



# Consolidated Learning

a domain-specific model-free optimization strategy with validation on metaMIMIC benchmarks

**Published in Machine Learning Journal**

**Katarzyna Woźnica** , M. Grzyb, Z. Trafas, P. Biecek  
MI2.AI, Warsaw University of Technology  
Poznan University of Technology



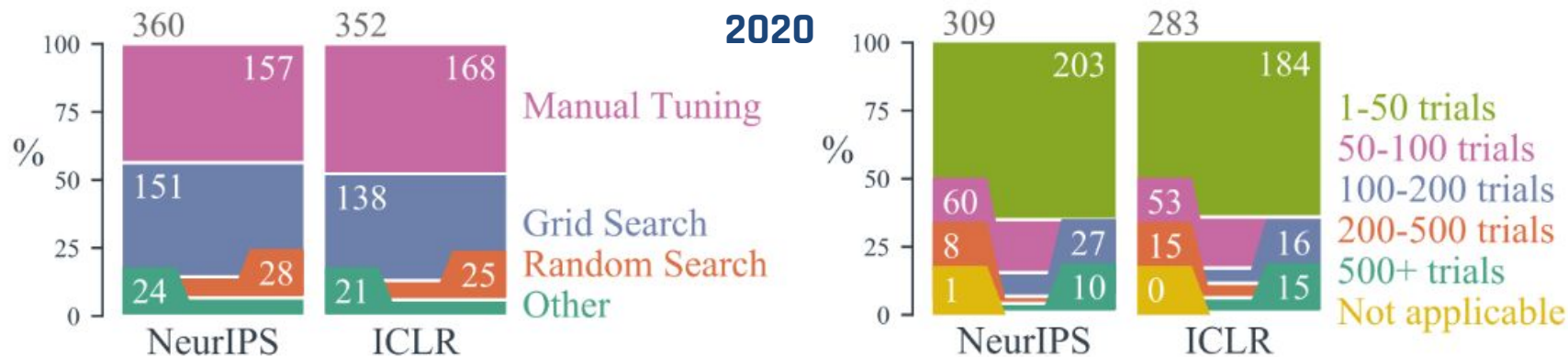
Neuro-symbolic Metalearning and AutoML | ECML 2023

**MI** RESEARCH

**Consolidated learning** is a new formulation of meta-learning in hyperparameter optimization motivated by practical challenges to build ML models for similar tasks (domain-specific).

# Importance of HPO

Algorithms are **complex**, depend on a multidimensional often hierarchical space of hyperparameters. **BUT** complex HPOs have low adoption among practitioners.



X Bouthillier, G Varoquaux, 2020. Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020

# Importance of HPO

Algorithms are **complex**, depend on a multidimensional often hierarchical space of hyperparameters. **BUT** complex HPOs have low adoption among practitioners.

**We need simple optimization methods providing  
anytime performance**

The diagram consists of three overlapping semi-circular shapes. The left shape is green and labeled 'Automation'. The right shape is blue and labeled 'Specialization'. The bottom shape is maroon and labeled 'Meta-learning'. The intersection of all three shapes is the central area.

## Automation

AutoML + HPO:  
Random Search  
Bayesian Optimization

## Specialization

Defaults based on user  
experience

## Meta-learning

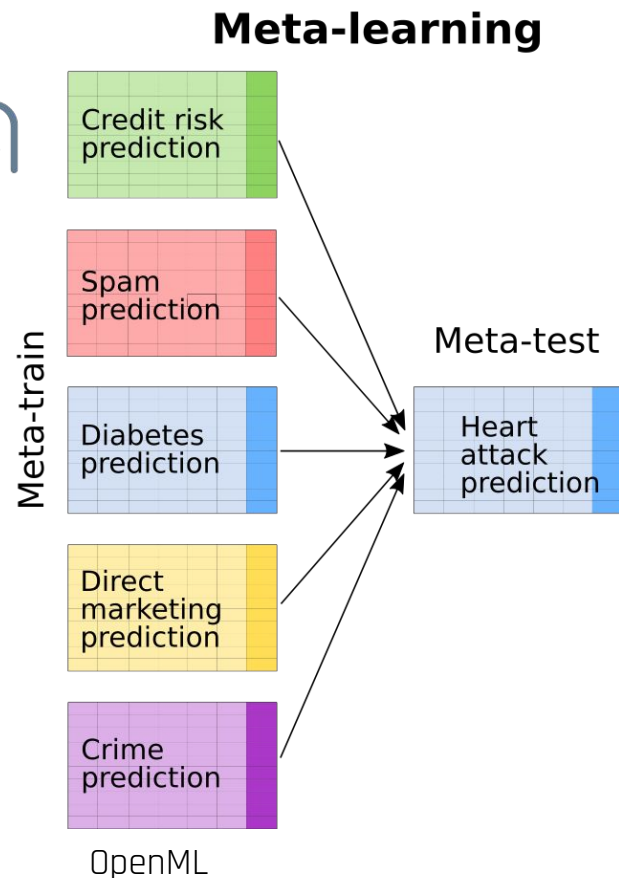
# How meta-learning for tabular dataset works

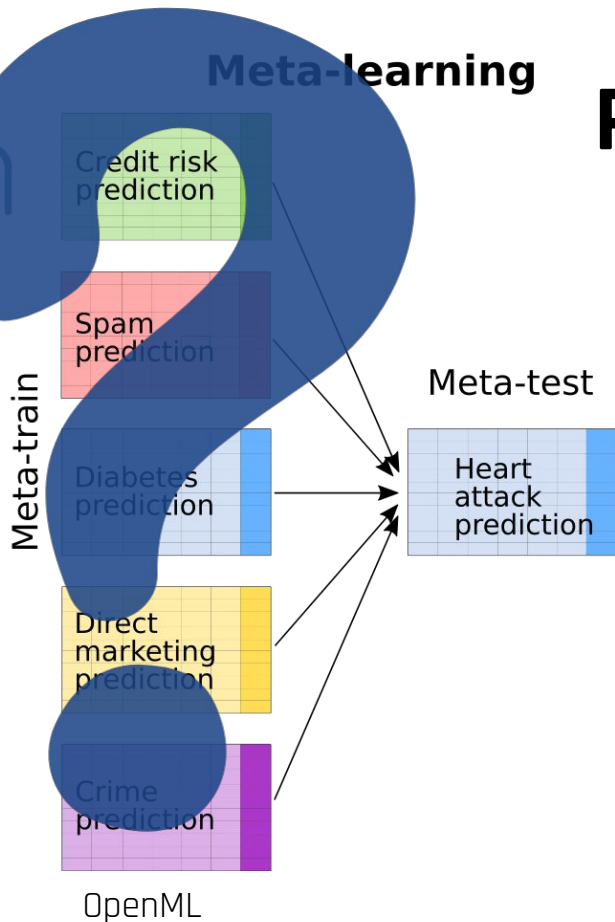


## What generally works

Idea: global (local) rankings of hyperparameters basing on meta-train

***AutoGluon, Auto-sklearn***





# Practical challenges

- **shared variables:** cognitive assessments in prediction of Alzheimer's disease;
- **related targets:** mortality prediction, as endpoints are often considered with differently defined short- and long-term mortality;
- **out-of-time data:** updating the set of observations and train the model anew.

**Domain knowledge**

**Automatization**

**Consolidated  
learning**

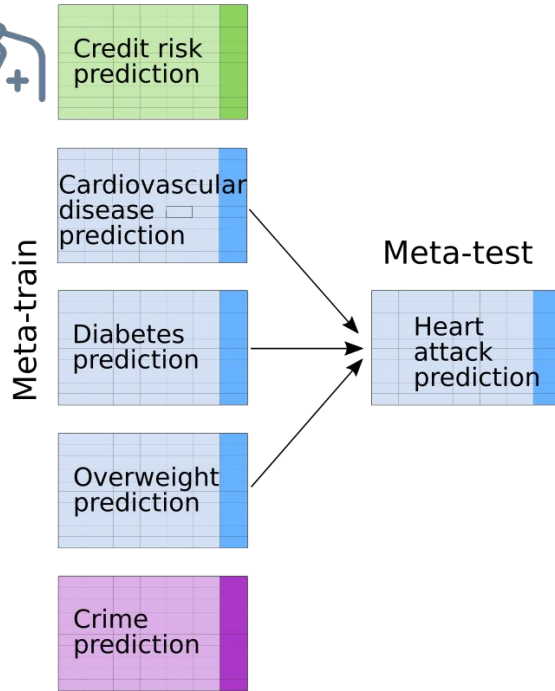
**Specialization**

**Meta-learning**





## Consolidated learning



# Consolidated Learning

- Consolidated *meta-train* - a constrained set in which common explanatory variables occur between the sets in the meta-train set and the meta-test set.
- Based on consolidated meta-training, a portfolio of hyperparameters is composed (according to any strategy) and this process is called **consolidated learning**.

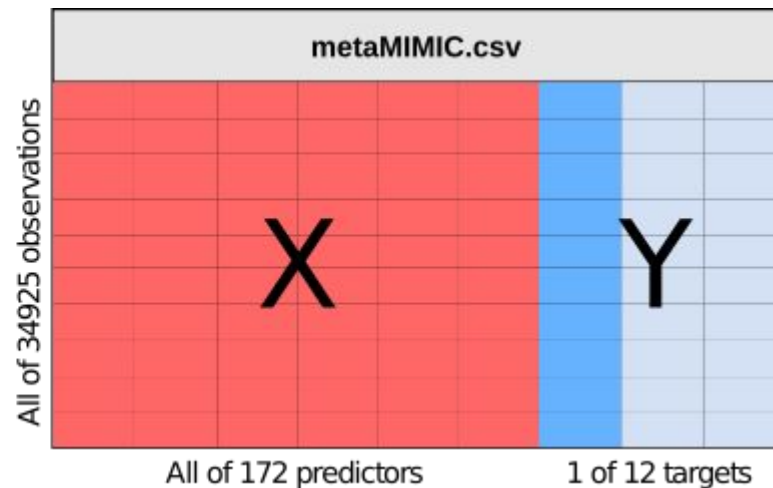
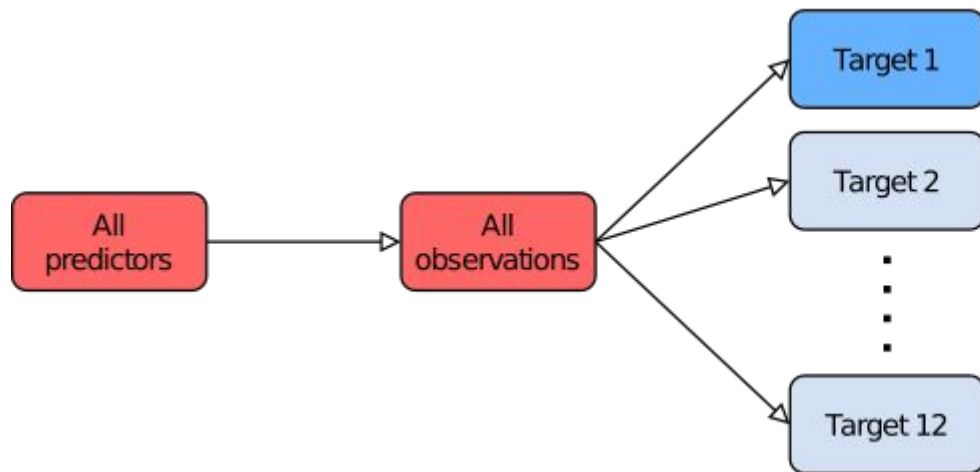
# metaMIMIC - first benchmark

MIMIC is single-center database comprising information relating to patients admitted to intensive care units (ICU)



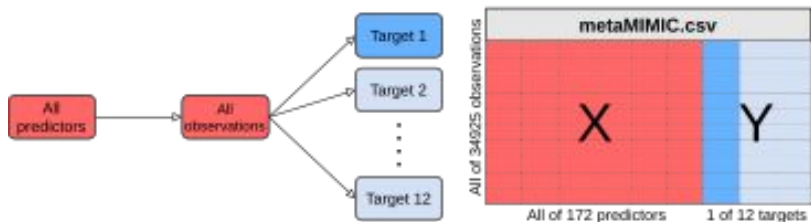
Targets	Frequency
Hypertensive disease	59.8%
Disorders of lipid metabolism	40.3%
Anemia	35.9%
Ischemic heart disease	32.8%
Diabetes	25.3%
...	

# metaMIMIC - scenarios

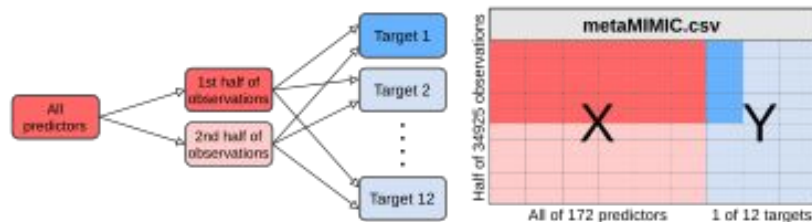


# metaMIMIC - first benchmark

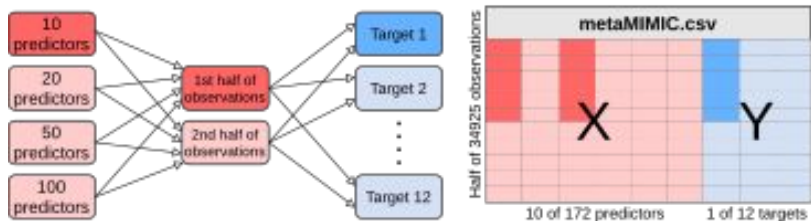
S1



S2



S3

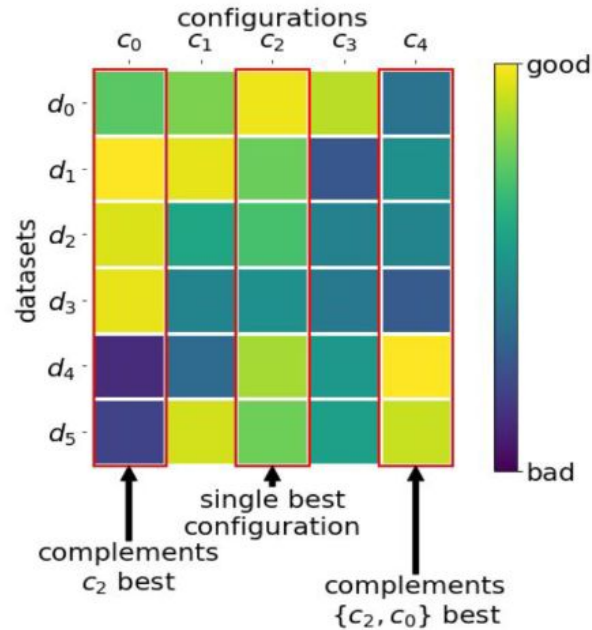


S4



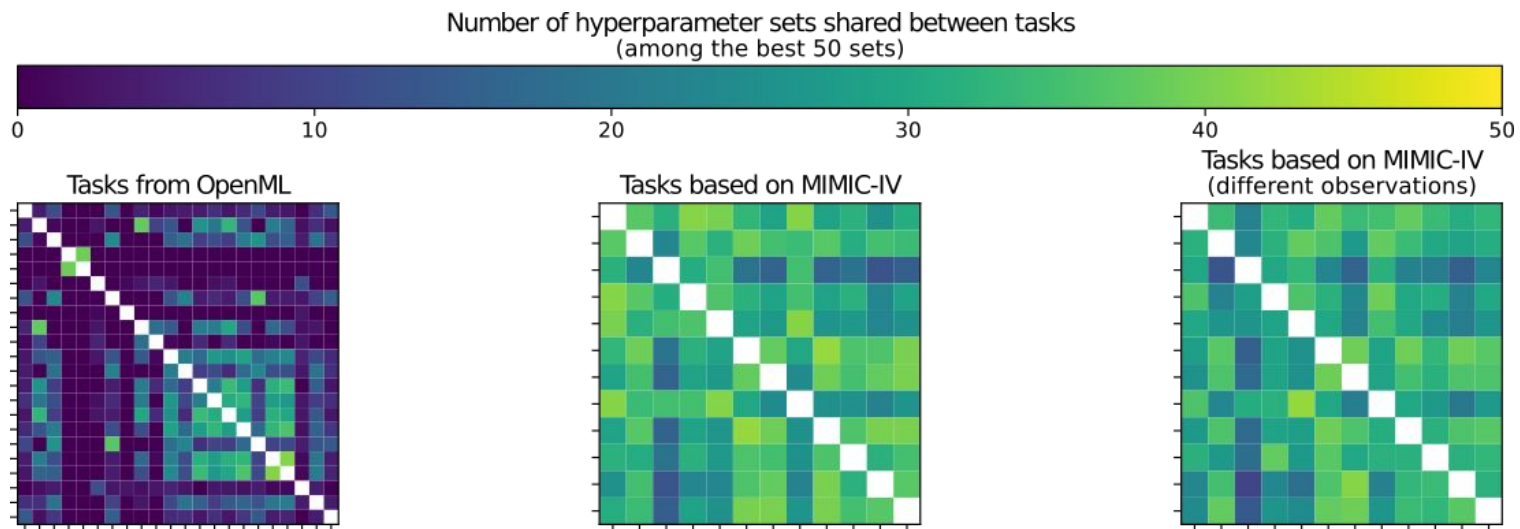
# metaMIMIC - experiments

- ⇒ XGBoost algorithm
- ⇒ 1000 random HP configurations
- ⇒ model-free portfolio of HPO: Greedy selection
- ⇒ validation:
  - ❑ one-dataset-out
  - ❑ validation measure: Average distance to maximum (ADTM)
  - ❑ comparison with baselines: RS, BO and model-free meta-learning from OpenML



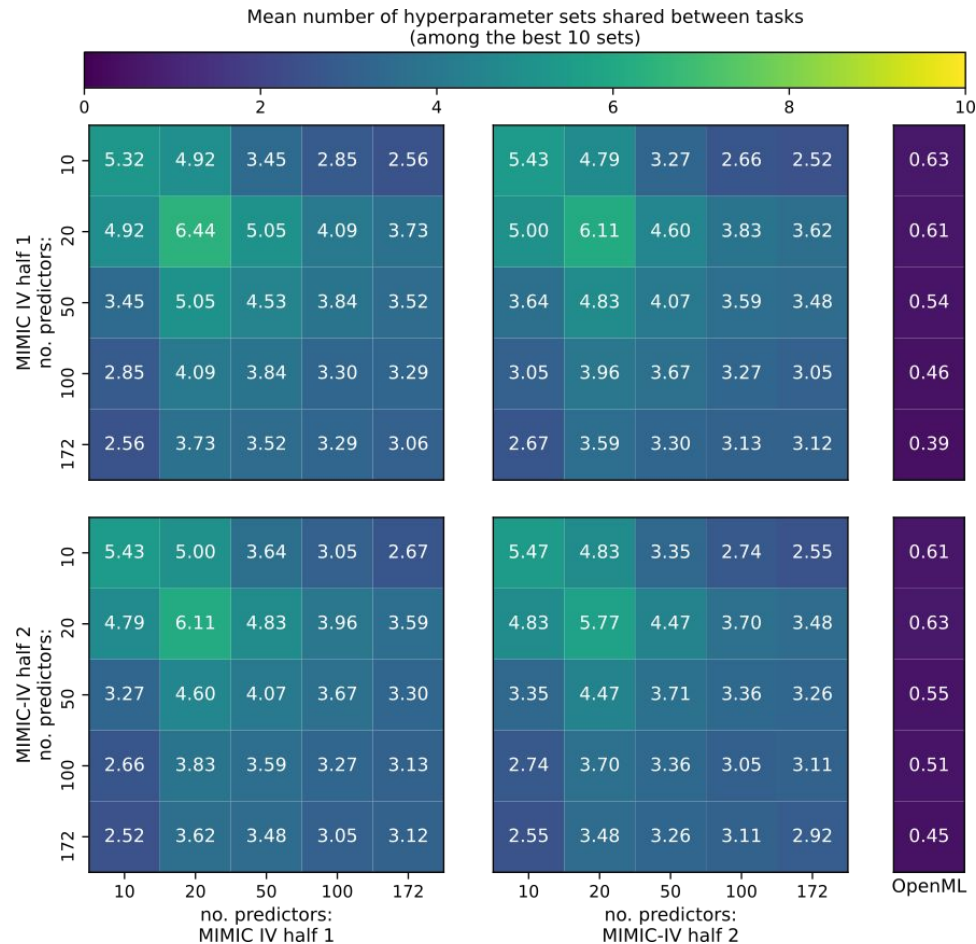
**Greedy selection**

# Results

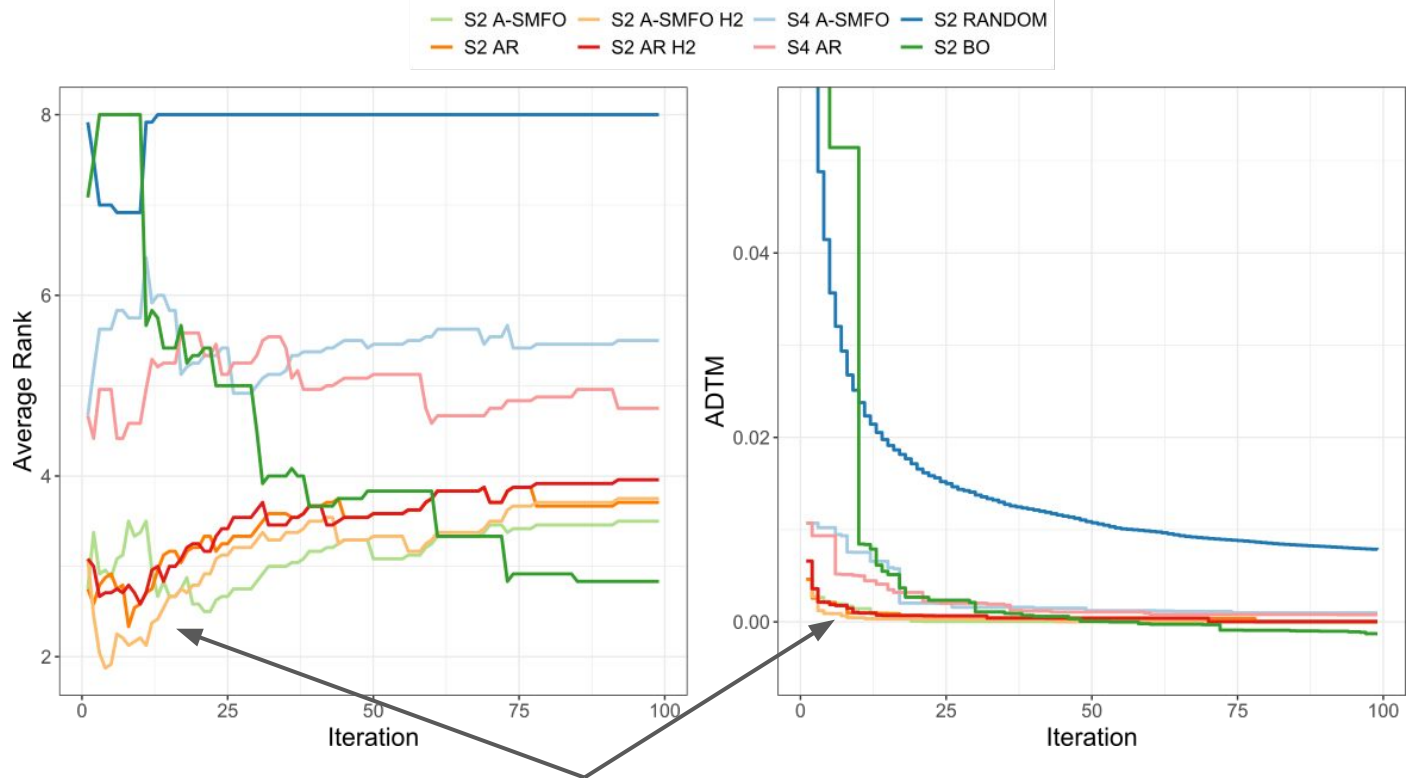


# Results

The definition-based (domain) similarity of tasks is positively related to hyperparameters' transferability between them.







# Global results



**anytime performance**



## Consolidated learning:

-  does not require any definitions of similarity based on statistical characteristics
-  easy to implement in any domain, benefits from specification of problems
-  provides anytime performance
-  increases trust in meta-learning

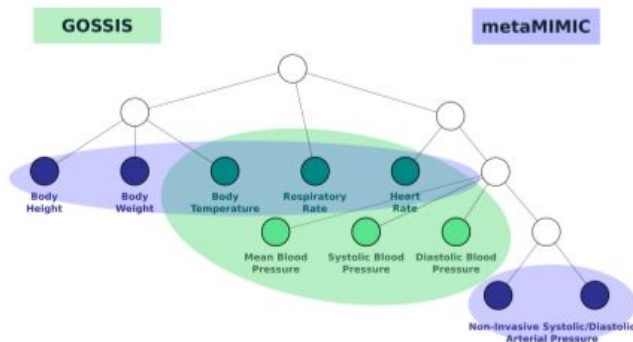
# Extension of CL with semantic similarity

## 1. Unrelated datasets

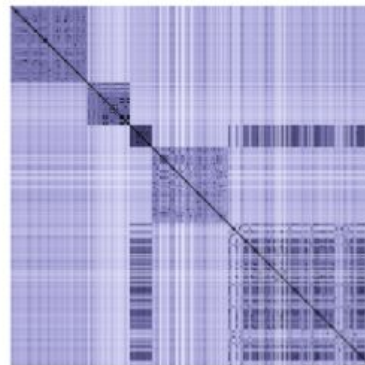
GOSSIS

metaMIMIC

## 2. SeFNet - Semantic Feature Net



## 3. Terms similarity Dataset similarity



# Semantic Feature Net - Healthcare

Welcome to Semantic Feature Net - Healthcare, a comprehensive collection of tabular datasets related to the applications of machine learning in predictive tasks for healthcare. All datasets have been structured considering the semantic meaning of their features with concepts derived from the SNOMED-CT ontology. Each dataset has an associated set of features' annotations, which can be used for sharing valuable insights between diverse predictive tasks.

dataset	category	instances	features	annotations	annotations %
Cardiovascular Study	Survey	4238	16	15	94%
Diagnosis of COVID-19 (Subset)	EHR	603	19	18	95%
Diabetes Health Indicators	Survey	253680	22	21	95%
Diabetes 130 US	EHR	101766	49	38	78%
GOSSIS-1-eICU Model Ready	EHR	131051	68	60	88%
Stroke Prediction	Survey	5110	11	11	100%
Heart Disease Indicators	Survey	253680	22	21	95%



# Questions?



Paper

# Consolidated Learning

**Katarzyna Woźnica** , M. Grzyb, Z. Trafas, P. Biecek

MI2.AI, Warsaw University of Technology



[woznicakatarzyna22@gmail.com](mailto:woznicakatarzyna22@gmail.com)



Neuro-symbolic Metalearning and AutoML | ECML 2023