Towards Interactive Curation & Automatic Tuning of ML Pipelines

Carsten Binnig, Benedetto Buratti, Yeounoh Chung, Cyrus Cousins, Tim Kraska, **Zeyuan Shang**, Eli Upfal, Robert Zeleznik, Emanuel Zgraggen

zeyuans@mit.edu

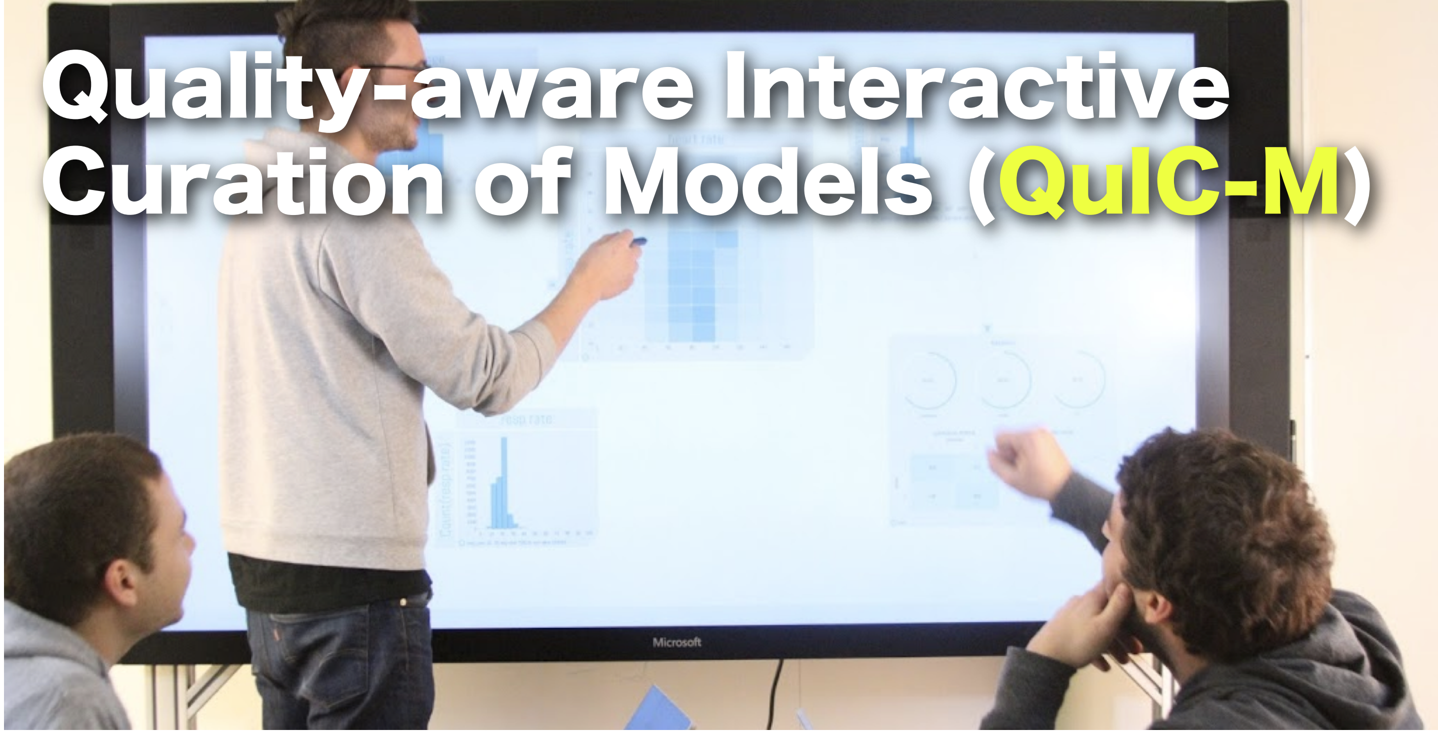
Motivation

Democratizing Data Science comes with challenges

Our Vision

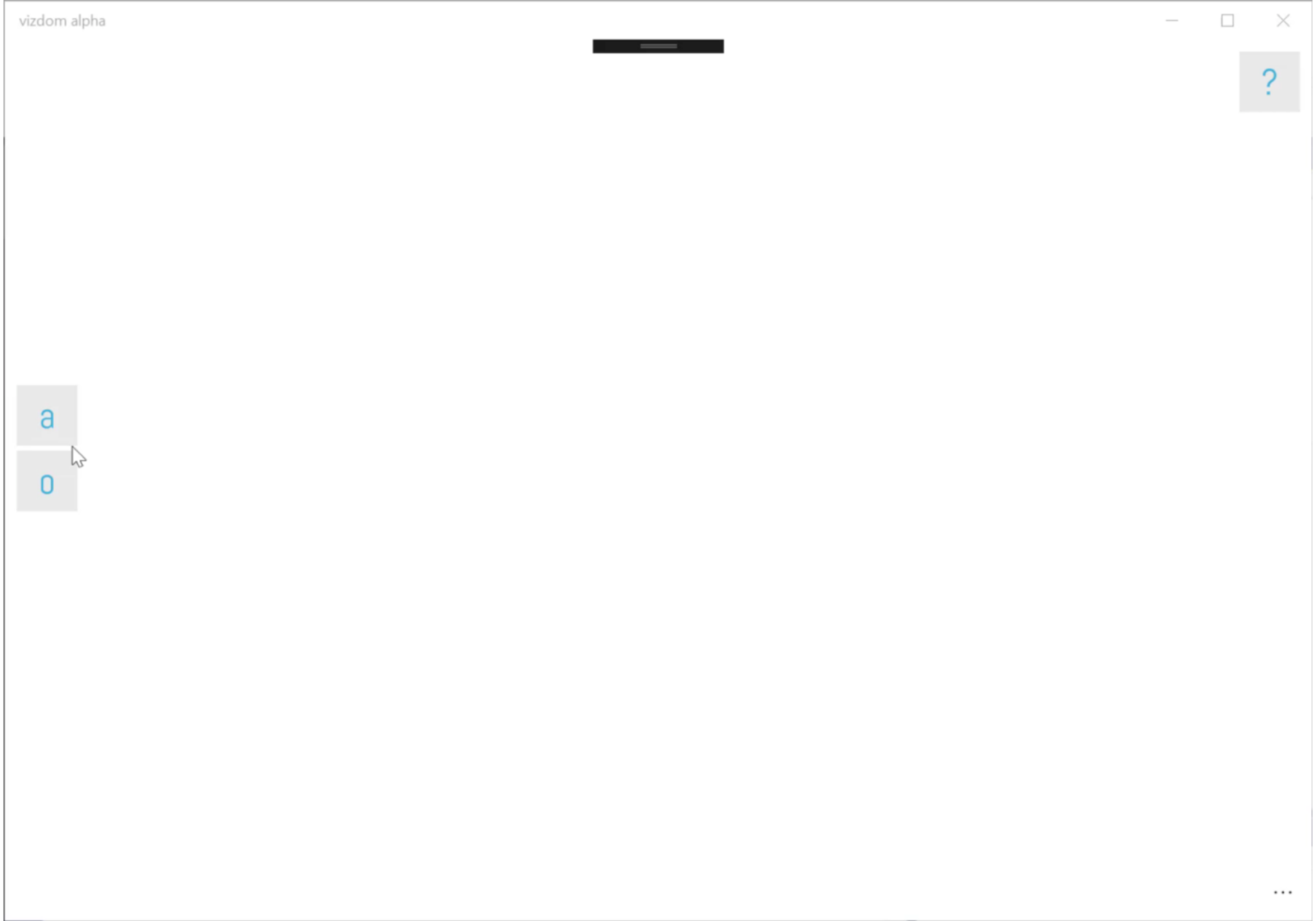
Interactive Data Science Tool to explore data and build models on the fly during a meeting and beyond

Quality-aware Interactive Curation of Models (QuIC-M)



Key Requirements:

- Enable non-experts
- Interactive (first response in seconds, progressive refinement)
- Prevent users from making false discoveries (not part of this talk, see our paper in SIGMOD 2017)


vizdom alpha
?
a
0
...

Related Works

Other AutoML Systems

- Auto-sklearn/ Auto-WEKA
- Spark TuPAQ
- Google Vizier

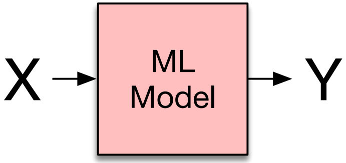
Require

- Well-trained data scientists
- Batch Execution (not interactive)

QuIC-M

- Interactive model exploration for non-experts
- Provides quality-aware curation

Design Goals

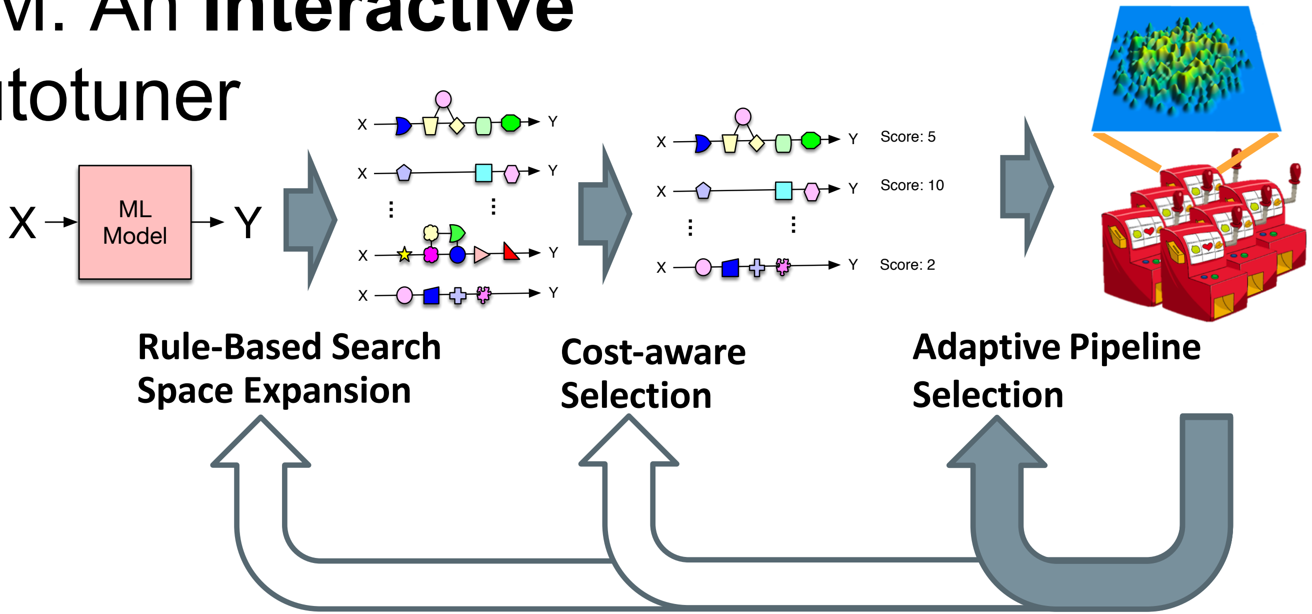


Automation

Progressiveness

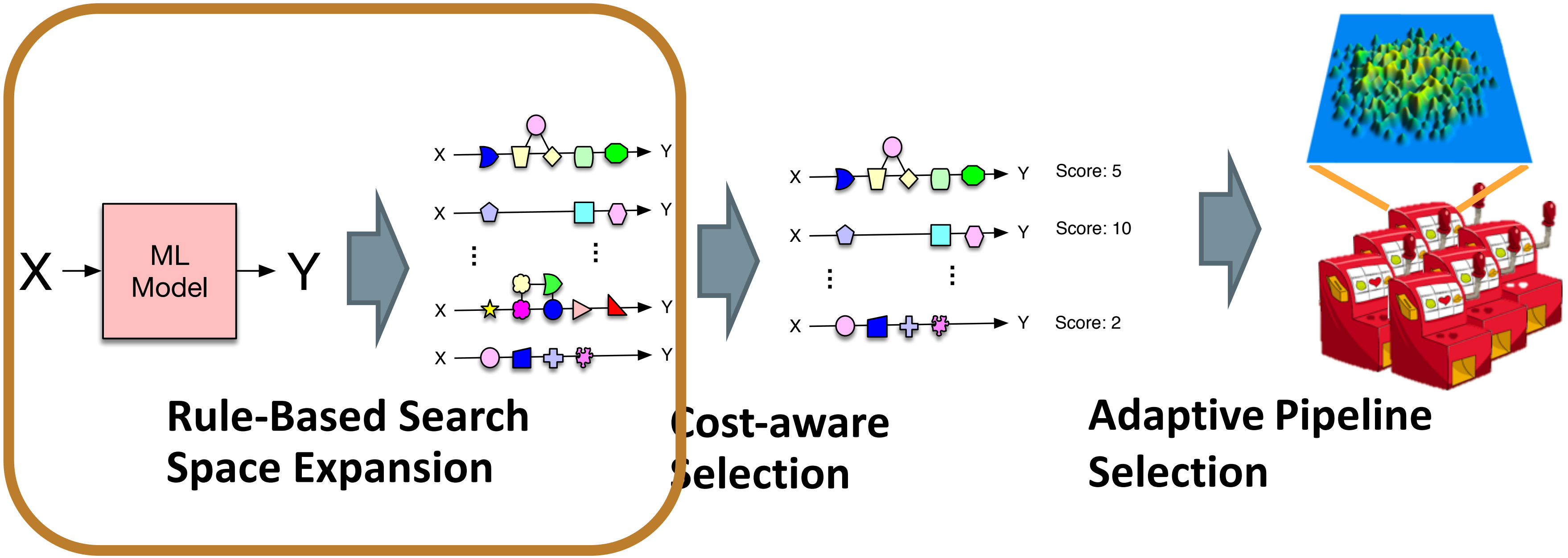
User-steered

QuIC-M: An **Interactive** ML-Autotuner

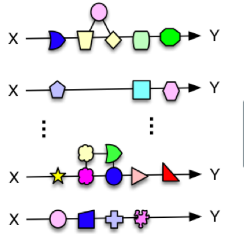


Example: given some stats of a base player, predict whether he will be selected into the hall of fame

Overview of Methods

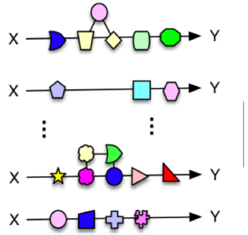


Rule-base Search Space Generation



- Added by experts based on best practices or learned from history
- Primitive Rule: the primitives of pipeline
  - E.g., if numeric feature use MinMaxScaler, StandardScaler
  - E.g., if classification use RandomForest, SVM
- Parameter Rule: the distribution (range) of hyper-parameters of pipeline
  - E.g., if SVM, learning rate is log-uniformly distributed between 0.001 and 1
- Enforcement Rule: validating pipeline
  - E.g., all categorical features should be encoded

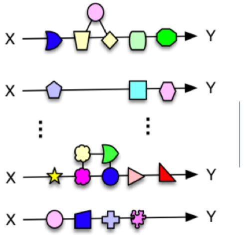
Rule-base Search Space Generation



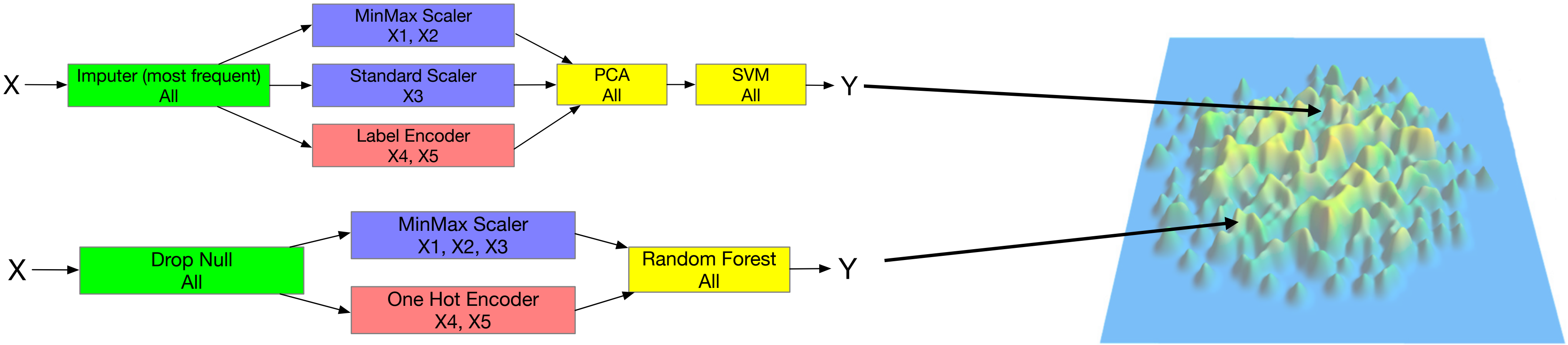
- Execution
  - Find all applicable rules (user-steered)
  - Primitive
  - Parameter
  - Enforcement
- Advantage
  - Easy to incorporate best practices from machine learning experts
  - Flexible to add and update

http://iridia.ulb.ac.be/~fmascia/graphvisualizer

Rule-base Search Space Generation

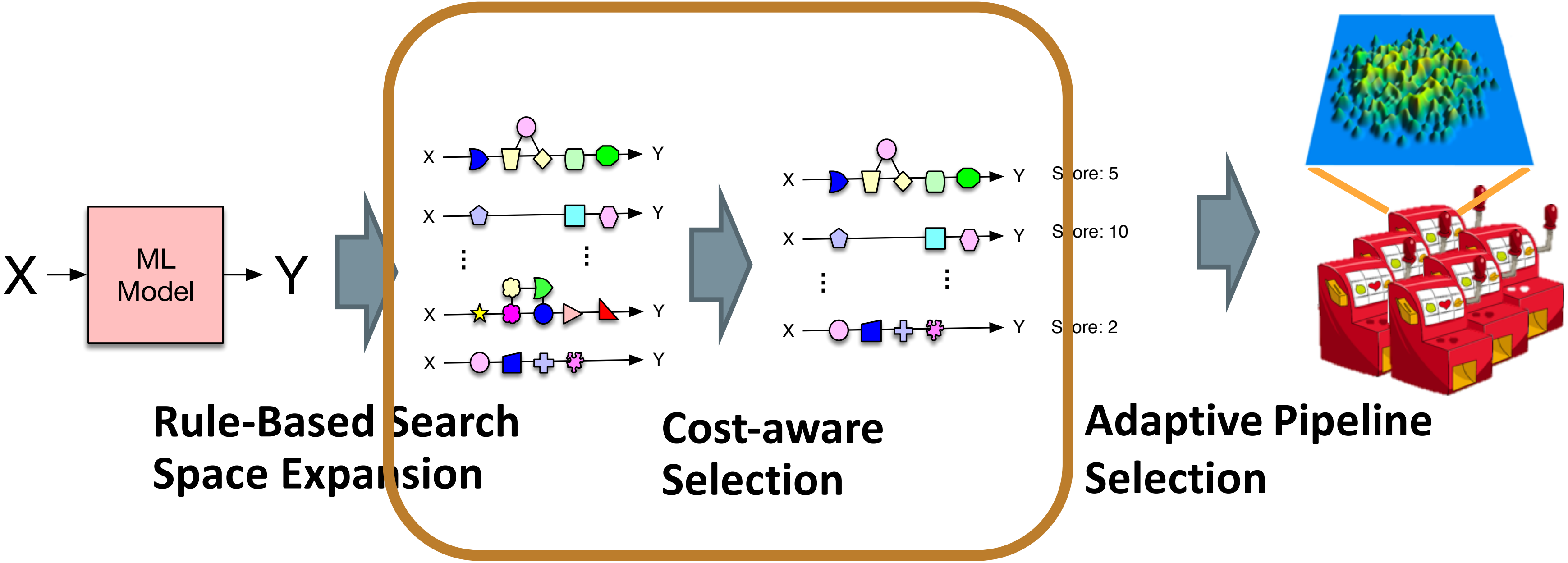


- Example
  - Input (X1,X2,X3) -> Numerical Features
  - Input (X4, X5) -> Categorical Features

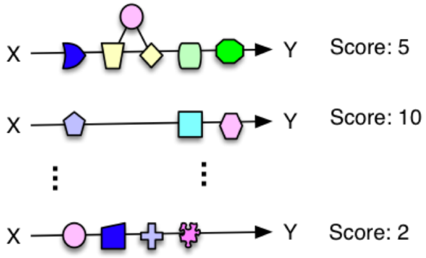


http://iridia.ulb.ac.be/~fmascia/graphvisualizer

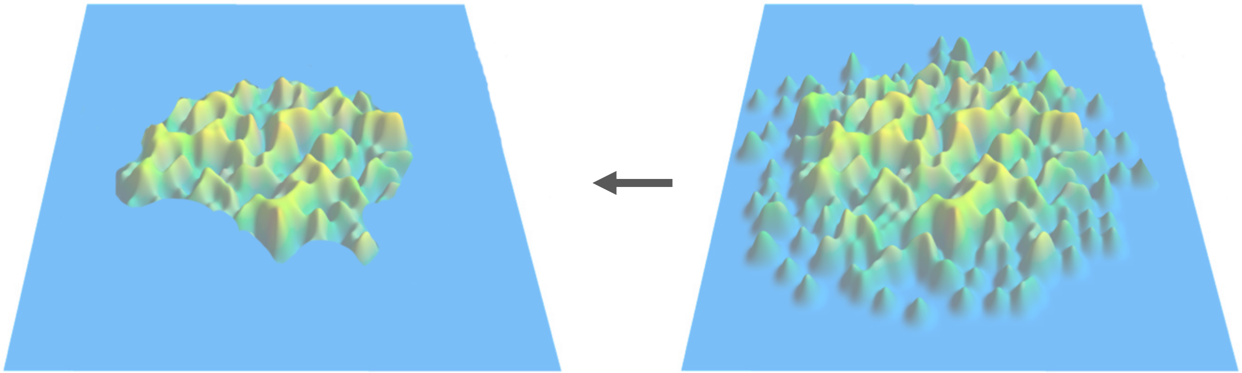
Overview of Methods



Cost-aware Search Space Selection

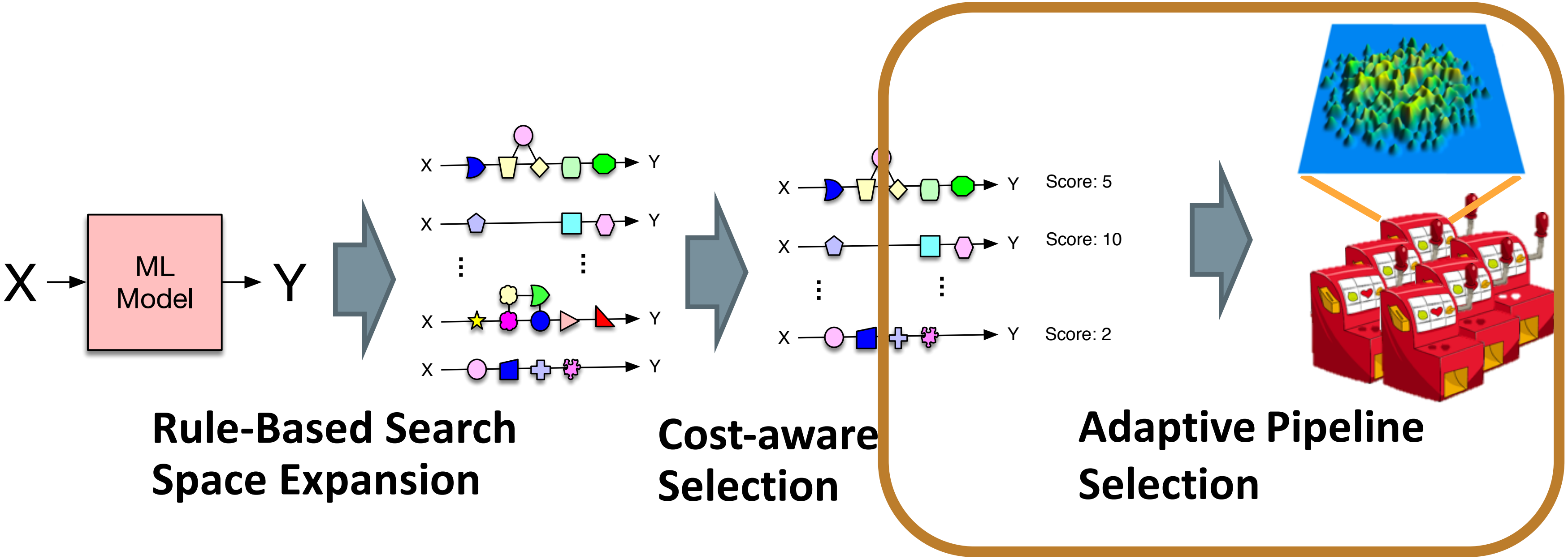


- Pruning the search space
- Input: pipeline, characteristics of data
- Output: cost(running time) and quality estimate
- User-steered
- Based on past history



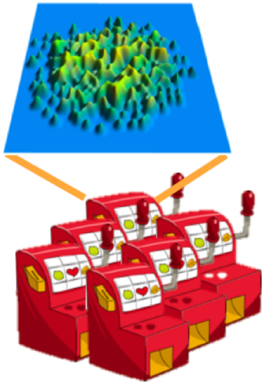
http://iridia.ulb.ac.be/~fmascia/graphvisualizer

Overview of Methods



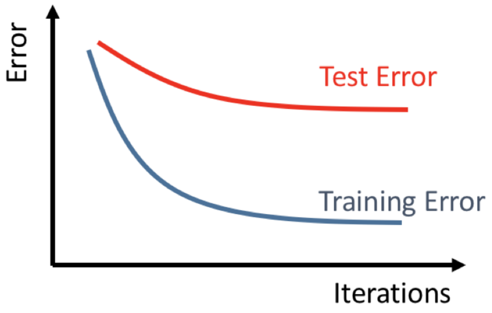
Adaptive Pipeline Selection

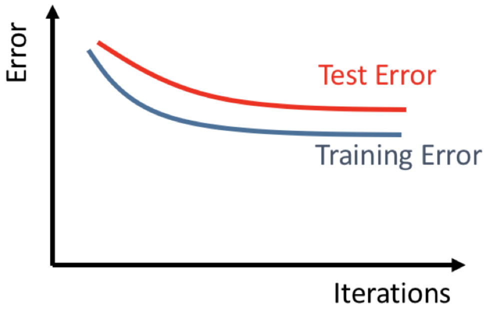
- Algorithm
  - Bayesian Optimization for hyper-parameter tuning
  - Bandit-based method on increasingly larger samples
  - Interactivity



Adaptive Pipeline Selection

- Training vs Test Error as A Signal





Model has **too much capacity** for the amount of data (high variance)

→ Postpone execution to runs with larger sample sizes

Model does not have enough **capacity** (high bias)

→ Prune pipeline (i.e., bandit)

DARPA D3M Competition

- DARPA Data-Driven Discovery of Models (D3M)
- Task: given a data description, predict X
  (e.g., hand-geometry, count crops in images, predict outcome of games,…)
- **For every problem DARPA provided a hand-tuned solution (Baseline)**

	Solved Problems	Better Than Baseline	Normalized Score
Baseline	100%	0%	0.00

DARPA D3M Competition

- Other teams include universities (e.g., UC Berkeley, Stanford, CMU, NYU, Harvard, Johns Hopkins University, University of Chicago, Cornell University, RPI, Tufts University) and companies (e.g,. Uncharted Software, Feature Labs)
- **Most teams involve more than one university/company**

	Solved Problems	Better Than Baseline	Normalized Score
Team MIT/Brown	100%	80%	0.42
Team 1	40%	27%	0.09
Team 2	40%	13%	0.02
Baseline	100%	0%	0.00
Team 3	20%	7%	-0.07
Team 4	87%	47%	-0.16
Team 5	27%	7%	-0.22
Team 6	60%	20%	-0.59
Team 7	87%	53%	-0.75
Team 8	60%	20%	-1.14
Team 9	60%	20%	-4.57

Future Work

- Extension of Rules
- Transfer-learning Opportunities
  - Cost Models
  - Hyper-parameter Tuning
- Execution of Pipelines
  - Caching / Scheduling
- More Benchmarks
- Managing risk (e.g., preventing over-use of hold-out)

Data System for AI Lab DSAIL@CSAIL

Research Area

Data Systems for AI

AI for Data Systems

System Faculty

ML Faculty

Founding Sponsors

Google

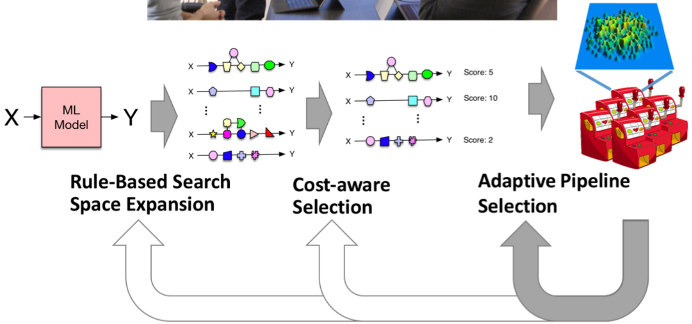
Microsoft

intel

Zeyuan Shang

<zeyuans@mit.edu>

- An interactive curator
- Fully integrated into our data exploration stack Vizdom/IDEA
- Very promising results as part of the DARPA D3M competition



Special thanks to: