Google Cloud

Launching into ML

Agenda

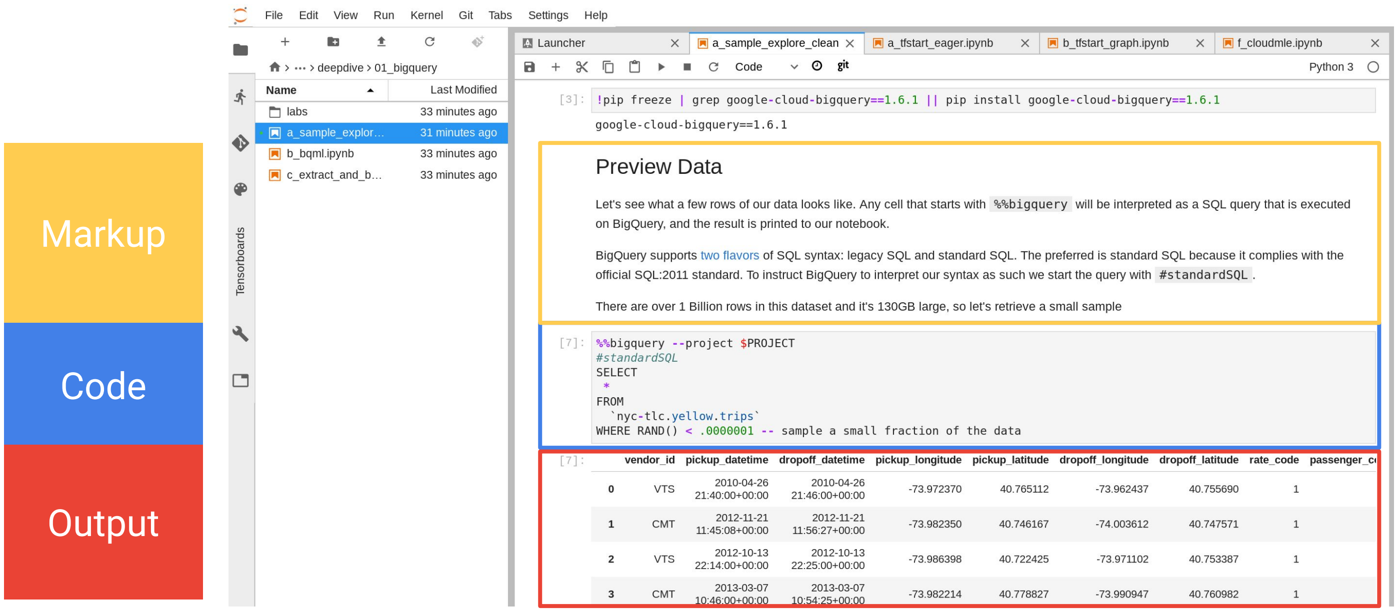
Python notebooks in the Cloud

Supervised Learning

Inclusive ML

Short History of ML

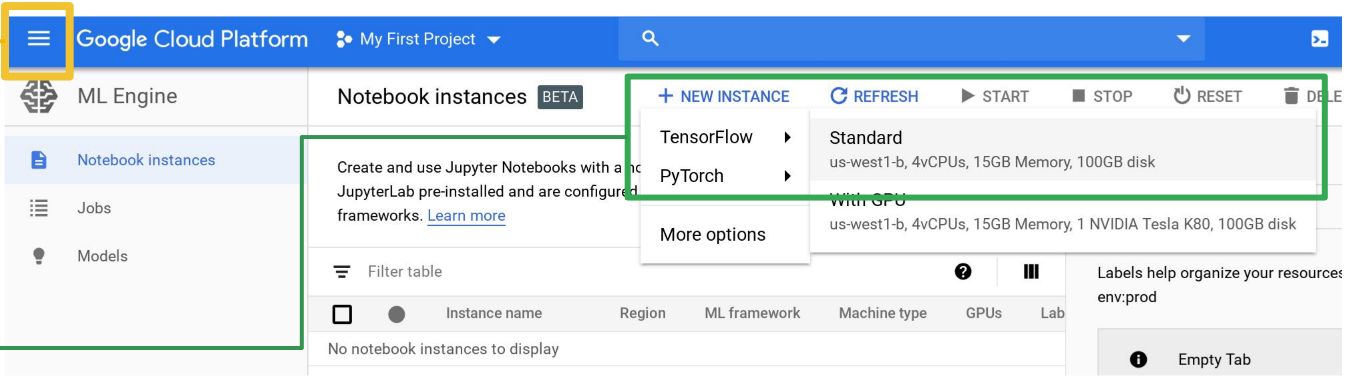
Increasingly, data analysis and ML are carried out in self-descriptive, shareable, executable notebooks



A typical notebook contains code, charts, and explanations.

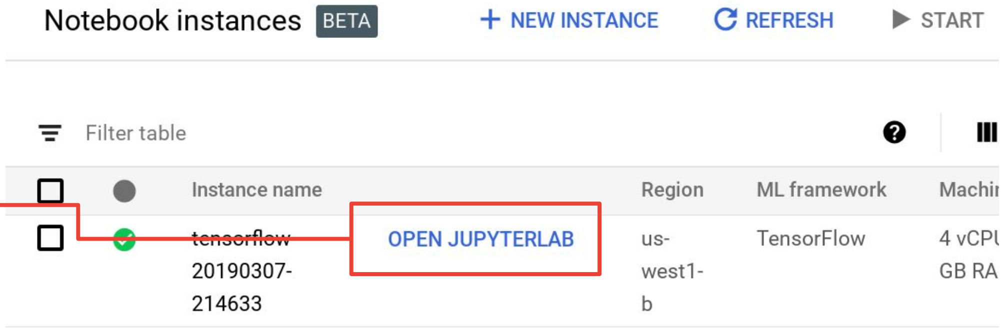
Follow-along: The Easy Way to Make a Notebook

CLME notebooks are found under [ML Engine -> Notebook Instances]



Click “+ New Instances” then TensorFlow -> Standard

Install NVIDIA GPU drivers automatically. The Click “OPEN JUPYTER LAB” after the VM is spun up.



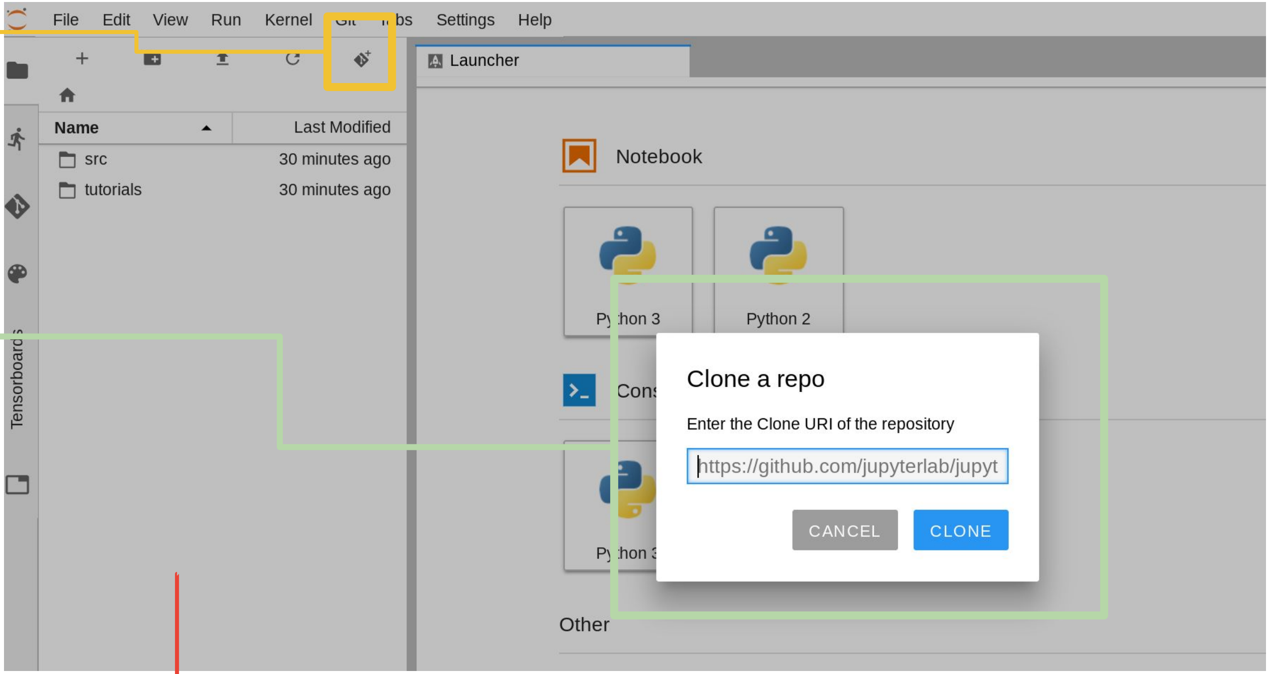
Follow-along: Connecting to Github

Click the git clone icon to clone a repository

Paste the following URL into the address box and click “Clone”

https://github.com/GoogleCloudPlatform/training-data-analyst.git

Double click the “training-data-analyst” folder when it appears here.



Agenda

Python notebooks in the Cloud

Supervised Learning

Inclusive ML

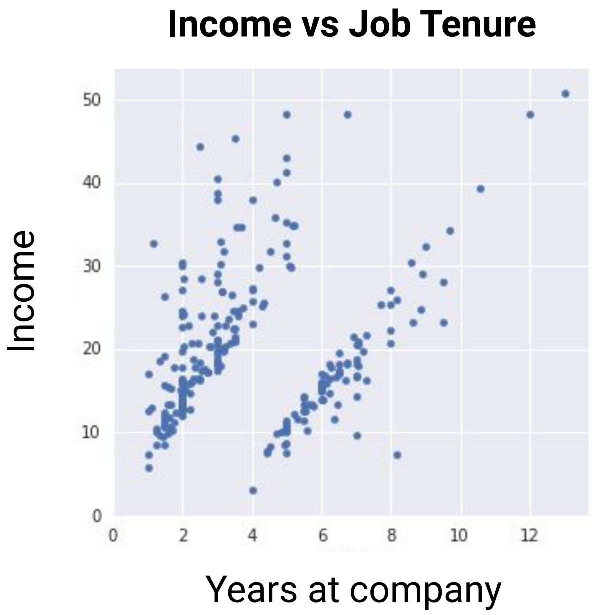
Short History of ML

Unsupervised and supervised learning are the two types of ML algorithms

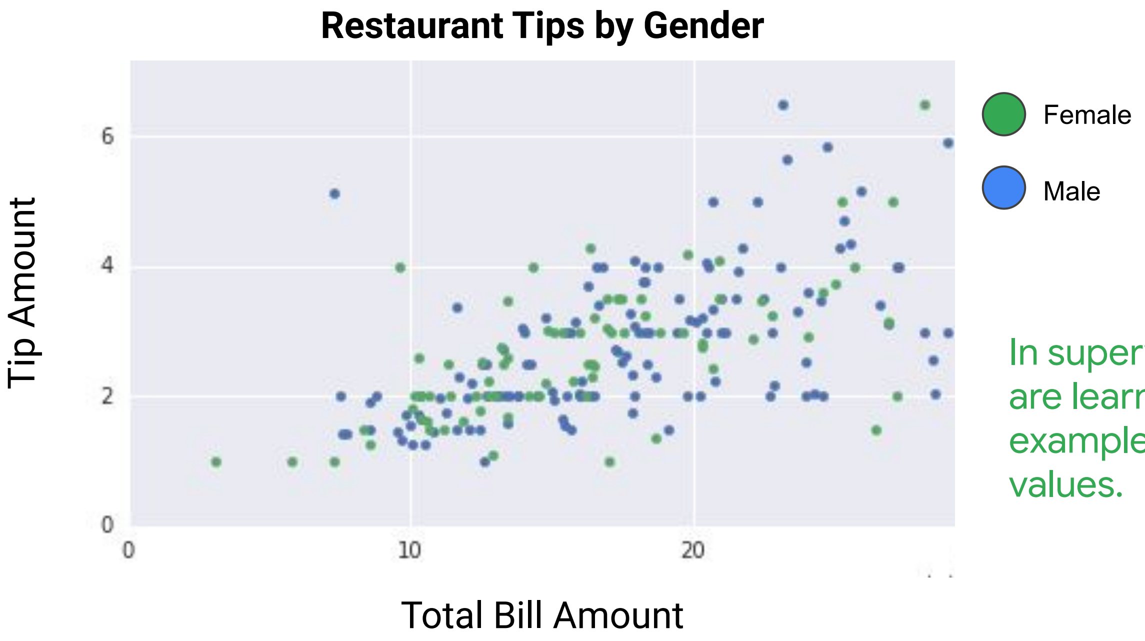
Example Model: Clustering

Is this employee on the “fast-track” or not?

In unsupervised learning, data is not labeled.



Supervised learning implies the data is already labeled



In supervised learning we are learning from past examples to predict future values.

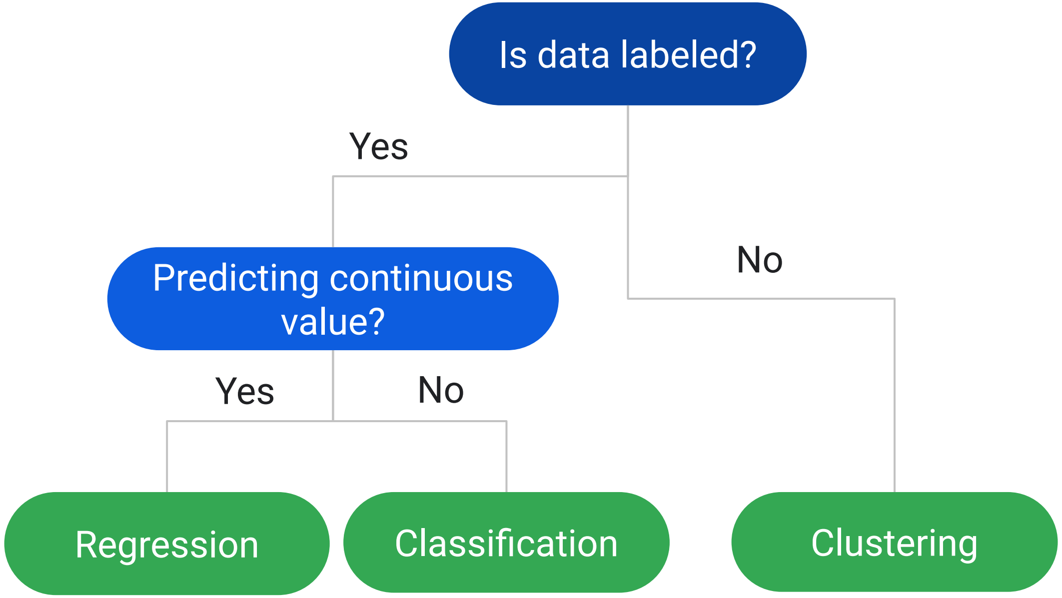
Regression and classification are supervised ML model types

	total_bill	tip	sex	smoker	day	time
1	total_bill	tip	sex	smoker	day	time
2	16.99	1.01	Female	No	Sun	Dinner
3	10.34	1.66	Male	No	Sun	Dinner
4	21.01	3.5	Male	No	Sun	Dinner
5	23.68	3.31	Male	No	Sun	Dinner
6	24.59	3.61	Female	No	Sun	Dinner
7	25.29	4.71	Male	No	Sun	Dinner
8	8.77	2	Male	No	Sun	Dinner
9	26.88	3.12	Male	No	Sun	Dinner

Option 1
Regression Model
Predict the tip amount

Option 2
Classification Model
Predict the sex of the customer

The type of ML problem depends on whether or not you have labeled data and what you are interested in predicting



Quiz: Supervised learning

Imagine you are in banking and you are creating an ML model for detecting if transactions are fraudulent or not. Is this classification or regression and why?

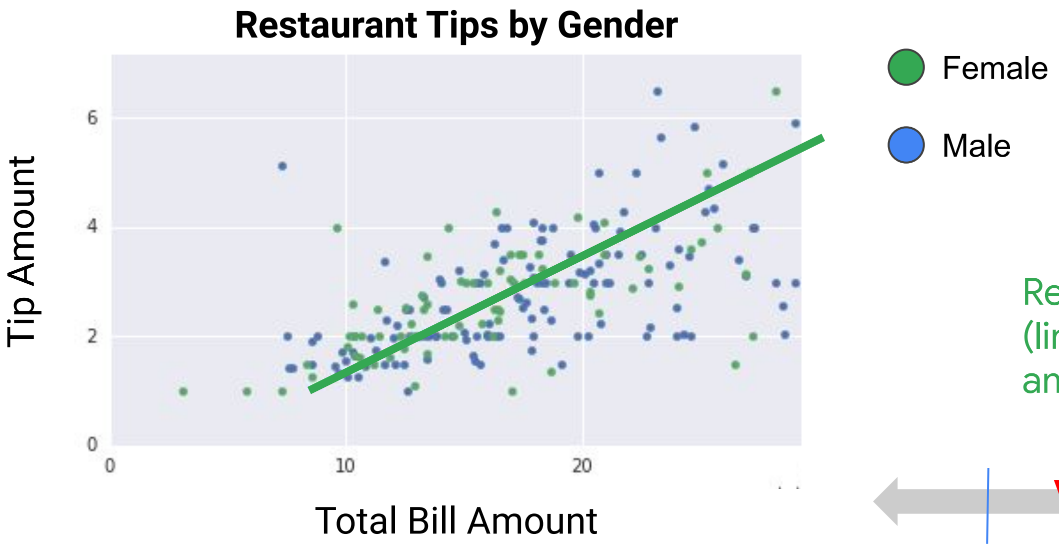
A. Regression, categorical label
B. Regression, continuous label
C. Classification, categorical label
D. Classification, continuous label

Quiz: Supervised learning

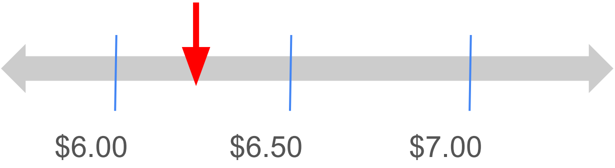
Imagine you are in banking and you are creating an ML model for detecting if transactions are fraudulent or not. Is this classification or regression and why?

A. Regression, categorical label
B. Regression, continuous label
C. $\boxed{\text{Classification, categorical label}}$
D. Classification, continuous label

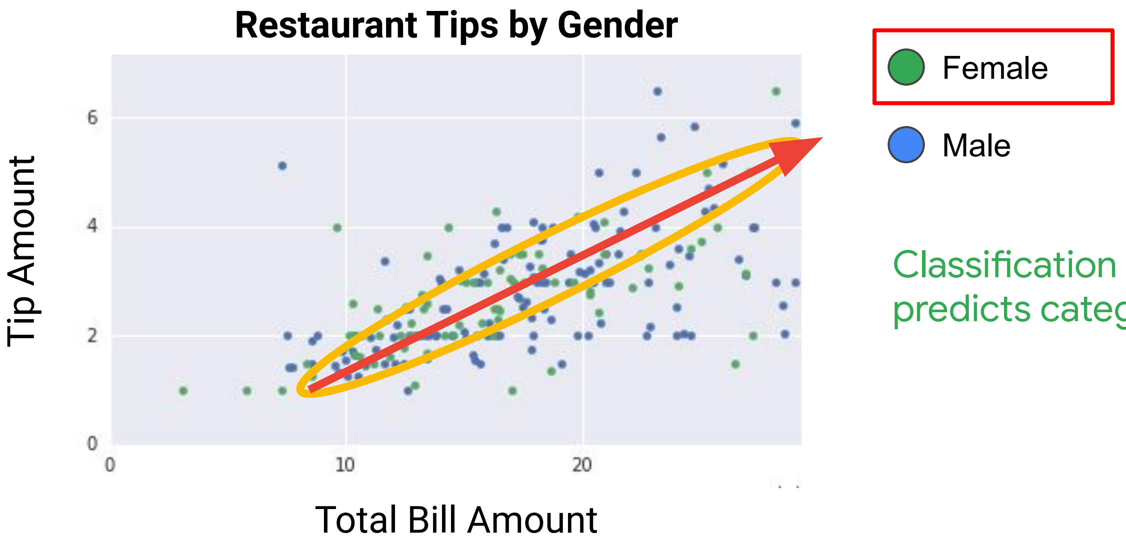
Use regression for predicting continuous label values



Regression Model (linear) predicts tip amount of $6.25.



Use classification for predicting categorical label values



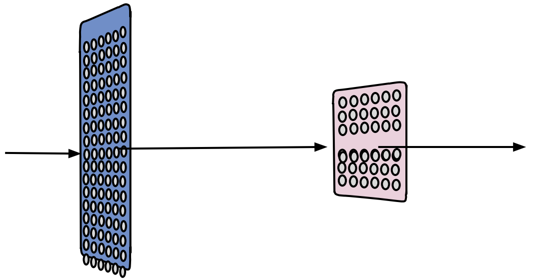
Classification Model (binary) predicts category of Female.

A data warehouse can be a source of structured data training examples for your ML model

```
SELECT
  gestation_weeks,
  mother_age,
  cigarette_use,
  alcohol_use,
  weight_gain_pounds
FROM
  `bigquery-public-data.samples.natality`
WHERE cigarette_use is not null AND alcoho
```

weight	year	mother_age	gestation_weeks	cigarette_use	alcohol_use
7.86	2003	25	39	false	false
7.5	2003	21	39	false	false
8.06	2004	29	40	false	false
7.56	2004	38	37	false	false
7.06	2003	22	38	false	false

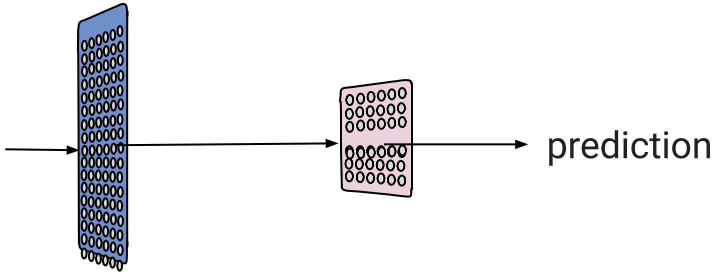
Data on births is sourced from our BigQuery Data Warehouse using SQL.



Since baby weight is a continuous value, use regression to predict

weight	year	mother_age	gestation_weeks	cigarette_use	alcohol_use
7.86	2003	25	39	false	false
7.5	2003	21	39	false	false
8.06	2004	29	40	false	false
7.56	2004	38	37	false	false
7.06	2003	22	38	false	false

Weight is stored as a floating point number, representing a continuous (real) value.

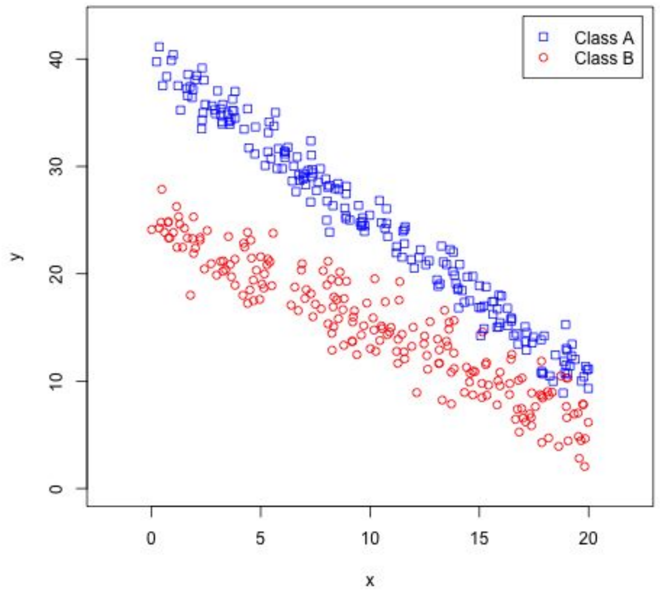


Regression DNN Model

Quiz: Regression/Classification

Is this dataset a good candidate for linear regression and/or linear classification?

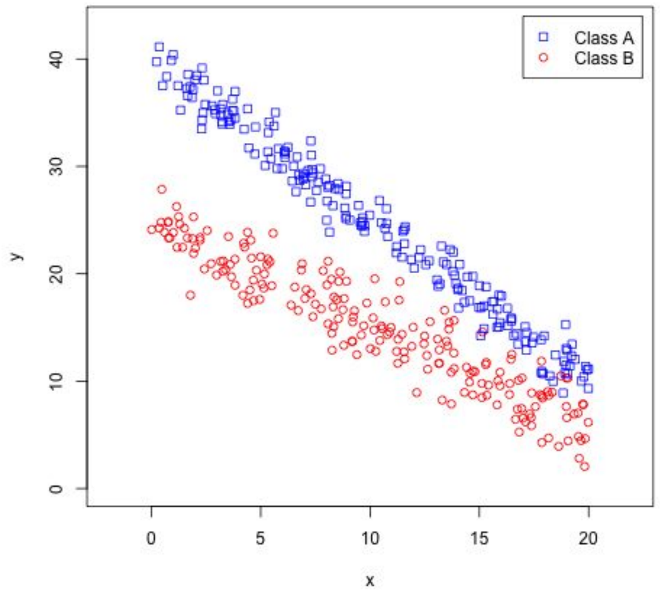
A. Linear classification
B. Both
C. None of the above



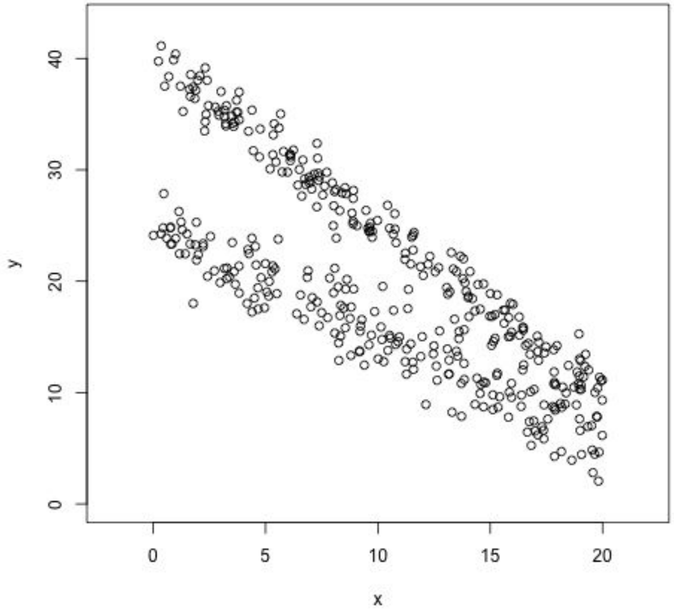
Quiz: Regression/Classification

Is this dataset a good candidate for linear regression and/or linear classification?

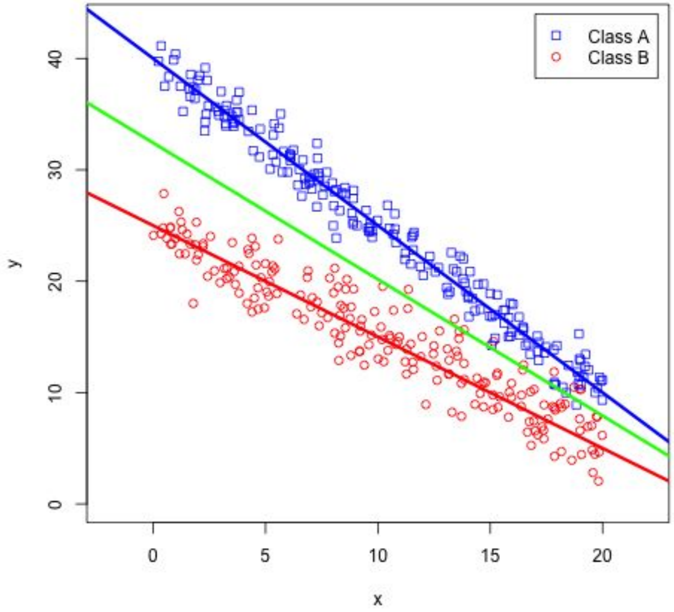
A. Linear classification
B. Both
C. None of the above



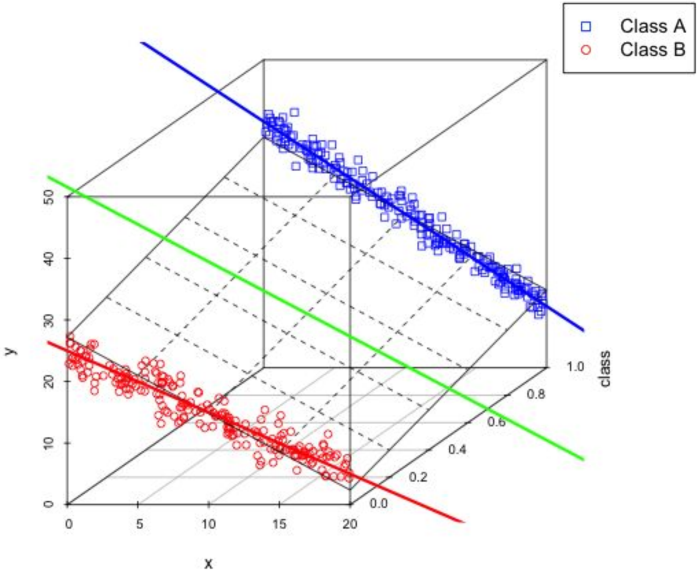
Is this dataset a good candidate for linear regression and/or linear classification?



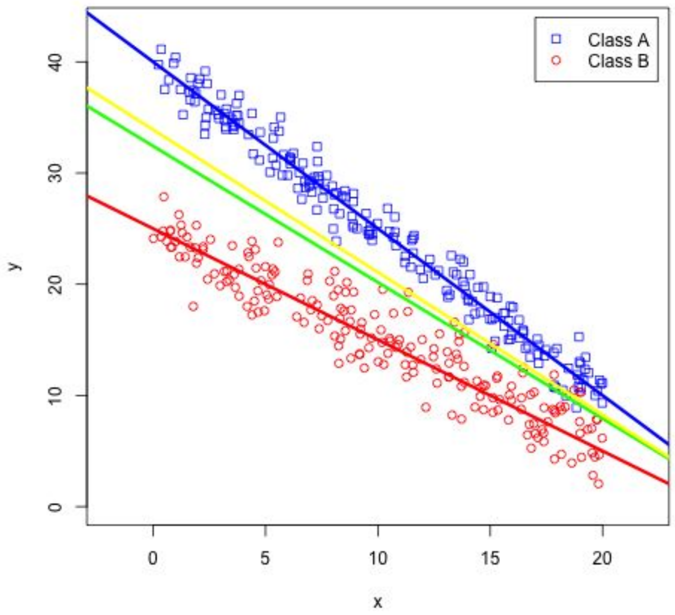
Is this dataset a good candidate for linear regression and/or linear classification?



Is this dataset a good candidate for linear regression and/or linear classification?



Is this dataset a good candidate for linear regression and/or linear classification?



Agenda

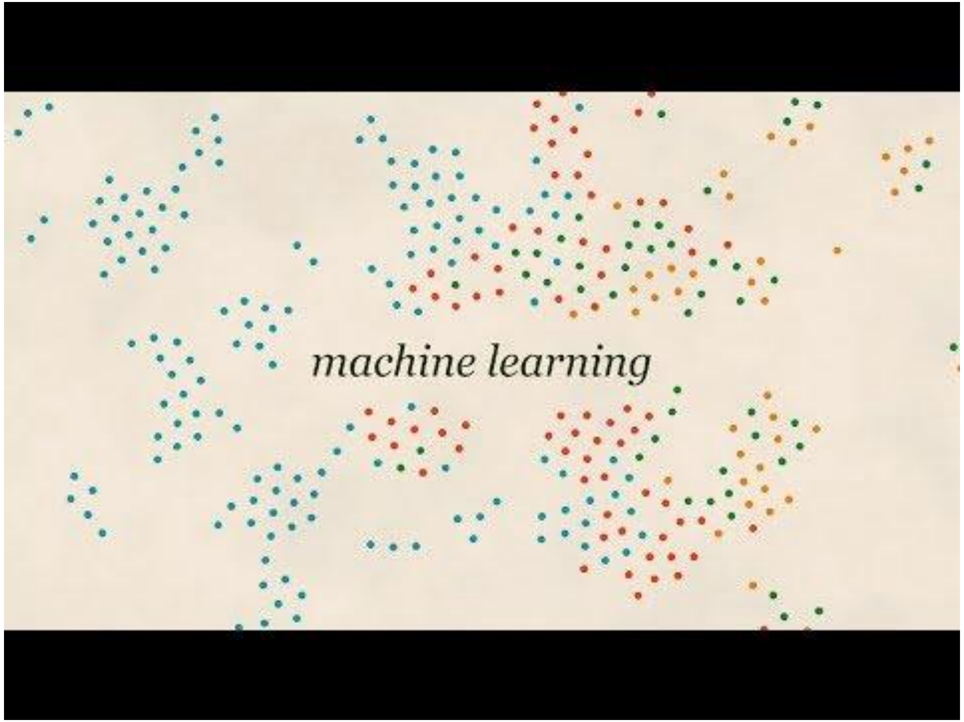
Python notebooks in the Cloud

Supervised Learning

Inclusive ML

Short History of ML

Human biases lead to biases in ML models



Unconscious biases exist in data

Unconscious bias from “the world” that we might reflect in ML when using existing data

Collecting data

Labeling data

Unconscious bias in our procedures that we might reflect in our ML

Examples of Human Biases in Data

Reporting bias

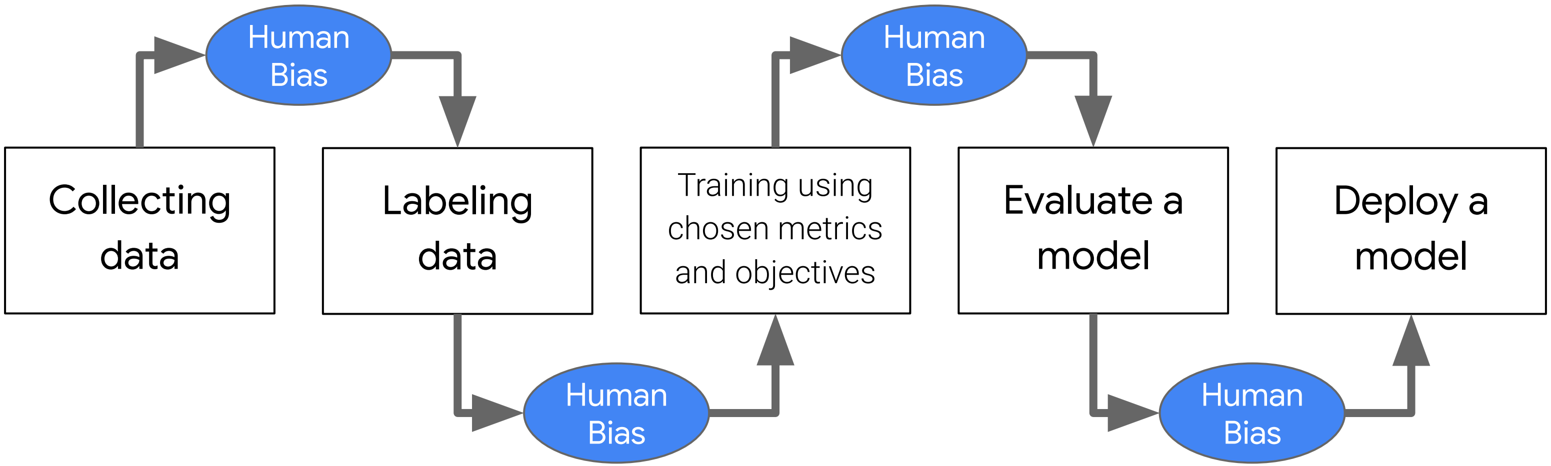
Selection bias

Examples of Human Biases in Collection and Labeling

Confirmation bias

Automation bias

A typical ML pipeline *with bias*

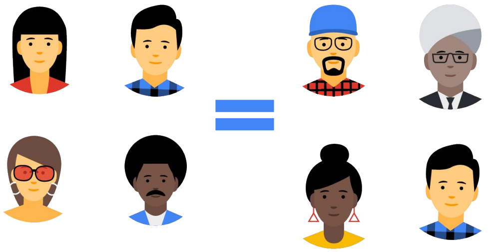


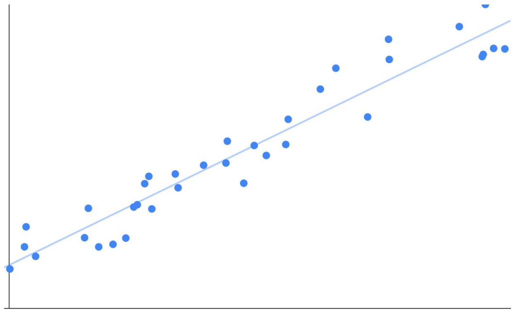
2 Avoid creating or reinforcing unfair bias

ML models learn from existing data collected from the real world, and so an accurate model may learn or even amplify problematic pre-existing biases in the data based on race, gender, religion, or other characteristics.

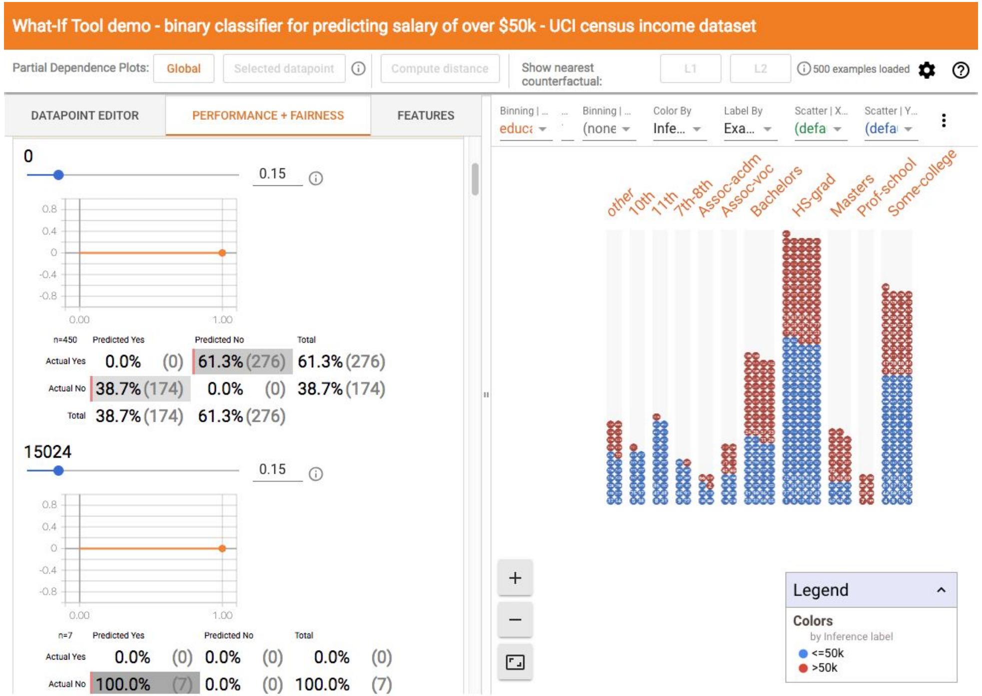
ai.google/principles

A Checklist for Bias-Related Issues





Tools for Responsible AI



Agenda

Python notebooks in the Cloud

Supervised Learning

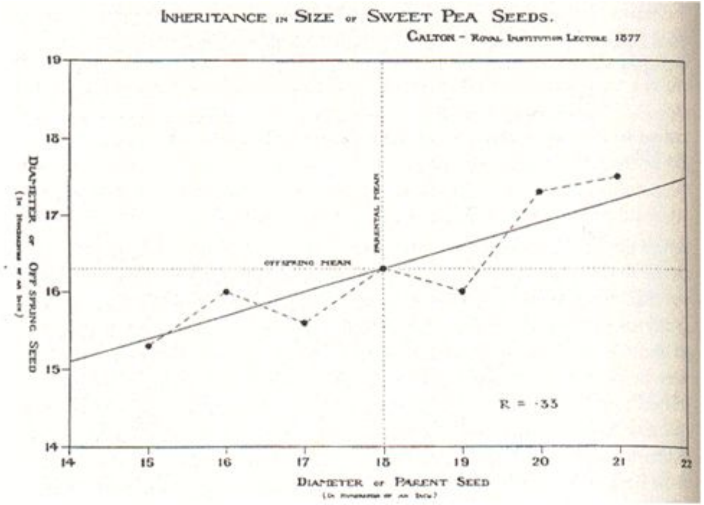
Inclusive ML

Short History of ML

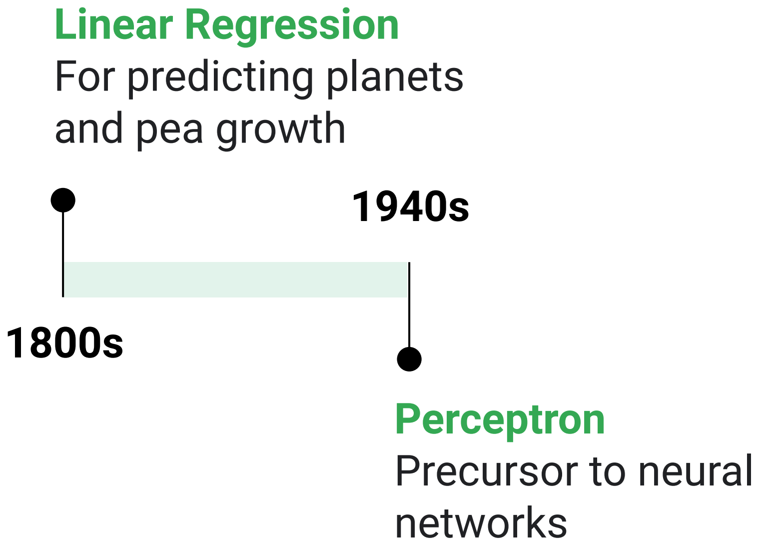
Linear regression was invented when computations were done by hand, but it continues to work well for large datasets

Linear Regression
For predicting planets and pea growth

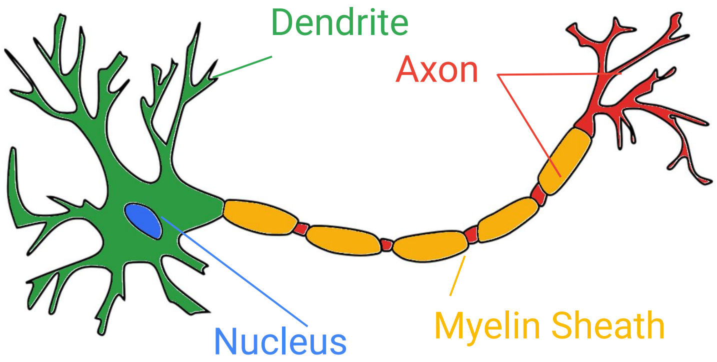
1800s

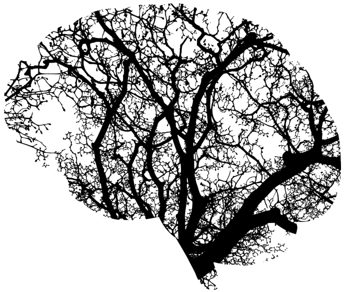


The perceptron was a computational model of a neuron

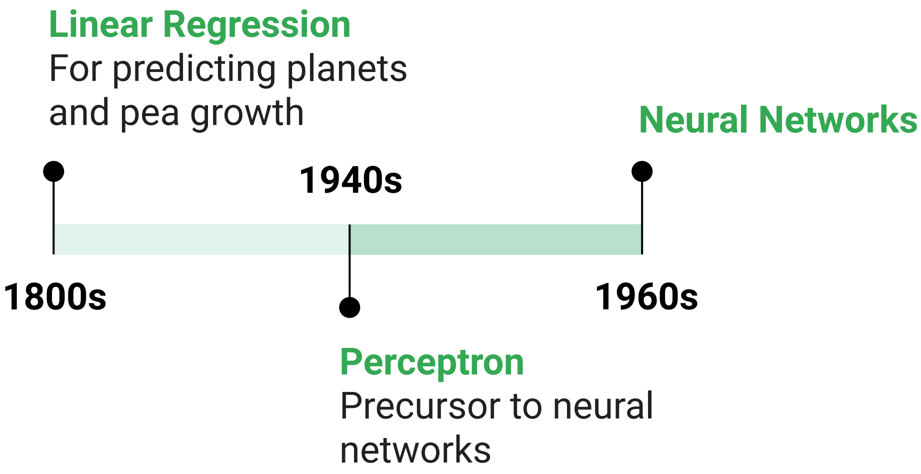


Perceptron motivation: Neurons

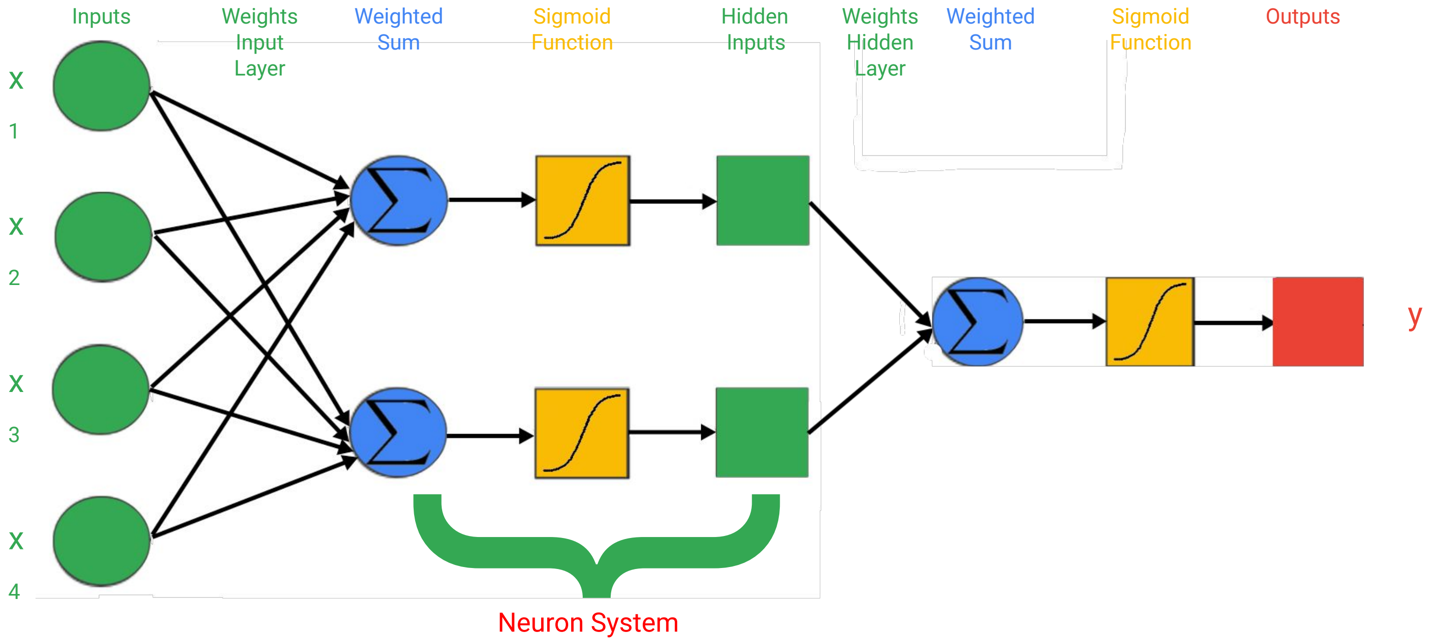




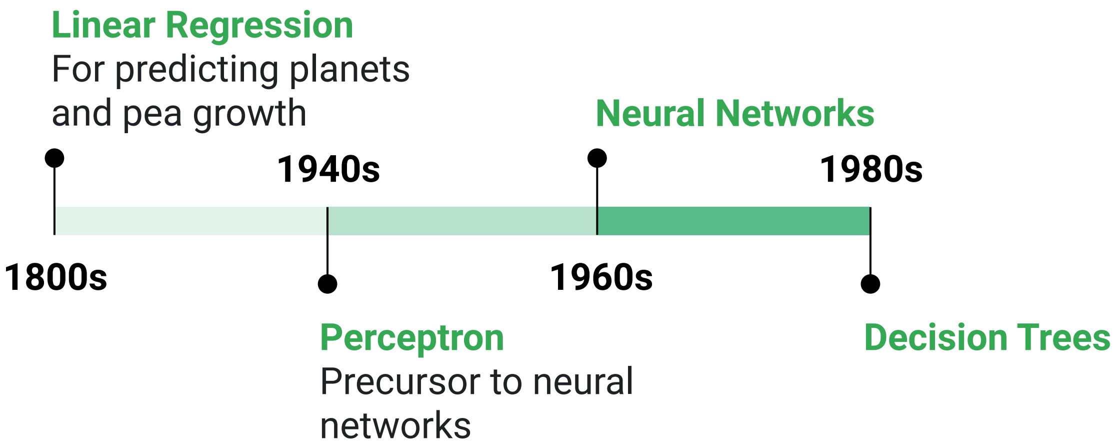
Neural networks combine layers of perceptrons, making them more powerful but also harder to train effectively



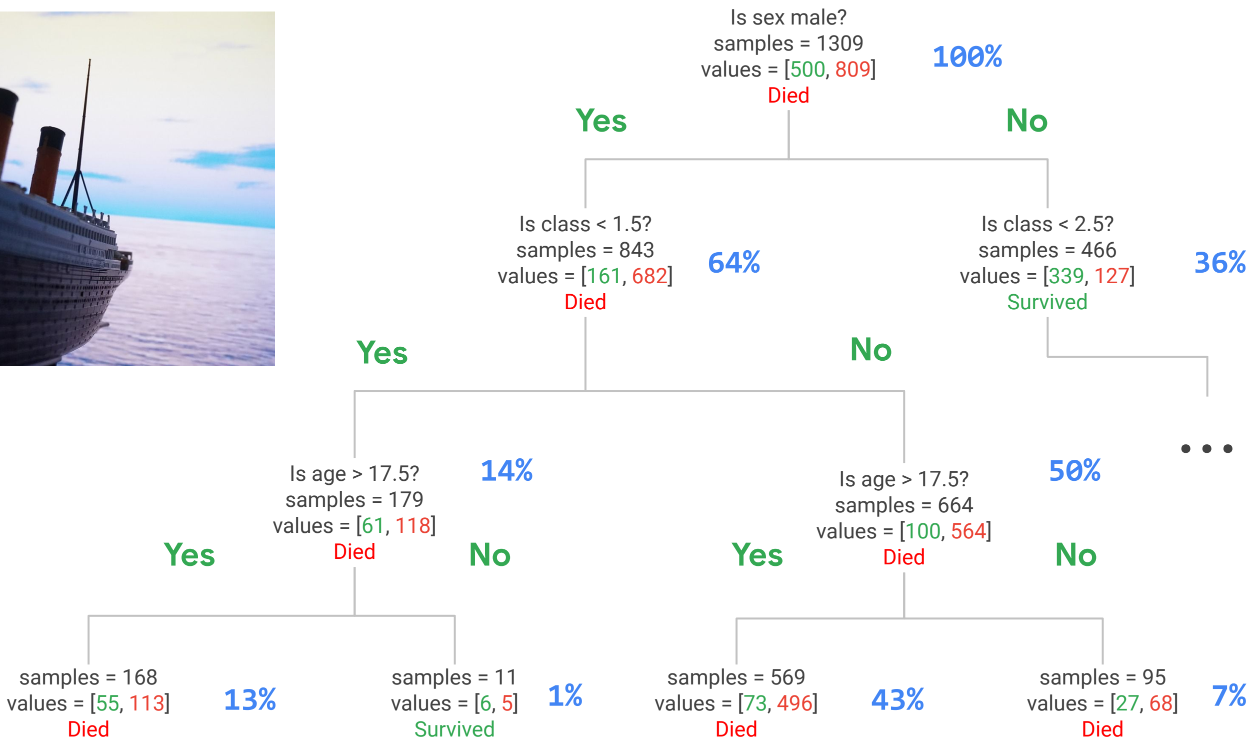
Neural networks: Multi-layer perceptron



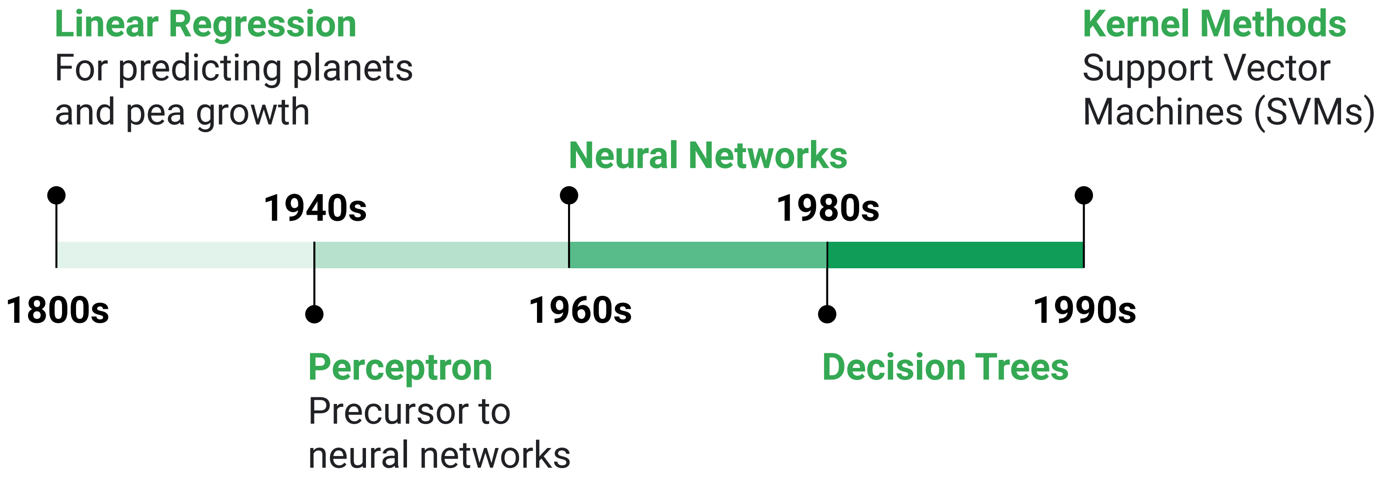
Decision trees build piecewise linear decision boundaries, are easy to train, and are easy for humans to interpret



Decision trees and the Titanic

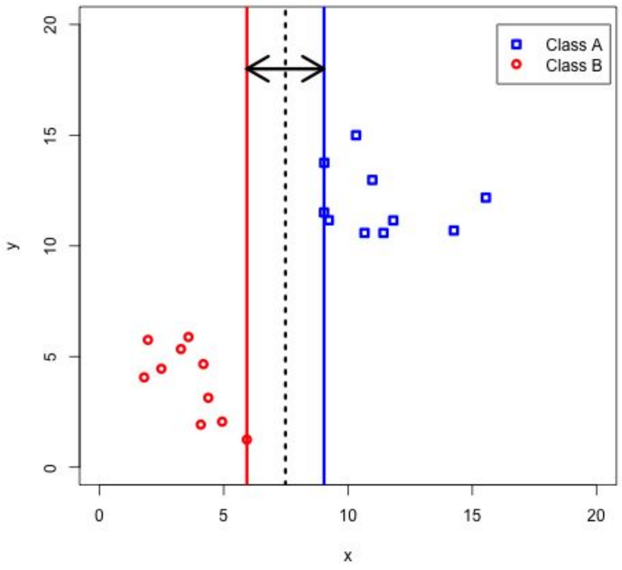


Support vector machines are nonlinear models that build maximum marginal boundaries in hyperspace

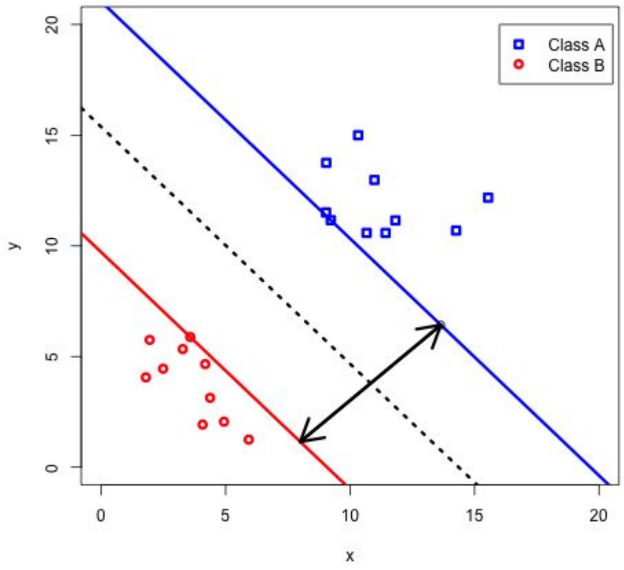


SVMs maximize the margin between two classes

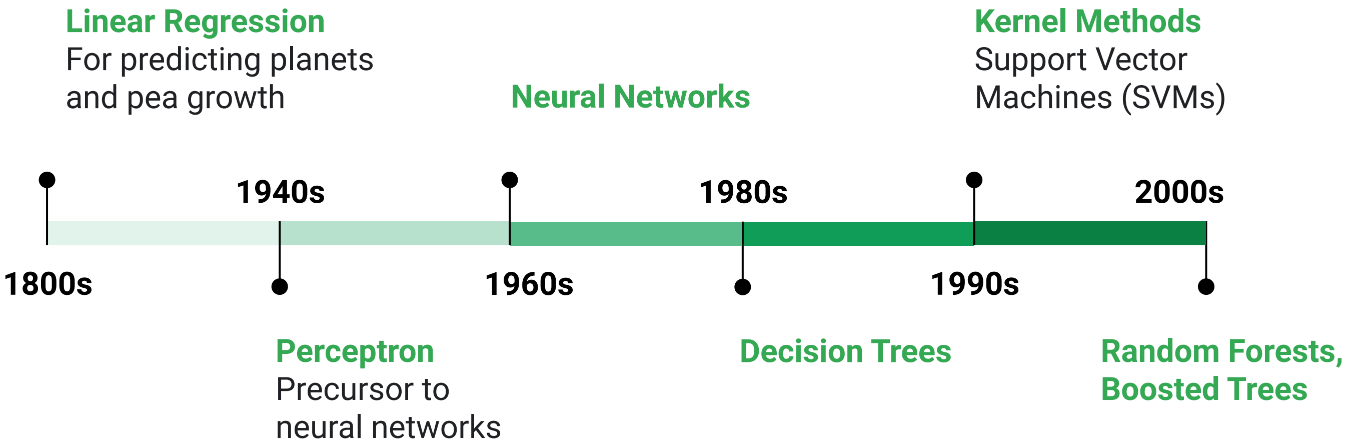
Small margin



Large margin

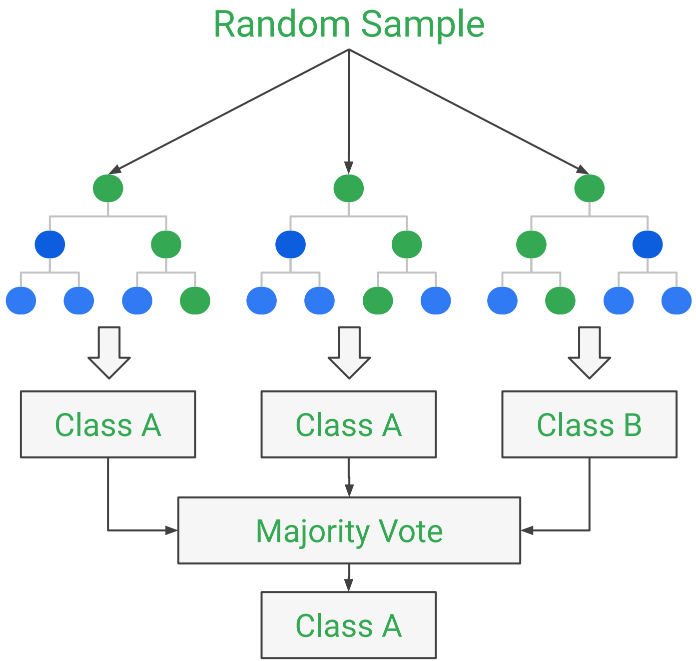


Random forests, bagging, and boosting are very effective predictors built by combining lots of very simple predictors

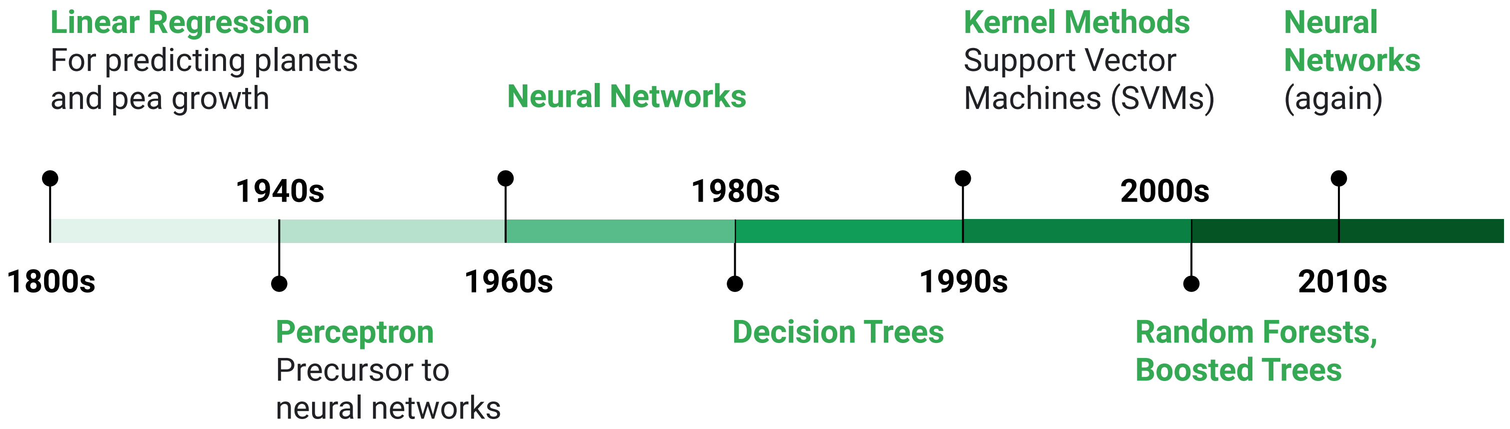


In many real-world problems, the highest quality is often attained with these methods

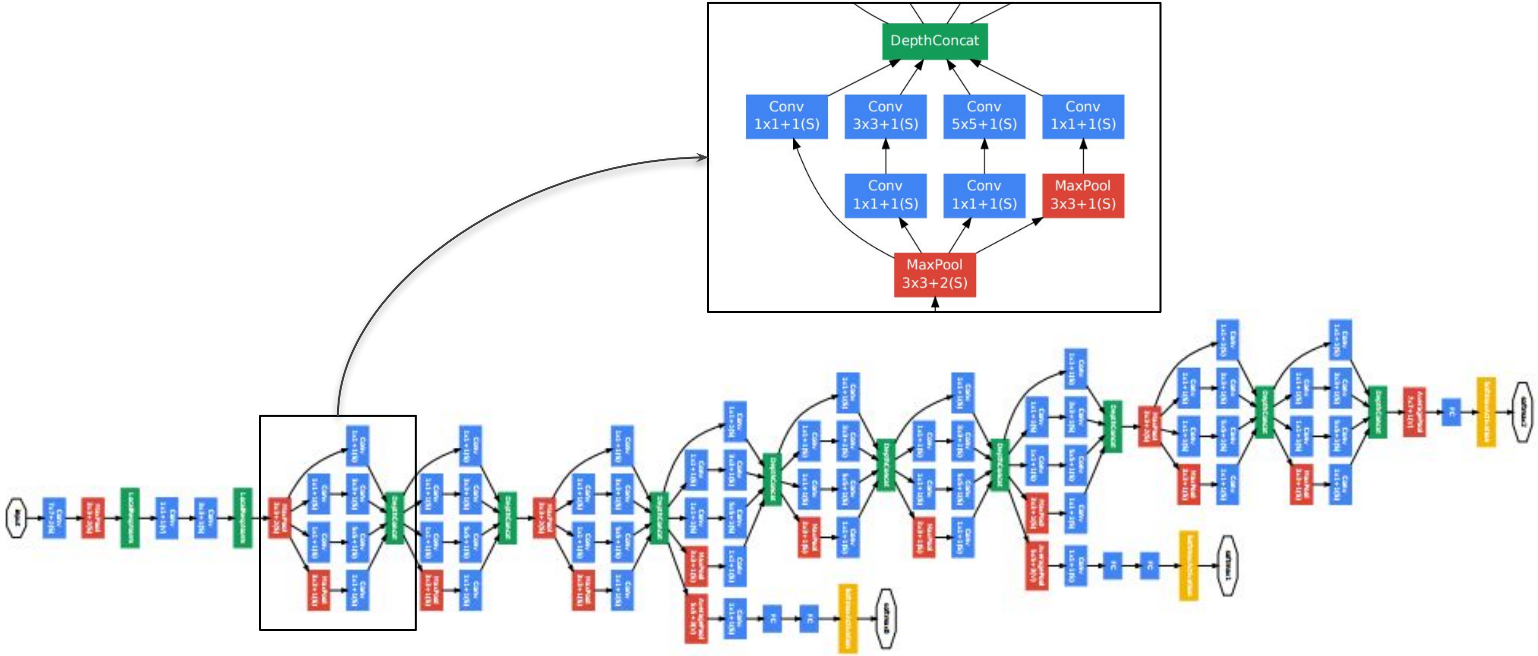
Random forest: Strong learner from many weak learners



With the advantage of technical improvements, more data, and computational power, neural networks made a comeback

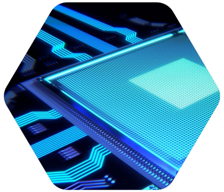


Inception/GoogLeNet Deep Neural Network



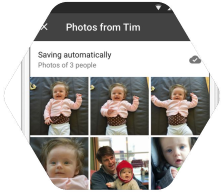
Neural networks are outperforming most other approaches in many domains

Large amounts of data



Available Computational Power

Available Infrastructure



Tasks and Goals we care about

Note that there are no models that are universally better, they're just different.