https://www.mpi-inf.mpg.de/vldb-2021-tutorial

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

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

https://www.mpi-inf.mpg.de/vldb-2021-tutorial

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

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

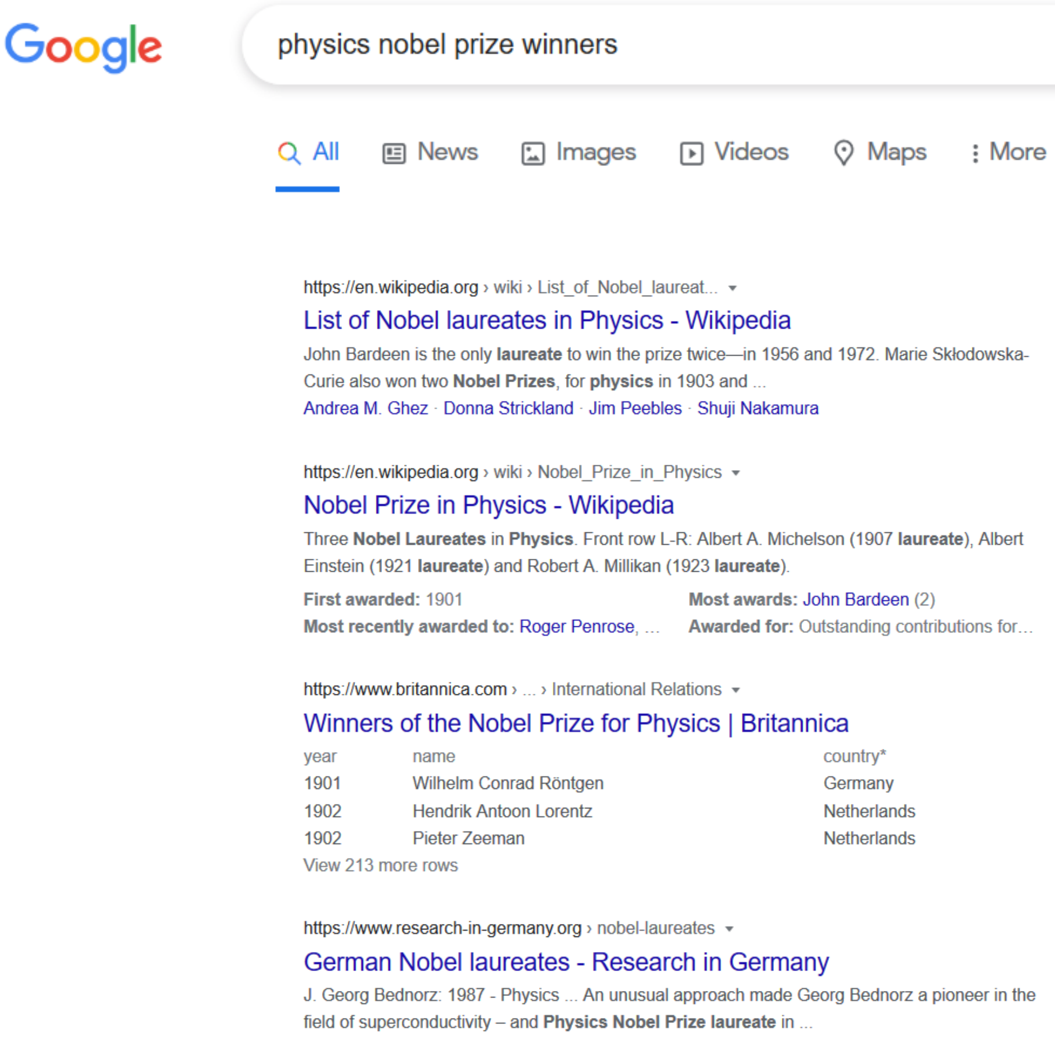
1. Introduction & Foundations (Simon) – 20 min
2. Predictive recall assessment (Fabian) – 20 min
3. Counts from text and KB (Shrestha) – 20 min
4. Negation (Hiba) – 20 min
5. Wrap-up (Simon) – 5 min

max planck institut
informatik

TELECOM ParisTech

Machine knowledge in action

Traditional search



physics nobel prize winners

https://en.wikipedia.org › wiki › List_of_Nobel_laureat...

List of Nobel laureates in Physics - Wikipedia

John Bardeen is the only **laureate** to win the prize twice—in 1956 and 1972. Marie Skłodowska-Curie also won two **Nobel Prizes**, for **physics** in 1903 and ...

Andrea M. Ghez · Donna Strickland · Jim Peebles · Shuji Nakamura

https://en.wikipedia.org › wiki › Nobel_Prize_in_Physics

Nobel Prize in Physics - Wikipedia

Three **Nobel Laureates** in **Physics**. Front row L-R: Albert A. Michelson (1907 **laureate**), Albert Einstein (1921 **laureate**) and Robert A. Millikan (1923 **laureate**).

First awarded: 1901

Most recently awarded to: Roger Penrose, ...

Most awards: John Bardeen (2)

Awarded for: Outstanding contributions for...

https://www.britannica.com › ... › International Relations

Winners of the Nobel Prize for Physics | Britannica

year	name	country*
1901	Wilhelm Conrad Röntgen	Germany
1902	Hendrik Antoon Lorentz	Netherlands
1902	Pieter Zeeman	Netherlands

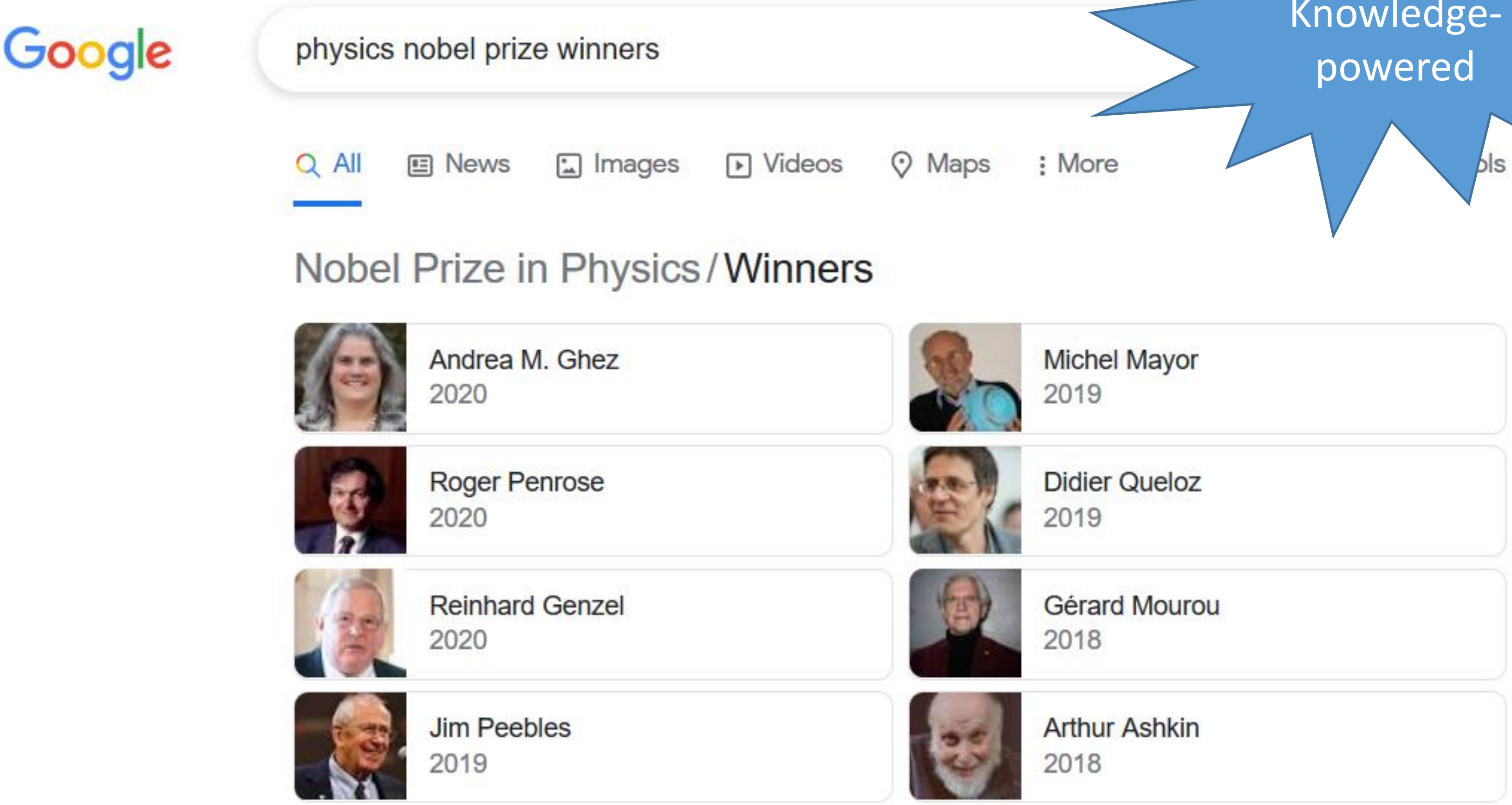
View 213 more rows

https://www.research-in-germany.org › nobel-laureates

German Nobel laureates - Research in Germany

J. Georg Bednorz: 1987 - Physics ... An unusual approach made Georg Bednorz a pioneer in the field of superconductivity – and **Physics Nobel Prize laureate** in ...

Machine knowledge in action



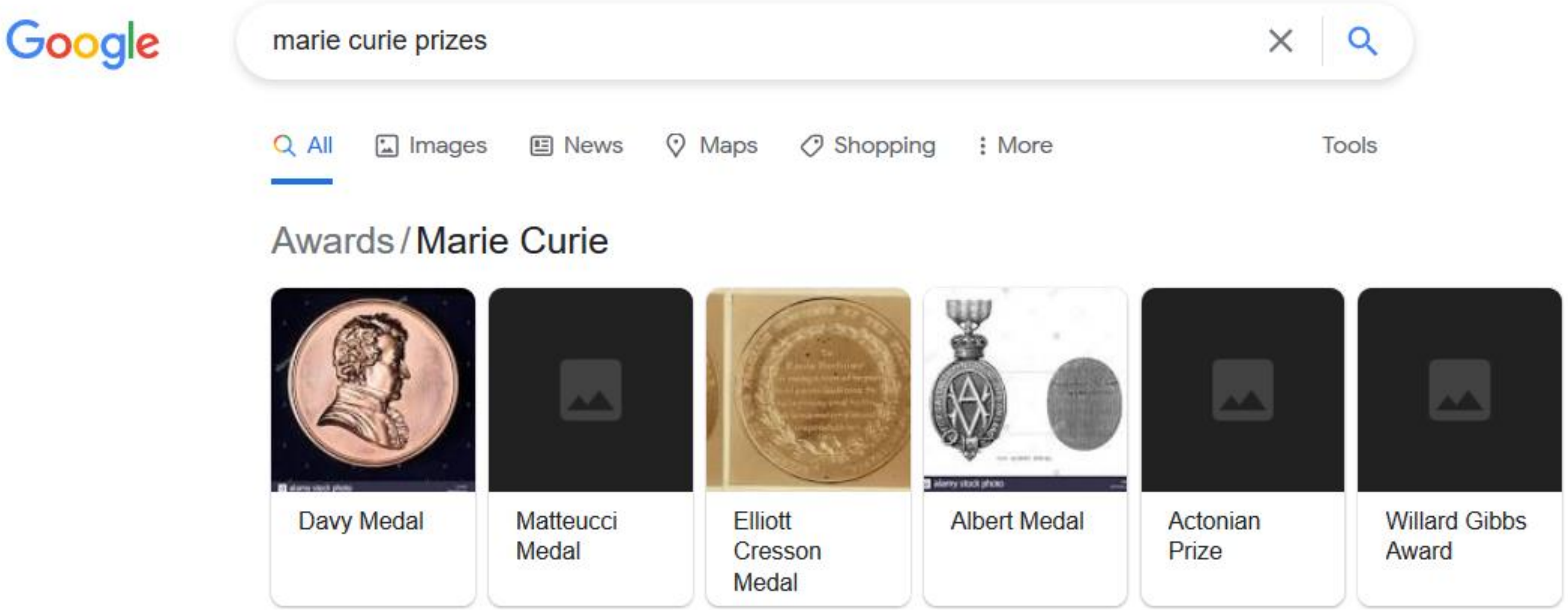
physics nobel prize winners

Nobel Prize in Physics / Winners

Andrea M. Ghez 2020	Michel Mayor 2019
Roger Penrose 2020	Didier Queloz 2019
Reinhard Genzel 2020	Gérard Mourou 2018
Jim Peebles 2019	Arthur Ashkin 2018

Knowledge-powered

Machine knowledge in action



marie curie prizes

Awards / Marie Curie

Davy Medal

Matteucci Medal

Elliott Cresson Medal

Albert Medal

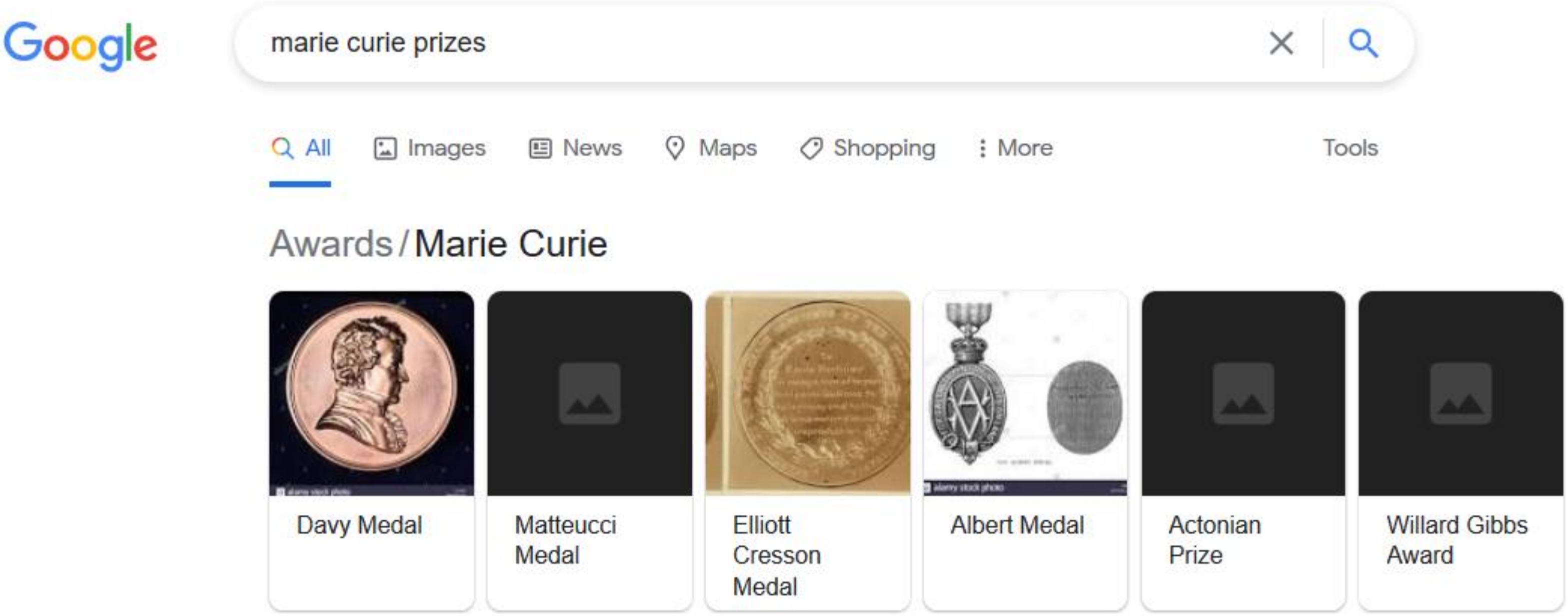
Actonian Prize

Willard Gibbs Award

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 commercial and public players
    - Wikidata, Google KG, Microsoft Satori, …
- Strongly rooted in database community
    - Data integration, data cleaning, conceptual modelling, storage, indexing and querying, …

But:
Machine Knowledge is incomplete



marie curie prizes

Awards / Marie Curie

- Davy Medal
- Matteucci Medal
- Elliott Cresson Medal
- Albert Medal
- Actonian Prize
- Willard Gibbs Award

Nobel Prize (2x)

Machine knowledge is incomplete (2)

Wikidata KB:

VLDB journal has only published 80 articles ever

- https://scholia.toolforge.org/venue/Q15760089

Most cited papers on data integration have <38 citations

- https://scholia.toolforge.org/topic/Q386824

But: Machine knowledge is one-sided

- In KB:
  - *Stephen Hawking won Presidential medal of freedom*
  - *Vietnam is a member of ASEAN*
  - *iPhone has 12MP camera*

- Not in KB:
  - *Stephen Hawking did not win the Nobel Prize*
  - *Switzerland is not a member of the EU*
  - *iPhone has no headphone jack*

Why is this problematic? (1) Querying

- Decision making more and more data-driven

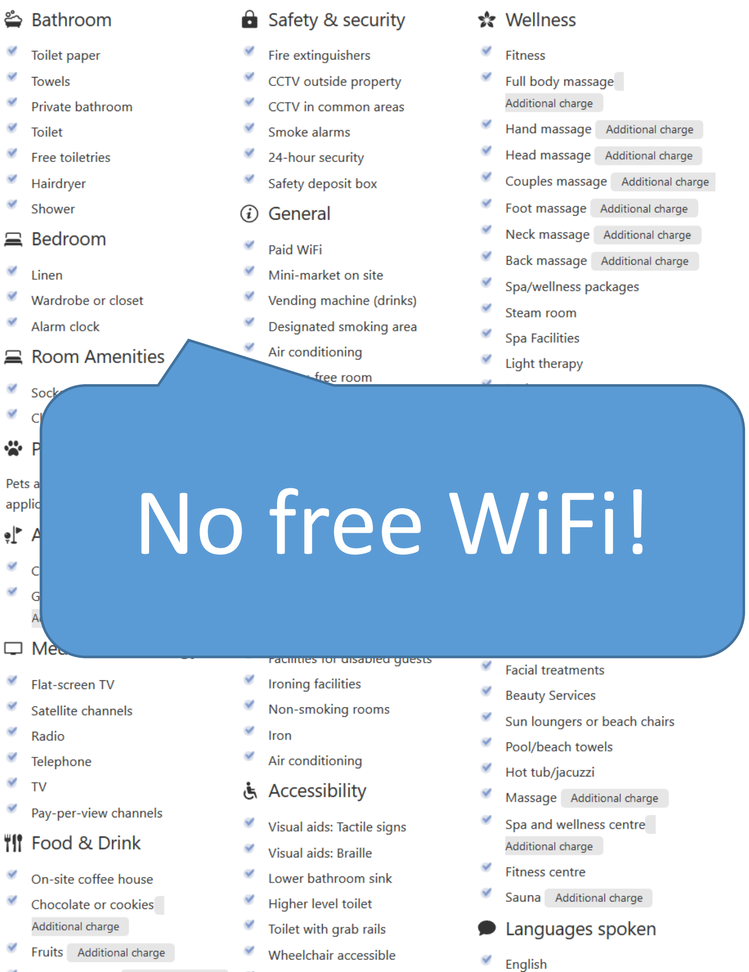
- Analytical queries paint wrong picture of reality
  - *E.g., VLDB journal deemed too small*

- Instance queries return wrong results
  - *E.g., wrongly assuming certain authors never published in VLDBJ*

Why is this problematic? (1) Data Curation

- Effort priorization fundamental challenge in human-in-the-loop curation
    - *Should we spend effort on obtaining data for VLDB or TKDE?*
- Risk of effort duplication if not keeping track of completed areas
    - *Data for TKDE complete up to 2020*

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



Bathroom
- Toilet paper
- Towels
- Private bathroom
- Toilet
- Free toiletries
- Hairdryer
- Shower

Bedroom
- Linen
- Wardrobe or closet
- Alarm clock

Room Amenities
- Sock
- C

P

Pets a
applic

A
- C
- G

A

Me
- Flat-screen TV
- Satellite channels
- Radio
- Telephone
- TV
- Pay-per-view channels

Food & Drink
- On-site coffee house
- Chocolate or cookies Additional charge
- Fruits Additional charge

Safety & security
- Fire extinguishers
- CCTV outside property
- CCTV in common areas
- Smoke alarms
- 24-hour security
- Safety deposit box

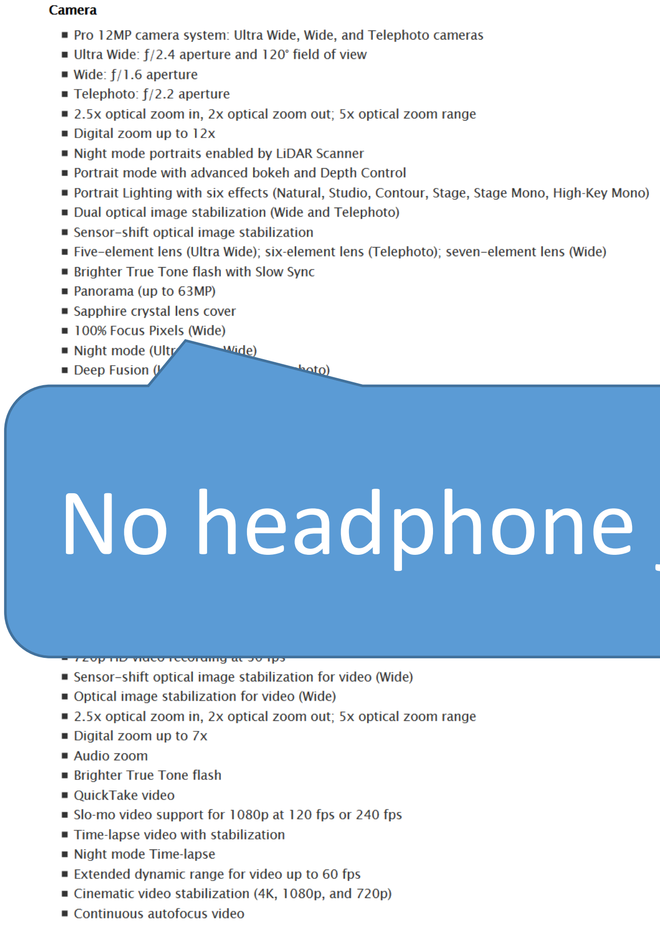
General
- Paid WiFi
- Mini-market on site
- Vending machine (drinks)
- Designated smoking area
- Air conditioning
- free room
- Facilities for disabled guests
- Ironing facilities
- Non-smoking rooms
- Iron
- Air conditioning

Accessibility
- Visual aids: Tactile signs
- Visual aids: Braille
- Lower bathroom sink
- Higher level toilet
- Toilet with grab rails
- Wheelchair accessible

Wellness
- Fitness
- Full body massage Additional charge
- Hand massage Additional charge
- Head massage Additional charge
- Couples massage Additional charge
- Foot massage Additional charge
- Neck massage Additional charge
- Back massage Additional charge
- Spa/wellness packages
- Steam room
- Spa Facilities
- Light therapy
- Facial treatments
- Beauty Services
- Sun loungers or beach chairs
- Pool/beach towels
- Hot tub/jacuzzi
- Massage Additional charge
- Spa and wellness centre Additional charge
- Fitness centre
- Sauna Additional charge

Languages spoken
- English

No free WiFi!



Camera
- Pro 12MP camera system: Ultra Wide, Wide, and Telephoto cameras
- Ultra Wide: ƒ/2.4 aperture and 120° field of view
- Wide: ƒ/1.6 aperture
- Telephoto: ƒ/2.2 aperture
- 2.5x optical zoom in, 2x optical zoom out; 5x optical zoom range
- Digital zoom up to 12x
- Night mode portraits enabled by LiDAR Scanner
- Portrait mode with advanced bokeh and Depth Control
- Portrait Lighting with six effects (Natural, Studio, Contour, Stage, Stage Mono, High-Key Mono)
- Dual optical image stabilization (Wide and Telephoto)
- Sensor-shift optical image stabilization
- Five-element lens (Ultra Wide); six-element lens (Telephoto); seven-element lens (Wide)
- Brighter True Tone flash with Slow Sync
- Panorama (up to 63MP)
- Sapphire crystal lens cover
- 100% Focus Pixels (Wide)
- Night mode (Ultr Wide)
- Deep Fusion (hoto)
- 720p HD video recording at 30 fps
- Sensor-shift optical image stabilization for video (Wide)
- Optical image stabilization for video (Wide)
- 2.5x optical zoom in, 2x optical zoom out; 5x optical zoom range
- Digital zoom up to 7x
- Audio zoom
- Brighter True Tone flash
- QuickTake video
- Slo-mo video support for 1080p at 120 fps or 240 fps
- Time-lapse video with stabilization
- Night mode Time-lapse
- Extended dynamic range for video up to 60 fps
- Cinematic video stabilization (4K, 1080p, and 720p)
- Continuous autofocus video

No headphone jack

Topic of this tutorial

How to know
how much a KB knows?

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

What this tutorial offers

- Logical foundations
  - Languages 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)

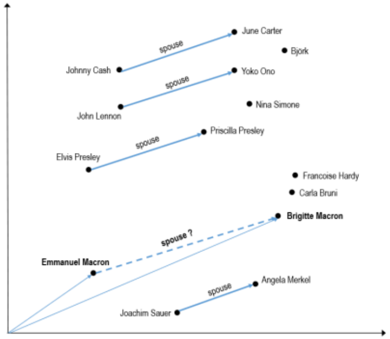
Goals:

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

Relevant research domains

- Databases
- Logics
- Statistics
- Machine Learning
- Natural language processing

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

Beatles members:

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

- 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*

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) – 20 min
2. Predictive recall assessment (Fabian) – 20 min
3. Counts from text and KB (Shrestha) – 20 min
4. Negation (Hiba) – 20 min
5. Wrap-up (Simon) – 5 min

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) $\in \mathbf{E} \times \mathbf{P} \times (\mathbf{E} \cup \mathbf{L})$
- An 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

History of knowledge bases

Cyc

WordNet

guitarist
$\subset$ {player,musician}
$\subset$ artist

{player,footballer}
$\subset$ athlete

$\forall$ x: human(x) $\Rightarrow$
($\exists$ y: mother(x,y) $\wedge$
$\exists$ z: father(x,z))

$\forall$ x,u,w: (mother(x,u) $\wedge$
mother(x,w)
$\Rightarrow$ u=w)

Wikipedia

6 Mio. English articles
40 Mio. contributors

Manual compilation

Automation and human-in-the-loop

DBpedia

yago select knowledge

freebase™

WolframAlpha computational knowledge engine

amazon

bing

WIKIDATA

DIFFBOT

Google Knowledge Graph

BOSCH

Bloomberg

Alibaba.com

1985 1990 2000 2005 2010 2020

KB scale and use cases

Wikidata (open)

- 95 M items
- 1.1 B statements

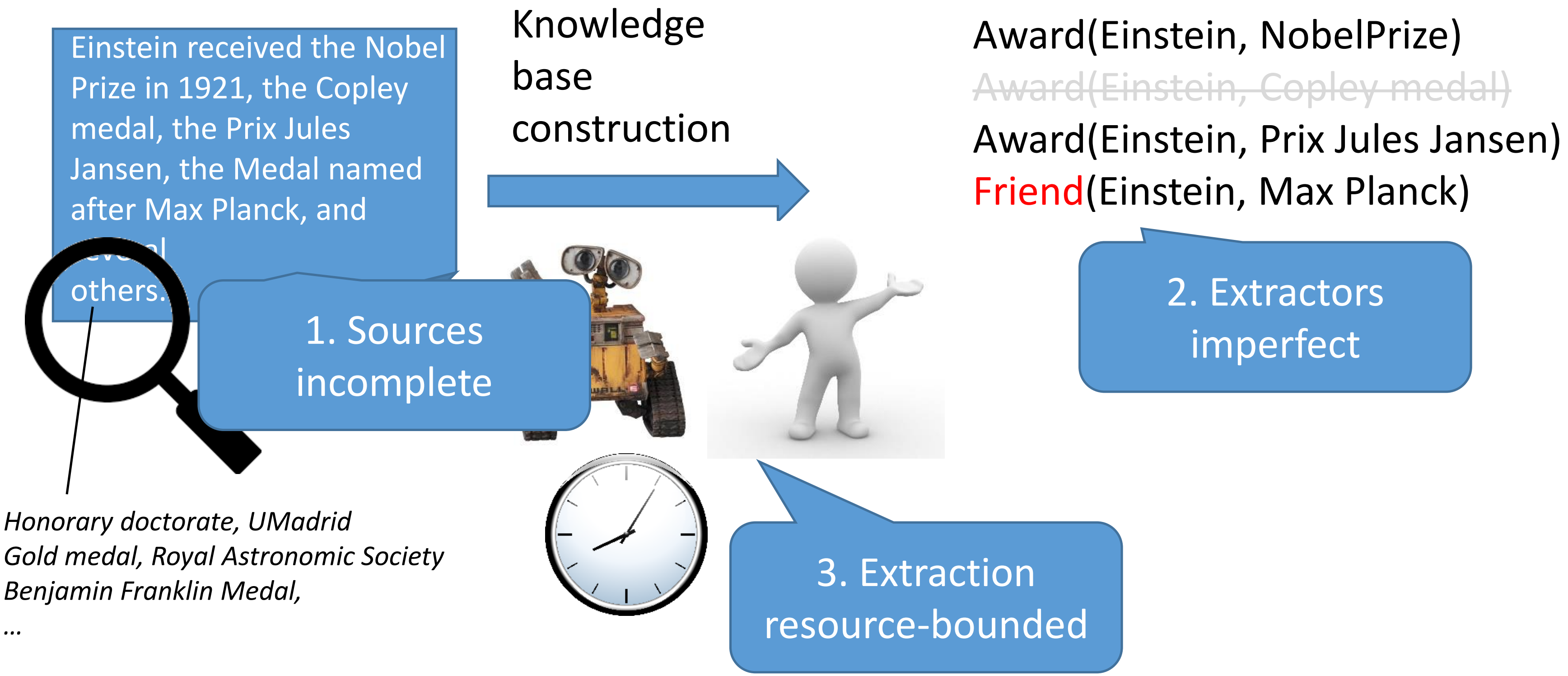
Google KG

- 5 B items
- 500 B statements

Major use cases:

- semantic search & QA
- language understanding
- distant supervision for ML
- data cleaning

KB incompleteness is inherent



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

2. Extractors imperfect

3. Extraction resource-bounded

Honorary doctorate, UMadrid
Gold medal, Royal Astronomic Society
Benjamin Franklin Medal,
…

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

Root 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 and open-world assumption

won	
name	**award**
Brad Pitt	Oscar
Einstein	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

ected by Shakespeare?

World-aware AI?
Practically useful paradigm?

- Q

  KB*: Maybe*

- Q: *Trump brother of Kim Jong Un*

  KB*: Maybe*

The logicians way out – completeness assertions

- 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 statements [Motro 1989]

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

Completeness statement:

`wonAward is complete for Nobel Prizes`

won(Pitt, Oscar)? → *Yes*

won(Pitt, Nobel)? → *No* (CWA)

won(Pitt, Turing)? → *Maybe* (OWA)

The power of completeness assertions

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 statements: Formal view

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

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

won^{i}(name, ‘Nobel’) → won^{a}(name, ‘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”*

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

→ Allows counting objects in **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 assertions

Problem: Query completeness reasoning

Input:

- Set of completeness assertions for base relations
- Query Q

Task:

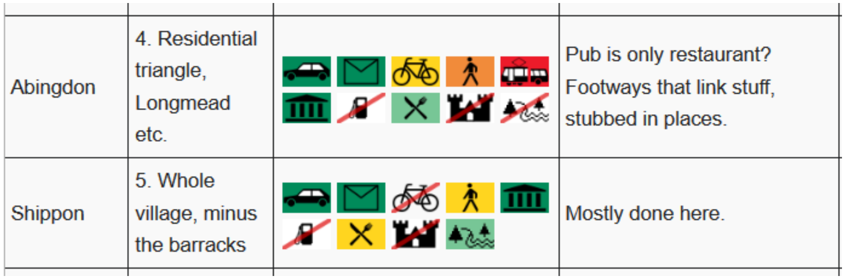
- Compute completeness assertions that hold for result of Q

Formal reasoning with completeness assertions

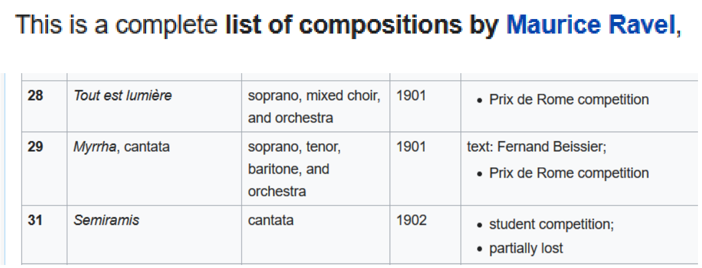
Work	Description Language	Results
Motro, TODS 1989	Views	Algorithm
Fan & Geerts, PODS 2009	Various query languages (CQ-Datalog)	Decidability/ Complexity
Razniewski & Nutt 2011	Join queries	Complexity
Lang et al., SIGMOD 2014	Selections	Algorithm
Razniewski et al., SIGMOD 2016	Selections	Algorithm, computational completeness

Where can completeness statements come from?

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



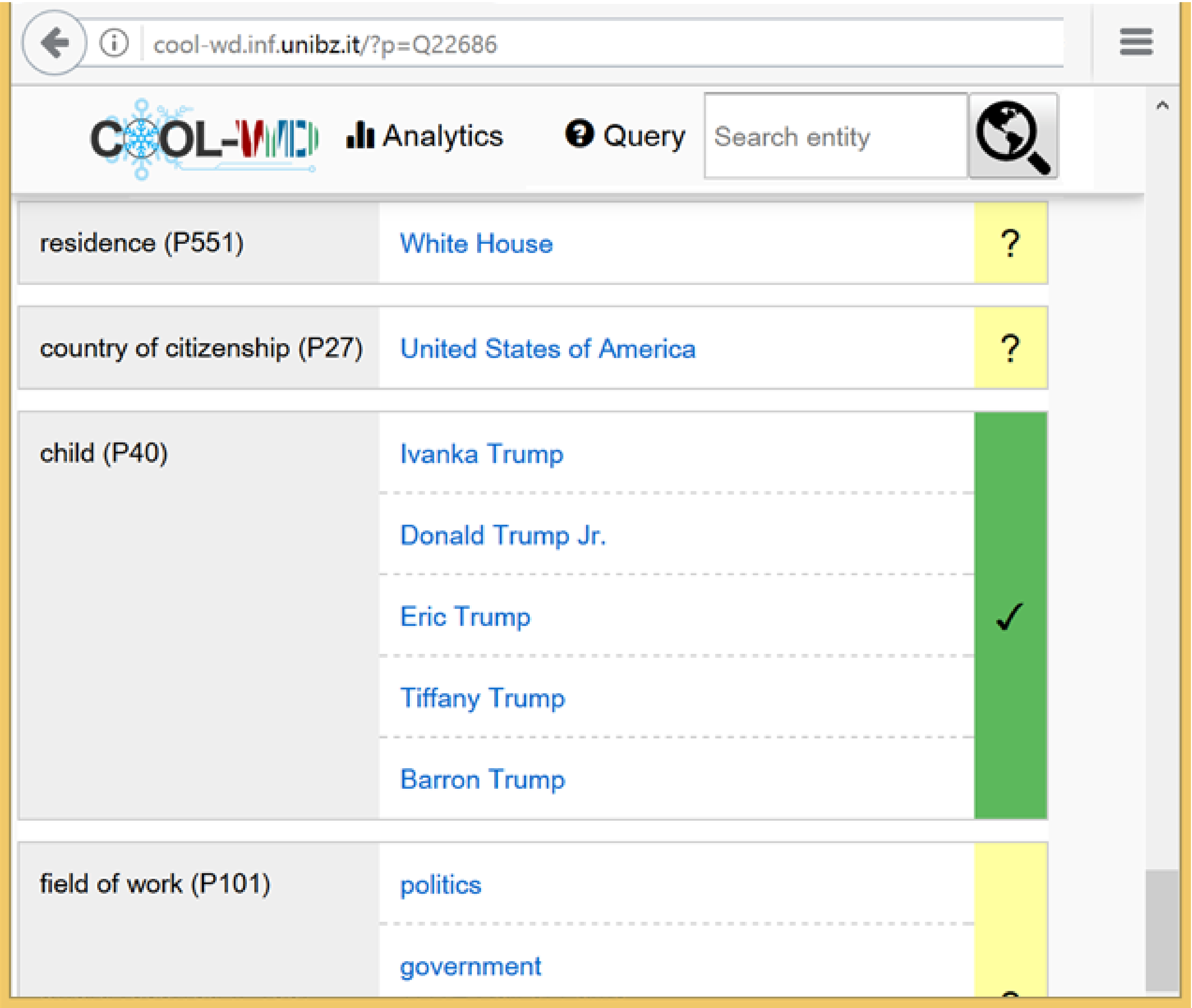
Abingdon	4. Residential triangle, Longmead etc.	🚗 ✉ 🚲 🚶 🚌 🏛 ~~⛽~~ 🍴 ~~🏰~~ ~~🌲~~	Pub is only restaurant? Footways that link stuff, stubbed in places.
Shippon	5. Whole village, minus the barracks	🚗 ✉ ~~🚲~~ 🚶 🏛 ~~⛽~~ 🍴 ~~🏰~~ 🌲	Mostly done here.



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 (**Co**mpleteness t**ool** for **Wikid**ata)



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
  - Editors are lazy
  - 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 for completeness assessment

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) – 20 min
2. Predictive recall assessment (Fabian) – 20 min
3. Counts from text and KB (Shrestha) – 20 min
4. Negation (Hiba) – 20 min
5. Wrap-up (Simon) – 5 min

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

Wrap-up: Recipe

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

- KB curation
  - Exploit KB-internal or textual cardinality assertions
  - Inspect statistical properties on density or distribution
  - 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?

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.**