

# Data Streams:

## *Introduction*

### Mining Massive Datasets

Prof. Carlos Castillo — <https://chato.cl/teach>



Universitat  
Pompeu Fabra  
*Barcelona*

# Sources

- Mining of Massive Datasets (2014) by Leskovec et al. (chapter 4)
  - Slides [part 1](#), [part 2](#)
- Tutorial: [Mining Massive Data Streams](#) (2019) by Michael Hahsler

# What is a data stream?

- A potentially infinite sequence of data points
  - Each data point can be a tuple or vector
- Examples:
  - web click-stream data → who clicks on what
  - computer network monitoring data
  - telecommunication connection data
  - readings from sensor nets
  - stock quotes

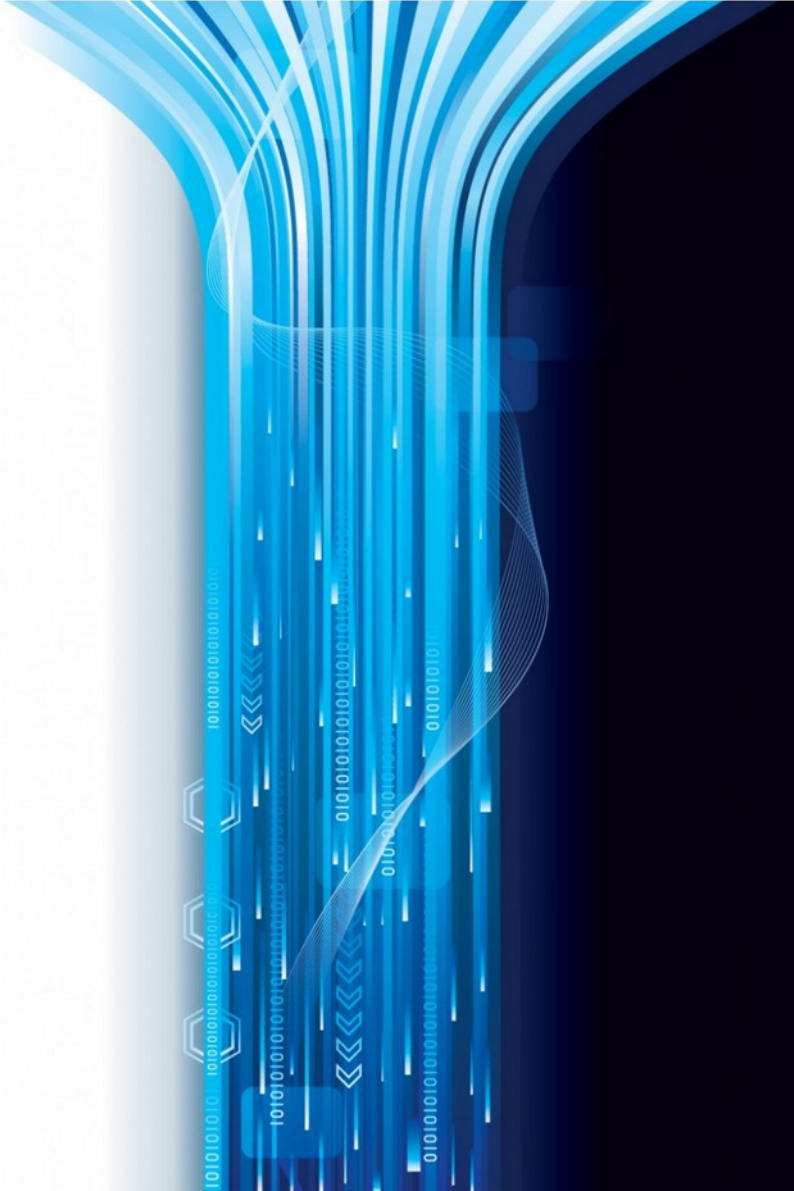
Do not confuse with “streaming,” which in vernacular typically means live video.

# Example: Apache server log

```
tecmint@TecMint ~ $ tailf /var/log/apache2/access.log
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET / HTTP/1.1" 200 729 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/blank.gif HTTP/1.1" 200 512 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/folder.gif HTTP/1.1" 200 512 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:37 +0530] "GET /icons/text.gif HTTP/1.1" 200 512 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:11:38 +0530] "GET /favicon.ico HTTP/1.1" 404 500 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:12:05 +0530] "GET /tecmint/ HTTP/1.1" 200 787 "http://127.0.0.1:80/" Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:12:05 +0530] "GET /icons/back.gif HTTP/1.1" 200 401 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /tecmint/Videos/ HTTP/1.1" 200 101 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /icons/compressed.gif HTTP/1.1" 200 101 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
127.0.0.1 - - [31/Oct/2017:11:13:58 +0530] "GET /icons/movie.gif HTTP/1.1" 200 101 Mozilla/5.0 (X11; Linux x86_64; rv:56.0) Gecko/20100101 Firefox/56.0"
```

# Key properties of data streams

- **Unbounded size**
  - Data cannot be persisted on disk
  - Only summaries can be stored
- **Transient**
  - Single pass over the data
  - Sometimes real-time processing is needed
- **Dynamic**
  - May require incremental updates
  - May require to forget old data
  - Concepts “drift”
- **Temporal order** is often important



# Applications

- **Mining query streams**
  - A search engine wants to know what queries are more frequent today than yesterday
- **Mining click streams**
  - A newspaper wants to know when one of its pages starts getting an unusual number of hits per hour
- **Mining social network news feeds**
  - A social media platform wants to show trending topics

# Applications (cont.)

- **Sensor Networks**
  - Many sensors feeding into a central controller
- **Telephone call records**
  - Data feeds into customer bills as well as settlements between telephone companies
- **IP packets monitored at a switch**
  - Gather information for optimal routing
  - Detect denial-of-service attacks

# Why not simply use a relational DB?

<b>Relational DBMS</b>	<b>DSMS (Stream)</b>
persistent relations	transient streams
only current state is important	history matters
not real-time	real-time
low update rate	stream!
one time queries	continuous queries

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom (2002). Models and issues in data stream systems. In PODS '02, pages 1–16, ACM Press.



# Why do we need **new algorithms**?

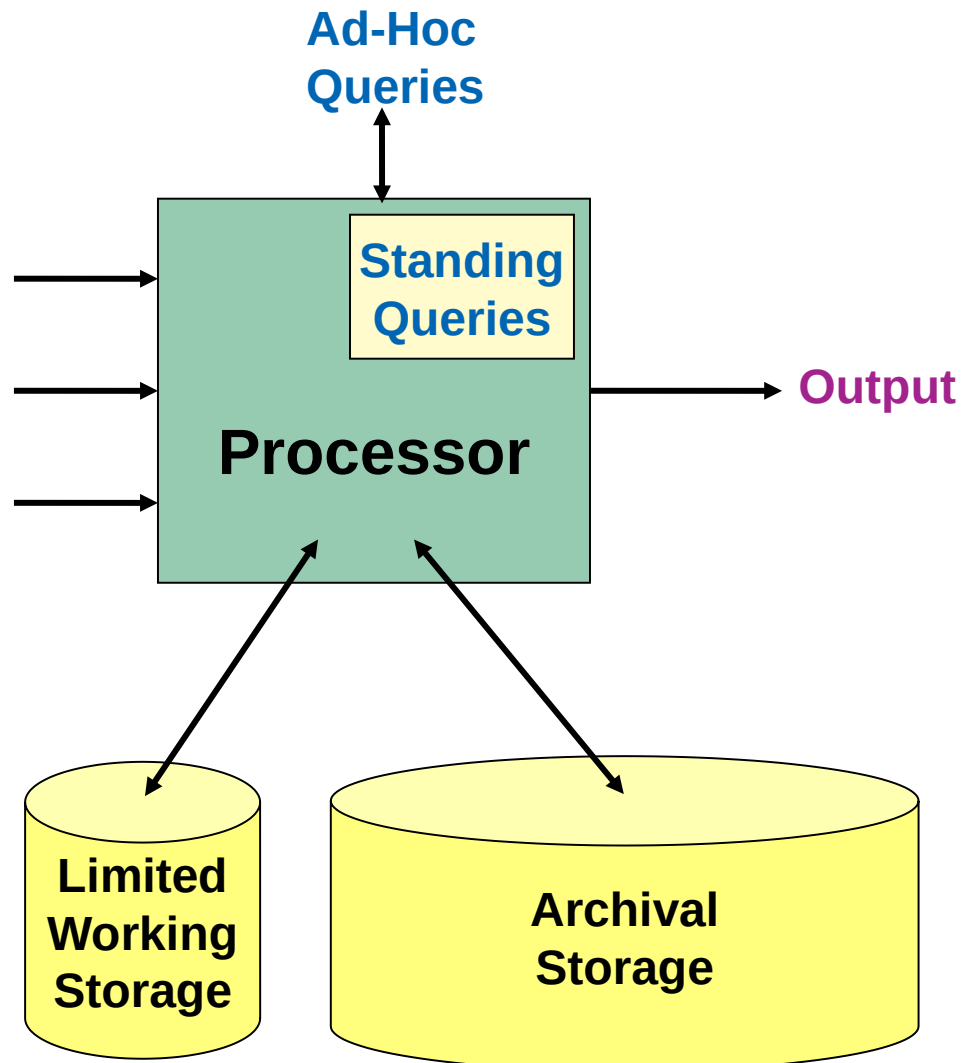
	<b>Traditional</b>	<b>Stream</b>
<b>passes</b>	multiple	single
<b>processing time</b>	unlimited	restricted
<b>memory</b>	disk	main memory
<b>results</b>	typically accurate	approximate
<b>distributed</b>	typically not	often

**Source:** Joao Gama, Data Stream Mining Tutorial, ECML/PKDD, 2007

# A generic stream-processing architecture

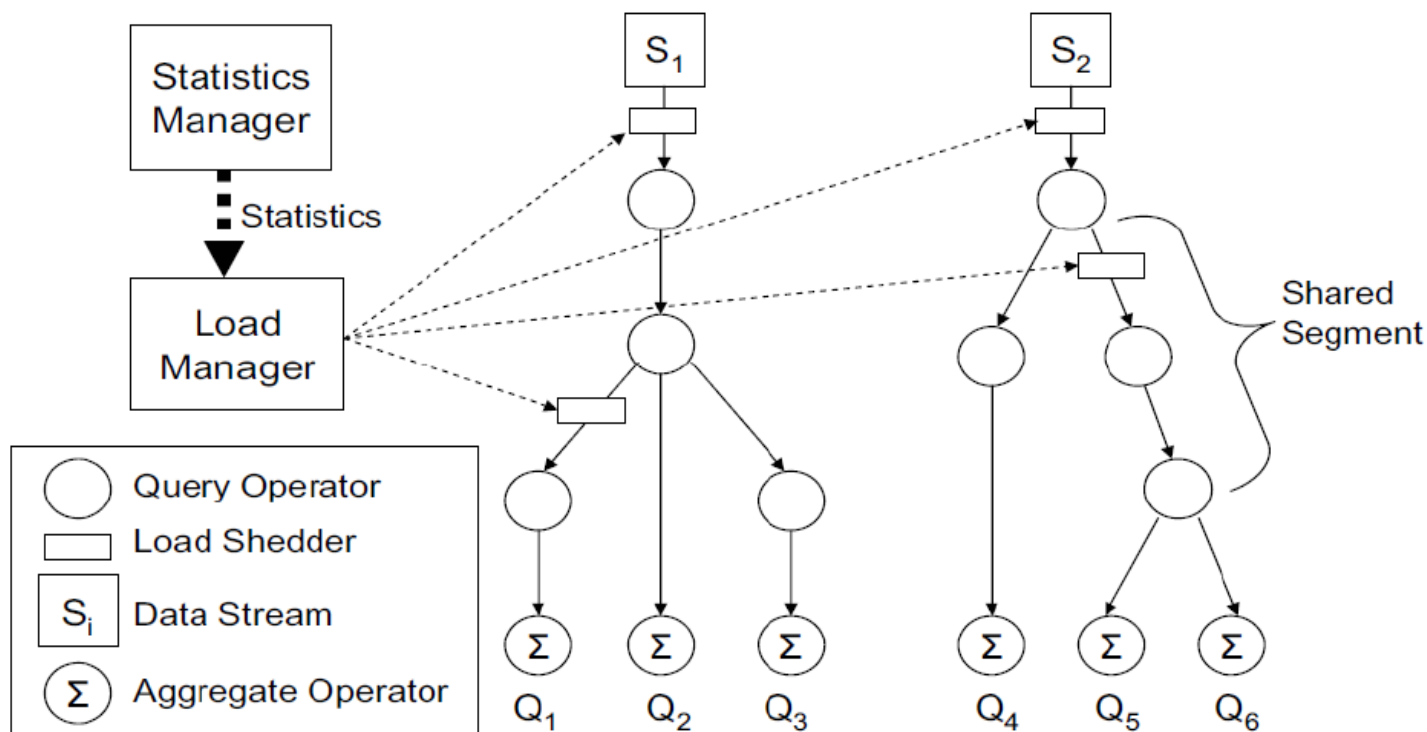
**Input streams**  
Each stream is composed of elements/tuples

... 1, 5, 2, 7, 0, 9, 3  
... a, r, v, t, y, h, b  
... 0, 0, 1, 0, 1, 1, 0  
← time



# Load shedding

# Too much data? Ignore some of it



# Sampling a fixed proportion

# Sampling a fixed proportion

- Example stream: `<user, query, timestamp>`  
from a search engine query log
- Suppose we have space to store  $1/r$  of the stream
  - E.g.: 1/10th, 1/100th, 1/1000th,
- Naïve solution:
  - Generate uniform random number in  $0 \dots (r-1)$   
`numpy.random.uniform(0, r)`
  - If the number is 0, keep the item

# What can we do with this sample?

- Approximate most frequent query
  - Pick the most frequent in the sample
- Approximate frequency of a query
  - Multiply observed frequency by  $r$
- Do people ask query  $q$ ?
  - Approximate answer (with some prob. of error)

# Exercise

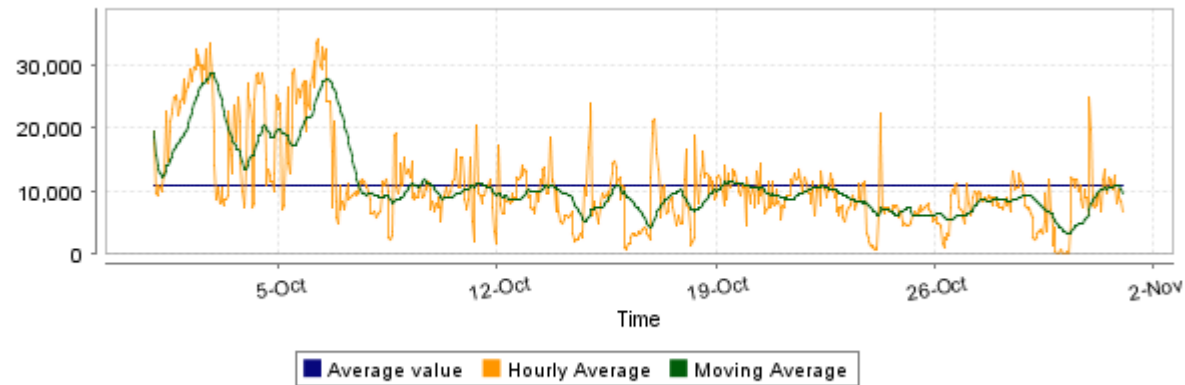
Answer in  
Nearpod Collaborate

- We want to tell if we have seen item  $q$
- Suppose we have seen  $n$  items
- Suppose we have sampled a fraction  $1/r$
- Suppose item  $q$  appears with probability  $p(q)$
- What is the probability of a:
  - False Positive? (*Item  $q$  was not in the stream but we said it was*)
  - False Negative? (*Item  $q$  was in the stream but we said it was not*)



# What can we do with this...? (cont.)

- Approximate num. queries per minute



- Peak frequency
  - Multiply observed peak by  $r$

# But there are questions we cannot answer well

- What fraction of queries by an average search engine user are duplicates?

- Suppose each user issues  $x$  queries once and  $d$  queries twice (total of  $x+2d$  queries)

- Correct answer:  $d/(x+d)$

- Proposed solution: We keep  $1/10^{\text{th}}$  of the queries ( $r=10$ )

- Sample will contain  $x/10$  of the singleton queries at least once

- Sample will contain  $2d/10$  of the duplicate queries at least once


- Sample will contain  $d/100$  pairs of duplicates

- $d/100 = 1/10 \cdot 1/10 \cdot d$

- Of the  $d$  duplicates,  $18d/100$  will be seen once\*

- $18d/100 = ((1/10 \cdot 9/10) + (9/10 \cdot 1/10)) \cdot d$

- So the sample-based answer is


$$\frac{\frac{x}{10} + \frac{18d}{100} + \frac{d}{100}}{\frac{d}{100}} = \frac{d}{10x + 19d}$$

\* Copy A is in the selected part, copy B in the unselected part, or viceversa

# But there are questions we cannot answer well (cont.)

- What fraction of queries by an average search engine user are duplicates?
  - Suppose each user issues  $x$  queries once and  $d$  queries twice (total of  $x+2d$  queries)
  - **Correct answer:**  $d/(x+d)$

- Proposed solution: We keep  $1/10^{\text{th}}$  of the queries ( $r=10$ )
  - Sample will contain  $x/10$  of the singleton queries at least once
  - Sample will contain  $2d/10$  of the duplicate queries at least once
  - Sample will contain  $d/100$  pairs of duplicates
    - $d/100 = 1/10 \cdot 1/10 \cdot d$
  - Of the  $d$  duplicates,  $18d/100$  will be seen once
    - $18d/100 = ((1/10 \cdot 9/10) + (9/10 \cdot 1/10)) \cdot d$

- So the sample-based answer is

$$\begin{array}{c}
 \text{Observed duplicates} \\
 \frac{\frac{x}{10} + \frac{18d}{100} + \frac{d}{100}}{\text{Observed singletons} \quad \text{Observed duplicates}} = \frac{d}{10x + 19d}
 \end{array}$$

**WRONG!**

# How do we solve it?

- We need to **sample  $1/r$  of users** and all of their actions

How do we do this?

```
<user1, action,
timestamp>
<user2, action,
timestamp>
<user2, action,
timestamp>
<user3, action,
timestamp>
<user1, action,
timestamp>
<user2, action,
```

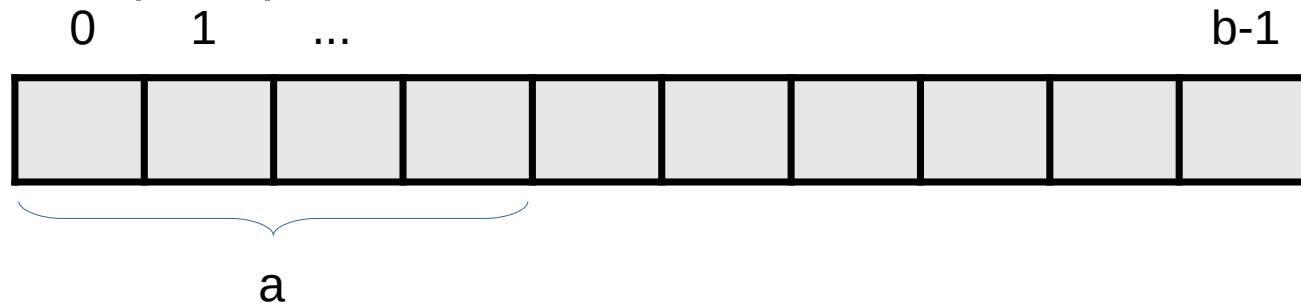
# How do we solve it?

- We need to **sample  $1/r$  of users** and all of their actions
- How do we do this?
  - **Hashing!**
  - Given  $\langle \text{user}, \text{action}, \text{timestamp} \rangle$
  - Compute  $h(\text{user}) \rightarrow 0, 1, \dots, (r-1)$
  - Keep tuple if hash value is 0

```
<user1, action,
timestamp>
<user2, action,
timestamp>
<user2, action,
timestamp>
<user3, action,
timestamp>
<user1, action,
timestamp>
user2, action
```

# In general ...

- To sample a fraction  $a/b$  of a stream by key
- Compute  $h(\text{key}) \rightarrow 0, 1, \dots, (b-1)$
- Keep if  $h(\text{key}) < a$



# Summary

# Things to remember

- What is a data stream
- How to sample a fixed percentage of values grouped by a key, using hashing



# Exercises for TT22-T26

- Mining of Massive Datasets (2014) by Leskovec et al.
  - Exercises 4.2.5
  - Exercises 4.3.4
  - Exercises 4.4.5
  - Exercises 4.5.6