Understanding your data to use for Machine Learning

What is Data and how do we use it?

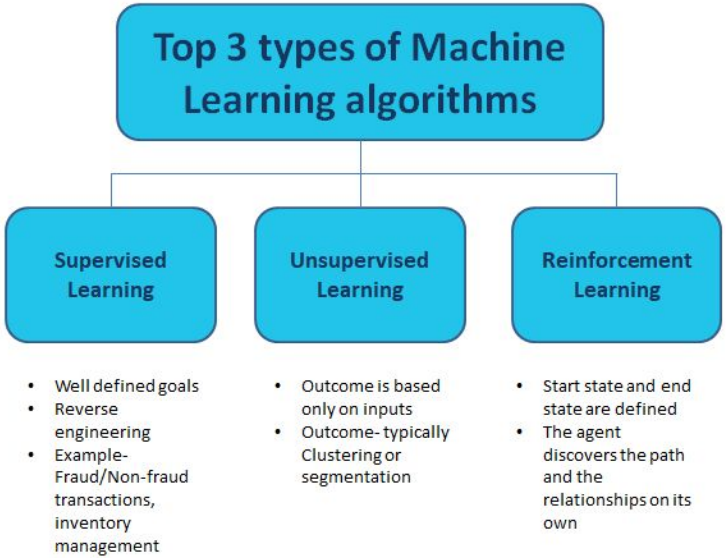
Data is a piece of raw information (facts; causes and consequences) that allows us to extract relevant knowledge for a particular event.

The 5 V's of data

1. *Volume* - The scale/quantity of Data.
2. *Variety* - Different forms of data – images, videos, audio, and so on.
3. *Velocity* - Rate of data streaming and generation. How often do we capture new data?
4. *Value* - Meaningfulness/relevance of data in terms of information that one might require to make a logical inference.
5. *Veracity* - Certainty and correctness in data we are working on.

Types of Machine Learning

1. **Supervised Learning** - Learning with existing information about the outcomes for an event (labeled data).
2. **Unsupervised Learning** - Finding similarities in the causal features and clusters information which may result in similar outcomes (unlabeled data).
3. **Reinforcement Learning** - Experience as you explore. Model uses trial-and-error to explore different techniques, and their outcomes define whether the model learns correctly. Favorable outputs are encouraged or 'reinforced', and non-favorable outputs are discouraged or 'punished'.



What kind of problems can we tackle using Supervised Learning?

Two types of learning objectives using labeled data:

1. ***Classification*** - Expected output is always from a discrete set of classes.

   e.g.- whether an image contains a dog or cat, credit card fraud detection, etc.

2. ***Regression*** - Output is continuous in nature. There can be infinite possible answers.

   e.g.- Price of a stock tomorrow, salary of an employee based on qualifications, etc.

Stages of Data Preprocessing

1. Data Exploration
2. Data Cleansing
3. Data Transformation
4. Feature Engineering

Data Exploration (Assessing your data)

1. What data is available and how much?
2. Do you have access to the ground truth, the values you're trying to predict?
3. What format will the data be in?
4. Which fields are most important?
5. Are the fields available in machine-readable form?
6. What important metrics are reported using this data?

Let's test it out!

Walmart Weekly Sales Dataset (https://www.kaggle.com/yasserh/walmart-dataset)

Data Cleansing

Detecting any incomplete or incorrect values.

- Dealing with missing values
- Dealing with outliers
- Correcting typos
- Grouping sparse classes
- Dropping duplicates

Data Transformation

Data you have available may not be in the right format or may require transformations to make it more useful.

- Categorical encoding
- Dealing with skewed data (Scaling)
- Bias mitigation (Correlation v/s Causation)

Feature Engineering

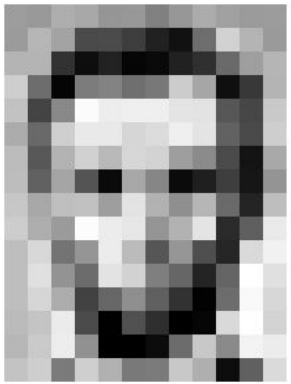
Feature engineering is the process of exploring new features based upon knowledge about current features and the required task.

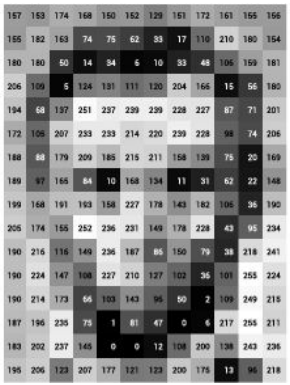
- Feature Extraction
- Capturing Feature Relationships

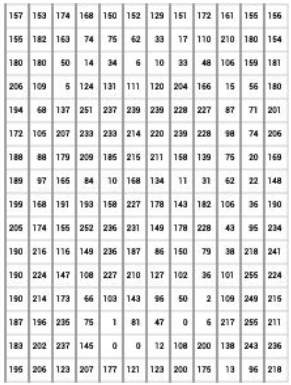
Image Processing & Computer Vision

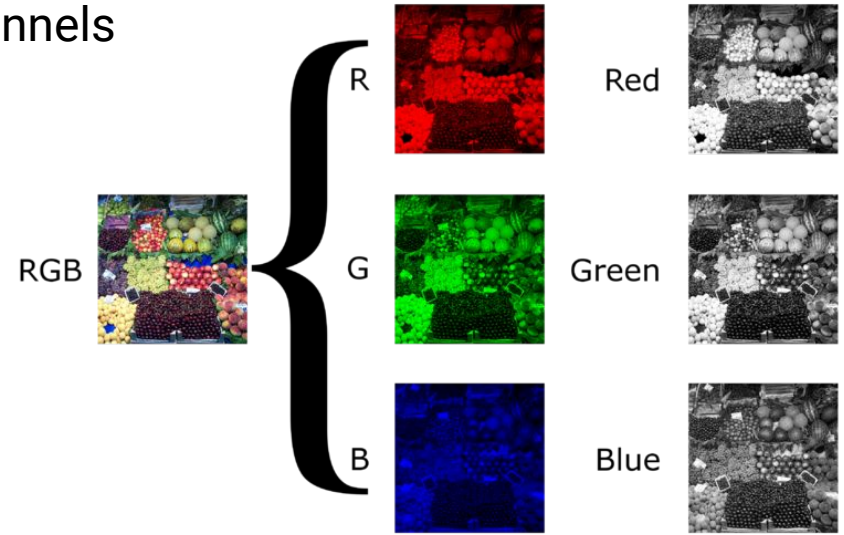
Representing the image data

- Array of pixels (represented as numbers) - colour channel.
- Images can be monochrome or coloured.
- Coloured images have 3 arrays - RGB channels



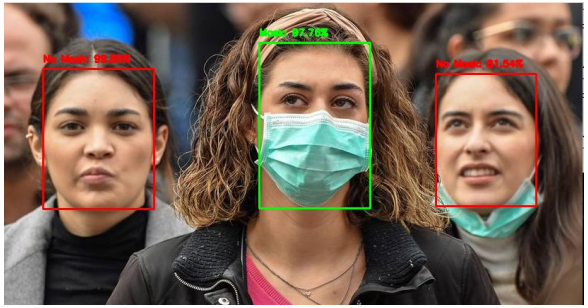


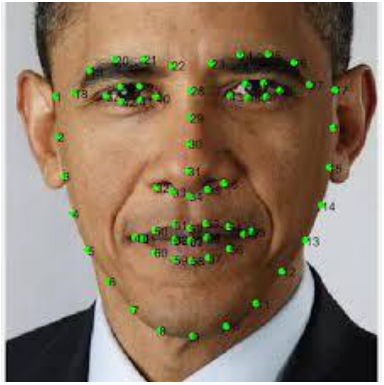


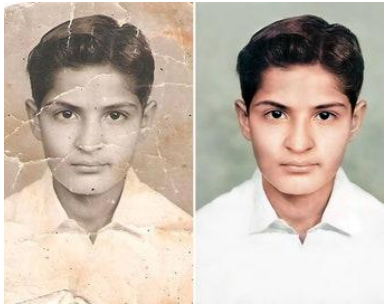


Applications of Computer Vision

- **Object Classification**
- **Object Recognition**
- **Object Verification**
- **Object Detection**
- **Object Landmark Detection**
- **Object Segmentation**
- **Video motion analysis**
- **Scene reconstruction**
- **Image restoration**

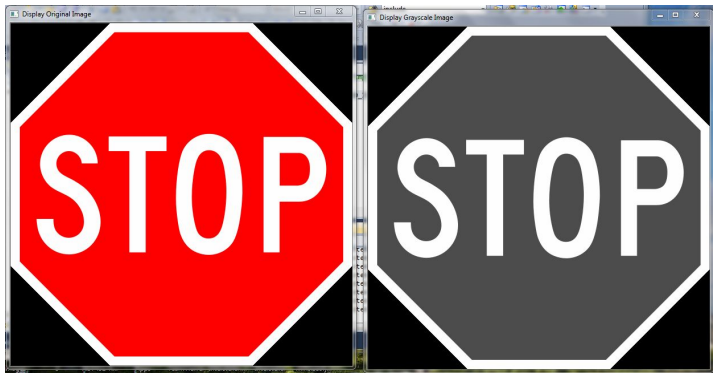


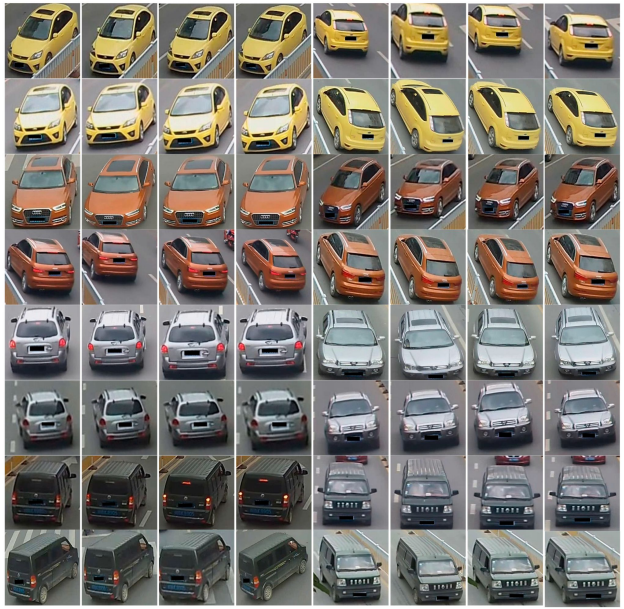




Difficulties while working with image data

- Size of the dataset
- Preprocessing overhead
- Quantity and quality of the data
- Eliminating redundant features





What other kinds of data?

More sources of data

- Simple Tabular Data - ML algorithms, DNN
- Image data - Convolution Networks (CNN)
- Time series data (Forecasting) - Recurrent Networks (RNN) / LSTMs
- Video data - CNN (Images) + RNN (Sequence)
- Textual data - Word embeddings, Vectorization (Word2Vec/Doc2Vec), LSTMs
- Audio data - CNN (Mel Spectrogram), LSTMs
- Reinforcement Learning ???

Thank you