Complexity vs. Performance: Empirical Analysis of Machine Learning as a Service

Yuanshun Yao, Zhujun Xiao, Bolun Wang*, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao

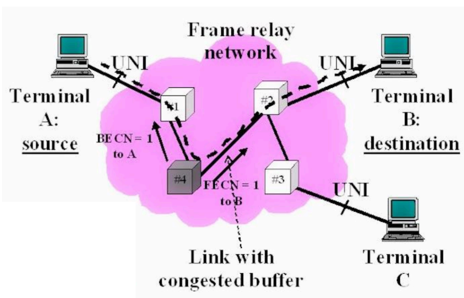
The University of Chicago

*University of California, Santa Barbara

ysyao@cs.uchicago.edu

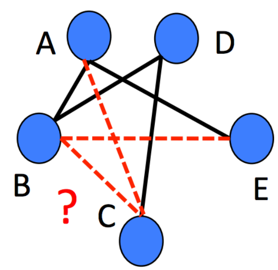
ML in Network Research

congestion control protocols



- Sivaraman et al., SIGCOMM'14
- Winstein & Balakrishnan, SIGCOMM'13

network link prediction



- Liu et al., IMC'16
- Zhao et al., IMC'12

user behavior analysis

- Wang et al., IMC'14
- Zannettouet al., IMC'17

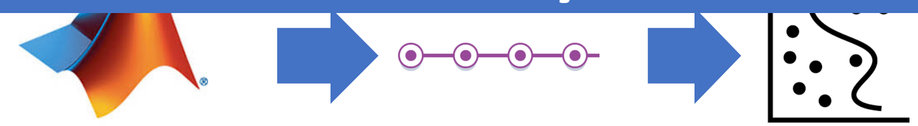
…

Running ML is Hard

Solution:

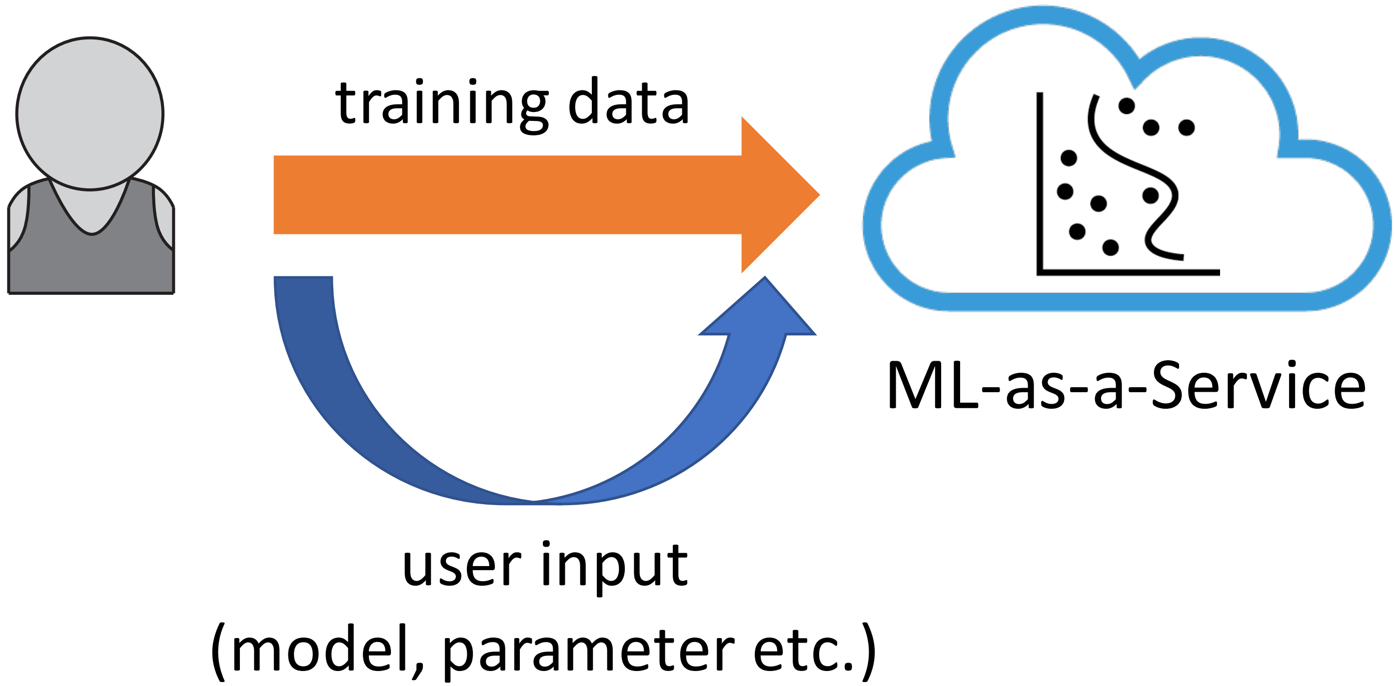
Machine Learning as a Service

(ML-as-a-Service)



model

ML-as-a-Service



Why Study ML-as-a-Service?

Is my model good enough?

Q: How well do they perform?

Q: How much does the amount of user control impact ML performance?

ML-as-a-Service Platforms

Google Prediction | Amazon ML | Microsoft ML

Azure ML

ABM | BigML | PIO

bigml

less — amount of user input — more

Control in ML

Control in ML

Data Cleaning

- Invalid/dup/missing data

Control in ML

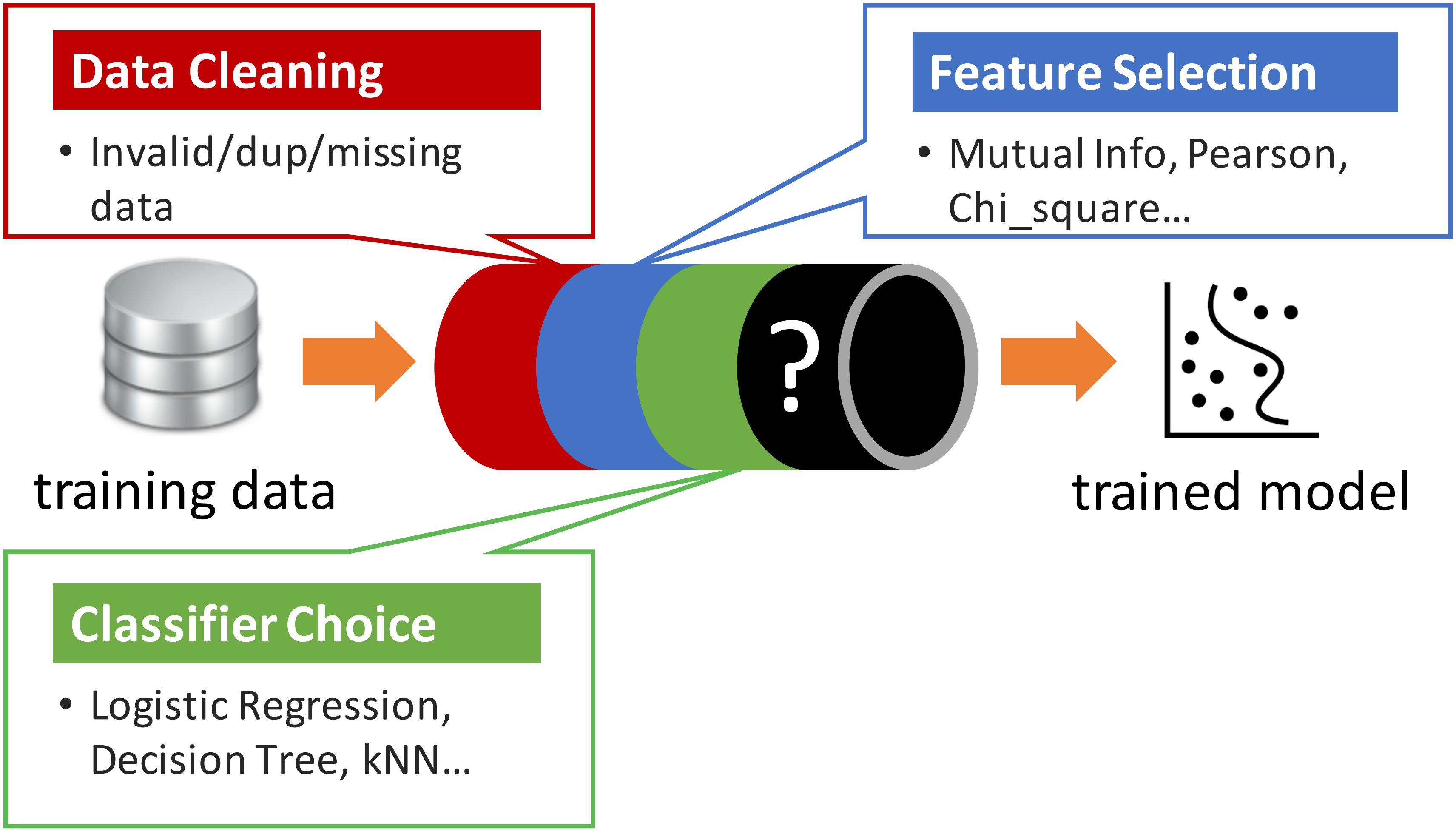
Data Cleaning

- Invalid/dup/missing data

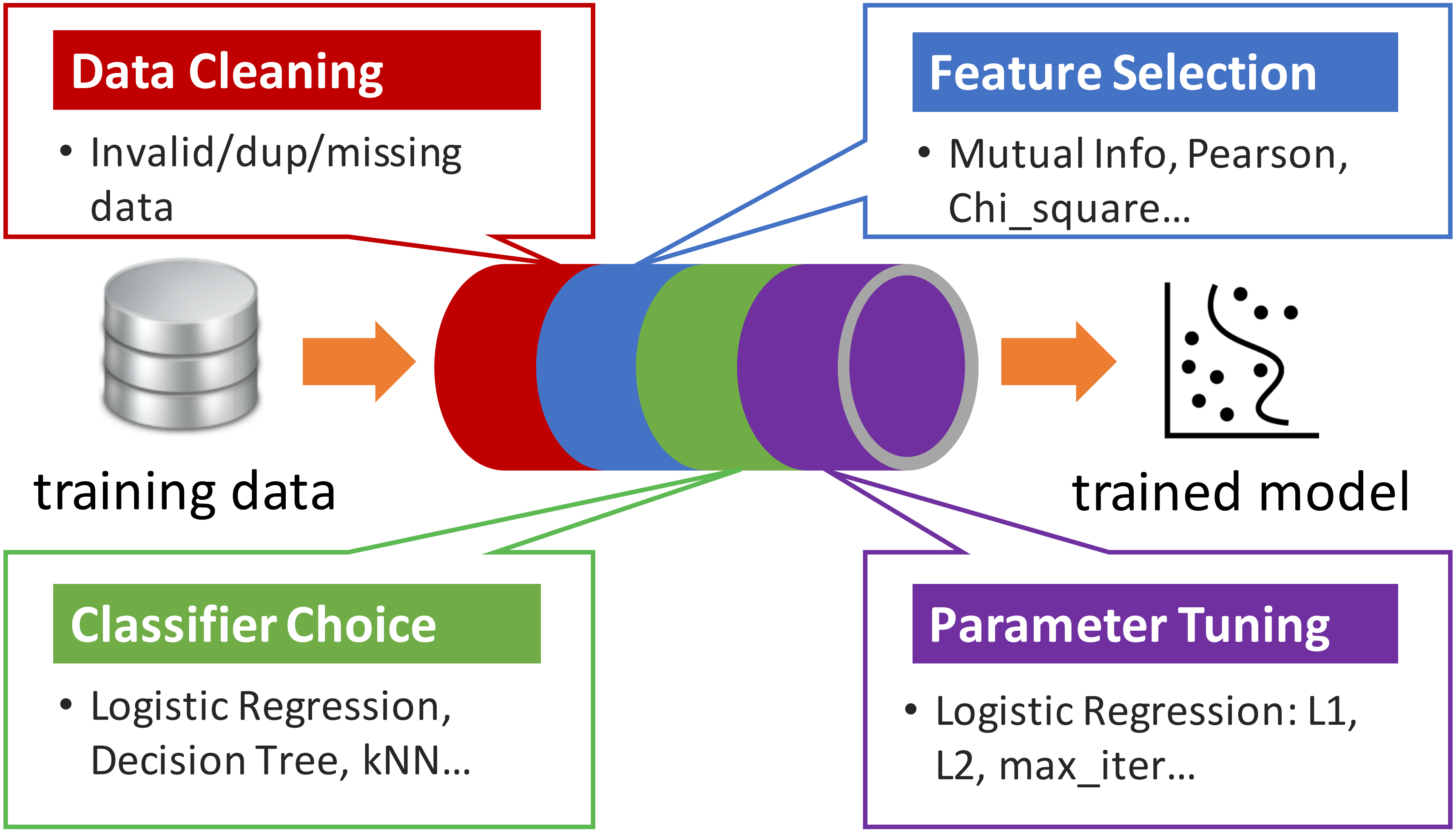
Feature Selection

- Mutual Info, Pearson, Chi…

Control in ML



Control in ML



Control in ML-as-a-Service

	Google	ABM	Amazon	PIO	BigML	Microsoft
Data Cleaning	✗	✗	✗	✗	✗	✗
Tuning	✗	✗	✓	✓	✓	✓

Complexity vs. Performance?

low — user control/complexity — high

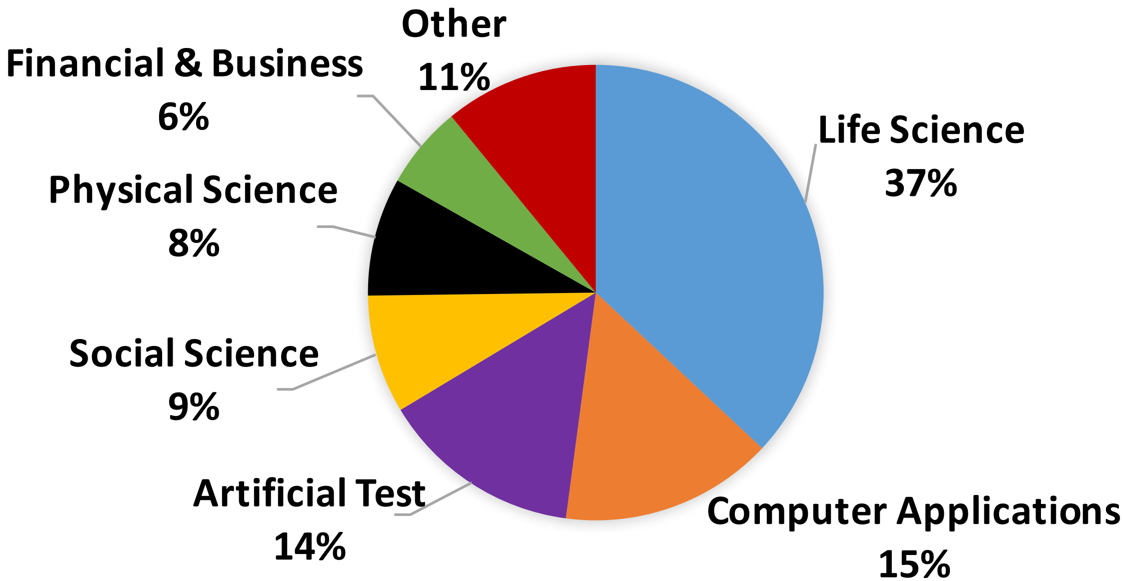
Performance Measurement

Characterizing Performance

- Theoretical modeling is hard
  - Output of ML model depends on dataset
  - No access to implementation details
- Empirical data-driven analysis
  - Simulate a real-world scenario from end to end
  - Need a large number of diverse datasets
- Focus on binary classification

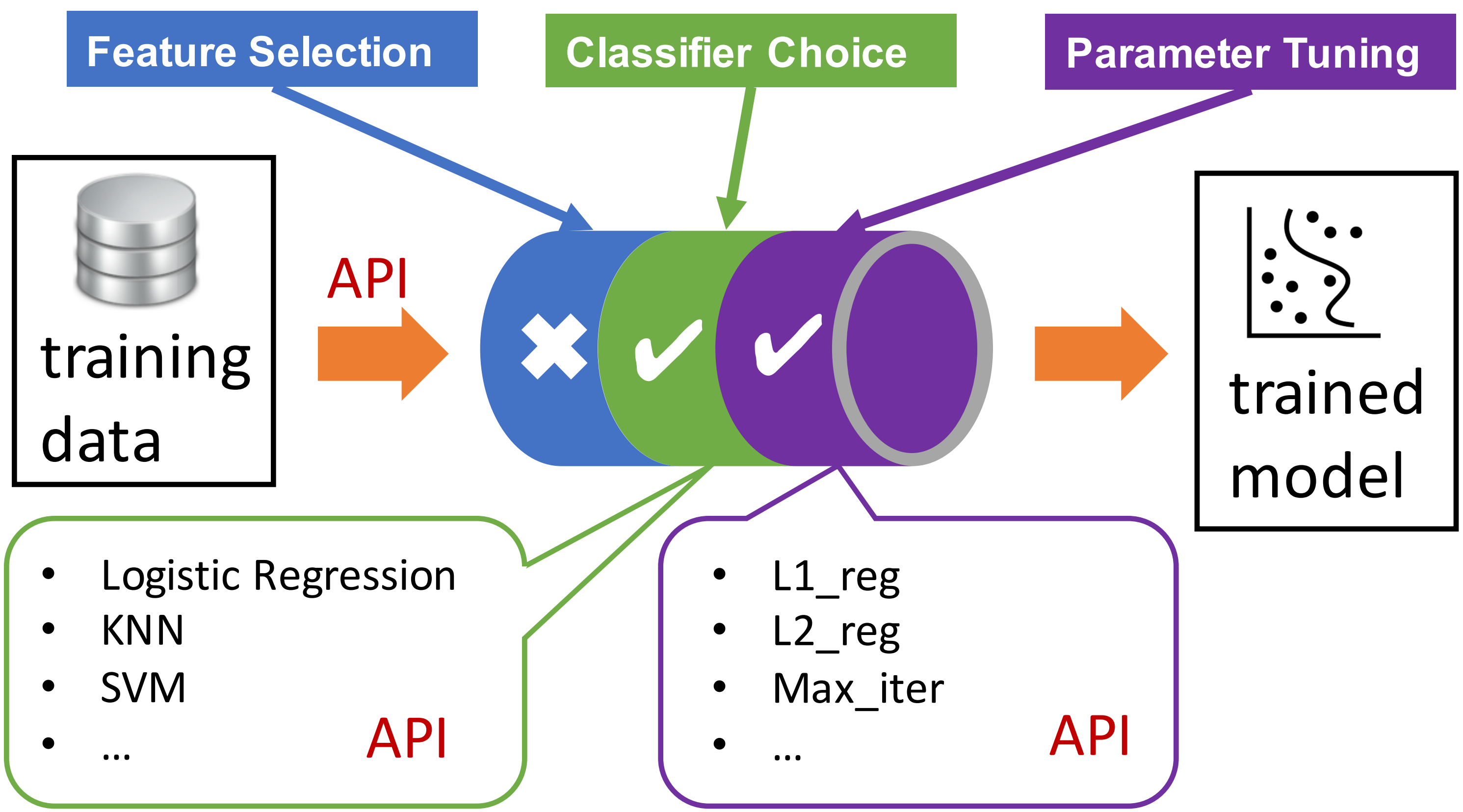
Dataset

- 119 datasets
  - From diverse application domains
  - Sample size: 15 - 245K, number of features: 1 - 4K
  - 79% of them are from UCI ML Repository



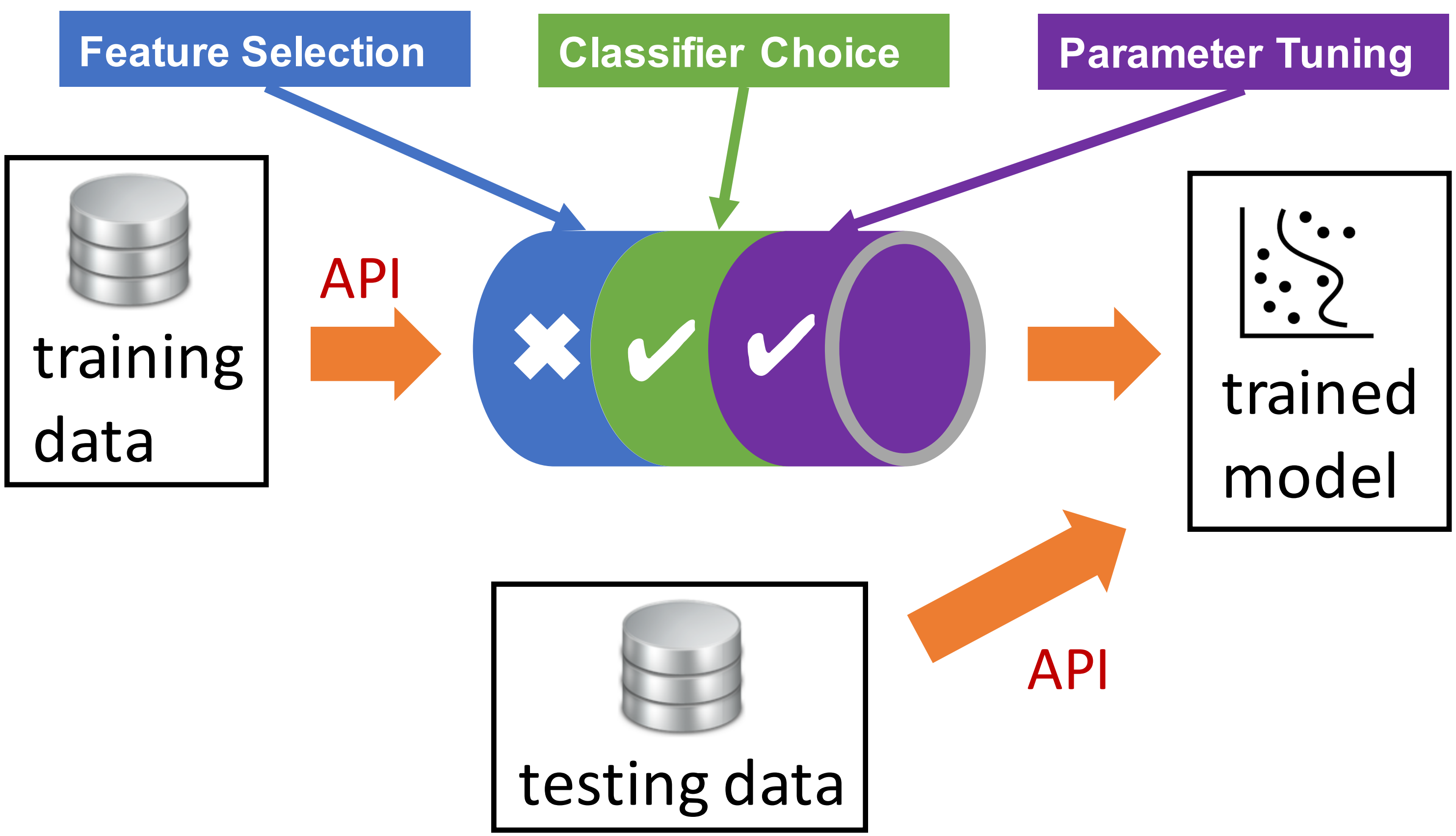
Methodology

- Tune all available control dimensions



Methodology

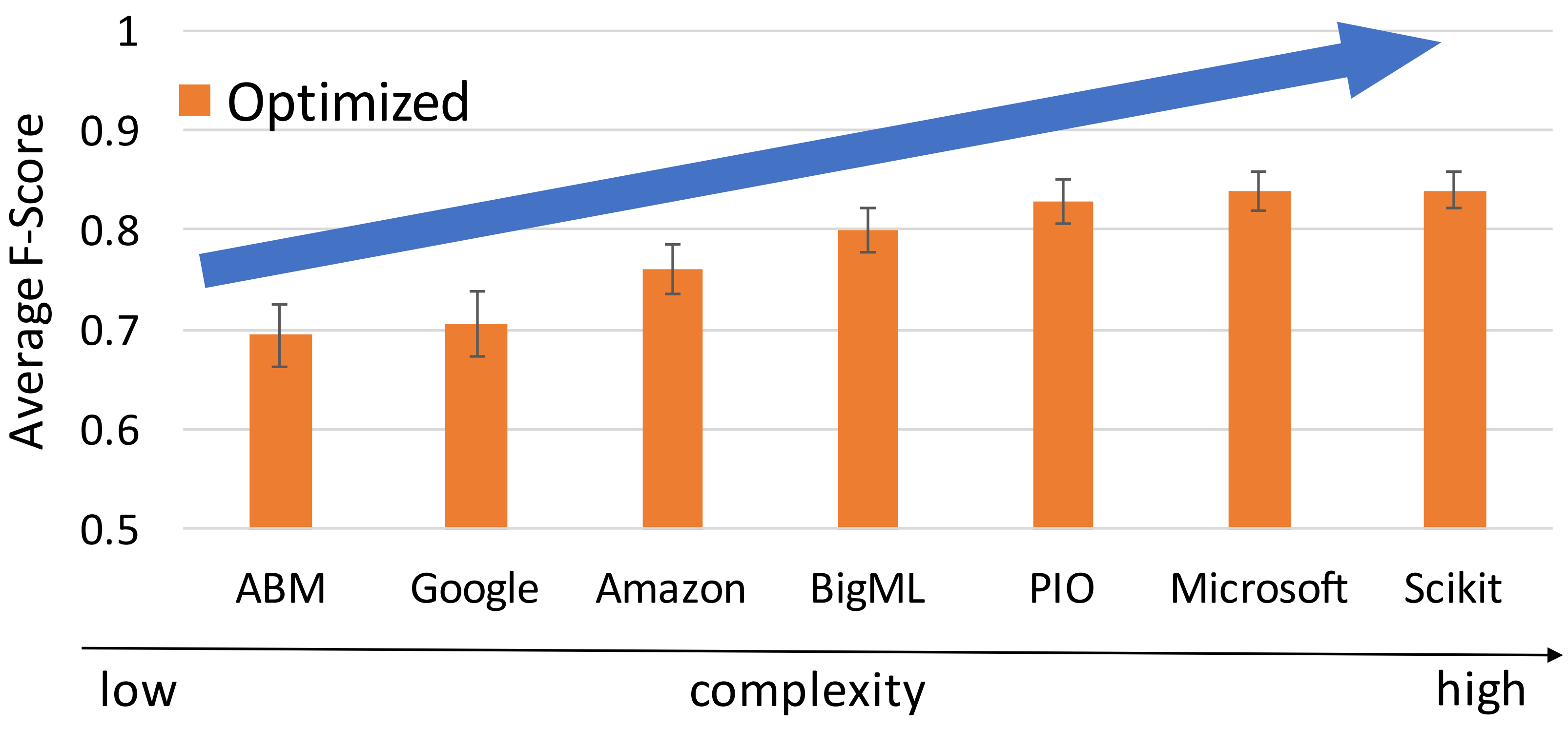
- Tune all available control dimensions



Trade-offs between Complexity and Performance

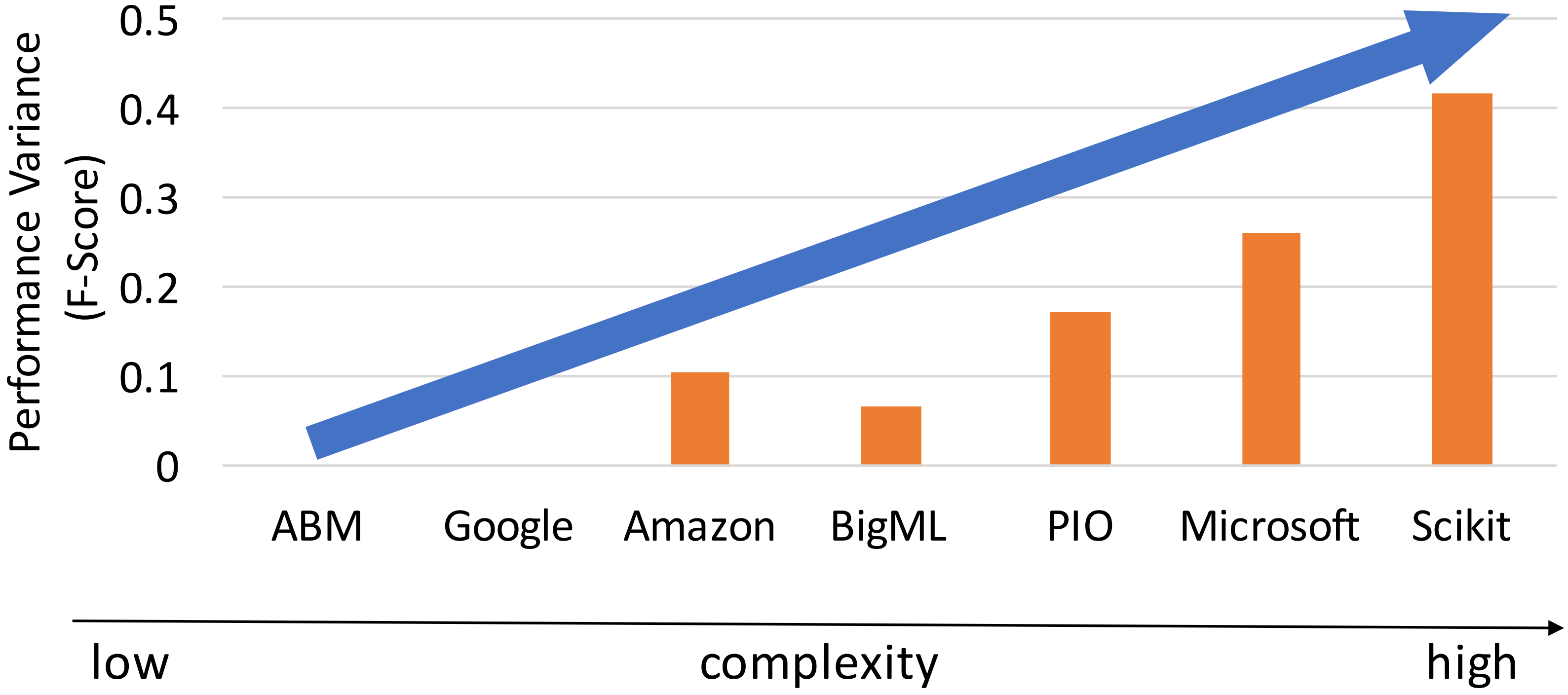
Complexity vs. Performance

- Q: How does the complexity correlate with performance?
  - High complexity -> high performance



Complexity vs. Risk

- Q: How does the risk correlate with complexity?
  - High complexity -> high risk

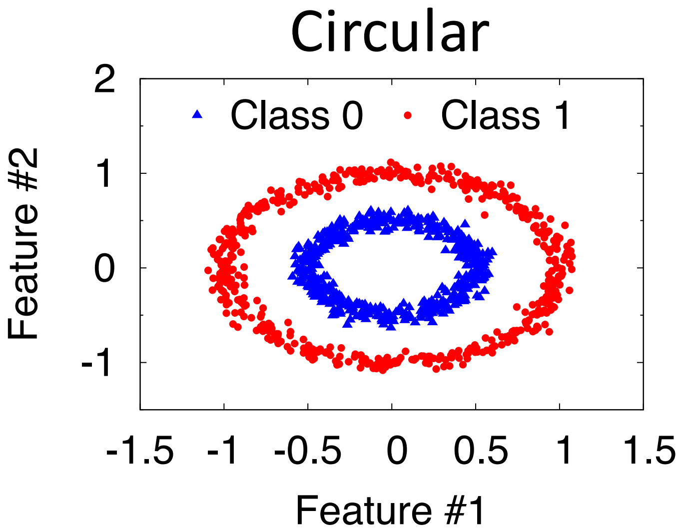


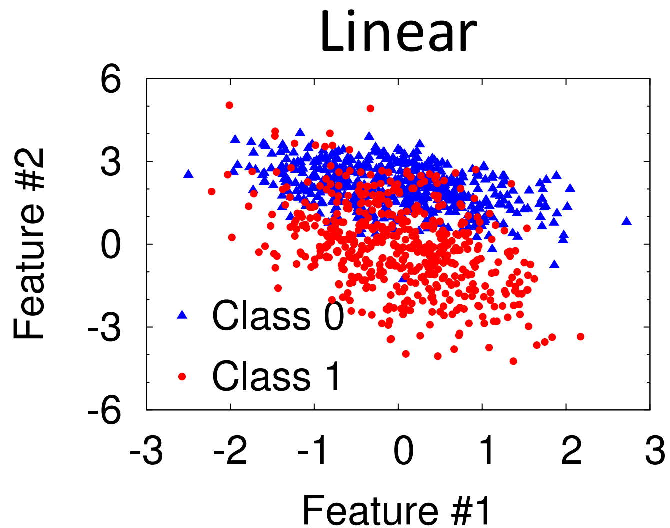
Understanding Server-side Optimization

Reverse-engineering Optimization

- Q: Does server-side adapt to different datasets?

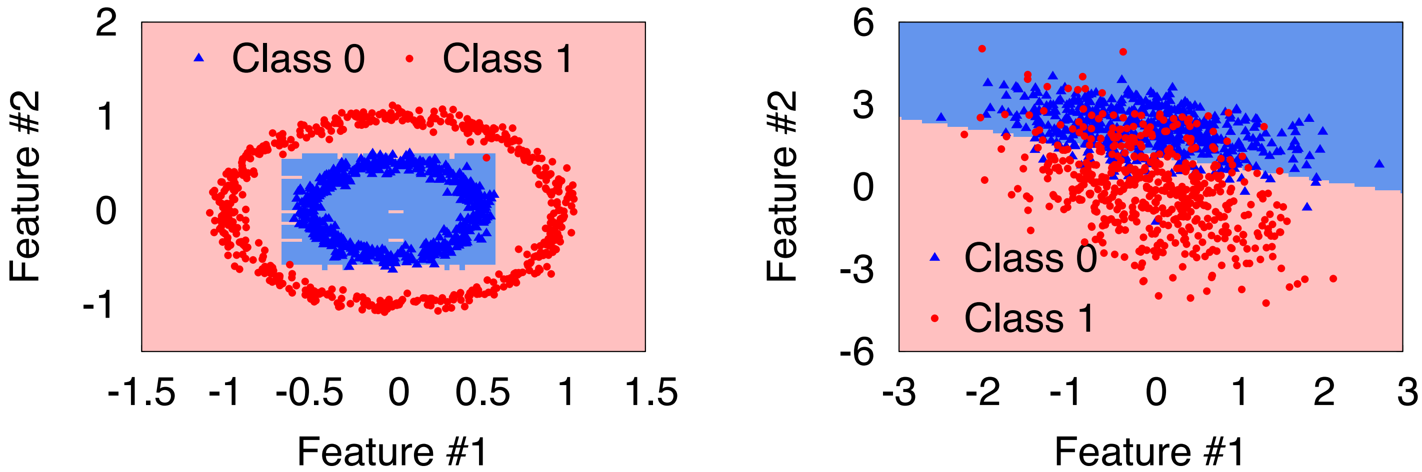
- Reverser-engineering using datasets
  - Create synthetic datasets
  - Use prediction results to infer classifier information





Understanding Optimization

Google decision boundaries



- Google switches between classifiers based on the dataset
- Use supervised learning to infer classifier family used

Takeaways

- ML-as-a-Service is an attractive tool to reduce workload
- But user control still has a large impact on performance
- Fully automated systems are less risky

Thank you!
Questions?