

YAHOO!

Streaming Data Mining

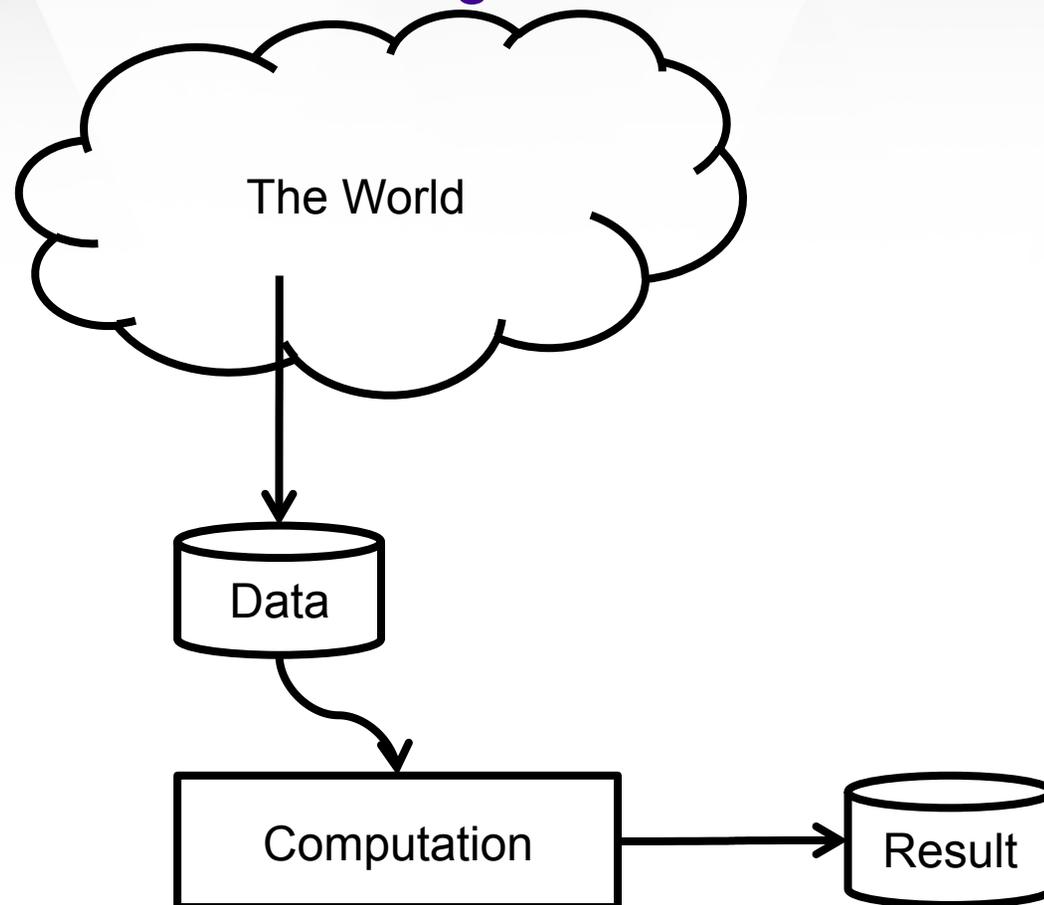
PRESENTED BY **Edo Liberty** | April 11, 2014

Copyright © 2014 Yahoo! All rights reserved. No reproduction or distribution allowed without express written permission.

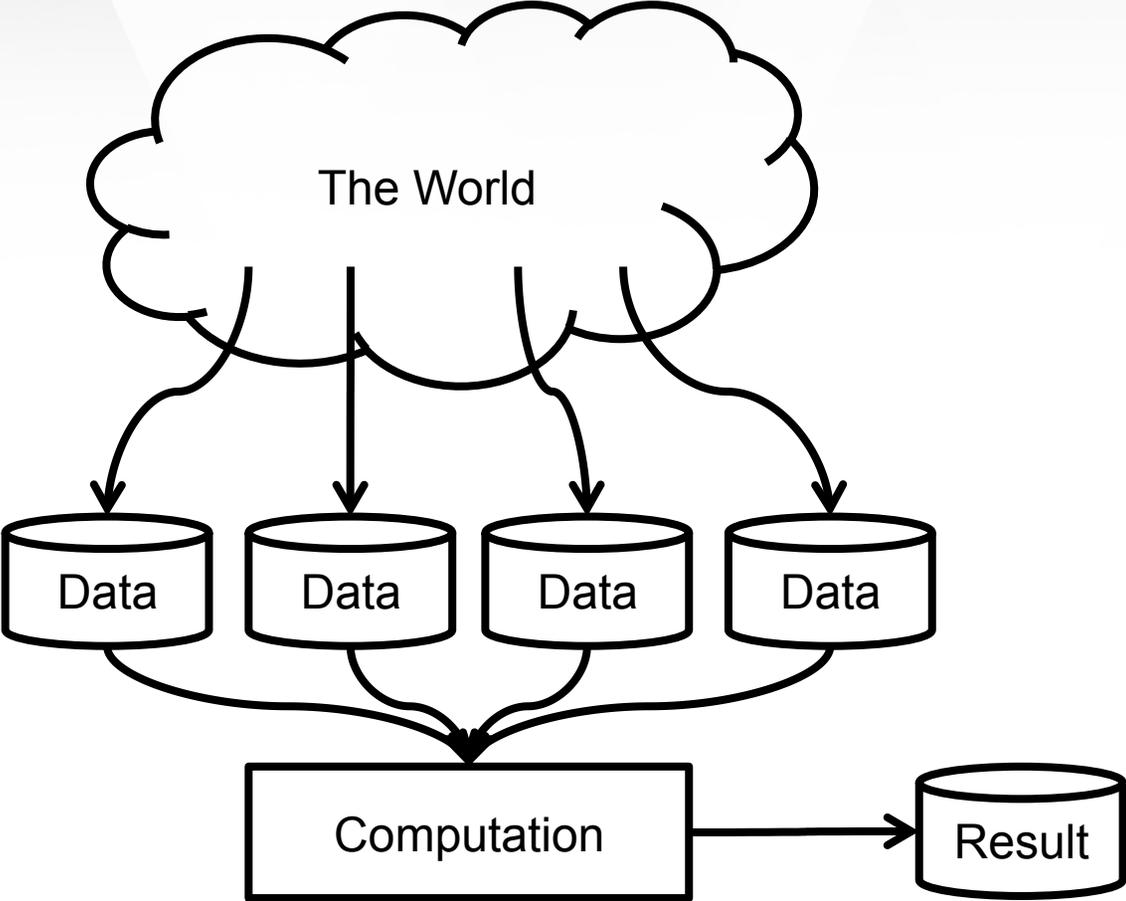


Parts of this presentation were given with Jelani Nelson (Harvard) as a KDD tutorial on streaming data mining.

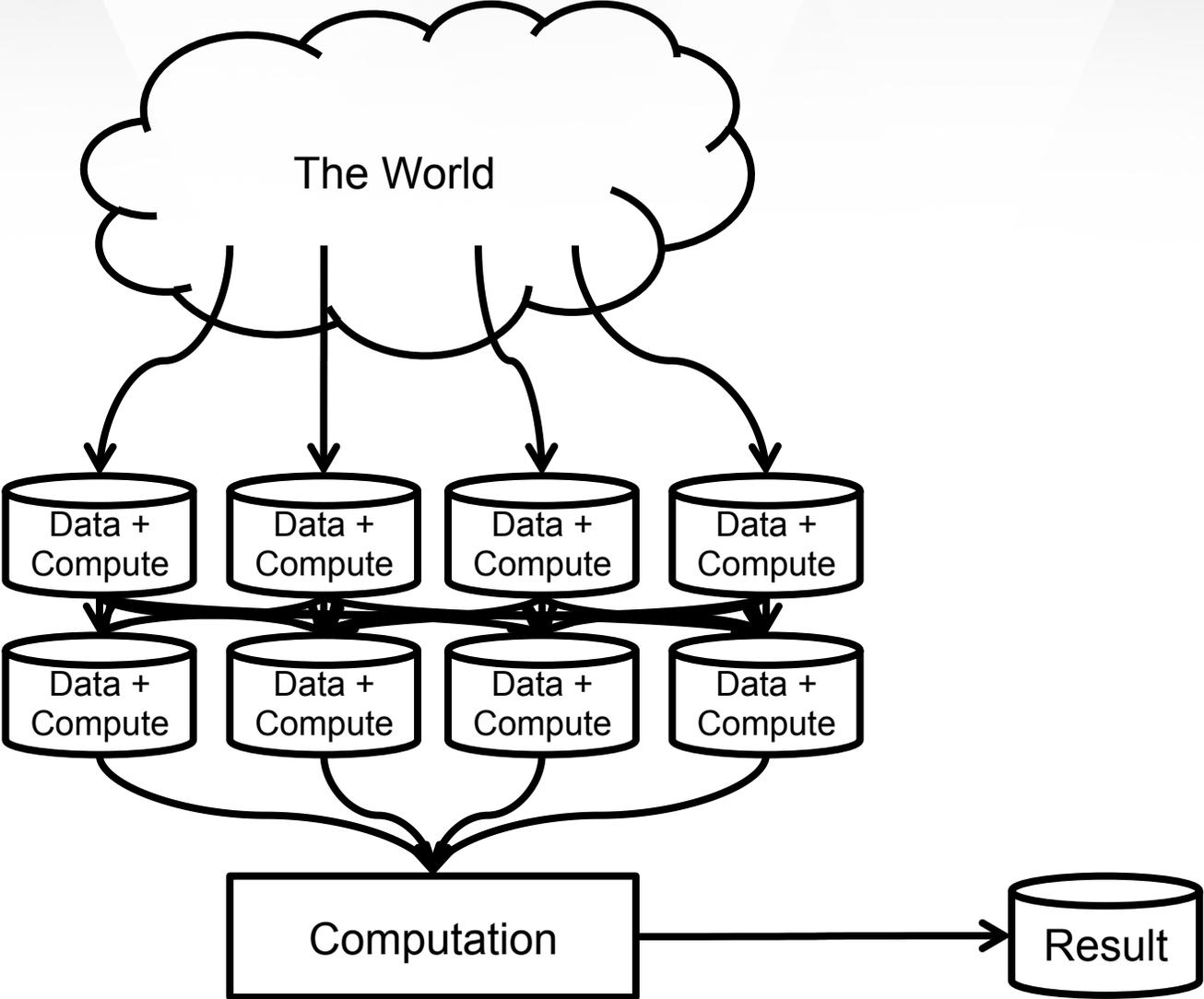
Single machine data mining



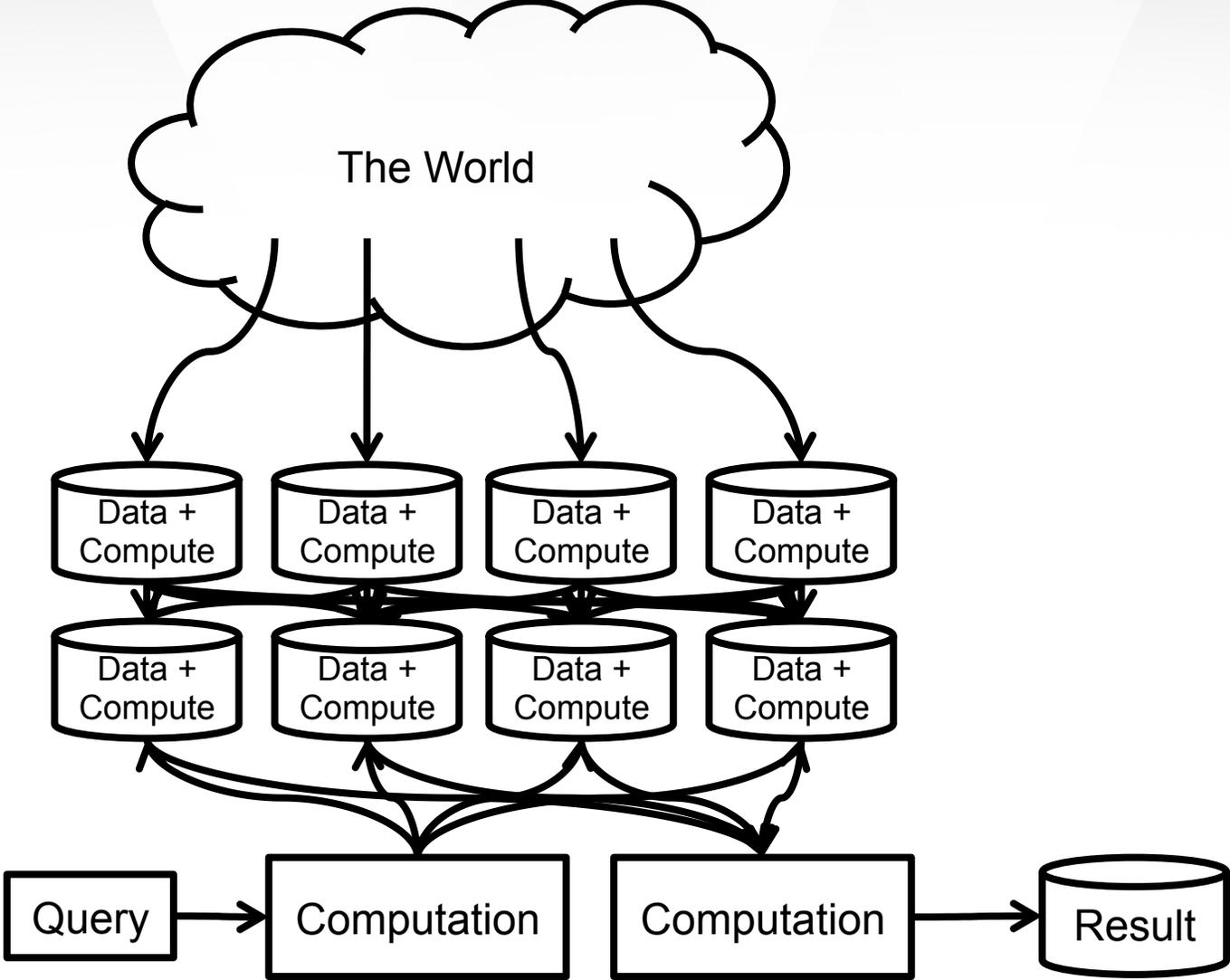
Distributed storage



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

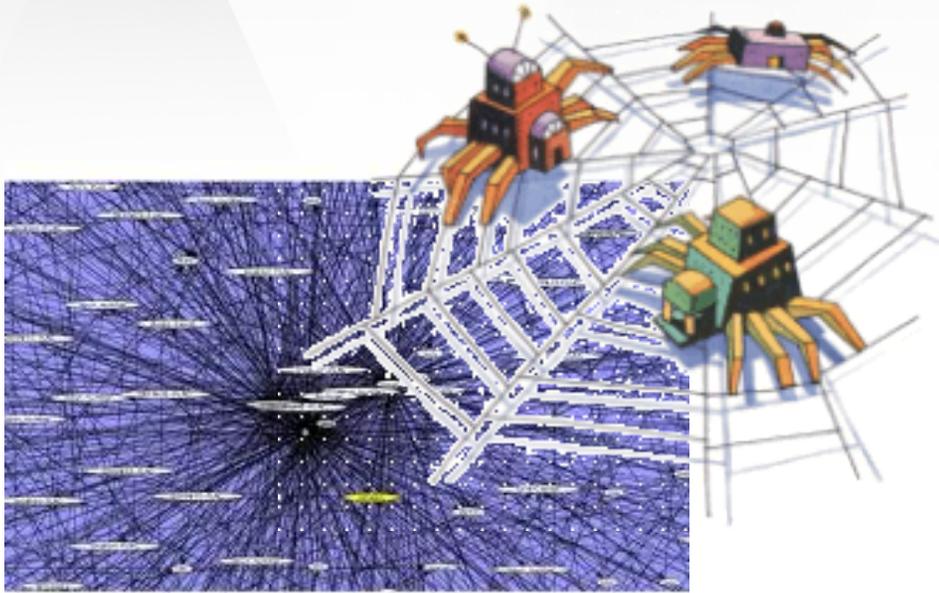


Distributed model (indexes, tables, databases, ...)

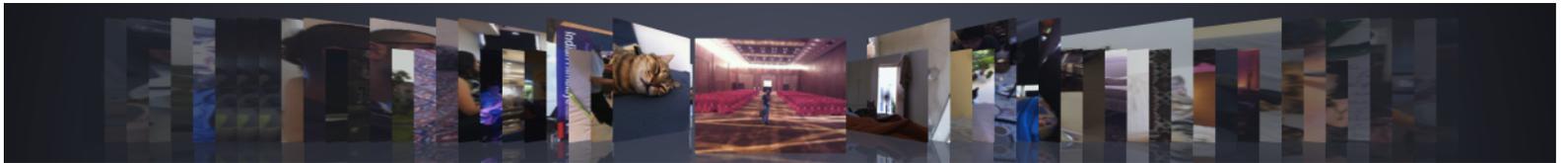


207 big-data infographics (meta infographic)

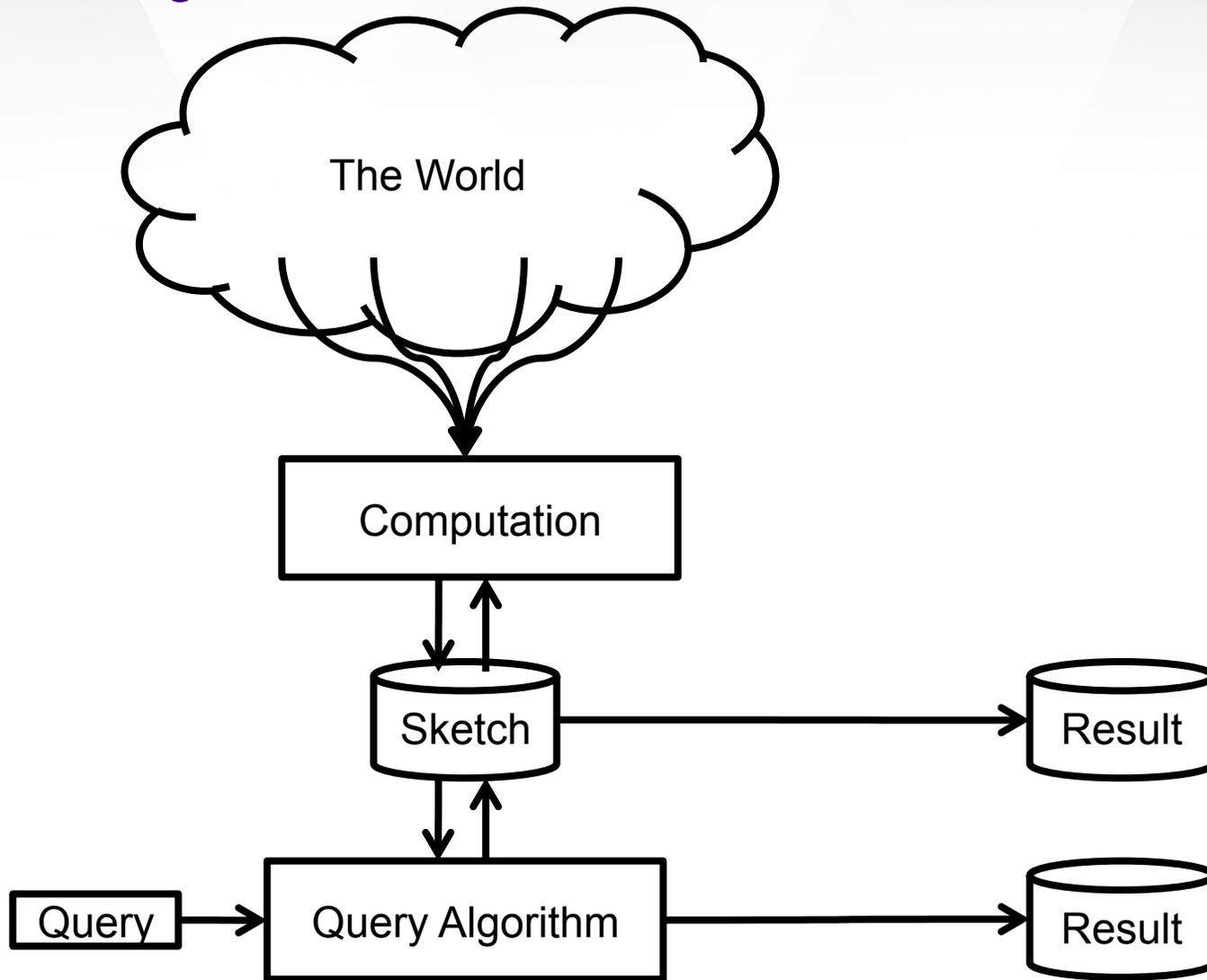




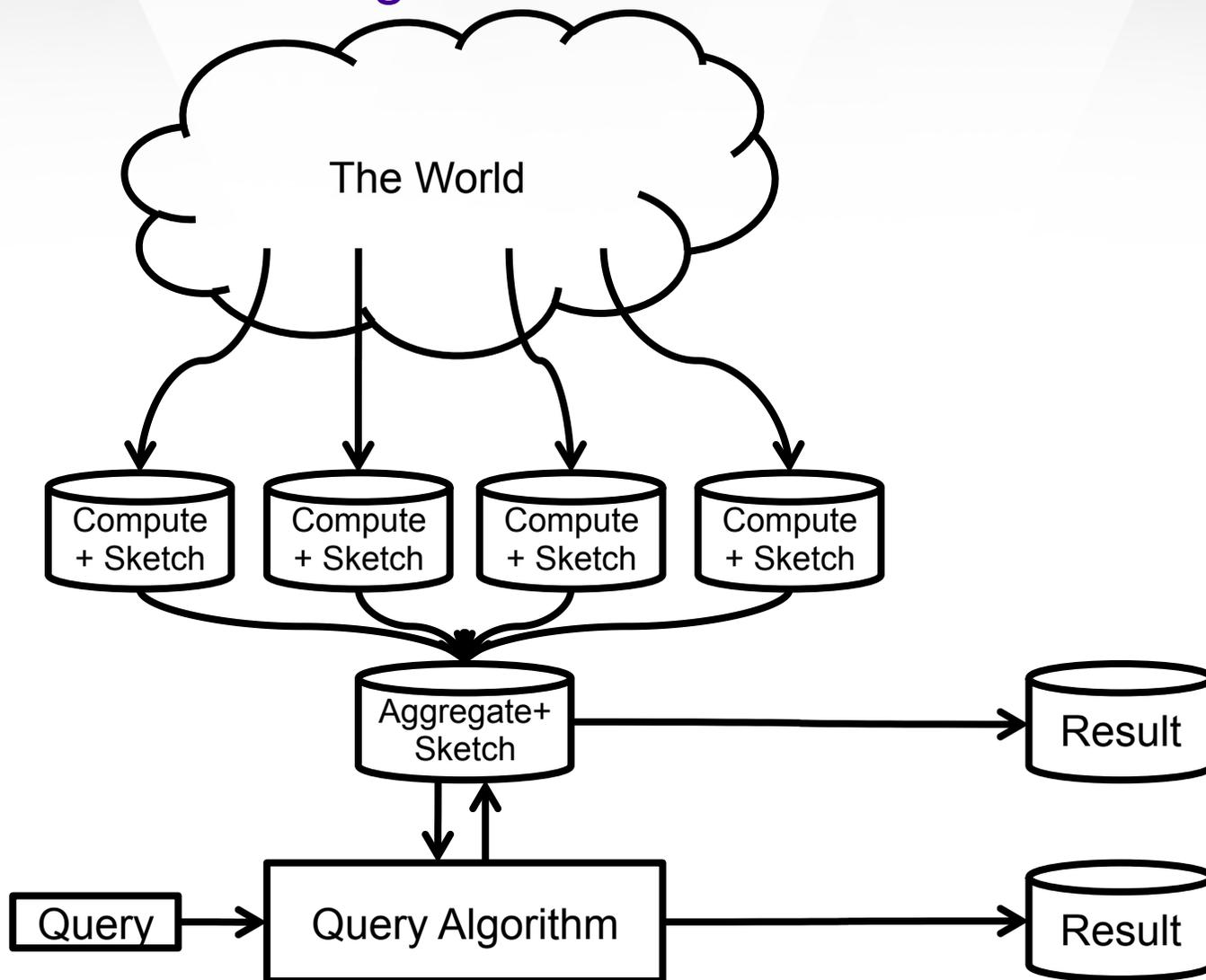
flickr



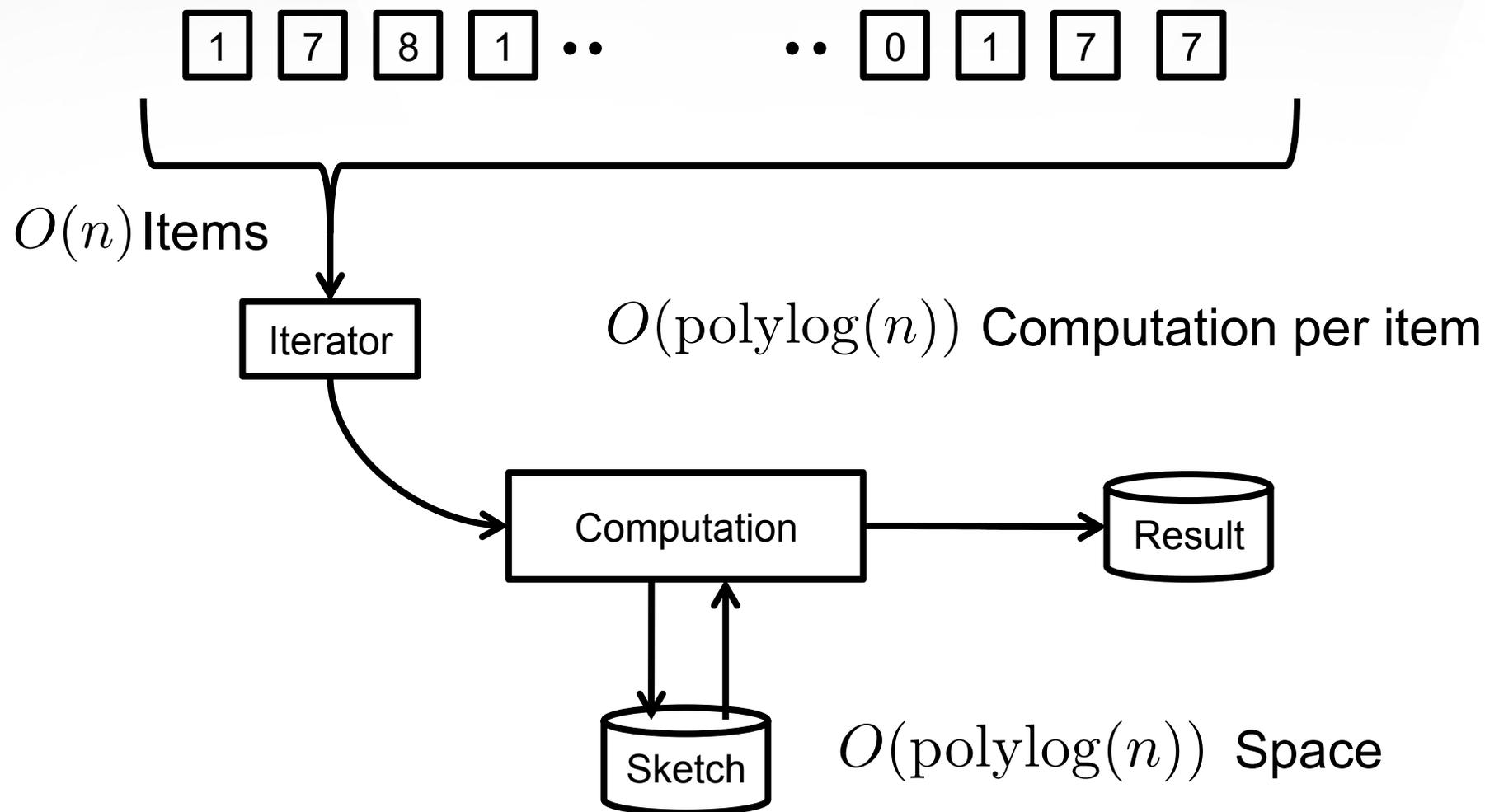
The streaming model



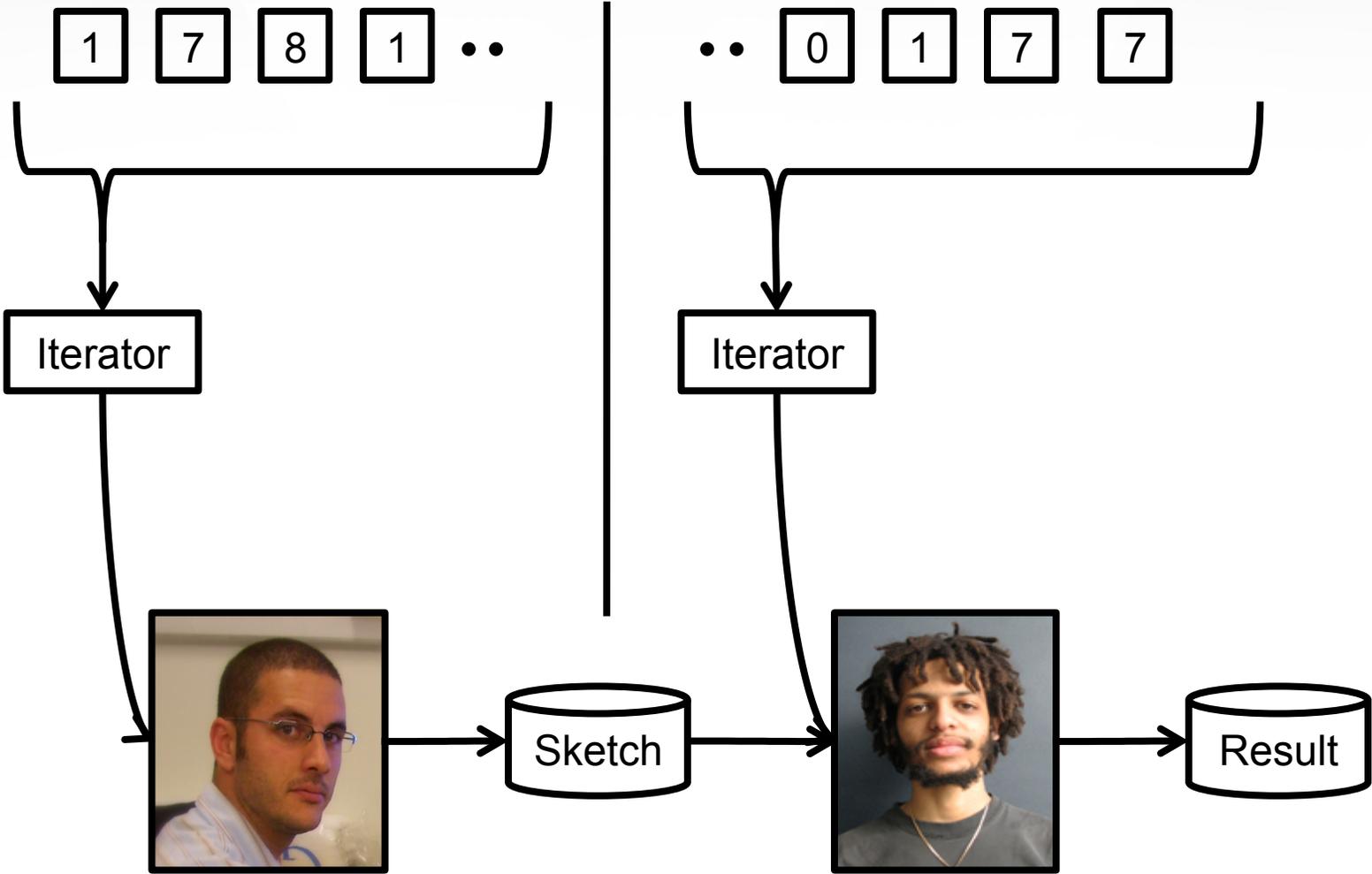
The parallel streaming model



The streaming model (more accurately)



Communication complexity



Frequent items

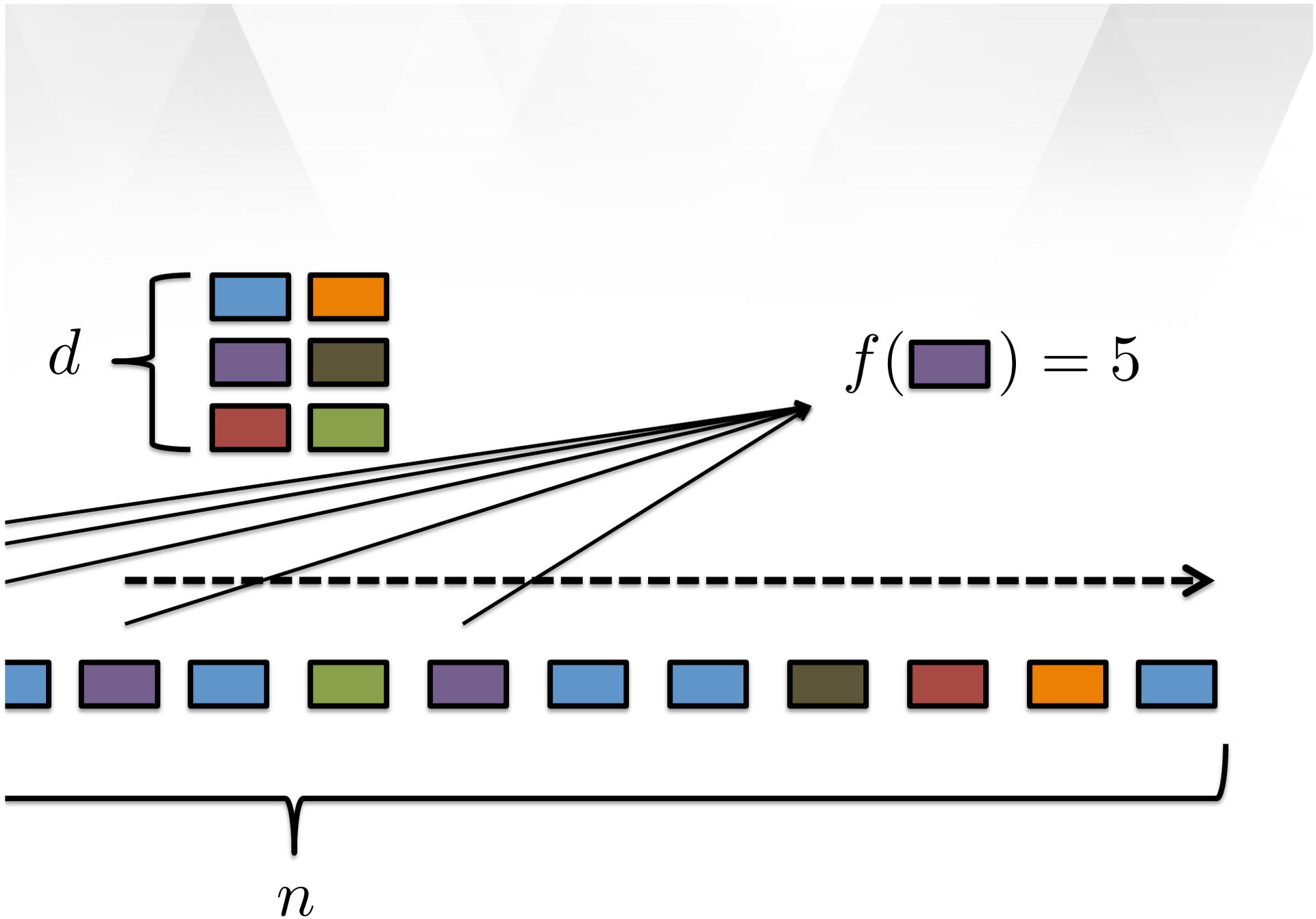
Misra, Gries. Finding repeated elements, 1982.

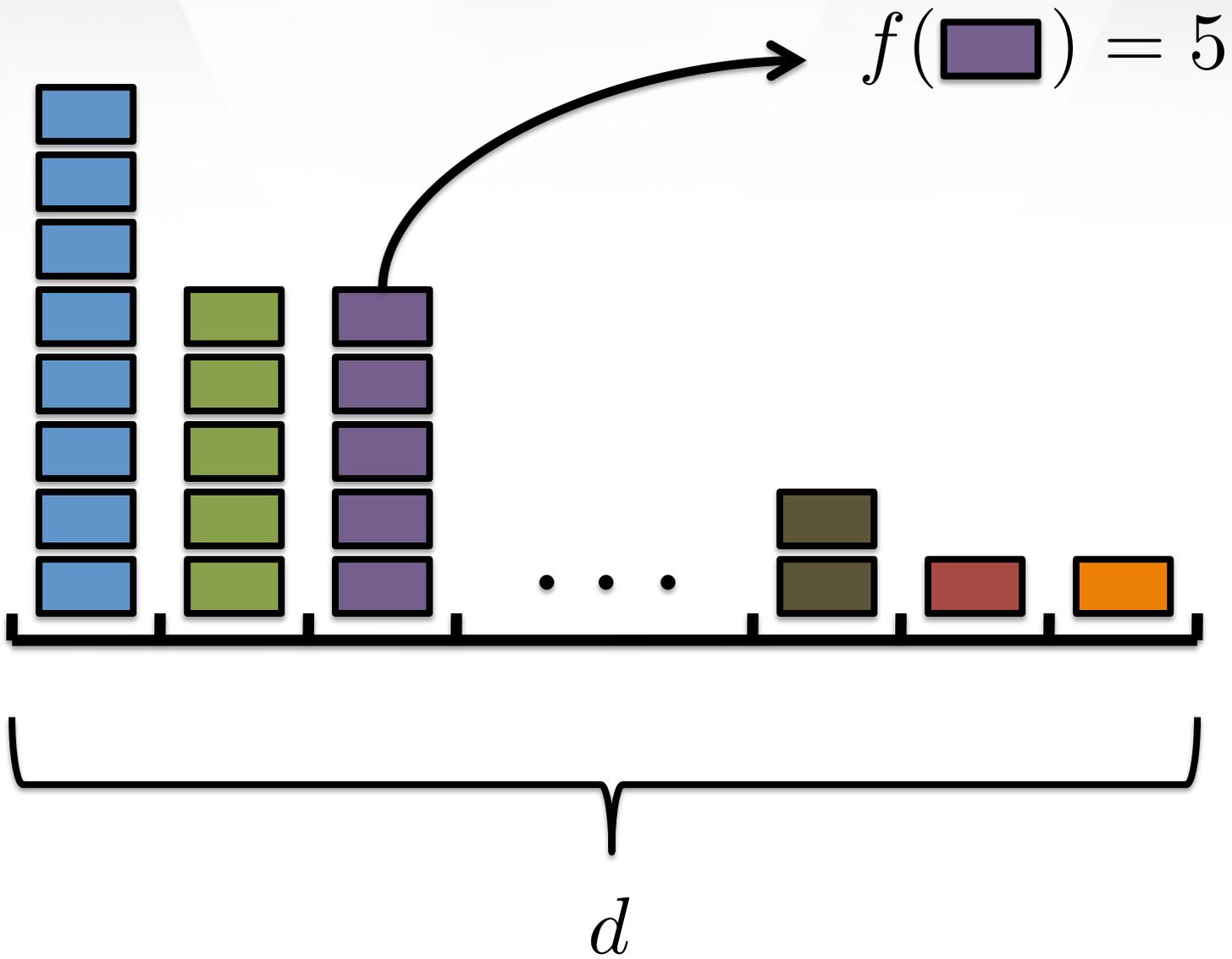
Demaine, Lopez-Ortiz, Munro. Frequency estimation of internet packet streams with limited space, 2002

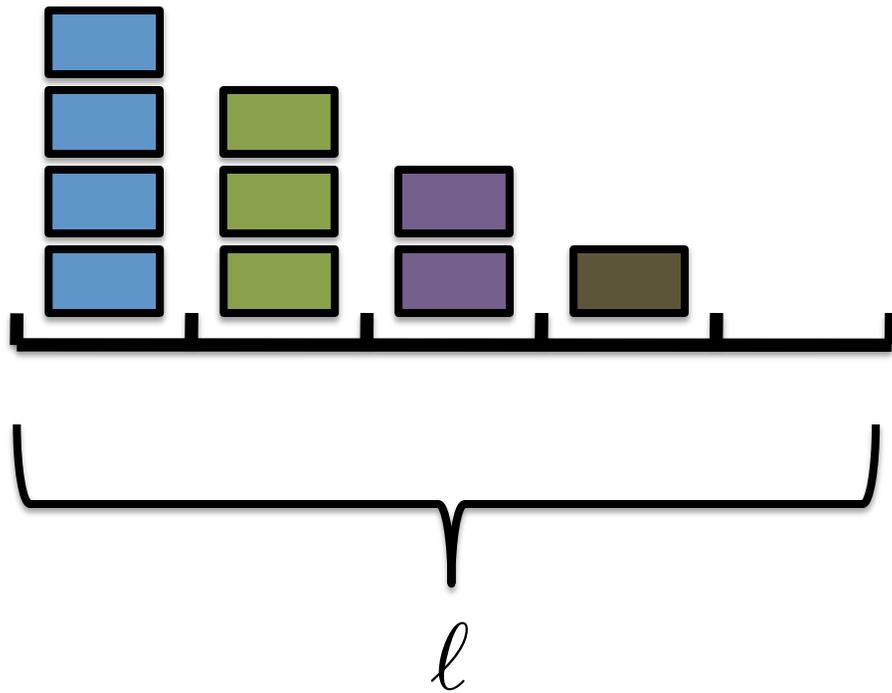
Karp, Shenker, Papadimitriou. A simple algorithm for finding frequent elements in streams and bags, 2003

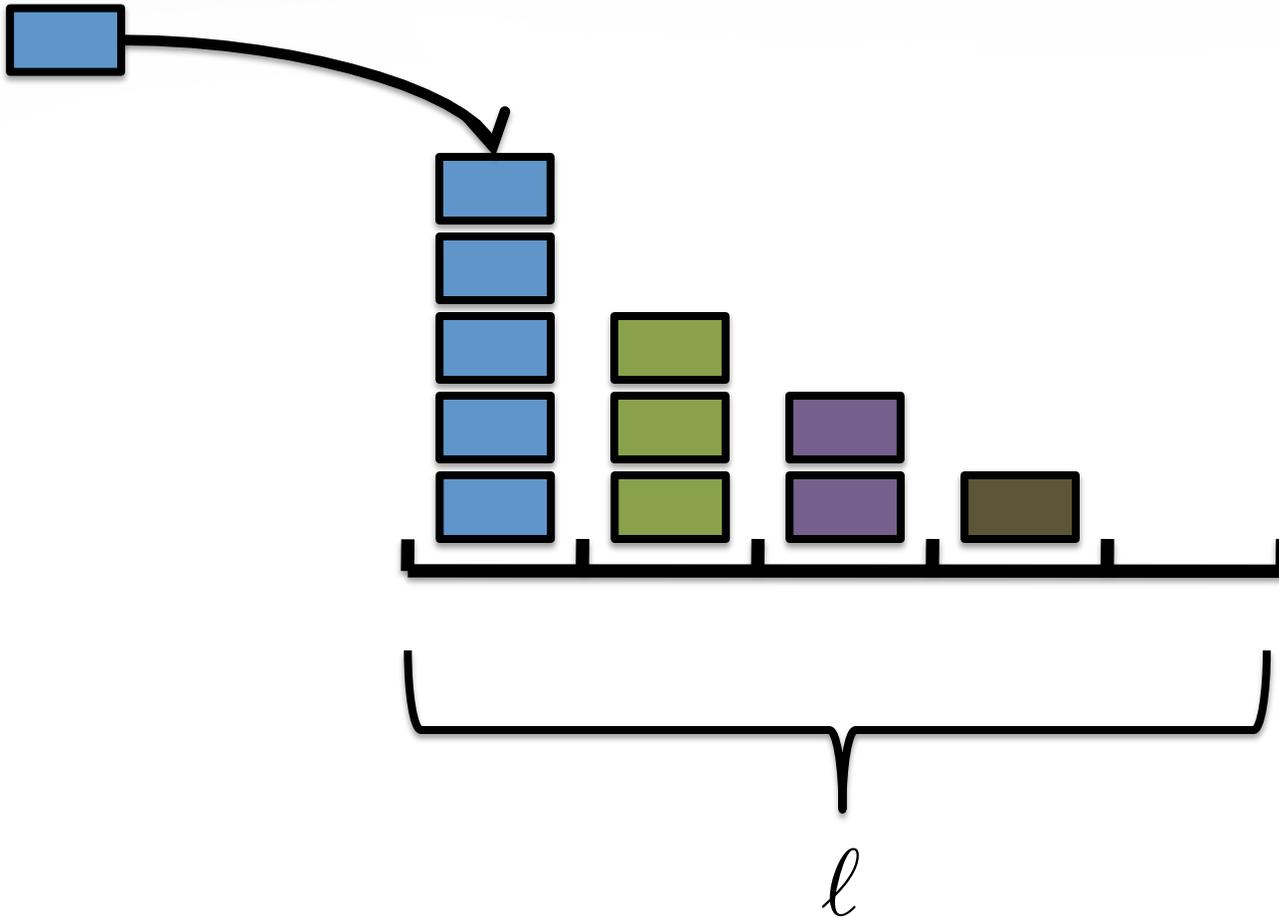
The name "Lossy Counting" was used for a different algorithm by Manku and Motwani, 2002

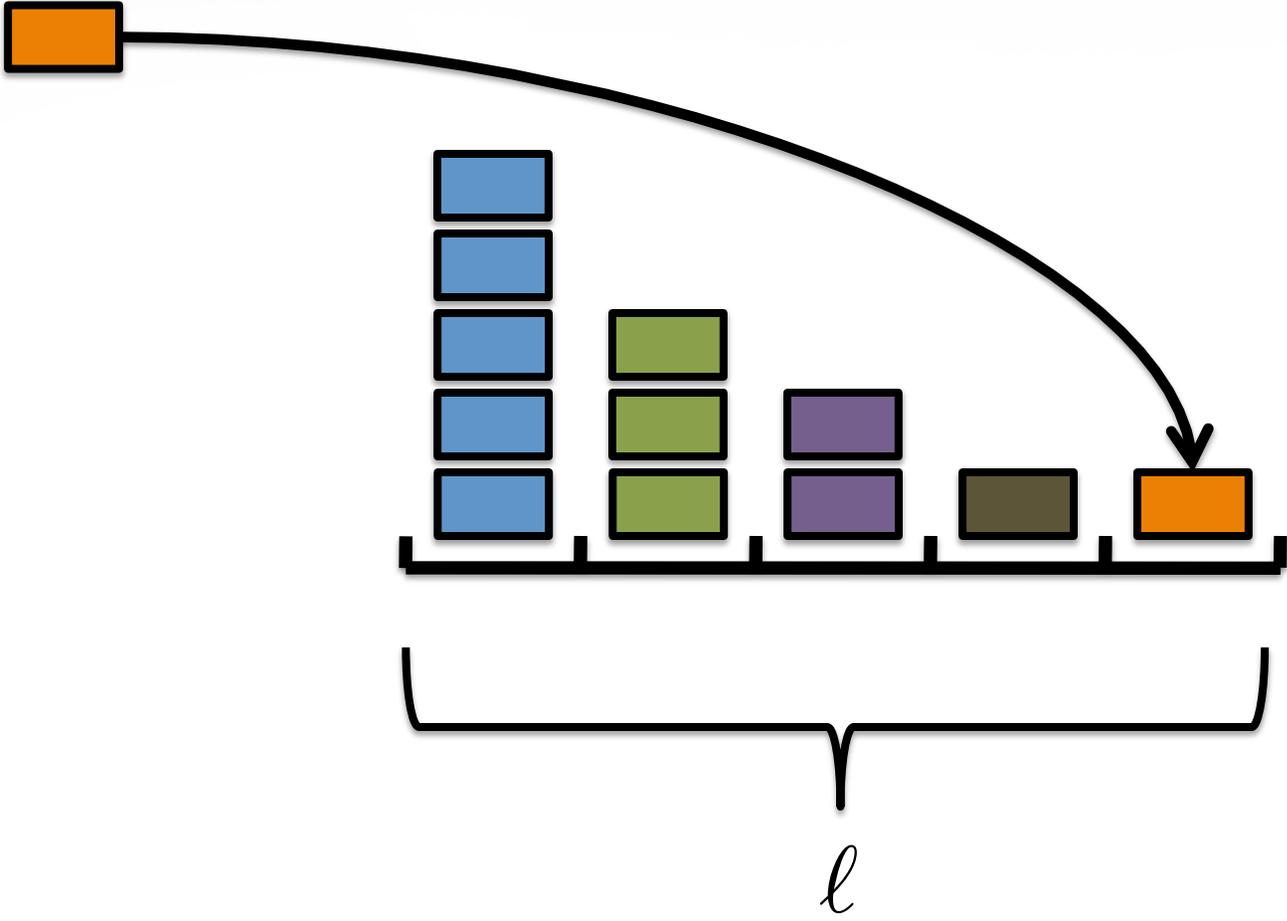
Metwally, Agrawal, Abbadi, Efficient Computation of Frequent and Top-k Elements in Data Streams, 2006

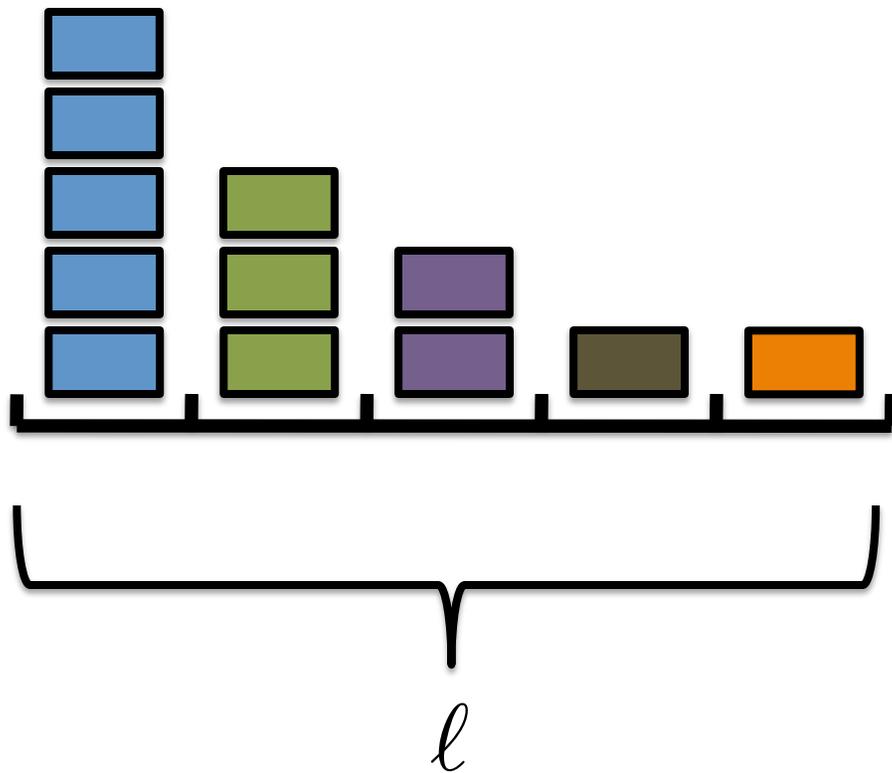


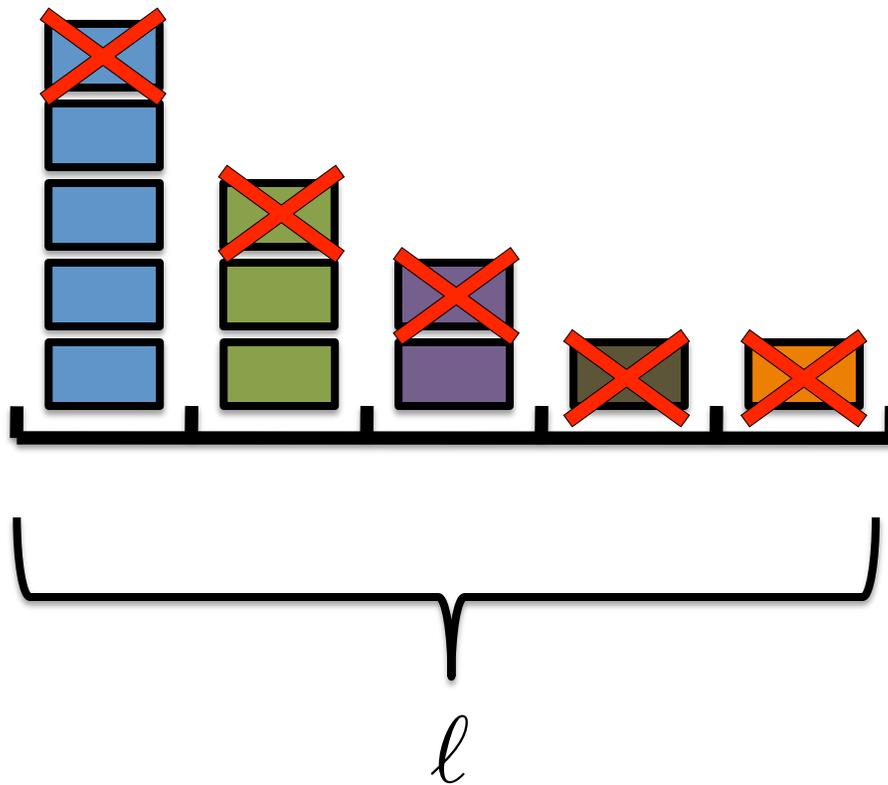


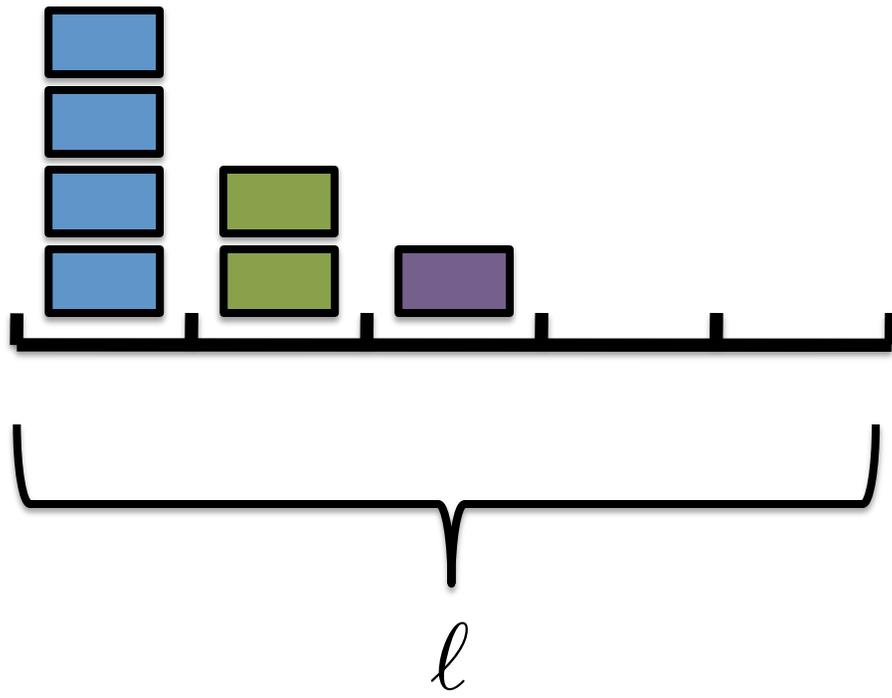


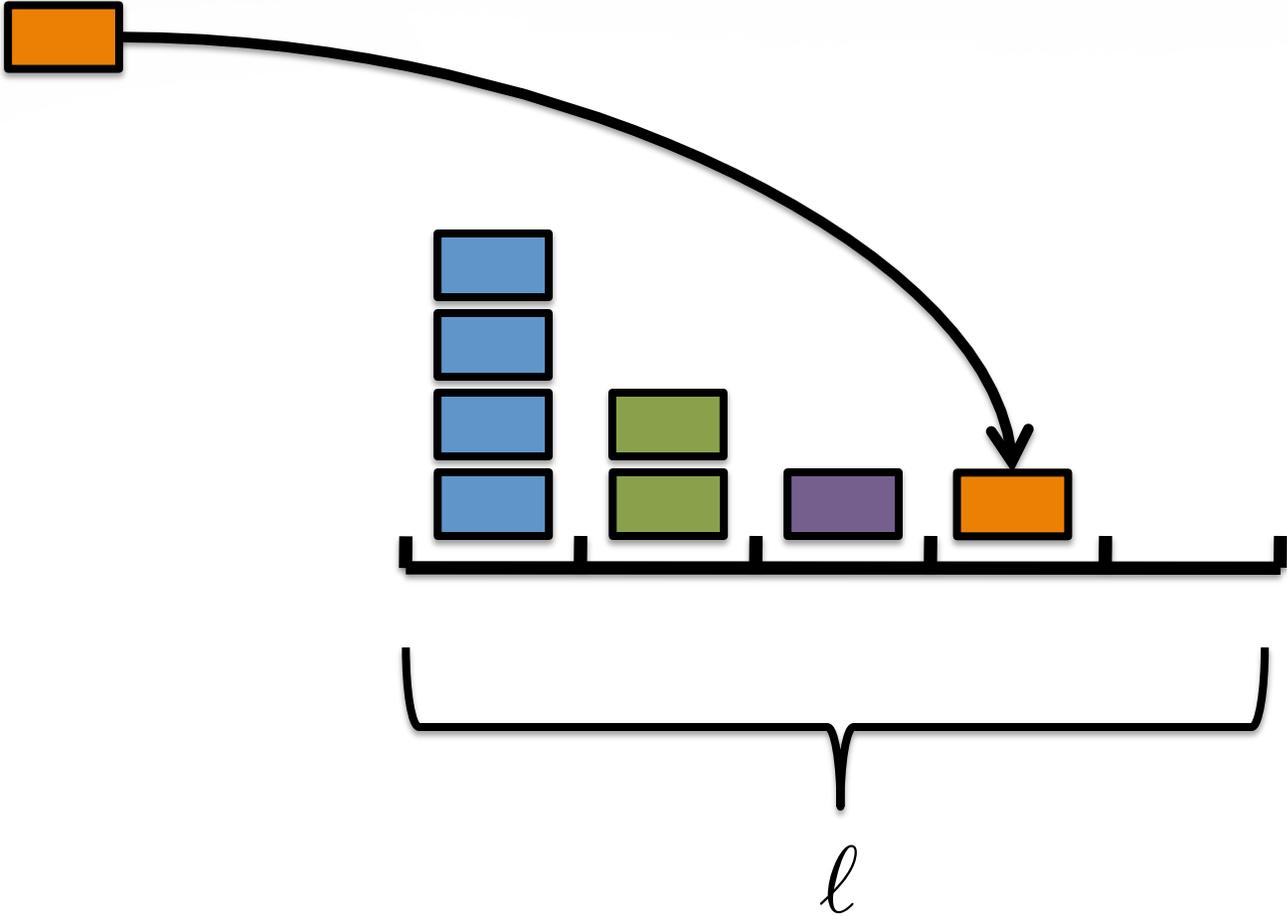


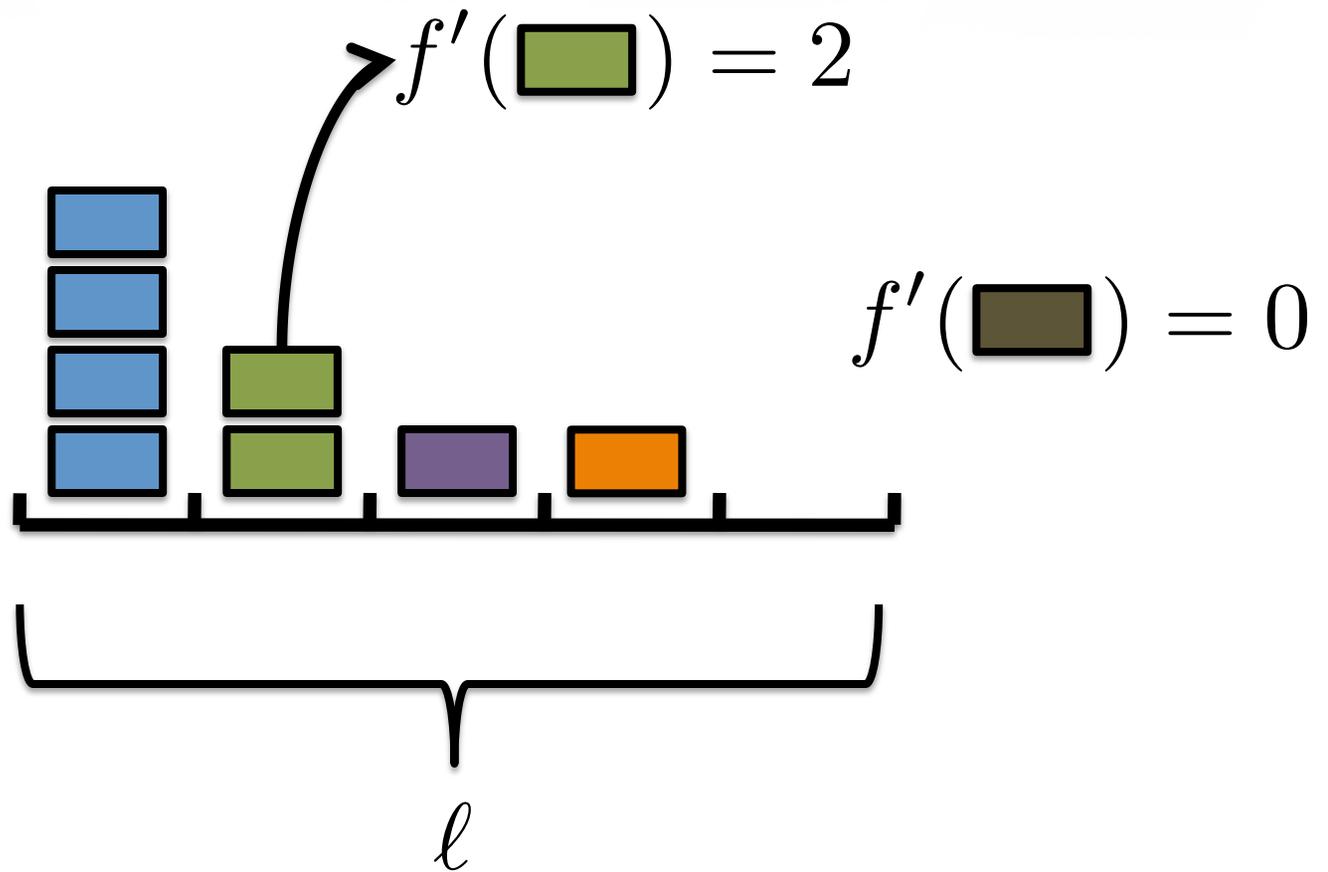








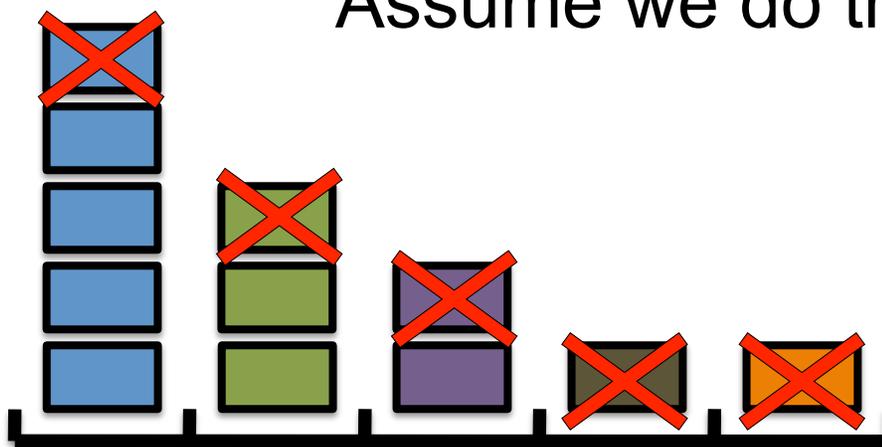




The proof (very short)

First fact: $f'(x) \leq f(x)$

Assume we do this t times



Second fact: $f'(x) \geq f(x) - t$

The proof (very short)

Third (not so obvious) fact:

$$0 \geq \sum f'(x) = \sum f(x) - t \cdot \ell = n - t \cdot \ell$$

Which gives $t \leq n/\ell$. In words:

We can only delete ℓ items n/ℓ times!

$$|f'(x) - f(x)| \leq n/\ell$$



Useful form...

Define $p(x) = f(x)/n$

And $p'(x) = f'(x)/n$

We get that

$$|p'(x) - p(x)| \leq 1/\ell$$

This is very useful for keeping approx' distributions!

Threading Machine Generated Email

Email threads

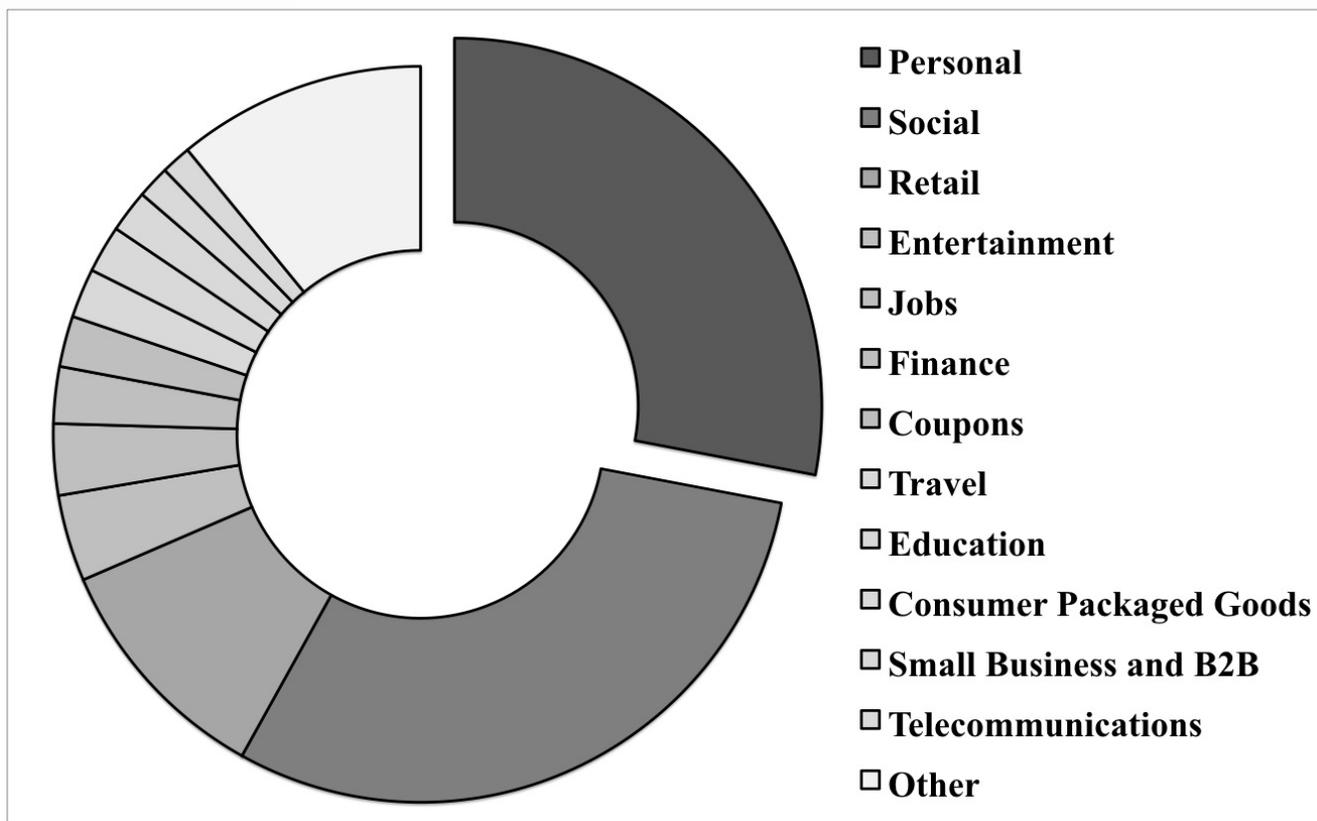
The screenshot shows the Yahoo! Mail interface. At the top, there is a navigation bar with links for Home, Mail, News, Sports, Finance, Weather, Games, Groups, Answers, Screen, Flickr, and Apps. Below this is the Yahoo! MAIL logo and a search bar with 'Search Mail' and 'Search Web' buttons. The main area shows an email thread titled 'following up from your CMU visit (3)'. The thread consists of three messages:

- **Emma Brunskill** Hi Edo, It was very interes Mar 3 ★
- **Me** Hi Emma, Thanks for reaching out, I ha Mar 5 ★
- **Emma Brunskill** To Me Mar 7 ★

A black arrow points from the 'Folders (22)' link in the left sidebar to the text below.

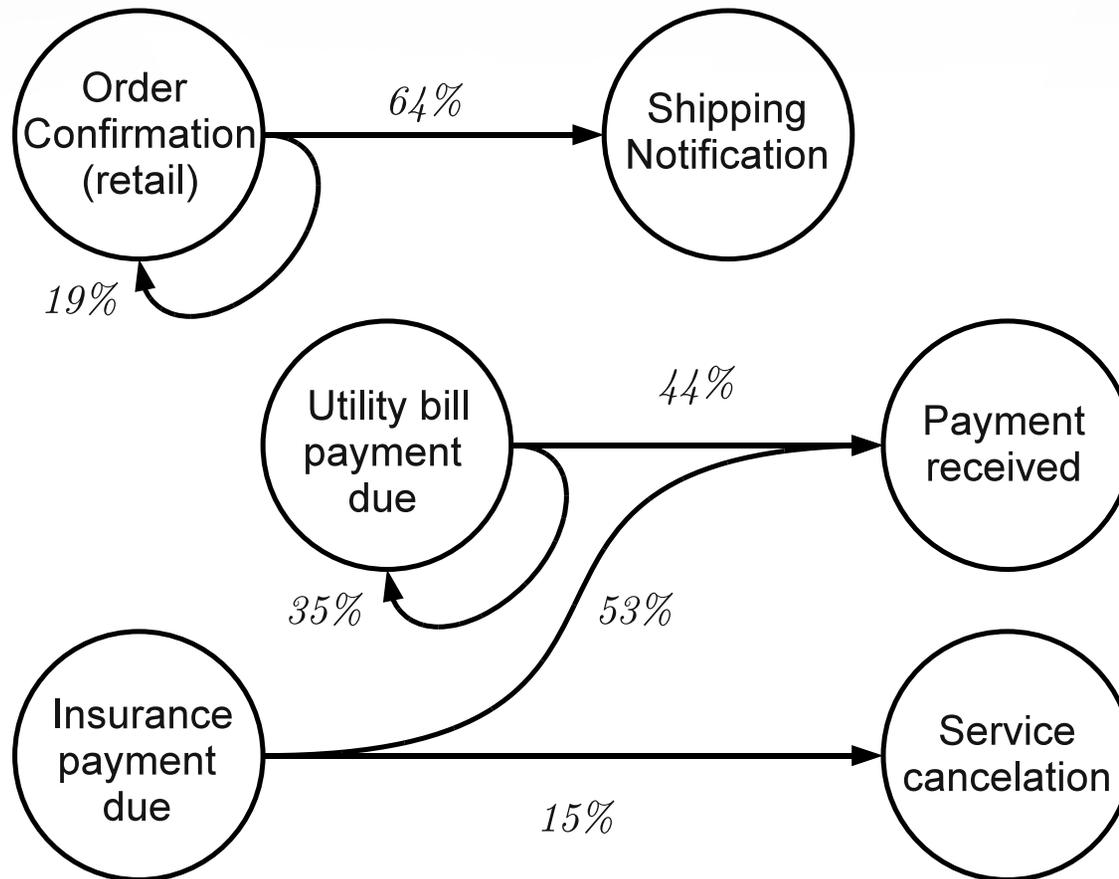
A simple email thread (that's not very hard to do...)

Threading Machine Generated Email

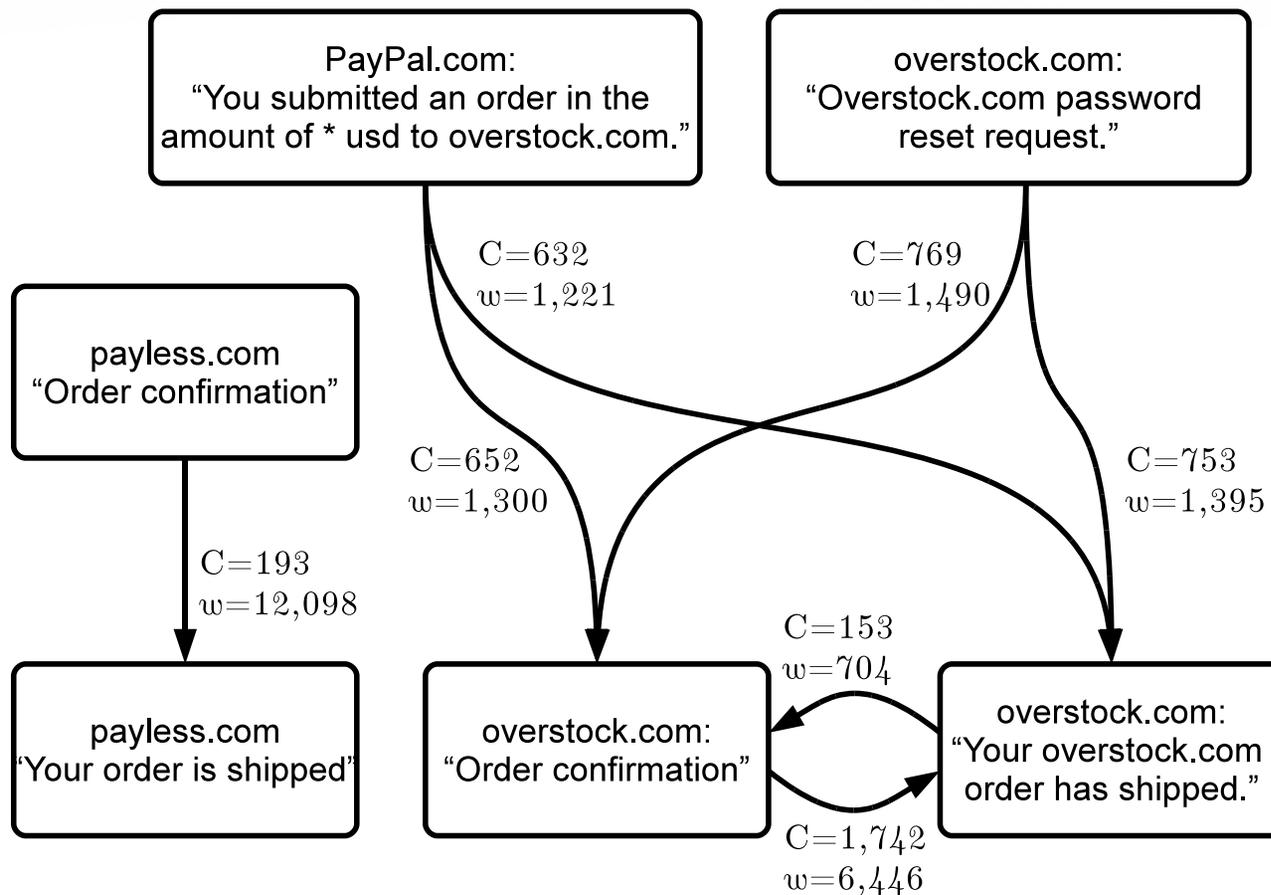


Ailon, Karnin, Maarek, Liberty, Threading Machine Generated Email, WSDM 2013

Threading Machine Generated Email



Threading Machine Generated Email



What else can we do in the streaming model...

Items (words, IP-adresses, events, clicks,...):

- Item frequencies
- Counting distinct elements
- Moment and entropy estimation
- Approximate set operations

Vectors (text documents, images, example features,...)

- Dimensionality reduction
- Clustering (k-means, k-median,...)
- Linear Regression
- Machine learning (some of it at least)

Matrices (text corpora, user preferences, graphs...)

- Covariance estimation matrix
- Low rank approximation
- Sparsification

Thanks!

Yahoo does big data algorithms, software and systems!

Speak to our Talent Team or visit [Careers.Yahoo.com](https://careers.yahoo.com) and explore our career opportunities in NYC or Sunnyvale, CA



Seth Tropper

satrapper@yahoo-inc.com



Doug DeSimone

desimone@yahoo-inc.com



Keith Daniels

kdn1@yahoo-inc.com