Streaming Algorithms

- Streaming Model of Computation
- Streaming Algorithms and DFA
- Stream: Motivation and Applications
- Synopsis: Sliding Window, Histogram, Wavelets
- Sampling from Stream: Reservoir Sampling
- Linear Sketch
- Count-Min Sketch
- AMS Sketch

Imdad ullah Khan

Randomized Stream Computation Model

Randomized Stream Computation Model

Since streaming algorithms have limited memory, exact algorithms are possible only for a few problems

We seek randomized approximate solutions

(ϵ, δ)-approximate algorithm

- Stream $\mathcal{S} := a_1, a_2, a_3, \ldots, a_m$ ▷ m may be unknown
- $f(\mathcal{S})$: Desired/Optimal output ▷ (a function of stream)
- $\mathcal{A}$: an algorithm to approximate $f(\mathcal{S})$
- $\mathcal{A}(\mathcal{S})$: output of $\mathcal{A}$ on $\mathcal{S}$

For $\epsilon > 0$, $0 \leq \delta \leq 1$, $\mathcal{A}$ is an (ϵ, δ)-approximation algorithm if

$$Pr\big[\, |\mathcal{A}(\mathcal{S}) - f(\mathcal{S})| \, > \, \epsilon f(\mathcal{S}) \,\big] \, \leq \, \delta$$

Randomized Stream Computation Model

Stream $\mathcal{S} := a_1, a_2, a_3, \ldots, a_m$ ▷ m may be unknown

Each $a_i \in [n]$ ▷ e.g., primary keys of complex data objects

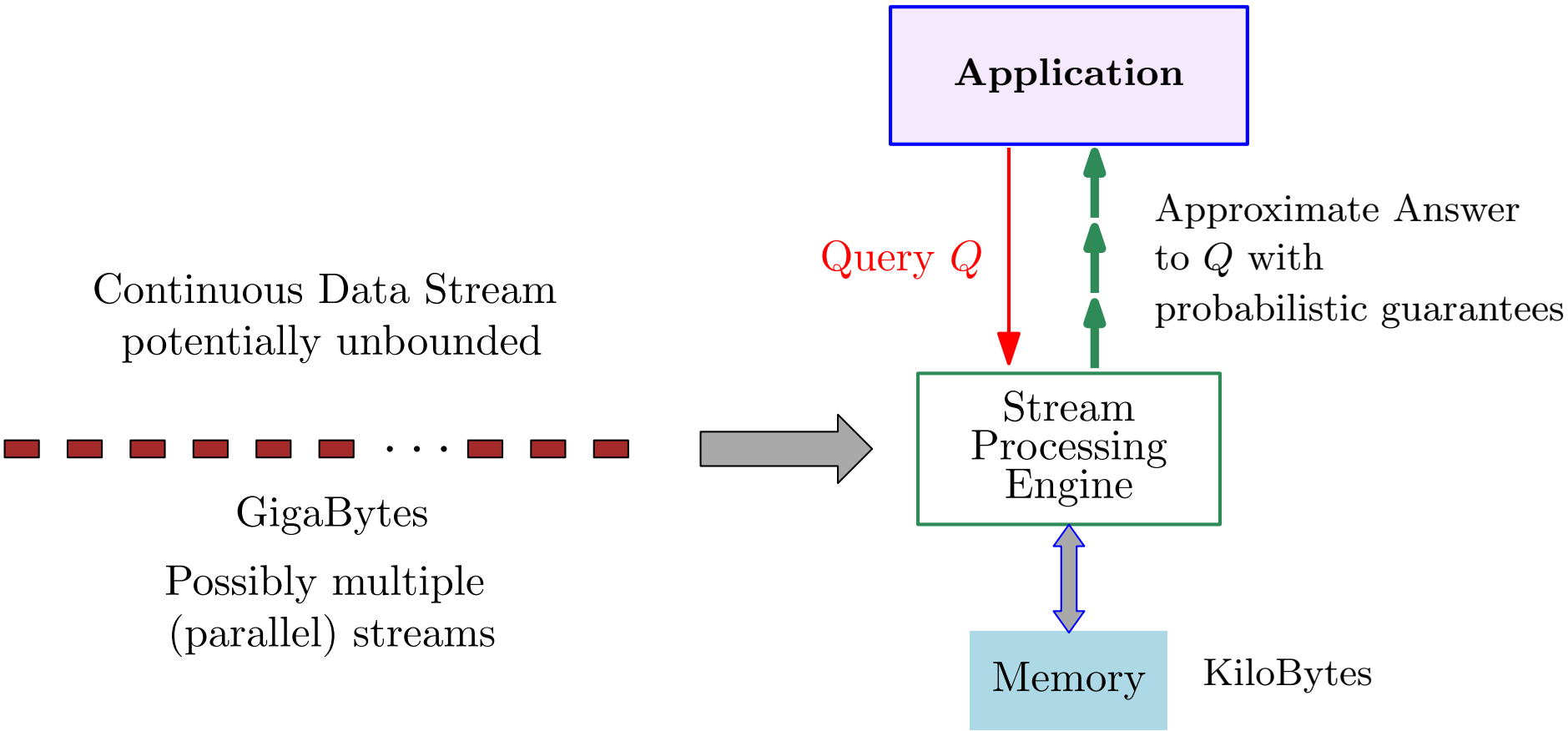
Goal: Compute a function of the stream $\mathcal{S}$

▷ e.g. mean, median, number of distinct elements, frequency moments,. . .

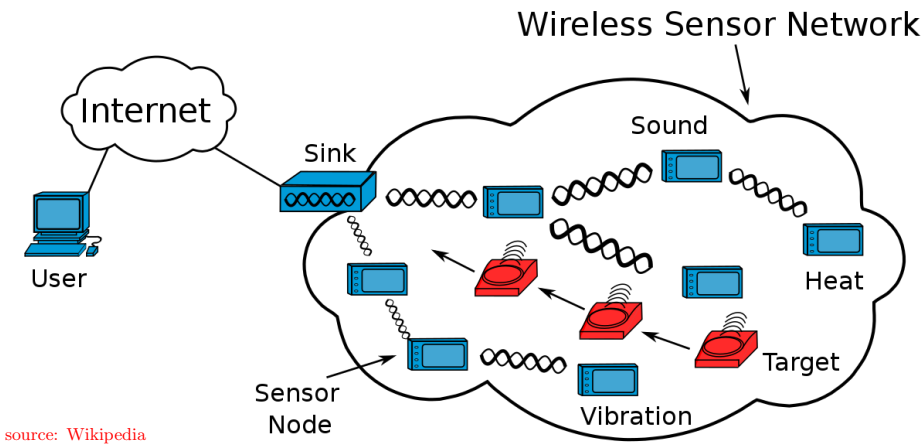
Subject to

- Single pass, read each element of $\mathcal{S}$ only once sequentially
- Per item processing time $O(1)$
- Use memory polynomial in $O(1/\epsilon, 1/\delta, \log n)$
- Return (ϵ, δ)-randomized approximate solution

Randomized Stream Computation Model



Randomized Stream Computation Applications: Sensor Networks



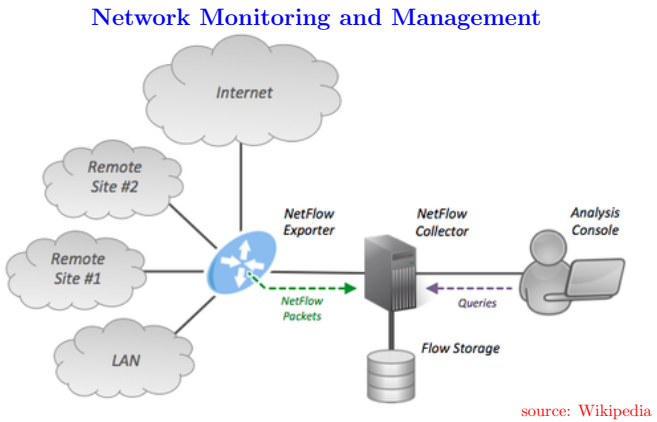
- Sensor nodes collect unlimited amount of data
- Have very limited computation power and memory
- Limited battery power constrain communication of all collected data

  1 bit transmission consumes power $\sim$ to executing 800 instructions[1]

Streaming algorithm deployed onto nodes are ideally suited for drawing analytics from sensed data

[1] Madden et.al. (2002)

Randomized Stream Computation Applications: Network Monitoring



NetFlow: A Cisco tool for network administrators (performance metrics, security analysis, detection and forensics). For each flow it reports (logs)

- Network Interface
- Source/Destination IP Addresses
- IP Protocol
- Source/Destination port
- TCP Flags
- Total packets/bytes in flow

AT&T processes over 567 billion flow records per day[2] ▷ $\sim$ 15 PBytes

Detects and characterizes approximately 500 anomalies per day

[2] Fred Stinger (AT&T) FloCon (2017) Netflow Collection and Analysis ..

Randomized Stream Computation Applications: Network Monitoring

Application Area

- Traffic engineering
- Traffic monitoring
- Volume estimation & analysis
- Load balancing
- Efficient resource utilization
- (D)DOS attack detection
- SLA violation

Queries

- How many bytes sent b/w IP-1 and IP-2?
- How many IP addresses are active?
- Top 100 IP's by traffic volume
- Average duration of IP session?
- Median number of bytes in each IP session
- Find sessions that transmitted $> 1k$ bytes
- Find sessions with duration $>$ twice average
- List all IP's with a sudden spike in traffic
- List all IP involved in more than 1k sessions

Randomized Stream Computation Applications: Click Stream

Tracking and analysis of websites visits

Figure credit: Alex Smola @Yahoo research & ANU

- Stream of user clicks on websites (tracked via cookies)
- Find hot links, frequent IP's, click probability
- Enhanced customer experience & conversion rates
- Digital marketing – Up-selling and cross-selling

Randomized Stream Computation Applications: Search Queries

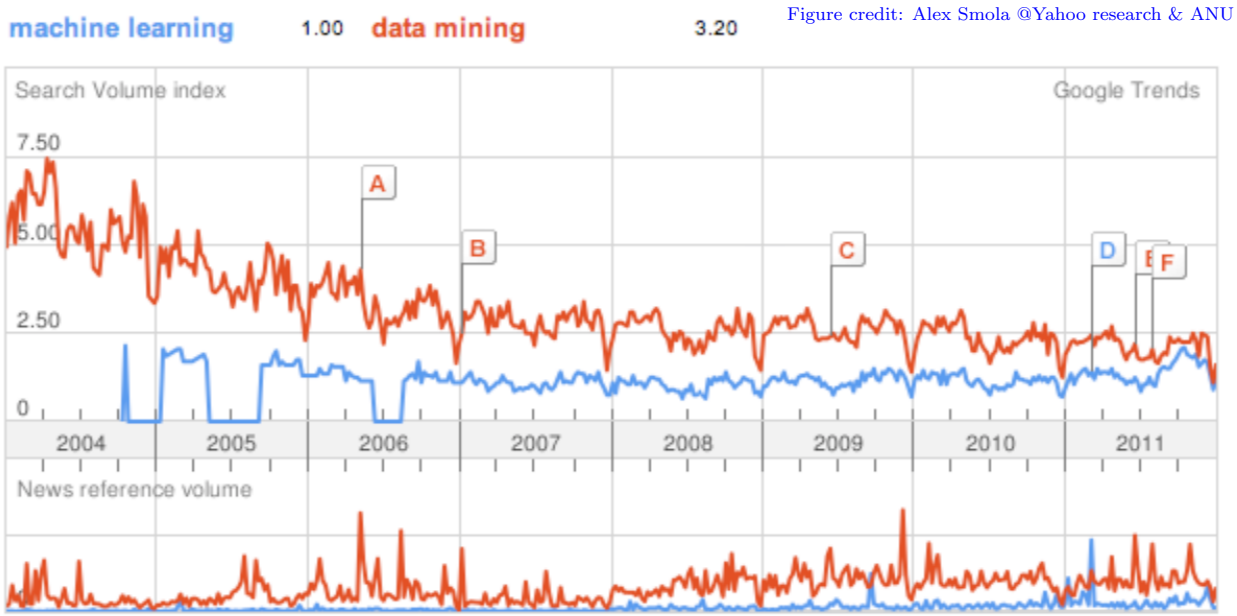
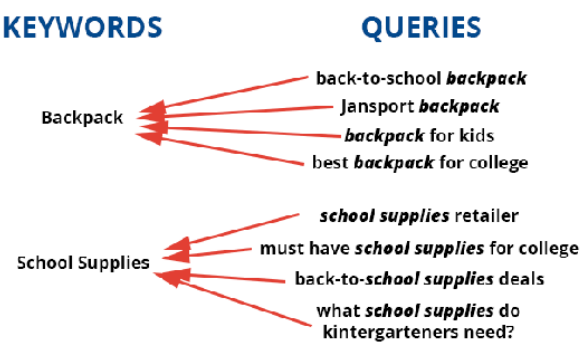


Figure credit: Alex Smola @Yahoo research & ANU

- Discover trends and patterns
- Relevant keywords for website
- Estimate competition scores or difficulty
- Estimate keywords CPC (cost per click)



Randomized Stream Computation Applications: AMI Data Analysis

Energy consumption Analysis

- Electricity consumption data from AMI (Automatic Metering Interface)
- Find average hourly load, load surges, anamoly
- Short term load forecast (total or for individual consumer)
- Identify faults, drops, failures

Randomized Stream Computation Applications: Financial Time Series

- Time stamped real time (multiple) stock data
- Need near real time prediction
- Algorithmic Trading

Randomized Stream Computation Applications: Query Execution

Query Execution Plan can be optimized using a synopsis of the database

Suppose we have data of $n = 1M$ people in a database and the query SELECT * from Table WHERE $25 \leq age \leq 35$ and $54 \leq weight \leq 60$

▷ **Runtime of brute force execution** is $2n$ comparisons

Suppose we have the following synopsis of distribution of attributes

Age	Freq
$0-10$	7%
$11-20$	8%
$21-30$	10%
$31-40$	12%
$41-50$	13%
$51-60$	25%
$61-70$	20%
$71+$	5%

First filter on Age, then on weight

Runtime: $1.22n$

Weight	Freq.
$0-20$	20%
$21-40$	25%
$41-60$	10%
$61-80$	15%
$81+$	30%

First filter on Weight, then on age

Runtime: $1.1n$