

AutoML: Introduction

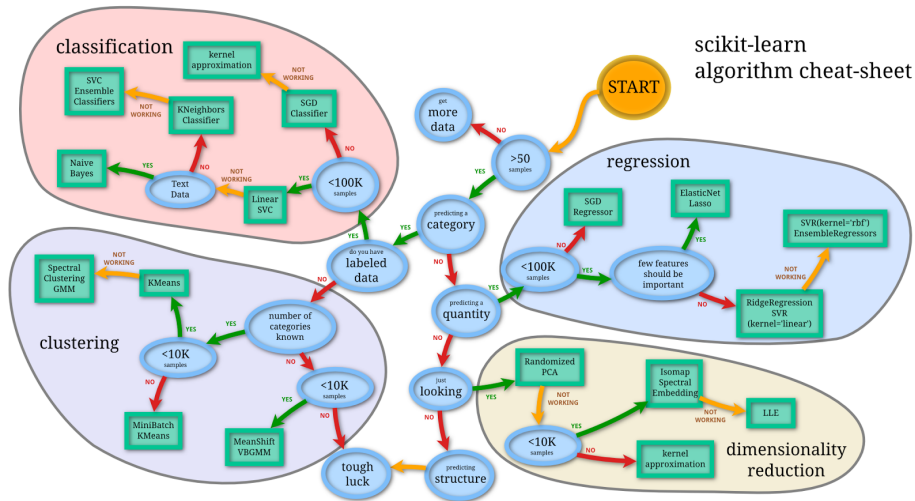
The Big Picture

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

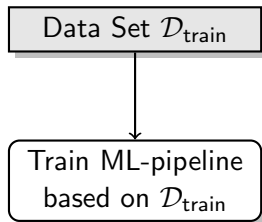
“Machine learning is the science of getting computers to act without being explicitly programmed.”

by Andrew Ng

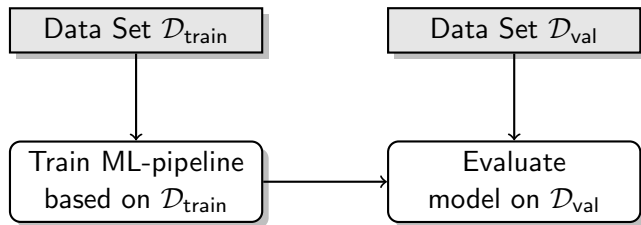
Machine Learning requires many design decisions



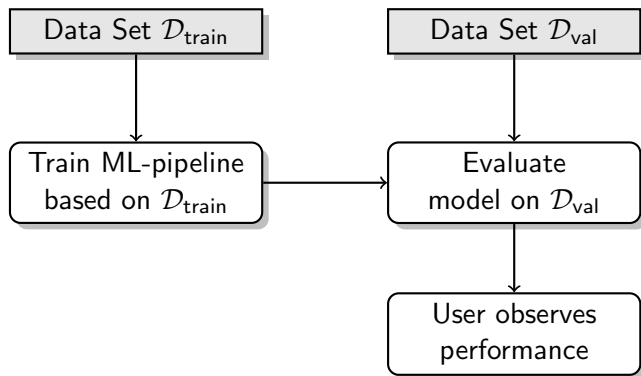
Machine Learning Workflow



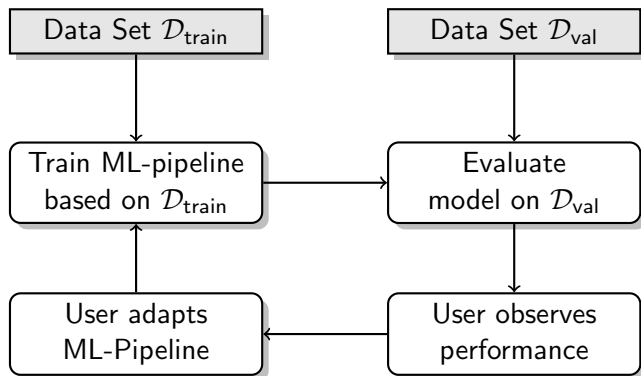
Machine Learning Workflow



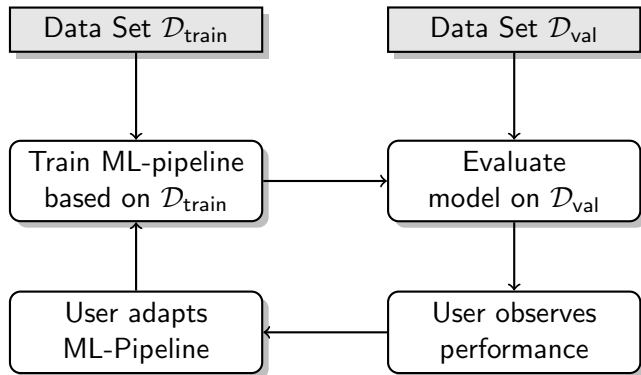
Machine Learning Workflow



Machine Learning Workflow



Machine Learning Workflow



⇒ Users indirectly teach machines how to learn.

Machine Learning does not scale up

- Basics in machine learning are not hard to grasp

Machine Learning does not scale up

- Basics in machine learning are not hard to grasp
- Achieving state-of-the-art performance is quite hard

Machine Learning does not scale up

- Basics in machine learning are not hard to grasp
- Achieving state-of-the-art performance is quite hard
- Design decisions are often not intuitive and require a lot of expertise
 - ▶ making these design decisions is a tedious and error-prone task

Machine Learning does not scale up

- Basics in machine learning are not hard to grasp
- Achieving state-of-the-art performance is quite hard
- Design decisions are often not intuitive and require a lot of expertise
 - ▶ making these design decisions is a tedious and error-prone task
- The job market for ML-experts is nearly empty

Machine Learning does not scale up

- Basics in machine learning are not hard to grasp
- Achieving state-of-the-art performance is quite hard
- Design decisions are often not intuitive and require a lot of expertise
 - ▶ making these design decisions is a tedious and error-prone task
- The job market for ML-experts is nearly empty
- Even with experts, developing new ML-applications takes time

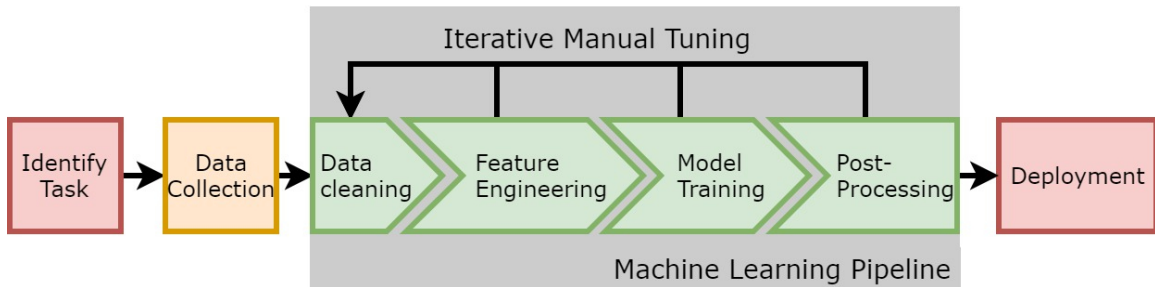
Machine Learning does not scale up

- Basics in machine learning are not hard to grasp
- Achieving state-of-the-art performance is quite hard
- Design decisions are often not intuitive and require a lot of expertise
 - ▶ making these design decisions is a tedious and error-prone task
- The job market for ML-experts is nearly empty
- Even with experts, developing new ML-applications takes time

Zoubin Ghahramani said that he often heard that:

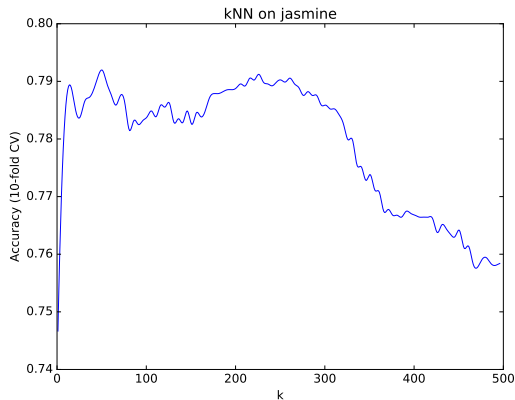
"I'd like to use machine learning, but I can't invest much time."

Why does ML development take a lot of time?



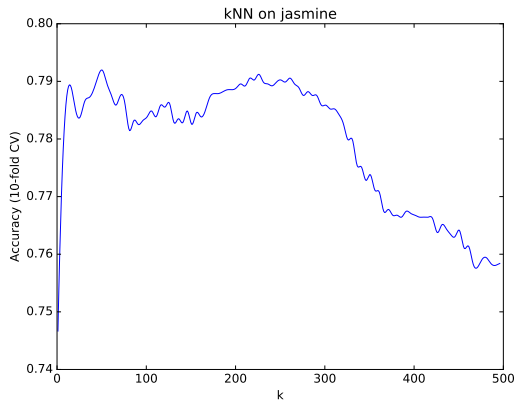
~> To achieve state-of-the-art performance,
this manual tuning has to be done for each new dataset again.

A Simple Example with k -NN



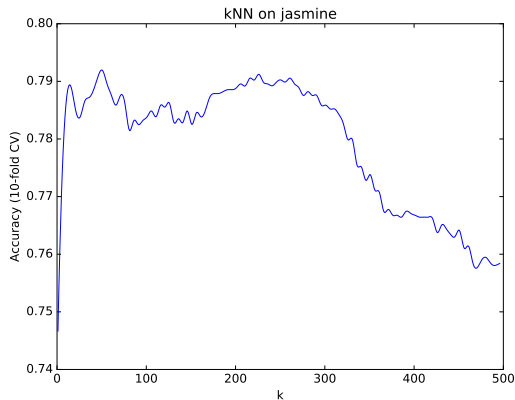
- k -nearest neighbors is one of the simplest ML algorithms

A Simple Example with k -NN



- k -nearest neighbors is one of the simplest ML algorithms
- Size of neighbourhood (k) is very important for its performance

A Simple Example with k -NN



- k -nearest neighbors is one of the simplest ML algorithms
- Size of neighbourhood (k) is very important for its performance
- The performance function depending on k is quite complex (not at all convex)

Goal of AutoML

AutoML

The goal of AutoML is to automate all parts of machine learning (as needed) to *support* users efficiently building their ML-applications.

Goal of AutoML

AutoML

The goal of AutoML is to automate all parts of machine learning (as needed) to *support* users efficiently building their ML-applications.

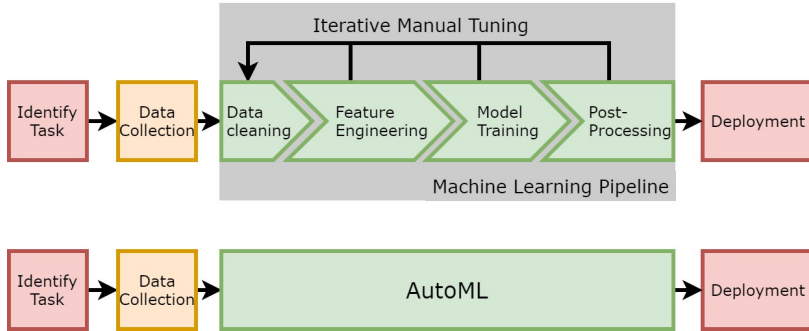
Informal Definition: AutoML System

Given

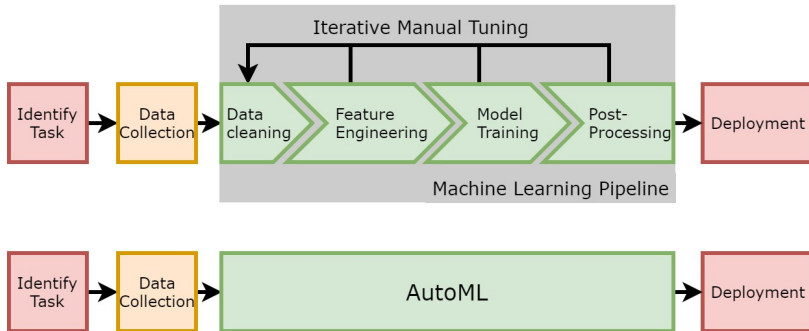
- a dataset
- a task (e.g., regression or classification)
- a cost metric (e.g., accuracy or RMSE)

an AutoML system automatically determines the approach that performs best for this particular application.

ML vs AutoML



ML vs AutoML



With AutoML, we ...

- support ML users
- improve the efficiency of developing new ML applications
- reduce the required ML-expertise
- might achieve better performance than developers w/o AutoML

AutoML in Research

AutoML enables:

- ① more efficient research
 - ▶ AutoML has shown on subproblems to outperform human experts

AutoML in Research

AutoML enables:

- ① more efficient research
 - ▶ AutoML has shown on subproblems to outperform human experts
- ② more systematic research
 - ▶ humans tend to be unsystematic which leads to errors

AutoML in Research

AutoML enables:

- ① more efficient research
 - ▶ AutoML has shown on subproblems to outperform human experts
- ② more systematic research
 - ▶ humans tend to be unsystematic which leads to errors
- ③ more reproducible research
 - ▶ human's unsystematic approaches cannot be reproduced, but AutoML is systematic

AutoML in Research

AutoML enables:

- ① more efficient research
 - ▶ AutoML has shown on subproblems to outperform human experts
- ② more systematic research
 - ▶ humans tend to be unsystematic which leads to errors
- ③ more reproducible research
 - ▶ human's unsystematic approaches cannot be reproduced, but AutoML is systematic
- ④ broader use of ML also in other disciplines
 - ▶ ML should not be limited to computer scientists;
 - ▶ the most amazing applications of ML are often done by either interdisciplinary teams or even non-computer scientists

Challenges in AutoML

- ① Each dataset potentially require different optimal ML-designs
 - ▶ Design decisions have to be made for each dataset again

Challenges in AutoML

- ① Each dataset potentially require **different optimal ML-designs**
 - ▶ Design decisions have to be made for each dataset again
- ② Training of a single ML model can be quite **expensive** (e.g., hours, days or weeks)
 - ▶ often, we cannot try many design decisions

Challenges in AutoML

- ① Each dataset potentially require **different optimal ML-designs**
 - ▶ Design decisions have to be made for each dataset again
- ② Training of a single ML model can be quite **expensive** (e.g., hours, days or weeks)
 - ▶ often, we cannot try many design decisions
- ③ the **mathematical relation** between design decisions and performance is (often) **unknown**
 - ▶ gradient-based optimization is not directly possible

Challenges in AutoML

- ① Each dataset potentially require **different optimal ML-designs**
 - ▶ Design decisions have to be made for each dataset again
- ② Training of a single ML model can be quite **expensive** (e.g., hours, days or weeks)
 - ▶ often, we cannot try many design decisions
- ③ the **mathematical relation** between design decisions and performance is (often) **unknown**
 - ▶ gradient-based optimization is not directly possible
- ④ optimization in **highly complex spaces**
 - ▶ incl. categorical choices, continuous parameters, conditional dependencies