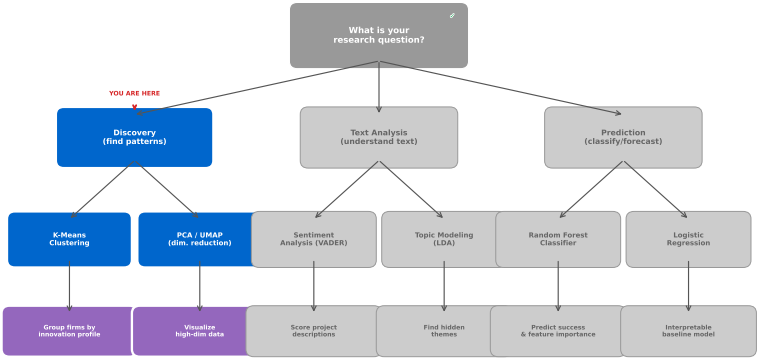
ML Technique Decision Tree



Your research question determines your method. We are exploring the Discovery branch.

Are There Distinct Types?

Among 500 innovation projects...

Are there distinct types of innovators?

Or is innovation one big blur?

The Discovery branch starts with an open question: do natural groups exist?
Clustering finds them automatically. No labels needed.

No hypothesis required. Let the data reveal its own structure.

The Clustering Intuition

The human analogy:

Imagine 500 project reports on a table. You start making piles of "similar" ones. Projects with high R&D, radical novelty, and VC funding naturally end up together.

That's clustering:

- Find natural groupings
- No pre-defined categories
- Algorithm decides what "similar" means

For innovation research:

- Discover innovator *archetypes*
- Segment by behavior, not demographics
- Data-driven persona creation
- Find outliers and unusual cases

Research question

"Are there distinct types of innovators in our sample, and do they differ in success rates?"

Clustering is unsupervised: the algorithm discovers groups you didn't know existed.

K-Means: Grouping by Similarity

How it works (3 steps, repeated):

1. **Place** k random center points
2. **Assign** each project to nearest center
3. **Move** centers to group averages

Repeat until stable.

You choose:

- k = number of clusters
- Which features to use

Strengths:

- Simple, fast, interpretable
- Works well for "blob-like" clusters
- Good starting point

Limitations:

- Assumes spherical clusters
- Sensitive to feature scales
  $\rightarrow$ always standardize first
- Need to choose k in advance

You tell it "find me 4 types" and it sorts 500 projects into 4 piles. The key question: do these groups make sense?

Seeing the Big Picture: Dimensionality Reduction

The problem:

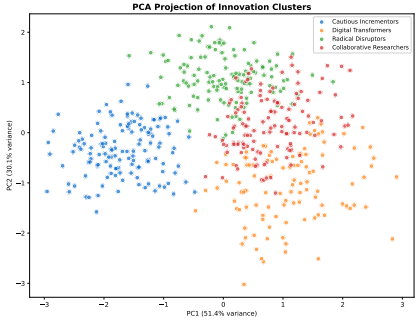
Your data has 16 features — 16 dimensions. You cannot see them all at once.

The solution:

Dimensionality reduction — project the data into 2D and step back to see the big picture.

Two popular methods:

- **PCA:** Linear projection, preserves global distances
- **UMAP:** Non-linear, preserves local neighborhoods



Important

Cluster on the original features, visualize in 2D.

PCA: Principal Component Analysis. UMAP: Uniform Manifold Approximation and Projection.

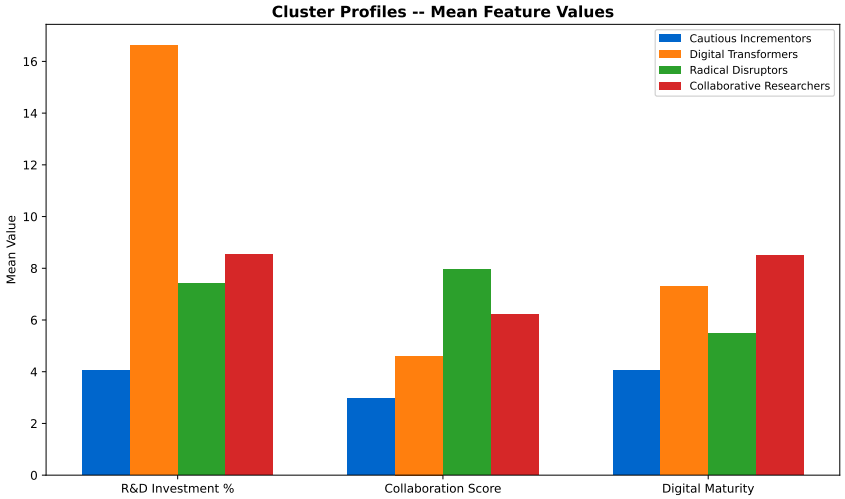
Live Demo: Clustering Innovation Projects

Jupyter Notebook: 02_clustering.ipynb

- Standardize features
- PCA + UMAP for 2D visualization
- K-Means with $k = 4$
- Interpret cluster profiles
- "The Cautious Incrementors" vs. "The Radical Disruptors"

Watch for: How do the clusters map to innovation theory? Are the profiles meaningful?

Cluster Profiles



Each bar group = one cluster. Compare across R&D investment, collaboration, digital maturity. These are your innovator archetypes.

Are the Groups Real? Validation

Validation metrics:

- **Silhouette score** (-1 to $+1$)
  > 0.3: reasonable structure
  > 0.5: strong structure
- **Elbow method:** Plot inertia vs. k
- **Gap statistic:** Compare to random data

For your DBA thesis:

- Always report silhouette scores
- Try multiple values of k
- Validate with external criteria
  ("Do clusters differ on success?")
- Cross-validate: split data, recluster
- Discuss: real segments or noise?

A good researcher asks: am I seeing a real pattern, or am I forcing connections?

Reporting Clustering in Academic Papers

What to include:

1. Algorithm used + parameters
2. Number of clusters + justification
3. Validation metrics
4. Cluster profiles (means/medians)
5. Visualization (2D projection)
6. Interpretation + theoretical link

Common mistakes:

- Choosing k "because it looks nice"
- Not standardizing features
- Ignoring cluster sizes (imbalanced)
- Over-interpreting small clusters
- Treating clusters as ground truth

Clustering is exploratory. Use it to generate hypotheses, then test them with confirmatory methods.

Your Turn: Where Does This Fit?

"If I told you there are exactly 4 types of innovators, would you believe me? What would convince you?"

Take 60 seconds. Discuss with your neighbor.

Map Check: Where Are We?

Branches explored:

- ✓ Discovery: 4 innovator archetypes identified via K-Means
- ✓ Cluster profiles validated with silhouette scores

Next on the map:

- Text branch: what does their language reveal?
- NLP, sentiment analysis, and topic modeling

We found the groups. Now we listen to what they are saying — the Text branch is next.