Democratizing AI, and Surviving Titanic with Automated Machine Learning

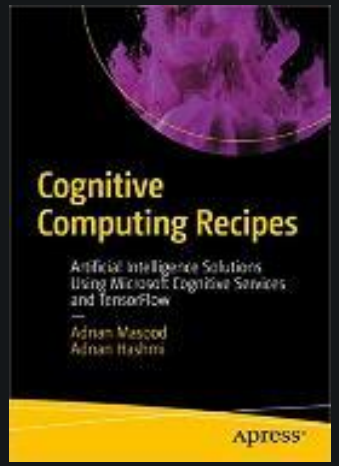
Adnan Masood, PhD.
@adnanmasood
adnanmasood@gmail.com

Microsoft Azure
+AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection

Microsoft® Most Valuable Professional

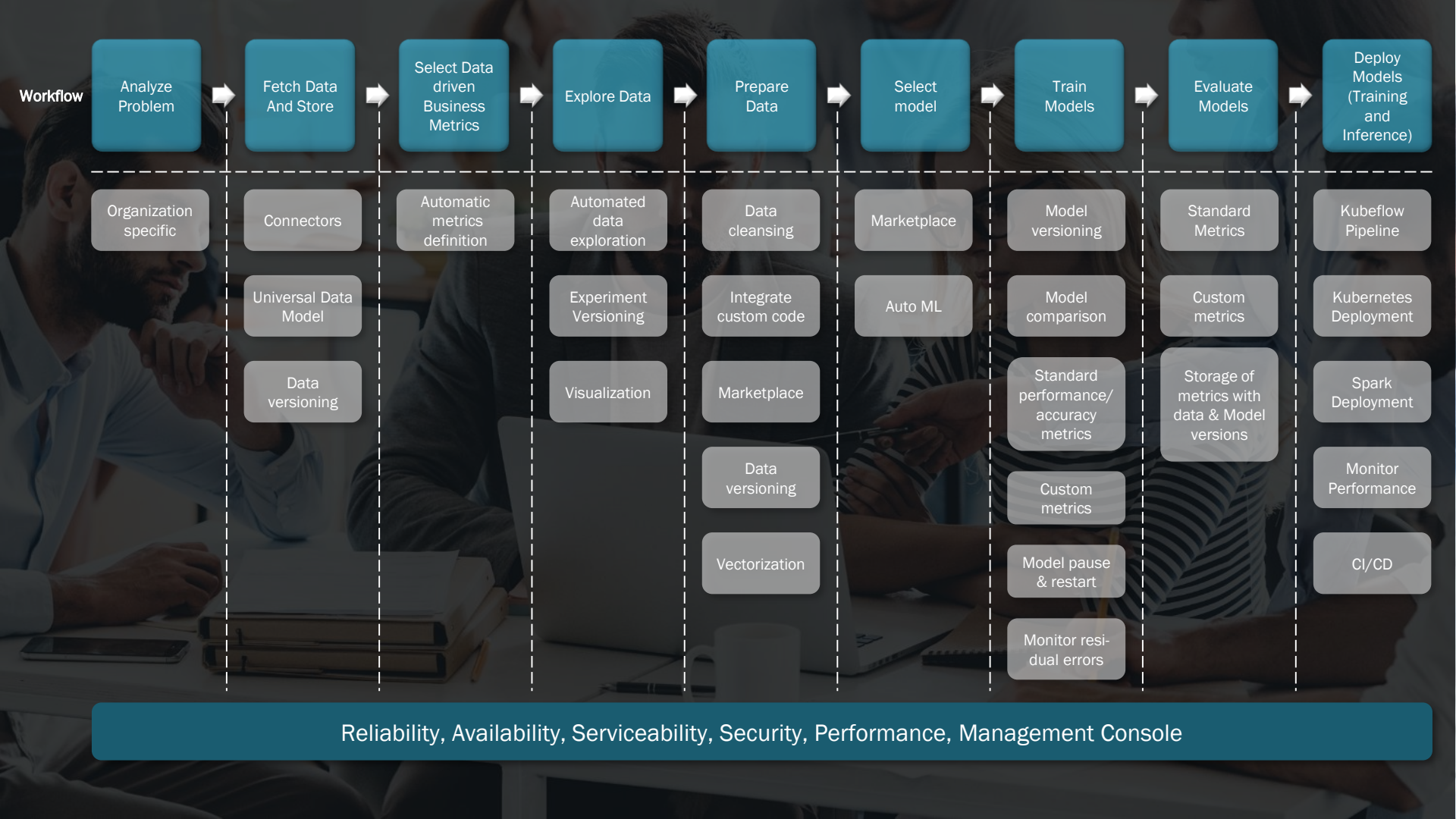
Dr. Adnan Masood

Adnan Masood, Ph.D. - Chief Architect - Artificial Intelligence and Machine Learning at UST Global, Visiting Scholar at Stanford University, and Microsoft MVP (Most Valuable Professional) for AI.



MIT CSAIL ALLIANCES

Stanford ARTIFICIAL INTELLIGENCE


Workflow
Analyze Problem
Fetch Data And Store
Select Data driven Business Metrics
Explore Data
Prepare Data
Select model
Train Models
Evaluate Models
Deploy Models (Training and Inference)
Organization specific
Connectors
Automatic metrics definition
Automated data exploration
Data cleansing
Marketplace
Model versioning
Standard Metrics
Kubeflow Pipeline
Universal Data Model
Experiment Versioning
Integrate custom code
Auto ML
Model comparison
Custom metrics
Kubernetes Deployment
Data versioning
Visualization
Marketplace
Standard performance/ accuracy metrics
Storage of metrics with data & Model versions
Spark Deployment
Data versioning
Custom metrics
Monitor Performance
Vectorization
Model pause & restart
CI/CD
Monitor resi-dual errors
Reliability, Availability, Serviceability, Security, Performance, Management Console

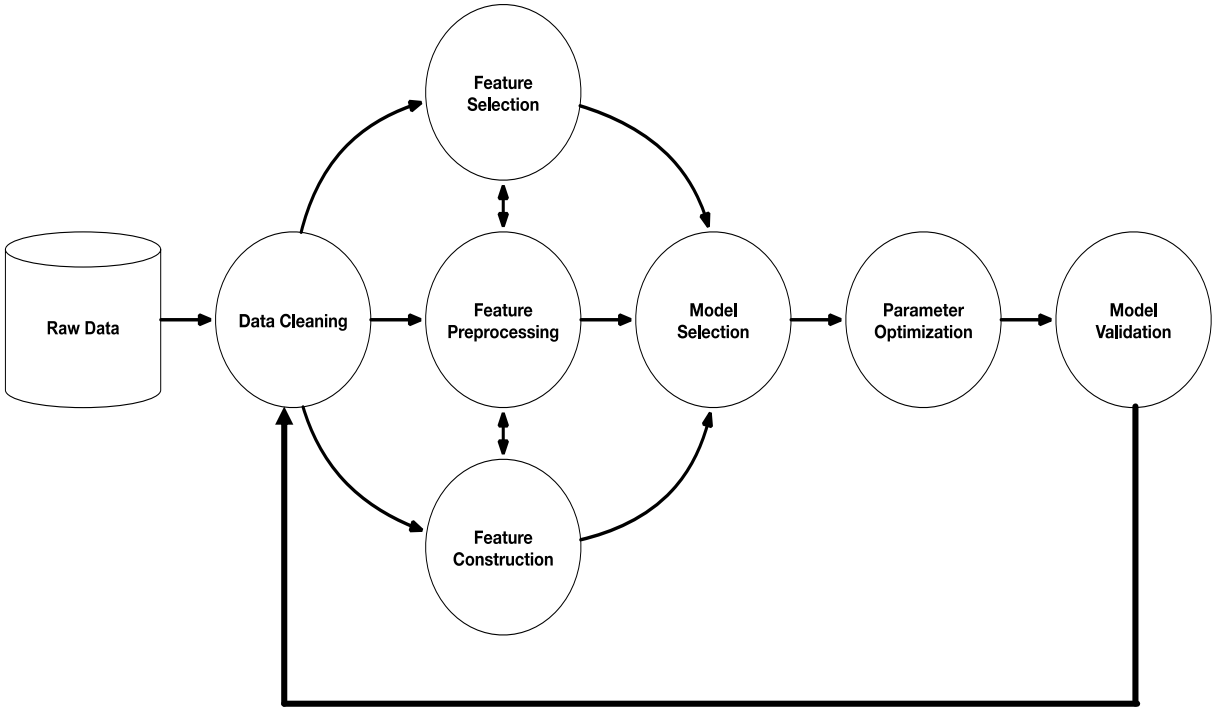
“AutoML is a quiet revolution in AI…”

Automated Machine Learning—A Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb

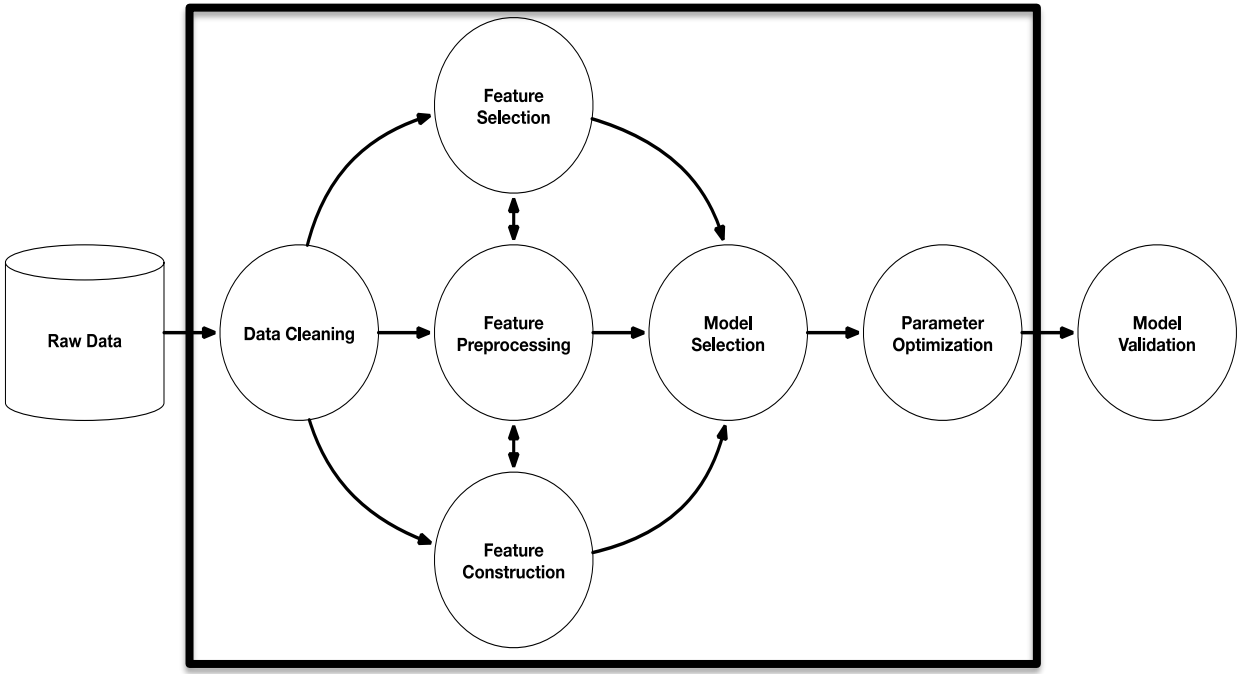
Building A.I. That Can Build A.I.

Google and others, fighting for a small pool of researchers, are looking for automated ways to deal with a shortage of artificial intelligence experts.

ML still requires a lot of manual programming

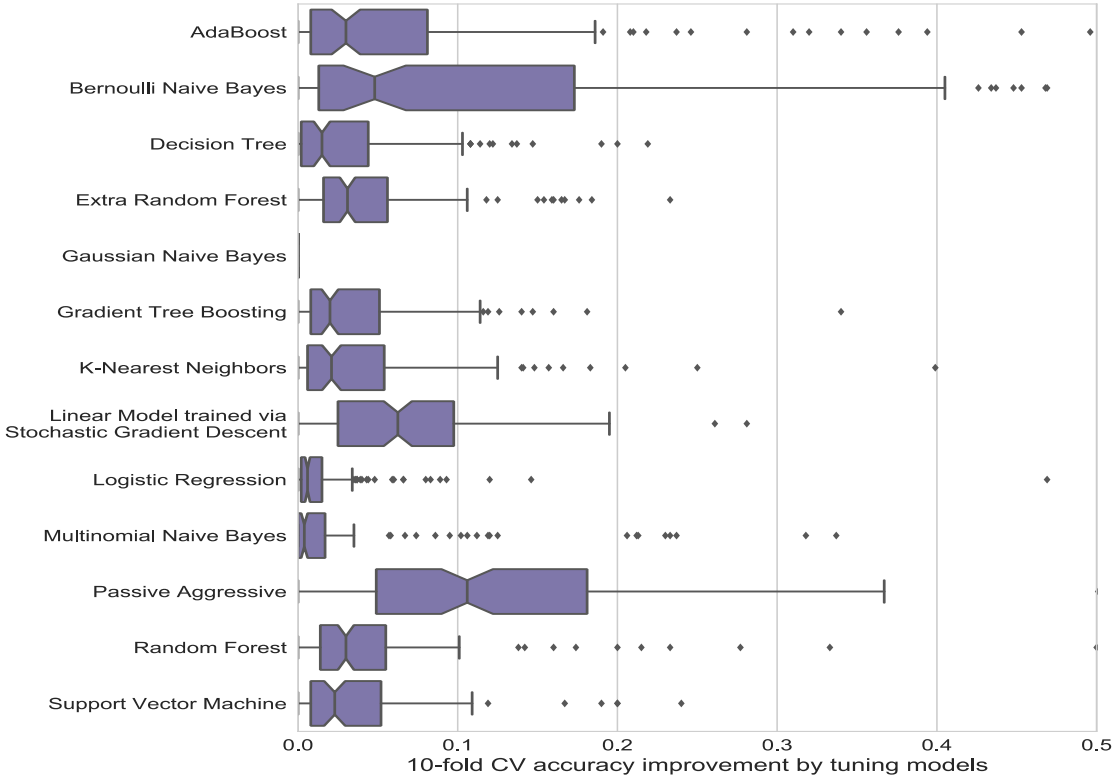


AutoML aims to automate the entire ML workflow



Default parameters are almost always bad

AutoML handles this for you!



AutoML is a huge time-saver

AutoML handles (some of) this for you!

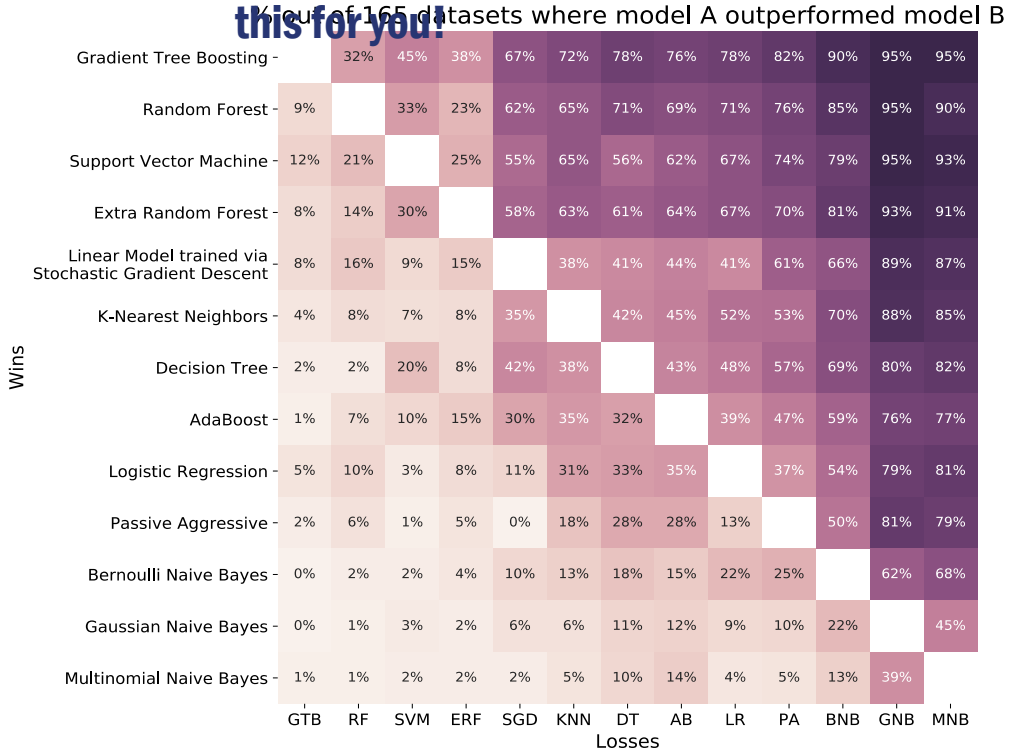


Image source: R. Olson & W. La Cava *et. al.* (2017) “Data-driven advice for applying machine learning to bioinformatics problems.”

AutoML is a huge time-saver

AutoML handles (some of) this for you!

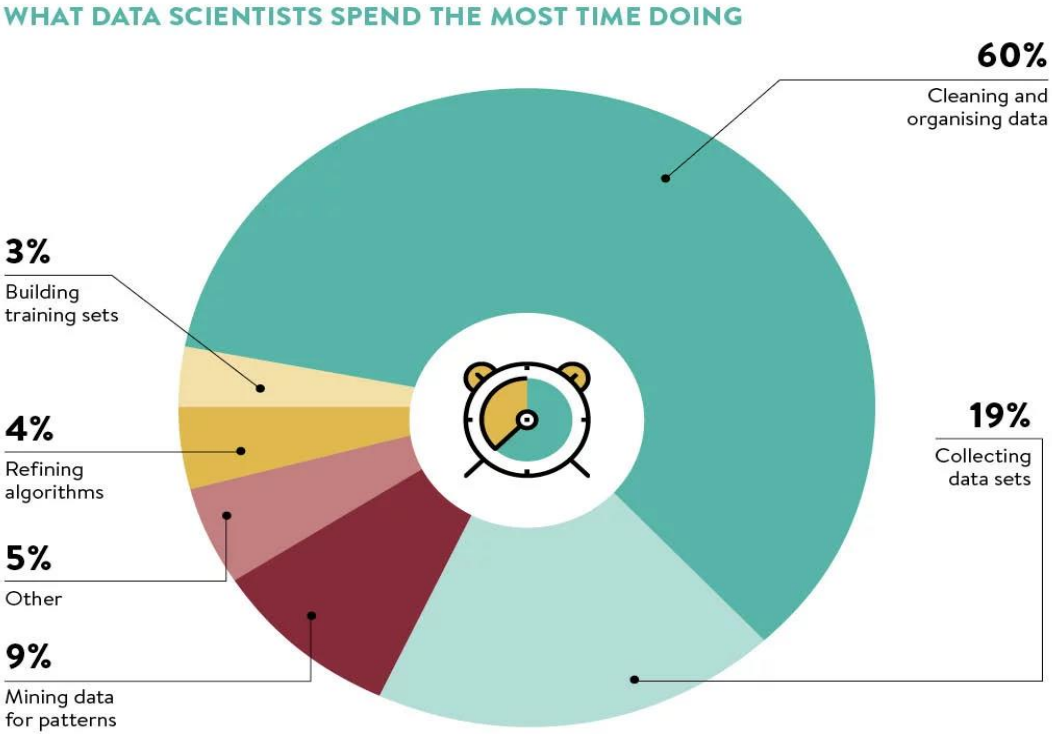


Image source: visit.crowdflower.com/data-science-report.html

The business case for AutoML

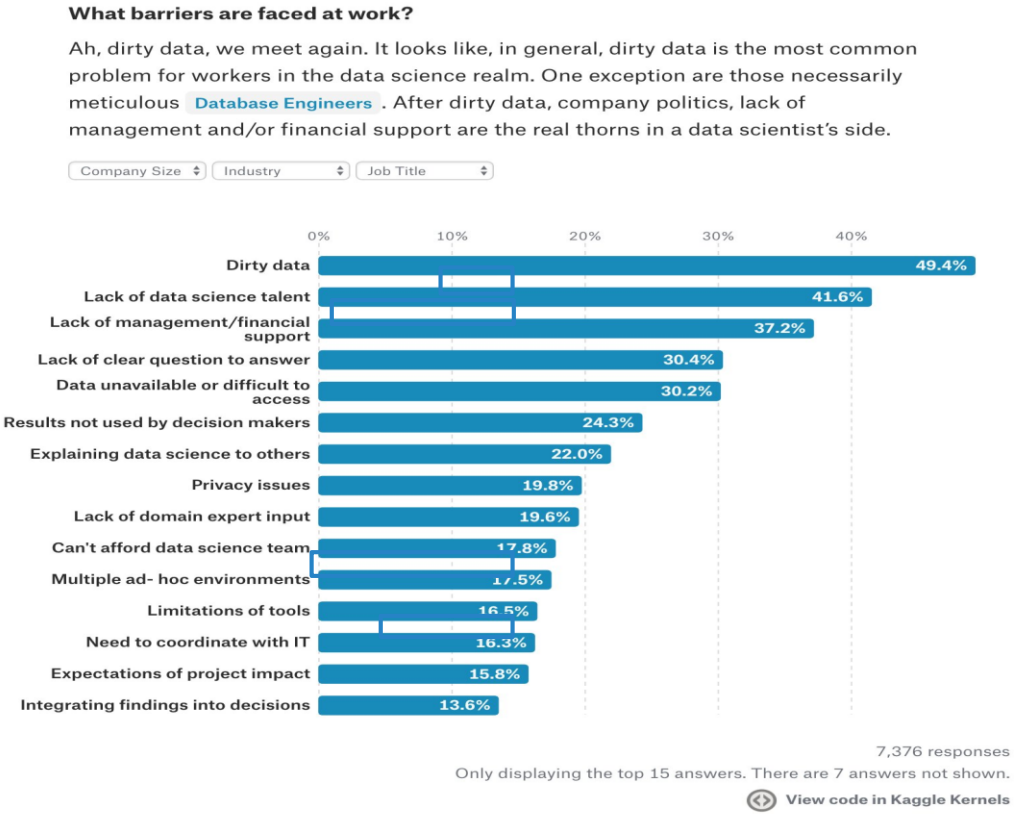


Image source: kaggle.com/surveys/2017

Early AutoML focused on only parameter tuning

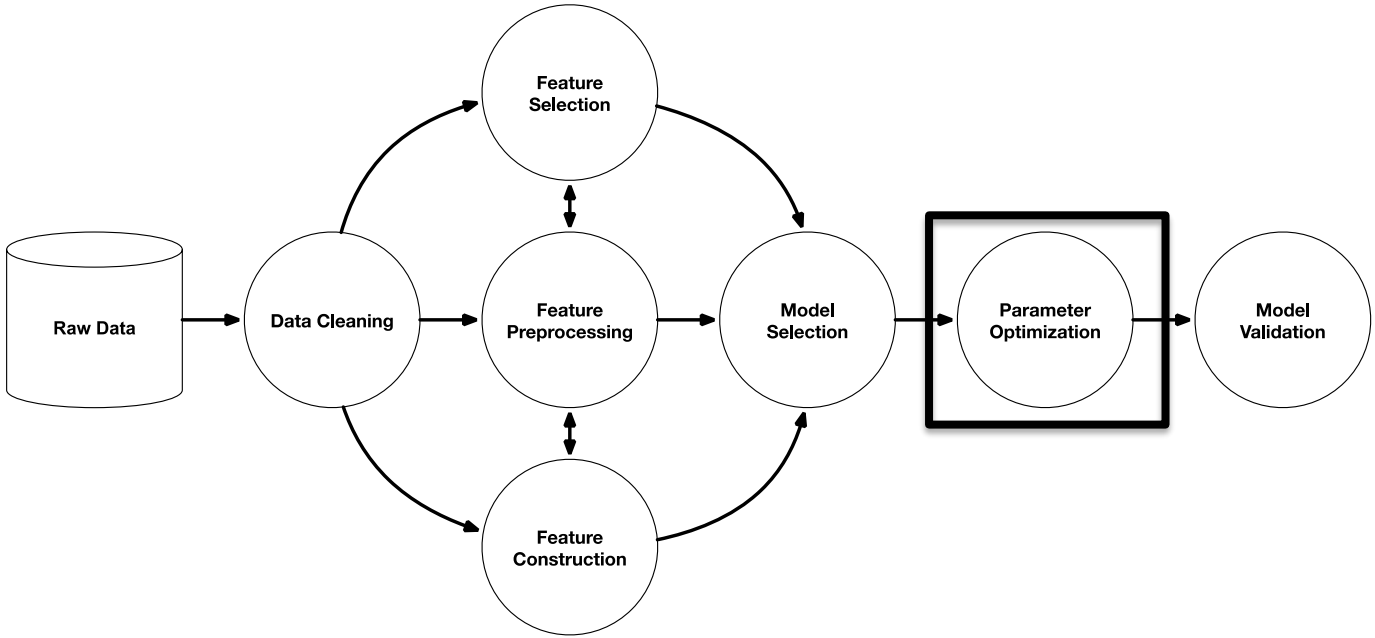
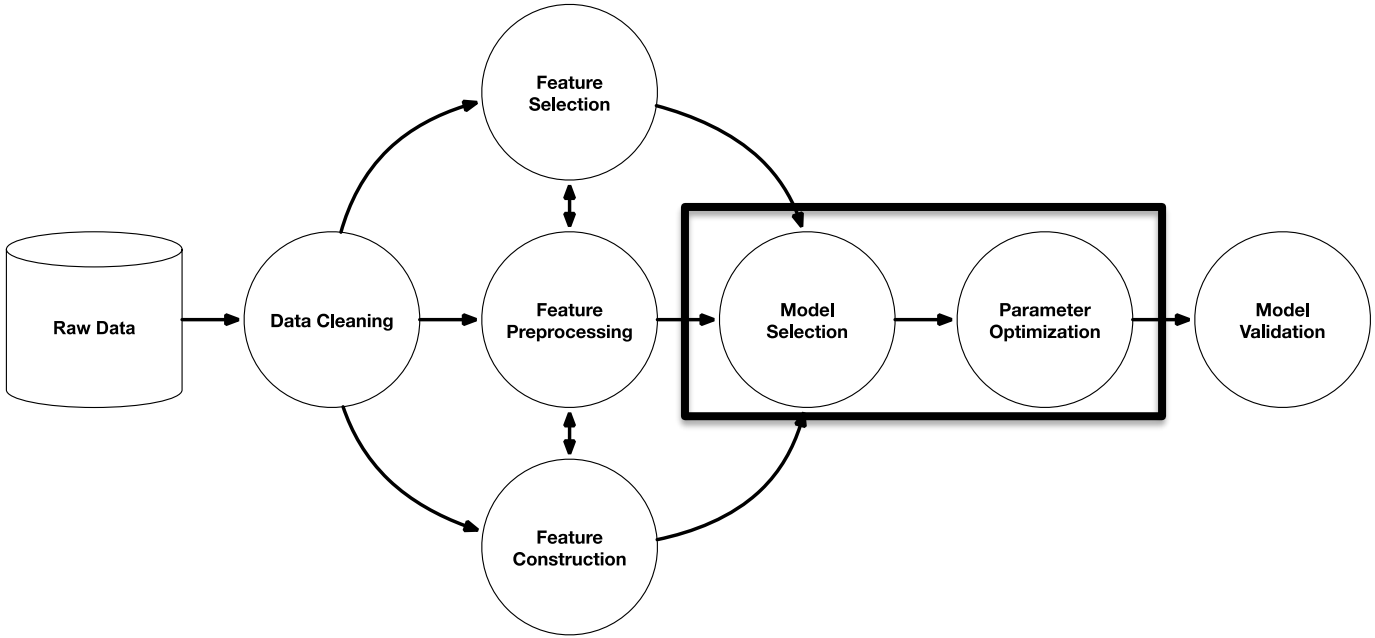


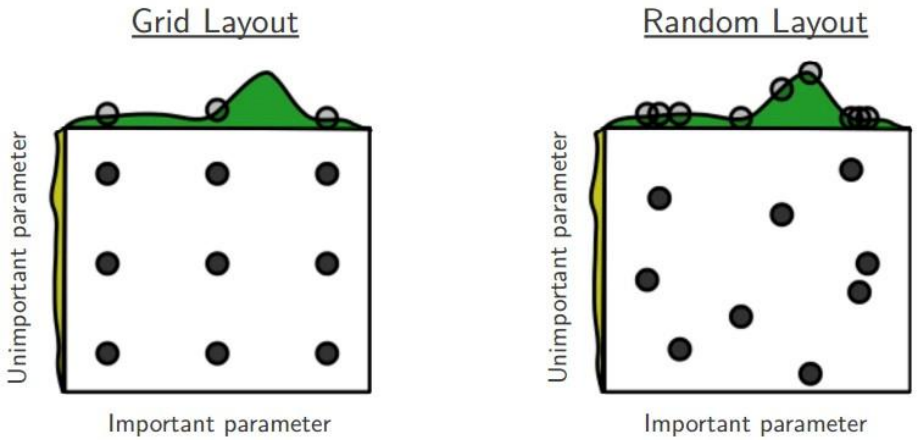
Image source: R. Olson e*t. al.* (2016) “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.”

Early AutoML focused on only parameter tuning



... and maybe (limited) model selection

We mostly used grid search and random search



Nowadays, we wouldn’t really call this AutoML

Modern AutoML optimizes the entire ML workflow

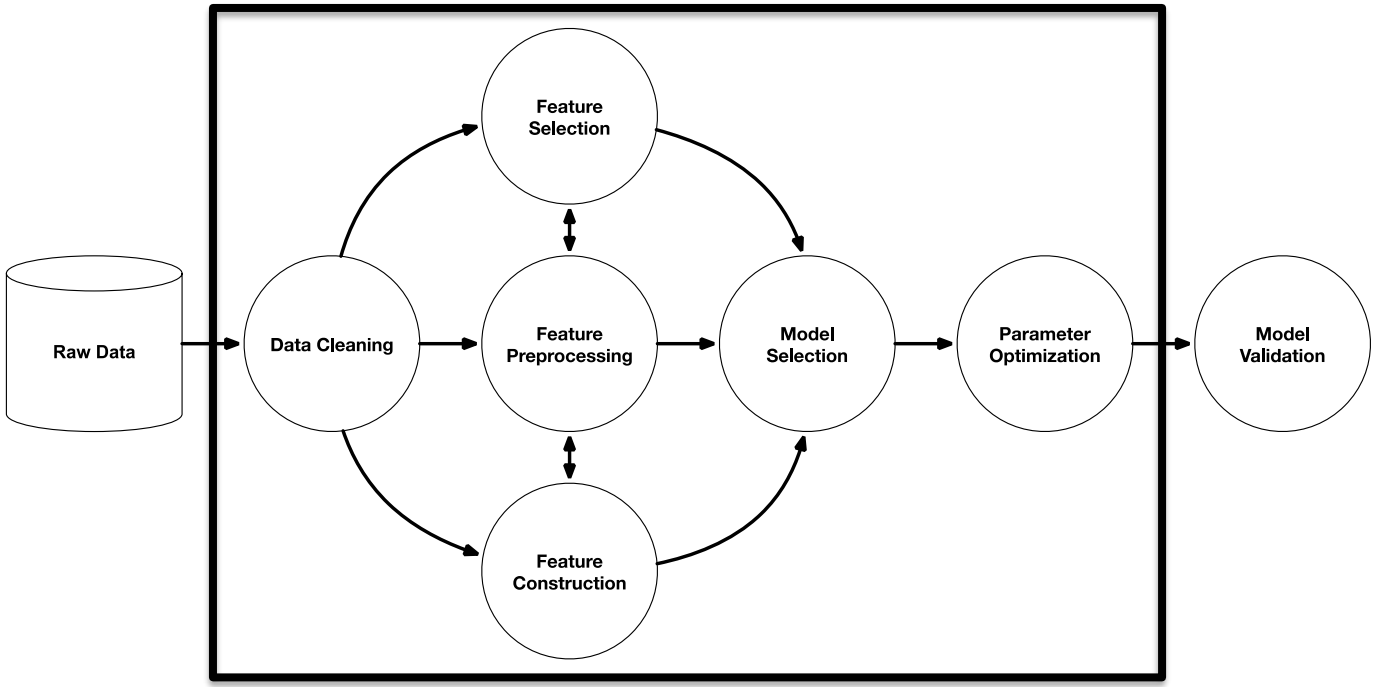
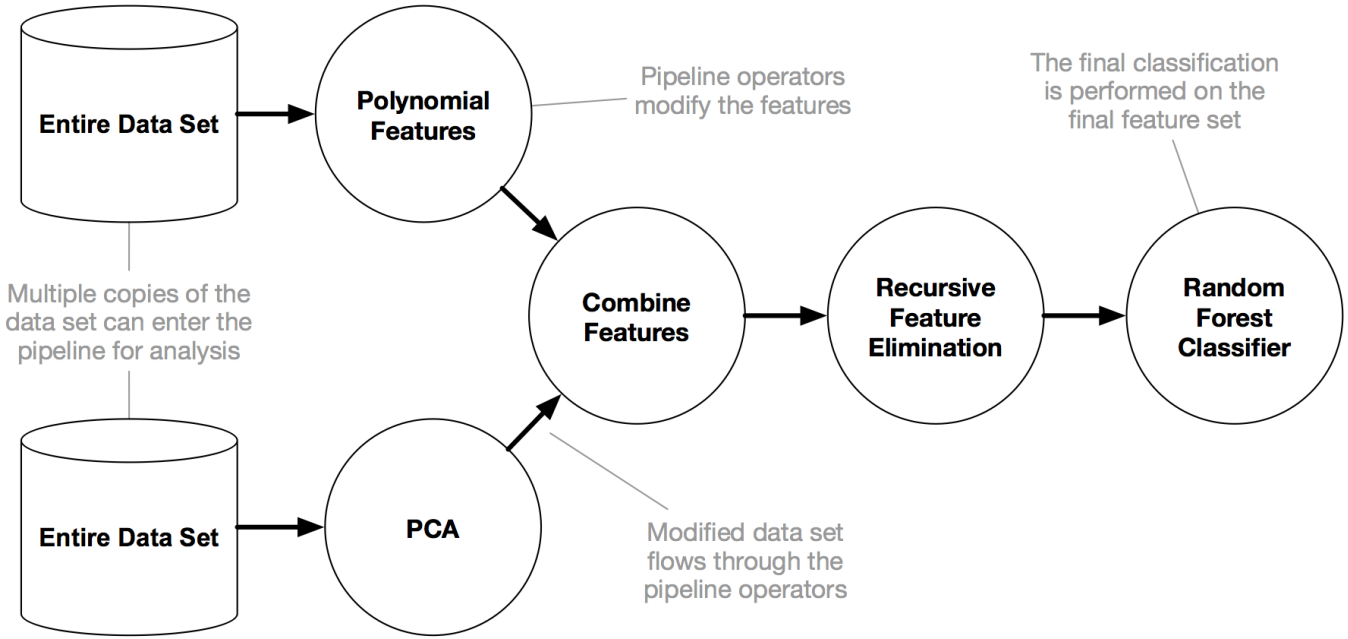


Image source: R. Olson et. *al.* (2016) “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science.”

Modern AutoML optimizes the entire ML workflow



Open source AutoML tools

- **auto-sklearn [Python]**
  - Bayesian optimization over a fixed 3-step ML pipeline
  - github.com/automl/auto-sklearn
- **auto-Weka [Java]**
  - Similar to auto-sklearn, but built on top of Weka
  - github.com/automl/autoweka
- **TPOT [Python]**
  - Genetic Programming over a configurable ML pipeline
  - github.com/rhiever/tpot
- **H2O.ai AutoML [Java w/ Python, Scala, & R APIs and web GUI]**
  - Basic data prep w/ mix of grid and random search over ML algorithms
  - github.com/h2oai/h2o-3
- **devol [Python]**
  - Deep Learning architecture search via Genetic Programming
  - github.com/joeddav/devol

AutoMLaaS: Commercial AutoML tools

- **DataRobot**
  - Web-based interface
  - Fixed search over thousands of ML pipelines
- **H2O.ai Driverless AI**
  - Web-based interface
  - H2O.ai AutoML + better feature construction
- **Google AutoML**
  - Integrated in the Google Cloud Compute platform
  - DNN architecture search
- **SAS Factory Miner**
  - Fixed search over a handful of ML methods
- **IBM SPSS Modeler**
  - Basic automated data preparation and ML modeling

AutoML in the near future

- **AutoML will also handle most of the data cleaning process**
  - Unstructured data → tabular data ready for analysis
  - Capture & automate human approaches to data cleaning
- **AutoML will vastly improve Deep Learning**
  - Automated DNN architecture design
  - Automated preprocessing of data prior to modeling
- **AutoML will scale to large datasets**
  - AutoML is very slow right now on “Big Data”
  - Spark, dask, TensorFlow, etc. will help bring AutoML to scale
- **AutoML will become human-competitive**
  - Already human-competitive on several Kaggle challenges
  - Already human-competitive in DNN architecture design (Google AutoML)

AutoML in the future

- **AutoML will transform the practice of data science as we know it**
  - "Data Science Assistant" → Junior Data Scientist level
  - Less focus on choosing the right ML workflow
  - More focus on posing the right questions, collecting & curating the right data, and "thinking like a data scientist"

- **AutoML will become productized**
  - Not AutoMLaaS!
  - "Siri, set an alarm for 6am" → "Siri, set an alarm for the best time for me to wake up"
  - "Siri, [given my personal medical history] should I worry about this rash on my face?"

- **AutoML is only a small part of a greater meta-learning movement**
  - Computer programming is focused on automating rote tasks
  - Machine learning is focused on automating the automation of rote tasks
  - Meta-learning is focused on *automating the automation of automation*
    - i.e., enabling the machine to learn *how* to learn in the best way possible

Questions

Please use EventsXD to fill out a session evaluation.

Thank you!

Microsoft Azure
+AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection