Dekonstrukcja AutoML, czyli co, jak i dlaczego?

Anna Kozak, Katarzyna Woźnica

Katarzyna Woźnica

PhD Candidate | AutoML | Meta-learning

7+ years of experience in data science
4+ years of experience in mentoring students

Skills: AutoML, meta-learning, R, Python,
data visualization

katarzyna.woznica.dokt@pw.edu.pl

MI RESEARCH

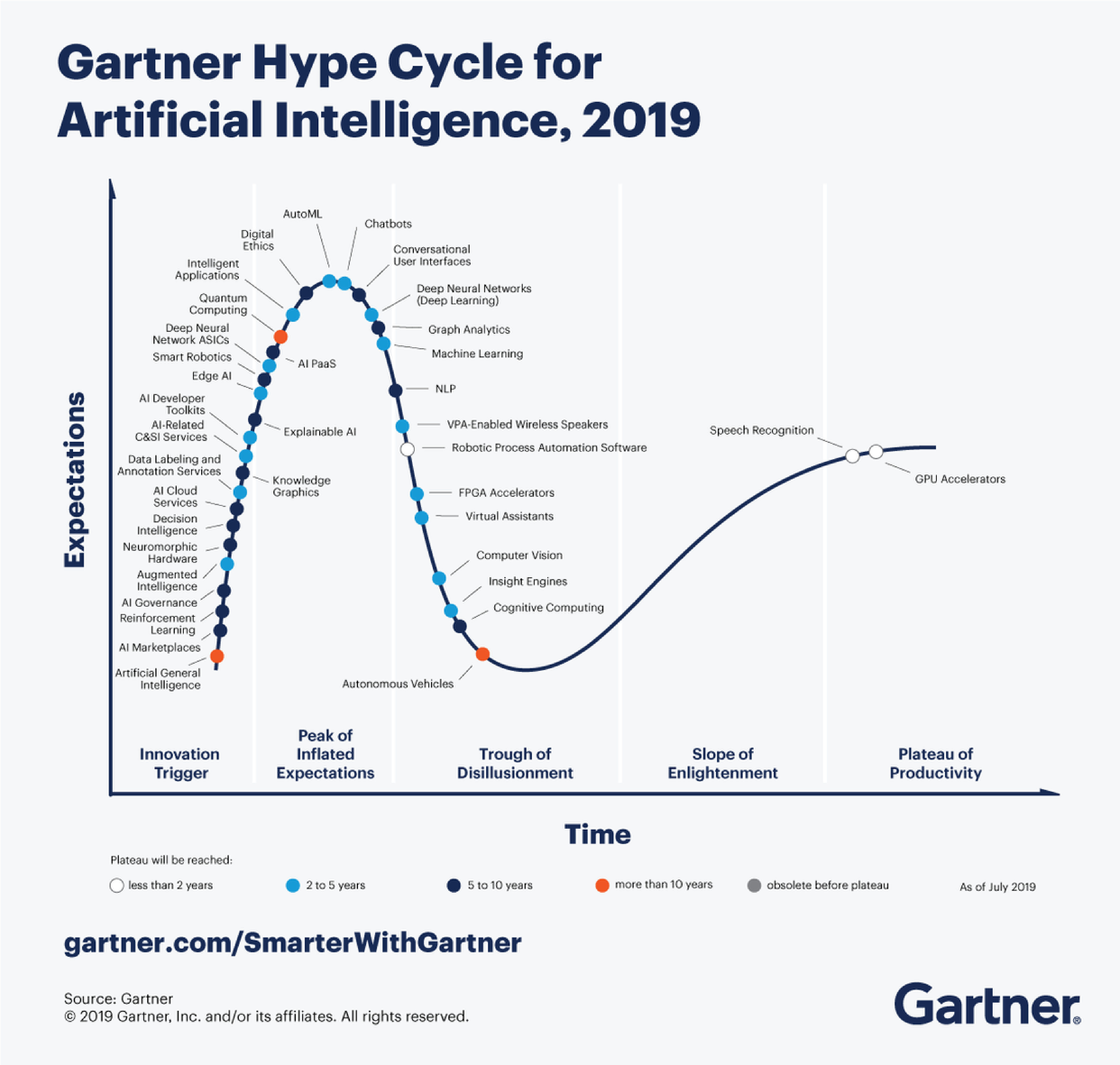
Anna Kozak

Data Scientist | Data Visualisation | Responsible Machine Learning

8+ years of experience in data science
4+ years of experience teaching

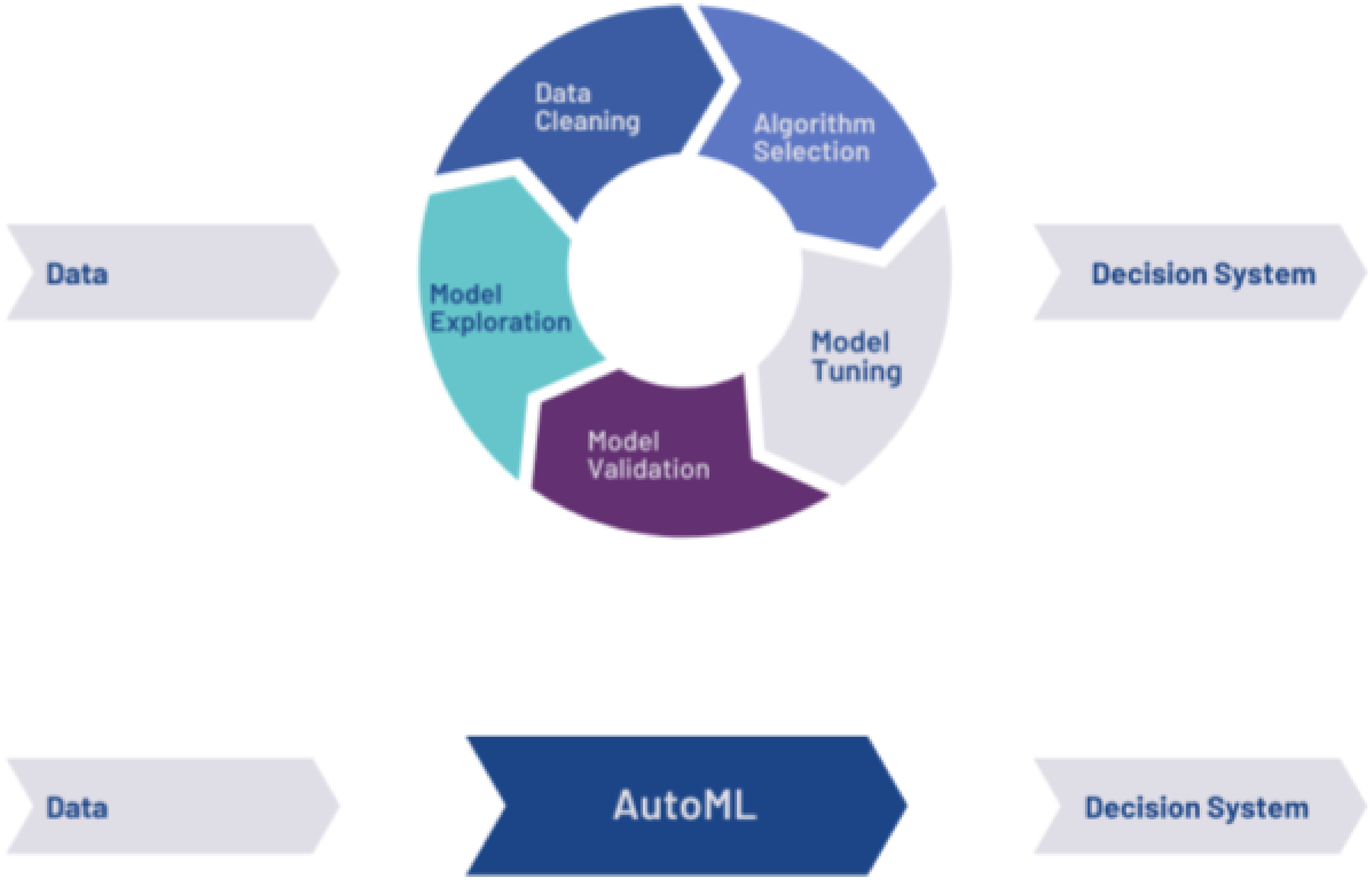
Skills: data visualization, data analysis, R, Python,
machine learning, AutoML, AutoEDA

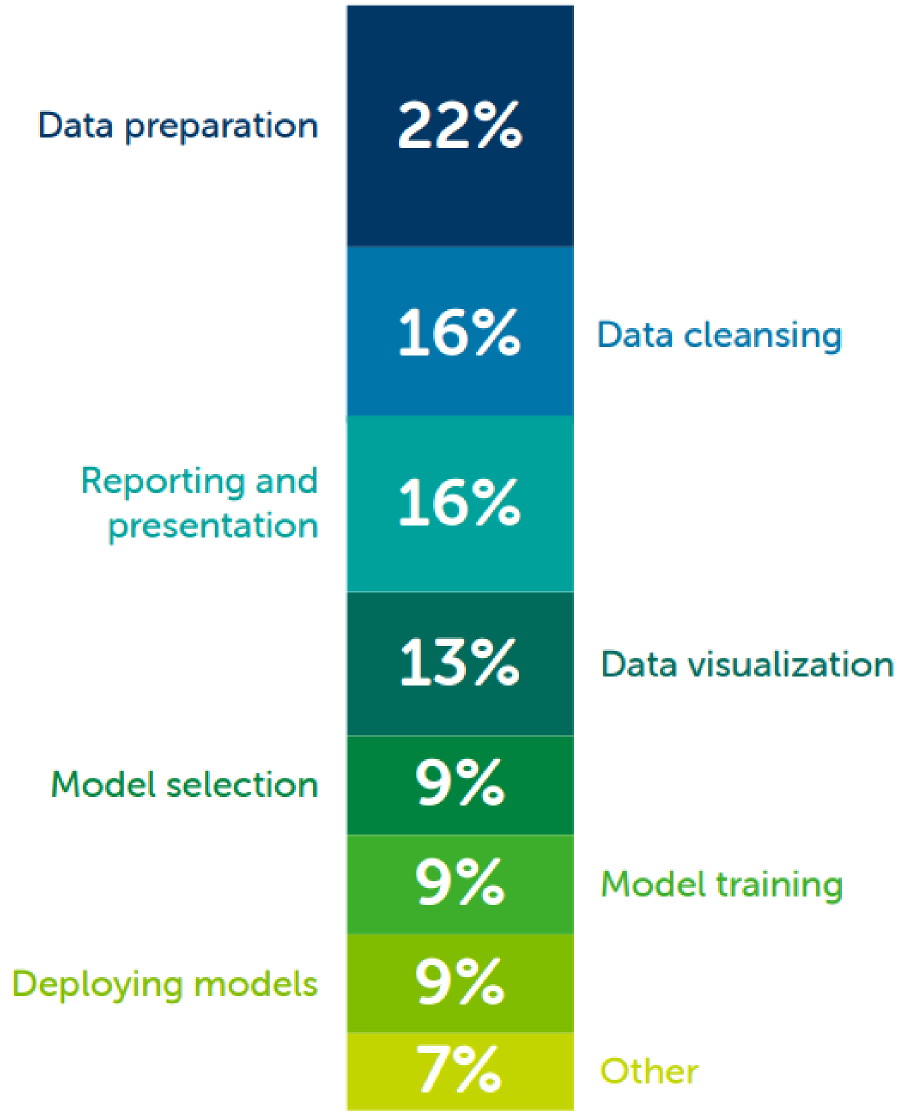
anna.kozak@pw.edu.pl



https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019

Czym jest AutoML?


Data
Data Cleaning
Algorithm Selection
Model Tuning
Model Validation
Model Exploration
Decision System
Data
AutoML
Decision System



n = 1,966

We asked our respondents how much time they spend on the above tasks, and for each item they entered a number reflecting the percentage of time spent relative to the other options. This is the average of the reported percentages.

Anaconda, State of Data Science 2022
(3,493 individuals from 133 countries)

1. Enable non-experts to train machine learning models (2.57)
2. Quickly and efficiently tune very many hyperparameters (2.75)
3. Help choose the best model types to solve specific problems (2.78)
4. Speed up the ML pipeline by automating certain workflows (data cleaning, etc.) (3.06)
5. Tune the model once performance (such as accuracy, etc.) starts to degrade (3.99)
6. Other (5.85)

We asked respondents to drag and rank the options from most to least important, with the first being most important.

n = 2,042

Anaconda, State of Data Science 2022
(3,493 individuals from 133 countries)

Who is AutoML end user?

Who is AutoML end user?

Traditionally, application's developers using statistical and learning methods choose algorithms and tune their parameters empirically, commonly by trial and error; or in the best case, by using prior knowledge of experts on the domain.

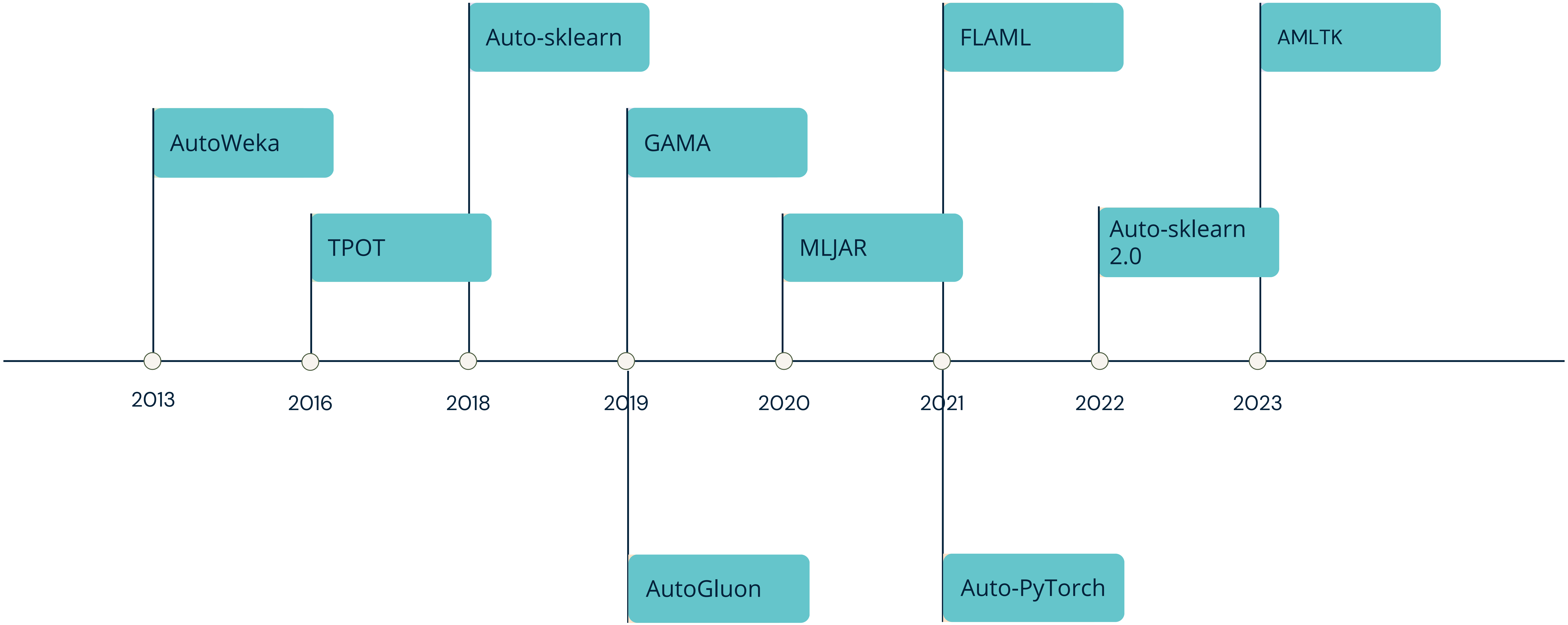
[PSMS for Neural Networks, 2007]

It can be challenging to make the right choice when faced with these degrees of freedom, leaving many users to select algorithms based on reputation or intuitive appeal, and/or to leave hyperparameters set to default values.

[AutoWEKA, 2013]

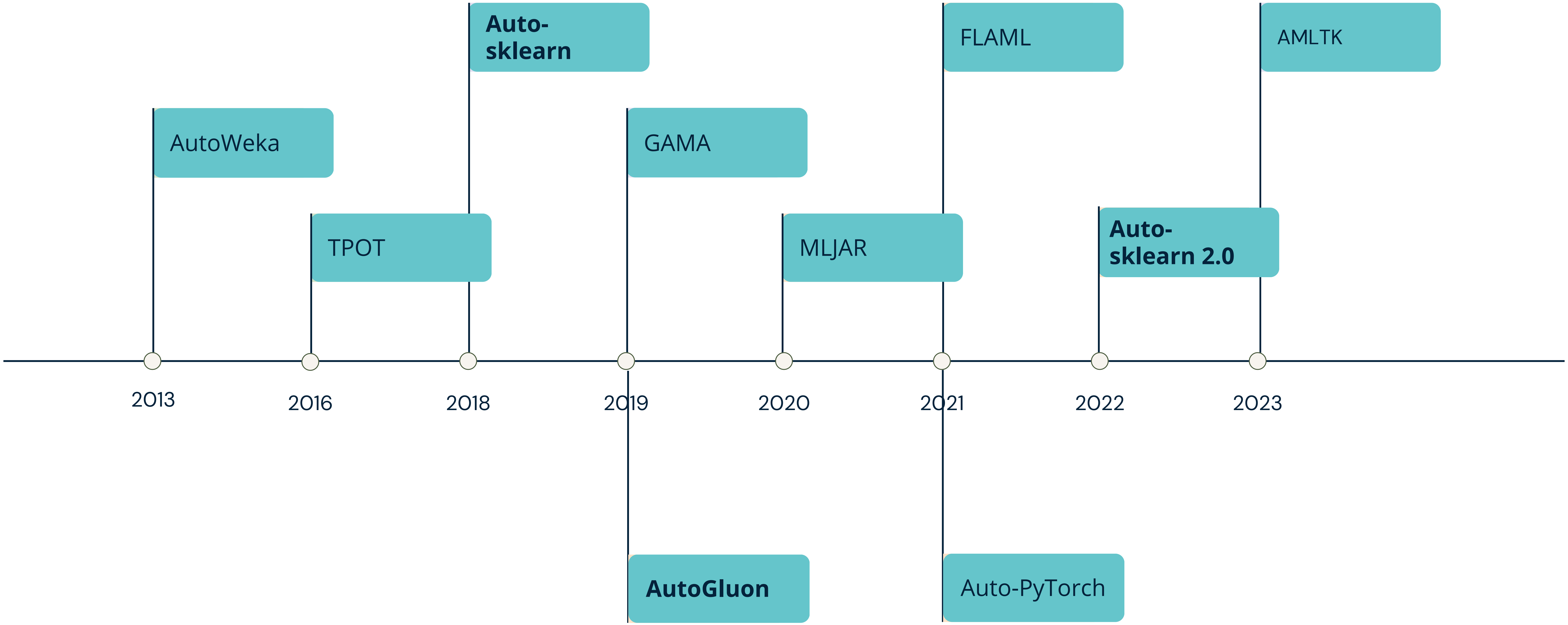
Automated Machine Learning (AutoML) supports practitioners and researchers with the tedious task of designing machine learning pipelines and has recently achieved substantial success.

[Auto-sklearn 2.0, 2022]


Auto-sklearn
FLAML
AMLTK
AutoWeka
GAMA
TPOT
MLJAR
Auto-sklearn 2.0
2013
2016
2018
2019
2020
2021
2022
2023
AutoGluon
Auto-PyTorch

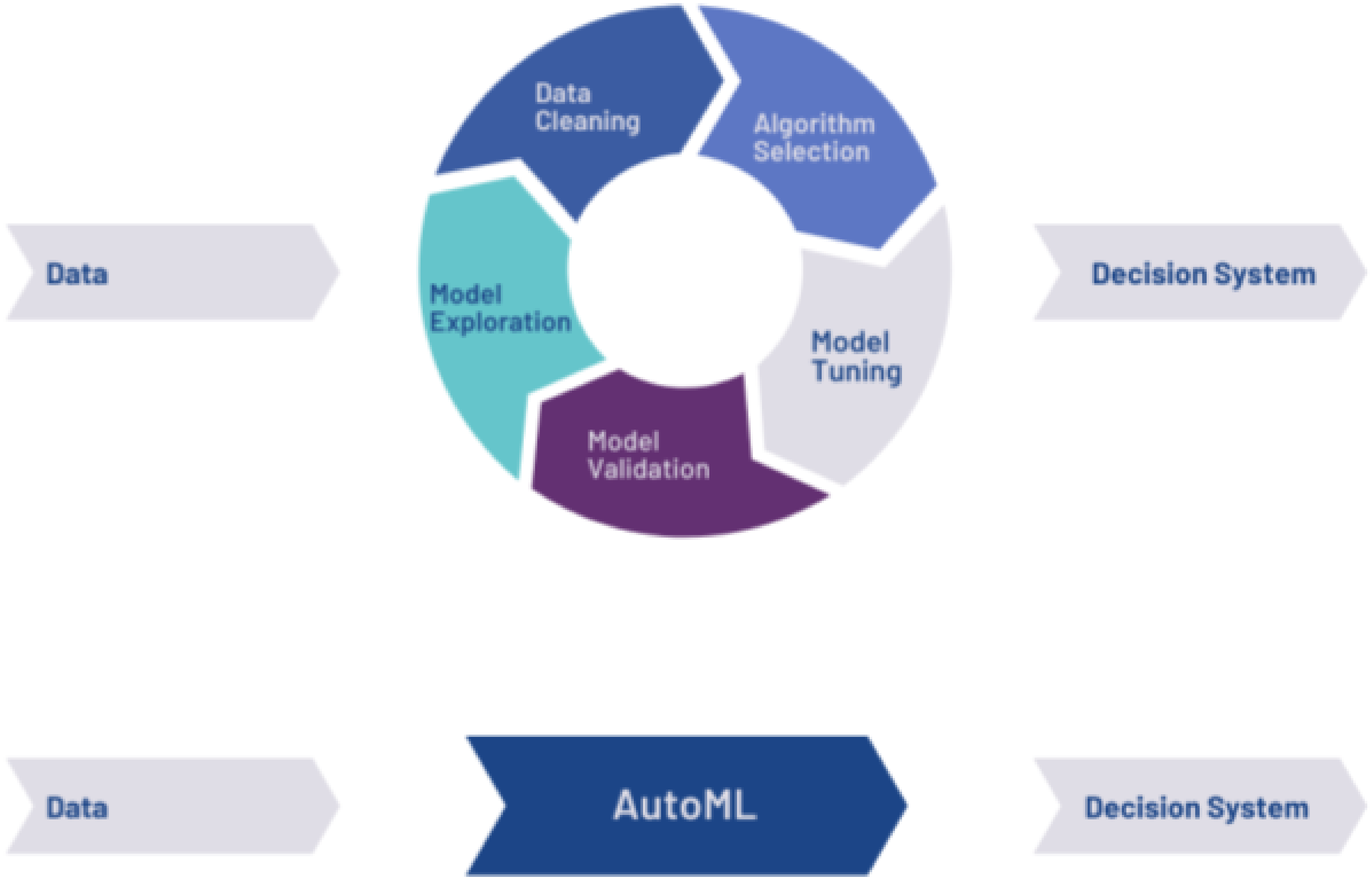
Nie ma najlepszego frameworku AutoML.

Zoptymalizowane Random Forest są zaskakująco skuteczne.


Auto-sklearn
FLAML
AMLTK
AutoWeka
GAMA
TPOT
MLJAR
Auto-sklearn 2.0
2013
2016
2018
2019
2020
2021
2022
2023
AutoGluon
Auto-PyTorch

Podejścia do danych tabelarycznych są najbardziej rozwinięte i na nich się skupimy.

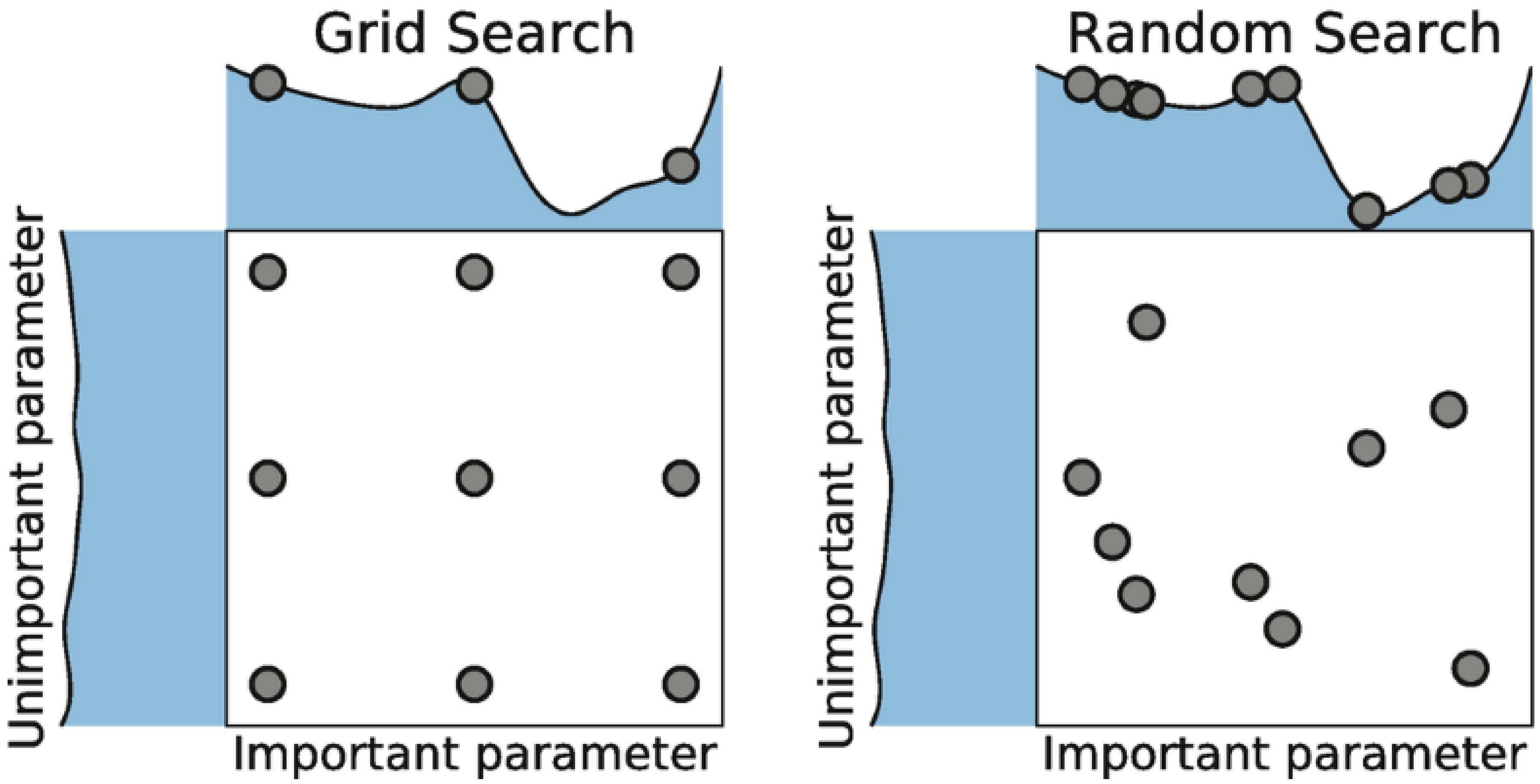
Są też podejścia do automatyzacji modelowania tekstu, szeregów czasowych.


Data
Data Cleaning
Algorithm Selection
Model Exploration
Model Tuning
Model Validation
Decision System
Data
AutoML
Decision System

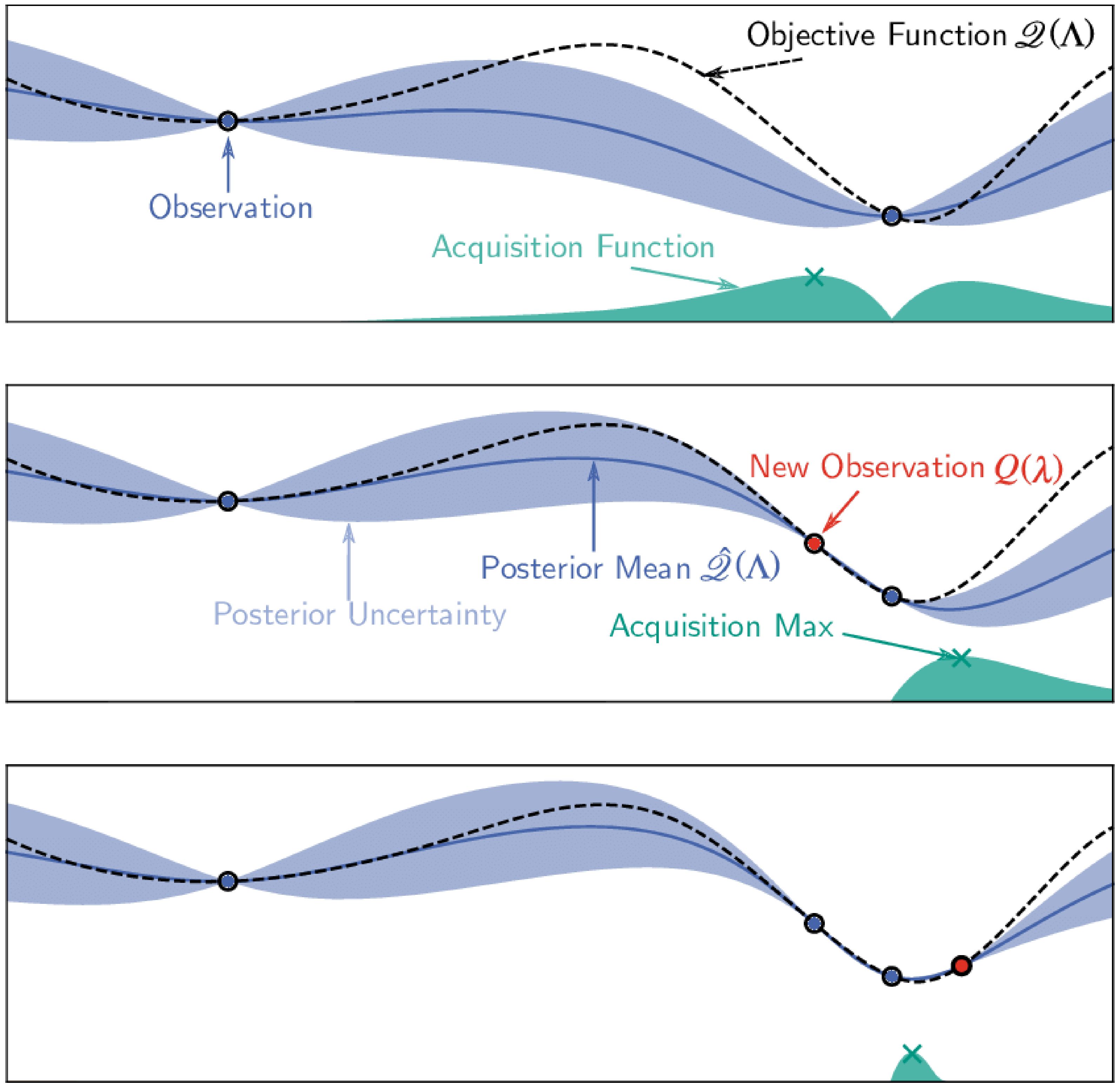
Optymalizacja hiperparametrów

Auto-sklearn

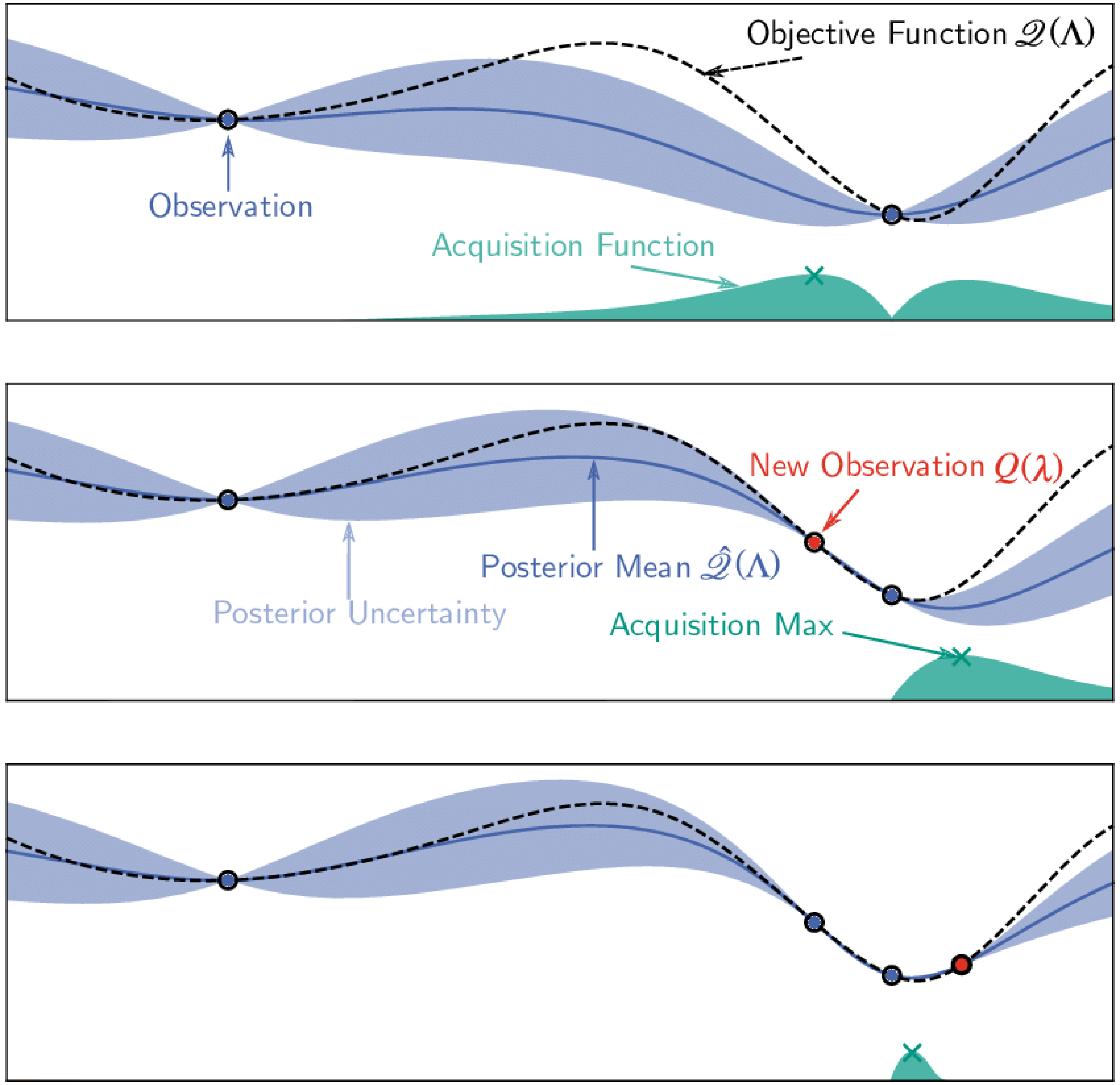
Grid/Random Search - task-agnostic



Bayesian Optimization - task-specific



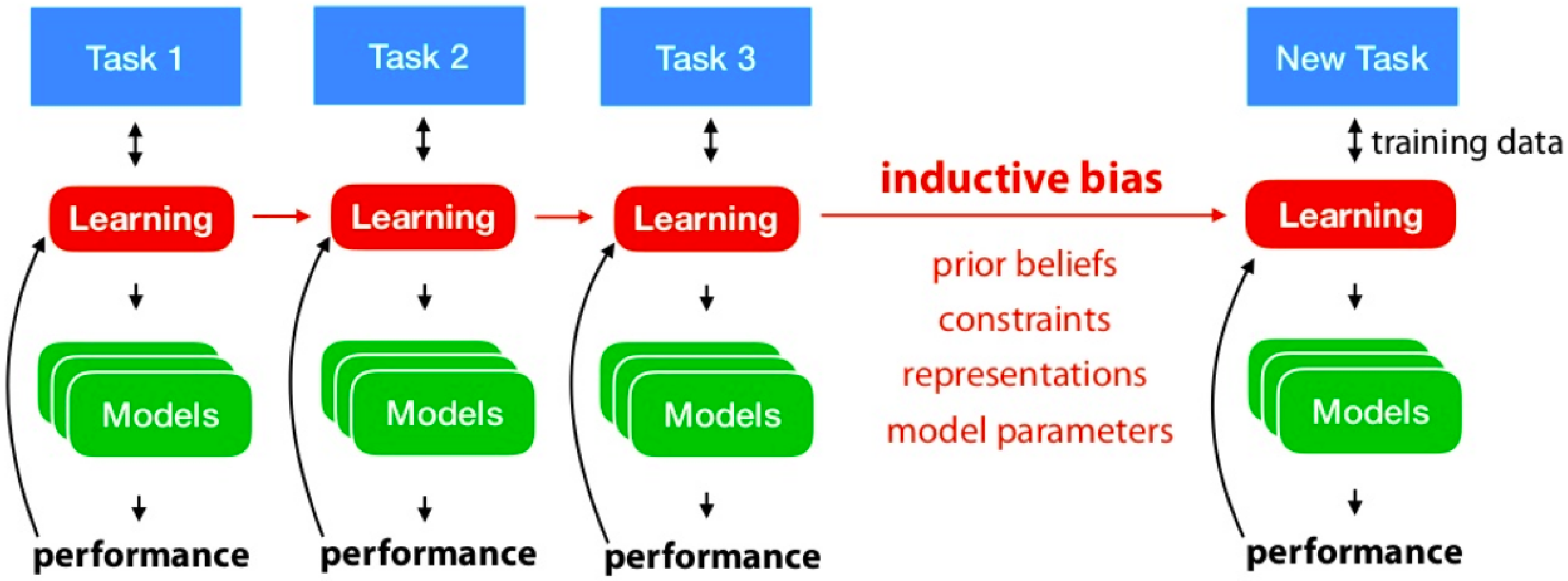
Bayesian Optimization - task-specific



Jak dodać informację o wcześniejszych eksperymentach?

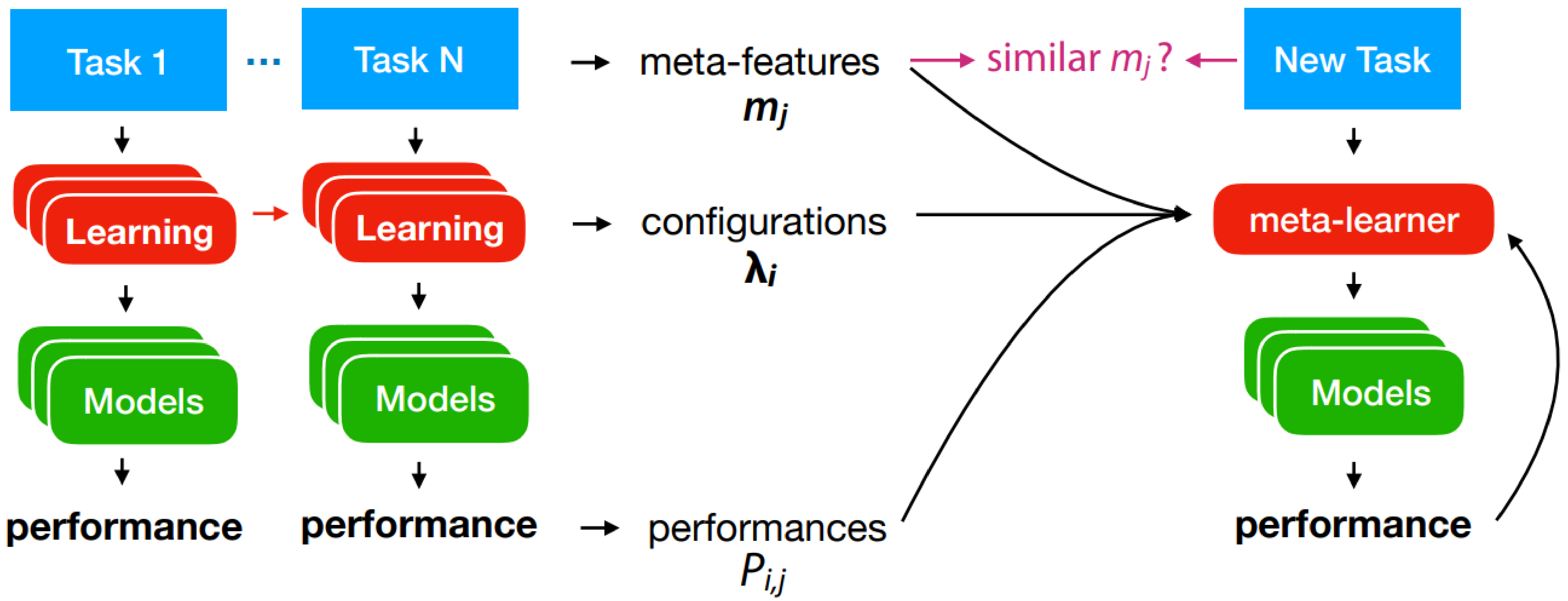
Meta-learning

Meta-learning



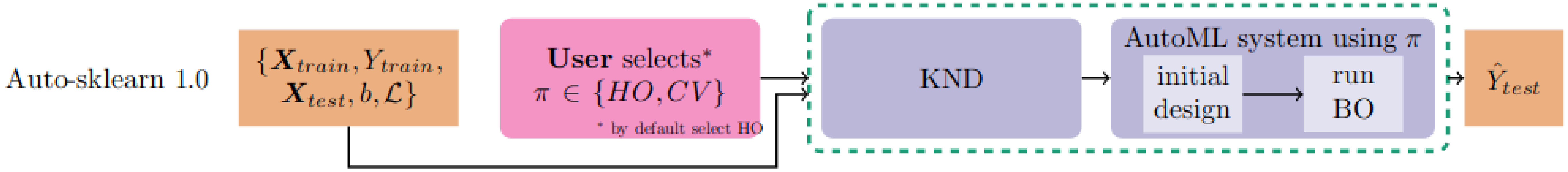
Joaquin Vanschoren, AAAI 2021 Meta learning tutorial

Learning from Task Properties



Joaquin Vanschoren, AAAI 2021 Meta learning tutorial

Meta-learning in auto-sklearn 1.0



Given a new dataset, we compute its meta-features, rank all datasets by their L1 distance to new dataset in meta-feature space and select the stored ML framework instantiations for the k = 25 nearest datasets for evaluation before starting Bayesian optimization with their results.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum and Frank Hutter,
Efficient and Robust Automated Machine Learning, Advances in Neural Information Processing Systems 28 (2015)

Meta-features

- hand-crafted, a priori defined statistical measures
- landmarkers
- model-based meta-features
- active testing - model performance correlation

Name	Formula	Rationale	Variants
Nr instances	n	Speed, Scalability [99]	p/n, $log(n)$, $log(n/p)$
Nr features	p	Curse of dimensionality [99]	$log(p)$, % categorical
Nr classes	c	Complexity, imbalance [99]	ratio min/maj class
Nr missing values	m	Imputation effects [70]	% missing
Nr outliers	o	Data noisiness [141]	o/n
Skewness	$\frac{E(X-\mu_X)^3}{\sigma_X^3}$	Feature normality [99]	min,max,μ,σ,q_1, q_3
Kurtosis	$\frac{E(X-\mu_X)^4}{\sigma_X^4}$	Feature normality [99]	min,max,μ,σ,q_1, q_3
Correlation	$\rho_{X_1X_2}$	Feature interdependence [99]	min,max,μ,σ,ρ_{XY}[158]
Covariance	$cov_{X_1X_2}$	Feature interdependence [99]	min,max,μ,σ,cov_{XY}
Concentration	$\tau_{X_1X_2}$	Feature interdependence [72]	min,max,μ,σ,τ_{XY}
Sparsity	sparsity(X)	Degree of discreteness [143]	min,max,μ,σ
Gravity	gravity(X)	Inter-class dispersion [5]	
ANOVA p-value	$p_{val_{X_1X_2}}$	Feature redundancy [70]	$p_{val_{XY}}$ [158]
Coeff. of variation	$\frac{\sigma_Y}{\mu_Y}$	Variation in target [158]	
PCA ρ_{λ_1}	$\sqrt{\frac{\lambda_1}{1+\lambda_1}}$	Variance in first PC [99]	$\frac{\lambda_1}{\sum_i \lambda_i}$ [99]
PCA skewness		Skewness of first PC [48]	PCA kurtosis [48]
PCA 95%	$\frac{dim_{95\%var}}{p}$	Intrinsic dimensionality [9]	
Class probability	$P(C)$	Class distribution [99]	min,max,μ,σ
Class entropy	$H(C)$	Class imbalance [99]	
Norm. entropy	$\frac{H(X)}{log_2 n}$	Feature informativeness [26]	min,max,μ,σ
Mutual inform.	$MI(C, X)$	Feature importance [99]	min,max,μ,σ
Uncertainty coeff.	$\frac{MI(C,X)}{H(C)}$	Feature importance [3]	min,max,μ,σ
Equiv. nr. feats	$\frac{H(C)}{\overline{MI(C,X)}}$	Intrinsic dimensionality [99]	
Noise-signal ratio	$\frac{\overline{H(X)}-\overline{MI(C,X)}}{\overline{MI(C,X)}}$	Noisiness of data [99]	
Fisher's discrimin.	$\frac{(\mu_{c1}-\mu_{c2})^2}{\sigma_{c1}^2-\sigma_{c2}^2}$	Separability classes c_1, c_2 [64]	See [64]
Volume of overlap		Class distribution overlap [64]	See [64]
Concept variation		Task complexity [180]	See [179, 180]
Data consistency		Data quality [76]	See [76]
Nr nodes, leaves	$\|\eta\|$, $\|\psi\|$	Concept complexity [113]	Tree depth
Branch length		Concept complexity [113]	min,max,μ,σ
Nodes per feature	$\|\eta_X\|$	Feature importance [113]	min,max,μ,σ
Leaves per class	$\frac{\|\psi_c\|}{\|\psi\|}$	Class complexity [49]	min,max,μ,σ

Problems in auto-sklearn 1.0

- It is time-consuming since it requires to compute meta-features describing the characteristics of datasets.
- It adds complexity to the system as the computation of the meta-features must also be done with a time and memory limit.
- Many meta-features are not defined with respect to categorical features and missing values, making them hard to apply for most datasets.
- It is not immediately clear which meta-features work best for which problem.

Forget about meta-features - task independent portfolio

Algorithm 1: Greedy Portfolio Building

Input: Set of candidate ML pipelines $\mathcal{C}$, $\mathbf{D}_{\text{meta}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{|\mathbf{D}_{\text{meta}}|}\}$, maximal portfolio size p, model selection strategy S

$\mathcal{P} = \emptyset$

while $|\mathcal{P}| < p$ **do**

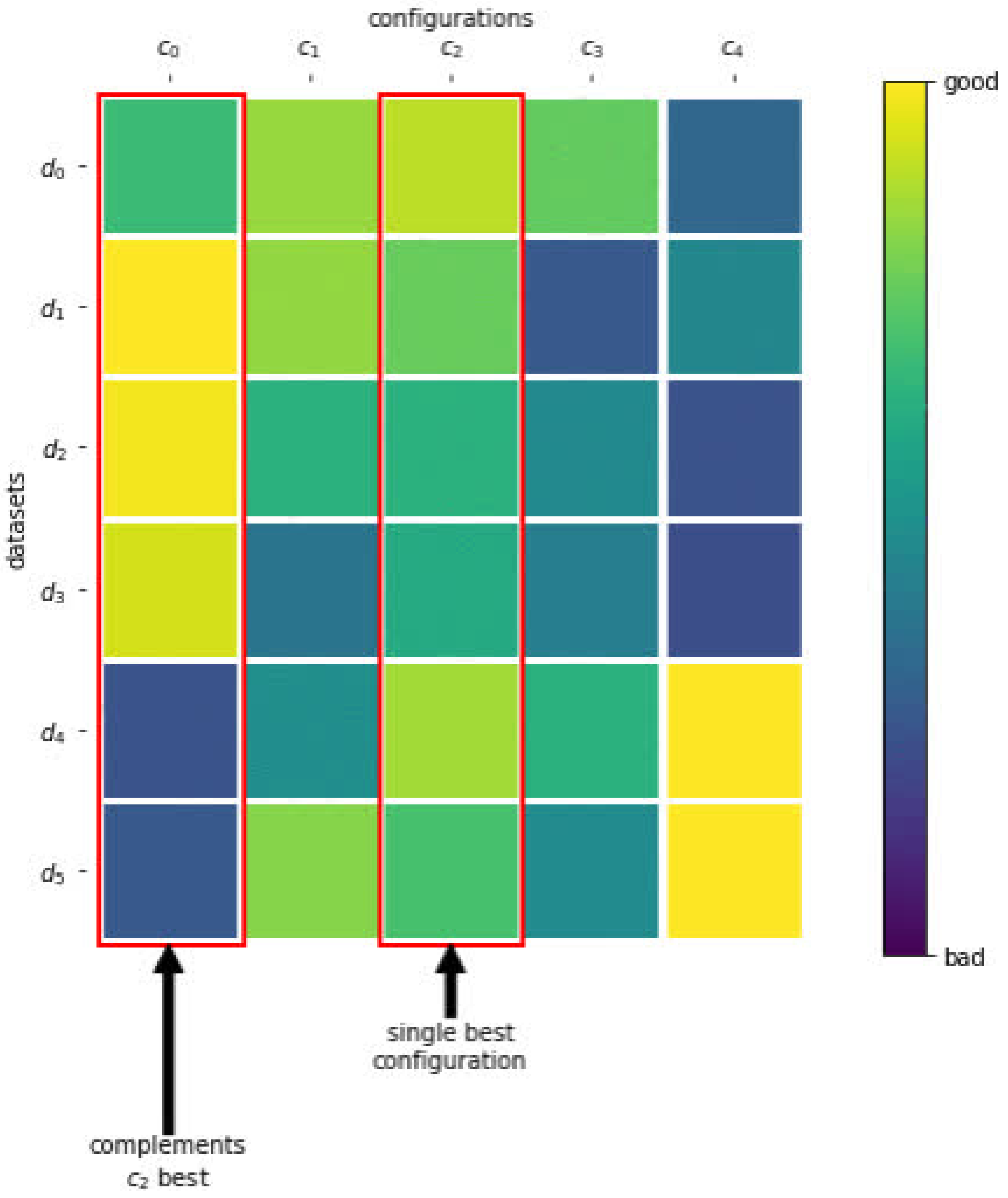
  $\boldsymbol{\lambda}^+ = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathcal{C}} \widehat{GE}_S(\mathcal{P} \cup \{\boldsymbol{\lambda}\}, \mathbf{D}_{\text{meta}})$

  `// Ties are broken favoring the model trained first.`

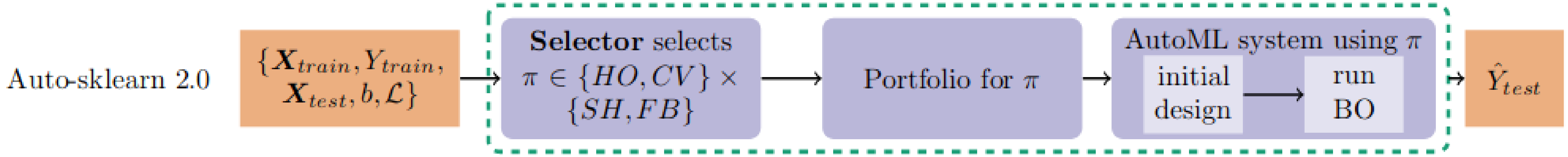
  $\mathcal{P} = \mathcal{P} \cup \boldsymbol{\lambda}^+,\ \mathcal{C} = \mathcal{C} \setminus \{\boldsymbol{\lambda}^+\}$

end while

return Portfolio $\mathcal{P}$



Meta-learning in auto-sklearn 2.0



"… we propose a meta-feature-free approach that does not warmstart with a set of configurations specific to a new dataset, but which uses a static portfolio – a set of complementary configurations that covers as many diverse datasets as possible and minimizes the risk of failure when facing a new task…"

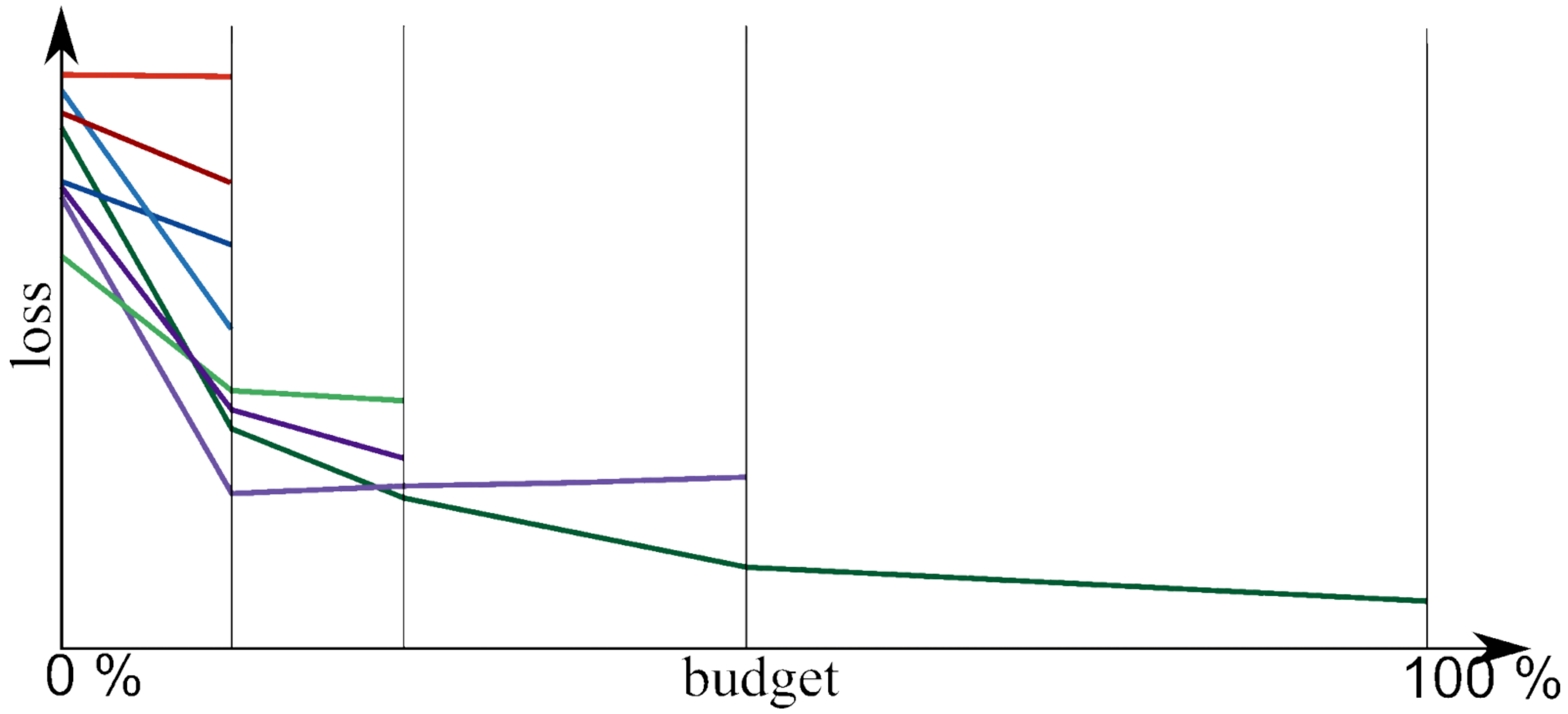
Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer and Frank Hutter, Auto-Sklearn 2.0: The Next Generation, 2022

Successive halving and HyperBand

Idea: szybko eliminujmy konfiguracje, które mają słaby performance, a bardziej eksploatujmy te które dają obiecujące wyniki

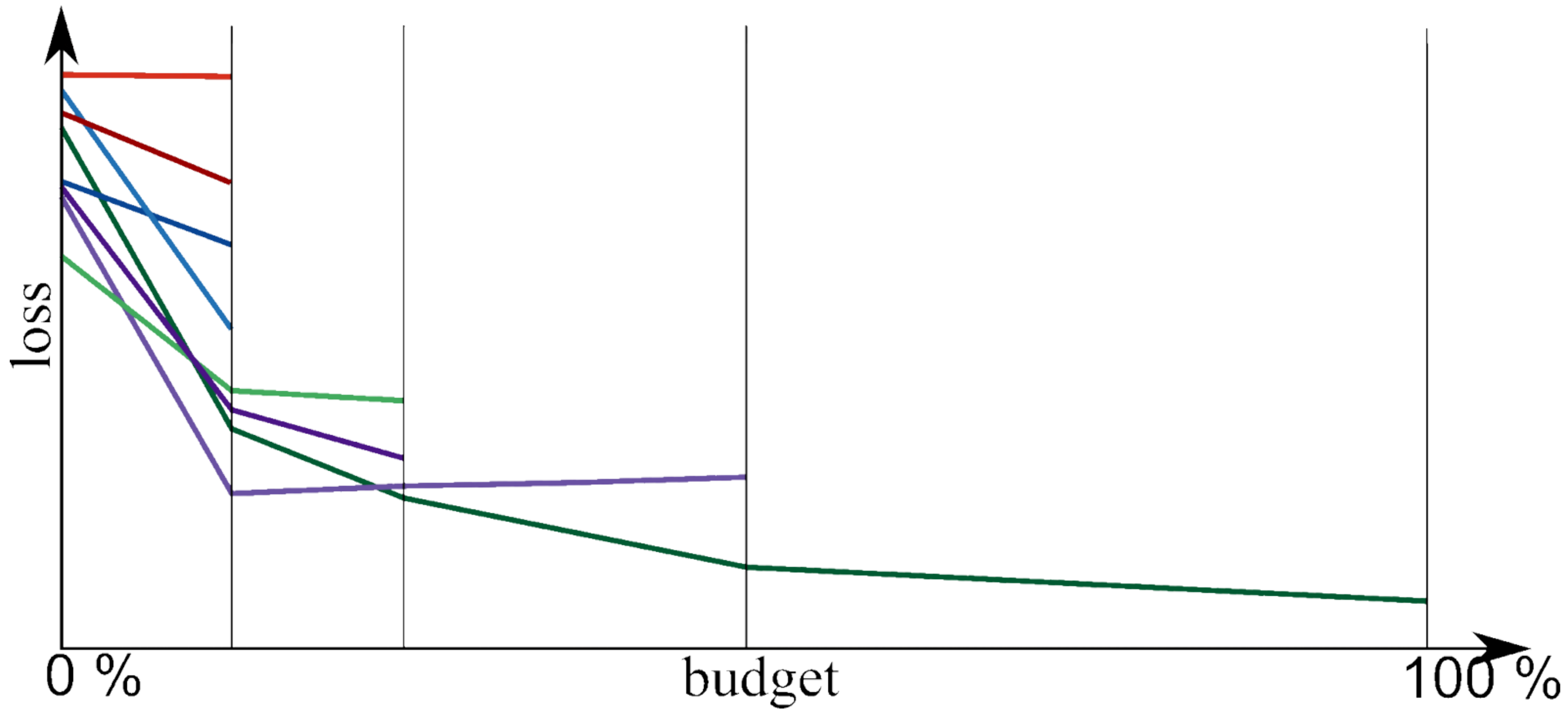
Jak wprowadzać różny budżet?

- liczba epok/ iteracji
- rozmiar danych treningowych
- liczba drzew w algortymie RF, XGBoosting
- liczba zmiennych (features)

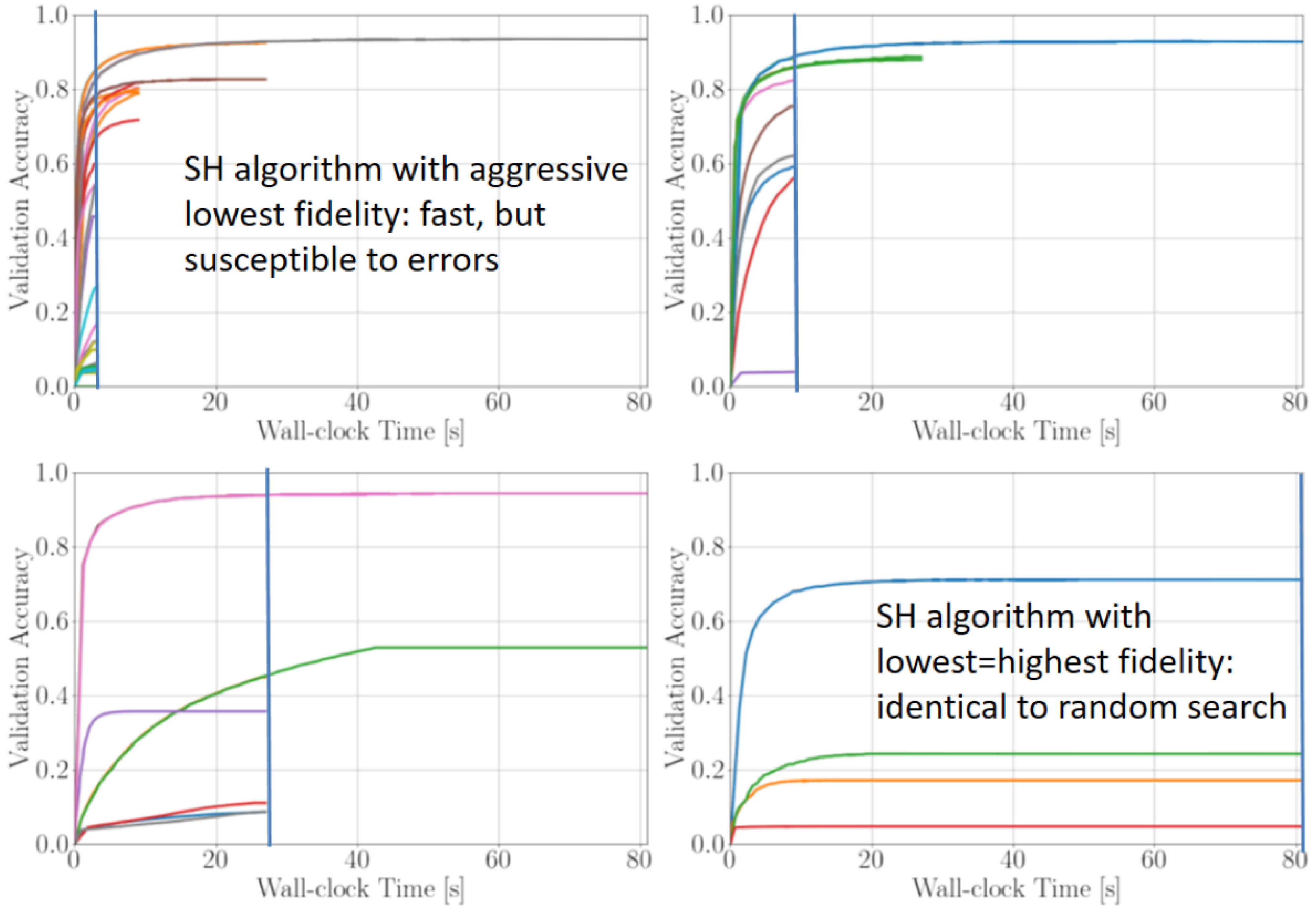


Successive Halving

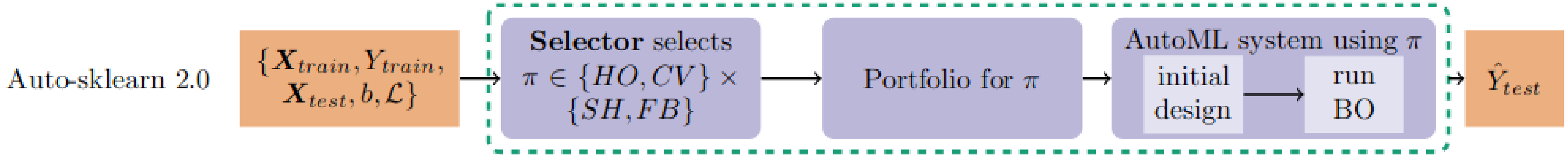
The search strategy starts evaluating all the candidates with a small amount of resources and iteratively selects the best candidates, using more and more resources.



Hyperband


SH algorithm with aggressive lowest fidelity: fast, but susceptible to errors
SH algorithm with lowest=highest fidelity: identical to random search
Validation Accuracy
Wall-clock Time [s]
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80

Budget allocation strategy in auto-sklearn 2.0



"… we introduce budget allocation strategies as a complementary design choice to model selection strategies (holdout (HO) and cross-validation (CV)) for AutoML systems. We suggest using the budget allocation strategy successive halving (SH) as an alternative to always using the full budget (FB) to evaluate a configuration to allocate more resources to promising ML pipelines…"

Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer and Frank Hutter, Auto-Sklearn 2.0: The Next Generation, 2022

Multi-layer stacking

Budowa ensemblingów

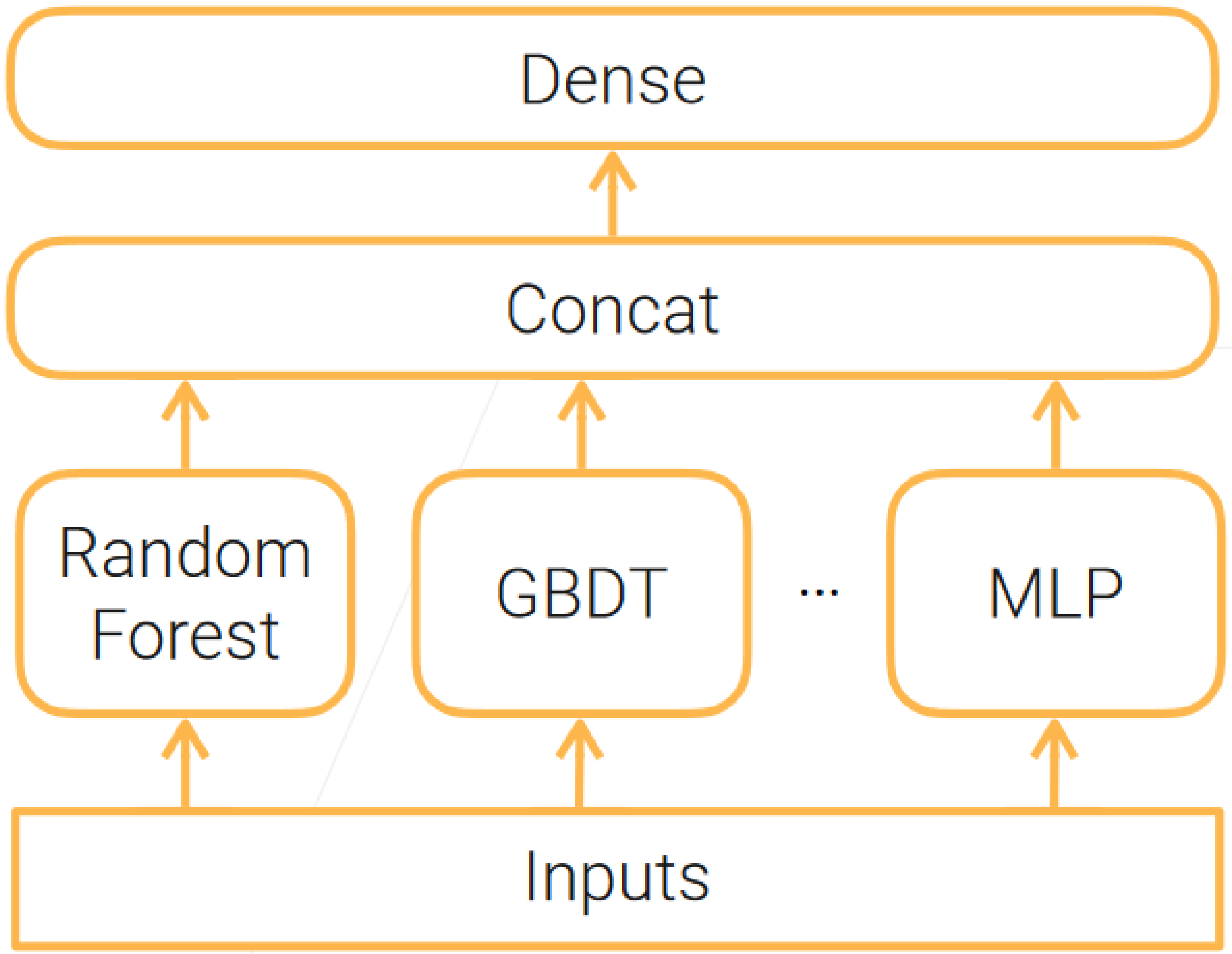
Ensembles

“Ensembles that combine predictions from multiple models have long been known to outperform individual models, often drastically reducing the variance of the final predictions”.

Dietterich, T. G. Ensemble methods in machine learnng. In International Workshop on Multiple Classifier Systems, pp. 1–15. Springer, 2000.

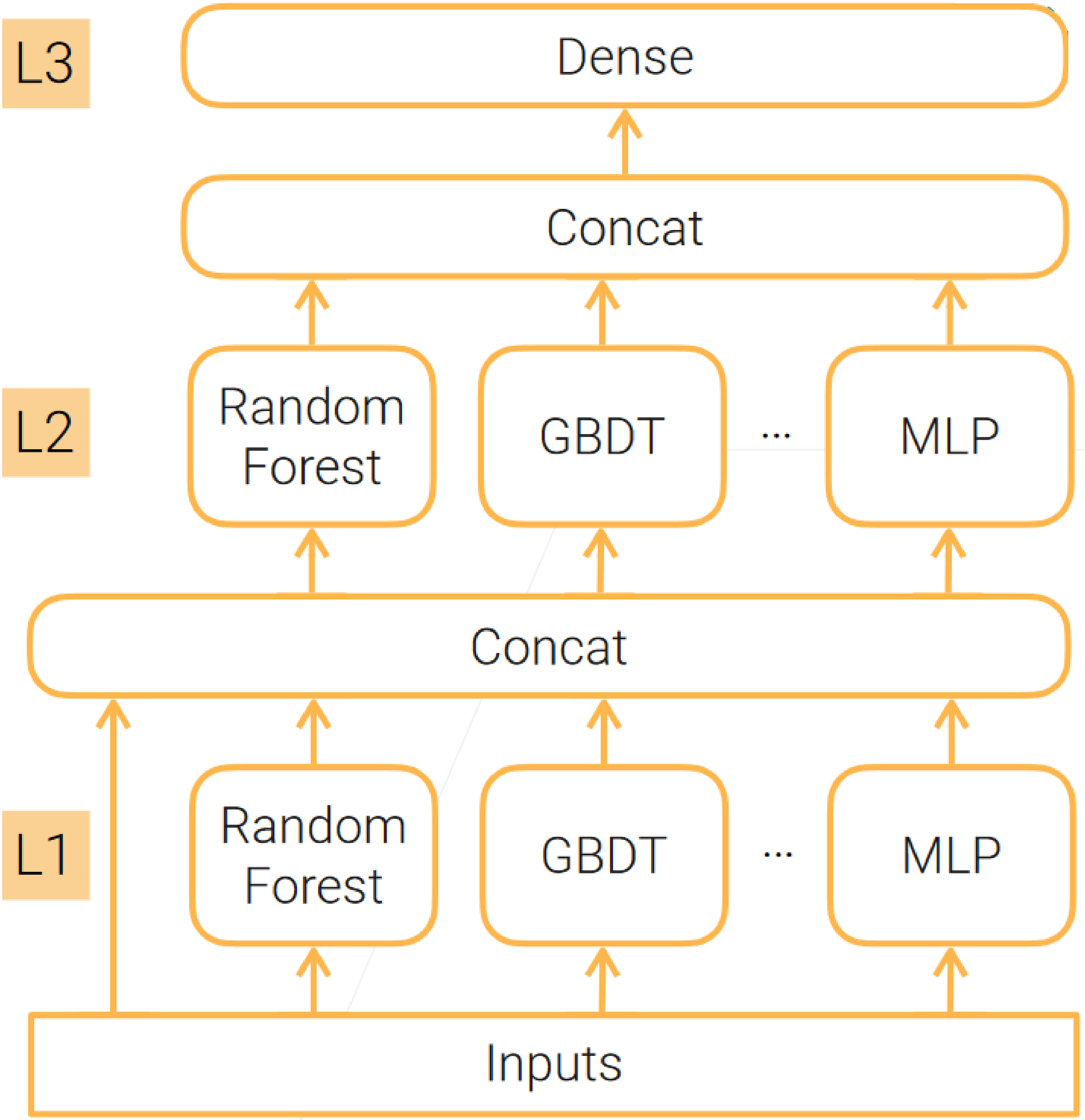
Stacking

Stacking



https://c.d2l.ai/stanford-cs329p/_static/pdfs/cs329p_slides_6_4.pdf

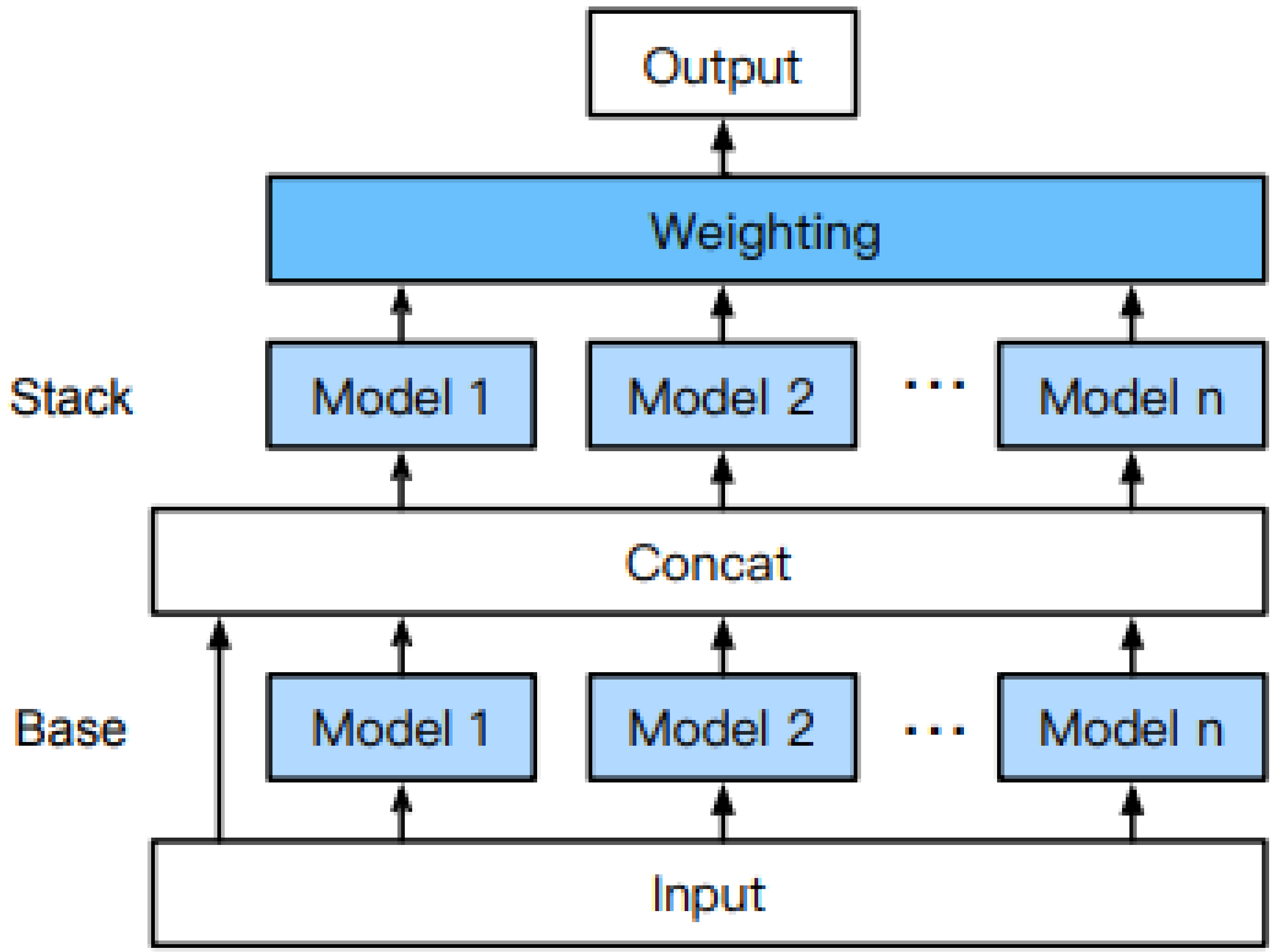
Multi-Layer Stacking



https://c.d2l.ai/stanford-cs329p/_static/pdfs/cs329p_slides_6_4.pdf

Multi-Layer Stack Ensembling

- korzysta z tych samych typów modeli i hiperparametrów we wszystkich warstwach
- w pierwszej warstwie modele są budowane na predykcjach i oryginalnych danych
- ostatnia warstwa agreguje predykcje biorąc pod uwagę wagi
- stosowany jest k-fold bagging



Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, Alexander Smola: AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data, 2020; arXiv:2003.06505

Inne sposoby budowy ensembles
auto-sklearn

The basic ensemble selection procedure is very simple:

- Start with the empty ensemble.
- Add to the ensemble the model in the library that maximizes the ensemble’s performance to the error metric on a validation set.
- Repeat Step 2 for a fixed number of iterations or until all the models have been used.
- Return the ensemble from the nested set of ensembles that has maximum performance on the validation set.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models, 2004.

Improving Ensemble Selection

1. Selection with Replacement

- performance drops because the best models in the library have been used and selection must now add models that hurt the ensemble,
- the loss in performance can be significant if the peak is missed.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models, 2004.

Improving Ensemble Selection

2. Sorted Ensemble Initialization

- forward selection sometimes overfits early in selection when ensembles are small
- starting with empty ensemble, sort the models in the library by their performance, and put the best N models in the ensemble
- we have 5-25 of the best models in ensemble before greedy stepwise selection begins

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models, 2004.

Improving Ensemble Selection

3. Bagged Ensemble Selection

- the number of models in a library increases, the chances of finding combinations of models that overif the validation set increases - *bagging can minimize this problem*
- random sample of models from the library, combination of M models overfits, the probability of those M models being in a random bag of models in less than (1-p)^M for p the fraction of models in the bag

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models, 2004.

AutoML ⟶ AutoDS

Automated Data Science

Automated Data Science

Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...

Model Building:
algorithm selection,
parameter optimization,
performance evaluation,
model selection,
...

Less dependent on domain context

Data Exploration:
domain understanding,
goal exploration,
data aggregation,
data visualization,
...

Exploitation:
model interpretation and visualization,
reporting and narratives,
predictions and decisions,
monitoring and maintenance,

More dependent on domain context

More open-ended

Less open-ended

Pytania?