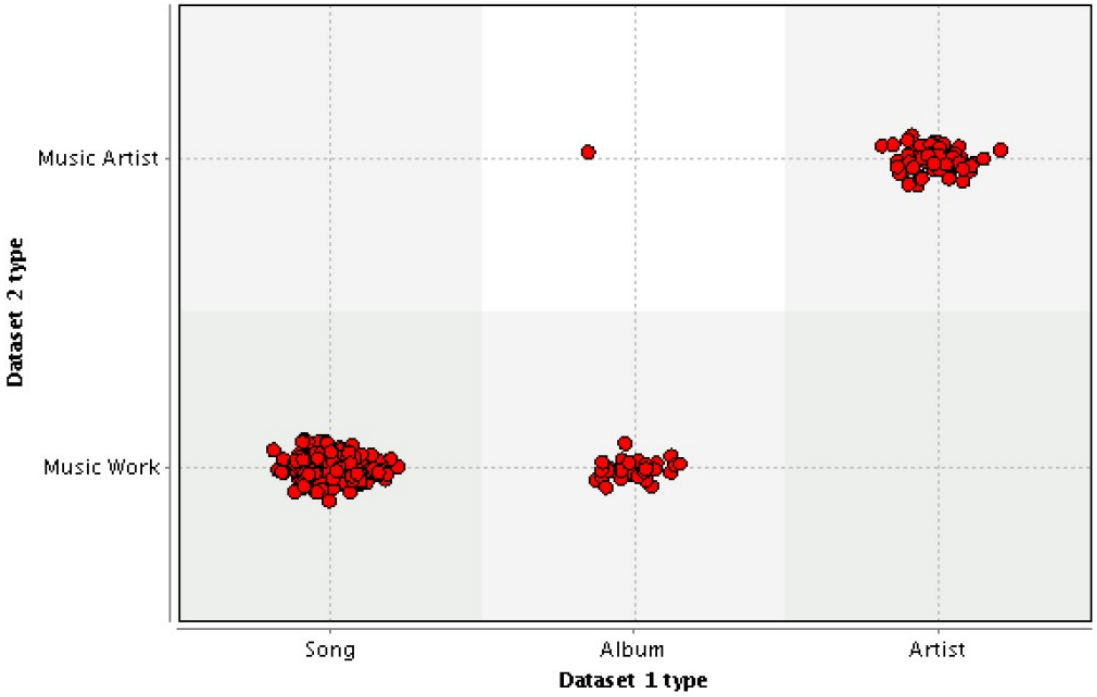
What is Special about Hilversum?

- Compare Hilversum to other Cities in the Netherlands
  - find distinctive features

- Hilversum is
  - a city where the modern Pentathlon olympics have been held
  - the headquarter of many media companies
  - a place where many music recordings have been made

Other Use Cases

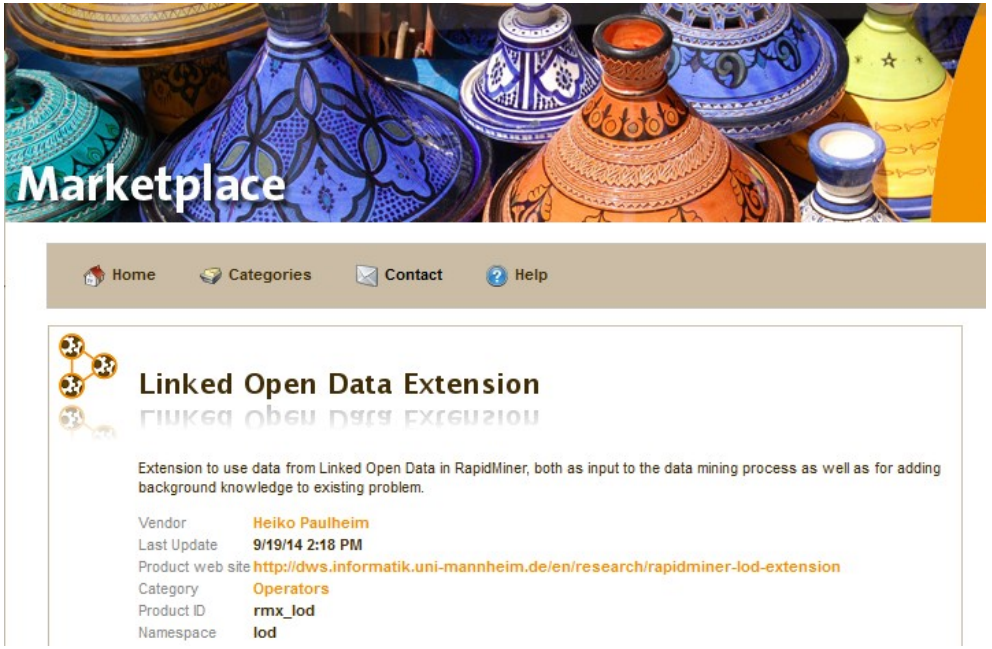
- Debugging Linked Open Data
  - loading a set of statements
  - augment with additional features
  - run outlier detection
    - again: a special extension
- Example: identify wrong dataset interlinks (WoDOOM'14)
  - AUC up to 85%



Summary

- The RapidMiner LOD Extension
  - brings data analysis to the web of data
  - can be used by data analysts without learning SPARQL

- Availability
  - on the RapidMiner marketplace
  - installable from inside RapidMiner
  - >9,000 installations and counting



Take Home Messages

- The Web is full of data
  - ...and more and more becomes Linked Data
- Intelligent data processing
  - helps unlocking the potential of that data
  - enables intelligent applications
- A good fit
  - Sophisticated analytics platforms (e.g., RapidMiner), and
  - Linked Open Data

Feedback?

Heiko Paulheim

heiko@informatik.uni-mannheim.de
@heikopaulheim

UNIVERSITÄT MANNHEIM

Linked Open Data enhanced Knowledge Discovery

Introducing the RapidMiner Linked Open Data Extension

Heiko Paulheim