Big Data Research in the AMPLab: BDAS and Beyond

Michael Franklin
UC Berkeley

1st Spark Summit
December 2, 2013

AMPLab: Collaborative Big Data Research

Launched: January 2011, 6 year planned duration
Personnel: ~60 Students, Postdocs, Faculty and Staff
Expertise: Systems, Networking, Databases and Machine Learning
In-House Apps: Crowdsourcing, Mobile Sensing, Cancer Genomics

amplab UC BERKELEY

amazon.com web services, Google, SAP, NSF, DARPA, Berkeley Lab

Apple, Inc., Cisco, ClearStory, cloudera, Ericsson, facebook, FitWave

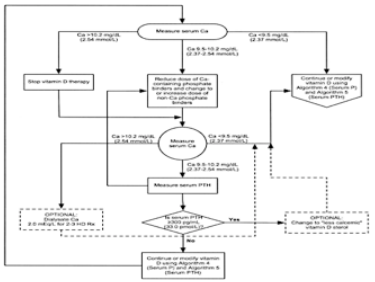
intel, Microsoft, NetApp, GE imagination at work, Hortonworks, Huawei, ORACLE

Samsung, splunk>, vmware, wanDISCO, YAHOO!

AMPLab: Integrating Diverse Resources

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence



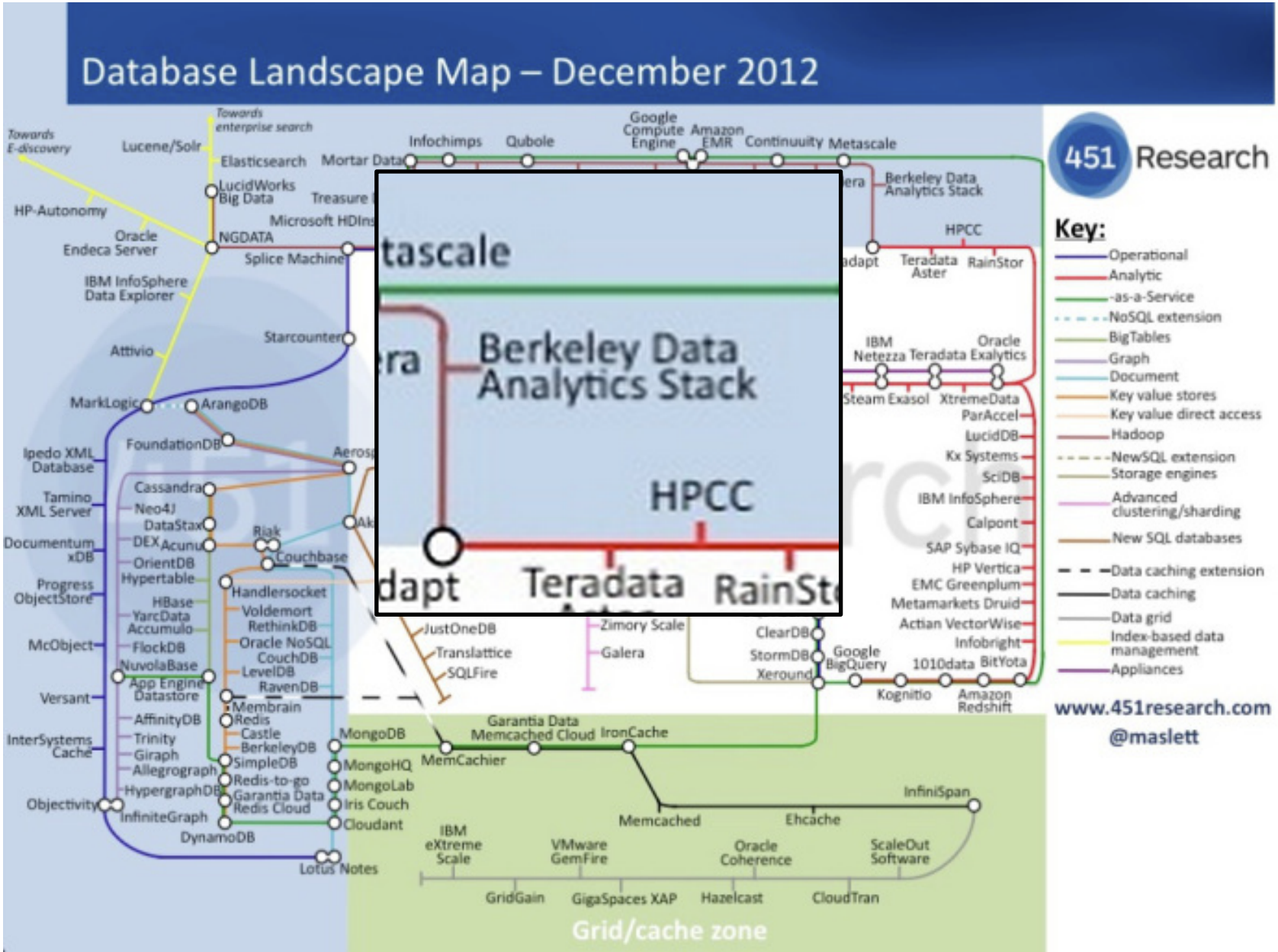
Machines

- Clusters and Clouds
- Warehouse Scale Computing

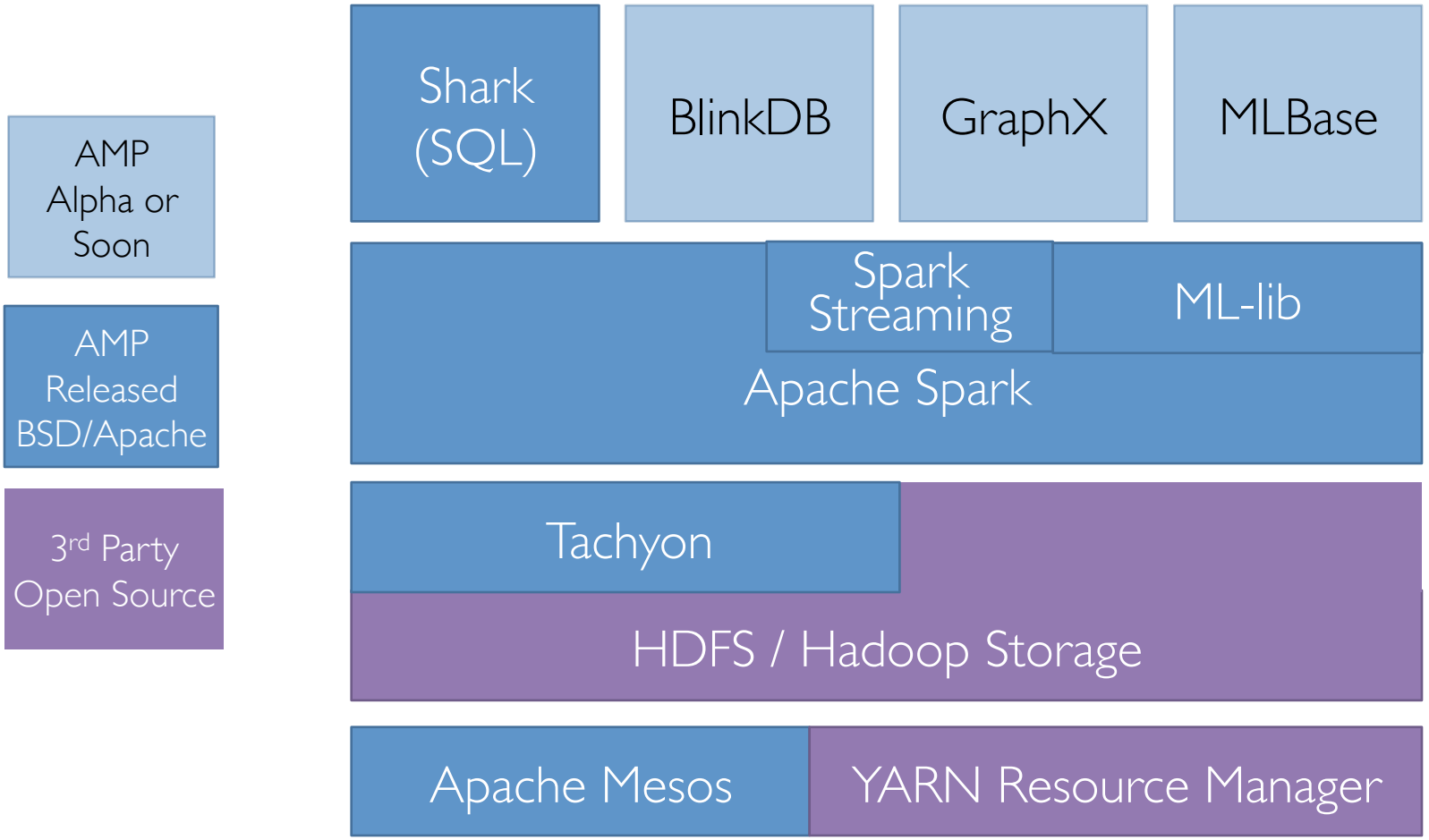
People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts

Big Data Landscape – Our Corner

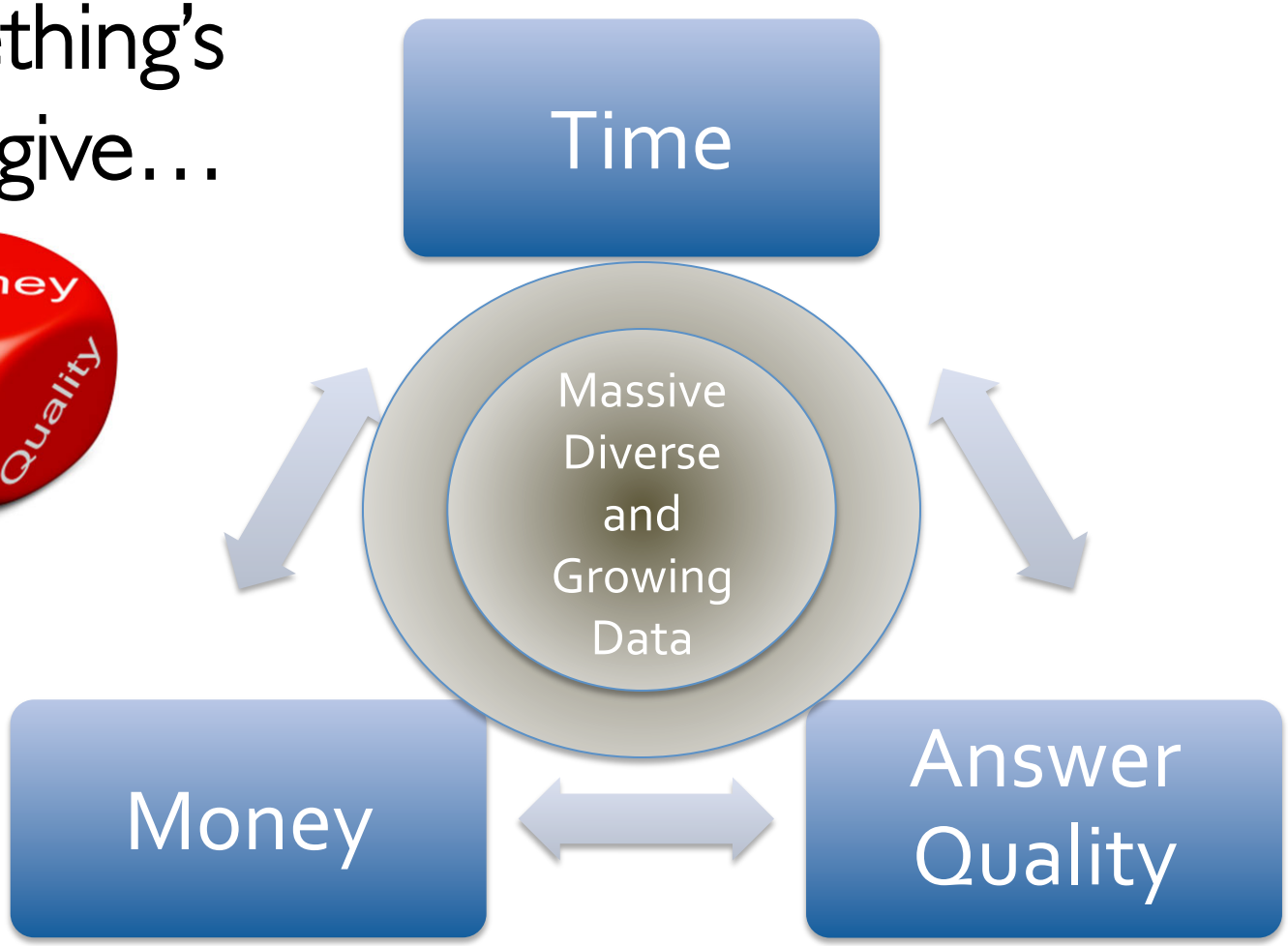


Berkeley Data Analytics Stack

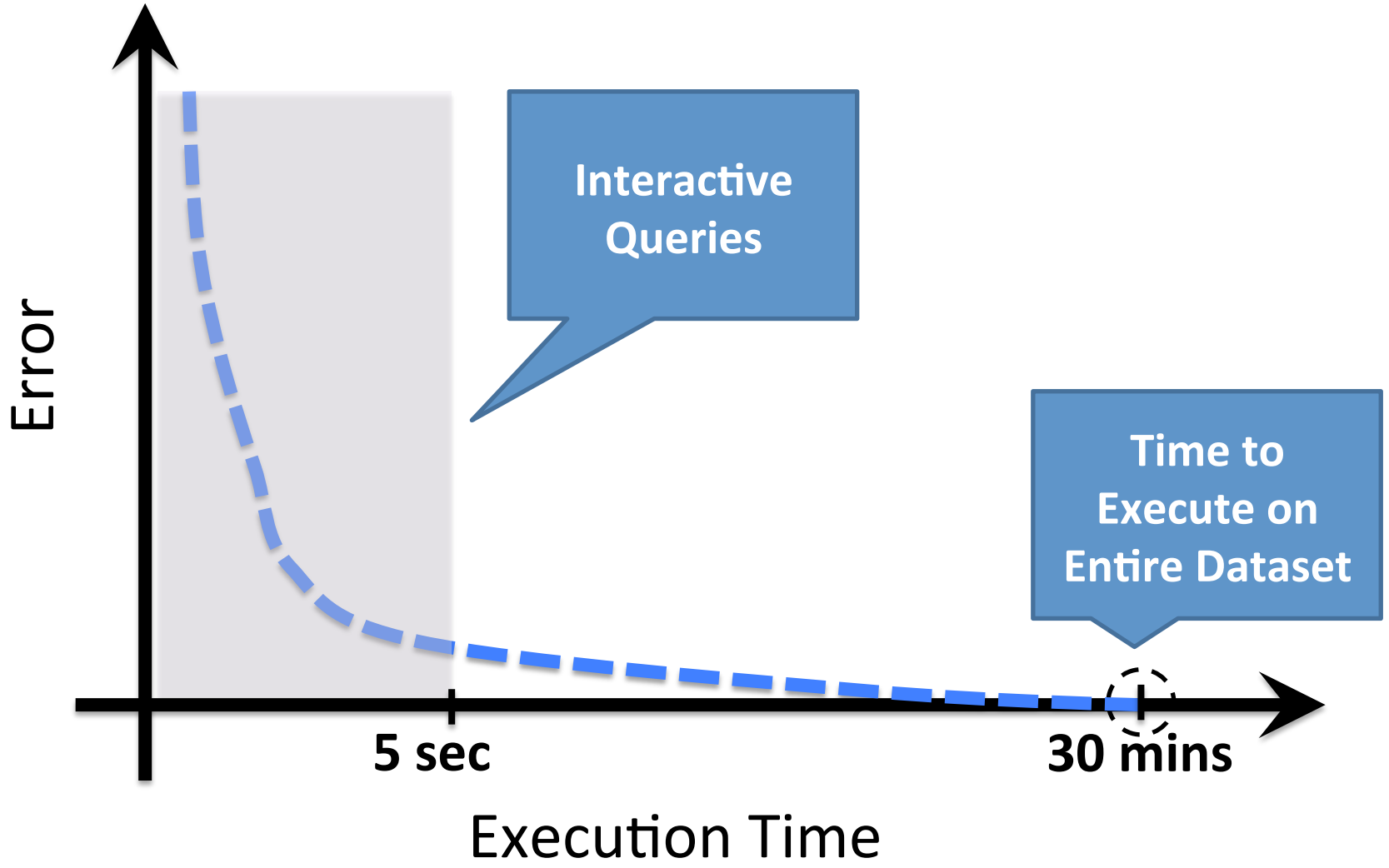


Our View of the Big Data Challenge

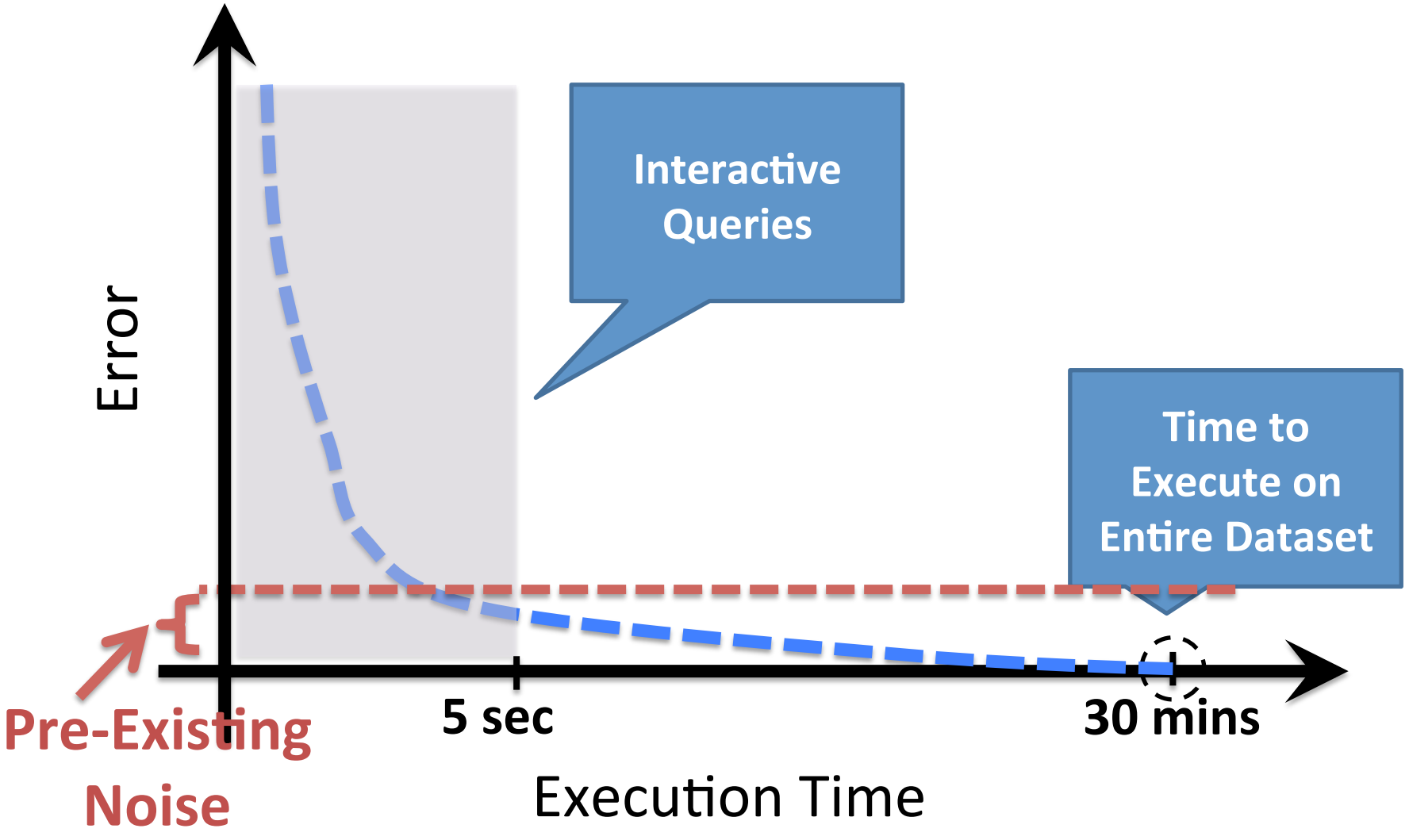
Something's gotta give…



Speed/Accuracy Trade-off



Speed/Accuracy Trade-off

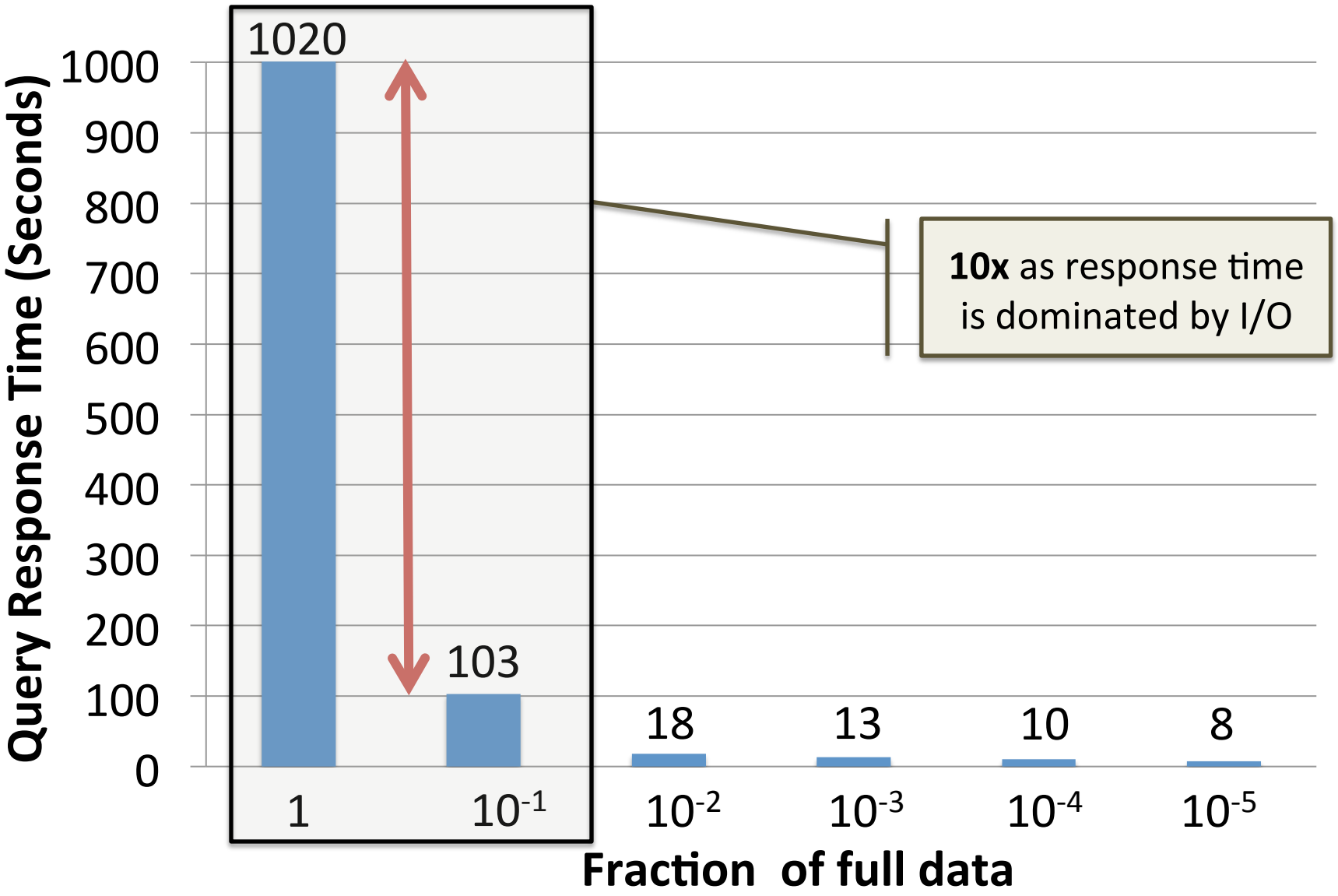


A data analysis (warehouse) system that …

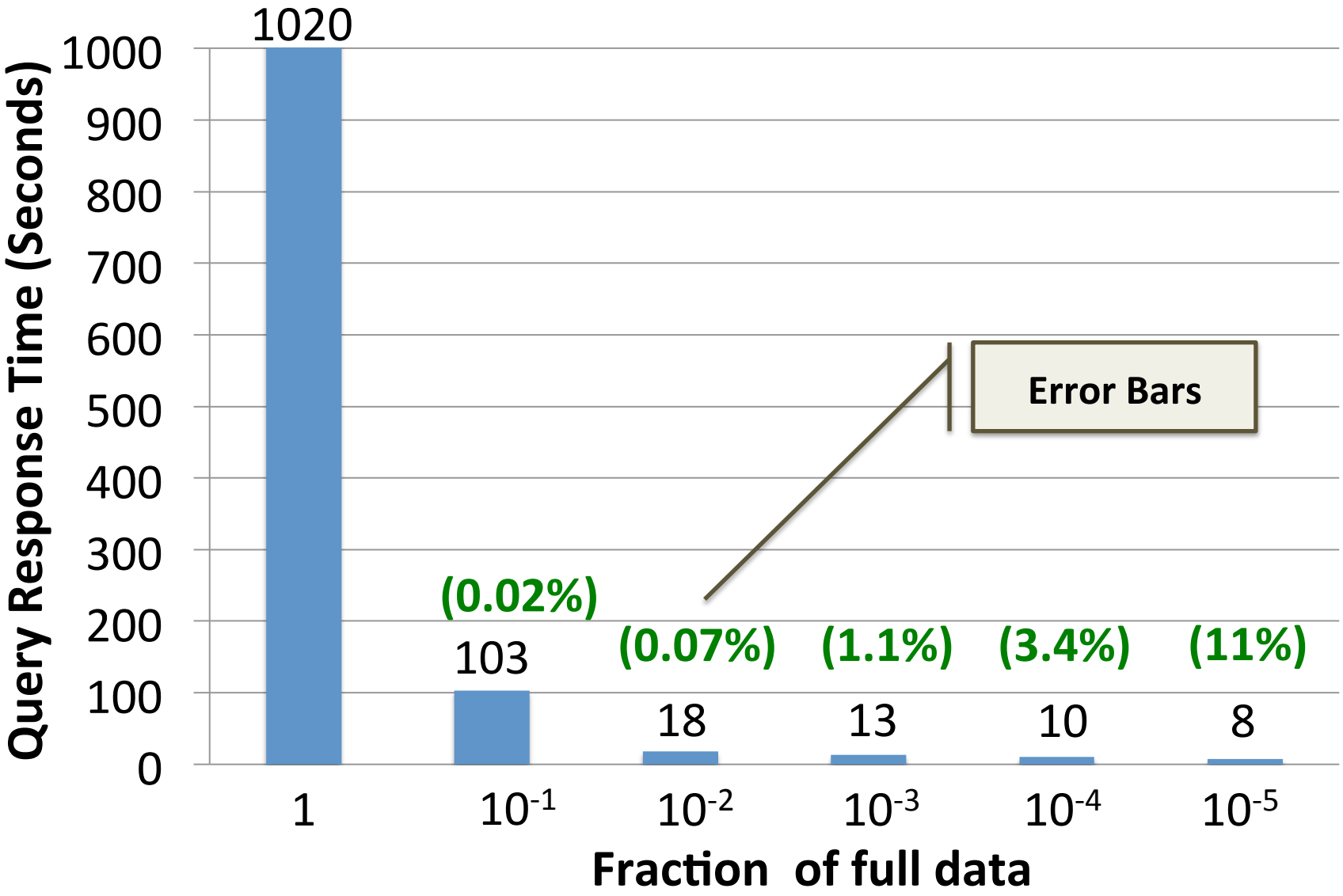
- builds on Shark and Spark
- returns fast, approximate answers with error bars by executing queries on small samples of data
- is compatible with Apache Hive (storage, serdes, UDFs, types, metadata) and supports Hive's SQL-like query structure with minor modifications

Agarwal et al., BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data. *ACM EuroSys 2013, Best Paper Award*

Sampling Vs. No Sampling



Sampling Vs. No Sampling



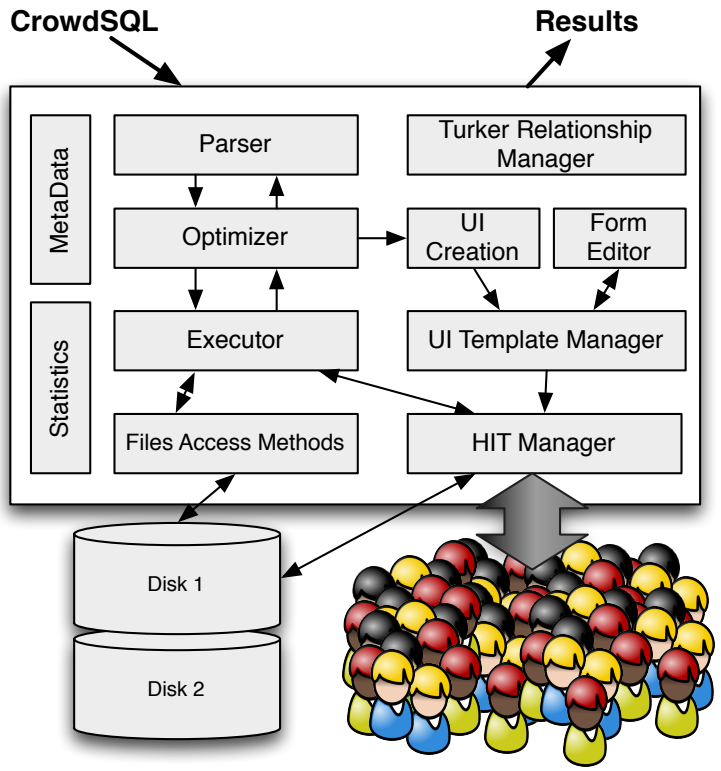
People Resources

Hybrid Human-Machine Computation

- Data Cleaning
- Active Learning
- Handling the last 5%

Supporting Data Scientists

- Interactive Analytics
- Visual Analytics
- Collaboration

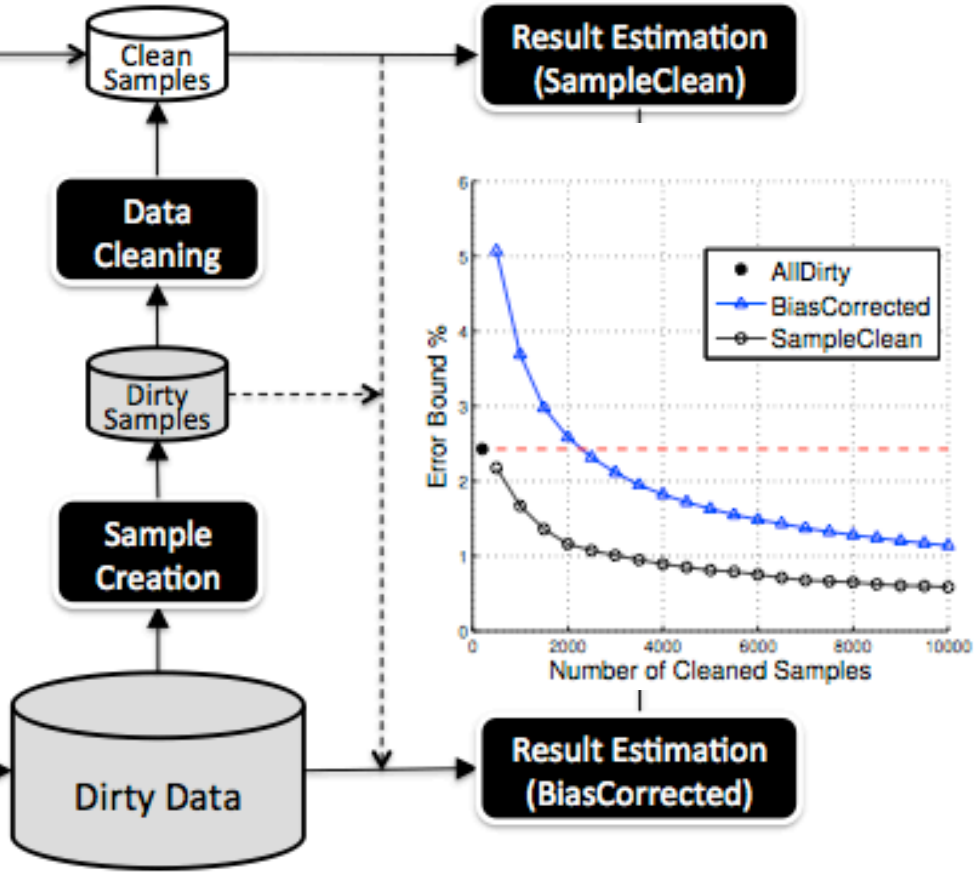


Franklin et al., CrowdDB: Answering Queries with Crowdsourcing, *SIGMOD 2011*
Wang et al., CrowdER: Crowdsourcing Entity Resolution, VLDB 2012
Trushkowsky et al., Crowdsourcing Enumeration Queries, *ICDE 2013 Best Paper Award*

Less is More?
Data Cleaning + Sampling

Aggregation Queries

BlinkDB



J. Wang et al., *Work in Progress*

Working with the Crowd

≠

Incentives

Fatigue, Fraud, & other Failure Modes

Latency & Prediction

Work Conditions

Interface Impacts Answer Quality

Task Structuring

Task Routing

The 3E's of Big Data: Extreme Elasticity Everywhere

Algorithms

- Approximate Answers
- ML Libraries and Ensemble Methods
- Active Learning

Machines

- Cloud Computing – esp. Spot Instances
- Multi-tenancy
- Relaxed (eventual) consistency/ Multi-version methods

People

- Dynamic Task and Microtask Marketplaces
- Visual analytics
- Manipulative interfaces and mixed mode operation

The Research Challenge

AMP Integration +

Extreme Elasticity +

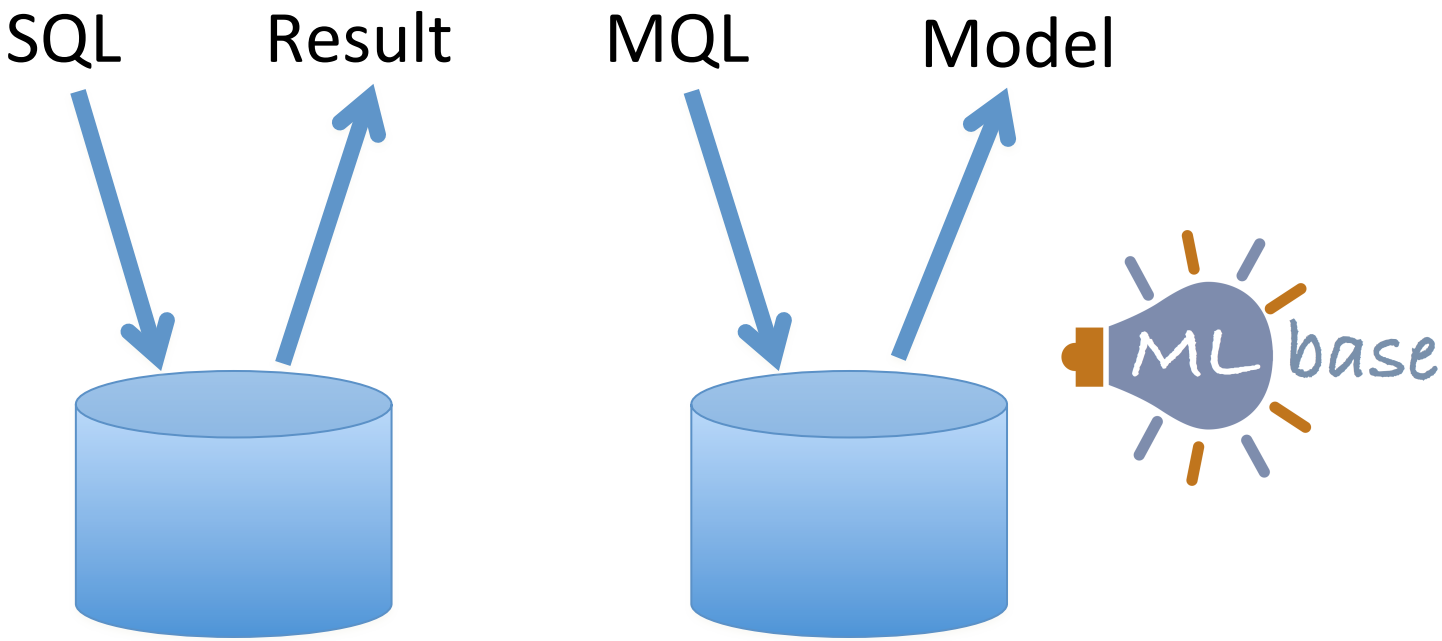
Tradeoffs +

More Sophisticated Analytics

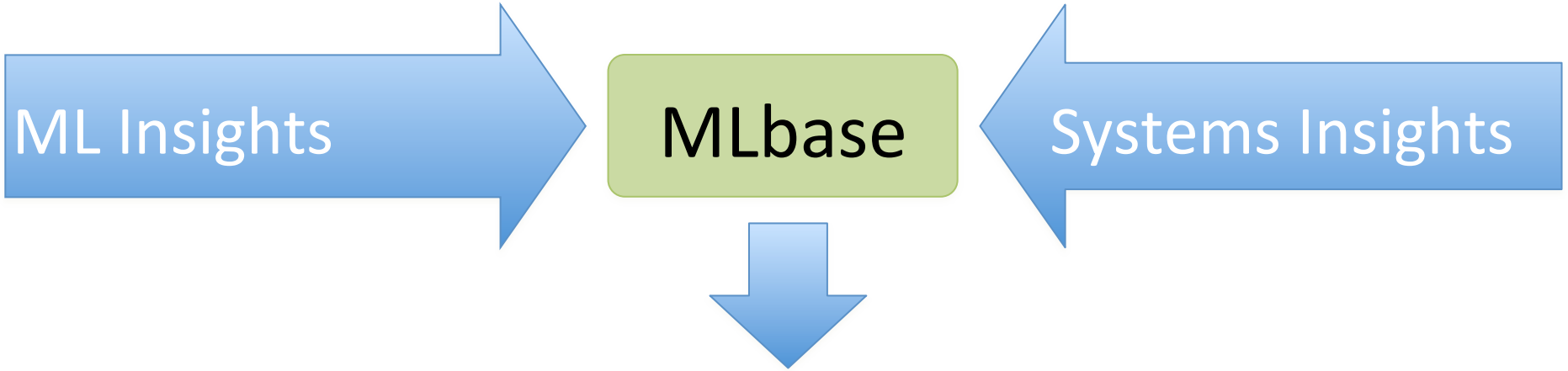
= Extreme Complexity

Can we Take a Declarative Approach?

- Can reduce complexity through automation
- End Users tell the system what they want, not how to get it



Goals of MLbase



1. Easy scalable ML development (ML Developers)
2. Easy/user-friendly ML at scale (End Users)

Along the way, we gain insight into data intensive computing

A Declarative Approach

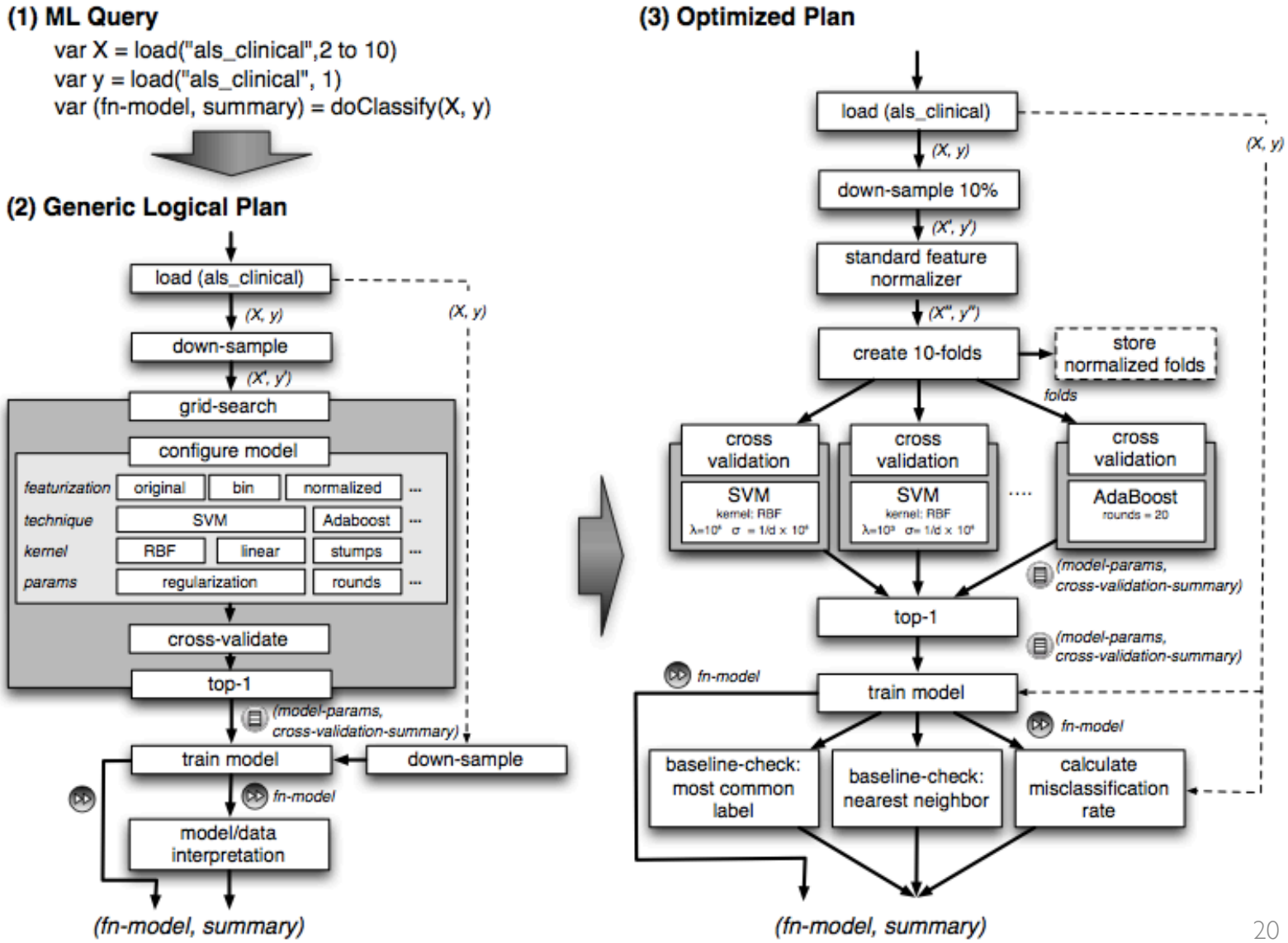
- End Users tell the system what they want, not how to get it

Example: Supervised Cla[illegible]ation

```
var X = load("als_cl[illegible], 2 to 10)
var y = load("[illegible]nical", 1)
var (fn[illegible], summary) = doClassify(X, y)
```

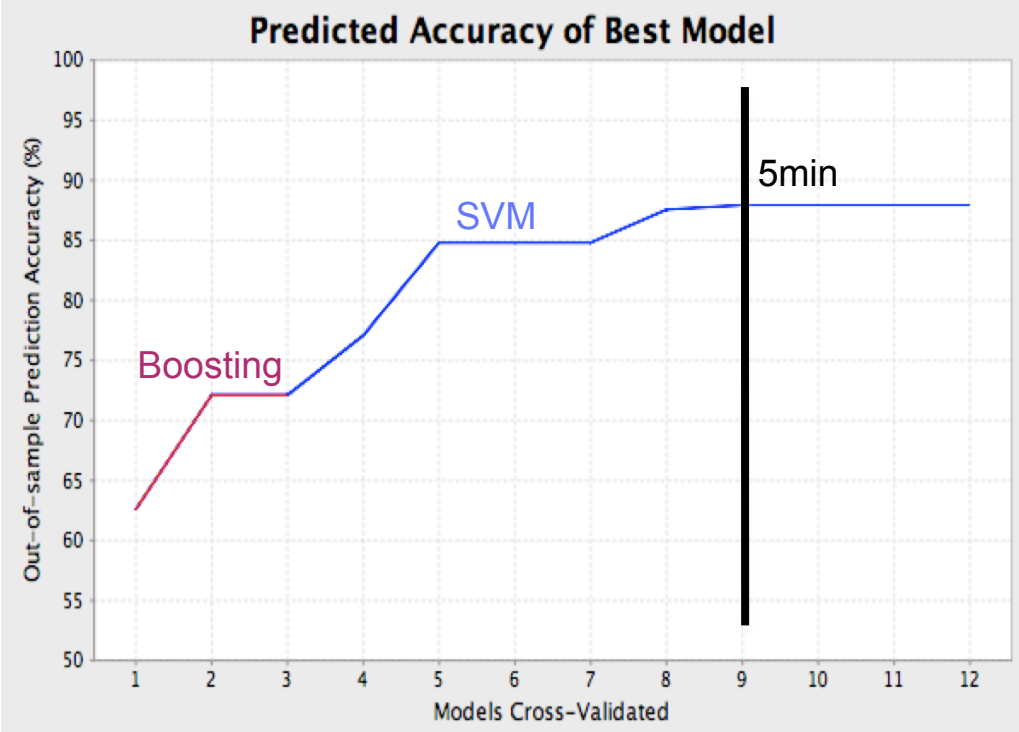
Algorithm Independent

MLBase – Query Compilation

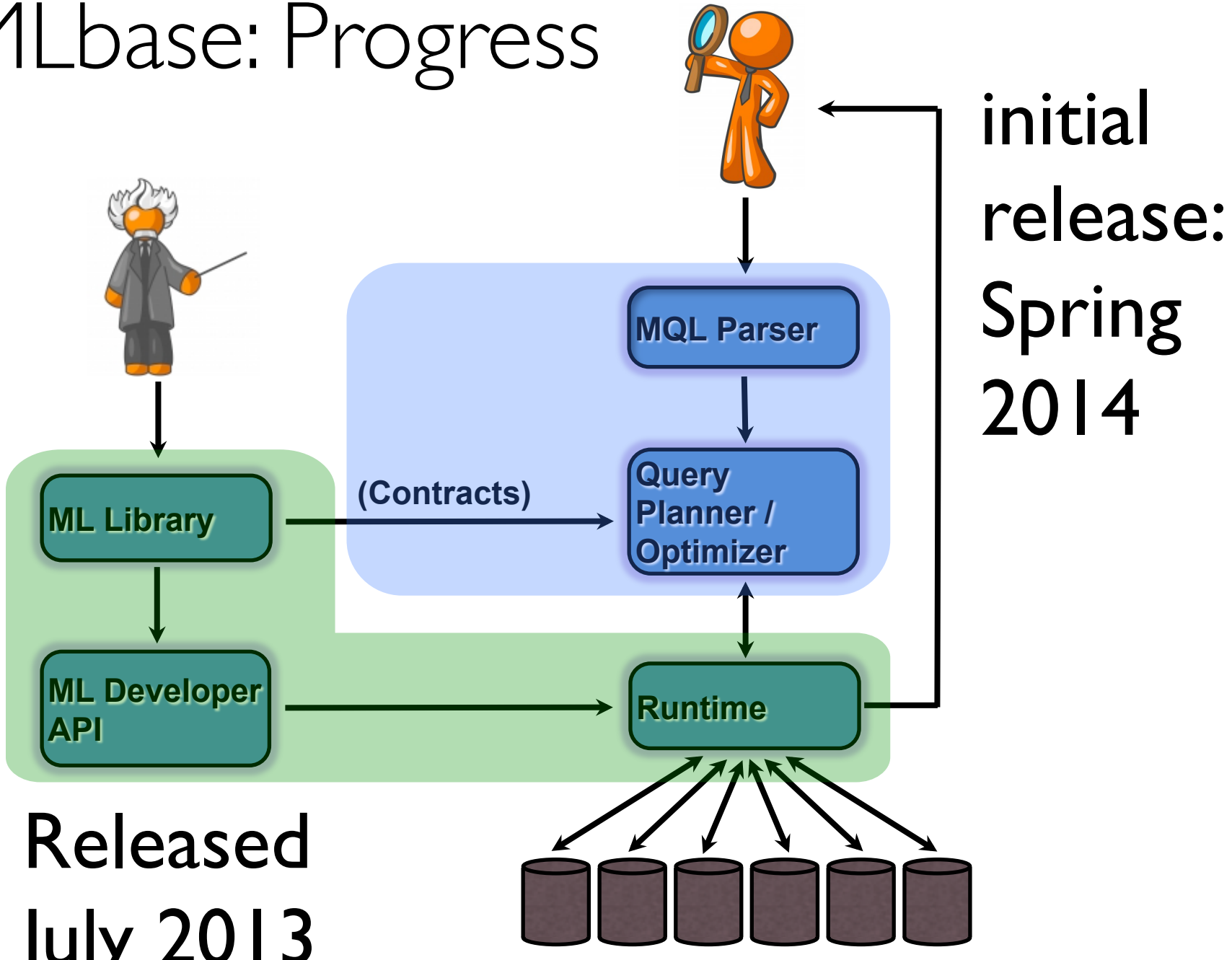


Query Optimizer: A Search Problem

- System is responsible for searching through model space
- Opportunities for physical optimization



MLbase: Progress



initial release: Spring 2014

Released July 2013

Other Things We're Working On

- GraphX: Unifying Graph Parallel & Data Parallel Analytics
- OLTP and Serving Workloads
  - MDCC: Mutli Data Center Consistency
  - HAT: Highly-Available Transactions
  - PBS: Probabilistically Bounded Staleness
  - PLANET: Predictive Latency-Aware Networked Transactions
- Fast Matrix Manipulation Libraries
- Cold Storage, Partitioning, Distributed Caching
- Machine Learning Pipelines, GPUs,
- …

It's Been a Busy 3 Years

Be Sure to Join us for the Next 3

amplab.cs.berkeley.edu

@amplab