Module 3 - Creating a Dataset
Lesson 1

1. Answering Questions With Data

Video: https://youtu.be/z7A44YnJqCw

Example used: Improve search results

Data is needed to:

- Return relevant results, OR

- Return results which are tailored to individual preferences

Building a dataset:

- Does the dataset fit the problem?

- Is the dataset complete and contains enough information to represent all real-world cases?

- How to annotate a dataset and ensure data quality and user experience over time

Quiz
Which of the following is NOT a requirement for building a dataset?
1. Dataset should be as complete as possible.
2. Dataset must contain enough information to represent all real-world cases.
3. Dataset must be already annotated with labels of interest.

Answer: 3

2. As Good As the Data

Video: https://youtu.be/LDXAmUJN-x4

How a model performs depends heavily on the training data

Considerations:

- Do you have enough data?

Machine learning (ML) algorithms need many different examples of different types of data to be able to distinguish between different classes and find patterns.

Example:

Movie reviews: To distinguish between positive and negative reviews, a lot of review data is required to gather the specific words/phrases for review categorisation.

Few data points will not provide enough information, resulting in inaccuracy and bias in model.

More data gives the ML algorithm more context and information to learn from, hence able to generalise better when faced with new user data.

Also be mindful of garbage in, garbage out (GIGO), which means flawed, or nonsense input data produces nonsense output or 'garbage'.
Source: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out

A deep learning (DL) algorithm will require a larger data size compared to traditional ML algorithms.

DL aka neural networks need to see many data examples before it can distinguish between them and find general patterns.

Sampling bias is introduced when data points are too few or there are uneven data representation in certain categories.

Quiz

What does garbage in, garbage out (GIGO) refer to?

1. Erroneous inputs will lead inevitably to false and misleading outputs.
2. Erroneous inputs will be rejected from being used in the model.
3. Erroneous inputs and outputs will be auto-corrected by the model.

Answer: 1

i. Data Distribution

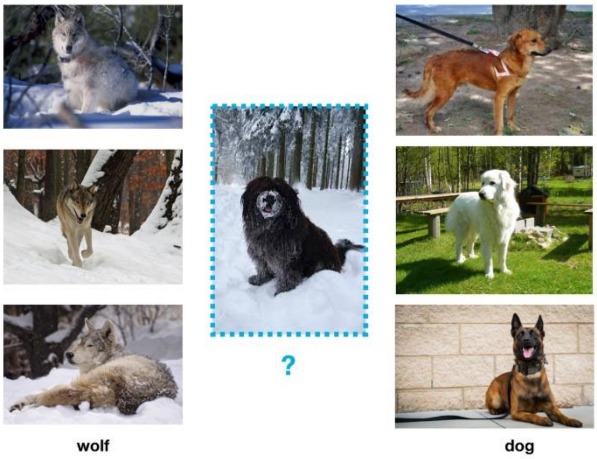
Credit card fraud detection:

Thousands of valid examples and very few examples of fraudulent transaction data, hence we need to take steps to account for this data imbalance, else a model may classify all new data as valid.

ii. Pattern Detection

For a classifier to distinguish wolves from dogs, if images of wolves are all shown with snowy backgrounds, a model may wrongly link snow with wolves.

To create a more accurate model, more varied data is needed.



3. Data Fit

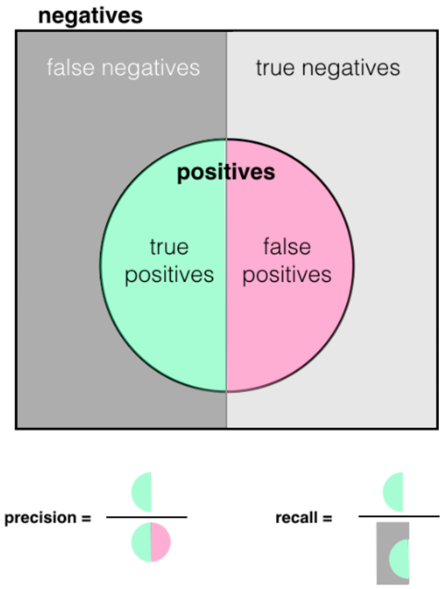
Video link: https://youtu.be/veFVZp4SpdU

i. Use production data to ensure training data matches real-world scenarios.

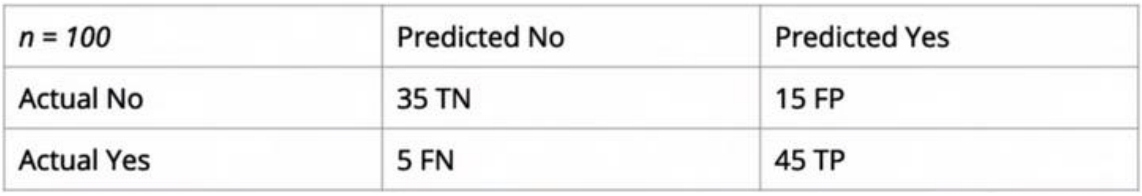
ii. Success criteria for a trained model:

- Precision: No. of True Positives (TP) over True Positives plus False Positives (TP + FP)
  - Precision is higher when FP is low
  - “Of those items selected, how many are relevant”
- Recall: No. of True Positives (TP) over True Positives plus False Negatives (TP + FN)
  - Recall is higher when FN is low.
  - “How many relevant items are selected”
- F1 score
  Formula: F1 score = 2 × (precision × recall) / (precision + recall)

If criteria are not met, retraining is needed.



A confusion matrix is used for calculating precision, recall and the F1 score.



n = 100	Predicted No	Predicted Yes
Actual No	35 TN	15 FP
Actual Yes	5 FN	45 TP

Precision = TP / (TP + FP) = 45 / (45 + 15) = 0.75
Recall = TP / (TP + FN) = 45 / (45 + 5) = 0.90
F1 score = 2 × (precision × recall) / (precision + recall) = 2 x (0.75 x 0.90)/(0.75 + 0.90) = 0.82

Quiz
What is the difference between Precision, and Recall?
1. Precision refers to how many relevant items are selected; Recall refers to among the selected items, how many are relevant.
2. Precision refers to sensitivity of the model; Recall refers to specificity of the model.
3. Precision refers to among the selected items, how many are relevant; Recall refers to how many relevant items are selected.
Answer: 3

4. Data Collection & Relevance Quiz

5. Data Completeness

Video link: https://youtu.be/R-d2U20vFqU

Data collected needs to represent all scenarios that comes up in real world data.
Consider:
i. What is the problem, and how does end users benefit?
ii. What data will help you solve that problem?

- Collect data and observe patterns/relationships in the data
- Identify missing data / potential anomalies

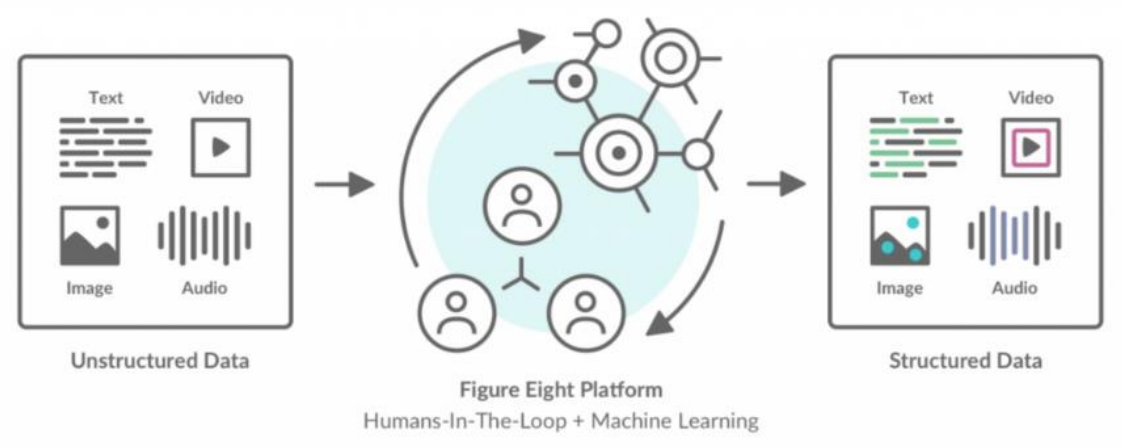
iii. Conduct research and get the best data for your use case.

Quiz
Which of the following is false?
1. To handle missing data, we can remove observations with missing data, or imputing the missing values based on other observations.
2. Outliers in data should always be removed.
3. We can label missing data as missing (categorical variable) or 0 (for numerical variable).

Answer: 2

6. Appen's Data Annotation Platform

More info.: https://www.figure-eight.com/platform/



7. Template Jobs Quiz

8. Case Study: Parking Signs & Figure Eight

Video link: https://youtu.be/aPHshUsA-FU

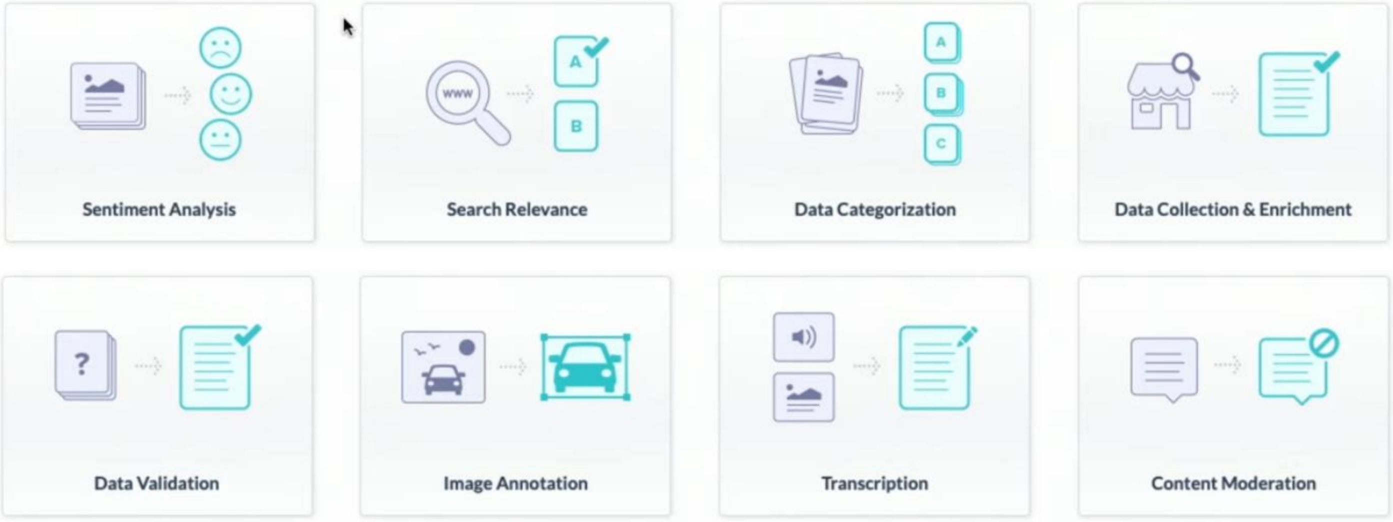
Annotated data is needed to train models. Labelling/annotating data can be in the form of categorising text, outlining objects/images or other labelling requirements.

Figure Eight platform for data annotation:
- Uploading data
- Designing an annotation job
- Creating test questions
- Monitoring results

9. The Platform

Video link: https://youtu.be/SiEhtRmtDr0

Fight Eight platform consists of the template library:



We go to 'Image Annotation', and then select the 'Image Categorisation' template.

Quiz
Which of the following are true?
1. Appen's data annotation platform supports all kinds of raw data (text, video, audio, images).
2. Job templates are useful for building jobs in projects, as most of the steps in the job has been completed.
3. All of the above.

Answer: 3

10. Job Design

Video link: https://youtu.be/jDd7sgJL25M

Steps:
i. Upload source data
ii. Design your job using CML*
iii. Human annotators will choose the appropriate radio button based on the image shown.

*CML (Custom Markup Language) (HTML based language) specific to Figure Eight is used to create the HTML template that will define how human annotators interact with images from the dataset.
It is easiest to use a template CML and then modify it for a specific use case.

Quiz
Which of the following about CML is false?
1. CML supports Python, R and Julia.
2. CML is derived from HTML.
3. CML was customised by Figure Eight to be used for creating HTML templates.

Answer: 1