Machine Learning 101

Outline

What is machine learning?

Importance of data

Learning types and algorithms

Examples and demo

About me

Mark Kalal

Software development / technology solutions

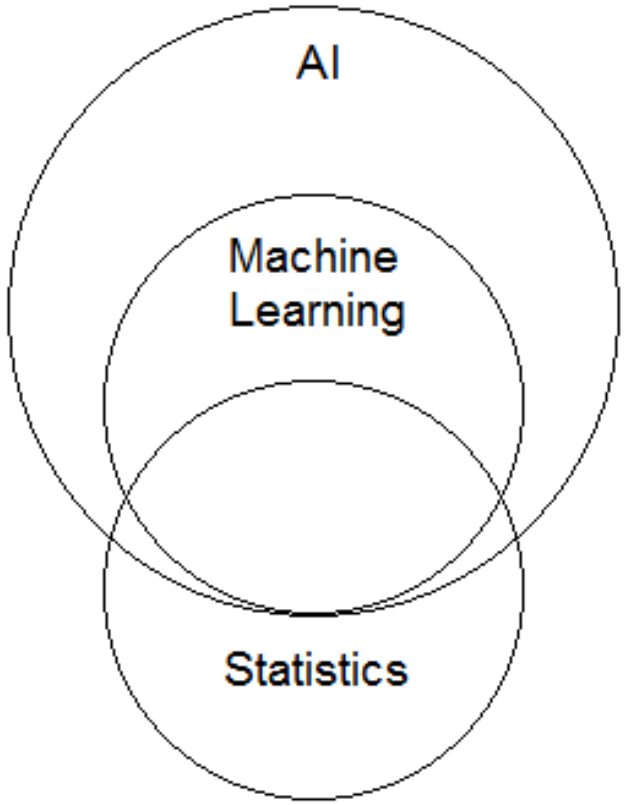
mdkalal@gmail.com

@MarkKalal

What is Machine Learning

Statistics?

Artificial Intelligence?

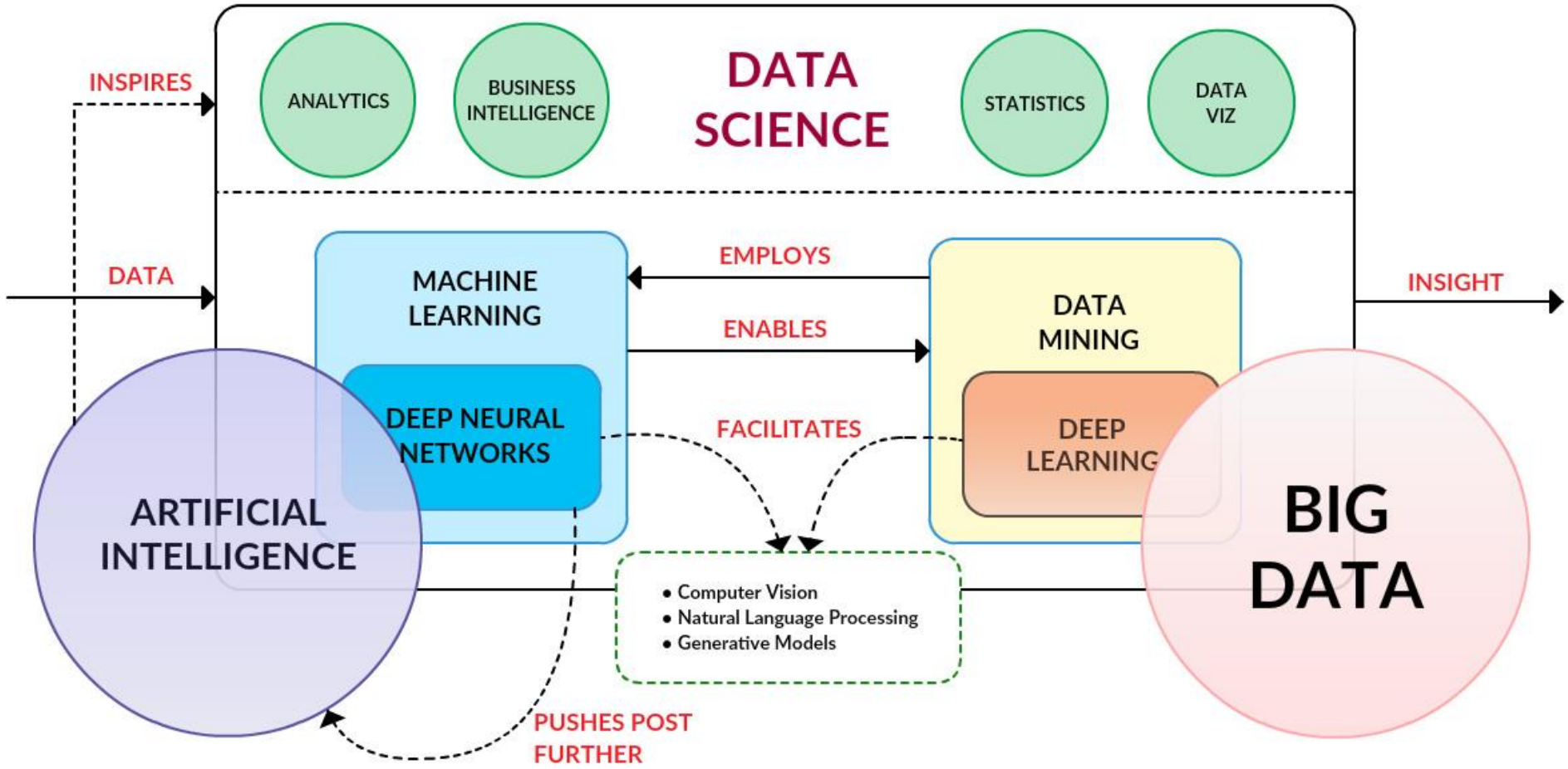


Learns by experience

Learns by ~~experience~~ data

Gets specific instructions

Application of artificial intelligence (AI) that provides systems the ability to learn and improve from “experience” (data) without being explicitly programmed


INSPIRES
ANALYTICS
BUSINESS INTELLIGENCE
DATA SCIENCE
STATISTICS
DATA VIZ
DATA
MACHINE LEARNING
EMPLOYS
ENABLES
DATA MINING
INSIGHT
DEEP NEURAL NETWORKS
FACILITATES
DEEP LEARNING
ARTIFICIAL INTELLIGENCE
BIG DATA
• Computer Vision
• Natural Language Processing
• Generative Models
PUSHES POST FURTHER

Nand Kishor

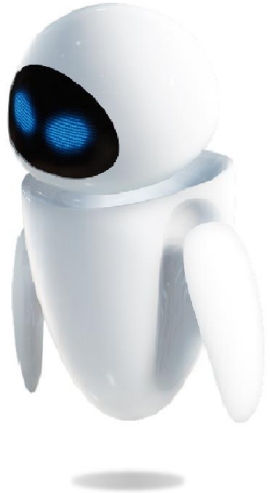
So what?

Many benefits

- Speed of analyzing complex data, revolutionizing business and data processing
- Greatly increased memory handing and computational powers (past barriers to implementation)

Some concerns

- Consequences, potential for misuse – intentional and non-intentional
- Opaque processing



My computer suddenly started singing “Hello from the other side”!

Of course it did, after all …

It’s *A Dell!*

It’s all about the data

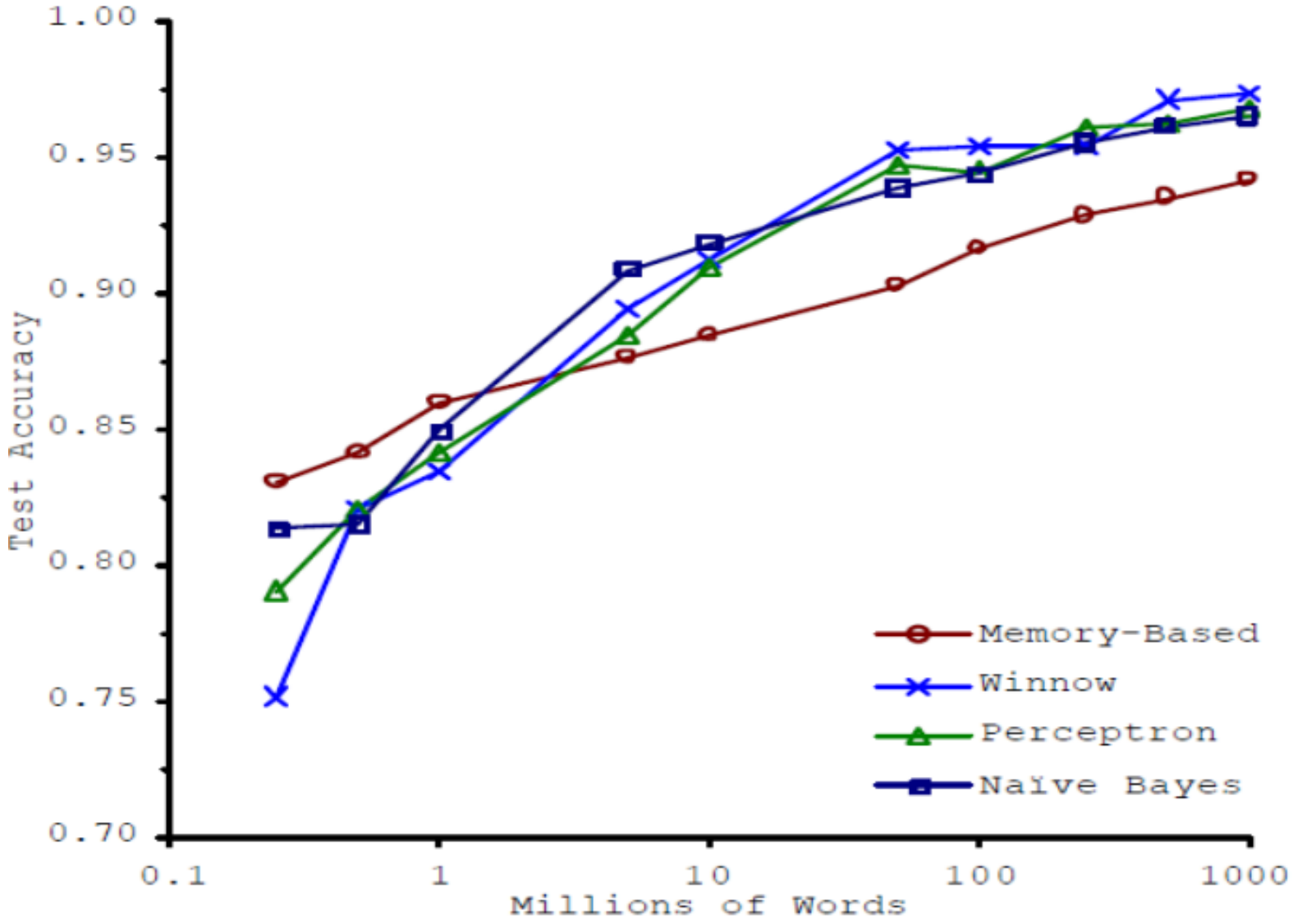
“Fuel” for machine learning

Data, data, and more data

Many useful things are being done


SPAM
Classic
Ingredients: Pork with Ham, Salt, Water, Modified Potato Starch, Sugar, Sodium Nitrite.
Serving Suggestion


Go to basket
Checkout



Banko and Brill, 2001

“Big Data”

Volume

Velocity

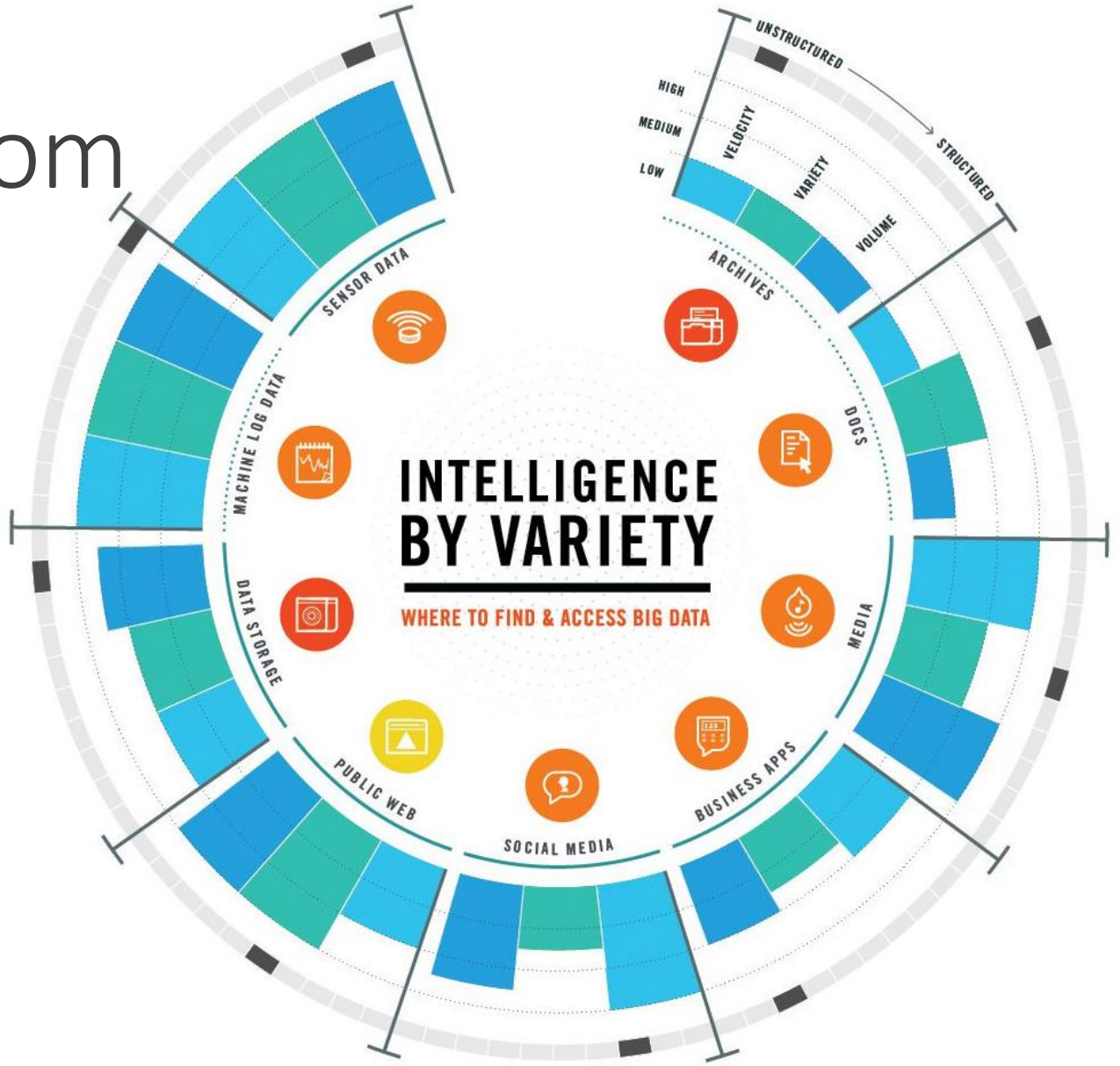
Variety

Where does it come from

Some you already have
- Database
- Logs

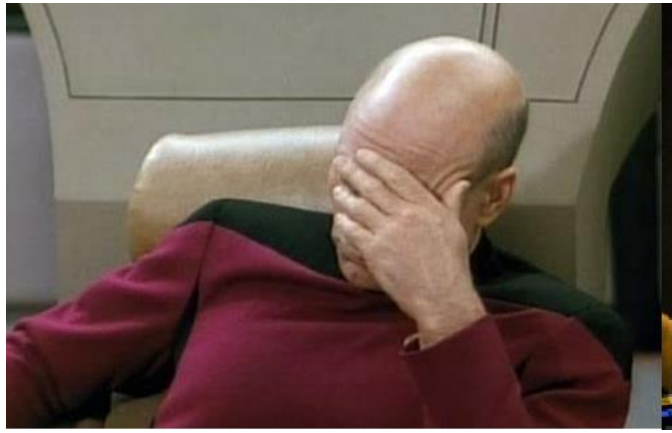
Some you can get
- Web, public sources
  (data.gov, data.worldbank.org)
- IOT sensors

Some you can ask for
- Social media
- Anything a user can provide

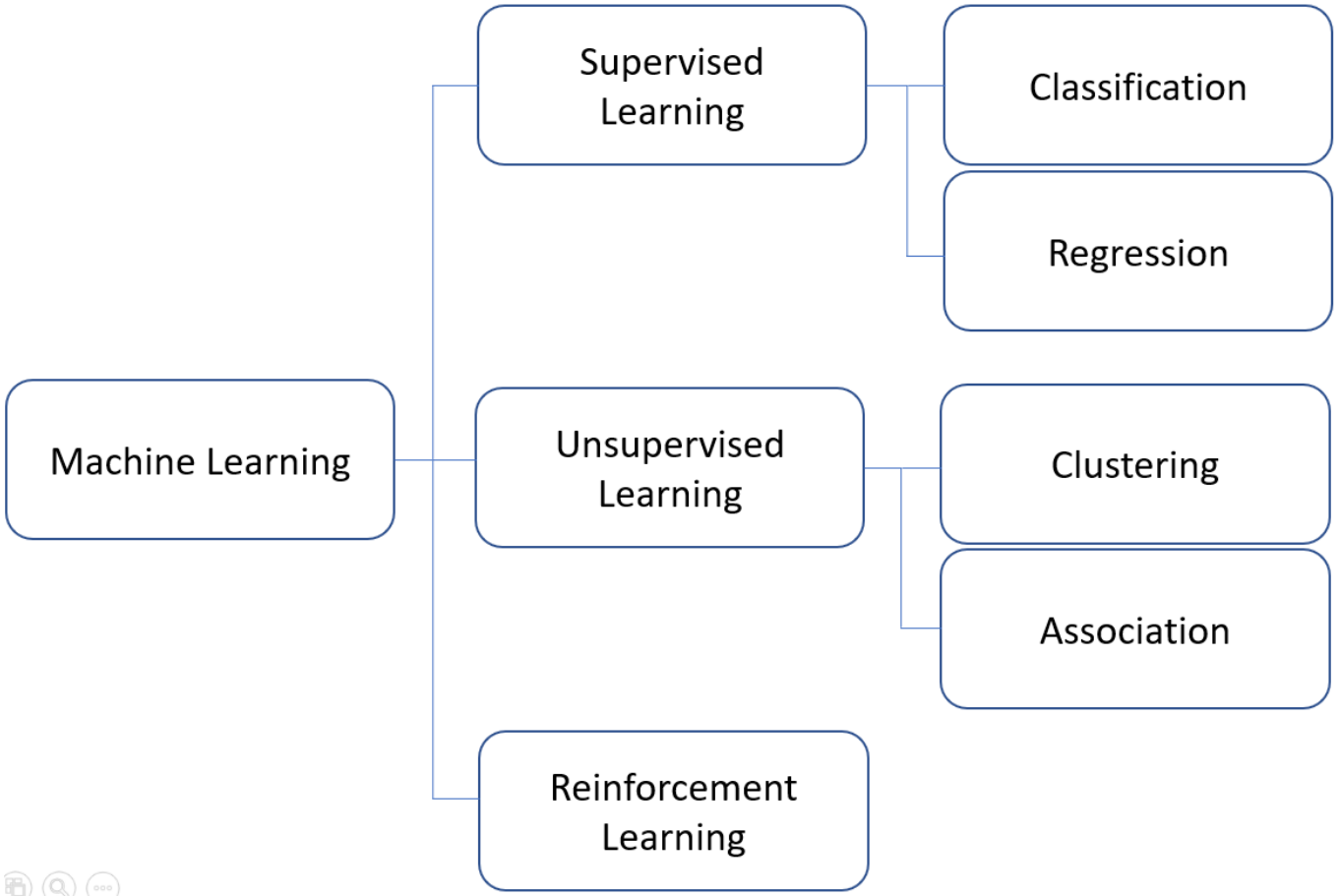


The journalist asked a programmer: “What makes bad code”? His reply?

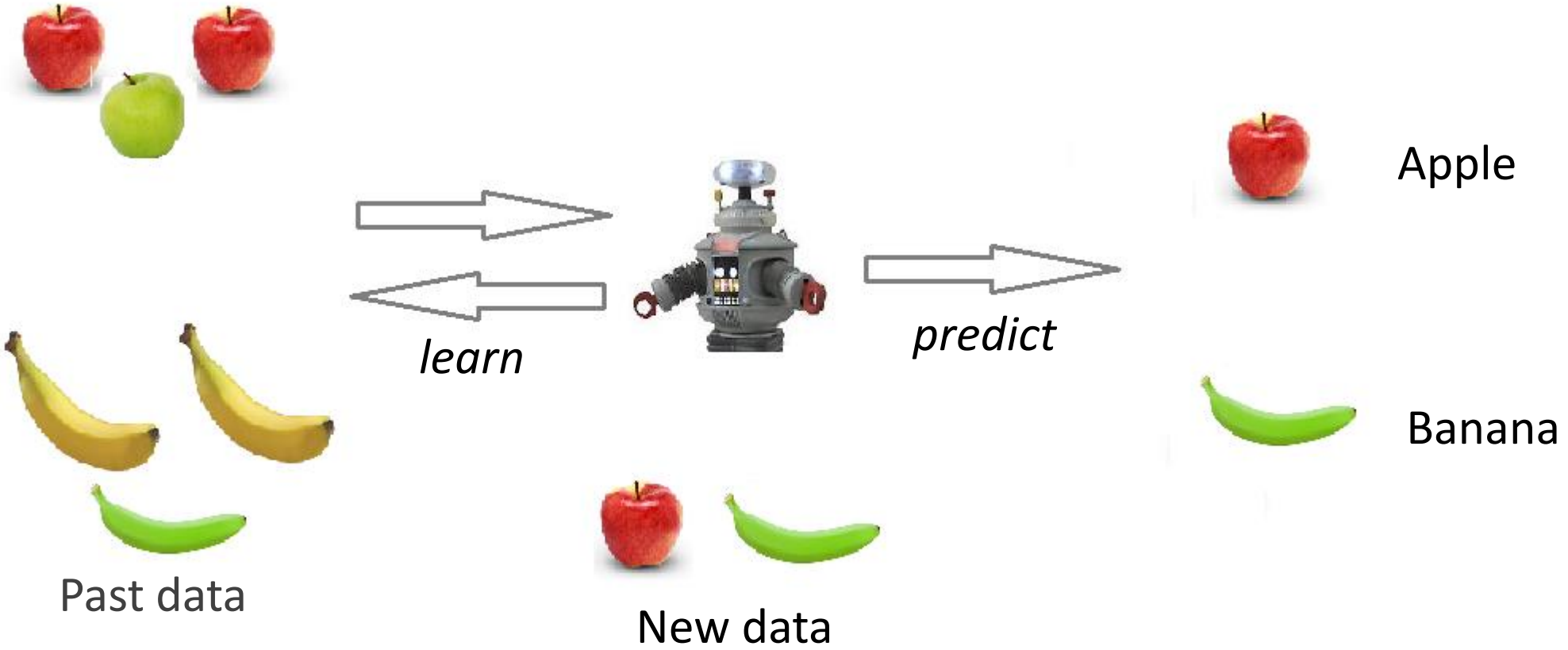
No Comment!



Learning types



Supervised Learning – make predictions



Supervised Learning

Classification

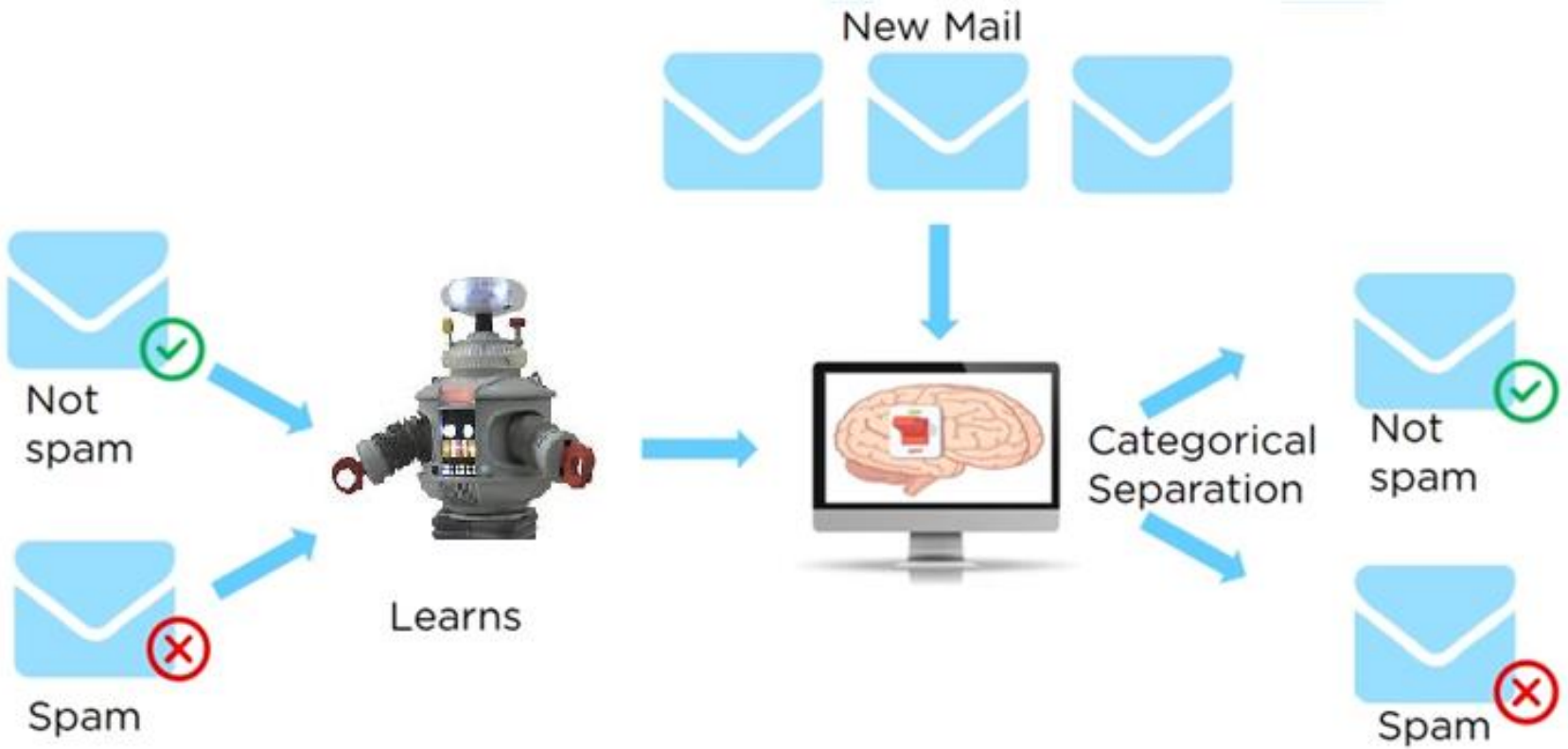
- Question or output is categorical, i.e. True/False

Regression

- Question or output is a real or continuous value

Supervised Learning

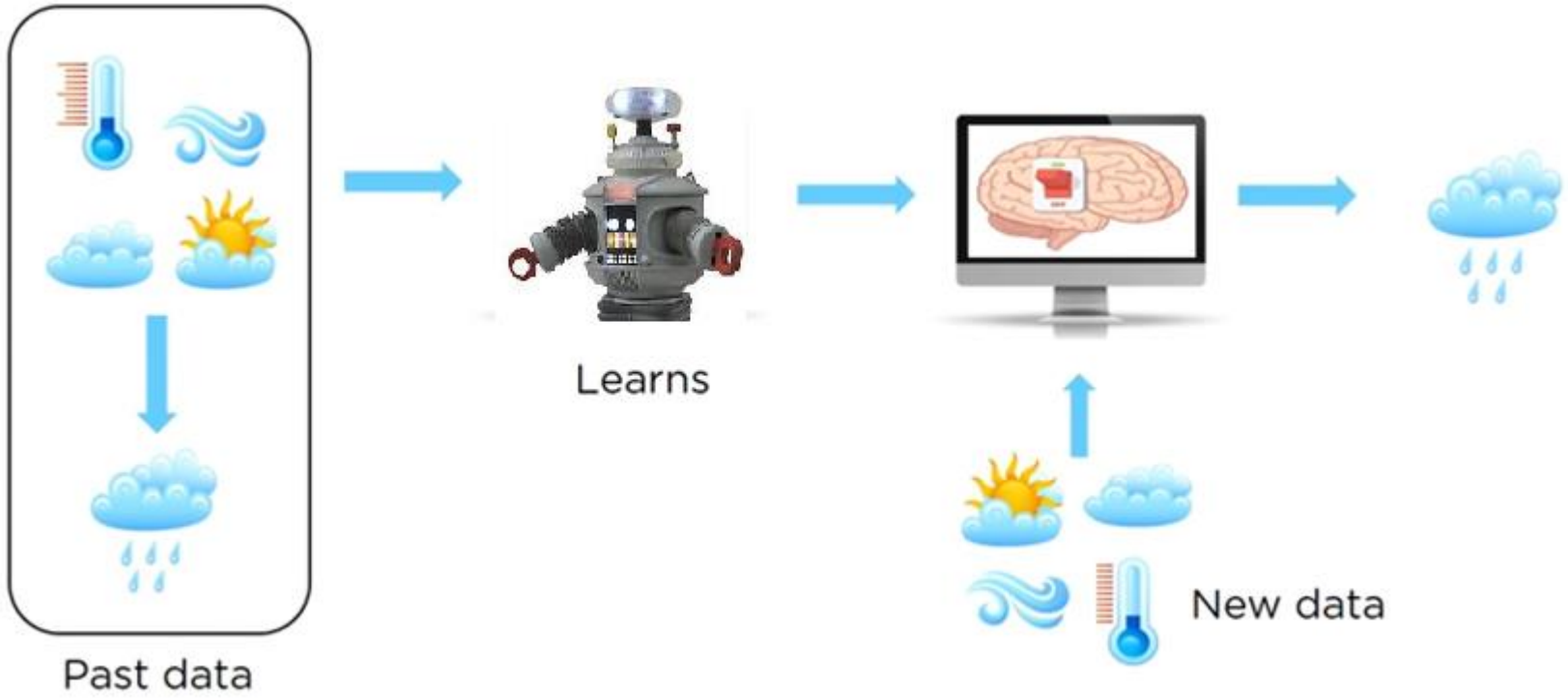
Classification



Simplilearn, 2018

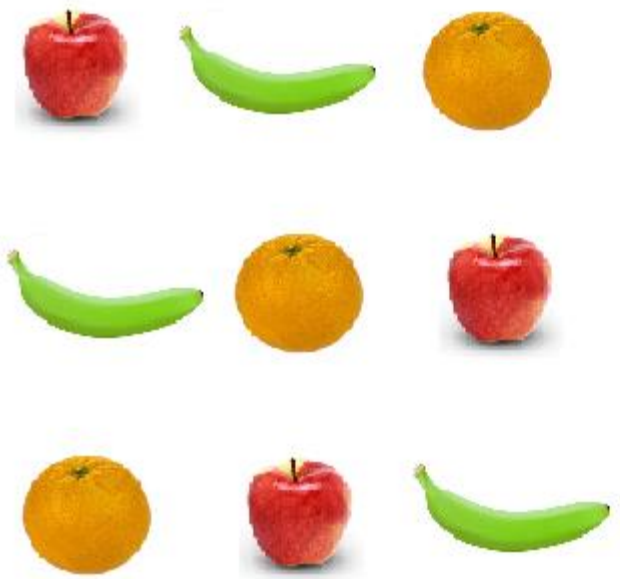
Supervised Learning

Regression



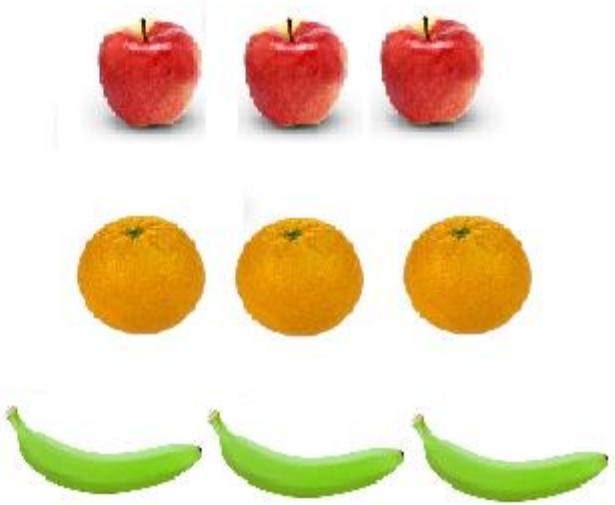
Simplilearn, 2018

Unsupervised Learning – finding hidden patterns



Input

segregate *organize*



Output

Unsupervised Learning

Clustering

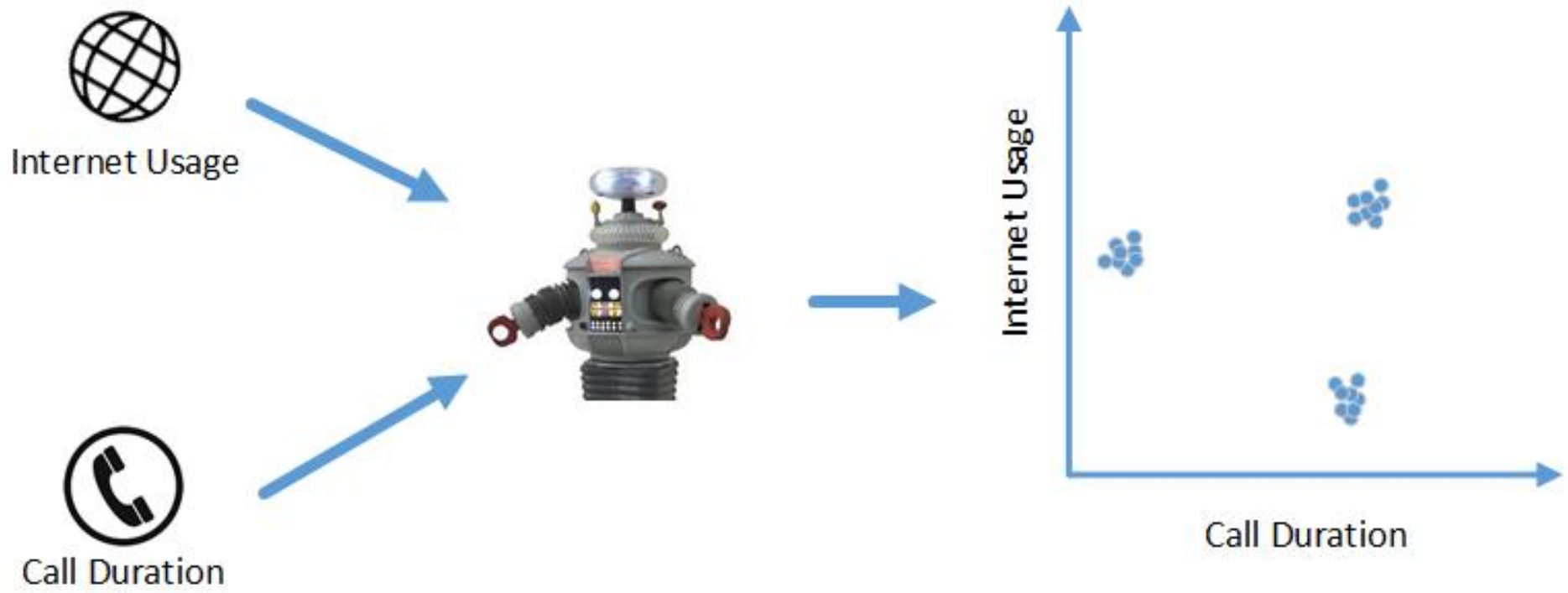
- Groups things based on similarities between them, and differences between others

Association

- Discovers relations or probability of occurrences within data

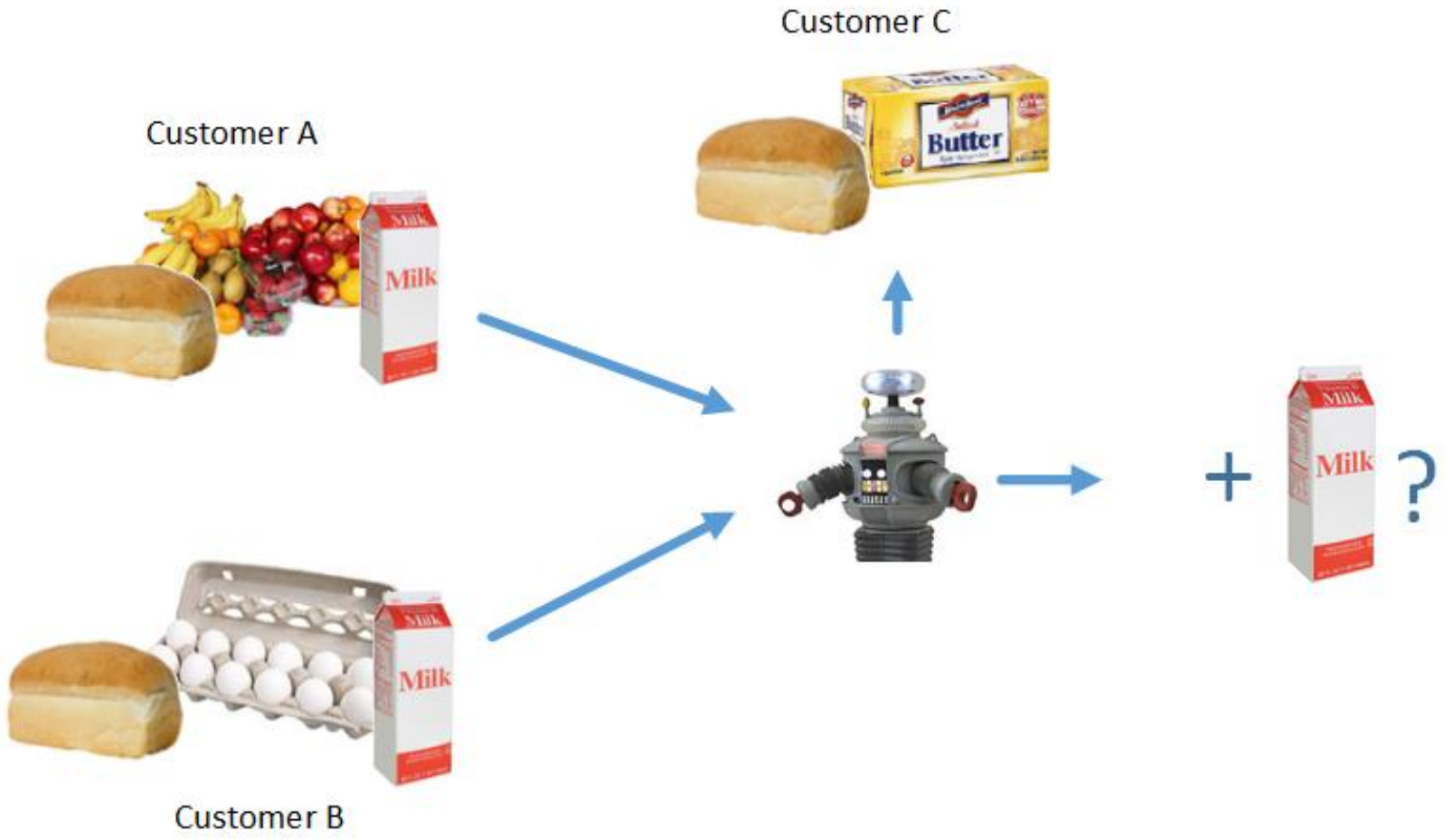
Unsupervised Learning

Clustering



Unsupervised Learning

Association



Reinforcement Learning - Decisions based on rewards for past actions

learns

good move

bad move

use feedback

Algorithms


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Algorithms

Decision Tree

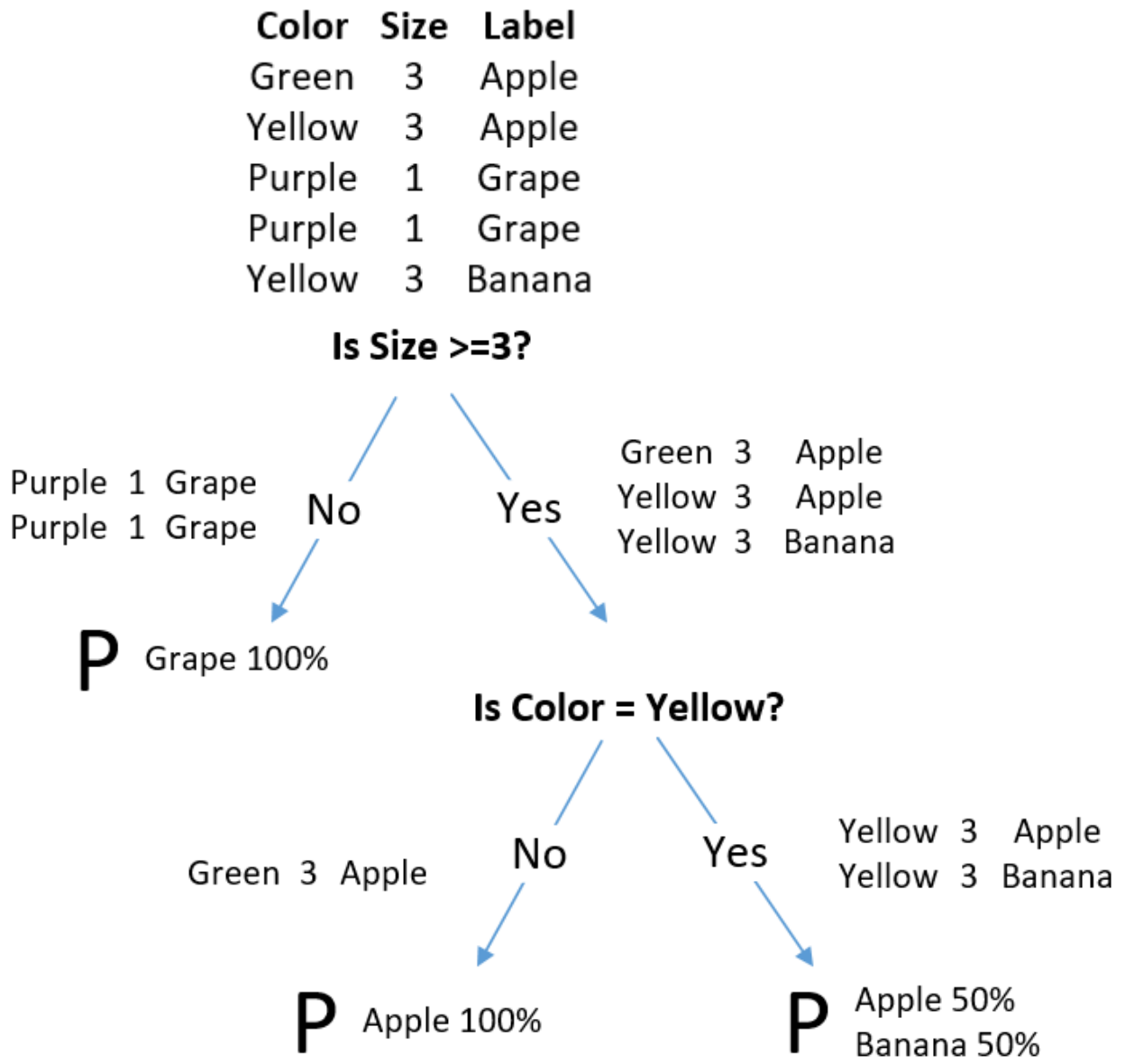
- Represents data that is divided/”branched” by conditions (questions and answers)

Linear Regression

- Represents and expresses the relationship between data with a line (X-Y grid)

Decision Tree

What fruit is this? Grape? Apple? Banana?

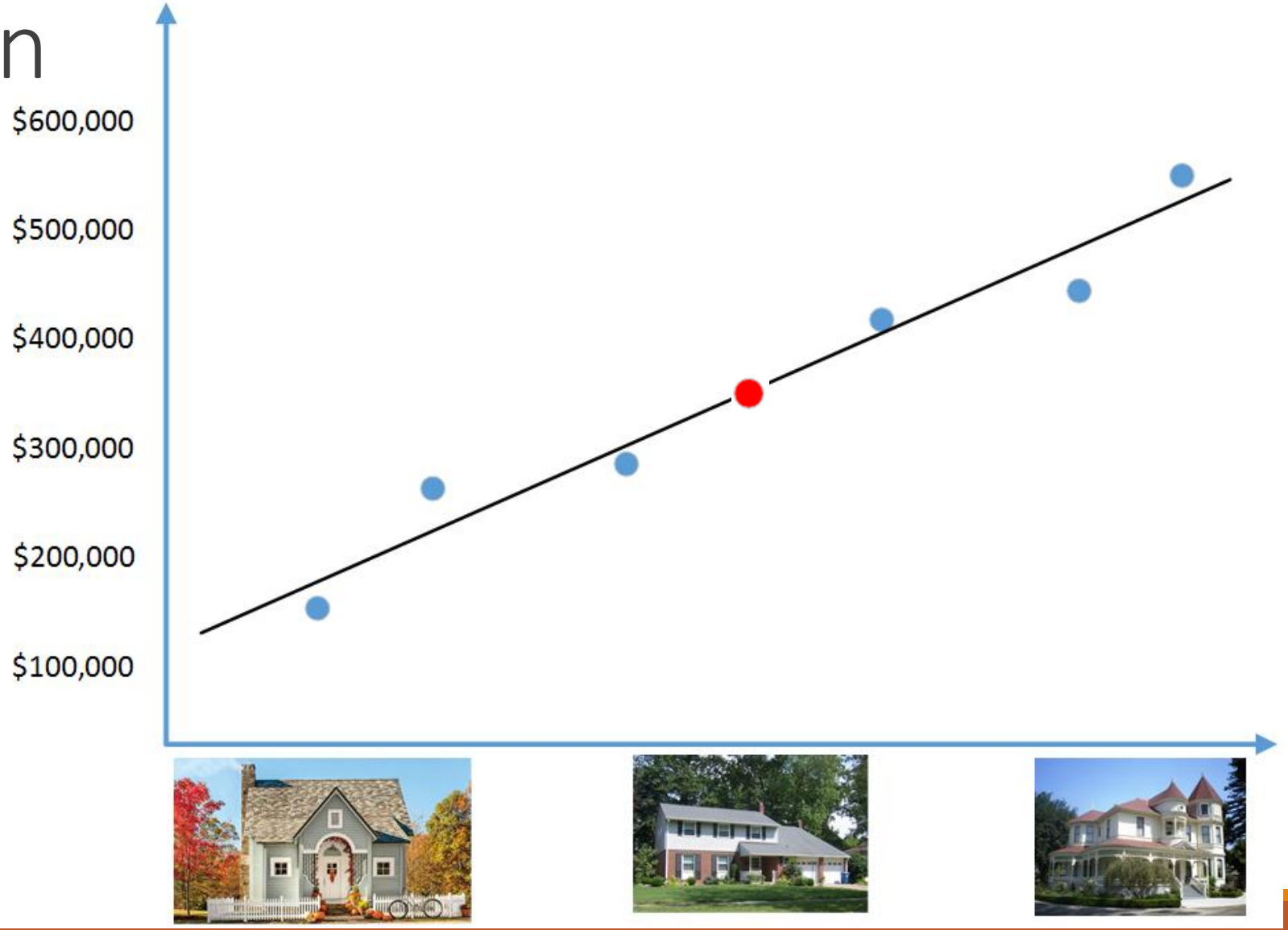

Color Size Label
Green 3 Apple
Yellow 3 Apple
Purple 1 Grape
Purple 1 Grape
Yellow 3 Banana
Is Size >=3?
No
Yes
Purple 1 Grape
Purple 1 Grape
Green 3 Apple
Yellow 3 Apple
Yellow 3 Banana
P
Grape 100%
Is Color = Yellow?
No
Yes
Green 3 Apple
Yellow 3 Apple
Yellow 3 Banana
P
Apple 100%
P
Apple 50%
Banana 50%

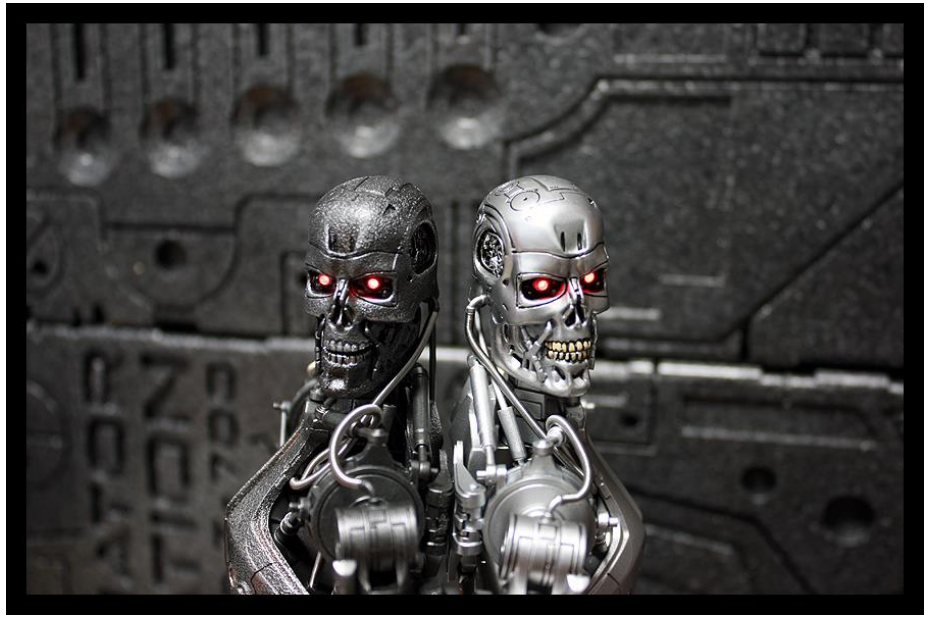
Linear Regression

What does this house cost?

Linear Regression

What does this house cost?





I walked down a street where the houses were numbered 64k, 128k, 256k, 512k, and 1mb

It was a trip down *memory lane*!

Machine Learning Lifecycle/Pipeline

Start with a question or problem

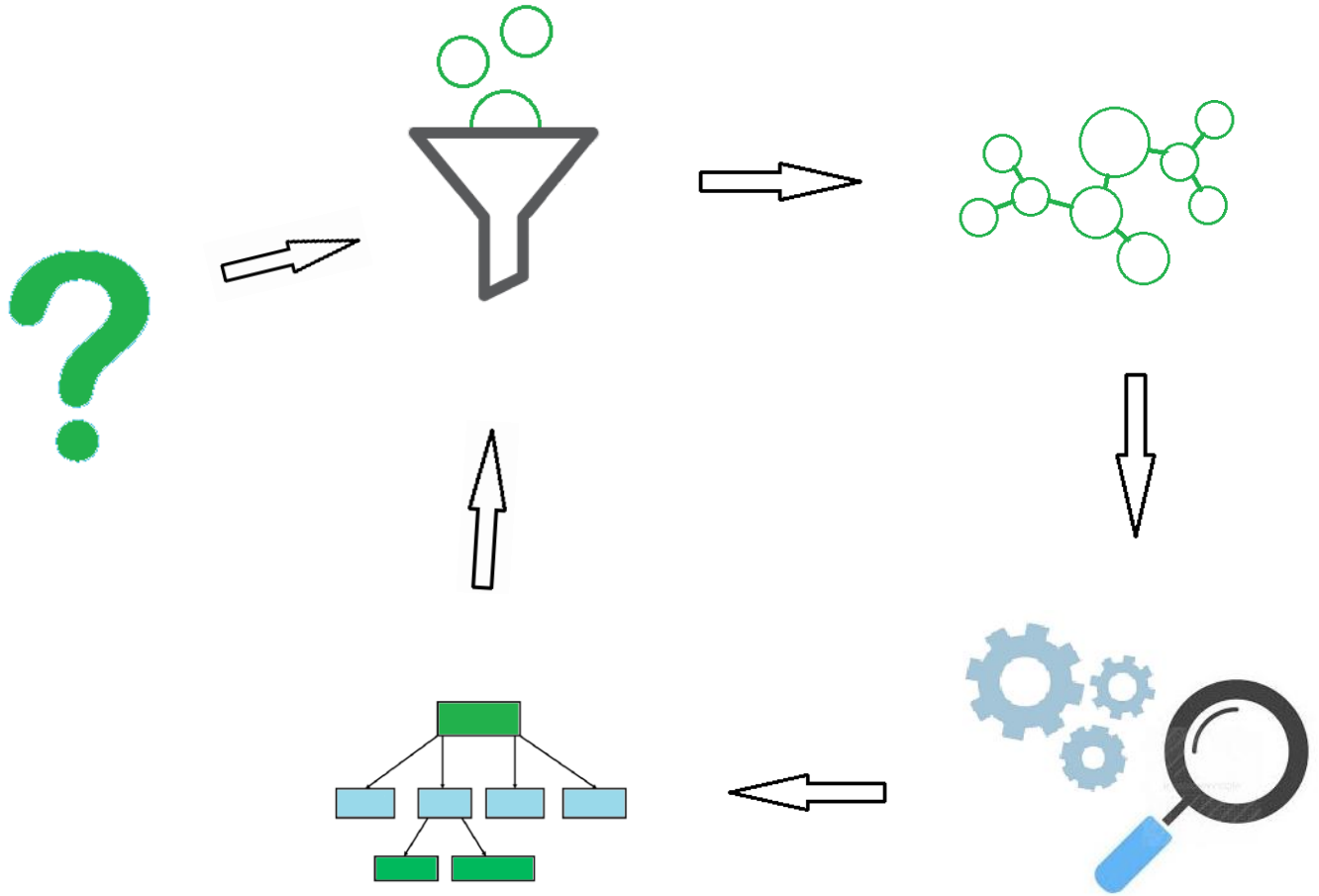
Collect data

Model

Train (and test)

Classify/Recommend/Predict

Re-train



Tools

Microsoft
Cognitive Services

Machine Learning

PROPHET

Weka Demo

Weka

- Installation
- Data files and formats
- Explorer
- Java API
- Sample apps and code

Summary

- ✓What is machine learning
- ✓Importance of data
- ✓Learning types and algorithms
- ✓Examples and demo

For more information

Weka - https://www.cs.waikato.ac.nz/~ml/weka

Test datasets - https://www.cs.waikato.ac.nz/ml/weka/datasets.html

Affective Tweets - https://affectivetweets.cms.waikato.ac.nz/

US open data - https://www.data.gov/

World Bank open data - https://data.worldbank.org/

Josh Gordon ML Recipes - https://www.youtube.com/watch?v=cKxRvEZd3Mw

Introduction to AI - https://www.coursera.org/learn/ai-for-everyone

Tensorflow - https://www.tensorflow.org/

This slide deck – https://github.com/mdkalal/ml101.git

Thank you!

Feedback welcome

Mark Kalal - mdkalal@gmail.com

@MarkKalal