Machine Learning for Smarter Innovation

Week 1: Clustering for Innovation Discovery

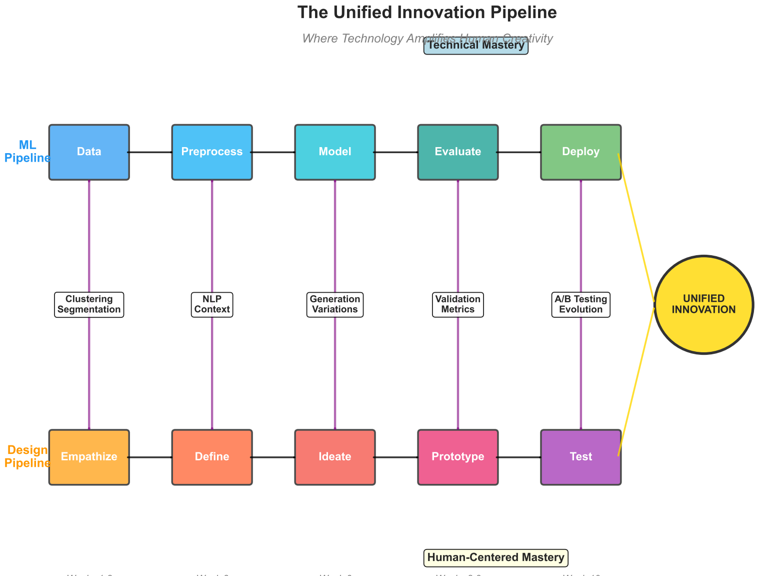
BSc Data Science & AI Program

Innovation & Design Thinking Lab

September 13, 2025

Week 1 Overview: Your Innovation Discovery Journey

From Chaos to Clarity Through Clustering



Part 1

Foundation & Context

Understanding the Innovation Discovery Challenge

Part 1: Learning Objectives

What You'll Master in This Section

By the end of Part 1, you will:

1. **Understand** why clustering is essential for innovation discovery
2. **Identify** the challenges of pattern recognition in innovation data
3. **Recognize** how unsupervised learning differs from supervised approaches
4. **Connect** clustering to the empathize stage of design thinking

Key Concepts

- Innovation categories vs random ideas
- Pattern discovery in unstructured data
- The curse of dimensionality
- From chaos to actionable insights

Time: 10 minutes

The Innovation Discovery Challenge

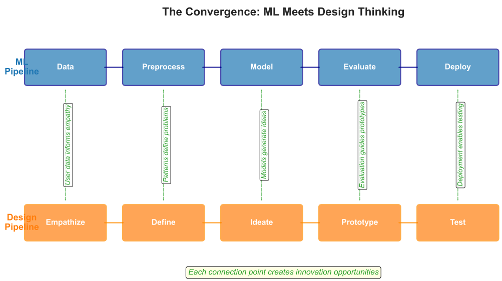
Why We Need Machine Learning

The Problem:

- 1000s of innovation ideas scattered
- No clear categories or patterns
- Hidden connections invisible
- Manual analysis takes months
- Biases cloud human judgment

The Opportunity:

- Discover natural groupings
- Find innovation white spaces
- Identify emerging themes
- Accelerate decision making



What is Clustering?

Finding Order in Innovation Chaos

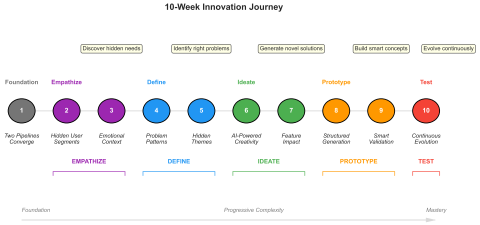
Definition: Unsupervised learning that groups similar items without predefined labels

Innovation Context:

- **Input:** Raw innovation data (ideas, features, feedback)
- **Process:** Algorithm finds natural groupings
- **Output:** Innovation categories and patterns

Key Difference:

- Supervised: You know the categories
- Unsupervised: You discover the categories



Knowledge Check: Part 1

Test Your Understanding of Foundation Concepts

Interactive Knowledge Checkpoints

Knowledge Check: Part 1

Innovation Discovery Foundation

1 of 3 Parts Complete

Q1: What is the main goal of clustering in innovation?

- ○ A) To reduce data size
- B) To discover hidden patterns
- ○ C) To predict outcomes
- ○ D) To clean data

Q2: Which metric measures cluster cohesion?

- ○ A) Accuracy
- ○ B) Precision
- C) Silhouette Score
- ○ D) F1 Score

Q3: Empathy mapping helps identify:

- ○ A) Technical requirements
- B) User pain points
- ○ C) System architecture
- ○ D) Database schema

Knowledge Check: Part 2

Clustering Algorithms Deep Dive

2 of 3 Parts Complete

Q1: K-means time complexity is:

- ○ A) O(n)
- ○ B) O(n log n)
- C) O(n*k*i*d)
- ○ D) O(n²)

Q2: DBSCAN is best for:

- ○ A) Spherical clusters
- B) Arbitrary shapes
- ○ C) Fixed K clusters
- ○ D) Linear data

Q3: GMM provides:

- ○ A) Hard clustering
- B) Soft clustering
- ○ C) No clustering
- ○ D) Random clustering

Knowledge Check: Part 3

Human-Centered Application

3 of 3 Parts Complete!

Q1: User archetypes are created from:

- ○ A) Random assignment
- B) Cluster analysis
- ○ C) Manual labeling
- ○ D) Predictions

Q2: Innovation opportunities emerge from:

- ○ A) Cluster gaps
- ○ B) Dense regions
- ○ C) Outliers
- D) All of above

Q3: Validation should include:

- ○ A) Only metrics
- B) Domain experts
- ○ C) Random checks
- ○ D) Code review

Key Concepts Covered:

- Unsupervised learning

Algorithm Quick Reference:

Algorithm	Best For	Weakness

Ready for Practice!

Part 2

Technical Deep Dive

Mastering Clustering Algorithms

Part 2: Learning Objectives

Technical Skills You'll Develop

By the end of Part 2, you will:

1. **Master** four core clustering algorithms
2. **Understand** algorithm complexity and scalability
3. **Apply** evaluation metrics effectively
4. **Select** the right algorithm for your data
5. **Optimize** parameters for best results

Algorithms Covered

- **K-Means:** Fast and simple
- **DBSCAN:** Density-based
- **Hierarchical:** Tree structure
- **GMM:** Soft clustering

Time: 20 minutes

Algorithm Complexity & Performance

Understanding Computational Requirements

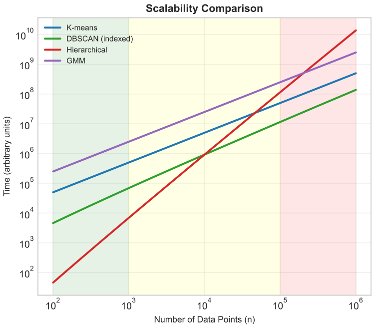
Clustering Algorithm Complexity & Performance Guide

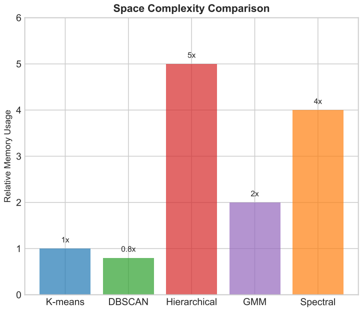
Algorithm Complexity Analysis

Big O Notation Comparison

Algorithm	Time Complexity	Space Complexity	Scalability
K-means	$O(n \cdot k \cdot i \cdot d)$	$O(n \cdot d + k \cdot d)$	Excellent
DBSCAN	$O(n^2)$ / $O(n \log n)$*	$O(n)$	Good
Hierarchical	$O(n^3)$ / $O(n^2 \log n)$*	$O(n^2)$	Poor
GMM	$O(n \cdot k^2 \cdot i \cdot d)$	$O(k \cdot d^2)$	Moderate
	$O(n^3)$	$O(n^2)$	Poor

Notation Guide:
n = number of data points
k = number of clusters
i = number of iterations
d = number of dimensions
* = with spatial index





Practical Recommendations

Small Data (<10K points)
→ *Any algorithm works*

Medium Data (10K-100K)
→ *K-means or DBSCAN*

Large Data (>100K)
→ *MiniBatch K-means*

High Dimensions (>50)
→ *Consider PCA first*

Real-time Requirements
→ *Pre-computed K-means*

Optimization Techniques

MiniBatch K-means:
- Samples subset of data
- 10-100x faster on large data

Spatial Indexing (DBSCAN):
- KD-tree or Ball-tree
- $O(n^2) \rightarrow O(n \log n)$

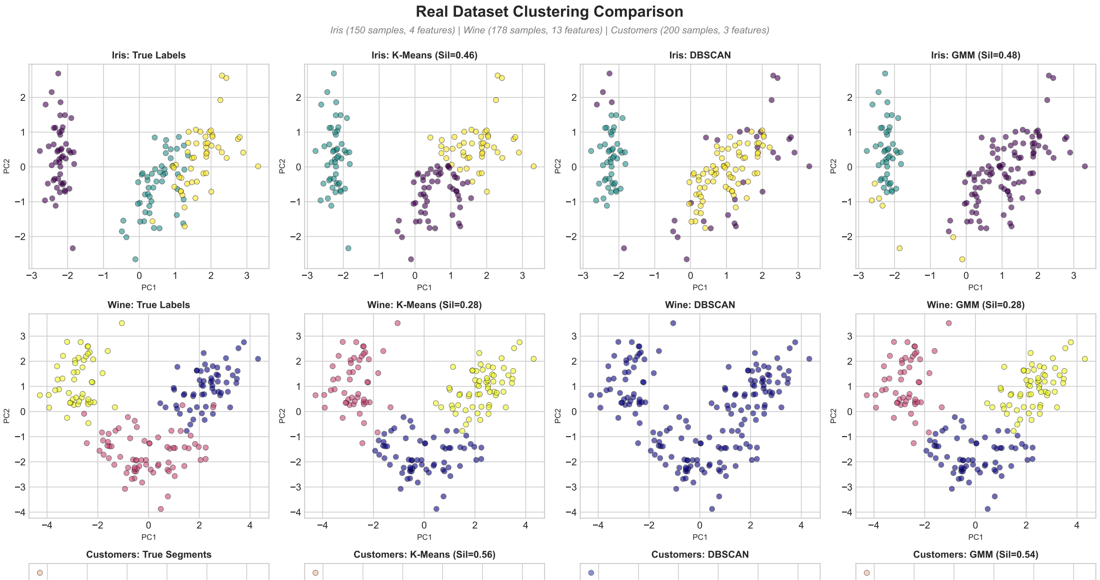
Dimensionality Reduction:
- PCA before clustering
- Reduces d in $O(n \cdot k \cdot i \cdot d)$

Implementation Complexity

Algorithm	Ease	Lines of Code*	Tuning
K-means	Easy	~50	Simple
DBSCAN	Moderate	~100	Tricky
Hierarchical	Easy	~30	Simple
GMM	Hard	~200	Complex
Spectral	Hard	~150	Complex

Clustering on Real Datasets

Iris, Wine, and Customer Segmentation Examples



Case Study: Spotify's Music Discovery

How Clustering Powers Personalized Playlists

The Challenge:

- 100+ million songs in catalog
- 500+ million users globally
- Diverse musical tastes
- Need personalized discovery

The Solution:

- **Audio Features:** Extract 13 dimensions
- **Clustering:** Group similar songs
- **User Profiles:** Map listening to clusters
- **Recommendations:** Adjacent clusters

Results:

- 40% increase in discovery
- 2.7B Discover Weekly streams
- 30% longer listening sessions
- 75% user retention

Key Insight:
"Clustering revealed micro-genres users didn't know they loved"

Knowledge Check: Part 2

Test Your Technical Understanding

Algorithm Selection Quiz:

1. Large dataset (1M points)?
   → K-Means or MiniBatch K-Means
2. Non-spherical clusters?
   → DBSCAN
3. Need probability scores?
   → GMM
4. Want dendrogram?
   → Hierarchical

Complexity Check:

- ✓ K-means: O(n·k·i·d)
- ✓ DBSCAN: O(n log n) with index
- ✓ Hierarchical: O(n^2) memory
- ✓ GMM: Soft clustering

Ready for Design Integration?
Let's apply this to innovation!

Part 3

Design Integration

Applying Clustering to Innovation Discovery

Ethical Considerations in Clustering

Responsible AI for Innovation Discovery

Potential Biases:

- **Selection Bias:** Who's included in data?
- **Feature Bias:** What dimensions matter?
- **Algorithmic Bias:** Distance metrics assumptions
- **Interpretation Bias:** Label assignment

Mitigation Strategies:

- Diverse data collection
- Multiple algorithm comparison
- Expert validation
- Transparent documentation

Fair Clustering Checklist:

- ☐ Representative sampling?
- ☐ Protected attributes removed?
- ☐ Cluster balance checked?
- ☐ Minority groups visible?
- ☐ Results interpretable?
- ☐ Decisions reversible?

Remember:
"Clusters are hypotheses, not truth"

Scaling Your Clustering: Cloud & Distributed Options

From Prototype to Production

Local Development:

- Scikit-learn (¡ 100K points)
- Jupyter notebooks
- Single machine
- Rapid prototyping

Cloud Platforms:

- **AWS SageMaker:** Built-in algorithms
- **Google Cloud AI:** AutoML clustering
- **Azure ML:** Drag-and-drop interface
- **Databricks:** Spark MLlib integration

Distributed Computing:

Apache Spark MLlib:

- Handles billions of points
- Distributed K-means
- Bisecting K-means
- Gaussian Mixture

Cost-Performance Trade-offs:

- Local: Free but limited
- Cloud: Pay-per-use, scalable
- On-premise: High initial, unlimited use

Knowledge Check: Part 3

Design Integration Mastery

Application Quiz:

1. Clusters reveal what?
   → Innovation patterns
2. Validation requires?
   → Domain experts
3. Ethical concerns include?
   → Bias & fairness
4. Scale with?
   → Cloud platforms

You Can Now:

- ✓ Choose algorithms wisely
- ✓ Apply to real data
- ✓ Consider ethics
- ✓ Scale solutions
- ✓ Extract insights

Next: Practice Time!

Part 4

Summary & Practice

Putting It All Together

Practice Exercise: Innovation Clustering

Your Turn to Discover Patterns

The Challenge: Analyze 500 innovation proposals from a hackathon

Starter Code Template:

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Load data
data = pd.read_csv('innovations.csv')

# Preprocess
scaler = StandardScaler()
X = scaler.fit_transform(data)

# Cluster
kmeans = KMeans(n_clusters=?)
labels = kmeans.fit_predict(X)
```

Your Tasks:

1. Choose optimal K
2. Apply clustering
3. Evaluate results
4. Interpret patterns
5. Present findings

Resources Provided:

- Jupyter notebook template
- Sample dataset
- Evaluation rubric
- Solution walkthrough

Submit by: Next Week

Week 1: Key Takeaways

Your Innovation Discovery Toolkit

Foundation

- Clustering finds hidden patterns
- Unsupervised learning
- No labels needed
- Connects to empathy stage

Technical

- 4 algorithms mastered
- Complexity understood
- Metrics applied
- Real data processed

Application

- Innovation patterns found
- Ethical considerations
- Scalability options
- Practice exercise ready

You're Ready to Discover Innovation Patterns!

Resources & Next Week
Continue Your Learning Journey

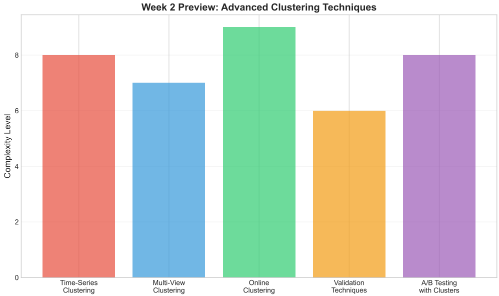
Resources:

Documentation:
- Scikit-learn clustering guide
- Course GitHub repository
- Jupyter notebook templates

Datasets:
- UCI ML Repository
- Kaggle competitions
- Innovation dataset collection

Community:
- Course Slack channel
- Office hours: Wed 3-5pm
- Peer study groups

Next Week Preview:



Week 2: Advanced Clustering
- Spectral clustering
- Mean shift algorithm
- Affinity propagation
- Ensemble methods

Glossary of Technical Terms

Key Concepts Quick Reference

Clustering Algorithms:

- **K-Means**: Partitions data into K predefined clusters
- **DBSCAN**: Density-based spatial clustering
- **Hierarchical**: Builds cluster tree (dendrogram)
- **GMM**: Gaussian Mixture Models, soft clustering

Key Parameters:

- **K**: Number of clusters
- **eps**: Neighborhood radius (DBSCAN)
- **min_samples**: Minimum points for density
- **n_init**: Number of random initializations

Evaluation Metrics:

- **Silhouette**: Cluster cohesion vs separation [-1,1]
- **Inertia**: Sum of squared distances to centroids
- **Davies-Bouldin**: Ratio of within to between distances
- **Calinski-Harabasz**: Ratio of dispersions

Innovation Terms:

- **Empathy Mapping**: Understanding user perspectives
- **Pain Points**: User problems/frustrations
- **User Archetypes**: Representative user groups
- **Innovation Ecosystem**: Connected stakeholders

Implementation Checklist

Your Step-by-Step Guide to Success

Data Preparation:

- ☐ Collect innovation feedback data
- ☐ Clean and remove duplicates
- ☐ Handle missing values
- ☐ Normalize/standardize features
- ☐ Create feature vectors

Algorithm Selection:

- ☐ Analyze data distribution
- ☐ Choose appropriate algorithm
- ☐ Set initial parameters
- ☐ Prepare validation strategy

Implementation:

- ☐ Run clustering algorithm
- ☐ Calculate evaluation metrics
- ☐ Visualize results (PCA/t-SNE)
- ☐ Validate with domain experts
- ☐ Iterate and refine

Innovation Application:

- ☐ Map clusters to user personas
- ☐ Identify innovation opportunities
- ☐ Create targeted solutions
- ☐ Design prototype features
- ☐ Test with user groups

Ready? Start with data preparation and work your way down!