Machine Learning for Smarter Innovation

Week 1: Foundations & Clustering

Discovering Innovation Patterns with ML

BSc Course in AI-Enhanced Innovation

Prerequisites & What You Need

Setting You Up for Success

What You Need to Know

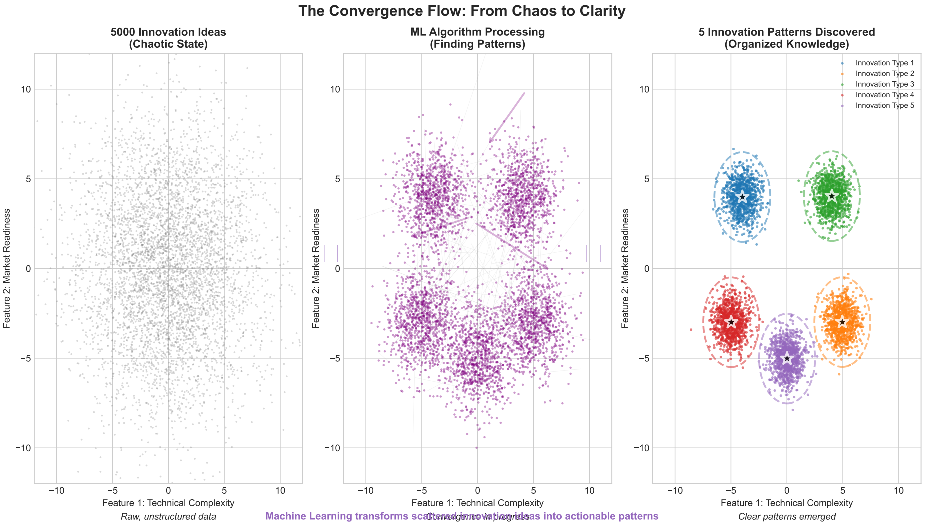
- Basic Python (variables, loops, functions)
- High school math (averages, distances)
- How to use Jupyter notebooks
- Basic data concepts (tables, rows, columns)

What We'll Provide

- All code templates
- Step-by-step instructions
- Visual explanations
- Practice datasets

Machine Learning + Innovation + Design Thinking

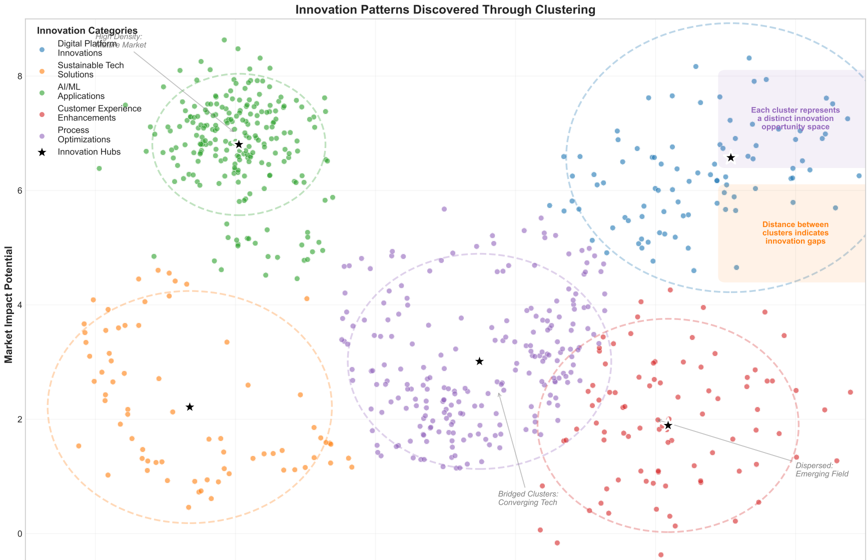
The Power of Convergent Methodologies



Where Data Science Meets Human Creativity

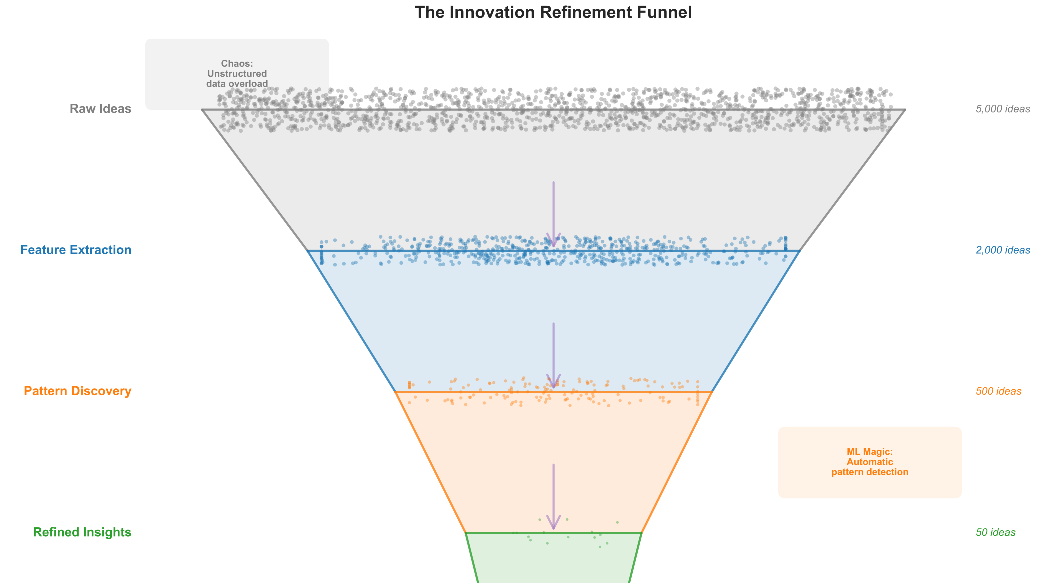
From Data Points to Innovation Insights

Bridging the Technical-Human Gap



The Innovation Refinement Funnel

From Chaos to Clarity Through Feature Analysis



PART 1

Foundation & Context

What we'll explore:

- Why traditional design hits limits
- How ML amplifies human insight
- The dual pipeline approach
- Your learning journey ahead

Setting the stage for transformation

Part 1: Learning Objectives

What You'll Learn in This Section

By the end of Part 1, you will be able to:

- **Understand** the limitations of traditional innovation approaches
- **Recognize** how ML enhances human creativity
- **Explain** the dual pipeline methodology
- **Navigate** the 10-week learning journey
- **Identify** Week 1's role in the overall course

Success Criteria

- Can articulate 3+ traditional design limitations
- Can describe ML's value proposition
- Can map ML pipeline to design pipeline
- Understand clustering's role in innovation

PART 1

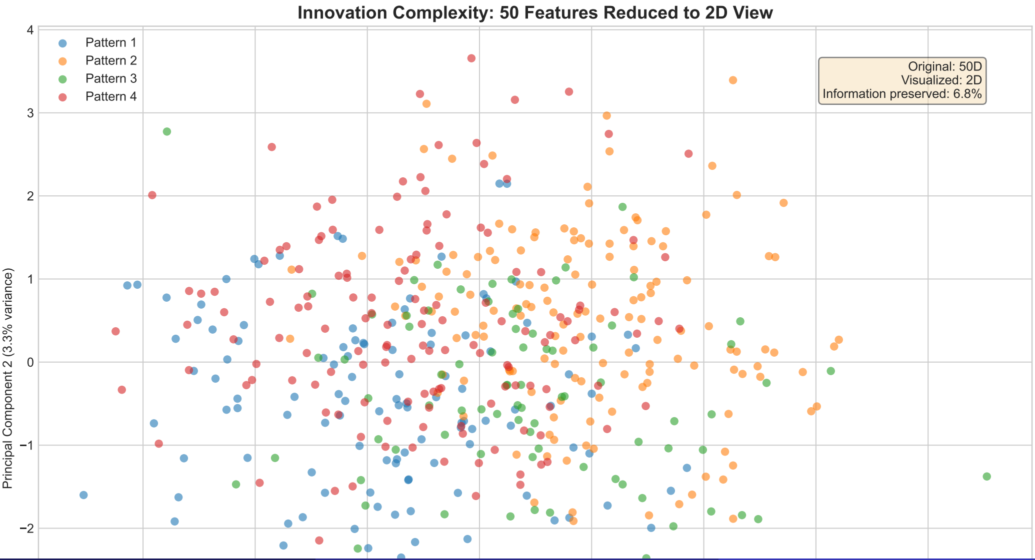
Foundation & Context

Understanding the Innovation Challenge


Innovation Discovery: Where Do We Start?
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
Market Opportunity
?
?
?
?
10
8
6
4
2
Innovation Potential Score

The Hidden Complexity

Each Innovation Depends on Hundreds of Features



The Innovation Challenge

Why Traditional Design Needs AI Enhancement

Traditional Design Limits

- **Scale**: Can analyze 50 ideas, not 50,000
- **Speed**: Months for insights
- **Bias**: Designer's perspective dominates
- **Patterns**: Miss hidden connections
- **Iteration**: Slow feedback loops

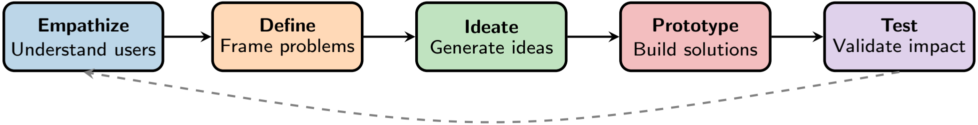
AI-Enhanced Innovation

- **Scale**: Analyze millions of data points
- **Speed**: Real-time insights
- **Objectivity**: Data-driven discovery
- **Patterns**: Find non-obvious relationships
- **Iteration**: Continuous learning

The Promise: 100x more insights, 10x faster innovation

Quick Recap: The Design Thinking Process

You've Seen This Before - Let's Connect It to ML



Iteration is key

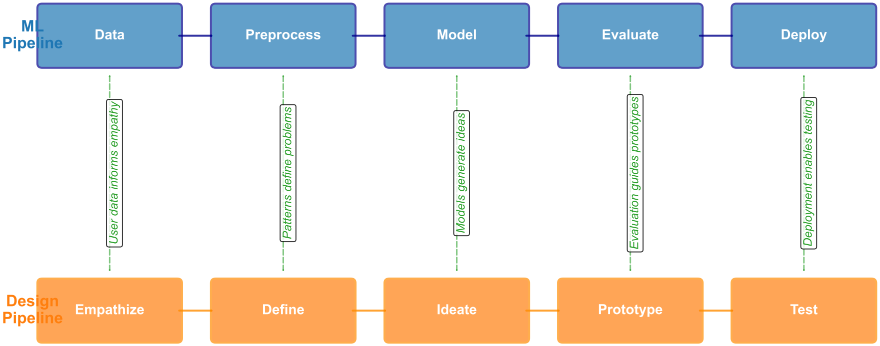
Traditional Approach

- Manual interviews
- Limited sample size
- Qualitative insights
- Slow iteration

ML-Enhanced Approach

- Data-driven discovery
- Massive scale analysis
- Quantitative patterns
- Real-time adaptation

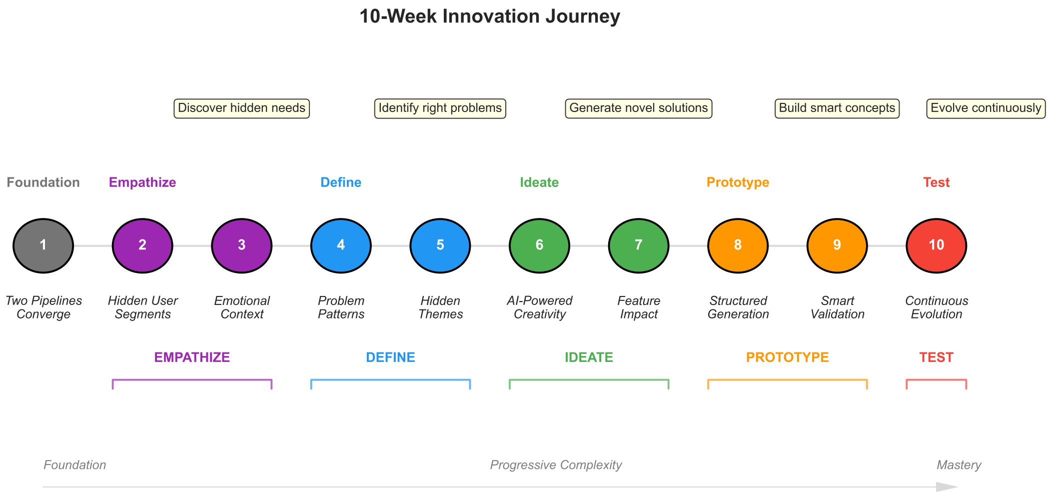
The Convergence: ML Meets Design Thinking

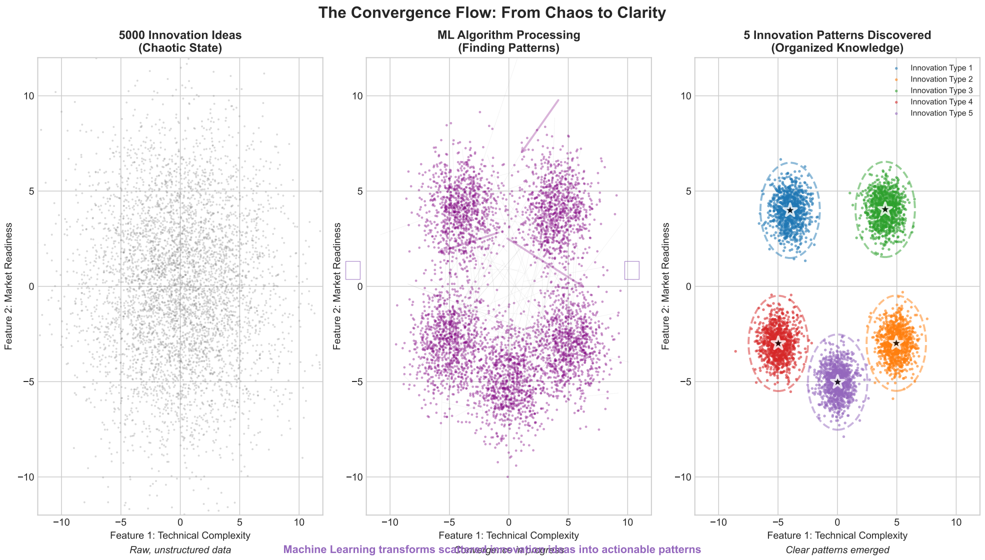


Each connection point creates innovation opportunities

Your Innovation Journey

10 Weeks to Understanding AI-Powered Design





The Convergence Flow: Order from Chaos

Watch 5000 innovation ideas self-organize into meaningful patterns

PART 2

Technical Core

Machine Learning Algorithms & Implementation

The Innovation Classification Problem

5000 Ideas - How Do They Connect?

The Pain

Current Reality:

- One-size-fits-all solutions
- Generic innovation categories
- Missed opportunities
- Unhappy edge cases

The Cost:

- Most innovations get misclassified
- Features with low adoption rates
- Inefficient resource allocation

The Question

What if we could...

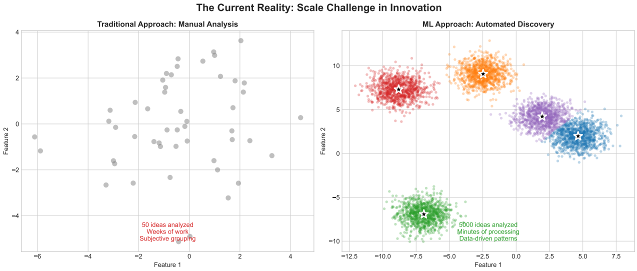
- Find natural innovation clusters?
- Discover innovation patterns?
- Innovate at scale?
- Identify opportunity gaps?

We can!

Solution: Clustering

Current Reality: The Problem

Why One-Size-Fits-All Doesn't Work



The Fragmentation Crisis

- **70%** of innovations fail due to misalignment
- **85%** miss their target audience
- **60%** duplicate existing solutions

Why This Happens:

- Treating all ideas the same
- Missing subtle patterns
- No systematic categorization
- Human cognitive limits

ML clustering reveals the hidden structure in innovation chaos

Discovery Exercise: Which Archetype?

Match Each Innovation to Its Type

Innovation Examples:

1. **Uber** - Connecting drivers with riders via app
2. **Tesla Model 3** - Affordable electric vehicle
3. **Amazon Prime** - Fast delivery subscription
4. **iPhone Camera** - Annual improvements
5. **ChatGPT** - AI conversation interface

Think: What makes each one similar or different?

Match to Type:

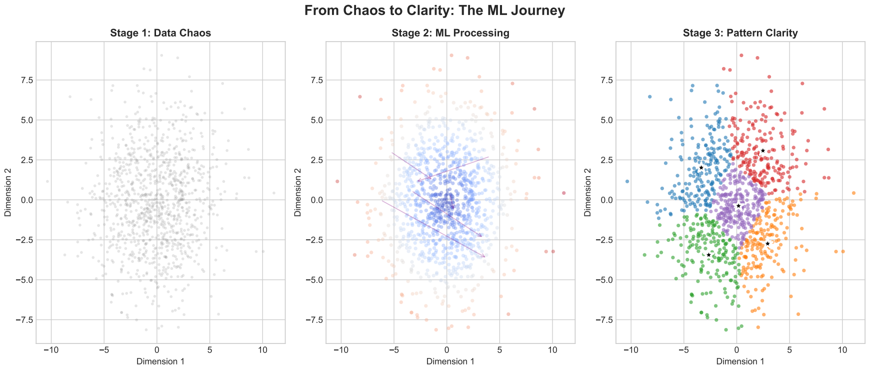
A. Disruptive Innovation
B. Incremental Innovation
C. Platform Innovation
D. Service Innovation
E. Business Model Innovation

Answers:
(Discuss with neighbor first)
1$\longrightarrow$C (Platform), 2$\longrightarrow$A (Disruptive), 3$\longrightarrow$E (Business Model), 4$\longrightarrow$B (Incremental), 5$\longrightarrow$D (Service)

ML can do this matching at scale - for thousands of innovations

What is Clustering?

Like Organizing a Messy Room - Finding Things That Belong Together



Clustering Finds:

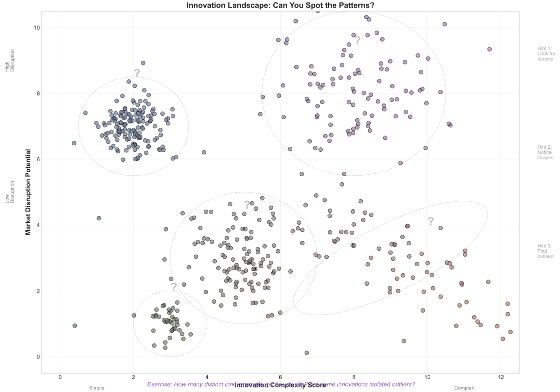
- Natural groupings
- Similar approaches
- Hidden patterns
- Innovation relationships

Key Insight:
Things that look similar often belong in the same group
(Just like organizing books by topic on a shelf)

Discovery: How Many Groups Do You See?

Visual Pattern Recognition Exercise

Look at this innovation data:



Questions:

- How many distinct groups?

Your Observations:

1. Groups you see: _____
2. Main pattern: _______
3. Outliers: _________

Key Insight:
Humans are good at 2D patterns.
But innovation has 100+ dimensions!
That's where ML helps.

For Innovations, What Features Matter?

Innovation	Cost	Impact	Time
Smart Thermostat	Low	Medium	Quick
Electric Car	High	High	Long
Mobile App	Low	Low	Quick
Solar Panels	High	High	Long
AI Chatbot	Medium	Medium	Medium

Which innovations group together?

- By cost? (Low vs High)
- By impact? (Low vs High)
- By timeline? (Quick vs Long)
- All combined?

Discovery Exercise

Group these by similarity:

1. Smart Thermostat + ?
2. Electric Car + ?
3. Mobile App + ?

The Challenge:
Real innovations have 100+ features!
- Market size - Technology readiness - Regulatory requirements - User demographics - Competition level - And many more...

Discovery: Manual Clustering Exercise

Try Clustering Yourself - Then See How ML Does It

Your Task:

Group these 12 innovations into 3 clusters:

1. Blockchain payment system
2. Voice-activated assistant
3. Renewable energy storage
4. Social media platform
5. Autonomous vehicle
6. Health tracking wearable
7. Cloud computing service
8. 3D printing technology
9. Virtual reality training
10. Drone delivery system
11. Gene editing tool
12. Quantum computing

Your Groups:

Group 1: ________
Group 2: ________
Group 3: ________

Think About:

- What features did you consider?
- How did you decide on groups?
- Was it difficult to classify some items?
- Did any items fit multiple groups?

Let's see how ML approaches this...

Discovery: How ML Clusters These Innovations

Based on 50+ Hidden Features

Digital Platforms

4. Social media platform
7. Cloud computing service
2. Voice-activated assistant
1. Blockchain payment system

Physical Innovation

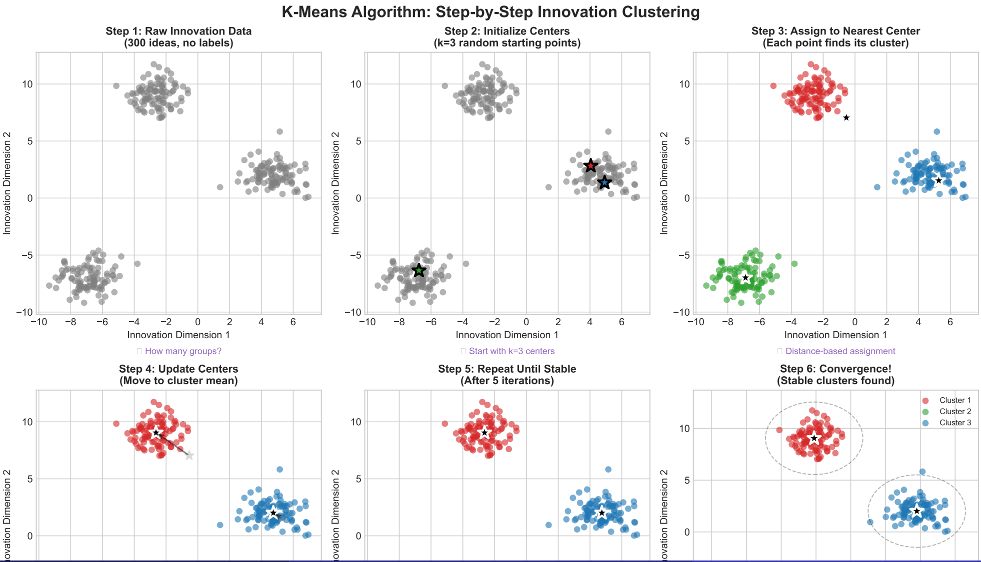
5. Autonomous vehicle
10. Drone delivery system
3. Renewable energy storage
8. 3D printing technology

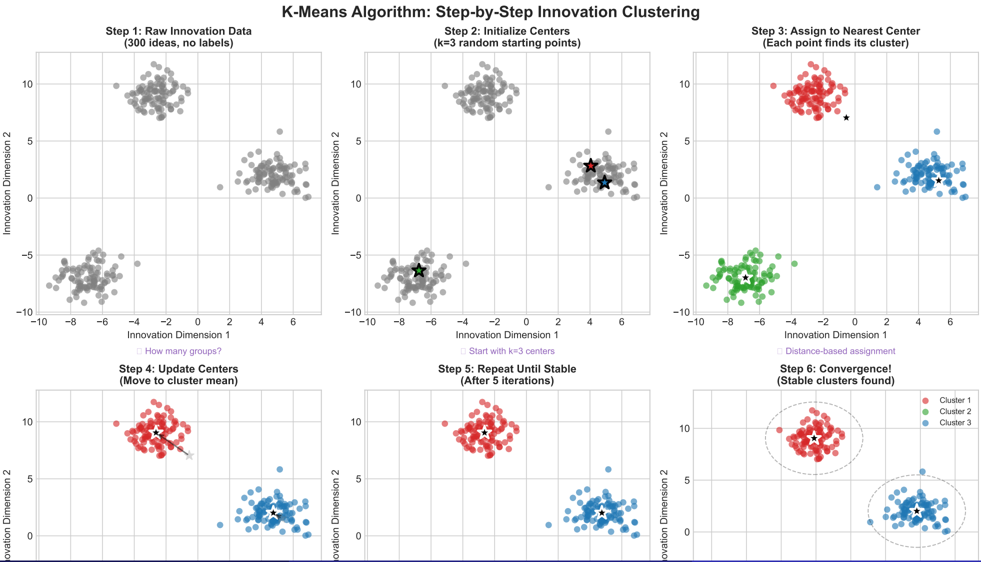
Frontier Tech

11. Gene editing tool
12. Quantum computing
9. Virtual reality training
6. Health tracking wearable

ML considers features you might not think of:

- Development complexity
- Market readiness level
- Infrastructure requirements
- Regulatory complexity
- User behavior patterns
- Technology stack similarity
- Investment requirements
- Innovation lifecycle stage


K-Means Algorithm: Step-by-Step Innovation Clustering
Step 1: Raw Innovation Data
(300 ideas, no labels)
How many groups?
Step 2: Initialize Centers
(k=3 random starting points)
Start with k=3 centers
Step 3: Assign to Nearest Center
(Each point finds its cluster)
Distance-based assignment
Step 4: Update Centers
(Move to cluster mean)
Step 5: Repeat Until Stable
(After 5 iterations)
Step 6: Convergence!
(Stable clusters found)
Cluster 1
Cluster 2
Cluster 3
Innovation Dimension 1
Innovation Dimension 2



K-Means Algorithm: Step-by-Step Innovation Clustering
Step 1: Raw Innovation Data
(300 ideas, no labels)
How many groups?
Step 2: Initialize Centers
(k=3 random starting points)
Start with k=3 centers
Step 3: Assign to Nearest Center
(Each point finds its cluster)
Distance-based assignment
Step 4: Update Centers
(Move to cluster mean)
Step 5: Repeat Until Stable
(After 5 iterations)
Step 6: Convergence!
(Stable clusters found)
Cluster 1
Cluster 2
Cluster 3
Innovation Dimension 1
Innovation Dimension 2

The Goldilocks Problem

Too Few vs. Too Many Groups

Too Few (K=2)

Oversimplification

- Mixed segments
- Lost nuance
- Generic solutions

Just Right (K

Optimal Balance

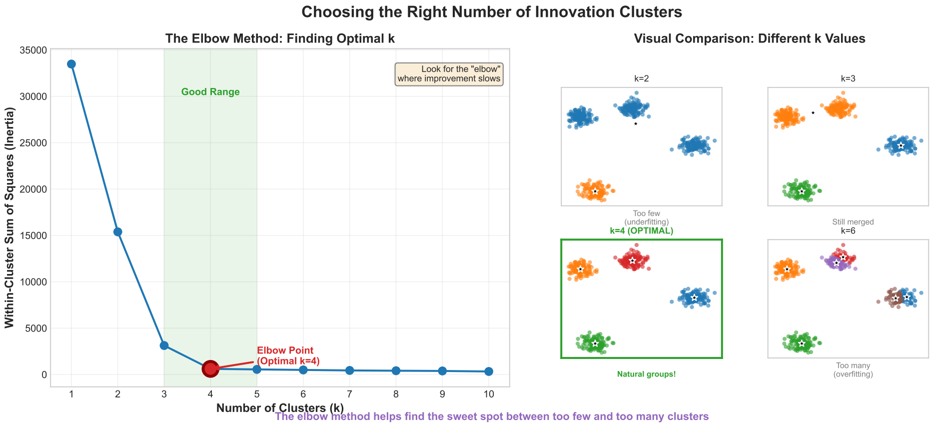
- Clear segments
- Actionable insights
- Manageable complexity

Too Many (K

Analysis Paralysis

- Overfitting
- Tiny segments
- Impossible to act on

How do we find the sweet spot?

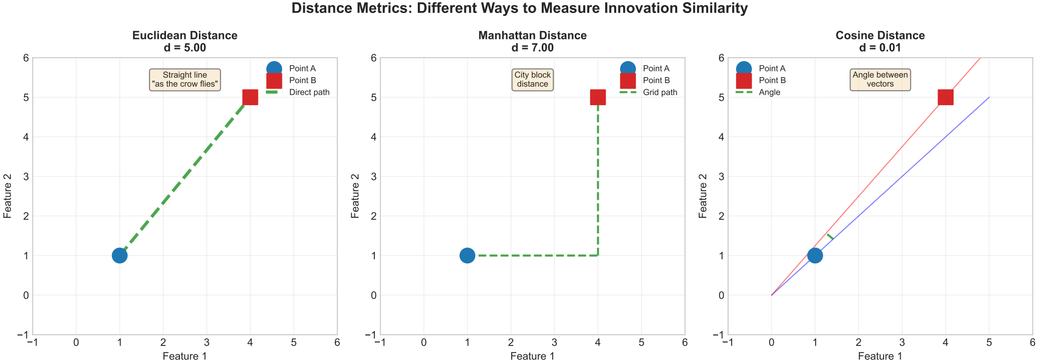


Finding the Elbow:

- Plot inertia vs K
- Look for the "elbow"

Optimal K = 5
Best trade-off point

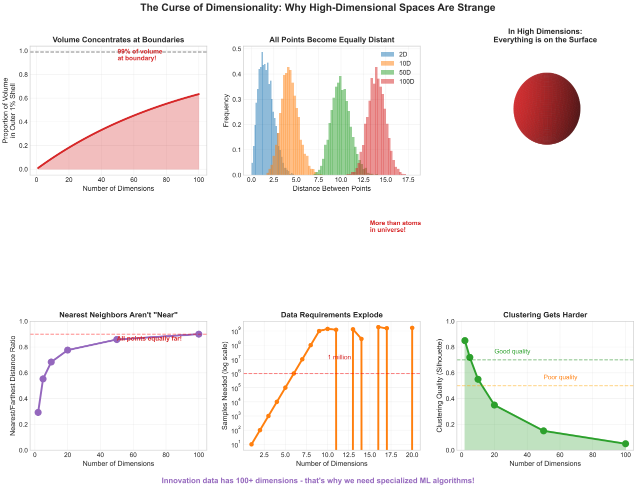
Different Ways to Measure "How Close" Things Are



Each metric reveals different patterns in your data

Sidestep: The Curse of Dimensionality

Why High-Dimensional Spaces Are Strange and Empty



The Paradox

In 100 dimensions:

- 99.99% of space is empty
- All points are outliers
- Nearest neighbors aren't near

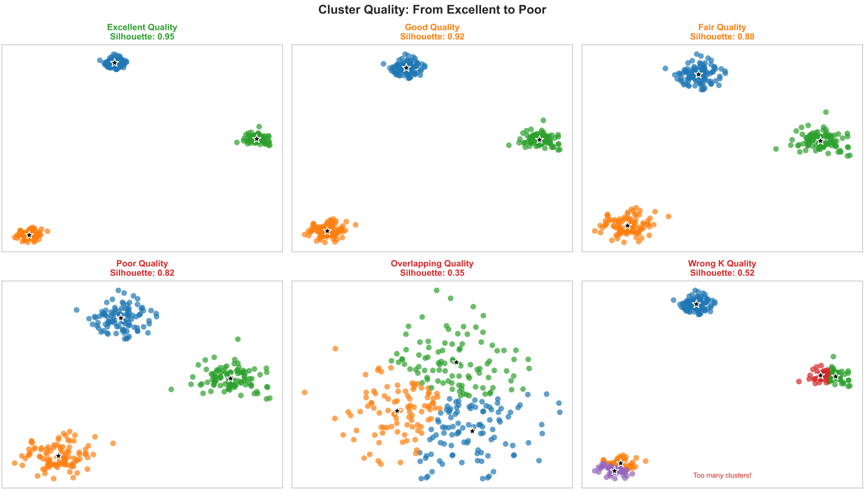
Why This Matters

Innovation has 100+ features!

- Distance metrics break down
- Need special techniques
- Dimensionality reduction crucial
- That's why we use PCA/t-SNE

As dimensions increase:

- Points become equally distant

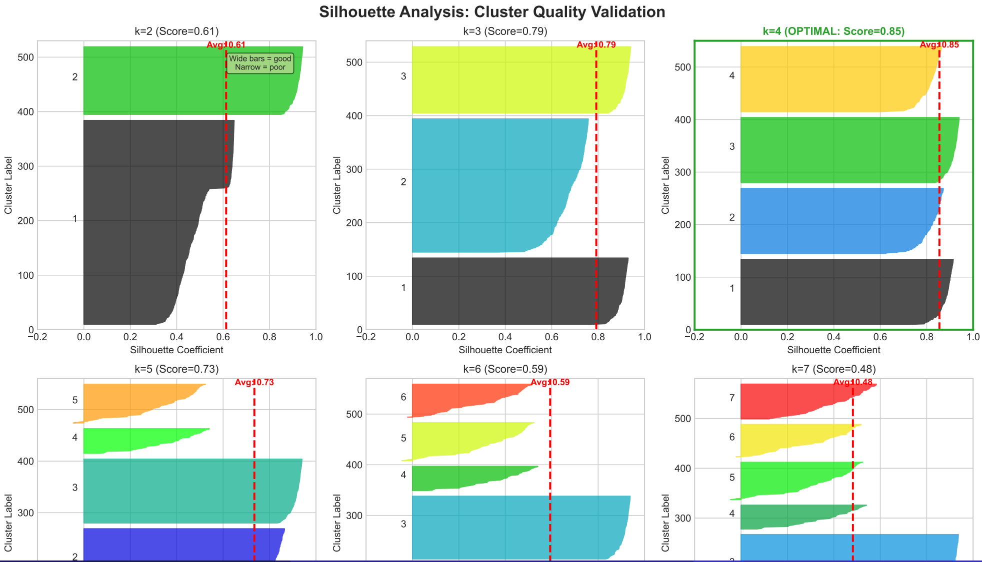


Silhouette Score:

What it measures:

Evaluation Metric 1: Silhouette Score

Measuring Cluster Cohesion and Separation



Discovery: Finding the Right K

What Happens With Different Numbers of Clusters?

Experiment with K:

K	What Happens
K=2	Everything too mixed
K=4	Natural groups emerge
K=8	Some groups split unnecessarily
K=20	Too fragmented to use

Your Turn:
If you have 100 customer types, what K would you choose?

- K=100? (one per type)
- K=5? (major groups)
- K=20? (detailed segments)

The Trade-offs:

Too Few (Under-fit)

- Mixed segments
- Lost insights
- Generic solutions

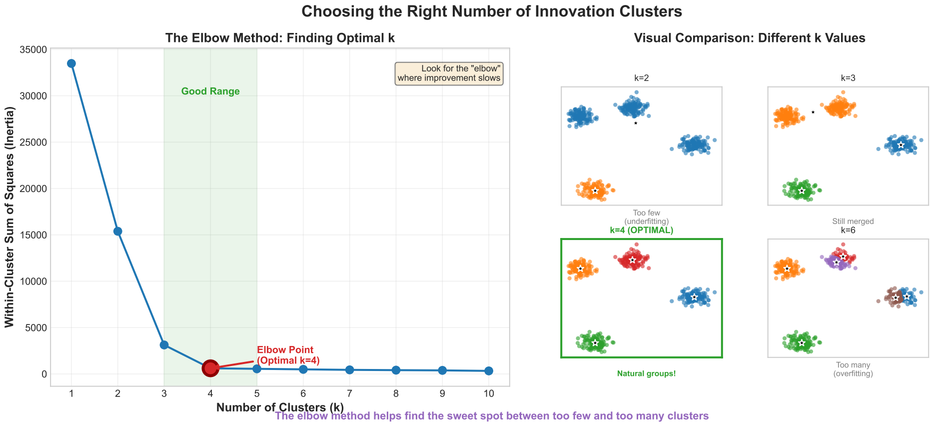
Just Right

- Clear segments
- Actionable groups
- Meaningful patterns

Too Many (Over-fit)

- Fragmented insights
- Hard to implement
- Statistical noise

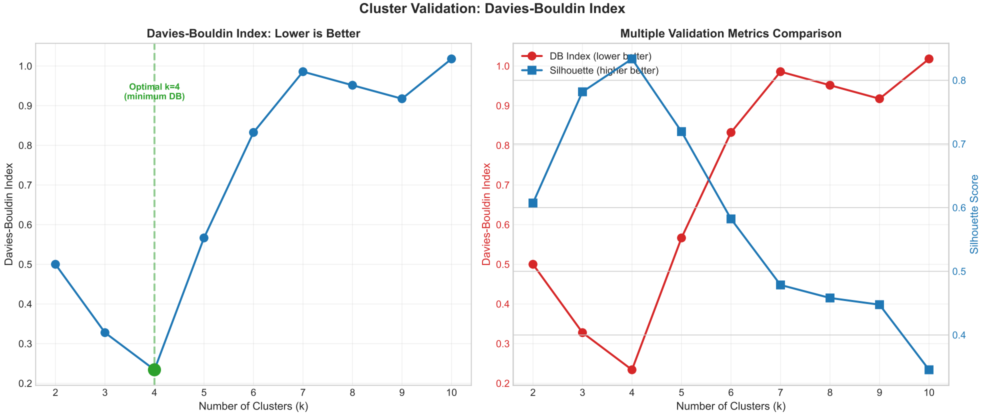
The Elbow Method helps find the sweet spot automatically



Choosing the Right Number of Innovation Clusters
The Elbow Method: Finding Optimal k
Within-Cluster Sum of Squares (Inertia)
Number of Clusters (k)
Good Range
Look for the "elbow" where improvement slows
Elbow Point (Optimal k=4)
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
Visual Comparison: Different k Values
k=2
Too few (underfitting)
k=3
Still merged
k=4 (OPTIMAL)
Natural groups!
k=6
Too many (overfitting)
The elbow method helps find the sweet spot between too few and too many clusters

Evaluation Metric 3: Davies-Bouldin Index

Balancing Within and Between Cluster Distances



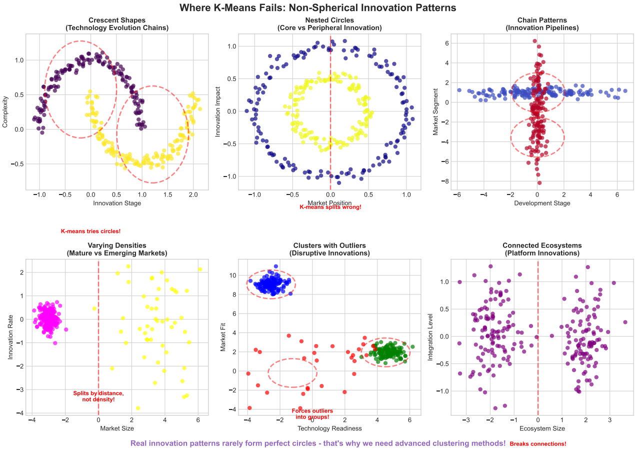
K-Means Assumes Spherical Clusters

But what about:

- Innovations connected through technology stacks
- Domain-specific innovation clusters
- Evolution patterns (incremental, disruptive)
- Outliers and noise points

K-Means Forces Round Pegs into Round Holes

Solution: Density-Based Clustering



Look at these patterns:

- Crescent shapes

K-Means Problems

K-Means assumes:

- Spherical (round) clusters
- Similar sizes
- Similar densities
- No outliers

Real Innovation Patterns

But innovations have:

- Evolution chains
- Technology ecosystems
- Varying market sizes
- Disruptive outliers

Exercise:

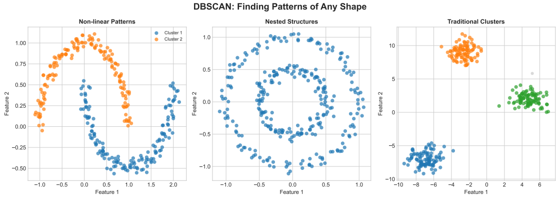
Draw clusters on the left image.
Where does K-means fail?

DBSCAN Advantages:

- No need to specify K *(finds groups automatically)*
- Finds arbitrary shapes *(not just circles)*
- Identifies outliers *(points that don't belong)*
- Handles noise well *(robust to random points)*

Perfect for:

- Non-spherical patterns
- Varying densities
- Outlier detection
- Exploratory analysis



DBSCAN: Understanding Parameters

Two Simple Settings Control Everything

Epsilon (Distance)

What it does:
Sets the maximum distance to consider points as neighbors

Think of it as:
How far can points be apart and still be friends?

Too small: Many tiny clusters
Too large: Everything merges

MinPts (Density)

What it does:
Minimum neighbors needed to form a dense region

Think of it as:
How many friends make a group?

Too small: Noise becomes clusters
Too large: Small clusters vanish

Rule of thumb: MinPts = 2 × dimensions

Algorithm	Speed	Shape	Outliers	Params	Best For
K-Means	Fast $O(nkt)$	Spherical clusters	Sensitive	K only	Quick segments
DBSCAN	Medium $O(n \log n)$	Any shape	Robust (detects)	eps, MinPts	Complex shapes
Hierarchical	Slow $O(n^2)$	Any shape	Moderate	Distance threshold	Multi-level analysis
GMM	Medium $O(nkt)$	Elliptical clusters	Moderate	K, covariance	Overlapping groups

Each algorithm has its strengths - choose wisely!

When to Use Each Algorithm

Practical Decision Guide

K-Means

Perfect when:

- Speed is critical
- Clusters are roughly equal size
- You know K in advance
- Data has spherical patterns

Hierarchical

Perfect when:

- Need multiple granularities
- Want to visualize relationships
- Small to medium datasets
- Exploring data structure

DBSCAN

Perfect when:

- Clusters have irregular shapes
- Outliers need identification
- Density varies across data
- You don't know K

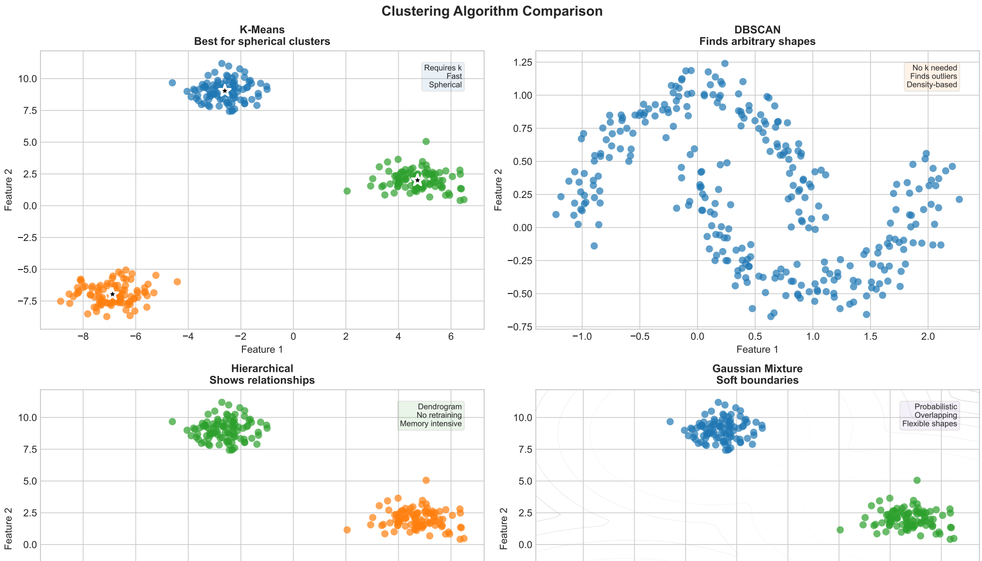
GMM

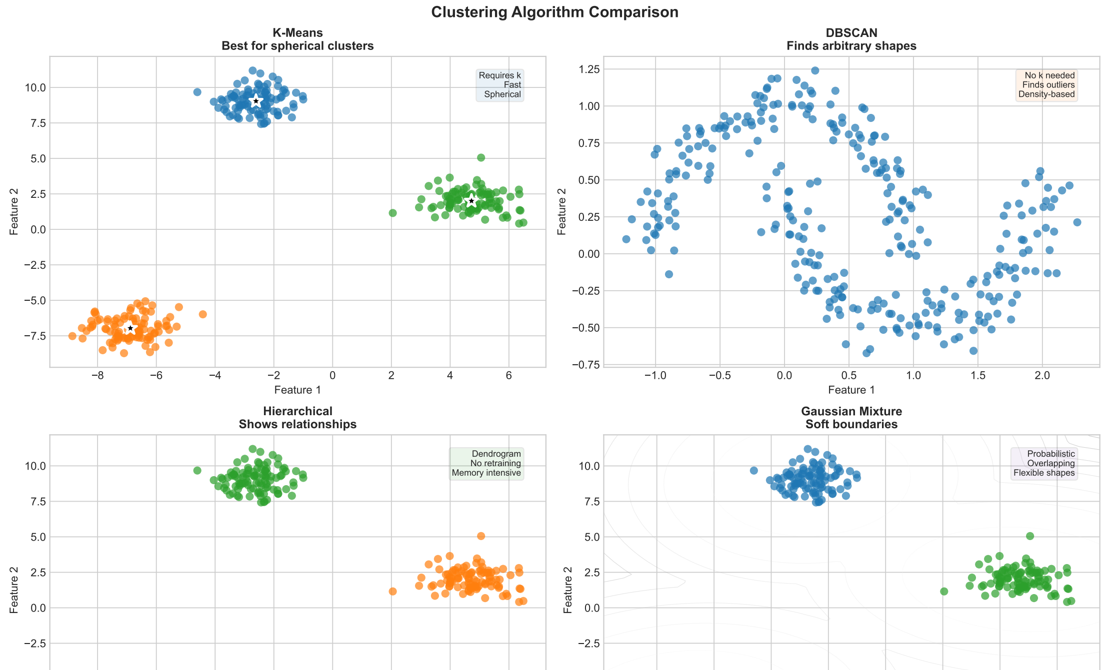
Perfect when:

- Groups overlap
- Need probability scores
- Elliptical cluster shapes
- Soft assignments needed

Algorithm Visual Comparison

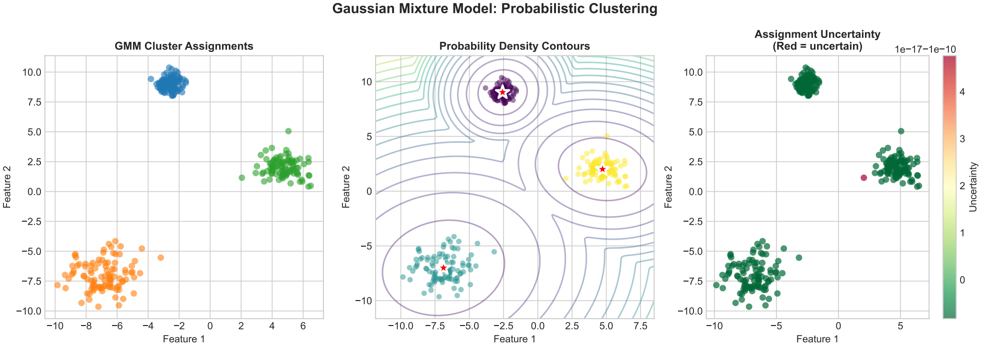
Same Data, Different Approaches

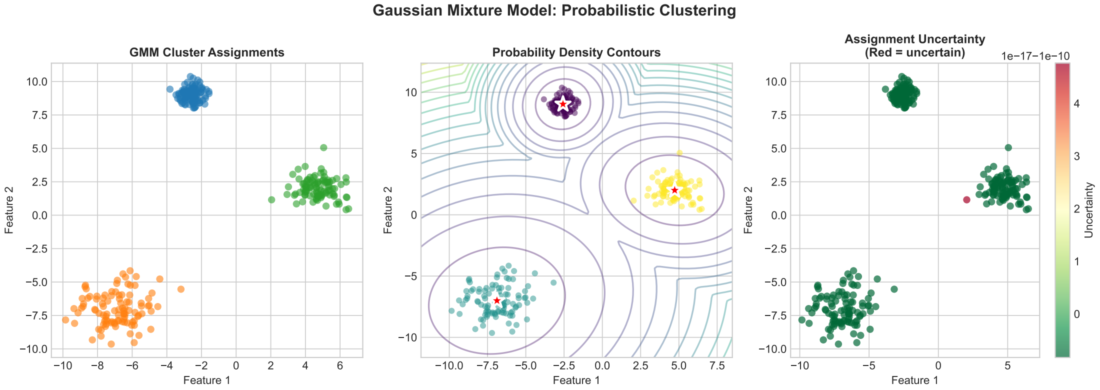



Clustering Algorithm Comparison
K-Means
Best for spherical clusters
Requires k
Fast
Spherical
Feature 1
Feature 2
DBSCAN
Finds arbitrary shapes
No k needed
Finds outliers
Density-based
Feature 1
Feature 2
Hierarchical
Shows relationships
Dendrogram
No retraining
Memory intensive
Feature 2
Gaussian Mixture
Soft boundaries
Probabilistic
Overlapping
Flexible shapes
Feature 2

Gaussian Mixture Models (GMM)

Soft Clustering for Overlapping Innovation Categories




Gaussian Mixture Model: Probabilistic Clustering
GMM Cluster Assignments
Probability Density Contours
Assignment Uncertainty
(Red = uncertain)
Feature 1
Feature 2
Uncertainty
1e−17−1e−10

Fixed K Gives One View

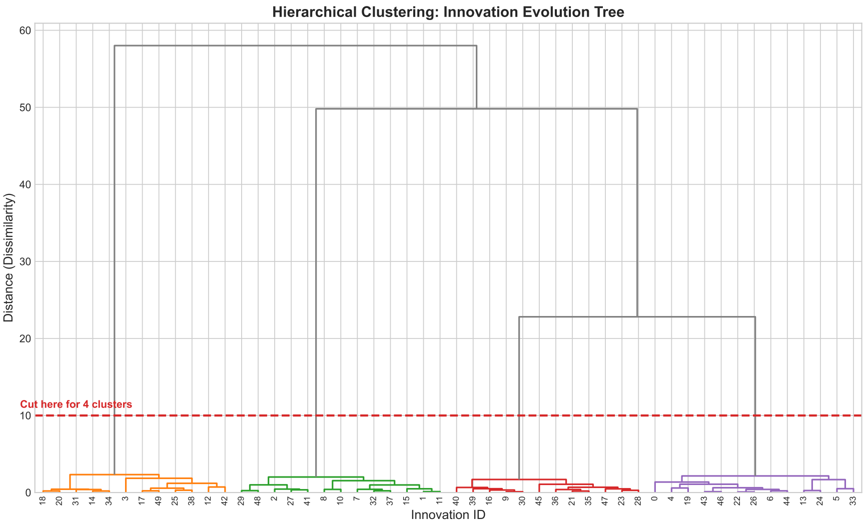
But real relationships are hierarchical:

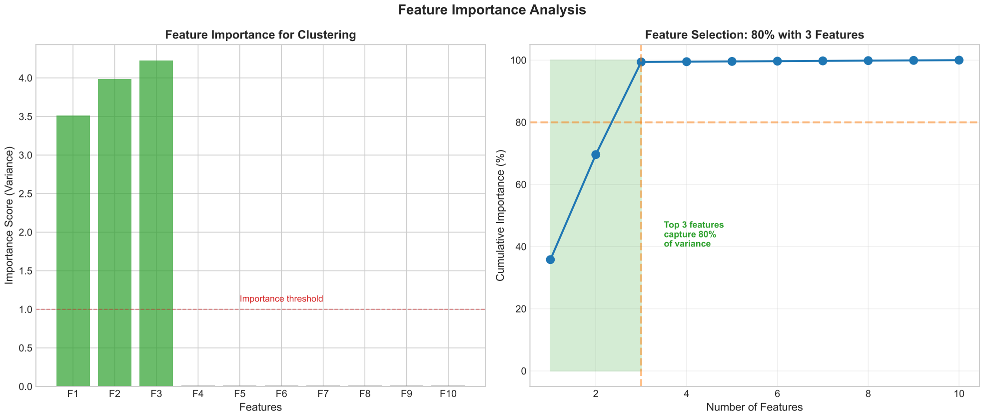
- Organization: Company → Department → Team → Individual
- Geography: Country → Region → City → Neighborhood
- Products: Category → Subcategory → Brand → SKU
- Innovations: All → Categories → Sub-types → Specific solutions

K-means: Pick 5 groups and that's it

What if we need flexibility?

Solution: See the full hierarchy, cut where needed

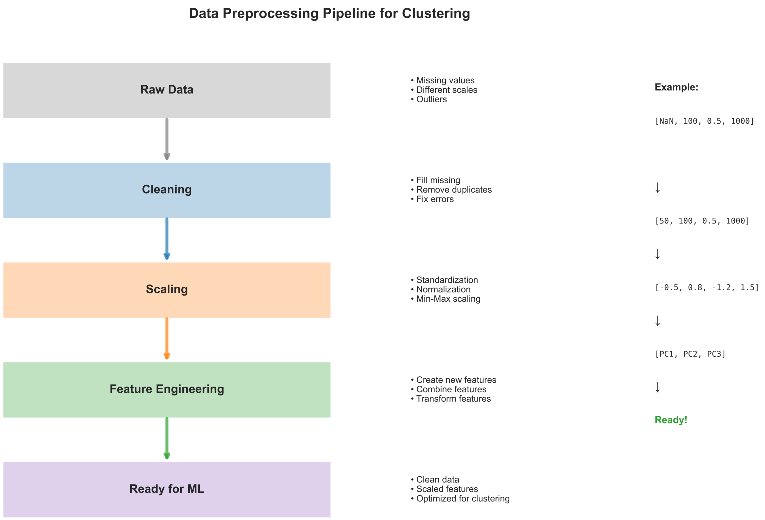

Hierarchical Clustering: Innovation Evolution Tree
Distance (Dissimilarity)
0
10
20
30
40
50
60
Cut here for 4 clusters
18 20 31 14 34 3 17 49 25 38 12 42 29 48 2 27 41 8 10 7 32 37 15 1 11 40 39 16 9 30 45 36 21 35 47 23 28 0 4 19 43 46 22 26 6 44 13 24 5 33
Innovation ID

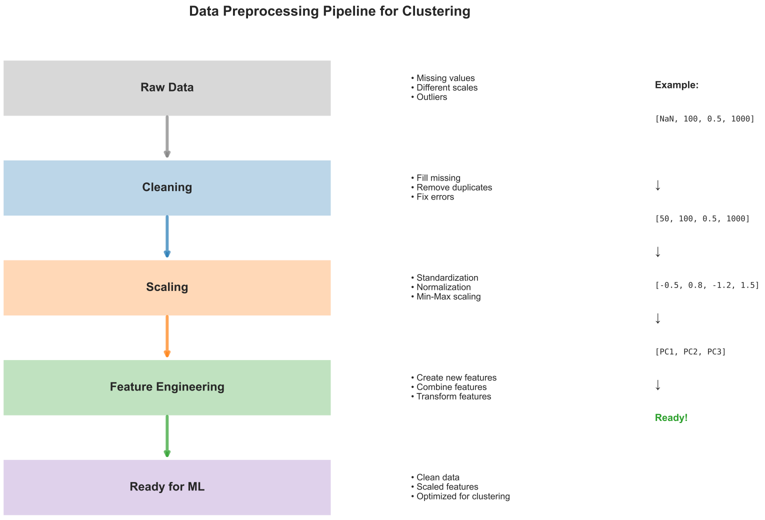


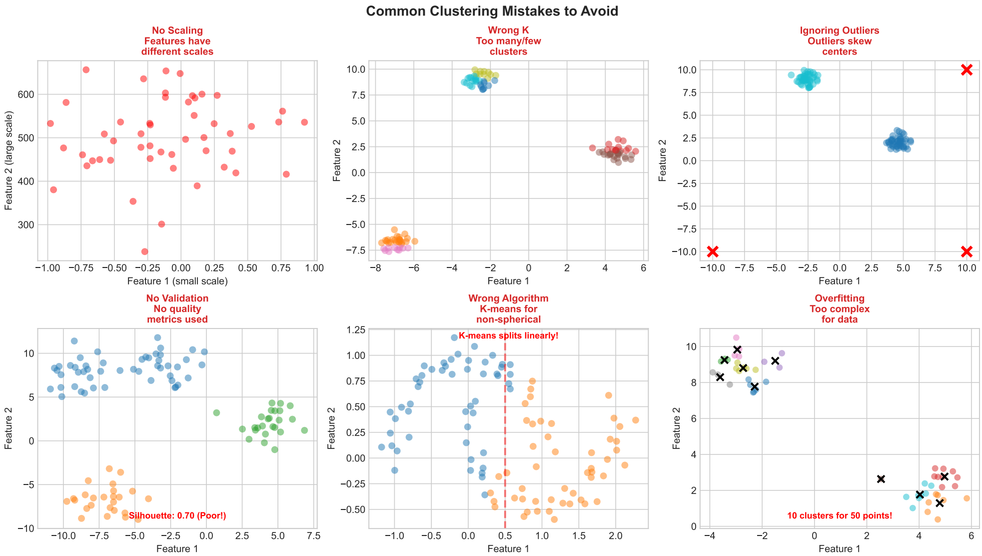
Key Insight: Usage frequency matters most!

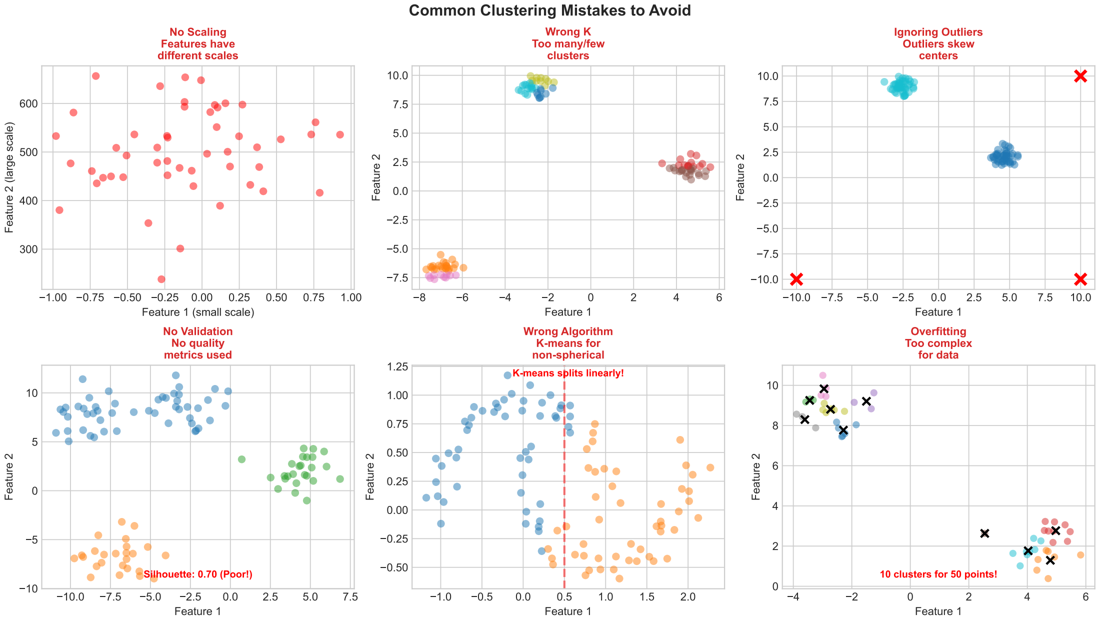
Data Preprocessing Pipeline

From Raw Data to Clustering-Ready Features




Data Preprocessing Pipeline for Clustering
Raw Data
• Missing values
• Different scales
• Outliers
Cleaning
• Fill missing
• Remove duplicates
• Fix errors
Scaling
• Standardization
• Normalization
• Min-Max scaling
Feature Engineering
• Create new features
• Combine features
• Transform features
Ready for ML
• Clean data
• Scaled features
• Optimized for clustering
Example:
[NaN, 100, 0.5, 1000]
↓
[50, 100, 0.5, 1000]
↓
[-0.5, 0.8, -1.2, 1.5]
↓
[PC1, PC2, PC3]
↓
Ready!

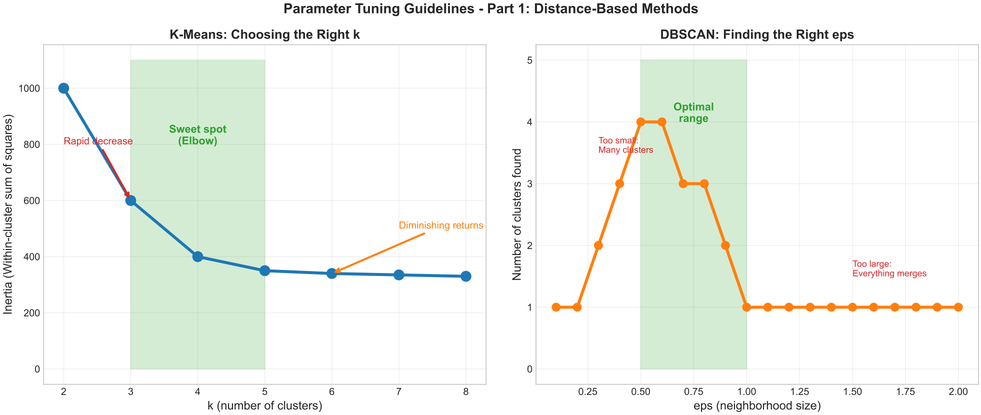

Common Clustering Mistakes to Avoid
No Scaling
Features have different scales
Feature 2 (large scale)
Feature 1 (small scale)
Wrong K
Too many/few clusters
Feature 2
Feature 1
Ignoring Outliers
Outliers skew centers
Feature 2
Feature 1
No Validation
No quality metrics used
Silhouette: 0.70 (Poor!)
Feature 2
Feature 1
Wrong Algorithm
K-means for non-spherical
K-means splits linearly!
Feature 2
Feature 1
Overfitting
Too complex for data
10 clusters for 50 points!
Feature 2
Feature 1

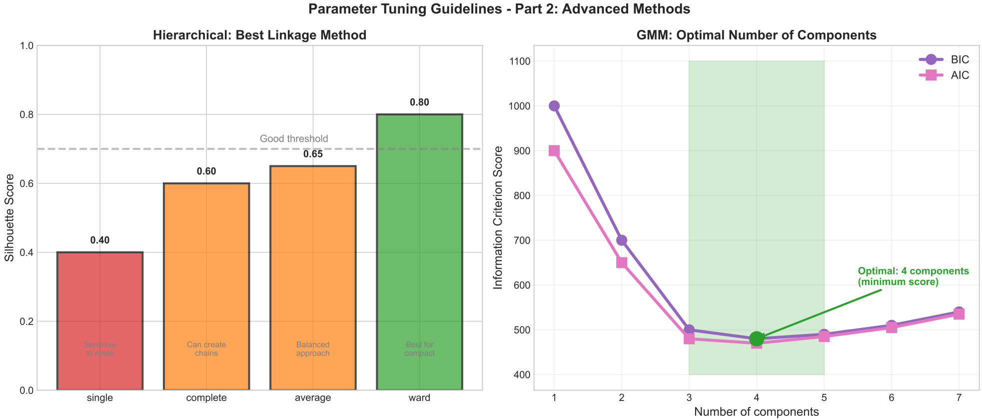


Common Clustering Mistakes to Avoid
No Scaling
Features have
different scales
Feature 2 (large scale)
Feature 1 (small scale)
Wrong K
Too many/few
clusters
Feature 2
Feature 1
Ignoring Outliers
Outliers skew
centers
No Validation
No quality
metrics used
Silhouette: 0.70 (Poor!)
Wrong Algorithm
K-means for
non-spherical
K-means splits linearly!
Overfitting
Too complex
for data
10 clusters for 50 points!

Parameter Tuning Guidelines (Part 1)

K-Means and DBSCAN Parameters





Parameter Tuning Guidelines - Part 2: Advanced Methods
Hierarchical: Best Linkage Method
Silhouette Score
1.0
0.8
0.6
0.4
0.2
0.0
Good threshold
0.40
0.60
0.65
0.80
Sensitive to noise
Can create chains
Balanced approach
Best for compact
single
complete
average
ward
GMM: Optimal Number of Components
BIC
AIC
Information Criterion Score
1100
1000
900
800
700
600
500
400
Optimal: 4 components (minimum score)
1
2
3
4
5
6
7
Number of components

Check Your Understanding - Part 2

Technical Concepts Review

Progress: 2/3

Quick Quiz

1. K in K-means stands for:
   - ☐ Kernel
   - ✓ Number of clusters
   - ☐ Constant
2. DBSCAN finds:
   - ☐ Only circles
   - ✓ Any shape clusters
   - ☐ Exactly K groups

Can You Calculate?

If Silhouette Score = 0.75:

- Is this good? Yes!
- Range is [-1, 1]
- Higher = better separation

Remember:

- Elbow method finds optimal K
- Scale your data first!

Great job! Now let's apply these concepts!

Next: Design integration, innovation patterns, and real-world applications

We've learned the technical tools:

Clustering, metrics, quality measures

But clusters are just numbers...

Until we connect them to innovation opportunities

Let's transform data into innovation insights

Each cluster represents innovation opportunities and patterns

PART 3

Innovation Pattern Analysis

What we'll create:

- Data-driven innovation archetypes
- Innovation pattern maps per category
- Cluster-specific journeys
- Opportunity heat maps
- Design priority matrices

Where ML reveals innovation patterns

Part 3: Learning Objectives

Innovation Applications You'll Explore

By the end of Part 3, you will be able to:

- **Create** innovation archetypes
- **Map** innovation patterns
- **Design** opportunity matrices
- **Analyze** innovation lifecycles
- **Build** ecosystem maps
- **Prioritize** innovation efforts

Design Outcomes

- Innovation taxonomy framework
- Cluster-based strategies
- Data-driven prioritization
- Opportunity identification
- Pattern recognition skills
- Ecosystem understanding

PART 3

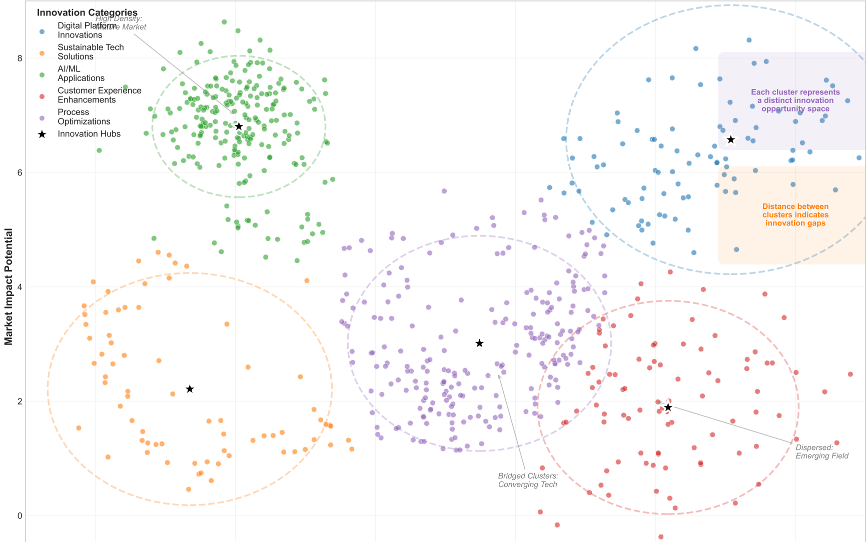
Design Integration

Bridging Technology & Human Experience

From Data Points to Innovation Insights

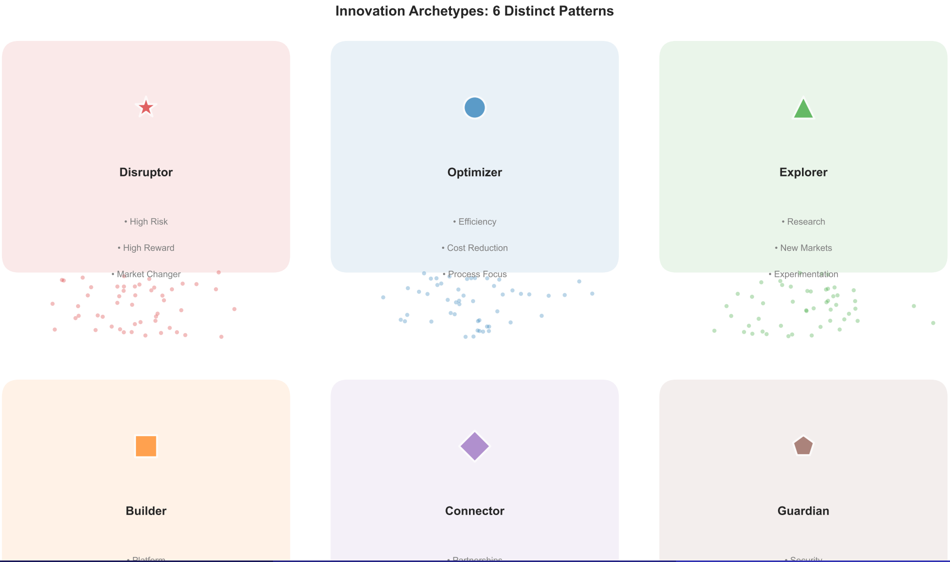
Bridging the Technical-Human Gap

Innovation Patterns Discovered Through Clustering



AI-Generated Innovation Archetypes

Data-Driven Character Development

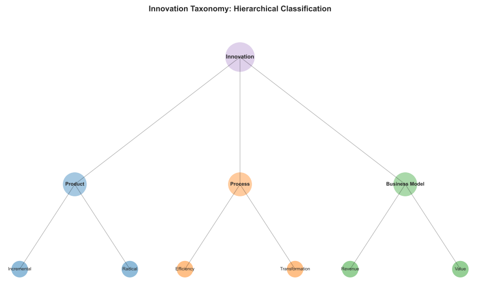


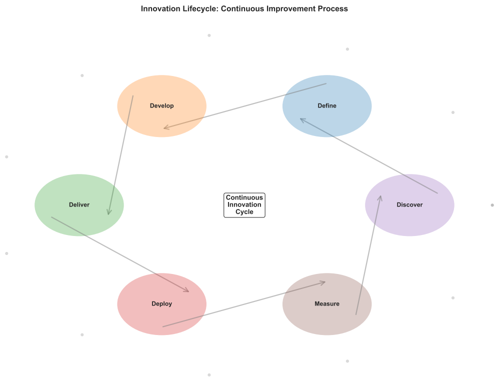
Innovation Pattern Maps

Cluster-Specific Insights

Innovation Framework

Taxonomy and Lifecycle Stages





Framework Levels:

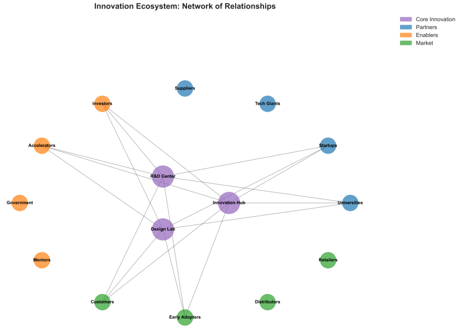
- Types & relationships
- Impact measurements
- Strategic positioning

Lifecycle Stages:

- Ideation & discovery
- Development & testing
- Launch & scaling

Innovation Ecosystem & Journey Mapping

From Networks to Evolution Paths



Evolution Paths:

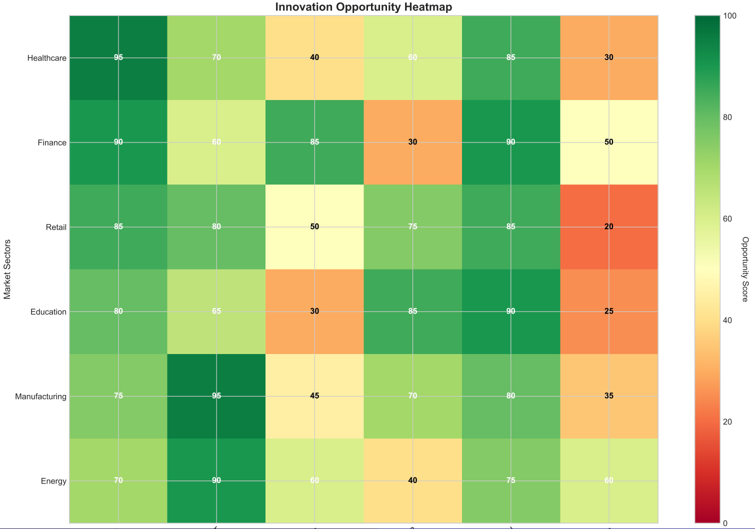
- Different speeds
- Varying trajectories
- Unique milestones

Ecosystem Elements:

- Network connections
- Stakeholder clusters
- Value flows

Innovation Opportunities by Cluster

Where Each Category Has Potential



Key Findings:

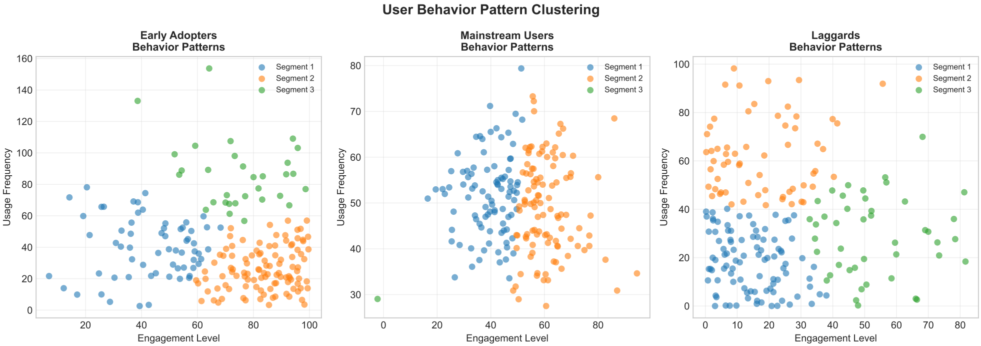
- Emerging tech: Early stage
- Disruptive: Scalability
- Incremental: Integration
- Platform-based: Network effects

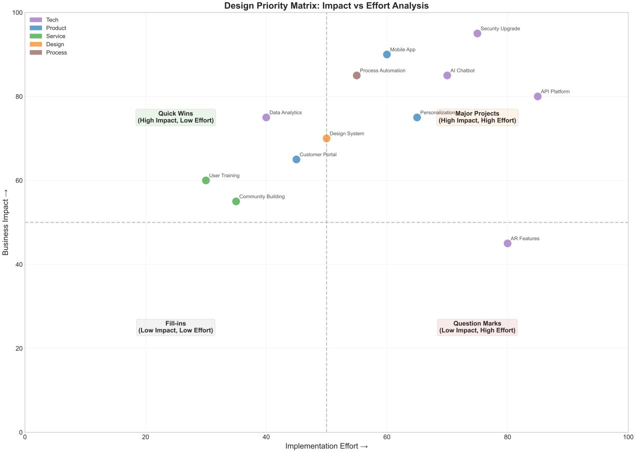
Design implication:

One solution won't fit all!

Innovation Patterns Revealed

What Clusters Tell Us About Evolution



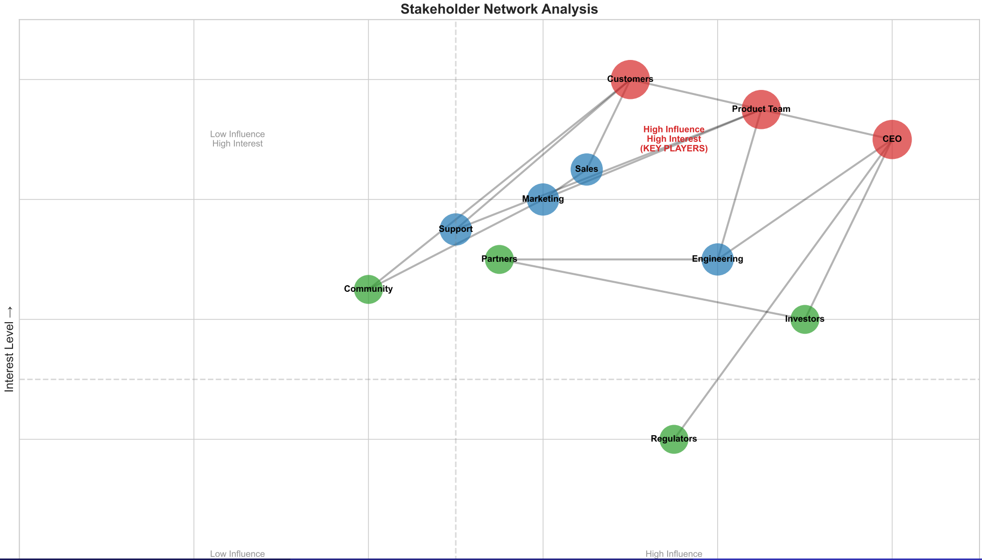


Priority Quadrants:

- **High Impact + High Effort**
  Strategic initiatives
- **High Impact + Low Effort**
  Quick wins
- **Low Impact + Low Effort**
  Fill-ins
- **Low Impact + High Effort**
  Avoid

Understanding Innovation Ecosystems

Network Analysis of Innovation Connections



Match the Application

Match algorithm to use case:

1. Customer segmentation → K-means
2. Finding outliers → DBSCAN
3. Creating taxonomy → Hierarchical
4. Overlapping groups → GMM

Design Thinking

How does clustering help in:

- **Empathize**: Find user groups
- **Define**: Identify patterns
- **Ideate**: Discover opportunities
- **Prototype**: Target solutions
- **Test**: Validate segments

Excellent! Ready to practice with real data?

Next: Summary, real-world case studies, and hands-on practice exercise

You've learned:

- The clustering algorithms
- How to validate quality
- Design applications

Now let's see it in action

How these techniques work in practice
to find patterns in data