

On the Democratization of AI:

Bridging the Gap with AutoML

Candidate:

Joseph Giovanelli

Supervisor:

Matteo Golfarelli

Coordinator:

Ilaria Bartolini

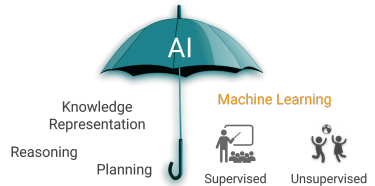


**ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA**

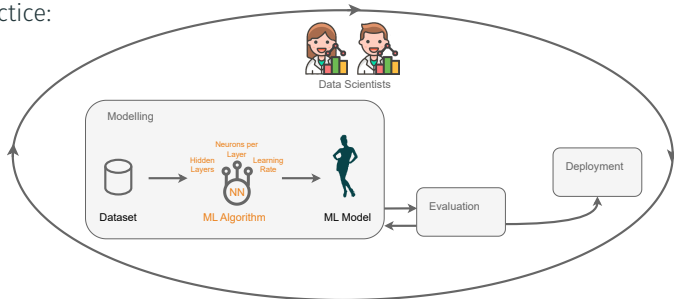
What Do We Mean By AI?

Machine Learning¹

A **computer program** is said to learn in some class of **tasks**, with respect to a **performance measure**, if it improves with **experience**.



In practice:

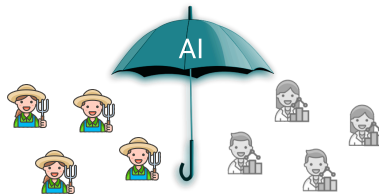


¹Mitchell, T. M. 1997. Machine Learning.

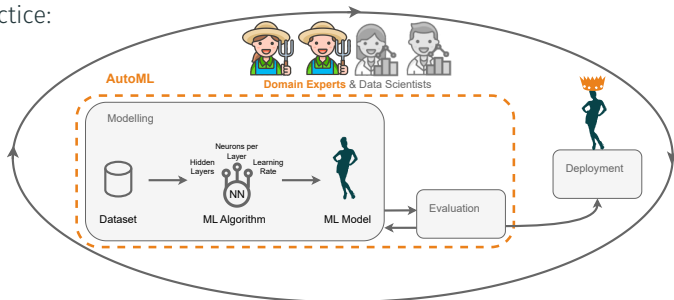
Bridging the Gap with AutoML

Democratization of AI²

Making AI **accessible to a broader audience**, allowing domain experts to apply it in their own fields.



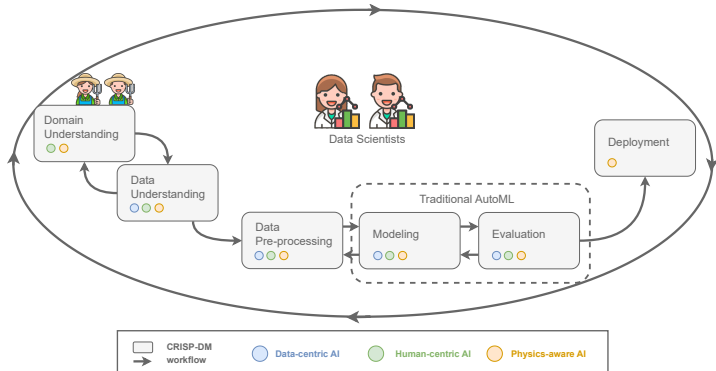
In practice:



²Thornton, C. et al. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of ACM SIGKDD.

Real-case problem scenarios

Cross-Industry Standard Process for Data Mining (CRISP-DM)
is a process model for dealing with problem complexity.

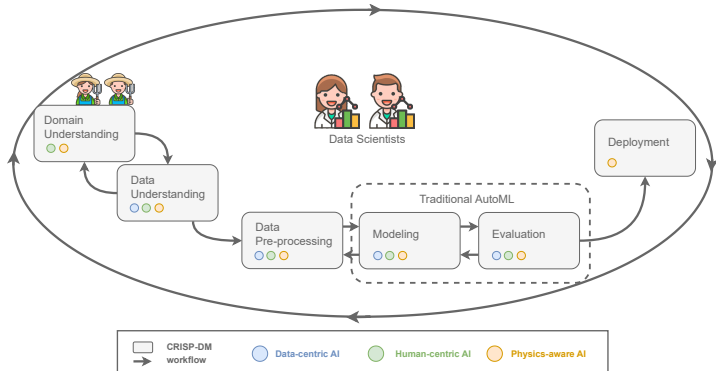


Data-centric AI

systematically engineers data used to build an AI system.

Real-case problem scenarios

Cross-Industry Standard Process for Data Mining (CRISP-DM)
is a process model for dealing with problem complexity.

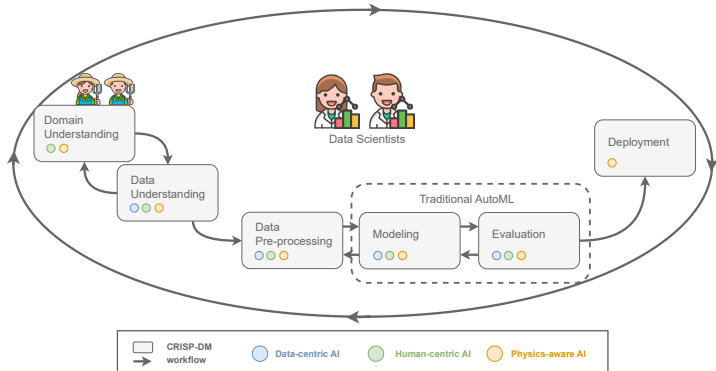


Human-centric AI

aims at complementing, instead of replacing, human intelligence.

Real-case problem scenarios

Cross-Industry Standard Process for Data Mining (CRISP-DM)
is a process model for dealing with problem complexity.



Physics-aware AI

focuses on coupling and enhancing physical simulators with AI.

Contribution Overview

Data-centric AI

1. Effective Data

Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering Pipelines via AutoML and Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_{\lambda}(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the Large Language Models

Physics-aware AI

6. Multi-sensor Profiling for Soil-Moisture Monitoring



7. Enhancing Process-Based Models for Soil Moisture Forecasting



Contribution Overview

Data-centric AI

1. Effective Data

**Pre-processing Pipelines
in Supervised Learning**^{3,4}

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering Pipelines via AutoML and Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_{\lambda}(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the Large Language Models

Physics-aware AI

6. Multi-sensor Profiling for Soil-Moisture Monitoring



7. Enhancing Process-Based Models for Soil Moisture Forecasting



³Giovanelli J., Bilalli B., and Abelló A. (2022). Data pre-processing pipeline generation for AutoETL. *Information Systems* 108 (2022): 101957.

⁴Giovanelli J., Bilalli B., and Abelló A., et al. (2023). Reproducible experiments for generating pre-processing pipelines for AutoETL. *Information Systems* (2023): 102314.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification⁵

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁵Francia M., [Giovanelli J.](#), and Golfarelli M. (2024). [AutoClues: Exploring Clustering Pipelines via AutoML and Diversification](#). In [Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining \(PAKDD\)](#). Springer Nature Singapore.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat. ⁶

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁶Francia M., Giovanelli J., and Pisano G. (2022). HAMLET: A framework for Human-centered AutoML via Structured Argumentation. *Future Generation Computer Systems* 142 (2023): 182-194.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning ⁷

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁷ Giovanelli J., Tornede A., Tornede T., and Lindauer M. (2024). Interactive hyperparameter optimization in multi-objective problems via preference learning. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 11. 2024.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models ⁸

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁸Tornede A., Difan D., Giovanelli J., Mohan A., Ruhkopf T., Segel S., Theodorakopoulos D., Tornede T., Wachsmuth H., and Lindauer M. (2024). Automl in the age of large language models: Current challenges, future opportunities and risks. *Transaction on Machine Learning Research*. ISSN 2835-8856 2024.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring ⁹



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁹Francia M., Giovanelli J., and Golfarelli M. (2022). Multi-sensor profiling for precision soil-moisture monitoring. *Computers and Electronics in Agriculture*. 197 (2022): 106924.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting ¹⁰



¹⁰Bitelli M., Francia M., Giovanelli J., Golfarelli M., and Tomei F. An Auto-Tuning Three-Dimensional Numerical Model Coupled with Data Assimilation from a Sensor Grid to Forecast Irrigation Demand in Kiwifruit. Submitted to **Computers and Electronics in Agriculture**.

Contribution Overview

Data-centric AI

1. Effective Data
Pre-processing Pipelines
in Supervised Learning

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X} \times \mathcal{Y}$$

2. Exploring Clustering
Pipelines via AutoML and
Diversification

$$\mathcal{D} = \{(x_i)\}_{i=0}^N \in \mathbb{D} \subset \mathcal{X}$$

Human-centric AI

3. Human-centric AutoML
via Logic and Argumentat.

$$\lambda^* \in \operatorname{argmin}_{\lambda \in \Lambda} \mathcal{L}(A_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{out}})$$

4. Interactive HPO
via Preference Learning ⁷

$$\mathcal{L}_1, \dots, \mathcal{L}_m$$

5. AutoML in the Age of the
Large Language Models

Physics-aware AI

6. Multi-sensor Profiling
for Soil-Moisture Monitoring



7. Enhancing Process-Based
Models for Soil Moisture
Forecasting



⁷ Giovanelli J., Tornede A., Tornede T., and Lindauer M. (2024). Interactive hyperparameter optimization in multi-objective problems via preference learning. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 11. 2024.

Interactive
Hyperparameter Optimization
via Preference Learning
in Multi-Objective Problems

Hyperparameter Optimization (HPO)

HPO Problem. Given a machine learning (ML) algorithm A and corresponding hyperparameter space of $\Lambda = \Lambda_1 \times \dots \times \Lambda_M$, the goal is to determine the configuration $\lambda^* \in \Lambda$ with optimal loss function \mathcal{L} .

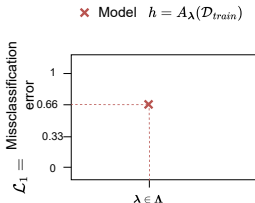
$$\lambda^* \in \arg \min_{\lambda \in \Lambda} \mathcal{L}(A_{\lambda}(\mathcal{D}_{train}), \mathcal{D}_{val})$$

\mathcal{L} quantifies how well the trained model $h = A_{\lambda}(\mathcal{D}_{train})$ performs a disjoint split \mathcal{D}_{val} .

Example NN hyperparameter conf.

$\lambda \in \Lambda =$

<i>learning rate:</i>	0.05
<i>hidden layers:</i>	5
<i>neurons per layer:</i>	256



SOTA. Bayesian Optimization (BO)¹¹ drives the exploration toward new promising configurations via a surrogate trained on past evaluations.

¹¹Brochu E., Vlad M. Cora, et al. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. (2010).

Hyperparameter Optimization (HPO)

HPO Problem. Given a machine learning (ML) algorithm A and corresponding hyperparameter space of $\Lambda = \Lambda_1 \times \dots \times \Lambda_M$, the goal is to determine the configuration $\lambda^* \in \Lambda$ with optimal loss function \mathcal{L} .

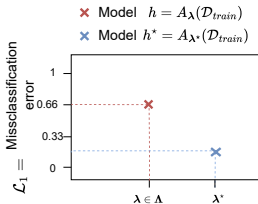
$$\lambda^* \in \arg \min_{\lambda \in \Lambda} \mathcal{L}(A_{\lambda}(\mathcal{D}_{train}), \mathcal{D}_{val})$$

\mathcal{L} quantifies how well the trained model $h = A_{\lambda}(\mathcal{D}_{train})$ performs a disjoint split \mathcal{D}_{val} .

Example Best NN hyperparameter conf.

$\lambda^* \in \Lambda =$

<i>learning rate:</i>	0.01
<i>hidden layers:</i>	10
<i>neurons per layer:</i>	256

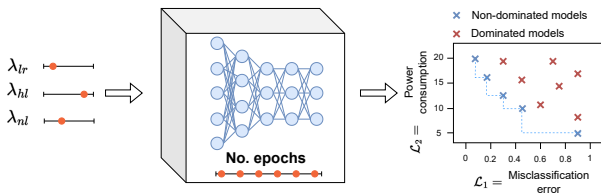


SOTA. Bayesian Optimization (BO)¹¹ drives the exploration toward new promising configurations via a surrogate trained on past evaluations.

¹¹Brochu E., Vlad M. Cora, et al. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. (2010).

Multi-Objective Machine Learning (MO-ML)

MO-ML algorithms. When optimizing multiple objectives $\mathcal{L}_1, \dots, \mathcal{L}_M$, MO-ML algorithms $A_{\lambda}(\mathcal{D}_{train})$ return a **Pareto front** $P_{\mathcal{D}_{val}}(\mathcal{H})$.



Quality Indicators. quantify the goodness of the Pareto front by measuring specific characteristics —e.g., hypervolume (HV)¹³, maximum spread (MS)¹³, spacing (SP)¹⁴, closeness to reference point (R2)¹⁵.

¹³Zitler E., Thiele L. **Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach.** IEEE transactions on Evolutionary Computation, 1999.

¹⁴Schott J. R. **Fault tolerant design using single and multicriteria genetic algorithm optimization.** 1995. PhD Thesis. Massachusetts Institute of Technology.

¹⁵Hansen M. P., Jaszewicz A. **Evaluating the quality of approximations to the non-dominated set.** Copenhagen, Denmark: IMM, Technical University of Denmark, 1994.

Interactive HPO via Preference Learning

Challenge. Choosing the **quality indicator** leading to a Pareto front which has a **desired shape** requires deep expert knowledge.

Approach. Learning a quality indicator via user preferences

1. Preliminary Sampling

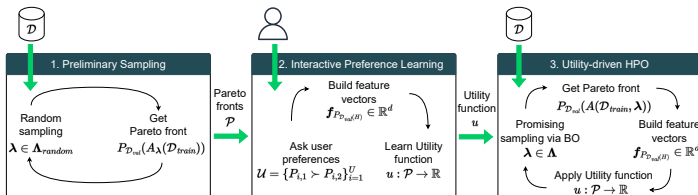
Collect Paretos $\mathcal{P} = \{P_i | P_i = P_{\mathcal{D}_{val}}(A_{\lambda}(\mathcal{D}_{train})) : \lambda \in \Lambda\}$

2. Interactive Preference Learning

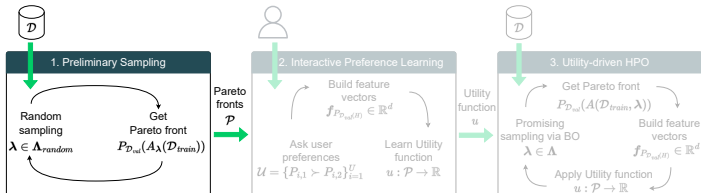
- Build preference dataset $\mathcal{U} = \{P_{i,1} \succ P_{i,2}\}_{i=1}^U$
- Learn utility function $u : \mathcal{P} \rightarrow \mathbb{R}$

3. Utility-driven HPO

Solve HPO problem $\lambda^* \in \arg \min u(P_{\lambda}) : P_{\lambda} = A_{\lambda}(\mathcal{D}_{val})$



1. Preliminary Sampling

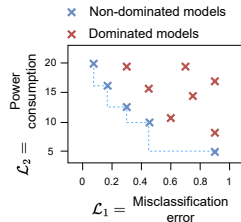


Example $\lambda_1 \in \Lambda$

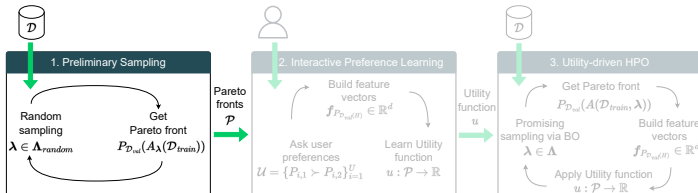
Example $P_{\mathcal{D}_{\text{val}}}(A_{\lambda}(\mathcal{D}_{\text{train}}))$

NN hyperparameter configuration:

learning rate: 0.01
hidden layers: 10
neurons per layer: 256



1. Preliminary Sampling

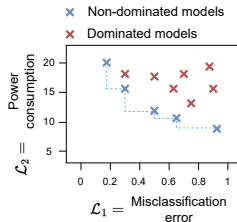


Example $\lambda_2 \in \Lambda$

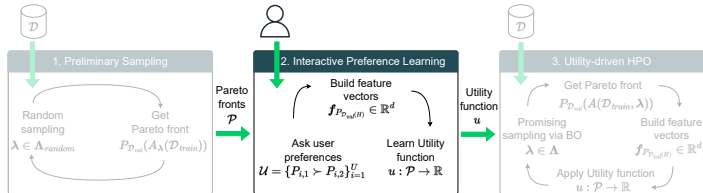
Example $P_{\mathcal{D}_{\text{val}}}(A_{\lambda}(\mathcal{D}_{\text{train}}))$

NN hyperparameter configuration:

learning rate: 0.05
hidden layers: 5
neurons per layer: 256



2. Interactive Preference Learning – Pairwise Preferences

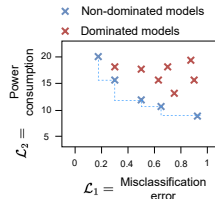
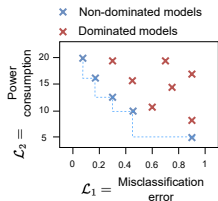


Example $U = \{P_{i,1} \succ P_{i,2}\}_{i=1}^U$

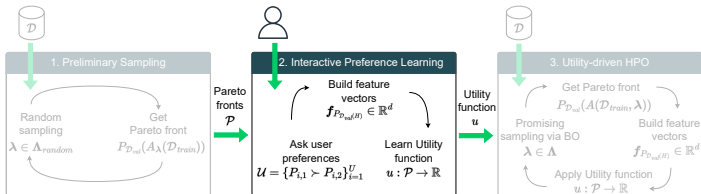
$P_{i,1}$

\succ

$P_{i,2}$

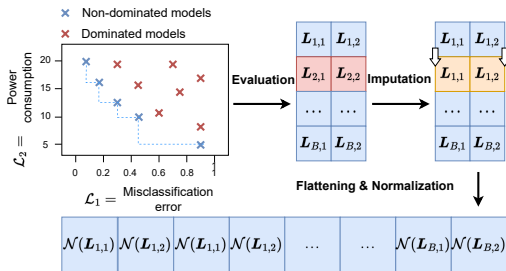


2. Interactive Preference Learning – Feature Building

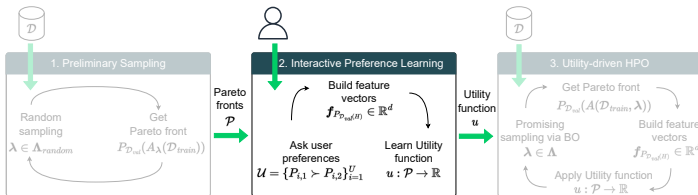


Example

$$f_{P_{\mathcal{D}_{\text{val}}(H)}} \in \mathbb{R}^d$$



2. Interactive Preference Learning – Learn Utility

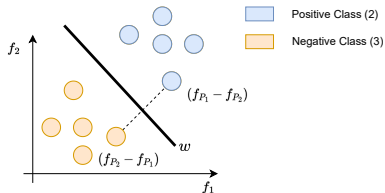


Example Building $u : \mathcal{P} \rightarrow \mathbb{R}$ through RankSVM¹⁶

$$P_1 \succ P_2 \Leftrightarrow \vec{w}^T \vec{f}_{P_1} > \vec{w}^T \vec{f}_{P_2} \quad (1)$$

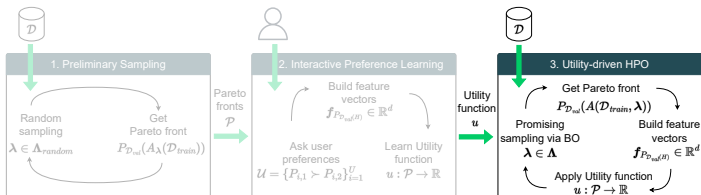
$$\Leftrightarrow \vec{w}^T (\vec{f}_{P_1} - \vec{f}_{P_2}) > 0 \quad (2)$$

$$\Leftrightarrow \vec{w}^T (\vec{f}_{P_2} - \vec{f}_{P_1}) < 0. \quad (3)$$



¹⁶Joachims T. **Optimizing search engines using clickthrough data**. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. (2002).

3. Utility-driven HPO



Example.

Promising sampling
 $\lambda \in \Lambda$



Example.

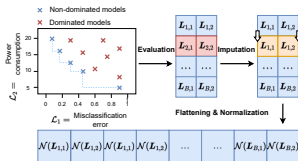
Get Pareto and build feature vector
 $P_{\mathcal{D}_{val}}(A(\mathcal{D}_{train}, \lambda)) \rightarrow f_{P_{\mathcal{D}_{val}}}$



Example.

Apply utility
 $u(f_{P_{\mathcal{D}_{val}}})$

learning rate: 0.001
hidden layers: 5
neurons layer: 256



$$W^T \cdot f_{P_{\mathcal{D}_{val}}(H, \lambda)}$$

Evaluation – End-to-end Performance

Preference-Based (PB): HPO process driven by the utility function trained with the indicator in the row

Indicator-Based (IB): HPO process driven by the indicator in the column

LCBench¹⁷:

- funnel-shaped MLP from **Auto-pytorch**;
- 35 datasets from **OpenML CC-18** suite.

PB\IB	HV	SP	MS	R2
HV	98.70%	146.15%	146.15%	98.70%
SP	300.00%	100.00%	400.00%	400.00%
MS	321.05%	321.05%	93.85%	265.22%
R2	95.65%	204.35%	195.65%	95.65%

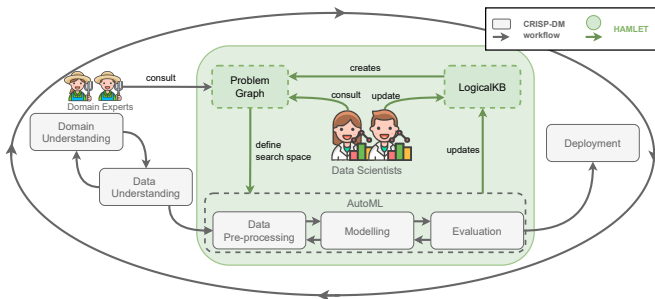
→ **PB** performs **better or equal** in 11/16 cases;

→ **IB** performs **slightly better** in only 5/16 cases.

¹⁷Zimmer L, Lindauer M, and Hutter F. **Auto-pytorch: Multi-fidelity metalearning for efficient and robust autodl**. IEEE transactions on pattern analysis and machine intelligence 43.9 (2021).

Conclusions and Future Works

Data-centric and Human-centric AI

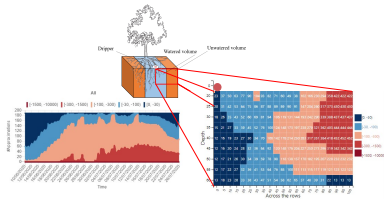


Main Contributions

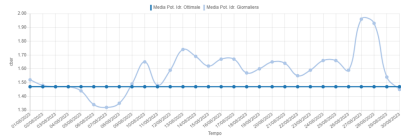
- Develop **effective data pipelines** for supervised and unsupervised learning.
- Propose a **HAMLET**: a **human-centric AutoML framework** allowing **explainability** and **interactive interventions**;

Future Works

- Provide insight on **data bias in pre-processing** for **fairness** through HAMLET.
- Integrate HAMLET with **multi-objective** and **cross-cutting constraints** (e.g., ethical, legal) to allow **fairness interventions**.



Potenziale Idrico Ottimale e Potenziale Idrico Medio Giornaliero



Main Contributions

- Integration of **physical models** with **AI** through **AutoML** for **monitoring** and **forecasting** tasks.
- During the whole campaign:
Water saving: 44%
Vine productivity: **unaffected**
Fruit quality: **increased** (+1 brix)

Future works:

- Integration of a **smart-irrigation** algorithm based on **control theory**, dynamically adjusting the water plan.
- Application of **transfer learning** with a pre-trained model to transfer the knowledge from different conditions, supporting a **wider range of crops**.

Thanks for the attention :)