https://www.mpi-inf.mpg.de/www-2022-tutorial

Completeness, Recall, and Negation in Open-World Knowledge Bases




Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

max planck institut informatik

TELECOM ParisTech

On the Limits of Machine Knowledge: Completeness, Recall and Negation in Web-scale Knowledge Bases

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

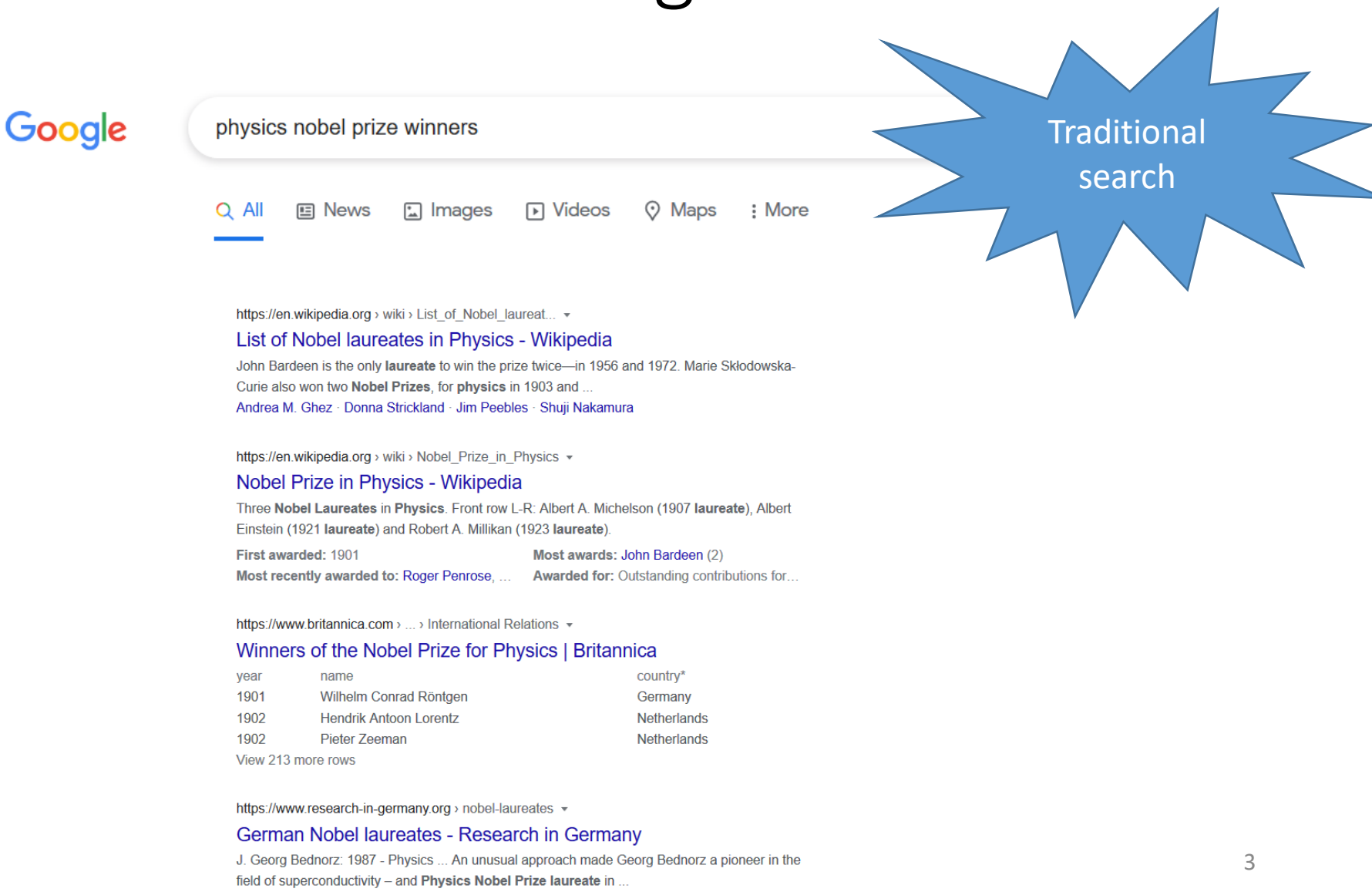
1. Introduction & Foundations (Simon) – 15:45-16:00 CEST
2. Predictive recall assessment (Fabian) – 16:00-16:20
3. Counts from text and KB (Shrestha) – 16:20-16:40
4. Negation (Hiba) – 16:40-17:00
5. Wrap-up (Simon) – 17:00-17:10

THE WEB CONFERENCE ACM

max planck institut informatik

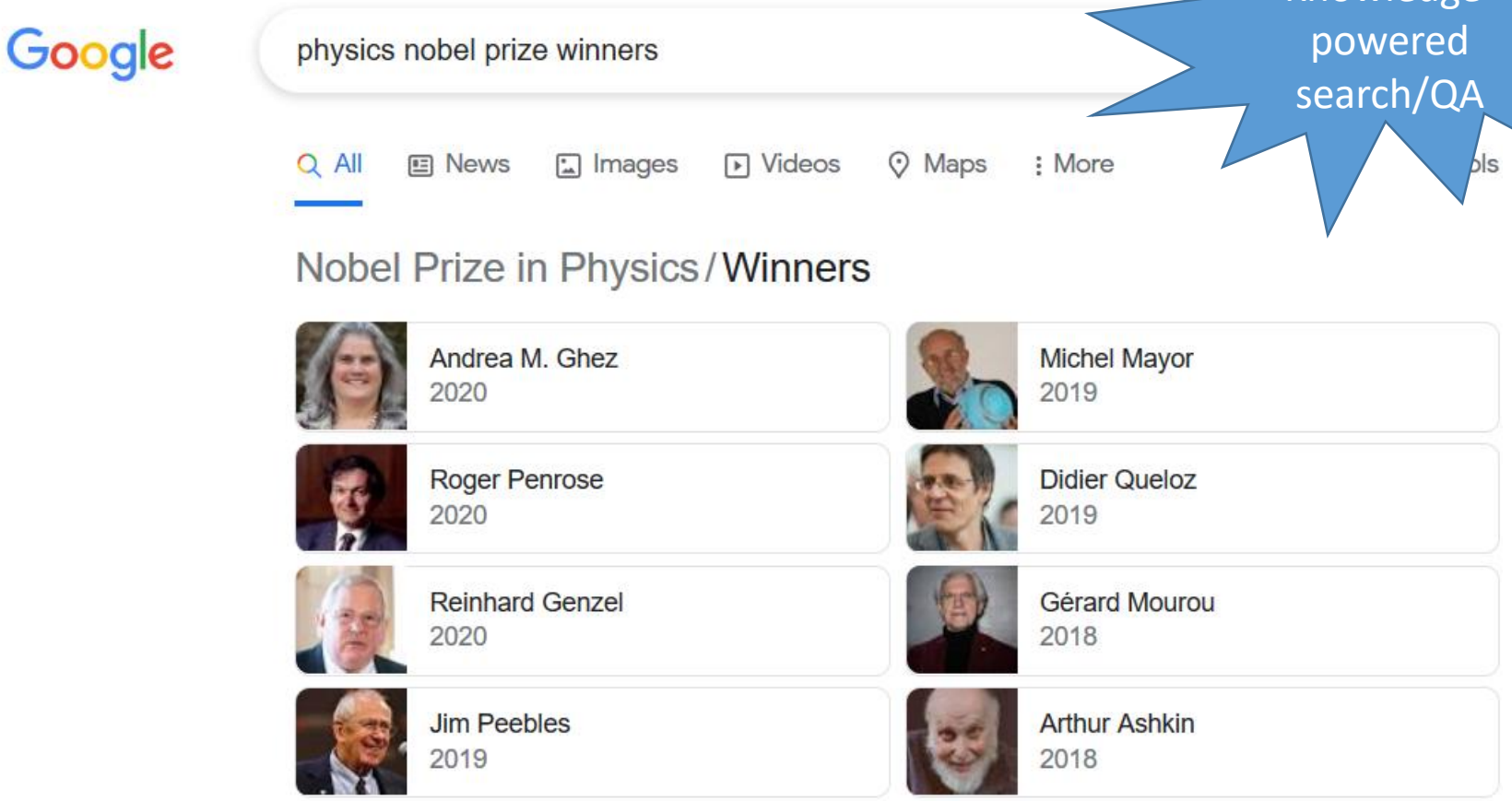
TELECOM ParisTech

Machine knowledge in action

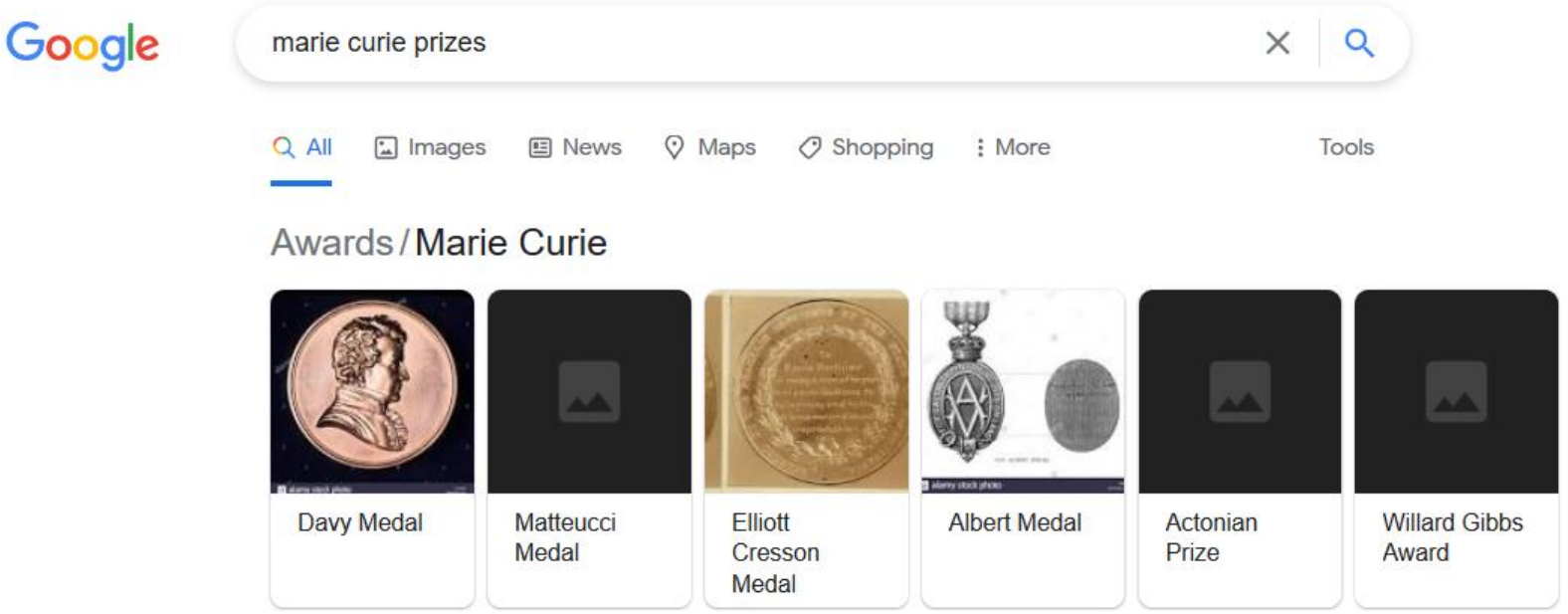


Machine knowledge in action

Knowledge-powered search/QA



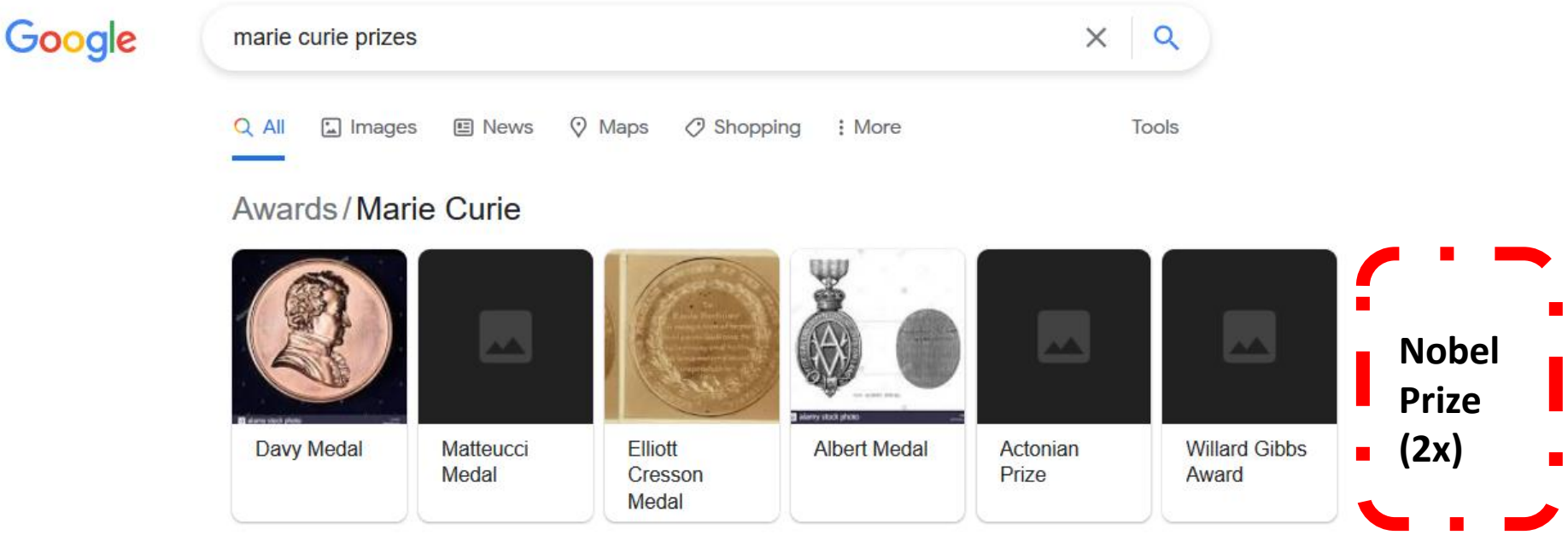
Machine knowledge in action



Machine knowledge is awesome

- Reusable, scrutable asset for knowledge-centric tasks
  - Semantic search & QA
  - Entity-centric text analytics
  - Distant supervision for ML
  - Data cleaning
- Impactful projects at major public and commercial players
  - Wikidata, Google KG, Microsoft Satori, …
- Strongly rooted in semantic web community
  - Linked data, vocabularies, ontologies, indexing and querying, …

But:
Machine Knowledge is incomplete



Machine knowledge is incomplete (2)

Wikidata KB:

Semantic Web Journal has only published 84 articles ever

- https://scholia.toolforge.org/venue/Q15817015

Only 7 papers ever deal with the topic “web science”

- https://scholia.toolforge.org/topic/Q579439

But: Machine knowledge is one-sided

- In KB:
  - *Nicola Tesla received title of IEEE fellow*
  - *Vietnam is a member of ASEAN*
  - *iPhone has 12MP camera*

- Not in KB:
  - *Nicola Tesla did not receive the Nobel Prize*
  - *Switzerland is not a member of the EU*
  - *iPhone 12 has no headphone jack*

Why is this problematic? (1) Querying

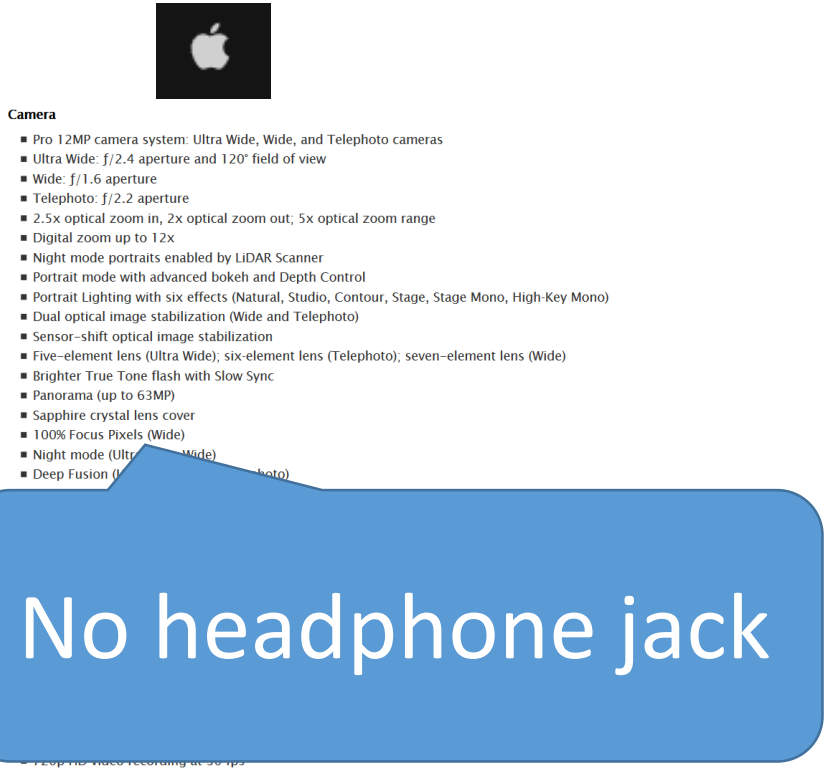
- Decision making more and more data-driven
- Analytical queries paint wrong picture of reality
  - *E.g., SW journal deemed too small*
- Instance queries return wrong results
  - *E.g., wrongly assuming certain topic is of no interest*

Why is this problematic? (2) Data curation

- Effort priorization fundamental challenge in human-in-the-loop curation
  - *Should we spend effort on obtaining data for SWJ or TWeb?*

- Risk of effort duplication if not keeping track of completed areas
  - *Spending effort on collecting data … already present*

Why is this problematic? (3) Summarization and decision making



Topic of this tutorial

How to know how much a KB knows?

How to = techniques
How much knows = completeness/recall/coverage bookkeeping/estimation
KB = General world knowledge repository

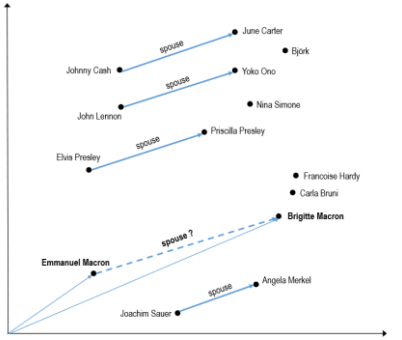
What this tutorial offers

- Logical foundations
  - Setting and formalisms for describing KB completeness (part 1)
- Predictive assessment
  - How (in-)completeness can be statistically predicted (Part 2)
- Count information
  - How count information enables (in-)completeness assessment (Part 3)
- Negation
  - How salient negations can be derived from incomplete KBs (Part 4)

Goals:

1. Systematize the topic and its facets
2. Lay out assumptions, strengths and limitations of approaches
3. Provide a practical toolsuite

What this tutorial is NOT about



TransE - a KBC model

- Knowledge base completion (KBC)
  - "How to make KBs more complete"
- Related: Understanding of completeness is needed to know when/when not to employ KBC
  - KBC naively is open-ended

  → Understanding of completeness needed to "stop"
- But:
  - Heuristic, error-prone KBC not always desired
  - Completeness awareness != actionable completion
- Literature on knowledge graph completion, link prediction, missing value imputation, etc.
  - E.g., Rossi, Andrea, et al. Knowledge graph embedding for link prediction: A comparative analysis *TKDD 2021*

Beatles members:

John Lennon	36%
Paul McCartney	23%
George Harrison	18%
Bob Dylan	5%
Ringo Starr	3%
Elvis Presley	2%
Yoko Ono	2%

On the Limits of Machine Knowledge: Completeness, Recall and Negation in Web-scale Knowledge Bases

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

1. Introduction & Foundations (Simon) – 15:45-16:00 CEST
2. Predictive recall assessment (Fabian) – 16:00-16:20
3. Counts from text and KB (Shrestha) – 16:20-16:40
4. Negation (Hiba) – 16:40-17:00
5. Wrap-up (Simon) – 17:00-17:10

THE WEB CONFERENCE ACM

max planck institut informatik

TELECOM ParisTech

Knowledge base - definition

Given set **E** (entities), **L** (literals), **P** (predicates)

- Predicates are positive or negated properties
  - *bornIn, notWonAward, …*

- An assertion is a triple (s, p, o) $\in$ **E** $\times$ **P** $\times$ (**E**$\cup$**L**)
- A practically available KB $\mathbf{K}^{a}$ is a set of assertions
- The ``ideal'' (complete) KB is called $\mathbf{K}^{i}$
- Available KBs are incomplete: $\mathbf{K}^{a} \subseteq \mathbf{K}^{i}$

Knowledge bases (KBs aka. KGs)

subject-predicate-object triples about entities, attributes of and relations between entities

+ composite objects

predicate (subject, object)

---

type (Marie Curie, physicist)
subtypeOf (physicist, scientist)

taxonomic knowledge

placeOfBirth (Marie Curie, Warsaw)
residence (Marie Curie, Paris)
¬placeOfBirth (Marie Curie, France)

factual knowledge

discovery (Polonium, 12345)
discoveryDate (12345, 1898)
discoveryPlace (12345, Paris)
discoveryPerson (12345, Marie Curie)

spatio-temporal & contextual knowledge

atomicNumber (Polonium, 84)
halfLife (Polonium, 2.9 y)

expert knowledge

KB incompleteness is inherent

Reality

Why?

Einstein received the Nobel Prize in 1921, the Copley medal, the Prix Jules Jansen, the Medal named after Max Planck, and several others

Knowledge base construction

Award(Einstein, NobelPrize)
~~Award(Einstein, Copley medal)~~
Award(Einstein, Prix Jules Jansen)
Friend(Einstein, Max Planck)

1. Sources incomplete

3. Extractors imperfect

4. Extraction resource-bounded

Honorary doctorate, UMadrid
Gold medal, Royal Astronomic Society
Benjamin Franklin Medal,
…
¬NobelPrizeFor(Einstein, RelativityTheory)
¬NobelPrizeFor(Einstein, ElectricToaster)
…

2. Negations quasi-infinite

Weikum et al.
Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases
FnT 2021

Resulting challenges

1. Available KBs are incomplete

$$K^a << K^i$$

2. Available KBs hardly store negatives

$$K^{a^-} \approx \emptyset$$

Formal semantics for incomplete KBs: Closed vs. open-world assumption

won	
name	**award**
Brad Pitt	Oscar
Marie Curie	Nobel Prize
Berners-Lee	Turing Award

	Closed-world assumption	**Open-world assumption**
won(BradPitt, Oscar)?	→ *Yes*	→ *Yes*
won(Pitt, Nobel Prize)?	→ *No*	→ ***Maybe***

- Databases traditionally employ closed-world assumption
- KBs (semantic web) necessarily operate under open-world assumption

Open-world assumption

Thrones directed by Shakespeare?

World-aware AI?
Practically useful paradigm?

- Q: *Brad Pitt brother of*

  KB*: Maybe*

The logicians way out – completeness metadata

- Need power to express both maybe and no
  (*Some paradigm which allows both open- and closed-world interpretation of data to co-exist)*

- Approach: Completeness assertions [Motro 1989]

won	
name	**award**
Brad Pitt	Oscar
Marie Curie	Nobel Prize
Berners-Lee	Turing Award

Completeness assertion:

```
wonAward is
complete for
Nobel Prizes
```

won(Pitt, Oscar)? → *Yes*

won(Pitt, Nobel)? → *No* (CWA)

won(Pitt, Turing)? → *Maybe* (OWA)

The power of completeness metadata

Know what the KB knows:

→ Locally, $K^a = K^i$

Absent assertions are really false:

→ Locally, $s \neg\in K^a$ implies $s \neg\in K^i$

Completeness metadata: Formal view

Complete (won(name, award); award = ‘Nobel’)

Implies constraint on possible state of $\mathbf{K^a}$ and $\mathbf{K^i}$

$won^i(name, \text{‘Nobel’}) \rightarrow won^a(name, \text{‘Nobel’})$

(tuple-generating dependency)

Darari et al.
Completeness Statements about RDF Data Sources and Their Use for Query Answering
ISWC 2013

Cardinality assertions: Formal view

- *"Nobel prize was awarded 603 times"*

→ $|won^{i}(name, \text{'Nobel'})| = 603$

→ Allows counting objects in $\mathbf{K}^{a}$
  - Equivalent count → Completeness assertion
  - Otherwise, fractional coverage/recall information
    - *"93% of awards covered"*

- Grounded in number restrictions/role restrictions in Description Logics

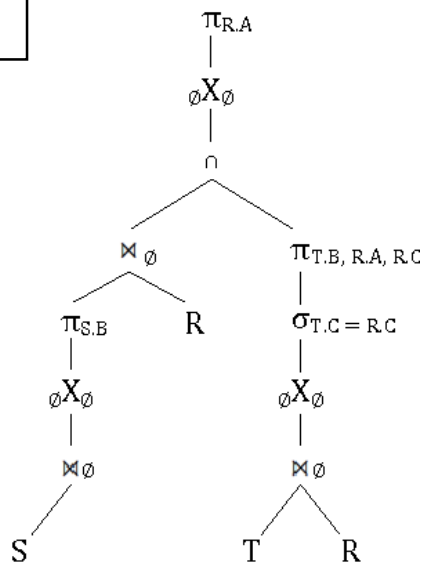
B. Hollunder and F. Baader
Qualifying Number Restrictions in Concept Languages
KR 1991

Formal reasoning with completeness metadata

Problem: **Query completeness reasoning**
Input:
- Set of completeness assertions for base relations
- Query Q

Task:
- Compute completeness assertions that hold for result of Q

Long-standing problem in database theory
[Motro, 1989, Fan and Geerts, 2009,
Razniewski and Nutt, 2011, …]



Where can completeness metadata come from?

- Data creators should pass them along as metadata
- Or editors should add them in curation steps

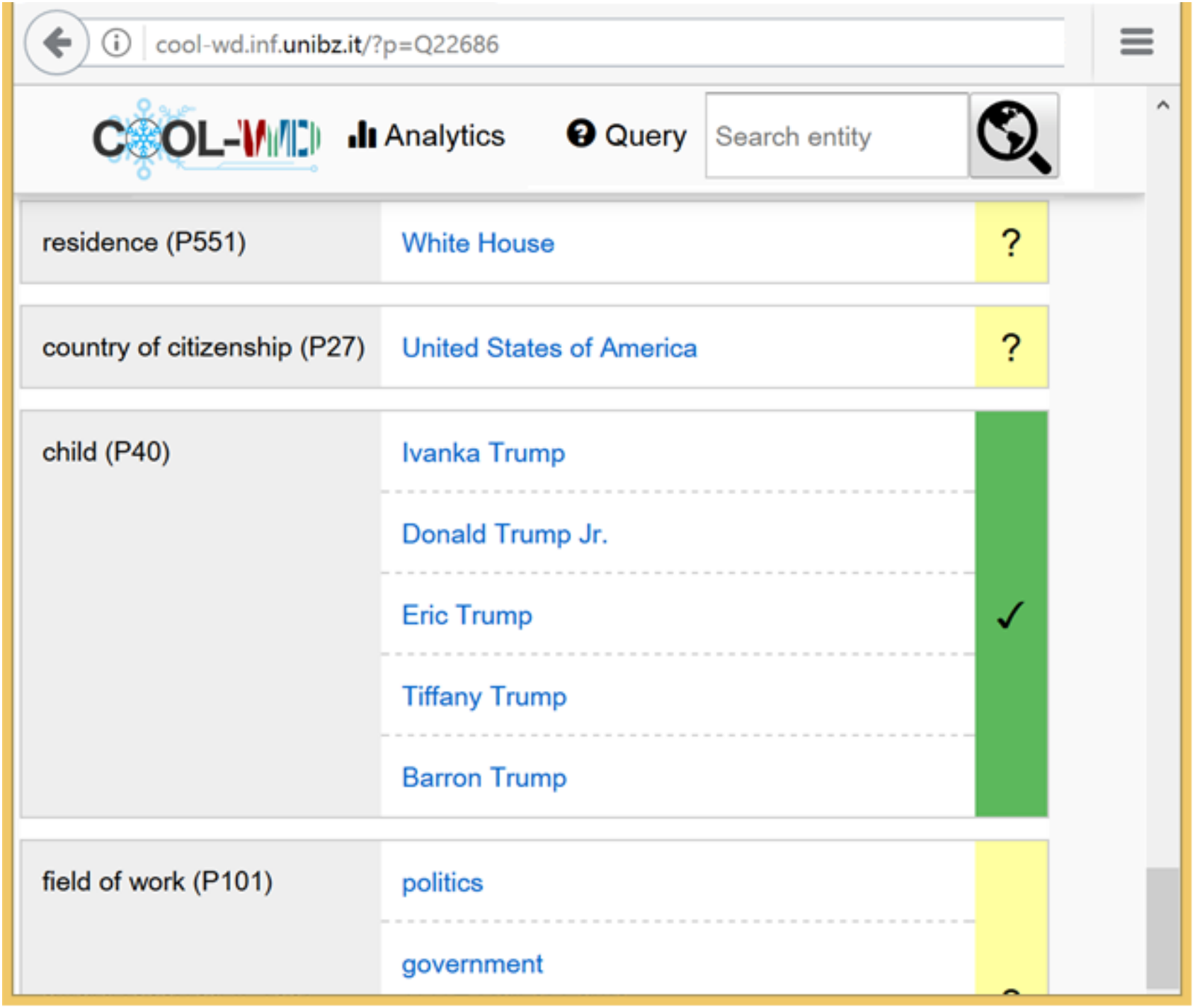
Abingdon	4. Residential triangle, Longmead etc.		Pub is only restaurant? Footways that link stuff, stubbed in places.
Shippon	5. Whole village, minus the barracks		Mostly done here.

This is a complete **list of compositions by Maurice Ravel**,

28	*Tout est lumière*	soprano, mixed choir, and orchestra	1901	• Prix de Rome competition
29	*Myrrha*, cantata	soprano, tenor, baritone, and orchestra	1901	text: Fernand Beissier; • Prix de Rome competition
31	*Semiramis*	cantata	1902	• student competition; • partially lost

- E.g., COOL-WD tool

Darari et al.
COOL-WD: A Completeness Tool for Wikidata
ISWC 2017


cool-wd.inf.unibz.it/?p=Q22686
COOL-WD
Analytics
Query
Search entity
residence (P551)
White House
?
country of citizenship (P27)
United States of America
?
child (P40)
Ivanka Trump
Donald Trump Jr.
Eric Trump
Tiffany Trump
Barron Trump
✓
field of work (P101)
politics
government

But…

- Requires human effort
  - Soliciting metadata more demanding than data
  - Automatically created KBs do not even have editors

Remainder of this tutorial:

How to automatically acquire information about what a KB knows

Takeaway Part 1: Foundations

- KBs are pragmatic collections of knowledge
  - Issue 1: Inherently incomplete
  - Issue 2: Hardly store negative knowledge
- Open-world assumption (OWA) as formal interpretation leads to counterintuitive results
- Metadata about completeness or counts as way out

Next: How to use predictive models to derive completeness metadata

On the Limits of Machine Knowledge: Completeness, Recall and Negation in Web-scale Knowledge Bases

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

1. Introduction & Foundations (Simon) – 15:45-16:00 CEST
2. Predictive recall assessment (Fabian) – 16:00-16:20
3. Counts from text and KB (Shrestha) – 16:20-16:40
4. Negation (Hiba) – 16:40-17:00
5. Wrap-up (Simon) – 17:00-17:10

THE WEB CONFERENCE ACM

max planck institut informatik

TELECOM ParisTech

Wrap-up: Take-aways

1. KBs are incomplete and limited on the negative side
2. Predictive techniques work from a surprising set of paradigms
3. Count information a prime way to gain insights into completeness/coverage
4. Salient negations can be heuristically materialized
5. Relative completeness tangible alternative

Wrap-up: Recipes

- Ab-initio KB construction
  1. Intertwine data and metadata collection
  2. Human insertion: Provide tools
  3. Automated extraction: Learn from extraction context

- KB curation
  1. Exploit KB-internal or textual cardinality assertions
  2. Inspect statistical properties on density or distribution
  3. Compute overlaps on pseudo-random samples

Open research questions

1. How are entity, property and fact completeness related?

2. How to distinguish salient negations from data modelling issues?

3. How to estimate coverage of knowledge in pre-trained language models?

4. How to identify most valuable areas for recall improvement?

Wrap-up: Wrap-up

- KBs major drivers of knowledge-intensive applications
- Severe limitations concerning completeness and coverage-awareness
- This tutorial: Overview of problem, techniques and tools to obtain awareness of completeness

Takeaway Part 1: Foundations

- KBs are pragmatic collections of knowledge
  - Issue 1: Inherently incomplete
  - Issue 2: Hardly store negative knowledge
- Open-world assumption (OWA) as formal interpretation leads to counterintuitive results
- Metadata about completeness or counts as way out

Takeaway: Predictive recall assessment

Using statistical techniques, we can predict more or less

- the recall of facts
  - are we missing objects for a subject?
  - do all subjects have an attribute in the real world?
  - does a text enumerate all objects for a subject?
- the recall of entities
  - is the distribution of entities representative?
  - how many entities are in the real world?

Takeaway: Counts from text and KB

1. Count information comes in two variants
   - Counting predicates - store integer counts
   - Enumerating predicates - store entities
2. Count information in text
   - occurs as cardinals, ordinals, non-numeric noun phrases
   - occurs with compositional cues
3. Count information in KBs
   - is expressed in two variants
   - occurs semantically related count predicates
4. Count information
   - can enrich KB
   - highlight inconsistencies

Takeaway: negation

- **Current KBs lack negative knowledge**
- **Rising interest in the explicit addition of negation to OW KB.**
- **Negations highly relevant in many applications including:**
  - **Commercial decision making (e.g., hotel booking)**
  - **General-domain question answering systems (e.g., is Switzerland a member of the EU?)**
- **Methodologies include:**
  - **Statistical inference**
  - **Text extraction**
  - **Pretrained LMs.**