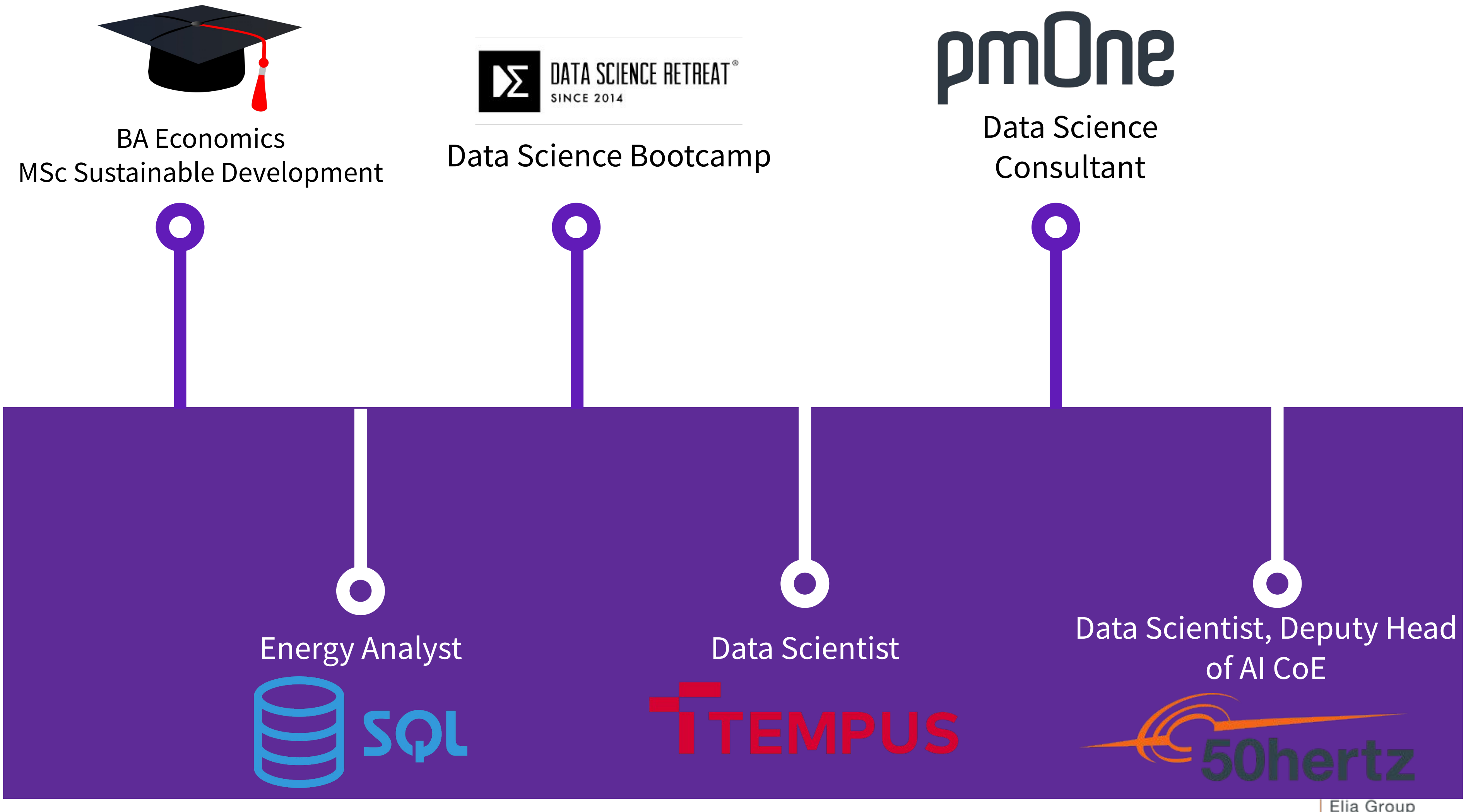
Data Science Fundamentals

Data Science Retreat - 2022
Rachel Berryman



BA Economics
MSc Sustainable Development

DATA SCIENCE RETREAT®
SINCE 2014

Data Science Bootcamp

pmOne

Data Science
Consultant

Energy Analyst

SQL

Data Scientist

TEMPUS

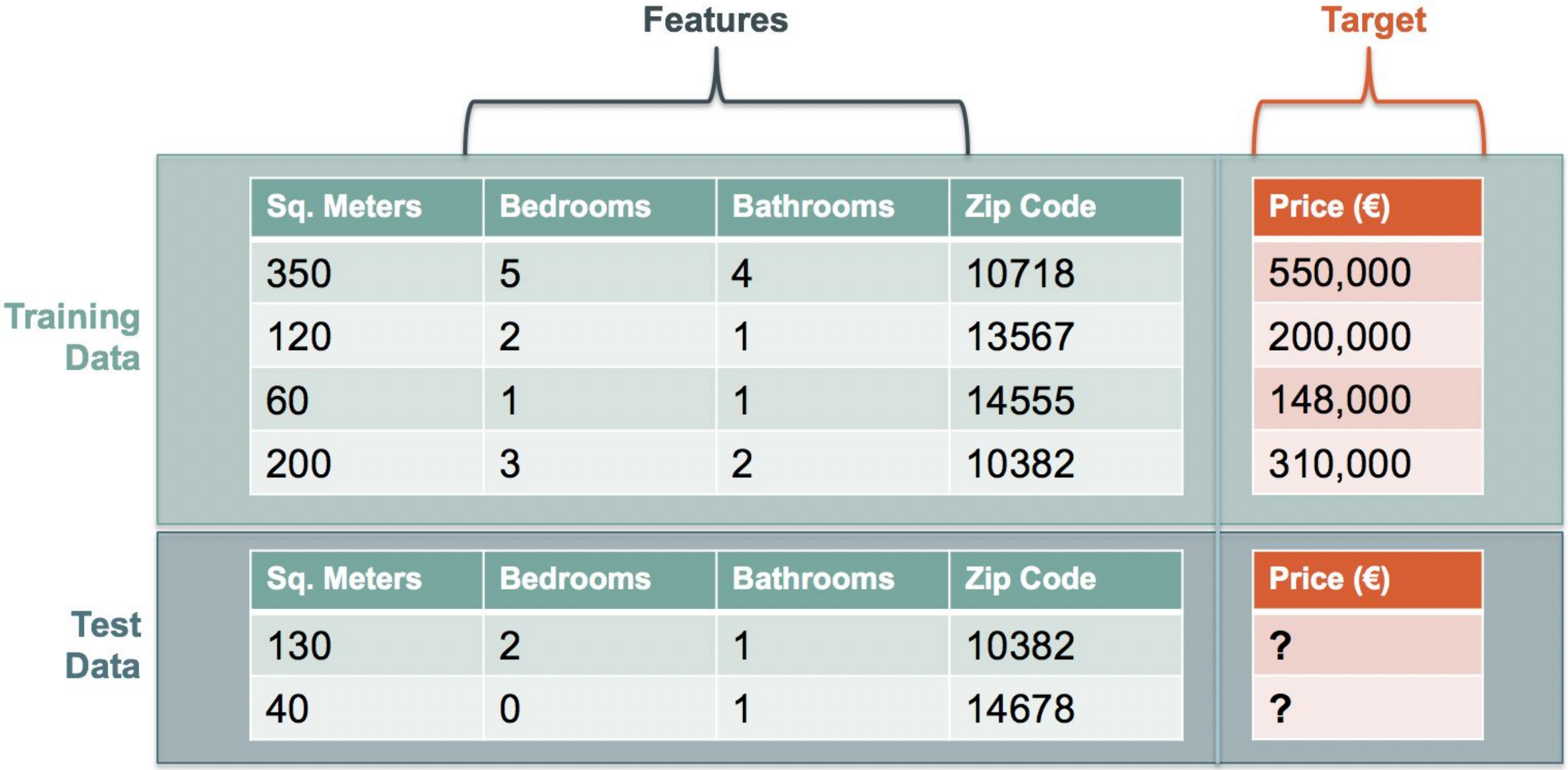
Data Scientist, Deputy Head
of AI CoE

50hertz
Elia Group

Machine Learning

Machine Learning is an application of artificial intelligence where a computer/machine learns from the past experiences (input data) and makes future predictions.

Example dataset



Features

Target

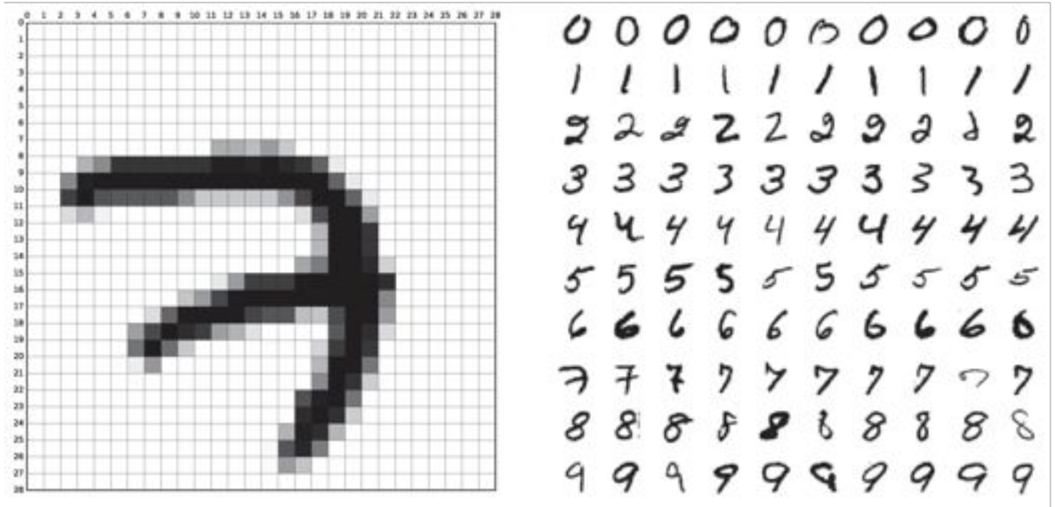
Training Data

Sq. Meters	Bedrooms	Bathrooms	Zip Code	Price (€)
350	5	4	10718	550,000
120	2	1	13567	200,000
60	1	1	14555	148,000
200	3	2	10382	310,000

Test Data

Sq. Meters	Bedrooms	Bathrooms	Zip Code	Price (€)
130	2	1	10382	?
40	0	1	14678	?

Example dataset



(a) MNIST sample belonging to the digit '7'. (b) 100 samples from the MNIST training set.

Labels

0

1

2

3

4

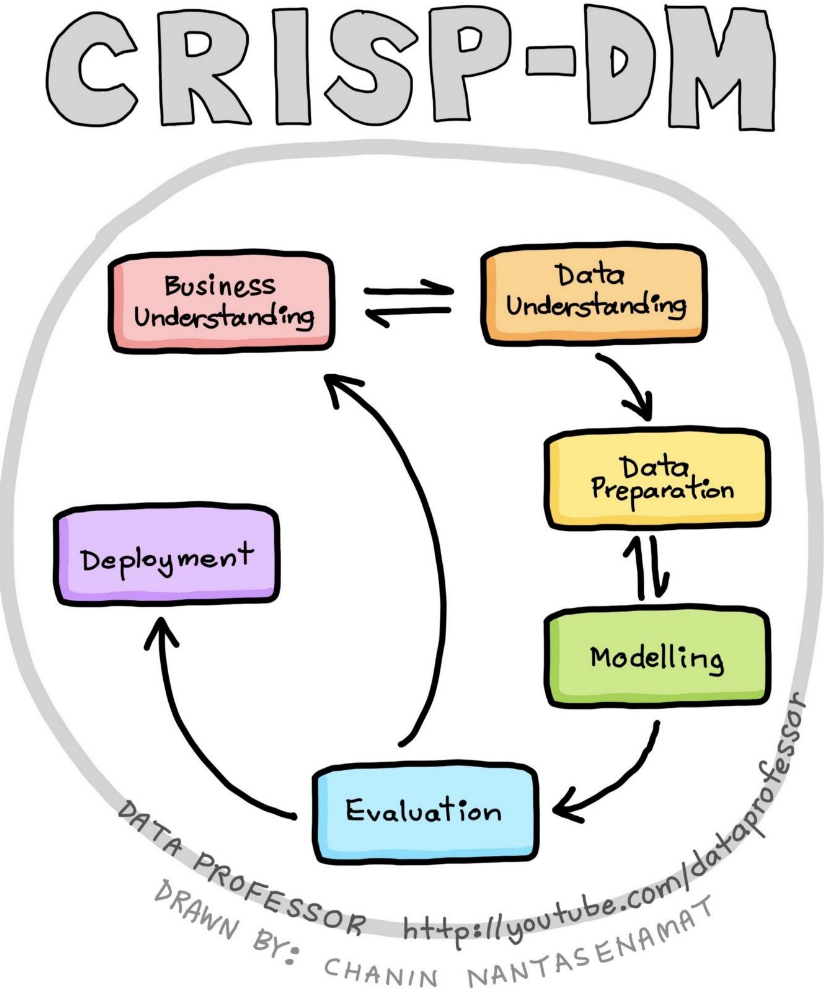
5

6

7

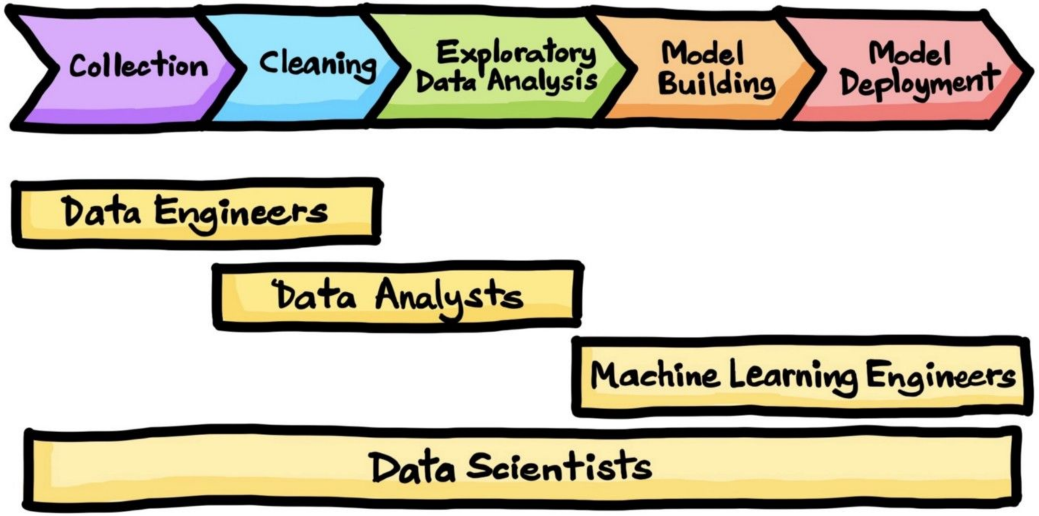
8

9

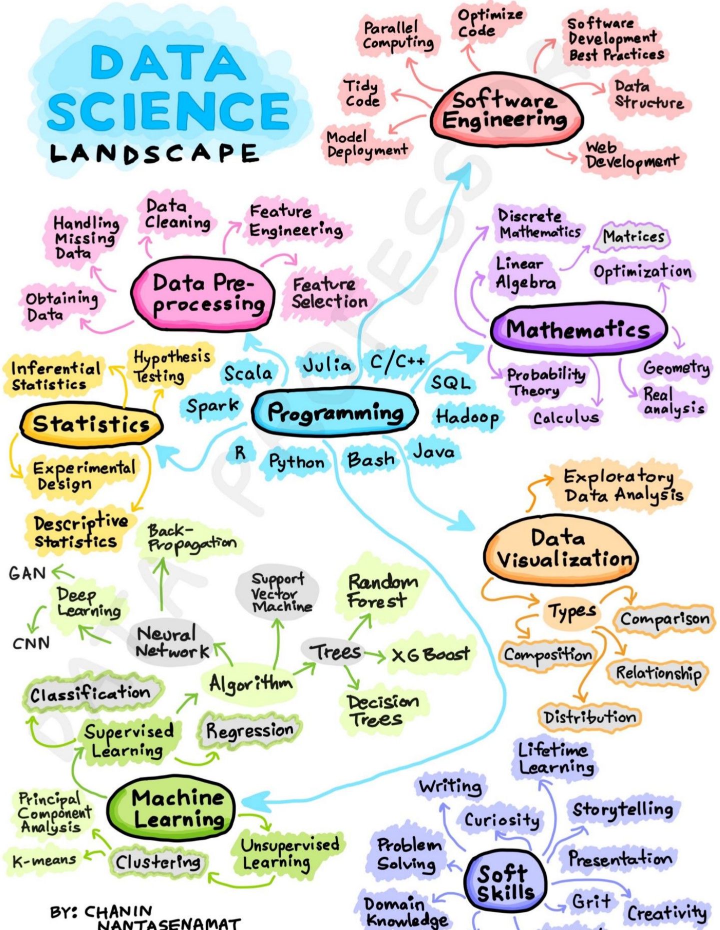


Source: https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

THE DATA SCIENCE PROCESS



DATA SCIENCE LANDSCAPE



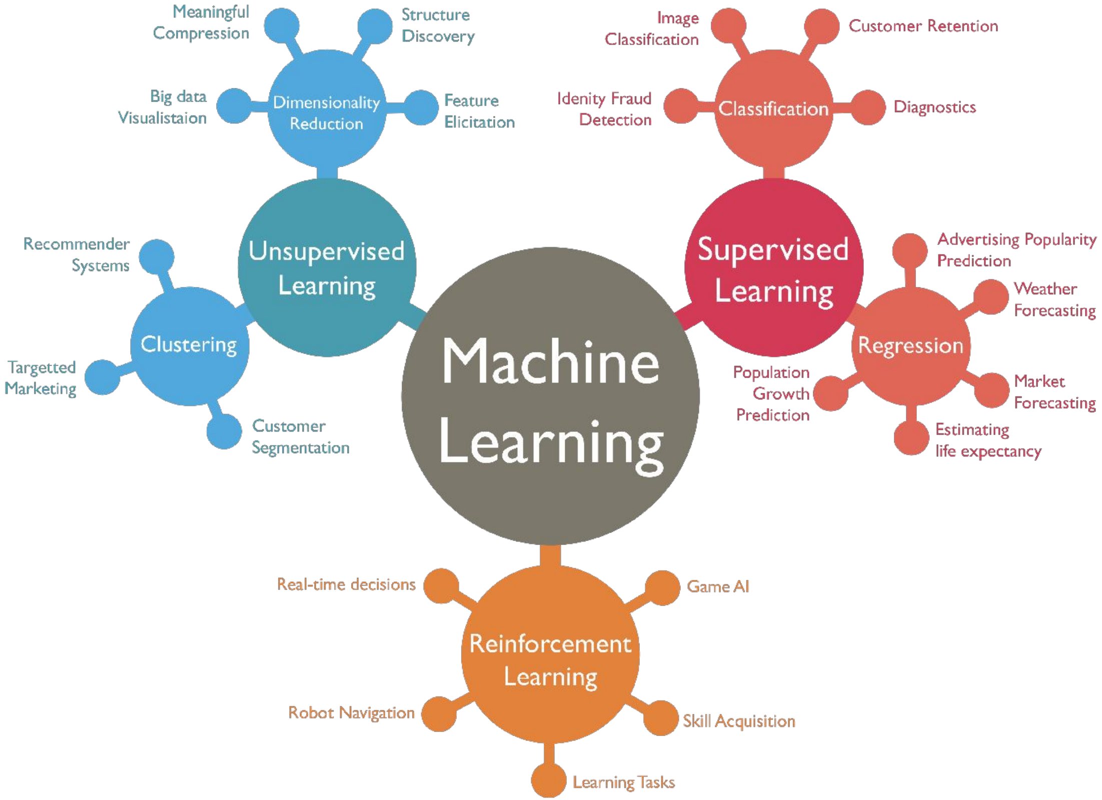
BY: CHANIN NANTASENAMAT

DATA PROFESSOR

http://youtube.com/dataprofessor

FEBRUARY 14, 2020

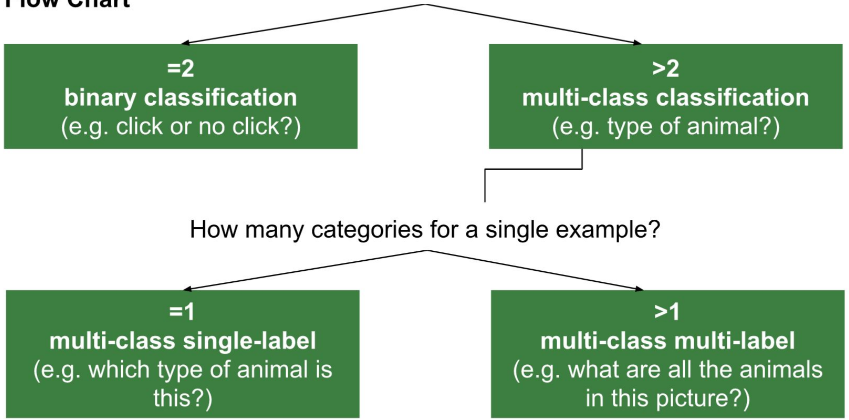
0. Framing the Problem



0. Framing the Problem

Classification Flow Chart

How many categories to pick from?



=2
binary classification
(e.g. click or no click?)

>2
multi-class classification
(e.g. type of animal?)

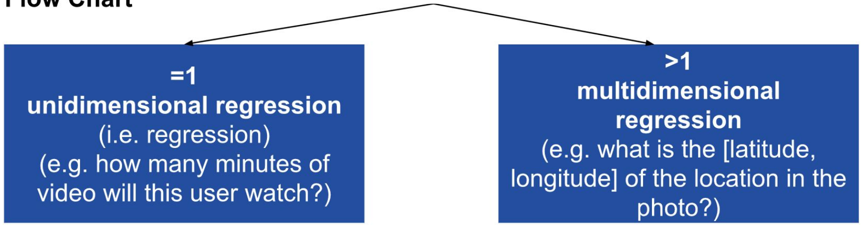
How many categories for a single example?

=1
multi-class single-label
(e.g. which type of animal is this?)

>1
multi-class multi-label
(e.g. what are all the animals in this picture?)

Regression Flow Chart

How many numbers are output?



=1
unidimensional regression
(i.e. regression)
(e.g. how many minutes of video will this user watch?)

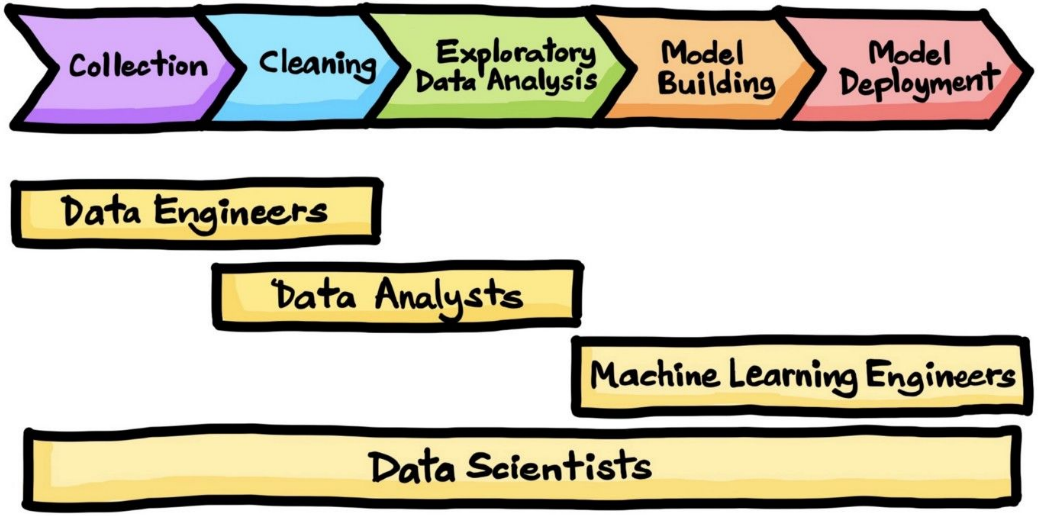
>1
multidimensional regression
(e.g. what is the [latitude, longitude] of the location in the photo?)

0. Framing the Problem

“Translate” your problem into an ML problem.

- Spam email filter
- Predicting popularity of newly posted youtube video
- Train a robot how to stand on its own
- Identify bias in tweets

THE DATA SCIENCE PROCESS



1. Collecting Data

Have to find the right data for your problem

- Both Target & Features

Is your dataset labeled? Do you need labels?

Where will you get data? Is it publicly available?

What form is your data in?

1. Collecting Data

Tabluar data formats:

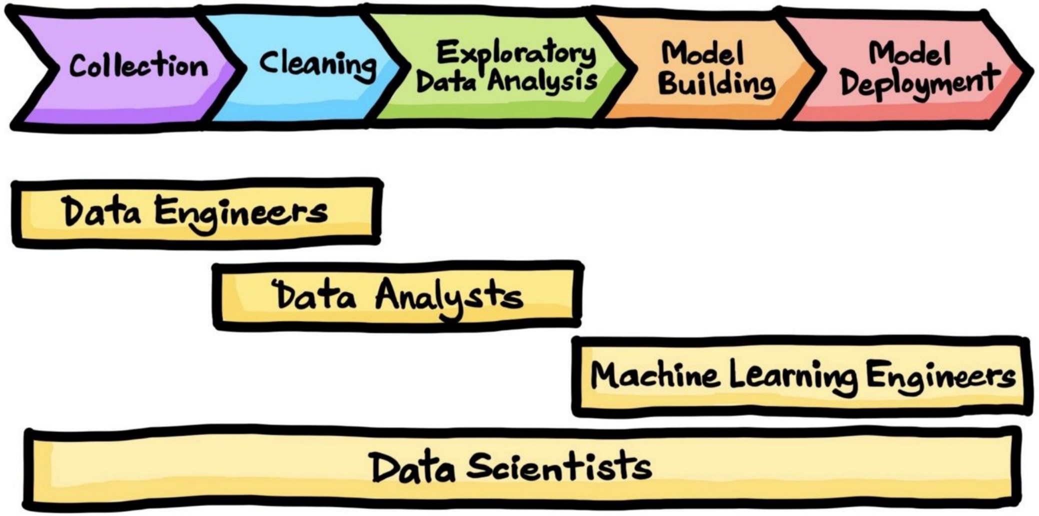
- Continuous
- Categorical
- Ordinal
- Binary
- Time

1. Collecting Data

Consider when starting your new ML project…

- Do I need to get data from other external sources?
- Do I actually need to use all the data I have?
- Is this data likely to help the model learn?
- Do you have enough data?
- Do you have enough positive labels?

THE DATA SCIENCE PROCESS



2. Cleaning Data

Data Characterization

- Quality: outliers / missing values
- Quantity: rows & columns
- Diversity: does the distribution match the test set
- Cardinality: number of unique values
- Dimensionality
- Sparsity

2. Cleaning Data

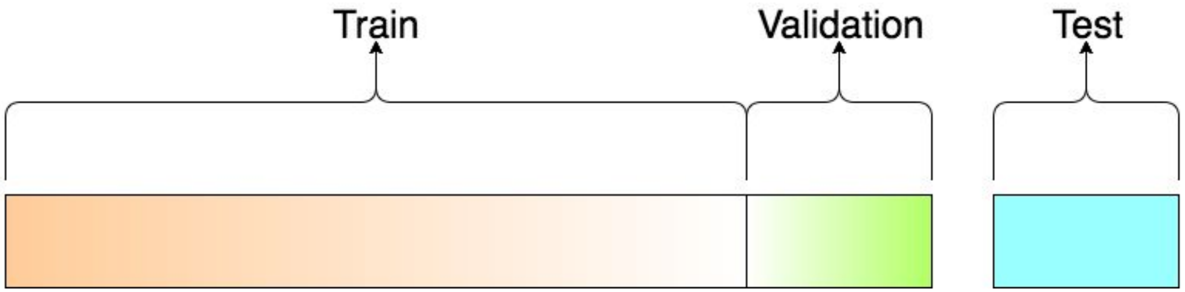
Data Characterization

- Stationarity
  - iterating on data
  - new / different / more customers
  - environment (interest rates changing)
  - model predictions influencing the data (recommendation, fraud)
- Duplicates
- Class imbalance
- Biased sampling

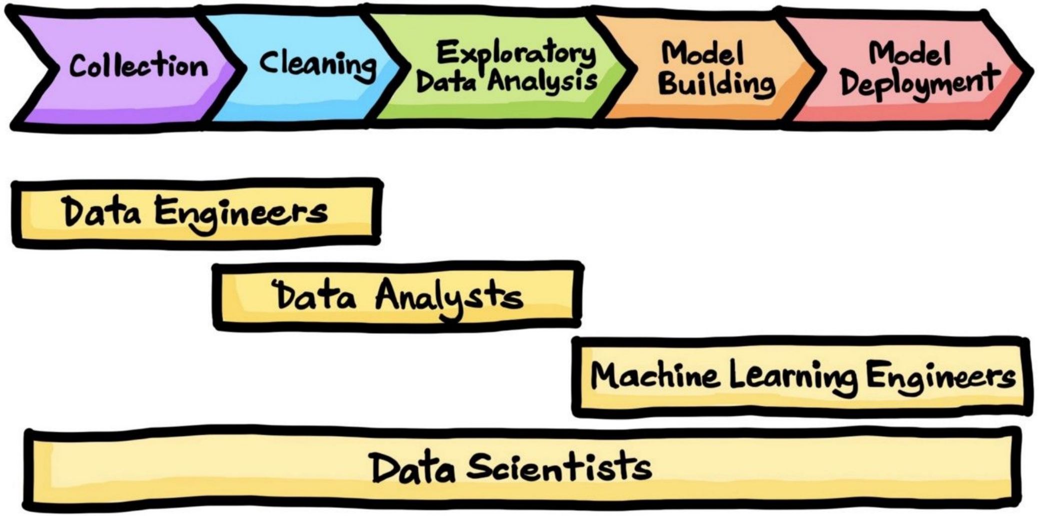
2. Train-test-validation split

Need to hold out the test set right away to prevent data leakage

Best to do this before making any new data/columns



THE DATA SCIENCE PROCESS



3. Data Exploration & Analysis

Understanding your data and features

Removing Outliers

Uni- and bi-variate analysis

Business understanding

VISUALIZATIONS

3. Data Exploration & Analysis

Visualization libraries

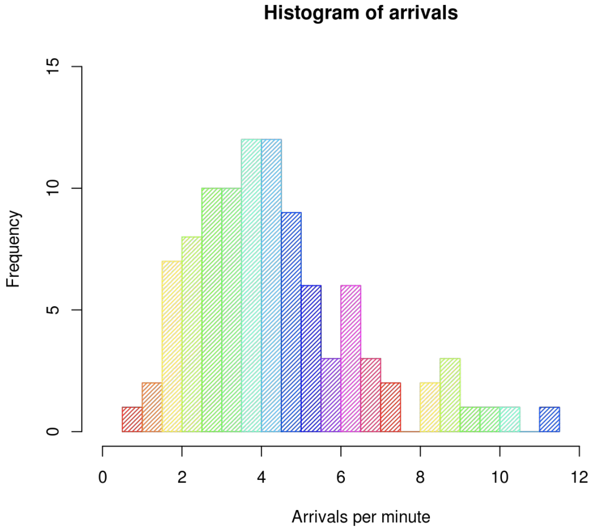
- Matplotlib
- Seaborn
- plotly

Visualization notebook on getting started

3. Data Exploration & Analysis

Visualization to always do

- Correlation matrix
- Plot the target



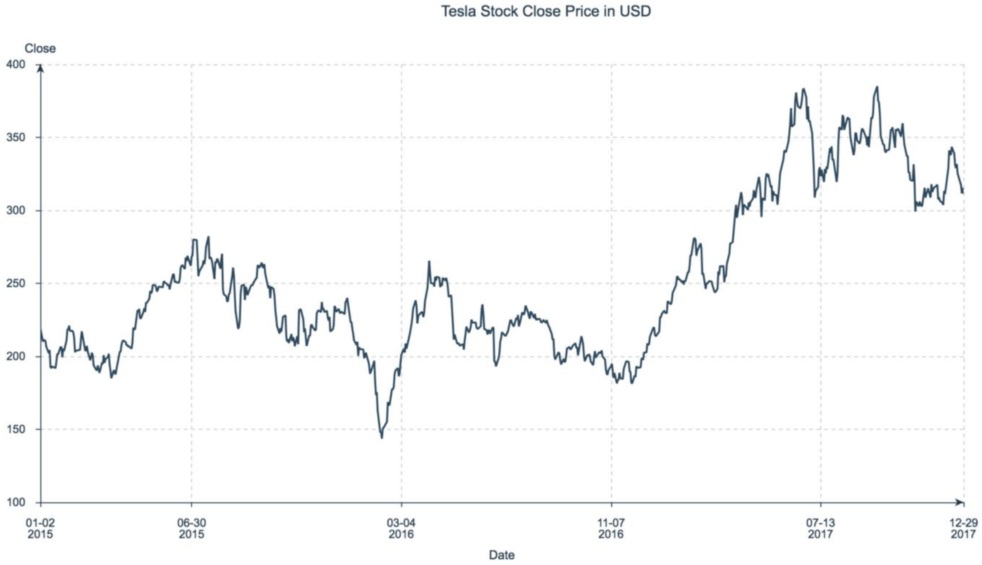
3. Data Exploration & Analysis

Tailor your visualizations to the problem at hand

Hernandez
Robinson
Jackson
Garcia
Wilson
Hall
Evans
Thomas
Williams
Anderson
Davis
Smith
Martin
Miller
Clarke
Davies
White
Brown
Johnson
Walker
Wood
Rodriguez
Jones
Taylor
Roberts
Wright
Moore
Martinez
Green
Thompson

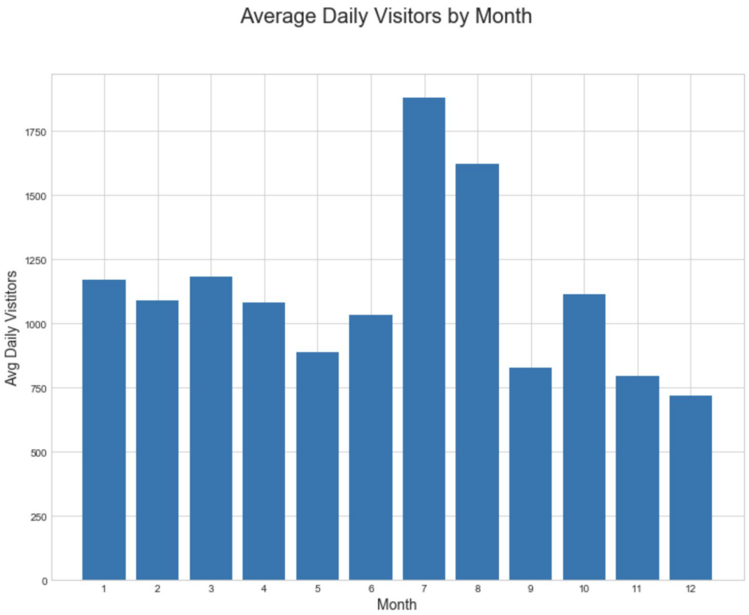
3. Data Exploration & Analysis

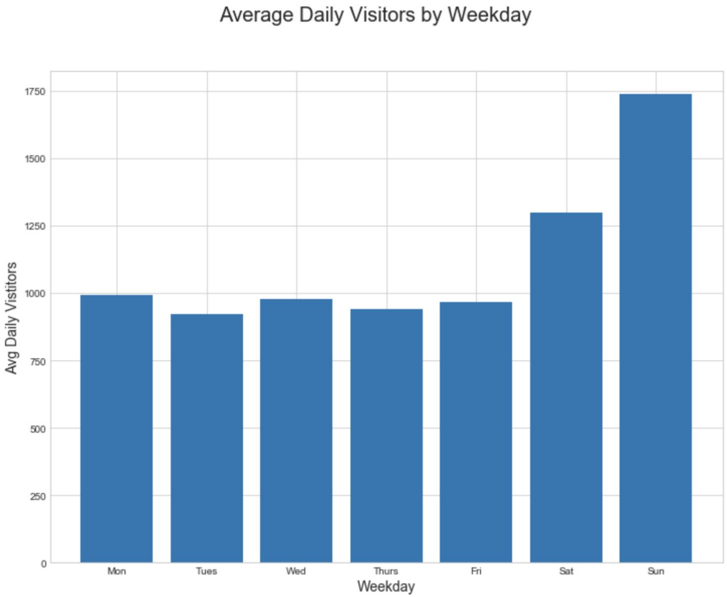
Tailor your visualizations to the problem at hand



3. Data Exploration & Analysis

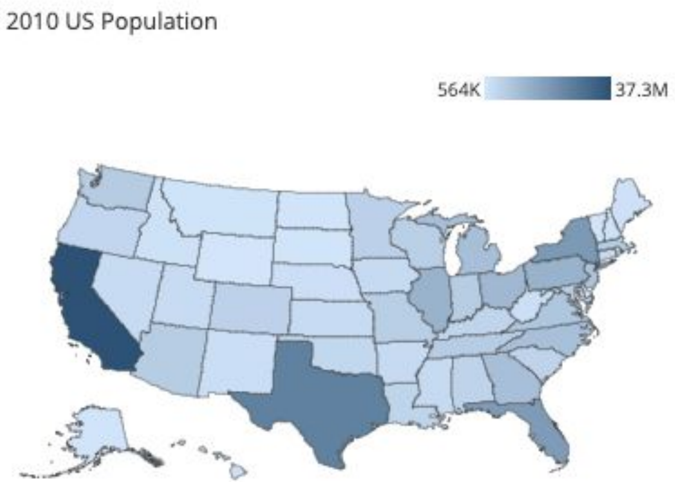
Tailor your visualizations to the problem at hand



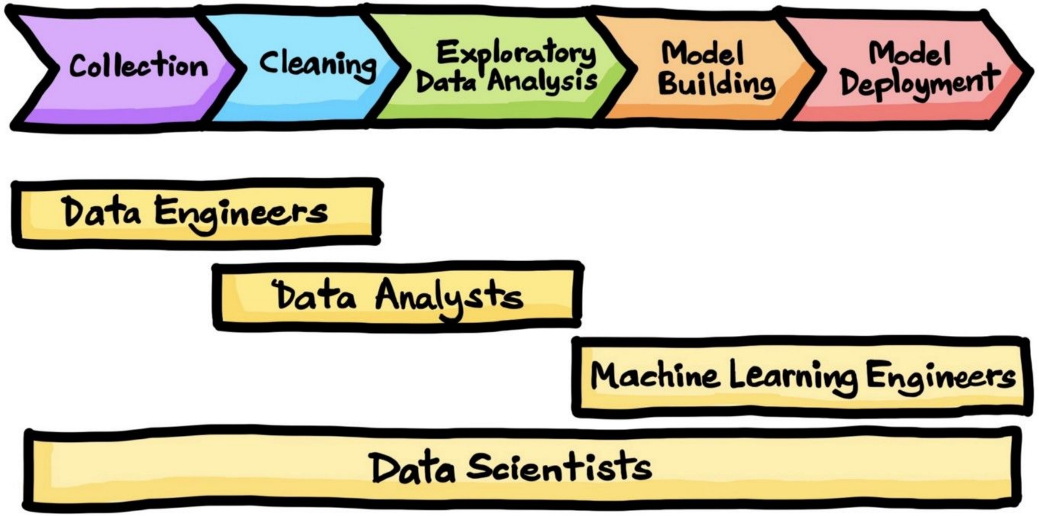


3. Data Exploration & Analysis

Tailor your visualizations to the problem at hand



THE DATA SCIENCE PROCESS



4. Model Building

Is your data model ready?

All data must be numeric!

Transforming non-numeric columns into numeric is called **encoding**.

4. Model Building - Data Encoding

Types of encoding:

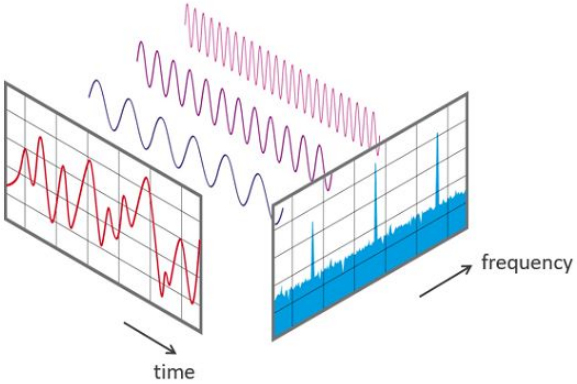
- One-Hot encoding
- Category encoding
- Ordinal encoding
- Frequency encoding
- Binary encoding
- Mean encoding

4. Model Building - Data Encoding for NLP

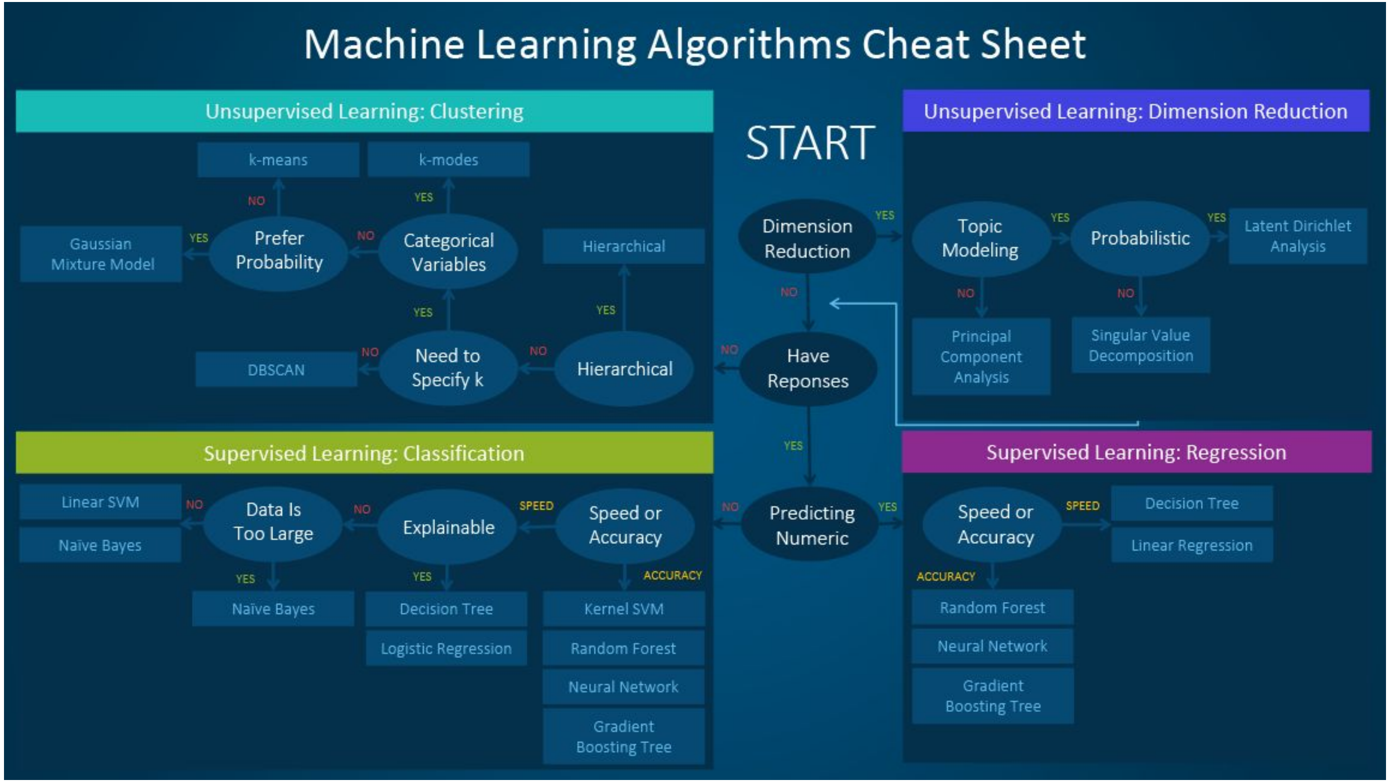
- Have to represent words as numbers
  - Tokenization
  - Removing stopwords
  - Lemming/stemming
  - N-grams
  - NLTK, SpaCy
  - TF-IDF matrix
- Blog on text encoding for real estate price prediction:
  https://medium.com/@data4help.contact/nlp-with-real-estate-advertisements-part-1-55200e0cb33c

4. Model Building - Data Encoding for Sound

- Have to represent sound as numbers
  - Fourier transformation
  - Feature extraction
- Blog on sound encoding for predicting car make from engine sounds:
  https://medium.com/@data4help.contact/signal-processing-engine-sound-detection-a88a8fa48344



4. Model Building - Algorithm Selection



Source: https://www.kdnuggets.com/2017/06/which-machine-learning-algorithm.html

4. Model Building - Algorithm Selection

What is important for your task and model?

- Accuracy?
- Explainability?
- Speed?

4. Model Building - Algorithm Selection

Lazy Estimator

Predict sensible value

Mean/median for regression problems, most common class for classification

Baseline Model

Easy-to-implement ML Model

Linear/logistic regression, random forest

Further Models

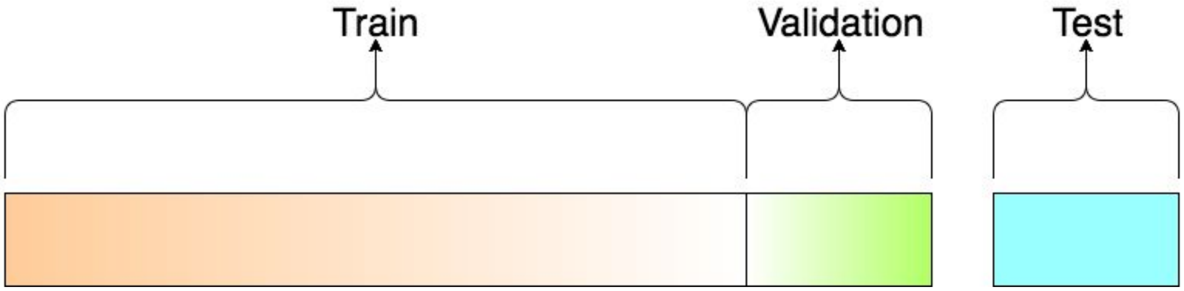
Sophisticated Algos

NNs, further feature engineering, hyperparameter tuning

4. Model Building - Model Evaluation

How do you know if your model is “good”?

- Hold data out to test on!



4. Model Building - Model Evaluation

To test how well your model is doing, have to use the correct metric for the problem!

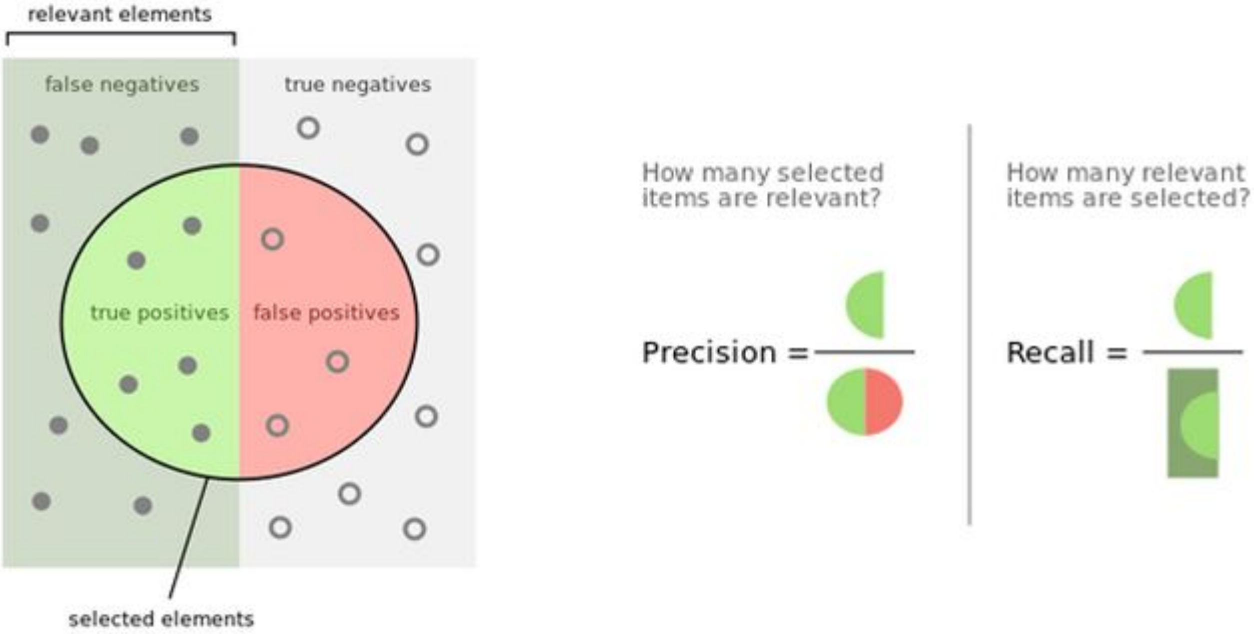
Classification:

- Accuracy
- Precision
- Recall
- Confusion Matrix
- F-Score
- ROC-curve

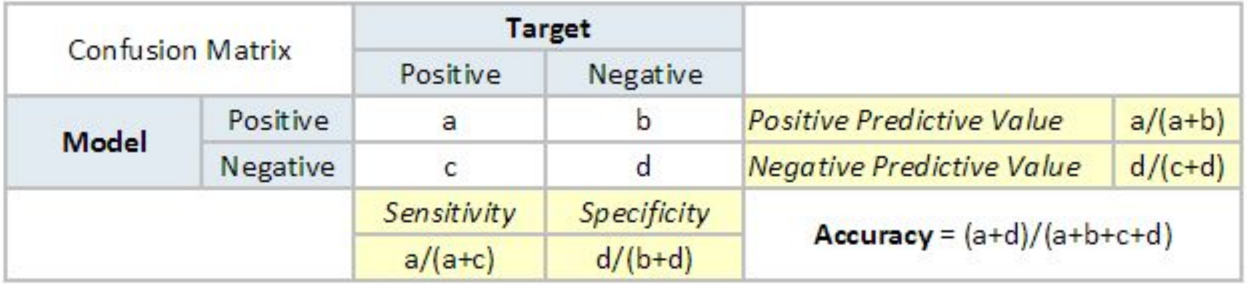
Regression:

- MAE
- MSE
- RMSE
- MAPE
- MASE
- Explained variance (infamous R2)
  - the proportion to which a model accounts for the variation (dispersion) of data
  - scaled
  - 0 = chance, 1 = perfect
  - only compare on the same dataset

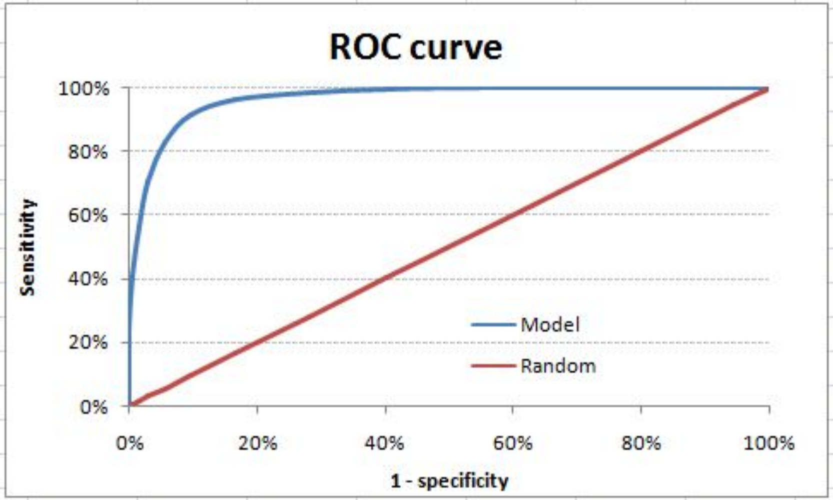
4. Model Building - Model Evaluation



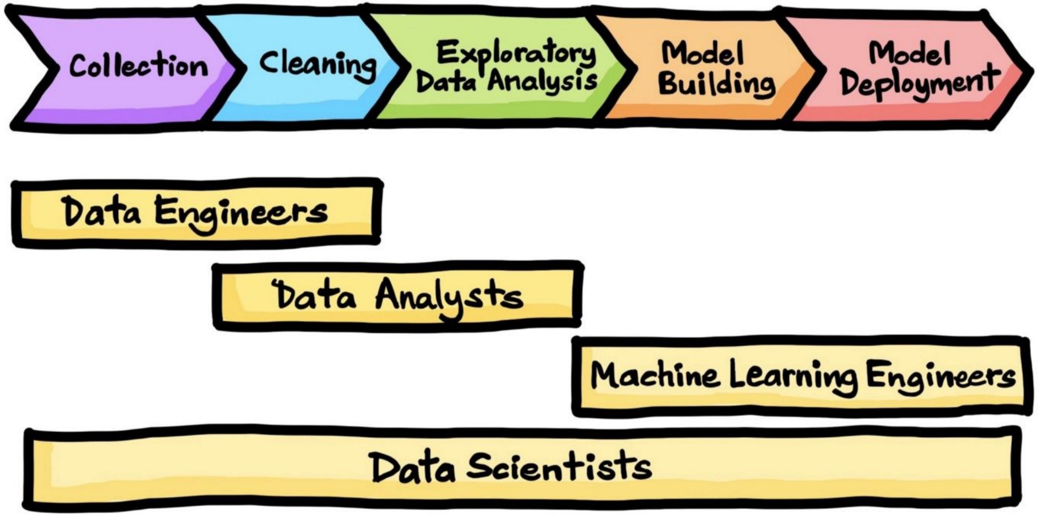
4. Model Building - Model Evaluation



Confusion Matrix		**Target**			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		



THE DATA SCIENCE PROCESS



5. Model Deployment

Deployment is how you make your model, and its predictions, available.

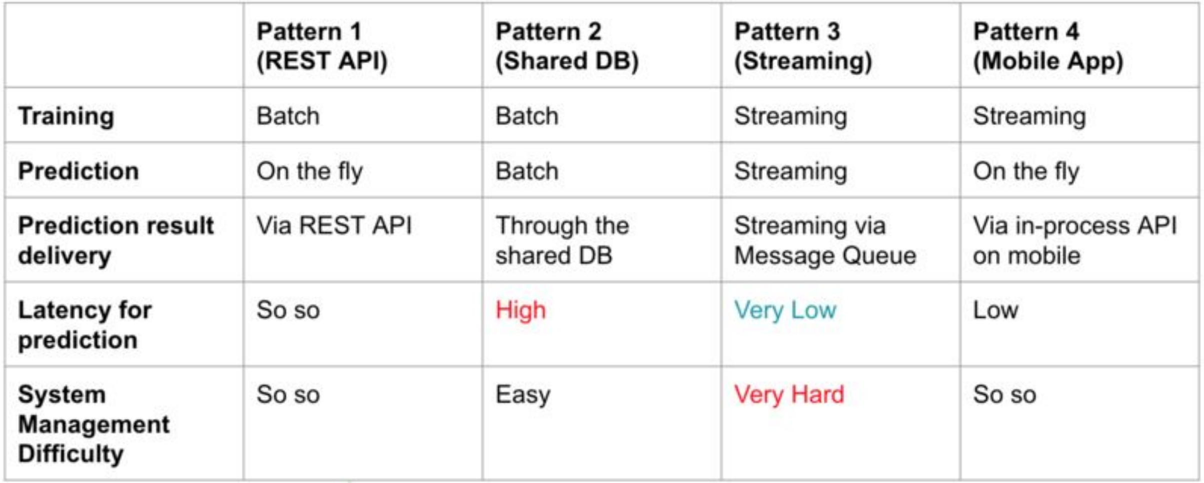
Batch vs. Realtime predictions

Creating an API endpoint

- Flask
- Python Anywhere
- Heroku

Cloud:

- AWS
- Microsoft Azure



	Pattern 1 (REST API)	**Pattern 2 (Shared DB)**	**Pattern 3 (Streaming)**	**Pattern 4 (Mobile App)**
Training	Batch	Batch	Streaming	Streaming
Prediction	On the fly	Batch	Streaming	On the fly
Prediction result delivery	Via REST API	Through the shared DB	Streaming via Message Queue	Via in-process API on mobile
Latency for prediction	So so	High	Very Low	Low
System Management Difficulty	So so	Easy	Very Hard	So so