

K-MEANS & FRIENDS

EXPLORING CLUSTERING WITH PHOTOGRAPHS

shaurya agarwal

Outline

- motivation
- movie
- math
- method
- more

motivation

what are we talking about again?

quantization, clustering, k-means and friends

Image Compression, Customer Segmentation, Document Clustering, Anomaly Detection, Feature Learning and Dimensionality Reduction, Medical Imaging, Genomics and Bioinformatics, Speech Recognition, Astronomical Data Analysis, Pattern Recognition and Classification, even enabling efficient training for LLMs

...and unusually enough, a movie

The background is a dark, textured surface with intricate, wavy, light-colored lines that resemble topographical contours or fluid motion. Scattered throughout are small, faint dots, some of which are connected by thin, delicate lines, giving the impression of a star map or a complex data visualization.

movie

cinematography: a rather unusual use case for k-means

i got to do this

cinematography

on

a

budget

powered

by

python

AKRITI SINGH
SURYA RAO
ARSHAD MUMTAZ
SHAHROUKH, SAMIN, HIMANSHU, RONIT, AFZAL,
SHANKAR, YIKAS, SADIQ, VARUN, ASIM, SHOAB

Music by
RABBI SHERGILL

eight down toofaan mail



in cinemas this *february*



AN AKRITI SINGH FILM

storia
senza
storia

W A L
il.com

first principles

the best cinematography has exceptional control on color

first principles

showcase what inspired the film's cinematography

first principles - colors

- one, or two or a few

first principles - colors

- one, or two or a few but no more

first principles - colors

the director ascribes meaning to these one, two or a few to tell the story

first principles - colors

these colors must remain (approximately, perceptually) consistent

throughout the process, across all devices and screens

pre-production

or else*

* I had a **great** director and a very very dedicated team, use your imagination

pre-production

there are expensive cameras and studio setup that offer great control over the process

pre-production

a typical day may cost 100s of 1000s of \$\$\$ for camera and lighting

pre-production

but for a 10-day shoot, interior and exterior

pre-production

we had USD 359 total

(yes, that's three hundred and fifty-nine dollars, at the current exchange rate)

key insight

light – waves – wavelength, wavelength, wavelength

key insight

light – colors – long wavelength, medium wavelength, short wavelength

how do you pick colors?

i used k-means++ to do it

The background is a dark, textured surface with intricate, wavy, light-colored lines that resemble topographical contours or fluid motion. Scattered throughout are small, bright white dots, some of which are connected by thin, faint lines, giving the impression of a starry night sky or a complex data visualization.

let's see how

there really is no math to it...

The background of the slide is a dark, textured surface. It features a series of light-colored, wavy, horizontal lines that flow from the left towards the right, creating a sense of movement. Scattered throughout the background are numerous small, light-colored dots, some of which are connected by thin, faint lines, resembling a network or a constellation.

math*

how clustering works

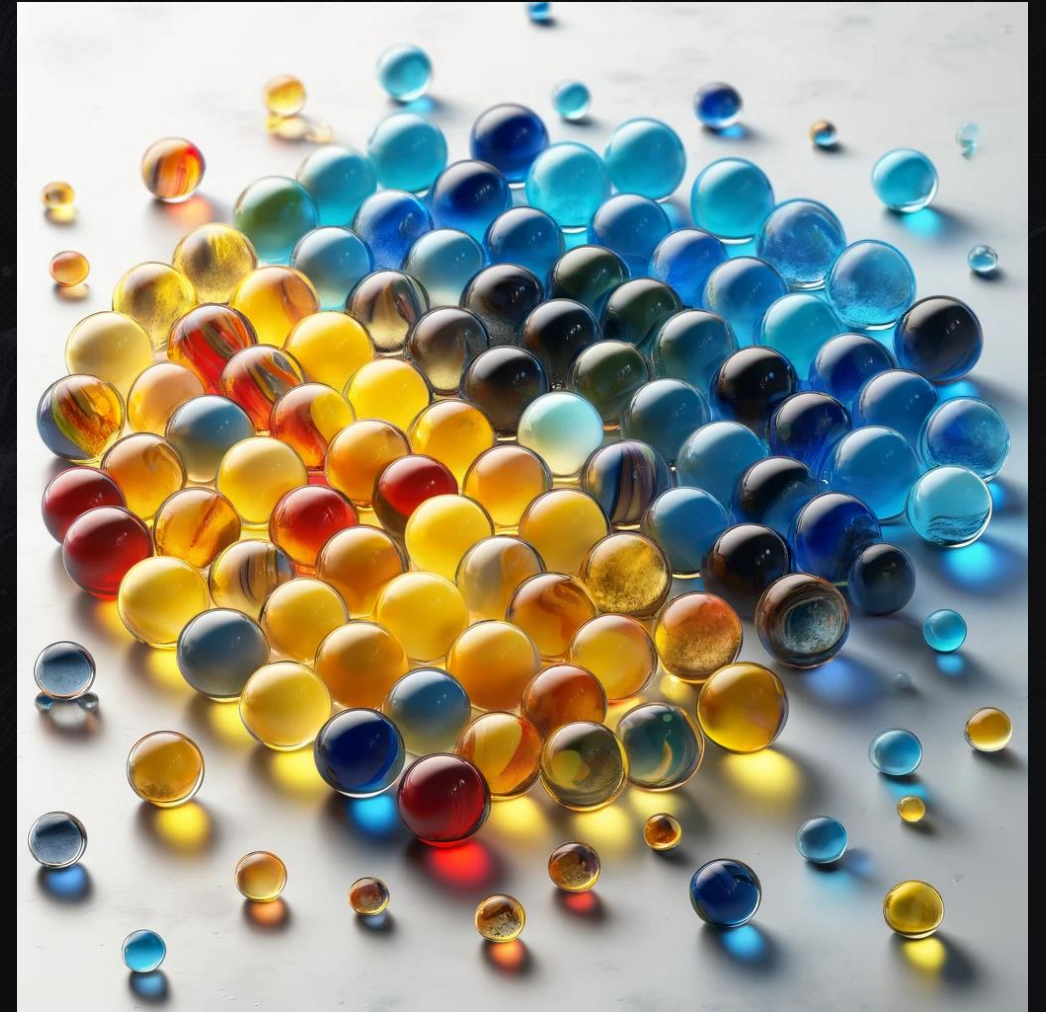
naïve explainer

imagine you have a bunch of marbles of
different colors, and you want to organize
them into groups



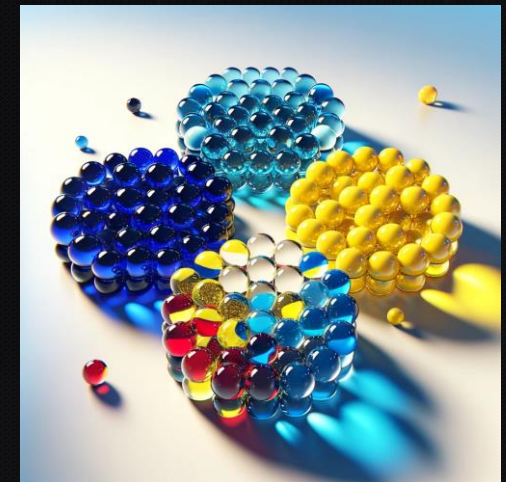
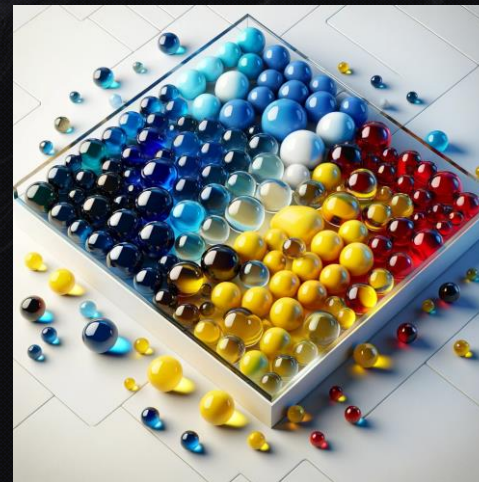
naïve explainer

the k-means method is like
deciding to group the marbles
based on how close they are to
each other (in color)



naïve explainer

the "k" in k-means is deciding how many
groups you want



The background of the slide is a dark, textured surface. It features a series of light-colored, wavy, horizontal lines that flow from the left side towards the right. Scattered throughout the background are numerous small, light-colored dots, some of which are connected by thin, faint lines, giving the impression of a starry sky or a complex network.

steps

Initialize

Assign

Update

Repeat

steps

Initialize

Start by selecting k initial centroids, where k is a predefined number of clusters

Initialize

Assign

Update

Repeat

steps

Assign

each data point to the **closest** centroid,
creating clusters.

Initialize

Assign

Update

Repeat

“distance” is how you find the closest centroid, define distance to mean the unit measure of the feature(s) you cluster on

steps

Update

Recalculate the centroids as the mean of
all points in each cluster

Initialize

Assign

Update

Repeat

steps

Repeat

the assignment and update steps until the
centroids no longer change significantly,
indicating convergence

Initialize

Assign

Update

Repeat

steps

Limit

Often the centroids do not seem to converge, but dance around the convergence points, this is when we stop the algorithm by specifying the maximum number of iterations

Initialize

Assign

Update

Repeat

Limit

framework

Representation

Distance Measure

K

framework

Representation

feature comes in focus

unit distances mean the same thing everywhere.

Representation

Distance Measure

K

framework

Distance Measure

Cartesian, Manhattan, Cosine or custom
based on the feature(s) you want to use to
create the clusters.

Representation

Distance Measure

K

framework

The Right **K**: Elbow Method

Plot the cost (e.g., within-cluster sum of squares) against different k values. The "elbow" point, where the rate of decrease sharply changes, can indicate a good k value.

Representation
Distance Measure
K

framework

The Right **K**: Silhouette Score

Measure how similar an object is to its own cluster compared to other clusters. A high silhouette score suggests the object is well matched to its own cluster and poorly matched to neighboring clusters. The k that maximizes the average silhouette score may be chosen.

Representation

Distance Measure

K

framework

The Right **K**: Gap Statistic

Compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The k with the highest gap statistic suggests the optimal clustering.

Representation
Distance Measure

K

how?

let's see this in action

method

let's look at some code

tech stack

this is compute intensive

we leverage vectorized, just-in-time approach

tech stack

JAX parallelizes and optimizes the compute DAG over NumPy

code

switch to code

The background is a dark, textured surface with intricate, wavy, light-colored lines that resemble topographical contours or fluid motion. Scattered throughout are small, faint dots, giving it a cosmic or digital feel.

MORE

JOIN THE DISCUSSION, GITHUB, MAKE THIS BETTER

open-source

Eight Down Toofaan Mail is now on YouTube

<https://www.youtube.com/watch?v=VnHPtozfhRU>

this project on github: <https://github.com/shauryashaurya/kandinsky>

join me

