MAX-PLANCK-GESELLSCHAFT

max planck institut
informatik

Knowledge Bases

16.11.2017

Block seminar, winter semester 17/18

Simon Razniewski

Paramita Mirza

Outline

1. We&You
2. Organisation
3. Introduction to Knowledge Bases
4. Topics

- Next week: Seminar survival skills
  - How to …
    - ... research a topic
    - … read a paper
    - … write a report
    - … give a presentation

Outline

1. We&You
2. Organisation
3. Introduction to knowledge bases
4. Topics

Simon Razniewski

Campus E1 4, Room 425

srazniew@mpi-inf.mpg.de

- Researcher at MPII, D5 since November 2017
- Assistant professor at FU Bozen-Bolzano, Italy, 2014-2017
- PhD same place, 2014
- Diplom at TU Dresden, 2010

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

Paramita Mirza

Campus E1 4, Room 420

paramita@mpi-inf.mpg.de

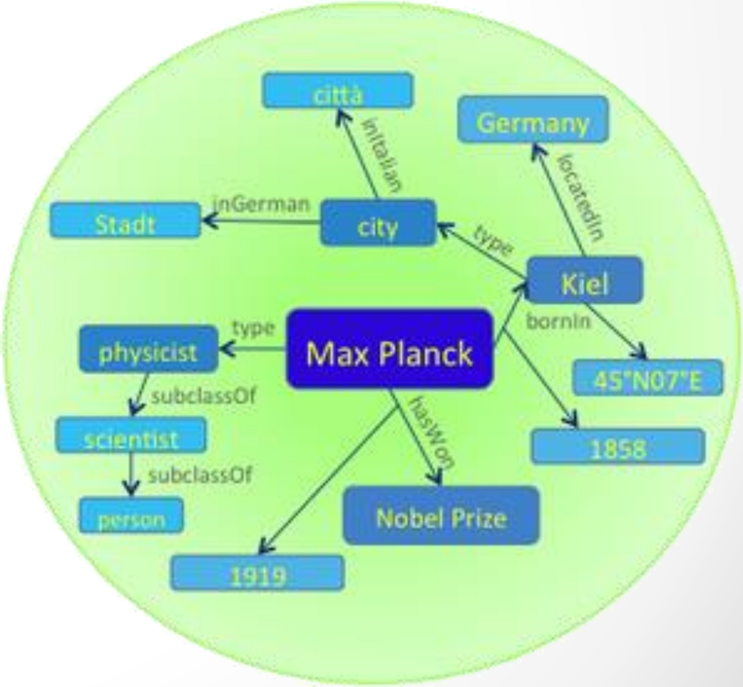
- Researcher at MPII, D5 since July 2016
- PhD study at FBK-ICT (Information and Communication Technology) Doctoral School and University of Trento, Italy
  - Dissertation: *"Extracting Temporal and Causal Relations between Events"*
  - Newsreader project (newsreader-project.eu) → structured knowledge of events from financial news for decision making support
- Research interests:
  - Semantics of numbers in texts
    - Existential information extraction (e.g., "Trump has 5 children")
  - Causation and common sense knowledge harvesting
    - Emotional aspect of relation extraction

max planck institut
informatik

D5: Databases and Information Systems

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

- YAGO select knowledge Knowledge Base
  - contains more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities



And you are?

- Name
- Course of study
- Why this seminar?
- …

Outline

1. We&you
2. Organisation
3. Introduction to knowledge bases
4. Topics

Block Seminar

- Credit points: 7, hours: 210 (!)
- Registration
  - Register in HISPOS (till 7.12.)
- What to do?
  - Attend the lectures and block seminar days
  - Hand in a seminar paper
  - Give a presentation

Seminar paper should be broader and deeper!

- Block seminar days
  - 2 days seminar, 8 presentations per day (~25 min talk, ~10 min discussion each)
  - Dates: Please fill the Doodle (email)

Block Seminar (2)

- Credit points: 7, hours: 210 (!)
- Grading based on
  - Seminar paper
  - Presentation
  - Knowledge on the subject → from the discussion
  - Ability to stick to deadlines!

Schedule

- 16. 11. 17
  - Kick-off meeting
- 23. 11. 17
  - "How to" lecture
  - Topic assignment
- 13. 12. 17
  - Students send extended outline of their seminar paper
- Middle of December/early January
  - Individual meetings to discuss outline and/or draft
- 31. 01. 18
  - Students submit their final seminar paper
- Two weeks before the first block seminar
  - Students send their preliminary slides
- Two days before the first block seminar
  - Students send their final slides

Extended outline

...
2. Methodology
   2.1 KB Representation
   - Discussion on how to represent the knowledge
   - Comparison with other representations
   - Benefits and limitations of such representations
   - ...

Learning Outcome

- Understanding the area of Knowledge Bases
  - Various existing KBs, differences, applications, benefits, limitations, …
- Learn to
  - …research a topic
  - …read papers
  - …structure and write up a summary about a scientific field
  - …give a high-level scientific presentation

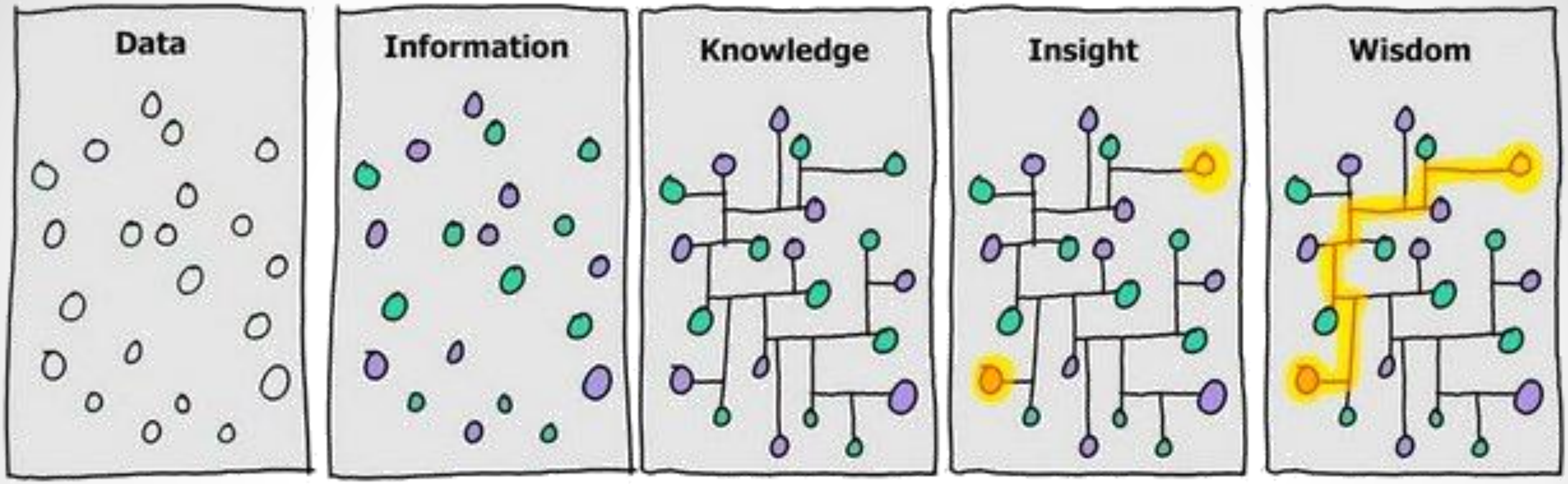
→ Almost a thesis!

  - Only missing the own original technical contribution

Outline

1. We&you
2. Organisation
3. Introduction to knowledge bases
4. Topics

Data(base) vs Knowledge(base)



- Database: flat data representation (tabular)
- Knowledge base: *"The ideal representation for a knowledge base is an object model (ontology) with classes, subclasses, and instances."*

--Wikipedia

Max Planck

From Wikipedia, the free encyclopedia

Max Karl Ernst Ludwig Planck, FRS[2] (/plɑːŋk/;[3] 23 April 1858 – 4 October 1947) was a German theoretical physicist whose discovery of energy quanta won him the Nobel Prize in Physics in 1918.[4]

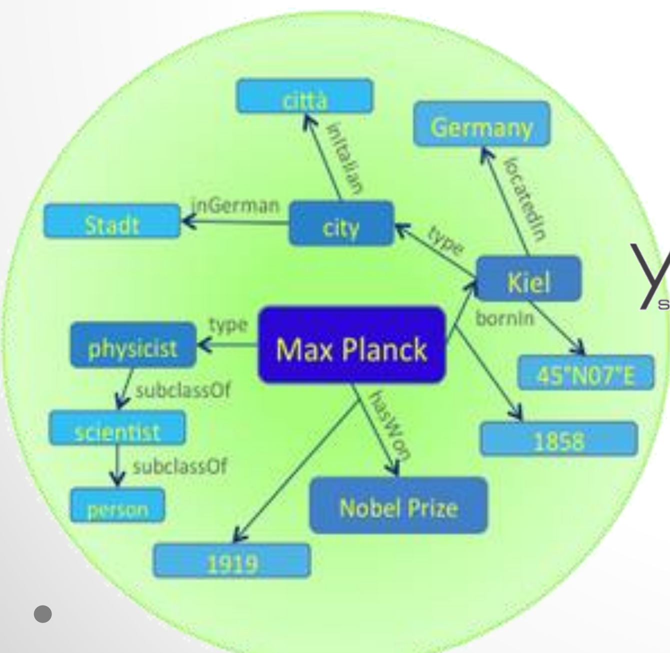
Knowledge Engineering

WIKIDATA

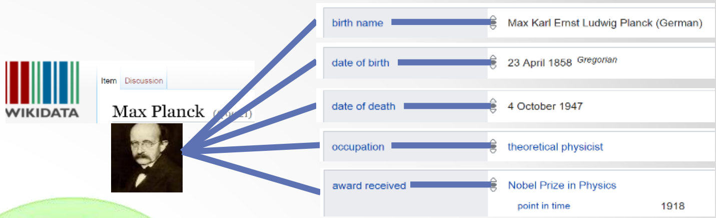
Item Discussion

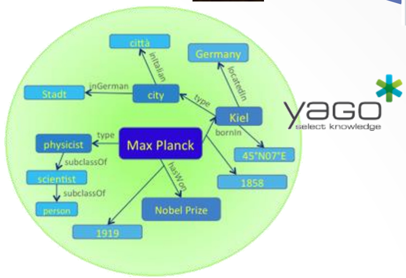
Max Planck (Q9021)

birth name	Max Karl Ernst Ludwig Planck (German)
date of birth	23 April 1858 *Gregorian*
date of death	4 October 1947
occupation	theoretical physicist
award received	Nobel Prize in Physics point in time 1918



yago
select knowledge





Knowledge Bases:

Pragmatic Definition

A knowledge base (KB) is a comprehensive semantically organized machine-readable collection of universally relevant or domain-specific entities, classes, and facts (attributes, relations)

- Plus spatial and temporal dimensions
- Plus commonsense properties and rules
- Plus contexts of entities and facts, e.g., textual/visual evidences
- Plus…

History of Digital Knowledge Bases

Cyc

- Objective: to codify, in machine-usable form, millions of pieces of knowledge that compose human common sense

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

"For every chordate, there exists a female animal which is its mother"

- "In 1986, Doug Lenat estimated the effort to complete Cyc would be 250,000 rules and 350 man-years of effort."

1984 2001 2007 2012

History of Digital Knowledge Bases

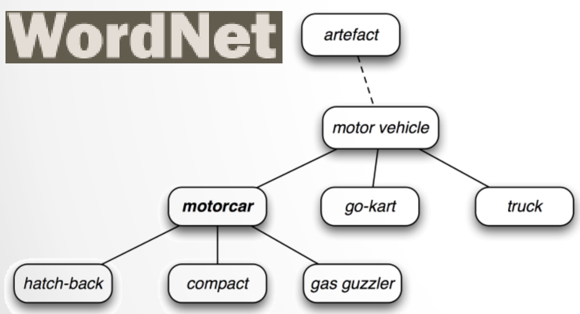
From humans for humans

Cyc

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

WIKIPEDIA
The Free Encyclopedia

WordNet

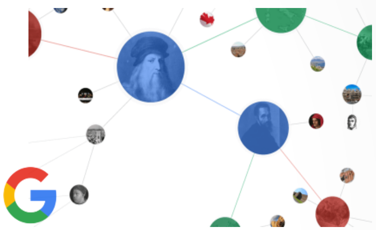


From algorithms for machines

WolframAlpha

yago
select knowledge

DBpedia



Knowledge Graph

Freebase
(collaborative)

WIKIDATA

ARISTO

OpenIE

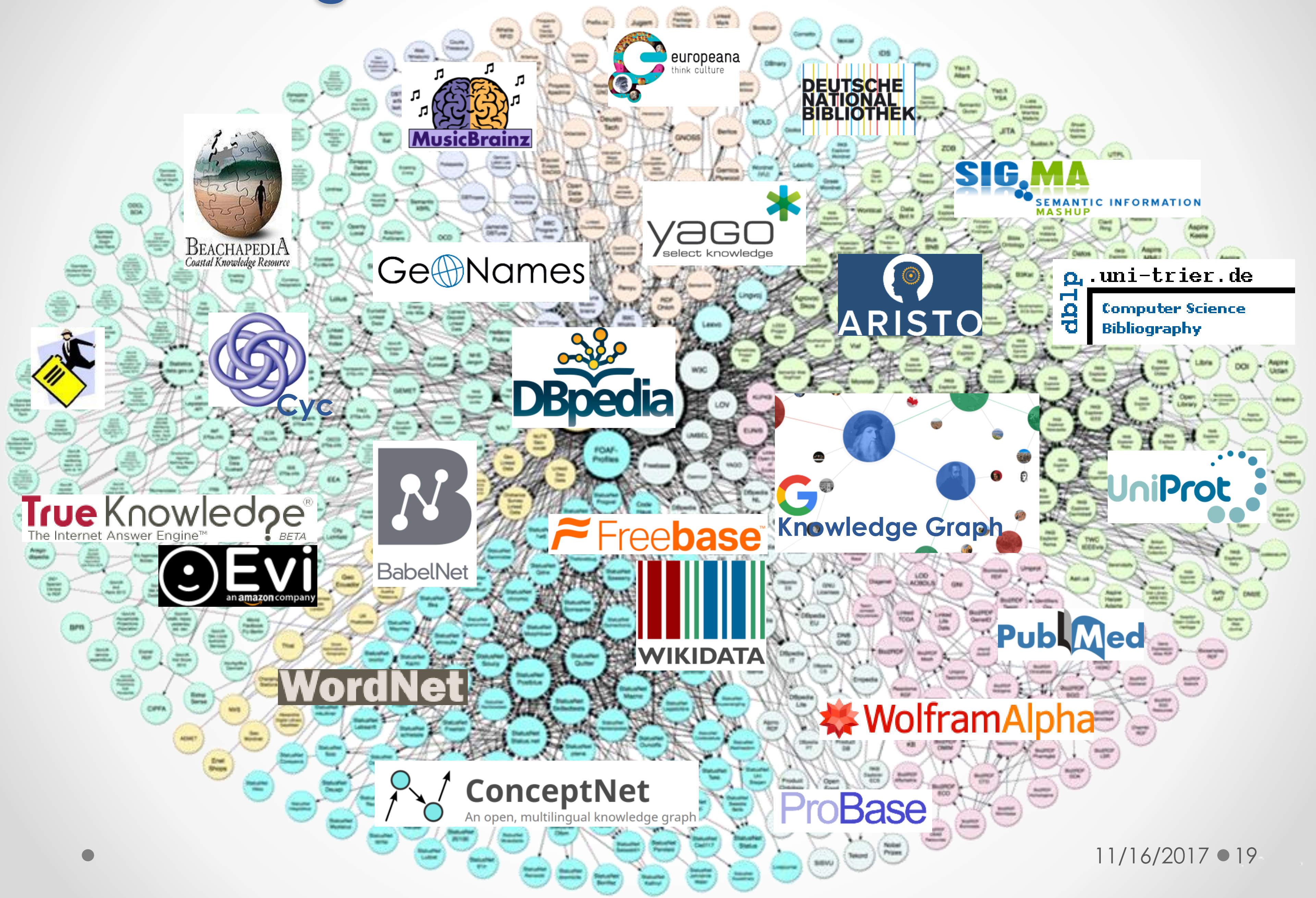
TextRunner

"Albert Einstein was born in Ulm and died in Princeton"

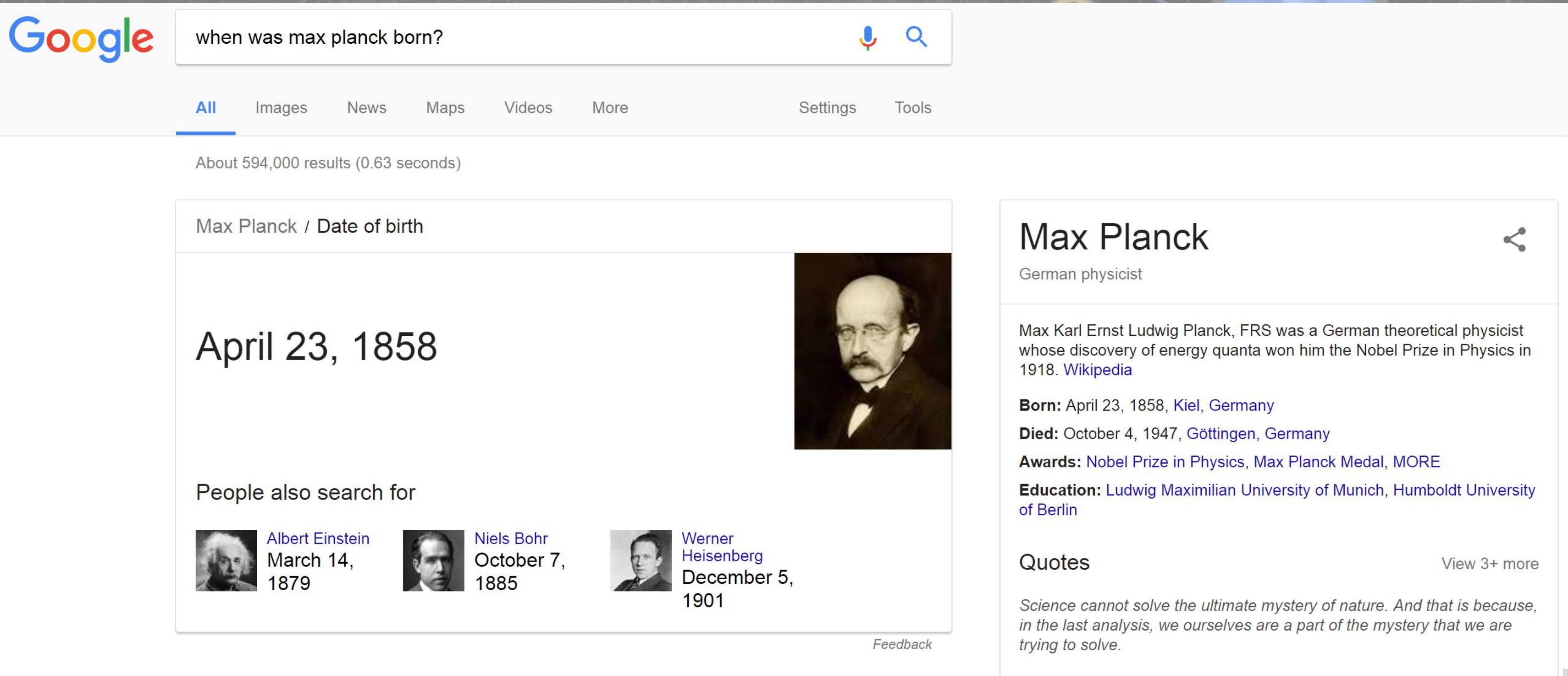
- (Albert Einstein, was born in, Ulm)
- (Albert Einstein, died in, Princeton)

1984 2001 2007 2012 2016

Knowledge Bases & Semantic Web



Use case: Internet Search



Use case: Question Answering

This town is known as "Sin City"
and its downtown as "Glitter Gulch"

Question classification
and decomposition

+ **KBs**

WIKIPEDIA
The Free Encyclopedia

WIKIDATA

DBpedia

yago
select knowledge

Las Vegas

Computer Wins on 'Jeopardy!': Trivial, It's Not

Carol Kaelson/Jeopardy Productions Inc., via Associated Press

Two "Jeopardy!" champions, Ken Jennings, left, and Brad Rutter, competed against a computer named Watson, which proved adept at buzzing in quickly.

By JOHN MARKOFF
Published: February 16, 2011

Use case: Domain-targeted Question Answering

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

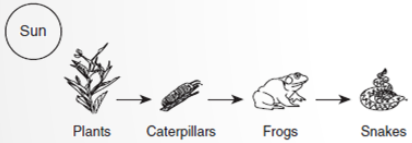
ARISTO

Answering Science Questions using Machine Reading

What is Aristo?

YOU ASKED ARISTO:

Base your answers on the diagram of a food chain below and on your knowledge of science.



(Not drawn to scale)

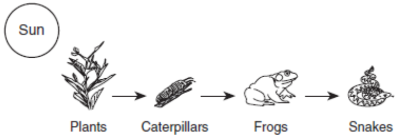
If the population of snakes increases, the population of frogs will most likely

(A) decrease
(B) increase
(C) remain the same

ARISTO ANSWERED:

Question: Base your answers on the diagram of a food chain below and on your knowledge of science.

Hide



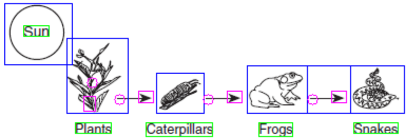
(Not drawn to scale)

If the population of snakes increases, the population of frogs will most likely

Answer: (A) decrease

Because:

Diagram elements detected:



Extracted food chain: **dependsOn**(Caterpillars, Plants) & **dependsOn**(Snakes, Frogs) & **dependsOn**(Frogs, Caterpillars)

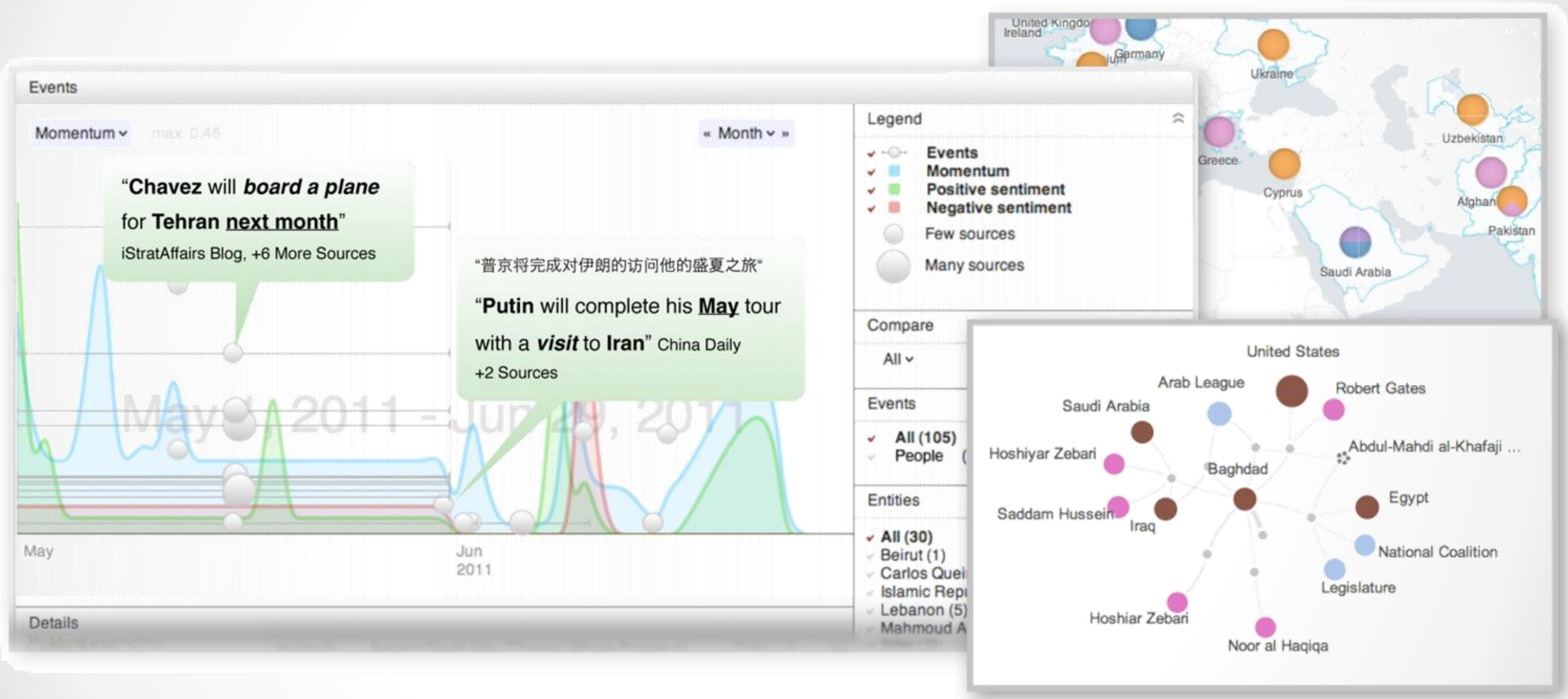
Given: **qChange**(Snakes, Increase)

Query: **qChange**(Frogs, X)

Answer: X = Decrease

Use case: Analysis of History

https://www.recordedfuture.com/



Outline

1. We&you
2. Organisation
3. Introduction to knowledge bases
4. Topics

Seminar Topics

Area 1: Knowledge Bases

Area 2: Learning over Knowledge Bases

Area 3: Using Knowledge Bases

Seminar Topics

Area 1:
Knowledge Bases

1. Collaborative KBs: Wikidata -- Simon

- Possible aspects
  - History
  - Data model
  - State of project and community
  - *How does the collaborative ontology engineering process work*
  - *What is the adoption across Wikipedia*
  - *Debates around the project*
- References
  - *Wikidata: a free collaborative knowledge base* (Vrandečić & Krötzsch, CACM 2014)
  - *Introducing Wikidata to the linked data web* (Erxleben et al., ISWC 2014)

2. Structured information extraction: DBpedia and YAGO -- Simon

- Possible aspects
  - Structured knowledge harvesting
  - Implementation of both
  - Quality and limitations
  - *Discussion on their class hierarchy*
  - *Extensions on both KBs (e.g., multilinguality, temporal dimension)*
- References
  - *DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia* (Lehmann, et al., Semantic Web 2015)
  - *YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia* (Hoffart et al., Artificial Intelligence 2013)

3. General common-sense KBs: ConceptNet and WebChild -- Paramita

- Possible aspects
  - Definition of common-sense knowledge
  - Knowledge representation
  - Comparison on knowledge acquisition
  - Discussion on possible benefits and use-cases
  - *Extensions on both KBs (e.g., multilinguality)*
- References
  - *Representing general relational knowledge in ConceptNet 5* (Speer & Harvasi, LREC 2012)
  - *WebChild: Harvesting and Organizing Commonsense Knowledge from the Web* (Tandon et al., WSDM 2014)

4. Domain-specific activity KBs -- Paramita

- Possible aspects
  - Definition of ‘human activity’
  - Knowledge acquisition
  - Discussion on possible benefits and use-cases
  - *Task knowledge vs script knowledge*
- References
  - *HowTo KB -- Distilling task knowledge from How-To communities* (Chu et al., WWW 2017)
  - *A hierarchical bayesian model for unsupervised induction of script knowledge* (Frermann et al., EACL 2014)

5. The Allen AI science challenge & building a science KB: Aristo -- Paramita

- Possible aspects
  - The Allen AI science challenge
  - What kind of knowledge is needed to pass this challenge
  - How is the science KB built
  - Limitations and future directions
  - *The Allen AI science challenge in comparison with other AI benchmarks (e.g., Turing test)*
- References
  - *Domain-targeted, high precision knowledge extraction* (Mishra et al., TACL 2017)
  - *Moving beyond the Turing test with the Allen AI Science Challenge* (Schoenick et al., CACM 2016)

6. A WordNet-based image KB: ImageNet -- Paramita

- Possible aspects
  - WordNet
  - ImageNet representation
  - Applications
  - *Limitations and possible extensions*
- References
  - *Imagenet: A large-scale hierarchical image database* (Deng et al., CVPR 2009)
  - *WordNet: A lexical database for English* (Miller, CACM 1995)

Seminar Topics

Area 2:
Learning over Knowledge Bases

7. KB association rule mining -- Simon

- Possible aspects
  - Challenges for association rule mining over KBs
  - AMIE-Implementation
  - *How to deal with time*
  - *Critical comparison with other learning techniques*
- References
  - *Fast rule mining in ontological knowledge bases with AMIE+* (Galárraga et al., VLDB 2015)
  - *Discovering Graph Temporal Association Rules* (Nakami et al., CIKM 2017)

8. Inferring new facts with vector space embeddings -- Simon

- Possible aspects
  - What is embeddings
  - Benefits of embeddings over traditional representations
  - Methodology in the TransE paper
  - *Look at the whole population of Trans* papers, how did they progress, what are main evolution steps*
  - *Critical comparison with other learning techniques*
- References
  - *Translating embeddings for modeling multi-relational data* (TransE) (Bordes et al., NIPS 2013)
  - *Learning entity and relation embeddings for knowledge graph completion* (TransR) (Lin et al., AAAI 2015)

Seminar Topics

Area 3:
Using Knowledge Bases

9. KB question answering -- Simon

- Possible aspects
  - Question answering task
  - Approach
  - Evaluation
  - *Which classes of queries are easy, which ones are tough*
  - *How to deal with queries that involve subjectivity*
- References
  - *Automated template generation for question answering over knowledge graphs* (Abujabal et al., WWW 2017)
  - *Robust question answering over the web of linked data* (Yahya et al., CIKM 2013)

10. Hybrid question answering using KBs and text -- Paramita

- Possible aspects
  - QA on unstructured vs QA on structured data
  - Advantages over QA with only KB or text
  - Approach
  - Evaluation
- References
  - *Question answering on Freebase via relation extraction and textual evidence* (Xu et al., ACL 2016)
  - *Open question answering over curated and extracted knowledge bases* (Fader et al., KDD 2014)

11. Non-encyclopedic QA in the science domain -- Simon

- Possible aspects
  - What is the difference to classic QA
  - Approach
  - Limitations
  - *Look at other papers taking part in the AllenAI science challenges, what are other approaches, ideas, limitations*
- References
  - *Answering complex questions using open information extraction* (Khot et al., ACL 2017)

12. Coreference resolution -- Paramita

- Possible aspects
  - Coreference resolution task
    - *"Martha Stewart is hoping people don't run out on her. The celebrity indicted on charges stemming from . . ."*
  - Approach
  - Evaluation
  - Limitations and future directions
  - *State-of-the-art methods in coreference resolution*
- References
  - *Coreference resolution with world knowledge* (Rahman & Ng, ACL 2011)

13. Referent prediction -- Paramita

- Possible aspects
  - Referent prediction task
    - *"I decided to take a bath yesterday. Once I got home, I walked to my bathroom and scrubbed the tub clean. Then I plugged […] and began filling it with water."* → *the bathroom? the tub?*
  - How is the script knowledge exploited
  - Evaluation
  - Limitations and future directions
- References
  - *Modeling semantic expectation: Using script knowledge for referent prediction* (Modi et al., TACL 2017)

14. Biography generation -- Paramita

- Possible aspects
  - Natural language text generation
  - Approach
  - Evaluation
  - Limitations and future directions
  - *State-of-the-art approach on text generation*
- References
  - *Learning to generate one-sentence biographies from Wikidata* (Chisholm et al., EACL 2017)

15. Fact ranking -- Simon

- Possible aspects
  - Problem
  - Approach
  - Evaluation
  - Related works
  - Critical evaluation of the notion of "interestingness"
- References
  - *The unusual suspects: Deep learning based mining of interesting entity trivia from knowledge graphs* (Fatma et al., AAAI 2017)

Outline

1. We&You
2. Organisation
3. Introduction to Knowledge Bases
4. Topics

- Next week: Seminar survival skills
  - How to …
    - … research a topic
    - … read a paper
    - … write a report
    - … give a presentation
  - Topic assignment