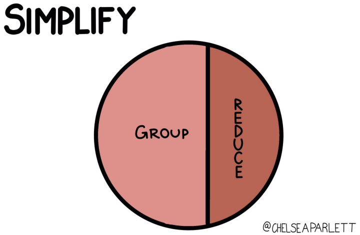
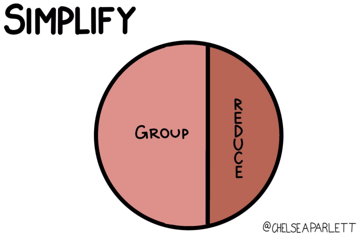


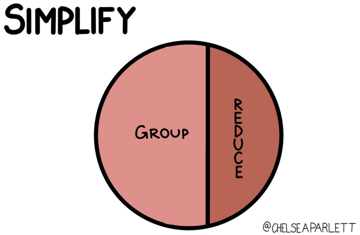
K-Means and Gaussian Mixtures

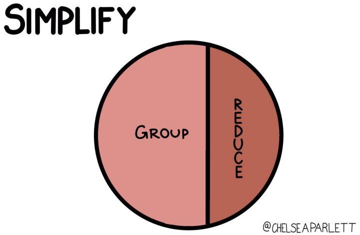
Dr. Chelsea Parlett-Pelleriti

Unsupervised Machine Learning



Clustering





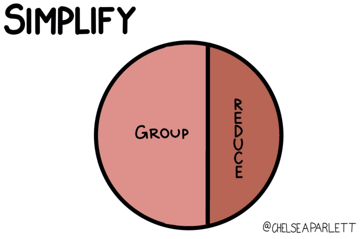
@CHELSEAPARLETT

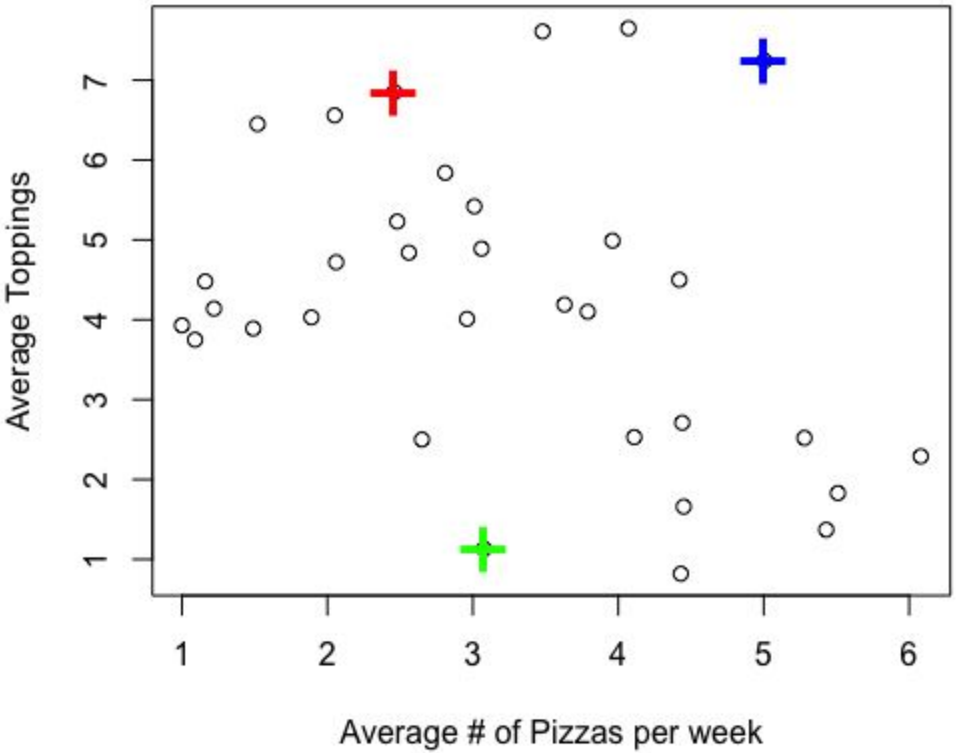
K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers
4. Repeat 2 and 3 until either:
    a. Cluster membership does not change
    b. Centers change only a tiny amount

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

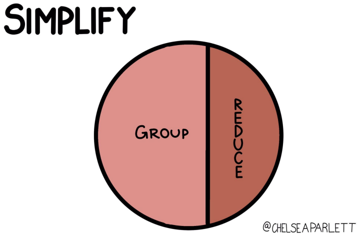


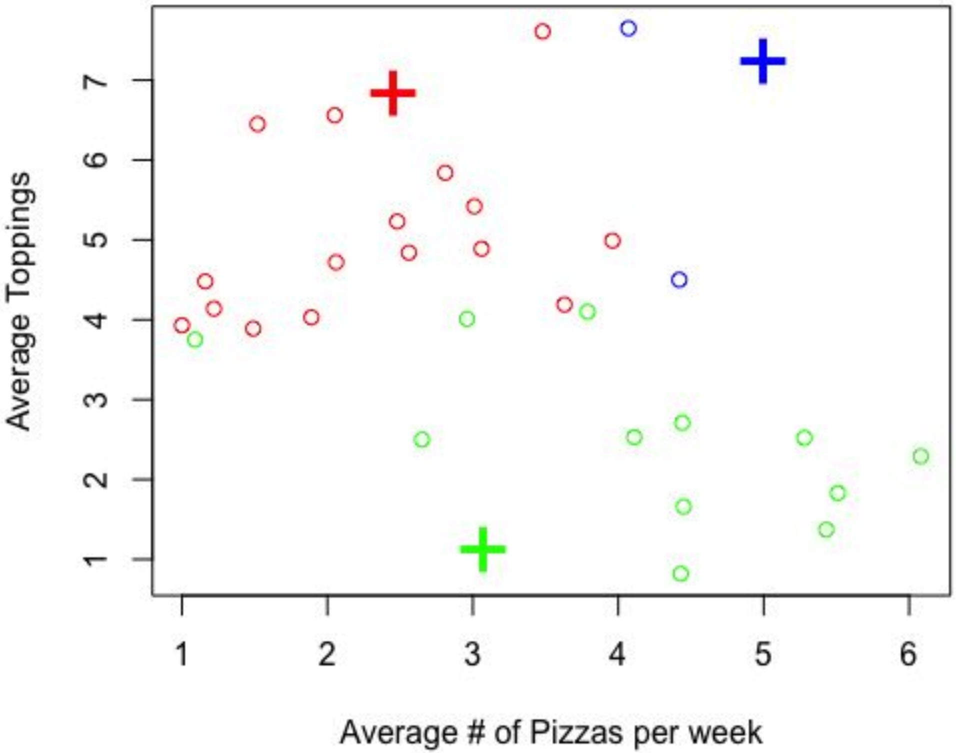


1

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

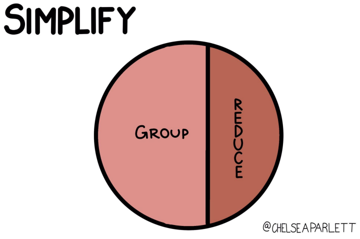




2

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

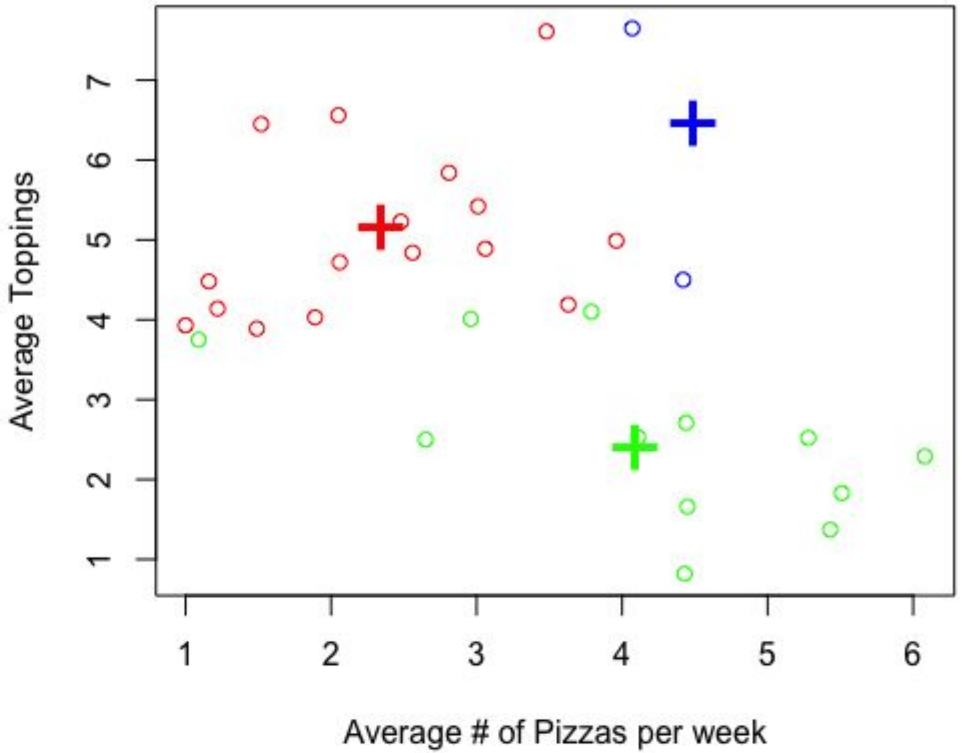


SIMPLIFY

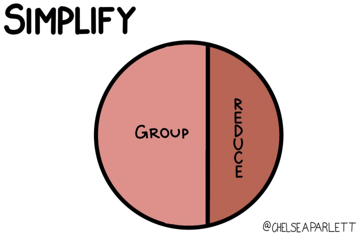
GROUP

REDUCE

@CHELSEAPARLETT

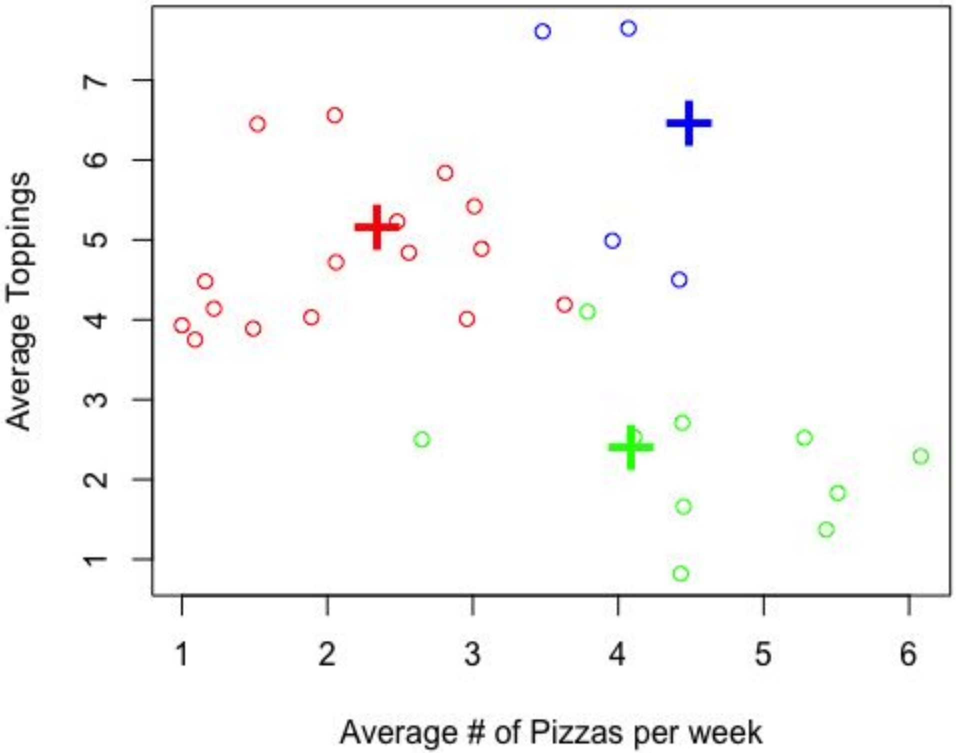


3



K-Means

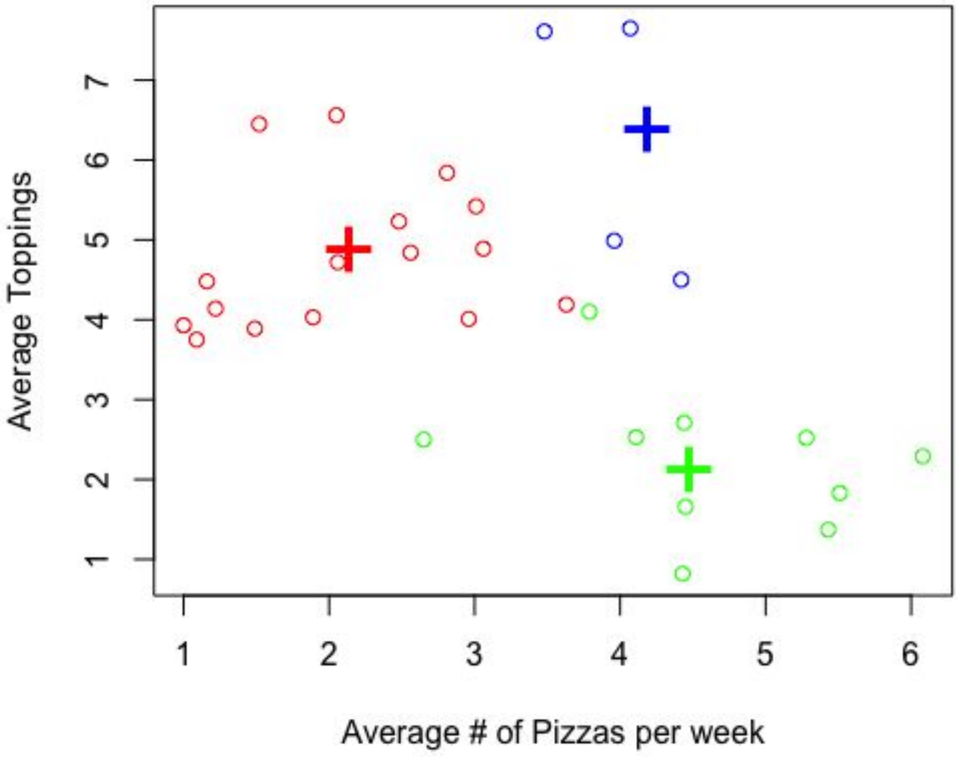
1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

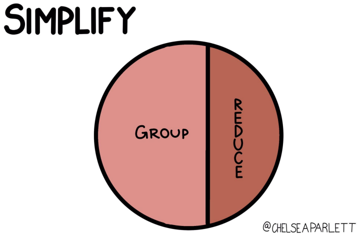


2

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

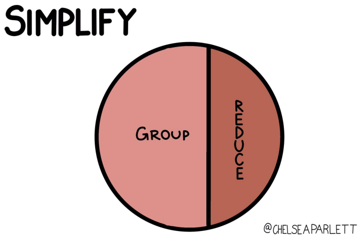


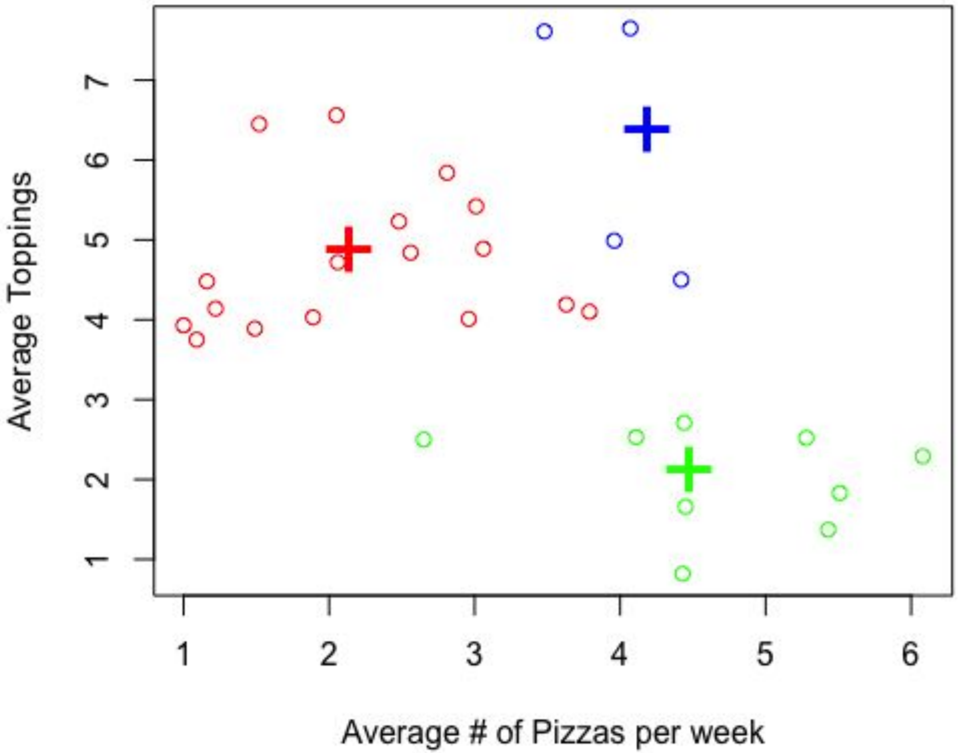


3

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

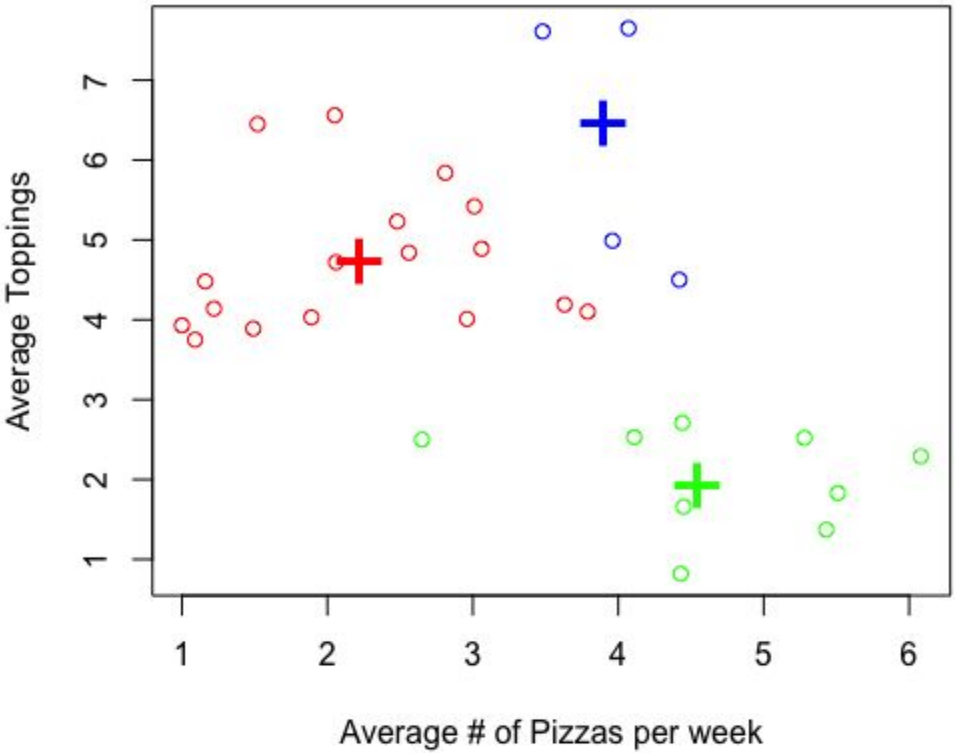


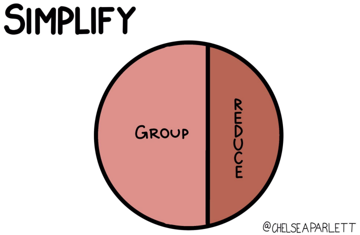


2

K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

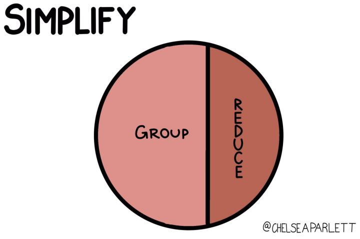


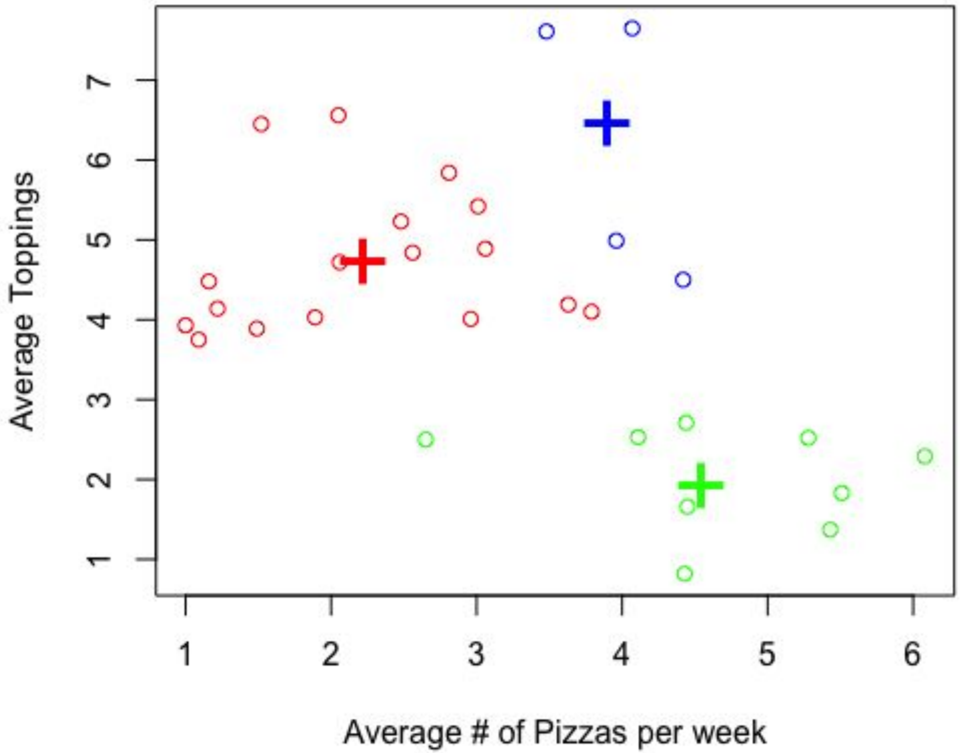


3

K-Means

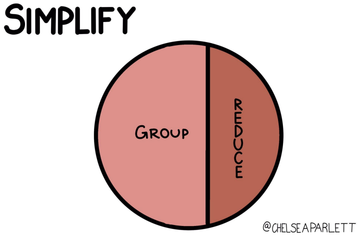
1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

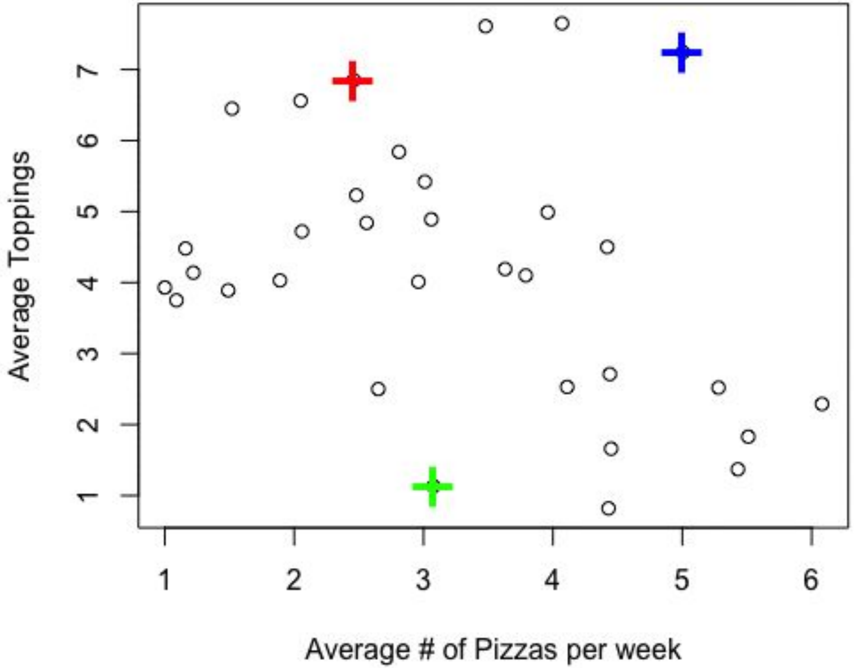




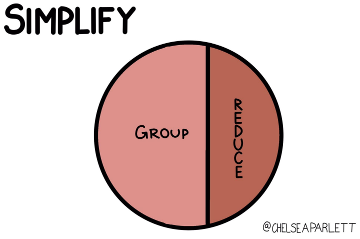
K-Means

1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers





K-Means

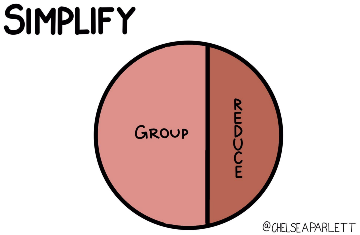


Assumptions

Spherical Clusters

Roughly the same # in each cluster

Evaluating Unsupervised Models



Cohesion:

Separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Applications

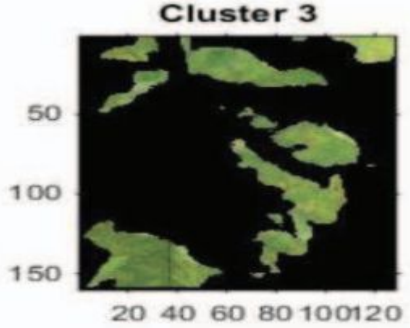


Fig. 8: Only Leaf

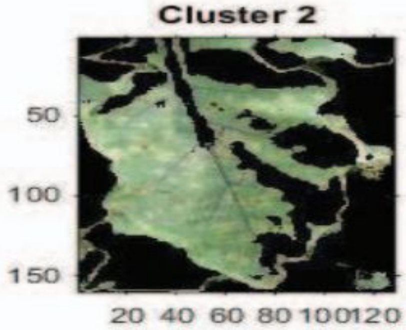


Fig. 7: Both Brinjal and Leaf

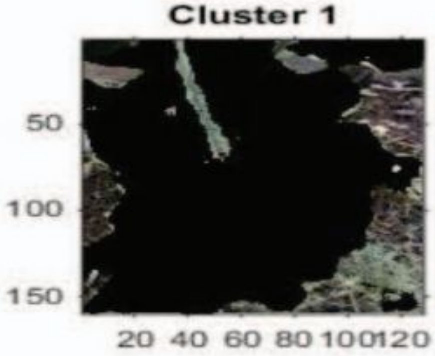


Fig. 6: Only Brinjal

https://ieeexplore.ieee.org/document/7569531

Applications

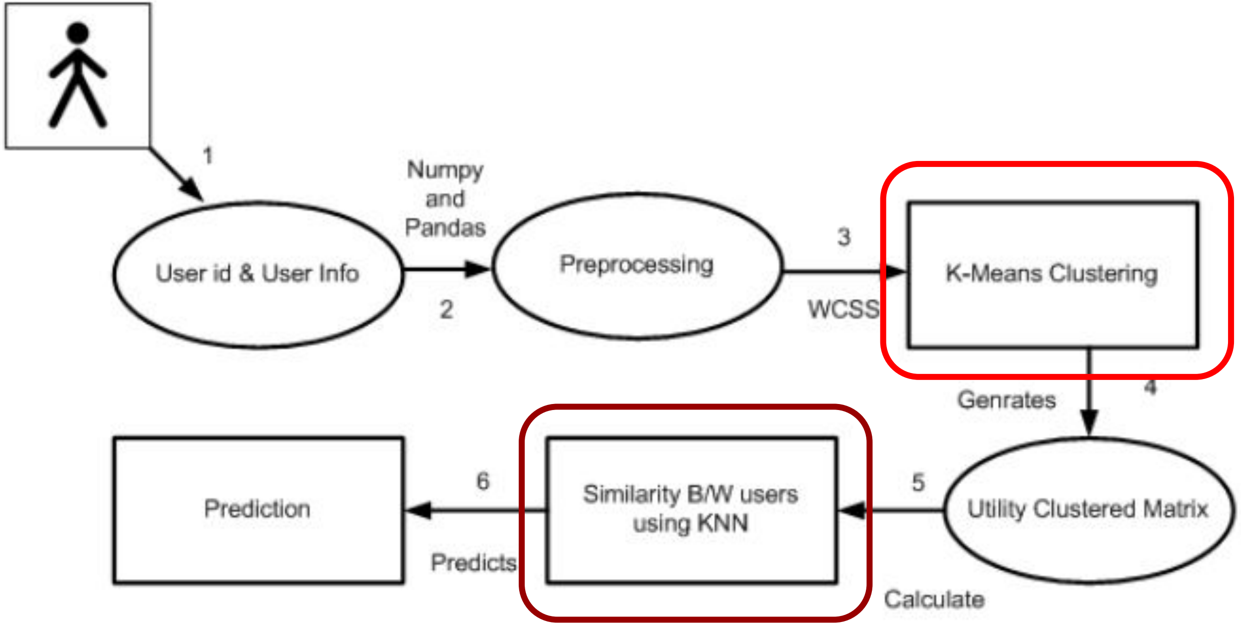
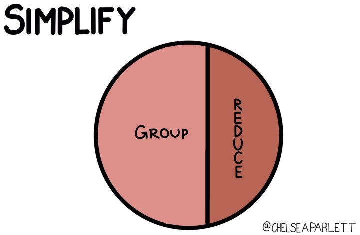


Fig. 2: Process Flow Diagram

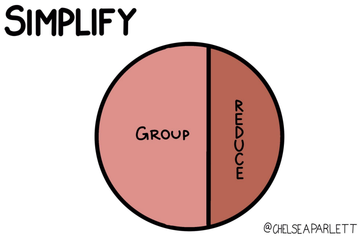
https://www.researchgate.net/profile/Arun-Solanki/publication/334763301_Movie_Recommender_System_Using_K-Means_Clustering_AND_K-Nearest_Neighbor/links/5ed727a7299bf1c67d34e356/Movie-Recommender-System-Using-K-Means-Clustering-AND-K-Nearest-Neighbor.pdf

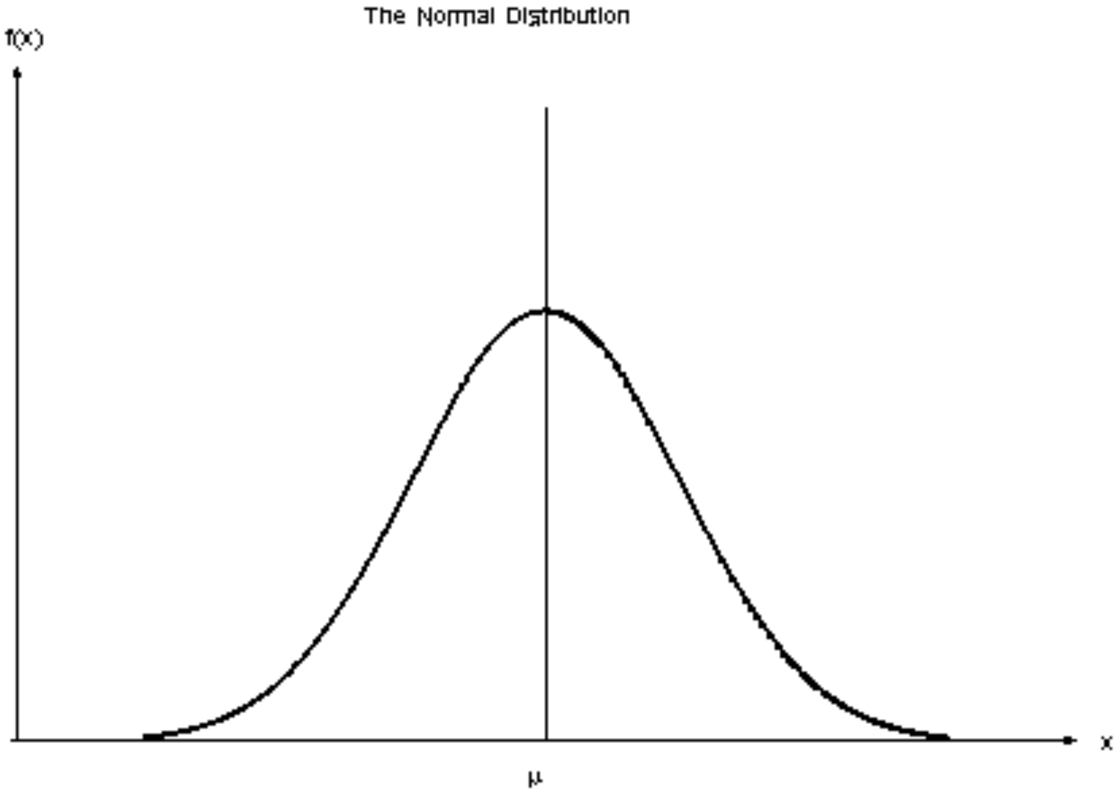


K-Means and Gaussian Mixtures

Dr. Chelsea Parlett-Pelleriti

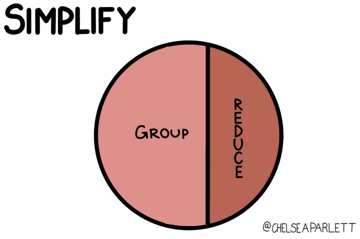
Normal (Gaussian) Distribution





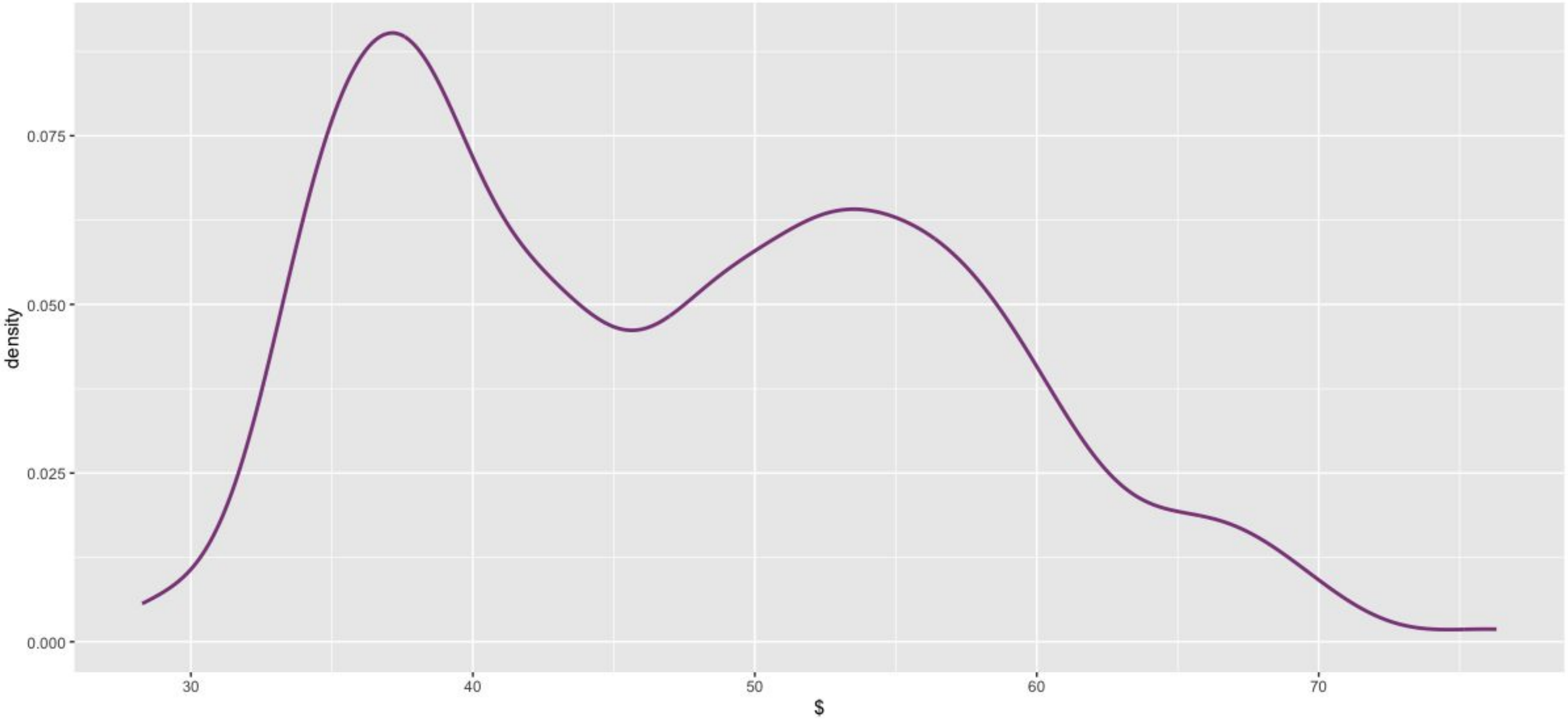
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu &= \text{ Mean} \\ \sigma &= \text{ Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$

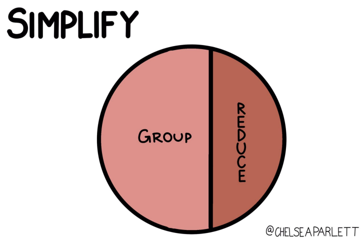


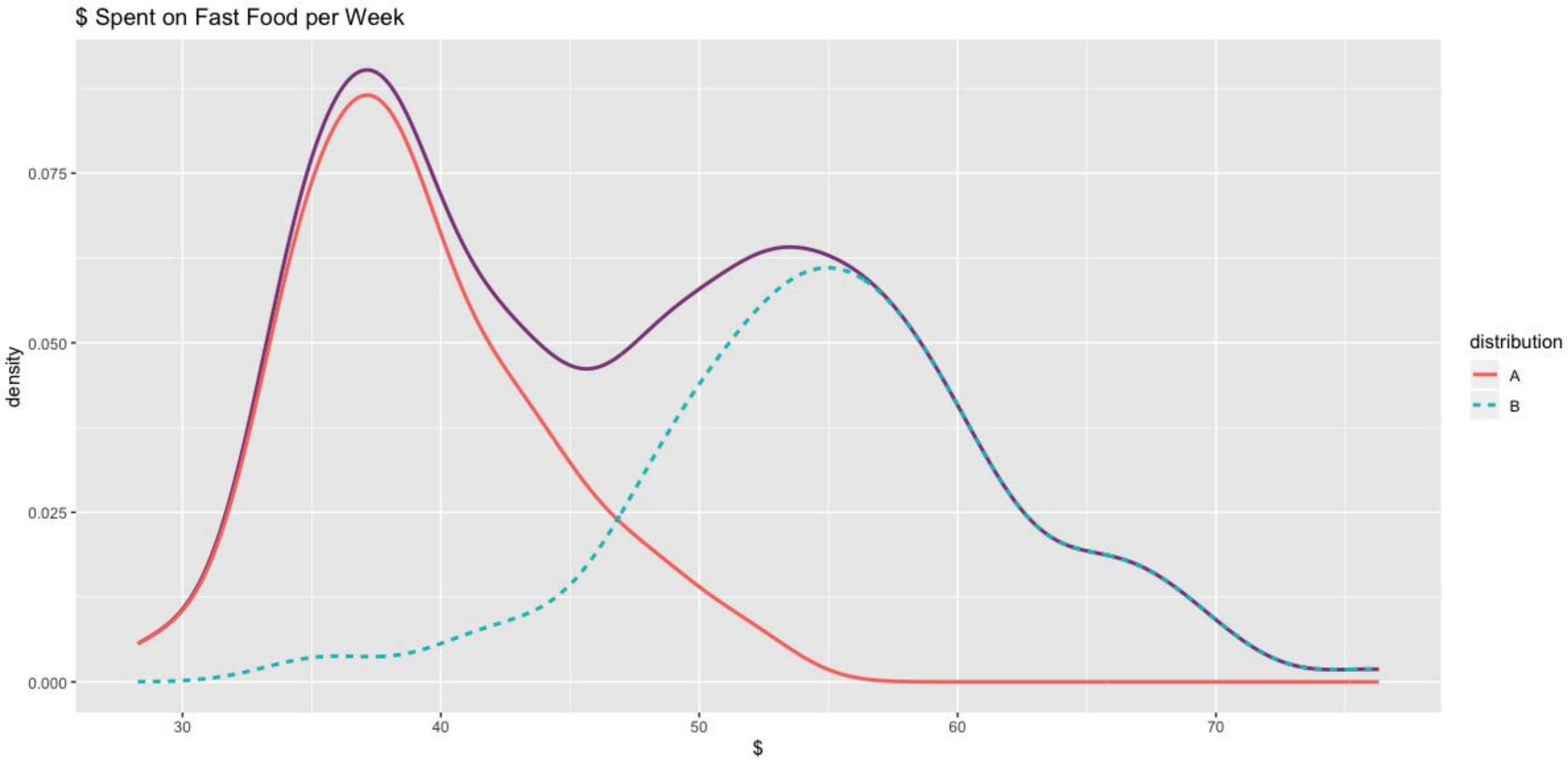
GMM

$ Spent on Fast Food per Week

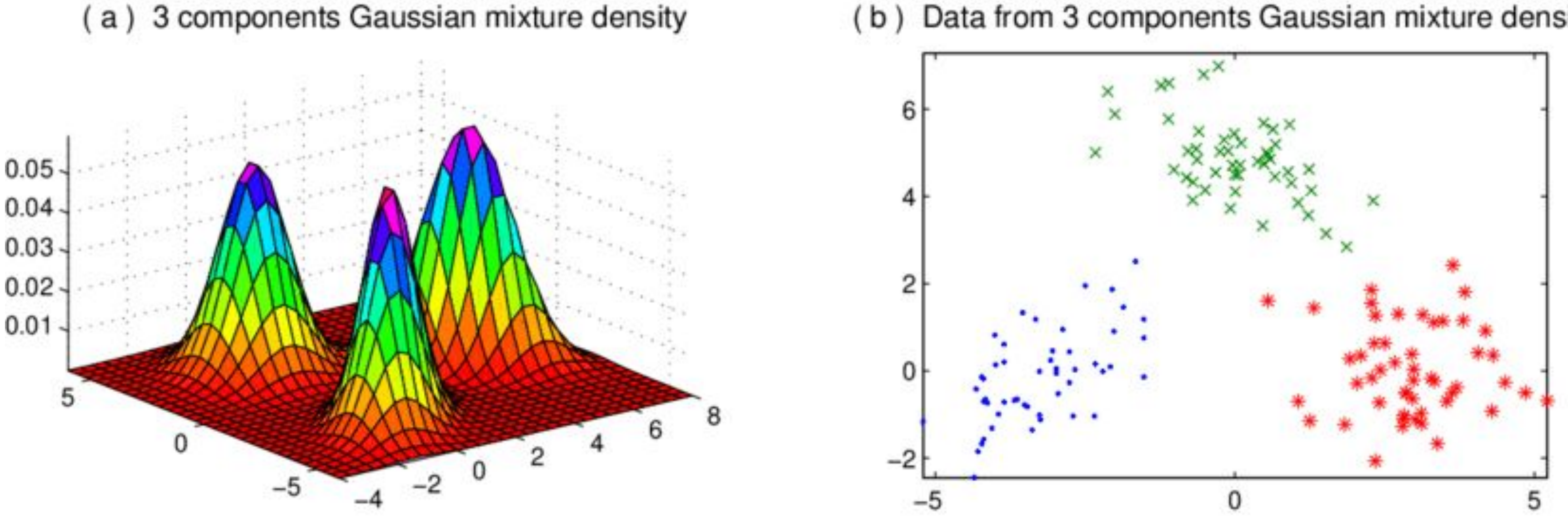


GMM


SIMPLIFY
GROUP
REDUCE
@CHELSEAPARLETT


$ Spent on Fast Food per Week
density
0.075
0.050
0.025
0.000
30
40
50
60
70
$
distribution
A
B

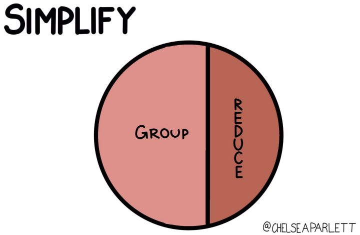
Multivariate Normal Distributions



(a) 3 components Gaussian mixture density

(b) Data from 3 components Gaussian mixture density

GMM



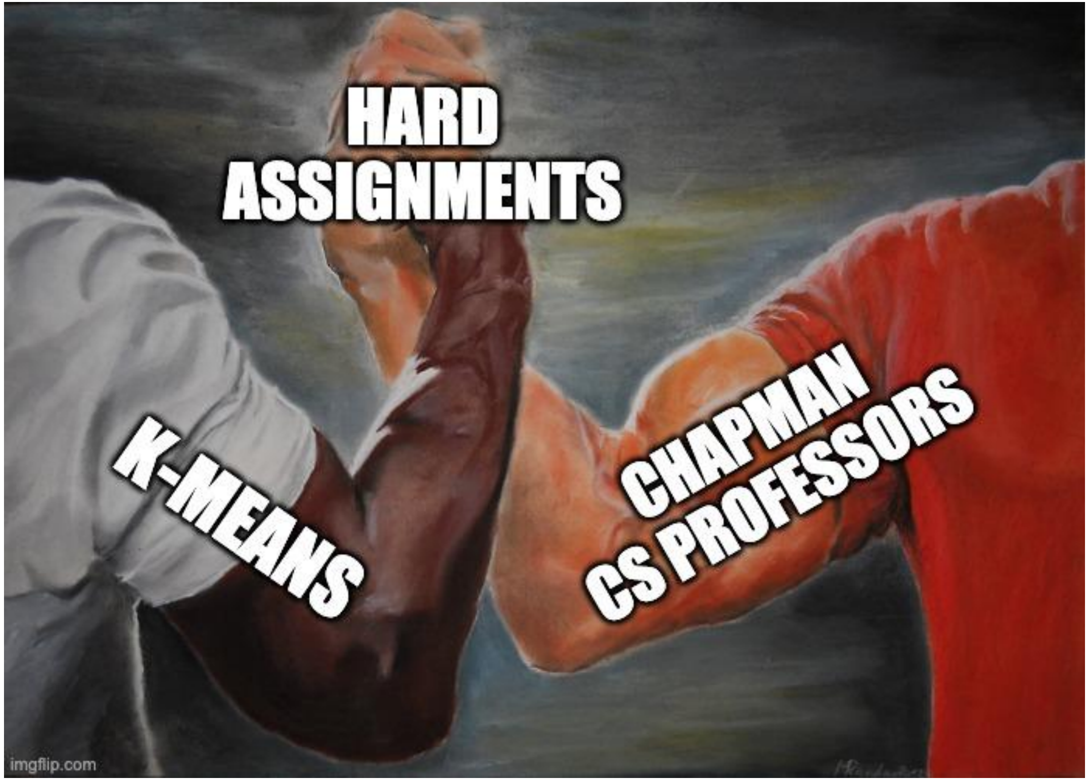
K means

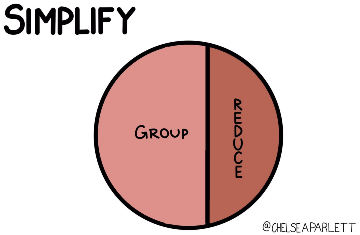
- Hard Assignment
- All Variances the Same

GMM

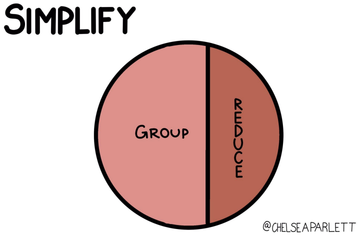
- Soft (probabilistic) Assignment
- Variances can be different

GMM


HARD ASSIGNMENTS
K-MEANS
CHAPMAN CS PROFESSORS
imgflip.com

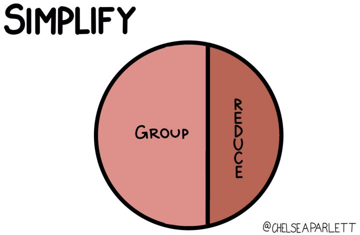

SIMPLIFY
GROUP
REDUCE
@CHELSEAPARLETT

K-Means Review



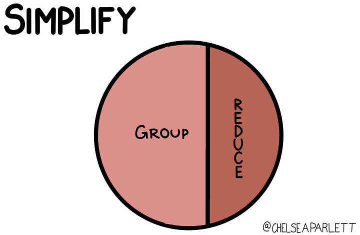
1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the **centers**
4. Repeat 2 and 3 until either:
   a. Cluster membership does not change
   b. Centers change only a tiny amount

GMM



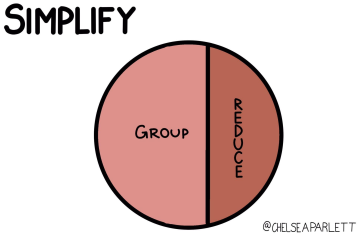
1. Choose **k** random points to be cluster centers (**or estimate using k-means...etc)**
2. For each data point, calculate the **probability** of belonging to each cluster
3. Using these probability weights, recalculate the **means + variances**
4. Repeat 2 and 3 until **distributions converge.**

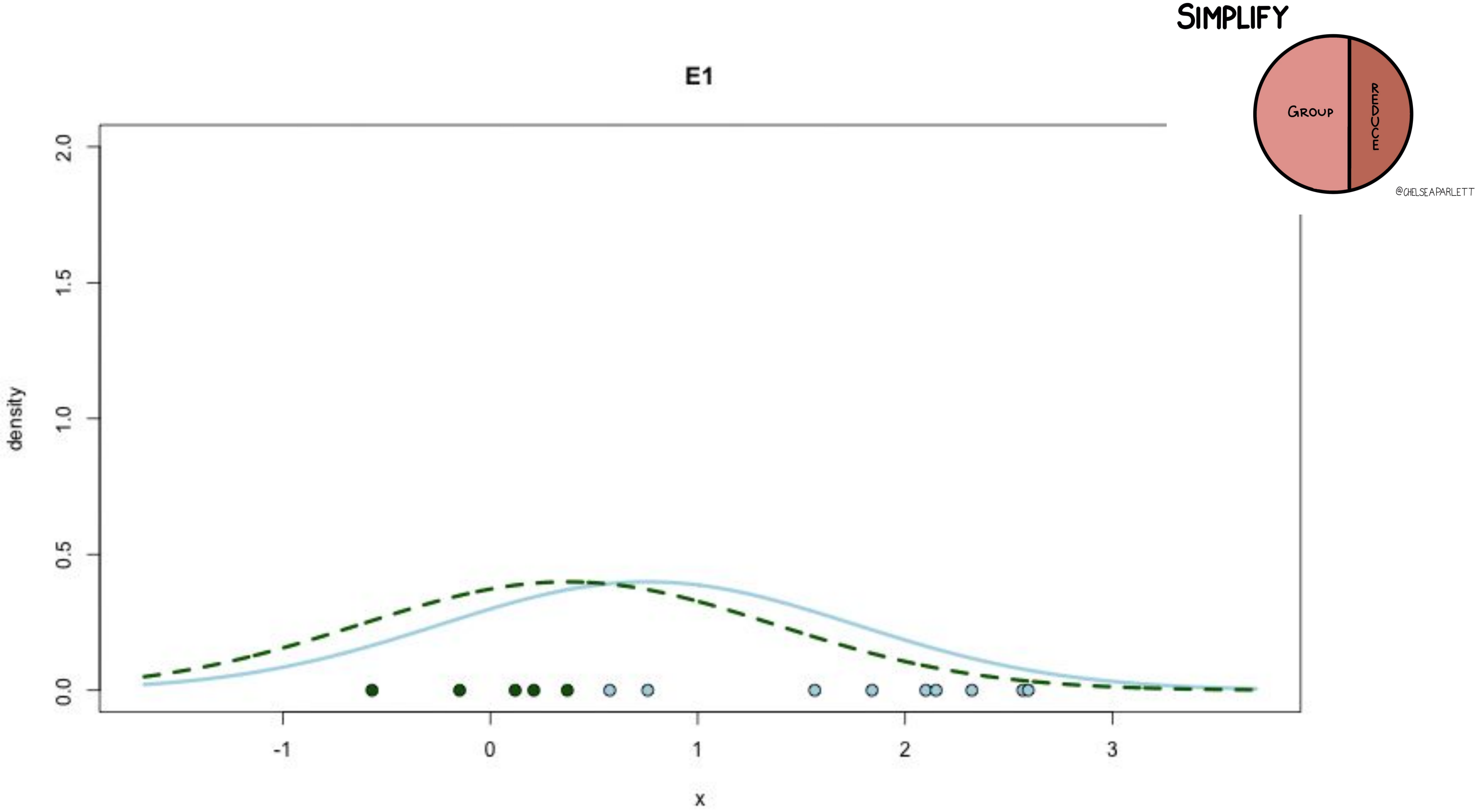
Formulas (E-Step)



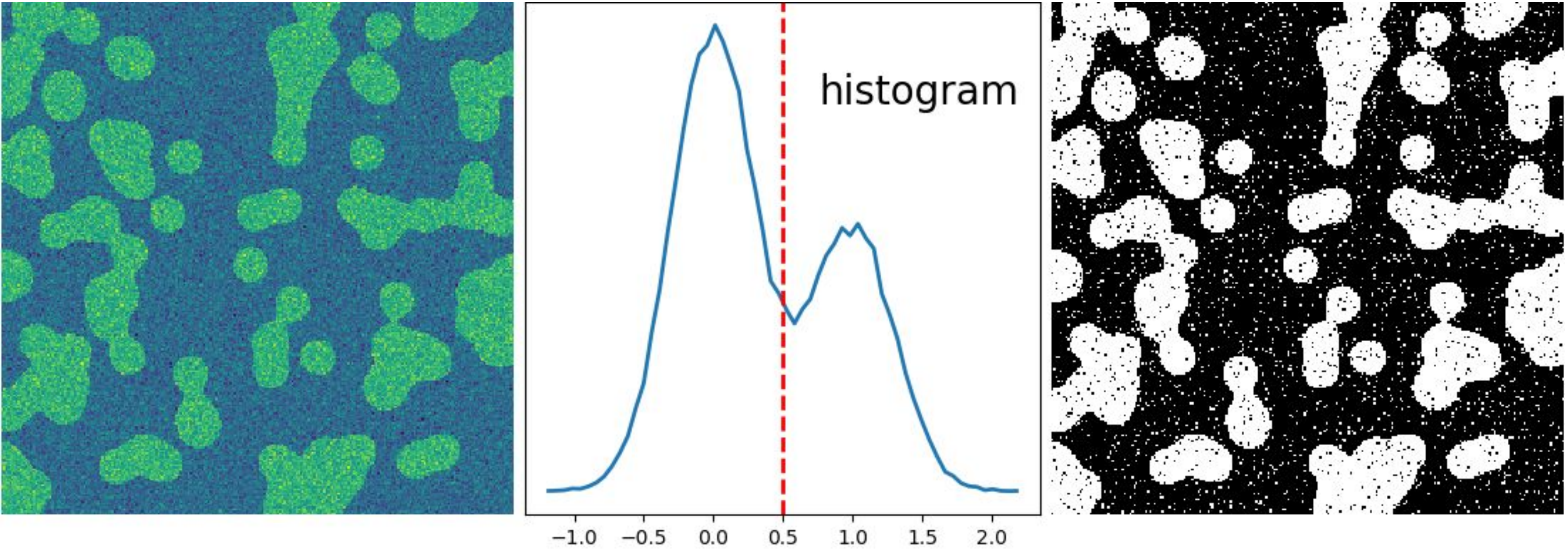
$$p_k(\underline{x}|\theta_k) \;=\; \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_k)^t \Sigma_k^{-1}(\underline{x}-\underline{\mu}_k)}$$

Formulas (M-Step)



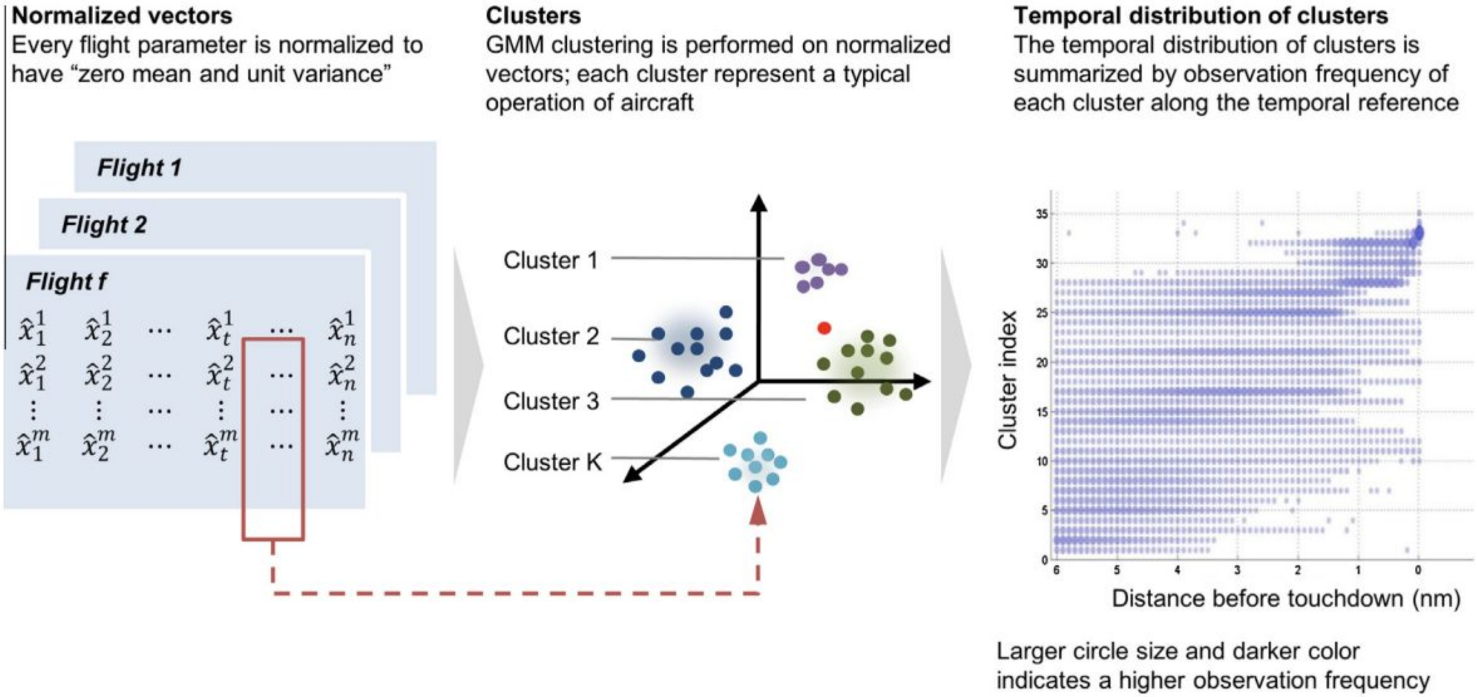

SIMPLIFY
GROUP
REDUCE
@CHELSEAPARLETT
E1
density
0.0
0.5
1.0
1.5
2.0
-1
0
1
2
3
x

Applications



https://scipy-lectures.org/advanced/image_processing/auto_examples/plot_GMM.html

Applications



Normalized vectors
Every flight parameter is normalized to have “zero mean and unit variance”

Clusters
GMM clustering is performed on normalized vectors; each cluster represent a typical operation of aircraft

Temporal distribution of clusters
The temporal distribution of clusters is summarized by observation frequency of each cluster along the temporal reference

Flight 1
Flight 2
Flight f
$\hat{x}_1^1$ $\hat{x}_2^1$ … $\hat{x}_t^1$ … $\hat{x}_n^1$
$\hat{x}_1^2$ $\hat{x}_2^2$ … $\hat{x}_t^2$ … $\hat{x}_n^2$
$\vdots$ $\vdots$ … $\vdots$ … $\vdots$
$\hat{x}_1^m$ $\hat{x}_2^m$ … $\hat{x}_t^m$ … $\hat{x}_n^m$
Cluster 1
Cluster 2
Cluster 3
Cluster K
Cluster index
35
30
25
20
15
10
5
0
6
5
4
3
2
1
0
Distance before touchdown (nm)

Larger circle size and darker color indicates a higher observation frequency

Fig. 3. Cluster analysis: identify typical operations and temporal distribution.

https://www.sciencedirect.com/science/article/abs/pii/S0968090X16000188

Applications

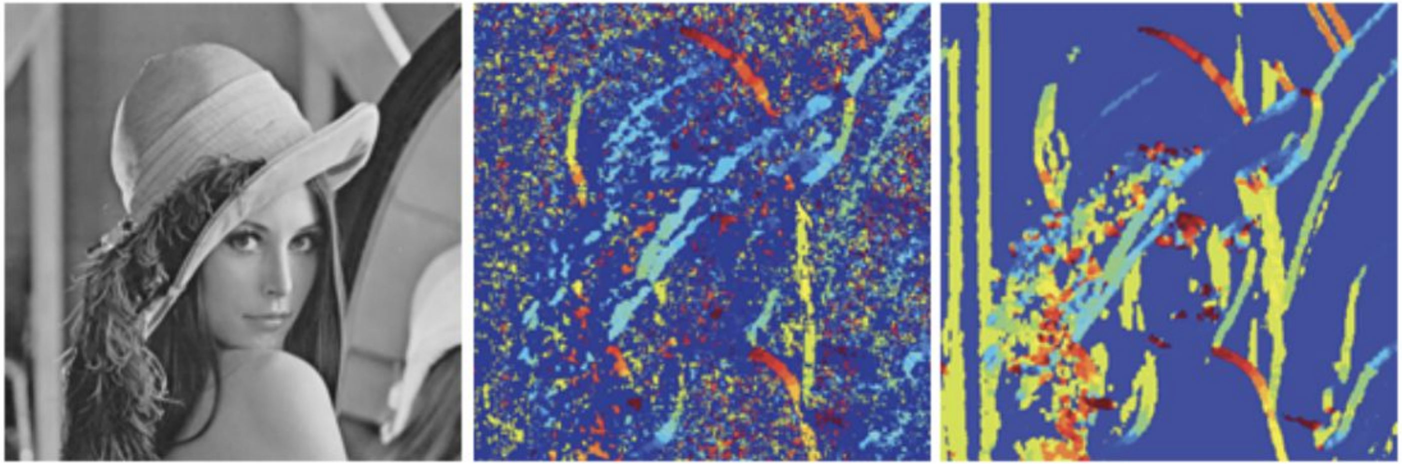
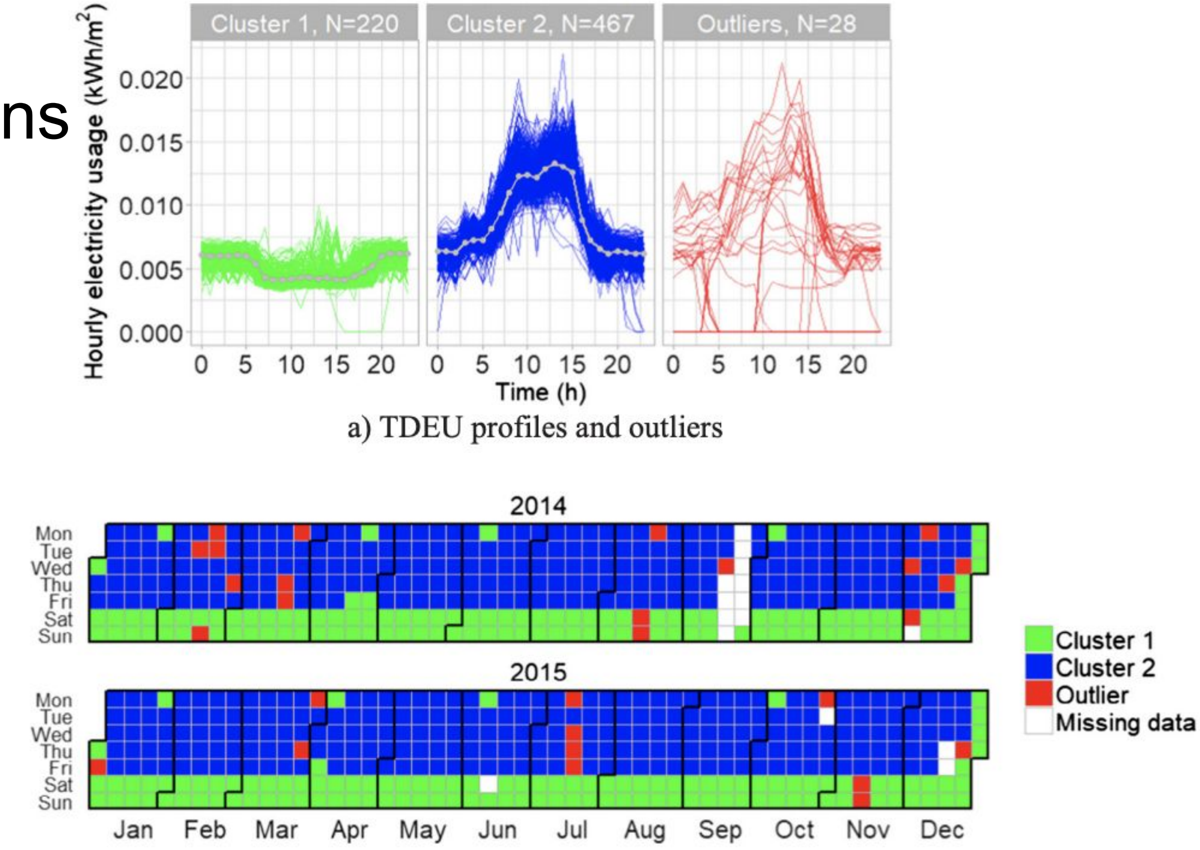


Fig. 1. Illustration of clustering of patches in the PLE method for the Lena image. LEFT: Original image; RIGHT: Clustered image; The pixels in the same color indicate that 8×8 patches around them are in the same cluster. It can be seen that patches from different parts of image are grouped into one cluster [17].

https://ieeexplore.ieee.org/document/7128671

Applications



a) TDEU profiles and outliers

b) Distribution of the TDEU profiles

Fig. 7. Visualisation of the intra-building clustering result of Building #16.

https://www.sciencedirect.com/science/article/pii/S0306261918313606?casa_token=ISbTRYcOzcYAAAAA:p2OBUhbFBenVK6aXGaVbJ4p2WRsmpa0y6bB8Uw3LSmjx8kPkhLyKVCx8gtOJ0cXDUFf07fNdDdHr