DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ARTIFICIAL INTELLIGENCE FOR SCIENCE

Sašo Džeroski
Head of Department of knowledge technologies
Jožef Stefan Institute, Ljubljana, Slovenia

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

WHY AI FOR SCIENCE IS IMPORTANT

Science has huge influence on our everyday lives

Science generates new knowledge and

- *Provides us with evidence-based understanding of the world*
- *It enables modern technologies*
  (incl. communication, transportation, energy)
- *It improves our quality of life*
  (through medicine, environmental protection, and food safety)

Artificial Intelligence provides us with ever more powerful tools

Fueled by

- *Ever faster hardware*
- *Ever more extensive data collection*
- *More and more powerful software*

AI can help science solve the global challenges better and faster

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

AI FOR SCIENCE: HISTORY

The field of AI is changing with lightning speed. *Today, AI (in the public eye) is all about Generative AI and Large Language Models (foundation models), underpinned by transformer architecture.*

But, AI has a long history. *And so has AI for science, starting with expert systems (DENDRAL, late 1960s) and computational scientific discovery (BACON, late 1970s).*

Lessons learned in computational scientific discovery *(early 2000s):*

- Output should be easy to communicate to domain scientists
- We should take advantage of background/ domain knowledge
- Should be able to infer knowledge from small data sets
- Produce models that provide explanations of data
- Support interaction with domain scientists

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ARTIFICIAL INTELLIGENCE FOR SCIENCE
Talk outline

A birds-eye view of my research on AI for science topics, including

- *Explainable ML for science,*
- *Foundation models for science,*
- *Automated scientific modelling, and*
- *Semantic technologies for open science.*

Example applications in life sciences and materials sciences

Projects and Infrastructure for AI in Science

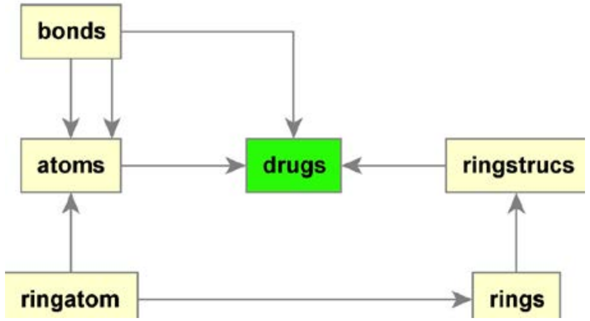
How this all fits in the broader European landscape:
Policy, events, community

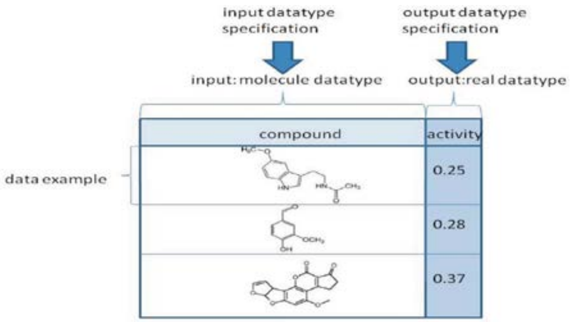
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

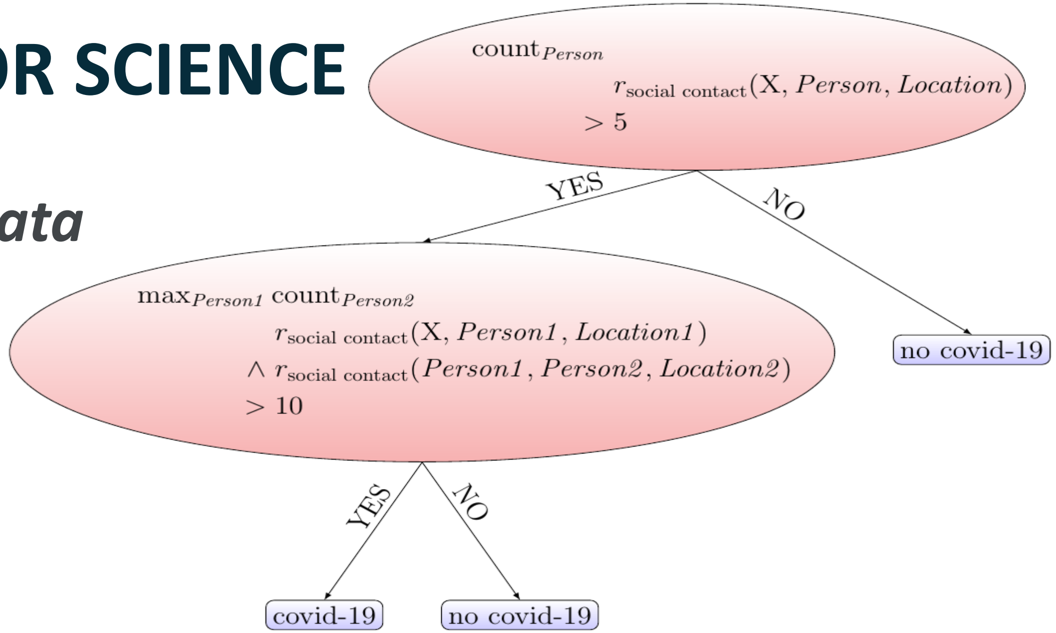
EXPLAINABLE ML FOR SCIENCE

EXPLAINABLE MACHINE LEARNING FOR SCIENCE

Learning interpretable models from complex data





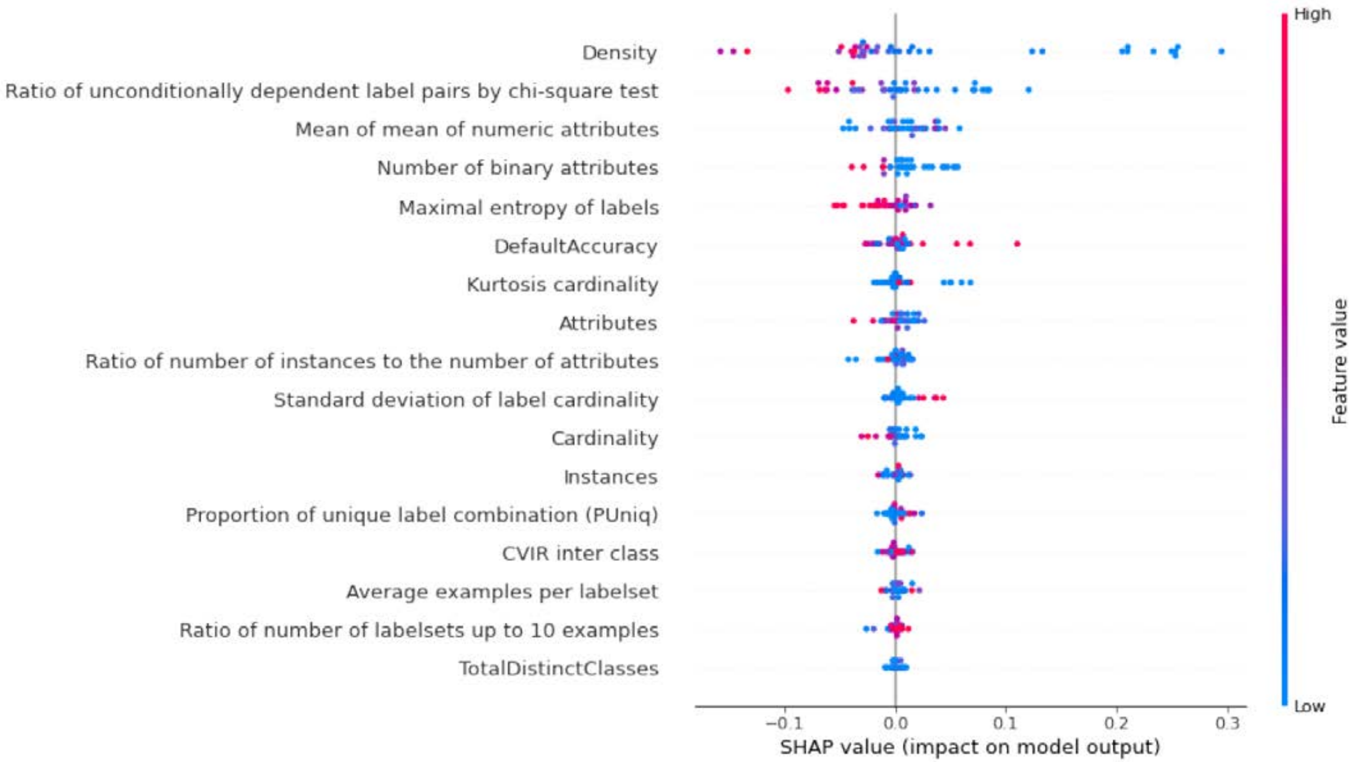


	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	?
Example 3	1	FALSE	0.08	0.07	?
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	?
…	…				…

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

EXPLAINABLE MACHINE LEARNING FOR SCIENCE

Explaining (uninterpretable models) and their predictions



MULTI-TARGET PREDICTION: DIMENSIONS

Different types of structured outputs

- MT/ML Classification, MT Regression, Hierarchical MLC/MTR

Different degrees of supervision

- Fully supervised
- Semi-supervised

Trees & ensembles

Batch vs. streaming MTP

(CLUSplus, iSOUP@MOA)

Data in context

- Spatial, temporal, spatio-temporal
- Relational (Re3Py)

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	?	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	?	?	?
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	?	?
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
...							

SoftwareX 24 (2023) 101526

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

ELSEVIER

Original software publication

CLUSPLUS: A decision tree-based framework for predicting structured outputs

Matej Petković [a,b,1], Jurica Levatić [a,1], Dragi Kocev [a], Martin Breskvar [a,2], Sašo Džeroski [a,*,2]

[a] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Faculty of Mathematics and Physics, Jadranska 21, 1000 Ljubljana, Slovenia

LEARNING TREES FOR MULTI-TARGET PREDICTION WITH PREDICTIVE CLUSTERING

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

To construct a tree T from a training set E:

If **the examples in E have low variance**,

construct a leaf labeled *target(prototype(E))*

Otherwise:

- Select the best attribute A with values v1, …, vn, which **reduces the most the variance Var(E)** (*measured according to a given distance function d*)
- Partition E into E1, …, En according to A
- Recursively construct subtrees T1 to Tn for E1 to En
- Result: a tree with root A and subtrees T1, …, Tn

The variance is assessed across the multiple targets

Semi-Supervised Predictive Clustering Trees for (Hierarchical) Multi-Label Classification

Jurica Levatić,[1,2] Michelangelo Ceci,[1,3] Dragi Kocev,[1,2] and Sašo Džeroski[1,2]

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3]Department of Computer Science, University of Bari, Bari, Italy

SEMI-SUPERVISED LEARNING WITH PCTs

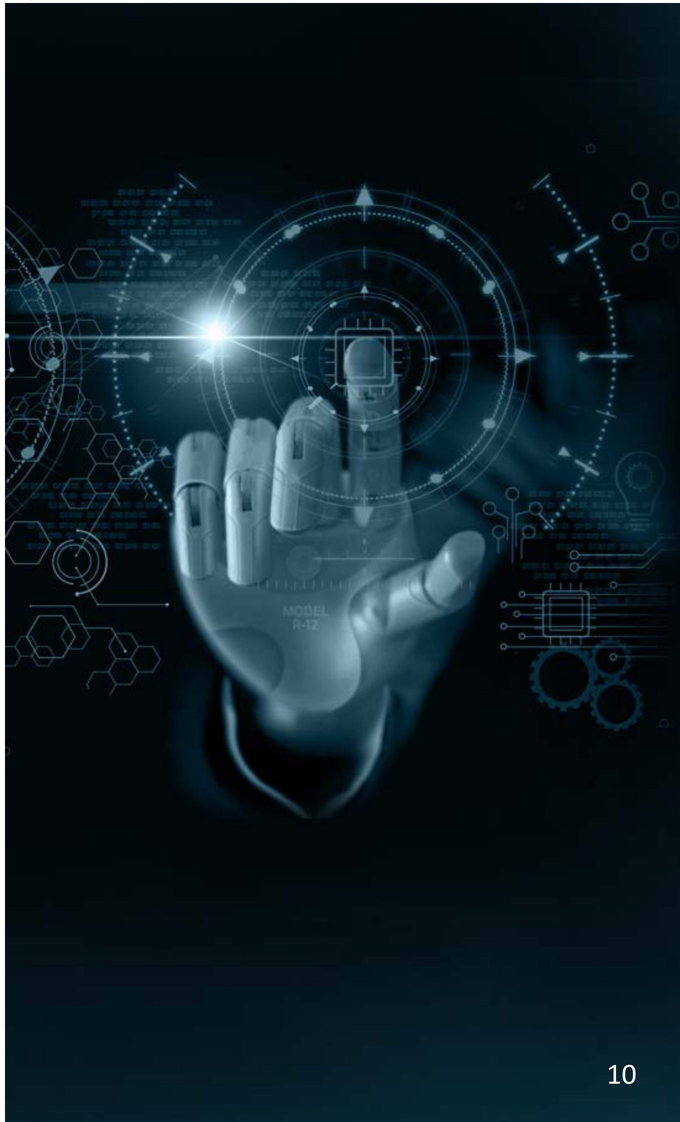
New definition of variance that includes both targets and attributes, e.g., for MTR

$$Var(E) = \frac{1}{T + D} \cdot \left(w \cdot \sum_{i=1}^{T} Var(Y_i) + (1 - w) \cdot \sum_{j=1}^{D} Var(X_j) \right)$$

T = #target attributes, D = #descriptive attributes

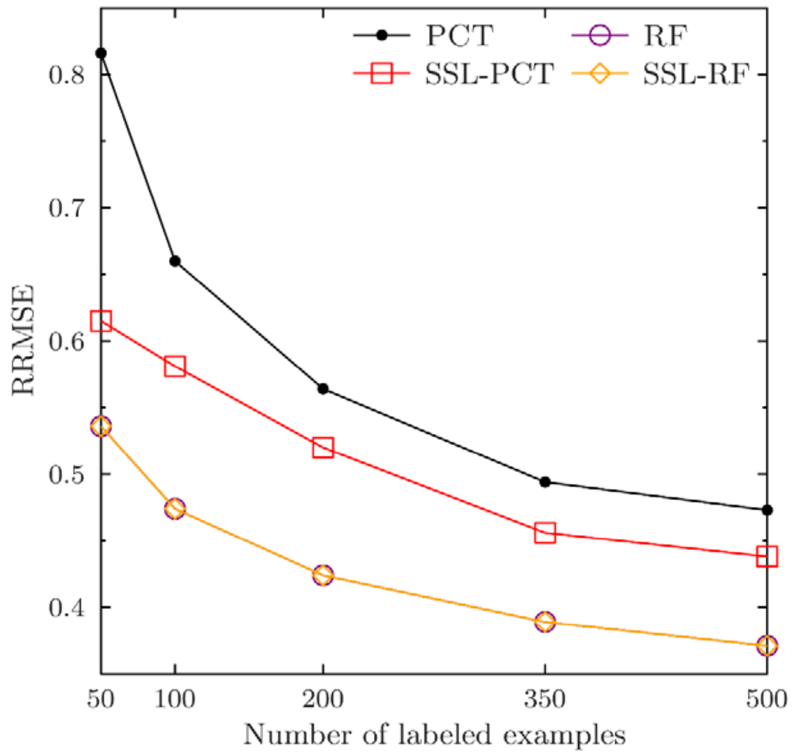
$$E = E_{\text{Labeled}} \cup E_{\text{Unlabeled}}$$

Variances only calculated for non-missing values

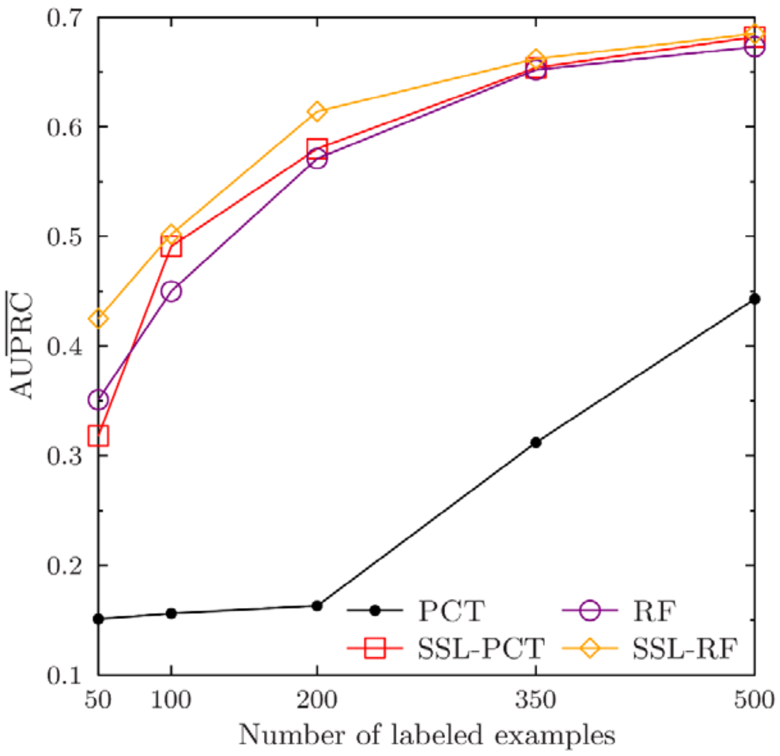


SSL vs. SL PERFORMANCE

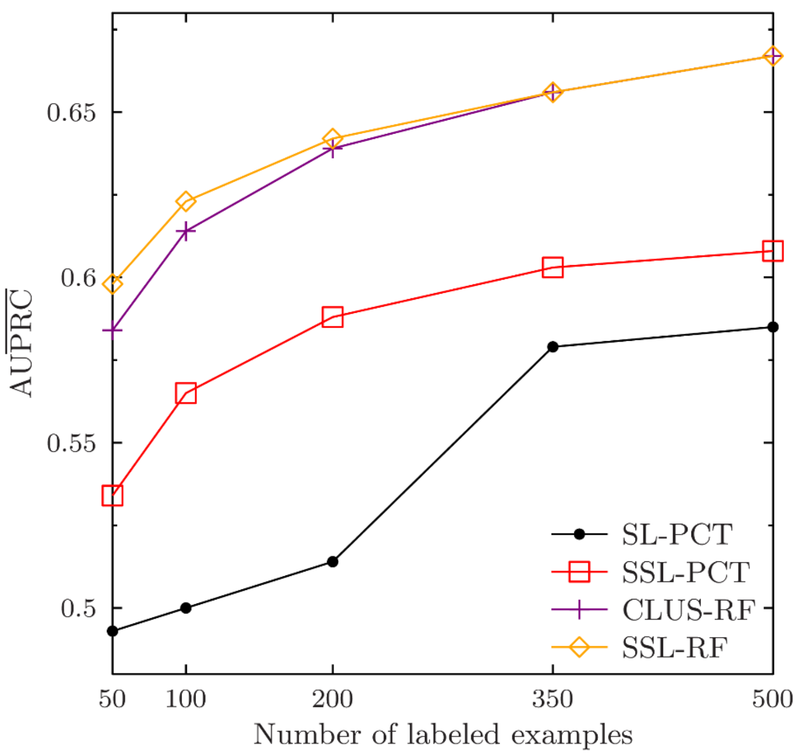
RF1 (MTR)



Medical (MLC)

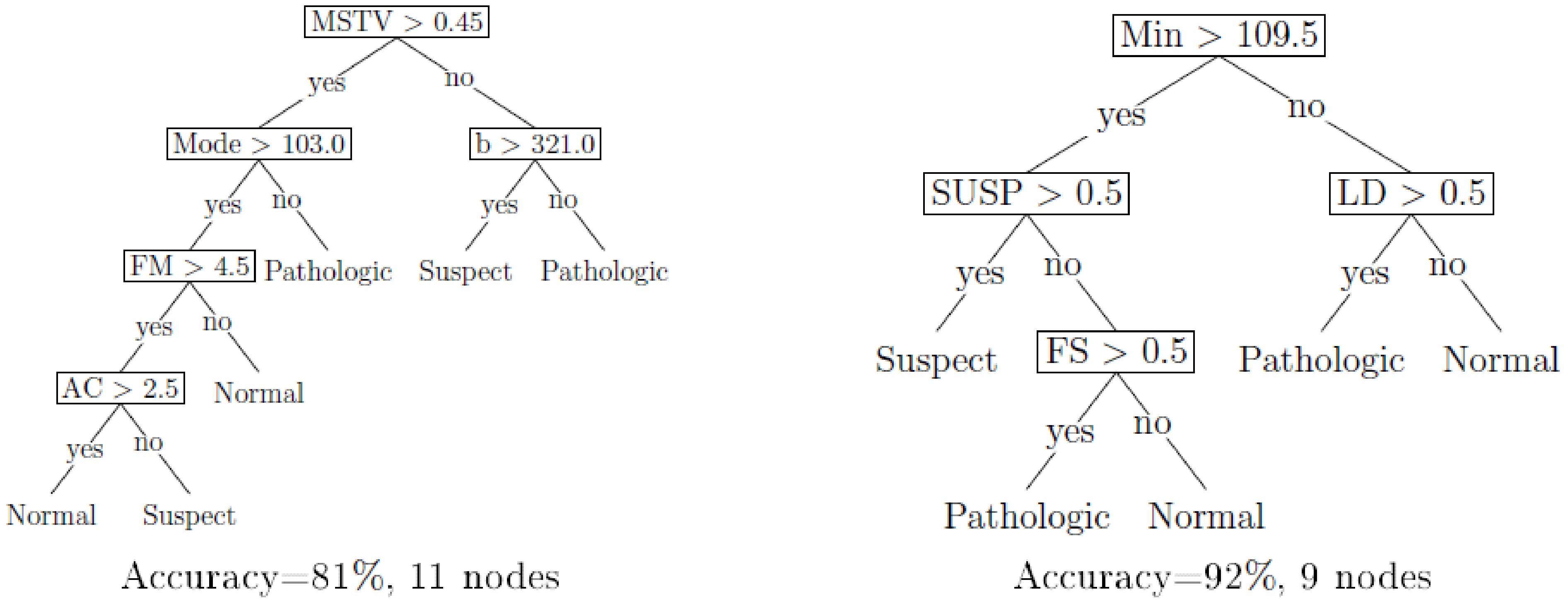


Enron (HMLC)



SSL OF DECISION TREES: ACCURACY & INTERPRETABILITY

Multi-class classification (Cardiotocogramy3 Dataset)



(c) SL-PCT, 50 labeled examples

(d) SSL-PCT, 50 labeled and 2076 unlabeled examples

Learning relational trees

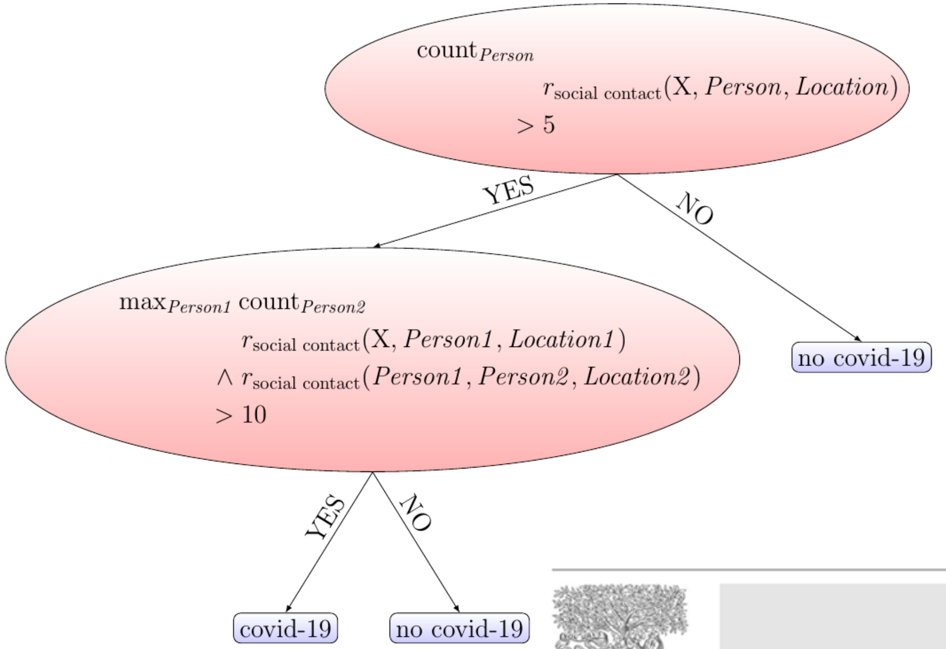
Relational data reside in multiple inter-related tables

The main part of relational tree induction is feature construction

Since examples can be related to different numbers of objects, aggregation of object property values is necessary

patients						
ID	gender	age	…	smoker	weight	COVID19
P1	male	31		yes	89	NO
P2	female	99		yes	54	YES
P3	female	12		no	32	YES
…	…	…		…		…

social contacts			
ID	patient1	patient2	location
SC1	P1	P3	L1
SC2	P1	P4	L2
…	…	…	



Recent work:

Ensembles, feature importance estimation, semi-supervised


Knowledge-Based Systems 251 (2022) 109254

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

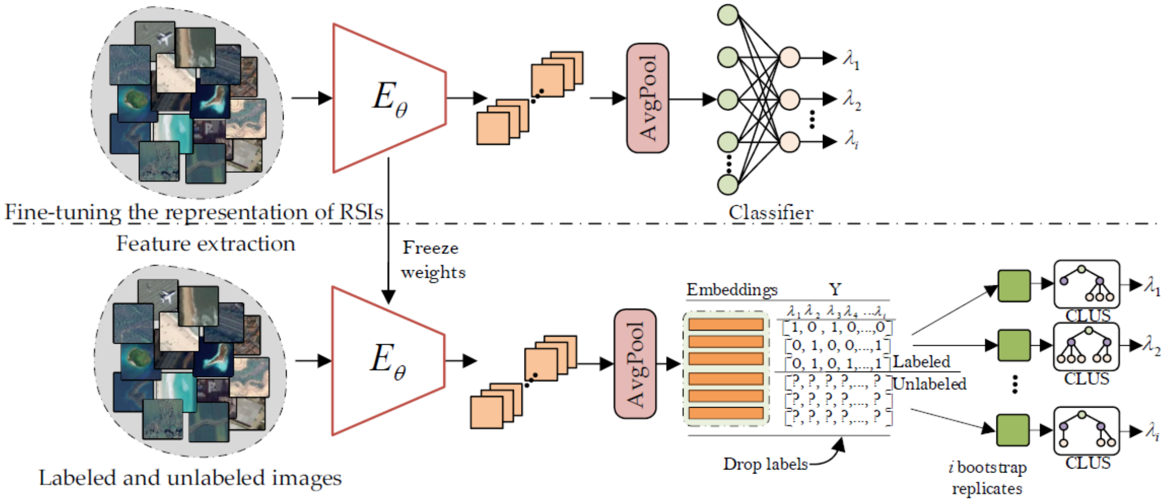
ELSEVIER

Relational tree ensembles and feature rankings

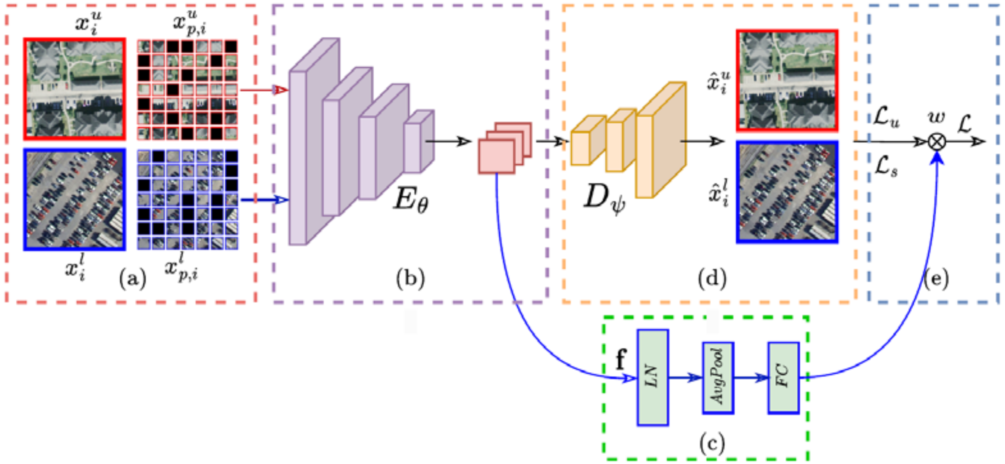
Matej Petković [a,b,e], Michelangelo Ceci [c,a,f,*], Gianvito Pio [c,f], Blaž Škrlj [a,b], Kristian Kersting [d], Sašo Džeroski [a,b]

Combining *representation learning (embeddings)* and multi-target prediction (PCTs)

Extract features with DNNs, then apply PCTs and ensembles



Combine the key ideas of DNNs and SSL MTP in end-to-end context



14882 IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, VOL. 18, 2025

SSL-MAE: Adaptive Semisupervised Learning Framework for Multilabel Classification of Remote Sensing Images Using Masked Autoencoders

Marjan Stoimchev, Jurica Levatić, Dragi Kocev, and Sašo Džeroski

APPLICATIONS IN THE LIFE SCIENCES

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- ***Predicting gene functions (HMLC=Hierarchical MLC)***

  In model organisms and bacterial genomes (genome-wide)

- ***Virtual compound screening for drug repurposing/ design***

  Tuberculosis & Salmonella; Fibrosis in myocardial infarction

- ***Relating environmental factors to biota structure***

  Microbiota in chicken/human gut; Funghi in beach sand

- ***Using machine learning to help select peptide insertion sites for regulation of protein function***

Plaper et al. *Cell Discovery* (2024)10:8
https://doi.org/10.1038/s41421-023-00635-y

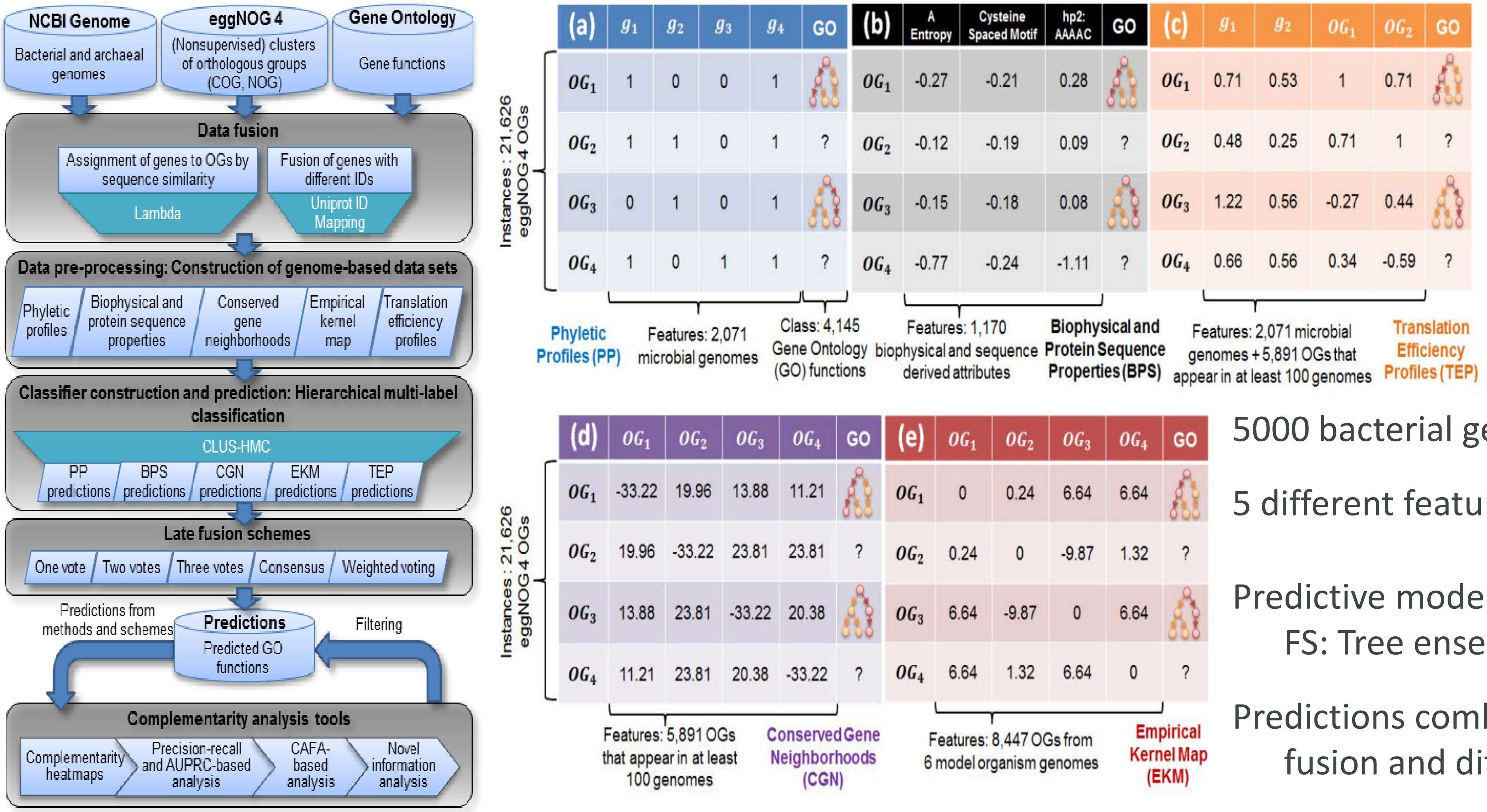
Cell Discovery
www.nature.com/celldisc

ARTICLE Open Access

Designed allosteric protein logic

Tjaša Plaper[1], Estera Merljak[1], Tina Fink[1], Tadej Satler[1,2], Ajasja Ljubetič[1], Duško Lainšček[1], Vid Jazbec[1,2], Mojca Benčina[1,3], Sintija Stevanoska[4], Sašo Džeroski[4] and Roman Jerala[1,3]✉

GENOME-WIDE GENE FUNCTION PREDICTION IN BACTERIA WITH PCT ENSEMBLES FOR HMLC ON DIFFERENT FEATURES



5000 bacterial genomes

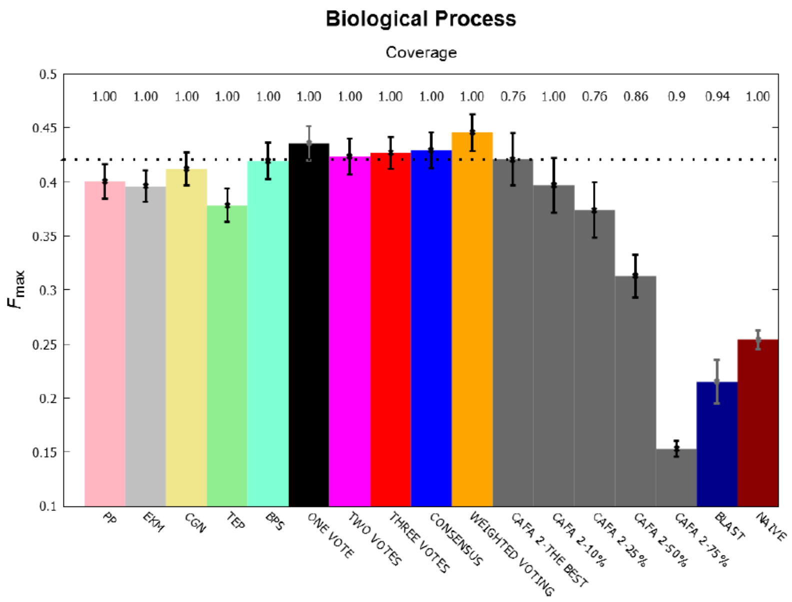
5 different feature sets

Predictive models learned from each FS: Tree ensembles for HMLC

Predictions combined with late fusion and diff. voting schemes

GENOME-WIDE GENE FUNCTION PREDICTION IN BACTERIA USINNG META-GENOME PHYLETIC PROFILES AS FEATURES

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute



Excellent results, comparable to the best in the CAFA 2 challenge

Further improved by using metagenomic phyletic profiles (MPP)

Metagenome Phyletic Profiles (MPP)

Features: metagenomes

	m_1	m_2	m_3	m_4	GO
OG_1	0	10E-5	0	0	
OG_2	10E-6	0	10E-7	10E-9	?
OG_3	0.008	0.02	0	0.01	
OG_4	0	0	0.003	0	?

Feature values: sum of OG member genes abundances in metagenomes

Phyletic Profiles (PP)

Features: microbial genomes

	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Feature values: presence/absence of genes in genomes

MPP can predict hundreds of gene functions that would not be predicted using only PP

VIRTUAL COMPOUND SCREENING FOR DRUG DESIGN & REPURPOSING

Analyzing data from compound screens, build predictive models to predict outcomes for new compounds

Descriptive variables refer to compound structure

- Functional groups
- Fingerprints
- Bulk properties

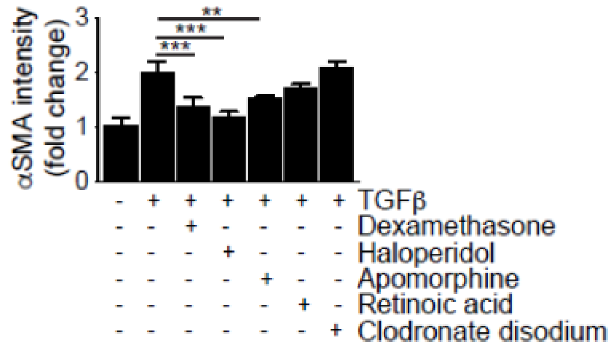
May also describe the compound in terms of the proteins it targets (e.g. from PubChem)

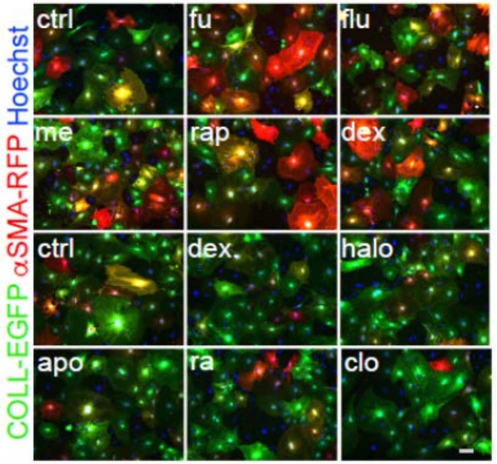
- Their functional annotations
- Pathways they are involved in
- Proteins that the targets interact with (and/or their functional annotations, pathways they are involved in)

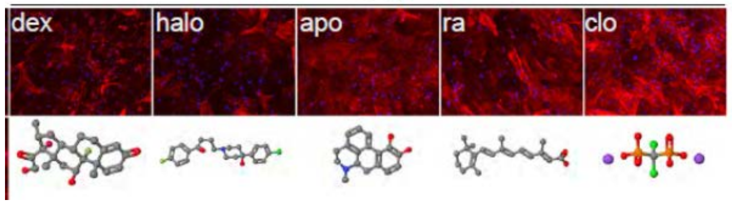
Target variables describe compound activity/toxicity

Practical examples

- Drug repurposing for Tuberculosis & Salmonella
- Drug design/repurposing to help recover after heart attack (high content scr.)







MACHINE LEARNING FOR MATERIALS DESIGN

Data

Parameters of the material synthesis process

- Temperature, Precursor concentration, Medium (acidity, ligands), Time, …

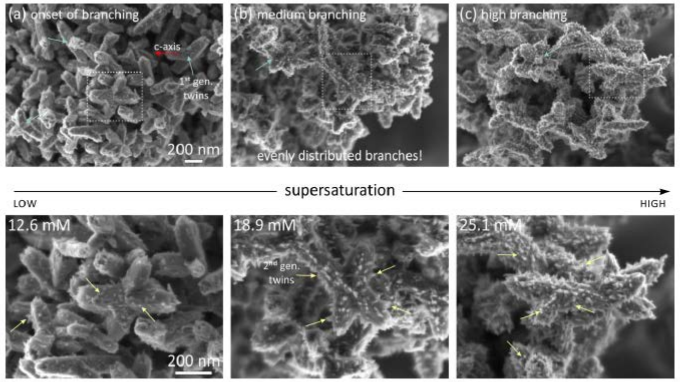
Quantum-level properties

- DFT simulation results
- Band-gap

Morphology of the material (images)

Functional properties of materials

- Electrical conductivity
- Catalytic activity
- Magnetic properties
- Anti-corrosive properties



Machine learning tasks

Find relations between the different kinds of data

Primarily predict functional properties of materials from the other kinds of data (synth. parameters, morphology, quantum-level properties)

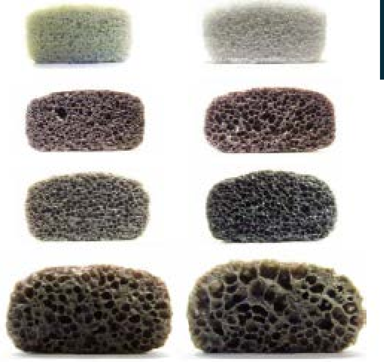
But also find relations between other pairs/ combinations of data, e.g.,

- *Parameters of synthesis and*
- *Morphology of the material*

or

- *Morphology of the materials and*
- *Properties of the material*

EXAMPLE APPLICATION: FOAMED GLASSES



The ML task and the data

Predict five properties of the foamed glass, based on the eight parameters of the synthesis process

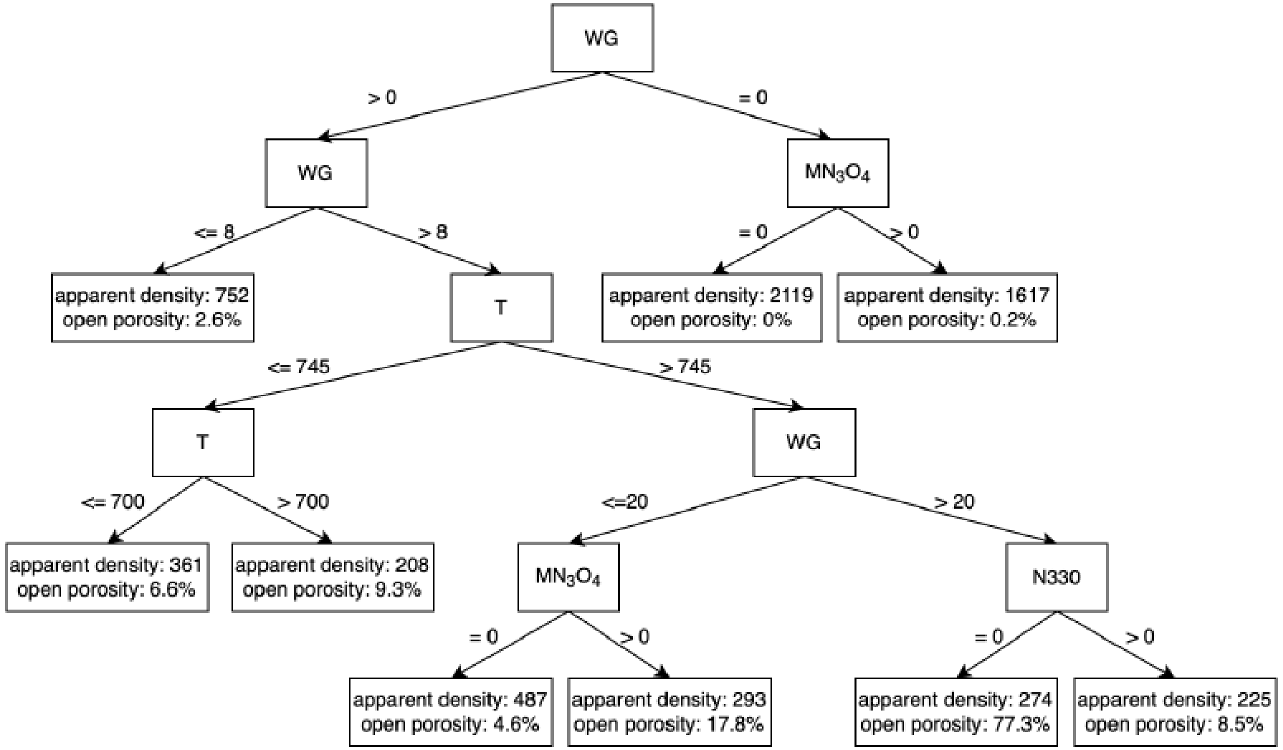
Process parameter
Water glass content [wt.%]
Carbon black content [wt.%]
Mn_3O_4 content [wt.%]
Furnace temperature [°C]
Foaming time [min]
K_3PO_4 content [wt.%]
Drying
Mixing

Material property
Apparent density ρ_{app}
Pycnometric density ρ_{pyc}
Overall porosity ε_{total}
Closed porosity ε_{closed}
Open porosity ε_{open}

A multi-target prediction tasks, where we have applied semi-supervised learning, as well as active learning

The learned model

Multi-target regression tree



COMBINING ML & OPTIMIZATION FOR MATERIALS DESIGN

The approach

First, use **machine learning** to learn a model **M** predicting material properties (**P**), from synthesis parameters (**S**)

P = M(S)

The task of material design is to find a set of synthesis parameters' values ***s*** that optimize the values of the properties **P**

For this task, we can use an **optimization** algorithm over the space of synthesis parameter values

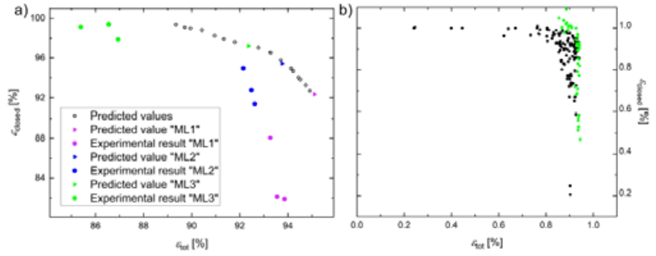
- If we are optimizing the value of one parameter, we can use single-criterion optimization
- If we are looking at multiple properties, we need to use multi-criteria/objective optimization (MOO)

Application to foamed glasses

Learn neural network predicting two properties

Use the NN to search for synthesis parameters that lead to non-dominated sets of properties

This suggested sets of values that were used by domain experts to synthesize new materials



Materials & Design 257 (2025) 114459

Contents lists available at ScienceDirect

Materials & Design

journal homepage: www.elsevier.com/locate/matdes

Optimizing foamed glass production with machine learning

Uroš Hribar [a,*], Sintija Stevanoska [a,b], Christian L. Camacho-Villalón [a,c], Matjaž Spreitzer [a], Jakob König [a,1], Sašo Džeroski [a,b,1]

[a] Jožef Stefan Institute, Jamova cesta 39, Ljubljana, 1000, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana, 1000, Slovenia

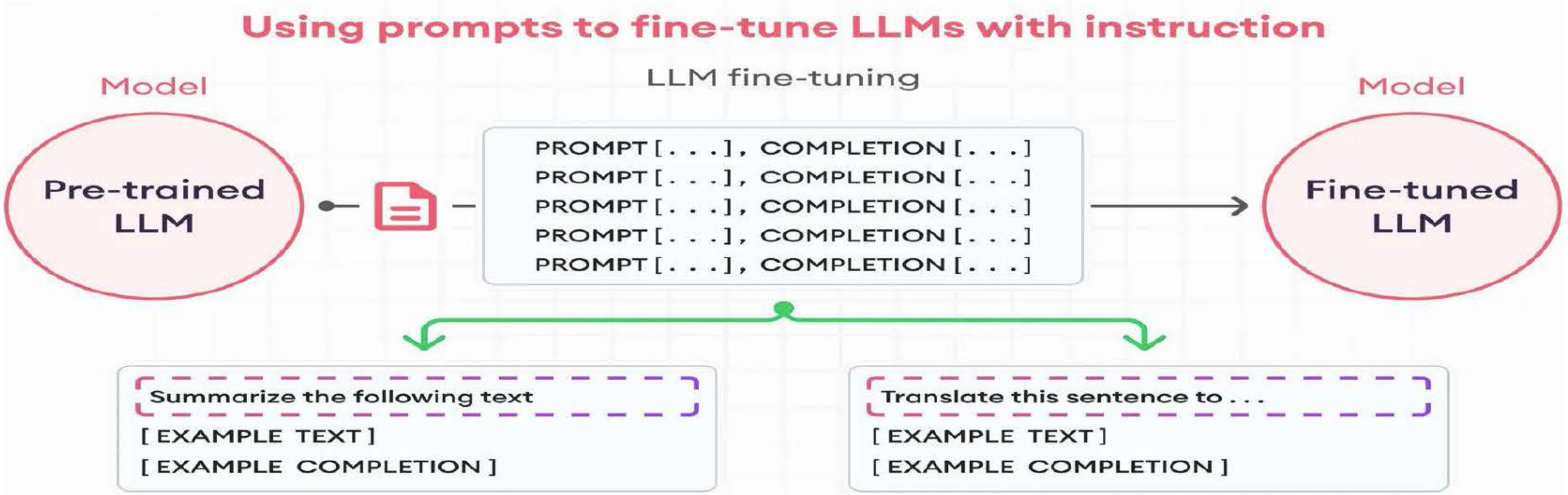
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES

Jožef Stefan Institute

FOUNDATION MODELS FOR SCIENCE

FOUNDATION MODELS FOR SCIENCE

Foundation models (FMs) are large models (e.g., large language models=LLMs) generated by applying ML (deep learning) to a broad collection of data at scale that can be adapted for use in a wide range of downstream tasks

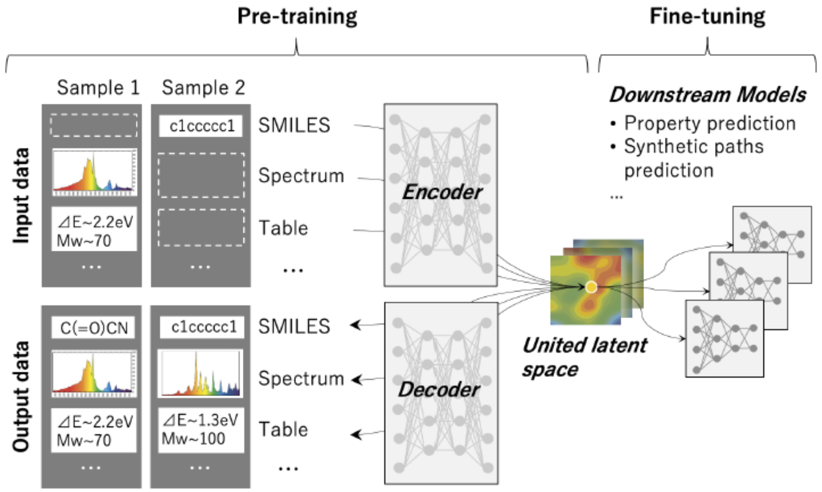


MULTI-MODAL FOUNDATION MODELS FOR SCIENCE

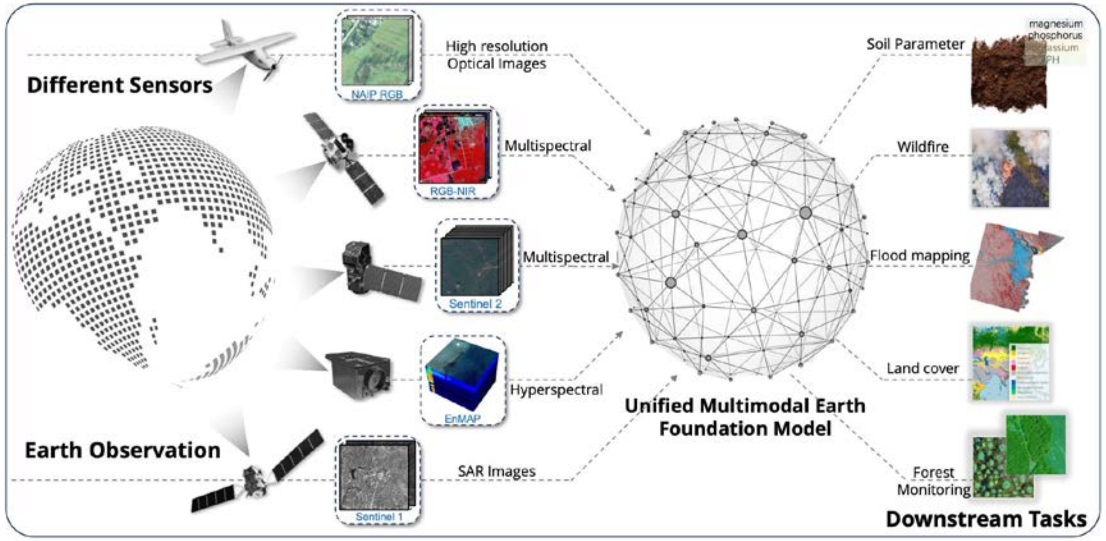
Built and operate on different modalities (text, different kinds of images, spectra, …)

- *Align the representations of different modalities*
- *Most typically bimodal (e.g. vision-language models)*
- *Scientific data is typically complex and multi-modal*

Materials science



Environmental/ Earth Science



FOUNDATION MODELS FOR SCIENCE (Example from nutrition science)

LLMs can be adapted (with own data) to specific scientific domains:
As an example, we have fine-tuned the Llama 3 LLM model to nutrition science

The LLM was fine-tuned on several nutrition datasets to be able to:

- Extract food related named entities (NE): NER (recognition) / NEL (linking)
- Classify food entities according to several food taxonomies (e.g., FoodON)
- Retrieve nutritional values for ingredients and recipes

```
EXAMPLE:

Input: Compute the nutrient values per 100 grams in a recipe with the following ingredients: 250 g cream, whipped, cream topping, pressurized, 250 g yogurt, greek, plain, nonfat, 50 g sugars, powdered.

Output: Nutrient values per 100 g listed: energy - 179.00, fat - 10.28, protein - 6.09, salt - 0.05, saturates - 6.34, sugars - 14.00
```

FOUNDATION MODELS FOR SCIENCE (Example from nutrition science)

The tasks:

- ***Traffic Light Classification of Recipes:*** *healthy (green), moderate (orange), limit consumption (red)*
  - *Two datasets: Ingredients, Title+Ingredients*
  - *Four dimensions: Salt, Sugar, Fat, Saturates*
- ***Assessing Recipe Nutrient Values***
  - *Two datasets: Ingredients, Title+Ingredients*
- ***Making food data interoperable***
  - *Named entity linking (NEL)*
  - *Named entity recognition (NER) + NEL*
  - *Three datasets: CafeteriaFCD (recipes), CafeteriaSA (scientific abstracts), Artificial*
  - *Three ontologies: FoodOn, SnomedCT, Hansard*

EXAMPLE PROMPTS FOR THE THREE TASKS

Zero-shot prompt:
[INST]Review the nutrient values per 100 grams in a recipe using these ingredients: 180 g wheat flours, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table.[/INST]
Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33

One-shot prompt:
[INST]*The following are examples of questions (with answers) about nutrition.* Question: Determine the nutritional profile per 100 grams in a recipe that uses these ingredients: 1 cup cheese, gouda, 4 tablespoon butter, without salt, 3/4 cup wheat flour, white, all-purpose, unenriched, 1/2 teaspoon salt, table, 1/2 teaspoon spices, pepper, red or cayenne, 1 tablespoon cream, fluid, heavy whipping Answer: Nutrient values highlighted for 100 grams: energy - 426.12, fat - 29.33, protein - 14.52, salt - 1.92, saturates - 18.33, sugars - 1.18 *Respond to the following question in the same manner as seen in the examples above.* Question: Review the nutrient values per 100 grams in a recipe using these ingredients: 180 g wheat flours, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table Answer:[/INST]
Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33

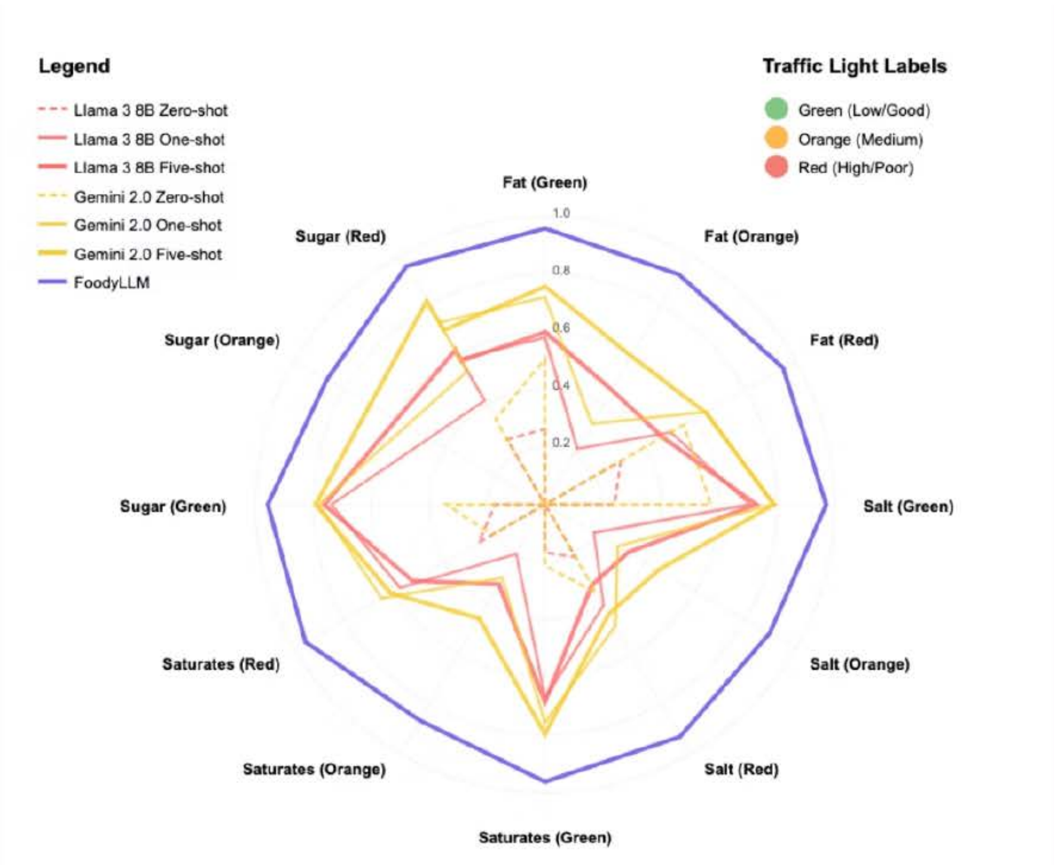
Five-shot prompt:
[INST]*The following are examples of questions (with answers) about nutrition.* Question: Gauge the nutrient values per 100 grams in a recipe prepared with the following ingredients: 2 cup cream, fluid, heavy whipping, 1 tablespoon spices, cardamom Answer: Per 100 g, the nutrient values are: energy - 339.02, fat - 35.36, protein - 3.04, salt - 0.07, saturates - 22.49, sugars - 2.85 Question: Establish the nutrient profile per 100 g in a recipe containing these ingredients: 1 tablespoon shallots, raw, 2 teaspoon spices, garlic powder, 12 cup peanut butter, smooth style, without salt, 3 tablespoon oil, sesame, salad or cooking, 2 tablespoon soy sauce made from soy (tamari), 1 teaspoon spices, ginger, ground, 1 teaspoon roland, seasoned rice wine vinegar, upc: 041224705142, 1/4-1/2 teaspoon spices, pepper, red or cayenne, 13 cup soup, chicken broth or bouillon, dry Answer: Nutrient profile for every 100 g: energy - 494.83, fat - 40.58, protein - 20.22, salt - 17.01, saturates - 8.28, sugars - 12.29 Question: Verify the nutrient values per 100 g in a recipe prepared with these ingredients: 16 ounce milk, fluid, 1% fat, without added vitamin a and vitamin d, 8 ounce beverages, almond milk, unsweetened, shelf stable, 13 cup sugars, granulated, 14 cup cornstarch, 12 teaspoon vanilla extract, 14 teaspoon shortening confectionery, coconut (hydrogenated) and or palm kernel (hydrogenated) Answer: Nutrient facts per 100 grams: energy - 340.40, fat - 1.30, protein - 0.41, salt - 0.03, saturates - 1.11, sugars - 50.82 Question: Identify the nutritional composition per 100 grams in a recipe with these ingredients: 500 g ground turkey, raw, 1 cup onions, raw, 12 cup bread crumbs, dry, grated, plain, 12 cup carrots, raw, 12 cup sauce, barbecue, 2 teaspoon sauce, worcestershire, 1 teaspoon spices, garlic powder, 34 teaspoon spices, pepper, black Answer: Nutrient profile for each 100 g: energy - 180.22, fat - 1.96, protein - 4.72, salt - 1.63, saturates - 0.42, sugars - 18.20 Question: Find the nutritional breakdown per 100 g in a recipe that uses the following ingredients: 12 pound pretzels, soft, unsalted, 21 ounce corn, sweet, white, raw, 12 ounce cookies, graham crackers, plain or honey, lowfat, 32 ounce nuts, walnuts, english, 7 ounce cookies, graham crackers, plain or honey, lowfat, 12 ounce cookies, graham crackers, plain or honey, lowfat, 1 pound butter, without salt, 12 drop sauce, ready-to-serve, pepper, tabasco, 8 tablespoon sugars, brown, 2 teaspoon spices, chili powder, 2 teaspoon sauce, worcestershire, 2-3 tablespoon spices, garlic powder Answer: The nutrient breakdown per 100 g is: energy - 383.98, fat - 14.10, protein - 7.92, salt - 0.59, saturates - 3.85, sugars - 3.92 *Respond to the following question in the same manner as seen in the examples above.* Question: Review the nutrient values per 100 grams in a recipe using these ingredients: 180 g wheat flours, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table Answer:[/INST]
Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33

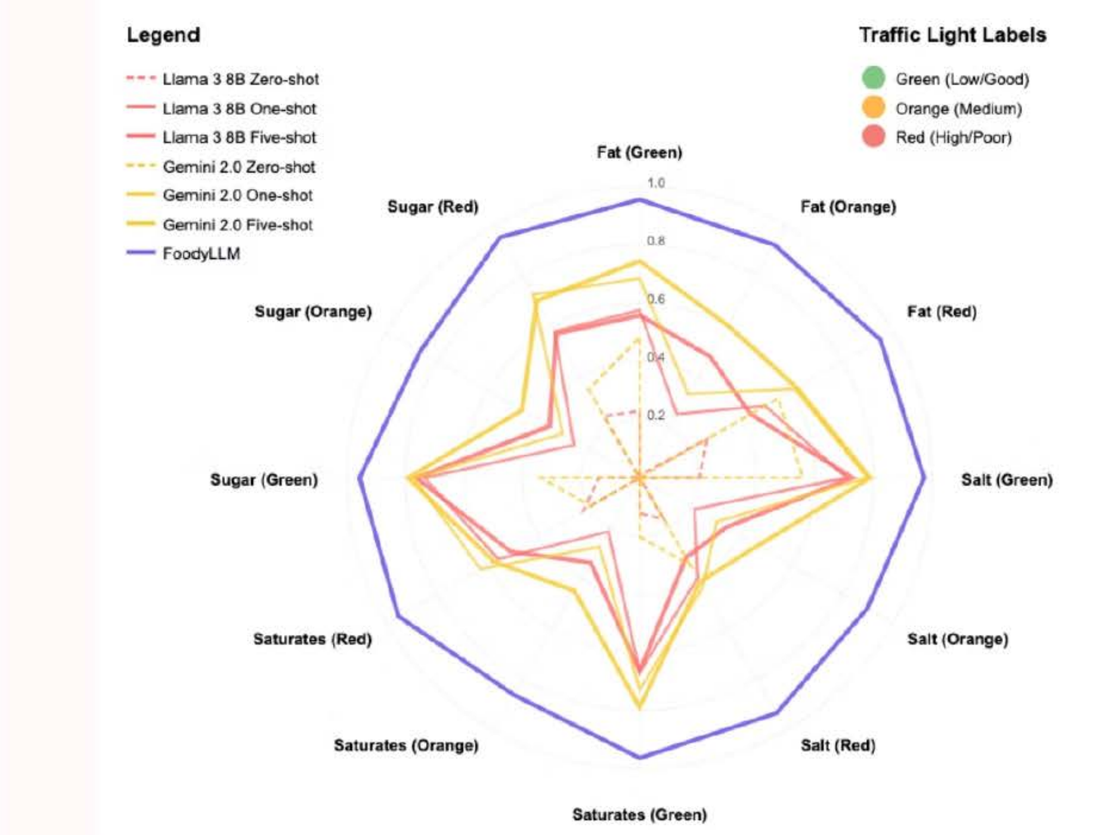
[INST] Verify the FSA traffic light labels per 100 grams in a recipe that has these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST]
Food Standards Agency traffic lights per 100 g presented: fat - green, salt - red, saturates - green, sugars - green

[INST] Evaluate the FSA traffic light labels per 100 grams in *Easy Pizza Dough for Bread Machine* using these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST]
The fsa traffic lights per 100 g listed are: fat - green, salt - red, saturates - green, sugars - green

[INST] Find all food entities present in the text: 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. [/INST]
Certainly, the entities connected with food are outlined as follows: cream cheese, beef, olives, onion, Worcestershire sauce, walnuts, cheese ball.
[INST] Is it within your capabilities to link the extracted food entities to a Hansard taxonomy? [/INST]
Certainly, the entities are associated properly: cream cheese - AG.01.e [Dairy produce]; AG.01.e.02 [Cheese]; AG.01.n [Dishes and prepared food]; AG.01.n.18 [Preserve], beef - AG.01.d.03 [Beef], olives - AG.01.h.01.e [Fruit containing stone], onion - AG.01.h.02.e [Onion/leek/garlic], Worcestershire sauce - AG.01.h [Fruit and vegetables]; AG.01.l.04 [Sauce/dressing], walnuts - AG.01.h.01.f [Nut], cheese ball - AG.01.e.02 [Cheese]; AG.01.n.18 [Preserve].
[INST] Might I trouble you to connect the extracted food entities to a FoodOn ontology, if possible? [/INST]
Definitely, the entities are linked suitably: cream cheese - FOODON-03301889; FOODON-00001013, onion - FOODON-03301704; NCBITaxon-4679, worcestershire sauce - FOODON-03305003; FOODON-03311146, walnuts - NCBITaxon-16718, cheese - FOODON-00001013.
[INST] Might it be possible for you to link them to a SNOMEDCT ontology, if needed? [/INST]
Absolutely, the entities are related as such: cream cheese - SNOMEDCT-226849005; SNOMEDCT-255621006; SNOMEDCT-102264005, beef - SNOMEDCT-226916002, olives - SNOMEDCT-227436000, onion - SNOMEDCT-735047000, worcestershire sauce - SNOMEDCT-443701000124100; SNOMEDCT-227519005, cheese - SNOMEDCT-102264005.

FoodyLLM: Results for Traffic Light Classification



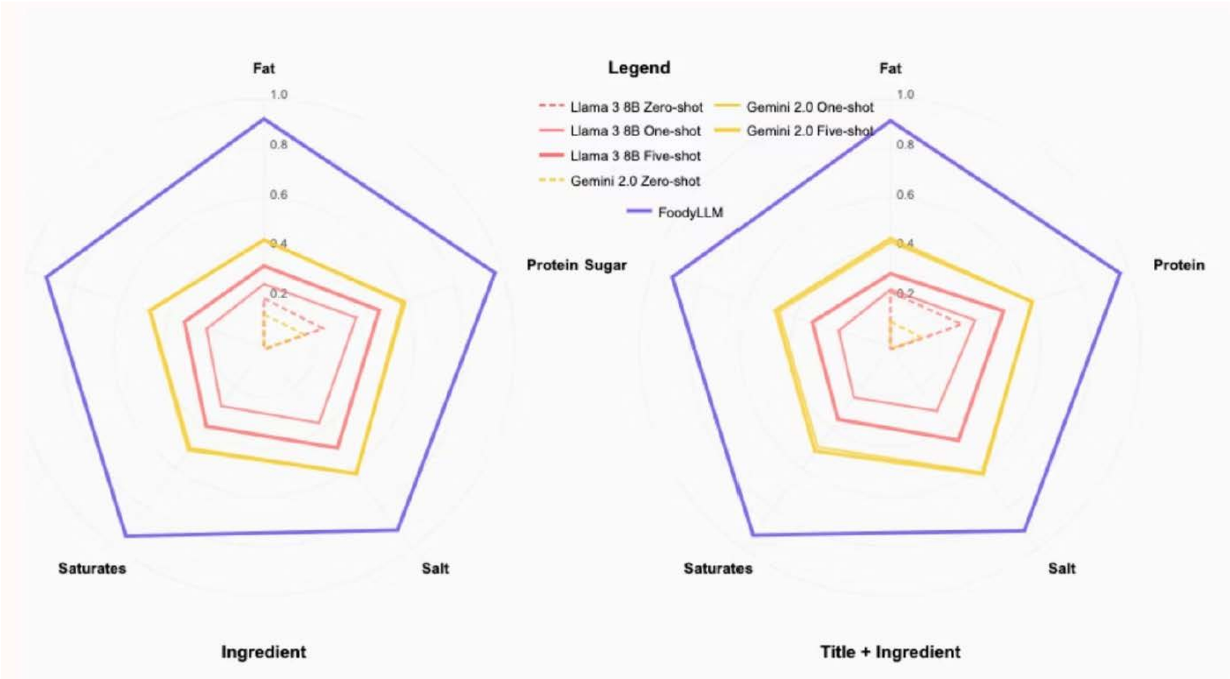


Ingredient dataset; Title+Ingredient dataset

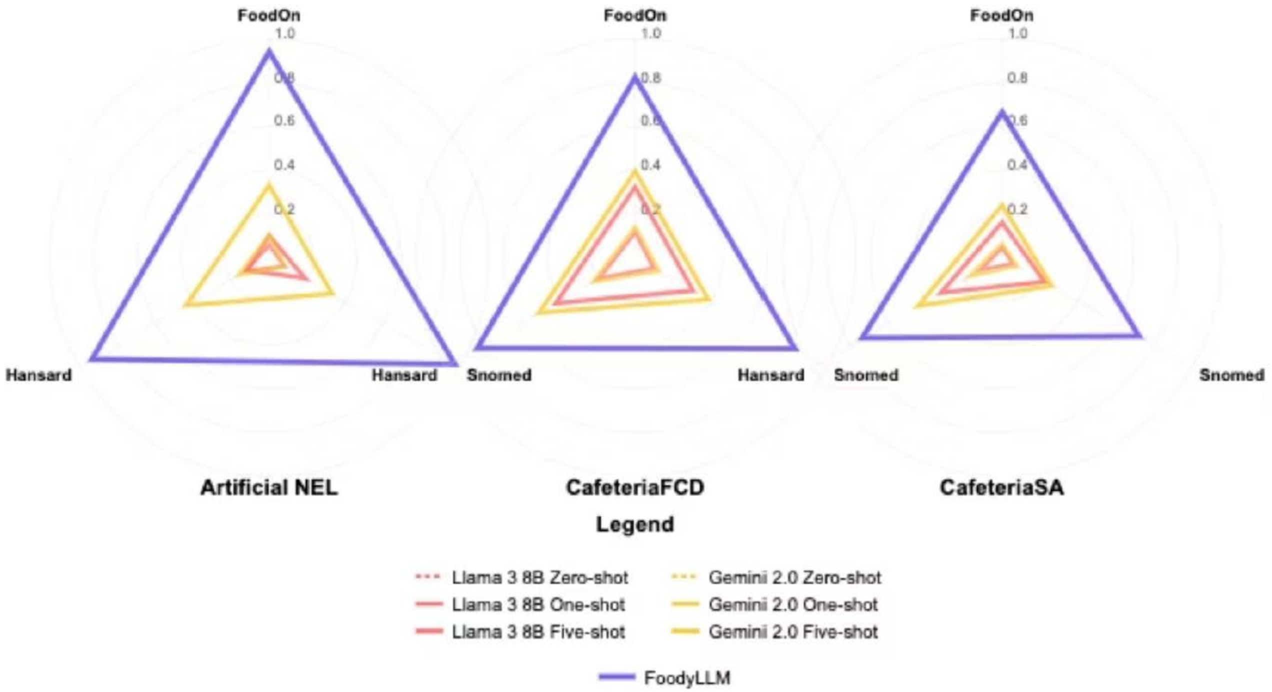
FoodyLLM: A FAIR-aligned specialized large language model for food and nutrition analysis

Ana Gjorgjevikj[1†], Matej Martinc[2†], Gjorgjina Cenikj[1,3], Riste Stojanov[4], Jan Drole[1,3], Gordana Ispirova[5], Giulia Menichetti[5,6,7], Nives Ogrinc[8], Dimitar Trajanov[4,9], Sašo Džeroski[2], Barbara Koroušić Seljak[1,10], Tome Eftimov[1*]

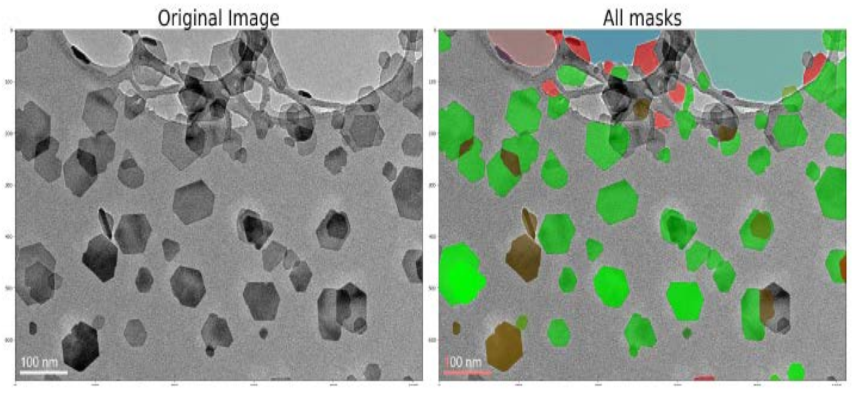
FoodyLLM:
Results for
Assessing Recipe Nutrient Values



Named Entity Linking



USING VISUAL FOUNDATION MODELS IN MATERIALS SCIENCE



Segment Anything Model: *Used to analyze Electron Microscopy Images*

Designing magnetic materials: *Transmission Electron Microscopy images of Barium hexaferrite nanoplatelets*

Particle diameter distribution predicts *magnetic properties*

Can Segment Anything Model Segment Just Some Things? Improving Zero-shot Segmentation of Electron Microscopic Images with Clustering

1st Matej Martinc
Jožef Stefan Institute
Ljubljana, Slovenia
matej.martinc@ijs.si

2nd Anastasija Manojlovska
University of Ljubljana
Ljubljana, Slovenia
anastasija.manojlovska@fe.uni-lj.si

3rd Sašo Džeroski
Jožef Stefan Institute
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Coming up next

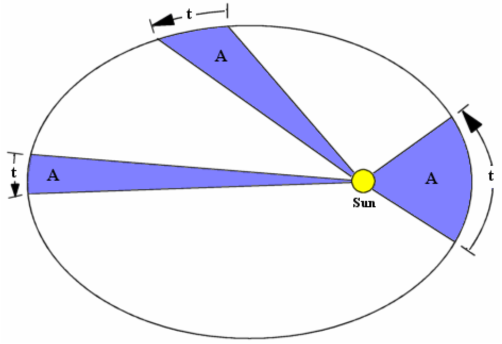
Multi-modal FMs for materials, e.g., visual-language models

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

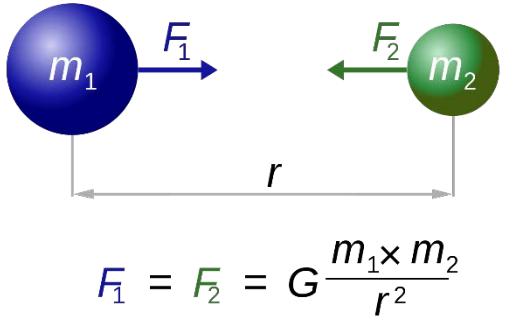
AUTOMATED SCIENTIFIC MODELLING

COMPUTATIONAL SCIENTIFIC DISCOVERY

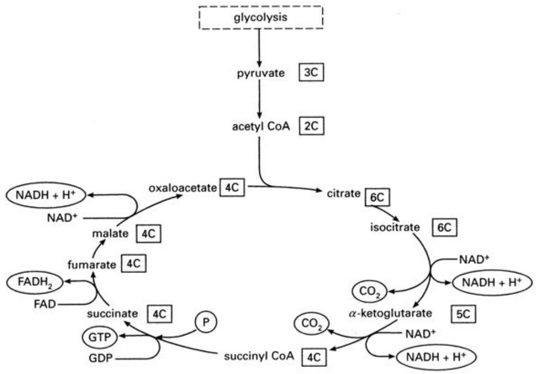
Find scientific knowledge, represented in scientific formalisms (e.g., equations, pathways), introduced and routinely used by scientists (hence accessible to them)



Kepler's laws of planetary motion



Newton's theory of gravitation



Krebs' citric acid cycle

Equation discovery (Symbolic regression): Finding equations from data

Moon	d	p
A	5.67	1.77
B	8.67	3.57
C	14.00	7.16
D	24.67	16.69

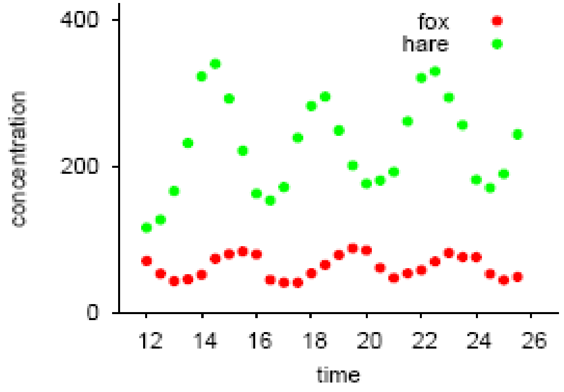
$$d^3/p^2 = k$$

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

COMPUTATIONAL SCIENTIFIC DISCOVERY

Automated modeling of dynamic systems

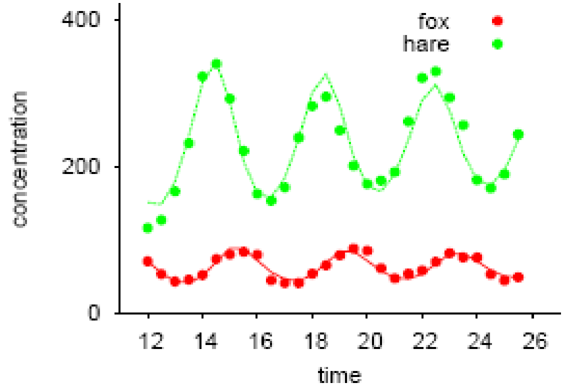
- Input: Observed behavior of dynamic system



Time	System variables			
	v_1	v_2	$\dots$	v_n
t_0	$v_{1,0}$	$v_{2,0}$	$\dots$	$v_{n,0}$
t_1	$v_{1,1}$	$v_{2,1}$	$\dots$	$v_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	$v_{1,m}$	$v_{2,m}$	$\dots$	$v_{n,m}$

- Output: System of ODEs

$$\frac{d}{dt} hare = 2.5 \cdot hare - 0.3 \cdot hare \cdot fox$$
$$\frac{d}{dt} fox = 0.1 \cdot 0.3 \cdot hare \cdot fox - 1.2 \cdot fox$$



AUTOMATED SCIENTIFIC MODELLING

$$
\begin{array}{rcl}
PPModel & \rightarrow & PreyChange,\ PredatorChange \\
\\
PreyChange & \rightarrow & PreyGrowth - Interaction \\
PredatorChange & \rightarrow & const * Interaction + PredatorLoss \\
\\
PreyGrowth & \rightarrow & const * V_{Prey} \\
PreyGrowth & \rightarrow & const * V_{Prey} * (1 - V_{Prey}/const) \\
\\
Interaction & \rightarrow & const * V_{Predator} * V_{Prey} \\
Interaction & \rightarrow & const * V_{Predator} * V_{Prey}/(V_{Prey} + const) \\
\\
PredatorLoss & \rightarrow & -const * V_{Predator} \\
\\
V_{Prey} & \rightarrow & hare \\
V_{Predator} & \rightarrow & fox
\end{array}
$$

Integrating data-driven and knowledge-driven modeling

*Use both data and domain knowledge as input: **Grammars***

Grammars can be used to represent different kinds of domain knowledge

- *Basic building blocks of systems/ models in the domain*
- *Existing models in the domain (that need to be revised)*
- *Incomplete models, than need to be completed*

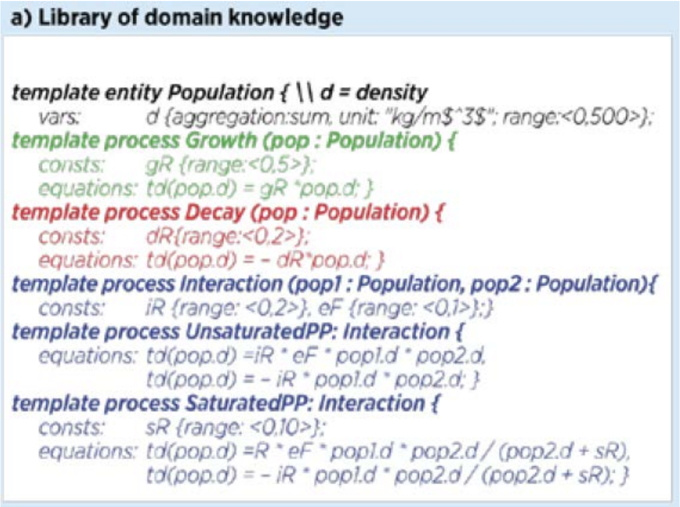
Context-free grammars (CFGs) encode language bias / hard constraints

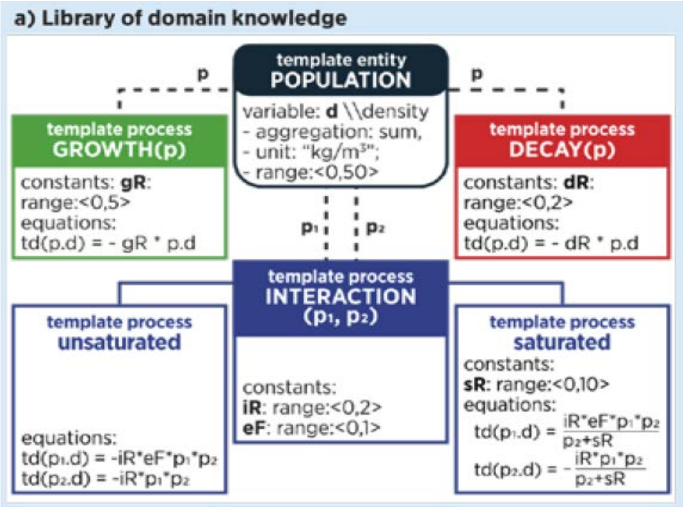
Probabilistic CFGs can be used to express preferences / soft constraints, e.g., parsimony

AUTOMATED SCIENTIFIC MODELLING

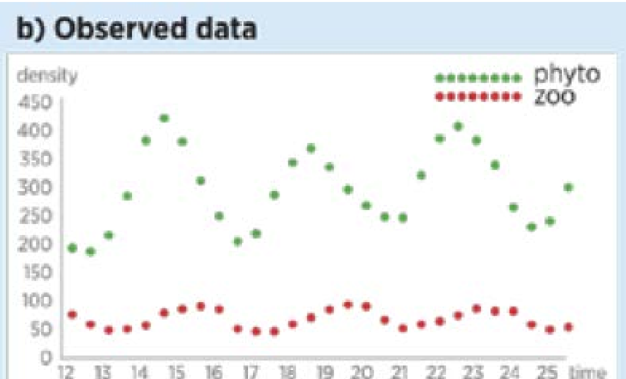
Integrating data-driven and knowledge-driven modeling

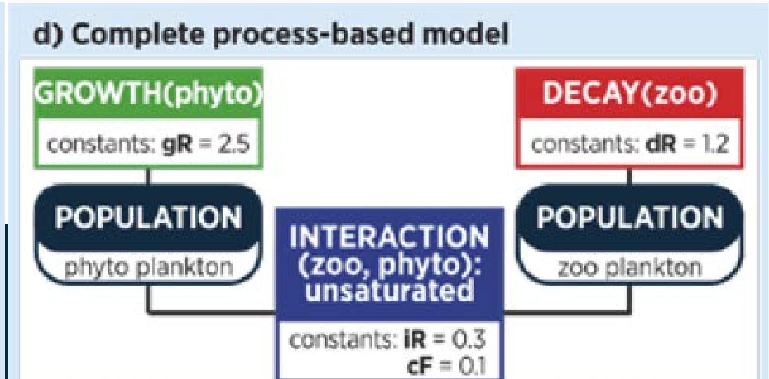
Use both data and domain knowledge as input: ***Process-based Libraries***

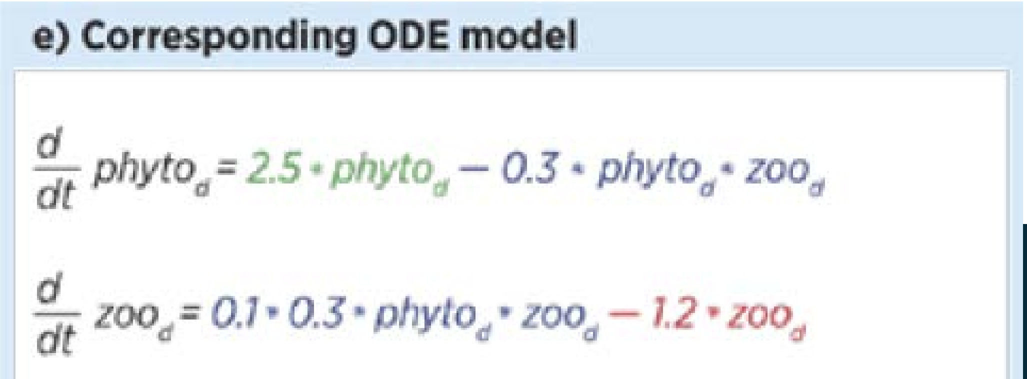




Using both data and domain knowledge yields interpretable models



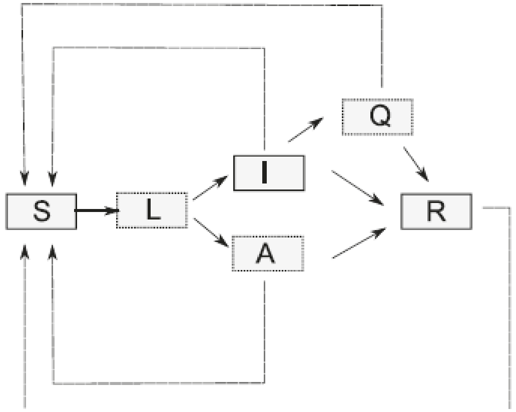


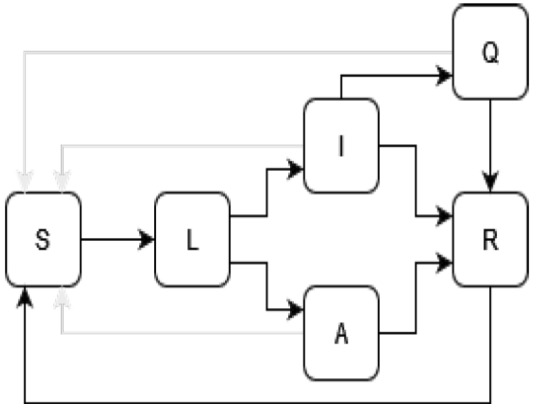


EXAMPLE APPL. OF AUTOMATED SCI. MODELLING

Learning epidemiological models

- Eyam plague outbreak (SIR, SLIR)
- Tristan da Cunha influenza outbreak (SLIR, SIR)
- COVID-19 in different countries (SLIAQRS): Slovenia, 2020





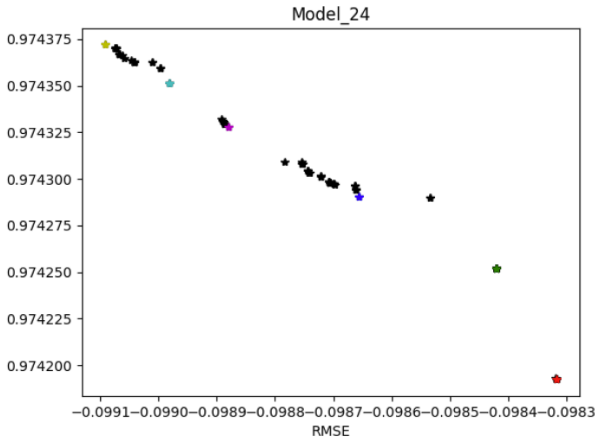
Tanevski et al. BMC Systems Biology (2016) 10:30
DOI 10.1186/s12918-016-0273-4

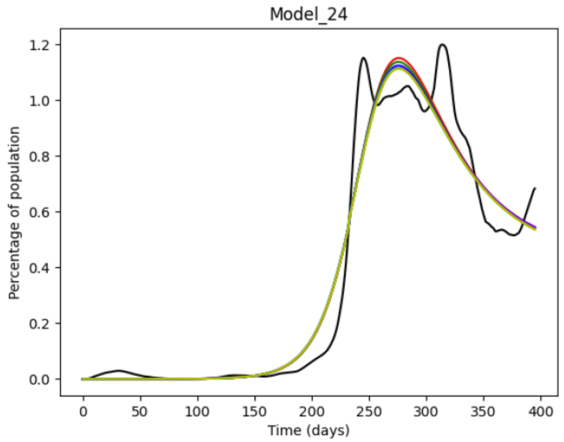
BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Learning stochastic process-based models of dynamical systems from knowledge and data

Jovan Tanevski[1,2*], Ljupčo Todorovski[3] and Sašo Džeroski[1,2]





Tuning Models for Predicting COVID-19 Infections: A Multiobjective Optimization Approach

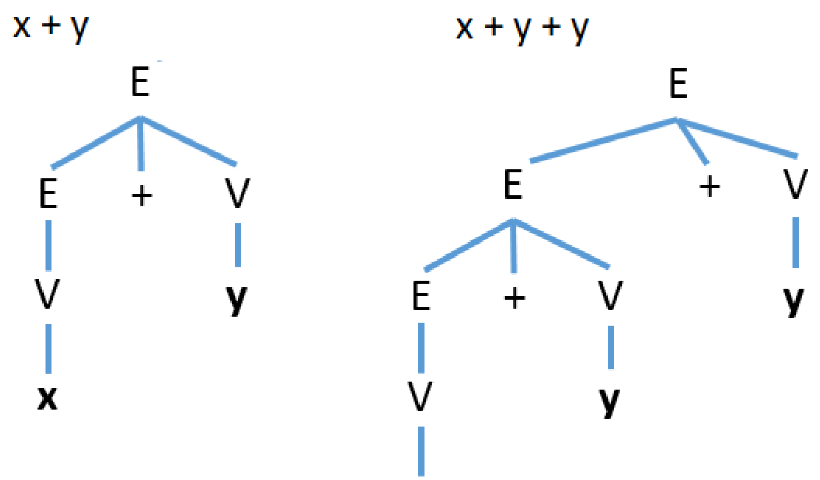
Andrejaana Andova[1,4], Jovan Tanevski[2,3], Sašo Džeroski[3,4] and Bogdan Filipič[1,4*]

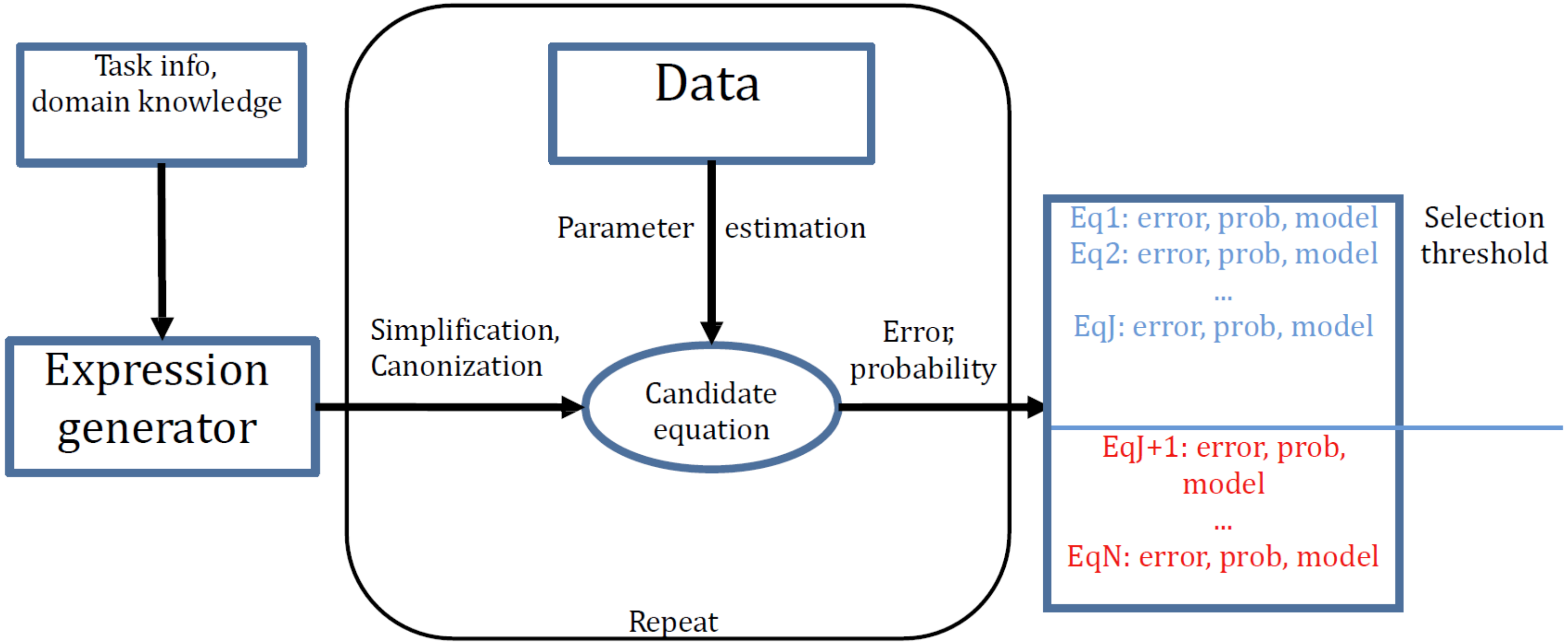
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

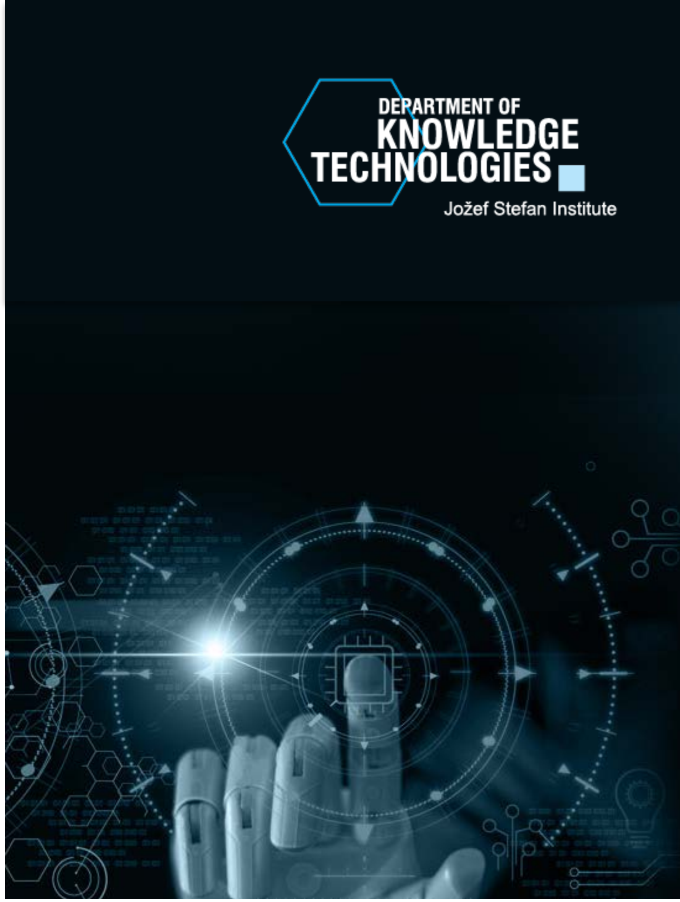
GENERATIVE MODELS IN EQUATION DISCOVERY

Probabilistic context-free grammars (PCFG) + Monte-carlo sampling

$$
\begin{aligned}
E &\rightarrow E + V\ [p] \\
E &\rightarrow V\ [1-p] \\
V &\rightarrow x\ [q] \\
V &\rightarrow y\ [1-q]
\end{aligned}
$$







Probabilistic grammars for equation discovery

Jure Brence [a,b,*], Ljupčo Todorovski [c,a], Sašo Džeroski [a,b]

[a] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
[c] University of Ljubljana, Faculty of Public Administration, Gosarjeva ulica 5, 1000 Ljubljana, Slovenia

Machine Learning (2024) 113:7689–7721
https://doi.org/10.1007/s10994-024-06522-1

Probabilistic grammars for modeling dynamical systems from coarse, noisy, and partial data

Nina Omejc[1,2] · Boštjan Gec[1,2] · Jure Brence[1,2] · Ljupčo Todorovski[1,3] · Sašo Džeroski[1]

LEARNING GENERATIVE MODELS FOR EQUATION DISCOVERY

Learn (probabilities in) probabilistic grammars from corpora of equations

Initial grammar

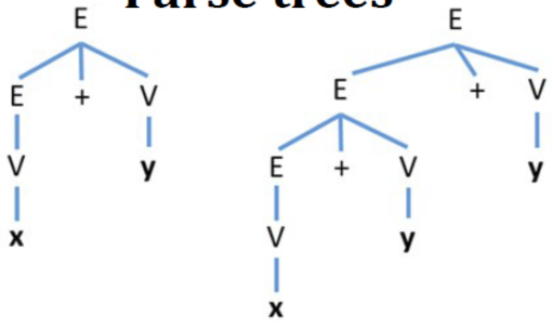
$$
\begin{aligned}
E &\rightarrow E + F\ [0.2]\ |\ E - F\ [0.2]\ |\ F\ [0.6] \\
F &\rightarrow F * T\ [0.2]\ |\ F / T\ [0.2]\ |\ T\ [0.6] \\
T &\rightarrow R\ [0.2]\ |\ V\ [0.4]\ |\ c\ [0.4] \\
R &\rightarrow (E)\ [0.6]\ |\ \sin(\mathrm{E})\ [0.1]\ |\ \cos(\mathrm{E})\ [0.1] \\
&\rightarrow \sqrt{E}\ [0.1]\ |\ \exp(E)\ [0.1]
\end{aligned}
$$

Parse

Corpus of equations

n_rho*mom*tanh(mom*B/(kb*T))
mom*H/(kb*T)+(mom*alpha)/(epsilon*c**2*kb*T)*M
mom*(1+chi)*B
Y*A*x/d
Y/(2*(1+sigma))
1/(exp((h/(2*pi))*omega/(kb*T))-1)
(h/(2*pi))*omega/(exp((h/(2*pi))*omeg

Parse trees



Production frequencies

Updated grammar

$$
\begin{aligned}
E &\rightarrow E + F\ [0.10]\ |\ E - F\ [0.15]\ |\ F\ [0.75] \\
F &\rightarrow F * T\ [0.36]\ |\ F / T\ [0.24]\ |\ T\ [0.40] \\
T &\rightarrow R\ [0.15]\ |\ V\ [0.72]\ |\ c\ [0.13] \\
R &\rightarrow (E)\ [0.56]\ |\ \sin(\mathrm{E})\ [0.12]\ |\ \cos(\mathrm{E})\ [0.09] \\
&\rightarrow \sqrt{E}\ [0.14]\ |\ \exp(E)\ [0.09]
\end{aligned}
$$

Learn Hierarchical Variational Autoencoders from sample generated from PCFG

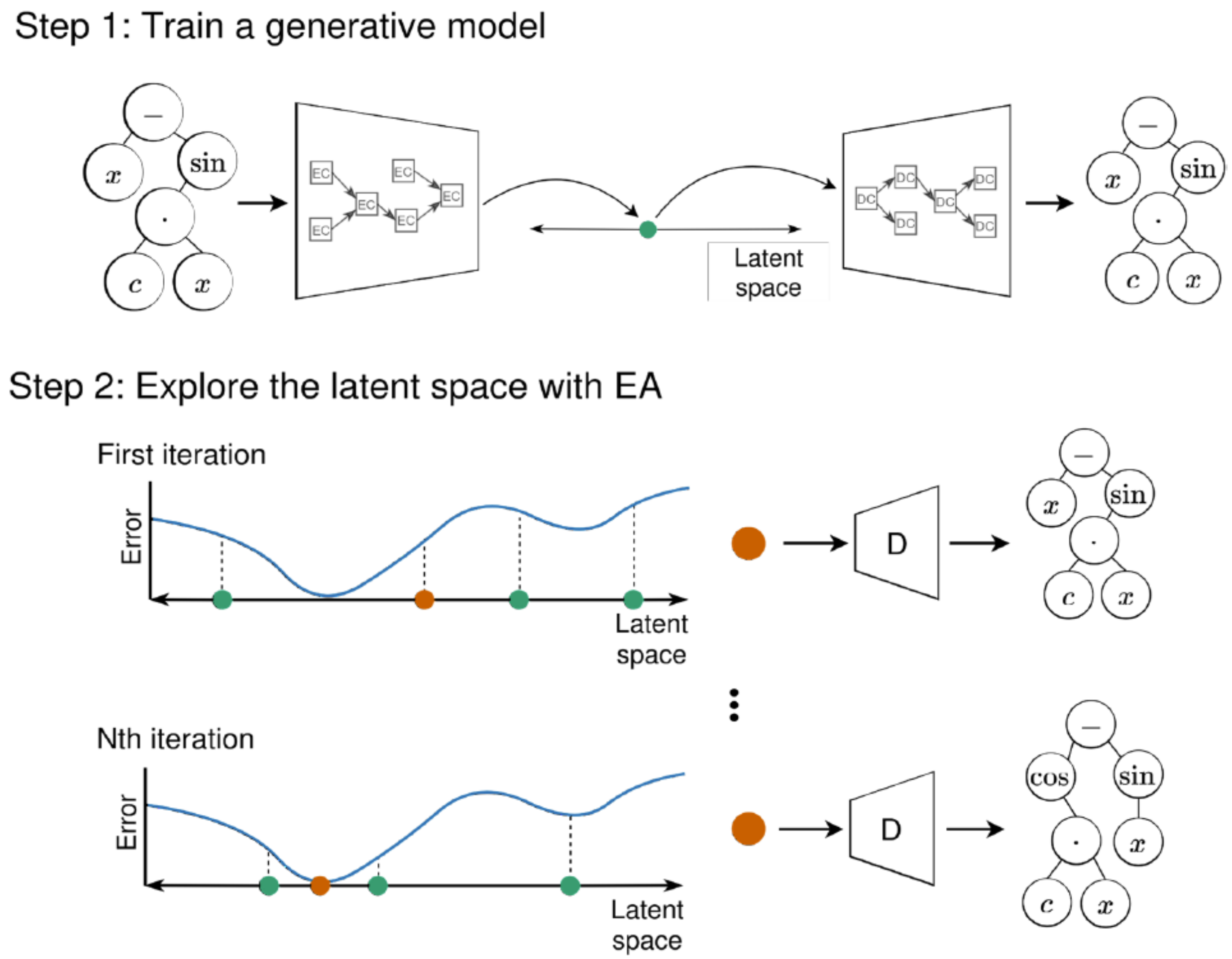
Then use HVAE for searching the latent space

Machine Learning (2023) 112:4563–4596
https://doi.org/10.1007/s10994-023-06400-2

EQUATION DISCOVERY WITH HVAEs

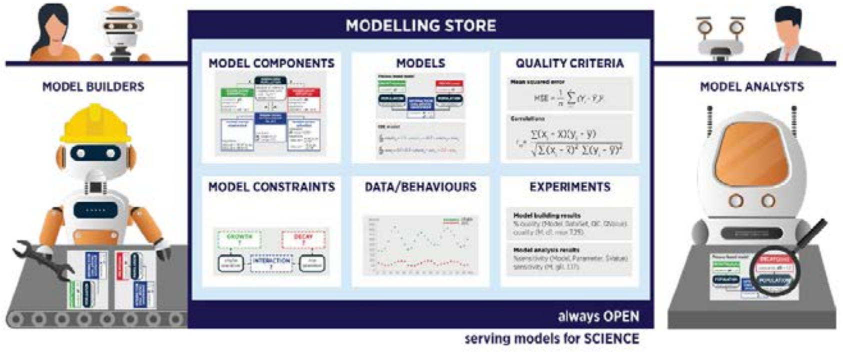
Efficient generator of mathematical expressions for symbolic regression

Sebastian Mežnar[1,2] · Sašo Džeroski[1] · Ljupčo Todorovski[1,3]



- Use a hierarchical variational autoencoder
- Embed expressions into a low-dimensional vector space
- Search the latent space with evolutionary algorithms to find best-fitting expressions.

A COLLABORATIVE VISION OF AUTOMATED SCIENTIFIC MODELLING



AI/Machine learning approaches should be capable of functioning within the scientific knowledge ecosystem and contributing to the pool of scientific knowledge

Using both observations and existing knowledge/models

Producing new knowledge (models) that can be used further (by both humans and machines)

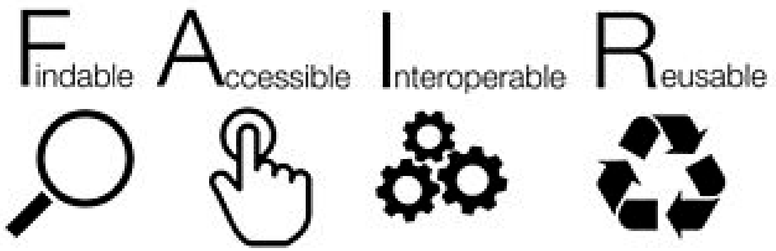
To this end, we need to

- Integrate knowledge-driven and data-driven modeling
- Data and knowledge need both to be first-class citizens, so that one can store, query/find, retrieve and reuse them
- Representations for models and knowledge should be close to those used by humans in scientific modeling

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

SEMANTIC TECHNOLOGIES FOR OPEN SCIENCE

THE PHILOSOPHY OF OPEN SCIENCE



Scientific knowledge should be shared public good, *accessible to all and not restricted by paywalls, proprietary systems, or institutional silos. By making research outputs openly available, reproducible, and reusable, open science aims to accelerate discovery, foster trust, ensure benefits of science reach society as a whole.*

Yes, scientific data should be FAIR. *But this is an understatement.*

Not only data, all research outputs/ artefacts should be FAIR/ openly accessible. This includes publications, code, models, experimental protocols, results of experiments…

We need to represent, annotate and store them, so that they can be found and re-used. *All outputs of the scientific process must be formally described and recorded.*

This is good for collaboration between scientists, whether they are humans or machines.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

FORMALISATION OF SCIENCE

Increasing effort is invested into automating (parts of science)

Robot scientists (aka self-driving labs) are computer/ robotic systems, capable of originating their own experiments, physically executing them, interpreting the results, and then repeating the cycle.

To automate an activity, we need to be able to thoroughly understand it and describe it

To automate science, we need to formalize it

- *The goal of science is to increase our knowledge of the natural world through the performance of experiments*
- *This knowledge should be expressed in formal logical languages*
- *Formal languages promote semantic clarity, which supports free exchange of scientific knowledge and simplifies scientific reasoning*

Ontologies provide controlled vocabularies for describing experiments and their outcomes

ONTOLOGIES FOR EMPIRICAL COMPUTER SCIENCE

Ontologies for data mining/ machine learning

Ontologies for optimization

Describe generic and specific tasks, methods/ algorithms, performance

Controlled vocabularies for describing experiments and their outcomes, e.g., which ML algorithms run on which dataset with what results (different metrics, cca 20 for MLC)

Essential for experiment databases, such as Open ML

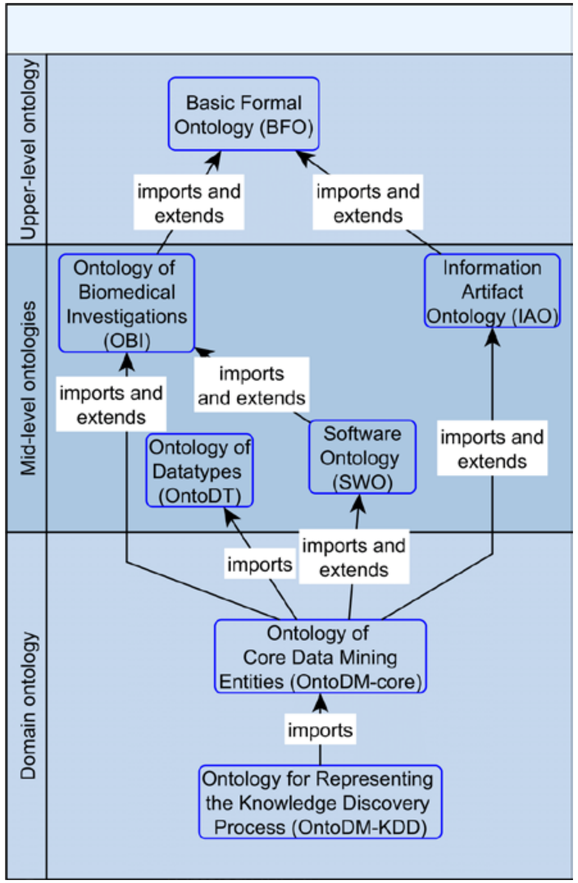
Facilitate meta-learning, Auto ML; AutoOpt; AutoLLM; AutoAI

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OntoDM: Ontologies for Machine Learning

- OntoDM is a modular ontology
- Modules can be used together or independently depending on the use case

- OntoDT – Ontology of datatypes
- OntoDM-core – Ontology of core data mining entities
- OntoDM-KDD – Ontology of the knowledge discovery process



P. Panov, L. N. Soldatova, S. Džeroski (2014) "Ontology of core data mining entities", Data Mining and Knowledge Discovery 28(5-6):1222-1265

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OntoDM: Ontologies for Machine Learning

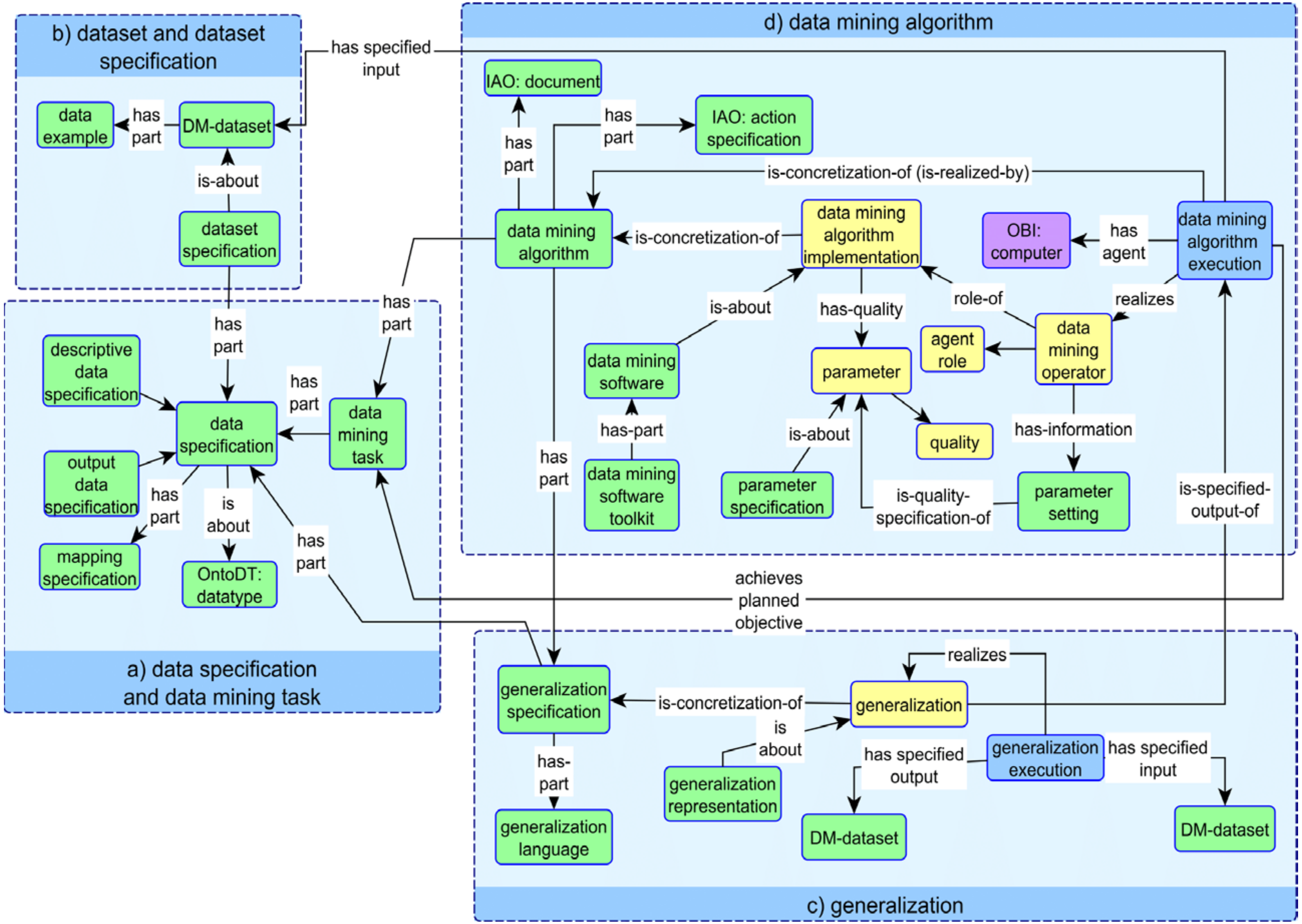
- ***OntoDT represents knowledge about datatypes***
  - *The notion of a datatype is very important in ML/DM*
  - *Characterizes the kind of data contained in a dataset*
  - *Allow to define data mining tasks on data of different datatypes*
  - *Determines applicability of a data mining algorithm on a dataset*
  - *OntoDT can support a wide range of applications, not only ML/DM*
- ***Ontology of core data mining entities - OntoDM-core***

  Contains the most essential data mining entities
  - *Data specification*
  - *Dataset*
  - *Data mining task*
  - *Data mining algorithm*
  - *Generalizations (patterns, models)*

  Includes taxonomies of datasets, data mining tasks, generalizations, data mining algorithms based on the type of data

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OntoDM: Ontologies for Machine Learning



OntoDM Use Case: Ontologies for Machine Learning

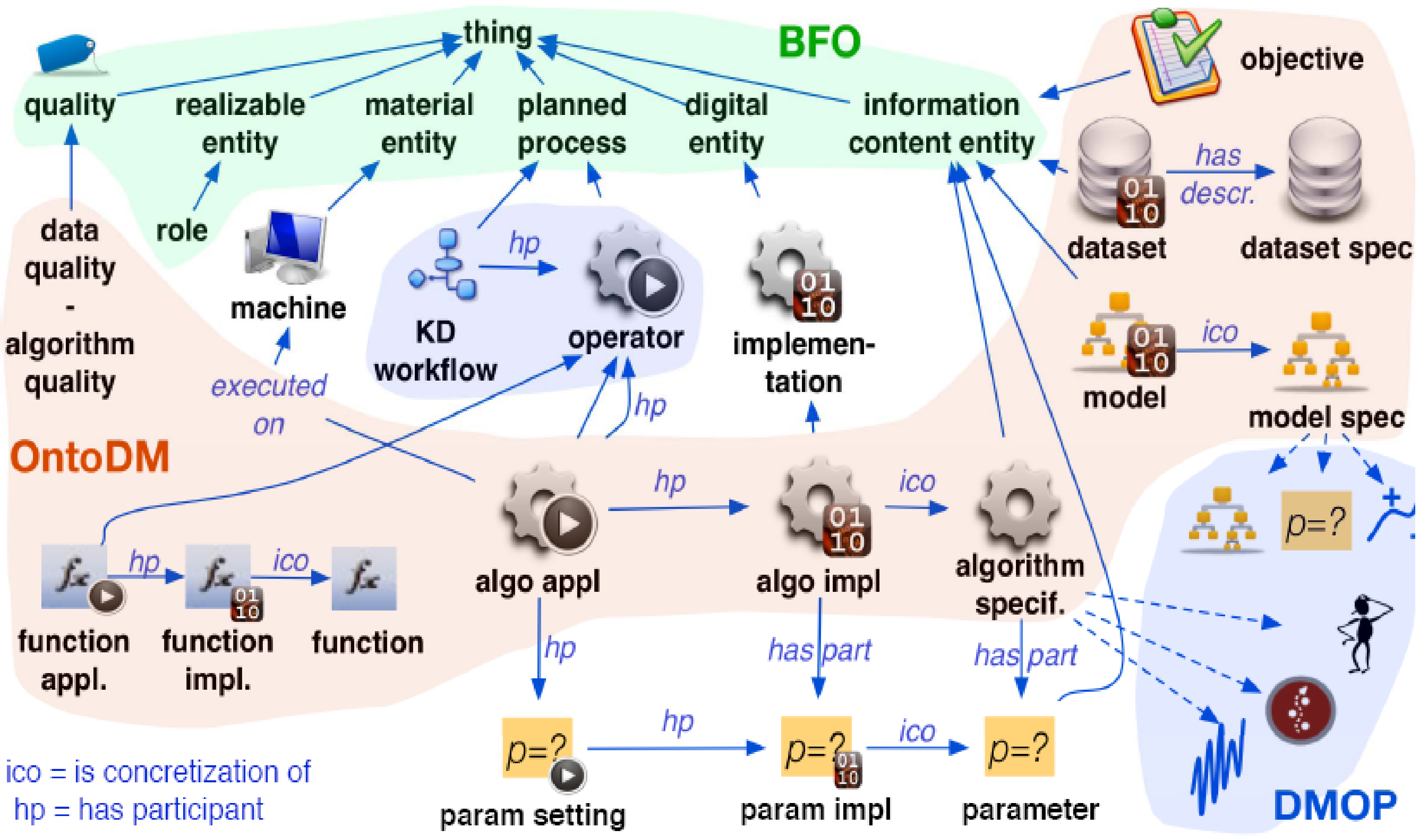
- ***Store the results of machine learning experiments***
  - *Many different machine learning algorithms*
  - *Run on many different datasets*
  - *With many different parameter settings*
  - *Described in enough detail to be reproducible*

- ***The machine learning experiments are described***
  - *In a mark-up language (EXP-ML), which is based on*
  - *An ontology (EXPOSE), whish in turn is largely based on OntoDM and imports many of its classes*

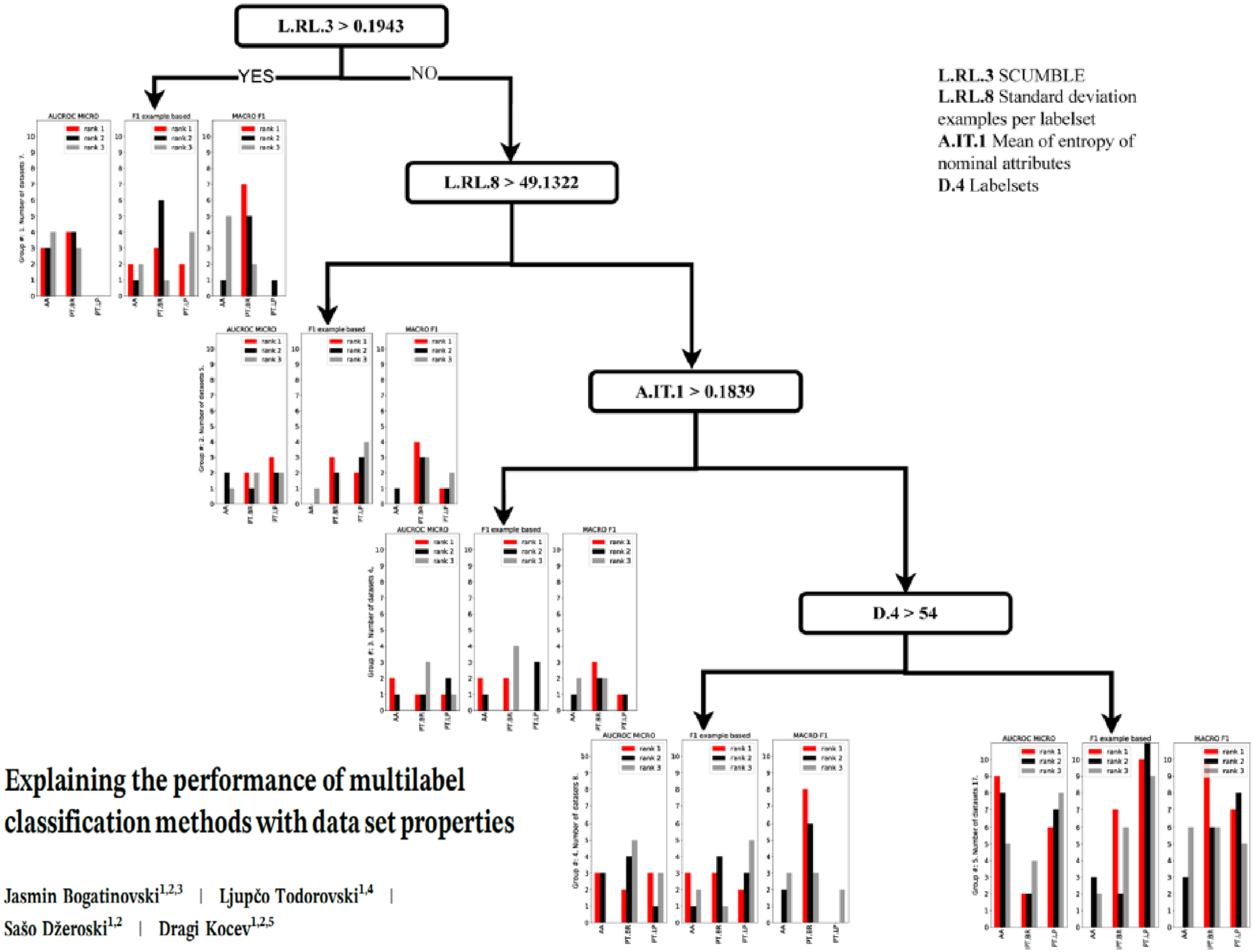
- ***Initial version: http://expdb.cs.kuleuven.be***
- ***Current version: http://www.openml.org/***
- ***OpenML data used for meta-learning, driving AutoML, based on the type of data***

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OntoDM Use Case: EXPOSE/ ONTO-DM



OntoDM Use: Meta-Learning in MLC Benchmarking



Explaining the performance of multilabel classification methods with data set properties

Jasmin Bogatinovski[1,2,3] | Ljupčo Todorovski[1,4] |
Sašo Džeroski[1,2] | Dragi Kocev[1,2,5]

Predict relative performance of classes of MLC methods from dataset properties

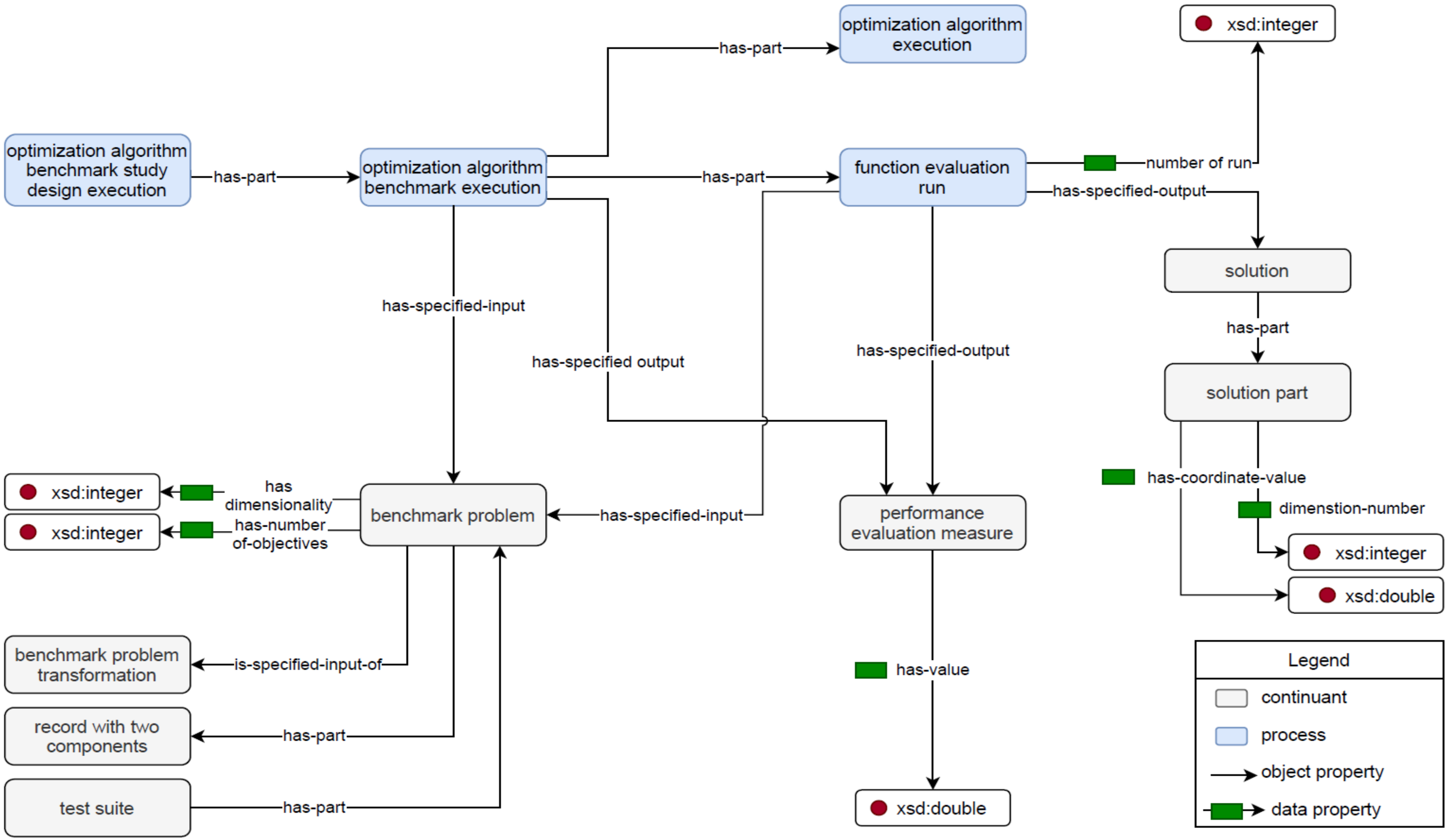
- *Algorithm adaptation (AA)*
- *Problem transformation (PT)*
  - *Binary relevance (BR)*
  - *Label powerset (LP)*

Trying to understand

- *Which algorithms*
- *Work best for which classes of problems*
- *According to which metrics*

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

OPTION: Ontologies for Optimization



IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 27, NO. 6, DECEMBER 2023

OPTION: OPTImization Algorithm Benchmarking ONtology

Ana Kostovska, *Graduate Student Member, IEEE*, Diederick Vermetten, Carola Doerr, Sašo Džeroski, Panče Panov, and Tome Eftimov, *Member, IEEE*

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

How does it all fit together?

- ***The notion of explanation is of central importance to the scientific enterprise***
    - *Trees and ensembles, for MTP, built in fully and semi-supervised fashion, are either interpretable themselves (can be small and accurate) or provide explanations of predictions*
    - *The use of domain knowledge facilitates interpretation/ explanation (both in relational trees and models of dynamical systems)*
- ***Scientific data can be small, but combination with domain knowledge helps***
    - *Both in explainable ML and the use of foundation models*
    - *Semi-supervised learning also helps, can be combined with DK*
- ***Foundation models can be used when data do not abound, and can also help in making data interoperable***
- ***Semantic technologies (ontologies) help in formalizing & sharing scientific knowledge and collaboration (for humans and machines)***

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

AI FOR SCIENCE@JSI: PROJECTS

ARTIFICIAL INTELLIGENCE FOR SCIENCE @ JSI
National projects, Gravity framework

Artificial Intelligence for Science (GC-001): AI4Sci

The project focuses on the development of AI approaches along the four major directions outlined in my talk and their use in materials sciences and life sciences

AI & Materials Sciences (GC-004): CHRONOSTORE

Chemical Energy Storage Solutions Across Temporal Scales for Climate-Resilient Renewable Energy Systems

AI & Life Sciences/ Personalized Medicine (GC-005): BOOST

Bridging oligonucleotide-based therapeutic upregulation of translation

AI & Environmental/Earth Sciences (GC-006): GeoAI

Geospatial information technologies for resilient and sustainable society

ARTIFICIAL INTELLIGENCE FOR SCIENCE @ JSI
EU funded projects

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ELIAS:

European Lighthouse of Artificial Intelligence for Sustainability

LLMs4EU:

Large Language Models for the European Union

AutoLearn-SI:

Leveraging Benchmarking Data for

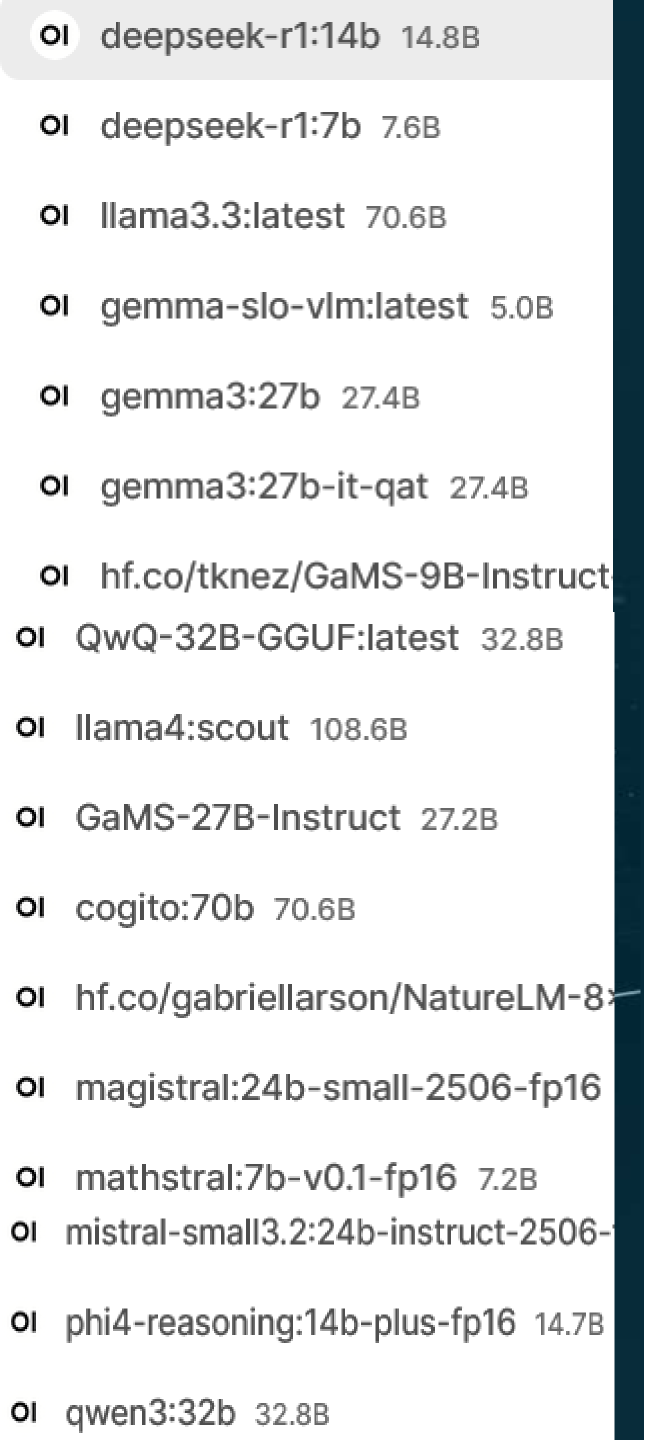
Automated Machine Learning and Optimization

Slovenian Artificial Intelligence Factory (SLAIF)

Funded by
the European Union

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

AI FOR SCIENCE: INFRASTRUCTURE



AI AS INFRASTRUCTURE FOR SCIENCE (llm.ijs.si)

Local infrastructure at JSI, serving LLMs to researchers at JSI

At the end of March 2024, we released a preliminary service, based on the LLM model Mistral7B

By October 2024, we had added a few additional LLMs

We have a local setup of open-sourced LLMs, hosted on JSI infrastructure, available via a chat assistant & programmatic API

- Fast and scalable implementation using Paged Attention
- Up to few thousand concurrent connections

Different LLMs available at llm.ijs.si for different purposes

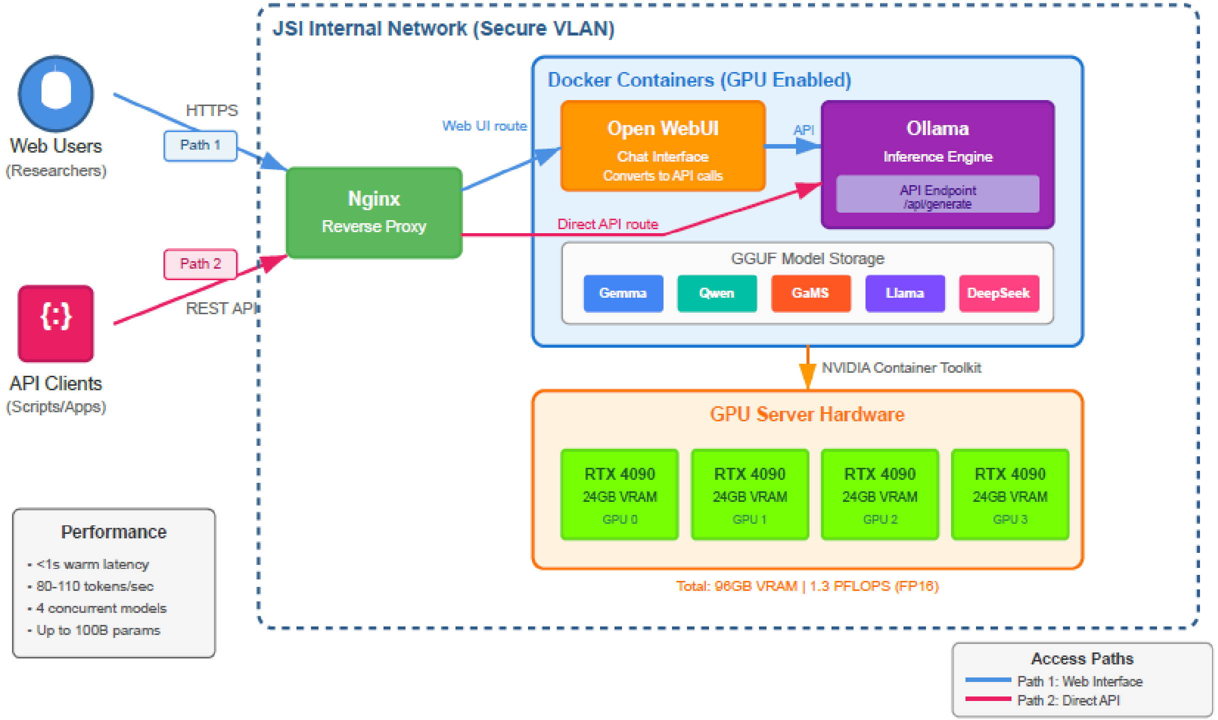
- Multiple models are implemented concurrently, including DeepSeek and MathModels – Mathstral, which are specialized in custom domains.
- Current list (left) includes GAMS, a LLM for Slovene
- This specialization allows them to deliver tailored solutions for specific applications, enhancing their effectiveness and performance in niche areas.

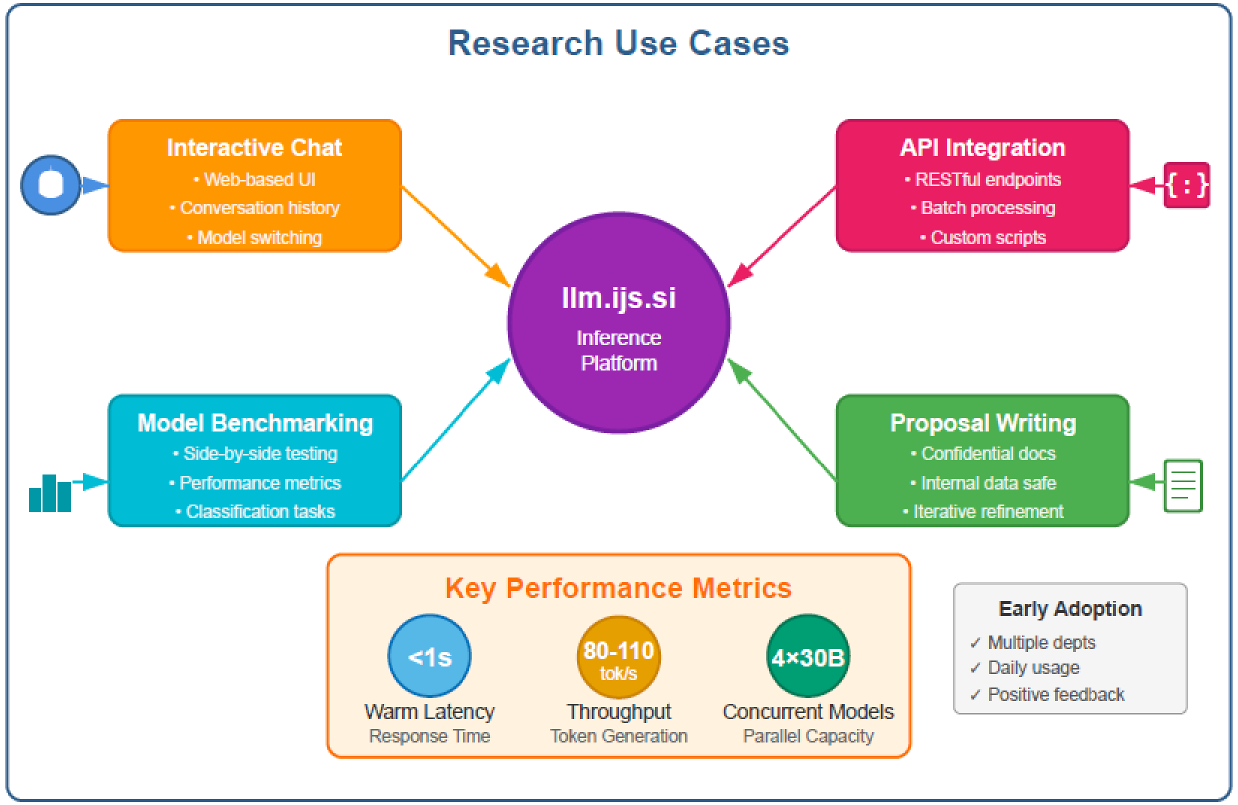
llm.ijs.si: MOTIVATION, IMPLEMENTATION, USE CASES

Motivation

- Need for secure, on-premise LLM infrastructure for scientific research.
- Data privacy concerns with cloud-based LLM services.
- Requirements for reproducible benchmarking of open-source models.
- Support for sensitive/proprietary data processing within institutional boundaries.

llm.ijs.si System Architecture





AI IN EUROPE: REGULATION, INFRASTRUCTURE & INNOVATION

The AI Act & The AI Innovation Package

The 2024 AI Act

- The **world's first** comprehensive AI law
- A **horizontal** act with a **risk-based** approach – different rules for different risk levels
- Protection of **health, safety** and **fundamental rights**
- **Innovation friendly**
- Will apply to **public and private** actors, **inside and outside the EU** (as long as the AI system or general-purpose AI model is placed on the Union market), **providers and deployers**
- Sets up the **AI Office -** centre of AI expertise ; foundation of a **single European AI governance** system; key role in implementing AIA

The 2024 AI Innovation Package

- **Boosting startups and innovation in trustworthy AI**
- **Two main components**

(1) AI Factories:

Making available HPC computing capacity to facilitate the development of GenAI models/ Applications

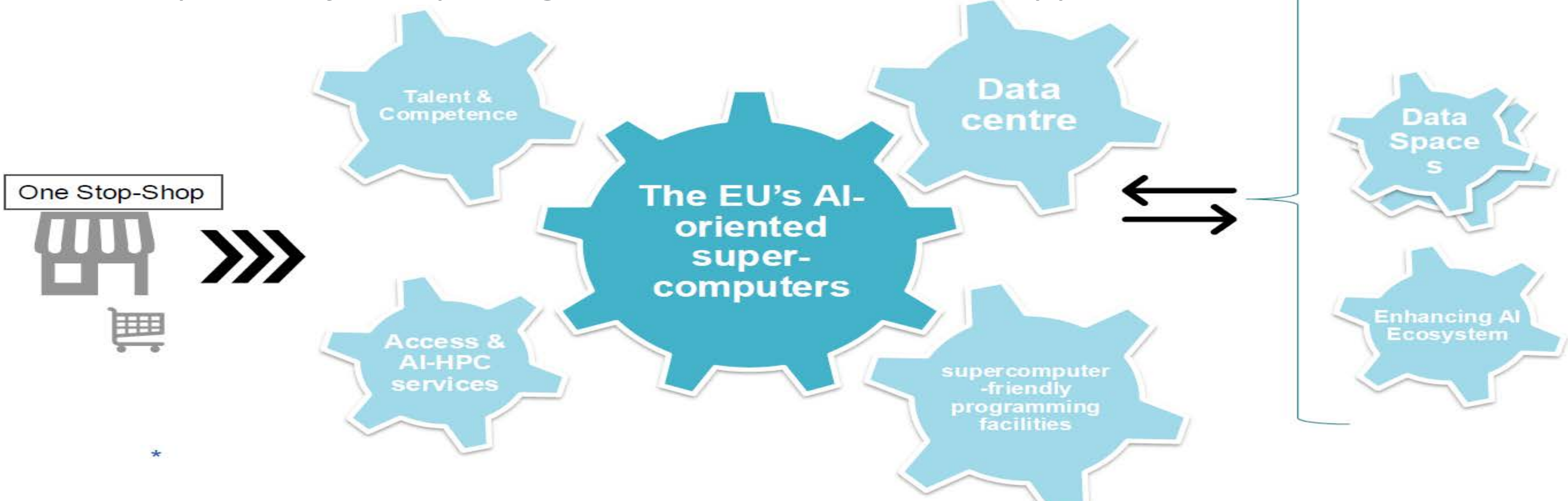
(2) AI support measures with GenAI4EU:

Stimulating the development in strategic sectors of novel and innovative applications based on GenAI models and facilitating their uptake

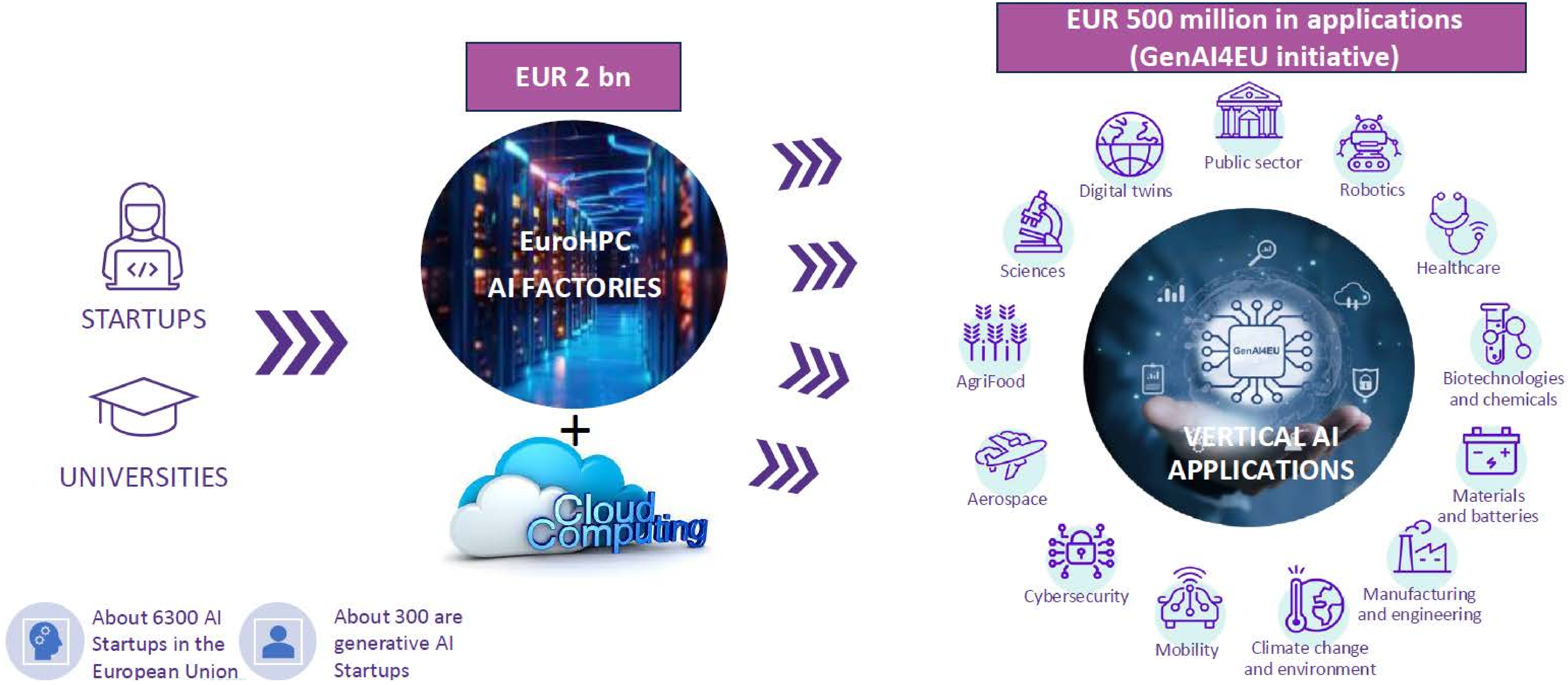
JUST WHAT ARE AI FACTORIES?

AI factories have two main components

- *AI-optimized supercomputer (AI-HPC)*
- *AI-factory activities and services*
- *AI Factories will provide the computing power, data, support and talent, to offer a wide and exhaustive range of services to AI startups and researchers needed for the development of European generative AI models and applications*

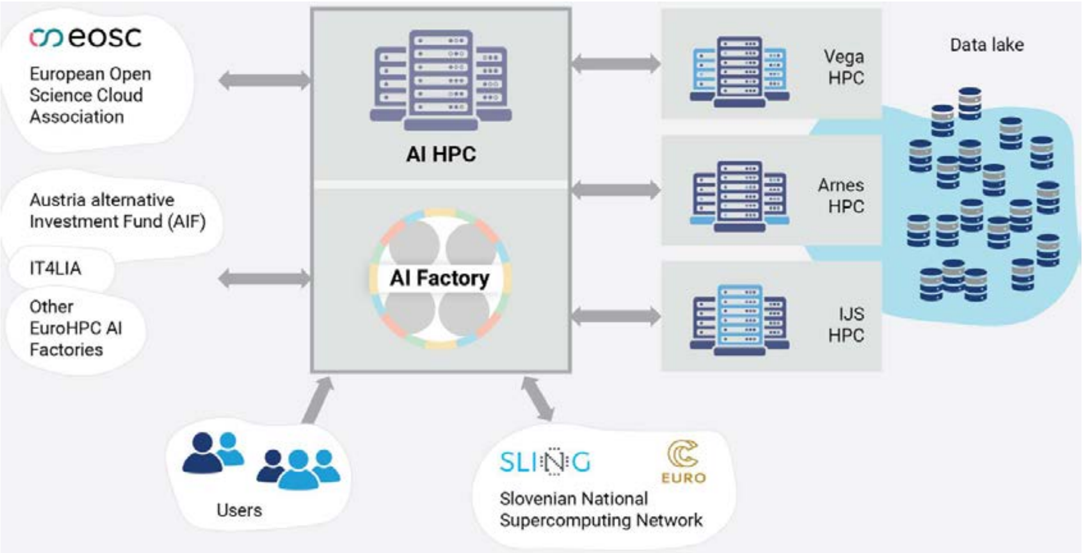


AI FACTORIES: STRENGTHENING THE AI ECOSYSTEM



THE SLOVENIAN AI FACTORY: INFRASTRUCTURE FOR SCIENCE & SOCIETY

The Slovenian AI factory will support the growth of the Slovenian AI ecosystem



EuroHPC
Joint Undertaking

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PREOBRAZBO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO, ZNANOST IN INOVACIJE

At JSI, we have viewed AI as infrastructure for science for quite a while

SLAIF will support industry, public sector, science: AI infrastructure for the whole society

*AI HPC Consortium: **IZUM (hosting entity)**, JSI, ARNES*

*AI Factory Consortium: **JSI (technical coordinator)**, UniLj, UniMb, FIS NM, UniNG, UniPr, IZUM (administrative coordinator), ARNES, GZS, TPLj*

Key workpackages of the SLAIF project

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

- **Workflow and Data Orchestration Infrastructure**:
  Develops robust infrastructure integrating HPC and cloud resources for scalable execution of AI workflows and establishes a national data lake enabling automated data exchange.

- **Core AI Platform**: Delivers AI workflows as generic horizontal services, maintains semantic catalogue of AI components, and supports development and deployment of AI workflows.

- **User-centric AI Services**: Focuses on making HPC/AI clusters more user-friendly for first-time users, providing libraries, tools, and interfaces for seamless connection to services.

- **Vertical Services/ Sectorial Applications**

- **AI Skills and Training**

APPLICATION AREAS FOR THE SLOVENIAN AI FACTORY

Demonstrate the impact of AI across multiple sectors, developing specialized applications that address specific needs of industry, society and science. Leverage core AI platform to deliver tailored solutions.

AI for Green Transition

Revolutionizing agriculture with precision farming, environmental monitoring with EO data, optimizing energy systems, and enhancing smart manufacturing for sustainability.

AI for Health and Biotechnology

Analyzing complex biosignals for early disease detection, delivering personalized treatment plans, accelerating drug discovery, and developing tools for medical decision-making.

AI for Digital Society

Adapting language models for Slovene and other less-resourced languages, empowering creative industries, streamlining public administration, and transforming education.

AI for Science

Automating scientific model discovery, accelerating life sciences research, facilitating materials science innovation, supporting environmental sciences, and advancing digital humanities.

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

AI FOR SCIENCE: POLICY, EVENTS, COMMUNITY

EU and National Policies on AI in Science

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

EUROPEAN STRATEGY FOR AI IN SCIENCE

The European Commission is currently crafting a European strategy for AI in Science

- *A **call for evidence** and a **targeted questionnaire** (open until June 2025) were used to gather input from researchers, funders, infrastructure providers, and other stakeholders*
- ***High-level roundtables** (the first one in December 2024, the second one in June 2025): Commissioner and VP discussed concrete actions for implementation with prominent scientists*
- *The scientific community (incl. CERN & EMBL) has proposed*
  - *The formation of a **high-level AI Research Council** that would guide overall vision, strategy, and governance*
  - *Accompanied by a distributed network of **Centres of Excellence**, some possibly hosted in large European research organizations and interconnected with existing initiatives like CAIRNE, ELLIS*

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

RAISE: Resource for AI Science in Europe

"The Strategy will also lead to the creation of a European AI Research Council. This Council will take the form of a Resource for AI Science in Europe (RAISE). …"

The purpose and goals of RAISE are to

- *Accelerate the responsible adoption of AI in scientific research.*
- *Improve access to state-of-the-art AI tools, infrastructure, and computing capacity.*
- *Attract talent, investment, and foster collaboration across Member States.*

Focus areas *include: health & drug discovery,*
climate change & clean technologies, materials science, …

Currently open call: HORIZON-CL4-INDUSTRY-2025-01-DIGITAL-61
AI Foundation models in science (GenAI4EU) (RIA)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

COMING UP NEXT

- ***Scientific conference***

  Artificial Intelligence for Science

  22-26 September, Ljubljana, Slovenia

- ***Policy conference***

  AI in Science Summit

  3-4 November, Copenhagen, Denmark

- *Edited book*

  S. Dzeroski, Y. Choi, N. Kutz, P. Langley, L. Soldatova (editors)

  Computational Approaches to Scientific Discovery

  Based on the very successful AAAI Symposium on

  Computational Approaches to Scientific Discovery,

  held March 27-29, 2023 at San Francisco Airport

  (as part of the AAAI Spring Symposium Series)

DS 2025

ai4science.si

Ljubljana, Slovenia, September 22-26, 2025

Artificial Intelligence for Science

Includes the **Discovery Science** conference
and many additional tracks

International Conference
AI for SCIENCE 2025

including DS-2025, the

Tracks

- 28th Discovery Science Conference
- AI & Space
- AI & Material Science
- DaFab Summer School
- AI Factories

SMASH Tracks

- AI & Digital Humanities
- AI & Environmental Science
- AI & Life Sciences
- AI & Physics

Opening Keynote

Ross King

The Automation of Science: Past, Present, and Future.

AI & Space

Lorenzo Bruzzone

Remote Sensing and the Labeled Data Challenge in the Foundation Model Era

Gilberto Câmara

Machine Learning and AI for Remote Sensing Application: Promises and Challenges

28th Discovery Science Conference

Iryna Gurevych

Please meet AI, our dear new colleague. In other words: can scientists and machines truly cooperate?

Mikel Landajuela

Deep Symbolic Optimization: Reinforcement Learning for Equation Discovery

Joaquin Vanschoren

Auto-continual learning

AI & Environmental Science

Alexander Barth

SMASH supervisor

Generative Deep Learning for Satellite Data Reconstruction

Joaquim Comas Matas

Integrating AI in Urban Water Management: Intelligent Decision Support Systems for the Twin Transition

AI & Life Sciences

Martin Michalowski

Explainable AI for Clinical Decision Support in Multimorbidity: Toward Transparent and Personalized Care

AI & Material Science

Tejs Vegge

Leveraging AI4Science and Self-Driving Laboratories to Accelerate Materials Discovery

Luka Suhadolnik

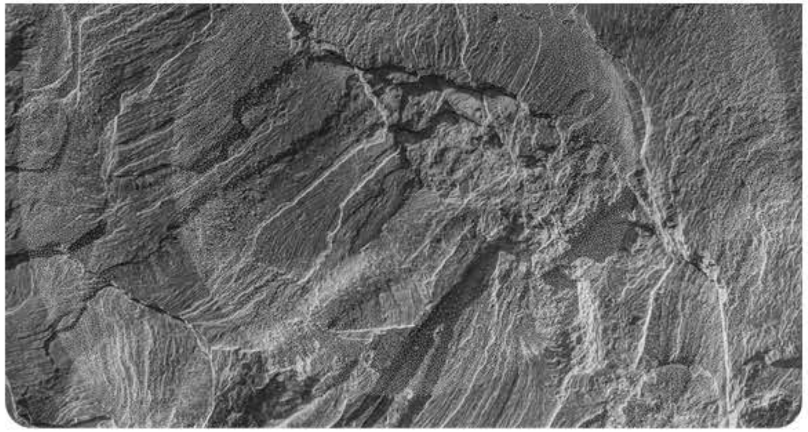
Building bridges between experiments and digital workflows: lessons from Quipnex

Julija Zavadlav

Multiscale Materials Modeling with Machine Learning Potentials

AIS25 (ais25.eu)
AI in Science Summit 2025, Copenhagen, 3-4 November 2025

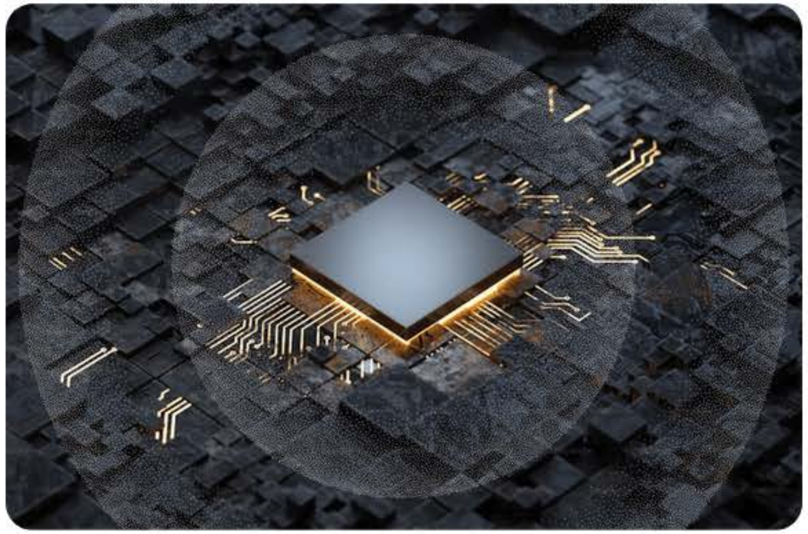
Life Science



Materials Science

Planet & Climate

Society & Community



Science of AI

Policy for AI in Science

Plenary

Yoshua Bengio
Professor, Université de Montréal

As one of the world's leading experts in AI and deep learning, Professor Yoshua Bengio will speak about ensuring a powerful and safe application of AI in science.

Christina Egelund
Danish Minister for Higher Education and Science

Ekaterina Zaharieva
Commissioner for Startups, Research and Innovation at the European Commission

Climate

Markus Reichstein

Sašo Džeroski

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

MANY THANKS

FOR YOUR ATTENTION

WE ARE HIRING! PHD STUDENTS, POSTDOCS, …

CONTACT: SASO.DZEROSKI@IJS.SI

ALSO: WHOVA JOB POST BY DRAGI KOCEV