

# Machine Learning for Smarter Innovation

## Week 2: Clustering for Deep Empathy

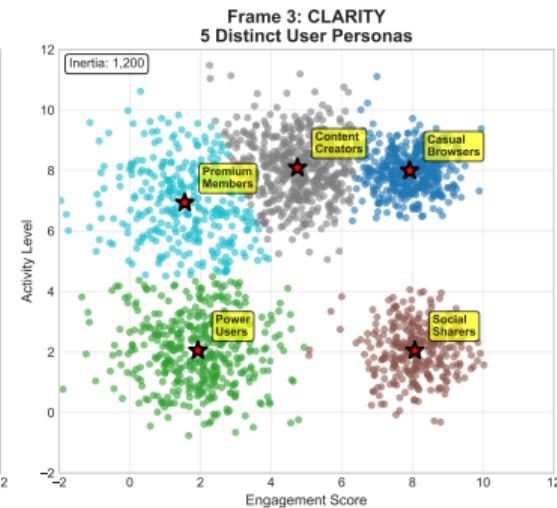
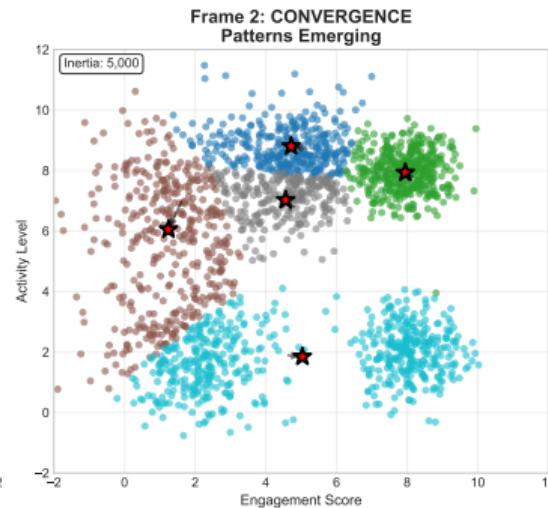
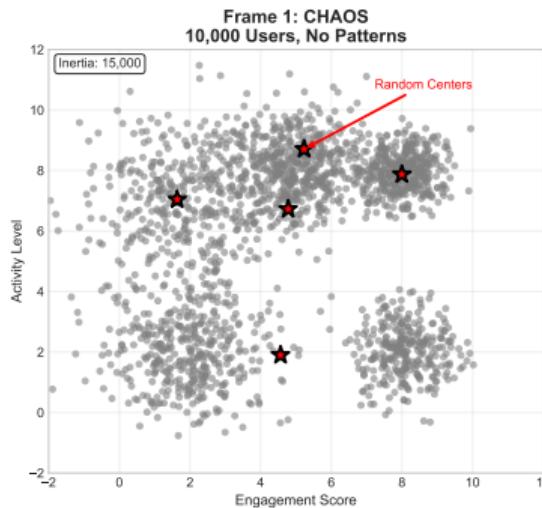
Discovering Hidden User Segments at Scale

BSc Course in AI-Enhanced Innovation

# Watch 10,000 Users Organize Themselves

K-Means Evolution: From Chaos to Clarity

K-Means Evolution: From Chaos to User Understanding



What if users could tell us their natural groups without asking?

From Assumptions

To Discovery

*“One size fits all” actually fits no one well*

# The User Understanding Challenge

Why "Average User" Thinking Fails

## The Problem

### Traditional Approach Failures:

- Design for “average user” → satisfies no one
- Manual personas → based on assumptions
- Small sample surveys → miss edge cases
- Demographics only → ignore behavior patterns
- Static segments → miss evolution

## The Opportunity

### What if we could:

- Find natural user groups automatically?
- Base segments on actual behavior?
- Discover unexpected patterns?
- Track segment evolution?

## One Product

:(: :(: :(: :(: :(:  
:(: :(: :(: :(: :(:  
:(: :(: :(: :(: :(:  
:(: :(: :(: :(: :(:  
:(: :(: :(: :(: :(:  
:(: :(: :(: :(: :(:



# Traditional vs ML-Driven Personas

From Assumptions to Data-Driven Discovery

## Traditional Personas

### Process:

- Interview 10-20 users
- Create fictional characters
- Based on demographics
- Static over time

### Example:

- “Sarah, 35, Marketing Manager”
- “Lives in suburbs”
- “2 kids, busy lifestyle”
- “Values convenience”

### Limitations:

- Confirmation bias
- Small sample size
- Stereotypes

## ML-Driven Segments

### Process:

- Analyze 10,000+ users
- Find natural groupings
- Based on behavior patterns
- Evolve with data

### Discovery:

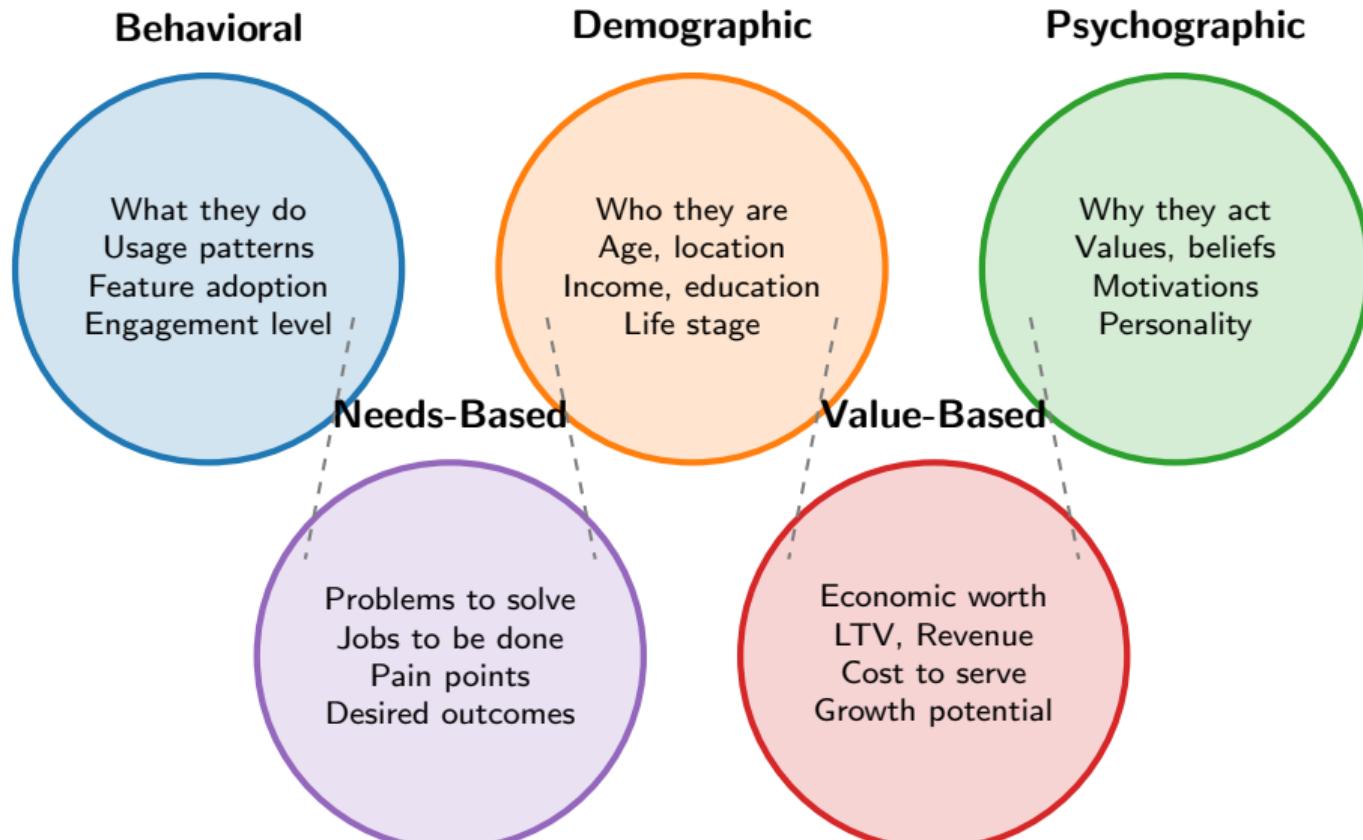
- “Power Feature Users”
- “High engagement, all features”
- “Cross demographics”
- “Worth 10x average user”

### Advantages:

- Data-driven truth
- Full population
- Unexpected insights

# Types of User Segmentation

Different Lenses for Understanding Users

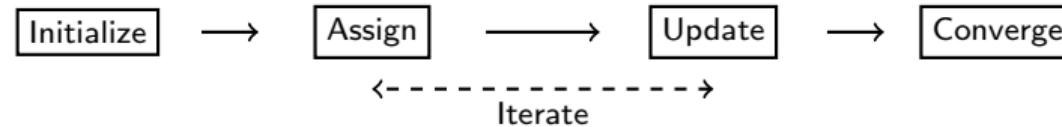


## Part 2: Technical Deep Dive

### The Mathematics Behind User Understanding

# K-Means Algorithm

Simple Rules → Complex Insights



# K-means Algorithm Mechanics

## The Four-Step Dance

### The Algorithm:

- ① **Initialize:** Choose  $k$  random centers
- ② **Assign:** Each point  $\rightarrow$  nearest center
- ③ **Update:** Centers move to mean
- ④ **Repeat:** Until centers stop moving

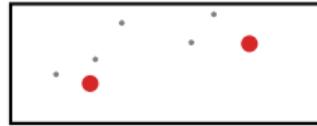
### Mathematical Objective:

$$\min \sum_{i=1}^n \sum_{j=1}^k w_{ij} \|x_i - \mu_j\|^2$$

Where:

- $x_i$  = data point  $i$
- $\mu_j$  = center of cluster  $j$
- $w_{ij} = 1$  if  $x_i$  belongs to cluster  $j$

Step 1: Initialize



Step 2: Assign



Step 3: Update



Step 4: Converged



Centers stable!

# Distance Metrics & Centroids

How We Measure "Similar"

## Euclidean



"As crow flies"

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

**Use when:**

- Physical distance
- Continuous features
- Equal scale

## Manhattan



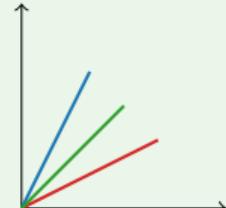
"City blocks"

$$d = |x_2 - x_1| + |y_2 - y_1|$$

**Use when:**

- Grid-like data
- Feature differences
- Outlier robust

## Cosine



"Direction"

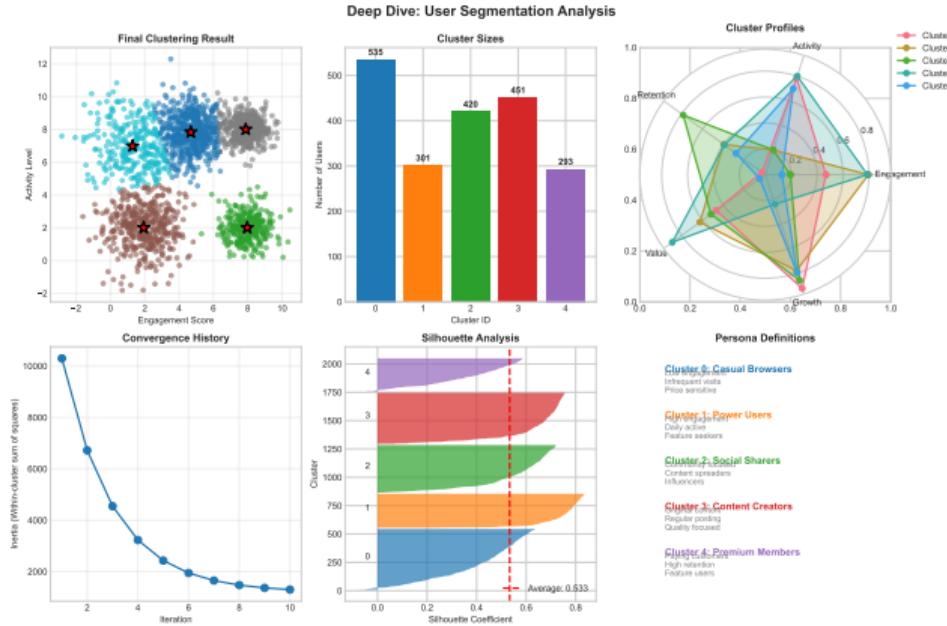
$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

**Use when:**

- Text/documents
- High dimensions
- Magnitude varies

# Convergence & Optimization

When Do We Stop?



## Convergence Criteria:

- Centers move  $\downarrow$  threshold
- Inertia plateaus
- Max iterations reached
- No reassignments

## Inertia (Within-cluster SSE):

$$J = \sum_{i=1}^n \min_j \|x_i - \mu_j\|^2$$

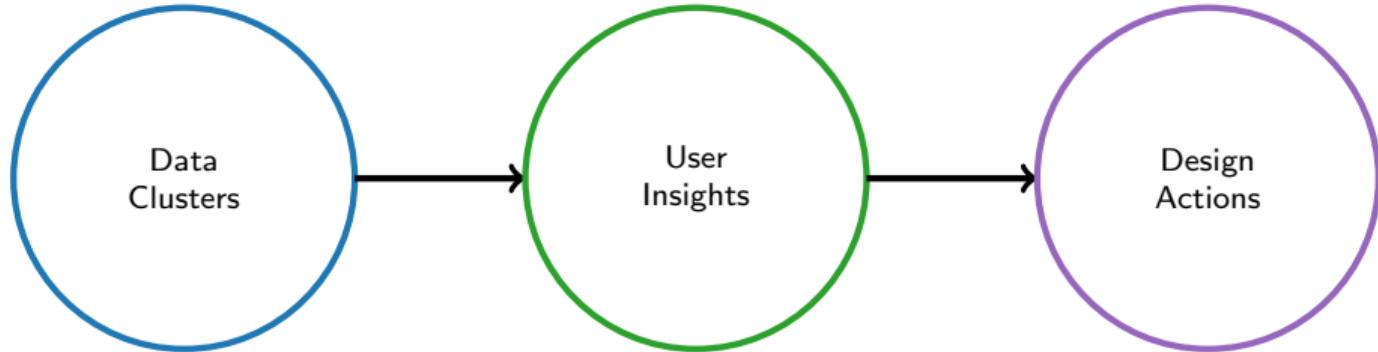
**Warning:** K-means can get stuck in local minima!  
Run multiple times with different initializations.

## Optimization Tricks:

- k-means++ initialization

## Part 3: Design Integration

Transforming Data Clusters into Human Understanding



*“Numbers tell you what, stories tell you why”*

## Part 4: Ethics & Practice

### Clustering with Responsibility

**With Great Patterns**

**Comes Great Responsibility**

# Case Study: Spotify Discover Weekly

30 Million Personalized Playlists Every Monday

## The Challenge:

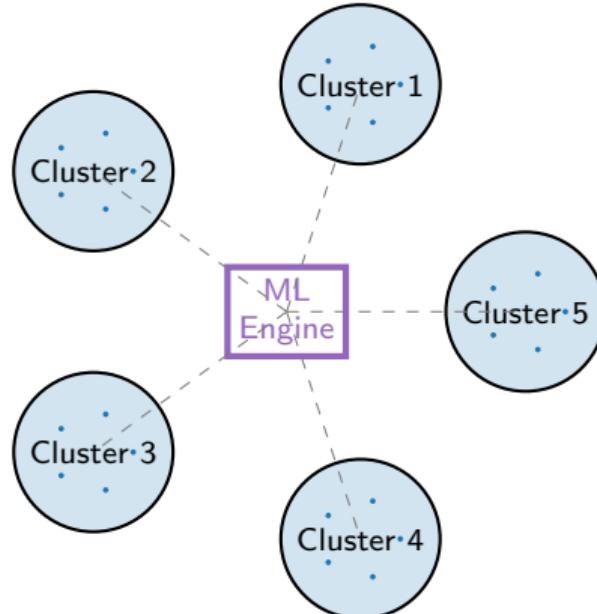
- 500M+ users globally
- 100M+ songs available
- Diverse music tastes
- Discovery paralysis

## The Solution:

- ① Cluster users by listening patterns
- ② Find “taste twins” in same cluster
- ③ Recommend unheard songs from twins
- ④ Personalize 30 songs weekly

## The Impact:

- 40M+ weekly active users
- 60% listen to 10+ songs
- 80% save at least 1 song
- \$1B+ value creation



**Key Insight:** Users with similar taste profiles discover music through each other

# Week 2 Summary

## Key Takeaways

### What We Learned

- K-means finds natural user groups
- Distance metrics matter for meaning
- Clusters evolve into personas
- Multiple methods for different needs
- Ethics crucial for segmentation

### Next Week Preview

#### Week 3: NLP for Emotional Context

- Sentiment analysis at scale
- BERT and transformers
- Emotion detection
- Sarcasm and context
- Voice of customer analysis

### Practical Skills

- Choose optimal k with elbow method
- Validate with silhouette analysis
- Transform clusters to empathy maps
- Detect and handle outliers
- Track segment evolution

### Practice Exercise

#### This Week's Challenge:

- ① Take any dataset with 100+ users
- ② Apply K-means with k=3,4,5
- ③ Use elbow method to find optimal k
- ④ Create personas for each cluster