Information extraction
1. Introduction

Simon Razniewski

Winter semester 2019/20

Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. Preliminaries & Lab 1

Simon Razniewski

- Senior Researcher at MPII, Department 5
- Heading "Knowledge Base Construction and Quality" area

- Background
  - Assistant professor at FU Bozen-Bolzano, Italy, 2014-2017
  - Research stays at AT&T Labs-Research, University of Queensland, UC San Diego
  - PhD FU Bozen-Bolzano, 2014
  - Diplom at TU Dresden, 2010

- Expertise:
  - Logics, databases, Semantic Web
  - More recently IR, (applied) NLP, ML, …

- Research focus:
  - Analyzing what knowledge bases know, and what they don't

Cuong Xuan Chu

- Doctoral researcher at D5, MPII
- Focus on information extraction for fictional domains and commonsense knowledge

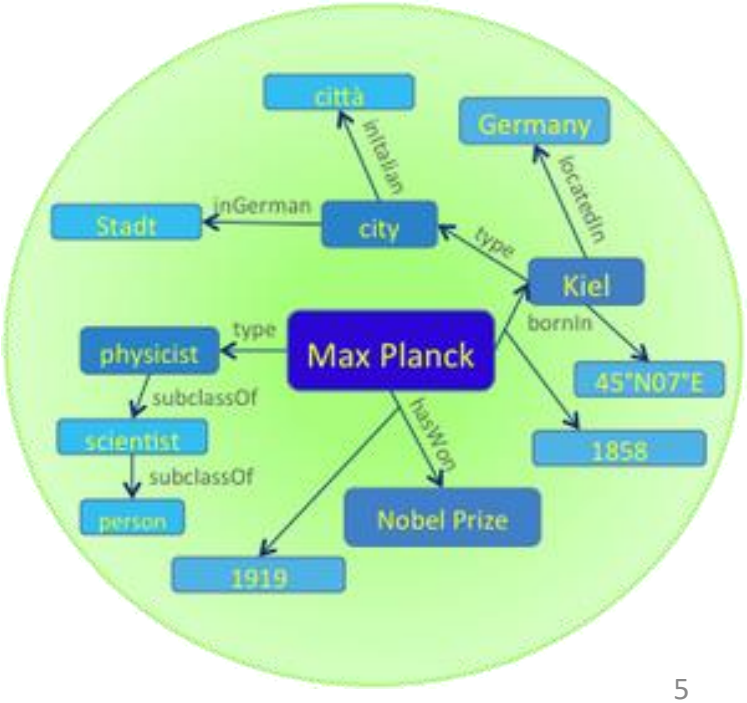
Department 5

- Department 5: Database and information systems, ~25 members

- Knowledge discovery: *extracting*, *organizing*, *searching*, *exploring* and *ranking* facts from *structured*, *semi-structured*, *textual* and *multimodal* information sources

- YAGO Knowledge Base
  - Earliest prominent machine-generated knowledge base (2007)
  - Contains more than 10 million entities and more than 120 million facts

- Gerhard Weikum 259th most cited computer scientist worldwide



And you?

- Course of study
- Preknowledge
- …
- Comments?

- https://tinyurl.com/ie-uds

Outline

1. Introducing each other
2. **Organization of the course**
3. What&why
4. Preliminaries & Lab 1

Learning outcomes

- Knowledge
  - What IE is about (“What”)
  - What IE is good for (“Why”)
  - What main tasks and challenges in IE are
  - What standard approaches to IE are (“How”)

- Skills
  - Analyze potentials and limitations of IE approaches
  - Learn to choose right datasource and method for right task
  - Implement simple solutions for main problems in IE
    - Scraping, typing, linking, …

- Abilities
  - Build your own IE pipeline for an IE problem

→ Very practical focus!

Prerequisites

- Basics of ML
  - We won’t go deep
- Python programming
  - Essential
  - Still time to learn

- Helpful but not required
  - Basic notions of information retrieval (IRDM?)
  - Computational linguistics (SNLP?)

Formal organization

- Credit points: 6, hours: 180 (!)
- Registration
  - Subscribe to the mailing list https://groups.google.com/d/forum/ie1920
  - Register in HISPOS timely before the exam
- When?
  - Lecture (9x): Tuesday 10:00-12:00
  - Lab (9x): Tuesday 16:00-18:00
- How to pass this course?
  - 8 small practical assignments
    - Pass/fail
    - To be admitted to exam, pass at least 6
  - Oral exam

Assignments

- Published on lecture day (Tuesday)
- Due Saturday 23:59 same week

- Labs are there to start solving the assignments

- Discussing assignments together is allowed, but **each student must write their own solution**
  - No sharing of code!
  - Plagiarism = course failed for both
  - Avoid triangular plagiarism = cite sources
    - *"Approach for NER adapted from stackoverflow.com/how-to-…"*

- Libraries that solve core tasks not allowed
  - In doubt ask..

- Weekly assignments are evil!?
  - Psychological trick to help you learn and pass!

Assignment content

- Coding
- 3/7 are assignments in competition format
  - Crisp input/output problem specification
    - "From the first sentence of Wikipedia, extract the type of an entity"
  - Labelled training data set
  - Unseen (hidden) evaluation dataset
    - To avoid overfitting

  → Ranked list by a standard metric, e.g., precision or F1-score
    - But pass/fail does not depend on relative performance

Schedule

	Tentative date	Lecture	Lab
1	15.10.	Introduction	Dataset familiarization (pdf)
2	22.10.	Knowledge representation	Domain modelling
3	29.10.	Crawling and Scraping	Infobox scraping
4	12.11.*	NER, typing and taxonomy induction	Entity typing from Wikipedia first sentence
5	19.11.	Disambiguation	Disambiguation
6	26.11.	Fact extraction	Pattern-fact duality exploration
7	3.12.	OpenIE and evaluation	OpenIE coding
8	10.12.	Rule Mining	Exhaustive short rule evaluation, crowdsourcing
9	17.12.	Applications	Exam preparation
	(7.1.2020)	(Backup slot)	
	14.+15.1.2020	Oral exam	

* Note: No lecture/lab on 5.11.

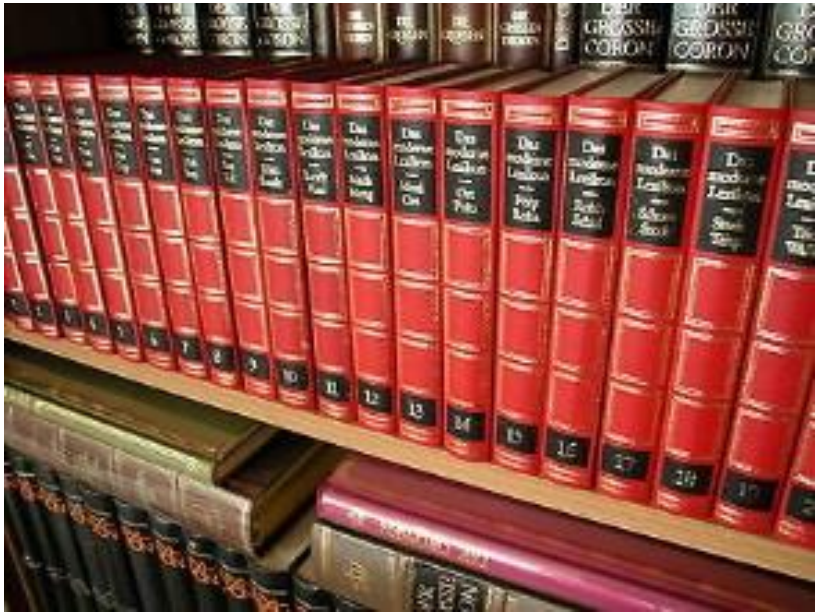
Outline

1. Introducing each other
2. Organization of the course
3. **What&why**
4. Preliminaries & Lab 1

3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future

I. Motivation



- https://en.wikipedia.org/wiki/Max_Planck_Institute_for_Informatics

- https://www.wikidata.org/wiki/Q565400

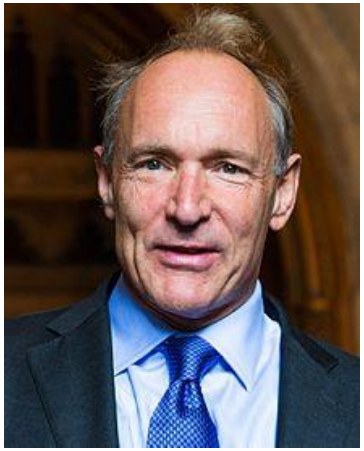
What for?

- One central hub for interlanguage interlinking of 100+ Wikipedia editions
- Your AI chatbot wants to know where MPII, MIT and KAIST are located? → structured query
- A library wants to distinguish which of the 100+ literary John Smiths wrote *"A description of New England"*? → Wikidata ID

Samples of advanced queries

- Who discovered the most planets: http://tinyurl.com/y7rldyqc
- Distribution of places ending with “-weiler” in Germany: https://w.wiki/67o
- Living relatives of Charlemagne: https://w.wiki/67n

The Semantic Web



- Term coined by Tim Berners-Lee for a machine-readable Web
  - Crucial for intelligent agents

- Web content originally from humans for humans

→ Make machines read human language, or make humans write machine-readable structured data?

Machine reading vs. information extraction/ knowledge base construction

3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future

Definitions

Information extraction is the task of transforming **semi/unstructured information** into a machine readable format.

Collections of **machine-readable** information about the **general world** are called **knowledge bases/graphs.**

Common types of machine knowledge

- Lexical knowledge
  - *<shout, isA, verb>*
  - *<shout, subformOf, communicate>*
- Instance knowledge ("Encyclopedic KBs"):
  - *<Paris, capitalOf, France>*
  - *<MPII, foundedIn, 1988>*
  - *<Angela Merkel, major, Physics>*
- Class knowledge ("Commonsense"):
  - *<Pizza, is, tasty>*
  - *<Elephant, color, grey>*
  - *<turnOnPC, requires, power>*

Lexical KBs

- WordNet (1995)
- FrameNet (1998)
- (Wiktionary (2002))
- SenticNet (2010)
- …

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: shout [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) cry, outcry, call, yell, **shout**, vociferation (a loud utterance; often in protest or opposition) *"the speaker was interrupted by loud cries from the rear of the audience"*

Verb

- S: (v) **shout** (utter in a loud voice; talk in a loud voice (usually denoting characteristic manner of speaking)) *"My grandmother is hard of hearing--you'll have to shout"*
- **S: (v) shout, shout out, cry, call, yell, scream, holler, hollo, squall (utter a sudden loud cry)** ***"she cried with pain when the doctor inserted the needle"; "I yelled to her from the window but she couldn't hear me"***
  - *direct troponym* / *full troponym*
  - *verb group*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
  - *phrasal verb*
  - *sentence frame*
- S: (v) exclaim, cry, cry out, outcry, call out, **shout** (utter aloud; often with surprise, horror, or joy) *"`I won!' he exclaimed"; "`Help!' she cried"; "`I'm here,' the mother shouted when she saw her child looking lost"*
- S: (v) abuse, clapperclaw, blackguard, **shout** (use foul or abusive language towards) *"The actress abused the policeman who gave her a parking ticket"; "The angry mother shouted at the teacher"*

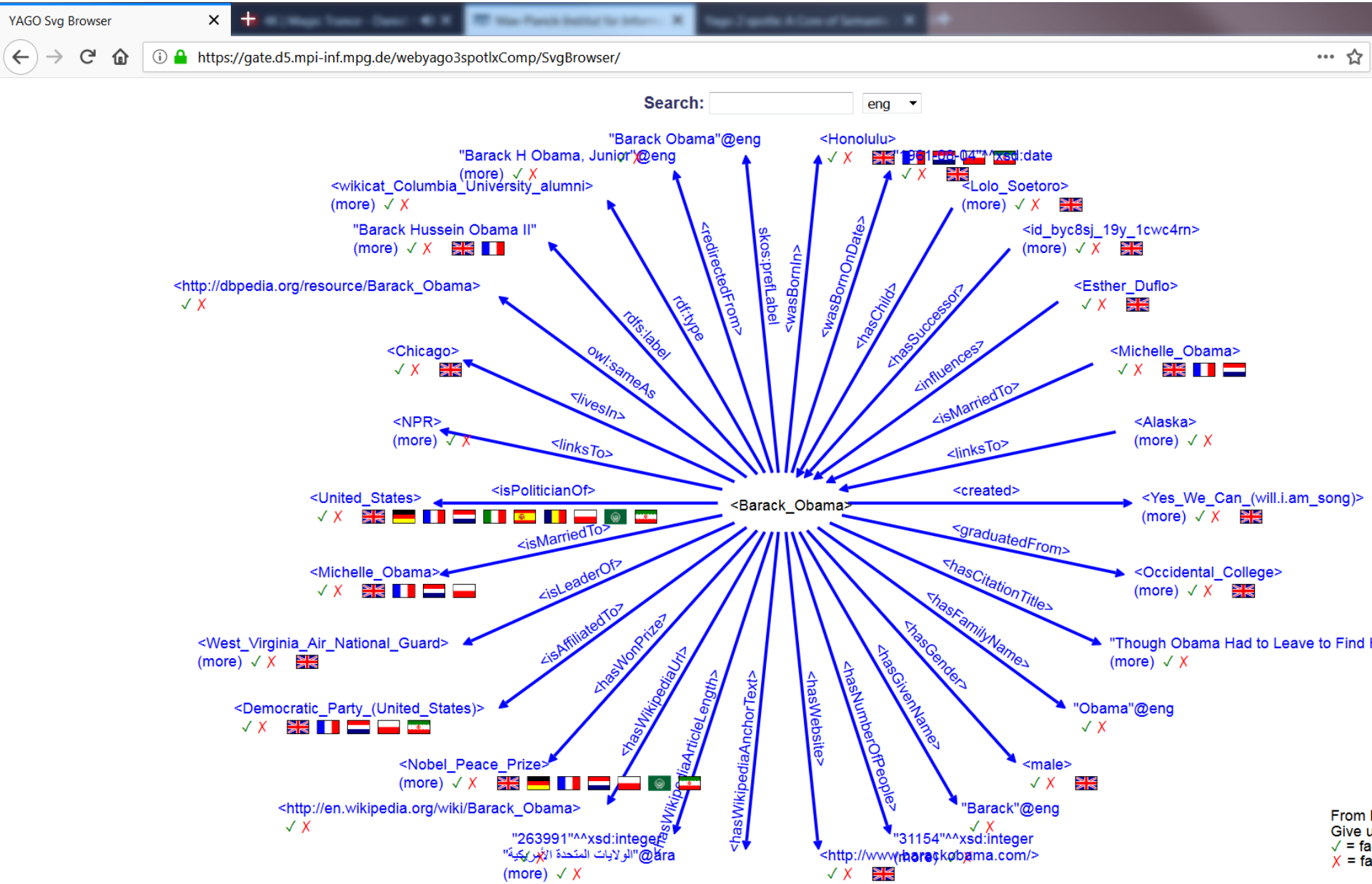
FrameNet

- Example Frame – "Revenge": Because of some **injury** to something-or-someone important to an **avenger** (maybe himself), the **avenger** inflicts a **punishment** on the **offender**. The **offender** is the person responsible for the **injury**.
- Frame elements:
  - **avenger, offender, injury, injured_party, punishment.**
- Invoking terms:
  - Nouns: *revenge, vengeance, reprisal, retaliation*
  - Verbs: *avenge, revenge, retaliate (against), get back (at), get even (with), pay back*
  - Adjectives: *vengeful, vindictive*

Encyclopedic KBs (“Instance-oriented KBs”)

- Cyc (1984)
- YAGO (2007)*
- DBpedia (2007)
- Wikidata (2012)

** developed at MPII*


YAGO Svg Browser
https://gate.d5.mpi-inf.mpg.de/webyago3spotlxComp/SvgBrowser/
Search:
eng
<Barack_Obama>
"Barack Obama"@eng
"Barack H Obama, Junior"@eng
(more) ✓ X
<wikicat_Columbia_University_alumni>
(more) ✓ X
"Barack Hussein Obama II"
(more) ✓ X
<http://dbpedia.org/resource/Barack_Obama>
✓ X
<Chicago>
✓ X
<NPR>
(more) ✓ X
<Honolulu>
✓ X
✓ X
<Lolo_Soetoro>
(more) ✓ X
<id_byc8sj_19y_1cwc4rn>
(more) ✓ X
<Esther_Duflo>
✓ X
<Michelle_Obama>
✓ X
<Alaska>
(more) ✓ X
<Yes_We_Can_(will.i.am_song)>
(more) ✓ X
<Occidental_College>
(more) ✓ X
"Though Obama Had to Leave to Find
(more) ✓ X
"Obama"@eng
✓ X
<male>
✓ X
"Barack"@eng
✓ X
"31154"^^xsd:integer
(more) ✓ X
<http://www.barackobama.com/>
✓ X
<United_States>
✓ X
<Michelle_Obama>
✓ X
<West_Virginia_Air_National_Guard>
(more) ✓ X
<Democratic_Party_(United_States)>
✓ X
<Nobel_Peace_Prize>
(more) ✓ X
<http://en.wikipedia.org/wiki/Barack_Obama>
✓ X
"263991"^^xsd:integer
"الولايات المتحدة الأمريكية"@ara
(more) ✓ X
owl:sameAs
rdfs:label
rdf:type
<redirectedFrom>
skos:prefLabel
<wasBornIn>
<wasBornOnDate>
<hasChild>
<hasSuccessor>
<influences>
<isMarriedTo>
<linksTo>
<livesIn>
<linksTo>
<isPoliticianOf>
<isMarriedTo>
<isLeaderOf>
<isAffiliatedTo>
<hasWonPrize>
<hasWikipediaUrl>
<hasWikipediaArticleLength>
<hasWikipediaAnchorText>
<hasWebsite>
<hasNumberOfPeople>
<hasGivenName>
<hasGender>
<hasFamilyName>
<hasCitationTitle>
<graduatedFrom>
<created>
From
Give u
✓ = fa
X = fa

About: Barack Obama

dbpedia.org/page/Barack_Obama

DBpedia Browse using Formats

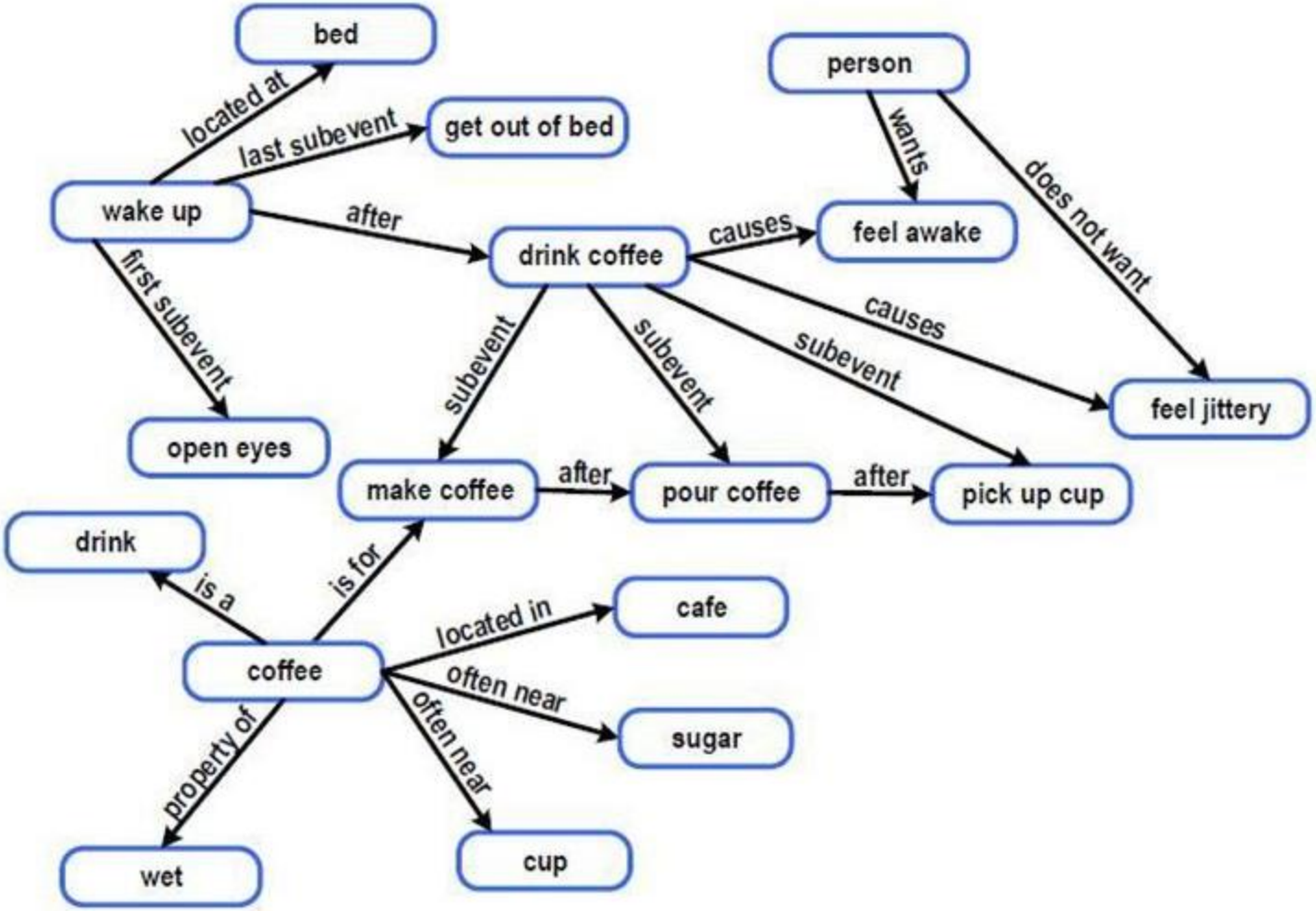
dbo:activeYearsEndDate	▪ 2004-11-04 (xsd:date) ▪ 2008-11-16 (xsd:date)
dbo:activeYearsStartDate	▪ 1997-01-08 (xsd:date) ▪ 2005-01-03 (xsd:date) ▪ 2009-01-20 (xsd:date)
dbo:almaMater	▪ dbr:Occidental_College ▪ dbr:Columbia_College,_Columbia_University ▪ dbr:Harvard_Law_School
dbo:award	▪ dbr:Nobel_Peace_Prize
dbo:birthDate	▪ 1961-08-04 (xsd:date) ▪ 1961-8-4
dbo:birthPlace	▪ dbr:Hawaii ▪ dbr:Honolulu ▪ dbr:Kapiolani_Medical_Center_for_Women_and_Children
dbo:orderInOffice	▪ 44th President of the United States
dbo:party	▪ dbr:Democratic_Party_(United_States)
dbo:region	▪ dbr:Illinois

Commonsense KBs (class-oriented)

- Cyc (1984)
- ConceptNet (1999)
- WebChild (2014)*
- TupleKB (2017)
- Quasimodo (2019)*

** Developed at MPII*

ConceptNet


bed
located at
last subevent
get out of bed
wake up
after
drink coffee
first subevent
open eyes
subevent
subevent
subevent
make coffee
after
pour coffee
after
pick up cup
person
wants
does not want
feel awake
causes
causes
feel jittery
drink
is a
is for
coffee
located in
cafe
often near
sugar
often near
cup
property of
wet

WEBCHILD Commonsense Browser

bicycle

Guess the concept

- Domain
- Comparable
- Physical Part
- Activity
- Property
- Location

Ask me!

bicycle

a wheeled vehicle that has two wheels and is moved by foot pedals

TYPE OF	wheeled_vehicle
	Related to **artifact**, under the category of **cycling**
COMPARABLES	bicycle,bike · bicycle,motorcycle · unicycle,bicycle · bicycle,wheel · bicycle,mountain_bike · More
ACTIVITIES	ride bicycle · buy bicycle · use bicycle · sell bicycle · steal bicycle
HAS PHYSICAL PARTS	axle · bicycle seat · bicycle wheel · brake · casing · More
HAS SUBSTANCE	suspension · hydrogen · oxygen · air · water · More
IN SPATIAL PROXIMITY WITH	street · chain · park · city · rack · More
PHYSICAL PROPERTIES	sensitive · fast · cool · light · small · More
ABSTRACT PROPERTIES	welcome · old · safe · good · important · More
OTHER PROPERTIES	cheap · dangerous · lucky · wobbly · hard · More
ASSOCIATED WITH COUNTRY	united_states · denmark · europe · vietnam · germany · More

Examples

- tiger-n-2
- boat
- car,bicycle
- a:ride bicycle

Related Concepts

- mountain_bike
- ordinary
- safety_bicycle
- velocipede
- bicycle-built-for-two
- wheeled_vehicle
- push-bike

Download Dataset!

3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future

Facts (triples) and their constituents

- **Entities**: Objects about which statements can be made
  Paris; Trump; Irony
- **Property**/predicate/relation/attribute: What can be said
  locatedIn(entity, location), worksAt(person, organization), antonymOf(term, term)
- **Fact**/statement/claim/triple: Core building block of KBs
  <Paris, locatedIn, France>

→General form:

<subject, predicate, object>

<s, p, o>

Subjects and objects

- Machine-generated identifiers
  - Wikidata: *Q4262, Q67245*
- Canonical name strings
  - DBpedia, YAGO: *"John_Smith_(politician)"*
- Internationalized resource identifier (IRI)
  - Semantic web: *http://dbpedia.org/resource/Max_Planck*
- General phrases
  - TupleKB: *<industry, grow over, past few decade>*
- Literals: Attribute values that are no entities
  - *www.mpi-inf.mpg.de*
  - Often with units: *1.63m; 54.85° N*

- Same for predicates, sometimes canonicalized, sometimes just text

Classes and class hierarchies

- **Classes**/types: Allow to group similar entities
  Presidents, nouns, Greek gods

- **Type/property hierarchy**: Tree-like hierarchy among types/properties (cf. inheritance in object-oriented programming)
  <Town, subclassOf, Administrative_unit>

Classes

WIKIDATA

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Reasonator

Saarbrucken (Q1724)

capital of the German state of Saarland edit

Saarbrucken

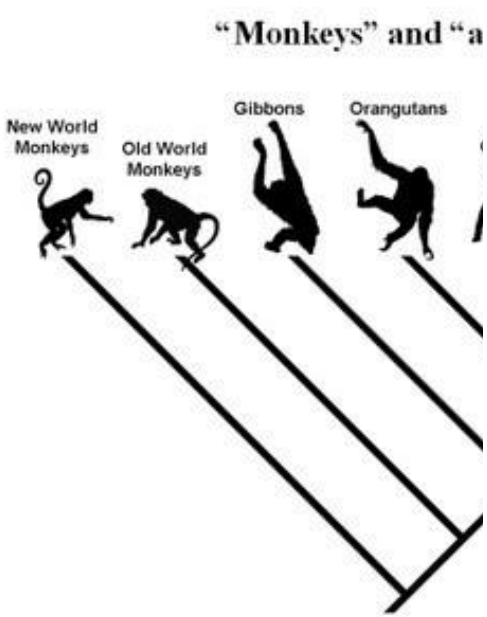
Most relevant properties which are absent

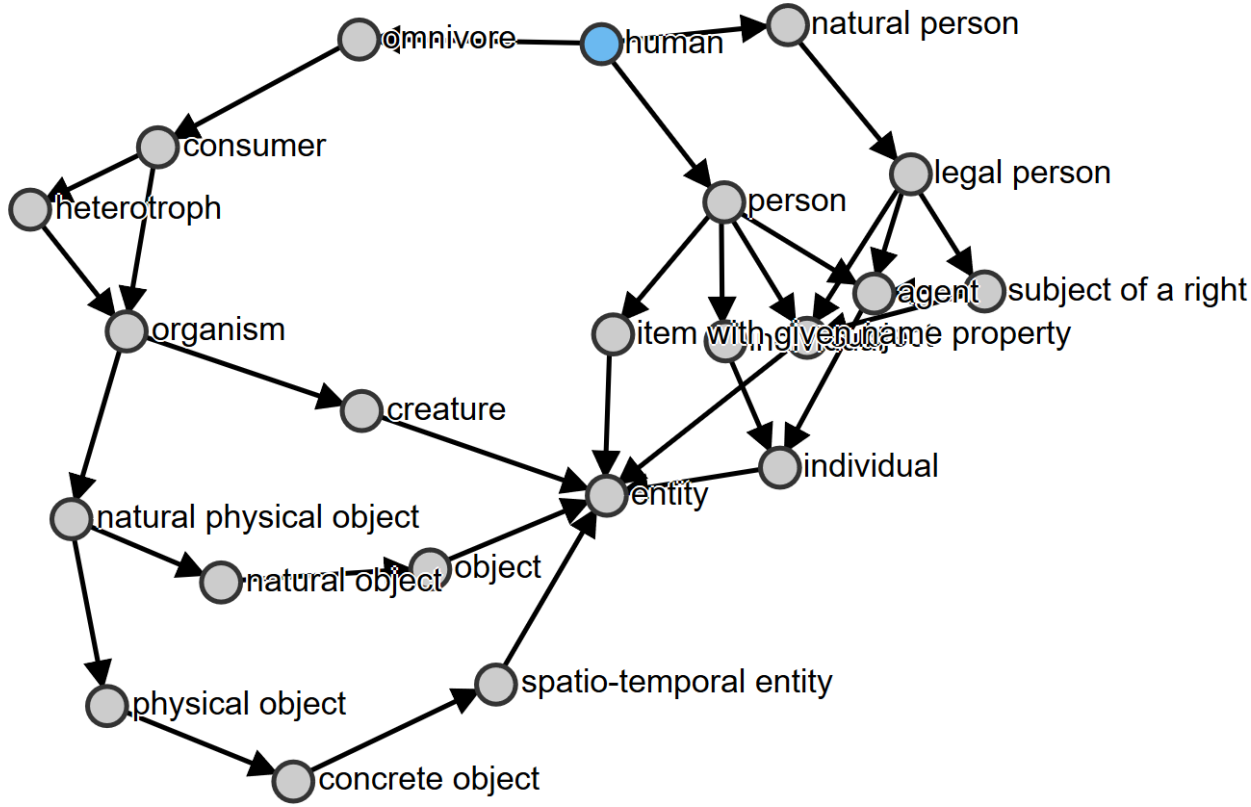
In more languages

Statements

instance of	big city 0 references	edit + add reference
	college town 0 references	edit + add reference
	urban municipality of Germany 0 references	edit + add reference
	state capital in Germany 0 references	edit + add reference
	municipality of Germany 0 references	edit

Taxonomies





https://angryloki.github.io/wikidata-graph-builder/?property=P279&item=Q5

Embedding-based knowledge

- *Apple (0.72 0.35 0.91)*
- *Pear (0.80 0.33 0.55)*
- *Penguin (0.12 0.58 0.27)*

→ Not human-readable

→ Limited machine-readable (meaning of dim. 2?)

- Often impressive performance (e.g., analogies)

3. Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Extraction techniques
V. Technologies
VI. Applications
VII. Past, present and future

How to extract information?

Possible approaches

A. Humans (CYC, ConceptNet, Wikidata)

B. Structured extraction (YAGO, DBpedia)

C. Text extraction (NELL, Textrunner)

D. Constraints and pattern mining

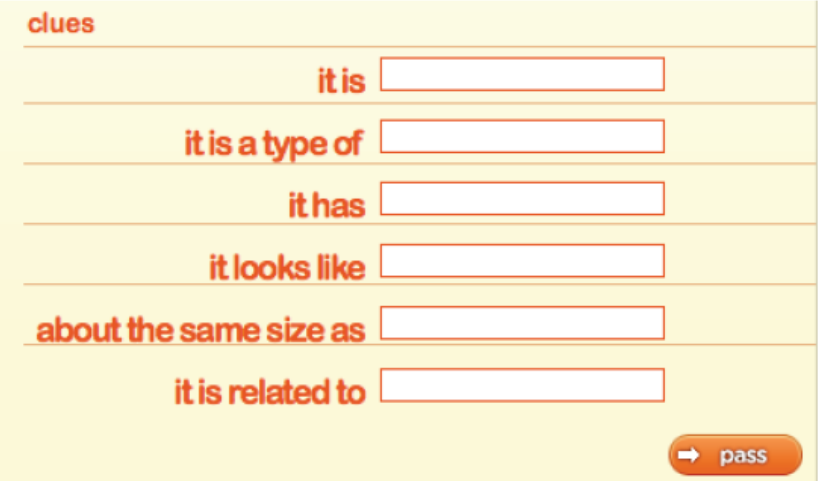
A. Humans: Experts

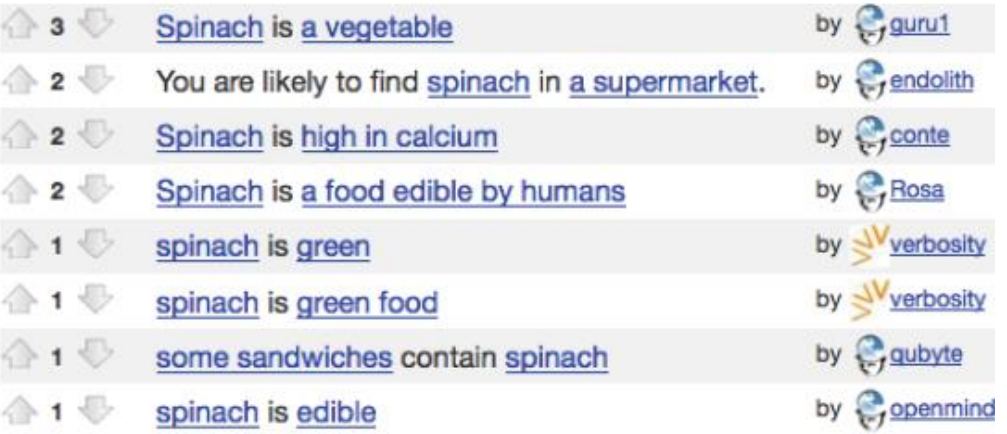
- Potentially best quality

- Difficult to scale
  - CYC: “In 1986, Doug Lenat estimated the effort to complete the KB to be 250,000 rules and 350 man-years of effort.”

Humans: Crowdsourcing/Gamification

- Make work fun (?)





Humans: Volunteers

- Wikidata: 18k active users
- Intrinsic motivation achieves great things
- Broad expertise, compared with selected experts or paid crowdsourcing
- https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

Humans: Challenges

- ConceptNet:
  - Common knowledge, normalization
- Crowdsourcing: Quality assurance
- Wikidata: Modelling and agreement
  - E.g., ethnicity, notable_work, …
  - Multilingual concept alignment

elephant is capable of…

- en carry a trunk →
- en forget to go on the paper →
- en lift logs from the ground →
- en to lift the tree →
- en remember water sources →
- en visit the grocery store →
- en weigh up to 14000 pounds →
- en weight 1000 kilos →

B. Structured extraction

- Wikipedia already provides structured data

- All we need to do is harvest…

Bill Gates

Gates at the United States Department of Health and Human Services in March 2018

Born	William Henry Gates III October 28, 1955 (age 62) Seattle, Washington, U.S.
Residence	Medina, Washington, U.S.
Years active	1968–present
Net worth	US$95.4 billion[1] (August 2018)
Title	Co-Founder and Technology Advisor of Microsoft Co-Chairman of the Bill & Melinda Gates Foundation CEO of Cascade Investment Chairman of Branded Entertainment Network Chairman of TerraPower
Board member of	Microsoft Berkshire Hathaway
Spouse(s)	Melinda French (m. 1994)
Children	3
Parent(s)	William H. Gates Sr. Mary Maxwell Gates
Website	www.gatesnotes.com

Signature

```
{{Infobox person
| name                = Bill Gates
| image               = Bill Gates 2018.jpg
| alt                 = Head and shoulders photo of Bill Gates
| caption             = Gates at the [[United States Department of Health and Human Services]]
2018
| birth_name          = William Henry Gates III
| birth_date          = {{birth date and age|1955|10|28}}
| birth_place         = [[Seattle, Washington]], U.S.
| residence           = [[Medina, Washington]], U.S.
| occupation          = {{hlist|Technology entrepreneur|investor|philanthropist}}
| net_worth           = [[US$]]97.9 billion<ref name="Forbes profile">{{cite web|title=Bill
Gates|url=https://www.forbes.com/profile/bill-gates/|website=Forbes|accessdate=September 12,
</ref> (September 2018)
```

Work done?

- Noise
- Canonicalization of entities and predicates
- Usage of category system

Examples: YAGO, DBpedia

C. Text extraction

- In principle most powerful
  - No need for humans
  - No restriction to Wikipedia existence
- In practice very noisy
  - Canonicalization
  - Consistency
  - …
- Examples: NELL, Textrunner

William Henry Gates III (born October 28, 1955),[2] commonly known as **Bill Gates**, is an American businessman, co-founder and chairman of Microsoft. He is the second richest person in the world just behind Jeff Bezos as of October 2017.[3]

IE demo

- https://www.rosette.com/capability/relationship-extraction/#try-the-demo

- *Merkel is of German and Polish descent. Her paternal grandfather, Ludwik Kasner, was a German policeman of Polish ethnicity, who had taken part in Poland's struggle for independence in the early 20th century.[22] He married Merkel's grandmother Margarethe, a German from Berlin, and relocated to her hometown where he worked in the police. In 1930, they Germanized the Polish name Kaźmierczak to Kasner.[23][24][25][26] Merkel's maternal grandparents were the Danzig politician Willi Jentzsch, and Gertrud Alma née Drange, a daughter of the city clerk of Elbing (now Elbląg, Poland) Emil Drange. Since the mid 1990s, Merkel has publicly mentioned her Polish heritage on several occasions and described herself as a quarter Polish, but her Polish roots became better known as a result of a 2013 biography.*
- *In 1968, Merkel joined the Free German Youth (FDJ), the official communist youth movement sponsored by the ruling Marxist–Leninist Socialist Unity Party of Germany.[30][31][32] Membership was nominally voluntary, but those who did not join found it difficult to gain admission to higher education.[33] She did not participate in the secular coming of age ceremony Jugendweihe, however, which was common in East Germany. Instead, she was confirmed.[34] During this time, she participated in several compulsory courses on Marxism-Leninism with her grades only being regarded as "sufficient".*

Challenges

- Entity identification
- Entity disambiguation
- Relation identification
- Relation normalization
- …

NLP pipeline

- End-to-end models can alleviate these to some extent, but are specific to their training data
  - E.g., DeepDive

D. Constraints

Databases

- Key, foreign key, range, …

Knowledge bases:

- *Events start earlier than they end*
- *Every human must have two parents*
- *Mayors of cities must be humans*
- *The parent of a person's sibling is the person's parent*

- Can be used to…

  … reject KB modifications
  … indicate missing information
  … infer new facts

- But reality is messy..

Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction techniques
V. Technologies
VI. Applications
VII. Past, present and future

Which technologies every information extraction engineer should know about?

Technologies (1): Scraping

- BeautifulSoup for Python web scraping

Beautiful Soup Tutorial — bash — 103×40

```
          YOUNG, Pierce Manning Butler
         </a>
        </td>
        <td>
         1836-1896
        </td>
        <td>
         Representative
        </td>
        <td>
         Democrat
        </td>
        <td align="center">
         GA
        </td>
        <td align="center">
         43
         <br>
          (1873-1874)
         </br>
        </td>
       </tr>
      </tbody>
     </table>
    </br>
   </br>
  </b>
 </center>
 <p align="left">
  <a href="http://bioguide.congress.gov/biosearch/biosearch.asp">
   <b>
    Search Again
   </b>
  </a>
 </p>
 <iframe border="0" height="270" id="rdbIndicator" src="./43rd-congress_files/indicator.html" style="d
isplay: none; border: 0; position: fixed; left: 0; top: 0; z-index: 2147483647" width="100%">
 </iframe>
</body>
</html>
```

Technologies (2): Storing

- RDF for representing data
  - Resource description framework
  - Turtle syntax for triples and data types:

    <Mark_Twain> <author> <Huckleberry_Finn>.
    <Huckleberry_Finn> <description> “A 19th century classic novel”.

  IRIs for unique identification of entities:

    <http://yago-knowledge.org/resource/Mark_Twain>

  Prefixes for shorthand notation:

    @prefix yago: <http://yago-knowledge.org/resource>
    yago:Mark_Twain yago:dateOfBirth 30.11.1835

Technologies (3): Querying

- SPARQL for posing queries
  - Query language inspired by SQL

Wikidata cats: https://w.wiki/33a

Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction techniques
V. Technologies
VI. Applications
VII. History and future

What KBs are good for

- Master data
- Data mining
- Search enhancements
- Question answering
- Language generation
- Entity linking
- Learning more knowledge
- ….

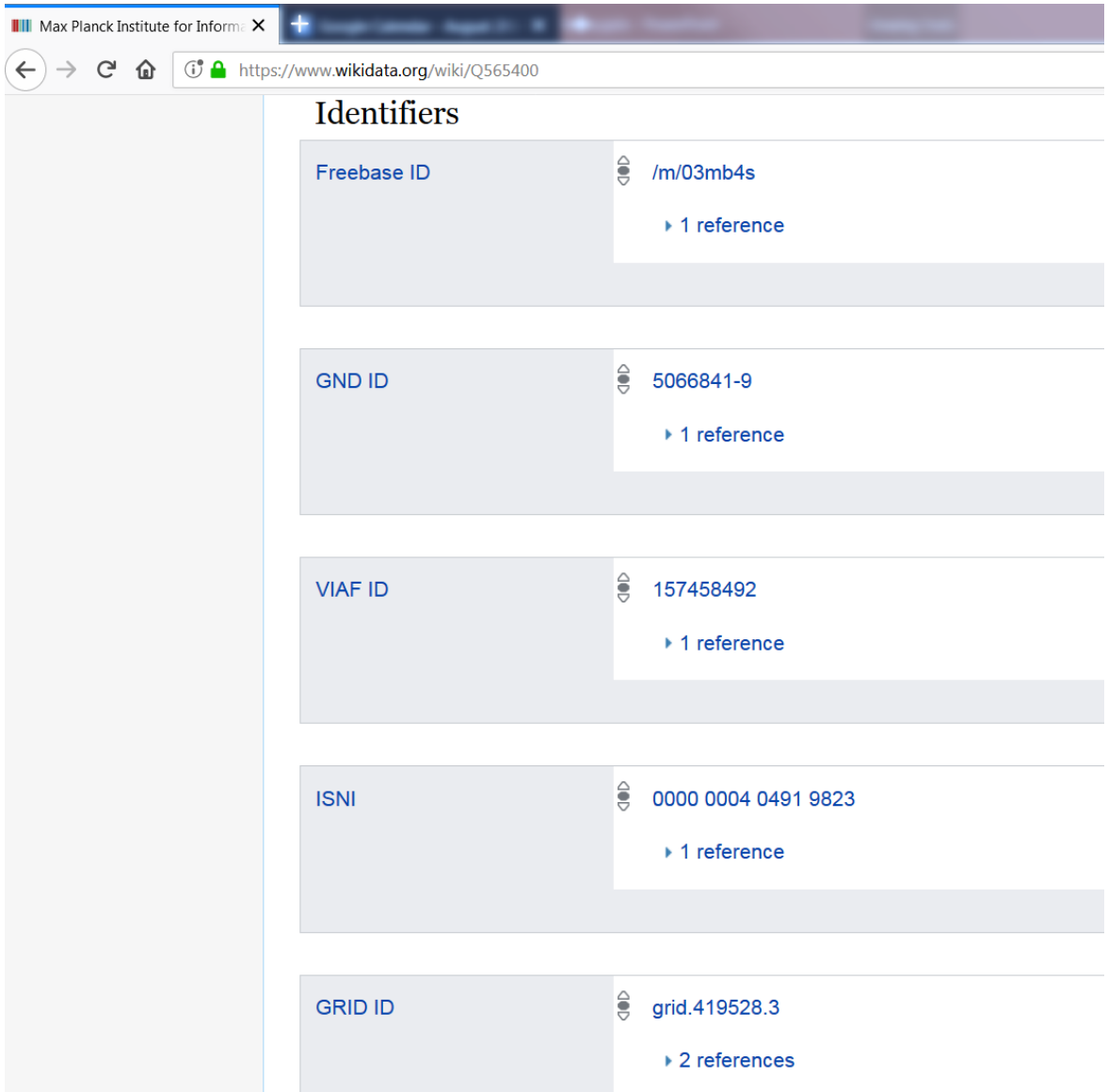
Master data (1)

wd:Q6258248	John Smith
wd:Q6258251	John Smith
wd:Q6258255	John Smith
wd:Q6258259	John Smith
wd:Q6258261	John Smith
wd:Q6258263	John Smith
wd:Q6258265	John Smith
wd:Q6258267	John Smith
wd:Q6258270	John Smith
wd:Q6258271	John Smith
wd:Q6258276	John Smith
wd:Q6258278	John Smith
wd:Q6258281	John Smith
wd:Q6258284	John Smith
wd:Q6258286	John Smith
wd:Q6258288	John Smith
wd:Q6258290	John Smith
wd:Q6258293	John Smith
wd:Q6258294	John Smith
wd:Q6258296	John Smith

(300 more)

https://w.wiki/A4Z

Master data (2)



Relevant for:

- Museums
- Libraries
- Scientific publications

....

Data mining

- Use input facts to extract patterns that allow to predict new facts

$$isCitizenOf(x, y) \Rightarrow livesIn(x, y)$$
$$hasAdvisor(x, y) \wedge graduatedFrom(x, z) \Rightarrow worksAt(y, z)$$
$$wasBornIn(x, y) \wedge isLocatedIn(y, z) \Rightarrow isCitizenOf(x, z)$$
$$hasWonPrize(x, G.\ W.\ Leibniz) \Rightarrow livesIn(x, Germany)$$

isCitizenOf(John, France) → livesIn(John, France)

- Various approaches based on association rule mining and latent models

Entity linking

https://gate.d5.mpi-inf.mpg.de/webaida/

Search enhancements

Google

max planck

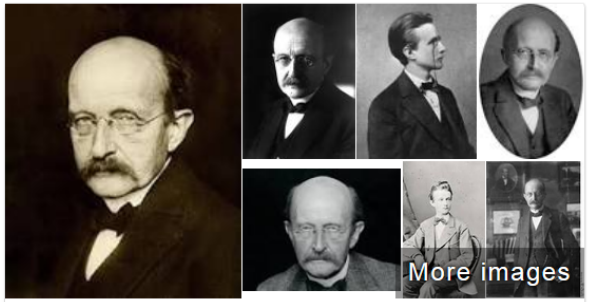
All Images News Videos Maps More Settings Tools

About 158.000.000 results (0,65 seconds)

Max Planck Institutes and Experts | Max-Planck-Gesellschaft
https://www.mpg.de/11741001/research_page
There is no such thing as "the" **Max Planck** Institute. In fact, the **Max Planck** Society operates a number of research institutions in Germany as well as abroad.

Your past visits Sort by

A Max-Planck-Institut für Informatik
900,0 m · 66123, Campus E1 4, Stuhlsatzenhausweg · 0681 93250
Closed · Opens 6AM Tue
WEBSITE DIRECTIONS



Max Planck

German physicist

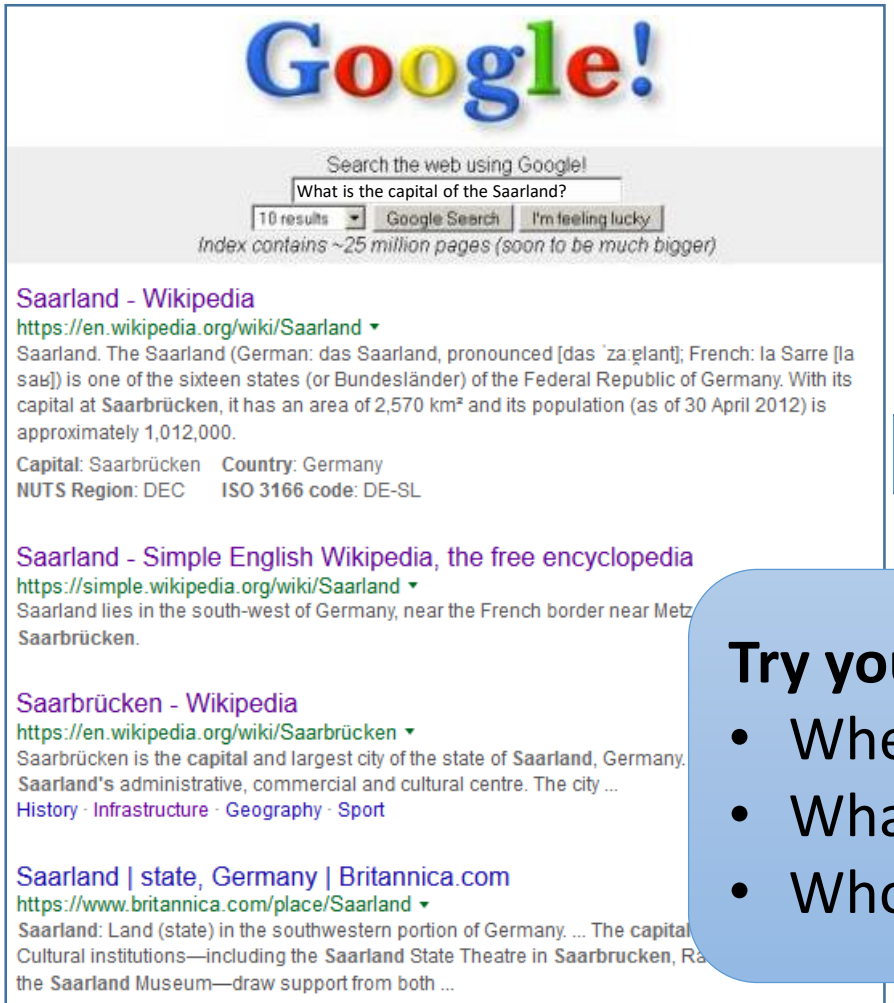
Max Karl Ernst Ludwig Planck, FRS was a German theoretical physicist whose discovery of energy quanta won him the Nobel Prize in Physics in 1918. Wikipedia

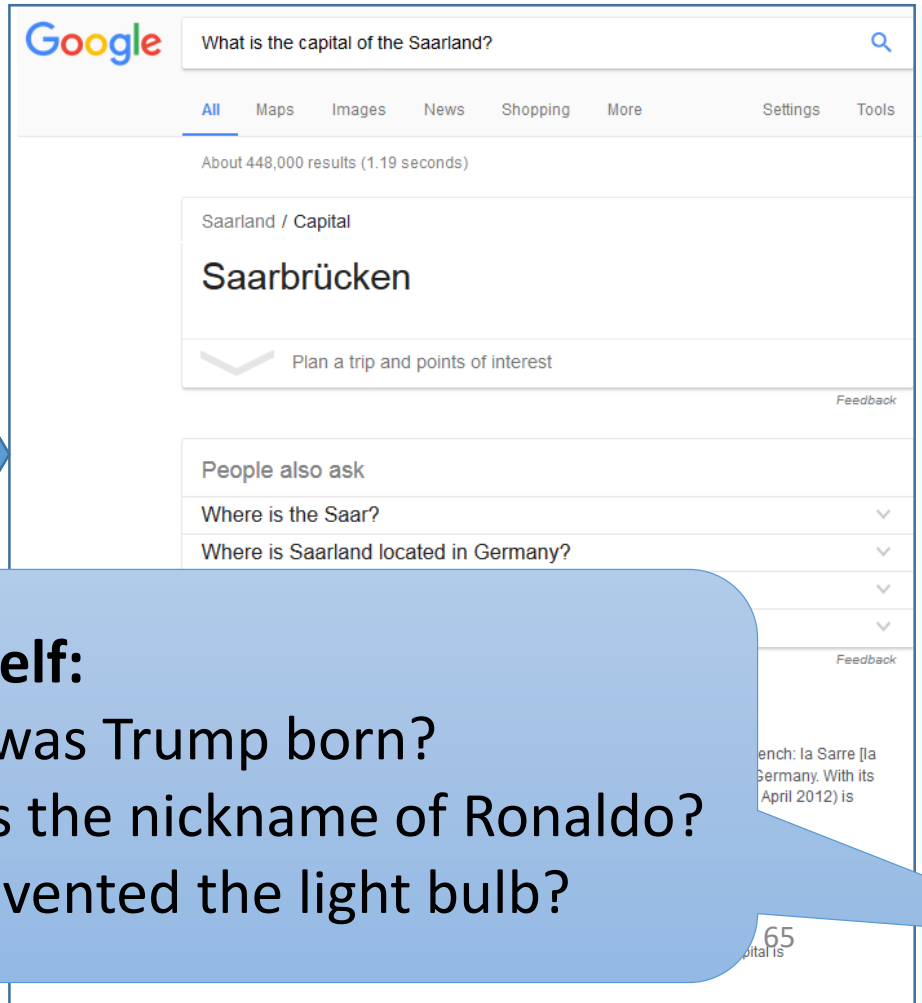
Born: April 23, 1858, Kiel

Died: October 4, 1947, Göttingen

Known for: Planck constant, Planck postulate, Planck's law, Third law of thermodynamics, Fokker–Planck

Question answering





Try yourself:

- When was Trump born?
- What is the nickname of Ronaldo?
- Who invented the light bulb?

Question answering (2)

- Knowledge bases key component in question answering systems
  - E.g., IBM Watson

- AllenAI science challenge: Computers currently in 8th grade
  - Knowledge acquisition still major bottleneck

Language generation

Douglas Adams was a British playwright, screenwriter, novelist,
children' on
March net
Adams
Brentw
marrie l,
2001).
myoca
buried

- Wikipedia in world's most spoken language: **1/10** as many articles as English Wikipedia
- World's fourth most spoken language: **1/100**

→ Wikidata intended to help resource-poor languages

https://tools.wmflabs.org/autodesc?q=9021&lang=&mode=long&links=reasonator&redlinks=reasonator&format=html

Introduction to Information Extraction

I. Motivation
II. Definition and topics
III. Formal foundations
IV. Construction and maintenance
V. Technologies
VI. Applications
VII. Past, present and future

Past

Cyc

```
(#$relationAllExists
   #$biologicalMother
   #$ChordataPhylum
   #$FemaleAnimal)
```

WIKIPEDIA
The Free Encyclopedia

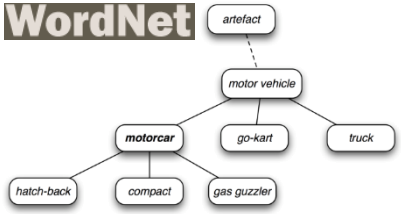
WolframAlpha

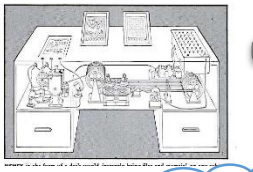
yago
select knowledge

DBpedia

ALEXANDRIA

WordNet





Memex
(1945)

Freebase
(collaborative)

Knowledge Graph

WIKIDATA

1984 2001 2007 2012 2018

Present

- IE and KBs at most major tech companies and beyond
  - Google, Microsoft, Alibaba, Bloomberg, …
- Feb 2018: $125 million investment by Microsoft cofounder Paul Allen into non-profit research on common sense knowledge extraction and reasoning
- Research: Major part of NLP conferences taken up by IE research

Future

- ?

Outline

1. Introducing each other
2. Organization of the course
3. What&why
4. **Lab 1**

Lab 1

- Information extraction where from?
  - Actual web crawling nontrivial
  - Wikipedia a popular high-quality resource
- For a change, we work on a Wiki about Game of Thrones (data dump)
- Task 1: Find pages of certain types
- Task 2: Find the different surface forms of links to a page
- Task 3: Formulate and run some SPARQL queries over Wikidata

Regular expressions

- Search patterns for String data

```
import re

str = "No pain no gain"

x = re.findall("\Sain", str)

print(x)
```

```
['pain', 'gain']
```

https://www.w3schools.com/python/python_regex.asp

Take home

- Information extraction translates unstructured/semistructured content into machine-readable structured formats
- Structured data is relevant for a range of knowledge-intensive and AI tasks
- More about how to do IE follows..