Information extraction

2. Knowledge representation

Simon Razniewski
Winter semester 2019/20

Announcements

- Assignment results online
- Thanks to all that provided additional info on the survey
- Registration
    - Pass ≥ 6 assignments
    - Register in HISPOS till 7.1.2020
    - No other registration
- Further reading: now on website

Goal today: Model anything

- Anything? Cats (Q146), submarines (Q2811), philosophical schools (Q16895642), …
- What's different from databases?
  → Enormously rich schemas
  → Dynamics
- What follows is the standard data model of web-scale KBs, and the semantic web
  - Builds upon Database 101

Motivation (CACM 2019)

“**Knowledge representation is a difficult skill to learn on the job**. The pace of development and the scale at which knowledge-representation choices impact users and data do not foster an environment in which to understand and explore its principles and alternatives. The importance of knowledge representation in diverse industry settings [..] should reinforce the idea that **knowledge representation should be a fundamental part of a computer science curriculum** – as fundamental as data structures and algorithms.”

- industry experts behind Google, Microsoft, Facebook, Amazon, IBM knowledge graphs

Outline

- Entities and classes
- Relations
- Binary relations
- Schema
- Knowledge graphs
- Reification
- Canonic entities
- Open-world assumption
- Lab 2

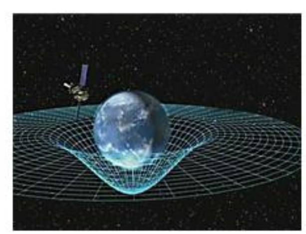
Acknowledgment

- Slides courtesy of Fabian Suchanek (Telecom ParisTech University)

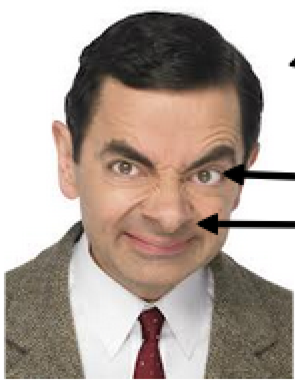
Entity

An entity (also resource, item, object) is any particular object of the world or of imagination, be it abstract or concrete




Is this a good definition?



Is this an entity?

Or this?

How many entities are there?

Entity permanence?

Over time, all parts of a ship are replaced at some point of time. Then, is it still the same ship?

see: Theseus's ship on Wikipedia

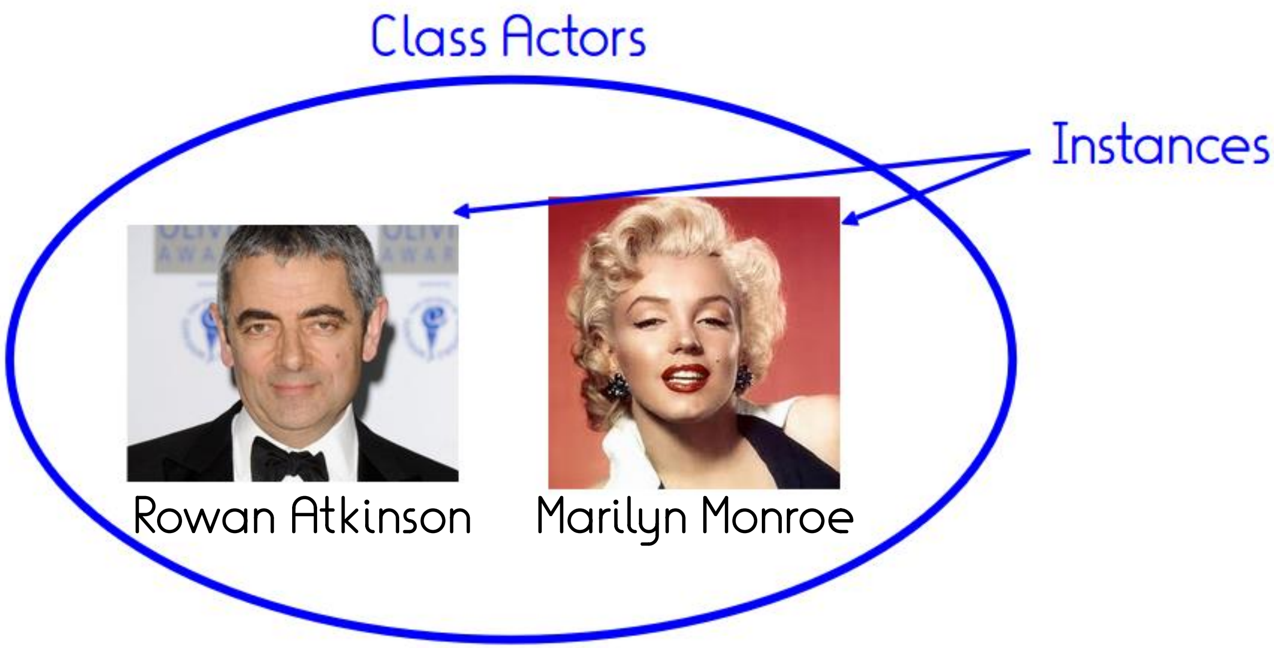
Humans replace their cells every 7 years

Class

A class (also: concept) is a set of similar entities.

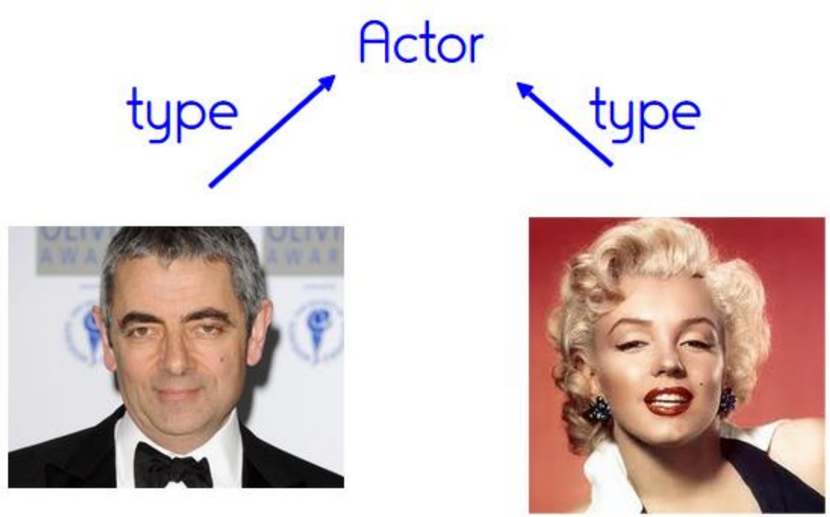
Entities that are not classes (and not literals, relations, ids) are called instances (or common entities).

Classes:
- Actors
- Cars
- Cities
- Rivers
- Universities
- Theories
- ...



Def: Instance of a class

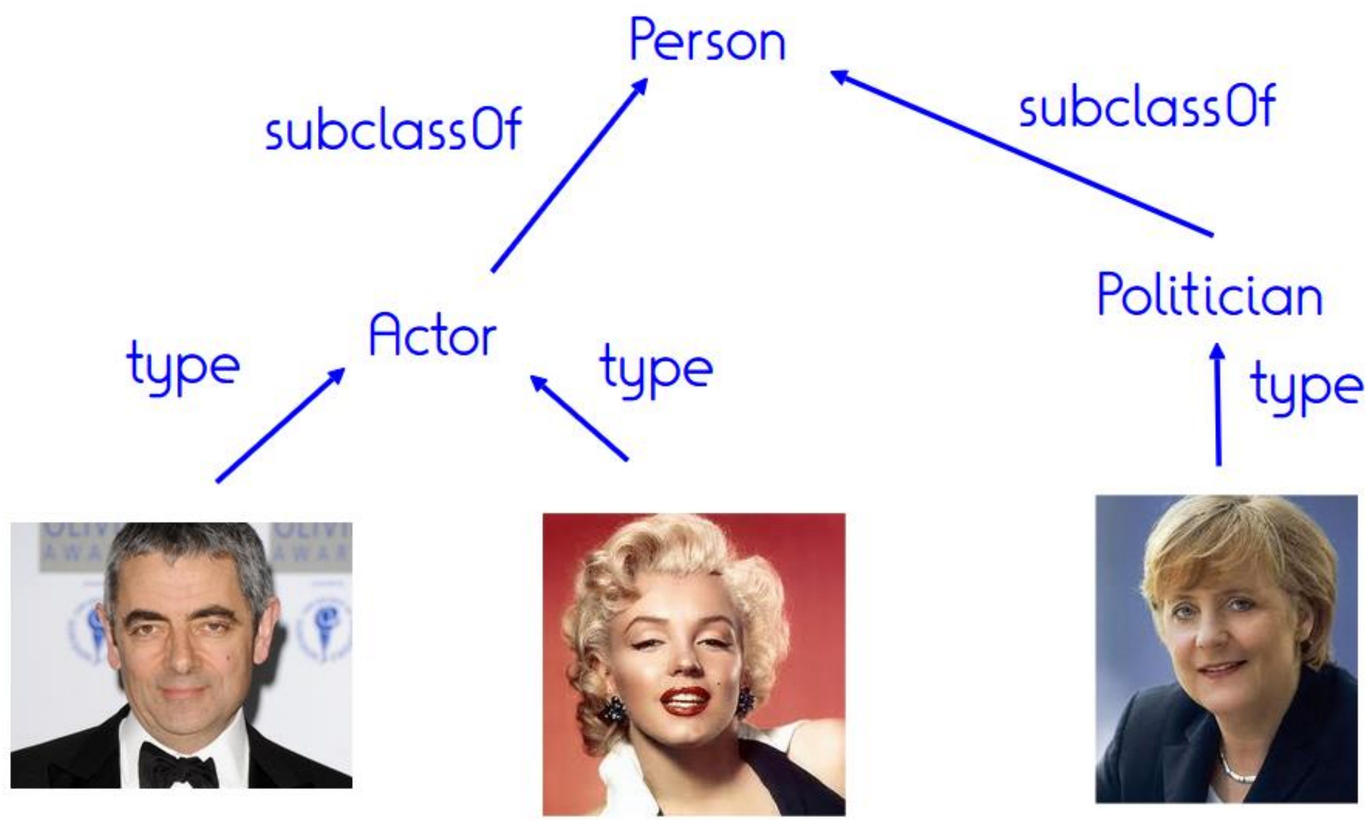
An entity is an instance of a class
(also: belongs to a class, has the type, is of the class),
if the entity is an element of that class.



Def: Subclass, Taxonomy

A class is a subclass of another class, if all instances of the first class are also instances of the second class.

A taxonomy is a hierarchy of classes.



Instance vs. class?

If we can say…	then…
• "a/an X", "every X"	X is a class
• "Xs" (plural)	X is a class
• "This is X"	X is an instance of some class
• "X is a Y"	X is an instance of Y
• "Every X is a Y"	X is a subclass of Y

Try it out: city, Elvis, Coli bacteria, Ford, time

[Finding Needles in an Encyclopedic Haystack: Detecting Classes Among Wikipedia Articles, Pasca, WWW 2018]

Examples

iPhone -> smartphone
finger -> hand
apple -> orange
flower -> plant
Paris -> city
fruit -> food
France -> Europe

YAGO examples

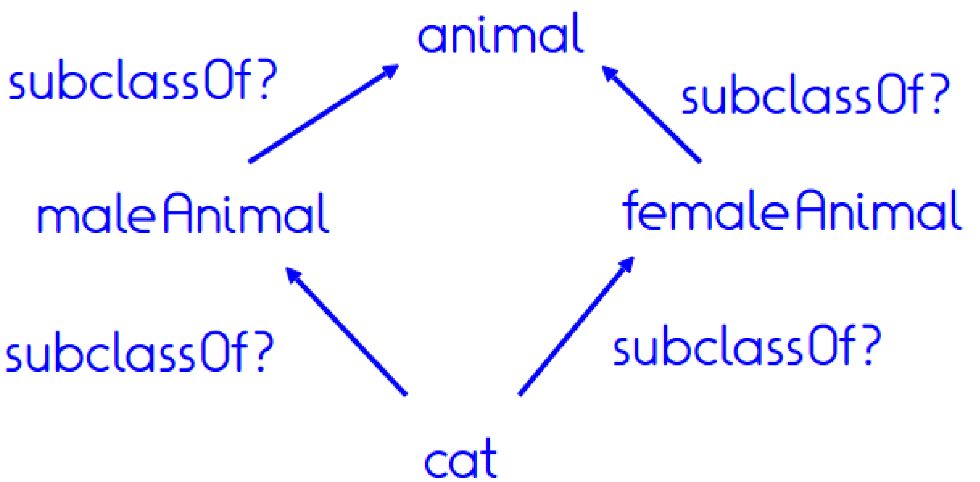
Subject	Property	Object
<Barack Obama>	rdf:type	<wikicategory Audio book narrators>
<Barack Obama>	rdf:type	<wikicategory Columbia University alumni>
<Barack Obama>	rdf:type	<wikicategory Community organizers>
<Barack Obama>	rdf:type	<wikicategory Democratic Party Presidents of the United States>
<Barack Obama>	rdf:type	<wikicategory Democratic Party United States Senators>
<Barack Obama>	rdf:type	<wikicategory Harvard Law School alumni>
<Barack Obama>	rdf:type	<wikicategory Illinois lawyers>
<Barack Obama>	rdf:type	<wikicategory Illinois State Senators>
<Barack Obama>	rdf:type	<wikicategory Living people>
<Barack Obama>	rdf:type	<wikicategory Nobel Peace Prize laureates>

(61 more)

Limitations

Consider a taxonomy of the animal kingdom.

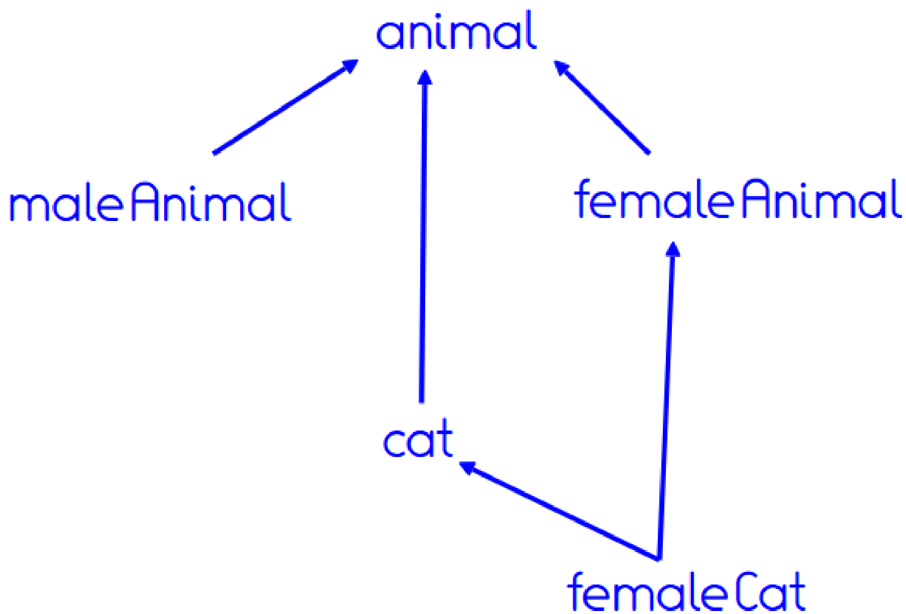
How do we deal with "male" and "female"?



Limitations

Consider a taxonomy of the animal kingdom.

How do we deal with "male" and "female"?



Intuition: Relations

A relation is like a table.

Relation "born":

Person	City	Year
Atkinson	Consett	1955
Monroe	Los Angeles	1926
...		

Def: Relation

A relation (also: predicate) over classes is a subset of their cartesian product. The classes are called the domains of the relation.
The number of classes is called the arity of the relation.

$$R \subseteq C_1 \times C_2 \times \ldots \times C_n$$

$$born \subseteq person \times city \times year$$

$$born = \{<Atkinson, Consett, 1955>,$$
$$<Monroe, LosAngeles, 1926>, \ldots\}$$

Def: Binary Relation, Triple

A binary relation is a relation of arity 2.

$$bornInCity \subseteq person \times city$$

For binary relations, the first class is called the domain and the second class is called the range.

An element of a binary relation is called a fact (or: triple), and we usually visualize it by an arrow:

$$bornInCity(Atkinson, Consett)$$

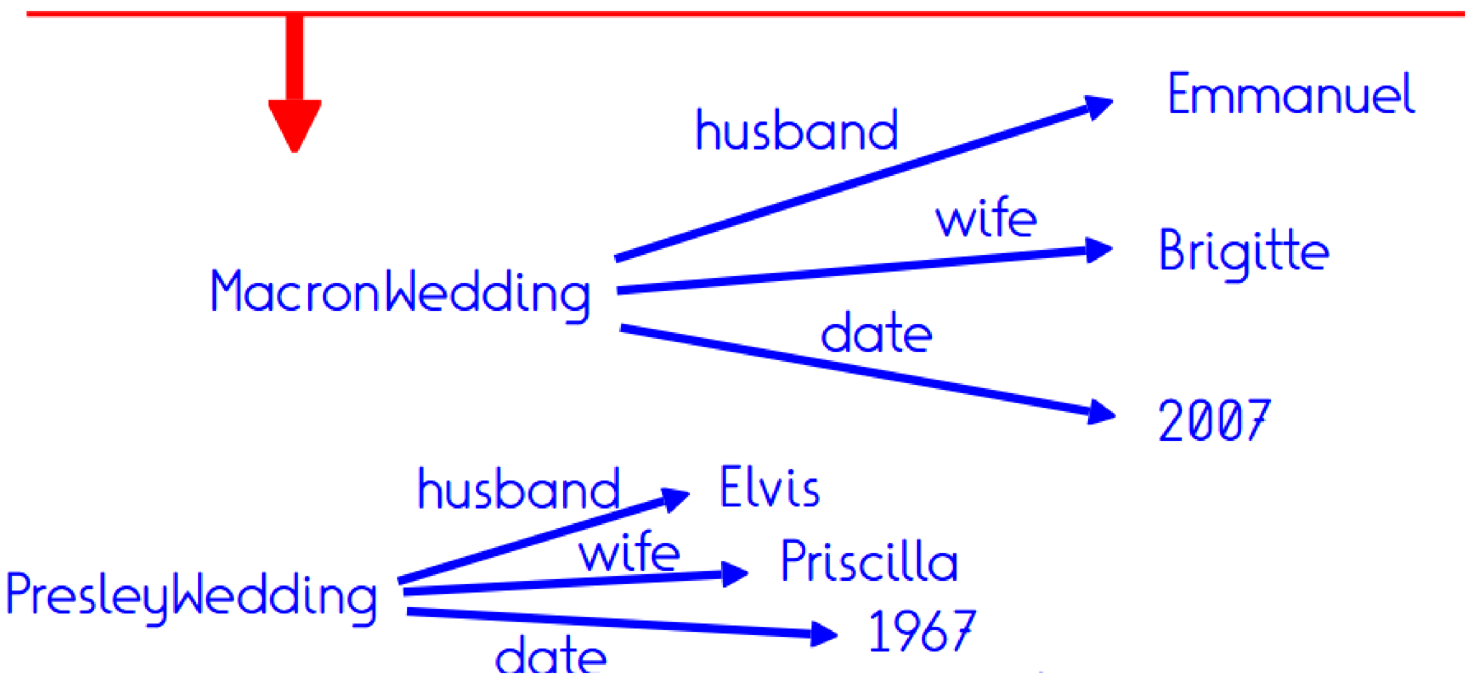
The first argument of a fact is the subject, the second the object.

n-ary facts as binary facts

Every n-ary fact can be represented as binary facts.

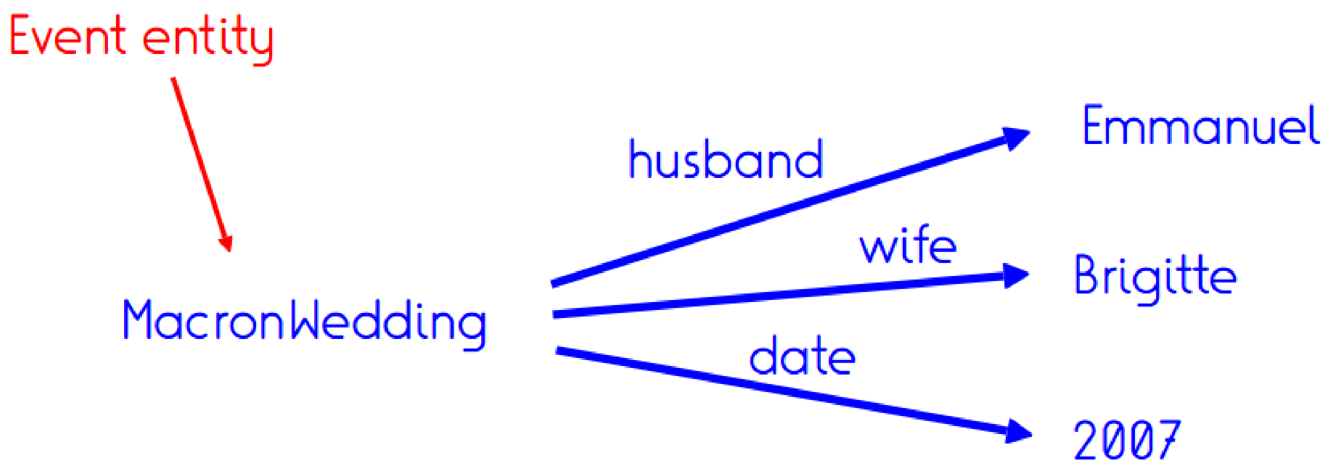
WEDDINGS

Husband	Wife	WeddingDate
Emmanuel	Brigitte	2007
Elvis	Priscilla	1967
...	...	...



Def: Event Entity

An event entity represents an n-ary fact.



Task: Event Entities

Draw a knowledge graph for the following facts.

Irma loves Mr. Bean since 1955.
Mr. Bean drives with Irma to the cinema.
Irma and Mr. Bean watch "Titanic".
The movie is about the trip of the ship "Titanic" from Europe to New York.

(There may be multiple solutions)

Binary relations are flexible

n-ary relations enforce the presence of all arguments:

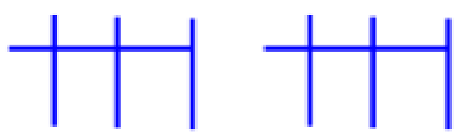
(And nulls blow up the data for high-arity relations)

born	Person	City	Year
	Atkinson	Consett	1955

Binary relations don't:

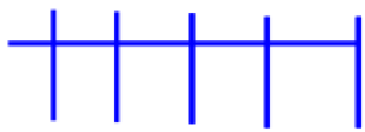
Binary vs n-ary

Binary and n-ary relations can represent the same facts.



binary

- more relations
- less arity
- more flexibility



n-ary

- less relations
- more arity
- more control

Def: Inverse

The inverse of a binary relation r is a relation r', such that $r'(x,y)$ iff $r(y,x)$.

$livesInCity \subseteq person \times city$
$livesInCity(Atkinson, Consett)$

inverses of each other

$hasInhabitant \subseteq city \times person$
$hasInhabitant(Consett, Atkinson)$

Def: Function

A function (also: functional relation) is a binary relation that has at most 1 object for each subject.

$$r \ functional \equiv \forall\, x: |\{y : r(x,y)\}| \leq 1$$

Examples:

- hasBirthPlace
- hasTaxID
- hasNumberOfTeeth

Def: Inverse Functional Rel.

An inverse functional relation is a relation whose inverse is functional.

$$r\ inv.\ functional \equiv \forall\, y{:}\ |\{x : r(x,y)\}| \leq 1$$

Examples:

- hasTaxID
- hasEmailAddress

Functions and inverse functions

- Function+inverse function = identifier
  - hastaxCode, VIAF_Identifier
- Use a relation or its inverse?
  - Preference for "more functional" direction
  - Or add both (Wikidata: hasPart/part of, head of government/position held, …)

Equality

If two entities share the same object of an inverse functional relation, they are equal.

hasPassportNumber(Bean, 29640617)

hasPassportNumber(MrBean, 29640617)

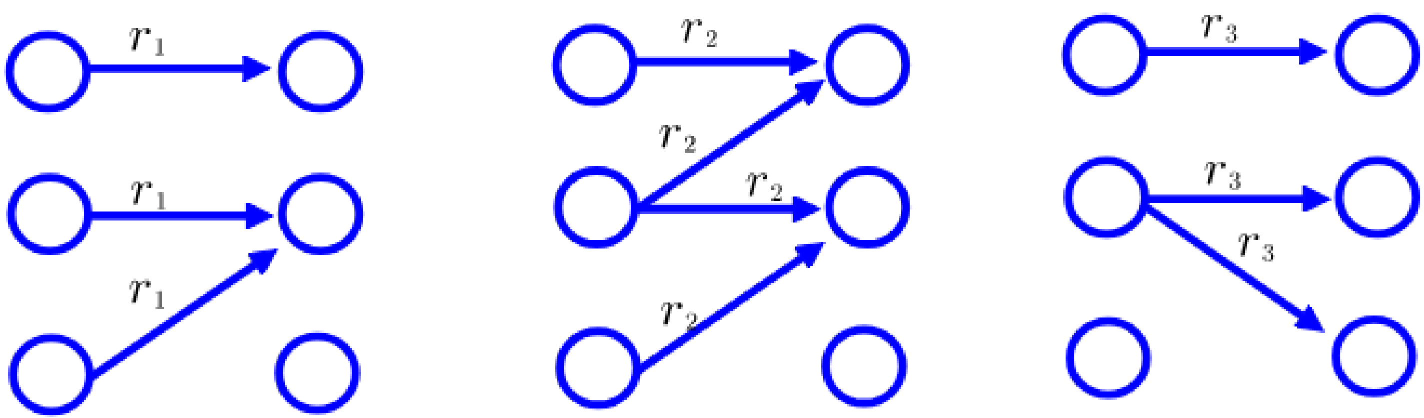
$$\Rightarrow MrBean = Bean$$

$born(Bean, 1955)$

$born(MrBean, 1955)$ (Nothing follows)

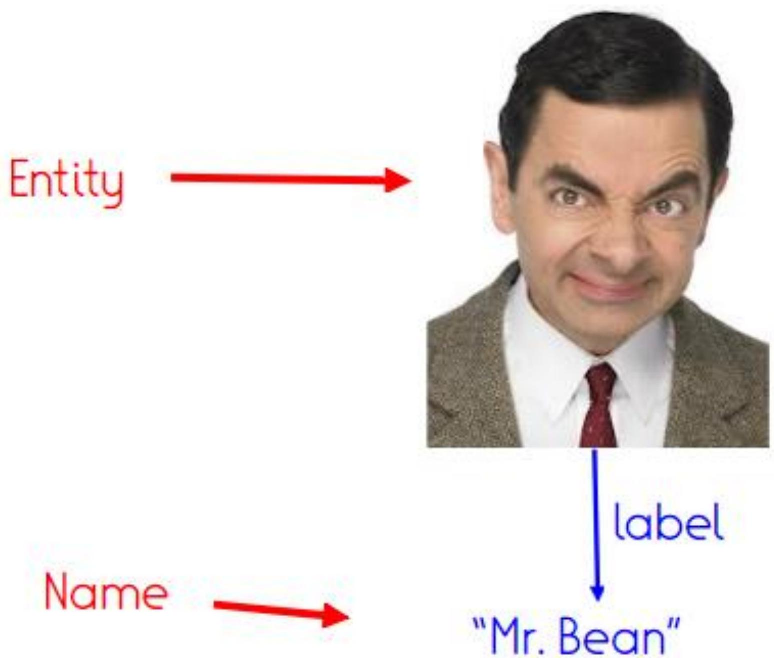
Task: Functions

Which of the following relations are functional?



Def: Name

A name (also: label) of an entity is a human-readable string attached to that entity. The entity is called the meaning of the name.



Def: Synonymy

If an entity has multiple names,
the names are called synonymous.

(The adjective for the names is "synonymous", each name is a "synonym", the phenomenon is called "synonymy")

Donald Trump (Q22686)

45th and current president of the United States

edit

Donald John Trump | Donald J. Trump | Trump | The Donald | POTUS 45 | Donald J Trump | President Donald Trump | President Trump | President Donald J. Trump | President Donald John Trump | DJT

▸ Recoin: Most relevant properties which are absent

▾ In more languages

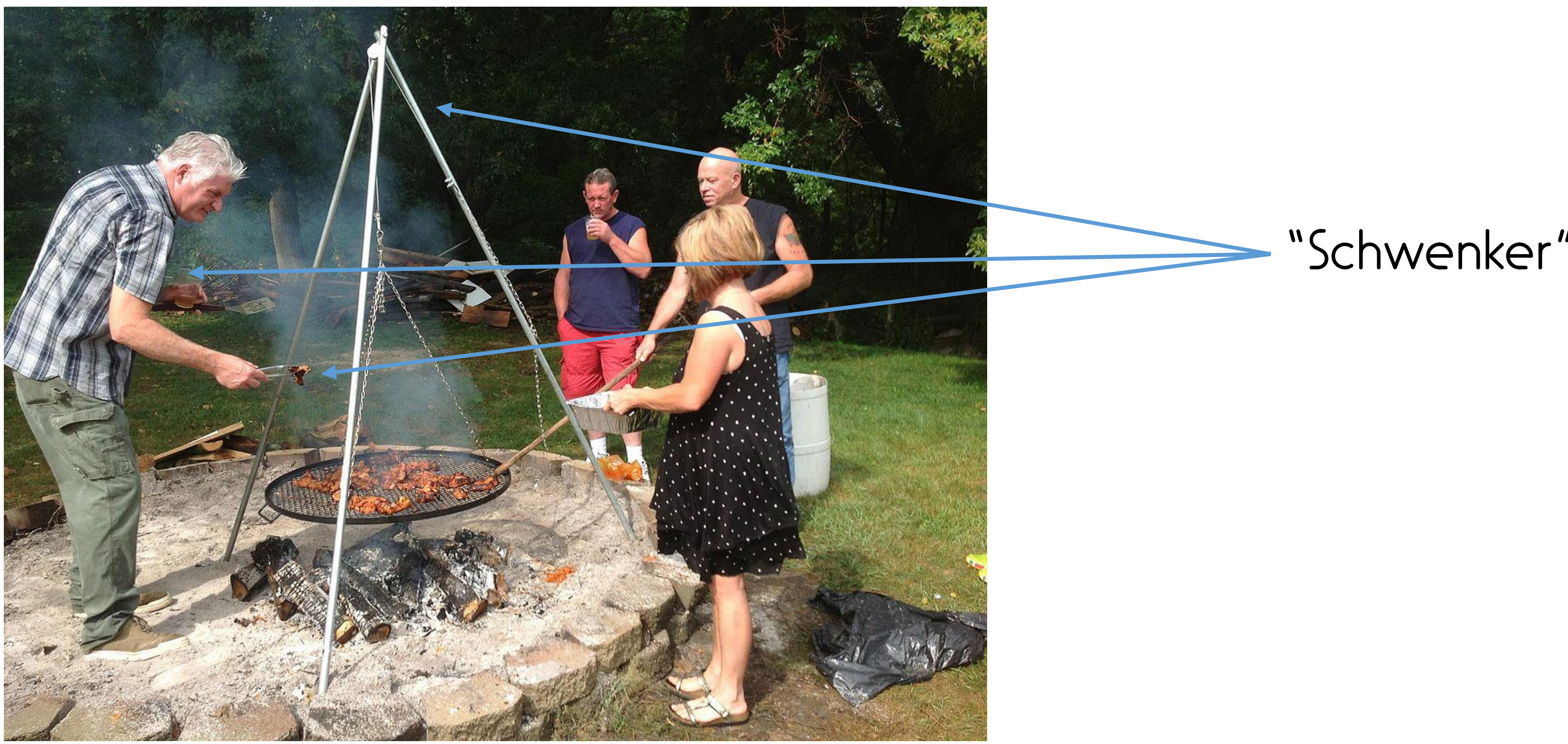
Configure

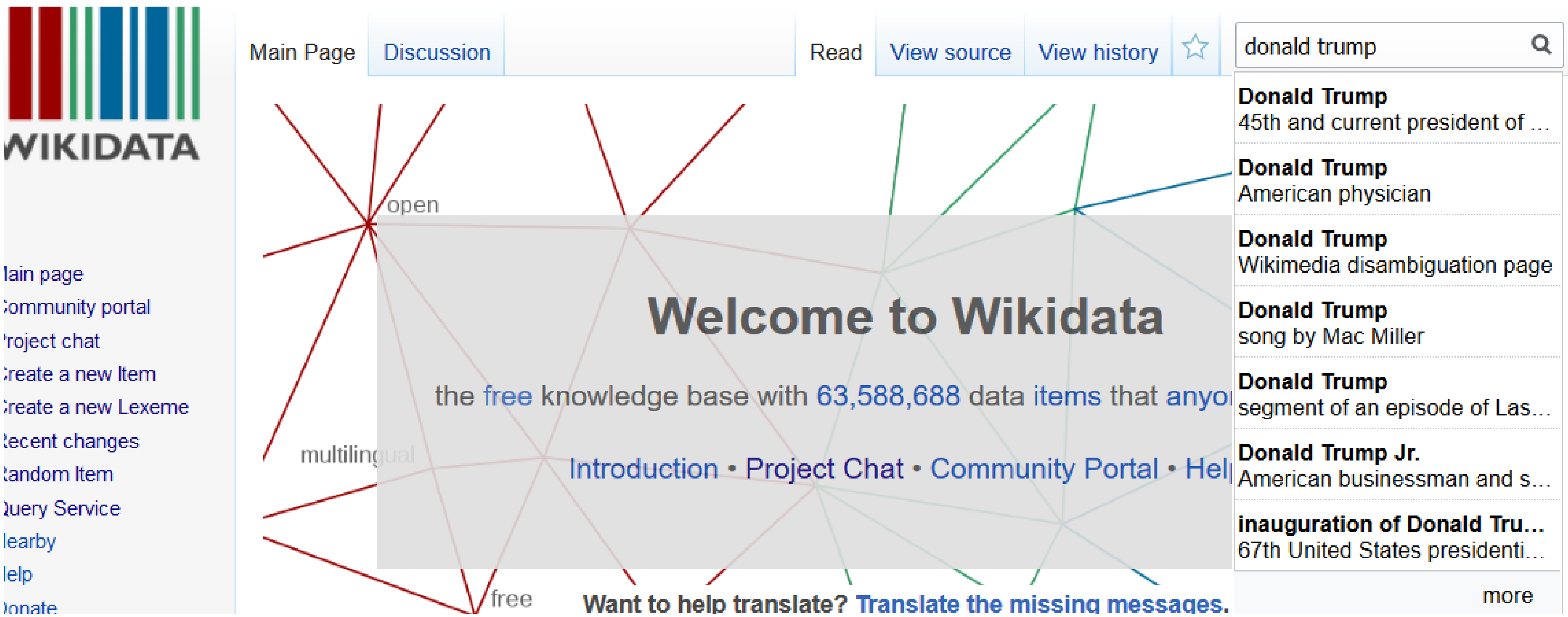
Language	Label	Description	Also known as
English	Donald Trump	45th and current president of the United States	Donald John Trump Donald J. Trump Trump The Donald POTUS 45 Donald J Trump President Donald Trump President Trump President Donald J. Trump President Donald John Trump DJT

Def: Ambiguity

If a name is attached to multiple entities,
the name is called ambiguous.

(The adjective for the names is "ambiguous", the phenomenon is called "ambiguity")

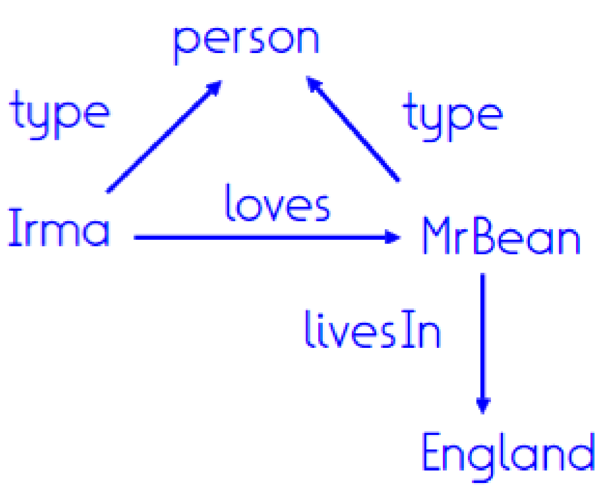



WIKIDATA
Main Page
Discussion
Read
View source
View history
donald trump
Donald Trump
45th and current president of ...
Donald Trump
American physician
Donald Trump
Wikimedia disambiguation page
Donald Trump
song by Mac Miller
Donald Trump
segment of an episode of Las...
Donald Trump Jr.
American businessman and s...
inauguration of Donald Tru...
67th United States presidenti...
more
open
multilingual
free
Welcome to Wikidata
the free knowledge base with 63,588,688 data items that anyo
Introduction • Project Chat • Community Portal • Hel
Want to help translate? Translate the missing messages.
lain page
:ommunity portal
'roject chat
:reate a new Item
:reate a new Lexeme
:ecent changes
:andom Item
)uery Service
learby
lelp
)onate

Def: Knowledge Graph

A knowledge graph (also: Entity-Relationship graph, Knowledge base, KB) is a directed labeled multi-graph that has an edge x->y with label r, iff r(x,y).

$$loves(Irma, MrBean)$$
$$type(Irma, person)$$
$$type(MrBean, person)$$
$$livesIn(MrBean, England)$$



Def: Triple Store

A triple store is a table that contains a KB of binary relations in the form of 3 columns: subject, relation, object.

Subject	Relation	Object
Irma	loves	MrBean
Irma	type	person

(The middle column is often called "Predicate")

Popular triple stores are:

- BlazeGraph
- Jena
- Virtuoso
- ...or classical databases

Classes as binary relations

One way to represent a class is by the binary relations $type$, $subclassOf$.

$$type \subseteq entity \times class$$

$$type(Atkinson, actor)$$

$$subclassOf \subseteq class \times class$$

$$subclassOf(actor, person)$$

Person

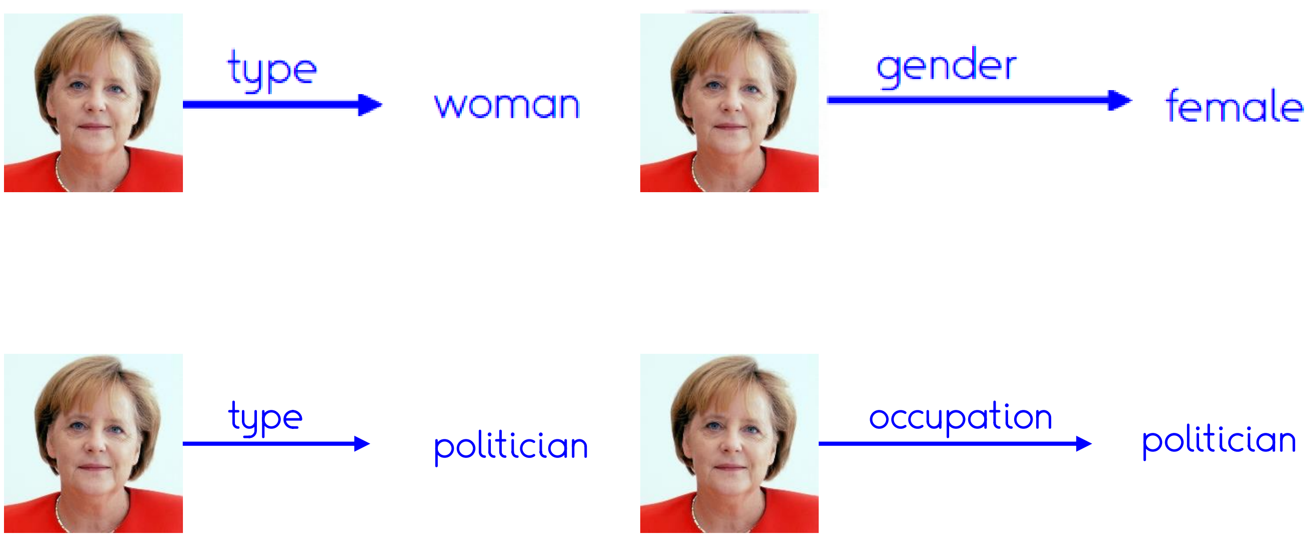
$\uparrow$ subclassOf

Actor

$\uparrow$ type

Digression: Classes and Relations

A fact can be modeled as a class or as a relation.



Domains as binary relations

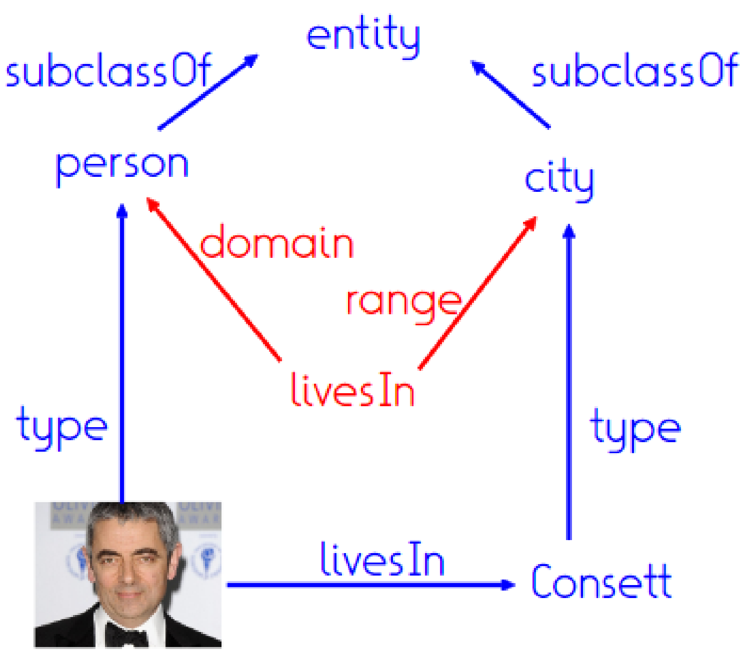
The domain and range of relations can be expressed by binary relations $domain$ and $range$.

$domain \subseteq relation \times class$

$domain(livesIn, person)$

$range \subseteq relation \times class$

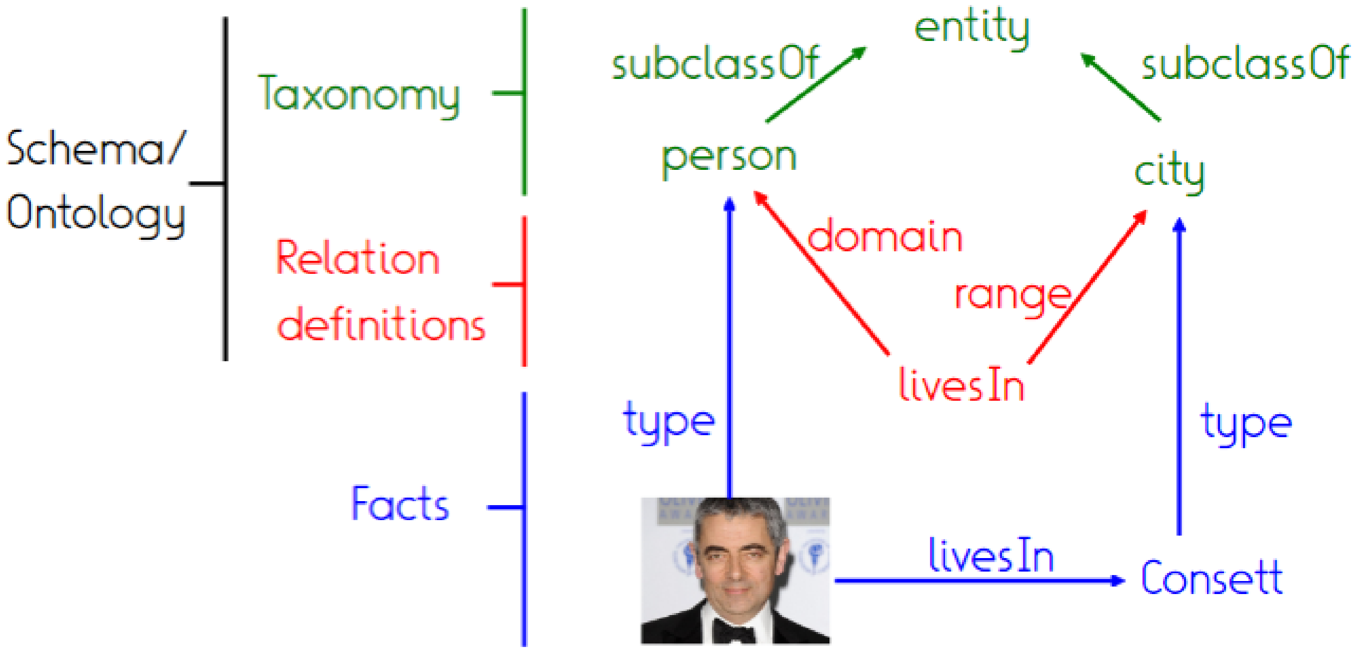
$range(livesIn, city)$



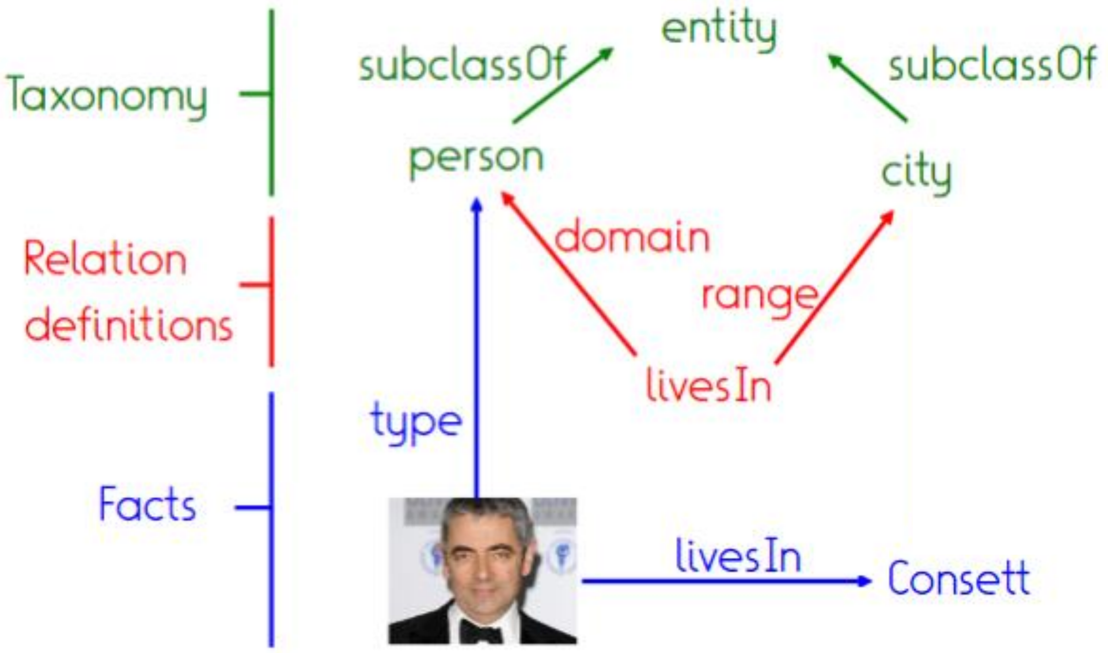
Def: Schema

The schema/ontology is the part of a knowledge graph that consists of

- the taxonomy (= set of classes with their subclassOf-links)
- relation definitions (= a set of relations with domains and ranges)



Inferences



Further reading:
- RDFS
- OWL
- Description logics
- ...

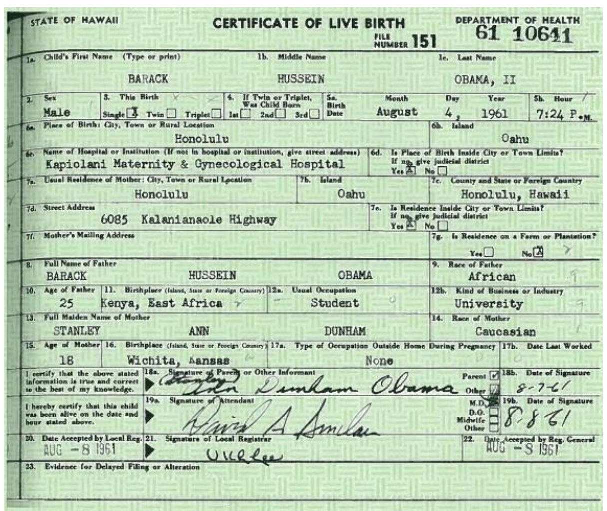
Task: Schema

1. Define a schema for the domain of movies (guided by statements below).

2. In that schema, express that The Audition is a short film, that De Niro and DiCaprio acted in it, and that Scorsese directed it.

Reified statements

A reified statement is an entity that represents a statement. This phenomenon is called reification.


STATE OF HAWAII — CERTIFICATE OF LIVE BIRTH — DEPARTMENT OF HEALTH

FILE NUMBER 151 — 61 10641

1a. Child's First Name (Type or print): BARACK — 1b. Middle Name: HUSSEIN — 1c. Last Name: OBAMA, II

2. Sex: Male — 3. This Birth: Single [X] Twin [ ] Triplet [ ] — 4. If Twin or Triplet, Was Child Born 1st [ ] 2nd [ ] 3rd [ ] — 5a. Birth Date: Month August, Day 4, Year 1961 — 5b. Hour: 7:24 P.M.

6a. Place of Birth: City, Town or Rural Location: Honolulu — 6b. Island: Oahu

6c. Name of Hospital or Institution (If not in hospital or institution, give street address): Kapiolani Maternity & Gynecological Hospital — 6d. Is Place of Birth Inside City or Town Limits? If no, give judicial district. Yes [X] No [ ]

7a. Usual Residence of Mother: City, Town or Rural Location: Honolulu — 7b. Island: Oahu — 7c. County and State or Foreign Country: Honolulu, Hawaii

7d. Street Address: 6085 Kalanianaole Highway — 7e. Is Residence Inside City or Town Limits? If no, give judicial district. Yes [X] No [ ]

7f. Mother's Mailing Address — 7g. Is Residence on a Farm or Plantation? Yes [ ] No [X]

8. Full Name of Father: BARACK HUSSEIN OBAMA — 9. Race of Father: African

10. Age of Father: 25 — 11. Birthplace (Island, State or Foreign Country): Kenya, East Africa — 12a. Usual Occupation: Student — 12b. Kind of Business or Industry: University

13. Full Maiden Name of Mother: STANLEY ANN DUNHAM — 14. Race of Mother: Caucasian

15. Age of Mother: 18 — 16. Birthplace (Island, State or Foreign Country): Wichita, Kansas — 17a. Type of Occupation Outside Home During Pregnancy: None — 17b. Date Last Worked

I certify that the above stated information is true and correct to the best of my knowledge. — 18a. Signature of Parent or Other Informant: Ann Dunham Obama — Parent [X] Other [ ] — 18b. Date of Signature: 8-7-61

I hereby certify that this child was born alive on the date and hour stated above. — 19a. Signature of Attendant: David A Sinclair — M.D. [X] D.O. [ ] Midwife [ ] Other [ ] — 19b. Date of Signature: 8-8-61

20. Date Accepted by Local Reg.: AUG -8 1961 — 21. Signature of Local Registrar — 22. Date Accepted by Reg. General: AUG -8 1961

23. Evidence for Delayed Filing or Alteration

represents

Obama

bornIn

Honolulu

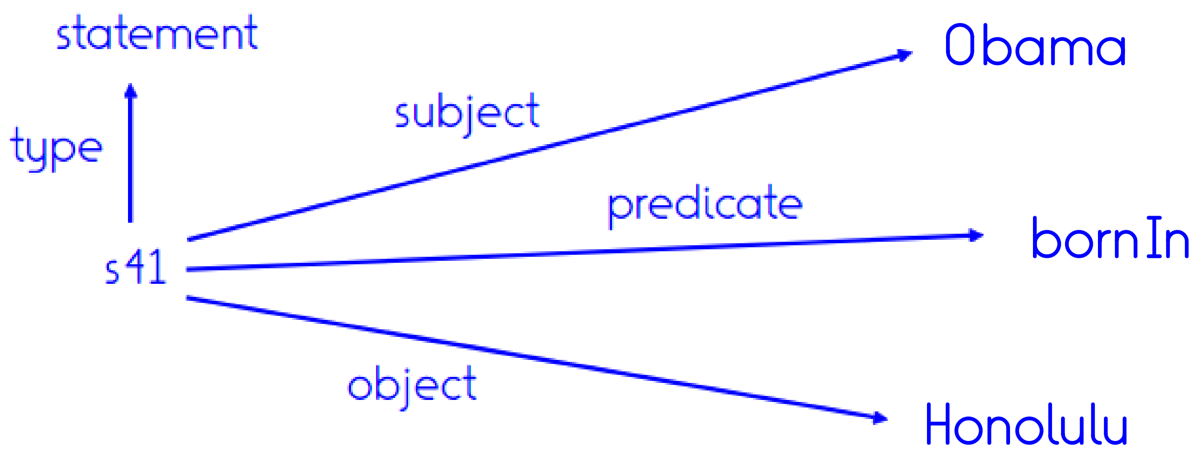
Reification Vocabulary

$statement$ = class of reified statements

$subject \subseteq statement \times entity$

$predicate \subseteq statement \times relation$

$object \subseteq statement \times entity$



Example: Reification

thinks(Trump, s42)

subject(s42, Johnson)

predicate(s42, type)

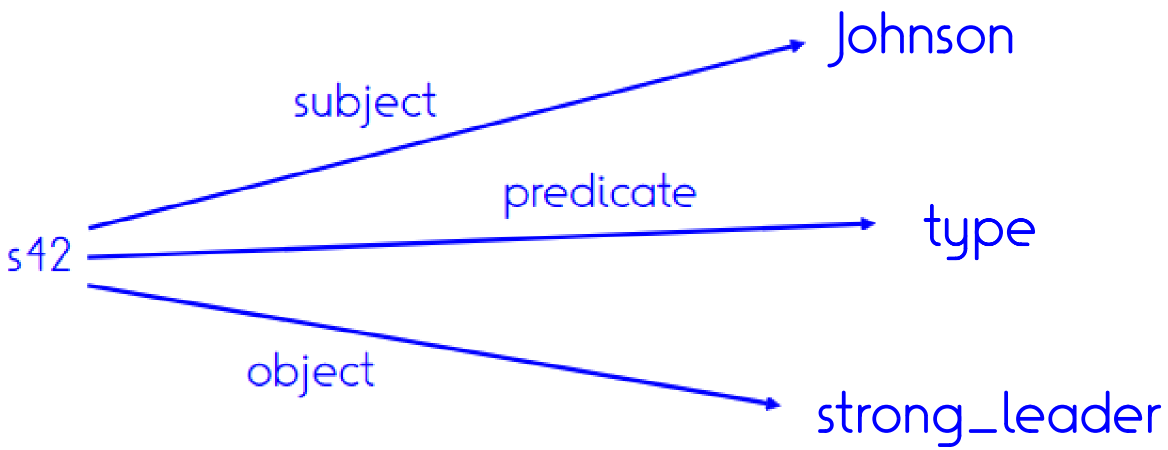
object(s42, strong_leader)

Simplified notation:

thinks(Trump, type(Johnson, strong_leader))

Reification and Event Entities

Just as event entities, reification allows higher-order relations and nesting



Task: Reification

Write down a knowledge base
with some reified facts.
Can you reify facts that have reified arguments?

Def: Canonic Entities

An entity is canonic in a KB, if there is no other entity in the KB that represents the same real-world object.

Alizee	produced	Gourmandises
A. Jacotey	hasProduced	Psychédélices
...	...	...

not canonical

Def: Canonic Relations

An relation is canonic in a KB, if there is no other relation in the KB that represents the same real-world relation.

Alizee	produced	Gourmandises
Alizee	hasProduced	Psychédélices
...	...	...

not canonical

Use of Canonicity

Canonicity is essential for

- Counting
- Confidence consolidation
- Constraint satisfaction

Alizee	produced	Gourmandises
Alizee	hasProduced	Psychédélices
...	...	...

not canonical

Canonicity and Names

A canonic entity can have multiple names.

Alizee	produced	Gourmandises
Alizee	produced	Psychédélices
Alizee	label	"Alizee"
Alizee	label	"A. Jacotey"
produced	label	"produced"
produced	label	"has produced"
...	...	...

Example: Non-canonicity

Open Information Extraction

Hosted by AI2 Created at

Argument 1: entity:Donald Trump **Relation:**

Argument 2: All Search

138 answers from **568 sentences** (results truncated)

You were directed to the entity "Donald Trump".

Show all results for "Donald Trump"

Donald Trump

all person (11) tv actor (9) employer (8) author (7) celebrity (6) misc.

more types

- endorsed **Mitt Romney** (56)
- is running for president (30)
- said in a statement (12)
- is **Idiot** (11)
- is a man (11)
- run for President (10)
- runs **Miss Universe Organization** (10)
- is a joke (9)
- owns the pageant (9)
- is a billionaire (9)
- has endorsed **Mitt Romney** (9)
- said in an interview (8)
- owns **Miss Universe Organization** (8)
- was in **Audience** (8)
- owns **Miss USA** (7)
- is the last person (6)
- attended **Fordham University** (6)
- is endorsing **Mitt Romney** (6)
- is at the top (6)
- is a jerk (6)

Example: Canonicity

WIKIDATA

itemLabel	type	typeLabel
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human
John Smith	wd:Q5	human

Example: Non-Canonicity

Open Information Extraction

"Who built the pyramids?"

Argument 1: **Relation:** built

192 answers from **865 sentences**

all | person (29) | deceased person (16) | location (13) | monarch (12)

Egyptians (278) correct

Pharaoh (41) not bad

useful

Aliens (35) less likely

the Ancient Egyptians (31) duplicate

people (25)

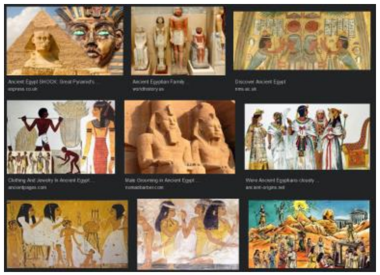
https://openie.allenai.org/

Example: Canonicity

elevation above sea level	95 metre 0 references
architect	Hemiunu 1 reference
architectural style	ancient Egyptian architecture 0 references
heritage designation	UNESCO World Heritage Site start time 1979 0 references

→ No answer

Canonicity as Trade-Off



non-canonic

canonic

- easier to extract
- less easy to use
- more noise
- more data

- difficult to extract
- easy to use
- less noise
- less data

What is the meaning of data?

won	
name	**award**
John	Oscar
Mary	FieldsMedal
Bob	DijkstraAward

	Closed-world assumption	Open-world assumption
won(John, Oscar)?	→ *Yes*	→ *Yes*
won(Ellen, DijkstraAward)?	→ *No*	→ *Maybe*

- (Relational) databases traditionally employ the closed-world assumption (CWA)
- KBs necessarily operate under the open-world assumption (OWA)

Open-world assumption

- Q: *Hamlet written by Goethe?*
  KB*: Maybe*
- Q: *Schwarzenegger lives in Dudweiler?*
  KB*: Maybe*
- Q: *Trump brother of Kim Jong Un?*
  KB*: Maybe*

→ Open-world assumption can be absurd

How to proceed?

- Formal solution:

  Partial-closed world assumption
  - *Uses additional metadata to record where OWA/CWA should be applied*

- Practical implementation:
  - Obtaining metadata not trivial
  - Application-specific

Outline

- Entities and classes
- Relations
- Binary relations
- Schema
- Knowledge graphs
- Reification
- Canonic entities
- Open-world assumption
- Lab 2

Lab 2

- Goals:
  - 1. Model a domain
  - 2. Get to know SpaCy

POS tagging

- Libraries: spaCy

```
import spacy

nlp = spacy.load("en_core_web_sm")

text = ("We import the import that was like a like.")
doc = nlp(text)

for token in doc:
    print(token.text, token.pos_)
```

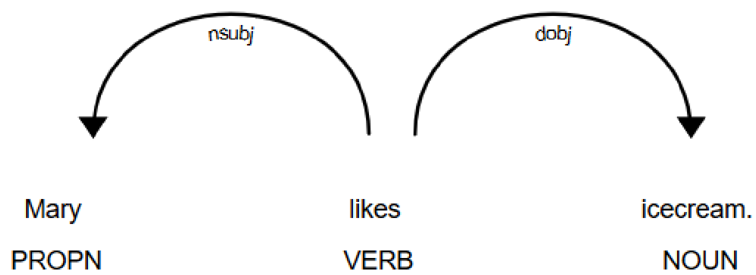
```
We PRON
import VERB
the DET
import NOUN
that DET
was VERB
like ADP
a DET
like INTJ
. PUNCT
```

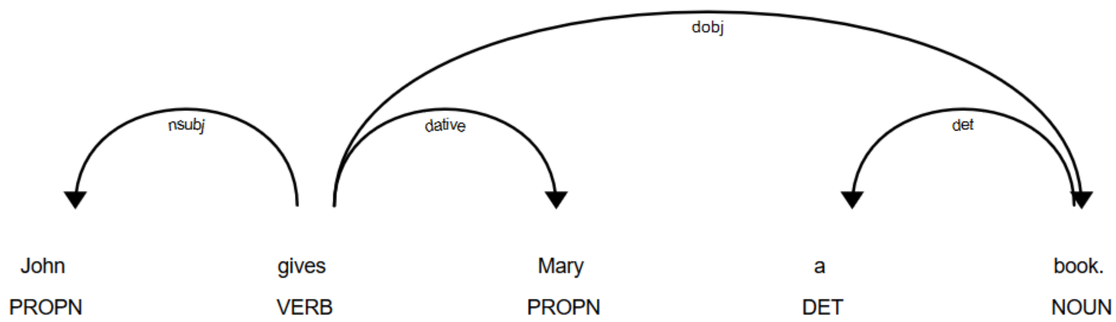
https://spacy.io/usage/spacy-101
https://spacy.io/api/annotation#pos-tagging

Dependency parsing

```
#print(token.text, token.pos_, token.dep_)

import spacy
nlp = spacy.load("en_core_web_sm")

text = ("Mary likes icecream. John gives Mary a book.")
doc = nlp(text)

for token in doc:
    print(token.text, token.dep_)
    for w in token.children:
            print("   " + w.text, w.dep_)

spacy.displacy.render(doc, style='dep')
```

```
Mary nsubj
likes ROOT
   Mary nsubj
   icecream dobj
   . punct
icecream dobj
. punct
John nsubj
gives ROOT
   John nsubj
   Mary dative
   book dobj
   . punct
Mary dative
a det
book dobj
   a det
. punct
```





- https://spacy.io/api/annotation#dependency-parsing -> English

Take home

- Triples can express everything
  - Event entities
  - Reification
- Schema as part of the data
- Canonicity vs. redundancy
- Interpretation of KB data needs caution