



AQM – SciProg September 2017

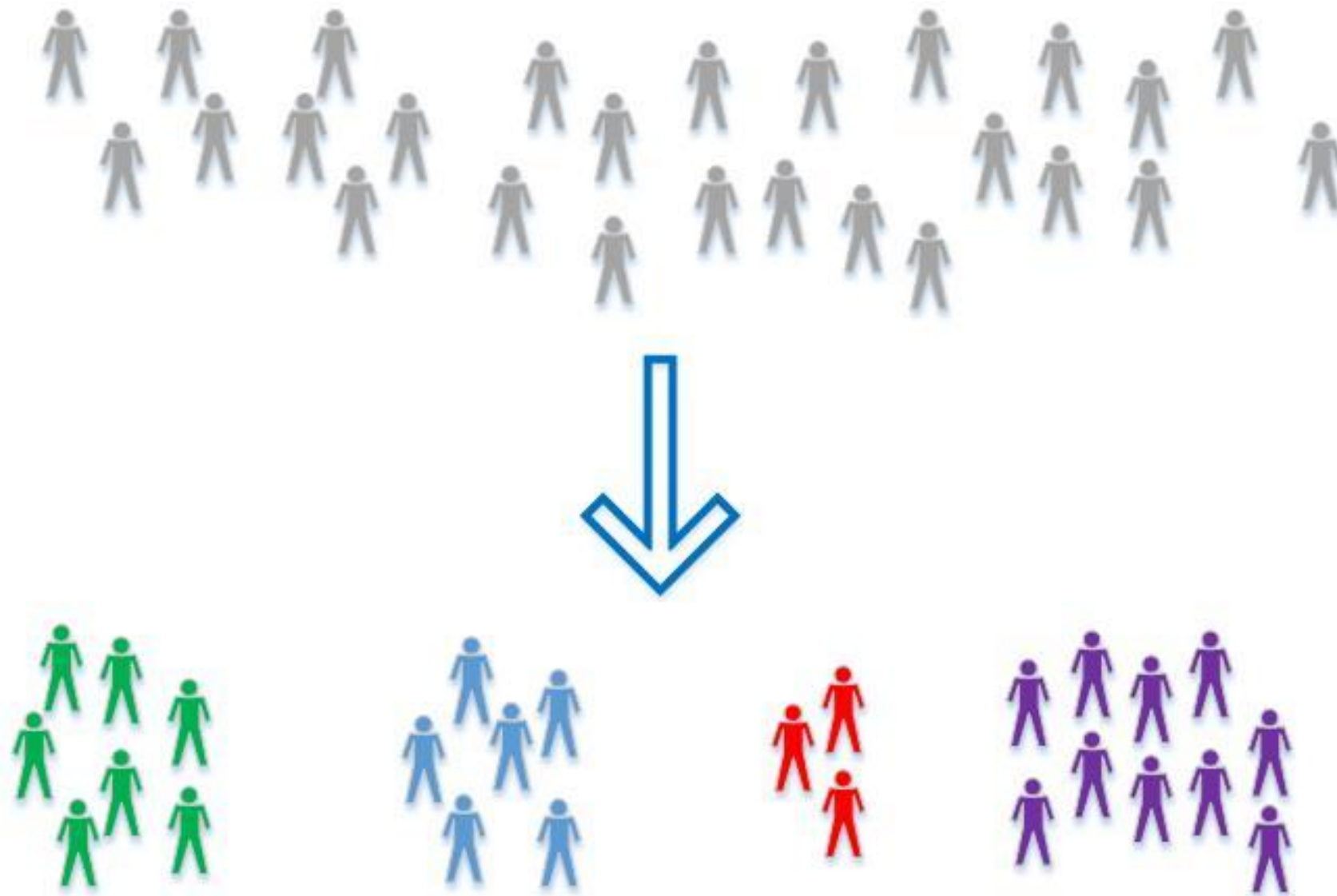
Unsupervised Ensemble Learning

Agenda

1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

Unsupervised Learning

“Unsupervised learning involves determining a hidden structure (typically groupings or clusters) from unlabelled data.”



Some Unsupervised Algorithms

- ▶ **K-means:** An algorithm that clusters data by trying to separate samples into n groups of equal variance, minimizing within-cluster sum-of-squares
- ▶ **Spectral Clustering:** a low-dimension embedding of the affinity matrix between samples, followed by a KMeans in the low dimensional space.
- ▶ **Hierarchical Clustering:** A general family of clustering algorithms that build nested clusters by merging or splitting them successively.
- ▶ **Mixture Models:** A probabilistic model that assumes all the data points are generated from a mixture of a finite number of probability distributions with unknown parameters.
- ▶ **DBSCAN:** An algorithm that views clusters as areas of high density separated by areas of low density.

Limitations of Unsupervised Learning

- ▶ How many clusters are optimal for my data?
- ▶ How stable are my clusters? Will they remain the same with new data?
- ▶ Which algorithm should I use? Is there an optimal clustering algorithm unique to my data?

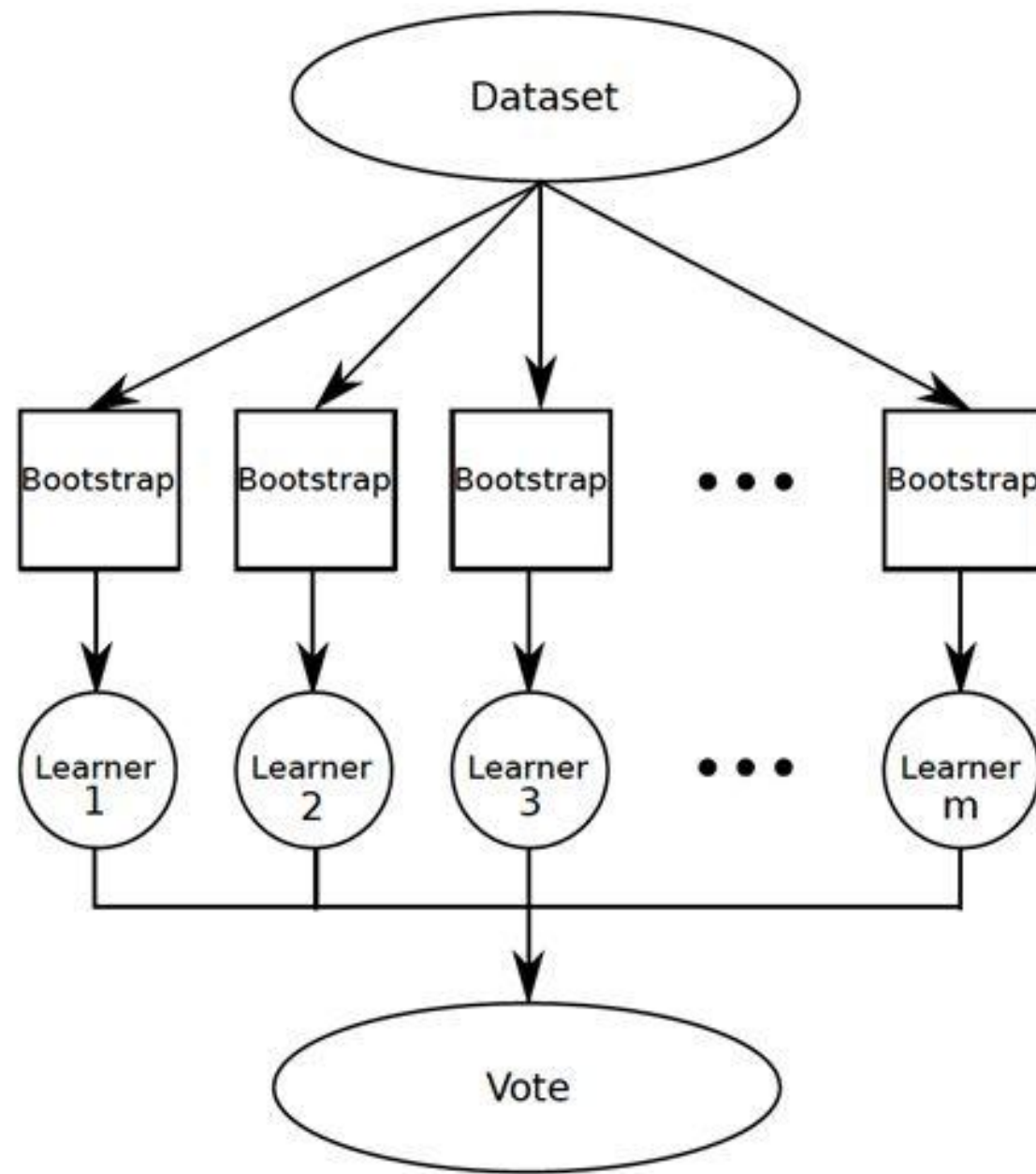
Agenda

1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

Bootstrap Aggregation

- ▶ If we had a bunch of new samples from the population, we could see how stable or robust our clustering algorithm performs over these different.
- ▶ In reality, we don't have new samples, but we can pretend **we do!** (Asymptotic theory ensures this works under certain conditions)
 - a. Create many (e.g. 100) random sub-samples of our dataset with replacement.
 - b. Train an unsupervised algorithm on each sample
 - c. Let the consensus of an observation belonging to a certain cluster win. We can also soft-assign observations to clusters.

Bootstrap Aggregation



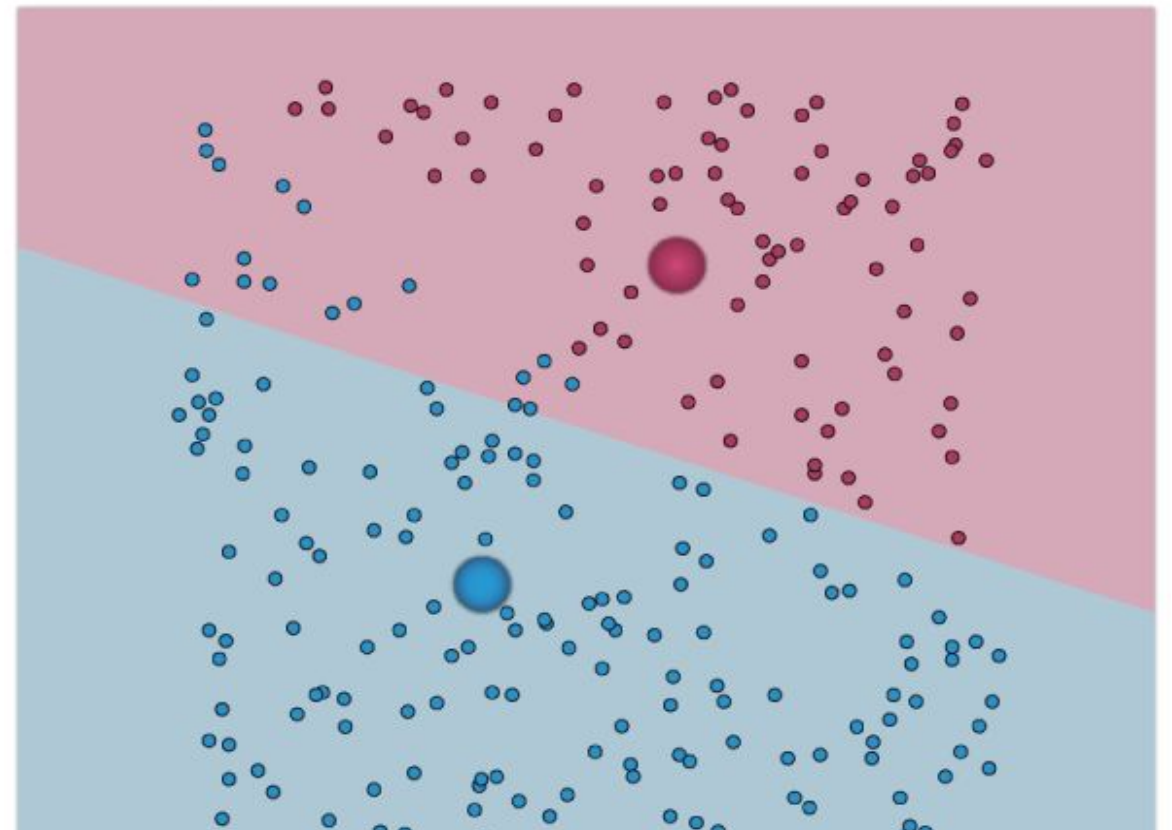
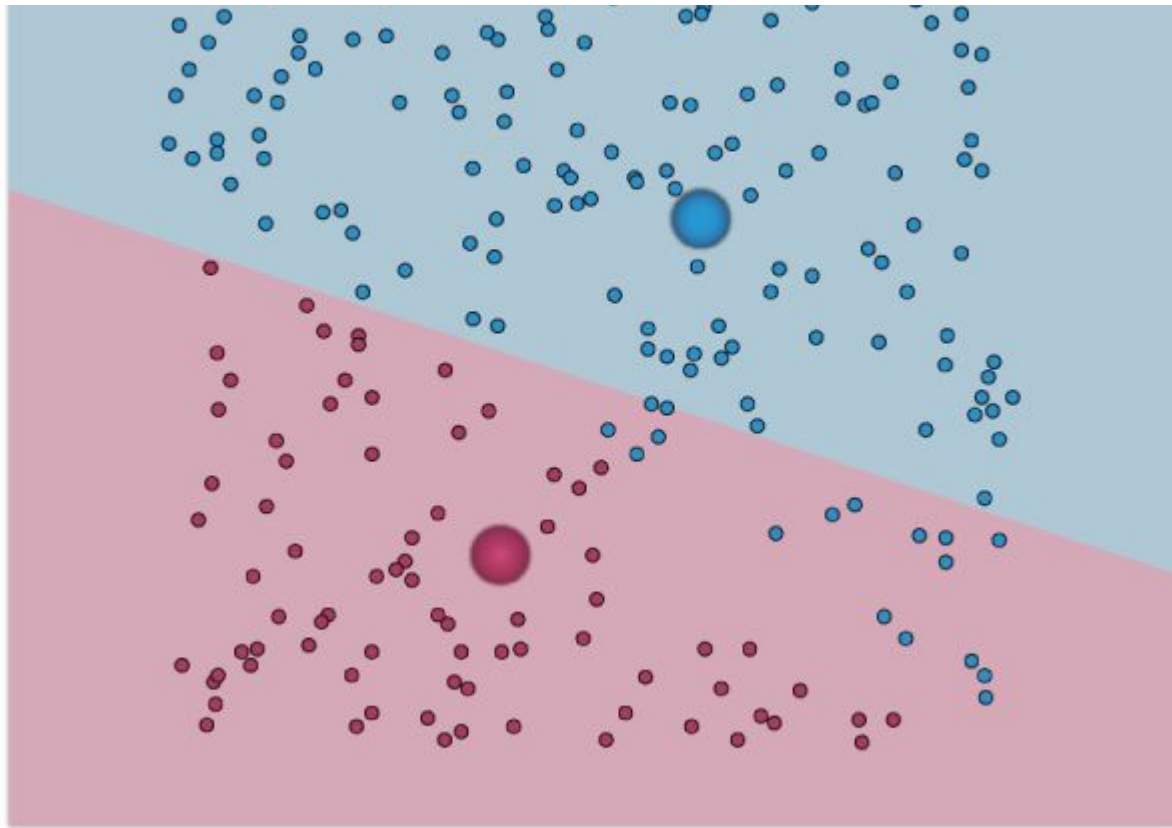
Bootstrap Aggregation

Let's Code!

Agenda

1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

Hungarian Algorithm



Do the labels match?

Hungarian Algorithm

- ▶ Do our cluster labels across bootstraps match?
 - a. i.e. does **label 1** in one sample coincide with **label 1** in another?

- ▶ We have an ‘Assignment Problem’:
 - a. Q: How do we assign the correct labels to one another?
 - b. A: Use the [Hungarian Algorithm](#)!

Hungarian Algorithm

REFERENCE LABEL					REFERENCE LABEL						
PREDICTED LABEL	[0	48	2]	→	[50	0	0]
	[50	14	0			[0	48	2	
	[0	0	36]			[0	14	36]	

optimize the diagonal of the confusion matrix

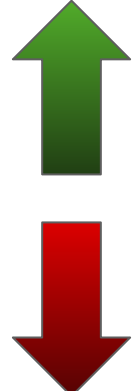
Hungarian Algorithm

Let's Code!

Agenda

1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

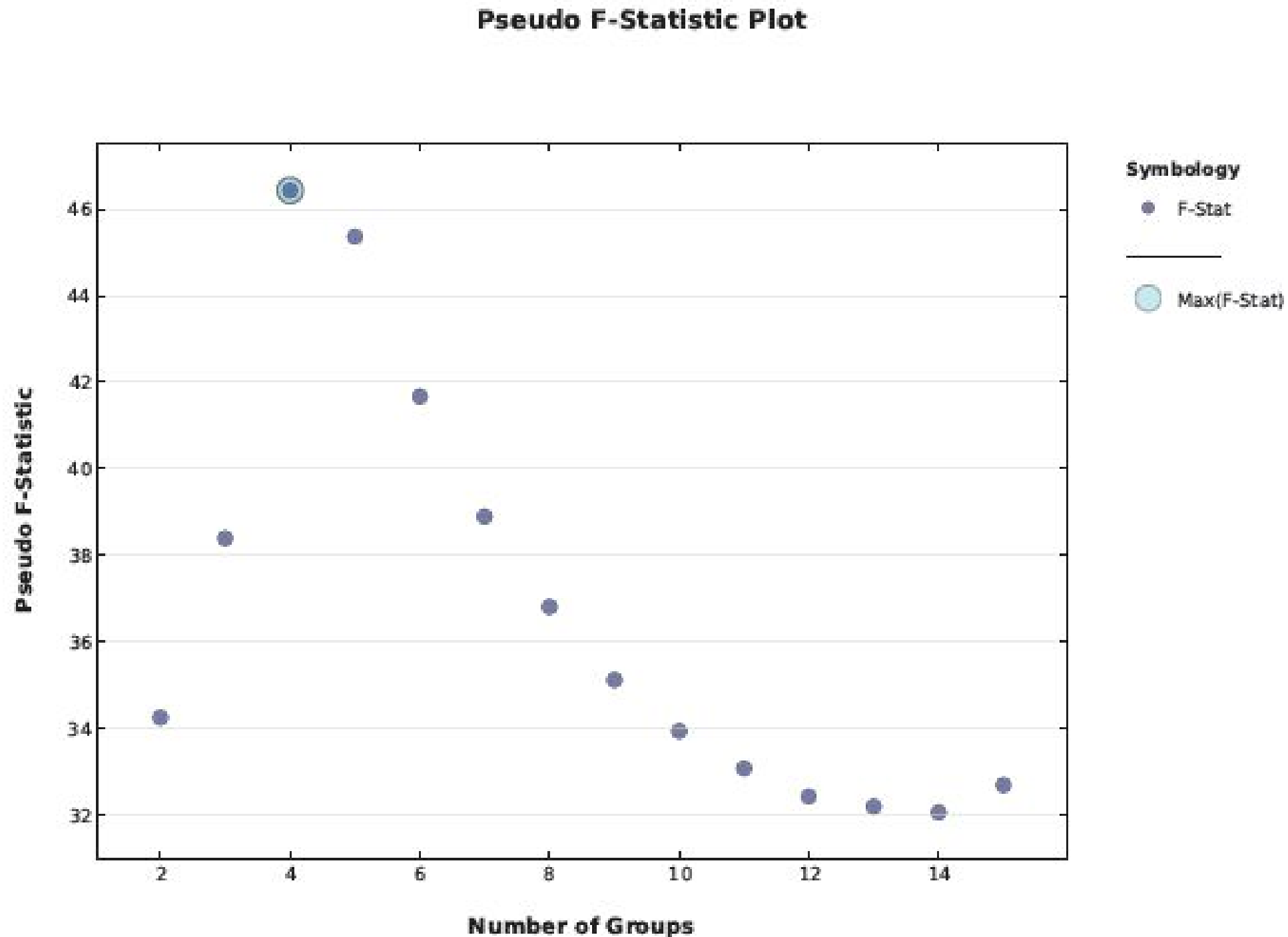
Deciding Optimal K Clusters

$$\text{Pseudo } F = \frac{(\text{GSS}) / (K - 1)}{(\text{WSS}) / (N - K)}$$


- ▶ GSS = between-group sum of squares
- ▶ WSS = within group sum of squares
- ▶ N = Number of observations
- ▶ K = Number of clusters

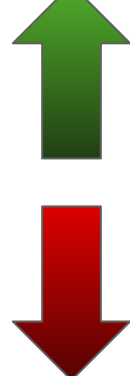
We want clusters that have the highest density and highest separation between each other

Deciding Optimal K Clusters



Deciding Optimal K with Bootstraps?

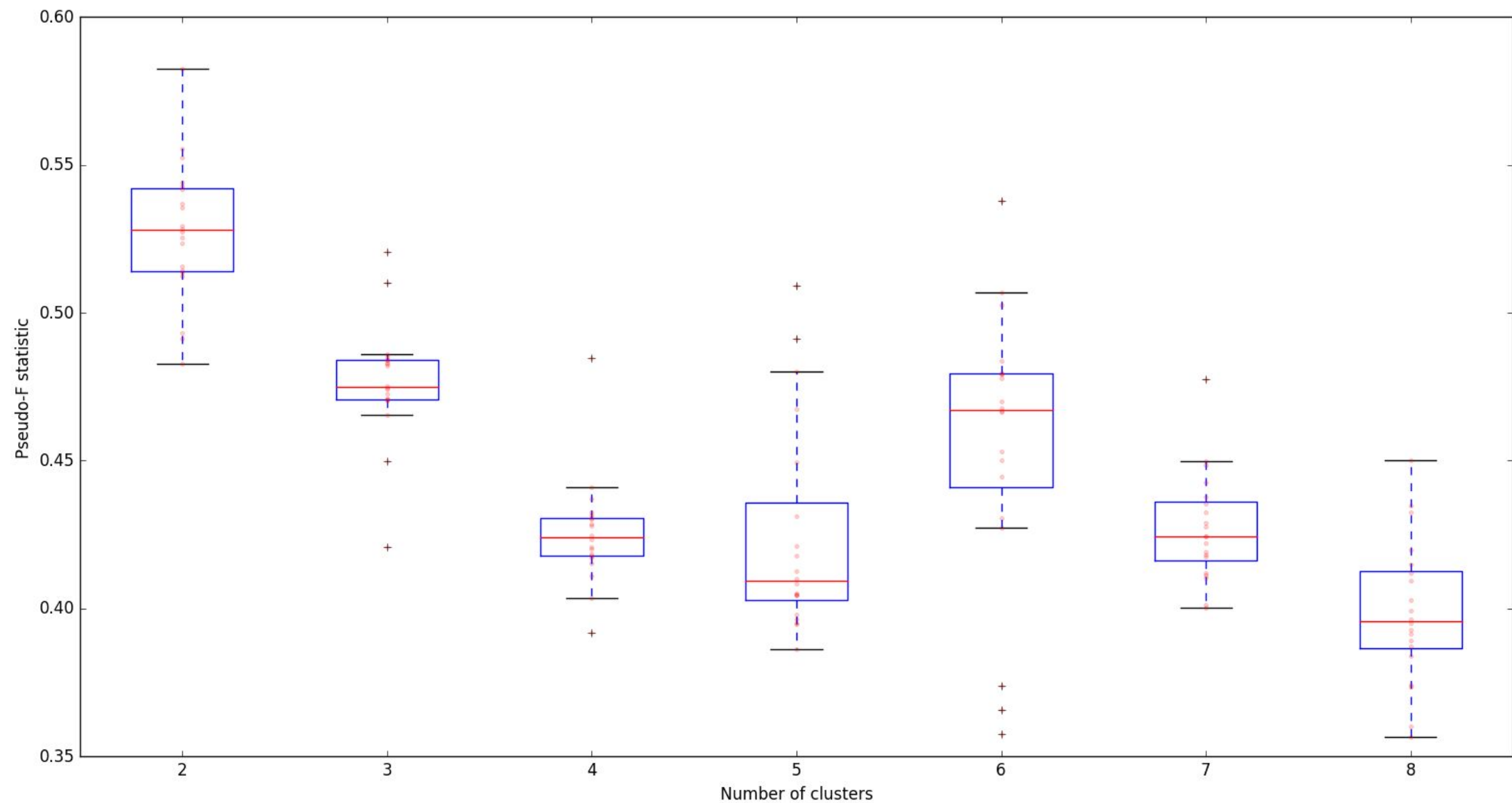
Average of Pseudo-F's

$$\text{Pseudo } F = \frac{(\text{GSS}) / (K - 1)}{(\text{WSS}) / (N - K)}$$


- ▶ GSS = between-group sum of squares
- ▶ WSS = within group sum of squares
- ▶ N = Number of observations
- ▶ K = Number of clusters

For each K, we have M samples from the M bootstraps - we have a distribution of Pseudo-F's for each K

Deciding Optimal K With Bootstraps?



Deciding Optimal K With Bootstraps?

Let's Code!

Agenda

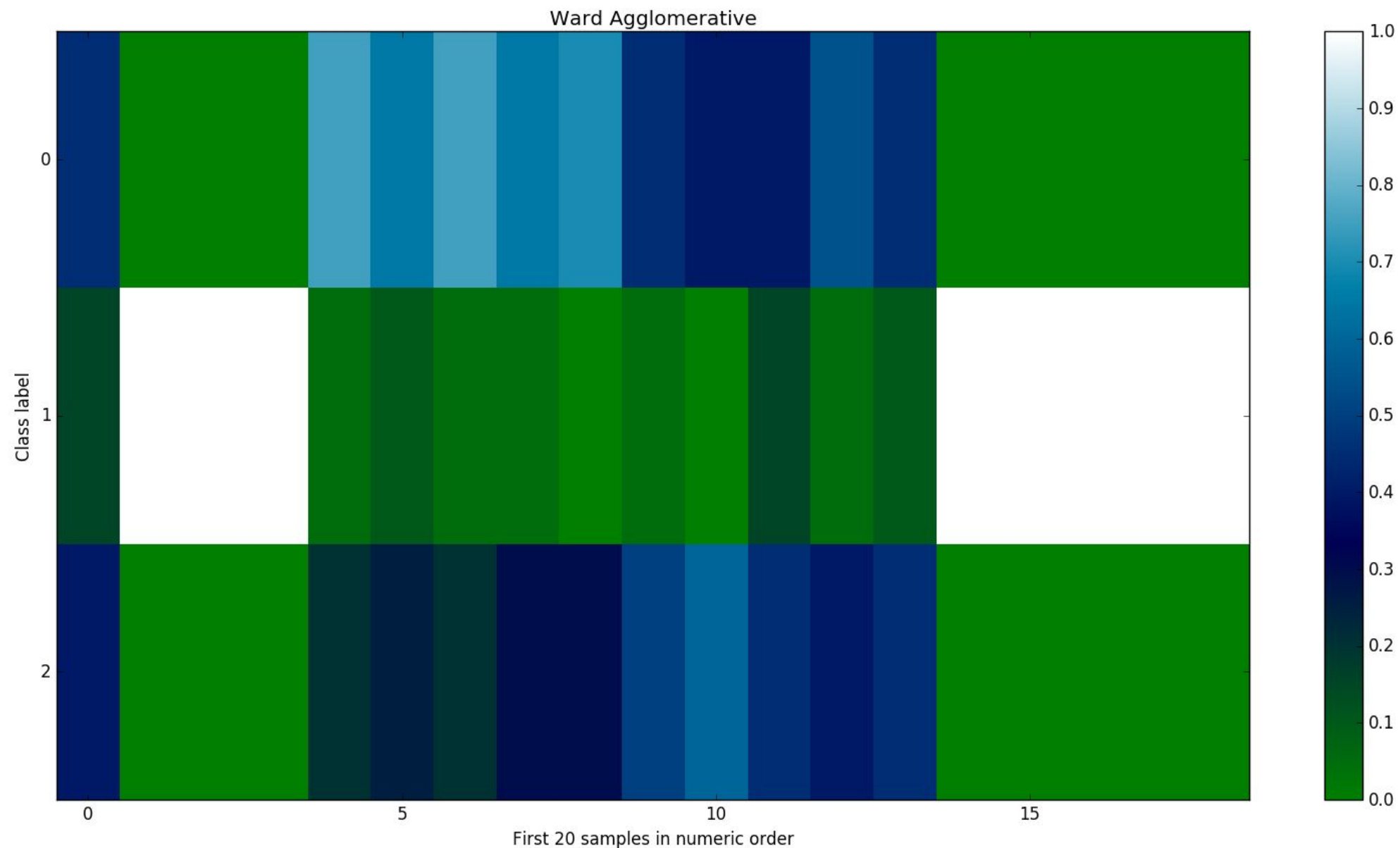
1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

Consensus of Algorithms

- ▶ We can do many types of voting strategies for unsupervised algorithms:
 - a. Vote for optimal K
 - b. Vote for optimal labels (let the models decide)
 - c. Winner takes all (let the data decide)
 - d. Different clusters get different experts
- ▶ Ultimately, we want stability and confidence
- ▶ Aggregate algorithms together in a way that makes sense

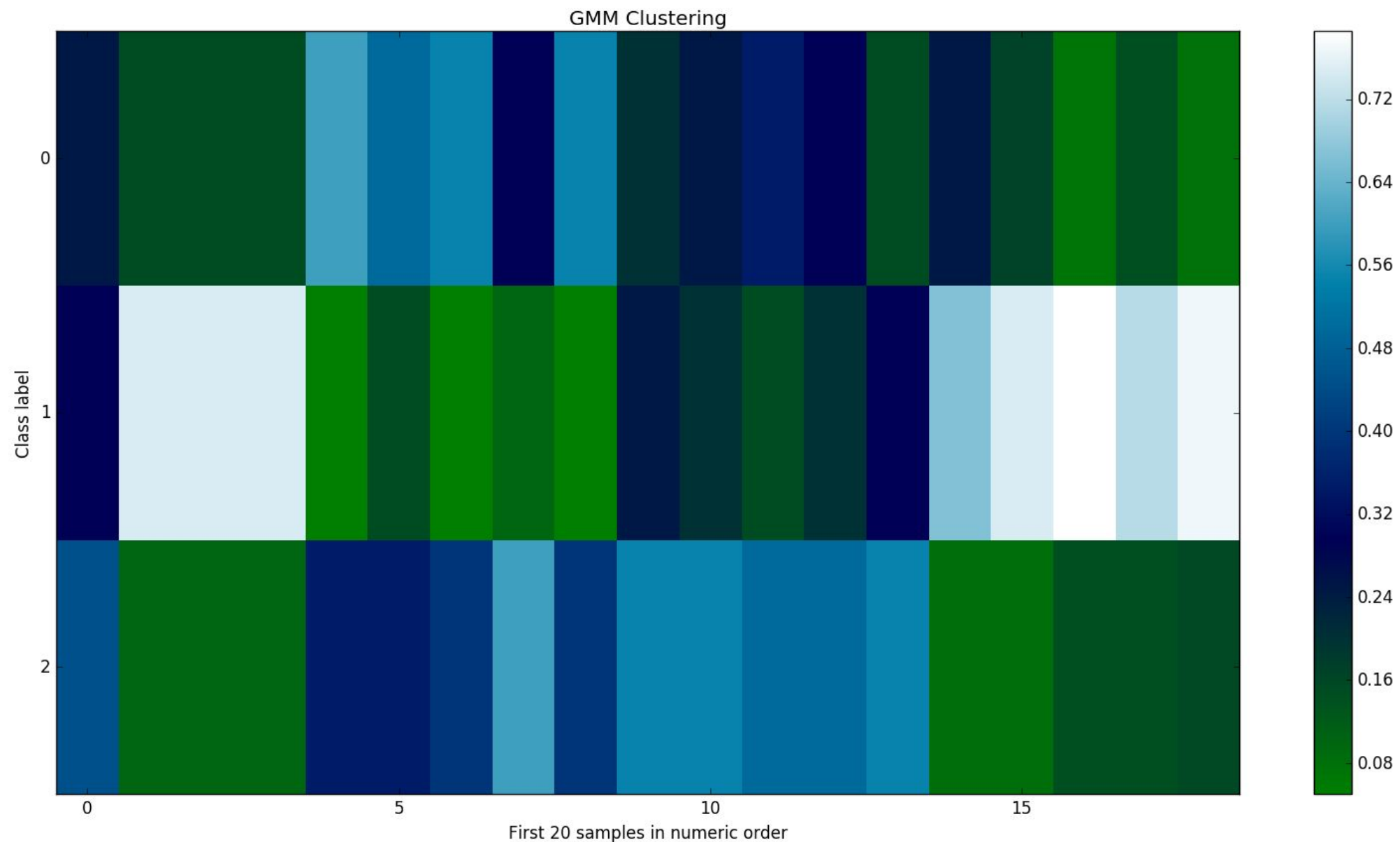
Consensus of Algorithms

- ▶ Stability of the Ward algorithm



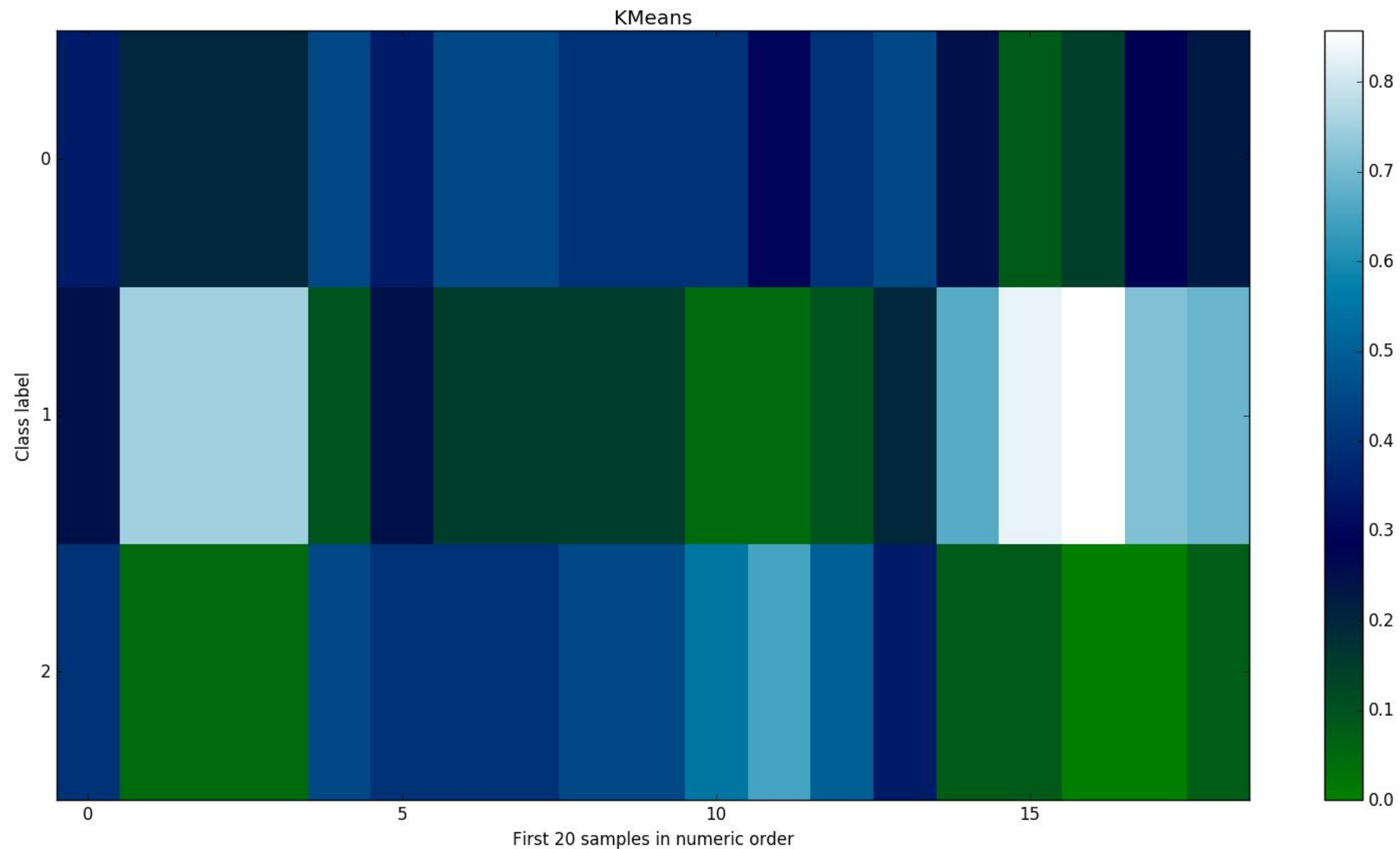
Consensus of Algorithms

- ▶ Stability of the GMM algorithm



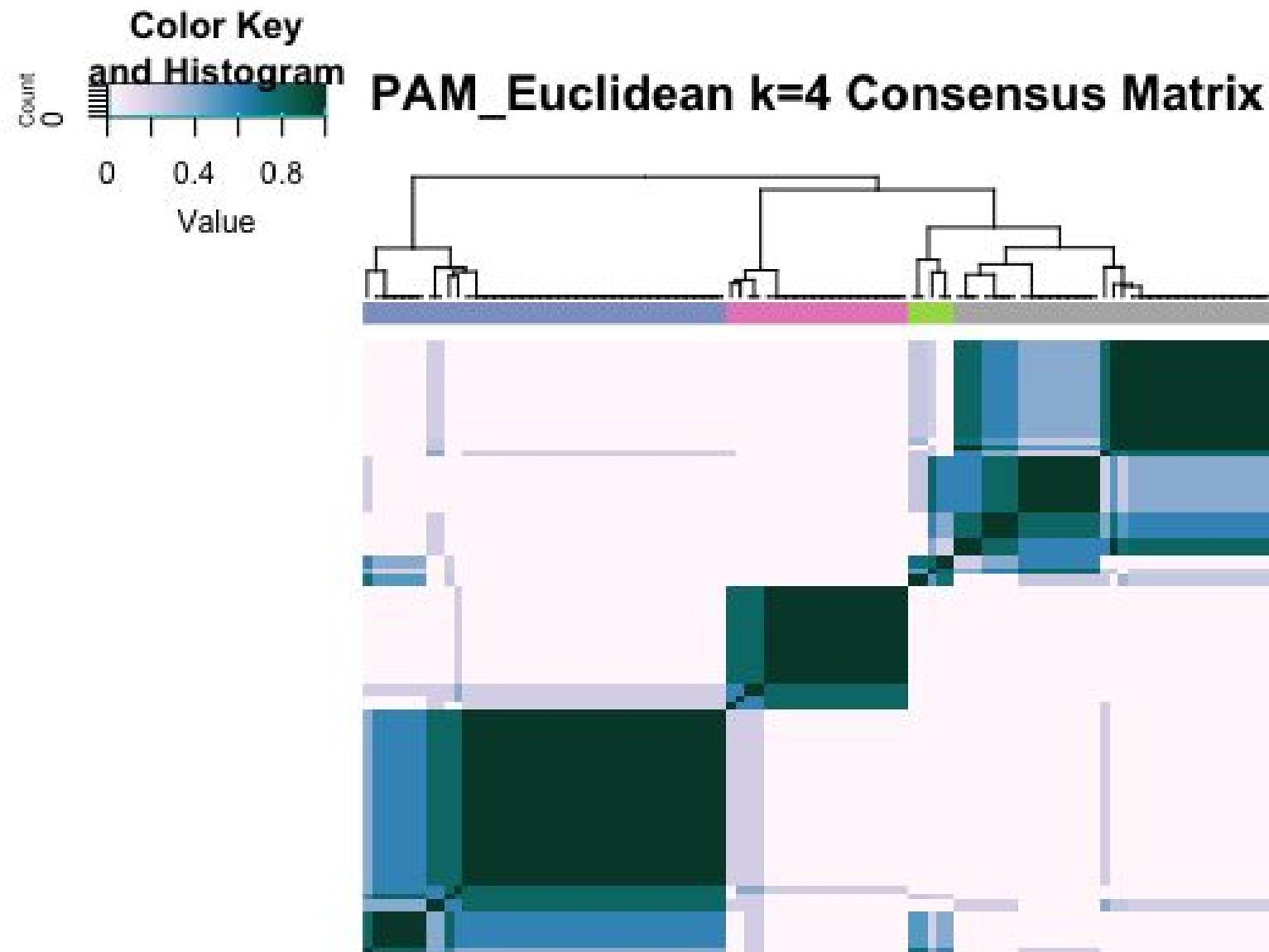
Consensus of Algorithms

- ▶ Stability of the Kmeans algorithm



Consensus of Algorithms

- Stability of the PAM algorithm



Consensus of Algorithms

Let's Code!

Agenda

1. Introduction to Unsupervised Learning for Clustering
2. Bootstrap Aggregation (Bagging)
3. Hungarian Algorithm
4. Determining Optimal Number of Clusters
5. Consensus of Unsupervised Algorithms
6. AQM Program Introduction
7. Q&A

The AQM Program



What is AQM?

AQM is a highly quantitative, rigorous 10-month data science training program designed for graduates, postgraduates and experienced professionals interested in making the transition to the lucrative field of Data Science.

- ▶ 4 months of training in all areas of Data Science expected in the field from cloud computing, data storage and processing to computational statistics, machine learning, and visualization
- ▶ 6 months focused on a real-world project for a large firm
- ▶ Supported by the Faculty of Operations Research at UBC
- ▶ Led and taught by experienced graduates and professionals
- ▶ A network of students and industry professionals advancing the data science presence in Vancouver

Why AQM?

- ▶ Immediately immersed into handling complex, real-world data while undergoing training in all aspects of Data Science
- ▶ Collaborative environment of graduates, post graduates and professionals with diverse backgrounds working together to solve problems
- ▶ Unconstrained curriculum; free to explore bleeding edge topics
- ▶ Hands-on workshops with instructors in relevant fields
- ▶ Personalized letter of recommendation from AQM and a leading firm
- ▶ Career opportunities with some of Canada's largest companies

Topics Covered



Data
Processing



Machine
Learning



Computational
Statistics

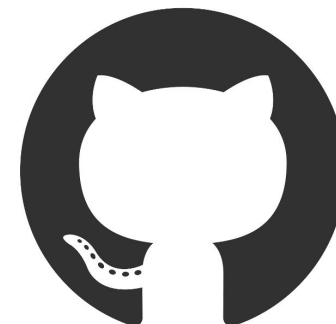
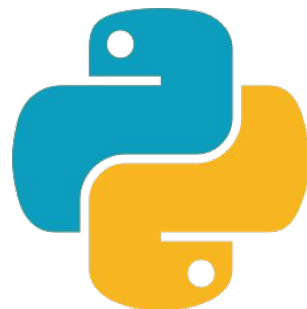


Visualization

Program Structure

Weekly 2 hour meetings @ UBC. Bi-weekly mini-projects.
Online collaboration with Github and Slack.

- ▶ **September - December:** The AQM candidates delve into exploratory data analysis, statistical computing, data pipelines and machine learning with real, complex data sets to explain and model real-world phenomena. During this stage, the value of data management and the ETL process will be enforced.
- ▶ **January - May:** The AQM team begins the well anticipated project upon a meeting with the partnered firm of that particular year. More advanced topics related to machine learning and probability are introduced relating to the the given project. Students spend this period developing a methodology to best serve the purpose of the project. At the end of the project, the team presents its findings at the company's headquarters and receives feedback. If accepted, the method with be deployed within the company and further development may be undertaken.



2017 Capstone Project



- ▶ Working with BestBuy Canada's Service Analytics team to analyze social media content that will add insights to buyer behavior and sentiment
- ▶ A four month project starting approximately in January focusing on text mining and Natural Language Processing (NLP)
- ▶ 3GB of social media data and counting with a soon-to-be deployed "listener" retrieving up to a 1GB a day.

2017 Capstone Project



- ▶ Working with TransLink's forecasting department to analyse 5GB of bus data
- ▶ A four month project starting approximately in January focusing on exploratory and modelling projects addressing the core problem posed by TransLink
- ▶ Learning additional theory, reading research papers and exploring tools necessary to answer
- ▶ Final presentations at TransLink headquarters followed by research-grade papers

Building a Community

- ▶ Have held Google Hangouts with Hadley Wickham (Chief Scientist, R Studio) and Peter Norvig (Director of Research, Google)
- ▶ Going to be holding panel talks with speakers from BCCDC, PHEMI etc
- ▶ Holding individual speaker series just for AQM team
- ▶ Hands-on learning sessions with firms
- ▶ Hackathons

Leadership Team



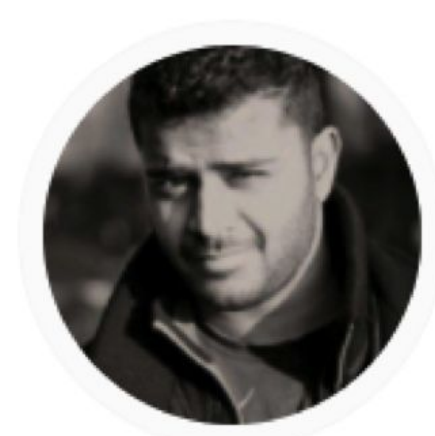
Dustin
MSc. Statistics, UBC



Sanjith
PhD. Operations Research, UBC



Haihan
BASc. Electrical Engineering, UBC



Dharu
BCom. Operations Research, UBC



Mirko
PhD. Theoretical Nuclear Physics, UBC

Program Investment

\$700 CAD

What does this pay for?

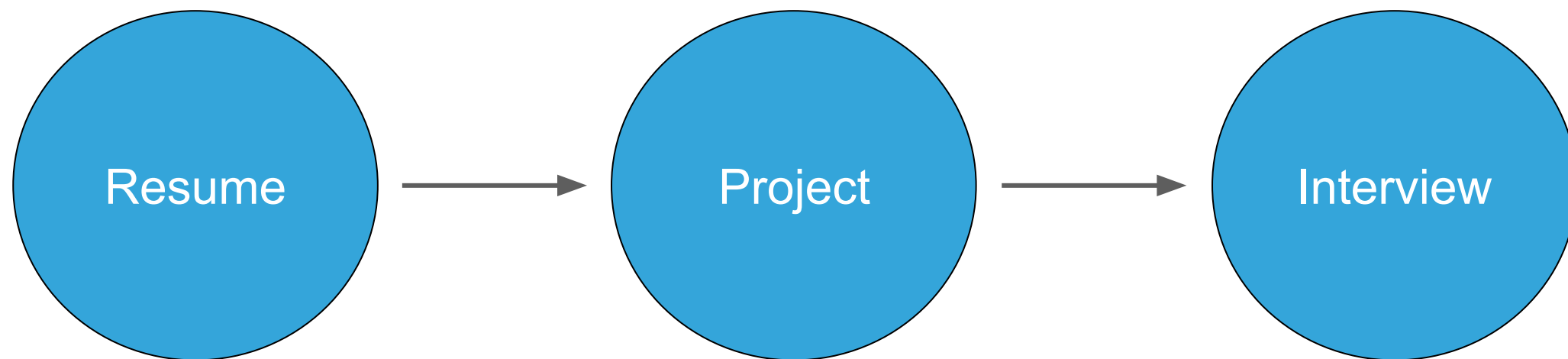
- ▶ Supporting Teaching Assistants who will be responsible for running lectures, marking assignments, evaluating students and coordinating the capstone project
- ▶ Paying for a variety of third-party software licenses used for large scale computing
- ▶ Holding panel talks and hosting industry speakers
- ▶ Scaling AQM into the best data science program

What do I get?

- ▶ A strong skill-set in data science (data exploration, data pipelines, large scale ML development, algorithm deployment)
- ▶ An extensive portfolio on GitHub
- ▶ A certificate of completion from AQM signed by the managers and faculty program sponsor
- ▶ A letter of recommendation outlining the accomplishments on the project from the firm
- ▶ Possible career opportunities

How to Apply

- ▶ Visit aqm.io to get a better idea of the program
- ▶ Navigate to the Fees & Registration section and fill out your details + attach your resume and CV in one file
- ▶ Wait for an email from us outlining the interview process



Q&A

- Question and Answer session

For Further Information

www.aqm.io

