MODULE 12: INTRODUCTION TO AI BIAS & FAIRNESS

BA713 - Machine Learning & AI

CONTENTS

AI BIAS WITH EXAMPLES

SOURCES & TYPES OF AI BIAS

MEASURING & MITIGATING AI BIAS

IMPORTANCE OF REDUCING BIAS IN AI ALGORITHMS

Discrimination of certain individuals, groups, religion, age, gender, skin color, race, culture, economic condition, marital status

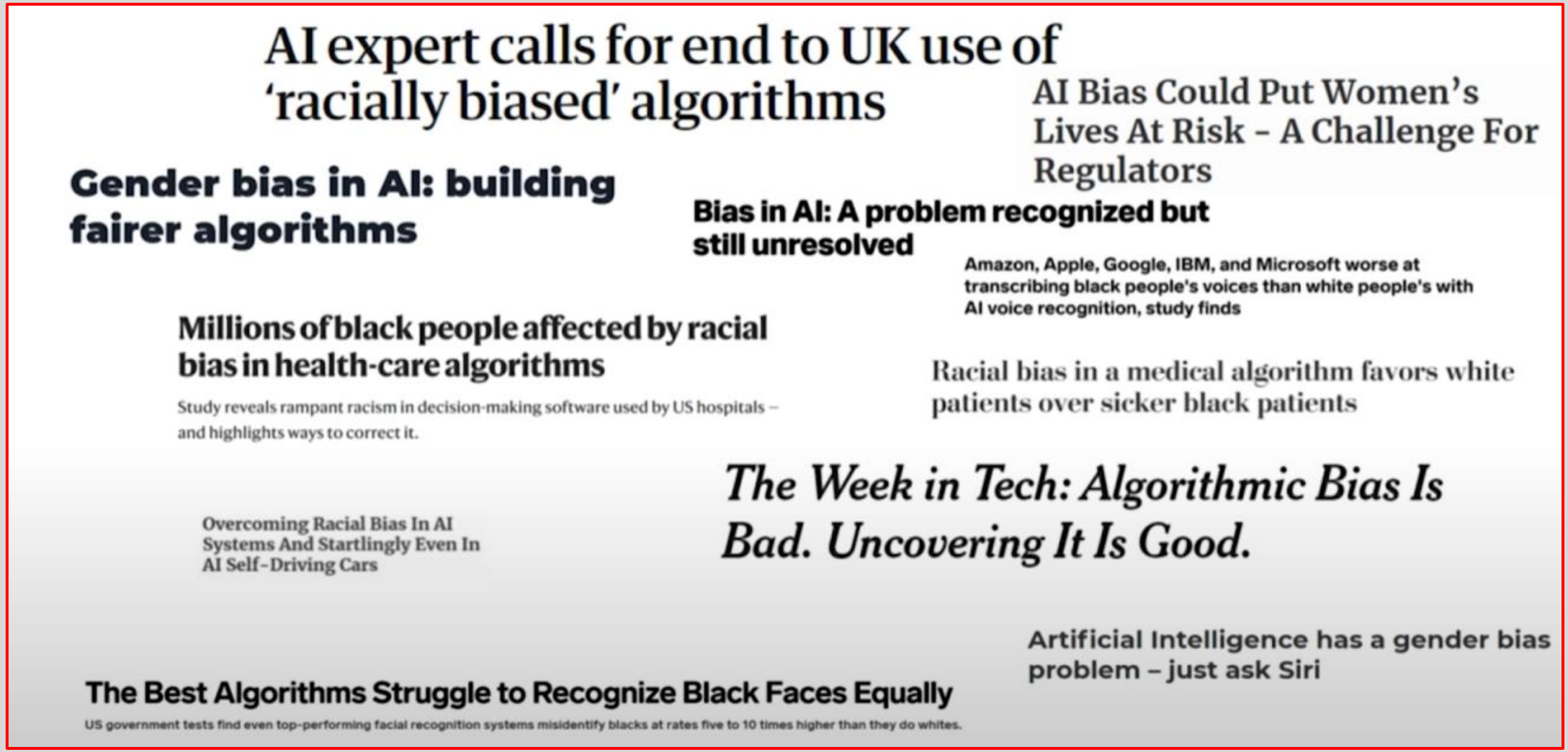
Disparate Impact→ unintentional unfairness that occurs when a decision has widely different outcomes for different groups

AI used to determine important real-world outcomes such as loan approval, pay rates

It is important on AI community to minimize unintentional discrimination

Source: https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/

AI BIAS IS VERY COMMON



Source: introtodeeplearning.com

WHAT IS AI BIAS?

- An anomaly in the output of an AI algorithm
- Exist in many shapes and forms
- prejudiced assumptions
- Introduced at any stage of the AI pipeline
- **Bias** is present inherently in the world around us, in our society
- We cannot directly solve the bias in the world
- We can take precautions to remove bias from different datasets used to train the AI algorithms
- Since AI has the potential to help humans to make fair decisions
- We need to work towards the **fairness** of the AI algorithms itself
- **Fairness** depends upon the situation in addition to the representation of your values, ethics and legal regulations

Source: https://www.datarobot.com/trusted-ai-101/ethics/bias-and-fairness/

EXPLAINING BIAS

What do you see in this image?

Image Source: https://www.medicalnewstoday.com/articles/279176

EXPLAINING BIAS

- cantaloupe (melon)
- cantaloupe slices
- cantaloupe with seeds
- Juicy cantaloupe
- cantaloupe slices next to each other
- But what about **orange cantaloupe** ?

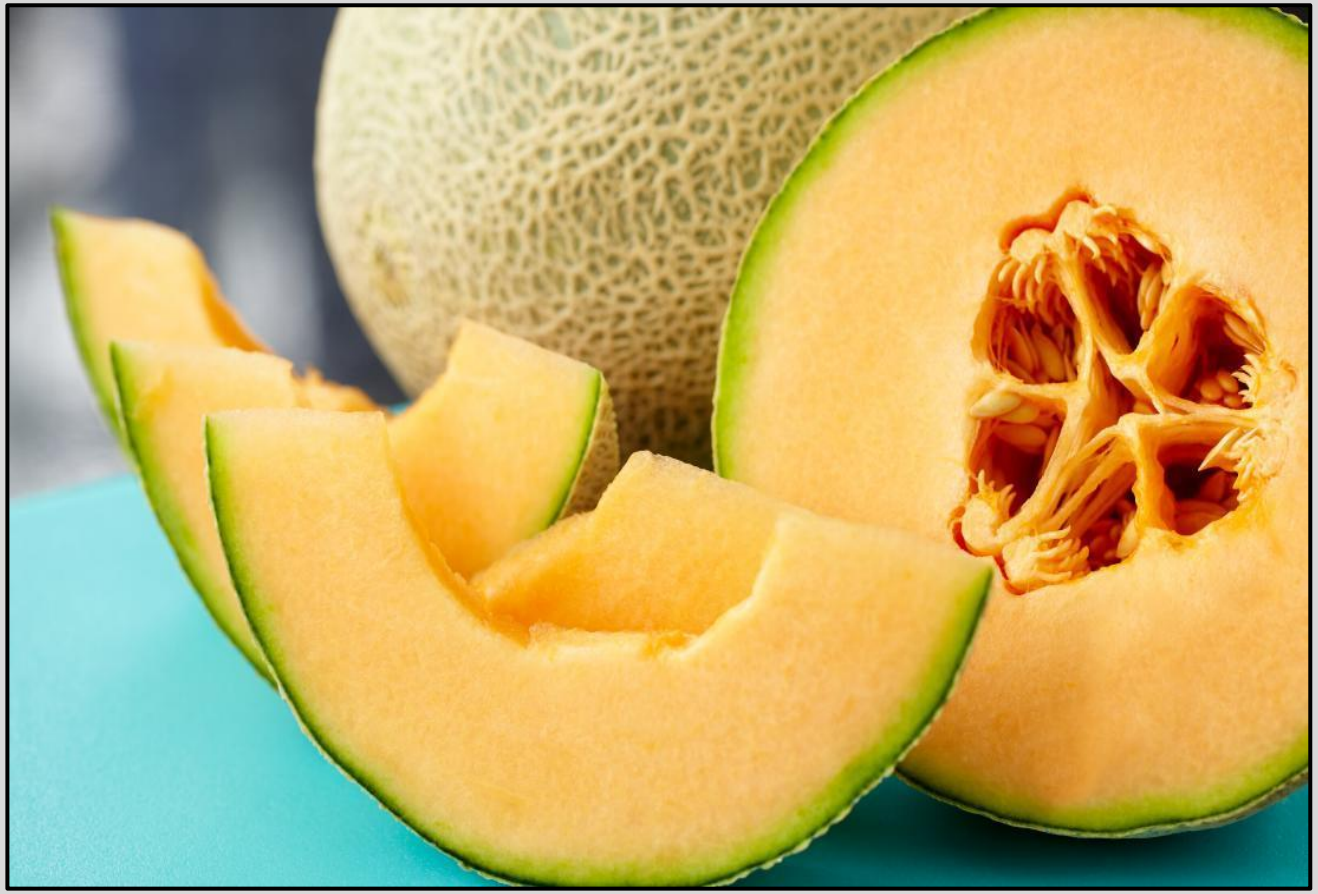
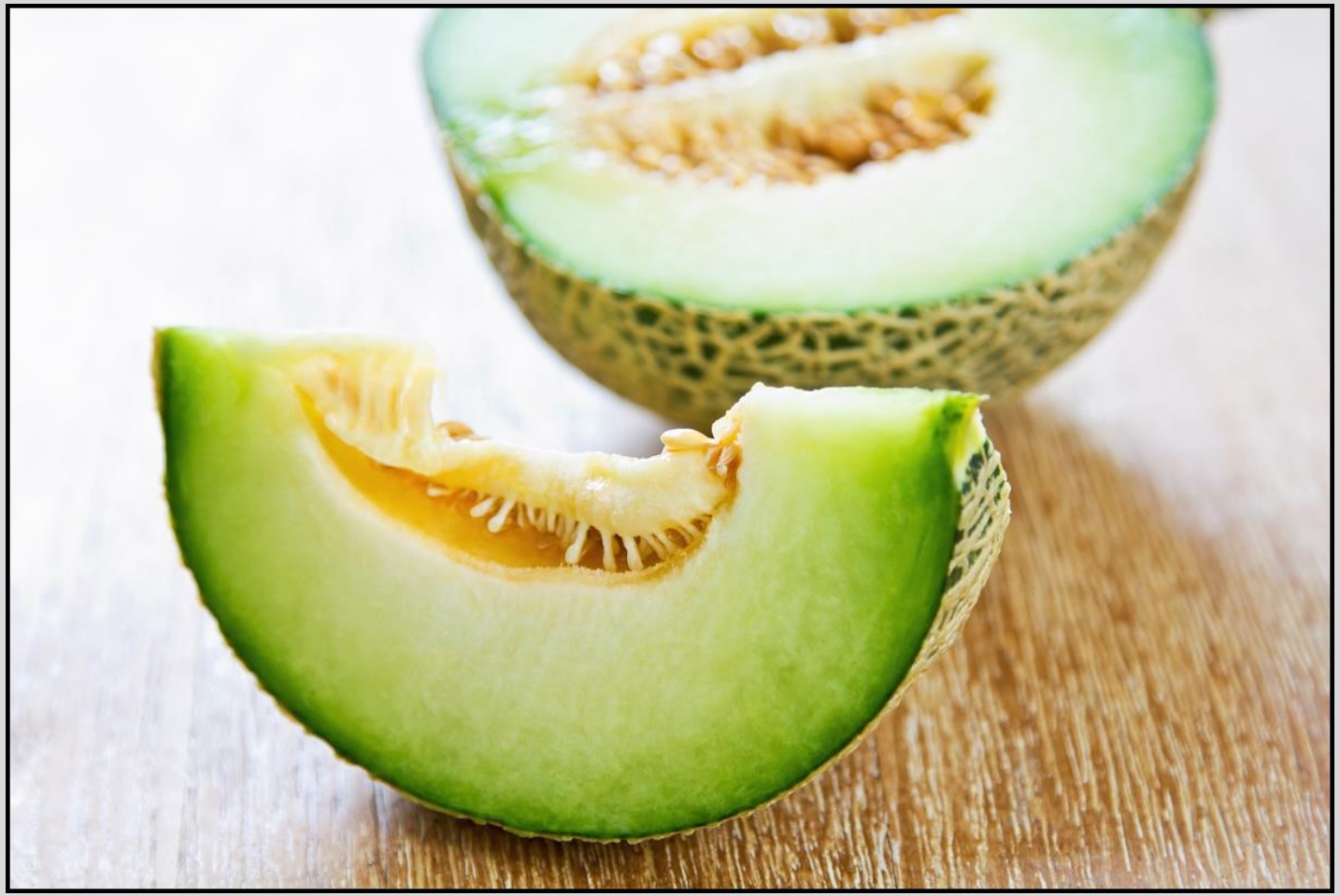


Image Source: https://www.medicalnewstoday.com/articles/279176

EXPLAINING BIAS

What do you see in this image?

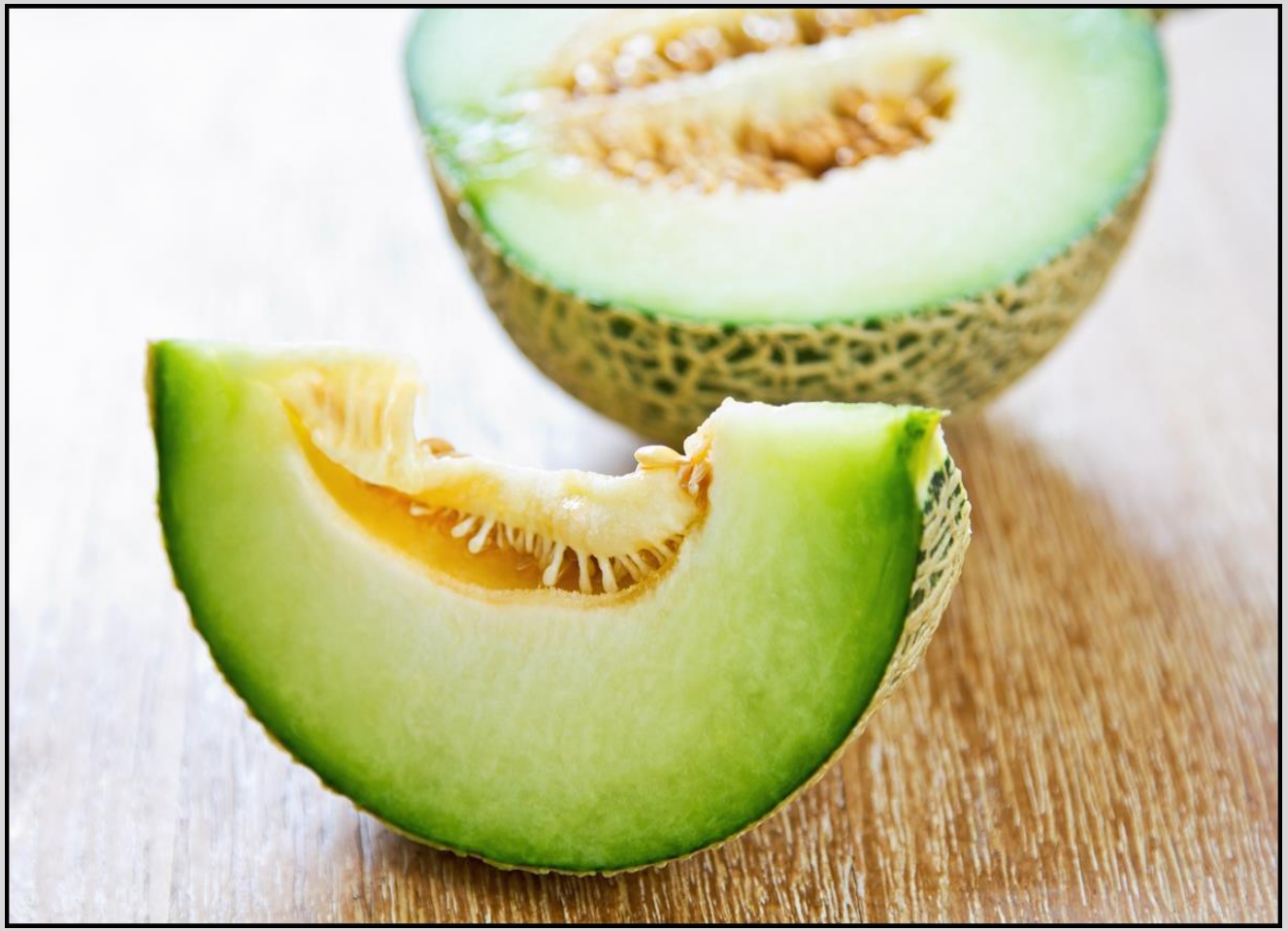


Source: https://servingjoy.com/health-benefits-of-honeydew-melon/

EXPLAINING BIAS

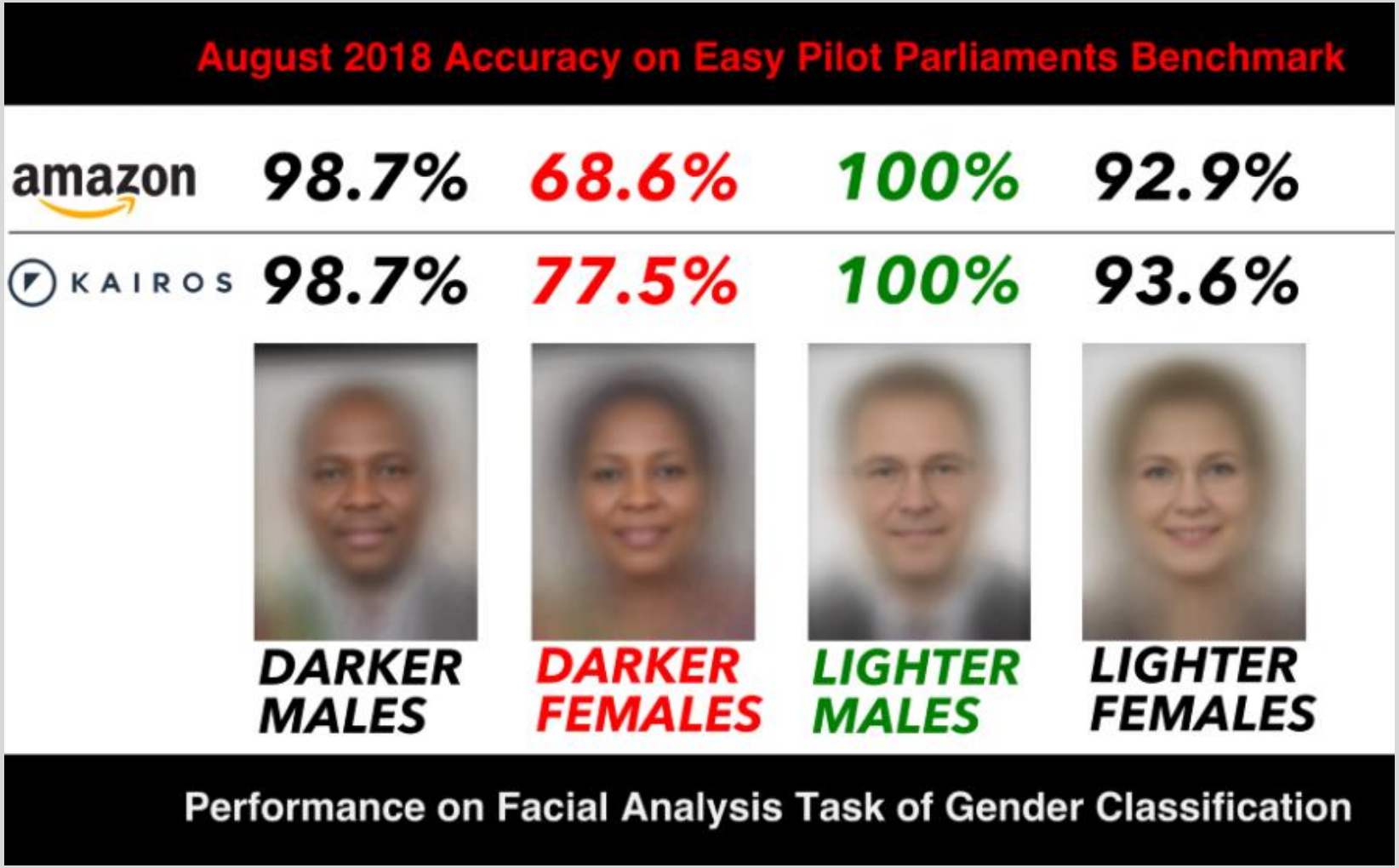
What do you see in this image?

- **Green** cantaloupe (melon)
- **Green** cantaloupe slices
- **Green** cantaloupe with seeds
- **Green** Juicy cantaloupe
- **Green** cantaloupe slices next to each other



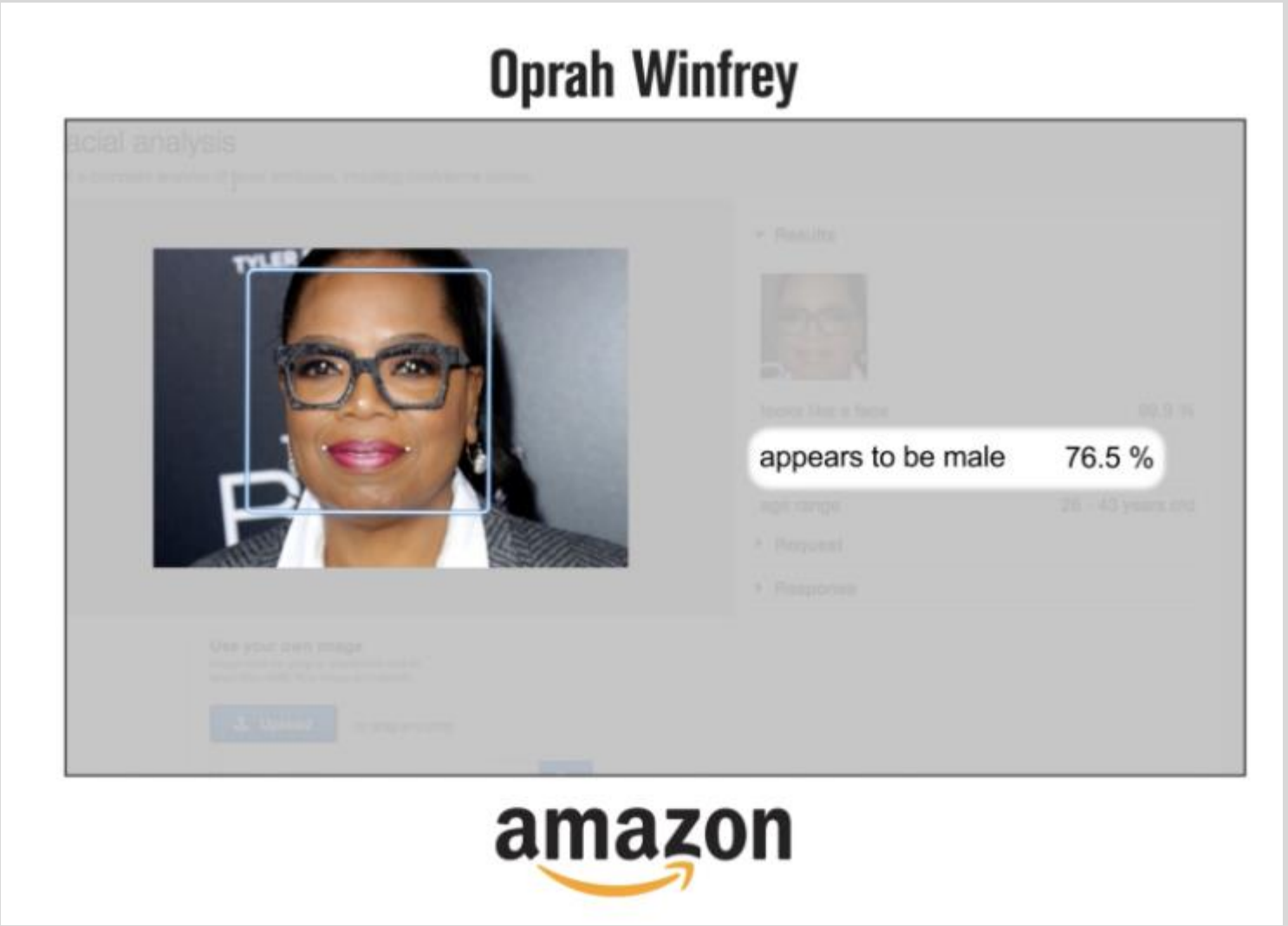
Source: https://servingjoy.com/health-benefits-of-honeydew-melon/

BIAS IN FACIAL DETECTION



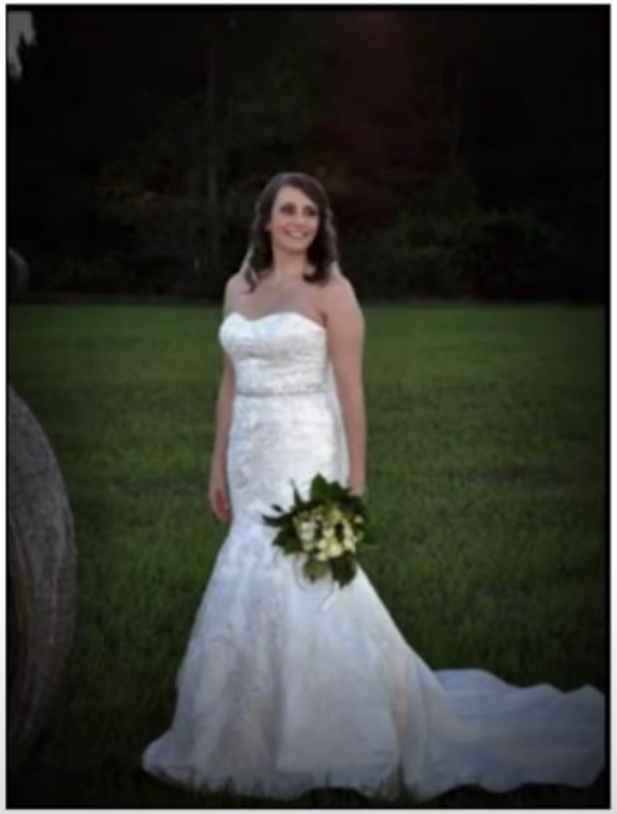
Source: https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

BIAS IN FACIAL DETECTION



Source: https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

BIAS IN IMAGE CLASSIFICATION



Ground Truth: Bride

CNN

CNN for image classification.

Predicted Classes

Bride
Dress
Ceremony
Woman
Wedding

Source: introtodeeplearning.com

BIAS IN IMAGE CLASSIFICATION

Ground Truth: Bride

CNN

CNN for image classification.

Predicted Classes

Clothing
Event
Costume
Red
Performance art

Source: introtodeeplearning.com

REAL-LIFE EXAMPLES OF AI BIAS

COMPAS:
software used by the US courts to judge the probability of a defendant→ Biased towards black defendants

Amazon's AI Recruiting System (2014):
discriminatory against women, use data from past 10 years, most selected applicants were men due to the male dominance in the tech industry, scraped in 2018

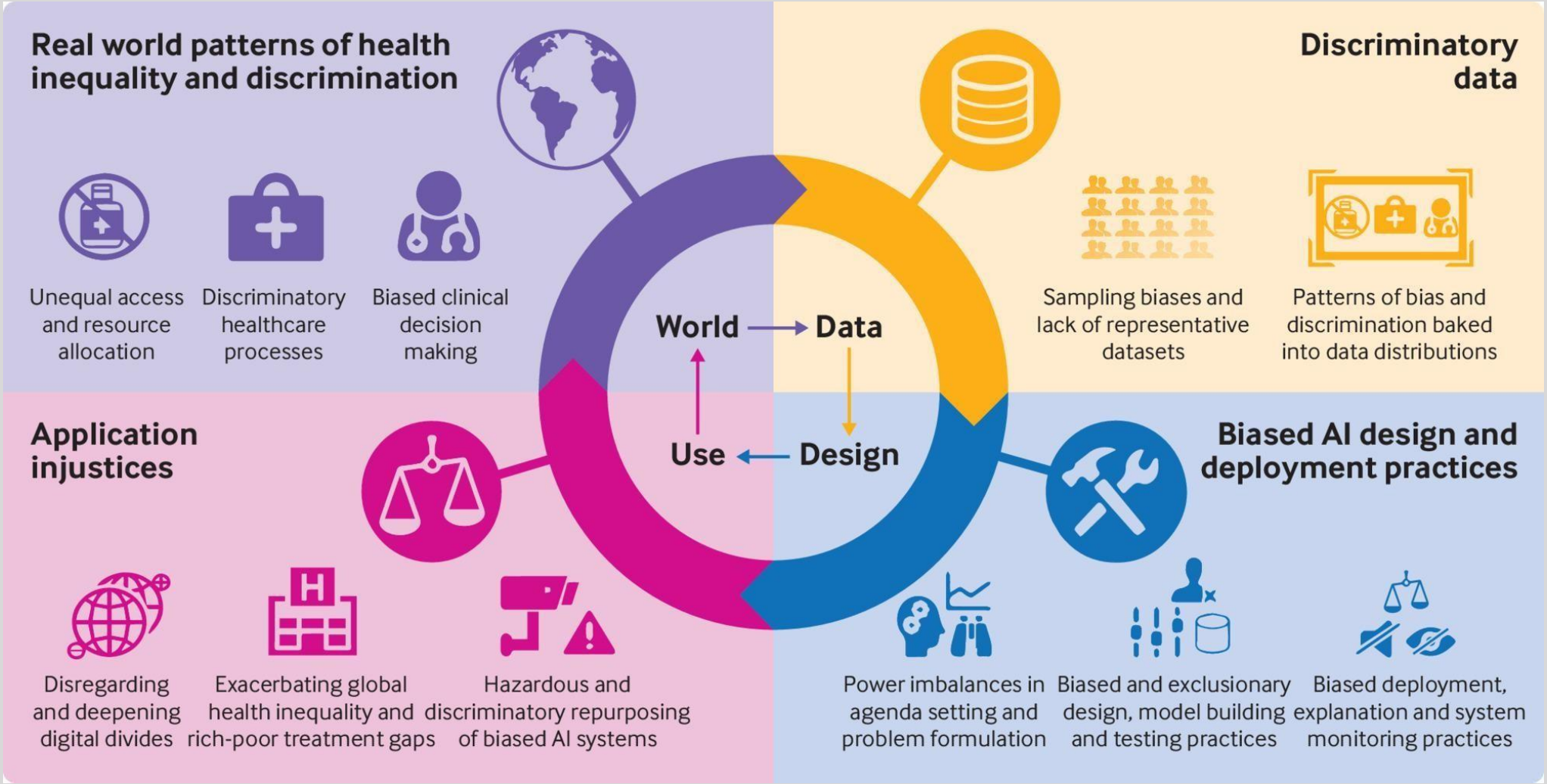
US Healthcare:
Since it is perceived, that Black people are less able to pay the costs, the AI ranked their health risk lower than white people for the same disease

Twitter Image Cropping (2020):
image cropping algorithm favored white faces over black faces

Facebook's Advertisement Algorithm (2019):
advertisers target people based on their race, gender, and religion
For example, jobs like nursing and secretary being targeted to women while jobs like janitor and taxi drivers targeted to men of color

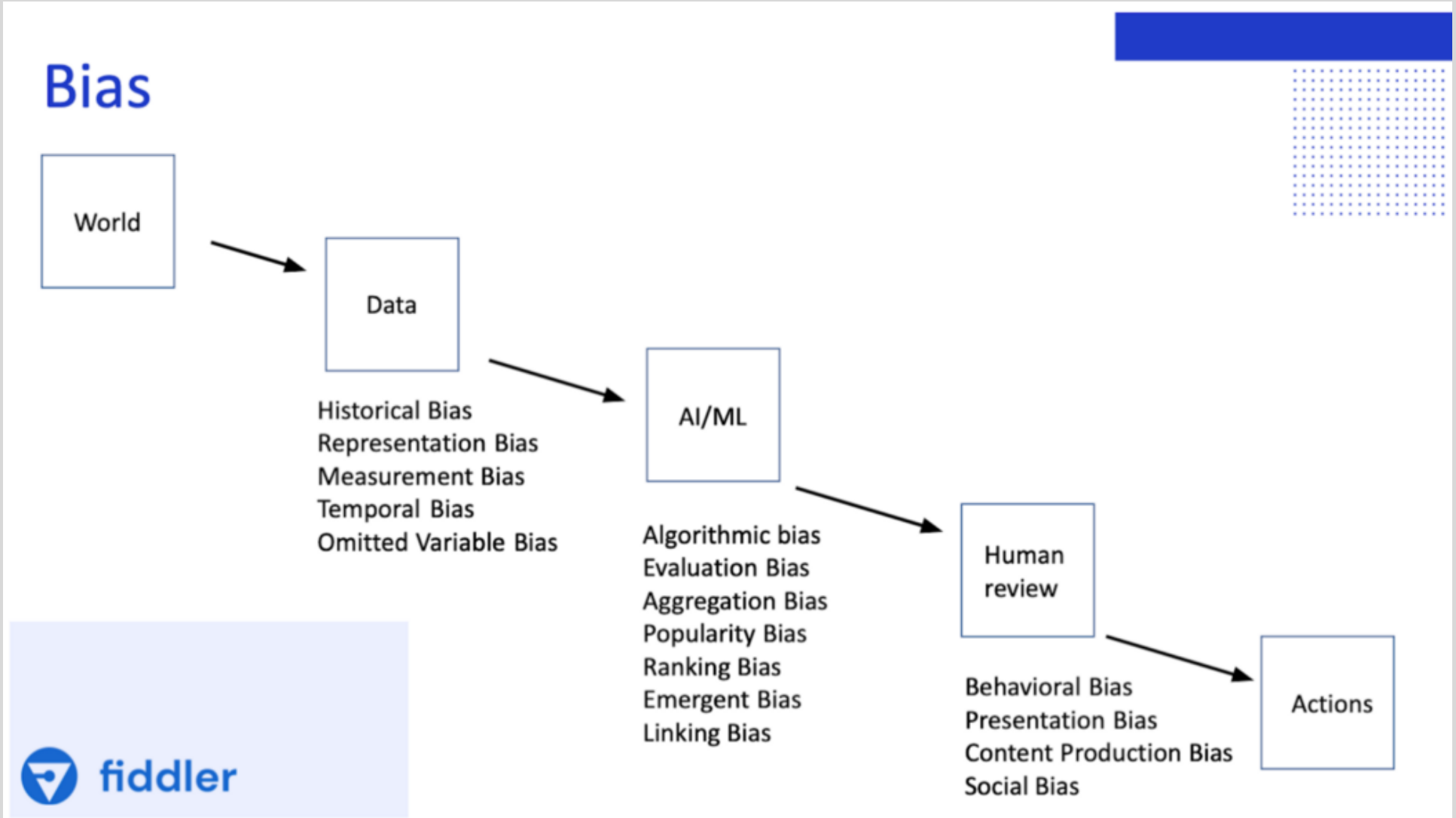
Source: https://towardsdatascience.com/ai-is-flawed-heres-why-3a7e90c48878

SOURCES OF AI BIASES



Source: https://research.aimultiple.com/ai-bias/

TYPES OF AI BIAS



Source: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

BIAS IN THE AI TRAINING DATASETS

- **Data imbalance, class labels, features, input structure**

1) Historical Bias:
- already exists in the world and is present in our training data
- **For example**, 2016 paper "**Man is to Computer Programmer as Woman is to Homemaker**"

2) Representation bias:
- the way we define and sample a population to create a dataset
- **For example**, the data used to train Amazon's facial recognition was mostly based on white faces

3) Measurement bias:
- When collecting features or labels to use in predictive models
- **For example**, predictive policing, likelihood that a black defendants will commit another crime leading to harsher sentences for them as compared to white defendants

Source: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

BIAS IN AI MODELS

- **uncertainty, interpretability, and performance metrics, training feedback loops perpetuate biases**
- **lack of systematic analysis with respect to data subgroups**

1) Evaluation Bias:

- Occurs during model iteration and evaluation
- Benchmark models do not represent the general population, or are not appropriate for the way the model will be used
- **For example,** evaluation metrics such as balanced accuracy and FPR must be consistent for different social groups (gender, ethnicity or age)

2) Aggregation bias:

- Occurs during model construction
- distinct populations are inappropriately combined
- For the population is heterogenous, a single model will not suit the needs of all groups
- **For example,** in medical applications such as diagnosing and monitoring diabetes, AI models use Hemoglobin levels as a predictor. But research studies show that these levels may differ across different ethnic groups

Source: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

BIAS IN HUMAN REVIEW

- **human errors and biases distort the meaning of AI predictions**

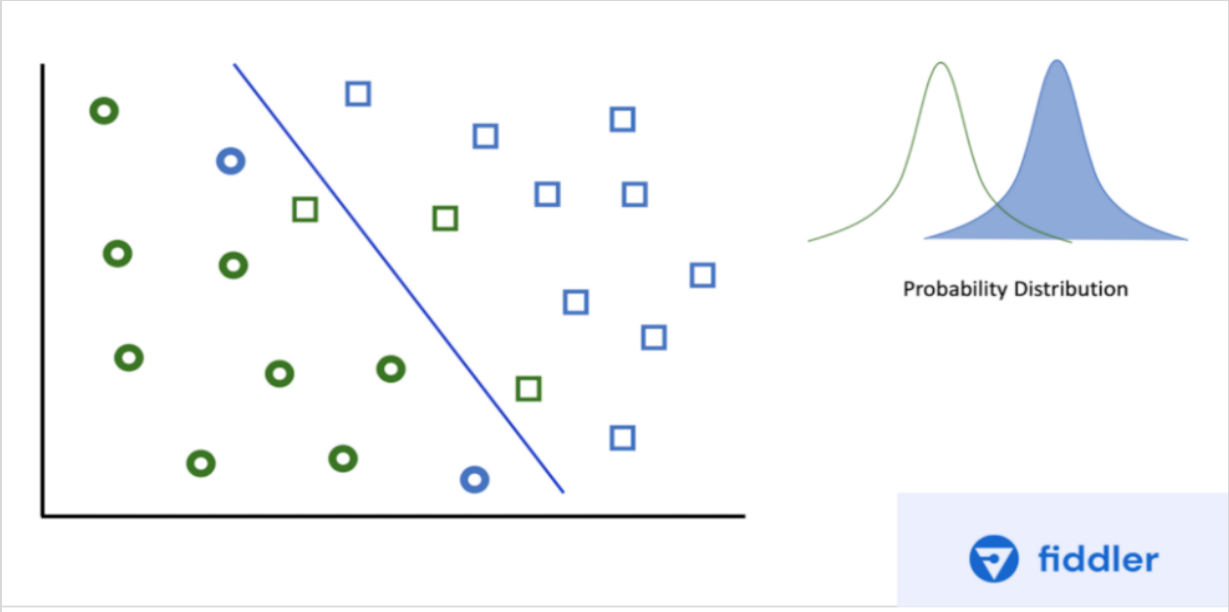
Social Bias:

- A human reviewer may introduce their own biases when accepting or disregarding the AI model's predictions
- **For example,** a human reviewer might override a correct prediction done by the AI model based on their social bias. "**I know since this defendant is black, he will commit another crime**"

Source: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

WHAT IS FAIRNESS?

- **Fairness:** absence of prejudice or preference for an individual or group based on their characteristics
- **For example,** in the figure, data included two different underlying groups→ green and blue
- represent different ethnicities, genders, or even geographical or temporal differences like morning and evening users
- In this case, using a single threshold for these two groups of diabetes patients would lead to poor health outcomes
- predictions are calibrated for each group properly



Source: https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

Fairness of error:

- similar performance (**accuracy**) across different groups, no group has more error than the other

Fairness of representation:

- model is equitable in its assignment of favorable outcomes to each group
- **Proportional parity**: metric→ requires that each group receives the favorable outcome the same proportion of the time
- **For example,** female and male candidates in automated resume review are each accepted 50% of the time

MEASURING AI BIAS

Source: https://www.datarobot.com/trusted-ai-101/ethics/bias-and-fairness/

MEASURING FAIRNESS

Predictive parity:

- Measures the raw accuracy score between groups
- Number of times the outcome is correctly identified

False positive and negative rate parity:

- measures the model's false positive rate and false negative rate for each class

Source: https://www.datarobot.com/trusted-ai-101/ethics/bias-and-fairness/

MITIGATING AI BIAS

AWARENESS & UNDERSTANDING OF DATA & AI ALGORITHM

IMPROVING DATA COLLECTION, RESAMPLING DATA USING GENERATIVE MODELS, REDUCING CLASS IMBALANCE PROBLEM

IMPROVING HUMAN CENTRIC DESIGN APPROACH

ESTABLISH A DEBIASING STRATEGY SUCH AS, EVALUATION METRICS, SUBGROUPS AND COMBINATIONS

MAINTAINING A DIVERSE AI TEAM, ENGAGE FACT-BASED CONVERSATIONS ABOUT POTENTIAL AI BIASES

MONITORING AND UPDATING YOUR MODEL DURING TRAINING & AFTER DEPLOYMENT

INVEST MORE IN BIAS RESEARCH, MULTI-DISCIPLINARY APPROACH

Minimizing AI bias is important for AI systems to flourish and increase people's trust in such systems

Source: https://research.aimultiple.com/ai-bias/

AUTOMATIC RESAMPLING USING GENERATIVE MODELS

- Use latent distribution to generate new data
- Reduce AI biases

Homogeneous skin color, pose

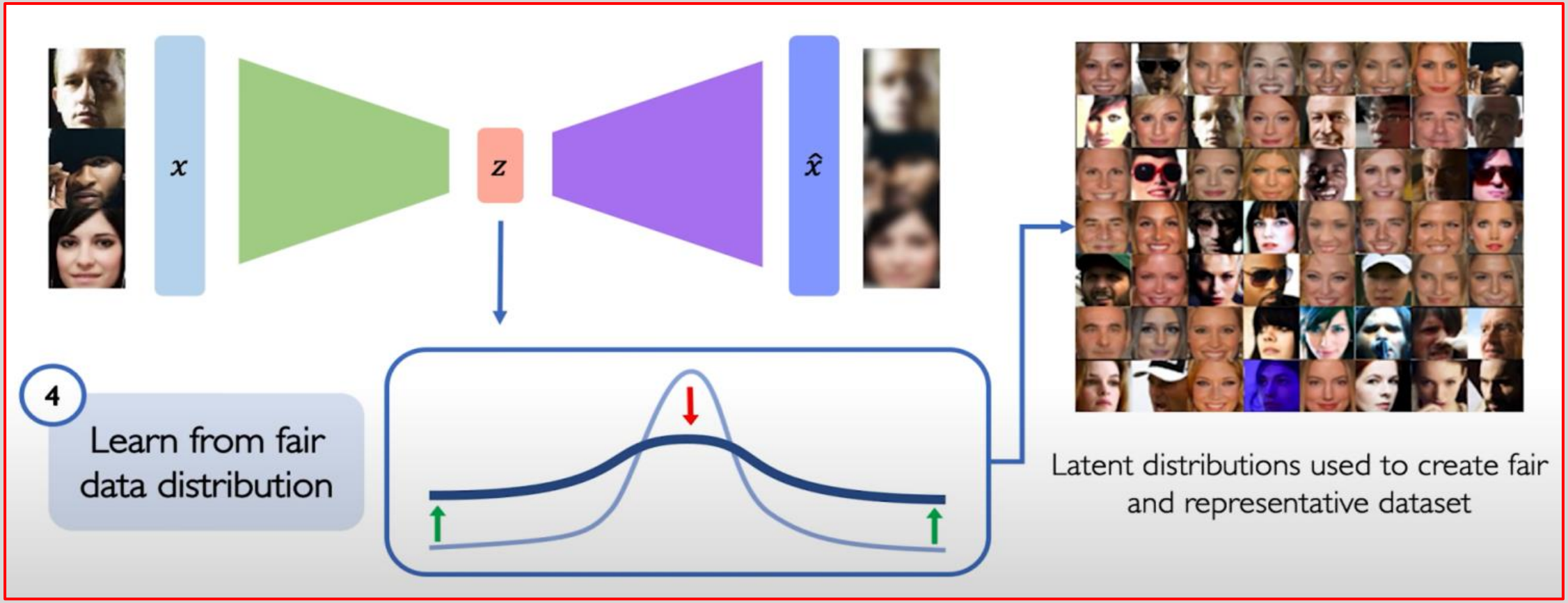
VS

Diverse skin color, pose, illumination

Source: introtodeeplearning.com

MITIGATING DATASET BIAS USING VAE

- Generate a fairer and more representative dataset for training



Source: introtodeeplearning.com

BALANCED DATASET FOR EVALUATION

- **Balanced Dataset:** equal or approximate number of samples from class A and B
- **Oversampling** and **Undersampling** techniques may be employed

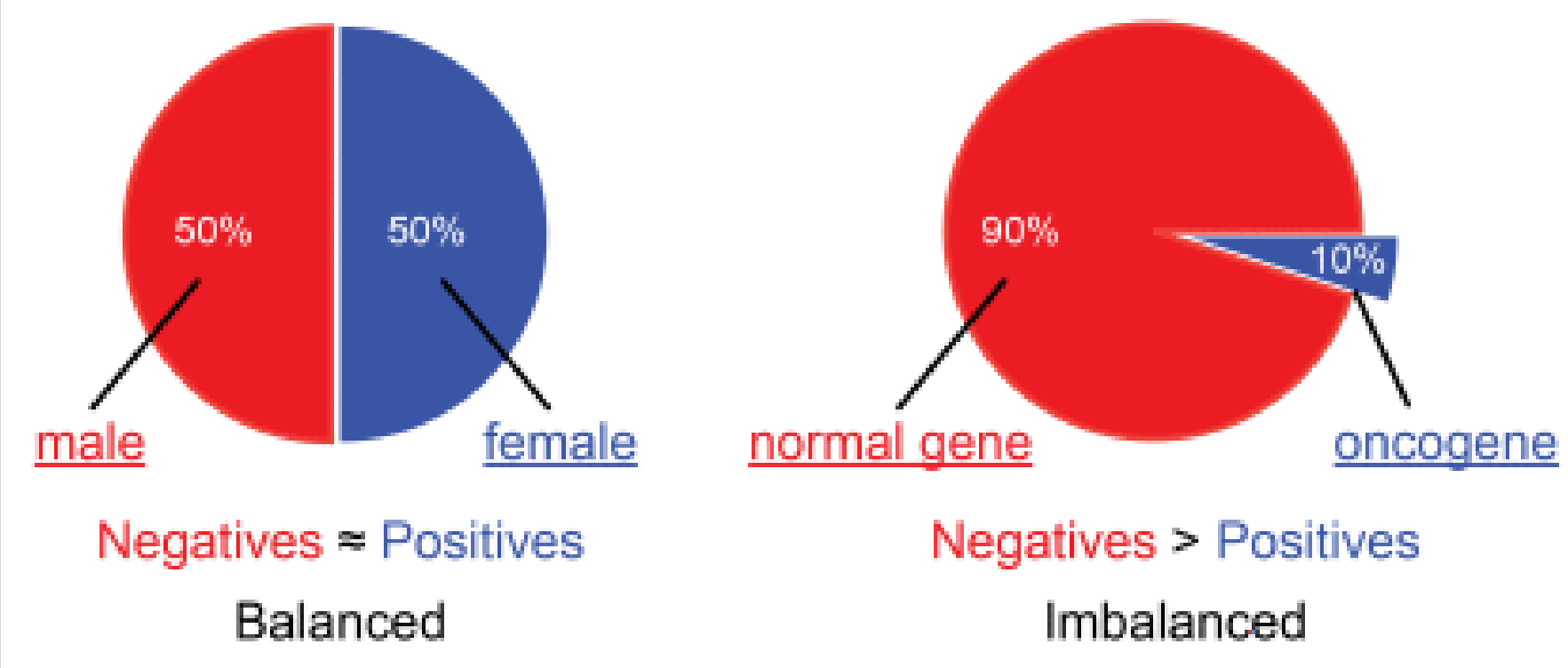
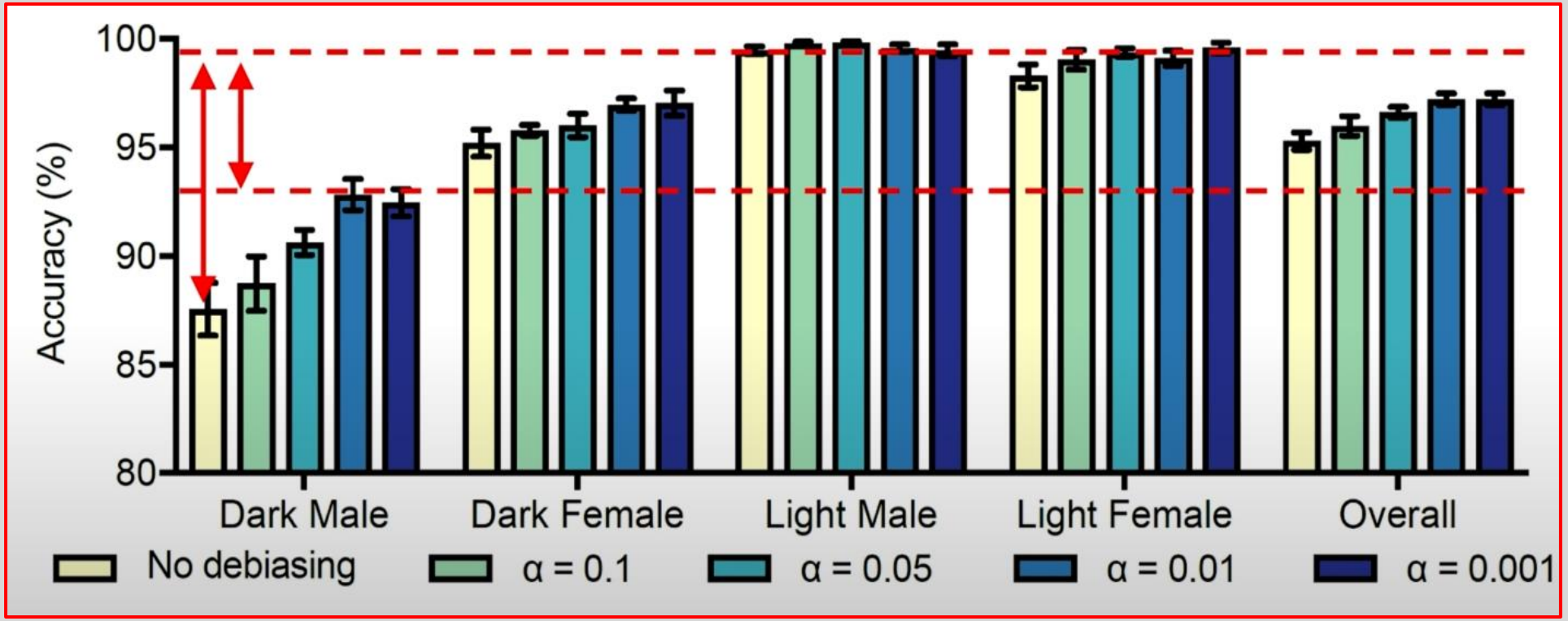


Figure Source: https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

EVALUATION TECHNIQUE TO REDUCE CATEGORICAL BIAS

- Decrease categorical bias
- Evaluate across subgroups and combinations of subgroups
- Disaggregated and intersectional evaluation



Source: introtodeeplearning.com

TOOLS TO REDUCE AI BIAS & DEVELOP A FAIR AI

AI Fairness 360:
- IBM
- Open source
- detect and mitigate biases in unsupervised learning
- test biases in models and datasets with a comprehensive set of metrics
- binary classification problems only

IBM Watson OpenScale:
- IBM
- bias checking and mitigation in real time
- when AI is making its decisions.

Google's What-If Tool:
- can test performance in hypothetical situations
- analyze the importance of different data features
- visualize model behavior across multiple models and subsets of input data
- for different ML fairness metrics

FATE:
- Fairness, Accountability, Transparency, and Ethics in AI (FATE)
- Microsoft
- assess visualization dashboards and bias mitigation algorithms
- compare trade-offs between fairness and performance of the system

Source: https://research.aimultiple.com/ai-bias/

THANKS!

Do you have any questions?