Becoming one with the data

Sayak Paul (@RisingSayak)

$whoami

- I call `model.fit()` @ PyImageSearch
- Netflix Nerd 👀
- My coordinates are here - https://sayak.dev/

Ideal audience

- ML enthusiasts and practitioners looking to understand data better.

What are we up to today?

- Why become one with the data?
- Data transformation
- Exploratory data analysis (EDA)
- Human baselines

This talk is basically a reflection of

- A Recipe for Training Neural Networks (by Andrej Karpathy)
- The Al-Dente Neural Network: Part I (by Sairam)
- Becoming One With the Data (by me)

Fundamental premise

The first step to training a neural net is to not touch any neural net code at all and instead begin by thoroughly inspecting your data.

- Andrej Karpathy (Source)

Fundamental co-premise

- 💵 What’s the business value of the project?

Fundamental co-premise

- 💵 What’s the business value of the project?
  - Need of more targeted questions.

Fundamental co-premise

- 💵 What’s the business value of the project?
  - Need of more targeted questions.
    - Significance of the data features w.r.t the problem statement.

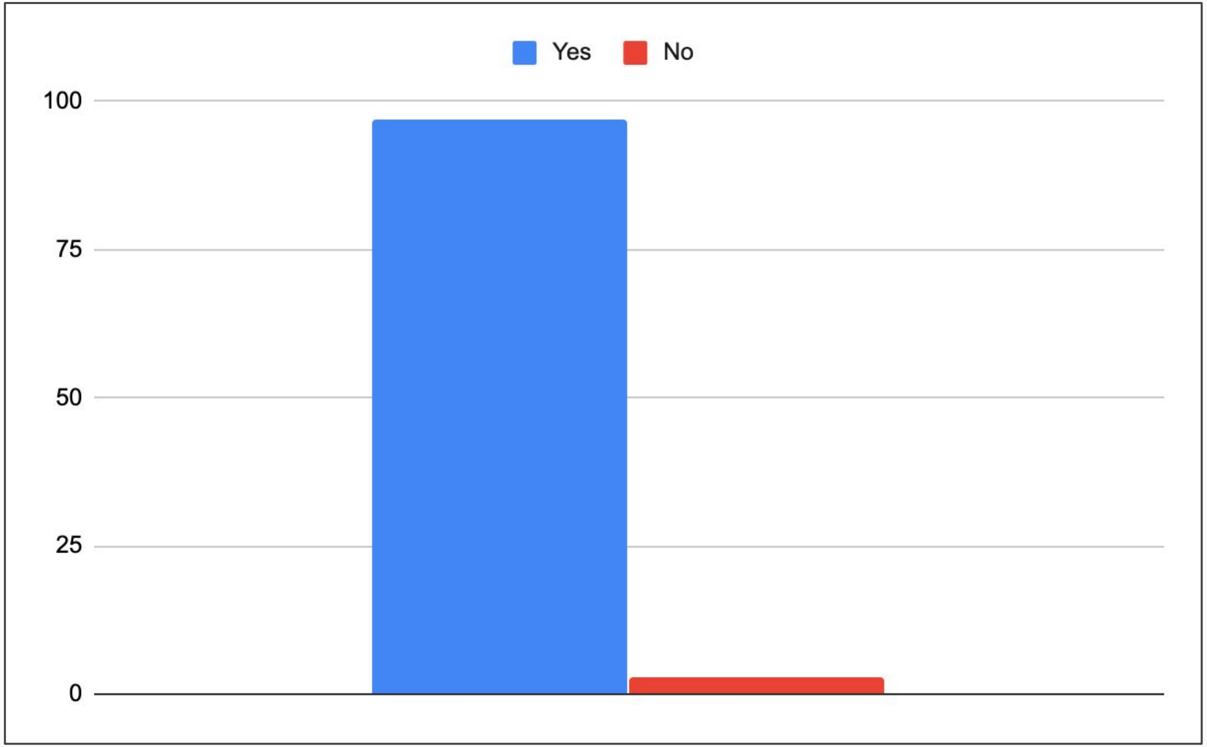
Fundamental co-premise

WHY RACE IS NEEDED TO
PREDICT CREDIT CARD APPROVAL?
imgflip.com

Fundamental co-premise

- 💵 What’s the business value of the project?
  - Need of more targeted questions.
    - Significance of the data features w.r.t the problem statement.
    - Under-representation/Over-representation?

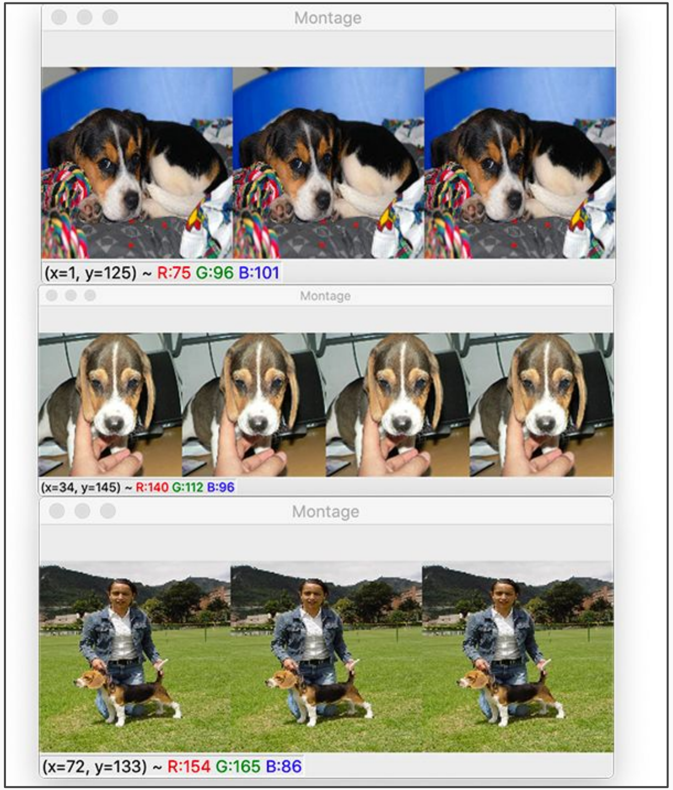
Fundamental co-premise



Fundamental co-premise

- 💵 What's the business value of the project?
    - Need of more targeted questions.
        - Significance of the data features w.r.t the problem statement.
        - Under-representation/Over-representation?
        - Data duplicacy?

Fundamental co-premise



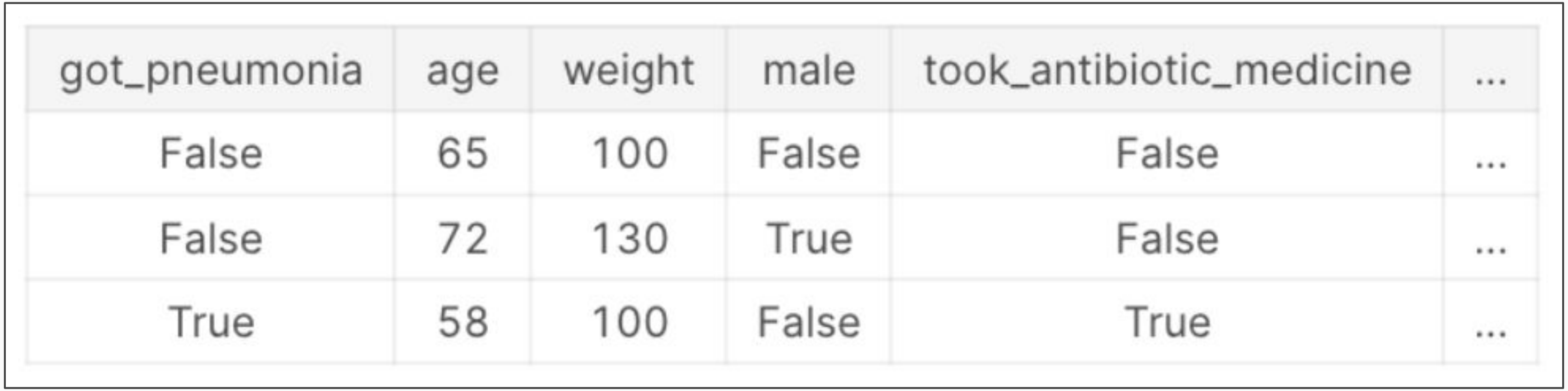
```
(x=1, y=125) ~ R:75 G:96 B:101
(x=34, y=145) ~ R:140 G:112 B:96
(x=72, y=133) ~ R:154 G:165 B:86
```

Source

Fundamental co-premise

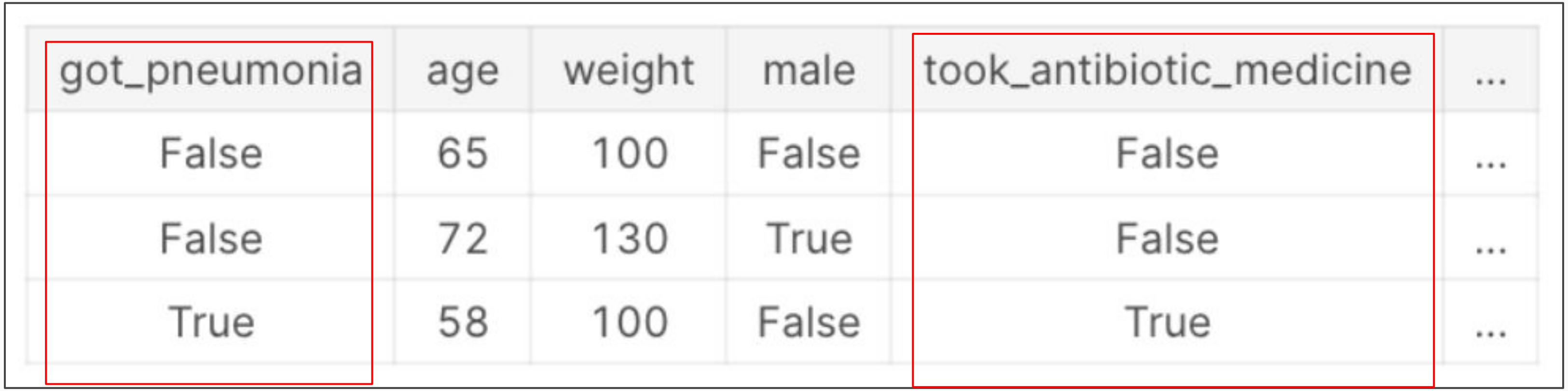
- 💵 What’s the business value of the project?
  - Need of more targeted questions.
    - Significance of the data features w.r.t the problem statement.
    - Under-representation/Over-representation?
    - Data duplicacy?
    - Leakage?

Fundamental co-premise



got_pneumonia	age	weight	male	took_antibiotic_medicine	...
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

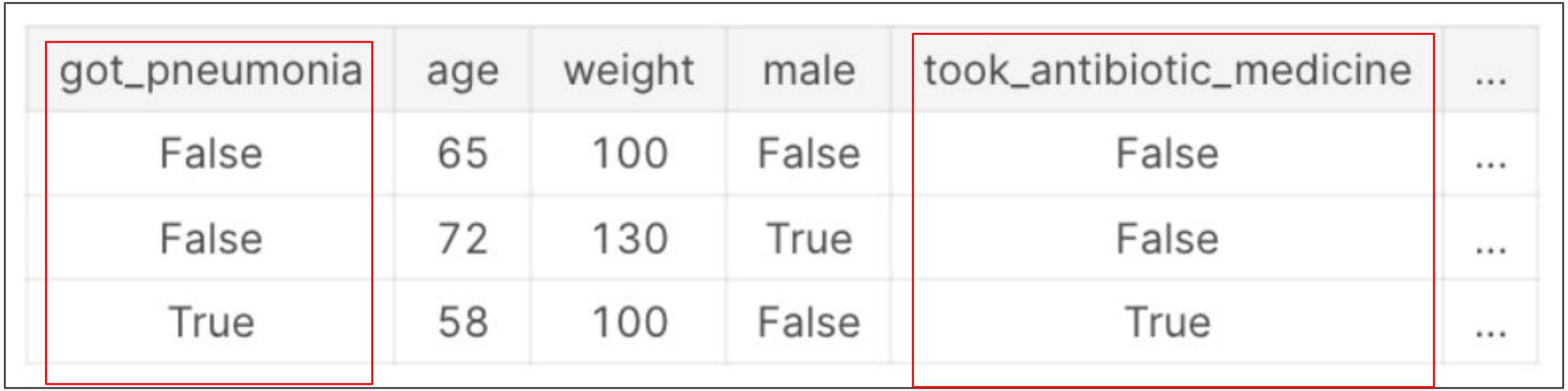
Fundamental co-premise



got_pneumonia	age	weight	male	took_antibiotic_medicine	...
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

- **Strong relationship between** `got_pneumonia` **and** `took_antibiotic_medicine`.

Fundamental co-premise



got_pneumonia	age	weight	male	took_antibiotic_medicine	...
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

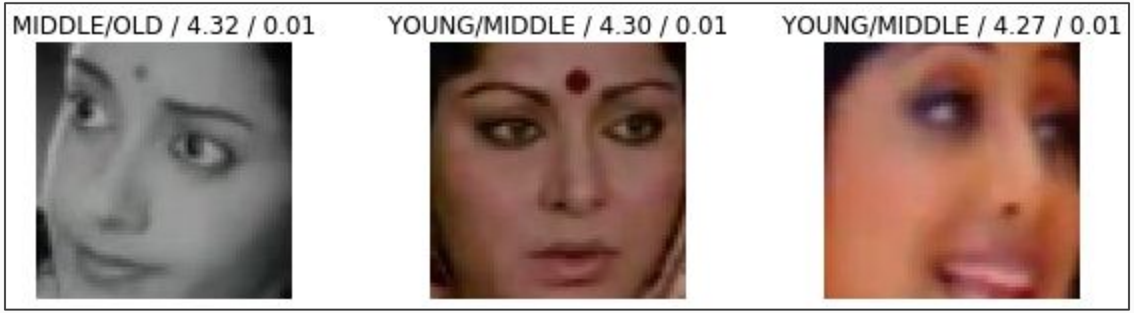
- **Strong relationship between** `got_pneumonia` **and** `took_antibiotic_medicine`.
- **So,** `took_antibiotic_medicine=False` **means no** `got_pneumonia`?

Fundamental co-premise

- 💵 What's the business value of the project?
  - Need of more targeted questions.
    - Significance of the data features w.r.t the problem statement.
    - Under-representation/Over-representation?
    - Data duplicacy?
    - Leakage?
    - Confusing data-points, outliers?

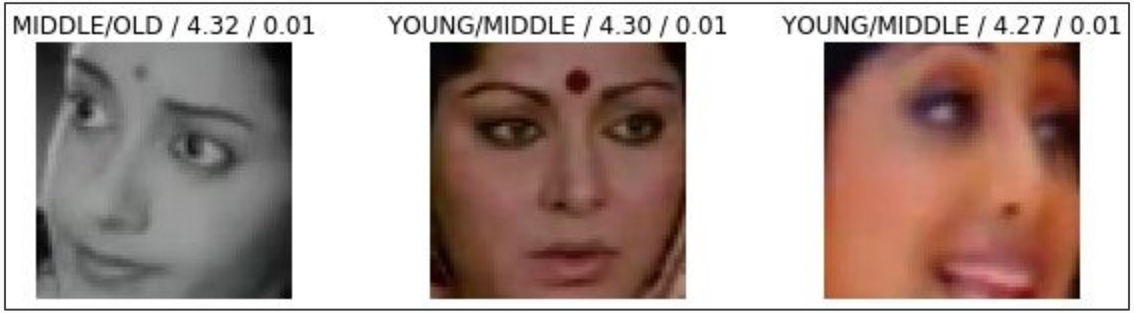
Fundamental co-premise

Prediction/Actual/Loss/Probability



Noisy labels (labeled based on colors 🤥)

Fundamental co-premise



[Are these good representations of airplanes?](#) 😜

Fundamental co-premise

- 💵 What's the business value of the project?
    - Need of more targeted questions.
        - Significance of the data features w.r.t the problem statement.
        - Under-representation/Over-representation?
        - Data duplicacy?
        - Leakage?
        - Confusing data-points, outliers?
        - Bias? (can creep in innumerable ways)

Data investigation should be done in various phases

Data transformation

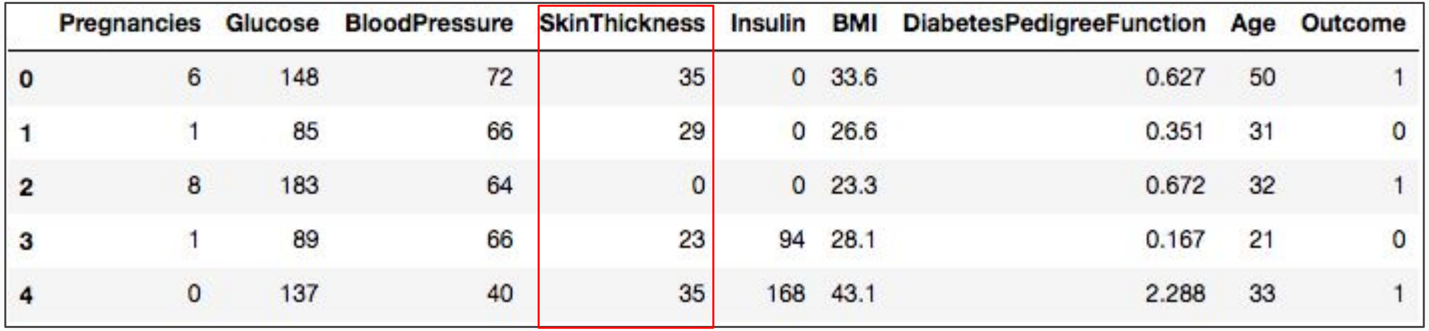
- Missing values

Data transformation

- Missing values
    - What if it is not instantly catchable?

Data transformation

- Missing values
    - What if it is not instantly catchable?



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Skin thickness zero? 😳 (Pima Indians' Diabetes dataset)

Data transformation

- Missing values
  - What if it is not instantly catchable?
  - Understand why data might have been missing and then impute if necessary!

Data transformation

- Missing values
- Typing

Data transformation

- Missing values
- Typing
    - Were the features recorded in correct data types?

Data transformation

- Missing values
- Typing
  - Were the features recorded in correct data types?
  - Significantly impacts the data loading time.

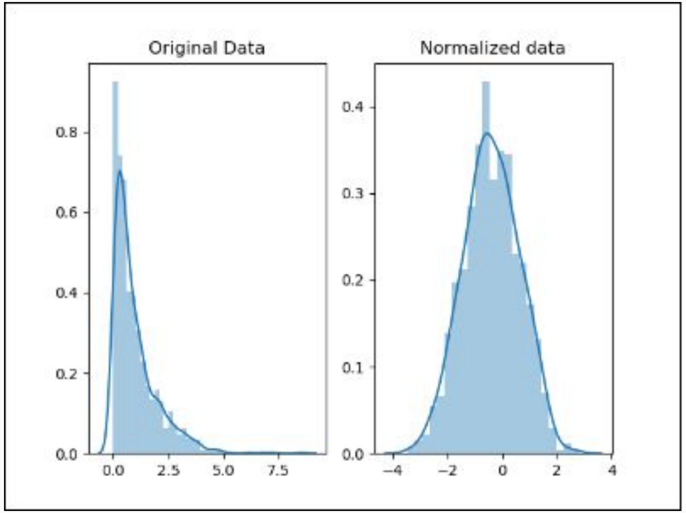
Data transformation

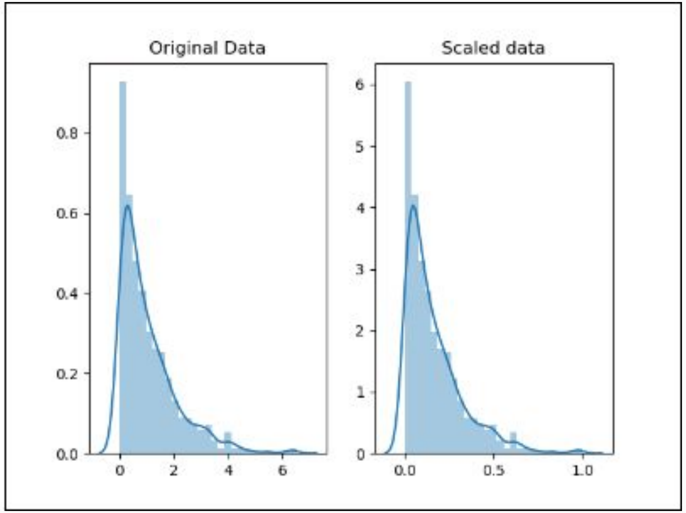
- Missing values
- Typing
  - Were the features recorded in correct data types?
  - Significantly impacts the data loading time.

    What if `int8` features were recorded in `float64` ❗

Data transformation

- Missing values
- Typing
- **Scaling and normalization**





Source

Data transformation

- Missing values
- Typing
- Scaling and normalization
    - What about categorical features?

Data transformation

- Missing values
- Typing
- Scaling and normalization
    - What about categorical features?
    - Normalization stats from the training set only ❗

Data transformation

Not gonna cover in the interest of time:

- Representation of categorical variables
- Handling inconsistent data entries
- Fighting data leakage
- Fighting data imbalance

Data transformation

Not gonna cover in the interest of time:

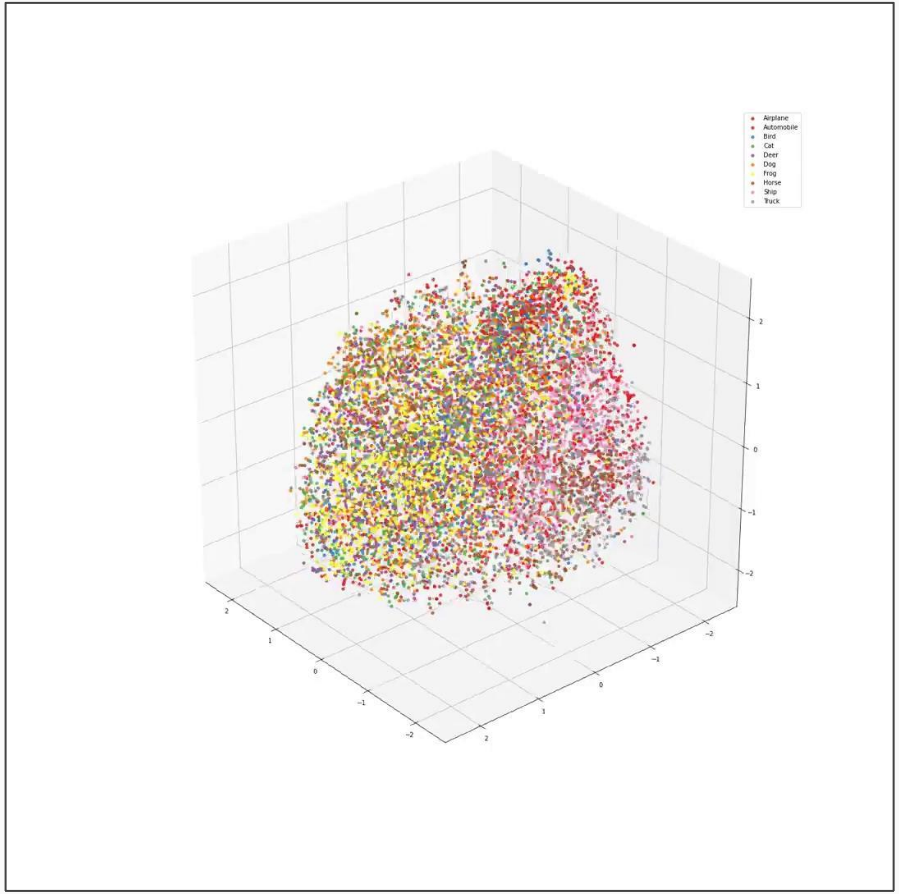
- Representation of categorical variables
- Handling inconsistent data entries
- Fighting data leakage
- Fighting data imbalance

Would encourage checking out the blog post Becoming one with the data.

EDA helping with

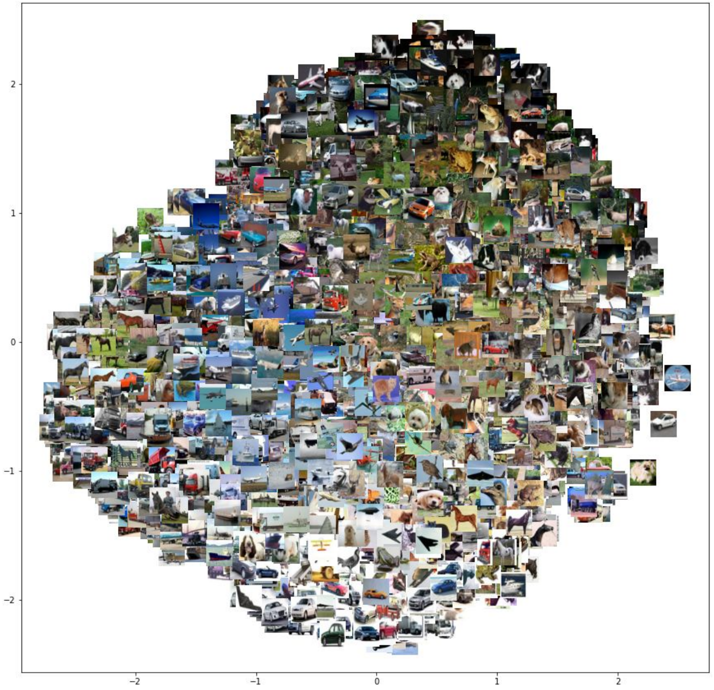
- Discovering interesting patterns from the data.
- Understanding how well the data represents the problem at hand.
- Identification of outliers that may be present in the data.

CIFAR-10 classes are not well separated!



Comes from here

Negative effects of image backgrounds



Comes from here

Negative effects of image backgrounds

- Horses in the middle left but see them mixed with cars, trucks ❗
- Same for cars, birds, dogs, frogs ❗

Thanks to Sairam for these amazing discoveries! For more, check out his report.

A few good stuff to consider

- Human baselines

A few good stuff to consider

- Human baselines
  - How would ***you*** classify a set of images?

A few good stuff to consider

- Human baselines
  - How would ***you*** classify a set of images?
  - Would it be consistent with that of a model?

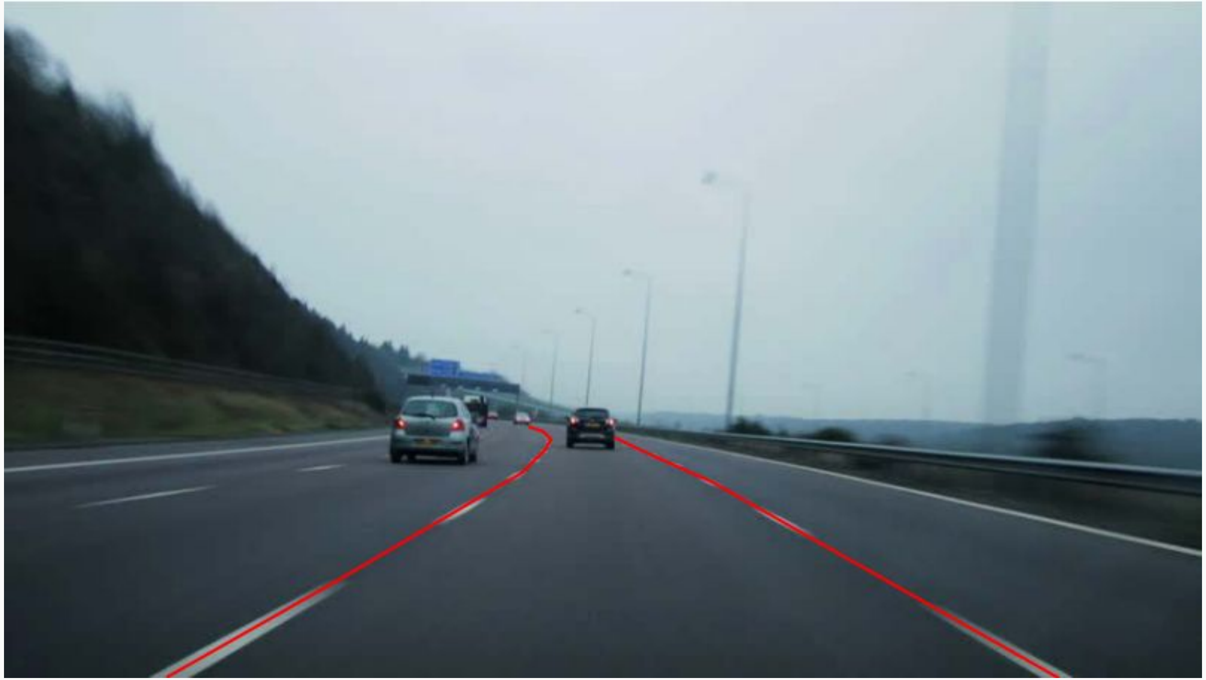
A few good stuff to consider

- Human baselines
- Focus on the data collection

A few good stuff to consider

- Human baselines
- Focus on the data collection
    - Handling cases like ...

A few good stuff to consider



Lane detection for self-driving cars

A few good stuff to consider

Lane detection for self-driving cars

Comes from here

A few good stuff to consider

- Human baselines
- Focus on the data collection
  - Focus on the ***long tail*** of the distribution.

A few good stuff to consider

- Human baselines
- Focus on the data collection
  - Focus on the ***long tail*** of the distribution.
  - Incorporate *active learning* if possible.

Deck available here: https://bit.ly/one-data

Let's get connected on Twitter! I am @RisingSayak.