AutoML Meets Materials Science

Automating Insights and Innovation

Agenda

Machine Learning in Materials Science

What is AutoML?

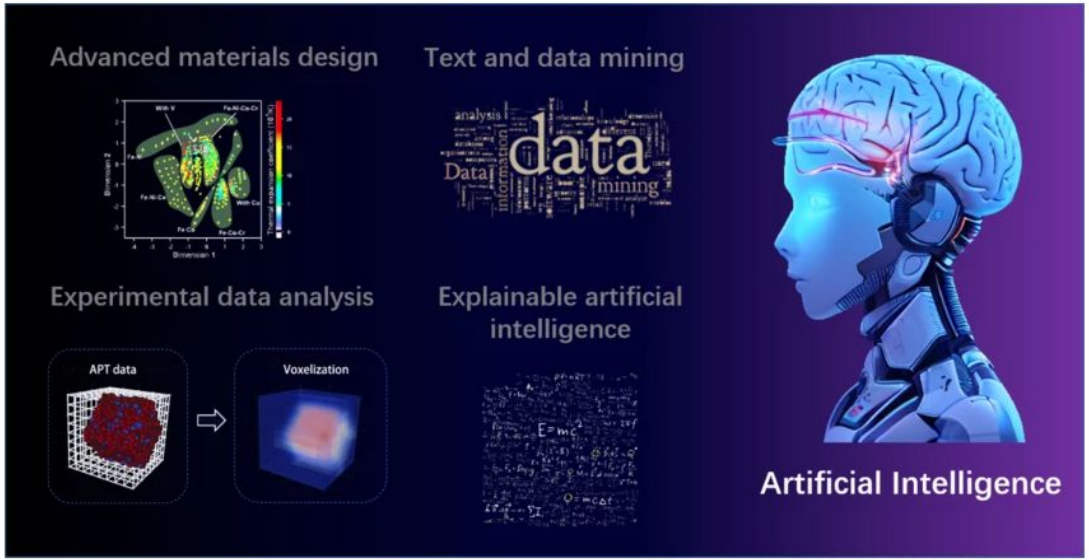
Prospects for AutoML in Materials Science

How can Materials Science help AutoML?

Machine Learning in Materials Science

Machine Learning already helps Materials Science!

Machine Learning already helps Materials Science!



Machine Lear
Material Des
By Steven Davies / Decem

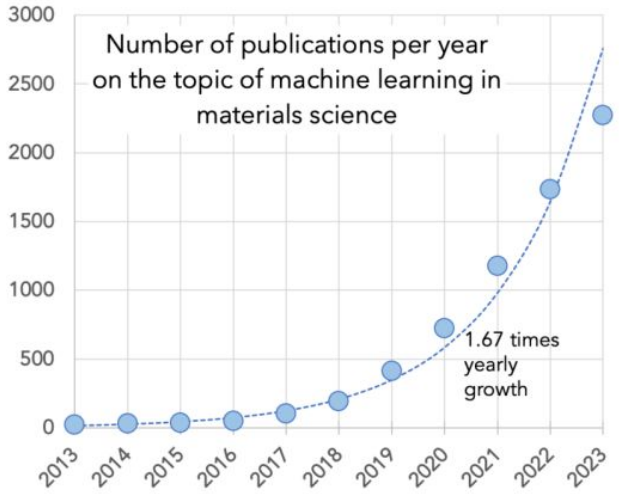
ials science: AI
ritic growth in
icts dendritic growth in thin films,

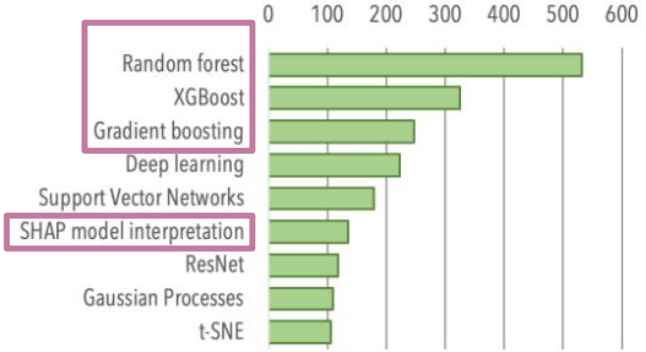
Millior

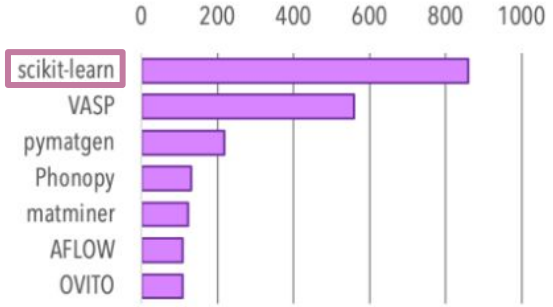
Artificial Intelligence for Materials Science

Welcome to the Artificial Intelligence for Materials Science group, where we are at the forefront of integrating cutting-edge AI technologies into the world of materials discovery and design. Our research is dedicated to transforming how materials are conceptualized, characterized, and optimized by developing next-generation AI tools and frameworks tailored to the unique challenges of materials science

Some Statistics...





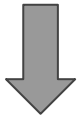


Jain A. Machine learning in materials research: developments over the last decade and challenges for the future. ChemRxiv. 2024

Machine Learning as a *tool*

*“[…] our new deep learning **tool** that dramatically increases the speed and efficiency of discovery by predicting the stability of new materials.”* – Google DeepMind Blog on GNoME

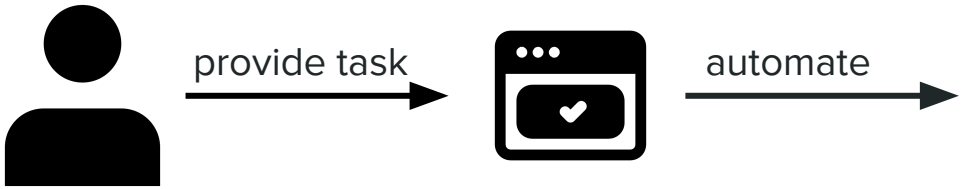
Thus, materials science researchers should be able to use ML as an every day tool! We shouldn’t expect them to get ML experts!

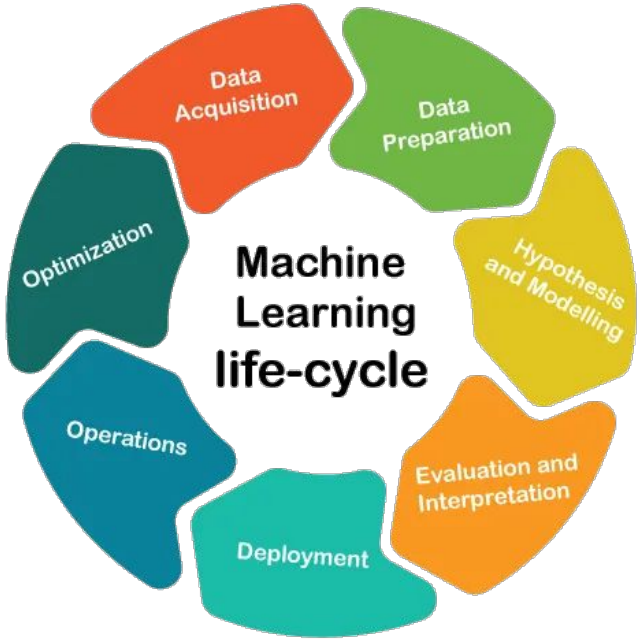


That’s where AutoML can help!

What is AutoML?

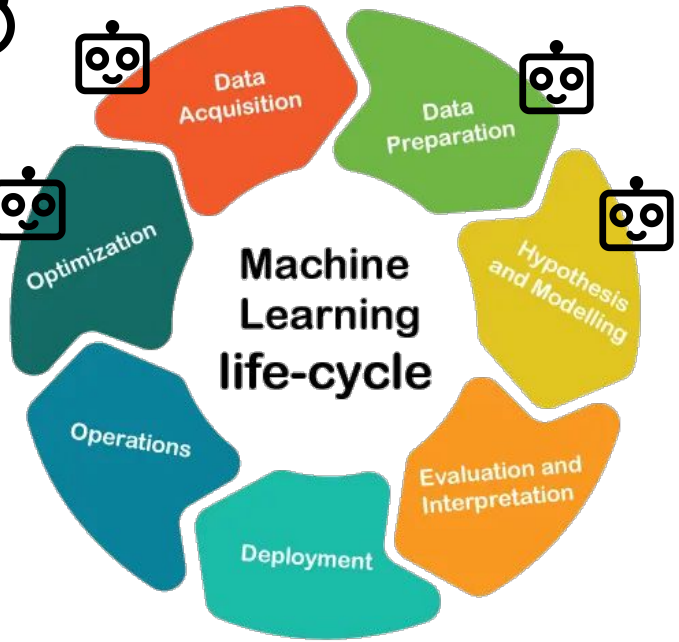
Goals of AutoML





ML Lifecycle image from: https://amansingh-javatpoint.medium.com/machine-learning-life-cycle-7b0e5f940097

What can AutoML do already?



CAFFE

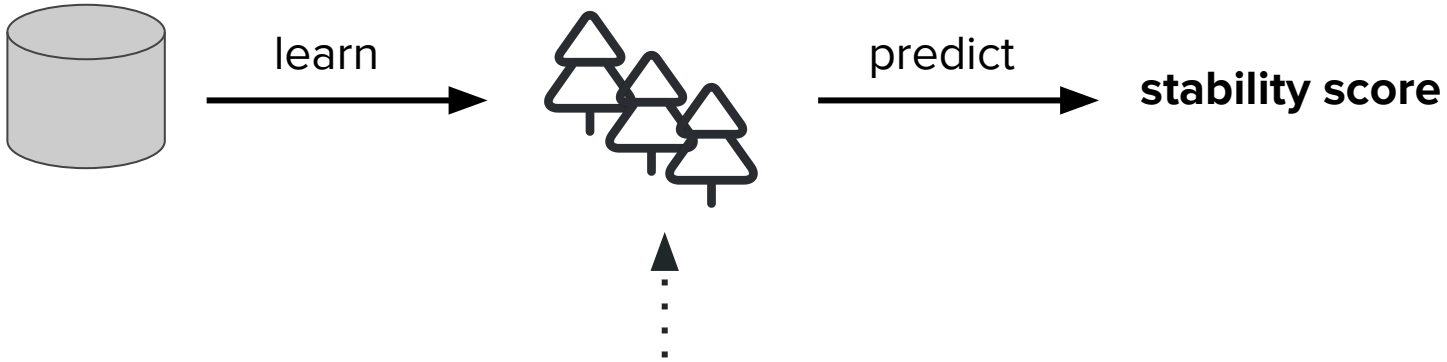
OPTUNA

Neural Network Intelligence

HuggingGPT

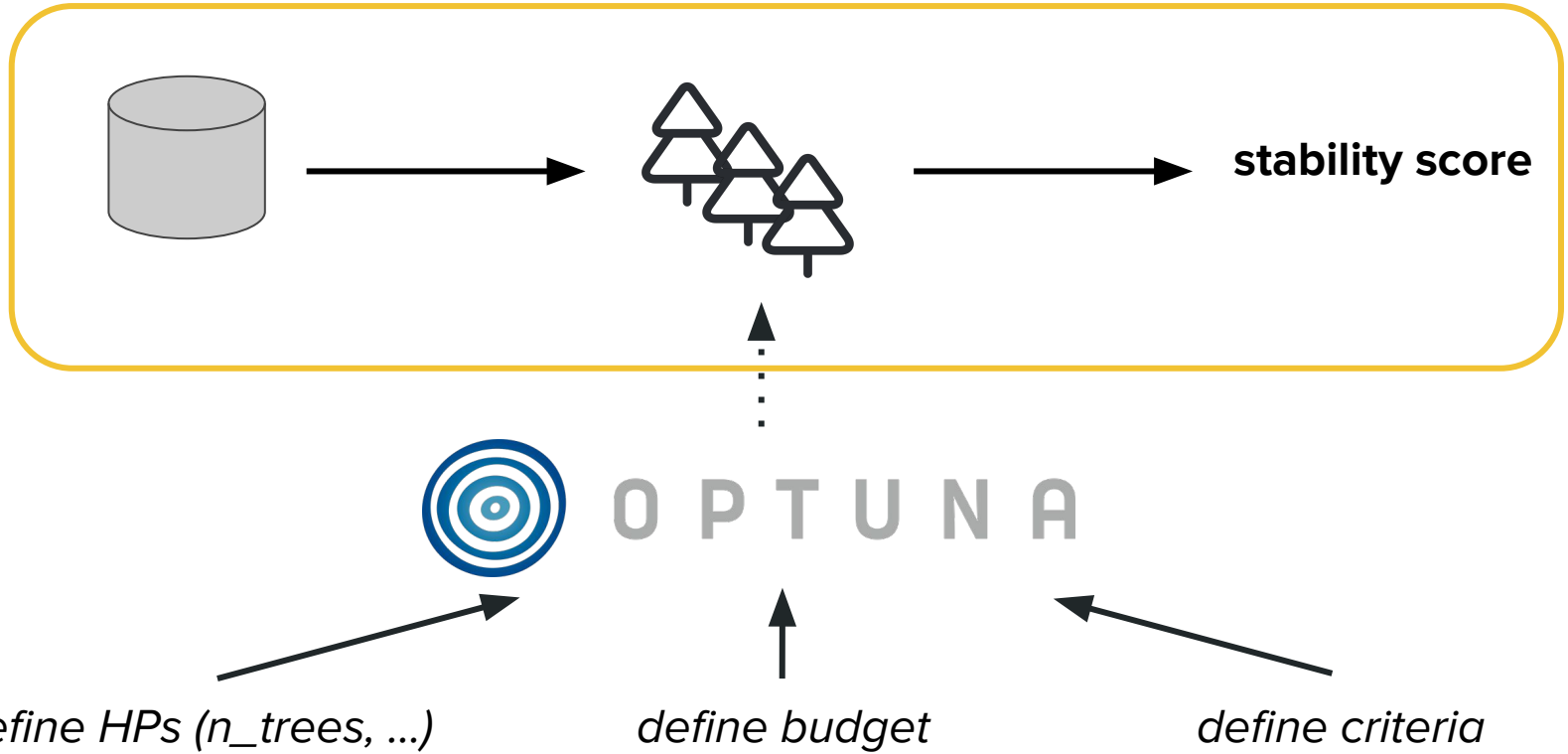
ML Lifecycle image from: https://amansingh-javatpoint.medium.com/machine-learning-life-cycle-7b0e5f940097

Example: Hyperparameter Optimization (Optuna)



n_trees, tree_depth, min_leaves, ...

Example: Hyperparameter Optimization (Optuna)



Example: Hyperparameter Optimization (Optuna)

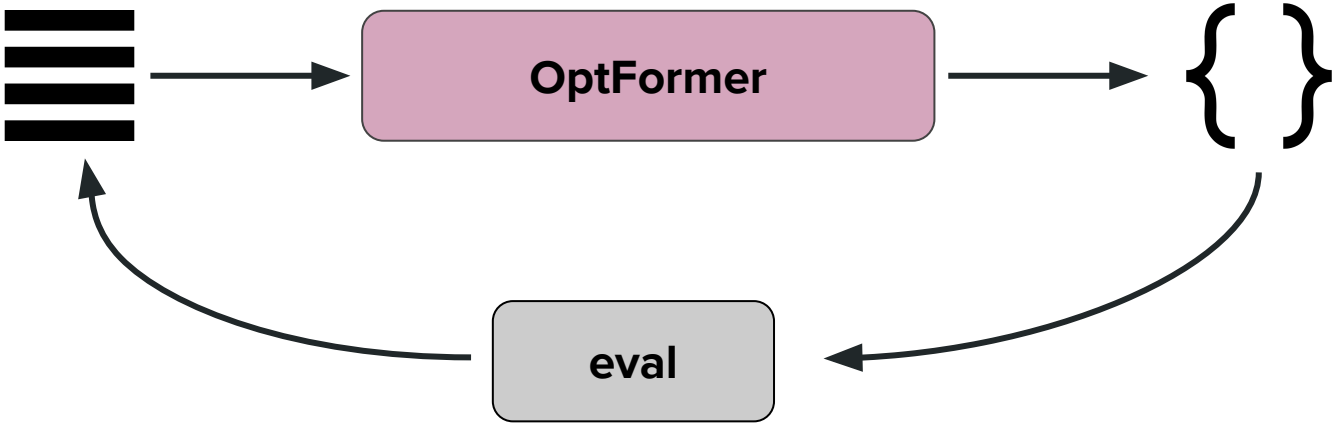
This only costs you a few lines of code!

```python
import optuna

import sklearn.datasets
import sklearn.ensemble
import sklearn.model_selection
import sklearn.svm


def objective(trial):
    iris = sklearn.datasets.load_iris()
    x, y = iris.data, iris.target

    rf_max_depth = trial.suggest_int("rf_max_depth", 2, 32, log=True)
    classifier_obj = sklearn.ensemble.RandomForestClassifier(
            max_depth=rf_max_depth, n_estimators=10
    )

    score = sklearn.model_selection.cross_val_score(classifier_obj, x, y, n_jobs=-1, cv=3)
    accuracy = score.mean()
    return accuracy


if __name__ == "__main__":
    study = optuna.create_study(direction="maximize")
    study.optimize(objective, n_trials=100)
    print(study.best_trial)
```

Example: Hyperparameter Optimization (OptFormer)

Idea: Learn to mimic HPO algorithms.

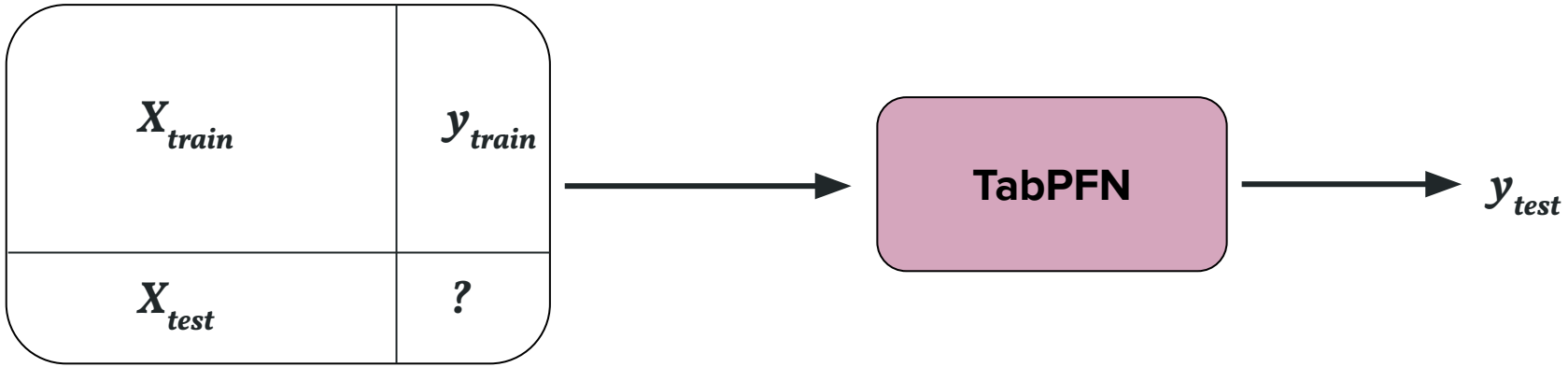
Approach: Treat it as a language modeling task where the next token to predict is the next configuration to test!



Example: TabPFN

Idea: Learn to directly mimic entire ML algorithms/pipelines for tabular data.

Approach: Treat regression/classification as a language modeling task where the next token to predict is the next column of the dataset!



Challenges for AutoML: Acquiring Users

Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020

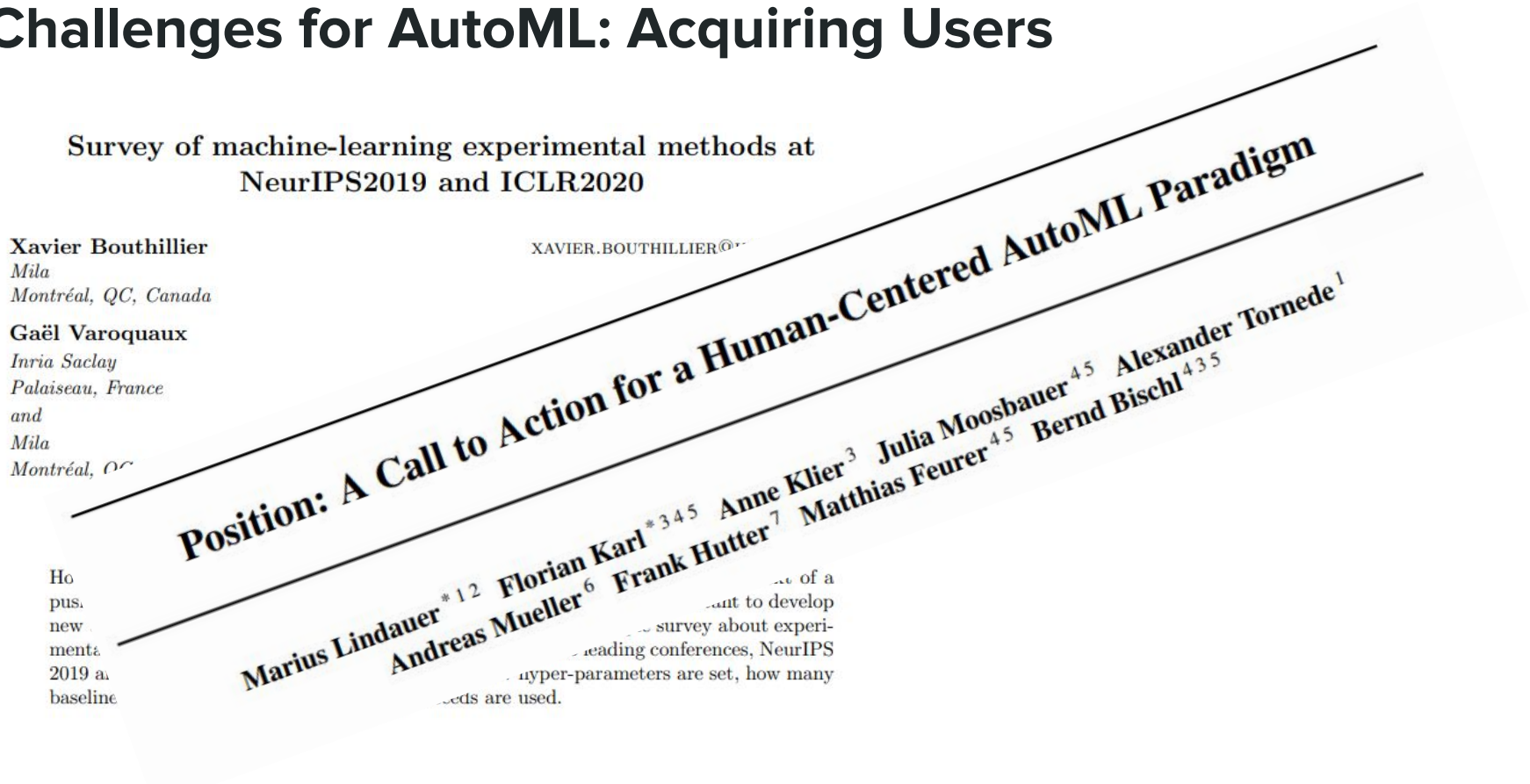
Xavier Bouthillier XAVIER.BOUTHILLIER@UMONTREAL.CA
Mila
Montréal, QC, Canada

Gaël Varoquaux GAEL.VAROQUAUX@INRIA.FR
Inria Saclay
Palaiseau, France
and
Mila
Montréal, QC, Canada

Abstract

How do machine-learning researchers run their empirical validation? In the context of a push for improved reproducibility and benchmarking, this question is important to develop new tools for model comparison. This document summarizes a simple survey about experimental procedures, sent to authors of published papers at two leading conferences, NeurIPS 2019 and ICLR 2020. It gives a simple picture of how hyper-parameters are set, how many baselines and datasets are included, or how seeds are used.

Challenges for AutoML: Acquiring Users



Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020

Xavier Bouthillier XAVIER.BOUTHILLIER@
Mila
Montréal, QC, Canada

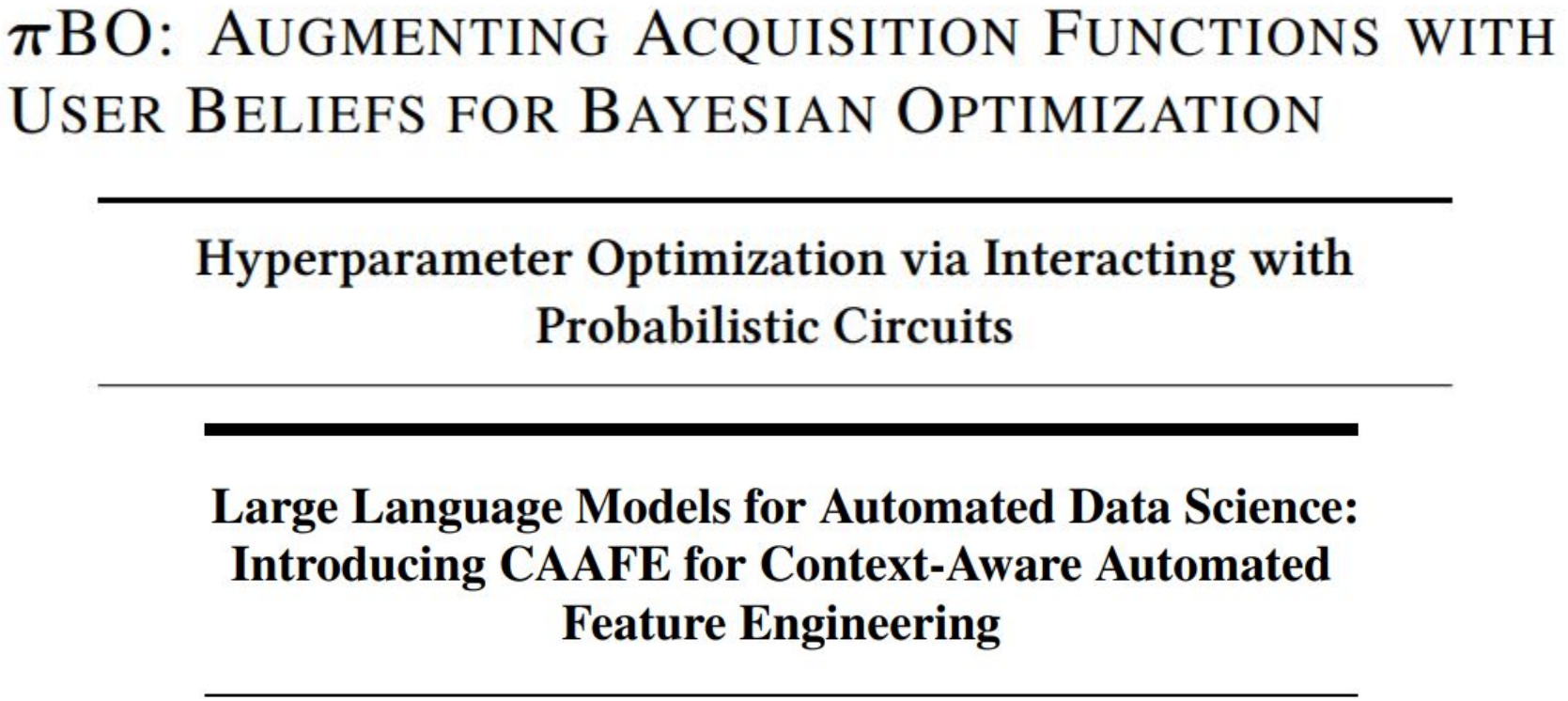
Gaël Varoquaux
Inria Saclay
Palaiseau, France
and
Mila
Montréal, QC

Position: A Call to Action for a Human-Centered AutoML Paradigm

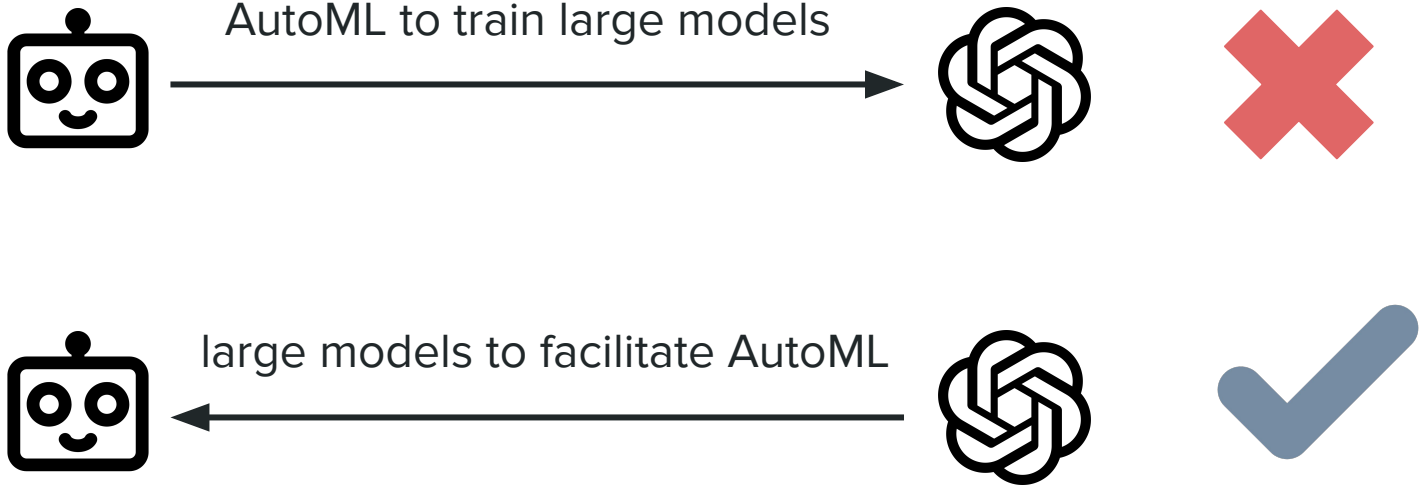
Marius Lindauer [*12] **Florian Karl** [*345] **Anne Klier** [3] **Julia Moosbauer** [45] **Alexander Tornede** [1]
Andreas Mueller [6] **Frank Hutter** [7] **Matthias Feurer** [45] **Bernd Bischl** [435]

Ho … of a
pus… …nt to develop
new … survey about experi-
menta… …leading conferences, NeurIPS
2019 a… …hyper-parameters are set, how many
baseline… …eds are used.

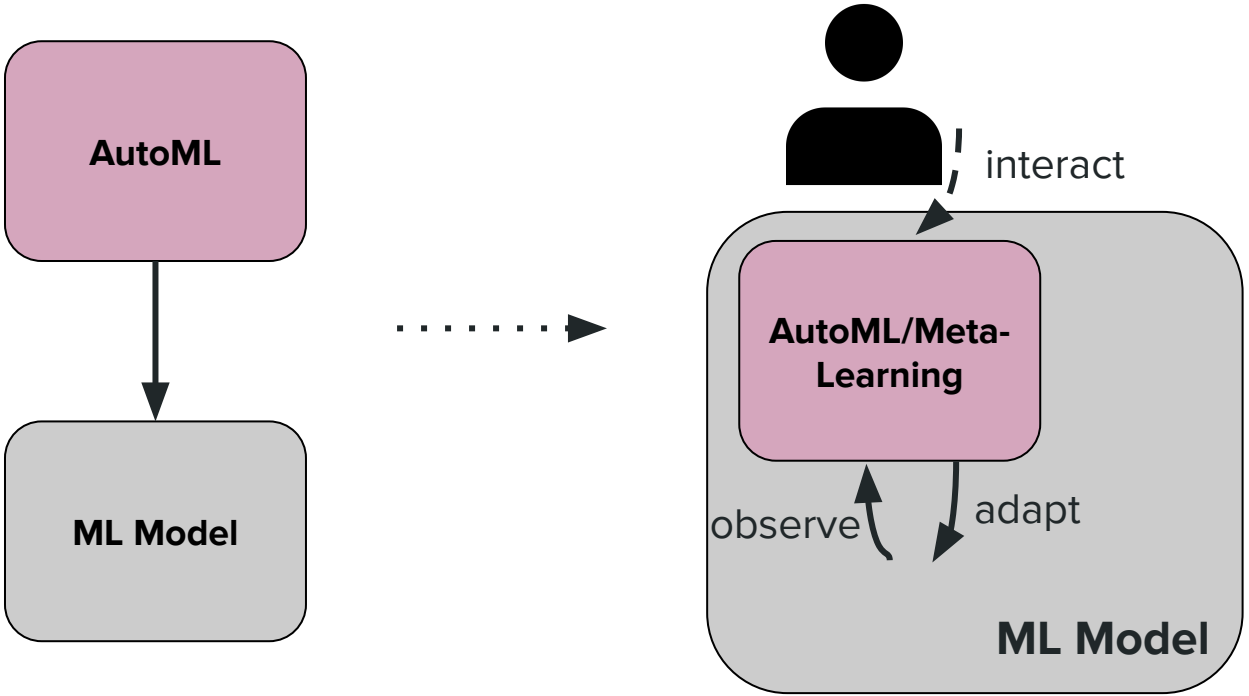
Challenges for AutoML: Giving Control to Users



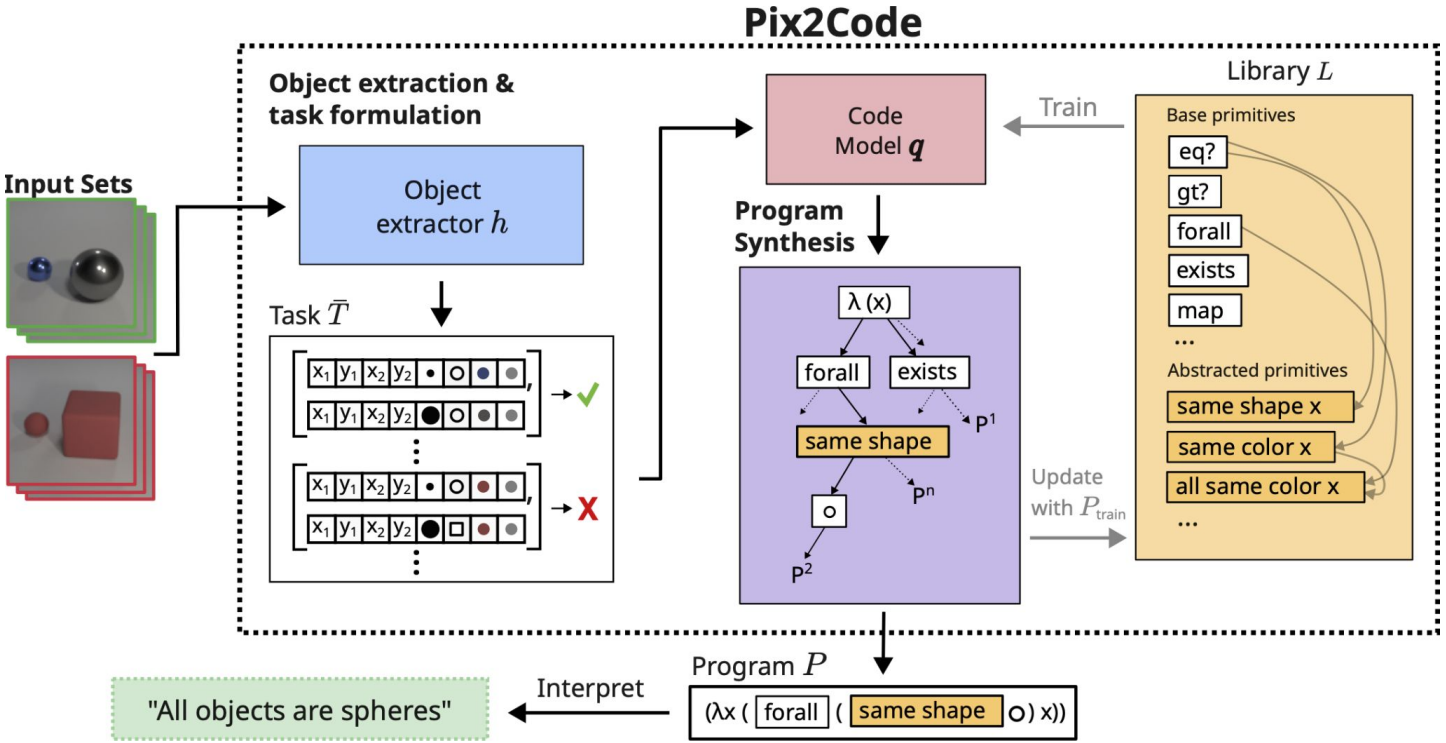
Challenges for AutoML: Large Models



What’s next in AutoML?



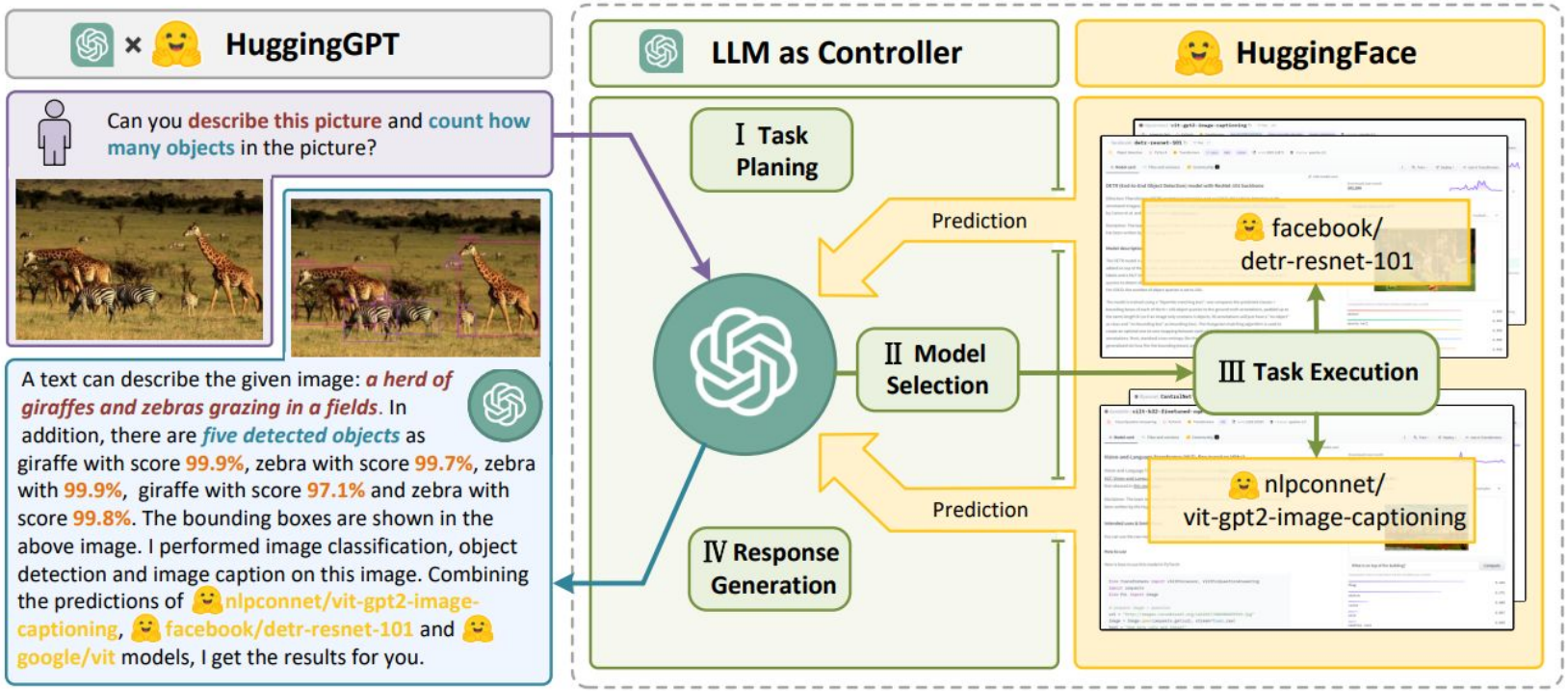
Example: Pix2Code



Wüst et al. Pix2Code: Learning to Compose Neural Visual Concepts as Programs. UAI 2024.

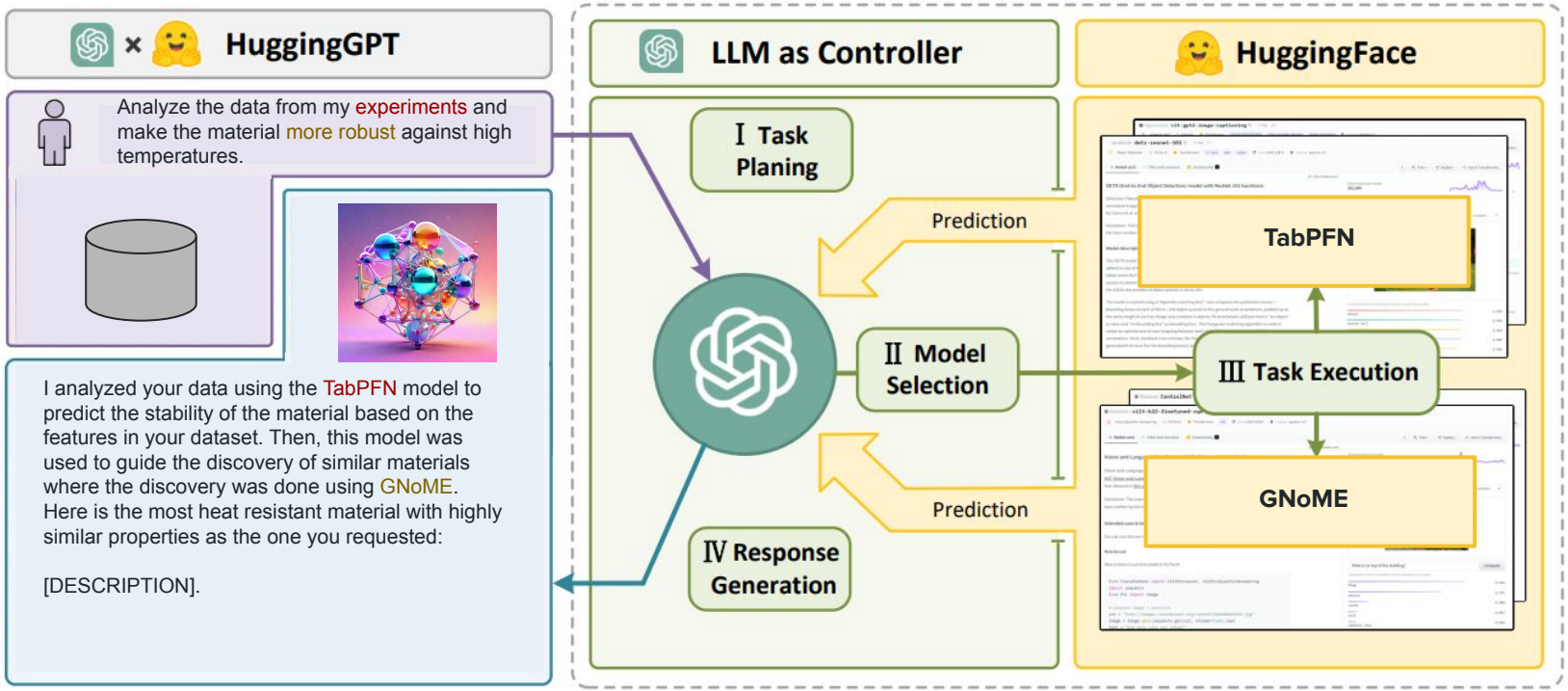
Prospects for AutoML in Materials Science

HuggingGPT for Materials Science?



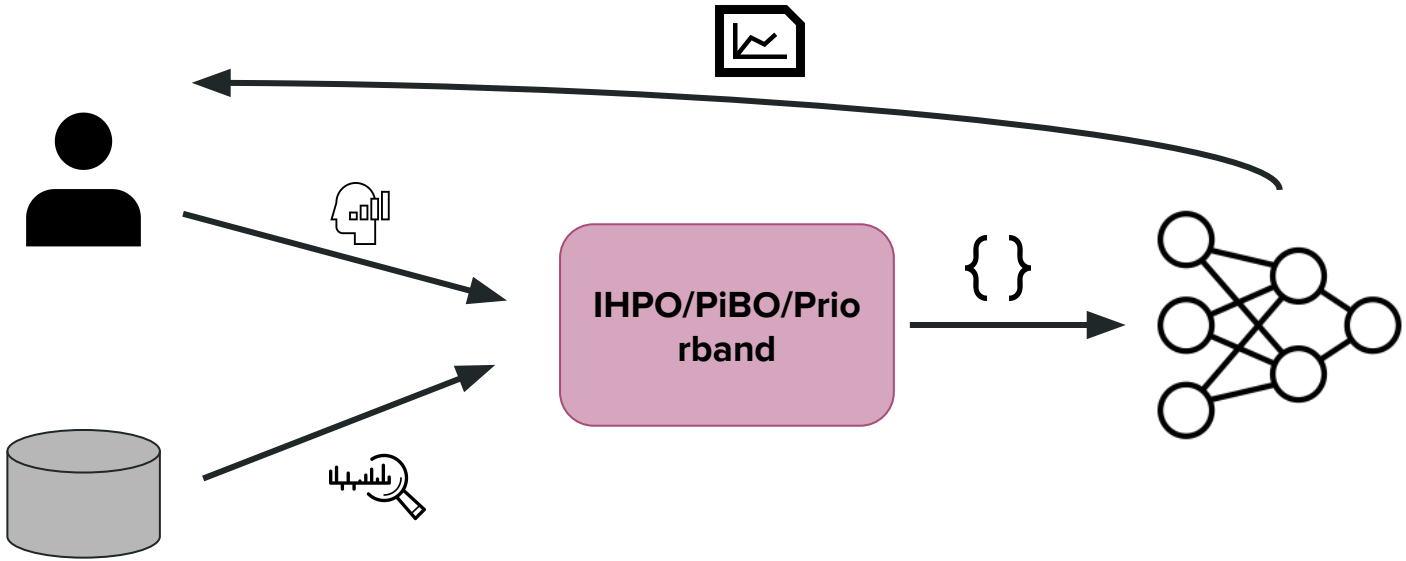
Shen et al. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. NeurIPS 2023.

HuggingGPT for Materials Science?

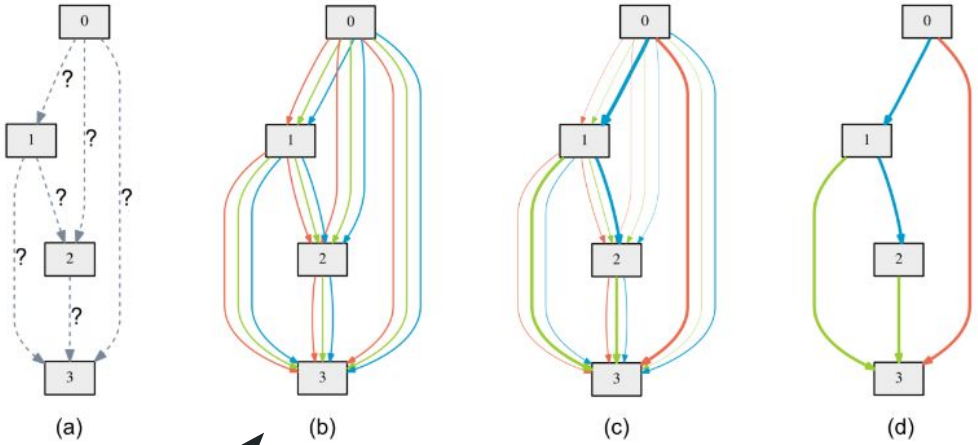


Shen et al. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. NeurIPS 2023.

Interactive Bayesian Optimization for HPO



Neural Architecture Search (NAS) for Material Discovery



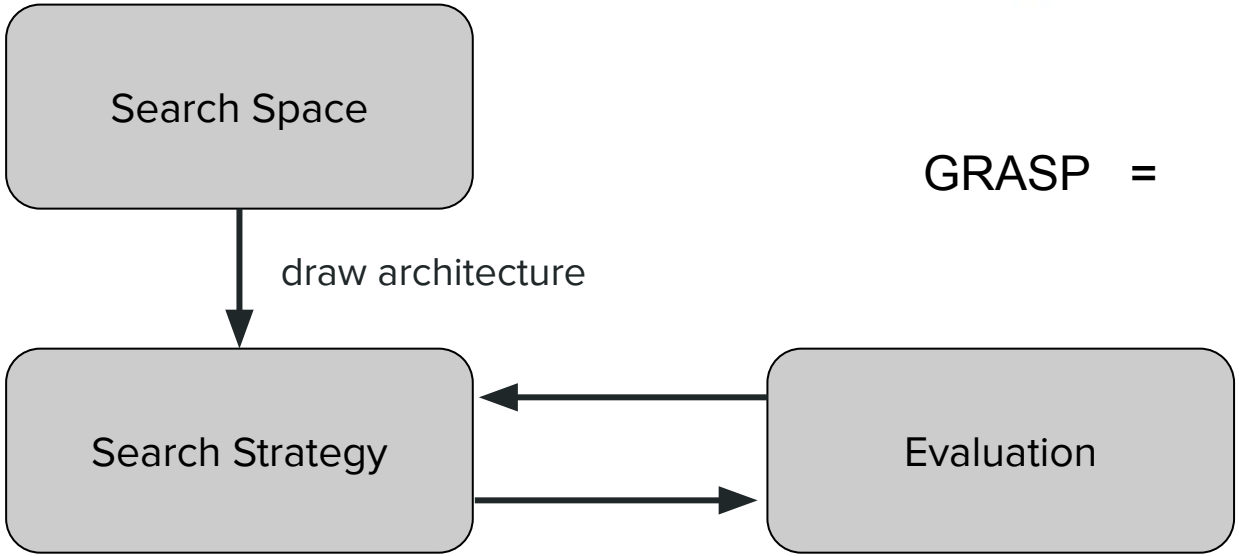
final architecture

search space of architectures

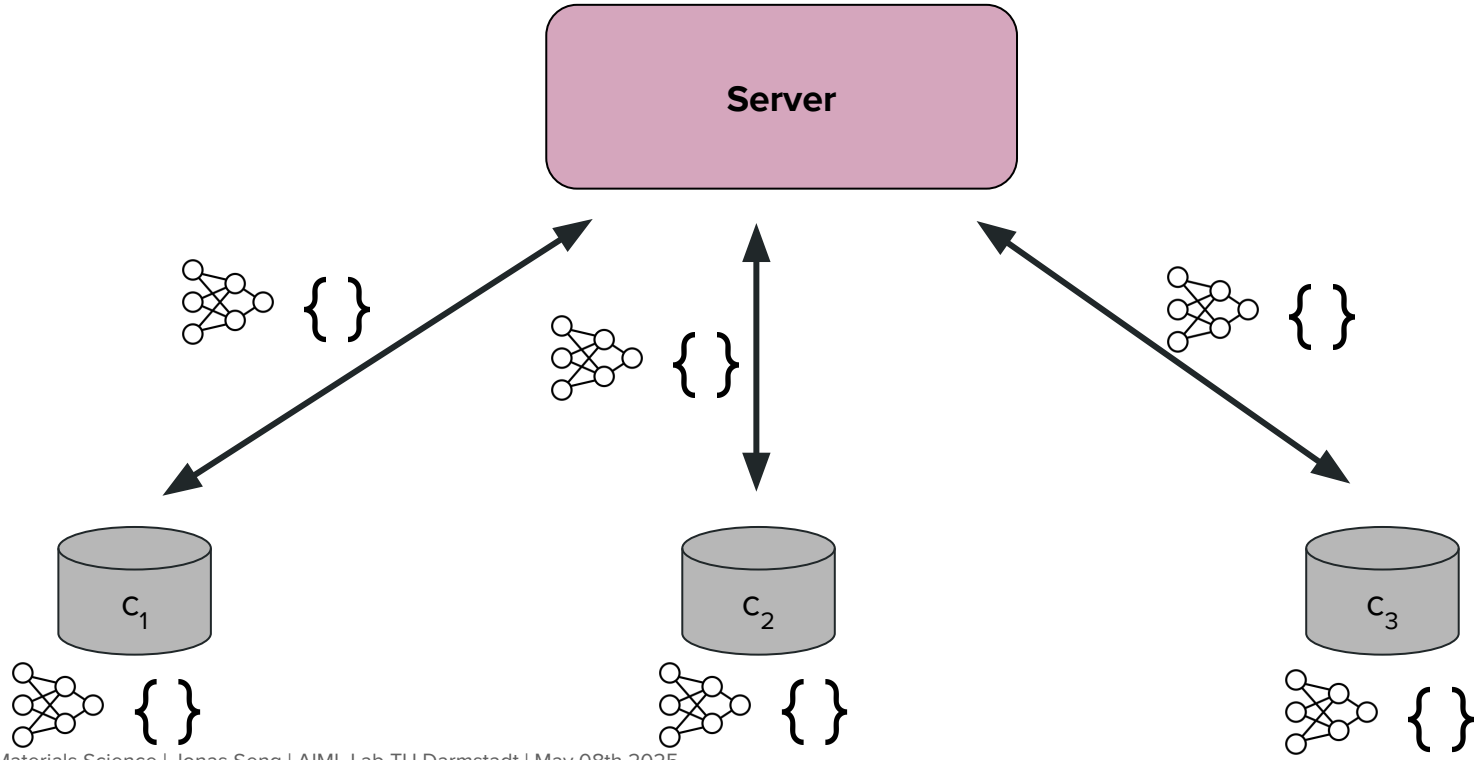
Neural Architecture Search (NAS) for Material Discovery

$$\text{Synflow} \triangleq \sum_{i}^{D} \langle \boldsymbol{\theta}_i, \nabla_{\boldsymbol{\theta}_i} L \rangle$$

$$\text{GRASP} = \sum_{i}^{D} -\langle \boldsymbol{H}_i \nabla_{\boldsymbol{\theta}_i} L, \boldsymbol{\theta}_i \rangle$$



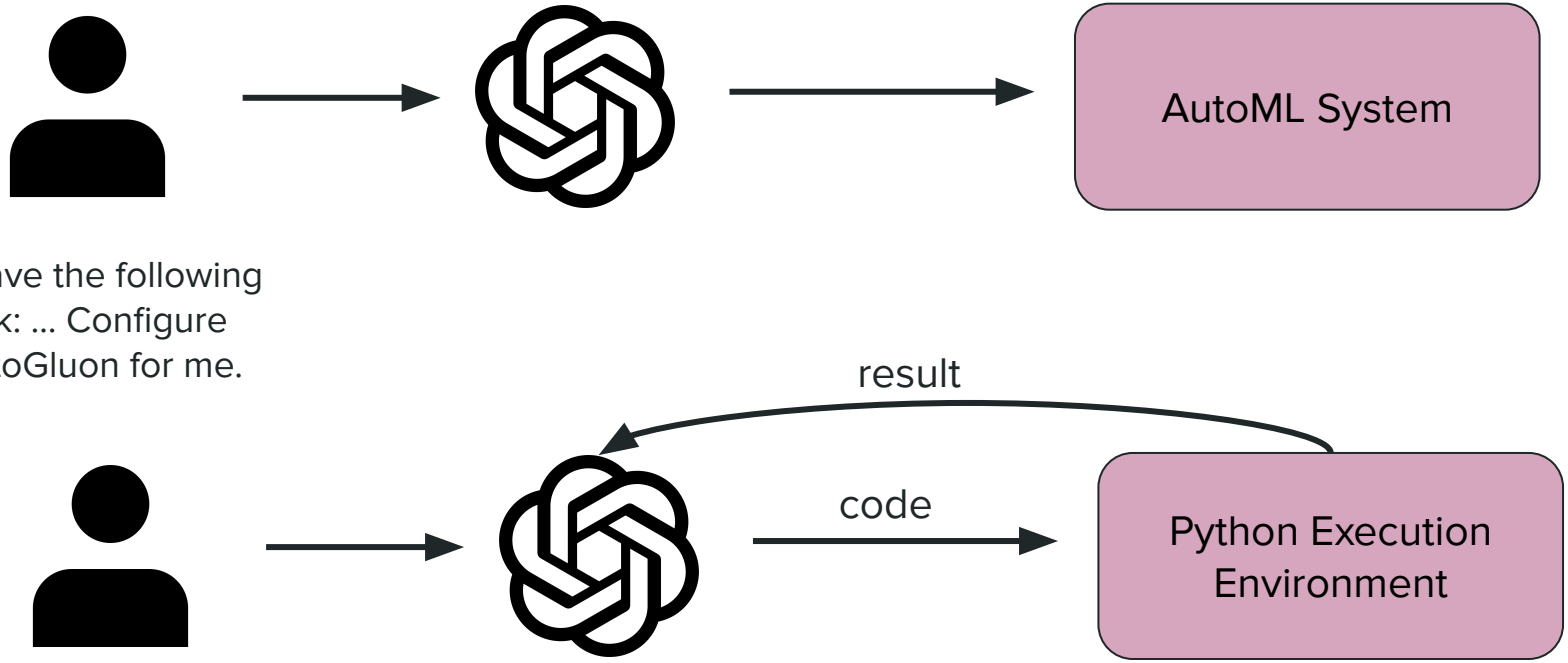
Distributed NAS and HPO (FEATHERS)



End-to-End AutoML Systems

Framework	Based on	Language
auto-sklearn	BO, Meta-Learning	Python
H20	Stacking	Python/R
AutoKeras	BO based NAS	Python
AutoGluon	Multi-Layer Stacking, BO	Python
…	…	…

LLMs + End-to-End AutoML Systems

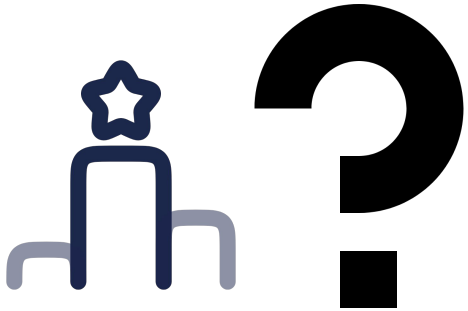


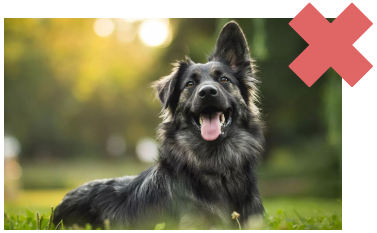
Use what AutoML has to offer!

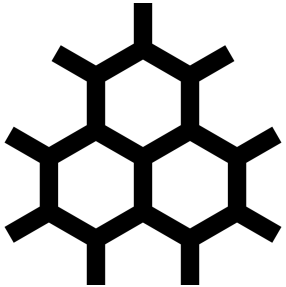
- there are robust libraries and frameworks for HPO ➜ this can substantially decrease efforts for you and yield better results!
- similarly, NAS helps you in designing neural networks efficiently
- end2end systems like AutoWeka can act as a good baseline for your evaluation ➜ make use of it!
- on tabular data, TabPFN can automate large parts of the ML lifecycle already with quite high accuracy

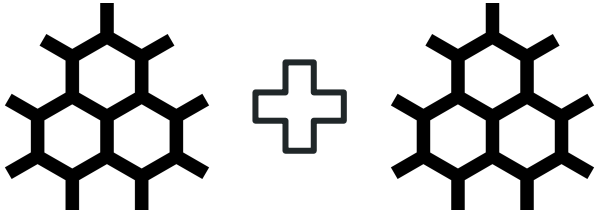
How can Materials Science help AutoML?

Materials Science gives AutoML access to new tasks



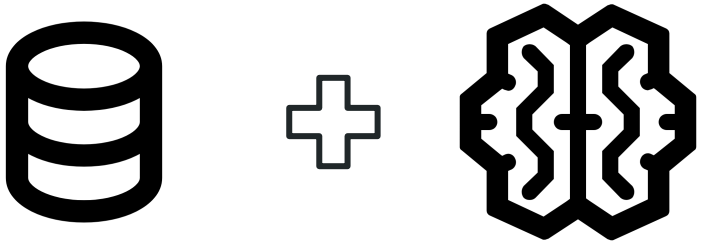






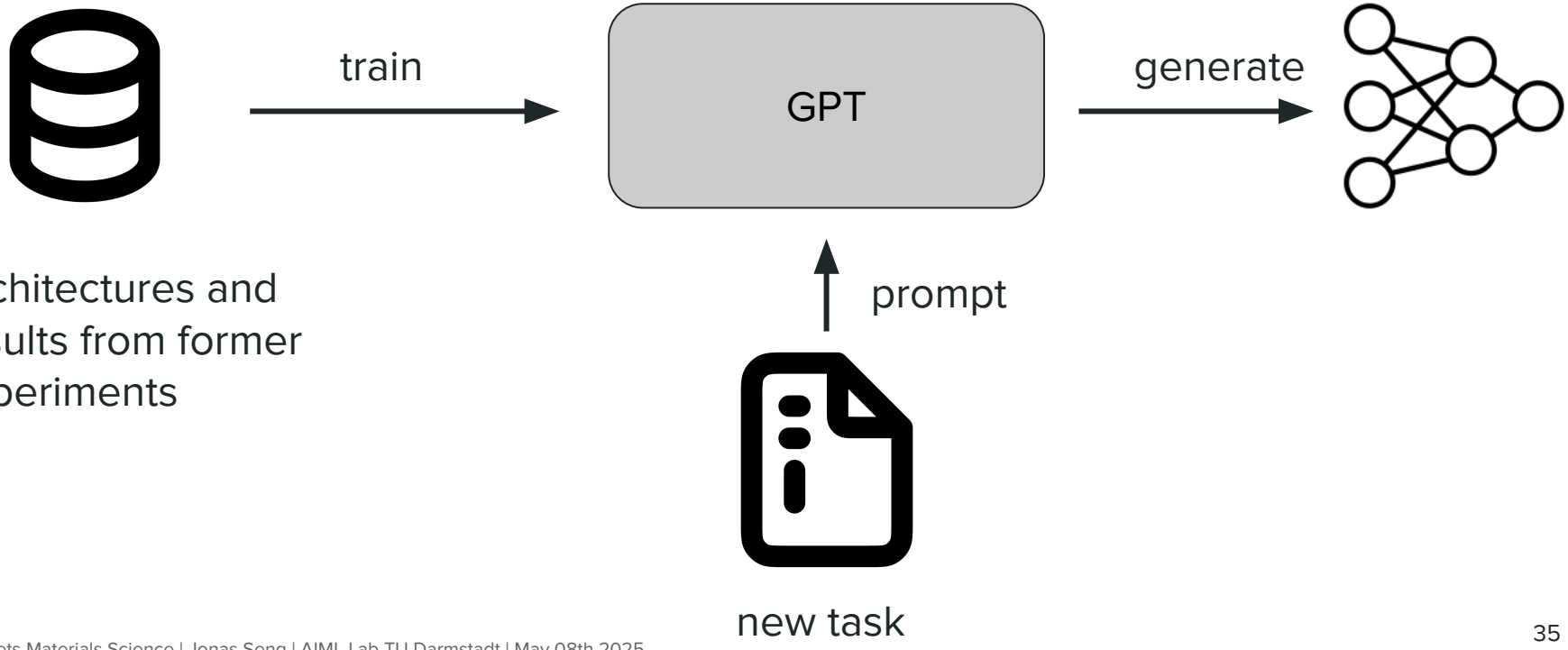
Structured Search Spaces

this should be reflected in search spaces of AutoML systems!

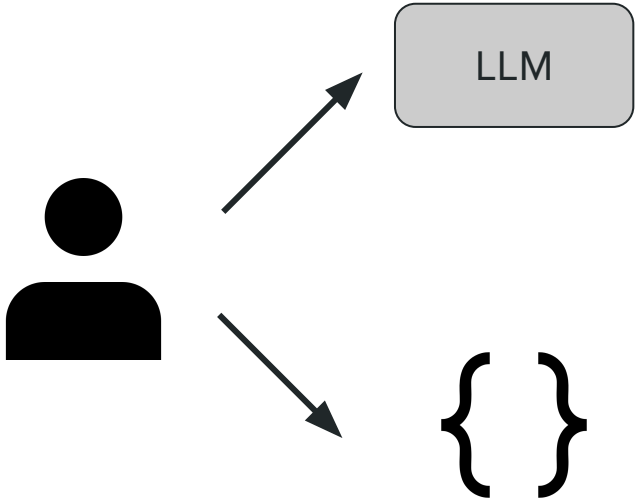


we can learn priors over such structured search spaces based on experimental data

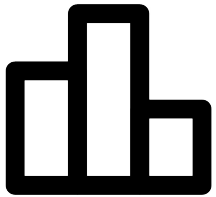
Structured Search Spaces



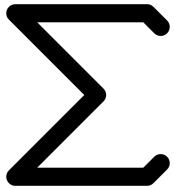
Materials Science can teach us about AutoML's usability



Getting Insights into AutoML decisions: SOTA

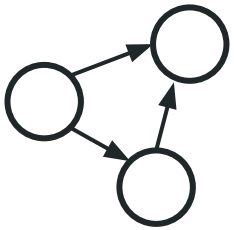


assess importance of hyperparameters on the final model performance

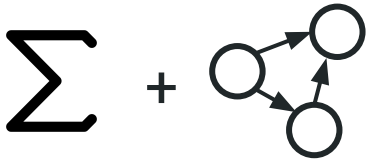


get a symbolic representation of hyperparameter influences on model performance

Getting Insights into AutoML decisions: Next steps



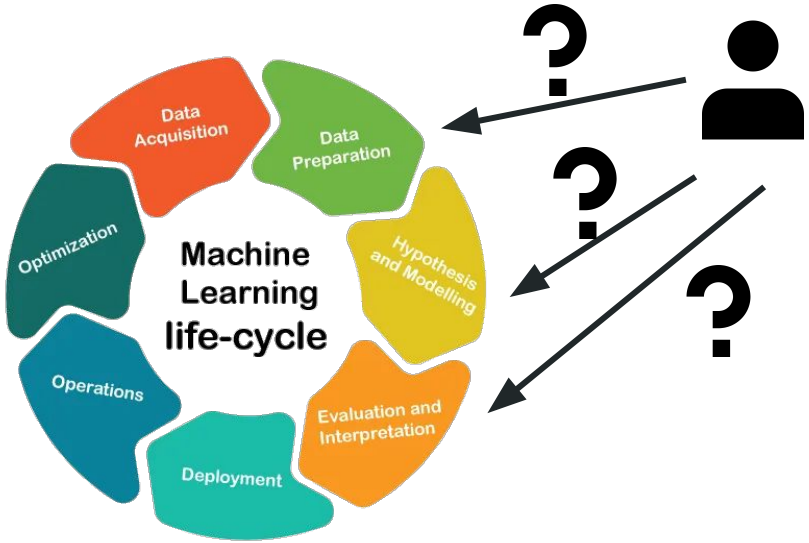
causal explanations: why was a certain algorithm configuration chosen? What if we had decided differently?



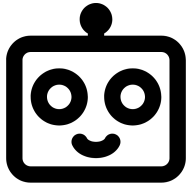
symbolic causal explanation ➔ provides not just the structure but also a symbolic representation of dependencies among decisions

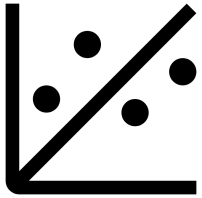
Materials Science can teach us about user needs

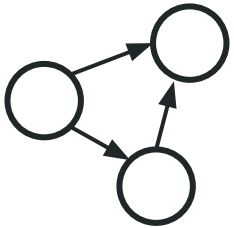
- assumptions made in human-centered AutoML are often based on intuition and common sense, but are not based on empirical findings



Causal AutoML



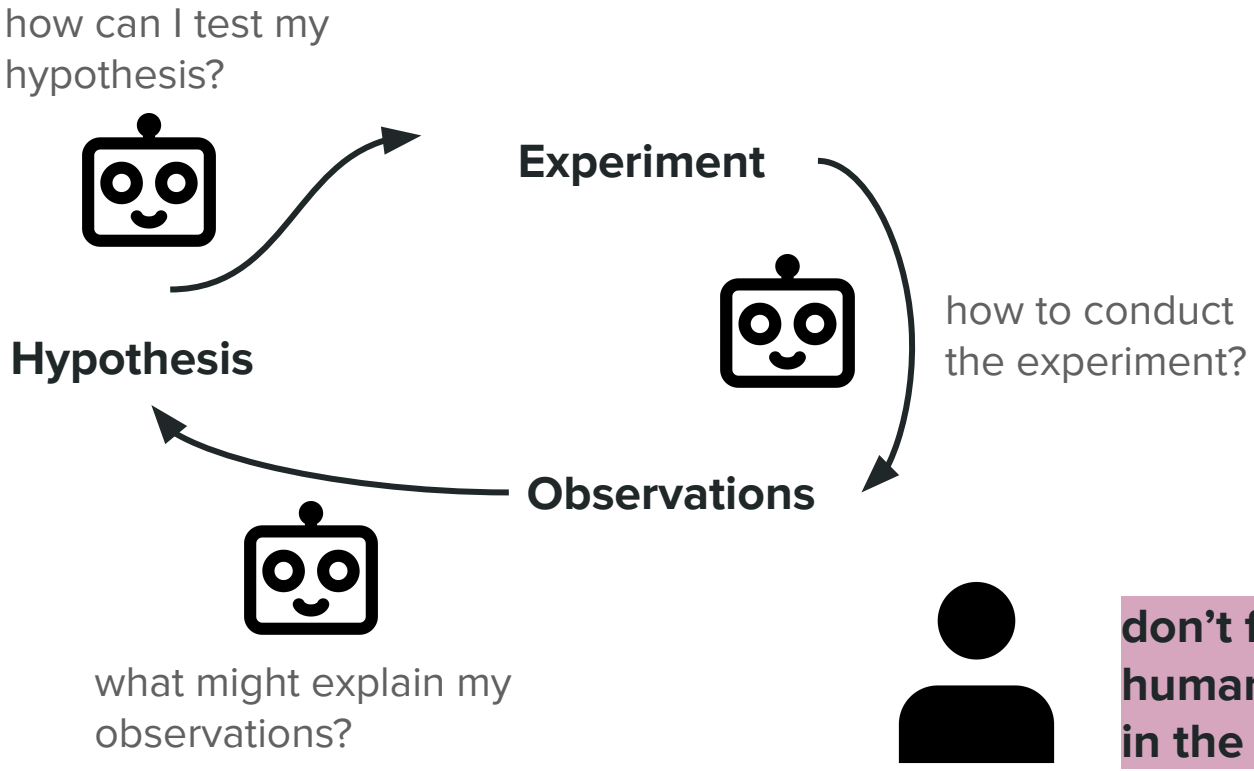




Science is all about causation: Materials Science could help the AutoML community to strive for AutoML systems for causal learning, enabling automatic discovery of cause-effect relations in experimental data!

The Automated Statistician

The Automated ~~Statistician~~ Researcher?



don’t forget the human researcher in the loop!

Questions?