Weeks 14–15 Summary: Data Streams (FM, AMS, and DGIM)

Z3003 Data Mining: Scaling to Large Data (Part 2 · Dr. Mahendran V)

1. Distinct-Element Counting (F_0) and Second-Moment Sketches (F_2)

When a stream exceeds RAM, probabilistic sketches process each arrival in $O(1)$ space:

- **Flajolet-Martin:** Hash items to bitstrings and track maximum trailing-zero run R; 2^R estimates F_0. Taking the median of small group means stabilizes variance.
- **AMS Sketch:** Sample random stream positions t and maintain exact subsequent counts c_t; $n(2c_t - 1)$ estimates the second frequency moment $F_2 = \sum_x m_x^2$ unbiasedly.

2. Sliding-Window Bit Counting (DGIM)

DGIM partitions the active N-bit window into non-overlapping buckets of exponentially growing sizes $1, 2, 4, \ldots$, merging whenever three buckets share the same size. Using $O(\log^2 N)$ space, summing all complete buckets plus half of the oldest bucket guarantees relative error $\leq 50\%$.