WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Open Data as the fuel for complexity science?

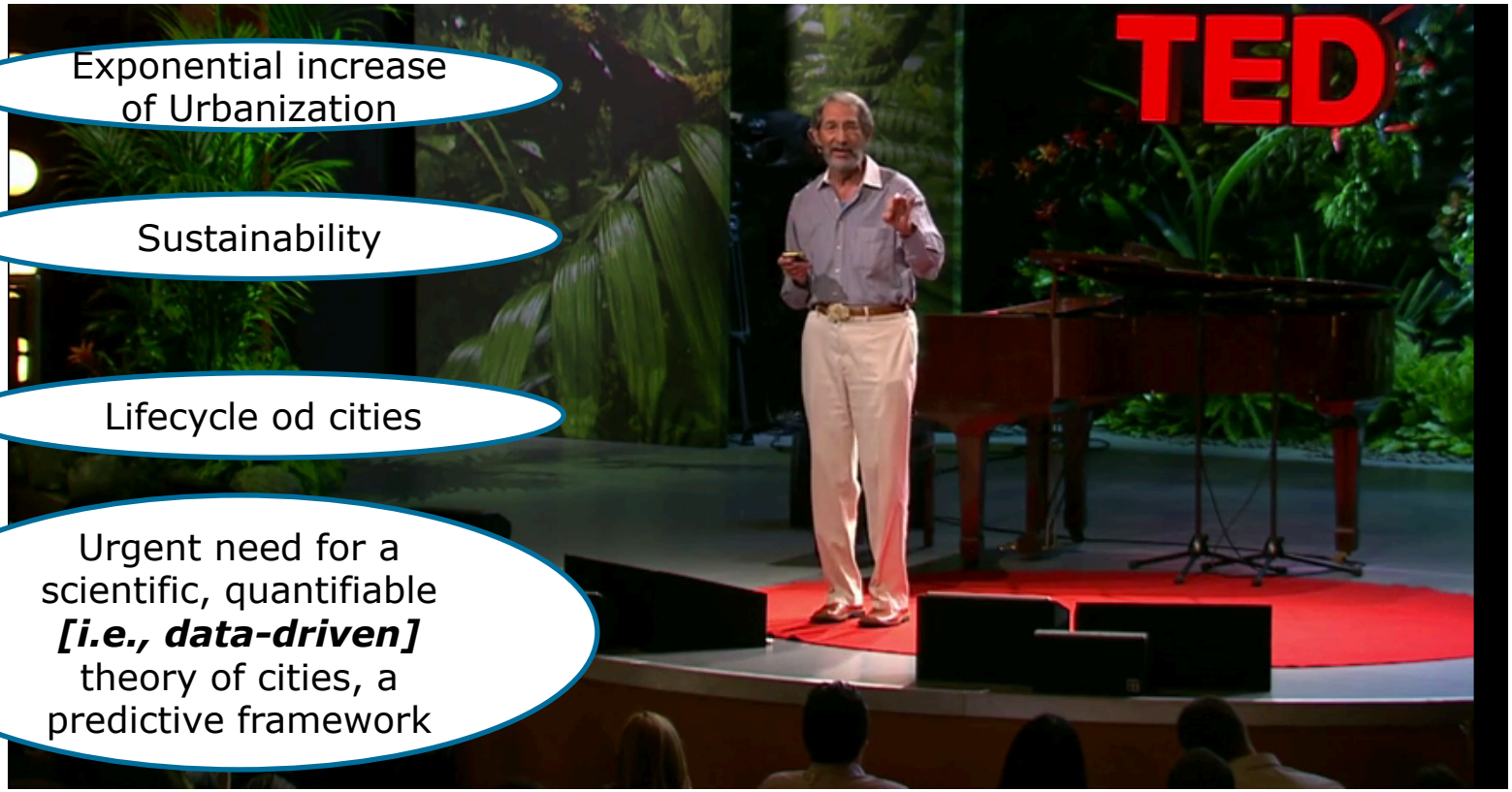
Axel Polleres,

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

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

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



Exponential increase of Urbanization

Sustainability

Lifecycle od cities

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

https://www.ted.com/talks/geoffrey_west_the_surprising_math_of_cities_and_corporations/

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

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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


Megacities


London

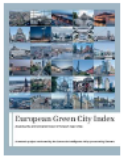

Munich


Dublin


Vienna


Trondheim

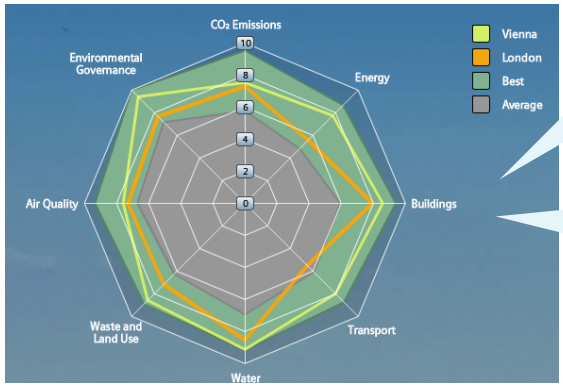

US Mayors


European Green City Index

… however, these are often **outdated** before even published!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

EQUIS ACCREDITED

City Data Pipeline (started 2012)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

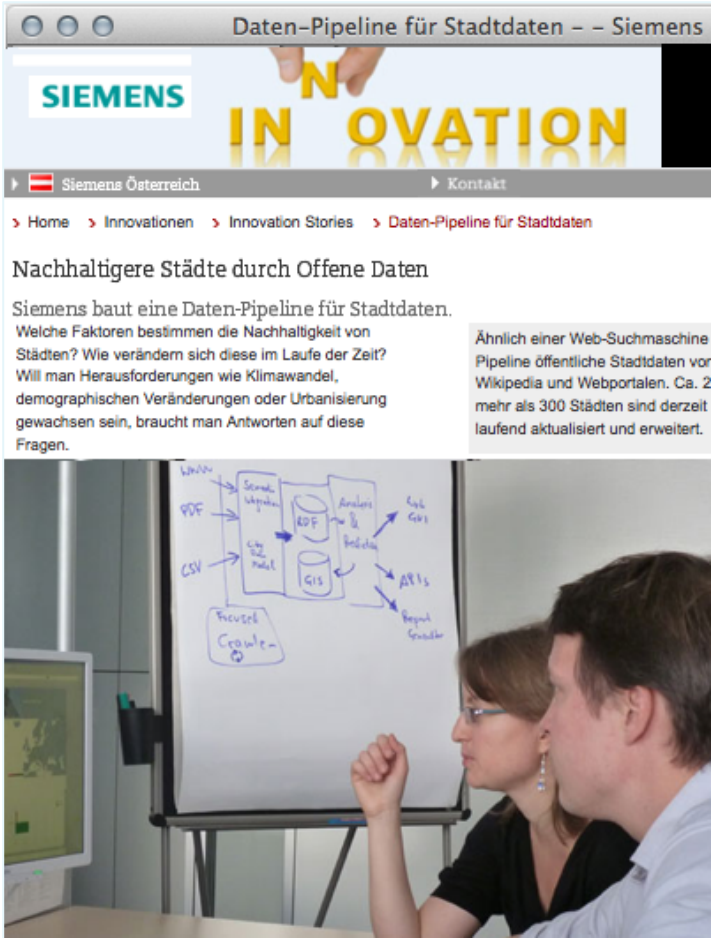
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

SIEMENS

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.

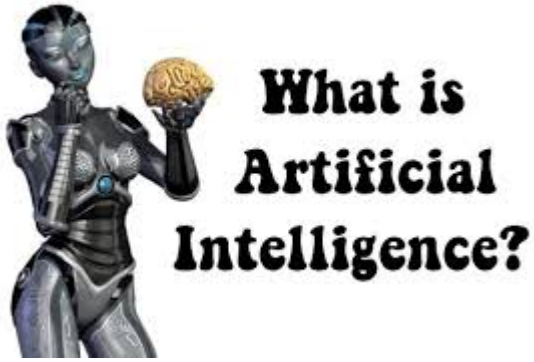
Launch Open City Data Pipeline



My background

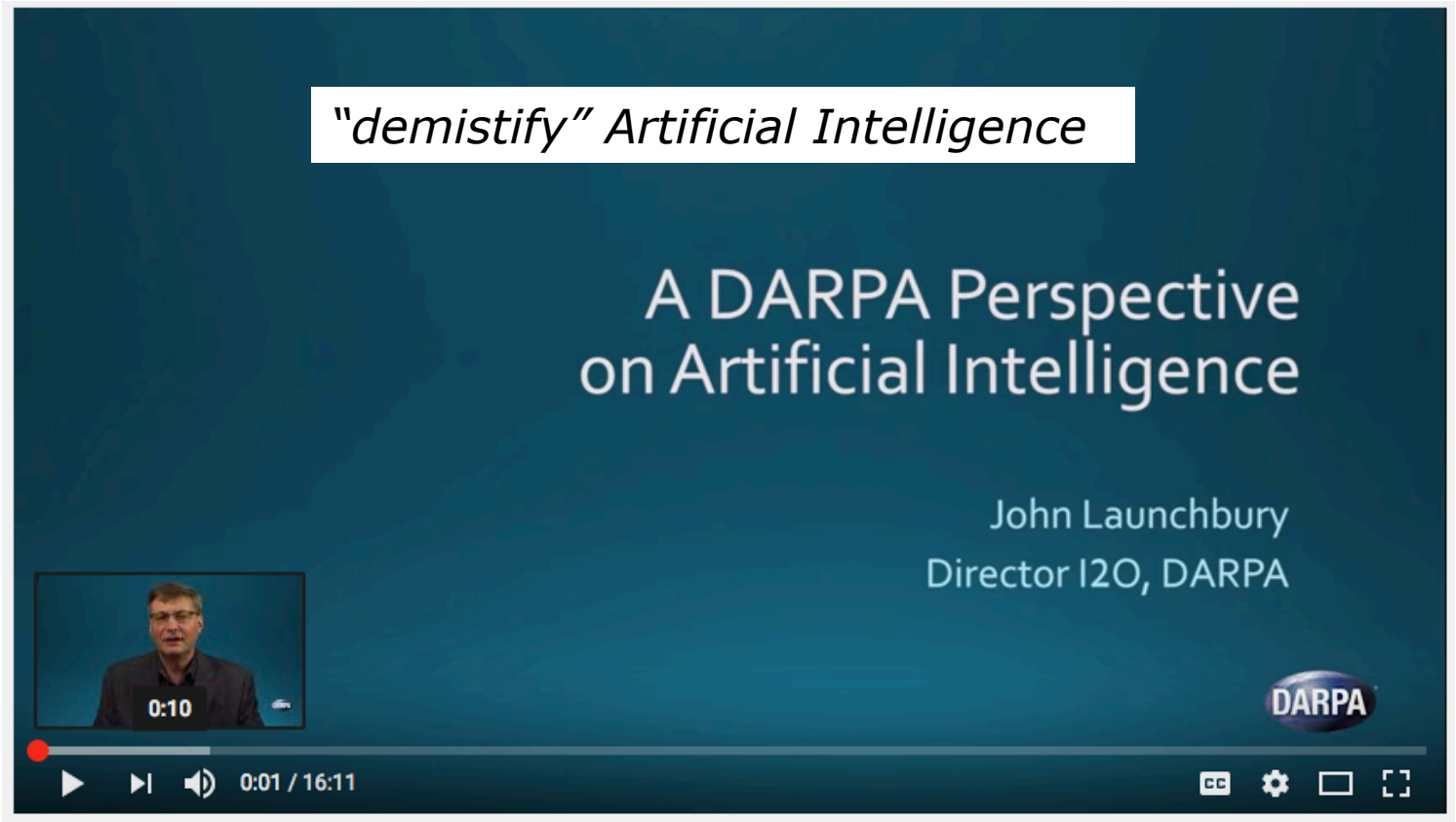
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



2. What is Artificial Intelligence?

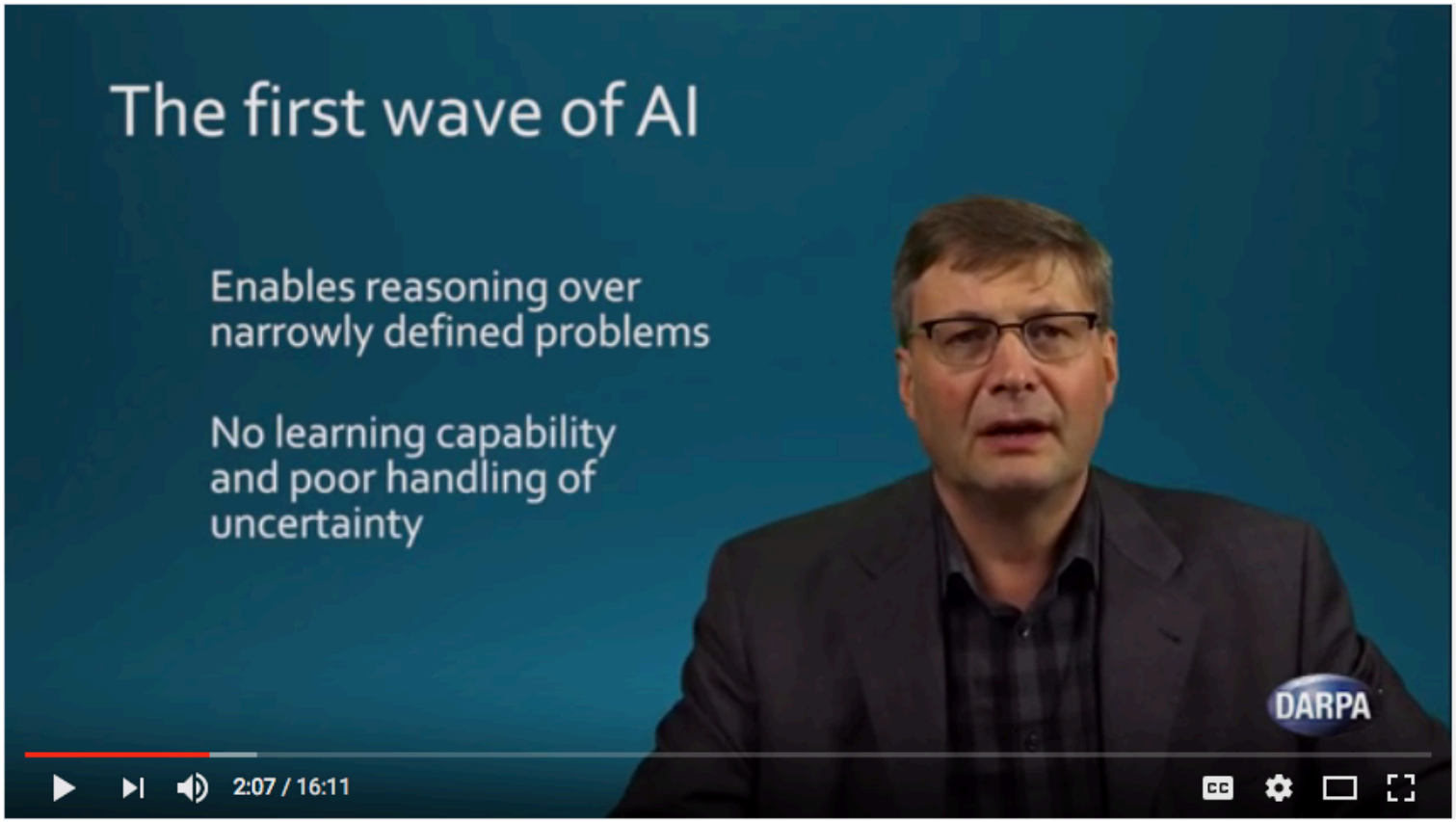
John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017



2. What is Artificial Intelligence?

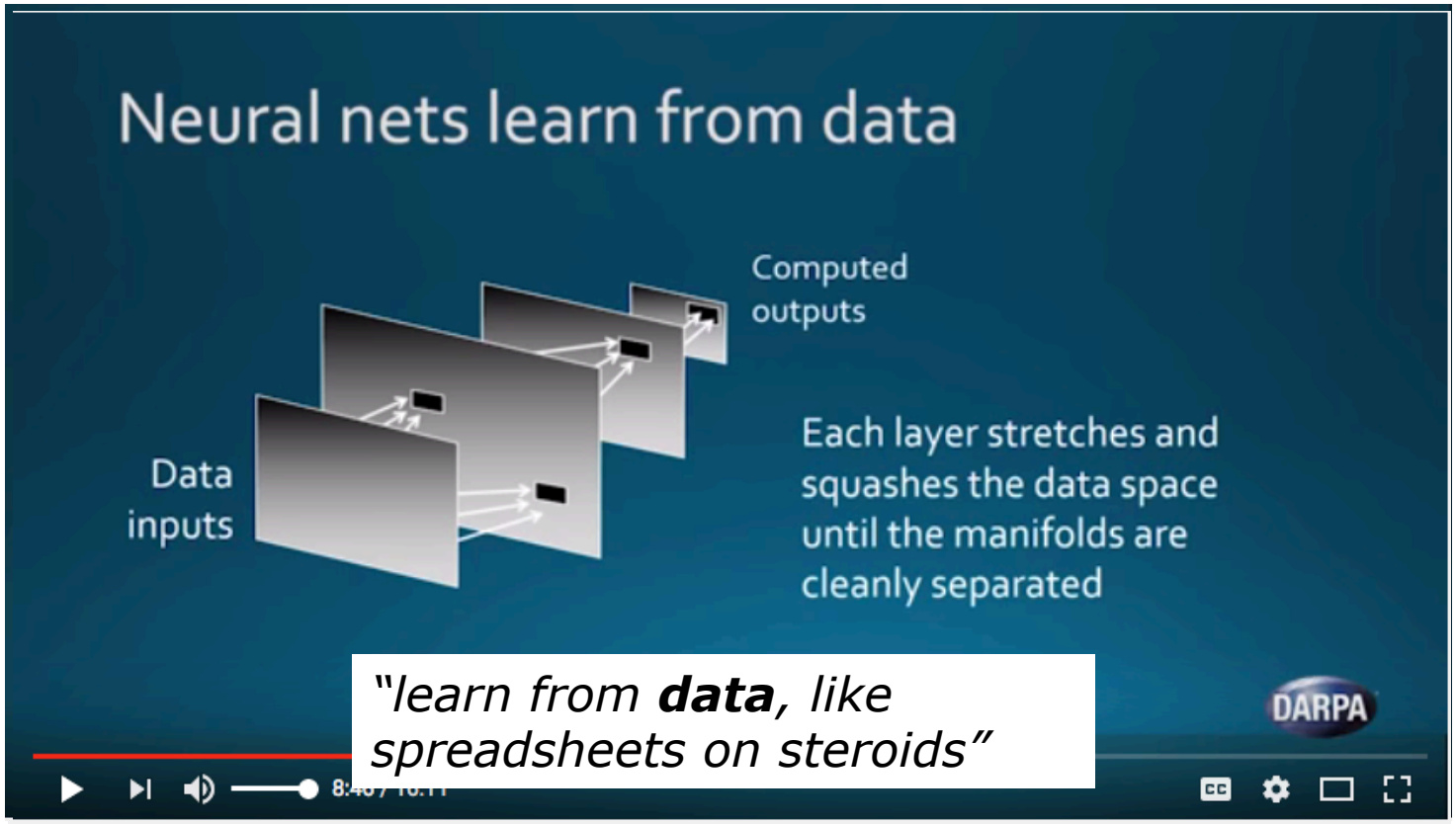
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

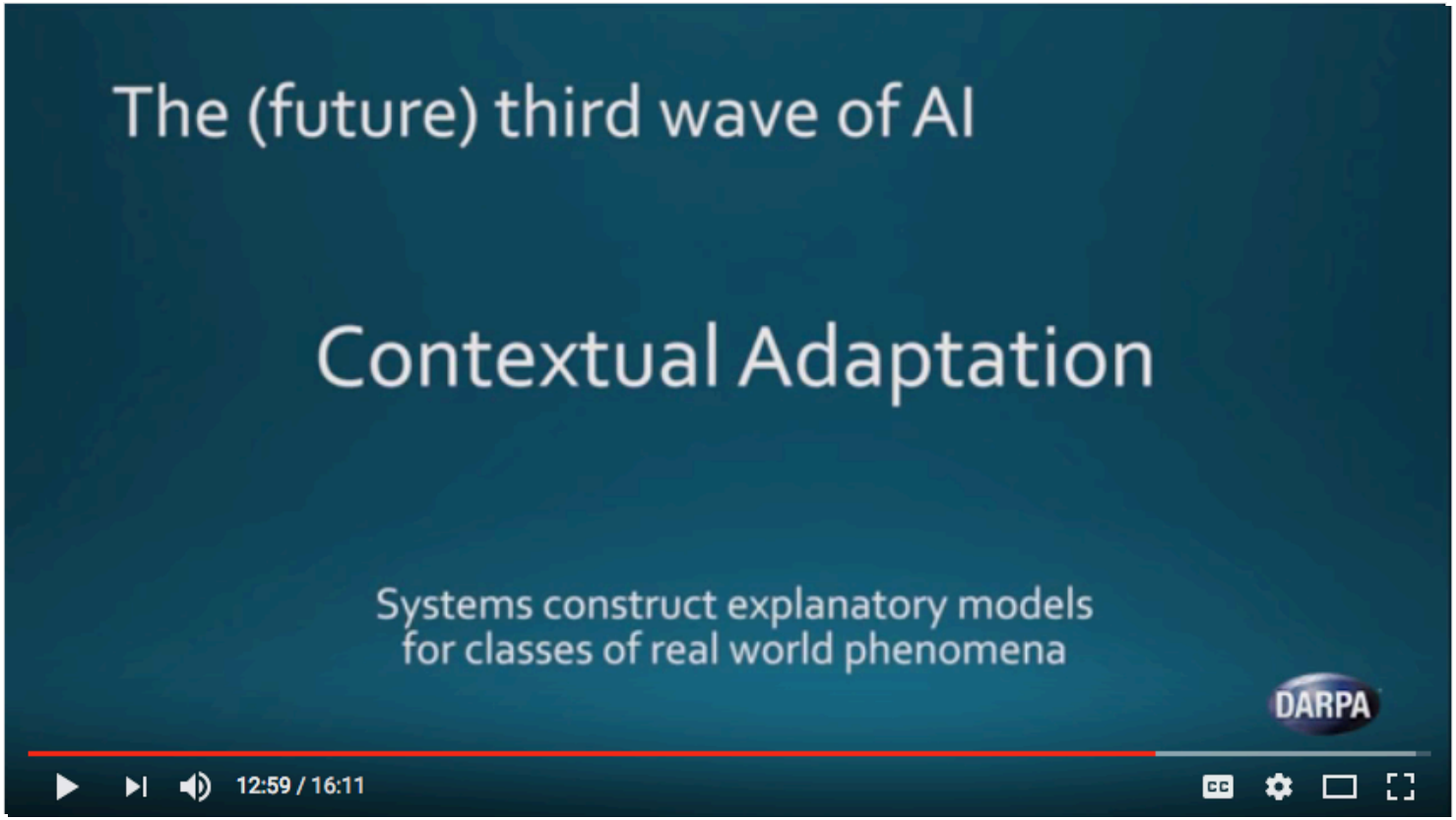


*"learn from **data**, like spreadsheets on steroids"*

2. What is Artificial Intelligence?

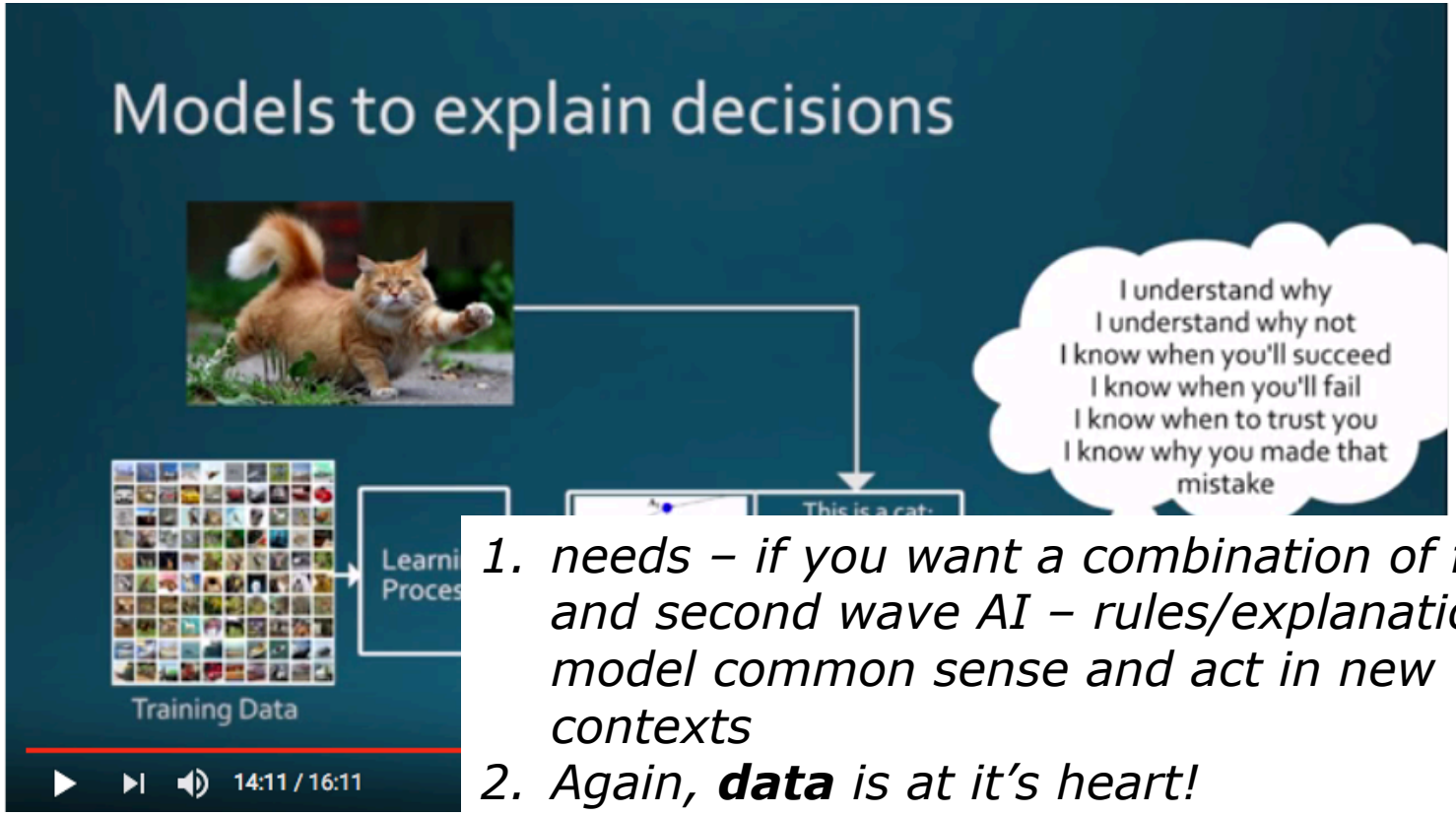
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



2. What is Artificial Intelligence?

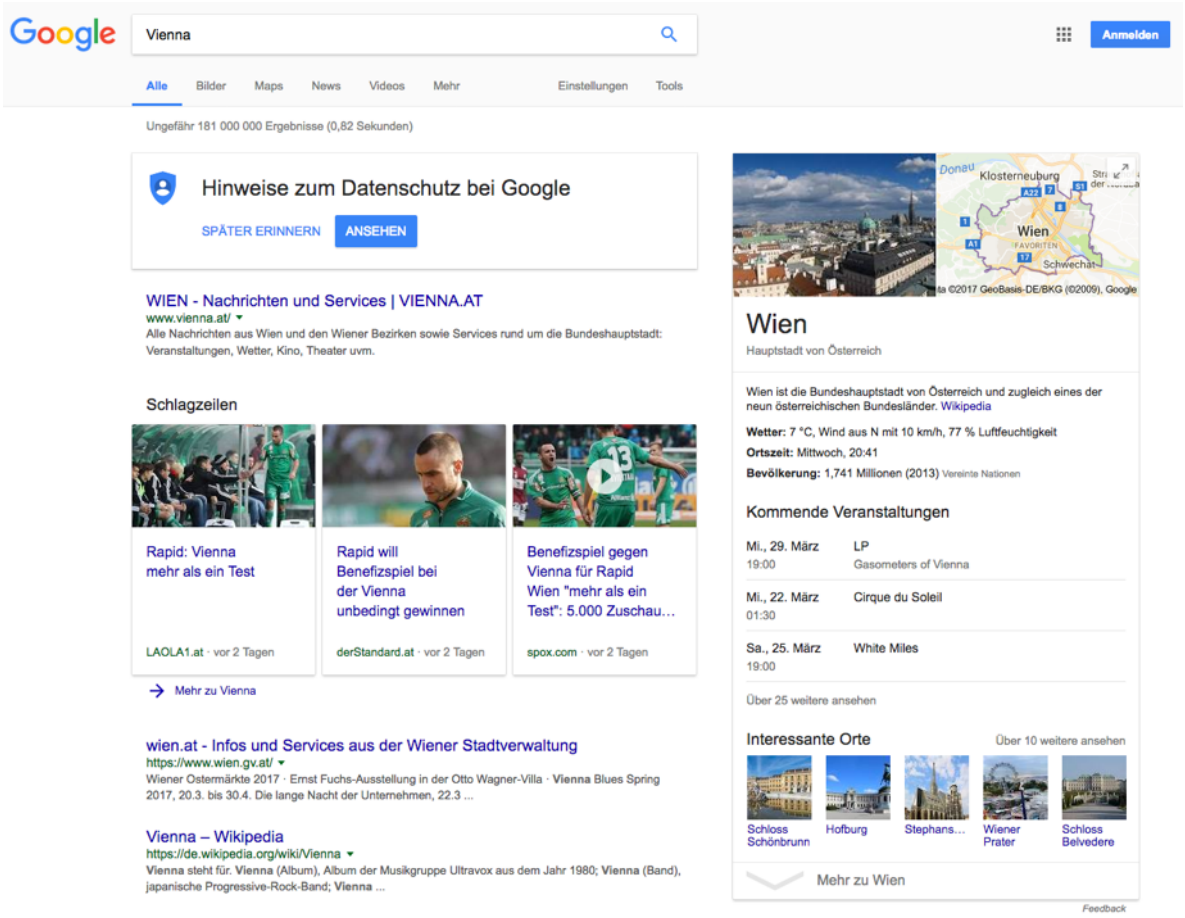
John Launchbury, the Director of DARPA's Information Innovation Office
Video, published on Feb 15, 2017



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!*

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

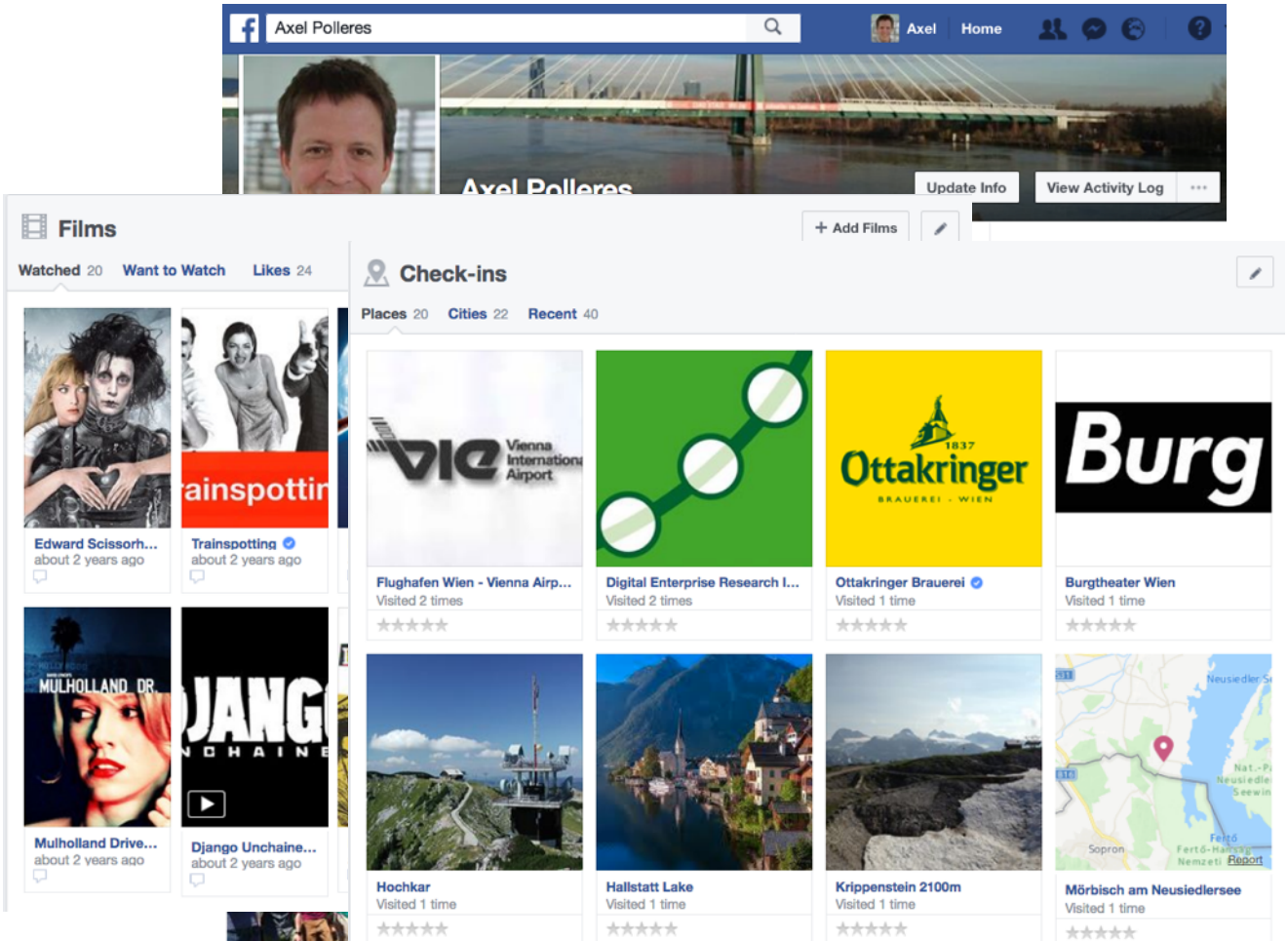
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



Example 1:
Google's
Knowledge Graph

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

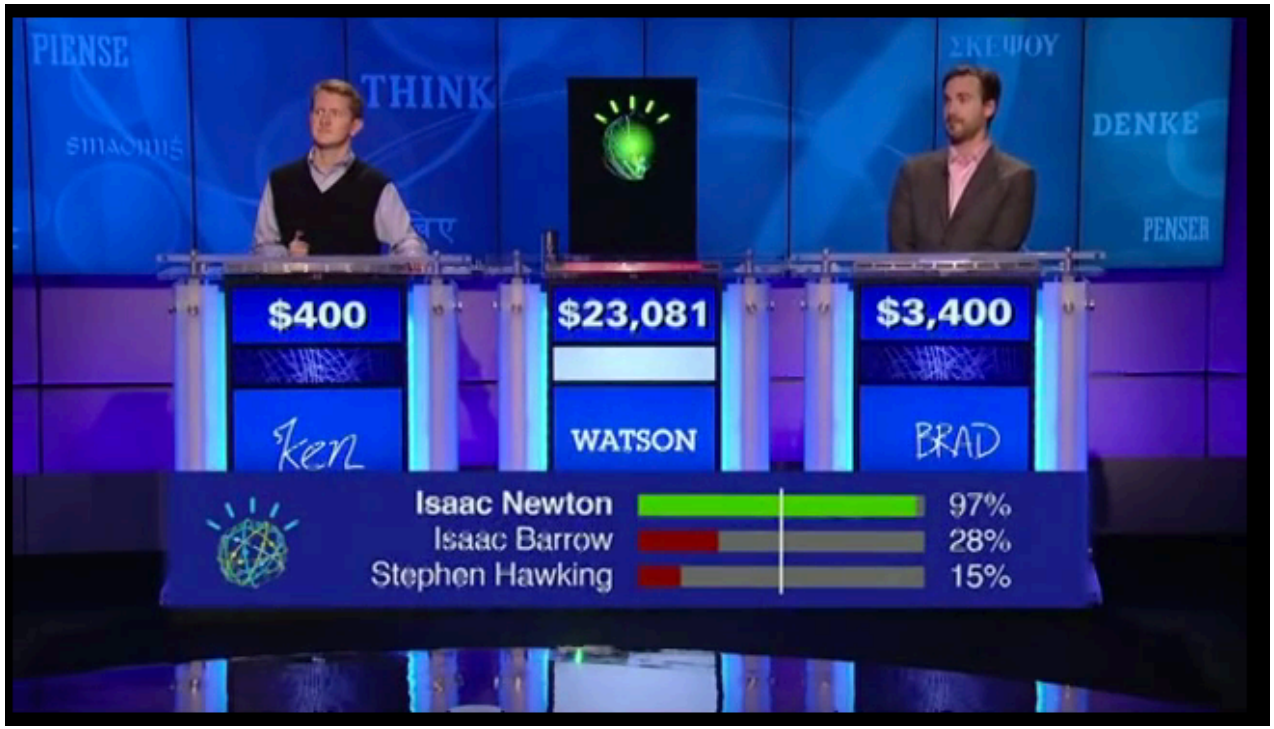
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



Example 2: FB's Social Graph & News Recommendations

Also uses a knowledge graph...

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



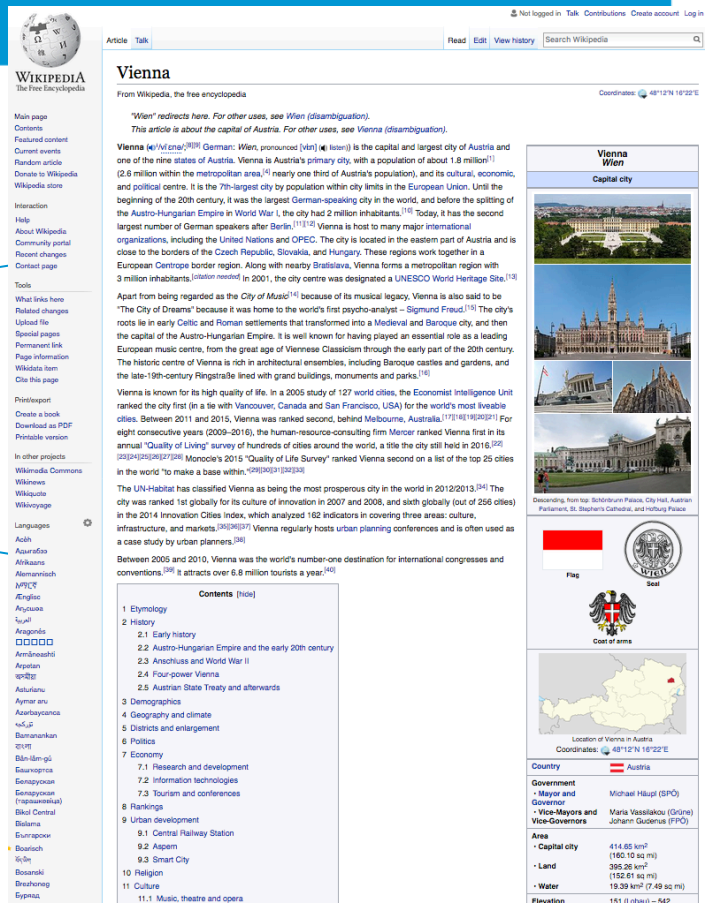
Example 3: *IBM* Watson!

Also uses a knowledge graph...

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

This is the knowledge Graph that IBM Watson used:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



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):

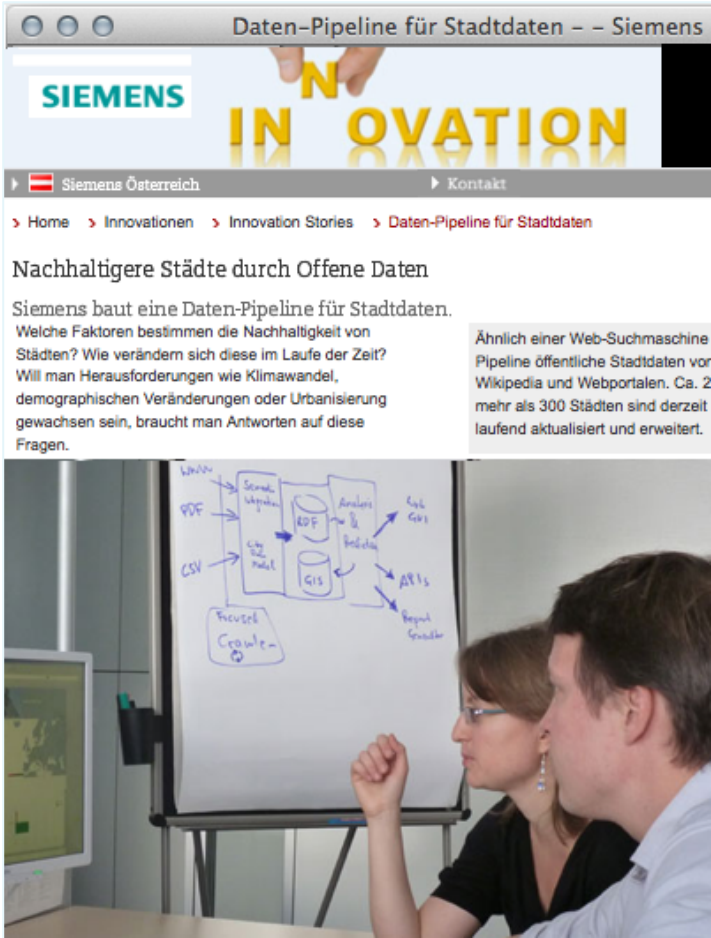
- Cities, International Organizations, National and European Portals, Int'l. Conferences:

Ok, now… how can I use it?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Recall my background:

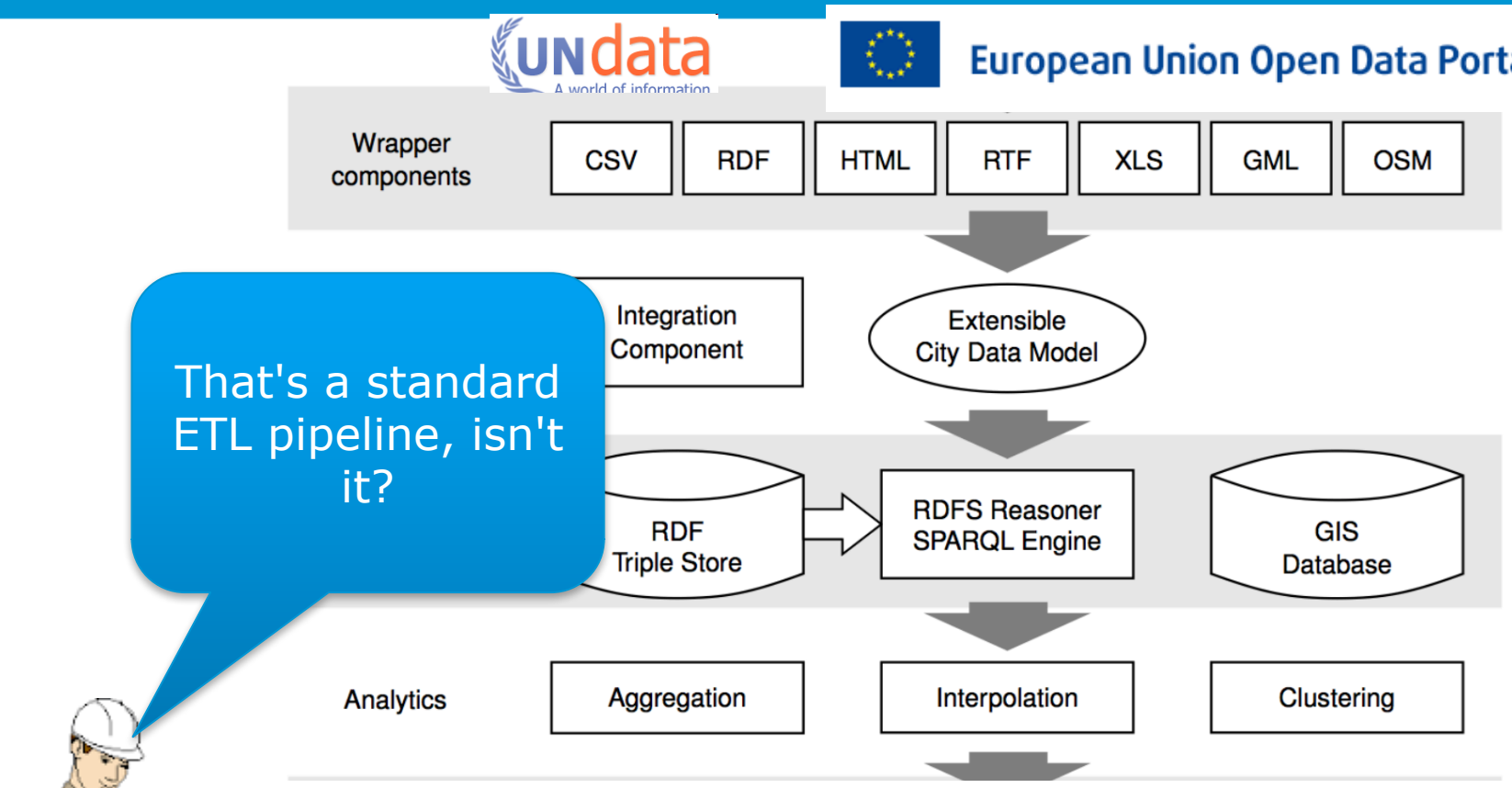
- Logic Programming
- Artificial Intelligence
- Knowledge Representation
- Semantic Web
- Web Data Integration



Attempt 1: use “first wave AI”

A concrete use case: The "City Data Pipeline"

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

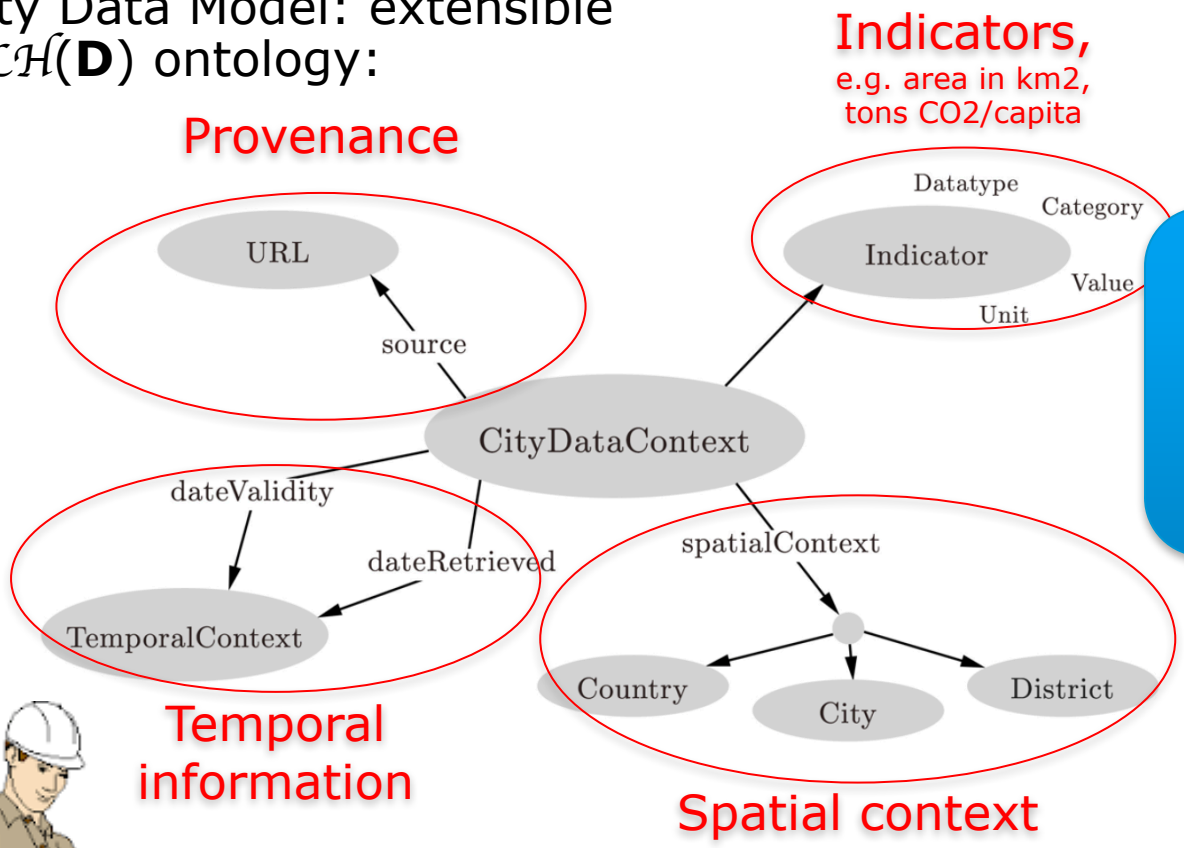


That's a standard ETL pipeline, isn't it?

A concrete use case: The "City Data Pipeline"

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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"

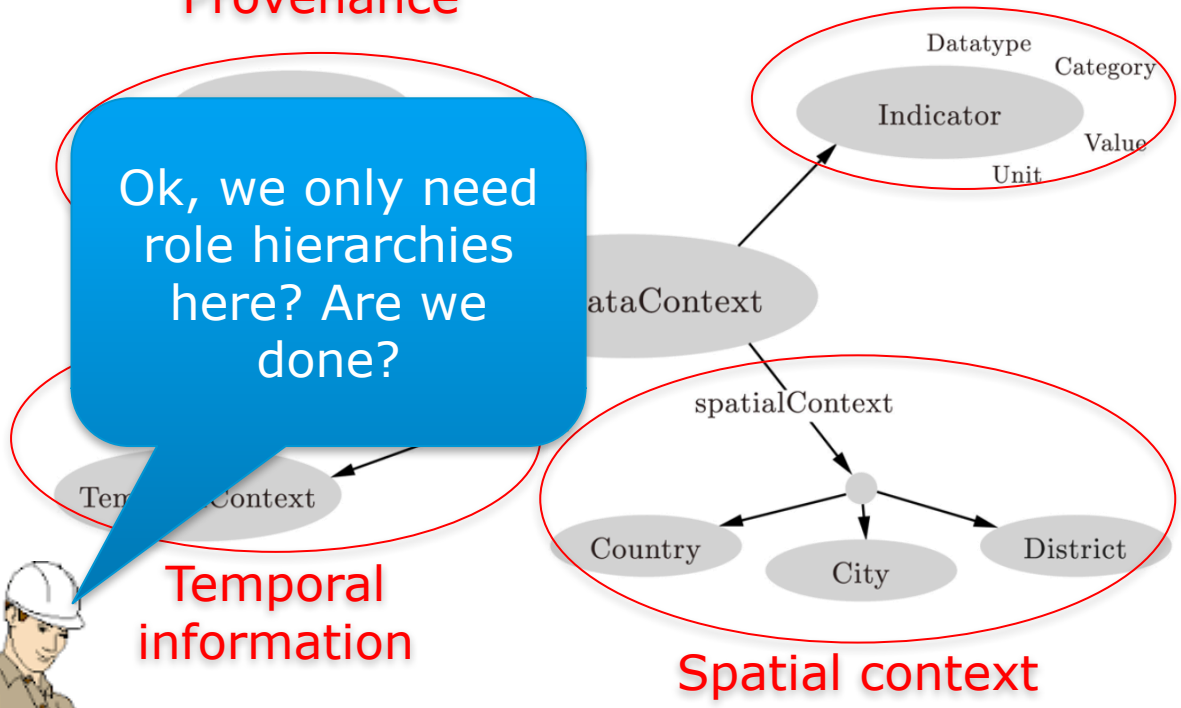
City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:

dbpedia:areakm $\sqsubseteq$:area

eurostat:area $\sqsubseteq$:area

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

Provenance



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

Temporal information

Spatial context

A concrete use case: The "City Data Pipeline"

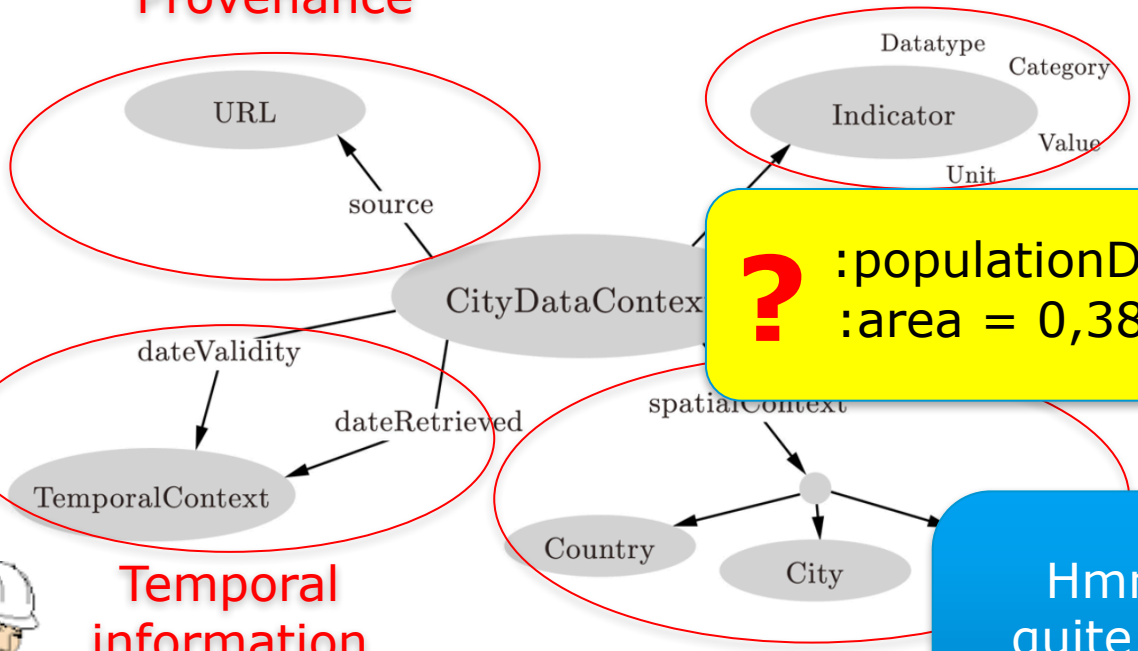
City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:

dbpedia:areakm2 $\sqsubseteq$:area

eurostat:area $\sqsubseteq$:area

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

Provenance



Temporal information

Spatial conte

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

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

Can equational knowledge co-exist with OWL?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

RDFS with Attribute Equations via SPARQL Rewriting

Stefan Bischof[1,2] and Axel Polleres[1]

[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

Stefan Bischof, Axel Polleres. ESWC2013

EFMD EQUIS ACCREDITED

Can equational knowledge co-exist with OWL?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- *Can equational knowledge co-exist with OWL?*
  - *We need a syntax & define a formal semantics*

- *Syntax:* :populationDensity = :population/:area
  :area = 0,386102 * 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:

:area = 0,386102 * dbpedia:areaMi2

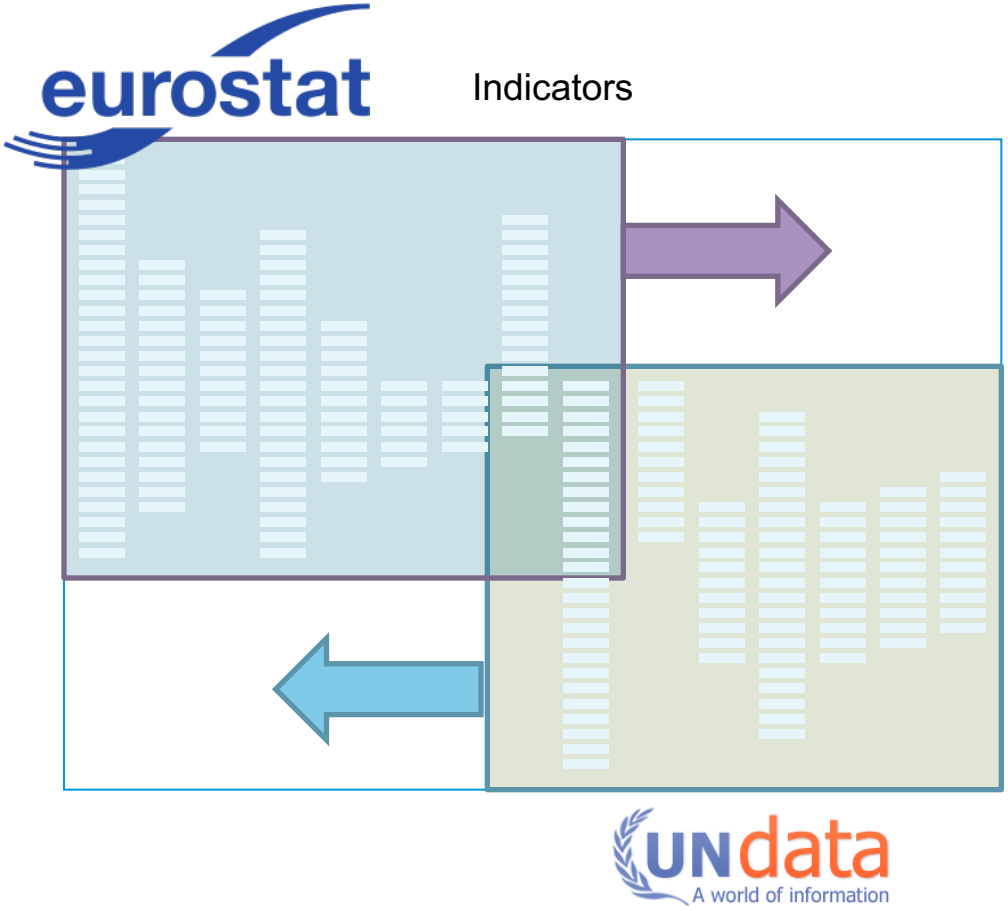
:areaMi2 = 2,589988 * :area

EFMD EQUIS ACCREDITED

Challenges – Too many Missing values

Goal: euqational 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

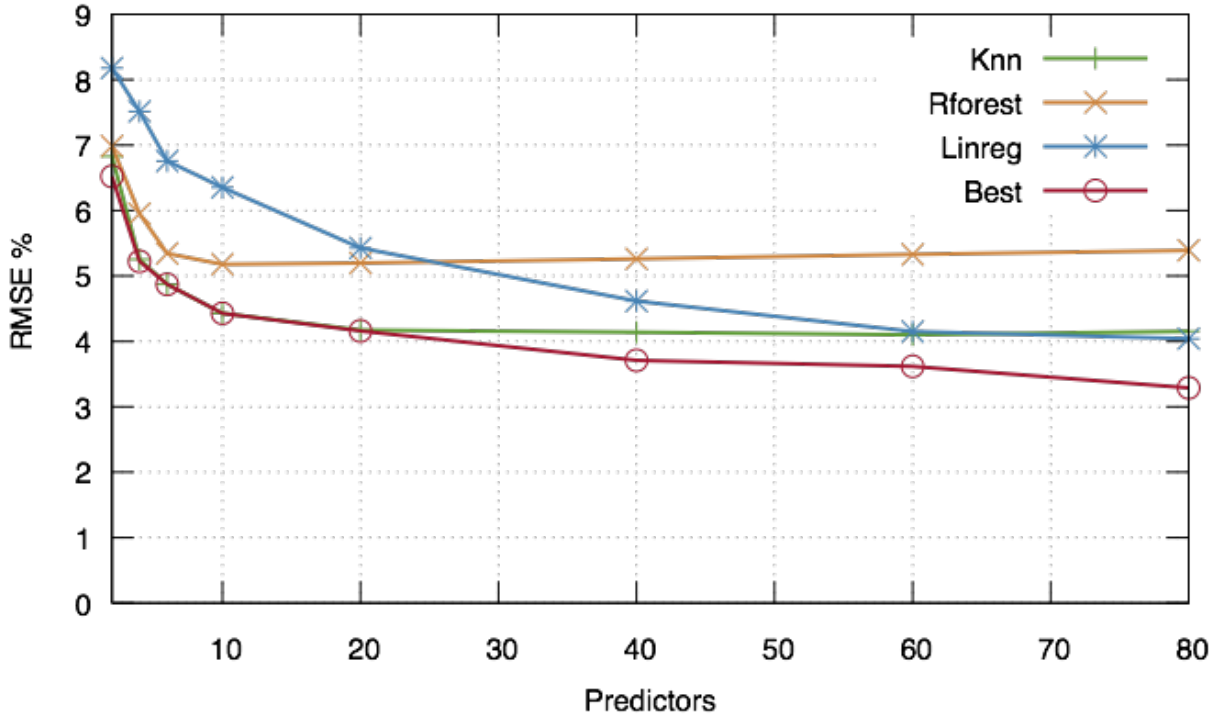
	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858			
Nationals	1342704	216289	203657			
Women per 1000 Men	109.8	108.7	101.9			
Available Beds per 1000				138.3	14969	166.1
Average area of living				36.42	37.24	38.16
Cinema Seats				4691	12751	2676

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)
  -
  -

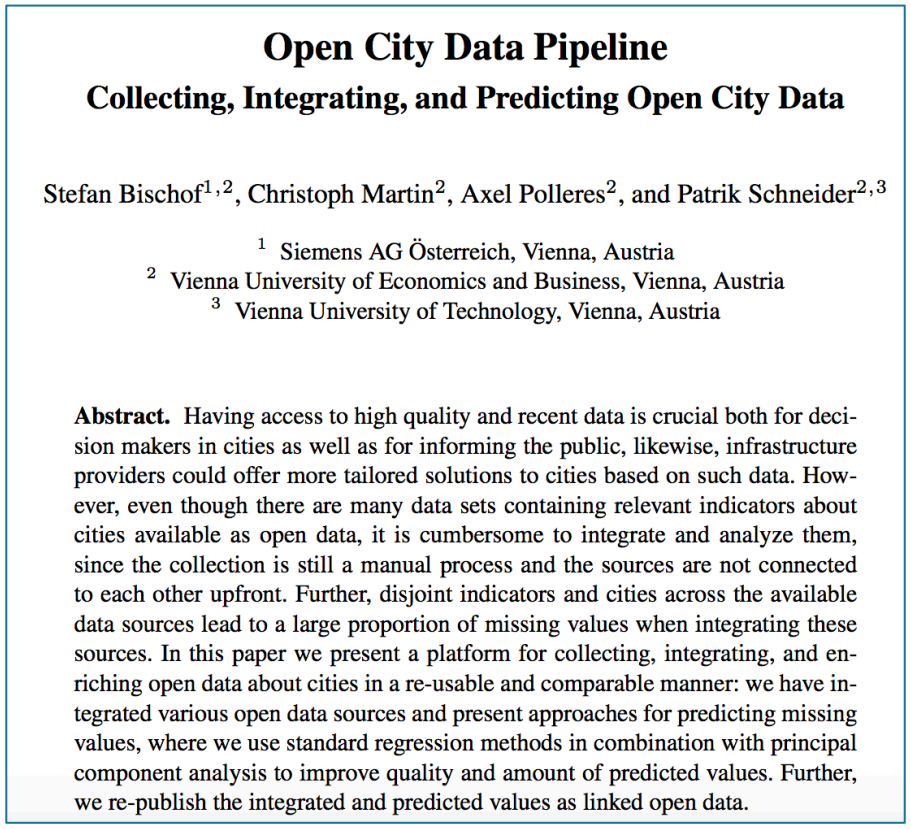
Missing Values – Hybrid approach choose best prediction method per indicator:

- Instead of using indicators directly we use Principle Components, built from the indicators
- For buidling 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 Martin[2], Axel Polleres[2], 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:

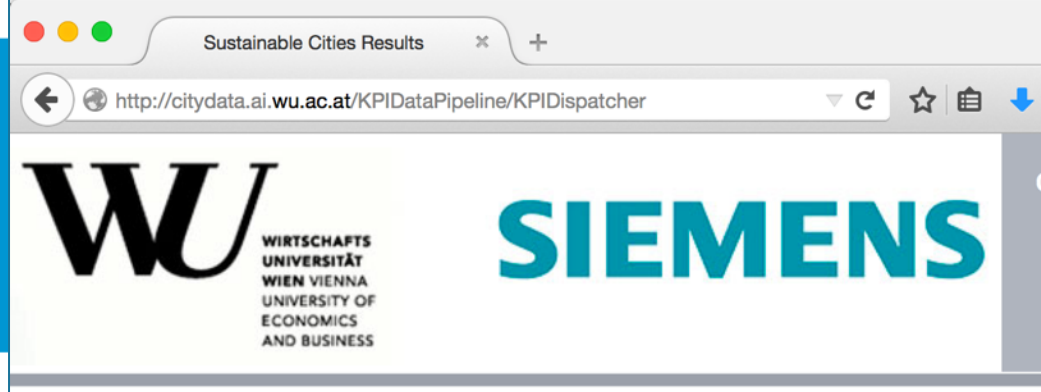
Combine ML and equations "iteratively" (under submission)

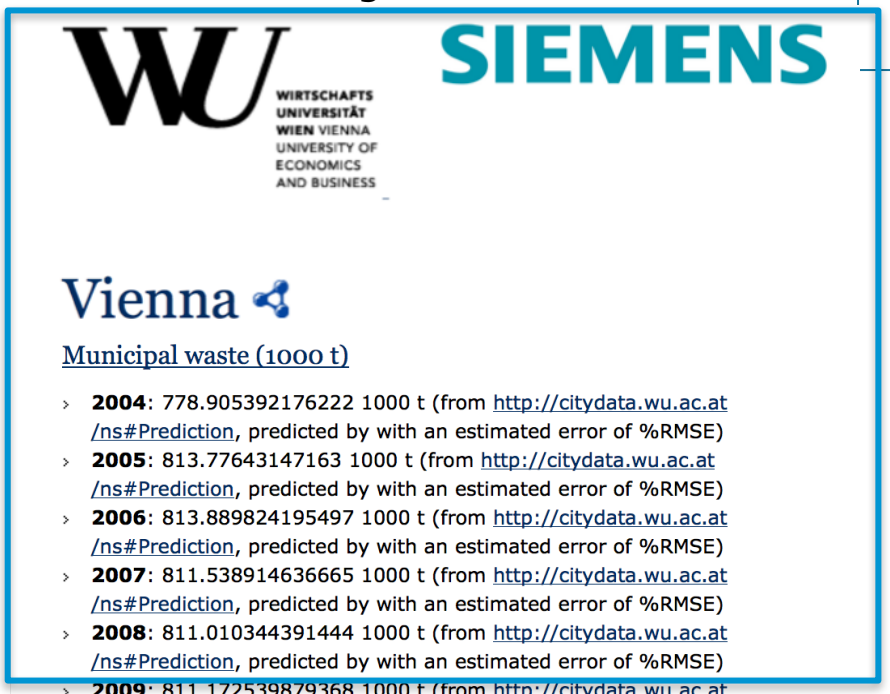
http://epub.wu.ac.at/5438/

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...





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

However:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

(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?

EFMD EQUIS ACCREDITED

Open Data Portals

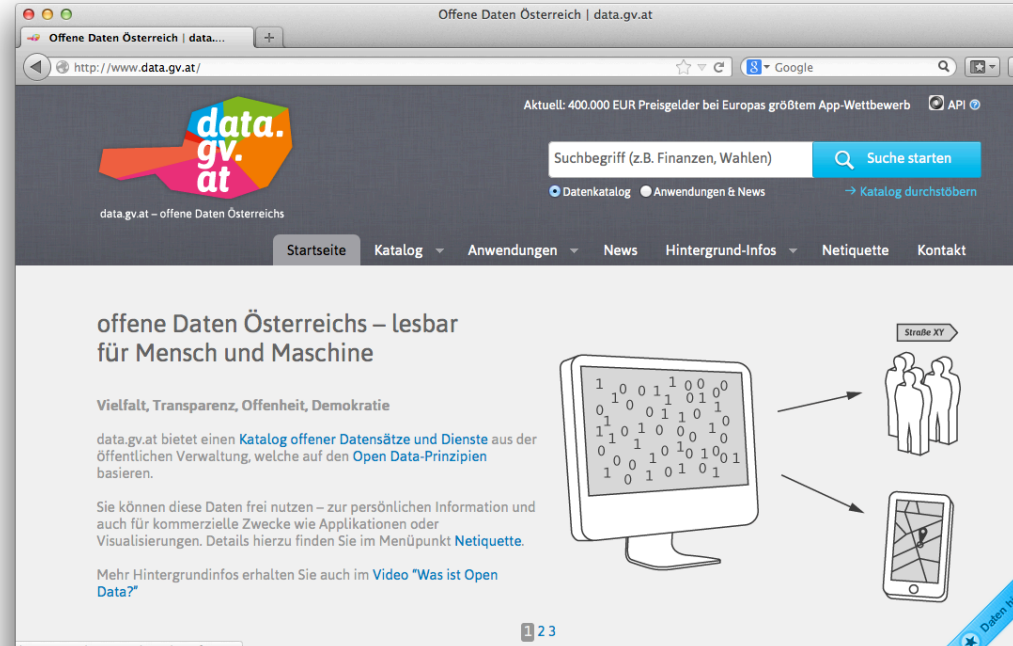
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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**



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 Congitive Intelligence based on Open Data?

Ongoing Projects (data.wu.ac.at)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

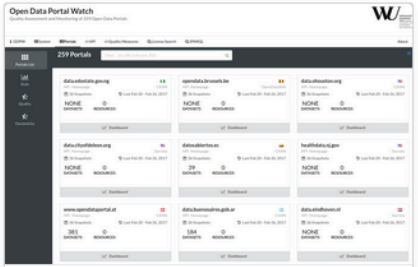
Projects

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

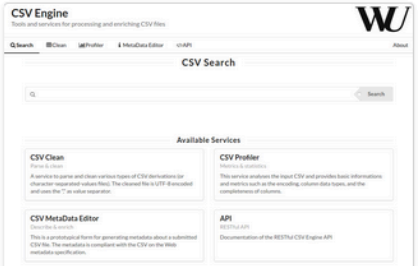


Open Data Portal Watch

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

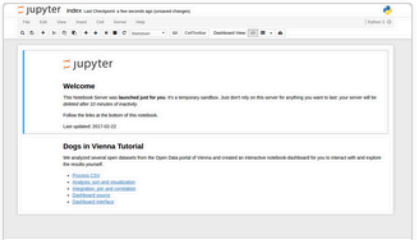
Search & enrich CSVs

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.

EFMD EQUIS ACCREDITED

Open Data Portal Watch

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

EFMD EQUIS ACCREDITED

Portalwatch Example:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

A

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

EFMD EQUIS ACCREDITED

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**!

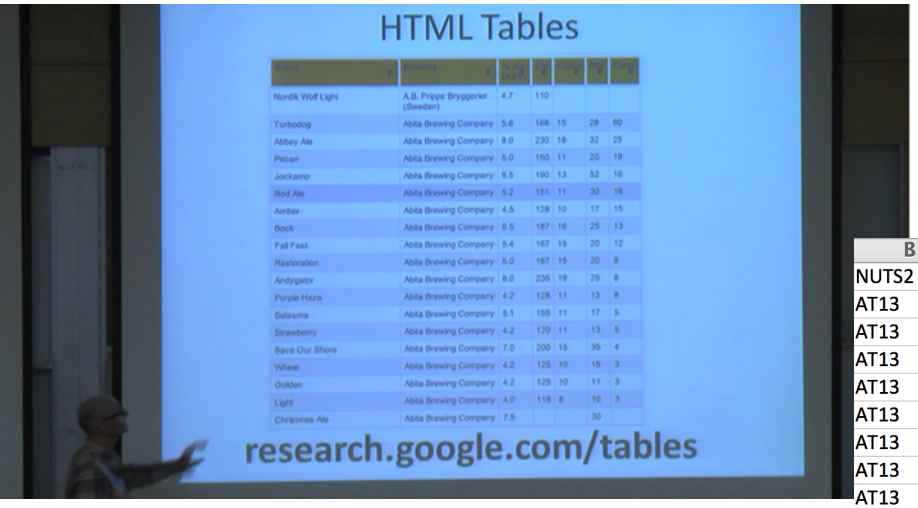
- A striving **Data Economy** needs no silos… re-democratise the Web by Congitive Intelligence based on Open Data?

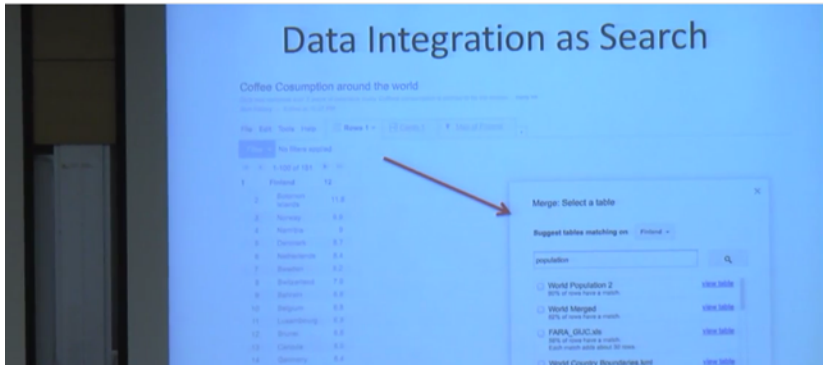
Why is Search in Open Data a problem?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

https://www.youtube.com/watch?v=kCAymmbYIvc

Structured Data in Web Search by Alon Halevy





vs.

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"

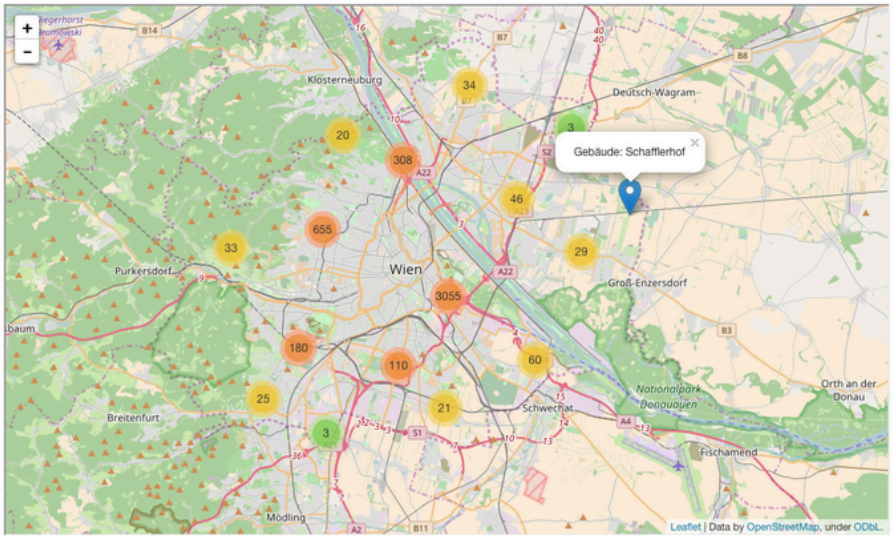
c) Open Data is not properly geo-referenced

Some starting points:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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 Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[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.

EFMD EQUIS ACCREDITED

Towards linking Open Data to a Knowledge Graph

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

International Semantic Web conference 2016:

Multi-level semantic labelling of numerical values

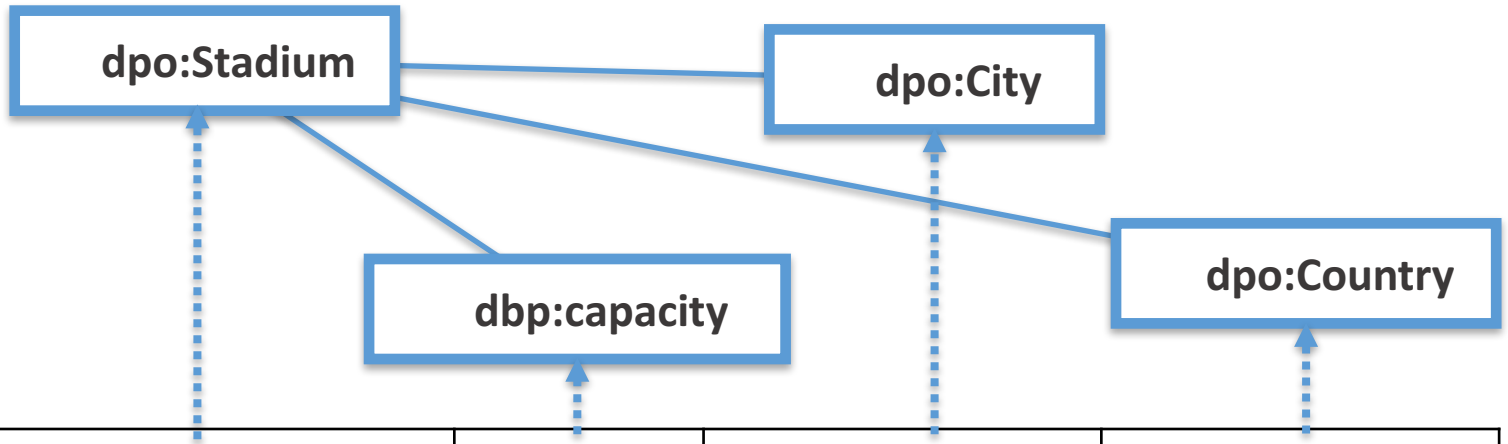
Sebastian Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[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.

EFMD EQUIS ACCREDITED

Example



stadium name	*capacity*	*city*	*country*
Emirates Stadium	60361	London	England
Villa Park	42785	Birmingham	England
Ewood Park	31154	Blackburn	England
…	…	…	…

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

Wohnungen in den 250 Zaehlbezirken in Wien - Registerzaehlung 2011 | Housing units in 250 su					
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)

stadium	***capacity***	***city***	***country***
Emirates Stadium	60361	London	England
Villa Park	42785	Birmingham	England
Ewood Park	31154	Blackburn	England
…	…	…	…

Example (Cont'd)

	TOTAL	*DISTRICT_CODE*	*ISO_2*
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

	TOTAL	*DISTRICT_CODE*	*ISO_2*
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
…	…	…	…

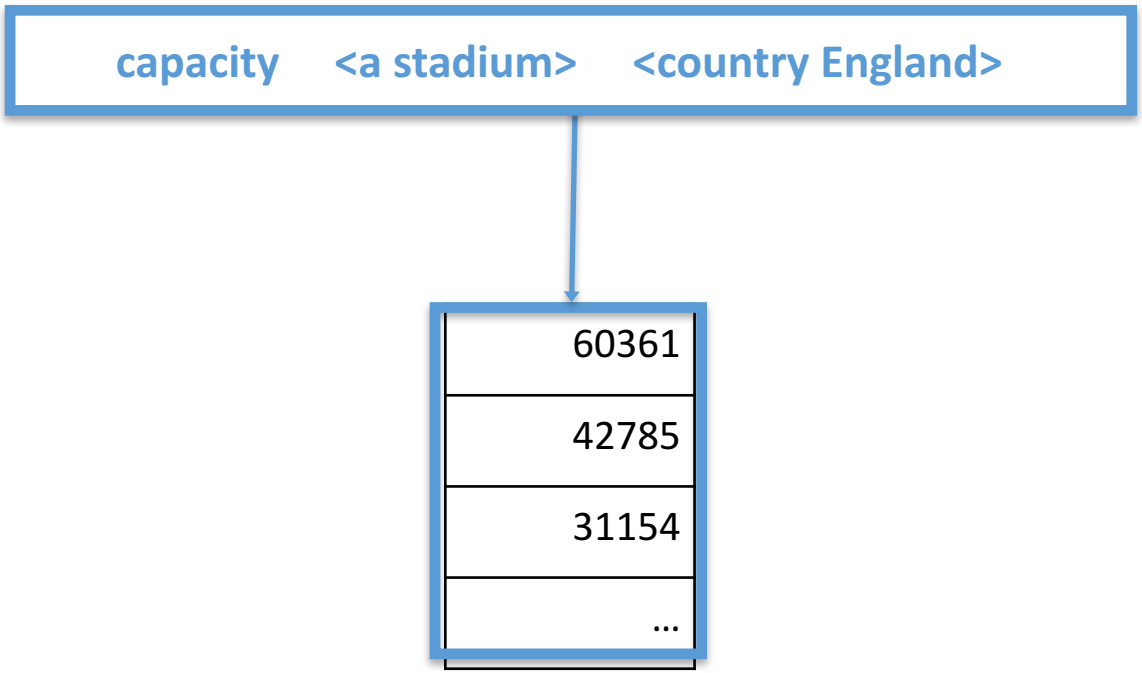
Why not use numeric values?

- 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

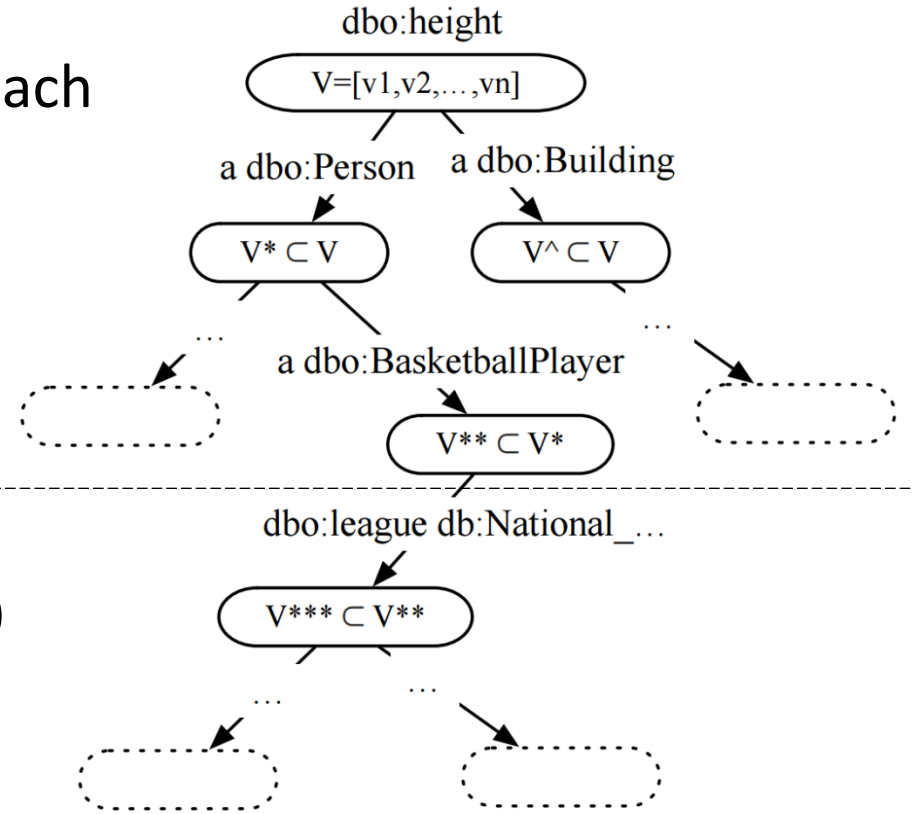


Our Approach

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

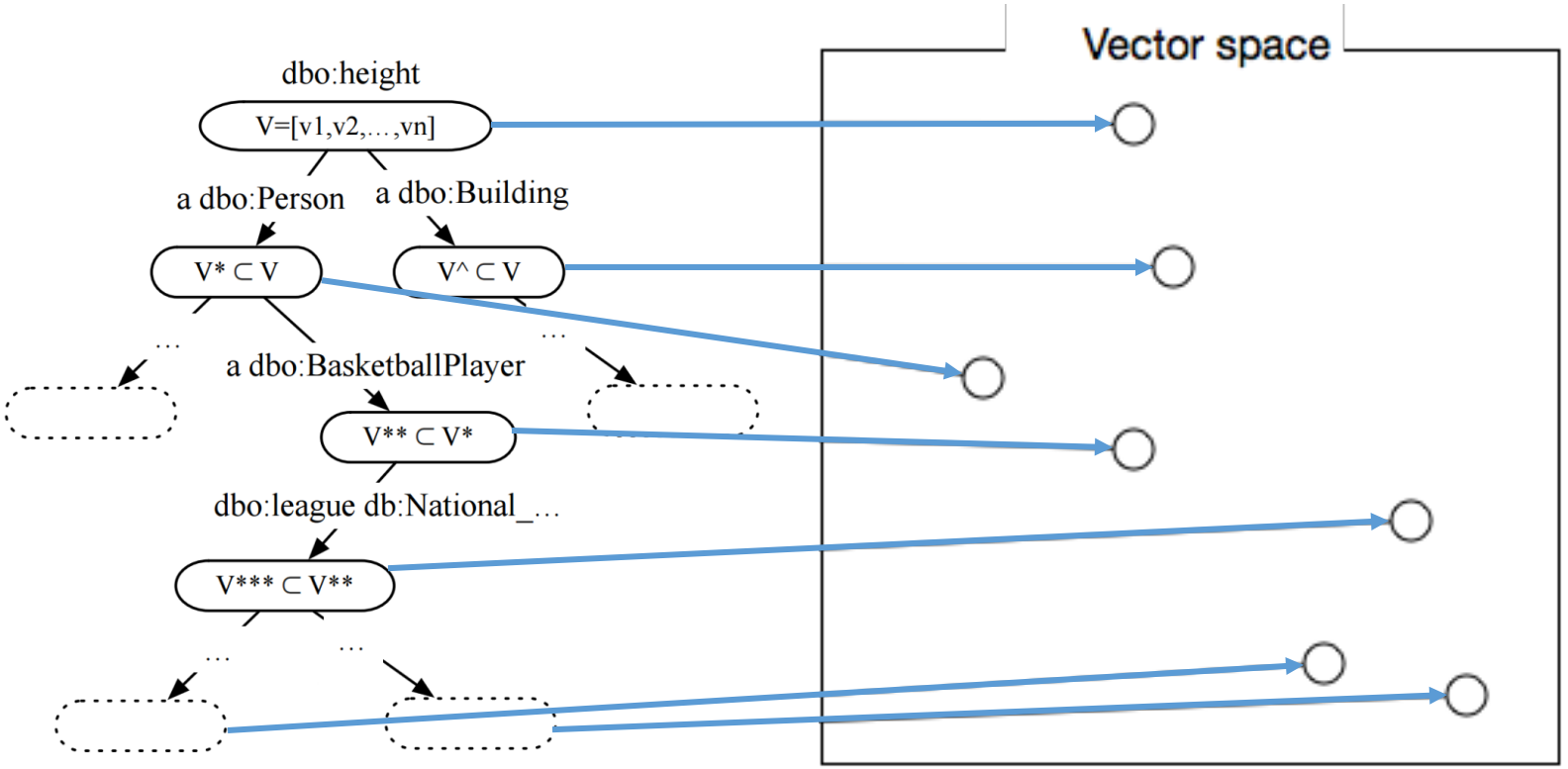
1. Background Knowledge Graph

- 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)



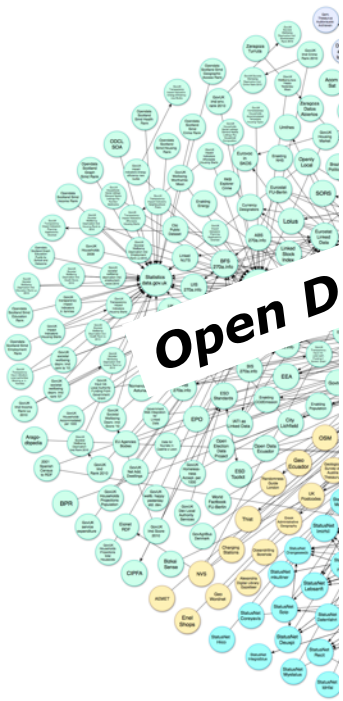
Our research:
data.wu.ac.at

- ***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 Congitive Intelligence based on Open Data?

This is t
that IBM



Quora

Search for questions, people, and topics

Quora uses cookies to improve your experience. Read more

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

"*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!

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.

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

Linking Op … n P. McCrae, http://l…

This is a fundamental threat to the Web itself:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

20,966 520

Tim Berners-Lee

Sunday 12 March 2017 00.01 GMT

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

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

Open-Data-Fueled Complexity Science to the rescue?

EFMD EQUIS ACCREDITED

Still Open Questions (with some starting points presented…)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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?

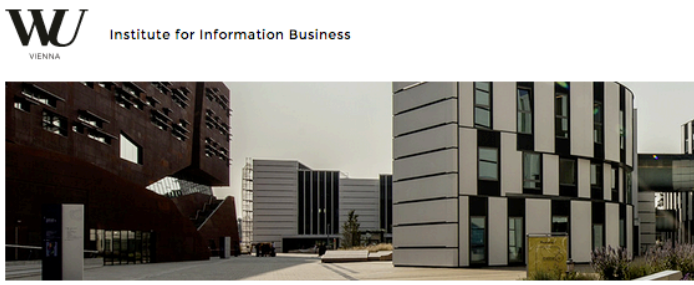
- What is the right form of **Knowledge Representation** for Knowledge graphs?
  - OWL, Rules, Equations, Property-domain pairs?)
  - How to represent models in an exchangeable manner?

- 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:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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

https://www.wu.ac.at/en/infobiz/