CERN openlab

Smart Data Analytics Platform for Science

ECRMA Short Talks 12.02.2018

Taghi Aliyev

12/02/2018

Outline

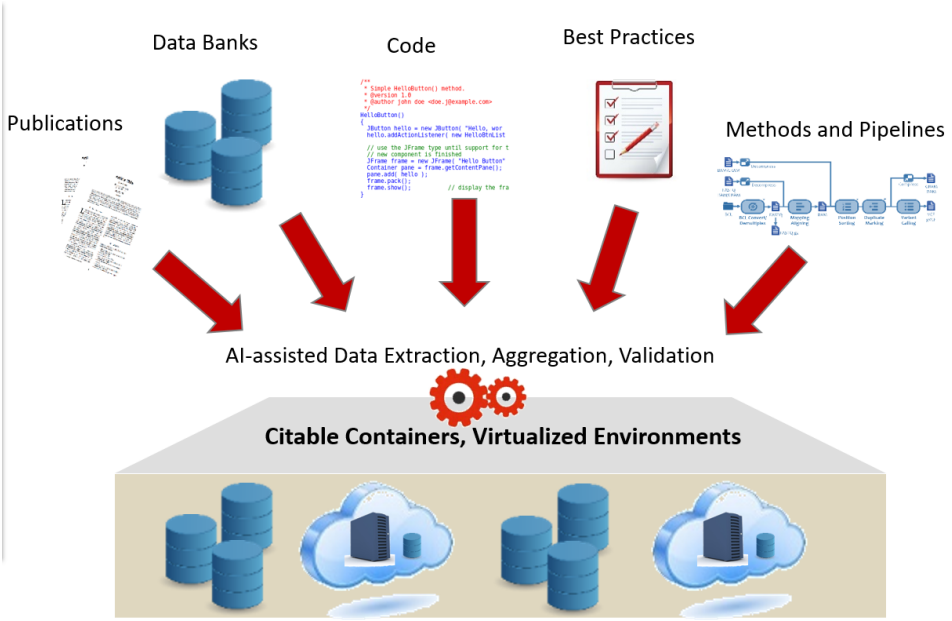
- Background and Motivation
- Introduction to the Platform
- Chatbots and Natural Language Processing
- Use Cases and the initial Core Team
- Future Work

Background and Motivation

- A lot of replicative work in Life Sciences
  - Non-reproducible research
  - Many different data structures and conventions → Need for parsers…
- High barriers to enter the research fields
- Lack of common ground, all-in-one environments
- Sparked out off discussion with the members of Medical Community
  - Genomics Analysis Experts, Professors in Bio-Informatics, personal experiences

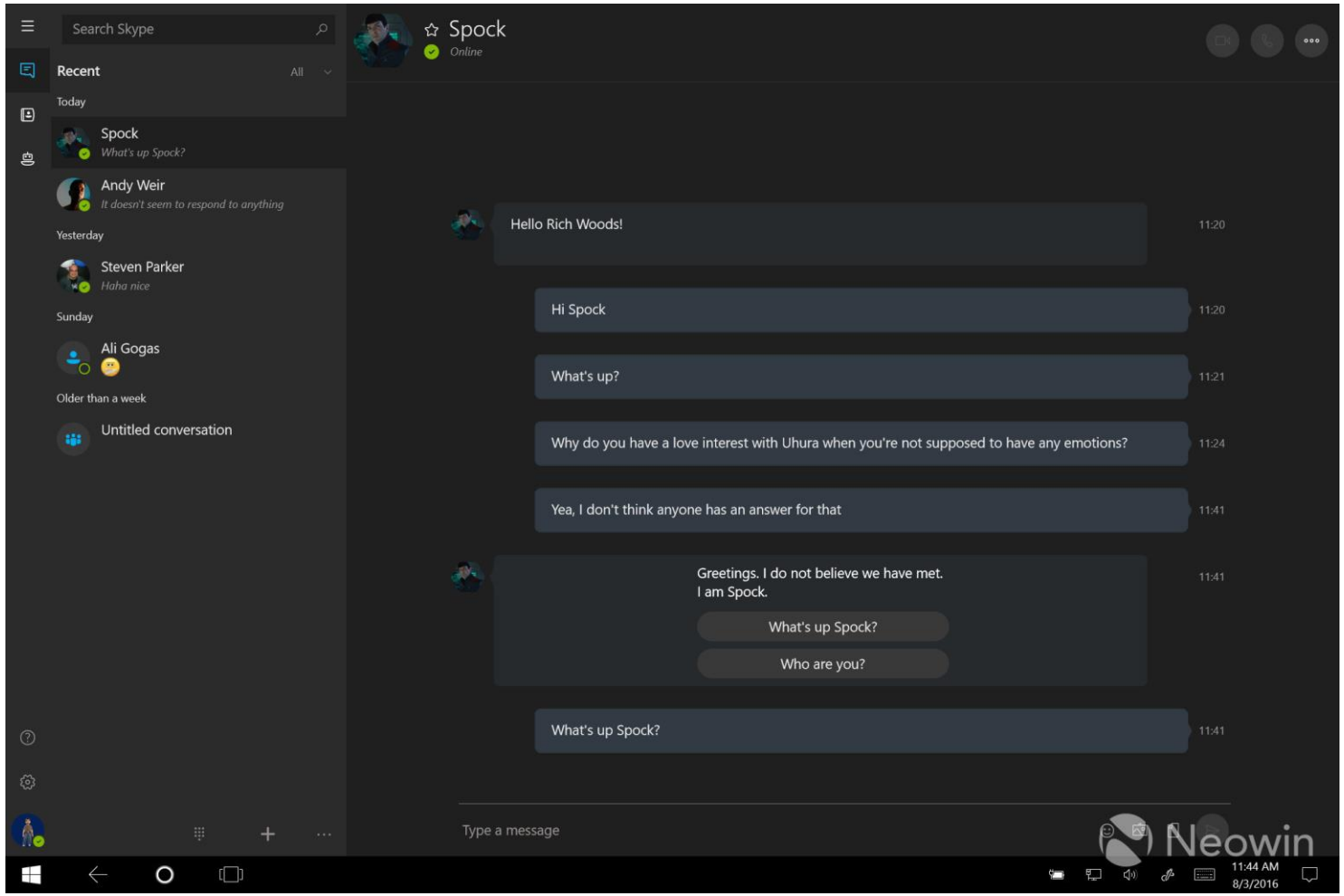
Introduction to the Platform

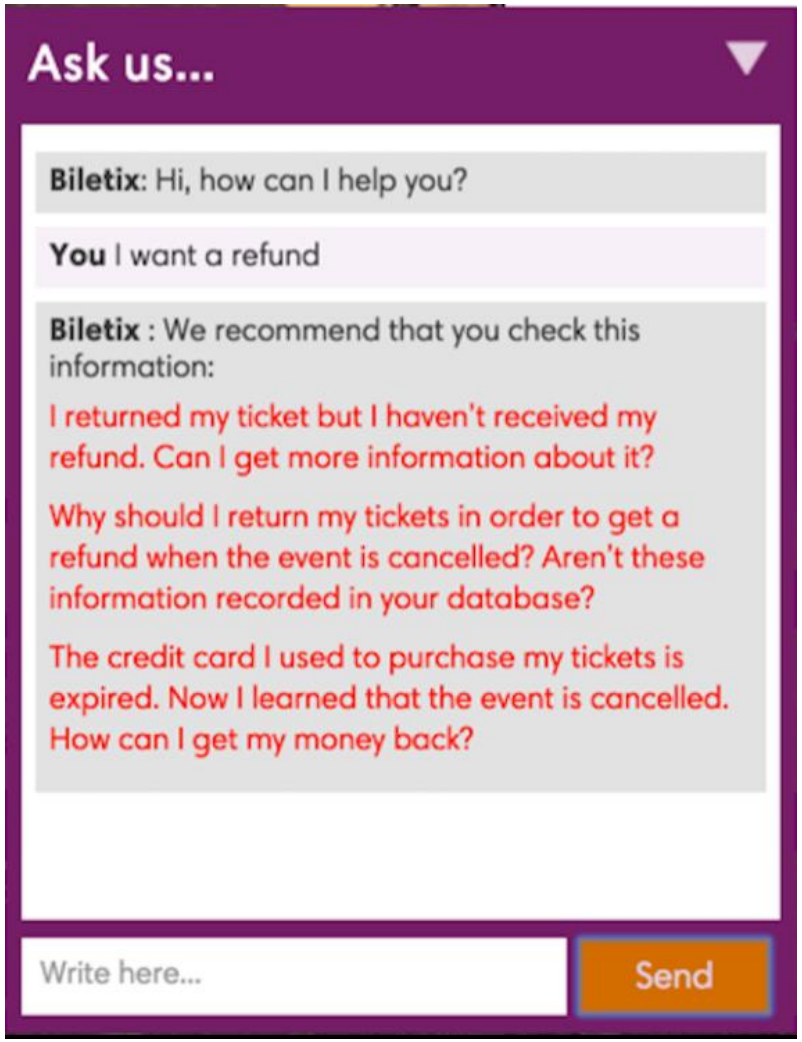
- Large-scale collaborative research platform
- Main focus on ease-of-use, reproducibility of research
- Use of Machine Learning for Narrative interfaces
  - Information Retrieval
  - Natural Language Processing (Chatbots)
- Provide and host in-house solutions and projects



Natural Language Processing

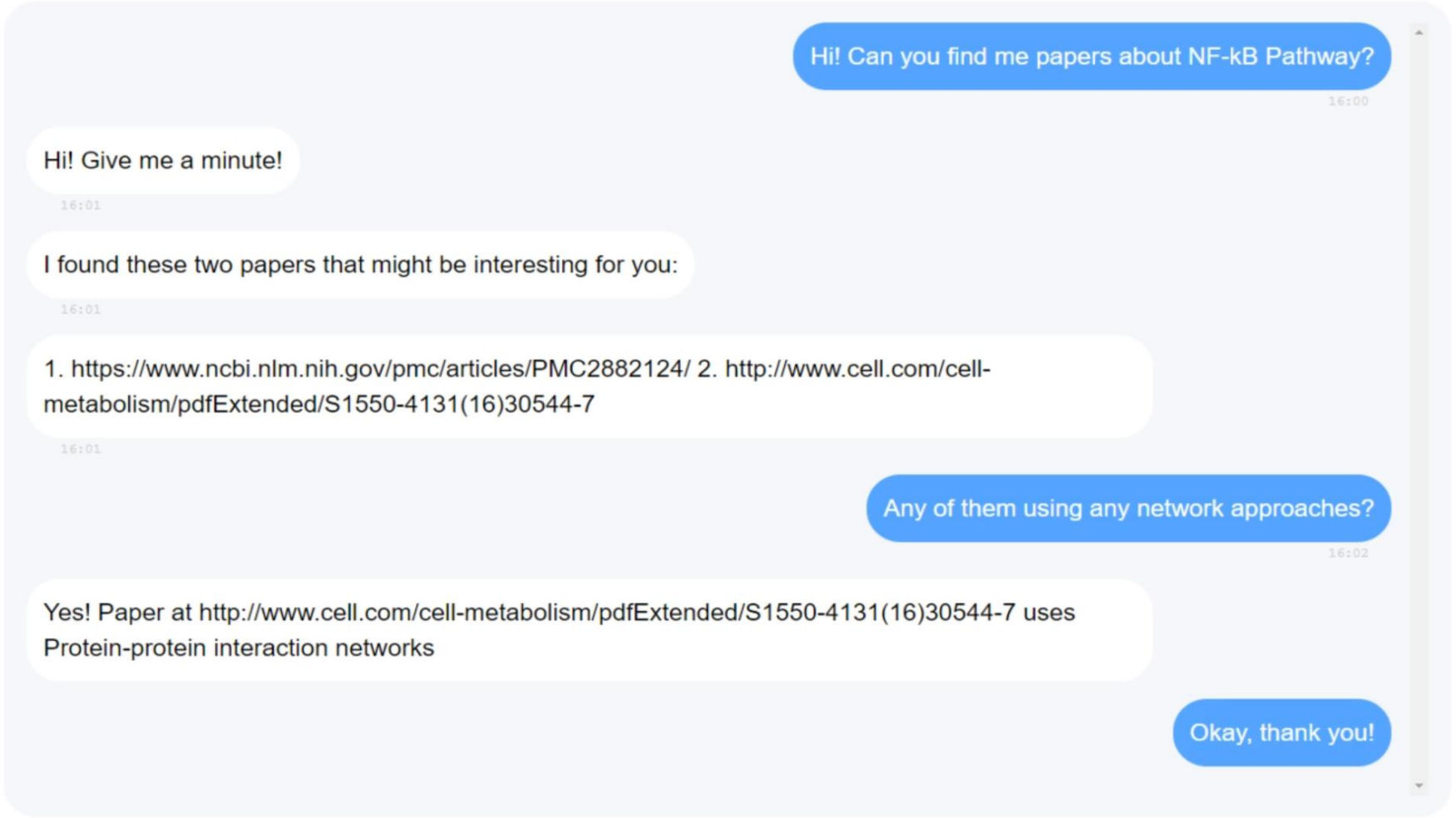
Chat bots – As seen commonly





Natural Language Processing

Chatbots – How we see it



Natural Language Processing

Chatbots and Information Retrieval

- Lower the barriers for junior researchers
- Enhance the way research is done for everyone
- Chatbots as Personal Assistants
- Information Retrieval:
  - Open vs Closed-Domain Retrieval

Natural Language Processing

Open vs Closed-Domain

- Two subgroups of relevant tasks:
  - Open Domain: Global Knowledge-based Information Retrieval
  - Closed Domain: Scan a given text/publication to find an answer to specific query

- Researched separately

- Open-domain for any generic queries and questions

- Closed-domain for training and newcomers

Natural Language Processing

Open Domain based models

- Latest advancements focus on uses of Deep Neural Network models and Logic Graphs

- Different tasks being solved with different DNN models:
  - Machine Translation
  - Topic Identification
  - Information Retrieval

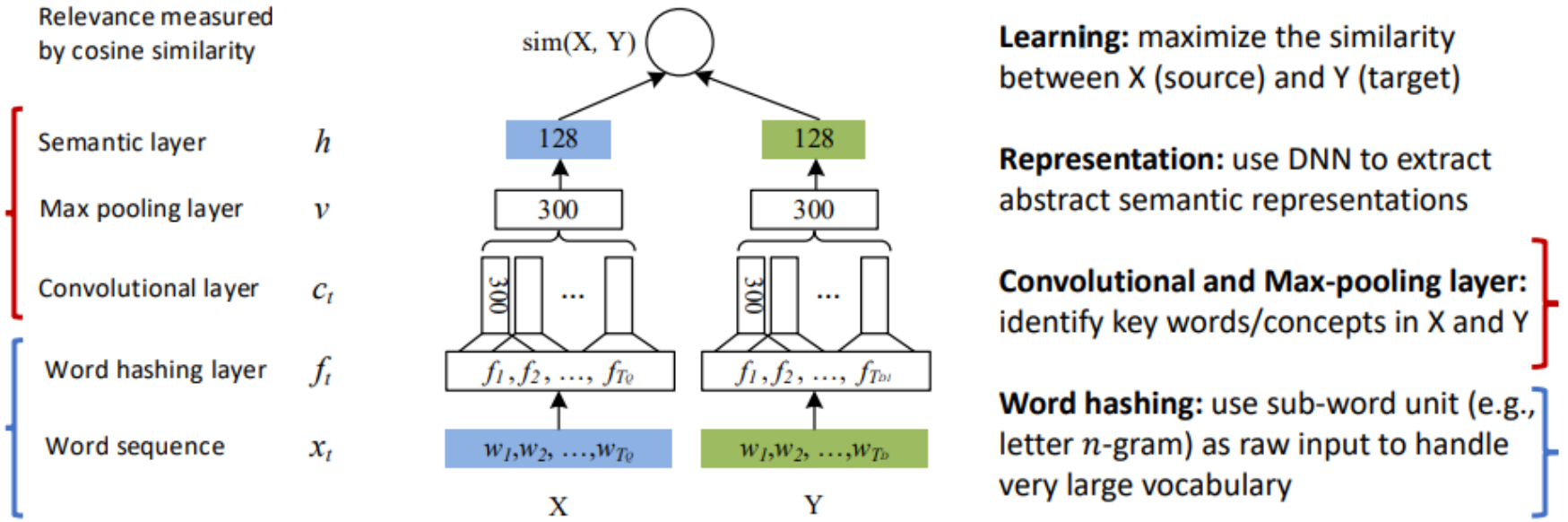
- Our initial interests focus on Information Retrieval and Response Generation

- DSSM as current model that is being investigated
  - Deep Semantic Similarity Measure Model (Shen, 2014)

Natural Language Processing

Open Domain based models

- Cosine Similarity in Semantic Space
- Fits a model to maximize the distance between query and irrelevant documents while minimizing the distance to relevant ones:
  - $\Delta = \text{sim}_{\boldsymbol{\theta}}(X, Y^+) - \text{sim}_{\boldsymbol{\theta}}(X, Y^-)$



Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic representations

Convolutional and Max-pooling layer: identify key words/concepts in X and Y

Word hashing: use sub-word unit (e.g., letter n-gram) as raw input to handle very large vocabulary

1 - https://www.microsoft.com/en-us/research/project/dssm/

Natural Language Processing

Closed Domain based models

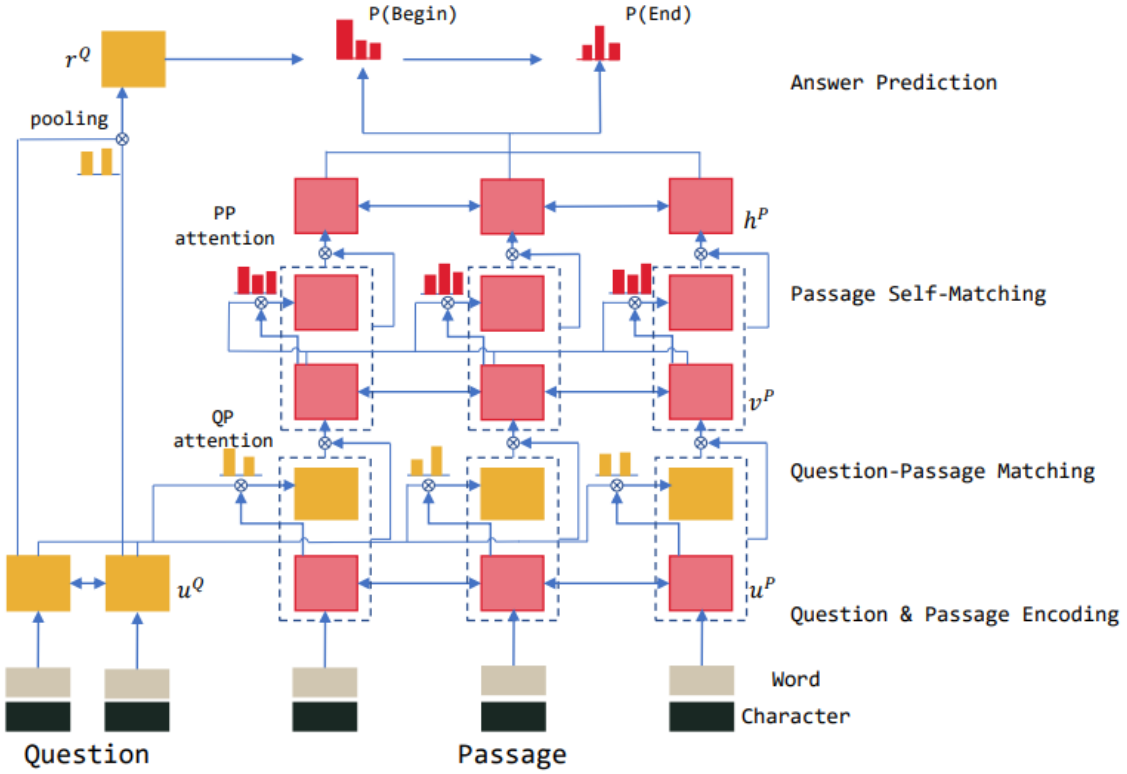
- Gathered interest in last 2-3 years
- Generation of SQuAD Dataset[1] and Google NIPS 2015[2] paper as main sparking points
- More specialized answer generation task
  - Focus on single passage of text/publication
  - Questions do not always hold answers in the given text
- SQuAD Challenge still ongoing
  - Recent models outperforming humans on specific Measures

1 - https://rajpurkar.github.io/SQuAD-explorer/
2 - https://arxiv.org/pdf/1506.03340.pdf

Natural Language Processing

Closed Domain based models

- R-net[1] the most popular and better performing model



Taghi Aliyev, ECRMA Short Talks.

1 - https://www.microsoft.com/en-us/research/wp-content/uploads/2017/05/r-net.pdf

Natural Language Processing

What are the current challenges?

- Current benchmarks focus on news articles or more factual knowledge
- Lack of Data sets and benchmarks for scientific publications
  - Need for compilation of large data sets for use in Deep NLP
  - Currently investigating automated ways of generating labeled Data Sets
- Plenty of room to improve on model performances
- Current models also require a higher level of connections due to the complex nature of scientific publications and facts

Use Cases

How to achieve initial designs?

- Core team concept
- Minimal Viable Platform
  - To initiate the future talks and the iterative process
- Generate an awareness
- Two ongoing and two upcoming projects:
  - Ongoing: KCL + SIDRA on benchmarking of CNV tools. Maastricht University: Target Lipid Identification
  - Upcoming: EBI and Cambridge University

Future Tasks

- Design and Implementation of Minimal Viable Platform
  - Together with Community members
- Machine Generated Benchmark Data Sets
- Testing of Initial NLP Models on the Generated Cases
- Deployment of the Modules together with other CERN Technologies for a more completed prototype

Thanks!

Taghi.aliyev@cern.ch

Twitter: @TaghiAliyev

Backup Slides

Background and Motivation

Example – Target Lipid Exploration

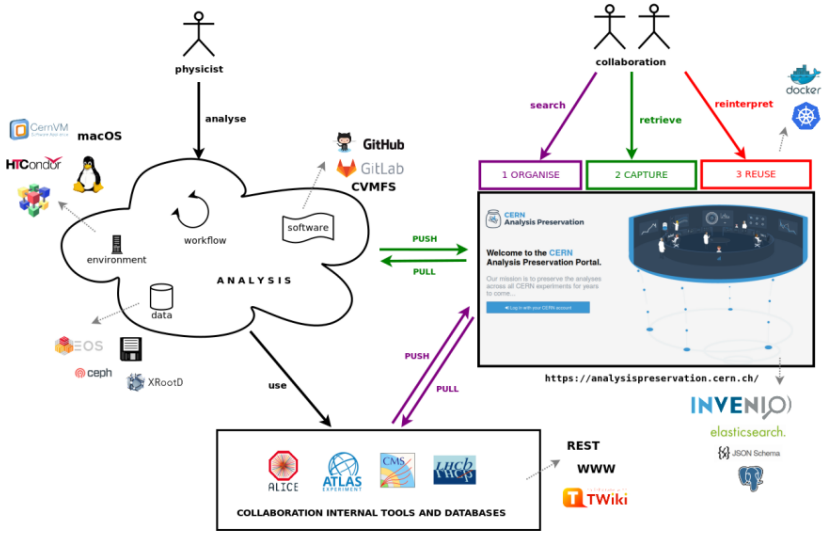
- Project with Maastricht University
  - Goal is to find relevant target lipids for better modeling and explanation of CVD
- Data consists of a lot of missing values
  - Machine Learning approaches used to deal with different multiple issues (Imputation, Overfitting, Classification)
- Currently, no available software to look for relevant publications within a field
  - Requires a lot of manual search and comprehension
  - Similarly for the identified Lipids as well

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

Natural Language Processing

Quick Introduction

- Field of AI focusing on human interaction and understanding human (natural) language

- In last years, approach to NLP Tasks have changed from classical pipelines

- Deep Neural Network Models as the new approach