Week 1: How do we truly understand our users?

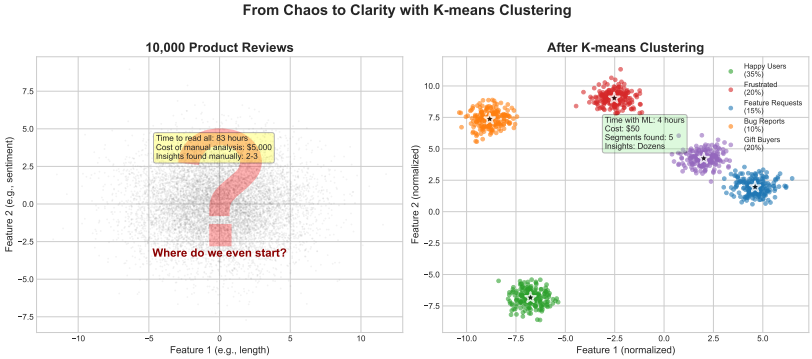
From overwhelming data to clear insights

ML/AI/GenAI for Design Thinking

BSc Course - 12 Week Program

2024

The Challenge: Information Overload



- A startup has **10,000 product reviews**
- Reading all = 83 hours = **2 full weeks**
- Critical insights hidden in the noise

Traditional Approach: Manual Sampling

What we typically do:

- Read 100 random reviews (1%)
- Take notes on patterns
- Draw conclusions
- Hope it represents all users

The problems:

- Miss 99% of insights
- Confirmation bias
- Random sampling misses minorities
- Time-consuming even for 1%

Result: Incomplete understanding, missed opportunities

The Hidden Cost: What We're Missing

Hidden Segments

- Power users (5%)
- Edge cases (8%)
- Silent majority (40%)

Unexpected Uses

- Creative workarounds
- Unintended features
- New market opportunities

Critical Issues

- Rare but severe bugs
- Accessibility problems
- Cultural differences

We're designing blind!

Today's Mission: ML-Powered Understanding

From This...

Reading 100 reviews → Guessing about 10,000 users

...To This

ML analyzes 10,000 reviews → Clear user segments

What you'll learn:

- K-means clustering
- Digital empathy at scale
- Pattern discovery

What you'll achieve:

- Hear ALL voices
- Find hidden patterns
- Make data-driven decisions

What is Clustering?

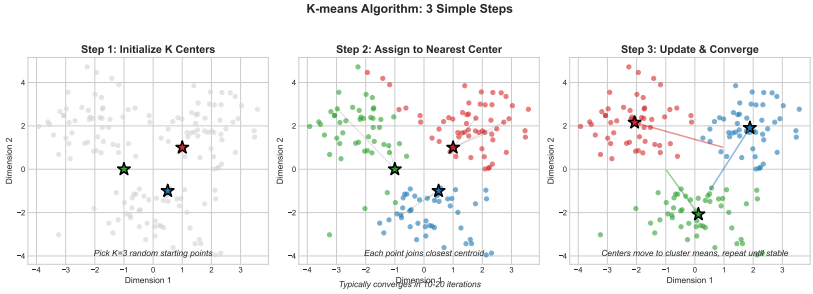
Finding natural groups without labels

Real-world analogy: Sorting cookies

- You have 100 mixed cookies
- No labels or categories given
- You naturally group by similarity:
  - Chocolate chip together
  - Sugar cookies together
  - Oatmeal raisin together
- Groups emerge from characteristics

Clustering does this automatically with data!

How K-means Works: Step 1 - Initialize



Step 1: Pick K random centers

- K = number of groups we want
- Start with random positions
- These are "centroids" (group centers)

Example: K=3 for three user types

How K-means Works: Step 2 - Assign

Step 2: Assign points to nearest center

For each data point:

1. Calculate distance to all centers
2. Assign to closest center
3. Point joins that cluster

Distance = Similarity

- Close = similar
- Far = different
- Uses Euclidean distance

Points naturally group with similar neighbors

How K-means Works: Step 3 - Update and Repeat

Step 3: Recalculate centers and repeat

1. Find new center of each cluster (mean position)
2. Centers move to better positions
3. Repeat assignment step
4. Continue until centers stop moving

Convergence:

- Usually 10-20 iterations
- Centers stabilize
- Clusters are final

Result: Natural groups discovered automatically!

The Math (Simplified)

Only one concept to understand: Distance = Similarity

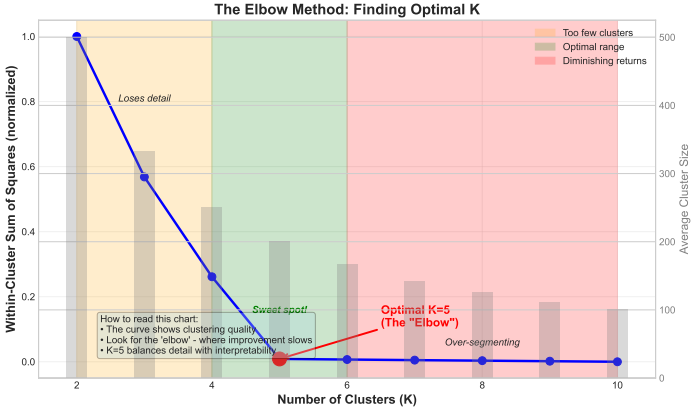
$$\text{Distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

In plain English:

- How different are two reviews?
- Smaller distance = more similar
- We minimize total distance within clusters

That's all the math you need!

Choosing K: The Elbow Method



How many clusters?

- Try different K values (2, 3, 4, 5...)
- Measure cluster quality
- Look for the “elbow” - where improvement slows
- Usually between 3-8 for user segmentation

Real Example: Clustering Product Reviews

Input: 10,000 smartphone reviews

K-means discovers 5 clusters:

1. **Happy Users** (35%): "love", "perfect", "amazing"
2. **Frustrated Users** (20%): "broken", "disappointed", "waste"
3. **Feature Requests** (15%): "wish", "should", "needs"
4. **Bug Reports** (10%): "crash", "freeze", "error"
5. **Comparison Shoppers** (20%): "better than", "compared to", "vs"

Found without reading a single review!

What Clustering Reveals

Patterns invisible to humans:

Expected findings:

- Happy vs unhappy
- New vs returning users
- Price-sensitive segments

Surprise discoveries:

- Gift purchasers (different language)
- Weekend vs weekday users
- Seasonal patterns
- Cultural differences

ML sees what we miss

Unsupervised Learning: No Labels Needed

Supervised Learning

- Needs labeled examples
- "This is spam" / "This is not"
- Learns from answers
- Predicts categories

Unsupervised Learning

- No labels required
- Discovers structure
- Finds patterns
- Reveals unknown groups

Perfect for exploration when you don't know what you're looking for!

When to Use Clustering in Design

Perfect for:

- **User Segmentation**: Find distinct user groups
- **Behavior Analysis**: Discover usage patterns
- **Content Organization**: Group similar items
- **Anomaly Detection**: Find outliers
- **Feature Discovery**: Identify themes

Not great for:

- When you need specific categories
- Very small datasets (¡100 points)
- When groups overlap heavily

What is Empathy in Design?

Understanding users' needs, feelings, and contexts

The foundation of human-centered design:

- Walk in users' shoes
- Feel their frustrations
- Understand their goals
- Recognize their constraints

Why it matters:

- Solve real problems, not assumed ones
- Create products people actually want
- Build emotional connections

"Fall in love with the problem, not the solution"

Traditional Empathy Methods

Interviews

- Deep insights
- Personal stories
- 5-20 people
- 2 weeks effort

Observations

- Natural behavior
- Context matters
- Time-intensive
- Small sample

Surveys

- Larger scale
- Structured data
- Surface-level
- Response bias

The Tradeoff: Depth vs Scale

The Scale Problem in Modern Design

Digital Empathy: Data as Voice

"Every click, review, and interaction tells a story"

Digital footprints reveal:

- What users do (behavior)
- What they say (reviews)
- When they struggle (errors)
- What they want (searches)
- How they feel (sentiment)
- What they value (choices)
- Where they quit (dropoff)
- Why they return (loyalty)

ML helps us listen to thousands of voices simultaneously

Human + Machine Partnership

What ML Does Well:

- Process volume
- Find patterns
- Remove bias
- Work 24/7
- Quantify sentiment

What Humans Do Well:

- Interpret meaning
- Understand context
- Feel emotion
- Make connections
- Create solutions

Together: Scalable Empathy

- ML finds the patterns
- Humans understand why they matter
- Result: Deep insights at scale

The 4-Step Process

1. **COLLECT** - Gather all user data
   - Reviews, support tickets, analytics
   - The more diverse, the better
2. **CLUSTER** - Find natural groups
   - Apply K-means or similar
   - Discover segments automatically
3. **INTERPRET** - Understand each segment
   - What defines each group?
   - What are their needs?
4. **ACT** - Design for each segment
   - Tailored solutions
   - Prioritized roadmap

Real Example: Spotify's Discovery

The Challenge:

- Millions of users, diverse tastes
- How to personalize for everyone?

The ML Solution:

- Clustered listening patterns
- Found unexpected segment: "Cooking Music" (12% of users)
- Never explicitly requested
- Pattern: Upbeat, no explicit lyrics, 30-45 min sessions

The Result:

- Created "Cooking" playlist category
- 8 million+ followers
- Increased engagement 23%

ML revealed what users never told them directly

Avoiding Digital Empathy Pitfalls

Remember:

- **Data shows WHAT, not WHY**
  - Always validate findings with real users
  - Follow up clusters with interviews
- **Correlation is not causation**
  - Ice cream sales and crime both increase in summer
  - Doesn't mean ice cream causes crime!
- **Don't lose the human touch**
  - Numbers tell a story, but people live it
  - Always connect data back to real humans
- **Beware of echo chambers**
  - Vocal minorities can dominate
  - Silent majority matters too

Live Demo: Analyzing 10,000 Amazon Reviews

Our Dataset:

- 10,000 product reviews for wireless headphones
- Mix of 1-5 star ratings
- Various review lengths
- Real customer feedback

Preprocessing Steps:

1. Convert text to numbers (TF-IDF)
2. Normalize features
3. Apply K-means with K=5
4. Analyze results

Let's see what patterns emerge...

Demo Results: 5 Clear Segments Emerge

K-means discovers:

- **Cluster 1 (22%)**: Price-conscious buyers
  - Keywords: “expensive”, “worth”, “cheap”, “value”
- **Cluster 2 (28%)**: Quality enthusiasts
  - Keywords: “sound”, “bass”, “quality”, “crisp”
- **Cluster 3 (18%)**: Convenience seekers
  - Keywords: “easy”, “quick”, “simple”, “portable”
- **Cluster 4 (17%)**: Technical users
  - Keywords: “bluetooth”, “battery”, “connection”, “range”
- **Cluster 5 (15%)**: Gift buyers
  - Keywords: “gift”, “husband”, “daughter”, “birthday”

Surprise Discovery: The Hidden Gift Segment

The 15% we never knew about:

- Never explicitly mentioned "gift" in many reviews
- Found through language patterns:
  - Third-person references ("he loves", "she uses")
  - Occasion mentions ("Christmas", "graduation")
  - Less technical language
  - Focus on packaging and presentation

Why this matters:

- Different decision criteria
- Price less sensitive
- Value presentation and packaging
- Buy multiple units
- Seasonal patterns

A whole market segment hiding in plain sight!

From Discovery to Design Decisions

Segment-specific strategies:

For Price-conscious (22%):

- Highlight value propositions
- Compare to competitors
- Offer bundles

For Quality enthusiasts (28%):

- Technical specifications
- Audio samples
- Expert reviews

For Gift buyers (15%):

- Gift wrapping options
- Gift message features
- Holiday promotions
- Gift guides

For Technical users (17%):

- Detailed specs
- Compatibility info
- Firmware updates

From Clusters to Action: Design Implications

Immediate actions based on clustering:

1. **Personalized landing pages**
   - Different messaging per segment
   - Relevant features highlighted
2. **Targeted email campaigns**
   - Price alerts for value seekers
   - New features for tech enthusiasts
3. **Product development priorities**
   - 28% care most about sound quality
   - Focus R&D on audio improvements
4. **Customer support training**
   - Different segments, different needs
   - Tailored support approaches

What We Achieved Today

Traditional Approach:

- 100 reviews read
- 2 segments identified
- 2 weeks of work
- $5,000 cost
- Lots of guessing

ML-Powered Approach:

- 10,000 reviews analyzed
- 5 segments discovered
- 4 hours total
- $50 compute cost
- Data-driven insights

100x more data, 50x faster, 100x cheaper

Key Learning: ML + Empathy = Understanding at Scale

Next Week's Challenge

We found 5 user segments...

We know they exist. We see their patterns.
We understand their keywords.

But...

What are they actually FEELING?

Happy? Frustrated? Confused? Excited?

Next Week: Transformers and NLP - Understanding Emotion in Text
How BERT reads between the lines to understand sentiment