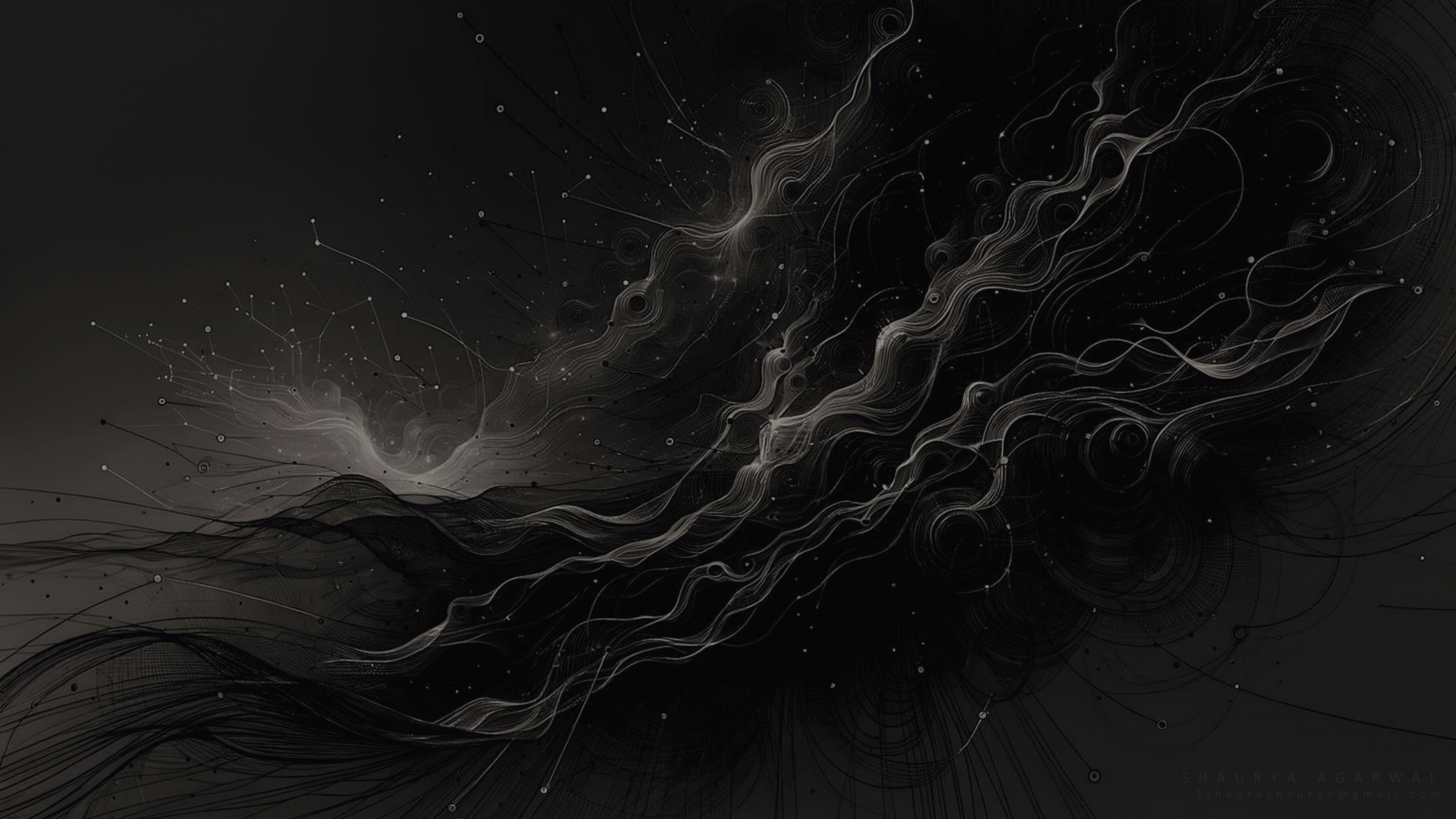

SHAURYA AGARWAL
3shaurashaurya@gmail.com

K-MEANS & FRIENDS

EXPLORING CLUSTERING WITH PHOTOGRAPHS

shaurya agarwal

SHAURYA AGARWAL
3shaurashaurya@gmail.com

Outline

- motivation
- movie
- math
- method
- more

SHAURYA AGARWAL
3shaurashaurya@gmail.com

motivation

what are we talking about again?

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

movie

cinematography: a rather unusual use case for k-means

SHAURYA AGARWAL
3shaurashaurya@gmail.com

i got to do this

cinematography

on

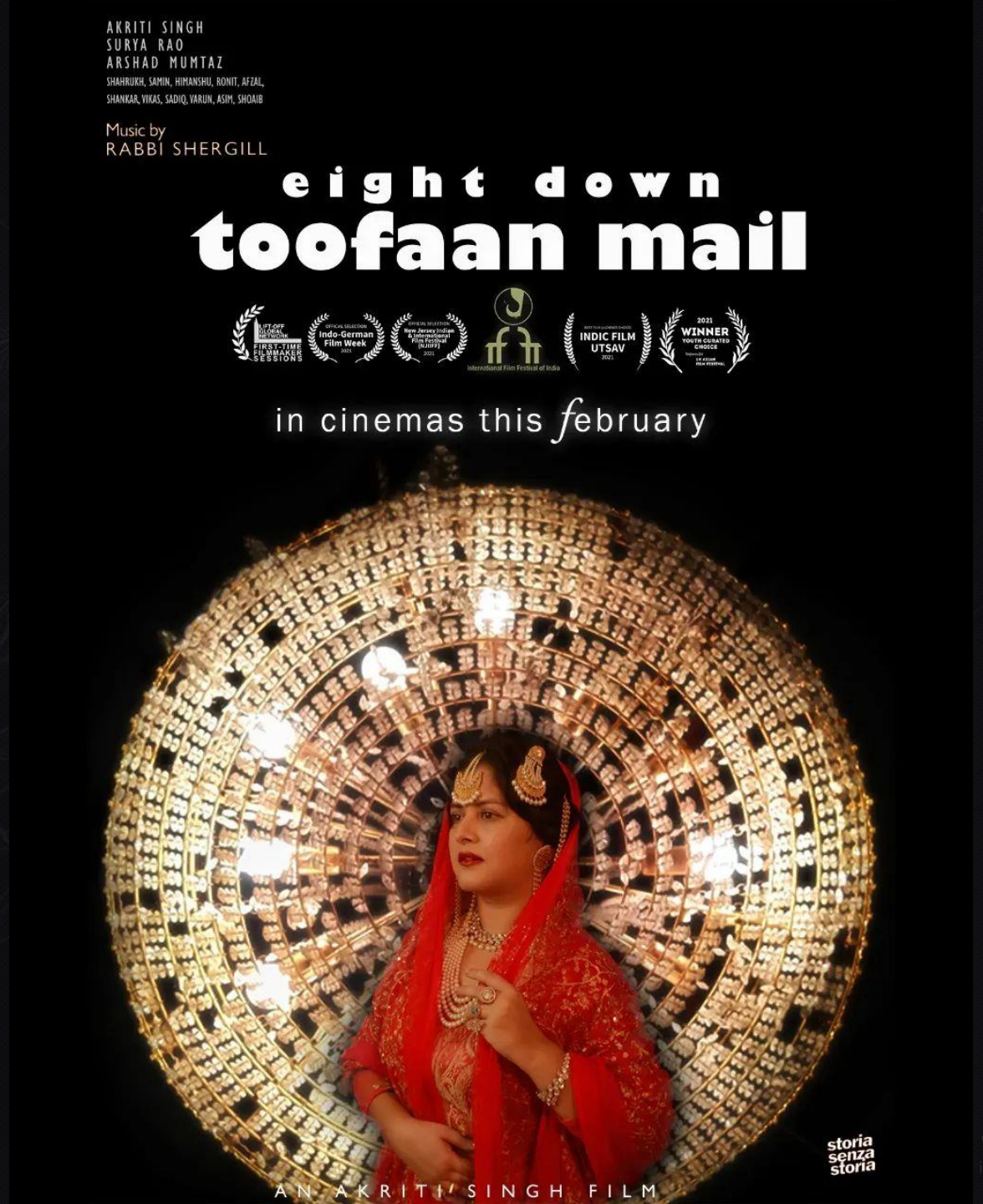
a

budget

powered

by

python



first principles

the best cinematography has exceptional control on color

SHAURYA AGARWAL
3shaurashaurya@gmail.com

first principles

showcase what inspired the film’s cinematography

SHAURYA AGARWAL
3shaurashaurya@gmail.com

first principles - colors

- one, or two or a few

SHAURYA AGARWAL
3shaurashaurya@gmail.com

first principles - colors

- one, or two or a few but no more

SHAURYA AGARWAL
3shaurashaurya@gmail.com

first principles - colors

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

first principles - colors

these colors must remain (approximately, perceptually) consistent

throughout the process, across all devices and screens

SHAURYA AGARWAL
3shaurashaurya@gmail.com

pre-production

or else*

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

pre-production

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

pre-production

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

pre-production

but for a 10-day shoot, interior and exterior

SHAURYA AGARWAL
3shaurashaurya@gmail.com

pre-production

we had USD 359 total

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

key insight

light – waves – wavelength, wavelength, wavelength

SHAURYA AGARWAL
3shaurashaurya@gmail.com

key insight

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

how do you pick the colors?

I used k-means++ to do it

we'll cover K-means first, the ++ is a way to better select the initial conditions, you'll see

SHAURYA AGARWAL
3shaurashaurya@gmail.com

let’s see how

there really is no math to it...

SHAURYA AGARWAL
3shaurashaurya@gmail.com

math*

how clustering works

SHAURYA AGARWAL
3shaurashaurya@gmail.com

naïve explainer

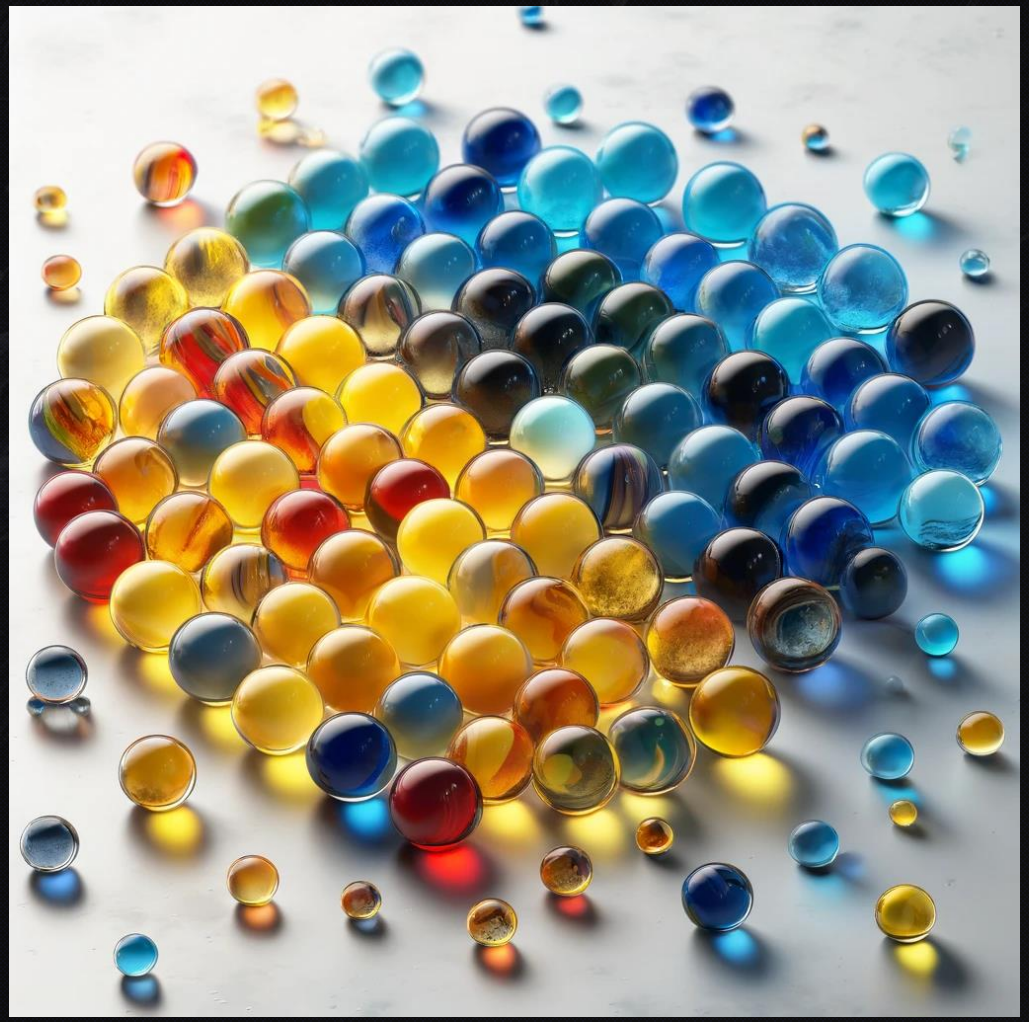
imagine you have a bunch of marbles of different colors, and you want to organize them into groups

SHAURYA AGARWAL
3shaurashaurya@gmail.com

naïve explainer

the k-means method is like deciding to group the marbles

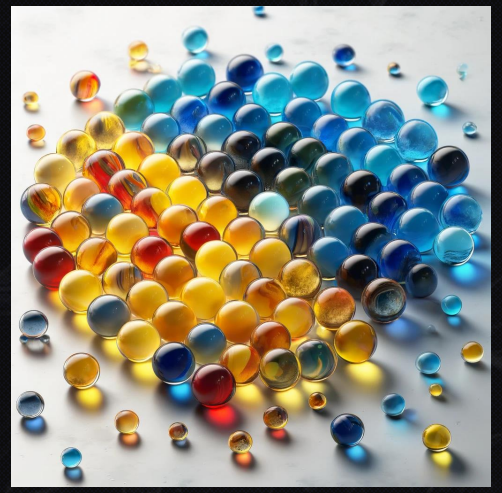
based on how close they are to each other (in color)

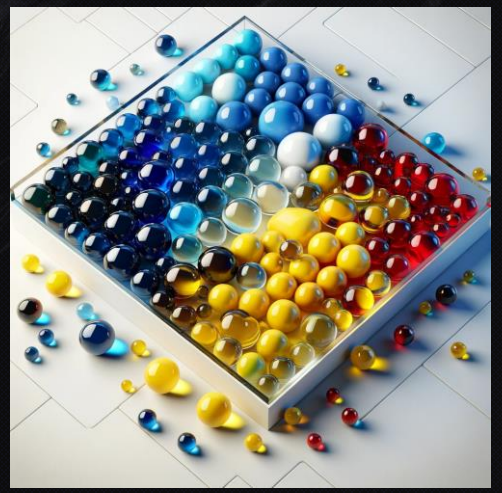


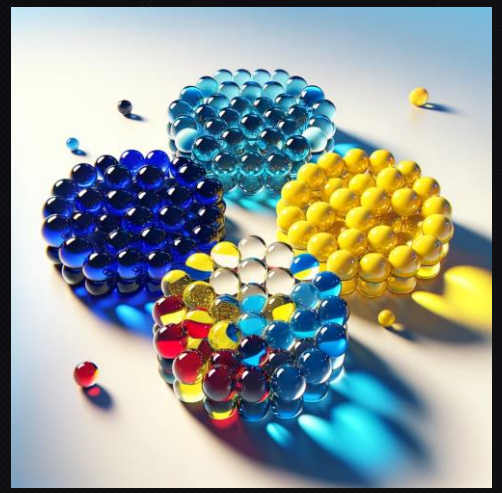
SHAURYA AGARWAL
3shaurashaurya@gmail.com

naïve explainer

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







SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Initialize

Assign

Update

Repeat

SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Initialize

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Initialize

Assign

Update

Repeat

Assign

each data point to the **closest** centroid,

creating clusters.

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Initialize

Assign

Update

Repeat

Update

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Repeat

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

Initialize

Assign

Update

Repeat

SHAURYA AGARWAL
3shaurashaurya@gmail.com

steps

Initialize

Assign

Update

Repeat

Limit

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

framework

Representation

Distance Measure

K

SHAURYA AGARWAL
3shaurashaurya@gmail.com

framework

Representation

feature comes in focus

unit distances mean the same thing everywhere.

Representation

Distance Measure

K

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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.

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com

how?

let’s see this in action

SHAURYA AGARWAL
3shaurashaurya@gmail.com

method

let’s look at some code

SHAURYA AGARWAL
3shaurashaurya@gmail.com

tech stack

this is compute intensive

we leverage vectorized, just-in-time approach

SHAURYA AGARWAL
3shaurashaurya@gmail.com

tech stack

JAX parallelizes and optimizes the compute DAG over NumPy

SHAURYA AGARWAL
3shaurashaurya@gmail.com

code

switch to code

SHAURYA AGARWAL
3shaurashaurya@gmail.com

MORE

JOIN THE DISCUSSION, GITHUB, MAKE THIS BETTER

SHAURYA AGARWAL
3shaurashaurya@gmail.com

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

SHAURYA AGARWAL
3shaurashaurya@gmail.com


SHAURYA AGARWAL
3shaurashaurya@gmail.com