

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



Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

# 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

# Machine knowledge in action



physics nobel prize winners

[All](#) [News](#) [Images](#) [Videos](#) [Maps](#) [More](#)

[https://en.wikipedia.org/wiki/List\\_of\\_Nobel\\_laureates\\_in\\_Physics](https://en.wikipedia.org/wiki/List_of_Nobel_laureates_in_Physics)

## 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](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 awards:** [John Bardeen](#) (2)

**Most recently awarded to:** [Roger Penrose](#), ...

**Awarded for:** Outstanding contributions for...

<https://www.britannica.com/topic/nobel-prize/physics>

## 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/en/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 ...

Traditional  
search

# Machine knowledge in action



physics nobel prize winners

All

News

Images

Videos

Maps

More

Tools

Knowledge-  
powered

## 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



# Machine knowledge in action



marie curie prizes



All

Images

News

Maps

Shopping

More

Tools

## Awards / Marie Curie



Davy Medal



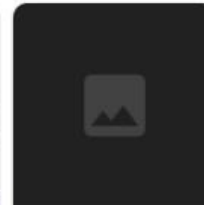
Matteucci  
Medal



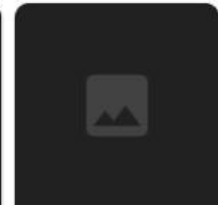
Elliott  
Cresson  
Medal



Albert Medal



Actonian  
Prize



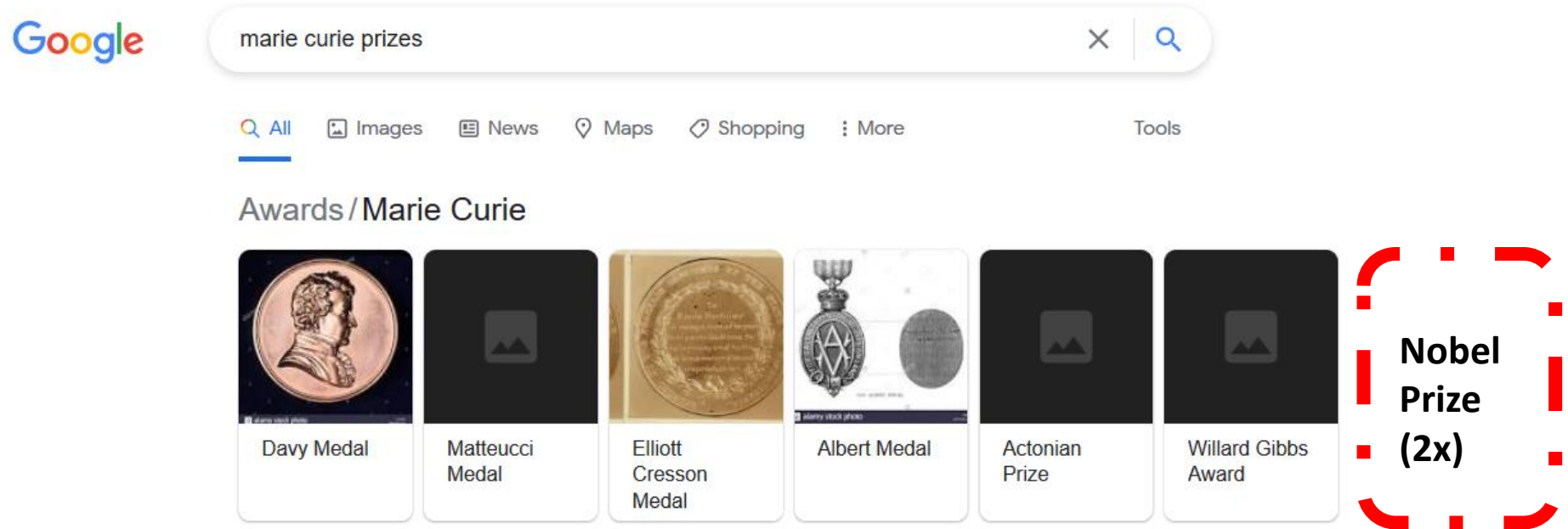
Willard Gibbs  
Award

# 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 public and commercial 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



# 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 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., VLDB journal deemed too small*
- Instance queries return wrong results
  - *E.g., wrongly assuming certain authors never published in VLDBJ*

# Why is this problematic? (2)

## Data Curation

- Effort prioritization 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
  - *Spending effort on collecting data ... already present*

# Why is this problematic? (3)

## Summarization and decision making

Booking.com

- Bathroom**
  - ✓ Toilet paper
  - ✓ Towels
  - ✓ Private bathroom
  - ✓ Toilet
  - ✓ Free toiletries
  - ✓ Hairdryer
  - ✓ Shower
- Bedroom**
  - ✓ Linen
  - ✓ Wardrobe or closet
  - ✓ Alarm clock
- Room Amenities**
  - ✓ Socks
  - ✓ Cleaning products
  - ✓ Pets and applicable
  - ✓ Air conditioning
- Medicine**
  - ✓ Flat-screen TV
  - ✓ Satellite channels
  - ✓ Radio
  - ✓ Telephone
  - ✓ TV
  - ✓ Pay-per-view channels
- Food & Drink**
  - ✓ On-site coffee house
  - ✓ Chocolate or cookies
  - ✓ Fruits
- 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
- Wellness**
  - ✓ Fitness
  - ✓ Full body massage
  - ✓ Hand massage
  - ✓ Head massage
  - ✓ Couples massage
  - ✓ Foot massage
  - ✓ Neck massage
  - ✓ Back massage
  - ✓ Spa/wellness packages
  - ✓ Steam room
  - ✓ Spa Facilities
  - ✓ Light therapy
- 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
- Facilities for disabled guests**
  - ✓ Facial treatments
  - ✓ Beauty Services
  - ✓ Sun loungers or beach chairs
  - ✓ Pool/beach towels
  - ✓ Hot tub/jacuzzi
  - ✓ Massage
  - ✓ Spa and wellness centre
  - ✓ Fitness centre
  - ✓ Sauna
- Languages spoken**
  - ✓ English

No free WiFi!



### Camera

- Pro 12MP camera system: Ultra Wide, Wide, and Telephoto cameras
- Ultra Wide: f/2.4 aperture and 120° field of view
- Wide: f/1.6 aperture
- Telephoto: f/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 (Ultra Wide, Wide, Telephoto)
- Deep Fusion (Ultra Wide, Wide, Telephoto)

No headphone jack

- 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



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
  - 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)

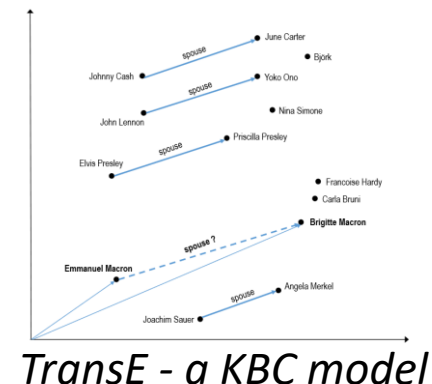
## Goals:

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



- **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) – 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

# Knowledge base - definition

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

- Predicates are positive or negated properties
  - *bornIn, notWonAward, ...*
- An **assertion** is a triple  $(s, p, o) \in \mathbf{E} \times \mathbf{P} \times (\mathbf{EUL})$
- 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

# History of knowledge bases



**Cyc**

**WordNet**



Manual compilation

Automation and  
human-in-the-loop

guitarist  
 $\subset \{\text{player}, \text{musician}\}$   
 $\subset \text{artist}$   
 $\{\text{player}, \text{footballer}\}$   
 $\subset \text{athlete}$

**Wikipedia**



6 Mio. English articles  
 40 Mio. contributors

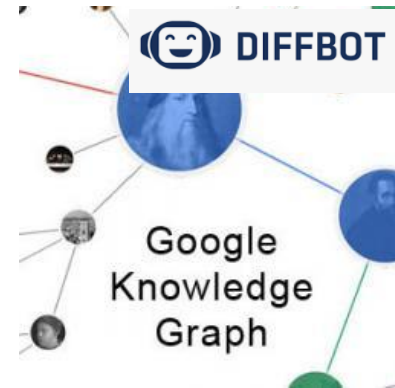
**WolframAlpha**™ computational knowledge engine

**amazon**



WIKIDATA

**DIFFBOT**



**BOSCH**



**Alibaba.com**



**freebase**™

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

Why?

Reality

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

Doc

1. Sources incomplete

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

Knowledge base construction



3. Extraction resource-bounded

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

2. Extractors imperfect

Weikum et al.

[Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases](#)

FnT 2021

# Resulting challenges

1. Available KBs are incomplete

$$K^a \ll K^i$$

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
Marie Curie	Nobel Prize
Berners-Lee	Turing Award

**Closed-world  
assumption**

*won(BradPitt, Oscar)?* → *Yes*

*won(Pitt, Nobel Prize)?* → *No*

**Open-world  
assumption**

→ *Yes*

→ ***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?**

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
Marie Curie	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  $K^a$  and  $K^i$

$won^i(name, 'Nobel') \rightarrow won^a(name, 'Nobel')$

(tuple-generating dependency)



# Cardinality assertions: Formal view

- *“Nobel prize was awarded 603 times”*  
→  $|\text{won}^i(\text{name}, \text{'Nobel'})| = 603$
- 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

# Formal reasoning with completeness assertions

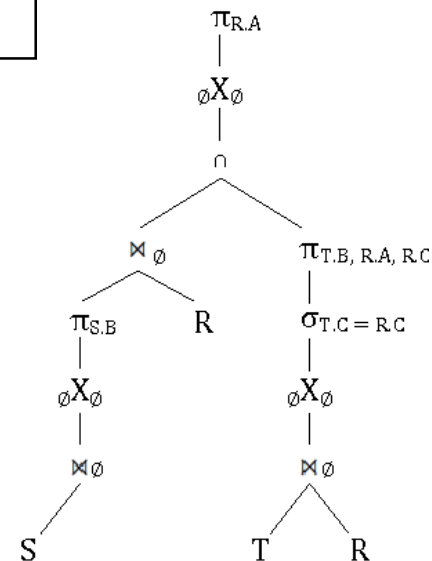
**Problem:** Query completeness reasoning

**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	<i>Tout est lumière</i>	soprano, mixed choir, and orchestra	1901	<ul style="list-style-type: none"><li>• Prix de Rome competition</li></ul>
29	<i>Myrrha</i> , cantata	soprano, tenor, baritone, and orchestra	1901	text: Fernand Beissier; <ul style="list-style-type: none"><li>• Prix de Rome competition</li></ul>
31	<i>Semiramis</i>	cantata	1902	<ul style="list-style-type: none"><li>• student competition;</li><li>• partially lost</li></ul>

- E.g., COOL-WD tool  
(**Completeness tool** for **Wikidata**)



cool-wd.inf.unibz.it/?p=Q22686



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
  - 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**

# 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 (Simon) – 10 min
2. Foundations (Simon) – 10 min
3. Predictive recall assessment (Fabian) – 20 min
4. Counts from text and KB (Shrestha) – 20 min
5. Negation (Hiba) – 20 min



# Predictive recall assessment

How can we find out if a knowledge base is complete?

- Recall of facts
  - Do we have all objects for a subject?
  - Can we use text to determine completeness?
- Recall of entities
  - Do we have all entities of the real world?

# Are we missing objects?



marriedTo



In the KB, and correct

# Are we missing objects?



marriedTo



In the KB, and correct



# Are we missing objects?



marriedTo



In the KB, and correct

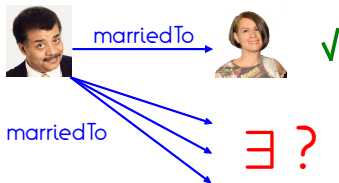
marriedTo



Not in the KB,  
but maybe also correct?



# Missing Object Problem



**Problem:** Missing Object Problem

**Input:**

- a knowledge base  $K$
- a subject  $s$
- a relation  $r$

**Task:** Determine if there is one or more  $o$  with  $r(s, o)$  in the real world, but  $r(s, o) \notin K$  (no matter which  $o$ , or how many  $o$ ).

# Signals for missing objects

## Closed World Assumption:

There are no missing objects (cf. first part of the tutorial).

## Partial Completeness Assumption:

If there are 1+ objects in the KB, then no object is missing.

## Popularity assumption:

If an entity is popular, it has no missing objects.

## No-change assumption:

If the number of objects did not change, none is missing.

# Complex signals for missing objects

## Class pattern oracle:

If the subject is in some class  $c$ , then there are (no) missing objects.

Example: Instances of "LivingPeople" are not missing a death date

## Star pattern oracle:

If the subject has one (or no) relationship  $r'$ , then there are no missing objects for relationship  $r$ .

Example: If you don't have a death place, you don't need a death date.

Can we combine and learn these signals?

# Learning signals for missing objects

Rule mining systems can learn (weighted) rules such as

$$\text{marriedTo}(x,y) \wedge \text{hasChild}(x,z) \Rightarrow \text{hasChild}(y,z)$$

Idea:

- 1) Add the ground truth on a sample of entities by crowdsourcing

We have all spouses of Elvis:  $\text{complete}(\text{Elvis}, \text{marriedTo})$

- 2) Add signals for missing objects as facts to the KB

Elvis is a popular entity:  $\text{popular}(\text{Elvis})$

Elvis has one spouse in the KB:  $\text{cardinalityIsNot0}(\text{Elvis}, \text{marriedTo})$

- 3) Use the rule miner to learn rules about missing objects

$$\text{cardinalityIsNot0}(x, \text{marriedTo}) \wedge \text{popular}(x) \Rightarrow \text{complete}(x, \text{marriedTo})$$

- 4) Use the rules to predict completeness

$$\begin{aligned} &\text{cardinalityIsNot0}(\text{Neil}, \text{marriedTo}) \wedge \text{popular}(\text{Neil}) \\ &\Rightarrow \text{complete}(\text{Neil}, \text{marriedTo}) \end{aligned}$$

->results

>results



# Learning rules for completeness

Artificially added assertions:

- $complete(x, r)$ : if  $x$  is complete on relation  $r$  on ground truth sample
- $incomplete(x, r)$ : same for incomplete
- $isPopular(x)$ :  $x$  is among the top 5% entities for number of facts
- $hasNotChanged(x, r)$ : no difference in objects between YAGO 1 and YAGO 3
- $notype(x, t)$ : entity  $x$  is not in class  $t$
- $lessThan_n(x, r)$ : entity  $x$  has less than  $n$  objects for relation  $r$
- $moreThan_n(x, r)$ : same for more

Example for rules learned with the AMIE system:

$dateOfDeath(x, y) \wedge lessThan_1(x, placeOfDeath) \Rightarrow incomplete(x, placeOfDeath)$

$IMDbId(x, y) \wedge producer(x, z) \Rightarrow complete(x, director)$

$notype(x, Adult) \wedge type(x, Person) \Rightarrow complete(x, hasChild)$

$lessThan_2(x, hasParent) \Rightarrow incomplete(x, hasParent)$

# Signals for Incompleteness (F1)

Relation	CWA	PCA	card <sub>2</sub>	Popularity	No change	Star	Class	AMIE
diedIn	60%	22%	—	4%	15%	50%	<b>99%</b>	96%
directed	40%	96%	19%	7%	71%	0%	0%	<b>100%</b>
graduatedFrom	89%	4%	2%	2%	10%	89%	<b>92%</b>	87%
hasChild	71%	1%	1%	2%	13%	40%	<b>78%</b>	<b>78%</b>
hasGender	78%	<b>100%</b>	—	2%	—	86%	95%	<b>100%</b>
hasParent*	1%	54%	<b>100%</b>	—	—	0%	0%	<b>100%</b>
isCitizenOf*	4%	98%	11%	1%	4%	10%	5%	<b>100%</b>
isConnectedTo	87%	34%	19%	—	—	68%	88%	<b>89%</b>
isMarriedTo*	55%	7%	0%	3%	12%	37%	<b>57%</b>	46%
wasBornIn	28%	<b>100%</b>	—	5%	8%	0%	0%	<b>100%</b>



Relation	CWA	PCA	card <sub>2</sub>	Popularity	Star	Class	AMIE
alma_mater	<b>90%</b>	14%	5%	1%	87%	87%	87%
brother	93%	1%	—	1%	<b>94%</b>	<b>96%</b>	<b>96%</b>
child	70%	1%	—	1%	<b>79%</b>	72%	73%
country_of_citizenship*	42%	97%	10%	3%	0%	0%	<b>98%</b>
director	81%	<b>100%</b>	—	3%	94%	89%	<b>100%</b>
father*	5%	<b>100%</b>	6%	9%	89%	8%	<b>100%</b>
mother*	3%	<b>100%</b>	3%	10%	67%*	5%	<b>100%</b>
place_of_birth	53%	<b>100%</b>	7%	5%	55%	0%	<b>100%</b>
place_of_death	89%	35%	1%	2%	81%	81%	<b>96%</b>
sex_or_gender	81%	<b>100%</b>	6%	3%	92%	91%	<b>100%</b>
spouse*	<b>57%</b>	7%	—	1%	54%	54%	55%



\* = biased training sample

# Missing Object Problem



marriedTo

$\exists$  ?

Are there objects in the real world that are missing from the KB?

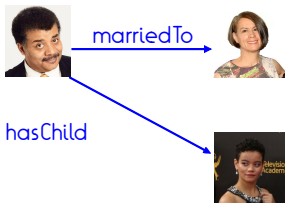
=> By help of supervised learning, we can learn rules that predict if an object is missing (although not which one, or how many).

$cardinalityIsNot0(x, marriedTo) \wedge popular(x) \Rightarrow complete(x, marriedTo)$

Luis Galárraga, Simon Razniewski, Antoine Amarilli, Fabian M. Suchanek:  
"Predicting Completeness in Knowledge Bases "

International Conference on Web Search and Data Mining (WSDM) 2017

# Missing Object Problem



=> By help of supervised learning, we can learn rules that predict if an object is missing (although not which one, or how many).

$$cardinalityIsNot0(x, marriedTo) \wedge popular(x) \Rightarrow complete(x, marriedTo)$$

Luis Galárraga, Simon Razniewski, Antoine Amarilli, Fabian M. Suchanek:  
"Predicting Completeness in Knowledge Bases "

>married

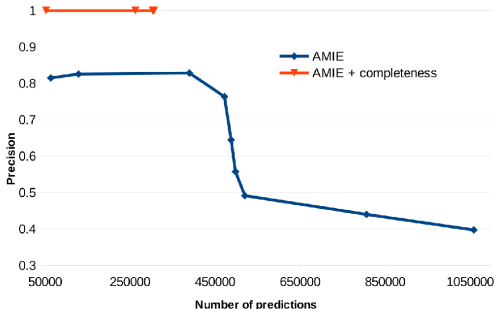
International Conference on Web Search and Data Mining (WSDM) 2017

# Missing Object Problem: Application

Fact prediction is a method that uses rules such as

$$\text{marriedTo}(x,y) \wedge \text{hasChild}(x,z) \Rightarrow \text{hasChild}(y,z)$$

to predict new facts. If we restrict fact prediction to those subjects where objects are missing, the precision increases:



>married

# Are all people married?



# Are all people married?



# Are all people married?



Obligatory for people:

- hasBirthPlace
- hasNationality

Not obligatory:

- isMarriedTo
- hasChild

**Problem:** Obligatory Attribute Problem

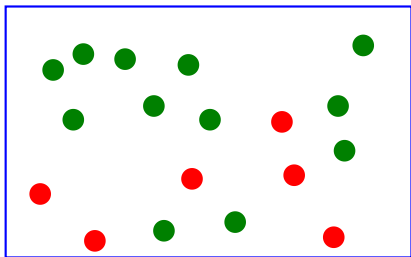
**Input:**

- a knowledge base  $K$
- a class  $c$
- a relation  $r$

**Task:** Determine if all instances of  $c$  have the relation  $r$  in the real world



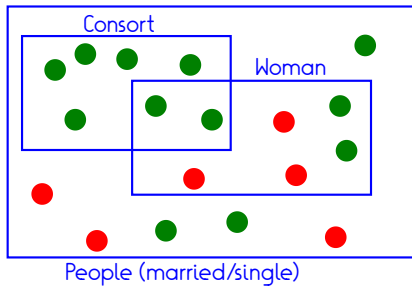
# Are all people married?



Real World

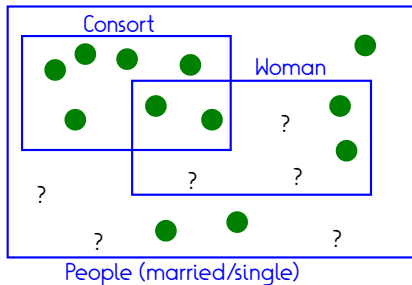
People (married/single)

# Are all people married?



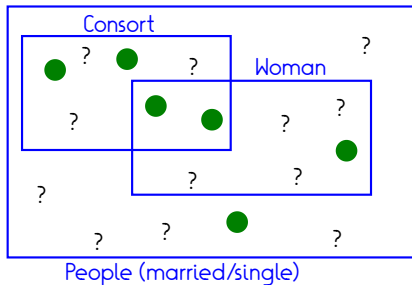
Real World

# Are all people married?



Knowledge base  
without negative facts

# Are all people married?

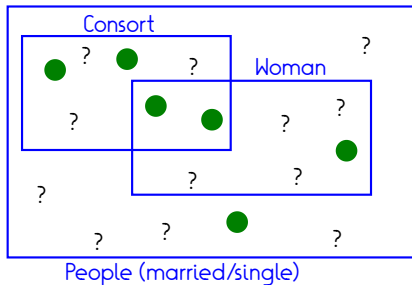


Knowledge base  
without negative facts  
and with incompleteness

In YAGO 3, only 2% of people have a nationality (obligatory attribute),  
and only 2% of people are married (non-obligatory attribute).

>assumptions

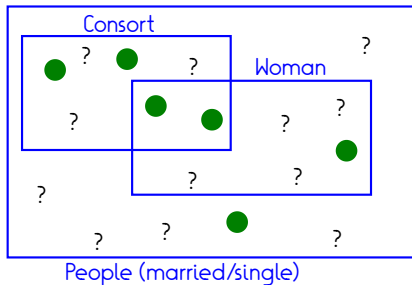
# Are all people married?



Assumptions:

- the KB is correct, i.e., every fact in the KB is in the real world
- the classes of the KB are correct and complete
- the partial completeness assumption
- the facts are a uniform random sample of the facts in the real-world

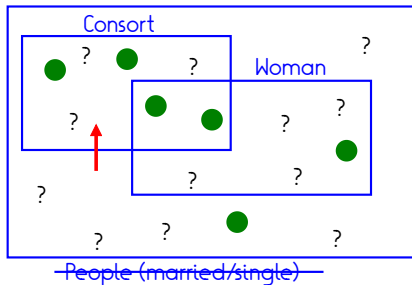
# Are all people married?



Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

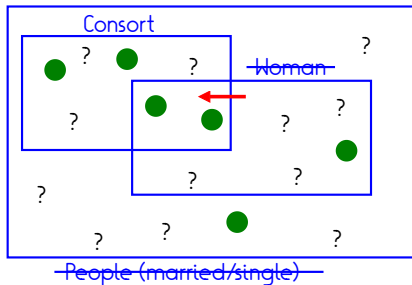
$p$  obligatory in class  $c \Rightarrow \forall c': E(\text{ratio of } p \text{ in } c \setminus c') = E(\text{ratio of } p \text{ in } c \cap c')$

# Are all people married?



Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

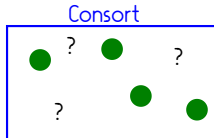
# Are all people married?



Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.



# Obligatory attributes problem



● = married in our KB

In the real world, do all instances of a class have the attribute?

=> By help of Density-difference-based estimators,  
we can predict the obligatory attributes of a class purely from the KB

(although the work does not actually predict attributes that are obligatory, but rather excludes attributes that cannot be obligatory)

Jonathan Lajus, Fabian M. Suchanek:

"Are All People Married? Determining Obligatory Attributes in KBs "

Web Conference (WWW) 2018

# Predictive recall assessment

How can we find out if a knowledge base is complete?

- Recall of facts
  - Do we have all objects for a subject?
  - Can we use text to determine completeness?
- Recall of entities
  - Do we have all entities of the real world?

# Text can help assess completeness

Marie brought her child Irène to school.

How many children does Marie have?



# Text can help assess completeness

Marie brought her child Irène to school.

How many children does Marie have?



Marie has two daughters, Irène and Ève.

How many children does Marie have?



->problem

# Text can help assess completeness

Marie brought her child Irène to school.

How many children does Marie have?



Marie has two daughters, Irène and Ève.

How many children does Marie have?



Natural language utterances imply a range of assertions that are not explicitly stated – the **implicatures** (based on work by Grice in 1975).

**Scalar implicatures** say that no more facts are true than those that are explicitly stated.

# Text Coverage Problem

Marie brought her child Irène to school.

Sentence incomplete



Marie has two daughters, Irène and Ève.

Sentence (probably) complete

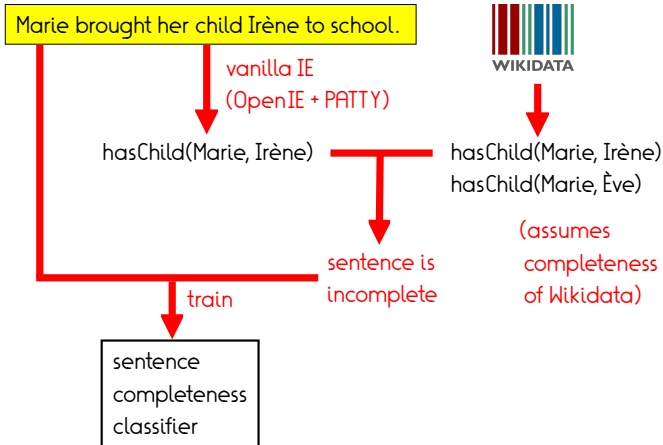


**Problem:** Text Coverage Problem

**Input:** A sentence about a subject  $s$  and a relation  $r$

**Task:** Determine if the sentence is complete, i.e.,  
if it enumerates all objects  $o$  with  $r(s,o)$ .

# Text Coverage Problem



# Text Coverage Problem

Does a sentence list all objects for a given subject and relation?

⇒ The Gricean maxims of conversation allow us to train a classifier.

What indicates completeness?

their daughters *list*  
her grandsons *list*  
his *number* children *list*

What indicates incompleteness?

her surviving (sons | daughters ...) *list*  
succeeded by her (daughters | sons ...) *list*  
in addition a (daughter | son ...) *name*

Simon Razniewski, Nitisha Jain, Paramita Mirza, Gerhard Weikum:

"[Coverage of Information Extraction from Sentences and Paragraphs](#) "

Empirical Methods in Natural Language Processing (EMNLP) 2019



# Predictive recall assessment

How can we find out if a knowledge base is complete?

- Recall of facts
  - Do we have all objects for a subject?
  - Can we use text to determine completeness?
- Recall of entities
  - Do we have all entities of the real world?

# Missing Entities Problem

Assume we're building a knowledge base about scientists:



**Problem:** Missing Entities Problem

**Input:** A set of entities of a given class

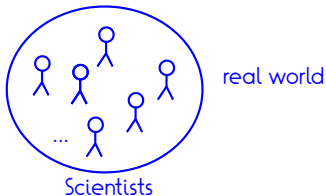
**Task:** Determine how many entities are missing compared to the real world.

But how many are there in the real world?

# Mark and recapture

**Mark-and-Recapture** is a method to estimate the number of animals in a population by capturing some animals, marking them, releasing them, capturing animals again, and observing the percentage of marked ones.

Example:



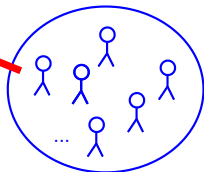
# Mark and recapture

**Mark-and-Recapture** is a method to estimate the number of animals in a population by capturing some animals, marking them, releasing them, capturing animals again, and observing the percentage of marked ones.

Example:



1. Capture  
and mark



Scientists

# Mark and recapture

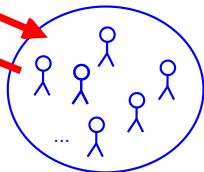
**Mark-and-Recapture** is a method to estimate the number of animals in a population by capturing some animals, marking them, releasing them, capturing animals again, and observing the percentage of marked ones.

Example:



1. Capture  
and mark

2. Release  
into the wild

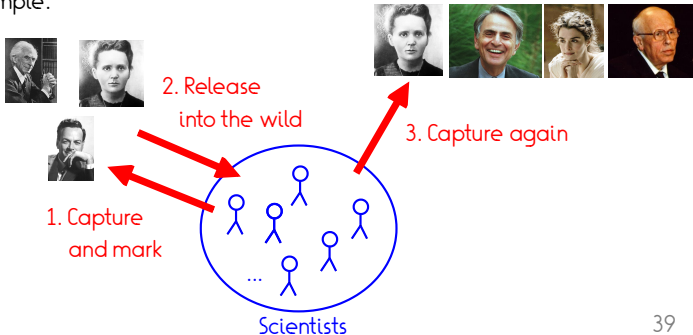


Scientists

# Mark and recapture

**Mark-and-Recapture** is a method to estimate the number of animals in a population by capturing some animals, marking them, releasing them, capturing animals again, and observing the percentage of marked ones.

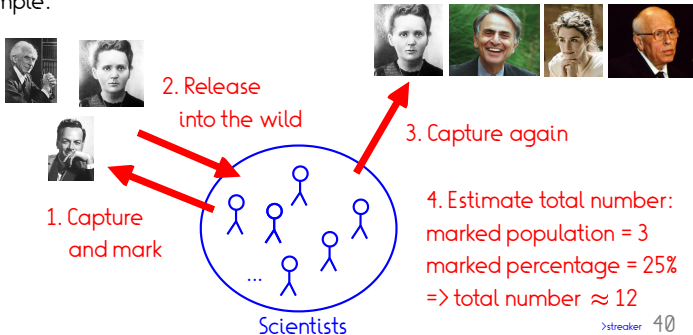
Example:



# Mark and recapture

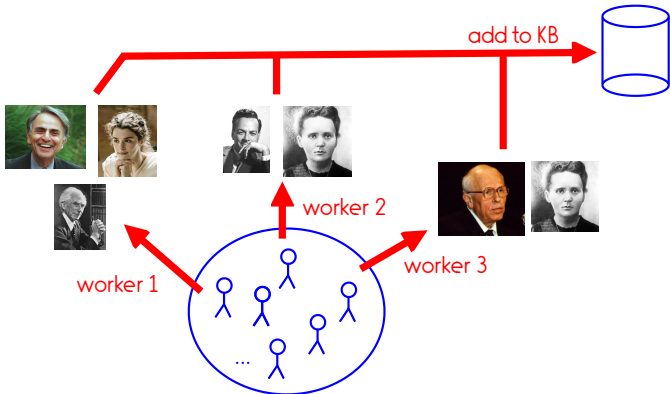
**Mark-and-Recapture** is a method to estimate the number of animals in a population by capturing some animals, marking them, releasing them, capturing animals again, and observing the percentage of marked ones.

Example:



# Crowd-sourced KBs

In a crowd-sourced KB, the workers create a "sample" from the world. The entities that appear more than once are the "re-captured" ones.

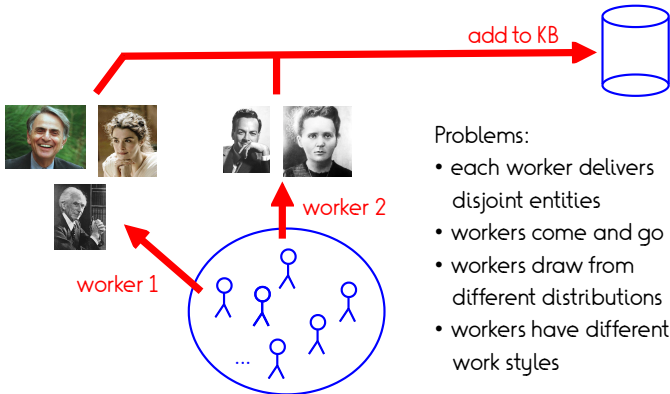


[Rachel Weisz as Hypatia from 2009 Agora movie, Bertrand Russell from Nationaal Archief, Andrei Sakharov from TheFamousPeople, Richard Feynman from Nobel Foundation, Carl Sagan from NASA] >streaker



# Crowd-sourced KBs

In a crowd-sourced KB, the workers create a "sample" from the world. The entities that appear more than once are the "re-captured" ones.



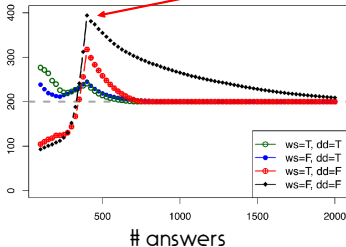
Problems:

- each worker delivers disjoint entities
- workers come and go
- workers draw from different distributions
- workers have different work styles

# The Streaker Problem

If one worker adds many (disjoint) entities in one go, the estimators over-estimate the total number of entities.

estimation  
of total  
number  
of entities



streaker arrives,  
estimators over-estimate

correct number of entities

different  
estimators

>solution

# Solving the Streaker Problem

To estimate the total number of entities at some time point, we create the multi-set of all worker responses that we received so far.

Chao92 estimator (simplified):

estimated total number:

$$\frac{c}{1 - f_1/n}$$

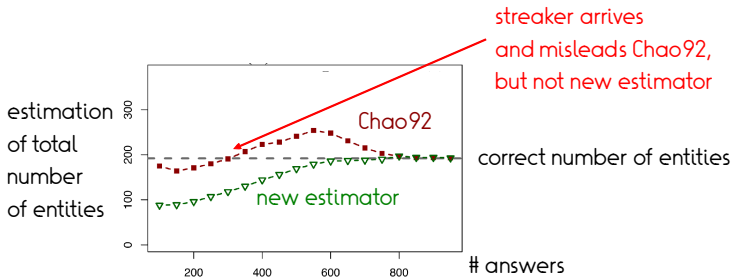
observed number  
of unique entities  
in the set

size of the set

number of entities that appear  
exactly once in the set

Idea: Replace  $f_1$  by a new number  
that ignores unique entities contributed by one worker  
beyond 2 standard deviations from the mean (= streakers).

# Solving the Streaker Problem



=> In a crowd-sourced KB, the total number of entities can be estimated

Beth Trushkowsky, Tim Kraska, Michael J. Franklin, Purnamrita Sarkar:

"[Crowdsourced Enumeration Queries](#)"

[wikidata](#)

International Conference on Data Engineering (ICDE) 2013

# Mark and recapture in Wikidata

How can we sample if the entities are already in the KB?

Idea: user edits in Wikidata "sample" from the real world.

time



sample period 1

hasChild(MarieCurie, Eve)

type(BertrandRussell, Humanist)

married(Arline, RichardFeynman)

sample period 2

hasChild(MarieCurie, Irène)

livedIn(Hypatia, Alexandria)

nationality(CarlSagan, USA)

namedAfter(SakharovPrize,...)

# Mark and recapture in Wikidata

How can we sample if the entities are already in the KB?

Idea: user edits in Wikidata "sample" from the real world.

time



sample period 1

hasChild(MarieCurie, Eve)

type(BertrandRussell, Humanist)

married(Arlene, RichardFeynman)

sample period 2

hasChild(MarieCurie, Irène)

livedIn(Hypatia, Alexandria)

nationality(CarlSagan, USA)

namedAfter(SakharovPrize,...)

Sample 1



Sample 2



# Mark and recapture in Wikidata

$k$  = number of sample periods (here: 2)

$n$  = number of observations (here: 7)

$c$  = current number of entities in the KB

$f_i$  = frequency of entities observed  $i$  times (here:  $f_1=5$ )

Try several estimators, e.g.

- JackKnife:  $c + \frac{k-1}{k} f_1$
- Streaker
- Chao92:  $\frac{c}{1-f_1/n} [\dots]$

Sample 1



Sample 2



# Estimators: Jackknife

$k$  = number of sample periods (here: 2)

$n$  = number of observations (here: 7)

$c$  = current number of entities in the KB

$f_i$  = frequency of entities observed  $i$  times (here:  $f_1=5$ )

Jackknife estimator: The number of unseen entities is the number of distinct entities seen in one sample period  $j$  ( $f_1^j$ ), multiplied by the number of other samples ( $k-1$ ). Average across all sample periods:

$$\text{Jackknife} = \text{AVERAGE}_{j=1..k}: (k-1)f_1^j + D$$

Sample 1



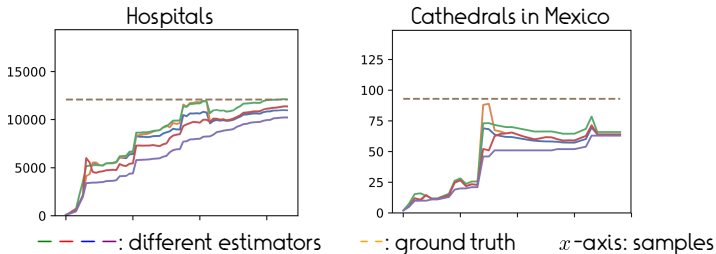
Sample 2





# Missing Entities Problem in edited KB

Can we estimate the total number of entities in the real world?



=> If the population is large, its size can be estimated

M. Luggen, D. Difallah, C. Sarasua, G. Demartini, P. Cudré-Mauroux:

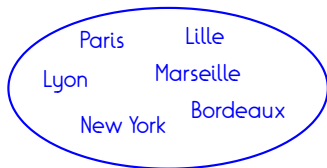
"Non-Parametric Class Completeness Estimators for Collaborative KGs "

International Semantic Web Conference (ISWC) 2019

>representativeness

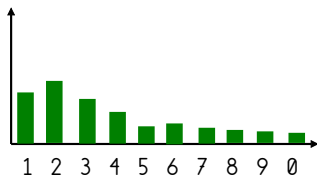
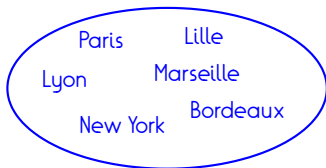
# Missing Entities Problem in static KBs

If the KB is static, the mark-and-recapture estimators do not work.



# Missing Entities Problem in static KBs

How can we estimate the missing entities in a static KB?



- 1) Take the number of inhabitants of each city
- 2) Take the first digit
- 3) Plot the number of cities per first digit

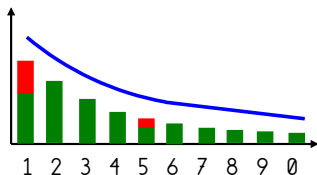
# Missing Entities Problem in static KBs

Benford's Law says that the first digit  $d$  appears with probability

$$\log_{10}\left(1 + \frac{1}{d}\right)$$

=> We can "fill up" the missing digits

It is also possible to parameterize the law, and learn the parameter.



[>details](#)

# Benford's Law explained

Benford's Law says that the first digit  $d$  appears with probability

$$\log_{10}\left(1 + \frac{1}{d}\right)$$

This holds only for quantities that grow by multiplicative factors:

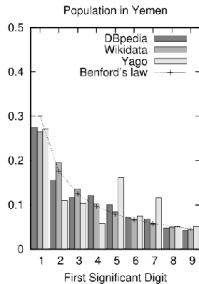
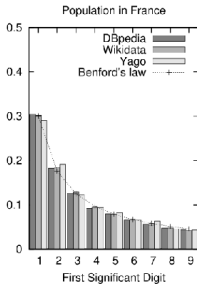
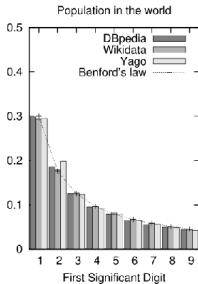
- number of inhabitants of cities
- the size of a lake
- other natural processes

Illustration: inhabitants of a village that grows by 50% each year

1000	5062
1500	7593
2250	11390
3375	...

[>details](#)

# Benford's Law examples



Not representative

[>details](#)

# Parameterized Benford's Law

A set of numbers satisfies a generalized Benford's law with exponent  $\alpha$ , if the first digit  $d \in [1..9]$  occurs with probability

$$B_d^\alpha = \frac{(1+d)^\alpha - d^\alpha}{10^\alpha - 1}$$

1. Transform a relation to a numerical relation, e.g.,  
by counting the number of objects:  $numMovies(x) = \# y: actedIn(x,y)$
2. Determine  $\alpha$  by a weighted least square measure
3. Run a MAD (Mean Absolute Deviation) test to see if Benford's Law could be applied
4. If so, compute number of entities that have to be added to conform to Benford's Law

# Missing Entities Problem in static KBs

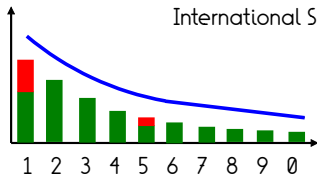
How can we know how many entities are missing in our KB,  
if the KB is static (i.e., not updated by edits)?

=> Benford's Law allows us to give a minimum numbers of entities that  
are missing to make the distribution representative of the real world.

A. Soulet, A. Giacometti, B. Markhoff, F. M. Suchanek:

"[Representativeness of KBs with Benford's Law](#)"

International Semantic Web Conference (ISWC) 2018





# Takeaway: Predictive recall assessment

Using statistical techniques, we can predict more or less:



marriedTo

→ ∃ ?

Are we missing objects in the KB?  
(Supervised learning of rules)

Neil came with his wife Alice.

Does a text enumerate all objects?  
(Train a classifier based on  
Grice's maxims of conversation)



How many entities are missing?  
(Mark and recapture)

# 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 and Foundations (Simon) – 10 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

# What is count information?

Relation between an entity and a set of entities



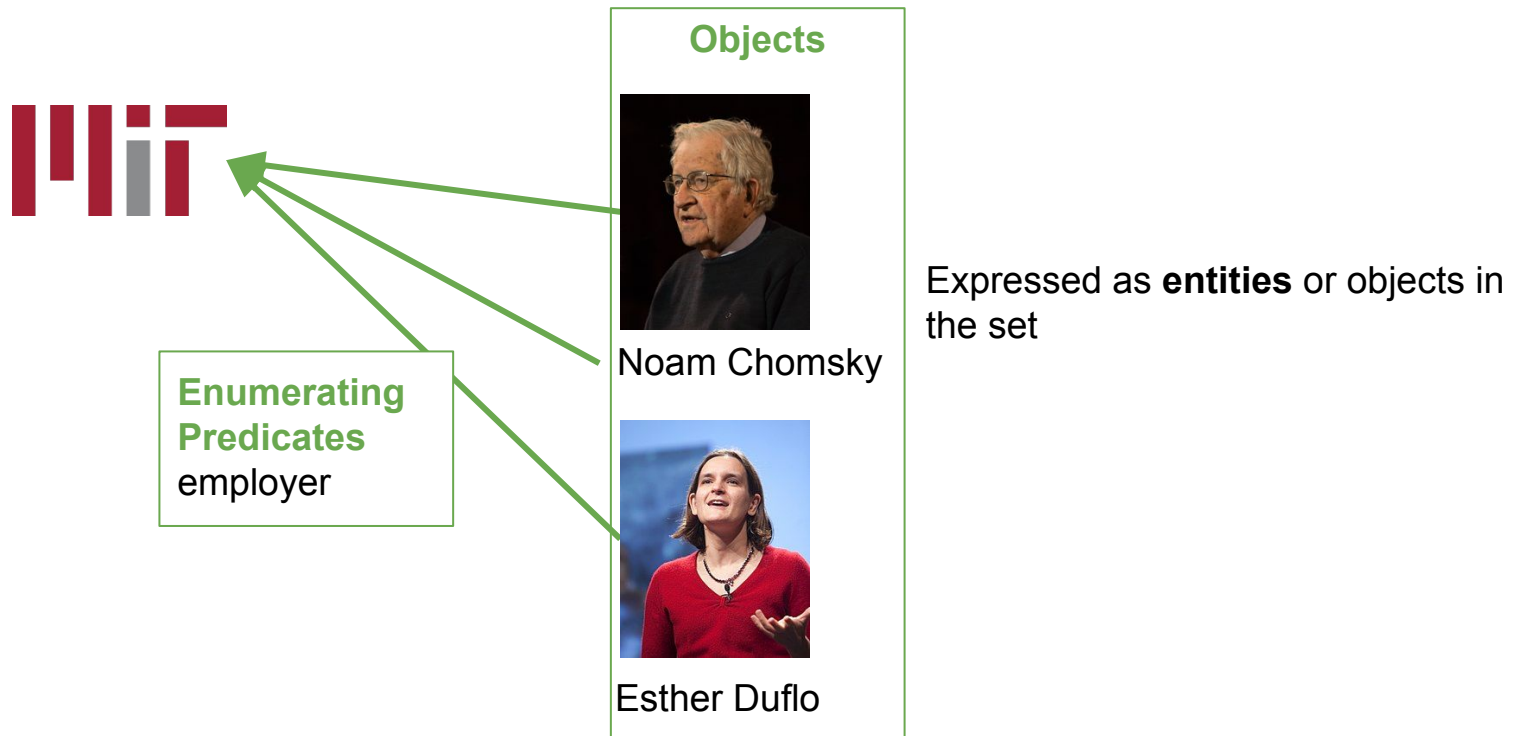
Noam Chomsky



Esther Duflo

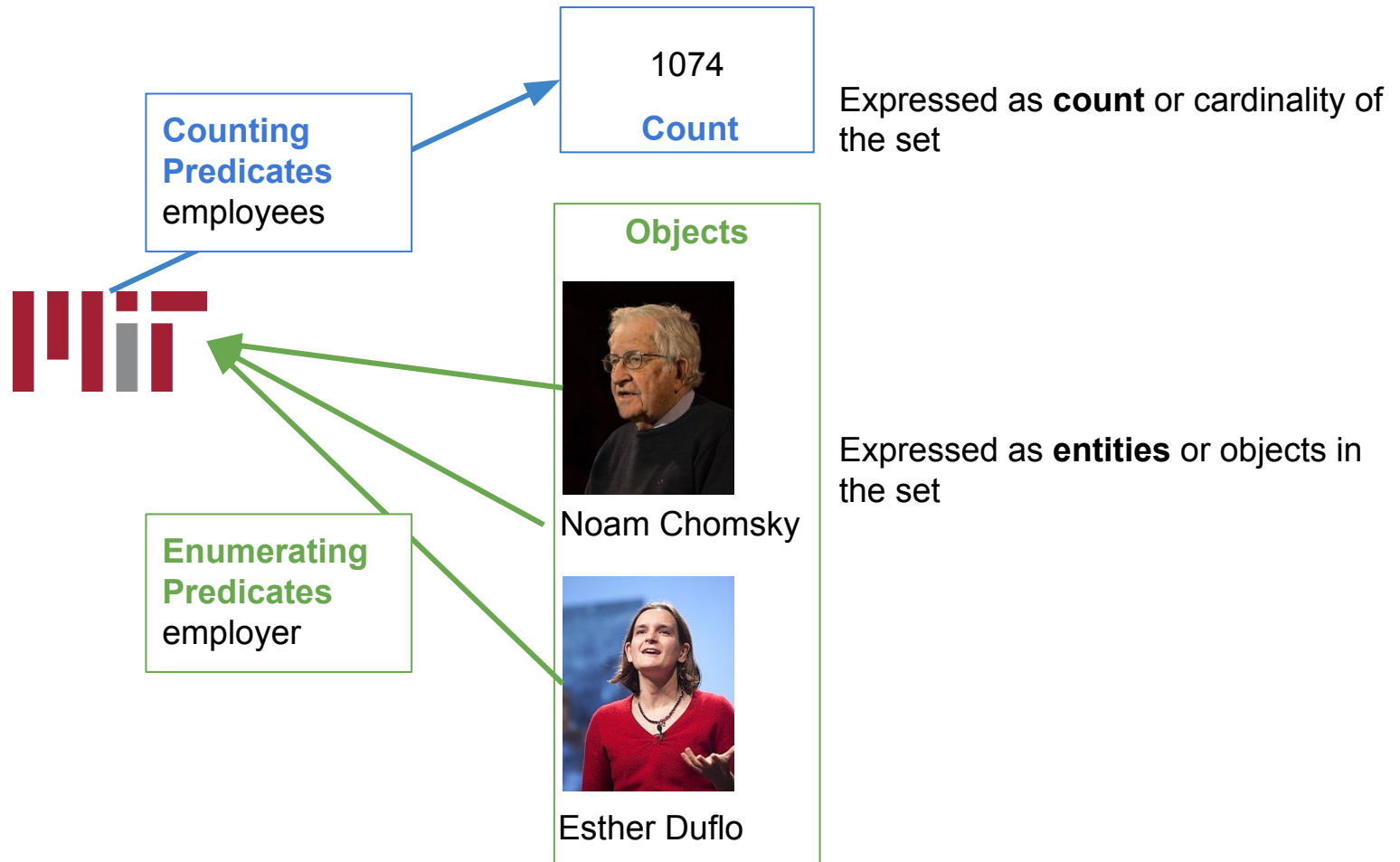
# What is count information?

Relation between an entity and a set of entities



# What is count information?

Relation between an entity and a set of entities



1. Count information for recall assessment
2. How can we extract count information from text?
3. Variants of count information in KB
4. Counts for KB curation

# Count information for recall assessment

Counts and entities benefit from each other

## Only entities

(?x, **employer**, MIT)

returns a handful of names from KB



**Enumerating  
Predicates**  
employer

## Objects



Noam Chomsky



Esther Duflo

# Count information for recall assessment

Counts and entities benefit from each other

## Only entities

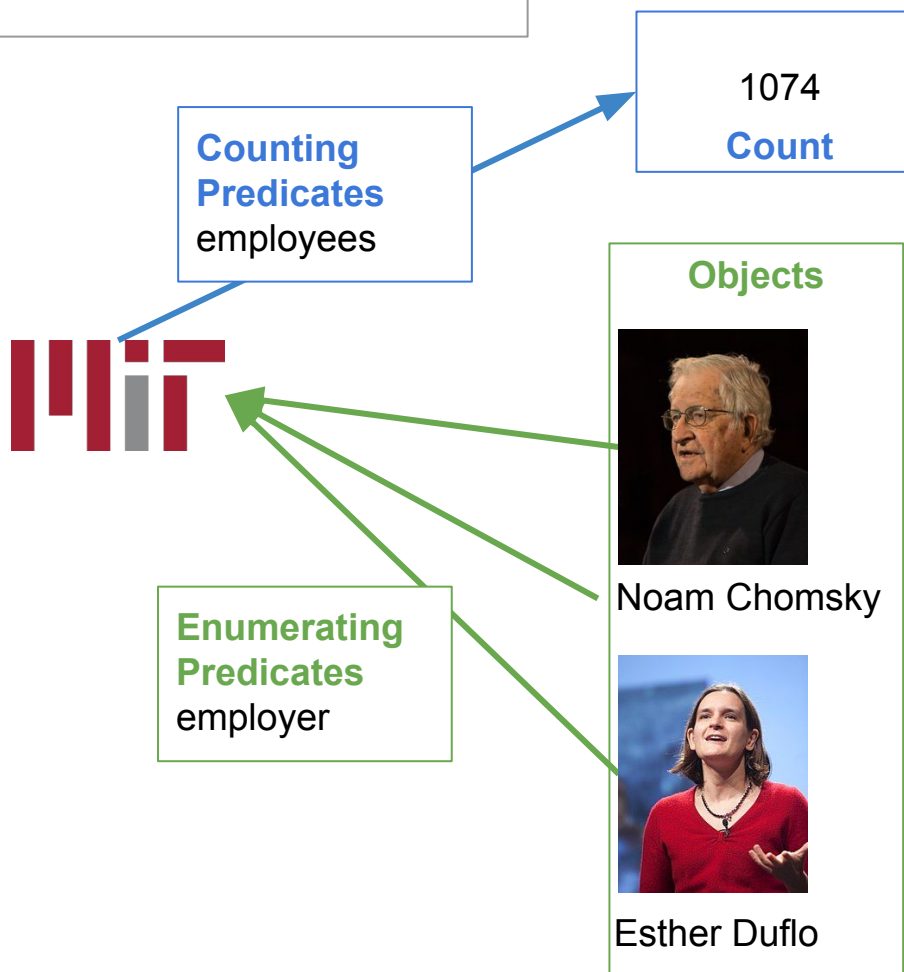
(?x, **employer**, MIT)

returns a handful of names from KB

## Only counts

(MIT, **employees**, ?y)

gives no insight about the entities





# Count information for recall assessment

Counts and entities benefit from each other

## Only entities

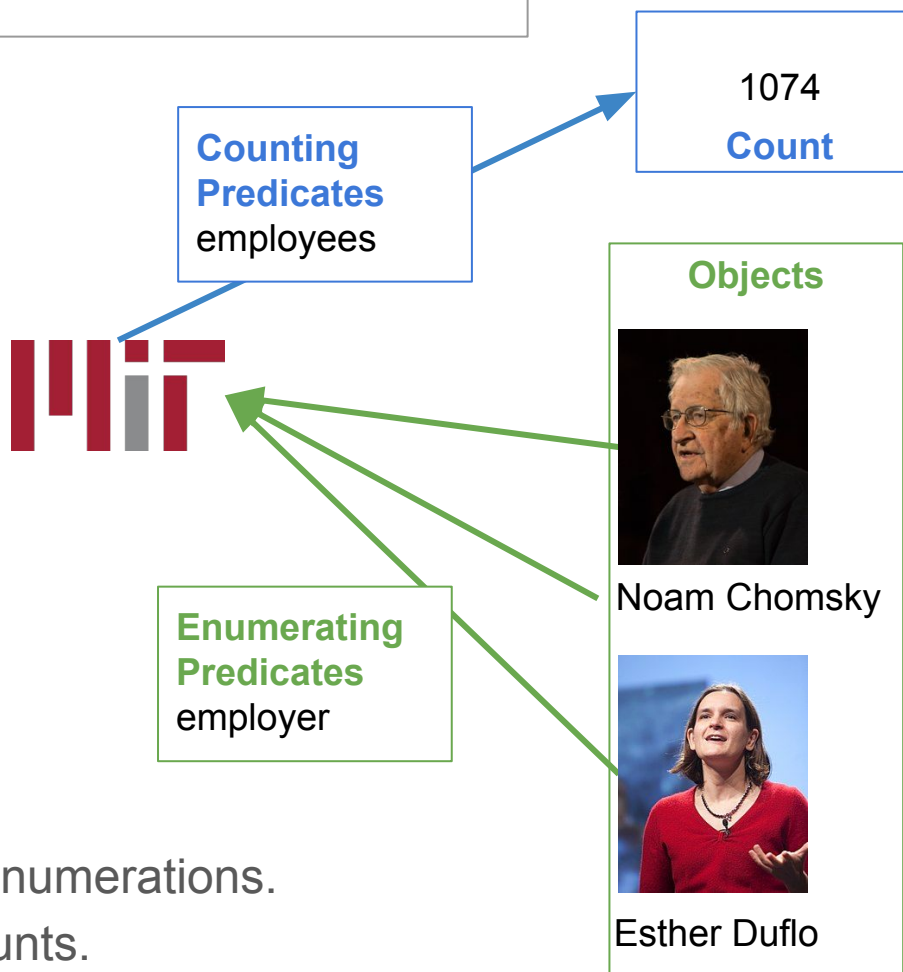
(?x, **employer**, MIT)

returns a handful of names from KB

## Only counts

(MIT, **employees**, ?y)

gives no insight about the entities



- Counts enhance incomplete entity enumerations.
- Representative entities enhance counts.

# Count information for recall assessment

KB mixes counts with standard facts



number of children

2

Tim Berners-Lee

How many children does Tim Berners-Lee have?

2 (KB fact)



child

Anne Blunt

Ralph King-Milbanke

Byron King-Noel

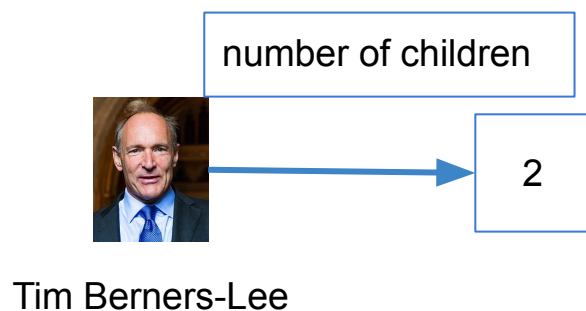
Ada Lovelace

How many children did Ada Lovelace have?

3 (Maybe?)

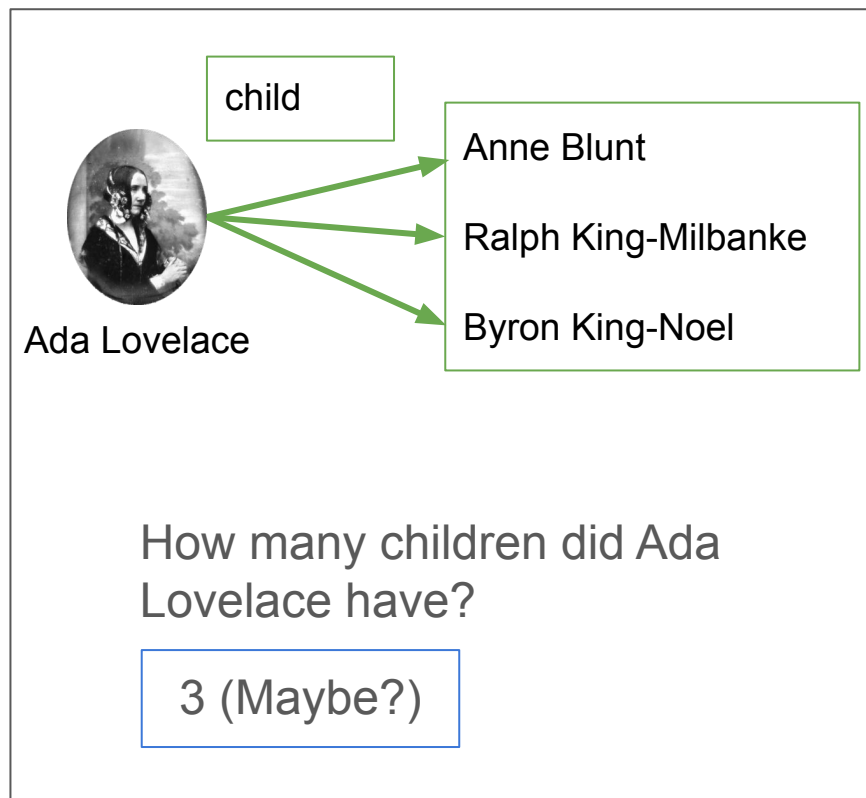
# Count information for recall assessment

KB mixes counts with standard facts



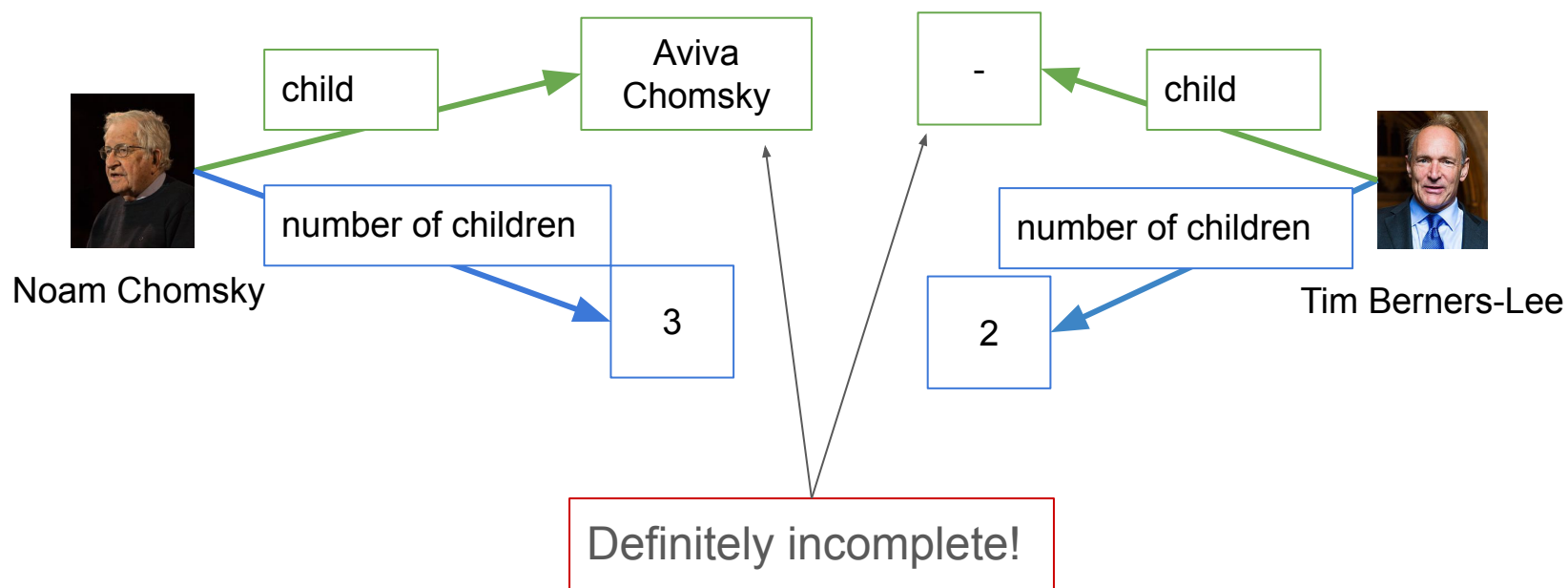
How many children does Tim Berners-Lee have?

2 (KB fact)



# Count information for recall assessment

Improving KB recall



1. Count information for recall assessment
2. How can we extract count information from text?
3. Variants of count information in KB
4. Counts for KB curation

# Count information from text

**Problem:** Counting Quantifier Extraction

**Input:**

- a text about a subject S
- a predicate P

**Task:** Determine the number of objects in which S stands in relation with P

**Subject:** Noam Chomsky

**Predicate:** number\_of\_children



Chomsky was married to Carol. They had **three children** together

**3**



# Count information from text

**Task 1:** Identify the **count tokens** and the **compositional cues**.

**Sequence Labelling of tokens** in a sentence on subject S and predicate P with:

- COUNT - for counts
- COMP - for compositional cues
- O - all other tokens



**Subject:** Noam Chomsky  
**Predicate:** number\_of\_children

Chomsky was married to Carol. They had **three** children together

O O O O O O O **COUNT** O O



# Count information from text

**Task 1:** Identify the **count tokens** and the **compositional cues**.

**Sequence Labelling of tokens** in a sentence on subject S and predicate P with:

- COUNT - for counts
- COMP - for compositional cues
- O - all other tokens



**Subject:** Angelina Jolie

**Predicate:** number\_of\_children

Jolie	has	three	sons	and	three	daughters.
O	O	COUNT	O	COMP	COUNT	O





# Count information from text

**Task 1:** Identify the **count tokens** and the **compositional cues**.

**Linguistic Diversity** while expressing counts in texts

## Cardinals

two sons,  
three books

## Ordinals

second son,  
third book

## Number-related terms

twins, trilogy

## Indefinite Articles

a son,  
the book

**Compositional cues** for counts

- consecutive count tokens
- comma-separated, and-separated counts

**Subject:** Angelina Jolie

**Predicate:** number\_of\_children

Jolie brought her **twins** , **one** daughter **and** **three** adopted children to the gala.

COMP

COMP

# Count information from text

## Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. **Sum up compositional cues**
2. Select prediction per type
3. Rank mention types

6

Jolie brought her **six** children: **twins** , **one** daughter **and** **three** adopted children to the gala.

---

**Subject:** Angelina Jolie

**Predicate:** number\_of\_children

# Count information from text

## Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. Sum up compositional cues
2. **Select prediction per type**
3. Rank mention types

6 (cardinal)

6 (cardinal)

Jolie brought her **six** children: **twins** , **one** daughter **and three** adopted children to the gala.

---

**Subject:** Angelina Jolie

**Predicate:** number\_of\_children

6 (cardinal)

# Count information from text

## Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. Sum up compositional cues
2. Select prediction per type
3. **Rank mention types**

cardinal	>>	number-related terms	>>	ordinals	>>	indefinite article
two children	>>	twins	>>	second child	>>	a child

Jolie brought her **six** children: **twins** , **one** daughter **and** **three** adopted children to the gala.

**Subject:** Angelina Jolie

**Predicate:** number\_of\_children

**6 (cardinal)**

# Count information from text

Relation	Baseline [22]			CINEX-CRF			CINEX-CRF (per type)					
							Cardinals		Numt.+Art.		Ordinals	
	P	Cov	MAE	P	Cov	MAE	P	Contr	P	Contr	P	Contr
containsWork	42.0	<b>29.0</b>	3.7	<b>49.2</b>	<b>29.0</b>	<b>2.6</b>	55.0	33.9	62.5	40.7	20.0	25.4
hasMember	11.8	6.0	3.8	<b>64.3</b>	<b>18.0</b>	<b>1.2</b>	62.5	28.6	65.0	71.4	0	0
containsAdmin	51.8	14.5	7.3	<b>78.6</b>	<b>22.0</b>	<b>1.7</b>	85.7	87.5	33.3	10.7	0	1.8
hasChild	37.0	<b>22.0</b>	<b>2.2</b>	<b>50.0</b>	19.5	2.3	67.3	70.5	6.3	20.5	14.3	9.0
hasSpouse	26.8	11.0	1.3	<b>58.1</b>	<b>12.5</b>	<b>0.5</b>	75.0	18.6	43.8	37.2	63.2	44.2
hasZeroChild				92.3	18.8	-						
hasZeroSpouse				71.9	13.7	-						

Performance of CINEX in consolidation of counting quantifier mensions on Wikidata.

Paramita Mirza, Simon Razniewski, Fariz Darari, Gerhard Weikum

[Enriching Knowledge Bases with Quantifiers](#)

International Semantic Web Conference (ISWC) 2018.

- Count information for recall assessment
- How can we extract count information from text?
- Variants of count information in KB
- Counts for KB curation

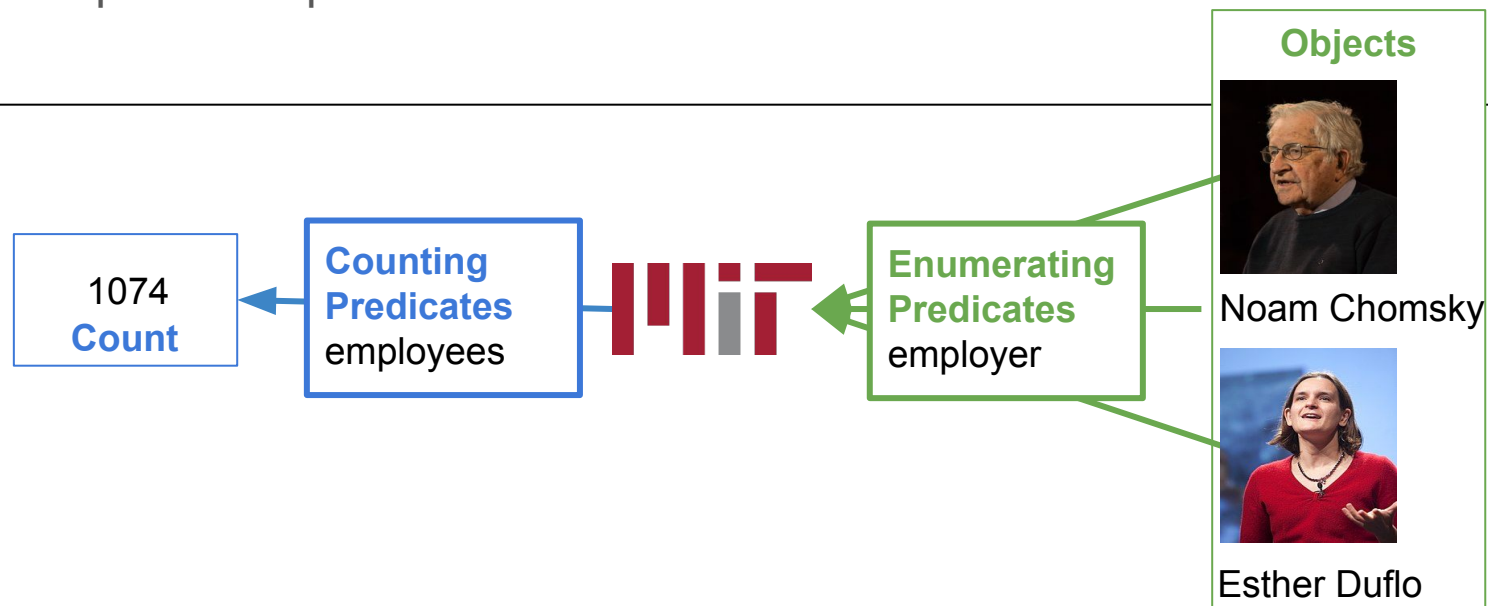
# Count information in KB

**Problem:** Identification of semantically related count predicates

**Input:**

- a set of knowledge base triples  $(s,p,o)$

**Task:** Determine counting and enumerating predicates and semantically related predicate pairs.



# Count information in KB

**Task 1:** Identify the two variants of count predicates

## Counting Predicates

academic\_staff, staff,  
faculty

number\_of\_children

...

wins, doubles\_titles,  
singles\_titles

## Enumerating Predicates

work\_institution<sup>-1</sup>, workplace<sup>-1</sup>,  
work\_institutions<sup>-1</sup>

child

...

gold<sup>-1</sup>



# Count information in KB

## Counting Predicates

academic_staff, staff, faculty	number_of_children	...	wins, doubles_titles, singles_titles
-----------------------------------	--------------------	-----	---

## Enumerating Predicates

work_institution <sup>-1</sup> , workplace <sup>-1</sup> , work_institutions <sup>-1</sup>	child	...	gold <sup>-1</sup>
---	-------	-----	--------------------

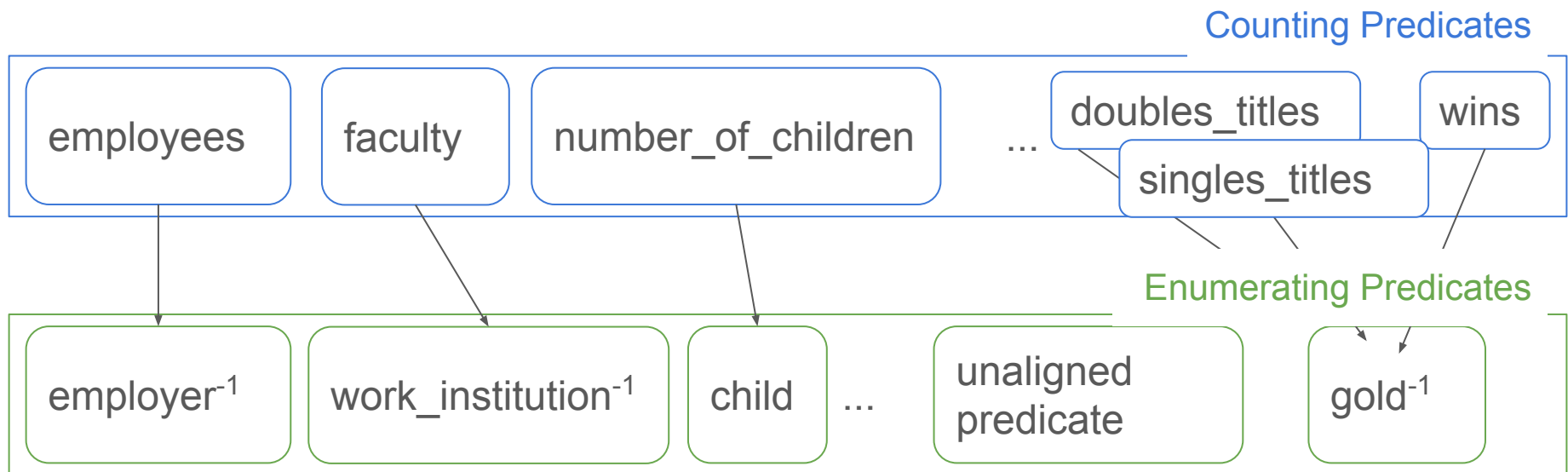
Supervised Classification using:

- **Textual Features** - count predicates are more often used in singular form
- **Type Information** - classes of subject and objects
- **KB statistics** - #objects per subject, datatype distribution of the objects

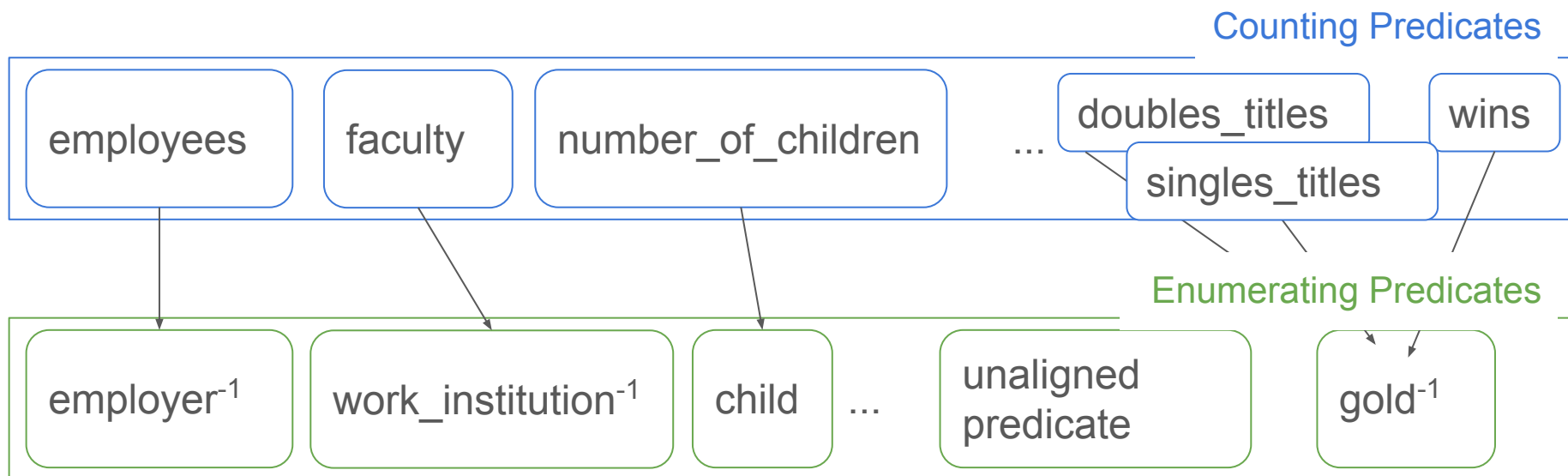
# Count information in KB

**Task 2: Heuristically align** predicates from two classes

Generate **pairwise alignment** scores



# Count information in KB



Heuristics used for the predicate pair  $(\mathbf{e}, \mathbf{c})$ , where  $\mathbf{e}$  stores entities and  $\mathbf{c}$  counts.

1. Predicate pair co-occurrences - #subjects  $\mathbf{e}$  and  $\mathbf{c}$  co-occur
2. Value distribution - number of objects of  $\mathbf{e}$  compared to count in  $\mathbf{c}$
3. Linguistic similarity - do  $\mathbf{e}$  and  $\mathbf{c}$  talk share topical similarity?

# Count information in KB

KB	Input	Output	Filtered
DBP-raw	16,635	4,090	4,090 (24.5%)
DBP-map	1,670	308	308 (18.4%)
WD-truthy	4,067	216	203 (4.9%)
Freebase	13,872	7,752	7,614 (54.8%)
<b>Total</b>	<b>36,244</b>	<b>12,366</b>	<b>12,215 (33.7%)</b>

Model	Recall	Precision	F1
Random	40.6	40.6	40.6
Logistic	<b>55.6</b>	51.7	53.5
Prior	<b>55.6</b>	51.0	53.5
Lasso	51.1	<b>59.6</b>	<b>55.0</b>
Neural	53.0	49.6	51.2

Enumerating predicates extracted by CounQER and their scores

KB	Input	Output	Filtered
DBP-raw	13,394	5,853	5,853 (43.6%)
DBP-map	1,127	898	898 (79.6%)
WD-truthy	3,346	1,922	1,067 (31.8%)
Freebase	8,289	1,723	1,687 (20.3%)
<b>Total</b>	<b>26,156</b>	<b>10,396</b>	<b>9,505 (36.3%)</b>

Model	Recall	Precision	F1
Random	12.8	12.8	12.8
Logistic	51.2	19.0	27.7
Prior	48.7	20.2	28.5
Lasso	<b>71.7</b>	<b>23.3</b>	<b>35.1</b>
Neural	35.8	20.8	26.3

Counting predicates extracted by CounQER and their scores

Shrestha Ghosh, Simon Razniewski, Gerhard Weikum

[Uncovering Hidden Semantics of Set Information in Knowledge Bases](#)

Journal of Web Semantics (JWS) 2020.

# Counts from text and KB

- What is count information?
- Count information for recall assessment
- How can we extract count information from text?
- Variants of count information in KB
- Counts for KB curation

# Counts for KB curation

The screenshot displays the COUNQER web interface. At the top, there is a header bar with a refresh icon, an 'Entity' field containing 'Noam Chomsky: Amer', a 'Set Predicate' button, and a 'P40: child (1)' field. Below this, a green notification box states '!! Hope the results satisfy your curiosity!' with a close button. The main content area shows two rows of results. The first row displays 'Noam Chomsky' on the left, a blue button labeled 'P40: child' in the center, and 'Aviva Chomsky' on the right, with a share icon to its right. The second row is titled 'Related Counting Predicates' and shows 'Noam Chomsky' on the left, an orange button labeled 'P1971: number of children' in the center, and '(no instantiations)' on the right, also with a share icon. Two arrows from an external text box point to the 'Aviva Chomsky' and '(no instantiations)' results, indicating they are KB inconsistencies.

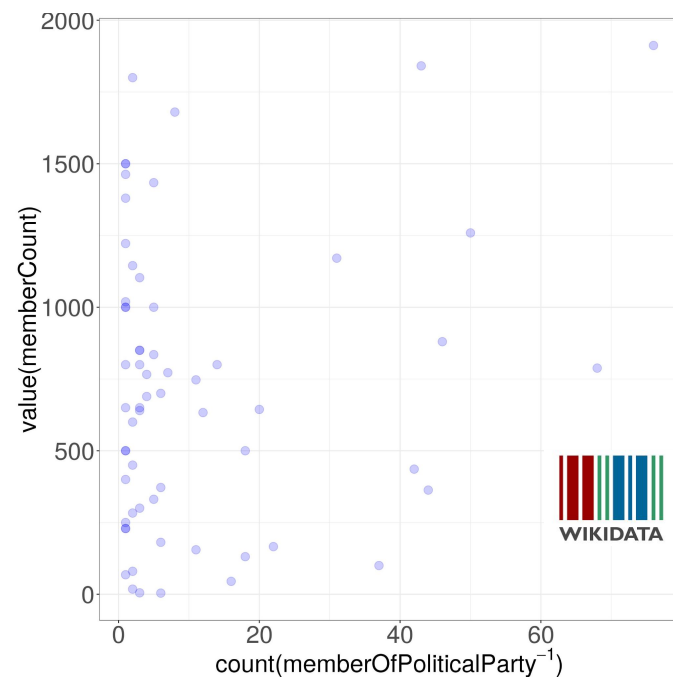
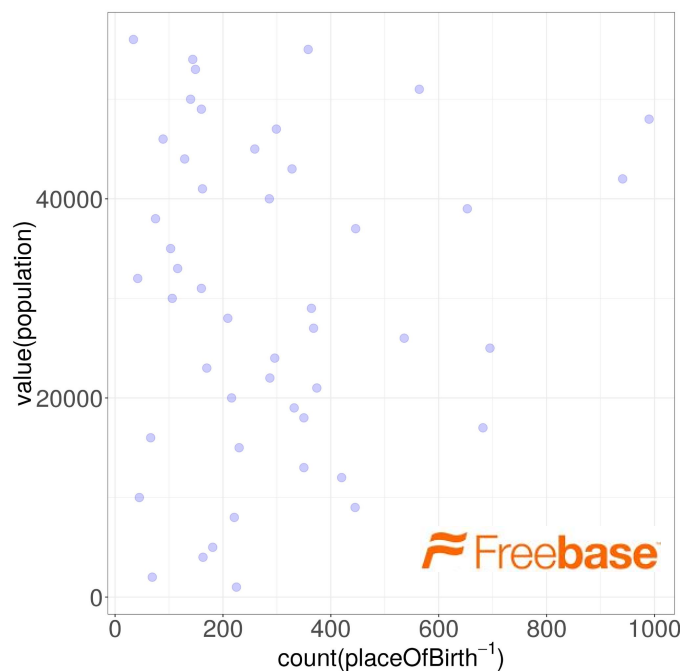
Entity	Predicate	Count
Noam Chomsky	P40: child	1
Noam Chomsky	P1971: number of children	0

KB inconsistencies are highlighted

<https://counqer.mpi-inf.mpg.de/spo>

# Counts for KB curation

Value distribution of aligned predicates show incompleteness



# Takeaway: Counts from text and KB

1. Count information for recall assessment
  - Counts and entities benefit from each other
  - KB mixes counts with standard facts
  - Counts can improve KB recall
2. Count information in text
  - is linguistically diverse (cardinals, ordinals, ..)
  - used to get the #objects for a given subject and predicate
3. Count information in KBs
  - can be identified by supervised classification
  - occurs as semantically related counting and enumerating predicates
4. KB curation using counts
  - highlights inconsistencies
  - gives value distribution of aligned predicates



# 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



## 42 awards

Adams Prize

Albert Einstein Medal

Pius XI Medal

Oskar Klein Medal

Michelson–Morley Award

Hughes Medal

Royal Society Science Books Prize

Wolf Prize in Physics



## 42 awards

Adams Prize

Albert Einstein Medal

Pius XI Medal

Oskar Klein Medal

Michelson–Morley Award

Hughes Medal

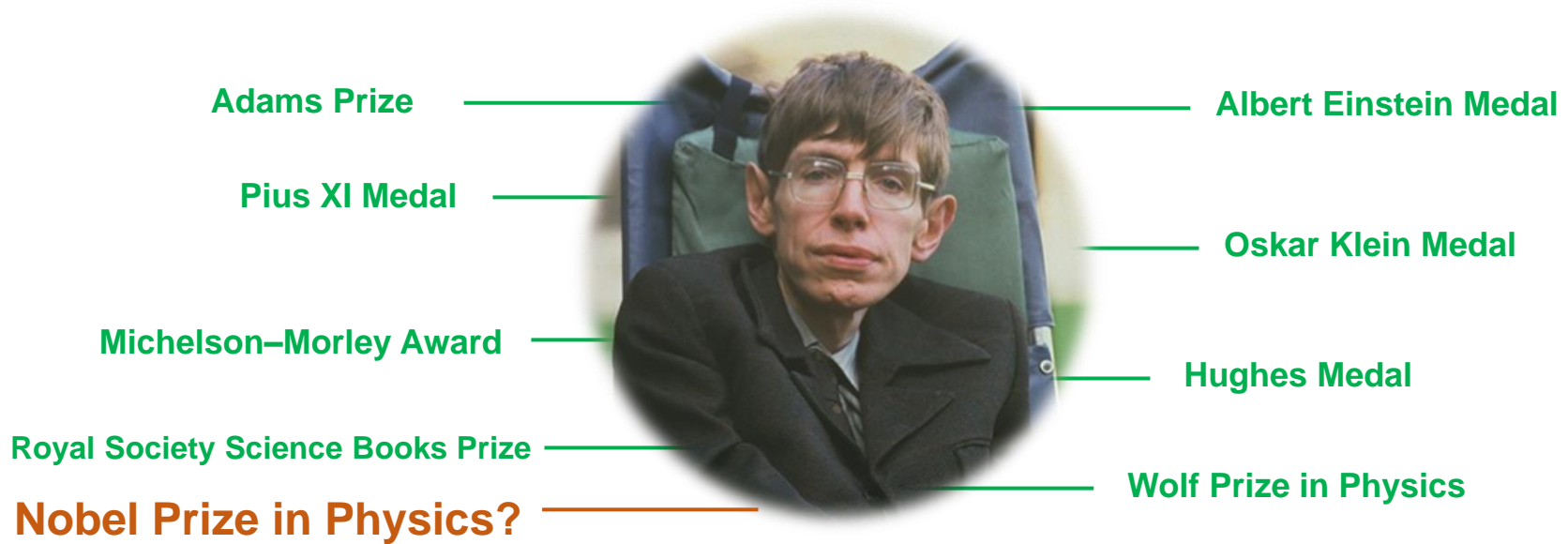
Royal Society Science Books Prize

Wolf Prize in Physics

**Nobel Prize in Physics?**



## 42 awards



**Wikidata doesn't know!**

## 42 awards



Existing positive-only KBs are unaware of negation.

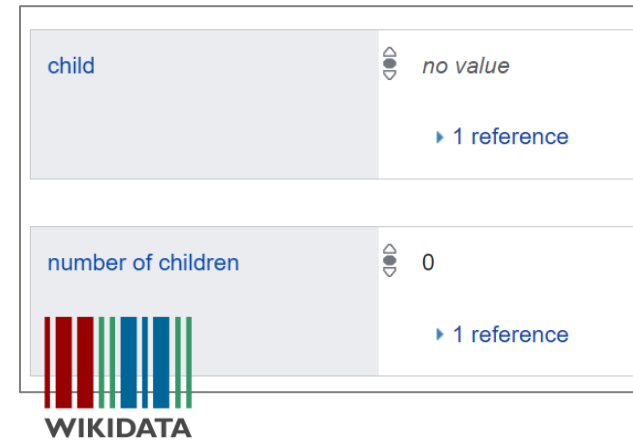
## 42 awards, 30000 awards



Existing positive-only KBs are unaware of negation.  
Set of negative statements is quasi-infinite!

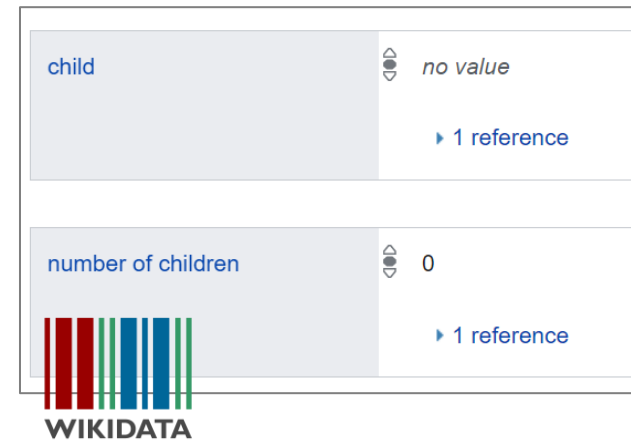
- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications

- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications
- **Count predicates**
  - instance-based predicates?





- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications
- **Count predicates**
  - instance-based predicates?
- **Negated predicates**
  - DBpedia e.g., never exceed alt (for airplanes)
  - Knowlife e.g., not caused by, not healed by



- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications
- **Count predicates**
  - instance-based predicates?
- **Negated predicates**
  - DBpedia e.g., never exceed alt (for airplanes)
  - Knowlife e.g., not caused by, not healed by
- **Object = No-value**



- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications
- **Count predicates**
  - instance-based predicates?
- **Negated predicates**
  - DBpedia e.g., never exceed alt (for airplanes)
  - Knowlife e.g., not caused by, not healed by
- **Object = No-value**
- **Deprecated rank**
  - statements that are known to include errors



- **Edit history**
  - Collaborative KBs, e.g., Wikidata
  - Deleted statements
  - 82% ontology modifications
- **Count predicates**
  - instance-based predicates?
- **Negated predicates**
  - DBpedia e.g., never exceed alt (for airplanes)
  - Knowlife e.g., not caused by, not healed by
- **Object = No-value**
- **Deprecated rank**
  - statements that are known to include errors



**Advantages:** formalizes syntax for explicit negation addition, & some allows querying them (e.g., Wikidata SPARQL with `o = no-value`)

**Limitations:** inherit same challenges from positive KBC, covers small domains, no active collection of useful negations

**Problem:**

Existing positive-only KBs are unaware of negation.

**Input:**

Open-world KB.

**Task:**

Explicitly add *salient* negative statements to KB.

# Identify Interesting Negative Knowledge

## Problem:

Existing positive-only KBs are unaware of negation.

## Input:

Open-world KB.

## Task:

Explicitly add *salient* negative statements to KB.

¬ (award; Nobel Prize in Physics)



¬ (award; Academy Awards for Best Actress)

¬ (headquarters location; Silicon Valley)



# How to identify interesting negation?

## **PART1: Statistical Inferences**

## **PART2: Text Extraction**

## **PART3: Pretrained Language Models**

## PART1: Statistical Inferences

- ★ Infer from *existing* positive statements:  
Peer-based negation inference method.

## PART2: Text Extraction

## PART3: Pretrained Language Models



# **PART1: Statistical Inferences**

## **Peer-based Negation Inference**

### **Input:**

**Given entity  $e$  from KB.**

### **Overview:**

- 1. Peer-based candidate retrieval**
- 2. Correctness filtering by local completeness assumption**
- 3. Supervised ranking for higher saliency**

### **Output:**

**Top interesting negative statements about  $e$ .**

What is a similar entity (peer) ?

What is a similar entity (peer) ?

Class-based

- **Stephen Hawking: Physicist**

What is a similar entity (peer) ?

Class-based

- **Stephen Hawking: Physicist**

Jaccard-similarity

- predicate-object pairs shared by entities:  
**Hawking AND Einstein = 423/750**

What is a similar entity (peer) ?

## Class-based

- **Stephen Hawking: Physicist**

## Jaccard-similarity

- predicate-object pairs shared by entities:  
**Hawking AND Einstein = 423/750**

## Embedding-based similarity

- Cosine of low-dimensional latent representations  
**Wikipedia embeddings**

What is a similar entity (peer) ?

## Class-based

- **Stephen Hawking: Physicist**

## Jaccard-similarity

- predicate-object pairs shared by entities:  
**Hawking AND Einstein = 423/750**

## Embedding-based similarity

- Cosine of low-dimensional latent representations  
**Wikipedia embeddings**

## Confounding factors:

- Popularity
- Sequences

What is a similar entity (peer) ?

## Class-based

- **Stephen Hawking: Physicist**

## Jaccard-similarity

- predicate-object pairs shared by entities:  
**Hawking AND Einstein = 423/750**

## Embedding-based similarity

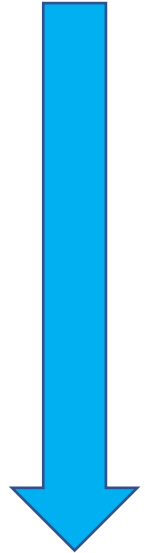
- Cosine of low-dimensional latent representations  
**Wikipedia embeddings**

## Confounding factors:

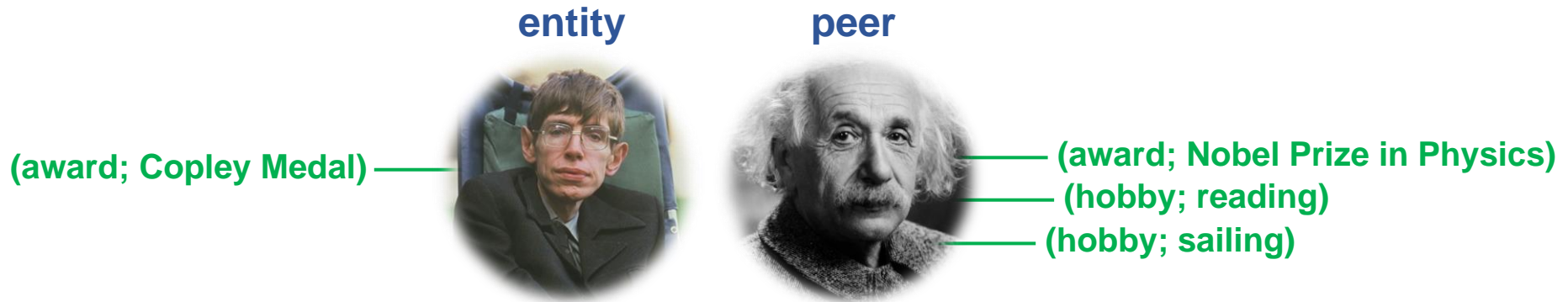
- Popularity
- Sequences



Interpretable

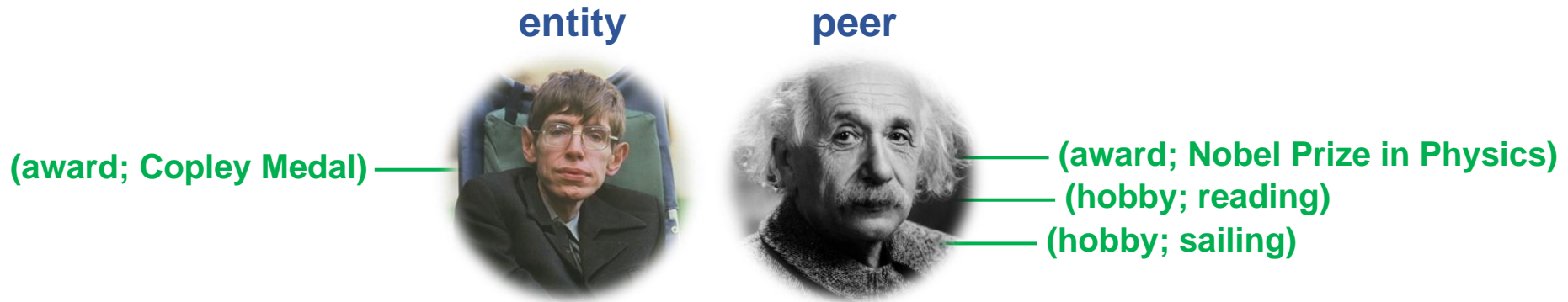


Accurate

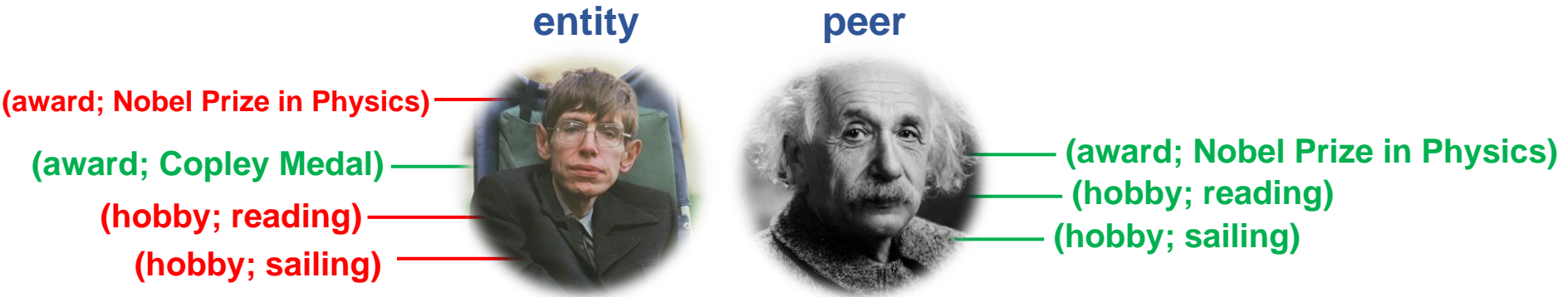




Every statement that applies to at least one peer is a *candidate negation*.



Every statement that applies to at least one peer is a *candidate negation*.



Every statement that applies to at least one peer is a *candidate negation*.



Challenge: *correctness* of inferred negations.

Every statement that applies to at least one peer is a *candidate negation*.



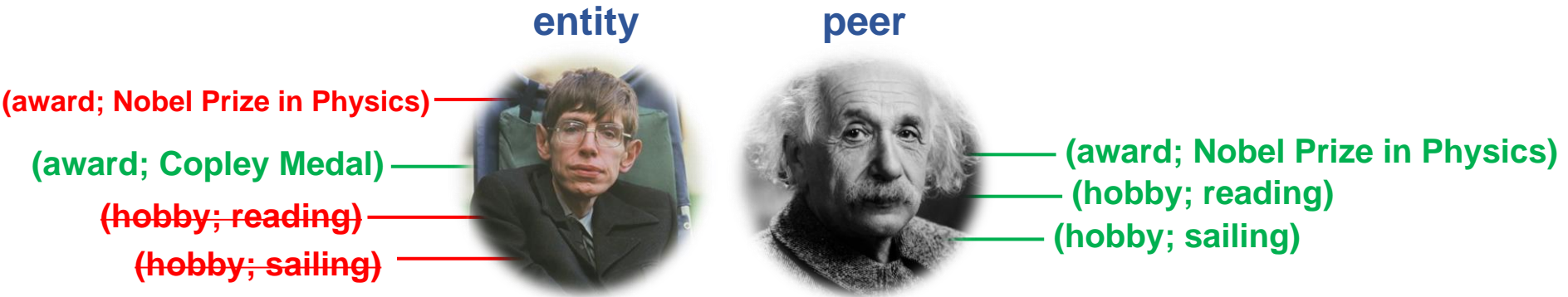
Challenge: *correctness* of inferred negations.

Retain candidate *only in presence of other values*

(Hawking, award, {Copley Medal, ...})  $\models \neg$  (award, Nobel Prize in Physics)

(Hawking, hobby,  $\emptyset$ )  $\not\models \neg$  (sailing, reading)

Every statement that applies to at least one peer is a *candidate negation*.



Challenge: *correctness* of inferred negations.

Retain candidate *only in presence of other values*

(Hawking, award, {Copley Medal, ...})  $\models \neg$  (award, Nobel Prize in Physics)

(Hawking, hobby,  $\emptyset$ )  $\not\models \neg$  (sailing, reading)

Significantly boosts correctness of deductions: **57 to 84%**.

# Supervised Learning-to-rank Model

30



Candidates = [  $\neg$  (handedness; left);  $\neg$  (citizen; U.S.);  $\neg$  (award; Nobel Prize in Physics)]



Candidates = [  $\neg$  (handedness; left);  $\neg$  (citizen; U.S.);  $\neg$  (award; Nobel Prize in Physics)]

- A. Scoring features include:  
peer frequency, object and predicate importance, and text signals.
- B. Pointwise L2R: Obtain annotator judgments for statement interestingness [0..1]  
Is it interesting that Stephen Hawking never received a Nobel in Physics?  
.. is not left-handed?
- C. Train supervised model to predict annotator scores  
Linear Regression
- D. Rank assertions by predicted score

Candidates = [  $\neg$  (handedness; left);  $\neg$  (citizen; U.S.);  $\neg$  (award; Nobel Prize in Physics)]

- A. Scoring features include:  
peer frequency, object and predicate importance, and text signals.
- B. Pointwise L2R: Obtain annotator judgments for statement interestingness [0..1]  
Is it interesting that Stephen Hawking never received a Nobel in Physics?  
.. is not left-handed?
- C. Train supervised model to predict annotator scores  
Linear Regression
- D. Rank assertions by predicted score



1.  $\neg$  (award; Nobel Prize in Physics)
2.  $\neg$  (citizen; U.S.)
3.  $\neg$  (handedness; left)



## PART1: Statistical Inferences

## PART2: Text Extraction

- ★ Pattern-based query log extraction.  
Mining common factual mistakes from Wikipedia updates.

## PART3: Pretrained Language Models

## PART2: Text Extraction

# Mine Negations from User Query Logs

## PART2: Text Extraction

# Mine Negations from User Query Logs

- **Wisdom of the crowd:**  
Search engine **autocomplete** provides access to **salient user assertions**

## PART2: Text Extraction

# Mine Negations from User Query Logs

- **Wisdom of the crowd:**  
Search engine **autocomplete** provides access to **salient user assertions**
- **Probing with negated prefixes**
  - Why didn't <e>
  - Why hasn't <e>
  - Why wasn't <e>
  - ...

# PART2: Text Extraction

## Mine Negations from User Query Logs

- **Wisdom of the crowd:**  
Search engine **autocomplete** provides access to **salient user assertions**
- **Probing with negated prefixes**
  - Why didn't <e>
  - Why hasn't <e>
  - Why wasn't <e>
  - ...



Q why didn't stephen hawking

Q why didn't stephen hawking **get a nobel prize**

Q why didn't stephen hawking **die**

Q why didn't stephen hawking **get a knighthood**

# PART2: Text Extraction

## Mine Negations from User Query Logs

- **Wisdom of the crowd:**  
Search engine **autocomplete** provides access to **salient user assertions**
- **Probing with negated prefixes**
  - Why didn't <e>
  - Why hasn't <e>
  - Why wasn't <e>
  - ...



Q why didn't stephen hawking

Q why didn't stephen hawking **get a nobel prize**

Q why didn't stephen hawking **die**

Q why didn't stephen hawking **get a knighthood**

Q why isn't Switzerland

Q why isn't switzerland **in the eu**

Q why isn't switzerland **part of germany**

Q why isn't switzerland **in nato**



# PART2: Text Extraction

## Mine Negations from User Query Logs

- **Wisdom of the crowd:**  
Search engine **autocomplete** provides access to **salient user assertions**
- **Probing with negated prefixes**
  - Why didn't <e>
  - Why hasn't <e>
  - Why wasn't <e>
  - ...
- **Advantage: High precision**
- **Limitation: Very low recall**



Q why didn't stephen hawking

Q why didn't stephen hawking **get a nobel prize**

Q why didn't stephen hawking **die**

Q why didn't stephen hawking **get a knighthood**

Q why isn't Switzerland

Q why isn't switzerland **in the eu**

Q why isn't switzerland **part of germany**

Q why isn't switzerland **in nato**



## **PART1: Statistical Inferences**

## **PART2: Text Extraction**

Pattern-based query log extraction.

★ Mining common factual mistakes from Wikipedia updates.

## **PART3: Pretrained Language Models**



# **PART2: Text Extraction**

## **Mine Text Revisions**

# PART2: Text Extraction

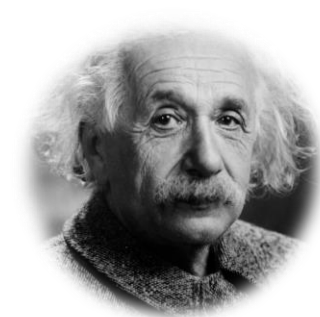
## Mine Text Revisions

- Anti-knowledge base (AKB)  
Create a knowledge base of *common factual mistakes*  
Complement the positive-only KB

# PART2: Text Extraction

## Mine Text Revisions

- Anti-knowledge base (AKB)  
Create a knowledge base of *common factual mistakes*  
Complement the positive-only KB
- Main idea:  
Exploit entity/number swaps in *Wikipedia update logs*  
Web hits for correctness score



Revision 505

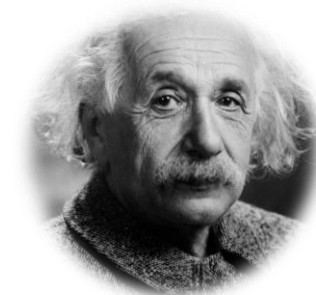
*Einstein was born in **Vienna**.*      Revision 506

*Einstein was born in **Ulm**.*

# PART2: Text Extraction

## Mine Text Revisions

- Anti-knowledge base (AKB)  
Create a knowledge base of *common factual mistakes*  
Complement the positive-only KB
- Main idea:  
Exploit entity/number swaps in *Wikipedia update logs*  
Web hits for correctness score
- Advantage: *High correctness*
- Limitation:  
*Updates occur for a variety of reasons*  
*60% are not factual corrections*  
*controversial, synonyms, spelling mistake, etc.*



Revision 505

*Einstein was born in **Vienna**.*

Revision 506

*Einstein was born in **Ulm**.*

## PART1: Statistical Inferences

## PART2: Text Extraction

## PART3: Pretrained Language Models

- ★ Generating meaningful commonsense negative knowledge:  
Generate corruptions & estimate contradictions.

## **PART3: Pretrained Language Models**

# **Generating Meaningful Negative Commonsense Knowledge**

## **PART3: Pretrained Language Models**

# **Generating Meaningful Negative Commonsense Knowledge**

- **Two-step framework:**

## PART3: Pretrained Language Models

### Generating Meaningful Negative Commonsense Knowledge

- Two-step framework:

- 1) Generate corruptions

plausible candidate negatives by corrupting positives

source: ConceptNet



## PART3: Pretrained Language Models

### Generating Meaningful Negative Commonsense Knowledge

- Two-step framework:

- 1) Generate corruptions

plausible candidate negatives by corrupting positives  
source: ConceptNet

- 2) Estimate contradiction

with fine-tuned BERT for commonsense classification

# Generating Meaningful Negative Commonsense Knowledge

- Two-step framework:

- 1) Generate corruptions

plausible candidate negatives by corrupting positives  
source: ConceptNet

- 2) Estimate contradiction

with fine-tuned BERT for commonsense classification

(horse, IsA, expensive pet)

(cat, IsA, expensive pet)

(goldfish, IsA, expensive pet)

(horse, IsA, expensive car)

**Wikinegata** (*online platform*)



Browse interesting negations about Wikidata entities

**Neguess** (*online quiz-game*) **Neguess?**

Entity guessing game with negative clues

**Anti-KB** (*dataset*)



Ranked common factual mistakes from Wikipedia

**ANION** (*dataset*)



Commonsense KB focusing on negated events

**Google Hotel Search** (*online platform*)



Hotel booking with negative features asserted

★ Wikinegata (*online platform*)



Browse interesting negations about Wikidata entities

Neguess (*online quiz-game*) *Neguess?*

Entity guessing game with negative clues

Anti-KB (*dataset*)



Ranked common factual mistakes from Wikipedia

ANION (*dataset*)



Commonsense KB focusing on negated events

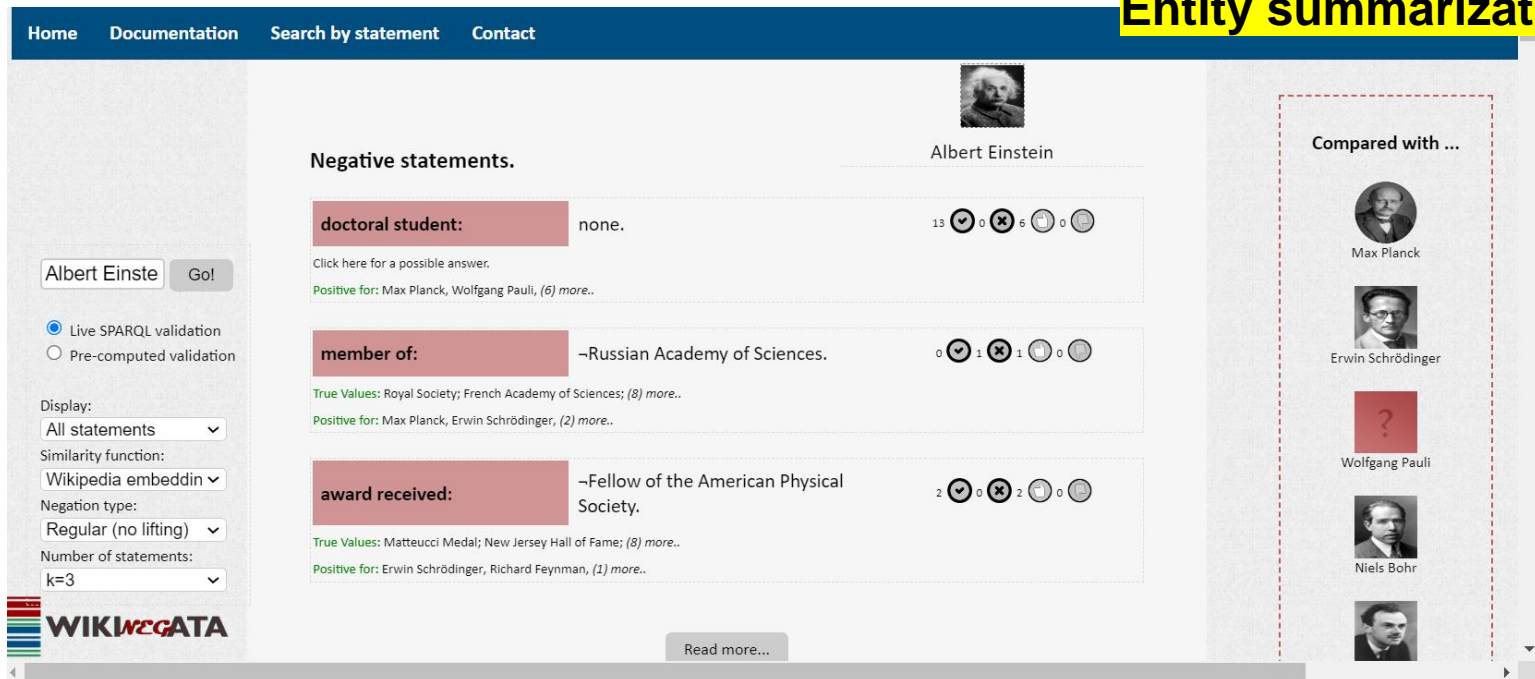
Google Hotel Search (*online platform*)



Hotel booking with negative features asserted

- Will be presented at VLDB this year:  
**Come to the demo session [Blocks 1 & 3]!!**
- Built upon the peer-based negation inference.
- Interesting negations about 0.5M Wikidata entities.

## Entity summarization



The screenshot displays the Wikinegata web application interface. At the top, there is a navigation bar with links: Home, Documentation, Search by statement, and Contact. The main content area is titled "Negative statements." and features a search bar with "Albert Einste" and a "Go!" button. Below the search bar, there are three sections of negative statements for Albert Einstein:

- doctoral student:** none. (13 thumbs down, 0 thumbs up, 5 neutral, 0 dislike)
- member of:** ~Russian Academy of Sciences. (0 thumbs down, 1 thumbs up, 1 neutral, 0 dislike)
- award received:** ~Fellow of the American Physical Society. (2 thumbs down, 0 thumbs up, 2 neutral, 0 dislike)

Each section includes a "Click here for a possible answer." link and a "Positive for:" list of entities. For example, for "doctoral student," the positive list includes Max Planck and Wolfgang Pauli. On the right side, there is a "Compared with ..." section showing a list of entities: Max Planck, Erwin Schrödinger, Wolfgang Pauli (with a red question mark icon), and Niels Bohr. The bottom of the page features the Wikinegata logo and a "Read more..." button.

- Will be presented at VLDB this year:  
**Come to the demo session [Blocks 1 & 3]!!**
- Built upon the peer-based negation inference.
- Interesting negations about 0.5M Wikidata entities.

Home
Documentation
Search by statement
Contact

Question Answering

(award received; Nobel Prize in Physics)

The statement is negative for...


Property: P166: award Entity: Nobel Prize in Physics

Similarity function:  
Wikipedia embedding

Entity type:  
People


Go!

Conditional: ☐ Yes ☒ No




Stephen Hawking - British theoretical physicist, cosmologist and author (1942-2018)

Sample Peer(s): Kip S. Thorne;




Alexander Graham Bell - scientist and inventor known for his work on the telephone

Sample Peer(s): Guglielmo Marconi;



Nikola Tesla - Serbian-American inventor

Sample Peer(s): Guglielmo Marconi;



**Wikinegata** (*online platform*)



Browse interesting negations about Wikidata entities

★ **Neguess** (*online quiz-game*) **Neguess?**

Entity guessing game with negative clues

**Anti-KB** (*dataset*)



Ranked common factual mistakes from Wikipedia

**ANION** (*dataset*)



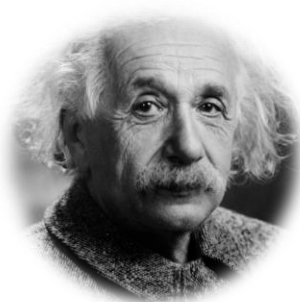
Commonsense KB focusing on negated events

**Google Hotel Search** (*online platform*)



Hotel booking with negative features asserted

- Entity-guessing game with interesting negations as clues.



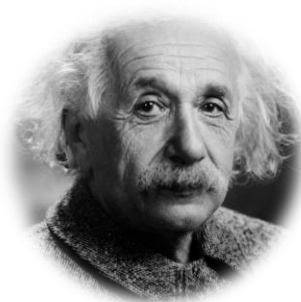
**Clue1:** was *not* educated at Trinity College.

**Clue2:** did *not* win Nobel Prize in Physics.

**Clue3:** is *not* German.



- Entity-guessing game with interesting negations as clues.



**Clue1:** was *not* educated at Trinity College.

**Clue2:** did *not* win Nobel Prize in Physics.

**Clue3:** is *not* German.

**Wikinegata** (*online platform*)



Browse interesting negations about Wikidata entities

**Neguess** (*online quiz-game*) **Neguess?**

Entity guessing game with negative clues

★ **Anti-KB** (*dataset*)



Ranked common factual mistakes from Wikipedia

**ANION** (*dataset*)



Commonsense KB focusing on negated events

**Google Hotel Search** (*online platform*)



Hotel booking with negative features asserted

- Dataset of common factual mistakes: mined from [Wikipedia change log](#).
- 116k likely mistakes where people confuse **entities or numbers**



Penicillin was discovered in 1928 by Scottish scientist **Alexander Baldwin**.



**Alexander Flemming**.



*Confidence score = 0.619*

**Wikinegata** (*online platform*)



Browse interesting negations about Wikidata entities

**Neguess** (*online quiz-game*) **Neguess?**

Entity guessing game with negative clues

**Anti-KB** (*dataset*)



Ranked common factual mistakes from Wikipedia

★ **ANION** (*dataset*)



Commonsense KB focusing on negated events

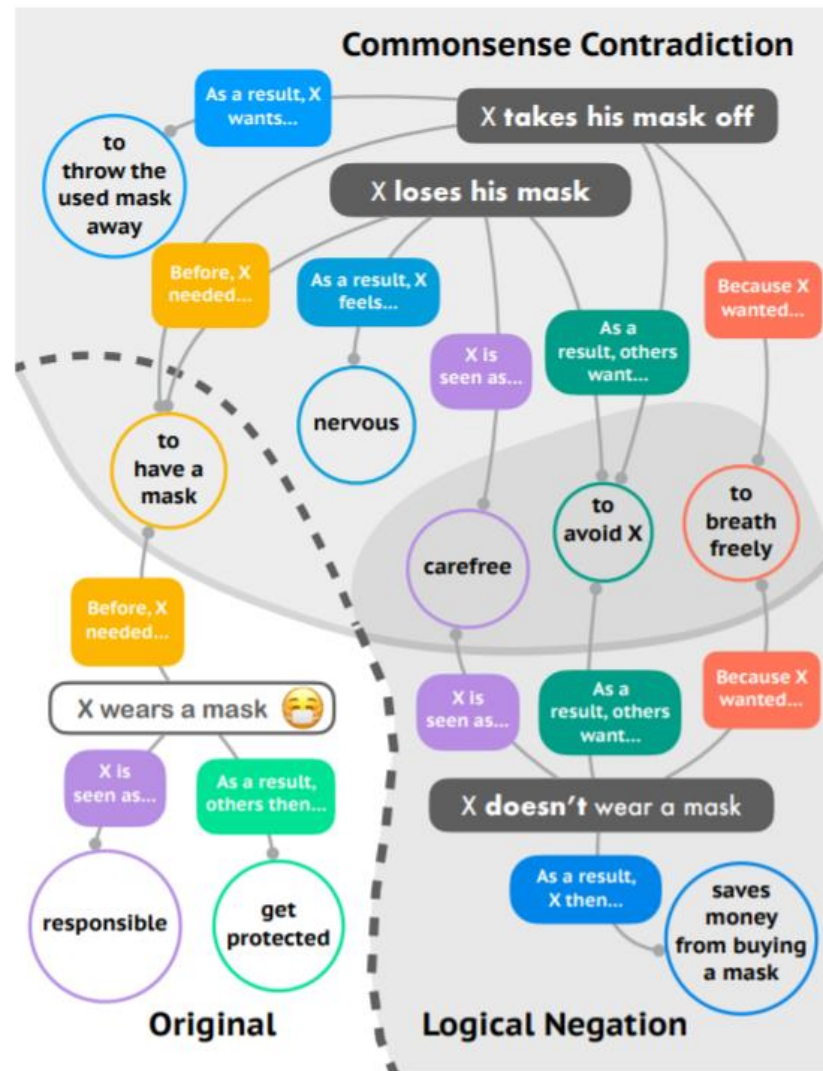
**Google Hotel Search** (*online platform*)



Hotel booking with negative features asserted

- A new commonsense knowledge graph with 624K if-then rules.

<https://github.com/liweijiang/anion>



**Wikinegata** (*online platform*)



Browse interesting negations about Wikidata entities

**Neguess** (*online quiz-game*) **Neguess?**

Entity guessing game with negative clues

**Anti-KB** (*dataset*)



Ranked common factual mistakes from Wikipedia

**ANION** (*dataset*)



Commonsense KB focusing on negated events

★ **Google Hotel Search** (*online platform*)



Hotel booking with negative features asserted



## Data crawled from:

- Hotel websites
- Third-party services
- User reviews



### Internet

- ✓ Wi-Fi **free**
- ✓ Wi-Fi in public areas

### Policies & payments

- ✓ Smoke-free property
- ✓ Credit cards
- ✓ Debit cards
- ✓ Cash

### Services

- ✓ Front desk **24-hour**
- ✓ Baggage storage
- ✓ Full-service laundry
- ✓ Lift
- ✓ Social hour
- ✓ Wake up calls
- ✓ Gift shop
- ✓ Housekeeping **daily**
- ✓ Turndown service

### Accessibility

- ✓ Accessible
- ✓ Accessible lift

### Food and drink

- ✓ Restaurant
- ✓ Bar
- ✓ Table service
- ✓ Room service
- ✓ Breakfast **extra charge**
- ✓ Breakfast buffet

### Activities

- ✓ Bicycle hire **extra charge**
- ✓ Boutique shopping

### Pools

- ☐ No pools
- ☐ No hot tub

### Parking & transport

- ✓ Parking **extra charge**
- ✓ Self parking **extra charge**

### Wellness

- ☐ No spa

### Pets

- ☐ No pets

- **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.**



# 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

# 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: 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**?

# 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

64

- **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.**