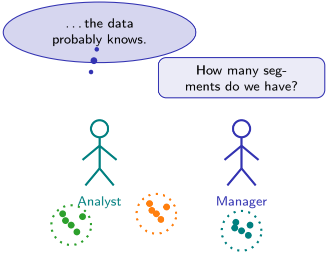
Why Would a Marketing Team Want to Group Customers Without Labels?

The Dilemma

- A bank has thousands of customers but no predefined categories
- Management asks "which customers are similar?" but nobody agrees
- The data has patterns but nobody labeled them

What if the structure is already in the data – waiting to be discovered?



Insight

K-Means formalizes unsupervised discovery: let the algorithm find groups that the data itself defines, rather than imposing human categories.

Unsupervised learning finds structure without labels – the algorithm discovers categories, not confirms them

Sorting a Crowd – Did Clustering Cross Your Mind?

Think Before You Compute

Imagine you are at a networking event with a hundred strangers. Within minutes, you mentally group people: the tech crowd near the coffee, the finance professionals by the window, the academics clustered around the speaker. You did not run an algorithm. You noticed patterns.

- How many groups did you identify?
- What features separated them – dress, language, location?
- Did some people seem to belong to two groups at once?

Pause and reflect:

When you last organized files on your desktop, did you sort them into folders based on perceived similarity – without anyone telling you the folder names?

That is clustering.

Reflection Prompt

Write down one situation where you mentally grouped items or people without being told the categories. How many clusters did you find?

Clustering mirrors how humans naturally organize: by perceived similarity, not by assigned labels

What Makes K-Means Different from DBSCAN, Hierarchical, and GMM?

Taxonomy of Clustering Algorithms

Property	**K-M**	**DBS**	**Hier.**	**GMM**
Choose K	Yes	No	Cut	Yes
Shape	Spher.	Arb.	Any	Ellip.
Outliers	All	Det.	All	Soft
Speed	nKt	$n\log n$	n^2	nK
Output	Hard	Hard+	Dend.	Soft

K-Means is the fastest and simplest, but assumes spherical clusters and assigns every point.

Insight

K-Means wins on speed and simplicity but pays a price: it cannot discover non-spherical clusters or flag outliers.

K-Means: Fast, spherical, hard

DBSCAN: Density, arbitrary shape

Hierarchical: Tree, any K post-hoc

GMM: Probabilistic, soft assign

Spherical assumption means K-Means minimizes within-cluster variance, equivalent to Voronoi tessellation

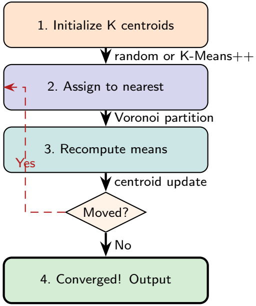
Follow One Iteration from Random Centers to Stable Clusters

One Iteration, Step by Step

- Start with K random centroids
- Assign every point to nearest centroid (Voronoi partition)
- Recompute each centroid as the mean of its assigned points
- Repeat until centroids stop moving
- Convergence guaranteed: each step reduces WCSS

Insight

K-Means always converges, but to a local minimum – not necessarily the global one.



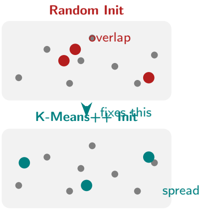
Each iteration is O(nKd): n points, K clusters, d dimensions. Typically converges in few iterations.

Who Should Pick the Starting Centers – Random, K-Means++, or Both?

Three Initialization Strategies

- **Random**: pick K points at random, fast but high variance
- **K-Means++**: distance-proportional sampling, spreads centers apart
- **Multiple restarts**: run R times, keep the run with lowest WCSS

K-Means++ is now the default in scikit-learn for good reason.



Insight

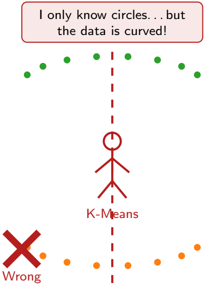
K-Means++ initialization reduces both the expected WCSS and the number of iterations.

K-Means++ guarantees O(log K)-competitive approximation to optimal WCSS (Arthur and Vassilvitskii)

What Could Go Wrong If You Choose the Wrong K?

Three Ways K-Means Fails Silently

- **Wrong K**: too few clusters merge real groups, too many split them
- **Non-spherical data**: K-Means forces round clusters on curved structure
- **Bad initialization**: trapped in a poor local minimum with high WCSS



Insight

K-Means always finds K clusters, even when the true number is different. The algorithm never says "I don't know."

Diagnostics: elbow method for K, silhouette score for cluster quality, visual inspection for shape assumptions

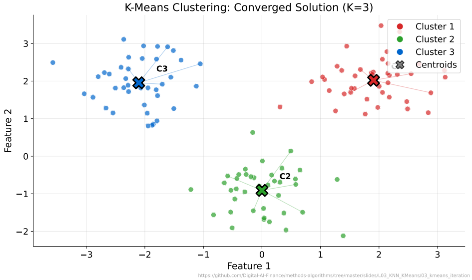
K-Means as a Baseline Clusterer

- Scales linearly with data size
- Simple to implement, explain, and debug
- Results easy to interpret: each cluster has a centroid
- Natural starting point before complex alternatives

The chart shows how cluster boundaries emerge from the iteration process.

Insight

K-Means owes its popularity to the same property as linear regression: the simplest reasonable solution, making it the natural baseline.



K-Means with K=2 is equivalent to finding the optimal split along the first principal component direction

Who Wins and Who Loses When Clusters Replace Categories?

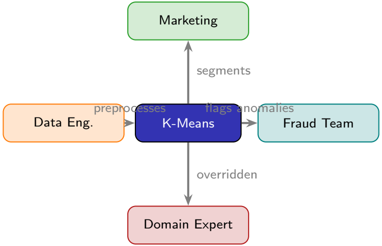
Stakeholder Analysis

- **Winners**: Marketing (data-driven segments), Fraud detection (anomaly = far from centroids), Data preprocessing (cluster features)
- **Losers**: Domain experts (categories overridden), Interpretability advocates, Anyone expecting stable segments over time

K-Means shifts power from domain intuition to data patterns.

Insight

K-Means clusters are not "real" categories – they are mathematical artifacts. The business meaning must be assigned after.



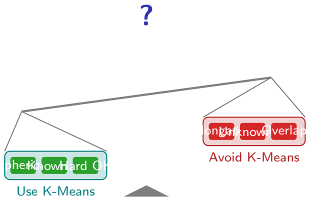
Cluster interpretation requires domain expertise: the algorithm finds groups, humans name them

3 Questions That Reveal Whether K-Means Is the Right Algorithm

The Decision Framework

1. **Are clusters spherical?** – If not, consider DBSCAN or spectral clustering
2. **Do you know how many clusters?** – If not, try elbow, silhouette, or hierarchical
3. **Every point in exactly one group?** – If overlap needed, use GMM

If all three answers are yes, K-Means is a strong candidate.



Insight

No clustering algorithm is universally best. K-Means excels on spherical, well-separated clusters with known K.

The "No Free Lunch" theorem applies to clustering too: no single algorithm dominates all data shapes

Can You Evaluate This Real Clustering Problem?

The Scenario

A retail bank wants to segment its customers. Features: income, transaction amount, transaction count, tenure, credit utilization. All numerical, no predefined categories.

- Apply the 3-question framework from the previous slide
- Decide: Is K-Means appropriate here?
- If yes: recommend K and a validation strategy
- If no: name a better algorithm and explain why

Deliverable

Fill in the table. Be prepared to defend your verdict to a skeptical marketing director.

Question	**Your Answer**
Spherical?	____________
Known K?	____________
Hard assignment OK?	____________
Verdict	____________
Recommended K	____________
Validation method	____________

Hint: consider the feature space shape, the business context, and how you would validate cluster quality