

HOBBIT

A Brief Overview

Michael Röder, Axel-Cyrille Ngonga Ngomo

DICE research group
Institute of applied Informatics, Leipzig, Germany
University Paderborn, Germany
(Horizon 2020, GA No 688227)



Presentation ZB MED
Paderborn, September 12th, 2019

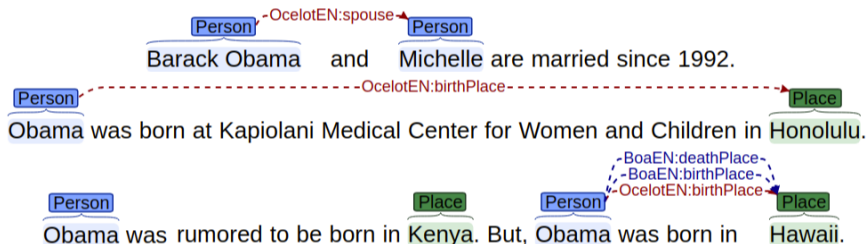
- 1 Fundamentals
- 2 Introduction
- 3 Project Highlights
- 4 Benchmarking Machine Learning
- 5 Future Directions

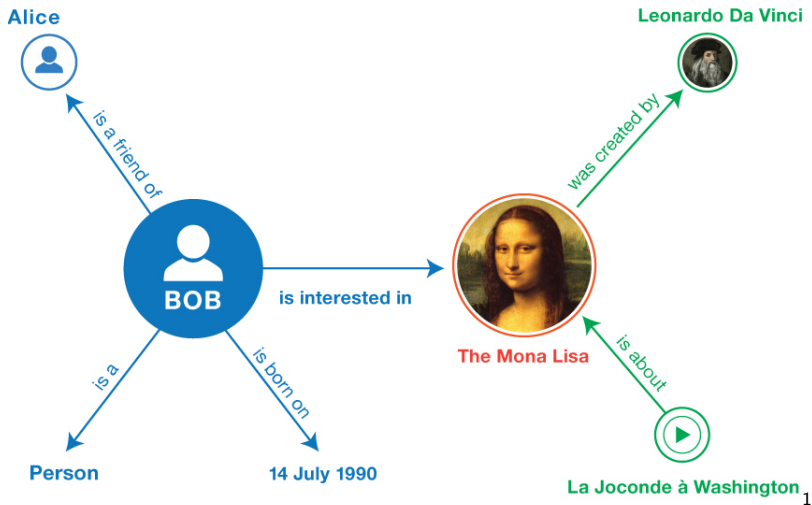
Section 1

Fundamentals

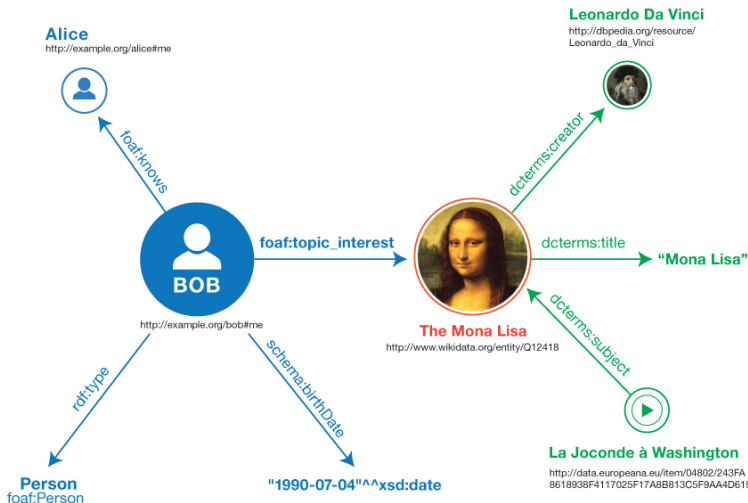
Barack Obama and Michelle are married since 1992. Obama was born at Kapiolani Medical Center for Women and Children in Honolulu. Obama was rumored to be born in Kenya. But, Obama was born in Hawaii.

Barack Obama and Michelle are married since 1992. Obama was born at Kapiolani Medical Center for Women and Children in Honolulu. Obama was rumored to be born in Kenya. But, Obama was born in Hawaii.



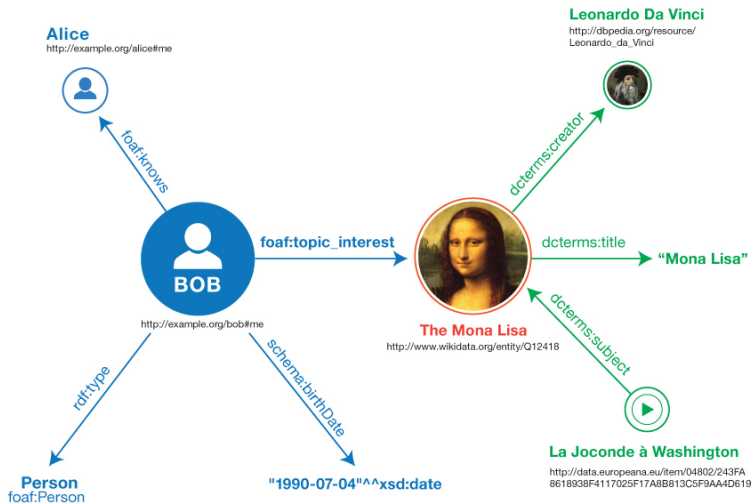


¹<https://www.w3.org/TR/rdf11-primer/>



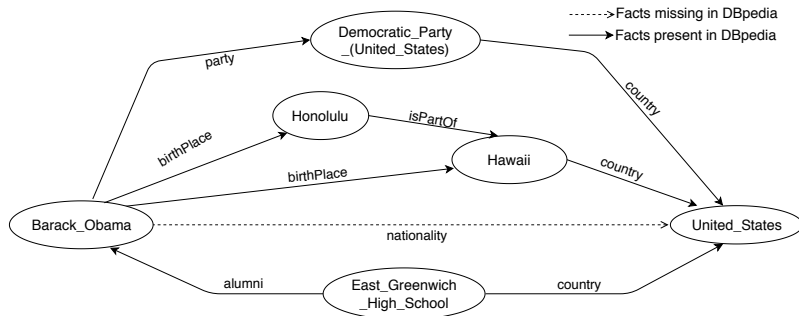
2

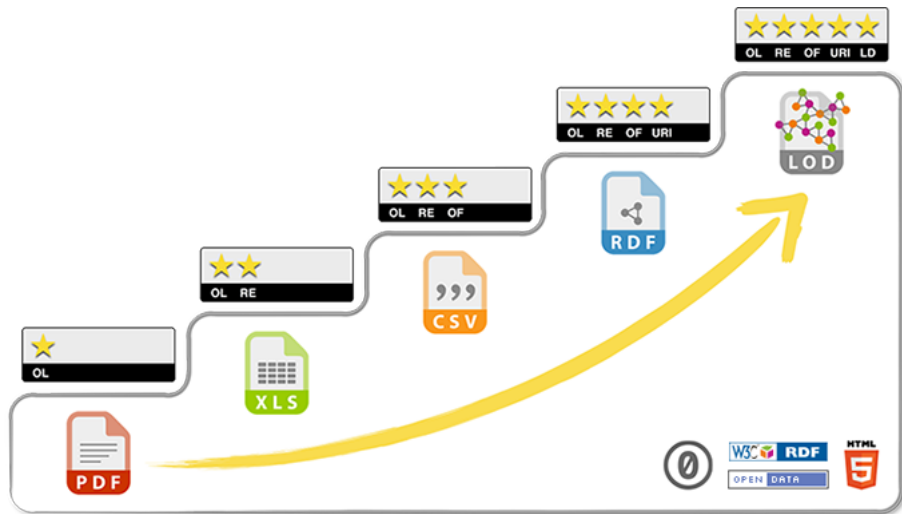
²<https://www.w3.org/TR/rdf11-primer/>

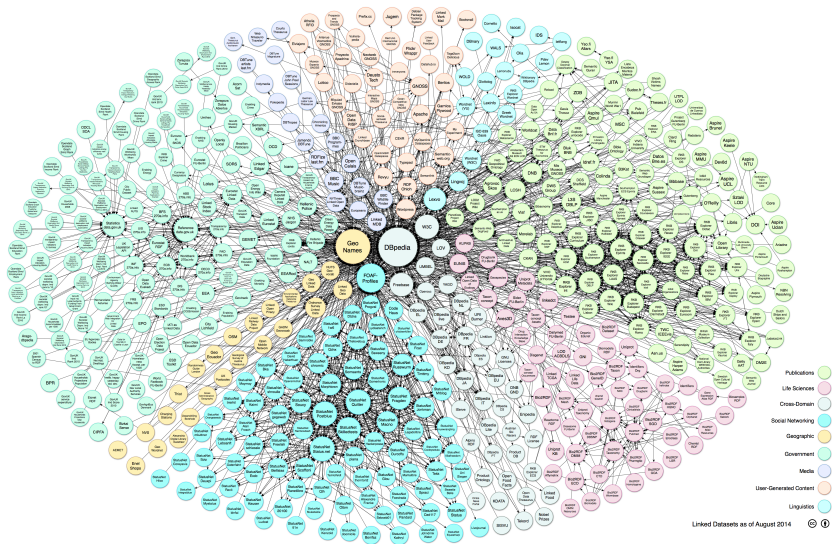


```
<http://example.org/bob#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://example.org/bob#me> <http://xmlns.com/foaf/0.1/knows> <http://example.org/alice#me> .
<http://example.org/bob#me> <http://schema.org/birthDate> "1990-07-04"^^<http://www.w3.org/2001/XMLSchema#date> .
```

...





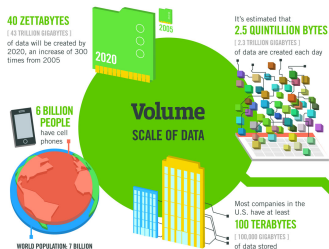


Section 2

Introduction

Introduction

A Lot of Data



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity ANALYSIS OF STREAMING DATA



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** (161 BILLION GIGABYTES)



30 BILLION PIECES OF CONTENT are shared on Facebook every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



400 MILLION TWEETS are sent per day by about 200 million monthly active users



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Veracity UNCERTAINTY OF DATA

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

³<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>

Introduction

A Lot of Tools

The Dataflog Open Source Landscape 2.0



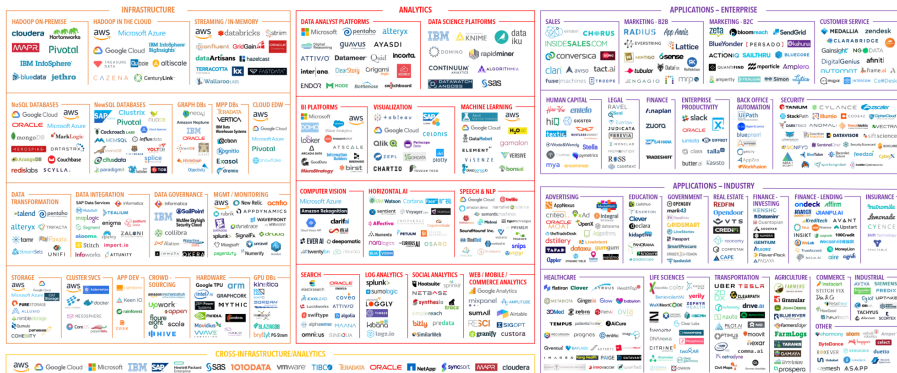
⁴https://cdn.dataflog.com/cms/os_big_data_open_source_tools-v2.png

Introduction

A Lot of Tools



BIG DATA & AI LANDSCAPE 2018



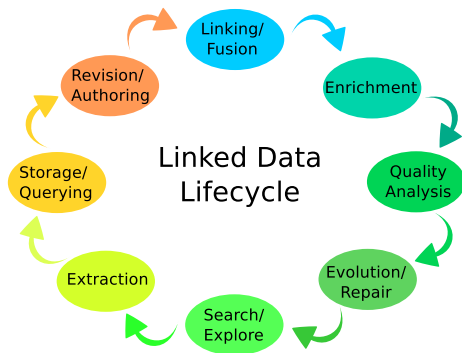
Which tool(s) should I use
for my use case?



- Which tool(s) should I use for my use case?
- Which key performance indicators are relevant?
- How do existing solutions perform w.r.t. relevant indicators?
- Where are the current bottlenecks?
- Which steps of the data lifecycle are critical?
- ...



- Research project from 2015 – 2018 (Horizon 2020, GA No 688227)
- Focus on Big Linked Data
- Cover the business-critical steps of the Linked Data lifecycle
- Used by a growing number of companies
- Mature and maturing technologies





1 Gathered real requirements

- Focussed on industrial requirements
- Gathered relevant performance indicators
- Gathered relevant performance thresholds
- Gathered real datasets



- ❶ Gathered real requirements
 - Focussed on industrial requirements
 - Gathered relevant performance indicators
 - Gathered relevant performance thresholds
 - Gathered real datasets
- ❷ Developed benchmarks based on real data



- ❶ Gathered real requirements
 - Focussed on industrial requirements
 - Gathered relevant performance indicators
 - Gathered relevant performance thresholds
 - Gathered real datasets
- ❷ Developed benchmarks based on real data
- ❸ Provided universal benchmarking platform
 - Comparable results
 - Hosted as a free-to-use online instance



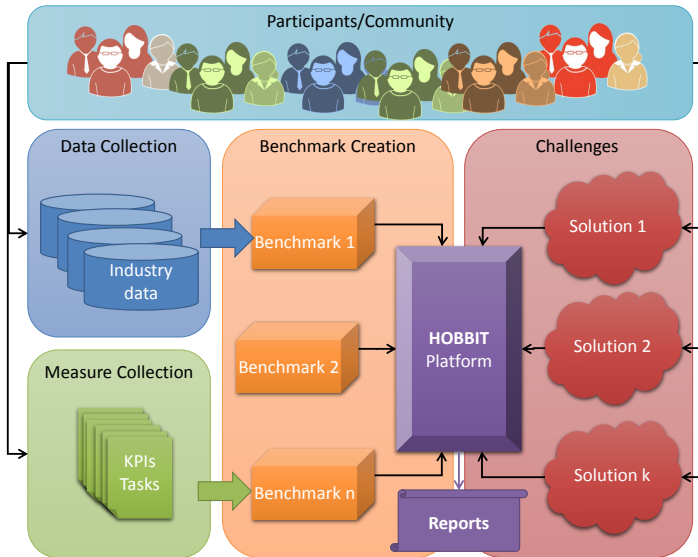
- ❶ Gathered real requirements
 - Focussed on industrial requirements
 - Gathered relevant performance indicators
 - Gathered relevant performance thresholds
 - Gathered real datasets
- ❷ Developed benchmarks based on real data
- ❸ Provided universal benchmarking platform
 - Comparable results
 - Hosted as a free-to-use online instance
- ❹ Periodic benchmarking challenges and reporting



- ① Gathered real requirements
 - Focused on industrial requirements
 - Gathered relevant performance indicators
 - Gathered relevant performance thresholds
 - Gathered real datasets
- ② Developed benchmarks based on real data
- ③ Provided universal benchmarking platform
 - Comparable results
 - Hosted as a free-to-use online instance
- ④ Periodic benchmarking challenges and reporting
- ⑤ Created an association (Special Group 7 of Task Force 6)

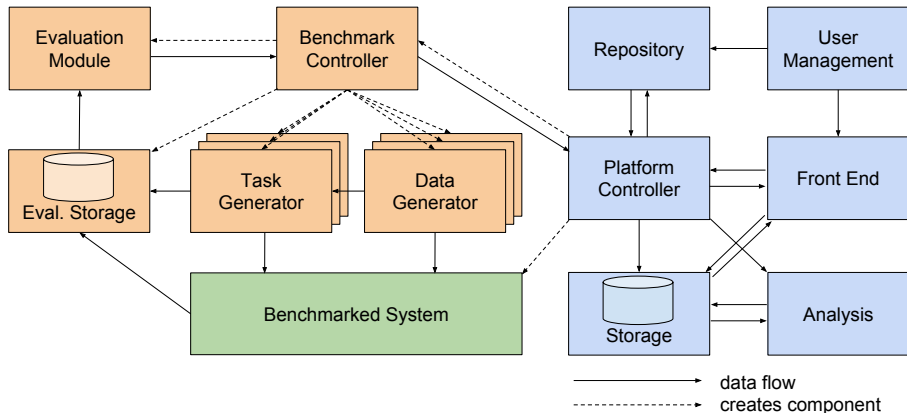
Overview

Architecture



Section 3

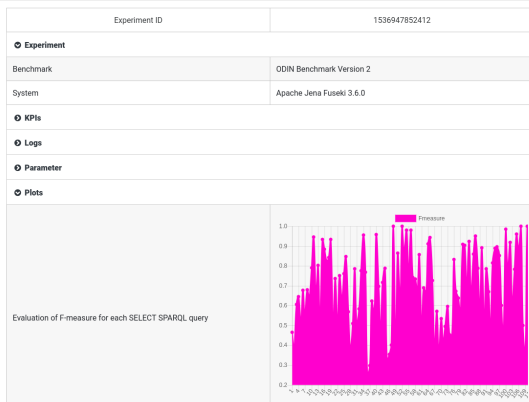
Project Highlights

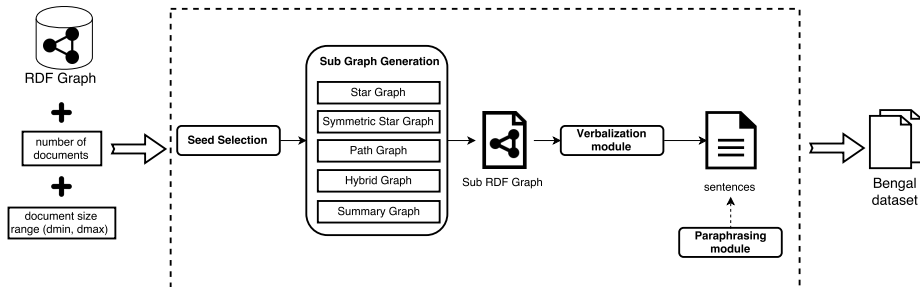


- Scalable open-source benchmarking platform
- Local, distributed and remote (AWS) deployment
- First FAIR platform for benchmarking Big Linked Data in a holistic manner

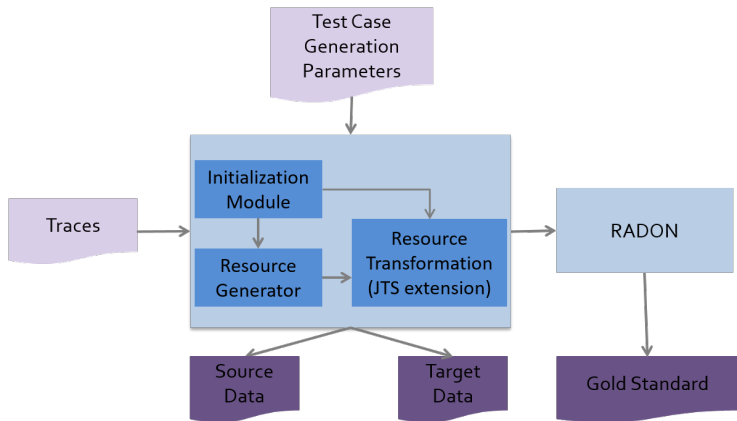
- ★ 5 mimicking algorithms
- ★ 52 benchmarks
- ★ 200+ systems
- ★ 14 challenges
DEBS GC 2017 and 2018
- ★ 300+ users
- ★ 13K+ experiments

Experiment Details

[Back](#)



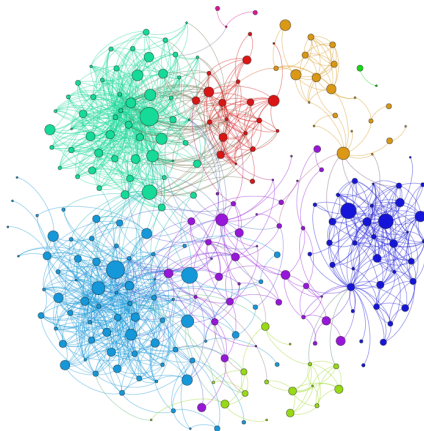
- ★ 21 systems benchmarked
- ★ 2 challenges organized
- ★ 2 mimicking algorithms (BENGAL and TWIG)
- ★ Integration with GERBIL



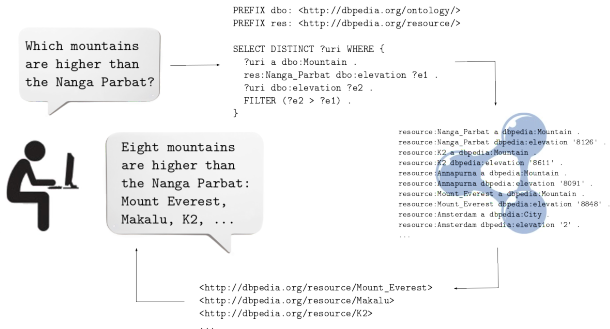
- ★ 7 systems benchmarked
- ★ 2 challenges (2× OAEI)
- ★ Collaboration with OAEI

Overview and Highlights

Storage and Versioning



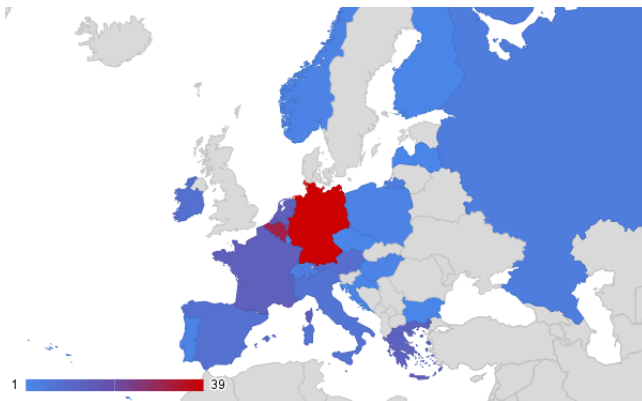
- ★ 9 systems benchmarked
- ★ 3 challenges organized
- ★ Improvement of world's most scalable triple store



- ★ 7 systems benchmarked
- ★ 3 challenges organized
- ★ More realistic benchmarks unveiled current systems' drawbacks



- ★ 14 challenges over 8 benchmarks organized (ESWC, ISWC and DEBS)
- ★ 77 systems benchmarked overall



- ★ Association created as Special Group 7 of Task Force 6 of BDVA
- ★ 300 members in industry-focused contact list
- ★ Participation in 40+ events with 10K+ participants
- ★ European focus with members from outside Europe

Section 4

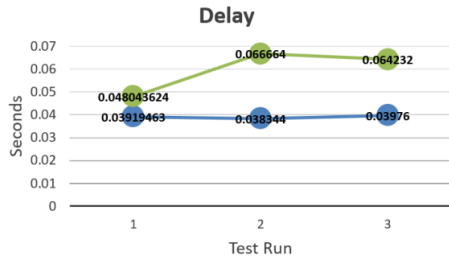
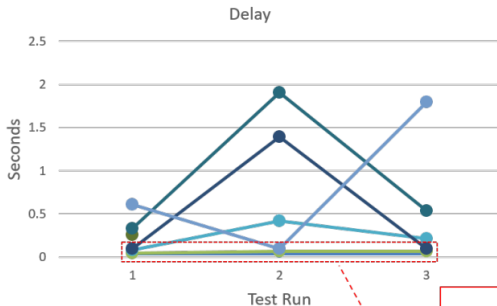
Benchmarking Machine Learning

- The task: find anomalies in molding machine sensor data to predict maintenance intervals (predictive maintenance).
- Mimicking algorithm based on real data
- Data was streamed as in the real world
- Participants had to use Markov Models to identify anomalies
- 14 Participants, 7 made it into the last round



Benchmarking Machine Learning

SML Benchmark v1 for DEBS GC 2017



- The task: predictions about ship routes based on AIS data
 - Spatio-temporal streaming data
 - Predictions for vessels' destinations and arrival times



Team	Earliness rate	A	Working time (sec)	B	Total Q1
University of Iasi	0.685	1	99	2	1.25
University of Illinois	0.672	2	86	1	1.75
Jean Monnet University	0.668	3	149	5	3.5
Chungnam National University	0.653	4	102	3	3.75
University of Iasi (2nd)	0.647	5	157	6	5.25
Israel Institute of Technology	0.5	6	129	4	5.5
Dresden University of Technology	-	-	-	-	-
Insight Centre	-	-	-	-	-
University of Carthage	-	-	-	-	-

Team	Mean Absolute Error (min.)	A	Working time (sec)	B	Total Q2
University of Iasi	959.839	1	100	2	1.25
Jean Monnet University	1099	2	145	4	2.5
Chungnam National University	1251.15	3	100	2	2.75
Israel Institute of Technology	1493.18	4	133	3	3.75
University of Illinois	5425.53	5	86	1	4.75
University of Iasi (2nd)	1705.35	6	164	5	5
Dresden University of Technology	-	-	-	-	-
Insight Centre	-	-	-	-	-
University of Carthage	-	-	-	-	-

Team	Q1	Q2	Total Score
University of Iasi	1.25	1.25	2.5
Jean Monnet University	3.5	2.5	6
Chungnam National University	3.75	2.75	6.5
University of Illinois	1.75	4.75	6.5
Israel Institute of Technology	5.5	3.75	9.25
University of Iasi (2nd)	5.25	5	10.25
Dresden University of Technology	-	-	-
Insight Centre	-	-	-
University of Carthage	-	-	-

Section 5

Future Directions

- KnowGraphs (Innovative Training Networks (ITN))
 - 4 years, starting in October 2019
 - 15 Early-Stage Researchers (ESRs)
 - HOBBIT will be used as central benchmarking platform
 - Further datasets will be integrated (e.g., UICML datasets)
- RAKI (BMW project)
 - 3 years, starting in September 2019
 - HOBBIT will be used for evaluation
- More projects pending

- KnowGraphs (Innovative Training Networks (ITN))
 - 4 years, starting in October 2019
 - 15 Early-Stage Researchers (ESRs)
 - HOBBIT will be used as central benchmarking platform
 - Further datasets will be integrated (e.g., UICML datasets)
 - RAKI (BMW project)
 - 3 years, starting in September 2019
 - HOBBIT will be used for evaluation
 - More projects pending
- Further development of the HOBBIT platform
- HOBBIT is open for the community! Benchmarks, systems, datasets can be added
 - Not limited to linked data

HOBBIT offers

- Scalable benchmarking
- Based on real world data in an
- Extendable,
- Open source platform
- Following the FAIR data principles

<http://project-hobbit.eu/>
<https://dice-research.org/about/>

SUMMARY

