MLbase: A Distributed Machine-learning System

Rahul Shanbhog
04/03/2015

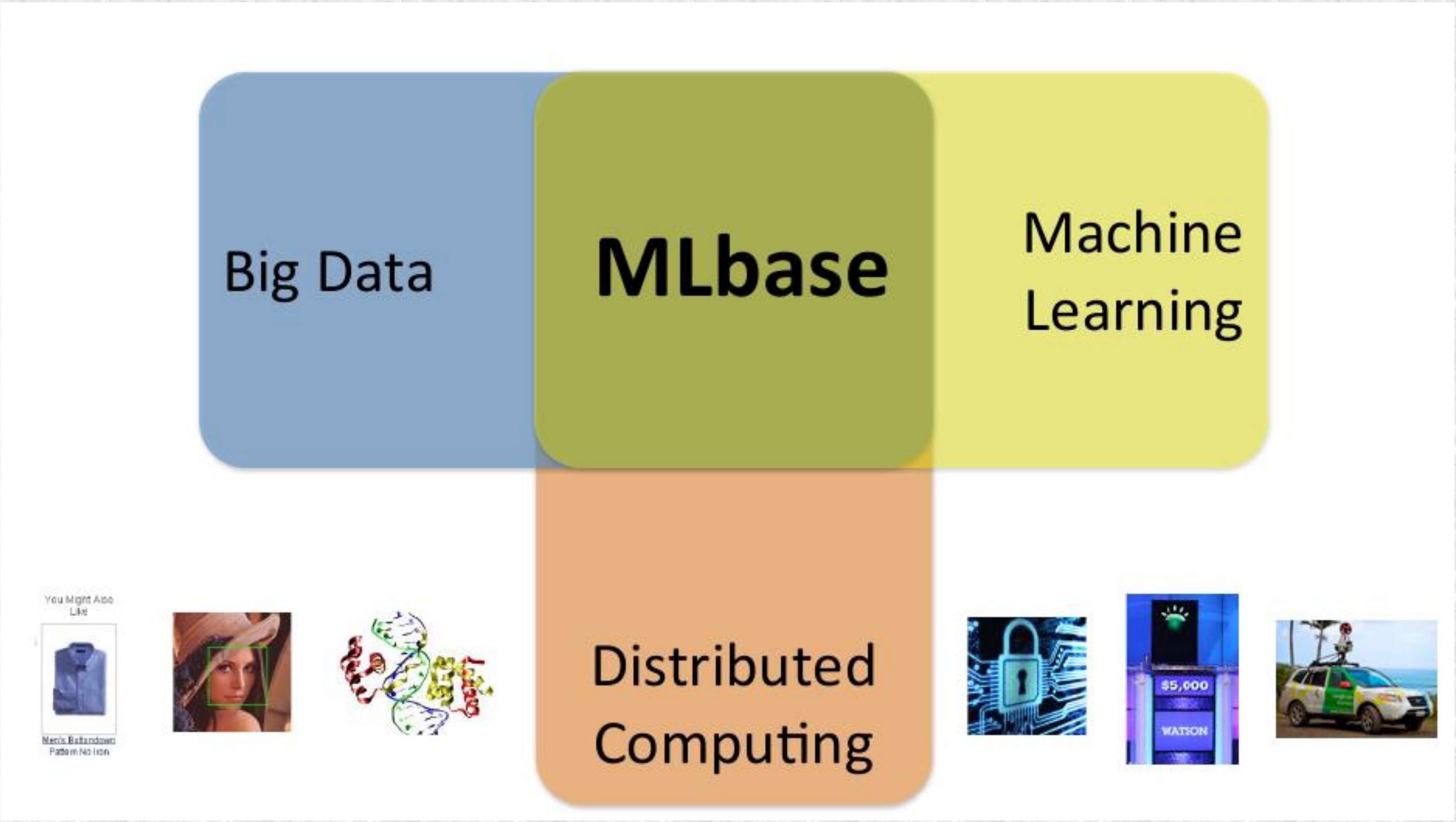
Three Converging Trends

- Big Data
- Distributed Computing
- Machine Learning

Challenges

- Machine learning is essential to transform big data into actionable knowledge
- Complexity of ML algorithms is overwhelming
- Users most often do not understand the tradeoffs and the challenges
- Existing systems demand the ML researchers to be strong on distributed systems background

Three Converging Trends



MLbase in a nutshell

- Simple declarative way to specify ML tasks
- An optimizer to select and dynamically adapt the choice of the learning algorithms
- High level operators to enable researchers to implement a wide range of ML methods without much knowledge
- New runtime optimized for the data-access

Use Cases

Set of functionality to end users :

- classification, regression, collaborative filtering
- exploratory data analysis techniques
  - dimensionality reduction, feature selection, and data visualization

Use Cases: Supervised Classification

- ALS Prediction:
- Using the largest database of clinical data for ALS patients, the ALS Prediction Prize challenges participants to develop a binary classifier to predict whether an ALS patient will display delayed disease progression.

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

Use Cases: Unsupervised Classification

- Twitter Analysis:
- Use snapshots of the Twitter network and associated tweets to perform a variety of unsupervised exploratory analyses to better understand the data.
- Advertisers may want to find features that best describe “hubs,” people with the most followers or the most retweeted tweets

```
var G = loadGraph("twitter_network")
var hubs-nodes = findTopKDegreeNodes(G, k =
```

Use Cases: Supervised Classification with Hints

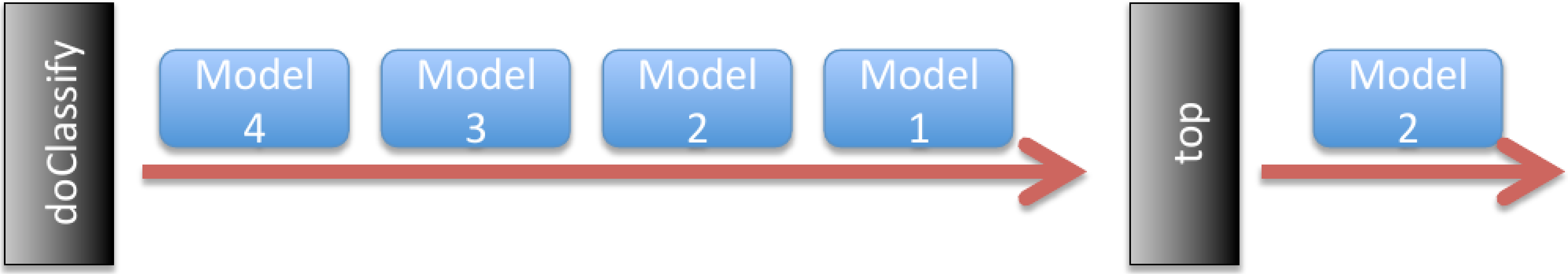
Offers Algorithm Independence

But can also take in suggestions

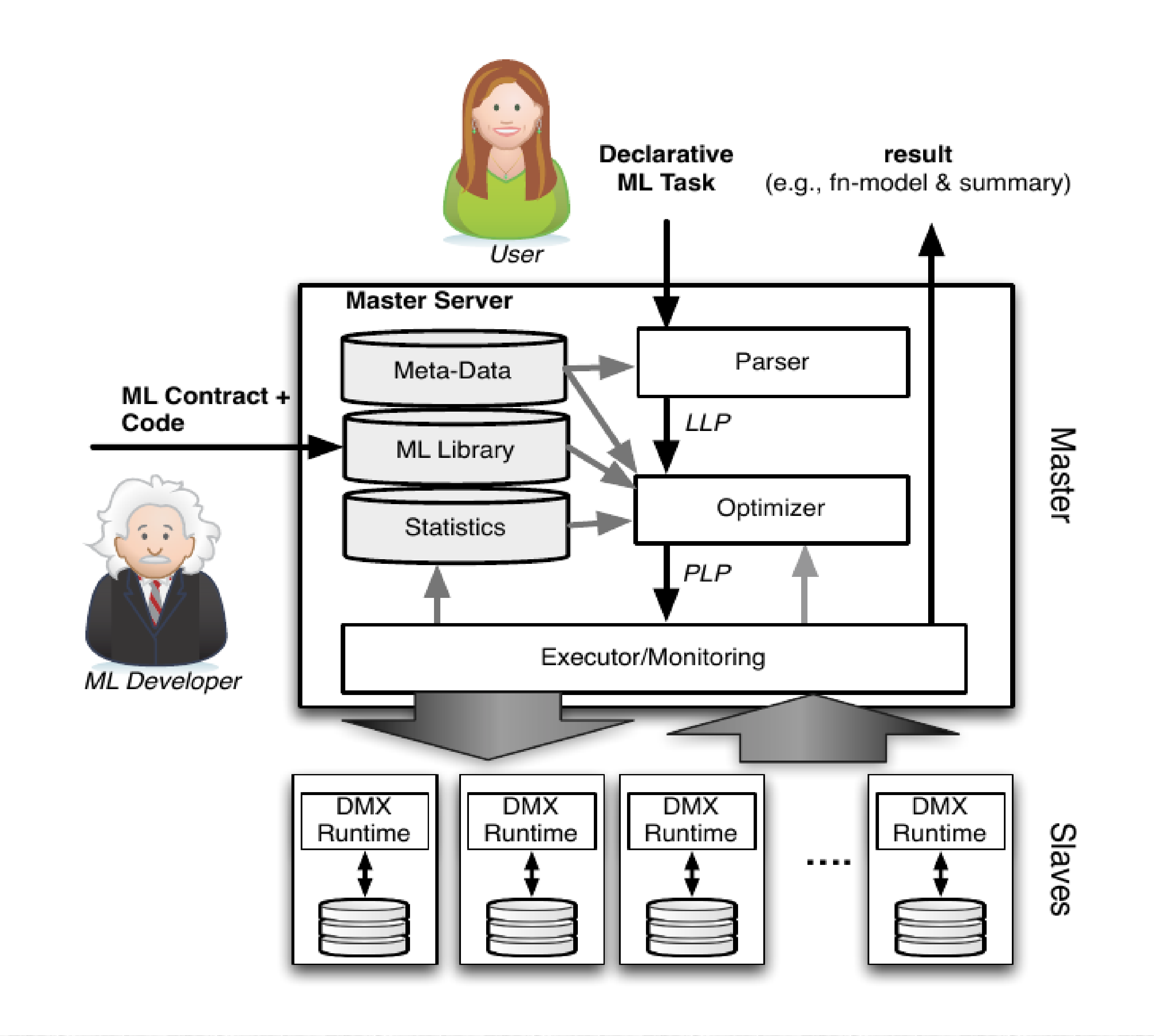
```
var X = load("als_clinical", 2 to 10)
var y = load("als_clinical", 1)
var (fn-model, summary) = top(doClassify(X, y, SVM), 5min)
```

Streaming-like Data Model

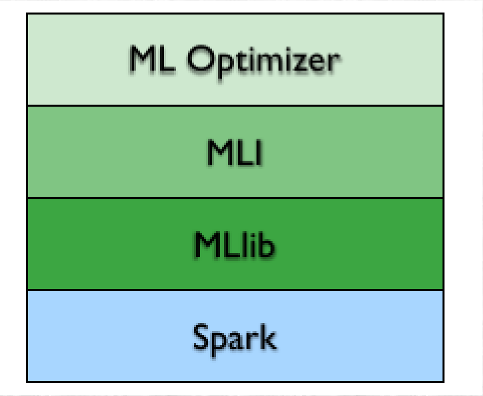
Infinite ordered stream of items, being either models (i.e., higher-ordered functions) or tuples



MLbase Architecture

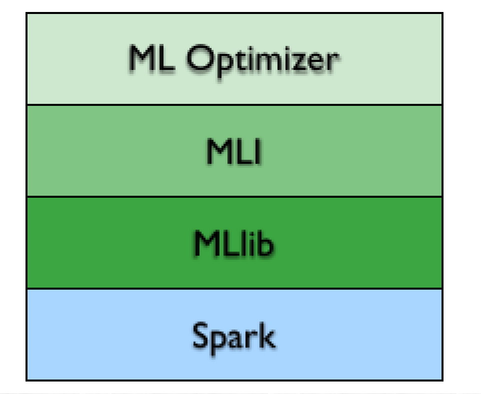


MLbase Stack



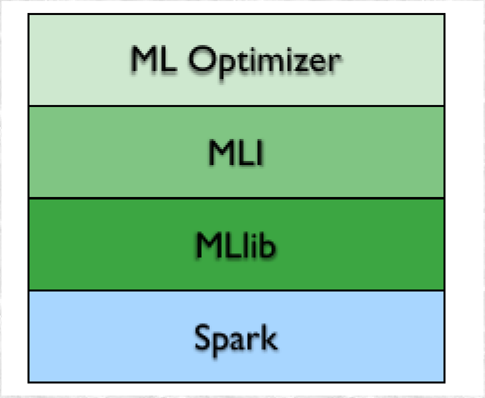
MLbase Stack

- Spark:
  - Base of the stack
  - Cluster computing system
  - Designed for machine learning
  - Easy to use
    - Setting up
    - Computing



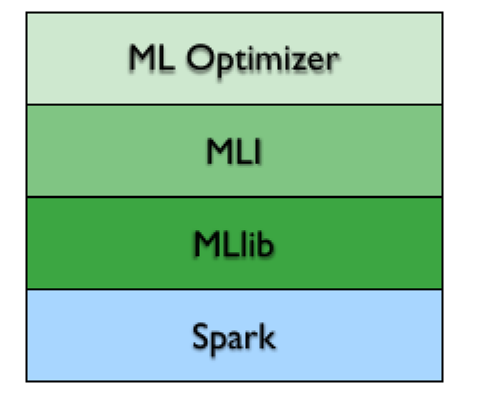
MLbase Stack

- MLlib
  - Lowest level of MLbase
  - Low level ML library
  - Present as part of the code ase of Spark
  - Callable from Scala / Java



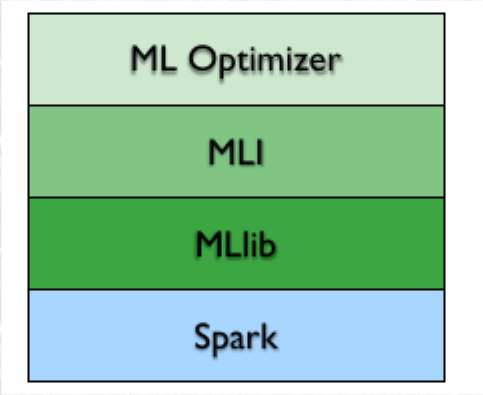
MLbase Stack

- MLI
  - Above MLlib
  - API / platform for feature extraction and algorithm development
  - Includes higher level functionality
  - Shield ML Developers from low-level details

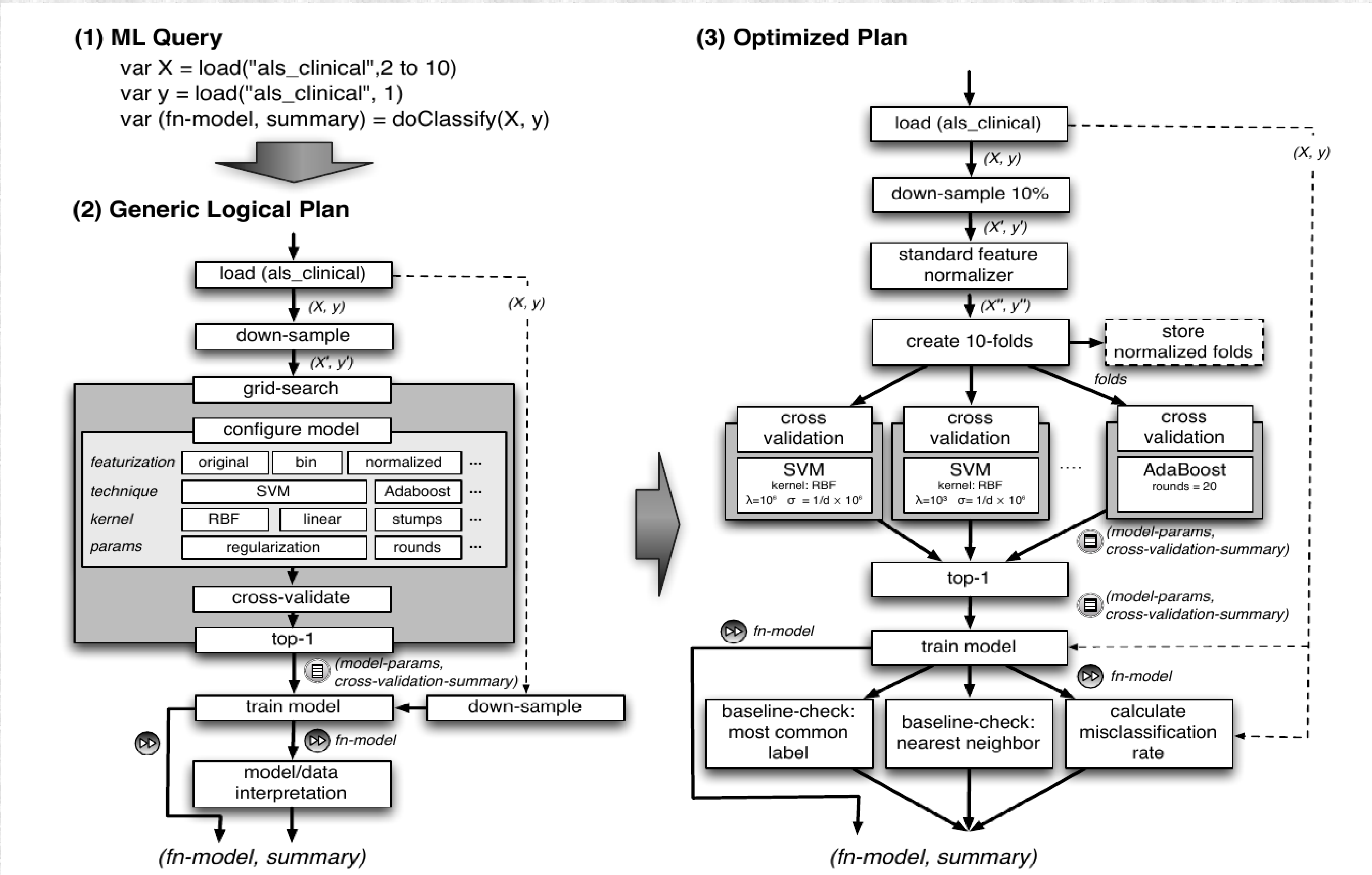


MLbase Stack

- ML Optimizer
  - Topmost layer in the stack
  - Automates the process of model selection
  - Designed to target the quality of the result and not only the timing



Optimization

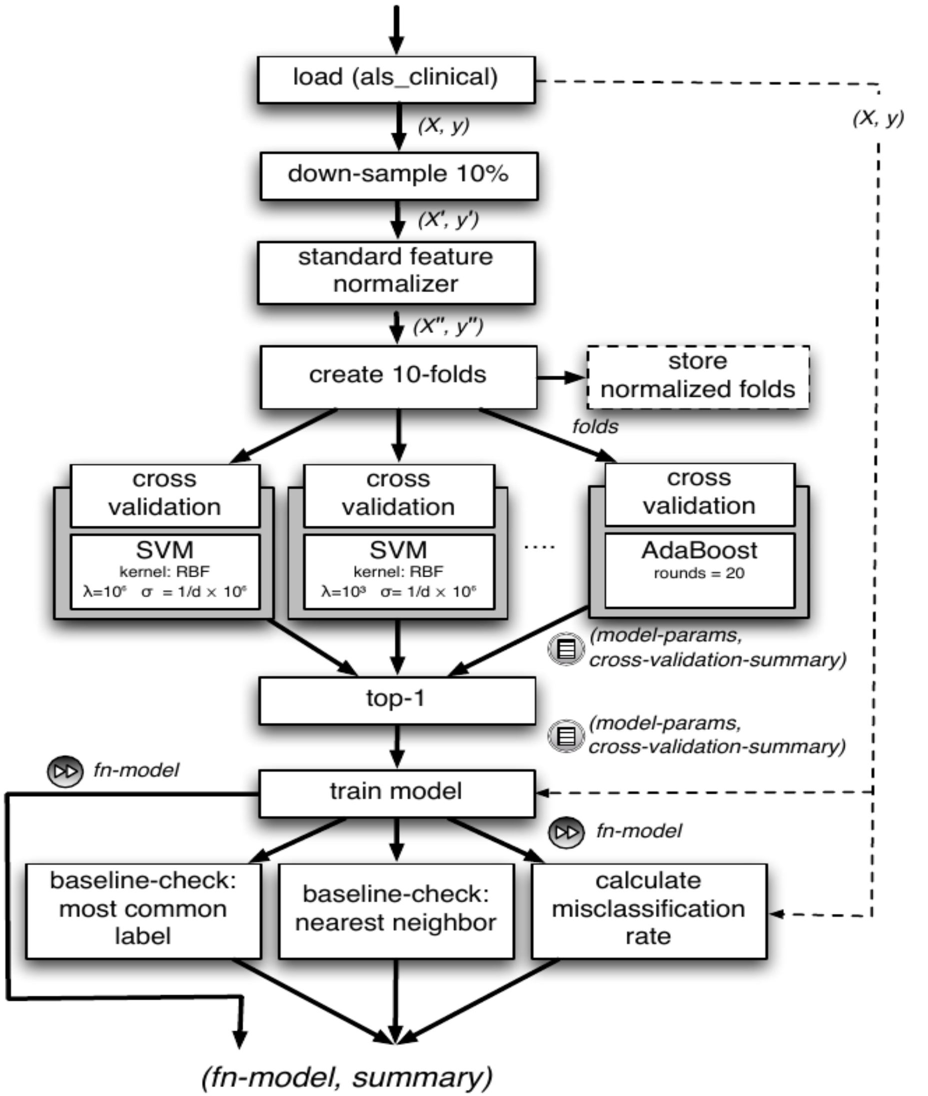


Optimization

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

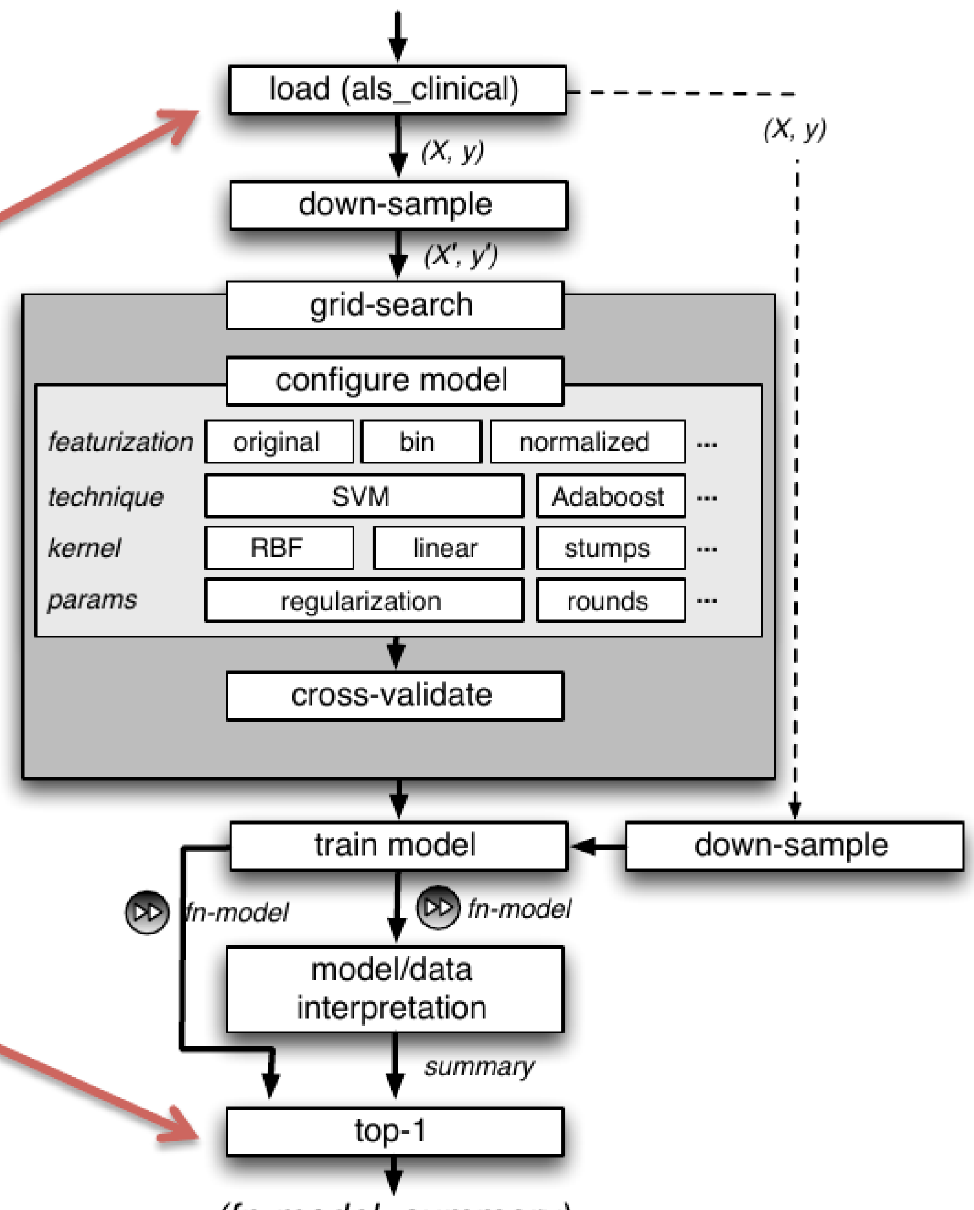


Optimization

(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

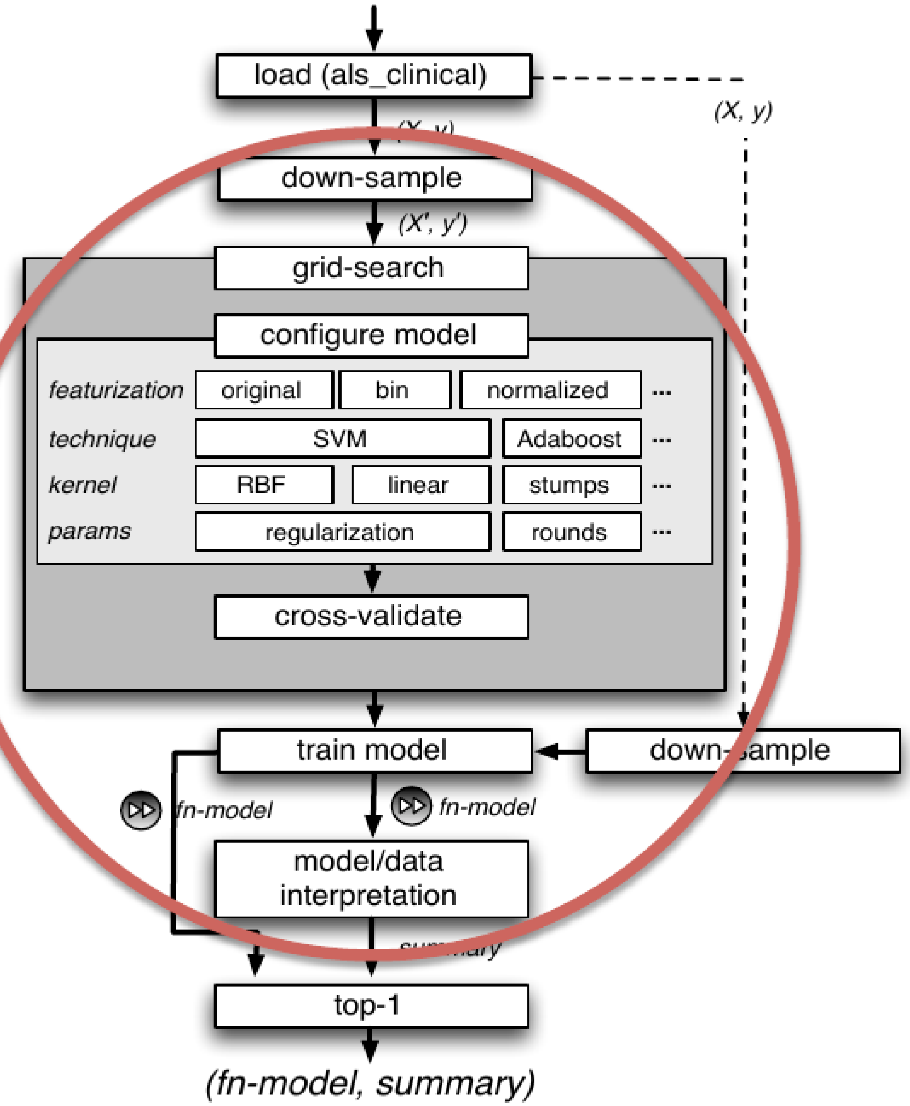


Optimization

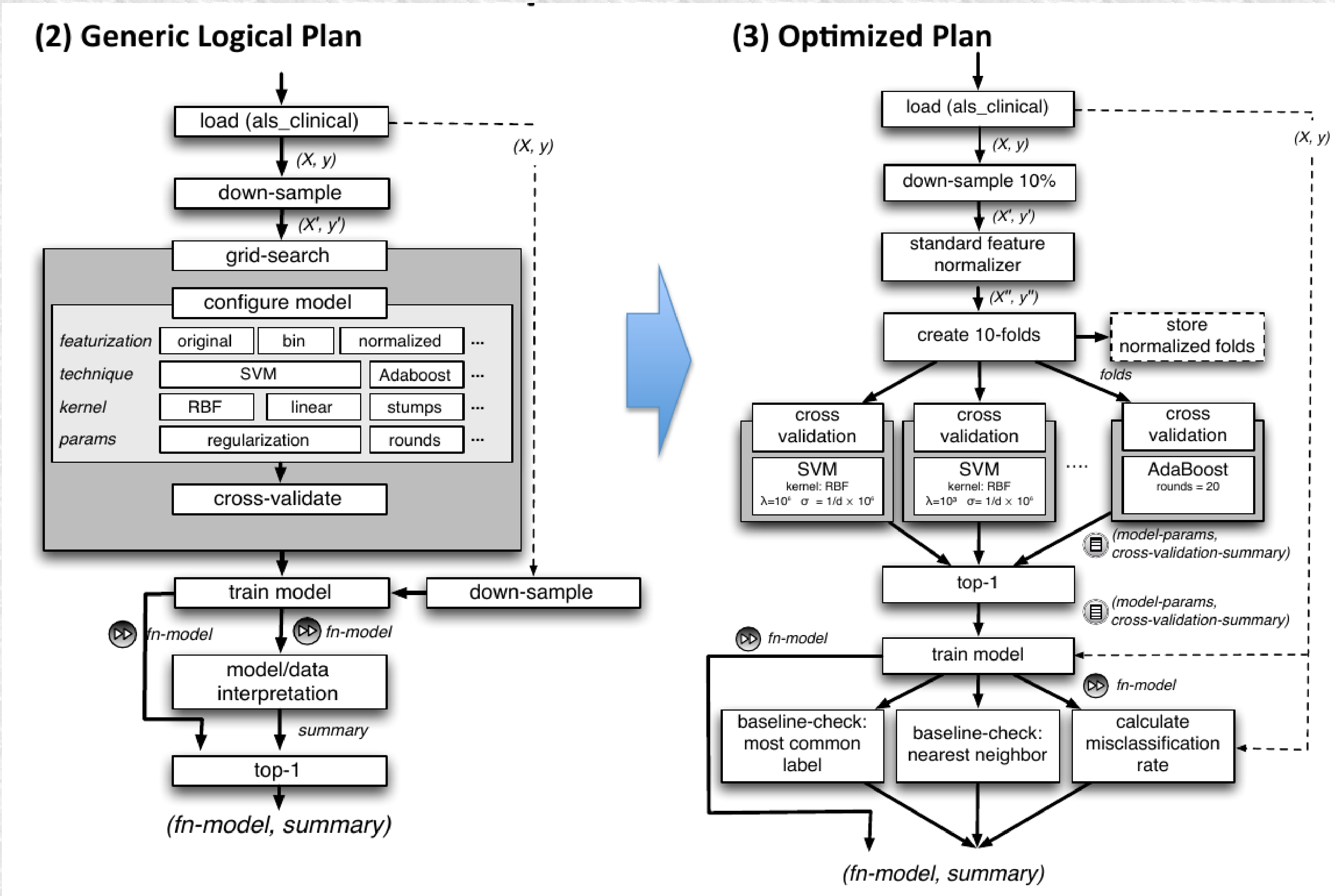
(1) MQL

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



Optimizer Example

6 datasets:

- ‘a1a’
- ‘australian’
- ‘breast-cancer’
- ‘diabetes’
- ‘fourclass’
- ‘splice’

Optimizer Example

Classifier accuracy using SVM with an RBF kernel and using AdaBoost

	SVM		AdaBoost
	original	scaled	
a1a	**82.93**	**82.93**	82.87
australian	85.22	85.51	**86.23**
breast	70.13	**97.22**	96.48
diabetes	76.44	**77.61**	76.17
fourclass	**100.00**	99.77	91.19
splice	88.00	87.60	**91.20**

Optimizer Example

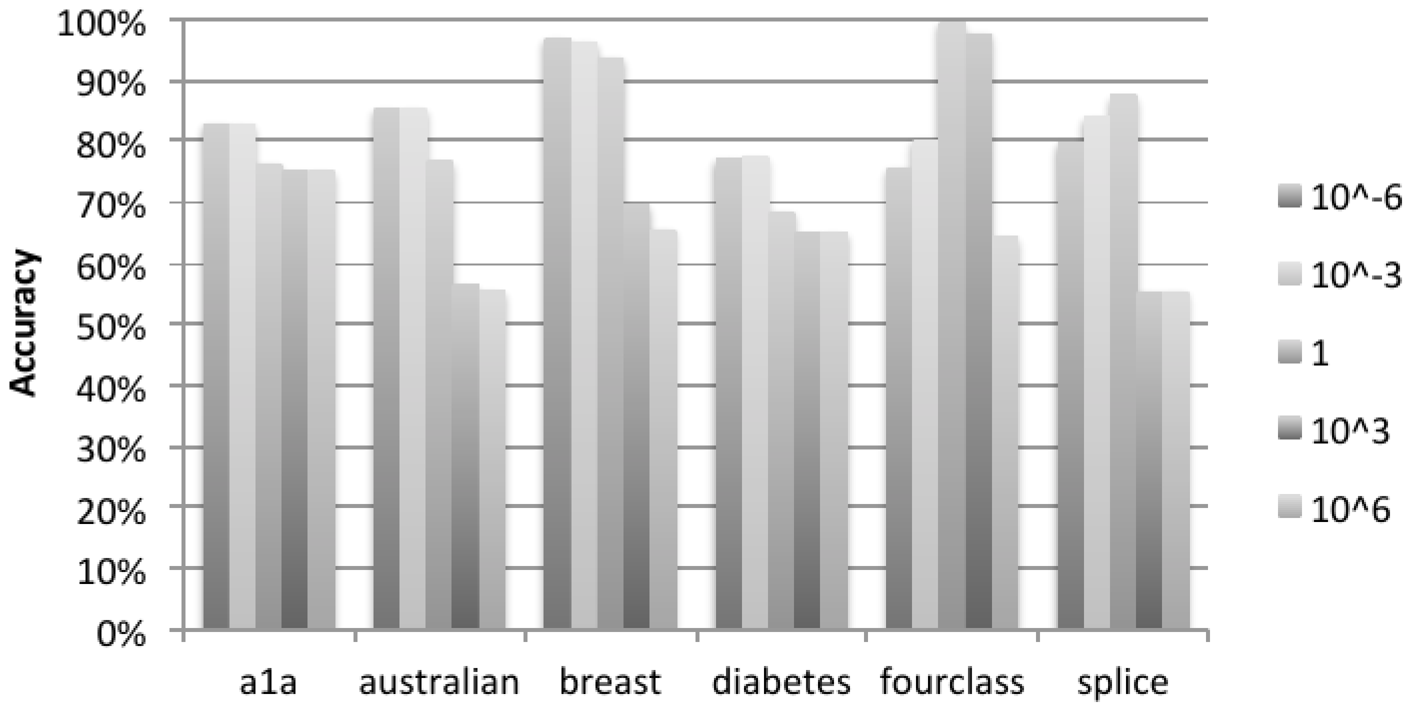


Figure 4: Impact of different $\sigma = \frac{1}{d} \times \{10^{-6}, 10^{-3}, 1, 10^{3}, 10^{6}\}$ on the SVM accuracy with an RBF kernel and $\lambda = 10^{-6}$ on LIBSVM data-sets

Optimizer Example

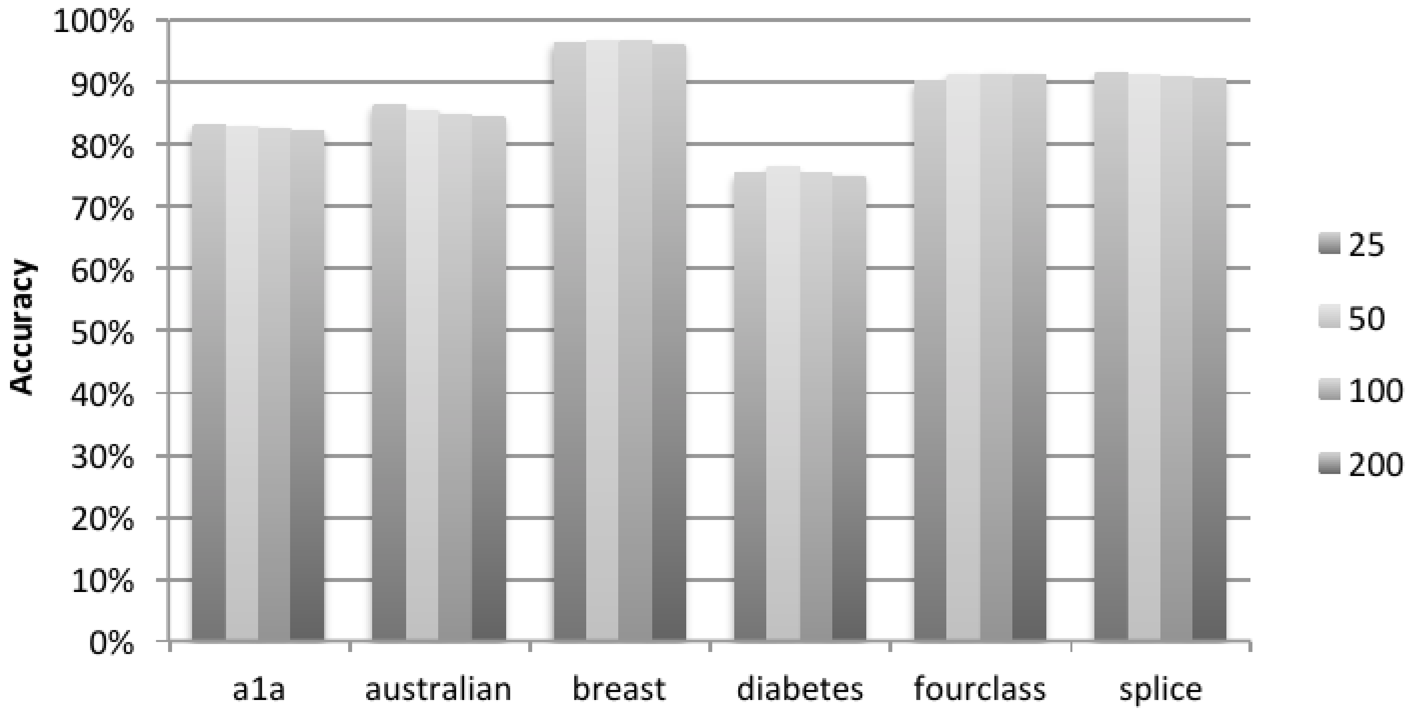


Figure 5: Impact of $r = \{25, 50, 100, 200\}$ on AdaBoost on LIBSVM data-sets

Direction

Released:

- MLI Interface
- A number of algorithms as part of Spark
- Simple feature extractors

Direction

Working on:

- Optimization Techniques
- Unified language for end users and ML developers
- Advanced ML capabilities

Questions?

Thank you