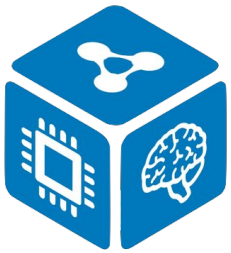
Project Group
Data Science Suite IV

Group: Data Science
Tutors: Stefan Heindorf and Michael Röder



DICE – Data Science Group, University Paderborn, Germany

April 15, 2020

What will we do today?

1. Clarify organizational things
2. Motivate the Data Science Suite
3. Present the different topics
4. Tasks until next week

Section 1

PG Organization

PG Organization

Setup

- The PG will be split up into sub groups. Every sub group
  - will work
    - selforganized,
    - as a team
    - on a single topic.

PG Organization

Setup

- The PG will be split up into sub groups. Every sub group
  - will work
    - selforganized,
    - as a team
    - on a single topic.
  - has a supervisor
    - as product owner and
    - as a contact point if something goes wrong
      1) During his office hours, 2) after you made an appointment. 3) Don't wait until the end of the project!

PG Organization

Setup

- The PG will be split up into sub groups. Every sub group
  - will work
    - selforganized,
    - as a team
    - on a single topic.
  - has a supervisor
    - as product owner and
    - as a contact point if something goes wrong
      1) During his office hours, 2) after you made an appointment. 3) Don't wait until the end of the project!
- The PG goes for 2 semesters (SoSe 20, WiSe 20/21)
- Every student should invest
  - 20h per week with vacations in-between
  - (15h per week without vacations)

→ You should learn

- manage a project
  - Transform goals into tasks, sub tasks, etc.
  - Understand dependencies between tasks
  - Create a project plan / sprint plan / milestones
  - Adapt your plan after you realized that it did not work
  - ...
- work in a team
  - How to communicate (!)
  - How to solve conflicts
  - ...

What do we expect from you?

Be a valuable member of your team!

- Manage your project
- Write code(!) and commit it
- Communicate with your team members
- Support each other where possible
- ...

We won't keep students if they are not participating in their team.

PG Organization

Timetable

- 1st meeting (today)
  Introduction and presentation of the topics
- 2nd meeting (next week)
  Setup of the single sub groups

PG Organization

Timetable

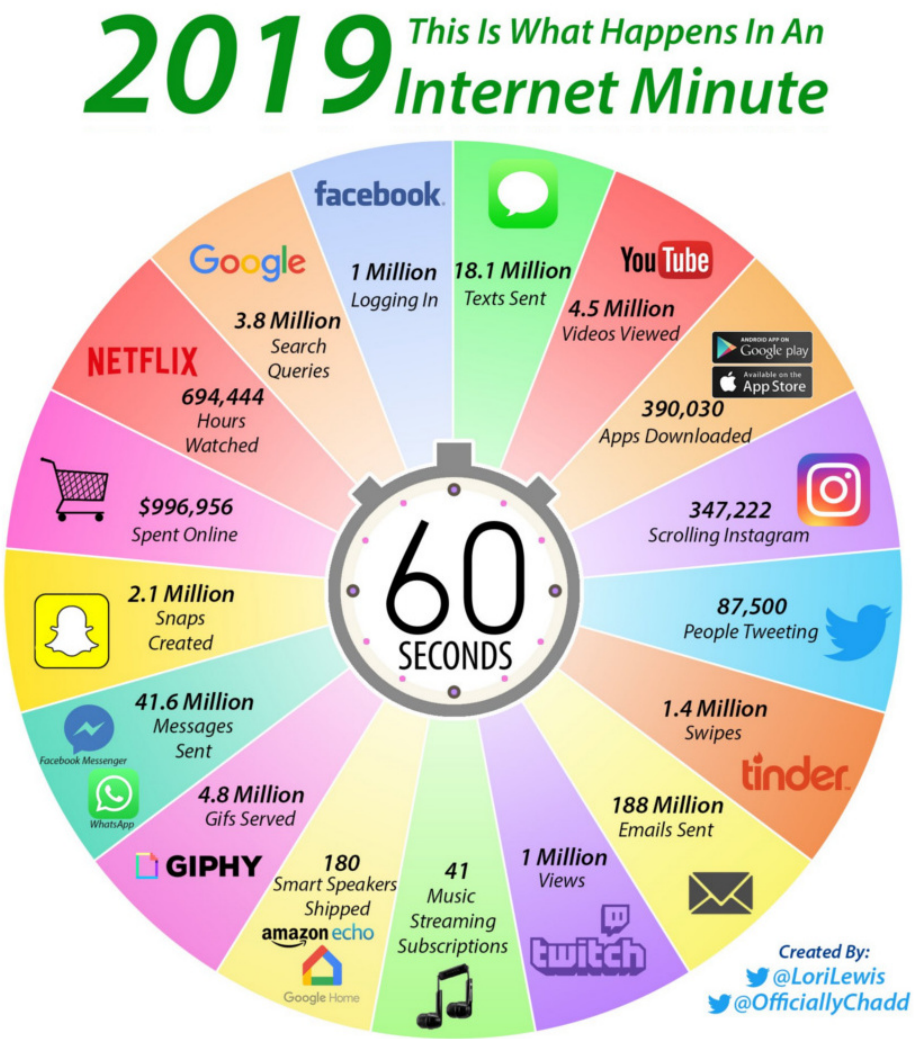
- 1st meeting (today)
  Introduction and presentation of the topics
- 2nd meeting (next week)
  Setup of the single sub groups
- until May 20th, 2020
  Registration as Studienleistung and exam—or leave the PG!
  `https://www.uni-paderborn.de/studium/paul-info/fristen-und-termine/pruefungsanmeldung/`
- End SoSe 20
  Status presentation and intermediate feedback round
- $\sim$ December 2020
  End of the implementation phase
- End WiSe 20/21
  Final presentation and end of the project

Section 2

Motivation

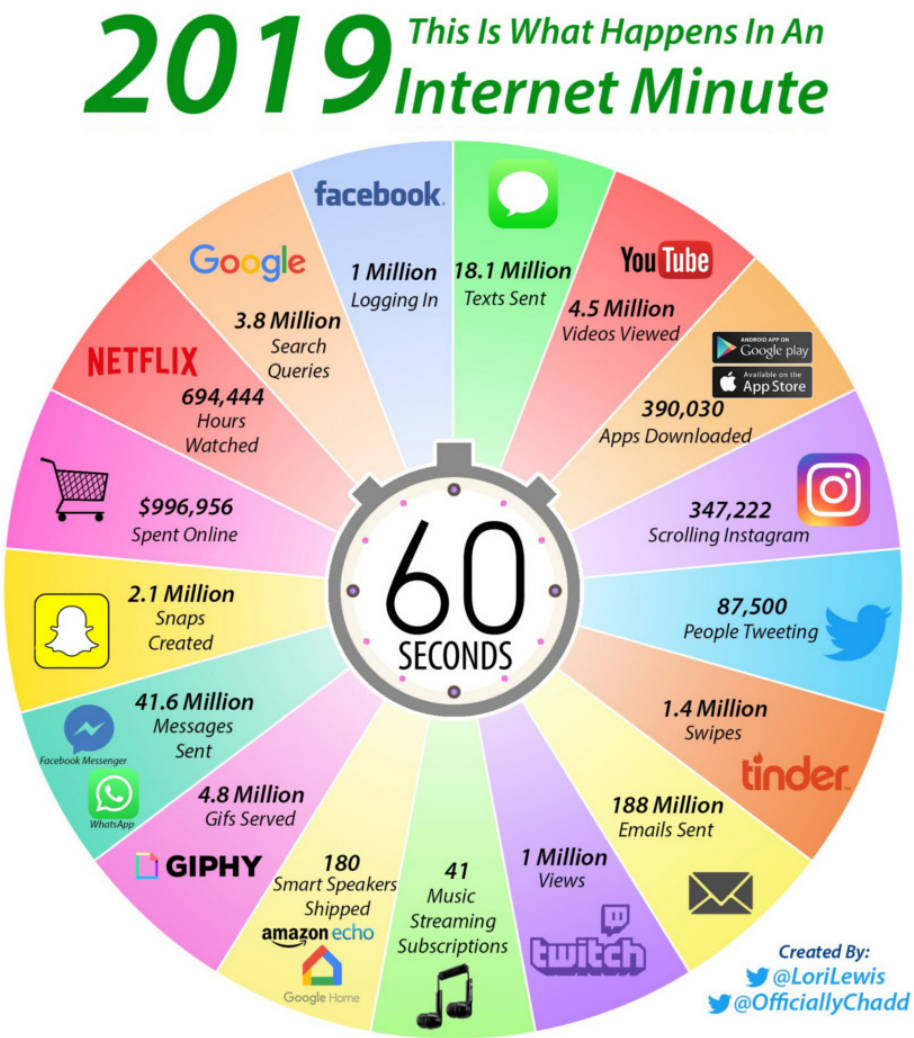
Motivation

One Minute on the Web



Motivation

One Minute on the Web

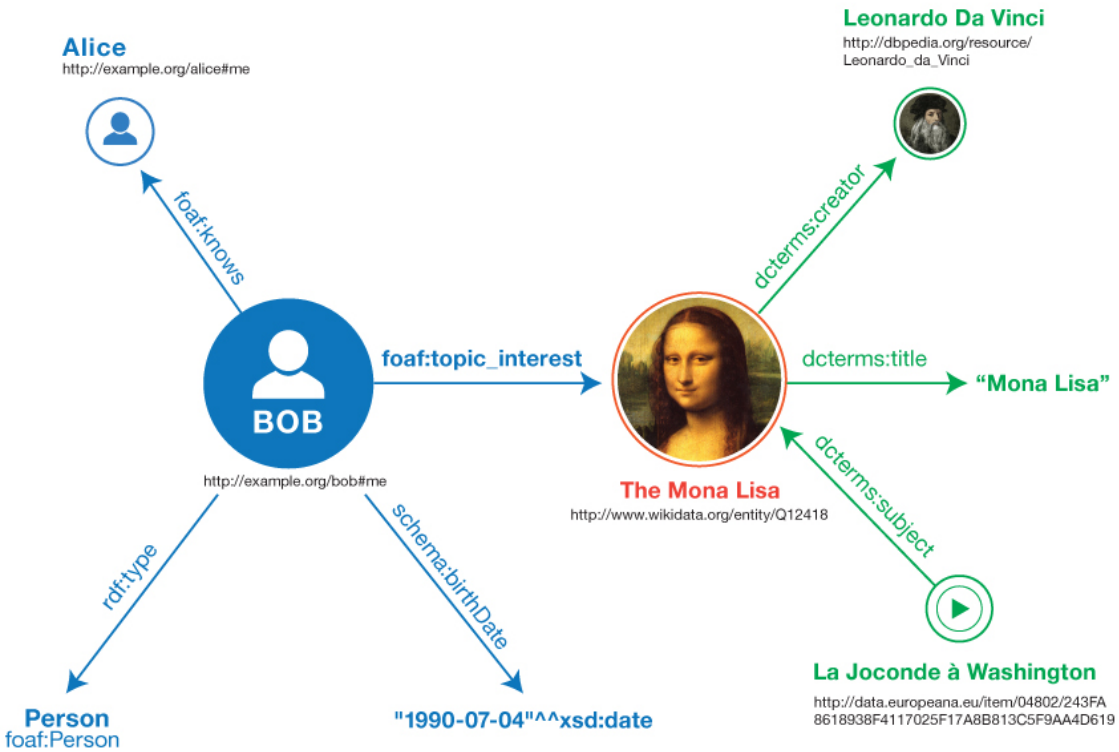


vs

http://cliparting.com/

Motivation

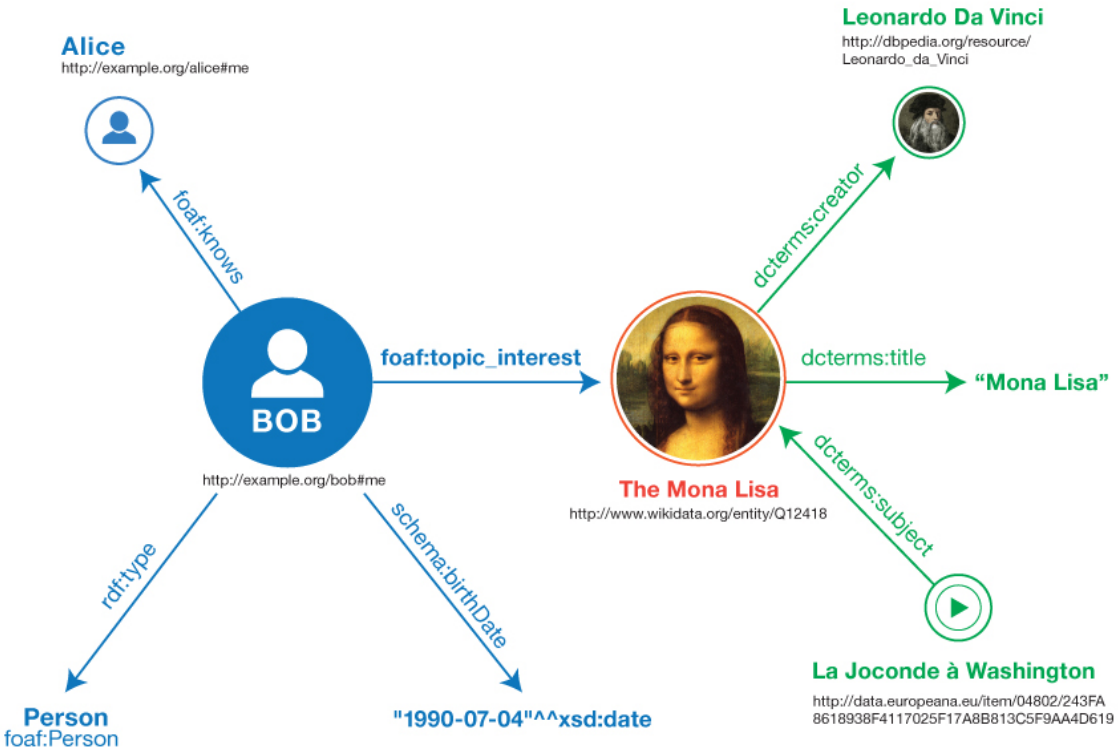
Knowledge Graphs

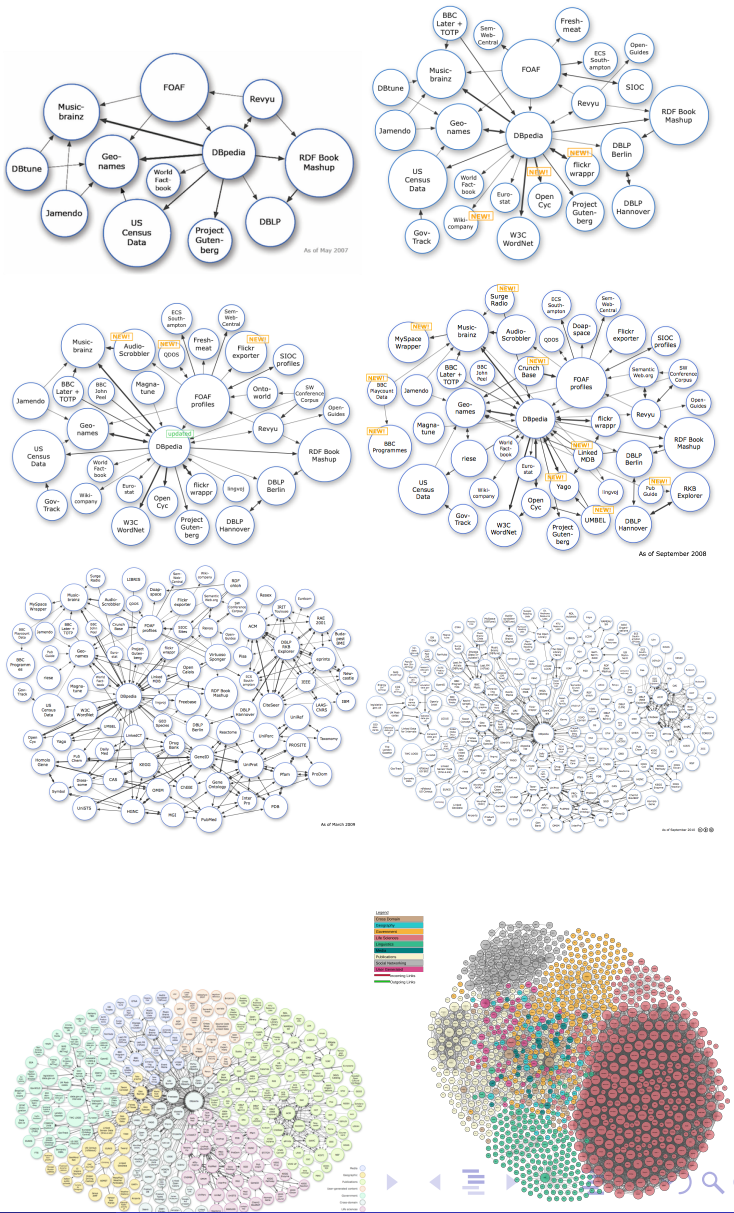


https://www.w3.org/TR/rdf11-primer/

Motivation

Knowledge Graphs



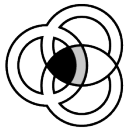


https://www.w3.org/TR/rdf11-primer/

Motivation

Systems

amazon alexa



DIESEL

QA KiS

Question Answering wiKiframework-based System

Google ASSISTANT

Qanary

QAMEL

Question Answering on Mobile Devices

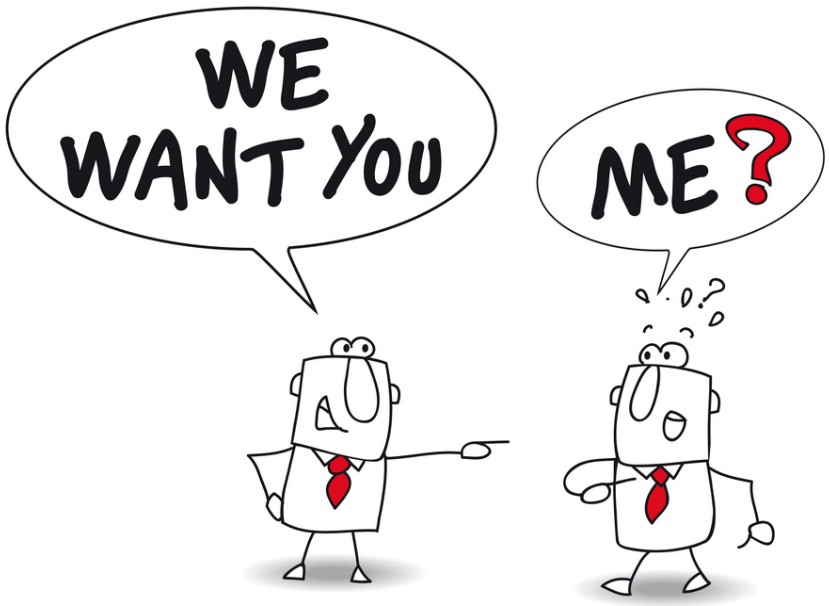
FOX

Tréo

WDAqua

We want you

- Create new software: Develop new software and research prototypes.
- Enhance code: Improve existing solutions.
- Participate: Bring your own ideas in.



We offer

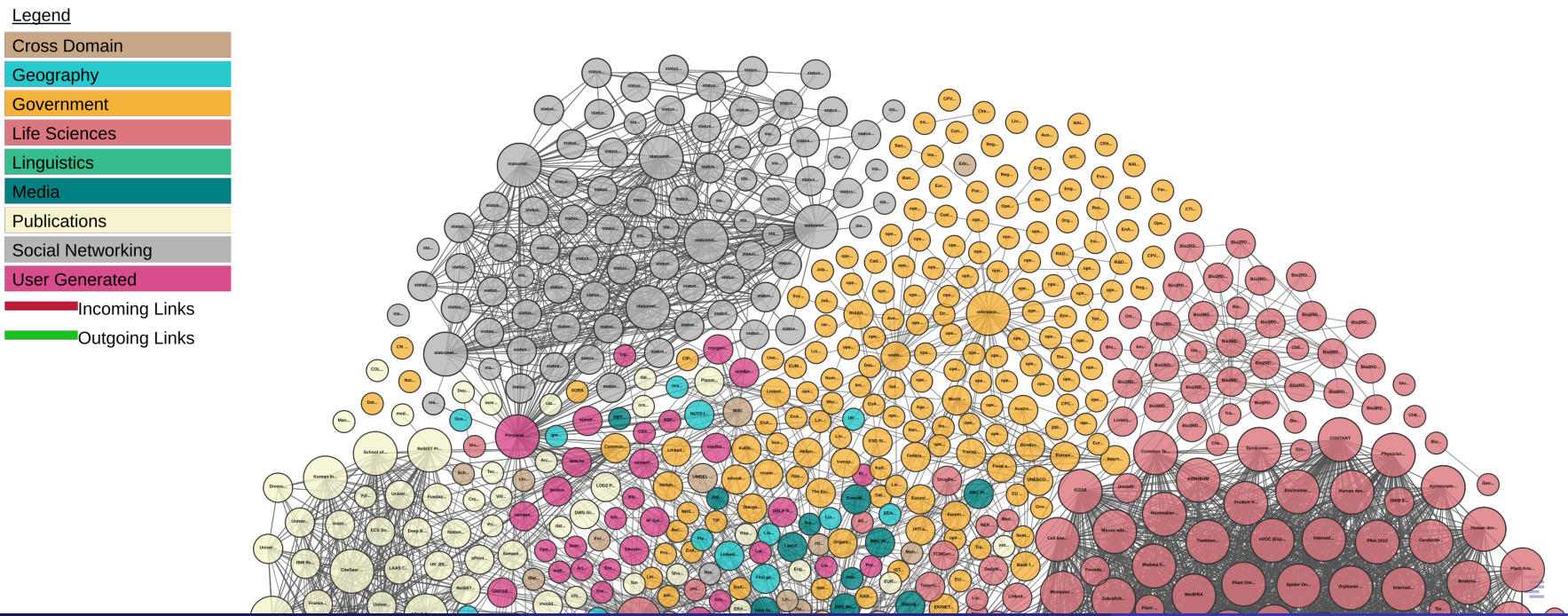
- Running software: Open-source, industry-grade solutions
- Real data: Billions of facts from Wikipedia, bio-medicine, etc.
- Expert tutors, who developed the core software
- Follow-up: Topics can be extended to MSc thesis
- Publications at top conferences (ISWC, ESWC, WWW)

Section 3

Topics

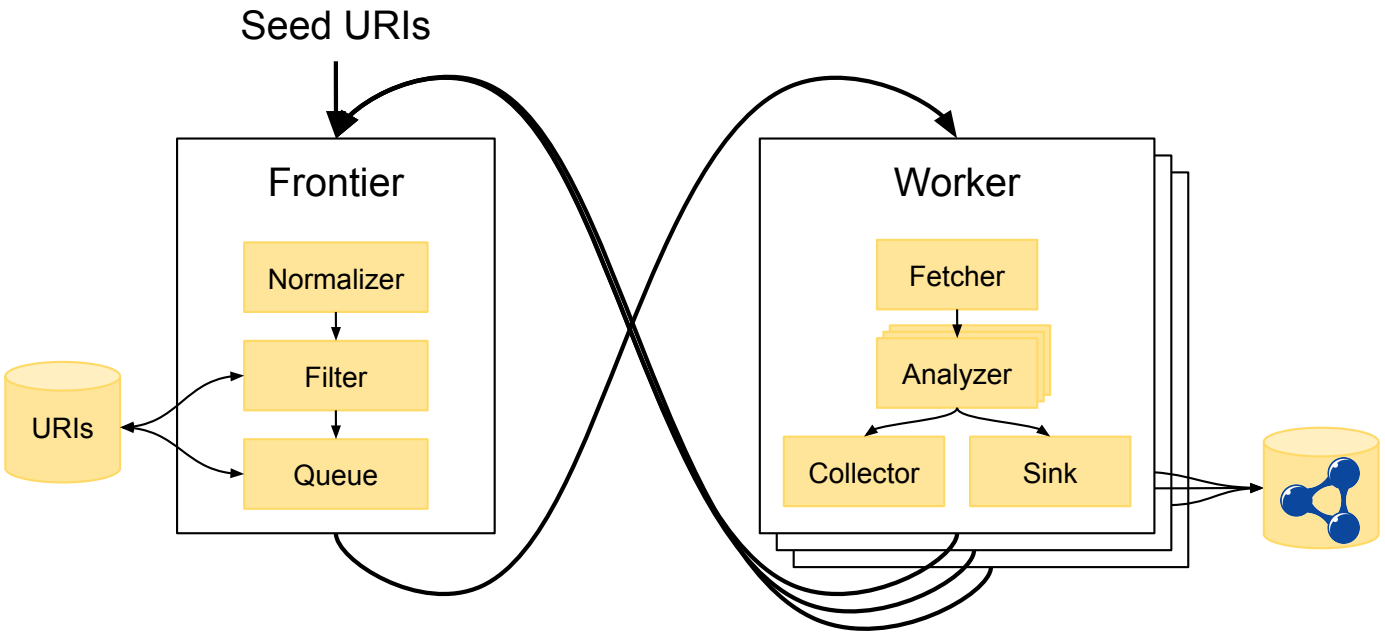
Summary

- Problem: We need to improve crawlers for gathering data from the Data Web.
- Solution: Create a benchmark for such crawlers.
- Goal: Improve our Data Web crawler benchamrk.



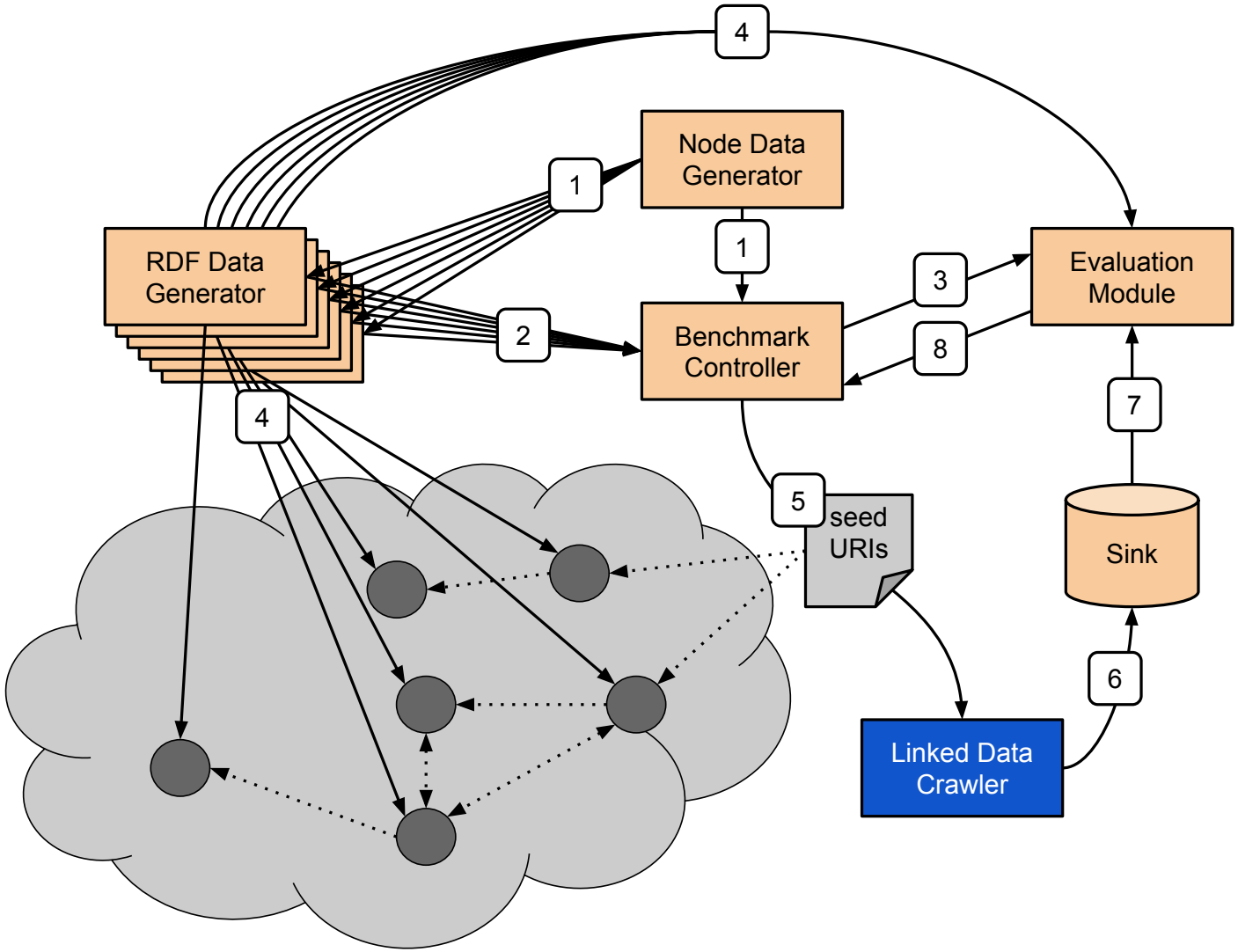
Summary

- Problem: We need to improve crawlers for gathering data from the Data Web.
- Solution: Create a benchmark for such crawlers.
- Goal: Improve our Data Web crawler benchamrk.



Topics

Benchmarking – ORCA



Further information:

- Paper on arxiv.org
  `https://arxiv.org/abs/1912.08026`

Github projects:

- `https://github.com/dice-group/orca`
- `https://github.com/dice-group/Squirrel`

Technologies:

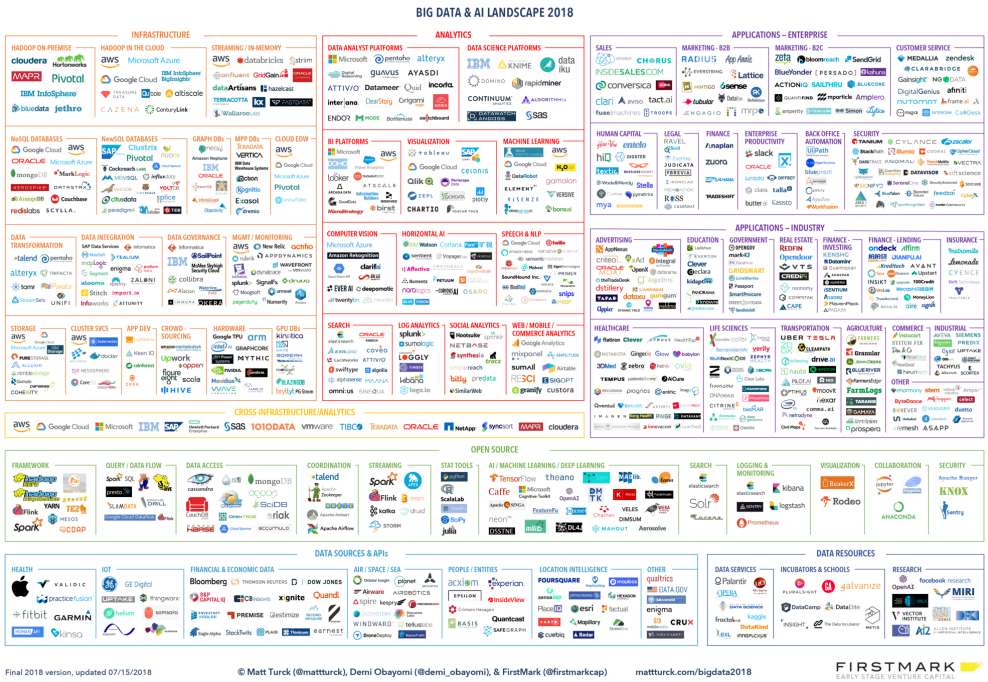
- Graph theory
- RDF, RDFa, Microformat, microdata, JSON-LD
- Java / Maven
- Docker

Summary

- Problem: Huge amount of available tools.
- Solution: Holistic benchmarking platform.
- Goal: Move the platform from Docker Swarm to Kubernetes.

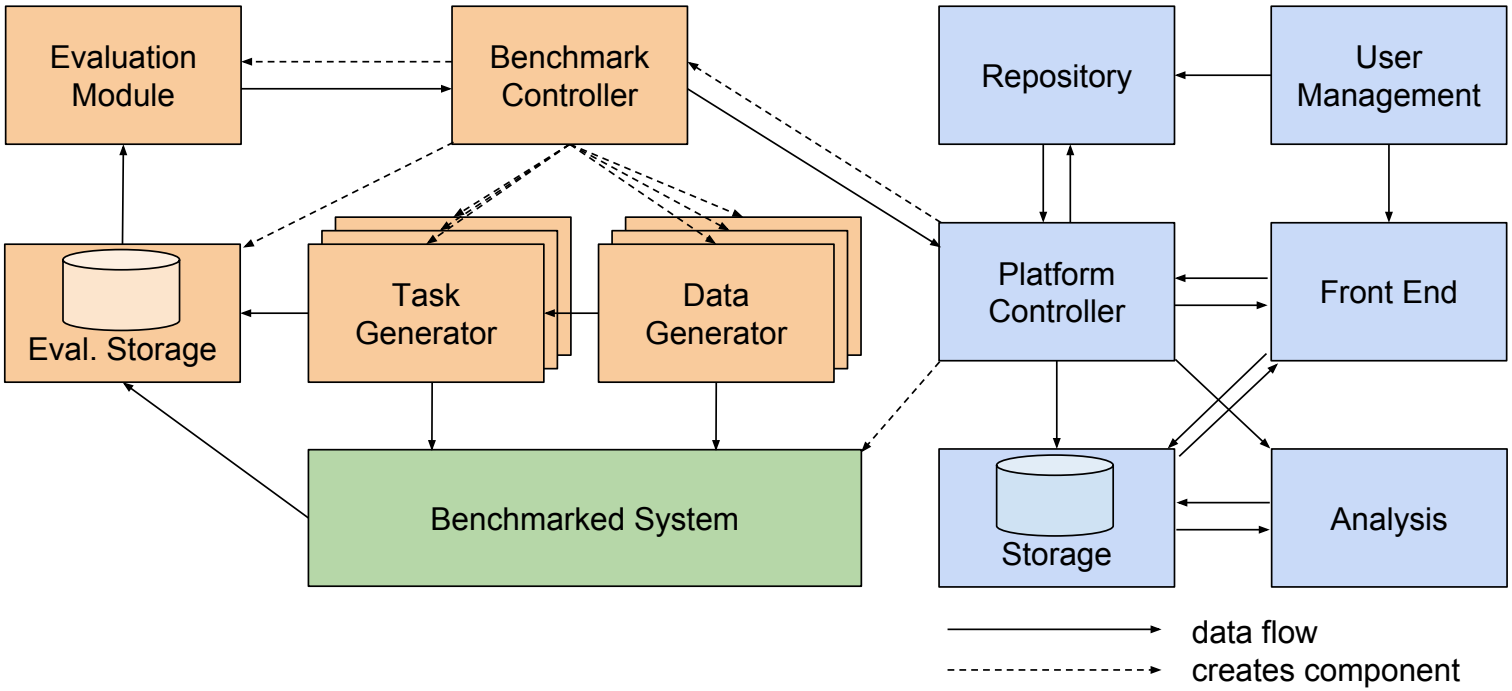
Summary

- Problem: Huge amount of available tools.
- Solution: Holistic benchmarking platform.
- Goal: Move the platform from Docker Swarm to Kubernetes.



Summary

- Problem: Huge amount of available tools.
- Solution: Holistic benchmarking platform.
- Goal: Move the platform from Docker Swarm to Kubernetes.



Further information:

- M. Röder, D. Kuchelev, and A.-C. Ngonga Ngomo: "HOBBIT: A platform for benchmarking Big Linked Data". Data Science, 2019. `https://content.iospress.com/articles/data-science/ds190021`
- Platform documentation `https://hobbit-project.github.io`

Github projects:

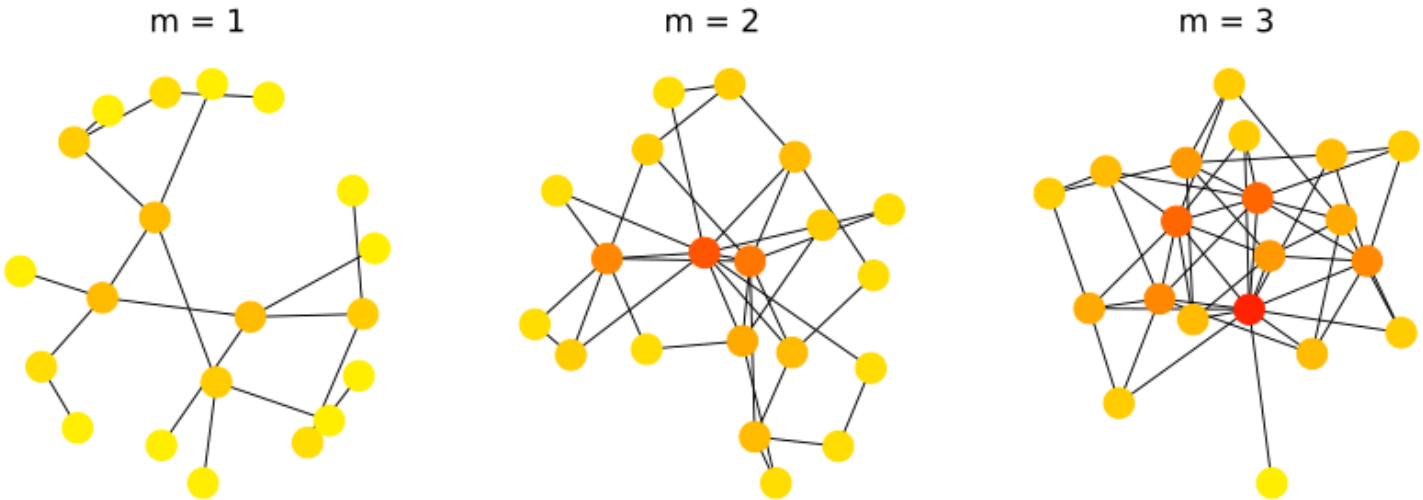
- `https://github.com/hobbit-project/platform`
- `https://github.com/hobbit-project/core`

Technologies:

- Java / Maven
- Kubernetes, Docker

Summary

- Problem: We would like to be able to scale knowledge graphs.
- Solution: Create an algorithm that is able to mimic real-world graphs.
- Goal: Improve this library with respect to runtime and functionalities.



Further information:

- Master thesis
  `https://hobbitdata.informatik.uni-leipzig.de/teaching/projectgroups/Thesis-Final-lemming.pdf`

Github projects:

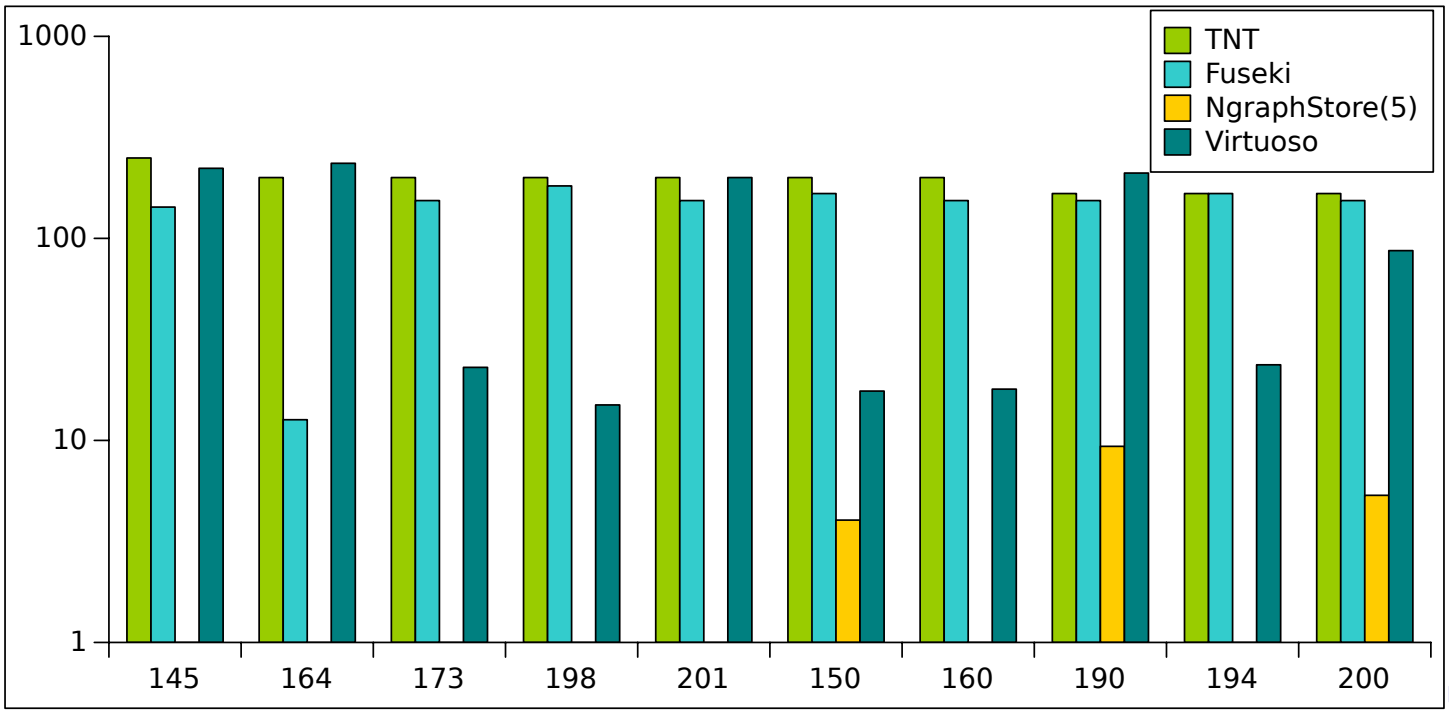
- `https://github.com/dice-group/Lemming`

Technologies:

- Graph theory
- RDF
- Java / Maven

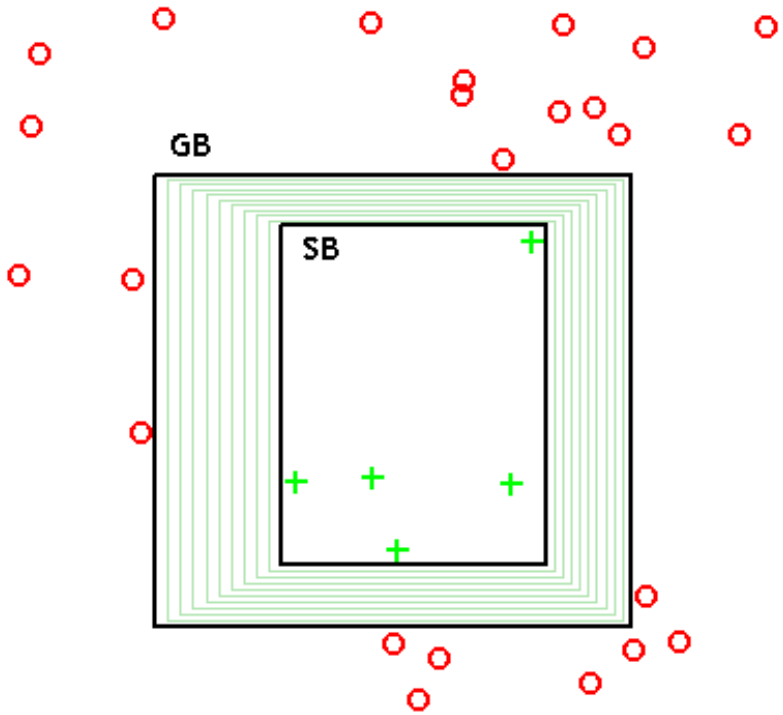
Summary

- Problem: Benchmark results are a bunch of numbers
- Solution: Search for explanations of them
- Goal: Create a framework for version space learning and apply it to the SPAB use case



Summary

- Problem: Benchmark results are a bunch of numbers
- Solution: Search for explanations of them
- Goal: Create a framework for version space learning and apply it to the SPAB use case



Summary

- Problem: Benchmark results are a bunch of numbers
- Solution: Search for explanations of them
- Goal: Create a framework for version space learning and apply it to the SPAB use case

- Implement a *general* framework for version space learning in Java.
- Find solutions if certain assumptions of the version space learning algorithm do not hold for the given data.
- Apply the framework in the SPAB use case and evaluate the approach

Further information:

- Tom M. Mitchell: “Generalization as Search”. Artificial Intelligence, Volume 18, Issue 2, 1982, Pages 203-226.
  (Available within the UPB Network)
- Tom M. Mitchell: “Machine Learning”. 1997.
  (Available in the library)

Github projects:

- `https://github.com/dice-group/SPAB`

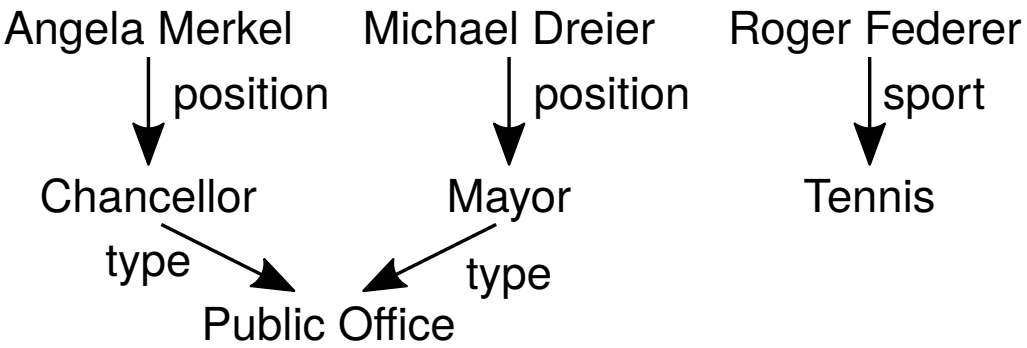
Technologies:

- Machine Learning
- Java / Maven
- SPARQL

Summary

- Problem: Neural networks not explainable, rule mining not accurate
- Solution: Combine neural networks and rule mining
- Goal: Explainable and accurate predictions

Knowledge Graph



Training Examples

Angela Merkel: Politician

Roger Federer: **not** Politician

Is Michael Dreier a politician?

Neural network: 0.95, **no explanation**

Rules: yes

$\exists$ position.public office $\sqsubseteq$ Politician

Summary

- Problem: Neural networks not explainable, rule mining not accurate
- Solution: Combine neural networks and rule mining
- Goal: Explainable and accurate predictions

Literature:

- Lehmann, J., & Hitzler, P. (2010). Concept learning in description logics using refinement operators. *Machine Learning*, 78(1-2), 203.
- Wang, Q., Wang, B., & Guo, L. (2015). Knowledge base completion using embeddings and rules. In *AAAI*.
- Ma, W., Zhang, M., Cao, Y., Jin, W., Wang, C., Liu, Y., ... & Ren, X. (2019). Jointly learning explainable rules for recommendation with knowledge graph. In *WWW* (pp. 1210-1221).

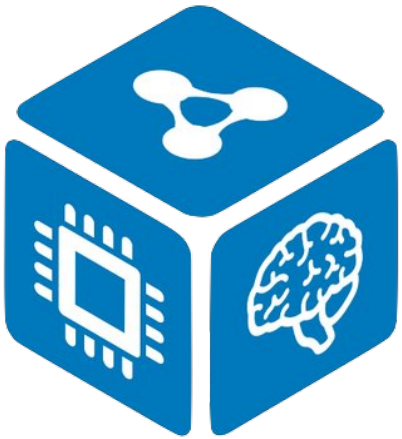
Section 4

Tasks until next week

Tasks until next week

- Until Wednesday (April 22nd, 23:59:59)
  1. Inform yourself about the topics
  2. Choose three topics (your 1st, 2nd and 3rd whish)
     send them to `michael.roeder@uni-paderborn.de`
- Next Monday (April 27th)
  - Setup of the single sub groups
- Until Monday (May 4th)
  - inform yourself about Scrum
  - inform yourself about Git
  - get a basic understanding of RDF and Semantic Web technologies

That's all Folks!



Thank you!
Questions?

Prof. Dr. Axel Ngonga
Data Science@UPB
mone@uni-paderborn.de
http://github.com/dice-group
http://dice-research.org
Follow us on Twitter @DiceResearch
Follow us on Twitter @NgongaAxel