Using Wikipedia categories for research

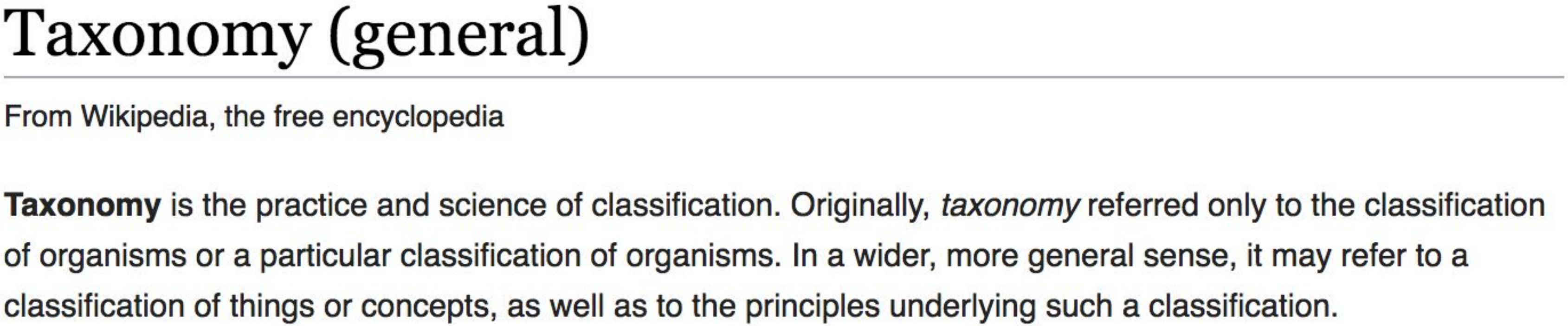
Opportunities, Challenges, and Solutions

Tiziano Piccardi, Michele Catasta, Robert West, Leila Zia

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Stanford
University

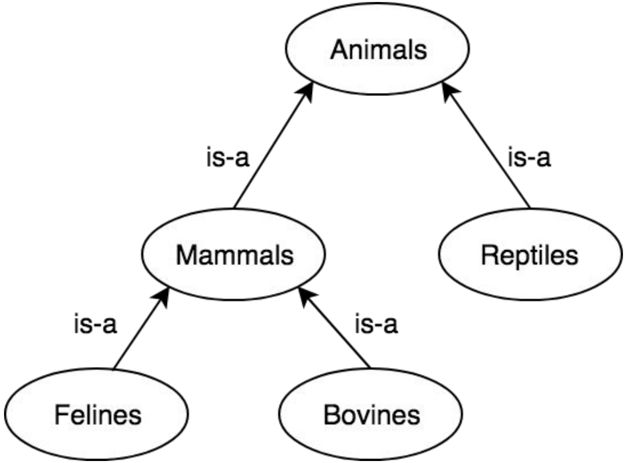
WIKIMEDIA
FOUNDATION



Taxonomy (general)

From Wikipedia, the free encyclopedia

Taxonomy is the practice and science of classification. Originally, *taxonomy* referred only to the classification of organisms or a particular classification of organisms. In a wider, more general sense, it may refer to a classification of things or concepts, as well as to the principles underlying such a classification.



Hierarchical organization of concepts

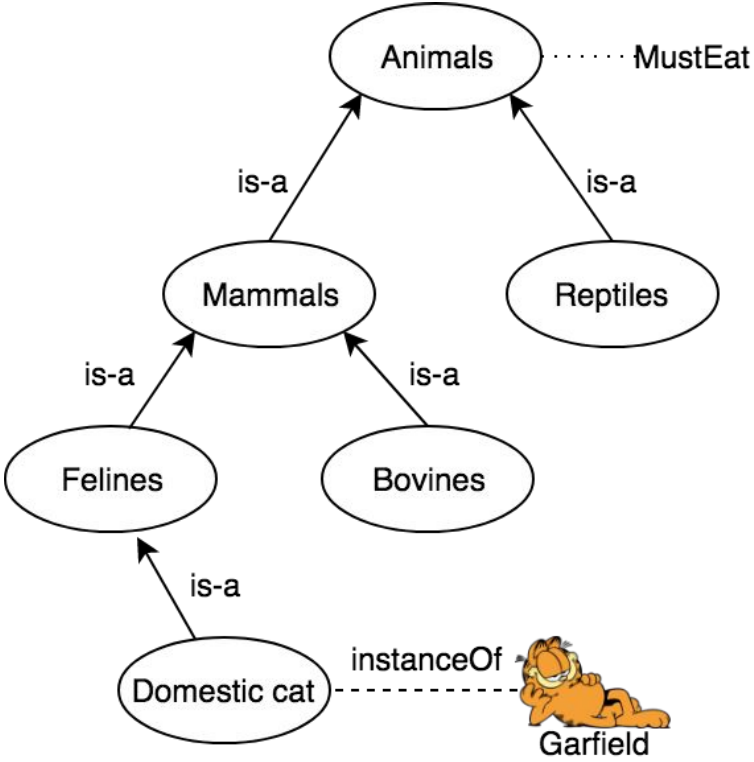
Taxonomies are useful in many different domains:

Image classification

Medical domain

Q&A bots

An automated reasoning tool can infer new knowledge



is-a*(Mammals, Animals)*

is-a*(Felines, Mammals)*

⇓

is-a*(Felines, Animal)*

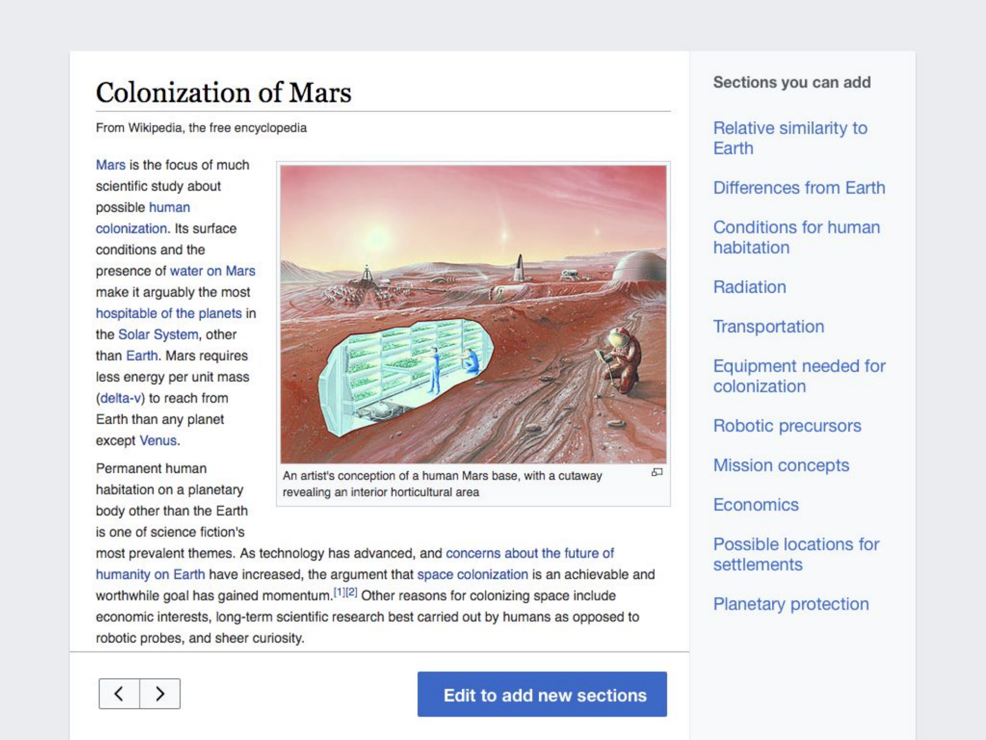
1

Statement: Animals must eat to live

Question: Does Garfield eat? ⇒ Yes

2

… and in sections recommendation!



Colonization of Mars

From Wikipedia, the free encyclopedia

Mars is the focus of much scientific study about possible human colonization. Its surface conditions and the presence of water on Mars make it arguably the most hospitable of the planets in the Solar System, other than Earth. Mars requires less energy per unit mass (delta-v) to reach from Earth than any planet except Venus.

An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area

Permanent human habitation on a planetary body other than the Earth is one of science fiction's most prevalent themes. As technology has advanced, and concerns about the future of humanity on Earth have increased, the argument that space colonization is an achievable and worthwhile goal has gained momentum.[1][2] Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.

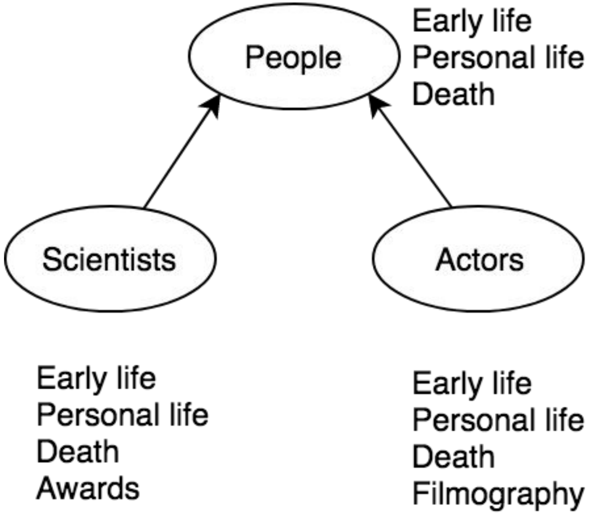
Sections you can add

- Relative similarity to Earth
- Differences from Earth
- Conditions for human habitation
- Radiation
- Transportation
- Equipment needed for colonization
- Robotic precursors
- Mission concepts
- Economics
- Possible locations for settlements
- Planetary protection

Given an article A in language L, recommend the list of sections to be added to the article considering similar articles to article A in language L.

** Leila's talk -- Showcase Dec 2017*

Generating concept-based templates



An article about a person can be represented with:

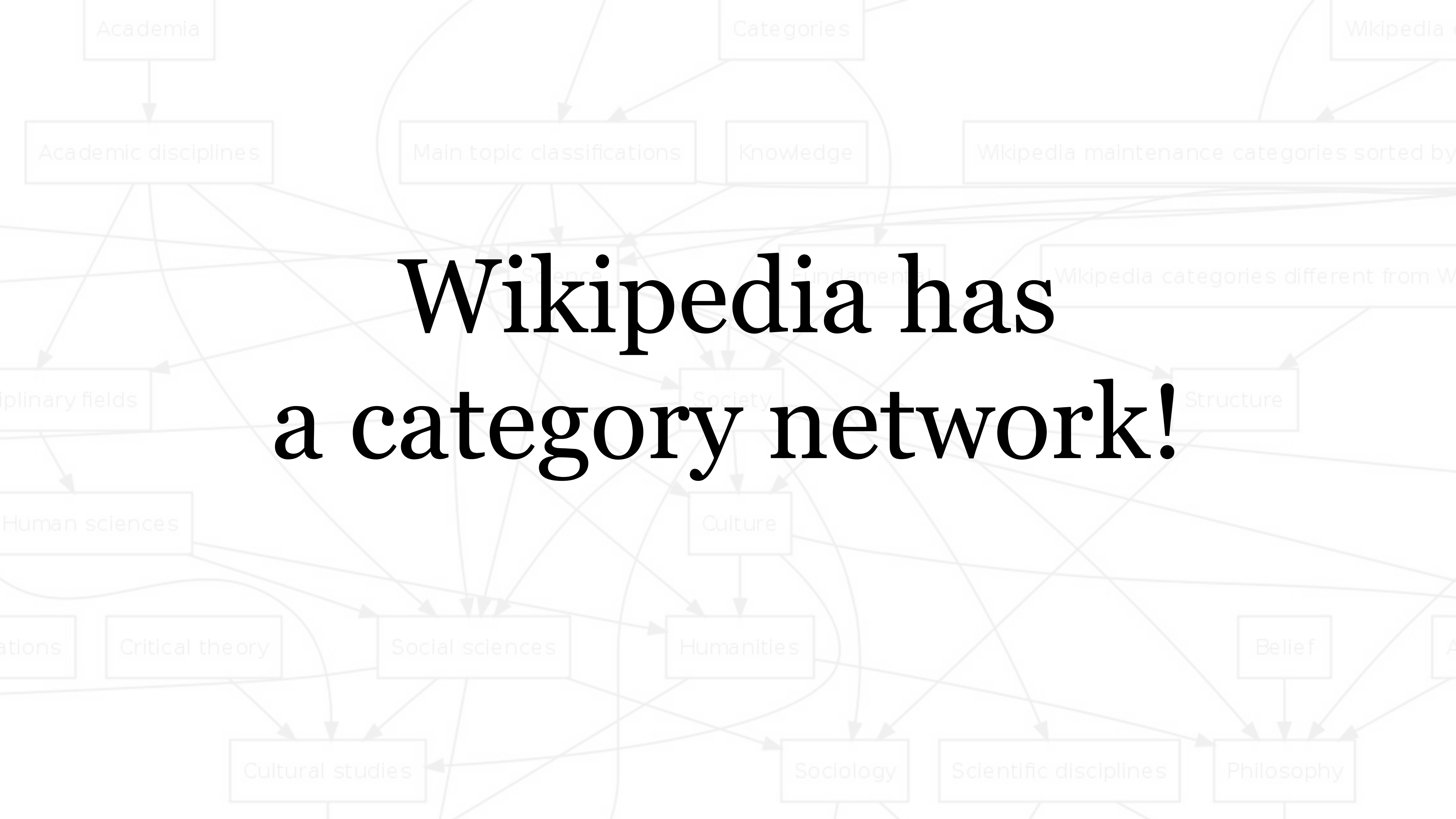
- Early life
- Personal life
- Death

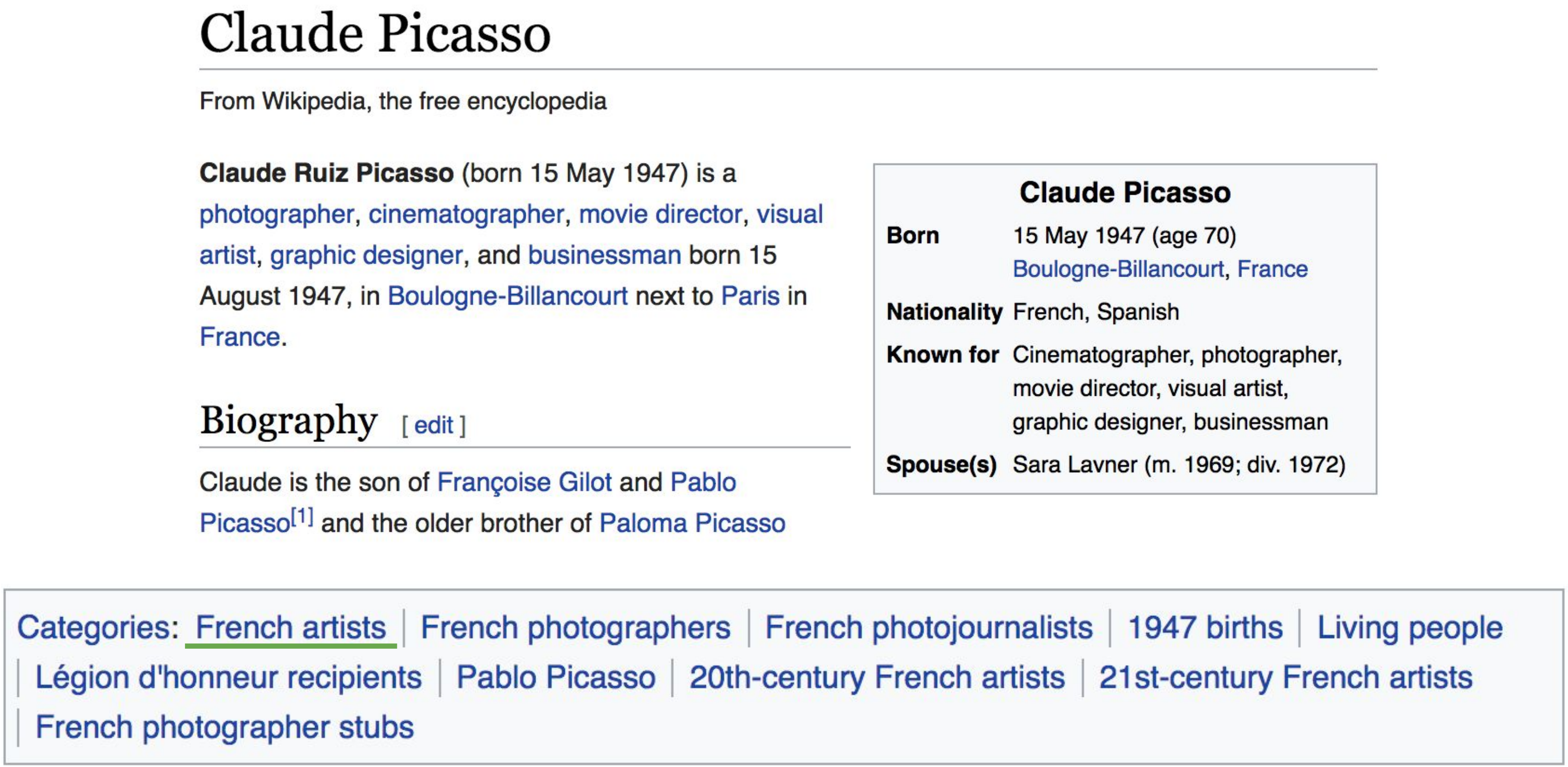
The structure shared between multiple children

Generating the complete taxonomy manually is not feasible!

288 languages
hundreds of thousands concepts

Wikipedia has a category network!





Claude Picasso

From Wikipedia, the free encyclopedia

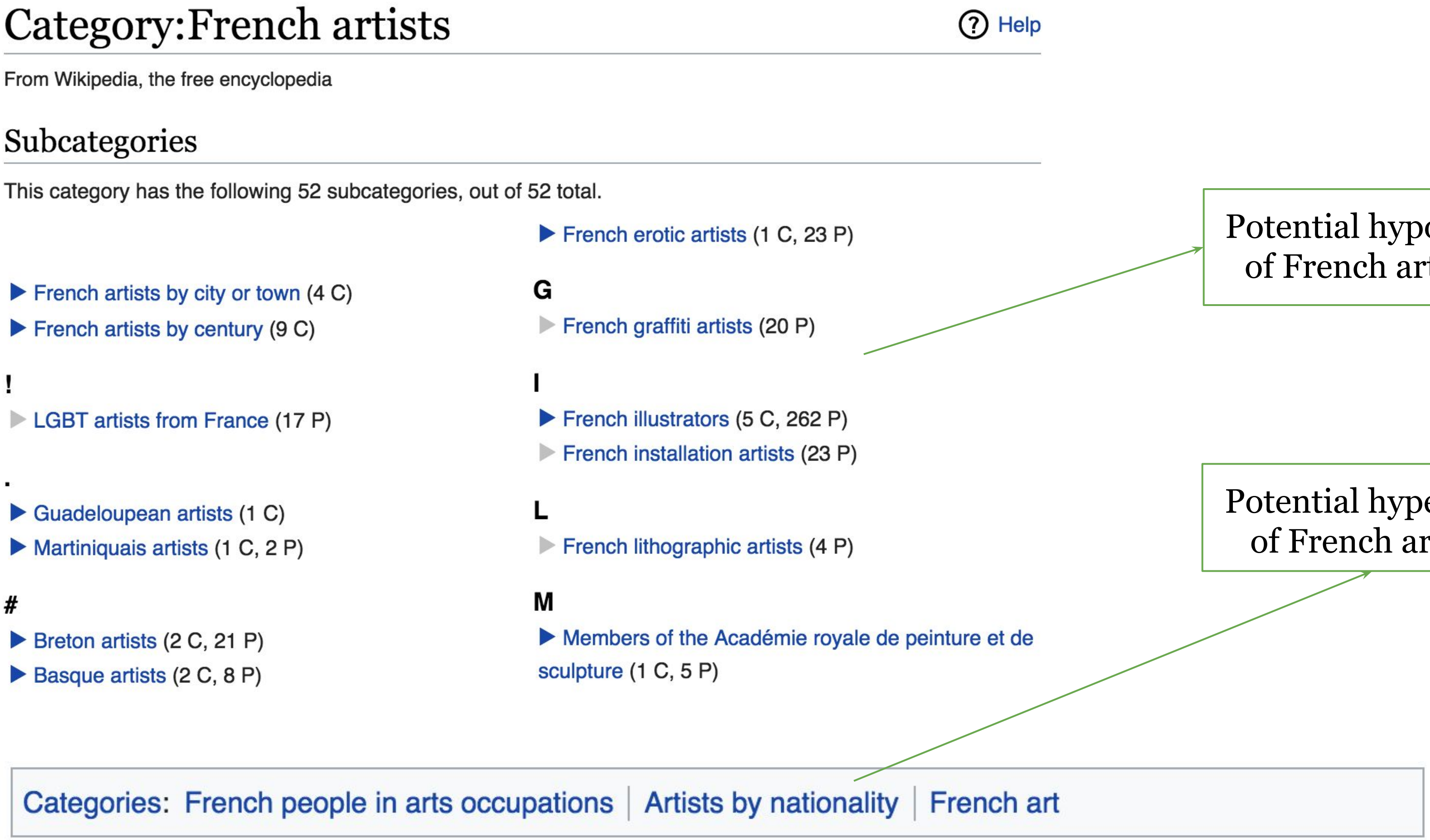
Claude Ruiz Picasso (born 15 May 1947) is a photographer, cinematographer, movie director, visual artist, graphic designer, and businessman born 15 August 1947, in Boulogne-Billancourt next to Paris in France.

Claude Picasso	
Born	15 May 1947 (age 70) Boulogne-Billancourt, France
Nationality	French, Spanish
Known for	Cinematographer, photographer, movie director, visual artist, graphic designer, businessman
Spouse(s)	Sara Lavner (m. 1969; div. 1972)

Biography [edit]

Claude is the son of Françoise Gilot and Pablo Picasso[1] and the older brother of Paloma Picasso

Categories: French artists | French photographers | French photojournalists | 1947 births | Living people | Légion d'honneur recipients | Pablo Picasso | 20th-century French artists | 21st-century French artists | French photographer stubs



Category:French artists

From Wikipedia, the free encyclopedia

Subcategories

This category has the following 52 subcategories, out of 52 total.

- ▶ French artists by city or town (4 C)
- ▶ French artists by century (9 C)

!

- ▷ LGBT artists from France (17 P)

.

- ▶ Guadeloupean artists (1 C)
- ▶ Martiniquais artists (1 C, 2 P)

#

- ▶ Breton artists (2 C, 21 P)
- ▶ Basque artists (2 C, 8 P)

- ▶ French erotic artists (1 C, 23 P)

G

- ▷ French graffiti artists (20 P)

I

- ▶ French illustrators (5 C, 262 P)
- ▷ French installation artists (23 P)

L

- ▷ French lithographic artists (4 P)

M

- ▶ Members of the Académie royale de peinture et de sculpture (1 C, 5 P)

Categories: French people in arts occupations | Artists by nationality | French art

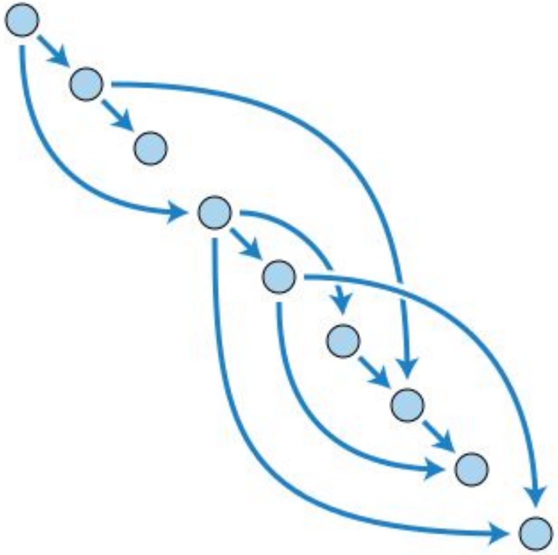
Potential hyponym of French artists

Potential hypernym of French artists

Wait, it’s not so easy…

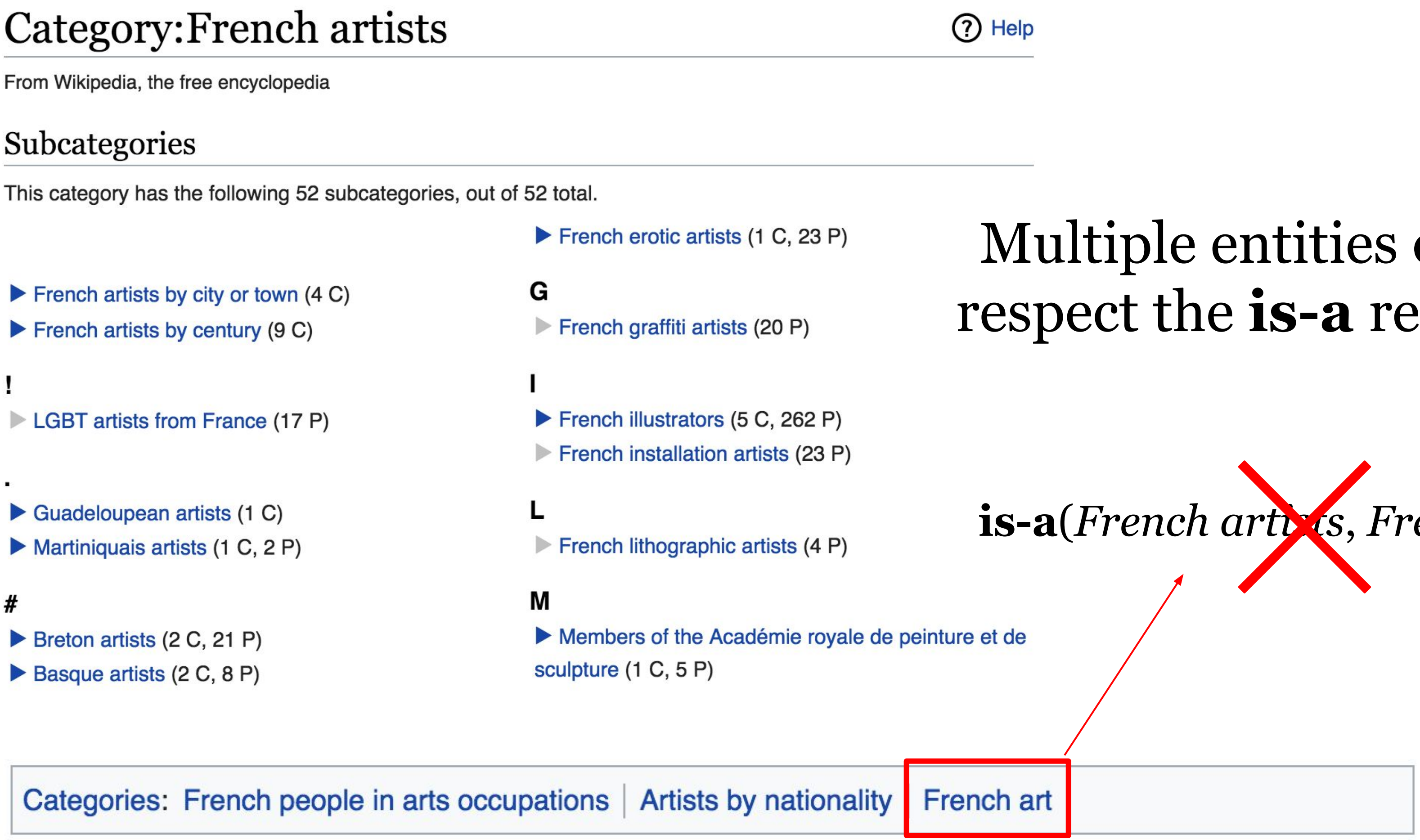
The category network has loops!

Government ➜ Public administration ➜ Public economics ➜ Economic policy ➜ Government



A fast and effective heuristic for the feedback arc set problem

Peter Eades, Xuemin Lin, W.F. Smyth



Category:French artists

From Wikipedia, the free encyclopedia

Subcategories

This category has the following 52 subcategories, out of 52 total.

▶ French artists by city or town (4 C)
▶ French artists by century (9 C)

!

▷ LGBT artists from France (17 P)

.

▶ Guadeloupean artists (1 C)
▶ Martiniquais artists (1 C, 2 P)

#

▶ Breton artists (2 C, 21 P)
▶ Basque artists (2 C, 8 P)

▶ French erotic artists (1 C, 23 P)

G

▷ French graffiti artists (20 P)

I

▶ French illustrators (5 C, 262 P)
▷ French installation artists (23 P)

L

▷ French lithographic artists (4 P)

M

▶ Members of the Académie royale de peinture et de sculpture (1 C, 5 P)

Categories: French people in arts occupations | Artists by nationality | French art

Multiple entities don’t respect the **is-a** relation

is-a(*French artists, French art*)

Related work

MultiWiBi

MultiWiBi: The multilingual Wikipedia bitaxonomy project.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, Roberto Navigli.

Head Taxonomy

Revisiting Taxonomy Induction over Wikipedia

Amit Gupta, Francesco Piccinno, Mikhail Kozhevnikov, Marius Pasca, Daniele Pighin

Generated with language-based heuristics

Goal:

Generate a set for sections to recommend based on the categories of the article

⇩

We need to preserve as much as possible the original topology and remove what could add "noise"

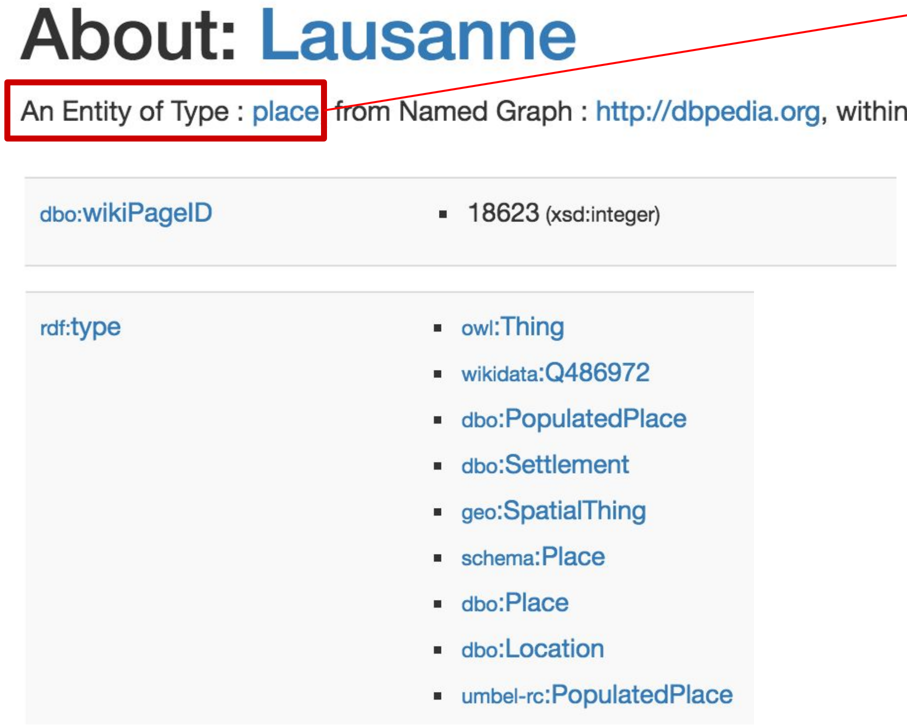
Our method!

Assuming we have a generic high-level type for each article in Wikipedia

Taxonomic categories should contain articles of the same type!

DBpedia - Types extraction

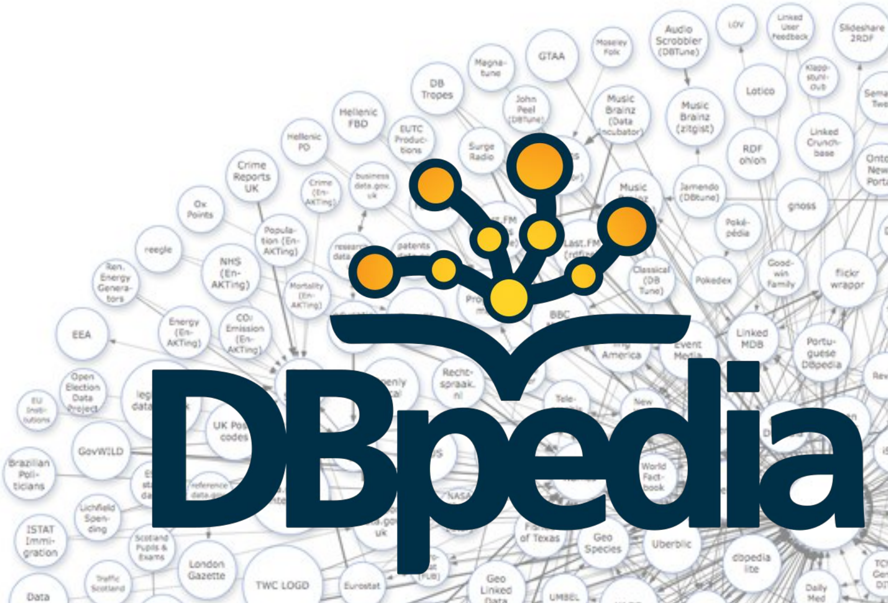
55 top-level types



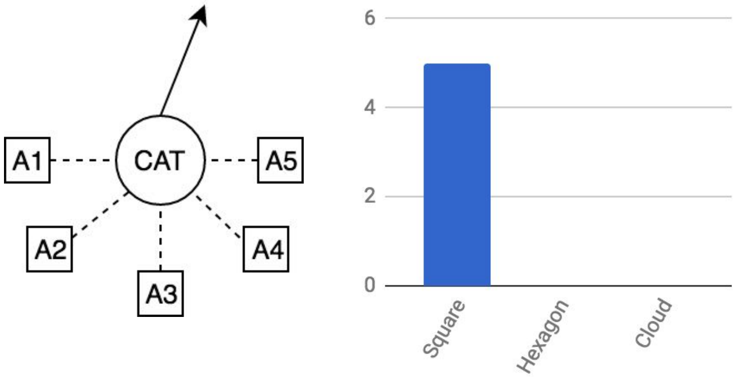
About: Lausanne

An Entity of Type : place from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

dbo:wikiPageID	18623 (xsd:integer)
rdf:type	owl:Thing wikidata:Q486972 dbo:PopulatedPlace dbo:Settlement geo:SpatialThing schema:Place dbo:Place dbo:Location umbel-rc:PopulatedPlace

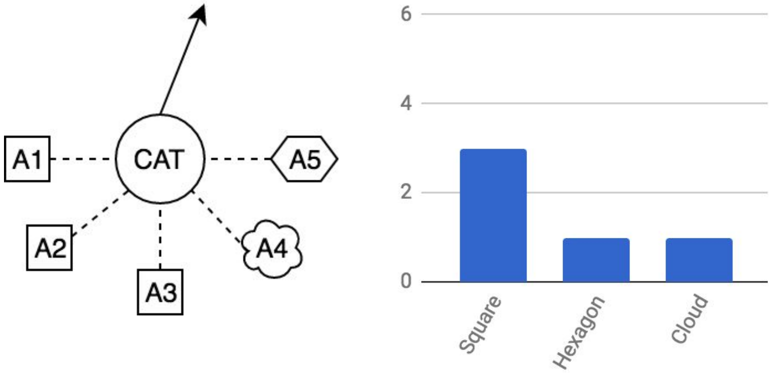


The concept of purity



Pure category
The distribution of types is homogeneous

Scientists → *“Person”*

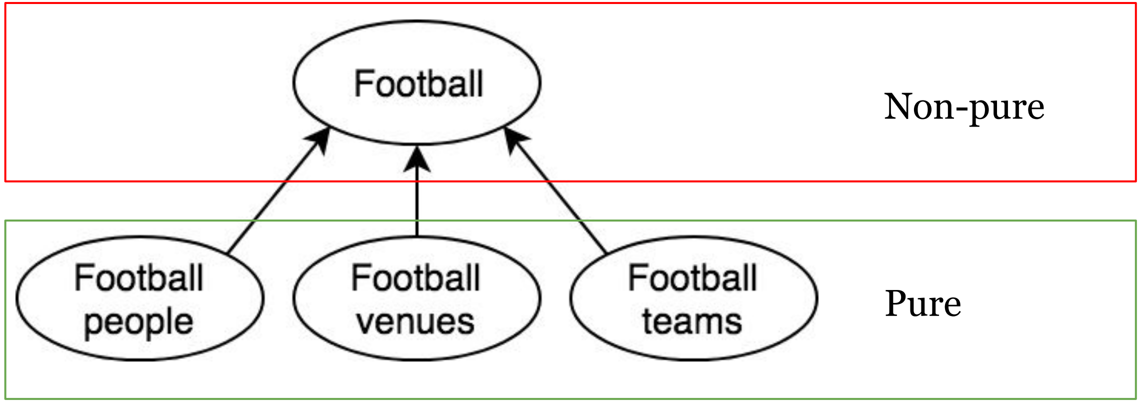


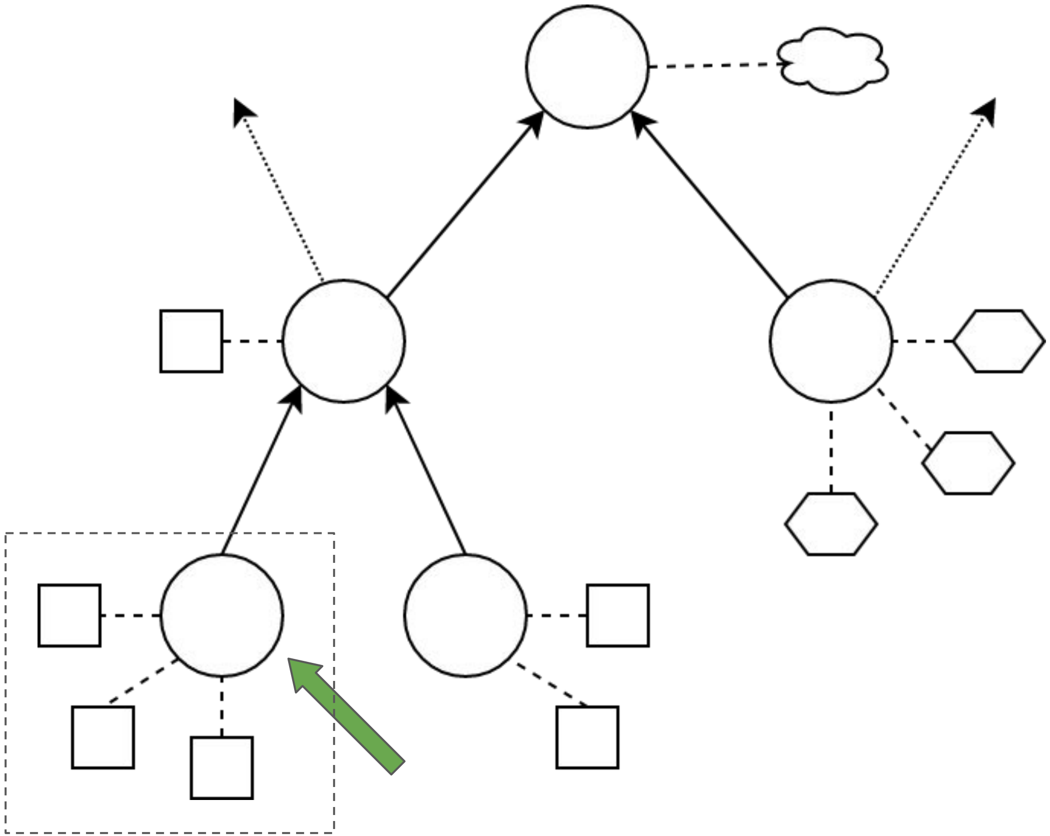
Non-pure category
The distribution of types is heterogeneous

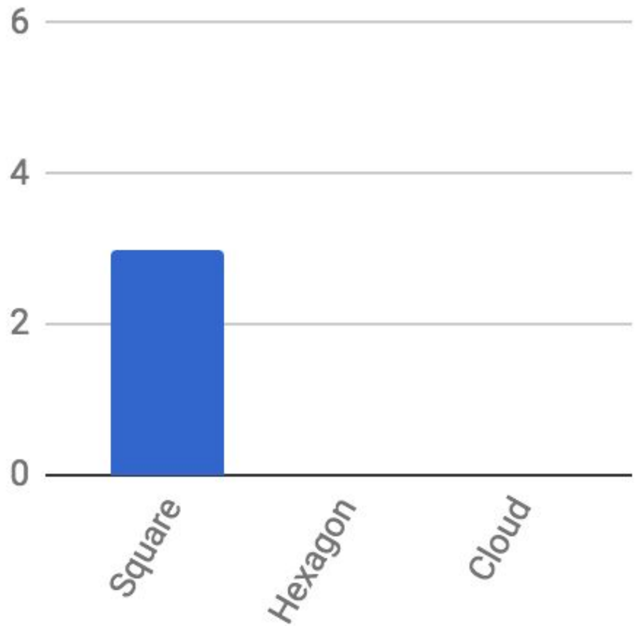
Football → *“Person”, “Event”, “Organization”*

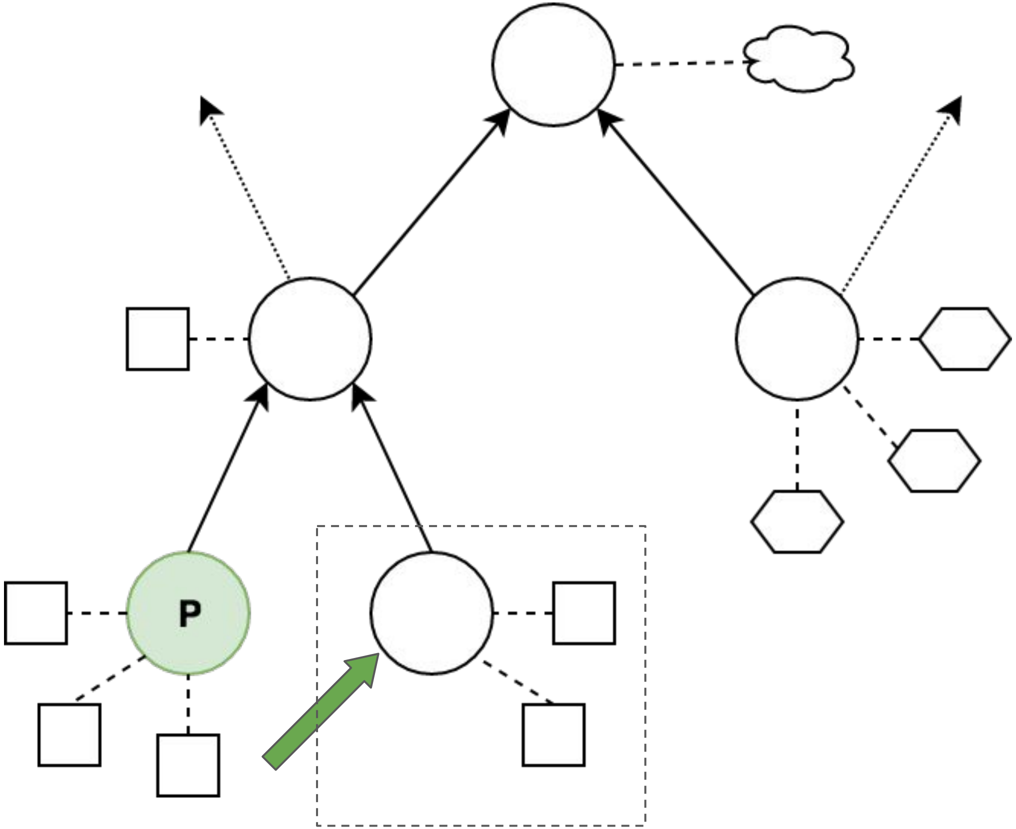
Idea:

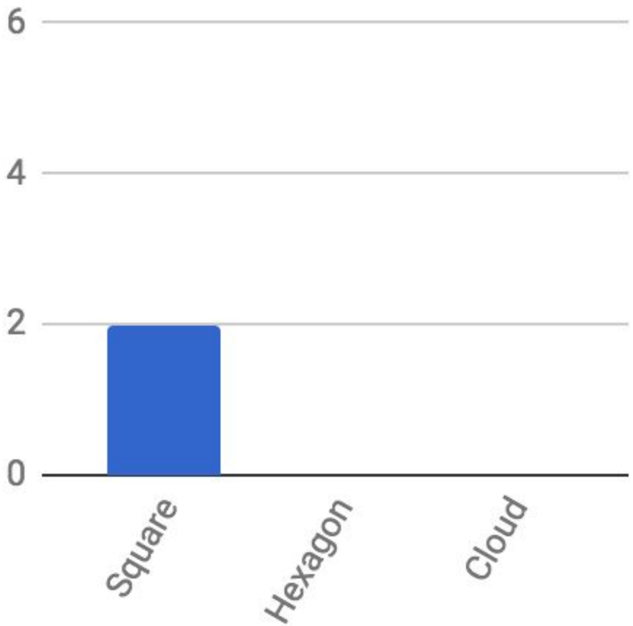
With a bottom-up approach, we can remove the categories where different types converge

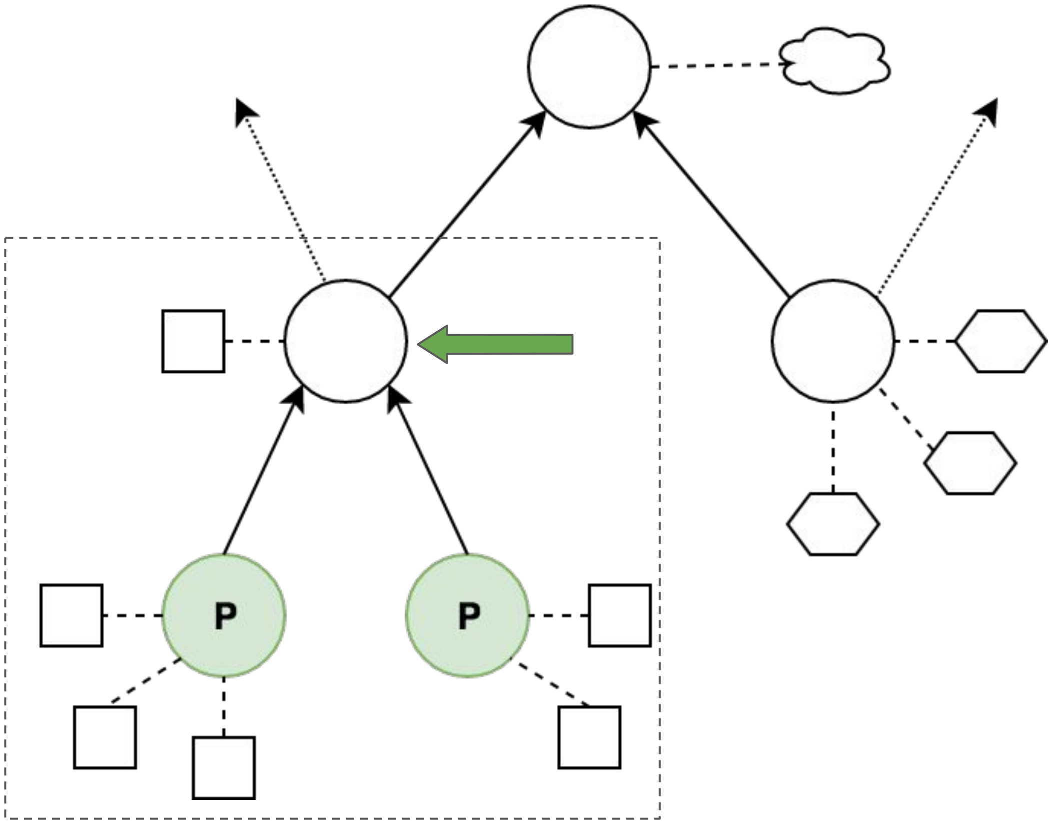


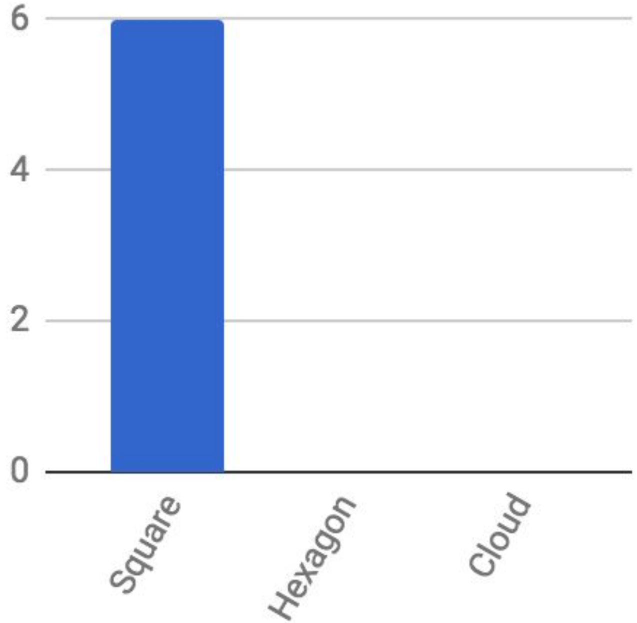


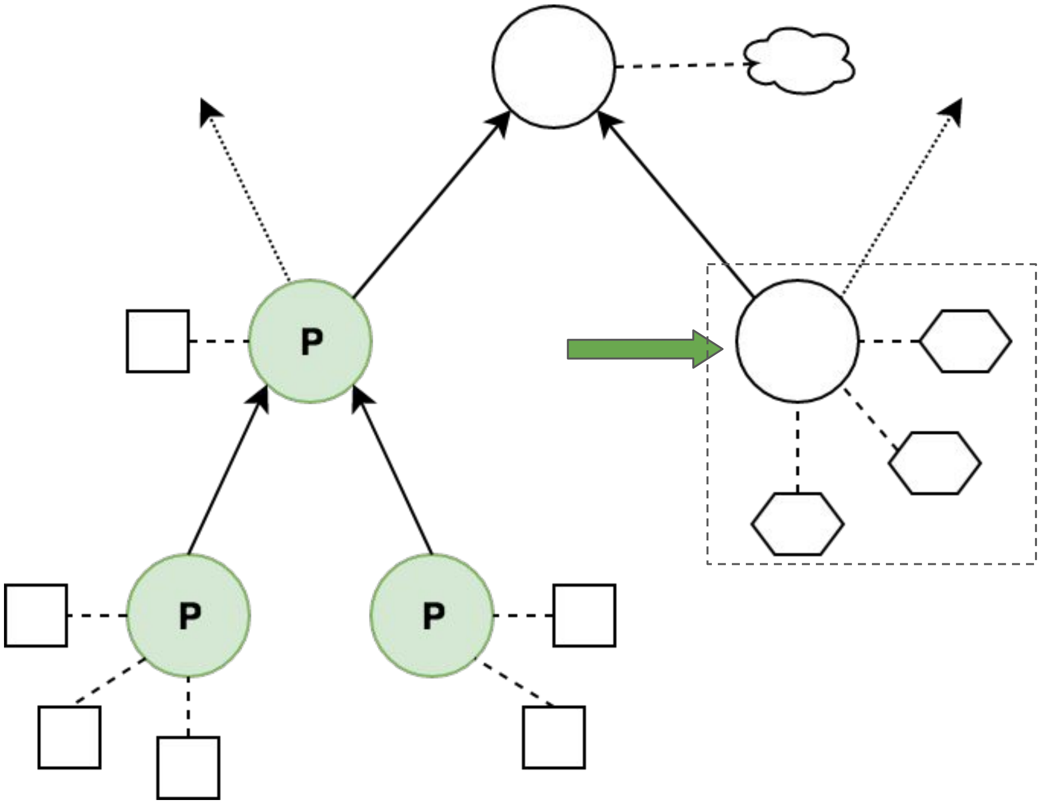

6
4
2
0
Square
Hexagon
Cloud

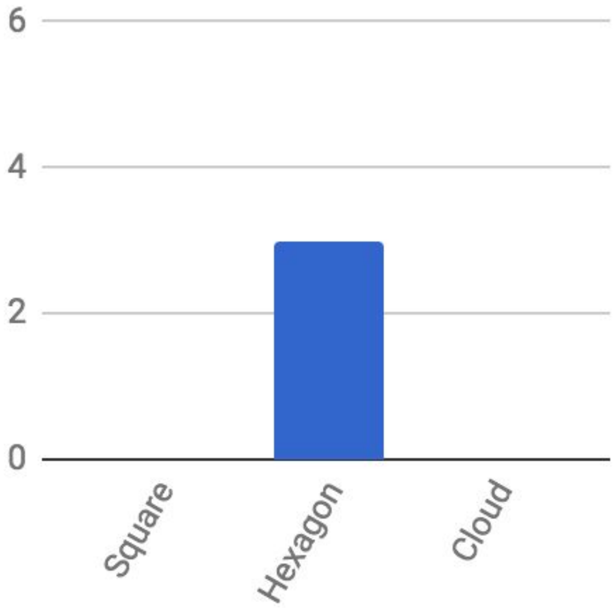

P

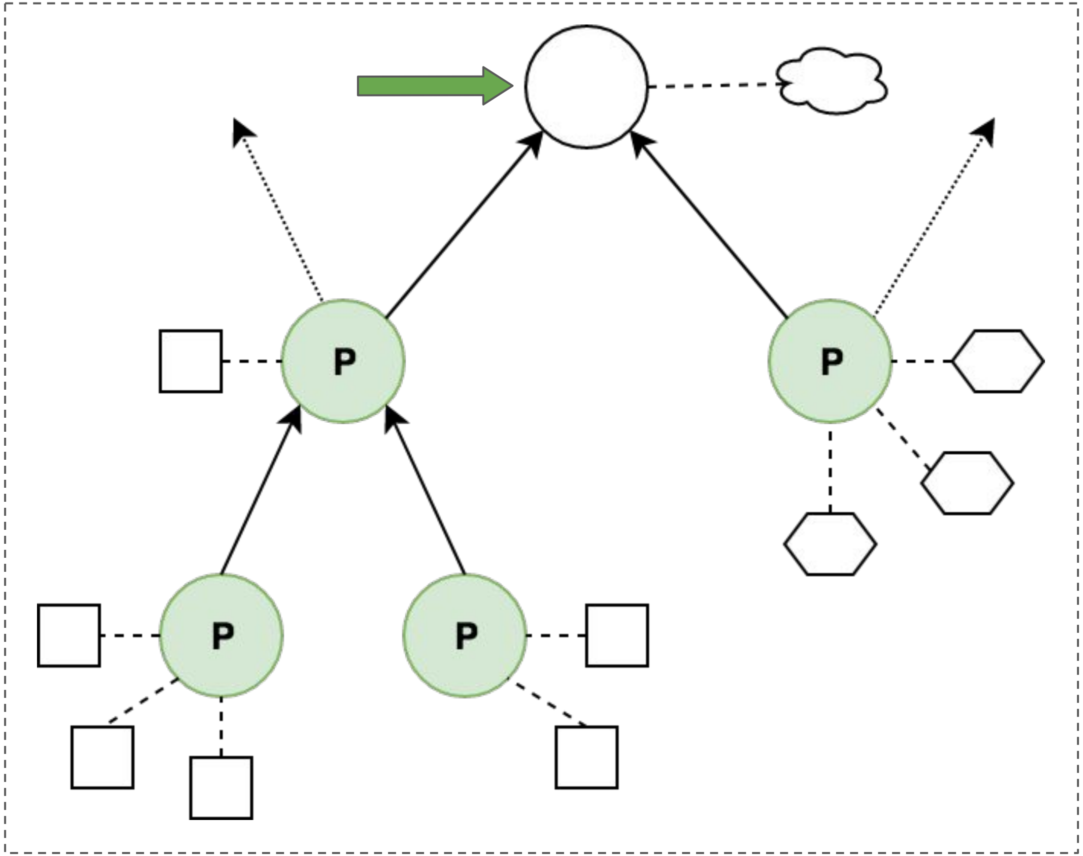

6
4
2
0
Square
Hexagon
Cloud

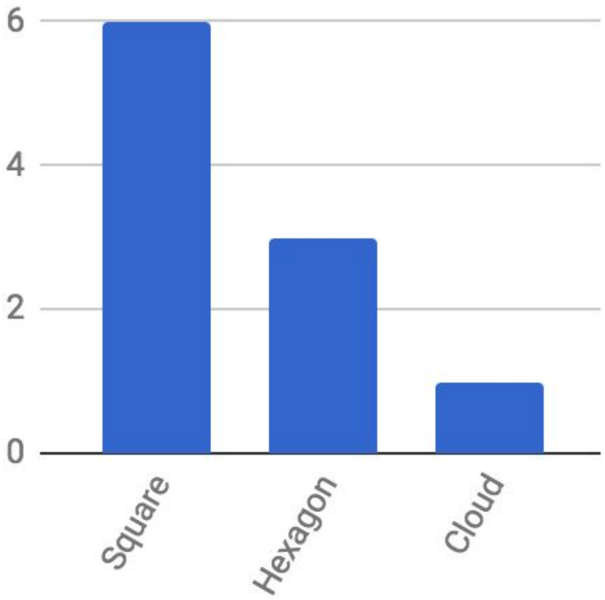

P
P

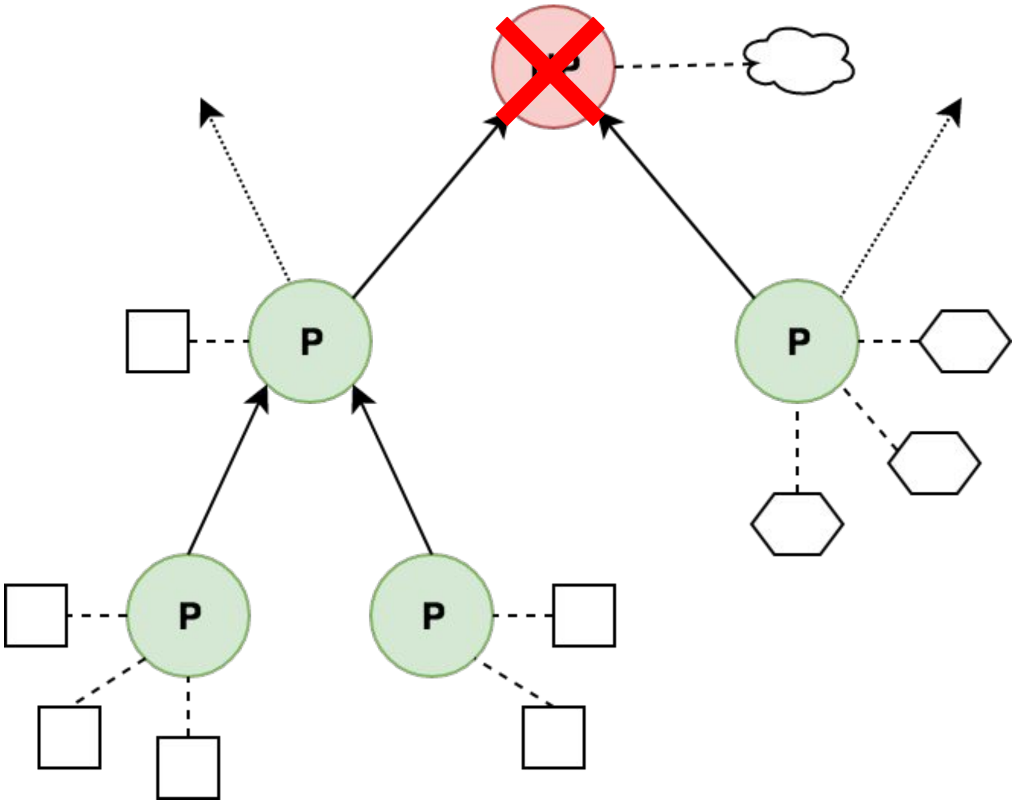

6
4
2
0
Square
Hexagon
Cloud


P
P
P

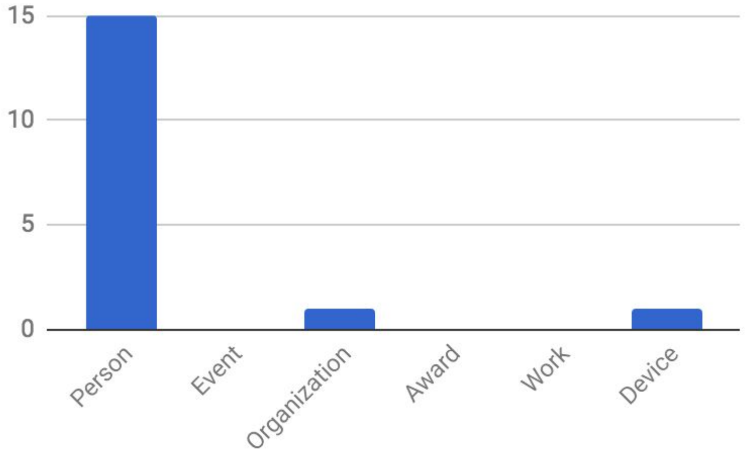

6
4
2
0
Square
Hexagon
Cloud


P
P
P
P
P


6
4
2
0
Square
Hexagon
Cloud



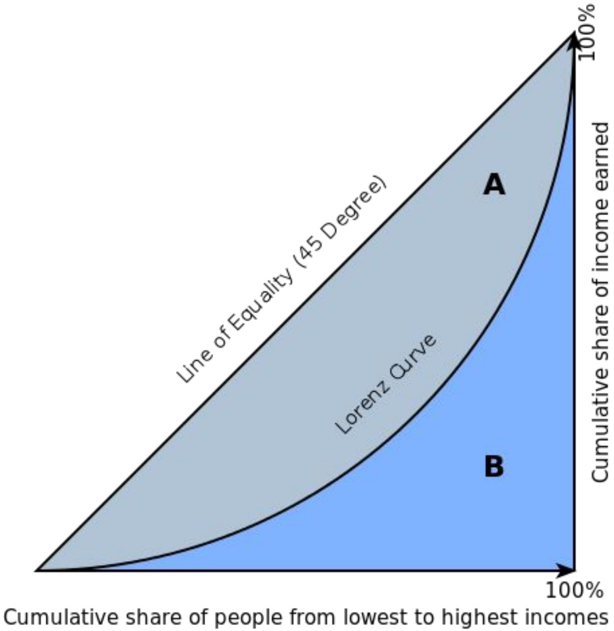
The category is marked as not-pure

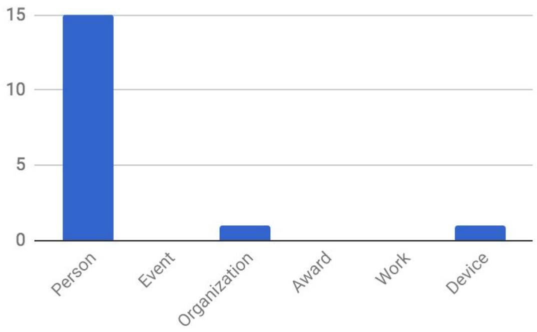


In the real world we don’t have perfect pure categories!

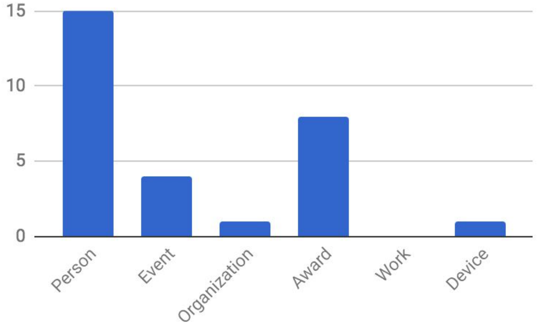
Pure categories have a high unbalanced types distribution

Gini coefficient


100%
Line of Equality (45 Degree)
A
Lorenz Curve
B
Cumulative share of income earned
100%
Cumulative share of people from lowest to highest incomes

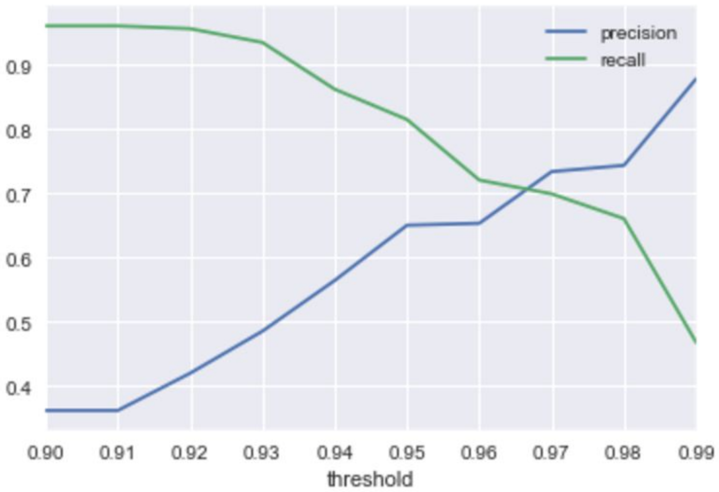


✓



✗

How can we select a Gini good threshold?



We manually annotated ~700 **is-a** (transitive)

is-a(*British scientist, People*) ✓

is-a(*Gaelic games grounds, Football*) ✗

We can prune the network with different thresholds

Conclusion

We presented a language-independent framework to clean the category network

Customizations:

Types from Wikidata, DBPedia, etc

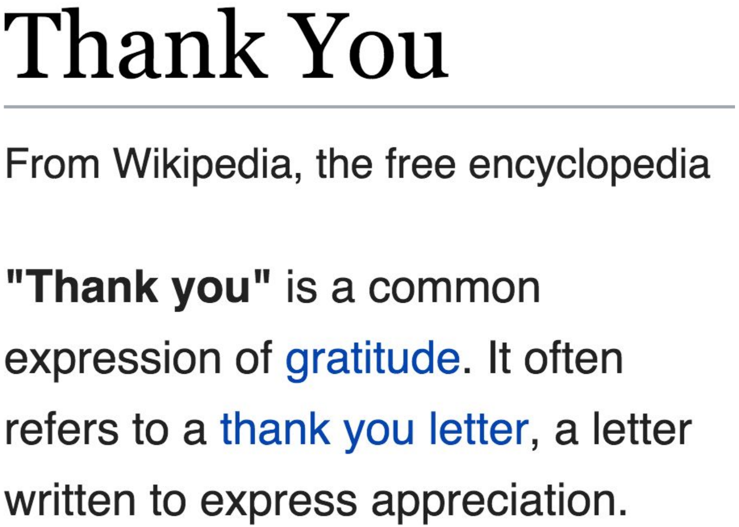
Dynamic threshold for the Gini coefficient

Gini coefficient selection

Deletion of edges vs. nodes

Read more about sections recommendation:

https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages



Thank You

From Wikipedia, the free encyclopedia

"Thank you" is a common expression of gratitude. It often refers to a thank you letter, a letter written to express appreciation.

✉ tiziano.piccardi@epfl.ch

🐦 @tizianopiccardi