MLbase

Tim Kraska <tim_kraska@brown.edu>


1 PetaByte reported every second by LHC

My Hidden Motivation

Why is it so damn hard?

Everybody thinks about

Data

...not Queries

Tool complexity

Multi-hypotheses
Pitfall

Brown Projects

DBNav

H-Store

SciDB

Data Tamer

CrowdDb

MLbase

TupleWare

CrowdDb

MLbase

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View CA	$210Bn
Intl. Business Machines	Armonk, NY	$200Bn
Microsoft	Redmond, WA	$250Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = "IBM"
```

```
Number of Rows: 0
```

```
Problem:
Entity Resolution
```

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View CA	$210Bn
Intl. Business Machines	Armonk, NY	$200Bn
Microsoft	Redmond, WA	$250Bn

```
SELECT Market_Cap
From Companies
where Company_Name = “Apple”
```

```
Number of Rows: 0
```

Problem:
Missing Data

DB-hard Queries

```
SELECT Image
From Pictures
Where Image contains
“professor with beard”
```



Number of Rows: 0

Problem:
Missing Intelligence

Easy Queries

```
SELECT Image
From Pictures
Where Image contains
“professor with beard”
```

Micro-Task CrowdSourcing

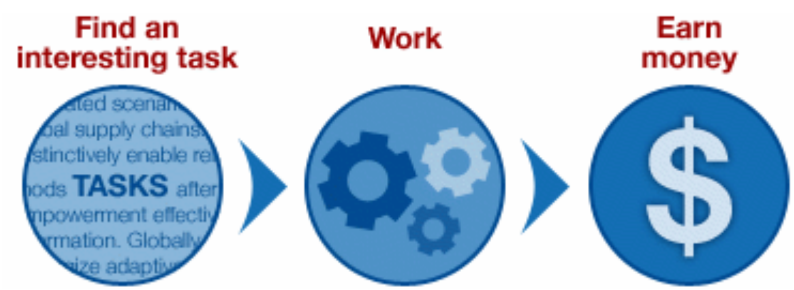
amazon mechanical turk™
Artificial Artificial Intelligence

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get started.

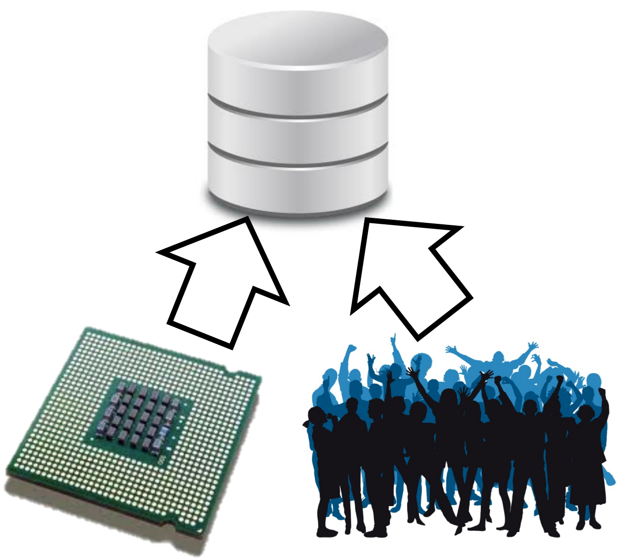
As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Get Started

Overview

Problem



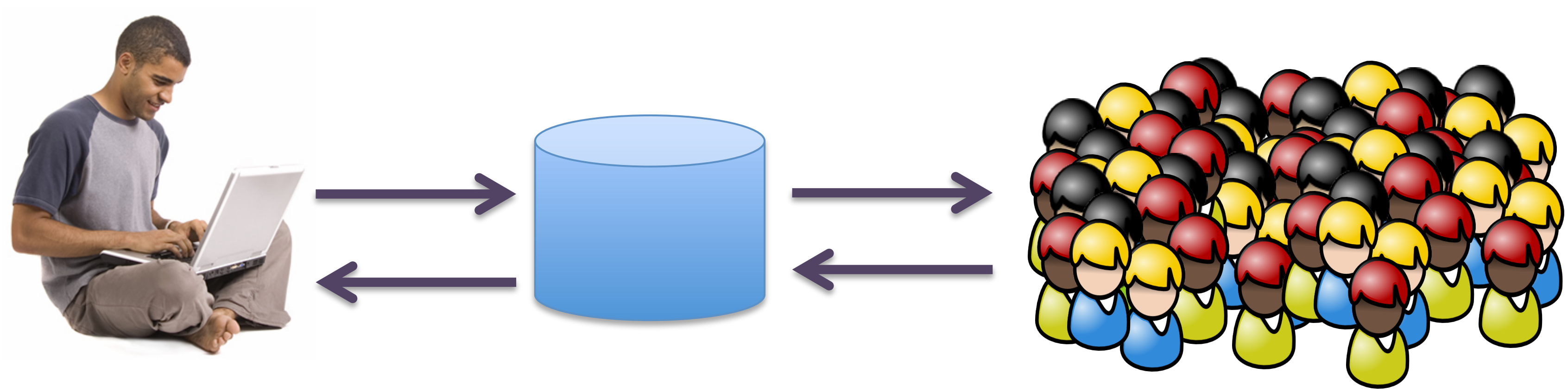
- How to integrate this new resource "humans" for DB-hard queries
- How to ensure high-quality results

Contributions

- **CrowdDb Systems**
  - Architecture
  - Query language
  - Query execution
- **Quality Control for Sets**

Queries in the Open World

```
CREATE CROWD TABLE PEOPLE(name,
age, picture, beard, occupation)
```



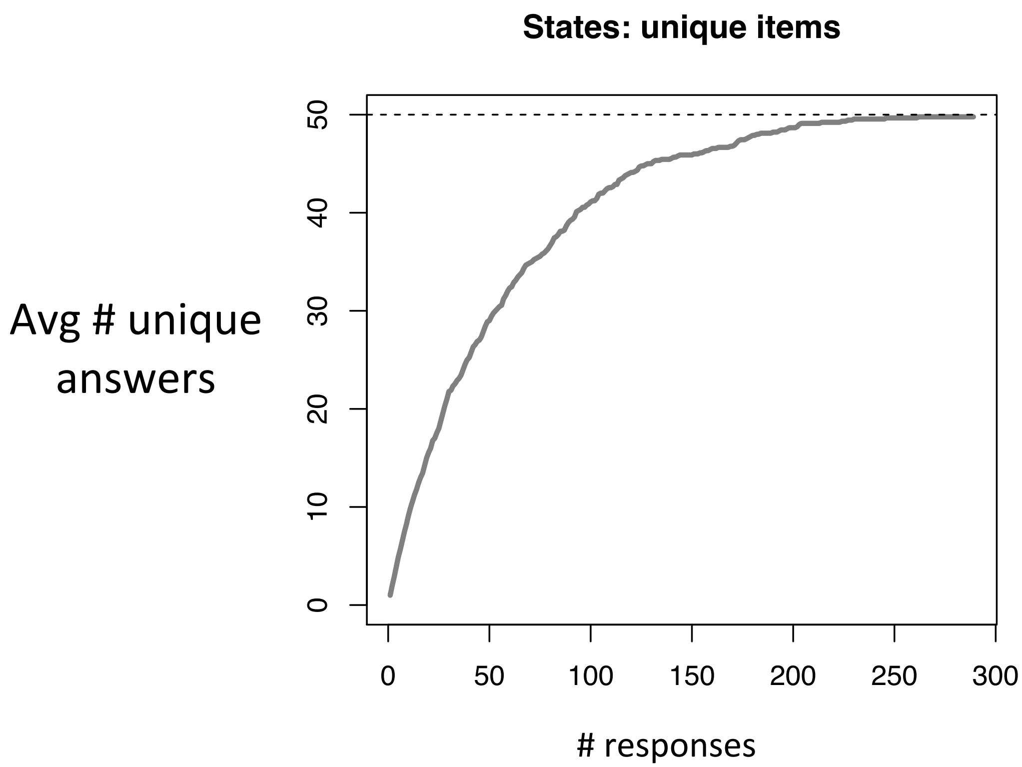
Big Questions

When should we **stop** asking **questions**?

Can we **estimate** query **result set size**?

Querying the crowd

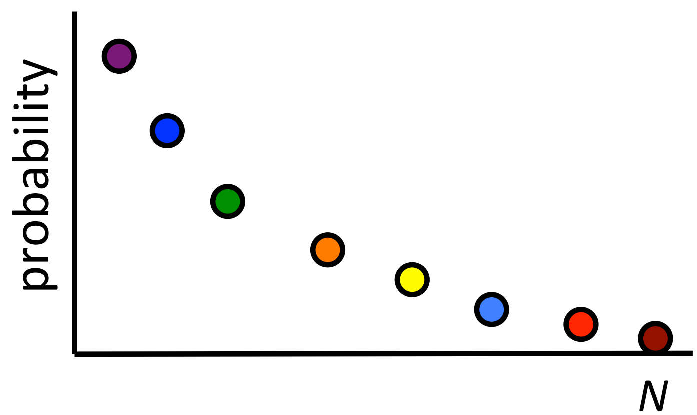
- SELECT name FROM US_States
  - Experiment runs on Mechanical Turk
  - Avg. “accumulation curve”

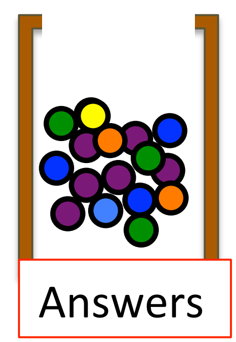


Species estimation

Species estimation

- Sample drawn from a population
  - There are ***N*** different types within the population, ***N*** unknown
  - Analog: worker answers are samples from item distribution

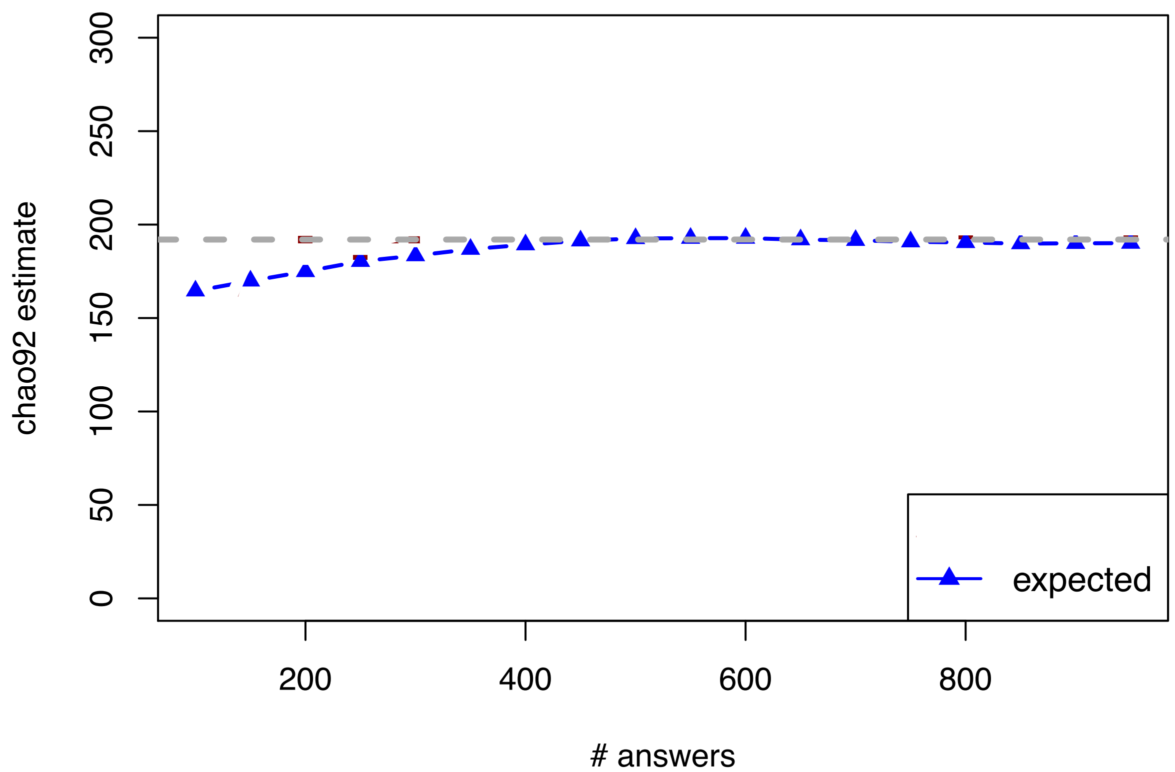




- Calculate query progress
  - based on estimate of ***N***
  - Use *Chao92* estimator, suitable for open-world

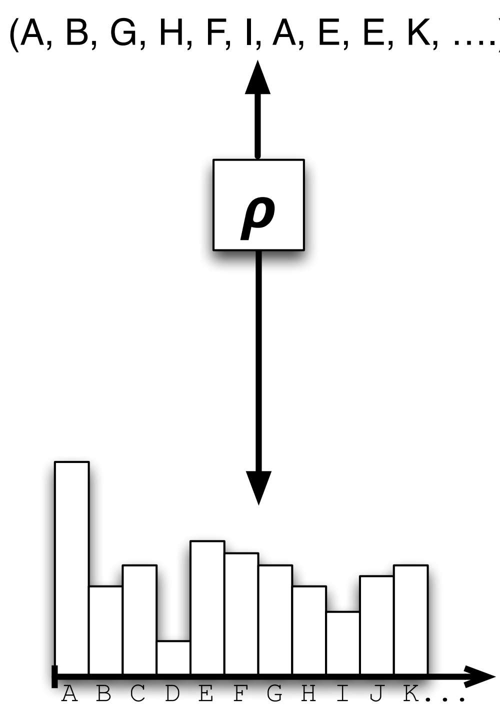
Worker behavior: example

- United Nations member countries (192)
  - Simulated vs. actual cardinality estimate



Worker behavior

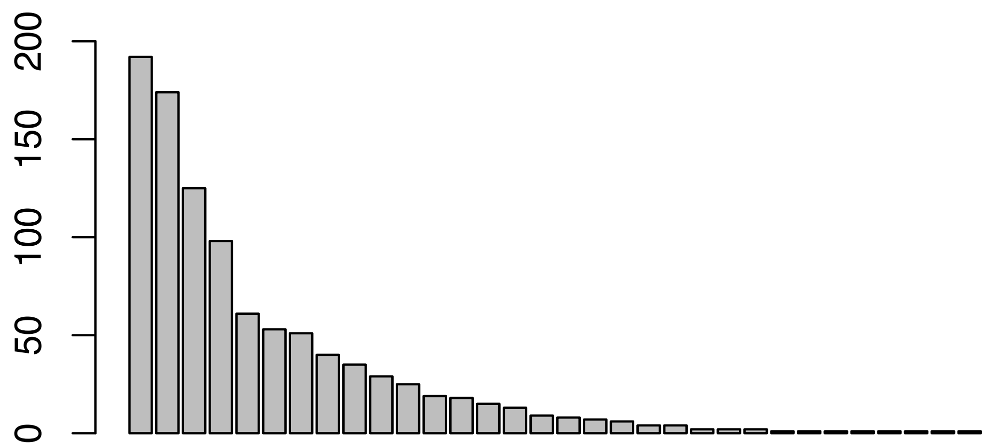
ρ = sampling process **with replacement**
λ = sampling process **without replacement**



(a) Database Sampling

“Streakers” [Heer10]

answers



workers

Streakers provide a lot of unique answers

Streaker-tolerant estimator

- Chao92 estimator
  - Non-parametric, “frequency of frequencies” statistic
    - f_1 = singletons, f_2 = doubletons, f_0 = unobserved
    - Uses notion of *sample coverage*: $\hat{C} = 1 - f_1/n$

$$\hat{N}_{chao92} = \frac{c}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\hat{\gamma}^2$$

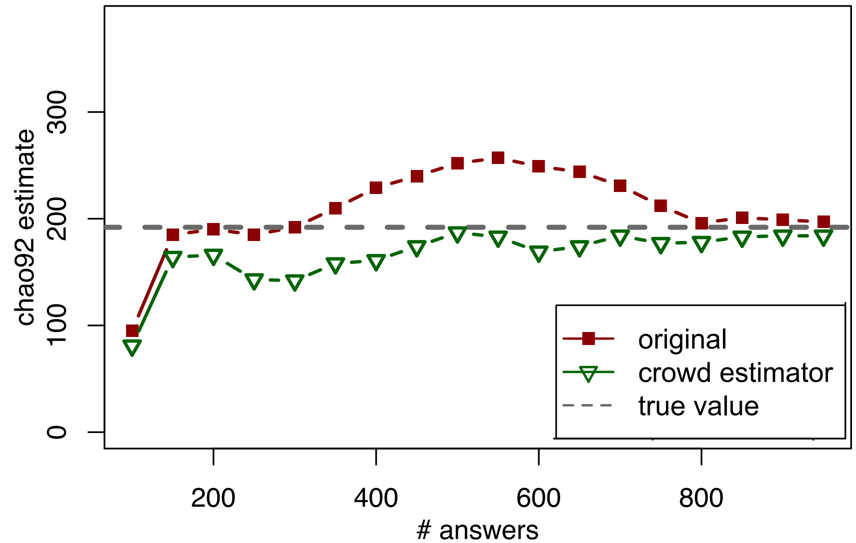
- Adding streaker-tolerance
  - Estimator over-predicts cardinality with abundance of unique answers (f_1)
  - Remove f_1 outliers

$$\hat{N}_{crowd} = \frac{cn}{n - \sum_i min(f_1(i), 2\hat{\sigma}_i + \bar{x}_i)}$$

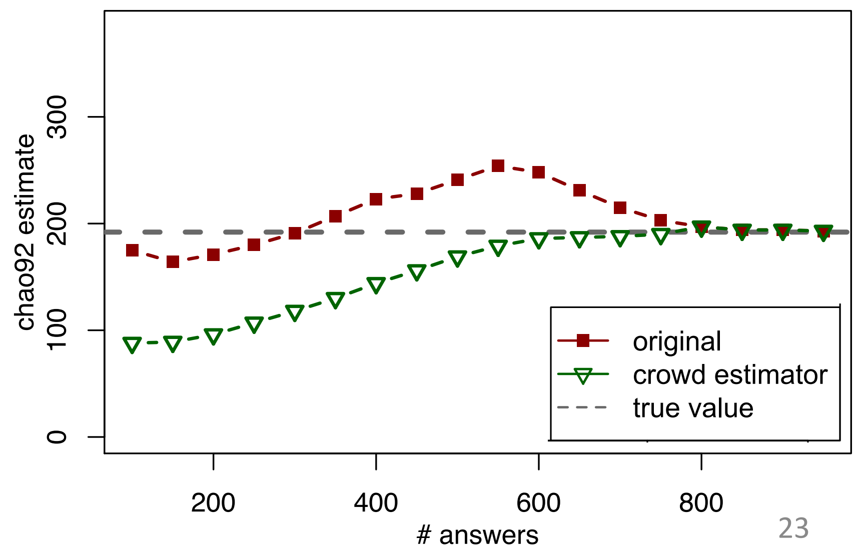
with coefficient of variance = 0

Streaker-tolerant estimator: results

- “UN member nations” (run 1)
  - Streaker during the middle ameliorated



- “UN member nations” (run 2)
  - Streaker at beginning
  - Other workers shared skewed distribution, yields low cardinality estimate



Now that we have the data…

…how do we analyze it

The Little Secret

Machine Learning is like Teenage Sex

- Everybody talks about it
- Nobody knows how to do it
- Everyone thinks everyone else is doing it
- So everyone claims they are doing it

The Problem

What you *want* to do

Build a Classifier

What you have to do

- **Learn the internals of ML classification algorithms, sampling, feature selection, X-validation,….**
- **Potentially learn Spark/Hadoop/…**
- **Implement 3-4 algorithms**
- **Implement grid-search to find the right algorithm parameters**
- **Implement validation algorithms**
- **Experiment with different sampling-sizes, algorithms, features**
- **….**

and in the end

Ask For Help

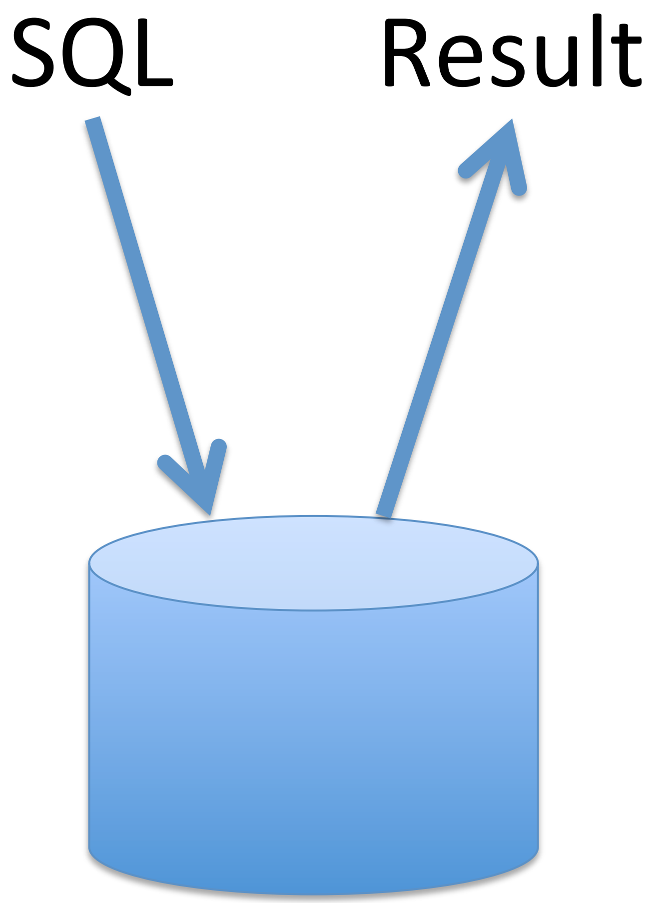
1st Goal: Simplify the use of ML algorithms

2nd Goal: Make it easier to implement distributed ML algorithms

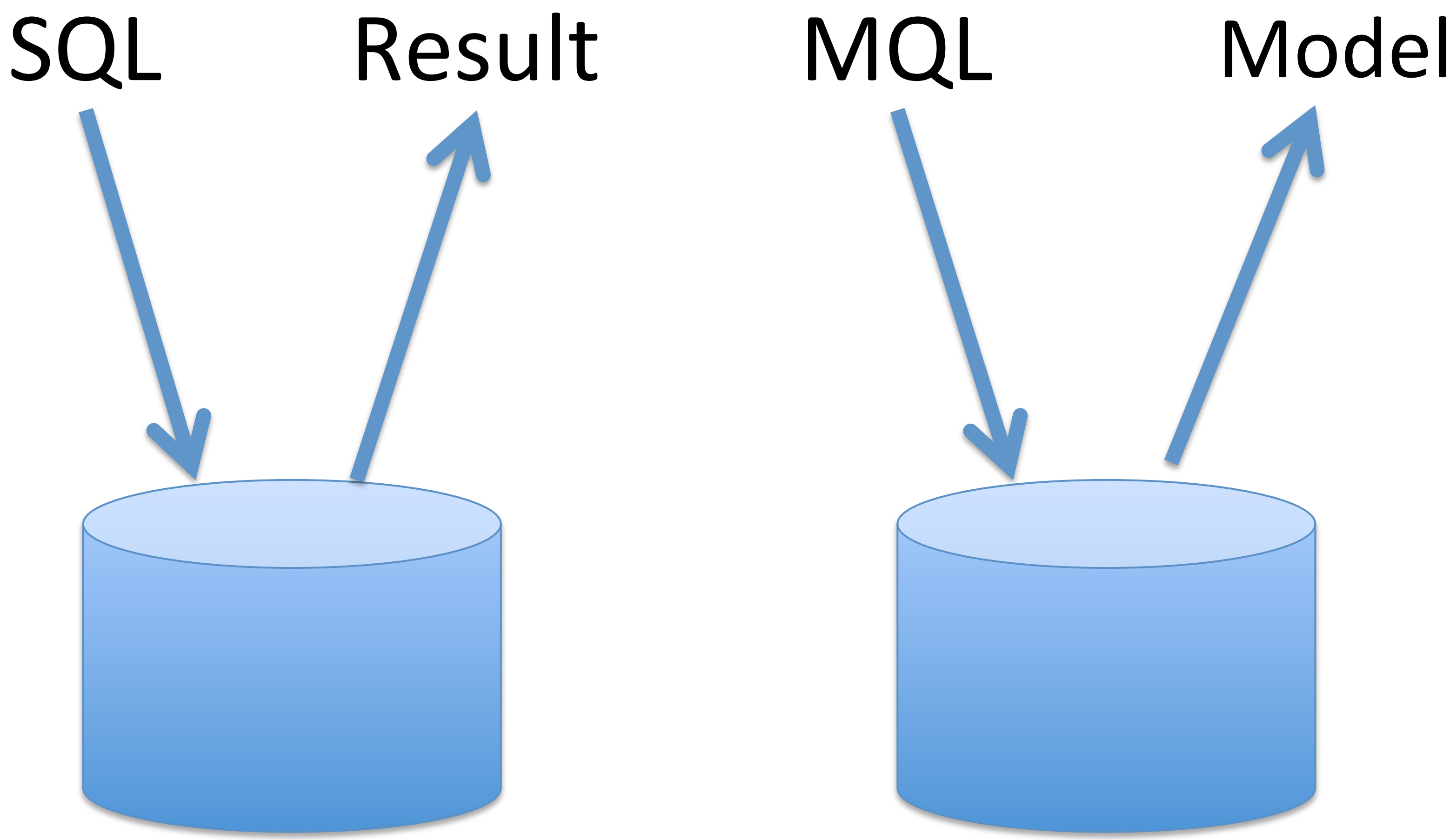
Collaborators

and others…..

A Declarative Approach to ML



A Declarative Approach to ML



Use Cases

Supervised Classification: ALS Prediction

```
var X = load("als_clinical", 2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) = top([illegible], y), 5min)
```

Unsupervis[illegible]e Extraction: Twitter

```
var G = lo[illegible]witter_network")
var hubs-n[illegible] = findTopKDegreeNodes(G, k = 1000)
var text-features = textFeaturize(load("twitter_tweet_data"))
var T-hub = join(hub-nodes, "u-id", text-features, "u-id")
findTopFeatures(T-hub)
```

Algorithm Independence

Use Cases

Supervised Classification: ALS Prediction

```
var X = load("als_clinical", 2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) = top(doClassify(X, y), 5min)
```

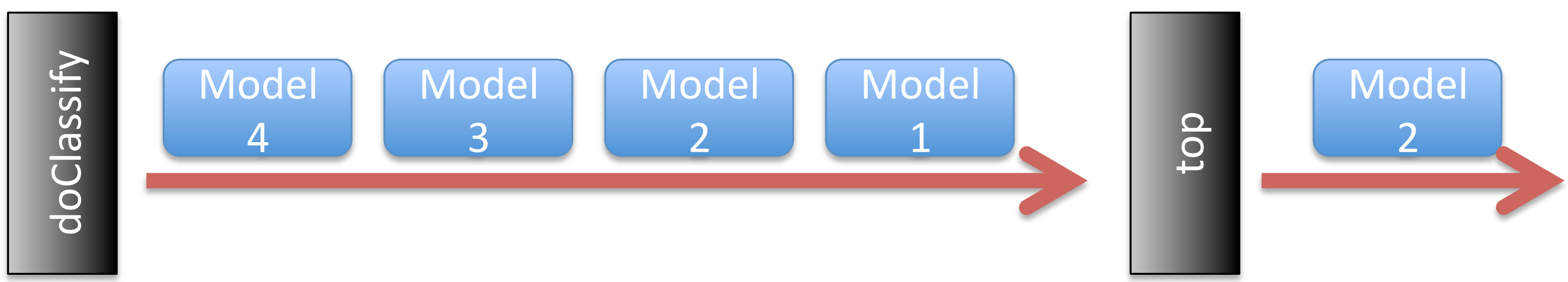
Hints

Supervised Classification: ALS Prediction

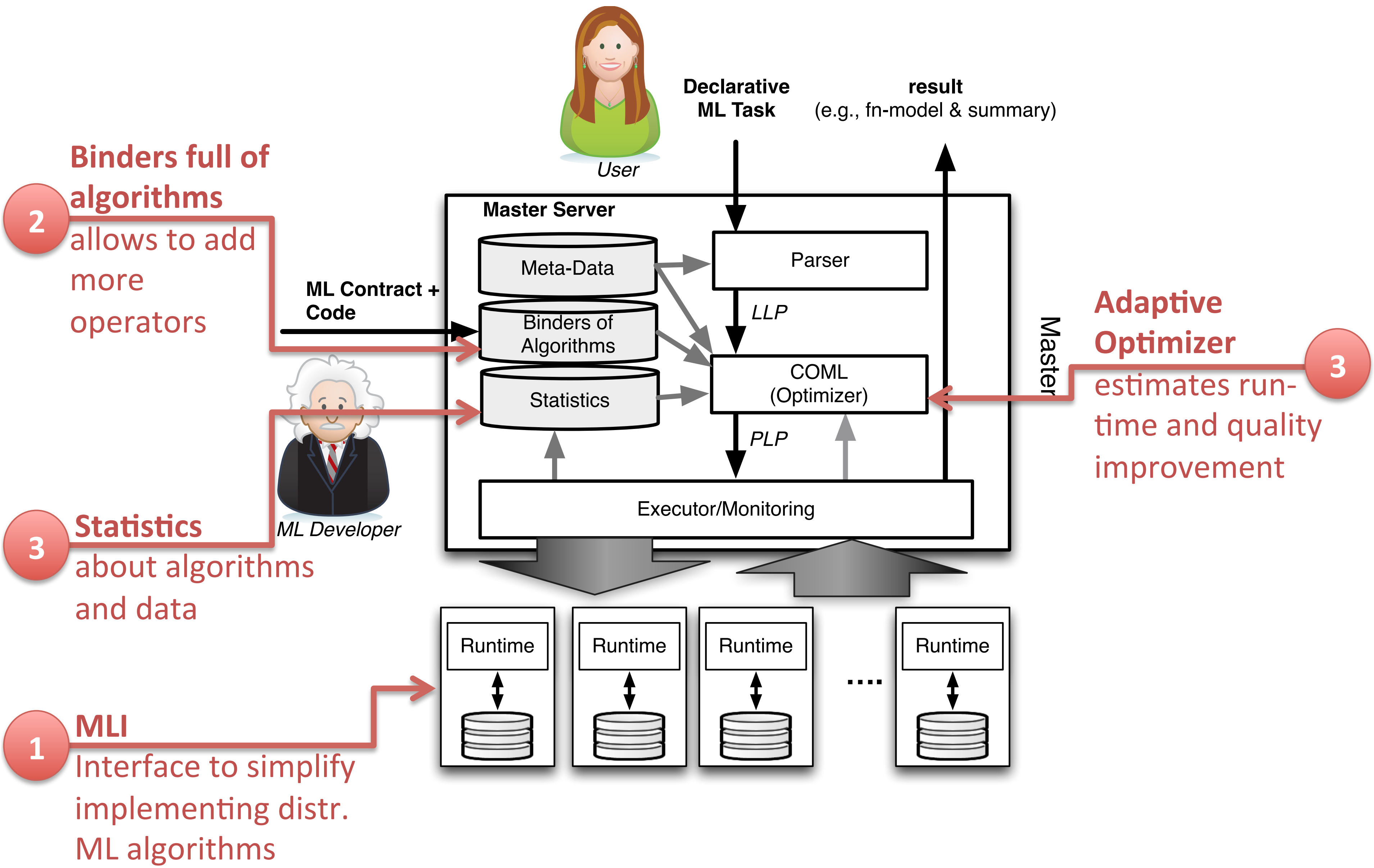
```
var X = load("als_clinical", 2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) = top(doClassify(X, y, SVM), 5min)
```

Streaming-like Data Model

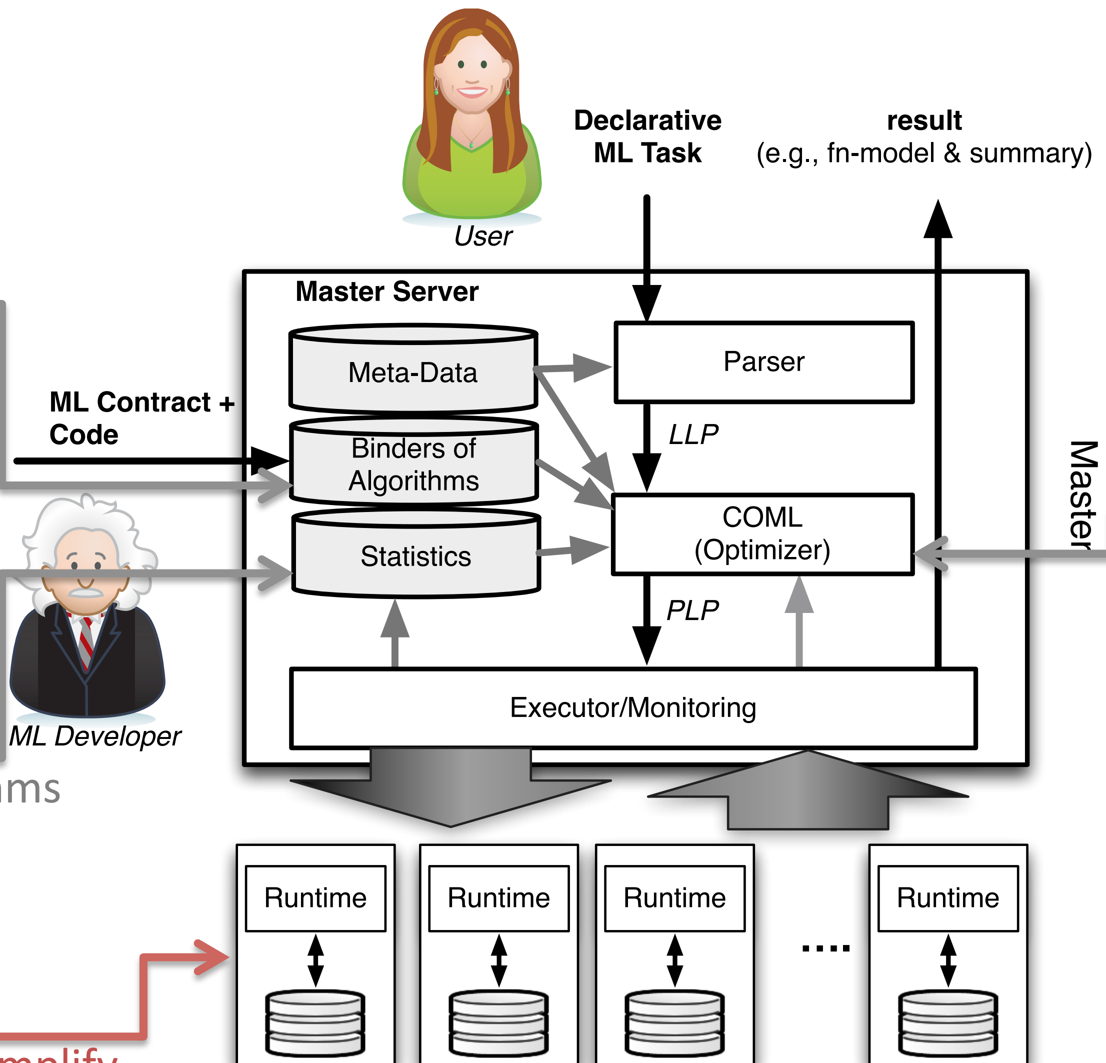
Infinite ordered stream of items, being either models (i.e., higher-ordered functions) or tuples



MLbase Architecture



MLbase Architecture



2 **Binders full of algorithms** allows to add more operators

3 **Statistics** about algorithms and data

1 **MLI** Interface to simplify implementing distr. ML algorithms

Adaptive Optimizer estimates run-time and quality improvement 3

1 MLI: Machine Learning Interface

- Shield ML Developers from low-level-details: provide familiar mathematical operators in distributed setting
- Physical independence between ML algorithm and run-time
- Initial abstractions: MLTable, MLMatrix, MLOpt
- Current supported run-times:

Spark

TupleWare

MLTable

- **Flexibility when loading data**
  - e.g., CSV,JSON,XML
  - Heterogeneous data across columns
  - Missing Data
  - Feature extraction
- **Common Interface**
- Supports MapReduce and Relational Operators
- Inspired by DataFrames (R) and Pandas (Python)

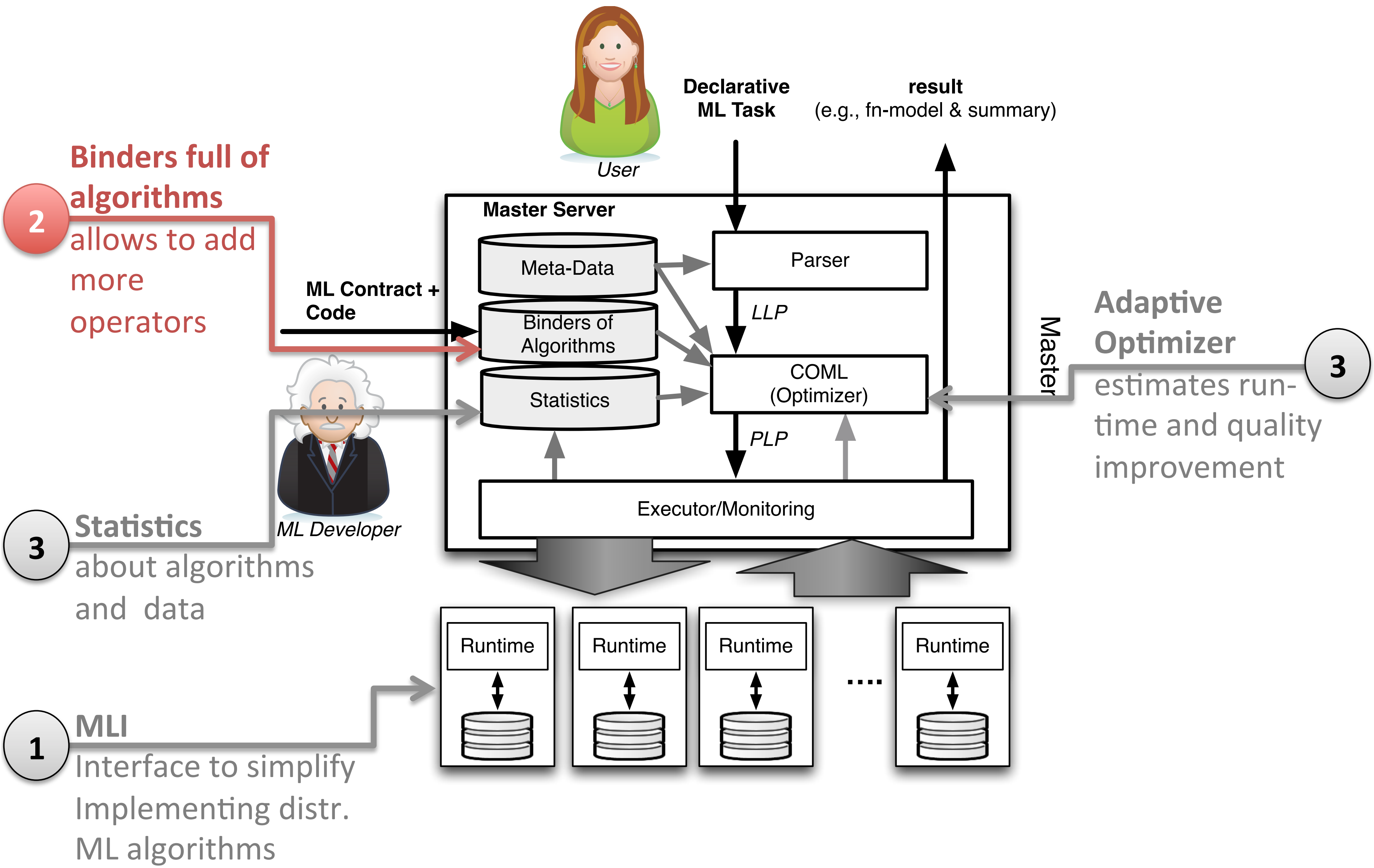
MLSubMatrix

- **Linear algebra on local partitions**
  - E.g.,matrix-vector operations for mini-batch logistic regression
  - E.g., solving linear systems of equations for Alternating Least Squares
- Sparse and Dense Matrix Support

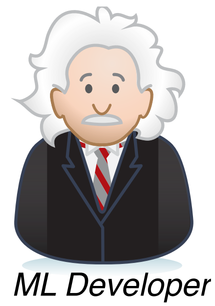
MLSolve

- **Distributed implementations of common optimization patterns**
  - E.g., Stochastic-Gradient-Descent: Applicable to summable ML losses
  - E.g., LBFGS: An approximate 2nd order optimization method
  - E.g., ADMM: Decomposition / coordination procedure

MLbase Architecture



2 Binders Full of Algorithms



Implementation

On top of MLI
(with optimization hints)

+

Contract

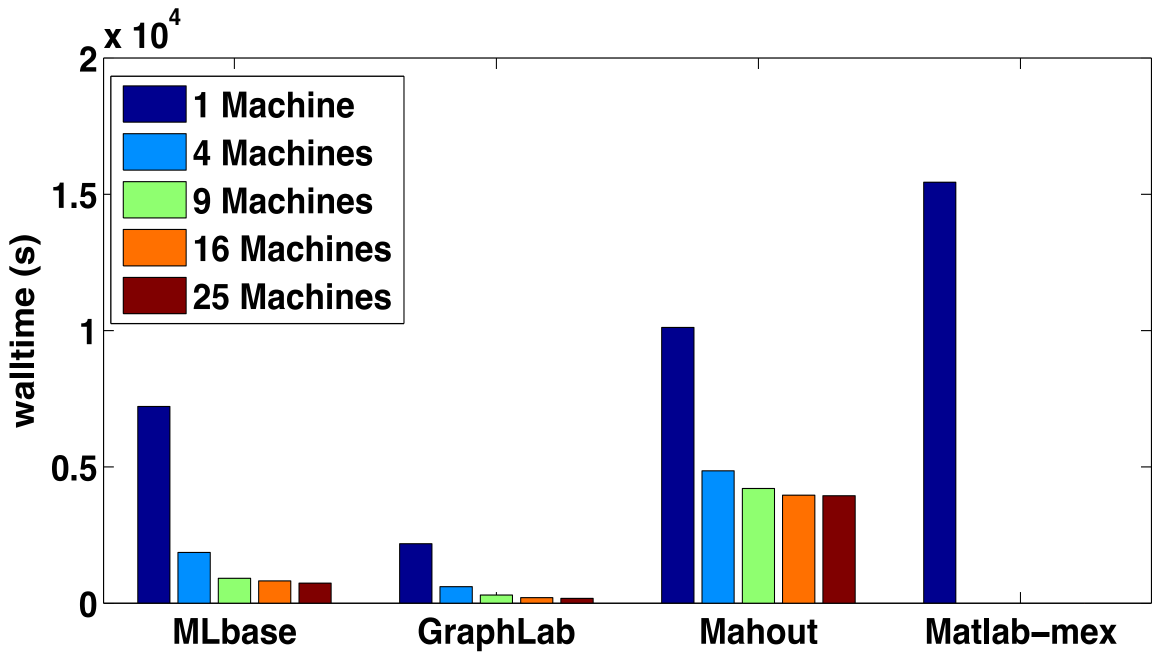
- Type (e.g., classification)
- Parameters
- Runtime (e.g., O(n))
- Input-Specification
- Output-Specification
- …

Today: Half-Full Binders

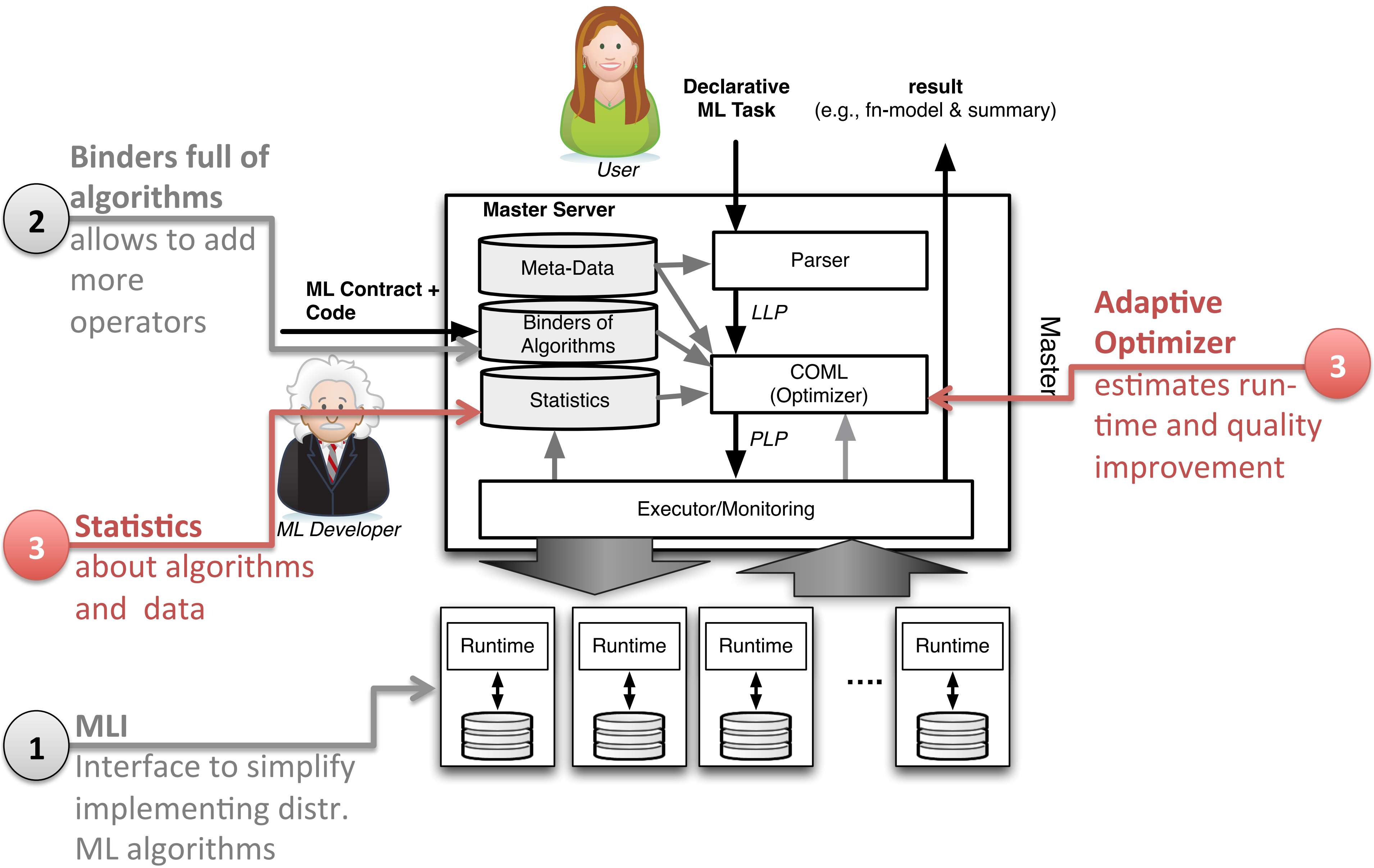
- **Regression:** Linear Regression (+Lasso, Ridge)
- **Classification:** Logistic Regression, Linear SVM (+L1, L2), Multinomial Regression, [Naïve Bayes, Decision Trees]
- **Collaborative Filtering:** Alternating Least Squares, [DFC]
- **Clustering:** K-Means, [DP-Means]
- **Optimization Primitives:** SGD, Parallel Gradient, [L-BFGS, ADMM, Adagrad]
- **Feature Extraction**: [PCA], N-grams, feature cleaning normalization
- **Other tools**: Cross Validation, Evaluation Metrics
- Released as part of Spark and MLlib

Example: Alternating Least Squares

System	Lines of Code
Matlab	20
Mahout	865
GraphLab	383
MLI	32



MLbase Architecture

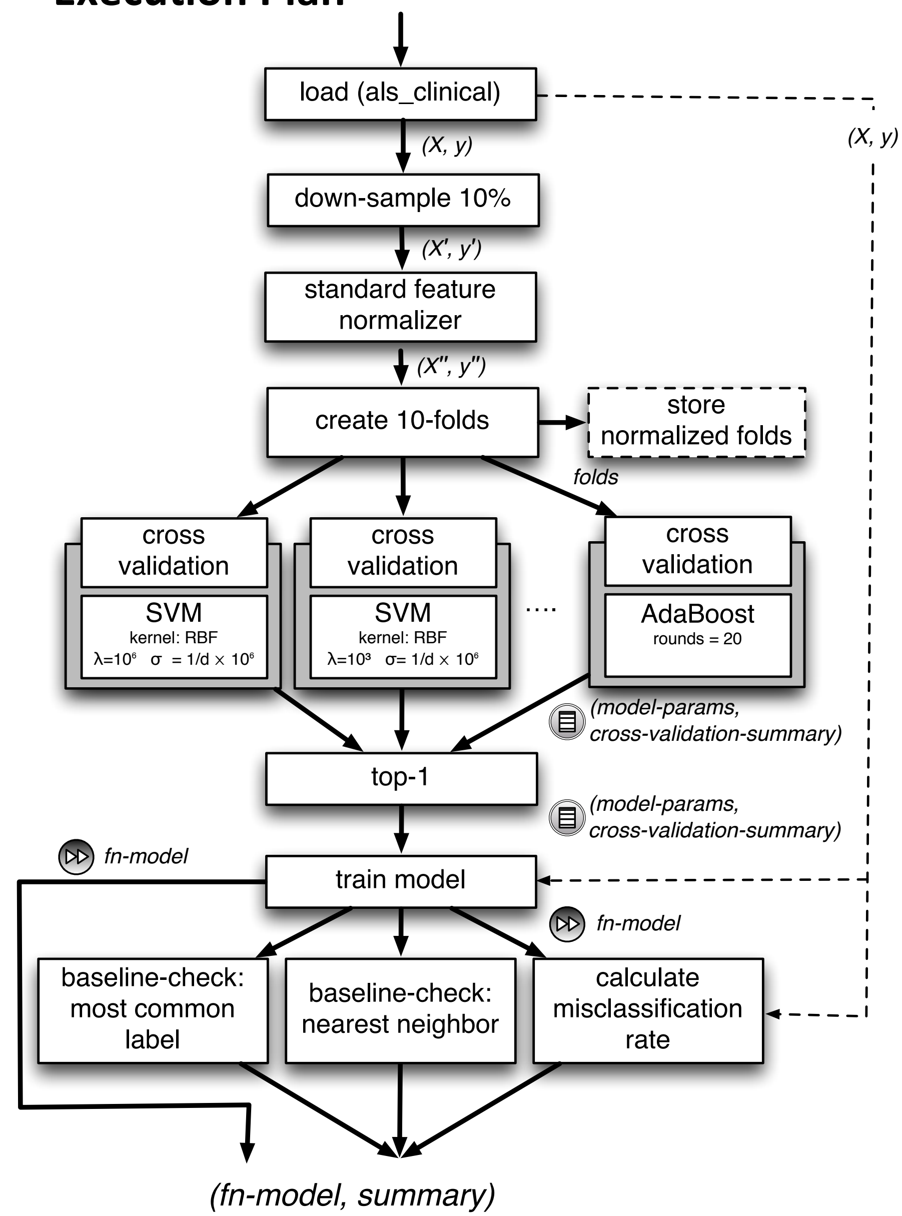


3 Optimization

MQL

```
var X = load("als_clinical",2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) =
      top(doClassify(X, y), 10min)
```

Execution Plan



Optimization Goals

1. Return **meaningful** **results**
2. Optimize the **whole processing** pipeline
3. Optimize **quality** **and** **time** simultaneously

Current Optimization Approach

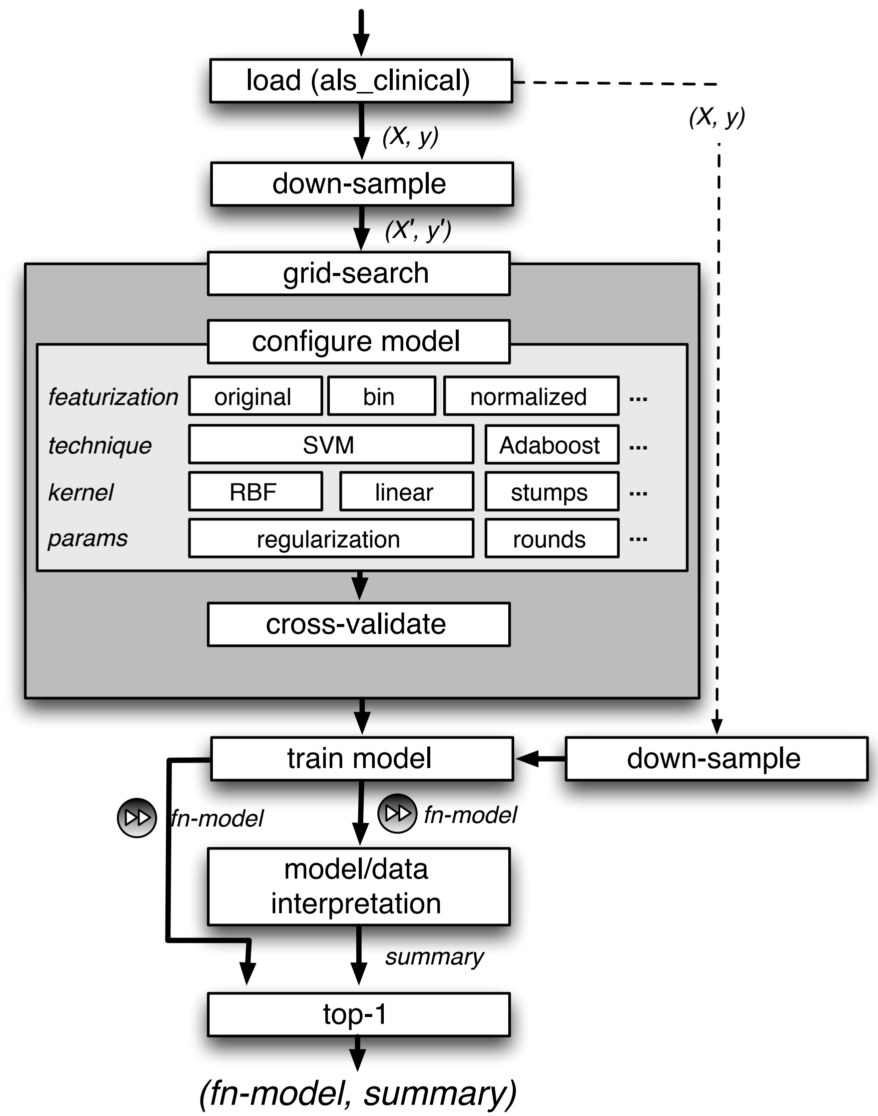
Idea: 3-Step Process

3 Optimization

(1) MQL

var X = load("als_clinical",2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) =
    top(doClassify(X, y), 10min)

(2) Generic Logical Plan

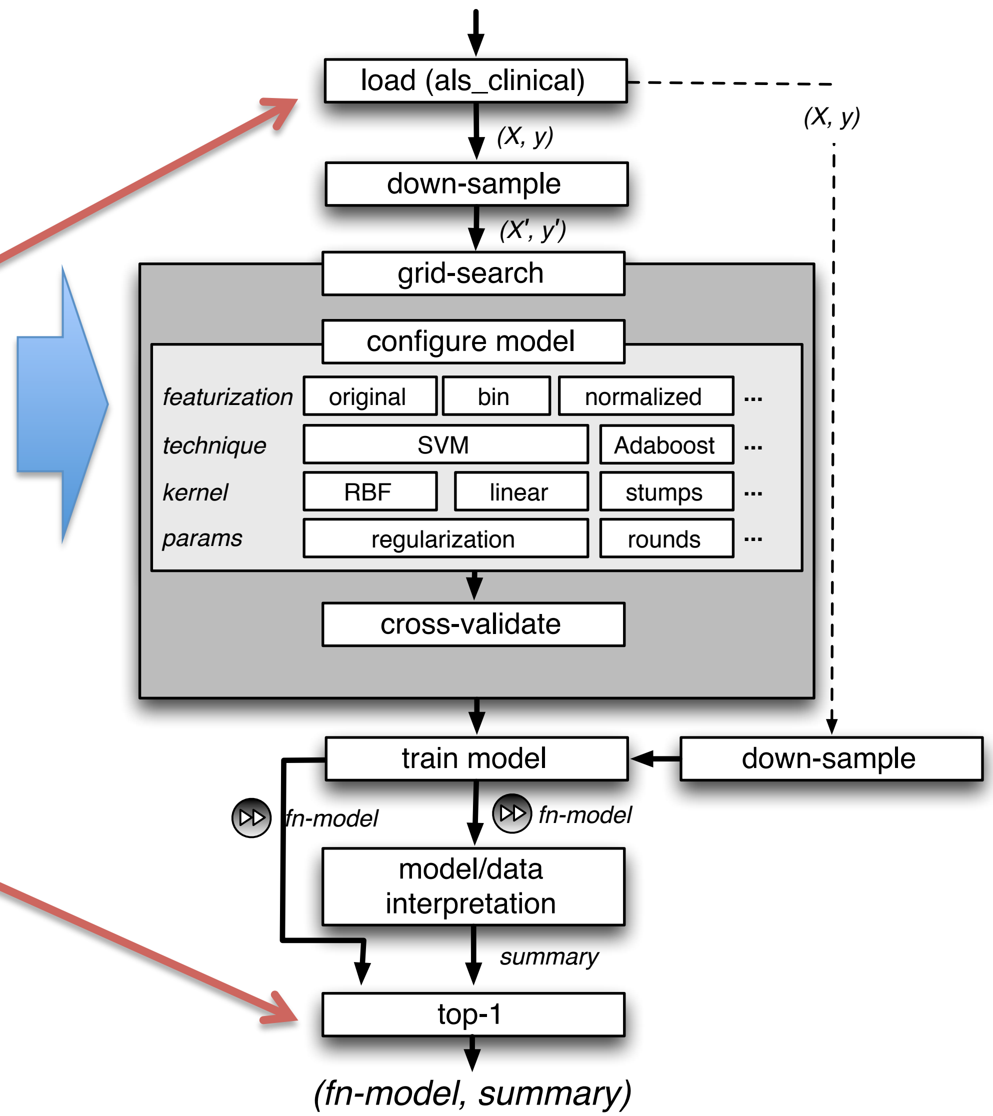


3 Optimization

(1) MQL

```
var X = load("als_clinical",2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) =
    top(doClassify(X, y), 10min)
```

(2) Generic Logical Plan

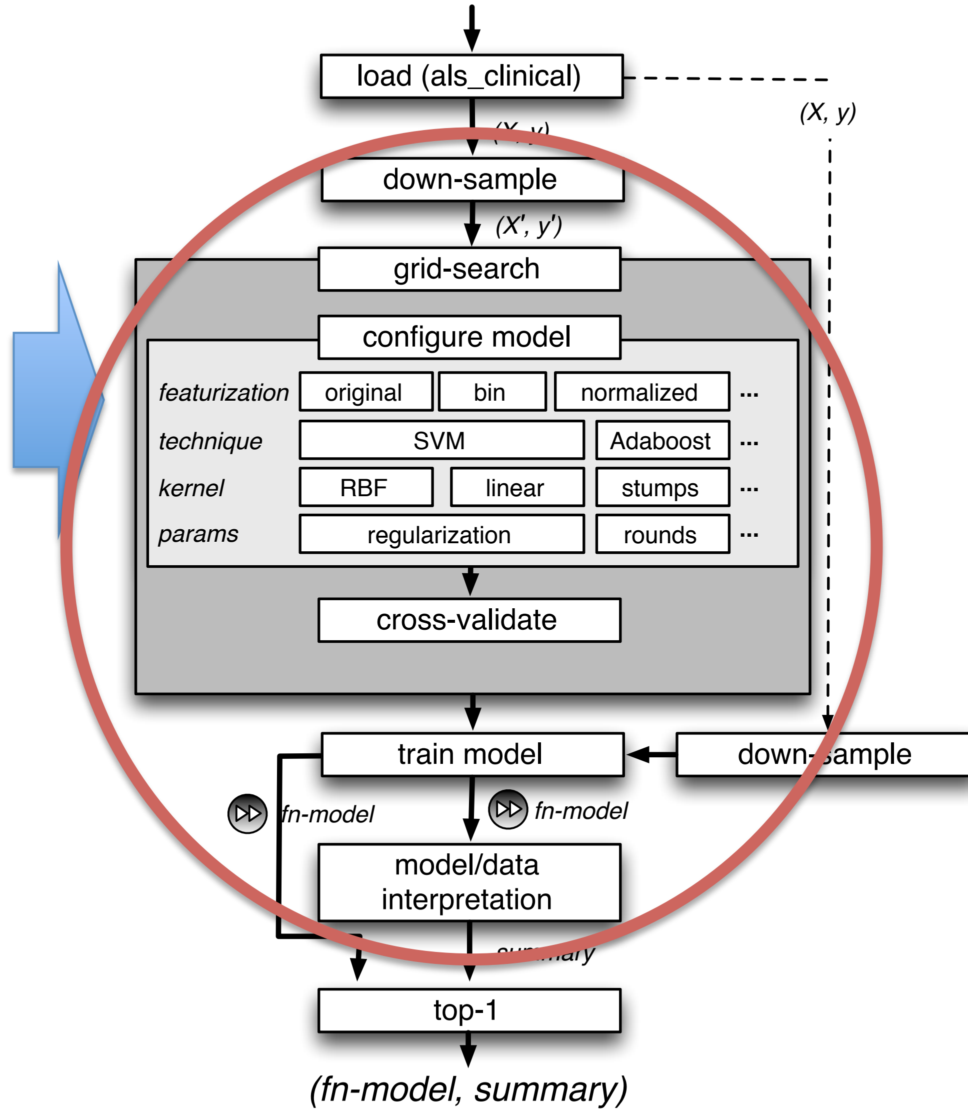


3 Optimization

(1) MQL

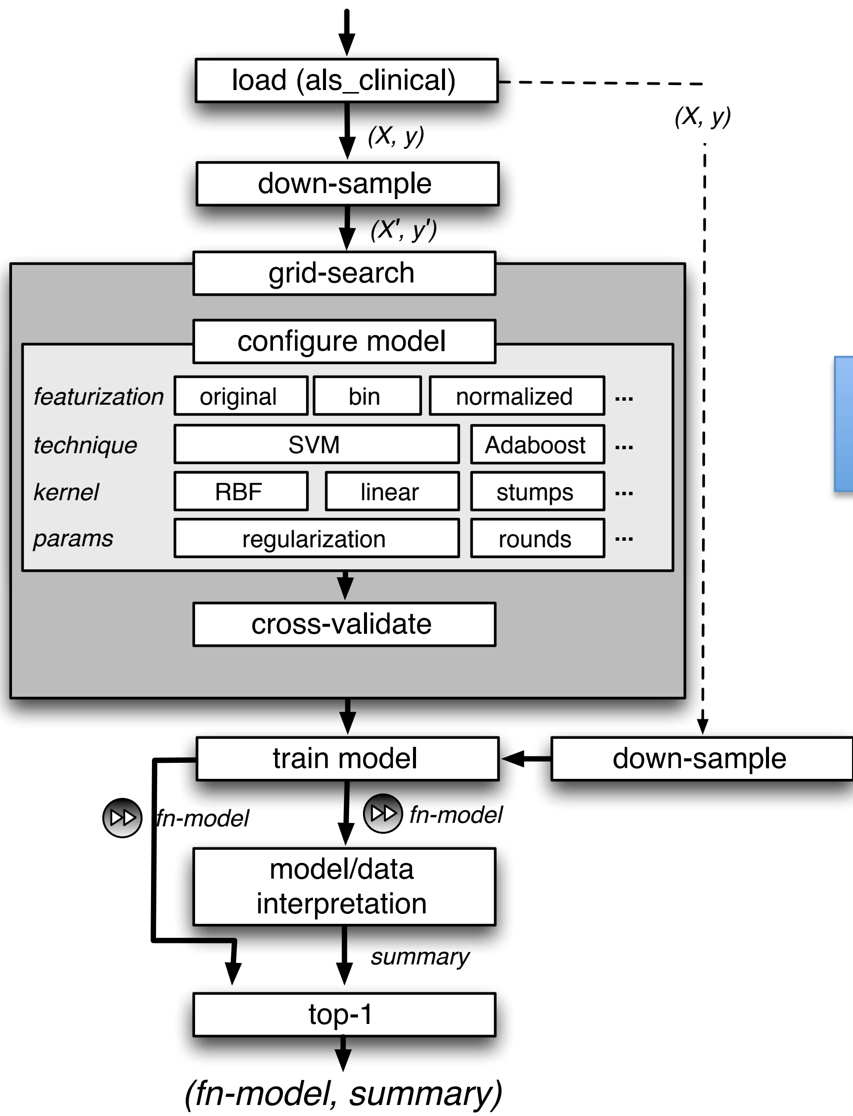
var X = load("als_clinical",2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) =
    top(**doClassify**(X, y), 10min)

(2) Generic Logical Plan

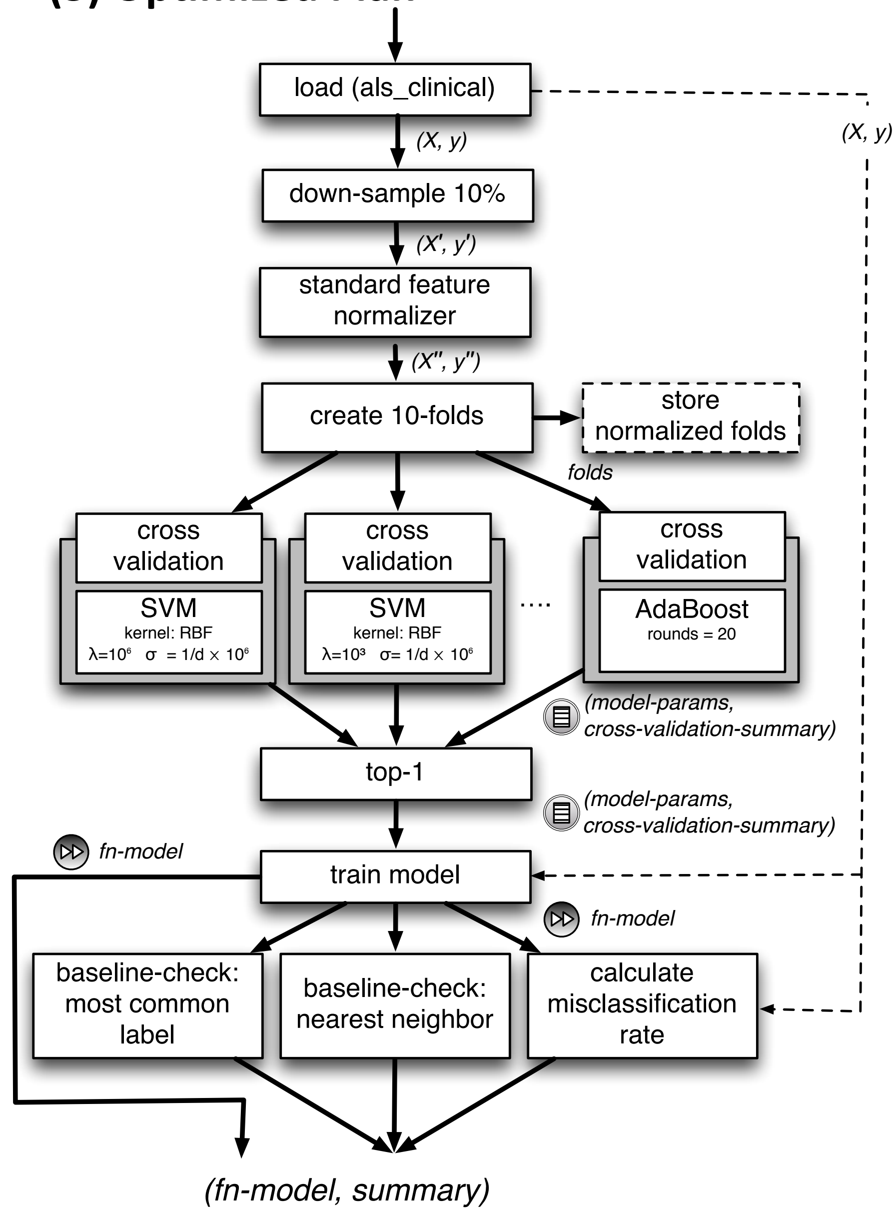


Optimization

(2) Generic Logical Plan



(3) Optimized Plan



DB Optimizer meets ML Parameter Tuning

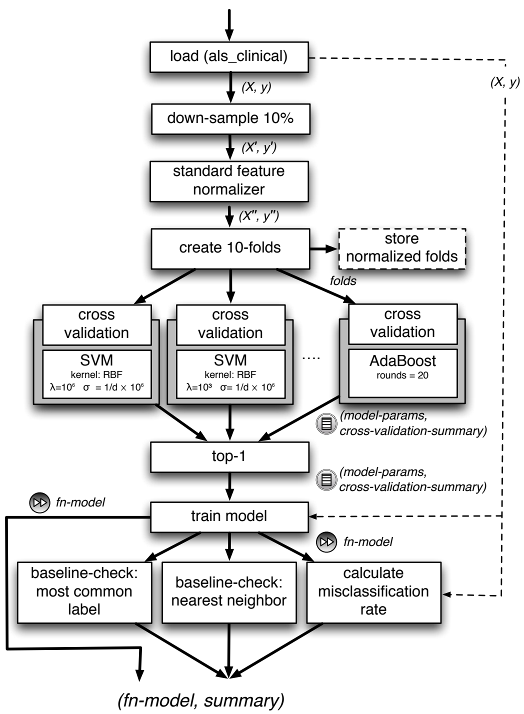
More than Grid-Search, more than Relational Query Optimization

MLbase cost-based optimization:

Quality & Time **(=budget)**

- **Considers algorithms, system techniques and best practice workflows together**
- **Statistics** about **data** and **algorithms**
  → Hope to find strong correlation between data statistics and the quality of an algorithm
- Optimization **across steps** (e.g., cleaning, feature extraction, classification,…)
- **Robustness/Avoiding Overfitting & Hypothesis Pitfall** (messing up quality is worse than time in traditional query optimization)

Possible Optimizations (classification)



Relational Optimizations (Top-K Pushdown, Join-Ordering,…)

Static ML Selection Rules

- Imbalance of labels
- SVMs are more sensitive to the scale-parameter than AdaBoost to rounds
- If *SVM* → normalize data between [-1, 1]
- If data contains outliers → pre-clean data or forego AdaBoost
- …

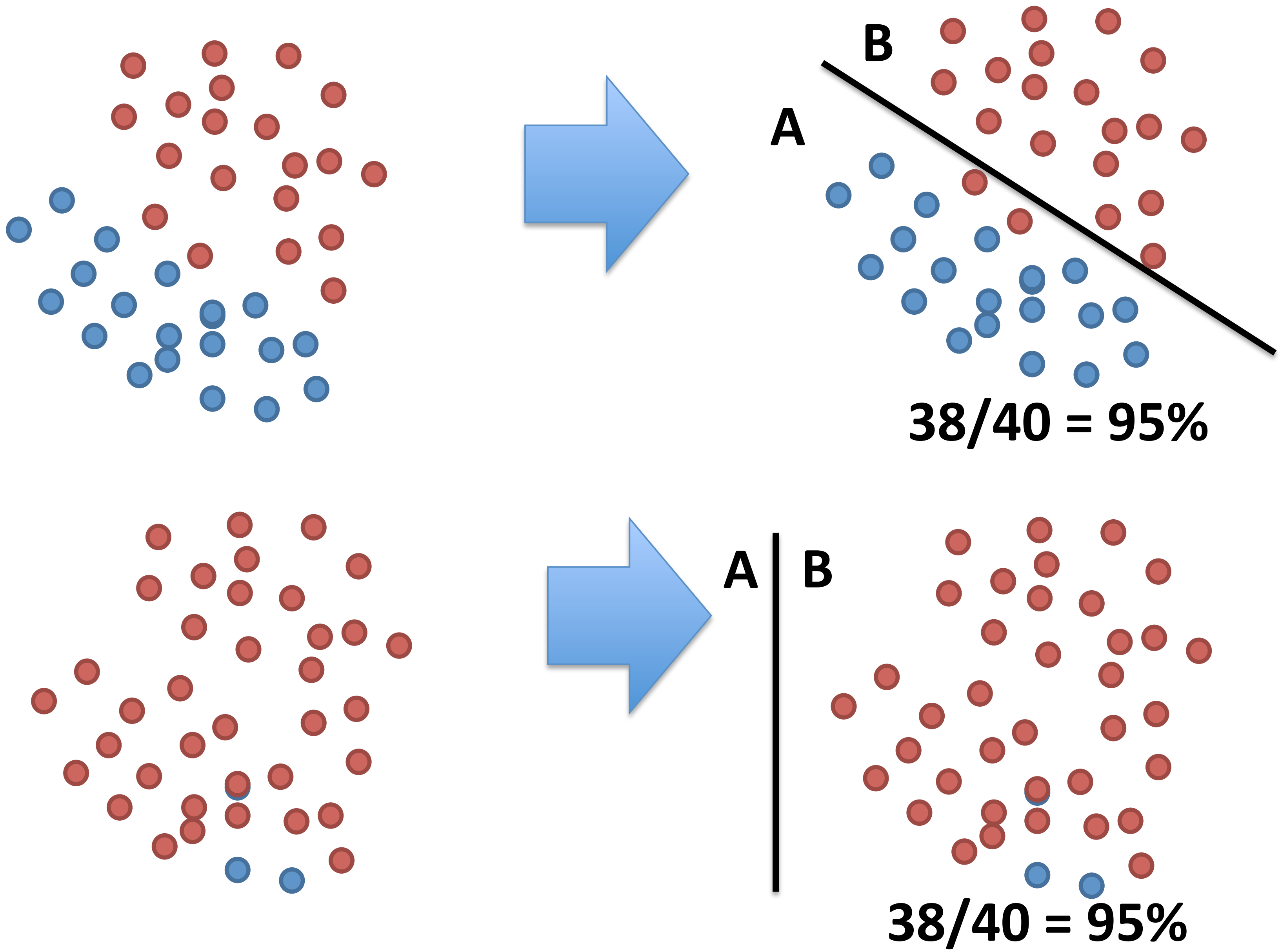
Run-Time Optimization Rules

- Caching: If 2nd run and deterministic, start with previously most successful model
- Set sample-size to fit Input-Data as well as intermediate result in memory
- Partition data according to cross-validation
- …

Cost-based Optimization Rules

- Materialization and indexing
- Expected quality improvement based on the history
- Consider cost of pre-cleaning, normalization, algorithm complexity,…
- …

Why Optimize? Pitfalls

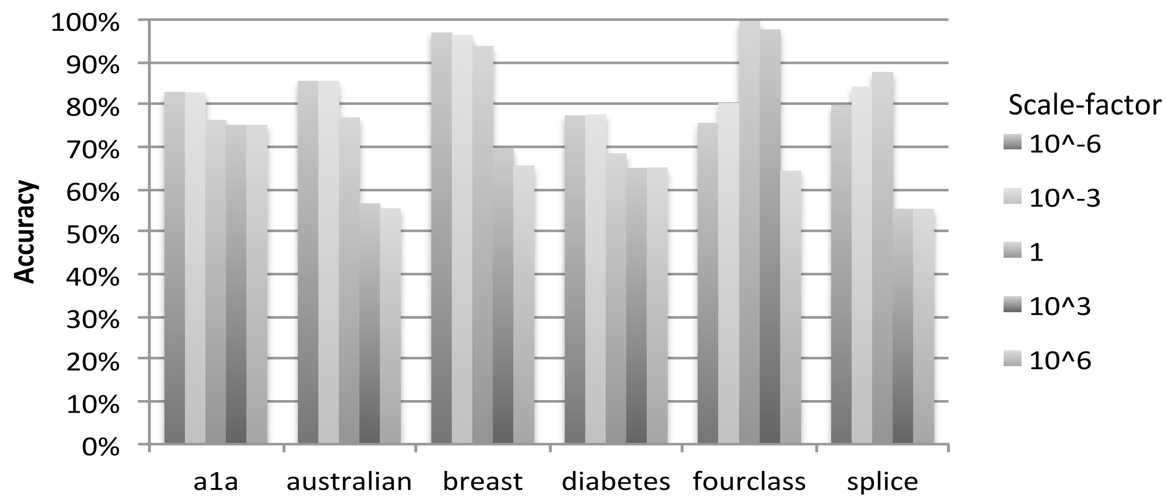


Why Optimize?
Quality

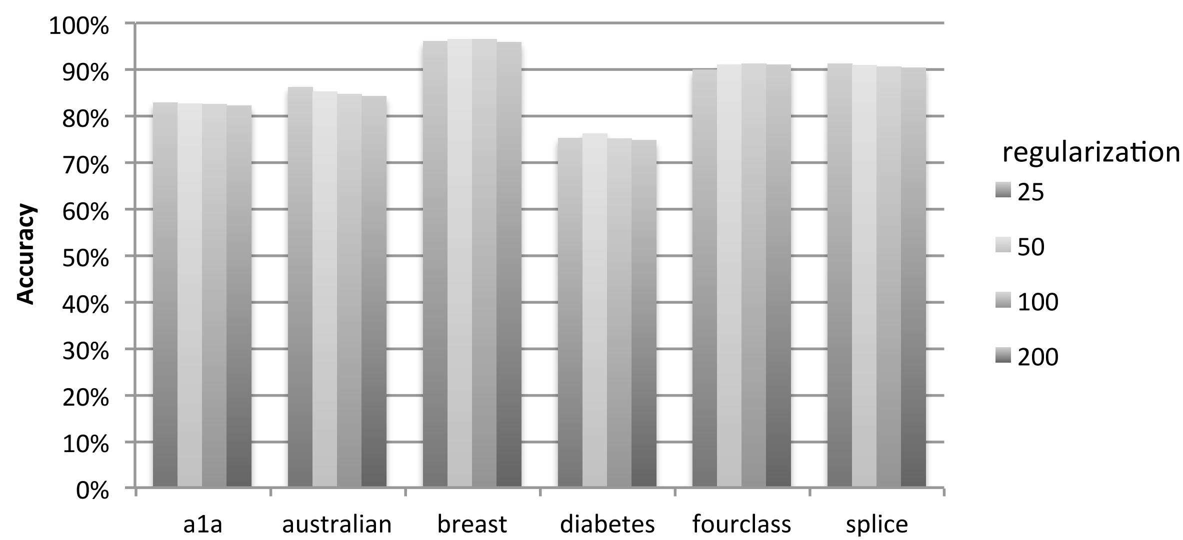
	SVM		AdaBoost
	original	scaled	
a1a	$\boxed{\mathbf{82.93}}$	$\boxed{\mathbf{82.93}}$	82.87
australian	85.22	85.51	$\boxed{\mathbf{86.23}}$
breast	70.13	$\boxed{\mathbf{97.22}}$	96.48
diabetes	76.44	$\boxed{\mathbf{77.61}}$	76.17
fourclass	$\boxed{\mathbf{100.00}}$	99.77	91.19
splice	88.00	87.60	$\boxed{\mathbf{91.20}}$

Why Optimize?
Quality

SVM



AdaBoost



Why Optimize?
Speed

- Running 1 algorithm tends to be **I/O bound**
- Idea: train in parallel with different algorithms and parameters → Similar to **shared cursors** in DB-world
- Questions:
  - How many models?
    → How to make it cache-aware
  - Impact of sampling?
  - How to leverage modern CPUs, in particular vectorization and CPU pipelining?

Direction

- Released:
  - MLI interface
  - Half-full binders as part of Spark
  - Some simple feature extractors
  - (End-to-end use cases)

- Working on:
  - Optimization techniques
  - Cost-based optimizer
  - Unified language for end users and ML developers
  - Advanced ML capabilities: Time-series algorithms, graphical models, advanced optimizations, online updates, sampling for efficiency
  - Integration into TupleWare: High-Performance analytic platform
  - Visualization

MLBase - Summary

- **MLbase is a first declarative machine-learning system**
- **It simplifies ML in the same way as databases simplify data management**
- Teaser: TupleWare will integrate Mlbase and leverage ideas from *programming languages* to significantly speed-up ML and explorative data analysis

Tim Kraska

tim_kraska@brown.edu