CERN openlab

Data Analytics Platform for Science

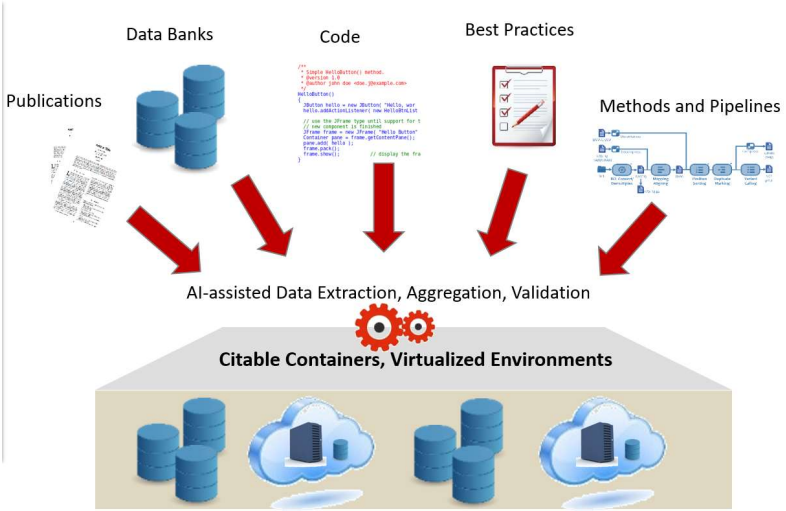
CERN openlab Technical Workshop 11.01.2018

Taghi Aliyev

11/01/2018

Introduction

- Large-scale collaborative research platform
- Main focus on ease-of-use, reproducibility of research
- Use of Machine Learning for Narrative interfaces
  - Information Retrieval
  - Natural Language Processing (Chatbots)
- Provide and host in-house solutions and projects

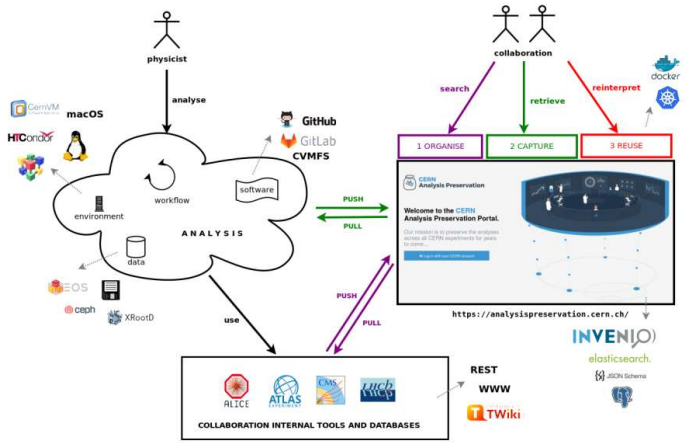


What do we mean as a Platform

- Idea is to not just provide tools to researchers
- Powerful ecosystem
  - Challenge the value chain and the ideas
- Focus on the ‘why’ of things rather than ‘how-do’
  - Enhance the way research is done

CERN Technologies

- Zenodo
  - Data Base of Publications and Presentations with links to social media
- CVMFS
  - Storage and distribution of tools and software
- REANA
  - Orchestration layer of the platform
  - Working closely with the team and Tibor Simko



@tiborsimko

Narrative Interfaces and Chatbots

- Personal Assistants
  - To ease the entry to the field
- Capturing the user behavior
  - Departing scientists challenge
- Textual Inference Problem
  - To be able to suggest from large number of publications
  - Answer scientific queries
  - Machine Comprehension Challenge
- Zenodo and other public repositories as training and validation points
  - NCBI (Dataset+Publication), Biostar, Arxiv.org

Relevant Research on NLP

Machine Comprehension

- Two main directions in the field of MC:
  - Answering/Asking Questions based on a piece (Supervised Learning)
  - Generation and Collection of Benchmark Data Sets
  - Supervised learning approaches only after 2015
- Different approaches for answering queries:
  - R-Net (Gated Self-matching Networks)
  - Smarnet
  - Google DeepMind NIPS 2015 paper
- Multiple data sets compiled from online information sources
  - SQuAD (Stanford Question Answering Dataset, Based on Wikipedia Articles)
  - NewsQA (Based on CNN/Daily Mail Articles)

Machine Comprehension in Platform

How we see it in our case?

- No relevant work in exploring scientific publications
  - No compiled Data Sets
  - Missing benchmark results of different algorithms
  - Current state-of-the-art algorithms performing well, but still worse than humans
- A smart bot that could both answer relevant queries or ask questions to improve learning based on User Profile
  - Act as a personal assistant
  - Virtual Teacher for newcomers in the field
  - Help in Training programs (Cambridge)

- Decrease and eliminate possible bias in research
  - Scan of multiple publications for relevant information

Use Cases

How to achieve initial designs?

- Core team concept
- Working with the members and representatives of the community directly
- Multiple initial use cases to define Minimal Viable Platform
  - To initiate the future talks and the iterative process
- Generate an awareness

A bit more on use cases and MVP

- Two ongoing projects:
  - King's College London and SIDRA : Genomics in ROOT and benchmarking of CNV tools
  - Maastricht University: Imputation-based Machine Learning for Target Lipid identification in Lipidomics
- More partners to come:
  - Cambridge University: Training of Future generations and automated benchmarking of newly deployed tools
  - EBI: Learning and Analysis of Web Logs for purposes of NLP

Future Tasks

- Meeting with members of the Community
- Implementation and Design of the Minimal Viable Platform
- Gathering of Feedback and initialization of the Feedback-loop with the users
- Design and Tests on NLP Models
  - Parsing and preparation of the Data sets
  - Designing test cases

Thanks!

Taghi.aliyev@cern.ch
Twitter: @TaghiAliyev