Module 08: Complete Toolkit — Cheat Sheet

ML for Innovation Research — Prof. J. Osterrieder

Seven Techniques at a Glance

Technique	Answers	Output
K-Means	What groups exist?	Cluster labels
PCA / UMAP	How to visualise?	2-D projection
VADER	What is the sentiment?	Polarity scores
TF-IDF	Which words matter?	Term weights
LDA	What themes emerge?	Topic distributions
Random Forest	Can we predict y?	Class predictions
GenAI (LLM)	Can we code at scale?	Structured labels

Decision Framework

Discovery / exploration: → Unsupervised (K-Means, PCA, LDA).
Prediction / classification: → Supervised (Random Forest, Logistic Reg.).
Text understanding: → NLP (VADER, TF-IDF, LDA).
Scaling annotation: → GenAI (structured prompts, JSON output).
Mixed methods: Combine two or more of the above.

DBA Thesis Patterns

Pattern	Approach
Exploratory	Cluster + profile + topic model
Predictive	Features → classify / regress
Mixed-methods	Cluster + classify + validate
Text-first	NLP → sentiment / topics → cross-tab

Quality Checklist

- ☐ **Reproducibility:** `random_state` set; environment pinned.
- ☐ **Data leakage:** test data never seen during training.
- ☐ **Scaling:** features standardised where required.
- ☐ **Multiple metrics:** not just accuracy.
- ☐ **Cross-validation:** used for model selection.
- ☐ **Interpretation:** results explained in domain terms.
- ☐ **Limitations:** stated honestly.
- ☐ **Ethics:** AI use disclosed; data privacy respected.

Reporting: TRIPOD Guidelines

- **T**itle — state prediction/classification goal.
- **R**ationale — why this model/approach.
- **I**nput — describe features and data source.
- **P**opulation — sample size, selection criteria.
- **O**utcome — define target variable.
- **D**evelopment — model training and validation process.
- Report: sample size, feature count, hyperparameters, all evaluation metrics, confidence intervals.

Final Tips

- Start simple; add complexity only if justified.
- Visualise early and often.
- Document every decision and its rationale.
- Your ML pipeline is part of your methodology chapter.