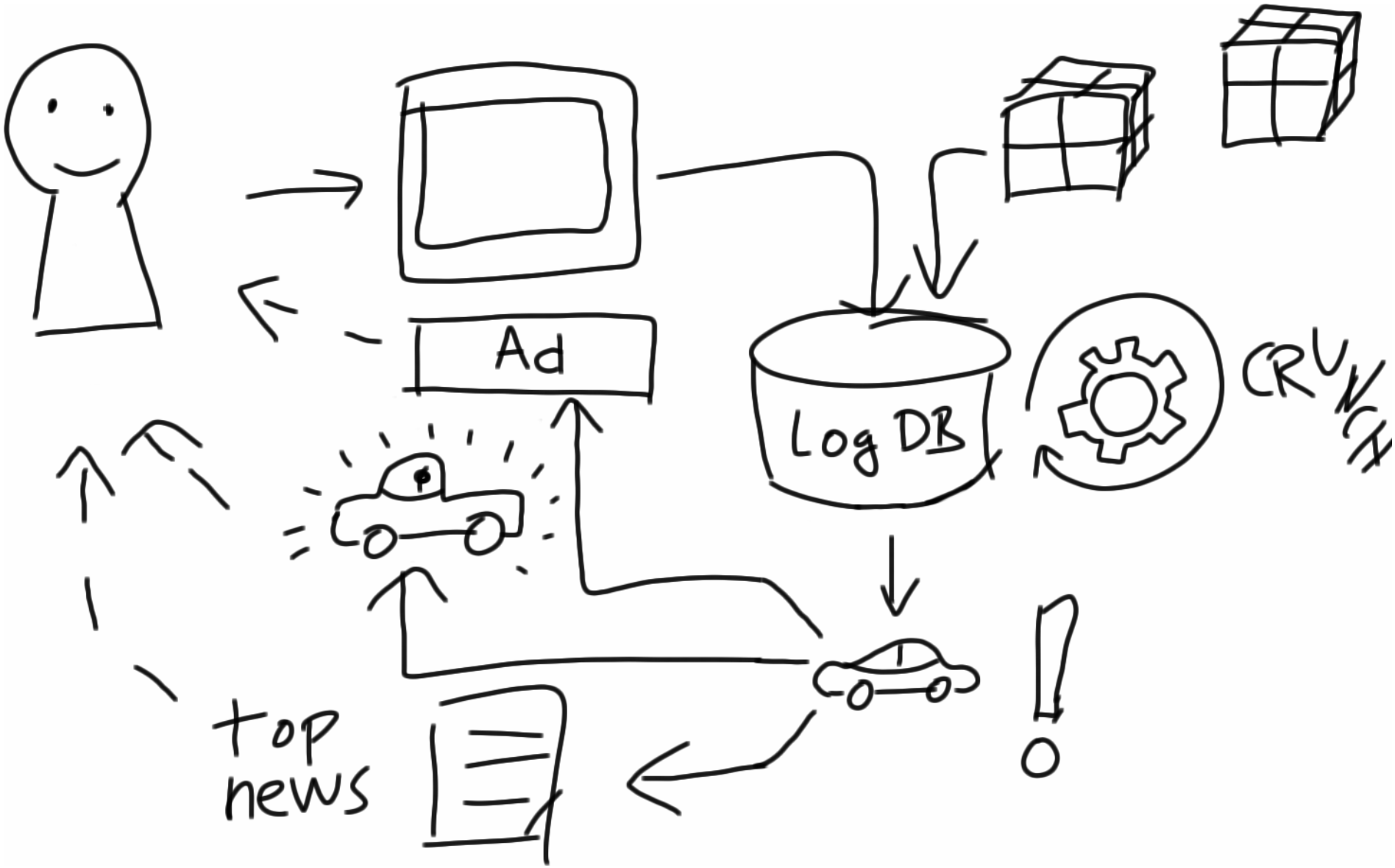
Realtime personalization and recommendation with stream mining

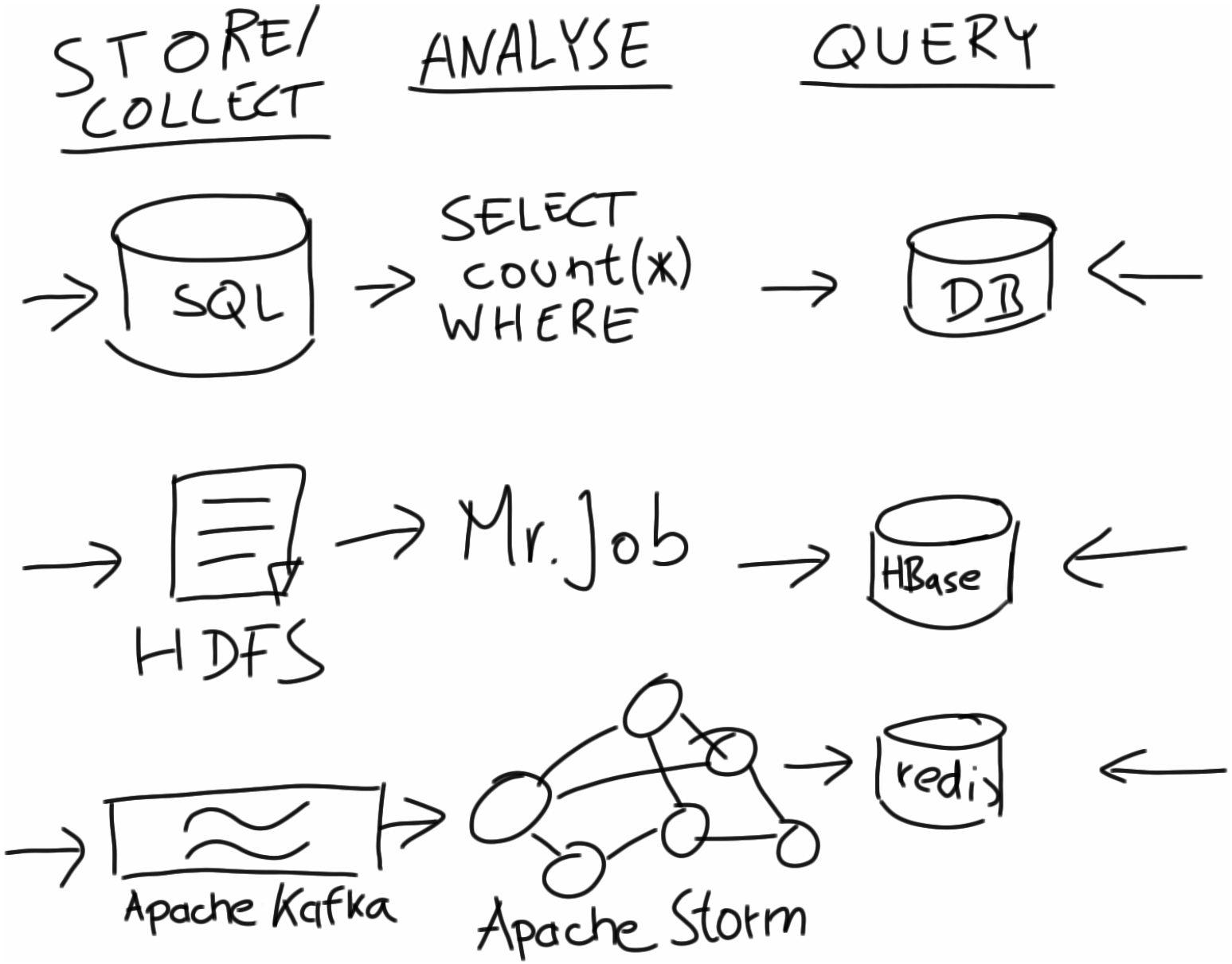
Mikio Braun
@mikiobraun
streamdrill

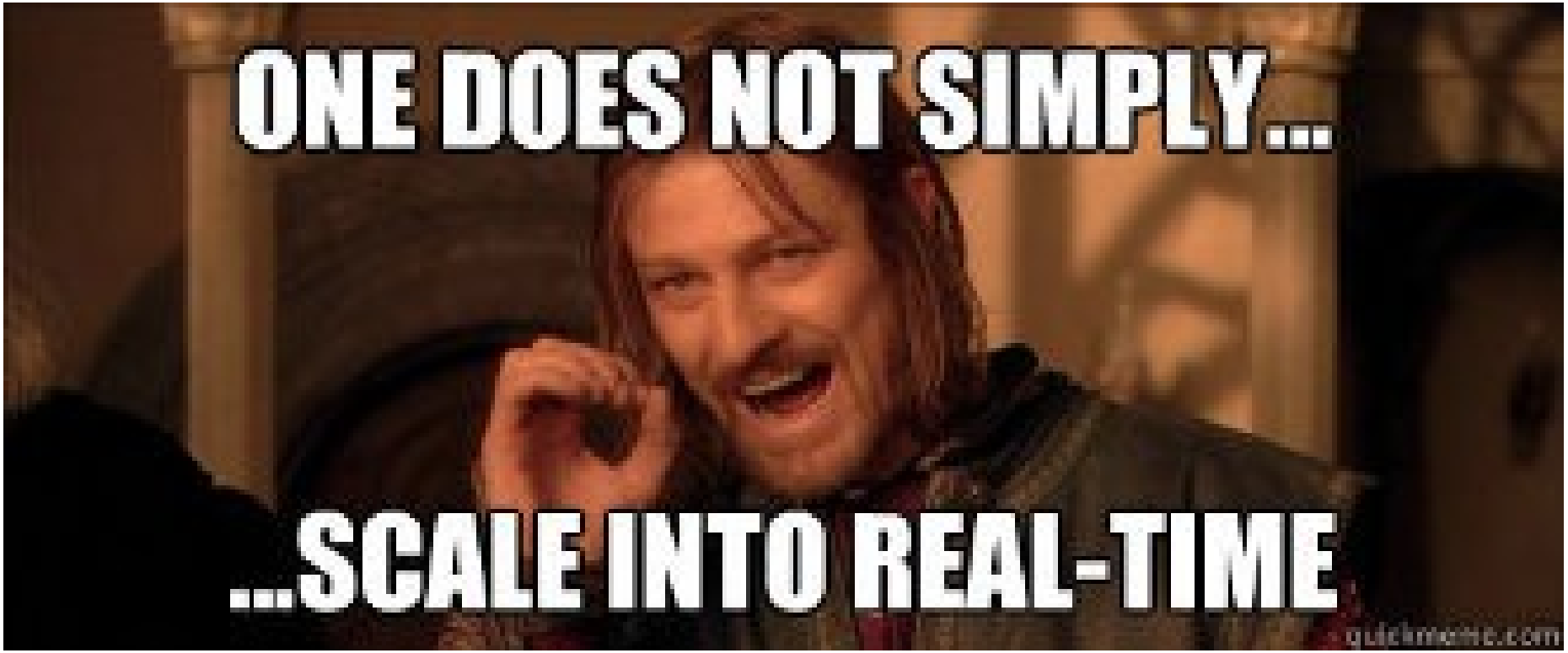
Berlin Buzzwords 2014

Reacting to user behavior

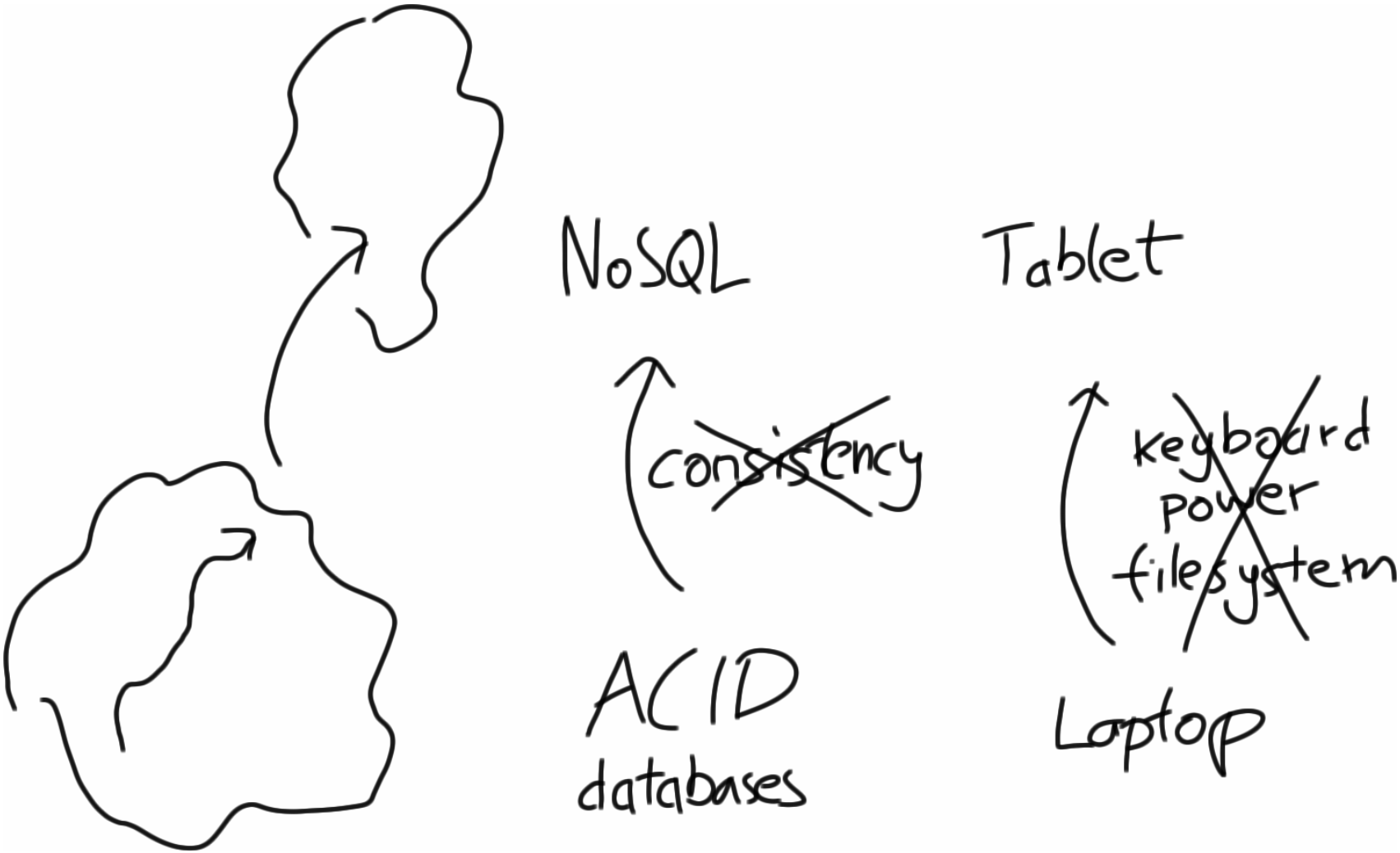


How?


STORE/ COLLECT
ANALYSE
QUERY
SQL
SELECT count(*) WHERE
DB
HDFS
Mr. Job
HBase
Apache Kafka
Apache Storm
redis


ONE DOES NOT SIMPLY...
...SCALE INTO REAL-TIME
quickmeme.com

Making progress



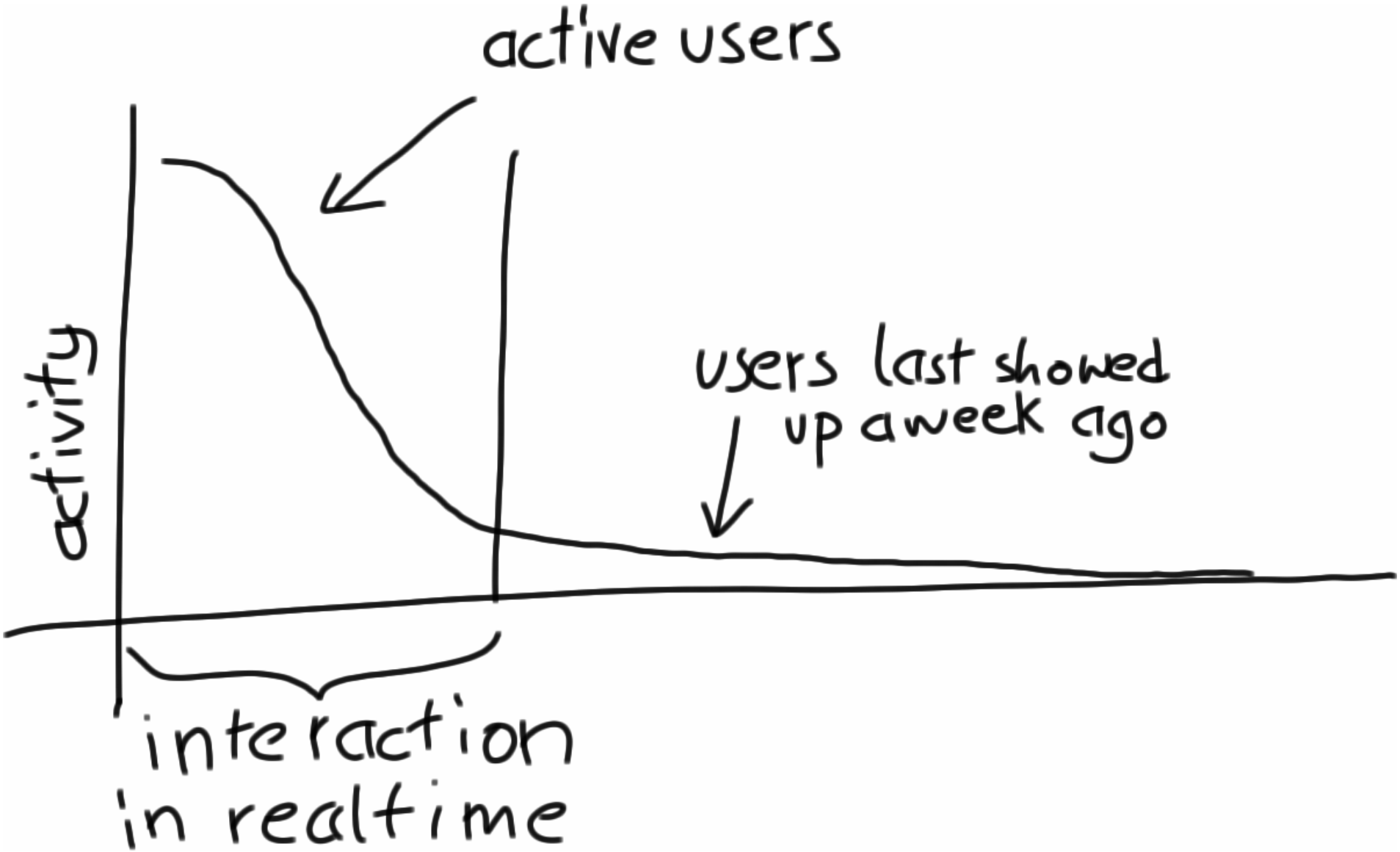
Getting rid of exactness

"classical"
Big Data

~~exactness~~ → stream mining

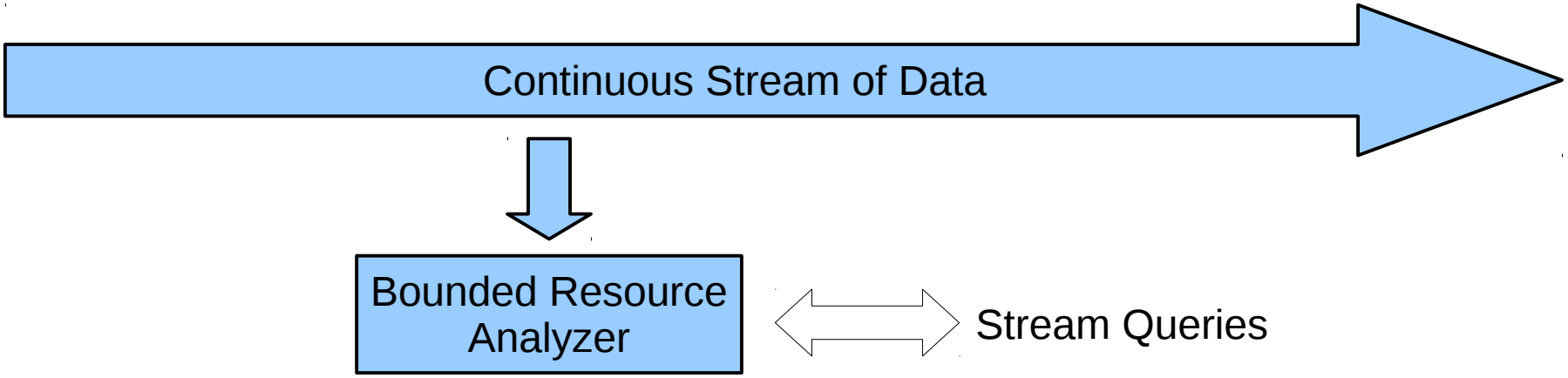
(in particular exact
aggregates
count
avg
...

Important users



Stream Mining to the rescue

- Stream mining algorithms:
  - answer “stream queries” with finite resources
- Typical examples:
  - **how often** does an item appear in a stream?
  - **how many distinct** elements are in the stream?
  - what are the **top-k most frequent** items?



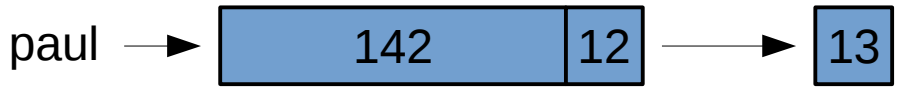
Heavy Hitters (a.k.a. Top-k)

- Count activities over large item sets (millions, even more, e.g. IP addresses, Twitter users)
- Interested in most active elements only.

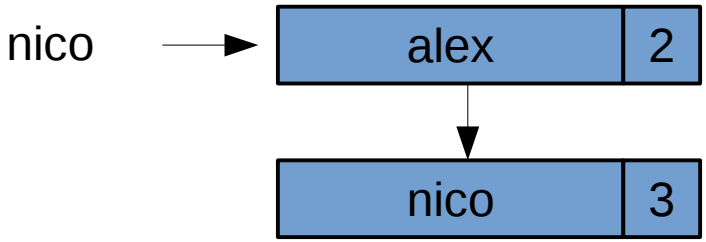
frank	15
paul	12
jan	8
felix	5
leo	3
alex	2

Fixed tables of counts

Case 1: element already in data base



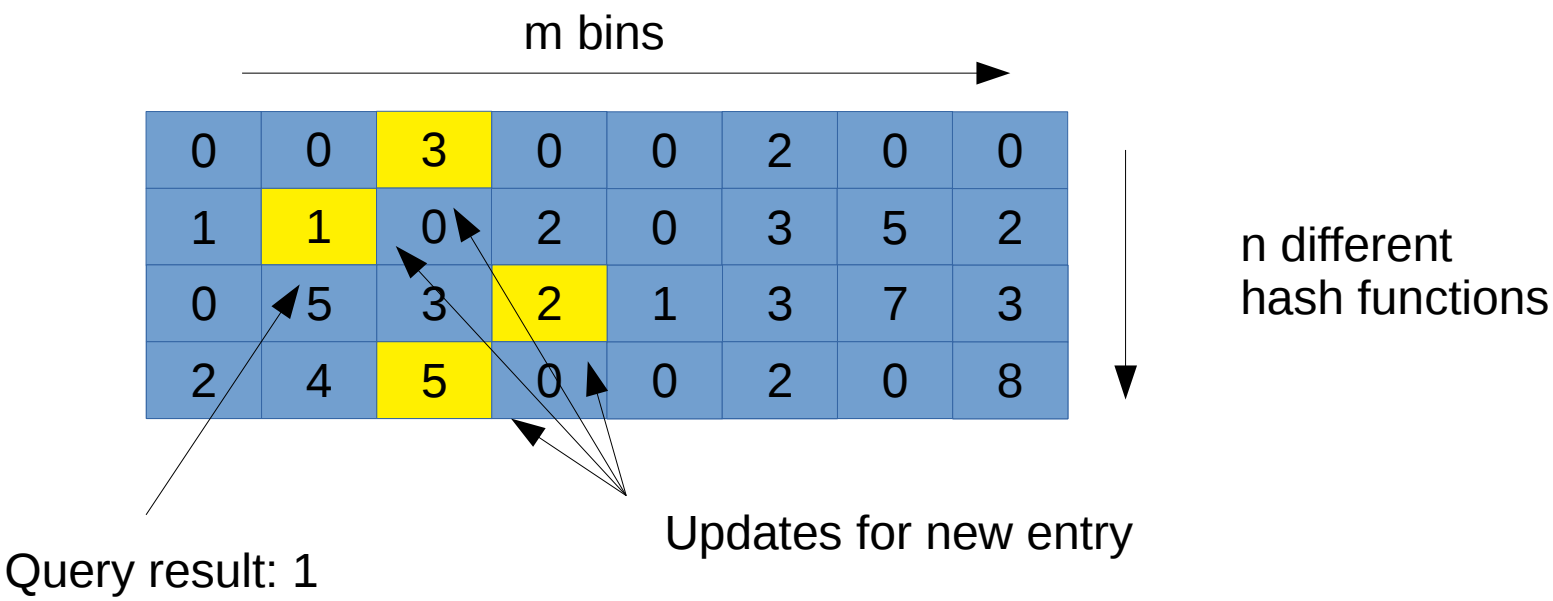
Case 2: new element



Metwally, Agrawal, Abbadi, *Efficient computation of Frequent and Top-k Elements in Data Streams, Internation Conference on Database Theory, 2005*

Count-Min Sketches

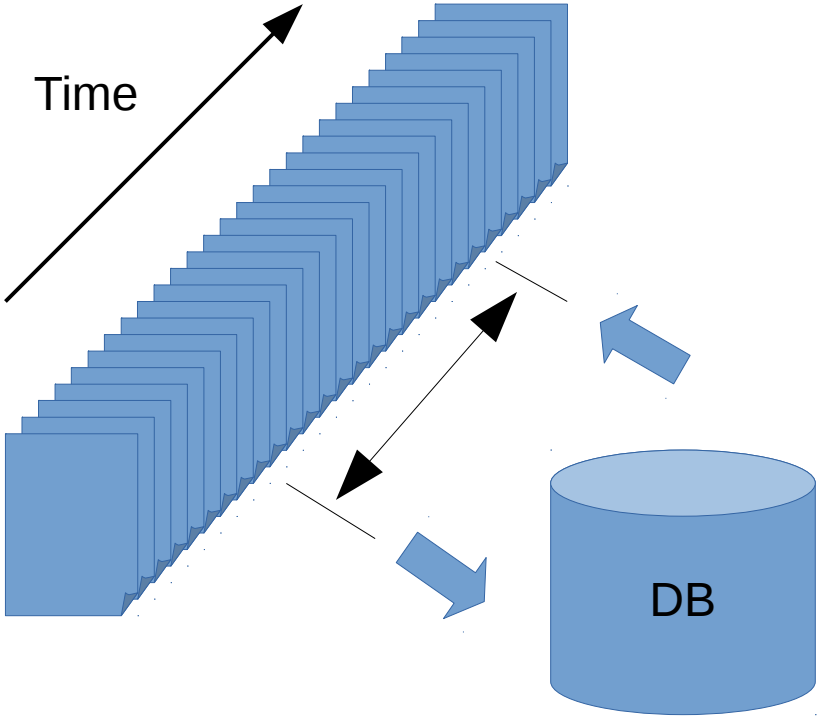
- Summarize histograms over large feature sets
- Like bloom filters, but better

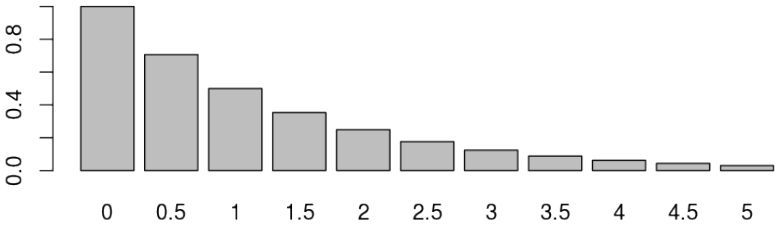


- Query: Take minimum over all hash functions

G. Cormode and S. Muthukrishnan. *An improved data stream summary: The count-min sketch and its applications.* LATIN 2004, J. Algorithm 55(1): 58-75 (2005) .

Heavy Hitters over Time-Window





- Keep old count data periodically
- Alternative: Exponential decay

Indexing Columns

page	referer	IP	score
/index.html	google	13.24.32.12	10
/post/123	facebook	43.13.43.67	9
/index.html	twitter	6.62.23.4	6
/about.html	google	13.24.32.12	3
...	...	...	..

Storing Data in Trends

c[x]=5
c[d]=2
c[a]=3

key	score
x	5
a	3
d	2

c[x]= {
a: 5, b: 1, c: 2}
c[y]= {
a: 8, d: 2}

K1	K2	score
y	a	8
x	a	5
x	c	2
y	d	2

Storing Data in Trends

$$
\begin{array}{c|ccc}
 & j=1 & 2 & 3 \\
\hline
i=1 & 5 & 0 & 3 \\
2 & 2 & 1 & 0 \\
3 & 0 & 0 & 8
\end{array}
\rightarrow
$$

i	j	score
3	3	8
1	1	5
1	3	3
2	1	2
2	2	1

column

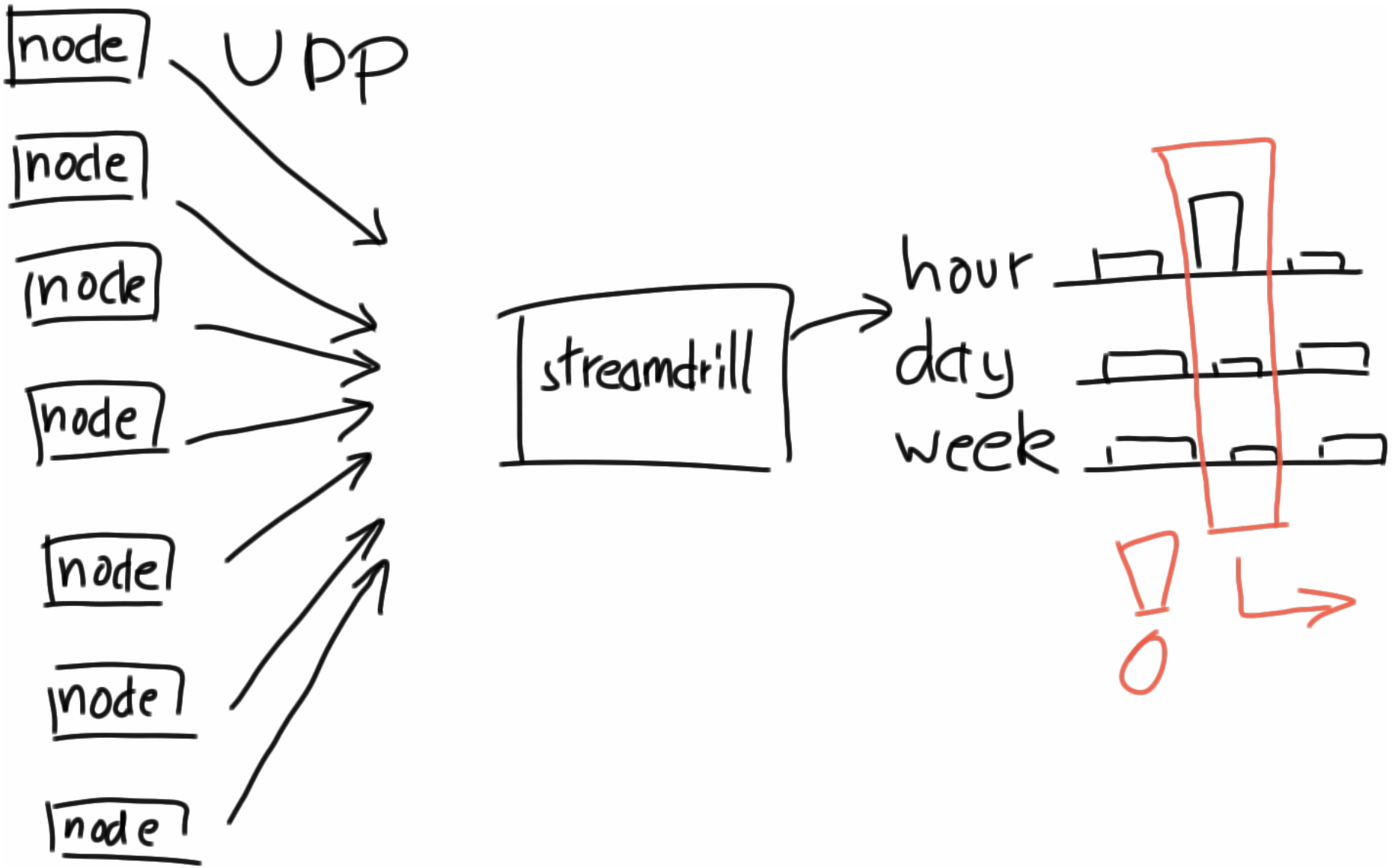
row

Streamdrill

- Realtime Analysis Solutions
- Core Engine:
  - Heavy Hitters counting + exponential decay
  - **Instant** counts & top-k results over time windows
  - In-Memory
  - written in Scala
- Modules
  - Profiling and Trending
  - Recommendations
  - Count Distinct

Realtime User Profiles

Realtime User Profiles



Realtime User Profiles

streamdrill Dashboard

1 | 1720.1m | 4.3k | 13GB | 90 | Documentation | Anonymous

« Dashboard / nuggad

Tue Apr 01 2014 16:42:10 GMT+0200 (CEST)

Real-Time User Profile *nuggad*

99%

Global Fingerprints more details »

week
day
hour

Activity by week day hour

Top 5 Sites more »

#	Site	Activity	Category Fingerprint	Top Categories
1		91,969,740.6		market place,football,lottery
2		69,572,802.1		regional,humour,ecommerce
3		59,012,708.1		car accesories,funbikes,lift truck
4		48,277,773.9		classified ad,market place
5		41,376,343.4		internal policy,obligatory,real

Top 5 Networks more »

#	Network	Activity	Category Fingerprint	Top Categories
1		94,357,907.6		jewellery,international,dating
2		76,881,622.0		high-performance cars,fashion (men),beer
3		69,572,934.9		regional,humour,ecommerce
4		59,012,818.0		car accesories,funbikes,lift truck
5		48,277,837.6		market place,classified ad

#	User	Activity	Category Fingerprint	Top Categories
		by week	low high	
1		56,726.5		
2		39,142.5		travel planning,red.content
3		38,179.7		news,red.content
4		37,251.3		news,red.content
5		34,492.0		classified ad,job offering
6		27,197.3		
7		24,856.3		lifestyle and leisure,red.content
8		23,868.5		news,red.content

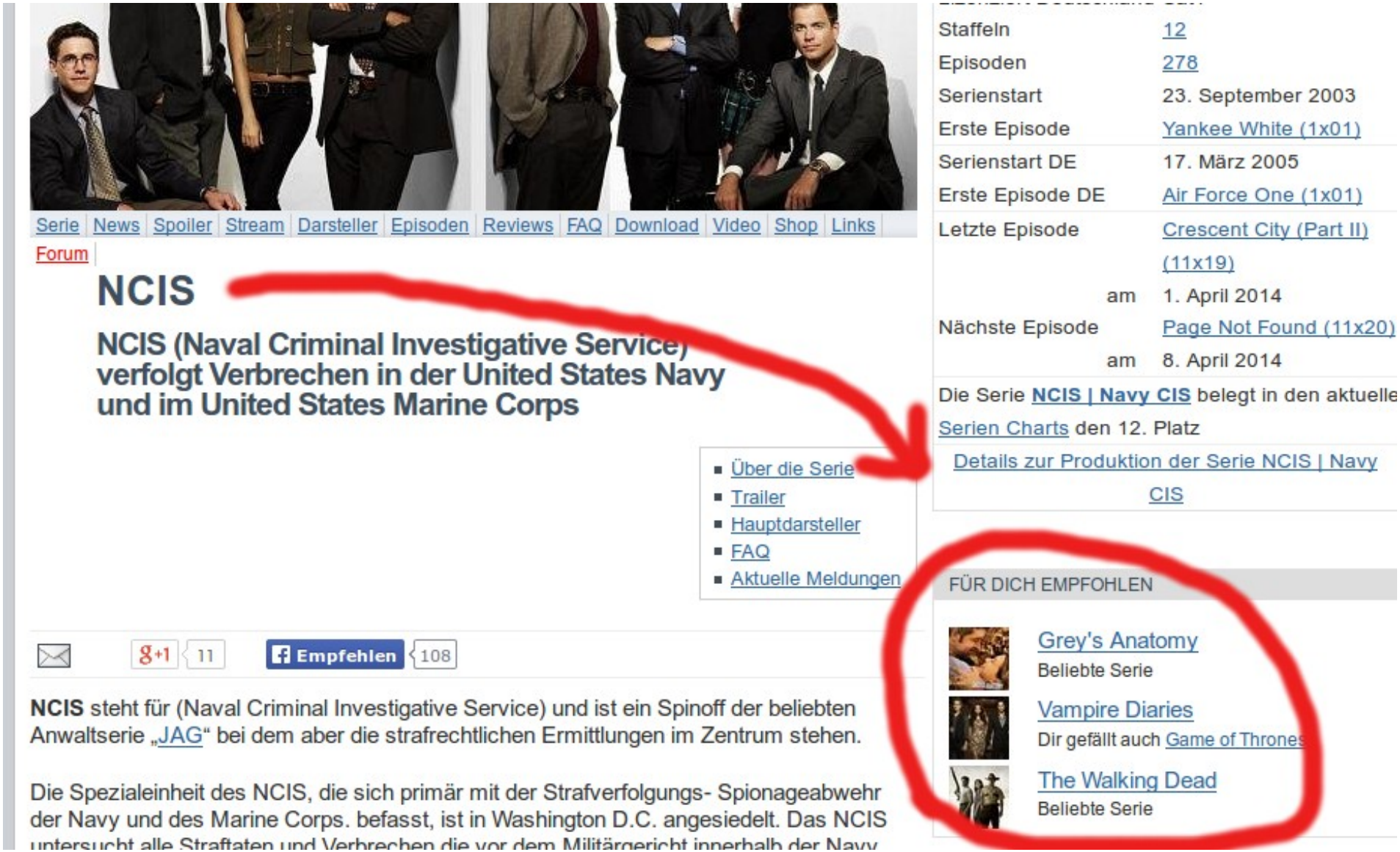
Realtime User Profiles

Category	Activity					
	week		day		hour	
chat	2.0	0.06%	1.9	0.45% ↓	0.0	
information	12.5	0.06%	9.9	0.38% ↓	0.9	
society	0.0	0.00%	0.0	1.14% ↑	0.5	
entertainment media	64.7	0.13% ↑	53.2	2.67% ↑	7.1	
television	174.5	0.15% ↓	141.2	3.12% ↓	14.1	
video gallery	253.7	0.04%	206.0	0.59% ↑	23.0	
communication boards	2.0	0.06%	1.9	0.45% ↓	0.0	

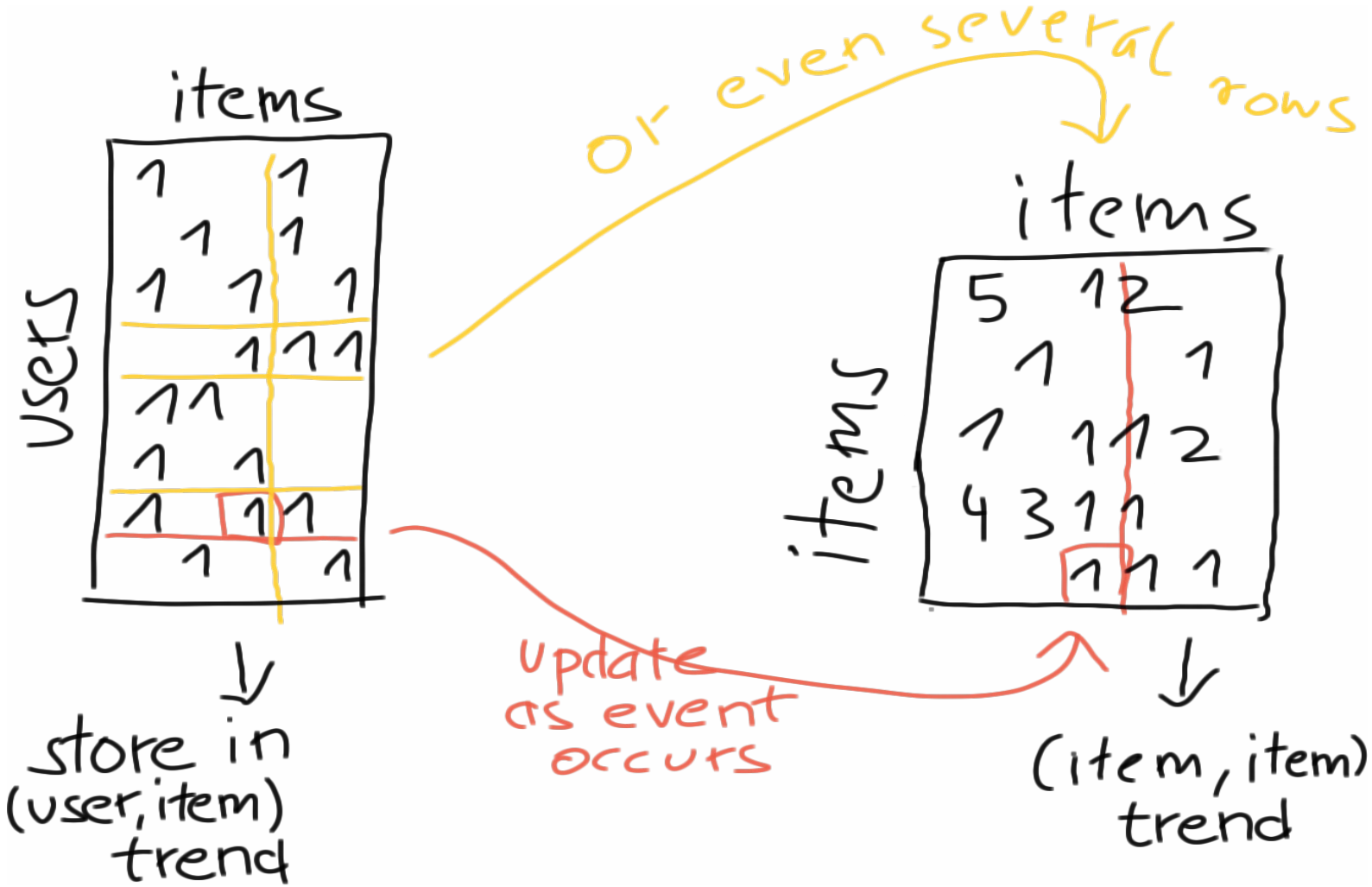
Realtime user profiles

- Process 10k events / second on one machine
- Track about 1 Million counts per 1 GB
- Shard by user for higher accuracy

Realtime Recommendation



Realtime Recommendation



Realtime Recommendation

- Pipe in events, get recommendations
- Seed by sampling past clicks
- Automatically adapts over time
- Number of alternatives (user based, item based, trends, categories)

Summary

- Ditch exactness → Approximate with stream mining
- Trends to store all kinds of counting structures
- React to realtime user behavior with managable resources