RICE | D2K
Data to Knowledge Lab

Welcome to DSCI 101

Introduction to Data Science

Week 13 Recap

- Decision trees
  - for classification
  - for regression

- Advantages of tree models
  - interpretability
  - non-linear relationship with interactions

- Ensemble models
  - Random Forest

Week 14 Preview

- Unsupervised learning – data driven discovery
  - No data label
  - Not focus on prediction

- Dimensionality reduction – visualize high dimensional data
  - Principal Component Analysis

- Clustering – finding subgroups in data
  - K-means

Why Dimensionality Reduction?

- Dimensionality = the columns of your tabular data

- reduce overfitting
- less noisy data
- more interpretable

- less data storage
- speed up training
- more efficient

- visualize big data
- overall data pattern
- detect outliers

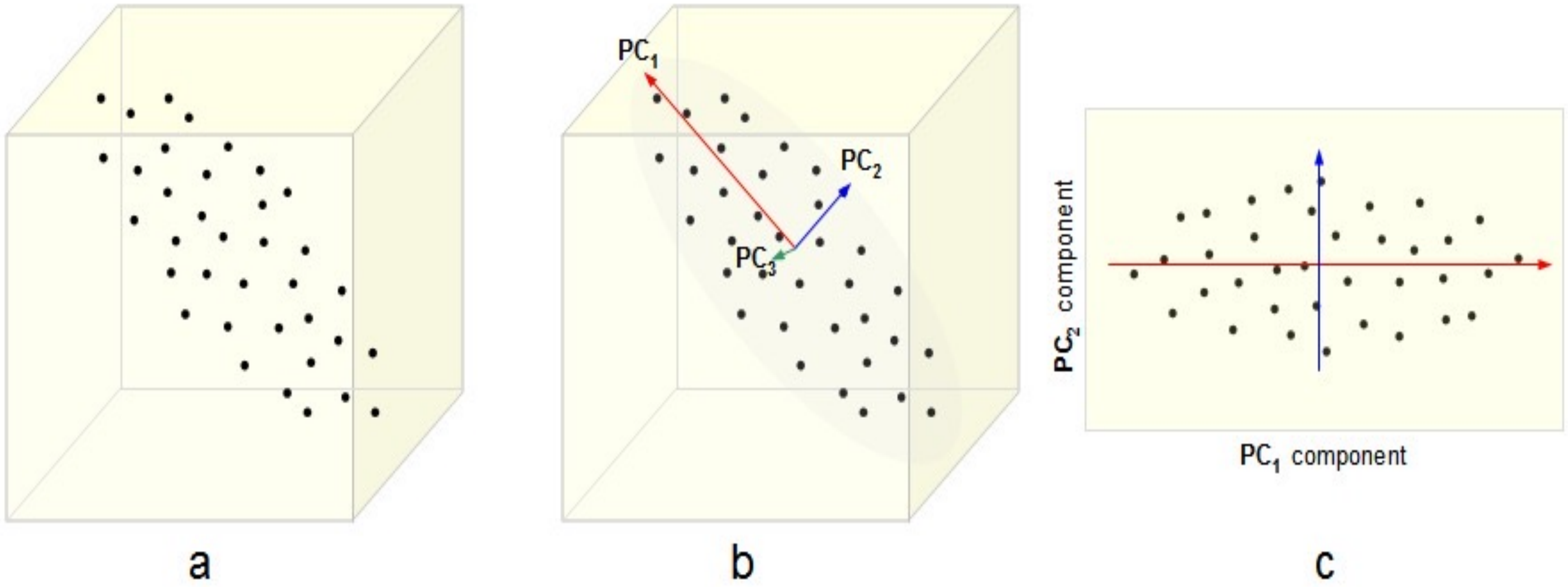
PCA overview

- How to visualize the entire dataset on one scatter plot
  - tabular data with n rows and p columns
  - n observations / samples / data points
  - p variables / features / dimensions

- Human eyes are not good at “seeing” > 3 variables at once
  - a scatter plot only shows two variables
  - could just plot the “most important” two variables, but can we do better?

- Principal Component Analysis (PCA)
  - the “best” low-dimensional representation of data

PCA illustration

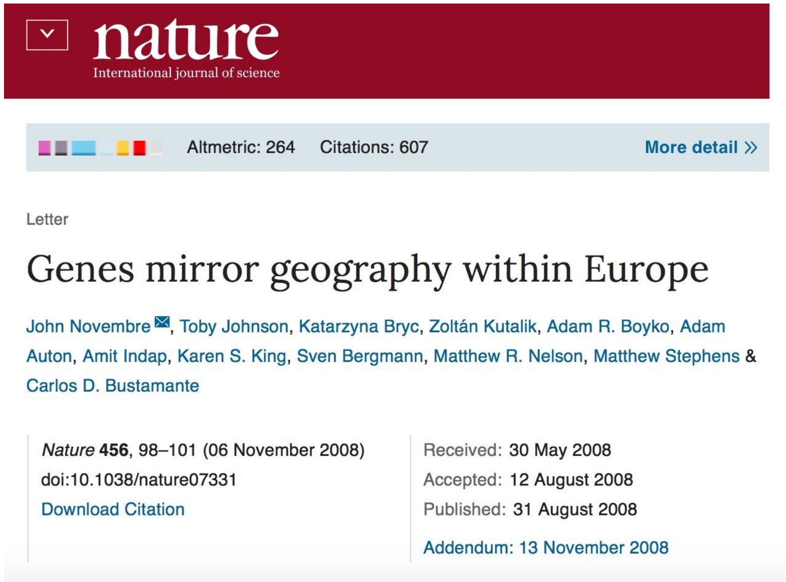


What are Principal Components?

- Some intuition:
  - Each PC is a linear combination (weighted average) of all the variables
  - Each PC is like a “created” variable, combines a little bit of every variable
  - PCs are ordered in decreasing importance: 1st PC > 2nd PC >…
  - PCs are no longer interpretable: they are not “real” variables

- PCA as data visualization tool:
  - Plot 1st and 2nd PCs on a scatter plot
  - The “best” scatter plot that captures the “most info” of data
  - Great for discover patterns in data

Delight



nature
International journal of science

Altmetric: 264 Citations: 607 More detail »

Letter

Genes mirror geography within Europe

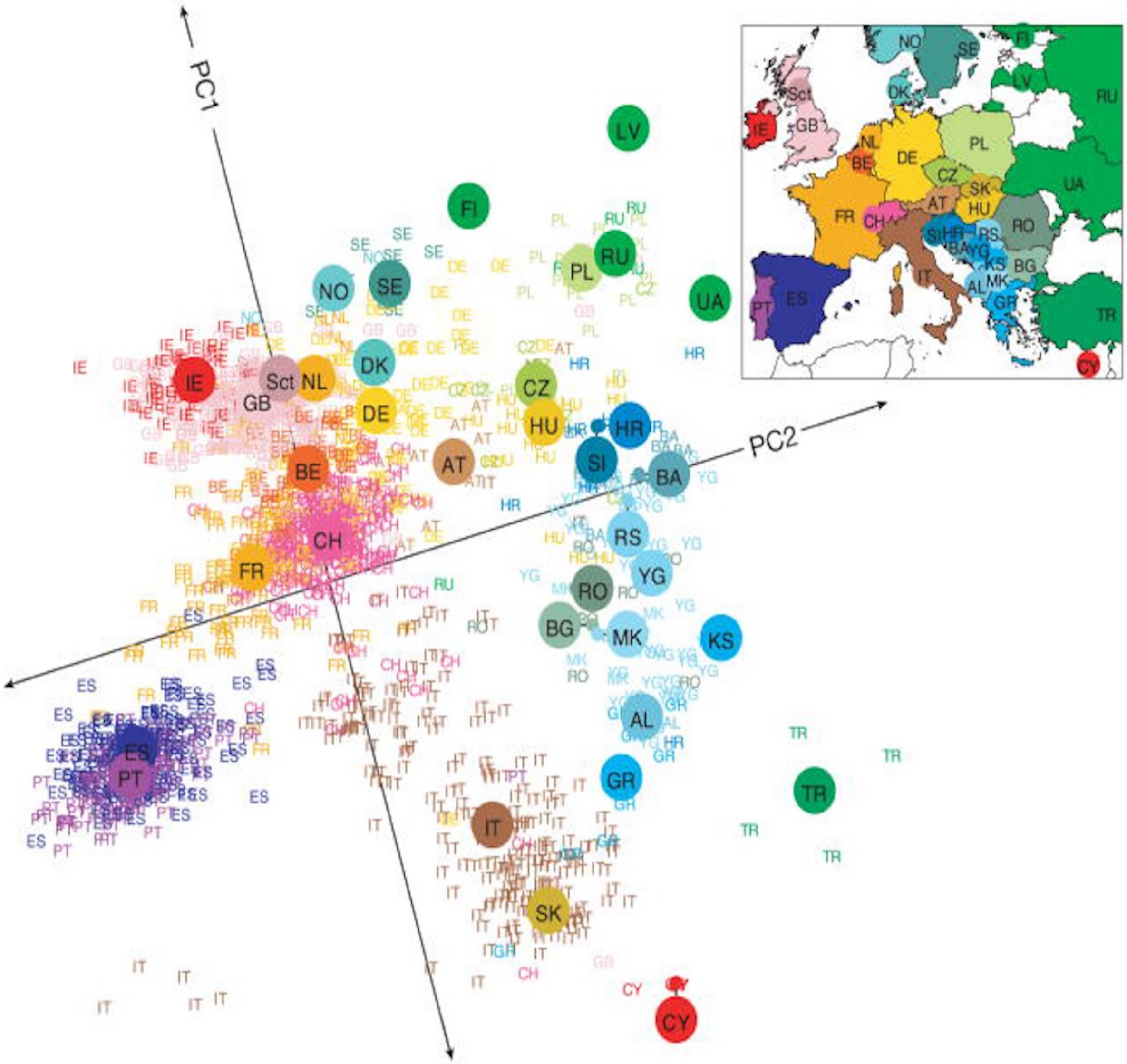
John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101 (06 November 2008)
doi:10.1038/nature07331
Download Citation

Received: 30 May 2008
Accepted: 12 August 2008
Published: 31 August 2008

Addendum: 13 November 2008

- Data matrix:
  - 1,387 rows (people from Europe)
  - 197,145 columns (gene measurement)
- PCA plot:
  - reduces to 2 dimensions!
  - reveal insightful data pattern



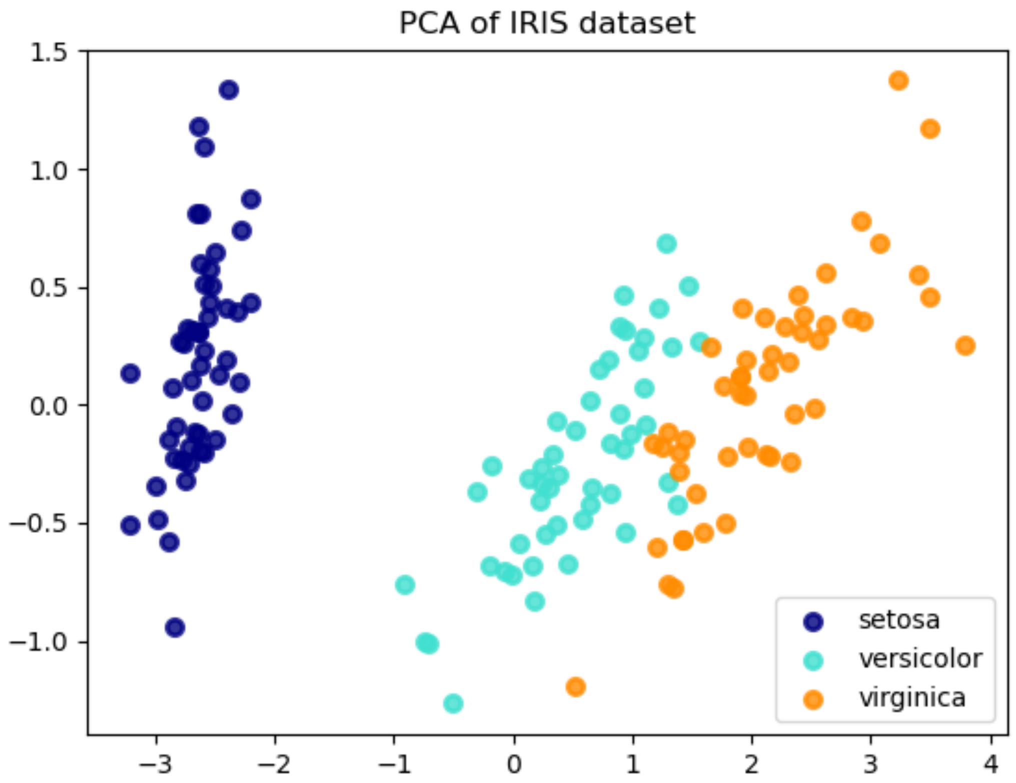
PCA in application

- Only applies to numerical variables!
  - watch out for categorical variables coded as numbers

- Can have at most p PCs for p-dimensional data!
  - mostly we just look at 1st, 2nd and maybe the 3nd PCs

- Standardize columns if variables are not comparable in scale
  - otherwise variables in large scale will dominate

Use PCA to discover data pattern

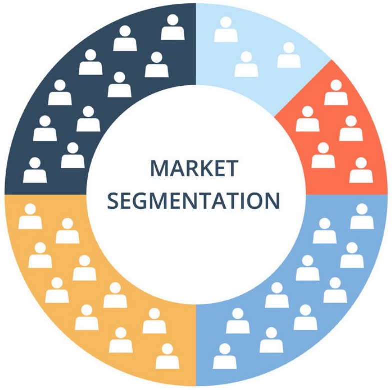


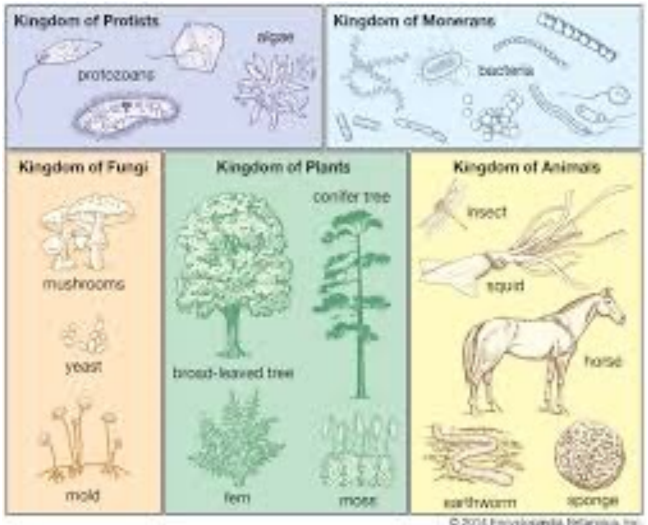
- A famous dataset of iris plants
- 4 dimensional: petal length/width, sepal length/width
- PCA scatter plot can completely separate the 3 species of iris, while any scatter plot of two variables can not

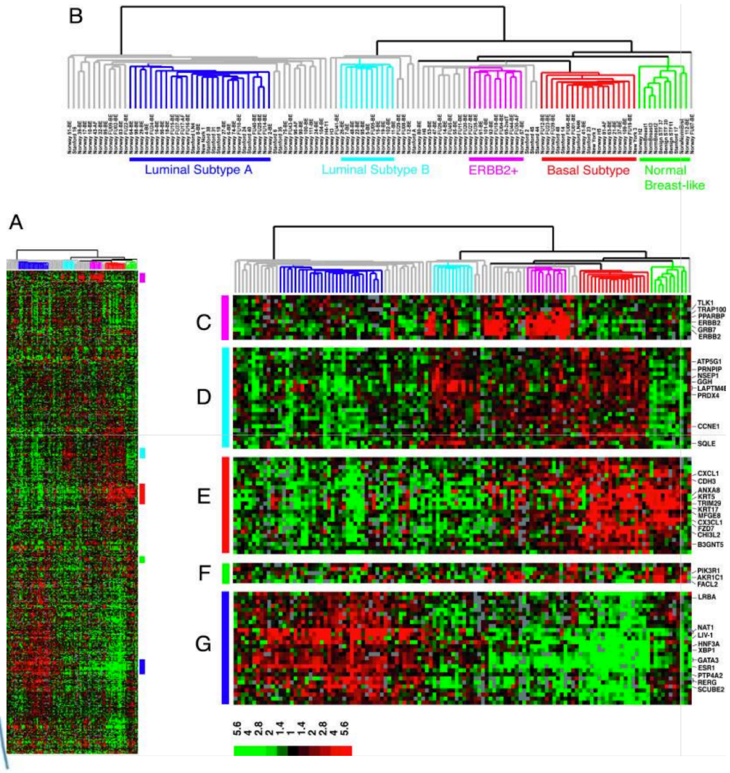
Clustering

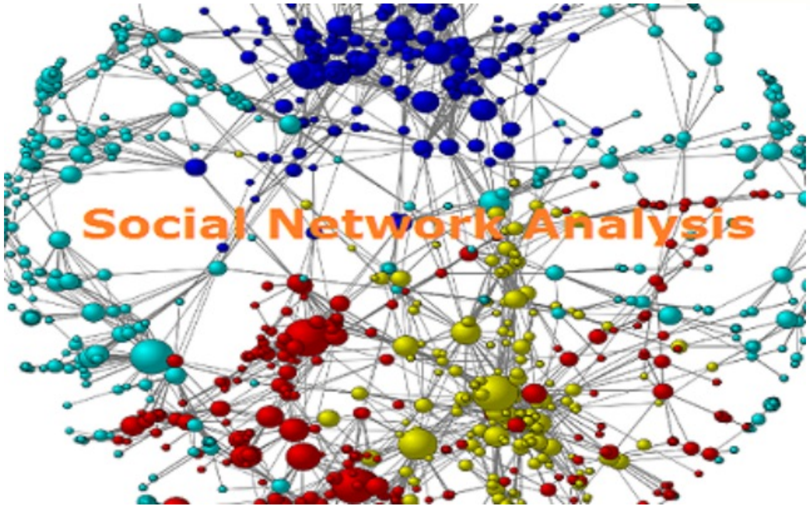
- Finding subgroups (clusters) in a dataset.
- The observations (rows) within each cluster are similar.
- How do we define two or more observations to be similar or different?
  - use some distance metric
  - often domain-specific

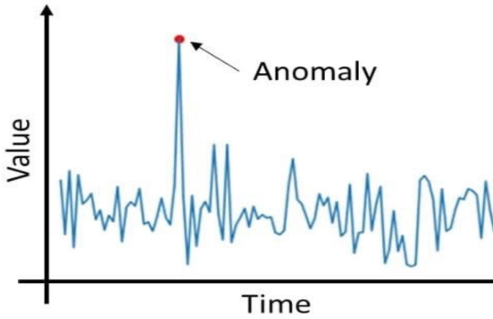
Applications





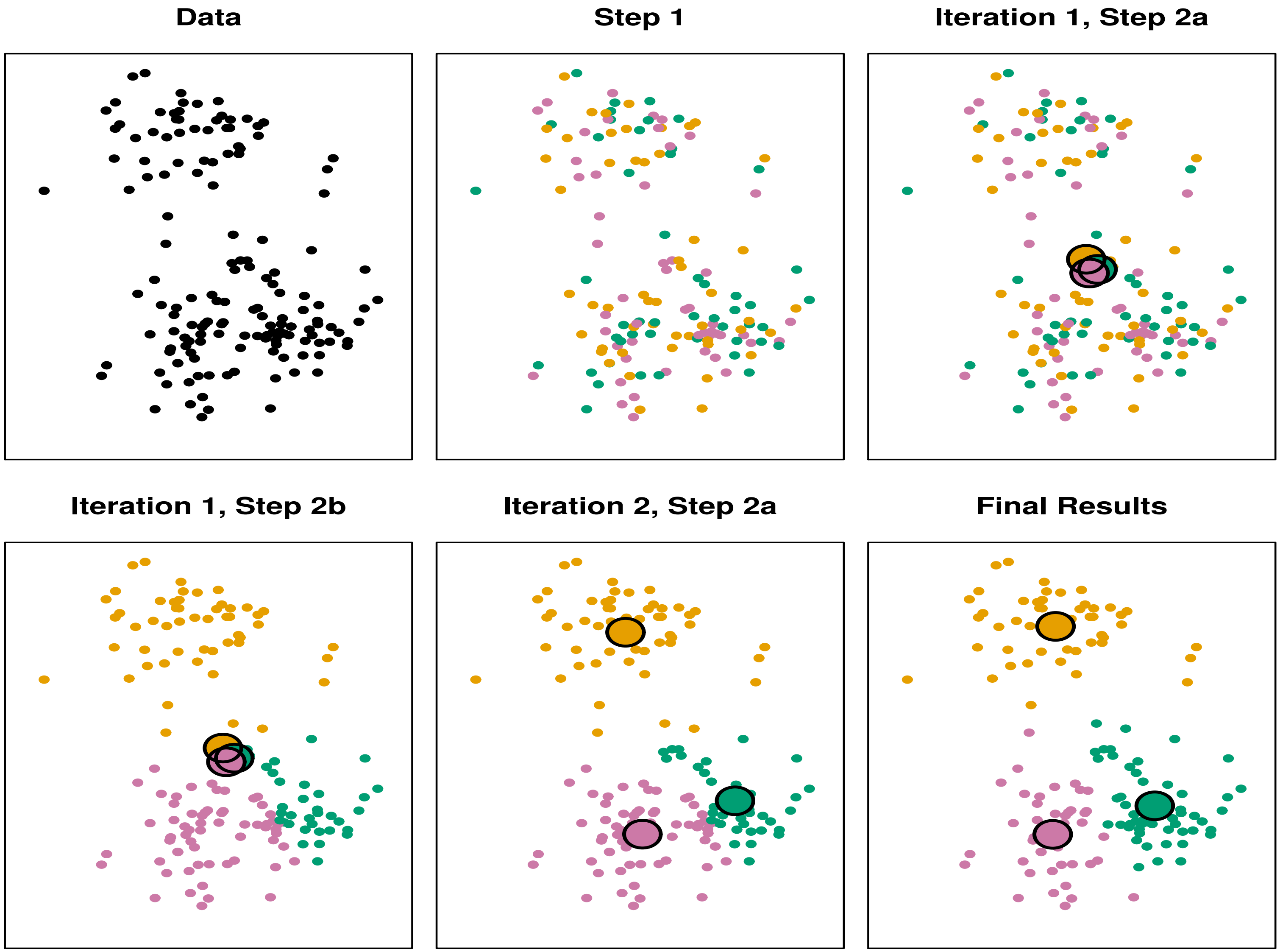






K-means

- Step 1: randomly assign cluster membership
- Step 2: iteratively update cluster centroids and membership
    - 2a: find cluster centroids
    - 2b: reassign cluster membership to the closest centroid
    - repeat until converge



Practical issues in clustering

- Should the features be standardized?
  - for each column, subtract its mean and divide by its standard deviation
  - standardized features should have mean 0 and SD 1
- How many clusters to choose???
  - can we take a peak of the entire data?
- Robustness: how to account for noise in observations?
- Can we cluster features instead of observations?
  - or cluster both!

Congratulations!
You did it!!!