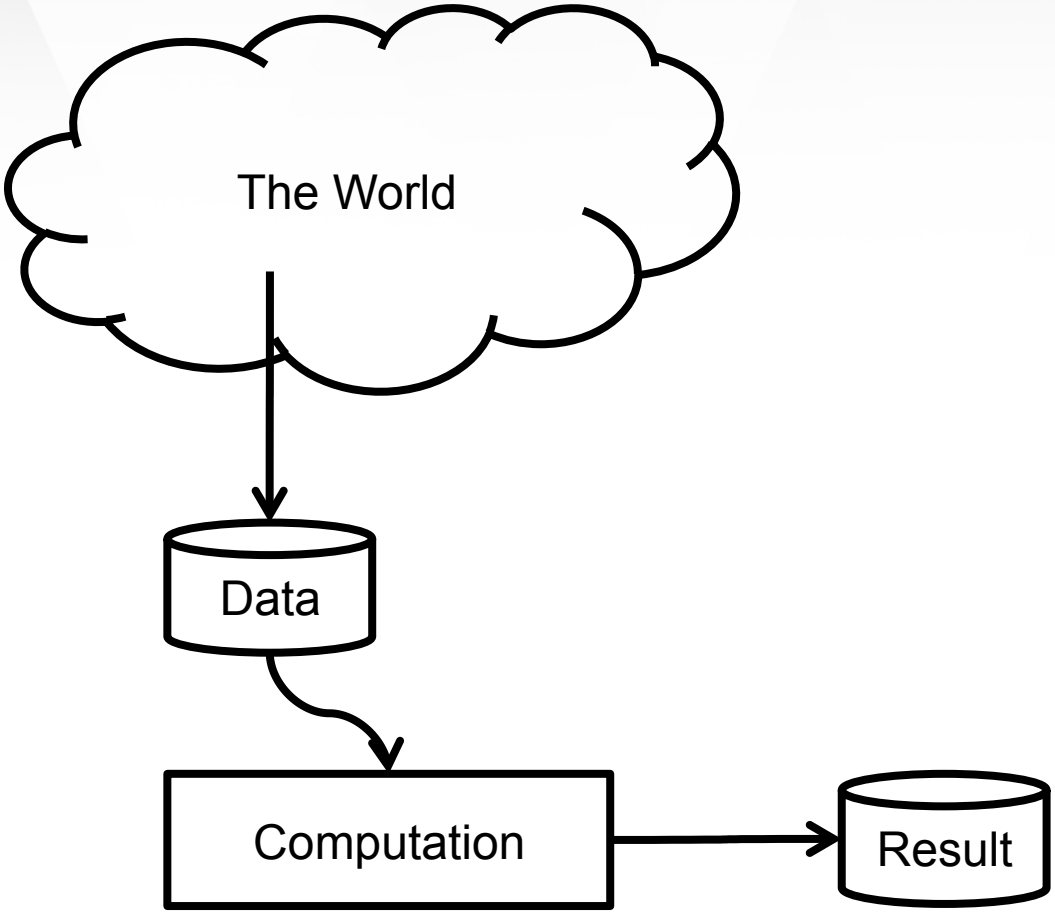
Online Data Mining

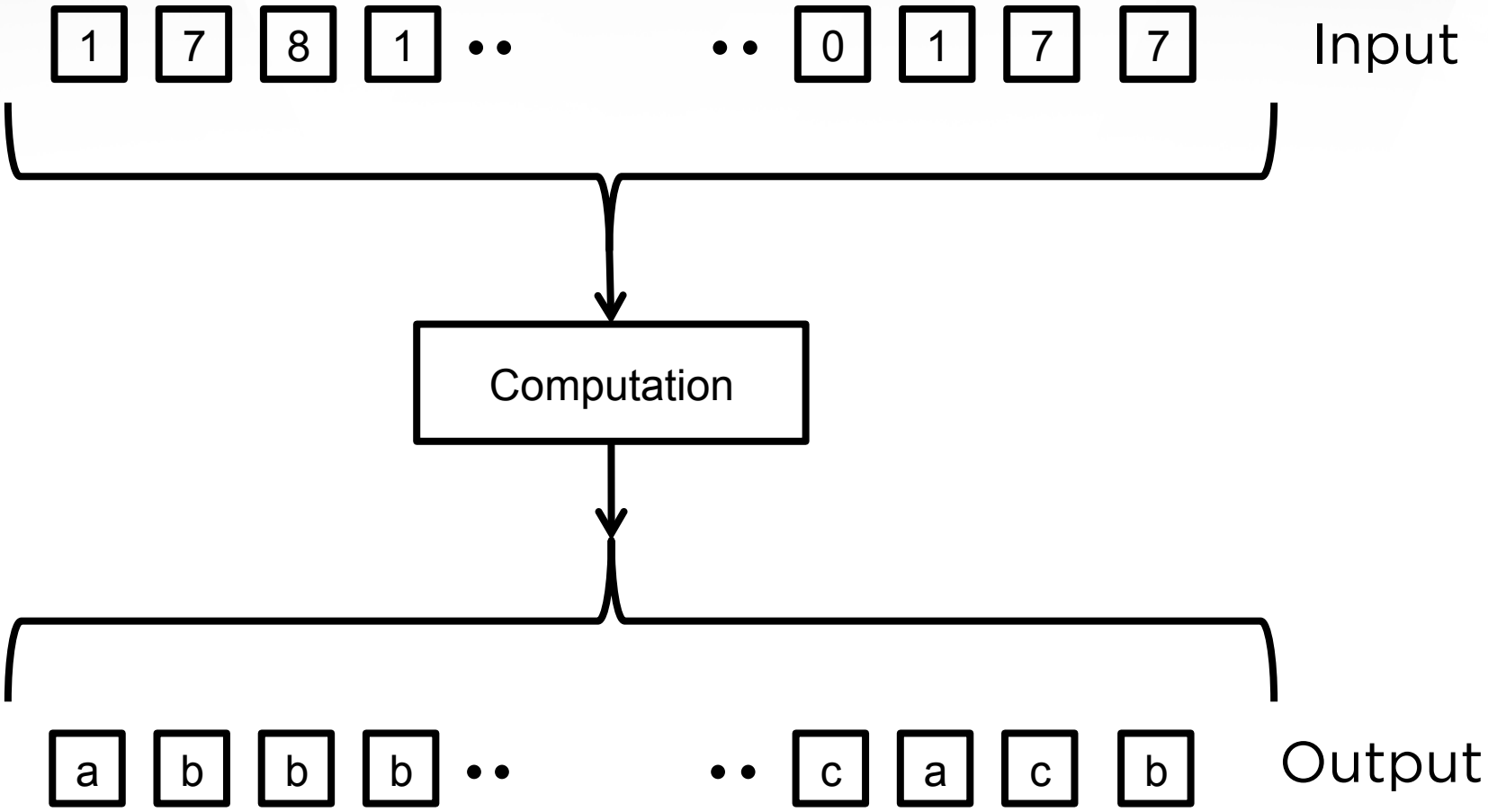
PRESENTED BY **Edo Liberty www.edoliberty.com**

Data mining in the batch setting



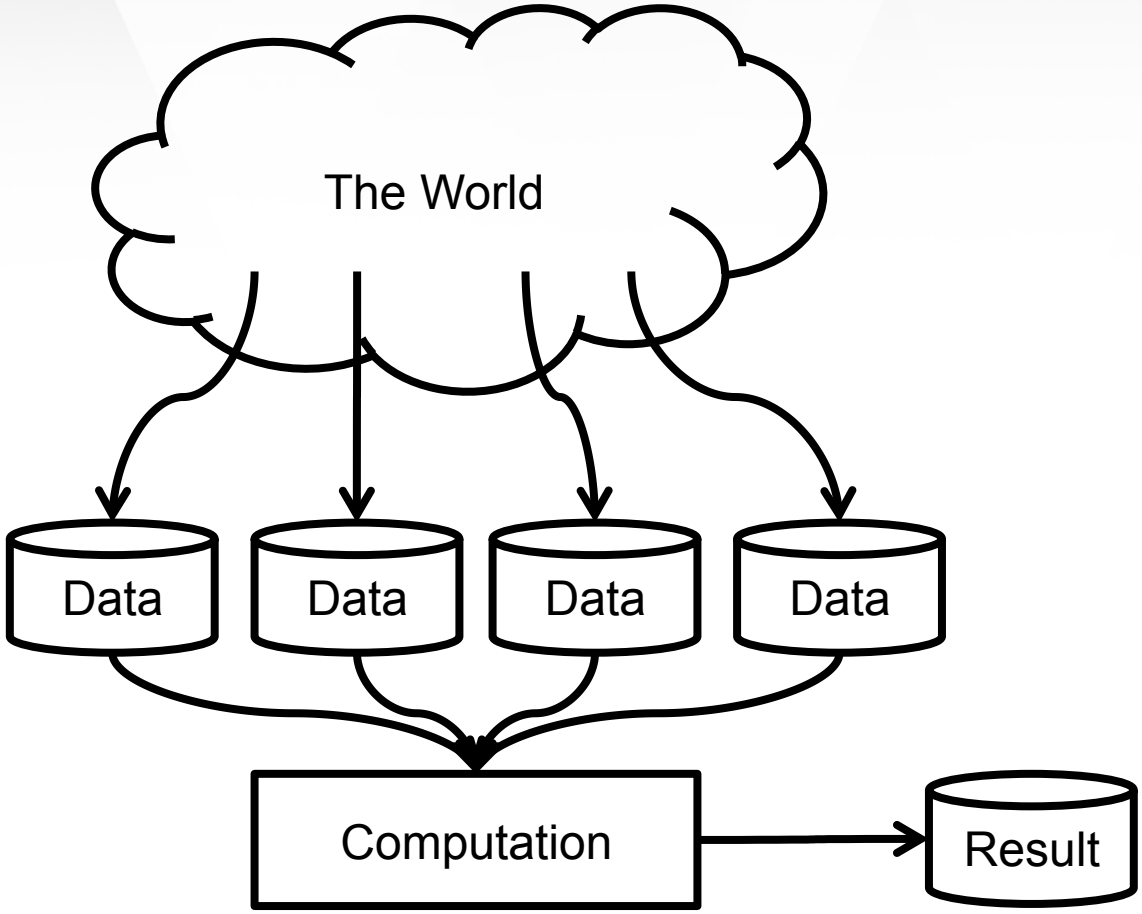
YAHOO!

Batch computational model



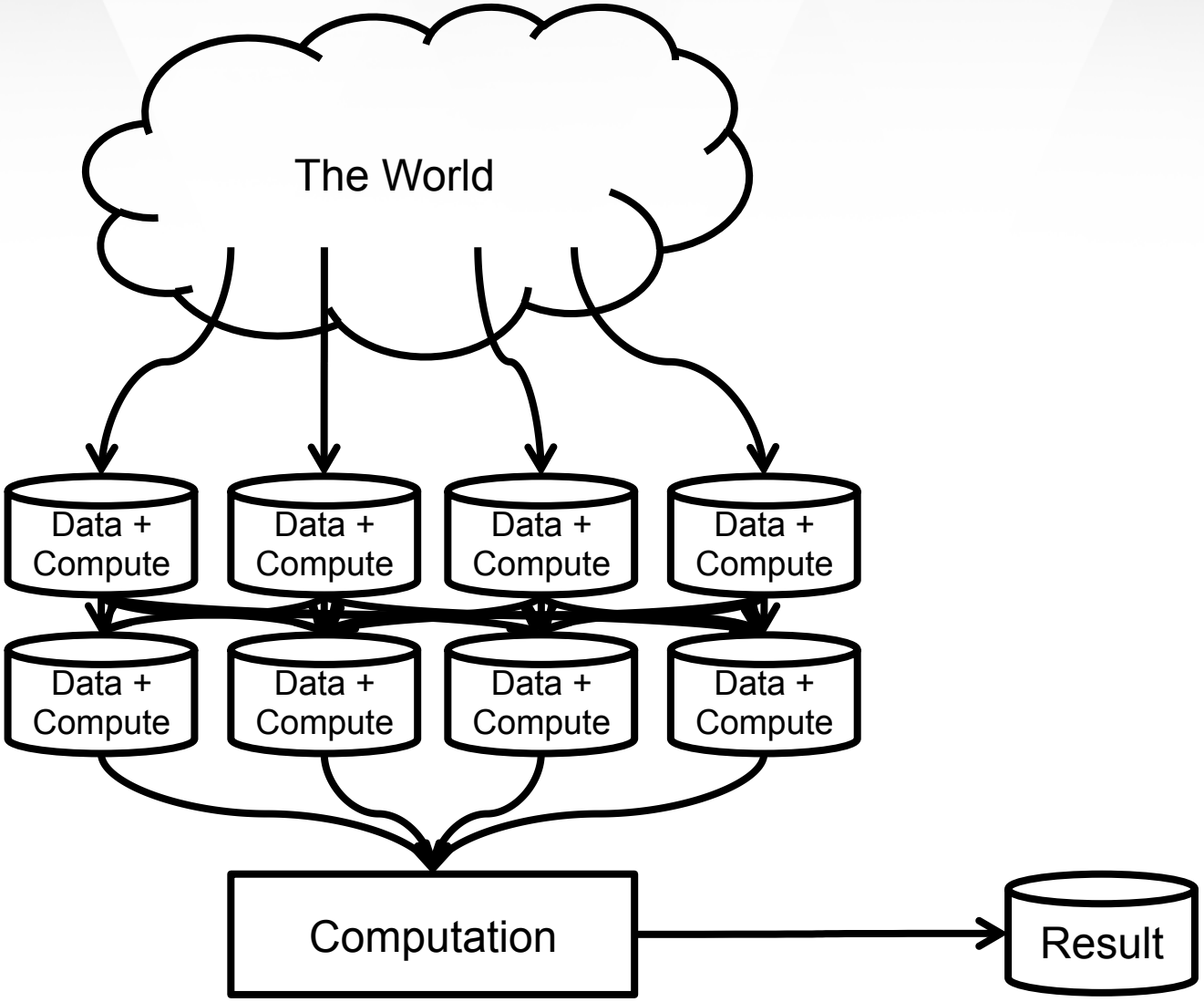
YAHOO!

Distributed storage



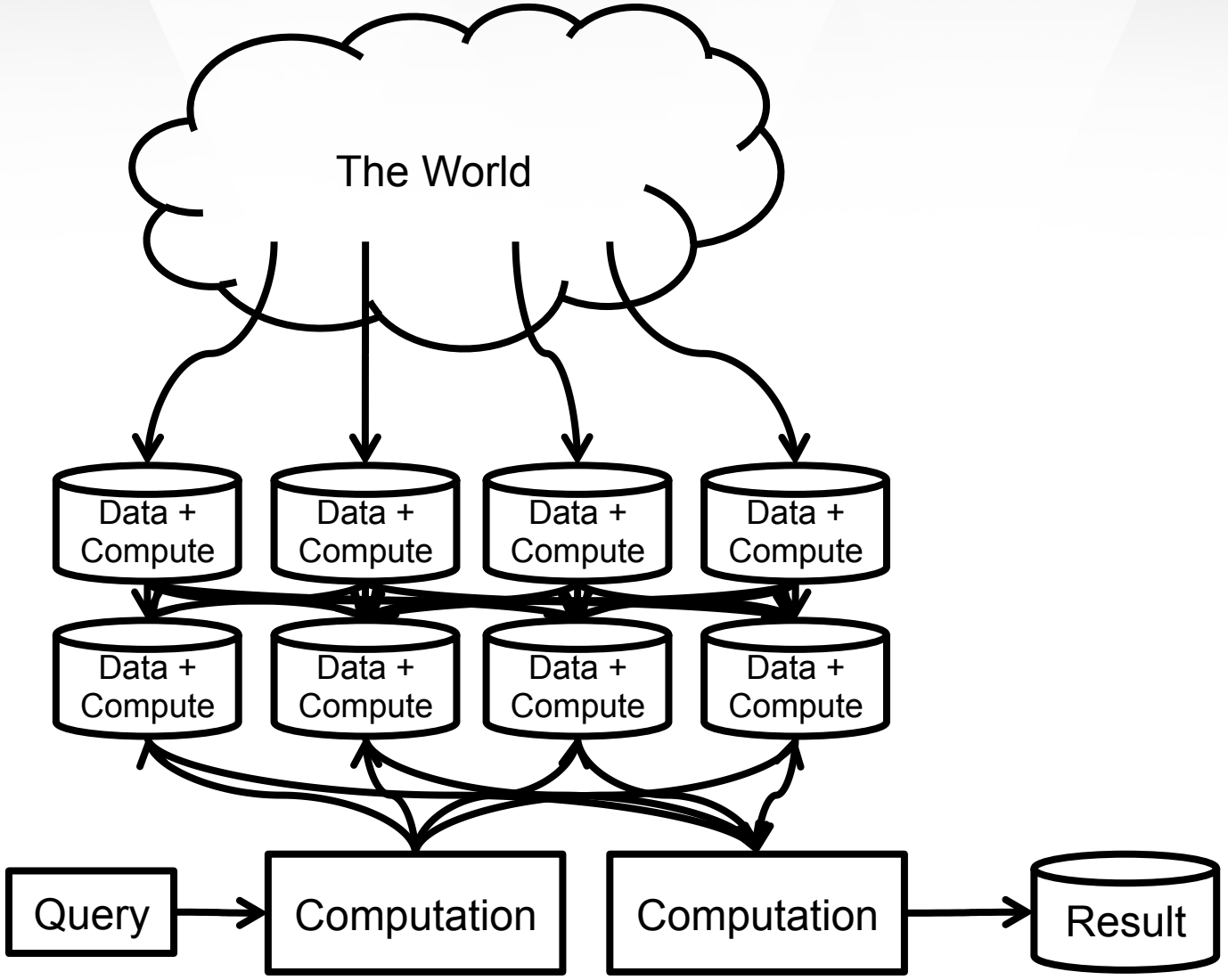
YAHOO!

Distributed model (map/reduce, message passing, …)



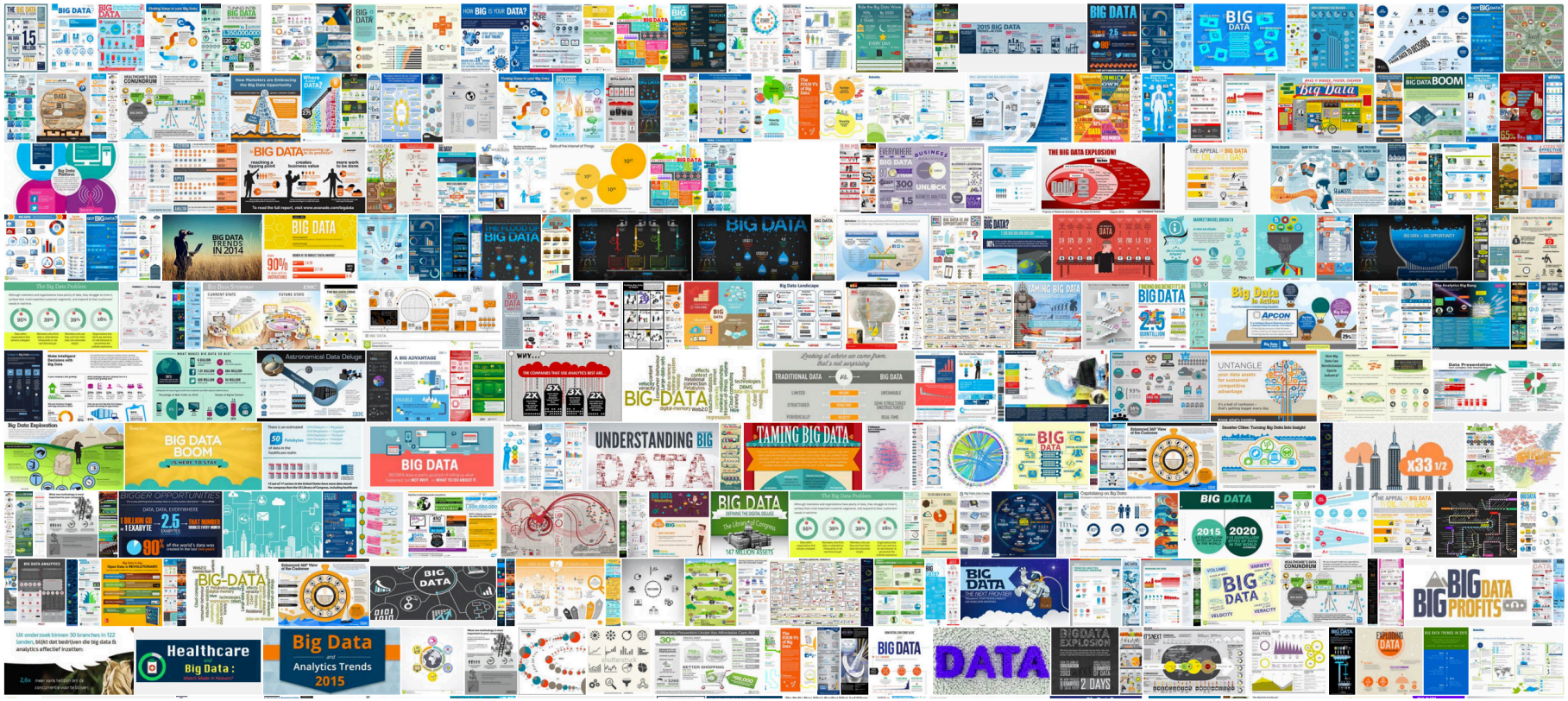
YAHOO!

Distributed model (indexes, tables, databases, …)



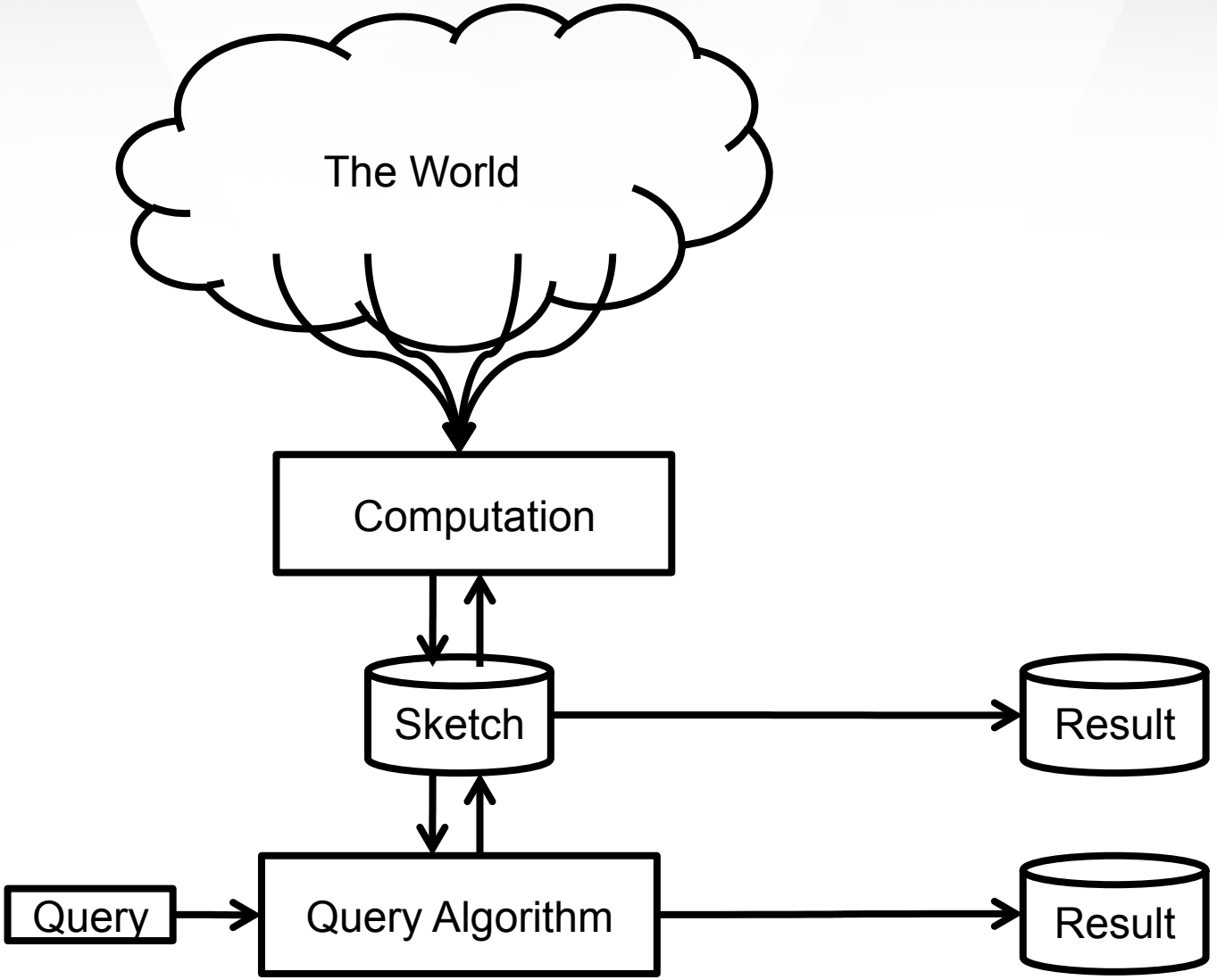
YAHOO!

Big-data *meta* infographic



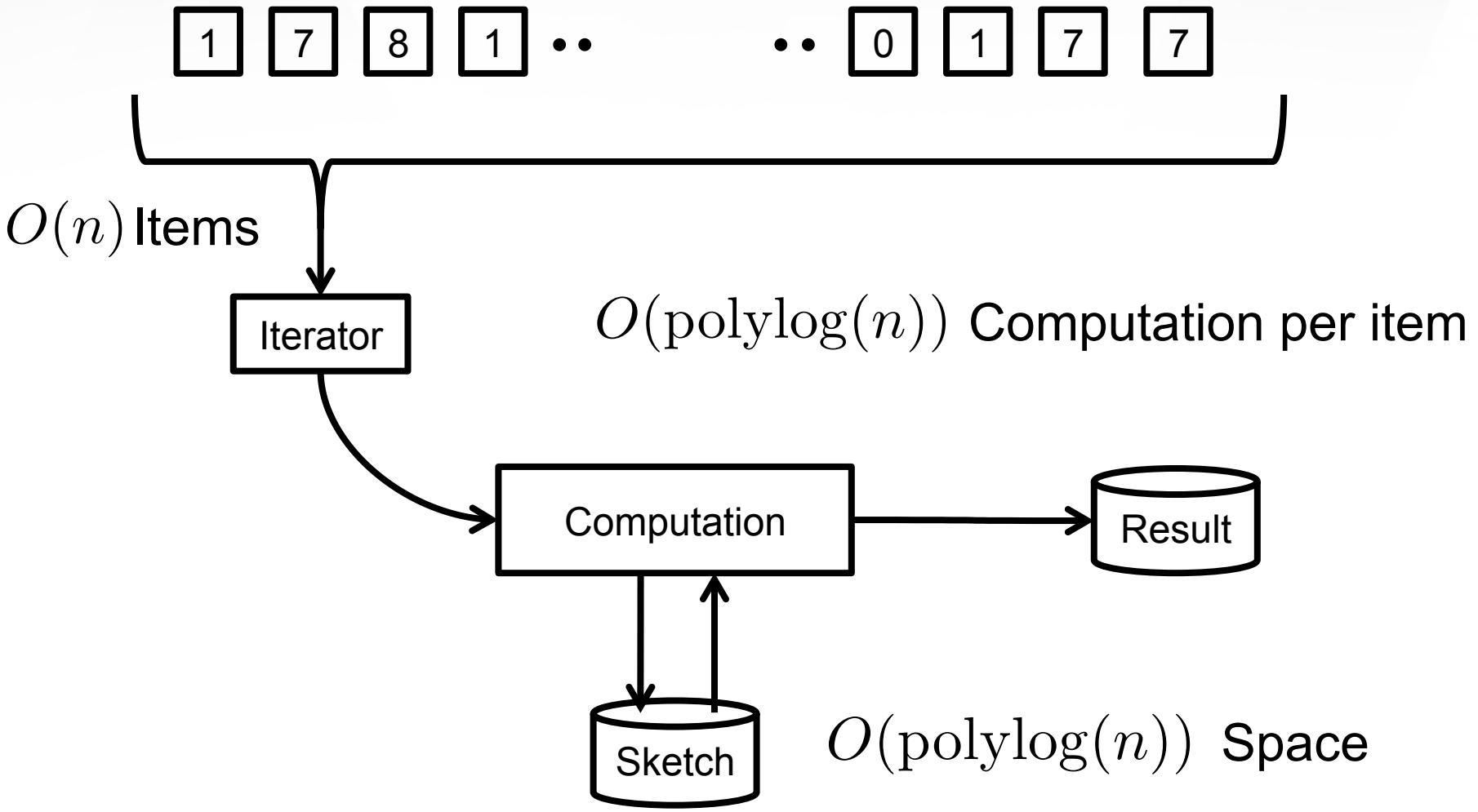
YAHOO!

The streaming computational model

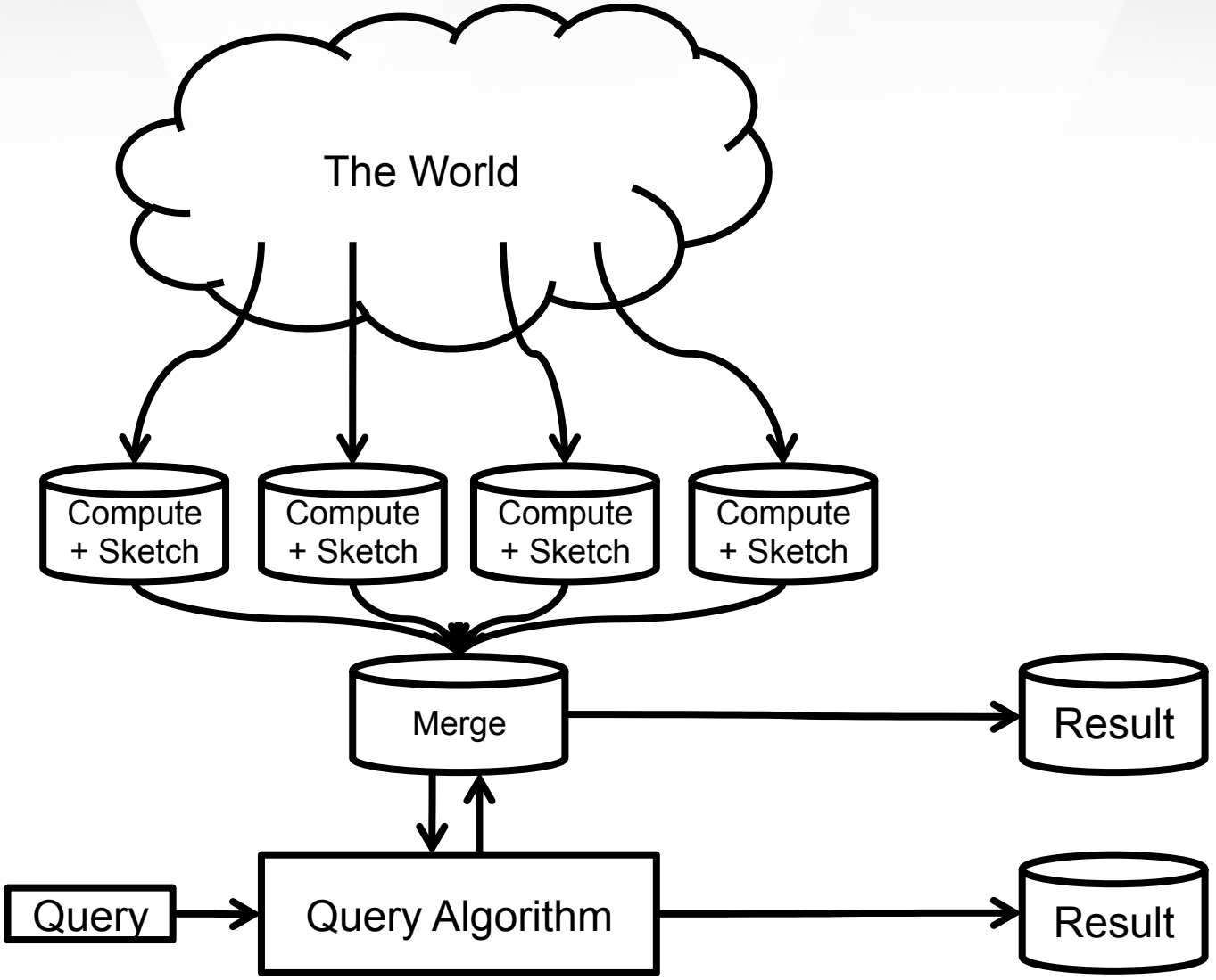


YAHOO!

The streaming computational model

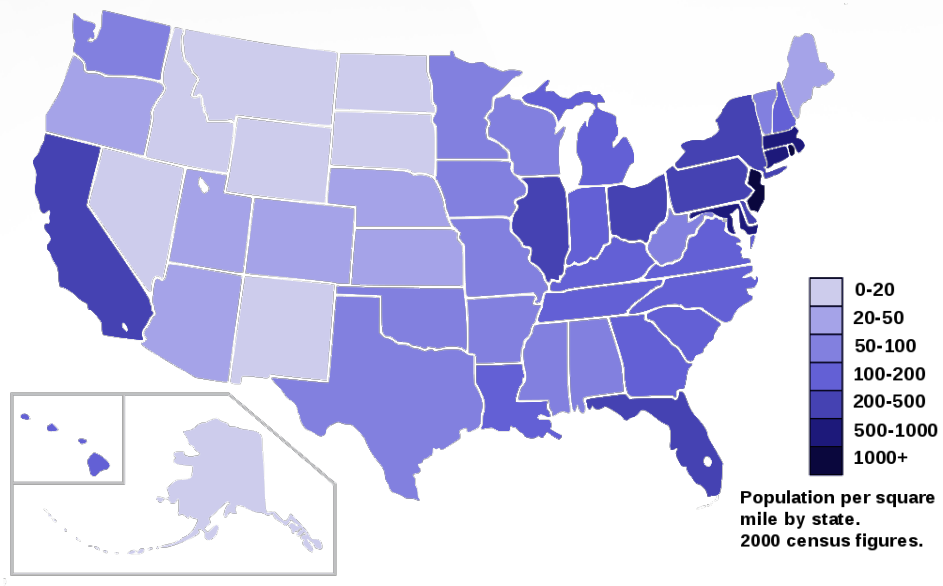


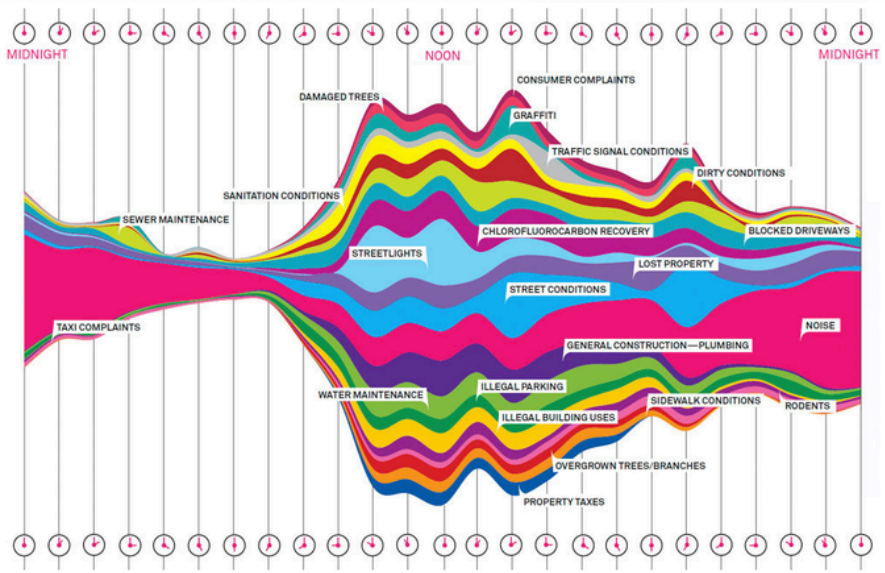
The distributed streaming computational model



YAHOO!

The distributed streaming computational model





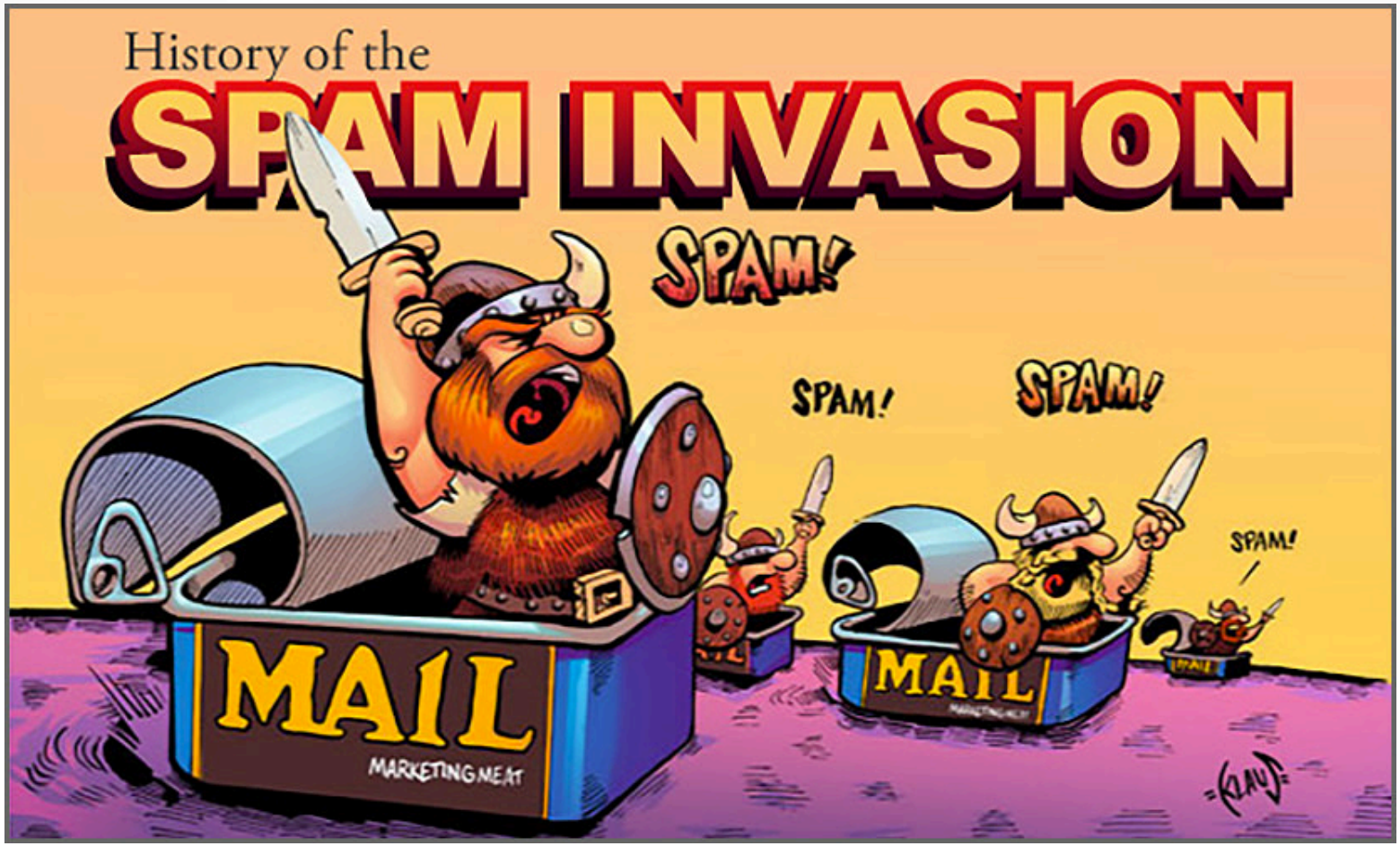
Sketches Library from YAHOO!

A Java software library of *stochastic streaming algorithms*

Overview | Download | GitHub | Comments

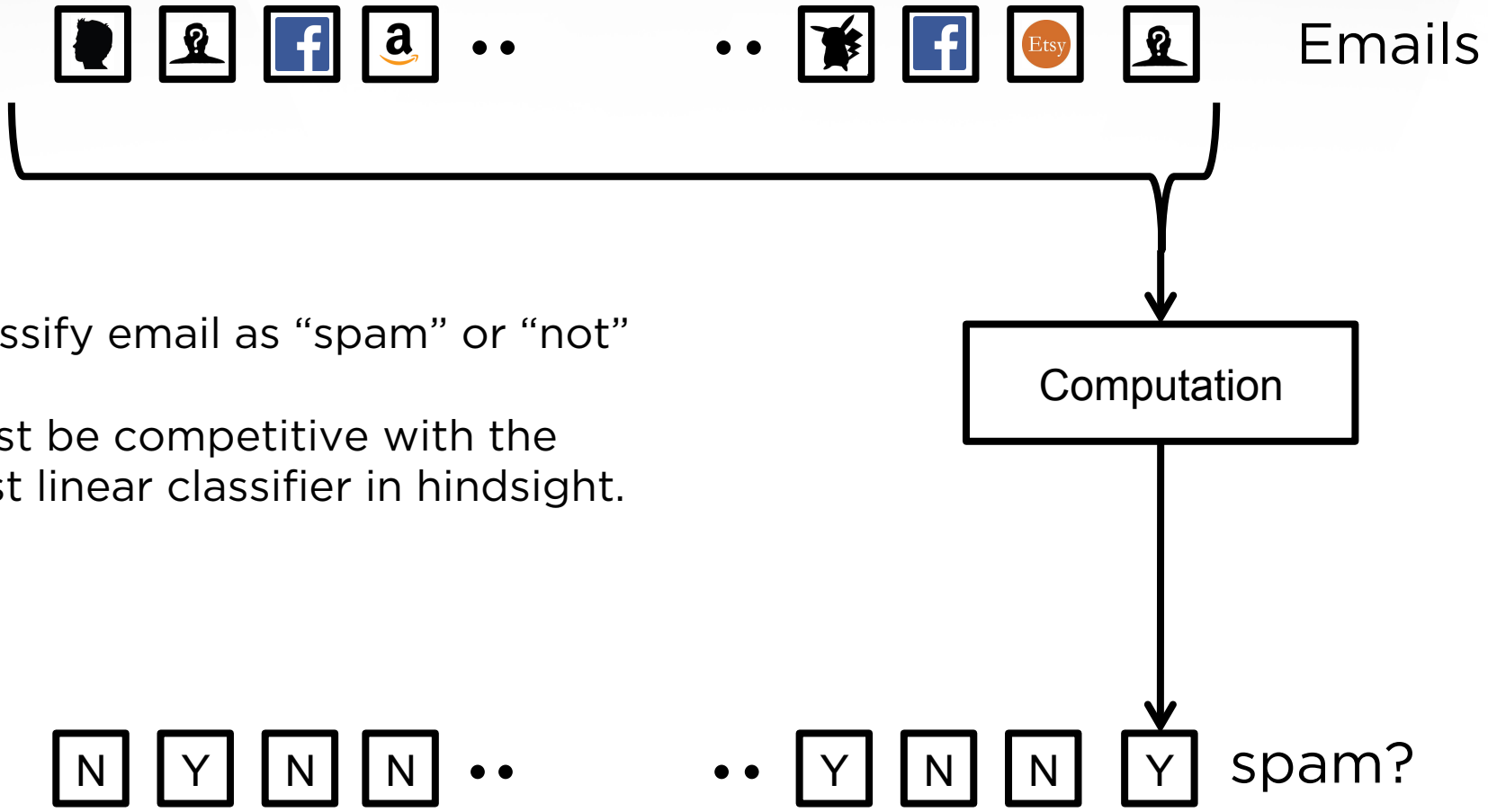
YAHOO!

Sometimes, you must take action immediately.



YAHOO!

Machine Learning



Classify email as “spam” or “not”

Must be competitive with the best linear classifier in hindsight.

Prediction, Learning, and Games, Cesa-Bianchi, Lugosi, 2006

Online Portfolio Management

- Manage your portfolio online
- Be competitive with best CRP model (constant rebalanced portfolio)

Elements of information theory, Cover, 1991
Efficient algorithms for universal portfolios, Kalai, Vempala, 2003
Efficient Algorithms for Online Game Playing and Universal Portfolio Management, Agarwal, Hazan, 2006

YAHOO!

Online Advertising

YAHOO! kitchen tables with benches Search

Web Images Video Anytime

Also try: **small** kitchen tables with benches

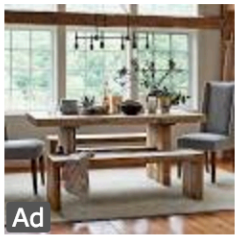
Ads related to: **kitchen tables with benches**

Deals on Dining Tables - Shop Dining Tables Online.
www.raymourflanigan.com/Dining-Room
Shop Dining **Tables** Online. Enjoy Store-Wide Savings Today!
490 Fulton Street, New York, NY (347) 416-5019 Directions

Financing Options - Store Locator - Dining Chairs - Glass **Tables**

Kitchen Benches on Sale - 20%-50% Off Kitchen Benches.
www.ATGStores.com/Dining**Benches**
20%-50% Off **Kitchen Benches**. 7 Day Customer Service & Free Shipping!
Brands: Homelegance, Liberty Furniture, Nuevo Living, Sunny Designs and more

- Dining **Benches With** Arms
- Dining **Benches** on Sale
- **Kitchen** Islands & Carts
- Traditional Dining Bench
- Modern Dining **Benches**
- Backless Dining **Benches**



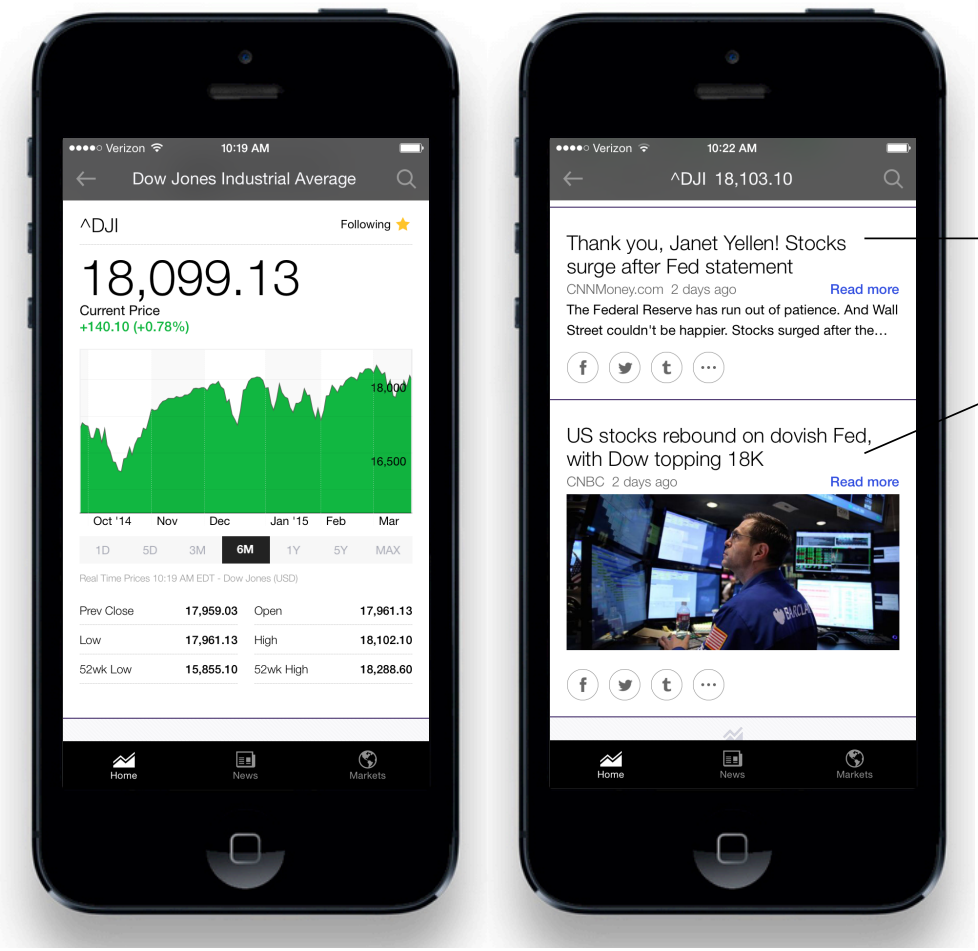
Emmerson 62" Dining Table,...

$799.00

west elm

YAHOO!

Yahoo Finance App

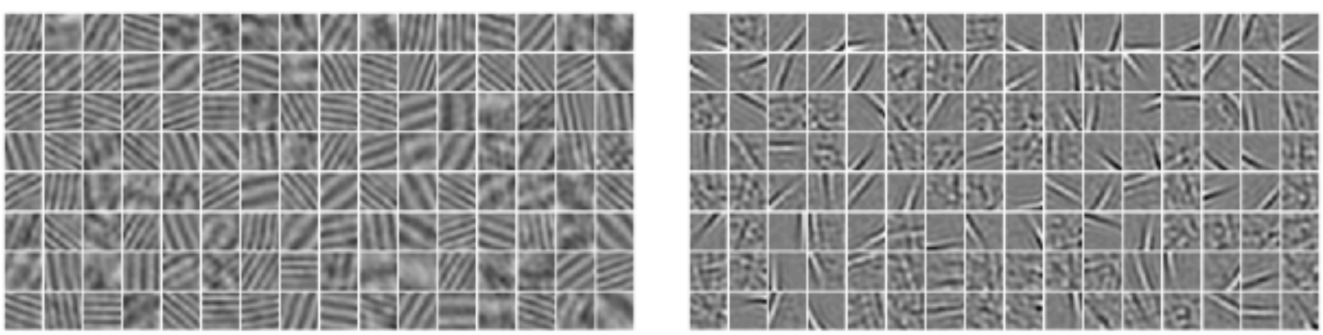


Same story line or not?

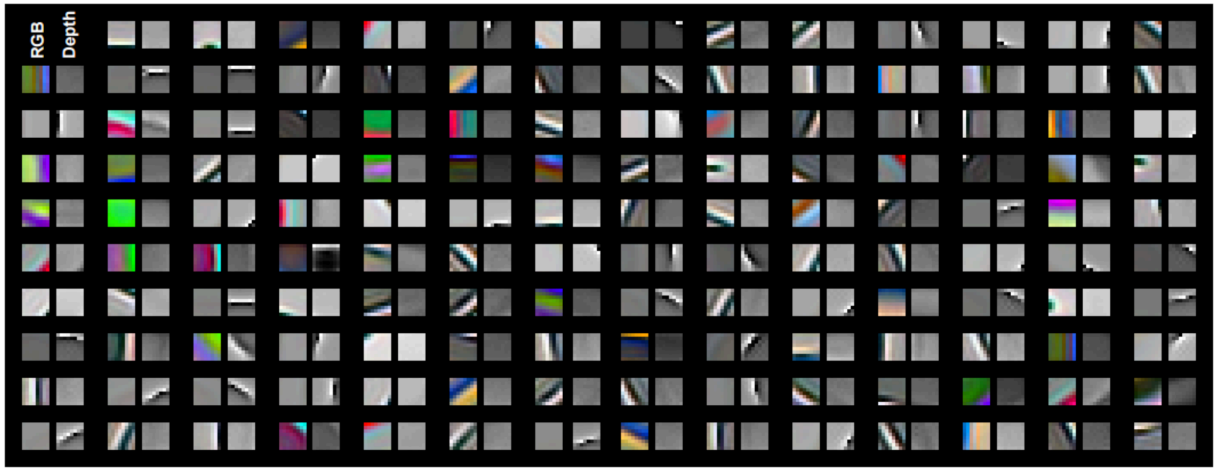
1) The answer depends on the future
2) We have to decide now…

YAHOO!

Online Feature Engineering

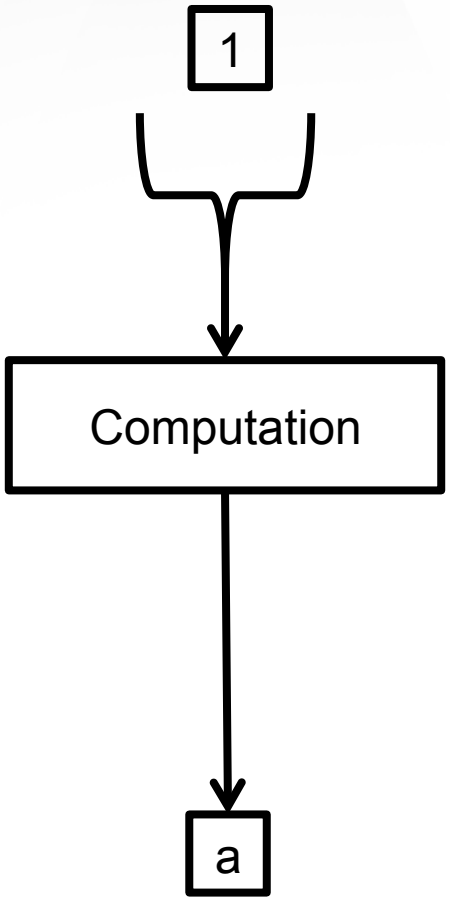


Learning feature representation with k-means, Adam Coates and Andrew Ng



On the Applicability of Unsupervised Feature Learning for Object Recognition in RGB-D Data, Manuel Blum, Jost Tobias Springenberg, Ja Wulfing, and Martin Riedmiller

Online model

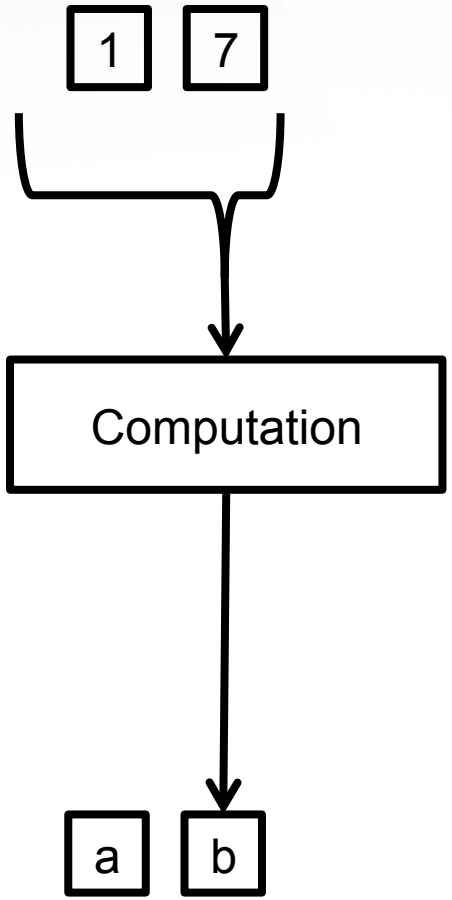


Input

Output

YAHOO!

Online model

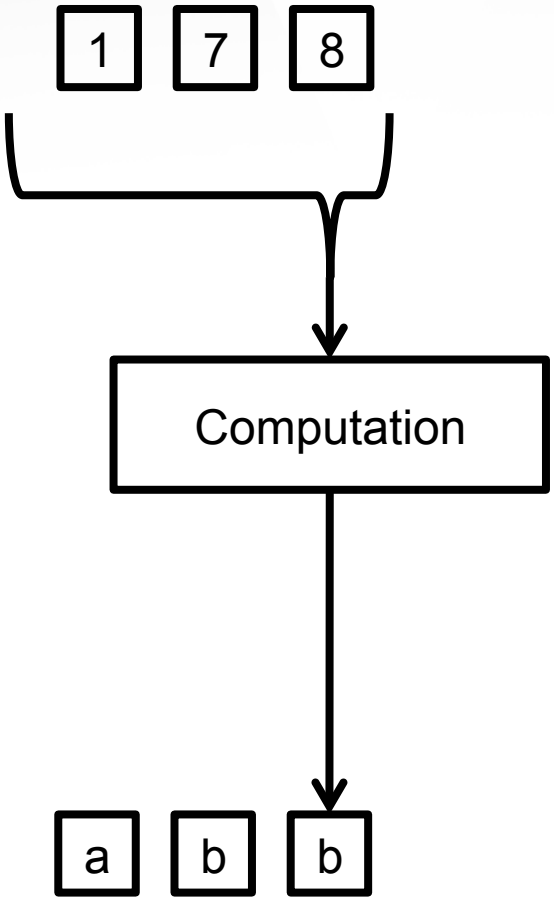


Input

Output

YAHOO!

Online model

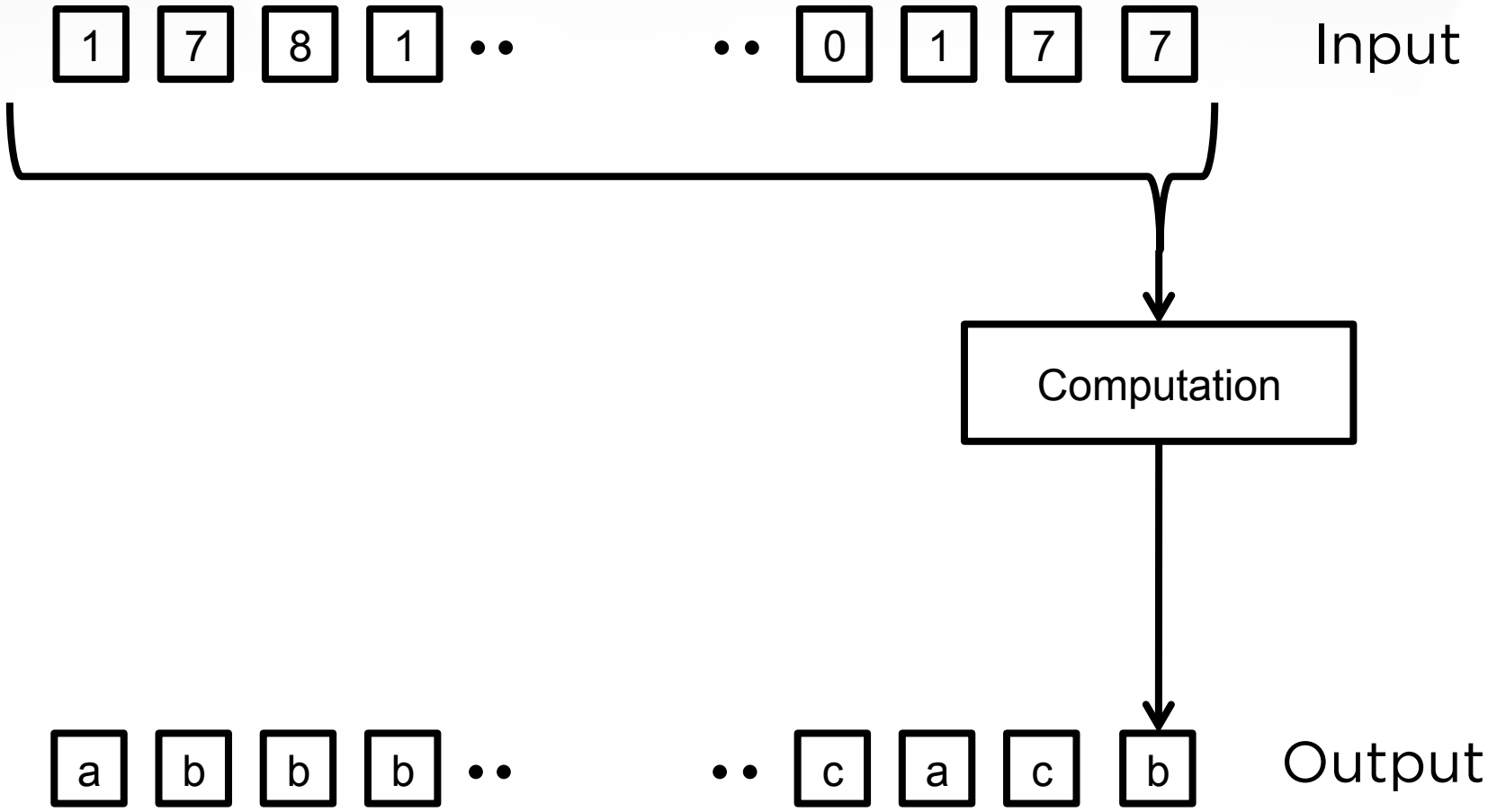


Input

Output

YAHOO!

Online model



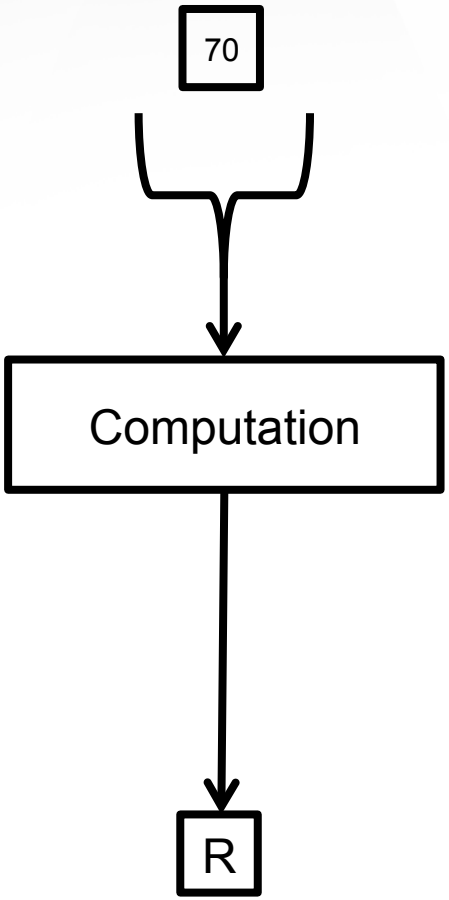
Ski Rental Problem Example

Ski Rental

Rent: x$ /day

Buy: 1000$

Ski Rental

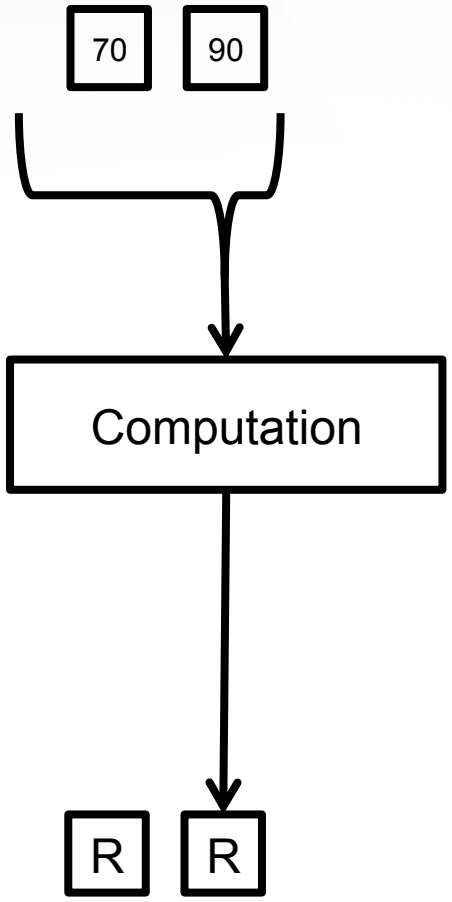


70
+

70

YAHOO!

Ski Rental

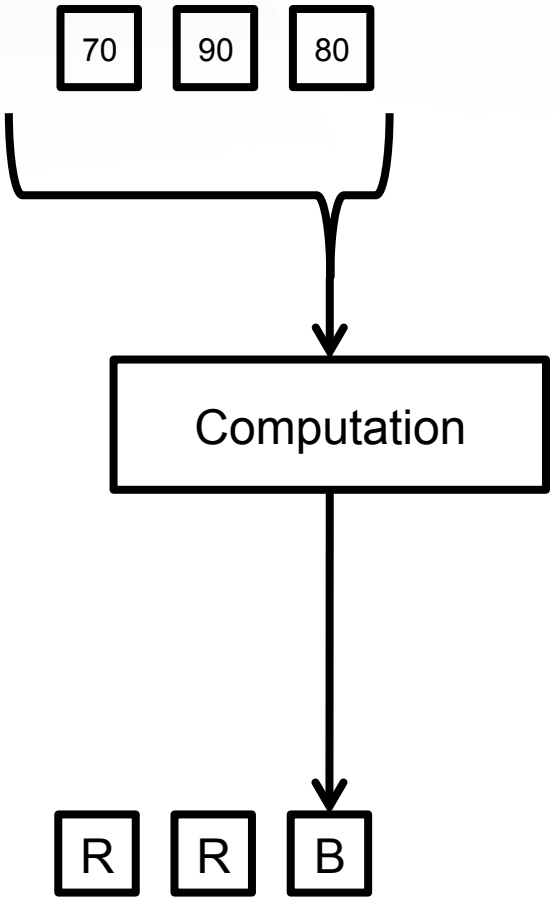


70
+
90

160

YAHOO!

Ski Rental

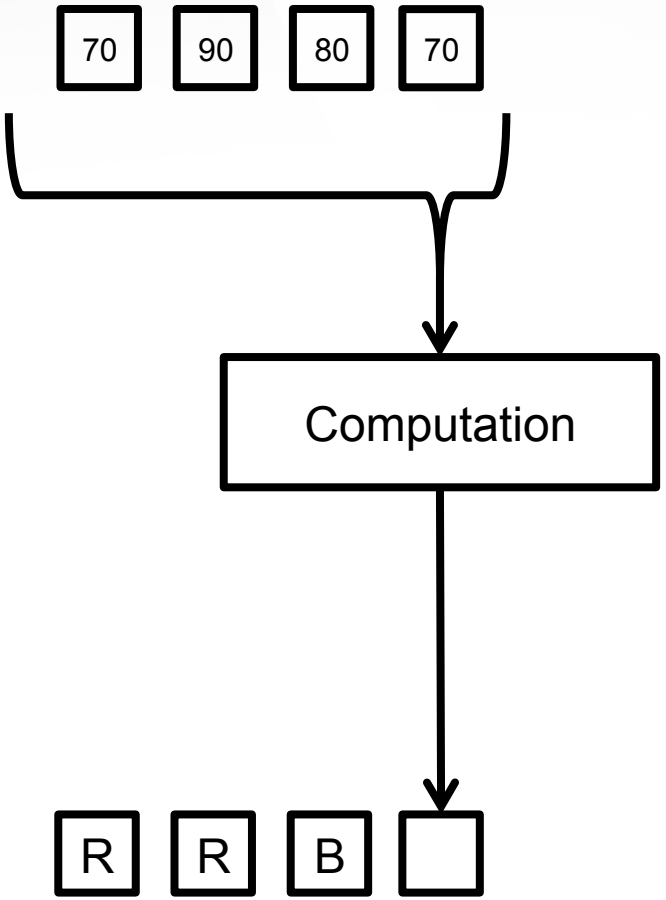


70
+
90
+
1000

1160

YAHOO!

Ski Rental

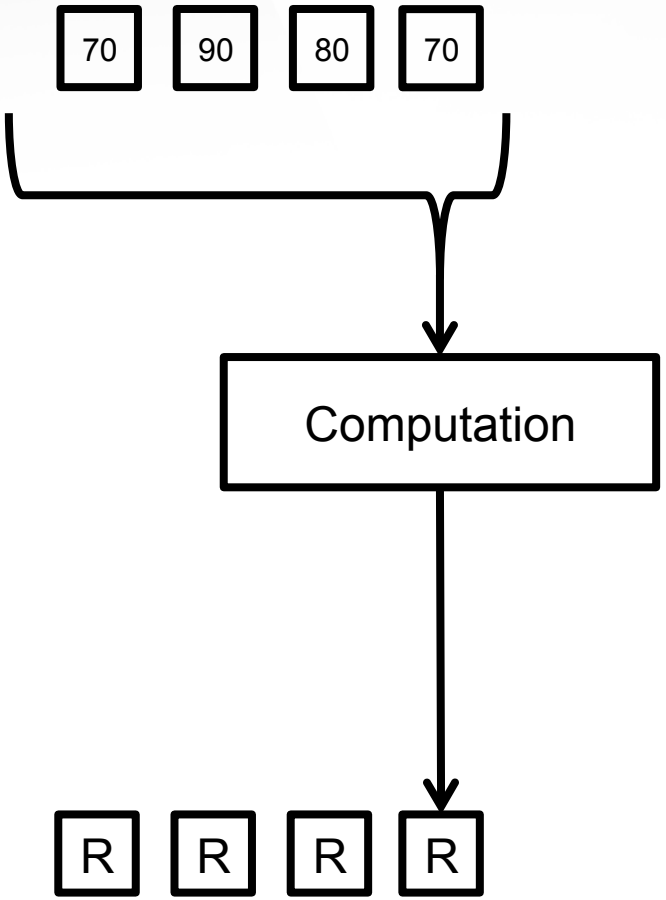


70
+
90
+
1000
+
0

1160

YAHOO!

Ski Rental



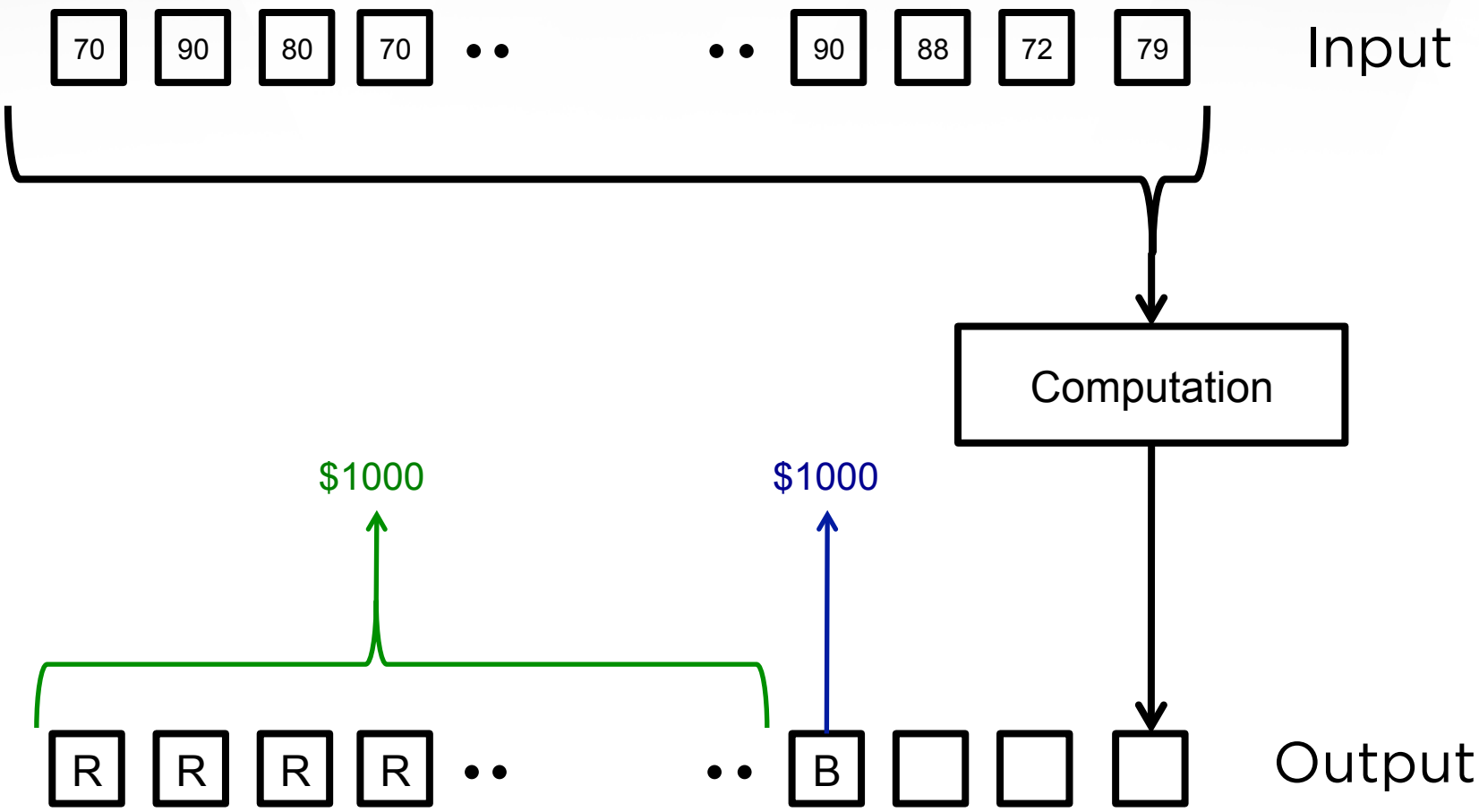
70
+
90
+
80
+
70

310

Much less than 1160…

YAHOO!

Ski Rental

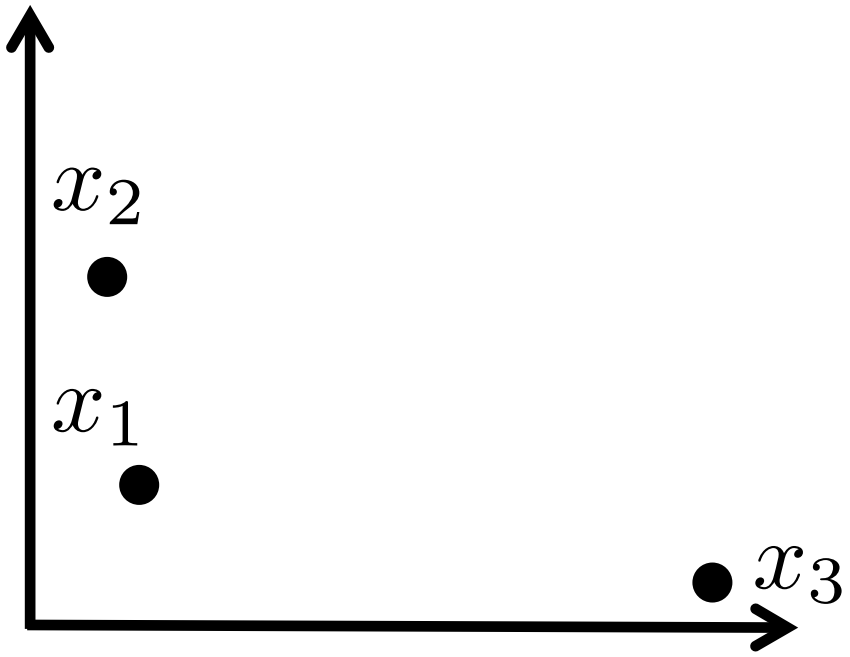


YAHOO!

Online Regression

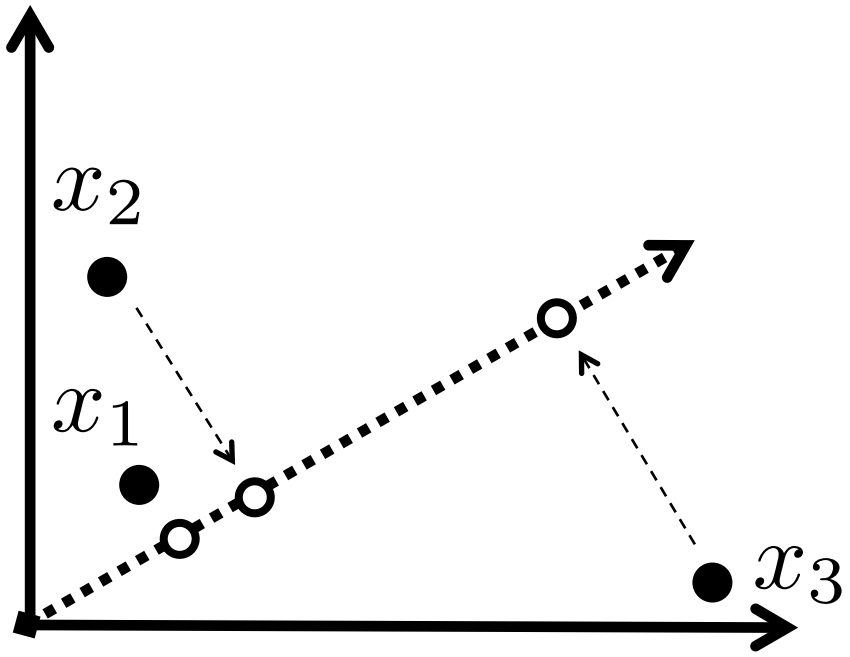
Online Principal Components Analysis, Boutsidis, Garber, Karnin, Liberty 2014
Online PCA with Spectral Bounds, Karnin, Liberty, 2015

Regression



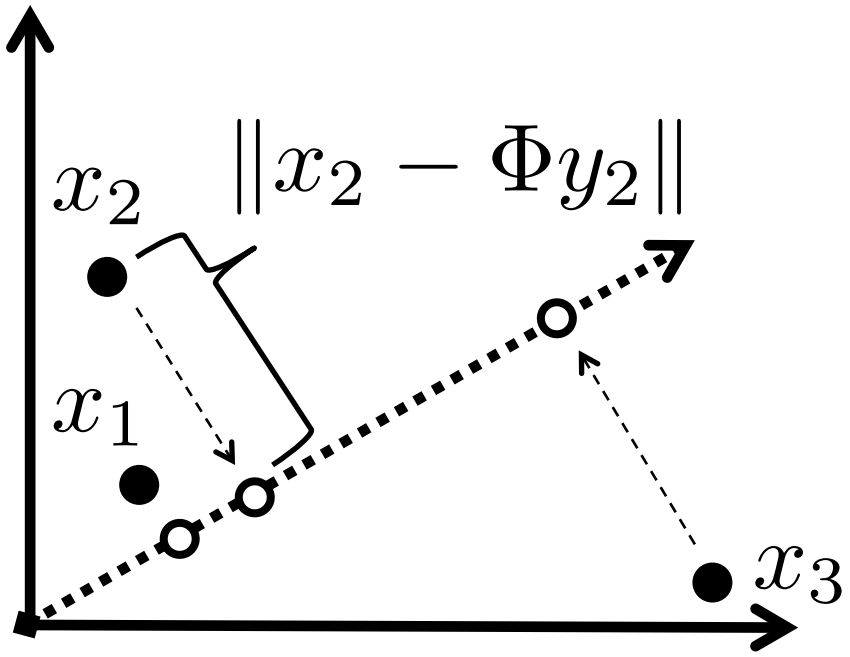
YAHOO!

Regression



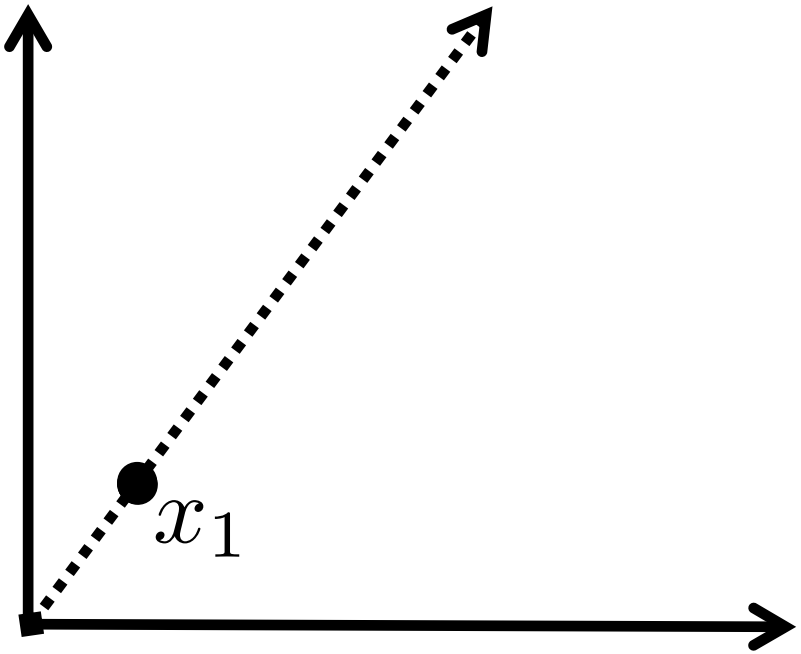
YAHOO!

Regression

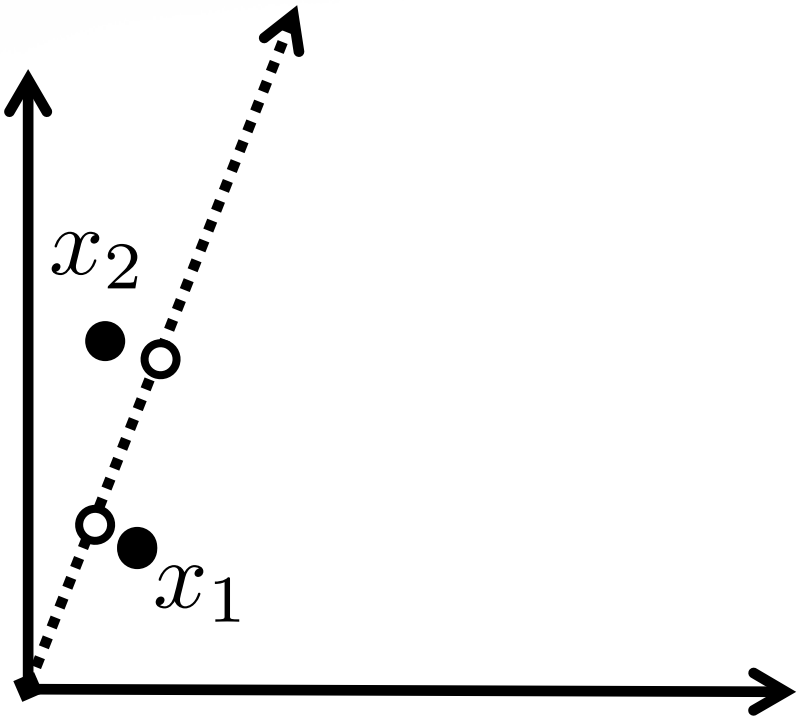


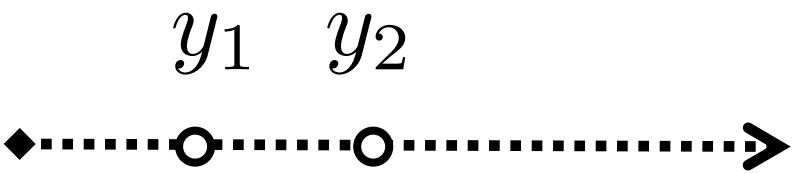
YAHOO!

Online Regression

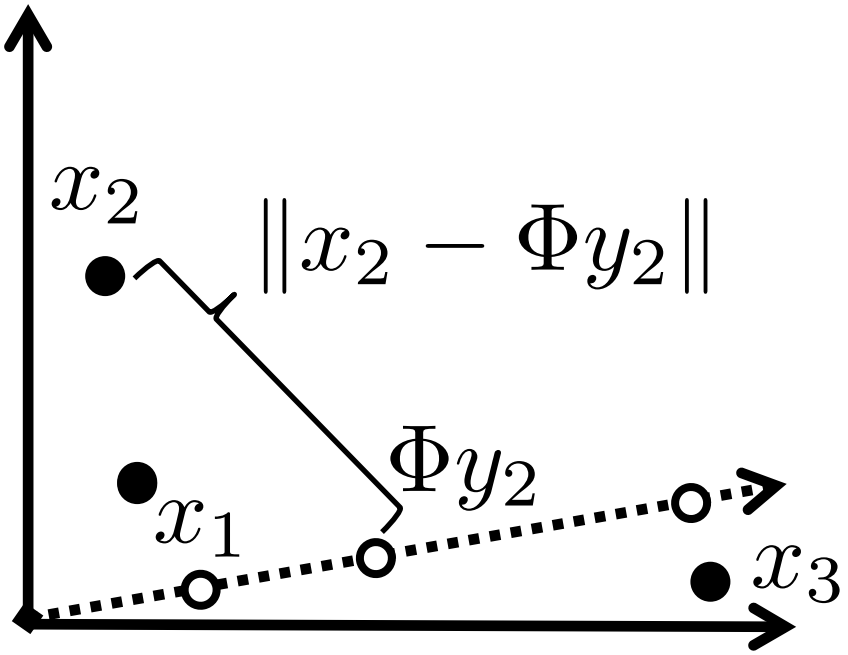


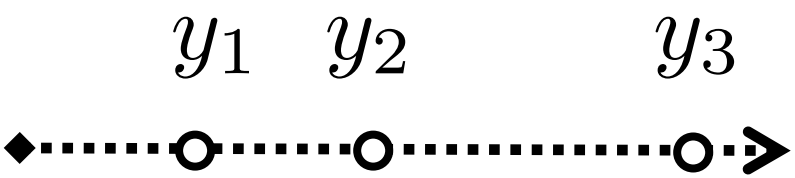
Online Regression





Online Regression



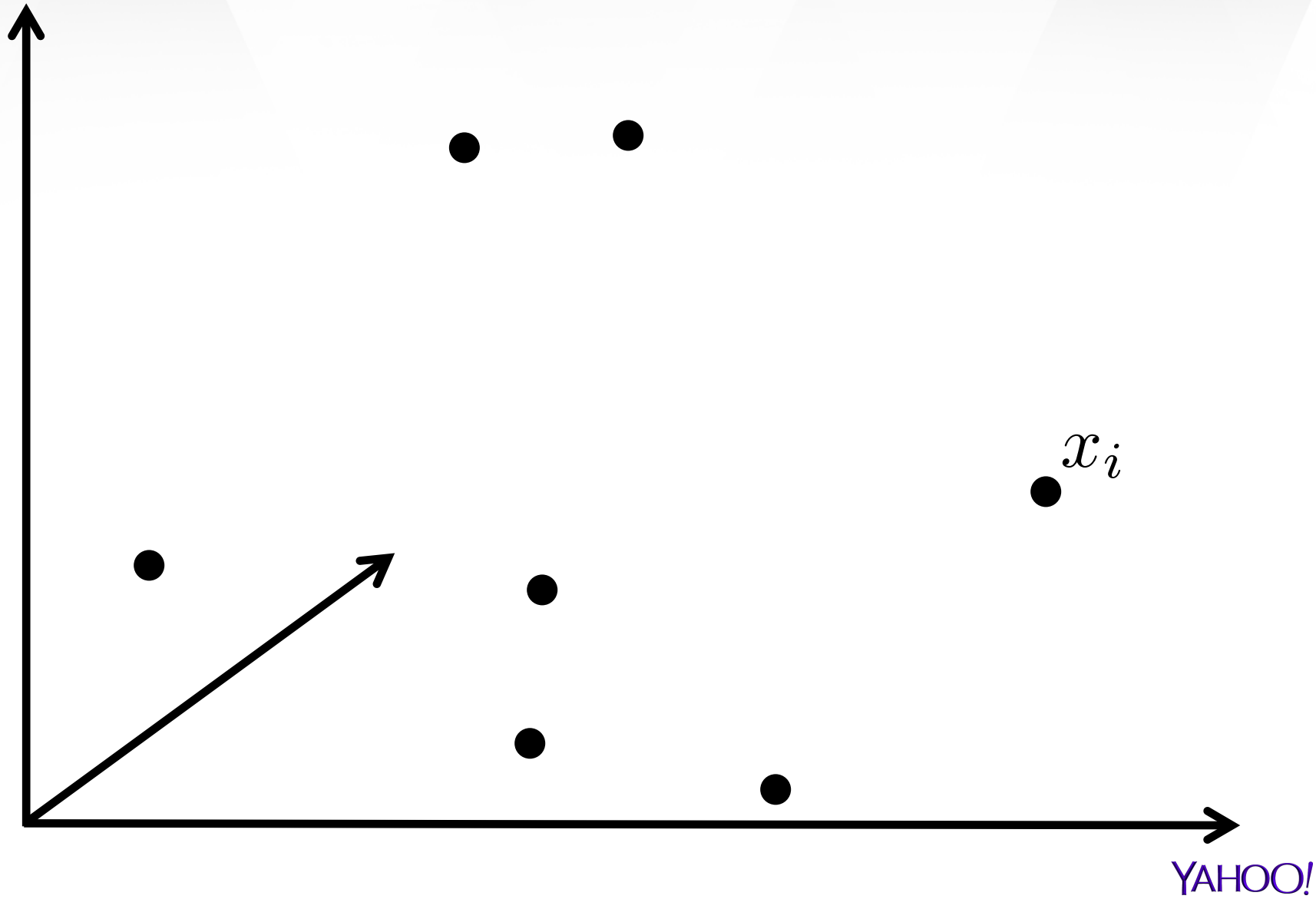


YAHOO!

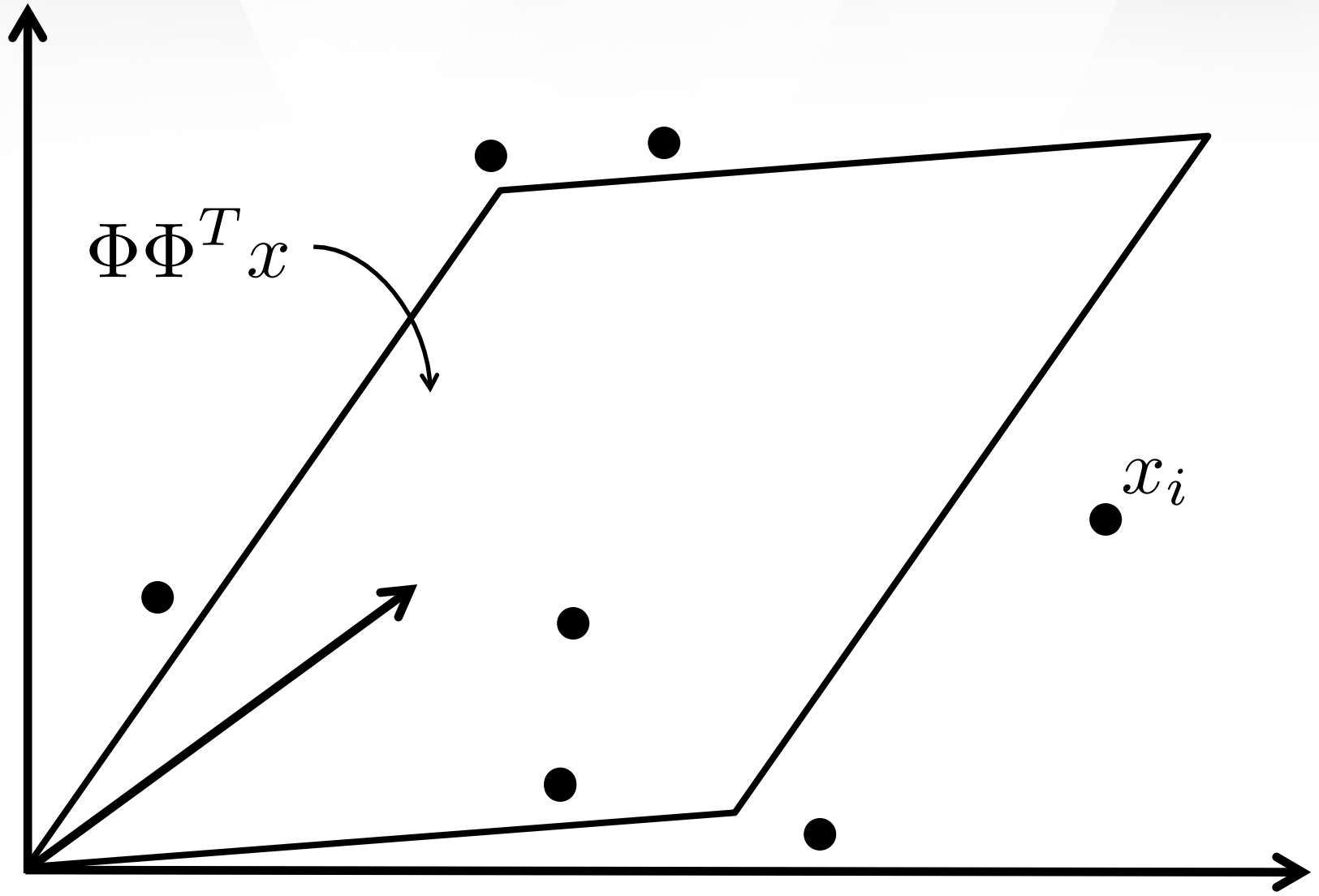
Online Principal Component Analysis

Online Principal Components Analysis, Boutsidis, Garber, Karnin, Liberty 2014
Online PCA with Spectral Bounds, Karnin, Liberty, 2015

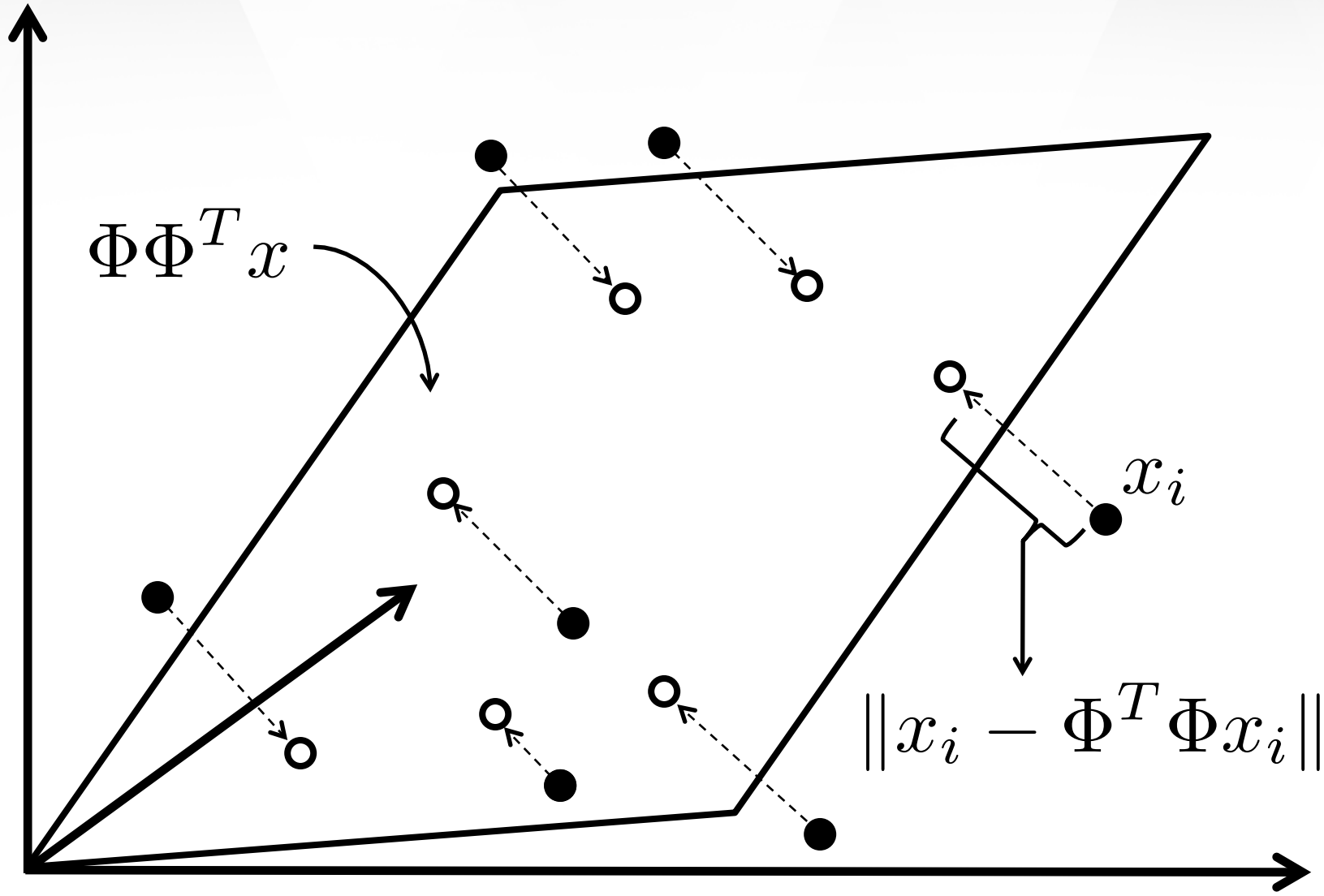
Principal Component Analysis



Principal Component Analysis



Principal Component Analysis



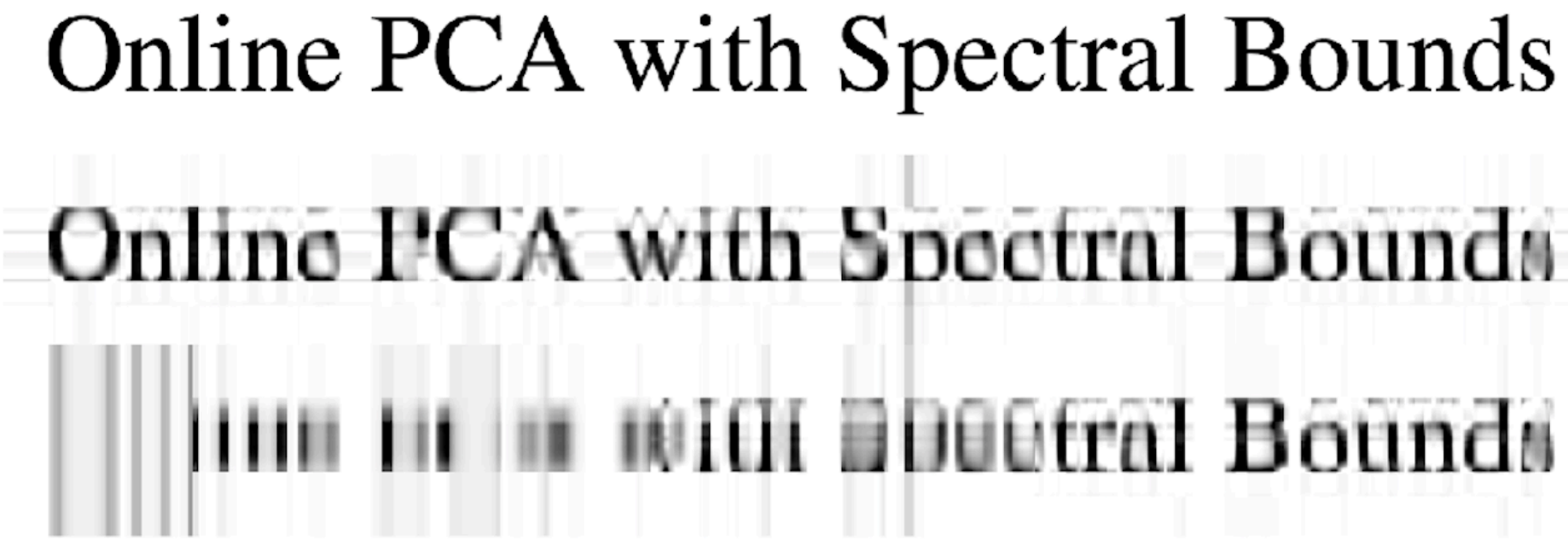
Online Principal Component Analysis

Algorithm 1 Fixed Error: Conceptual Algorithm

input: X, Δ
$U \leftarrow$ all zeros matrix
for $x_t \in X$ **do**
  if $\|(I - UU^T)X_{1:t}\|^2 \geq \Delta$
    Add the top left singular vector of $(I - UU^T)X_{1:t}$ to U
  yield $y_t = U^T x_t$
end for

Online PCA with Spectral Bounds, Karnin, Liberty, 2015

Online PCA, a visual example



YAHOO!

Online PCA combined with online learning

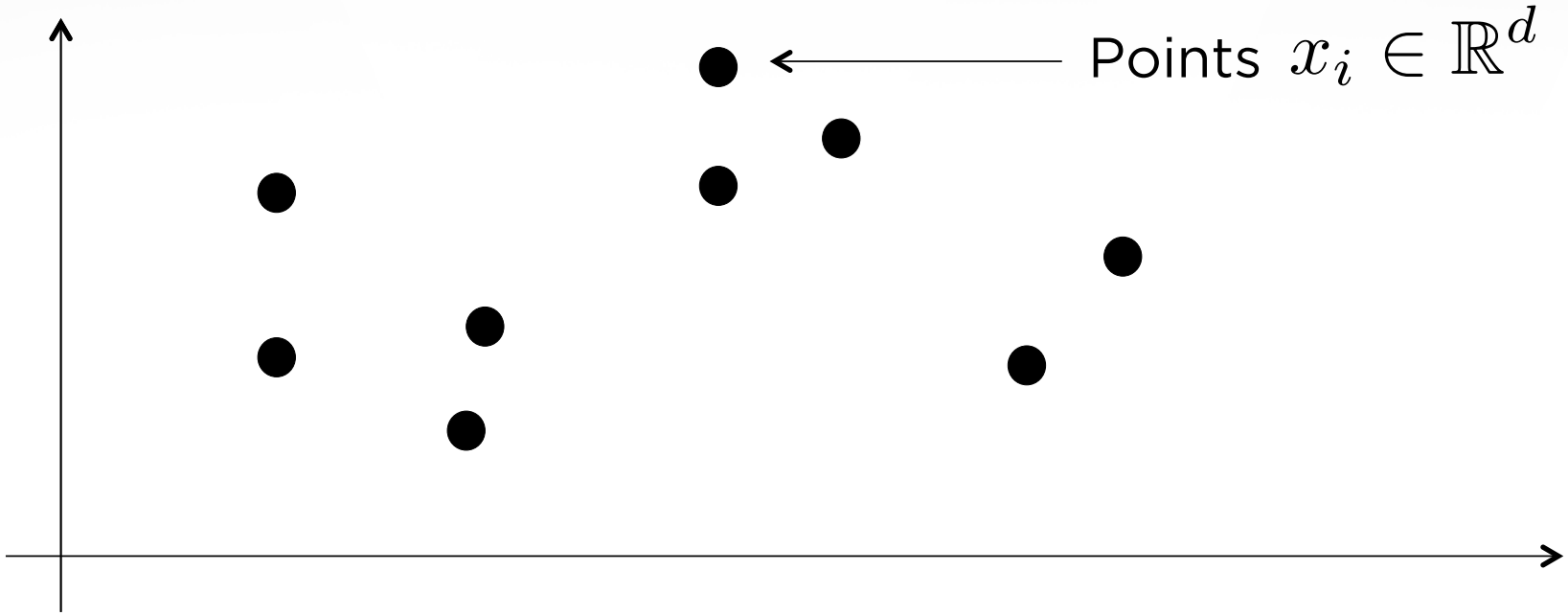
- Saves on storage and computation
- Does not hurt classification accuracy
- Is not worse than offline PCA and learning

	Before PCA		After PCA		
Dataset	**Dimension**	**Accuracy**	**Dimension**	**Online**	**Offline**
census	401	0.93	40	0.94	0.94
maptask	5944	0.78	20	0.75	-
nomao	174	0.58	5	0.59	0.58
rcv1	43001	0.86	500	0.88	-
letter	16	0.75	2	0.76	0.76

Online k-means clustering

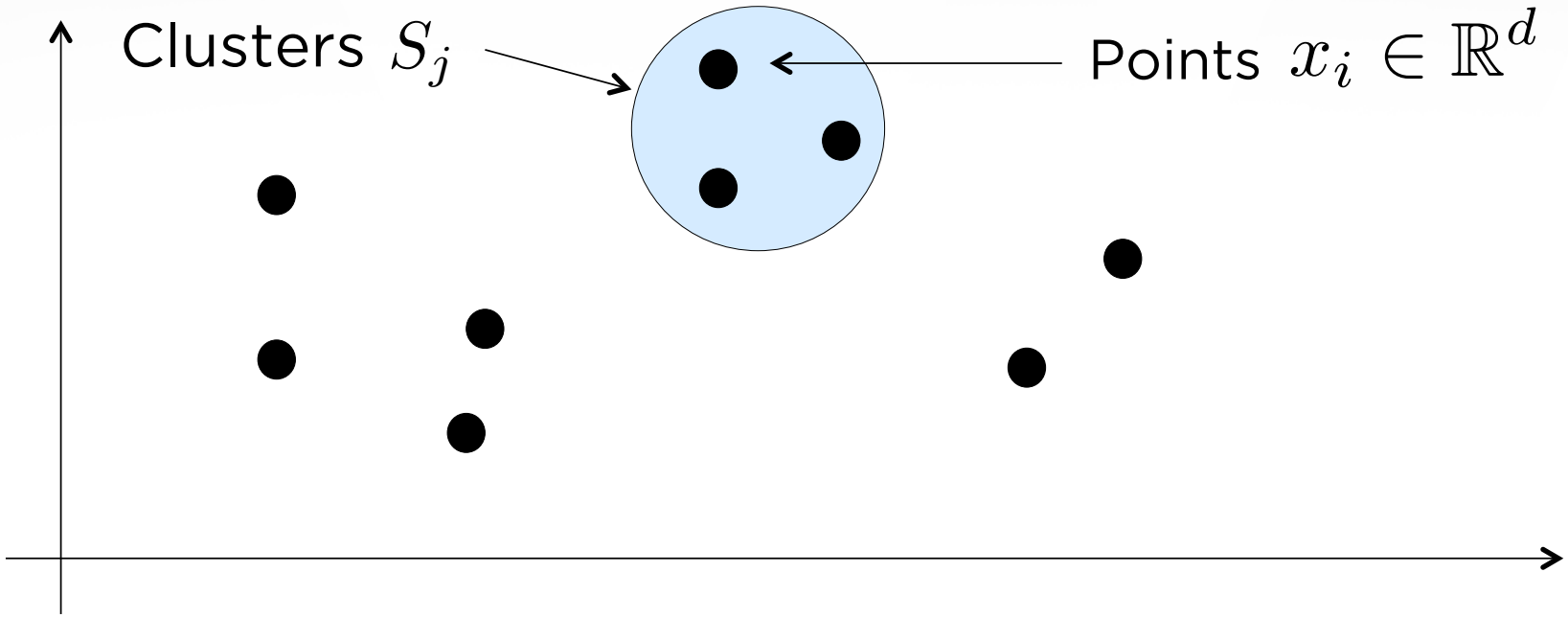
An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko, 2014

k-means clustering



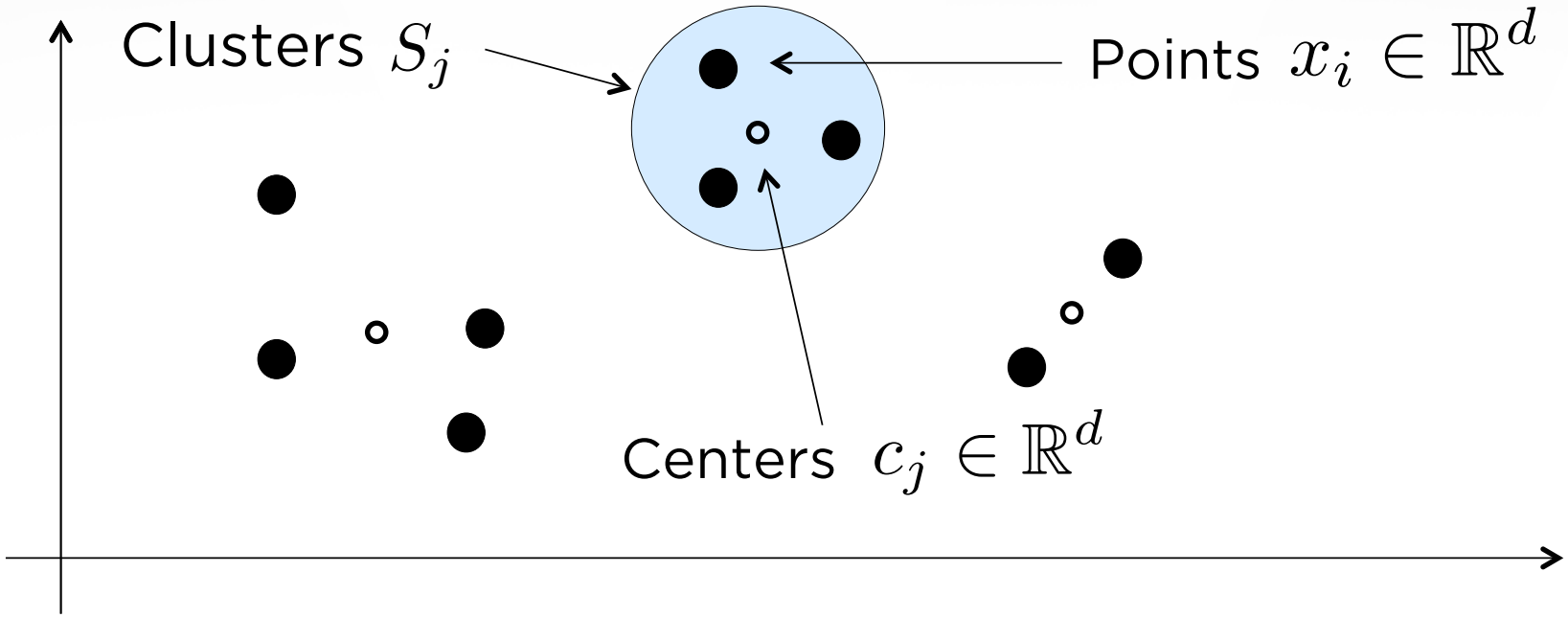
Small $\|x_i - x_j\|$ indicates the two points are "similar"

k-means clustering



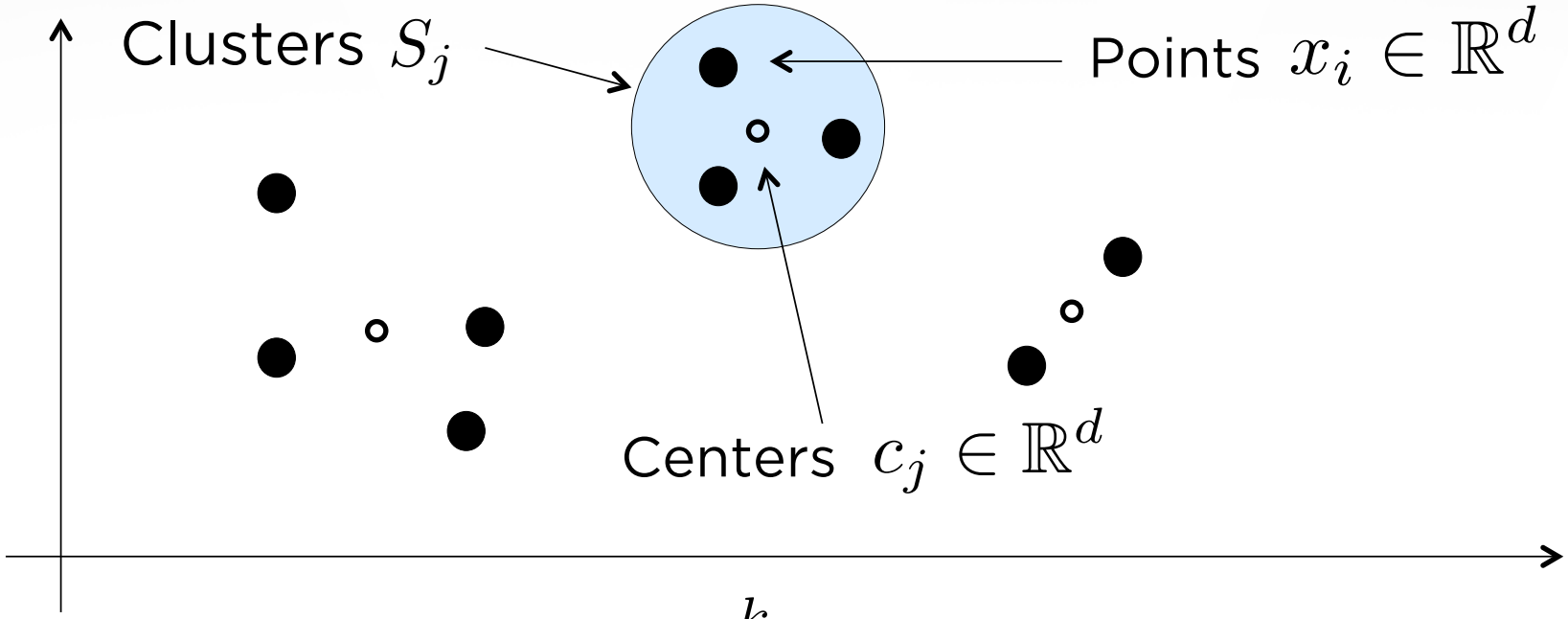
A cluster is a set of points

k-means clustering



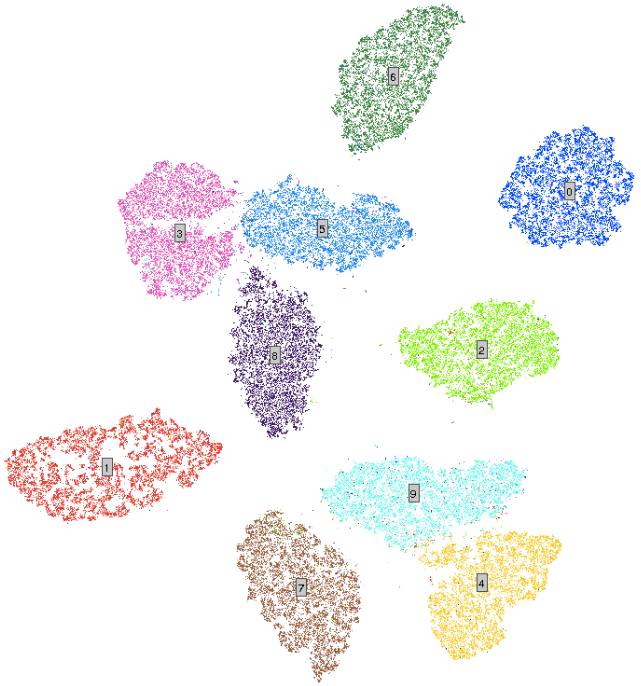
Each cluster has a cluster center

k-means clustering



K-means objective $\sum_{j=1}^{k} \sum_{i \in S_j} \|x_i - c_j\|_2^2$

Hand written letters example

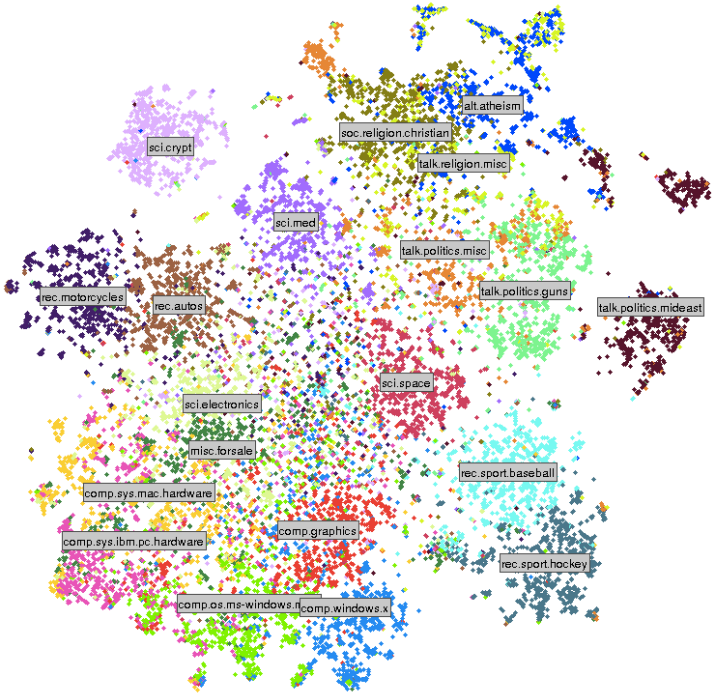


http://en.wikipedia.org/wiki/MNIST_database

http://research.ics.aalto.fi/mi/software/ne/

News groups example

- Roughly 20,000 documents
- 20 topics:
  - Graphics
  - PC hardware
  - Baseball
  - For-sale
  - Politics
  - …

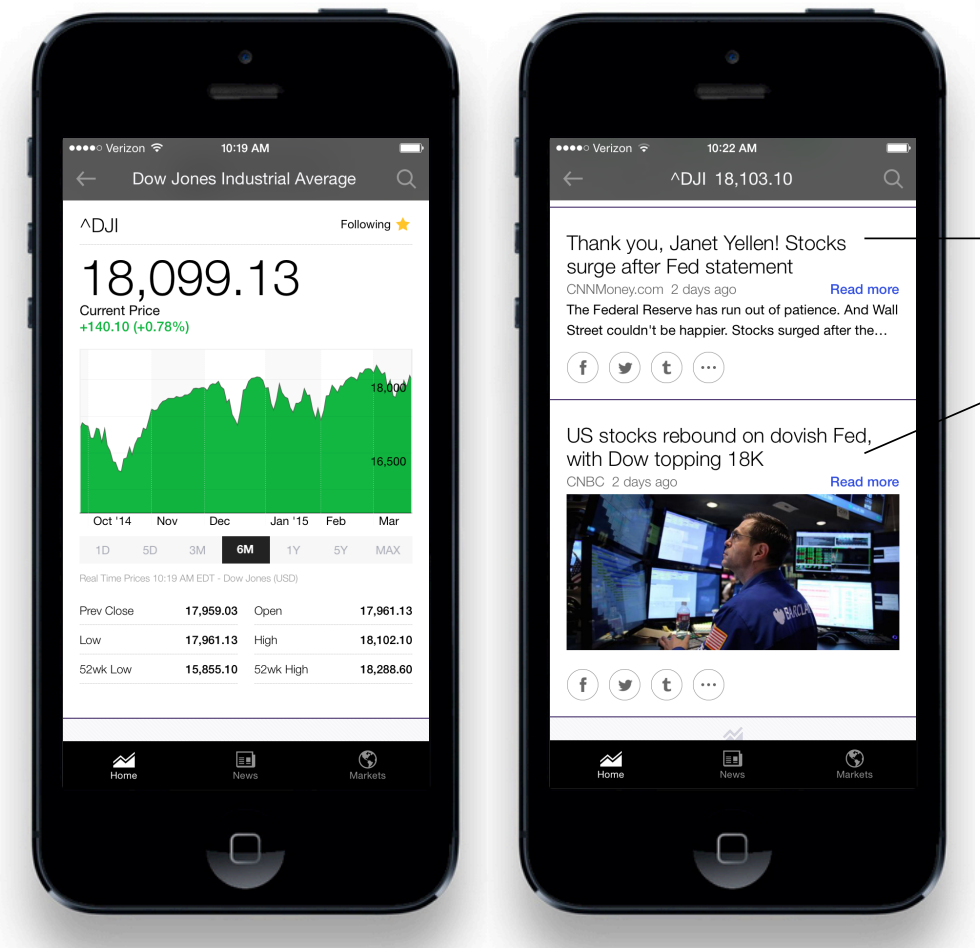


http://qwone.com/~jason/20Newsgroups/

http://research.ics.aalto.fi/mi/software/ne/

YAHOO!

Yahoo Finance App



Same story line or not?

1) The answer depends on the future
2) We have to decide now…

YAHOO!

Online k-means clustering

1) One can clustering points (documents) fully online
2) Create only slightly more than k centers (story lines)
3) Be competitive with the best clustering to k clusters

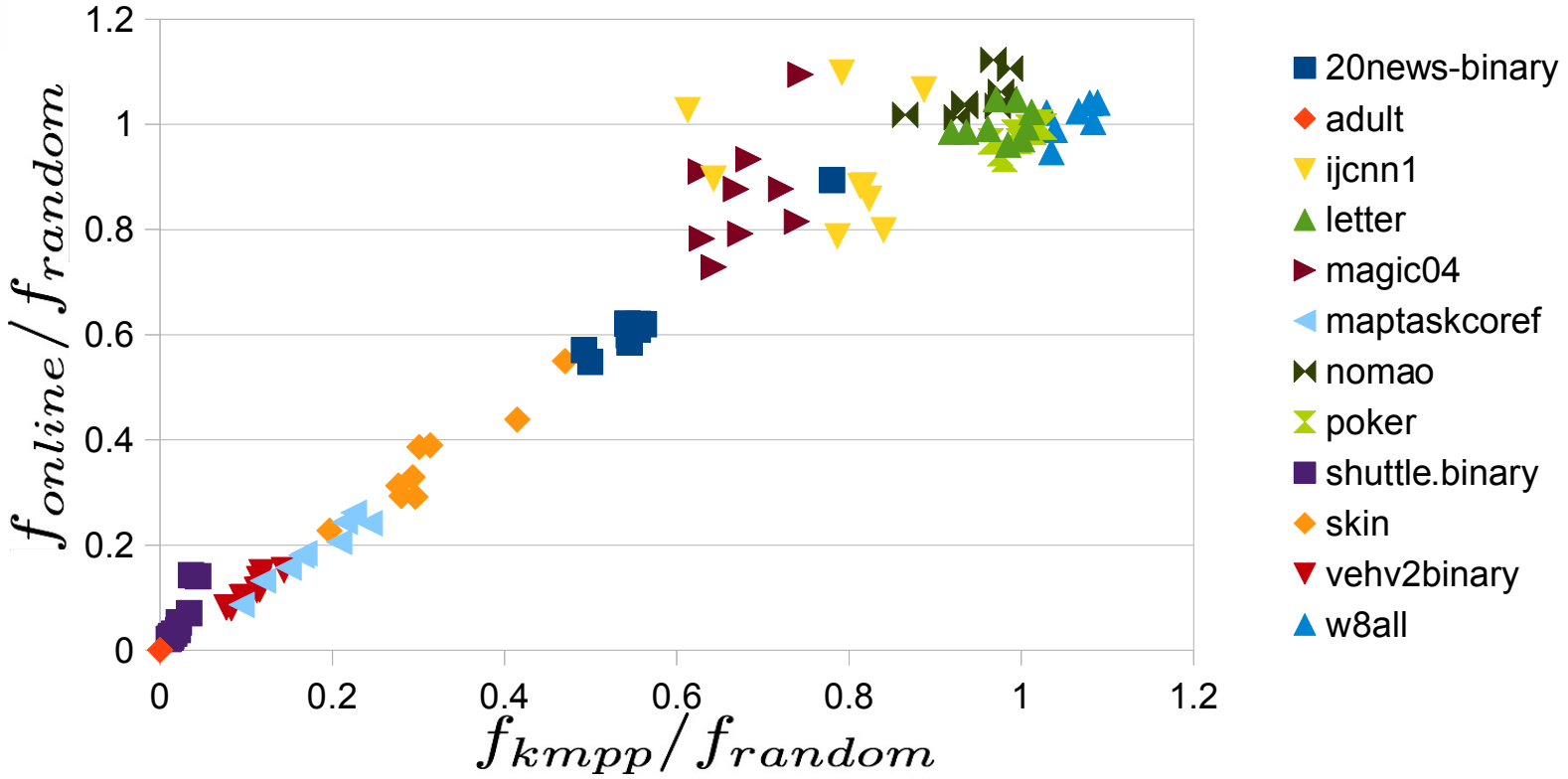
Algorithm 2 Online k-means algorithm

input: V, k
$C \leftarrow$ first $k+1$ distinct vectors in V; and $n = k+1$
(For each of these **yield** itself as its center)
$w^* \leftarrow \min_{v,v' \in C} \|v - v'\|^2/2$
$r \leftarrow 1$; $q_1 \leftarrow 0$; $f_1 = w^*/k$
for $v \in$ the remainder of V **do**
  $n \leftarrow n+1$
  with probability $p = \min(D^2(v,C)/f_r, 1)$
    $C \leftarrow C \cup \{v\}$; $q_r \leftarrow q_r + 1$
  if $q_r \geq 3k(1+\log(n))$ **then**
    $r \leftarrow r+1$; $q_r \leftarrow 0$; $f_r \leftarrow 2 \cdot f_{r-1}$
  end if
  yield: $c = \arg\min_{c \in C} \|v - c\|^2$
end for

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015

Online k-means clustering

Clustering online is competitive with batch k-means++



An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015
k-means++: the advantages of careful seeding, Arthur, Vassilvitskii, 2006

YAHOO!

Online k-means clustering

Clustering online allows for feature engineering for online machine learning!

dataset	# records	dimension	Before	After	Lift
letter	20000	15	0.7581	0.7653	0.94%
shuttle	43500	8	0.9247	0.9950	7.60%
skin	245057	2	0.9247	0.9957	7.67%
poker	946799	9	0.5436	0.6015	10.65%

An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015

Rule of thumb:

Sequential decision making in big data should be both **streaming** and **online**.

Thank you