GEORGE MCINTIRE

INTRODUCTION TO MACHINE LEARNING

TABLE OF CONTENTS

WHAT IS MACHINE LEARNING

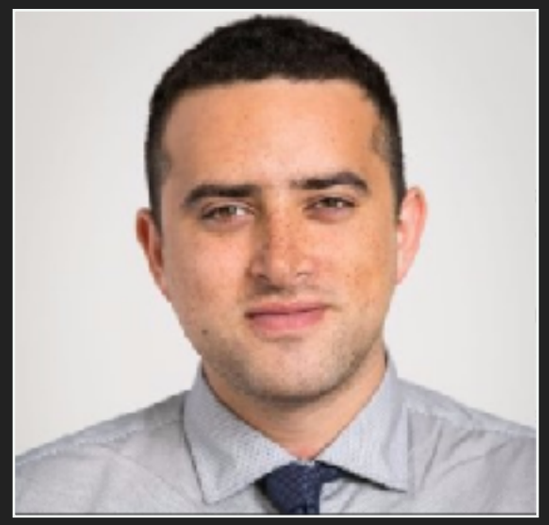
- “A field of study that gives computers the ability to learn without being explicitly programmed” (1959)
  - Arthur Samuel, AI pioneer, coined the term “Machine Learning“

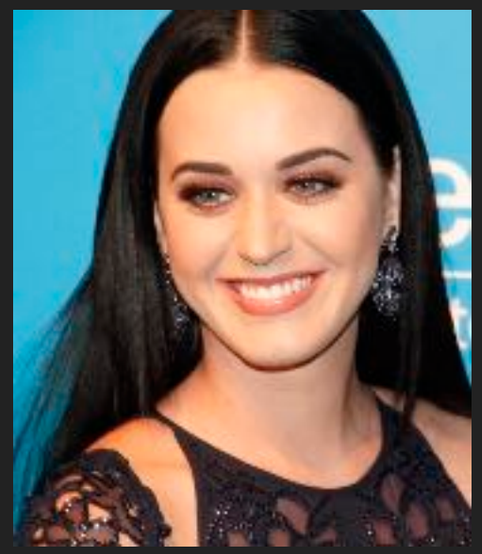
- “The automation of activities that we associate with human thinking, activities such as decision-making, problem solving, learning…” (1978)
  - Richard Bellman, applied mathematician

WHAT IS MACHINE LEARNING

- Examples in the form of data are passed through algorithms that look for patterns in that data in order to make predictions and decisions on future data.
- The computer observes that data of a certain category exhibits certain characteristics and data of another category exhibits a whole set of different characteristics. Allows the computer to properly classify data without that labeling.

FACE





FACE

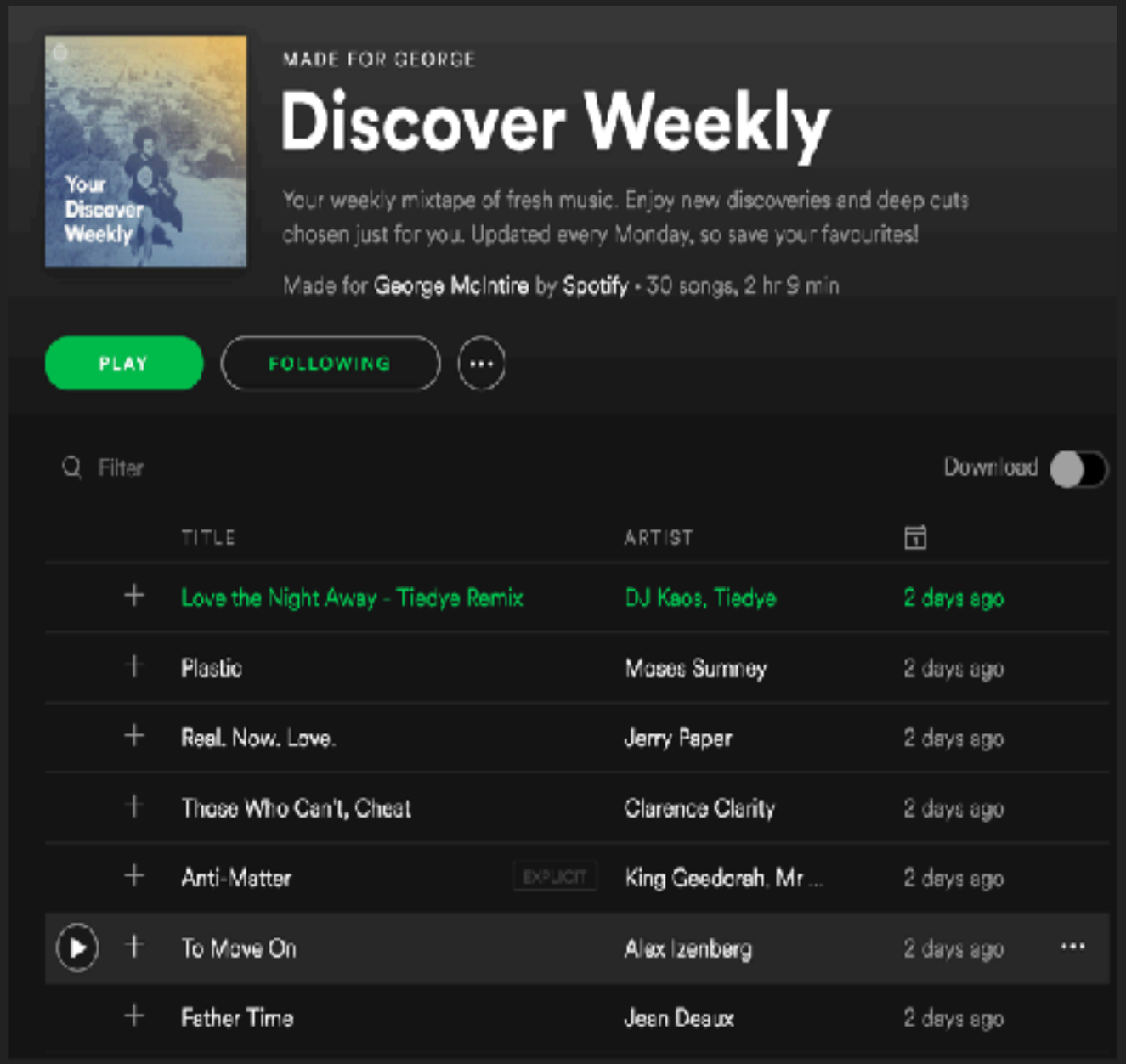
NOT A FACE

?

?

EXAMPLES

- Netflix and Spotify recommendations
- Credit card fraud detections
- Loan approvals
- Zillow’s Zestimate
- Email spam prevention
- Personal assistant machines: Siri, Alexa, etc…
- Crime pattern detections



TYPES OF MACHINE LEARNING

SUPERVISED

UNSUPERVISED

MAKING PREDICTIONS

FINDING STRUCTURES

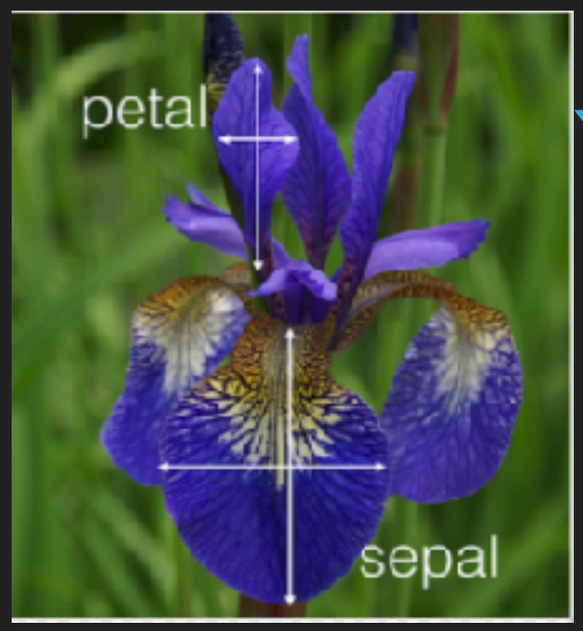
SUPERVISED LEARNING

SUPERVISED LEARNING

- Main goal is making predictions/classifications. Uses the past to predict the future.
- Data is composed of observations/events/instances.
- Predictors aka “X” aka the independent variables aka the features aka the input aka the attributes.
- Response variable aka “Y” aka the outcome aka the label aka the target aka the dependent variable.

SUPERVISED LEARNING DATASET

OBSERVATIONS



Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

PREDICTORS

RESPONSE

TYPES OF SUPERVISED LEARNING

CLASSIFICATION

- Outcome variable is a category:
  - good/bad
  - 1/0
  - sports/tech/politics/style
- Types of algorithms:
  - Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tress

REGRESSION

- Outcome variable is a continuous:
  - 3.6, 9.7, 2.3, 8.9,11.1,18.3, 23.6,4.2, 6.9
- Types of algorithms:
  - Linerar regression, Ridge regression, Lasso Regression

CLASSIFICATION EXAMPLE: LOAN DEFAULTS

- **Problem:** Lenders lose money when loanees fail to pay back loans.
- **Goal:** Develop a system that can efficiently identify high risk loans so lenders know which applications to reject
- **Data:** Records of previous loans marked as successful or failure that includes relevant information such as income, credit score, loan term, loan amount, etc…

REGRESSION EXAMPLE: HOUSING PRICES

- **Problem:** A home owner wants to sell her home but can't decide on an asking price.
- **Goal:** Accurately appraise the true value of the property
- **Data:** Home sale records labelled with their sale prices. Dataset features # bedrooms/bathrooms, sq ft, location, year sold, etc…

UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

- Absence of outcome variable/labels. Only features.
- Objective is looser and more exploratory.
  - Find groups of observations that exhibit similar characteristics.
  - Find combinations of features that explain the variation in the data.
- Useful as a preprocessing/exploratory data analysis step but to difficult to evaluate how well you're doing.

TYPES OF UNSUPERVISED LEARNING

CLUSTERING

- Make labels from unlabelled data.
- Examples: detect segment of users, derive micro-positions in sports
- Algorithms: KMeans, Hierarchal, DBScan

DIMENSIONALITY REDUCTION

- Deals with two many variables. Compresses data
- Great visualizing data with >=4 dimensions
- Algorithms: PCA, Truncated SVD, NMF

MACHINE LEARNING IN SCIKIT-LEARN