Text and Data Mining

Patrick Ruch
patrick.ruch@hesge.ch / @patrickruch
HES-SO HEG Genève

5th National Gathering of Medical Librarians
SAMW Bern, August 31 2017

SIB Swiss Institute of Bioinformatics

elixir SWITZERLAND

Hes·so Haute Ecole Spécialisée de Suisse occidentale
University of Applied Sciences Western Switzerland

heg Haute école de gestion de Genève
Geneva School of Business Administration

Genealogy…

How can I relevantly answer the invitation from the SAMS' medical librarians ?

Special greetings to Tamara Morcillo and Isabelle de Kaenel !

Today's agenda

Programme

9.00 am Registration, coffee and opportunity to visit sponsors' stalls

9.20 am **Opening**
Dr. Herman Amstad, Secretary General, SAMS, Bern

9.30 am **Keynote lecture: Open Science and Open Data**
Prof. Matthias Egger, Professor of Epidemiology and Public Health, University of Bern, President of the SNSF Research Council

10.30 am Coffee break and opportunity to visit sponsors' stalls

10.50 am **Research Data Management in Medical Contexts**
Dominic Tate, Head of Library Research Support, Edinburgh University Library, Scotland

11.35 am **Data and Text Mining**
Prof. Patrick Ruch, Head of Information Sciences Department, University of Applied Sciences HEG, Geneva

12.20 pm Lunch break and opportunity to visit sponsors' stalls

1.40 pm **Sponsor's session: Ovid Discovery - A Unified Discovery and Delivery Platform Focused and Specialized in Biomedical Content**
Charlotte Viken, Senior Consultant, Wolters Kluwer Health

2.00 pm **Living Systematic Reviews – from Theory to Implementation**
Dr. Michel Counotte and Dr. Phi Hung Nguyen, Institute of Social and Preventive Medicine, University of Bern

2.45 pm Coffee break and opportunity to visit sponsors' stalls

3.15 pm **Parallel sessions**

Workshop: Mine and Combine – Text Mining Tools Used for Search Term Identification
Jolanda Elmers and Cécile Jaques, Medical University Library, Lausanne

Q&A Session: Research Data Management in Medical Contexts
Dominic Tate, Head of Library Research Support, Edinburgh University Library, Scotland

4.35 pm **Wrapping up**
Gerhard Bissels, Head of Bühlplatz Library,

Count of bi-gram and tri-grams

1. and opportunity	4 *(1.5%)*
2. opportunity to	4 *(1.5%)*
3. to visit	4 *(1.5%)*
4. visit sponsors'	4 *(1.5%)*
5. sponsors' stalls	4 *(1.5%)*
6. head of	4 *(1.5%)*
7. university library	4 *(1.5%)*
8. text mining	3 *(1.1%)*
9. research data	3 *(1.1%)*
10. data management	3 *(1.1%)*

1. and opportunity to	4 *(1.5%)*
2. opportunity to visit	4 *(1.5%)*
3. to visit sponsors'	4 *(1.5%)*
4. visit sponsors' stalls	4 *(1.5%)*
5. research data management	3 *(1.1%)*
6. break and opportunity	3 *(1.1%)*
7. data and text	2 *(0.7%)*
8. and text mining	2 *(0.7%)*
9. university of bern	2 *(0.7%)*

… to support my presentation !

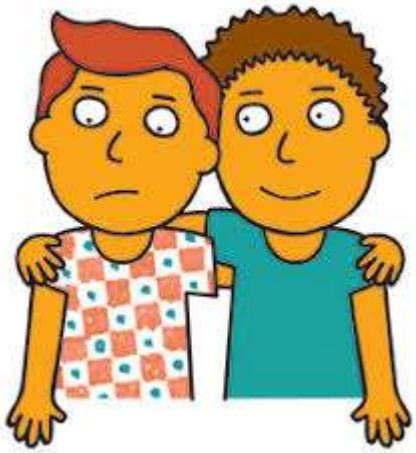
1. and opportunity	4 *(1.5%)*
2. opportunity to	4 *(1.5%)*
3. to visit	4 *(1.5%)*
4. visit sponsors'	4 *(1.5%)*
5. sponsors' stalls	4 *(1.5%)*
6. head of	4 *(1.5%)*
7. university library	4 *(1.5%)*
8. text mining	3 *(1.1%)*
9. research data	3 *(1.1%)*
10. data management	3 *(1.1%)*

1. and opportunity to	4 *(1.5%)*
2. opportunity to visit	4 *(1.5%)*
3. to visit sponsors'	4 *(1.5%)*
4. visit sponsors' stalls	4 *(1.5%)*
5. research data management	3 *(1.1%)*
6. break and opportunity	3 *(1.1%)*
7. data and text	2 *(0.7%)*
8. and text mining	2 *(0.7%)*
9. university of bern	2 *(0.7%)*

Objective of text and data mining

To support you

like TDM helped me today !



How text & data mining can support…

- … medical librarians ?
- … research data and open science ?
  [well covered by other speakers]
- … clinical knowledge ?

History

- Empirical medicine → Observation & Cooking
- Evidence-based medicine → Statistical power
- Personalized health → Deciphering omics

Blind spot ☺

- Empirical medicine → Observation & Cooking 😐
  [Mountains are made of stones]
- Evidence-based medicine → Statistical power

- Personalized health → Deciphering omics

- **Access to EHR evidences (80% narratives)**
  - Overcome publication paywalls
  - Fight silos [researchers, clinicians, institutions…]
  - Empower patients
  - Establish circle of trust and clarify legal basis

Basis of clinical practice…

- Empirical medicine → Observation & Cooking
- Evidence-based medicine → Statistical power
- Personalized health → Deciphering omics
- Access to EHR evidences (80% narratives) !

Evidence-based “holistic” (including phenotypes) personalized health !

Basis of clinical practice…

- Empirical medicine → Observation & Cooking
- Evidence-based medicine → Statistical power

Open Access, Open Science required !

- Personalized health → Deciphering omics
- Access to EHR evidences (80% narratives) !

How text & data mining can support…

- … medical librarians ?
- … research data and open science ?
- … clinical knowledge ?
- … Text and Data Mining Licenses
- … Institutional archives

How text & data mining can support…

- … medical librarians ?
- … research data and open science ?
- … clinical knowledge ?
- … Text and Data Mining Licenses
- … Institutional archives

Operate at national level
→ SwissUniversities-AKOA
cf. SONAR (lead by RERO)
ORCID, growth of FT…

How text & data mining can support…

- … medical librarians ?
- … research data and open science ?
- … clinical knowledge ?

Evidence-based medicine is dependent on (published / accessible) evidences !

FAIR

F.A.I.R

- Findable – indexing strategies
- Accessible – archiving + access rights
- Interoperable – terminologies
- Re-usable – licensing models

How librarian can improve compliance with FAIR principles ?

The solution is…

Recherche Google J'ai de la chance

Le domaine Google.ch est disponible en : Deutsch English Italiano Rumantsch

Why not Google ?

- **Atomic research unit is no more the article → datasets**

- Datasets are multimodal – text search is not sufficient !
  - Sequences
  - Texts
  - Images
  - Spreadsheets
  - […]

- Datasets require semantically-rich meta-data → Curation
- **Access must be monitored, de-identification is a myth !**

Librarians are needed

- To define standards
- To define terminology contents
- To define transcoding tables between terminologies
- To curate datasets (~indexing)

Librarians are needed if…

- They are data science skills
  - Onto-Terminology management *
  - Semantic web technologies *
  - Data management
    - *Databases, e.g. SQL… **
    - *Text processing pipelines, e.g.XML… **
    - *Search engines*
    - *Data Analytics…*
- They have some domain-specific expertise

Data search & analytics

- How far should be go ?
  - *Search engines*
  - *Text and Data Analytics…*

→ *Specialization at MSc level (2018) ?*

?

Overview

- Dataset access → Learning to rank !
- Lifecycle Management of Dataset
  - Primary Data Generation (DMP, SNF Oct 1st 2017)
  - Expert-level curation
  - Storage… archiving…
- Applications for decision-support in oncology

A Machine Learning Pipeline for Enhanced Question Answering over Biomedical Datasets

Swiss Institute of Bioinformatics

elixir SWITZERLAND

From traditional search to dataset search

- Search engine
- Automatic text categorizer (indexing)
- Question-answering, e.g. EAGLi

- Dataset search engines
  - Dataset categorization → Validation → Curation
  - Query expansion for dataset search

EGA: European Genome-Phenome Archive

- Data are stored locally (e.g. Research Libraries, Hospitals, SIB…)*
- Access policy is managed locally (ELSI, IRB…)*
- Meta-data are generated locally*
- Meta-data are exported and stored centrally (SIB, EBI, NIH, …)
- Search is currently possible only on meta-data but not sufficient
- Compatibility with NIH repository (dbGap) → standards*
  - Structuring
  - Transcoding [e.g. ICD-10 or ICD-O3 → MeSH]

*: Research libraries

Query types

1. Disease-based search across scales (phenotypes, MoA, Pathway, Proteins…)
2. Molecular-based search across organisms and scales
3. Molecular data/phenotype associations
4. Behavioural and environmental data

Example

1. *Search for **data** on **neural brain tissue** in **transgenic mice** related to **Huntington's disease***
2. *Search for **gene expression datasets** on **photo transduction** and **regulation of calcium** in **blind D. melanogaster***
3. *Find **data of all types** on the **regulation of DNA repair** related to the **estrogen signaling pathway** in **breast cancer patients** across all databases*
4. *Search for **protein aggregation** and **gene expression** data regarding **aging** across all databases*

Background

- Need for dataset retrieval engines/QA applied to datasets
  - Text → PubMed
  - Open data movement
  - Production of data in public and private sectors

- Text search vs. dataset search
  - Modality
  - Heterogeneity
  - Indexing

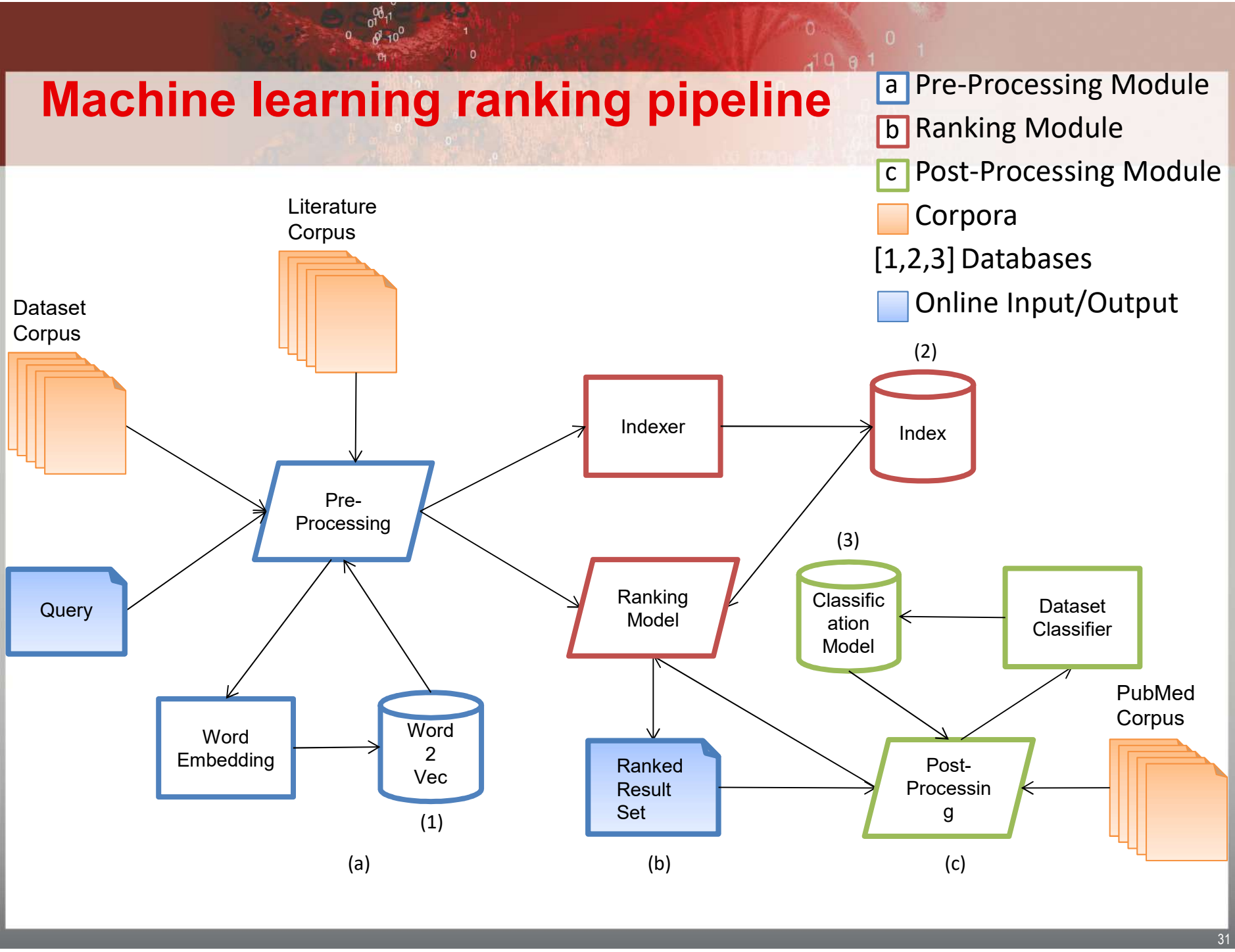
Specifics

- Query size
  - Lack of context
  - … but N terms ~ 5-10 !
- Query constraints
  - QA → Type of dataset
- Dataset formalization and search benchmark
  - Variety of formats
  - Lack of extensive gold standard queries

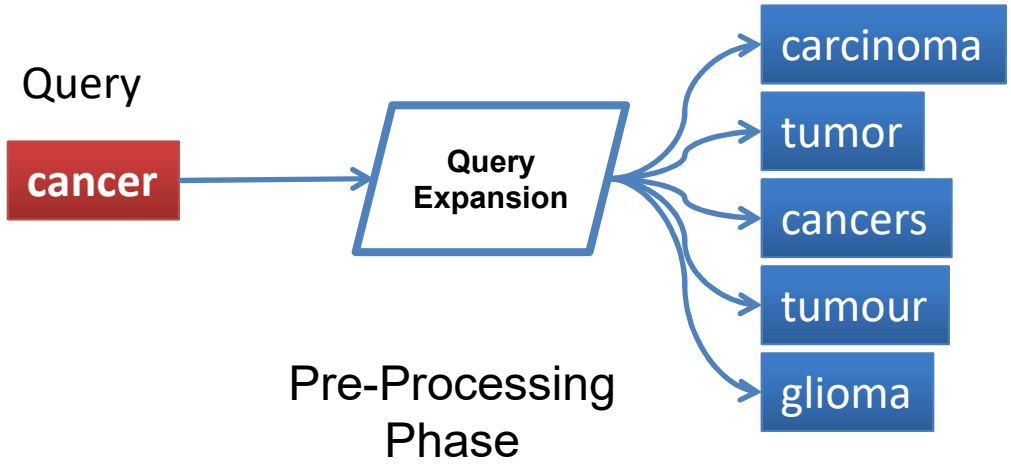
Scientific objective

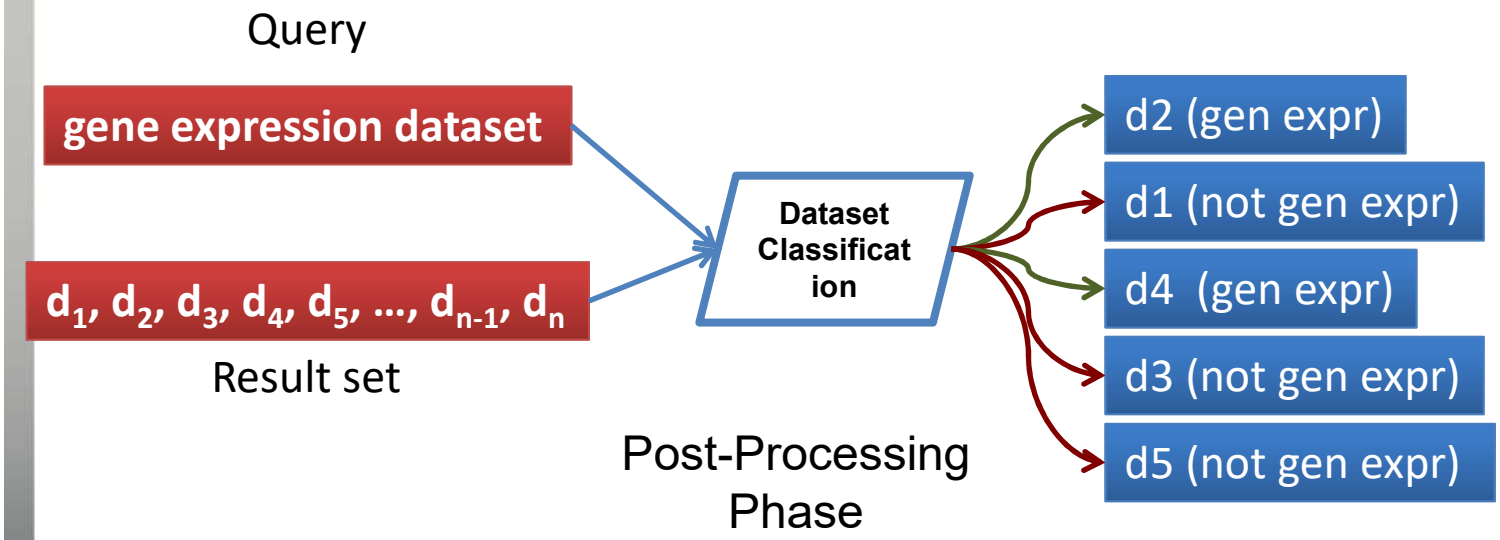
- To accelerate development of search strategies for biomedical datasets
  - Go beyond utilization of the metadata
    - *Assign meta-data automatically*
    - *Search without meta-data*
- To explore machine learning methods to enhance search
  - Increase query context
  - Constraint results to query context

Machine learning ranking pipeline



Machine learning ranking pipeline





Original rank	Final rank
2	1
1	2
4	3
3	4
5	5

Query Expansion and dataset categorization

Swiss Institute of Bioinformatics

elixir SWITZERLAND

Access to research data

- Query expansion – like PubMed
- Dataset indexing – like MeSH indexing

Proposal

- Allow 360 search
- Use semantically-rich contents (Institutional Archives) to bridge data contents
- Provide interactive data curation mechanisms

Backgroung: Doc2Vec

- Derive a semantic algebraic model on top of textual features
  - Derived from a Bag of Word representation (assume independence of words)
  - Generative model to recover from (too strong) independence
    - *Distributed Bag of Word*
    - *Skip-gram model*
  - Parametric model (must be tuned)

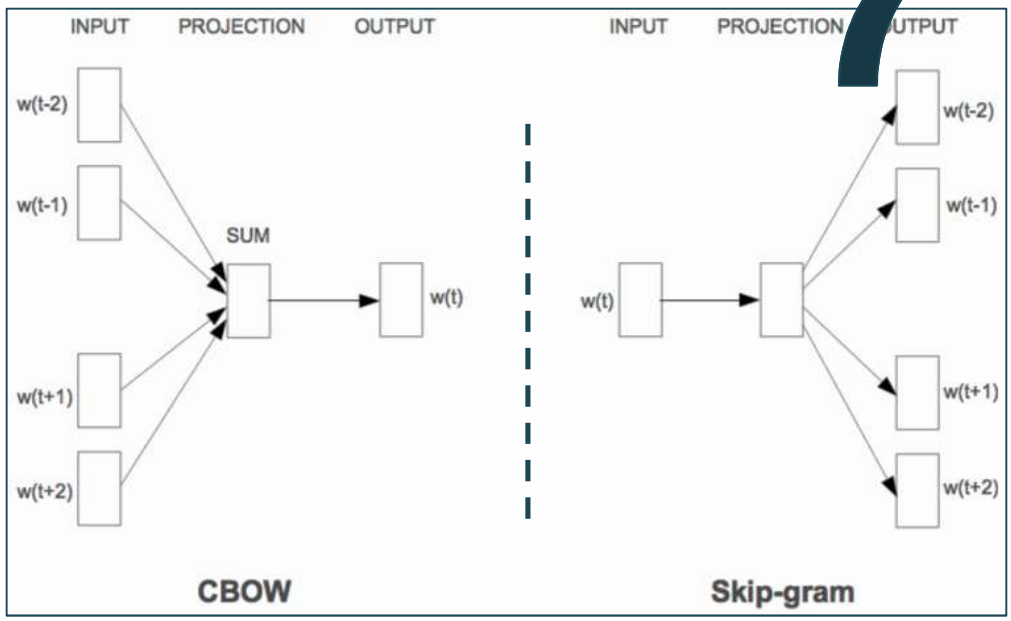
- Example:

  v(Paris + France)… v(London, UK) → Implicit representation of “capital city”

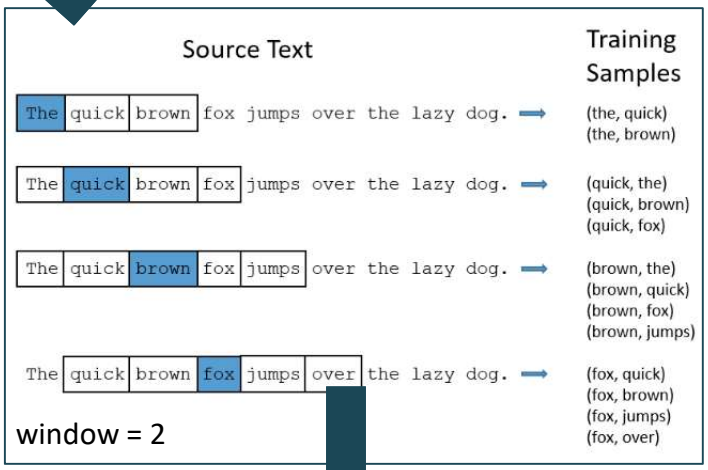
  v(Paris + France) – v(Paris + Italy) → “Rome”

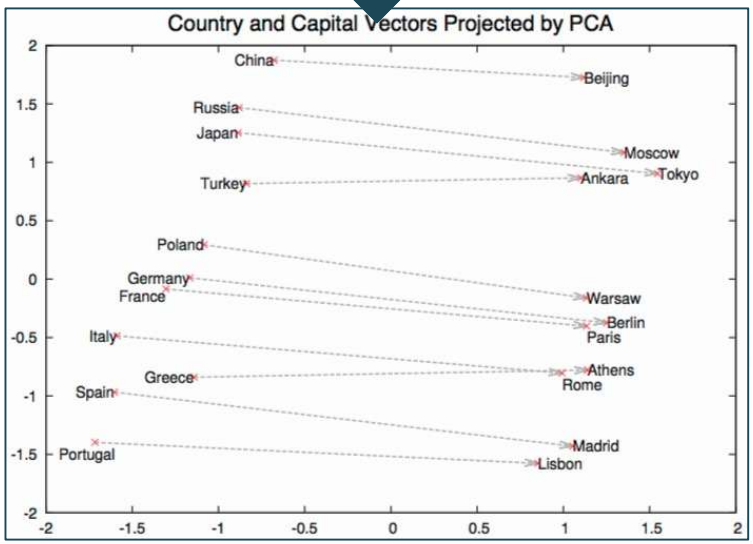
Query expansion – Word embedding

Word embedding



Skip-gram example





Model parameters

- Window size (training sliding window)
- Vector size (embedding space dimension)

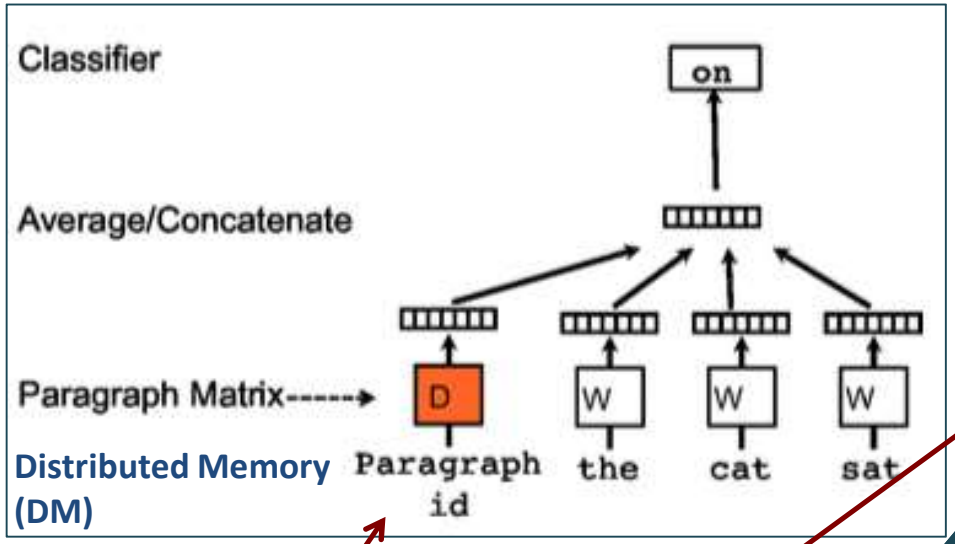
Input

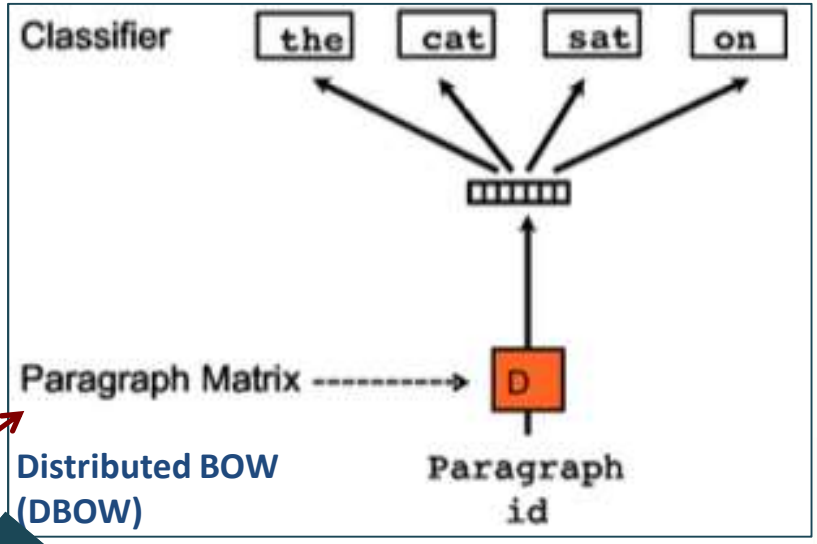
- Medline
- Others: local collection, CT

*Mikolov et al., Efficient Estimation of Word Representations in Vector Space (2013)

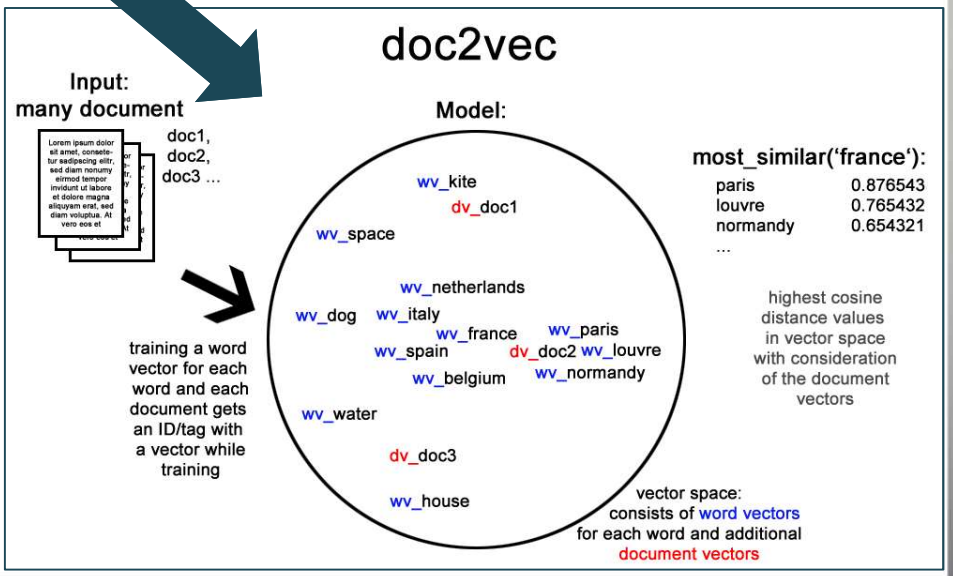
Query expansion – Text embedding

Doc embedding





Doc/Paragraph/Sentence provided as (part of) the word context



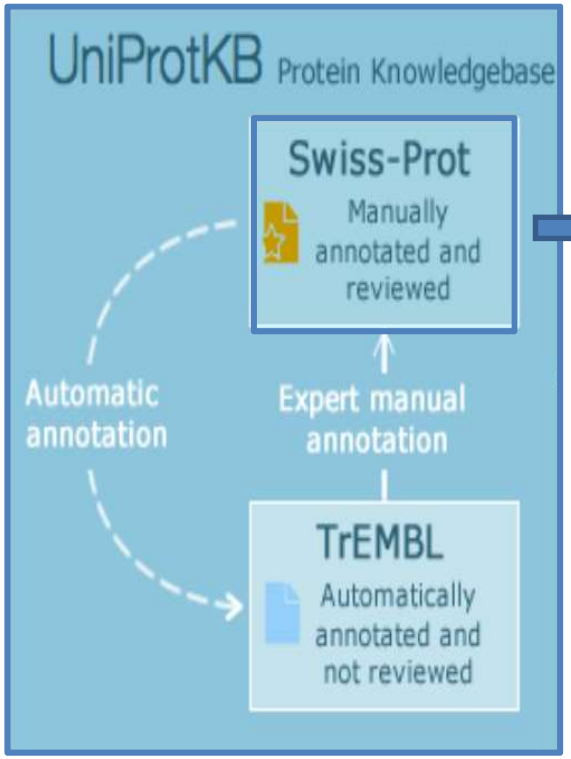
*Le and Mikolov, Distributed Representations of Sentences and Documents (2014)

Query expansion – Examples

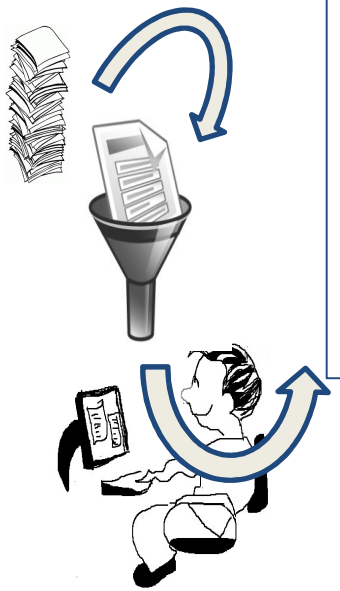
Term	Medline		PMC		bioCADDIE	
	expansion	score	expansion	score	expansion	score
cancer	breast	0.889	carcinoma	0.674	carcinoma	0.737
	cancers	0.855	tumor	0.616	cancers	0.720
	prostate	0.801	cancers	0.585	adenocarcinoma	0.669
	colorectal	0.794	tumour	0.583	malignancies	0.626
	lymphoma	0.621	glioma	0.559	tumor	0.779
human	mouse	0.756	mammalian	0.582	bovine	0.553
	mammalian	0.711	murine	0.441	porcine	0.542
	also	0.661	rat	0.428	murine	0.526
	humans	0.661	vertebrate	0.417	mouse	0.518
	murine	0.656	preeclamptic	0.400	humans	0.486
repair	damage	0.794	repairthe	0.597	closure	0.515
	excision	0.764	replication	0.570	metabolism	0.510
	double-strand	0.727	ssbr	0.543	formation	0.509
	nucleotide-excision	0.723	repairing	0.540	grafting	0.504
	damaged	0.717	damage	0.516	implantation	0.502

Dataset classification – UniProtKB

- A central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.



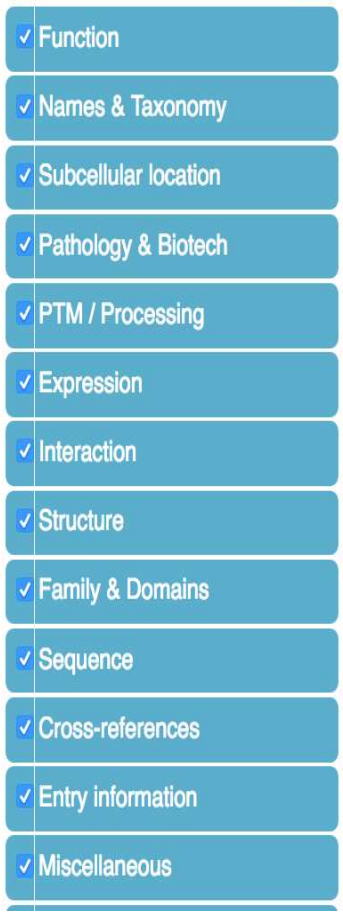
Source: http://www.uniprot.org/help/about



UniProtKB Entry

Entry annotation based on a representative literature subset to provide a complete overview of available information about a protein

Annotations organized in topics in Entry view



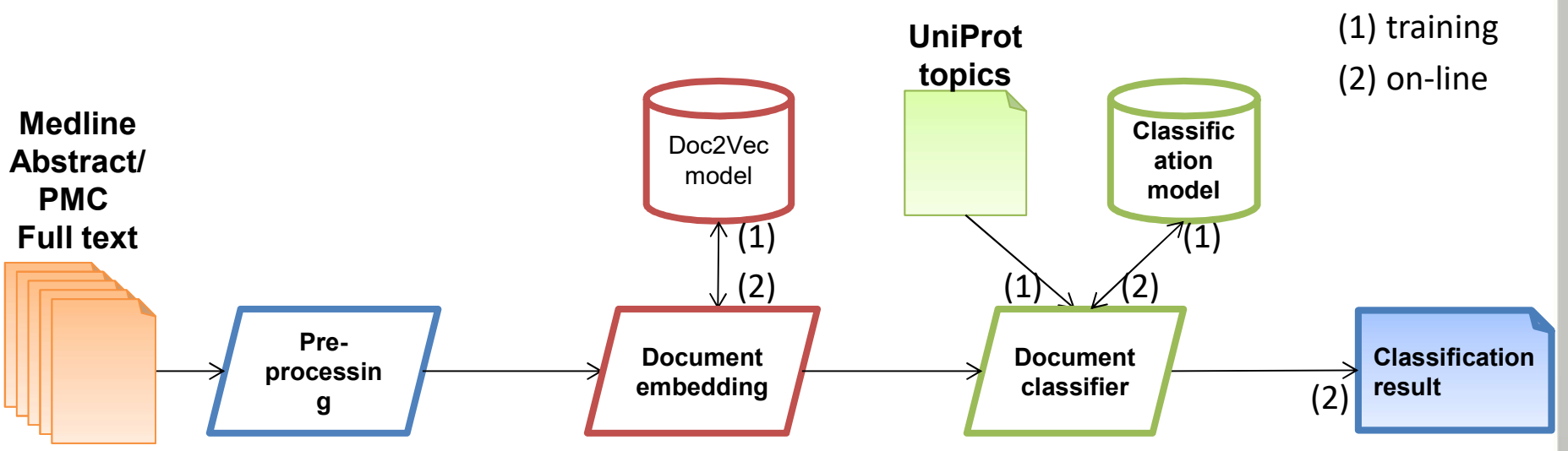
Dataset classification – Training

- Corpus → 100k abstracts (UniProt manually curated publications)

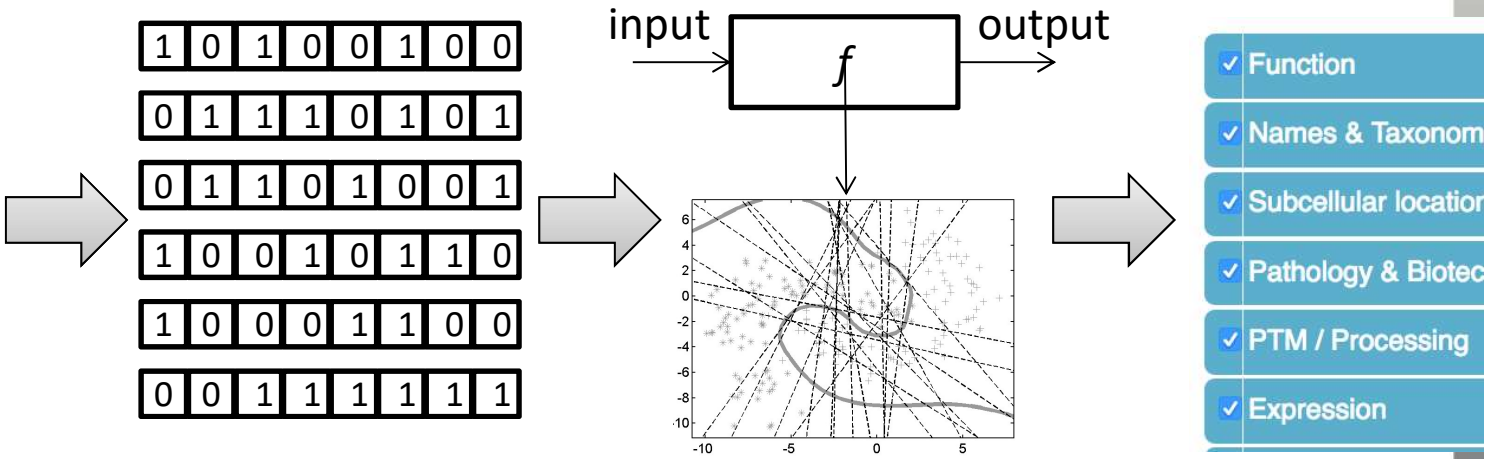
Train			Test		
Category	PMID		Category	PMID	
	#	%		#	%
Names	529	1	Names	9	1
Family & Domains	1595	2	Family & Domains	11	1
Miscellaneous	8152	8	PTM/processing	90	9
PTM/processing	8506	9	Miscellaneous	95	10
Structure	9580	10	Structure	100	10
Subcellular location	13754	14	Interaction	146	15
Interaction	14619	15	Subcellular location	148	15
Pathology & Biotech	15639	16	Pathology & Biotech	156	16
Expression	16456	16	Expression	162	16
Function	34753	35	Function	362	36
Sequences	55696	56	Sequences	570	57

- Training set → 99k abstracts
  - Validation set → 5k abstracts
- Test set → 1k abstracts

Dataset classification pipeline



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam a augue nunc. Aenean iaculis arcu id mi mollis malesuada ut id enim. Maecenas feugiat orci a fringilla laoreet



Dataset classification – Precision

- Expected list of UniProt categories associated with an abstract

Model	Precision	Recall	F-score
Naïve Bayes	0.7474	0.6329	**0.6618**
Random forest	0.8032	0.6642	**0.7015**
kNN	0.7394	0.6830	**0.7045**
Logistic regression	0.8110	0.7191	**0.7521**
MLP	0.8292	0.7749	**0.7980**

Baseline model: Naïve Bayes

Impact on search effectiveness

Swiss Institute of Bioinformatics

elixir SWITZERLAND

Pipeline assessment – bioCADDIE 2016 Challenge

- bioCADDIE Corpus
  - 800k datasets

Repository	Datasets		Repository	Datasets	
	#	%		#	%
ClinicalTrials	192500	24.257%	phenoDisco	429	0.054%
BioProject	155850	19.638%	NursaDatasets	389	0.049%
PDB	113493	14.301%	MPD	235	0.030%
GEO	105033	13.235%	PeptideAtlas	76	0.010%
Dryad	67455	8.500%	PhysioBank	70	0.009%
ArrayExpress	60881	7.672%	CIA	63	0.008%
Dataverse	60303	7.599%	CTN	46	0.006%
NeuroMorpho	34082	4.295%	OpenfMRI	36	0.005%
Gemma	2285	0.288%	CVRG	29	0.004%
ProteomeXchange	1716	0.216%	YPED	21	0.003%

- Query benchmarking
  - Train: 6 queries; Test: 15 queries

Complex queries !

- *"Search for gene expression and genetic deletion data that mention CD69 in memory augmentation studies across all databases"*

- *"Find data of all types on the regulation of DNA repair related to the estrogen signaling pathway in breast cancer patients treated with clopidogrel across all databases"*

Dataset formalization – DATS

<DOC>
<DOCNO>215676</DOCNO> <TITLE>VGlut-F-800286</TITLE> <REPOSITORY>neuromorpho_030116</REPOSITORY>
<METADATA>


```
{
  "dataItem": {
    "dataTypes": ["dataset", "organism", "anatomicalPart", "treatment", "cell", "studyGroup", "dimension", "dataRepository", "organization"]
  },
  "studyGroup": {
    "name": "Control"
  },
  "anatomicalPart": {
    "name": ["Left Antennal Lobe", "Not reported"]
  },
  "dataRepository": {
    "abbreviation": "NeuroMorpho",
    "homePage": "http://neuromorpho.org/",
    "name": "NeuroMorpho.Org",
    "ID": "SCR:002145"
  },
  "dataset": {
    "downloadURL": "http://neuromorpho.org/neuron_info.jsp?neuron_name=VGlut-F-800286",
    "note": "Cell types and Brain regions were assigned with a <a href=\"techDocFlyData.jsp?code=1\">heuristic process</a> based on available metadata. This dataset was processed with a <a href=\"techDocFlyData.jsp?code=2\">streamlined automated variant</a> of the standardization procedure, additional details of which are published <a href=\http://www.ncbi.nlm.nih.gov/pubmed/?term=25576225\ target=\"_blank\">here</a>. Digital reconstruction used a <a href=\"http://www.ncbi.nlm.nih.gov/pubmed/?term=23028271\" target=\"_blank\">custom method</a> after image segmentation by Amira.",
    "ID": "27187",
    "title": "VGlut-F-800286"
  },
  "cell": {
    "name": ["Principal cell", "Glutamatergic neuron", "day8 Born"]
  },
  "treatment": {
    "title": "Green fluorescent protein (GFP)"
  },
  "organization": {
    "abbreviation": "GMU",
    "homePage": "http://www.gmu.edu/"
```

Dataset retrieval – Best results

Group	infAP	P@10	NDCG@10 I	nfNDCG	P@10 (+partial)
SIBTextMin	**0.3664**	0.3467	0.6271	0.4258	0.7533
Elsevier	0.3283	**0.4267**	**0.6861**	0.4368	**0.8267**
UIUC GSIS	0.3228	0.2867	0.5569	0.4502	0.7133
OHSU	0.3193	0.3333	0.6122	0.4454	0.7600
UCSD	0.3169	0.3333	0.5877	**0.5132**	0.7600
Emory	0.2818	0.2667	0.5538	0.4241	0.7200
HiTSZ-ICRC	0.2576	0.2800	0.5472	0.3850	0.7000
BioMelb	0.2568	0.3333	0.6325	0.4017	0.7733
Mayo	0.1628	0.2600	0.5735	0.3933	0.7467
IAII_PUT	0.0876	0.1600	0.4265	0.3580	0.5333

+partial → partial answers are relevant
-partial → partial answers are not relevant
partial answer → does not contain all key query concepts (but more than 50%)

Dataset retrieval – Relative results

	Stats	infAP	infNDCG	P@10 (+partial)	NDCG@10	P@10 (-partial)	UIR
SIB Text Mining	**rank**	**1/10**	**5/10**	**5/10**	**3/10**	**2/10**	**2/10**
	score	0.3664	0.4258	0.7533	0.6271	0.3467	0.51
All participants	median	0.2994	0.4250	0.7500	0.5806	0.3100	0.13
	min	0.0876	0.3580	0.5333	0.4265	0.1600	-1.00
	1st quartile	0.2570	0.3954	0.7150	0.5546	0.2700	-0.43
	3rd quartile	0.3219	0.4433	0.7600	0.6234	0.3333	0.40
	max	0.3664	0.5132	0.8267	0.6861	0.4267	0.82

UIR → Unanimous Improvement Ratio

*Amigó et al., Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. (2011)

Query expansion – Corpus comparison

- Retrieval performance for different collections
  - word2vec training corpus
  - Baseline results use no query expansion

Collection	infAP	infNDCG	P@10 (+partial)
- (baseline)	0.3557	0.4235	0.7267
bioCADDIE	0.3545	0.4243	0.7178
PMC	0.3571	0.4216	0.7178
Medline	**0.3704**	**0.4377**	**0.7511**

K parameter

- bioCADDIE: 20
- PMC: 22
- Medline: 25

Performance improvement

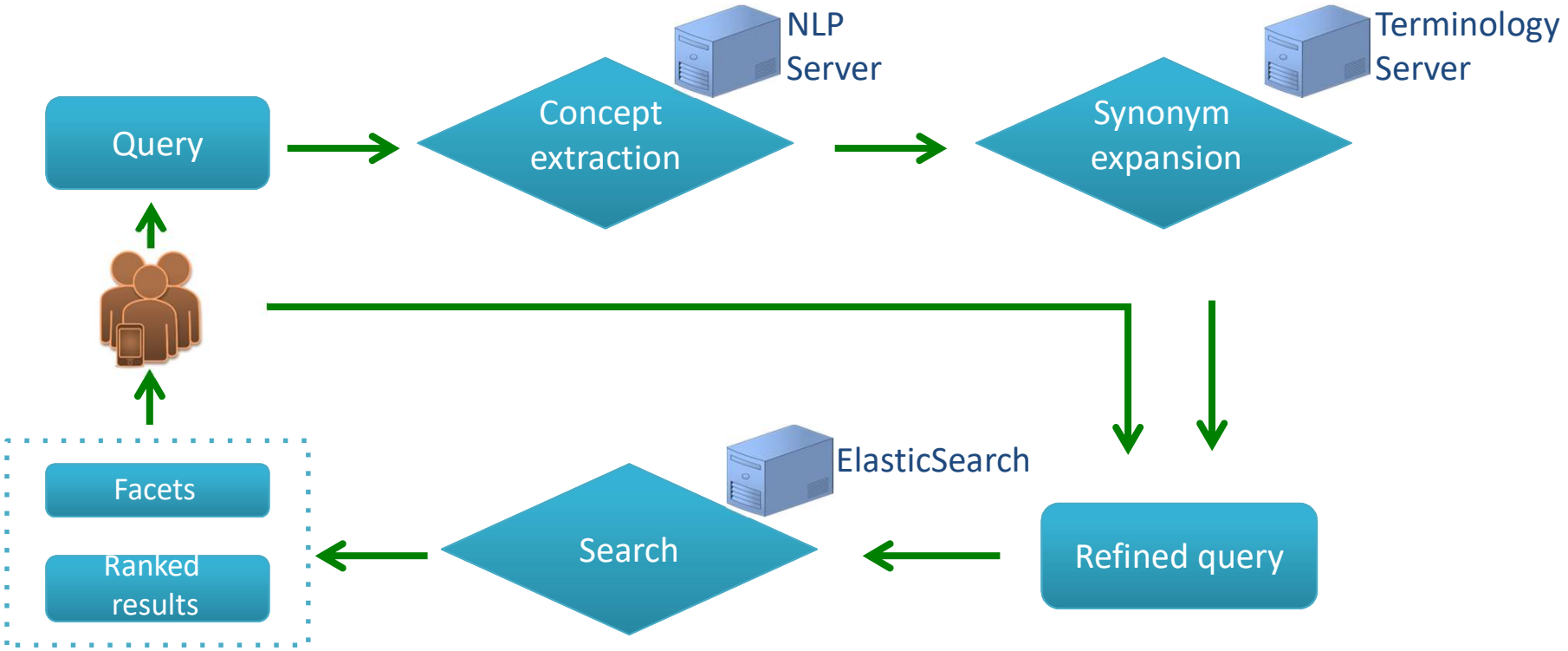
- infAP: +4.1%
- infNDCG: +3.4%
- P@10: +3.4%

Integration @ NIBR

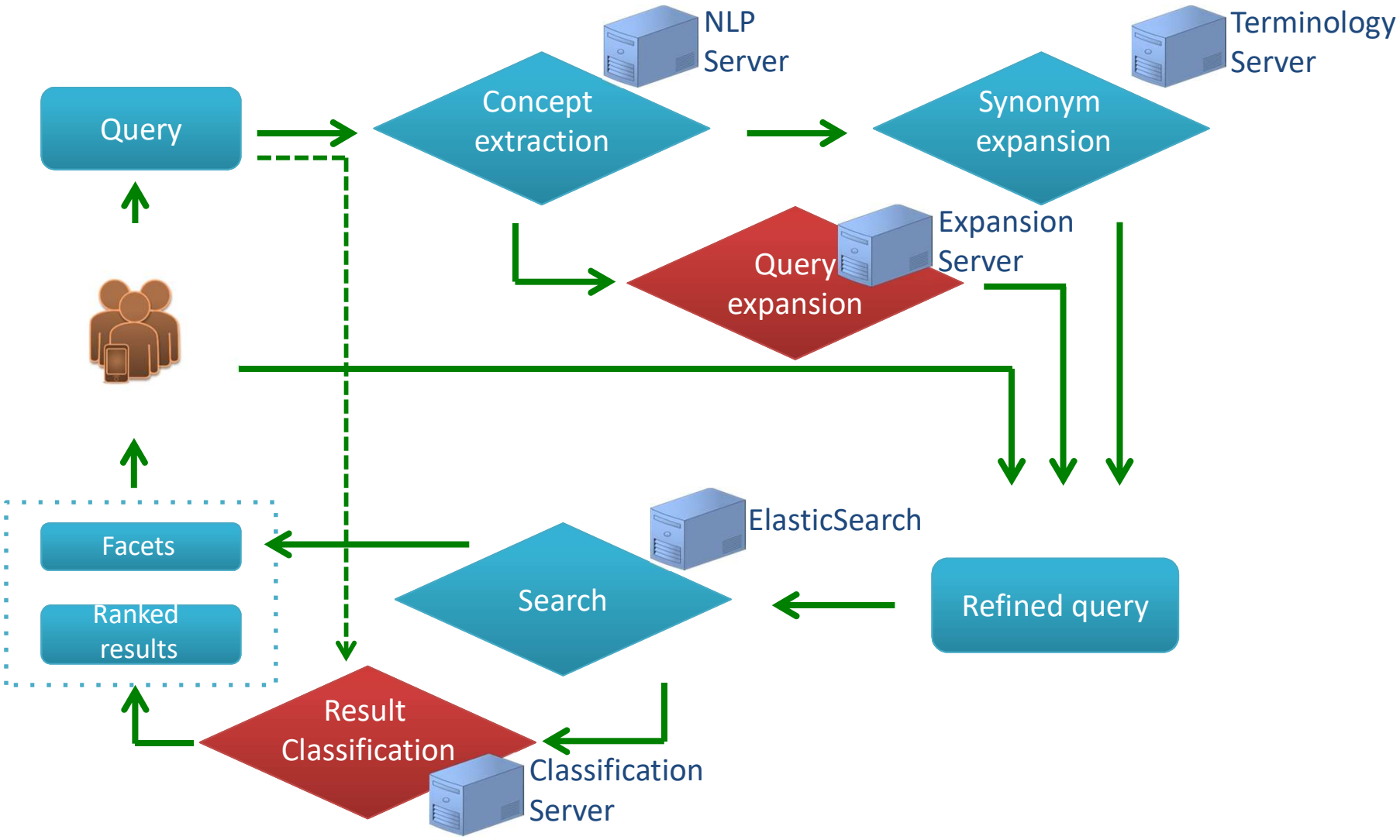
Swiss Institute of Bioinformatics

elixir SWITZERLAND

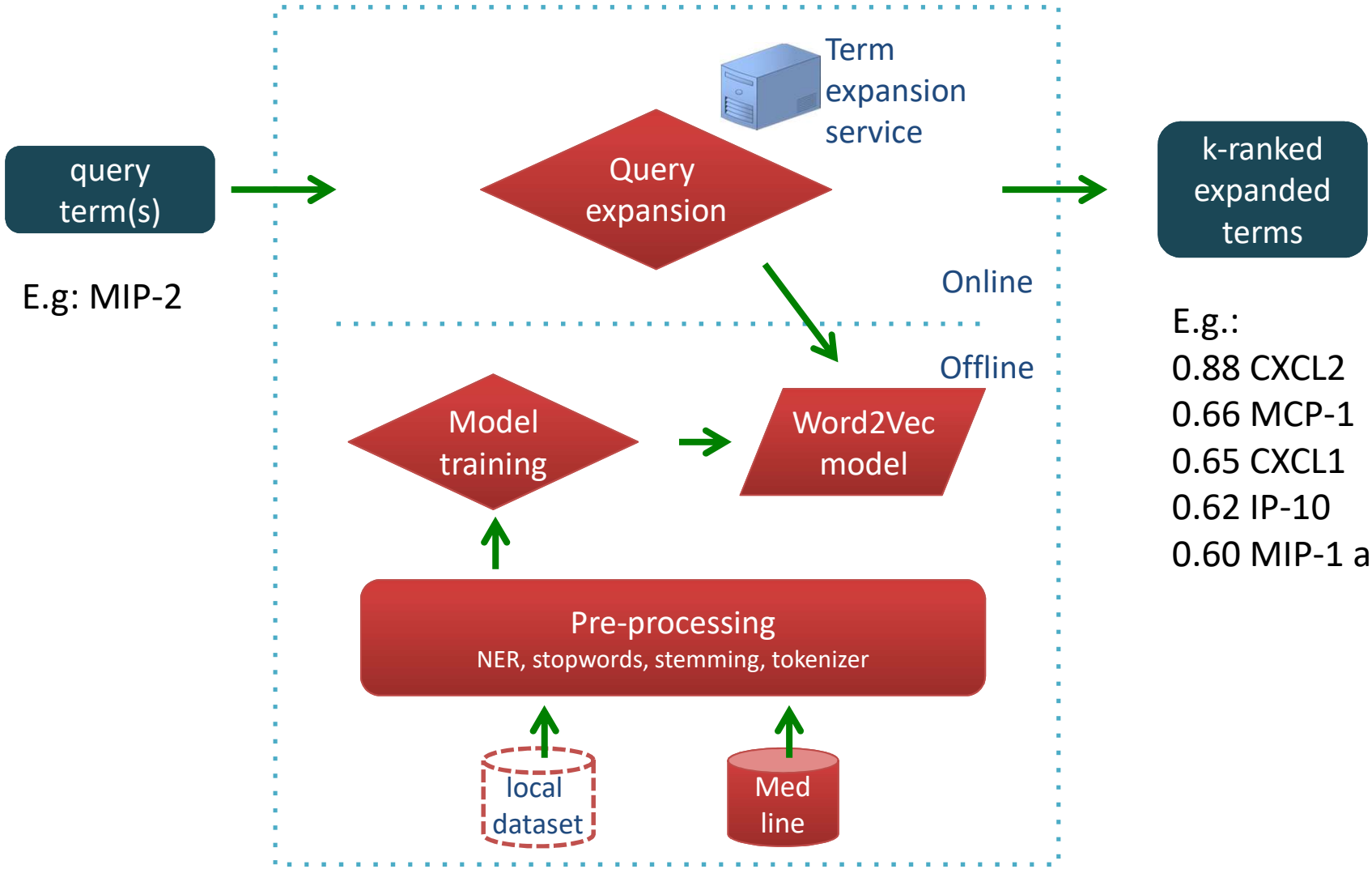
Classical search process



Enhance expansion and classification



Query expansion service



Query expansion service

goldorak.hesge.ch:8088/getebioserver

Generalized Term Expander in Bio

Query

cardiovascular diseases

Top N terms 10

Noun phrase model

Result type

JSON

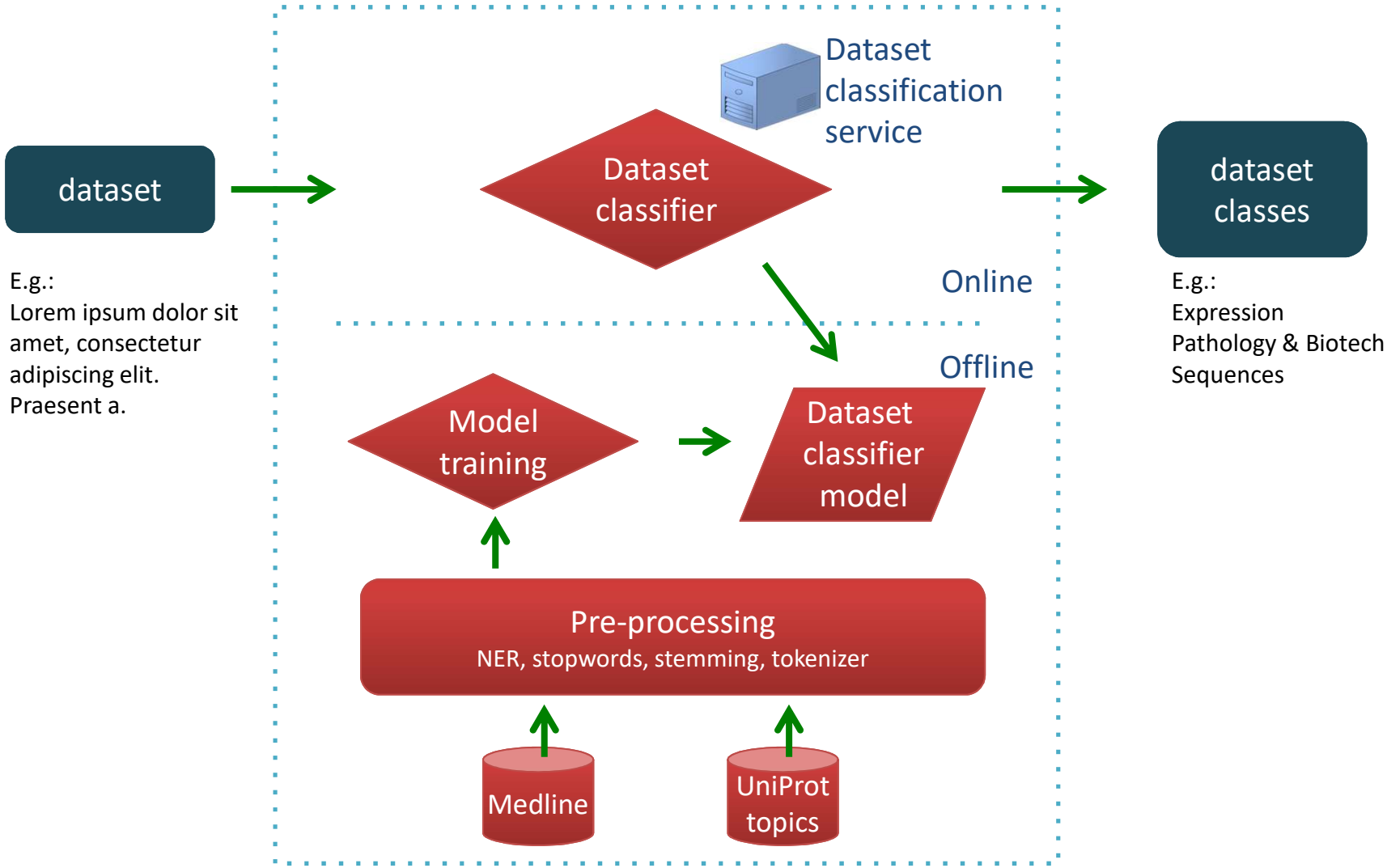
HTML

Expand

Expansion results

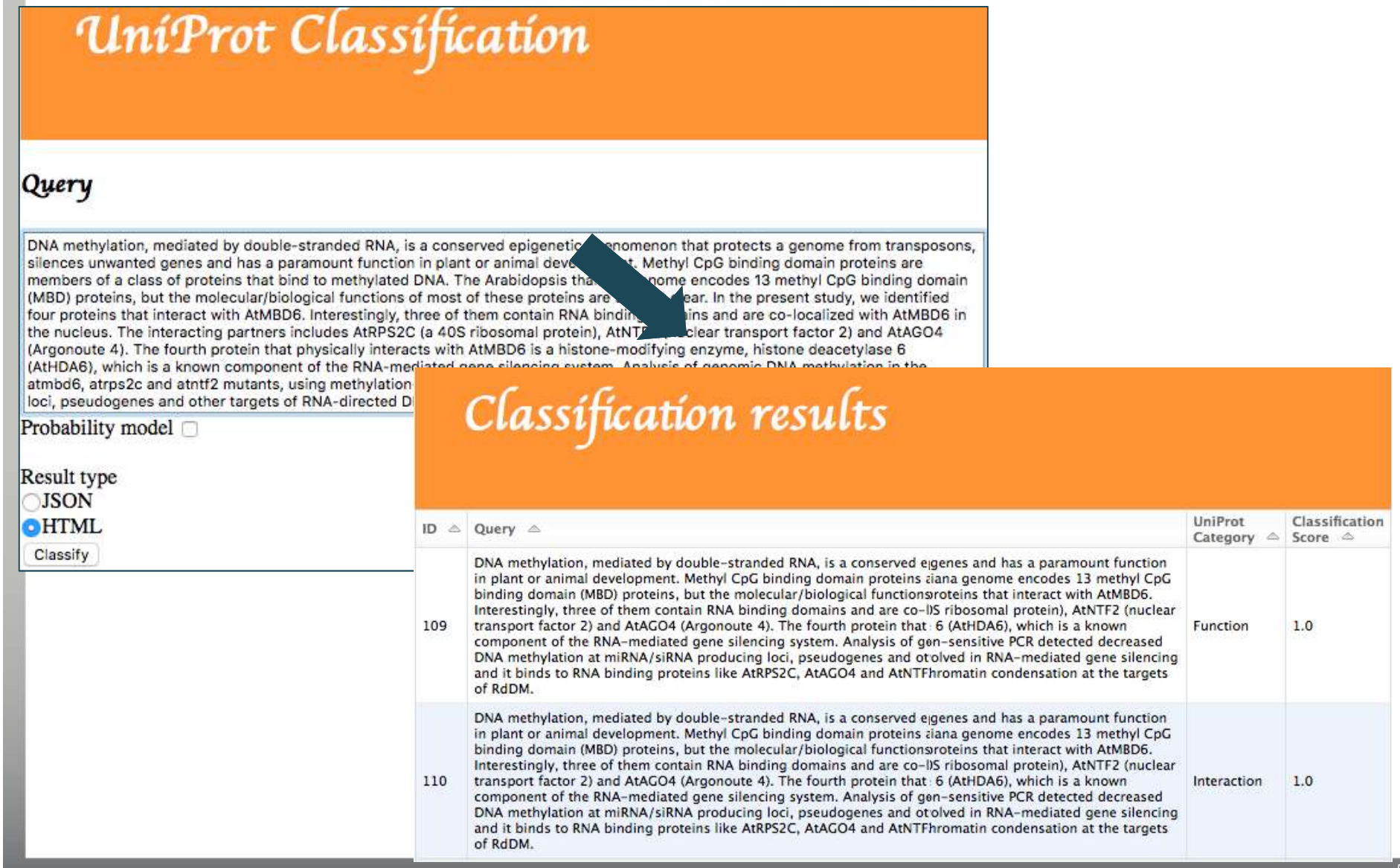
ID	Query	Expanded Term	Expanded Lemma	Proximity Score	Term Rank
341	cardiovascular_diseases|NOUN	cardiovascular disease	cardiovascular disease	0.8751567602157593	0
342	cardiovascular_diseases|NOUN	atherosclerosis	atherosclerosis	0.7965342402458191	1
343	cardiovascular_diseases|NOUN	vascular diseases	vascular diseas	0.7777147889137268	2
344	cardiovascular_diseases|NOUN	cardiovascular complications	cardiovascular complication	0.746307909488678	3
345	cardiovascular_diseases|NOUN	vascular disease	vascular disease	0.7457807064056396	4
346	cardiovascular_diseases|NOUN	coronary heart disease	coronary heart disease	0.7361394166946411	5
347	cardiovascular_diseases|NOUN	metabolic syndrome	metabolic syndrome	0.7336413860321045	6
348	cardiovascular_diseases|NOUN	hypertension	hypertension	0.7325003147125244	7

Dataset classification service



Dataset classification service

goldorak.hesge.ch:8088/upclass



Conclusion

- Doc2Vec improves categorization but no impact on search
  Hypothesis: expert validation would be needed
- Doc2Vec improves query expansion
- Dataset search is possible with ~75% precision

Future work

- Embedded search into EGA
  - Swiss EGA node for SPHN (BioMedIT)
  - Cross-link with Swiss Hospitals Clinical Data Warehouses (SPHN)

- EGA / ELIXIR
  - Central discovery tool
  - Embedded meta-indexing with validation by dataset submitter

EMBL-EBI

European Genome-phenome Archive

SPHN Swiss Personalized Health Network

Hes·so Haute Ecole Spécialisée de Suisse occidentale University of Applied Sciences Western Switzerland

h e g Haute école de gestion de Genève Geneva School of Business Administration

excelerate

SIB Swiss Institute of Bioinformatics

Acknowledgements

- SIB Text Mining / HES-SO

Emilie Pasche
Julien Gobeill
Luc Mottin
Arnaud Gaudinat
Romain Tanzer
Daniel Texeira
Pascale Gaudet (CALIPHO)
Aurore Britan (CALIPHO)
Amos Bairoch (CALIPHO)
Pierre-André Michel (CALIPHO)

- Novartis Institute for Biomedical Research

Fatma Oezdermir-Zaech
Pierre Parisot
Olivier Kreim
Therese Vachon

- EBI-EMBL

Thomas Keane
Jo McEntyre

Thank you

SIB
Swiss Institute of Bioinformatics

Triage by Ranking to Support the Curation of Protein Interactions

Patrick Ruch

SIB Text Mining
HES-SO / HEG Geneva and SIB Swiss Institute of Bioinformatics

Three step-curation

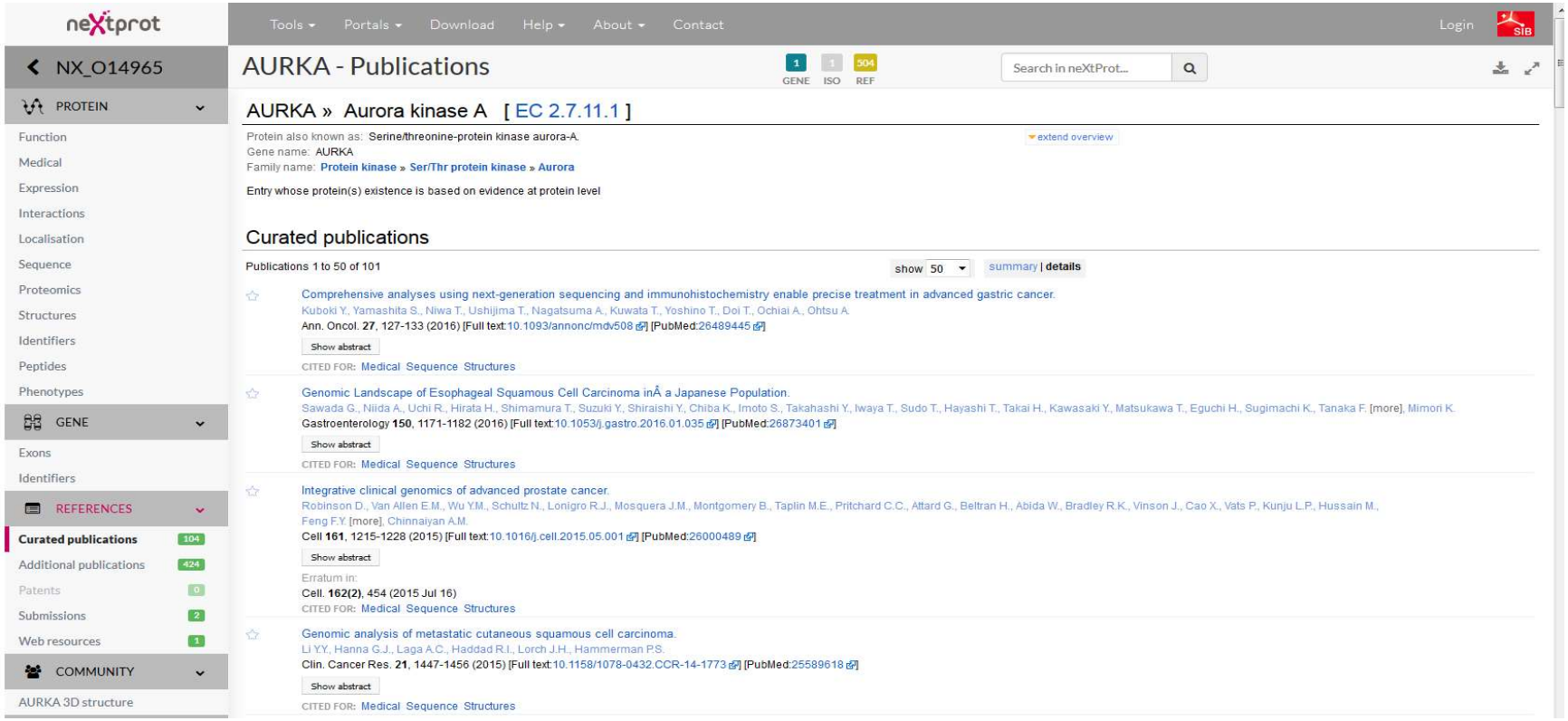
Find

Curate

Share/
Save

Objectives

- Improve triage to support annotation of two data types
  - Protein Interactions
  - Post-Translational Modifications



Methods

- Triage – Focus of the evaluation
  - 1 Pre-annotate PPIs/PTMs descriptors in MEDLINE/PMC → BioMed DB
  - 2 Search protein in a relevance-driven search engine (BioMed) → Ranked list
  - 3 Search protein + PTMs/PPIs specific keywords → Ranked list (neXtA5)
  - 4 Query-independent ranking of content-rich PPIs/PTMs papers → Ranked list
  - 5 Merge by linear combination to obtain a unique ranking
  - 6 Select of PMIDs / PMC by curators
- Annotation - Under evaluation
  - 6 Identify (and normalized) proteins and interactions
  - 7 Select relevant protein-protein relationships
  - 8 Save triples

    [REST web services available]

Protein Interactions

- Subset of 29 concepts instead of a full ontology from the Proteomics Standards Initiative
  - bind, link, …
- Query refinement : “binds + interacts + associates”

$$\text{Linear combination} = 0.9 \times \text{search engine score} + 1.5 \times \sum \text{distinct descriptor}$$

$$\text{Linear combination} = 1.0 \times \text{search engine score} + 0.1 \times \sum_{\text{descriptor}} \log(1 + \text{descriptor length} \times \text{term frequency of descriptor})$$

Post-Translational Modifications

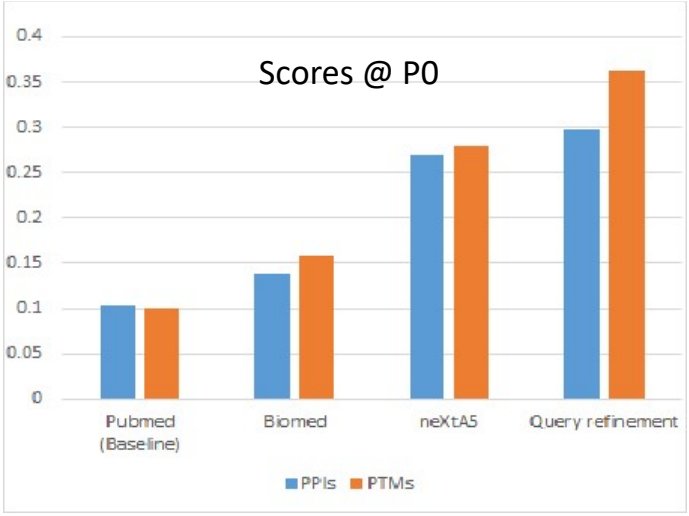
- 16 most frequent PTMs in literature
  - phosphorylation, glycosylation, …

$$\text{Linear combination} = 0.9 \times \text{search engine score} + 1.7 \times \sum \text{distinct descriptor}$$

- Query Refinement : “phosphorylation”

$$\text{Linear combination} = 1.4 \times \text{search engine score} + 1.3 \times \sum \text{distinct descriptor}$$

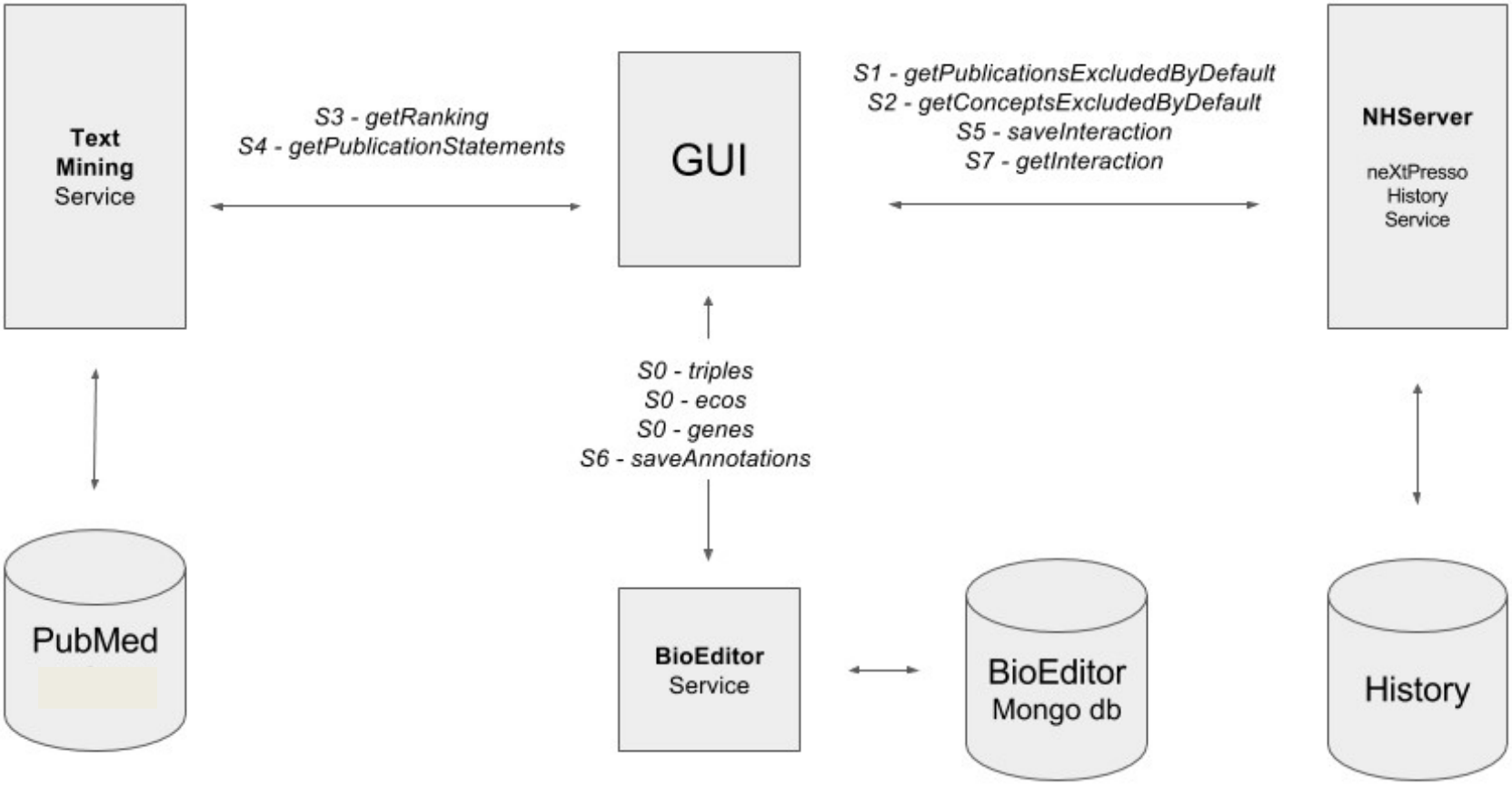
Results



	P0	P100
PubMed	-	-
BioMed	+34%	+30%
neXtA5	+170%	+101%
Query refinement	+191%	+66%

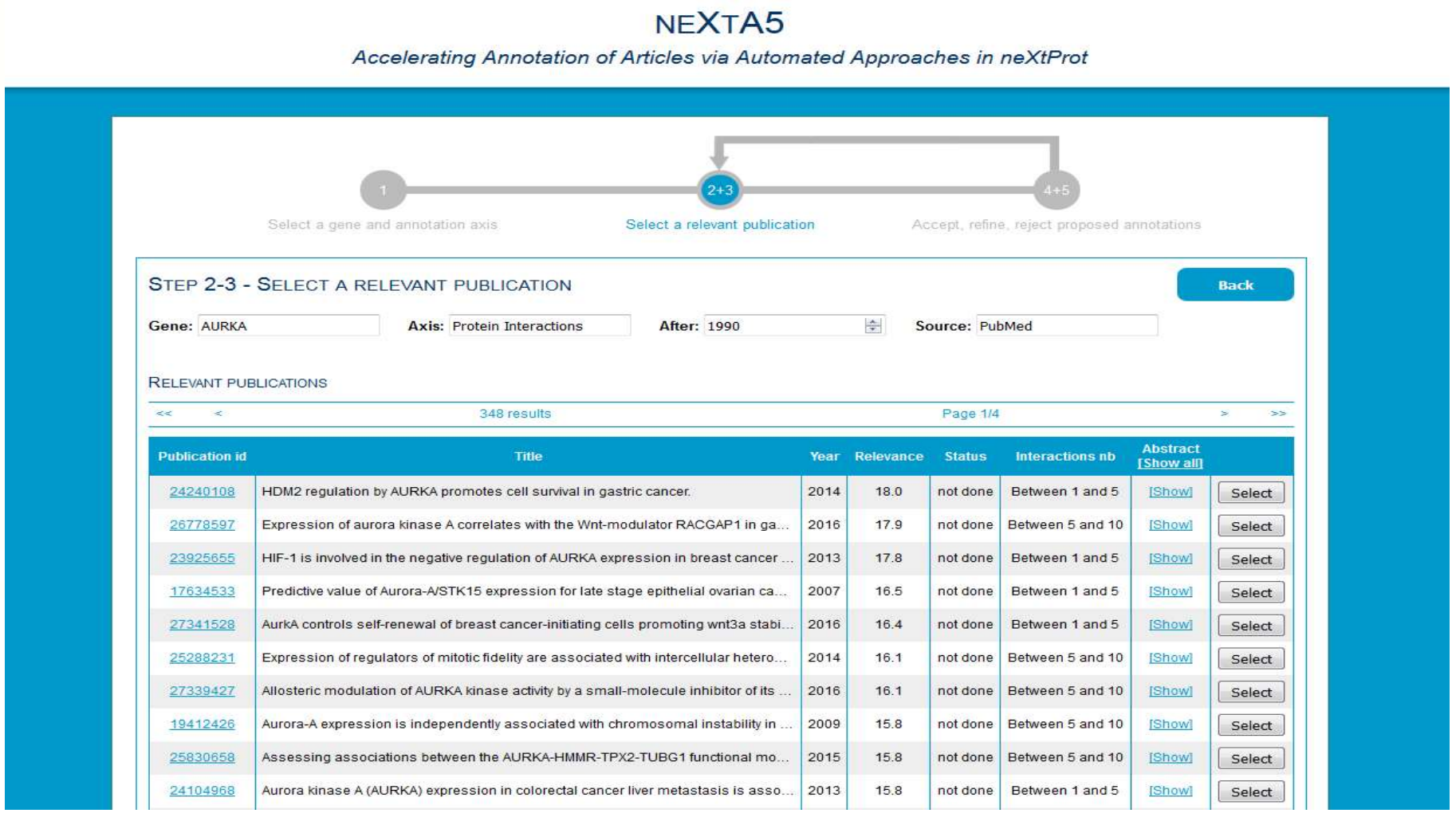
	P0	P100
PubMed	-	-
BioMed	+57%	+19%
neXtA5	+180%	+63%
Query refinement	+261%	+91%

Functional architecture

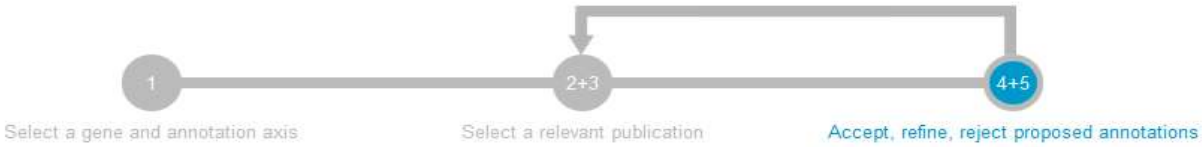


Online prototype

http://casimir.hesge.ch/nextA5/



Online prototype



STEP 4-5 - ACCEPT, REFINE, REJECT PROPOSED ANNOTATIONS

Back

Gene: AURKA **Axis:** Protein Interactions **Source:** PubMed **Id:** 27339427

Allosteric modulation of AURKA kinase activity by a small-molecule inhibitor of its protein-protein interaction with TPX2.

Abstract: The essential mitotic kinase Aurora A (**AURKA**) is controlled during cell cycle progression via two distinct mechanisms. Following activation loop autophosphorylation early in mitosis when it localizes to centrosomes, **AURKA** is allosterically activated on the mitotic spindle via binding to the microtubule-associated protein, TPX2. Here, we report the discovery of AurkinA, a novel chemical inhibitor of the **AURKA**-TPX2 interaction, which acts via an unexpected structural mechanism to inhibit **AURKA** activity and mitotic localization. In crystal structures, AurkinA binds to a hydrophobic pocket (the 'Y pocket') that normally accommodates a conserved Tyr-Ser-Tyr motif from TPX2, blocking the **AURKA**-TPX2 interaction. AurkinA binding to the Y- pocket induces structural changes in **AURKA** that inhibit catalytic activity in vitro and in cells, without affecting ATP binding to the active site, defining a novel mechanism of allosteric inhibition. Consistent with this mechanism, cells exposed to AurkinA mislocalise **AURKA** from mitotic spindle microtubules. Thus, our findings provide fresh insight into the catalytic mechanism of **AURKA**, and identify a key structural feature as the target for a new class of dual-mode **AURKA** inhibitors, with implications for the chemical biology and selective therapeutic targeting of structurally related kinases.

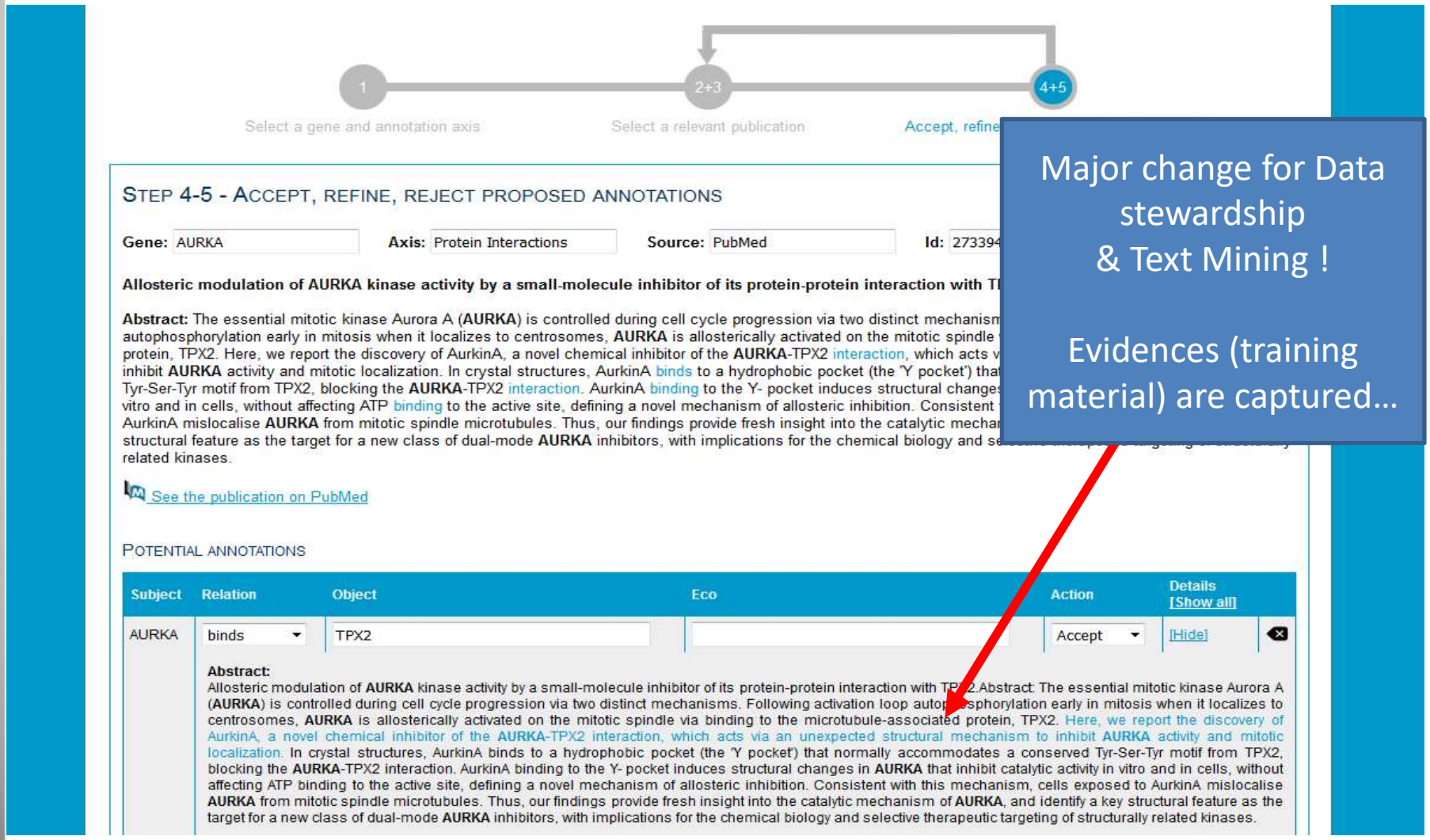
See the publication on PubMed

POTENTIAL ANNOTATIONS

Subject	Relation	Object	Eco	Action	Details [Show all]
AURKA	binds	TPX2		Accept	[Hide]

Abstract:
Allosteric modulation of **AURKA** kinase activity by a small-molecule inhibitor of its protein-protein interaction with TPX2.Abstract: The essential mitotic kinase Aurora A (**AURKA**) is controlled during cell cycle progression via two distinct mechanisms. Following activation loop autophosphorylation early in mitosis when it localizes to centrosomes, **AURKA** is allosterically activated on the mitotic spindle via binding to the microtubule-associated protein, TPX2. Here, we report the discovery of AurkinA, a novel chemical inhibitor of the **AURKA**-TPX2 interaction, which acts via an unexpected structural mechanism to inhibit **AURKA** activity and mitotic localization. In crystal structures, AurkinA binds to a hydrophobic pocket (the 'Y pocket') that normally accommodates a conserved Tyr-Ser-Tyr motif from TPX2, blocking the **AURKA**-TPX2 interaction. AurkinA binding to the Y- pocket induces structural changes in **AURKA** that inhibit catalytic activity in vitro and in cells, without affecting ATP binding to the active site, defining a novel mechanism of allosteric inhibition. Consistent with this mechanism, cells exposed to AurkinA mislocalise **AURKA** from mitotic spindle microtubules. Thus, our findings provide fresh insight into the catalytic mechanism of **AURKA**, and identify a key structural feature as the target for a new class of dual-mode **AURKA** inhibitors, with implications for the chemical biology and selective therapeutic targeting of structurally related kinases.

Online prototype



Major change for Data stewardship
& Text Mining !

Evidences (training material) are captured…

Conclusion

- Triage by pre-annotating literature is effective
  - PPIs +191%
  - PTMs +261%
  - Diseases – Functions – Cell location: factor 2-30 !
- UX and productivity gain for triage currently evaluated
  - Gain of time on the whole process ?
  - Usability to improve UX

Acknowledgements

SIB Swiss Institute of Bioinformatics

BiTeM group

- Patrick Ruch
- Julien Gobeill
- Emilie Pasche
- Douglas Teodoro
- Dina Vishnyakova

CALPIPHO group

- Pascale Gaudet
- Valentine Rech de Laval
- Anne Gleizes
- Isabelle Cusin
- Pierre-André Michel
- Amos Bairoch

Funding: [neXtPresso SNF 153437]

… and YOU !!

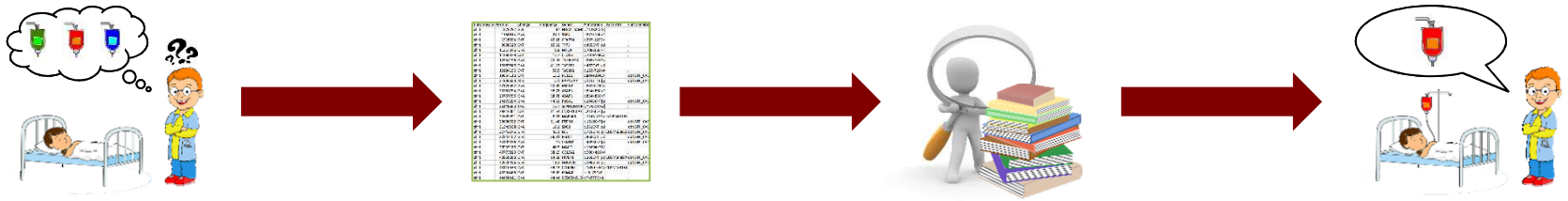
Decision-support for personalized health

a variome scale challenge for evidence-based prescription

Patrick Ruch – patrick.ruch@hesge.ch @patrickruch
SIB Text Mining Group & HES-SO – HEG Geneve

Context

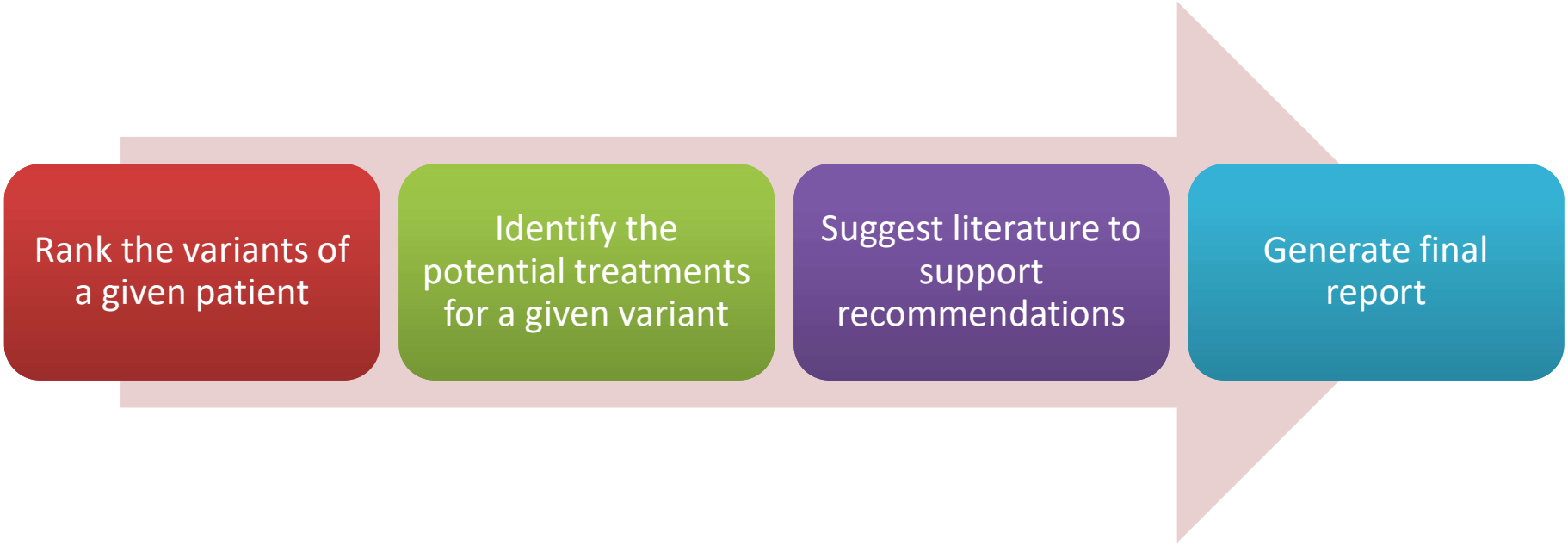
- Personalized medicine in oncology
- Characterization of variants is labour-intensive



Patient	Sequencing/Calling	Interpretation	Treatment
A patient is diagnosed with a cancer	Mutations are identified	Mutations are curated for pathogenicity → therapies	A personalized clinical report is generated

Objective

- System to automatically rank SNVs and chemotherapies
- Based on occurrences in the biomedical literature
- Acceleration of the variant analysis process



Ranking of variants

- Data

Patient	SNV total	CNV total	Clinic. relevant SNV	Clin. relevant CNV
1	4569	~4789	3	14
2	88	~1118	1	6
3	985	~2028	1	6
4	44	0	14	0
5	20	0	14	0

Patient 3 and patient 5 selected for tuning.

Ranking of variants

- Data

Patient	SNV total	CNV total	Clinic. relevant SNV	Clin. relevant CNV
1	4569	~4789	3	14
2	88	~1118	1	6
3	985	~2028	1	6
4	44	-	14	-
5	20	-	14	-

→Large-scale learning to rank / selection problem

Patients #3 and #5 selected for tuning

Funnel

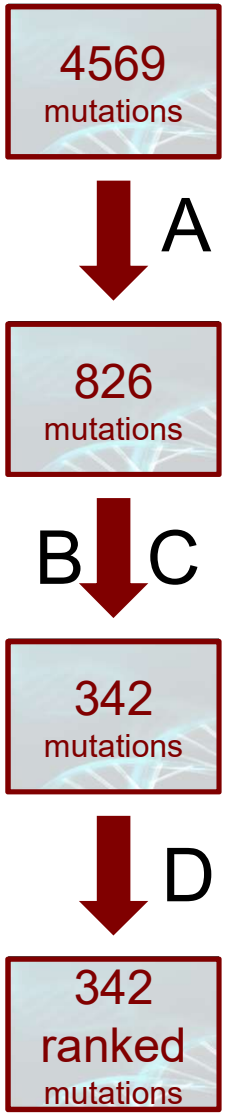
- Method

A. Preprocessing
   - Non coding SNVs are removed from the list
   - Reduction of SNVs by a factor 5

B. Query building
   - Generation of several queries
   - E.g. Disease + Gene + Variant ; Gene + Variant …

C. Query expansion
   - SNV generator, based on HVGS nomenclature and non standard formats and expressions encountered in literature

D. Ranking
   - Based on number of publications found

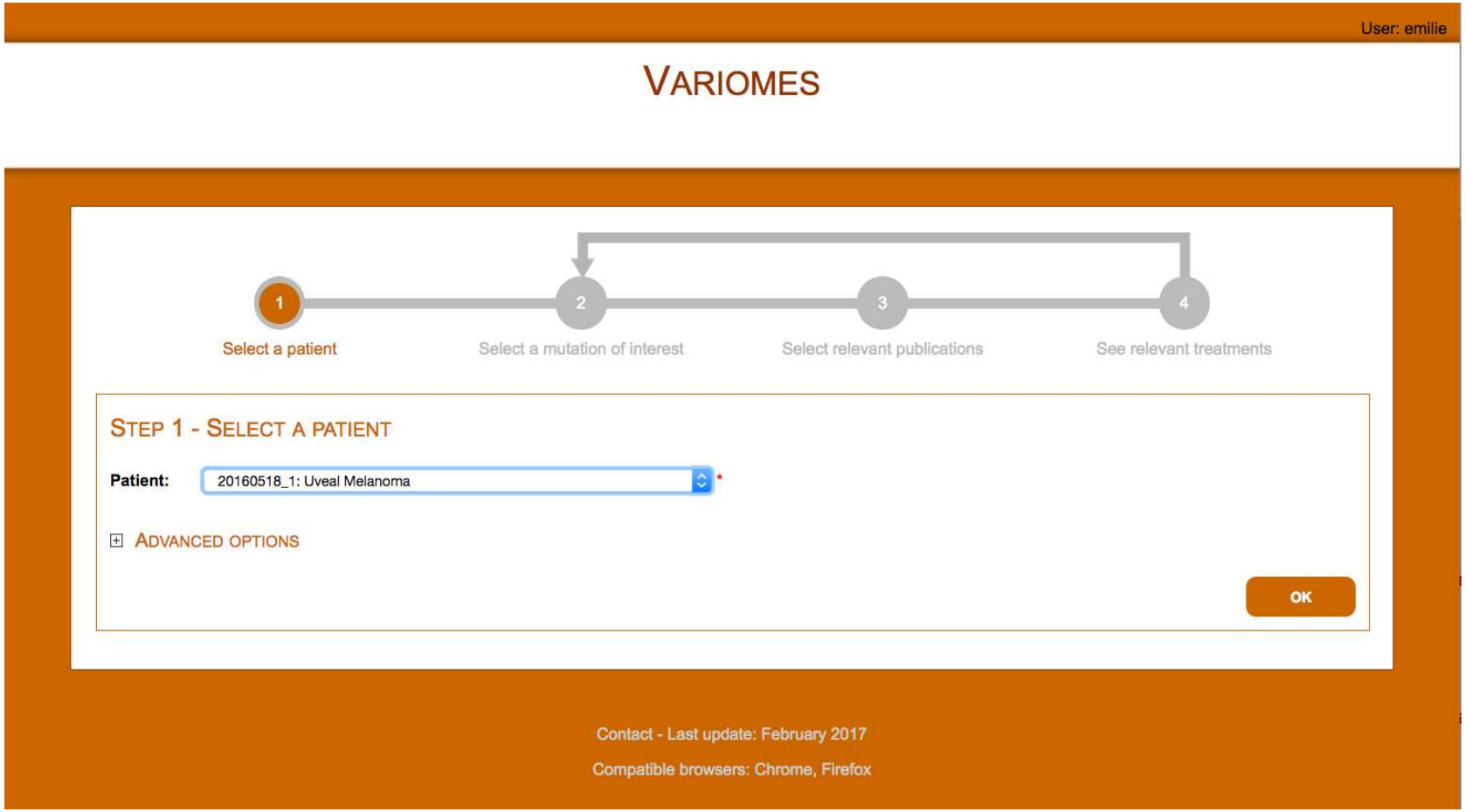


Evaluation

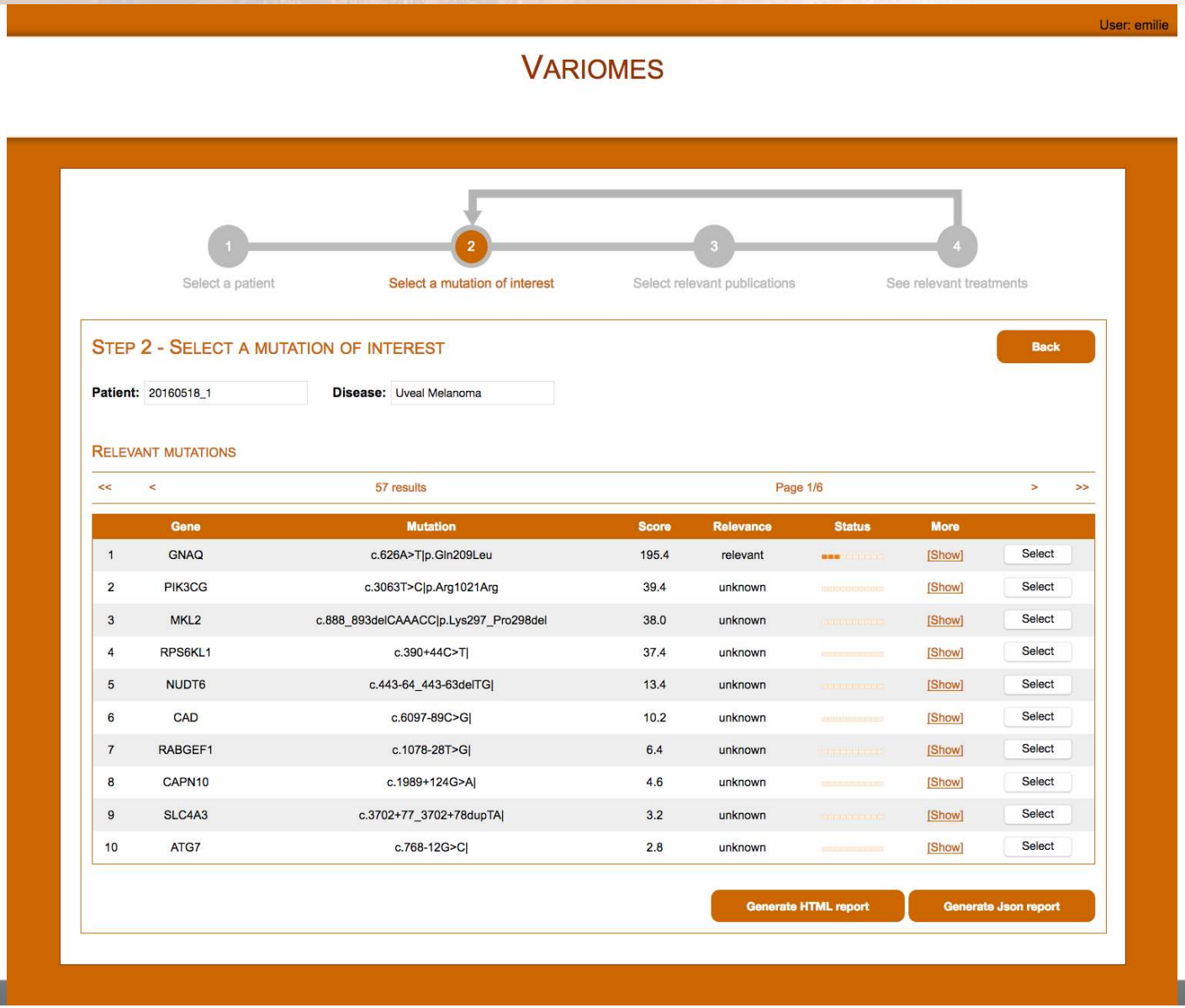
- Preliminary results
  - Mean reciprocal rank: 81%
  - Precision at rank 5: 62%
  - Recall at rank 5: 58%
  - +++ synonyms for variants
  - +++ full-text vs. abstract
  - +++ up to 1000 publications per SNVs
  - Benchmark is not “gold”

  ⇒ precision of the system is probably higher

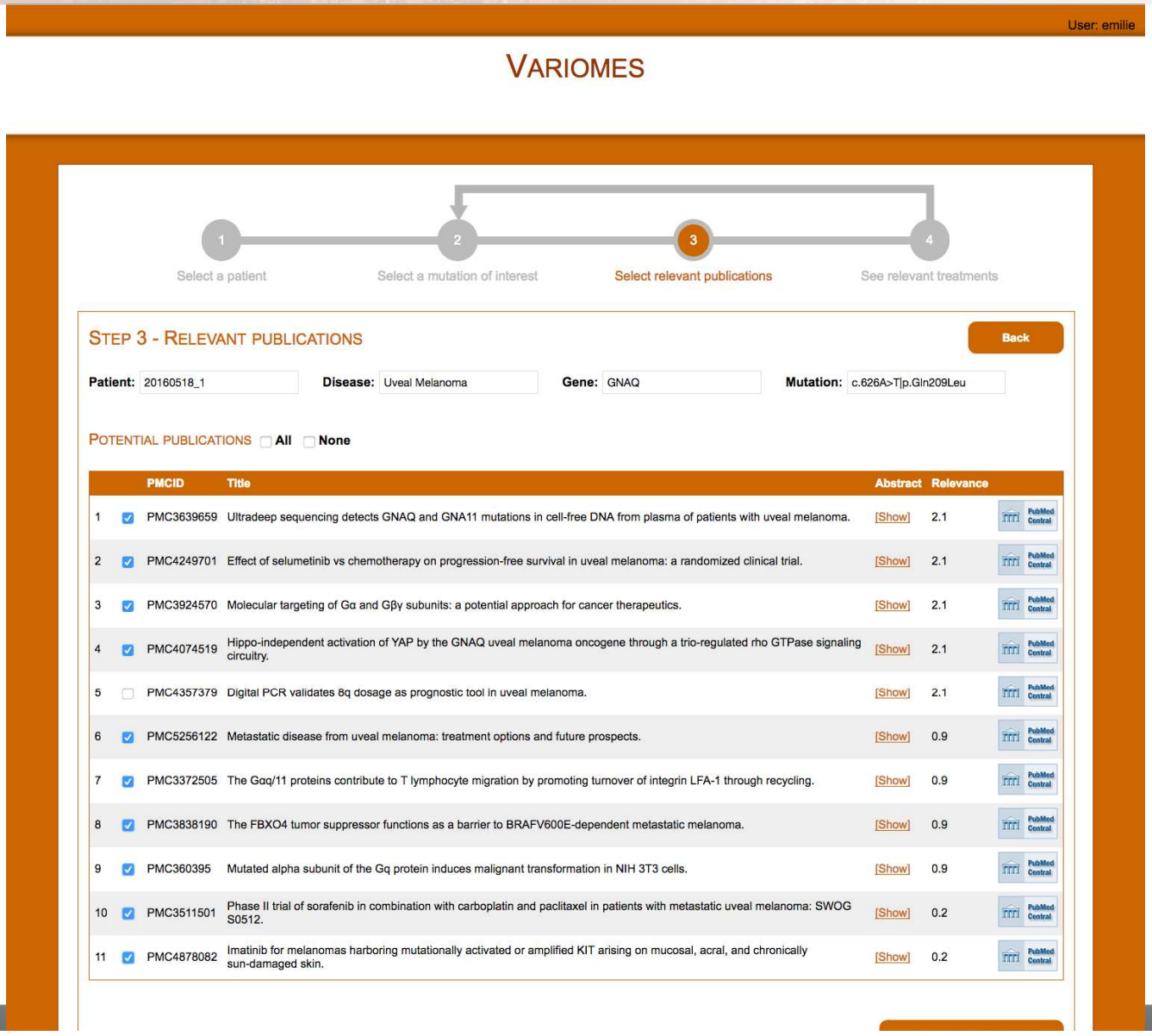
Graphical User Interface



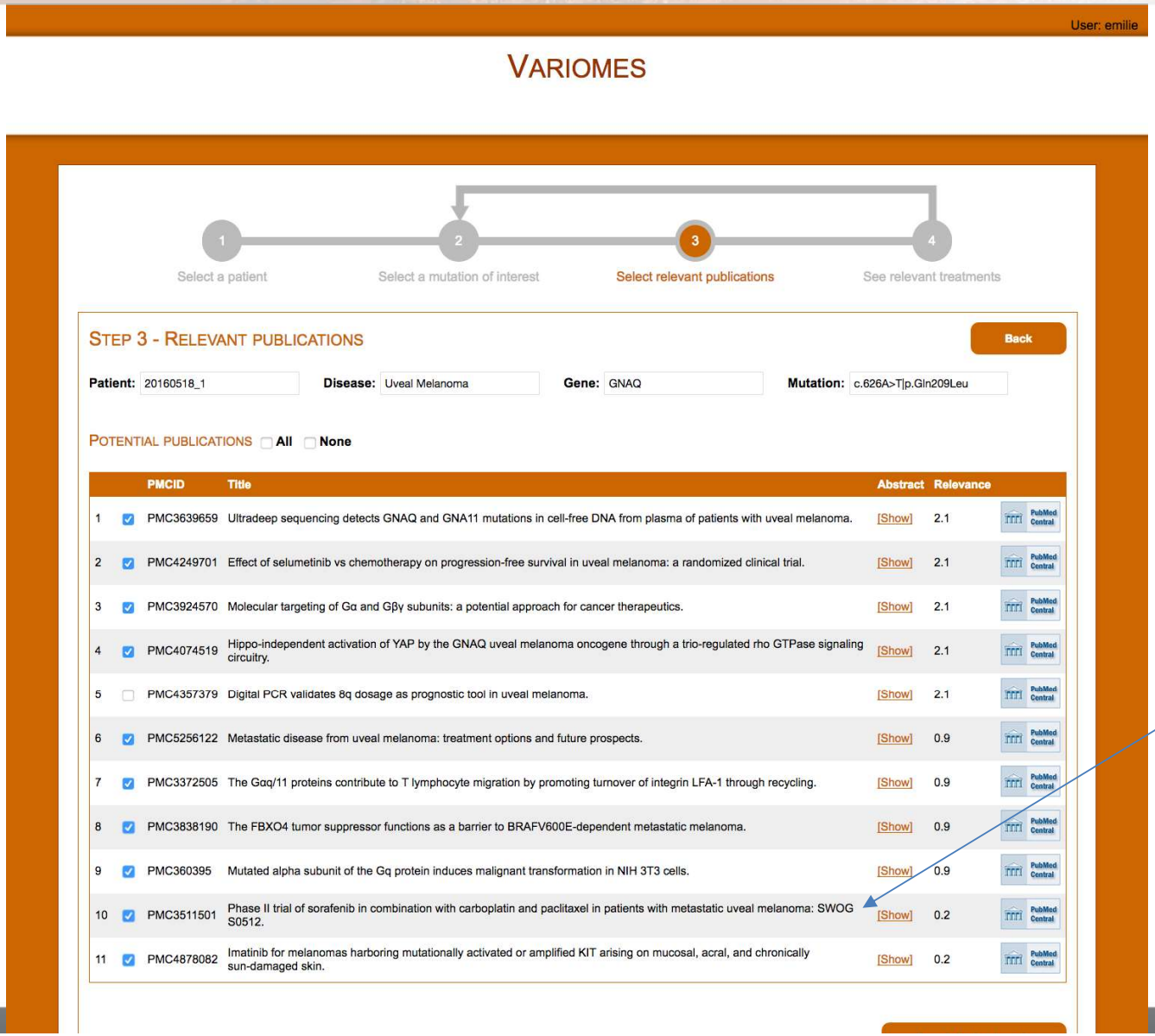
Graphical User Interface



Graphical User Interface

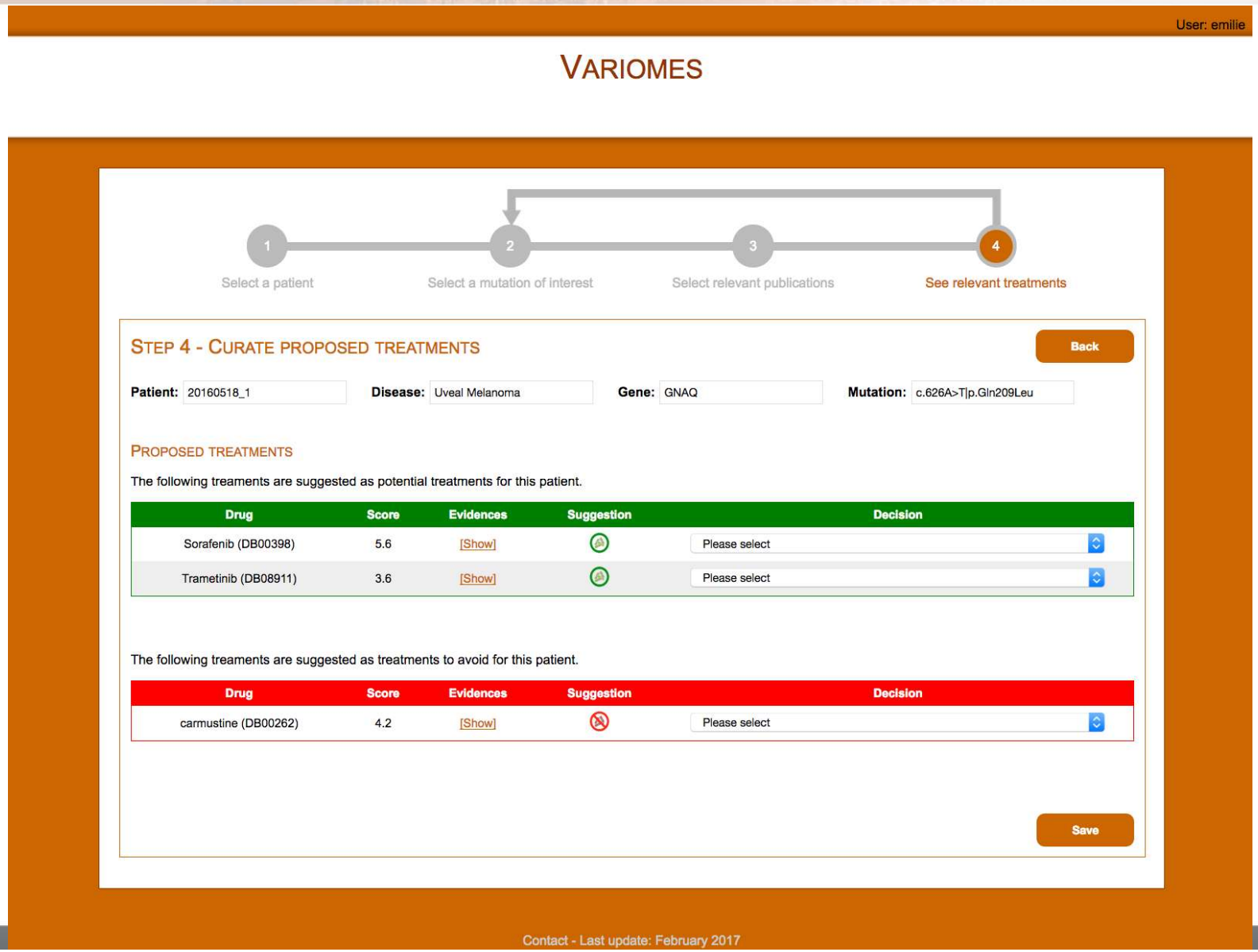


Graphical User Interface

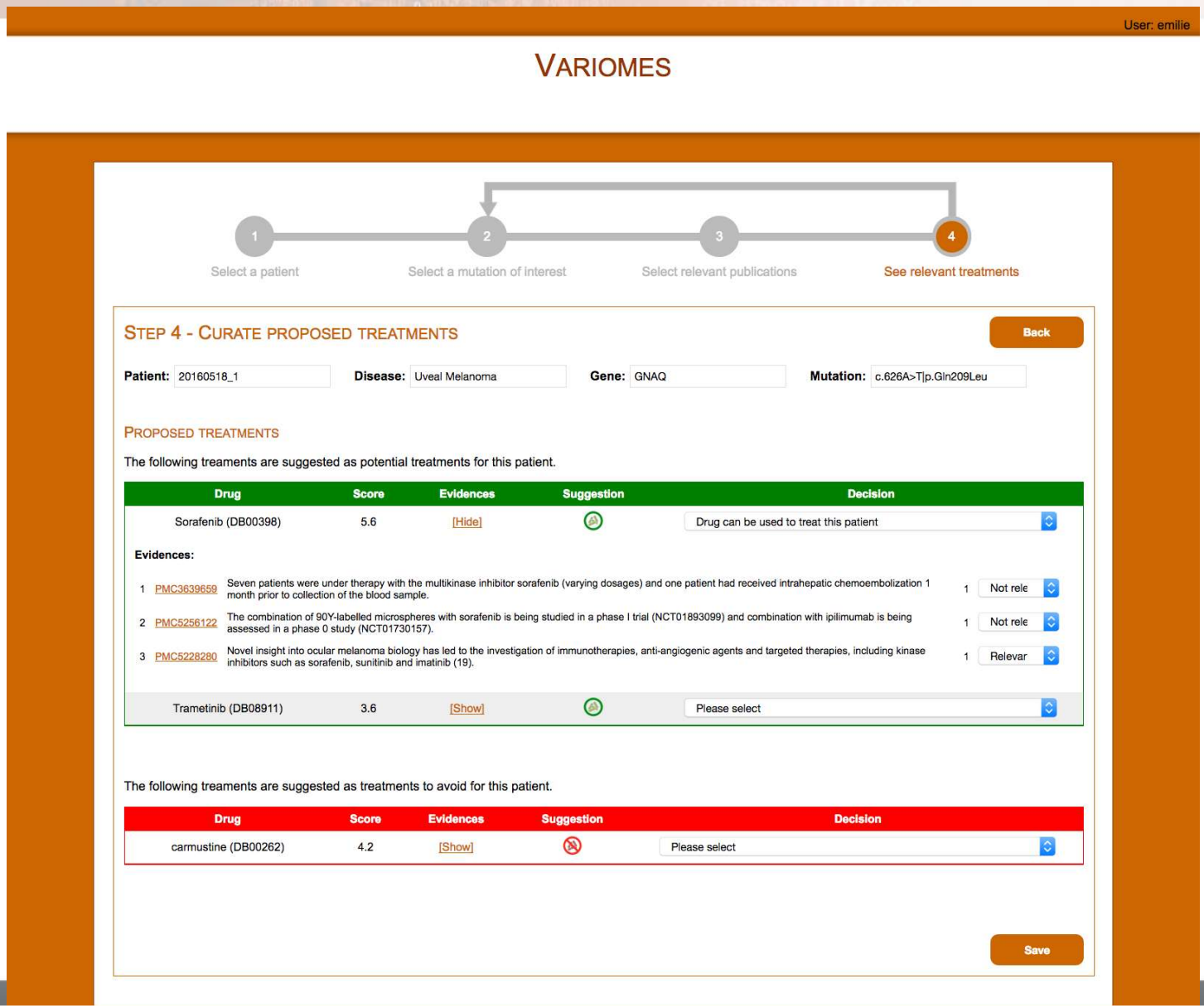


e.g. Sorafenib

Graphical User Interface



Graphical User Interface



Graphical User Interface

MOLECULAR TUMOR REPORT

generated by SIB Text mining

GENERALITIES

Report generated by: emilie

Patient identifier: 20160518_1

Disease: Uveal Melanoma (C7712)

SOMATIC MUTATED GENES, CANCER-TYPE SPECIFIC THERAPY

Gene	Mutation	Pathway	Treatment	Score	Références
GNAQ	c.626A>T|p.Gln209Leu	Acetylcholine regulates insulin secretion G alpha (q) signalling events ADP signalling through P2Y purinoceptor 1 Thromboxane signalling through TP receptor Acids bound to GPR40 (FFAR1) regulate insulin secretion Thrombin signalling through proteinase activated receptors (PARs)	DB00398 (Sorafenib)	5.6	PMC3639659 PMC5256122 PMC5228280
GNAQ	c.626A>T|p.Gln209Leu	Acetylcholine regulates insulin secretion G alpha (q) signalling events ADP signalling through P2Y purinoceptor 1 Thromboxane signalling through TP receptor Acids bound to GPR40 (FFAR1) regulate insulin secretion Thrombin signalling through proteinase activated receptors (PARs)	DB08911 (Trametinib)	3.6	PMC3639659 PMC5256122

THERAPIES WITH POTENTIAL LACK OF BENEFIT

Gene	Mutation	Pathway	Treatment	Score	Références
GNAQ	c.626A>T|p.Gln209Leu	Acetylcholine regulates insulin secretion G alpha (q) signalling events ADP signalling through P2Y purinoceptor 1 Thromboxane signalling through TP receptor Acids bound to GPR40 (FFAR1) regulate insulin secretion Thrombin signalling through proteinase activated receptors (PARs)	DB00262 (carmustine)	4.2	PMC4040458

Next steps

- Extraction of chemotherapies
  - Drug list + off-label (?)
- **Scale-up evaluation and control negative**
- Copy Number Variant (CNV)
- Scale patient population: CHUV, HUG, …

Infrastructure is needed: BioMedIT + SPHN

- Swiss Variant Interpretation Platform…
  and maybe beyond Switzerland (cf. Beacon)

- Participant-level / Patient-level Data + sequences
  [ELSI]

- Access-restricted evidence data (e.g. EGA)

→ **Trusted third party (SAMW) / Federated platform**

Acknowledgements

SIB Text Mining / HES-SO
Emilie Pasche
Anais Mottaz
Luc Mottin
Julien Gobeill

Nexus (ETH/USZ)
Franziska Singer
Nora Toussaint
Daniel Stekhoven

SIB ClinBio
Valérie Barbié
Aitana Lebrand

Theoretical oncogenomics (UZH)
Michael Baudis

SIB Tech
Heinz Stockinger
Danier Texeira

Special thanks to the

Librarians of the Swiss Academy of Medical Sciences