Automated Machine Learning

1. Tools & Techniques
2. Creating Value with Automation

Getting started

Check out the Github repository:

https://github.com/vikua/aml-class-20

Follow installation instructions

https://github.com/vikua/aml-class-20#requirements

Who am I?

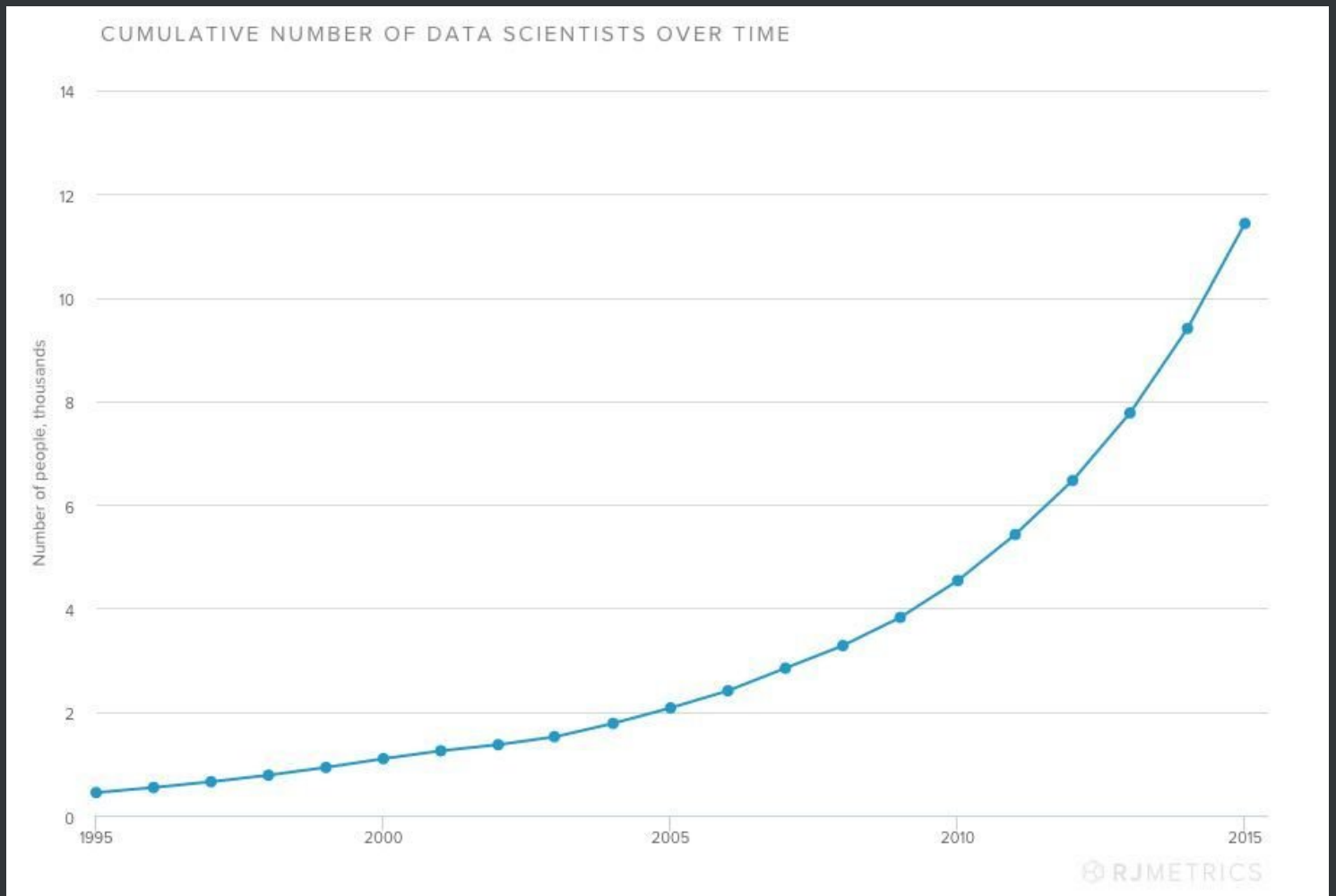
- Lead ML engineer at DataRobot (Automated ML vendor)
- Part of Time Series team
- Software Engineer by training, stumbled into Machine Learning

Source: https://www.nytimes.com/2017/11/05/technology/machine-learning-artificial-intelligence-ai.html

Why Automated Machine Learning?

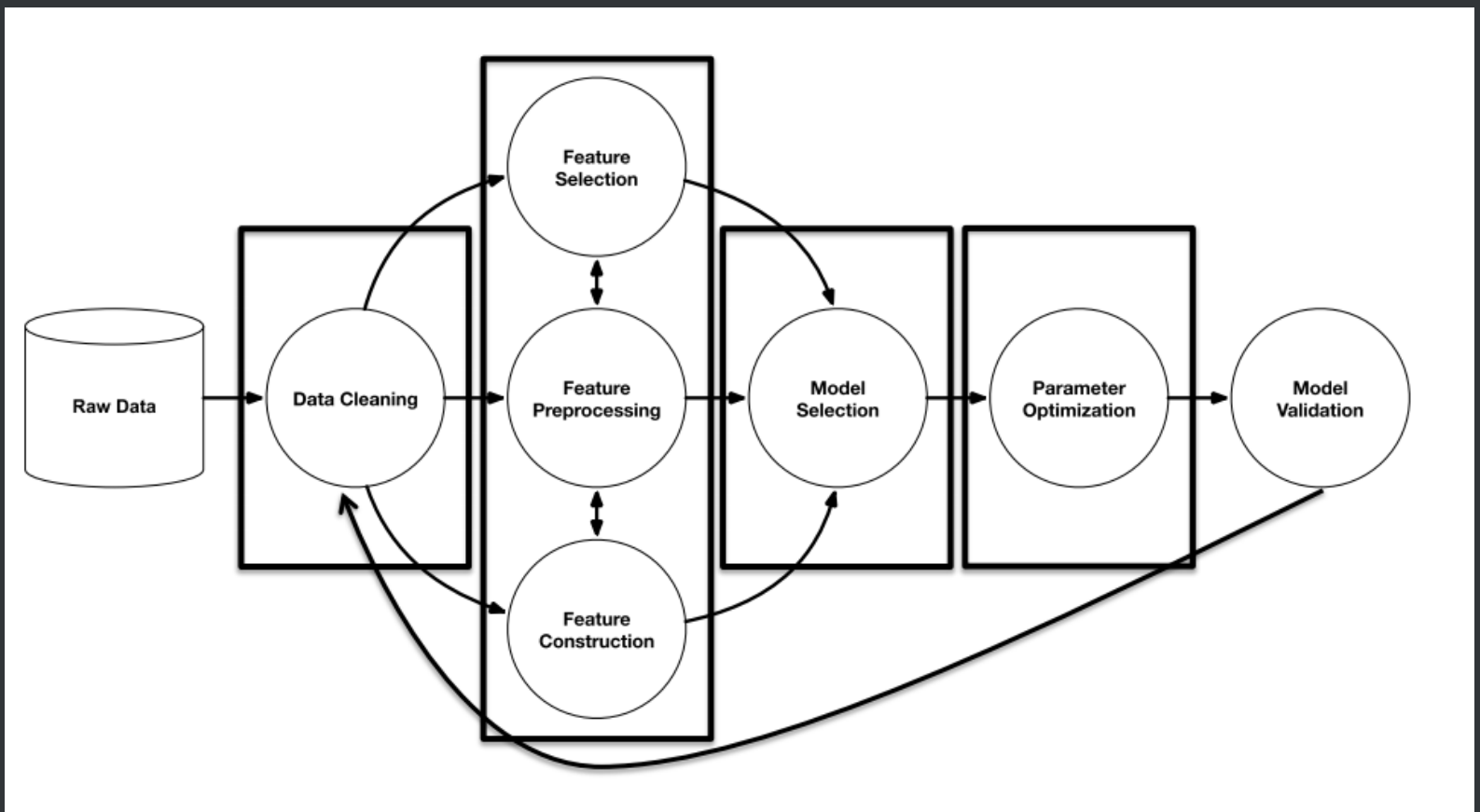
ML allows us to create software components without the need to explicitly program them,

however, creating software components via ML still requires substantial "meta-programming".



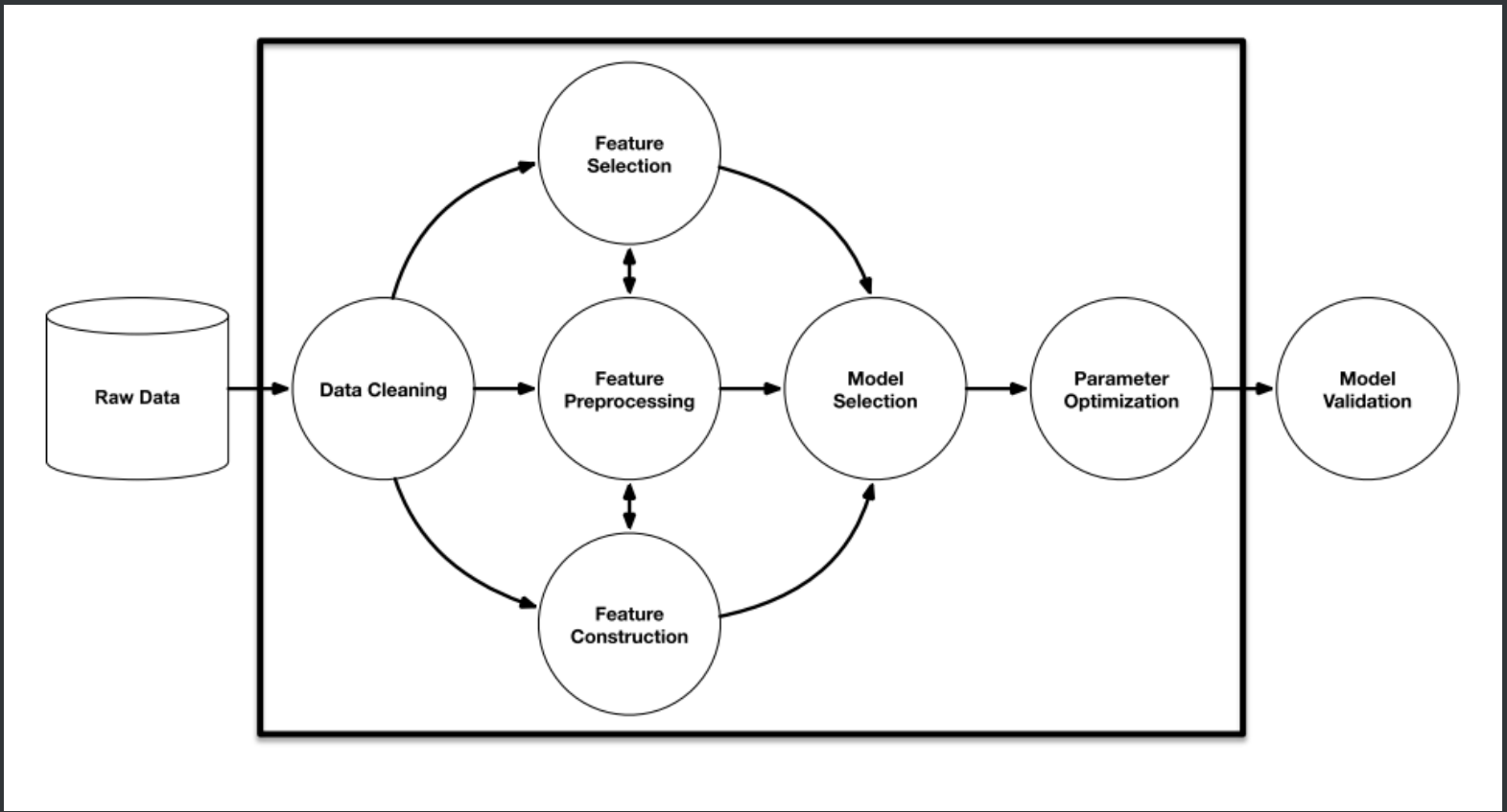
Source: globalnewswire.com | RJMetrics (2015)

ML still requires a lot of manual programming



Source: R. Olson et. al. (2016) "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science."

AML claims to automate the entire process



Source: R. Olson et. al. (2016) "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science."

Benefits of Automated Machine Learning?

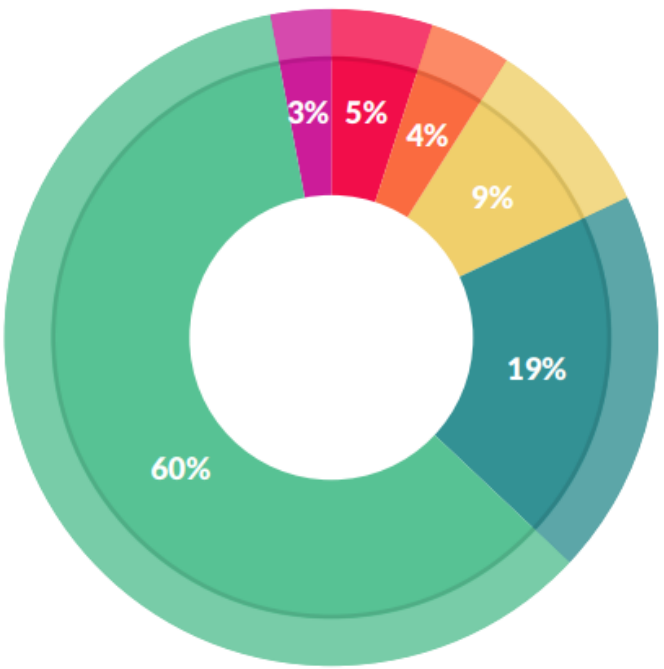
Efficiency gains

- Workforce shortage
- Cost optimization

Effectiveness gains

- Consistently good / great performance
- Guardrails

Efficiency: (state-of-the-art) AML can help with some of this ...

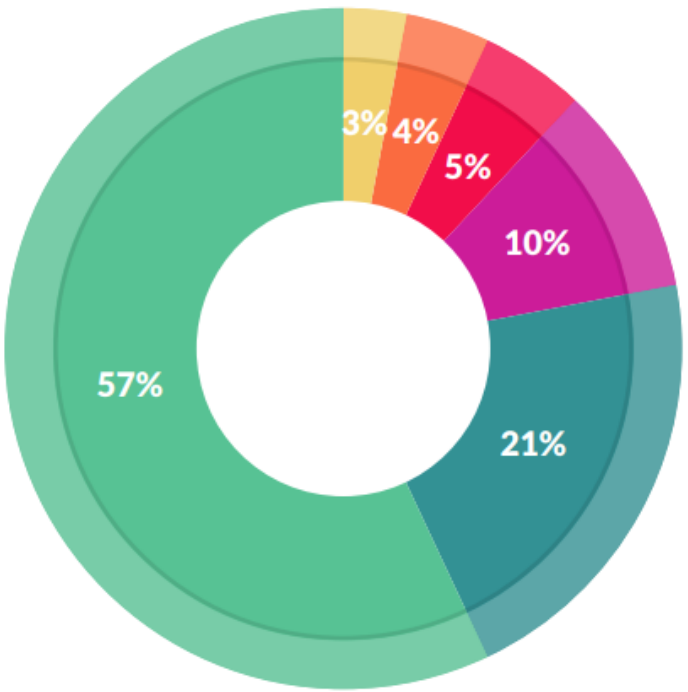


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: https://visit.figure-eight.com/data-science-report.html

Efficiency: ... but mostly the fun part.



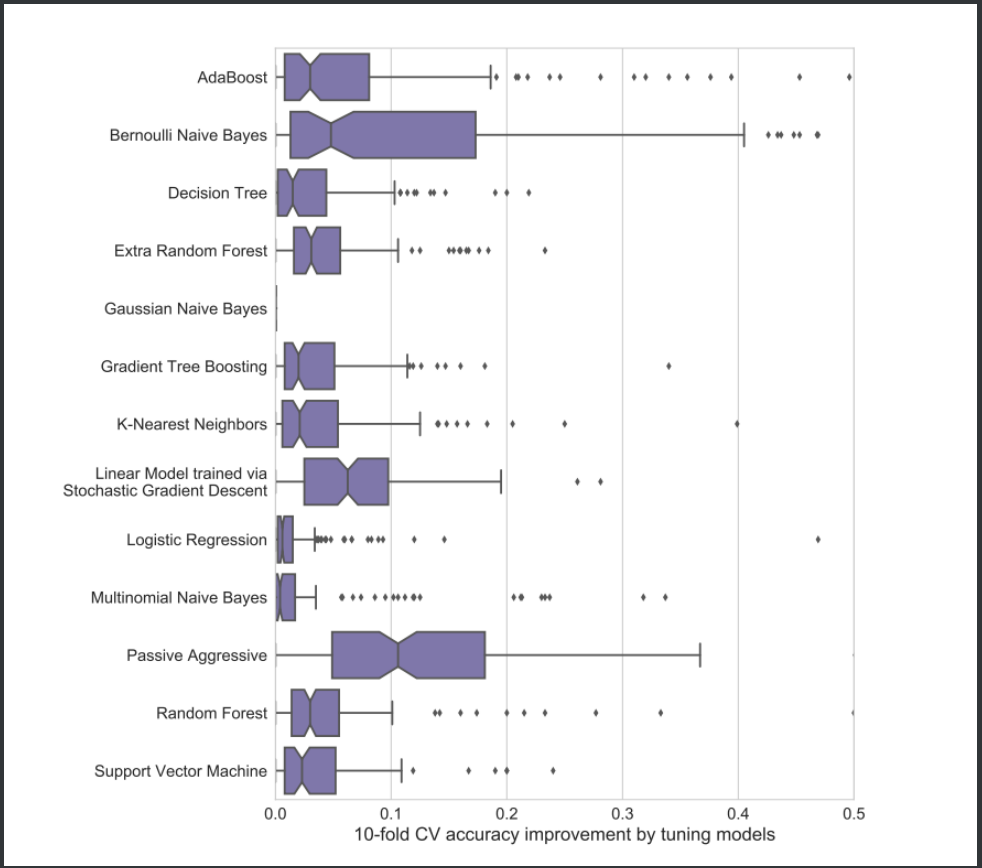
What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: https://visit.figure-eight.com/data-science-report.html

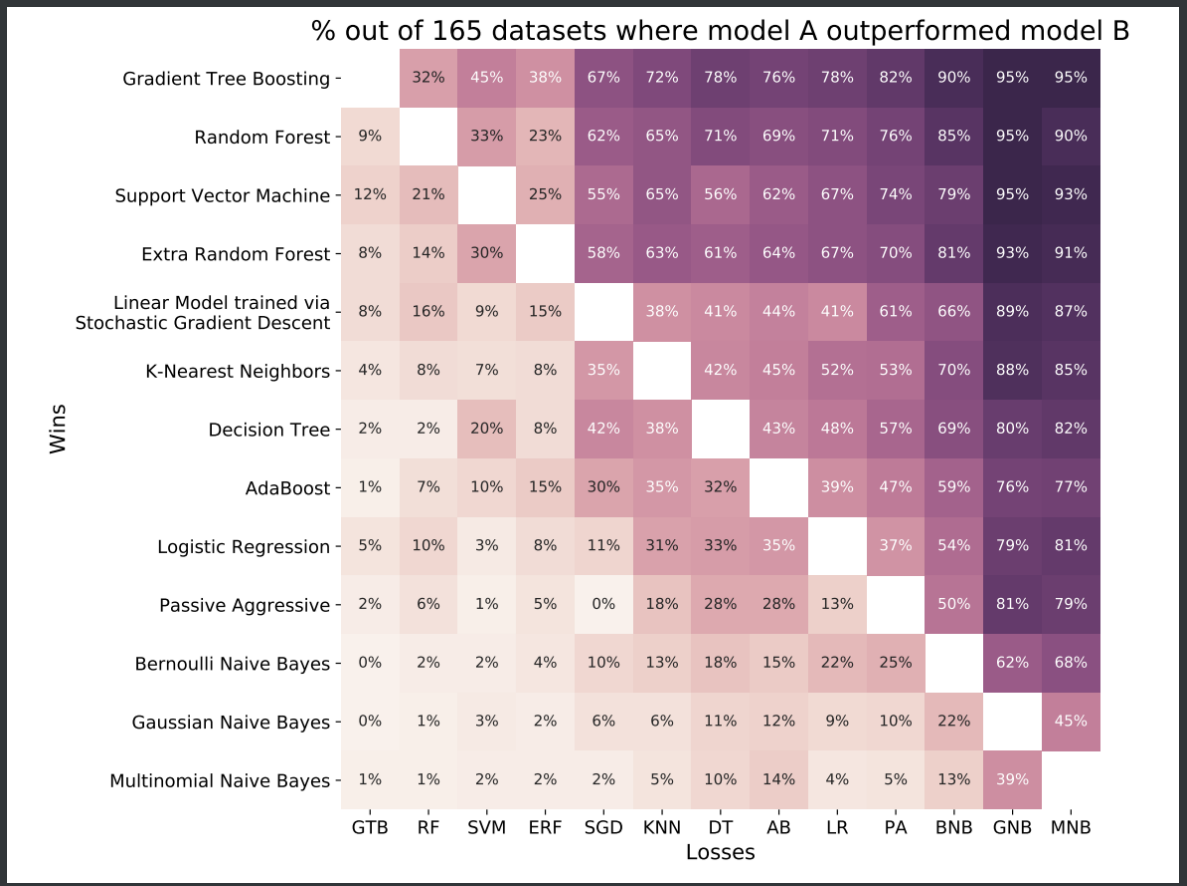
How does AML achieve this?

Efficient tuning of hyper-parameters



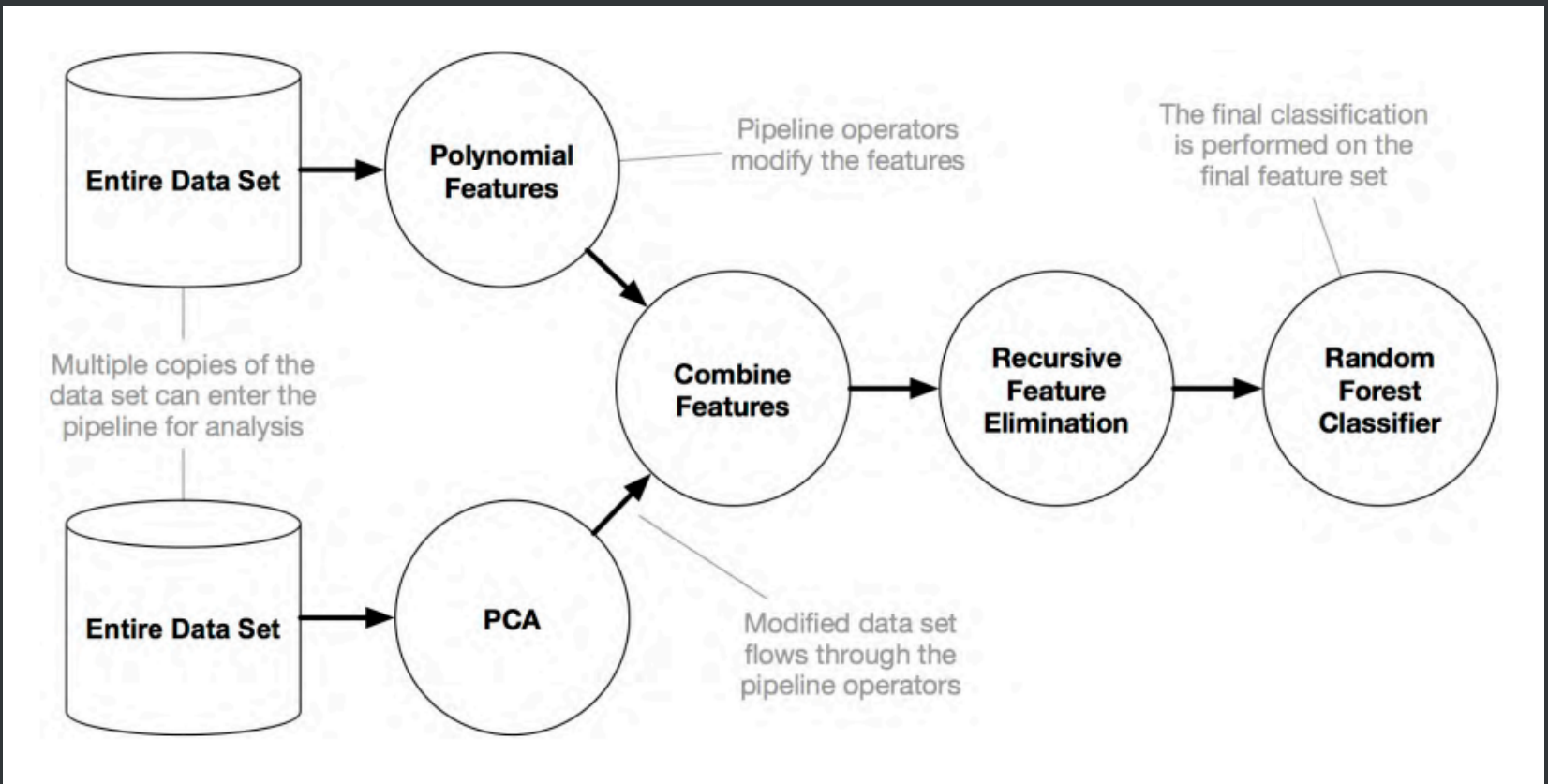
Source: R. Olson et. al. (2017) "Data-driven Advice for Applying Machine Learning to Bioinformatics Problems."

Selecting the right modeling approach



Source: R. Olson et. al. (2017) "Data-driven Advice for Applying Machine Learning to Bioinformatics Problems."

Modern ML is more complex than that



Source: R. Olson et. al. (2017) "Data-driven Advice for Applying Machine Learning to Bioinformatics Problems."

Past, Present and Future of AML

- Past: parameter optimization and model selection
- Present: pipeline optimization and architecture search
- (Near) Future: continous learning, data cleansing & fusion, AML at-scale
- Future: Autonomous Machine Learning

Agenda

1. Introduction
    1. What is AML
    2. Machine Learning 101
    3. Model Selection & Assessment
    4. Hyper-parameter Tuning
    5. Debugging and Improving ML model
2. AML Tools & Techniques
    1. ML Pipelines
    2. Bayesian Hyper-parameter Optimization
    3. Pipeline Optimization
    4. Advanced Topics
3. Creating value using Automation
    1. Democratization of ML
    2. Model Factories
    3. Continuous learning