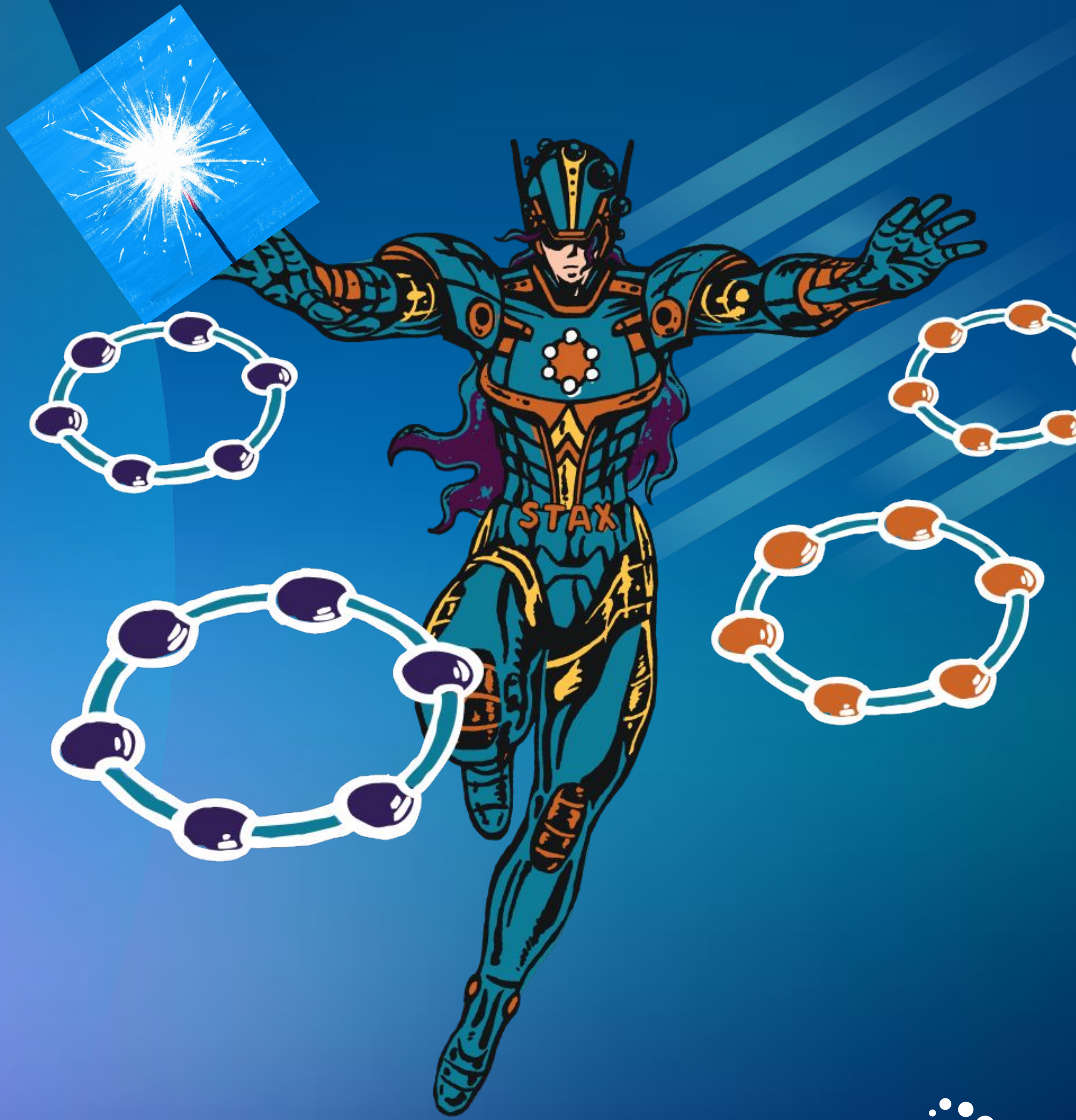
DATASTAX

ACCELERATE

Lighting a Spark with Machine Learning

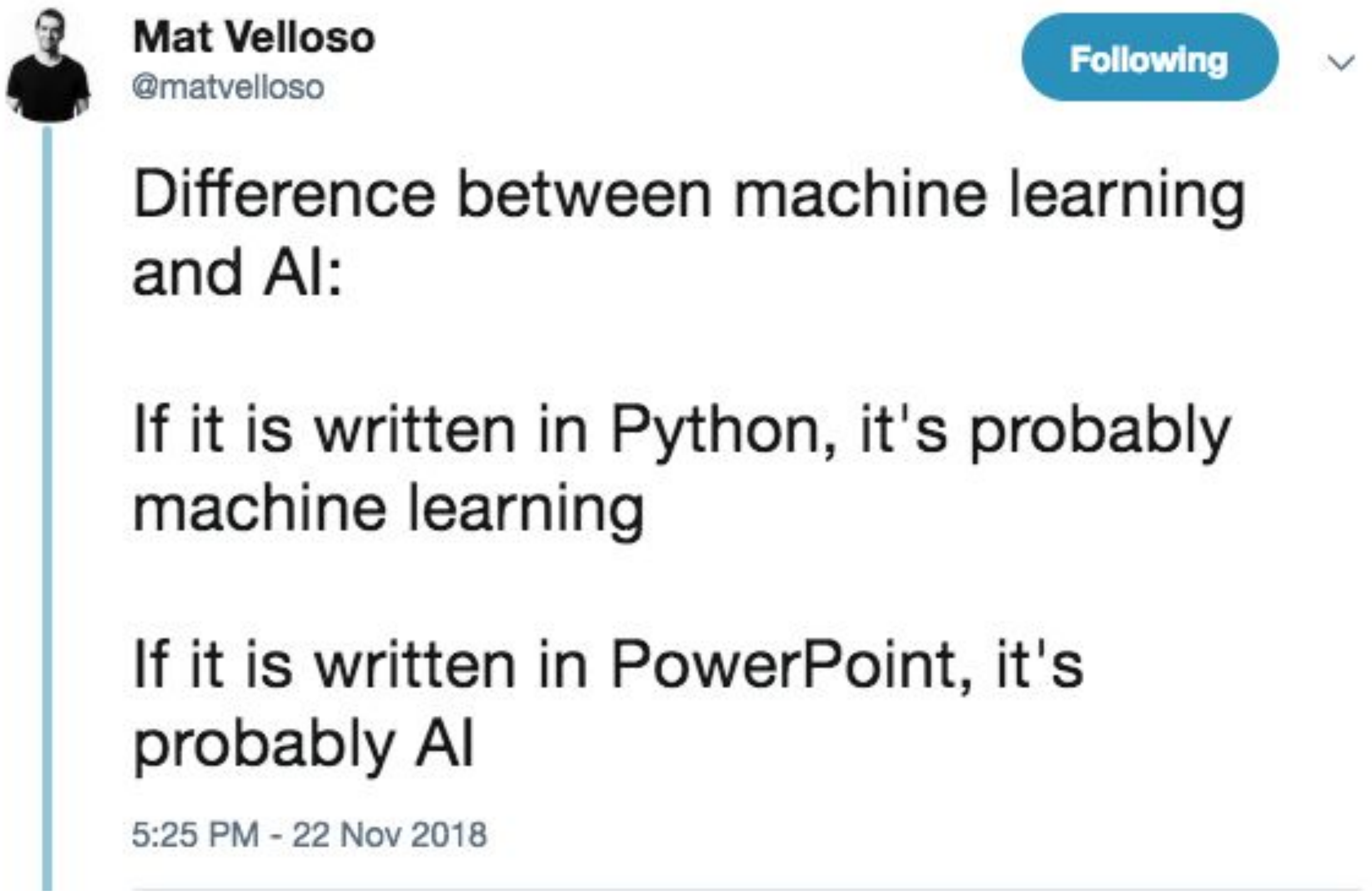
Amanda Moran
Developer Advocate, DataStax



DATASTAX®

WHY

The Importance of Machine Learning



Mat Velloso
@matvelloso

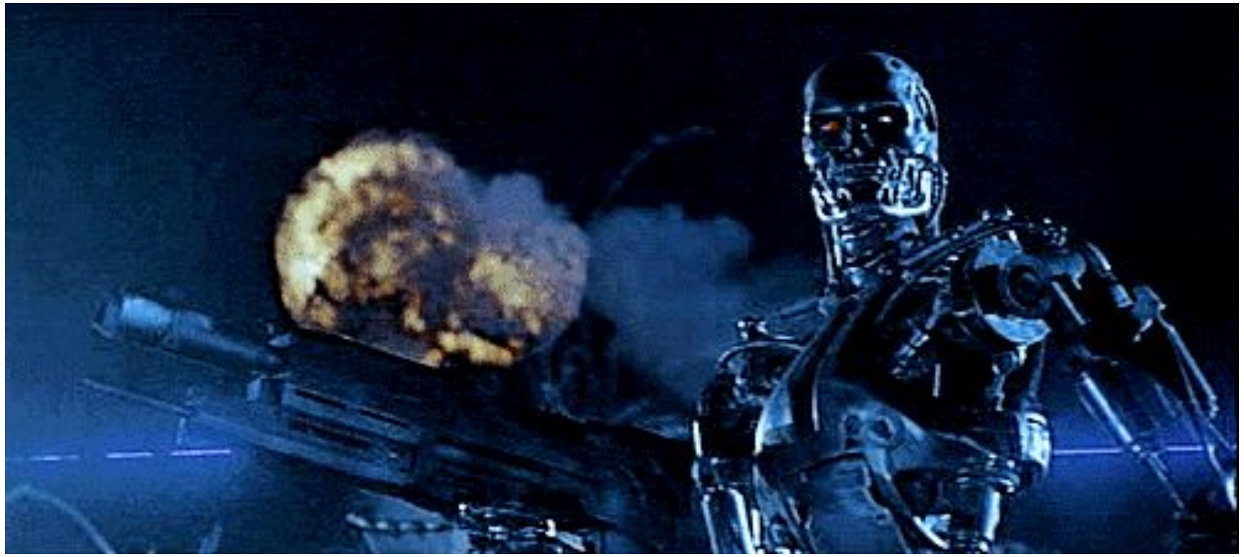
Difference between machine learning and AI:

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

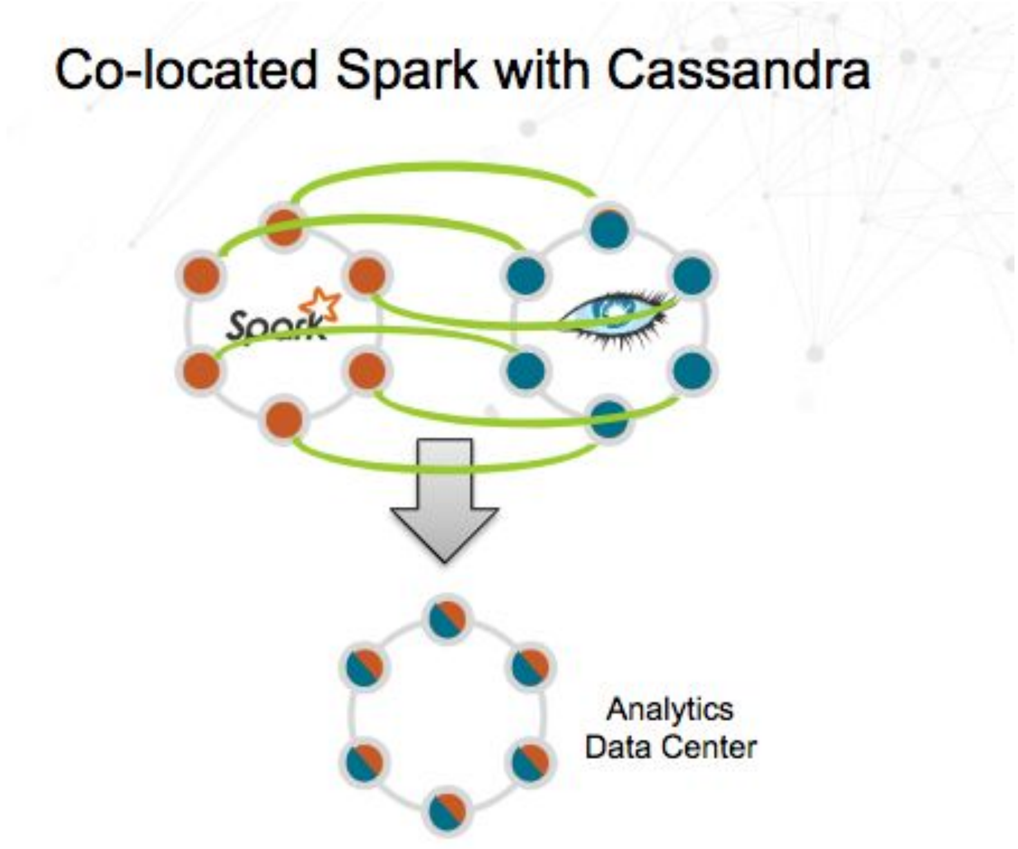
Focused on the Practical



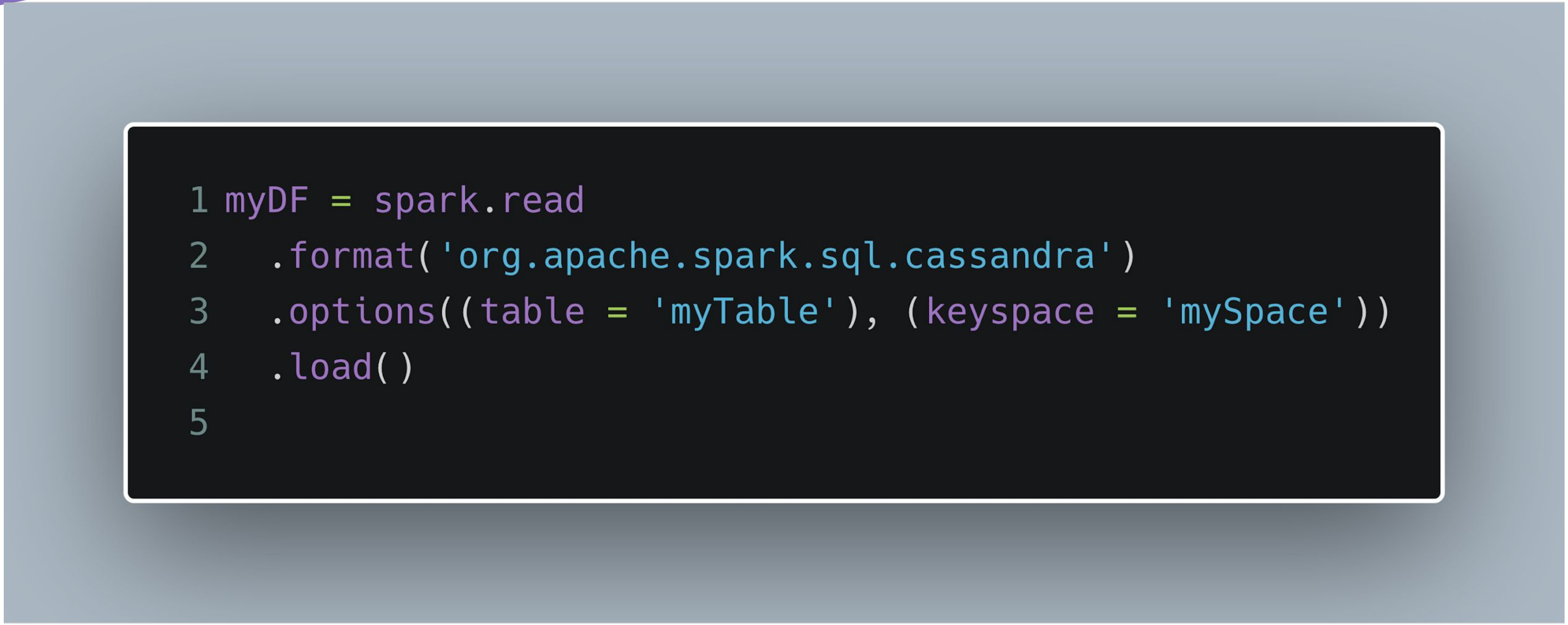
DataStax Analytics and Machine Learning

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

    MAGIC



It All Comes Together with the Beautiful Code



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

```

ABOUT ME

Who is Amanda?

WASHINGTON HUSKIES®

Santa Clara University

Apache Trafodion

APACHE Spark™

TERADATA. ASTER

TensorFlow

LOGISTICS

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

Our Top 5 Functions

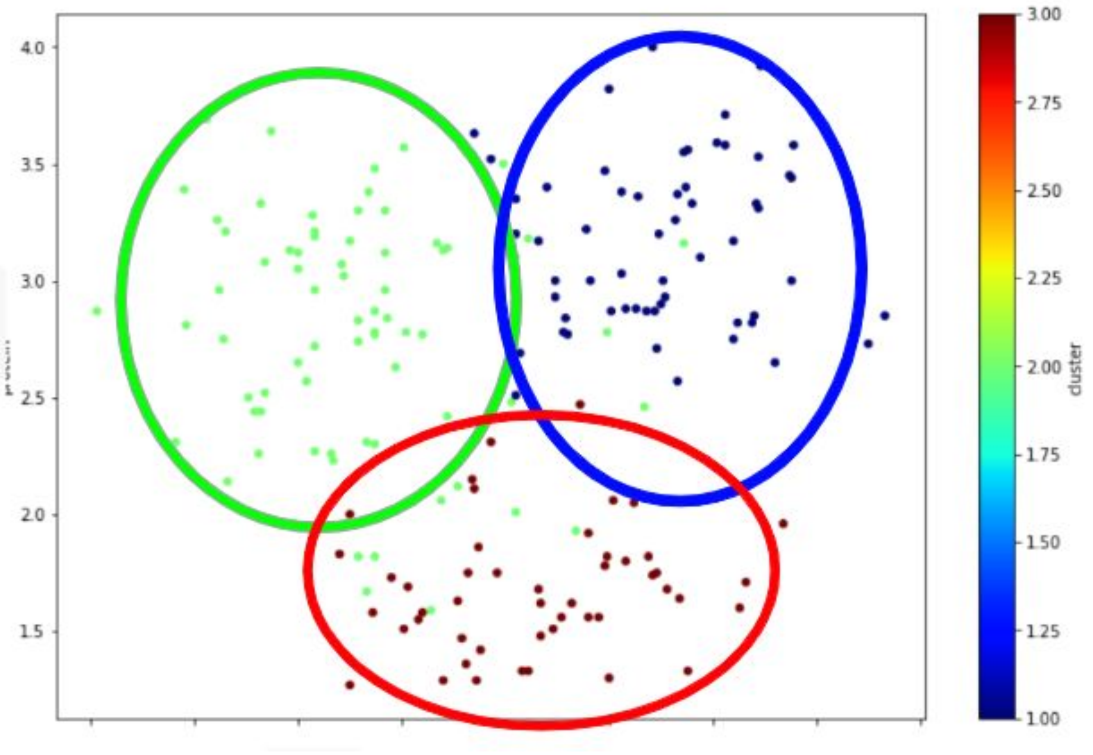
- K-Means
- Naive Bayes
- Random Forest
- FP-Growth
- Collaborative Filtering

TIMERS READY

K-MEANS

What is K-Means?

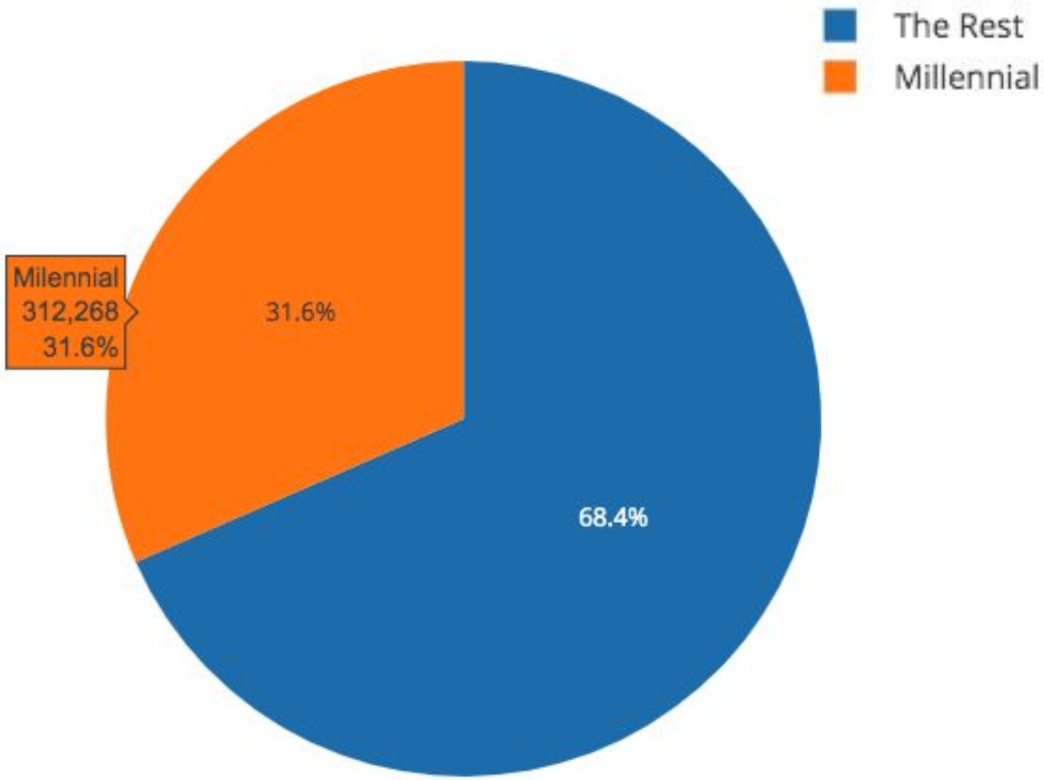
- **Clustering** is the task of grouping a set of objects in such a way that objects in the same group are more similar
- **K-means** clustering is a simple unsupervised learning algorithm that is used for clustering
- It follows a simple procedure of classifying a number of clusters, defined by the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



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



What Question Are We Asking?

Can K-Means be used to help decide what are the attributes of a car that will lead a customer to making a purchase?

Price of Car	Maintenance Cost	Doors	Capacity	Trunk Size	Safety

Car Evaluation Dataset

TIMERS READY

NAIVE BAYES

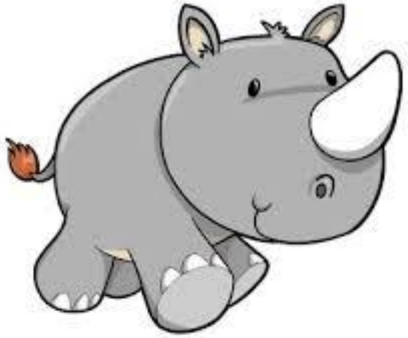
What is Naive Bayes?

- **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.

shutterstock.com • 225563614

Label: FRUIT



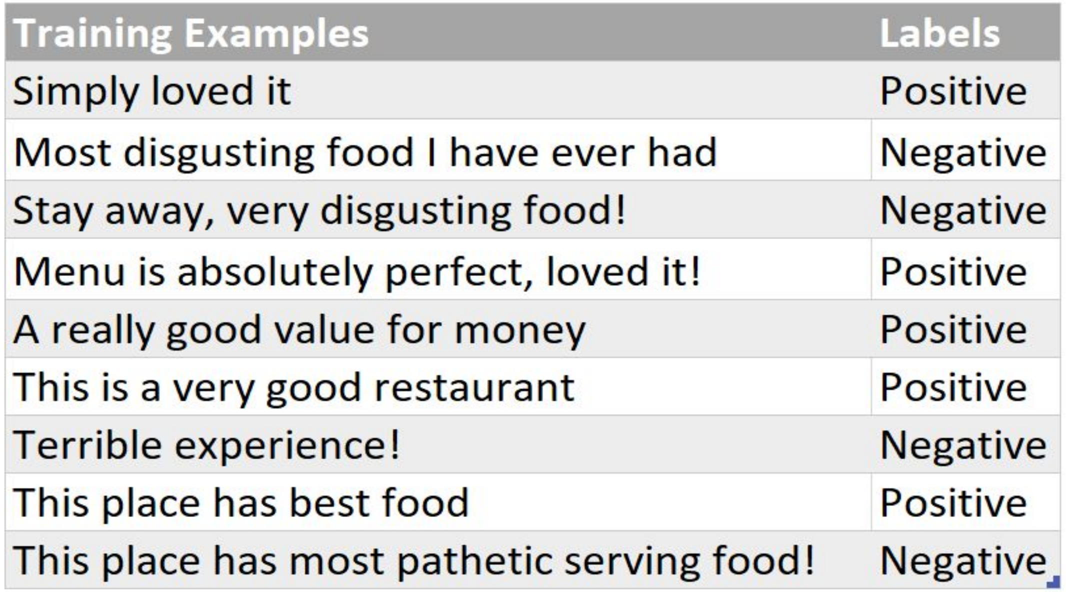
Label: ANIMAL

Label: FRUIT

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

What Question are we Asking?

Can Naive Bayes be used to classify if a wine is a good wine (score 9+) by its attributes?

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

Wine Quality Dataset

TIMERS READY

RANDOM FOREST

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.

Random Forest Use Cases

- Classification
- Regressions

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

What Question are we Asking?

Can Naive Bayes be used to classify if a wine is a good wine (score 9+) by its attributes?

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

Wine Quality Dataset

TIMERS READY

FP-GROWTH

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.

FP-Growth Use Cases

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

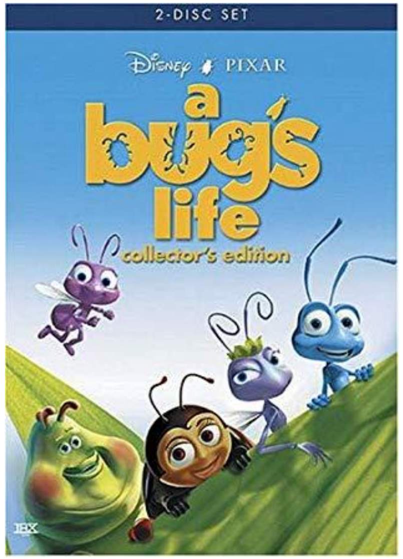
```
{bread, peanut butter} => {jelly}
```

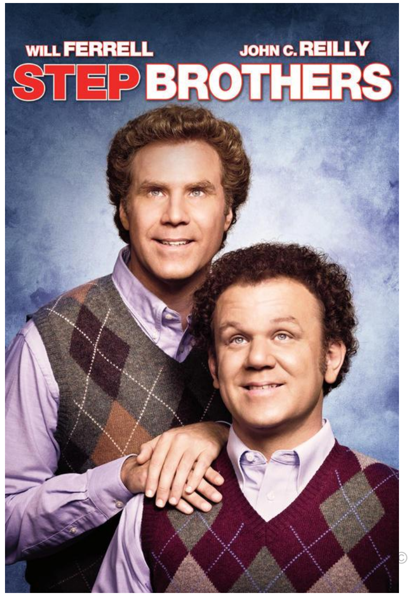
What Question are we Asking?

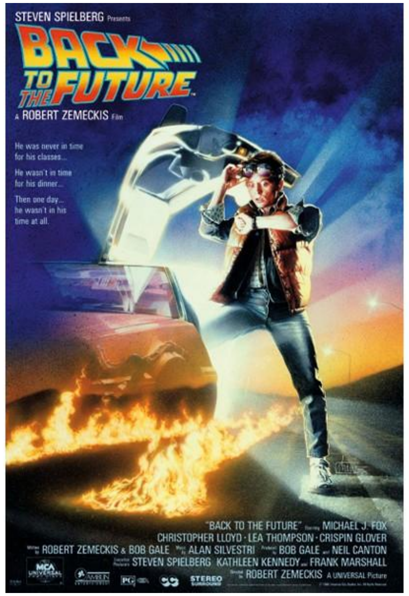
Can Fp-Growth be used to find which movies to recommend to our users?

User Id	Movie Id	Rating	TimeStamp

Movie Len Dataset





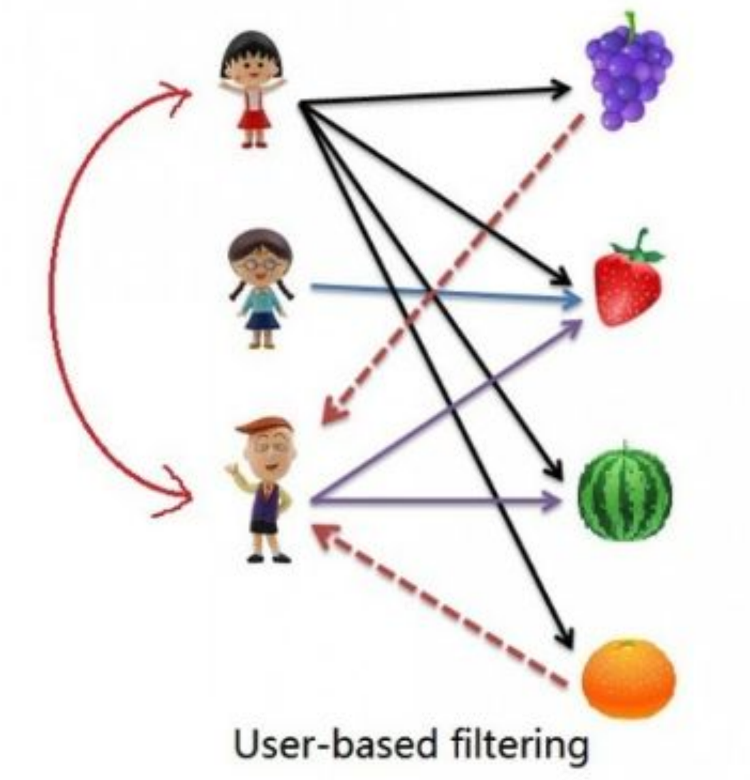


TIMERS READY

Collaborative Filtering

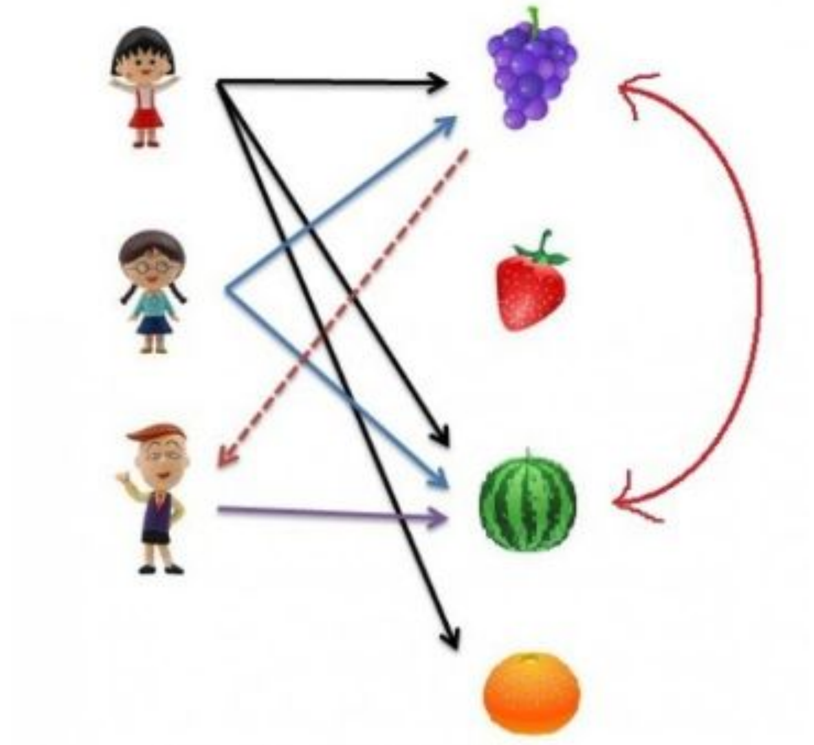
What is Collaborative Filtering?

- **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.



Collaborative Filtering Use Cases

- User Based Recommendations
- Item Based Recommendations



Item-based filtering

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

WE DID IT!

CLUSTERING

CLASSIFICATION

RECOMMENDATIONS

What’s Next for You!

- Set this up locally and try it out
  - https://academy.datastax.com/content/Apache-Cassandra-Apache-Spark-and-Jupyter
- All the code can be found
  - https://github.com/amandamoran/accelerate
- Learn more about Apache Cassandra, DSE, and Spark
  - https://academy.datastax.com/

DATASTAX

ACCELERATE

THANK YOU

DATASTAX®