



WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS



Open Data as the fuel for complexity science?

Axel Polleres,

joint work with Stefan Bischof, Sebastian Neumaier, Jürgen Umbrich, etc.

Geoffrey West (former director of the Santa Fe Institute) 2011

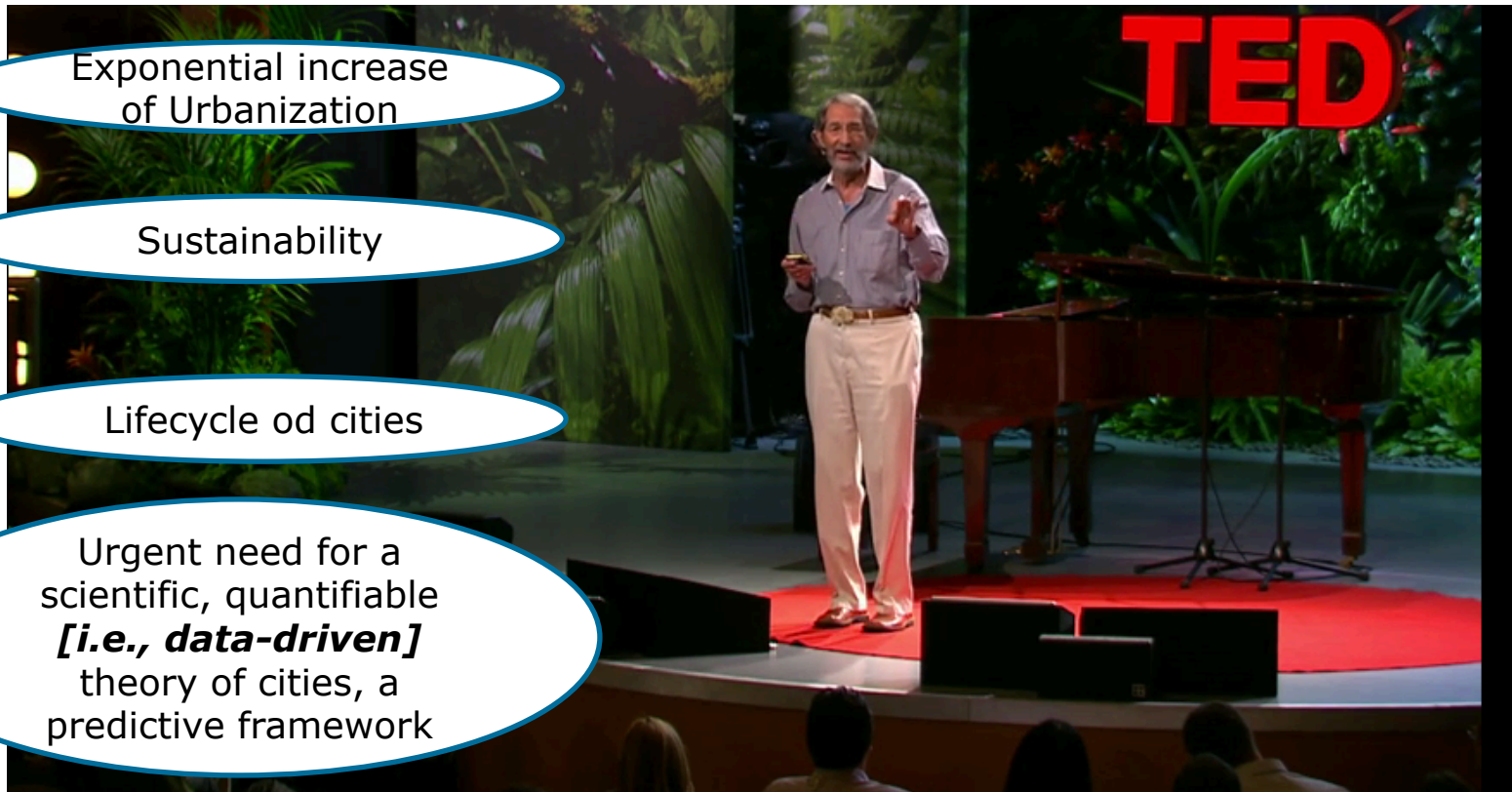
... my first impression of Complexity Science, if you want...

Exponential increase
of Urbanization

Sustainability

Lifecycle of cities

Urgent need for a
scientific, quantifiable
[i.e., data-driven]
theory of cities, a
predictive framework



Back at that time... City Data – Important for Infrastructure Providers & for City Decision Makers

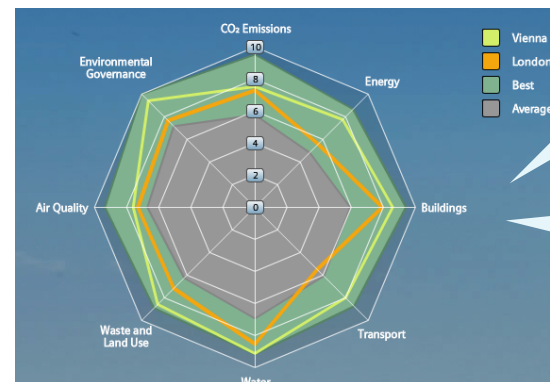
- City Assessment and Sustainability reports
- Tailored offerings by Infrastructure Providers



... however, these are often **outdated** before even published!

→Needs **up-to-date City Data** and **calculates City KPIs** in a way that allows to display the current state and run scenarios of different product applications.

e.g. towards a “Dynamic” Green City Index:



Goal (short term):

- Leverage Open Data for calculating a city' performance from public sources on the Web **automatically**

Goal (long term):

- Define and Refine KPI models to assess specific impact of infrastructural investments and gather/check input **automatically**

City Data Pipeline (started 2012)

- <http://citydata.wu.ac.at/>

Open City Data Pipeline

We present the City Data Pipeline – a system for gathering city performance indicators published as Open Data in order to ease the compilation of studies and reports used within Siemens. Under the assumption that Open Data provides means to automatise tedious data research tasks, we have built a system that integrates basic indicators for cities from various Open Data sources. The architecture is flexible, extensible, and natively based on RDF & SPARQL.

[Launch Open City Data Pipeline](#)

The screenshot shows a web browser window with the title 'Daten-Pipeline für Stadtstaaten -- Siemens'. The page features the Siemens logo and the word 'INNOVATION' in large yellow letters. Below the header, there is a navigation menu with 'Home', 'Innovationen', 'Innovation Stories', and 'Daten-Pipeline für Stadtstaaten'. The main content area has the heading 'Nachhaltigere Städte durch Offene Daten' and a sub-heading 'Siemens baut eine Daten-Pipeline für Stadtstaaten.' The text discusses factors of sustainability and the use of open data. A sidebar on the right contains a short summary. At the bottom, a photo shows two people, a woman and a man, looking at a whiteboard. The whiteboard has a hand-drawn diagram of the data pipeline architecture, showing data sources like 'VDF', 'CSU', and 'Smart Village' feeding into a central 'City Data Pipeline' box, which then outputs to 'Analytics & Reporting', 'GIS', and 'AR/VR'.

My background

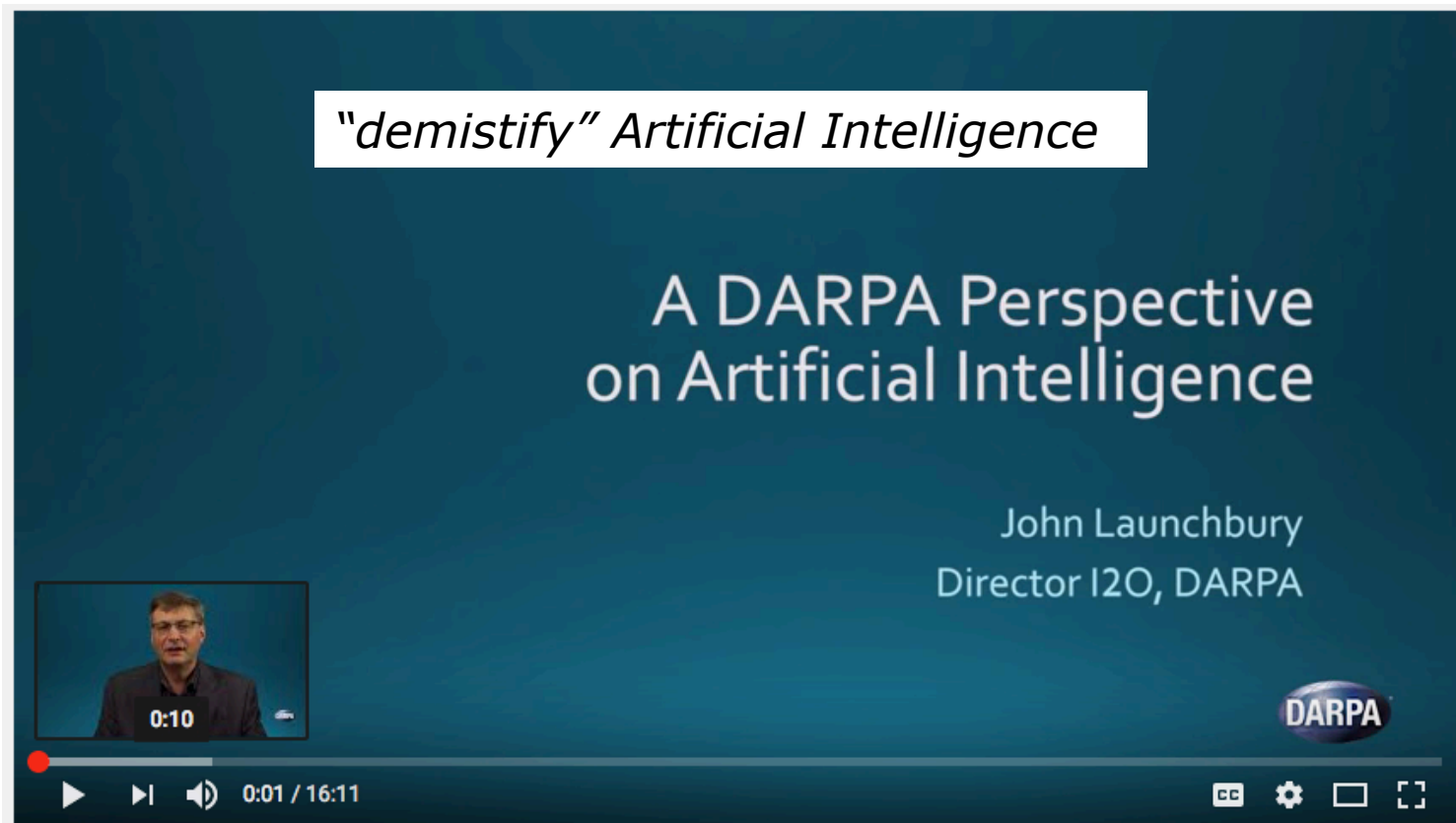
- Logic Programming
- Artificial Intelligence
- Knowledge Representation
- Semantic Web
- Web Data Integration



**What is
Artificial
Intelligence?**

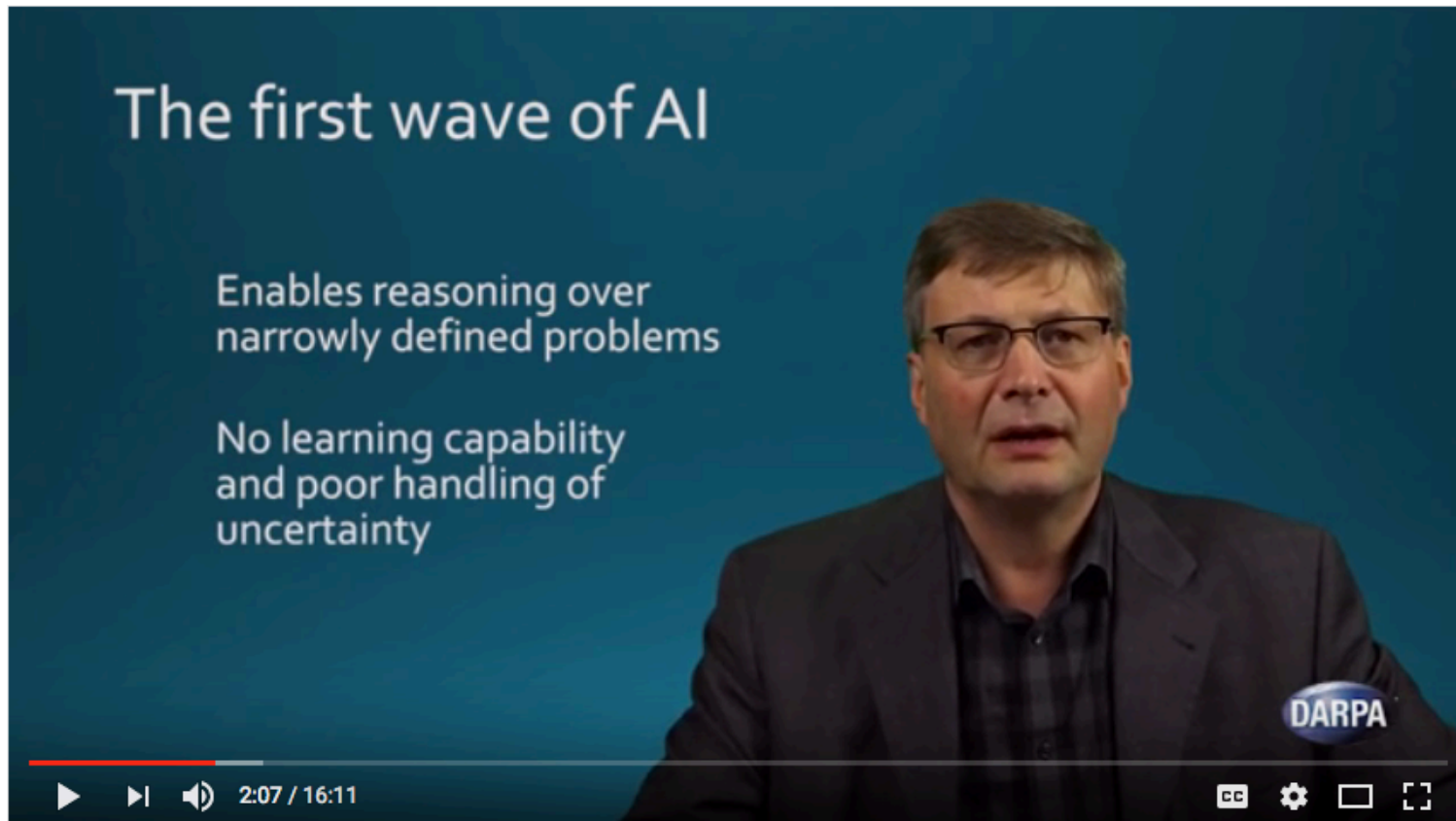
2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017



2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017



The first wave of AI

- Enables reasoning over narrowly defined problems
- No learning capability and poor handling of uncertainty

DARPA

2:07 / 16:11

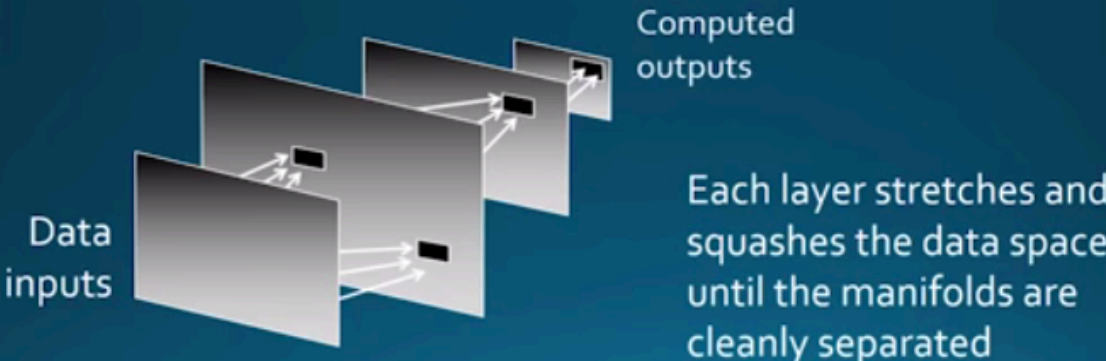
CC Settings Full Screen

The image shows a video player interface. The main content is a slide with a dark blue background. The slide title is 'The first wave of AI'. Below the title, there are two bullet points: 'Enables reasoning over narrowly defined problems' and 'No learning capability and poor handling of uncertainty'. A speaker overlay of John Launchbury is positioned in the lower right quadrant of the slide. The video player controls at the bottom include a play button, a progress bar showing 2:07 / 16:11, and icons for closed captions (CC), settings, and full screen.

2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017

Neural nets learn from data



Data inputs

Computed outputs

Each layer stretches and squashes the data space until the manifolds are cleanly separated

*"learn from **data**, like spreadsheets on steroids"*

DARPA

8:40 / 10:11

2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017

The (future) third wave of AI

Contextual Adaptation

Systems construct explanatory models
for classes of real world phenomena

DARPA

▶ ▶| 🔊 12:59 / 16:11 CC ⚙️ 🖥️ 🗉

2. What is Artificial Intelligence?

John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017

Models to explain decisions

Training Data

Learning Process

This is a cat:

I understand why
I understand why not
I know when you'll succeed
I know when you'll fail
I know when to trust you
I know why you made that mistake

1. *needs – if you want a combination of first and second wave AI – rules/explanations to model common sense and act in new contexts*
2. *Again, **data** is at it's heart!*

▶ | 🔊 14:11 / 16:11

Which "AI" solutions exist now (on the Web)?

The screenshot shows a Google search for "Vienna". The search bar contains "Vienna" and the results are displayed below. A notification banner at the top reads "Hinweise zum Datenschutz bei Google" with buttons for "SPÄTER ERINNERN" and "ANSEHEN".

WIEN - Nachrichten und Services | VIENNA.AT
www.vienna.at
Alle Nachrichten aus Wien und den Wiener Bezirken sowie Services rund um die Bundeshauptstadt: Veranstaltungen, Wetter, Kino, Theater uvm.

Schlagzeilen

- Rapid: Vienna mehr als ein Test**
LAOLA1.at · vor 2 Tagen
- Rapid will Benefizspiel bei der Vienna unbedingt gewinnen**
derStandard.at · vor 2 Tagen
- Benefizspiel gegen Vienna für Rapid Wien "mehr als ein Test": 5.000 Zuschauer...**
spox.com · vor 2 Tagen

wien.at - Infos und Services aus der Wiener Stadtverwaltung
https://www.wien.gv.at
Wiener Ostermärkte 2017 · Ernst Fuchs-Ausstellung in der Otto Wagner-Villa · Vienna Blues Spring 2017, 20.3. bis 30.4. Die lange Nacht der Unternehmen, 22.3. ...

Vienna – Wikipedia
https://de.wikipedia.org/wiki/Vienna
Vienna steht für: Vienna (Album), Album der Musikgruppe Ultravox aus dem Jahr 1980; Vienna (Band), japanische Progressive-Rock-Band; Vienna ...

Wien
Hauptstadt von Österreich

Wien ist die Bundeshauptstadt von Österreich und zugleich eines der neun österreichischen Bundesländer. [Wikipedia](#)

Wetter: 7 °C, Wind aus N mit 10 km/h, 77 % Luftfeuchtigkeit
Ortszeit: Mittwoch, 20:41
Bevölkerung: 1,741 Millionen (2013) Vereinte Nationen

Kommende Veranstaltungen

- Mi., 29. März 19:00 LP Gasometers of Vienna
- Mi., 22. März 01:30 Cirque du Soleil
- Sa., 25. März 19:00 White Miles

Über 25 weitere ansehen

Interessante Orte Über 10 weitere ansehen

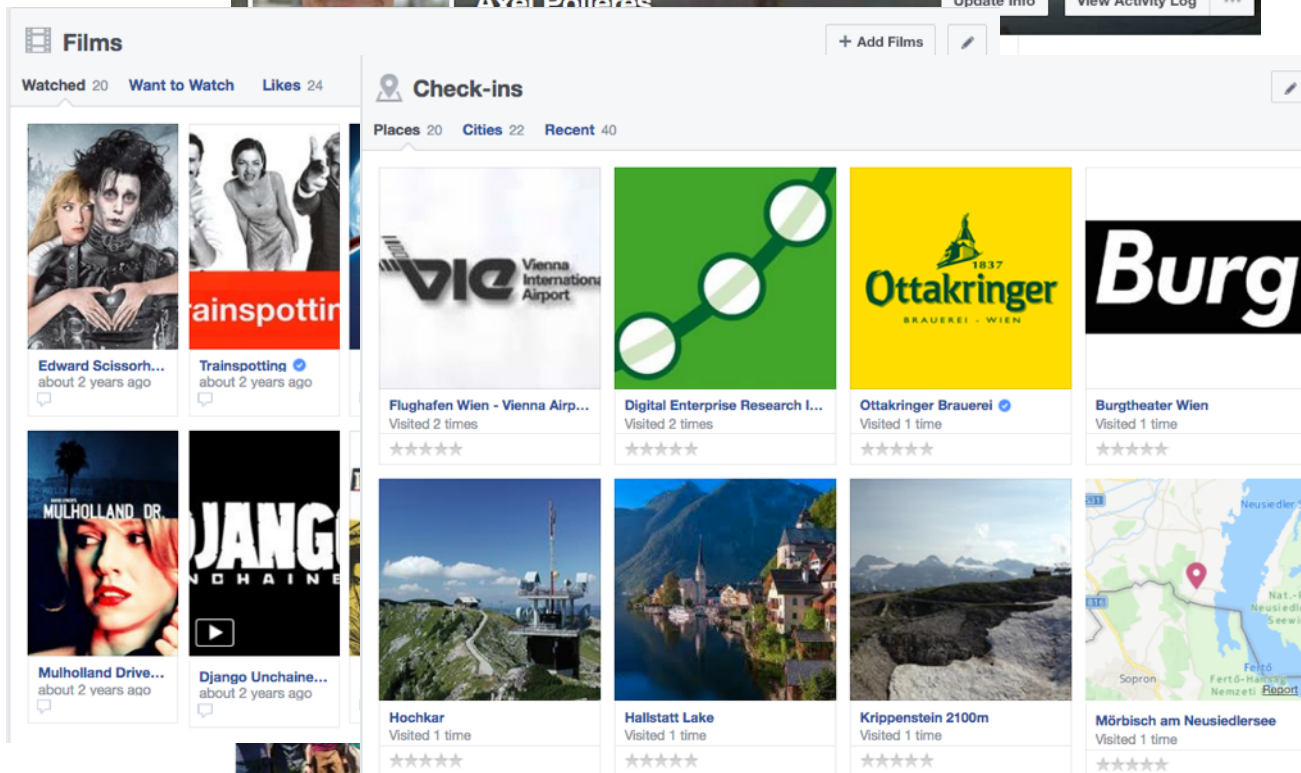
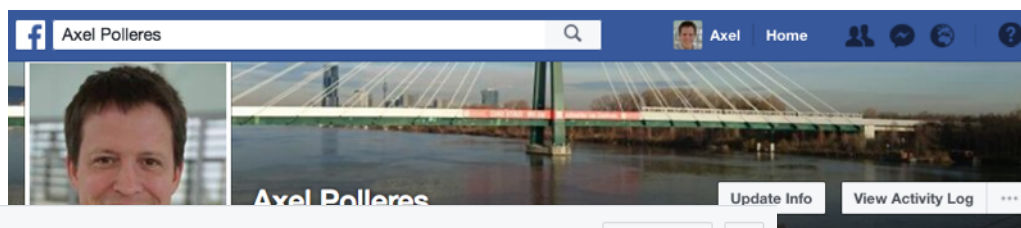
- Schloss Schönbrunn
- Hofburg
- Stephans...
- Wiener Prater
- Schloss Belvedere

Mehr zu Wien

Feedback

Example 1: Google's Knowledge Graph

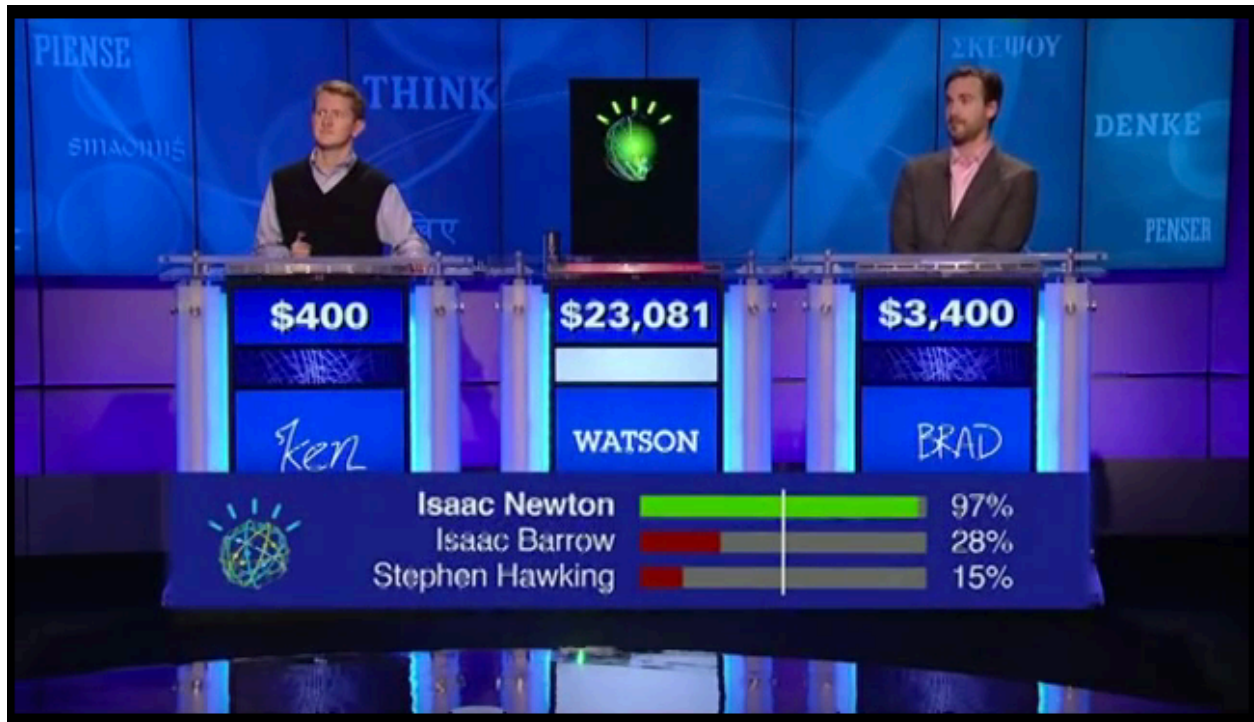
Which "AI" solutions exist now (on the Web)?



Example 2: FB's Social Graph & News Recommendations

Also uses a knowledge graph...

Which "AI" solutions exist now (on the Web)?



**Example 3: IBM
Watson!**

*Also uses a
knowledge
graph...*

<https://youtu.be/P0Obm0DBvwI?t=951>

This is the knowledge Graph that IBM Watson used:



WIKIPEDIA The Free Encyclopedia

Article Talk From Wikipedia, the free encyclopedia

Vienna

"Wien" redirects here. For other uses, see Wien (disambiguation).
This article is about the capital of Austria. For other uses, see Vienna (disambiguation).

Vienna (/ˈviːnə/ [[]listen[ⓘ]; German: Wien, pronounced [viːn]) is the capital and largest city of Austria and one of the nine states of Austria. Vienna is Austria's primary city, with a population of about 1.8 million^[1] (2.6 million within the metropolitan area,^[4] nearly one third of Austria's population), and its cultural, economic, and political centre. It is the 7th-largest city by population within city limits in the European Union. Until the beginning of the 20th century, it was the largest German-speaking city in the world, and before the splitting of the Austro-Hungarian Empire in World War I, the city had 2 million inhabitants.^[1] Today, it has the second largest number of German speakers after Berlin.^{[11][12]} Vienna is host to many major international organizations, including the United Nations and OPEC. The city is located in the eastern part of Austria and is close to the borders of the Czech Republic, Slovakia, and Hungary. These regions work together in a European Cereopros border region. Along with nearby Bratislava, Vienna forms a metropolitan region with 3 million inhabitants.^[13] In 2001, the city centre was designated a UNESCO World Heritage Site.^[13]

Apart from being regarded as the City of Music^[14] because of its musical legacy, Vienna is also said to be "The City of Dreams" because it was home to the world's first psychoanalyst – Sigmund Freud.^[15] The city's roots lie in early Celtic and Roman settlements that transformed into a Medieval and Baroque city, and then the capital of the Austro-Hungarian Empire. It is well known for having played an essential role as a leading European music centre, from the great age of Viennese Classicism through the early part of the 20th century. The historic centre of Vienna is rich in architectural ensembles, including Baroque castles and gardens, and the late-19th-century Ringstraße lined with grand buildings, monuments and parks.^[16]

Vienna is known for its high quality of life. In a 2005 study of 127 world cities, the Economist Intelligence Unit ranked the city first (in a tie with Vancouver, Canada and San Francisco, USA) for the world's most liveable cities. Between 2011 and 2015, Vienna was ranked second, behind Melbourne, Australia.^{[17][18][19][20][21]} For eight consecutive years (2009–2016), the human-resources-consulting firm Mercer ranked Vienna first in its annual "Quality of Living" survey of hundreds of cities around the world, a title the city still holds in 2016.^{[22][23][24][25][26][27][28]} Monocle's 2015 "Quality of Life Survey" ranked Vienna second on a list of the top 25 cities in the world "to make a base within."^{[29][30][31][32][33]}

The UN-Habitat has classified Vienna as being the most prosperous city in the world in 2012/2013.^[34] The city was ranked 1st globally for its culture of innovation in 2007 and 2008, and sixth globally (out of 256 cities) in the 2014 Innovation Cities Index, which analyzed 162 indicators in covering three areas: culture, infrastructure, and markets.^{[35][36]} Vienna regularly hosts urban planning conferences and is often used as a case study by urban planners.^[38]

Between 2005 and 2010, Vienna was the world's number-one destination for international congresses and conventions.^[39] It attracts over 6.8 million tourists a year.^[40]

Contents

- Eymology
- History
 - Early history
 - Austro-Hungarian Empire and the early 20th century
 - Anschluss and World War II
 - Four-power Vienna
 - Austrian State Treaty and aftermath
- Demographics
- Geography and climate
- Districts and enlargement
- Politics
- Economy
 - Research and development
 - Information technologies
 - Tourism and conferences
- Rankings
- Urban development
 - Central Railway Station
 - Aspm
 - Smart City
- Religion
- Culture
 - Music, theatre and opera

Publications: English, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Cross-Domain: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Social Networking: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Government: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Media: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

User-Generated Content: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Linguistics: Afrikaans, Albanian, Arabic, Armenian, Azerbaijani, Basque, Bengali, Bhojpuri, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, Esperanto, Estonian, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Zulu

Linked Datasets as of August 2014

Country: Austria

Government: Mayor and Governor: Michael Häupl (SPÖ); Vice-Mayors and Vice-Governors: Maria Vassilakou (Grüne); Johann Gudenus (FPÖ)

Area: Capital city: 414.65 km² (160.10 sq mi); Land: 365.26 km² (141.03 sq mi); Water: 19.39 km² (7.49 sq mi)

Population: 1,811,000 (as of 2016)

Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak.
<http://lod-cloud.net/>

Open Data is a global trend (also apart from Linked Data):

- Cities, International Organizations, National and European Portals, Int'l. Conferences:



Ok, now... how can I use it?

Recall my background:

- Logic Programming
- Artificial Intelligence
- Knowledge Representation
- Semantic Web
- Web Data Integration

Attempt 1: use "first wave AI"

The image shows a screenshot of a Siemens website article titled "Daten-Pipeline für Stadt- und Unternehmensdaten". The article discusses the use of open data for sustainable cities and mentions that the pipeline is similar to a search engine, pulling data from Wikipedia and web portals. Below the article is a photo of two people, a woman and a man, looking at a whiteboard. The whiteboard contains a hand-drawn diagram of a data pipeline. The diagram shows data sources on the left (WIKI, VDF, CSI, Focused Creative) feeding into a central processing stage (Semantic Web, RDF, GIS). From this stage, data flows to "Analog & Digital" and "IoT/GIS", which then leads to "APIs" and "Regel-Grund".

A concrete use case: The "City Data Pipeline"



European Union Open Data Portal

Wrapper components

CSV

RDF

HTML

RTF

XLS

GML

OSM

Integration Component

Extensible City Data Model

That's a standard ETL pipeline, isn't it?

RDF Triple Store

RDFS Reasoner
SPARQL Engine

GIS Database

Analytics

Aggregation

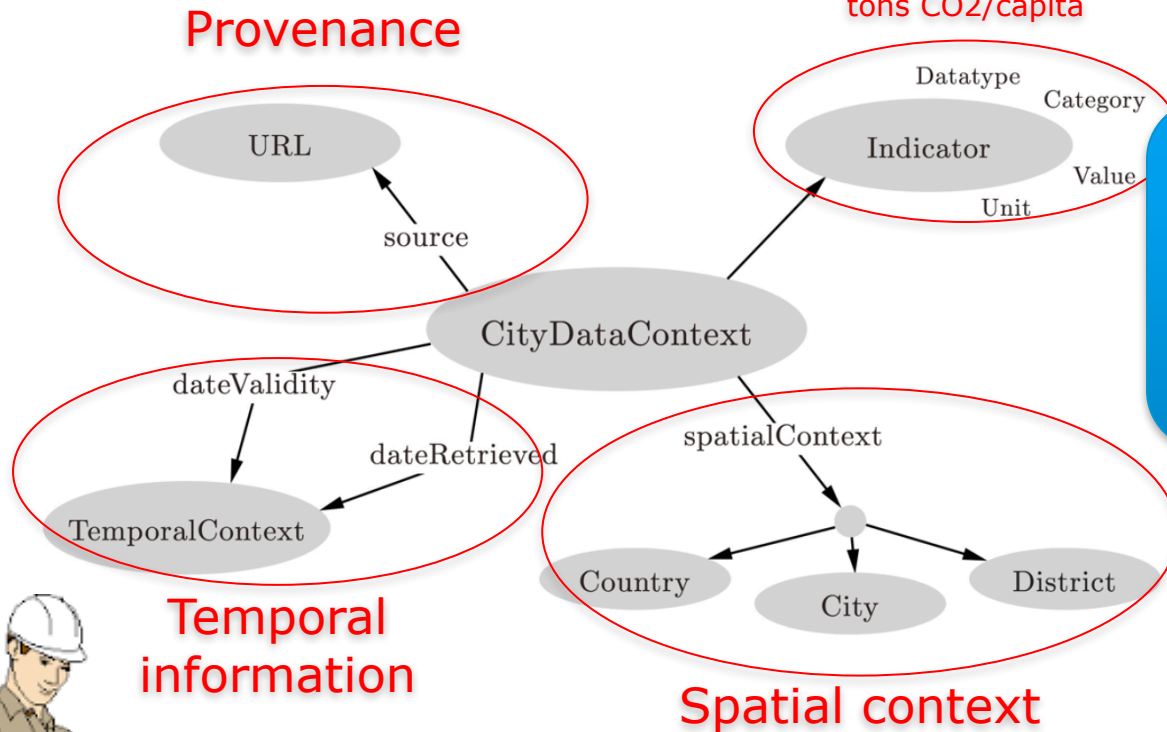
Interpolation

Clustering



A concrete use case: The "City Data Pipeline"

City Data Model: extensible
 $\mathcal{ALH}(\mathbf{D})$ ontology:



But we use and flexible Semantic integration using **ontologies** and **reasoning!**



A concrete use case: The "City Data Pipeline"

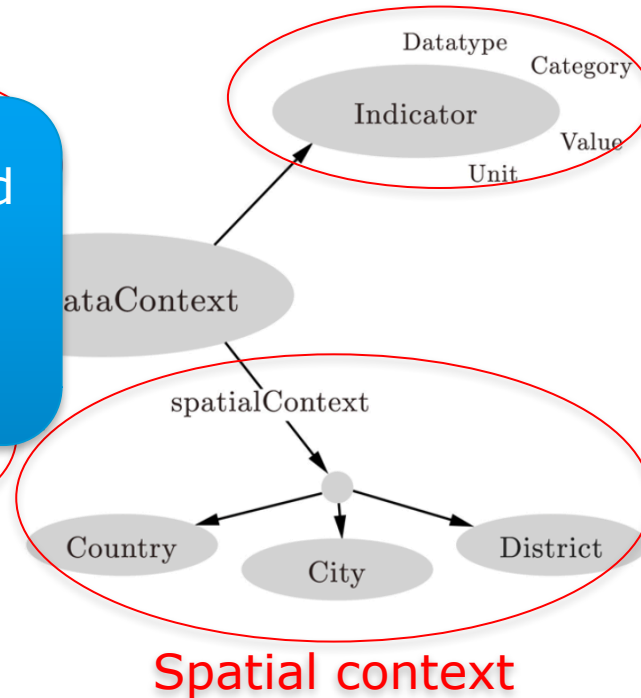
City Data Model: extensible
 $\mathcal{ALH}(\mathbf{D})$ ontology:

Provenance

Indicators,
e.g. area in km²,
tons CO₂/capita

dbpedia:areakm \sqsubseteq :area
eurostat:area \sqsubseteq :area

Ok, we only need
role hierarchies
here? Are we
done?



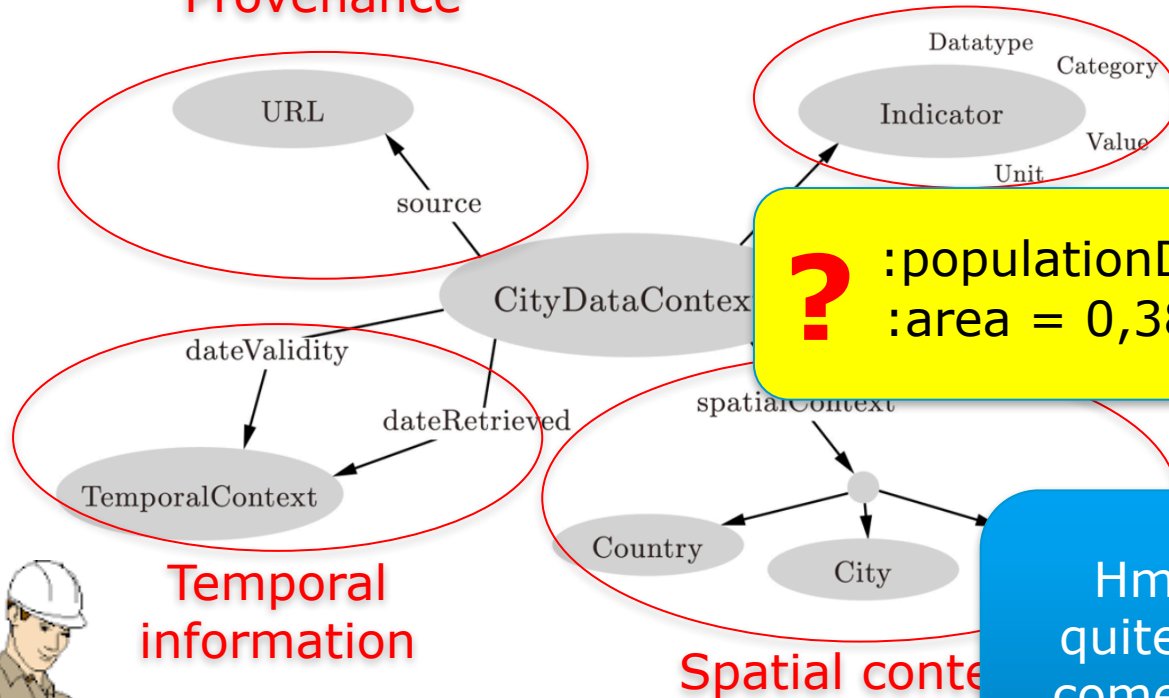
A concrete use case: The "City Data Pipeline"

City Data Model: extensible
 $\mathcal{ALH}(\mathbf{D})$ ontology:

Provenance

Indicators,
e.g. area in km²,
tons CO₂/capita

dbpedia:areakm2 \sqsubseteq :area
eurostat:area \sqsubseteq :area



? :populationDensity = :population/:area
:area = 0,386102 * dbpedia:areaMi2

Temporal
information

Spatial conte

Hmmm, not quite... Let me come up with a solution...



Can equational knowledge co-exist with OWL?

RDFS with Attribute Equations via SPARQL Rewriting

Stefan Bischof^{1,2} and Axel Polleres¹

¹ Siemens AG Österreich, Siemensstraße 90, 1210 Vienna, Austria

² Vienna University of Technology, Favoritenstraße 9, 1040 Vienna, Austria

Abstract. In addition to taxonomic knowledge about concepts and properties typically expressible in languages such as RDFS and OWL, implicit information in an RDF graph may be likewise determined by arithmetic equations. The main use case here is exploiting knowledge about functional dependencies among numerical attributes expressible by means of such equations. While some of this knowledge can be encoded in rule extensions to ontology languages, we provide an arguably more flexible framework that treats attribute equations as first class citizens in the ontology language. The combination of ontological reasoning and attribute equations is realized by extending query rewriting techniques already successfully applied for ontology languages such as (the DL-Lite-fragment of) RDFS or OWL, respectively. We deploy this technique for rewriting SPARQL queries and discuss the feasibility of alternative implementations, such as rule-based approaches.

1 Introduction

A wide range of literature has discussed completion of data represented in RDF with implicit information through ontologies, mainly through taxonomic reasoning within a hierarchy of concepts (classes) and roles (properties) using RDFS and OWL. However, a

Stefan Bischof, Axel Polleres. ESWC2013

Can equational knowledge co-exist with OWL?

- *Can equational knowledge co-exist with OWL?*
 - *We need a syntax & define a formal semantics*
- *Syntax:*
 - $\text{:populationDensity} = \text{:population} / \text{:area}$
 - $\text{:area} = 0,386102 * \text{dbpedia:areaMi2}$

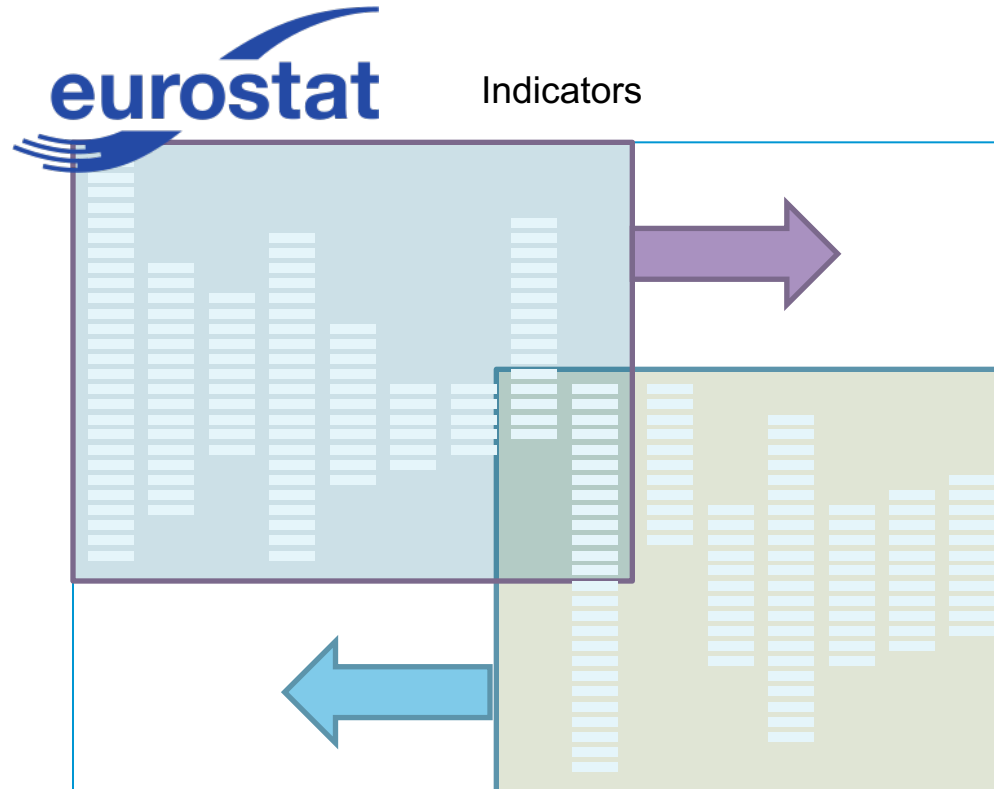
```
:populationDensity :defineByEquation "population/:area" .  
:area :defineByEquation "areaMi2 * 0,386102" .  
dbPedia:populationTotal :rdfs:subPropertyOf :population.
```

- **Semantics:**
 - **Requirements:**
 - "Fit" with common model-theoretic semantics for OWL and RDFS
 - Treat equivalent equations equivalently, combine with **query rewriting** and **rule-based reasoning** techniques:

$$\text{:area} = 0,386102 * \text{dbpedia:areaMi2}$$

$$\text{:areaMi2} = 2,589988 * \text{:area}$$

Challenges – Too many Missing values



Goal: equational knowledge is not enough...

Idea: using both first-wave and second wave AI methods

Challenges – Too many Missing values

- Individual datasets (e.g. from Eurostat) have missing values
- **Merging together datasets** with different indicators/cities adds sparsity

Data from Source 1

	Vienna	Augsburg	Valletta
Cars	655806	111561	95858
Nationals	1342704	216289	203657
Women per 1000 Men	109.8	108.7	101.9

Data from Source 2

	Marbella	Stockholm	Funchal
Available Beds per 1000	138.3	14969	166.1
Average area of living	36.42	37.24	38.16
Cinema Seats	4691	12751	2676



Combined data from Source 1 and Source 2

	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858			
Nationals	1342704	216289	203657			
Women per 1000 Men	109.8	108.7	101.9			
Available Beds per 1000				138.3	14969	166.1
Average area of living				36.42	37.24	38.16
Cinema Seats				4691	12751	2676

Missing Values – Hybrid approach choose best prediction method per indicator:

- Our **assumption**: every indicator has its own distribution and relationship to others.
- Basket of „**standard**“ **regression** methods:
 - K-Nearest Neighbour Regression (KNN)
 - Multiple Linear Regression (MLR)
 - Random Forest Decision Trees (RFD)

▪

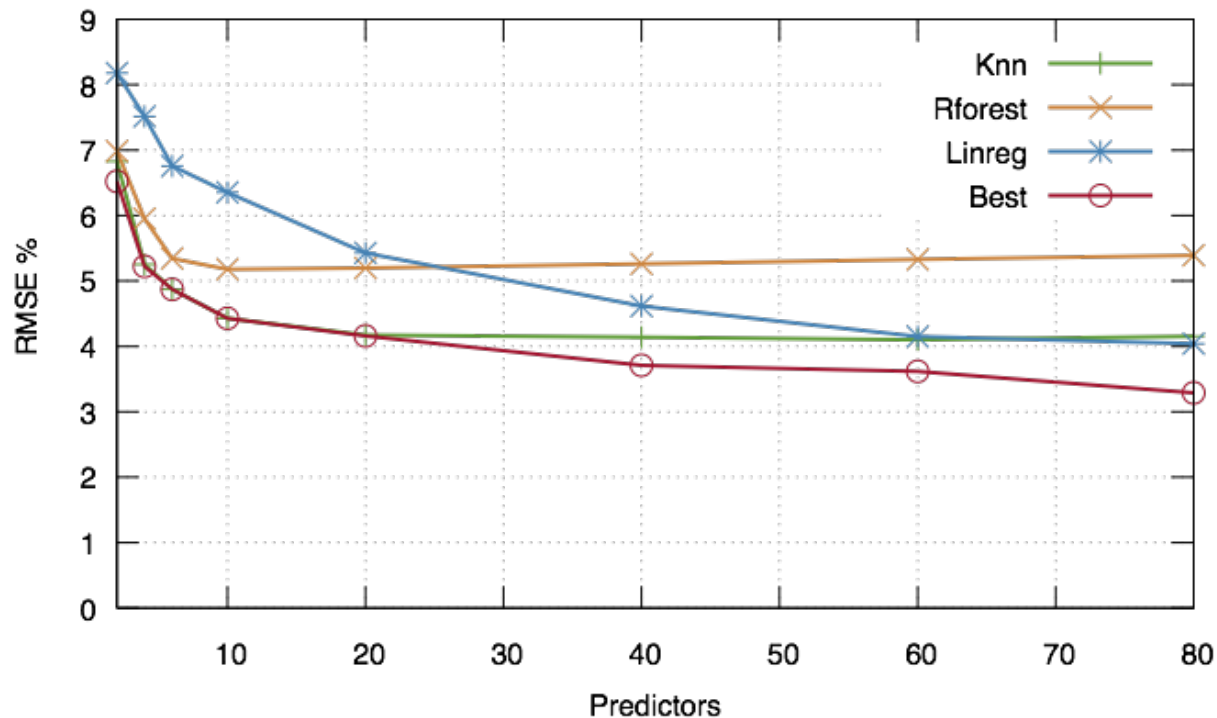
▪



Missing Values – Hybrid approach choose best prediction method per indicator:

- Instead of using indicators directly we use **Principle Components**, built from the indicators
- For building the PCs, **fill in** missing data points with **neutral values** → predict all rows

-
-



More Details:

Stefan Bischof, Christoph Martin, Axel Polleres, and Patrik Schneider. Open City Data Pipeline: Collecting, Integrating, and Predicting Open City Data. In 4th Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD), co-located with ESWC2015, Portoroz, Slovenia, May 2015.

Open City Data Pipeline

Collecting, Integrating, and Predicting Open City Data

Stefan Bischof^{1,2}, Christoph Martin², Axel Polleres², and Patrik Schneider^{2,3}

¹ Siemens AG Österreich, Vienna, Austria

² Vienna University of Economics and Business, Vienna, Austria

³ Vienna University of Technology, Vienna, Austria

Abstract. Having access to high quality and recent data is crucial both for decision makers in cities as well as for informing the public, likewise, infrastructure providers could offer more tailored solutions to cities based on such data. However, even though there are many data sets containing relevant indicators about cities available as open data, it is cumbersome to integrate and analyze them, since the collection is still a manual process and the sources are not connected to each other upfront. Further, disjoint indicators and cities across the available data sources lead to a large proportion of missing values when integrating these sources. In this paper we present a platform for collecting, integrating, and enriching open data about cities in a re-usable and comparable manner: we have integrated various open data sources and present approaches for predicting missing values, where we use standard regression methods in combination with principal component analysis to improve quality and amount of predicted values. Further, we re-publish the integrated and predicted values as linked open data.

Next step:

Combine ML and equations
“iteratively” (under submission)

<http://epub.wu.ac.at/5438/>

City Data Pipeline

citydata.wu.ac.at

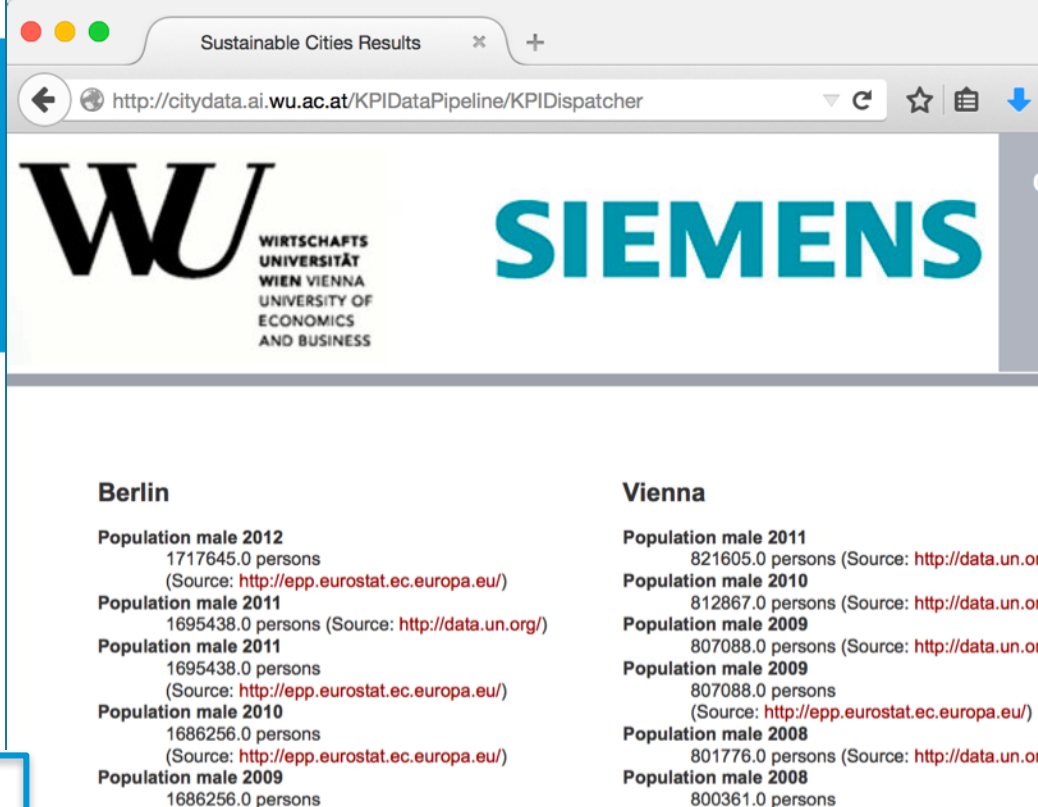
- Search for indicators & cities
- obtain results incl. sources
- Integrated data served as Linked Open Data
- Predicted values AND **estimated error rates** for missing data...



Vienna

Municipal waste (1000 t)

- › **2004:** 778.905392176222 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2005:** 813.77643147163 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2006:** 813.889824195497 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2007:** 811.538914636665 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2008:** 811.010344391444 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2009:** 811.172539879368 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)



The screenshot shows a web browser displaying the website <http://citydata.wu.ac.at/KPIDataPipeline/KPIDispatcher>. The page features the logos for WU (Wirtschaftsuniversität Wien) and Siemens. Below the logos, there are two columns of data for Berlin and Vienna. The Berlin data includes population male for the years 2010, 2011, and 2012, with sources cited as <http://epp.eurostat.ec.europa.eu/> and <http://data.un.org/>. The Vienna data includes population male for the years 2008, 2009, and 2010, with sources cited as <http://data.un.org/> and <http://epp.eurostat.ec.europa.eu/>.

City	Year	Population (persons)	Source
Berlin	2012	1717645.0	http://epp.eurostat.ec.europa.eu/
	2011	1695438.0	http://data.un.org/
	2010	1686256.0	http://epp.eurostat.ec.europa.eu/
Vienna	2011	821605.0	http://data.un.org/
	2010	812867.0	http://data.un.org/
	2009	807088.0	http://epp.eurostat.ec.europa.eu/

...it's not finished, but:
assumption: Predictions get better, the more Open data we integrate...



However:

(Strong) Limitations:

- We combined 3-4 specific OD sources (there are 100s of Open Data Portals out there)
- We manually created an ontology for mapping those sources and set of equations from eurostat?

Open Questions:

- How can I build a scalable repository of Open Data?
- How can I automate finding relevant data?
- How can I automatize building an Open Data Knowledge graph?

Open Data Portals

CKAN ... <http://ckan.org/>

- almost „de facto“ standard for Open Data Portals
- facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets

• <http://opendataportal.at/>

• <http://data.gv.at/>

- machine-processable? ...
... **partially**

The screenshot shows the homepage of data.gv.at. The browser address bar displays 'http://www.data.gv.at/'. The page features a search bar with the placeholder text 'Suchbegriff (z.B. Finanzen, Wahlen)' and a 'Suche starten' button. Below the search bar, there are navigation links for 'Datenkatalog', 'Anwendungen & News', and 'Katalog durchstöbern'. The main content area has a heading 'offene Daten Österreichs – lesbar für Mensch und Maschine' and a sub-heading 'Vielfalt, Transparenz, Offenheit, Demokratie'. It describes the portal as a 'Katalog offener Datensätze und Dienste' and provides information on how to use the data. A diagram on the right shows a computer monitor displaying binary code, with arrows pointing to a group of people and a smartphone, illustrating the accessibility of the data for both humans and machines.

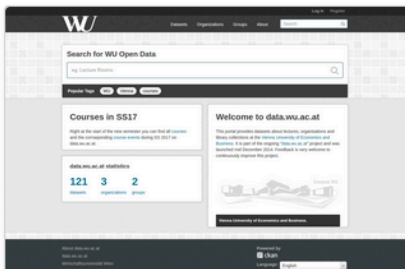
Our ongoing research: data.wu.ac.at



- ***What is the status of Open Data and what are the challenges using Open Data?***
 - OpenData PortalWatch – a project at WU
 - Improving Open Data Quality and Access: ADEQUATE (FFG)
- ***What's next?***
 - Making Open Data Searchable
 - Building an Open Data **Knowledge Graph!**
- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?

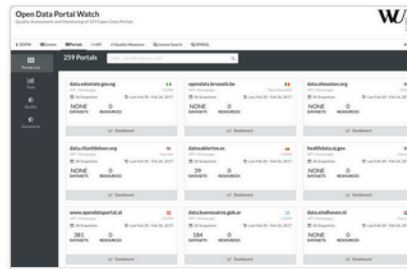
Ongoing Projects (data.wu.ac.at)

Projects



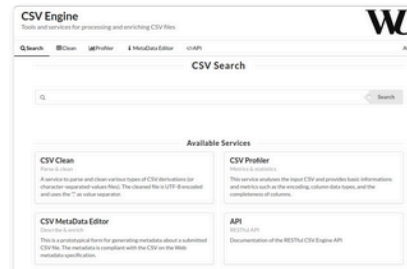
WU Open Data Portal
WU lectures, rooms and organizations
data.wu.ac.at is an Open Data portal where you can find data about lectures, rooms and organizations at WU.

121 datasets

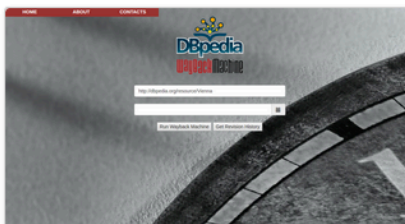


Open Data Portal Watch
Monitoring & exposing portals' metadata
Open Data Portal Watch assesses the evolution of the (meta) data quality of about 260 Open Data portals over since September 2014.

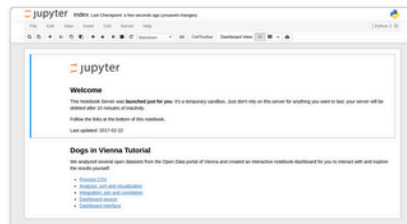
259 portals



CSV Engine
Tools and services for processing and enriching CSV files
Search & enrich CSVs
The CSV Engine is a collection of tools and services for processing and enriching CSV files.




DBpedia Wayback Machine
Extract past DBpedia versions
The DBpedia Wayback Machine aims at providing the wayback functionality for DBpedia based on the revisions of their Wikipedia article.



Jupyter Notebook Server
Programming & Documentation
Notebook documents are documents which contain both computer code (e.g. python) and human-readable rich text elements.

<> Only available within local WU Vienna network



Open Data AT Assistant
Search chatbot for Austrian datasets
The assistant will help you to explore the content of the austrian open data portals: data.gv.at and opendataportal.at.

OPEN DATA PORTAL WATCH

<http://data.wu.ac.at/portalwatch/>

- Periodically monitoring a list of Open Data Portals
 - 260 CKAN powered Open Data Portals worldwide
- Quality assessment
- Evolution tracking
 - Meta data
 - Data
 - Formats, growth

Portalwatch Example:

http://data.wu.ac.at/portalwatch/portal/data_gv_at/1724

Automated Quality Assessment of Metadata across Open Data Portals

SEBASTIAN NEUMAIER, Vienna University of Economics and Business
JÜRGEN UMBRICH, Vienna University of Economics and Business
AXEL POLLERES, Vienna University of Economics and Business

The Open Data movement has become a driver for publicly available data on the Web. More and more data – from governments, public institutions but also from the private sector – is made available online and is mainly published in so called Open Data portals. However, with the increasing number of published resources, there are a number of concerns with regards to the quality of the data sources and the corresponding metadata, which compromise the searchability, discoverability and usability of resources.

In order to get a more complete picture of the severity of these issues, the present work aims at developing a generic metadata quality assessment framework for various Open Data portals: we treat data portals independently from the portal software frameworks by mapping the specific metadata of three widely used portal software frameworks (CKAN, Socrata, OpenDataSoft) to the standardized DCAT metadata schema. We subsequently define several quality metrics, which can be evaluated automatically and in a efficient manner. Finally, we report findings based on monitoring a set of over 260 Open Data portals with 1.1M datasets. This includes the discussion of general quality issues, e.g. the retrievability of data, and the analysis of our specific quality metrics.

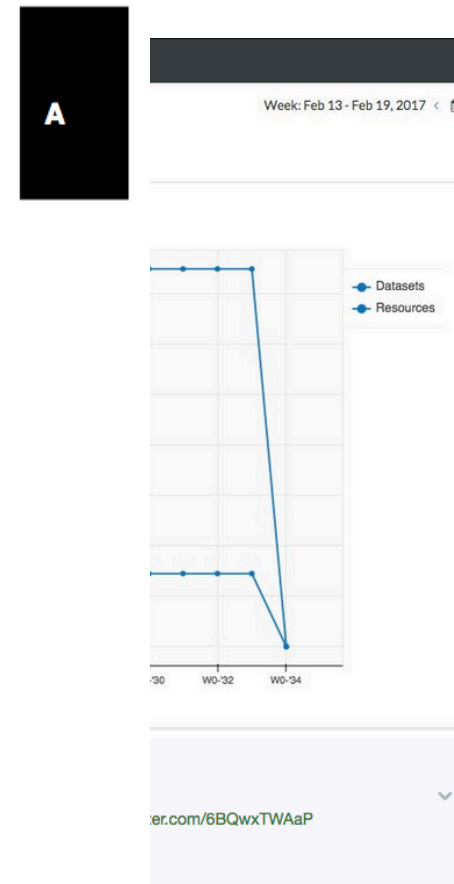
CCS Concepts: •General and reference → Measurement; Metrics; •Information systems → Web searching and information discovery; Digital libraries and archives;

Additional Key Words and Phrases: Open Data, quality assessment, data quality, data portal

ACM Reference Format:

Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres, 2015. Automated Quality Assessment of Metadata across Open Data Portals. *ACM J. Data Inform. Quality* V, N, Article A (January YYYY), 29 pages.

DOI: <http://dx.doi.org/10.1145/0000000.0000000>



Our research: data.wu.ac.at



- ***What is the status of Open Data and what are the challenges using Open Data?***
 - OpenData PortalWatch – a project at WU
 - Improving Open Data Quality and Access: ADEQUATE (FFG)
- ***What's next?***
 - Making Open Data Searchable
 - Building an Open Data **Knowledge Graph!**
- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?

Why is Search in Open Data a problem?

<https://www.youtube.com/watch?v=kCAymmbyIvc>

Structured Data in Web Search by Alon Halevy



VS.

HTML Tables

Beer	Company	ABV	IBU	Color	Style
Novik Wolf Light	A.B. Pilsen Bryggerier (Breweri)	4.7	110		
Turbodog	Abba Brewing Company	5.6	166	15	28
Abbey Ale	Abba Brewing Company	8.0	230	18	32
Piccan	Abba Brewing Company	5.0	150	11	20
Jockamo	Abba Brewing Company	6.5	190	13	52
Red Ale	Abba Brewing Company	5.2	151	11	30
Amber	Abba Brewing Company	4.5	128	10	17
Rock	Abba Brewing Company	6.5	187	16	25
Fat Feet	Abba Brewing Company	5.4	167	15	20
Razoration	Abba Brewing Company	5.0	167	15	20
Andygar	Abba Brewing Company	8.0	235	19	28
Purple Haze	Abba Brewing Company	4.2	128	11	13
Balsura	Abba Brewing Company	5.1	155	11	17
Strawberry	Abba Brewing Company	4.2	120	11	13
Save Our Shore	Abba Brewing Company	7.0	200	15	30
Wheat	Abba Brewing Company	4.2	125	10	15
Golden	Abba Brewing Company	4.2	125	10	11
Light	Abba Brewing Company	4.0	118	8	10
Christmas Ale	Abba Brewing Company	7.5			30

research.google.com/tables

Data Integration as Search

Coffee Consumption around the world

World Population 2

World Merged

FAMA, GACALB

World Countries - Resourcelinks list

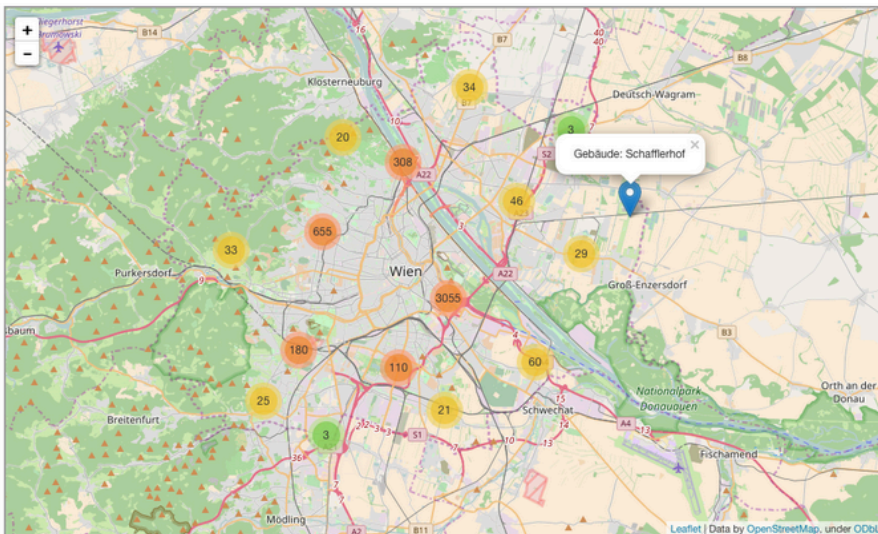
B	C	D	E	F	G	H	I
NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	POP_TOTAL	POP_MEN	POP_WOMEN	REF_DATE
AT13	AT130	90101		0	16131	7726	8405 01.01.2014
AT13	AT130	90201		0	99597	48650	50947 01.01.2014
AT13	AT130	90301		0	86454	41085	45369 01.01.2014
AT13	AT130	90401		0	31452	14903	16549 01.01.2014
AT13	AT130	90501		0	53610	26299	27311 01.01.2014
AT13	AT130	90601		0	30613	14833	15780 01.01.2014
AT13	AT130	90701		0	30792	14703	16089 01.01.2014
AT13	AT130	90801		0	24279	11855	12424 01.01.2014
AT13	AT130	90901		0	40528	19286	21242 01.01.2014
AT13	AT130	91001		0	186450	91638	94812 01.01.2014
AT13	AT130	91101		0	93440	45541	47899 01.01.2014
AT13	AT130	91201		0	90874	43752	47122 01.01.2014

Open Data Search is hard...

- a) No natural language „cues“ like in Web tables...
- b) Existing knowledge graphs don't cover the domain of "Open Data"
- c) Open Data is not properly geo-referenced

Some starting points:

- First baby steps on building an Open Data Knowledge Graph:
- Ongoing work to make
- Open Data **geo-searchable** e.g. in our project communidata.at:



International Semantic Web conference 2016:

Multi-level semantic labelling of numerical values

Sebastian Neumaier¹, Jürgen Umbrich¹, Josiane Xavier Parreira², and Axel Polleres¹

¹ Vienna University of Economics and Business, Vienna, Austria

² Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to “semantically label” such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual “cues” are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible “semantic contexts” for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.

Towards linking Open Data to a Knowledge Graph

- Attempt to link numeric Open data to the dbpedia knowledge graph...

International Semantic Web conference 2016:

Multi-level semantic labelling of numerical values

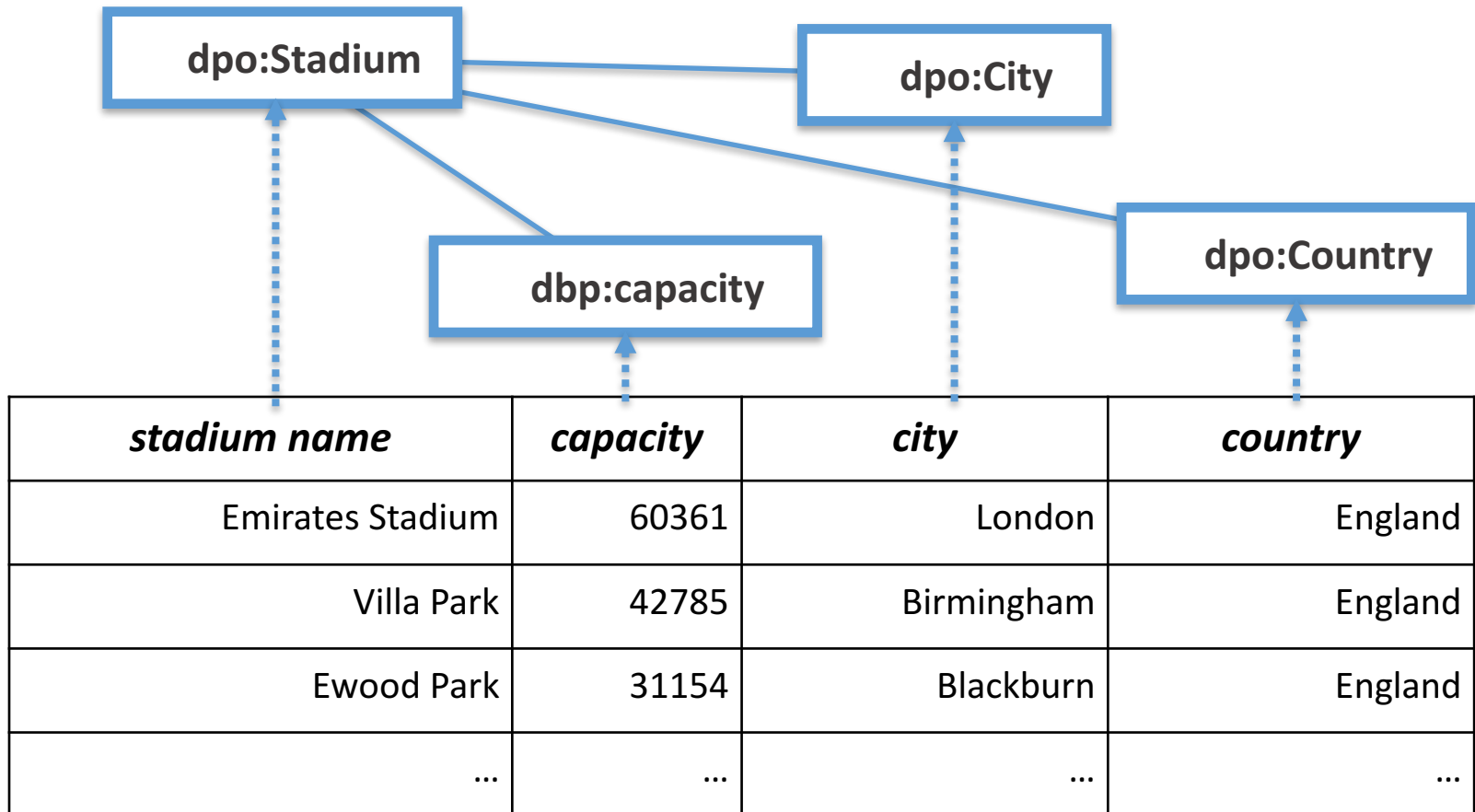
Sebastian Neumaier¹, Jürgen Umbrich¹, Josiane Xavier Parreira², and Axel Polleres¹

¹ Vienna University of Economics and Business, Vienna, Austria

² Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to “semantically label” such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual “cues” are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible “semantic contexts” for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.

Example



But:

Web/HTML tables differ from typical Open Data tables:

- **Domain:** e.g., public administration data, statistical data, weather data, elections, ...
- **Structure:** OD tables contain large amount of numerical columns

NUTS1	NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	WHG_TOTAL
AT1	AT13	AT130	90100	90101	3004
AT1	AT13	AT130	90100	90102	1049
AT1	AT13	AT130	90100	90103	1389
AT1	AT13	AT130	90100	90104	1014
AT1	AT13	AT130	90100	90105	1337
AT1	AT13	AT130	90100	90106	1915
AT1	AT13	AT130	90100	90107	2032
AT1	AT13	AT130	90200	90201	5178
AT1	AT13	AT130	90200	90202	6345
AT1	AT13	AT130	90200	90203	7549
AT1	AT13	AT130	90200	90204	8388
AT1	AT13	AT130	90200	90205	5358
AT1	AT13	AT130	90200	90206	4237
AT1	AT13	AT130	90200	90207	7812
AT1	AT13	AT130	90200	90208	1478
AT1	AT13	AT130	90200	90209	7547

Example (Cont'd)

<i>stadium</i>	<i>capacity</i>	<i>city</i>	<i>country</i>
Emirates Stadium	60361	London	England
Villa Park	42785	Birmingham	England
Ewood Park	31154	Blackburn	England
...

Example (Cont'd)

	<i>TOTAL</i>	<i>DISTRICT_CODE</i>	<i>ISO_2</i>
Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
...

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

	<i>TOTAL</i>	<i>DISTRICT_CODE</i>	<i>ISO_2</i>
Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
...

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
...

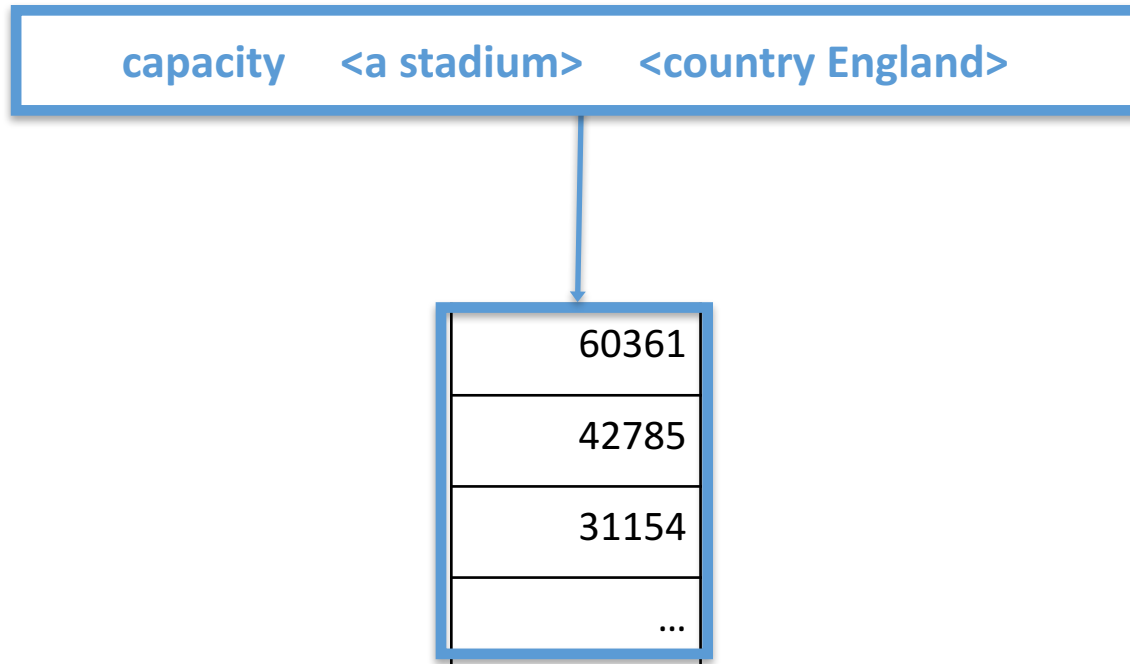
Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

60361
42785
31154
...

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings



Our Approach

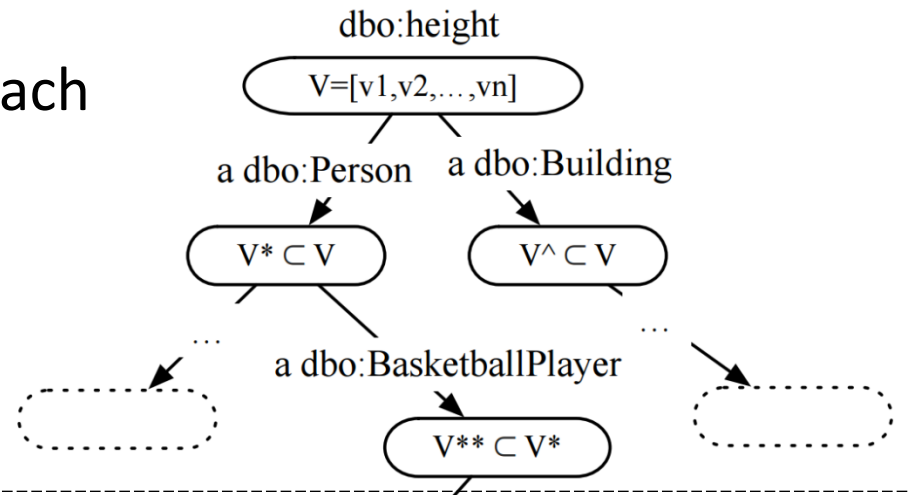
- 1. Hierarchical clustering** over an RDF knowledge base
 - to build background knowledge graph (**BKG**)
 - nodes consist of **typical numerical values**, annotated with context information, i.e.:
 - grouped by **properties** and their **shared domain (subject) pairs**
- 2. k-nearest neighbors search**
- 3. Aggregation of the results** at different levels to find the most likely context:
 - property
 - type
 - context

1. Background Knowledge Graph

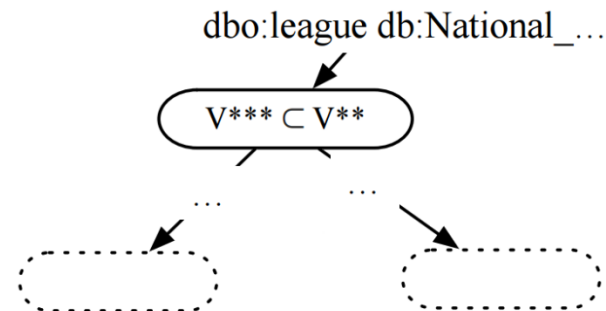
- Find properties with **numerical range**
- Hierarchical clustering approach

- Two hierarchical layers:

- **Type** hierarchy
(using OWL classes)

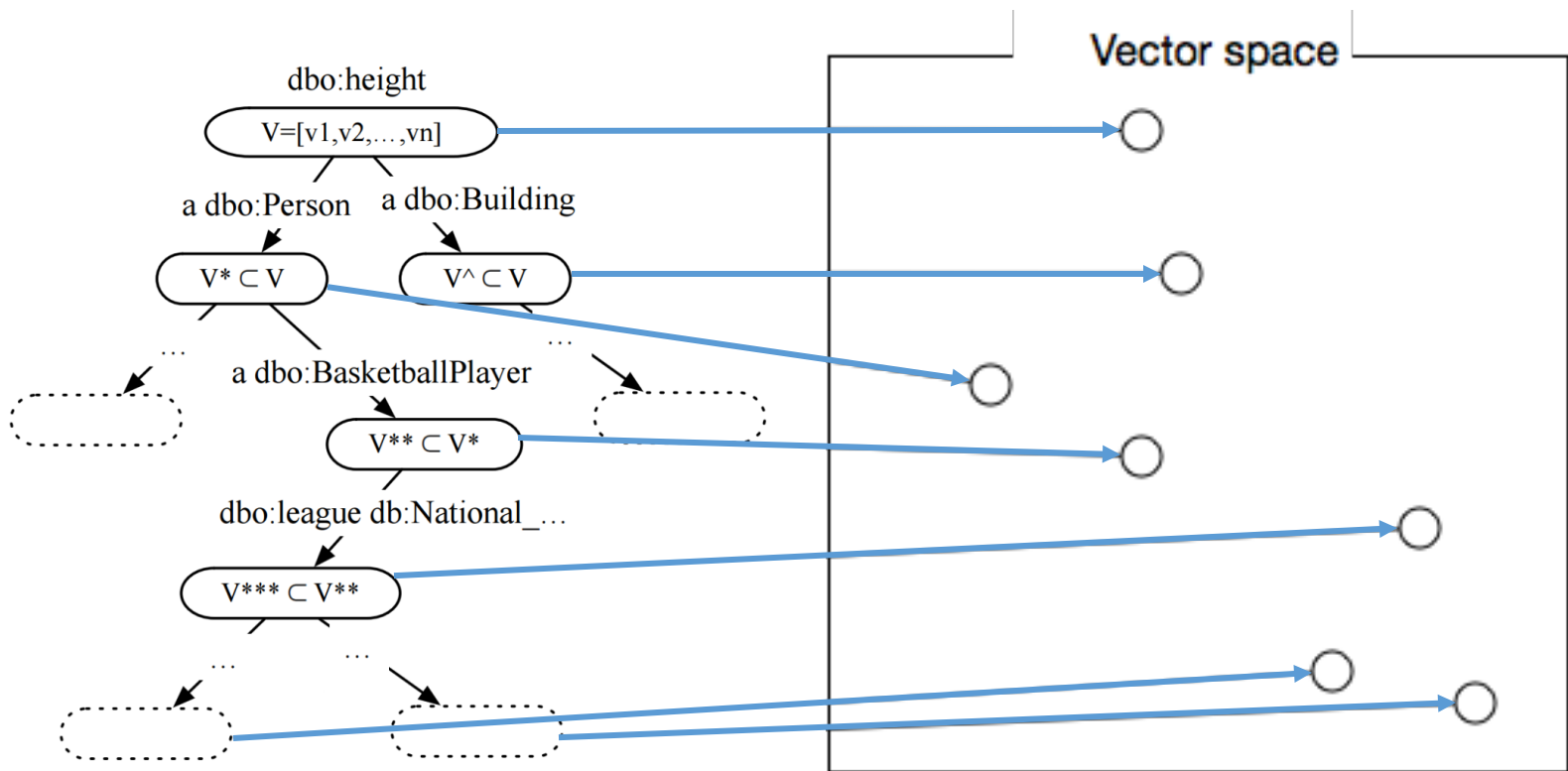


- **Property-object** hierarchy
(shared property-object pairs)



2. *k*-Nearest neighbor search

Mapping bags of numerical value to vector space (feature vector)



Our research: data.wu.ac.at



- ***What is the status of Open Data and what are the challenges using Open Data?***
 - OpenData PortalWatch – a project at WU
 - Improving Open Data Quality and Access: ADEQUATE (FFG)
- ***What's next?***
 - Making Open Data Searchable
 - Building an Open Data **Knowledge Graph!**
- A striving **Data Economy** needs no silos... re-democratise the Web by Cognitive Intelligence based on Open Data?

This is t
that IBM

[Google Knowledge Graph](#)
[Facebook Graph Search](#)

Out of Facebook Graph Search and Google Knowledge Graph, which is more revolutionary, creative and useful?

Say after both these graphs grow to their full extent

Compare and contrast Facebook's [Introducing Graph Search](#) and Google's

<http://www.google.co.in/insidese...> in terms of

- 1.Revolution to the internet
- 2.Creativity in their design
- 3.Usefulness to the users

3 Answers



Justin Moore, Engineering Manager at Facebook

Written Mar 19, 2013

You're comparing apples to oranges. Facebook has graph search *and* a knowledge graph (although we didn't give it a name externally that I know of). **Both our knowledge graph and google's have roots in Wikipedia and freebase** and both are semantic knowledge stores. Search for baseball (sport) on Facebook and scroll through the page to see our knowledge graph about players, teams, etc.

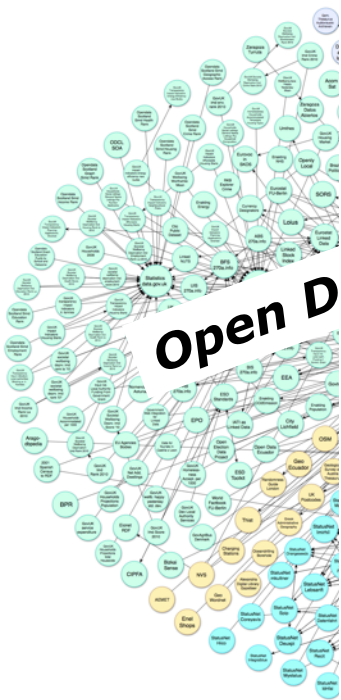
*"Both our knowldege graph and google's habe **roots in Wikipedia and freebase**" – but none of Google and FB make their knowledge graphs freely and openly available again as Open Data!*

Linking Op
McCrae,
<http://l>

Graph search is different. Its structured semantic search on top of structured data like knowledge graph but also all of your connections to people, photos, places. You can't really judge the two any more than comparing the Internet and google.

707 Views · View Upvotes

1 P.



This is a fundamental threat to the Web itself:

<https://www.theguardian.com/technology/2017/mar/11/tim-berners-lee-web-inventor-save-internet>

Internet

Tim Berners-Lee: I invented the web. Here are three things we need to change to save it

It has taken all of us to build the web we have, and now it is up to all of us to build the web we want - for everyone

1) We've lost control of our personal data

2) It's too easy for misinformation to spread on the web

3) Political advertising online needs transparency and understanding



© Sir Tim Berners-Lee, inventor of the worldwide web. Photograph: Sarah Lee for the Guardian

***Open-Data-Fueled Complexity
Science to the rescue?***

Still Open Questions (with some starting points presented...)

- How can I build a scalable repository of Open Data?
- How can I automate finding relevant data?
- How can I automatize building an Open Data Knowledge graph?

- What is the right form of **Knowledge Representation** for Knowledge graphs?
 - OWL, Rules, Equations, Property-domain pairs?)
 - How to represent models in an exchangeable manner?

- Eventually: How can I enable fact checking, verify information on the Web, understand cities,... by Open Data?

Thanks! Things I did NOT have time to talk about:

- Open Data and licences → [DALICC](#)
- Open Data Archiving → Javier
- Open Data adoption barriers → see our recent paper to be presented at [CEDEM2017](#)
- Privacy and data on the Web → <http://privacylab.at>
→ <http://specialprivacy.eu/>
- Something completely different (but maybe related to Ruggiero's work?) Resource allocation in BPM
→ <https://ai.wu.ac.at/shape-project/>
- Organizing Semantic Web conferences:
<https://iswc2017.semanticweb.org/>



Institute for Information Business

