DATASTAX

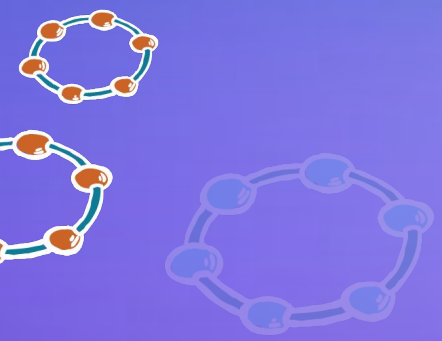
ACCELERATE

Lighting a Spark with Machine Learning

Amanda Moran
Developer Advocate, DataStax

DATASTAX®

WHY



DATASTAX®

The Importance of Machine Learning

Mat Velloso
@matvelloso

Following

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

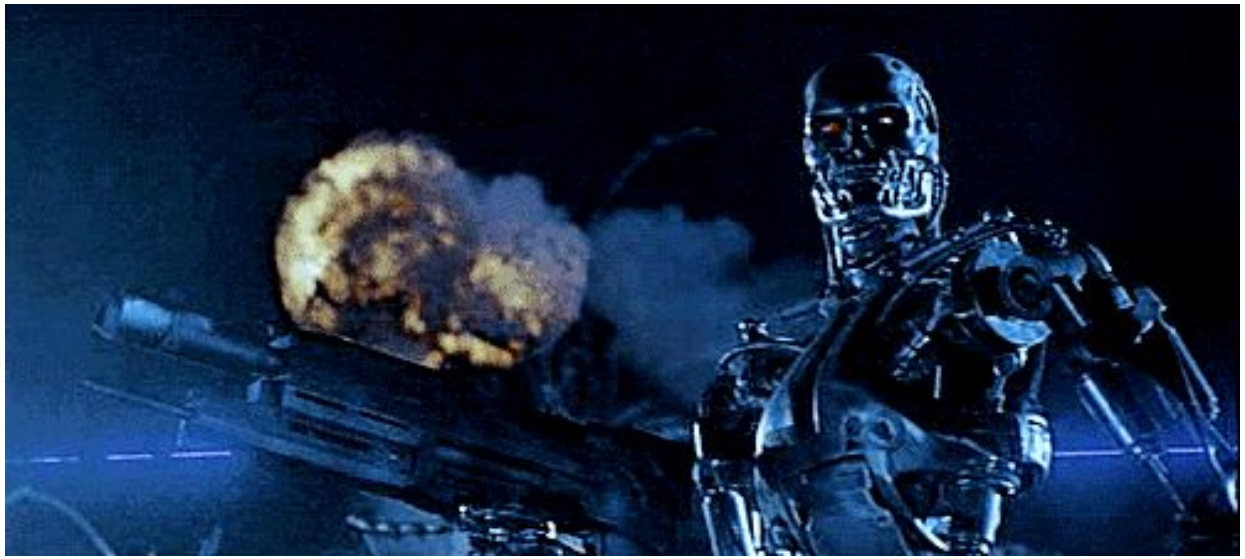
If it is written in PowerPoint, it's probably AI

5:25 PM - 22 Nov 2018

Confidential © DataStax, All Rights Reserved.

DATASTAX

Focused on the Practical



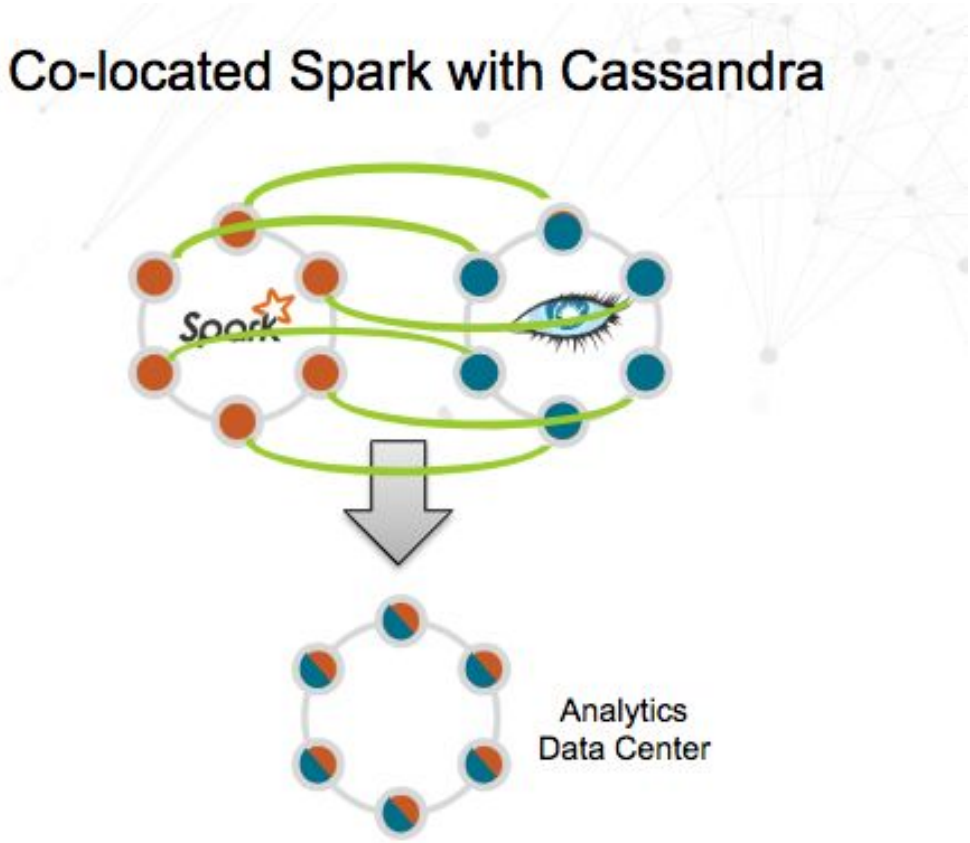
Confidential © DataStax, All Rights Reserved.

DATASTAX

DataStax Analytics and Machine Learning

- DataStax Analytics
  - Apache Cassandra
  - Apache Spark
  - 1 line of code

    MAGIC



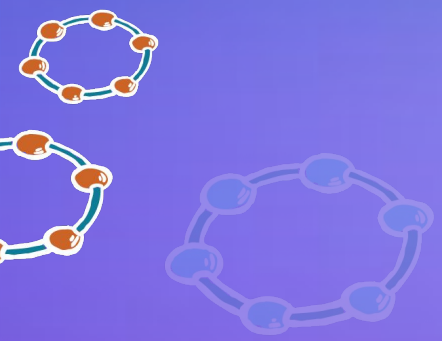
Confidential © DataStax, All Rights Reserved.
DATASTAX

It All Comes Together with the Beautiful Code

```
1 myDF = spark.read
2    .format('org.apache.spark.sql.cassandra')
3    .options((table = 'myTable'), (keyspace = 'mySpace'))
4    .load()
5
```

Confidential © DataStax, All Rights Reserved.

ABOUT ME

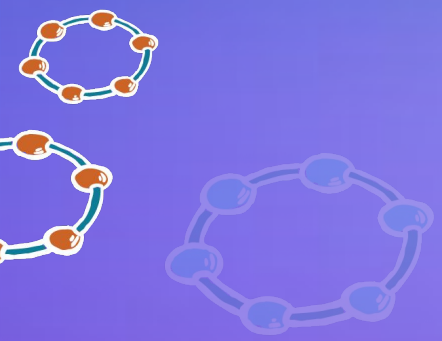


DATASTAX®

Who is Amanda?

Confidential © DataStax, All Rights Reserved.

LOGISTICS



DATASTAX®

5 Machine Learning Functions in 5 Easy Steps

- Explanation of the function
- Talk about use cases
- Review the Problem to Solve
- Review the Dataset
- DEMO

Confidential © DataStax, All Rights Reserved.
DATASTAX

Our Top 5 Functions

- K-Means
- Naive Bayes
- Random Forest
- FP-Growth
- Collaborative Filtering

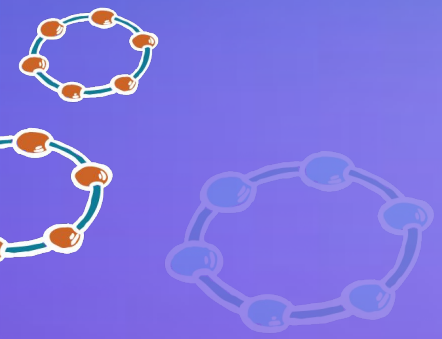
Confidential © DataStax, All Rights Reserved.

TIMERS READY

Confidential © DataStax, All Rights Reserved.

DATASTAX

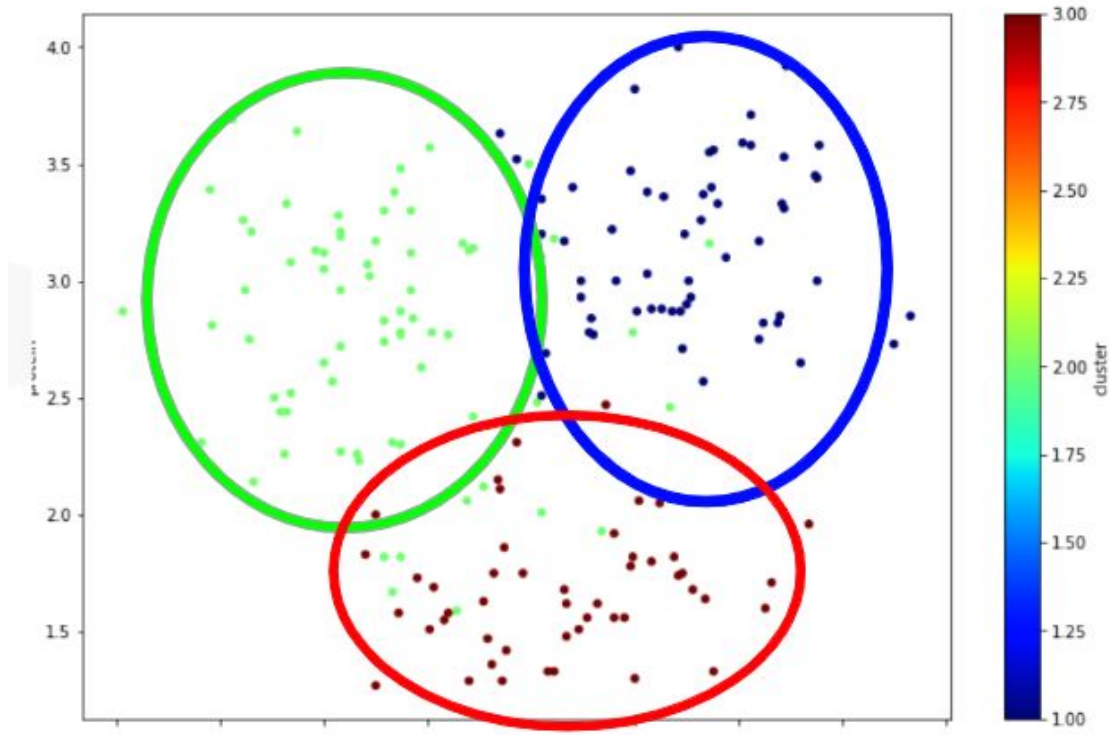
K-MEANS



DATASTAX®

What is K-Means?

- **Clustering** is the task of grouping a set of objects in such a way that objects in the same group are more similar
- **K-means** clustering is a simple unsupervised learning algorithm that is used for clustering
- It follows a simple procedure of classifying a number of clusters, defined by the letter "k"

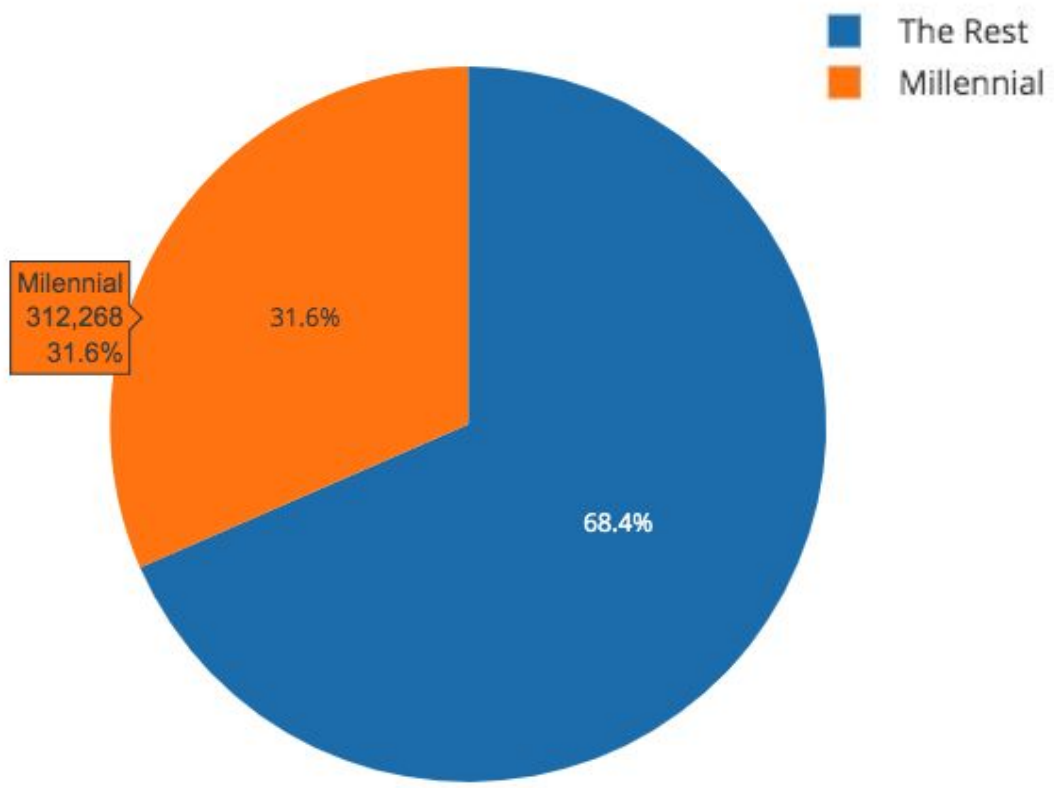


Confidential © DataStax, All Rights Reserved.
DATASTAX

K-Means Use Cases

- The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data.

- Behavior Segmentation of customers
- Buying Decisions
- Finding anomalies



Confidential © DataStax, All Rights Reserved.
DATASTAX

What Question Are We Asking?

Can K-Means be used to do social media analysis, can we group together different types of media by the reaction they received?

Status Id	Num Reaction	Num Comment	Shares	Likes	Loves	Wows	Hahas	Sad	Angry	Social Type

Social Media Dataset

Confidential © DataStax, All Rights Reserved.

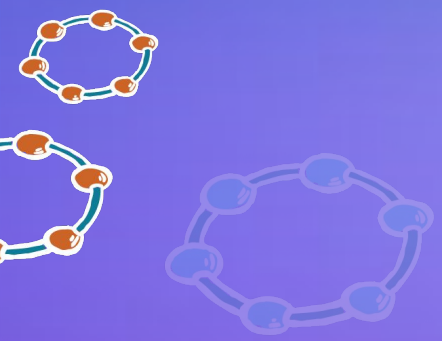
DATASTAX

TIMERS READY

Confidential © DataStax, All Rights Reserved.

DATASTAX

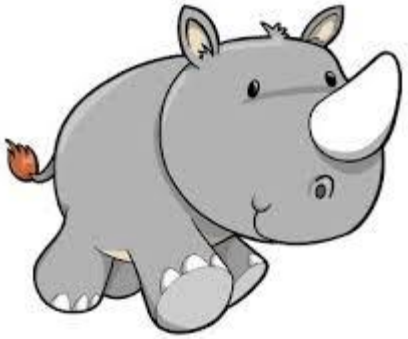
NAIVE BAYES



DATASTAX®

What is Naive Bayes?

- **Classification** identifies a category a new observation belongs on the basis of a training set of data containing data whose category membership is known
- **Naive Bayes** is a simple technique for constructing classifiers: assign class labels.



Label: FRUIT

Label: ANIMAL

Label: FRUIT

Confidential © DataStax, All Rights Reserved.

DATASTAX®

Naive Bayes Use Cases

- Good for real-time predictions
- Text Classification
- Spam Filtering
- Sentiment Analysis

SPAM Filters

Training Examples	Labels
Simply loved it	Positive
Most disgusting food I have ever had	Negative
Stay away, very disgusting food!	Negative
Menu is absolutely perfect, loved it!	Positive
A really good value for money	Positive
This is a very good restaurant	Positive
Terrible experience!	Negative
This place has best food	Positive
This place has most pathetic serving food!	Negative

Confidential © DataStax, All Rights Reserved.

What Question are we Asking?

Can Naive Bayes be used to classify a wine's rating score by its attributes?

Fixed Acidity	Volatile Acidity	Citric Acid	Residual sugar	Chloride	Free Sulfur	Total Sulfur	Density	pH	Sulphates	OH

Wine Quality Dataset

taStax, All Rights Reserved.

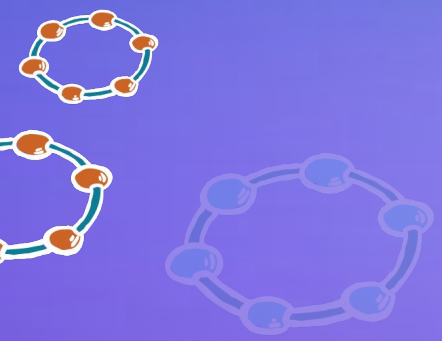
DATASTAX

TIMERS READY

Confidential © DataStax, All Rights Reserved.

DATASTAX

RANDOM FOREST



DATASTAX®

What is Random Forest?

- Random Forest models are built from Decision Trees and use a random sampling of data to build each tree and then merge them together.
- Decision Trees are built using intuitive modeling going through the data and asking yes and no questions until a classification can be made.

Confidential © DataStax, All Rights Reserved.

Random Forest Use Cases

- Classification
  - Discrete
- Regressions
  - Continuous

- Different than Naive Bayes:
  - Larger Model Size
  - Slower to Build
  - Can predict more advanced behavior
  - Better accuracy

Confidential © DataStax, All Rights Reserved.

What Question are we Asking?

Can Naive Bayes be used to classify a wine's rating score by its attributes?

Volatile Acidity	Fixed Acidity	Citric Acid	Residual sugar	Chloride	Free Sulfur	Total Sulfur	Density	pH	Sulphates	OH

Wine Quality Dataset

DataStax, All Rights Reserved.

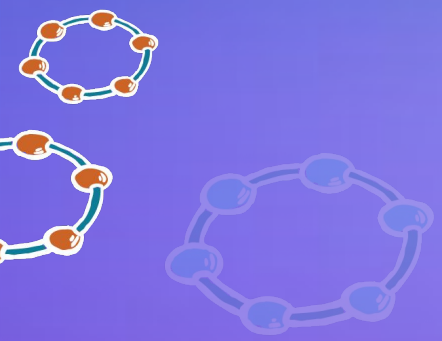
DATASTAX

TIMERS READY

Confidential © DataStax, All Rights Reserved.

DATASTAX

FP-GROWTH



DATASTAX®

What is FP-Growth?

- **Association rule learning** is a rule-based method for discovering interesting relations between variables in large databases
- **FP** stands for Frequent Pattern
- First, a set of attribute-value pairs in the dataset is found. Second, it builds the FP-tree structure for quick access.

Confidential © DataStax, All Rights Reserved.

FP-Growth Use Cases

- Shopping Cart Analysis
  - Promotions
  - Product Placement
- Web Traffic Usage

```
{bread, peanut butter} => {jelly}
```

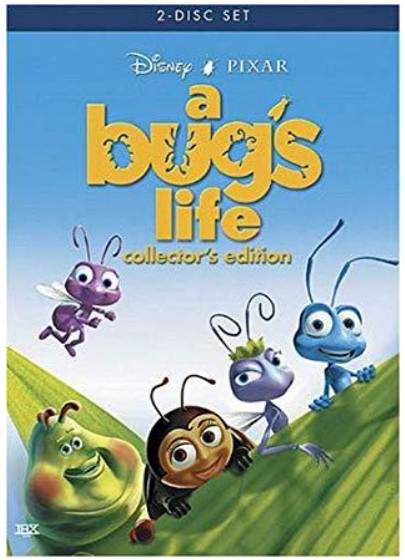
Confidential © DataStax, All Rights Reserved.

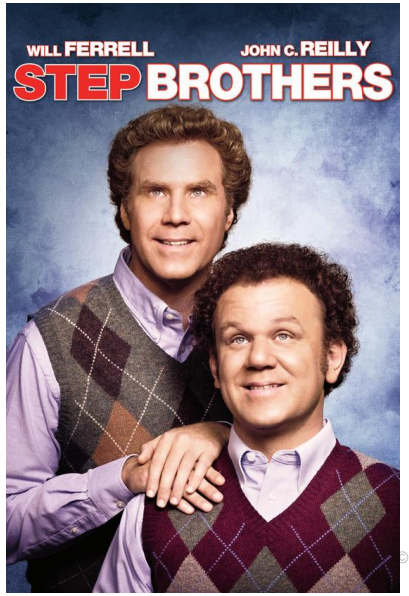
What Question are we Asking?

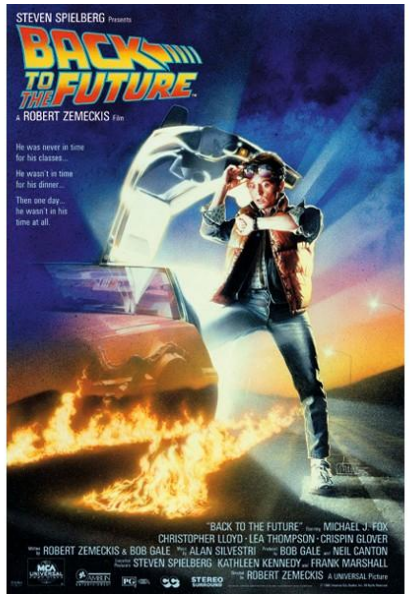
Can Fp-Growth be used to find which movies to recommend to our users?

User Id	Movie Id	Rating	TimeStamp

Movie Len Dataset







© DataStax, All Rights Reserved.

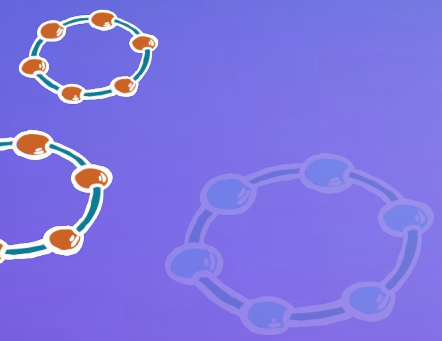
DATASTAX

TIMERS READY

Confidential © DataStax, All Rights Reserved.

DATASTAX

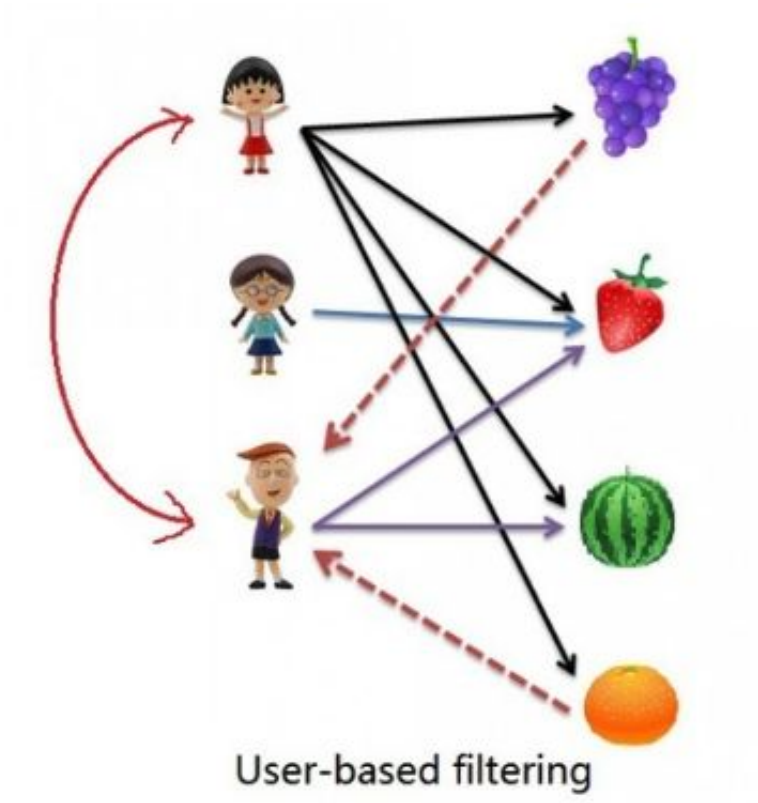
Collaborative Filtering



DATASTAX®

What is Collaborative Filtering?

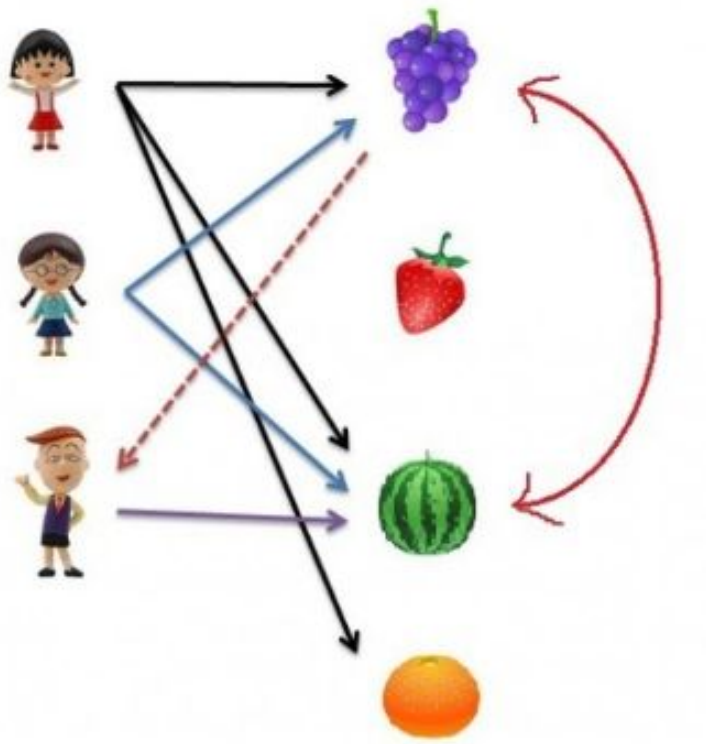
- **Collaborative filtering** is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Example:
  - *If A* is like *B*
  - A preference is more likely to equal B's for something we don't know about A.



Confidential © DataStax, All Rights Reserved.

Collaborative Filtering Use Cases

- User Based Recommendations
- Item Based Recommendations



Confidential © DataStax, All Rights Reserved.

DATASTAX

What Question are we Asking?

Can Collaborative Filtering be used to find which jokes to recommend to our users?

User Id	Joke Id	Rating

Jester Dataset

© DataStax, All Rights Reserved.

DATASTAX

WE DID IT!

Confidential © DataStax, All Rights Reserved.
DATASTAX

What's Next for You!

Install All The Things!

DataStax Academy

GitHub

Confidential © DataStax, All Rights Reserved.

DATASTAX

DATASTAX

ACCELERATE

THANK YOU

DATASTAX®