

# Cloud Computing

## CS 15-319

Apache Mahout  
Feb 13, 2012

Shannon Quinn

# MapReduce Review

- + Scalable programming model
- + Map phase
- + Shuffle
- + Reduce phase
- + MapReduce Implementations
  - + Google
  - + Hadoop

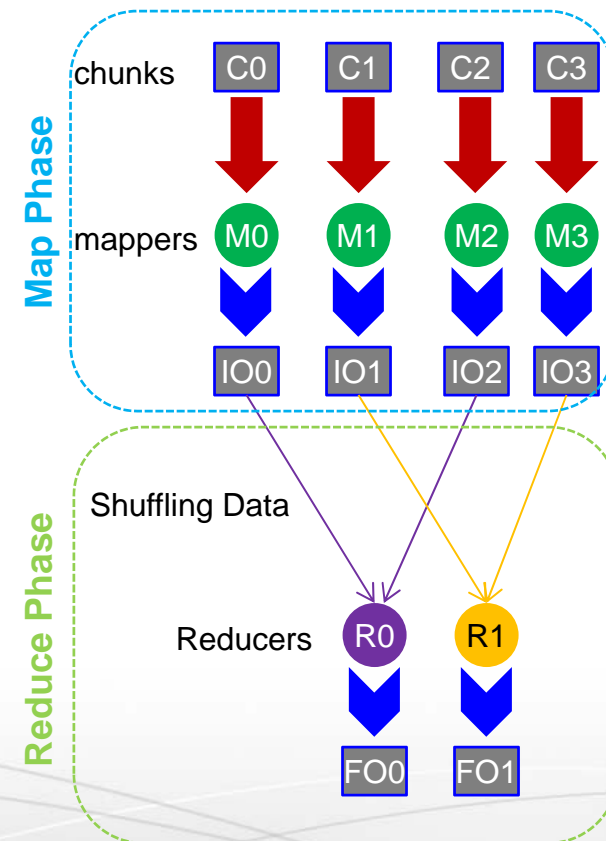


Figure from lecture 6: MapReduce

# MapReduce Review

- + Scalable programming model
- + Map phase
- + Shuffle
- + Reduce phase
- + MapReduce Implementations
  - + Google
  - + Hadoop

This is our focus!

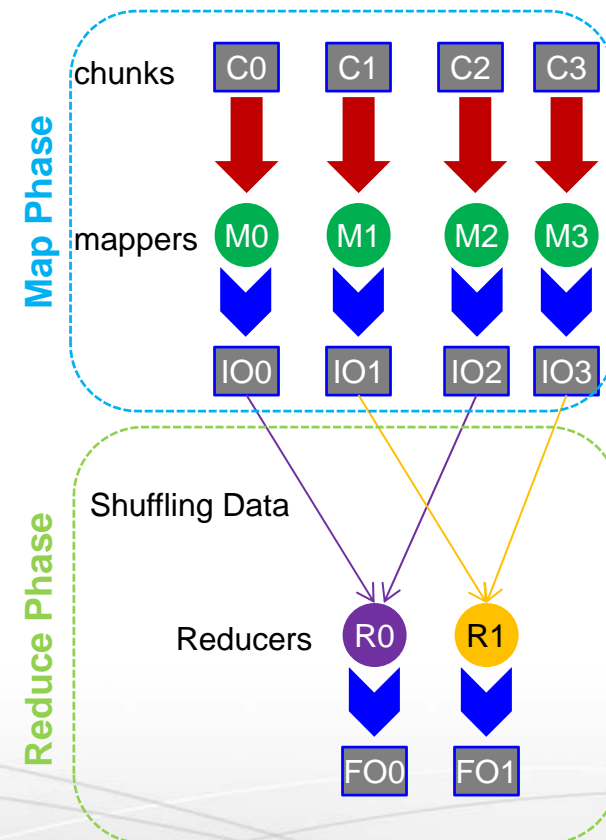


Figure from lecture 6: MapReduce

# Apache Mahout

+ A scalable machine learning library



# Apache Mahout



- + A scalable machine learning library
- + Built on Hadoop

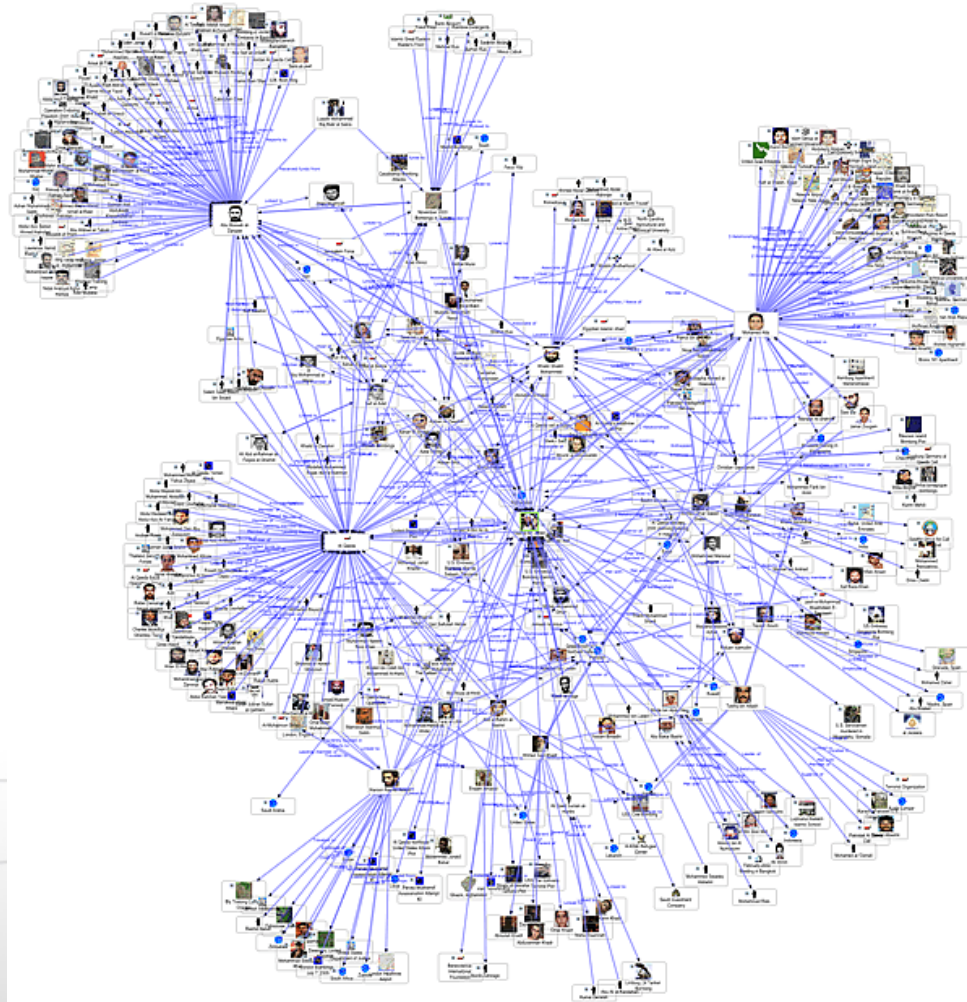


# Apache Mahout



- + A scalable machine learning library
- + Built on Hadoop
- + Philosophy of Mahout (and Hadoop by proxy)

# What does Mahout do?





# Recommendation

Who to follow · Refresh · View all



**Demetri Martin** ✓ @DemetriMartin  
Followed by waitwait and others  
[Follow](#)



**Josh Gates** ✓ @joshuagates  
Followed by David Blue and others  
[Follow](#)



**American Red Cross** ✓ @RedCr...  
Followed by James Cotton and ot...  
[Follow](#)

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).



[Programming in Python 3: A Cookbook Introduction to the Python Language](#) by Mark Summerfield  
★★★★☆ (11) \$26.98  
[Fix this recommendation](#)



[Algorithm Design](#) (Hardcover) by Jon Kleinberg  
★★★★☆ (25) \$103.24  
[Fix this recommendation](#)



[Econometrics](#) (Hardcover) by Fumio Hayashi  
★★★★☆ (19) \$74.55  
[Fix this recommendation](#)



[Mathematical Statistics, Basic and Modern Topics](#) (Paperback) by Peter J. Bickel  
★★★★☆ (4) \$68.34  
[Fix this recommendation](#)

## Critically-acclaimed Goofy Comedies

Your taste preferences created this row.

Comedies  
Goofy  
Critically-acclaimed.




Top Rated




Most Popular




Recommended »




**Family Guy Picard and Troi**  
by CristooRomania  
175,331 views



**Our three Guinea pigs fighting for a cucumber**  
by Gerald56  
2,752,485 views



**"Technically I outrank you"-WEST WING**  
by Phoenixfreak87  
44,668 views



**CJ CREGG TESTS YOUR MATH**  
by joao3xs  
89,157 views

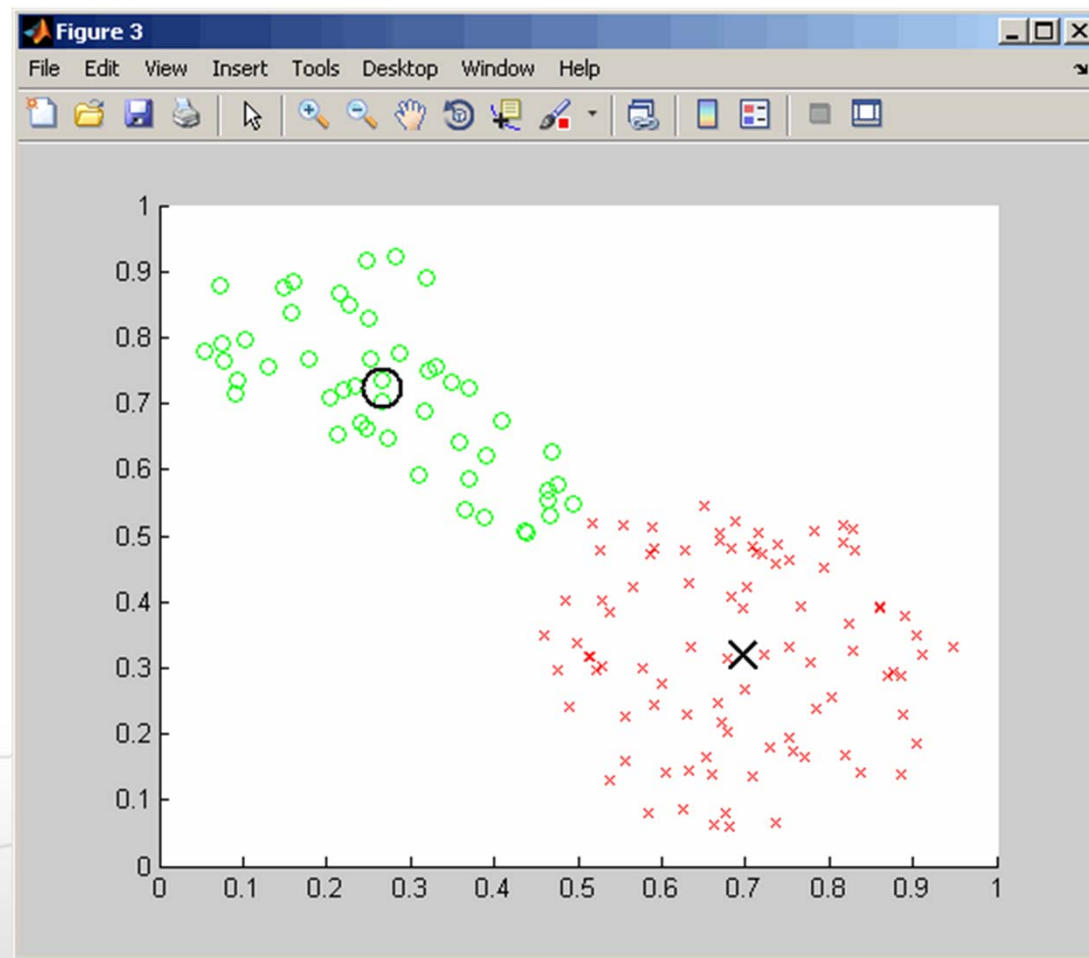
جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar



# Classification



# Clustering



# Other Mahout Algorithms

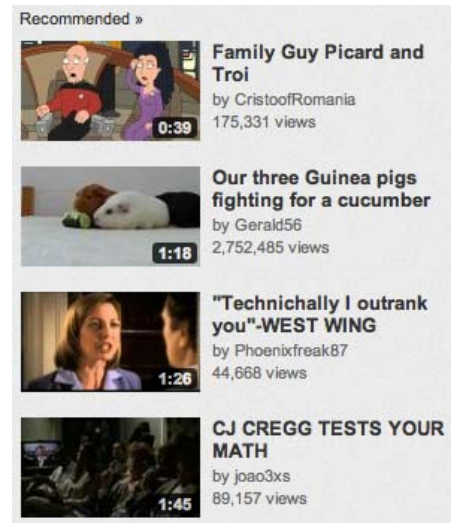
- + Dimensionality Reduction
- + Regression
- + Evolutionary Algorithms

# Mahout

1. Recommendation

2. Classification

3. Clustering



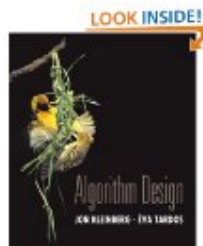
# Recommendation Overview

- + Help users find items they might like based on historical preferences

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).



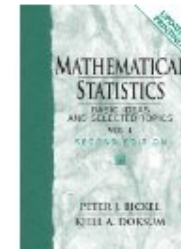
[Programming in Python 3: A C...](#) (Paperback) by Mark Summerfield  
★★★★☆ (11) \$26.98  
[Fix this recommendation](#)



[Algorithm Design](#) (Hardcover) by Jon Kleinberg  
★★★★☆ (25) \$103.24  
[Fix this recommendation](#)



[Econometrics](#) (Hardcover) by Fumio Hayashi  
★★★★★ (19) \$74.55  
[Fix this recommendation](#)



[Mathematical Statistics, Basi...](#) (Paperback) by Peter J. Bickel  
★★★★☆ (4) \$68.34  
[Fix this recommendation](#)

# Recommendation Overview

+ Mathematically



# Recommendation Overview



Alice

5

1

4

Bob

?

2

5

Peter

4

3

2

\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar



# Recommendation Overview



5

1

4

-

2

5

4

3

2

\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar

# Recommendation Overview



Bob

5

1

4

?

2

5

4

3

2

\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar

# Recommendation Overview



Bob

5

1

4

1.5

2

5

4

3

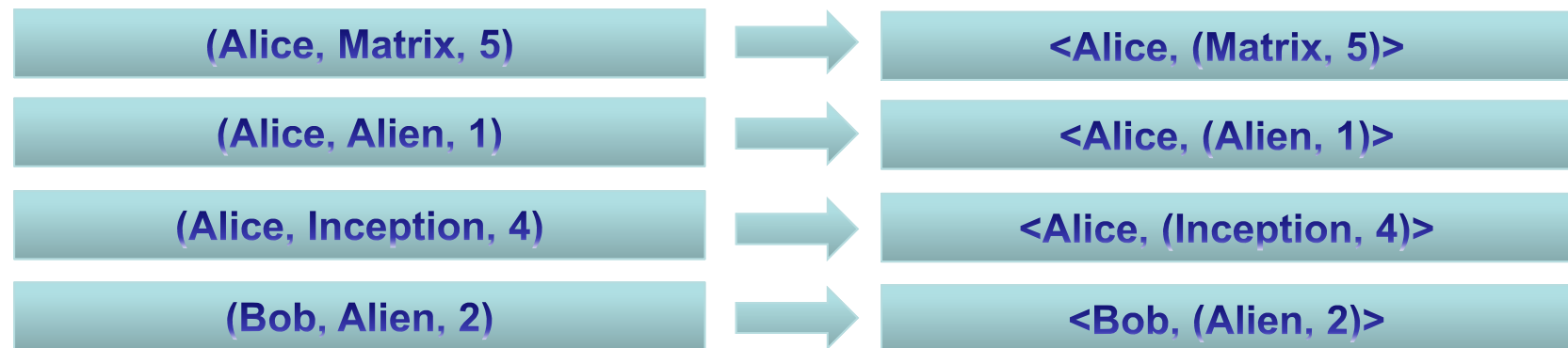
2

\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar

# Recommendation in Mahout

+ 1<sup>st</sup> Map phase: process input



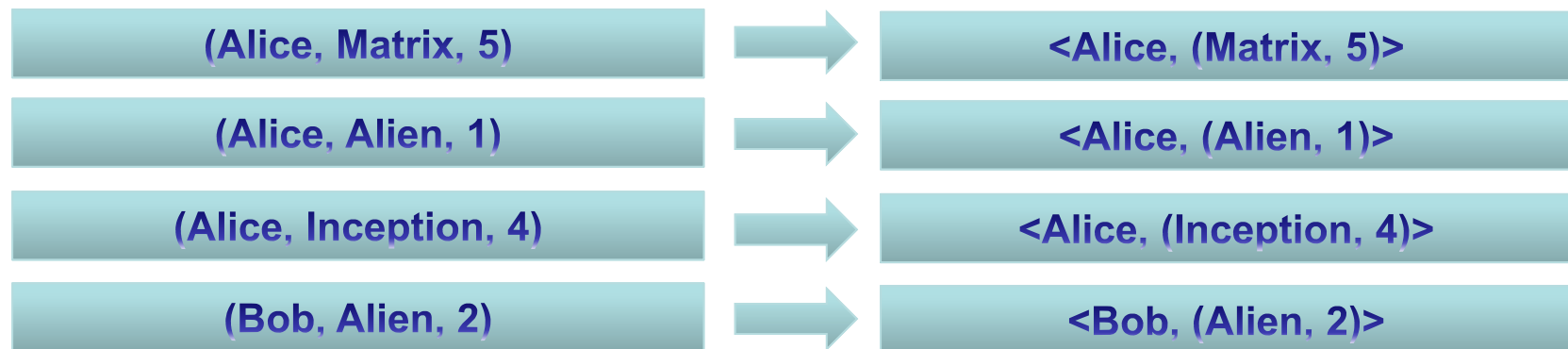
\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar

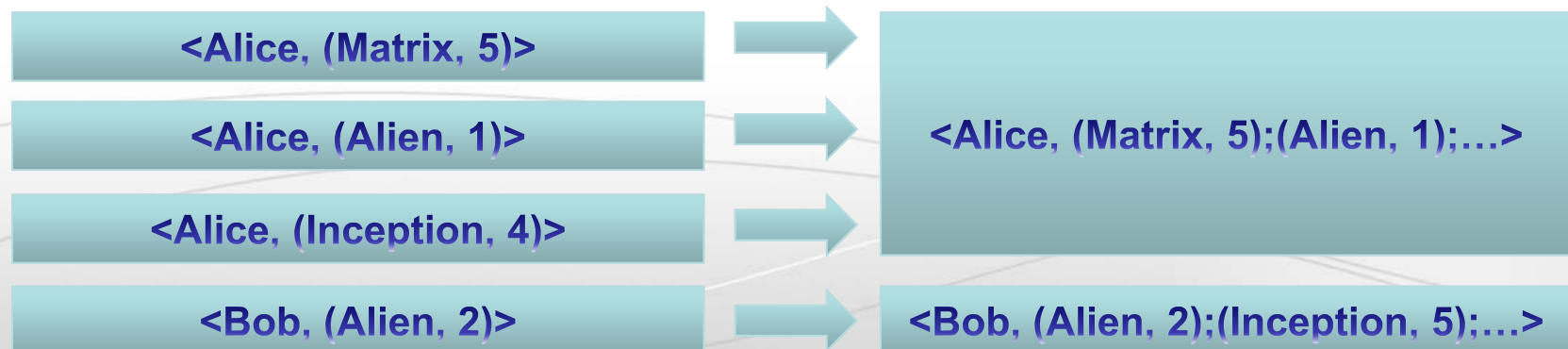
# Recommendation in Mahout

+ 1<sup>st</sup> Map phase: process input

\*based on example by  
Sebastian Schelter

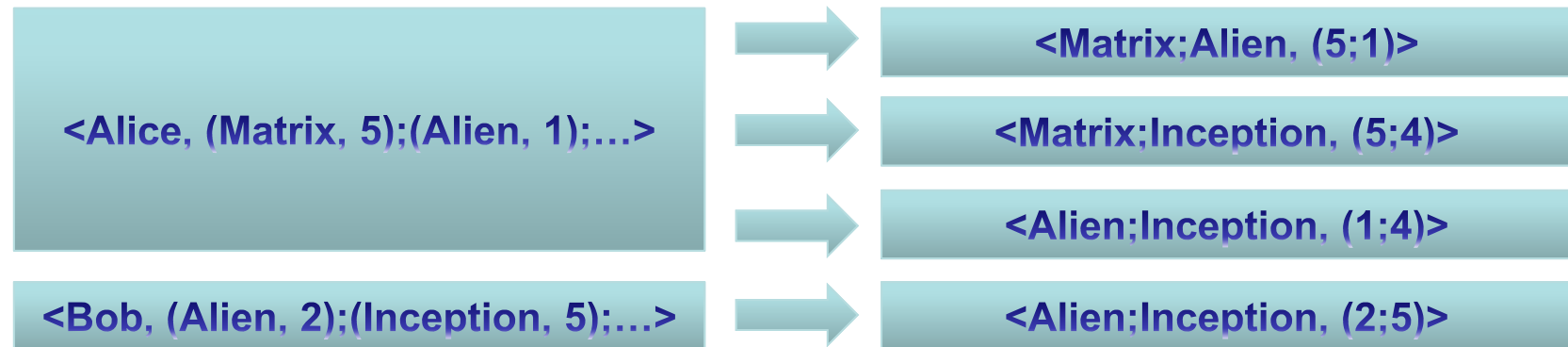


+ 1<sup>st</sup> Reduce phase: list by user



# Recommendation in Mahout

+ 2<sup>nd</sup> Map phase: Emit co-occurred ratings



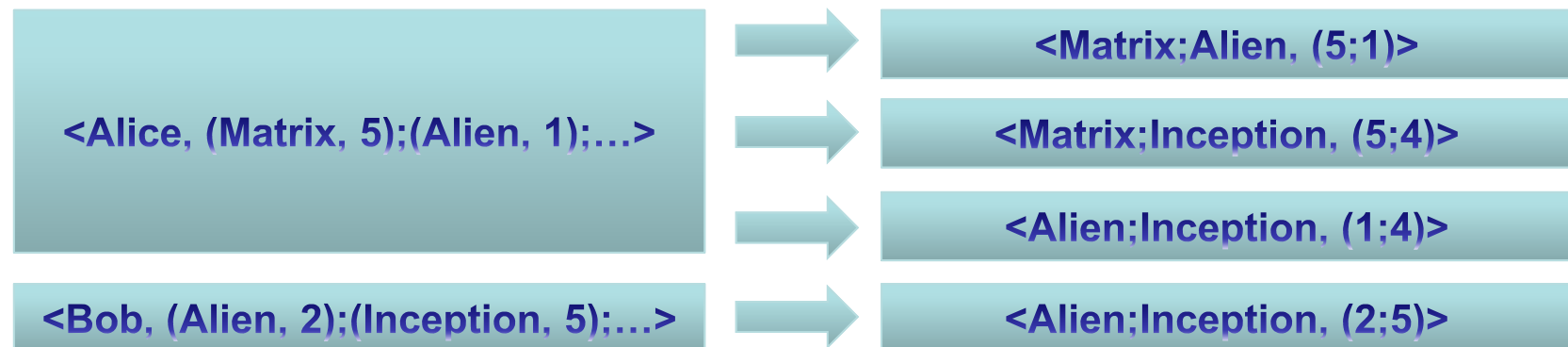
\*based on example by  
Sebastian Schelter

جامعة كارنيغي ميلون في قطر  
Carnegie Mellon Qatar

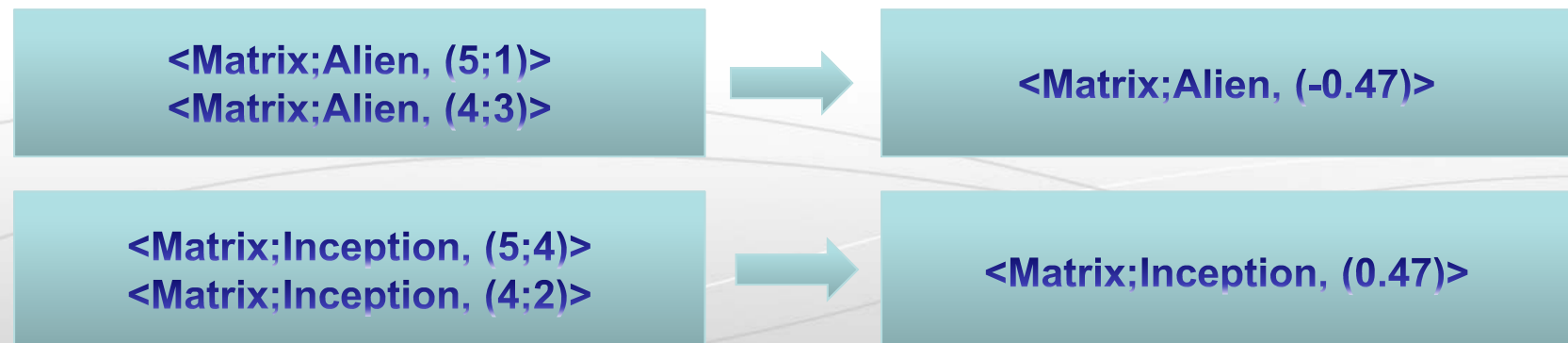
# Recommendation in Mahout

\*based on example by  
Sebastian Schelter

+ 2<sup>nd</sup> Map phase: Emit co-occurred ratings



+ 2<sup>nd</sup> Reduce phase: Compute similarities



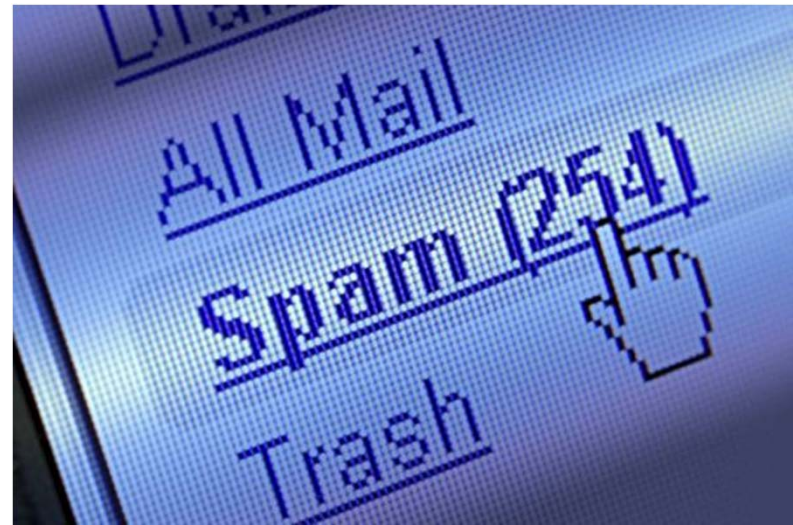


# Mahout

1. Recommendation

2. Classification

3. Clustering



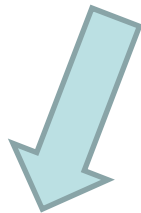
# Classification Overview

- + Assigning data to discrete categories

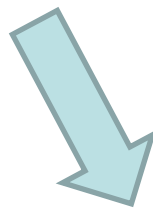
# Classification Overview



- + Assigning data to discrete categories
- + Train a model on labeled data

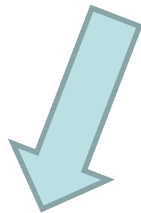


Spam

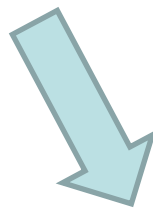


Not spam

# Classification Overview



Spam



Not spam

- + Assigning data to discrete categories
- + Train a model on labeled data
- + Run the model on new, unlabeled data

# Naïve Bayes Example

# Naïve Bayes Example

Prob (token | label) =

# Naïve Bayes Example

Not spam







# Naïve Bayes Example

Spam



# Naïve Bayes Example

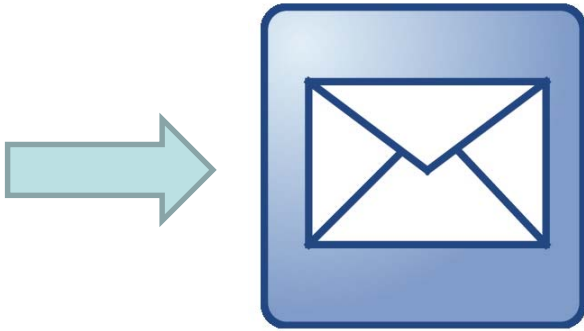
Spam



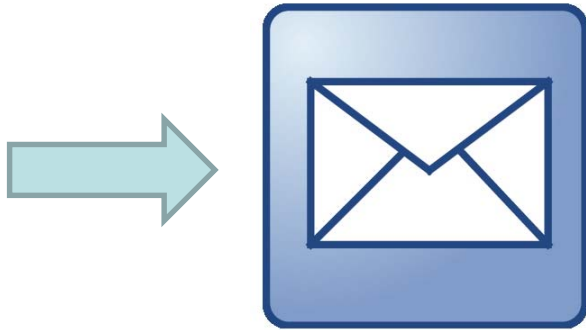
# Spam email content



# Naïve Bayes Example



# Naïve Bayes Example



“Order a trial Adobe chicken daily  
EAB-List new summer savings, welcome!”

# Naïve Bayes in Mahout

+ Complex!

# Naïve Bayes in Mahout

- + Complex!
- + Training
  1. Read the features



# Naïve Bayes in Mahout

- + Complex!

- + Training

1. Read the features
2. Calculate per-document statistics

# Naïve Bayes in Mahout

+ Complex!

+ Training

1. Read the features
2. Calculate per-document statistics
3. Normalize across categories

# Naïve Bayes in Mahout

+ Complex!

+ Training

1. Read the features
2. Calculate per-document statistics
3. Normalize across categories
4. Calculate normalizing factor of each label

# Naïve Bayes in Mahout

- + Complex!

- + Training

1. Read the features
2. Calculate per-document statistics
3. Normalize across categories
4. Calculate normalizing factor of each label

- + Testing

- + Classification

# Other Classification Algorithms

+ Stochastic Gradient Descent

# Other Classification Algorithms

- + Stochastic Gradient Descent
- + Support Vector Machines

# Other Classification Algorithms

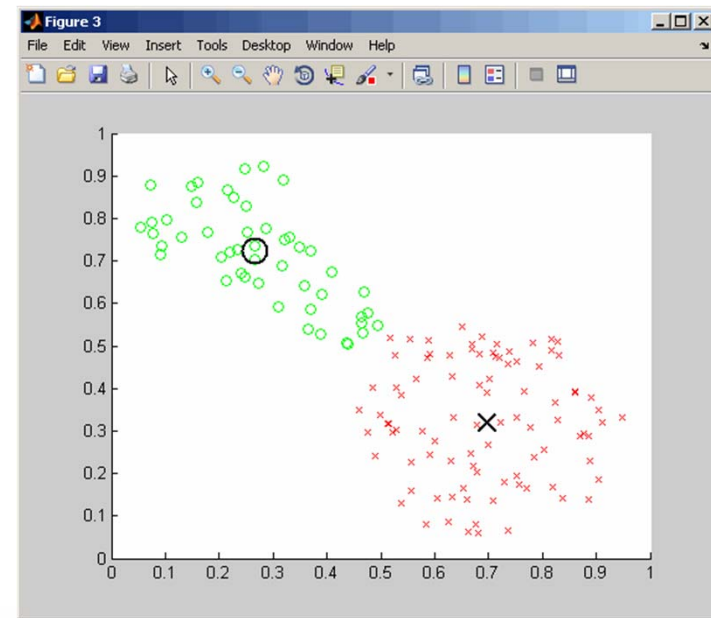
- + Stochastic Gradient Descent
- + Support Vector Machines
- + Random Forests

# Mahout

1. Recommendation

2. Classification

3. Clustering



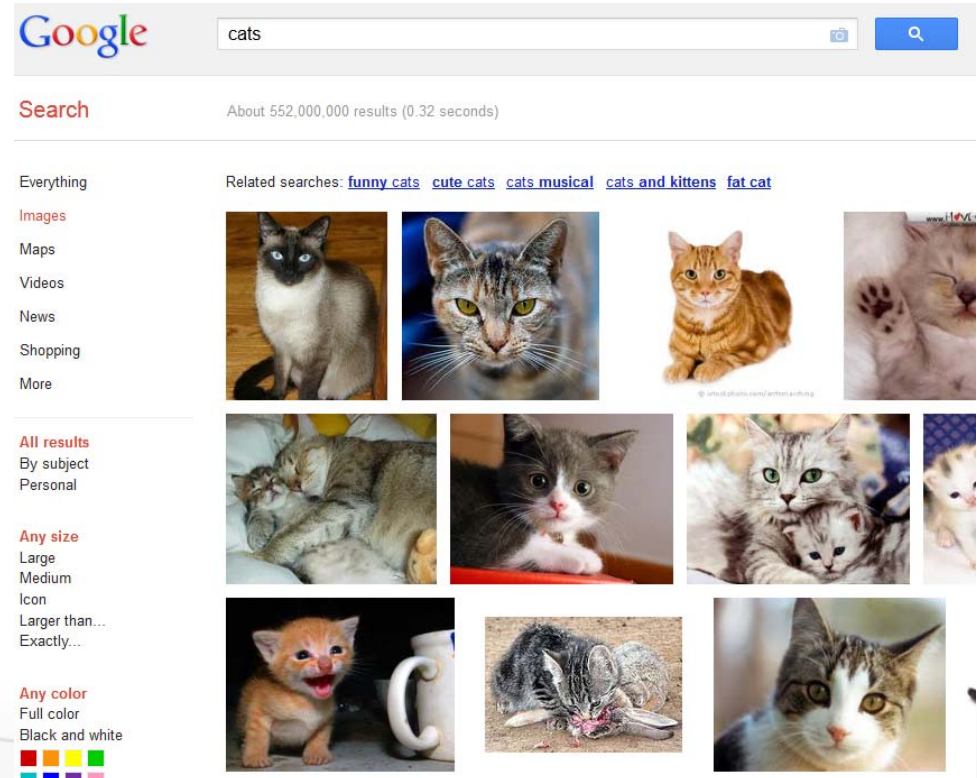


# Clustering Overview

+ Grouping  
unstructured data

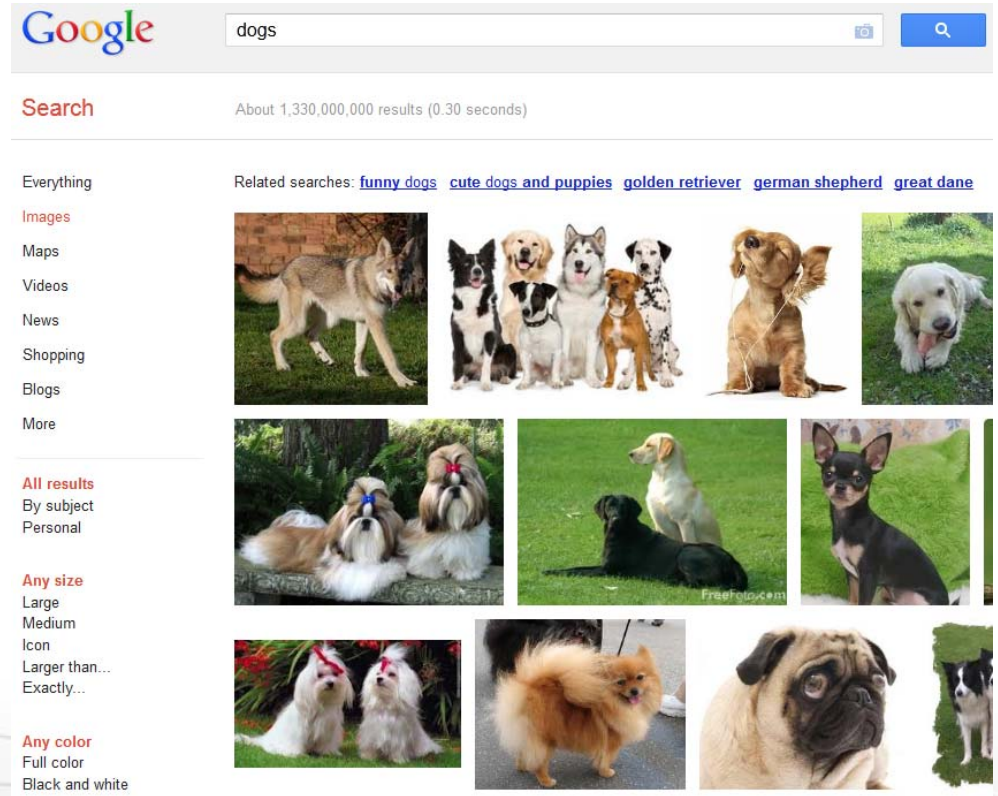
# Clustering Overview

- + Grouping unstructured data
- + Small intra-cluster distance

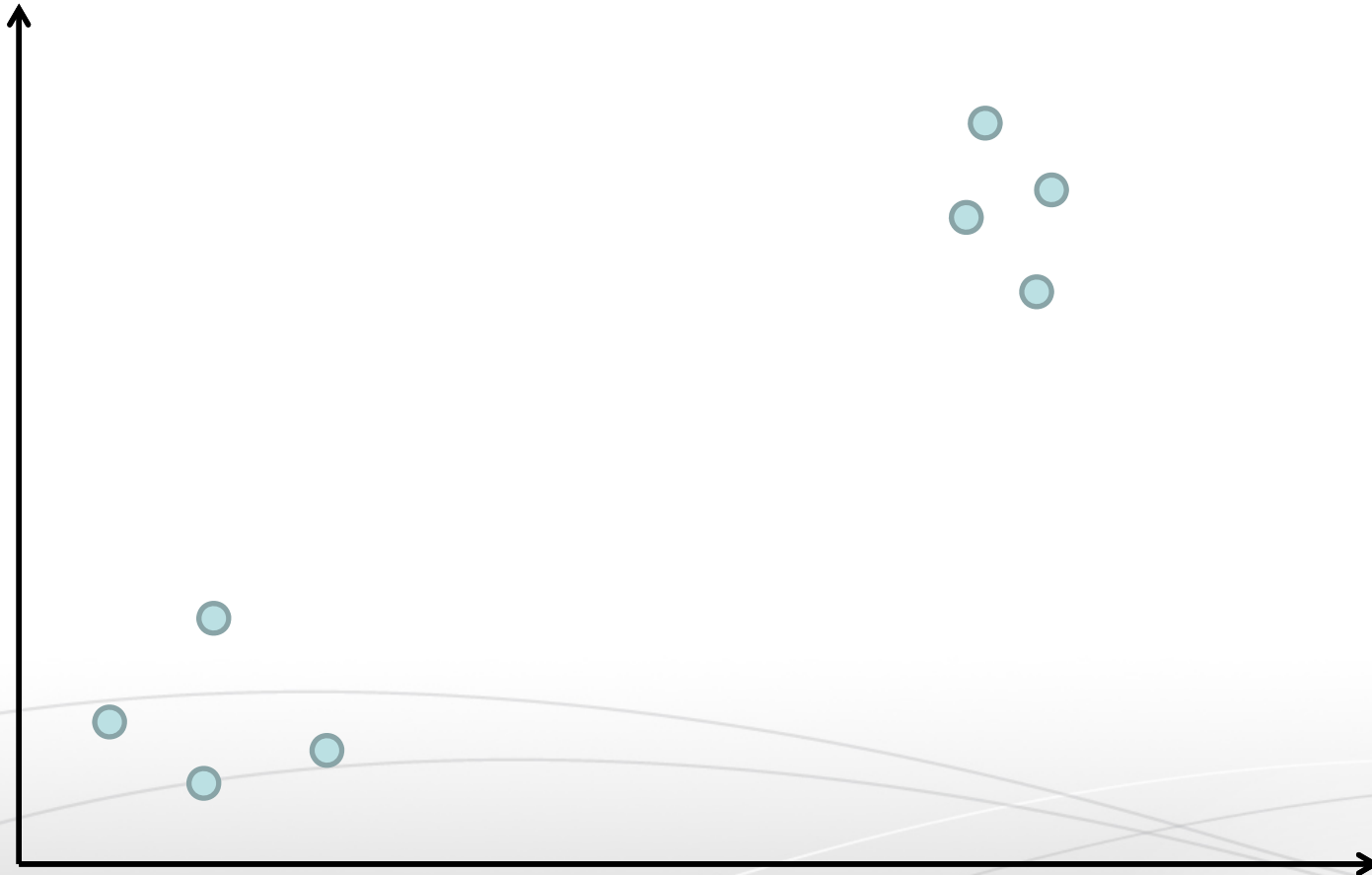


# Clustering Overview

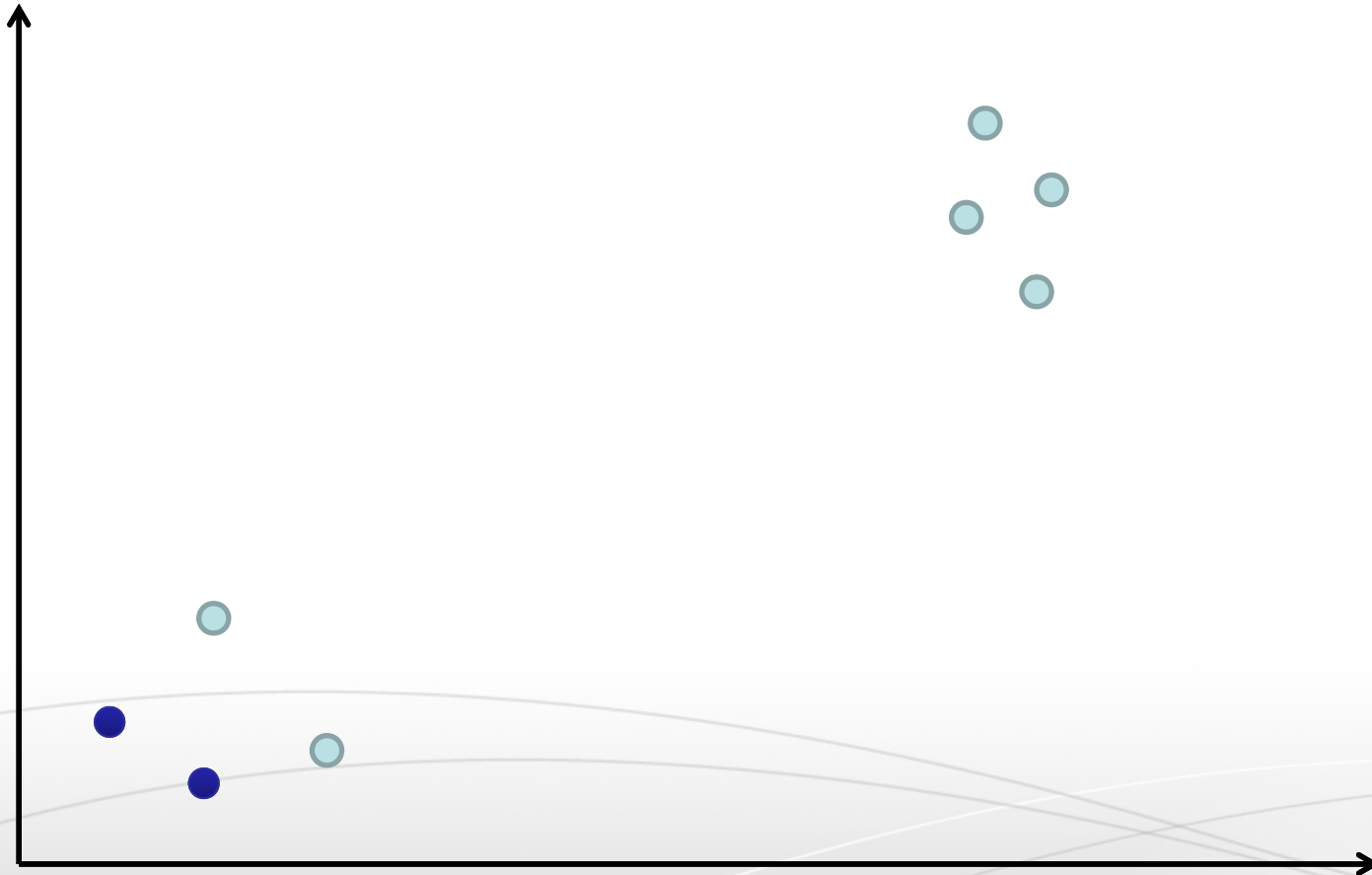
- + Grouping unstructured data
- + Small intra-cluster distance
- + Large inter-cluster distance



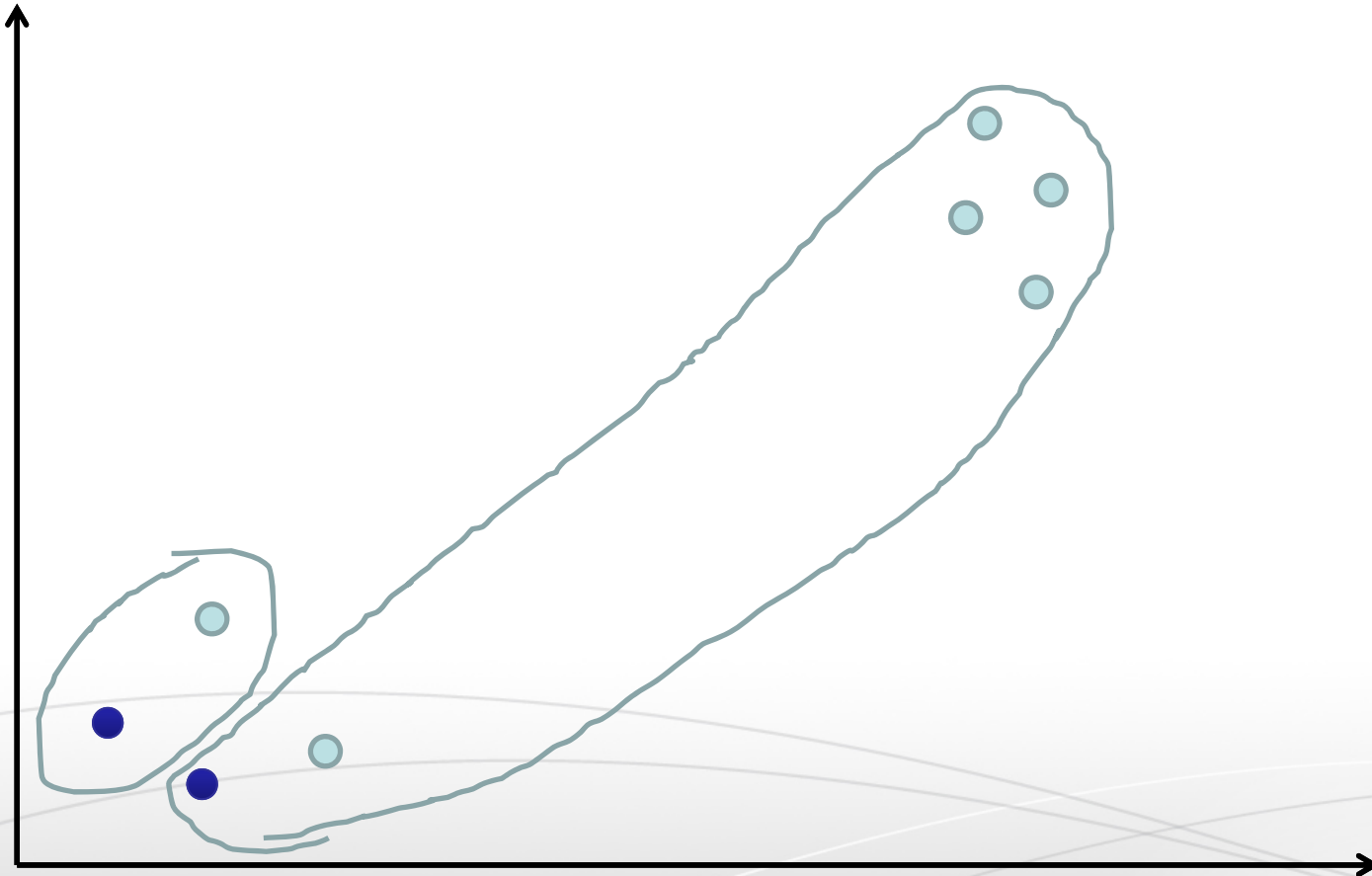
# K-Means Clustering Example



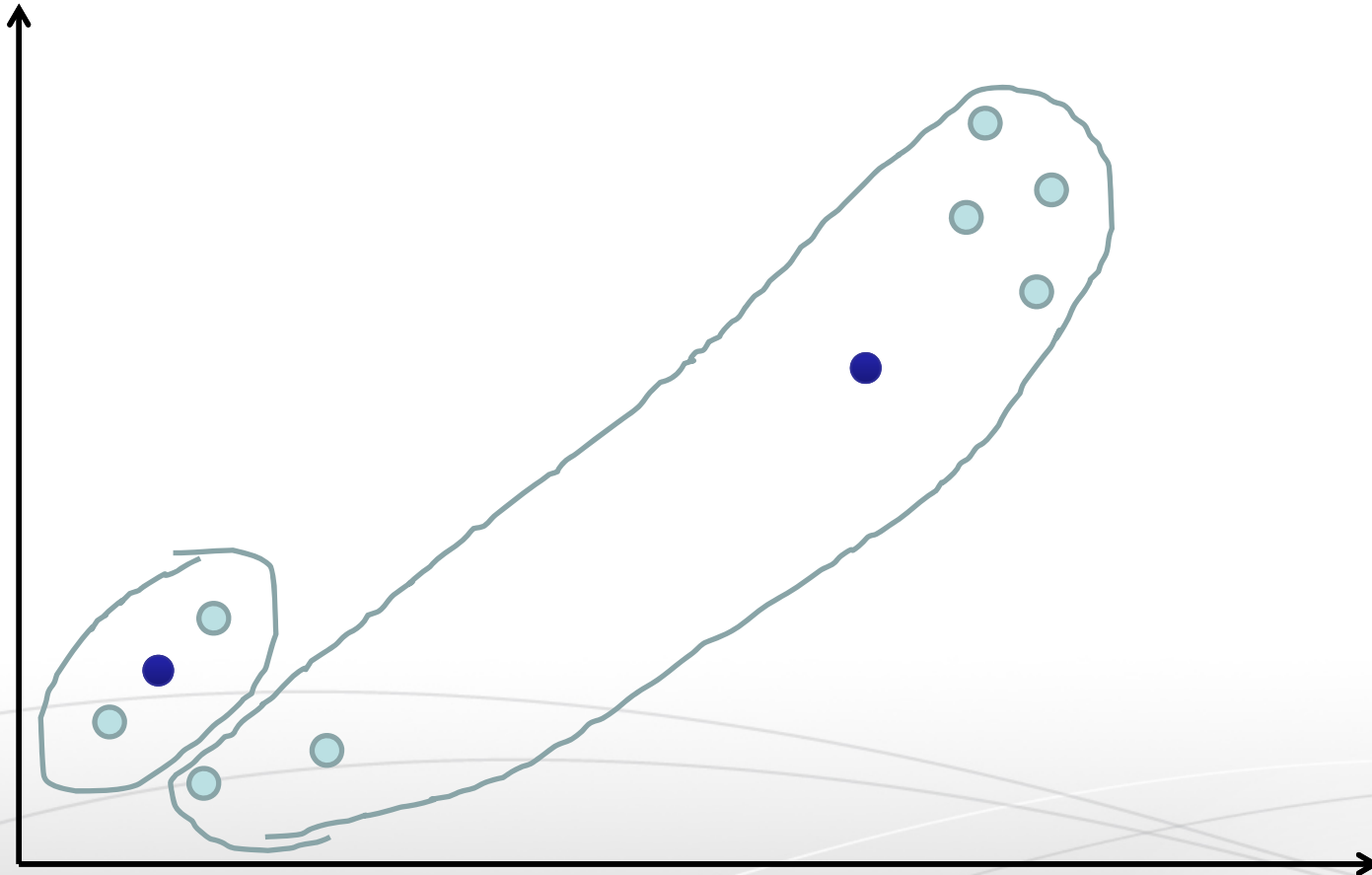
# K-Means Clustering Example



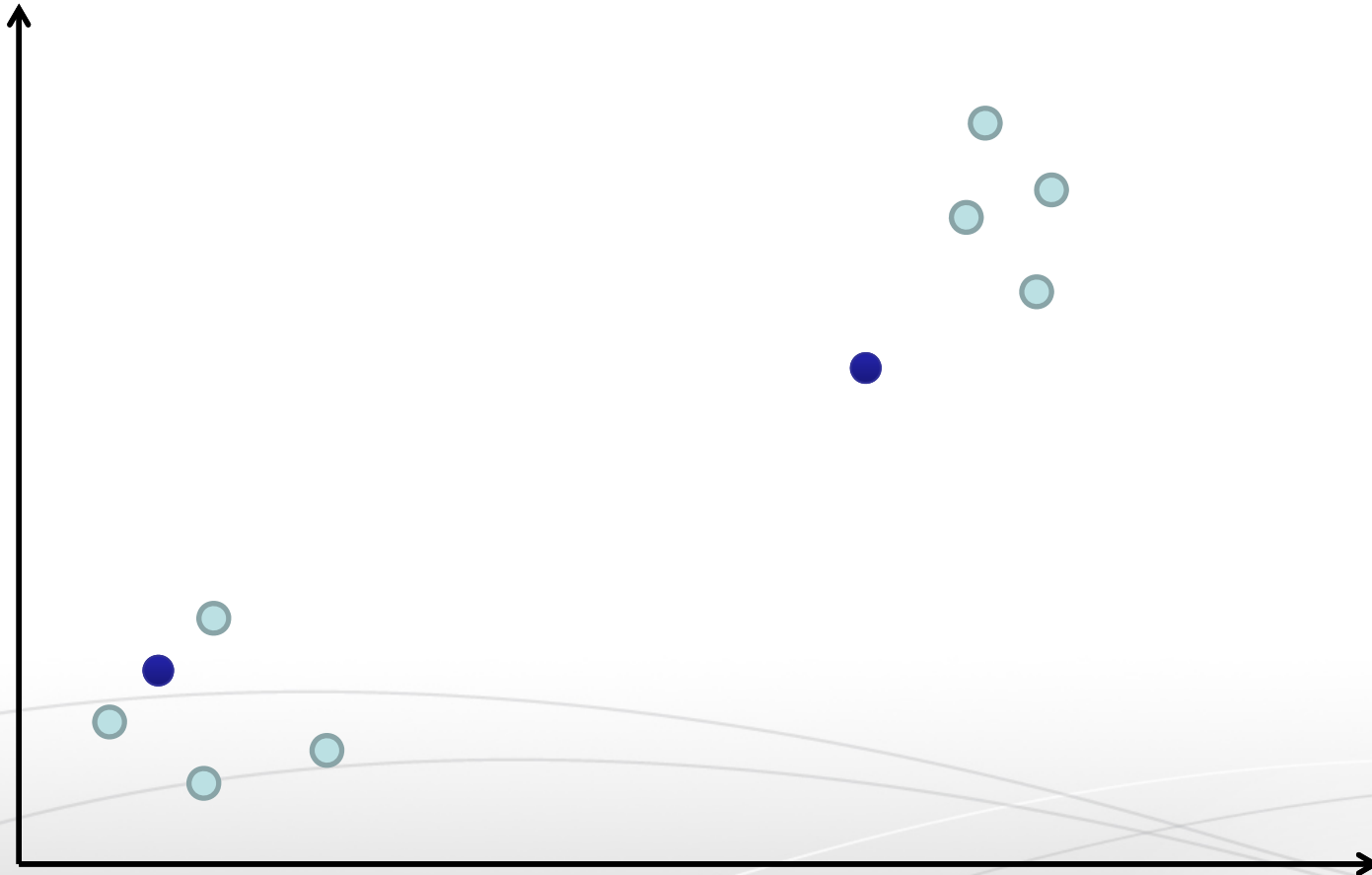
# K-Means Clustering Example



# K-Means Clustering Example

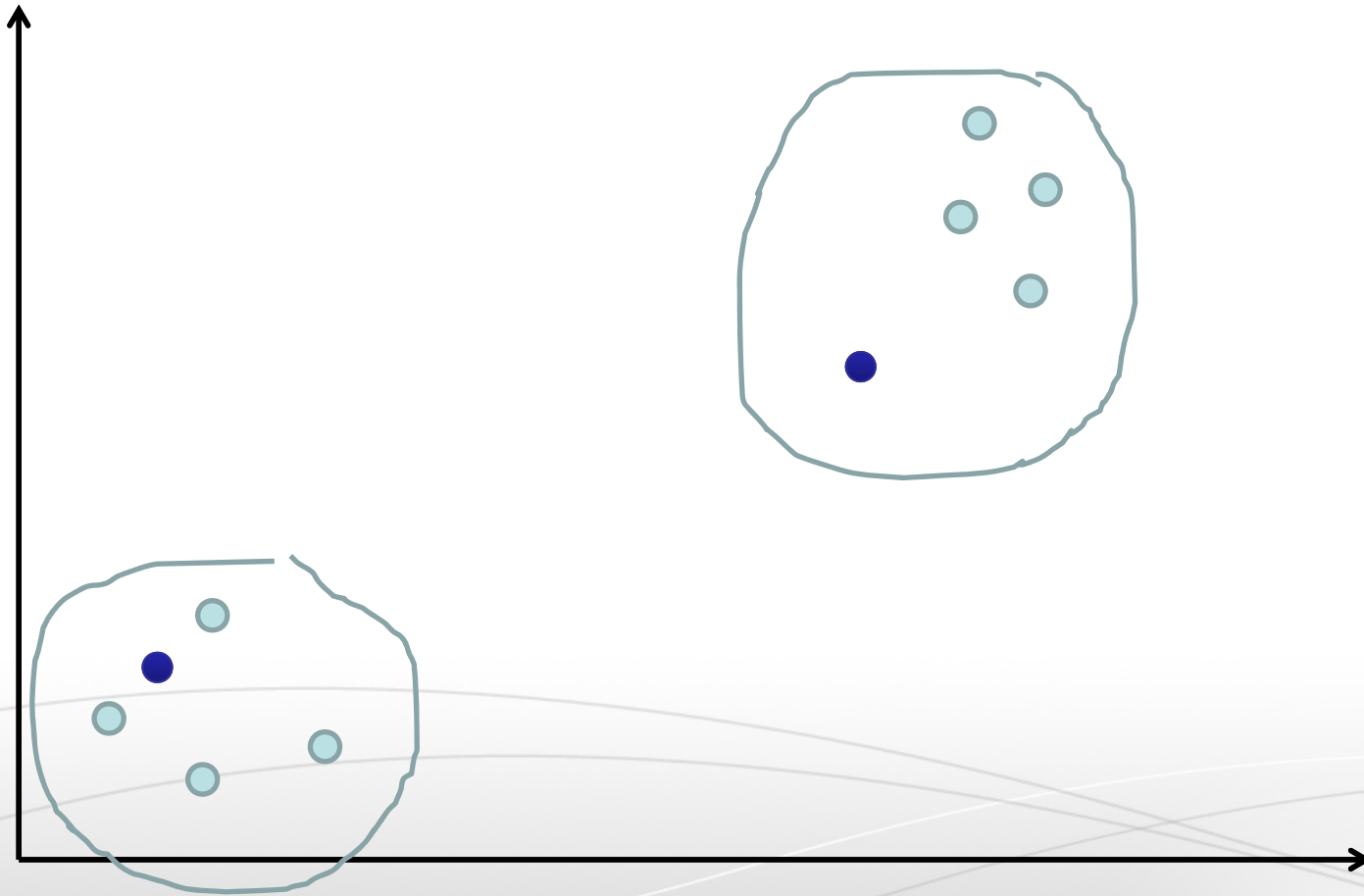


# K-Means Clustering Example

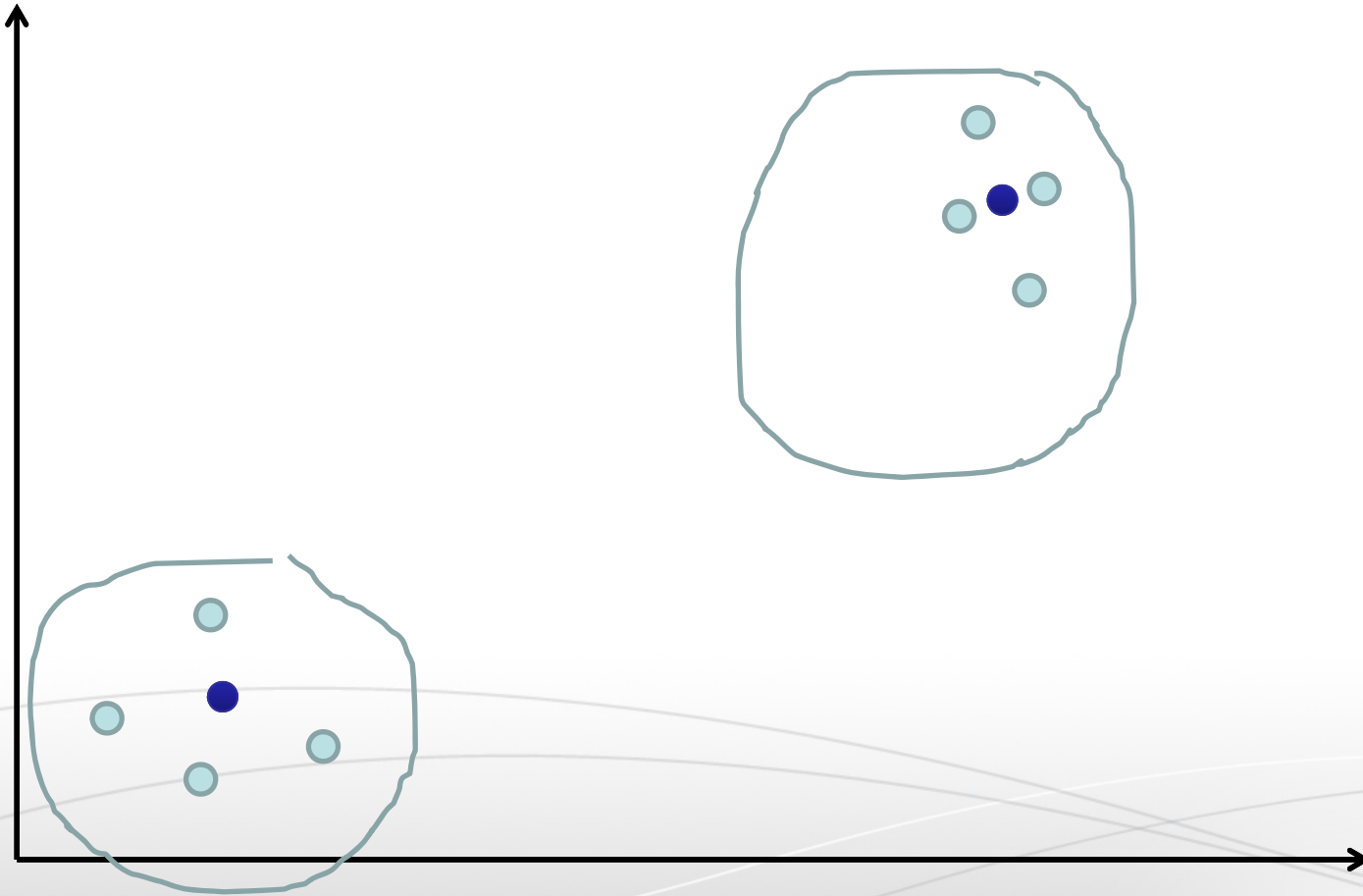




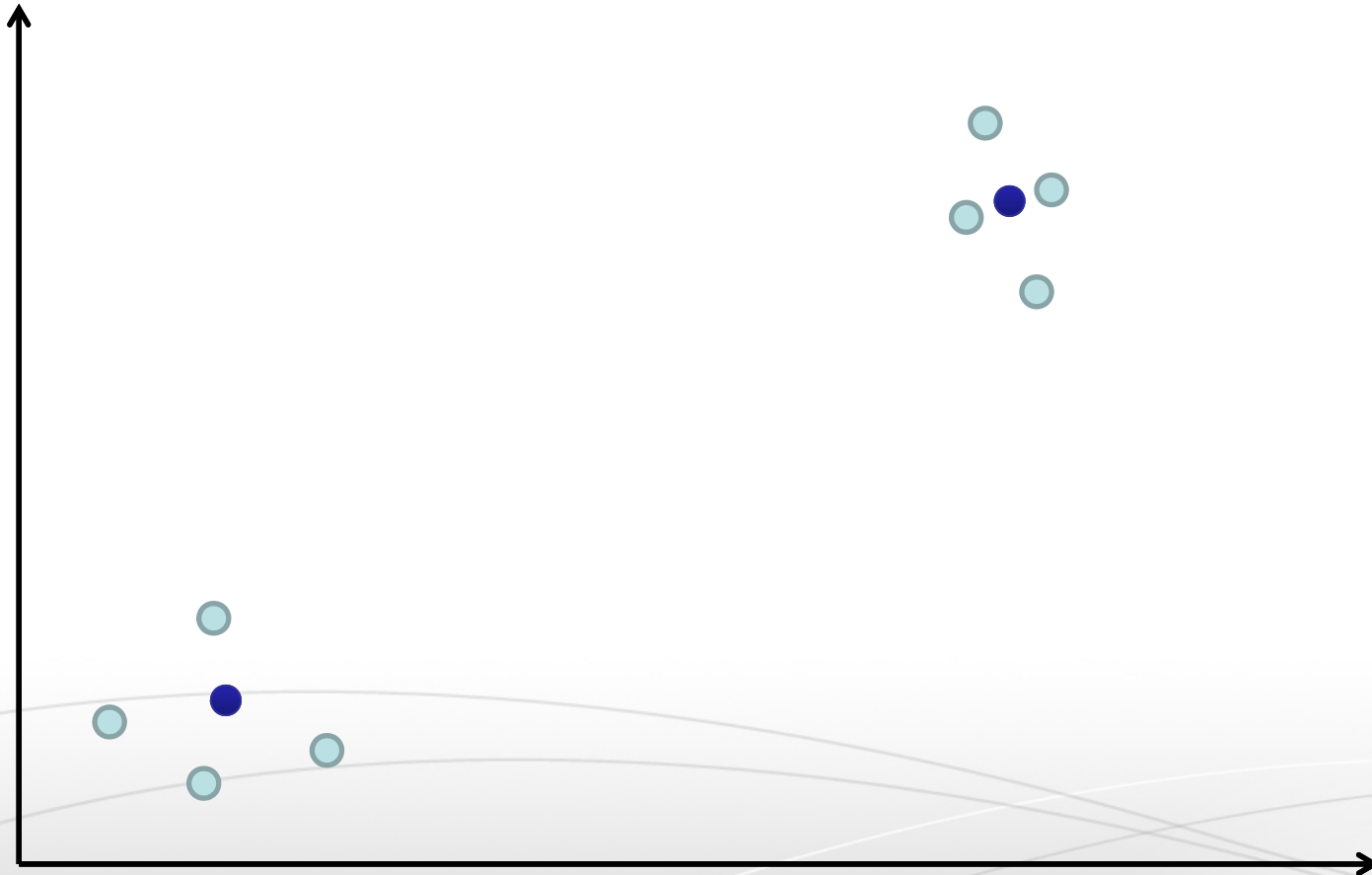
# K-Means Clustering Example



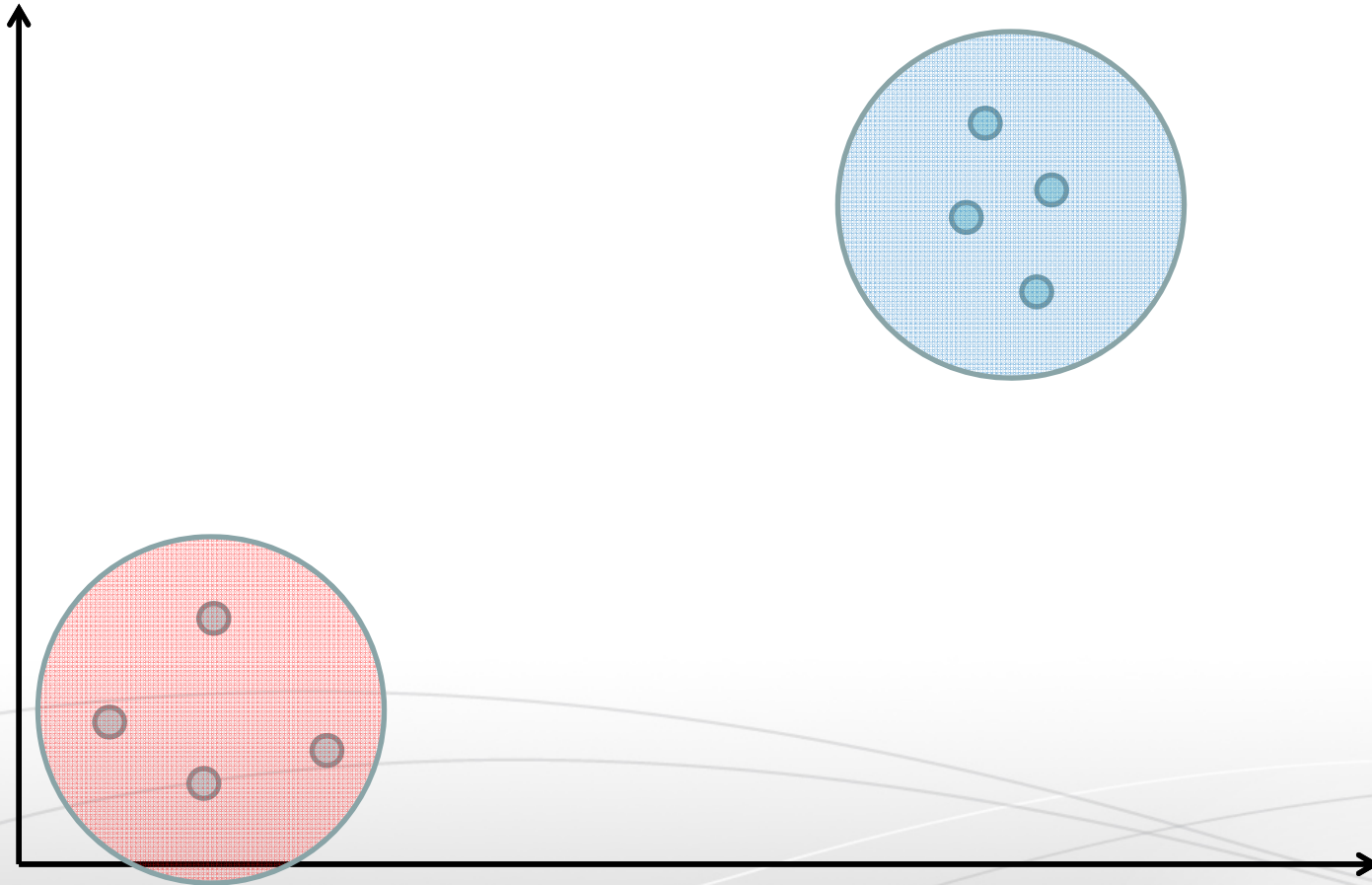
# K-Means Clustering Example



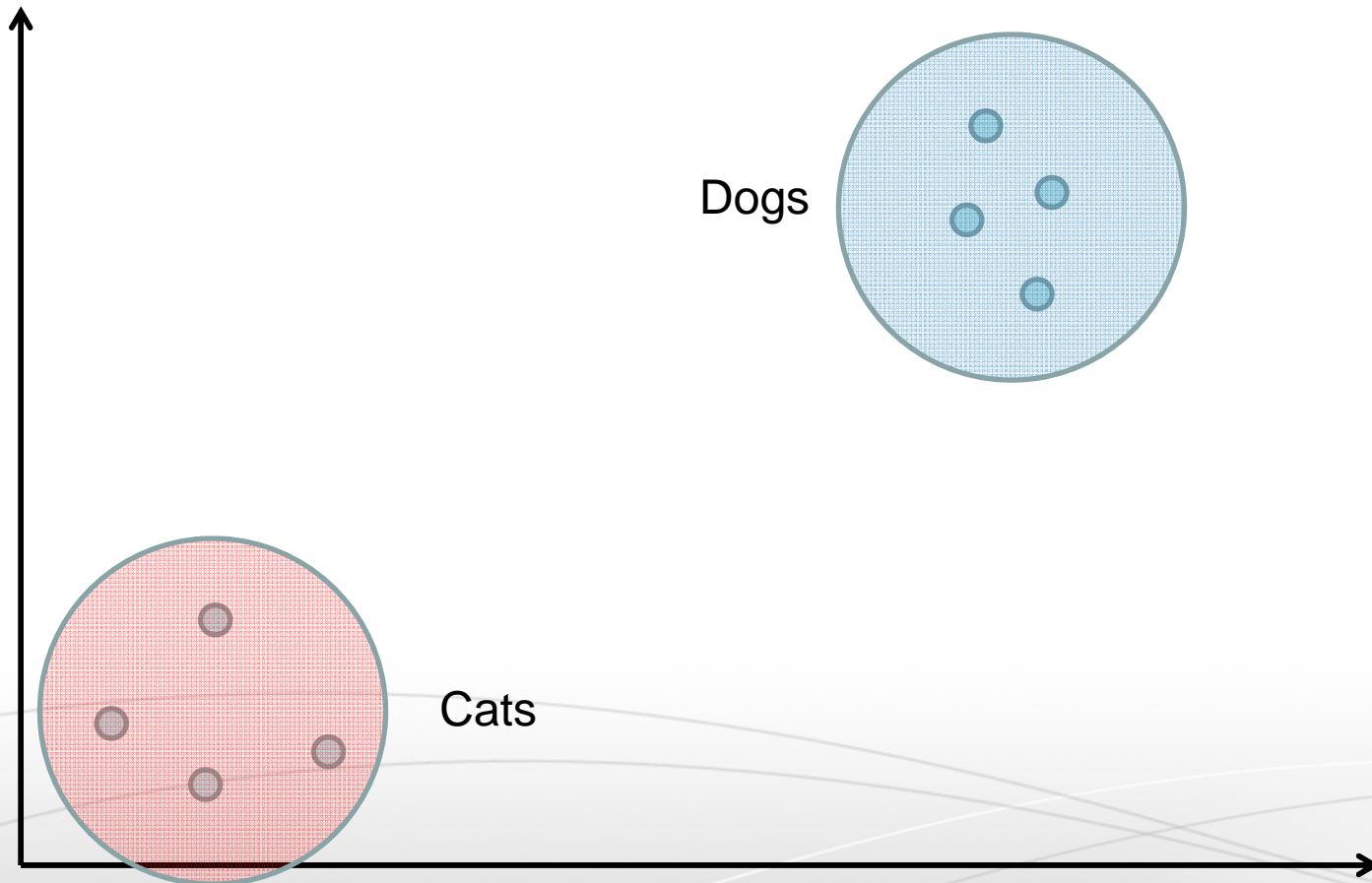
# K-Means Clustering Example



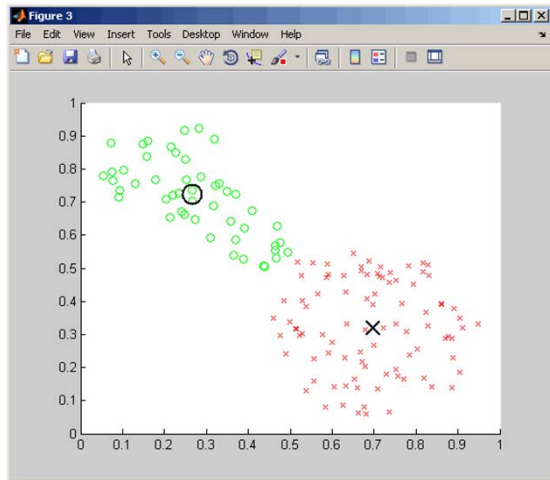
# K-Means Clustering Example



# K-Means Clustering Example



# K-Means Clustering in Mahout



+

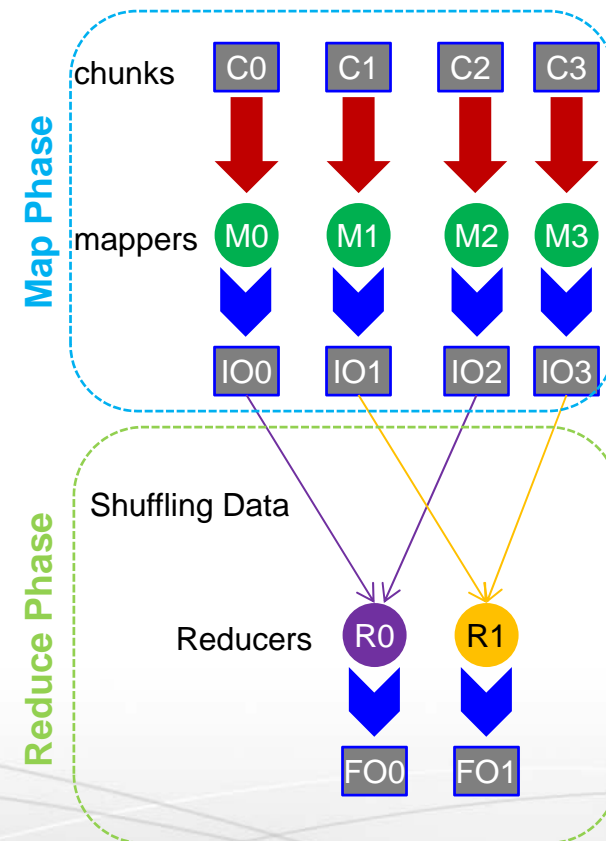


Figure from lecture 6: MapReduce

# K-Means Clustering in Mahout

+ Assume: # clusters  $\ll$  # points

# K-Means Clustering in Mahout

+ Assume: # clusters  $\ll$  # points





# K-Means Clustering in Mahout

+ Assume: # clusters  $\ll$  # points

M0 M1 M2 M3

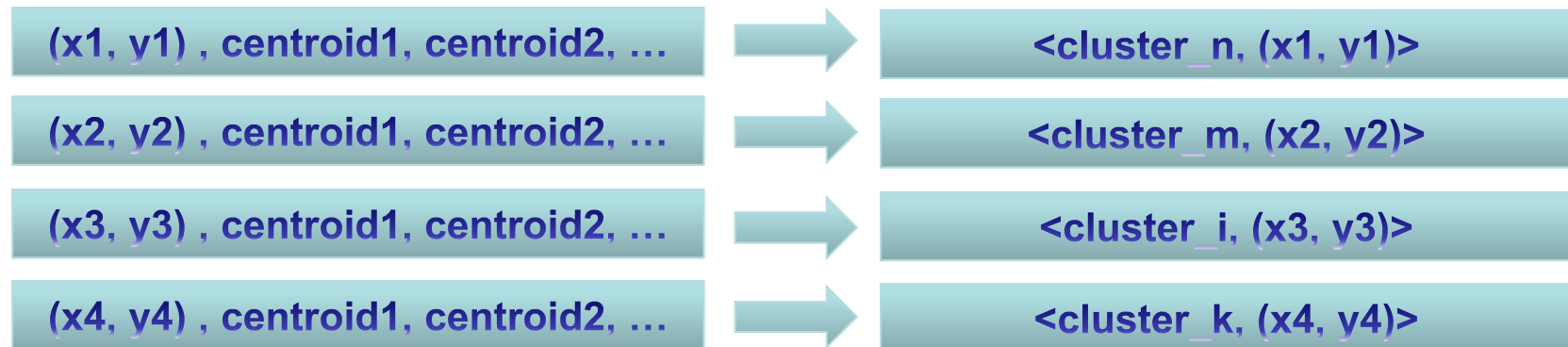
<clusterID, observation>



R0 R1

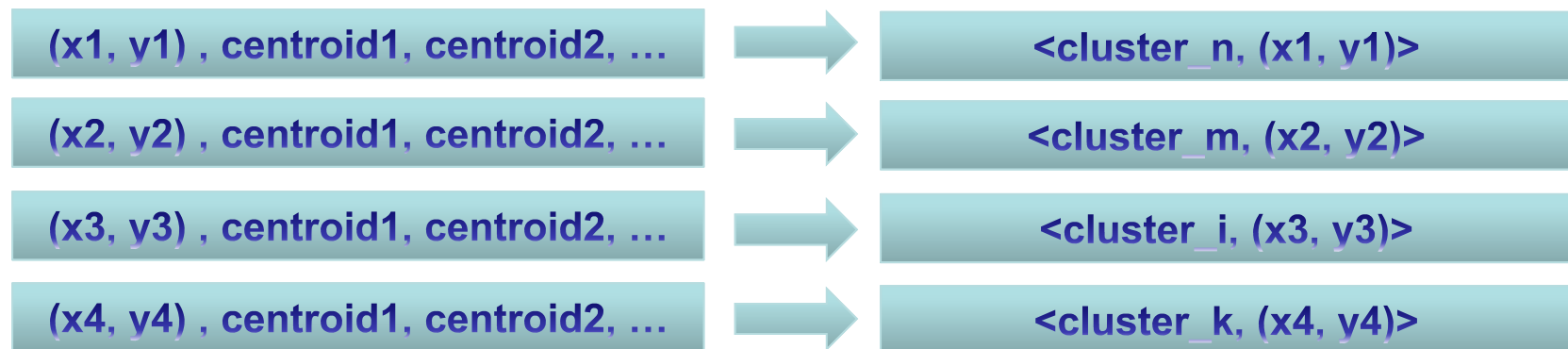
# K-Means Clustering in Mahout

+ Map phase: assign cluster IDs

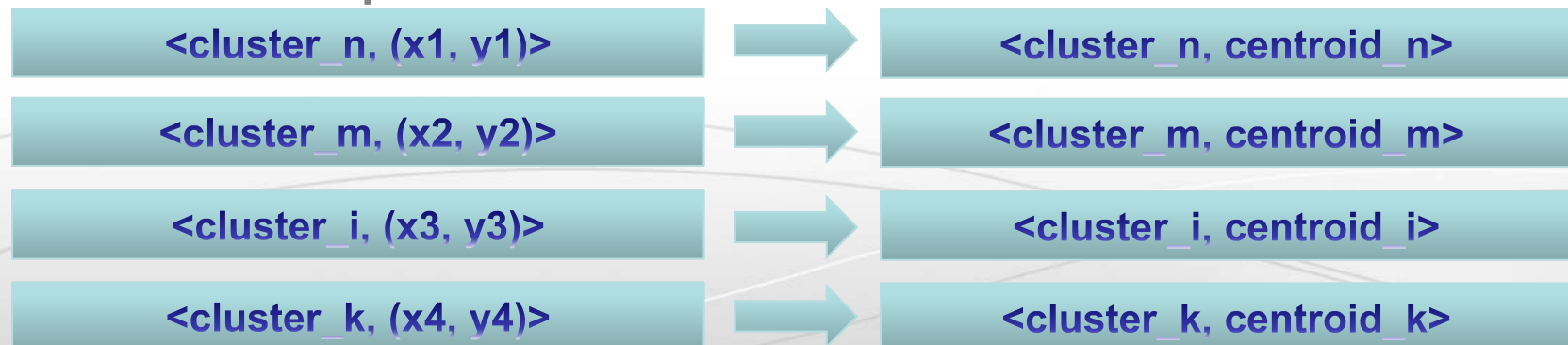


# K-Means Clustering in Mahout

+ Map phase: assign cluster IDs



+ Reduce phase: reset centroids



# K-Means Clustering in Mahout

- + Important notes
  - + --maxIter
  - + --convergenceDelta
  - + method

# Other Clustering Algorithms

- + Latent Dirichlet Allocation
  - + Topic models

# Other Clustering Algorithms

- + Latent Dirichlet Allocation
  - + Topic models
- + Fuzzy K-Means
  - + Points are assigned multiple clusters

# Other Clustering Algorithms

- + Latent Dirichlet Allocation
  - + Topic models
- + Fuzzy K-Means
  - + Points are assigned multiple clusters
- + Canopy clustering
  - + Fast approximations of clusters


# Other Clustering Algorithms

- + Latent Dirichlet Allocation
  - + Topic models
- + Fuzzy K-Means
  - + Points are assigned multiple clusters
- + Canopy clustering
  - + Fast approximations of clusters
- + Spectral clustering
  - + Treat points as a graph



# Other Clustering Algorithms

- + Latent Dirichlet Allocation
  - + Topic models
- + Fuzzy K-Means
  - + Points are assigned multiple clusters
- + Canopy clustering
  - + Fast approximations of clusters
- + Spectral clustering
  - + Treat points as a graph



K-Means &  
Eigencuts

# Mahout in Summary



# Mahout in Summary

+ Scalable library



# Mahout in Summary

- + Scalable library
- + Three primary areas of focus



# Mahout in Summary

- + Scalable library
- + Three primary areas of focus
- + Other algorithms



# Mahout in Summary



- + Scalable library
- + Three primary areas of focus
- + Other algorithms
- + All in your friendly neighborhood MapReduce

# Mahout in Summary



<http://mahout.apache.org/>

- + Scalable library
- + Three primary areas of focus
- + Other algorithms
- + All in your friendly neighborhood MapReduce

# Thank you!

