# Week 3 Summary: Normalization, Discretization, Sampling, PCA, and Warehousing

Z3003 Data Mining: Scaling to Large Data · 30 Sep – 2 Oct 2026 · Dr. Innocent Nyalala

## Lecture 7: Normalization, Leakage Prevention, and Entropy Discretization

Distance- and variance-based methods ($k$-NN, $k$-means, PCA) require comparable attribute scales. Always fit normalization parameters strictly on training data before transforming test records:

- **Min-Max:** $v' = (v - \min)/(\max - \min)$; **Z-score:** $z = (v - \mu)/\sigma$; **Modified Z-score:** $M = 0.6745\,(v - \text{median})/\text{MAD}$ (flag $|M| > 3.5$). On streaming data, Welford's one-pass recurrence updates mean and variance exactly.
- **Supervised Discretization:** Choose the numeric midpoint $t$ maximizing Information Gain $\text{IG} = H(S) - \left[\frac{n_L}{n}H(S_L) + \frac{n_R}{n}H(S_R)\right]$, where $H(S) = -\sum p_i \log_2 p_i$.

## Lecture 8: Sampling at Scale and Principal Component Analysis (PCA)

Sampling reduces rows ($n_h = nN_h/N$ for stratified allocation; standard error $\text{SE} = \sigma/\sqrt{n}$; reservoir sampling keeps item $t$ with probability $k/t$ in $O(k)$ stream memory). PCA reduces columns via eigenvectors of the standardized covariance matrix $\Sigma v = \lambda v$. Standardizing before PCA is essential: on raw Zanzibar farm tables, unscaled revenue produced a spurious 99.8% first component, whereas standardized columns yielded $42\%, 20\%, 20\%, 16\%, 1\%$.

## Lecture 9: Data Warehousing Concepts and Architecture

Following Inmon, a warehouse is Subject-oriented, Integrated, Time-variant, and Nonvolatile (S-I-T-N), separating analytical OLAP queries from live transactional OLTP systems via a three-tier ETL pipeline (Inmon top-down enterprise model vs. Kimball bottom-up data marts; row stores vs. compressed column stores).

**Worked Check: Entropy Split & $2 \times 2$ PCA Eigenvalues**

Eight farms ordered by rainfall 1000..1700 mm with labels $(L, L, L, H, H, H, H, L)$ have base entropy 1.0. Splitting at 1250 mm yields left $3L$ ($H = 0$) and right $4H, 1L$ ($H = 0.722$), giving $\text{IG} = 1.0 - \frac{5}{8}(0.722) = 0.549$ bits. For $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, eigenvalues are $\lambda_1 = 3$ (75%) and $\lambda_2 = 1$ (25%).