CS 360: Machine Learning

Sara Mathieson, Sorelle Friedler

Spring 2024

HAVERFORD
COLLEGE

Admin

- **Lab 8** due TODAY!
  - Extra office hours **4-5pm TODAY in H110** (Sara)
- **Midterm April 25** in class (next Thursday)
  - Can still do handout videos for extra credit! Up to 24 hours before exam
- **Project presentations**: last week of classes
- **Writeup** due by the end of finals period
  - May 11 for seniors (AND groups involving seniors)
  - May 17 for non-seniors

Final Project Deliverables

- **Presentation**
  - Last week of classes
  - 12 min per pair
  - Peer feedback
- **Writeup**
  - In the style of a research paper
  - Text can be brief but should describe your motivation, hypotheses, data, methods, experiments, results, interpretation, and conclusions
  - At least 3 figures

README.md should have command lines for reproducibility!

Lab 7 competition

- **Fejiro and Pranav**
  - Pooling layer, add FC layer with 1000 units, optimizer RMSprop
  - Test accuracy 66%

- **Gavin and Neha**
  - DenseNet layers
  - Test accuracy 70%

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Supervised Learning:
makes use of examples where we know the underlying "truth" (label/output)

Supervised learning [hide]
(**classification** · **regression**)

Apprenticeship learning · Decision trees · Ensembles (Bagging · Boosting · Random forest) · *k*-NN · Linear regression · Naive Bayes · Artificial neural networks · Logistic regression · Perceptron · Relevance vector machine (RVM) · Support vector machine (SVM)

Clustering [hide]

BIRCH · CURE · Hierarchical · *k*-means · Fuzzy · Expectation–maximization (EM) · DBSCAN · OPTICS · Mean shift

Dimensionality reduction [hide]

Factor analysis · CCA · ICA · LDA · NMF · PCA · PGD · t-SNE · SDL

Structured prediction [hide]

Graphical models (Bayes net · Conditional random field · Hidden Markov)

Anomaly detection [hide]

RANSAC · *k*-NN · Local outlier factor · Isolation forest

Artificial neural network [hide]

Autoencoder · Cognitive computing · Deep learning · DeepDream · Feedforward neural network · Recurrent neural network (LSTM · GRU · ESN · reservoir computing) · Restricted Boltzmann machine · GAN · Diffusion model · SOM · Convolutional neural network (U-Net) · Transformer (Vision) · Mamba · Spiking neural network · Memtransistor · Electrochemical RAM (ECRAM)

Reinforcement learning [hide]

Q-learning · SARSA · Temporal difference (TD) · Multi-agent (Self-play)

Unsupervised Learning:
Learn underlying structure or features without labeled training data

Note: **generative models** are typically unsupervised!

Image: wikipedia

Unsupervised learning: main areas subject to debate

1) Clustering: group data points into clusters based on features only

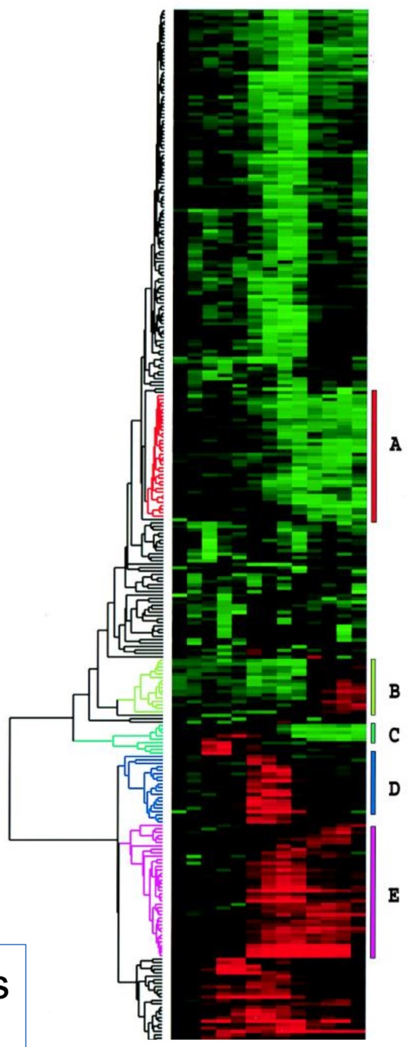
2) Dimensionality reduction: remove feature correlation, compress data, visualize data

3) Structured prediction: model latent variables (example: Hidden Markov Models)

4) Generative models: learn latent structure in order to generate novel examples

Applications of clustering

- Cluster genes with similar expression patterns



Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen,* Paul T. Spellman,* Patrick O. Brown,† and David Botstein*‡

Applications of clustering

- Image segmentation: cluster similar regions of an image

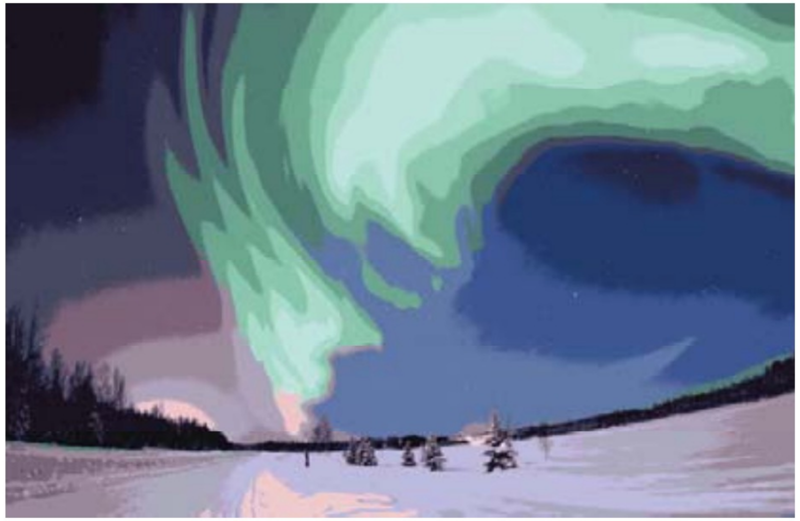


Image: Jessica Wu

Applications of clustering

- Clustering in social graphs

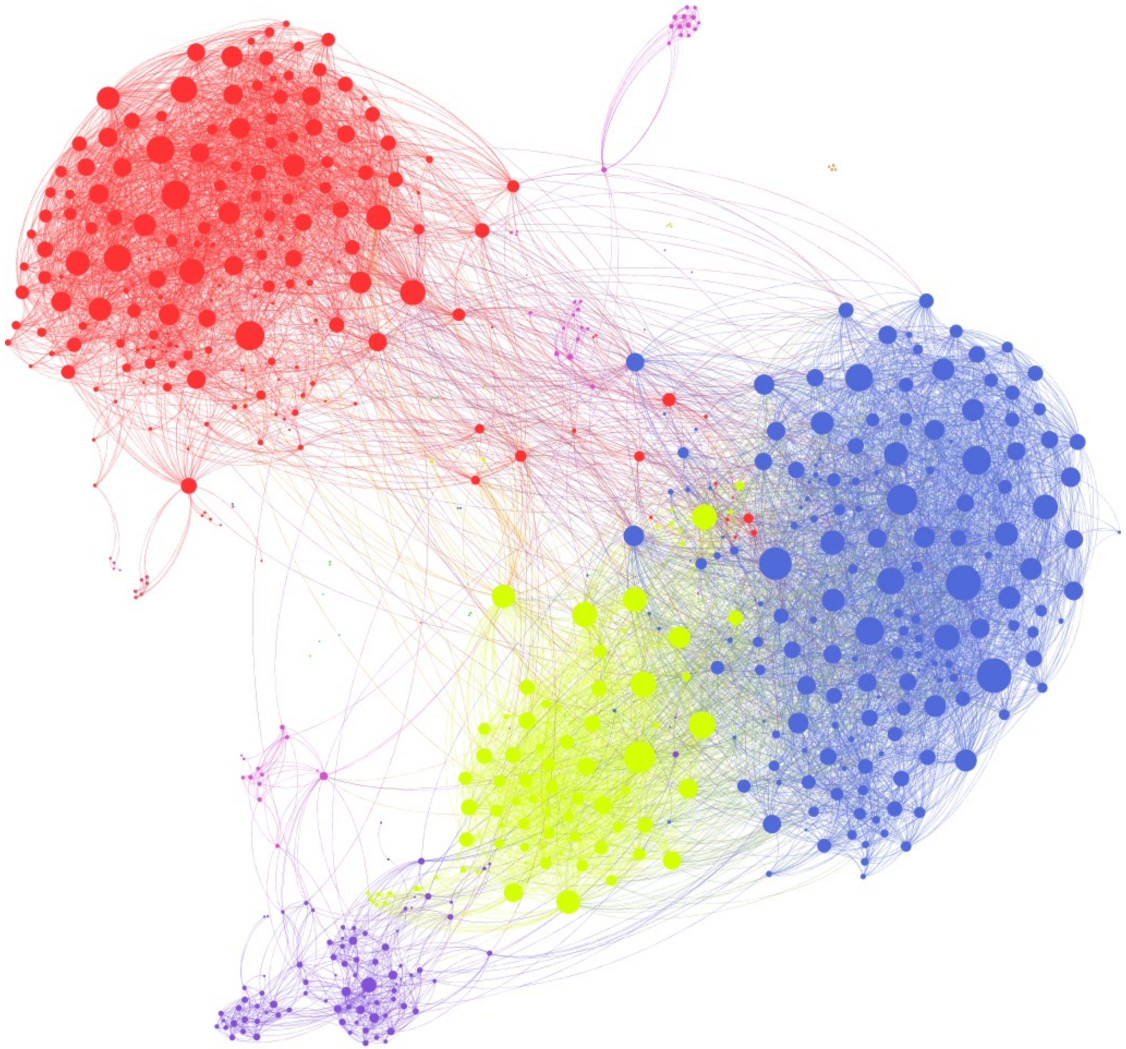
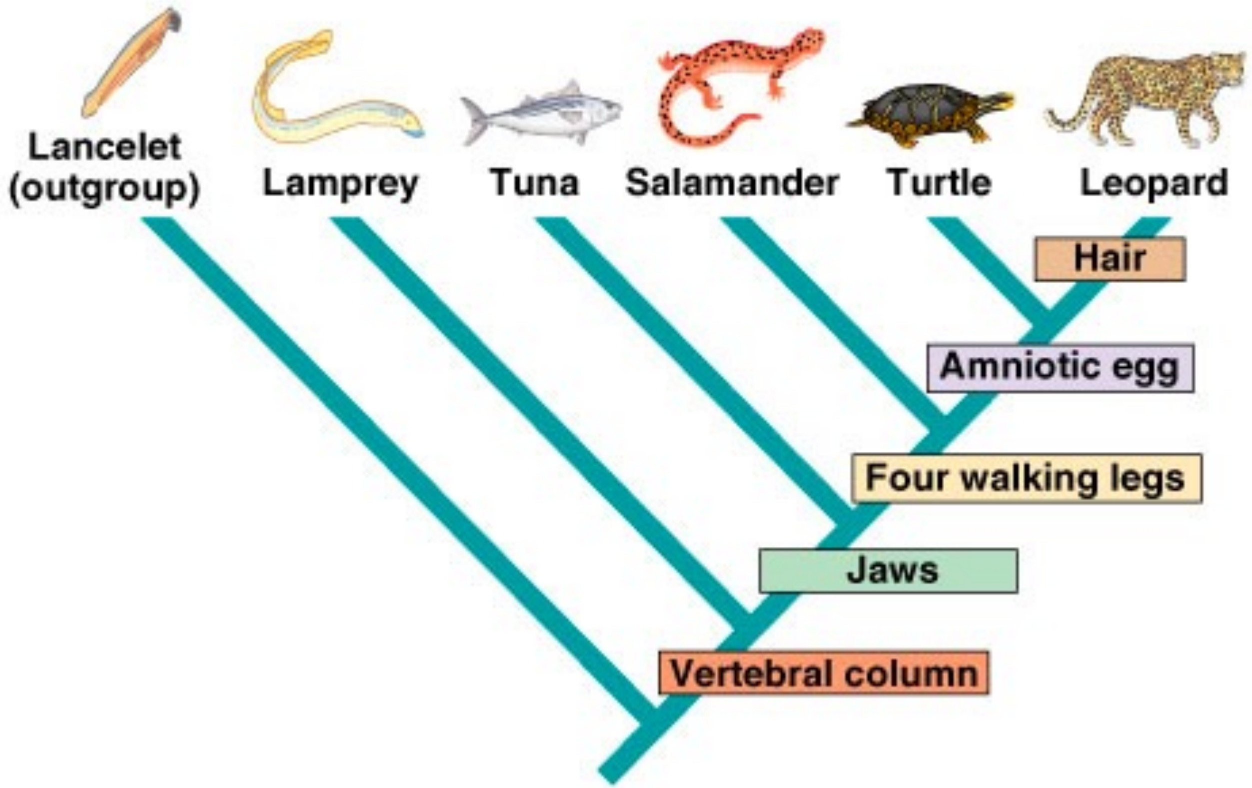


Image: https://griffsgraphs.wordpress.com/2012/07/02/a-facebook-network/

Two main types of clustering

- Flat/Partitional:
  - K-means
  - Gaussian mixture models
- Hierarchical:
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and Neighbor Joining

Hierarchical clustering example: trees



Credit: Pearson Education, Benjamin Cummings

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Review K-means from CS260

- Goal: learn about the structure in our data
- Goal: predict cluster membership for a new data point
- Method: minimize the within-cluster sum of squares (WCSS)

$$\mathcal{C} = \left\{ \mathcal{C}_1, \mathcal{C}_2 \cdots \mathcal{C}_K \right\}$$

$$\Downarrow \quad \Downarrow \quad \Downarrow$$

clusters $\quad \left\{ \vec{\mu}_1 \quad \vec{\mu}_2 \quad \vec{\mu}_K \right\}$ means

$$\mathcal{C}_k = \left\{ \vec{x}_7, \vec{x}_{12}, \vec{x}_{18} \right\}$$

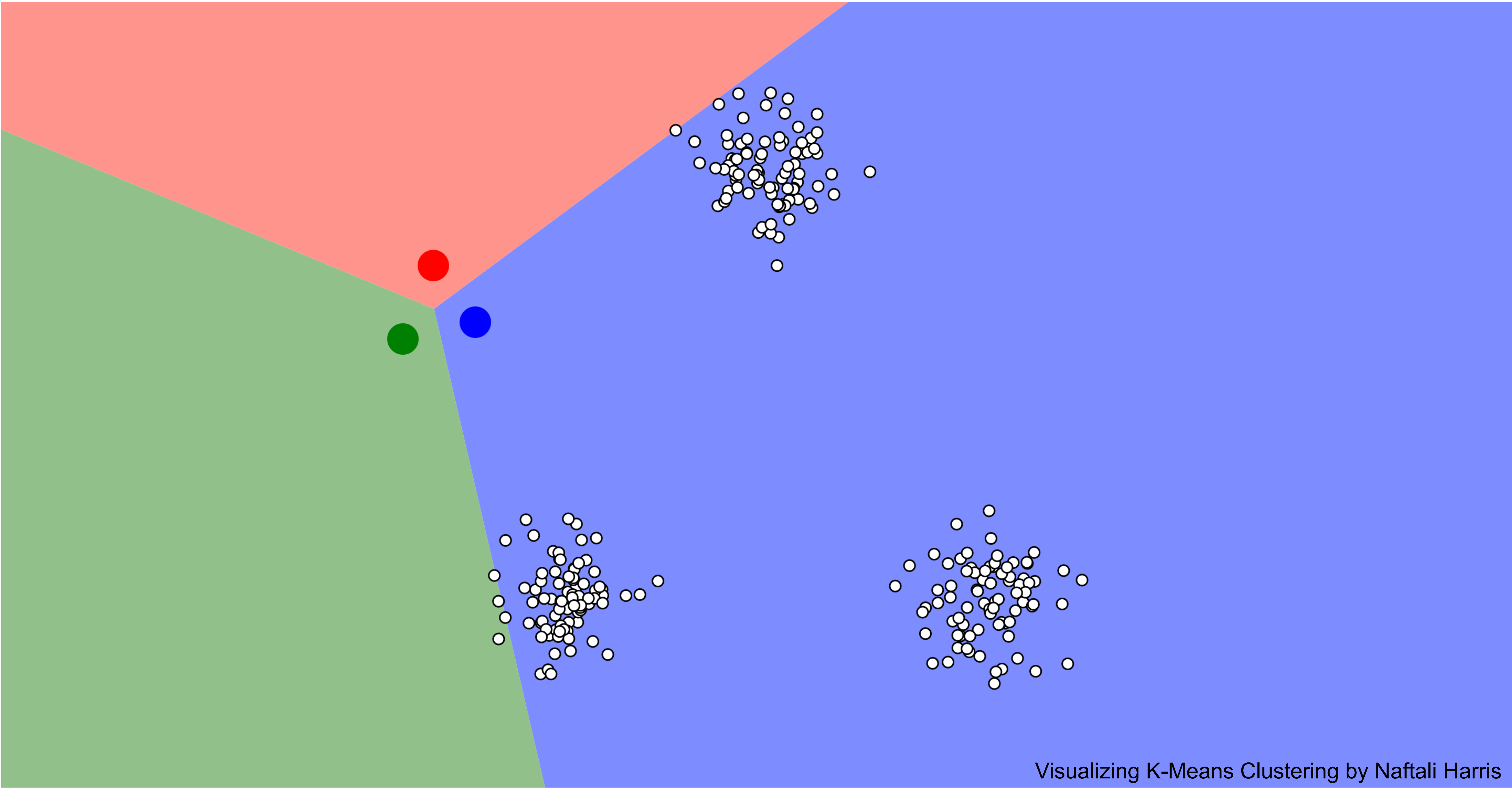
Minimize

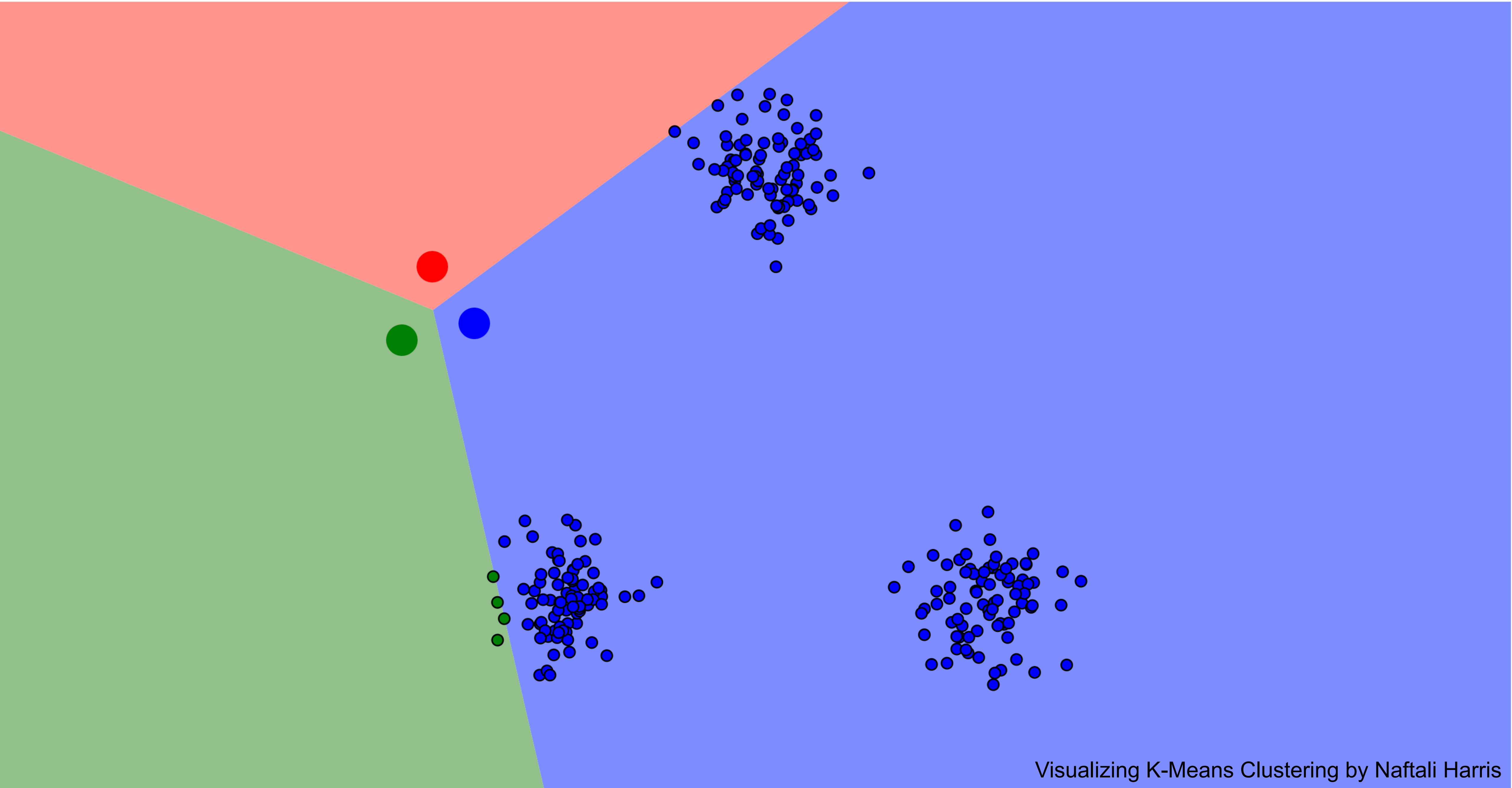
$$\uparrow \text{WCSS} = \boxed{\sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \| \vec{x}_i - \vec{\mu}_k \|^2}$$

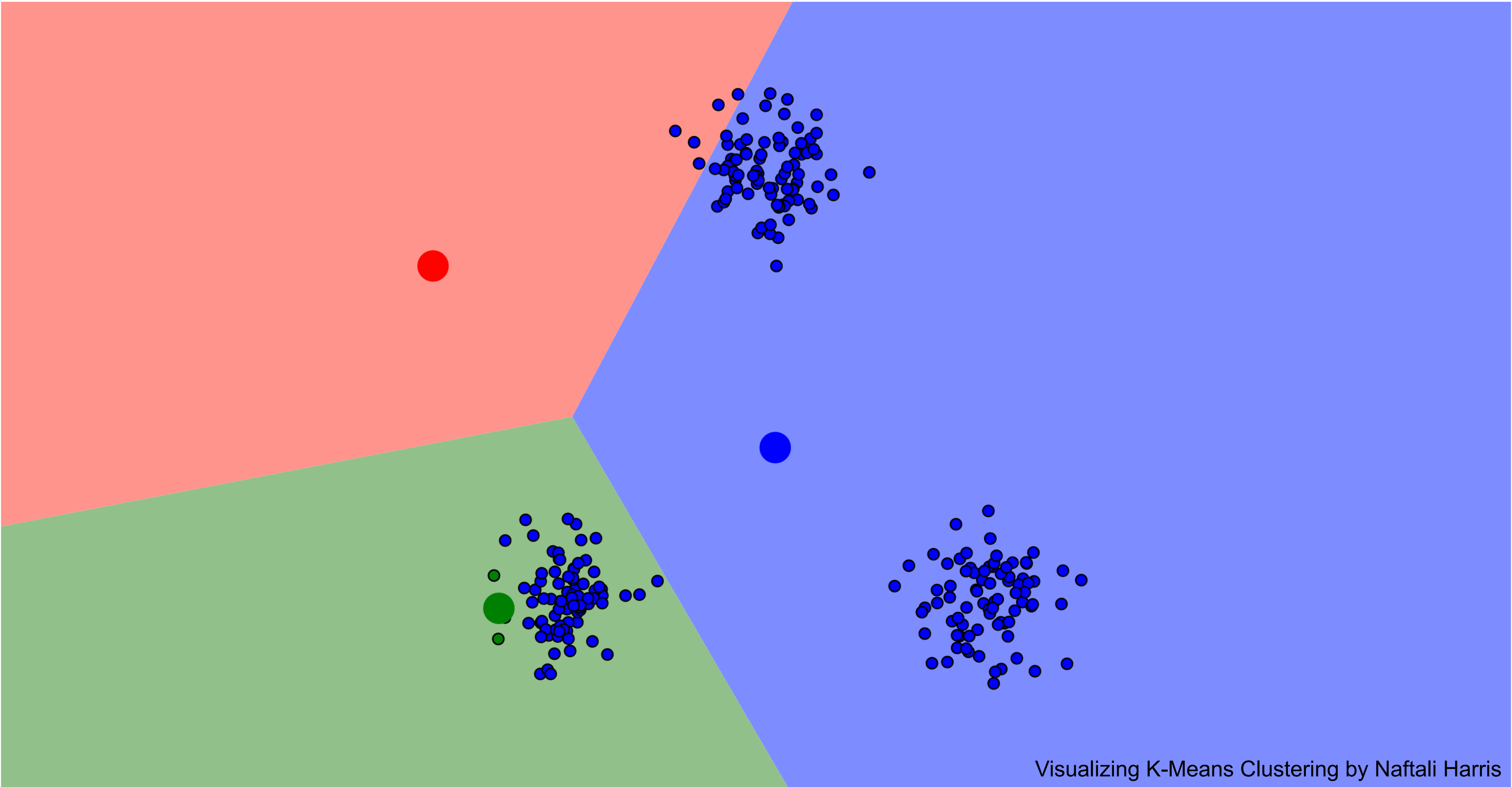
NP-hard

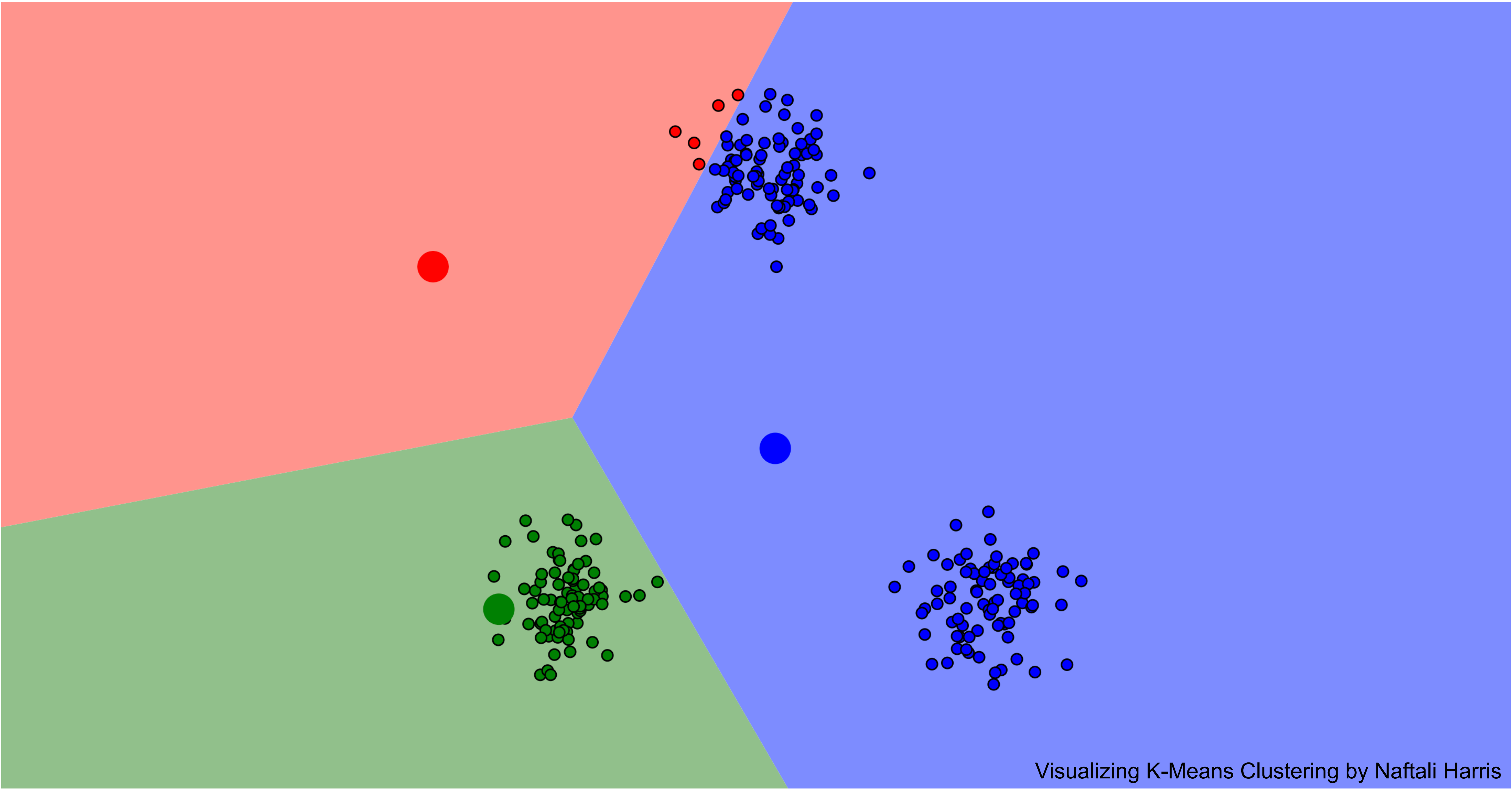
Review K-means from CS260

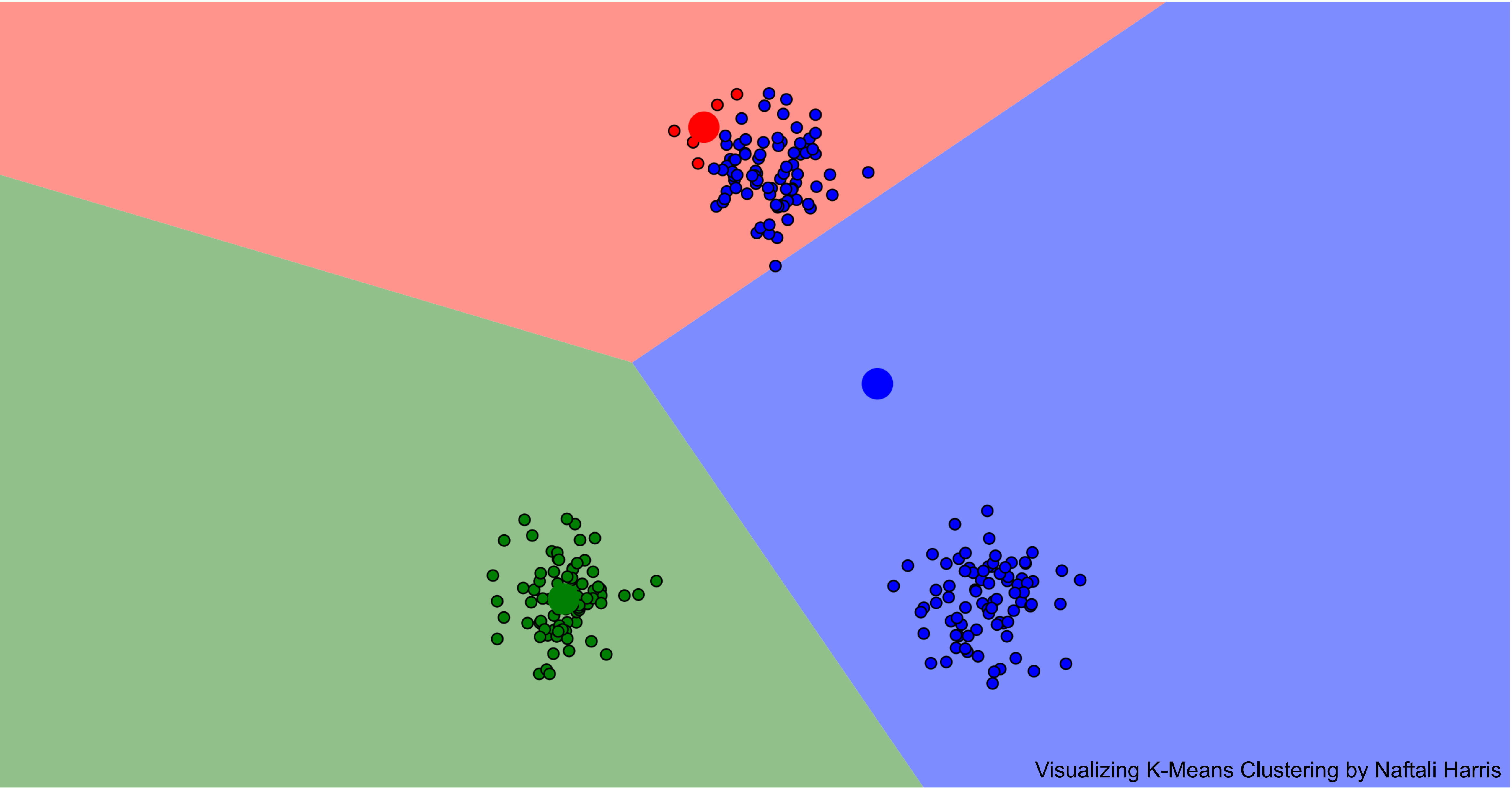
- **Initialization**: choose K means (cluster centers) randomly
  - Usually from among the training data
- **E-step**: assign each point to the closest mean
- **M-step**: update the means as the cluster averages

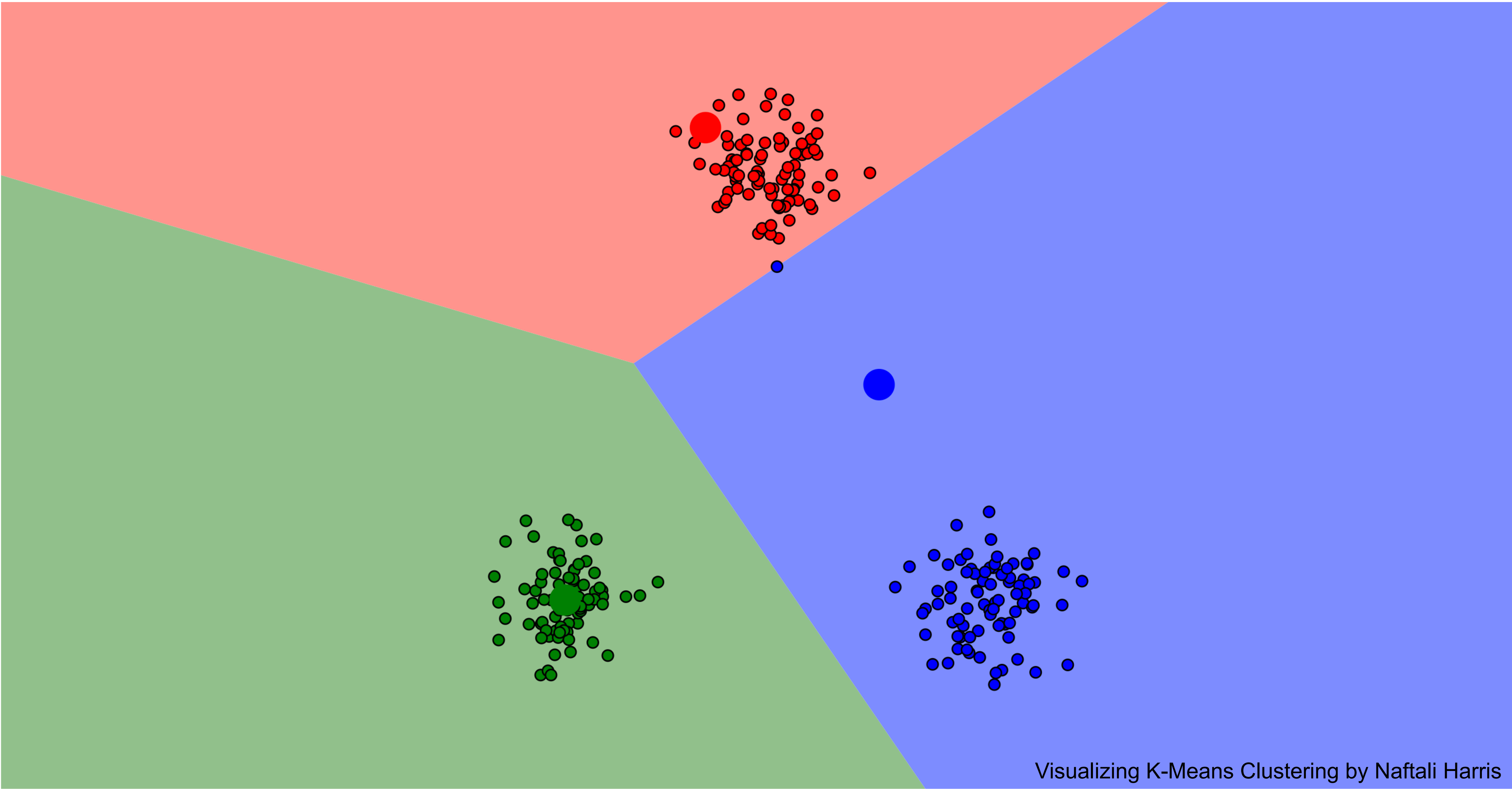

Visualizing K-Means Clustering by Naftali Harris

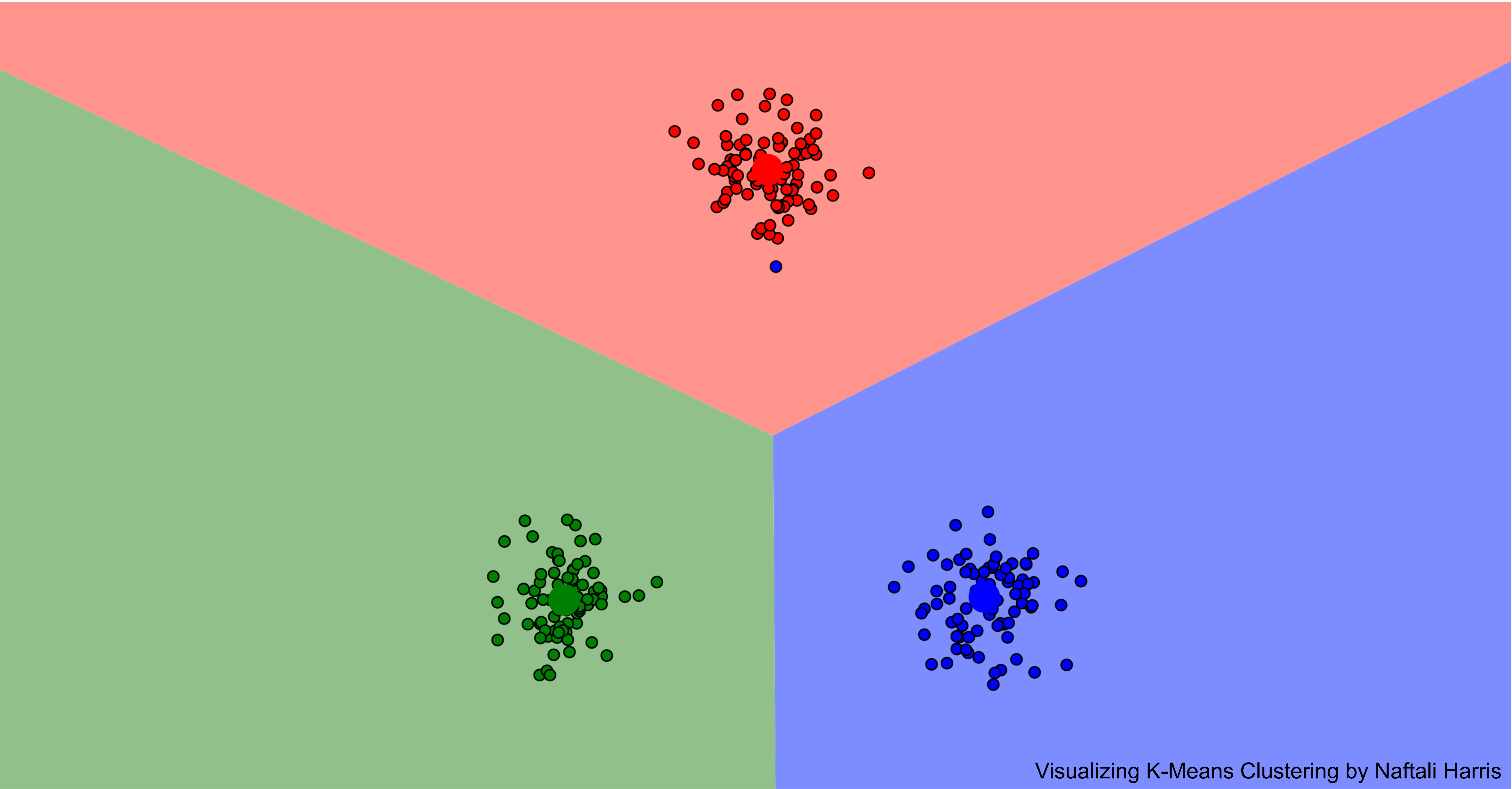

Visualizing K-Means Clustering by Naftali Harris

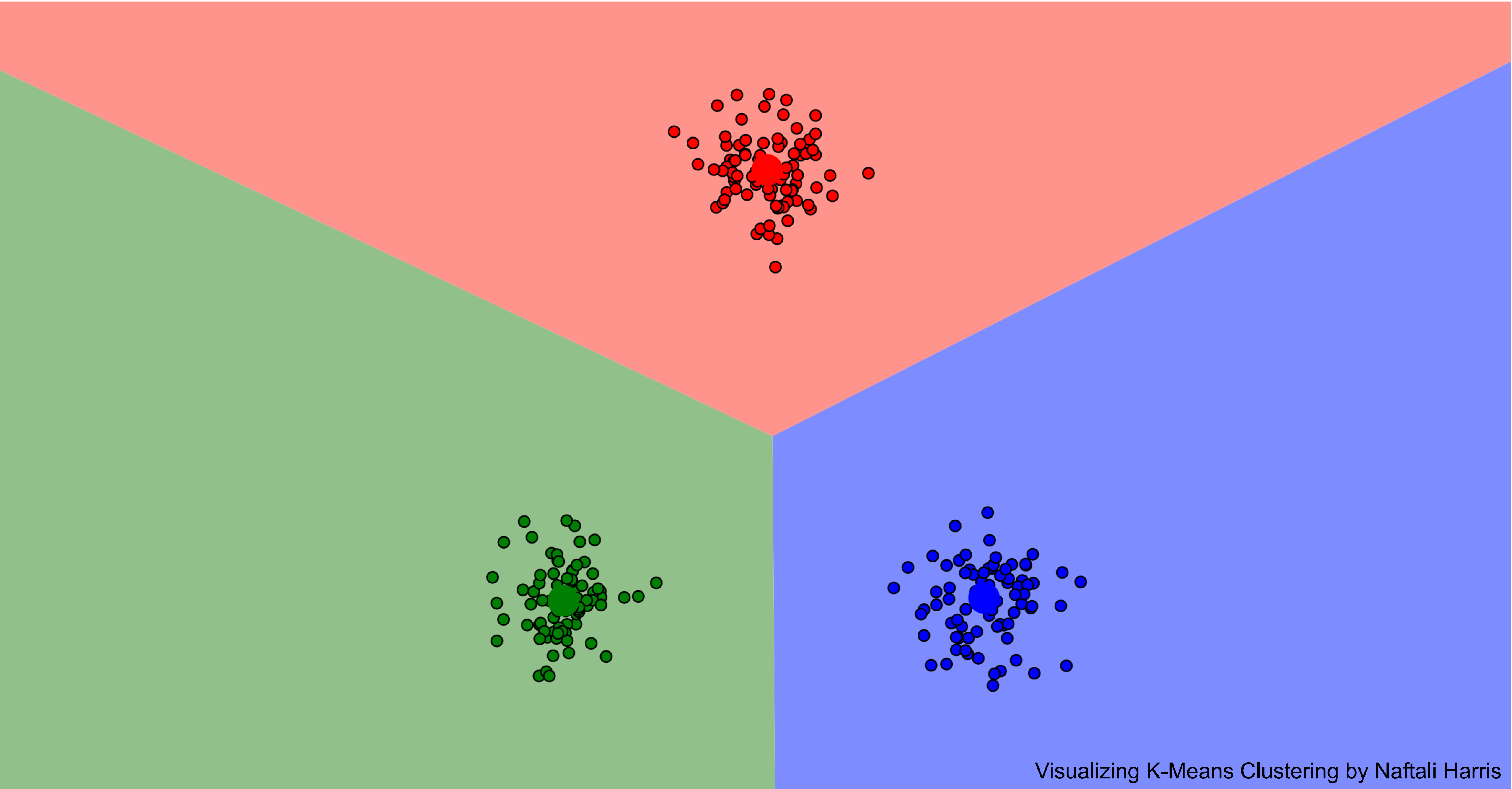

Visualizing K-Means Clustering by Naftali Harris


Visualizing K-Means Clustering by Naftali Harris


Visualizing K-Means Clustering by Naftali Harris


Visualizing K-Means Clustering by Naftali Harris


Visualizing K-Means Clustering by Naftali Harris


Visualizing K-Means Clustering by Naftali Harris

How to choose the number of clusters K?

- A larger K will always reduce the within-cluster sum of squares (WCSS)!
- Try a variety of K and choose the K value where the decreases in WCSS plateau
- **Question**: what is the WCSS when K=n? (the number of training points)

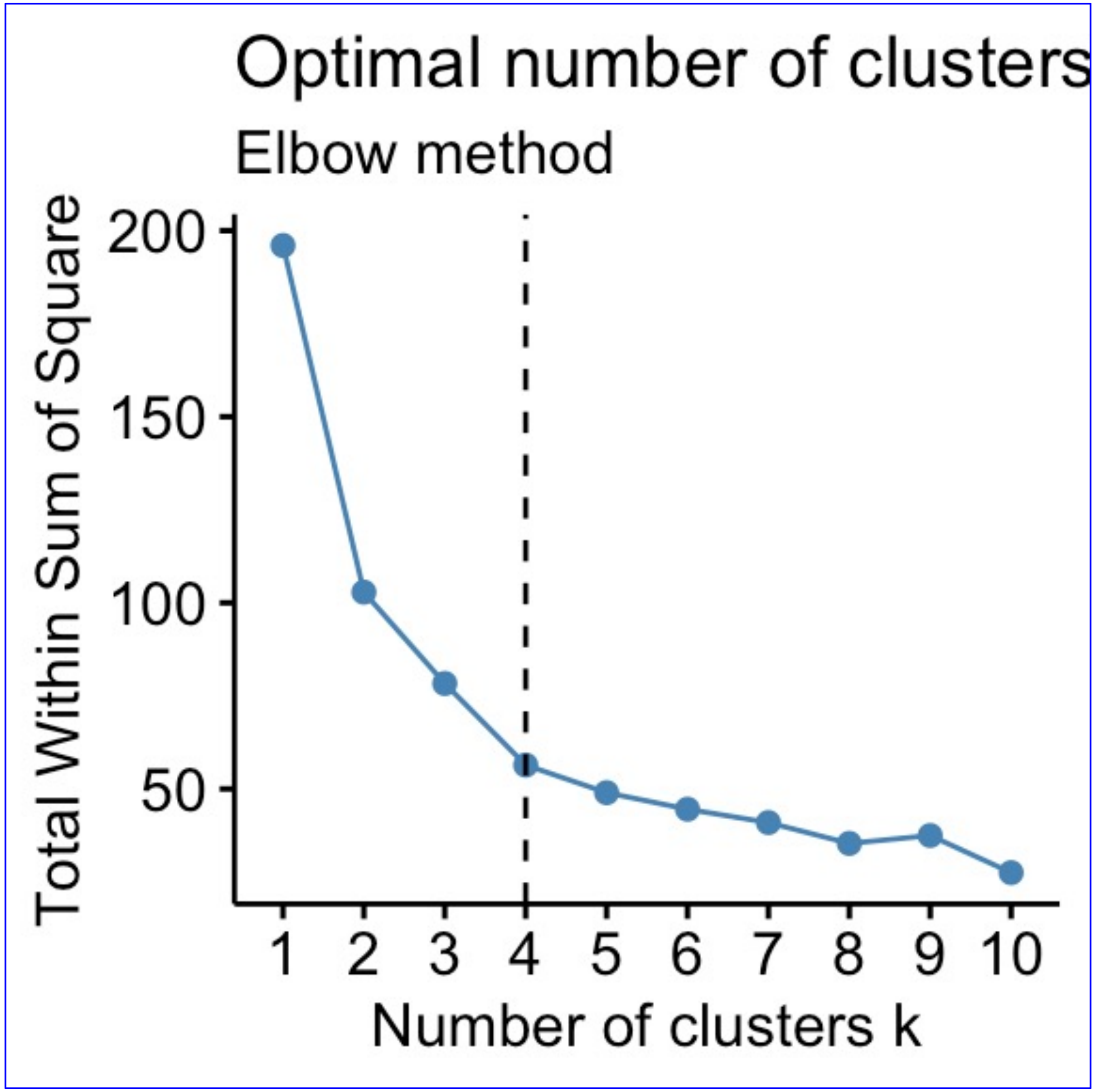


Image: Datanovia

K-means stopping criteria

- No cluster membership changes
- Max number of iterations exceeded
- See a configuration you’ve seen before (cycle)

Discriminative vs. Generative

- <u>Discriminative</u>: finds a decision boundary
  - Logistic regression, K-means
- <u>Generative</u>: estimates probability distributions
  - Naïve Bayes, Gaussian Mixture Models

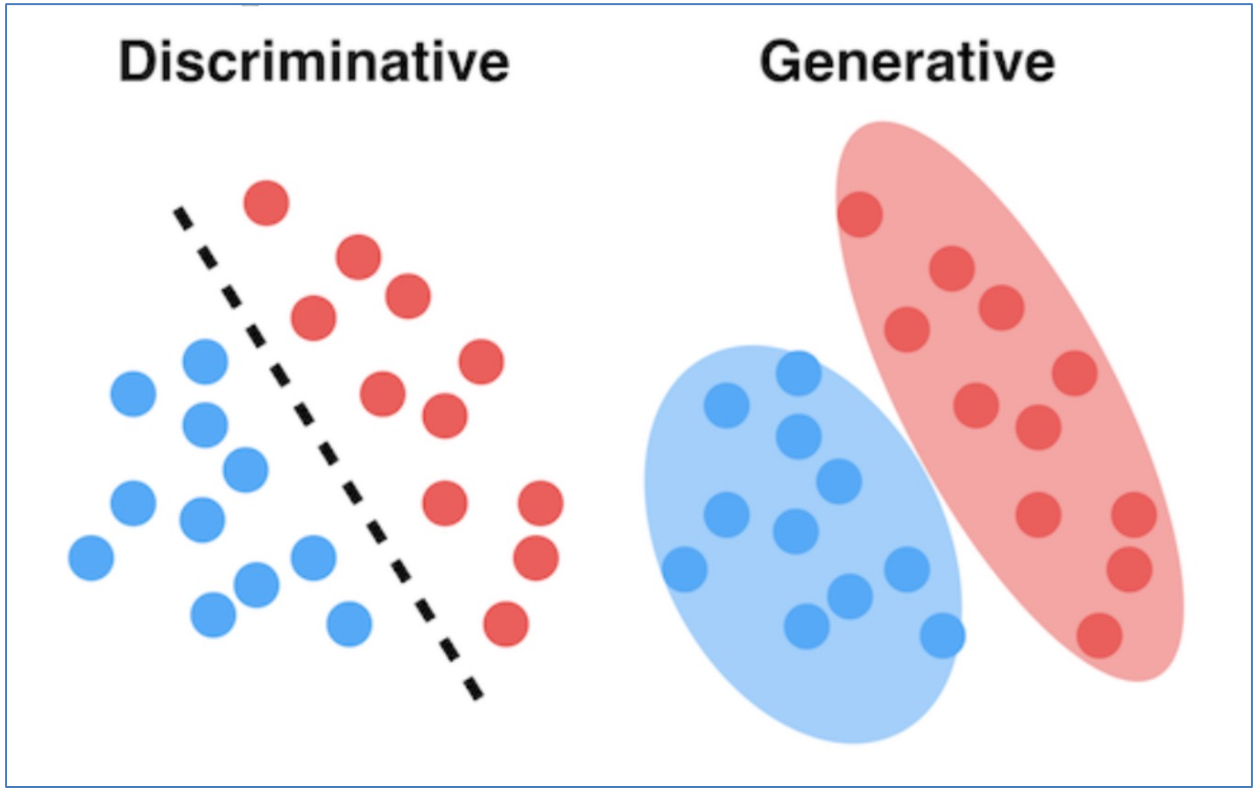


Figure: Ameet Soni

Problems with K-means

- Not generative (could not create a new data point)
- Does not account for different cluster sizes and variances
- Does not allow points to belong to multiple clusters

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Gaussian Mixture Models GMM

(K clusters)

Likelihood

$$p(\vec{x}) = \sum_{k=1}^{K} p(\vec{x}, \underset{\substack{\uparrow \\ \text{cluster} \\ \text{membership}}}{z} = k) \xrightarrow{\text{BAYES}} = \sum_{k=1}^{K} \underbrace{p(z=k)}_{\underbrace{\pi_k}_{\substack{\text{size} \\ \text{of cluster} \\ k}}} \underbrace{p(\vec{x} \mid z=k)}_{\substack{\text{given cluster,} \\ \text{prob of } \vec{x}}}$$

$$\underset{\substack{\uparrow \\ \text{all} \\ \text{train} \\ \text{data}}}{L(X)} = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(\vec{x}_i ; \underbrace{\vec{\mu}_k}_{\text{mean}}, \underbrace{\sigma_k^2}_{\text{variance}})}_{\text{normal distribution}}$$

Goal:

find

$$\boxed{\pi_k, \vec{\mu}_k, \sigma_k^2}$$

that maximize likelihood!

EM algorithm

- π_k = prob of cluster k
  $= \frac{1}{K}$ to start
- $\vec{\mu}_k$ = mean of cluster k
  = random data pt to start
- σ_k^2 = variance of cluster k
  = Sample variance of all points closest to $\vec{\mu}_k$

$\begin{bmatrix} w_{11} & w_{12} & w_{13} \end{bmatrix}$

E-s

let

w_{ik}

e

$W = \begin{bmatrix} 0. \end{bmatrix}$

$\boxed{\text{E-step}}$ (soft assignment)

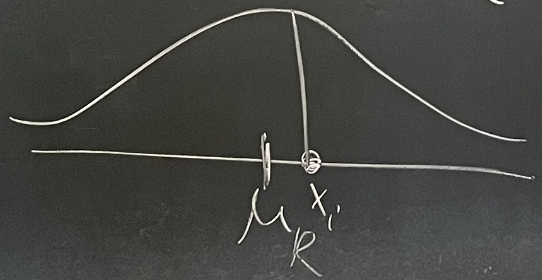
let w_{ik} = prob that $\vec{x}_i$ came from cluster k

$$w_{ik} = p(k \mid \vec{x}_i) = \frac{p(k)\,p(\vec{x}_i \mid k)}{p(\vec{x}_i)}$$

$$= \frac{\pi_k N(\vec{x}_i; \vec{\mu}_k, \sigma_k^2)}{\underbrace{\sum_{k'} \pi_{k'} N(\vec{x}_i; \vec{\mu}_{k'}, \sigma_{k'}^2)}_{\text{normalize}}}$$

w_{13}

$$W = \begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \underbrace{\left[\begin{array}{ccc} 0.5 & 0.3 & 0.2 \\ & & \\ & & \\ & & \end{array}\right]}_{K} \end{array} \Big\} n$$

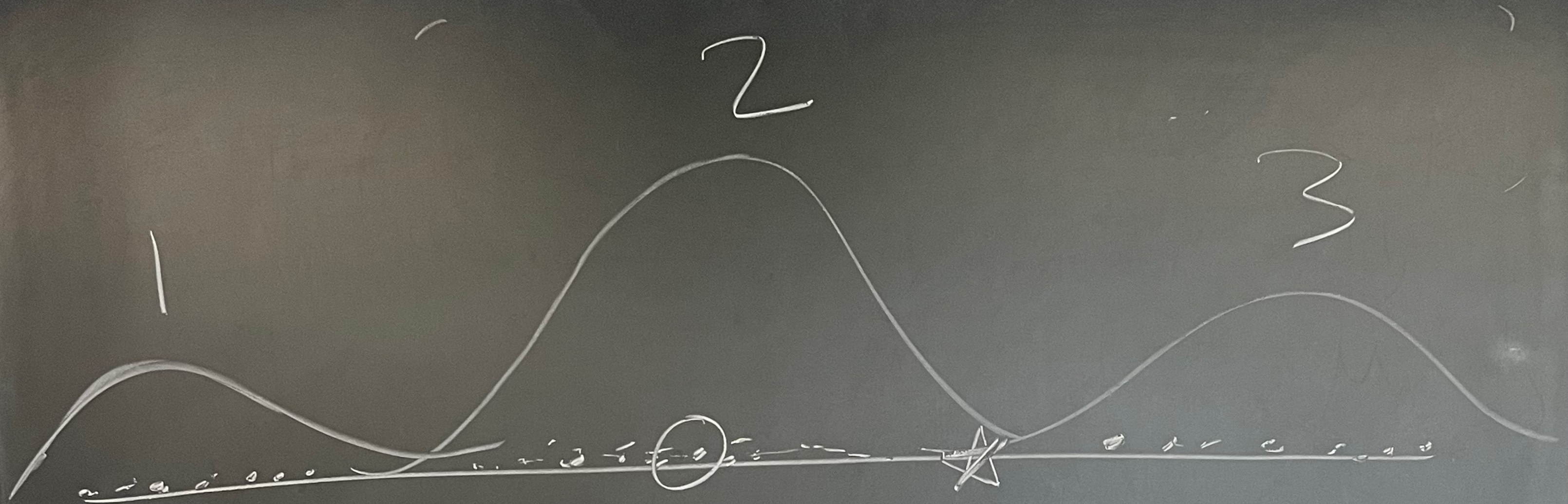


$\boxed{\text{M-step}}$ let $M_k = \underbrace{\sum_{i=1}^{n} w_{ik}}$ "#" of points in cluster k

$$\boxed{\pi_k = \frac{M_k}{n}}$$

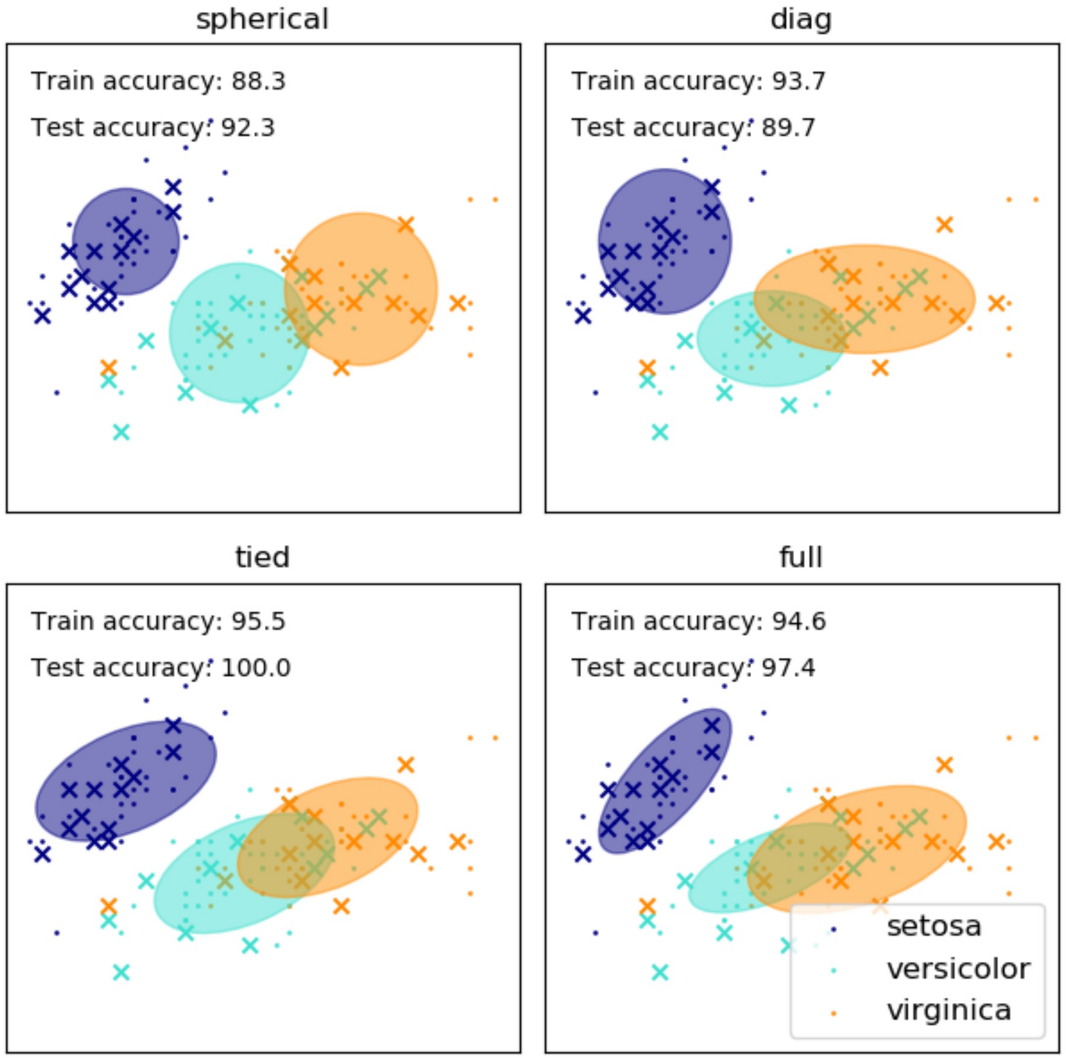
$$\boxed{\vec{\mu}_k = \frac{1}{M_k} \sum_{i=1}^{n} w_{ik} \vec{x}_i}$$

σ_k^2 = weighted sample variance

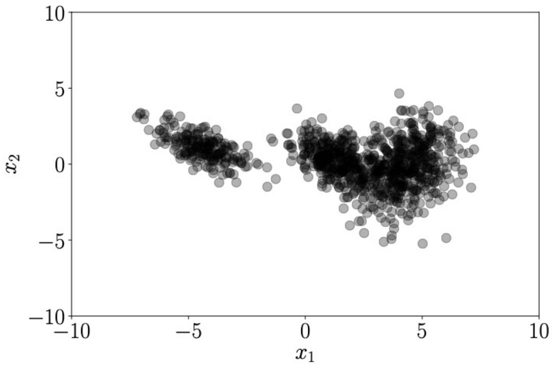


$$
\begin{array}{c}
\star \\
\bigcirc
\end{array}
\begin{bmatrix}
0.01 & 0.49 & 0.5 \\
0.01 & 0.99 & 0.0
\end{bmatrix}
$$

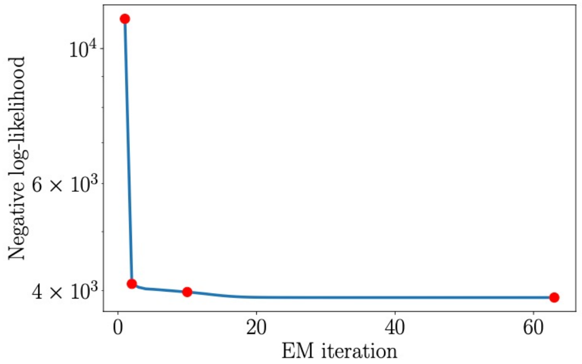
Example of GMMs with different covariance constraints on the Iris flower data



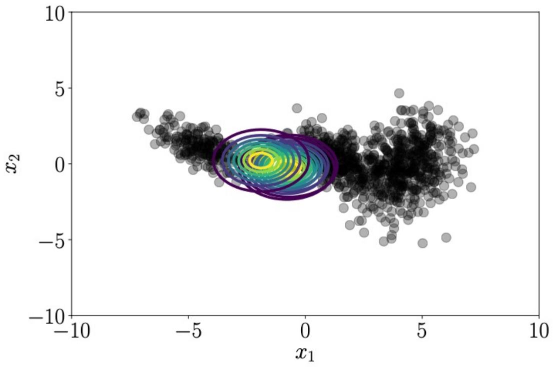
https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html#sphx-glr-auto-examples-mixture-plot-gmm-covariances-py



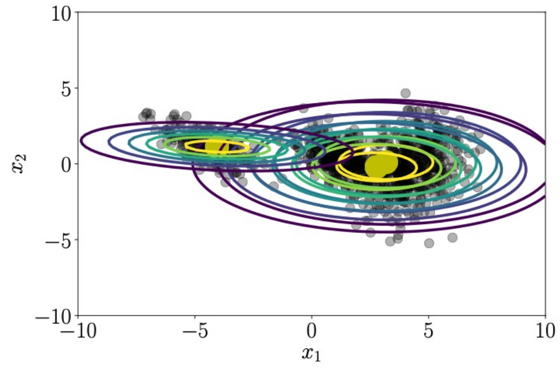
(a) Dataset.



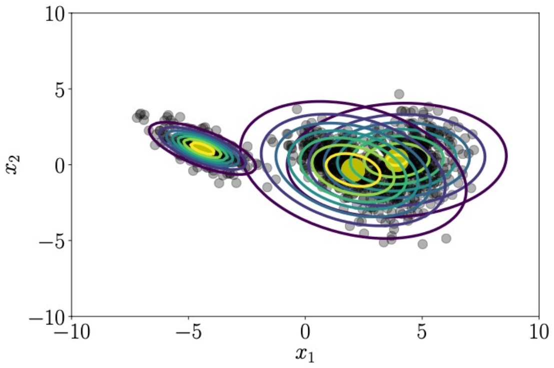
(b) Negative log-likelihood.



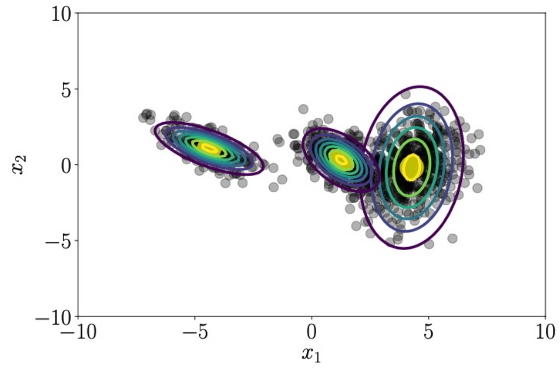
(c) EM initialization.



(d) EM after one iteration.



(e) EM after 10 iterations.



(f) EM after 62 iterations.

Figure 11.7 from MML textbook

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Principal Components Analysis (PCA)

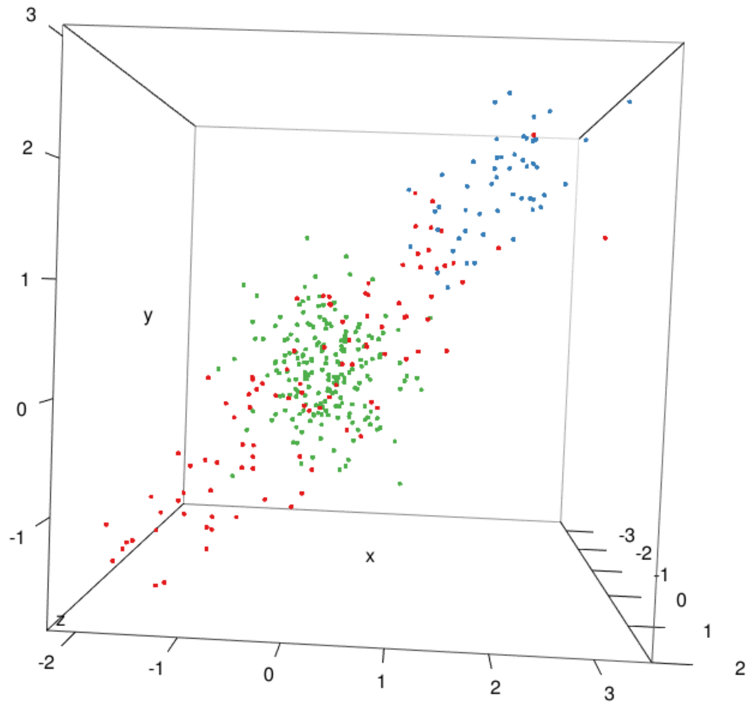
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

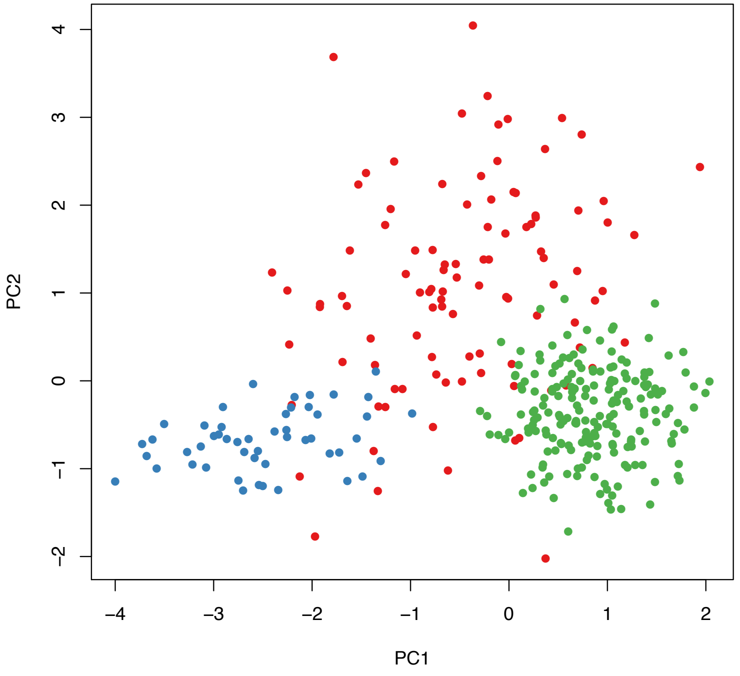
- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction and visualization

- PCA is a linear transformation

- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

Principal component analysis





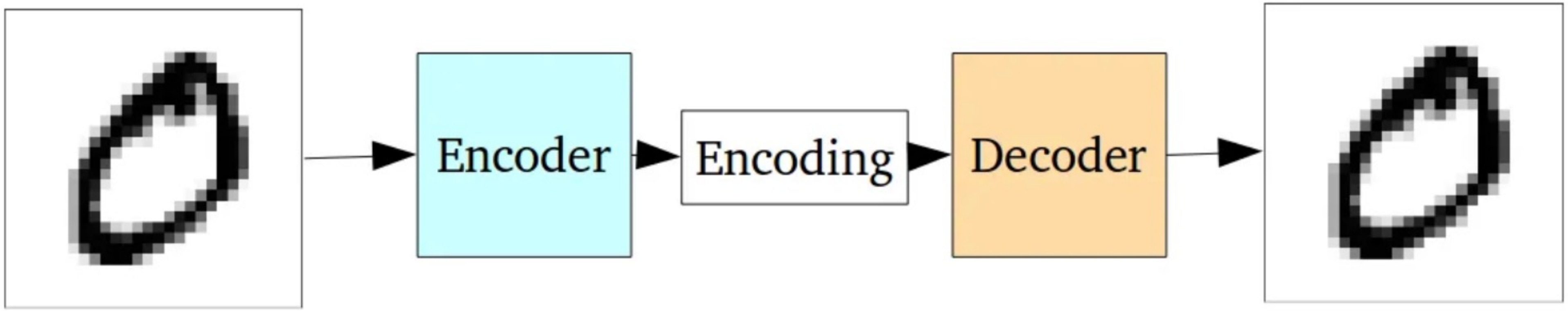
Slide: Iain Mathieson

Outline for April 18

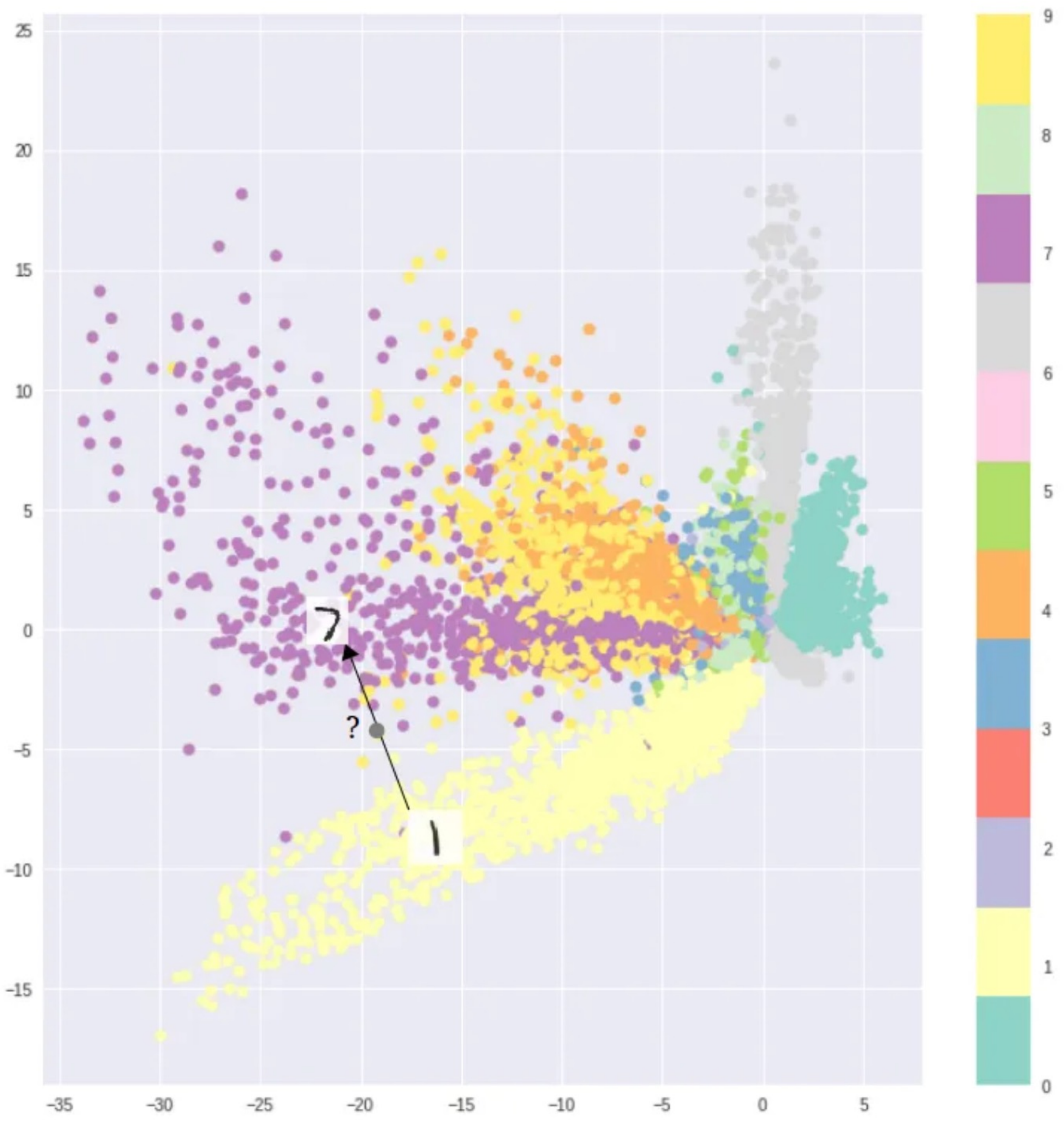
- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Traditional Autoencoder

- Project data in a lower dimension (encoder), then reconstruct using a decoder
- Loss function: minimize reconstruction error



Intuitively Understanding Variational Autoencoders by Irhum Shafkat

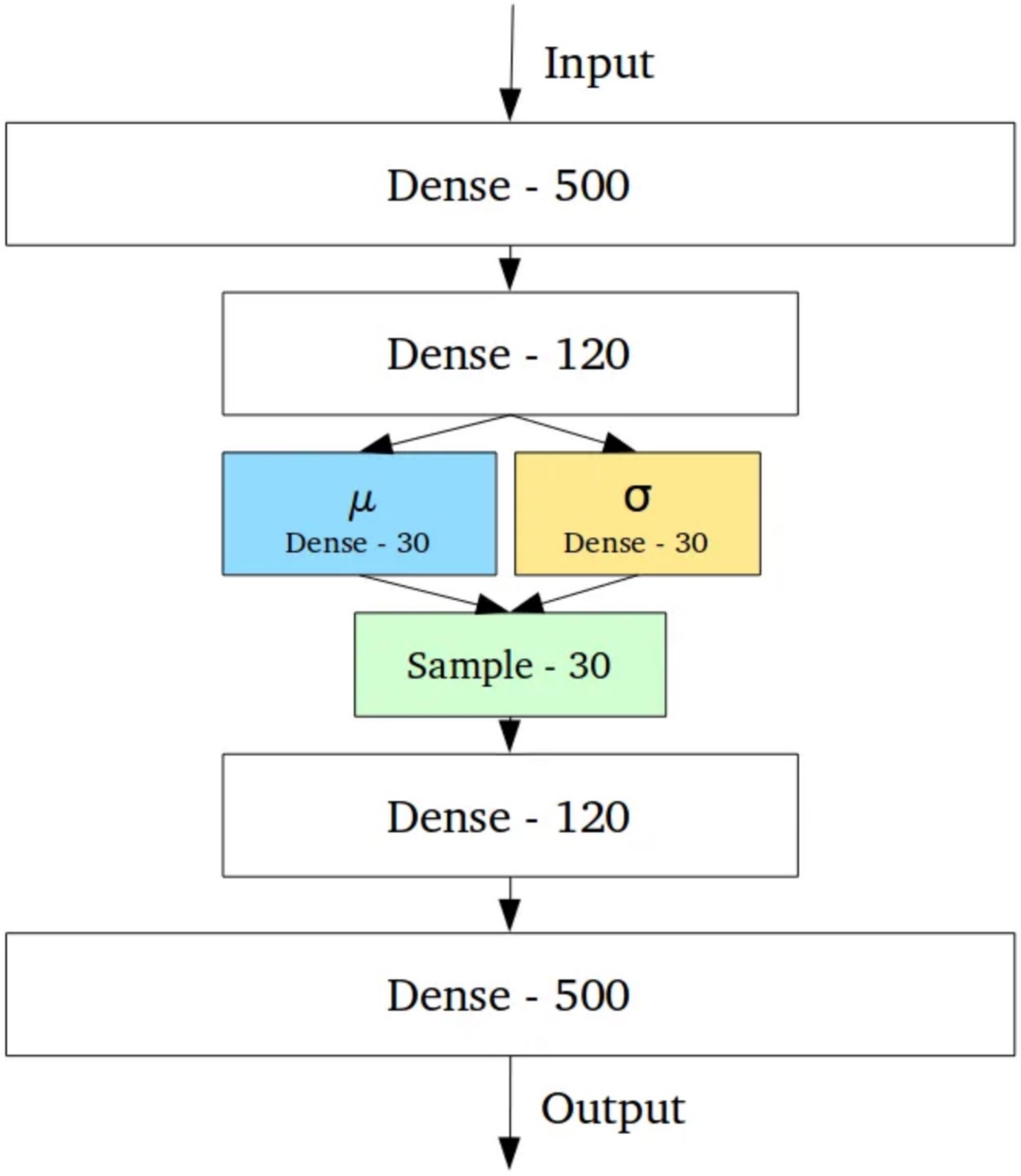


Optimizing purely for reconstruction loss

Traditional Autoencoder on MNIST

Intuitively Understanding Variational Autoencoders by Irhum Shafkat

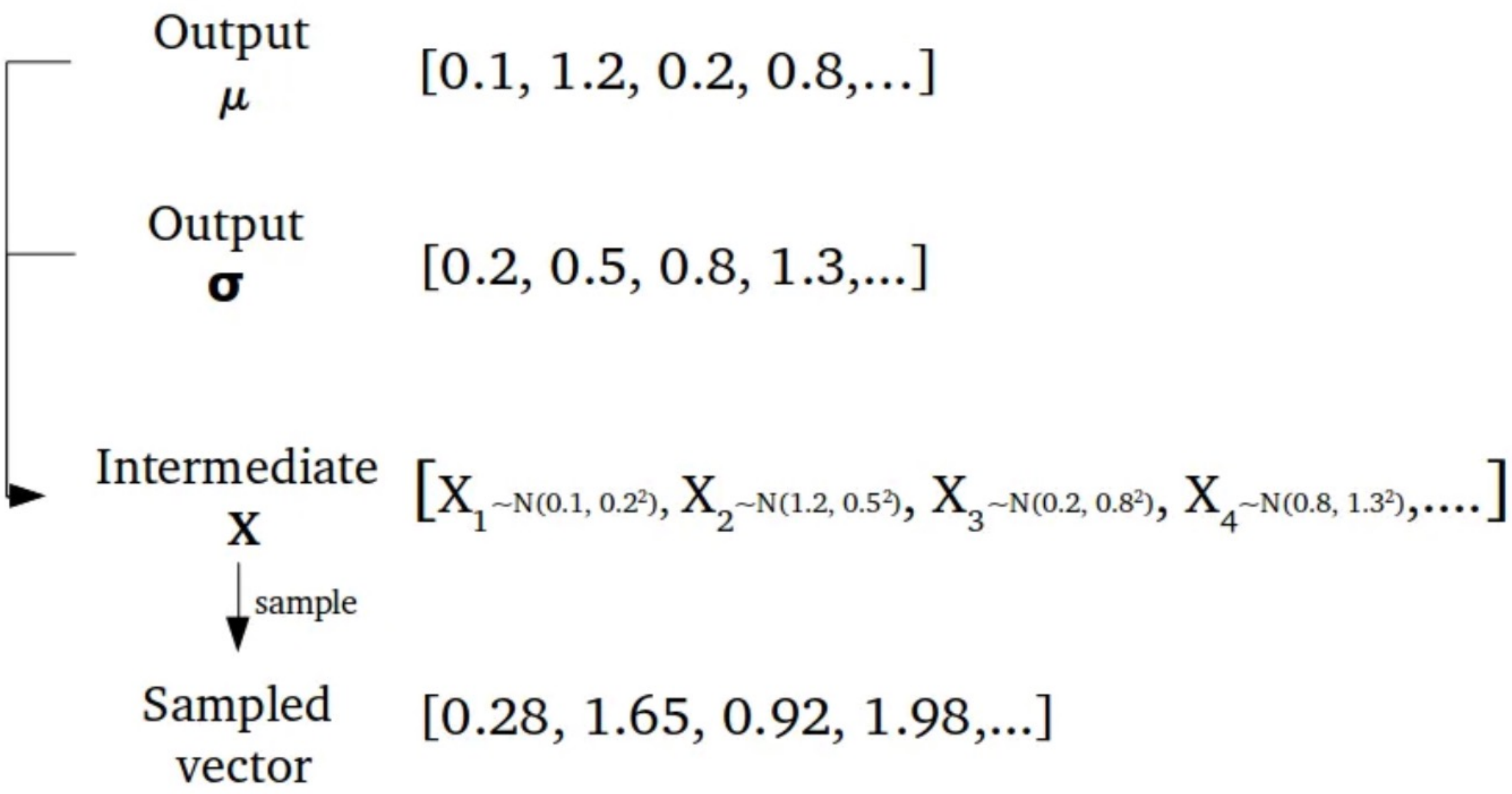
Variational Autoencoder: latent space has a set of means and variances



Variational Autoencoder

Intuitively Understanding Variational Autoencoders by Irhum Shafkat

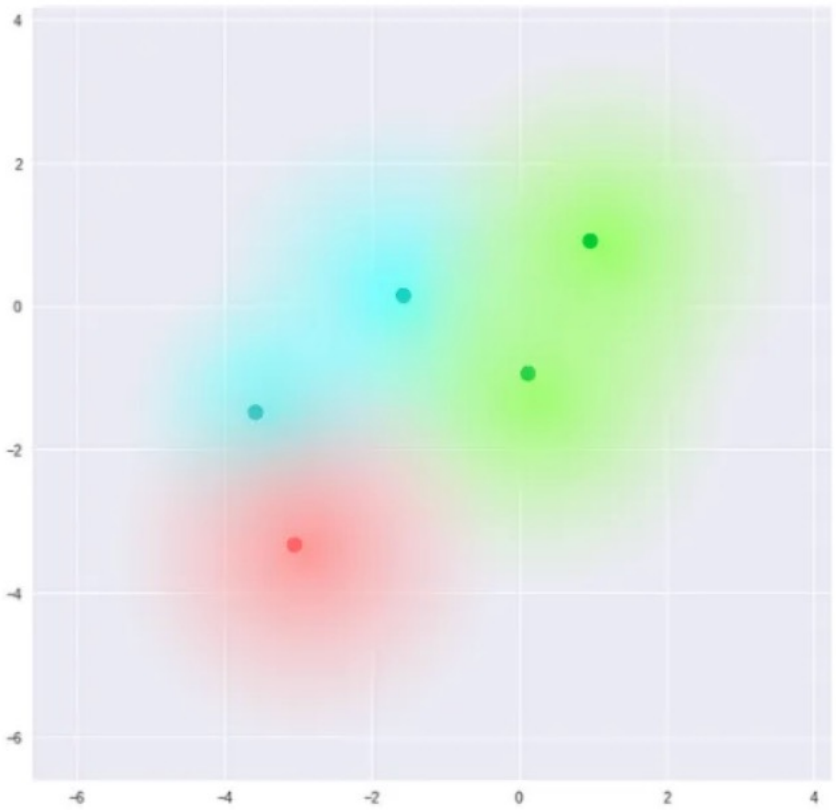
Example of a trained VAE



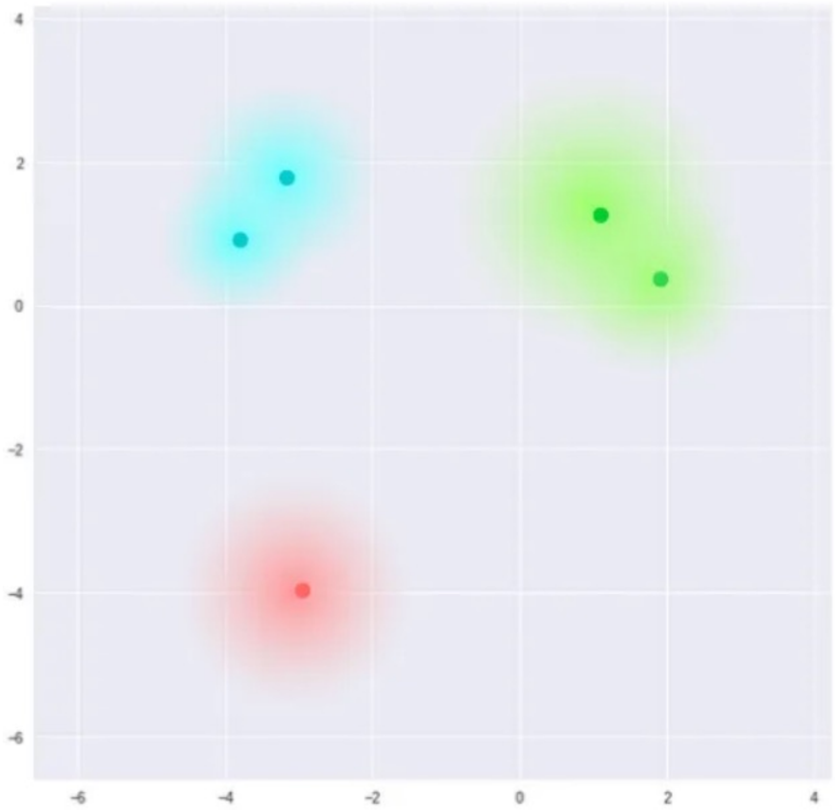
Stochastically generating encoding vectors

Intuitively Understanding Variational Autoencoders by Irhum Shafkat

VAE: don’t want “spaces” between our clusters

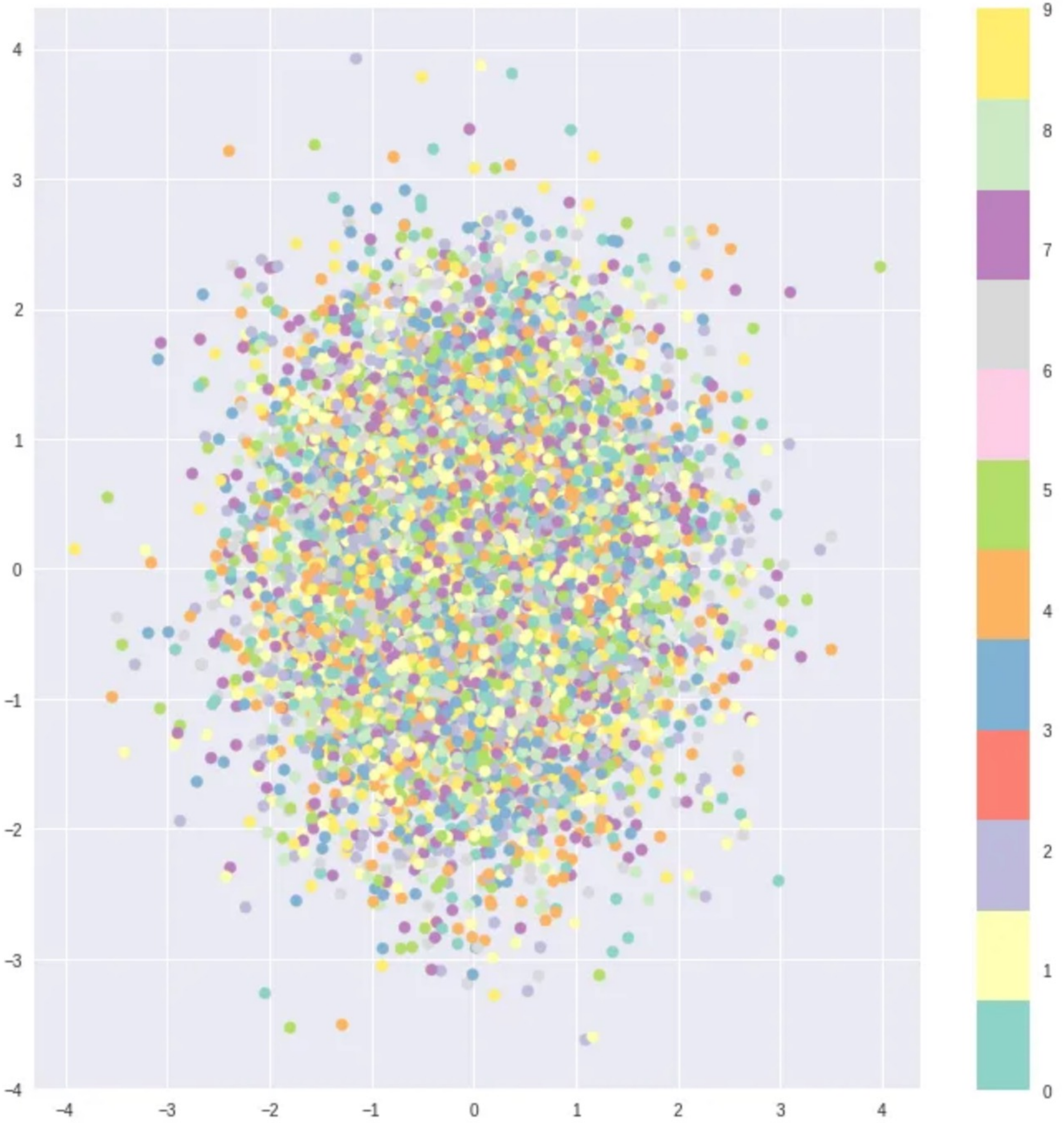


What we require



What we may inadvertently end up with

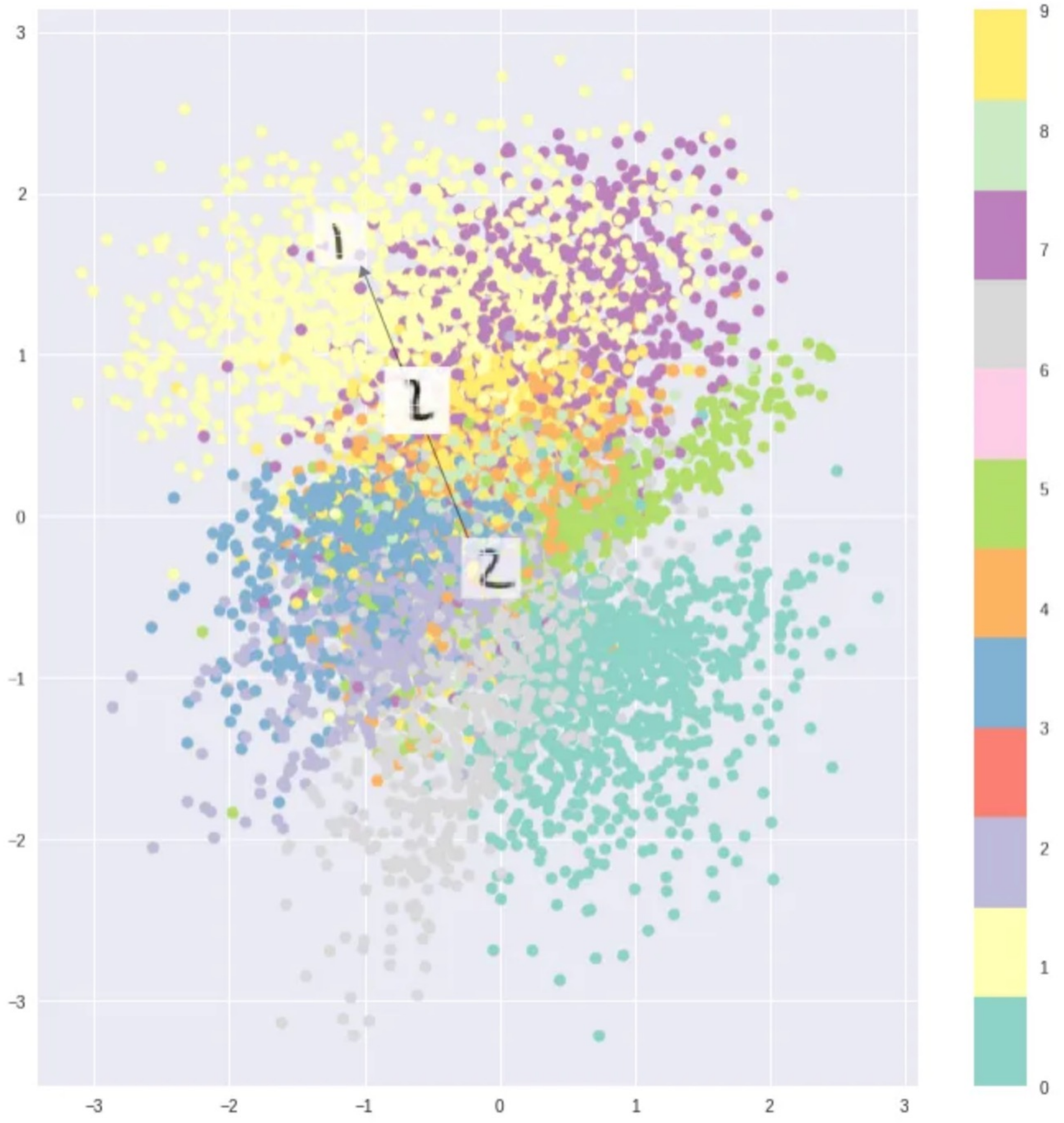
Intuitively Understanding Variational Autoencoders by Irhum Shafkat



Optimizing using pure KL divergence loss

Optimizing KL divergence only

Intuitively Understanding Variational Autoencoders by Irhum Shafkat



Optimizing using both reconstruction loss and KL divergence loss

Full VAE loss

Intuitively Understanding Variational Autoencoders by Irhum Shafkat

Biology Example: popvae

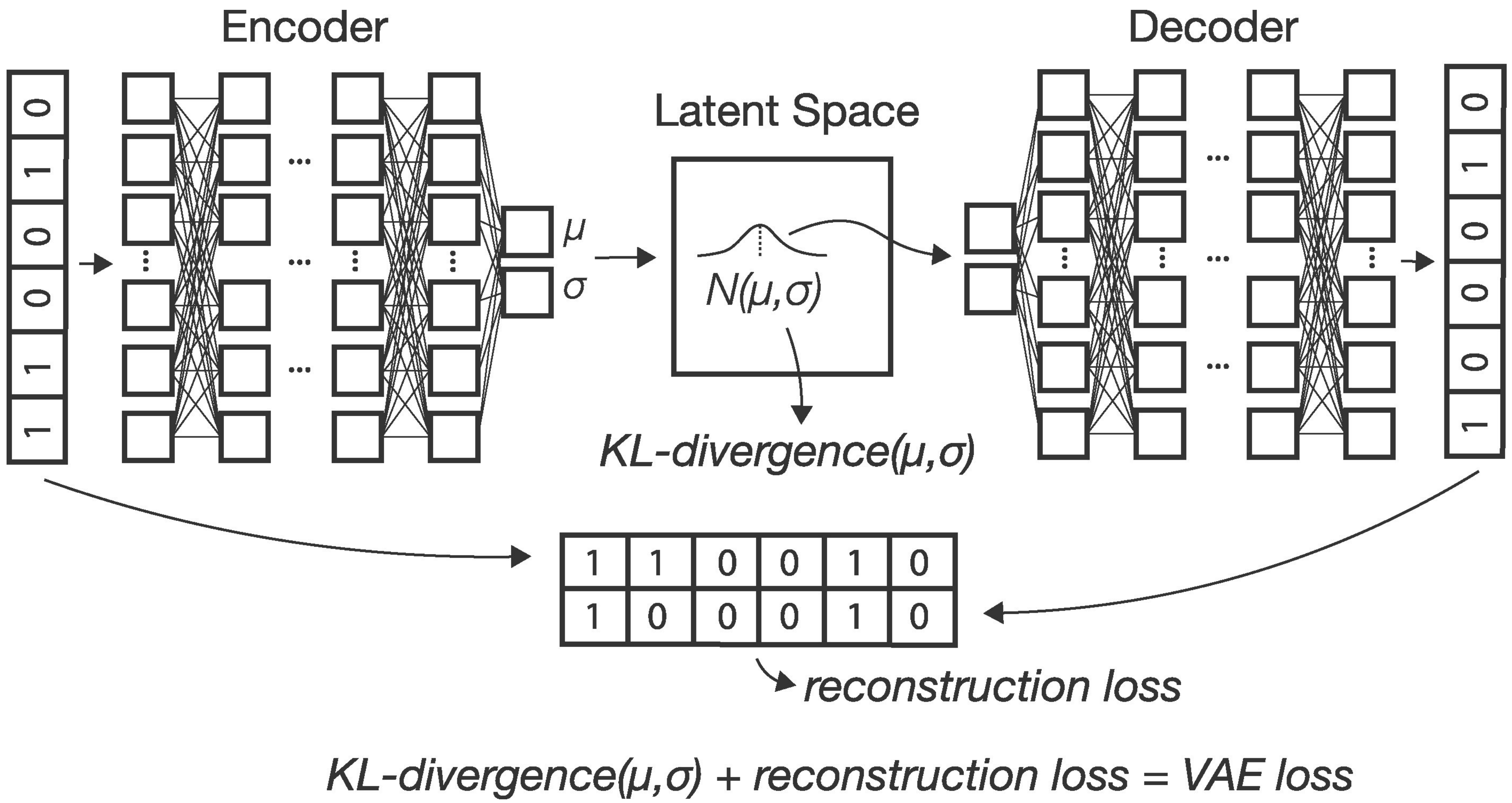


Figure 1 A schematic of the VAE architecture.

G3 Genes|Genomes|Genetics, Volume 11, Issue 1, January 2021, jkaa036, https://doi.org/10.1093/g3journal/jkaa036

Biology Example: popvae

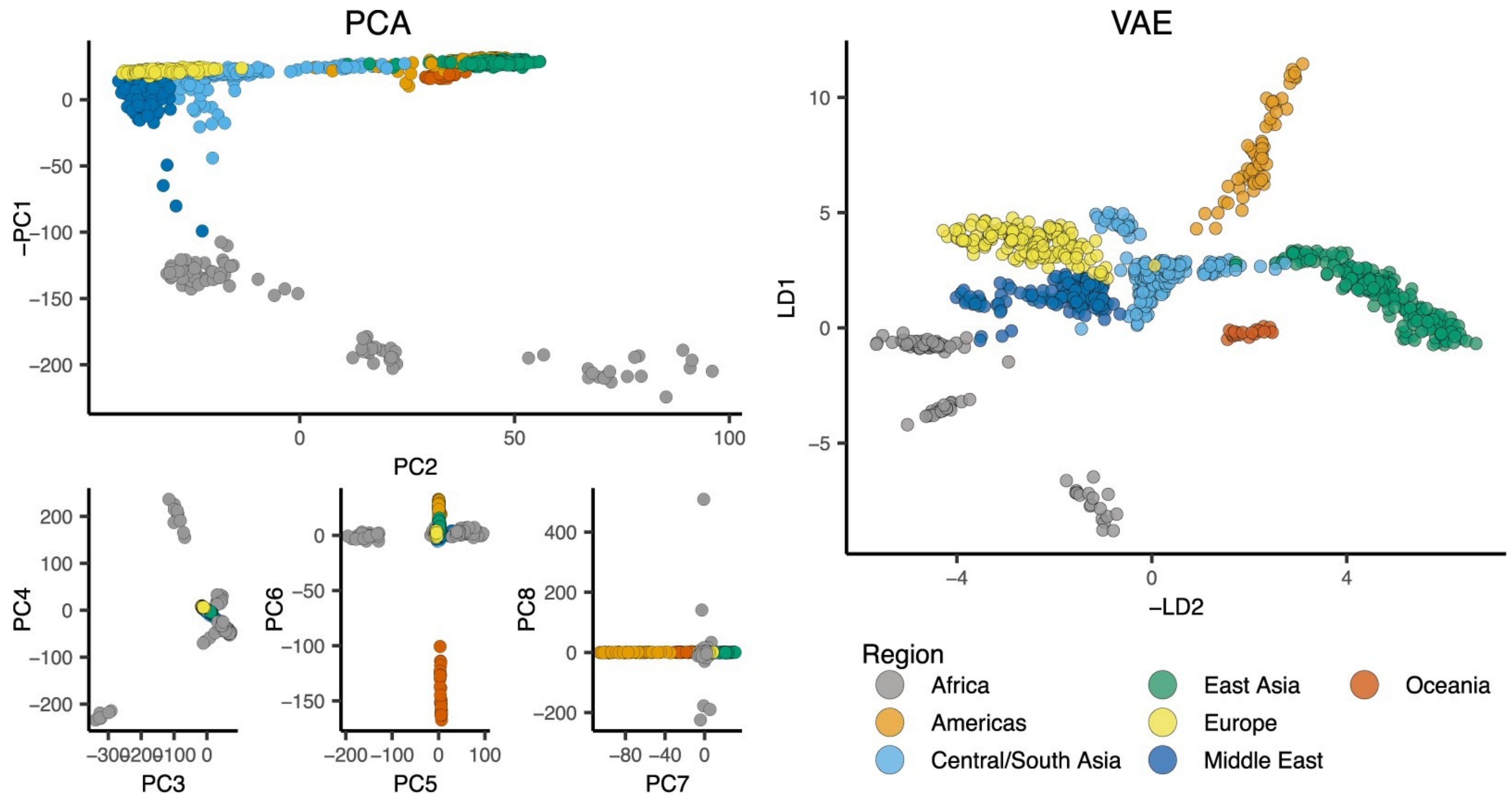


Figure 2 PCA axes 1–8 (left) and popvae run at default settings (right) for 100,000 random SNPs from chromosome 1

G3 Genes|Genomes|Genetics, Volume 11, Issue 1, January 2021, jkaa036, https://doi.org/10.1093/g3journal/jkaa036

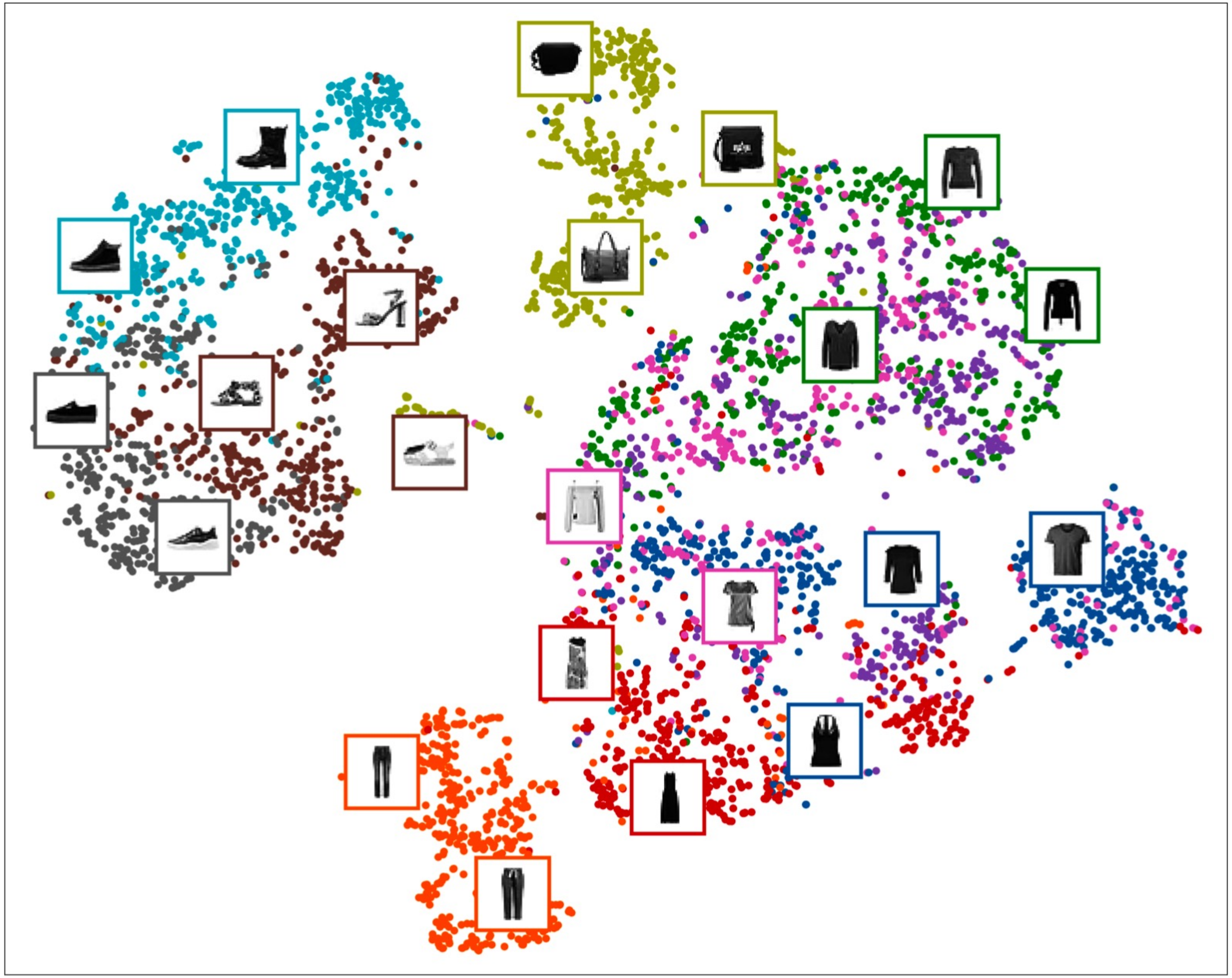
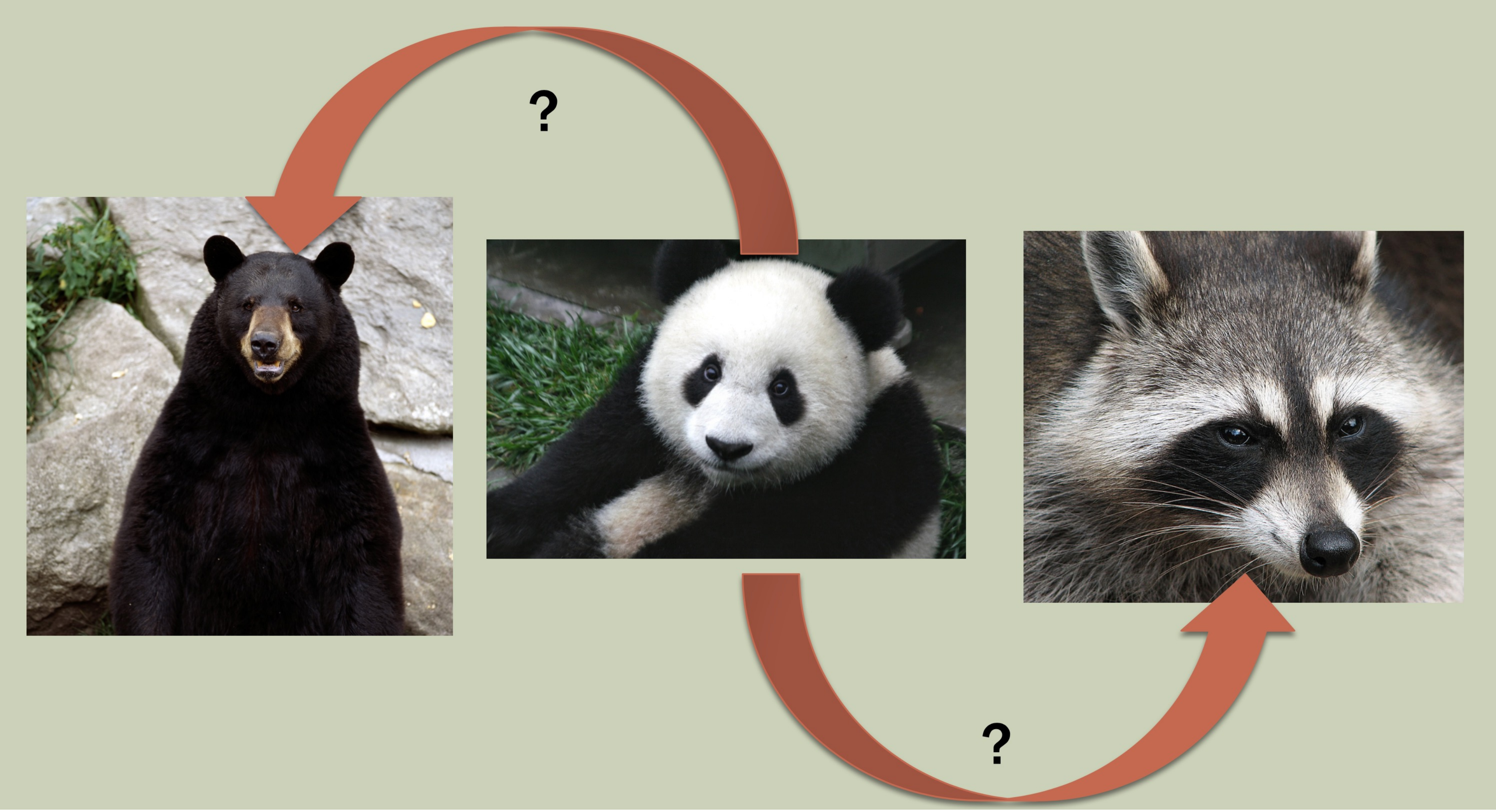


Figure 17-5. Fashion MNIST visualization using an autoencoder followed by t-SNE

Outline for April 18

- Introduction to unsupervised learning
- Review K-means (from CS260)
- Gaussian Mixture Models (GMMs)
- Review PCA (from CS260)
- Autoencoders
- Variational Autoencoders (VAEs)
- Hierarchical clustering (if time)

Are pandas more closely related to bears or raccoons?



Credit: Ameet Soni

UPGMA and Neighbor Joining

- Start with a dissimilarity map between examples (symmetric matrix)
- Say our examples are: A,B,C,D,E

δ	A	B	C	D	E
A	0	1	3	6	6
B		0	2	5	5
C			0	5	5
D				0	2
E					0

Hierarchical clustering example (UPGMA)

University of Southampton
School of Biological Sciences

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

A D B F G C E

Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

University of Southampton
School of Biological Sciences

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

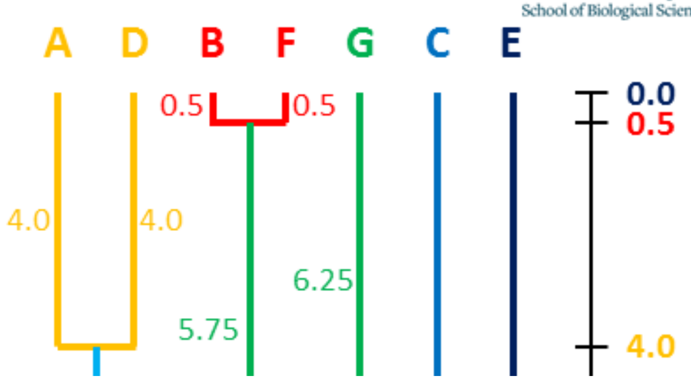


Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

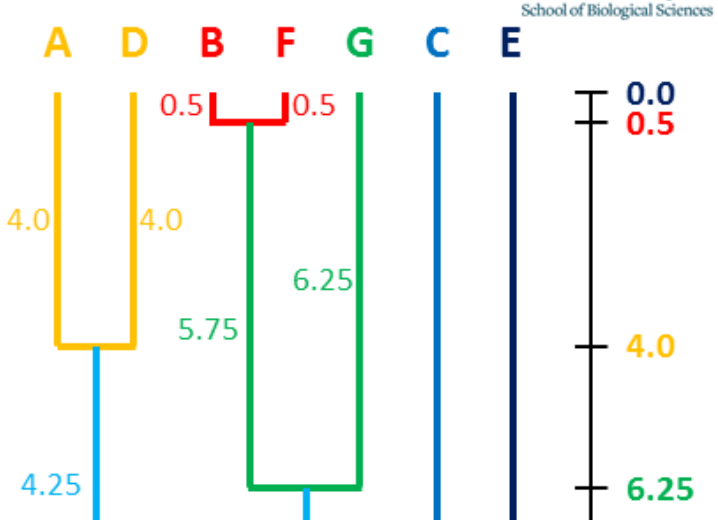


Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

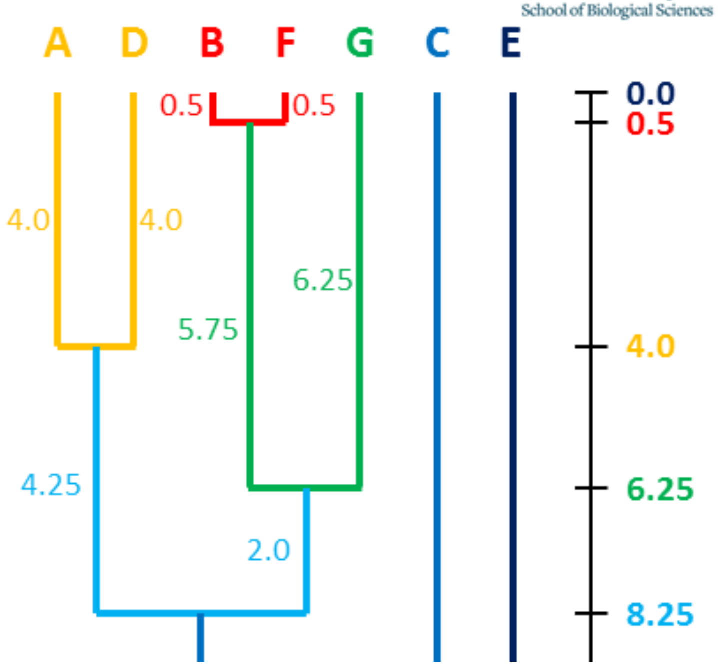
	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00



UNIVERSITY OF Southampton
School of Biological Sciences

Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**

UNIVERSITY OF Southampton
School of Biological Sciences

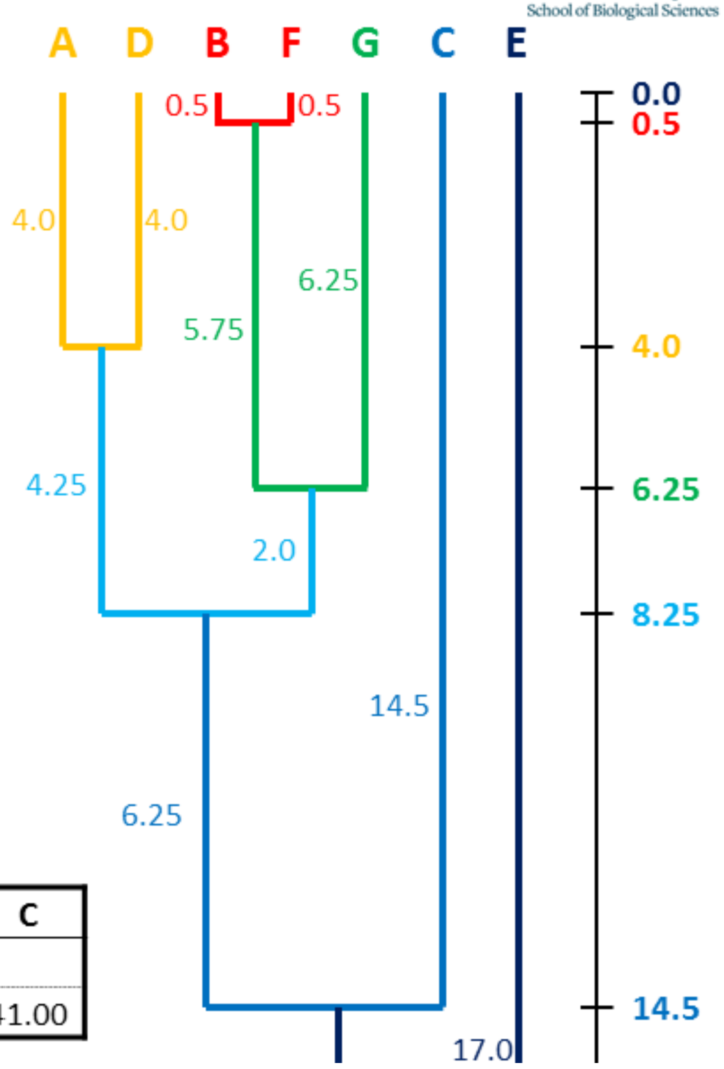


Figure: Dr. Richard Edwards

Hierarchical clustering example (UPGMA)

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**

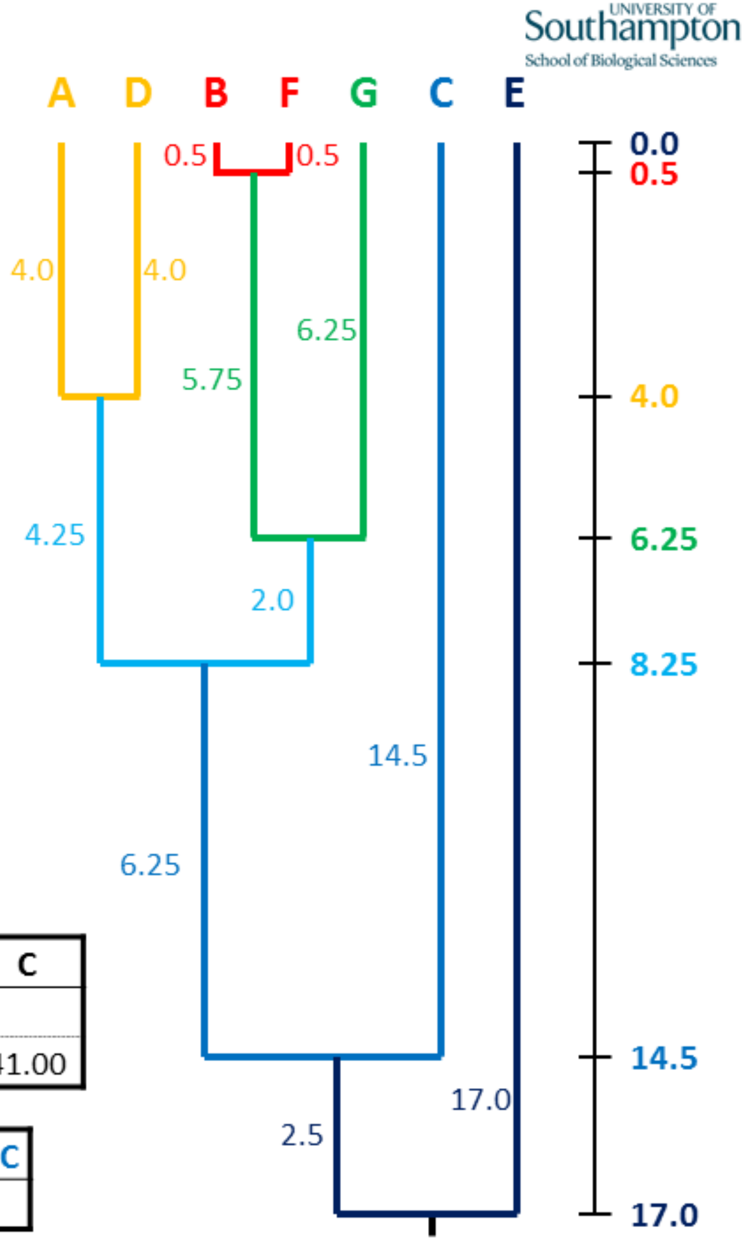
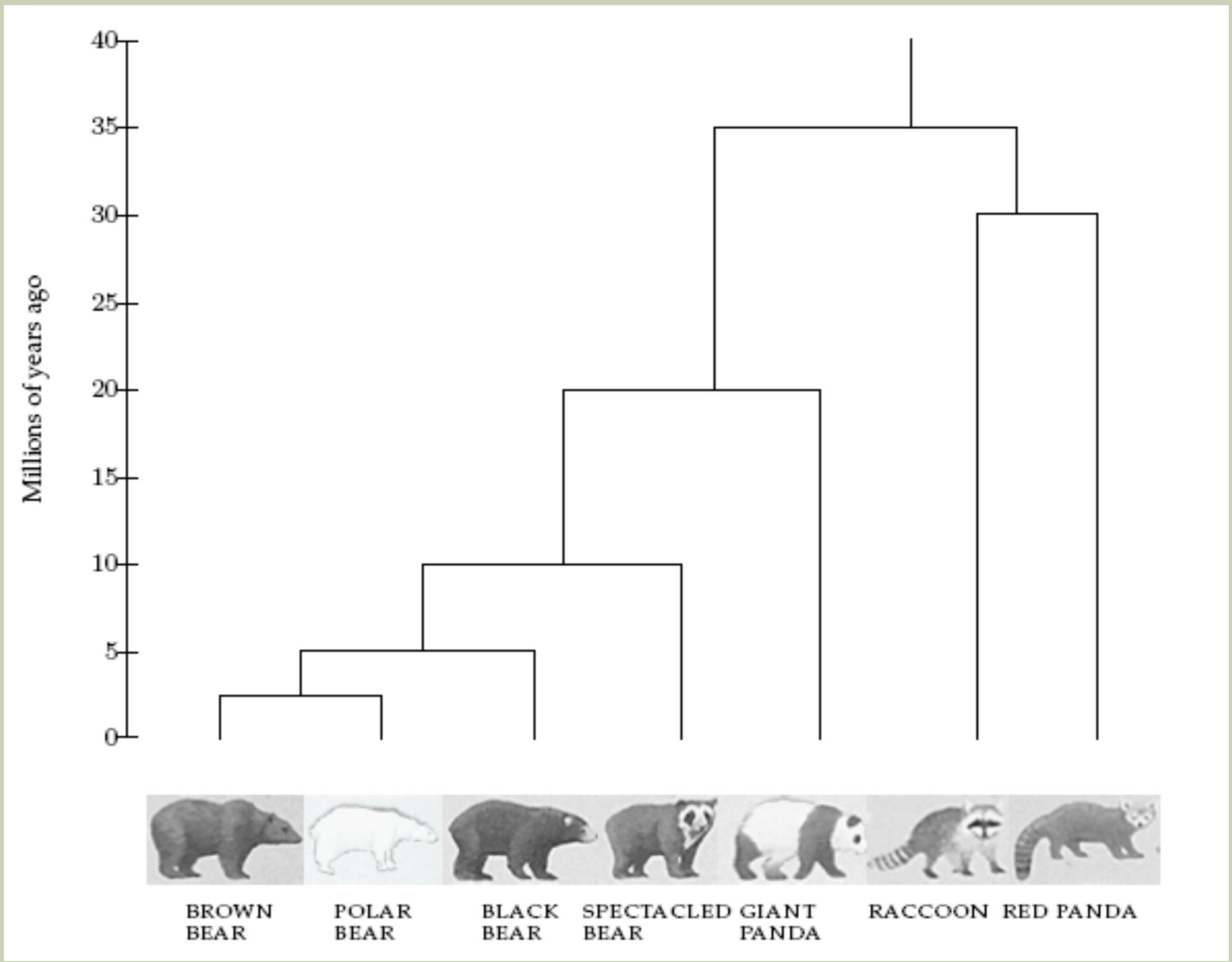


Figure: Dr. Richard Edwards

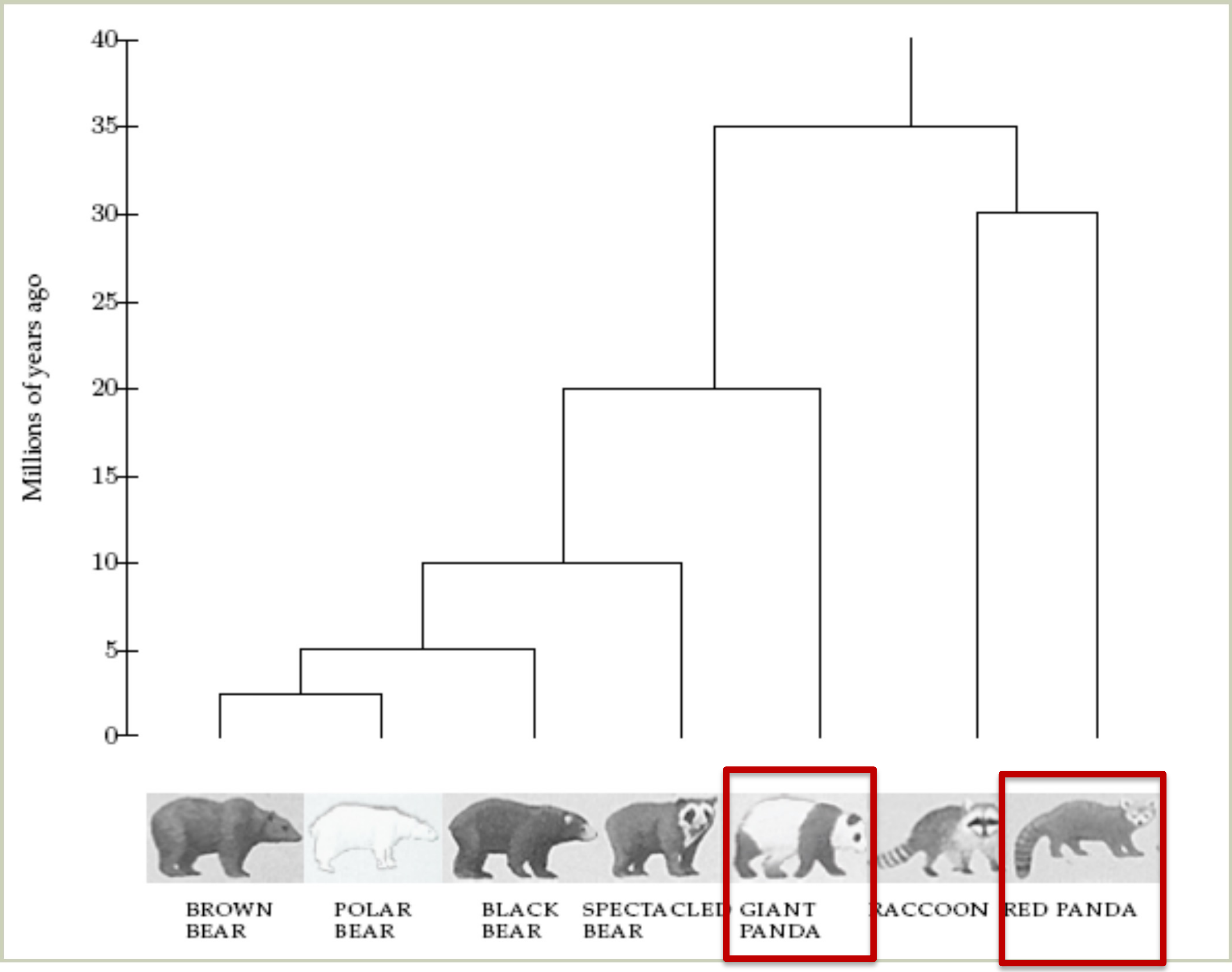
Back to the pandas….

Back to the pandas....



Credit: Ameet Soni

Back to the pandas….



Credit: Ameet Soni