



Multi-Science Data Analysis Platform

IT Technical Forum 08.12.2017

Taghi Aliyev

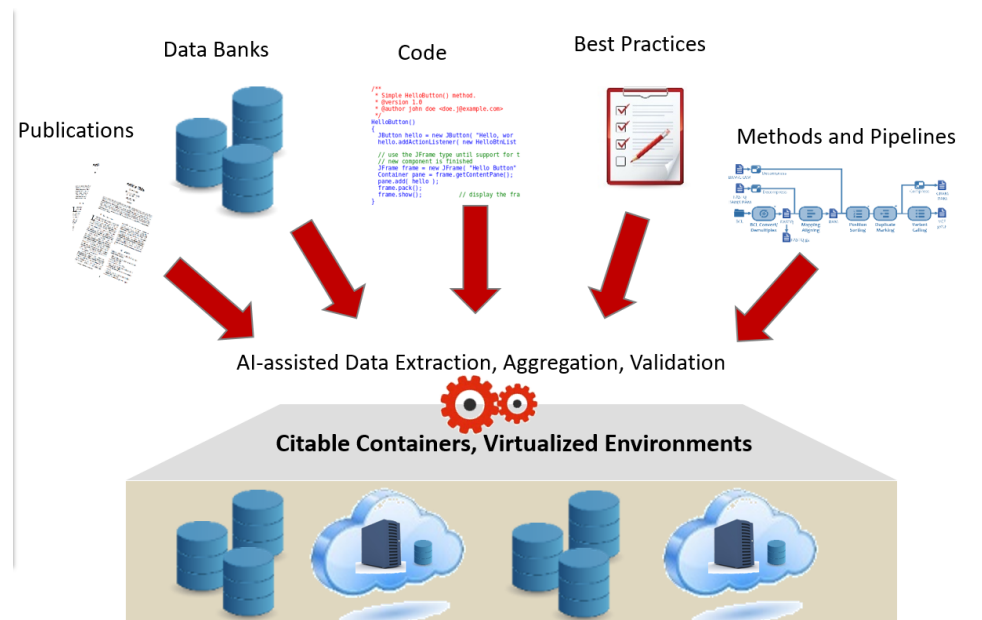
08/12/2017

Outline

- Introduction
- Motivation
- What do we mean as a Platform and Why at CERN?
- CERN Technologies in use
- Narrative Interfaces and Chatbots
- Current Use Cases

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

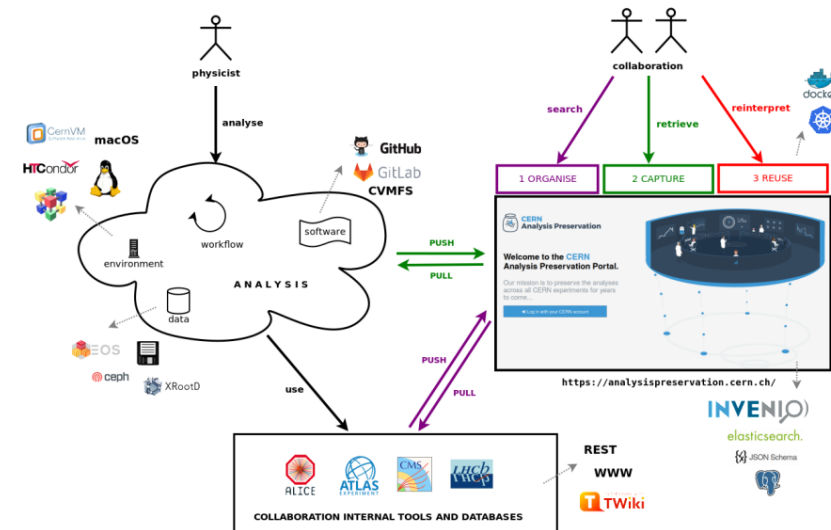
Motivation

What sparked the idea? Why do we do this at CERN?

- Talks with Genomics community last year
- Data → Commodity
- CERN and HEP community's experience and culture of sharing
 - Hypothesis Memory
 - More focused community-driven sharing tools

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



Narrative Interfaces and Chatbots

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

Relevant Research on NLP

Machine Comprehension

- What has been done in research already?
- NIPS 2015, Google DeepMind Paper
- Deep Neural Networks based models to learn to read documents

Teaching Machines to Read and Comprehend

Karl Moritz Hermann[†] Tomáš Kočiský^{†‡} Edward Grefenstette[†]
Lasse Espeholt[†] Will Kay[†] Mustafa Suleyman[†] Phil Blunsom^{†‡}

[†]Google DeepMind [‡]University of Oxford
{kmh,tkocisky,etg,lespeholt,wkay,mustafasul,pblunsom}@google.com

Abstract

Teaching machines to read natural language documents remains an elusive challenge. Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation. In this work we define a new methodology that resolves this bottleneck and provides large scale supervised reading comprehension data. This allows us to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure.

Taghi Aliyev, IT Technical Forum

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 (Fons)
 - 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



Questions?

Taghi.aliyev@cern.ch

Twitter: @TaghiAliyev