Week 9 Summary: Clustering at Scale, End-to-End Pipeline, and Part 2 Handover

Z3003 Data Mining: Scaling to Large Data · 11–13 November 2026 · Dr. Innocent Nyalala

Lecture 24: Mini-Batch k-Means, BIRCH, and DBSCAN

To break k-means memory and spherical-shape limits:

- **Mini-Batch k-Means:** Updates centroids from streaming batches ($\approx 1\%$ memory on 10^6 farms).
- **BIRCH:** Compresses points in one pass into Clustering Features CF $= (N, \mathrm{LS}, \mathrm{SS})$ that merge additively ($200,000$ farms summarized into 11 CFs in 2.3 s).
- **DBSCAN:** Discovers arbitrary-shape clusters via ε-neighborhood density ($\geq$ minPts core points, border points, and isolated noise points).

Lectures 25–26: End-to-End Synthesis and Handover to Part 2

Every scalable method relies on five design patterns: sampling, summarization, pruning, fewer passes, and parallelism. On the end-to-end failed-season pipeline (25 gauge errors of 9999 mm cleaned; $8,111$ missing-yield rows dropped), thresholding at the $9,000$-visit budget achieves 53% recall, 33% precision, and 0.83 ROC-AUC, while fairness checks verify recall across all five Zanzibar regions prior to Part 2 with Dr. Mahendran V.