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

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

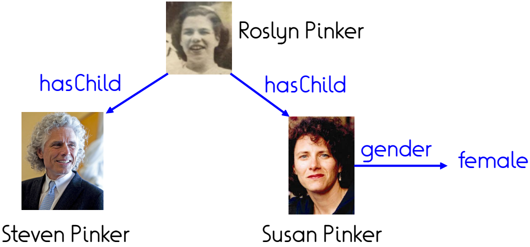
1. Introduction & foundations (Simon)
2. Predictive recall assessment (Fabian)
3. Counts from text and KB (Shrestha)
4. Negation (Hiba)
5. Relative completensss & wrap-up (Simon)

Predictive recall assessment

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

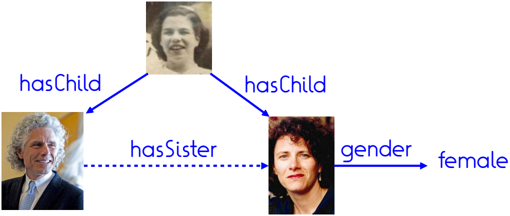
- The Basics: Predicting facts
- Recall of facts
  - Do we have all objects for a subject?
  - Can we use text to determine completeness?
- Recall of entities
  - Do we have all entities of the real world?

Fact Prediction Problem



[Roslyn Pinker from Pinterest, Steven Pinker by Rose Lincoln, Susan Pinker by De Standaard]

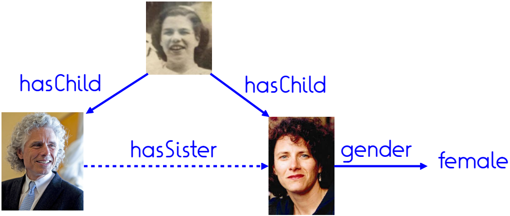
Fact Prediction Problem



We may be able to deduce some facts that are very likely to be true in reality, even though they are not in the KB.
=> This is where the KB must be incomplete

Problem: Fact Prediction Problem
Input: a knowledge base K
Task: Find facts $f \notin K$ that are true in the real world.

Fact Prediction by Rule Mining



Given a KB, rule mining automatically finds logical rules such as:

$$hasChild(x,y) \wedge hasChild(x,z) \wedge gender(z, female) \Rightarrow hasSister(y,z)$$
$$marriedTo(x,y) \wedge hasChild(x,z) \Rightarrow hasChild(y,z)$$
$$wasBornIn(x,y) \wedge hasLanguage(x,z) \Rightarrow speaks(x, z)$$

... usually with a confidence score. These can be used to predict facts.

Fact Prediction by Rule Mining

Bottom-up approaches

Start with rules for concrete instances, generalize them

C. Meilicke, M. Chekol, D. Ruffinelli, H. Stuckenschmidt:
"Anytime Bottom-Up Rule Learning for Knowledge Graph Completion"
(AnyBurl system), IJCAI 2019

Top-down approaches

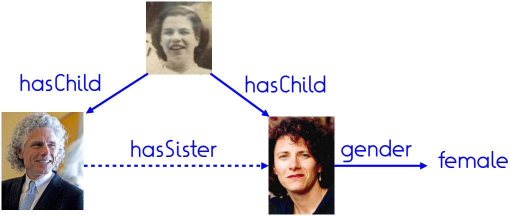
Start with short rules, make them longer

Jonathan Lajus, Luis Galárraga, Fabian M. Suchanek:
"Fast and Exact Rule Mining with AMIE 3", ESWC 2020

Stefano Ortona, Venkata Vamsikrishna Meduri, Paolo Papotti:
"Robust Discovery of Positive and Negative Rules in Knowledge Bases"
(Rudik system) ICDE 2018

>linkPred

Fact Prediction by Link Prediction



We can try to embed the entities in an n-dimensional vector space in such a way that their relative position corresponds to their relations:

>linkPred

[Serena Williams by Edwin Martinez, VEnus Williams by Ken Maynard]

Fact Prediction by Link Prediction

A fact $r(x,y)$ can be embedded in different ways:

TransE

Find $v(\cdot)$ such that $v(x)+v(r)\approx v(y)$.

TransH, TransR, TransD

Map each embedding $v(x)$ to a new vector $v_r(v(x))$ that is specific to the relation r, and impose $v_r(v(s))+v(r)\approx v_r(v(o))$.

RESCAL, DistMul, HolE, ComplEx, ANALOGY

Minimize $cos(v(s)+v(r),v(o))$

Fabian M. Suchanek, Jonathan Lajus, Armand Boschin, Gerhard Weikum: "Knowledge Representation and Rule Mining" Reasoning Web Summer School 2019

Predictive recall assessment

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

- The Basics: Predicting facts
- 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

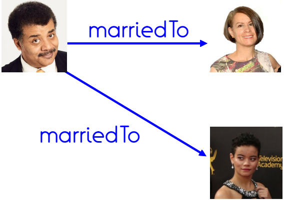
Neil deGrasse Tyson — Alice Young

[Neil deGrasse Tyson from his Facebook page, Alice Young from Wikinetworth]

Are we missing objects?

 $\xrightarrow{\text{marriedTo}}$  √ In the KB, and correct

Are we missing objects?

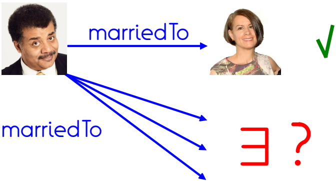


√ In the KB, and correct

? Not in the KB,
but maybe also correct?

[Miranda Tyson from Zimbio]

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

<u>Class pattern oracle:</u>

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

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

<u>Star pattern oracle:</u>

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

As we have seen, rule mining systems can learn (weighted) rules such as

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

Idea:

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

   We have all spouses of Elvis: $complete(Elvis,marriedTo)$

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

   Elvis is a popular entity: $popular(Elvis)$

   Elvis has one spouse in the KB: $cardinalityIsNot0(Elvis,marriedTo)$

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

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

4) Use the rules to predict completeness

   $cardinalityIsNot0(Neil,marriedTo) \wedge popular(Neil)$

   $\Rightarrow complete(Neil, marriedTo)$

->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:

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

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

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

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

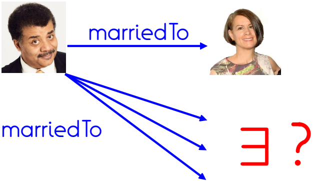
Signals for Incompleteness (F1)

Relation	CWA	PCA	$card_2$	Popularity	No change	Star	Class	AMIE
diedIn	60%	22%	—	4%	15%	50%	**99%**	96%
directed	40%	96%	19%	7%	71%	0%	0%	**100%**
graduatedFrom	89%	4%	2%	2%	10%	89%	**92%**	87%
hasChild	71%	1%	1%	2%	13%	40%	**78%**	**78%**
hasGender	78%	**100%**	—	2%	—	86%	95%	**100%**
hasParent*	1%	54%	**100%**	—	—	0%	0%	**100%**
isCitizenOf*	4%	98%	11%	1%	4%	10%	5%	**100%**
isConnectedTo	87%	34%	19%	—	—	68%	88%	**89%**
isMarriedTo*	55%	7%	0%	3%	12%	37%	**57%**	46%
wasBornIn	28%	**100%**	—	5%	8%	0%	0%	**100%**

Relation	CWA	PCA	$card_2$	Popularity	Star	Class	AMIE
alma_mater	**90%**	14%	5%	1%	87%	87%	87%
brother	93%	1%	—	1%	94%	**96%**	**96%**
child	70%	1%	—	1%	**79%**	72%	73%
country_of_citizenship*	42%	97%	10%	3%	0%	0%	**98%**
director	81%	**100%**	—	3%	94%	89%	**100%**
father*	5%	**100%**	6%	9%	89%	8%	**100%**
mother*	3%	**100%**	3%	10%	67%*	5%	**100%**
place_of_birth	53%	**100%**	7%	5%	55%	0%	**100%**
place_of_death	89%	35%	1%	2%	81%	81%	**96%**
sex_or_gender	81%	**100%**	6%	3%	92%	91%	**100%**
spouse*	**57%**	7%	—	1%	54%	54%	55%

• = biased training sample

Missing Object Problem



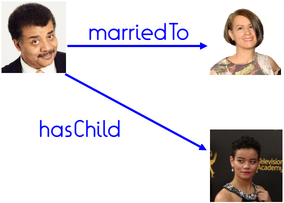
Are there objects in the real world that are missing from the KB?

=> By help of <u>supervised learning</u>, 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 <u>supervised learning</u>, 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

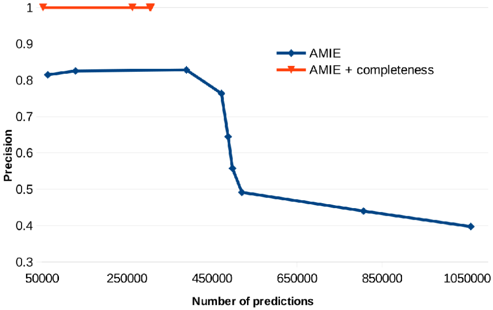
>married

Missing Object Problem: Application

As we have seen, fact prediction is a method that uses rules such as

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

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



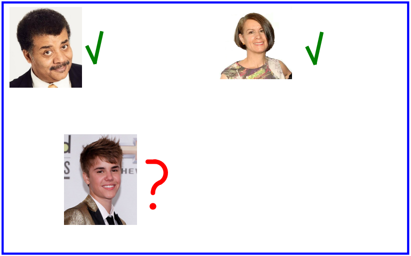
›married

Are all people married?

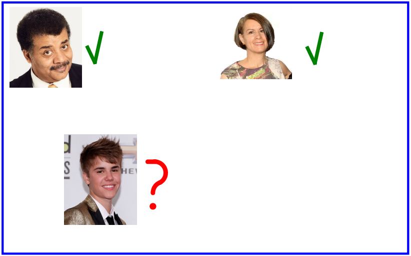
 √

 √

Are all people married?



Are all people married?



Obligatory for people:

- hasBirthPlace
- hasNationality

Not obligatory:

- isMarriedTo
- hasChild

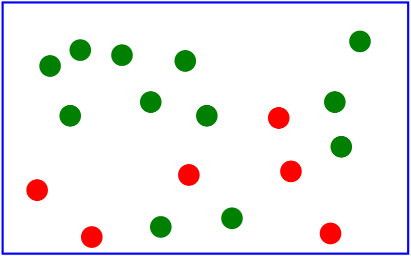
Problem: Obligatory Attribute Problem

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

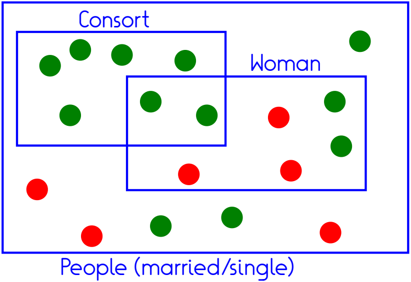
Are all people married?



People (married/single)

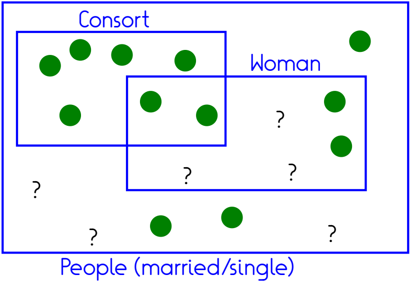
Real World

Are all people married?



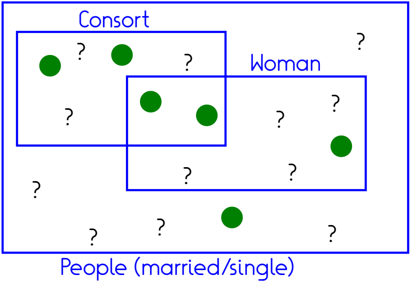
Real World

Are all people married?



Knowledge base
without negative facts

Are all people married?

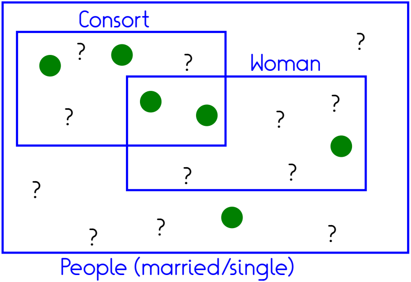


Knowledge base without negative facts and with incompleteness

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

>assumptions

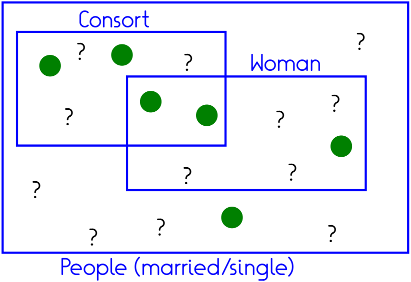
Are all people married?



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

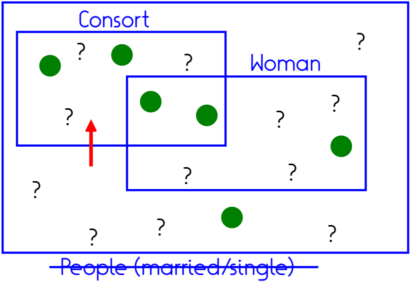
Are all people married?



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.

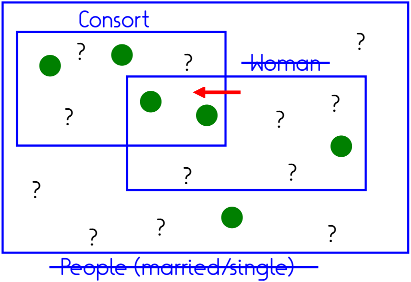
p obligatory in class $c \Rightarrow \forall\, c'$: $E(ratio\ of\ p\ in\ c \backslash c') = E(ratio\ of\ p\ in\ c \cap c')$

Are all people married?



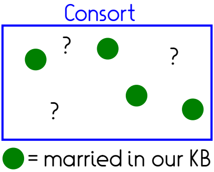
Theorem: If the KB is sampled randomly uniformly from the real world, and if the density of an attribute changes when we go into an intersecting class, then the attribute cannot be obligatory.

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



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?

- The Basics: Predicting facts
- 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?

Marie Curie

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

[Marie Curie from FineArtAmerica, Irène Curie from Twitter, Eve Curie from Valentine]

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.

[Marie Curie from FineArtAmerica, Irène Curie from Twitter, Eve Curie from Valentine]

Text Coverage Problem

Marie brought her child Irène to school.

Sentence incomplete

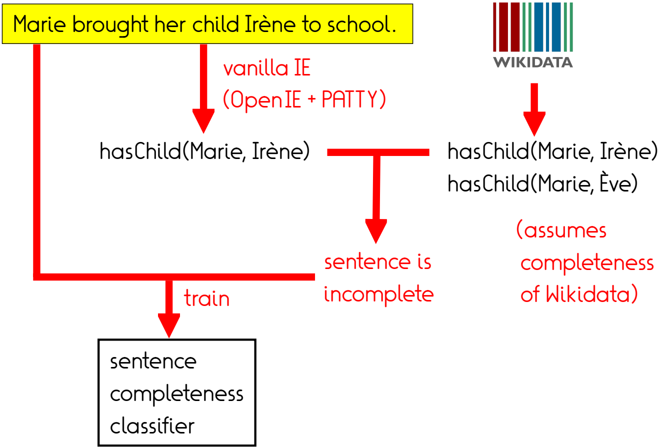
Marie has two daughters, Irène and Ève.

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 $r(s,o)$.

[Marie Curie from FineArtAmerica, Irène Curie from Twitter, Eve Curie from Valentine]

Text Coverage Problem



Text Coverage Problem

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

$\Rightarrow$ 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

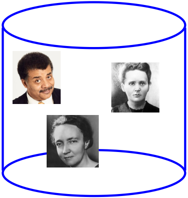
Predictive recall assessment

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

- The Basics: Predicting facts
- 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 builing 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?

>problemsOfDef

Missing Entities Problem

Classes are usually not well-defined:

- is anyone with a doctoral degree a scientist?



- what is the total number of cathedrals, if some are built/destroyed?
- what is the total number of islands? Do we also count islets, rocks...?
- what is the total number of inhabitants of a country? Do we also count deceased people? Do we count only famous people?

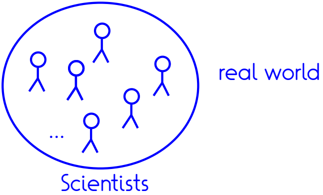
=> we can work only on very crisp and restricted classes

- countries recognized by the UN as of 2021
- mountains taller than 1000m
- ...

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.

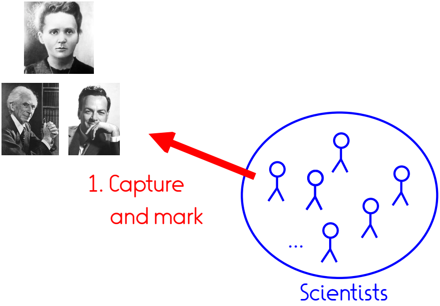
Example:



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.

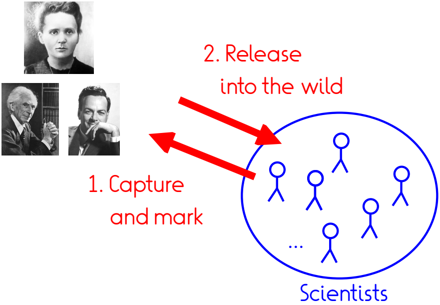
Example:



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.

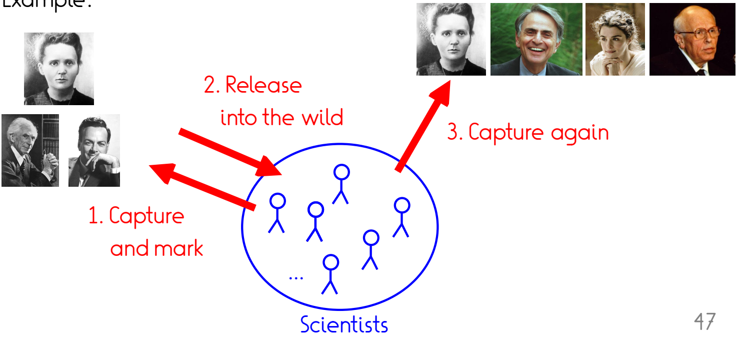
Example:



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.

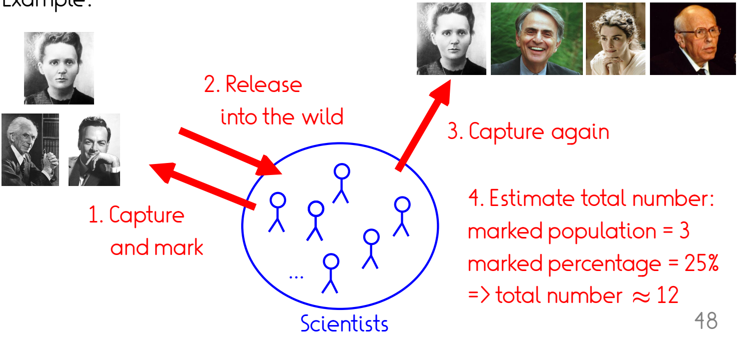
Example:



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 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	sample period 2
hasChild(MarieCurie, Eve)	hasChild(MarieCurie, Irène)
type(BertrandRussell,Humanist)	livedIn(Hypatia, Alexandria)
married(Arline,RichardFeynman)	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

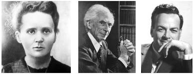
sample period 1

hasChild(MarieCurie, Eve)
type(BertrandRussell,Humanist)
married(Arline,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)

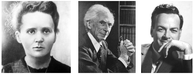
c = current number of entities in the KB

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

Try several estimators, e.g.

- JackKnife: $c + \frac{k-1}{k} f_1$
- Streaker
- Chao92, Good-Turing: $\frac{c}{1 - f_1/n}$ [...]

Sample 1



Sample 2

›estimators

Estimators: Good-Turing

k = number of sample periods (here: 2)

n = number of observations (here: 7)

c = current number of entities in the KB

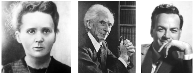
f_i = frequence of entities observed i times (here: $f_1 = 5$)

Good-Turing estimator: The fraction of items that we have not seen, out of the entire population is estimated as $\frac{f_1}{n}$. Therefore, the total number of items is

$$N \approx D \times \left(1 - \frac{f_1}{n}\right)^{-1}$$

If every item I see is new, $f_1 = n$, $N = \infty$

Sample 1



Sample 2

Estimators: Jackknife

k = number of sample periods (here: 2)

n = number of observations (here: 7)

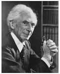
c = current number of entities in the KB

f_i = frequence 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:

Jacknife = $AVERAGE_{j=1..k}\,(k-1)\,f_1^j + D$

Sample 1

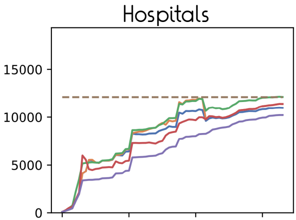



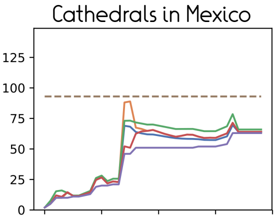

Sample 2




Missing Entities Problem in edited KB

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





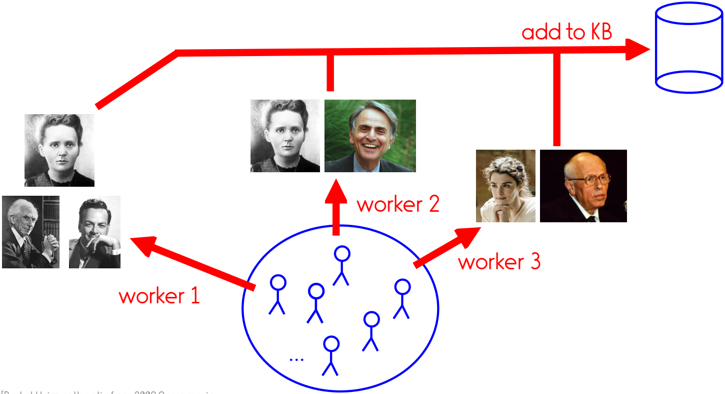
— — — —: different estimators   - -: ground truth x-axis: samples

=> 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

Missing Entities in Crowd-sourced KBs

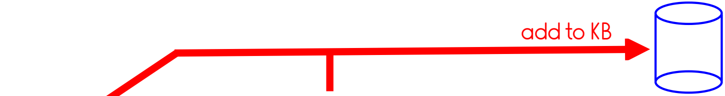
In a crowd-sourced KB, the workers create a "sample" from the world.
The entities that appear more than once are the "re-captured" ones.

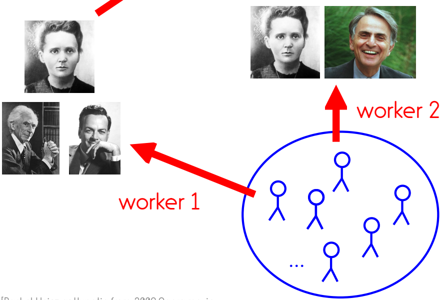


[Rachel Weisz as Hypatia from 2009 Agora movie,
Bertrand Russell from Nationaal Archief, Andrei Sakharov from TheFamousPeople, Richard Feynman from Nobel Foundation, Carl Sagan from NASA] >streaker

Missing Entities in 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

[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

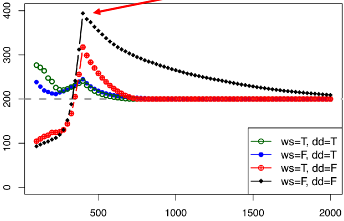
The Streaker Problem

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

streaker arrives,

estimators over-estimate

estimation of total number of entities



answers

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: $\dfrac{c}{1-f_1/n}$

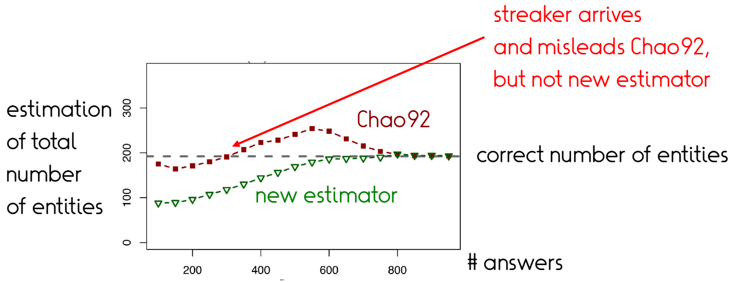
observed number of unique entities in the set

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

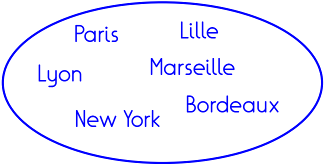


=> In a crowd-sourced KB, the total number of entities can be estimated

Beth Trushkowsky, Tim Kraska, Michael J. Franklin, Purnamrita Sarkar:
"Crowdsourced Enumeration Queries "
International Conference on Data Engineering (ICDE) 2013

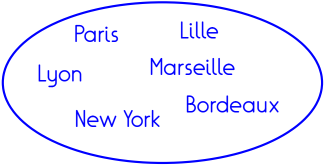
Missing Entities Problem in static KBs

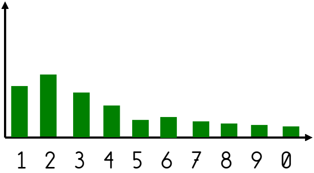
If the KB is static, the mark-and-recapture estimators do not work.



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

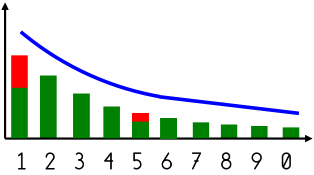
Missing Entities Problem in static KBs

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

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

=> 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

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

This holds only for quantities that grow by multiplicative factors:

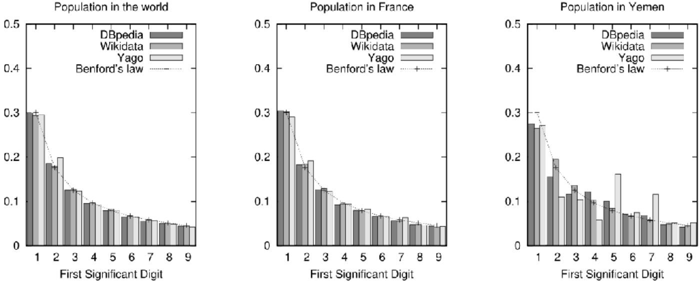
- number of inhabitants of cities
- the size of a lake
- other natural processes

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 α, 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 α by a weighted least square measure

3. Run a MAD (Mean Absolute Deviation) test to see if Benford's Law could be applied

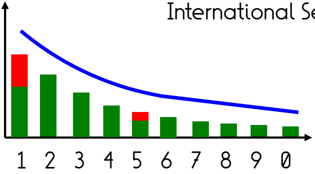
4. If so, compute number of entities that have to be added to conform to Benford's Law

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



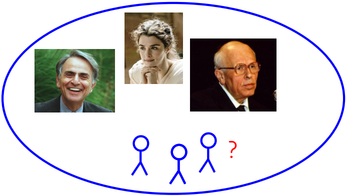
Takeaway: Predictive recall assessment

Using statistical techniques, we can predict more or less:

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)