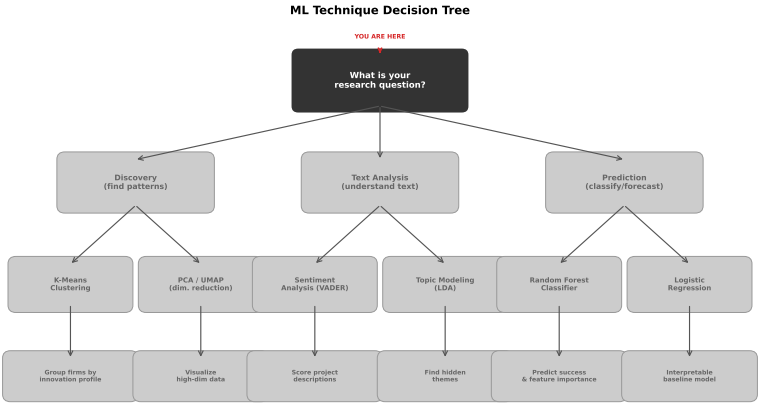


Your research question determines your method. Today we map all three branches.

Your Research Question Determines Your Method

Discovery: What groups exist? What themes emerge?

Text Analysis: What does the language reveal?

Prediction: What will happen next?

Every branch answers a different kind of question.
Today we map the entire tree — so you know which path to take.

Three branches, seven techniques. Your research question tells you where to start.

Why Machine Learning for Innovation Research?

Traditional Approach

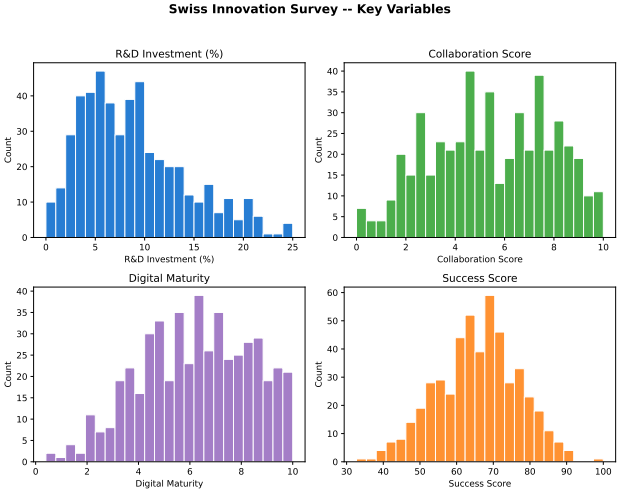
- Survey $\rightarrow$ SPSS $\rightarrow$ regression
- Small samples, few variables
- Theory-driven hypotheses only
- Text data? Manual coding

ML-Enhanced Approach

- Surveys + text + patents + web data
- Hundreds of features, pattern discovery
- Data-driven *and* theory-driven
- Automated text analysis at scale

ML does not create patterns — it reveals what was always there. The right tool makes hidden structure visible.

The Innovation Data Explosion



Community Innovation Surveys, patent databases, Crunchbase, annual reports — all analyzable with ML.

Questions the Map Can Answer

Discovery Branch

- What types of innovators exist?
  → Clustering
- What themes emerge from patent texts?
  → Topic Modeling
- How do innovators describe their work?
  → NLP / Sentiment

Prediction Branch

- What predicts innovation success?
  → Classification
- Which features matter most?
  → Feature Importance
- Can AI code qualitative data?
  → Structured Output

Session 1: Discovery and Text branches. Session 2: Prediction branch and beyond.

Two Learning Strategies

Imagine sorting a pile of 500 innovation project reports.

Supervised Learning

Someone already labeled 300 reports as "success" or "failure." You learn the pattern, then predict the remaining 200.

Key idea

Learn from *labeled* examples.
The Prediction branch of the tree.

Unsupervised Learning

No labels. You look for natural groupings — maybe some reports cluster together by topic, style, or industry.

Key idea

Discover *structure* without labels.
The Discovery branch of the tree.

Two sides of the tree: sometimes you explore what is there (no labels). Sometimes you test a hypothesis (labeled data).

The Research Protocol

Step-by-step:

1. **Define** your research question
2. **Prepare** your data (features, cleaning)
3. **Choose** an appropriate ML method
4. **Train** the model on your data
5. **Evaluate** with proper validation
6. **Interpret** — connect to theory

Critical for DBA:

- Step 1 is the most important
- Step 5: cross-validation, not just accuracy
- Step 6: feature importance $\neq$ causality
- Report transparently (TRIPOD guidelines)
- Reproducibility: share code and data

Every research project follows a protocol. Define $\rightarrow$ prepare $\rightarrow$ model $\rightarrow$ evaluate $\rightarrow$ interpret $\rightarrow$ report.

Choosing Which Features Matter

Think of it as: "What columns does your spreadsheet have?"

Numerical features

- R&D investment (%)
- Team size
- Collaboration score (0–10)
- Digital maturity (0–10)

Categorical features

- Industry (fintech, healthtech, ...)
- Company size (micro $\rightarrow$ large)
- Innovation type (product, process, ...)
- Market novelty (incremental $\rightarrow$ radical)

+ **Text features:** Project descriptions (50–200 words each)

You decide which features to examine. This choice shapes everything that follows on every branch of the tree.

The Dataset: 500 Innovation Projects

What we have:

- 500 innovation projects
- 16 variables per project
- 5 industries (fintech, healthtech, cleantech, edtech, manufacturing)
- Text descriptions (50+ words each)
- Success labels (binary + continuous)

Inspired by:

- Community Innovation Survey (CIS)
- Swiss innovation indicators
- Real-world startup ecosystem data

Live Demo

Let's explore this data together.

500 projects, 16 variables, text descriptions. The raw material for every branch of our analysis.

Live Demo: Data Exploration

Jupyter Notebook: 01_data_exploration.ipynb

- Load the dataset
- Distribution of key variables
- Correlation heatmap
- First look at text descriptions

Key question: What patterns can you already see? What surprises you?

Your Turn: Where Does This Fit?

"Think of your thesis topic. Which branch of the tree does it fall on? Discovery, Text, or Prediction?"

Take 60 seconds. Discuss with your neighbor.

Map Check: Where Are We?

Branches explored:

- ✓ The dataset: 500 projects, 16 variables, text descriptions
- ✓ Three branches mapped: Discovery, Text, Prediction

Next on the map:

- Discovery branch: what groups exist in the data?
- Clustering and dimensionality reduction

We have surveyed the landscape. Now the real exploration begins — starting with the Discovery branch.