

PRESENTED BY



strataconf.com

[#StrataHadoop](https://twitter.com/StrataHadoop)

SAMOA

A Platform for Mining Big Data Streams

Gianmarco De Francisci Morales

Yahoo Labs Barcelona

gdfm@apache.org

[@gdfm7](https://twitter.com/gdfm7)

Agenda

- **Streams**

- Applications, Model, Tools

- **SAMOA**

- Goal, Architecture, Advantages

Research Scientist @ Yahoo Labs

Web mining &
data-intensive
scalable computing

Committer @ Apache Pig

Contributor for Hadoop,
Giraph, S4, Grafos.ml

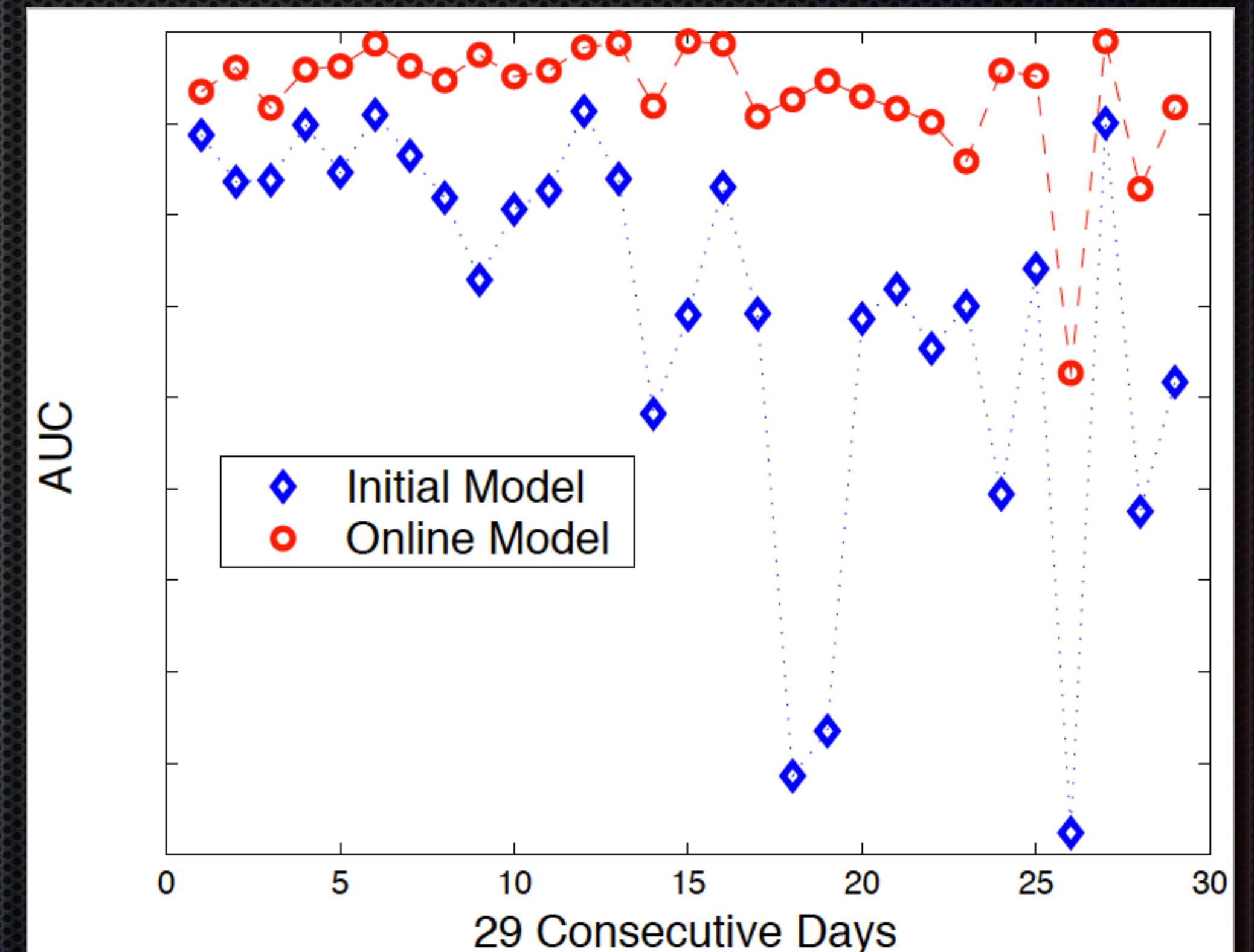


“Panta rhei” (everything flows)


–*Heraclitus*

Importance

- ✦ Spam detection in comments on Yahoo! News
- ✦ Trends change in time
- ✦ Need to retrain model with new data




Spam on Twitter



Michael Wharton @EarthquakeTest · 20 Sep 2011
ping.fm/a7mJR #napa auto parts #earthquake #blackfriday2011
#cybermonday2011 #sep11 #hallowinRT @deedee... bit.ly/reMvRZ

Expand

Reply Retweet Favorite Pocket



Abu Bander @aser44444 · Aug 22
#teamfollowback
islam-guide.com
quran.ksu.edu.sa/index.php?l=en...
#相互フォロー #vma
#iphone
#tbt
#sougofollow
#sex

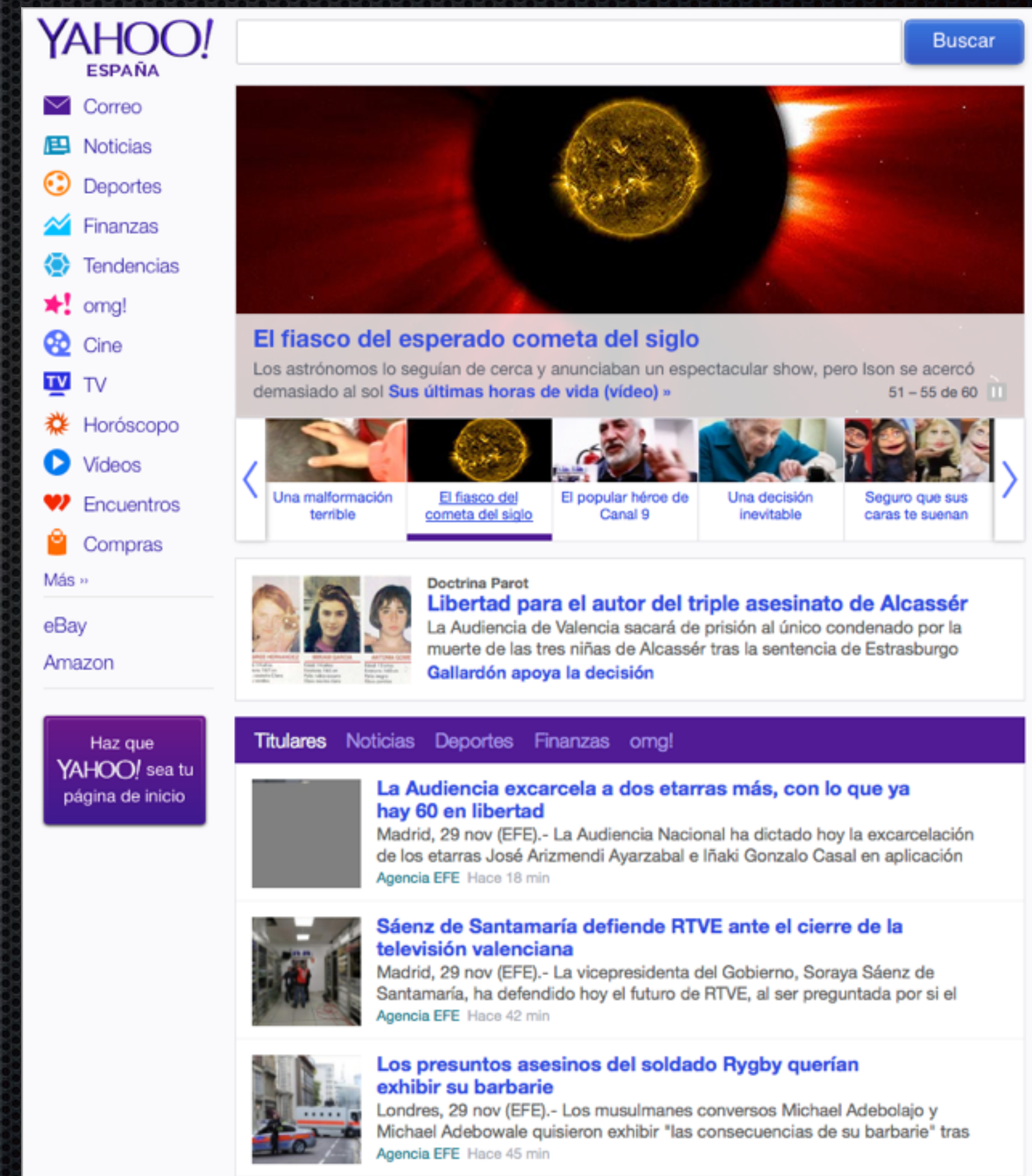
Expand

Reply Retweet Favorite

Applications

Applications

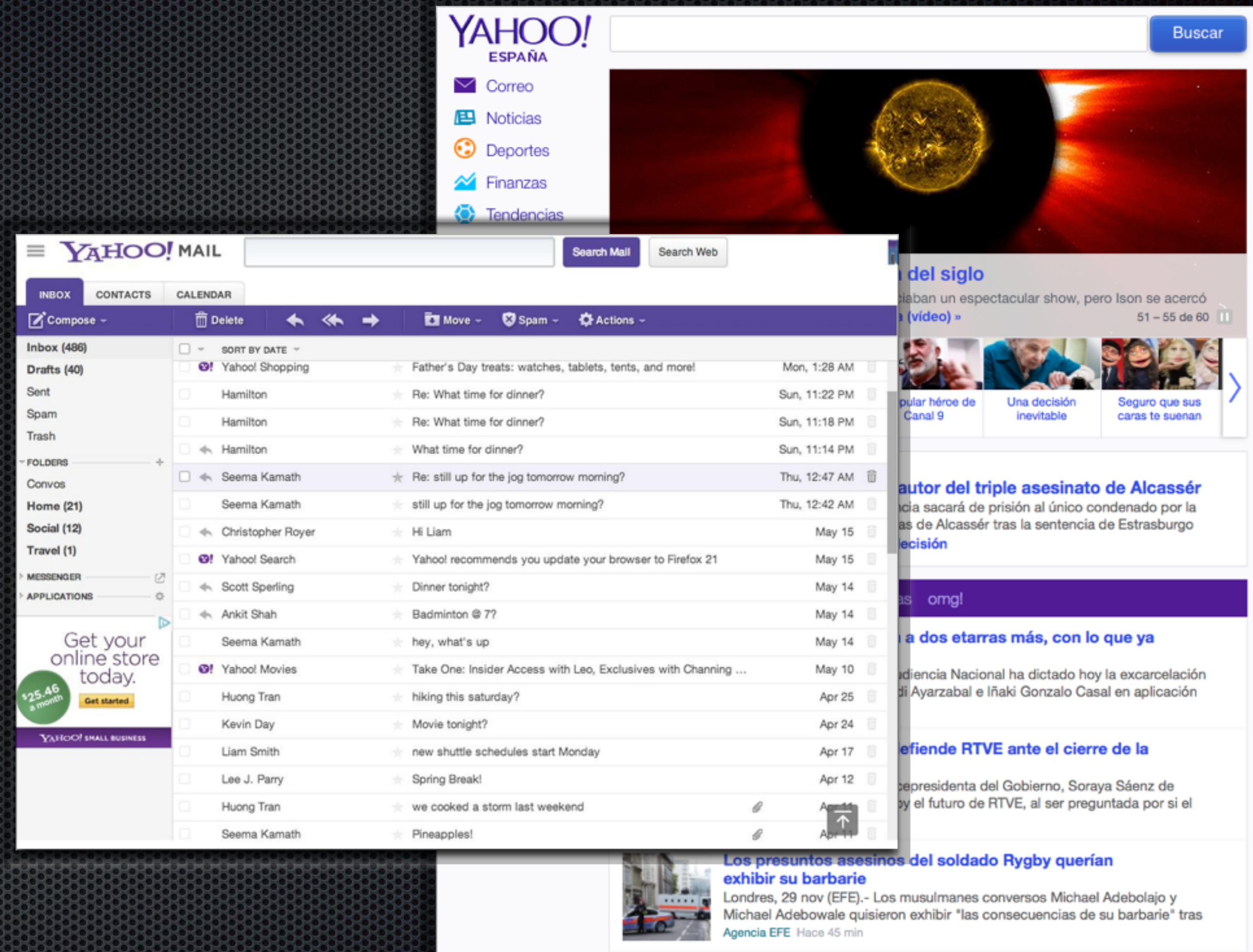
Personalization



Applications

Personalization

Spam detection

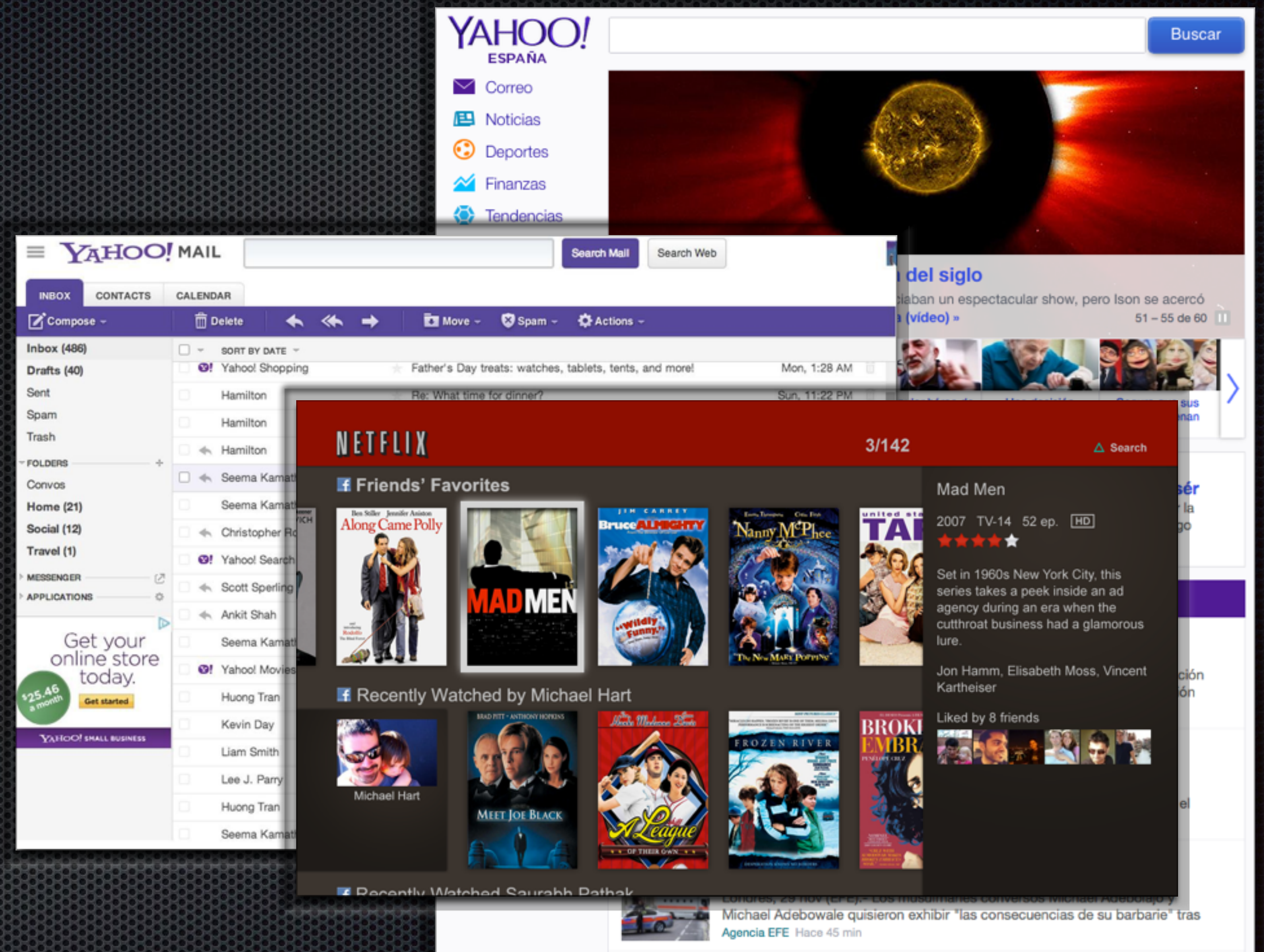


Applications

Personalization

Spam detection

Recommendation



Big Data Stream

- ✦ Volume + Velocity (+ Variety)
- ✦ Too large for single commodity server main memory
- ✦ Too fast for single commodity server CPU
- ✦ A solution should be:
 - ✦ Distributed
 - ✦ Scalable

Examples

- ✧ User clicks
- ✧ Search queries
- ✧ News
- ✧ Emails
- ✧ Tumblr posts
- ✧ Flickr photos
- ✧ Finance stocks
- ✧ Credit card transactions
- ✧ Wikipedia edit logs
- ✧ Facebook statuses
- ✧ Twitter updates
- ✧ Name your own...

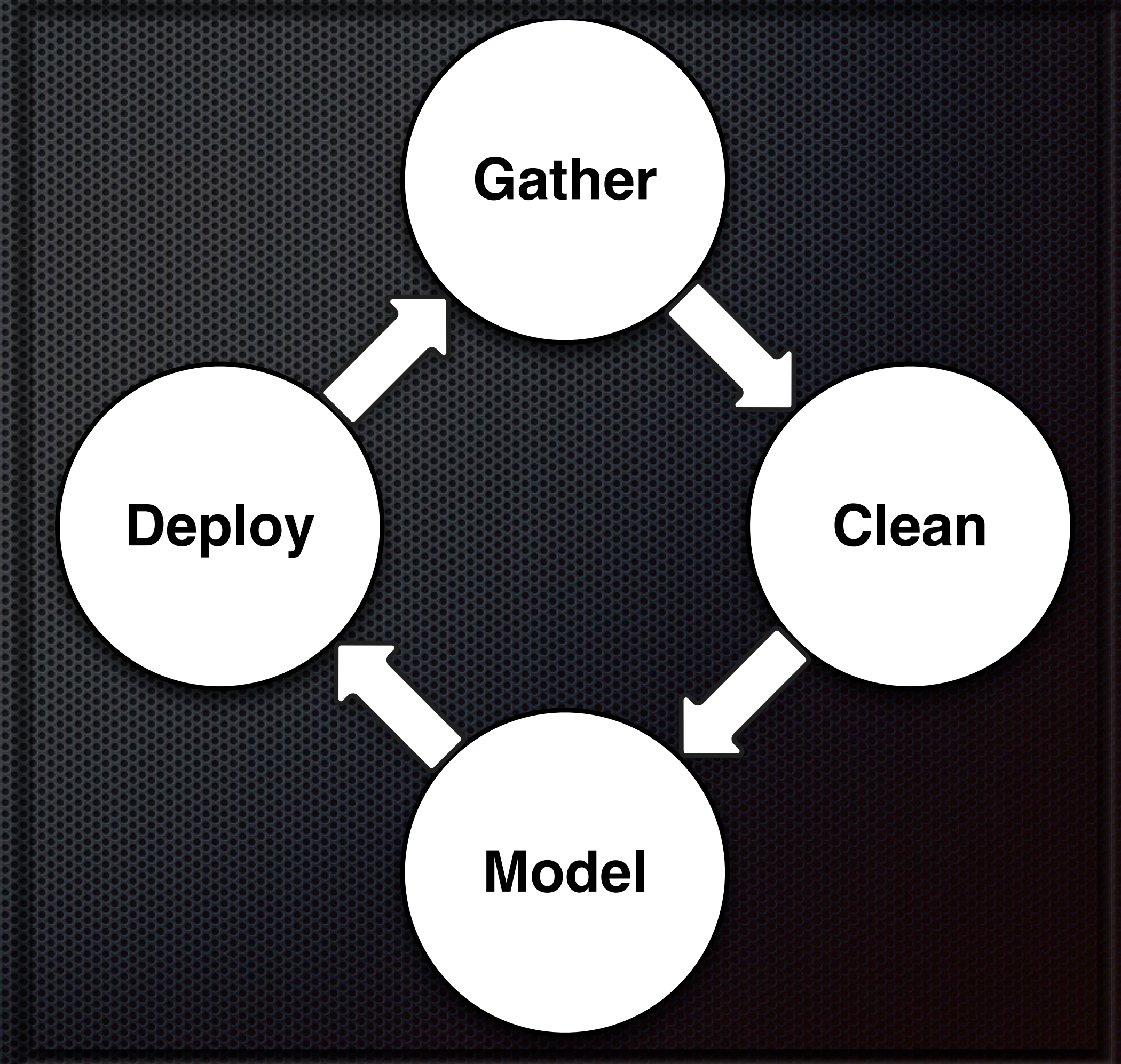
Stream

Batch data is
a snapshot of
streaming data

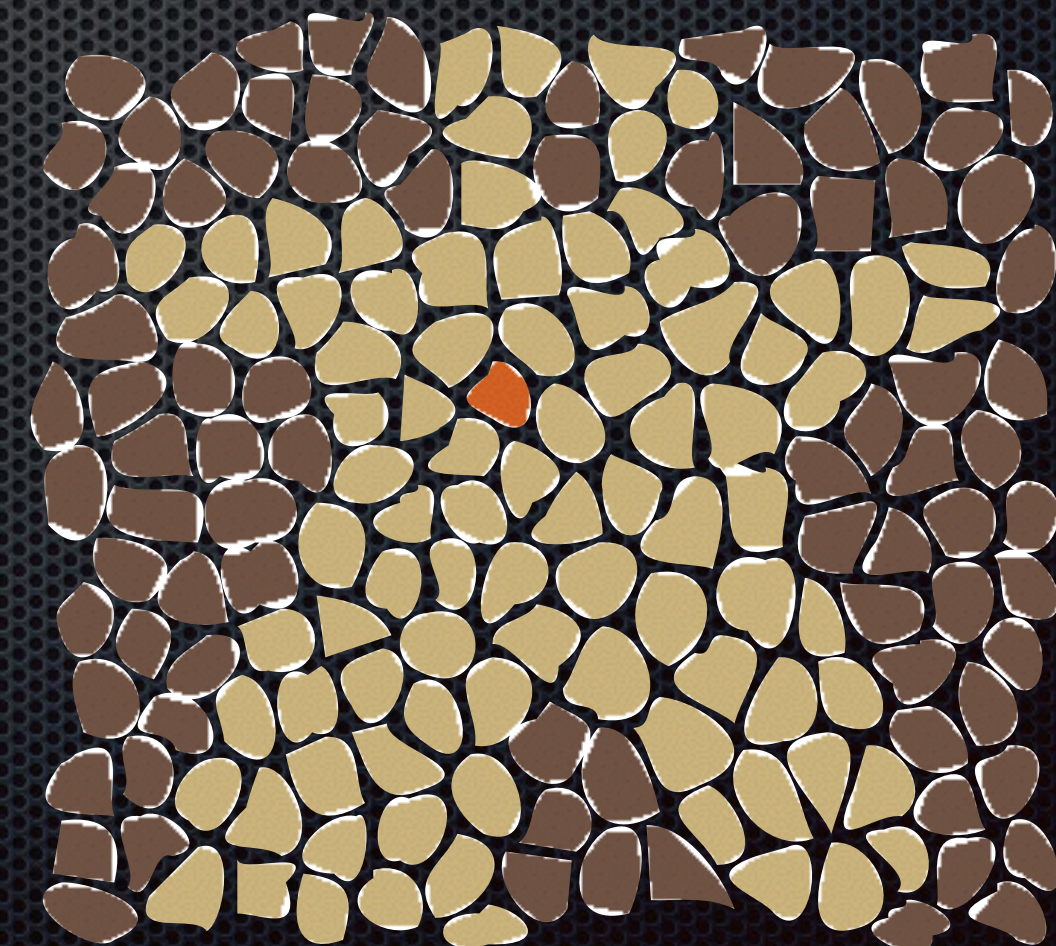
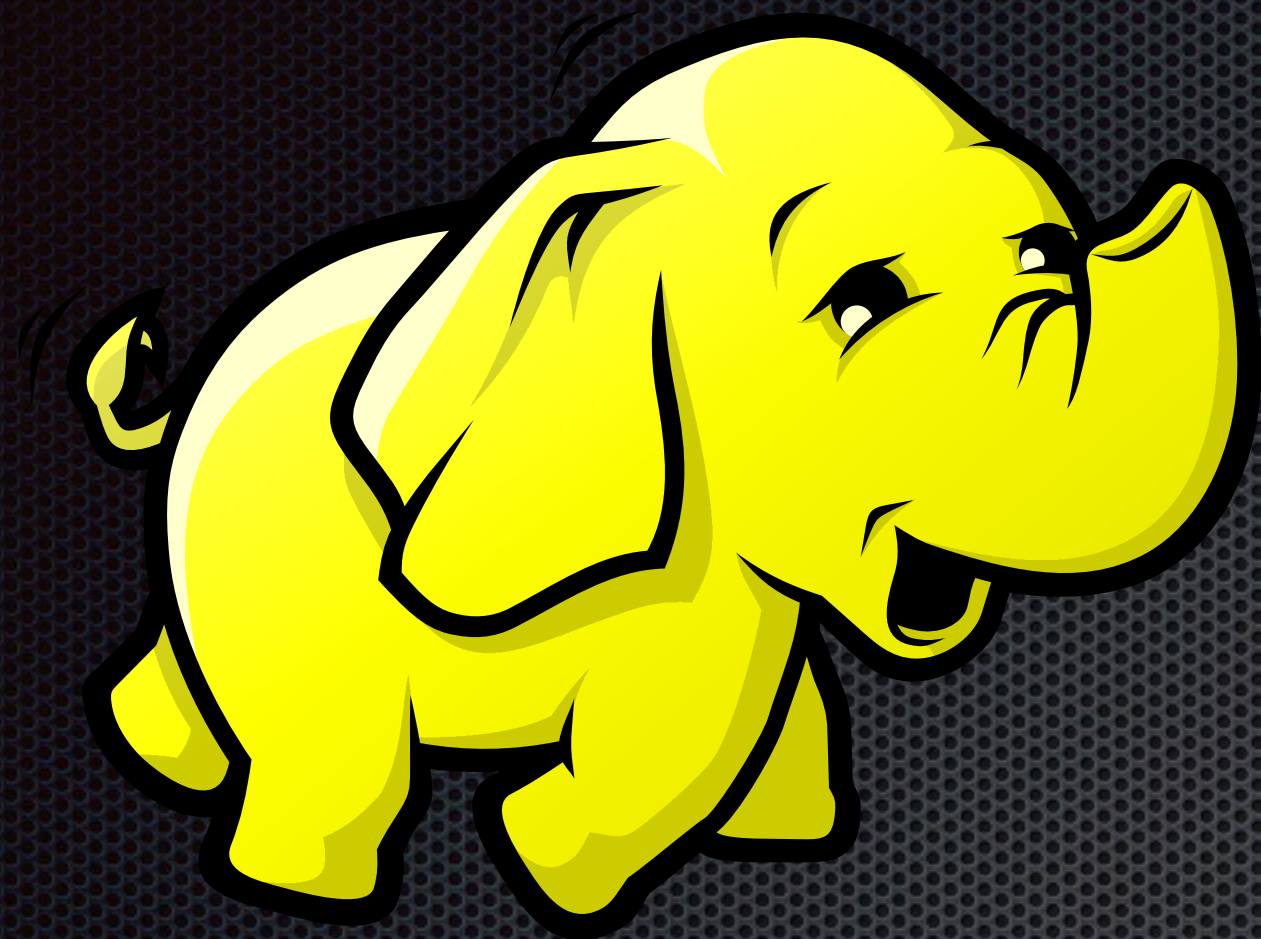


Data Science Lifecycle

- ✦ Old school's data mining
- ✦ From data to insight
- ✦ From insight to model
- ✦ From model to value
- ✦ And repeat!



Big Data Tools



Problems

- ✦ Operational
 - ✦ Need to rerun the pipeline and redeploy the model when new data arrives
- ✦ Paradigmatic
 - ✦ New data lies in storage without generating new value until the new model is retrained



Present of big data

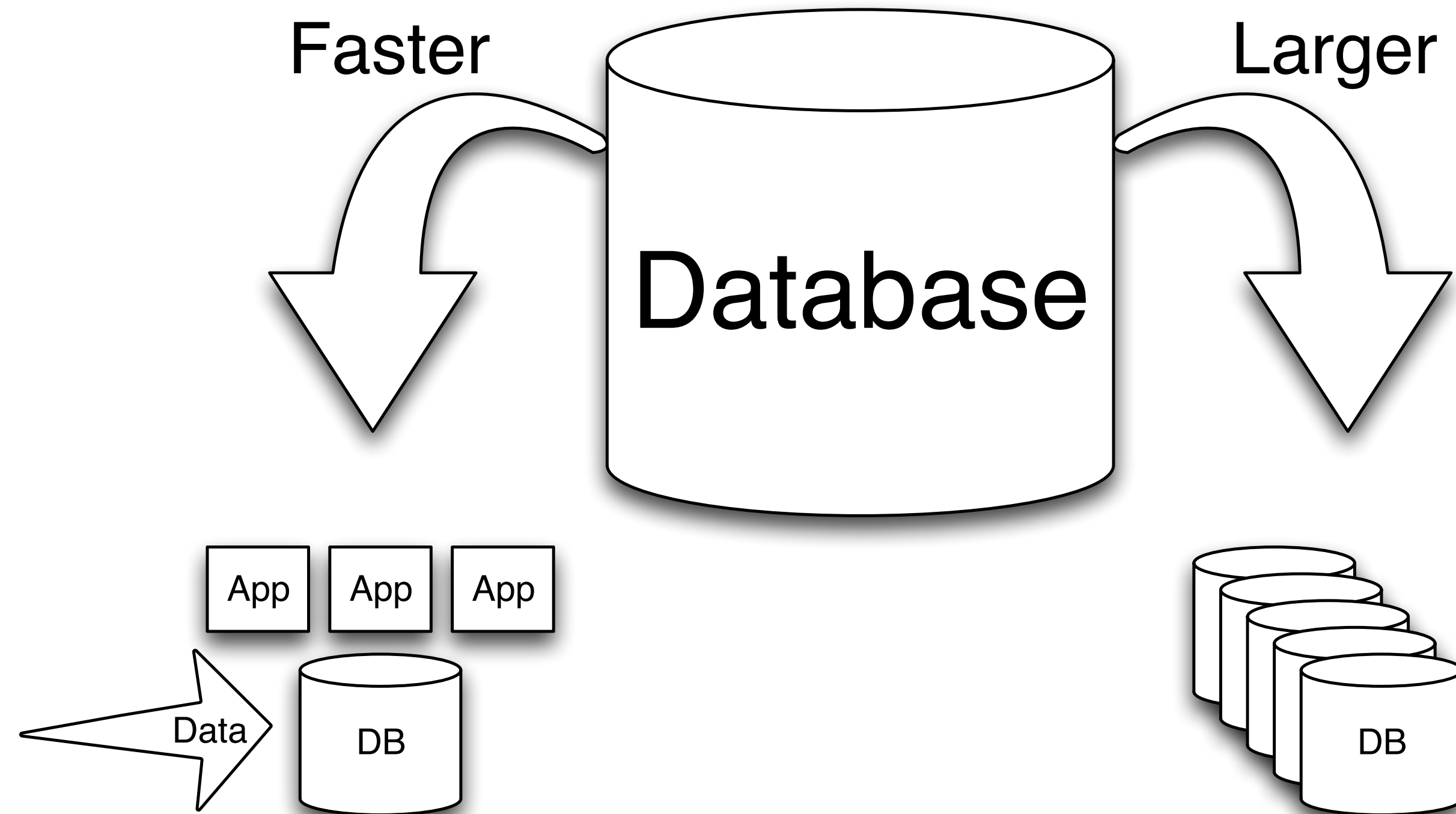
Too big to handle



Future of big data

Drinking from a firehose

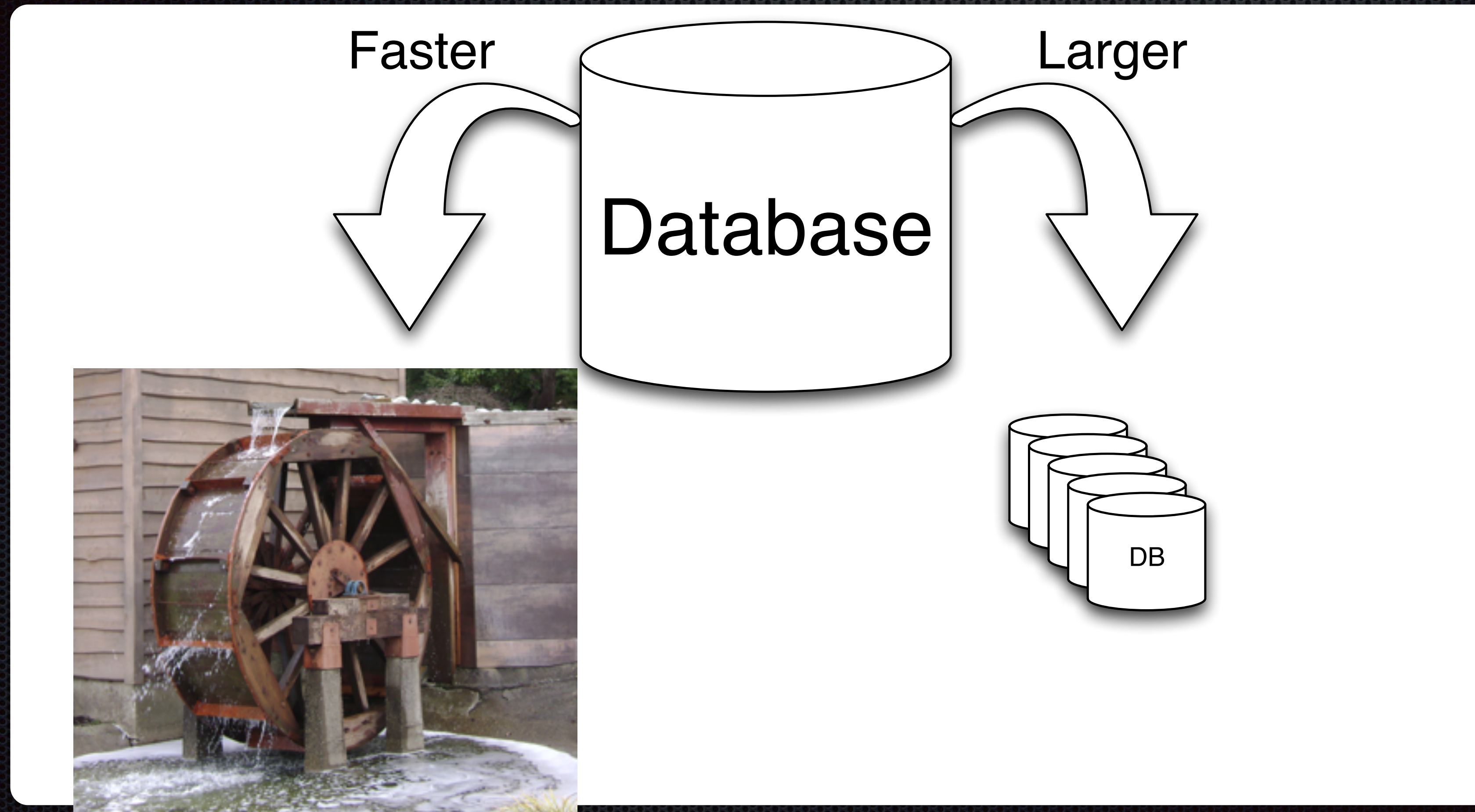
A Tale of Two Tribes



M. Stonebraker and U. Çetintemel, “‘One Size Fits All’: An Idea Whose Time Has Come and Gone,” in *ICDE '05*

A. Jacobs, “The Pathologies of Big Data,” *Communications of the ACM*, 52(8):36–44, Aug. 2009

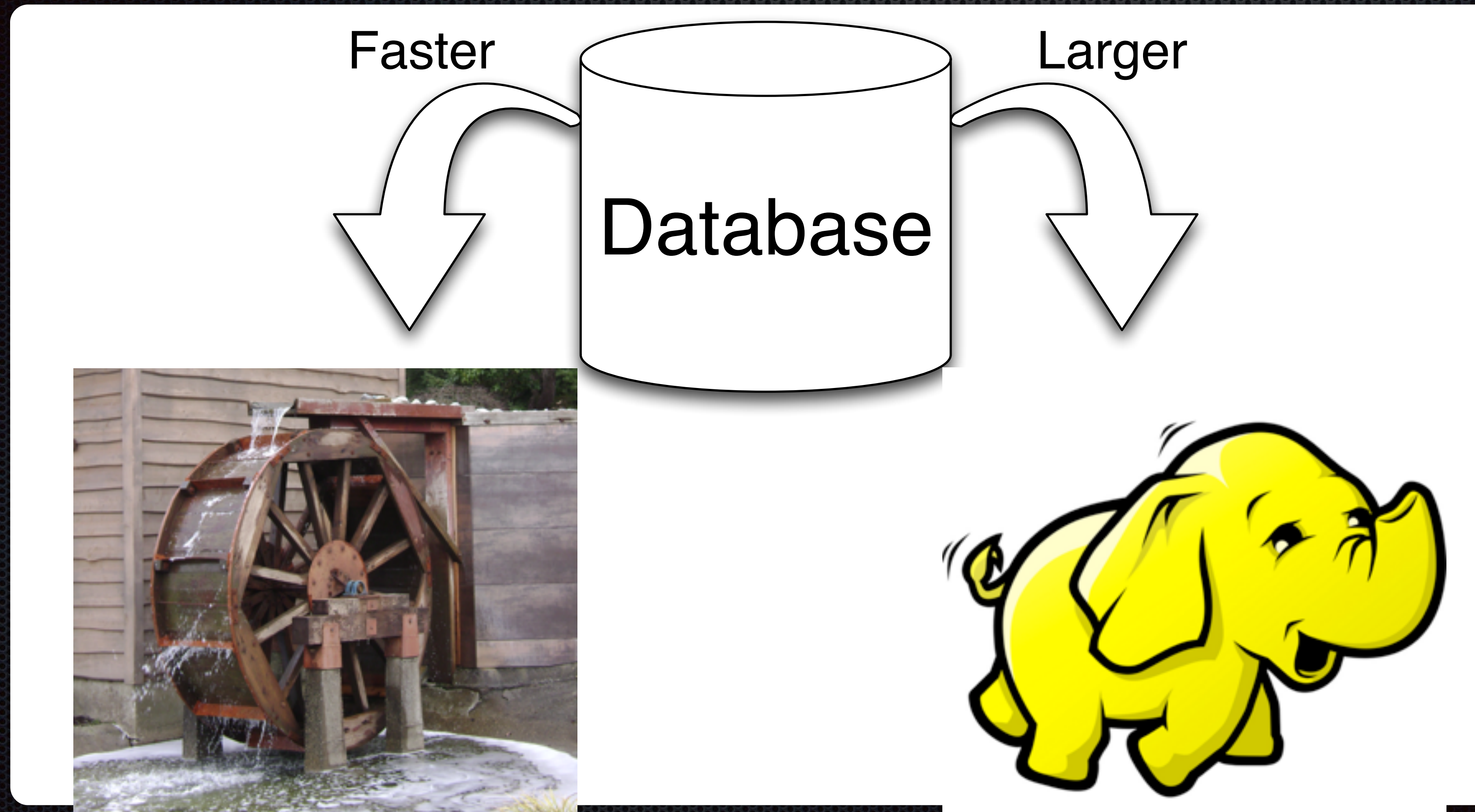
A Tale of Two Tribes



M. Stonebraker and U. Çetintemel, “‘One Size Fits All’: An Idea Whose Time Has Come and Gone,” in *ICDE '05*

A. Jacobs, “The Pathologies of Big Data,” *Communications of the ACM*, 52(8):36–44, Aug. 2009

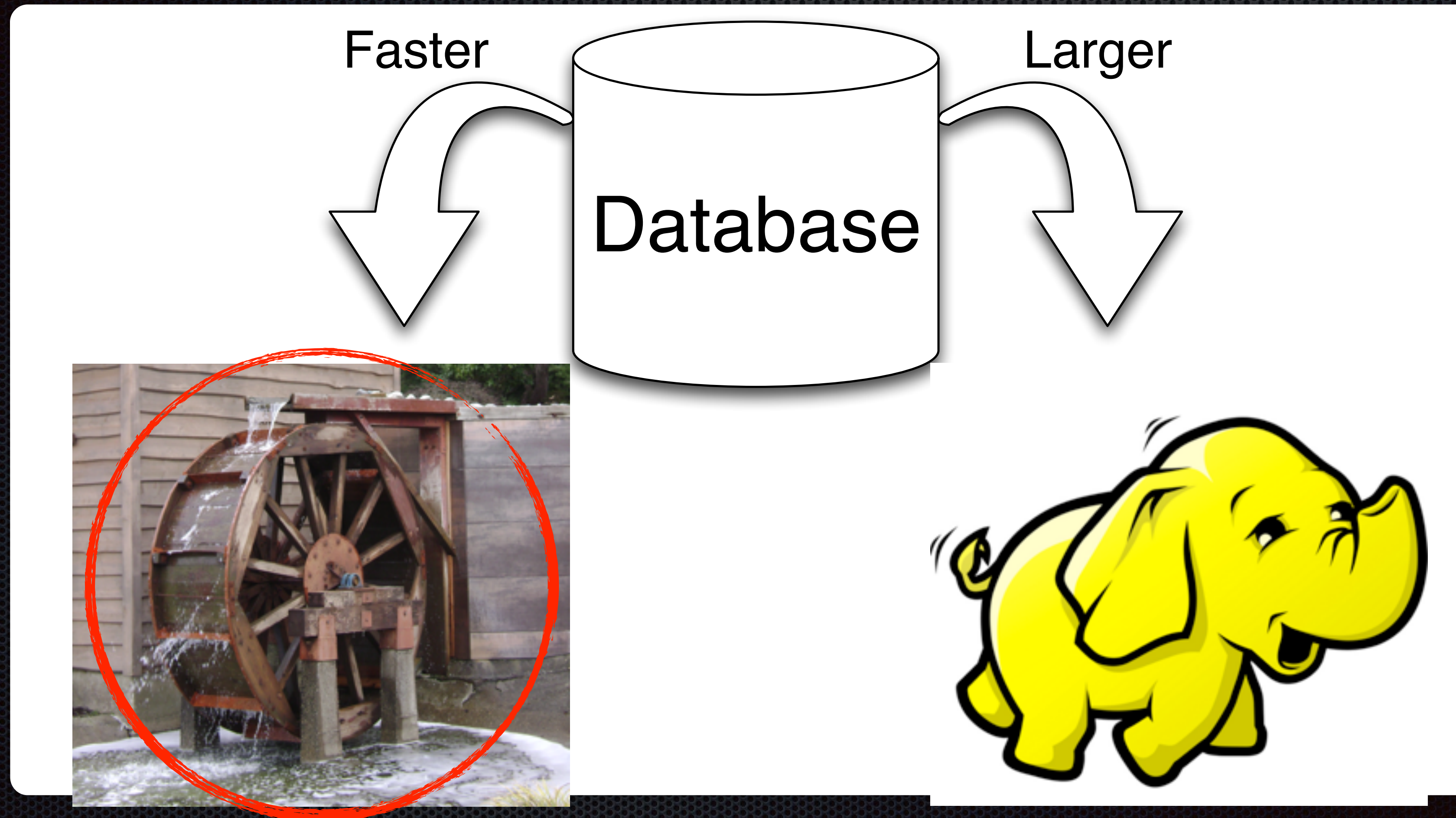
A Tale of Two Tribes



M. Stonebraker and U. Çetintemel, “‘One Size Fits All’: An Idea Whose Time Has Come and Gone,” in *ICDE '05*

A. Jacobs, “The Pathologies of Big Data,” *Communications of the ACM*, 52(8):36–44, Aug. 2009

A Tale of Two Tribes



M. Stonebraker and U. Çetintemel, “‘One Size Fits All’: An Idea Whose Time Has Come and Gone,” in *ICDE '05*

A. Jacobs, “The Pathologies of Big Data,” *Communications of the ACM*, 52(8):36–44, Aug. 2009

Evolution of SPEs

1st generation

2003

Aurora

Abadi et al., “Aurora: a new model and architecture for data stream management,” VLDB Journal, 2003

2004

STREAM

Arasu et al., “STREAM: The Stanford Data Stream Management System,” Stanford InfoLab, 2004.

2nd generation

2005

Borealis

Abadi et al., “The Design of the Borealis Stream Processing Engine,” in CIDR '05

2006

SPC

Amini et al., “SPC: A Distributed, Scalable Platform for Data Mining,” in DMSSP '06

2008

SPADE

Gedik et al., “SPADE: The System S Declarative Stream Processing Engine,” in SIGMOD '08

3rd generation

2010

S4

Neumeyer et al., “S4: Distributed Stream Computing Platform,” in ICDMW '10

2011

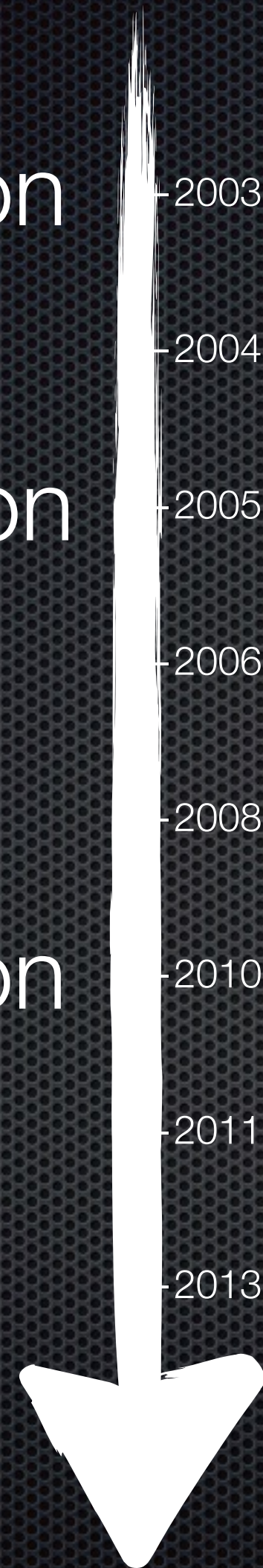
Storm

<http://storm.apache.org>

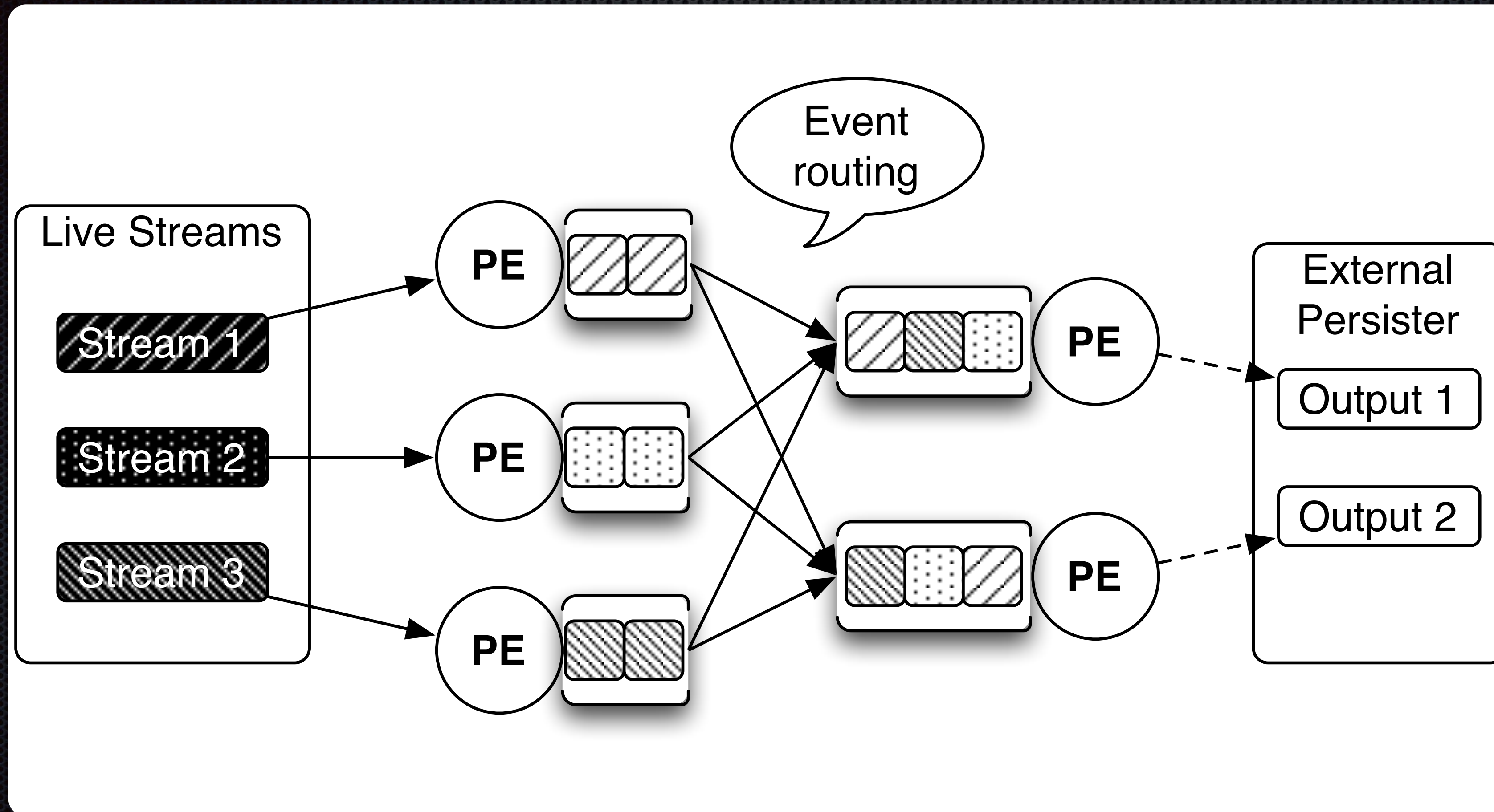
2013

Samza

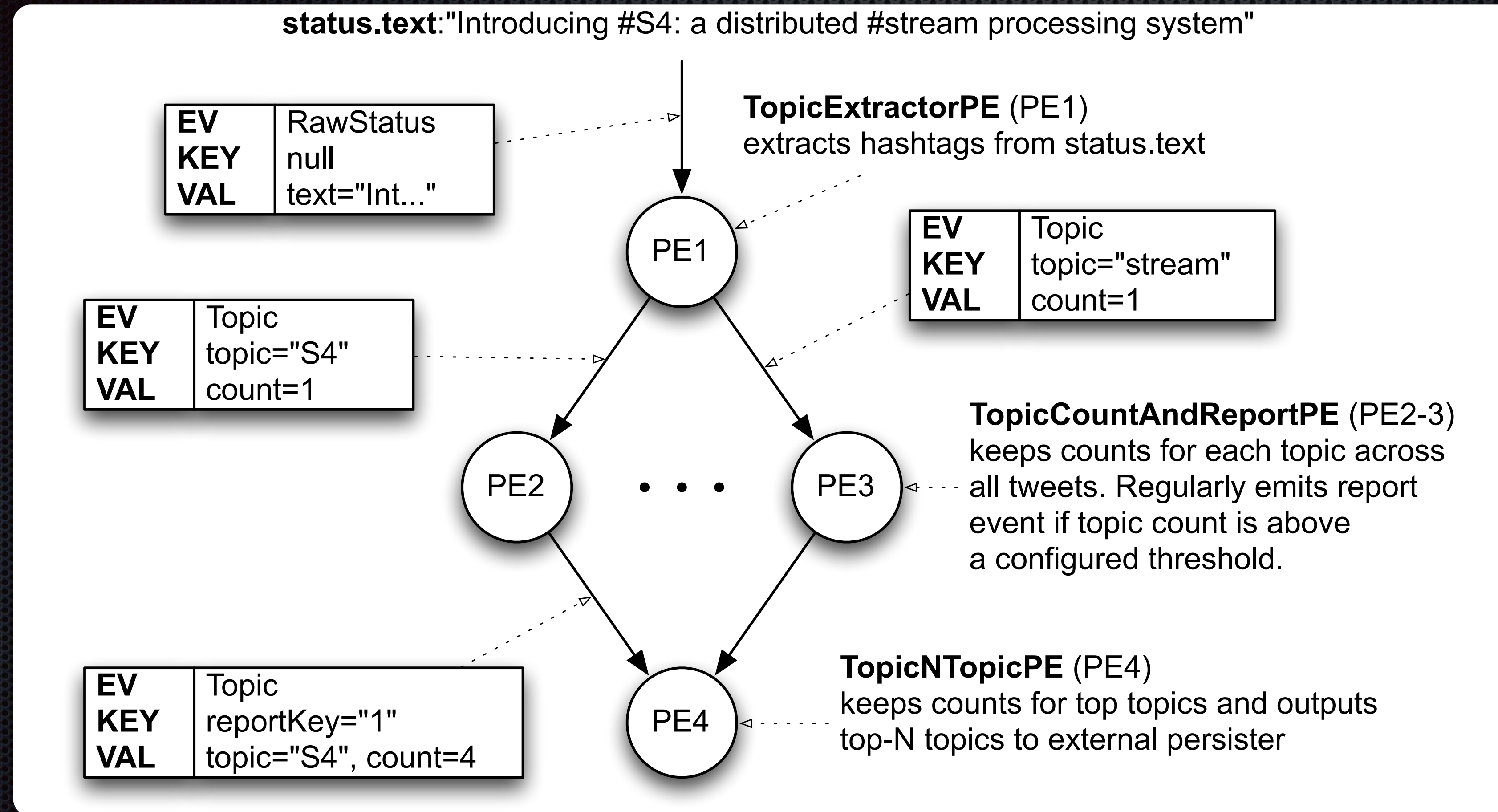
<http://samza.incubator.apache.org>



Actors Model



S4 Example



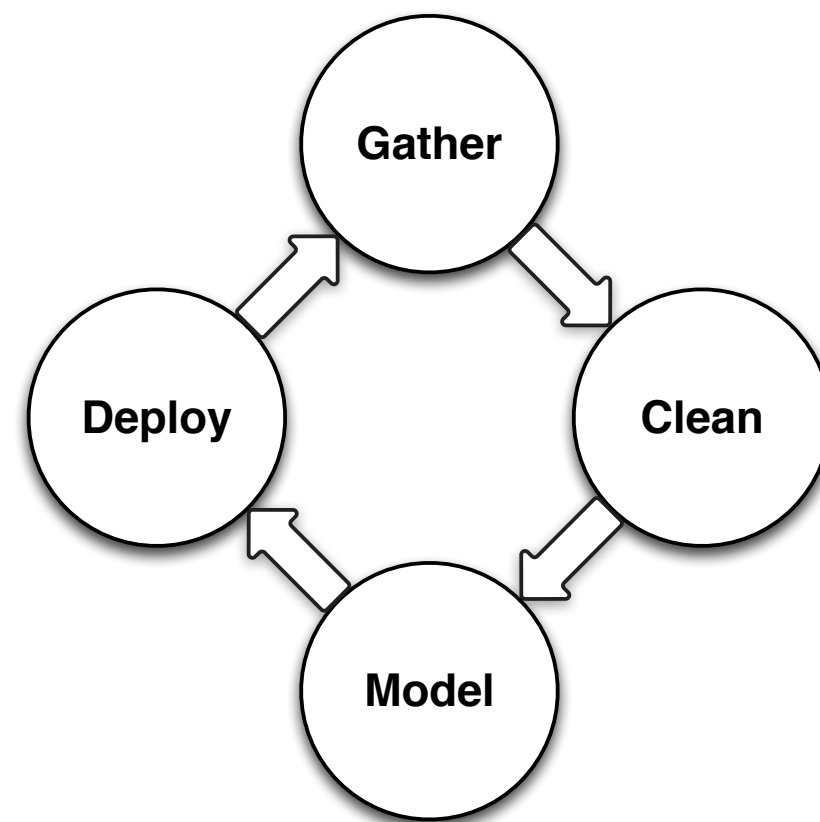
But we have Hadoop!

- ✦ “Mapreduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That’s Not a Nail!”
[J. Lin, in Big Data, 1(1):28–37, 2013]
- ✦ “Data whose characteristics forces us to look beyond the traditional methods that are prevalent at the time”
[A. Jacobs, in ACM Queue, 7(6):10,2009]

Paradigm Shift



+



=



Streaming Model

- ✦ Sequence is potentially infinite
- ✦ High amount of data, high speed of arrival
- ✦ Change over time (concept drift)
- ✦ Approximation algorithms
(small error with high probability)
 - ✦ Single pass, one data item at a time
 - ✦ Sub-linear space and time per data item

SAMOA

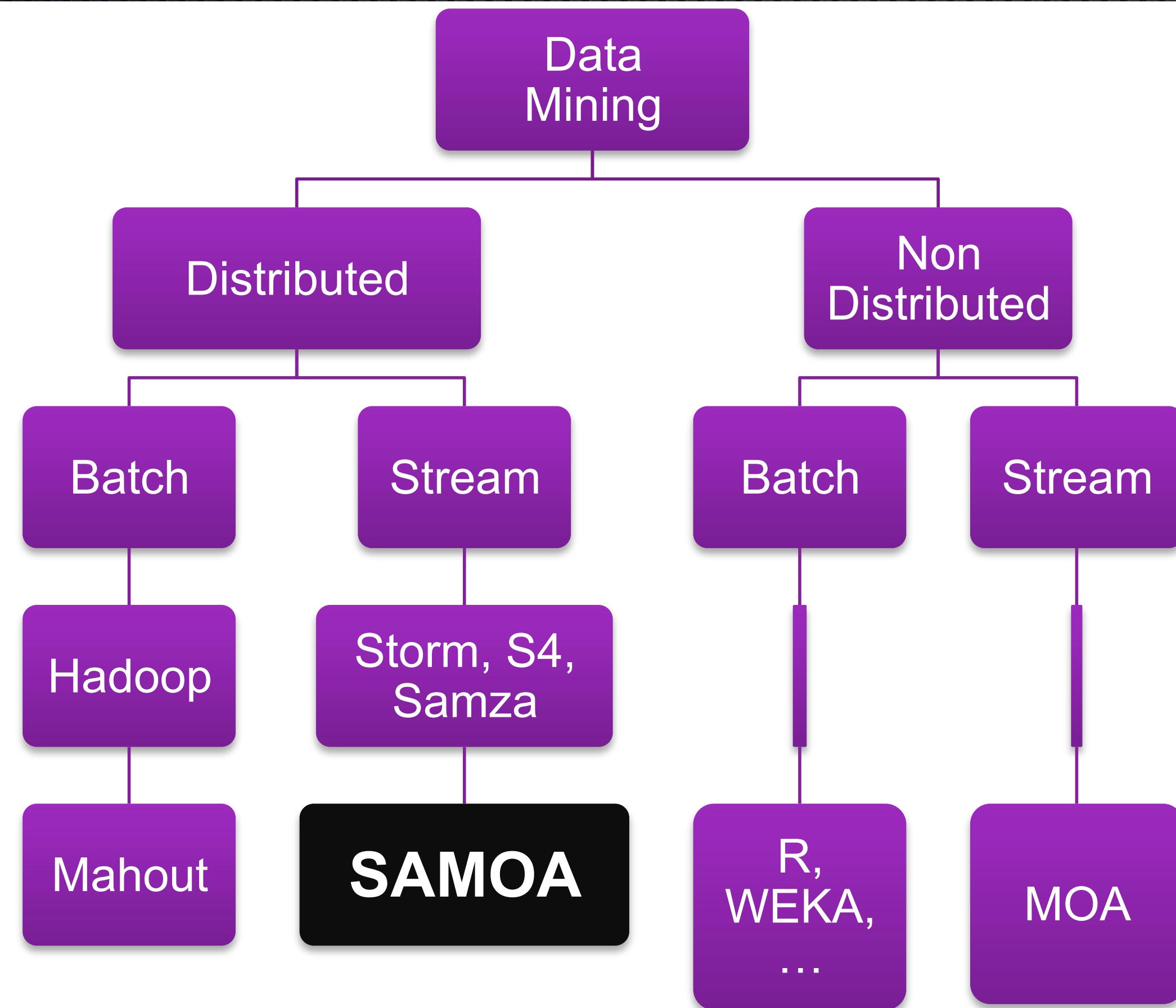
Scalable Advanced Massive Online Analysis

G. De Francisci Morales, A. Bifet
Journal of Machine Learning Research, 2014

Concept

- ✧ SAMOA is a platform
- ✧ Researchers
 - ✧ Framework for developing distributed stream mining algorithms
- ✧ Practitioners
 - ✧ Library of state-of-the-art distributed stream mining algorithms

Taxonomy



What about Mahout?

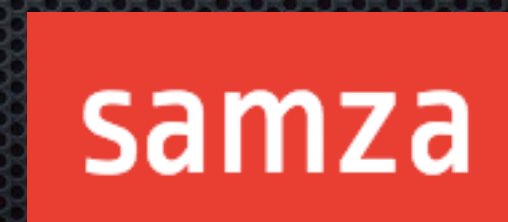


- ✦ Think SAMOA = Mahout for streaming
- ✦ But SAMOA...
 - ✦ More than JBoA (just a bunch of algorithms)
 - ✦ Provides a common platform
 - ✦ Easy to port to new computing engines

Status

<https://github.com/yahoo/samoa>

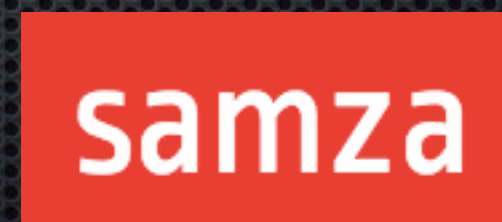
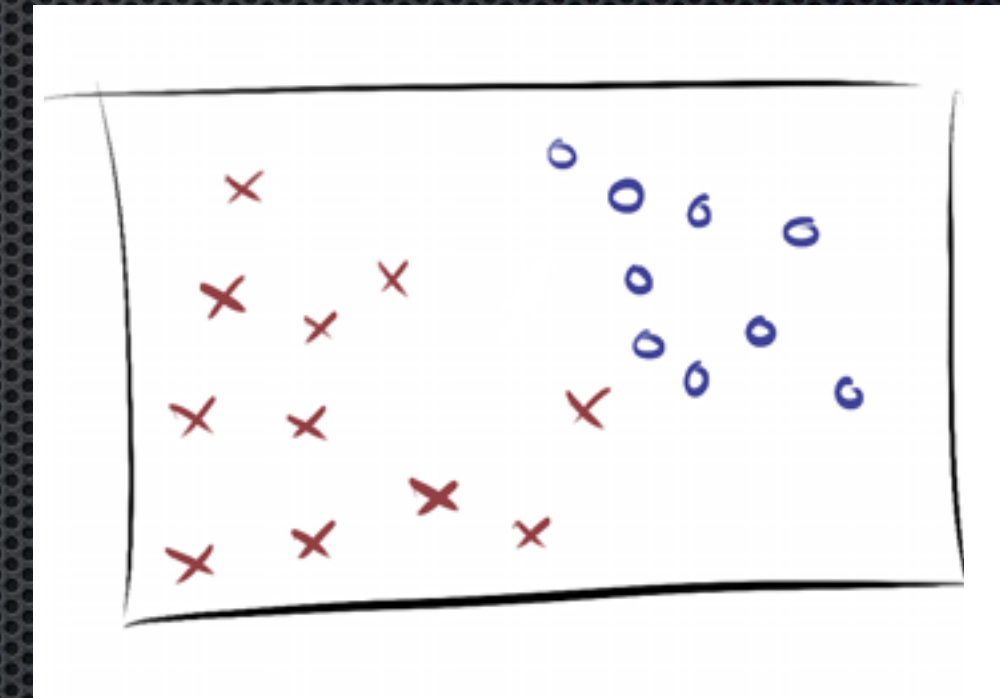
- ✦ Parallel algorithms
 - ✦ Vertical Hoeffding Tree (classification)
 - ✦ CluStream (clustering)
 - ✦ Adaptive Model Rules (regression)
 - ✦ PARMA (frequent pattern mining) [pending]
- ✦ Execution engines



Status

<https://github.com/yahoo/samoa>

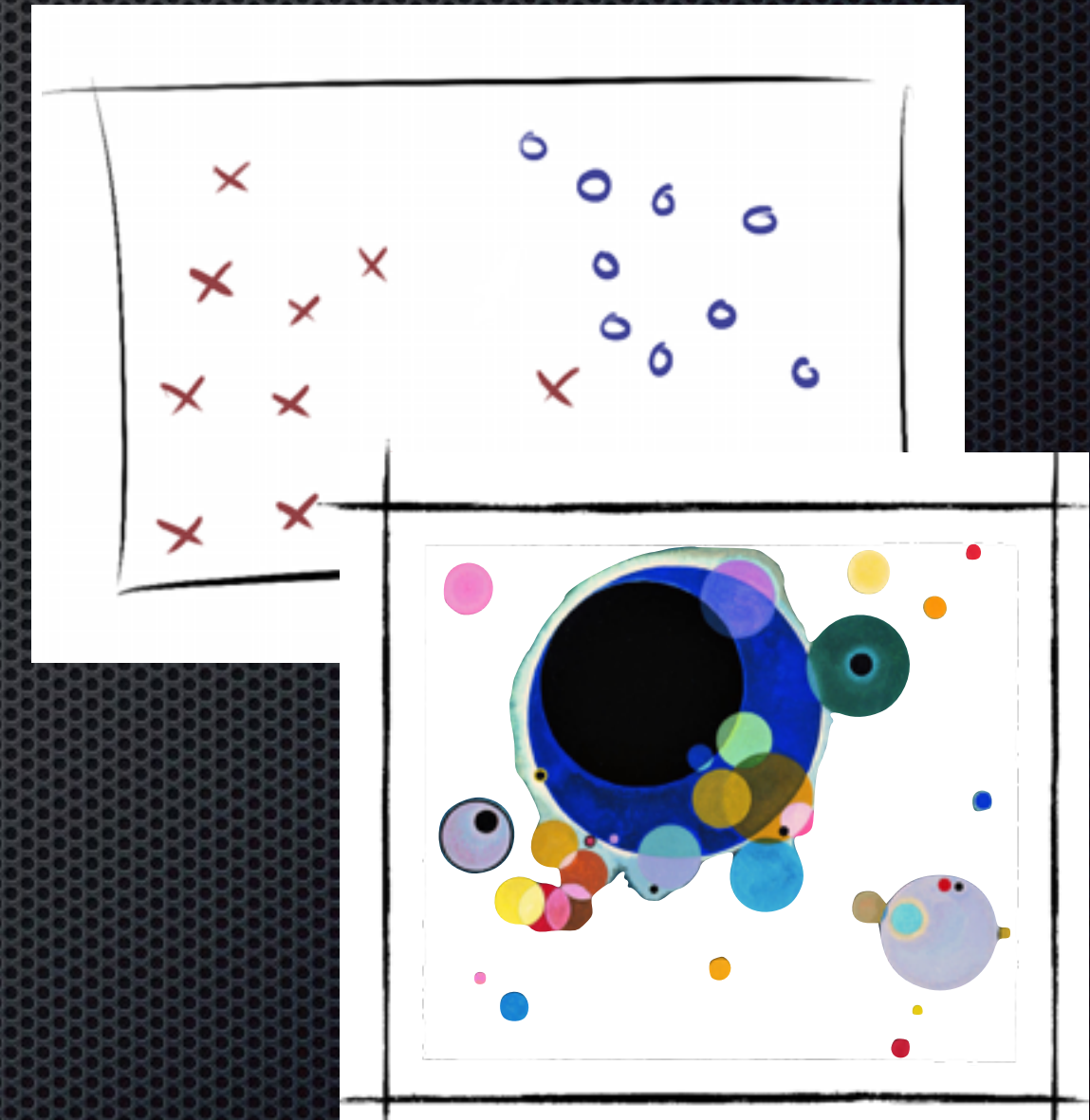
- ✧ Parallel algorithms
 - ✧ Vertical Hoeffding Tree (classification)
 - ✧ CluStream (clustering)
 - ✧ Adaptive Model Rules (regression)
 - ✧ PARMA (frequent pattern mining) [pending]
- ✧ Execution engines



Status

<https://github.com/yahoo/samoa>

- ✦ Parallel algorithms
 - ✦ Vertical Hoeffding Tree (classification)
 - ✦ CluStream (clustering)
 - ✦ Adaptive Model Rules (regression)
 - ✦ PARMA (frequent pattern mining) [pending]
- ✦ Execution engines



samza

Status

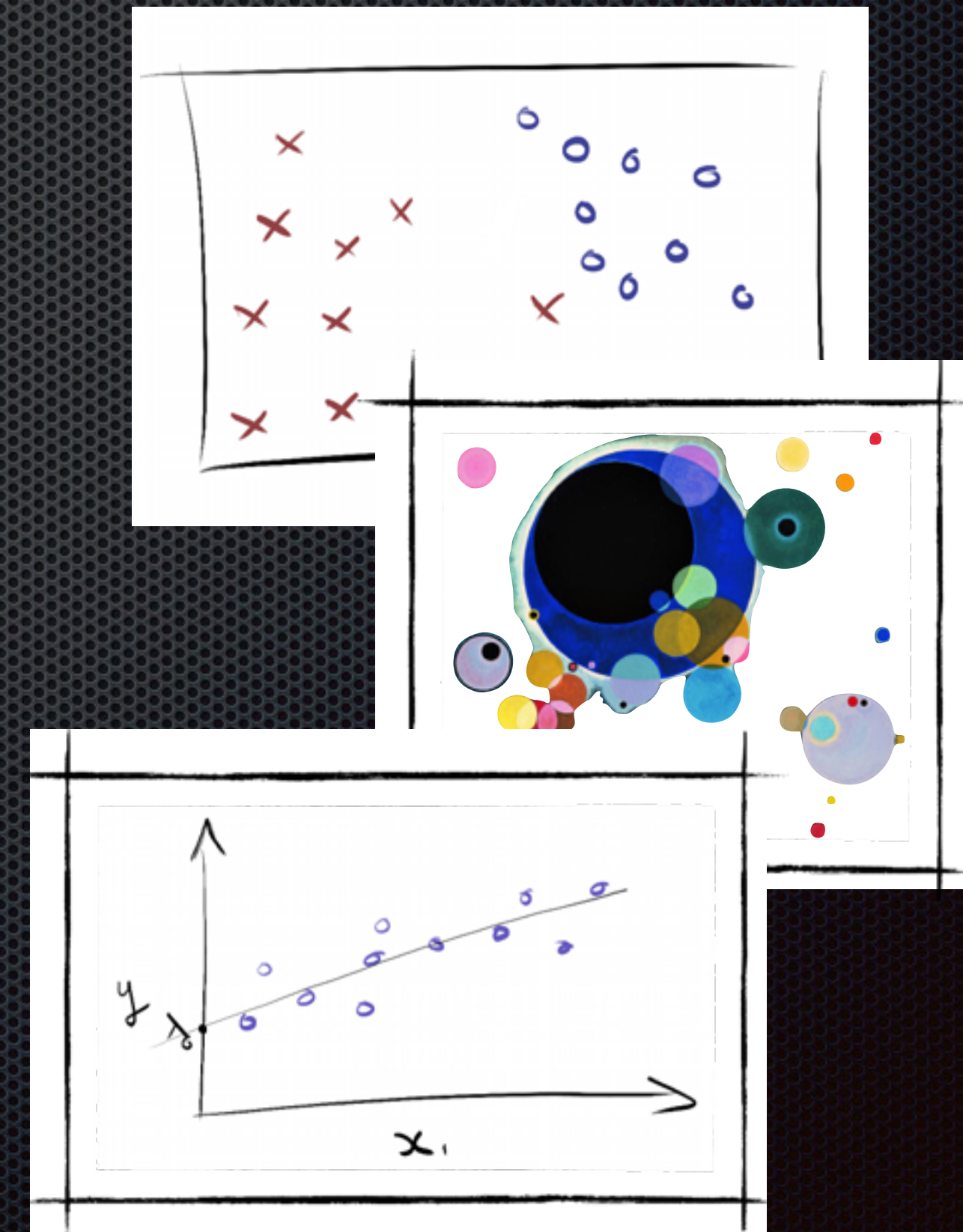
<https://github.com/yahoo/samoa>

- ✦ Parallel algorithms
 - ✦ Vertical Hoeffding Tree (classification)
 - ✦ CluStream (clustering)
 - ✦ Adaptive Model Rules (regression)
 - ✦ PARMA (frequent pattern mining) [pending]
- ✦ Execution engines

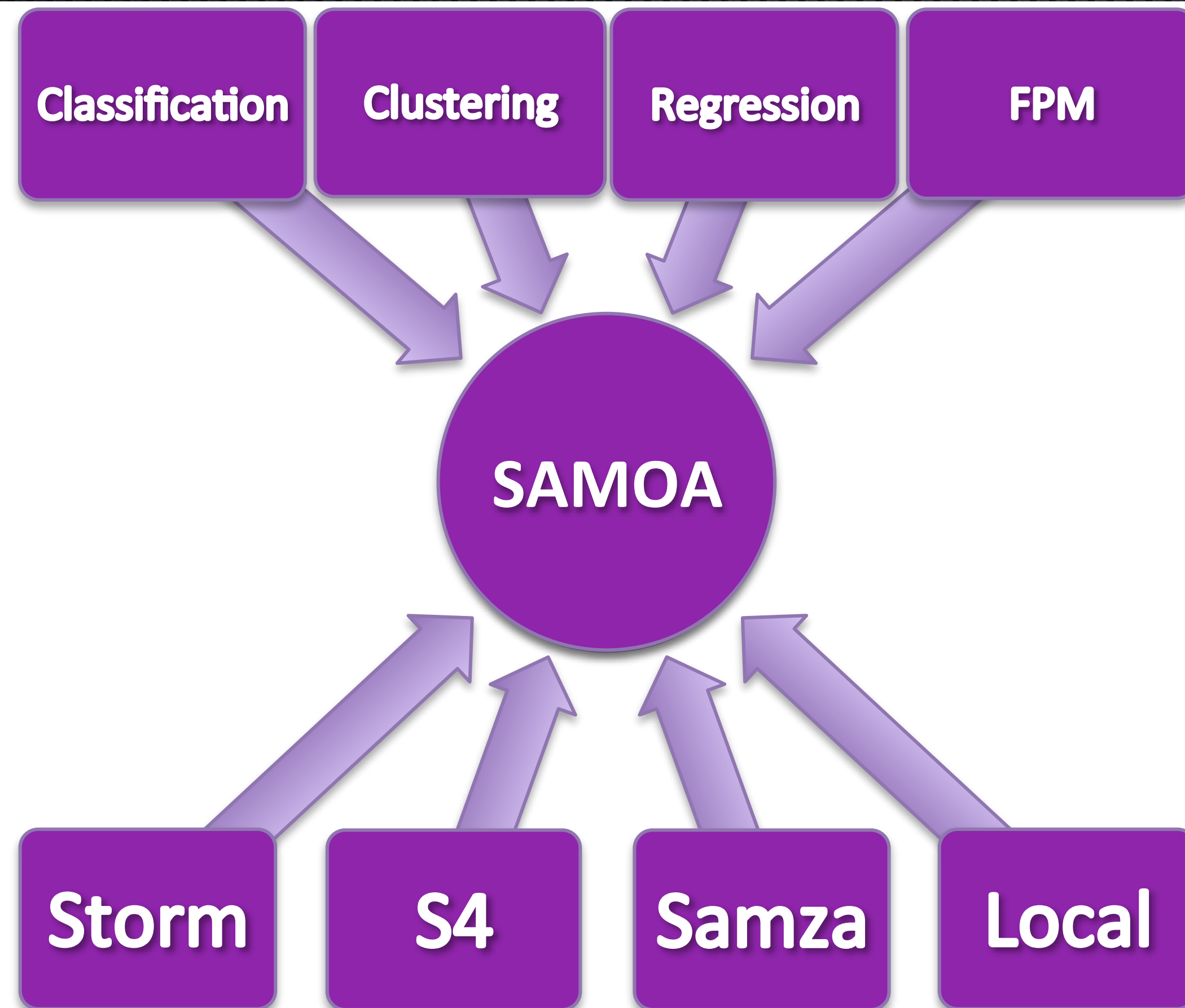


S4 distributed stream
computing platform

samza



Architecture



Is SAMOA useful for you?

- ✦ Only if you need to deal with:
 - ✦ **Big** *fast* data
 - ✦ Evolving data (model updates)
- ✦ What is happening now?
 - ✦ Use feedback in real-time
 - ✦ Adapt to changes faster

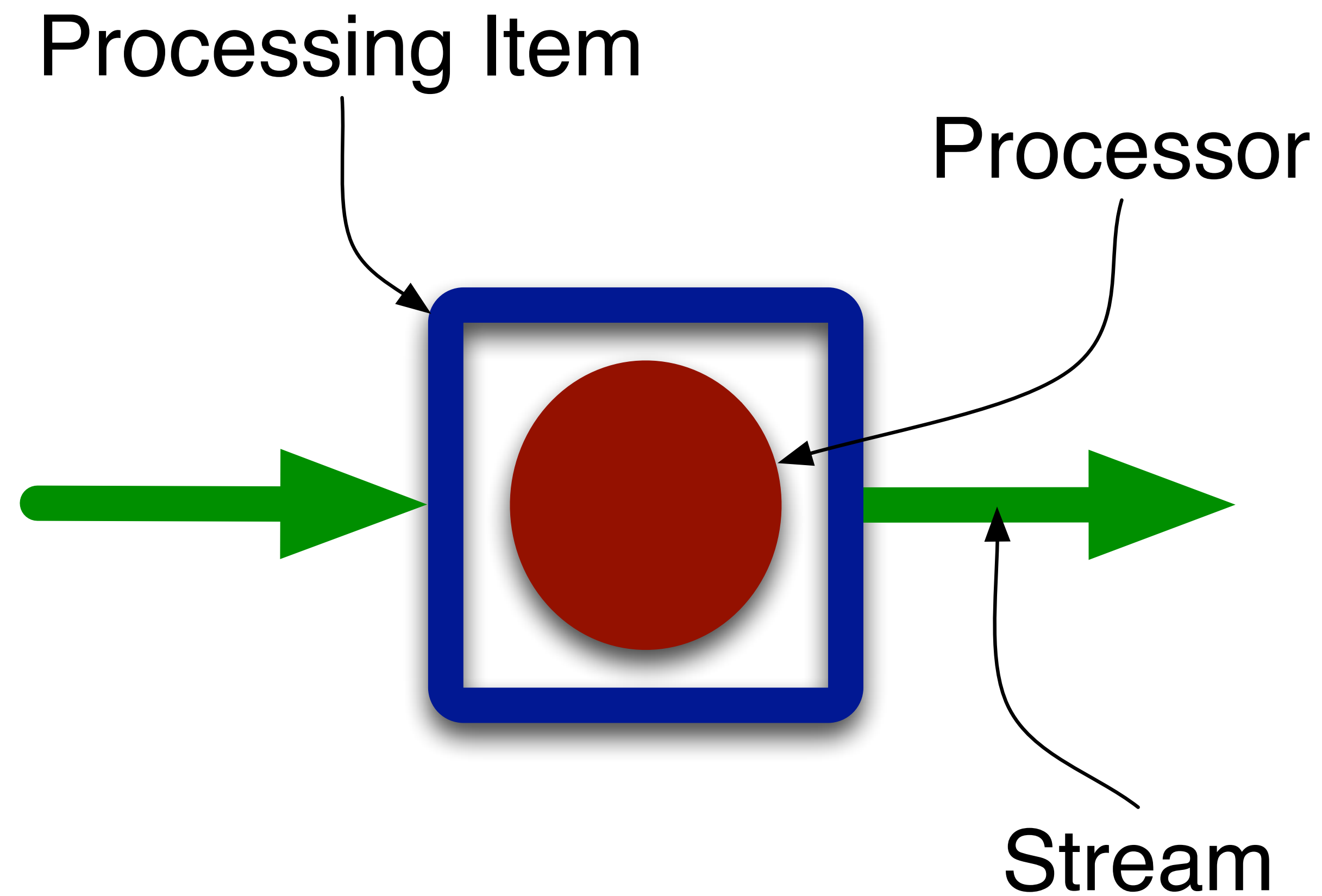
Advantages (operational)

- ✦ Program once, run everywhere
 - ✦ Reuse existing computational infrastructure
- ✦ Avoid deploy cycle
 - ✦ No system downtime
 - ✦ No complex backup/update procedures
 - ✦ No need to choose update frequency

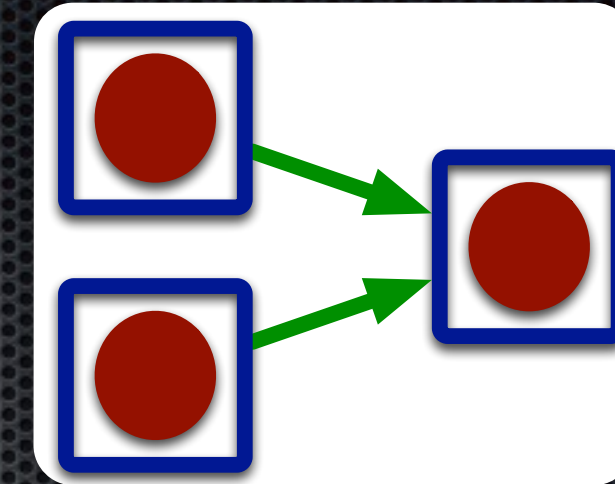
Advantages (paradigmatic)

- ✦ Model freshness
- ✦ No retraining
- ✦ Immediate data value
- ✦ No stream/batch impedance mismatch

ML Developer API

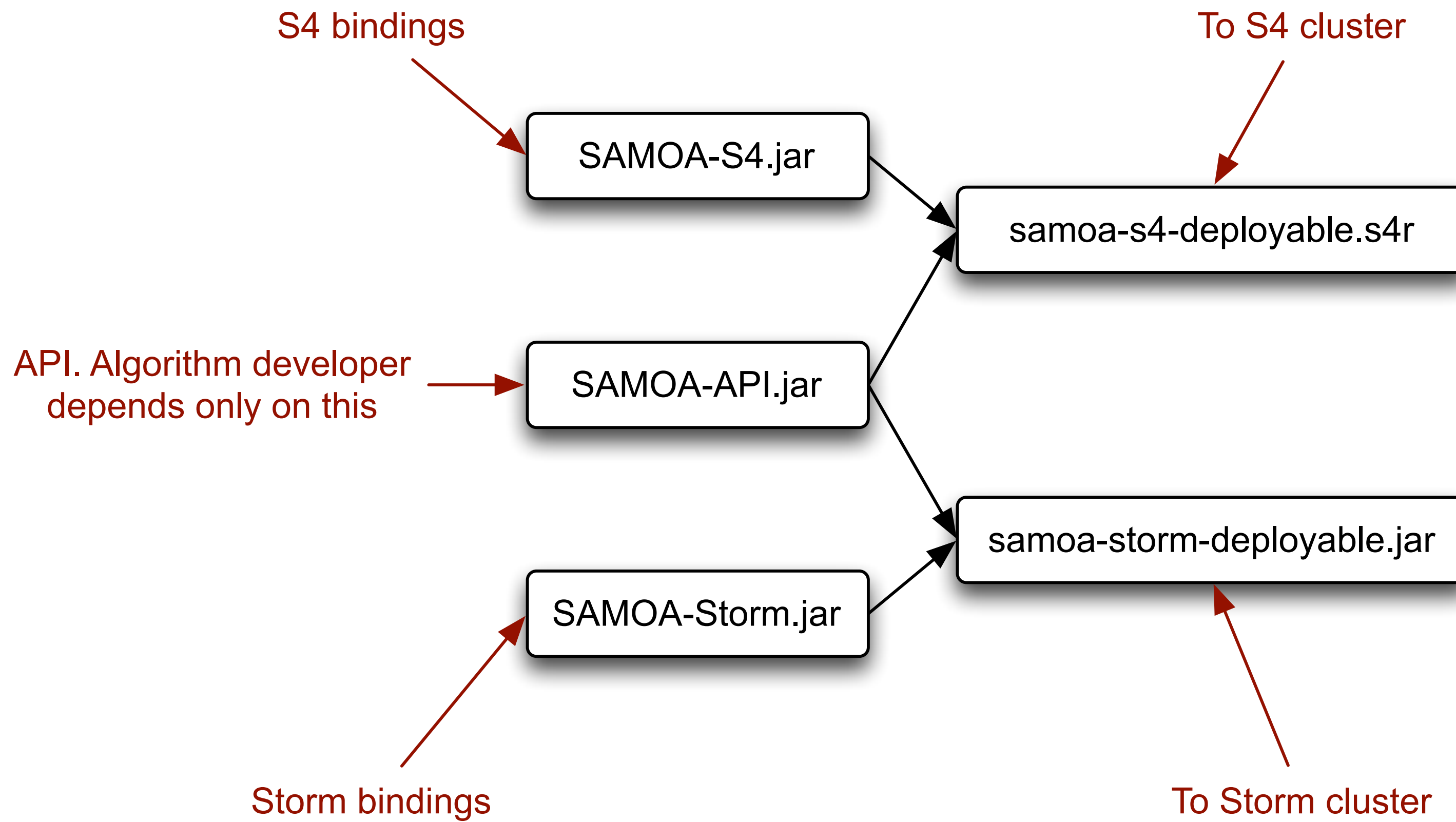


ML Developer API



```
TopologyBuilder builder;  
Processor sourceOne = new SourceProcessor();  
builder.addProcessor(sourceOne);  
Stream streamOne = builder.createStream(sourceOne);  
  
Processor sourceTwo = new SourceProcessor();  
builder.addProcessor(sourceTwo);  
Stream streamTwo = builder.createStream(sourceTwo);  
  
Processor join = new JoinProcessor();  
builder.addProcessor(join)  
    .connectInputShuffle(streamOne)  
    .connectInputKey(streamTwo);
```


Deployment



Conclusions

- ✦ Streaming is the future and is happening now
- ✦ SAMOA
 - ✦ Runs on existing DSPEs (Storm, S4, Samza)
 - ✦ Algorithms for classification, regression, clustering
 - ✦ Available and open-source <http://samoa-project.net>
- ✦ A platform for collaboration and research on distributed stream mining

The Team



**Albert
Bifet**



**Gianmarco
De Francisci Morales**



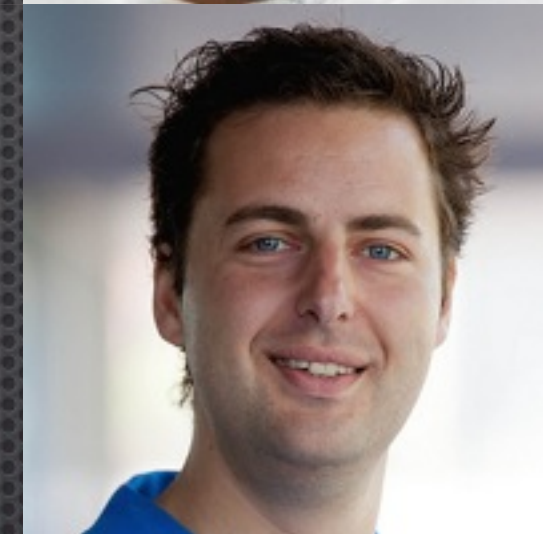
**Nicolas
Kourtellis**



**Matthieu
Morel**



**Arinto
Murdopo**



**Olivier
Van Laere**

Thanks!



@samoa_project

<https://github.com/yahoo/samoa>



@gdfm7

gdfm@apache.org