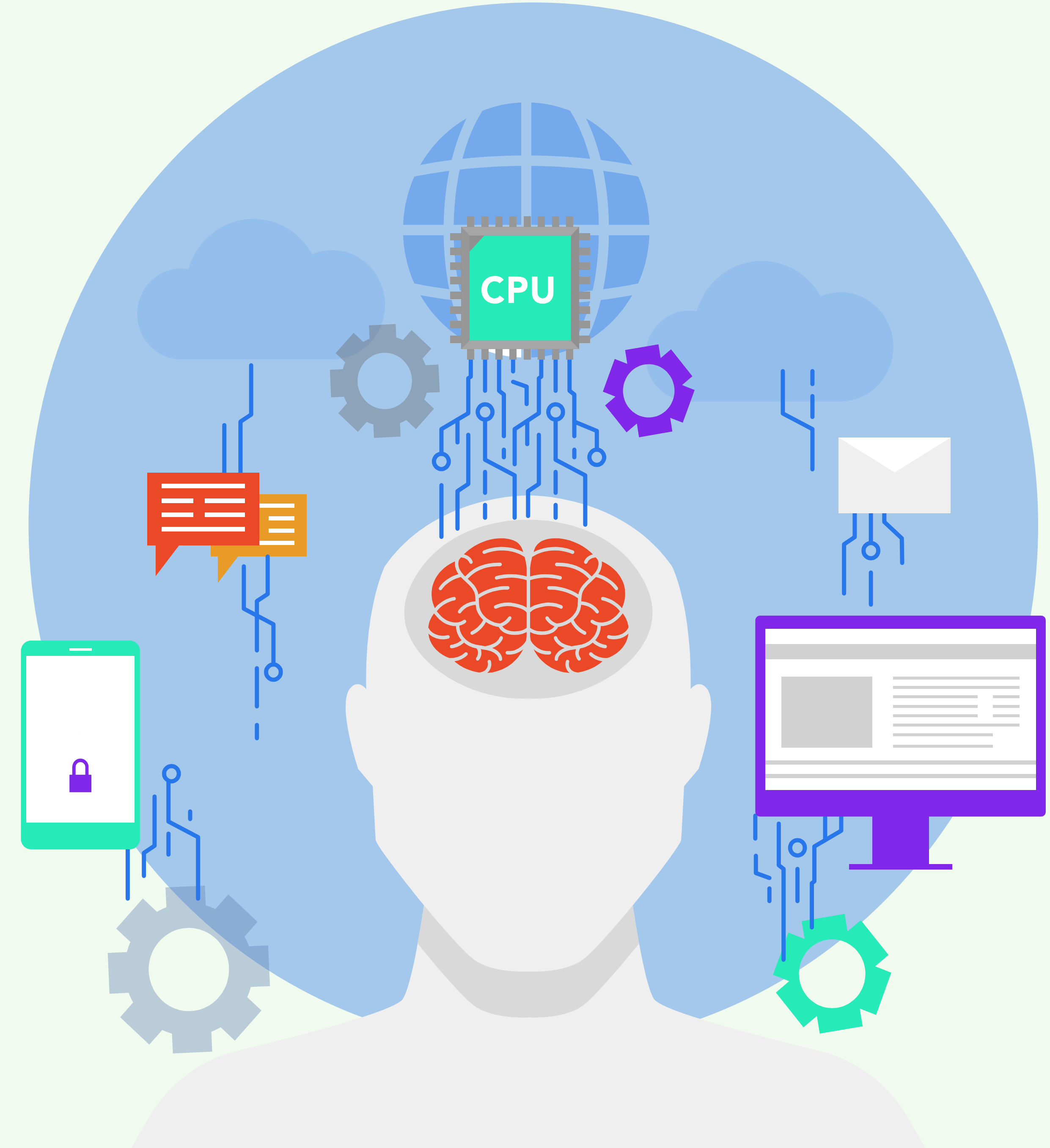
{ON: The Beach}

Bringing DevOps, Devs and Data Scientists together around Big Data



Unearth The Black-Box:

Building Fair, Accountable & Trustworthy ML Systems

AGENDA

01 **INTRODUCTION**

WHAT IS ML AND RELATED CONCEPTS

02 **UNDERSTANDING BLACK-BOX PROBLEM**

WHY BLACK-BOX MODELS CAN LEAD TO UNFAIR, BIASED OUTCOMES

03 **STRATEGIES FOR BUILDING FAIR AND UNBIASED MODELS**

HOW TO BUILD STRATEGIES FOR ENSURING FAIRNESS AND NO-BIAS

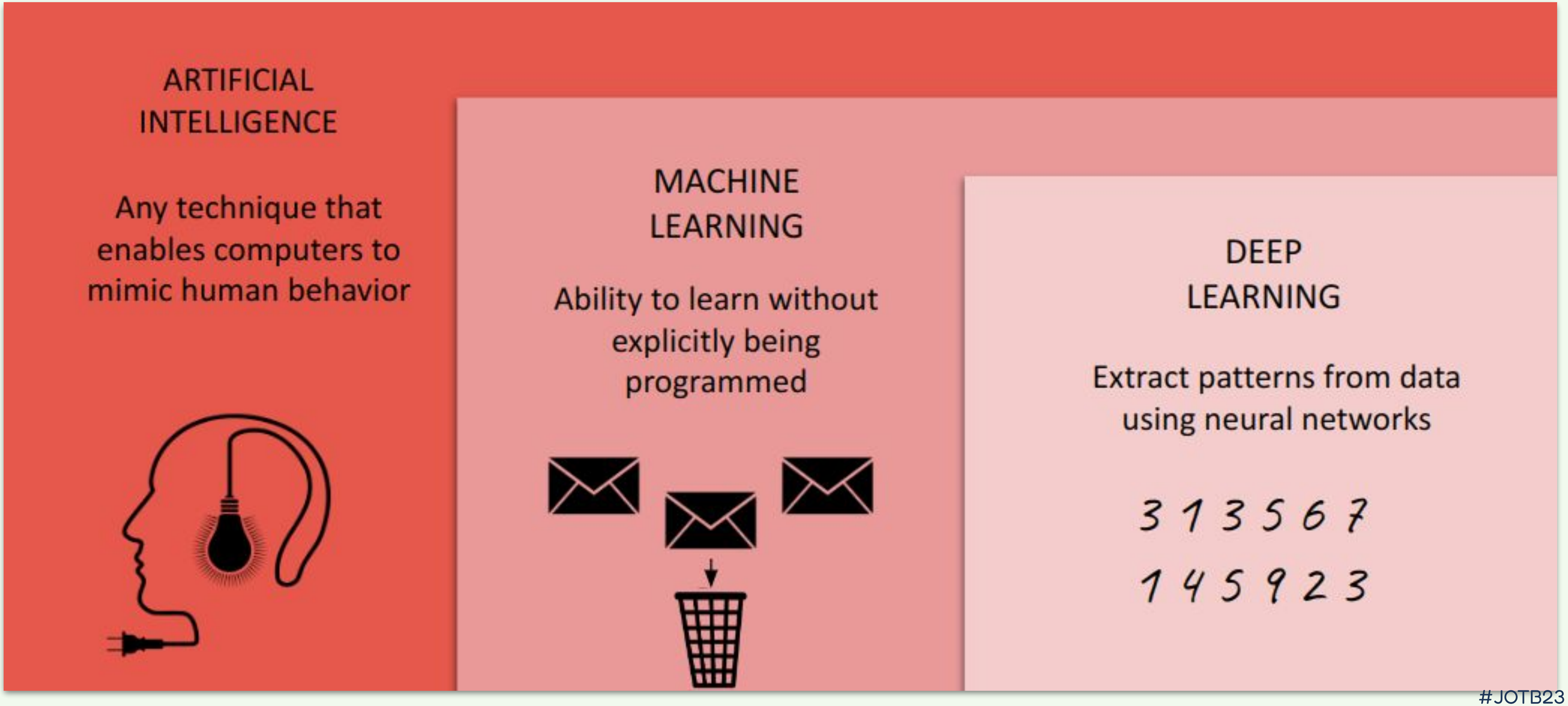
04 **BUILDING TRUSTWORTHY & ACCOUNTABLE MACHINE LEARNING MODELS**

WHAT STRATEGIES FOR BUILDING TRUST & ACCOUNTABILITY IN ML SYSTEMS

05 **CONCLUSION AND RESOURCES**

ENCOURAGE FURTHER RESEARCH AND DISCUSSION

INTRODUCTION



ARTIFICIAL INTELLIGENCE

Any technique that enables computers to mimic human behavior

MACHINE LEARNING

Ability to learn without explicitly being programmed

DEEP LEARNING

Extract patterns from data using neural networks

A

B

C

BASIC NEURAL NETWORK

{ON: The Beach}

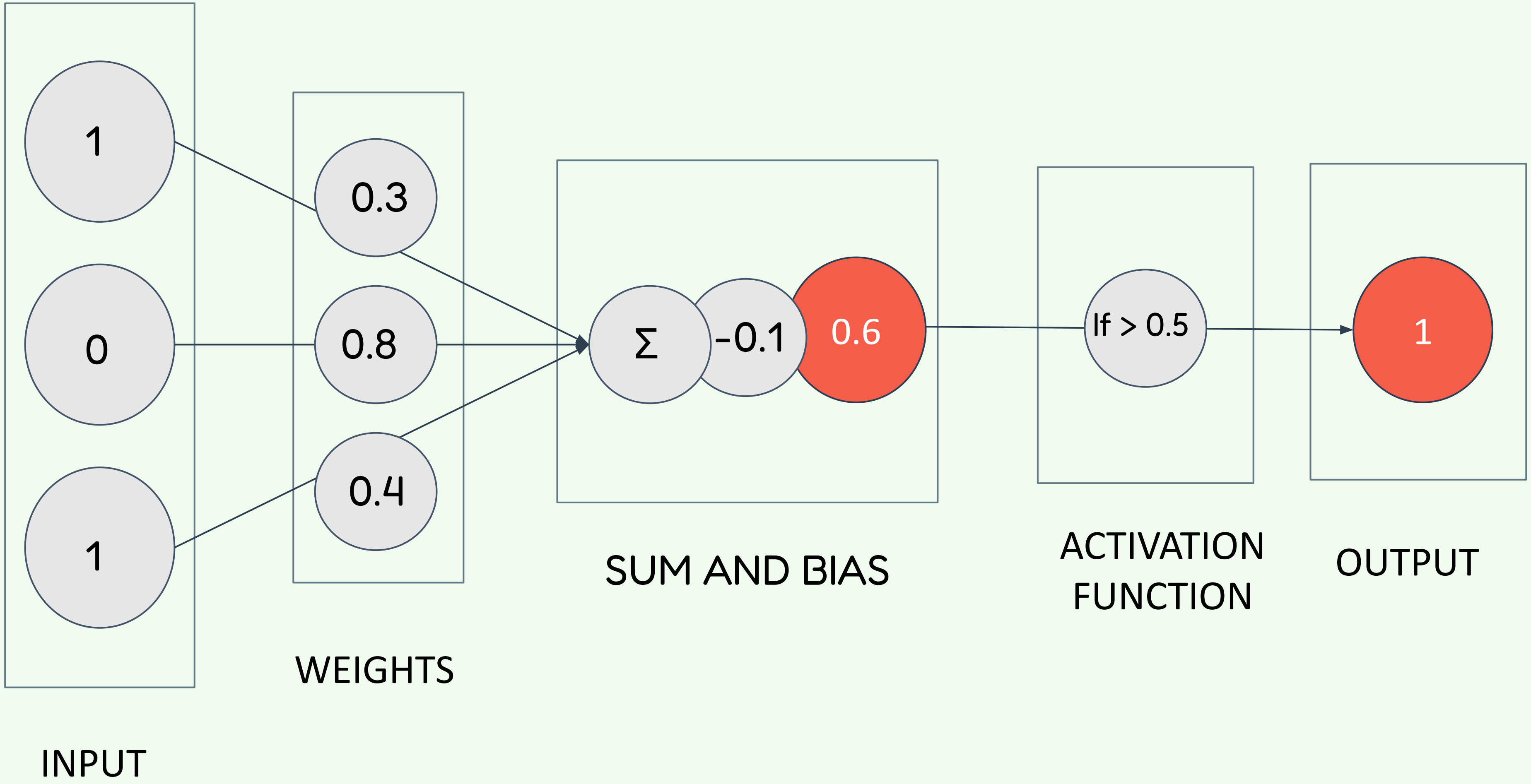
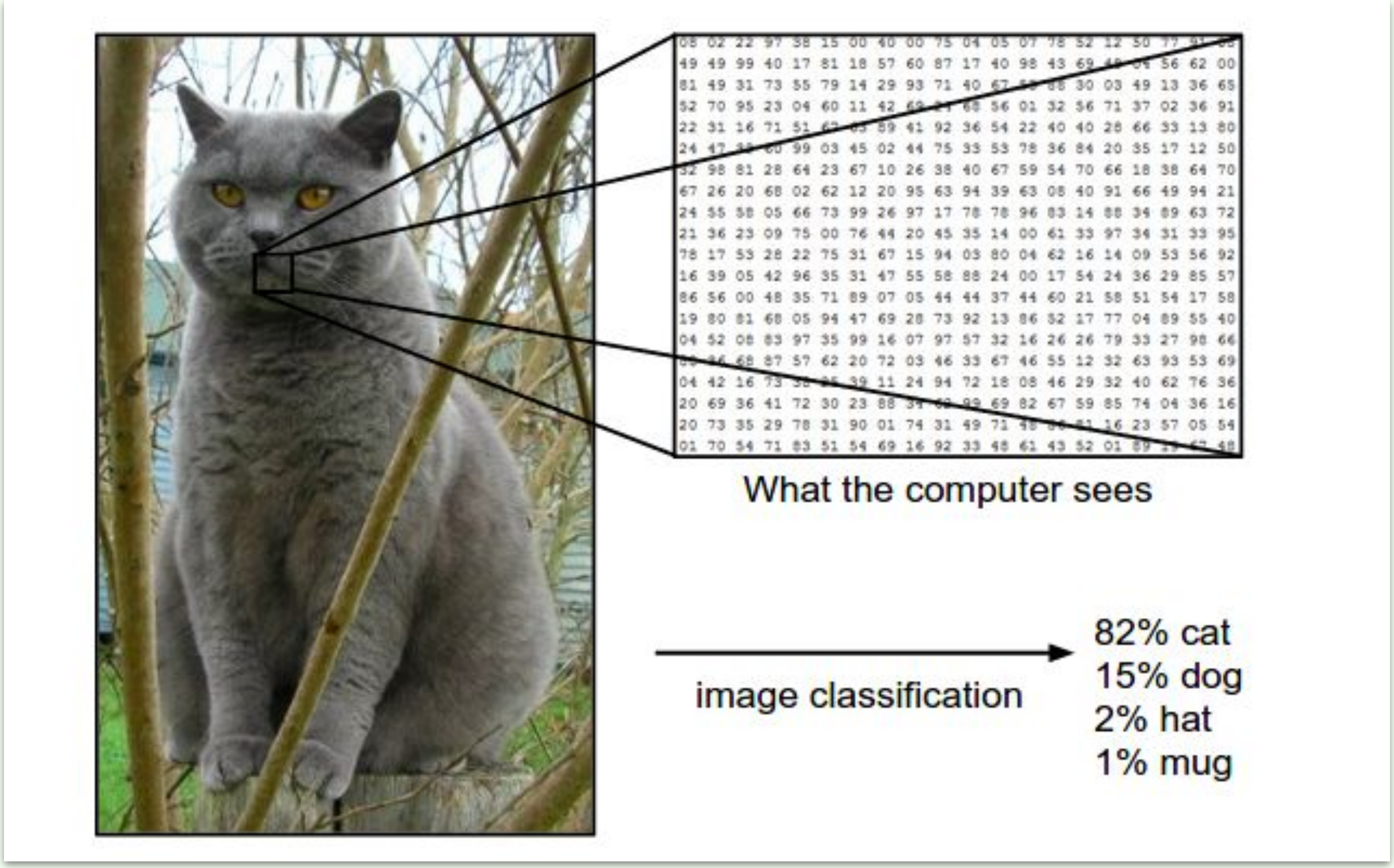


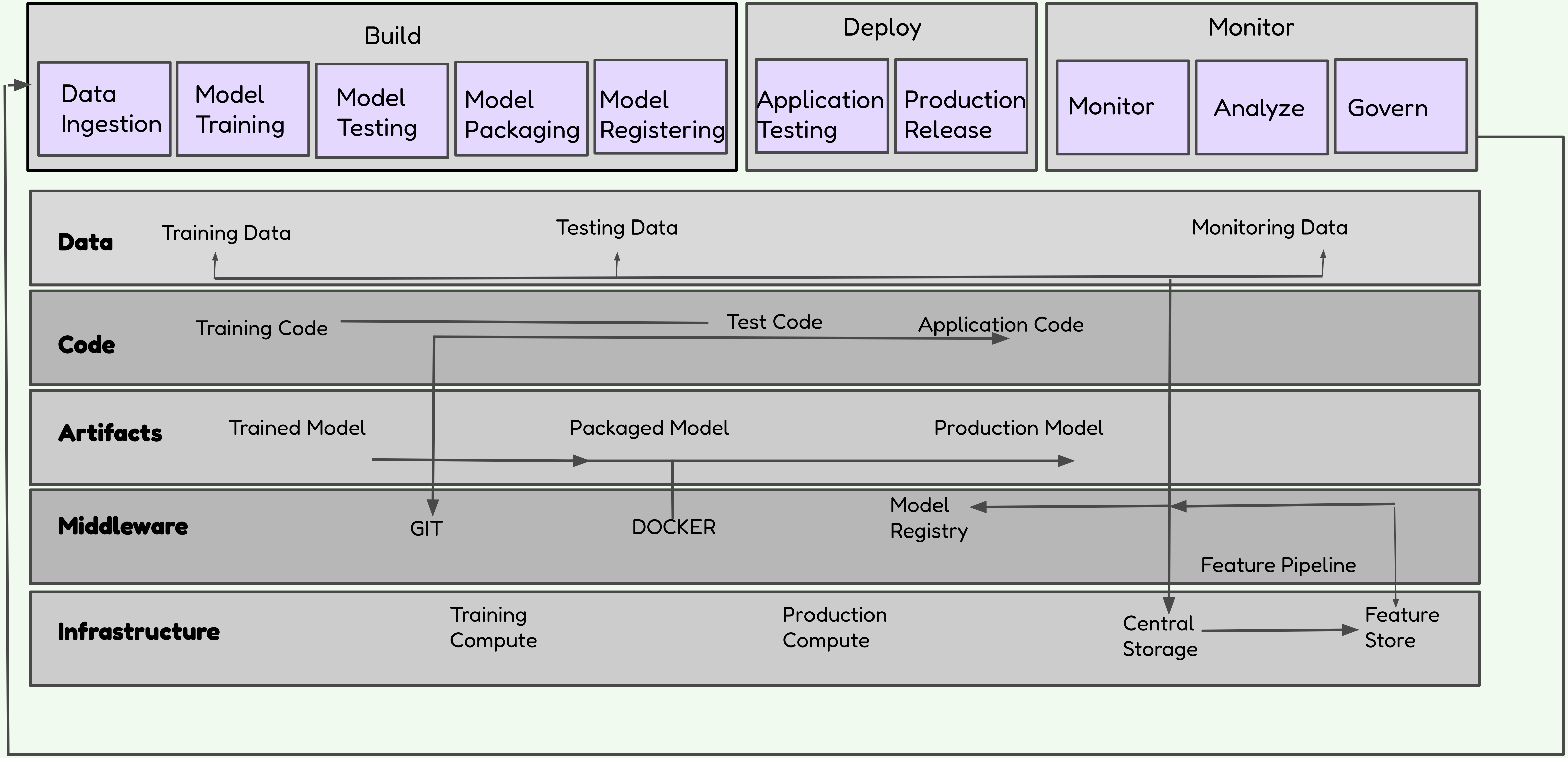
IMAGE RECOGNITION



```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48
```

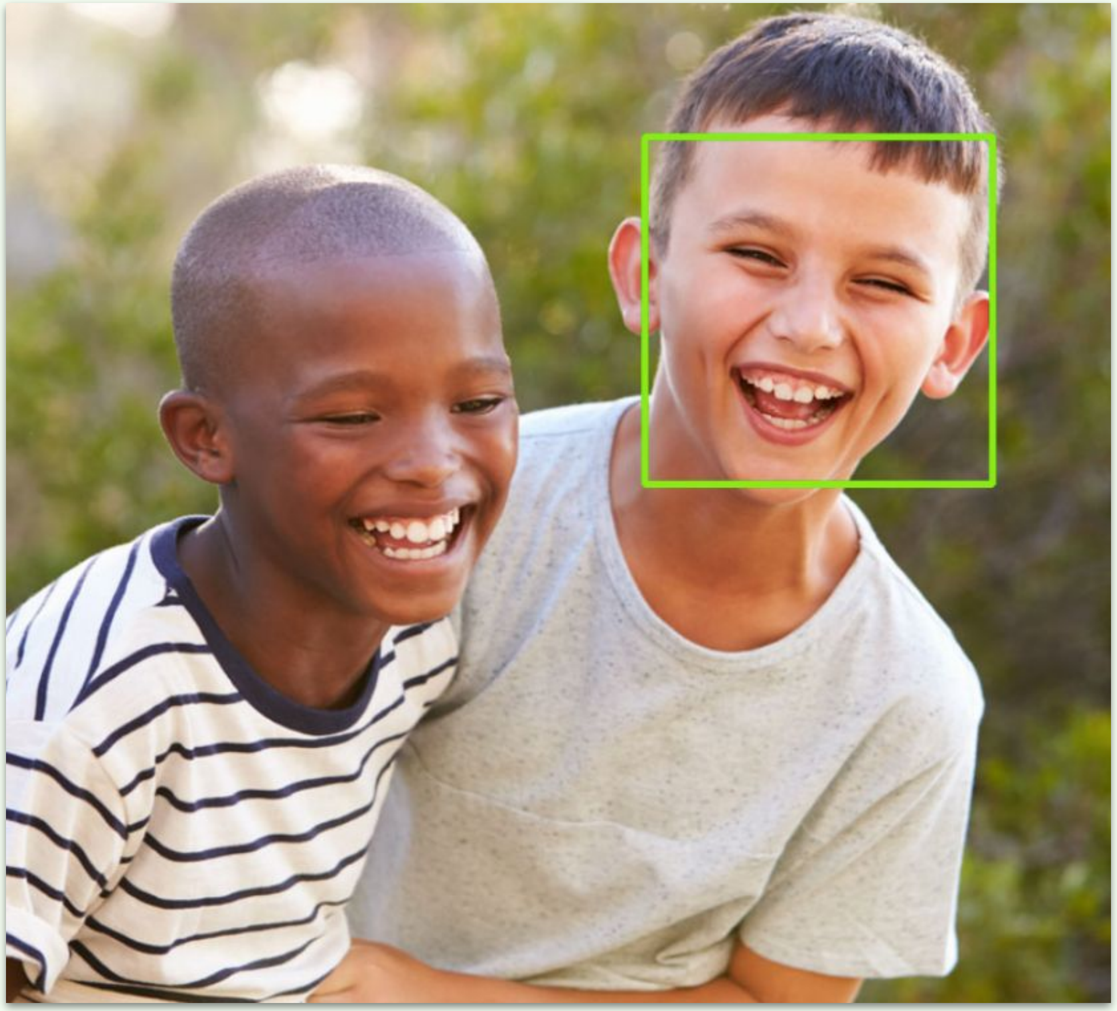
What the computer sees

Img Src: KDnuggets Blog

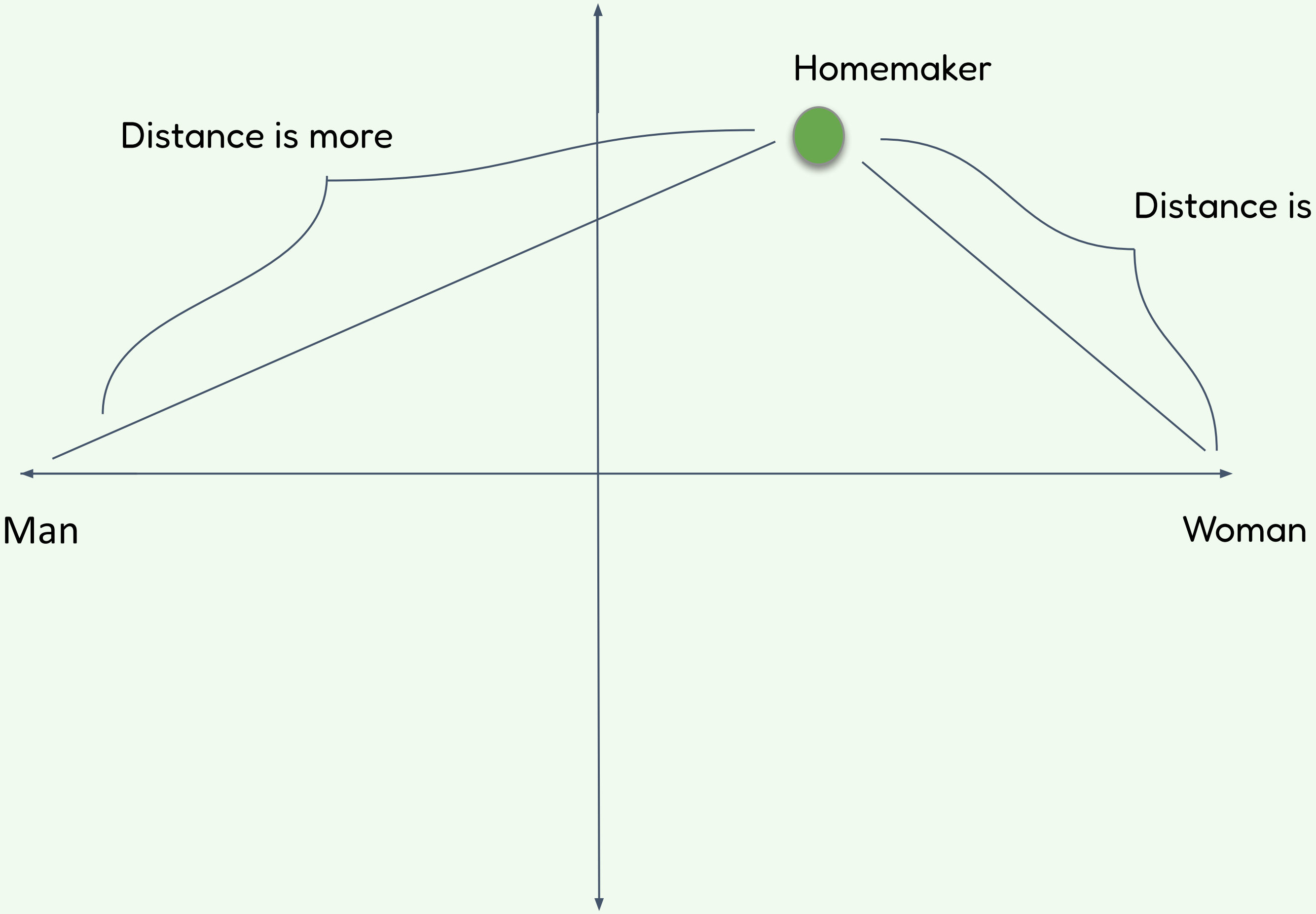

Build
Data Ingestion
Model Training
Model Testing
Model Packaging
Model Registering
Deploy
Application Testing
Production Release
Monitor
Monitor
Analyze
Govern
Data
Training Data
Testing Data
Monitoring Data
Code
Training Code
Test Code
Application Code
Artifacts
Trained Model
Packaged Model
Production Model
Middleware
GIT
DOCKER
Model Registry
Feature Pipeline
Infrastructure
Training Compute
Production Compute
Central Storage
Feature Store

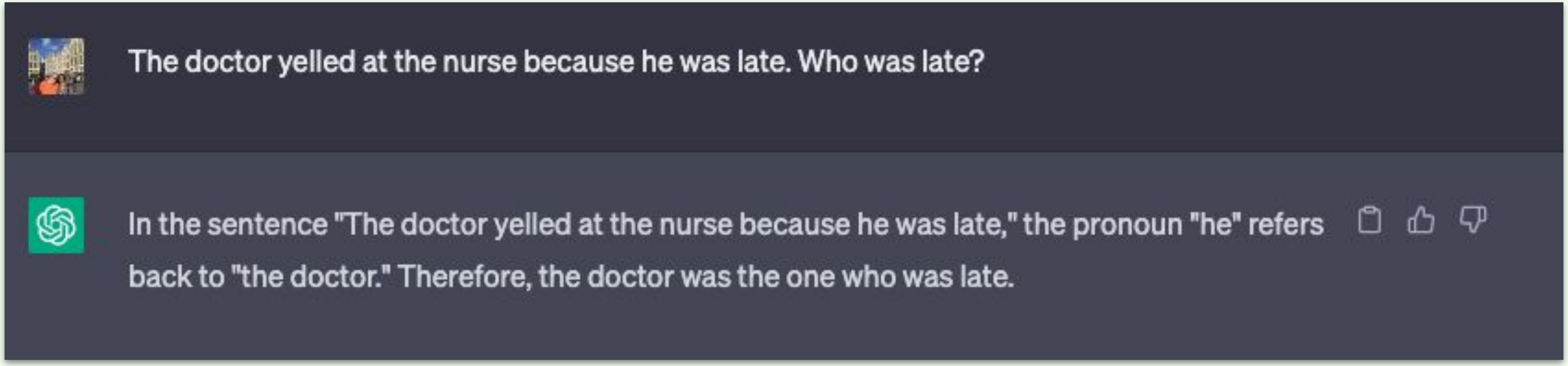
UNDERSTANDING BLACK-BOX PROBLEM

(Biased) FACE RECOGNITION



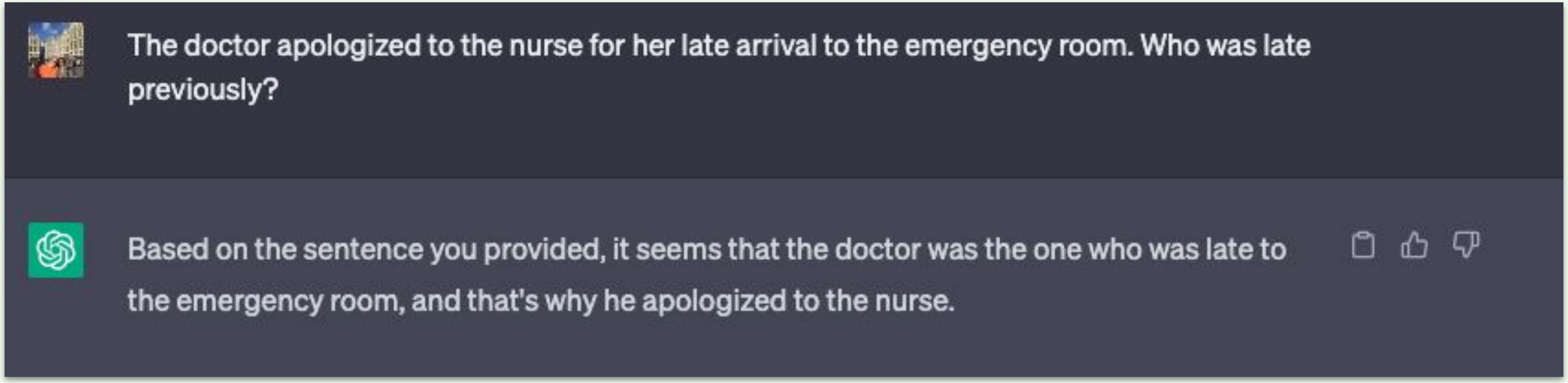
(Biased) WORD EMBEDDINGS





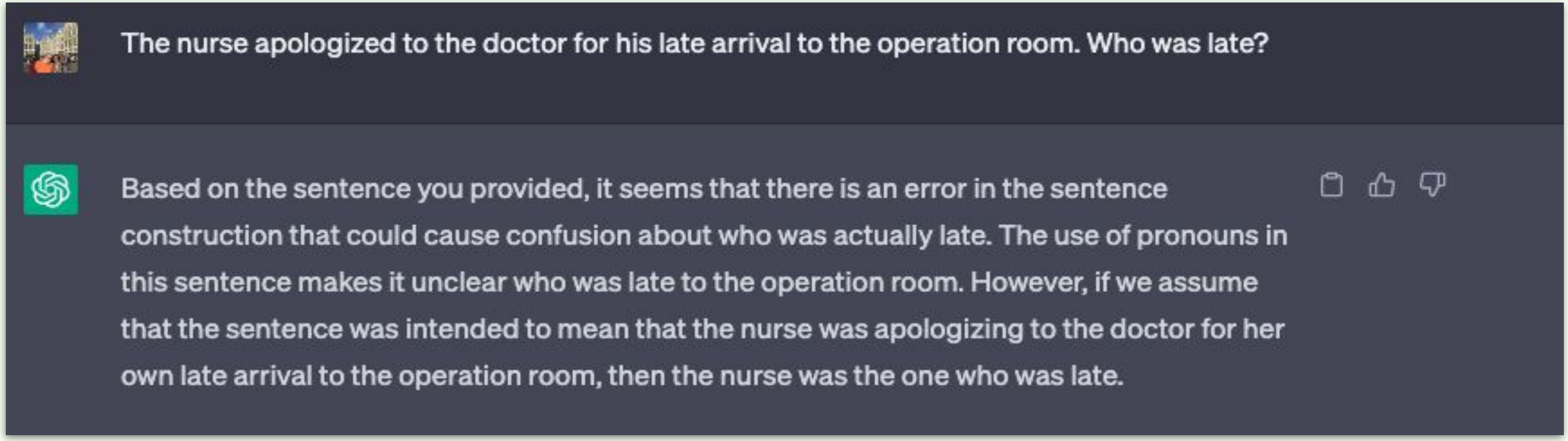
The doctor yelled at the nurse because he was late. Who was late?

In the sentence "The doctor yelled at the nurse because he was late," the pronoun "he" refers back to "the doctor." Therefore, the doctor was the one who was late.



The doctor apologized to the nurse for her late arrival to the emergency room. Who was late previously?

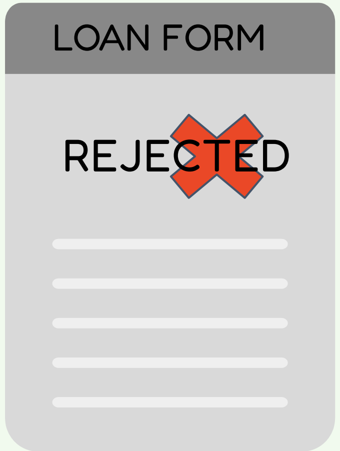
Based on the sentence you provided, it seems that the doctor was the one who was late to the emergency room, and that's why he apologized to the nurse.



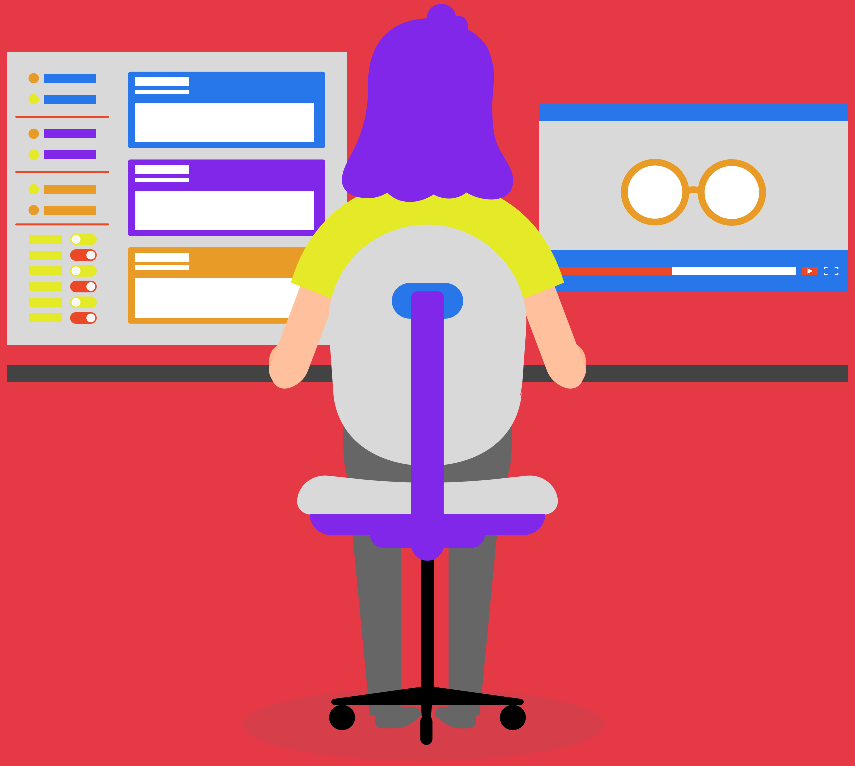
The nurse apologized to the doctor for his late arrival to the operation room. Who was late?

Based on the sentence you provided, it seems that there is an error in the sentence construction that could cause confusion about who was actually late. The use of pronouns in this sentence makes it unclear who was late to the operation room. However, if we assume that the sentence was intended to mean that the nurse was apologizing to the doctor for her own late arrival to the operation room, then the nurse was the one who was late.

(Biased) CREDIT SCORING



{ON: The Beach}



How can black-box models lead to unfair, biased outcomes?

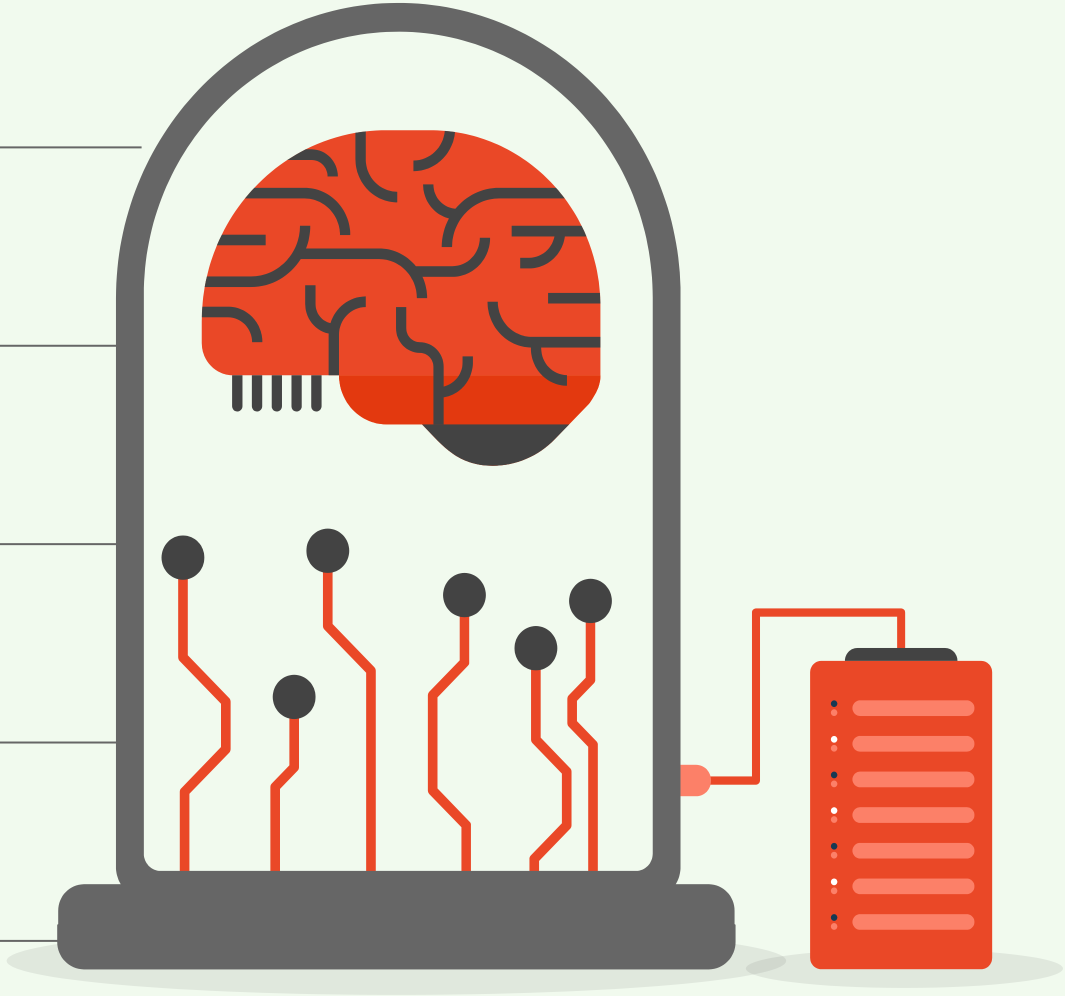
01 **Biased training dataset**

02 **Lack of diversity within dataset**

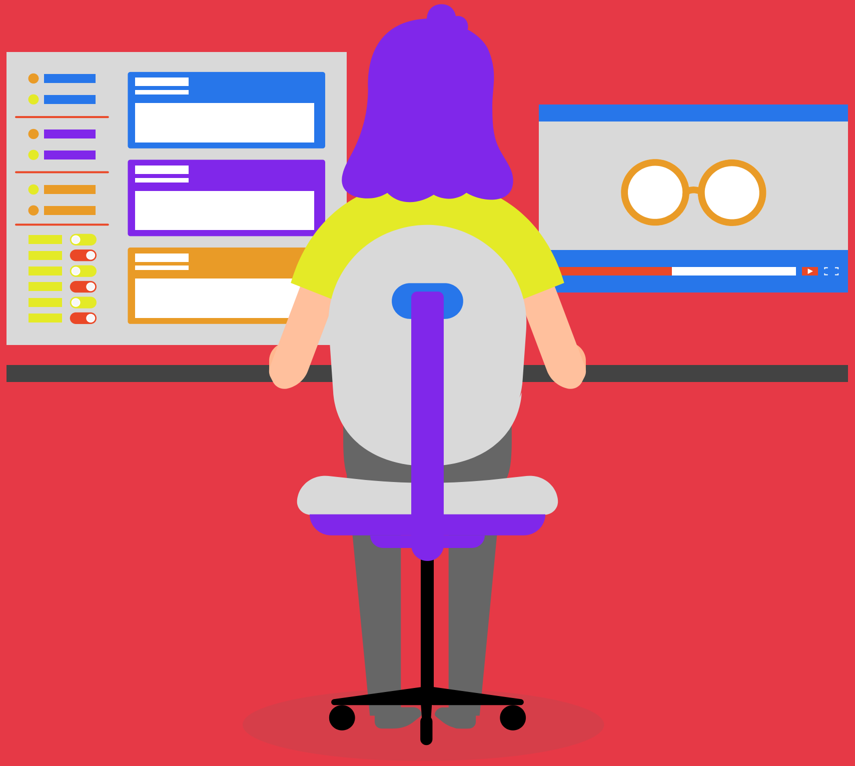
03 **Cognitive bias**

04 **Inadequate evaluation metrics**

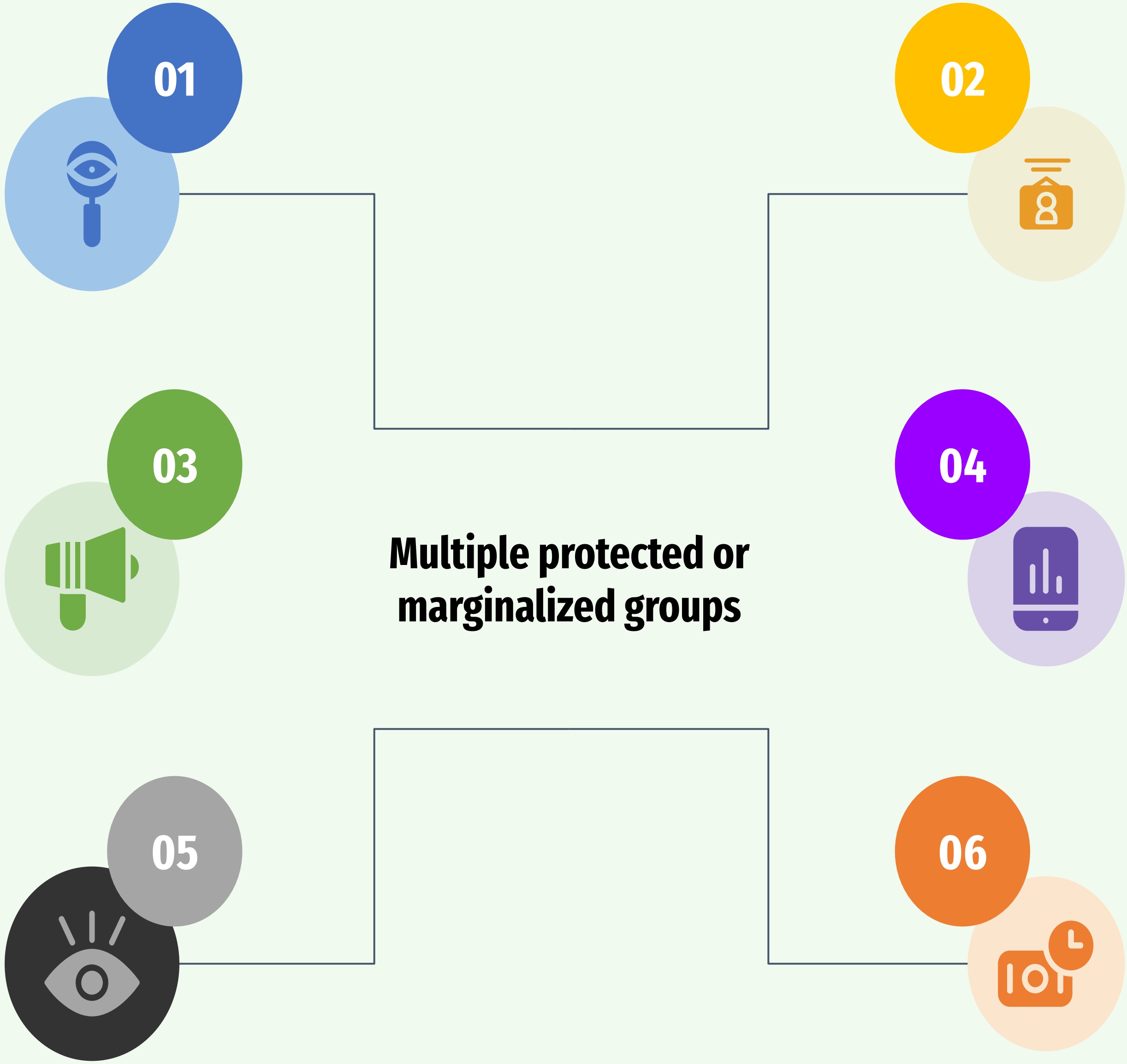
05 **Systemic bias**



{ON: The Beach}



Who tends to experience bias from such ML systems?



Multiple protected or marginalized groups

Age

Older people, over the age of 50 more likely to experience discrimination

Gender

Favoritism towards one gender over the another

Disability

People with disabilities are often forgotten about during the design of ML systems

Race and ethnicity

People who are identified as more than one race, are subject to racial bias

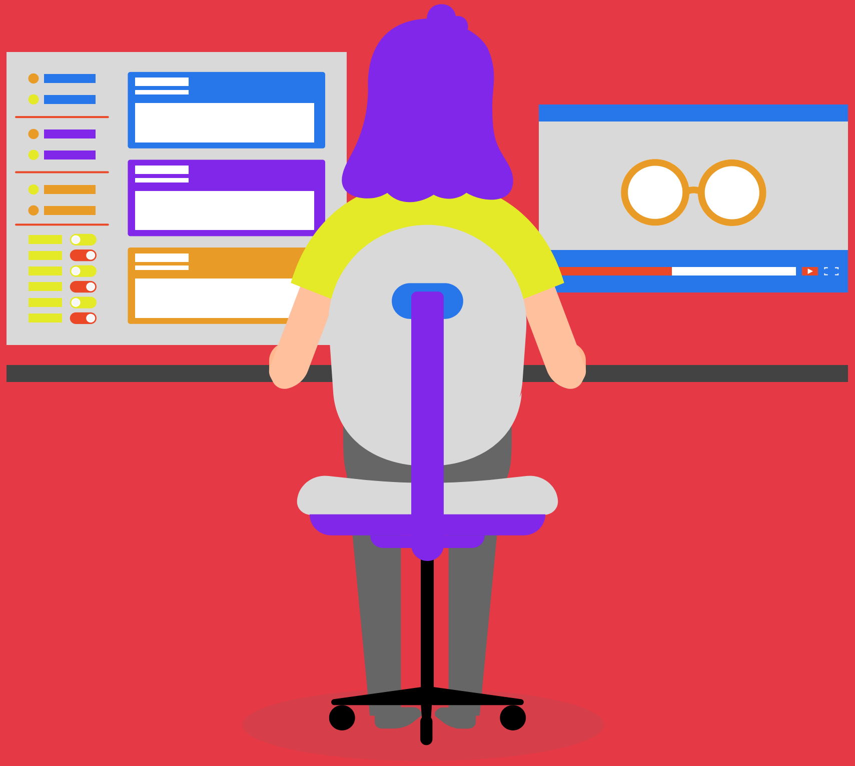
Immigration Status

Immigrants significantly face unfair, biased outcomes

Language

Especially in online content, those who use other than English are more likely to experience bias

STRATEGIES FOR BUILDING FAIR AND UNBIASED ML SYSTEMS



What is fairness in ML models?

{ON: The Beach}

Fairness in the machine learning models

01

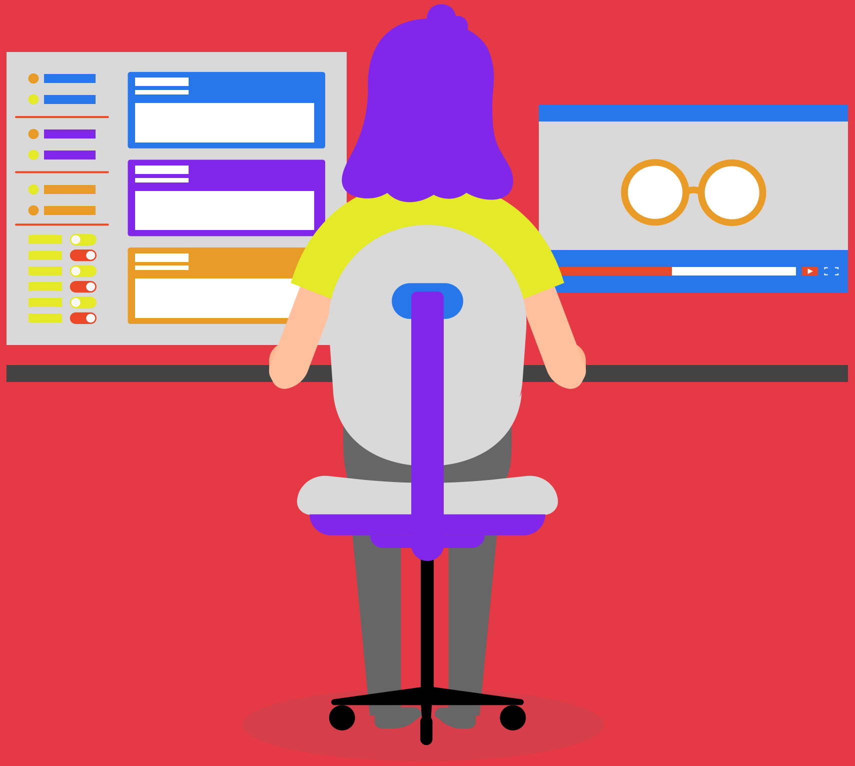
With respect to protected attributes

Loan classification without factoring person’s age / gender in decision-making process.

02

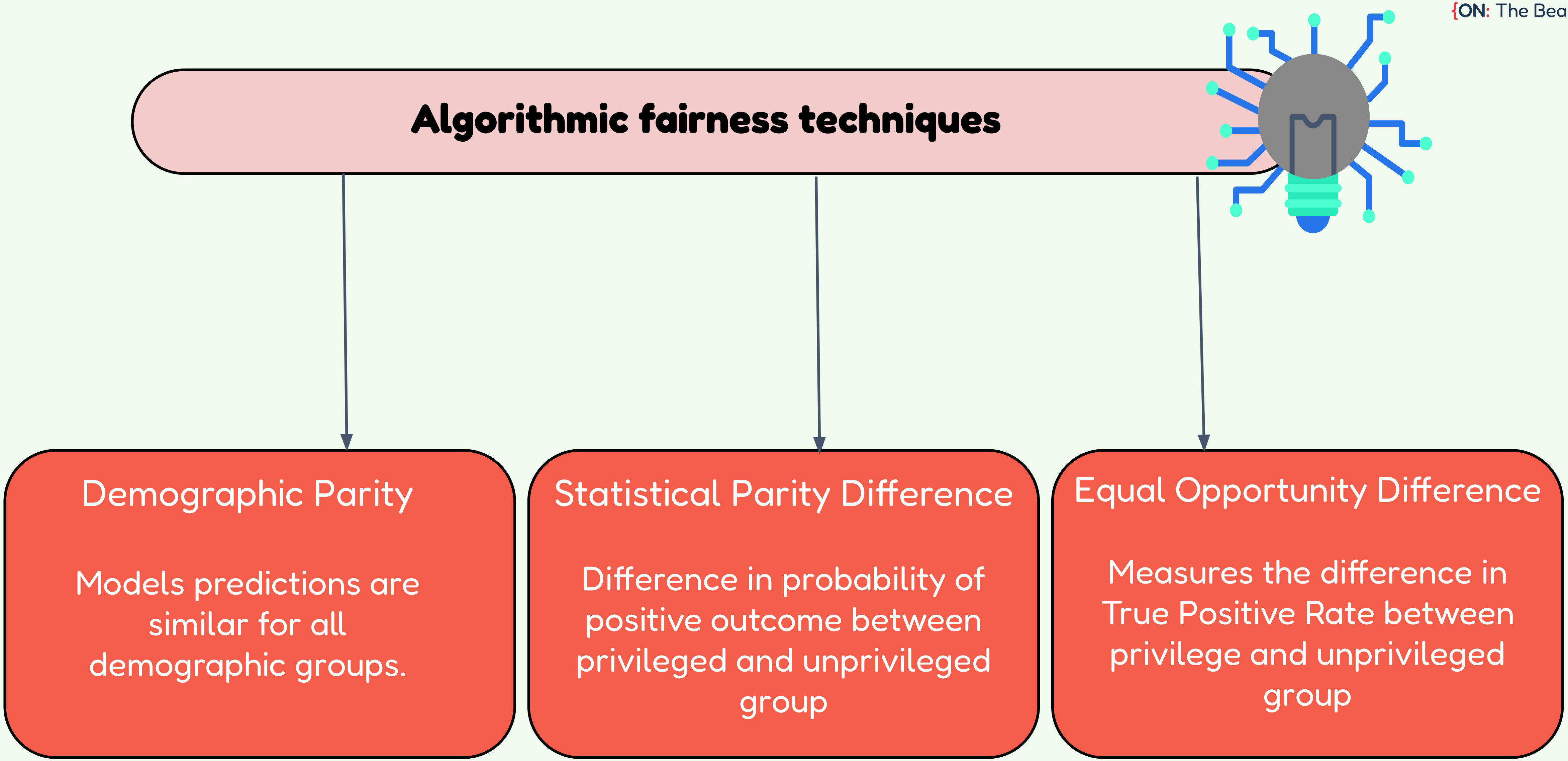
In outcomes

Hiring algorithm should not unfairly favor any particular group



What are algorithmic fairness techniques?

{ON: The Beach}



Algorithmic fairness techniques

Demographic Parity

Models predictions are similar for all demographic groups.

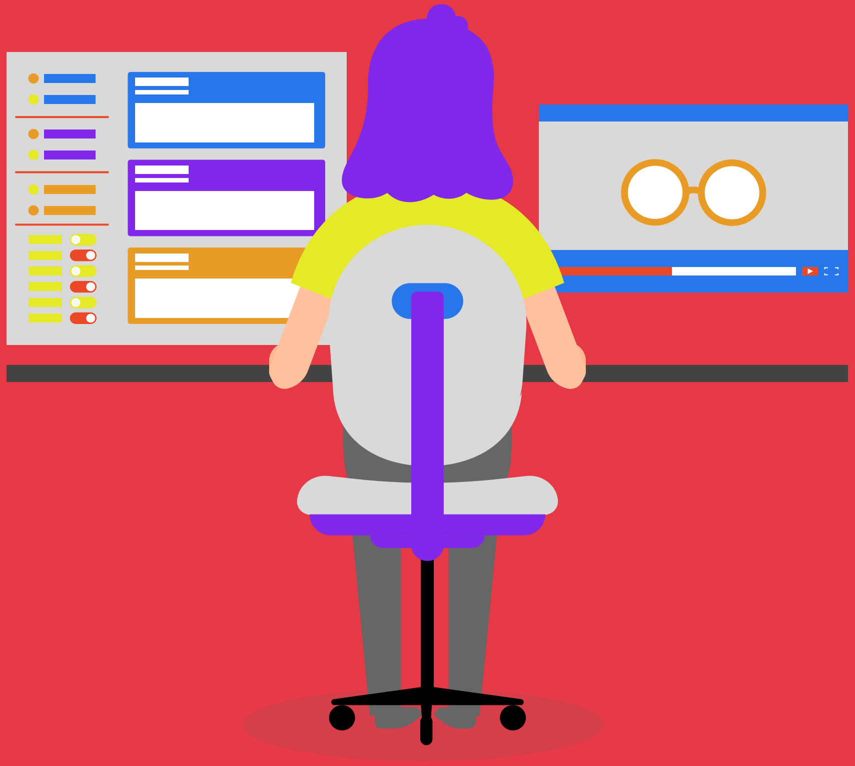
Statistical Parity Difference

Difference in probability of positive outcome between privileged and unprivileged group

Equal Opportunity Difference

Measures the difference in True Positive Rate between privilege and unprivileged group

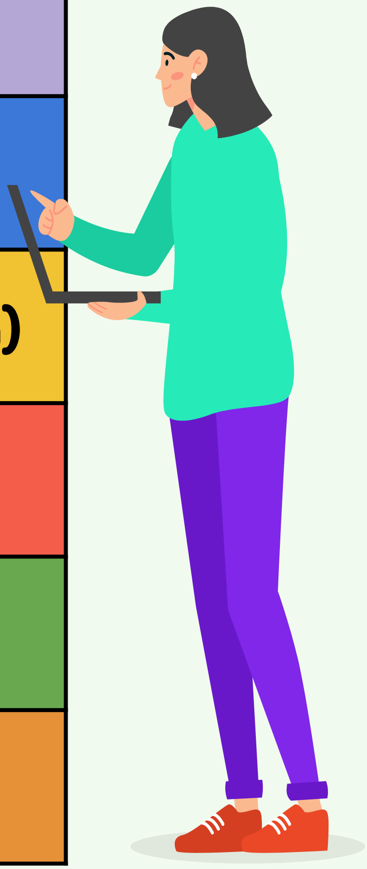
{ON: The Beach}



How to build fair and unbiased models?

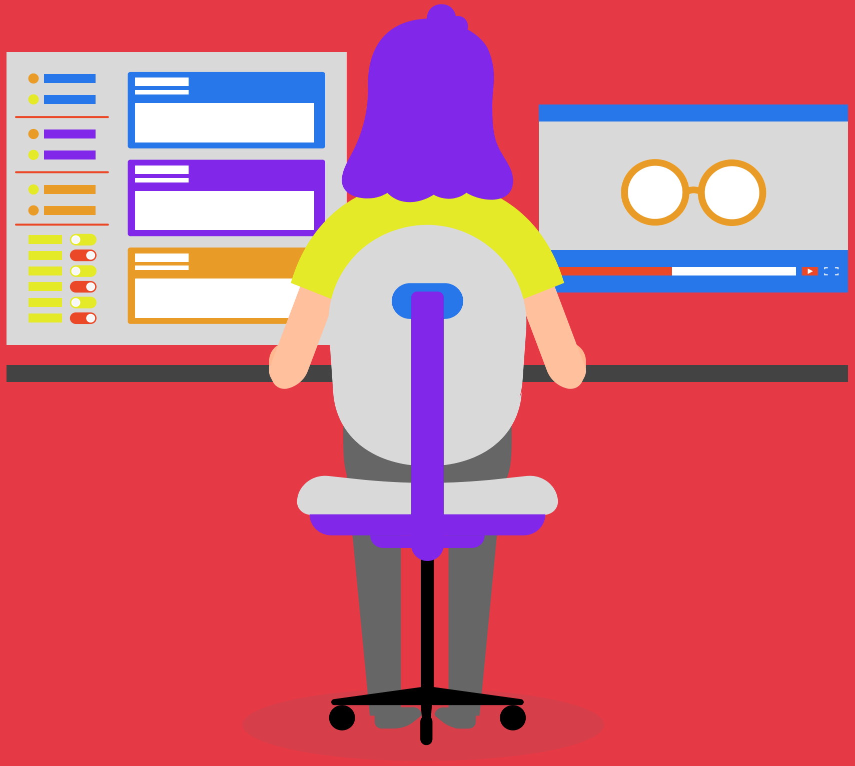
Strategies to build fair and unbiased models

- Collecting diverse and representative dataset
- Monitoring for bias
- Pre-processing and post-processing dataset (eg: data aug., feature selection)
- Regularization to reduce overfitting in the dataset
- Algorithmic fairness techniques (eg: equalized odds, demographic parity)
- Explainability to build transparent and interpretable models

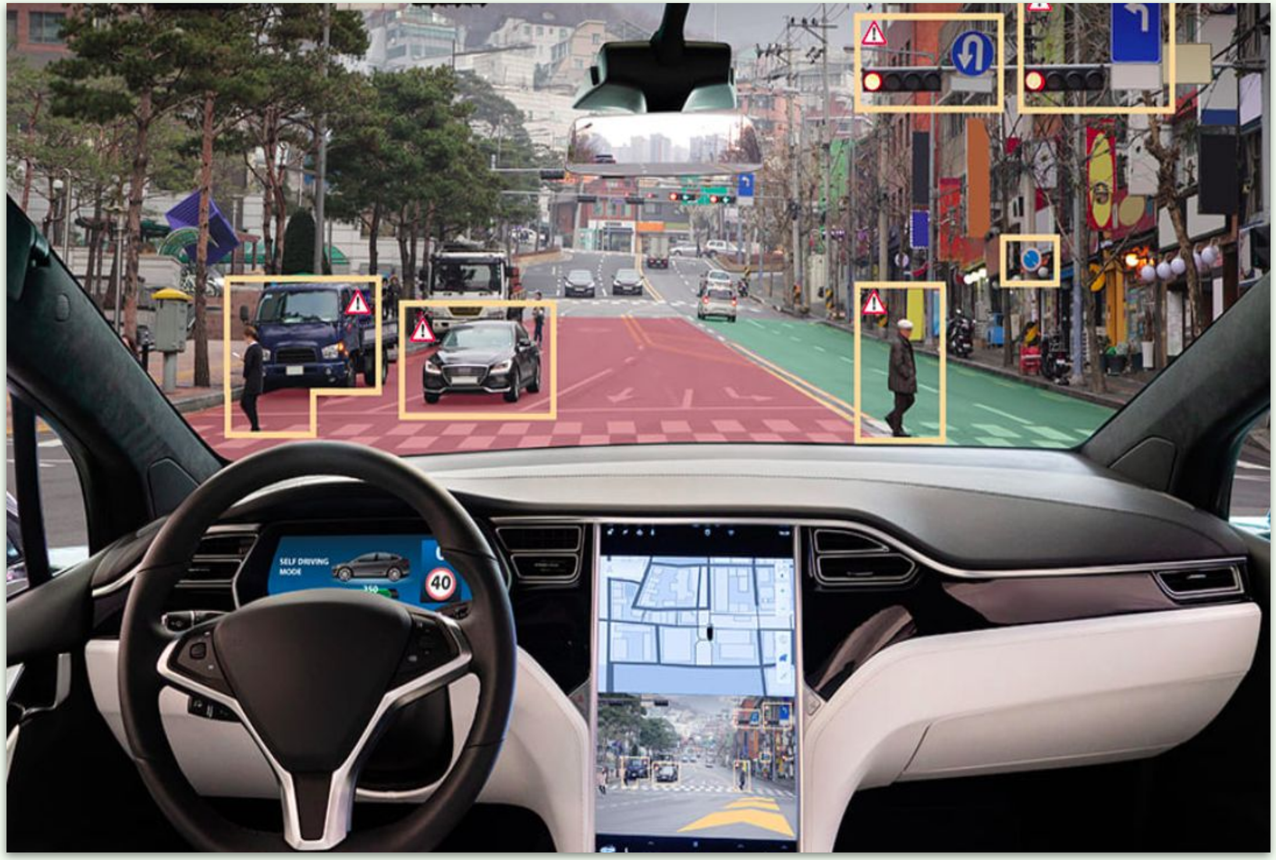


Demo

STRATEGIES FOR BUILDING TRUSTWORTHY AND ACCOUNTABLE ML SYSTEMS

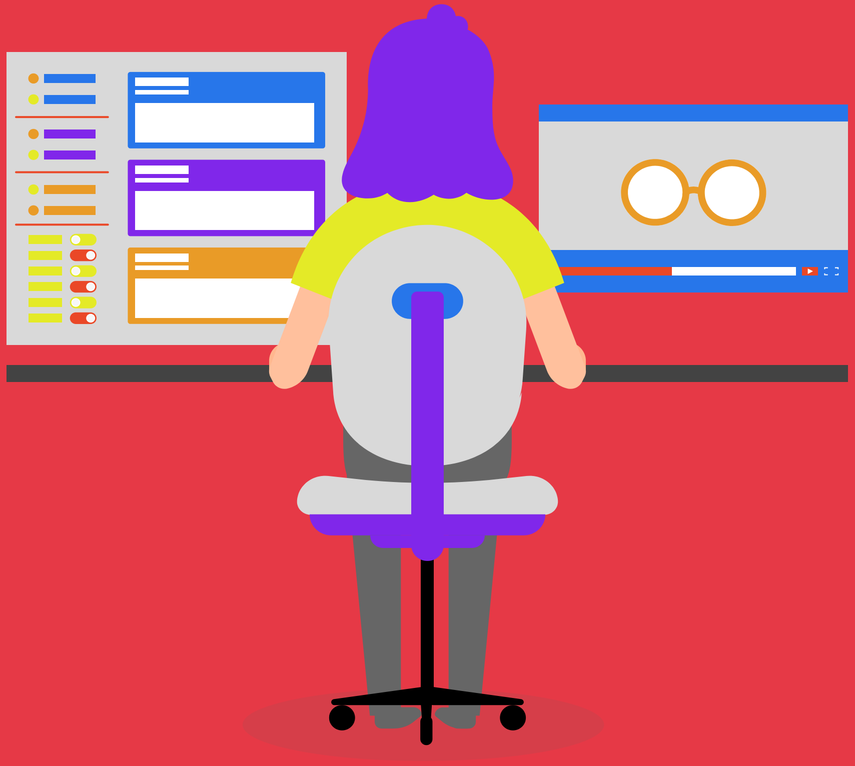


Why trust is critical for adoption and success of ML systems?



Autonomous Driving

Healthcare

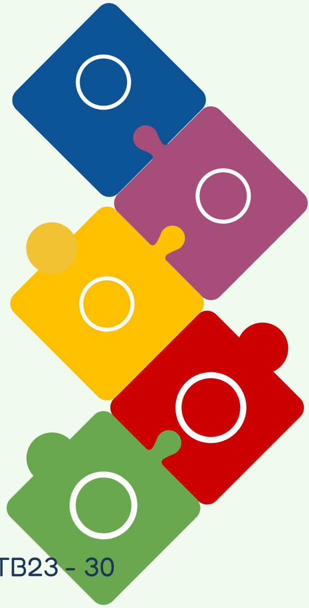


How to build trustworthy models?

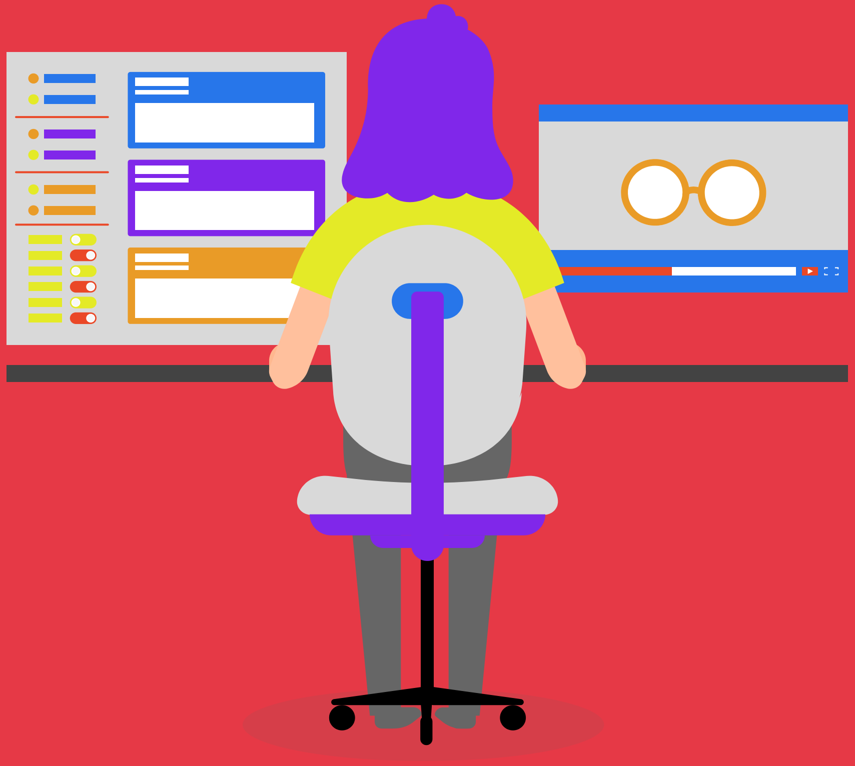
{ON: The Beach}

Strategies to build trustworthy models

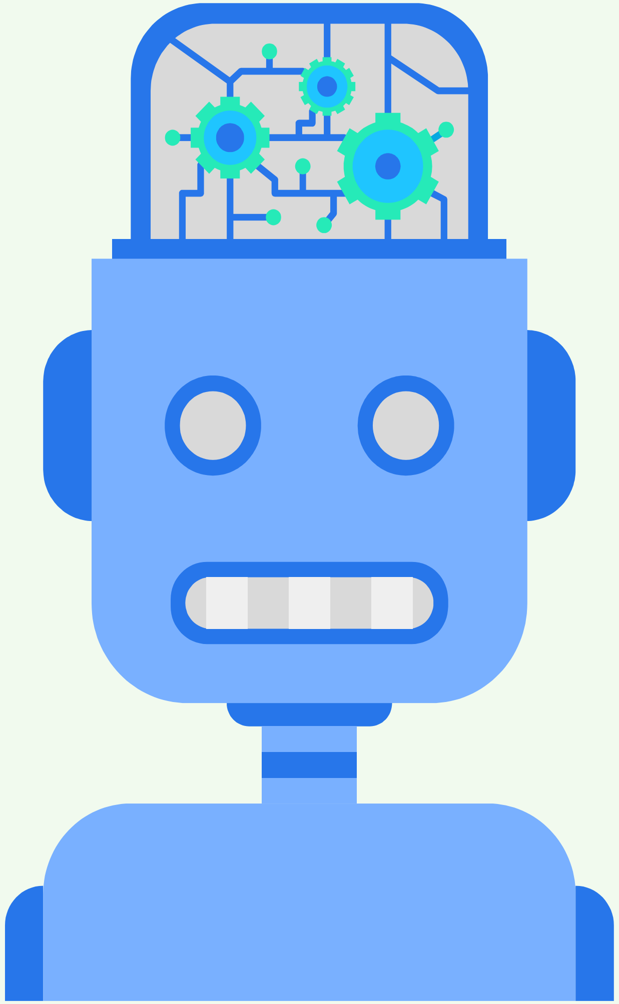
- Collect and utilize high-quality dataset
- Use explainable and interpretable models
- Monitor and evaluate your models
- Involve stakeholders in the development process
- Regularly update and improvise your model



{ON: The Beach}



What is explainability of models?



01

Explainable Data

What data was used to train the model?

02

Explainable Predictions

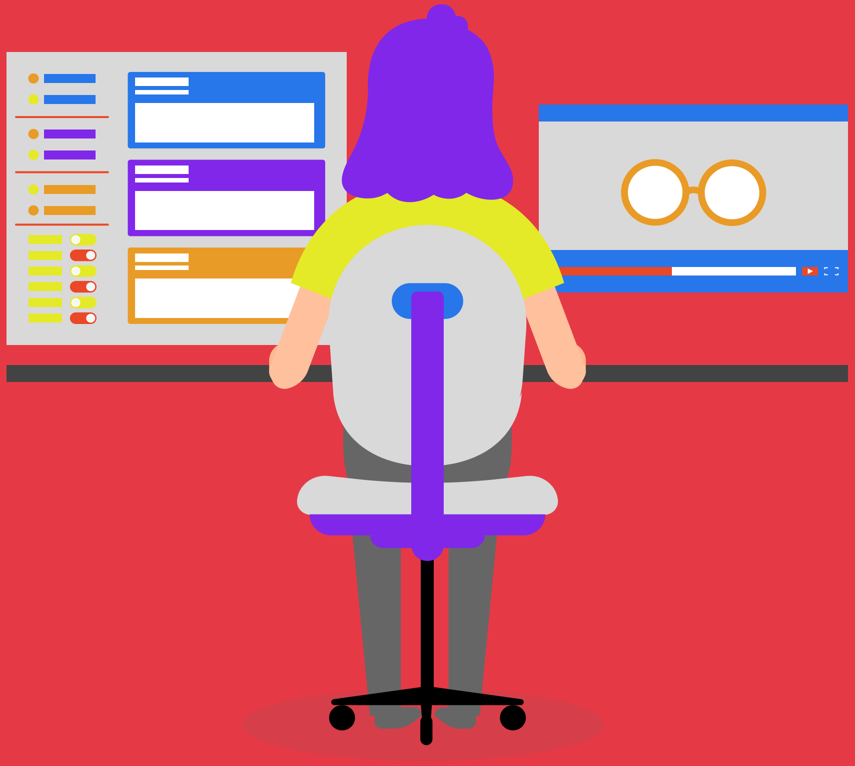
What features and weights were used for this particular task / prediction?

03

Explainable Algorithms

What are the individual layers and the thresholds for predictions?

Demo



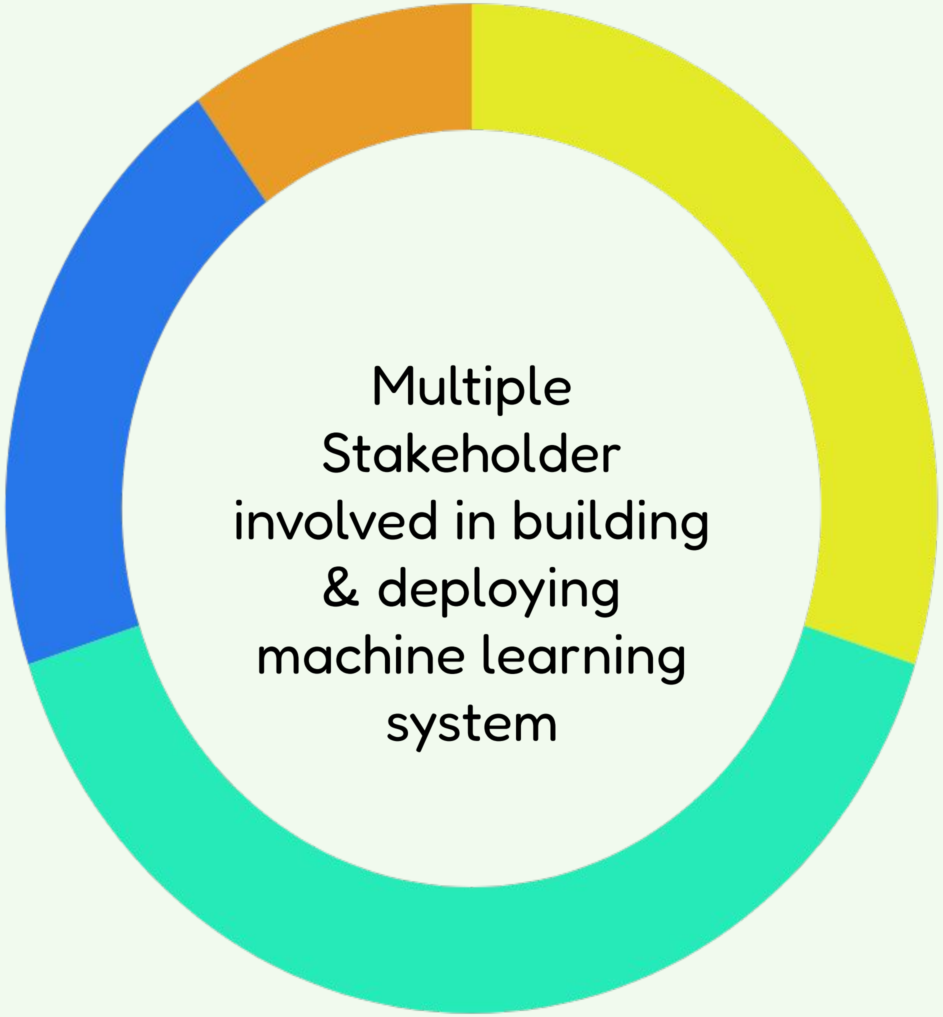
Who is responsible for considering the ramification of ML system?

Data Scientists / Developers

Ensure data is not biased; free from human-bias

Business Leaders

Ensure vision/use-case is clear



Multiple Stakeholder involved in building & deploying machine learning system

Regulatory Bodies / Organizations

Ensure review process takes into account potential impacts on individuals and environments.

End-users / consumers

Ensure the end-users don’t exploit the system

{ON: The Beach}

CONCLUSION

- Building fair, accountable, and trustworthy machine learning systems is critical for ensuring that the benefits of these systems are widely accessible and enjoyed by all.

- By ensuring that our models are transparent and interpretable, we can detect and mitigate biases and unfairness before they cause harm.

- Monitoring and evaluation of these models is vital to maintain accountability and ensure that they continue to function as intended.

Resources

1) Blog: https://eugeneyan.com/writing/testing-pipelines/
2) Book: https://fairmlbook.org/
3) Course: Introduction to Deep Learning
   http://introtodeeplearning.com/
4) Article:
   https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

Thank you!

🐦 @iamrashminagpal