UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim

Data vs. Knowledge

1982

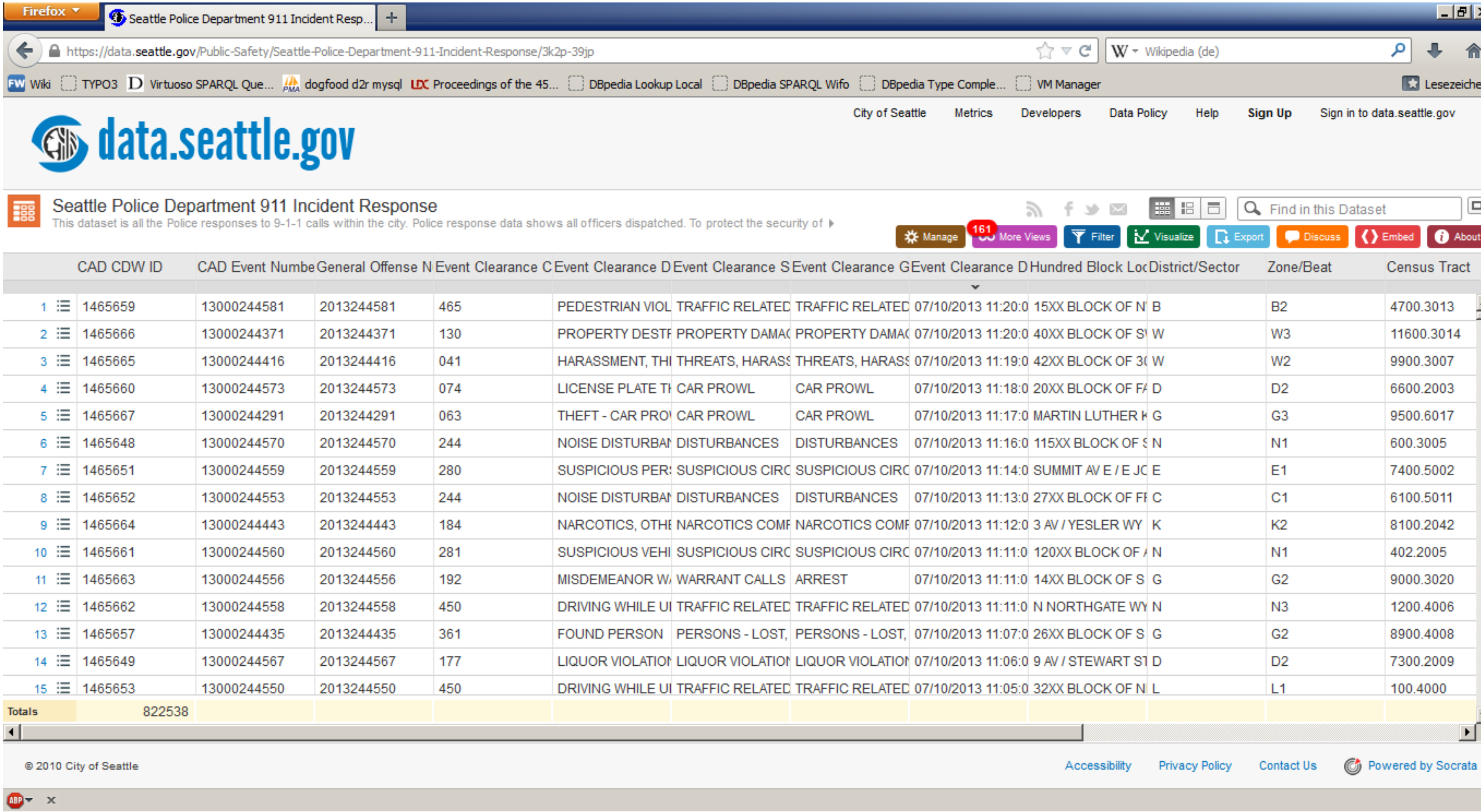
"We are drowning in data,
but starving for **knowledge**"

John Naisbitt

Data vs. Knowledge

- There is plenty of data
  - Linked (Open) Data
  - Government Data
  - Sensor Data
  - Social Networks
  - …
- ...but data is not knowledge
- Knowledge is required for
  - Assessments
  - Actions

Crime Data Live on the Web



	CAD CDW ID	CAD Event Numbe	General Offense N	Event Clearance C	Event Clearance D	Event Clearance S	Event Clearance G	Event Clearance D	Hundred Block Loc	District/Sector	Zone/Beat	Census Tract
1	1465659	13000244581	2013244581	465	PEDESTRIAN VIOL	TRAFFIC RELATED	TRAFFIC RELATED	07/10/2013 11:20:0	15XX BLOCK OF N	B	B2	4700.3013
2	1465666	13000244371	2013244371	130	PROPERTY DESTI	PROPERTY DAMA	PROPERTY DAMA	07/10/2013 11:20:0	40XX BLOCK OF S	W	W3	11600.3014
3	1465665	13000244416	2013244416	041	HARASSMENT, THI	THREATS, HARASS	THREATS, HARASS	07/10/2013 11:19:0	42XX BLOCK OF 3	W	W2	9900.3007
4	1465660	13000244573	2013244573	074	LICENSE PLATE TI	CAR PROWL	CAR PROWL	07/10/2013 11:18:0	20XX BLOCK OF F	D	D2	6600.2003
5	1465667	13000244291	2013244291	063	THEFT - CAR PRO	CAR PROWL	CAR PROWL	07/10/2013 11:17:0	MARTIN LUTHER K	G	G3	9500.6017
6	1465648	13000244570	2013244570	244	NOISE DISTURBAI	DISTURBANCES	DISTURBANCES	07/10/2013 11:16:0	115XX BLOCK OF	N	N1	600.3005
7	1465651	13000244559	2013244559	280	SUSPICIOUS PER	SUSPICIOUS CIRC	SUSPICIOUS CIRC	07/10/2013 11:14:0	SUMMIT AV E / E JC	E	E1	7400.5002
8	1465652	13000244553	2013244553	244	NOISE DISTURBAI	DISTURBANCES	DISTURBANCES	07/10/2013 11:13:0	27XX BLOCK OF FI	C	C1	6100.5011
9	1465664	13000244443	2013244443	184	NARCOTICS, OTHI	NARCOTICS COMI	NARCOTICS COMI	07/10/2013 11:12:0	3 AV / YESLER WY	K	K2	8100.2042
10	1465661	13000244560	2013244560	281	SUSPICIOUS VEHI	SUSPICIOUS CIRC	SUSPICIOUS CIRC	07/10/2013 11:11:0	120XX BLOCK OF	N	N1	402.2005
11	1465663	13000244556	2013244556	192	MISDEMEANOR W	WARRANT CALLS	ARREST	07/10/2013 11:11:0	14XX BLOCK OF S	G	G2	9000.3020
12	1465662	13000244558	2013244558	450	DRIVING WHILE UI	TRAFFIC RELATED	TRAFFIC RELATED	07/10/2013 11:11:0	N NORTHGATE WY	N	N3	1200.4006
13	1465657	13000244435	2013244435	361	FOUND PERSON	PERSONS - LOST,	PERSONS - LOST,	07/10/2013 11:07:0	26XX BLOCK OF S	G	G2	8900.4008
14	1465649	13000244567	2013244567	177	LIQUOR VIOLATIOI	LIQUOR VIOLATIOI	LIQUOR VIOLATIOI	07/10/2013 11:06:0	9 AV / STEWART ST	D	D2	7300.2009
15	1465653	13000244550	2013244550	450	DRIVING WHILE UI	TRAFFIC RELATED	TRAFFIC RELATED	07/10/2013 11:05:0	32XX BLOCK OF N	L	L1	100.4000
Totals	822538											

Crime Data Live on the Web

- data.seattle.gov:
  - Live 911 call data
  - Open for mashups
- Problem for the response team:
  - Quick decisions required
    - With severe effects...
  - Minimal information
  - Missing background knowledge
    - ...but lots of potential sources

There's a Fire!

- Knowledge:
  - What sort of fire?
  - Where?
- Assessment:
  - Relevant from irrelevant
  - Useful from useless
- Action:
  - Maybe send someone
  - Evacuate the neighborhood

There's a Fire!

- How to tell the severity of an (incoming) emergency call?
- What is it sort of emergency (fire alarm, first aid call, shooting)?
- What is its context? E.g., for fires
  - is it near a gas station or a pipeline?
  - are there any schools/kindergartens nearby?
  - is a hospital affected?
- What is its context? E.g., for shootings
  - what are possible escape routes?
  - are there any schools/kindergartens nearby?

There's a Fire!

- For answering these questions, background knowledge is required

- E.g., Linked Geo Data
  - Information about objects with coordinates
  - Queries such as: give me all objects within 50m of (lat,long)

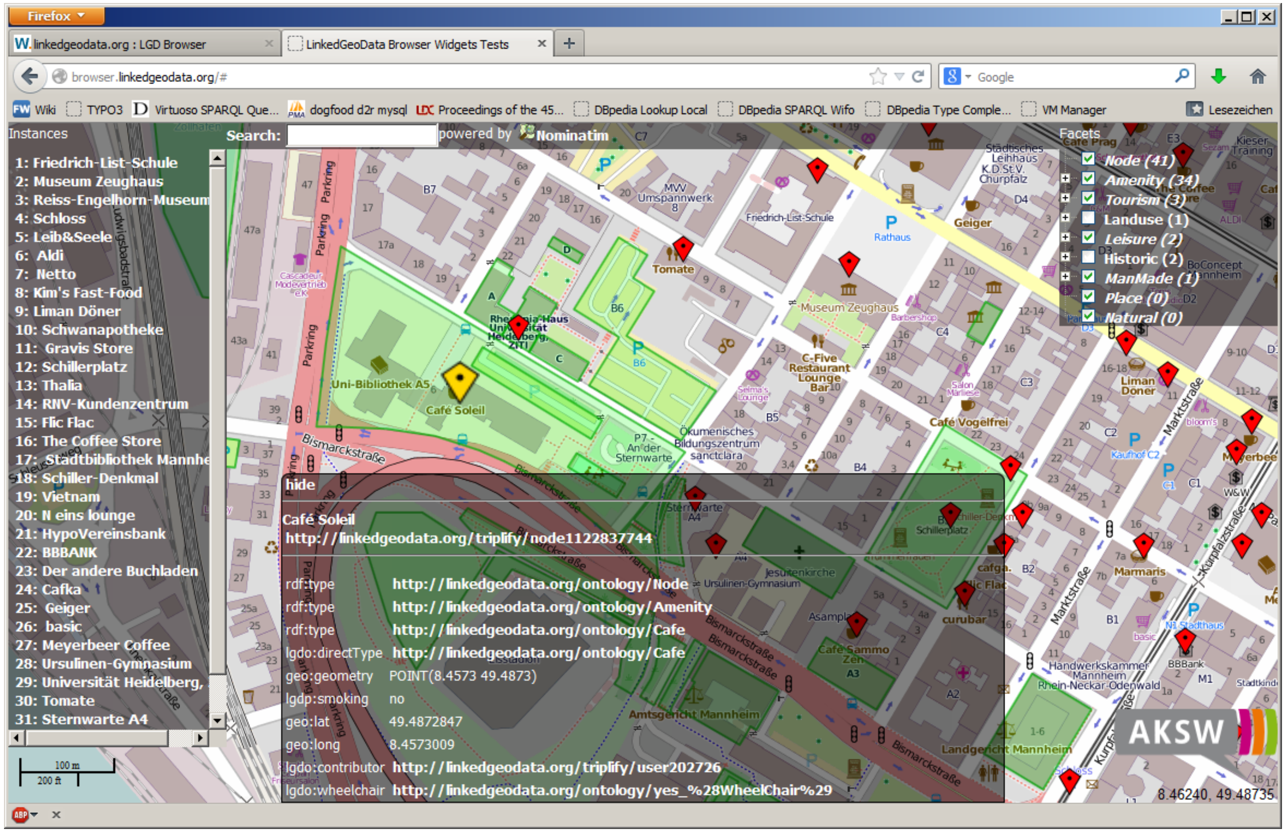
- As for incidents, relevance of Linked Geo Data needs to be assessed
  - gas stations and pipelines may be relevant
  - phone booths and statues probably are not
  - based on user-defined rules

Linked GeoData

www.linkedgeodata.org

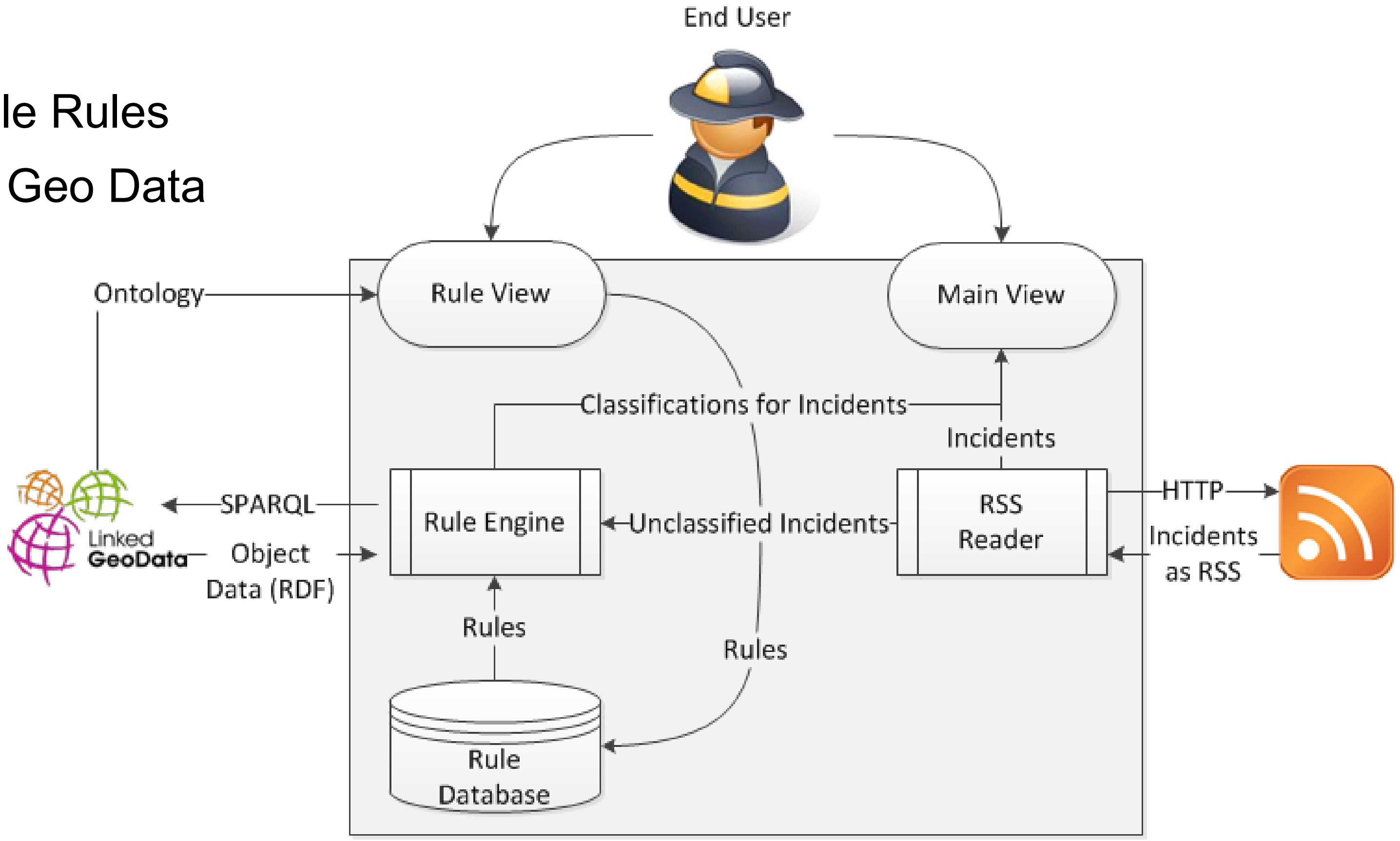
Interlude: Linked Geo Data

- Wraps data from Open Street Maps as LOD
- Objects with coordinates

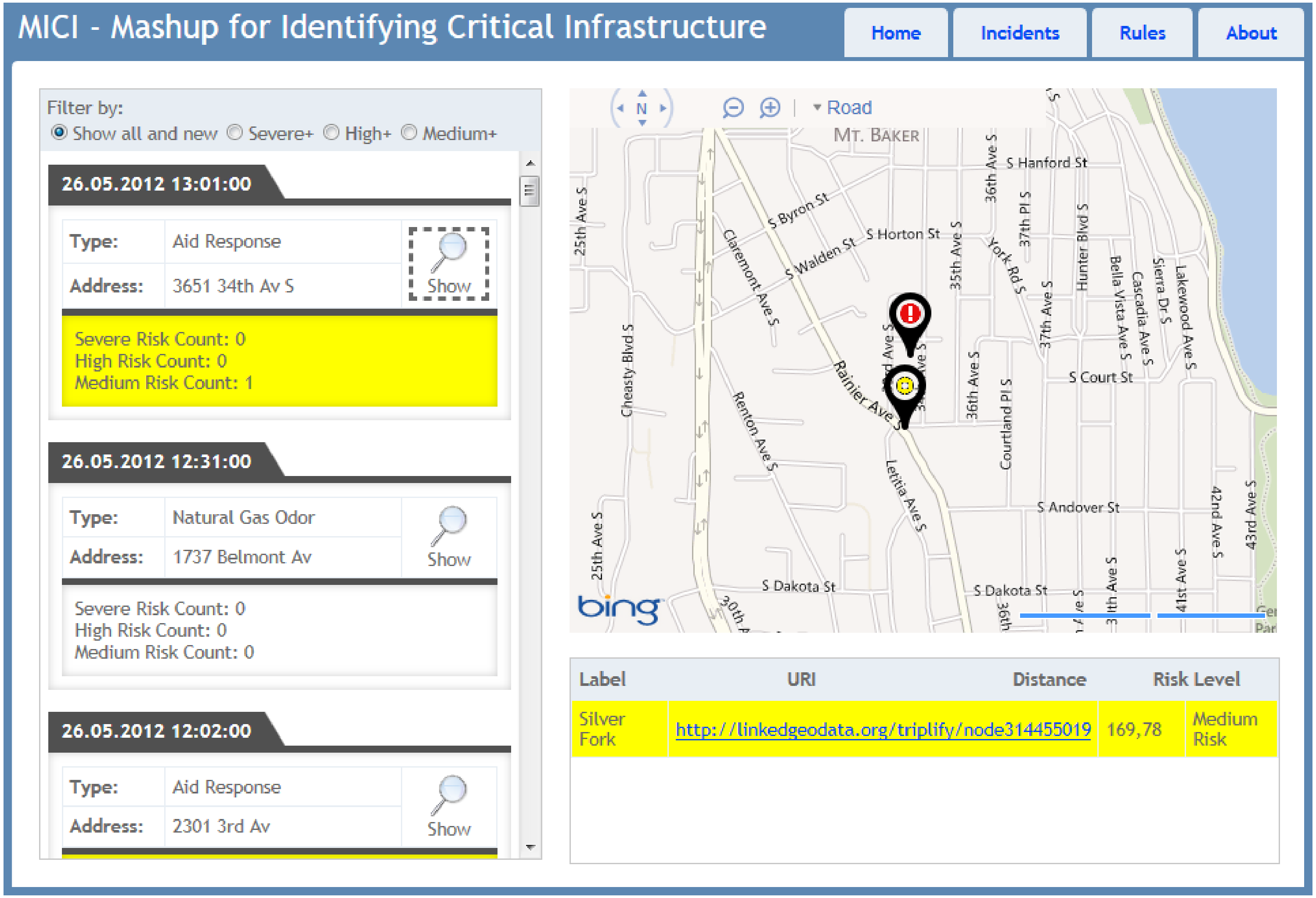


There's a Fire!

- MICI: Live emergency calls from the city of Seattle
  - provided as RSS
- Plus
  - Example Rules
  - Linked Geo Data



There's a Fire!



There's a Fire!

MICI - Mashup for Identifying Critical Infrastructure

Home | Incidents | Rules | About

Index

Create New

Incident Type	Radius	Object Types	Risk Level	Actions
Aid Response	50	http://linkedgeodata.org/ontology/Parking;fuel http://linkedgeodata.org/ontology/Parking;restaurant;fuel http://linkedgeodata.org/ontology/Fuel;carWash 1 more object types Details	Severe Risk	Edit | Delete | Duplicate
Aid Response	100	http://linkedgeodata.org/ontology/Line http://linkedgeodata.org/ontology/PowerSwitch http://linkedgeodata.org/ontology/SubStation,http://linkedgeodata.org/ontology/Substation 5 more object types Details	Severe Risk	Edit | Delete | Duplicate
Aid Response	100	http://linkedgeodata.org/ontology/OilPlatform http://linkedgeodata.org/ontology/Pump http://linkedgeodata.org/ontology/Pipeline 2 more object types Details	Severe Risk	Edit | Delete | Duplicate
Aid Response	250	http://linkedgeodata.org/ontology/Cafe http://linkedgeodata.org/ontology/Shop http://linkedgeodata.org/ontology/Restaurant 7 more object types Details	Medium Risk	Edit | Delete | Duplicate

There's a Fire!

- MICI live: http://mici.tk.informatik.tu-darmstadt.de/
- Recap:
  - Plenty of data (incoming 911 messages)
  - Massive background information (Linked Geo Data)
  - Filtering based on rules
  - Helps: assessing information and acting properly
- Limitations:
  - Data is already preprocessed (RSS from data.seattle.gov)
  - Rules are created manually

UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime ✓

Chris Bizer, Heiko Paulheim, University of Mannheim

Brief Interlude: Machine Learning Basics

- An essential ingredient to intelligent applications
  - Dealing with new pieces of knowledge
  - Handling unknown situations
  - Adapting to users' needs
  - Making predictions for the future

- Inductive vs. Deductive Reasoning:
  - Deductive: rules + facts → facts
  - Inductive: facts → rules

Brief Interlude: Machine Learning Basics

- Example: learning a new concept, e.g., "Tree"

"tree"

"tree"

"tree"

"not a tree"

"not a tree"

"not a tree"

Brief Interlude: Machine Learning Basics

- Example: learning a new concept, e.g., "Tree"
  - we look at (positive and negative) examples
  - ...and derive a *model*
    - e.g., "Trees are big, green plants"






- Goal: Classification of new instances

"tree?"

Warning:
Models are only approximating examples!
Not guaranteed to be correct or complete!

Brief Interlude: Machine Learning Basics

- Typical tasks:
  - Classification (binary or multi-label)
  - Regression (i.e., predicting numerical values)
  - Clustering (finding groups of objects)
  - Frequent Pattern Mining
  - ...

- Methods:
  - Statistical approaches (Naive Bayes, Support Vector Machines, …)
  - Symbolic approaches (Rules, Decision Trees, …)
  - ...

Linked Open Data for Machine Learning

- Example machine learning task: predicting book sales

ISBN	City	Sold

ISBN	City	Population	...	Genre	Publisher	...	Sold
3-2347-3427-1	Darm-stadt	144402	...	Crime	Bloody Books	...	124
3-43784-324-2	Mann-heim	291458	…	Crime	Guns Ltd.	…	493
3-145-34587-0	Roß-dorf	12019	...	Travel	Up&Away	...	14
	...						

→ Crime novels sell better in larger cities

Data Mining Framework "FeGeLOD", RapidMiner Plugins

The FeGeLOD Framework

ISBN	City	# sold
3-2347-3427-1	Darmstadt	124

Named Entity Recognition

ISBN	City	City_URI	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	124

Feature Generation

ISBN	City	City_URI	City_URI_dbpedia-owl:populationTotal	City_URI_...	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	141471	...	124

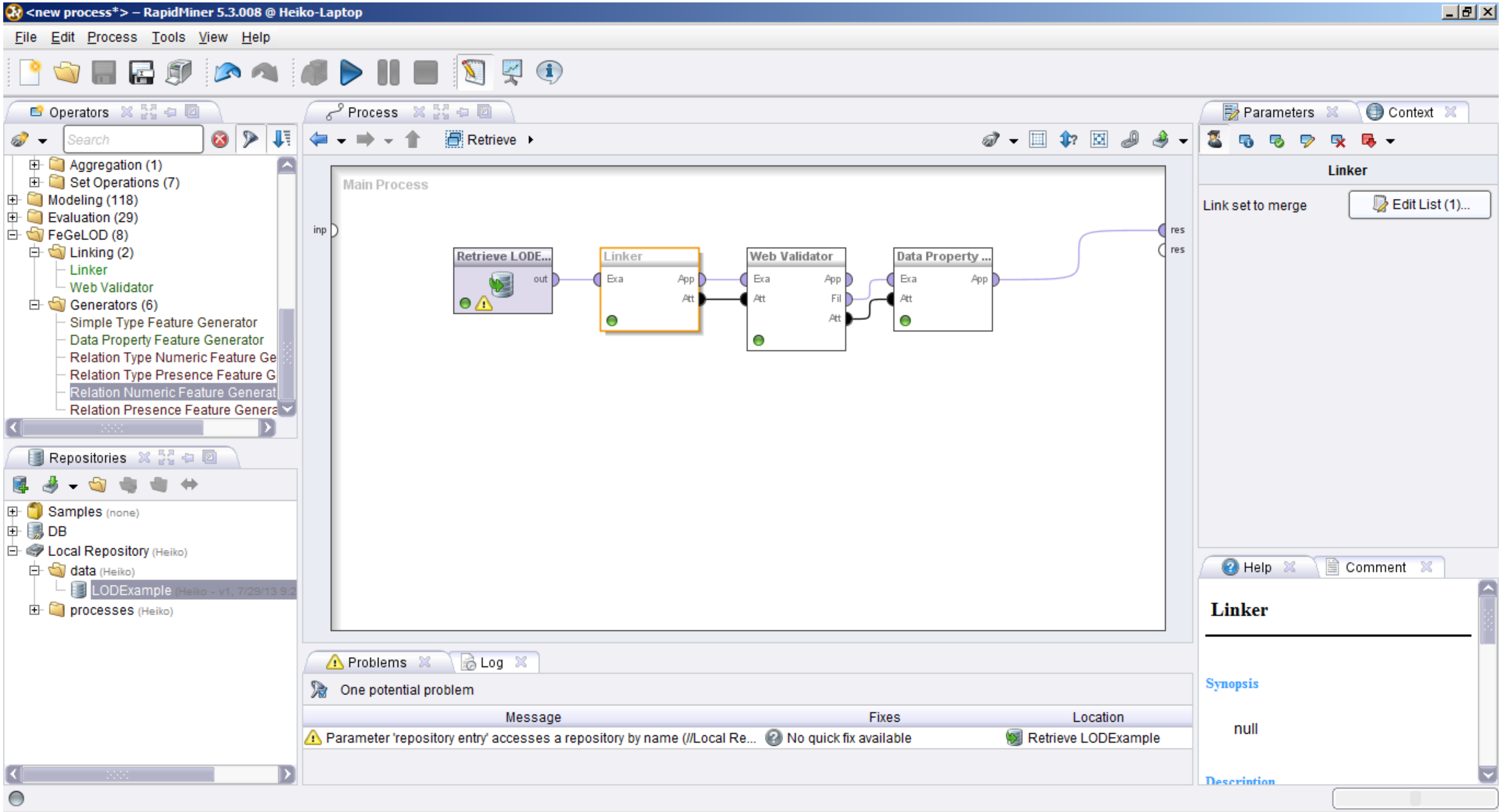
Feature Selection

ISBN	City	City_URI	City_URI_dbpedia-owl:populationTotal	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	141471	124

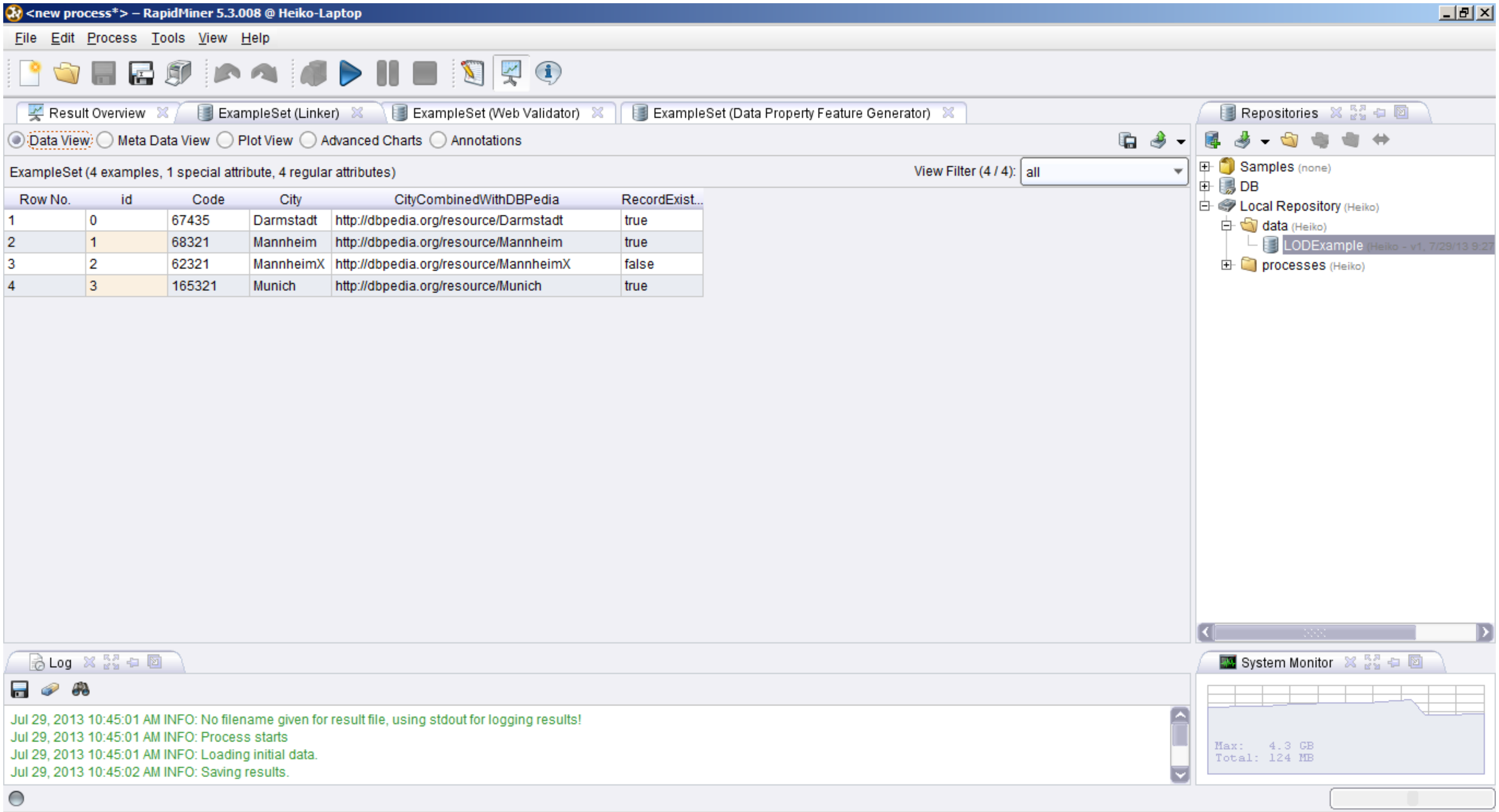
The FeGeLOD Framework

- Entity Recognition
  - Simple approach: guess DBpedia URIs
  - Hit rate >95% for cities and countries (by English name)
- Feature Generation
  - Different Generators
    - Data values (including heuristic numerical conversion)
    - Classes (plus transitive closure)
    - Quantifying Unqualified relations (boolean or numeric)
    - Quantifying Qualified relations (boolean or numeric)
- Feature Selection
  - Filter noise: >95% unknown, identical, or different nominals

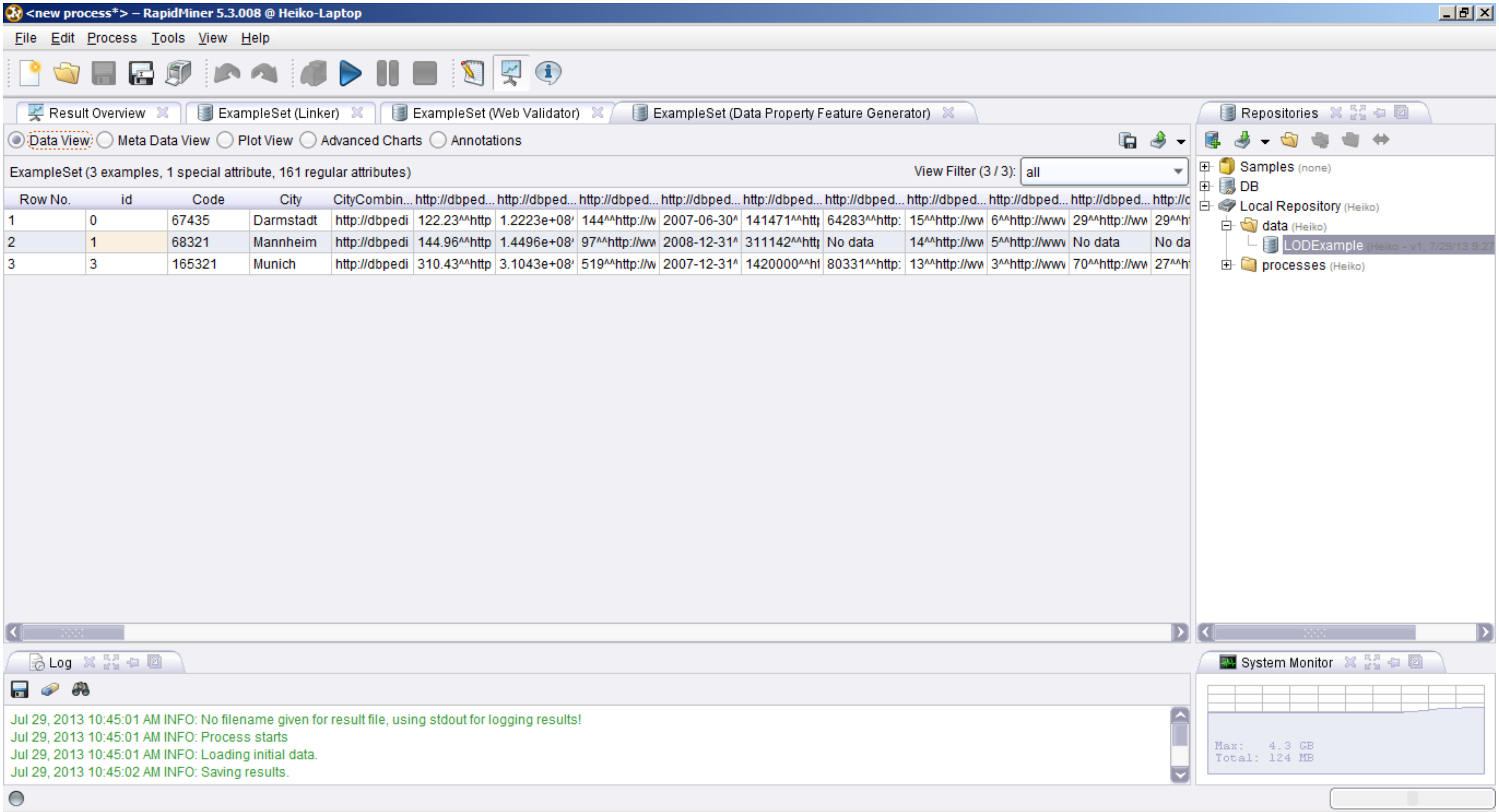
The FeGeLOD Prototype
(now: RapidMiner Linked Open Data Extension)



The FeGeLOD Prototype (now: RapidMiner Linked Open Data Extension)

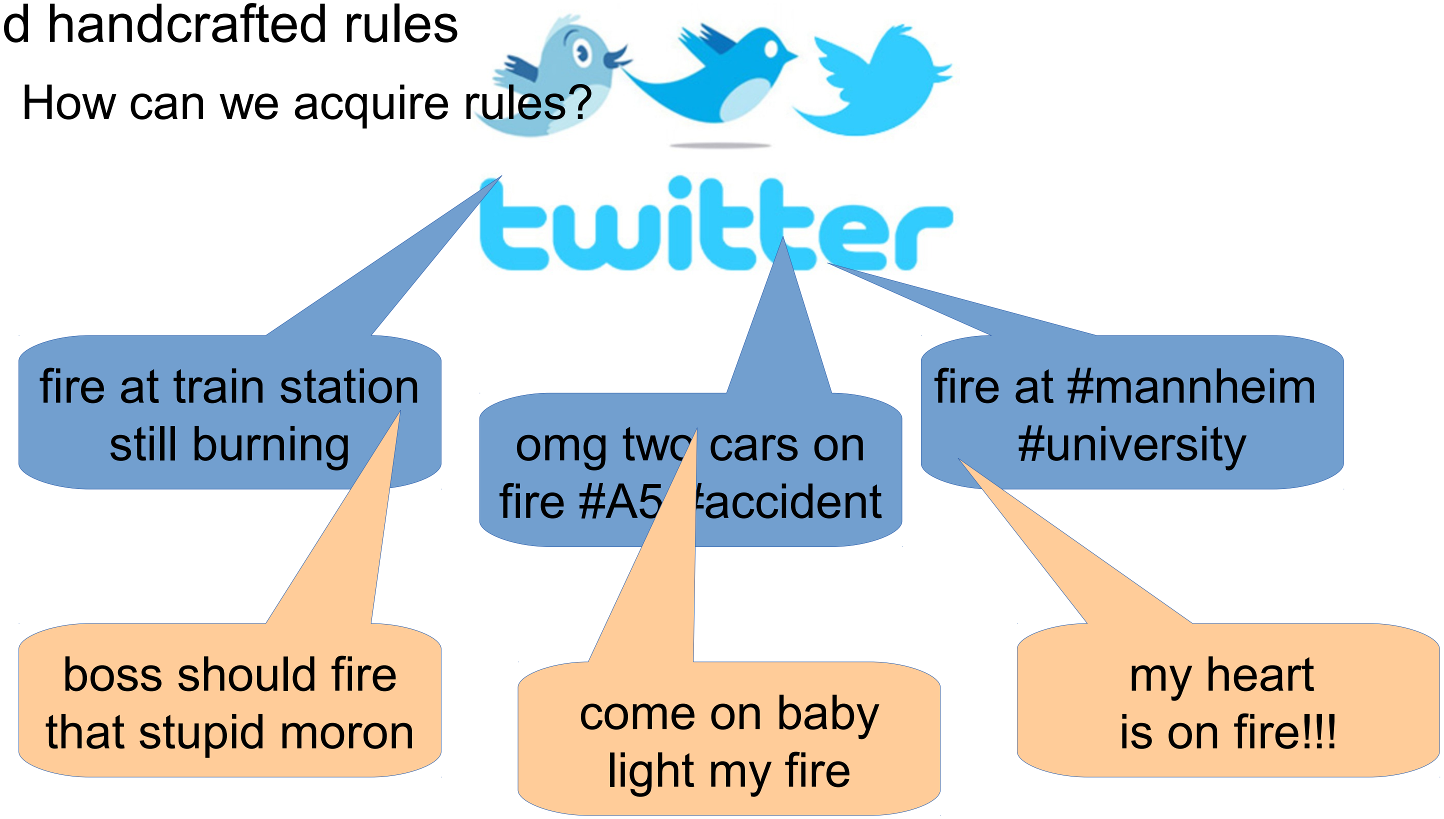


The FeGeLOD Prototype (now: RapidMiner Linked Open Data Extension)



Back to Incidents

- So far, we were using preprocessed RSS data
  - How can we detect incidents automatically?
- And handcrafted rules
  - How can we acquire rules?



Detecting Incidents from Social Media

- Social media contains data on many incidents
  - But keyword search is not enough
  - Detecting small incidents is hard
  - Manual inspection is too expensive (and slow)

- Machine learning could help
  - Train a model to classify incident/non incident tweets
  - Apply model for detecting incident related tweets

- Training data:
  - Traffic accidents
  - ~2,000 tweets containing relevant keywords (“car”, “crash”, etc.), hand labeled (50% related to traffic incidents)

Detecting Incidents from Social Media

- Learning to classify tweets:
  - Positive and negative examples
  - Features:
    - Stemming
    - POS tagging
    - Word n-grams
    - …
- Accuracy ~90%
- But
  - Accuracy drops to ~85% when applying the model to a different city

Brief Interlude: Model Overfitting

- What happens here?
  - Model is trained on a sample of labeled data
  - Tries to identify the characteristics of that data
- Possible effect:
  - Model is too close to training data

Brief Interlude: Model Overfitting

- Extreme example
  - Predict credit rating

- Possible (useful) model:
  - (job status = employed) && (debts<5000) → rating=positive

Name	Net Income	Job status	Debts	Rating
John	40000	employed	0	+
Mary	38000	employed	10000	-
Stephen	21000	self-employed	20000	-
Eric	2000	student	10000	-
Alice	35000	employed	4000	+

Brief Interlude: Model Overfitting

- Extreme example
  - Predict credit rating

- Possible overfit models:
  - (34000<income<36000) || (income>39000) → rating=positive
  - (name=John) || (name=Alice) → rating=positive

Name	Net Income	Job status	Debts	Rating
John	40000	employed	0	+
Mary	38000	employed	10000	-
Stephen	21000	self-employed	20000	-
Eric	2000	student	10000	-
Alice	35000	employed	4000	+

Brief Interlude: Model Overfitting

- All three models perfectly describe the training data
  - But only one is a useful generalization
- Two goals of machine learning (sometimes contradicting):
  - Explain training data as good as possible
  - Find a model that is as general as possible
- Strategies for preventing overfitting:
  - Cross validation
  - Model pruning
  - Stopping criteria in model building
  - Occam's Razor
  - ...

Detecting Incidents from Social Media

- Accuracy ~90%
- But
  - Accuracy drops to ~85% when applying the model to a different city
  - Model overfitting?

- Example set:
  - “Again crash on I90”
  - “Accident on I90”
- Model:
  - “I90” → indicates traffic accident

- Applying the model:
  - “Two cars crashed on I51” → not related to traffic accident

Using LOD for Preventing Overfitting

- Example set:
  - “Again crash on I90”
  - “Accident on I90”

dbpedia:Interstate_90

rdf:type

dbpedia-owl:Road

rdf:type

dbpedia:Interstate_51

- Model:
  - dbpedia-owl:Road → indicates traffic accident

- Applying the model:
  - “Two cars crashed on I51” → indicates traffic accident

- Using DBpedia Spotlight + FeGeLOD
  - Accuracy keeps up at 90%
  - Overfitting is avoided

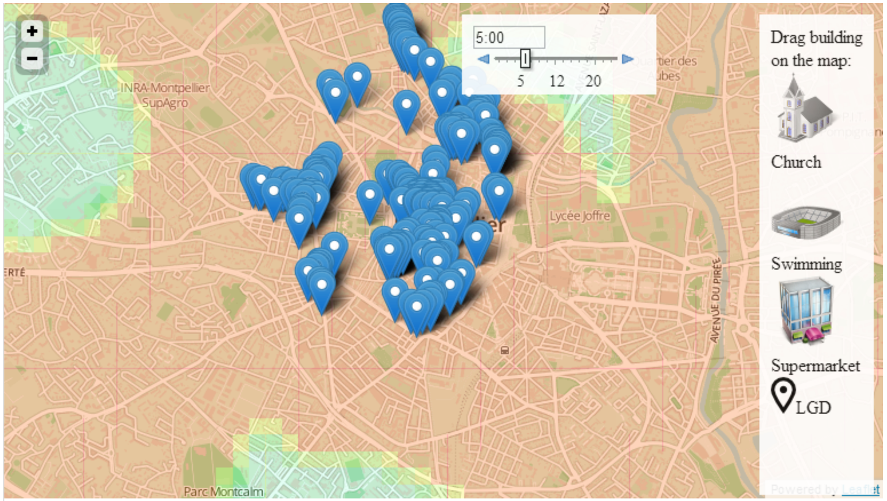
Back to Incidents

- So far, we were using preprocessed RSS data
  - How can we detect incidents automatically?
- And handcrafted rules
  - How can we acquire rules?

- Recap: Inductive vs. Deductive Reasoning:
  - Deductive: rules + facts → facts
  - Inductive: facts → rules

What's that Noise?

- A slightly different scenario:
  - Noise measurements from cities
- Create predictions
  - For the rest of the city
  - For new infrastructure
- Note:
  - No handcrafted rules
  - But a fully automatic prediction
    - Based on example measurements



Delete selected row

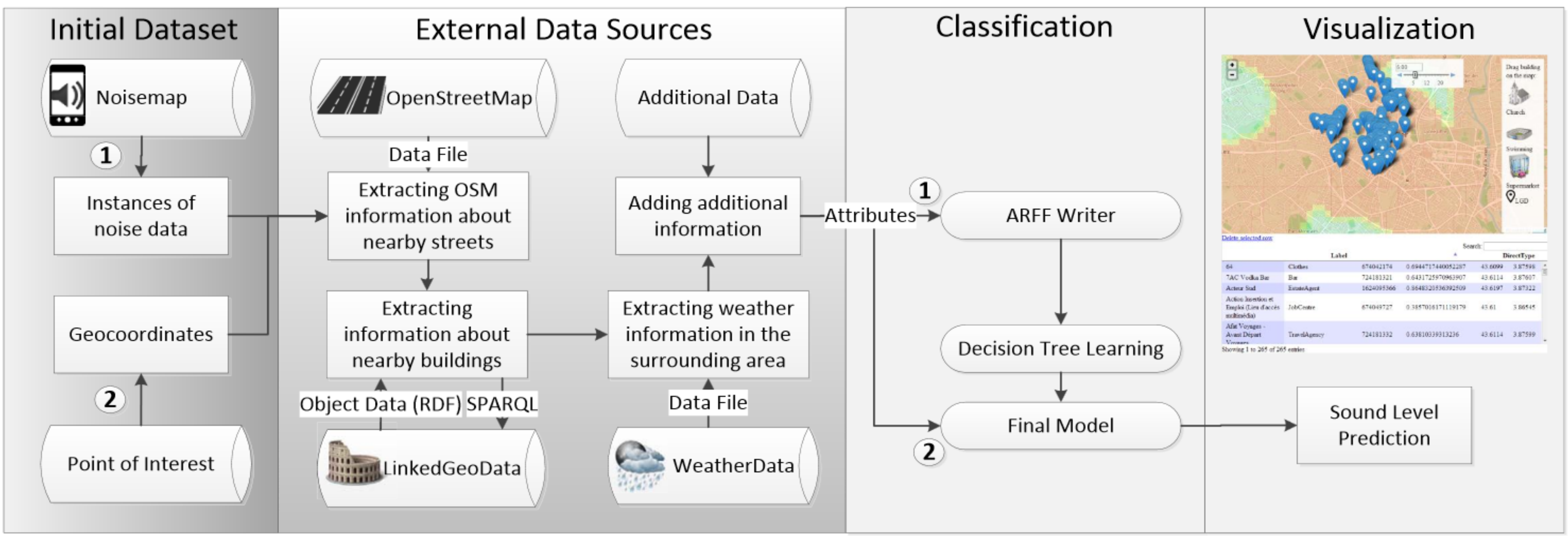
Search:

	Label					DirectType
64	Clothes	674042174	0.6944717440052287	43.6099	3.87598	
7AC Vodka Bar	Bar	724181321	0.6431725970963907	43.6114	3.87607	
Acteur Sud	EstateAgent	1624095366	0.8648320536392509	43.6197	3.87322	
Action Insertion et Emploi (Lieu d'accès multimédia)	JobCentre	674049727	0.3857016171119179	43.61	3.86545	
Afat Voyages - Avant Départ Voyages	TravelAgency	724181332	0.63810339313236	43.6114	3.87599	

Showing 1 to 265 of 265 entries

What's that Noise?

- Additional data comes from
  - Linked Geo Data
  - Open Street Maps (Streets: types, lanes and speed limits)
  - Deutscher Wetterdienst



What's that Noise?

- Results:
  - Machine trained model for noise
  - ~80% accuracy in predicting the noise level (six classes)
    - Mean absolute error only 0.077

- This allows to...
  - generate a whole noise map from a small set of observations
  - play “what if” with hypothetical changes
    - e.g., how do speed limits affect noise levels?

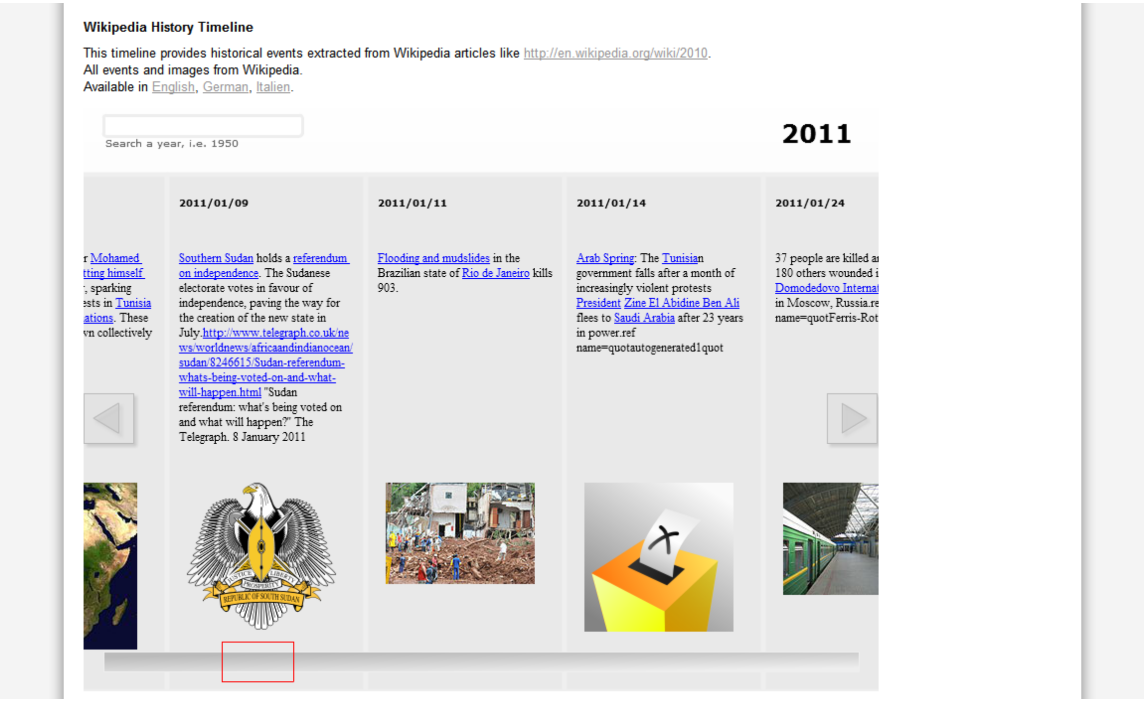
What's that Noise?

- From preprocessed RSS data
  - To automatic detection (e.g., Twitter)

- From hand-crafted rules
  - To automatically induced models

Supporting Information Extraction

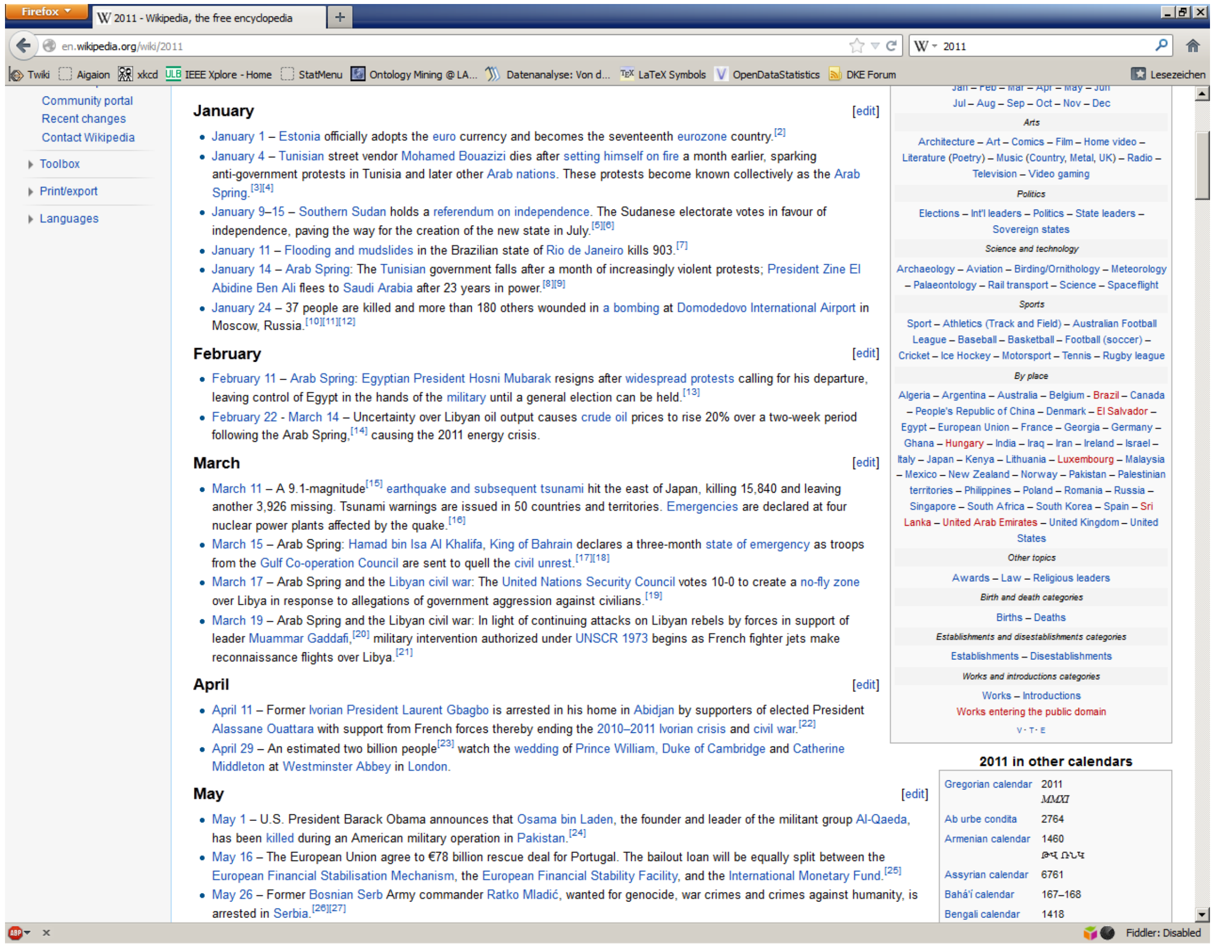
- Task: Event Extraction from Wikipedia
- Joint work with GESIS (Cologne)



http://www.vizgr.org/historical-events/timeline/

Supporting Information Extraction

- Source Material:



2011 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/2011

Community portal
Recent changes
Contact Wikipedia
Toolbox
Print/export
Languages

January [edit]

- January 1 – Estonia officially adopts the euro currency and becomes the seventeenth eurozone country.[2]
- January 4 – Tunisian street vendor Mohamed Bouazizi dies after setting himself on fire a month earlier, sparking anti-government protests in Tunisia and later other Arab nations. These protests become known collectively as the Arab Spring.[3][4]
- January 9–15 – Southern Sudan holds a referendum on independence. The Sudanese electorate votes in favour of independence, paving the way for the creation of the new state in July.[5][6]
- January 11 – Flooding and mudslides in the Brazilian state of Rio de Janeiro kills 903.[7]
- January 14 – Arab Spring: The Tunisian government falls after a month of increasingly violent protests; President Zine El Abidine Ben Ali flees to Saudi Arabia after 23 years in power.[8][9]
- January 24 – 37 people are killed and more than 180 others wounded in a bombing at Domodedovo International Airport in Moscow, Russia.[10][11][12]

February [edit]

- February 11 – Arab Spring: Egyptian President Hosni Mubarak resigns after widespread protests calling for his departure, leaving control of Egypt in the hands of the military until a general election can be held.[13]
- February 22 - March 14 – Uncertainty over Libyan oil output causes crude oil prices to rise 20% over a two-week period following the Arab Spring,[14] causing the 2011 energy crisis.

March [edit]

- March 11 – A 9.1-magnitude[15] earthquake and subsequent tsunami hit the east of Japan, killing 15,840 and leaving another 3,926 missing. Tsunami warnings are issued in 50 countries and territories. Emergencies are declared at four nuclear power plants affected by the quake.[16]
- March 15 – Arab Spring: Hamad bin Isa Al Khalifa, King of Bahrain declares a three-month state of emergency as troops from the Gulf Co-operation Council are sent to quell the civil unrest.[17][18]
- March 17 – Arab Spring and the Libyan civil war: The United Nations Security Council votes 10-0 to create a no-fly zone over Libya in response to allegations of government aggression against civilians.[19]
- March 19 – Arab Spring and the Libyan civil war: In light of continuing attacks on Libyan rebels by forces in support of leader Muammar Gaddafi,[20] military intervention authorized under UNSCR 1973 begins as French fighter jets make reconnaissance flights over Libya.[21]

April [edit]

- April 11 – Former Ivorian President Laurent Gbagbo is arrested in his home in Abidjan by supporters of elected President Alassane Ouattara with support from French forces thereby ending the 2010–2011 Ivorian crisis and civil war.[22]
- April 29 – An estimated two billion people[23] watch the wedding of Prince William, Duke of Cambridge and Catherine Middleton at Westminster Abbey in London.

May [edit]

- May 1 – U.S. President Barack Obama announces that Osama bin Laden, the founder and leader of the militant group Al-Qaeda, has been killed during an American military operation in Pakistan.[24]
- May 16 – The European Union agree to €78 billion rescue deal for Portugal. The bailout loan will be equally split between the European Financial Stabilisation Mechanism, the European Financial Stability Facility, and the International Monetary Fund.[25]
- May 26 – Former Bosnian Serb Army commander Ratko Mladić, wanted for genocide, war crimes and crimes against humanity, is arrested in Serbia.[26][27]

Jul – Aug – Sep – Oct – Nov – Dec
Arts
Architecture – Art – Comics – Film – Home video – Literature (Poetry) – Music (Country, Metal, UK) – Radio – Television – Video gaming
Politics
Elections – Int'l leaders – Politics – State leaders – Sovereign states
Science and technology
Archaeology – Aviation – Birding/Ornithology – Meteorology – Palaeontology – Rail transport – Science – Spaceflight
Sports
Sport – Athletics (Track and Field) – Australian Football League – Baseball – Basketball – Football (soccer) – Cricket – Ice Hockey – Motorsport – Tennis – Rugby league
By place
Algeria – Argentina – Australia – Belgium – Brazil – Canada – People's Republic of China – Denmark – El Salvador – Egypt – European Union – France – Georgia – Germany – Ghana – Hungary – India – Iraq – Iran – Ireland – Israel – Italy – Japan – Kenya – Lithuania – Luxembourg – Malaysia – Mexico – New Zealand – Norway – Pakistan – Palestinian territories – Philippines – Poland – Romania – Russia – Singapore – South Africa – South Korea – Spain – Sri Lanka – United Arab Emirates – United Kingdom – United States
Other topics
Awards – Law – Religious leaders
Birth and death categories
Births – Deaths
Establishments and disestablishments categories
Establishments – Disestablishments
Works and introductions categories
Works – Introductions
Works entering the public domain

2011 in other calendars

Gregorian calendar	2011 MMXI
Ab urbe condita	2764
Armenian calendar	1460
Assyrian calendar	6761
Bahá'í calendar	167–168
Bengali calendar	1418

Supporting Information Extraction

- Event data is automatically extracted
  - Date
  - Textual Description
  - Links to other entities (place, involved people, ...)
- Classification of events required
  - Politics, Culture, Sports, …
  - e.g., for better querying, filtering, …
- Approach: use Machine Learning for classification!

Supporting Information Extraction

- Positive Examples for class *politics*:
  - *2011, March 15 - German chancellor Angela Merkel shuts down the seven oldest German nuclear power plants.*
  - *2010, June 3 – Christian Wulff is nominated for President of Germany by Angela Merkel.*
- Negative Examples for class *politics*:
  - *2010, July 7 – Spain defeats Germany 1-0 to win its semi-final and for its first time, along with Netherlands make the 2010 FIFA World Cup Final.*
  - *2012, February 16 – Roman Lob is selected to represent Germany in the Eurovision Song Contest.*

Supporting Information Extraction

- Positive Examples for class *politics*:
  - *2011, March 15 - German chancellor Angela Merkel shuts down the seven oldest German nuclear power plants.*
  - *2010, June 3 – Christian Wulff is nominated for President of Germany by Angela Merkel.*
- Negative Examples for class *politics*:
  - *2010, July 7 – Spain defeats Germany 1-0 to win its semi-final and for its first time, along with Netherlands make the 2010 FIFA World Cup Final.*
  - *2012, February 16 – Roman Lob is selected to represent Germany in the Eurovision Song Contest.*
- Possible learned model:
  - "Angela Merkel" → Politics

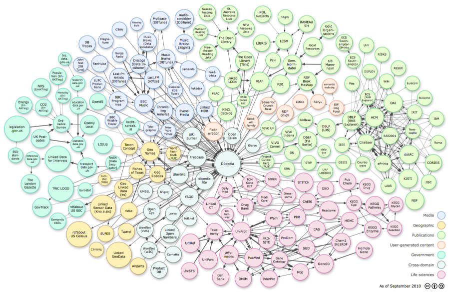
Supporting Information Extraction

- Possibly Learned Model:
  - "Angela Merkel" → Politics

- There are some problems with that model
- Missing generality (again: overfitting!)
  - *2012, May 13, Elections in North Rhine-Westphalia – Hannelore Kraft is elected to continue as Minister-President, heading an SPD-Green coalition.*
- Large amount of training examples required
  - At least one positive and one negative example per politician
  - but training examples are expensive...

Supporting Information Extraction

- Possibly Learned Model:
  - "Angela Merkel" → Politics



- How can we do better?
- Background knowledge from Linked Open Data
  - *2011, March 15 - German chancellor Angela Merkel [class: Politician] shuts down the seven oldest German nuclear power plants.*
  - *2012, May 13, Elections in North Rhine-Westphalia – Hannelore Kraft [class: Politician] is elected to continue as Minister-President, heading an SPD-Green coalition.*

- Model learned in that case:
  - "[class: Politician]" → Politics

Supporting Information Extraction

- Model learned in that case:
  - "[class: Politician]" → Politics

- Much more general
  - Can also classify events with politicians not contained in the training set
- Less training examples required
  - A few events with politicians, athletes, singers, ... are enough

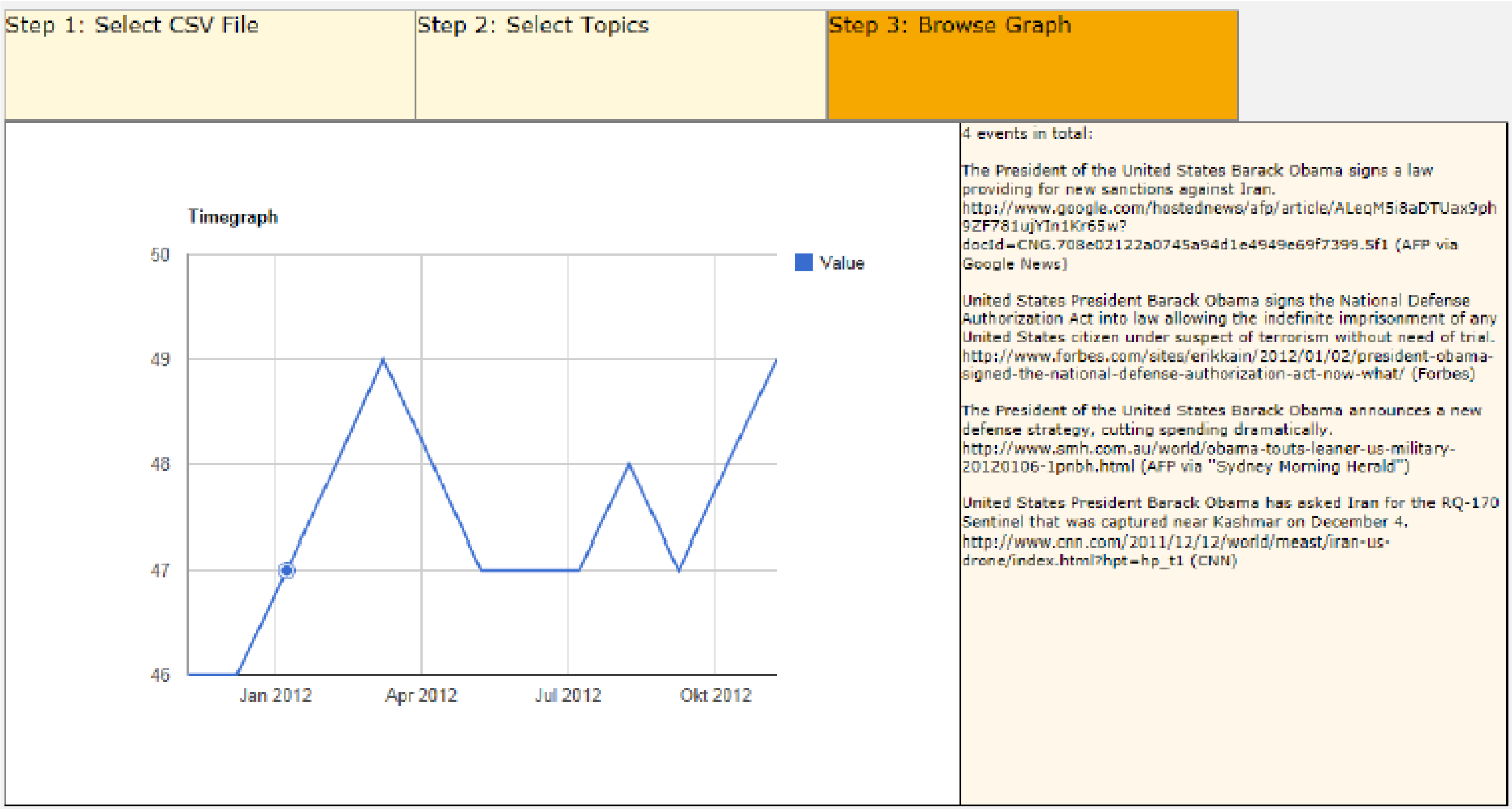
Supporting Information Extraction

- Experiments on Wikipedia data
  - >10 categories
  - 1,000 labeled examples as training set
  - Classification accuracy: 80%

- Plus:
  - We have trained a language-independent model!
    - often, models are like "elect*" → Politics
  - *22. Mai 2012: Peter Altmaier [class: Politician] wird als Nachfolger von Norbert Röttgen [class: Politician] zum Bundesumweltminister ernannt.*
  - *6 januari 2012: Jonas Sjöstedt [class: Politician] väljs till ny partiledare för Vänsterpartiet efter Lars Ohly [class: Politician].*

Using the Events

- E.g., for annotating time series graphs



Using the Events

- Annotating a Time Series
  - e.g., the stock market price of Apple_Inc.
- Starting point: a link to DBpedia
  - e.g., dbpedia:Apple_Inc.

- Simple approach: retrieve all the events for that entity
  - Problem: low recall

- Naive “improvement”:
  - Also include events for entities linked to dbpedia:Apple_Inc. in DBpedia
  - e.g.: dbpedia:IPhone
  - Problem: extremely low precision
  - Due to frequent entities such as dbpedia:United_States

Relatedness in DBpedia

- Required: entities that are *closely* related to dbpedia:Apple_Inc.

- Problem:
  - There is no notion of proximity, predicate weights, etc. in DBpedia
  - And in LOD in general

- Possible solution: let humans rank proximity
  - Subjective
  - Scales badly

- Required:
  - Automatic approximation

Relatedness in DBpedia

- Good approximation: Normalized Google Distance
  - How frequently do two terms co-occur in websites?
  - E.g., “Apple” and “iPhone” co-occur quite frequently
  - “Apple” and “United States” co-occur less frequently

- Problem:
  - Search engines are not for free (for machine requests)
  - Pairwise ranking of all DBpedia entities would cost much money

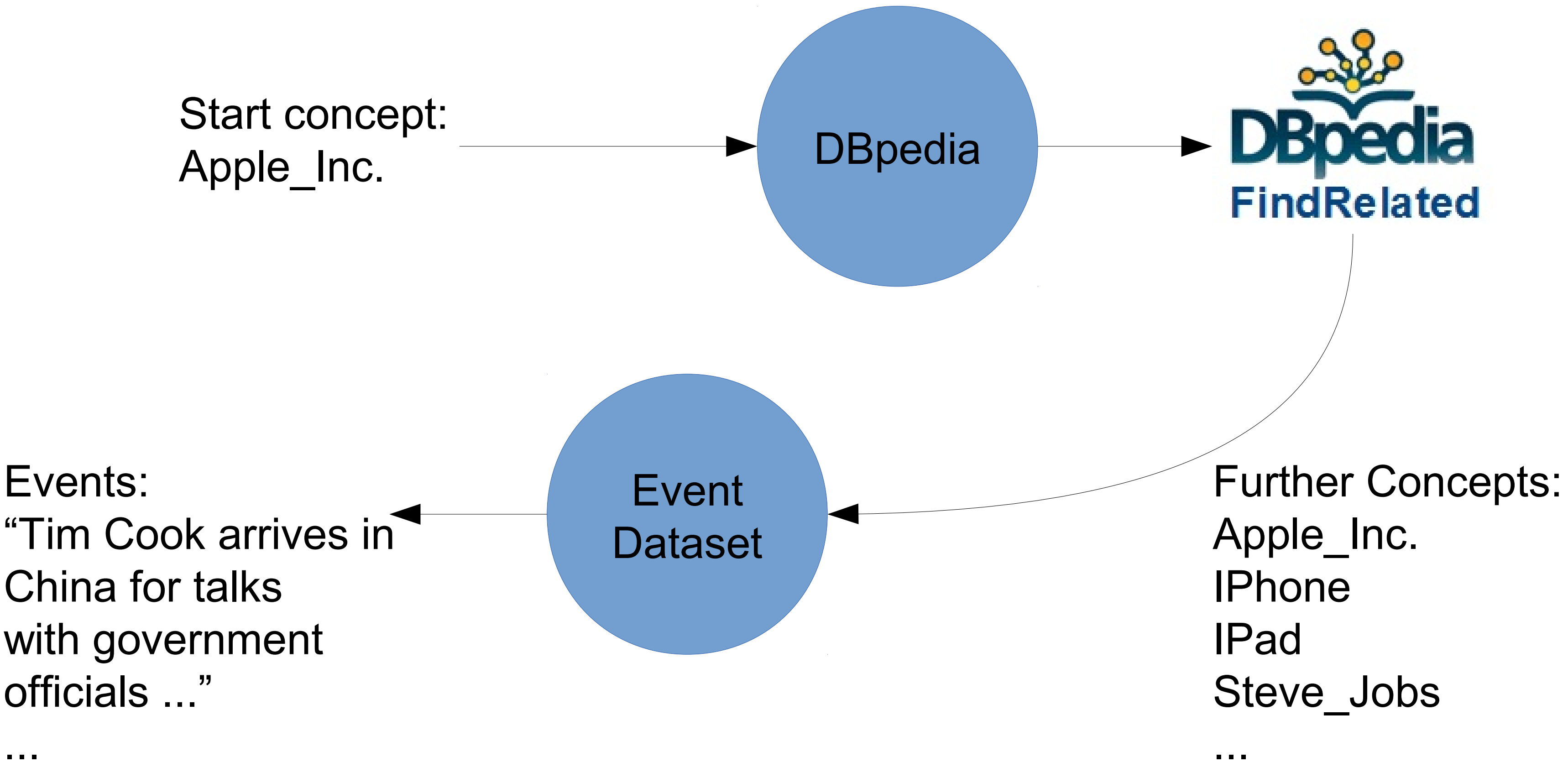
- Solution:
  - Retrieve search engine based rankings for a small sample
  - Approximate rankings by machine learning model

Relatedness in DBpedia

- DBpedia FindRelated Service
  - Trained on 10,000 statements labeled with NGD
  - >50 Features: network based, linguistic, dataset specific
  - Fair correlation with NGD
  - Live: http://wifo5-21.informatik.uni-mannheim.de:8080/DBpediaFindRelated/

- Use in Time Series Application:
  - Increases recall up to 25%
  - Fair tradeoff with precision (<10%)
  - Hundreds of entities added to the search!

Time Series Application – The Big Picture



Intermediate Recap

- What we have seen so far:
  - Linked Open Data as background knowledge in various tasks
  - Combination with Machine Learning for intelligent applications
  - Additional dimensions on the data (proximity measures)

UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime ✓

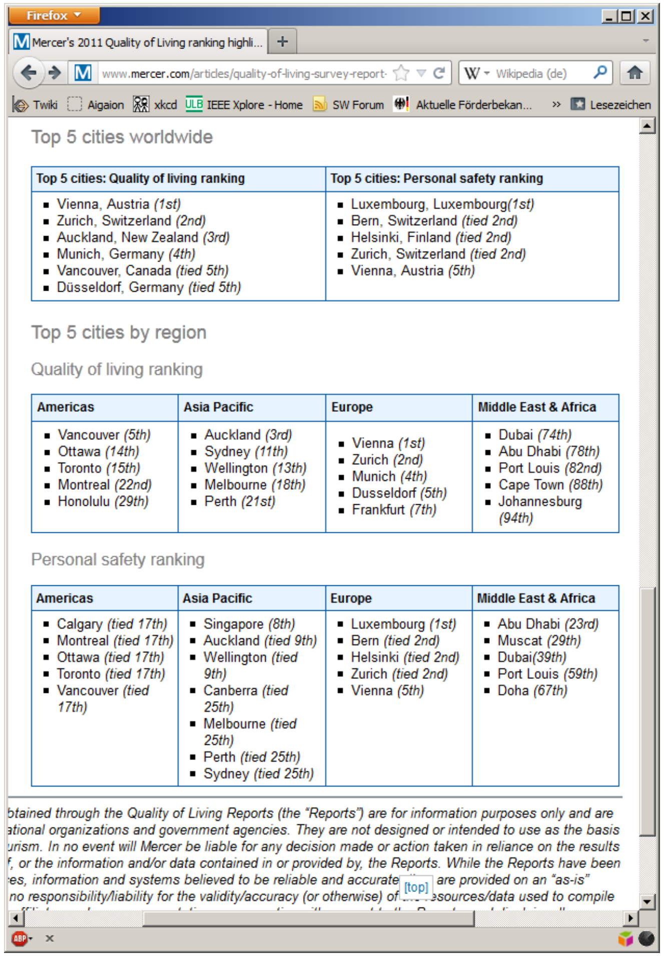
Chris Bizer, Heiko Paulheim, University of Mannheim

And now for Something Completely Different

- Who are these men?

Statistical Data

- Statistics are very wide spread
  - Quality of living in cities
  - Corruption by country
  - Fertility rate by country
  - Suicide rate by country
  - Box office revenue of films
  - ...

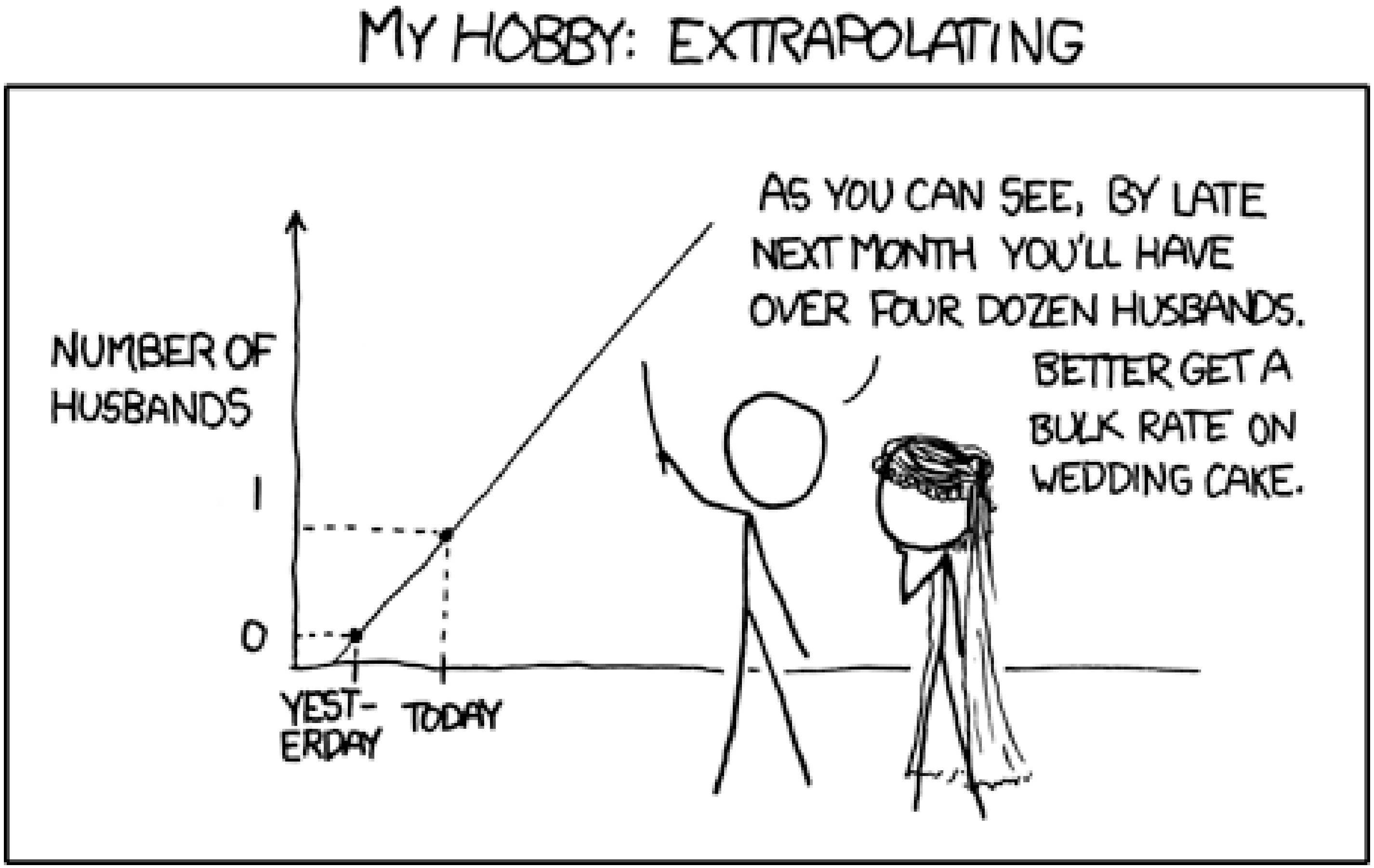


Statistical Data

- Questions we are often interested in
  - **Why** does city X have a high/low quality of living?
  - **Why** is the corruption higher in country A than in country B?
  - Will a **new film** create a high/low box office revenue?

- i.e., we are looking for
  - explanations
  - forecasts (e.g., extrapolations)

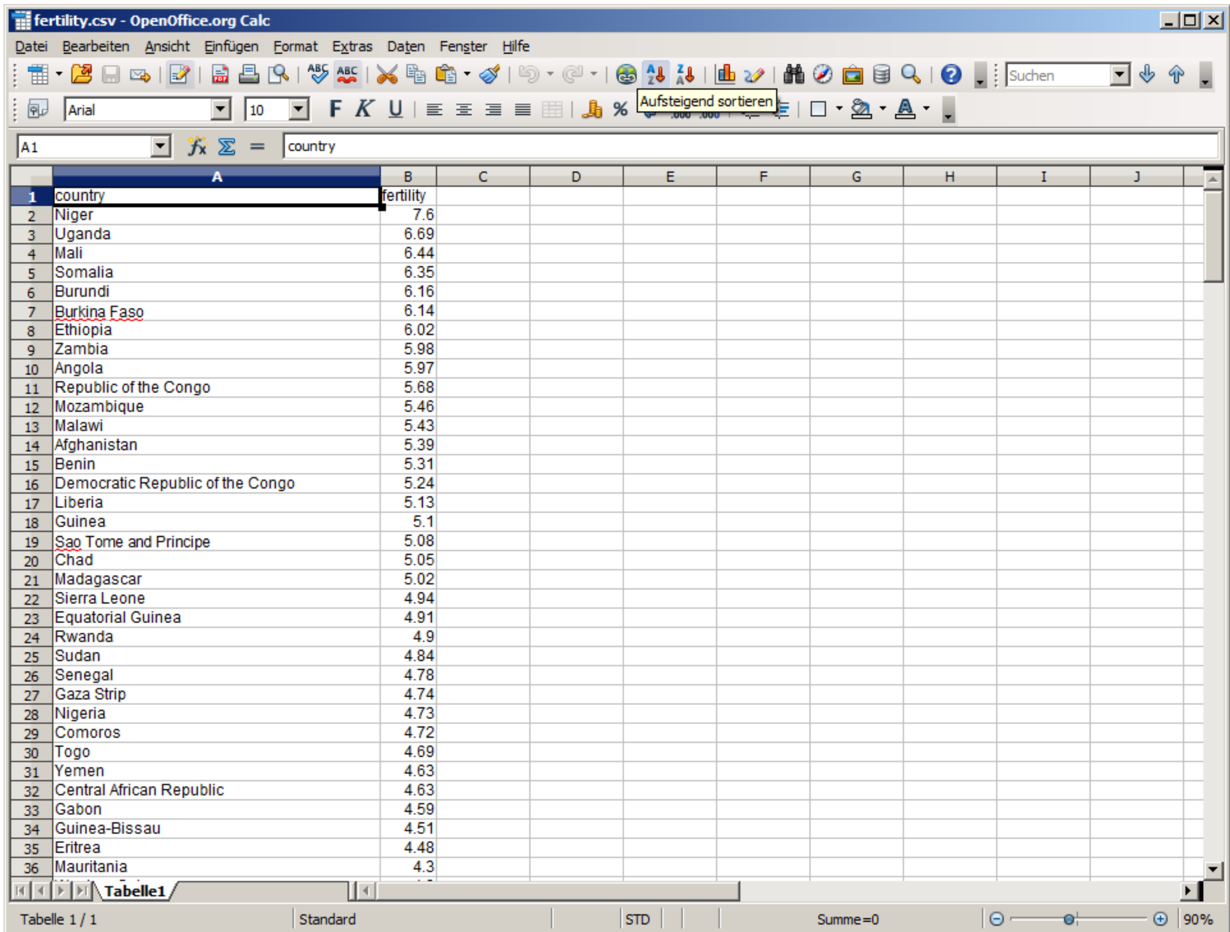
Statistical Data



http://xkcd.com/605/

Statistical Data

- What statistics typically look like



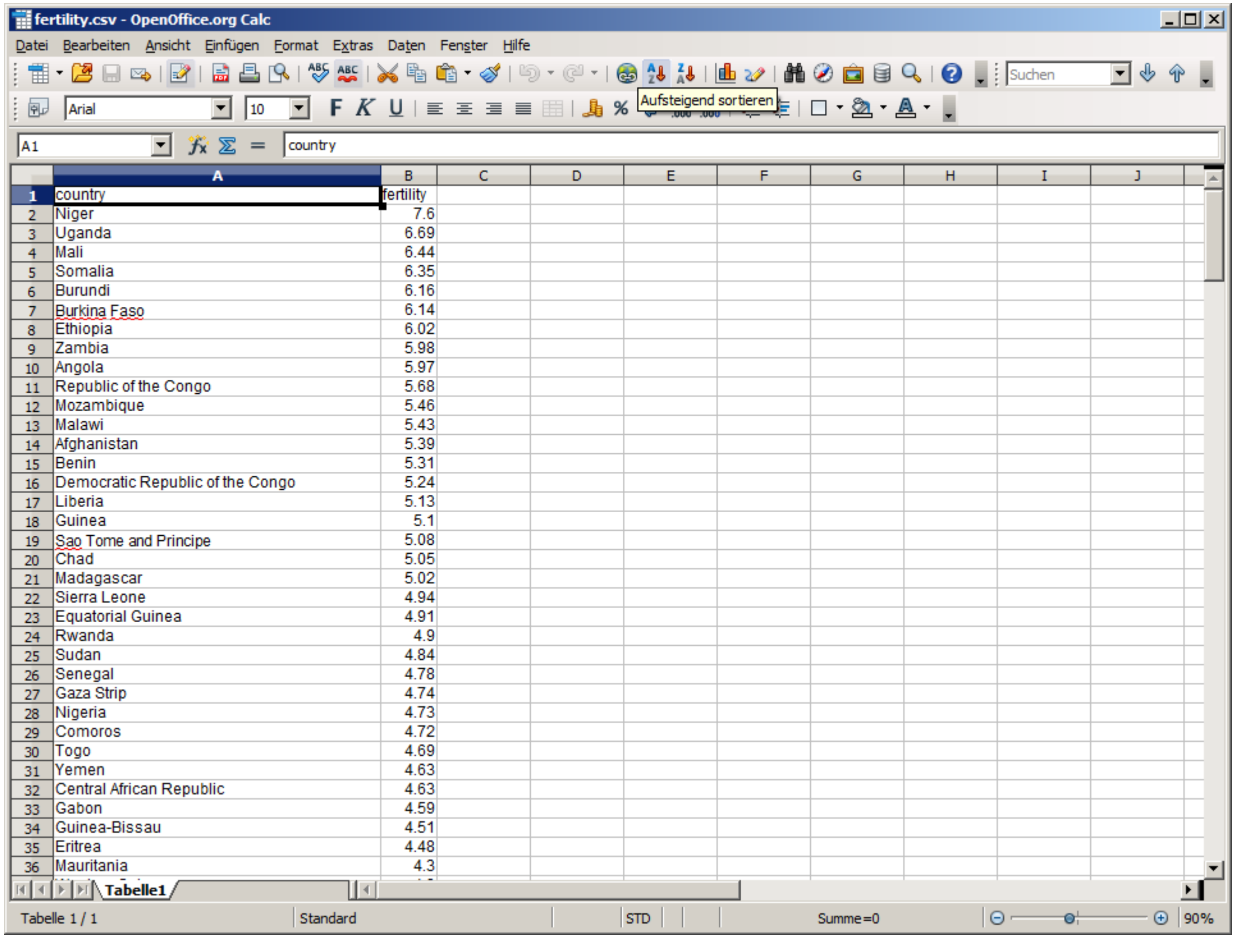
	A	B
1	country	fertility
2	Niger	7.6
3	Uganda	6.69
4	Mali	6.44
5	Somalia	6.35
6	Burundi	6.16
7	Burkina Faso	6.14
8	Ethiopia	6.02
9	Zambia	5.98
10	Angola	5.97
11	Republic of the Congo	5.68
12	Mozambique	5.46
13	Malawi	5.43
14	Afghanistan	5.39
15	Benin	5.31
16	Democratic Republic of the Congo	5.24
17	Liberia	5.13
18	Guinea	5.1
19	Sao Tome and Principe	5.08
20	Chad	5.05
21	Madagascar	5.02
22	Sierra Leone	4.94
23	Equatorial Guinea	4.91
24	Rwanda	4.9
25	Sudan	4.84
26	Senegal	4.78
27	Gaza Strip	4.74
28	Nigeria	4.73
29	Comoros	4.72
30	Togo	4.69
31	Yemen	4.63
32	Central African Republic	4.63
33	Gabon	4.59
34	Guinea-Bissau	4.51
35	Eritrea	4.48
36	Mauritania	4.3

Statistical Data

- There are powerful tools for finding correlations etc.
  - but many statistics cannot be interpreted directly
  - background knowledge is missing
- So where do we get background knowledge from?
  - with as little efforts as possible

Statistical Data

- What we have



	A	B
1	country	fertility
2	Niger	7.6
3	Uganda	6.69
4	Mali	6.44
5	Somalia	6.35
6	Burundi	6.16
7	Burkina Faso	6.14
8	Ethiopia	6.02
9	Zambia	5.98
10	Angola	5.97
11	Republic of the Congo	5.68
12	Mozambique	5.46
13	Malawi	5.43
14	Afghanistan	5.39
15	Benin	5.31
16	Democratic Republic of the Congo	5.24
17	Liberia	5.13
18	Guinea	5.1
19	Sao Tome and Principe	5.08
20	Chad	5.05
21	Madagascar	5.02
22	Sierra Leone	4.94
23	Equatorial Guinea	4.91
24	Rwanda	4.9
25	Sudan	4.84
26	Senegal	4.78
27	Gaza Strip	4.74
28	Nigeria	4.73
29	Comoros	4.72
30	Togo	4.69
31	Yemen	4.63
32	Central African Republic	4.63
33	Gabon	4.59
34	Guinea-Bissau	4.51
35	Eritrea	4.48
36	Mauritania	4.3

Statistical Data

- What we need

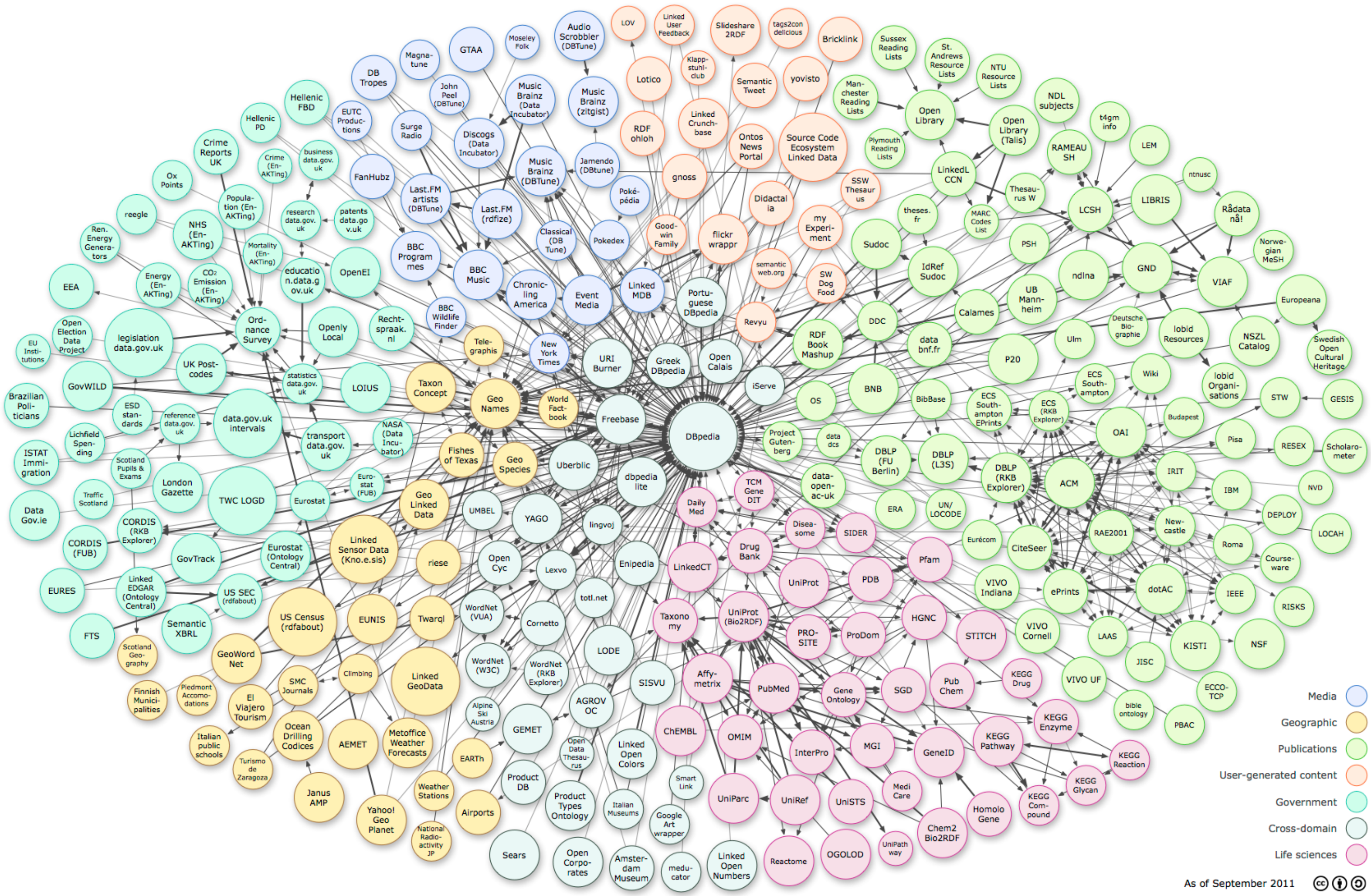
fertility-data.csv - OpenOffice.org Calc

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	country	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	country_uri_c	cour
2	Niger	1.28	32.389999	137.1	2.360E+011	15.39	236040	137.1	1998	514	4.219E+010	80	
3	Uganda	1.28	32.389999	137.1	2.360E+011	15.39	236040	137.1	1998	514	4.219E+010	80	
4	Mali	12.65	-8	11.698895	1.240E+012	1.6	1240187.32	11.7	1994	691	1.677E+010	215	
5	Somalia	2.033333	45.349998	13.899678	6.377E+011	?	637657	13.899678	?	?	5731000000	209	
6	Burundi	-3.5	30	322.974456	2.783E+010	7.8	27834	323	1998	180	3397000000	45	
7	Burkina Faso	12.333333	-1.666667	57.4	2.742E+011	0.146	274199.451	57.4	2007	597	1.999E+010	145	
8	Ethiopia	9.03	38.740002	74	1.104E+012	0.7	1104295.82	74.903819	1999	350	8.612E+010	123	
9	Zambia	-15.416667	28.283333	17.181546	7.526E+011	1	752616.875	17.181546	2002	1086	1.845E+010	?	
10	Angola	-8.833333	13.333333	14.8	1.247E+012	?	1246701.14	14.826323	2000	4477	1.073E+011	199	
11	Republic of t	-4.266667	15.283334	10.77225	3.420E+011	3.3	342000.16	10.8	?	2983	1.711E+010	204	
12	Mozambique	-25.950001	32.583332	28.68739	8.016E+011	2.2	801590	28.7	1996	458	2.181E+010	178	
13	Malawi	-13.95	33.700001	128.8	1.185E+011	20.6	118484	128.8	2008	322	1.298E+010	94	
14	Afghanistan	34.049999	69.133331	43.166221	6.475E+011	?	647500	43.166221	?	517	2.736E+010	150	
15	Benin	6.466667	2.6	78.069856	1.126E+011	0.02	112622	78.069856	2003	689	1.399E+010	120	
16	Democratic	-4.316667	15.316667	29.3	2.345E+012	4.3	2345409	29.3	?	186	2.312E+010	182	
17	Liberia	6.316667	-10.8	35.5	1.114E+011	13.514	111369.489	35.521399	?	226	1691000000	180	
18	Guinea	7.538055	-13.7	40.9	2.459E+011	?	245857	40.9	1994	448	1.081E+010	?	
19	Sao Tome ar	0.116667	6.566667	169.1	963475577	0	963.475577	169.1	?	1183	311000000	69	
20	Chad	12.1	16.033333	8	1.284E+012	1.9	1283994.38	8.030925	?	767	1.736E+010	212	
21	Madagascar	-18.916666	47.516666	35.173907	5.869E+011	0.13	586883.536	35.173907	2001	320	1.941E+010	174	
22	Sierra Leone	8.484445	-13.234445	79.382604	7.174E+010	1.1	71740	79.382604	2003	311	4585000000	114	
23	Equatorial G	1.5	8.783334	24.092775	2.805E+010	?	28049.5712	24.092775	?	15401	2.152E+010	187	
24	Rwanda	-1.943883	30.05945	419.770267	2.634E+010	5.3	26338	419.770267	2003	593	1.311E+010	29	
25	Sudan	15.633056	32.533054	16.370732	1.886E+012	?	1886068	16.370732	?	?	?	?	
26	Senegal	14.666667	-17.416666	69.652829	1.967E+011	2.1	196723	69.652829	1995	1026	2.327E+010	134	
27	Gaza Strip	31.416666	34.333332	4117.77952	360000000	?	360	4117.77952	?	?	770000000	6	
28	Nigeria	8	7.483333	164.788401	9.238E+011	1.4	923768	164.8	2003	1389	3.779E+011	?	
29	Comoros	?	?	?	?	?	?	?	?	?	?	?	?
30	Togo	6.116667	1.216667	116.564242	5.679E+010	4.2	56785.4893	116.564242	?	422	5612000000	93	
31	Yemen	15	44.200001	44.67202	5.280E+011	?	527966.486	44.67202	?	1061	5.822E+010	160	
32	Central Afric	4.366667	18.583334	7.1	6.230E+011	0	622984	7.10428	1993	436	3446000000	223	
33	Gabon	0.383333	9.45	5.5	2.677E+011	3.76	267667.501	5.521261	?	8724	2.248E+010	216	
34	Guinea-Biss	11.866667	-15.6	44.1	3.613E+010	22.4	36125.1542	44.594799	1993	508	1784000000	154	
35	Eritrea	15.333333	38.916668	43.1	1.176E+011	0.14	117598.41	43.1	?	397	3625000000	165	
36	Mauritania	18.15	-15.966666	3.166038	1.031E+012	0.03	1030700	3.2	2000	1195	6655000000	?	

Tabelle1 Tabelle2 Tabelle3

Tabelle 1 / 3 Standard STD Summe=0 90%

Possible Sources for Background Knowledge



...and we've already seen FeGeLOD

ISBN	City	# sold
3-2347-3427-1	Darmstadt	124

Named Entity Recognition

ISBN	City	City_URI	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	124

Feature Generation

ISBN	City	City_URI	City_URI_dbpedia-owl:populationTotal	City_URI_...	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	141471	...	124

Feature Selection

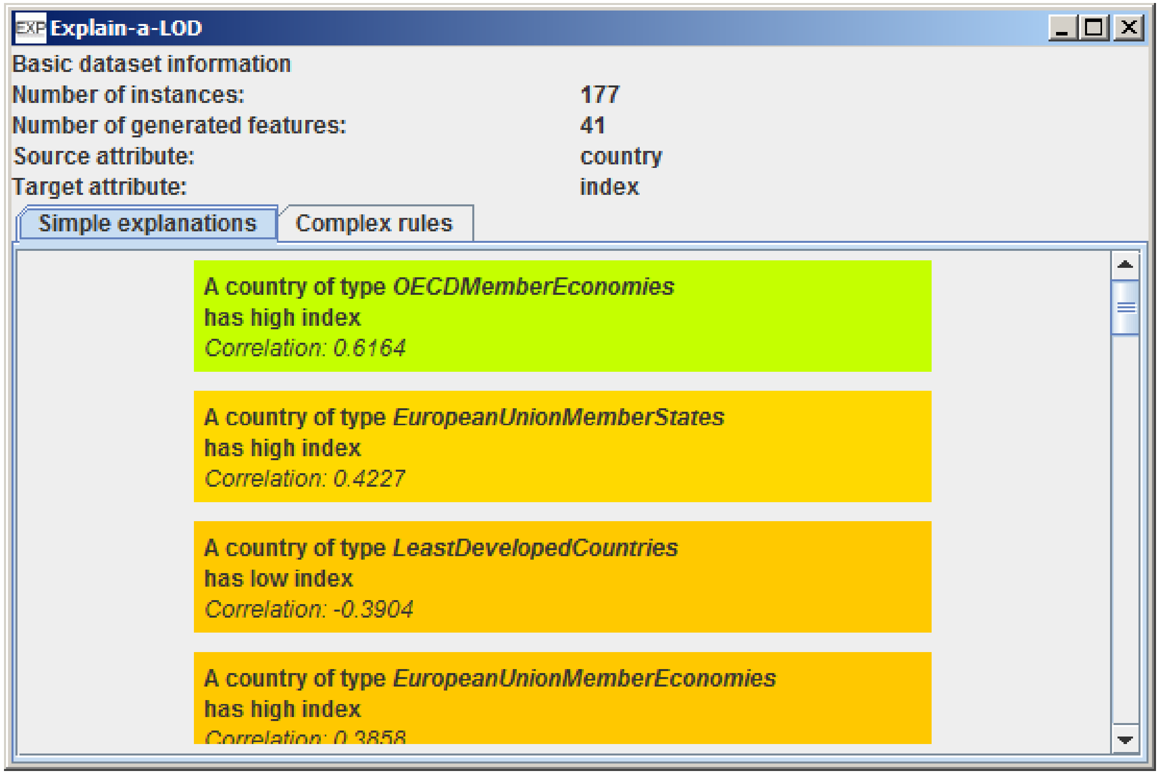
ISBN	City	City_URI	City_URI_dbpedia-owl:populationTotal	# sold
3-2347-3427-1	Darmstadt	http://dbpedia.org/resource/Darmstadt	141471	124

Statistical Data

- Adding background knowledge
  - FeGeLOD framework
- Correlation analysis
  - e.g., Pearson Correlation Coefficient
- Rule learning
  - e.g., Association Rule Mining
  - e.g., Subgroup Discovery
- Further data preprocessing
  - depending on approach
  - e.g., discretization

Prototype Tool: Explain-a-LOD

- Loads a statistics file (e.g., CSV)
- Adds background knowledge
- Presents explanations



Presenting Explanations

- Verbalization with simple patterns
  - e.g., negative correlation between *population* and *quality of living*
  - "A *city* which has a low *population* has a high *quality of living*"
- Color coding
  - By correlation coefficient, confidence/support of rules, etc.

Statistical Data: Examples

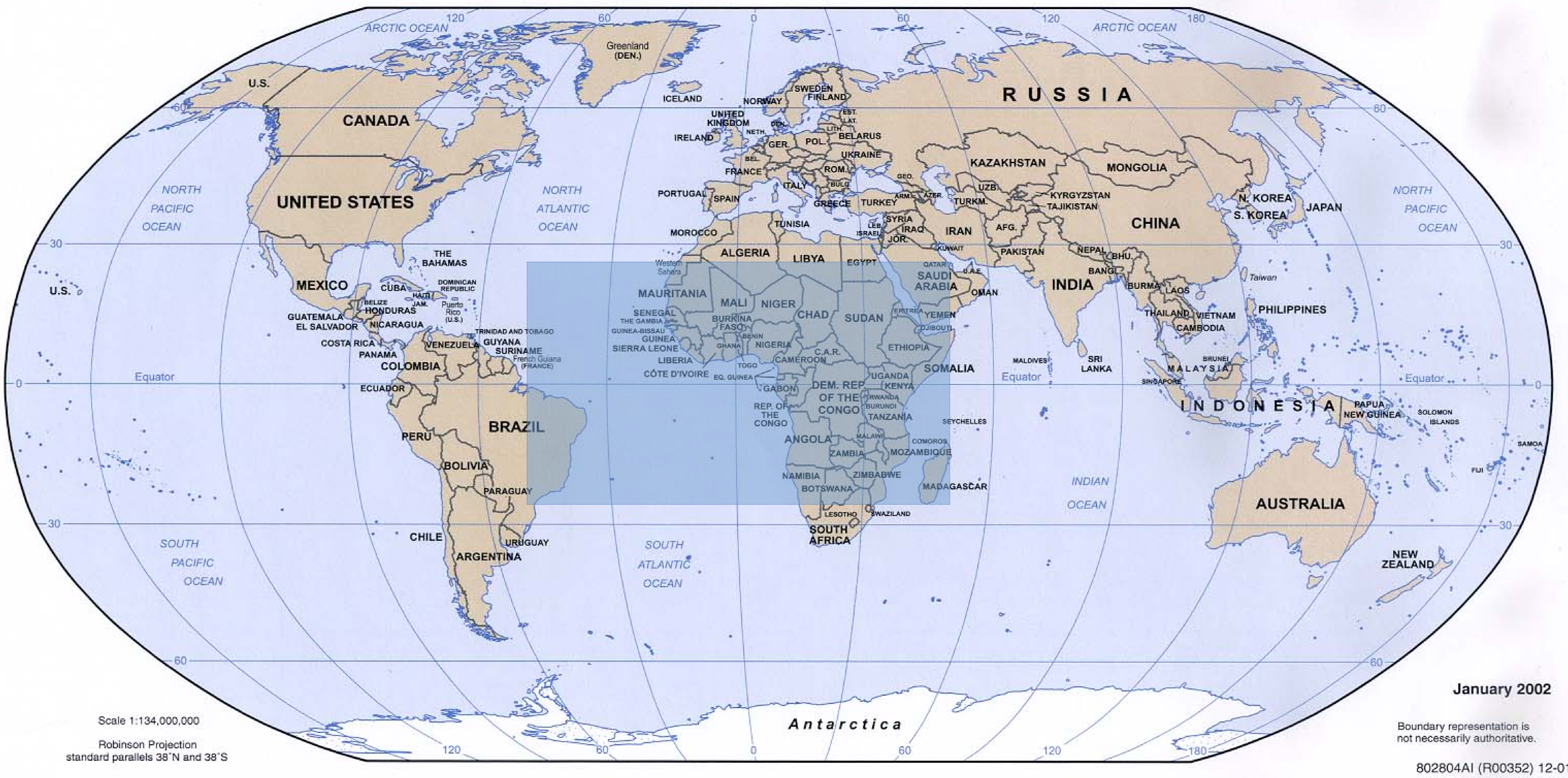
- Data Set: Mercer Quality of Living
  - Quality of living in 216 cities word wide
  - norm: NYC=100 (value range 23-109)
  - As of 1999
  - http://across.co.nz/qualityofliving.htm

- LOD data sets used in the examples:
  - DBpedia
  - CIA World Factbook for statistics by country

Statistical Data: Examples

- Examples for low quality cities
  - big hot cities (junHighC >= 27 and areaTotalKm >= 334)
  - cold cities where no music has ever been recorded (recordedIn_in = false and janHighC <= 16)
  - latitude <= 24 and longitude <= 47
    - a very accurate rule
    - but what's the interpretation?

Statistical Data: Examples



Statistical Data: Examples

- Data Set: Transparency International
  - 177 Countries and a corruption perception indicator (between 1 and 10)
  - As of 2010
  - http://www.transparency.org/cpi2010/results

Statistical Data: Examples

- Example rules for countries with low corruption
  - HDI > 78%
    - Human Development Index, calculated from live expectancy, education level, economic performance
  - OECD member states
  - Foundation place of more than nine organizations
  - More than ten mountains
  - More than ten companies with their headquarter in that state, but less than two cargo airlines

Statistical Data: Examples

- Data Set: Burnout rates
  - 16 German DAX companies
  - Absolute and relative numbers
  - As of 2011
  - http://de.statista.com/statistik/daten/studie/226959/umfrage/burn-out-erkrankungen-unter-mitarbeitern-ausgewaehlter-dax-unternehmen/

Statistical Data: Examples

- Findings for burnout rates
  - Positive correlation between turnover and burnout rates
  - Car manufacturers are less prone to burnout
  - German companies are less prone to burnout than international ones
    - Exception: Frankfurt

Statistical Data: Examples

- Data Set: Antidepressives consumption
  - In European countries
  - Source: OECD
  - http://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2011/pharmaceutical-consumption_health_glance-2011-39-en

Statistical Data: Examples

- Findings for antidepressives consumption
  - Larger countries have higher consumption
  - Low HDI → high consumption
  - By geography:
    - Nordic countries, countries at the Atlantic: high
    - Mediterranean: medium
    - Alpine countries: low
  - High average age → high consupmtion
  - High birth rates → high consumption

UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains
sex, drugs ✓, and crime ✓

Chris Bizer, Heiko Paulheim, University of Mannheim

Statistical Data: Examples

- Data Set: Suicide rates
  - By country
  - OECD states
  - As of 2005
  - http://www.washingtonpost.com/wp-srv/world/suiciderate.html

Statistical Data: Examples

- Findings for suicide rates
  - Democraties 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

Statistical Data: Examples

- Data set: sexual activity
  - Percentage of people having sex weekly
  - By country
  - Survey by Durex 2005-2009
  - http://chartsbin.com/view/uya

Statistical Data: Examples

- Findings on sexual activity
  - 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

UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim

Try it... but be careful!

- Download from

  http://www.ke.tu-darmstadt.de/resources/explain-a-lod

- including a demo video, papers, etc.

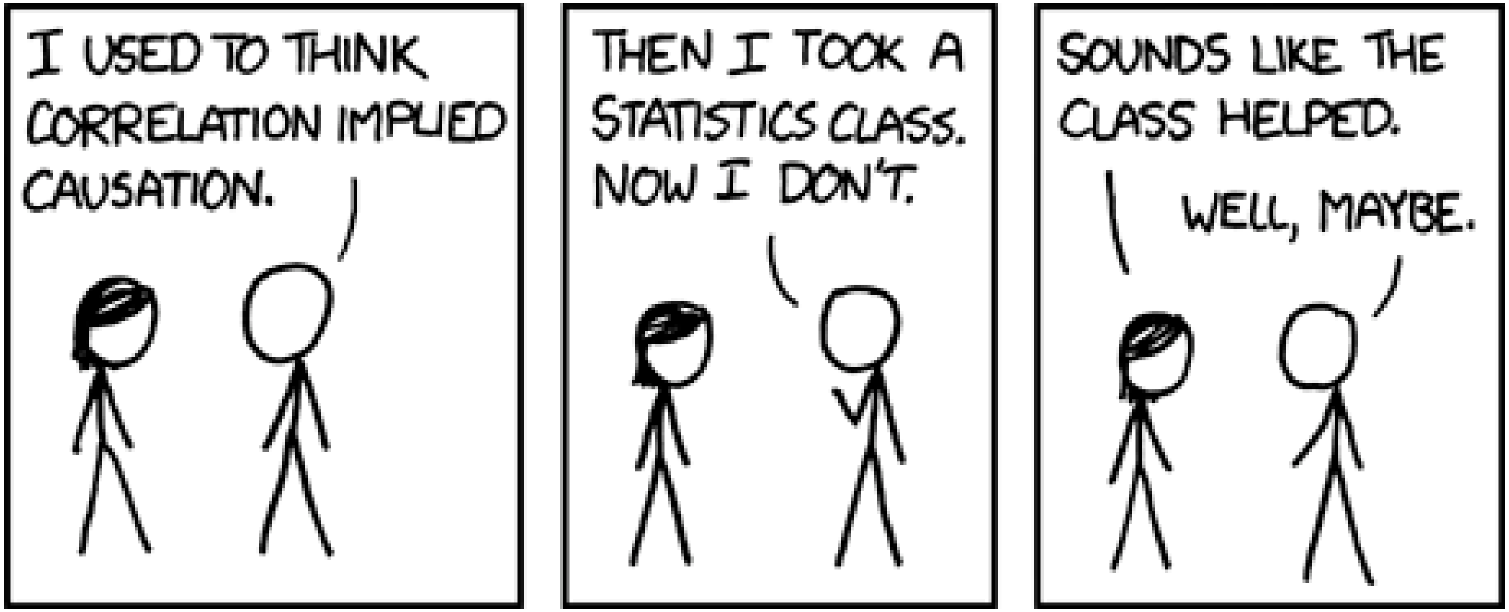
- Pitfalls
  - Open world assumption
  - LOD may be noisy
  - Biases
  - ...

Try it... but be careful!

- Download from

  http://www.ke.tu-darmstadt.de/resources/explain-a-lod

- including a demo video, papers, etc.



http://xkcd.com/552/

Conclusions

- Many tasks require massive background knowledge
  - Can be acquired from LOD
  - E.g., FeGeLOD framework
- Machine learning is often useful to...
  - make sense using Linked Open Data
  - answer non-trivial questions
  - Add additional knowledge dimensions (e.g., similarity)

Credits

- Daniel Hienert, GESIS (WikiEvents)
- Simon Holthausen, TU Darmstadt (Time Series Analysis)
- Frederik Janssen, TU Darmstadt (NoiseMap)
- Jacob Karolus, TU Darmstadt (NoiseMap)
- Evgeny Mitichkin, Uni Mannheim (FeGeLOD)
- Petar Ristoski, Uni Mannheim (Incident Detection from Twitter)
- Axel Schulz, SAP Research/TU Darmstadt
  (MICI, NoiseMap, Incident Detection from Twitter)
- Dennis Wegener, GESIS (WikiEvents)

UNIVERSITÄT MANNHEIM

Building Knowledge-Intensive Applications with Linked Open Data*

*using examples from the domains sex, drugs, and crime

Chris Bizer, Heiko Paulheim, University of Mannheim