DICE Project Groups SS-2021

Data Science Group (DICE)
Tutors: *Michael Röder, Mohamed Sherif and Stefan Heindorf*

PADERBORN UNIVERSITY
The University for the Information Society

DICE – Data Science Group, Paderborn University, Germany

February 15, 2021

Overview

Which topics will be presented?

- DICE – Data Science Group

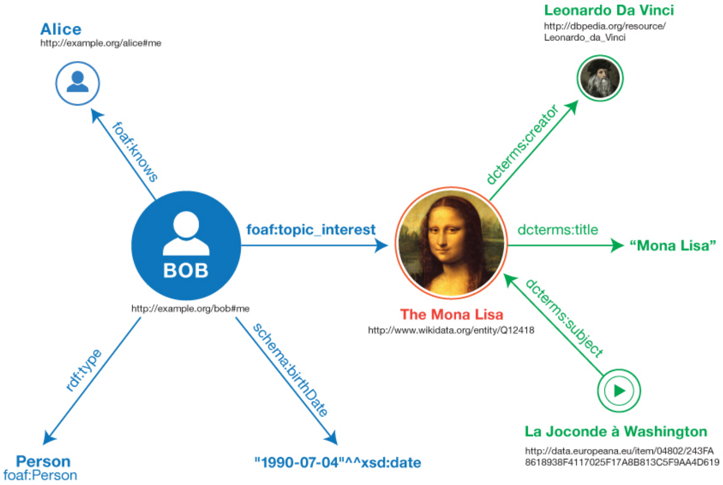
- LEMMING is an example mimicking graph generator
- ORCA: a crawler analysis benchmark
- Knowledge Graph Fusion (KGFusion)
- Explainable Artificial Intelligence (XAI)

Section 1

DICE – Data Science Group

Data Science

Knowledge Graphs (KG)

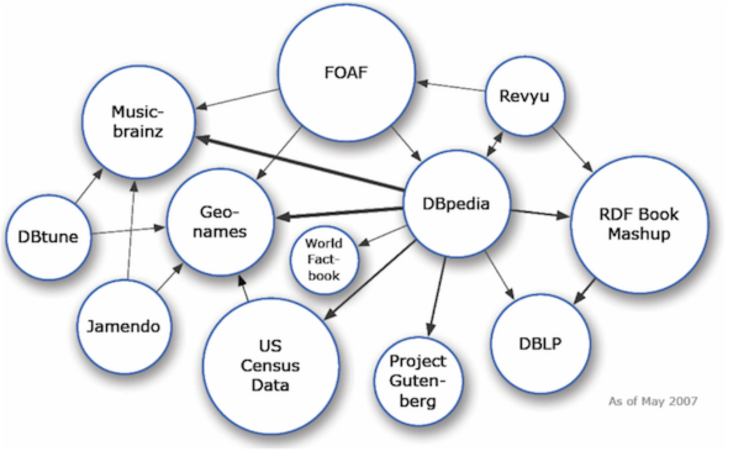


Alice
http://example.org/alice#me
foaf:knows
BOB
http://example.org/bob#me
foaf:topic_interest
rdf:type
Person
foaf:Person
schema:birthDate
"1990-07-04"^^xsd:date
Leonardo Da Vinci
http://dbpedia.org/resource/Leonardo_da_Vinci
dcterms:creator
dcterms:title
"Mona Lisa"
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:subject
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619

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

Data Science

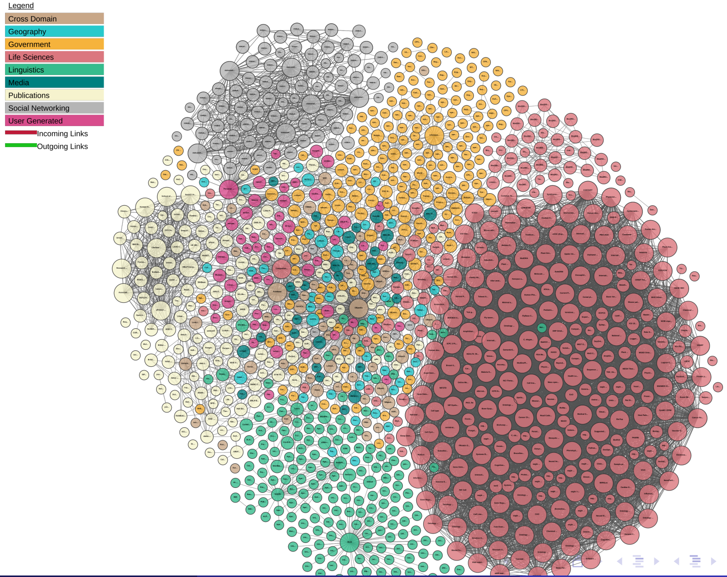
Linked Open Data (2007)



http://lod-cloud.net

Data Science

Linked Open Data (2021, `http://lod-cloud.net`)



Why we need KGs?

1. Knowledge reasoning
2. Explainable AI
3. Never ending learning
4. Natural language processing
5. Data integration
6. Intelligent Question answering
7. Fact checking
8. Digital assistants
9. . . .

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

- Machine Learning: State-of-the-art software (PyTorch, DEAP, ...)
- Real data: Millions of facts from Wikipedia (Wikidata, DBpedia)
- Expert tutors, who developed the core software
- Master theses: Topics can be extended accordingly
- Publications at top conferences (ISWC, ESWC, WWW)

Section 2

LEMMING

LEMMING

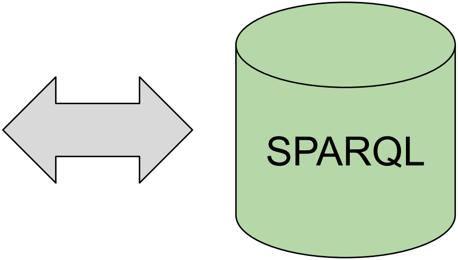
Data-driven applications

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...

LEMMING

Data-driven applications

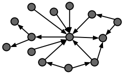
- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .



SPARQL

Industry company

- Owns a lot of data
- Wants high performance solutions for their data

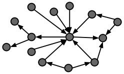


LEMMING

Data sharing problem

Industry company

- Owns a lot of data
- Wants high performance solutions for their data



Solution developer

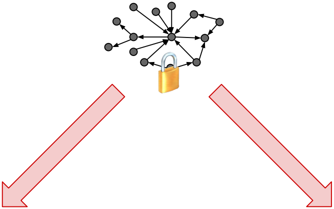
- Offers software solutions
- Can adapt it to the user's situation

LEMMING

Data sharing problem

Industry company

- Owns a lot of data
- Wants high performance solutions for their data
- Cannot share the data



Research institute

- Wants to research new approaches
- Has a limited set of data sets and generators

Solution developer

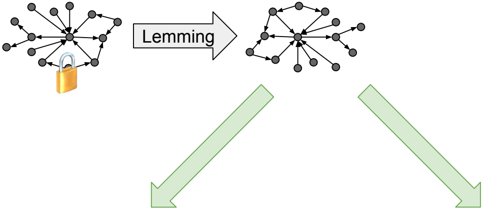
- Offers software solutions
- Can adapt it to the user’s situation

LEMMING

Data sharing problem

Industry company

- Owns a lot of data
- Wants high performance solutions for their data
- Cannot share the data



Research institute

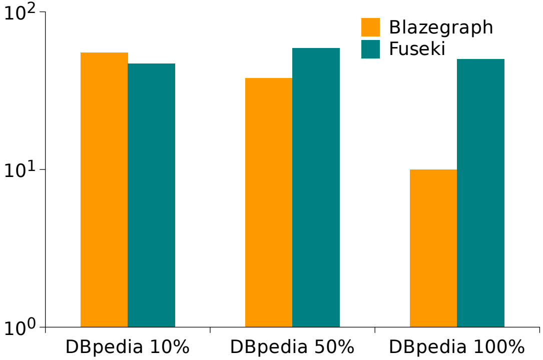
- Wants to research new approaches
- Has a limited set of data sets and generators

Solution Developer

- Offers software solutions
- Can adapt it to the user’s situation

LEMMING

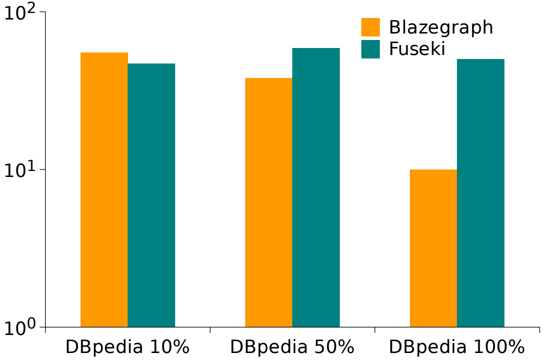
Growth of Knowledge Graphs



Conrads et al. "IGUANA : a generic framework for benchmarking the read- write performance of triple stores". ISWC 2017.

LEMMING

Growth of Knowledge Graphs

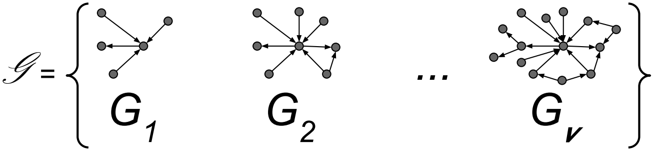


Conrads et al. "IGUANA : a generic framework for benchmarking the read- write performance of triple stores". ISWC 2017.

→ Predict the future performance of storage solutions given existing versions of a dataset.

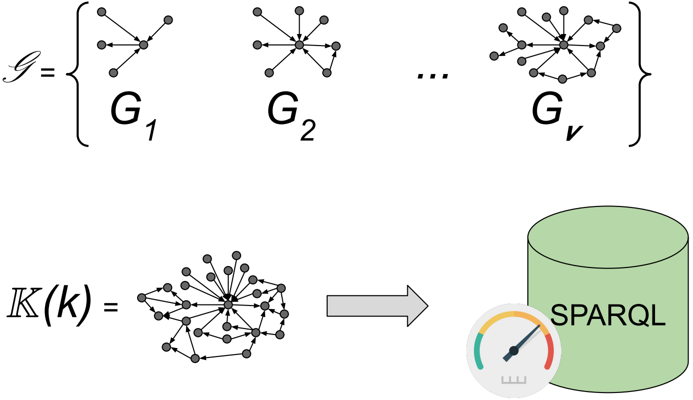
LEMMING

Goal



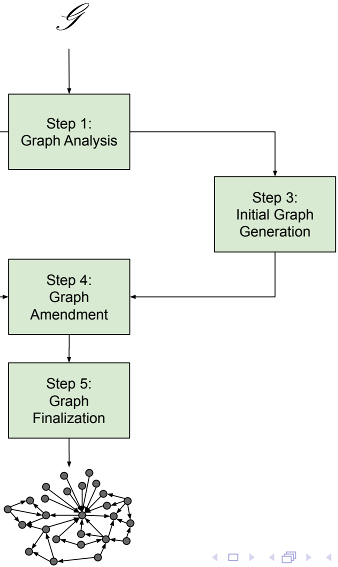
LEMMING

Goal



LEMMING

Overview



$\mathscr{G}$

Step 1: Graph Analysis

Step 2: Learn graph-specific Invariants

Step 3: Initial Graph Generation

Step 4: Graph Amendment

Step 5: Graph Finalization

Summary

- Problem: LEMMING is slow and its functionality is limited
- Solution: Enhance the existing LEMMING implementation
- Goal: Improved efficiency and effectiveness

Summary

- Problem: LEMMING is slow and its functionality is limited
- Solution: Enhance the existing LEMMING implementation
- Goal: Improved efficiency and effectiveness

- Parallelization
- Smarter metrics
- ...

- Different distribution types
- More metrics
- ...

Technologies:

- Java / Maven
- RDF (helpful)
- Graph theory (helpful)

Further information:
`https://dice-research.org/teaching/LemmingPG/`

Section 3

ORCA

ORCA

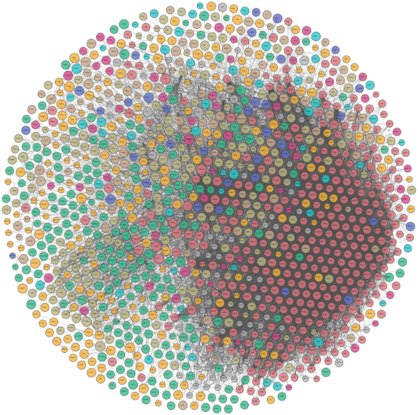
Why crawling?

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...

ORCA

Why crawling?

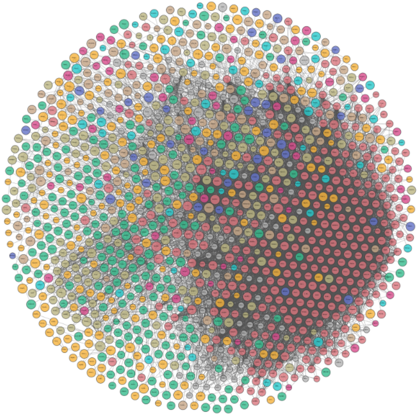
- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .



LOD cloud figures from `https://www.lod-cloud.net/`

- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .

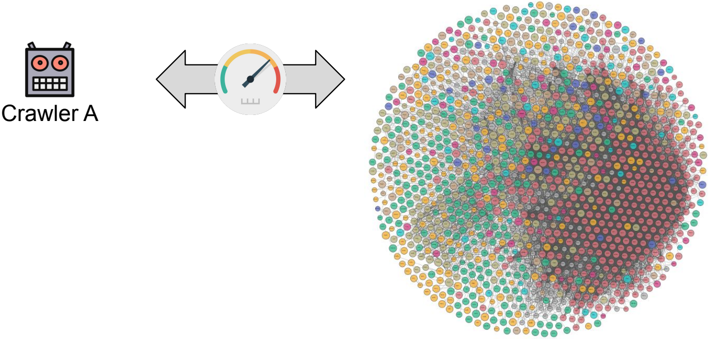
→ We need a crawler.



LOD cloud figures from https://www.lod-cloud.net/

ORCA

Why a new benchmark?

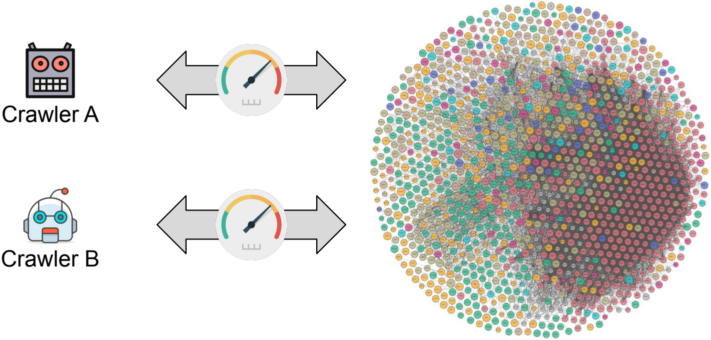


Crawler A

LOD cloud figures from `https://www.lod-cloud.net/`

ORCA

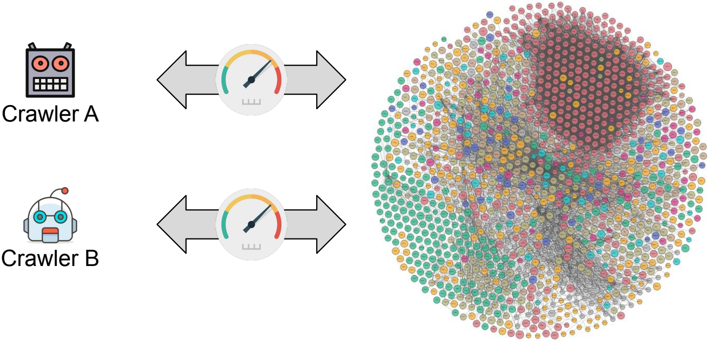
Why a new benchmark?



LOD cloud figures from `https://www.lod-cloud.net/`

ORCA

Why a new benchmark?



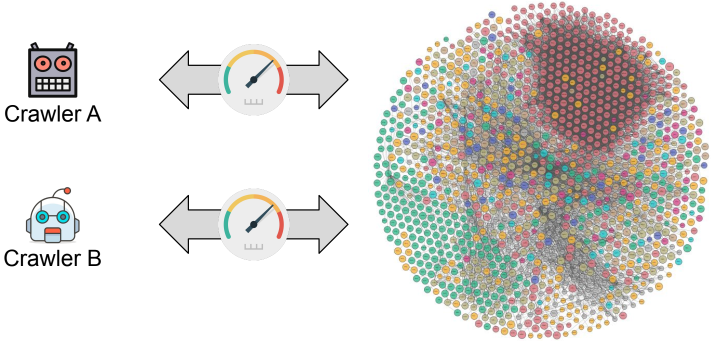
Crawler A

Crawler B

LOD cloud figures from `https://www.lod-cloud.net/`

ORCA

Why a new benchmark?



Crawler A

Crawler B

Problems

- Repeatability
- Unknown ground truth

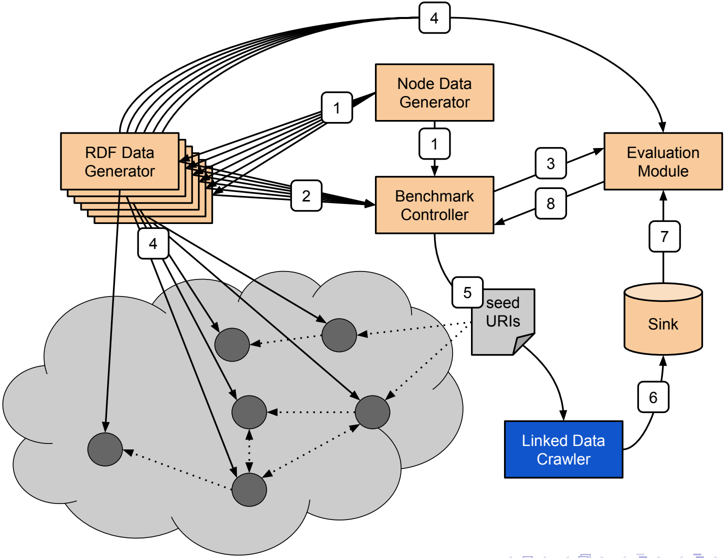
LOD cloud figures from `https://www.lod-cloud.net/`

Generate a synthetic Data Web

- Repeatable
- Scalable
- Configurable
- Ground truth is known

ORCA

Overview



ORCA

Formats

	RDF Serialisations										Comp.			HTML				
	RDF/XML	RDF/JSON	Turtle	N-Triples	N-Quads	Notation 3	JSON-LD	TriG	TriX	HDT	ZIP	Gzip	bzip2	RDFa	Microdata	Microformat	SPARQL	CKAN
ORCA	✓	(✓)	✓	✓	(✓)	✓	(✓)	(✓)	(✓)	–	✓	✓	✓	✓	–	–	✓	✓

ORCA

Goal

Summary

- Problem: ORCA does not reflect all major technologies
- Solution: Extend ORCA in various directions
- Goal: Evaluation results of a new ORCA version

ORCA

Goal

Summary

- Problem: ORCA does not reflect all major technologies
- Solution: Extend ORCA in various directions
- Goal: Evaluation results of a new ORCA version

- More compression algorithms
- Microdata, microformat, ...
- Existing RDF data generators

- More complex graph generators
- ...

ORCA

Summary

Technologies:

- Java / Maven
- RDF (helpful)
- Docker (helpful)

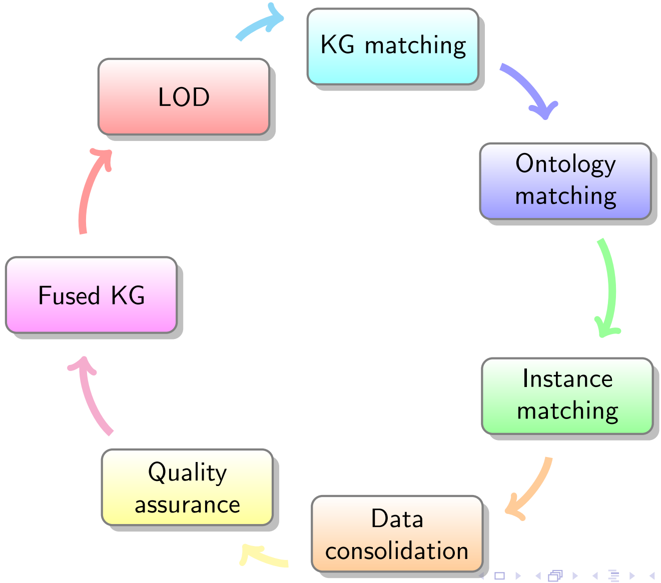
Further information:
`https://dice-research.org/teaching/LemmingPG/`

Section 4

Knowledge Graph Fusion (KG Fusion)

KG Fusion: Overview

What is KG fusion?



KG Fusion: Motivation

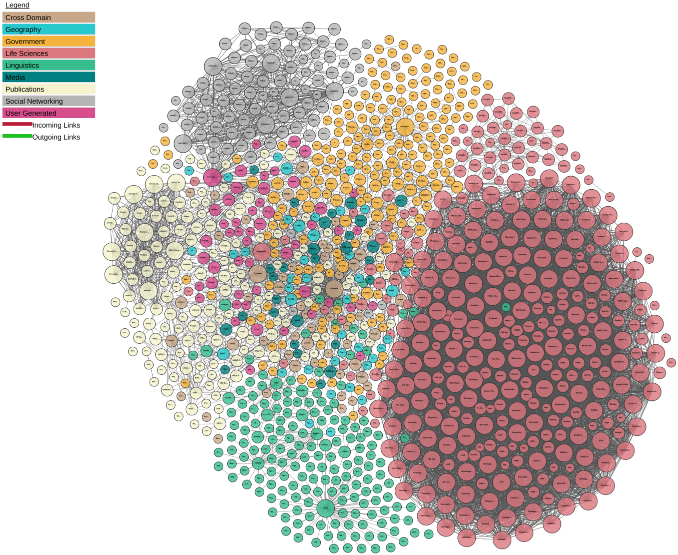
Why we need KG fusion?

Fused KG should be

1. more complete
2. more accurate
3. non redundant
4. richer
5. cleaner and
6. as universal description for the respective resources

KG Fusion: KG matching

Linked Open Data 2020 (http://lod-cloud.net)



KG Fusion

Example: Fusing *DBpedia* and *Yago*

1. KG matching
   - https://dbpedia.org
   - https://yago-knowledge.org

Summary

- Problem: KG topic(s) is not explicitly defined
- Solution: Apply KG matching techniques
- Goal: Limit next steps to deal only with similar KGs

KG Fusion

Example: Fusing *DBpedia* and *Yago*

2. Ontology matching
   - `https://dbpedia.org/ontology/Town`
   - `https://yago-knowledge.org/resource/schema:City`

Summary

- Problem: Classes have different labels, structure and ontologies
- Solution: Apply ontology matching techniques
- Goal: Next step match only instances of similar classes

KG Fusion

Example: Fusing *DBpedia* and *Yago*

3. Instance matching
   - `https://dbpedia.org/resource/Paderborn`
   - `https://yago-knowledge.org/resource/Paderborn`

Summary

- Problem: KG instances have different labels, structure and ontologies
- Solution: Apply link discovery techniques
- Goal: Next step fuse only similar instances

KG Fusion

Example: Fusing *DBpedia* and *Yago*

4. Data consolidation
   - Paderborn location in *DBpedia* is defined using `georss:point` to be `(51.718055555555556, 8.754166666666666)`
   - Paderborn location in *Yago* is defined using `schema:geo` to be `(51.7167701, 8.7666842)`
   - Fuse using the *keep most precise value* strategy

Summary

- Problem: KG instances have different properties labels and values
- Solution: Implement automatic fusion strategies
- Goal: Generate fused KG

KG Fusion

Example: Fusing *DBpedia* and *Yago*

5. Quality assurance
    - Benchmark the resulted fused KG

Summary

- Problem: No benchmark exist for KG fusion
- Solution: Generate our own benchmark for KG fusion
- Goal: Assure the quality of the fused KG

Section 5

Explainable Artificial Intelligence (XAI)

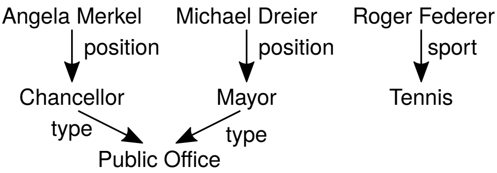
XAI: Motivation

Explainable Machine Learning on Knowledge Graphs

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?

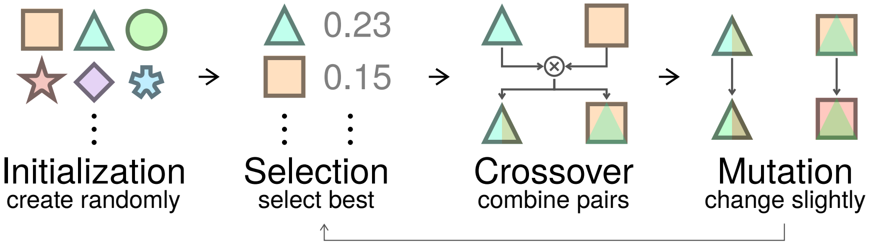
Neural network: 0.95, **no explanation**

Rules: yes

∃ position.public office ⊑ Politician

XAI: Motivation

Evolutionary Algorithms for Rule Mining



Initialization
create randomly

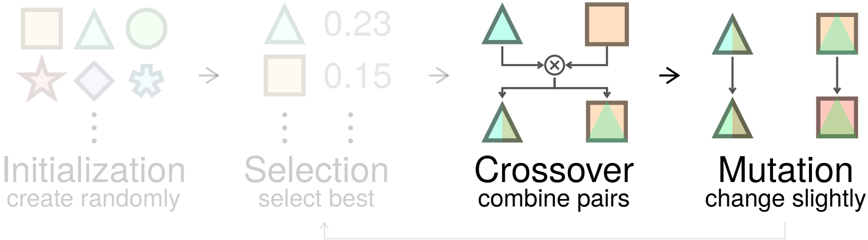
Selection
select best

Crossover
combine pairs

Mutation
change slightly

Summary

- Problem: Many (bad) candidates generated
- Solution: Guide crossover and mutation with neural network
- Goal: Generate promising candidates as soon as possible

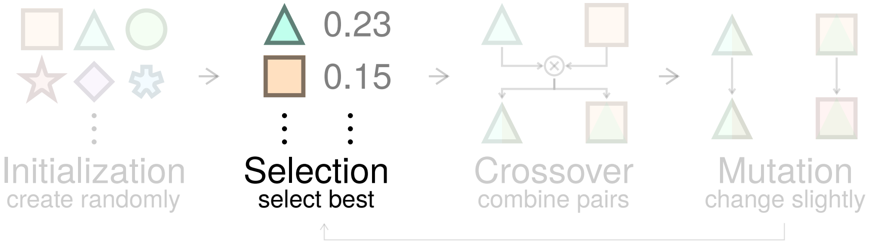


XAI: Topics

Surrogate Models for Fast Evaluation of Fitness Function

Summary

- Problem: Evaluation of fitness function takes long time
- Solution: Approximate fitness function with surrogate model
- Goal: Enable the evaluation of more candidates



XAI: Topics

Corpus Construction for Realistic Evaluation

Summary

- Problem: Existing benchmarking datasets artificial
- Solution: Construct realistic datasets for important use cases
- Goal: Realistic evaluation of rule miners

Important use cases:

- Type prediction
- Vandalism detection

XAI: Summary

Explainable Artificial Intelligence on Knowledge Graphs

Summary

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

Technical skills to learn

- Neural networks (PyTorch)
- Evolutionary algorithms (DEAP)
- Data analysis (Pandas)

Scientific skills to learn

- Literature review
- Scientific presentation
- Scientific writing

That’s all Folks!

Thank you!

Topics:

- Knowledge Graphs
- Machine Learning
- Explainability

The topics are subject to change.
More information at
`https://dice-research.org`