Automated integration of Genomics Metadata with Sequence-to-Sequence Models

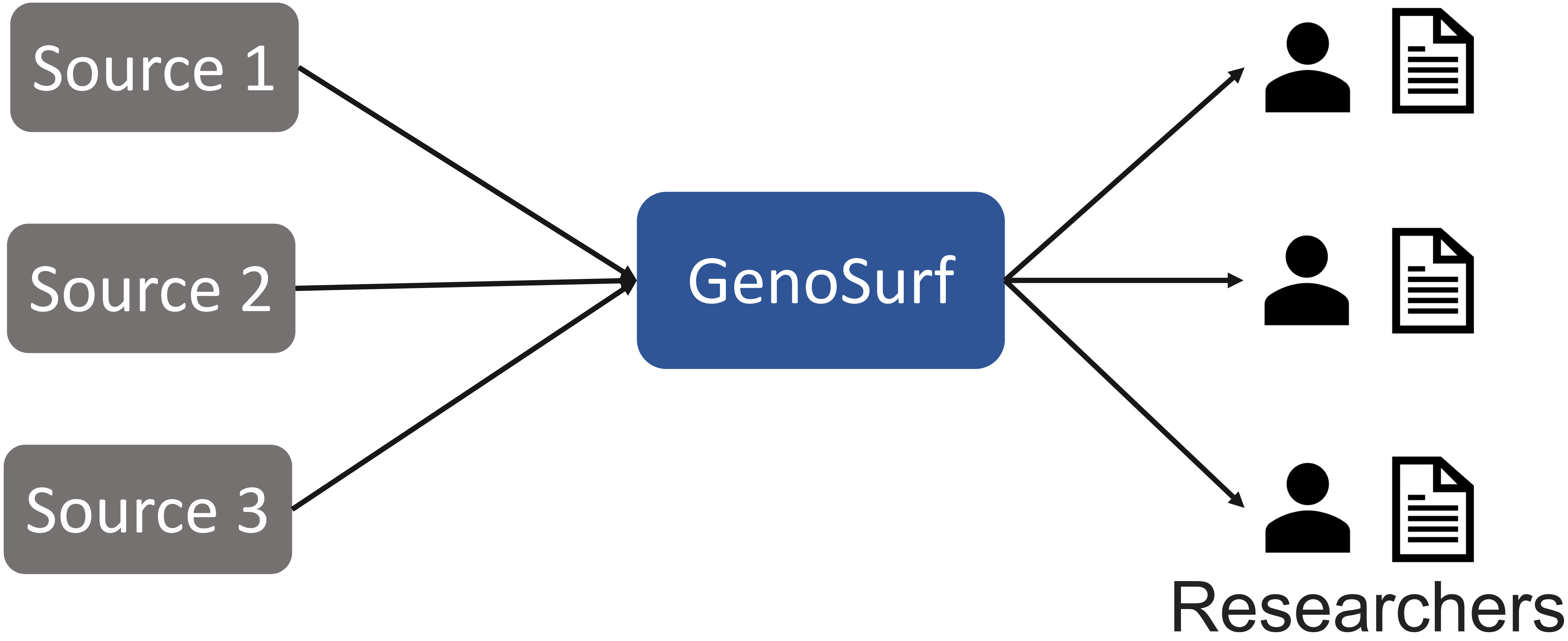
Giuseppe Cannizzaro – Master's student
Thesis submission: April 2020
Adv. Mark Carman, Stefano Ceri

Publication (almost) ready to be submitted: Automated integration of Genomics Metadata with Sequence-to-Sequence Models
G. Cannizzaro & M.Leone, A. Canakoglu, A. Bernasconi, M. Carman @ PKDD 2020

POLITECNICO
MILANO 1863

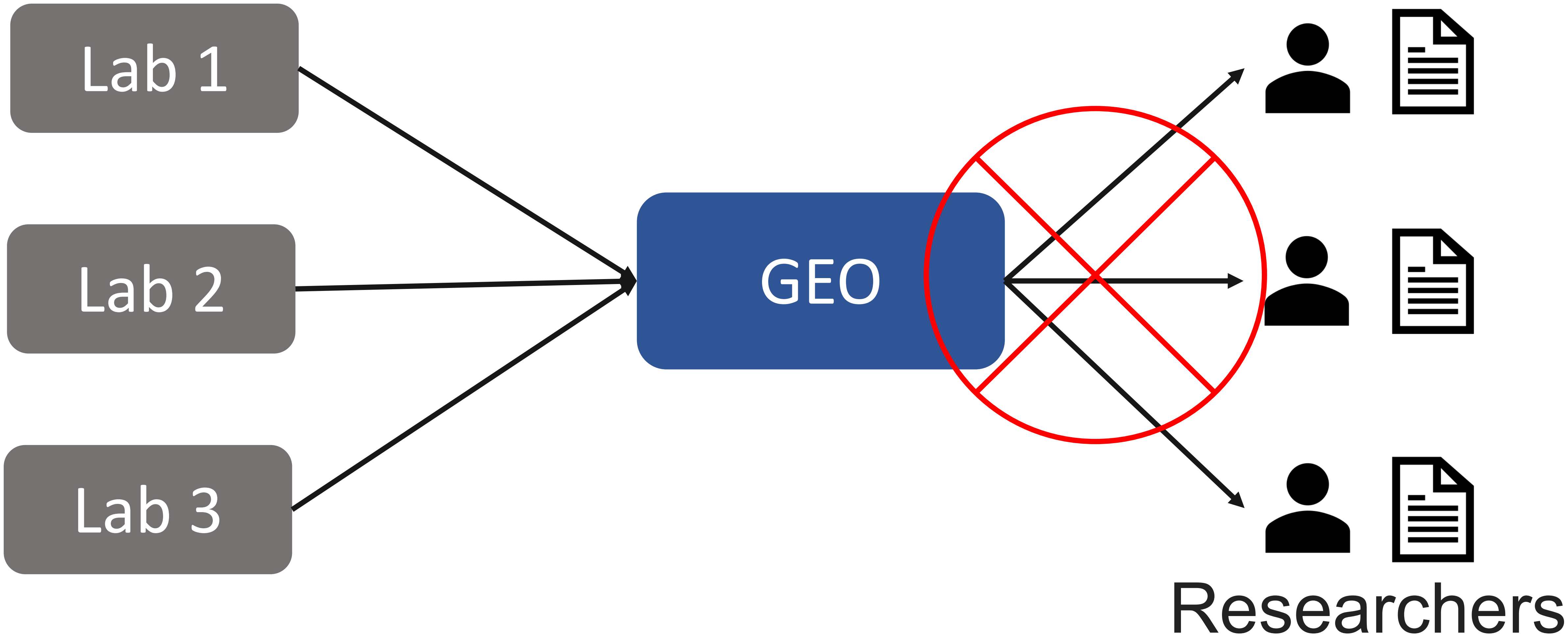
The integration problem

Biologists and bioinformaticians need access to **large structured datasets** to perform queries on biologic metadata, therefore **integration** of repositories is a central task in bioinformatics (GenoSurf objective)



Gene Expression Omnibus

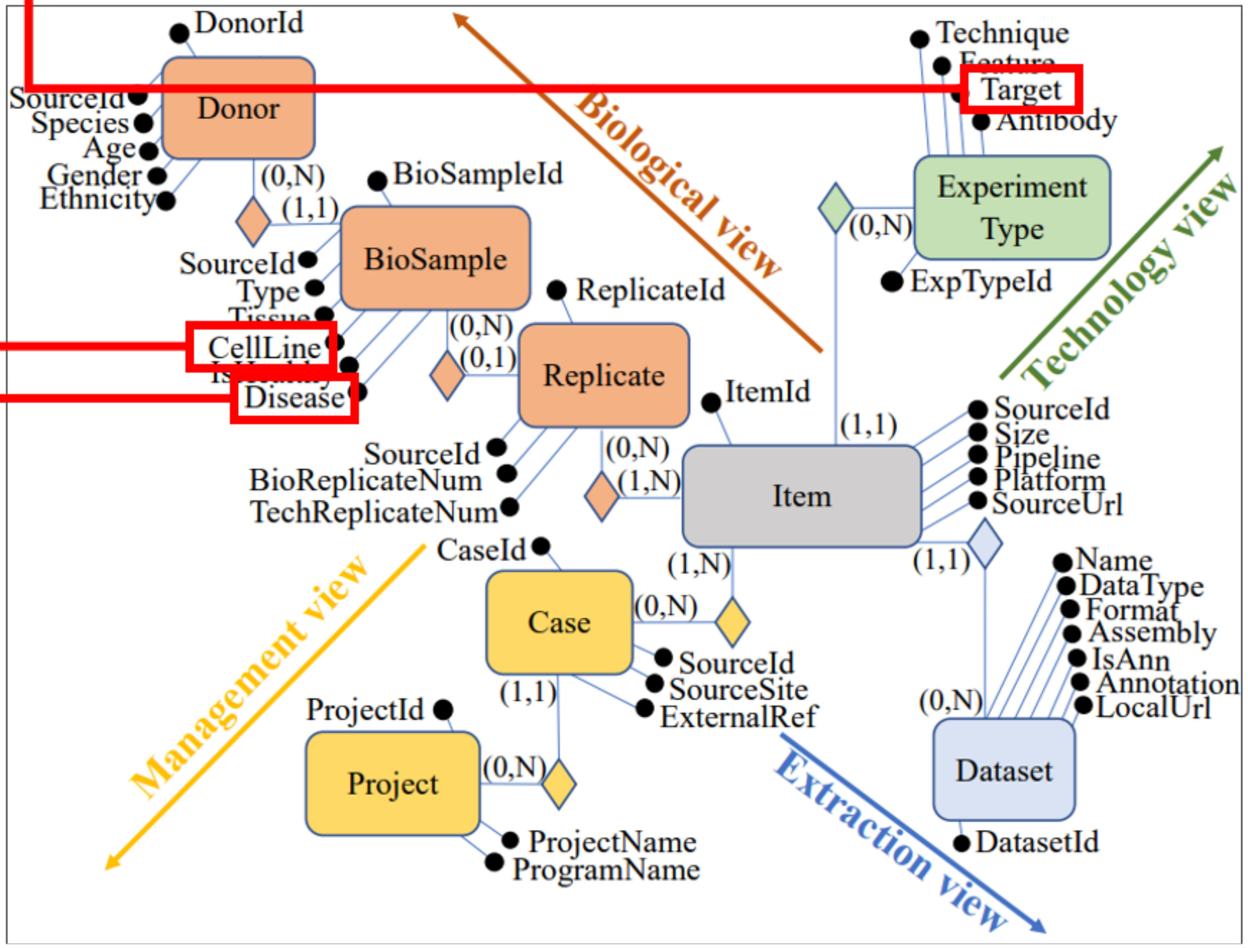
GEO is a very large repository (more than 3 million samples) which objective is the same, but due to **lack of structured metadata** associated to samples, very few types of query can be performed on it



Our task

Automatically extract useful information from plain text metadata

DC_MTB_H3K4me1_rep1 (ChIP-Seq)
Dendritic cells (DCs) were infected with a Mycobacterium tuberculosis (MTB) strain expressing green fluorescent protein (H37Rv) for 18 h at a multiplicity of infection of 1-to-1. Full details can be found in Barreiro et al. (2012). Peripheral blood mononuclear cells (PBMCs) were isolated from buffy coats by Ficoll-Paque centrifugation. Blood monocytes were then purified from PBMCs by positive selection with magnetic CD14 MicroBeads (Miltenyi Biotech). Pure monocytes were cultured for 5 days in RPMI 1640 (Invitrogen) supplemented with 10% heat-inactivated FCS (Dutscher), L-glutamine (Invitrogen), GM-CSF (20 ng/mL; Immunotools), and IL-4 (20 ng/mL; Immunotools). Cell cultures were fed every 2 days with complete medium supplemented with the cytokines previously mentioned. Before infection, we systematically checked the differentiation/activation status of the monocyte-derived DCs by flow cytometry, using antibodies against CD1a, CD14, CD83, and HLA-DR. Only samples presenting the expected phenotype for non-activated DCs – CD1a+, CD14-, CD83-, and HLA-DRlow – were used in downstream experiments.



GEO

GMQL

How?

Regular Expressions

- Needs text patterns
- Impossible to infer information from text
- Multiple matches problem

Named Entity Recognition

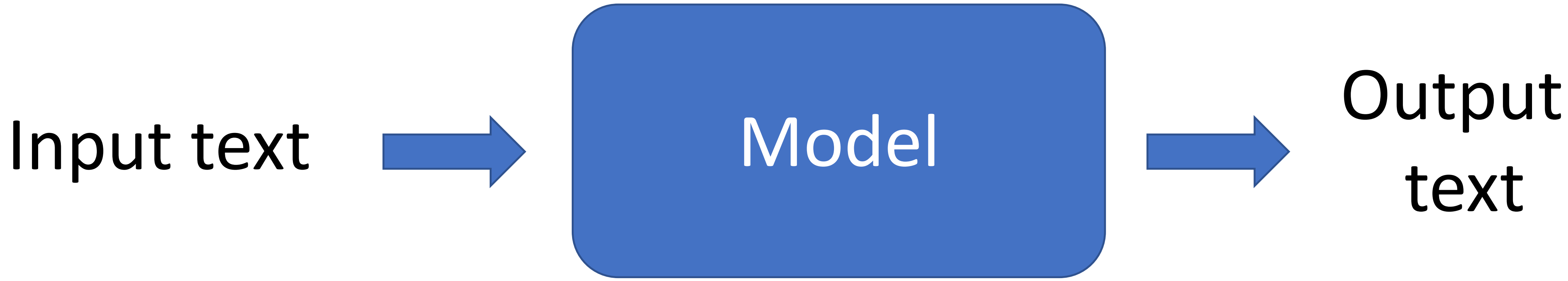
- Requires dataset created with automatic labelling (Regular Expressions)
- Impossible to infer information from text
- Multiple matches problem

Classification

- All possible values must be known apriori
- Can’t handle cases like multi-cells experiments

How?

Sequence-to-Sequence models!



Input text

Model

Output text

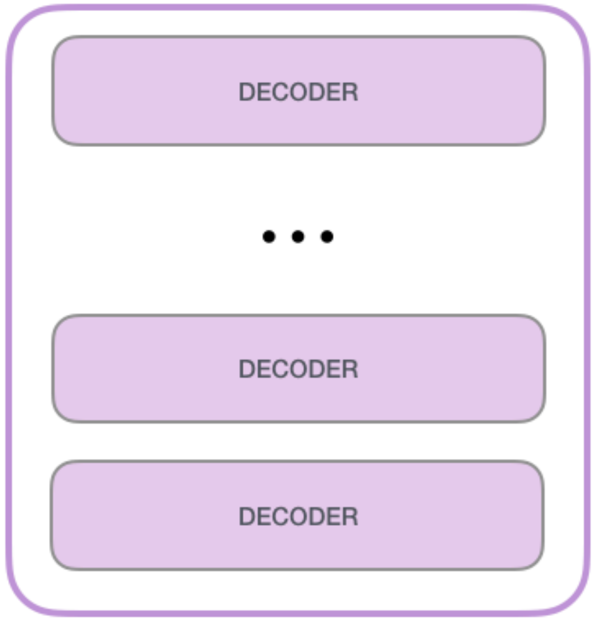
Title: Tcf4_3.nonsh - Description: none - Characteristics: cell line: LS174T - Protocol: Immunoprecipitated chromatin was sheared for a second time for 6 minutes using Covaris sonicator (6 x 16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping) to obtain suitable shorter fragments (75-125 bp). To exclude a shearing bias as a possible source of binding site ultrastructure, partial digestion using DNaseI was used […]

Cell Line: None –
Cell Type: None -
Tissue Type: Retina -
Factor: CTCF -
Disease: retinoblastoma

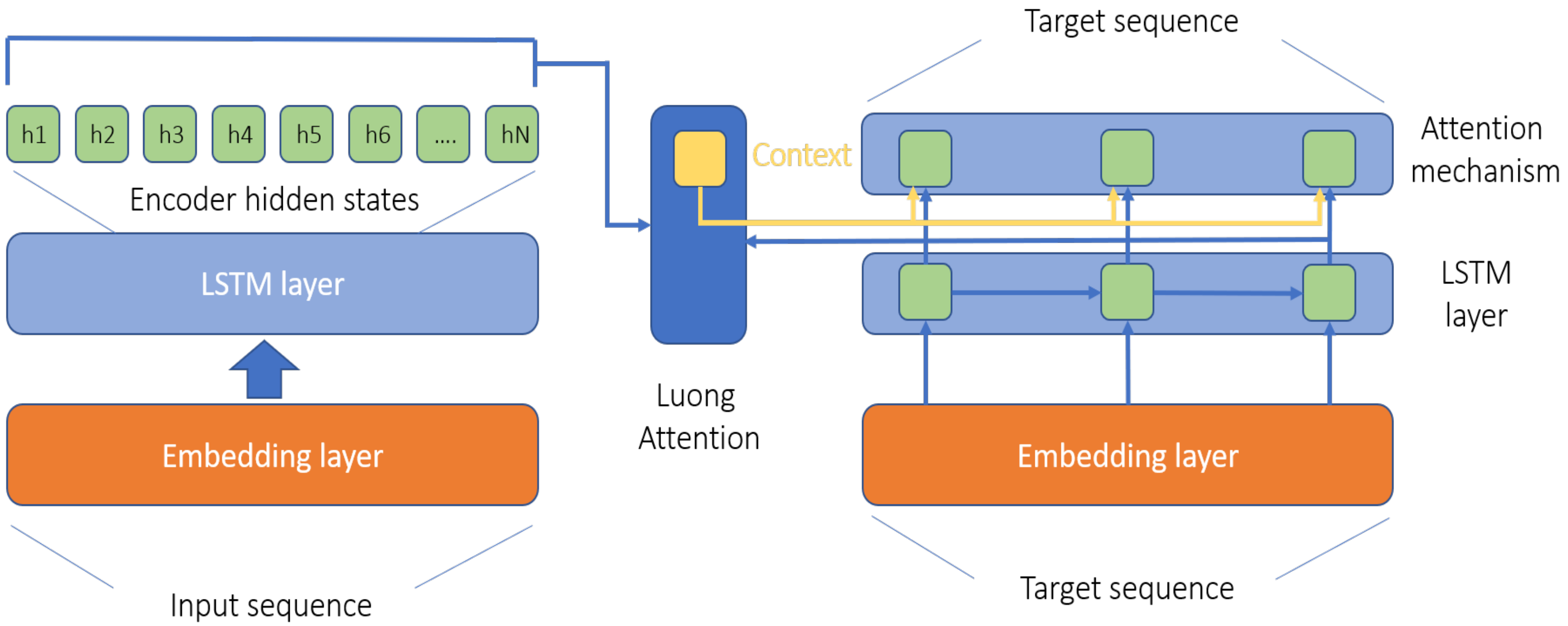
Bad structure

Nice structure

GPT-2



ENCODER - DECODER

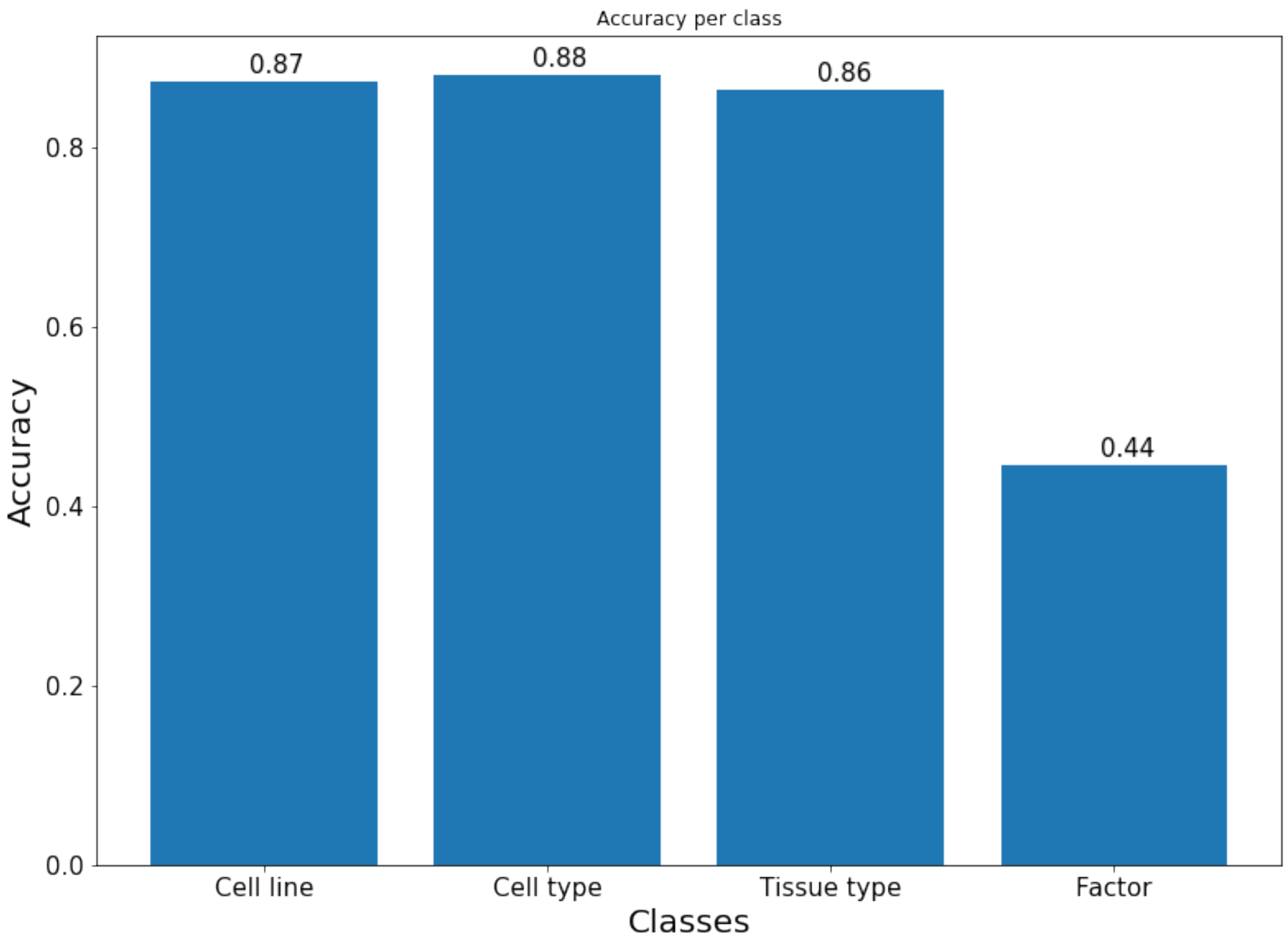


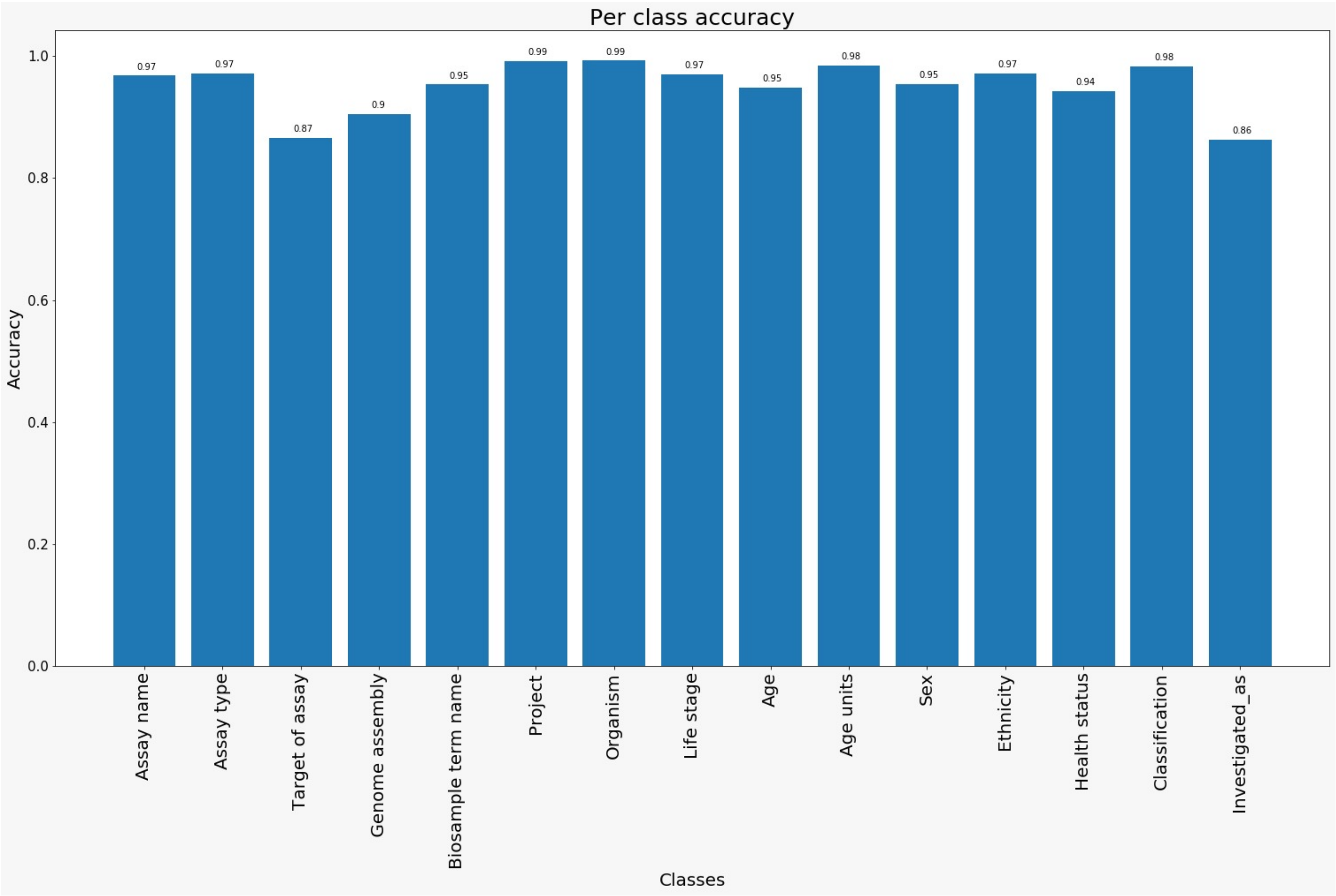
Models tested so far

Some results

Encoder-decoder

OpenAI – GPT2





CISTROME dataset

ENCODE dataset

Future works

Improve Performances

More general as possible

Restructured repositories

Using Machine Learning approaches, it’s important that the system we are going to build is **reliable**, to do that we need to get results that outperform possibly a human checker (promising results)

Future works

Improve Performances

More general as possible

Restructured repositories

Results are promising:

INPUT = HNRNPK ChIP-seq in K562 K562 HNRNPK ChIP-seq in K562
OUTPUT = Assay name: ChIP-seq - **Assay type**: DNA binding - **Target of assay**: HNRNPK - **Genome assembly**: GRCh38,hg19 - **Biosample term name**: K562 - **Project**: ENCODE - **Organism**: Homo sapiens - **Life stage**: adult - **Age**: 53 - **Age units**: year - **Sex**: female - **Ethnicity**: None - **Health status**: chronic myelogenous leukemia (CML) - **Classification**: cell line - **Investigated as**: transcription factor $ <pad>

Accuracy

Deduction

Future works

Improve Performances

More general as possible

Restructured repositories

The models adopted should be able to handle different types of input text and long input texts; to do that **more experiments** are required, on different datasets (possibly with human check for performances) and **powerful models** are required

Future works

Improve Performances

More general as possible

Restructured repositories

The final aim could be the restructuring of existing repositories, creating an infrastructure of databases easy to query and use, to facilitate the work of biologists and bioinformaticians.
GenoSurf could be able to wrap all existing repos and become the Genomic Leader of the world

Thanks for the attention