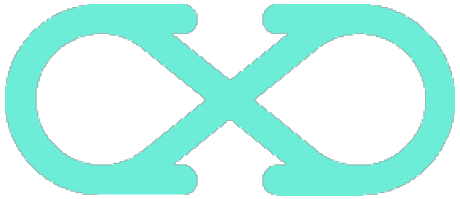


Analyzing Weather Data with Python

Greg Filla, Associate Offering Manager
Data Science Experience

BIG DATA UNIVERSITY

About me

- Working at IBM for ~2 years - Product Management, Analytics
- Started career as analyst…started coding
- I Like Python - Pandas, Sklearn, Jupyter, Flask, PySpark, etc.

- Twitter: @gdfilla
- Github: https://github.com/gfilla

BIG DATA UNIVERSITY

Introducing the Data Science Experience

Learn

Built-in learning to get started or go the distance with advanced tutorials

Create

The best of open source and IBM value-add to create state-of-the-art data products

Collaborate

Community and social features that provide meaningful collaboration

http://datascience.ibm.com

BIG DATA UNIVERSITY

Core Attributes of the Data Science Experience

IBM Data Science Experience

Community

- Find tutorials and datasets
- Connect with Data Scientists
- Ask questions
- Read articles and papers
- Fork and share projects

Open Source

- Code in Scala/Python/R
- Jupyter Notebooks
- RStudio IDE and Shiny apps
- Apache Spark
- Your favorite libraries

IBM Added Value

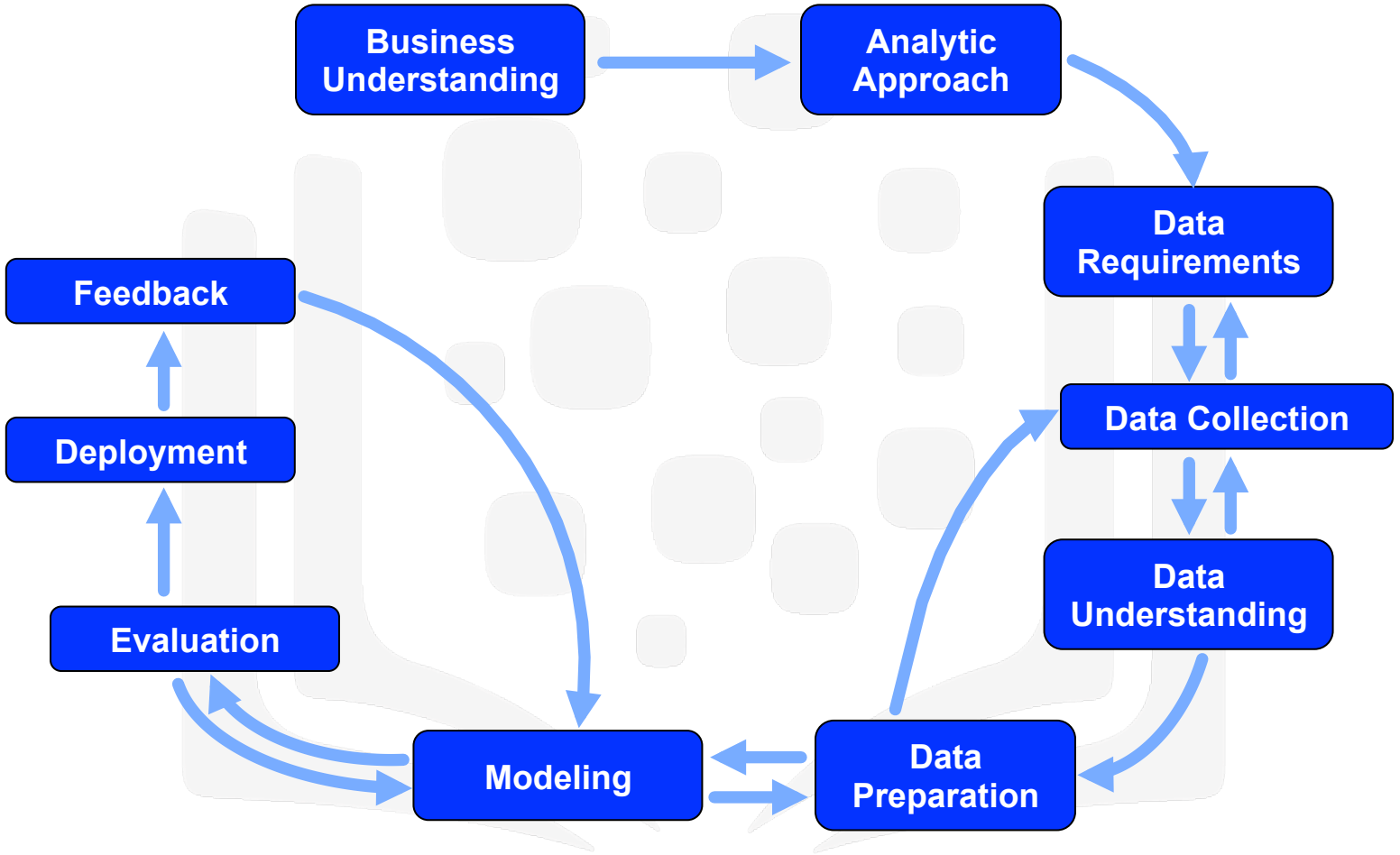
- Data Shaping/Pipeline UI *
- Auto-data preparation and modeling*
- Advanced Visualizations*
- Model management and deployment*
- Documented Model APIs*
- Spark as a Service

Powered by Watson Data Platform

BIG DATA UNIVERSITY

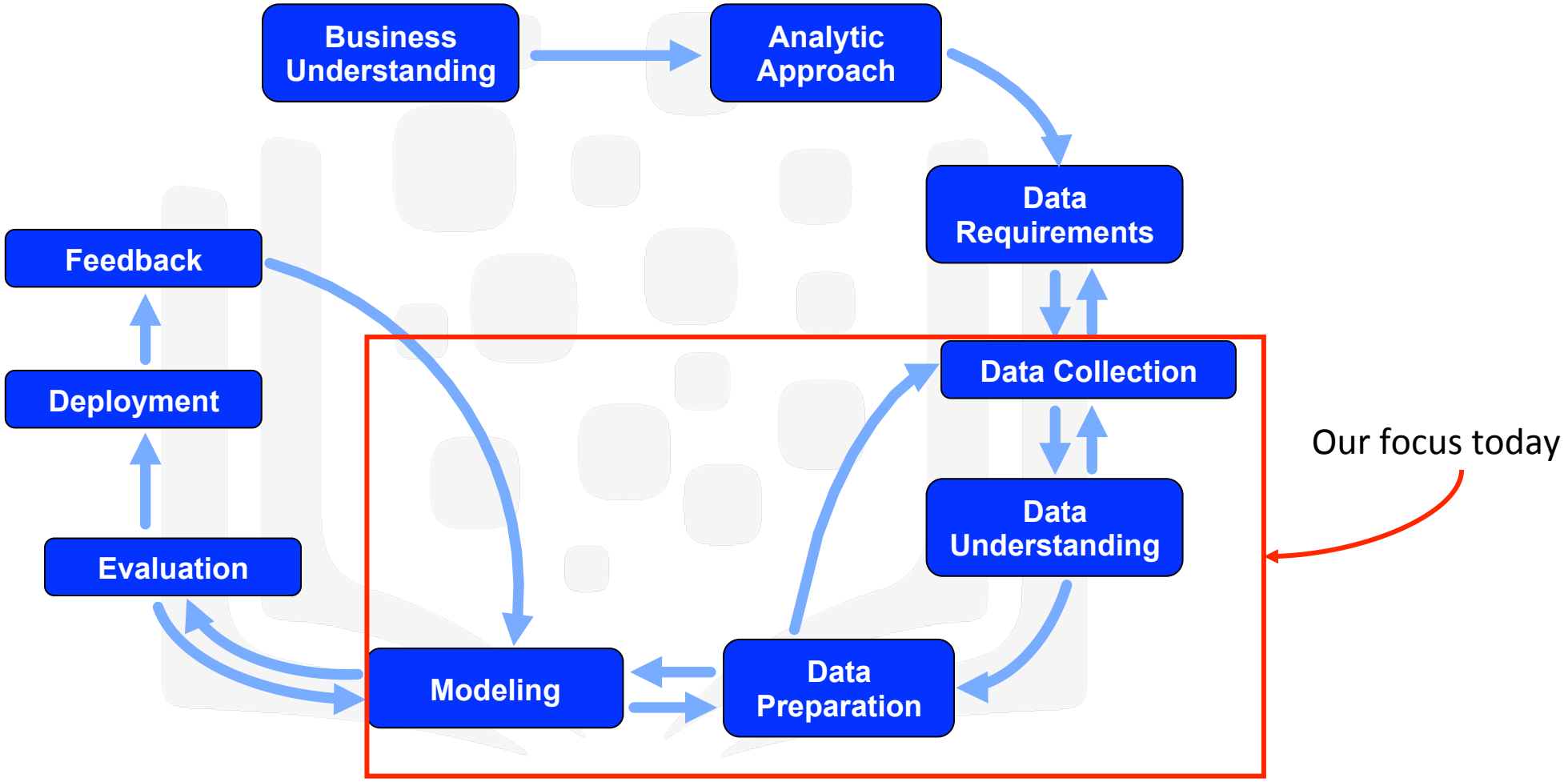
* DSX product roadmap items

Data Science Methodology Diagram



BIG DATA UNIVERSITY

Data Science Methodology Diagram



BIG DATA UNIVERSITY

Machine Learning

Supervised vs. Unsupervised

BIG DATA UNIVERSITY

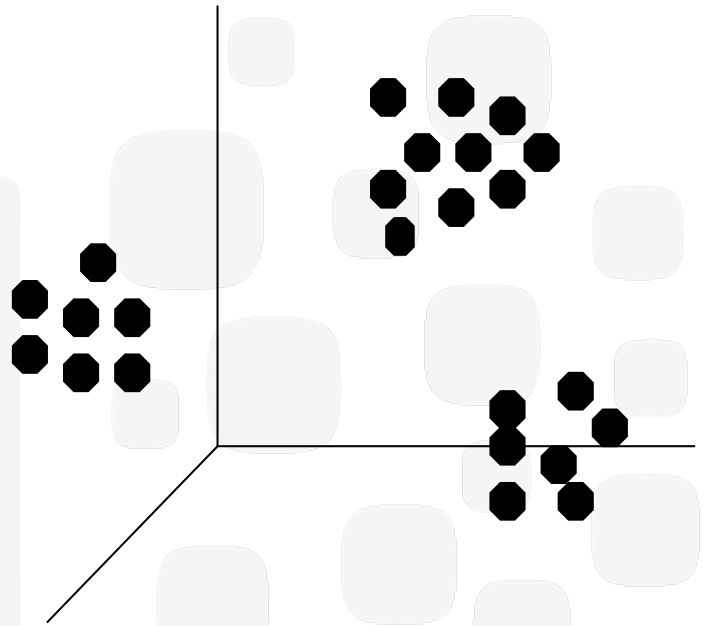
Toolbelt of the Data Scientist

- Linear regression (S)
- Logistic regression (S)
- Decision Trees (S)
- **Clustering (U)**
- Principal component analysis (U)
- Text analysis (S/U)
- SVM/SVR (S)
- Neural networks (S/U)
- Recommender Systems (S)

S == Supervised Learning Technique
U == Unsupersved Learning Technique
S/U == Can be combination of both

BIG DATA UNIVERSITY

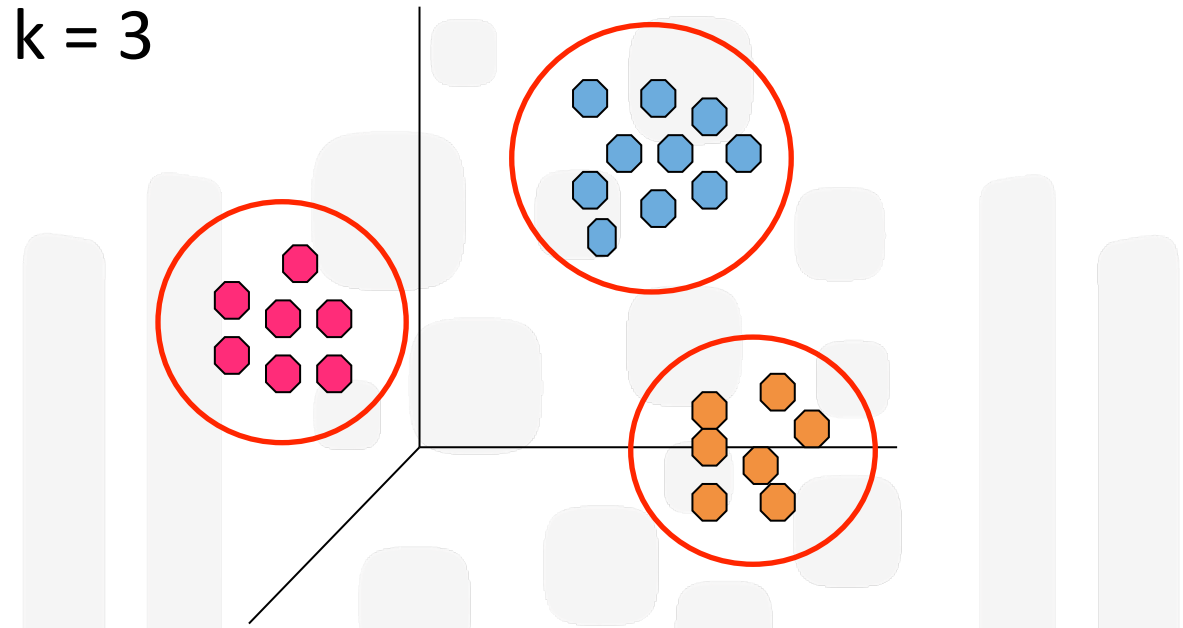
Which weather stations are similar to each other based on weather observations?



Group similar stations together into ***k*** number of clusters.

BIG DATA UNIVERSITY

Which weather stations are similar to each other based on weather observations?

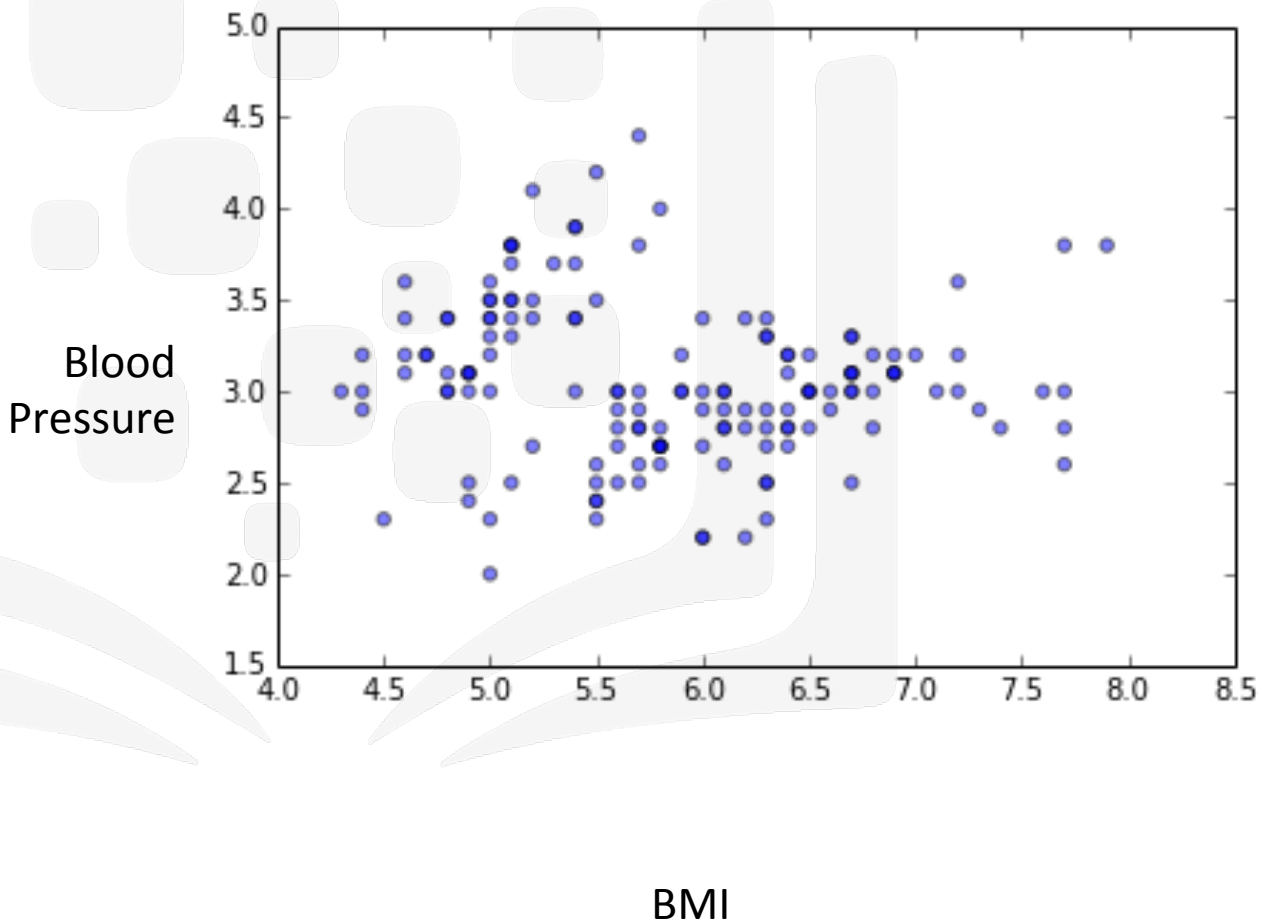


Group similar stations together into ***k*** number of clusters.

K-means Clustering

1) Load your data

	A	B
1	weight	BP
2	5.1	3.5
3	4.9	3
4	4.7	3.2
5	4.6	3.1
6	5	3.6
7	5.4	3.9
8	4.6	3.4
9	5	3.4
10	4.4	2.9
11	4.9	3.1
12	5.4	3.7
13	4.8	3.4

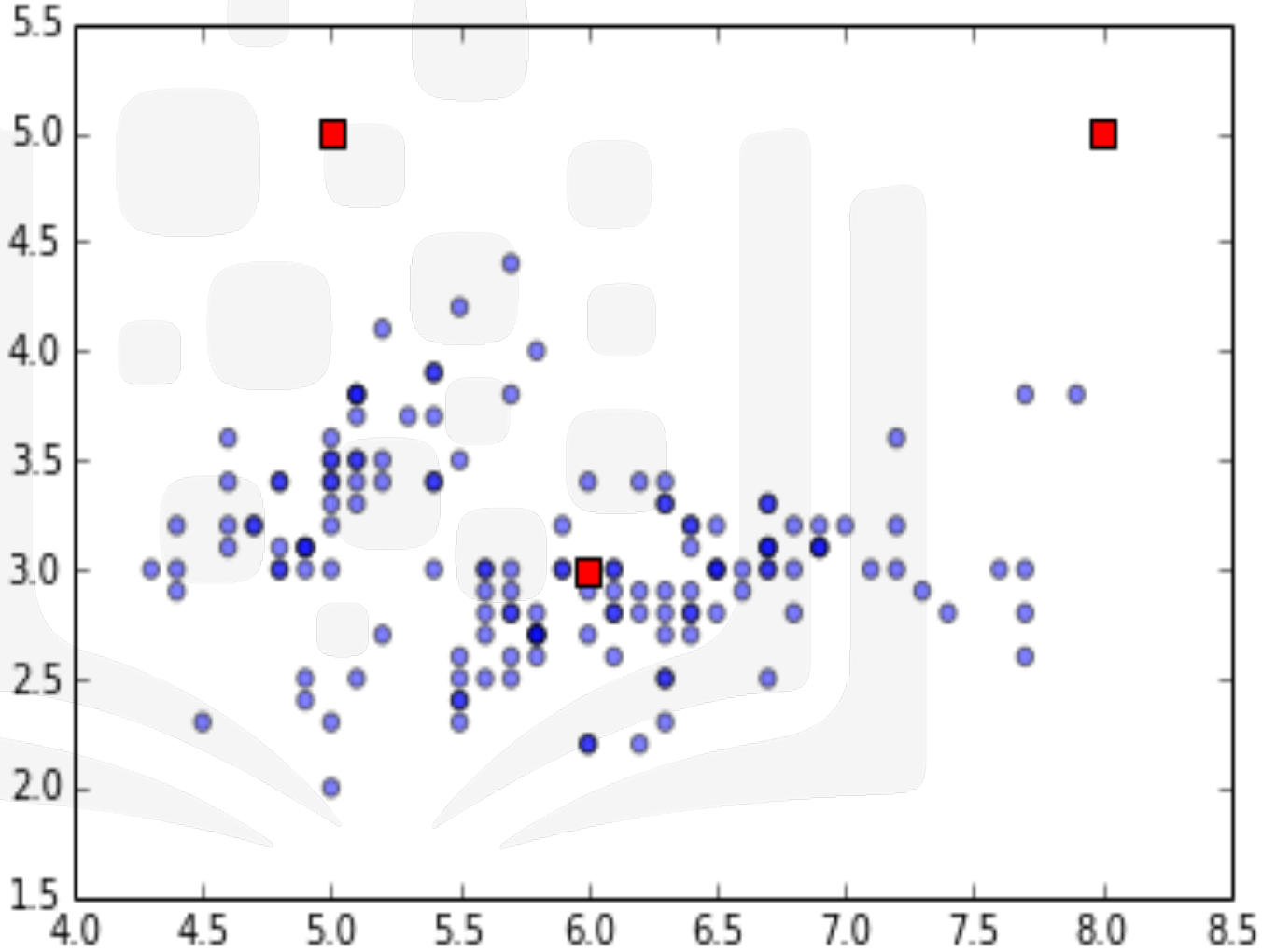


BIG DATA UNIVERSITY

K-means Clustering

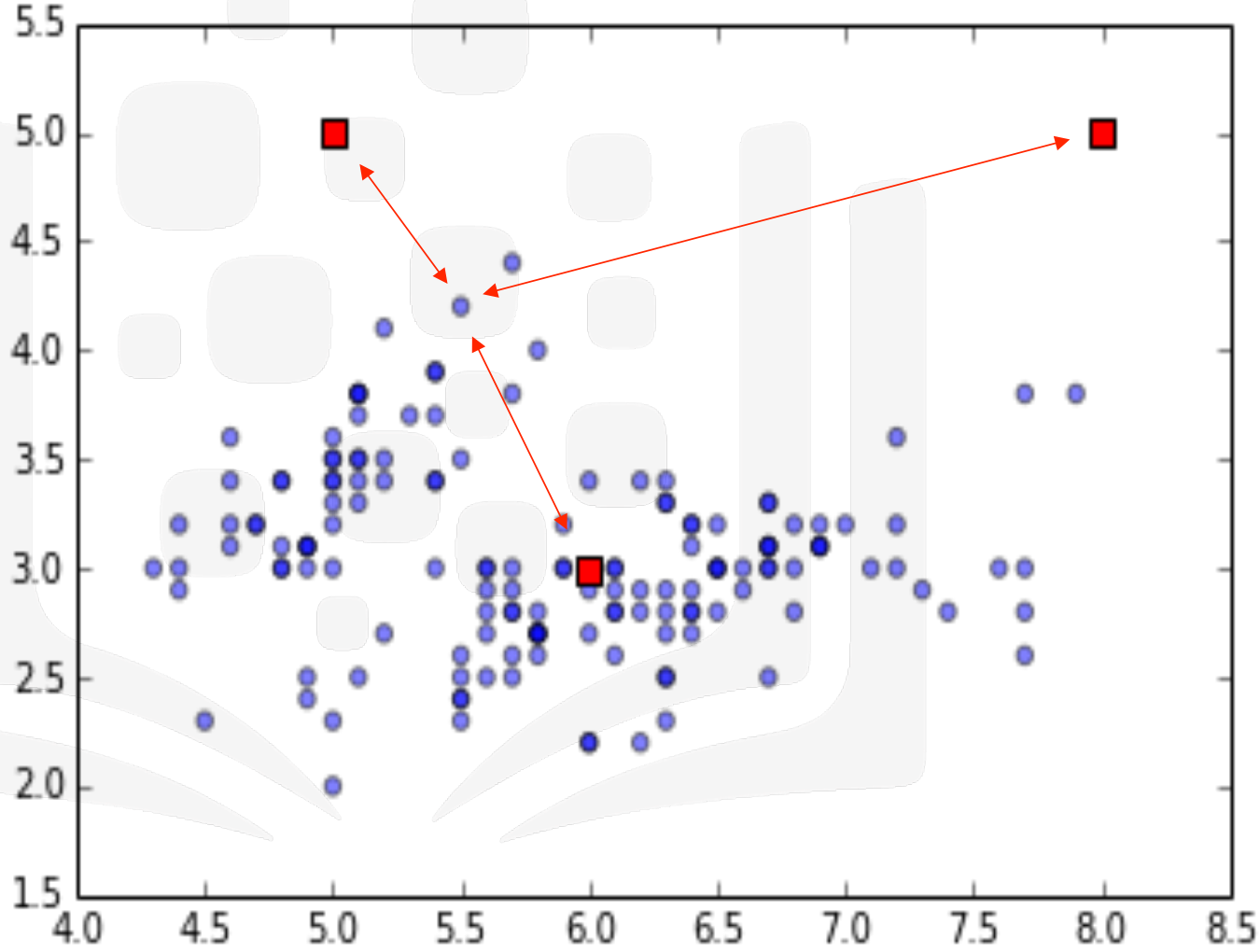
2) Initialize k=3 centroids randomly

[[8., 5.],
[5., 5.],
[6., 3.]]



K-means Clustering

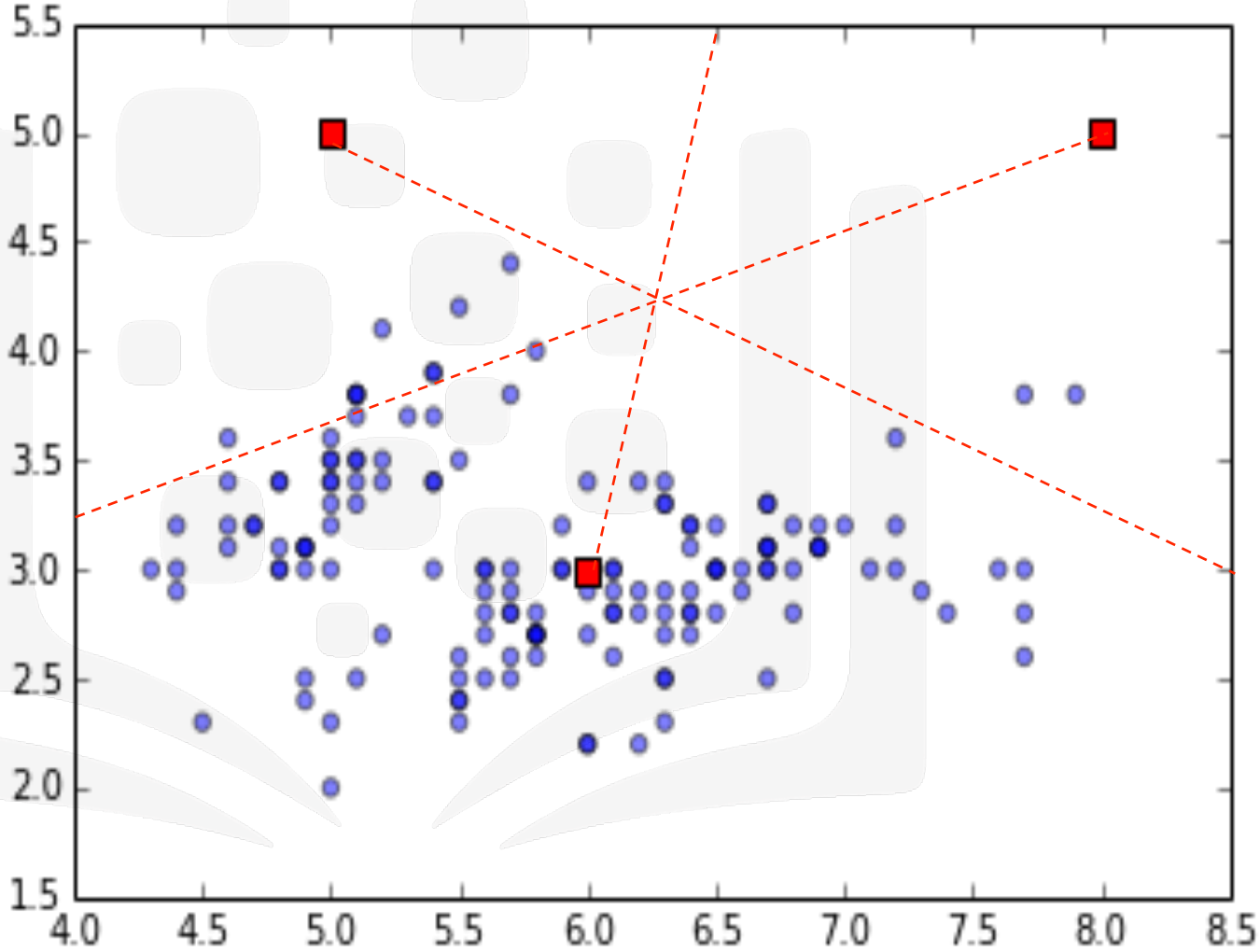
3) Calculate distance of all points from centroids



BIG DATA UNIVERSITY

K-means Clustering

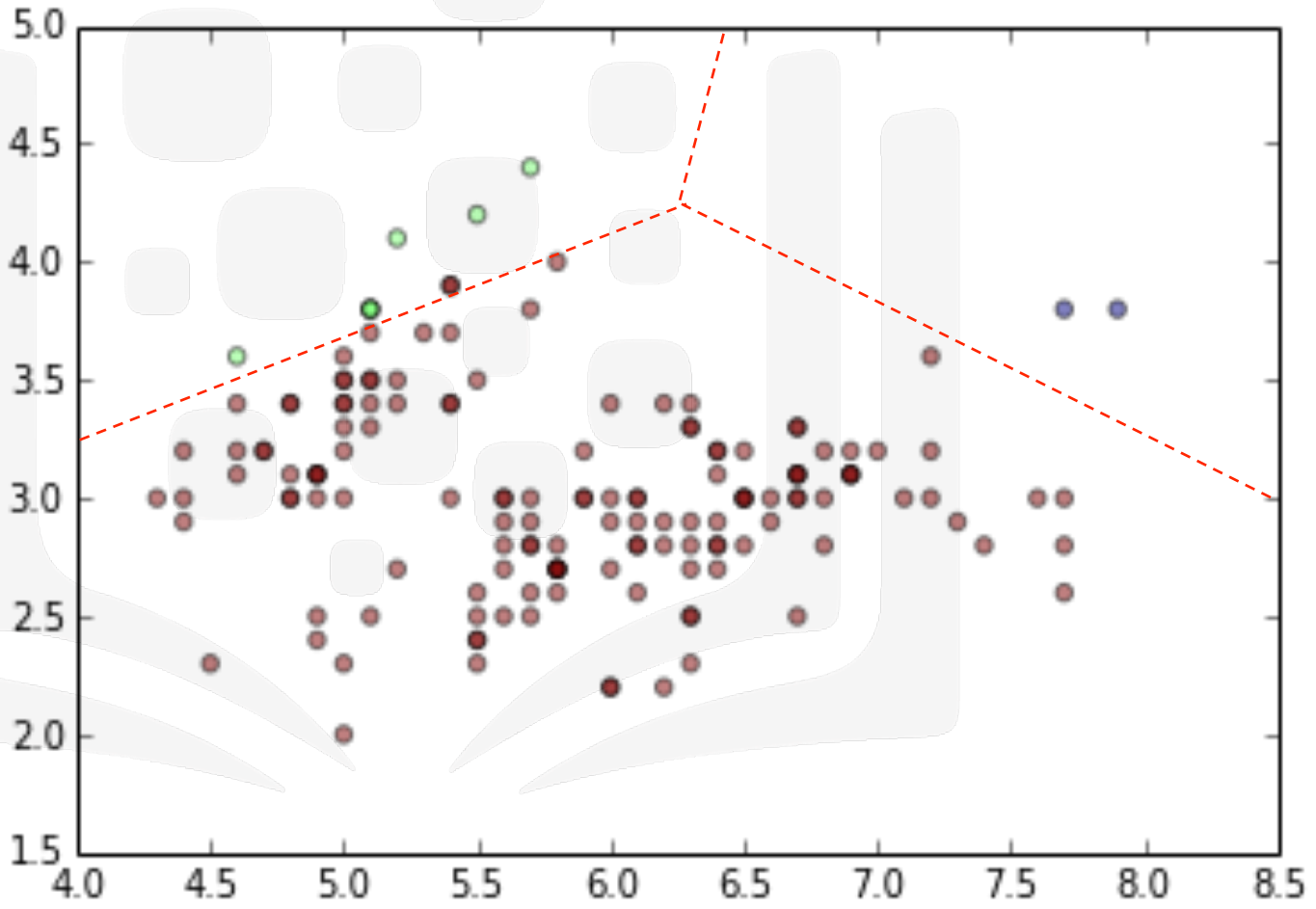
4) Find the closest center to each data point



BIG DATA UNIVERSITY

K-means Clustering

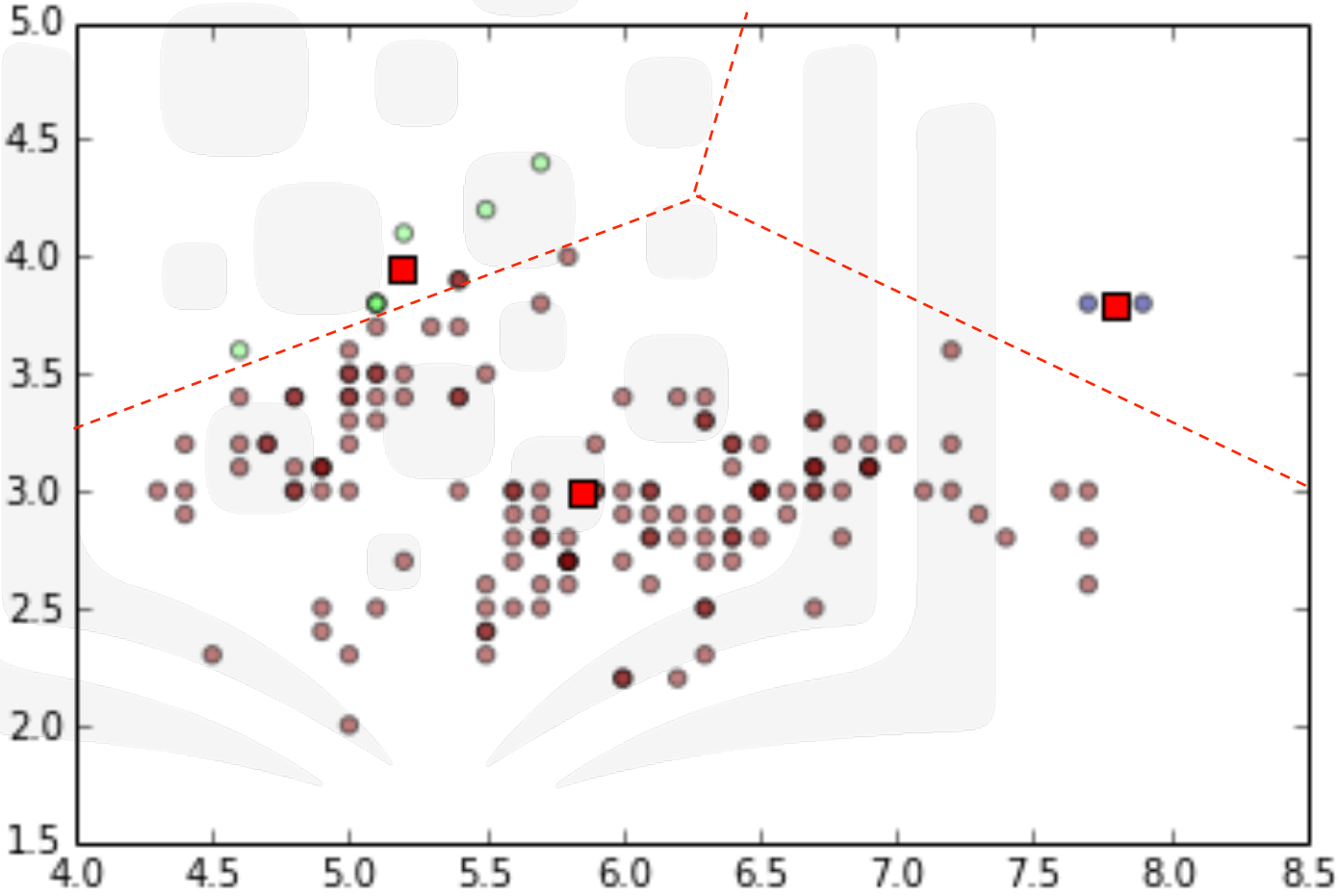
5) Assign each point to the closest centroid



BIG DATA UNIVERSITY

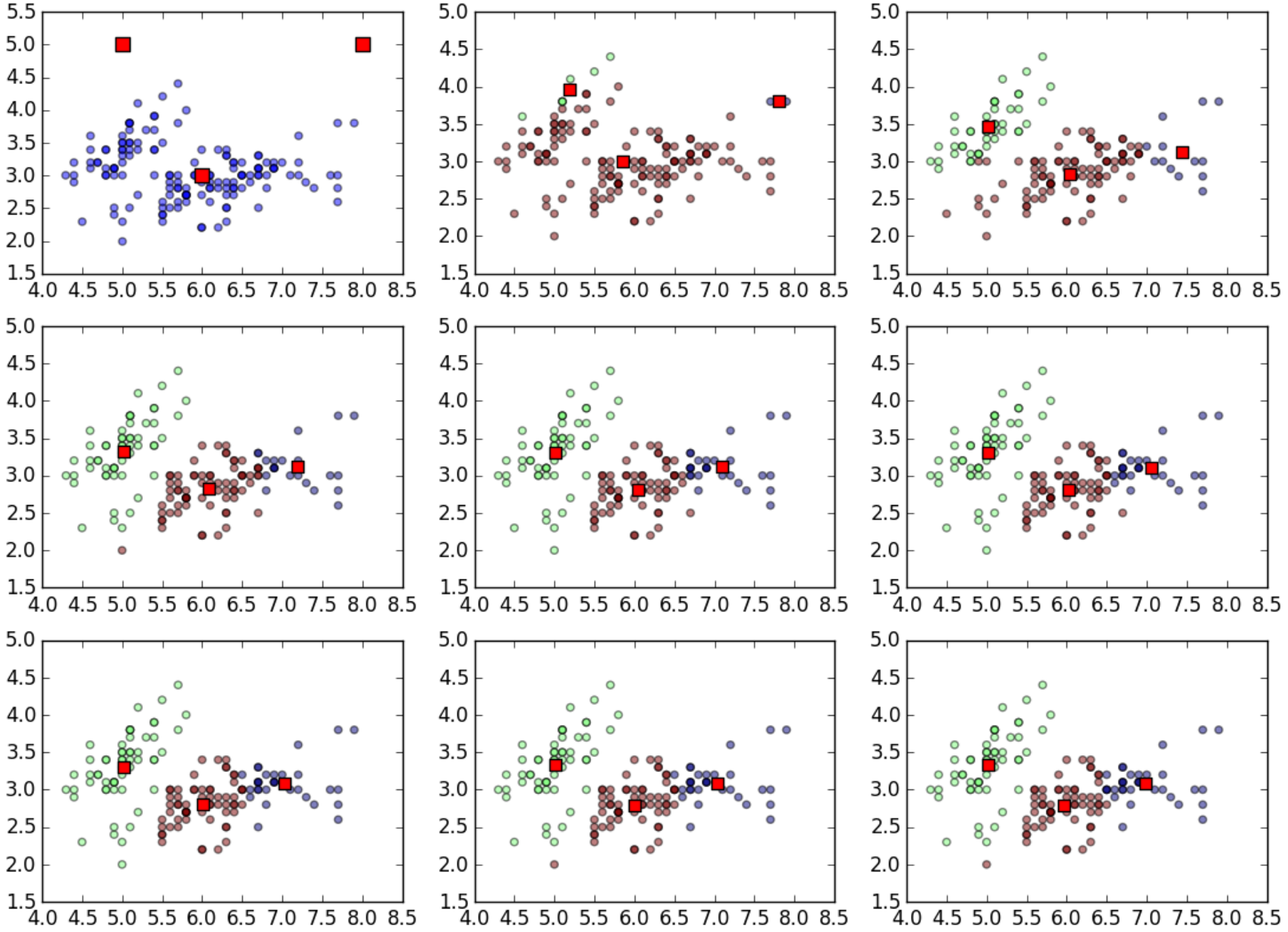
K-means Clustering

6) Compute the new centroids for each cluster.



BIG DATA UNIVERSITY

7) Repeat until there are no more changes



Free Online Courses

www.BigDataUniversity.com

Let’s jump into DSX to look at come code!