SampleClean: Bringing Data Cleaning into the BDAS Stack

Sanjay Krishnan and **Daniel Haas**

In Collaboration With: Juan Sanchez, Wenbo Tao, Jiannan Wang, Tim Kraska, Michael Franklin, Tova Milo, Ken Goldberg

Who publishes more?

Microsoft Academic Search

Paper Id	Affiliation
16	Computer Science Division--University of California Berkeley CA
101	University of California at Berkeley
102	Department of Physics Stanford University California
116	Lawrence Berkeley National Labs <ref>California</ref>

Microsoft Academic Search

Paper Id	Affiliation
16	Computer Science Division University of California Berkeley CA
101	University of California at Berkeley
102	Department of Physics Stanford University California
116	Lawrence Berkeley National Labs ~~<ref>California</ref>~~

Microsoft Academic Search

University of California at Berkeley

Computer Science Division
University of California at Berkeley

Department of Physics Stanford
University California

Enter SampleClean

- Data cleaning in BDAS.
  - Problem 1. Scale
  - Problem 2. Latency

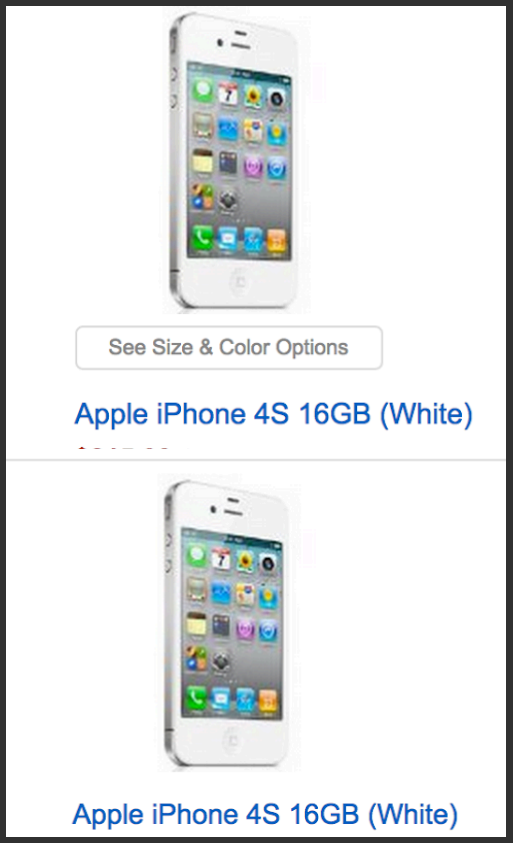
- Sampling to cope with scale.
- Asynchrony to cope with latency.

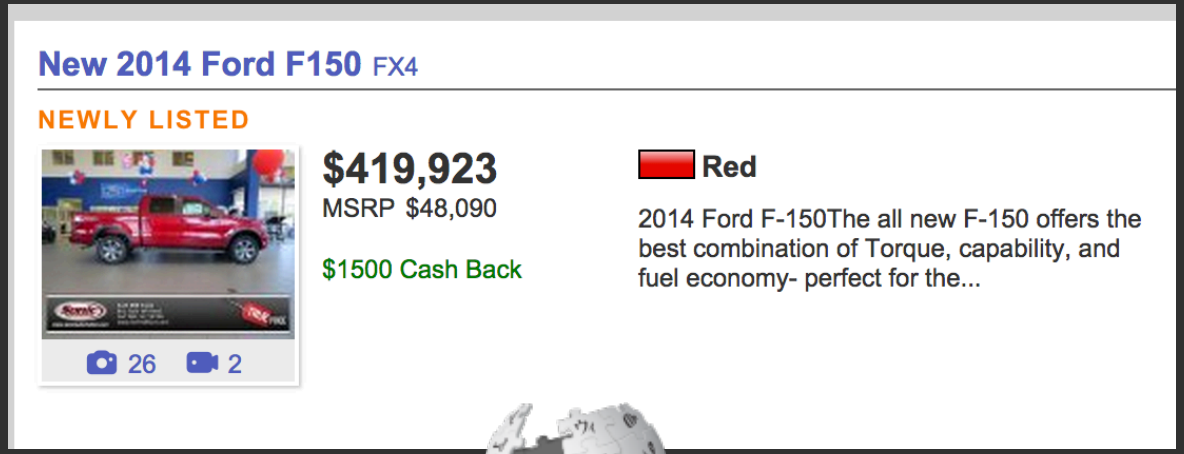
Now it’s your turn!

Be the crowd and help us decide

Dirty Data is Ubiquitous

Example: Missing, incomplete, inconsistent data





Country	UN R/P 10%[4]	UN R/P 20%[5]	World Bank Gini (%)[6]	WB Gini (year)	CIA R/P 10%[7]	Year	CIA Gini (%)[8]	CIA Gini (year)	GPI Gini (%)[9]
Seychelles			65.8	2007					
Comoros			64.3	2004					
Namibia	[illegible]	56.1	63.9	2004	129.0	2003	59.7	2010	
South Africa	33.1	17.9	63.1	2009	31.9	2000	65.0	2005	
Botswana	43.0	20.4	61.0	1994			63	1993	
Haiti	54.4	26.6	59.2	2001	68.1	2001	59.2	2001	
Angola			58.6	2000					62.0
Honduras	59.4	17.2	57.0	2009	35.2	2003	57.7	2007	

Data Cleaning is Hard

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Time consuming

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Data Cleaning is Hard

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Time consuming

Costly

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Data Cleaning is Hard

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Time consuming

Costly

Domain-specific

Data Cleaning is Hard

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Time consuming

Costly

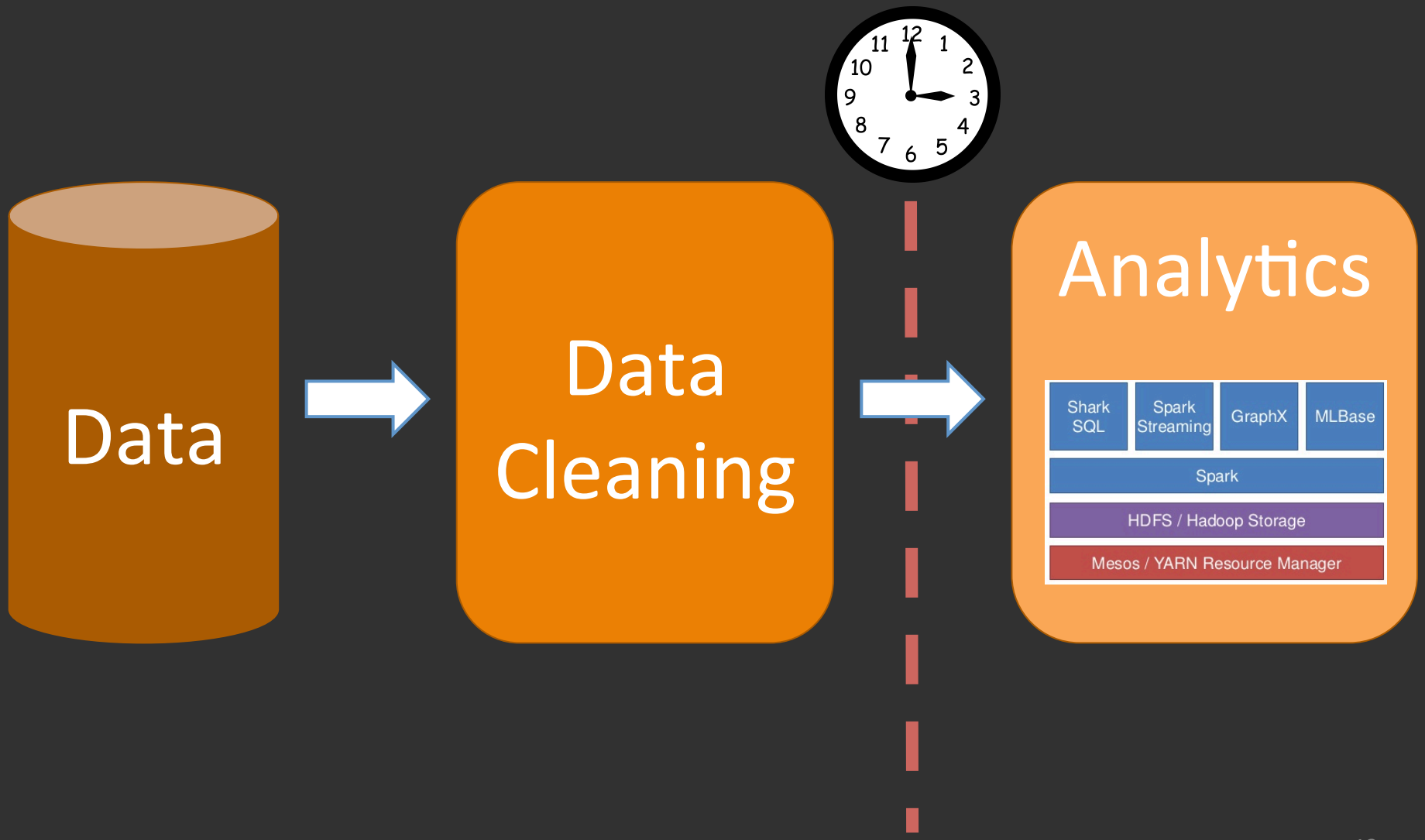
Domain-specific

amazon mechanical turk™
Artificial Artificial Intelligence

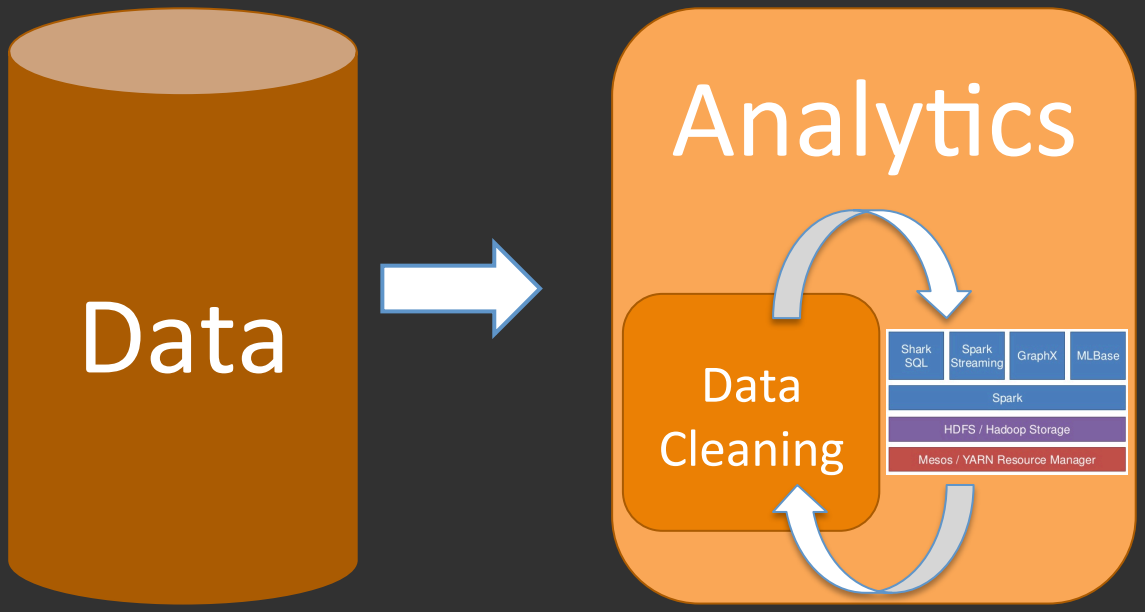
oDesk® samasource

mobileworks

A New Data Cleaning Architecture

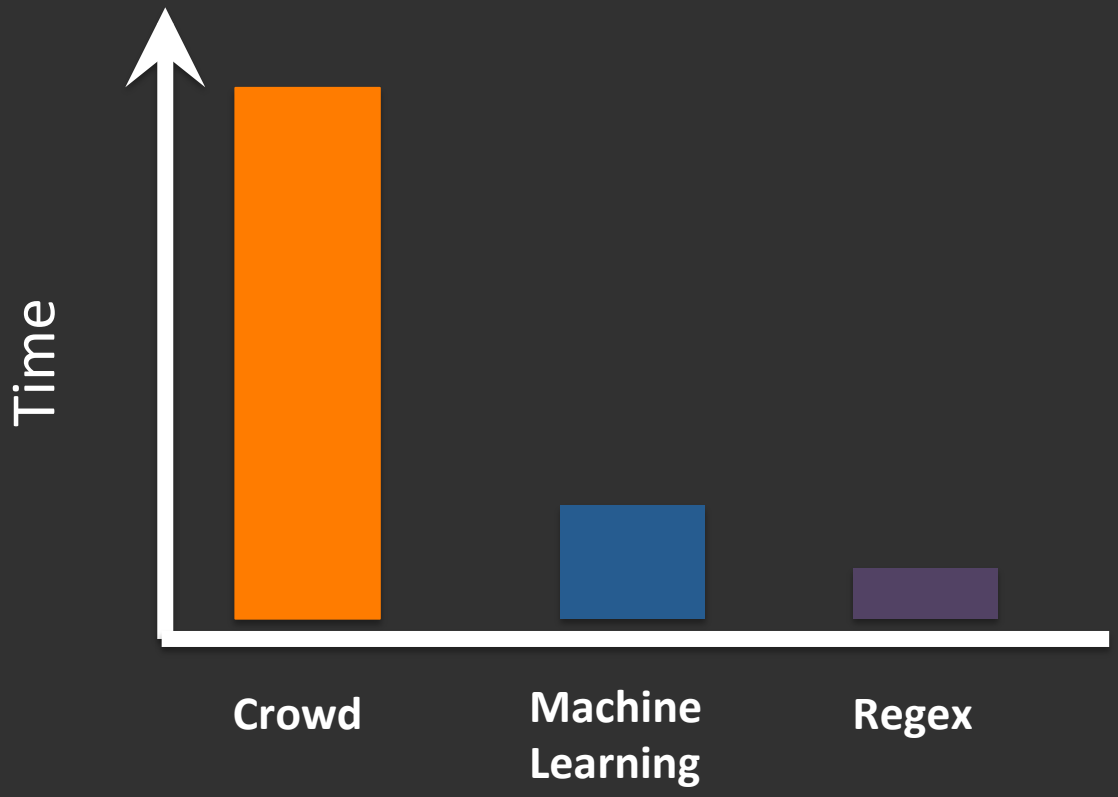


A New Data Cleaning Architecture

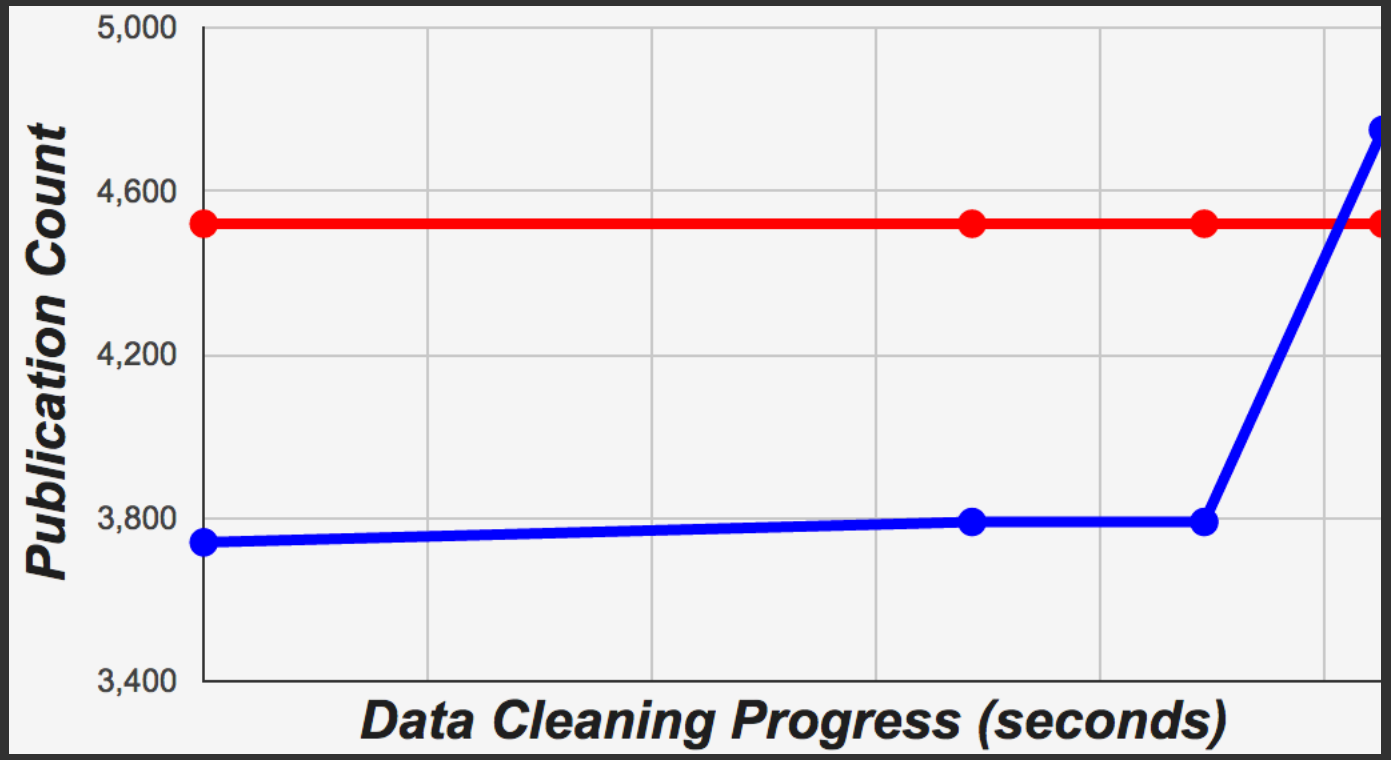


Can it Scale?

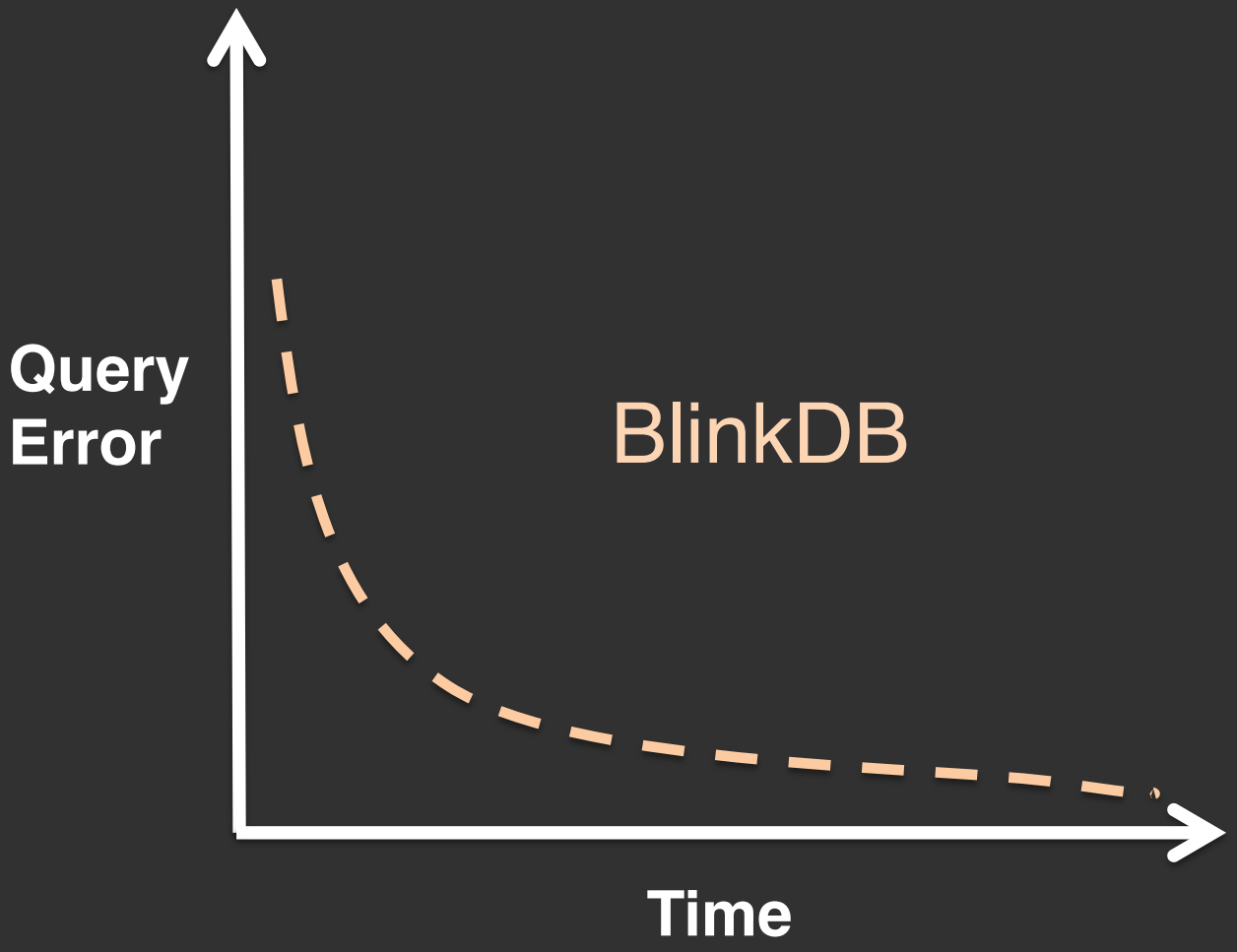
People are **slow** and **expensive**



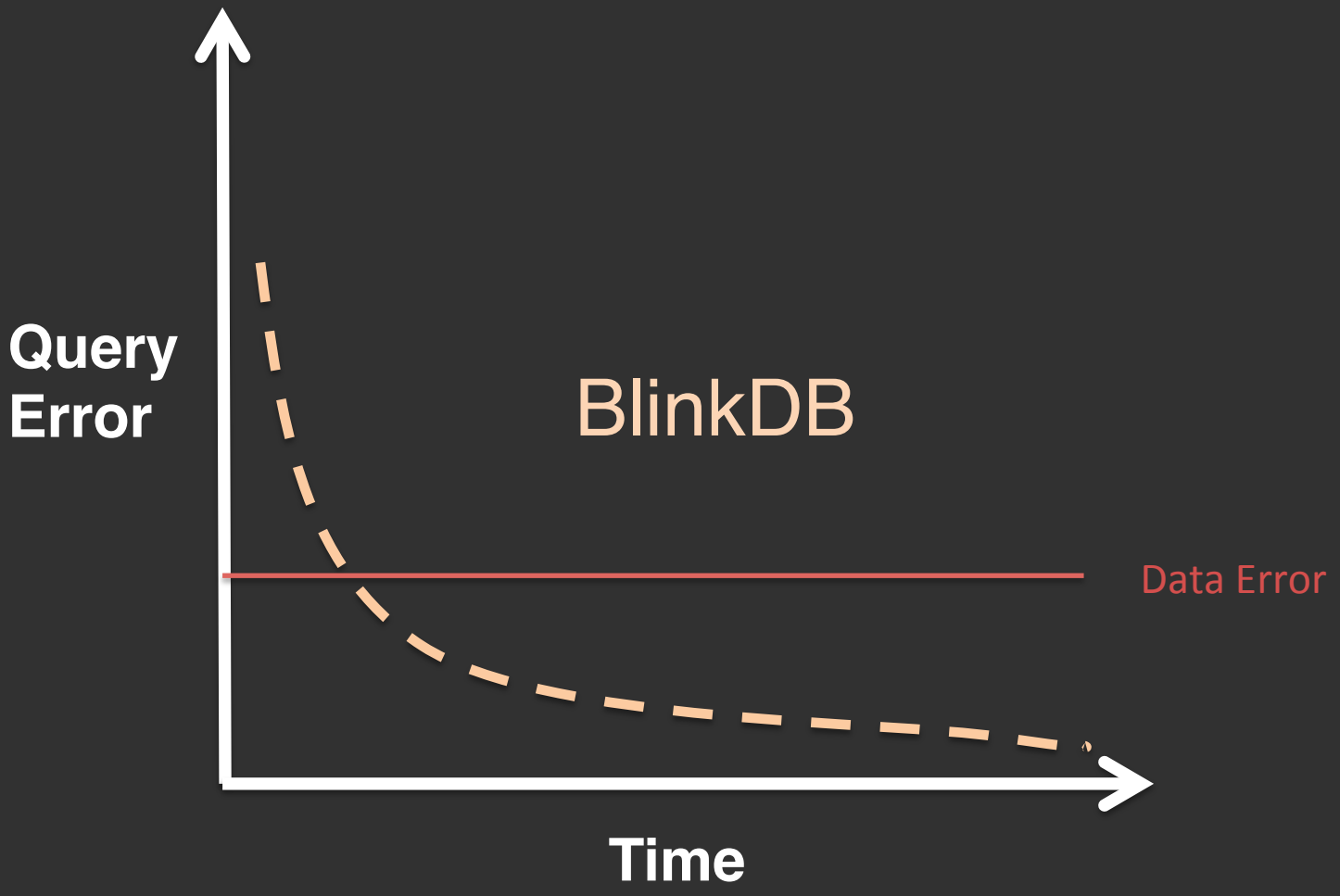
Insight 1: Asynchrony Hides Latency



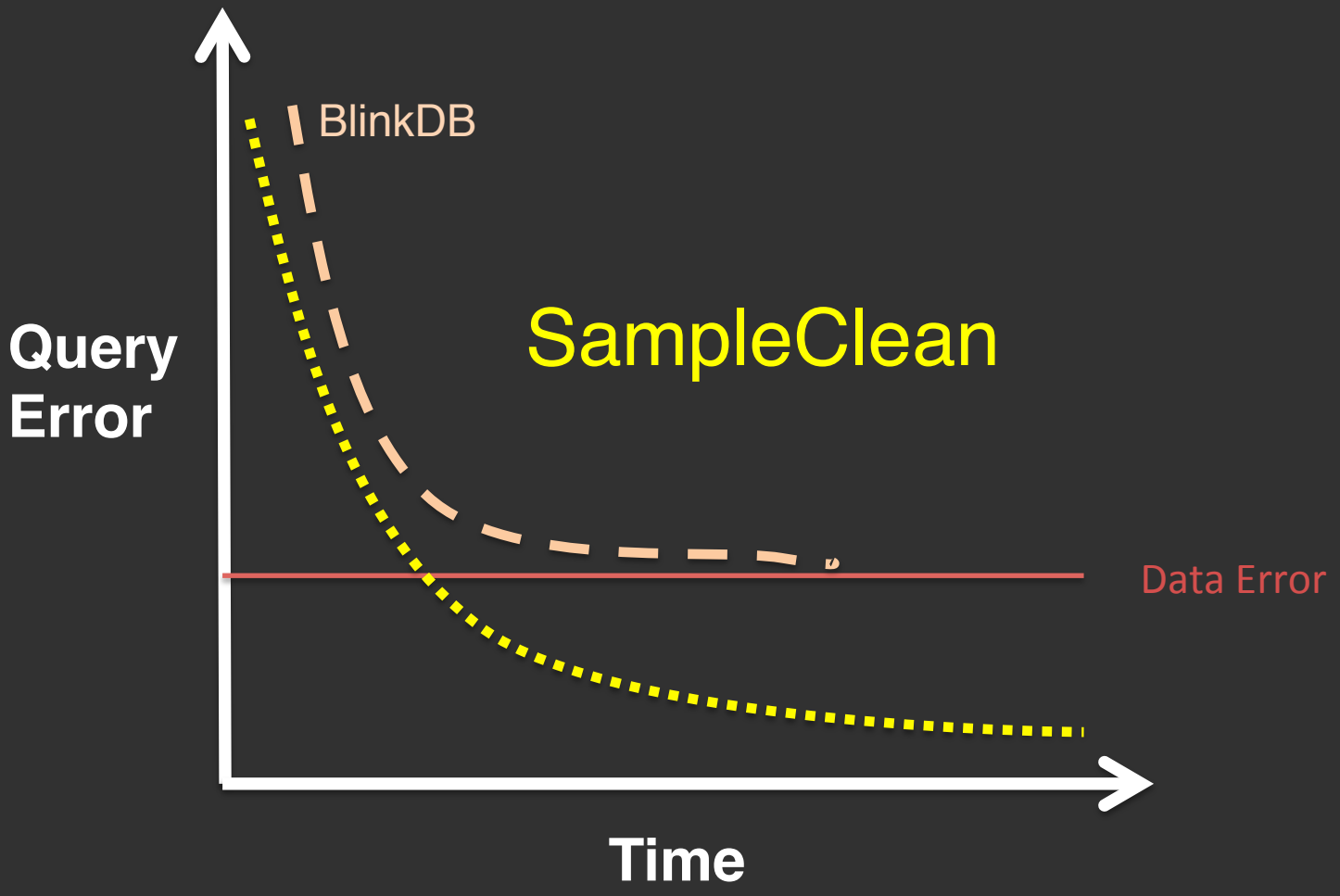
Insight 2: Sampling Hides Scale



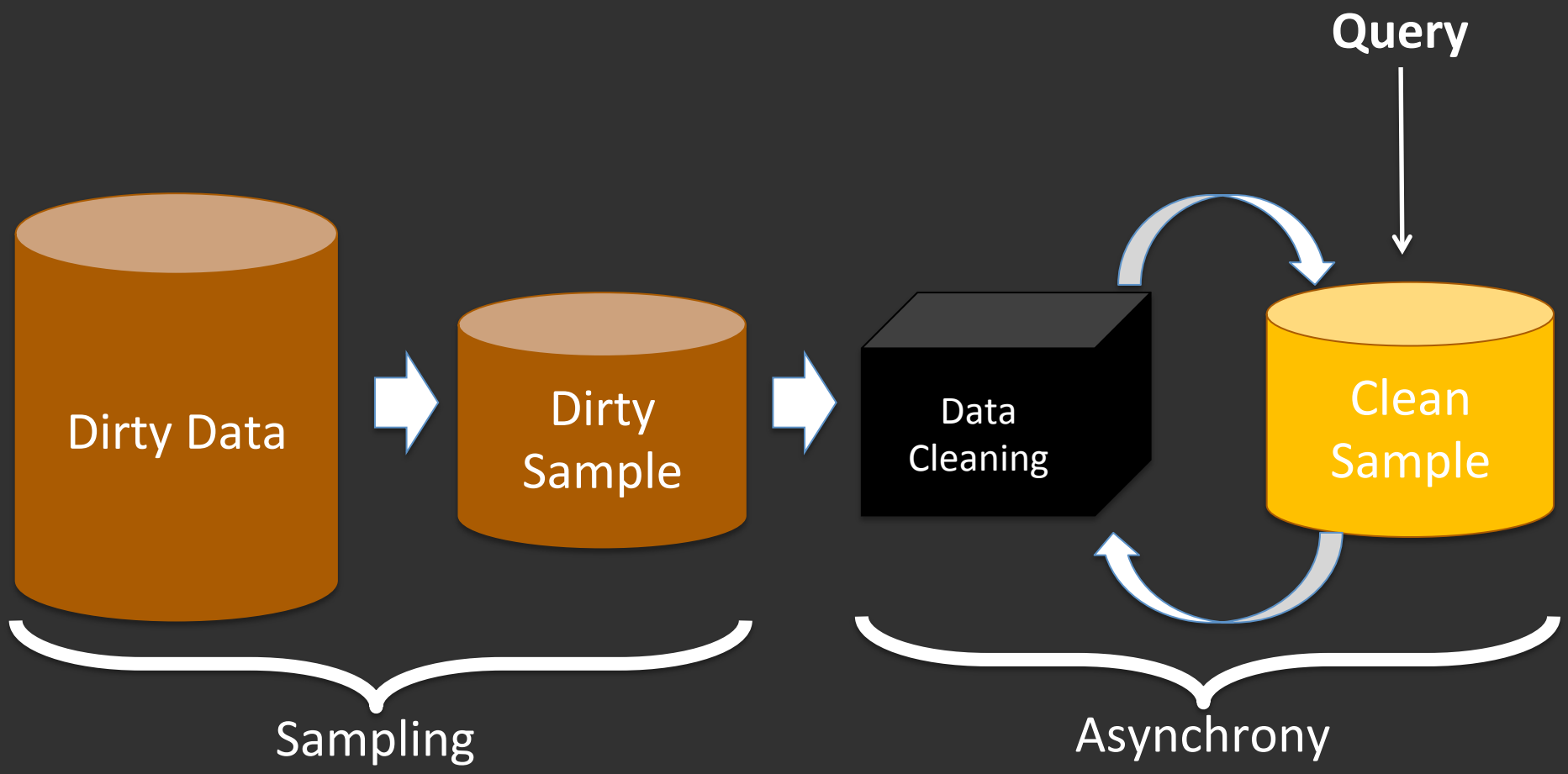
Insight 2: Sampling Hides Scale



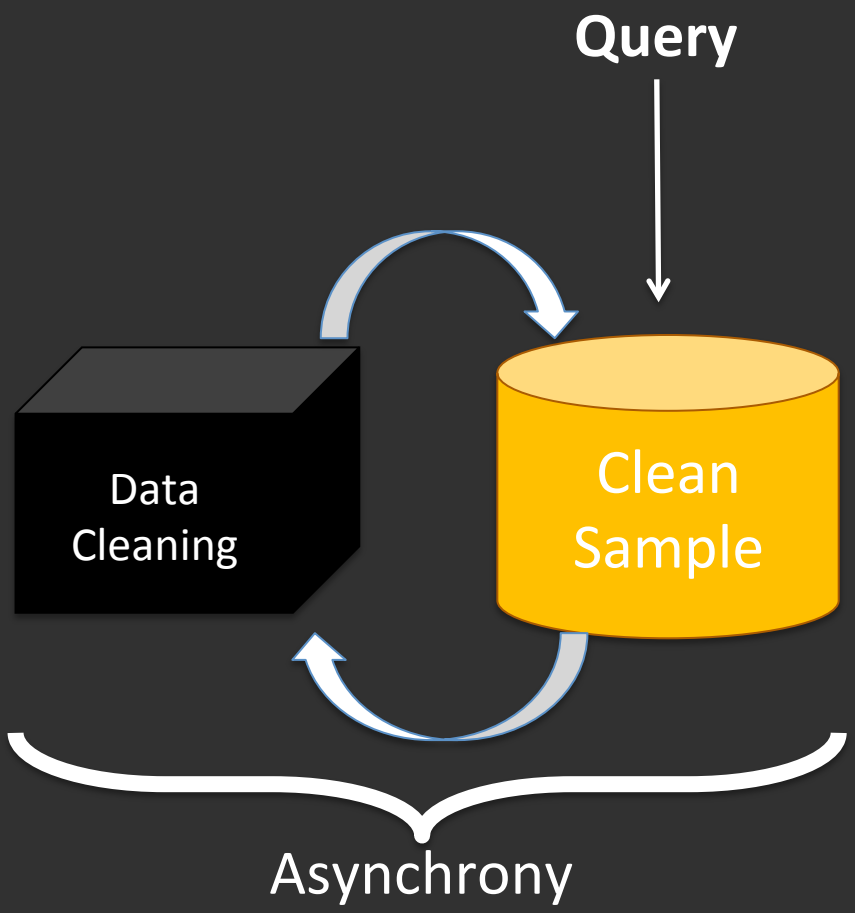
Insight 2: Sampling Hides Scale



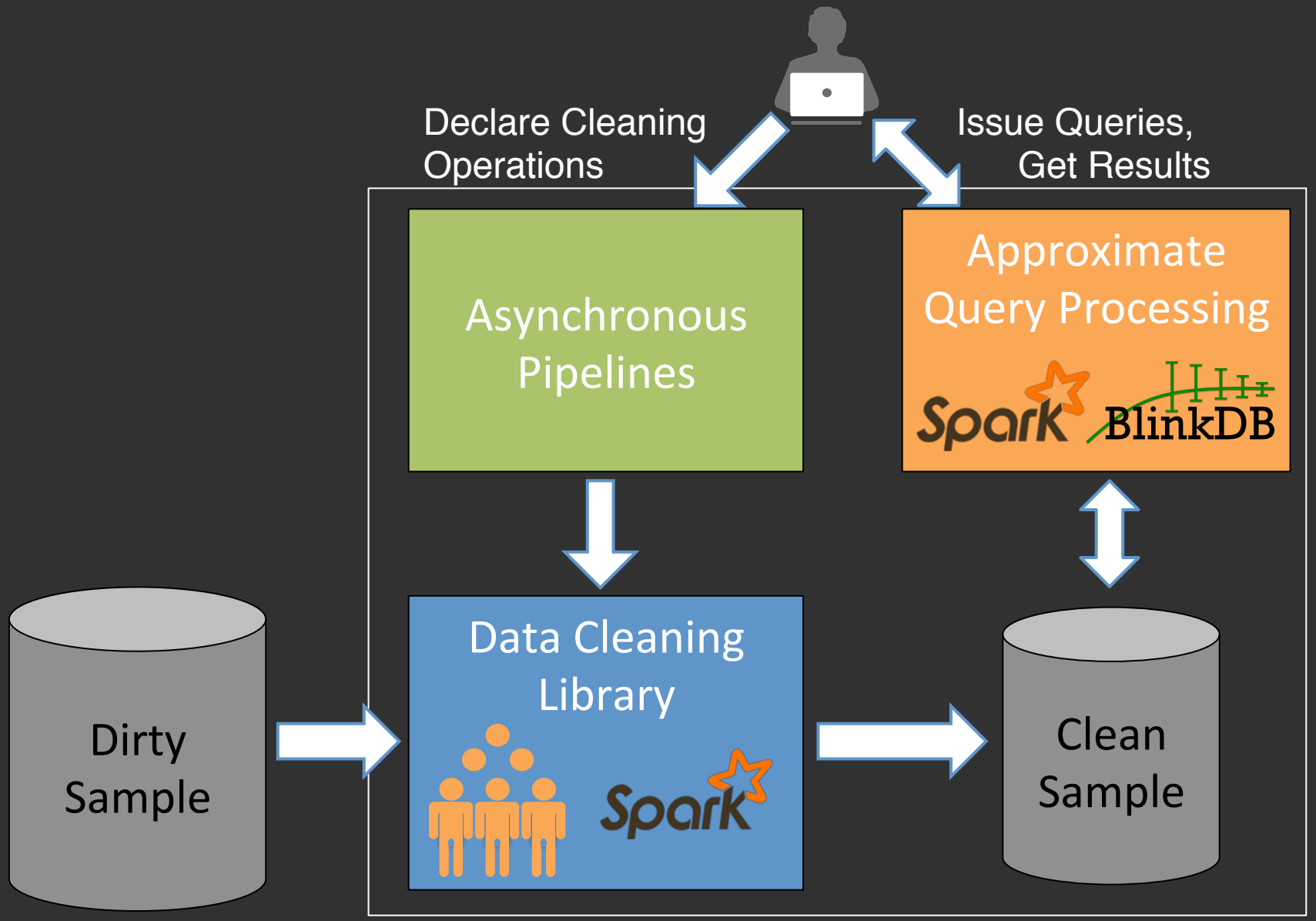
SampleClean Data Flow



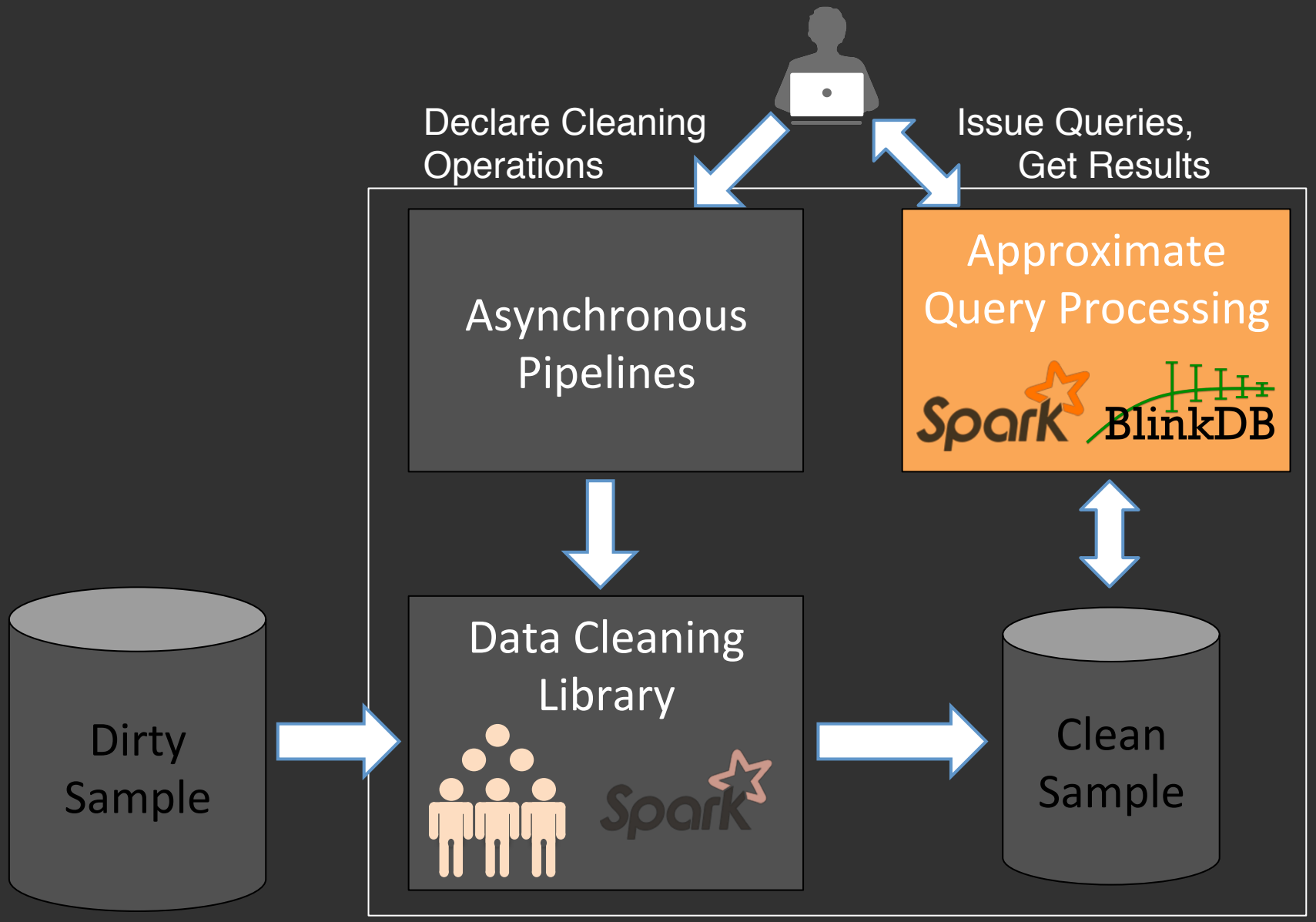
SampleClean Data Flow



The SampleClean Architecture

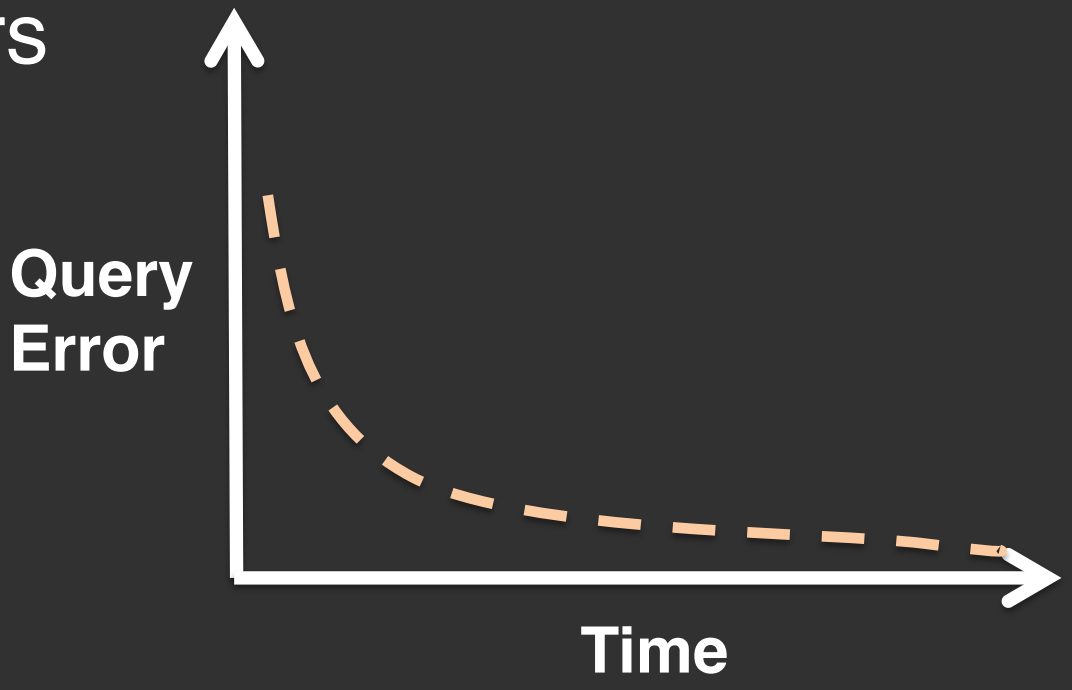


The SampleClean Architecture



Approximate Query Processing

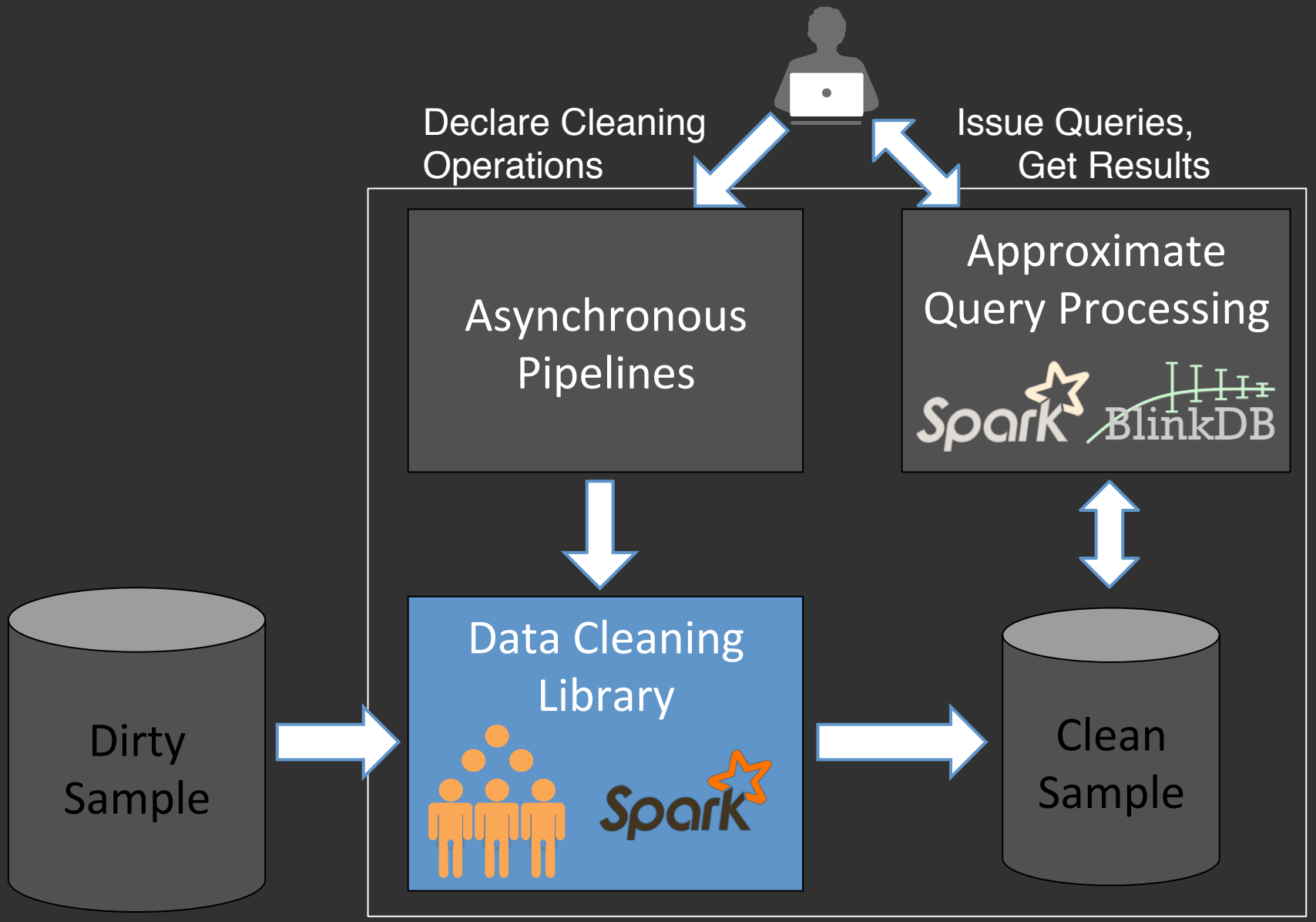
- Estimate early results and bound with error bars



SampleClean: Fast and Accurate Query Processing on Dirty Data. SIGMOD 2014

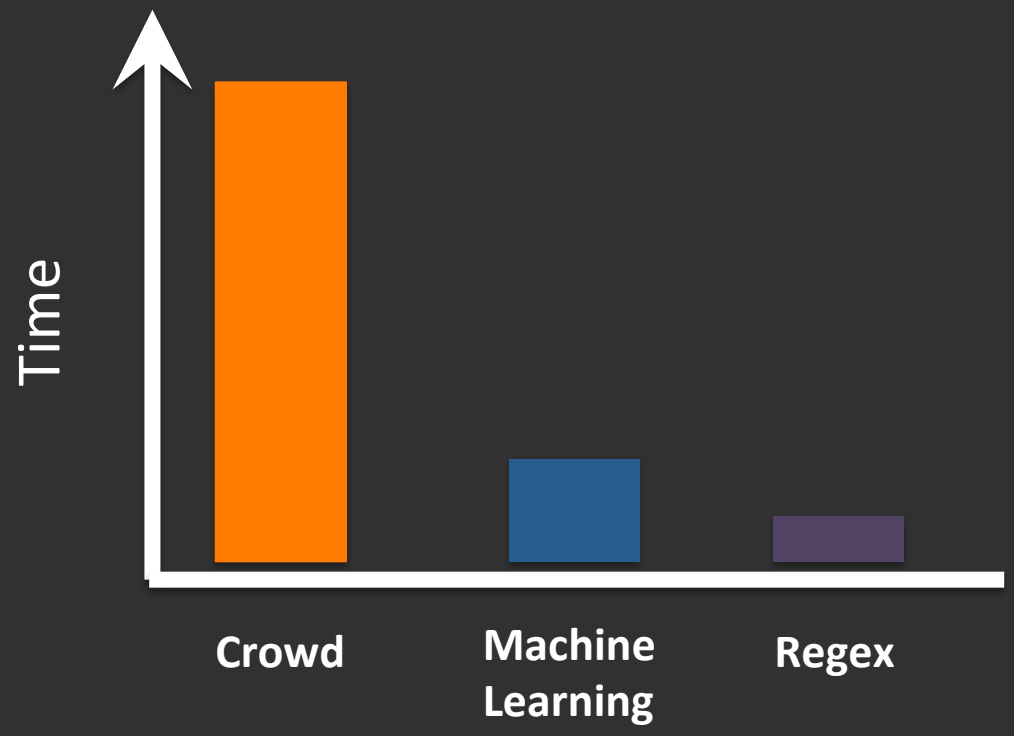
BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data. EuroSys 2013

The SampleClean Architecture



Crowds and Machines Work Together

- Extensible library of data cleaning tools
- Tools are:
  - Automated
  - Human-powered
  - Hybrid



Active Learning and Crowds

- Choose informative training points

Not Informative

Are these the same?
Stanford Department of IEOR

UC Berkeley Stats

- ○ Yes
- ◉ No

Informative

Are these the same?
Department of Mathematics Stanford University

University of California Berkeley Department of Mathematics

- ○ Yes
- ◉ No

Active Learning and Crowds

- Choose informative training points

Not Informative

Are these the same?
Stanford Department of IEOR

UC Berkeley Stats

○ Yes
⊙ No

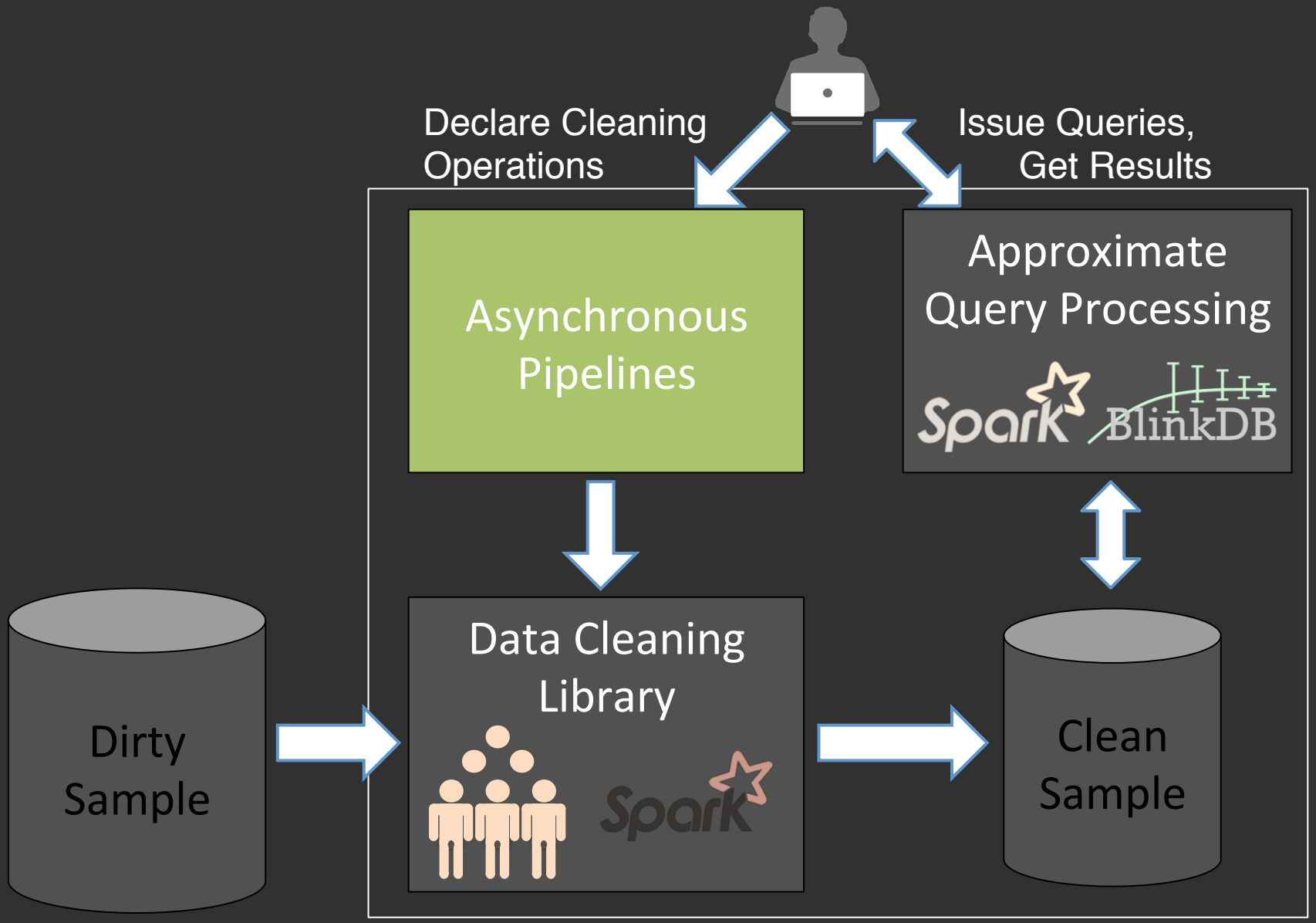
Informative

Are these the same?
Department of Mathematics Stanford University

University of California Berkeley Department of Mathematics

○ Yes
⊙ No

The SampleClean Architecture

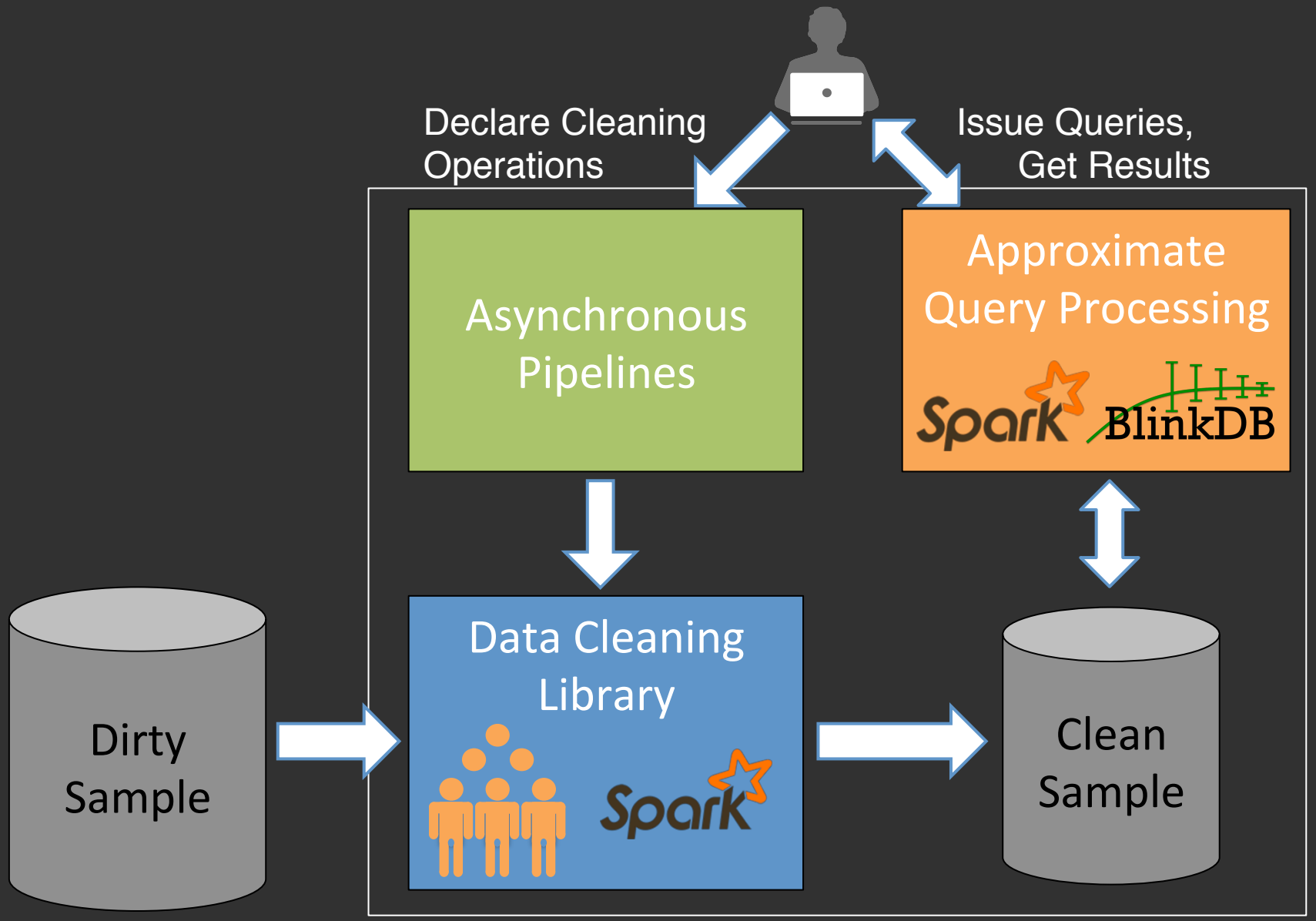


Putting it all together: Asynchronous Pipelines

- Users group data cleaning operations into pipelines

```
val p = new SampleCleanPipeline(fixstrings,
                                normalization)
p.exec("paper_sample")
p.register("SELECT count(*) FROM paper_sample
            GROUP BY affiliation",
            on_result_change_callback)
```

The SampleClean Architecture



Great, Now What?

- Prototype implementation complete!
- Significant research challenges remain:
  - Crowd worker performance and quality
  - Pipeline semantics and optimization
  - Programming model and interface
- Open source release targeted for next year

Summary

- Data Cleaning is **slow**, **costly**, and **domain-specific**
- **SampleClean** brings **data cleaning** into the **BDAS stack**
- SampleClean uses **asynchrony to hide latency**, and **sampling to hide scale**
- SampleClean combines **Algorithms**, **Machines**, and **People**, all in one system

Asynchrony in Spark

- The Spark abstraction: blocking BSP

- So how do we achieve asynchrony?
  - **Multithreaded master**
  - Intermediate results **materialized in Hive**
  - Standalone **Finagle HTTP server** for crowd work