Google AI

How to find data set and fairness practices

Sayak Paul

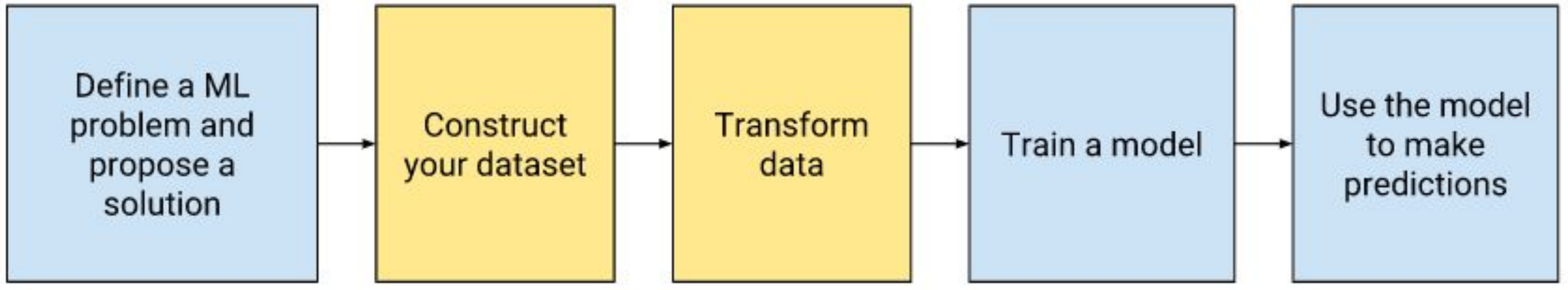
Deep Learning Associate at PyImageSearch

Explore ML Academy, Hyderabad (September 14, 2019)

Agenda

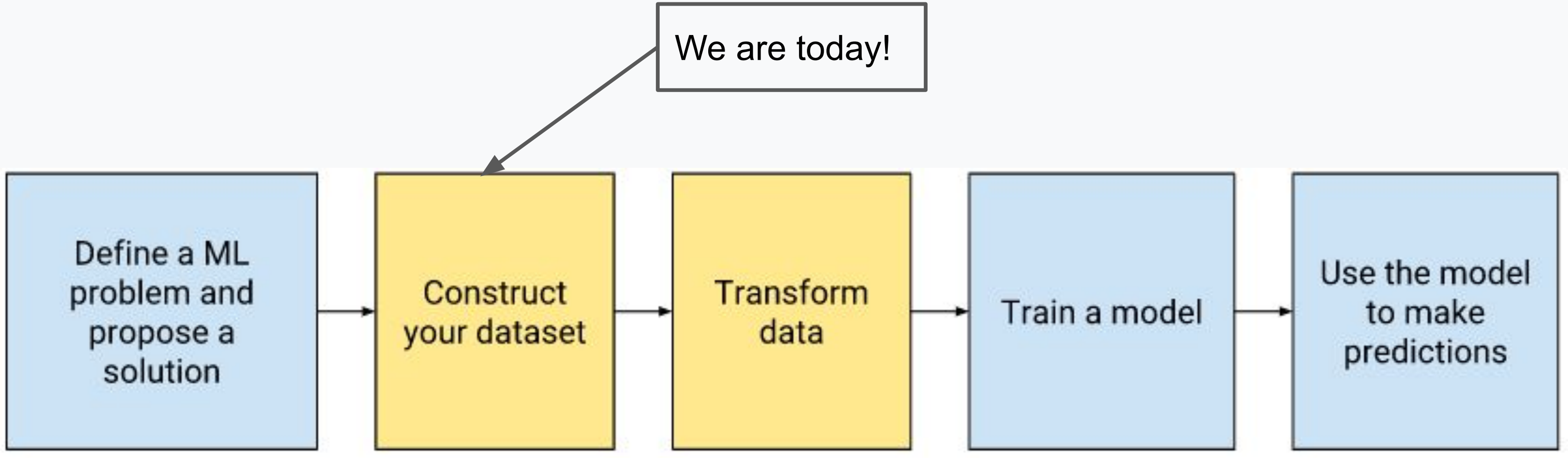
- Why is collecting a good dataset important?
- The process of data preparation
  - Collect raw data
  - Identify features and label sources
  - Select sampling strategy
  - Data splitting
- Incorporating fairness practices
- Guiding lights

The *typical* machine learning workflow


Define a ML problem and propose a solution
Construct your dataset
Transform data
Train a model
Use the model to make predictions

Source

The *typical* machine learning workflow



We are today!

Define a ML problem and propose a solution

Construct your dataset

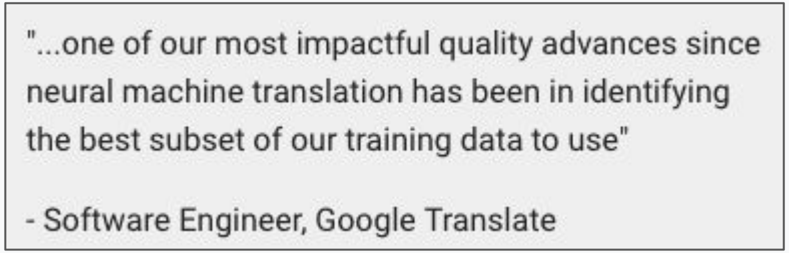
Transform data

Train a model

Use the model to make predictions

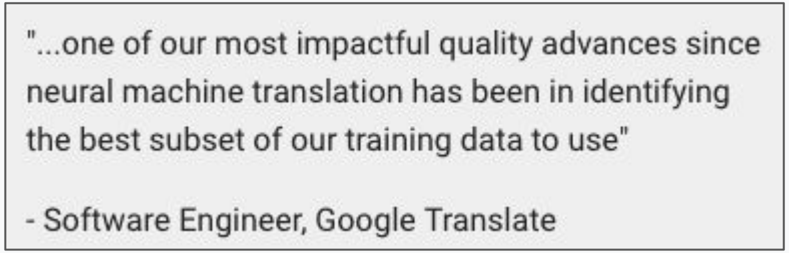
Source

Why is collecting a good dataset important?



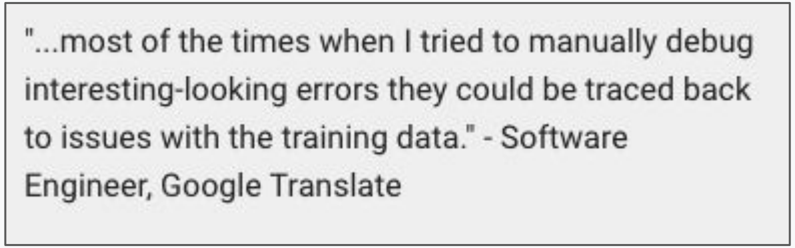
"...one of our most impactful quality advances since neural machine translation has been in identifying the best subset of our training data to use"

- Software Engineer, Google Translate

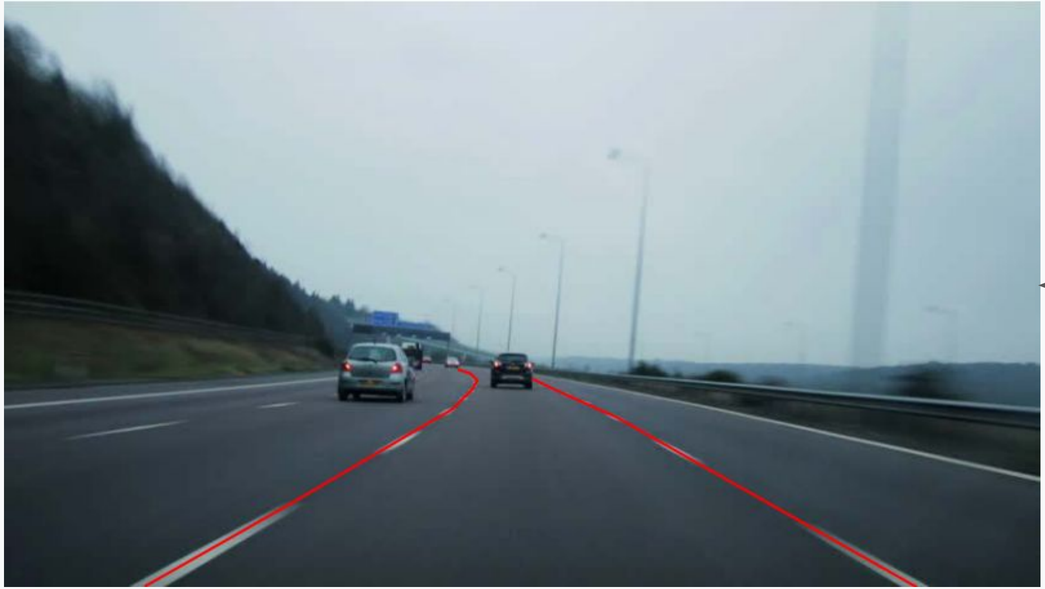


"...one of our most impactful quality advances since neural machine translation has been in identifying the best subset of our training data to use"

- Software Engineer, Google Translate



"...most of the times when I tried to manually debug interesting-looking errors they could be traced back to issues with the training data." - Software Engineer, Google Translate



This is a sample from the training data

Lane detection in self-driving cars (Source)

Source

A sample during the time of inference!

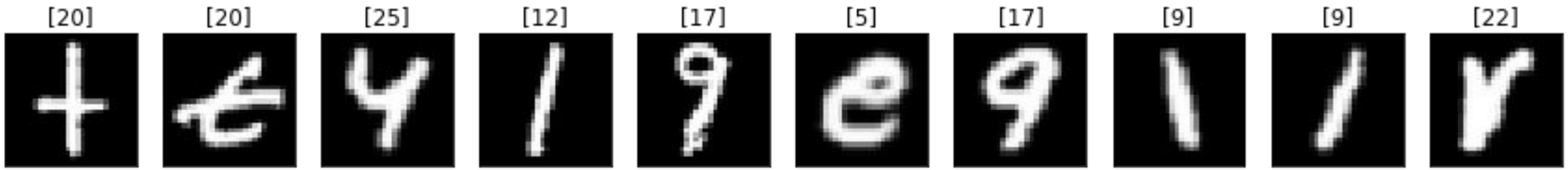
The process of data preparation

Step 1: Collect raw data

- Start with the problem statement
- Identify input elements of the problem
- Collect data that closely represents those elements

Step 1: Collect raw data

- Start with the problem statement
- Identify input elements of the problem
- Collect data that closely represents those elements



Source

The size of a dataset

- Your model should train on at least an order of magnitude more data points than trainable parameters
- Simple models on large data sets generally beat fancy models on small data sets

The quality of a dataset

The quality of a dataset

“Garbage in, garbage out”

The quality of a dataset

- Reliability

The quality of a dataset

- Reliability
- Feature representation

The quality of a dataset

- Reliability
- Feature representation
- Data leakage

Step 2: Identify features and label sources

Step 2.1: Identify features

- Make a hypothesis about a feature(s), test it and then repeat

Step 2.1: Identify features

- Make a hypothesis about a feature(s), test it and then repeat
- Identify the features such that data leakage is not introduced

Step 2.1: Identify features

- Make a hypothesis about a feature(s), test it and then repeat
- Identify the features such that data leakage is not introduced
- Sometimes more features might lead to Curse of Dimensionality

Step 2.1: Identify features

- Make a hypothesis about a feature(s), test it and then repeat
- Identify the features such that data leakage is not introduced
- Sometimes more features might lead to Curse of Dimensionality
- Prediction data sources — online vs. offline

Step 2.2: Label sources

- Direct label vs. derived labels

Step 2.2: Label sources

- Direct label vs. derived labels
- Human labelling

Step 3: Select sampling strategy

Two situations:

- Need to collect more data :(
- Plenty of data, hence need to use a subset ^_^

Step 3: Select sampling strategy

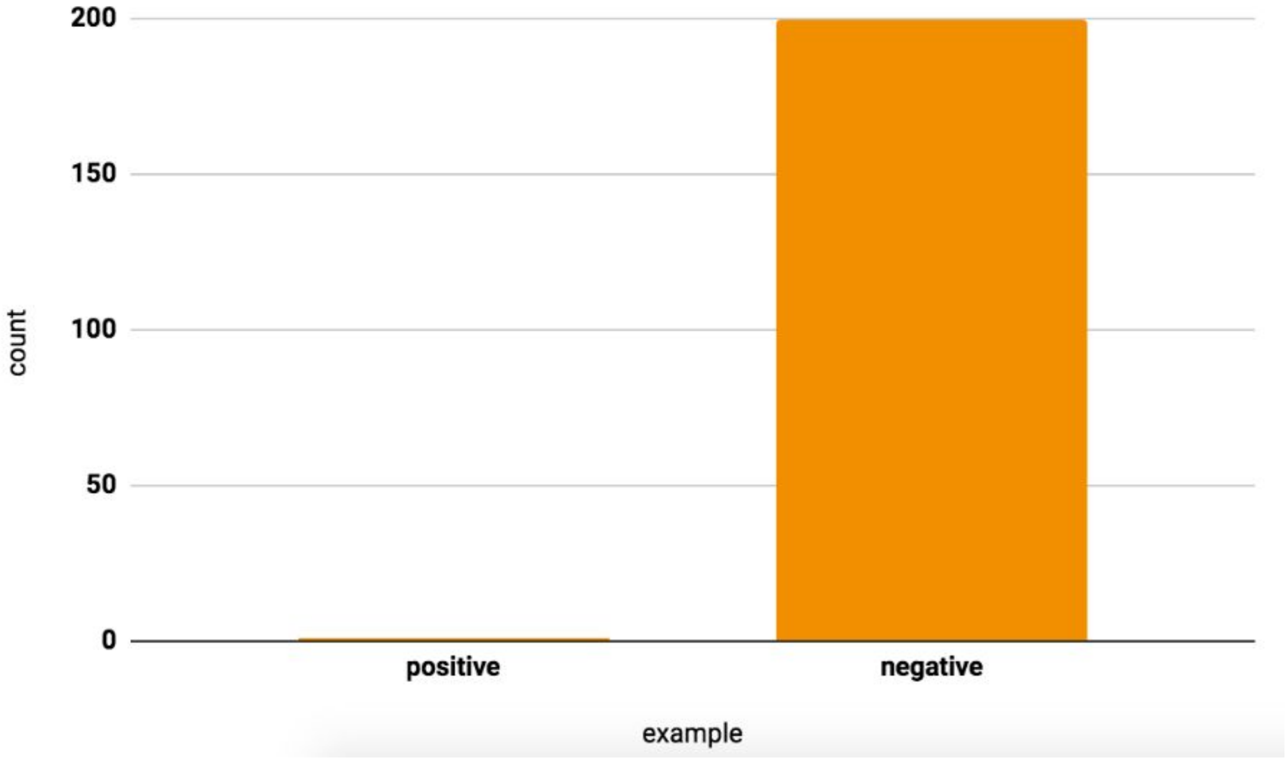
Two situations:

- Need to collect more data :(
- Plenty of data, hence need to use a subset ^_^

How do you do that? :O

The problem of Class Imbalance

Google AI



Source

Step 4: Data splitting

Three golden splits:

- Training
- Validation
- Testing

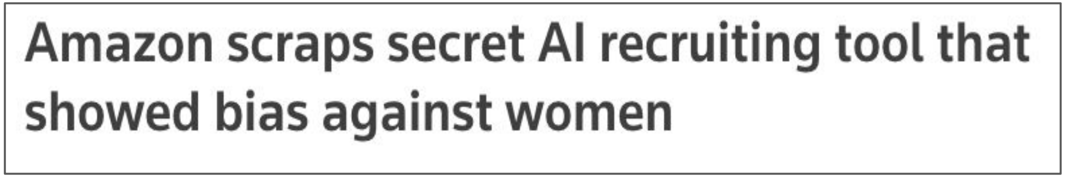
Step 4: Data splitting

Two important questions:

- When can we not split the data randomly?
- When can we do that?

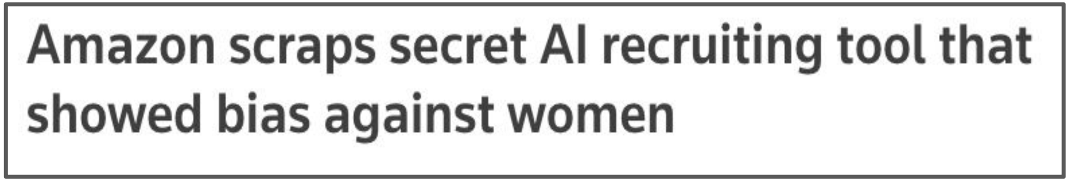
Incorporating fairness practices

Why care about it?

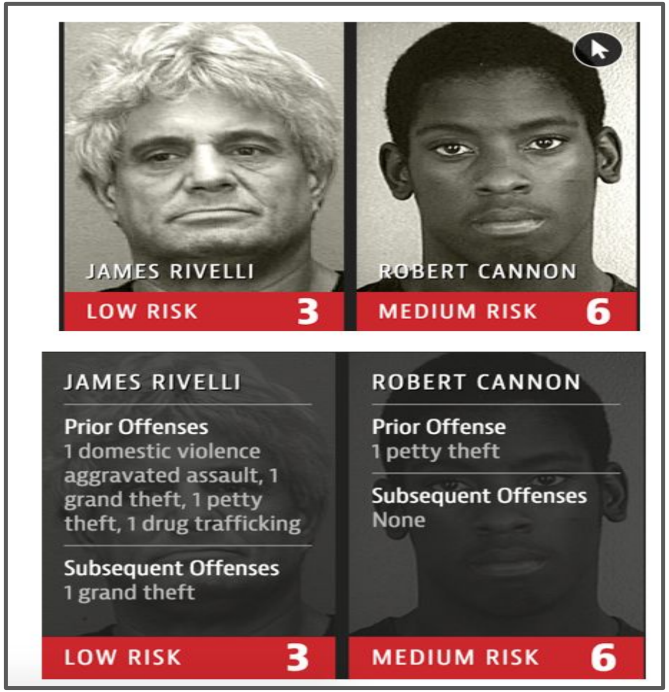


Amazon scraps secret AI recruiting tool that showed bias against women

Why care about it?



Amazon scraps secret AI recruiting tool that showed bias against women



JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6

JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

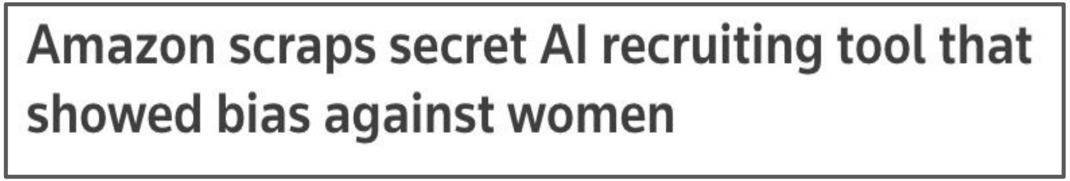
ROBERT CANNON

Prior Offense
1 petty theft

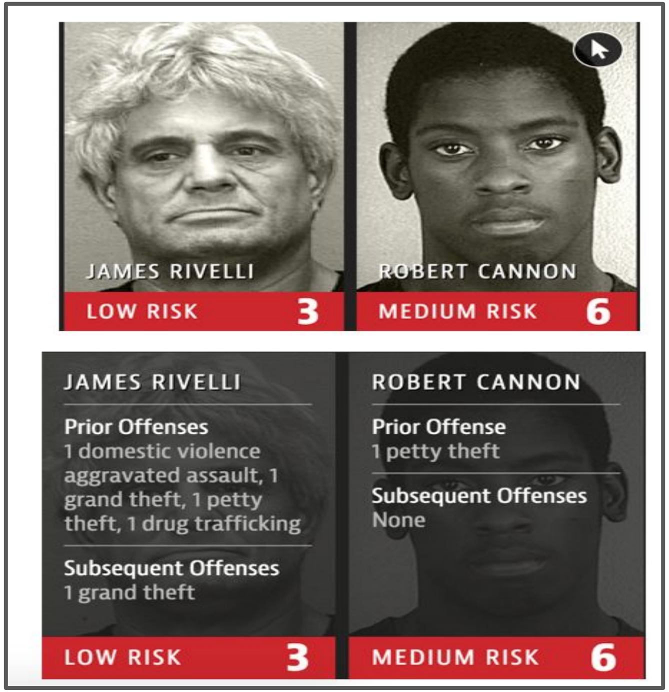
Subsequent Offenses
None

MEDIUM RISK 6

Why care about it?



Amazon scraps secret AI recruiting tool that showed bias against women



JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6

JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

And many more ...

Combating bias

- Identifying bias

Combating bias

- Identifying bias
  - Missing feature values in large numbers
  - Unusual feature values
  - Class imbalance

Combating bias

- Evaluating for bias

Combating bias

- Evaluating for bias
  - Don’t fall prey to the accuracy paradox
  - Determine the confusion matrix of the model
  - Investigate the false predictions made by the model

Wrapping up

- The process of data preparation
    - Collect raw data
    - Identify features and label sources
    - Select sampling strategy
    - Data splitting
- Incorporating fairness practices

References

- Data Preparation and Feature Engineering in ML
- How (and why) to create a good validation set
- Becoming One With the Data
- Data Science from Scratch
- Machine Learning Fairness

See you next time

Google AI

Find me here:
sayak.dev

Thank you very much :)

Experts

pyimagesearch