DSC 257R - UNSUPERVISED LEARNING

MAINTAINING A RANDOM SUBSET

SANJOY DASGUPTA, PROFESSOR



COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE



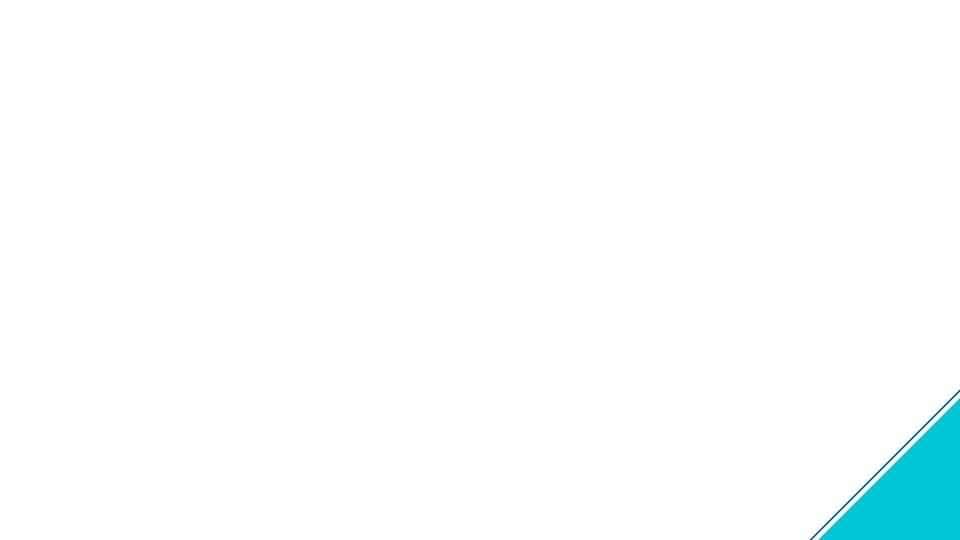
Useful Primitive: Maintaining a Random Subset

Goal: Keep a random sample of k of the points encountered so far.

Let's start with the k = 1 case:

- $s = x_1$ (this is our random sample)
- For t = 2, 3, ...:
 - Get x_t
 - Update s:

Why is this correct?



Maintaining k Random Samples (With Replacement)

- $s_1 = s_2 = \cdots = s_k = x_1$
- For t = 2, 3, ...:
 - Get x_t
 - For j = 1 to k:
 - With probability 1/t: Set $s_j = x_t$

Why is this correct?

Maintaining k Random Samples (Without Replacement)

- \bullet $(s_1, s_2, ..., s_k) = (x_1, x_2, ..., x_k)$
- For t = k + 1, k + 2, ...:
 - Get x_t
 - Update:

Why is this correct?

Approximate Median

- lacktriangle Maintain a random sample of k elements with replacement
- At any time t: let m_t be the median of these k elements

Here's what we can show using large deviation bounds:

Claim. Pick any $0 < \delta, \epsilon < 1$. If

$$k \ge \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$$
,

then for any time t, with probability at least $1 - \delta$, the value m_t is a $(1/2 \pm \epsilon)$ -fractile of $\{x_1, ..., x_t\}$.