Institut "Jožef Stefan" Ljubljana, Slovenija

INSTITUT "JOŽEF STEFAN" RAČUNALNIŠKI SISTEMI

SRIP PMiS

SRIP TOP
Strateško razvojno inovacijsko partnerstvo
TOVARNE PRIHODNOSTI

tetramax

Linking Food, Nutrition and Biomedical Data for Trustworthy AI in Predictive Healthcare

Online Workshop

17. 11. 2021, 12:30 – 15:00

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA GOSPODARSKI RAZVOJ IN TEHNOLOGIJO

The TETRAMAX project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement number 761349.

Jožef Stefan Institute

Heterogeneous Big Data in food, nutrition, and biomedicine

Barbara Koroušić Seljak

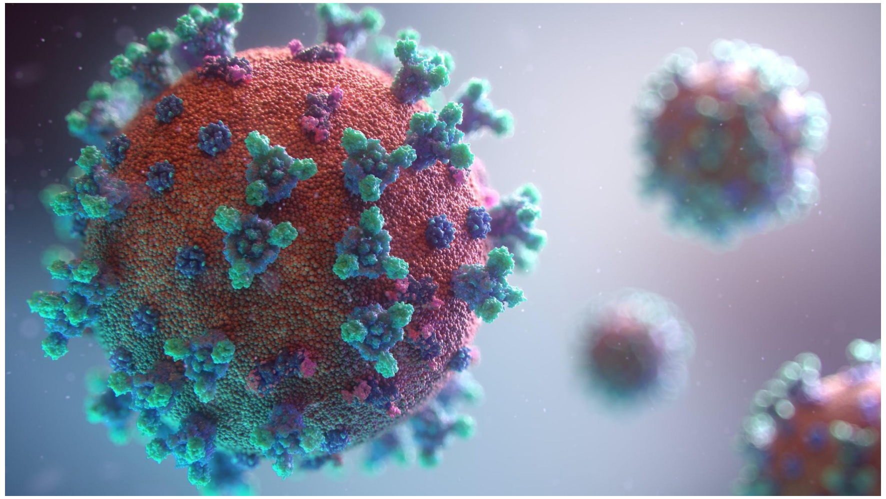
Computer Systems Department (http://cs.ijs.si)

JOŽEF STEFAN INSTITUTE COMPUTER SYSTEMS

Jožef Stefan Institute, Ljubljana

How Food and Nutrition can help to fight against COVID-19 Pandemic?

- EVIDENCE-BASED KNOWLEDGE: An optimal intake of food and nutrients **strengthens our immune system**.
- OPEN RESEARCH QUESTIONS:
  1. What is an optimal nutritional status under specific (personal and external) conditions?
  2. How to achieve an optimal nutrition?



How to find answers to such open research questions?

- We require knowledge based on scientific evidence (and not intuition)!
- Steps to acquire such knowledge:
    1. Conduct research
    2. Collect reliable *data* and corresponding *metadata*
    3. Extract *information* from data
    4. Assemble *knowledge* from multiple pieces of information

How to find answers to open research questions?

- We require knowledge based on scientific evidence (and not intuition)!
- Steps to acquire such knowledge:
  1. Conduct research
  2. Collect reliable *data* and corresponding *metadata*
  3. Extract *information* from data
  4. Assemble *knowledge* from multiple pieces of information

Our focus

The **Data** – Information – Knowledge hierarchy

- Data are the individual facts that have no meaning and are difficult to be understood.
  - For example: 30.85 mg

The **Data** – Information – Knowledge hierarchy

- Data are the individual facts that have no meaning and are difficult to be understood.
  - For example: 30.85 mg

- Once data are considered in specific context, they get a meaning and can be understood.
  - Context: *content of Cy-3-GE per mL of elderberry juice*
  - Data in this context: *30.85 mg of Cy-3-GE per mL of elderberry juice*

The Data – **Information** – Knowledge hierarchy

- Information is a set of data in context with relevance (to one or more people at a point of time or for a period of time):
    - *1 ml of the elderberry (Sambucus nigra L.) juice contains 30.85 mg of Cy-3-GE*

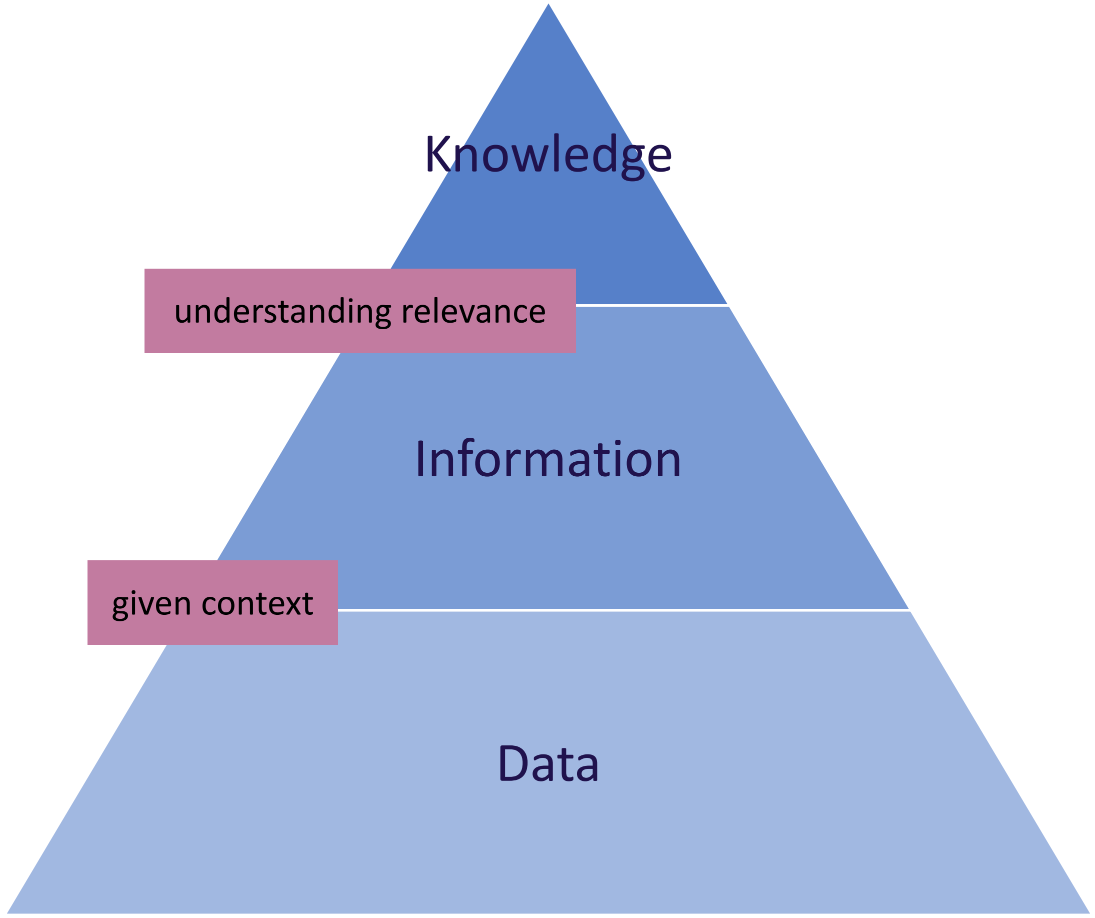
The Data – Information – **Knowledge** hierarchy

- Knowledge is information that has been retained with **an understanding about the significance of that information**.
  - Considering the pieces of information:
    - *1 ml of the elderberry (Sambucus nigra L.) juice contains 30.85 mg of Cy-3-GE.*
    - *Cy-3-GE is a term for the compound cyanidin-3-O-glucoside.*
    - *Anthocyanins have shown antimicrobial, antioxidative, anti-inflammatory, and anti-mutagenic properties.*
    - *The intake of even moderate amounts of anthocyanins (<50 mg) daily is associated with risk reduction for cardiovascular disease, type 2 diabetes mellitus, and neurological decline.*
  - we could conclude that drinking the elderberry juice is recommended in the prevention and treatment of the SARS-CoV-2 infections.

- However, there is an additional important piece of information:
  - *The elderberry increases the release of a cytokine (interleukin 1 beta), which is part of the inflammatory reaction to COVID-19 that can result in acute respiratory distress.*
- Therefore, elderberry shouldn't be taken by anyone who tests positive for the virus!

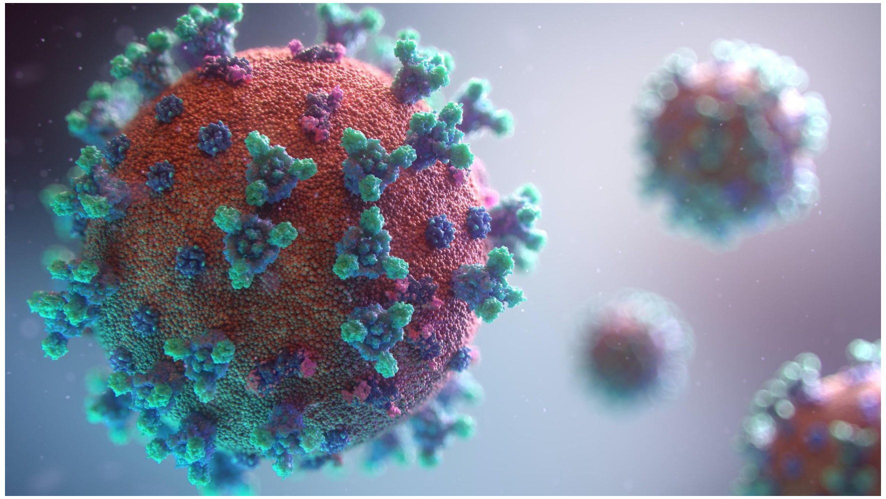
The Data – Information – **Knowledge** hierarchy

- We are focusing on *formal knowledge* that is codified and stored, and is capable of being shared with domain experts and information systems.



How Food and Nutrition can help to fight against COVID-19 Pandemic?

- EVIDENCE-BASED KNOWLEDGE: An optimal intake of food and nutrients strengthens our immune system.
- OPEN RESEARCH QUESTIONS:
  1. What is an **optimal nutritional status** under specific (personal and external) conditions?
  2. How to achieve an optimal nutrition?



Which factors influence the nutritional status?

- The nutritional status of an individual is affected by several factors, such as
  - age and sex,
  - health status,
  - life style (including dietary habits) and
  - medications.

- In order to find an answer to the first question (*"What is an optimal nutritional status under specific conditions?"*), knowledge based on information extracted from food, nutrition and biomedical data is required.

Food data

- Examples:
  - Food classification
  - Food composition (generic and branded food items)
- Sources:
  - Databases (e.g., EuroFIR) and catalogues (e.g., LanguaL, EFSA FoodEx2)
  - Repositories (e.g., FooDB) and food semantic resources (to be explained by Tome)
  - Peer-reviewed and gray literature (PubMed)
  - Web stores (branded food items)

- Who collects food data: Food scientists applying chemistry, biology, and other sciences to study the basic elements of food.

Nutrition data

- Examples:
  - Food consumption
  - Dietary reference values
  - Dietary recommendations and guidelines
- Sources:
  - Peer-reviewed and grey literature
  - Databases and knowledge bases (e.g., EFSA) - to be explained in the following presentations

- Who collects nutrition data: Nutrition scientists, in order to find information regarding the types and quantities of foods people eat and should eat.

Biomedical data

- Examples:
  - Routine medical data , e.g., height, mass, blood pressure, cholesterol levels, medications used, etc.
  - Specialized laboratory data , e.g., proteins, lipids, metabolites, imaging
  - Genetic data , e.g., genotype or sequencing
  - Gene expressions and epigenetic data
- Possible sources:
  - Medical records
  - Biological measurements
  - Diagnostic procedures
  - Questionnaires or interviews
  - Biomedical semantic resources (to be presented in the next presentations)

- Who collects biomedical data: Biomedical scientists, in order to support the diagnosis and treatment of disease.

The Data – Information – **Knowledge** hierarchy

- Knowledge is information that has been retained with **an understanding about the significance of that information.**
  - Considering the pieces of information:
    - *1 ml of the elderberry (Sambucus nigra L.) juice contains 30.85 mg of Cy-3-GE.*
    - *Cy-3-GE is a term for the compound cyanidin-3-O-glucoside.*
    - *Anthocyanins have shown antimicrobial, antioxidative, anti-inflammatory, and anti-mutagenic properties.*
    - *The intake of even moderate amounts of anthocyanins (<50 mg) daily is associated with risk reduction for cardiovascular disease, type 2 diabetes mellitus, and neurological decline.*
  - we could conclude that drinking the elderberry juice is recommended in the prevention and treatment of the SARS-CoV-2.

Info from food data

Info from nutrition data

- However, there is an additional important piece of information:
  - *The elderberry increases the release of a cytokine (interleukin 1 beta), which is part of the inflammatory reaction to COVID-19 that can result in acute respiratory distress.*
- Therefore, elderberry shouldn't be taken by anyone <u>who tests positive for the virus</u>!

Info from biomedical data

Complete (linked & integrated) information needs to be considered!!!

Information from non-scientific literature

Takoj, ko telo z začetkom potenja sporoči, da je virus »skuhalo«, je čas, da se bezeg ponovno vrne v igro. Z njim pospešimo hlajenje telesa in »izpiranje« trupel premaganih virusov ter preprečimo sekundarne infekcije. Povedano bolj znanstveno, Krawitz et al (2011) poroča, da zgoščen sok jagod črnega bezga poleg inhibicije virusov izkazuje protibakterijski učinek. Posebej pomembno je, da zavira rast bakterij, ki povzročajo pljučnico in vnetja gornjih dihalnih poti.

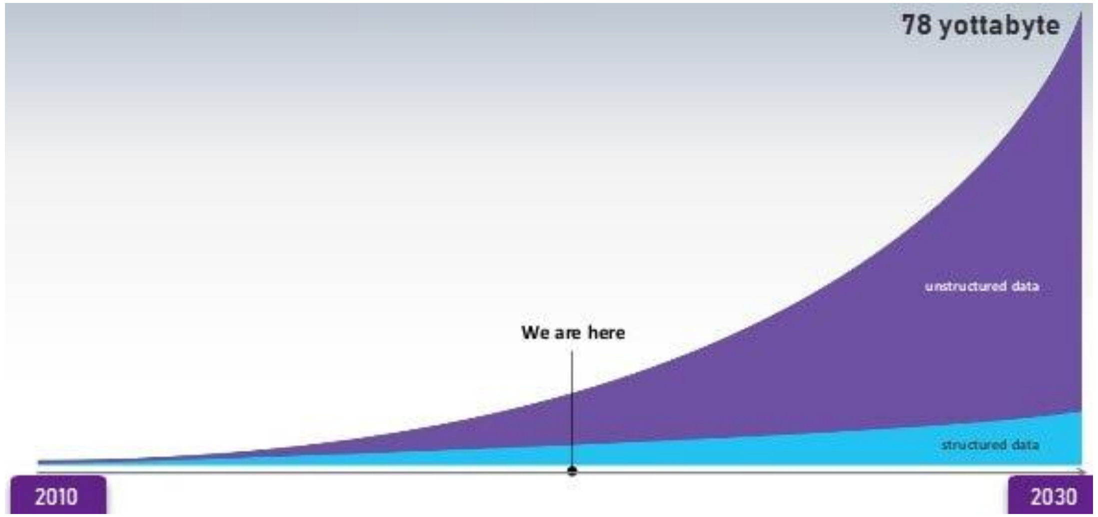
Učinkovitost bezga potrjuje tudi pred kratkim opravljena metaanaliza vseh razpoložljivih študij. Avtorji celo navajajo, da je bezeg dobra alternativa sinteznim zdravilom in dobra rešitev za preprečevanje zlorabe antibiotikov na tem področju (Ther Med 2019 Feb;42: 361–365. PMID: 30670267)

Challenges with data from heterogeneous sources

- Data is ‘dirty’ (different units, codings, standards, unreliable)
- Data can be of different types:
  - Structured (fixed format)
  - Unstructured (unfixed format, e.g. textual data, images, videos etc.)
  - Semi-structured (structured in form but without definition in relational DBMS, e.g. XML file)
- Data are missing (solution: discarding or imputation)
- Data are becoming ‘big’

Why Big Data are so relevant?

- BASIC DEFINITION: Big Data means data that are huge in size and yet growing exponentially with time.
  - They are collected by apps, gadgets, social media, IoT etc.
- WHY NEEDED: *Big data is necessary to isolate hidden patterns and to find answers without overfitting the data. (Wayne Thompson)*



INTERESTING: A yottabyte is the largest unit approved as a standard size by the International System of Units. 1 YB is approx. a million trillion megabytes.

Relevant ongoing bigger projects

- H2020 FNS-Cloud (https://www.fns-cloud.eu) – Food, nutrition, security cloud as part of EOSC
- EFSA CAFETERIA – semantic resources to support EFSA in automated extraction of information on food safety from literature
- H2020 COMFOCUS (https://comfocus.eu) – project trying to find an answer to our second question (*“How to achieve an optimal nutrition and what are current obstacles?”*)
- Tome’s Postdoc project MrBEC (http://cs.ijs.si/project/mrbec/) - ambitious project on advanced approaches for benchmarking in evolutionary computation

Organisation of international events

- AI & Food track at the Applied Machine Learning Days, EPFL (https://appliedmldays.org)
- BIOMA 2022 - The 10th International Conference on Bioinspired Optimization Methods and Their Applications, Maribor, 17-18 November 2022 (https://bioma2022.um.si/Committees/)

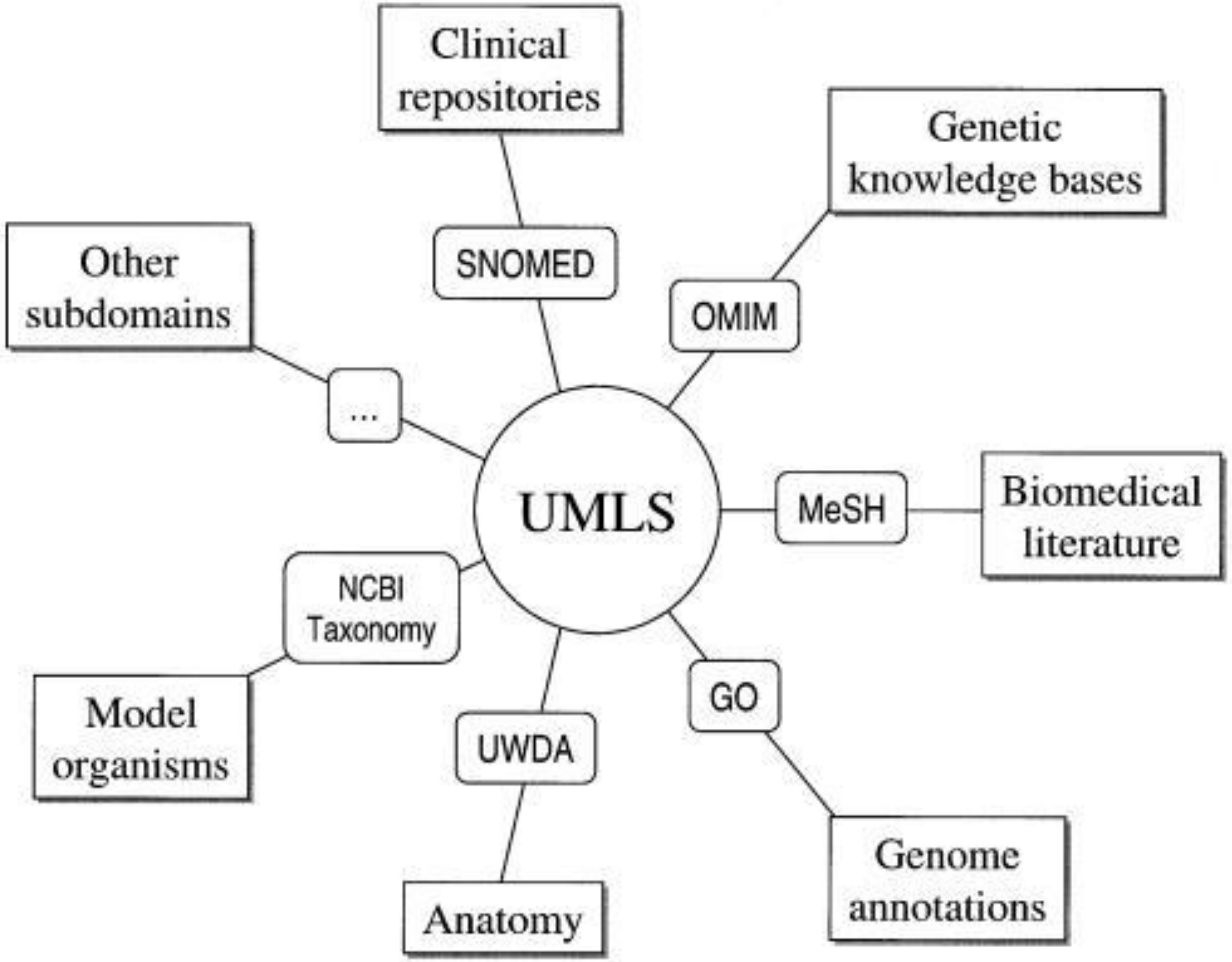
Food and biomedical semantic resources

Tome Eftimov

Computer Systems Department

Jožef Stefan Institute, Ljubljana

Biomedical semantic resources



Biomedical semantic resources

UMLS Points of Interest

Source	Number of Concepts
ICD09CM	20,997
ICD10CM	98,178
CPT	9,526
Meta CPT	1,036
HCPCS	5,651
CDT	587
RxNorm	204,081
NCI	90,135
DSM-IV	452
LOINC 236	140,633
CCS	1106
Snowmed CT	324,494

The UMLS "is organized by concept. One of its primary purposes is to connect different names for the same concept from many different vocabularies."

UMLS Concept Unique Identifier (cui) **UMLS Description**

C0155870 → "Influenza with pneumonia"

ICD09CM: 487.0

ICD10CM: J11.00

SnowmedCT: 195921001

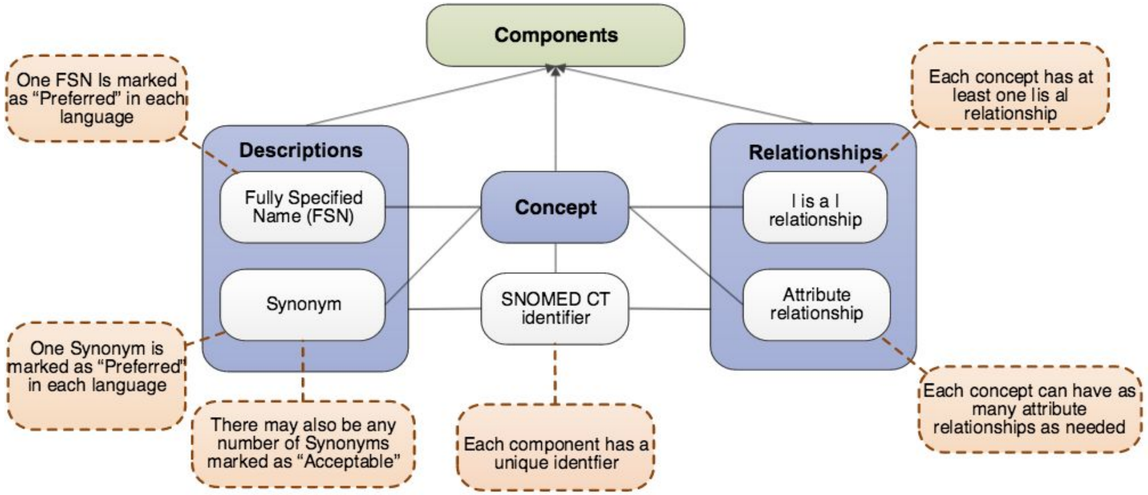
This is a good source for building code lists with associated and standardized descriptions.

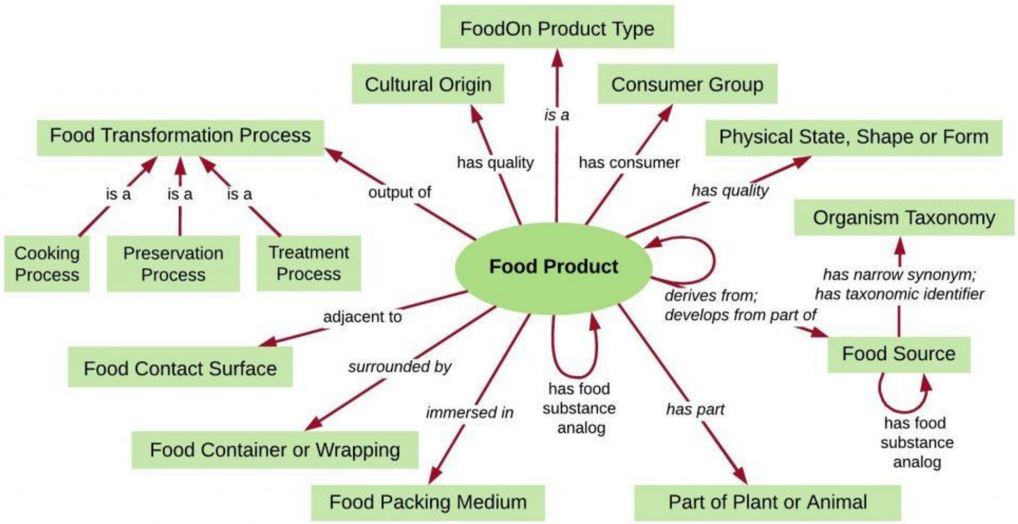
International Classification of Diseases (ICD)

Differences Between ICD-9-CM and ICD-10 Code Sets		
	ICD-9-CM	ICD-10 code sets
Procedure	3,824 codes	71,924 codes
Diagnosis	14,025 codes	69, 823 codes

ICD-10 Code Structure Changes (selected details)		
	Old	New
Diagnosis Structure	ICD-9-CM • 3 -5 characters • First character is numeric or alpha • Characters 2-5 are numeric	ICD-10-CM • 3 -7 characters • Character 1 is alpha • Character 2 is numeric • Characters 3 – 7 can be alpha or numeric
Procedure Structure	ICD-9-CM • 3-4 characters • All characters are numeric • All codes have at least 3 characters	ICD-10-PCS • ICD-10-PCS has 7 characters • Each can be either alpha or numeric • Numbers 0-9; letters A-H, J-N, P-Z

Food semantic resources





FoodEx2 efsa

Selecting the appropriate UMLS concepts about our study

- Selecting the appropriate semantic groups/semantic types
- Identify all concepts (identifiers) from the selected groups/types (MRSTY.RFF table)
- Identify all concepts’ names and synonyms for the selected identifiers (MRCONSO.RFF)

```
DISO|Disorders|T020|Acquired Abnormality
DISO|Disorders|T190|Anatomical Abnormality
DISO|Disorders|T049|Cell or Molecular Dysfunction
DISO|Disorders|T019|Congenital Abnormality
DISO|Disorders|T047|Disease or Syndrome
DISO|Disorders|T050|Experimental Model of Disease
DISO|Disorders|T033|Finding
DISO|Disorders|T037|Injury or Poisoning
DISO|Disorders|T048|Mental or Behavioral Dysfunction
DISO|Disorders|T191|Neoplastic Process
DISO|Disorders|T046|Pathologic Function
DISO|Disorders|T184|Sign or Symptom
GENE|Genes & Molecular Sequences|T087|Amino Acid Sequence
GENE|Genes & Molecular Sequences|T088|Carbohydrate Sequence
GENE|Genes & Molecular Sequences|T028|Gene or Genome
GENE|Genes & Molecular Sequences|T085|Molecular Sequence
GENE|Genes & Molecular Sequences|T086|Nucleotide Sequence
GEOG|Geographic Areas|T083|Geographic Area
LIVB|Living Beings|T100|Age Group
LIVB|Living Beings|T011|Amphibian
LIVB|Living Beings|T008|Animal
LIVB|Living Beings|T194|Archaeon
LIVB|Living Beings|T007|Bacterium
LIVB|Living Beings|T012|Bird
LIVB|Living Beings|T204|Eukaryote
LIVB|Living Beings|T099|Family Group
LIVB|Living Beings|T013|Fish
LIVB|Living Beings|T004|Fungus
LIVB|Living Beings|T096|Group
LIVB|Living Beings|T016|Human
LIVB|Living Beings|T015|Mammal
LIVB|Living Beings|T001|Organism
LIVB|Living Beings|T101|Patient or Disabled Group
LIVB|Living Beings|T002|Plant
LIVB|Living Beings|T098|Population Group
LIVB|Living Beings|T097|Professional or Occupational Group
LIVB|Living Beings|T014|Reptile
LIVB|Living Beings|T010|Vertebrate
```

Example: MRSTY table

Col.	Description
CUI	Unique identifier of concept
TUI	Unique identifier of Semantic Type
STN	Semantic Type tree number
STY	Semantic Type. The valid values are defined in the Semantic Network.
ATUI	Unique identifier for attribute
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

Sample Record

C0001175|T047|B2.2.1.2.1|Disease or Syndrome|AT17683839|2304|

Example: MRCONSO table

Concept Names and Sources (File = MRCONSO.RRF)

Col.	Description
CUI	Unique identifier for concept
LAT	Language of term
TS	Term status
LUI	Unique identifier for term
STT	String type
SUI	Unique identifier for string
ISPREF	Atom status - preferred (Y) or not (N) for this string within this concept
AUI	Unique identifier for atom - variable length field, 8 or 9 characters
SAUI	Source asserted atom identifier [optional]
SCUI	Source asserted concept identifier [optional]
SDUI	Source asserted descriptor identifier [optional]
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
TTY	Abbreviation for term type in source vocabulary, for example PN (Metathesaurus Preferred Name) or CD (Clinical Drug). Possible values are listed on the Abbreviations Used in Data Elements page.
CODE	Most useful source asserted identifier (if the source vocabulary has more than one identifier), or a Metathesaurus-generated source entry identifier (if the source vocabulary has none)
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag. Values = O, E, Y, or N O: All obsolete content, whether they are obsolesced by the source or by NLM. These will include all atoms having obsolete TTYs, and other atoms becoming obsolete that have not acquired an obsolete TTY (e.g. RxNorm SCDs no longer associated with current drugs, LNC atoms derived from obsolete LNC concepts). E: Non-obsolete content marked suppressible by an editor. These do not have a suppressible SAB/TTY combination. Y: Non-obsolete content deemed suppressible during inversion. These can be determined by a specific SAB/TTY combination explicitly listed in MRRANK. N: None of the above Default suppressibility as determined by NLM (i.e., no changes at the Suppressibility tab in MetamorphoSys) should be used by most users, but may not be suitable in some specialized applications. See the MetamorphoSys Help page for information on how to change the SAB/TTY suppressibility to suit your requirements. NLM strongly recommends that users not alter editor-assigned suppressibility, and MetamorphoSys cannot be used for this purpose.
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

Example: MRCONSO table

Sample Records

C0001175|ENG|P|L0001175|VO|S0010340|Y|A0019182||M0000245|D000163|MSH|PM|D000163|Acquired Immunodeficiency Syndromes|0|N||

C0001175|ENG|S|L0001842|PF|S0011877|N|A2878223|103840012|62479008||SNOMEDCT_US|PT|62479008|AIDS|9|N|2304|

C0001175|ENG|P|L0001175|VO|S0354232|Y|A2922342|103845019|62479008||SNOMEDCT_US|SY|62479008|Acquired immunodeficiency syndrome|9|N|2304|

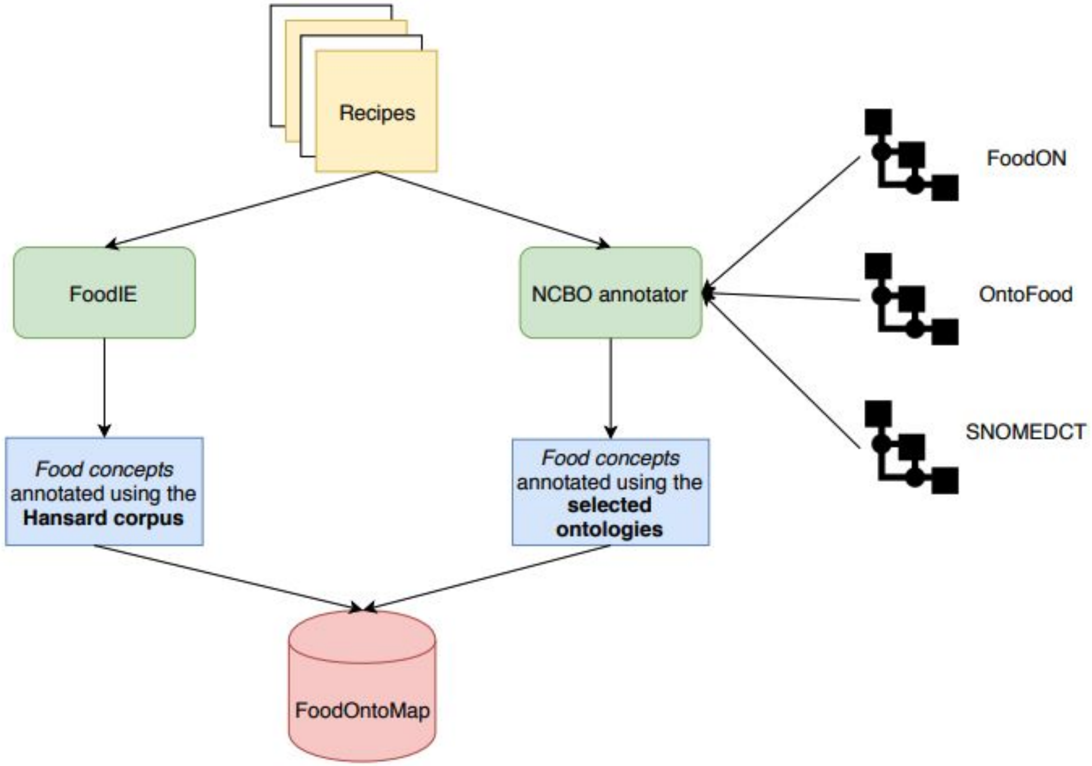
C0001175|FRE|S|L0162173|PF|S0226654|Y|A27478989||M0000245|D000163|MSHFRE|ET|D000163|SIDA|3|N||

C0001175|RUS|S|L0904943|PF|S1108760|Y|A13488500||M0000245|D000163|MSHRUS|SY|D000163|SPID|3|N||

FoodOntoMap

Popovski, G., Seljak, B. K., & Eftimov, T. (2020). A survey of named-entity recognition methods for food information extraction. IEEE Access, 8, 31586-31594.

Popovski, G., Korousic-Seljak, B., & Eftimov, T. (2019). FoodOntoMap: Linking Food Concepts across Different Food Ontologies. In KEOD (pp. 195-202).



FoodBase	FOODON	SNOMEDCT	OF	RCD	MESH	SNMI	NDDF
A000000	B000001;	C000001;	NULL	E000000;	F000000;	G000000;	H000000;
A000007	B000008;	C000010;	NULL	E000010;	NULL	NULL	NULL
A000014	B000012;	NULL	D000003;	NULL	NULL	NULL	NULL
A000016	B000011;	C000012;	D000002;	E000012;	F000002;	NULL	NULL
A000032	B000023;	C000020;	D000007;	E000020;	NULL	NULL	NULL
A000046	B000027;	C000026;	NULL	E000023;	F000009;	NULL	NULL
A000049	B000031;	C000031;	D000011;	E000028;	NULL	NULL	NULL
A000059	B000034;	C000035;	NULL	NULL	NULL	NULL	NULL
A000076	B000026;	C000023;	NULL	NULL	NULL	NULL	NULL
A000083	B000023;	C000020;	D000007;	E000020;	NULL	NULL	NULL
A000121	B000061;	NULL	NULL	NULL	NULL	NULL	NULL
A000123	B000065;B000012;	NULL	D000003;	NULL	NULL	NULL	NULL
A000141	B000069;	NULL	NULL	NULL	NULL	NULL	NULL

Food and biomedical rule-based information extraction (IE) from textual data

Gordana Ispirova, Matevž Ogrinc
Computer Systems Department
Jožef Stefan Institute, Ljubljana

Information Extraction (IE)

Named-Entity Recognition (NER)

Excessive salt intake has been associated with a higher incidence of heart disease.

Named-Entity Linking (NEL)

Excessive salt **[00002 (FOODB)]** intake has been associated with a higher incidence of heart disease **[0001 (UMLS)]**.

Rule-based NERs

1. Creating dictionaries
2. NER with Spacy
3. NER with NCBO

Creating dictionaries

1. FoodEx2 based dictionary

```
Term|Label
Teff grain|A000A
Finger millet grain|A000B
African millet grain|A000C
Foxtail millet grain|A000D
Little millet grain|A000E
Oat and similar-|A000F
Oat grain|A000G
Oat grain, red|A000H
Grains and grain-based products|A000J
```

Creating dictionaries

2. FooDB based dictionary
    a. NCBI taxonomy id
    b. FooDB id

```
Term|Scientific name|Label
Angelica|Angelica keiskei|357850
Savoy cabbage|Brassica oleracea var. sabauda|1216010
Silver linden|Tilia argentea|nan
Kiwi|Actinidia chinensis|3625
Allium|Allium|4678
Garden onion|Allium cepa|4679
Leek|Allium porrum|nan
Garlic|Allium sativum|4682
Chives|Allium schoenoprasum|74900
```

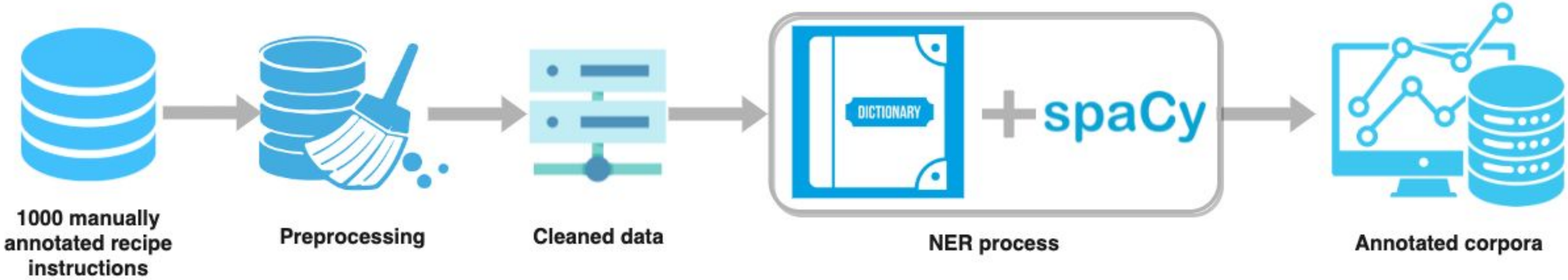
```
Term|Scientific name|Label
Angelica|Angelica keiskei|FOOD00001
Savoy cabbage|Brassica oleracea var. sabauda|FOOD00002
Silver linden|Tilia argentea|FOOD00003
Kiwi|Actinidia chinensis|FOOD00004
Allium|Allium|FOOD00005
Garden onion|Allium cepa|FOOD00006
Leek|Allium porrum|FOOD00007
Garlic|Allium sativum|FOOD00008
Chives|Allium schoenoprasum|FOOD00009
```

Creating dictionaries

3. UMLS Metathesaurus Data dictionary (MRSTY and MRCONSO tables)

```
Term|Label
alkoholické nápoje|C0001967
Alcoholische drank|C0001967
Alcoholische dranken|C0001967
Alcoholic Beverages|C0001967
Alcoholic Beverages|C0001967
Alcoholic Beverages|C0001967
Alcoholic beverages|C0001967
Alcoholic beverages|C0001967
alcoholic beverages|C0001967
Alcoholic beverage|C0001967
Alcoholic beverage|C0001967
Beverage, Alcoholic|C0001967
```

NER with Spacy



Evaluation results

Preheat skillet over medium heat. Generously **butter** one side of a slice of **bread**. Place **bread** butter-side-down onto skillet bottom and add 1 slice of **cheese**. Butter a second slice of **bread** on one side and place butter-side-up on top of **sandwich**. Grill until lightly browned and flip over; continue grilling until **cheese** is melted. Repeat with remaining 2 slices of **bread, butter** and slice of **cheese.**

Dictionary		Captured	Uncaptured
FoodEx2		'butter', 'bread', 'bread', 'cheese', 'bread', 'cheese', 'bread', 'butter', 'cheese'	'sandwich'
FooDB	FoodB id	'butter', 'cheese', 'cheese', 'butter', 'cheese'	'bread', 'bread', 'bread', 'sandwich', 'bread'
	NCBI id	'butter', 'cheese', 'cheese', 'butter', 'cheese'	'bread', 'bread', 'bread', 'sandwich', 'bread'
UMLS		'sandwich', 'butter','butter'	'bread', 'bread', 'cheese', 'bread', 'cheese', 'bread', 'cheese'

NER WITH NCBO

- Ontologies
- NCBO annotator demo
- Results

Ontologies

- **FOODON**
- MESH
- OCHV
- RCD
- CRISP
- **OF**
- LOINC
- MEDLINEPLUS
- **SNOMEDCT**
- NDDF
- NDFRT
- PDQ
- RCD
- RXNORM
- SNMI
- VANDF

Ontologies

What are we looking for?

Baking banana bread is one of my favorites, and I love nothing more than enjoying a slice with a nice cup of coffee. This was the inspiration for my recipe, which features a coffee infused loaf and a rich caramel glaze.

SNOMEDCT

Details | Visualization | Notes (0) | Class Mappings (24)

Preferred Name	Banana
Synonyms	Banana (substance)
ID	http://purl.bioontology.org/ontology/SNOMEDCT/256307007
Active	1
Active ingredient of	Banana diagnostic allergen extract
altLabel	Banana (substance)
CASE SIGNIFICANCE ID	900000000000448009
Causative agent of	Allergy to banana
CTV3ID	X79Q4
cui	C0004722
DEFINITION STATUS ID	900000000000074008
Effective time	20020131
notation	256307007
prefLabel	Banana
Subset member	900000000000509007~ACCEPTABILITYID~900000000000548007 900000000000508004~ACCEPTABILITYID~900000000000548007 900000000000497000~MAPTARGET~X79Q4
tui	T168
Type ID	900000000000013009 900000000000003001
subClassOf	Fresh fruit

FOODON

Class: Musa acuminata

Term IRI: http://purl.obolibrary.org/obo/NCBITaxon_4641

Annotations

- **database_cross_reference:** GC_ID:1
- **has_alternative_id:** NCBITaxon:214695
- **has_exact_synonym:** banana; dessert bananas; sweet banana; dwarf banana
- **has_obo_namespace:** ncbi_taxonomy
- **has_rank:** species
- **has_related_synonym:** Musa nana; Musa AA Group; Musa acuminata AA Group

Class Hierarchy

Thing
+ root

Ontologies

Is the annotator program limited to one domain?

```
clna|T201|Clinical Attribute
clnd|T200|Clinical Drug
cnce|T077|Conceptual Entity
comd|T049|Cell or Molecular Dysfunction
crbs|T088|Carbohydrate Sequence
diap|T060|Diagnostic Procedure
dora|T056|Daily or Recreational Activity
drdd|T203|Drug Delivery Device
dsyn|T047|Disease or Syndrome
edac|T065|Educational Activity
eehu|T069|Environmental Effect of Humans
elii|T196|Element, Ion, or Isotope
emod|T050|Experimental Model of Disease
emst|T018|Embryonic Structure
enty|T071|Entity
enzy|T126|Enzyme
euka|T204|Eukaryote
evnt|T051|Event
famg|T099|Family Group
ffas|T021|Fully Formed Anatomical Structure
fish|T013|Fish
fndg|T033|Finding
fngs|T004|Fungus
food|T168|Food
ftcn|T169|Functional Concept
genf|T045|Genetic Function
geoa|T083|Geographic Area
gngm|T028|Gene or Genome
gora|T064|Governmental or Regulatory Activity
grpa|T102|Group Attribute
grup|T096|Group
```

```
aapp|T116|Amino Acid, Peptide, or Protein
acab|T020|Acquired Abnormality
acty|T052|Activity
aggp|T100|Age Group
amas|T087|Amino Acid Sequence
amph|T011|Amphibian
anab|T190|Anatomical Abnormality
anim|T008|Animal
anst|T017|Anatomical Structure
antb|T195|Antibiotic
arch|T194|Archaeon
bacs|T123|Biologically Active Substance
bact|T007|Bacterium
bdsu|T031|Body Substance
bdsy|T022|Body System
bhvr|T053|Behavior
biof|T038|Biologic Function
bird|T012|Bird
blor|T029|Body Location or Region
bmod|T091|Biomedical Occupation or Discipline
bodm|T122|Biomedical or Dental Material
bpoc|T023|Body Part, Organ, or Organ Component
bsoj|T030|Body Space or Junction
celc|T026|Cell Component
celf|T043|Cell Function
cell|T025|Cell
cgab|T019|Congenital Abnormality
chem|T103|Chemical
chvf|T120|Chemical Viewed Functionally
chvs|T104|Chemical Viewed Structurally
clas|T185|Classification
```

NCBO annotator demo

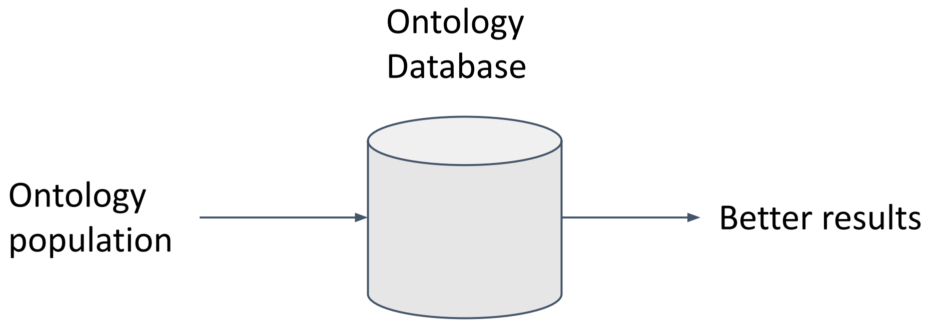
Results

Where is the problem?

Baking banana bread is one of my favorites, and I love nothing more than enjoying a slice with a nice cup of coffee. This was the inspiration for my recipe, which features a coffee infused loaf and a rich caramel glaze.

Results

	NCBO (SNOMED CT)	NCBO (OntoFood)	NCBO(FoodON)
F_1 Score	63.75%	32.62%	63.90%
Precision	91.53%	85.48%	79.22%
Recall	48.91%	20.16%	53.54%



Domain experts' information extraction evaluation using a Human-Computer Interaction tool

Eva Valenčič

Computer Systems Department

Jožef Stefan Institute, Ljubljana

tetramax

Human-computer interaction tool

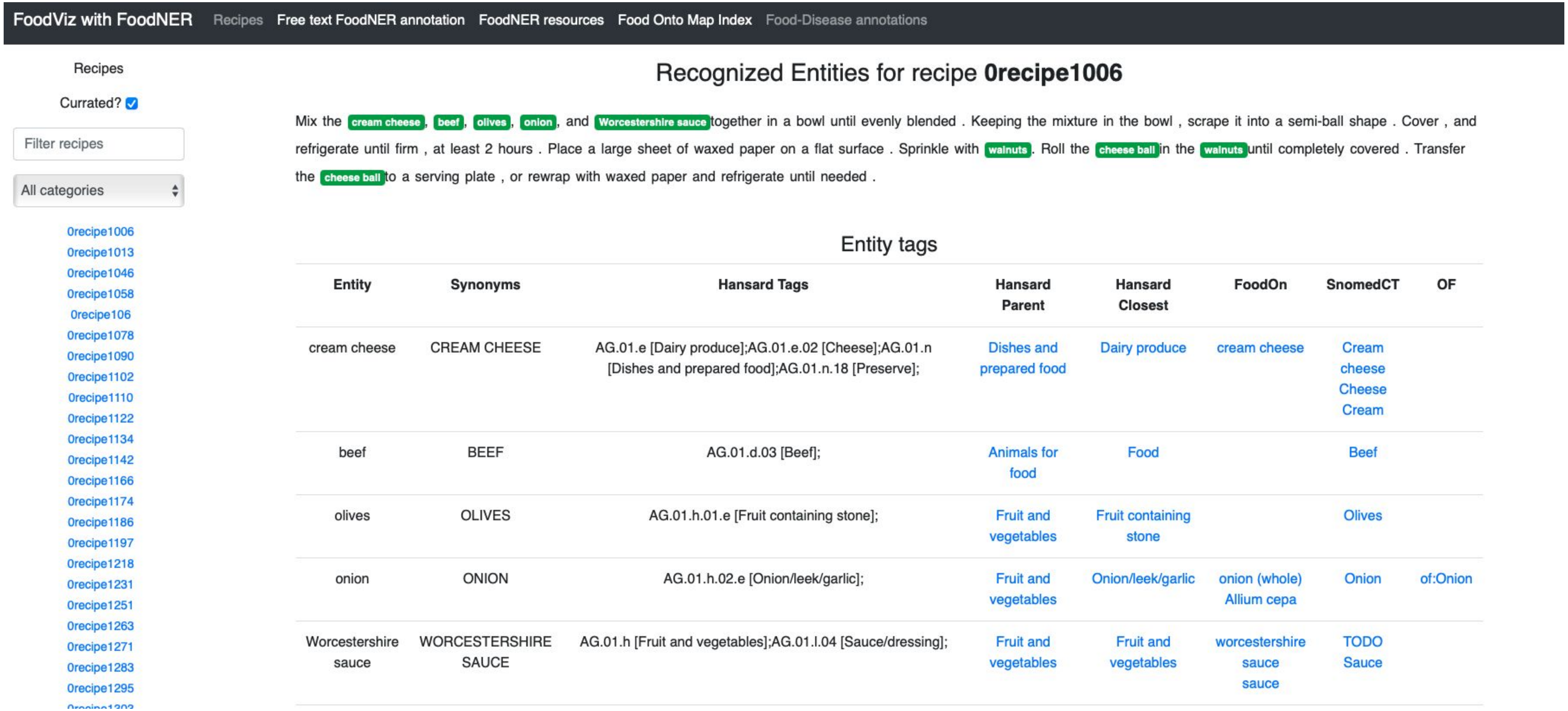
FoodViz

- computer methods do not always give perfect results
    - domain experts can manually make corrections

- tool for visualization of automatically annotated text on foods

- allows to explore the existing link between food standards (SNOMED CT, FoodOn, etc.)

FoodViz



FoodViz with FoodNER Recipes Free text FoodNER annotation FoodNER resources Food Onto Map Index Food-Disease annotations

Recipes

Currated? ☑

Filter recipes

All categories

0recipe1006
0recipe1013
0recipe1046
0recipe1058
0recipe106
0recipe1078
0recipe1090
0recipe1102
0recipe1110
0recipe1122
0recipe1134
0recipe1142
0recipe1166
0recipe1174
0recipe1186
0recipe1197
0recipe1218
0recipe1231
0recipe1251
0recipe1263
0recipe1271
0recipe1283
0recipe1295

Recognized Entities for recipe **0recipe1006**

Mix the cream cheese , beef , olives , onion , and Worcestershire sauce together in a bowl until evenly blended . Keeping the mixture in the bowl , scrape it into a semi-ball shape . Cover , and refrigerate until firm , at least 2 hours . Place a large sheet of waxed paper on a flat surface . Sprinkle with walnuts . Roll the cheese ball in the walnuts until completely covered . Transfer the cheese ball to a serving plate , or rewrap with waxed paper and refrigerate until needed .

Entity tags

Entity	Synonyms	Hansard Tags	Hansard Parent	Hansard Closest	FoodOn	SnomedCT	OF
cream cheese	CREAM CHEESE	AG.01.e [Dairy produce];AG.01.e.02 [Cheese];AG.01.n [Dishes and prepared food];AG.01.n.18 [Preserve];	Dishes and prepared food	Dairy produce	cream cheese	Cream cheese Cheese Cream	
beef	BEEF	AG.01.d.03 [Beef];	Animals for food	Food		Beef	
olives	OLIVES	AG.01.h.01.e [Fruit containing stone];	Fruit and vegetables	Fruit containing stone		Olives	
onion	ONION	AG.01.h.02.e [Onion/leek/garlic];	Fruit and vegetables	Onion/leek/garlic	onion (whole) Allium cepa	Onion	of:Onion
Worcestershire sauce	WORCESTERSHIRE SAUCE	AG.01.h [Fruit and vegetables];AG.01.l.04 [Sauce/dressing];	Fruit and vegetables	Fruit and vegetables	worcestershire sauce sauce	TODO Sauce	

FoodViz

Recipes

Currated? ☐

Filter recipes

All categories

0recipe10
0recipe100
0recipe1000
0recipe1001
0recipe1002
0recipe1003
0recipe1005
0recipe1007
0recipe1008
0recipe1009
0recipe101
0recipe1010
0recipe1011
0recipe1012
0recipe1014
0recipe1015
0recipe1016
0recipe1017
0recipe1018
0recipe1019
0recipe102
0recipe1020
0recipe1021
0recipe1022
0recipe1023
0recipe1024
0recipe1026
0recipe1027
0recipe1028

Recognized Entities for recipe **0recipe101**

Preheat oven to 350 degrees F (175 degrees C) . In a 12 inch skillet over medium heat , cook and stir the garlic and white parts of the green onions in canola oil until tender . Mix in shredded chicken , salt and pepper . Toss until well coated with oil . Stir in the salsa . Arrange tortilla chips on a large baking sheet . Spoon the chicken mixture over tortilla chips . Top with Cheddar / Monterey Jack cheese blend and tomato . Bake in the preheated oven 10 minutes , or until cheese has melted . Remove from heat and sprinkle with green onion tops before serving .

Entity tags

Entity	Synonyms	Hansard Tags	Hansard Parent	Hansard Closest	FoodOn	SnomedCT	OF
garlic x	GARLIC	AG.01.h.02.e [Onion/leek/garlic]	Fruit and vegetables	Onion/leek/garlic	Allium sativum	Garlic	of:Garlic
green onions x		AG.01.h.02.e [Onion/leek/garlic]					
canola oil x	CANOLA OIL	AG.01.f [Fat/oil]	Fat/oil	Fat/oil	oil canola oil	Canola oil Rapeseed oil	
shredded chicken x	SHREDDED CHICKEN CHICKEN	AG.01.d.06 [Fowls]	Animals for food	Fowls	Gallus gallus chicken	Chicken - meat	of:Chicken
salt x	SALT	AG.01.l.01 [Salt]	Additive	Food	salt		
pepper x		AG.01.l.03 [Spice]					
oil x	OIL	AG.01.f [Fat/oil]	Food	Food	oil		
salsa x	SALSA	AG.01.l.04 [Sauce/dressing]	Additive	Sauce/dressing	salsa		

Human-computer interaction tool

http://foodviz.env4health.finki.ukim.mk/#/recipes

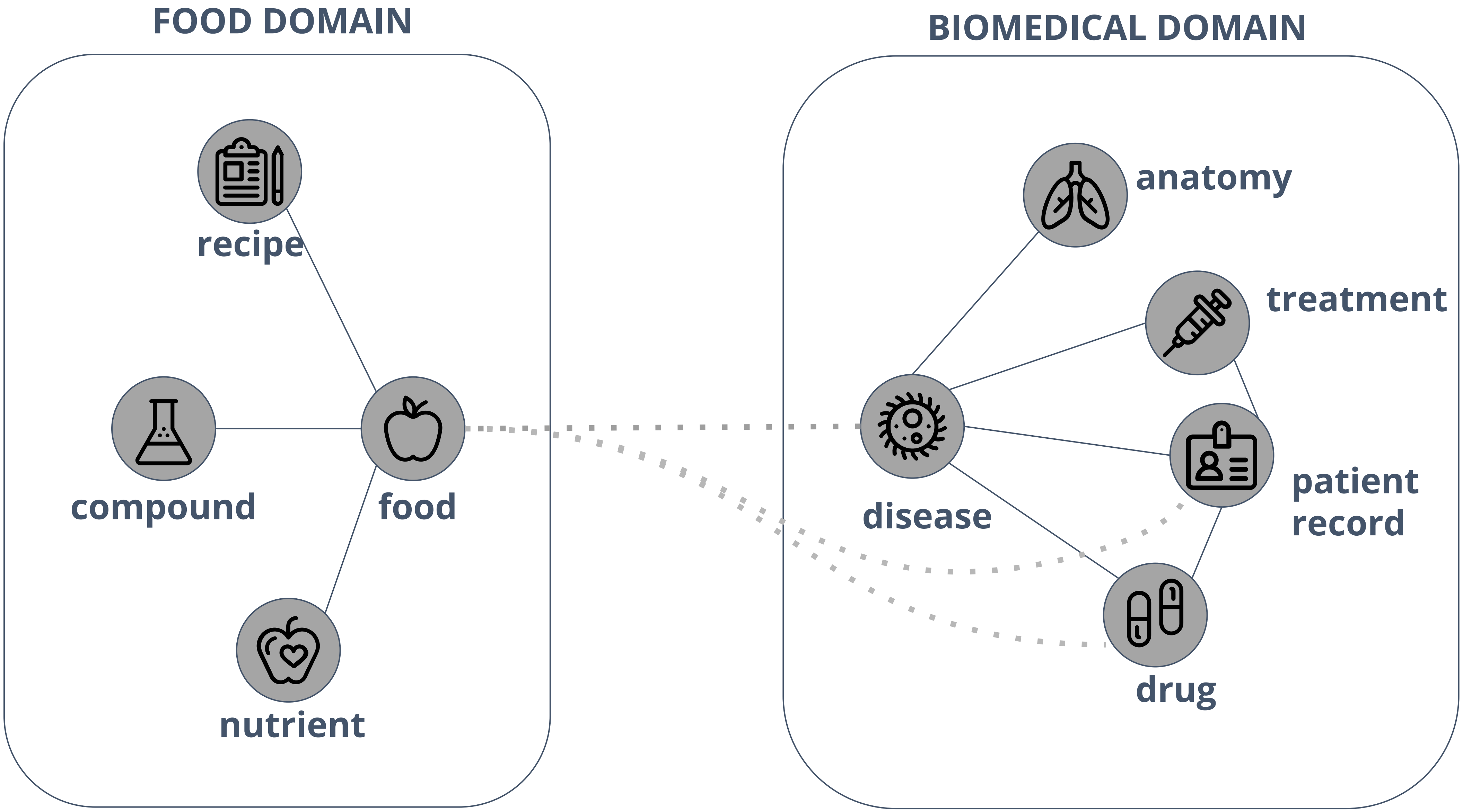
Food, chemical, and disease relation extraction

Gjorgjina Cenikj

Computer Systems Department

Jožef Stefan Institute, Ljubljana

Knowledge Graphs



IE pipelines for Knowledge Graph Construction

Example

Excessive salt intake has been associated with a higher incidence of heart disease.

SAFFRON

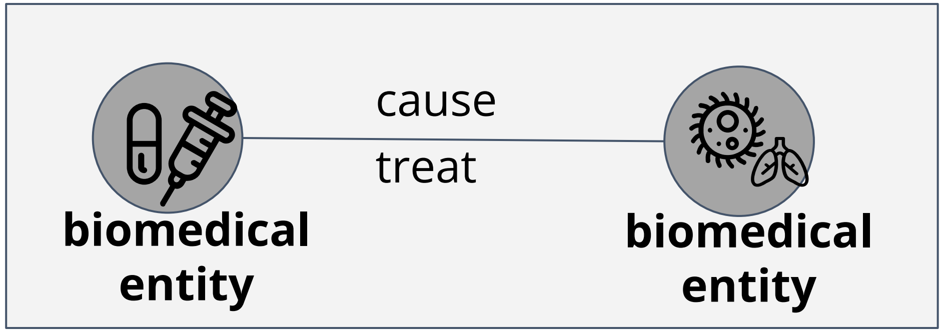
Tran**S**fer le**A**rning **F**or **F**ood-Disease **R**elati**O**n extractio**N**

- RE method for detecting **cause** and **treat** relations
- Main challenge: lack of annotated data with relations between food and disease entities
- Proposed solution: use **transfer learning** to repurpose existing resources in the biomedical domain for the food domain

Gjorgjina Cenikj, Tome Eftimov, Barbara Koroušić Seljak. "SAFFRON: tranSfer leArning For Food-Disease RelatiOn extractioN", Annual Conference of the North American Chapter of the Association for Computational Linguistics 2021

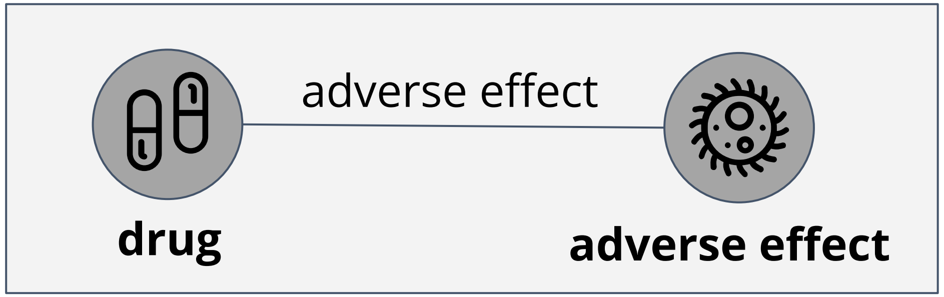
SAFFRON

Transfer Learning Data



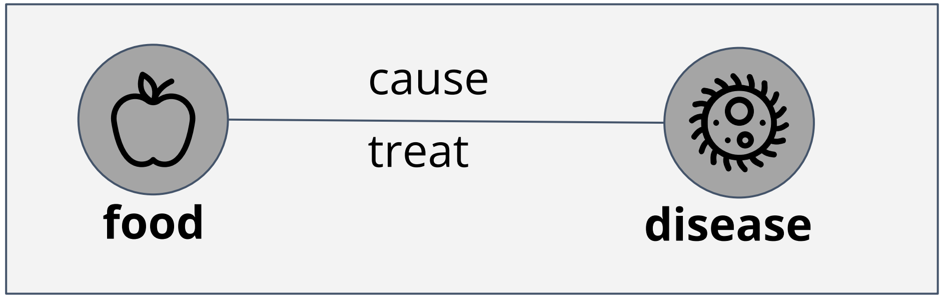
The CrowdTruth dataset - source

~4000 sentences



Adverse Drug Events (ADE) dataset - source

~6800 sentences



The FoodDisease dataset - source and target

~600 sentences

SAFFRON
Preprocessing

ORIGINAL SENTENCE:
Several epidemiological and preclinical studies supported the protective effect of **coffee** on **Alzheimer's disease.**

ENTITY MASKING:
Several epidemiological and preclinical studies supported the protective effect of **XXX** on **YYY**

CONTEXT EXTRACTION:
supported the protective effect of **XXX** on **YYY**

SAFFRON

Models

- Performs fine-tuning of BERT, RoBERTa and BioBERT models on data annotated for relations between different biomedical entities
- Relation extraction treated as binary classification, with each model producing a 0/1 indicator of the existence of a single relation
- Best models achieve macro averaged F1 scores of 0.847 and 0.900 for the cause and treat relations, respectively.

FooDis
A food-disease relation mining pipeline

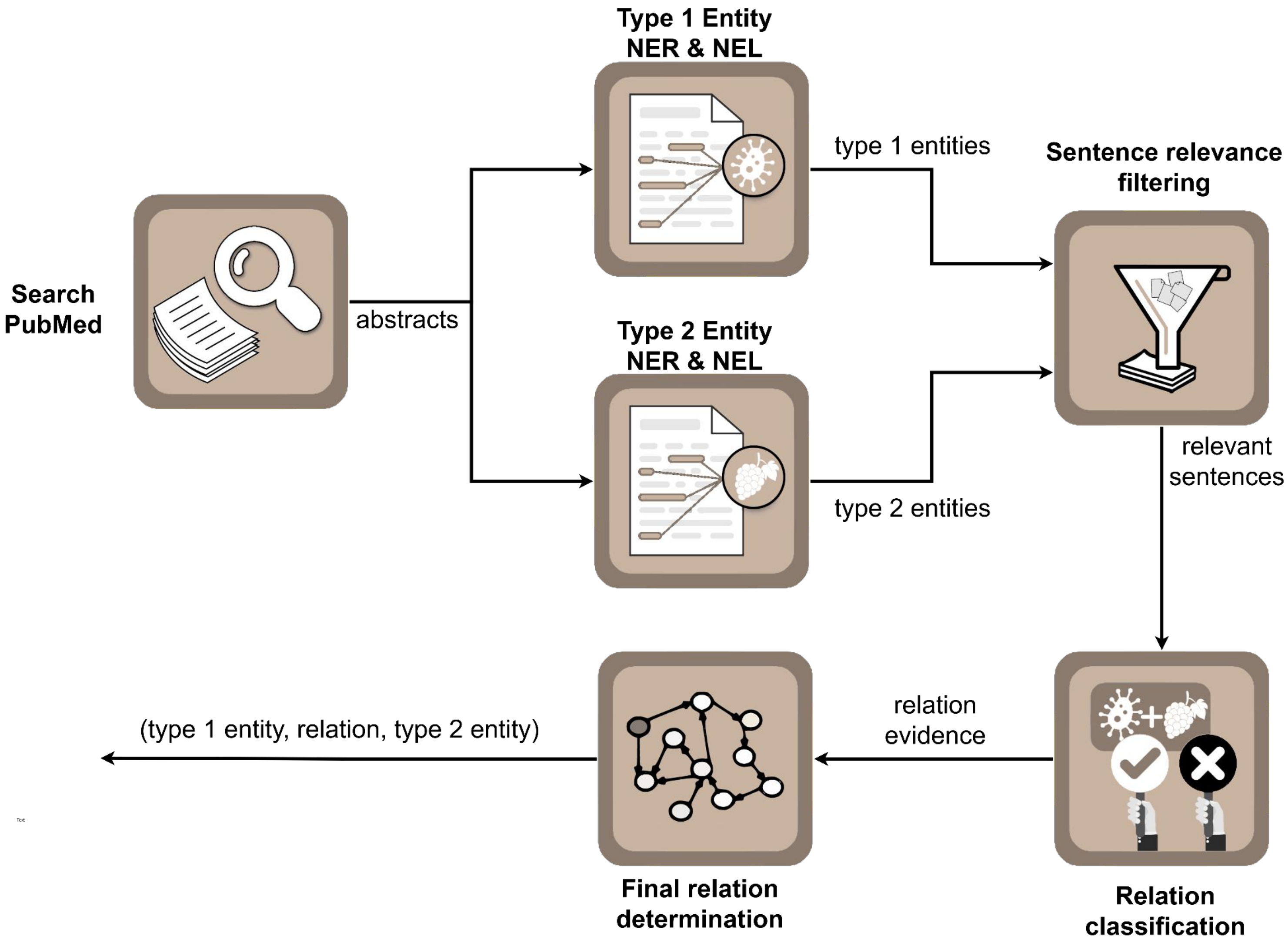
- Information extraction pipeline for semi-automatic relation mining
- Extracts **cause** or **treat** relations between **food** and **disease** entities from biomedical scientific literature
- Links entities to different knowledge bases in the biomedical and food domains

Gjorgjina Cenikj, Tome Eftimov, Barbara Koroušić Seljak. "FooDis: A food-disease relation mining pipeline", Journal of Medical Internet Research, 2021, In second round review

FooDis

Pipeline overview

	Type	Extractor
Entity 1	Food	Voting scheme
Entity 2	Disease	SABER-DISO
Relation	Cause/Treat	SAFFRON



SABER for biomedical NER&NEL

- Sequence Annotator for Biomedical Entities and Relations
- Biomedical NER&NEL tool, based on a BiLSTM-CRF neural architecture
- Pretrained NER models are provided for identifying diseases, genes, organisms and chemicals
- Disease entities are linked to concepts in the Disease Ontology, chemical entities are linked to PubChem

Limitations of food NER methods

Corpus-based methods:

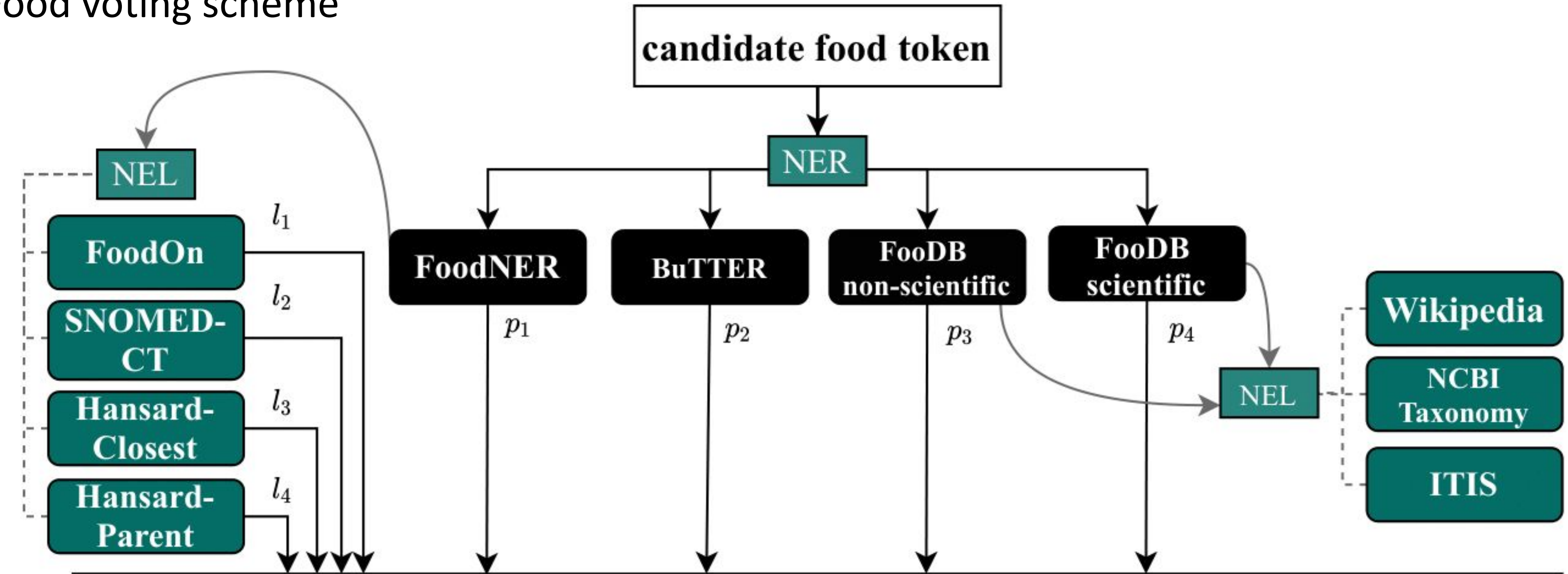
- Excessive salt intake has been shown to cause heart disease, while avocado oil consumption has been linked to lower risks of heart disease.

Dictionary-based methods:

- Excessive salt intake has been shown to cause heart disease, while avocado oil consumption has been linked to lower risks of heart disease.

FooDis

Food voting scheme



$$is\_food(token) = \begin{cases} true, & if \quad [\sum_{i=1}^{3} p_i >= 2 \vee p_4 == 1 \vee \sum_{i=1}^{4} l_i > 1] \\ & \wedge token \notin \{'food', 'foods', 'consumption', 'drug'\} \\ & \wedge contains\_noun(token) \\ false, & otherwise \end{cases}$$

FooDis

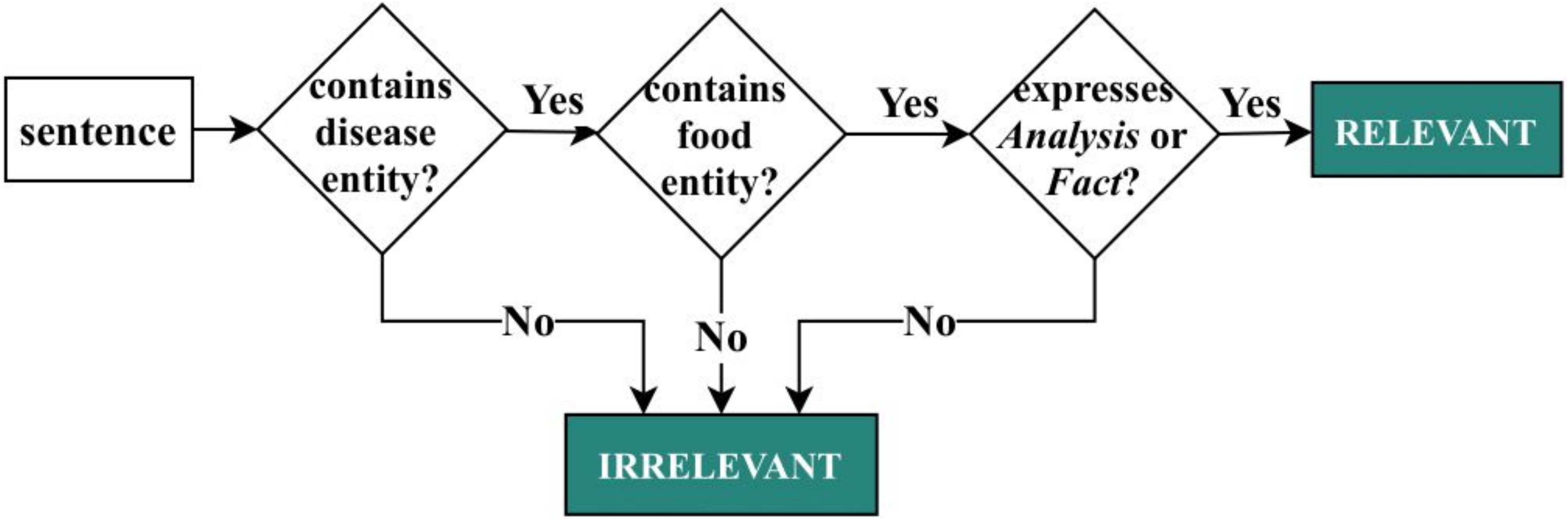
Sentence relevance filtering

Excessive **salt** intake increases the risk of **heart disease**.

Hypothesis	**We hypothesize that** excessive salt intake increases the risk of heart disease.	✗
Analysis	**Our results indicate that** excessive salt intake increases the risk of heart disease.	✓
Fact	**It has been shown that** excessive salt intake increases the risk of heart disease.	✓
Method	**We perform clinical trials and observe the reactions of 50 patients to test if** excessive salt intake increases the risk of heart disease.	✗

FooDis

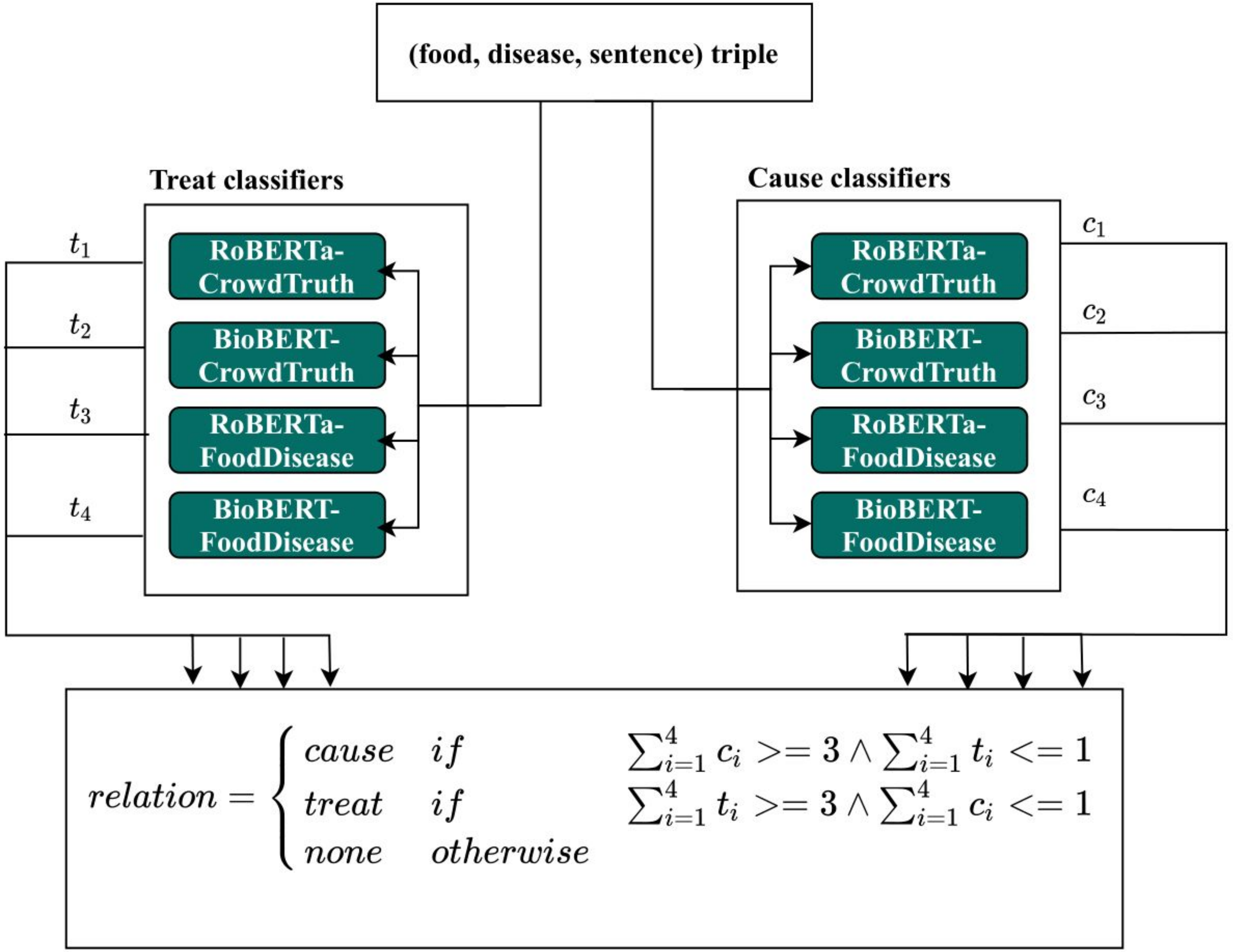
Sentence relevance filtering



FooDis

Relation extraction

A voting scheme implemented by combining 8 of the SAFFRON models, 4 per each of the 2 relations, *cause* and *treat*



FooDis

Number of extracted, linked relations

Table 1. Number of unique pairs linked with a cause relationship for each combination of food and disease resource.

	DO	SNOMED CT	UMLS	NCIT	OMIM	EFO	MESH
FoodOn	167	152	157	120	70	53	142
SNOMED CT	159	149	153	116	74	57	138
Hansard Closest	222	202	207	151	100	67	187
Hansard Parent	221	201	206	150	101	68	186
FooDB	148	134	140	109	66	50	122
ITIS	80	70	75	61	41	31	65
Wikipedia	139	125	131	103	63	47	113
NCBIT	82	72	77	63	43	33	66

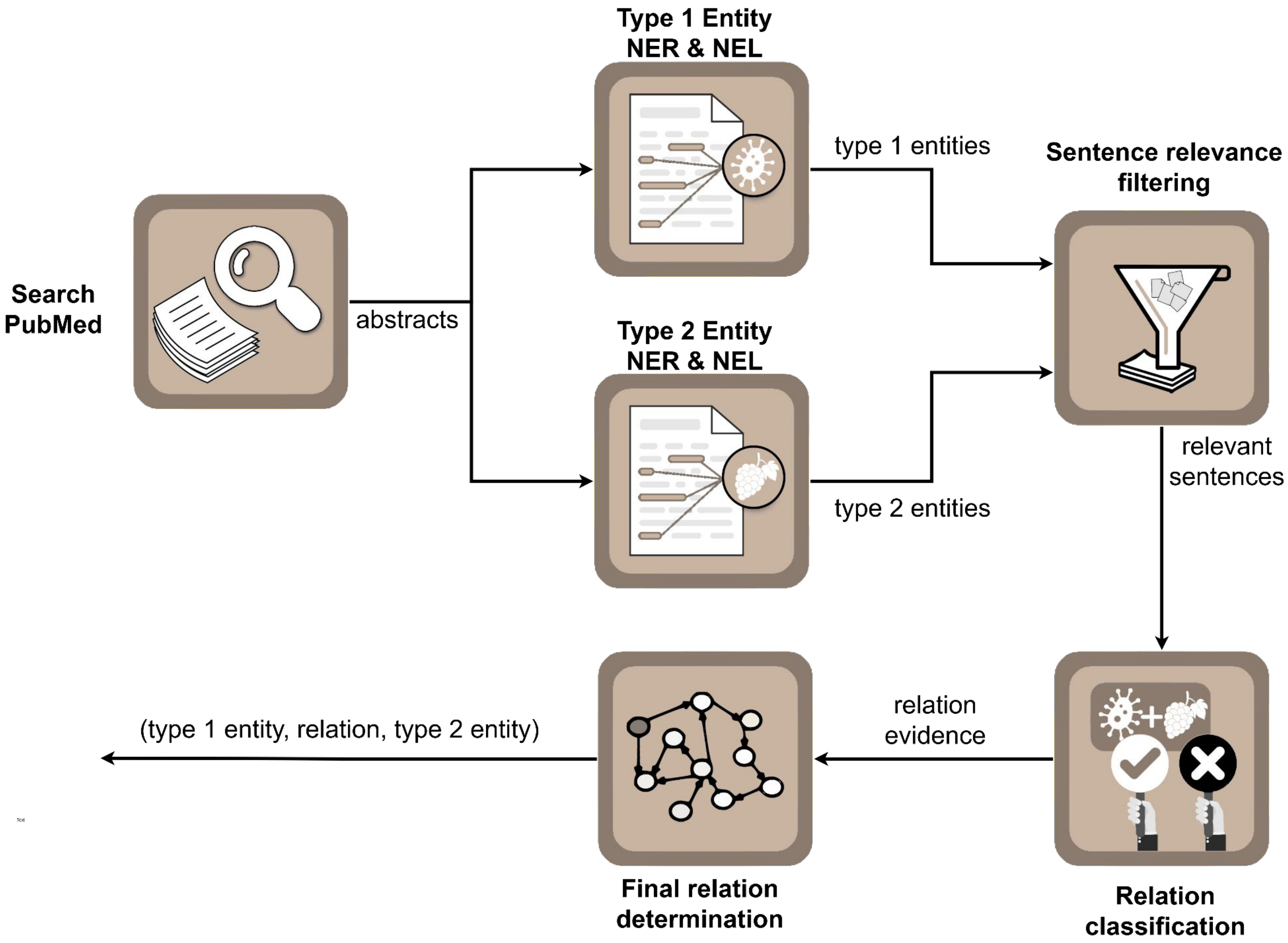
Table 2. Number of unique pairs linked with a treat relationship for each combination of food and disease resource.

	DO	SNOMED CT	UMLS	NCIt	OMIM	EFO	MESH
FoodOn	280	264	269	251	101	103	257
SNOMED CT	352	334	339	320	133	141	329
Hansard Closest	438	418	424	397	165	165	407
Hansard Parent	438	415	423	394	167	167	404
FooDB	503	472	484	432	173	184	464
ITIS	314	293	300	268	105	107	290
Wikipedia	423	396	405	363	144	147	389
NCBIT	313	292	299	267	104	107	289

FoodChem

Pipeline overview

	Type	Extractor
Entity 1	Food	Voting scheme
Entity 2	Chemical	SABER-CHED
Relation	Contains	FoodChem

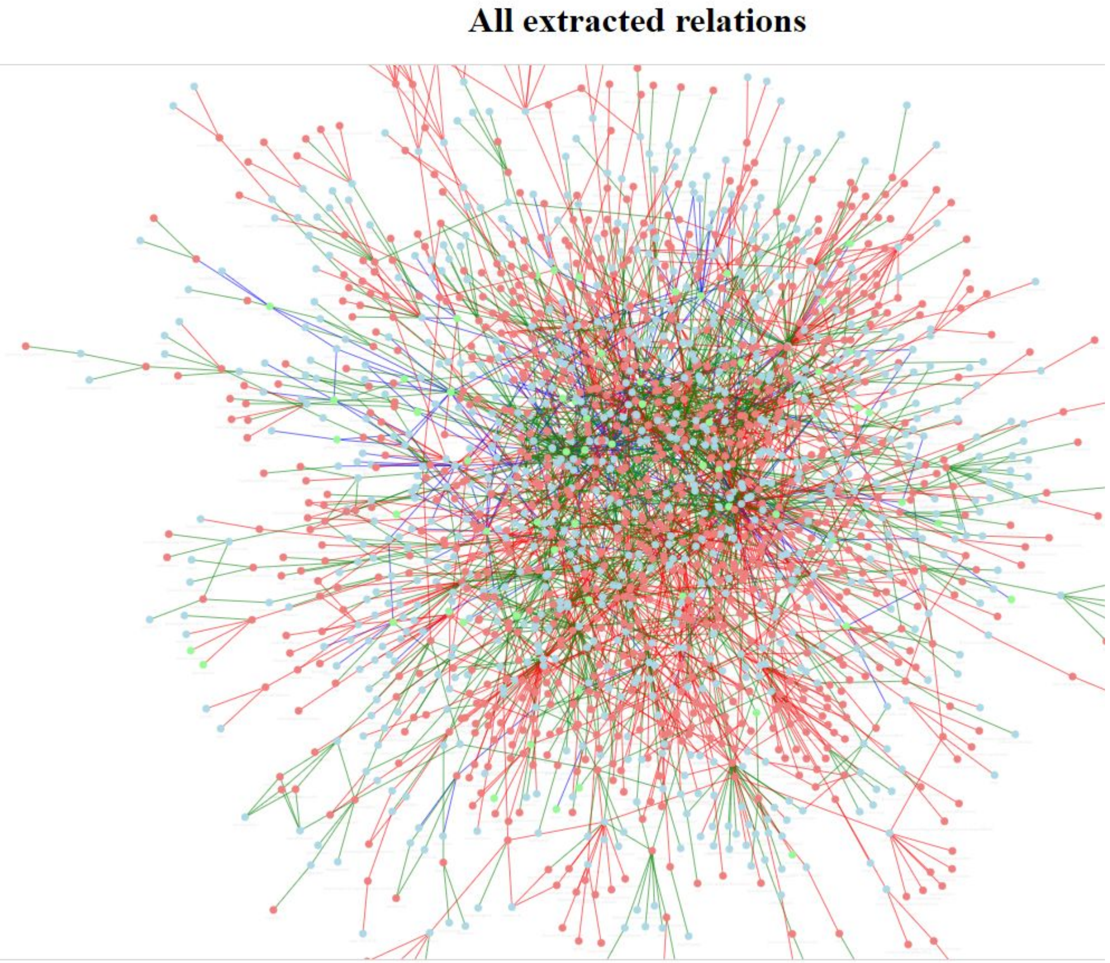


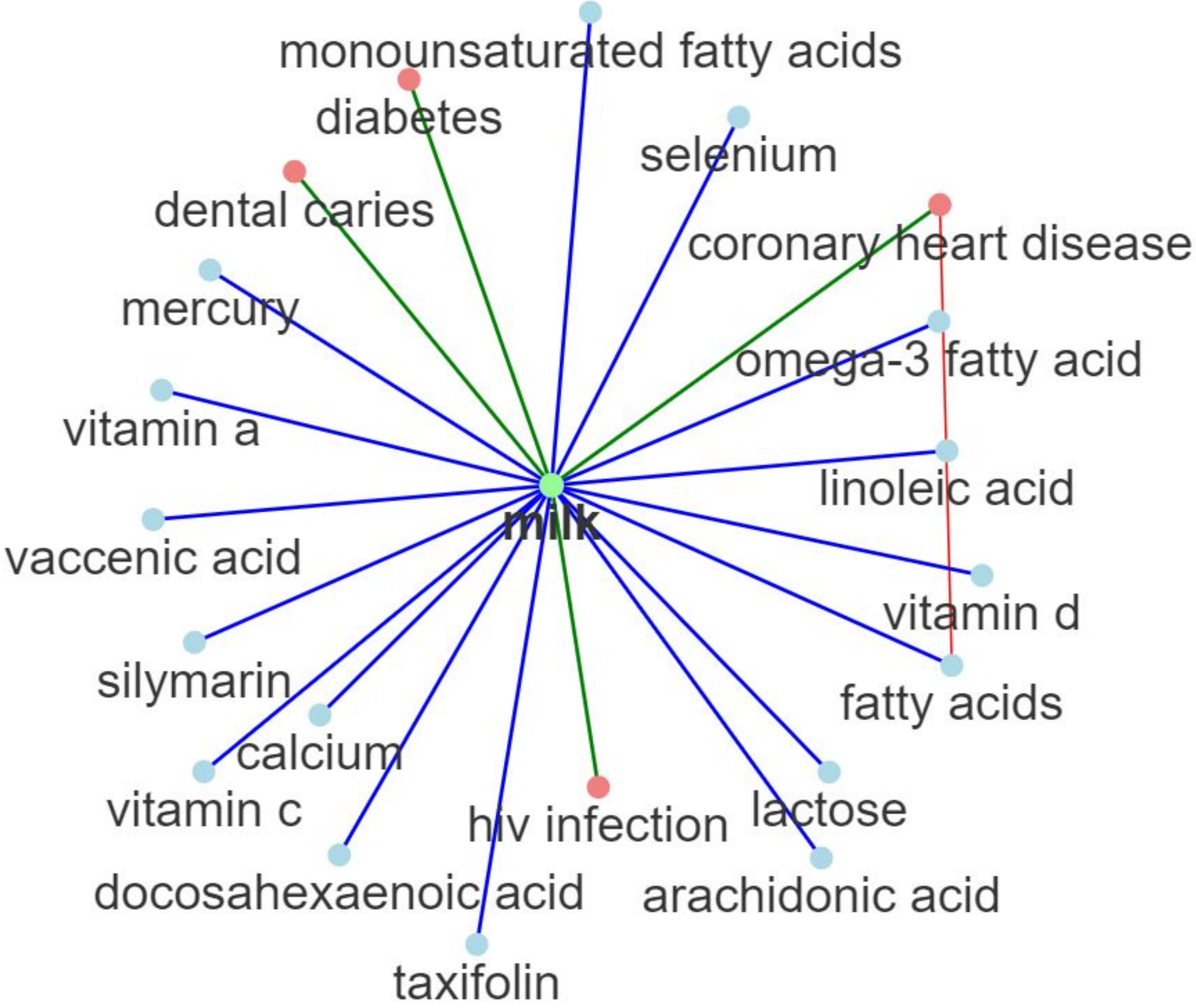
FoodChem
A food-chemical relation extraction model

- Extracts **contains** relations between **food** and **chemical** entities
- RE task is treated as a binary classification problem
- fine-tuning BERT, BioBERT and RoBERTa transformer models
- The BioBERT model achieves the best results, with a macro averaged F1 score of 0.902

Gjorgjina Cenikj, Tome Eftimov, Barbara Koroušić Seljak. "FoodChem: A food-chemical relation extraction model", IEEE Symposium Series on Computational Intelligence, 2021, IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2021)

Food, Chemical, and Disease KG





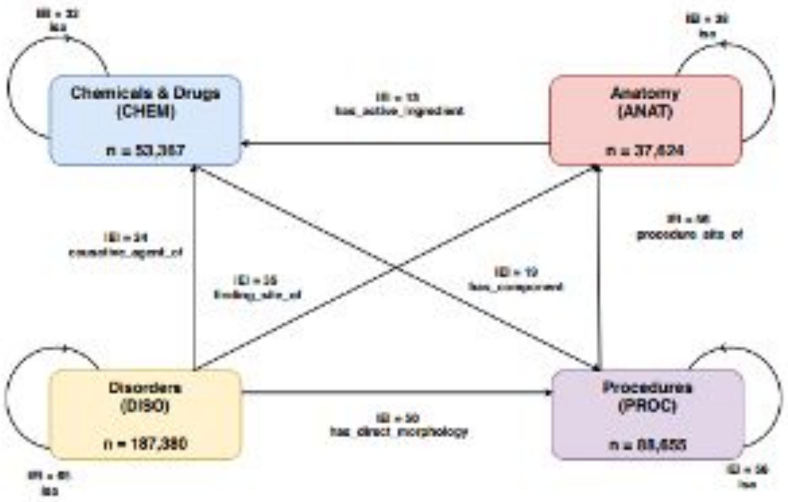
Patient diagnosis prediction using EHRs

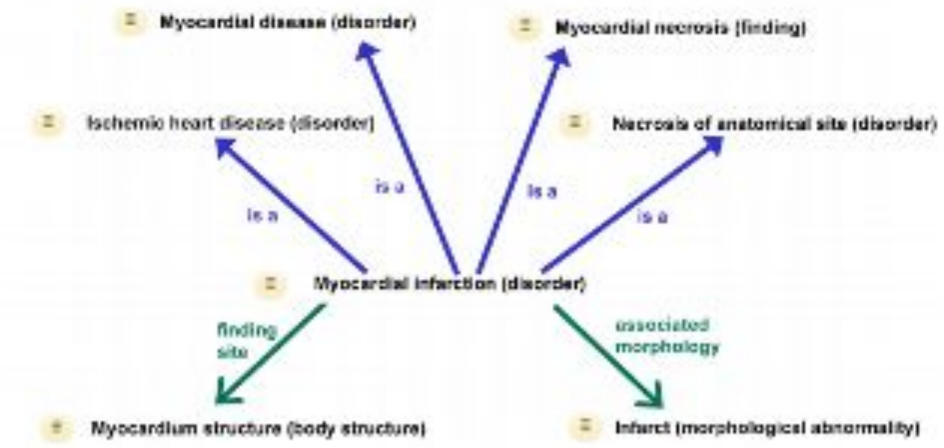
Tome Eftimov

Computer Systems Department

Jožef Stefan Institute, Ljubljana

SNOMED2vec representations (embeddings)





- Agarwal, K., Eftimov, T., Addanki, R., Choudhury, S., Tamang, S., & Rallo, R. (2019). Snomed2Vec: Random Walk and Poincar'e Embeddings of a Clinical Knowledge Base for Healthcare Analytics. In *2019 KDD Workshop on Applied Data Science for Healthcare (DSHealth '19)*.

SNOMED2vec representations (embeddings)

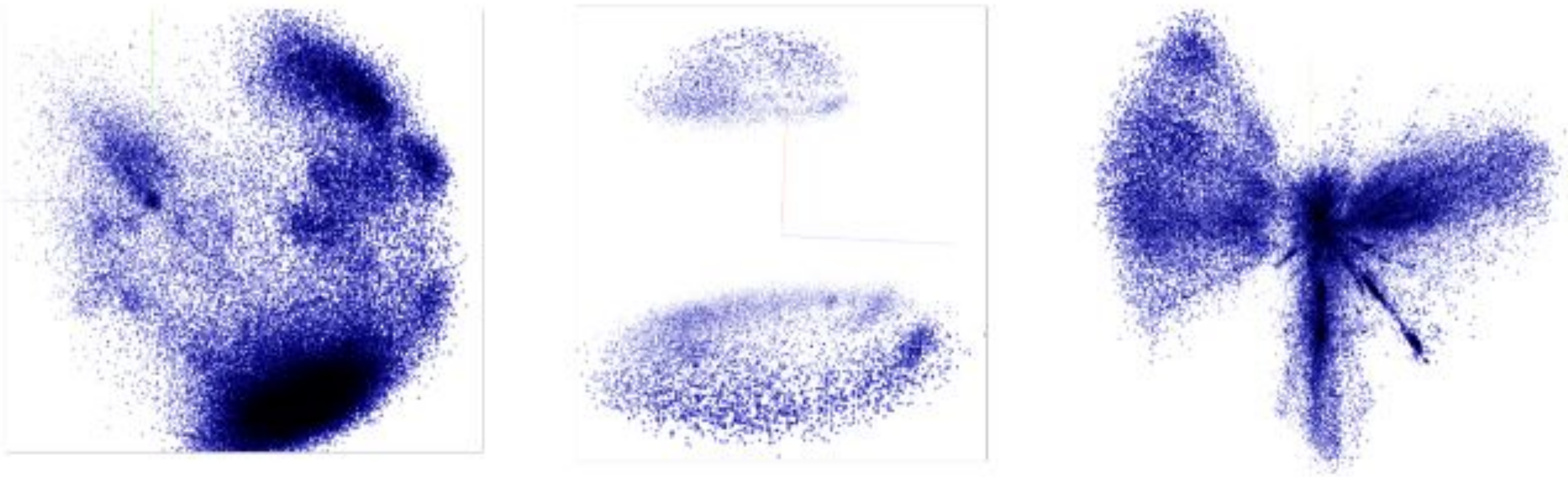


Figure 3: Visualization of the SNOMED-X graph embeddings (d=500) learned by Node2vec (top left), Metapath2vec (middle) and Poincaré (right). The shape of the visualizations demonstrate the distinct method objective and embedding characteristics (Node2vec: neighbourhood correlations; Metapath2vec: distinct node types; Poincare: hierarchical relations)

Diagnosis prediction

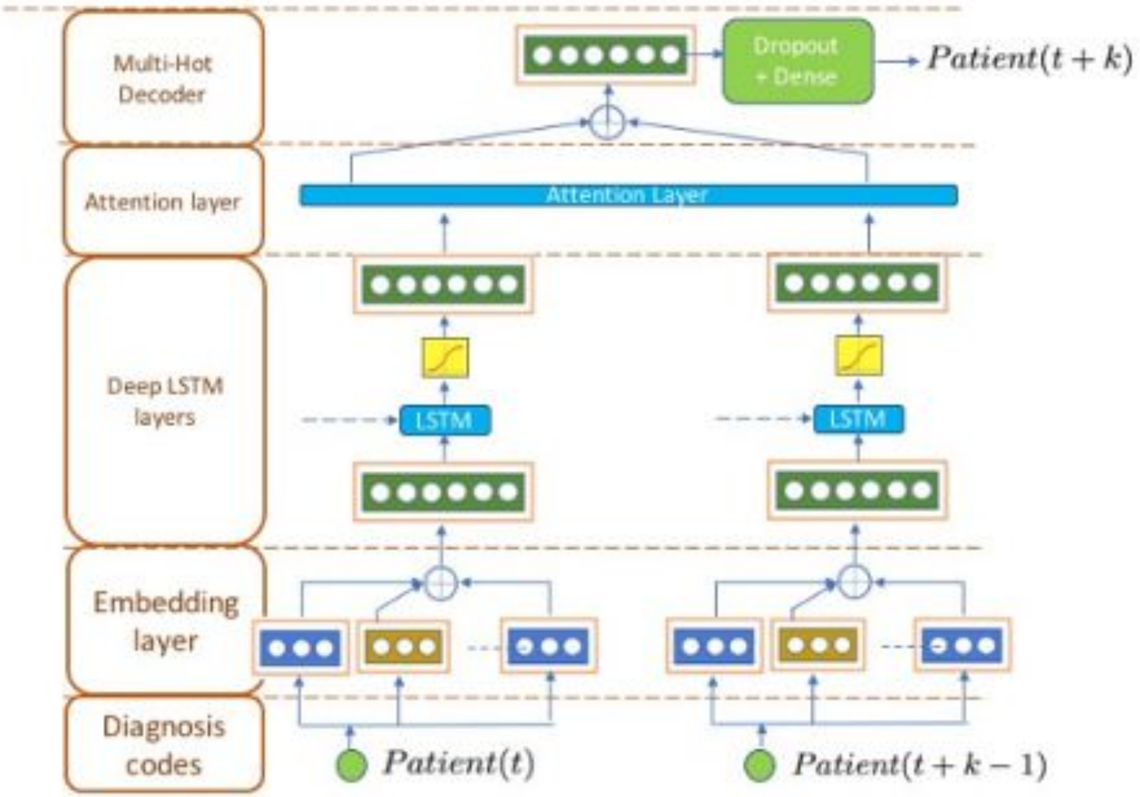


Figure 4: Architecture of the deep learning model to predict diagnostic codes from past EHR information

Diagnosis prediction

	Node2vec	Metapath2vec	Poincare	CUI2vec	Med2vec
Node Classification	0.817	0.3287	**0.8579**	0.5685	0.0409
Link Prediction	**0.986**	0.3988	0.7135	0.7222	0.8665
Concept Similarity (D1)	**0.79**	0.3	0.7	0.16	NA
Concept Similarity (D3)	**0.90**	0.46	0.31	0.15	NA
Concept Similarity (D5)	**0.81**	-0.32	-0.06	-0.01	NA
Patient State Prediction (All Diagnosis)	0.3938	0.3359	**0.4197**	0.3948	0.3881
Patient State Prediction (Frequent 20)	0.8465	**0.9749**	0.85	0.8035	0.7980
Patient State Prediction (Rare 20)	0.018	0.001	**0.019**	**0.019**	0.011

Table 1: Performance evaluation of embeddings on each task. Best performing method is highlighted for each task (using data from best performing embedding size for each method). Evaluation results show that knowledge graph-based embeddings outperform the state of art (Med2vec and CUI2vec) on all tasks.

Team

Gjorgjina Cenikj
Master Student
JSI

Gordana Ispirova
PhD Candidate
JSI

Eva Valenčič
PhD Candidate
JSI

Matevž Ogrinc
Master Student
JSI

Tome Eftimov, PhD
Senior Researcher
JSI

Prof. Barbara Koroušić Seljak, PhD
Senior Researcher
JSI