

# Automated Machine Learning (AutoML)

## – A Retrospective Review

Dr. Quanming Yao

Senior Scientist & Leader (machine learning research team)

[yaoquanming@4paradigm.com](mailto:yaoquanming@4paradigm.com) / [gyaoaa@connect.ust.hk](mailto:gyaoaa@connect.ust.hk)


4Paradigm Inc (Hong Kong).

2020/08/24



# Outline

- What is Machine Learning (ML)
- What is Automated Machine Learning (AutoML)
- Is AutoML Really New
- What Should We Focus Next



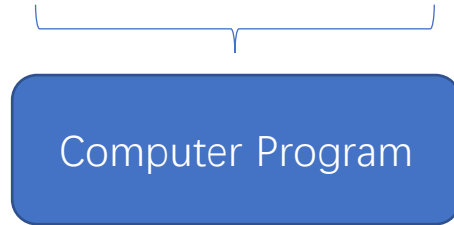
Summary of 1<sup>st</sup> Stage

Proposal of 2<sup>nd</sup> Stage

# What is Machine Learning (ML)?

**Definition 1** [1,2,3]. A computer program is said to learn from experience  $E$  with respect to some classes of task  $T$  and performance measure  $P$  if its performance can improve with  $E$  on  $T$  measured by  $P$ .

Experience Task Measurement



Improve performance



Image Classification

Cat / Dog / Car?

- Experience: Images
- Task: Classification
- Measurement: Accuracy

In short: A computer program specified by  $E$ ,  $T$  and  $P$ .

---

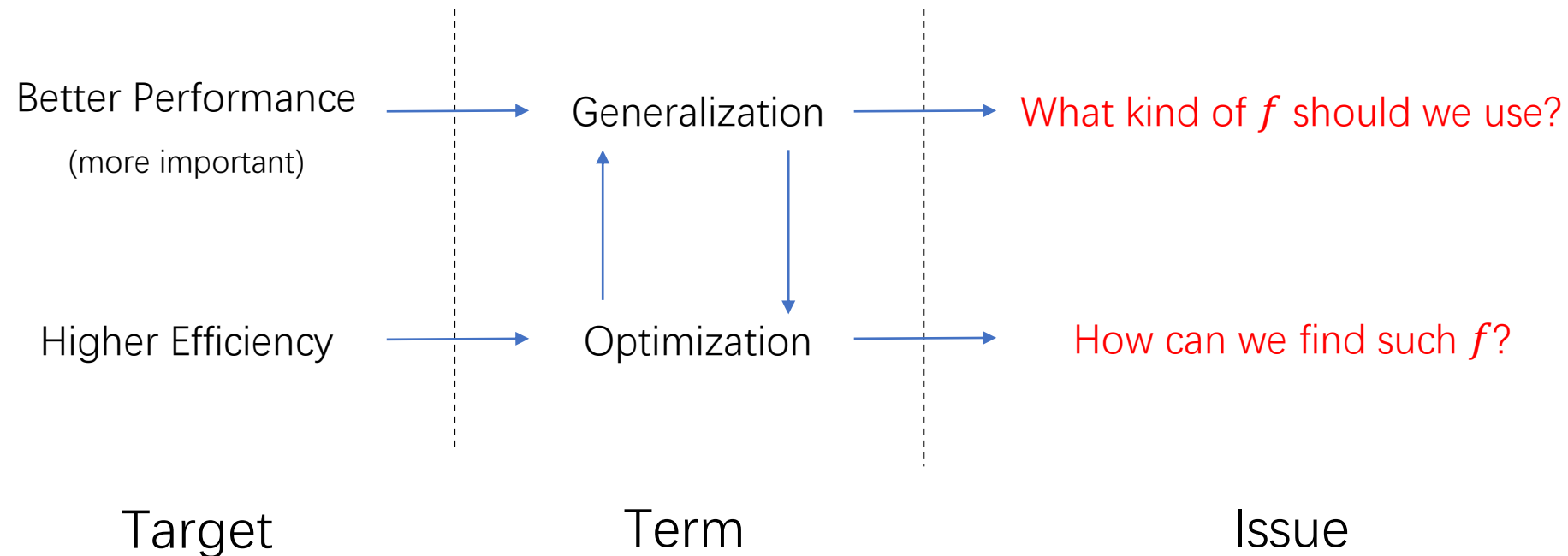
[1]. T. Mitchell. Machine learning. 1997.

[2]. M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning. 2018

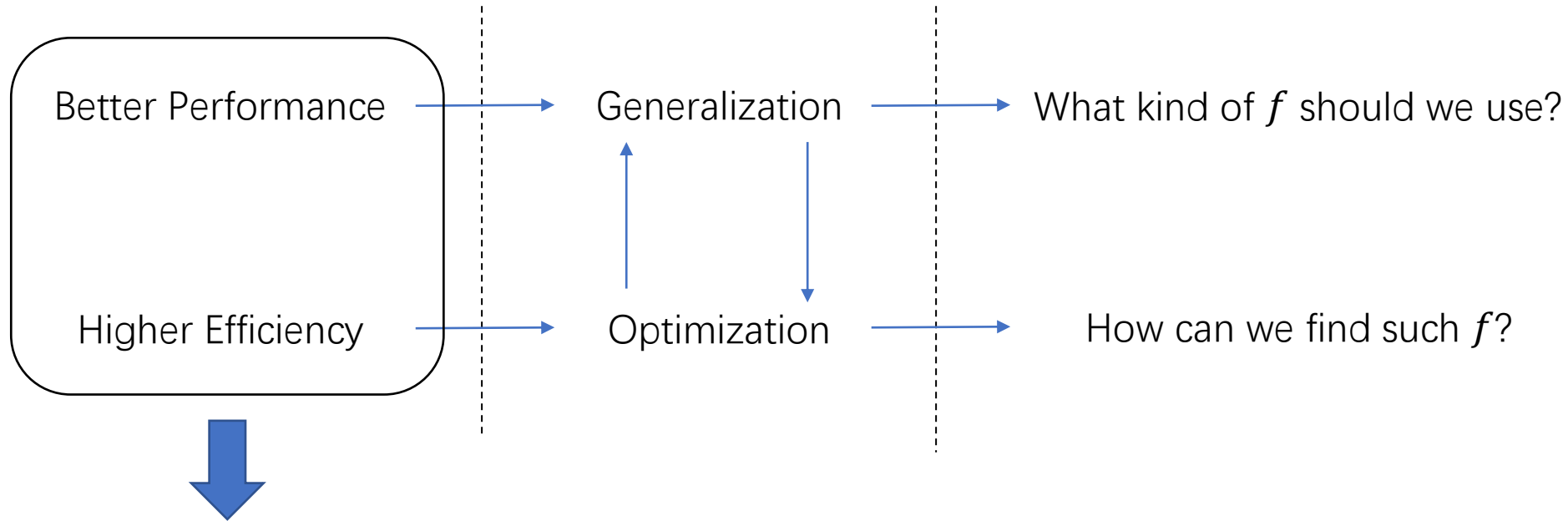
[3]. 周志华. 机器学习. 2016

# What are Core Issues in ML?

Usually, we need to find a hypothesis (function)  $f$  to perform the learning task



# Not Everything can be Learnt



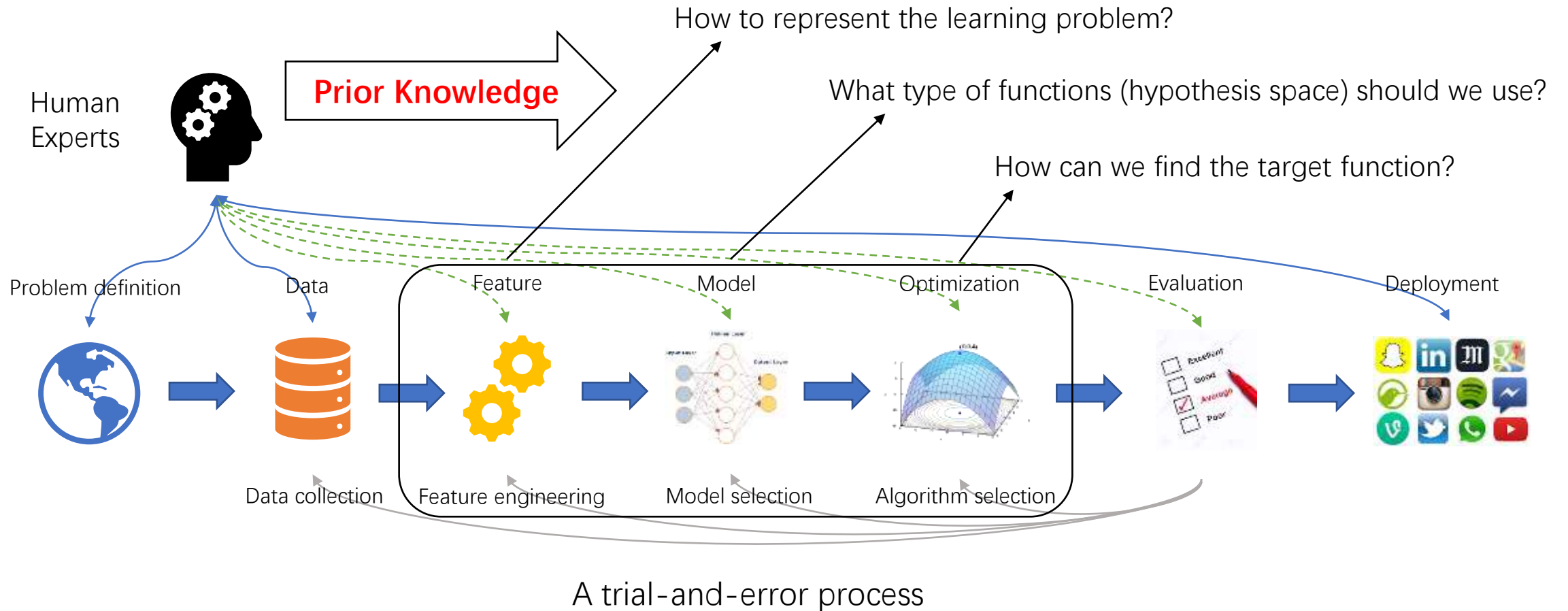
**PAC-Learning** (Definition 2.3 in [1]): What kind of problems can be solved in polynomial time

**No Free Lunch Theorem** (Appendix B [2]): No single algorithm can be good on all problems

[1]. M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning. 2018

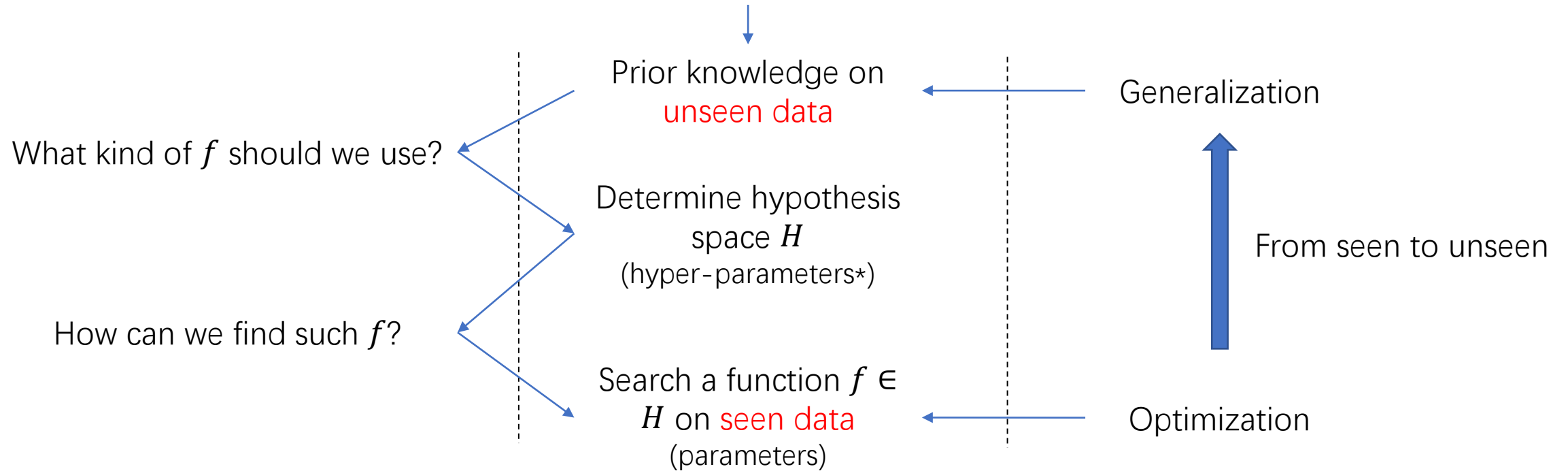
[2]. O. Bousquet, et.al. Introduction to Statistical Learning Theory. 2016

# How to use ML?



# ML = Data + Knowledge

Given a learning problem: human's understanding on target concept  $c \in \mathcal{C}$



\*Hyper-parameters: Free parameters that are not determined by the learning algorithm, but rather specified as inputs to the learning algorithm [Page 4. M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning. 2018]

# Trade-off underneath ML

**Theorem 2.13 (Learning bound — finite  $\mathcal{H}$ , inconsistent case)** Let  $\mathcal{H}$  be a finite hypothesis set. Then, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , the following inequality holds:

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}. \quad (2.20)$$

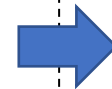
**Definition 2.1 (Generalization error)** Given a hypothesis  $h \in \mathcal{H}$ , a target concept  $c \in \mathcal{C}$ , and an underlying distribution  $\mathcal{D}$ , the generalization error or risk of  $h$  is defined by

$$R(h) = \mathbb{P}_{x \sim \mathcal{D}} [h(x) \neq c(x)] = \mathbb{E}_{x \sim \mathcal{D}} [1_{h(x) \neq c(x)}], \quad (2.1)$$

where  $1_\omega$  is the indicator function of the event  $\omega$ .<sup>2</sup>

**Definition 2.2 (Empirical error)** Given a hypothesis  $h \in \mathcal{H}$ , a target concept  $c \in \mathcal{C}$ , and a sample  $S = (x_1, \dots, x_m)$ , the empirical error or empirical risk of  $h$  is defined by

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m 1_{h(x_i) \neq c(x_i)}. \quad (2.2)$$



## Fundamental Trade-off

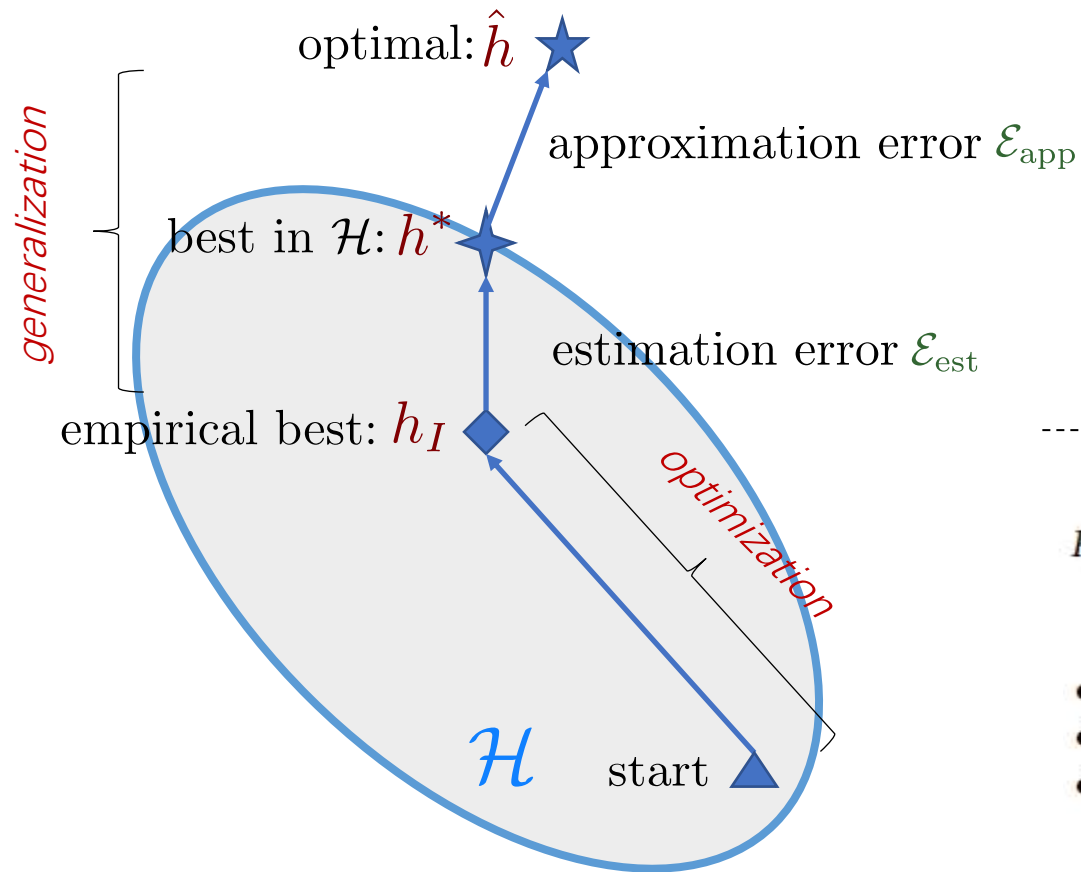
- Independent of distribution and algorithms



- More training samples are always desired
- In order to approximate target concept  $c$ , empirically we prefer using more complex, i.e., larger  $H$



# Error Decomposition in ML



Fundamental error decomposition:

$$\mathbb{E}[R(h_I) - R(\hat{h})] = \underbrace{\mathbb{E}[R(h^*) - R(\hat{h})]}_{\mathcal{E}_{\text{app}}(\mathcal{H})} + \underbrace{\mathbb{E}[R(h_I) - R(h^*)]}_{\mathcal{E}_{\text{est}}(\mathcal{H}, I)},$$

Determine by  
prior **knowledge**

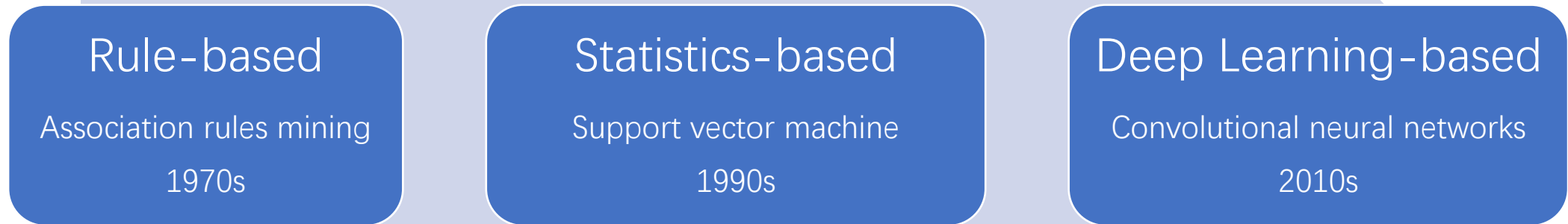
Determine by  
training **data**

$$R(h) = \int \ell(h(x), y) dp(x, y) = \mathbb{E}[\ell(h(x), y)]. \quad R_I(h) = \frac{1}{I} \sum_{i=1}^I \ell(h(x_i), y_i),$$

- $\hat{h} = \arg \min_h R(h)$  be the function that minimizes the expected risk;
- $h^* = \arg \min_{h \in \mathcal{H}} R(h)$  be the function in  $\mathcal{H}$  that minimizes the expected risk;
- $h_I = \arg \min_{h \in \mathcal{H}} R_I(h)$  be the function in  $\mathcal{H}$  that minimizes the empirical risk.

# Recent Trends

Computers being getting more powerful allowing collections of more samples



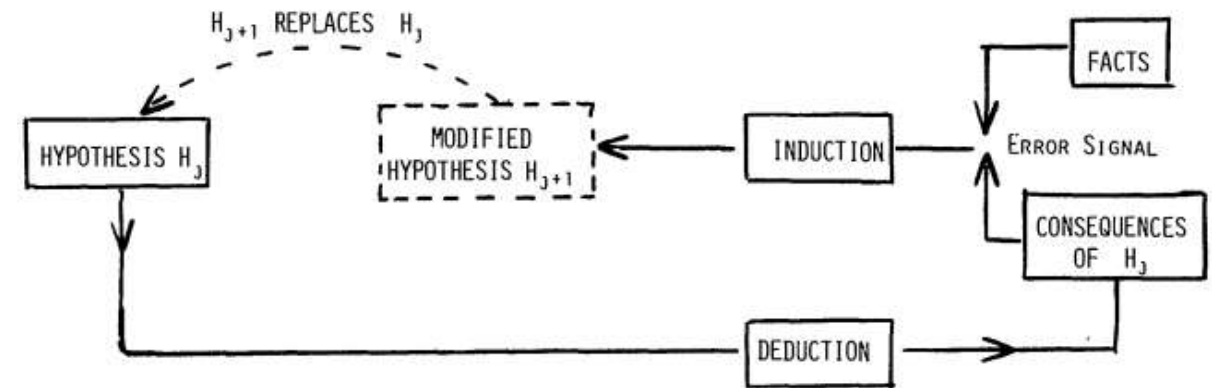
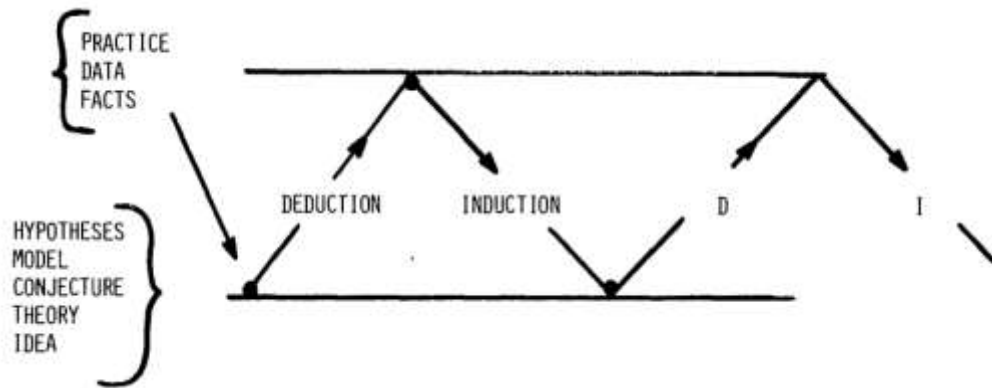
*Better performance*

Trends

- **Larger hypothesis** (more complex models) are being used
- Optimization is getting complex (even mixed up with generalization)
- The prior knowledge is imposed on more abstract level

# Principles underneath Trends

Ronald Fisher  
"a genius who almost single-handedly created the foundations for modern statistical science"



The Advancement of Learning

- Left: an iteration between theory and practice
- Right: a feedback loop

Better hypothesis (better performance) on the real data

*Prior knowledge*



"All models are wrong, but some are useful"[1]

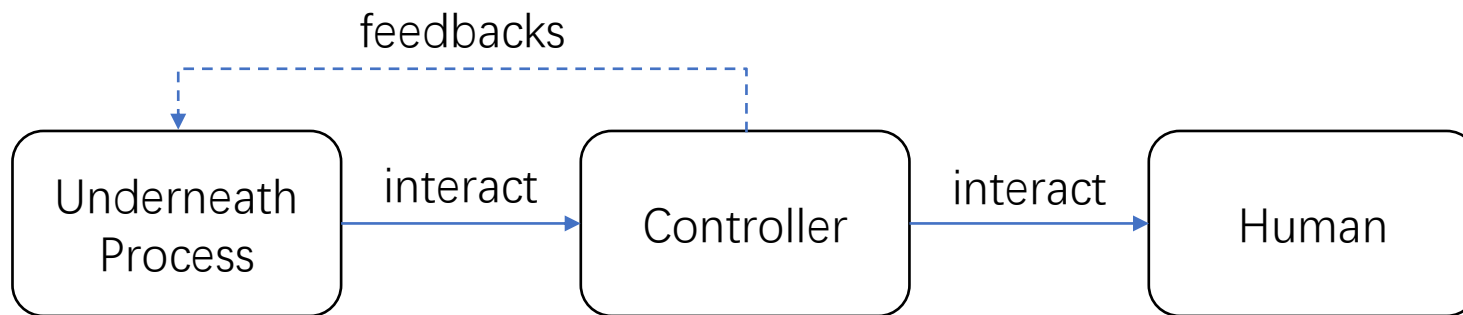
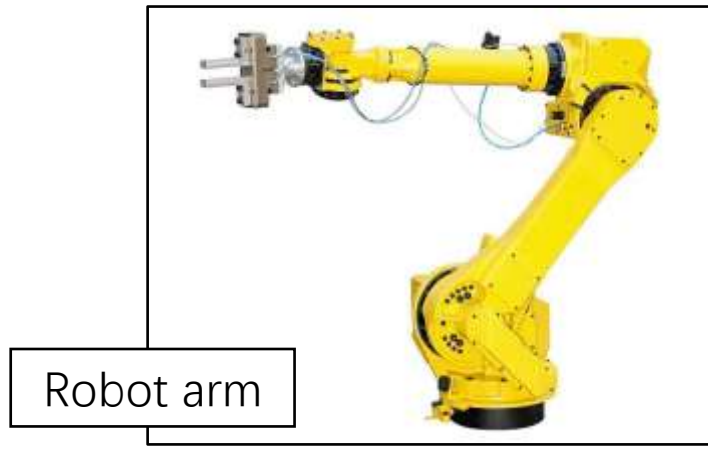
Figures are taken from [1] G. Box, Science and statistics, JASA 1976

# Outline

- What is Machine Learning
- What is Automated Machine Learning (AutoML)
- Is AutoML Really New
- What Should We Focus Next

# What is Automation?

Automation is the technology by which a process or procedure is performed with **minimal human assistance**



Automation (control with feedbacks):

- **Fewer and more understandable** interface exposed to **human**
- The **controller** interacts with underneath process in a **more robust and stable** way

# What can be Automated in ML?

Reduce the usage of humans **(low-level) prior knowledge** in the trial-and-error process of machine learning

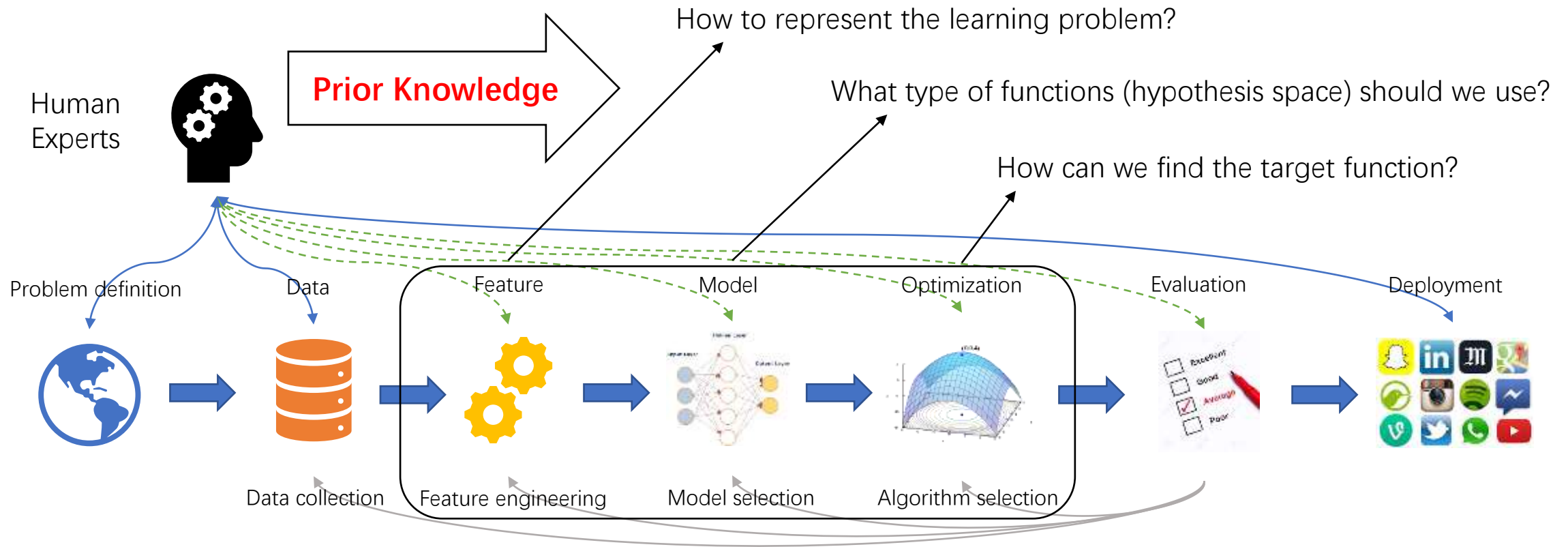
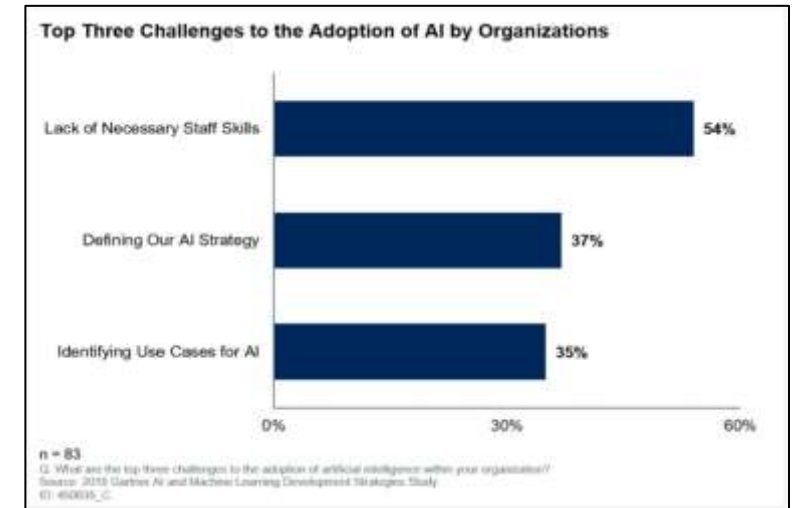
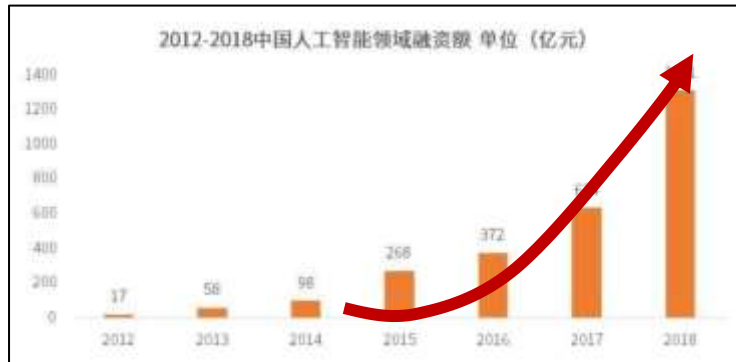


Figure is from "Q. Yao et.al. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv 2019"

# Why We need AutoML?



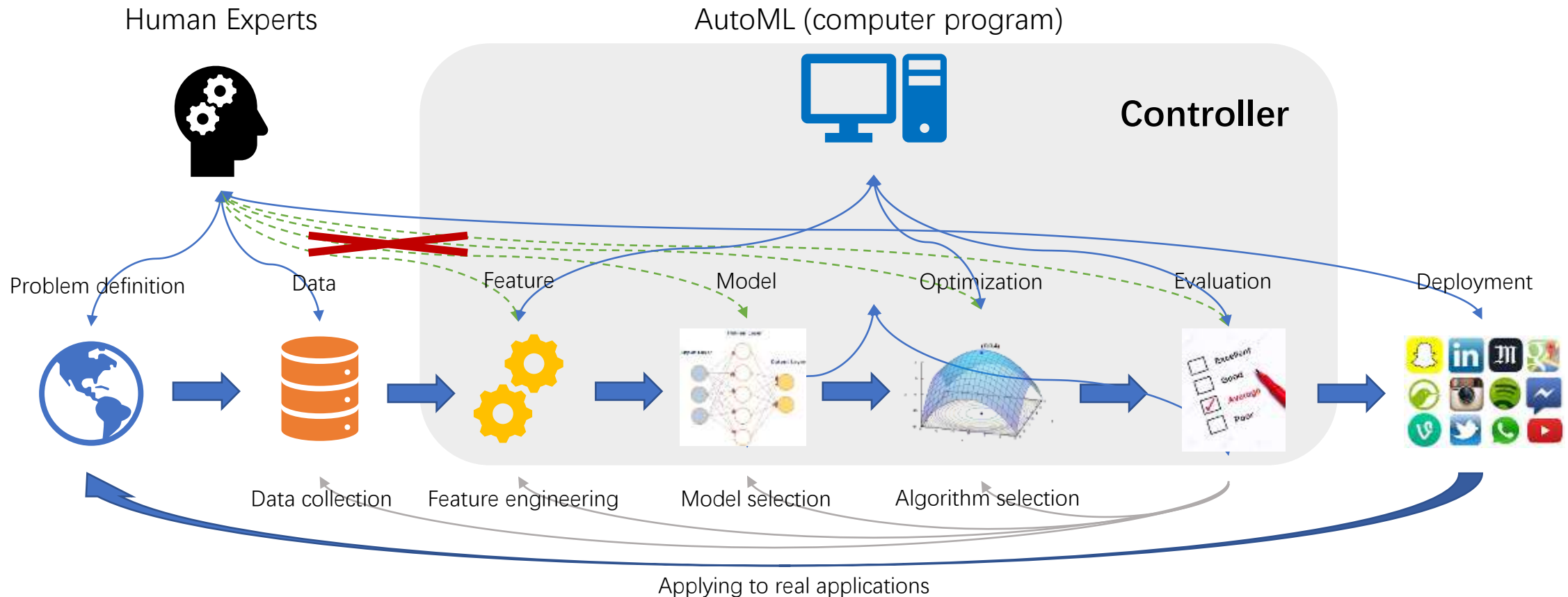
- **Industry** – reduce the expense, increase usage coverage – huge **market value** <sup>[1]</sup>
- **Academy** – understanding data science on a higher level – great **intelligence value** <sup>[2]</sup>

[1]. Gartner: <https://www.forbes.com/sites/janakirammsv/2020/03/02/key-takeaways-from-the-gartner-magic-quadrant-for-ai-developer-services/#a95b99ee3e5e>

[2]. Yoshua Bengio: From System 1 Deep Learning to System 2 Deep Learning | NeurIPS 2019

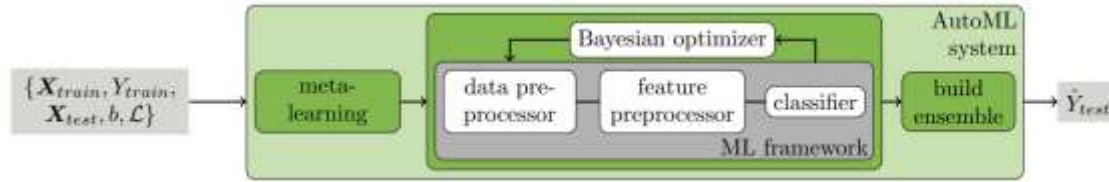
# AutoML – Industrial view

Taking machine learning as a black box – simply its exposed interface





# AutoML – Commercialized examples



AutoSklearn

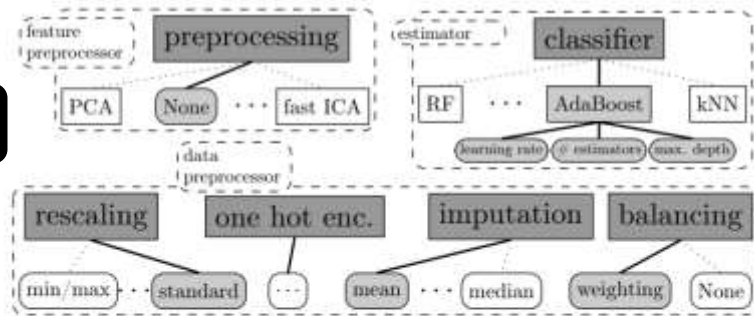
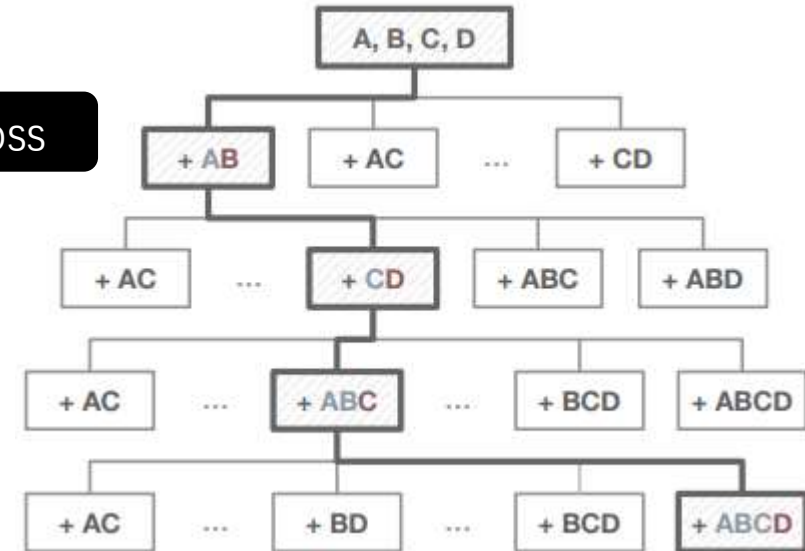


Figure 2: Structured configuration space. Squared boxes denote parent hyperparameters whereas boxes with rounded edges are leaf hyperparameters. Grey colored boxes mark active hyperparameters which form an example configuration and machine learning pipeline. Each pipeline comprises one *feature preprocessor*, *classifier* and up to three *data preprocessor* methods plus respective hyperparameters.

Tuning few hyper-parameters [1]

$$c_{i,j,\dots,k} = \text{vec} (f_i \otimes f_j \otimes \dots \otimes f_k),$$

AutoCross



Cross-product feature generation [2]

[1]. F. Matthias et.al. Efficient and Robust Automated Machine Learning. NIPS 2015

[2]. Y. Luo, et.al. AutoCross: Automatic Feature Crossing for Tabular Data in Real-World Applications. KDD 2019

# AutoML – Commercialized examples

A brief list of AutoML products in the industrials, and “—” indicated no official announcements are found.

	company	AutoML products	customer
public company	Google	Deployed in Google's Cloud	Disney, ZSL, URBN
	Microsoft	Deployed in Azure	—
	IBM	IBM Watson Studio	—
startup	H2O.ai	H2O AutoML Package	AWS, databricks, IBM, NVIDIA
	Feature Labs	Feature Labs' platform	NASA, MONSANTO, MIT, KOHL'S
	4Paradigm	AutoML platform	Bank of China, PICC, Zhihu

Some popular open-source research projects on Github (up to Nov. 2018). More stars indicates greater popularity.

Project	stars	Project	stars
TPOT	4326	hyperopt	2302
autokeras	3728	adanet	1802
H2O AutoML	3262	darts	1547
Auto-sklearn	2367	ENAS-pytorch	1297
MOE	1077	Spearmin	1124



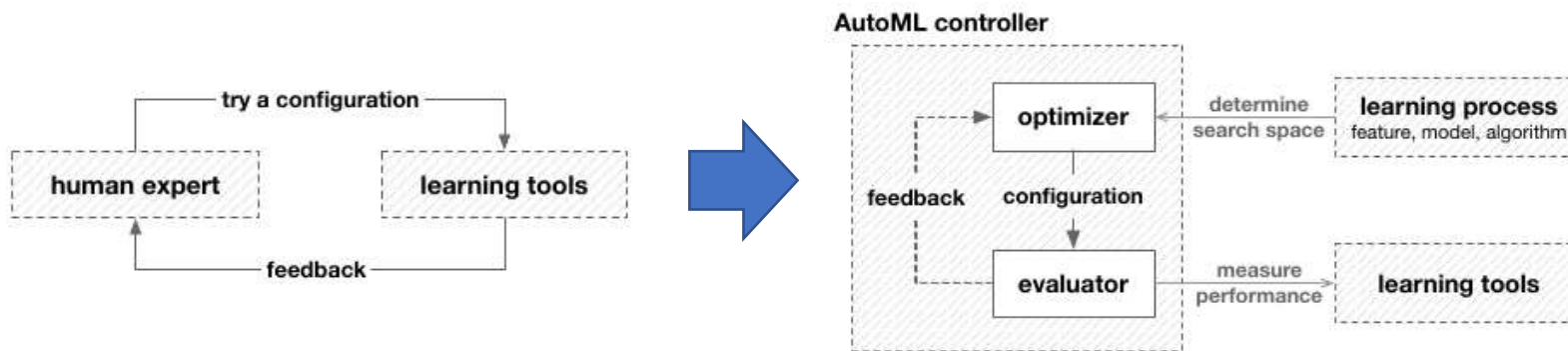
Google's AutoML



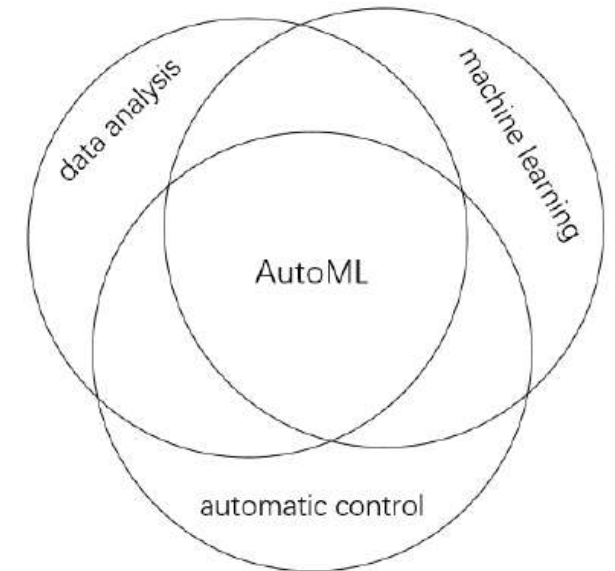
IBM Watson

# Automated ML (AutoML) – Definition

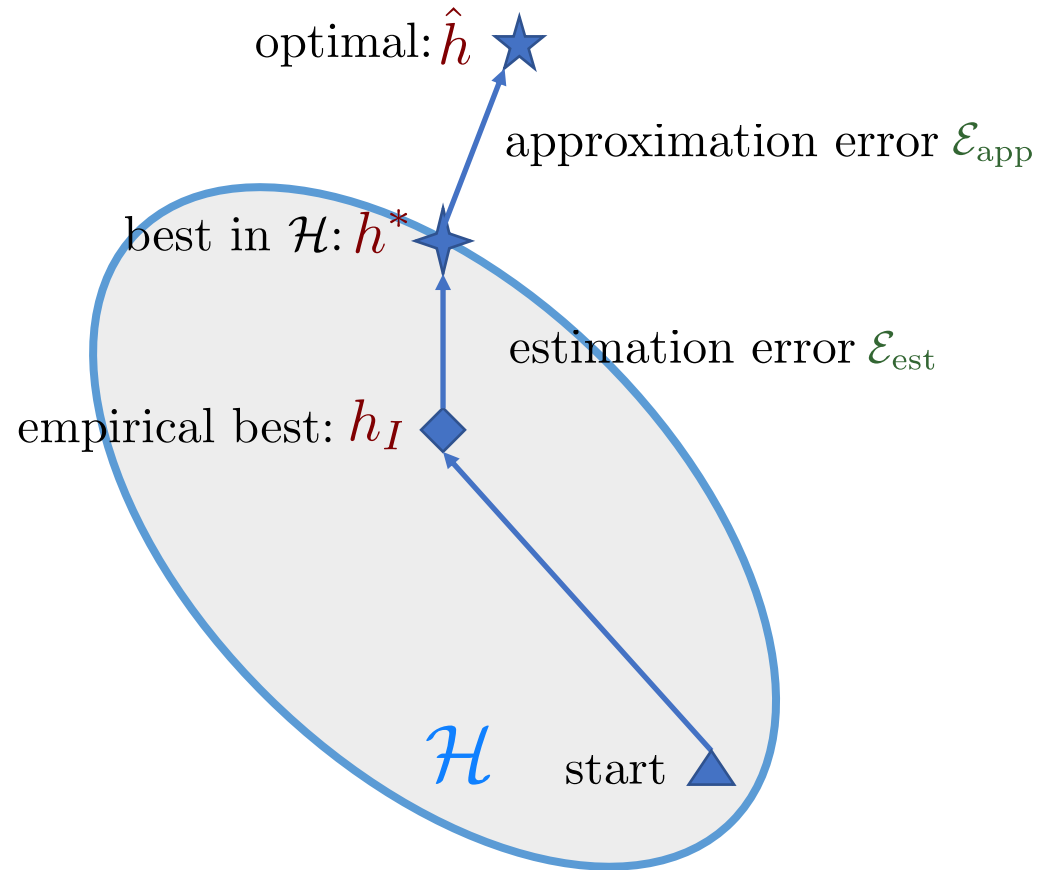
**Definition 2.** AutoML attempts to minimize the assistance from human on designing proper machine learning computer programs (specified by  $E$ ,  $T$  and  $P$  in Definition 1) which can satisfy certain requirements.



	classical machine learning	AutoML
feature engineering	humans design and construct features from data	automated by the computer program
	humans process features making them more informative	
model selection	humans design or pick up some machine learning tools based on professional knowledge	
	humans adjust hyper-parameters of machine learning tools based on performance evaluation	
algorithm selection	humans pick up some optimization algorithms to find parameters	
summary	human are involved in every aspect of learning applications	the program can be directly reused on other learning problems



# AutoML – Academic view



AutoML **directly** minimizes the **total error**

- Approximation error
  - Which classifier to be used
  - What are their hyper-parameters
- Estimation error
  - How to represent your training data
- Optimization error
  - Which algorithm to be used
  - How to tune its step-size

# AutoML – Academic view

Parameterized the prior knowledge that help meet needs of ML programs, e.g.,

- minimize the total error
- reduce parameter numbers

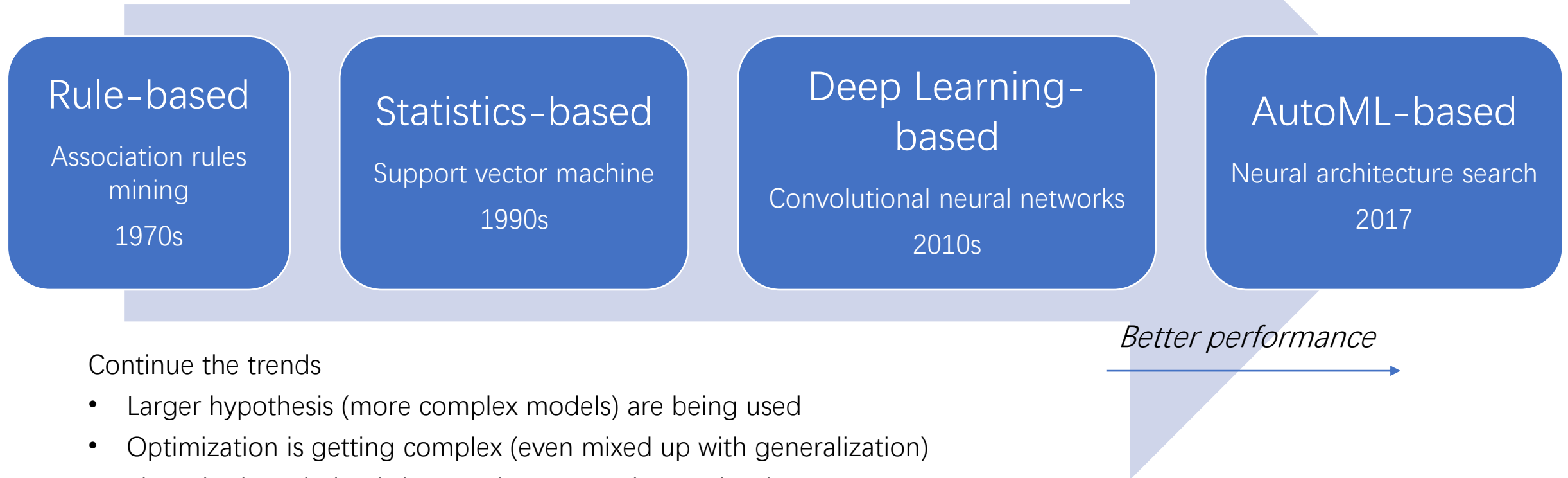
Perform efficient search in the designed (new) space





# AutoML – Successor of ML's trend

- Core Issue in Machine Learning: Improving learning performance (with higher efficiency)
- AutoML: an evolving way to improve learning performance

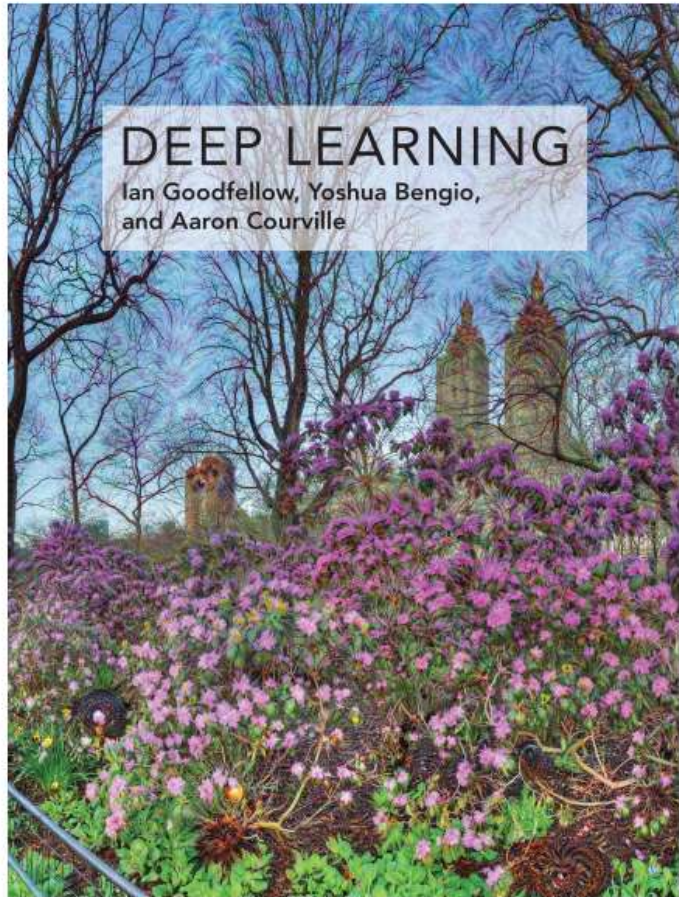


Continue the trends

- Larger hypothesis (more complex models) are being used
- Optimization is getting complex (even mixed up with generalization)
- The prior knowledge is imposed on more abstract level

**Low-level human knowledge on data / model are replacing by computation power**

# AutoML – Successor of ML's trend



## DEEP LEARNING FOR SYSTEM 2 PROCESSING YOSHUA BENGIO

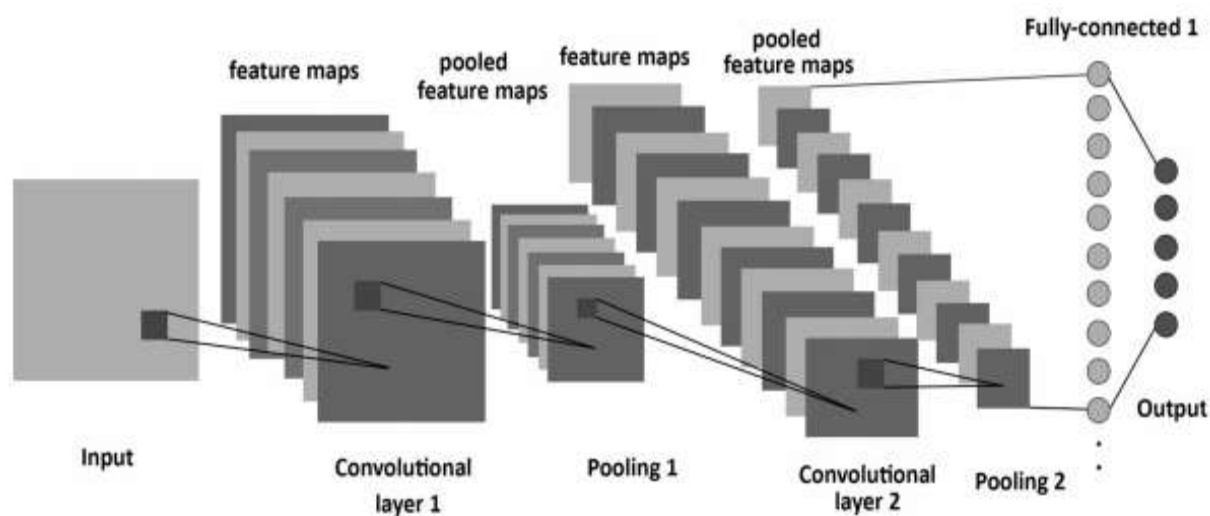
AAAI'2019 Invited Talk  
February 9th, 2020, New York City



Parameterized  
prior knowledge  
on a higher level

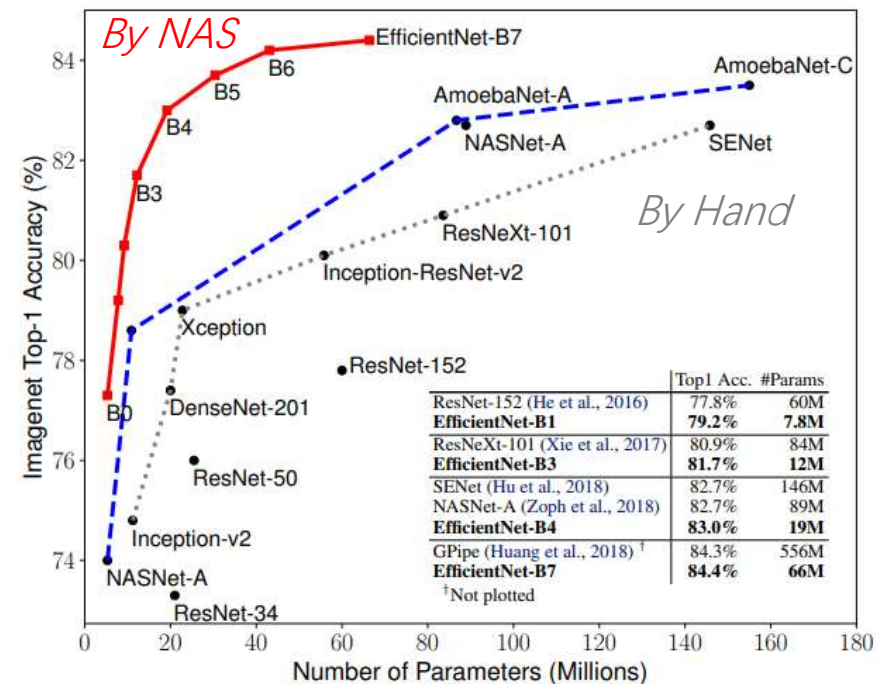
# AutoML – Research examples

Architecture of networks are **critical** to deep learning's performance but **hard to fine-tune**



Design choice in each layer

- number of filters
- filter height
- filter width
- stride height
- stride width
- skip connections



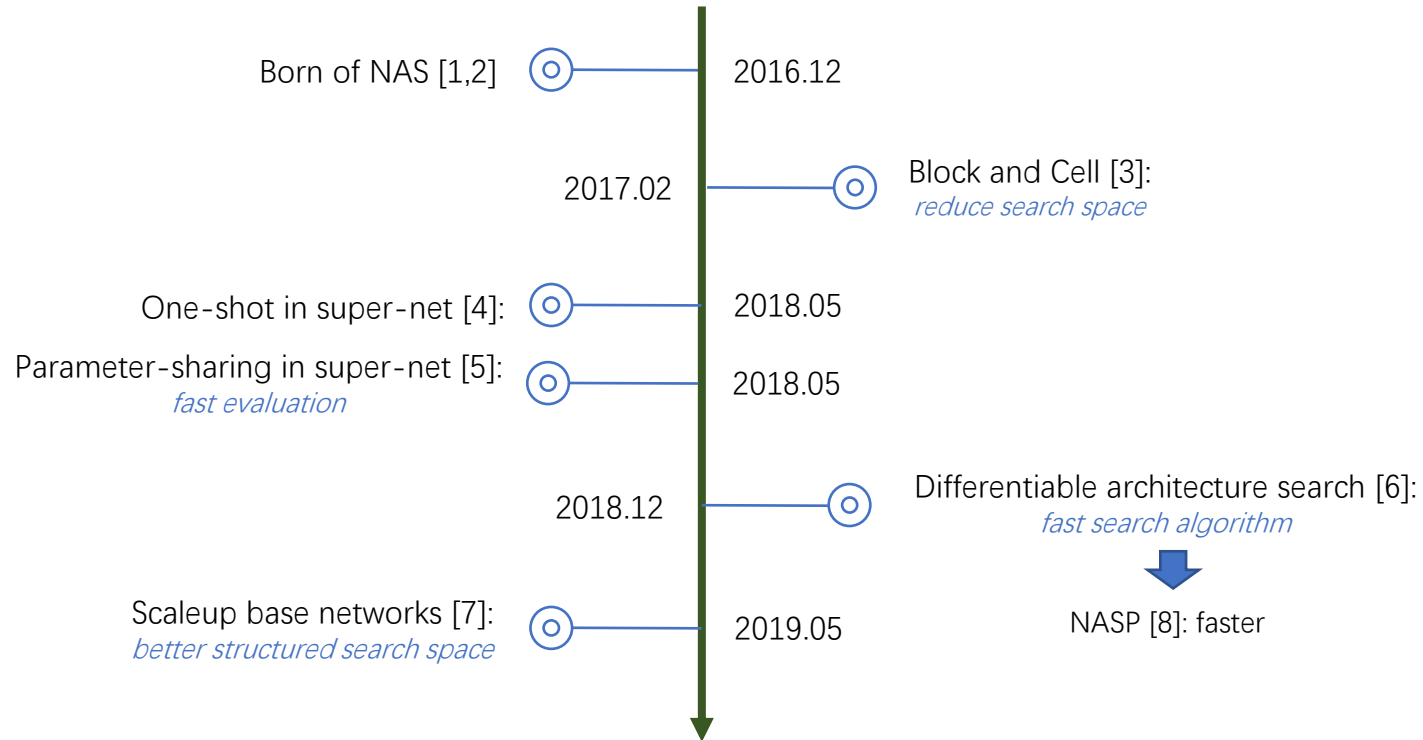
Much better than hand-designed ones

Neural Architecture Search (**NAS**) tries to directly **optimize network architecture using validation data sets**

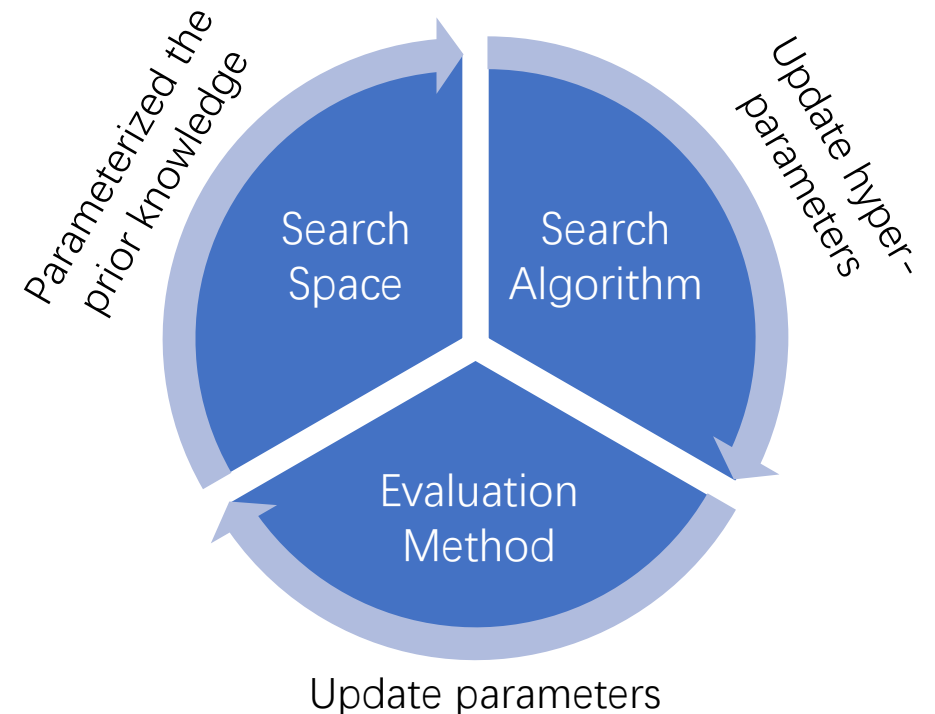
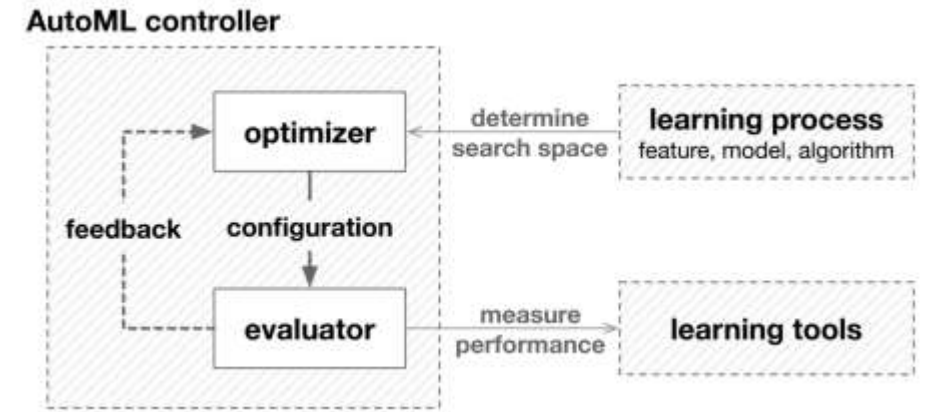


# NAS – Brief review

Core issue: *effectiveness v.s. efficiency*

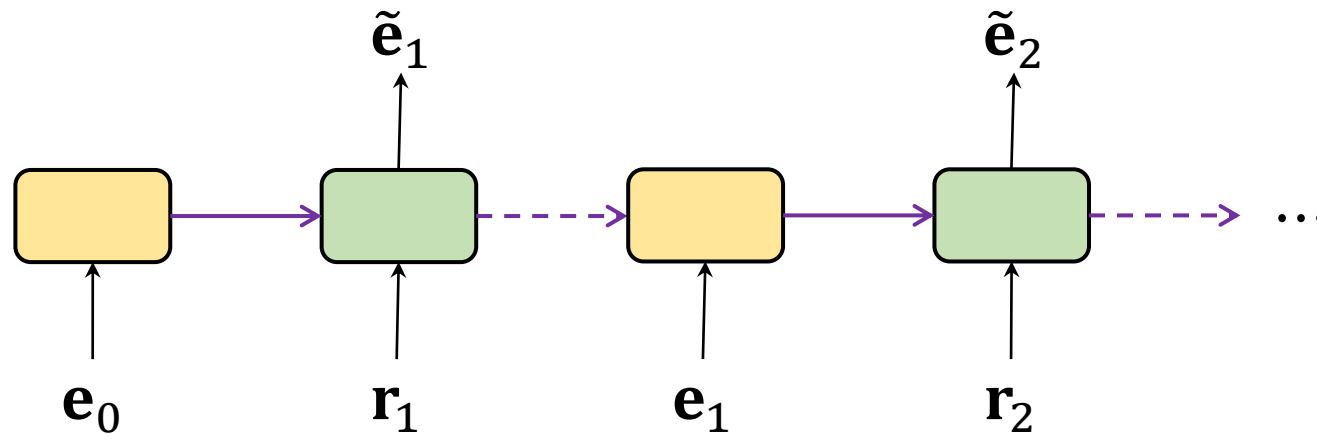


- [1] Neural architecture search with reinforcement learning. ICLR 2017 (1785 cites)
- [2] Designing neural network architectures using reinforcement learning. ICLR 2017 (599 cites)
- [3] Learning transferable architectures for scalable image recognition. CVPR 2017 (1736 cites)
- [4] Efficient Neural Architecture Search via Parameter Sharing. ICML 2018 (785 cites)
- [5] Understanding and Simplifying One-Shot Architecture Search. ICML 2018 (206 cites)
- [6] DARTS: Differentiable Architecture Search. ICLR 2019 (820 cites)
- [7] EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ICML 2019 (683 cites)
- [8] Efficient Neural Architecture Search via Proximal Iterations. AAAI 2020 (16 cites)

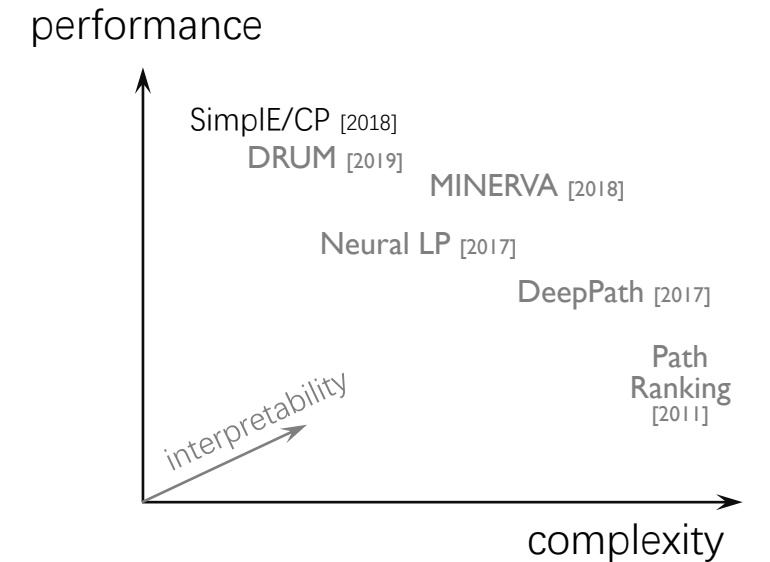
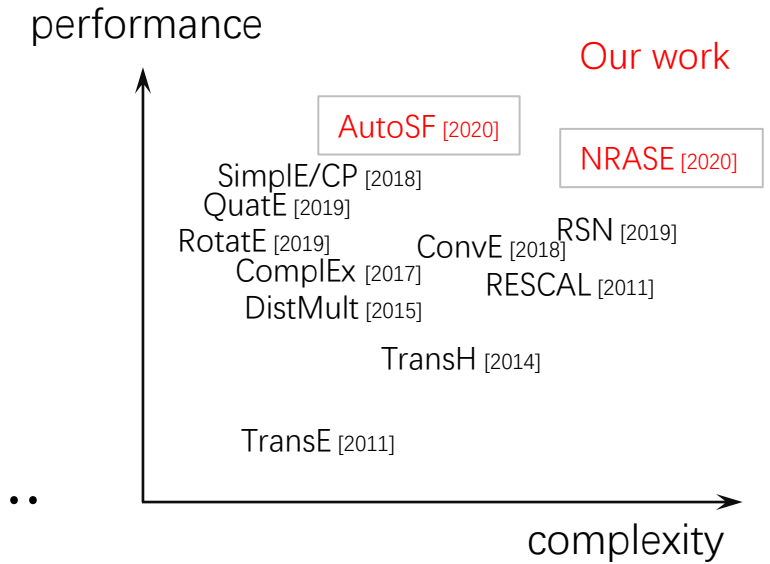
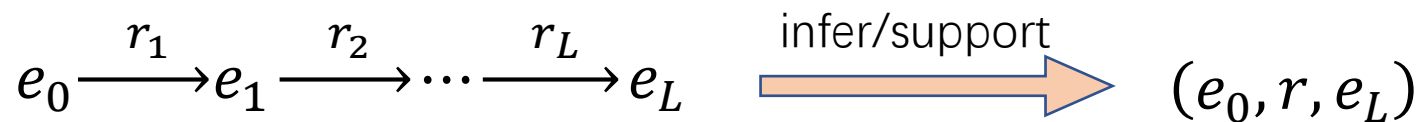


# AutoML – Usage in KG

- Knowledge Graph Embedding learning



- Knowledge Graph Rule learning



# Outline

- What is Machine Learning
- What is Automated Machine Learning (AutoML)
- Is AutoML Really New
- What Should We Focus Next

# Is AutoML really New? – No

## Feature generation

[\[PDF\] Genetic Algorithms as a Tool for Feature Selection in Machine Learning](#)

[H Vafaie, KA De Jong](#) - ICTAI, 1992 - [researchgate.net](#)

This paper describes an approach being explored to improve the usefulness of machine learning techniques for generating classification rules for complex, real world data. The approach involves the use of genetic algorithms as a "front end" to traditional rule induction ...

☆ [🔗](#) Cited by 288 [Related articles](#) [All 15 versions](#) [🔗](#)

## Bi-level optimization

[An overview of bilevel optimization](#)

[B Colson, P Marcotte, G Savard](#) - Annals of operations research, 2007 - Springer

This paper is devoted to **bilevel optimization**, a branch of mathematical programming of both practical and theoretical interest. Starting with a simple example, we proceed towards a general formulation. We then present fields of application, focus on solution approaches ...

☆ [🔗](#) Cited by 1034 [Related articles](#) [All 19 versions](#)

## Model selection

[\[BOOK\] Model selection.](#)

[H Linhart, W Zucchini](#) - 1986 - [psycnet.apa.org](#)

**Model selection.** Citation. Linhart, H., & Zucchini, W. (1986). Wiley series in probability and mathematical statistics. **Model selection.** Oxford, England: John Wiley & Sons. Abstract. This book describes a systematic way of selecting between competing statistical models ...

☆ [🔗](#) Cited by 999 [Related articles](#) [All 3 versions](#)

## Meta-learning

[A perspective view and survey of meta-learning](#)

[R Vilalta, Y Drissi](#) - Artificial intelligence review, 2002 - Springer

Different researchers hold different views of what the term **meta-learning** exactly means. The first part of this paper provides our own perspective view in which the goal is to build self-adaptive learners (ie learning algorithms that improve their bias dynamically through ...

☆ [🔗](#) Cited by 658 [Related articles](#) [All 22 versions](#)

## Hyper-parameter optimization

[Gradient-based optimization of hyperparameters](#)

[Y Bengio](#) - Neural computation, 2000 - MIT Press

Many machine learning algorithms can be formulated as the minimization of a training criterion that involves a hyperparameter. This hyperparameter is usually chosen by trial and error with a model selection criterion. In this article we present a methodology to optimize several hyper-parameters, based on the computation of the gradient of a model selection criterion with respect to the hyperparameters. In the case of a quadratic training criterion, the gradient of the selection criterion with respect to the hyperparameters is efficiently computed ...

☆ [🔗](#) Cited by 265 [Related articles](#) [All 15 versions](#)

## Neural architecture search

[Constructive algorithms for structure learning in feedforward neural network regression problems](#)

[TY Kwok, DY Yeung](#) - IEEE transactions on neural networks, 1997 - [ieeexplore.ieee.org](#)

In this survey paper, we review the constructive algorithms for structure learning in feedforward neural networks for regression problems. The basic idea is to start with a small network, then add hidden units and weights incrementally until a satisfactory solution is found. By formulating the whole problem as a state-space search, we first describe the general issues in constructive algorithms, with special emphasis on the search strategy. A taxonomy, based on the differences in the state transition mapping, the training algorithm ...

☆ [🔗](#) Cited by 588 [Related articles](#) [All 21 versions](#)

# AutoML v.s. Meta-Learning – Examples

**Definition 2 (CASH).** Let  $\mathcal{A} = \{A^{(1)}, \dots, A^{(K)}\}$  be a set of algorithms, and let the hyperparameters of each algorithm  $A^{(j)}$  have domain  $\Lambda^{(j)}$ . Further, let  $D_{\text{train}} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  be a training set which is split into  $K$  cross-validation folds  $\{D_{\text{valid}}^{(1)}, \dots, D_{\text{valid}}^{(K)}\}$  and  $\{D_{\text{train}}^{(1)}, \dots, D_{\text{train}}^{(K)}\}$  such that  $D_{\text{train}}^{(i)} = D_{\text{train}} \setminus D_{\text{valid}}^{(i)}$  for  $i = 1, \dots, K$ . Finally, let  $\mathcal{L}(A_{\lambda}^{(j)}, D_{\text{train}}^{(i)}, D_{\text{valid}}^{(i)})$  denote the loss that algorithm  $A^{(j)}$  achieves on  $D_{\text{valid}}^{(i)}$  when trained on  $D_{\text{train}}^{(i)}$  with hyperparameters  $\lambda$ . Then, the Combined Algorithm Selection and Hyperparameter optimization (CASH) problem is to find the joint algorithm and hyperparameter setting that minimizes this loss:

$$A^*, \lambda_* \in \underset{A^{(j)} \in \mathcal{A}, \lambda \in \Lambda^{(j)}}{\operatorname{argmin}} \frac{1}{K} \sum_{i=1}^K \mathcal{L}(A_{\lambda}^{(j)}, D_{\text{train}}^{(i)}, D_{\text{valid}}^{(i)}). \quad (1)$$

Auto-sklearn [1]

This implies a bilevel optimization problem (Anandalingam & Friesz, 1992; Colson et al., 2007) with  $\alpha$  as the upper-level variable and  $w$  as the lower-level variable:

$$\min_{\alpha} \mathcal{L}_{\text{val}}(w^*(\alpha), \alpha) \quad (3)$$

$$\text{s.t. } w^*(\alpha) = \underset{w}{\operatorname{argmin}} \mathcal{L}_{\text{train}}(w, \alpha) \quad (4)$$

**Definition 1 (AutoML Problem).** Let  $F(P; g)$  be a KGE model (with indexed embeddings  $P = \{h, r, t\}$  and structure  $g$ ).  $\mathcal{M}(F(P; g), S)$  measures the performance (the higher the better) of a KGE model  $F$  on a set of triplets  $S$ . The problem of searching the SF is formulated as:

$$g^* \in \arg \max_{g \in \mathcal{G}} \mathcal{M}(F(P^*; g), S_{\text{val}}) \quad (1)$$

$$\text{s.t. } P^* = \arg \max_P \mathcal{M}(F(P; g), S_{\text{tra}}), \quad (2)$$

where  $\mathcal{G}$  contains all possible choices of  $g$ ,  $S_{\text{tra}}$  and  $S_{\text{val}}$  denote training and validation data sets.

AutoSF [3]

In standard training, we aim to minimize the expected loss for the training set:  $\frac{1}{N} \sum_{i=1}^N C(\hat{y}_i, y_i) = \frac{1}{N} \sum_{i=1}^N f_i(\theta)$ , where each input example is weighted equally, and  $f_i(\theta)$  stands for the loss function associating with data  $x_i$ . Here we aim to learn a reweighting of the inputs, where we minimize a weighted loss:

$$\theta^*(w) = \arg \min_{\theta} \sum_{i=1}^N w_i f_i(\theta), \quad (1)$$

with  $w_i$  unknown upon beginning. Note that  $\{w_i\}_{i=1}^N$  can be understood as training hyperparameters, and the optimal selection of  $w$  is based on its validation performance:

$$w^* = \arg \min_{w, w \geq 0} \frac{1}{M} \sum_{i=1}^M f_i^v(\theta^*(w)). \quad (2)$$

It is necessary that  $w_i \geq 0$  for all  $i$ , since minimizing the negative training loss can usually result in unstable behavior.

Noisy Label Learning [4]

1. Efficient and robust automated machine learning. NIPS 2015
  2. DARTS: Differentiable Architecture Search. ICLR 2019
  3. AutoSF: Searching Scoring Functions for Knowledge Graph Embedding. ICDE 2020
  4. Learning to Reweight Examples for Robust Deep Learning. ICML 2018
- See more in "Bilevel programming for hyperparameter optimization and meta-learning. ICML 2018"



# What's Meta-Learning?

In the 1990s, the term metalearning started to appear in machine learning research, although the concept itself dates back to the mid-1970s (Rice 1976). A number of definitions of metalearning have been given, the following list cites the main review papers and books from the last decade:

1. Metalearning studies how learning systems can increase in efficiency through experience; the goal is to understand how learning itself can become flexible according to the domain or task under study (Vilalta and Drissi 2002a).
2. The primary goal of metalearning is the understanding of the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable (Giraud-Carrier 2008).
3. Metalearning is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning and data mining processes (Brazdil et al. 2009).
4. Metalearning monitors the automatic learning process itself, in the context of the learning problems it encounters, and tries to adapt its behaviour to perform better (Vanschoren 2010).

- Definition 1**
1. A metalearning system must include a learning subsystem, which adapts with experience.
  2. Experience is gained by exploiting metaknowledge extracted
    - (a) ...in a previous learning episode on a single dataset, and/or
    - (b) ...from different domains or problems.

## DARTS

This implies a bilevel optimization problem (Anandalingam & Friesz, 1992; Colson et al., 2007) with  $\alpha$  as the upper-level variable and  $w$  as the lower-level variable:

$$\min_{\alpha} \mathcal{L}_{val}(w^*(\alpha), \alpha) \quad (3)$$

$$\text{s.t. } w^*(\alpha) = \operatorname{argmin}_w \mathcal{L}_{train}(w, \alpha) \quad (4)$$

Meta-learner

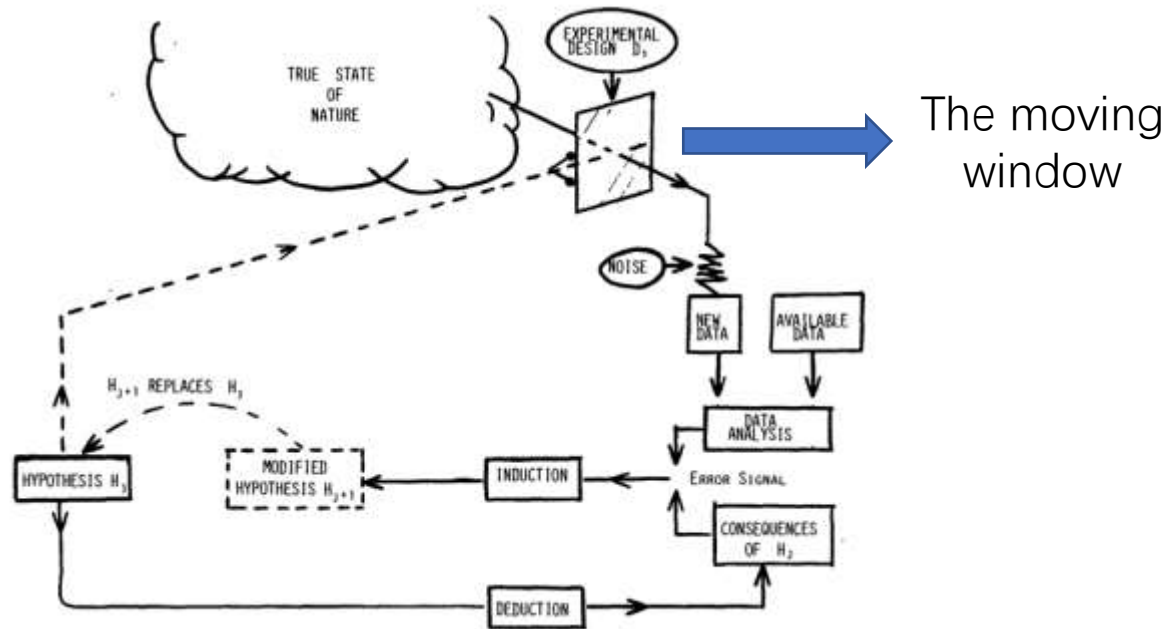
Base-learner

Just an application of meta-learning?

Model selection?

Bilevel programming?

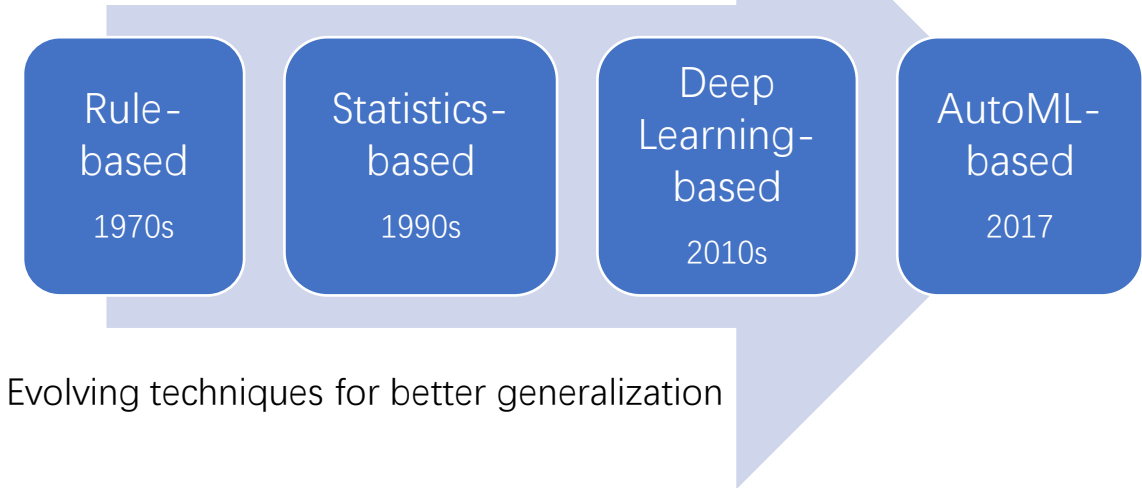
# Is AutoML New? – Yes!



The experimental design is here shown as a movable window looking onto the true state of nature. Its positioning at each stage is motivated by **current beliefs, hopes, and fears** [1]

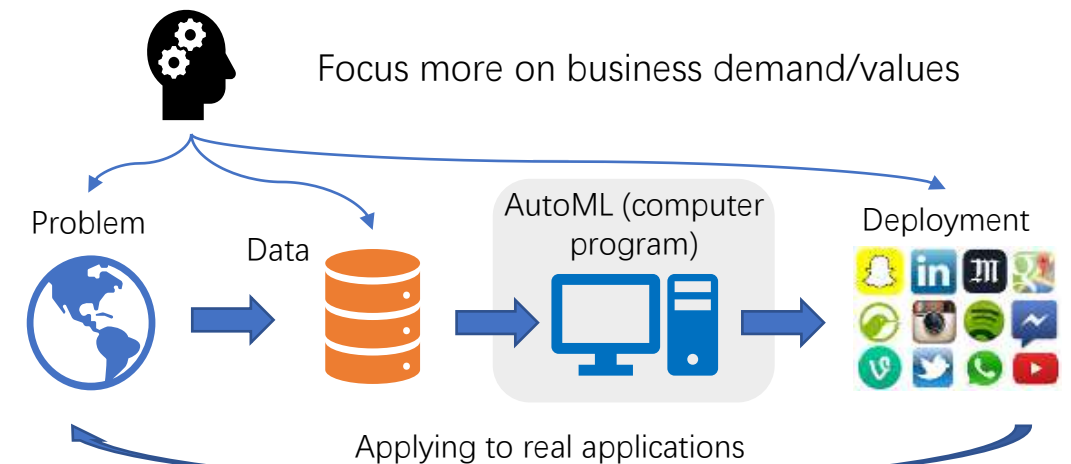
[1] G. Box, Science and statistics, JASA 1976

## Academy needs AutoML



Evolving techniques for better generalization

## Industry needs AutoML



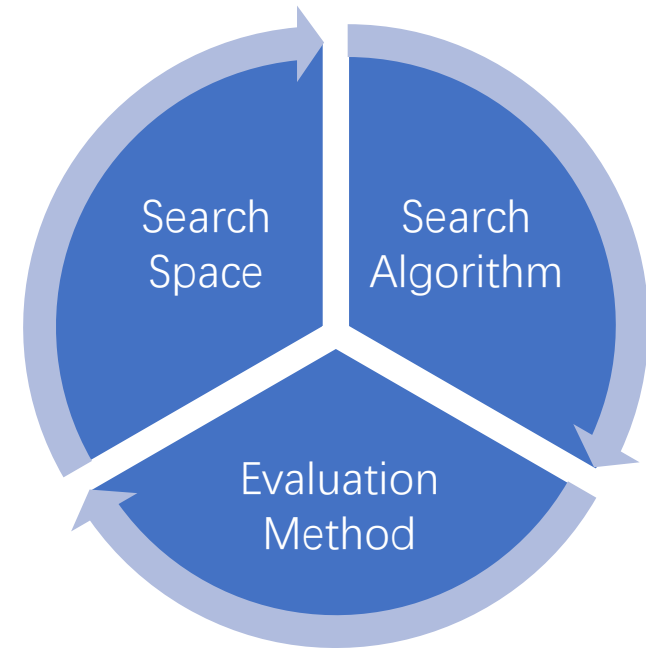
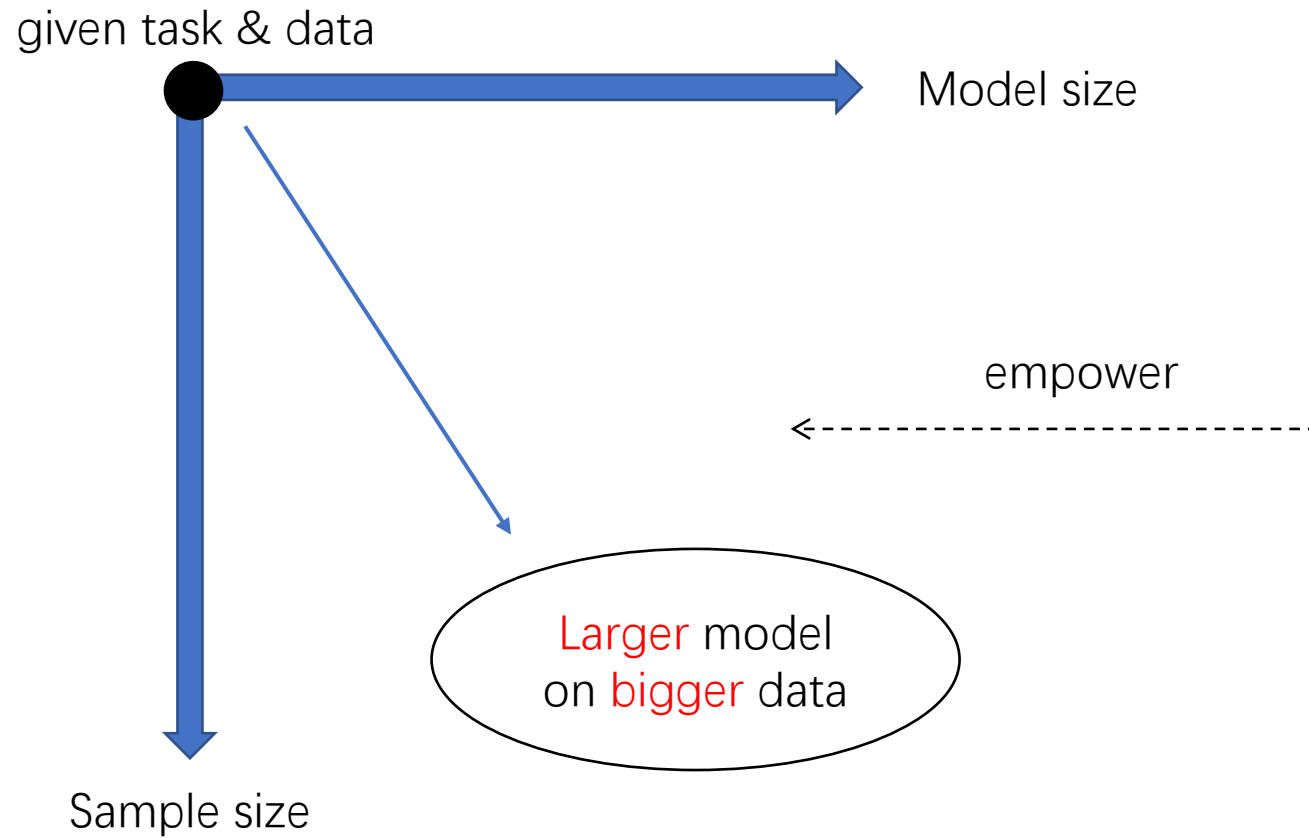
# Outline

- What is Machine Learning
- What is Automated Machine Learning (AutoML)
- Is AutoML Really New
- What Should We Focus Next

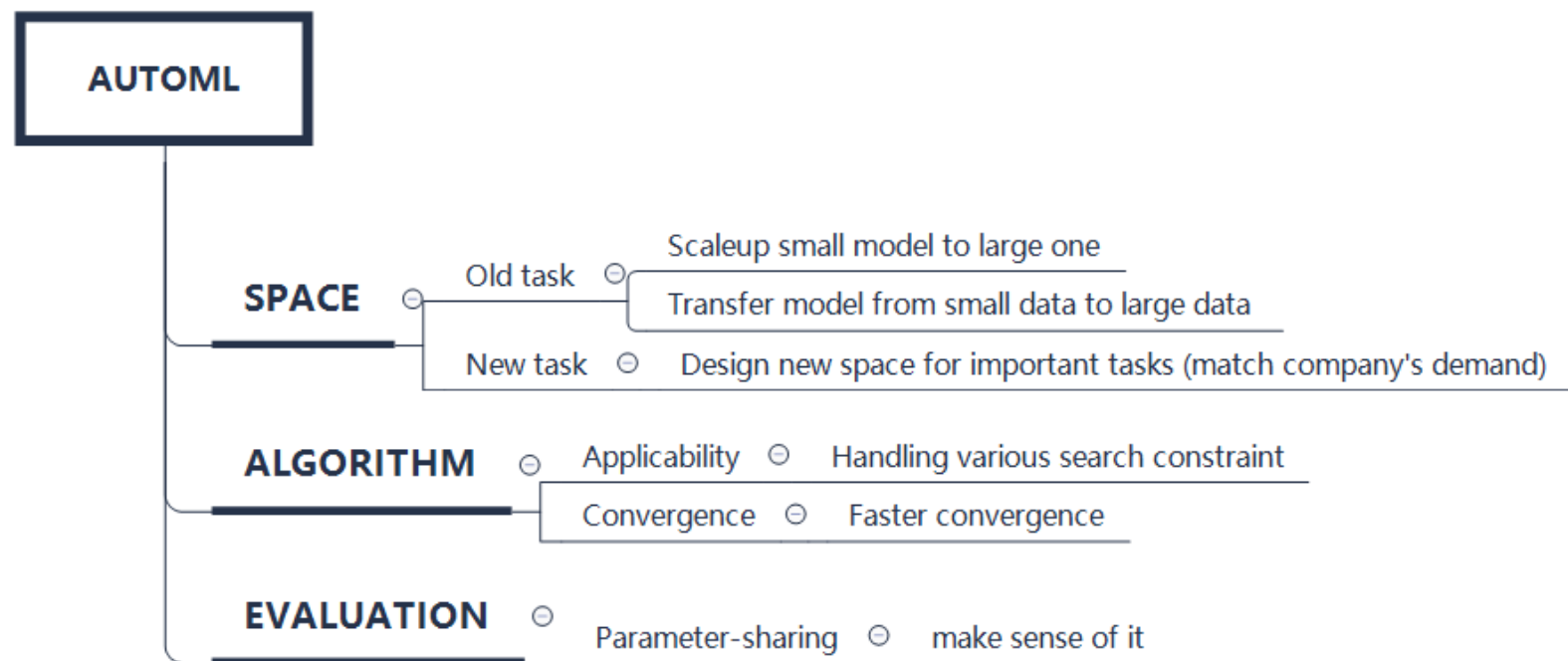
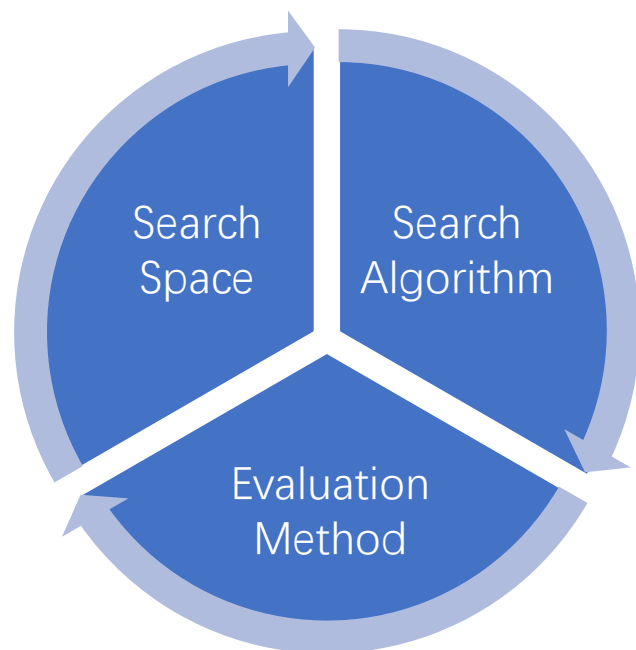
Proposal of 2<sup>nd</sup> Stage



# AutoML – Research landscape



# AutoML – Next focus



# Thanks!

yaoquanming@4paradigm.com / qyaoaa@connect.ust.hk