CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 16

Samson Zhou

Last Time: Sparse Recovery

- Suppose we have an insertion-deletion stream of length $m = \Theta(n)$ and at the end we are promised there are at most k nonzero coordinates
- Goal: Recover the k nonzero coordinates and their frequencies

Last Time: Sparse Recovery

- Suppose at the end we are promised there are at most k nonzero coordinates

- Algorithm: Keep $2k$ running sum of different linear combinations of all the coordinates

- We have $2k$ equations and $2k$ unknown variables

- Correctness can be shown (not quite linear algebra)

Last Time: Sparse Recovery

- Suppose at the end we are promised there are at most k nonzero coordinates
- Algorithm: Keep $2k$ running sum of different linear combinations of all the coordinates
- Space: $O(k)$ words of space

Previously: Chebyshev's Inequality

- Let X be a random variable with expected value $\mu := \mathrm{E}[X]$ and variance $\sigma^2 := \mathrm{Var}[X]$

- $\Pr[|X - \mathrm{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$ becomes $\Pr[|X - \mathrm{E}[X]| \geq t] \leq \frac{\sigma^2}{t^2}$

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

- "Bounding the deviation of a random variable in terms of its variance"

Distinct Elements (F_0 Estimation)

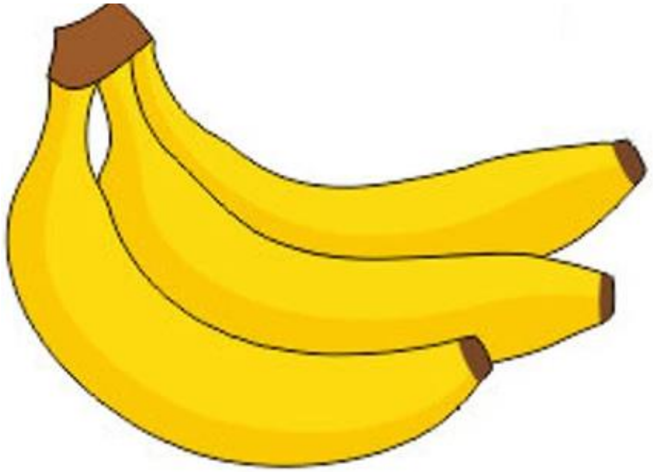
- Given a set S of m elements from $[n]$, let f_i be the frequency of element i. (How often it appears)
- Let F_0 be the frequency moment of the vector:

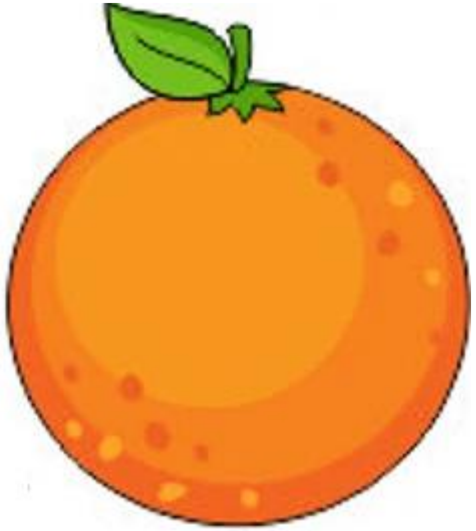
$$F_0 = |\{i : f_i \neq 0\}|$$

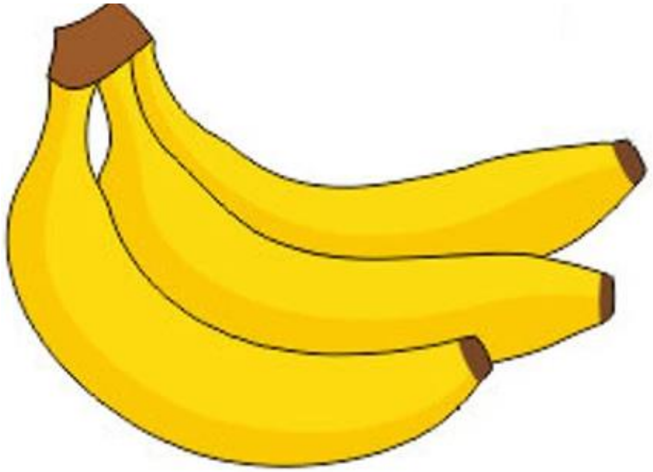
- Goal: Given a set S of m elements from $[n]$ and an accuracy parameter ε, output a $(1+\varepsilon)$-approximation to F_0

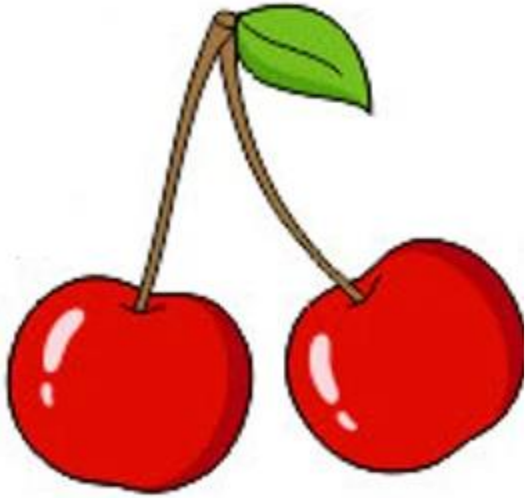


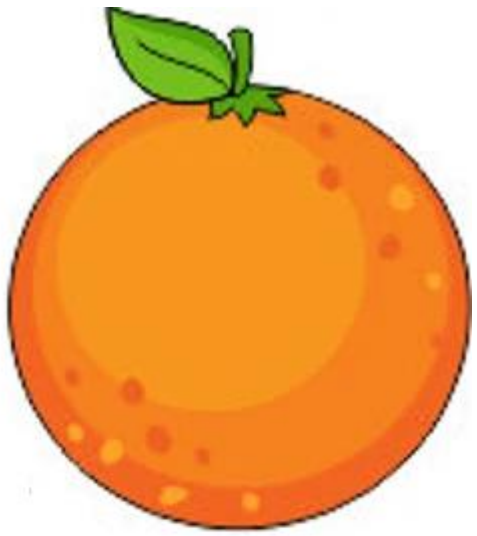


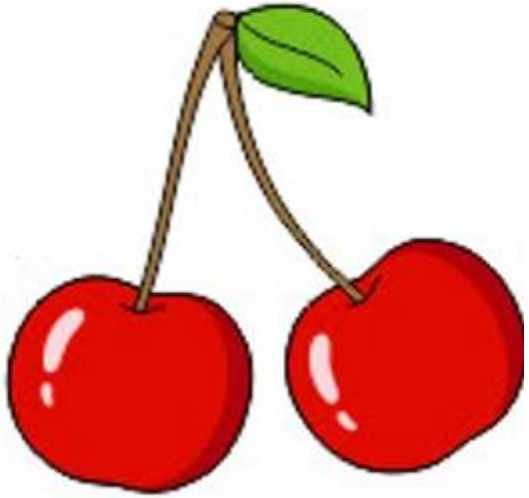


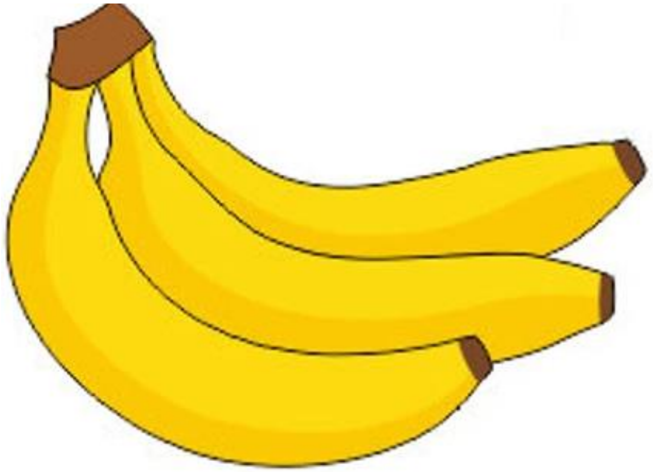


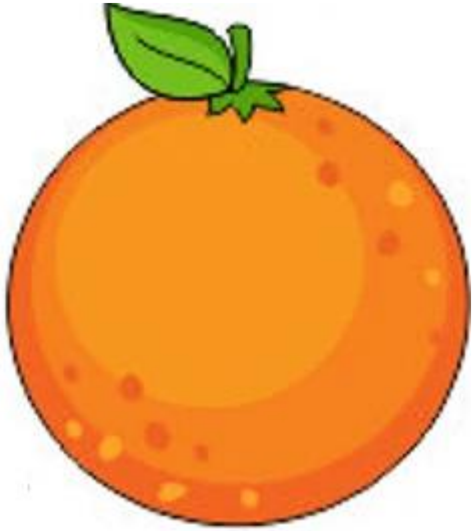


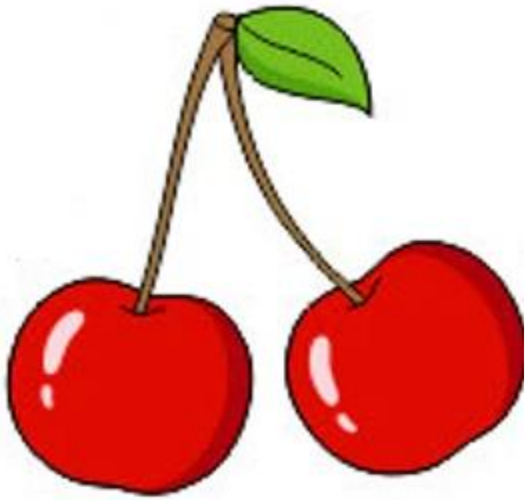


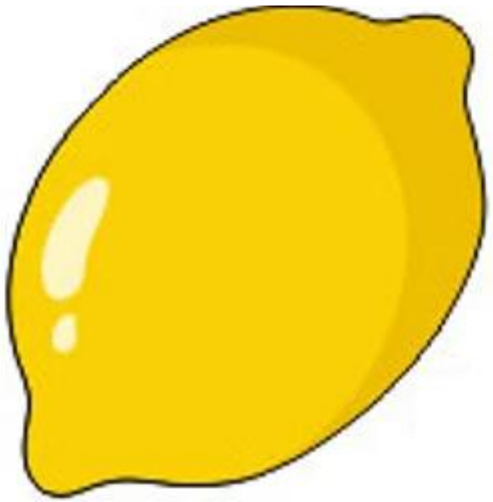











Distinct Elements (F_0 Estimation)

- How many different fruits left in fruit basket?

Distinct Elements (F_0 Estimation)

- How many different fruits left in fruit basket? 8

Distinct Elements (F_0 Estimation)

- Ad allocation: Distinct IP addresses clicking an ad

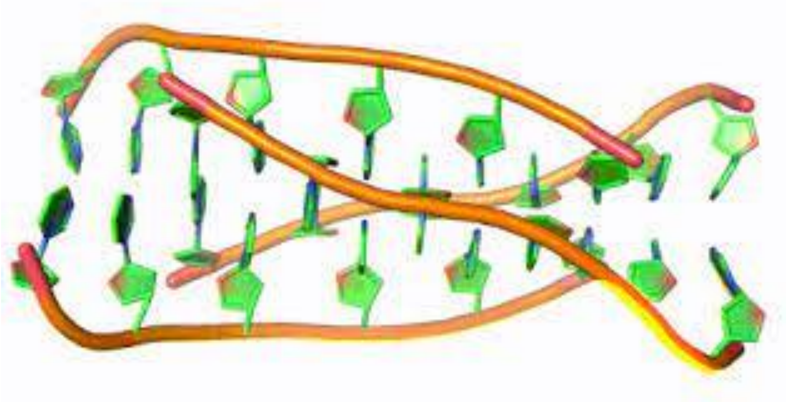
Distinct Elements (F_0 Estimation)

- Traffic monitoring: Distinct IP addresses visiting a site or number of unique search engine queries

3 billion monthly active users

Distinct Elements (F_0 Estimation)

- Computational biology: Counting number of distinct motifs in DNA sequencing



- Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers
- Suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$
- How many numbers are in S'?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers

- Suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$

- Can we use S' to get a good estimate of N?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers, suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$

- We have $\mathrm{E}[|S'|] = \frac{N}{2}$ and $\mathrm{Var}[|S'|] \leq \frac{N}{2}$

Distinct Elements (F_0 Estimation)

- What can we say about $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq t\right]$?

- By Chebyshev's inequality, we have $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq 100\sqrt{N}\right] \leq \frac{1}{10}$

Distinct Elements (F_0 Estimation)

- What can we say about $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq t\right]$?
- By Chebyshev's inequality, we have $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq 100\sqrt{N}\right] \leq \frac{1}{10}$
- With probability at least $\frac{9}{10}$,

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100\sqrt{N}$$

Distinct Elements (F_0 Estimation)

- With probability at least $\frac{9}{10}$,

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100\sqrt{N}$$

- Thus with probability at least $\frac{9}{10}$,

$$N - 200\sqrt{N} \leq 2|S'| \leq N + 200\sqrt{N}$$