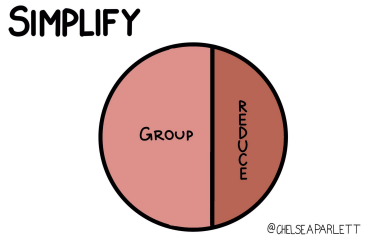
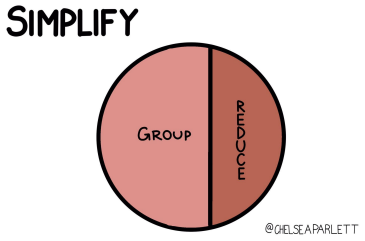


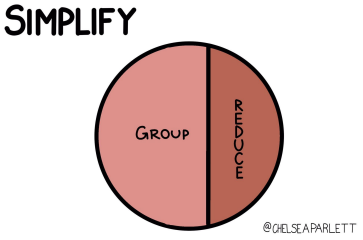
K-Means and Expectation Maximization

Chelsea Parlett-Pelleriti

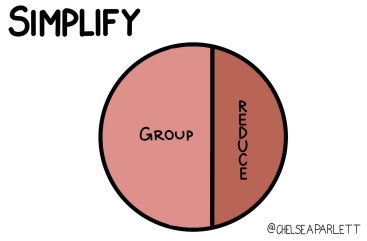
Unsupervised Machine Learning



Clustering



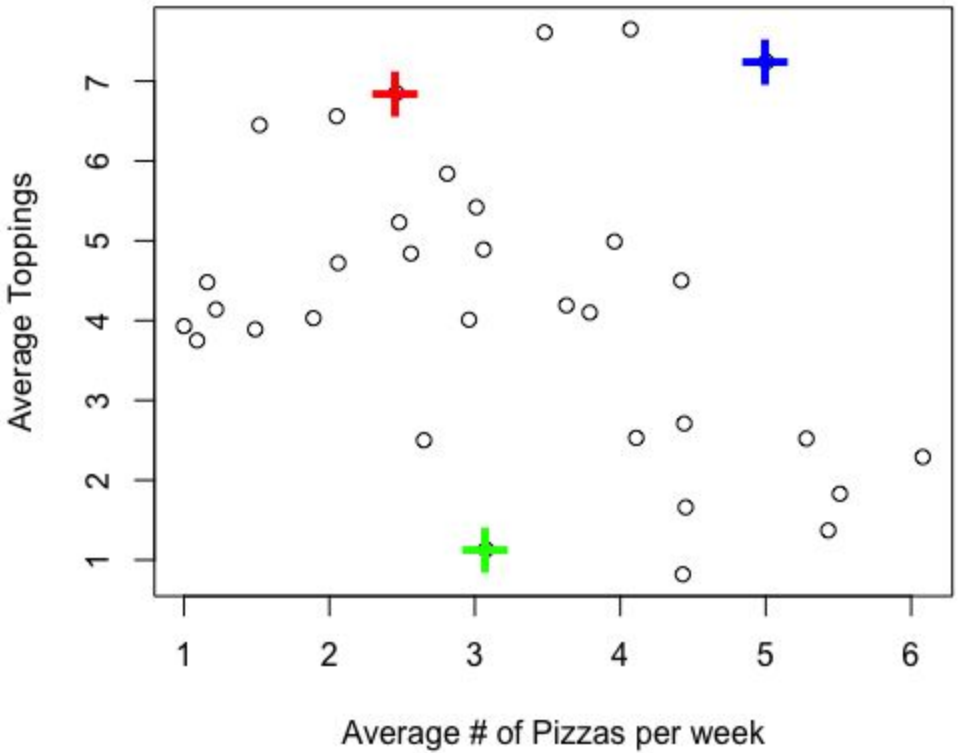
K-Means



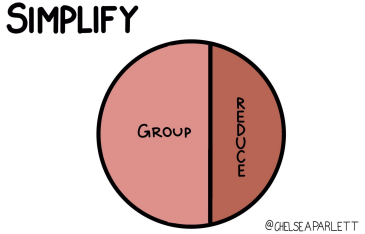
1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers
4. Repeat 2 and 3 until either:
    a. Cluster membership does not change
    b. Centers change only a tiny amount

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

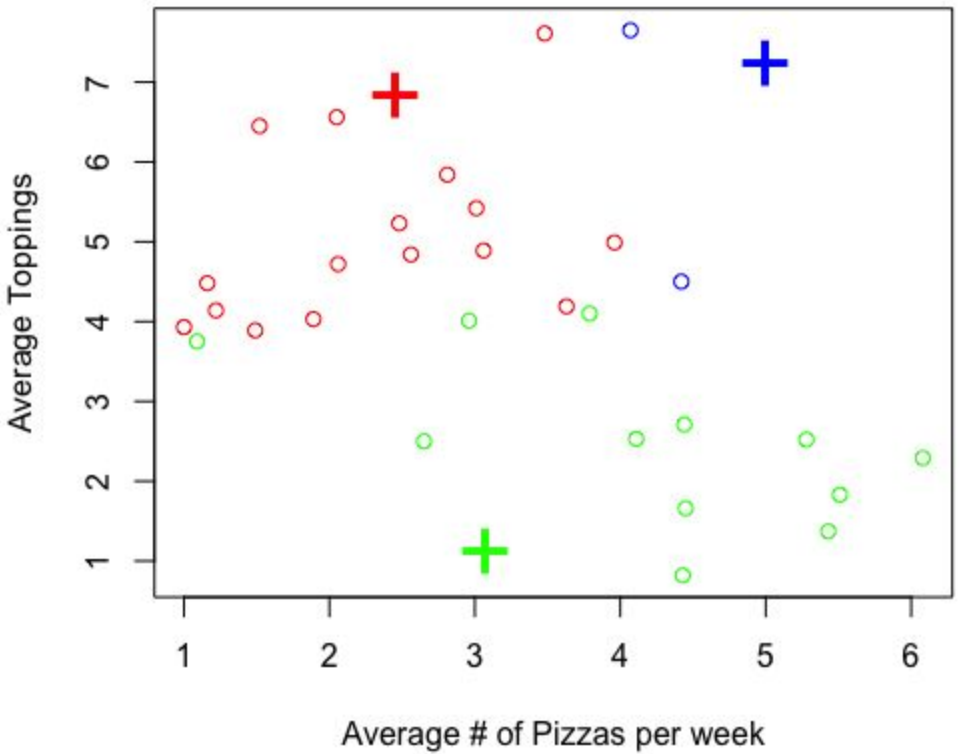


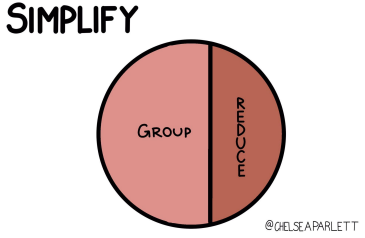
1



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

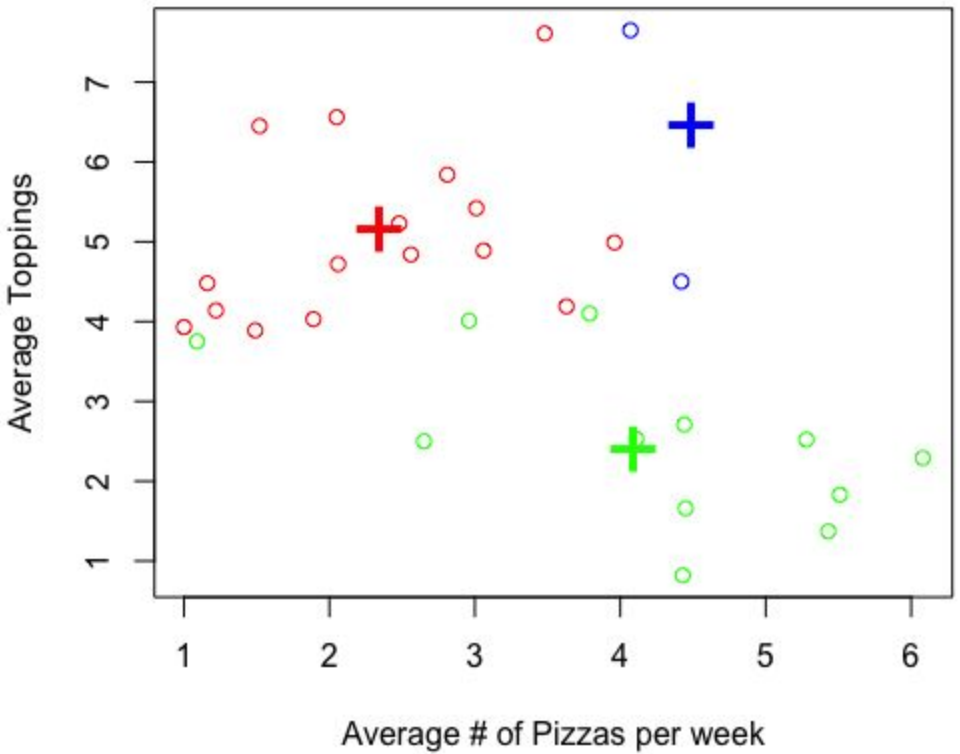


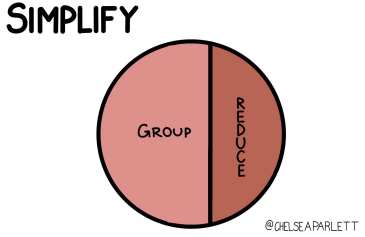


2

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

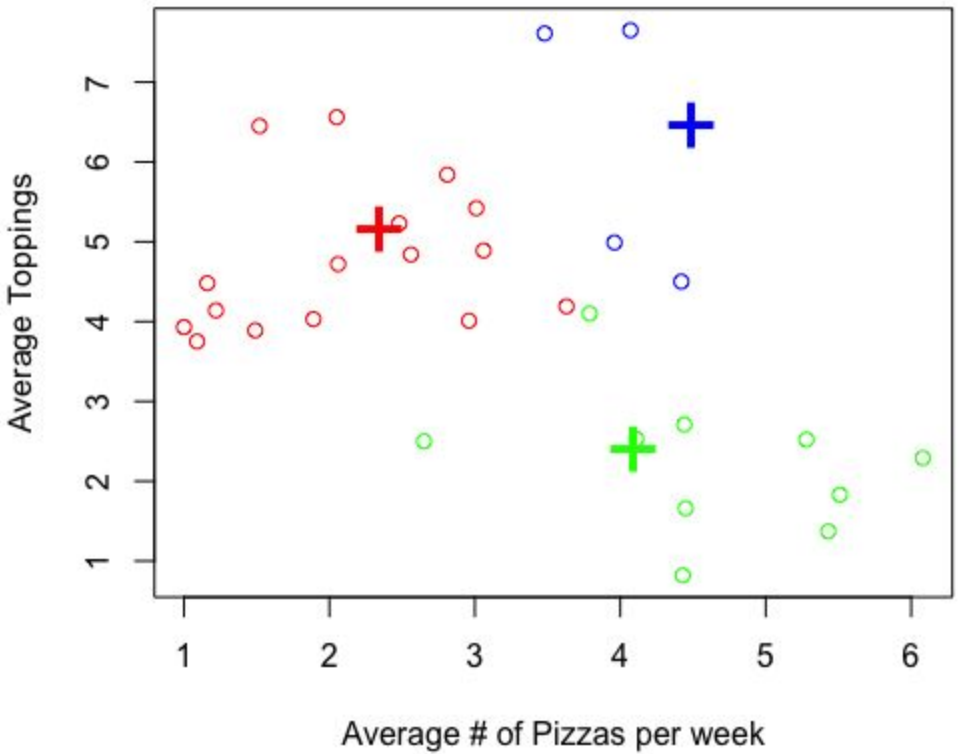


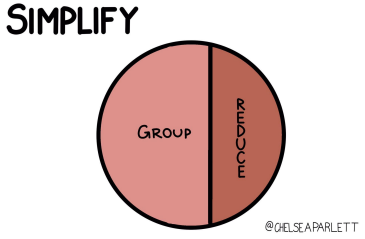


3

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

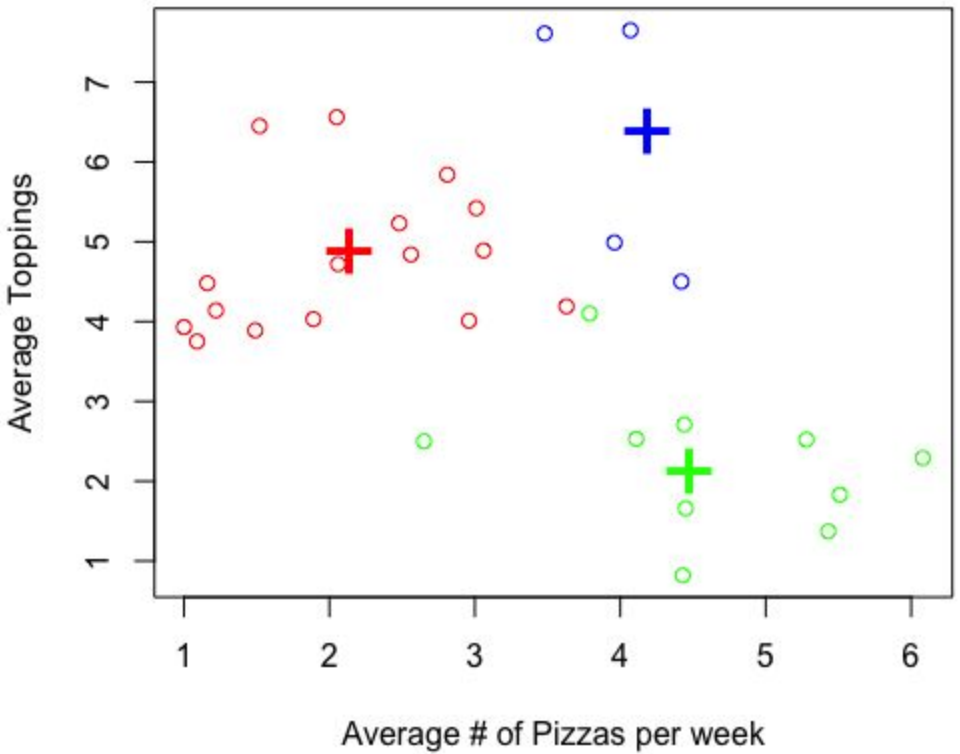


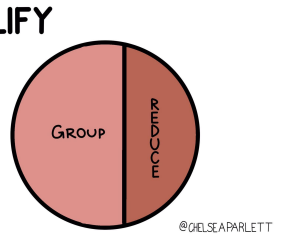


2

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

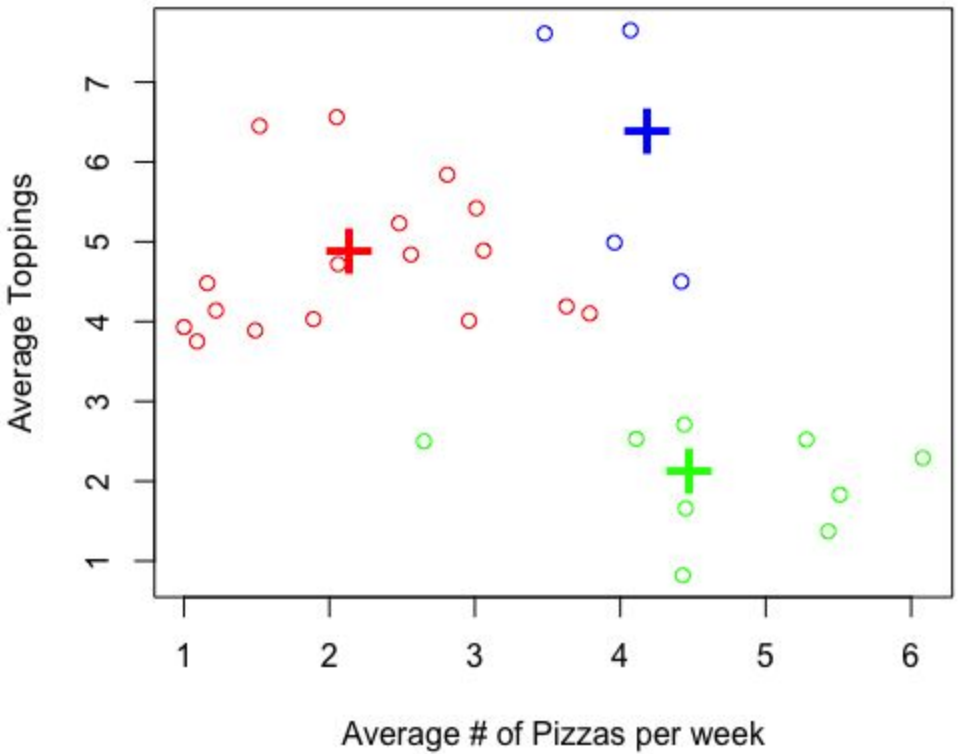


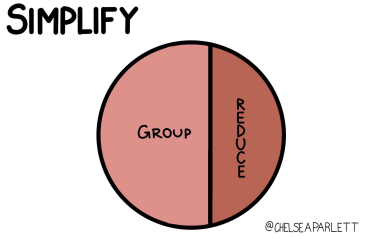


3

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

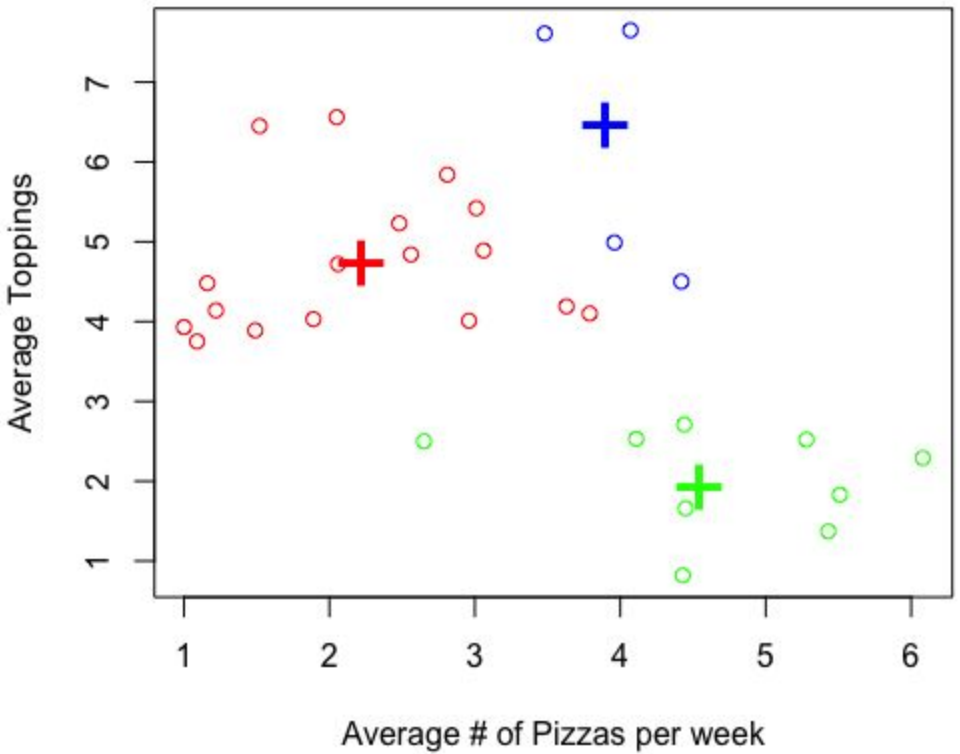


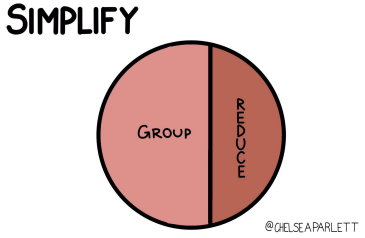


2

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

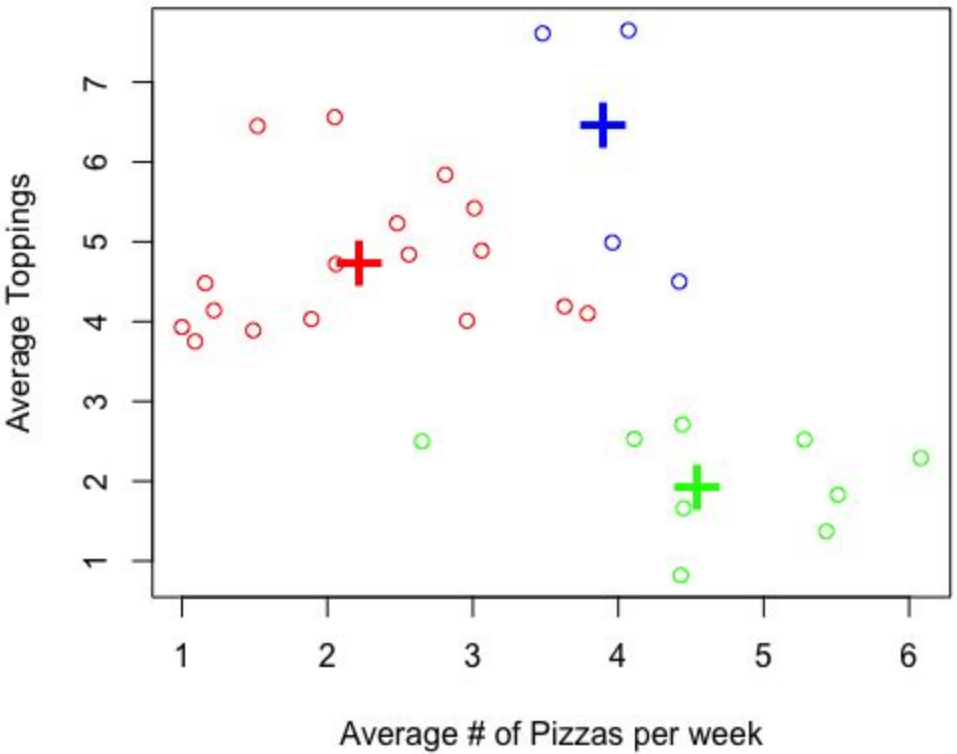


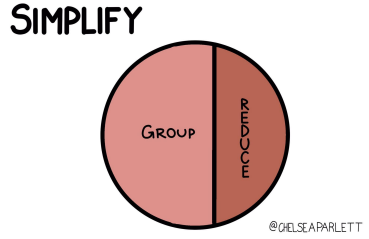


3

K-Means

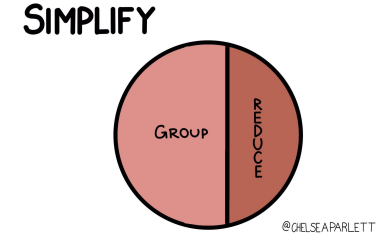
1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

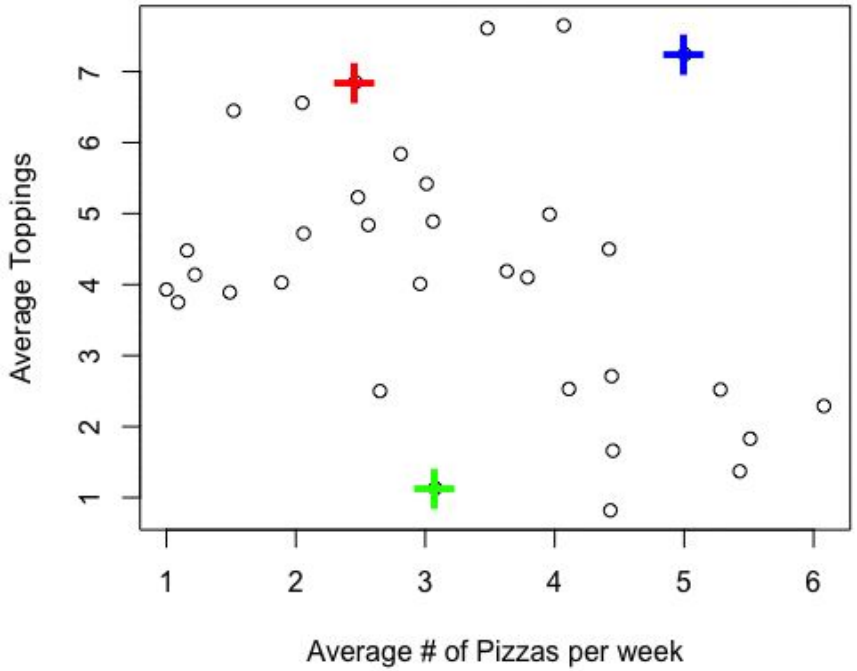




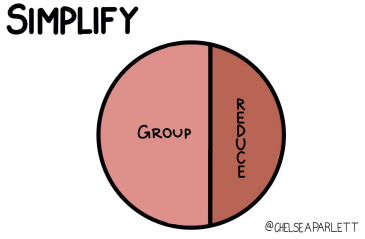
K-Means

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

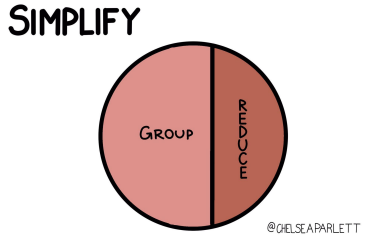


Assumptions

Spherical Clusters

Roughly the same # in each cluster

Evaluating Unsupervised Models

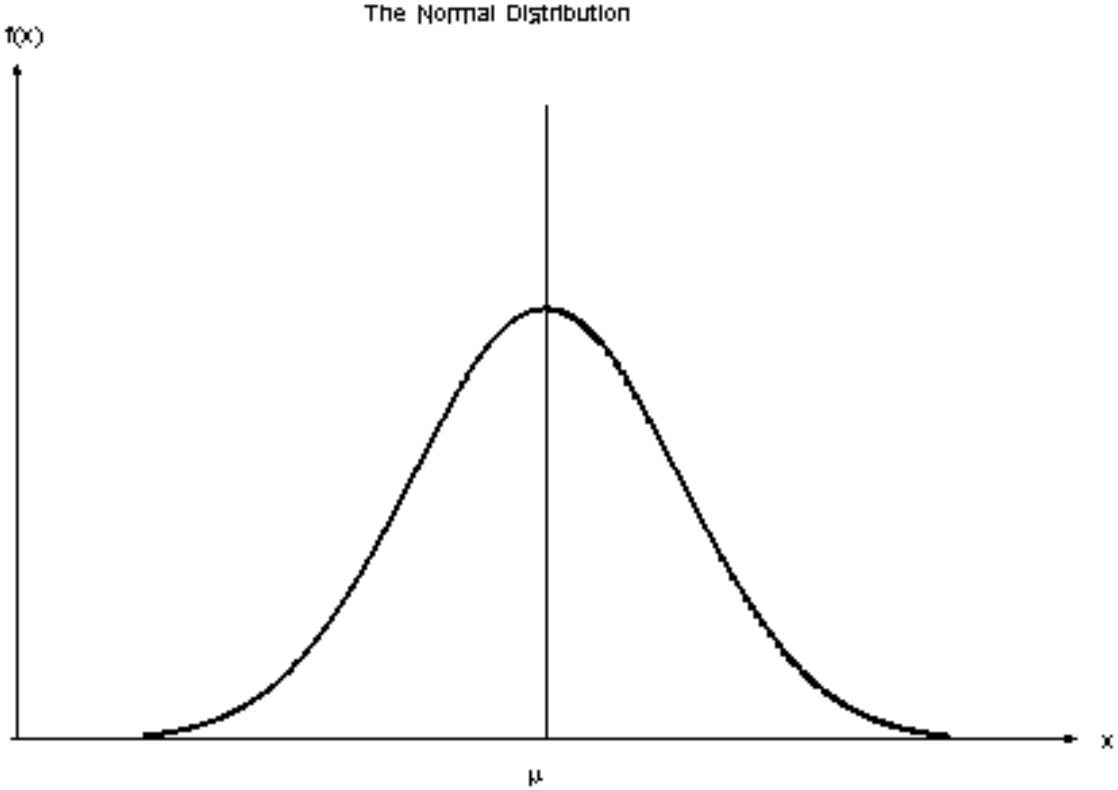


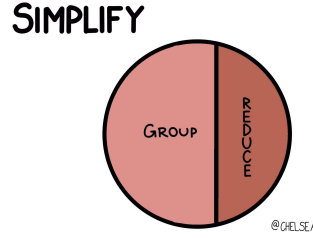
Cohesion:

Separation:

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

Normal (Gaussian) Distribution

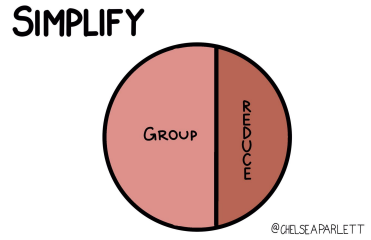


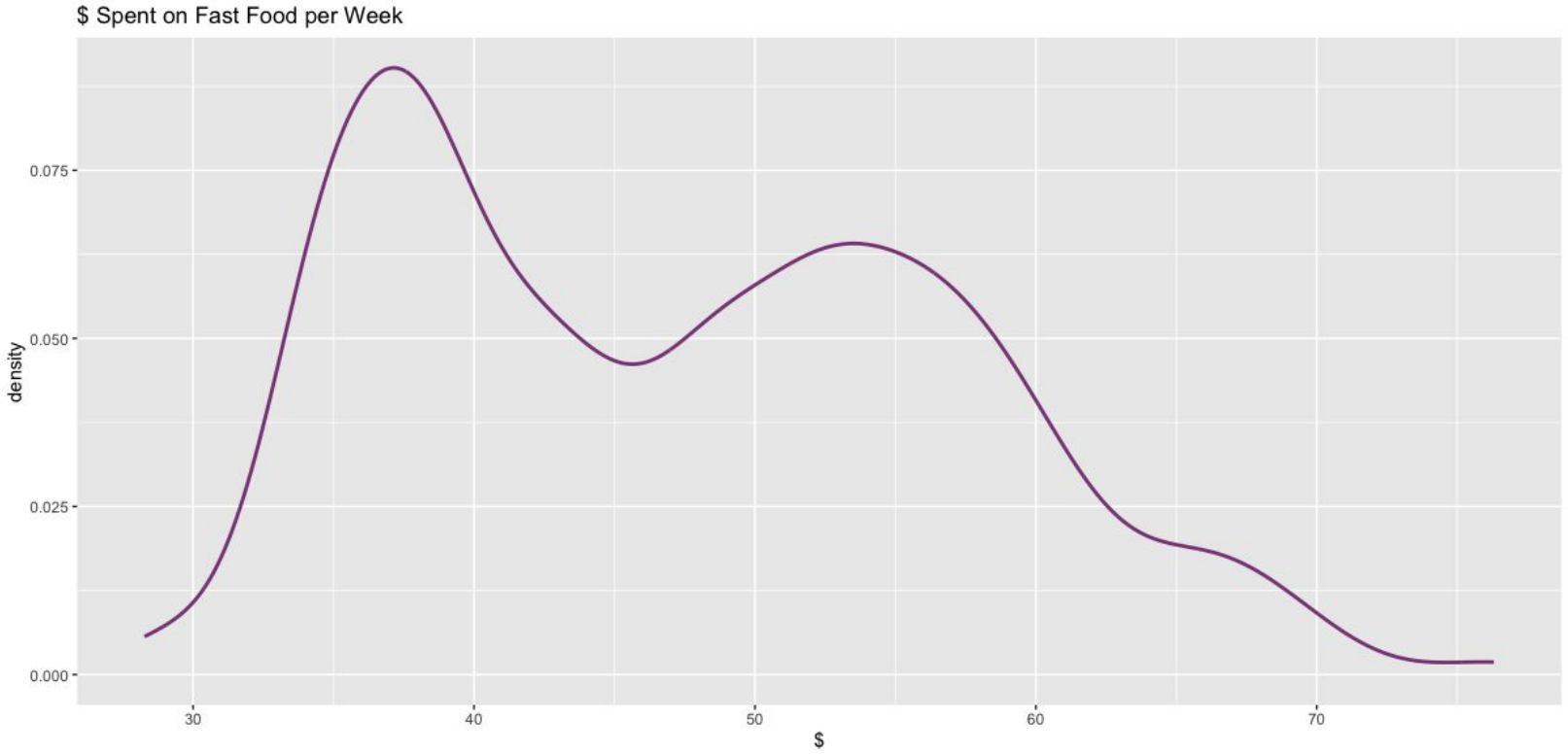


$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

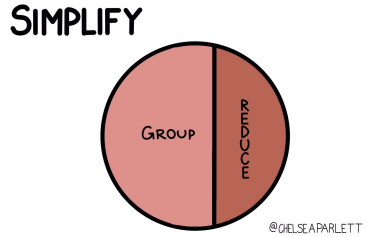
$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$

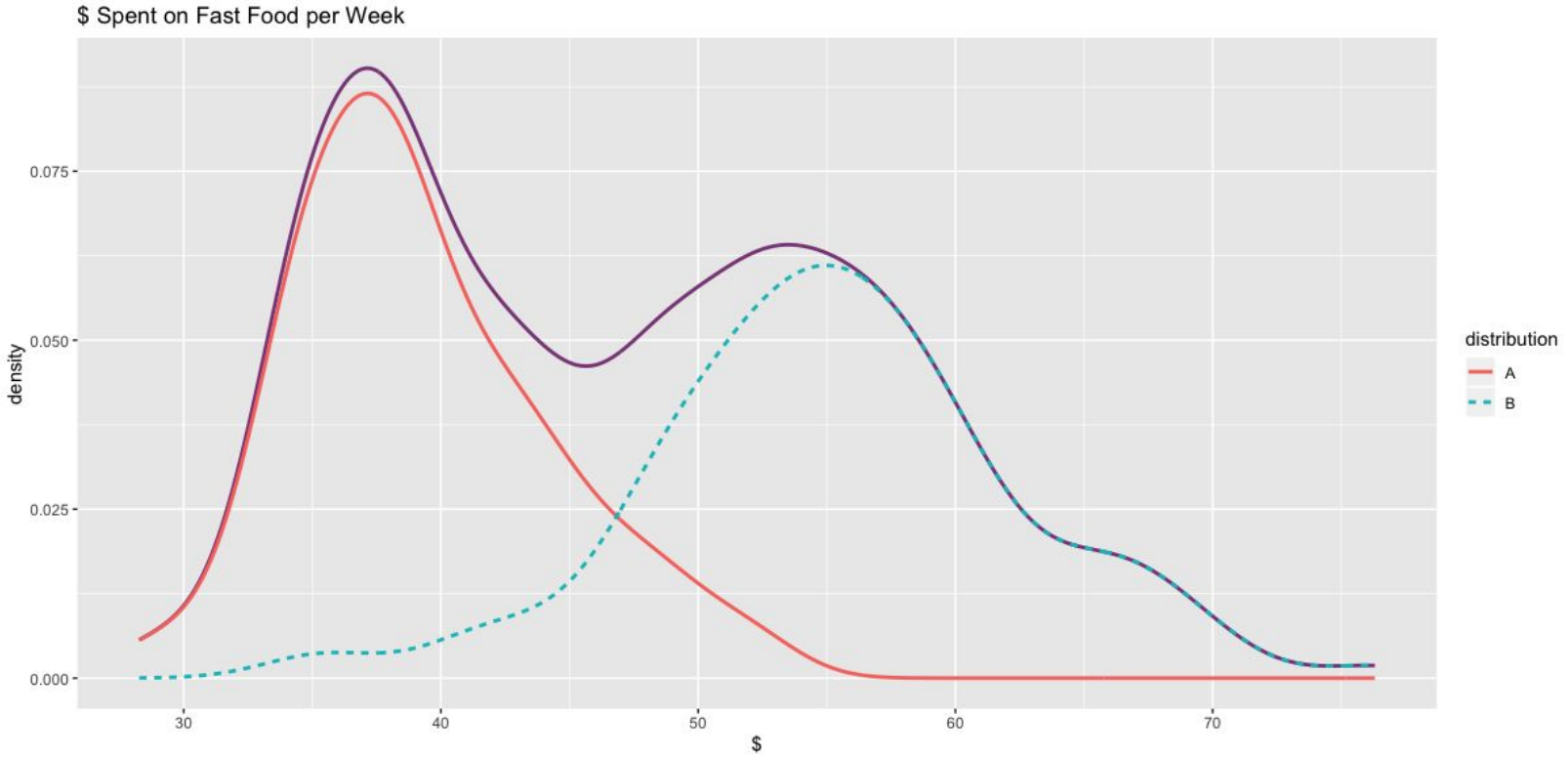
EM


SIMPLIFY
GROUP
REDUCE
@CHELSEAPARLETT

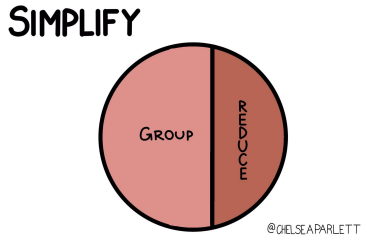

$ Spent on Fast Food per Week
density
0.075
0.050
0.025
0.000
30
40
50
60
70
$

EM


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

EM



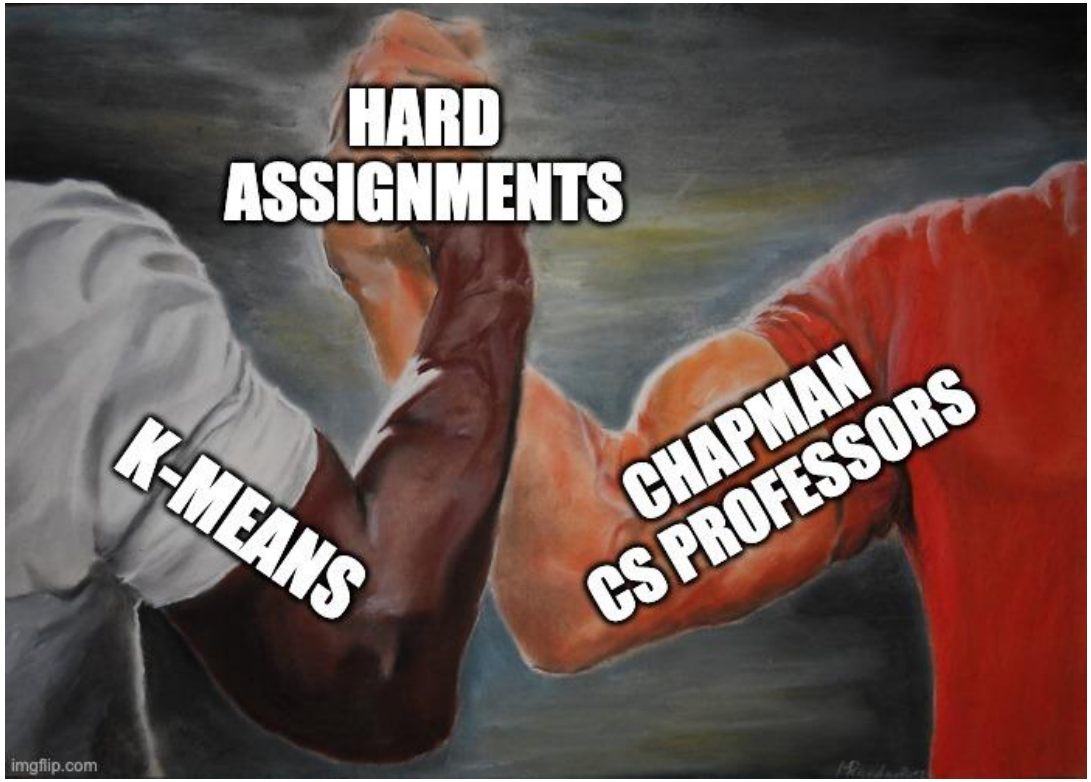
K means

- Hard Assignment
- All Variances the Same

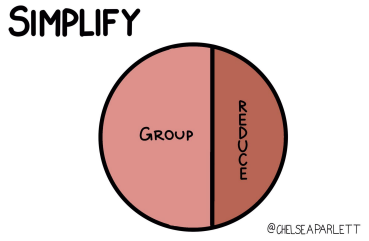
EM with mixtures of Gaussians

- Soft (probabilistic) Assignment
- Variances can be different

EM

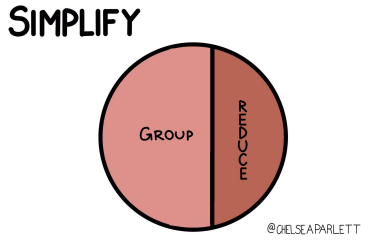

HARD
ASSIGNMENTS
K-MEANS
CHAPMAN
CS PROFESSORS
imgflip.com

K-Means Review

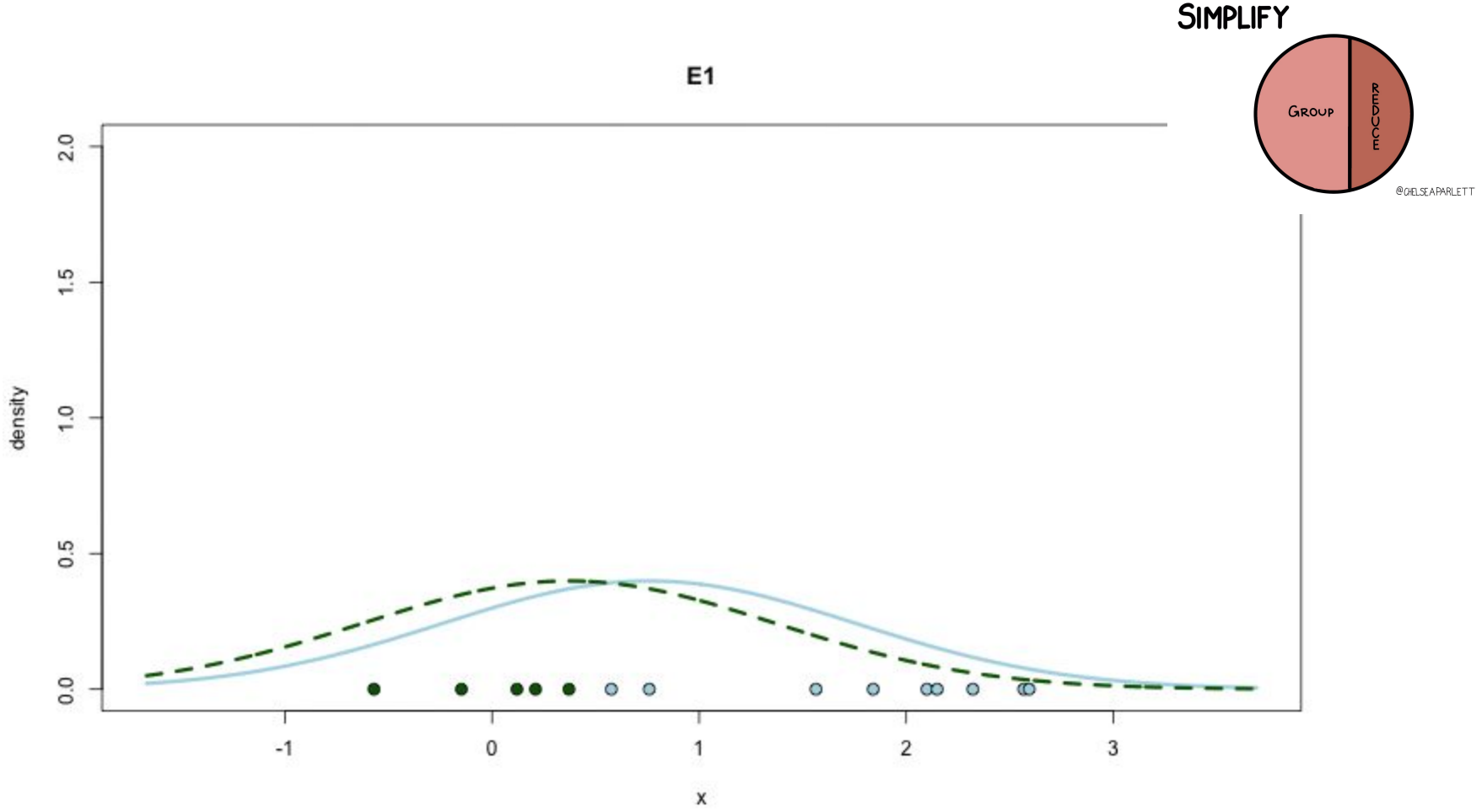


1. Choose **k** random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the **centers**
4. Repeat 2 and 3 until either:
    a. Cluster membership does not change
    b. Centers change only a tiny amount

EM



1. Choose **k** random points to be cluster centers (**or estimate using k-means…etc)**
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the **means + variances**
4. Repeat 2 and 3 until **distributions converge.**


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