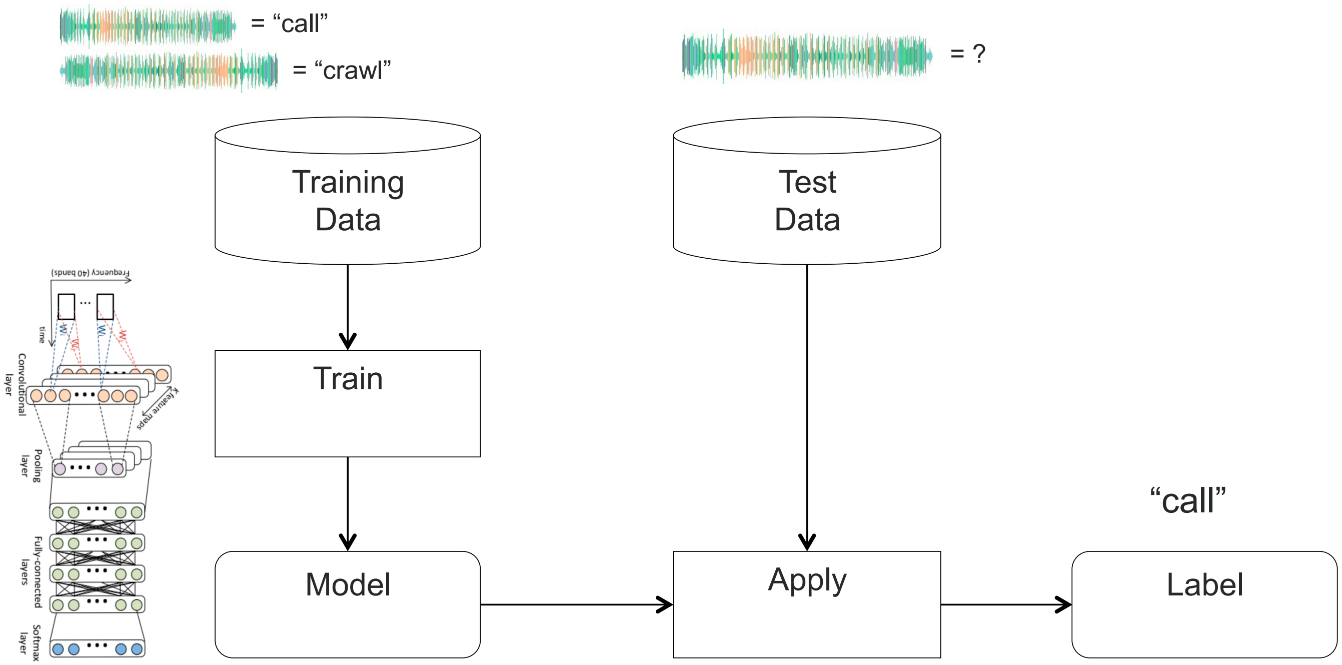
ONLINE MACHINE LEARNING AND DATA MINING

EDO LIBERTY

YAHOO!

STANDARD MACHINE LEARNING SETTING



= "call"

= "crawl"

= ?

Training Data

Test Data

Train

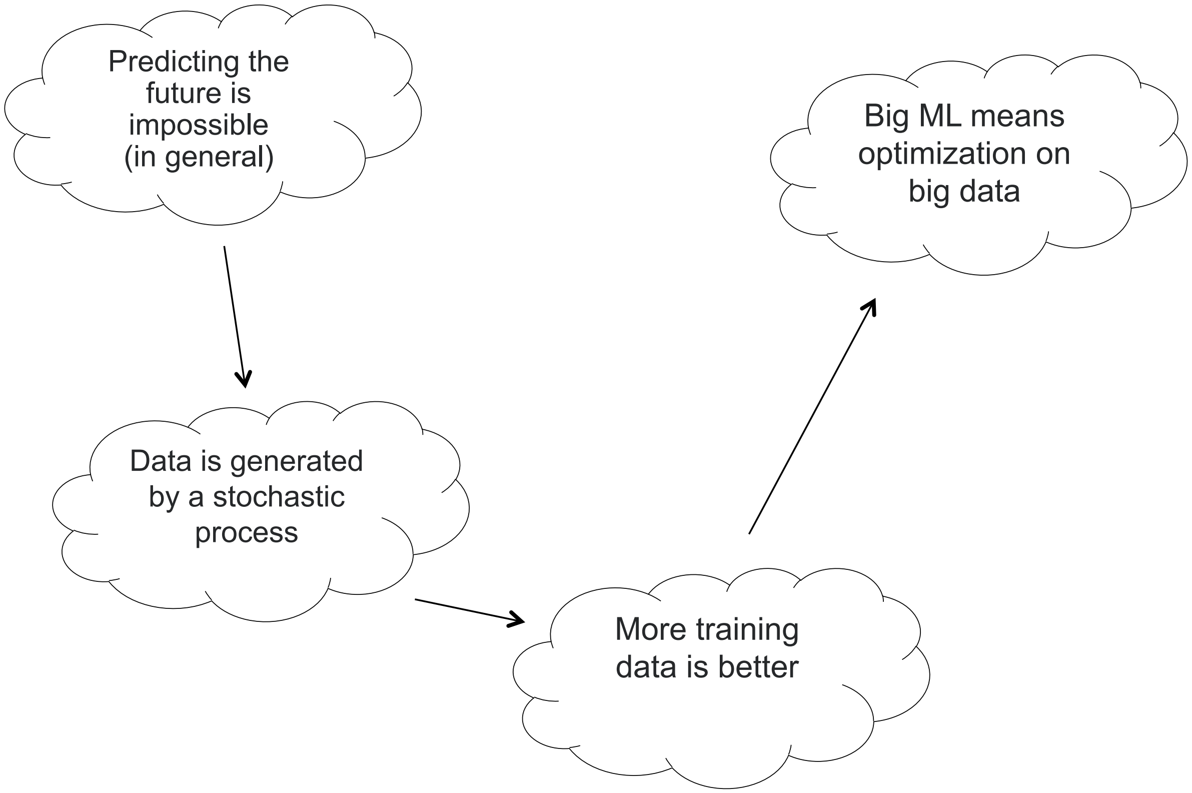
Model

Apply

"call"

Label

STANDARD MACHINE LEARNING SETTING



Predicting the future is impossible (in general)

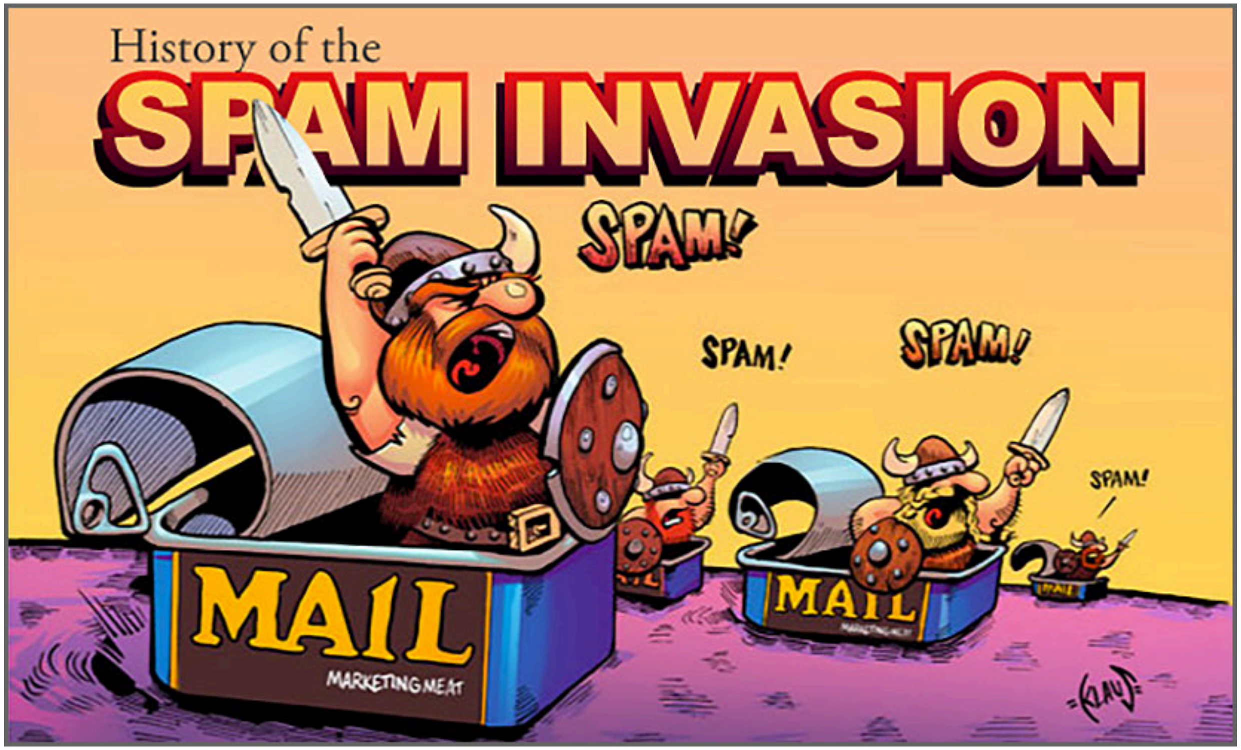
Data is generated by a stochastic process

More training data is better

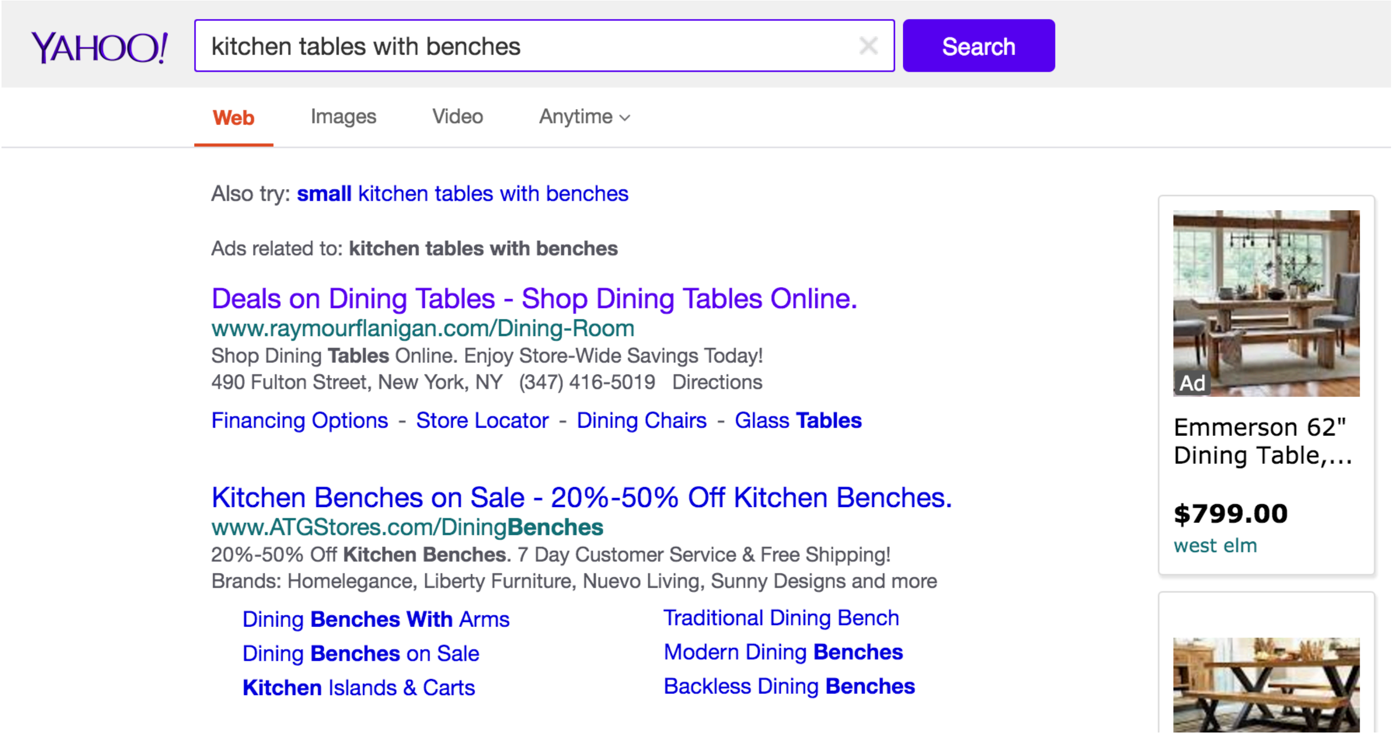
Big ML means optimization on big data

© Alamy

MORE DATA IS OFTEN WORSE (MORE DATA = OLDER DATA)



OUR ACTIONS HEAVILY INFLUENCE THE DATA



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**

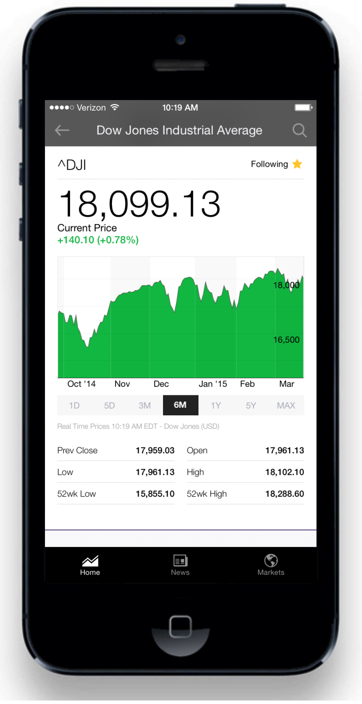
Kitchen Benches on Sale - 20%-50% Off Kitchen Benches.
www.ATGStores.com/Dining**Benches**
20%-50% Off **Kitchen Benches**. 7 Day Customer Service & Free Shipping!
Brands: Homelegance, Liberty Furniture, Nuevo Living, Sunny Designs and more

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

Ad
Emmerson 62" Dining Table,...

$799.00
west elm

THE FUTURE IS OFTEN NOT LIKE THE PAST!



●●●●○ Verizon 10:19 AM

Dow Jones Industrial Average

^DJI Following ★

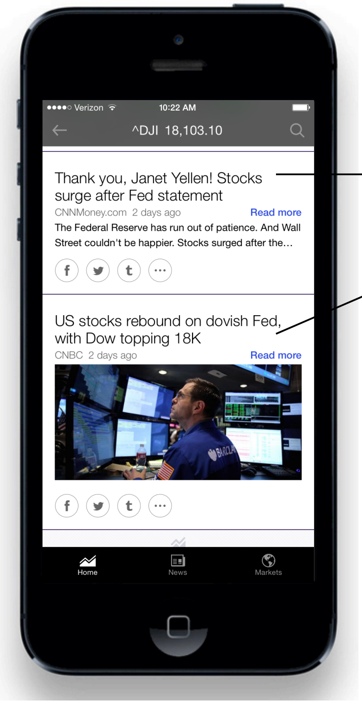
18,099.13

Current Price

+140.10 (+0.78%)

Real Time Prices 10:19 AM EDT - Dow Jones (USD)

Prev Close	17,959.03	Open	17,961.13
Low	17,961.13	High	18,102.10
52wk Low	15,855.10	52wk High	18,288.60



●●●●○ Verizon 10:22 AM

^DJI 18,103.10

Thank you, Janet Yellen! Stocks surge after Fed statement

CNNMoney.com 2 days ago Read more

The Federal Reserve has run out of patience. And Wall Street couldn't be happier. Stocks surged after the…

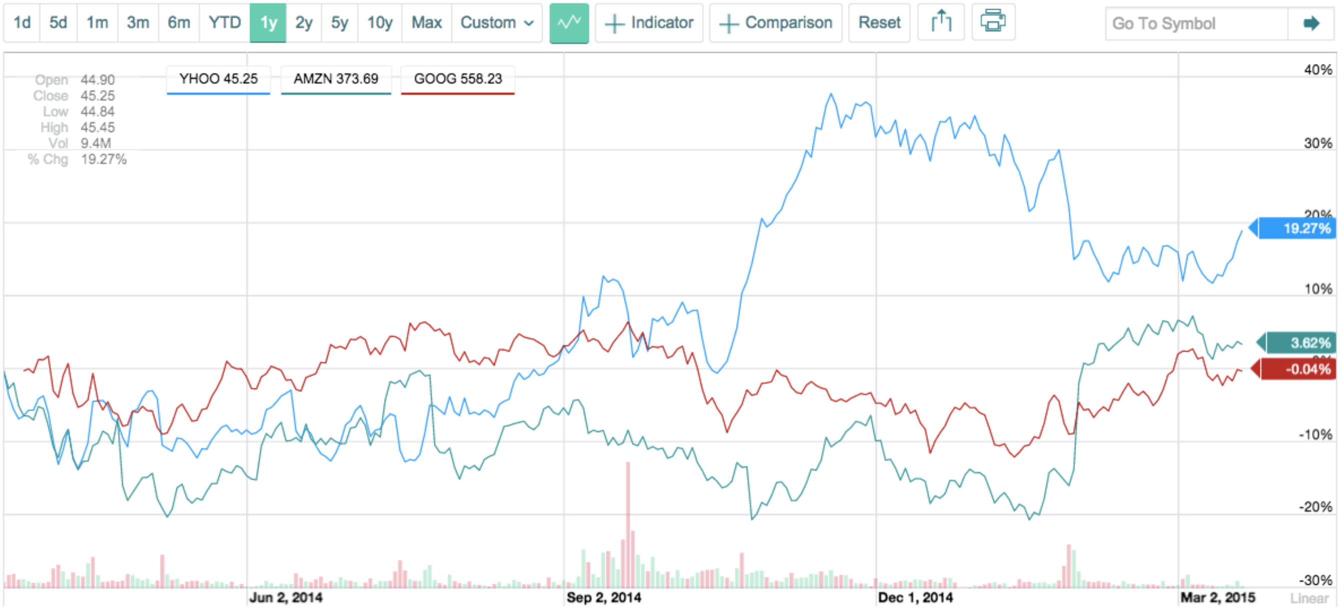
US stocks rebound on dovish Fed, with Dow topping 18K

CNBC 2 days ago Read more

Same story line or not?

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

HAVING “A MODEL” IS COMPLETELY UNIMPORTANT



Open	44.90
Close	45.25
Low	44.84
High	45.45
Vol	9.4M
% Chg	19.27%

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

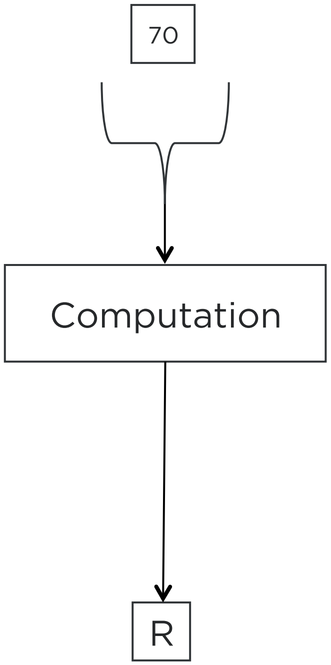
ONLINE ALGORITHMS
(DECISION MAKING WITHOUT PREDICTING)

THE SKI RENTAL PROBLEM

Rent: x$ /day

Buy: 1000$

THE SKI RENTAL PROBLEM



70

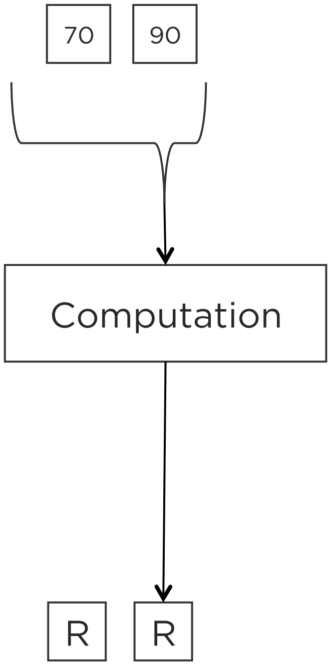
Computation

R

70
+

70

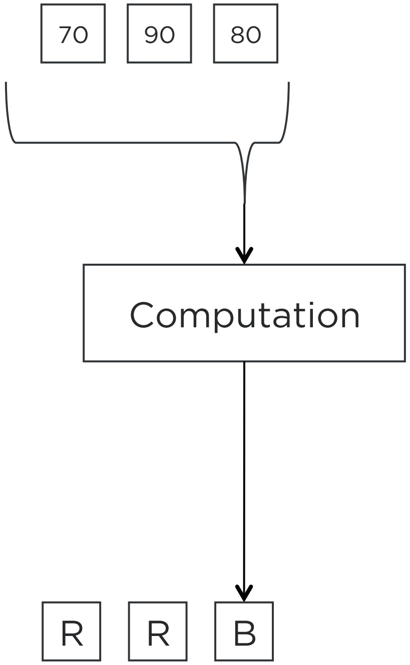
THE SKI RENTAL PROBLEM



70
+
90

160

THE SKI RENTAL PROBLEM



70 90 80

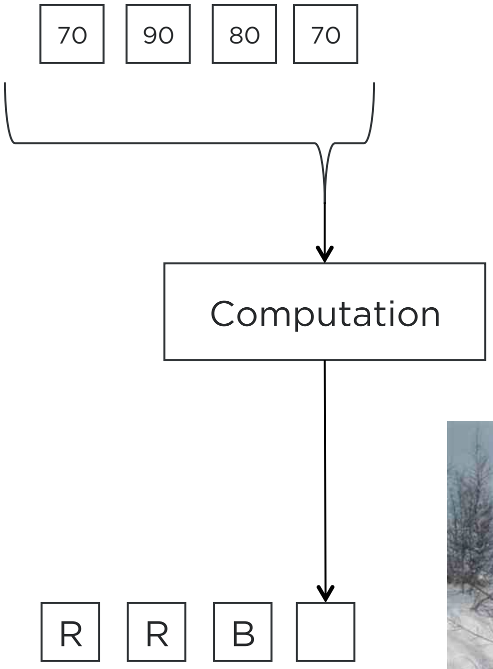
Computation

R R B

70
+
90
+
1000

1160

THE SKI RENTAL PROBLEM



70 90 80 70

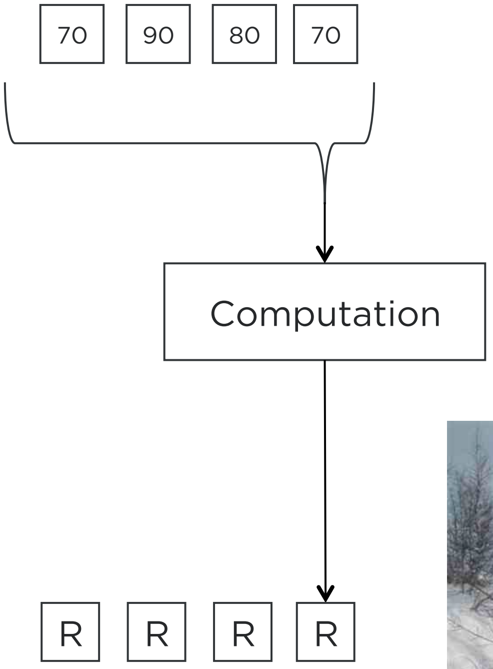
Computation

R R B

70
+
90
+
1000
+
0

1160

THE SKI RENTAL PROBLEM



70 90 80 70

Computation

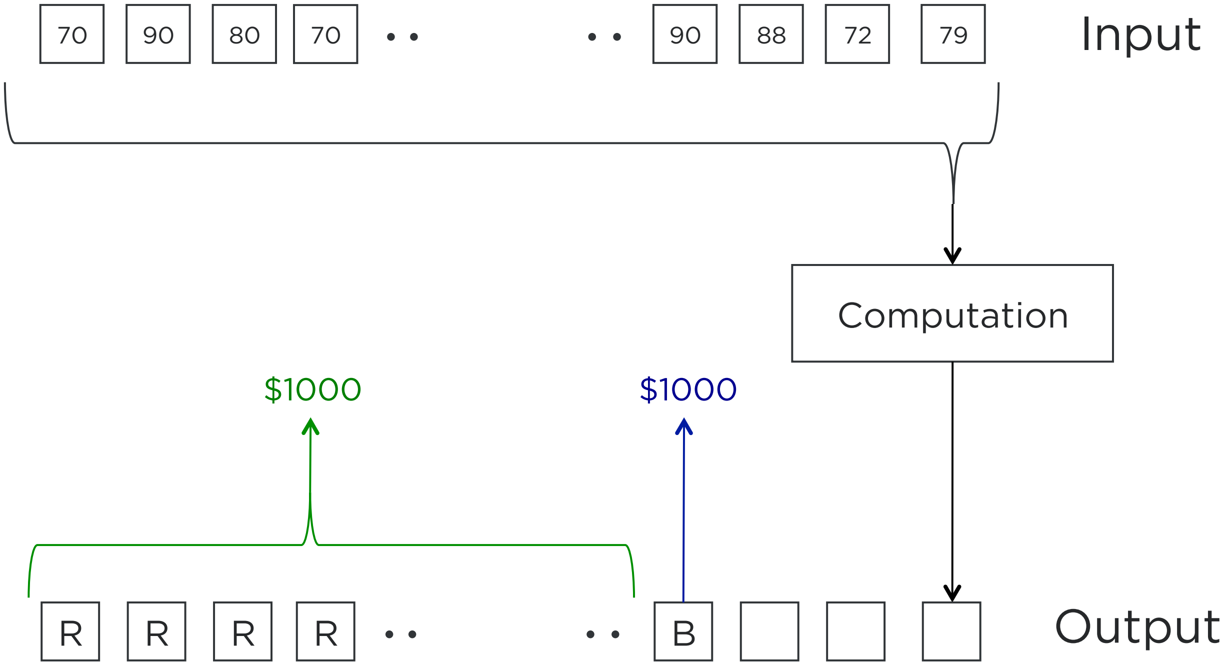
R R R R

70
+
90
+
80
+
70

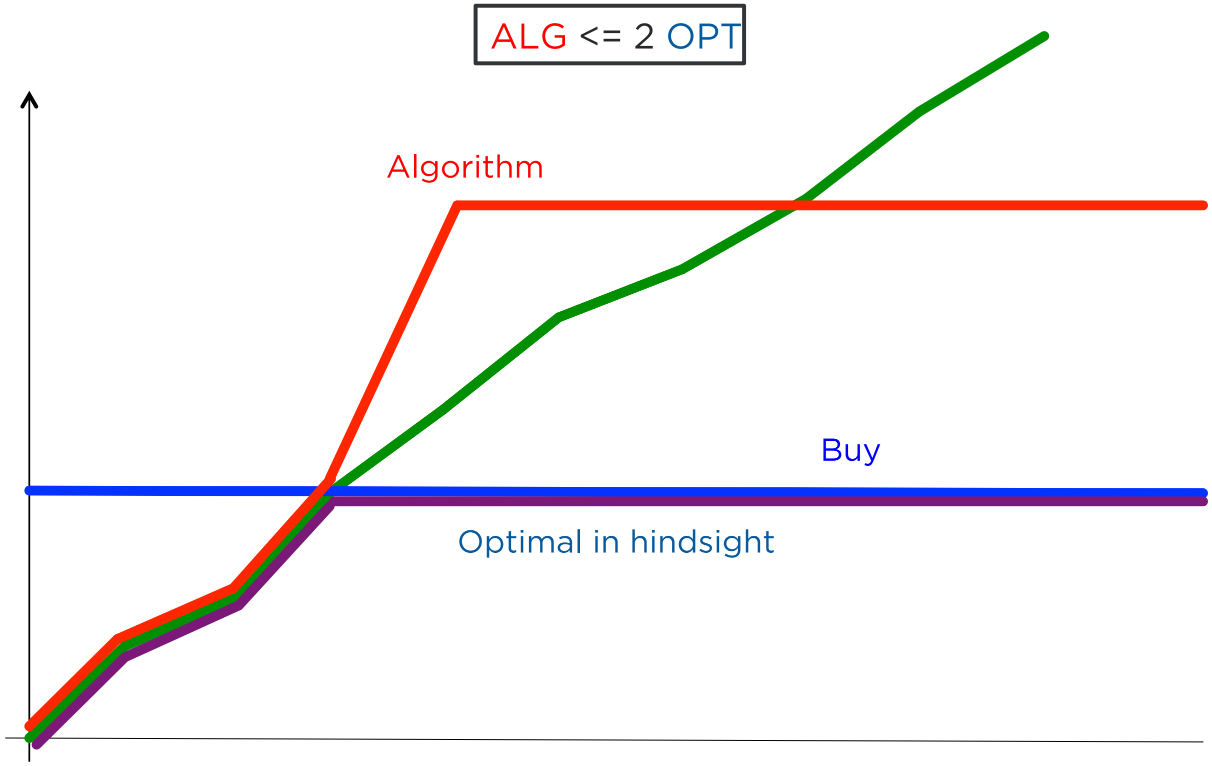
310

You should have rented all along...

THE SKI RENTAL PROBLEM

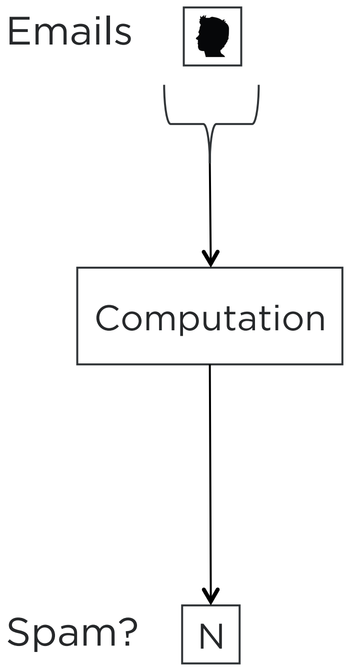


THE SKI RENTAL PROBLEM



ONLINE LINEAR CLASSIFICATION

ONLINE MACHINE LEARNING

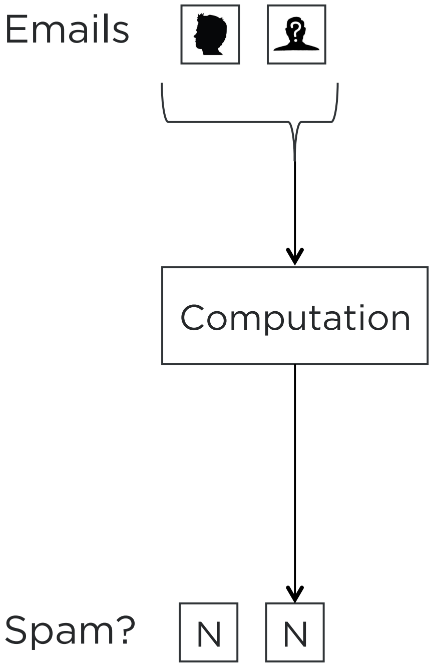


Emails

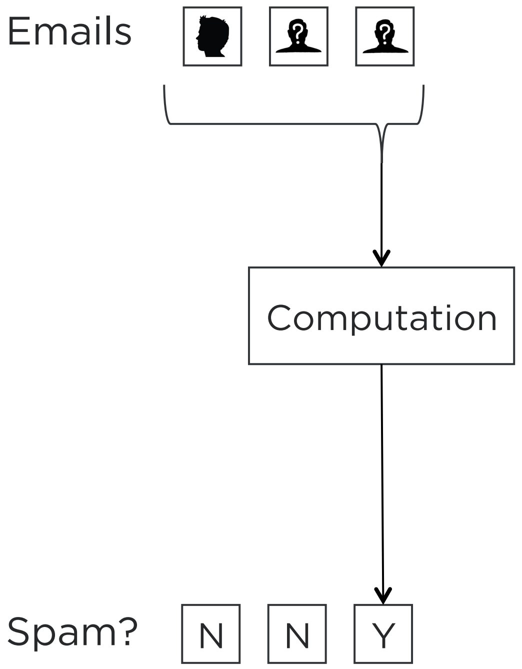
Computation

Spam?

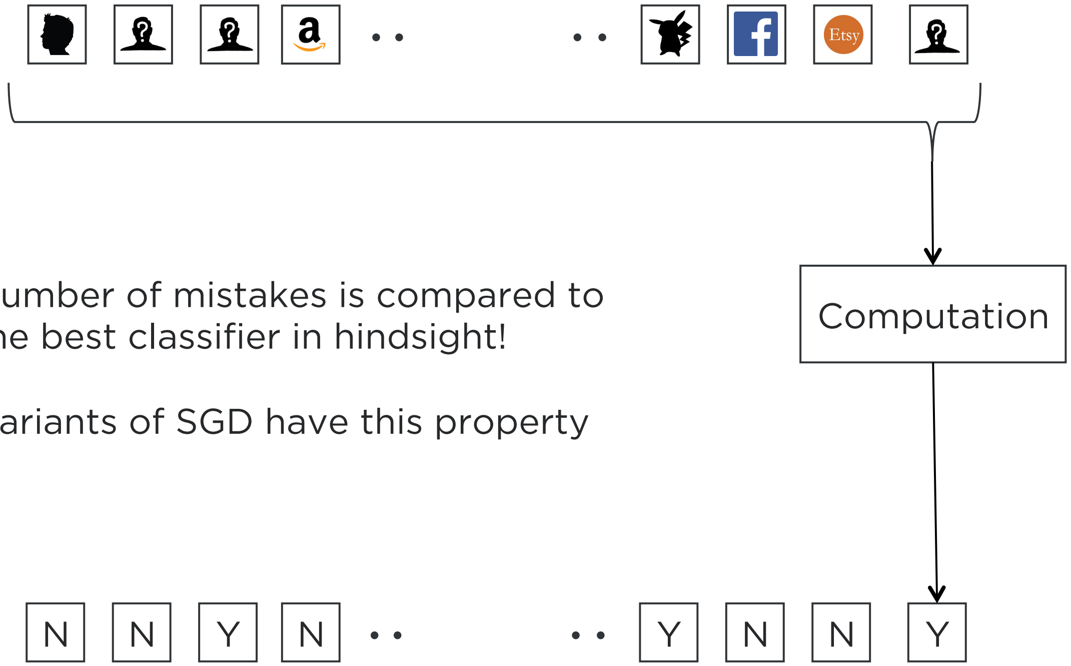
ONLINE MACHINE LEARNING



ONLINE MACHINE LEARNING



ONLINE MACHINE LEARNING



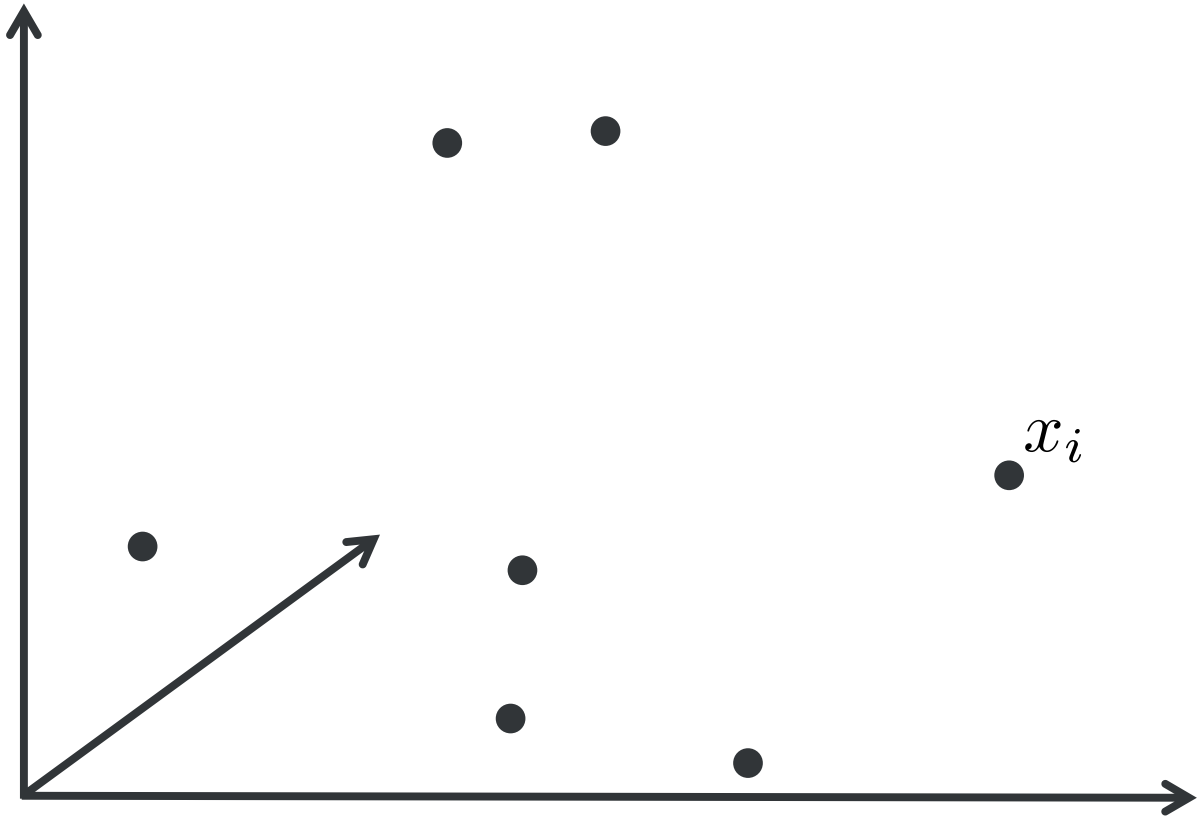
Number of mistakes is compared to the best classifier in hindsight!

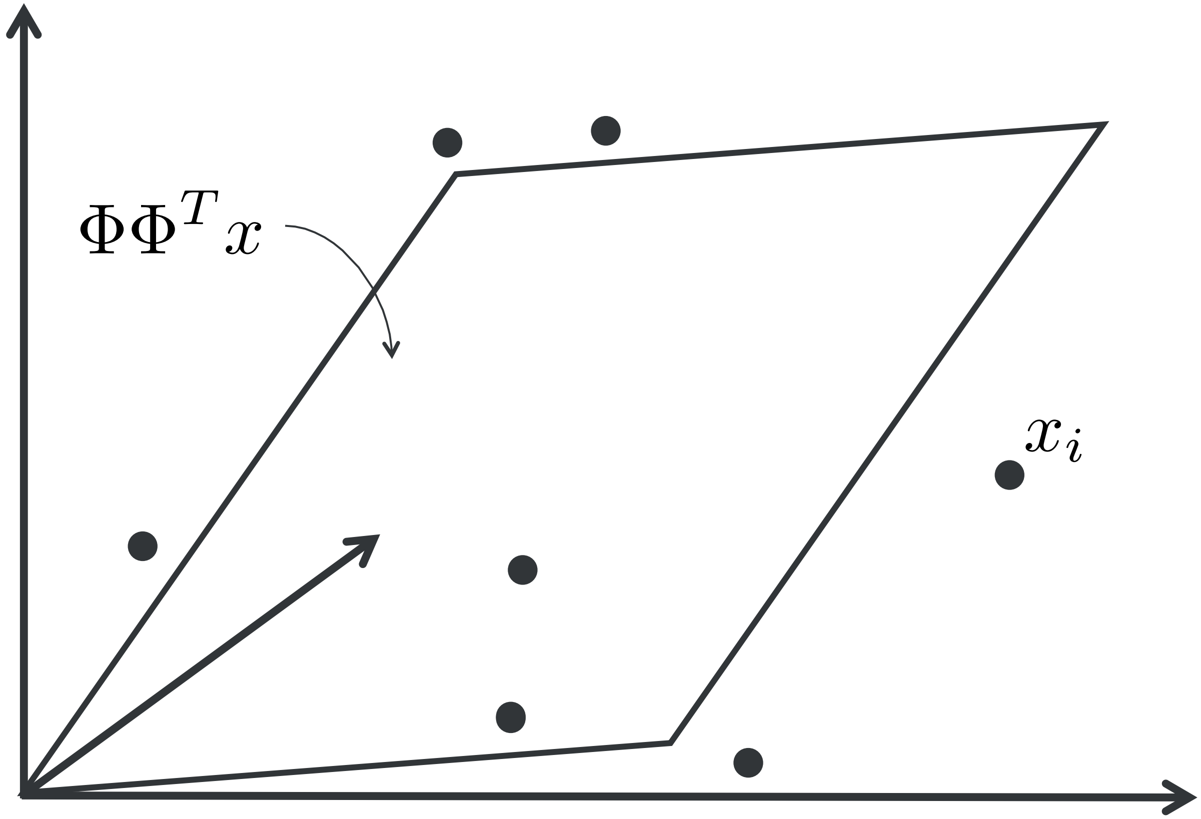
Variants of SGD have this property

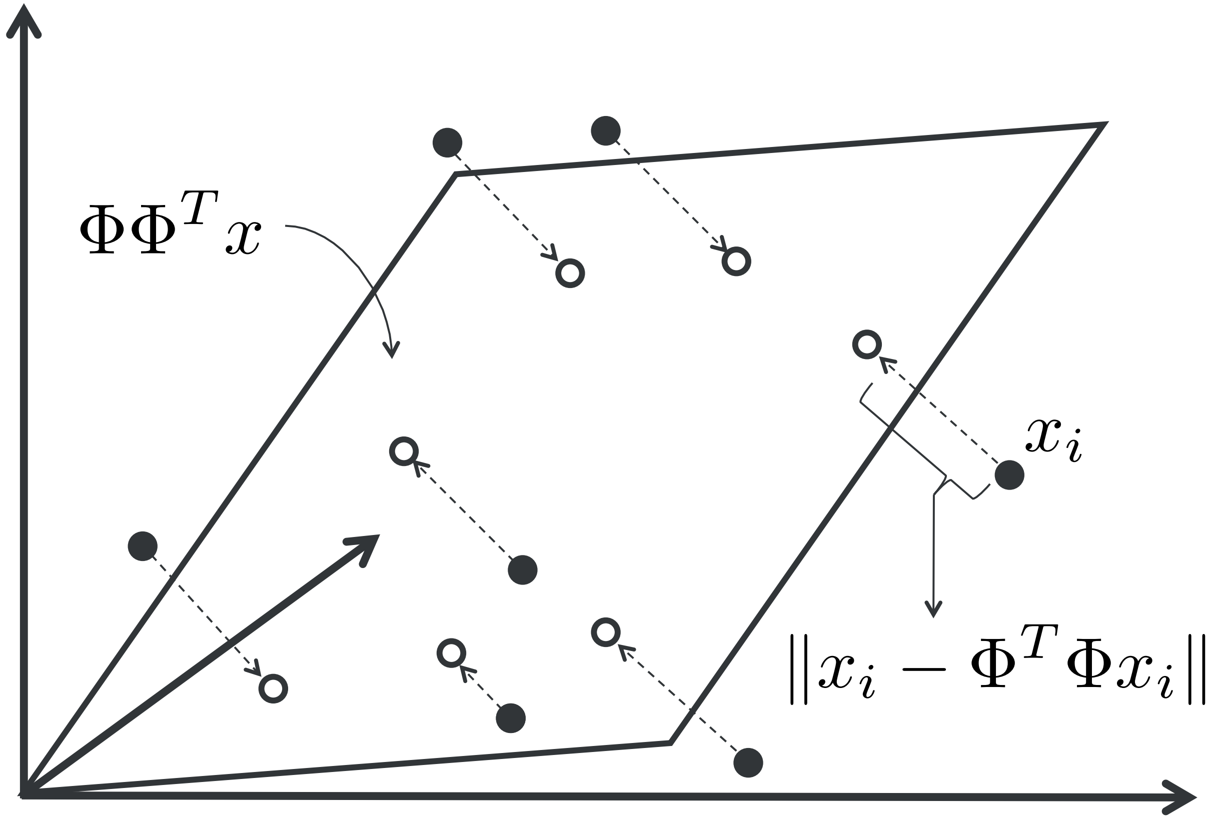
Prediction, Learning, and Games, Cesa-Bianchi, Lugosi, 2006

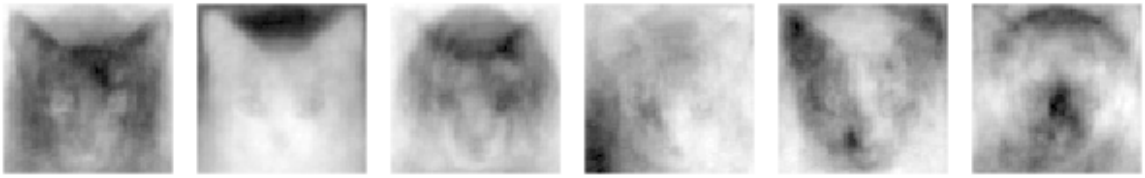
ONLINE PRINCIPAL COMPONENT ANALYSIS

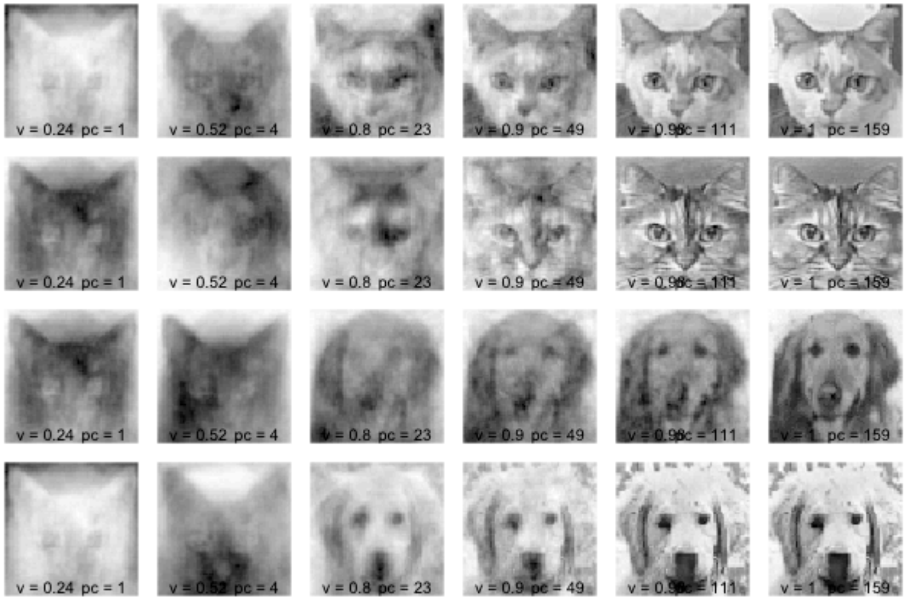
Online Principal Components Analysis, Boutsidis, Garber, Karnin, Liberty 2014
Online PCA with Spectral Bounds, Karnin, Liberty, 2015


x_i


$\Phi\Phi^T x$
x_i

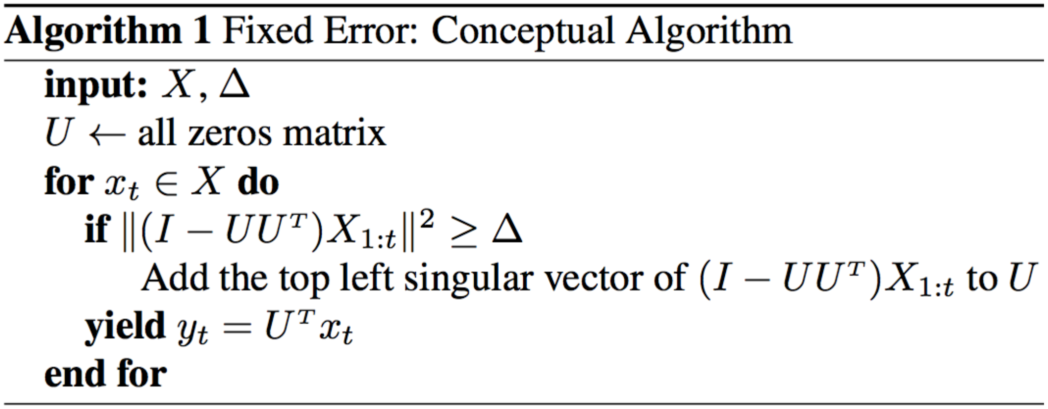

$\Phi\Phi^T x$
x_i
$\|x_i - \Phi^T\Phi x_i\|$





Eigenpets: https://bioramble.wordpress.com/2015/09/01/

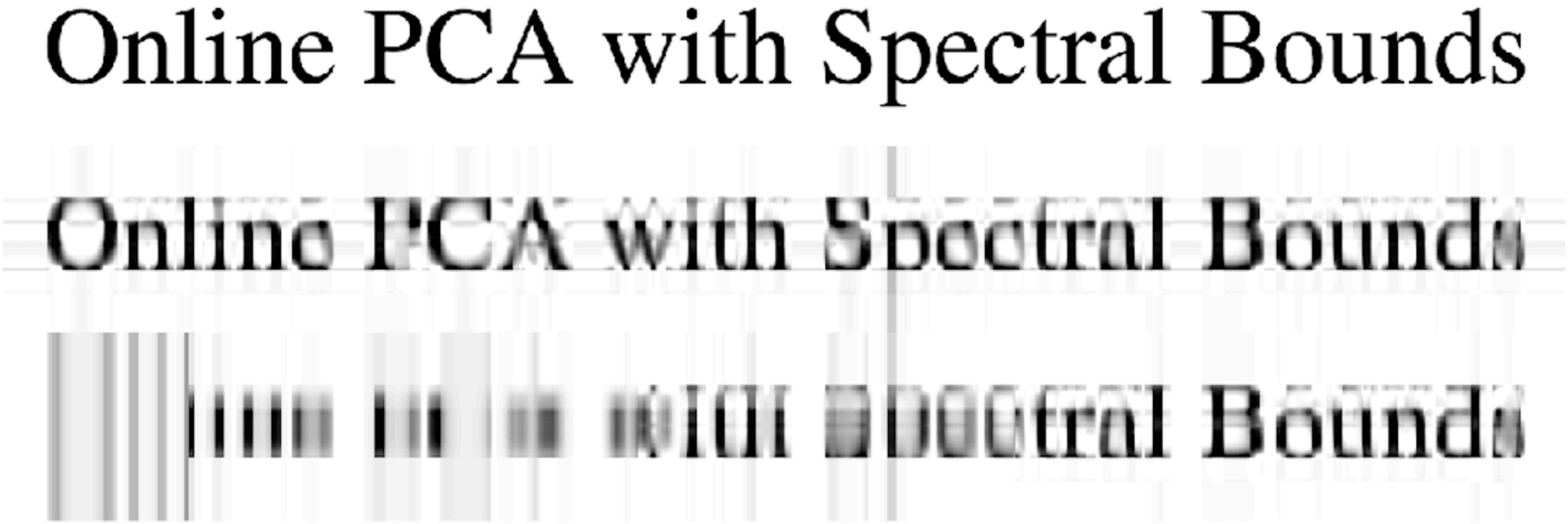
ONLINE PRINCIPAL COMPONENT ANALYSIS



Algorithm 1 Fixed Error: Conceptual Algorithm

```
input: $X, \Delta$
$U \leftarrow$ 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
```

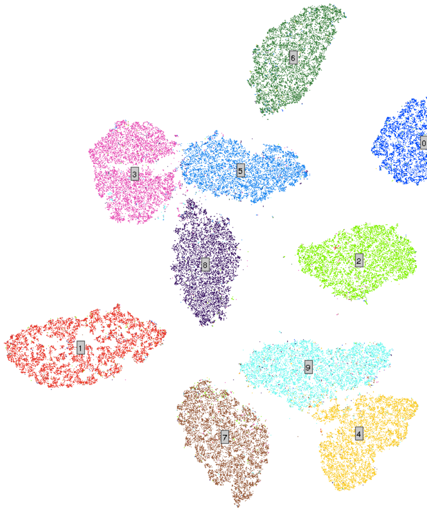
Online PCA with Spectral Bounds, Karnin, Liberty, 2015


Online PCA with Spectral Bounds
Online PCA with Spectral Bounds
Spectral Bounds

ONLINE K-MEANS CLUSTERING

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

K-MEANS CLUSTERING

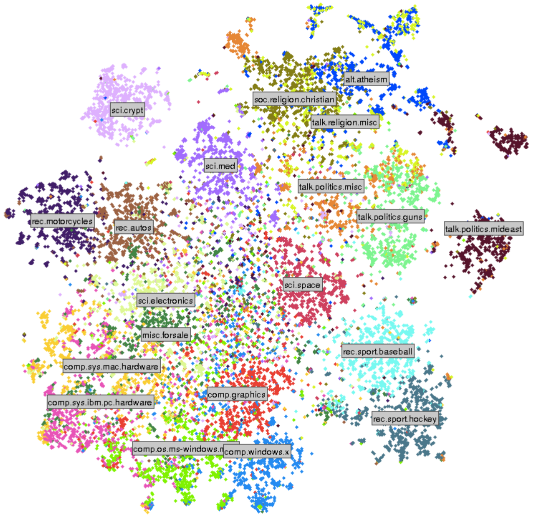


http://en.wikipedia.org/wiki/MNIST_database

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

K-MEANS CLUSTERING

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



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

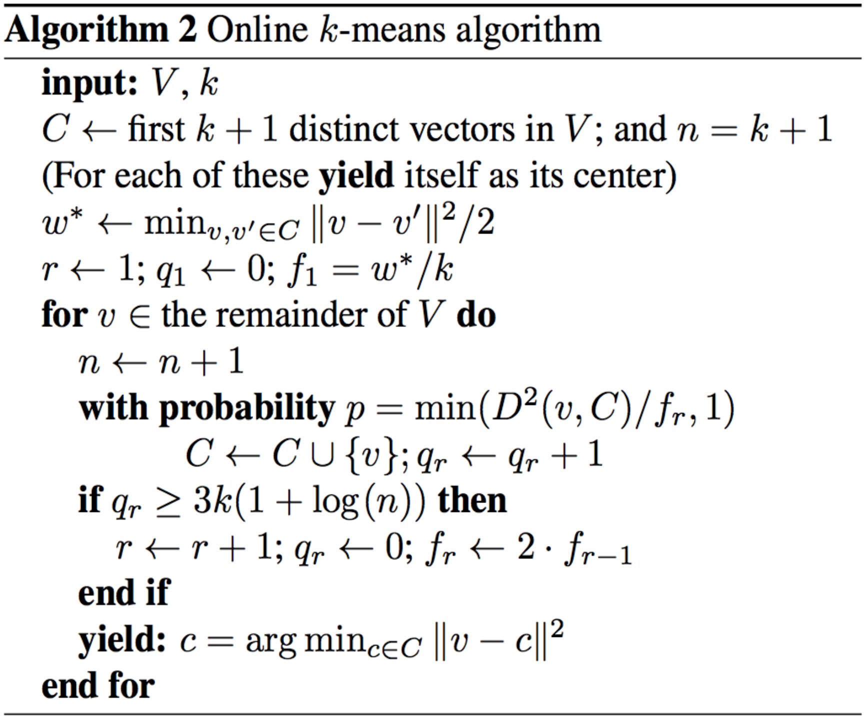
http://research.ics.aalto.fi/mi/software/ne/

K-MEANS CLUSTERING

1) One can cluster points fully online

2) Create only slightly more than k centers

3) Be competitive with the best offline clustering to k clusters

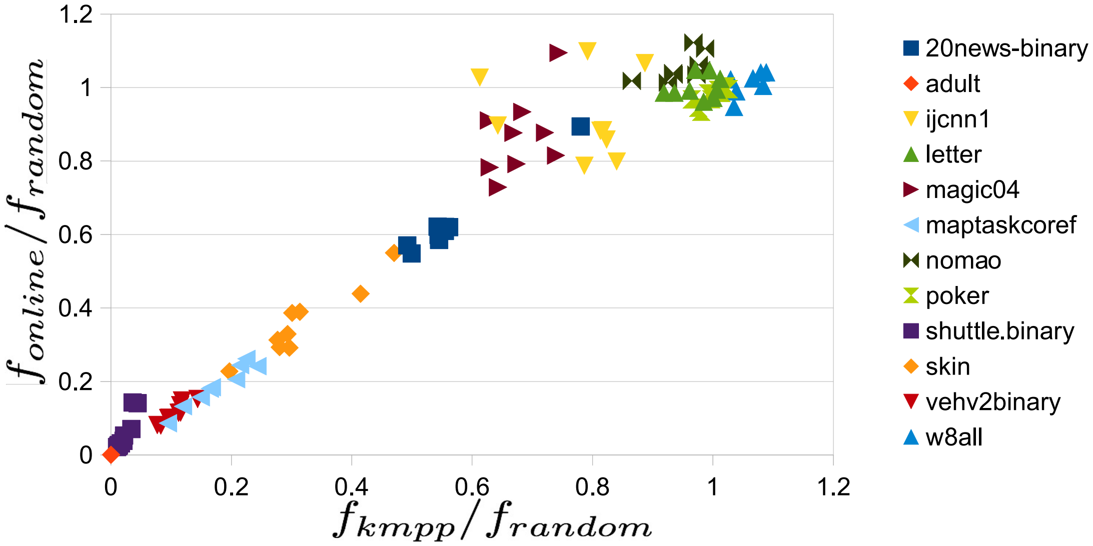


Algorithm 2 Online k-means algorithm

```
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



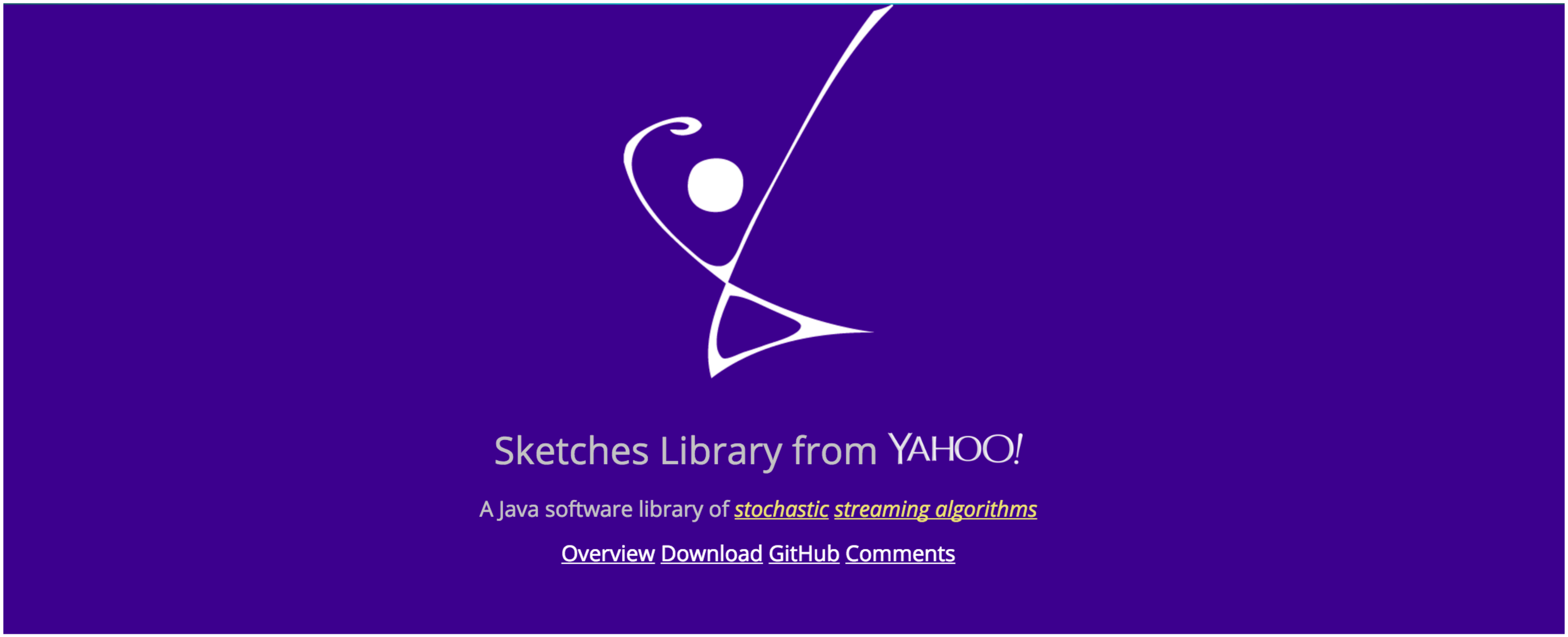
An Algorithm for Online K-Means Clustering, Liberty, Sriharsha, Sviridenko 2015
k-means++: the advantages of careful seeding, Arthur, Vassilvitskii, 2006

STREAMING ALGORITHMS

OPEN SOURCE FROM YAHOO

EDO LIBERTY

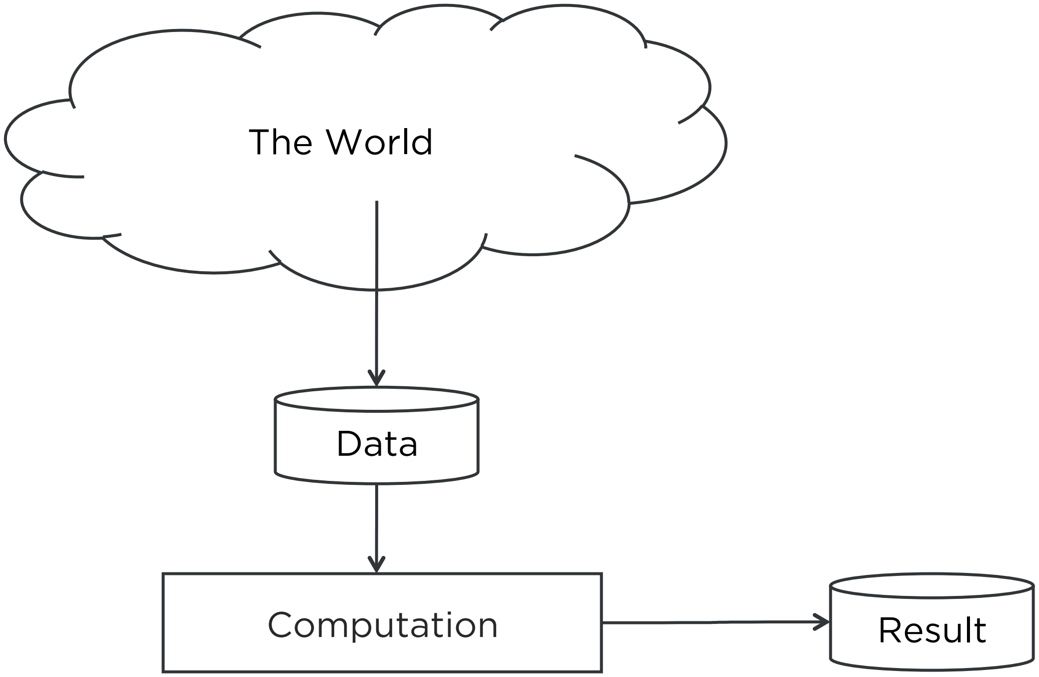
DATASKETCHES.GITHUB.IO



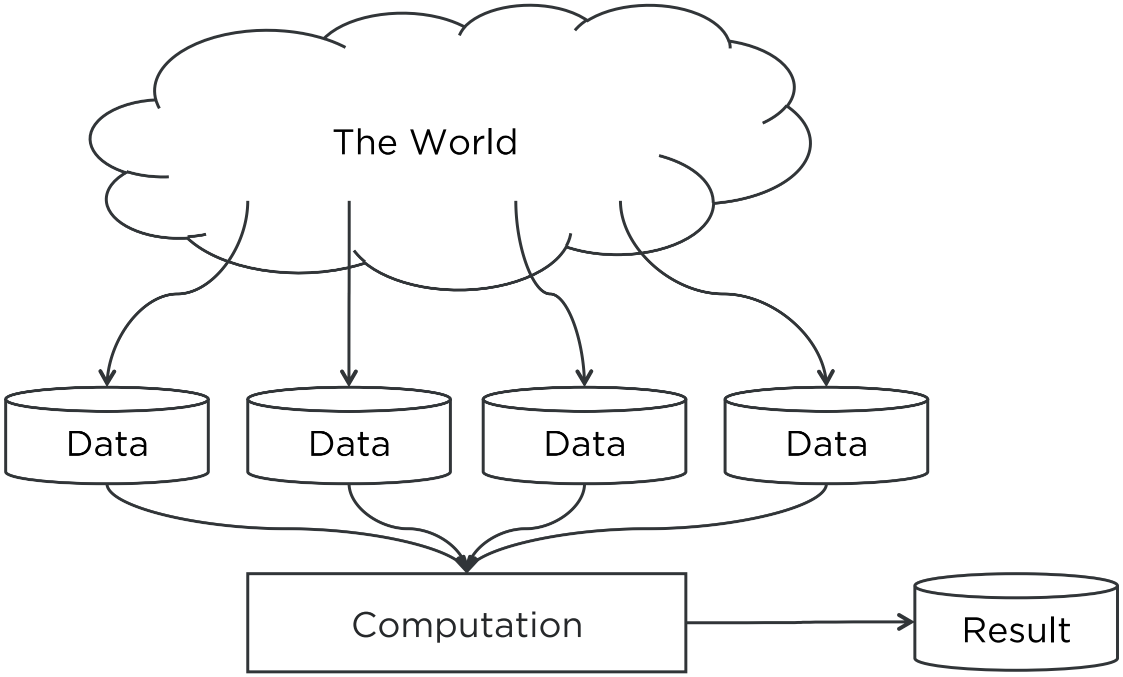
Sketches Library from YAHOO!

A Java software library of *stochastic* *streaming algorithms*

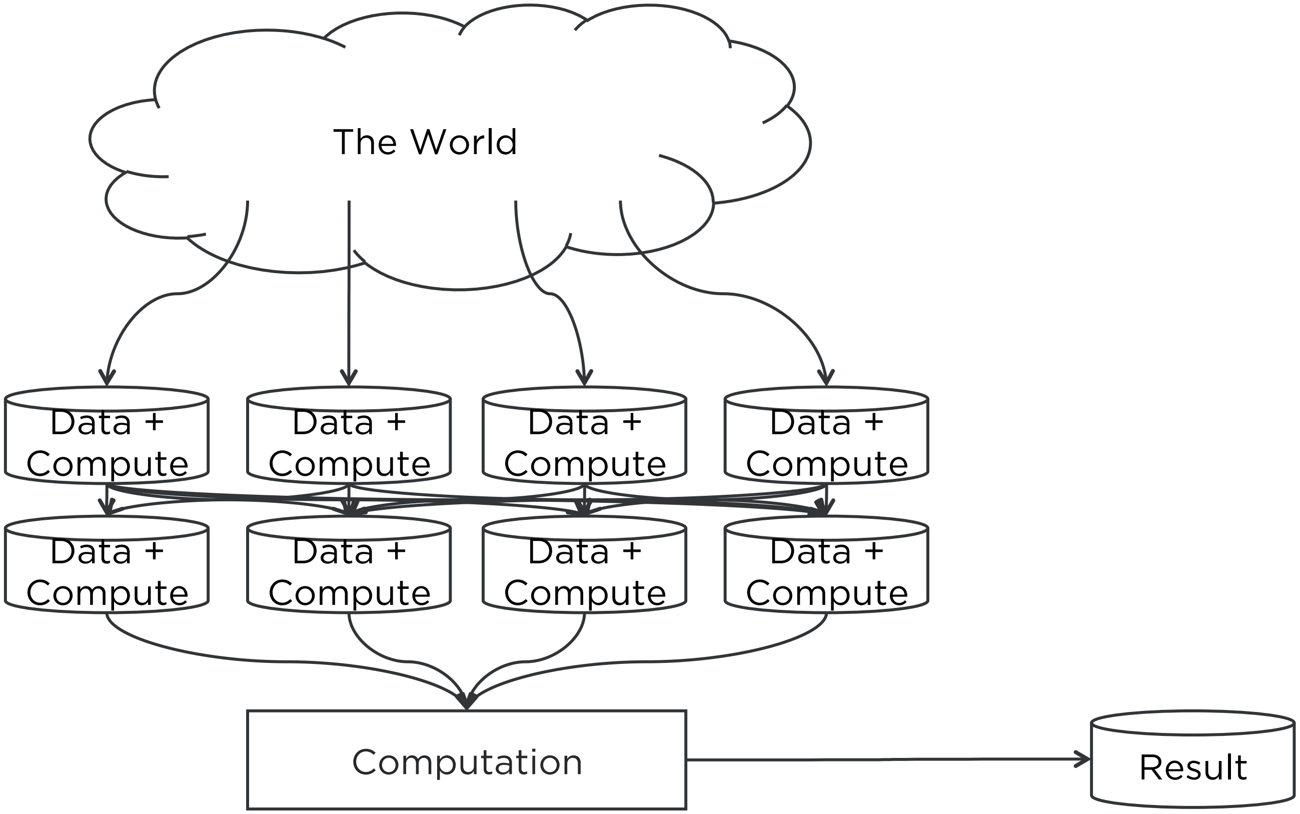
Overview Download GitHub Comments


The World
Data
Computation
Result

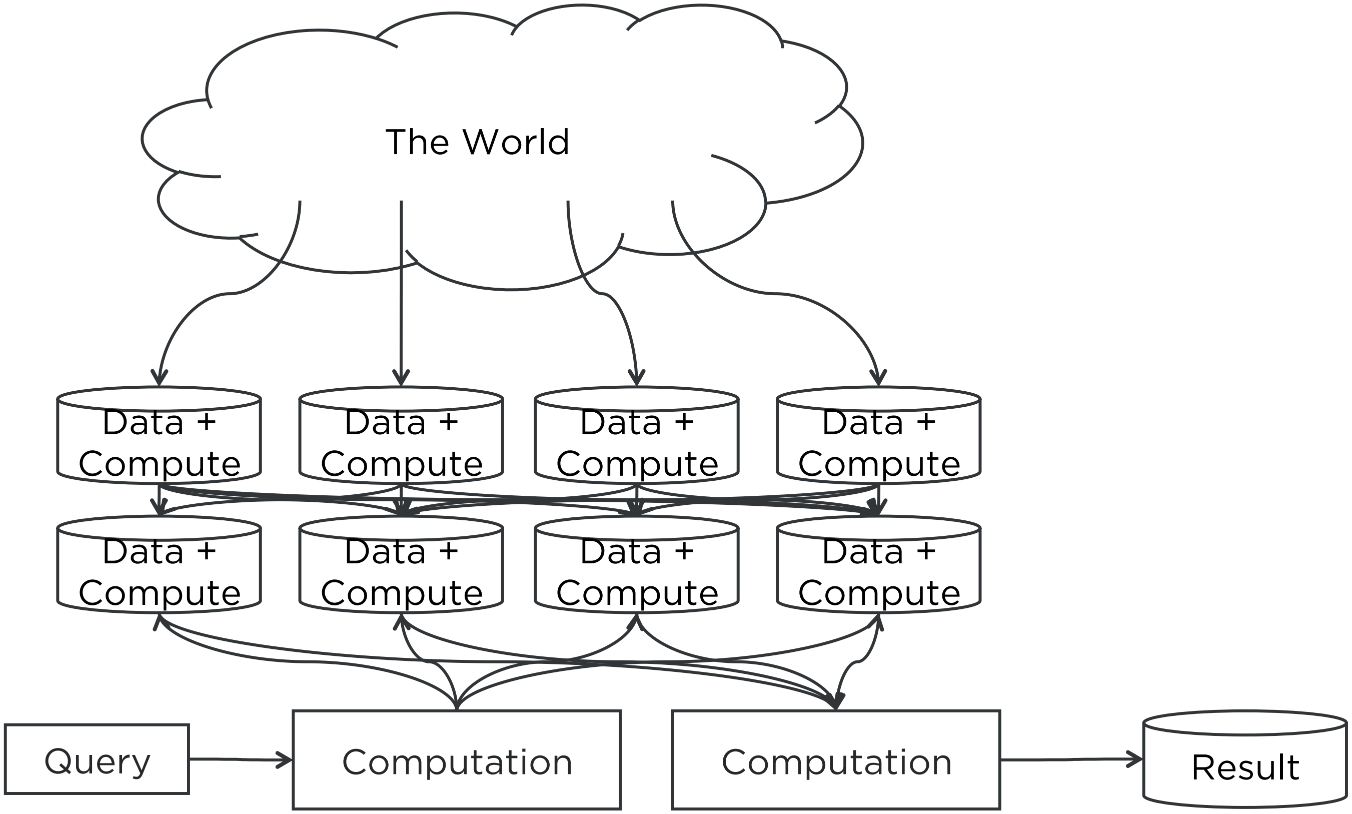
DISTRIBUTED STORAGE



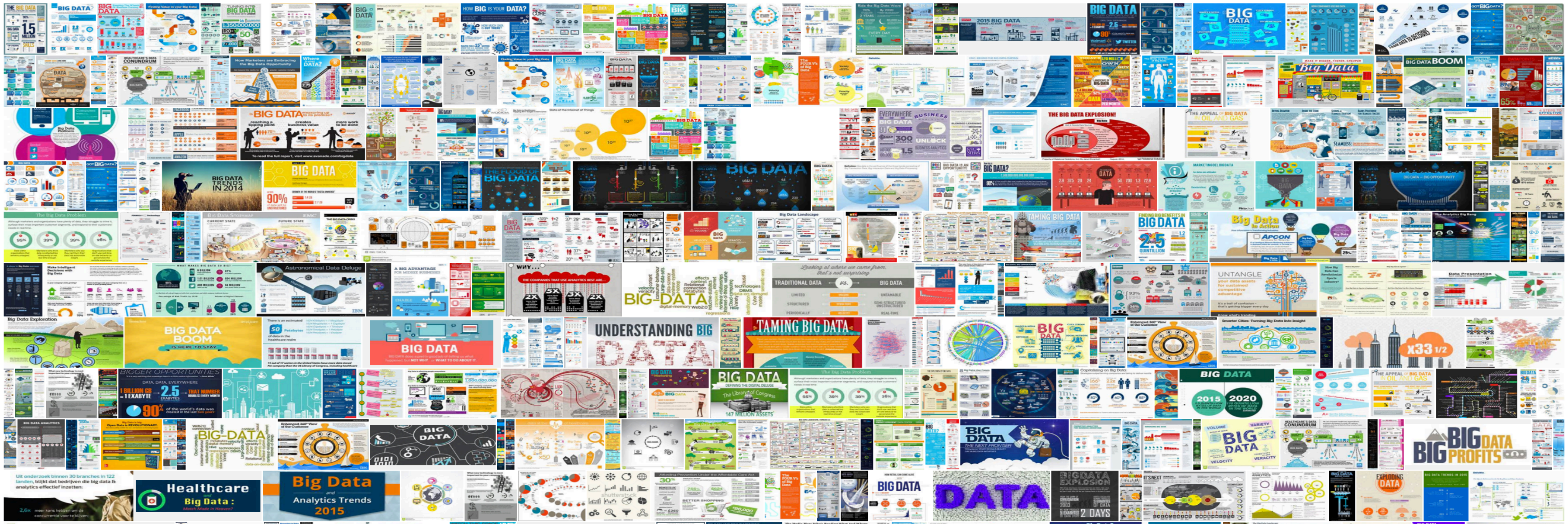
DISTRIBUTED MODEL (MAP/REDUCE, MESSAGE PASSING, ...)



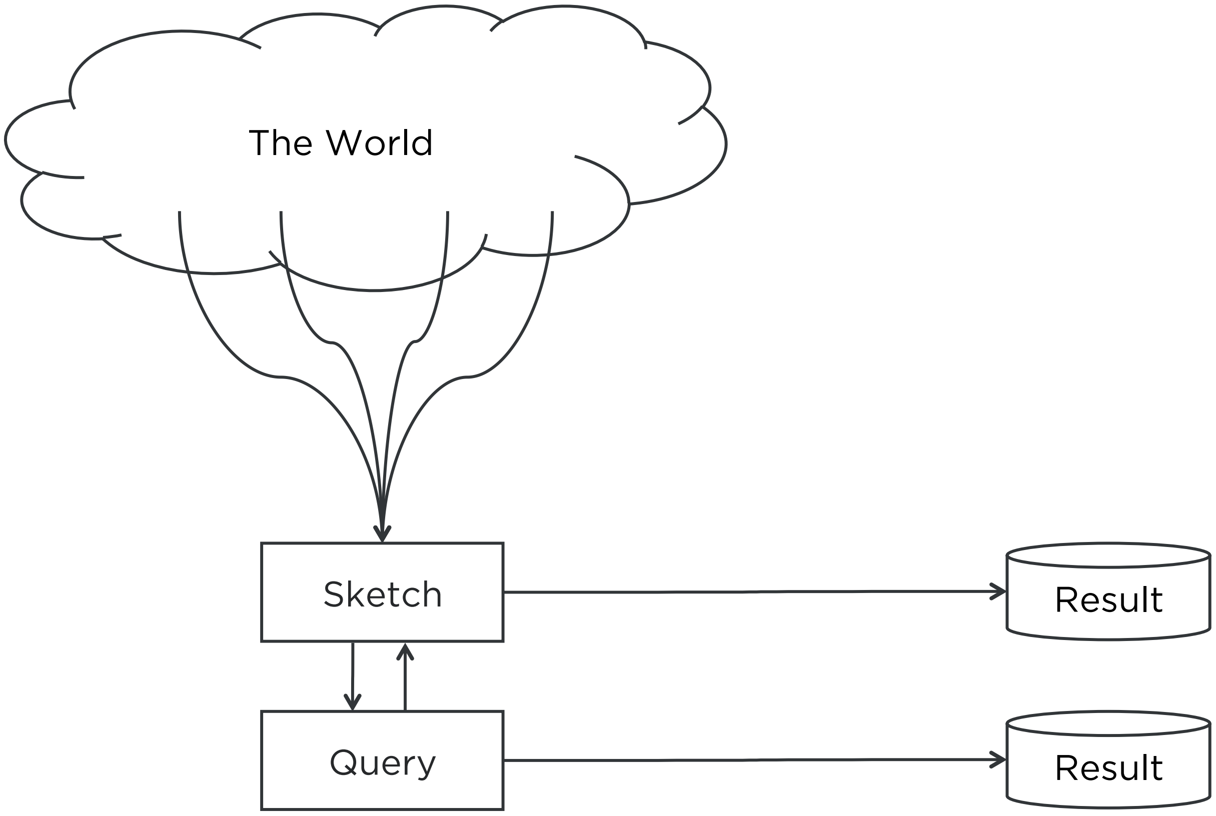
DISTRIBUTED MODEL (INDEXES, TABLES, DATABASES, …)



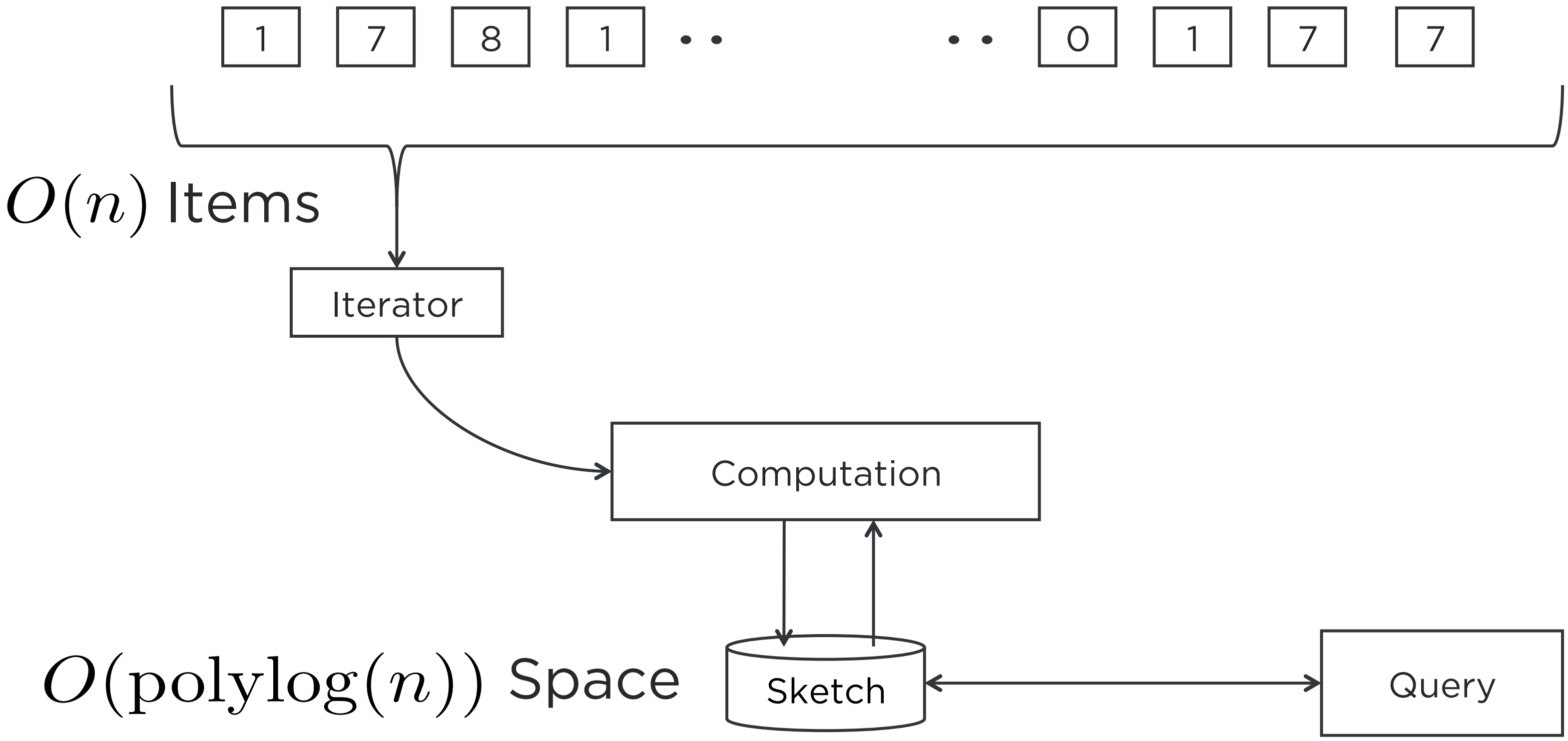
BIG-DATA *META* INFOGRAPHIC



THE STREAMING COMPUTATIONAL MODEL



THE STREAMING COMPUTATIONAL MODEL



1 7 8 1 • • • • 0 1 7 7

$O(n)$ Items

Iterator

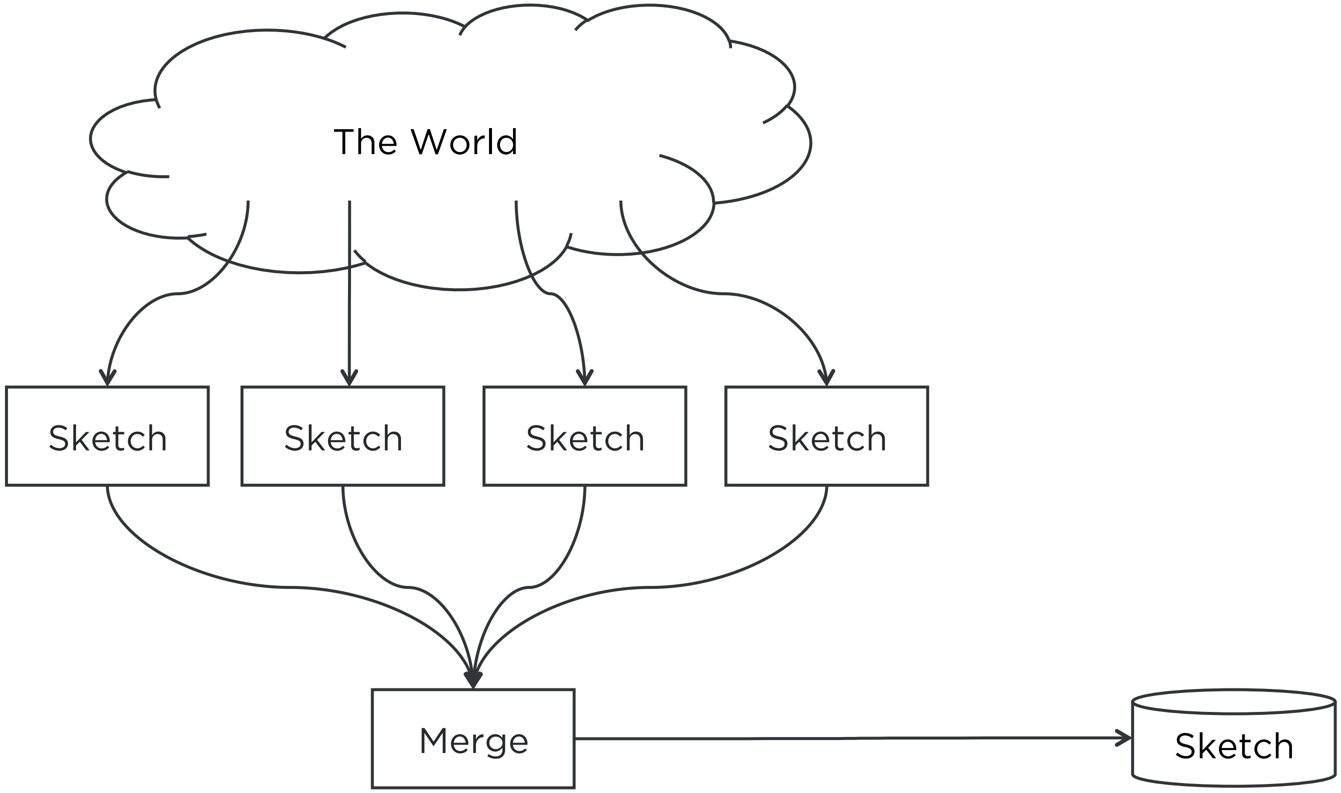
Computation

$O(\text{polylog}(n))$ Space

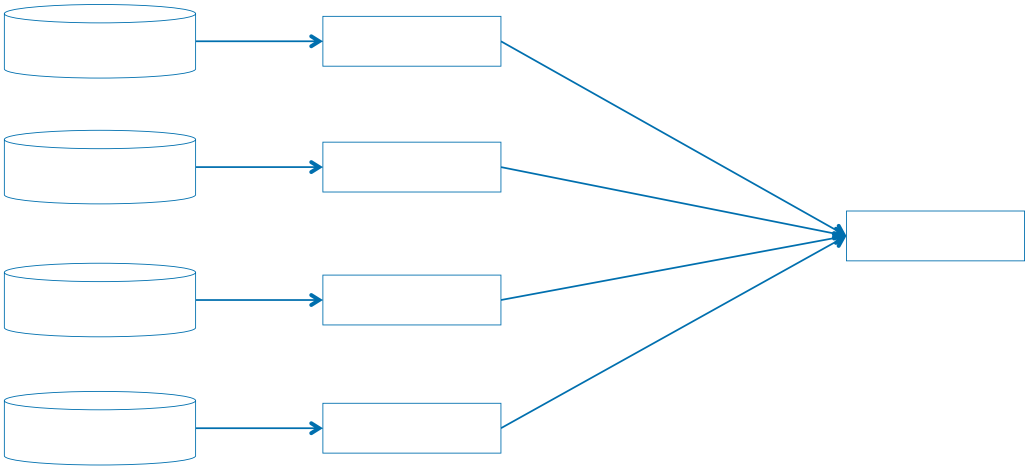
Sketch

Query

THE DISTRIBUTED STREAMING COMPUTATIONAL MODEL



Number of users (easy)



data

Map
(count)

Reduce
(sum)

Web Site Logs

Time	User ID	Site	Time Spent Sec	Items Viewed
9:00	U1	Apps	59	5
9:30	U2	Apps	179	15
10:00	U3	Music	29	3
1:00	U1	Music	89	10
…	…	…	…	…

Financial Transactions System Log

Time	User ID	Site	Purchased	Revenue
9:00	U1	Apps	FaceTune	$3.99
9:30	U2	Apps	Minecraft	$6.99
10:00	U3	Music	Purple Rain	$1.29
10:05	U3	Apps	Minecraft	$6.99
…	…	…	…	…

Unique User Queries

- Unique users viewing Apps since 9:45…?
- Unique users visiting Apps site AND Music site?
- Unique users visiting Apps site AND NOT Music site?

Quantile Queries

- The median and 95%ile Time Spent seconds by …?
- A Frequency Histogram of Time Spent by Split-Points specified at query time?

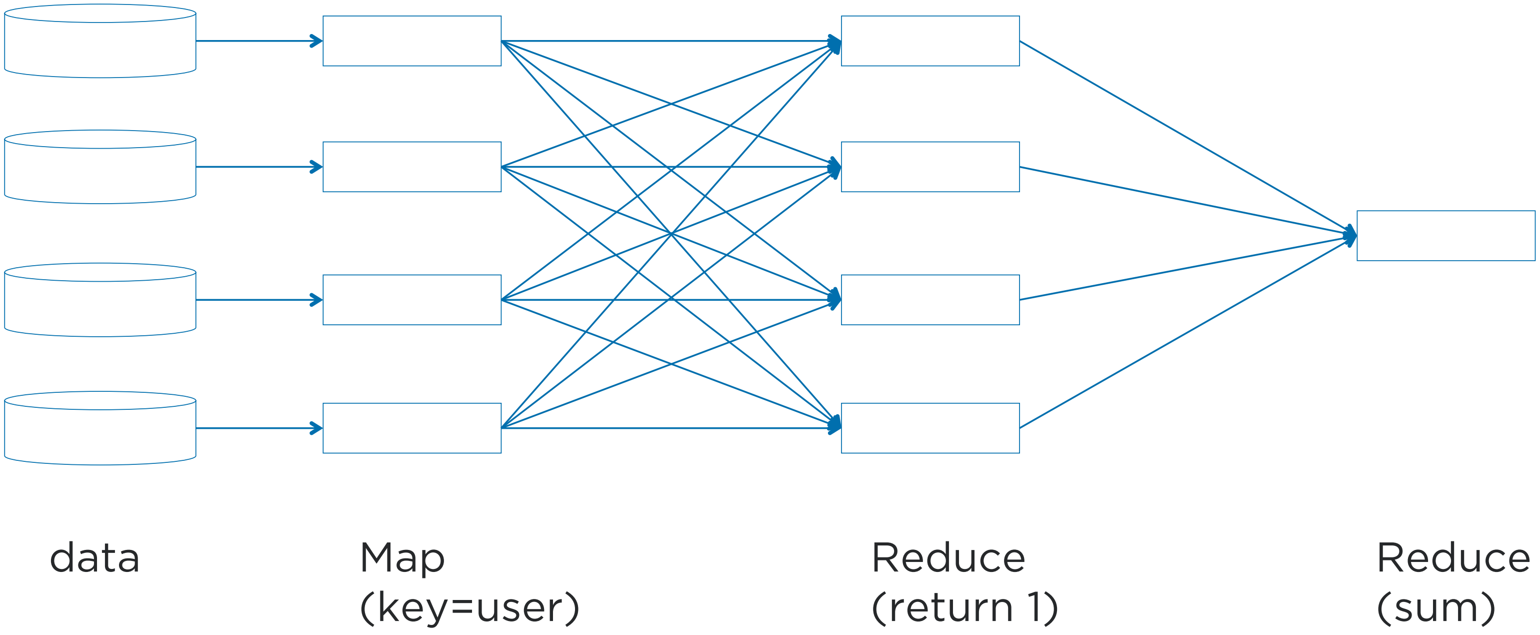
Frequency Queries

- The numbers of times each app was purchased

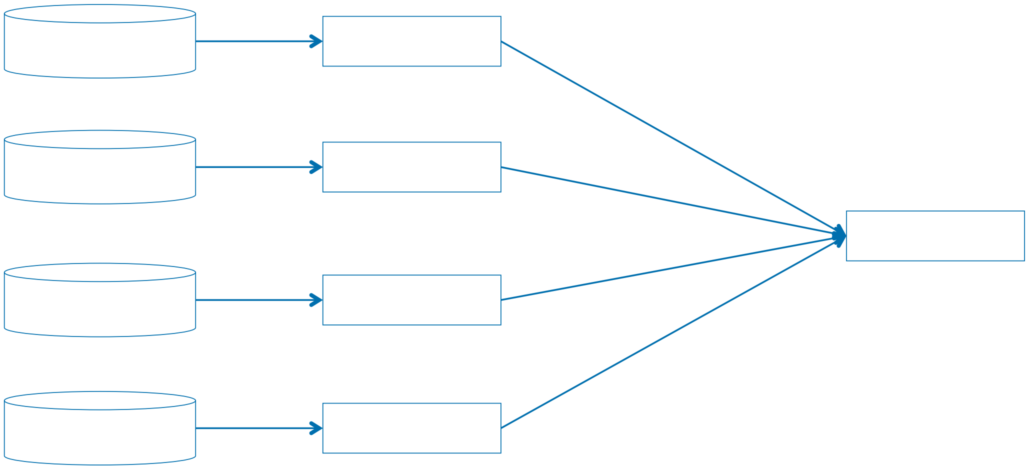
Join Queries

- For all users that purchased Apps, what is the average / median time spent?

Number of unique users (hard)



Number of unique users (made easy)



data

Map
(sketch)

Reduce
(merge)

Current Sketch Implementations

Count Unique Sketches

- Both Theta Sketches* and HLL Sketches
- **Estimating Cardinality** of a stream of identifiers with duplicates
- **Set Operations** (e.g., Union, Intersection, and Difference)
- Can be extended to produce approximate Joins

Quantiles Sketches

- Normal or Inverse PMF's, CDF's of streams of numeric values, using after-the-fact queries.

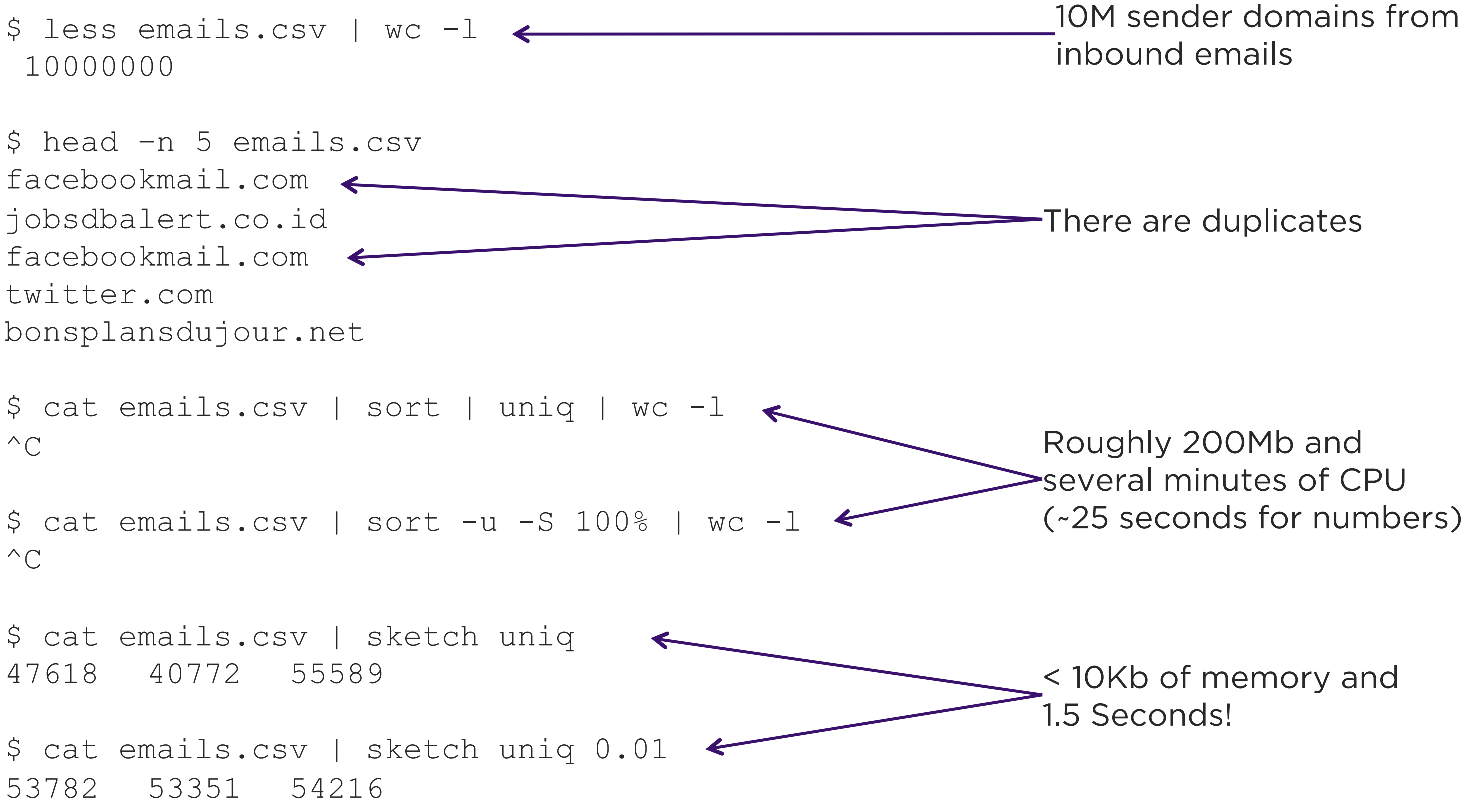
Frequent Item Sketches

- Identify the Heavy Hitters of arbitrary objects from a stream of objects
- Estimate the frequency of any item from the stream

DataSketches.GitHub.io Open Source Library

- Dedicated to production quality Sketch implementations.
  - These are not toy algorithms!
  - Heavily used within Yahoo
- Common Attributes
  - True streaming. Single pass, “one-touch” algorithms for either *real-time* or *batch*
  - All Sketches are Mergeable, which makes them highly parallelizable.
  - Designed for multiple large-scale computing environments:
    - Core of library is coded in Java with no external dependencies
    - Easy integration into virtually any system environment
    - Adaptors for Hadoop/Pig and Hadoop/Hive environments
    - Standard library promotes sharing across platforms and organizations
  - Maven deployable and registered with *Maven Central Repository*
    - *http://search.maven.org/#search|ga|1|datasketches*
  - Comprehensive unit tests and testing tools are provided
  - Extensive documentation with Systems Developers in mind
  - All algorithms are backed by published mathematical theory

Counting distinct elements example



```
$ less emails.csv | wc -l
 10000000

$ head -n 5 emails.csv
facebookmail.com
jobsdbalert.co.id
facebookmail.com
twitter.com
bonsplansdujour.net

$ cat emails.csv | sort | uniq | wc -l
^C

$ cat emails.csv | sort -u -S 100% | wc -l
^C

$ cat emails.csv | sketch uniq
47618   40772   55589

$ cat emails.csv | sketch uniq 0.01
53782   53351   54216
```

10M sender domains from inbound emails

There are duplicates

Roughly 200Mb and several minutes of CPU (~25 seconds for numbers)

< 10Kb of memory and 1.5 Seconds!

YAHOO!