Machine Learning for Smarter Innovation

Week 2: Clustering for Deep Empathy

Understanding Users Through Data-Driven Segmentation

Discovering Hidden Patterns in Human Behavior

BSc Course in AI-Enhanced Innovation

Department of Computer Science & Design

2025

This week: Transform behavioral data into deep user understanding through advanced clustering

Today's Journey: From Data Points to Human Insights

Your Roadmap to Understanding Users Through Clustering

Morning Session

Part 1: Foundation (25 min)

- Why clustering for empathy?
- Traditional vs data-driven personas
- Setting objectives

Part 2: Technical Deep Dive (35 min)

- Advanced K-means techniques
- Finding optimal clusters
- DBSCAN, Hierarchical, GMM
- Algorithm selection guide

Afternoon Session

Part 3: Design Integration (30 min)

- From clusters to personas
- Building empathy maps
- Journey mapping per segment
- Innovation opportunities

Part 4: Practice (20 min)

- Spotify case study
- 5 music personas discovered
- Implementation pipeline
- Your turn: exercise

Goal: Master data-driven empathy to understand users you've never met

Prerequisites & What You'll Build

Setting You Up for Success in User Understanding

Building on Week 1

You already know:

- Basic K-means clustering
- Elbow method for K
- Silhouette scores
- Distance metrics

New this week:

- User behavior features
- Persona creation methods
- Empathy map construction
- Journey differentiation

What You'll Create

By end of today:

- Data-driven user personas
- Behavioral segment profiles
- Empathy maps per segment
- Differentiated journey maps
- Innovation opportunity matrix

Tools we'll use:

- Python + scikit-learn
- Pandas for data processing
- Matplotlib/Seaborn for viz
- Jupyter notebooks

Remember: Every data point represents a real person with real needs

PART 1

Foundation: Why Clustering for Empathy?

Moving beyond assumptions to data-driven understanding

Key Questions We'll Answer:

- How do we truly understand millions of users?
- What patterns exist in user behavior?
- How can data reveal emotional needs?
- Where do traditional personas fail?

Let's discover hidden user segments

Part 1: Learning Objectives

Foundation Skills You'll Develop

By the end of Part 1, you will:

- **Understand** why clustering enables deep empathy
- **Recognize** limitations of assumption-based personas
- **Identify** behavioral patterns in user data
- **Explain** the value of data-driven segmentation
- **Connect** clustering to design thinking

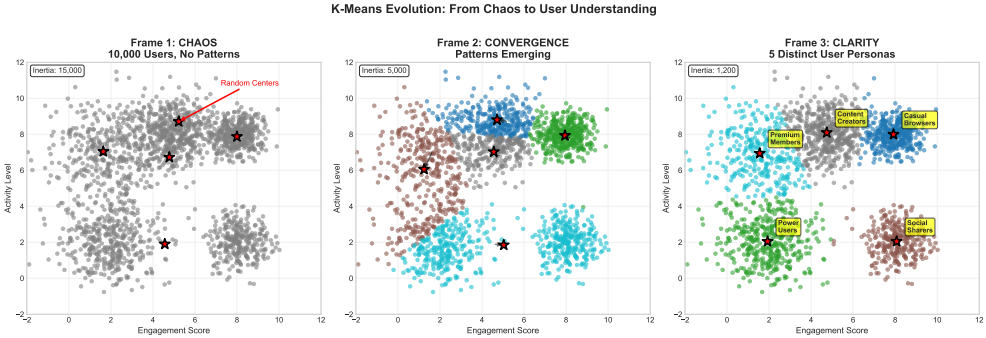
Success Criteria

- Can articulate clustering's role in empathy
- Understand behavioral vs demographic segmentation
- Know when to use data-driven personas
- Can identify suitable user features
- Ready for technical deep dive

Foundation first, then we'll dive into the algorithms

From Chaos to Clarity: The Power of User Clustering

Watch 10,000 Users Self-Organize into Natural Groups



Each dot = a real user, each cluster = shared needs

The User Understanding Challenge: A Deep Dive

Why Traditional Methods Fall Short at Scale

Traditional Challenges

Assumption-Based:

- "Millennials want X"
- "Power users need Y"
- Based on 5-10 interviews

Problems:

- Generic personas
- Missing hidden segments
- Biased by loud voices
- Static profiles
- Limited samples

Result:
70% of features unused

ML-Enhanced Solutions

Data-Driven:

- Behavioral patterns
- Usage analytics
- 1000s of data points

Benefits:

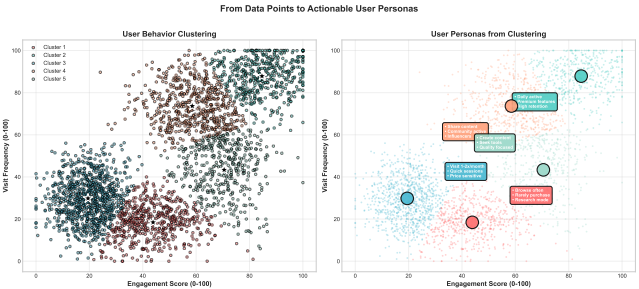
- Natural segments
- Unexpected patterns
- Balanced representation
- Dynamic insights
- Scale to millions

Result:
2x feature adoption

Think about it: How many user types are you missing with traditional personas?

Why Clustering Creates Deeper Empathy

Understanding the "Why" Behind User Behavior



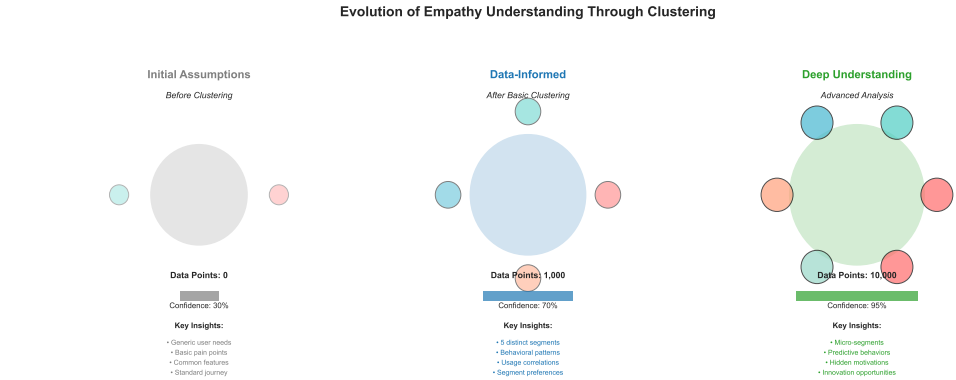
Clustering Reveals

1. **Natural Groupings**
   - Users cluster by behavior
   - Not by demographics
2. **Hidden Patterns**
   - Unexpected user types
   - Cross-demographic needs
3. **Emotional Context**
   - Usage = emotional state
   - Patterns = needs
4. **Evolution**
   - Users change over time
   - Clusters evolve

Key Insight: Behavior reveals needs better than demographics ever could

Evolution: From Assumptions to Data-Driven Insights

The Journey of User Understanding Methods



1990s
Demographics
Age, Gender, Income
0 behavioral data

2000s
Personas
5-10 interviews
Limited scale

2010s
Analytics
Usage tracking
Some patterns

2020s
ML Clustering
Behavioral segmentation
Deep understanding

We're here: Using ML to understand users at scale with empathy

True or False?

1. Clustering finds natural user groups (T)
2. Demographics ¿ behavior for segmentation (F)
3. Traditional personas scale well (F)
4. Clustering reveals hidden patterns (T)
5. Users stay in same segment forever (F)

Can You Explain?

- Why does behavior matter more than demographics?
- How does clustering create empathy?
- What patterns might we discover?
- When should we re-cluster users?

Ready for the technical deep dive?

Next: Advanced clustering algorithms for user understanding

PART 2

Technical Deep Dive

Advanced clustering for nuanced user understanding

What You'll Master:

- Advanced K-means techniques
- Optimal cluster validation
- DBSCAN for outlier users
- Hierarchical for user taxonomies
- GMM for overlapping segments

From algorithms to insights

Technical Skills

- **Implement** K-means variations
- **Validate** cluster quality
- **Choose** optimal K systematically
- **Apply** DBSCAN for outliers
- **Build** hierarchical taxonomies
- **Use** GMM for soft clustering

Application Skills

- Select right algorithm for user data
- Handle outlier users properly
- Validate segment stability
- Interpret cluster characteristics
- Scale to millions of users
- Update clusters dynamically

Each algorithm reveals different aspects of user behavior

K-Means for User Segmentation: Advanced Techniques

Beyond Basic Clustering - Understanding User Nuances

K-Means++ Initialization

Problem: Random init = poor segments
Solution: Smart initialization

1. Choose first center randomly
2. Next center: farthest from existing
3. Repeat until K centers

Result:

- Better convergence
- More stable segments
- Reproducible personas

Mini-Batch K-Means

Problem: Millions of users = slow
Solution: Batch processing

- Sample random batches
- Update centroids incrementally
- Converge faster

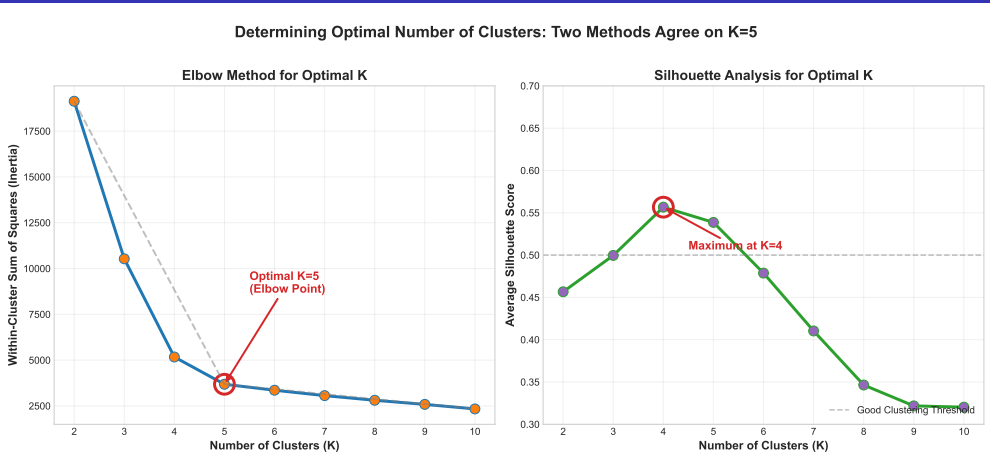
Performance:

- 3-10x faster
- Less than 5% quality loss
- Scales to billions

Pro Tip: Use K-means++ for quality, Mini-batch for scale, combine for production!

Finding the Right Number of User Segments

Data-Driven Approach to Persona Count



Elbow Method

Look for:

Silhouette

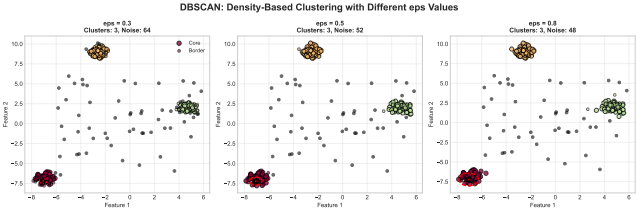
Measures:

Business Logic

Consider:

DBSCAN: Finding Unique Users and Edge Cases

Not Everyone Fits in a Box - And That's Valuable!



DBSCAN Benefits

Finds:

- Core user groups
- Edge users
- True outliers
- Irregular shapes

Perfect for:

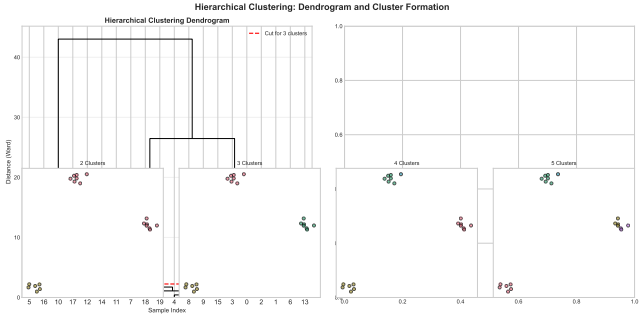
- Power users
- Early adopters
- Special needs
- Fraud detection

No K needed!
Algorithm finds natural groups

Insight: Your most valuable users might be outliers - DBSCAN finds them!

Hierarchical Clustering: Building User Taxonomies

Understanding User Relationships at Multiple Levels



Multi-Level Understanding

Level 1: Broad

- Active vs Passive users

Level 2: Categories

- Creators, Consumers, Curators

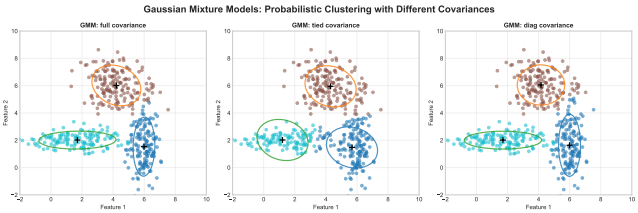
Level 3: Specific

- 8-10 detailed personas

Benefits:

- Flexible granularity
- Natural hierarchy
- Evolution tracking

Use Case: Perfect for creating user documentation at different detail levels



Soft Assignments

User A:

- 70% Power User
- 20% Creator
- 10% Casual

User B:

- 60% Consumer
- 40% Curator

Benefits:

- Realistic modeling
- Probability scores
- Transition detection
- Nuanced personas

Reality: Most users are combinations - GMM captures this complexity!

Clustering Method Selection Guide

K-Means

Pros:
Fast Scalable Simple

Cons:
Fixed K Spherical Sensitive

Well-separated, spherical clusters

DBSCAN

Pros:
No K needed Any shape Noise handling

Cons:
Parameters Density Memory

Arbitrary shapes, noise present

Hierarchical

Pros:
Dendrogram No K upfront Interpretable

Cons:
Slow Memory No undo

Need hierarchy, small datasets

GMM

Pros:
Soft clustering Flexible Probabilistic

Cons:
Complex Slow Assumptions

Overlapping, elliptical clusters

Mean Shift

Pros:
No K Robust Modes

Cons:
Very slow Bandwidth Memory

Mode seeking, computer vision

Implementation: User Clustering Pipeline

Production-Ready Code for User Segmentation

Data Preparation

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

# Load user behavior data
users = pd.read_csv('user_behavior.csv')

# Select features
features = ['sessions', 'duration',
            'actions', 'shares',
            'likes', 'comments']

# Scale features
scaler = StandardScaler()
X = scaler.fit_transform(users[features])

# Find optimal K
from sklearn.metrics import silhouette_score

scores = []
for k in range(2, 10):
    kmeans = KMeans(n_clusters=k)
    labels = kmeans.fit_predict(X)
    score = silhouette_score(X, labels)
    scores.append(score)

optimal_k = scores.index(max(scores)) + 2
```

Clustering & Analysis

```python
# Cluster with optimal K
kmeans = KMeans(
    n_clusters=optimal_k,
    init='k-means++',
    n_init=10,
    random_state=42
)
users['segment'] = kmeans.fit_predict(X)

# Analyze segments
for i in range(optimal_k):
    segment = users[users['segment'] == i]
    print(f"\nSegment {i} ({len(segment)} users):")
    print(segment[features].mean())

# Create persona profiles
personas = users.groupby('segment').agg({
    'sessions': ['mean', 'std'],
    'duration': ['mean', 'std'],
    'actions': ['mean', 'std']
})

# Export for design team
personas.to_csv('user_personas.csv')
users.to_csv('users_segmented.csv')
```

Match the Algorithm

Match use case to algorithm:

1. Finding power users
   → DBSCAN
2. Quick segmentation
   → K-means
3. User taxonomy
   → Hierarchical
4. Mixed behaviors
   → GMM

Can You Code?

Write the code to:

- Load user data (Y)
- Scale features (Y)
- Find optimal K (Y)
- Cluster users (Y)
- Analyze segments (Y)

Excellent! Now let's turn clusters into personas

Next: Design integration - from data to human stories

PART 3

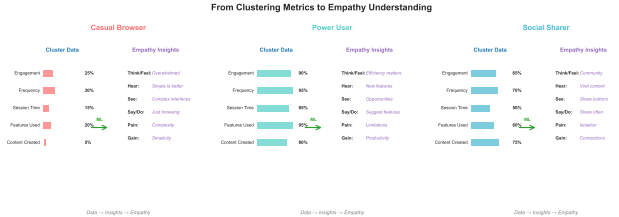
Design Integration

Transforming clusters into empathetic understanding

What You'll Create:

- Data-driven personas with narratives
- Empathy maps per segment
- Differentiated journey maps
- Pain point matrices
- Innovation opportunities

From numbers to human stories



Persona Building

Cluster 3 → "Creative Curator"
Data says:

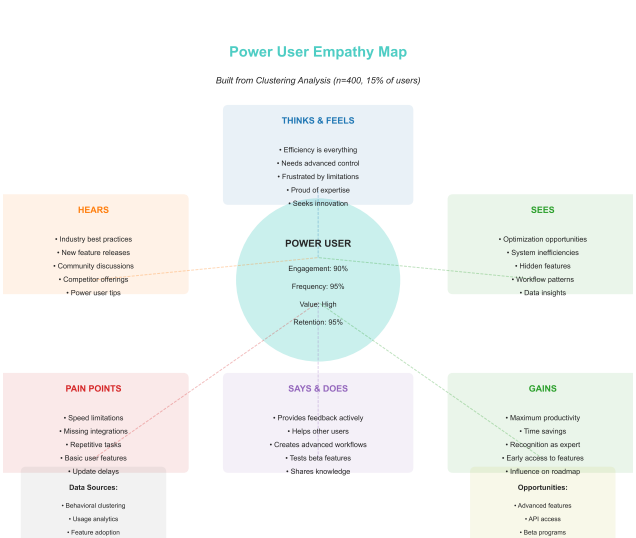
- High sharing rate
- Medium session length
- Peak usage evenings

Persona becomes:

- Sarah, 28, Designer
- Discovers & shares inspiration
- Values quality over quantity
- Needs: curation tools

Key: Data informs, empathy guides

Remember: Clusters are statistical, personas are human - bridge both worlds!



From Data to Empathy

Think (from search data):

- "Is there a better way?"
- "How do others do this?"

Feel (from behavior):

- Frustrated (rage clicks)
- Delighted (shares)

Say (from reviews):

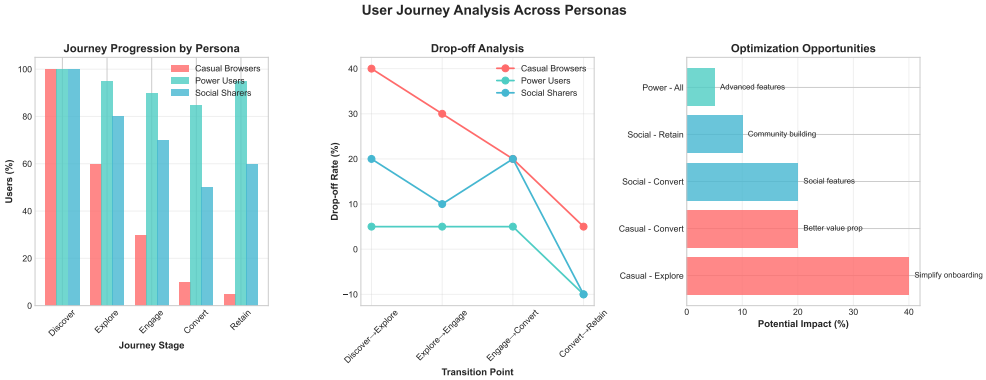
- "Love this feature!"
- "Too complicated"

Do (from analytics):

- Abandons after 3 clicks
- Returns daily

Differentiated Journey Maps: One Product, Many Paths

How Different User Segments Experience Your Product



Power Users

Journey:

- Skip onboarding
- Deep dive features

Casual Users

Journey:

- Need guidance
- Use basics only

Explorers

Journey:

- Try everything
- Test limits

Innovation Opportunity Matrix

Where Each Segment Needs Innovation

[Innovation Opportunity Matrix Chart]

Insight: Different segments = different innovation priorities

PART 4

Practice & Case Study

Real-world application: Spotify's data-driven personas

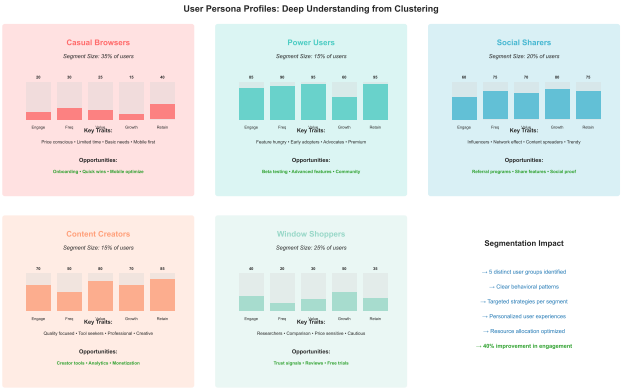
What We'll Explore:

- How Spotify segments 400M users
- 5 core music personas discovered
- Features designed per segment
- Measurable impact
- Your practice exercise

From theory to practice

Case Study: Spotify's Data-Driven Personas

How 400 Million Users Became 5 Core Personas



The Challenge

Problem:

- 400M users
- 180 countries
- Diverse tastes
- One app?

Solution:

- Behavioral clustering
- 5 core personas found
- Personalized features
- Dynamic adaptation

Impact:

- 2x engagement
- 40% less churn
- Higher satisfaction

Spotify's 5 Music Personas: Data-Driven Discovery

Each Persona Gets Different Features

Explorer

Behavior:

- New music daily
- Diverse genres
- Short sessions

Feature:
Discover Weekly

Loyalist

Behavior:

- Same artists
- Full albums
- Long sessions

Feature:
Artist Radio

Social

Behavior:

- Share playlists
- Follow friends
- Group sessions

Feature:
Blend Playlists

Focused

Behavior:

- Background music
- Mood playlists
- Repeat mode

Feature:
Focus Playlists

Curator

Behavior:

- Create playlists
- Organize library
- Edit metadata

Feature:
Enhanced Library

Result: Each user feels Spotify was made for them

Practice Exercise: Segment Your Users

Hands-On: From Raw Data to Personas

Your Task

Dataset: E-commerce users
Size: 10,000 users
Features: 15 behavioral metrics

Steps:

1. Load and explore data
2. Preprocess features
3. Find optimal K
4. Cluster users
5. Analyze segments
6. Create 3 personas
7. Build empathy maps
8. Design features

Time: 45 minutes
Deliverable: Persona cards

Starter Code

```
Load your data users = pd.read_csv('ecommerce_users.csv')
Explore print(users.info()) print(users.describe())
Select behavioral features features = ['visits', 'duration', 'pages_viewed', 'items_bought', 'cart_abandons', 'reviews']
Your code here...  1. Scale features  2. Find optimal K  3. Cluster users  4. Analyze segments
Template for persona persona_template = {'name': '', 'characteristics': [], 'needs': [], 'pain_points': [], 'opportunities': []}
```

Implementation Checklist

Your Step-by-Step Guide to User Segmentation Success

1. Data Prep

Collect:

- ☐ User behavior data
- ☐ Engagement metrics
- ☐ Feature usage
- ☐ Feedback data

Process:

- ☐ Clean data
- ☐ Handle missing
- ☐ Feature engineering
- ☐ Scale features

2. Cluster

Algorithm:

- ☐ Choose method
- ☐ Find optimal K
- ☐ Validate quality
- ☐ Check stability

Analysis:

- ☐ Segment profiles
- ☐ Size distribution
- ☐ Key differences
- ☐ Edge cases

3. Design

Personas:

- ☐ Create narratives
- ☐ Build empathy maps
- ☐ Map journeys
- ☐ Identify needs

Apply:

- ☐ Design features
- ☐ Prioritize roadmap
- ☐ Test with users
- ☐ Measure impact

Success Metric: Users say "This feels made just for me!"

Key Takeaways: Clustering for Deep Empathy

What You've Mastered Today

Technical

You can now:

- Implement K-means++
- Validate clusters
- Use DBSCAN
- Build hierarchies
- Apply GMM
- Scale to millions

Design

You create:

- Data personas
- Empathy maps
- Journey maps
- Pain matrices
- Opportunity maps
- Feature priorities

Impact

Results in:

- Better products
- Happy users
- Less churn
- More engagement
- Innovation focus
- Scalable empathy

Remember: Every cluster represents real people with real needs

You now have the tools to understand millions of users with empathy!

Resources & Next Week Preview

Continue Your Journey in Data-Driven Empathy

This Week's Resources

Readings:

- "Persona Lifecycle" - Pruitt & Adlin
- "K-means++: The Advantages" - Arthur
- Spotify Engineering Blog

Tools:

- Scikit-learn clustering guide
- Persona template toolkit
- Journey mapping tools

Practice:

- Kaggle customer segmentation
- UCI ML Repository datasets

Next Week: NLP for Emotion

Week 3 Preview:

- Sentiment analysis with BERT
- Understanding user emotions
- Context-aware NLP
- Voice of customer analysis

You'll learn:

- Extract emotions from text
- Detect sarcasm and nuance
- Analyze reviews at scale
- Build emotional journeys

Homework:
Segment a dataset of your choice!

Office Hours: Tuesday 2-4pm — Slack: #ml-empathy — Questions welcome!

This Week's Achievement:

You can now understand user segments at scale with deep empathy

Next Week's Challenge:

Understanding what users feel through their words

Your Homework

1. Choose a dataset
2. Apply clustering pipeline
3. Create 3-5 personas
4. Build empathy maps
5. Share in Slack!

Success Tips

- Start with K-means++
- Always validate clusters
- Combine with qualitative data
- Focus on actionable insights
- Remember: data serves empathy

Questions? Let's discuss!