ACSIP- Association of Chinese Senior IT Professionals

Big Data Event

Brought you By

ACSIP & IBM Big Data University

2015-12-06

Join ACSIP WeChat Group by scanning QR Code with a note "ACSIP Group" for invitation

ACSIP IT 资深人协会

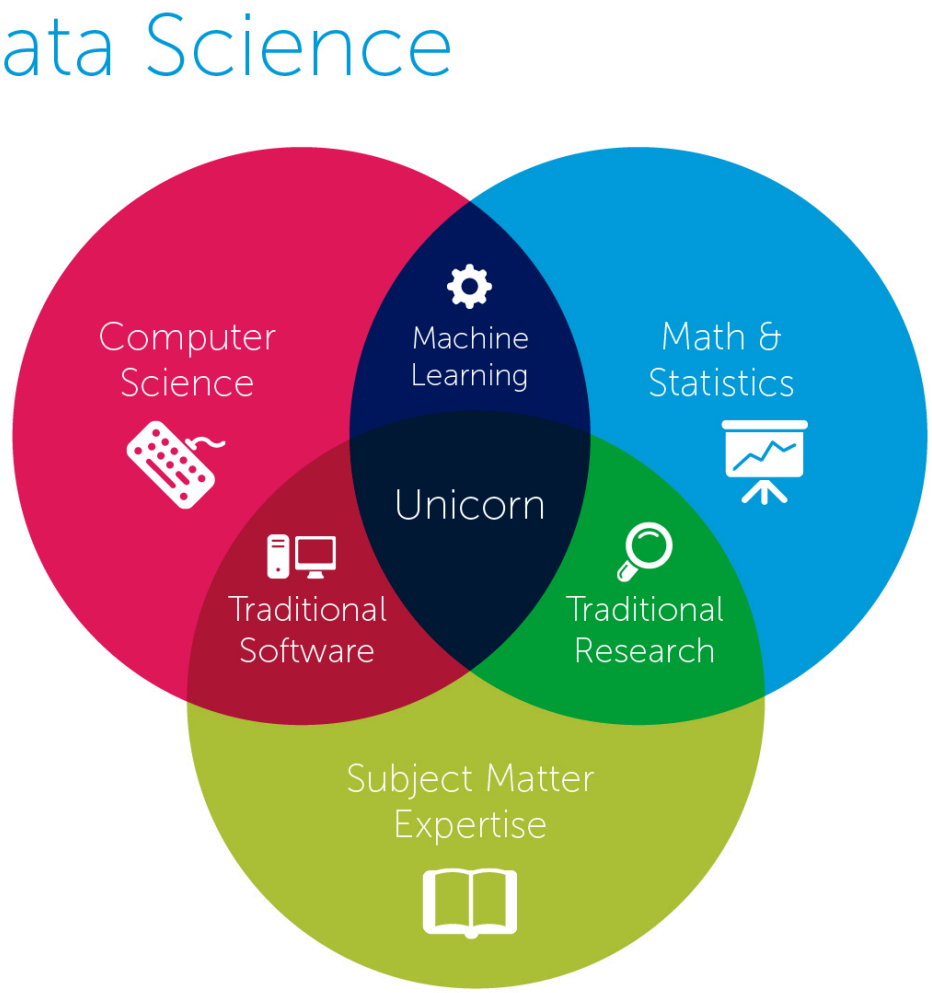
A Platform For Senior IT Professionals

ACSIP Board Of Directors & IBM Speakers

Join ACSIP WeChat Group by scanning QR Code with a note "ACSIP Group" for invitation

ACSIP
IT 资深人协会
A Platform
For Senior IT Professionals

WHAT IS DATA SCIENCE?



DATA SCIENCE METHODOLOGY

Business Understanding

Analytic Approach

Data Requirements

Data Collection

Data Understanding

Prediction
Interpretation
Justification
Testing

Feedback

Deployment

Evaluation

Modelling

Data Preparation

MAJOR ANALYTICAL APPROACHES & ALGORITHMS

- Associations
  - E.g. frequent co-occurring Items
  - Algorithms: a priori association rules
- Classification
  - E.g. prediction of item class
  - Algorithms: decision trees (ID3, C4.5, C5.0), CART, SVM, NN, Naïve Bayes, CHAID
- Estimation/Prediction
  - Predicting a continuous value
  - Algorithms: regression, SVM (support vector machines), KNN (K-nearest neighbours)
- Clustering
  - E.g. Finding cluster of patients
  - Algorithms: k-means, Hierarchical Clustering

MAJOR ANALYTICAL APPROACHES & ALGORITHMS

- Sequence mining
  - E.g. Click-stream
  - Algorithms: Markov Model, HMM
- Dimension Reduction
  - PCA
- Visualization
  - To facilitate human discovery, understanding
- Summarization
  - Describing a group
- Deviation Detection
  - Finding changes
- Link/Graph Analysis
  - Finding relationship

DATA SCIENCE METHODOLOGY

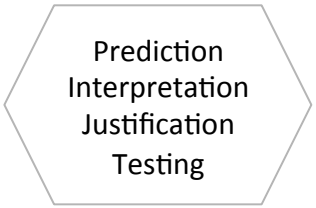
Business Understanding

Analytic Approach

Data Requirements

Data Collection

Data Understanding



Feedback

Deployment

Evaluation

Modelling

Data Preparation

DEMO:

Spark x Weather Data
x Exploratory Data Analysis

CLUSTER ANALYSIS

Business Understanding

Cluster Analysis

Analytic Approach

Data Requirements

Data Collection

Data Understanding

Prediction
Interpretation
Justification
Testing

Feedback

Deployment

Evaluation

K-means

Modeling

Data Preparation

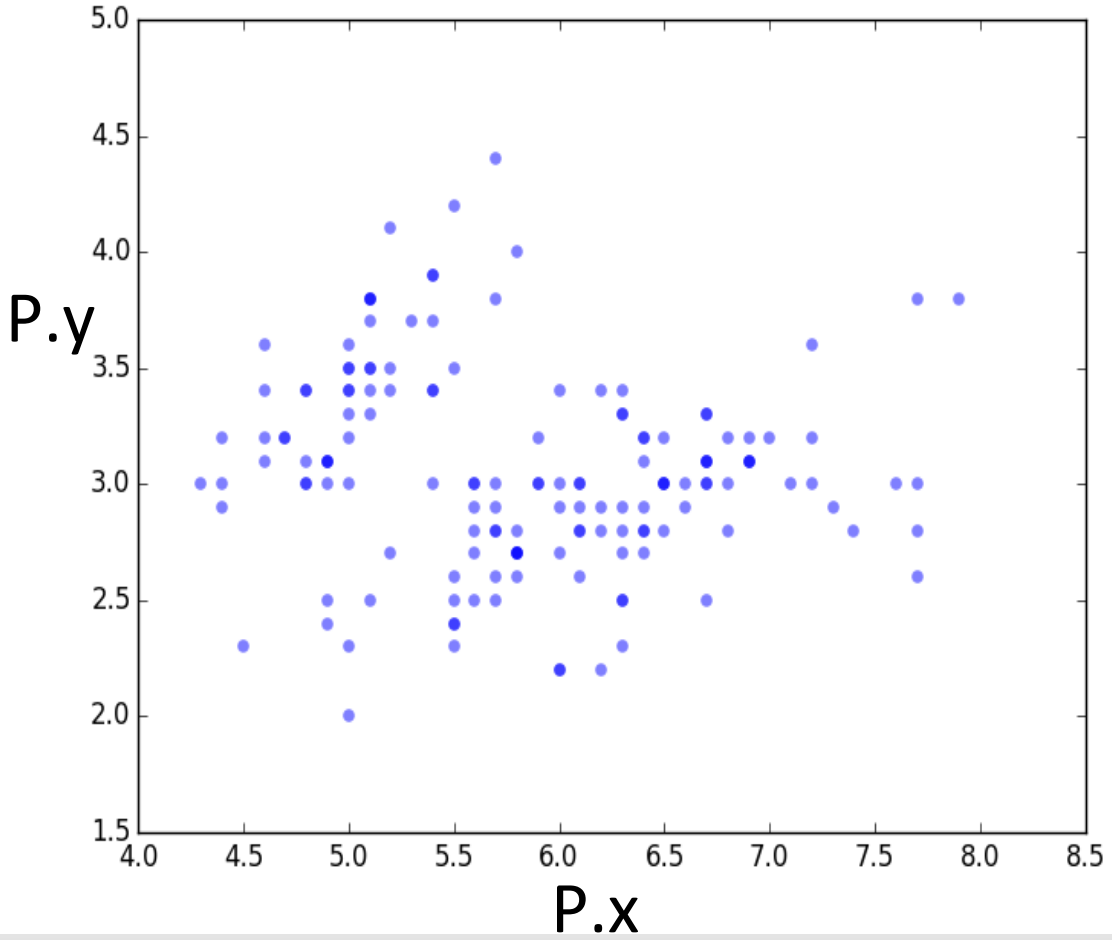
CLUSTER ANALYSIS

- What is a cluster?
  - A group of data points that are similar to other data points in the cluster, and dissimilar to data points in other clusters

CLUSTERING

Point1(2,3)
Point1(3,2)
Point1(1,1)
…

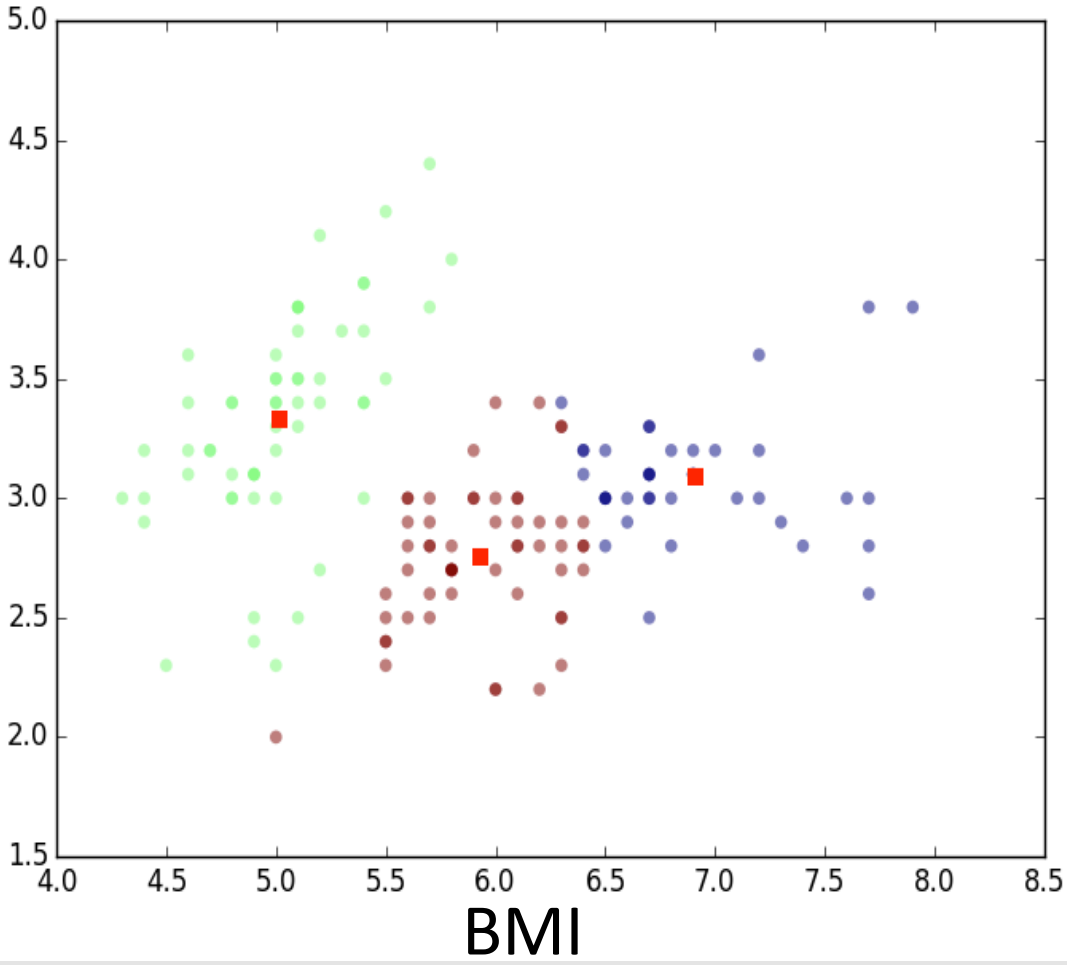
P1.x	P2.y
2	3
3	2
1	1
1	2
…	…
…	…



CLUSTERING

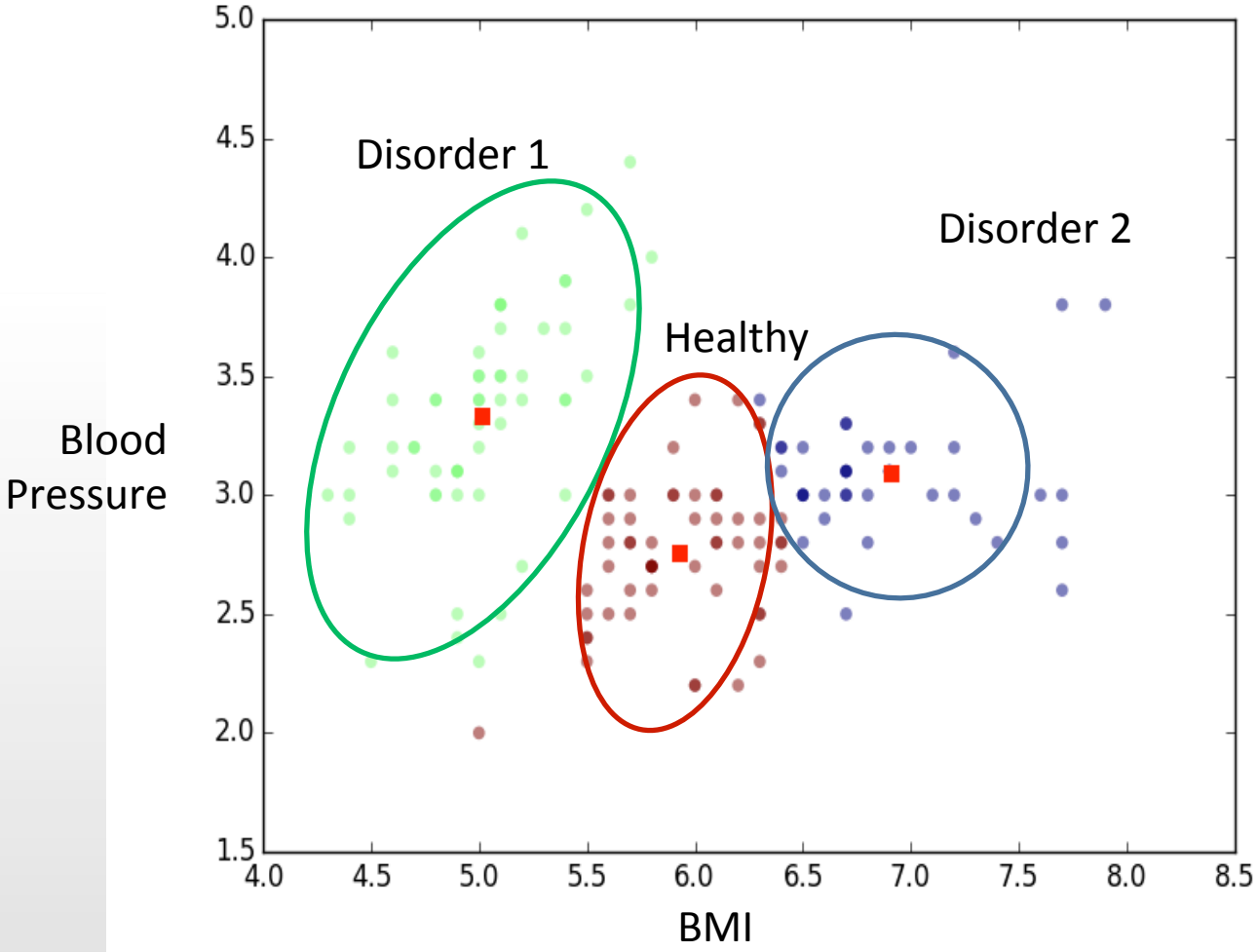
- Unsupervised learning

Blood Pressure



CLUSTERING

- What is the usage of Clustering?



CLUSTERING APPLICATIONS

- Example:
  - Finding cluster of customers based on their demographic characteristics, preferences, or their buying patterns
    - To recommend a new book, or to new customer by identifying clusters of books or clusters of customer preferences

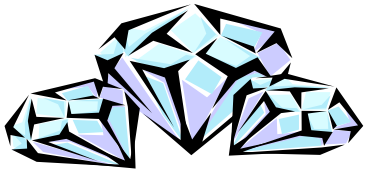
CLASSIFICATION VS CLUSTERING

BMI	Age	BP	…	Label
2	3	3	…	?
1	4	1	…	?
1	2	2	…	?
…	…	…	…	..

← Cluster A

← Cluster B

← Cluster B

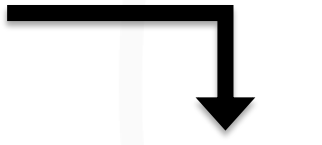


K-means

CLASSIFICATION VS CLUSTERING

Labeled dataset:

BMI	Age	BP	…	Label/ Actual val
2	3	3	…	disorder1
1	4	1	…	disorder2
1	2	2	…	healthy
…		…	…	…



Decision Tree

BMI	Age	BP	…	Label
4	3	3	…	?
5	2	1	…	?
1	4	2	…	?
…	…	…	…	…

← disorder2

← healthy

← disorder1

Testing dataset:

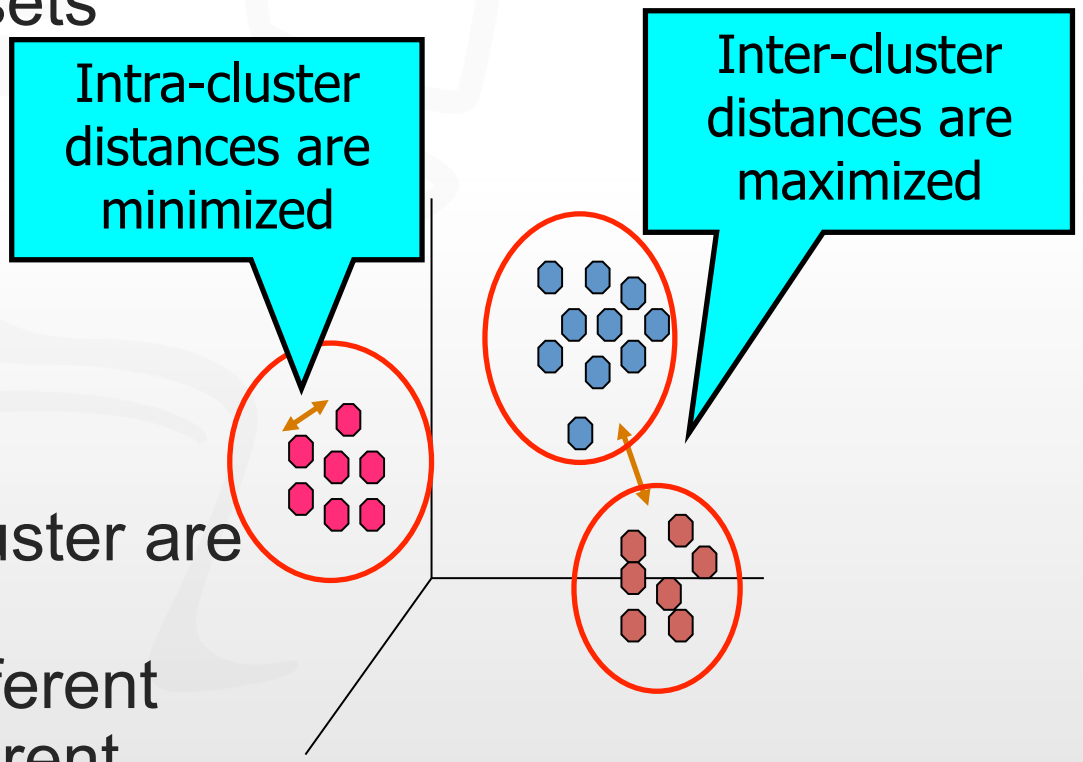
CLUSTERING ALGORITHMS

- **What kind of clustering algorithms?**
- k-means Clustering
- Fuzzy Clustering
- Hierarchical Clustering
- Density based Clustering

K-MEANS

what is k-Means?

- Partitioning Clustering
- K-Means divides the data into non-overlapping subsets (clusters) without any internal structure

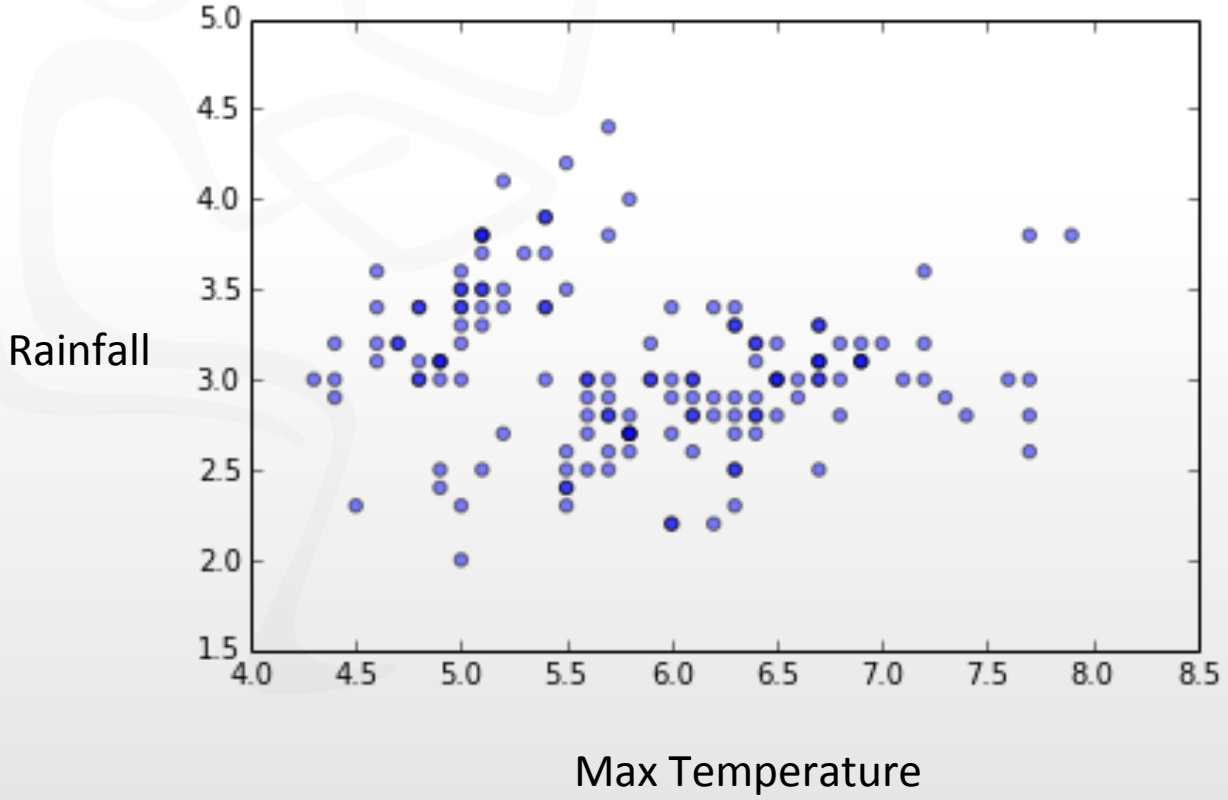


- Examples within a cluster are very similar
- Examples across different clusters are very different

K-MEANS CLUSTERING

1) Load your data

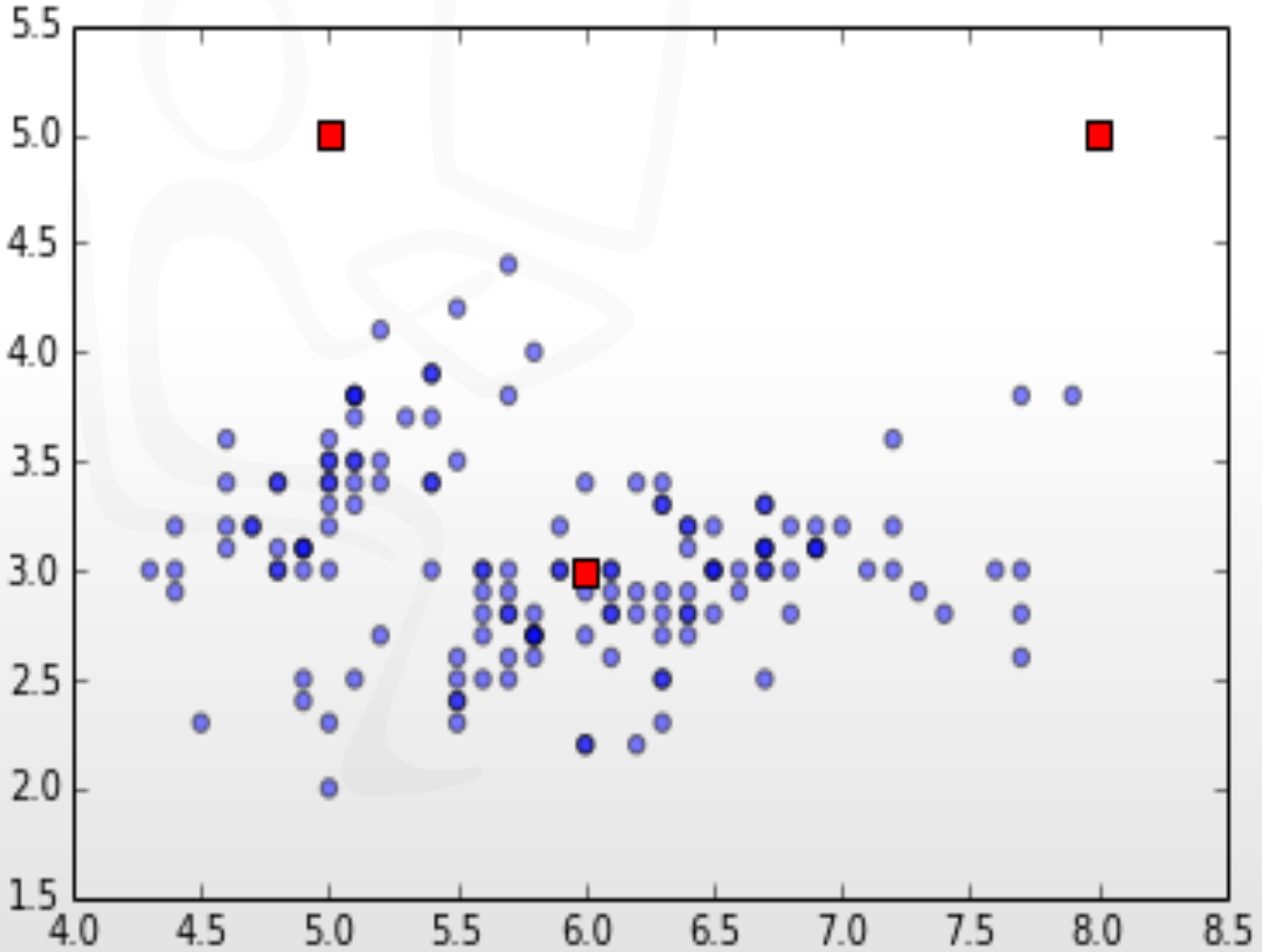
Rainfall	Max Temp
3 mm	4
2mm	6
3.5mm	2
…	..



K-MEANS CLUSTERING

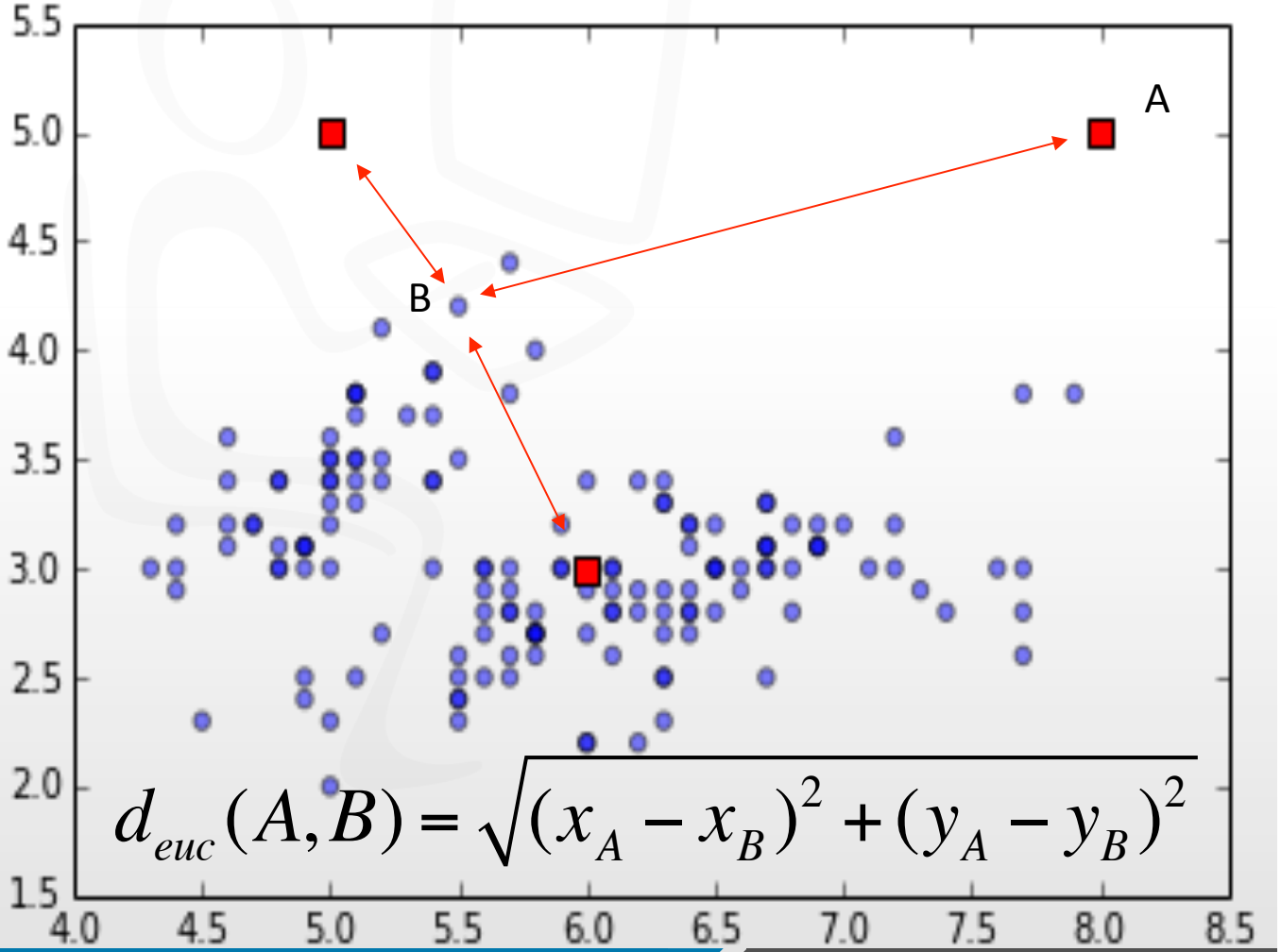
2) Initialize k=3 centroids randomly

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



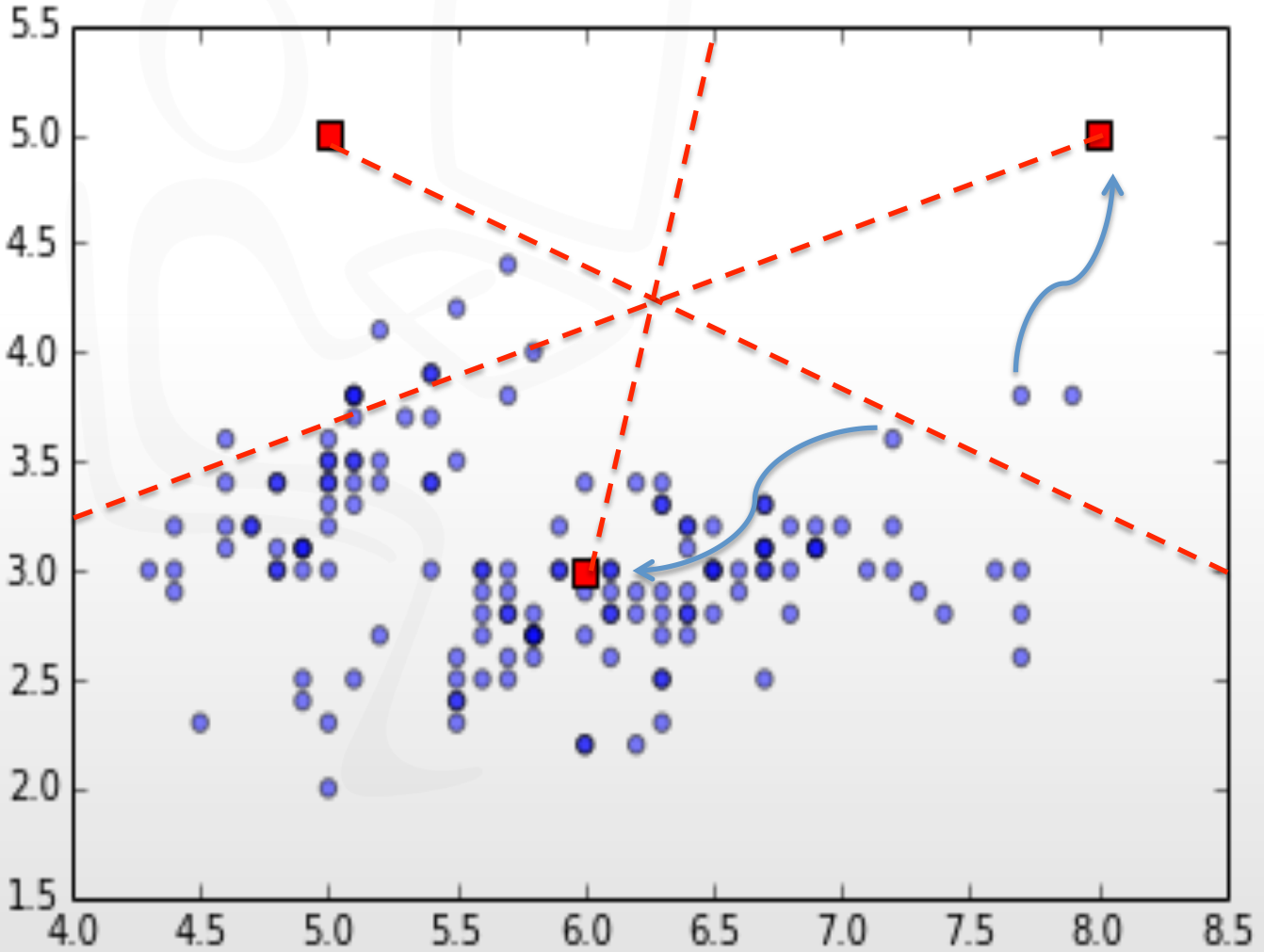
K-MEANS CLUSTERING

3) Calculate distance of all points from centroids
Similarity=1-distance



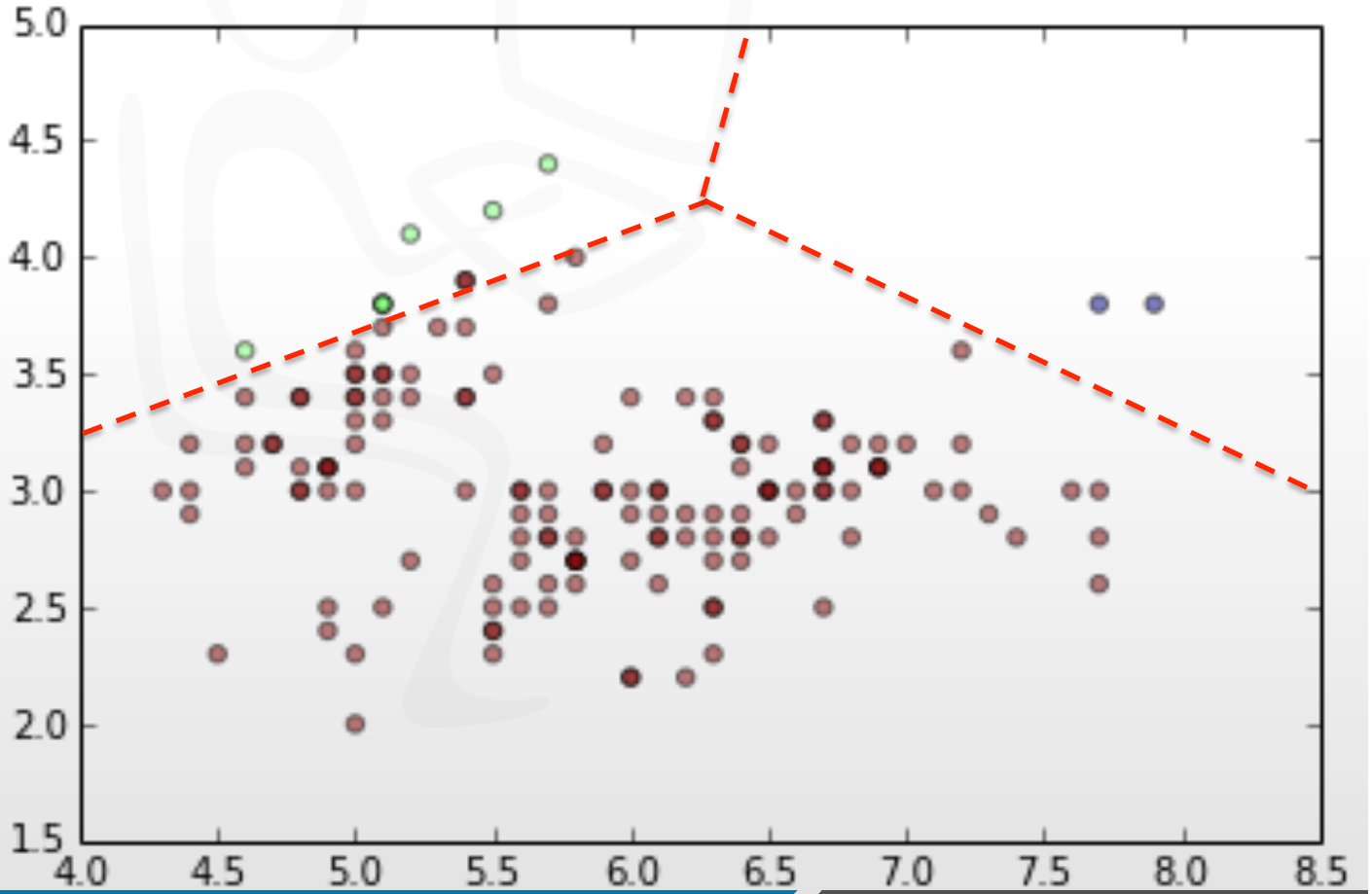
K-MEANS CLUSTERING

4) Find the closest center to each data point



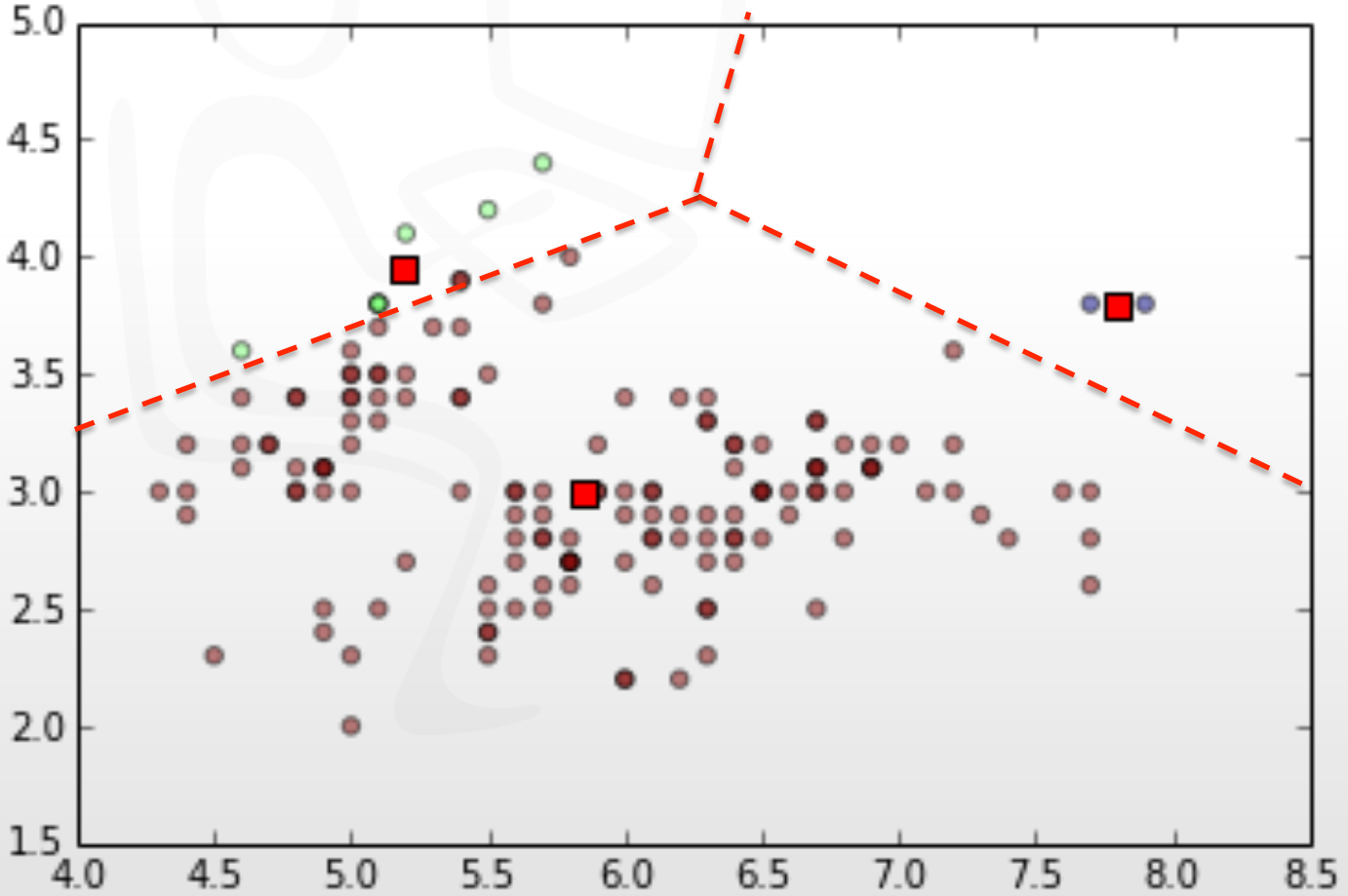
K-MEANS CLUSTERING

5) Assign each point to the closest centroid

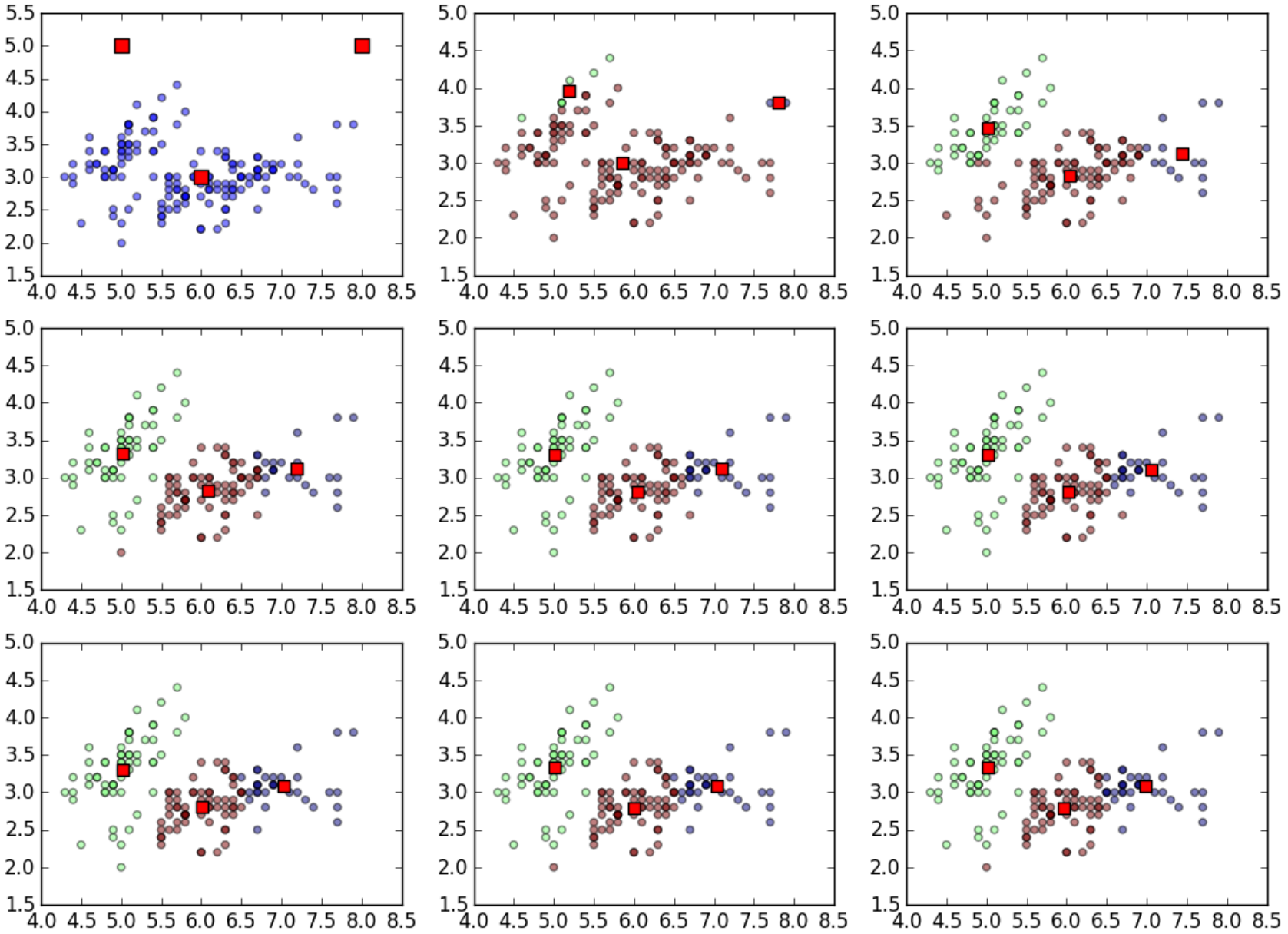


K-MEANS CLUSTERING

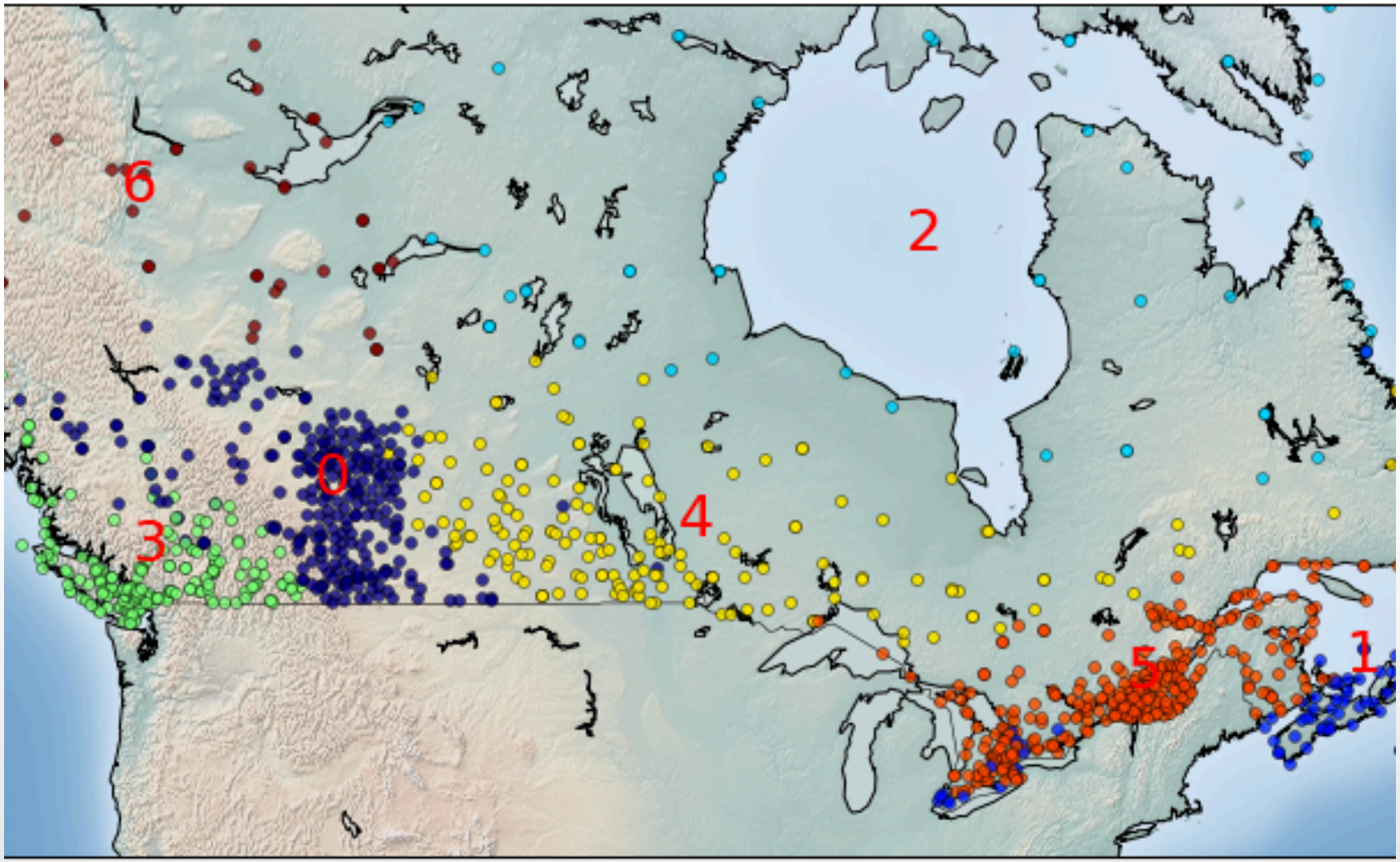
6) Compute the new centroids for each cluster.



7) Repeat until there are no more changes



DEMO: WEATHER STATION CLUSTERING



DEMO:

Spark x Weather Data x Cluster Analysis (MLlib)

CLASSIFICATION

Business Understanding

Classification

Analytic Approach

Data Requirements

Data Collection

Data Understanding

Prediction
Interpretation
Justification
Testing

Feedback

Deployment

Evaluation

Classification Trees

Modeling

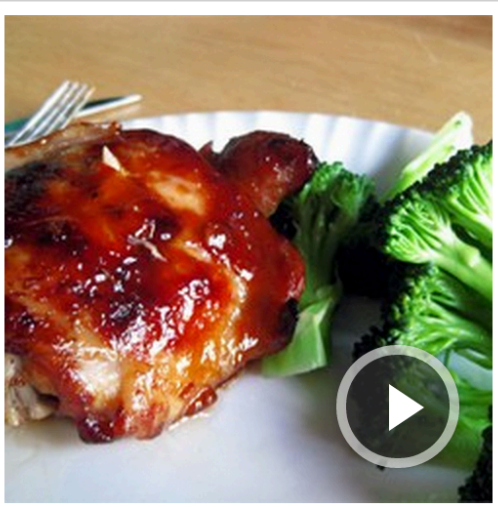
Data Preparation

BUSINESS UNDERSTANDING

Friend and I are looking at some food…

Friend: “How do you know this is Japanese food?”
Me: “Uh… because it has teriyaki sauce?”
Friend: “But how do you *know* that teriyaki = Japanese?”

"BUT HOW DO YOU *KNOW* THIS IS JAPANESE FOOD?"



See all 254 photos!

Baked Teriyaki Chicken

READY IN 1½ hrs

Read Reviews (3963)

Pin it 31K+ | Like 9.1k | Tweet 45 | +1 167

Recipe by Marian Collins

"A much requested chicken recipe! Easy to double for a large group. Delicious!"

+ Recipe Box | + Shopping List | + Menu | Email | Print

Ingredients Edit and Save

Original recipe makes 6 servings Change Servings

- 1 tablespoon cornstarch
- 1 tablespoon cold water
- 1/2 cup white sugar
- 1/2 cup soy sauce
- 1/4 cup cider vinegar
- 1 clove garlic, minced
- 1/2 teaspoon ground ginger
- 1/4 teaspoon ground black pepper
- 12 skinless chicken thighs

Check All | Add to Shopping List

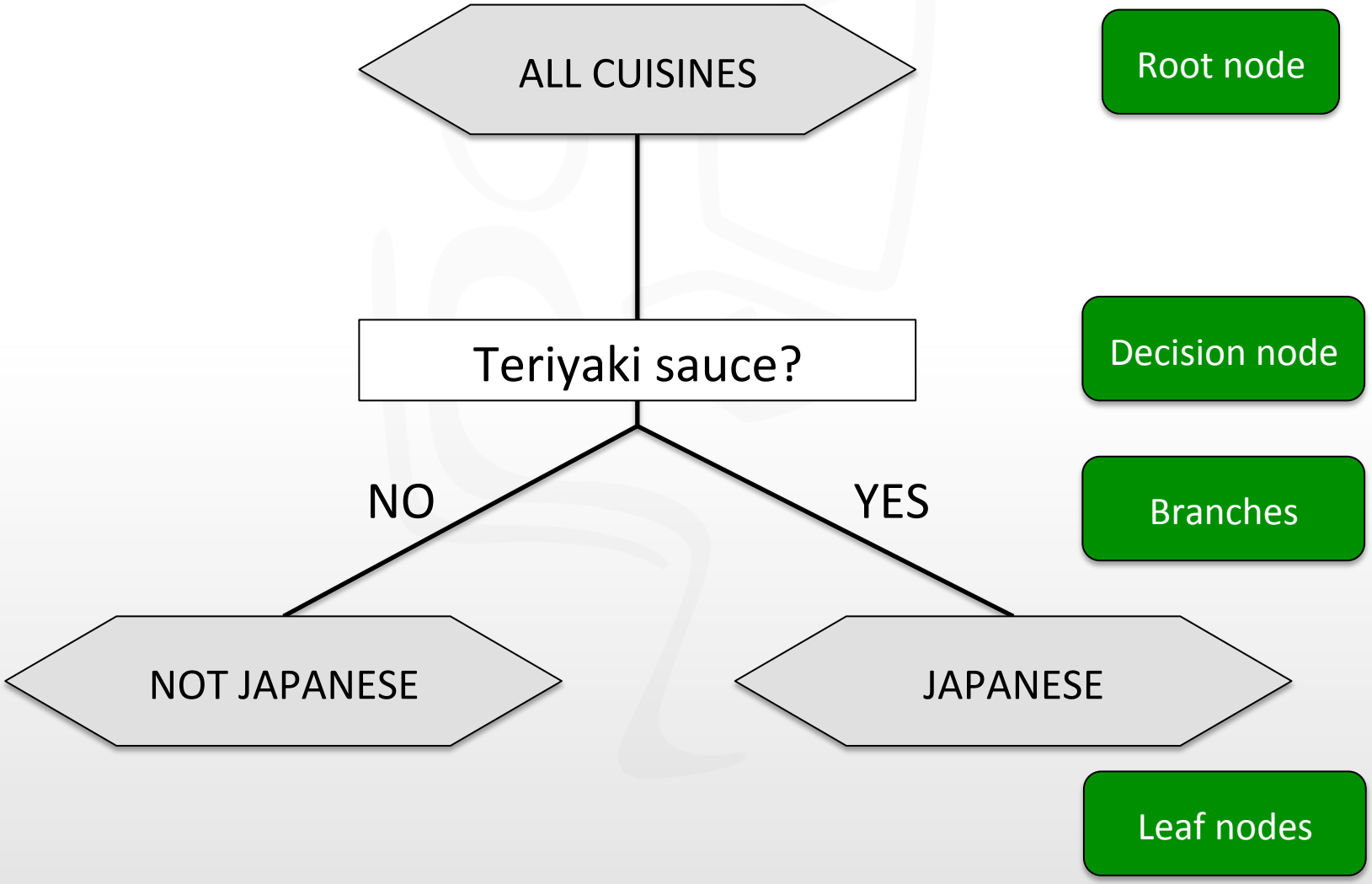
Watch video tips and tricks

Baked Teriyaki Chicken

Crispy and Tender Baked Chicken Thighs

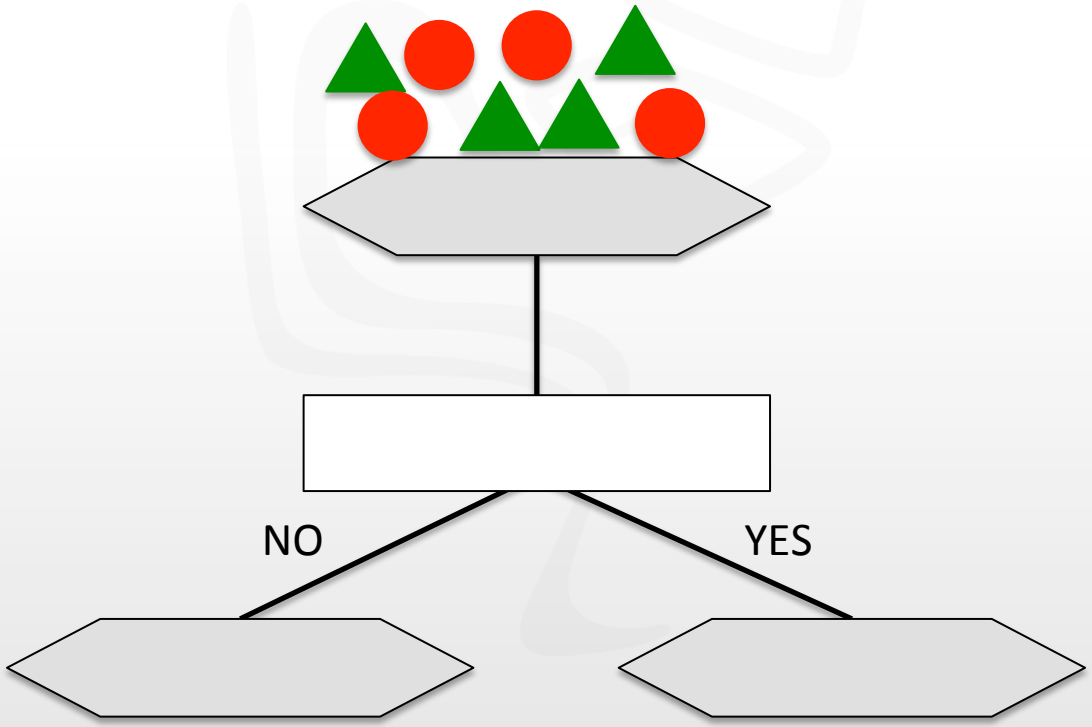
allrecipes.com

CLASSIFICATION - DECISION TREES



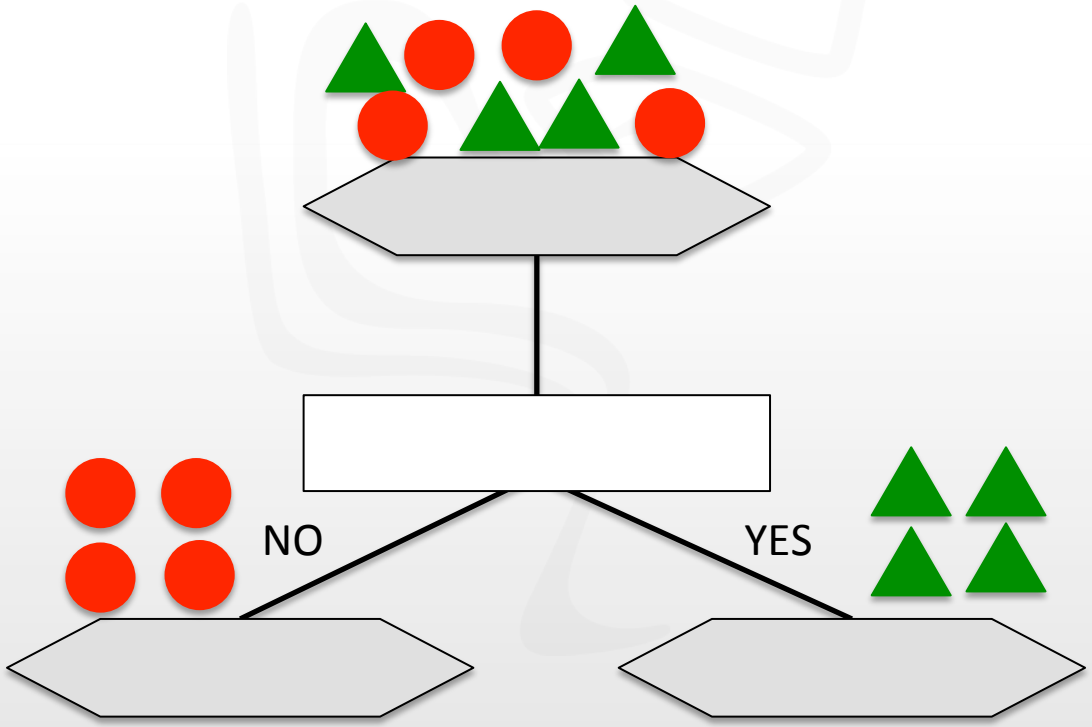
UNDERSTANDING DECISION TREES

- DTs are built using **recursive partitioning** to classify the data
- The algorithm chooses the most predictive feature to split the data on
- “predictiveness” is based on decrease in entropy (gain in information) or “impurity”



UNDERSTANDING DECISION TREES

- DTs are built using **recursive partitioning** to classify the data
- The algorithm chooses the most predictive feature to split the data on
- “predictiveness” is based on decrease in entropy (gain in information) or “impurity”



CHARACTERISTICS OF DECISION TREES

Pros	Cons
Easy to interpret	Easy to overfit or underfit the model
Can handle numeric or categorical features	Cannot model interactions between features
Can handle missing data	Large trees can be difficult to interpret
Uses only the most important features	
Can be used on very large or small data	

A tree stops growing at a node when…

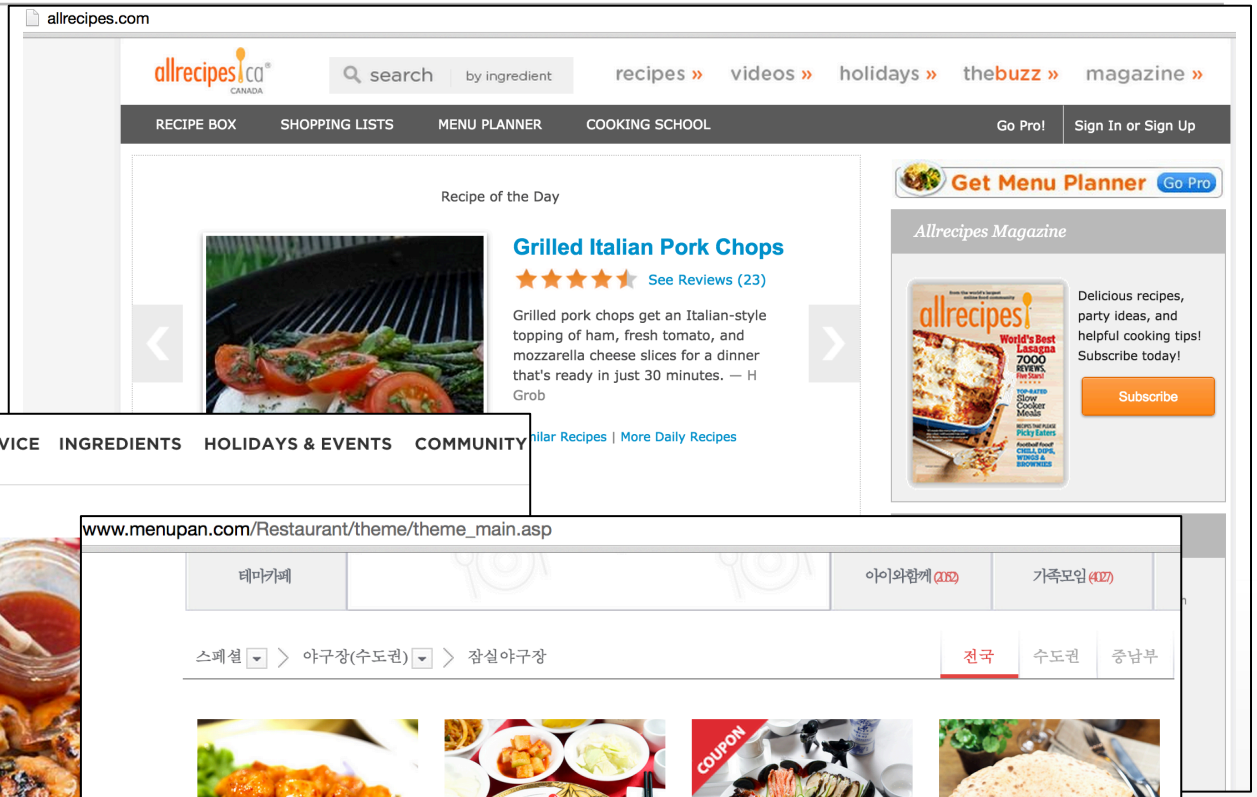
- pure or nearly pure
- no remaining variables on which to further subset the data
- the tree has grown to a preselected size limit

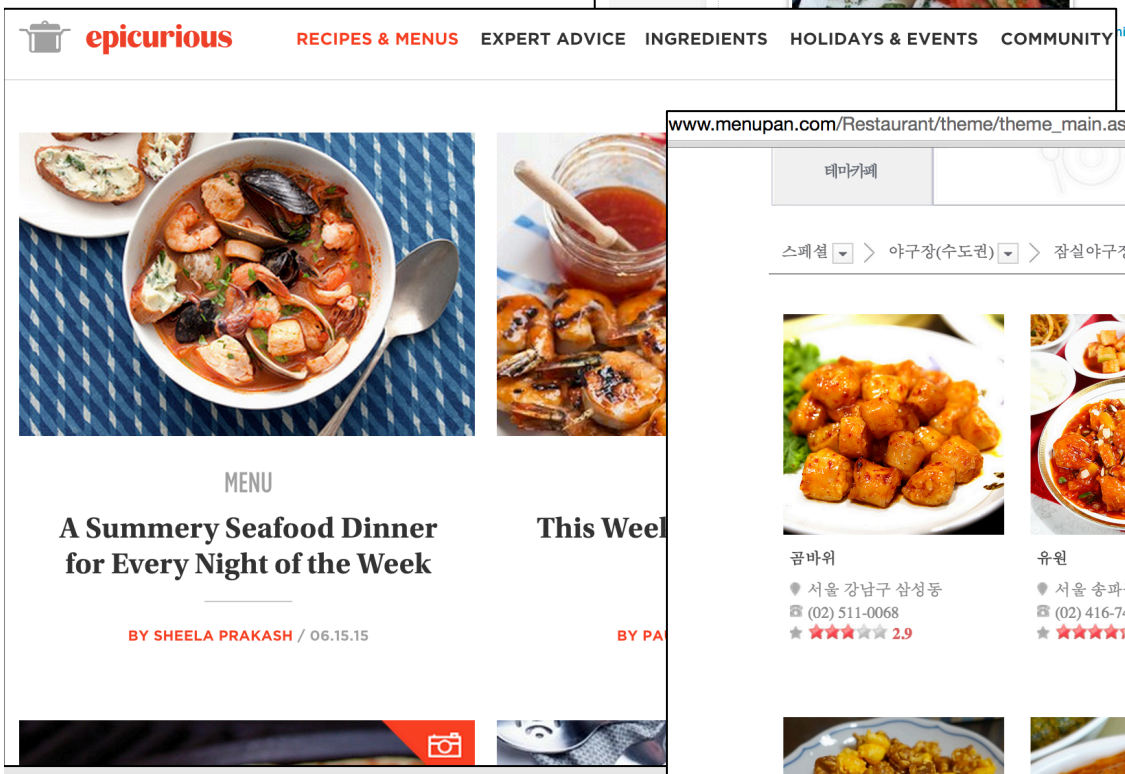
UNDERSTANDING THE ALGORITHMS IN DECISION TREES

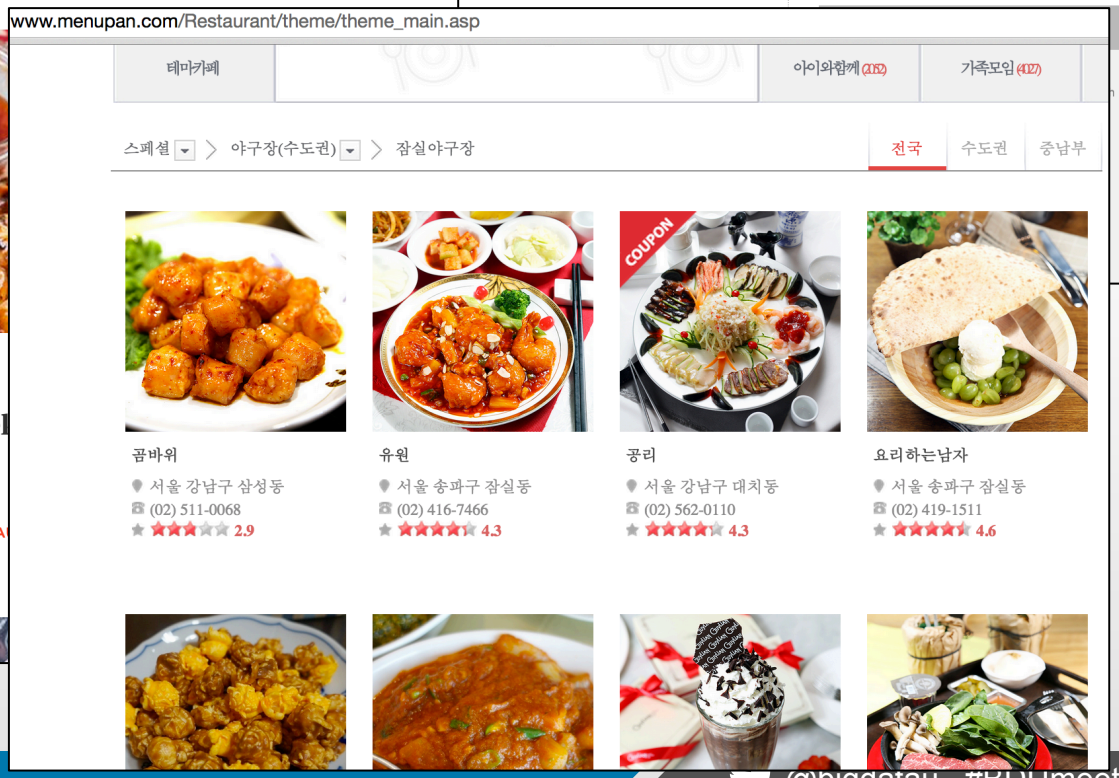
- How the `rpart` package in R uses recursive partitioning
  - http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf
- `scikit-learn` in Python
  - http://scikit-learn.org/stable/modules/tree.html
- "Machine learning – decision trees" by Professor Nando de Freitas
  - https://www.youtube.com/watch?v=-dCtJjlEEgM
- Datacamp's Kaggle R tutorial on Titanic survivorship
  - https://www.datacamp.com/courses/kaggle-tutorial-on-machine-learing-the-sinking-of-the-titanic

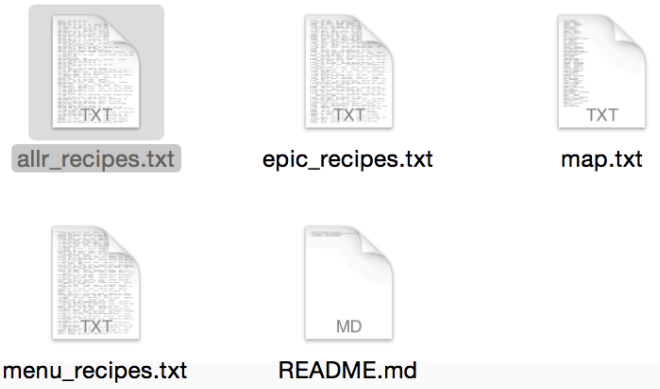
WEB-SCRAPE BY YONG-YEOL AHN

WWW.ALLRECIPES.COM
WWW.EPICURIOUS.COM
WWW.MENUPAN.COM









Goals:

- Combine three .txt files
- Convert to dataframe
- One row per recipe
- One column per ingredient
- Row values = “Yes” or “No”

allr_recipes.txt

```
Canada  egg     yeast   wheat   milk    lard
Canada  pork    carrot  pea     onion   potato
Canada  maple_syrup
Canada  wheat   yeast   almond  honey   oat     date    vegetable_oil   whole_grain_wheat_flour
Canada  butter  lovage  clam    wheat   onion   thyme   potato  yeast   black_pepper    parsley
ginger  bay     celery  cinnamon        milk    mustard
Canada  cane_molasses   butter  wheat   raisin  ginger  lard    egg     milk    cream
Canada  asparagus       olive_oil       pepper  garlic  tomato
Canada  butter  cilantro        tea     jasmine vegetable       brown_rice
Canada  vegetable_oil   wheat   egg     milk
Canada  lemon_juice     onion   soy_sauce       black_pepper    ginger  white_wine      garlic
vegetable_oil   chicken
Canada  butter  cane_molasses   wheat   vanilla egg     milk
Canada  butter  olive_oil       pepper  rice    mushroom        onion   chicken_broth   thyme
garlic  basil   porcini
Canada  tomato  onion   black_pepper    garlic  mozzarella_cheese       tuna    rice
Canada  wheat   lard
Canada  butter  pepper  lemon   beef    fish    parsley white_wine      basil   wheat   cream
Canada  butter  cane_molasses   wheat   vanilla cocoa   egg     milk
Canada  vegetable_oil   potato
Canada  tomato  olive_oil       pork    green_bell_pepper       onion   mushroom        thyme
white_wine      basil   garlic  oregano mozzarella_cheese       rosemary
Canada  cane_molasses   wheat   yeast   raisin  oat     cinnamon        lard
Canada  tomato  olive_oil       mushroom        black_pepper    garlic  oregano basil   rosemary
Canada  butter  coffee  wheat   vanilla cocoa   egg     banana
Canada  butter  cane_molasses   wheat   yeast   milk    whole_grain_wheat_flour
Canada  butter  onion   potato  chicken_broth   black_pepper    squash  celery  carrot
Canada  onion   mushroom        egg_noodle      black_pepper    celery  bell_pepper
mozzarella_cheese       chicken cheddar_cheese  broccoli
Canada  butter  pepper  wheat   onion   potato  asparagus       chicken_broth   black_pepper
```

DEMO:

Classifcation x Recipes Data

Thank you for listening!

Polong Lin
Leo Wu
Raul CHong

ACSIP- Association of Chinese Senior IT Professionals

Big Data Event

Brought you By

ACSIP & IBM Big Data University

2015-12-06

Join ACSIP WeChat Group by scanning QR Code with a note "ACSIP Group" for invitation

ACSIP IT 资深人协会

A Platform For Senior IT Professionals