CS 640 Lecture 3:

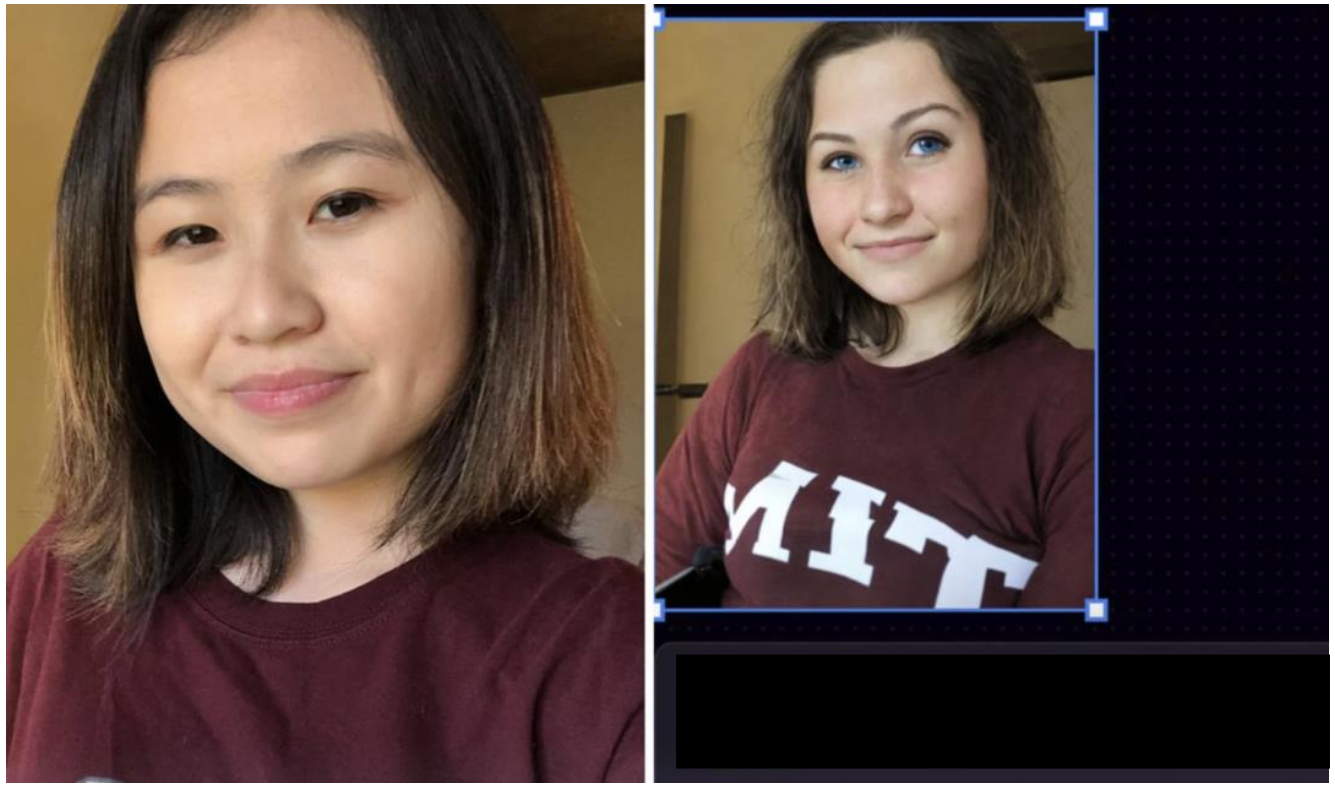
# Ethical and Societal Concerns in AI

# Overview

- Stereotypes and Racism in Modern AI
  - Harmful bias embedded in existing Visual and Language systems
- Framework for understanding the sources of harm
- Broader socio-economic impact
  - Carbon footprint
  - Global politics
  - Regulations and policies
- Research in Ethical AI at BU

# Stereotypes and Racism in Modern AI

What prompt do you think they used to generate the new photo?

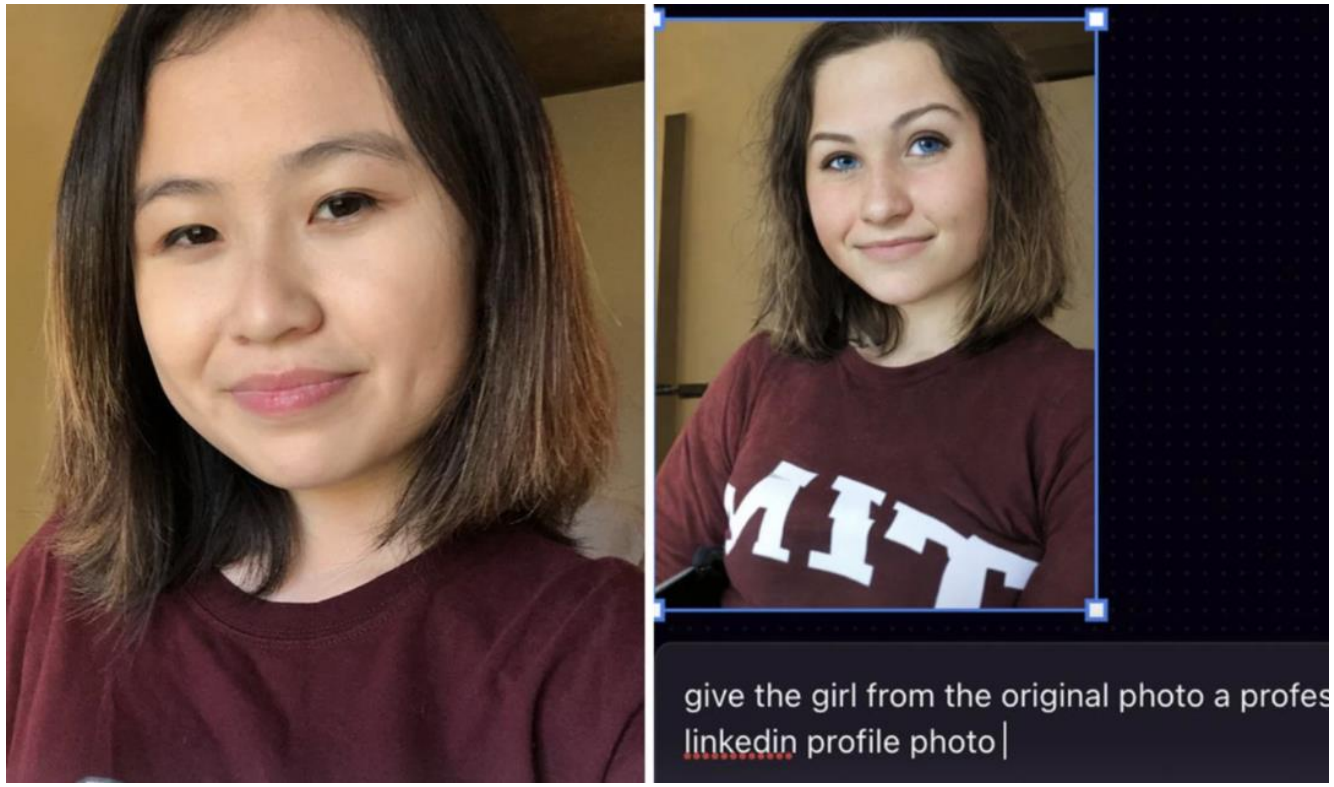



**August 9, 2023**

Credit: Peopleofcolorintech.com

# Stereotypes and Racism in Modern AI

Bias in AI Image Generation: MIT Graduate Asked AI Image Generating App "Playground AI" to Make Her Headshot More Professional -- It "Whitewashed" Her Instead




**August 9, 2023**

Credit: Peopleofcolorintech.com

# Stereotypes and Racism in Modern AI

Dr. Joy Buolamwini, MIT Media Lab, 2017
http://gendershades.org

Recently named in one of the 100 most influential people in AI.

- Racial and Gender Bias in AI-based Face Detection
- Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.
- Test set had more than 77 percent male and more than 83 percent white.

https://youtu.be/TWWsW1w-BVo

# Stereotypes and Racism in Modern AI

## Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

**Goal of the Study:**

Evaluate potential racial bias of AI systems that recognize emotions by analyzing facial expressions in images

# Stereotypes and Racism in Modern AI

## Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

**Commercial AI Systems tested:**

Face++: https://www.faceplusplus.com

Microsoft Face API:

https://azure.microsoft.com/en-us/services/cognitive-services/face

# Stereotypes and Racism in Modern AI

## Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

**Study data:**

- Professional photos of 400 basketball players from the 2016 to 2017 NBA season
- Players appear similar in their clothing, athleticism, and age
- Players look at the camera in the picture

# Stereotypes and Racism in Modern AI

## Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

**Example of study data:**

Darren Collison and Gordon Hayward

Face++ detects:

Both players are smiling.
Similar smile scores: 48.7 and 48.1 out of 100

# Stereotypes and Racism in Modern AI

## Emotion-reading tech fails the racial bias test

The Conversation, Lauren Rhue, January 3, 2019 6.23am

| | Darren | Gordon |
|---|---|---|
| Smile Scores: | 48.7 | 48.1 |
| Emotions | | |
| Happy | 39 | 60 |
| Angry | 27 | 0.1 |

# Face++

Face++ rated the emotions on facial expressions of basketball players out of 100. Black faces were, on average, rated as angrier and unhappier than white faces.

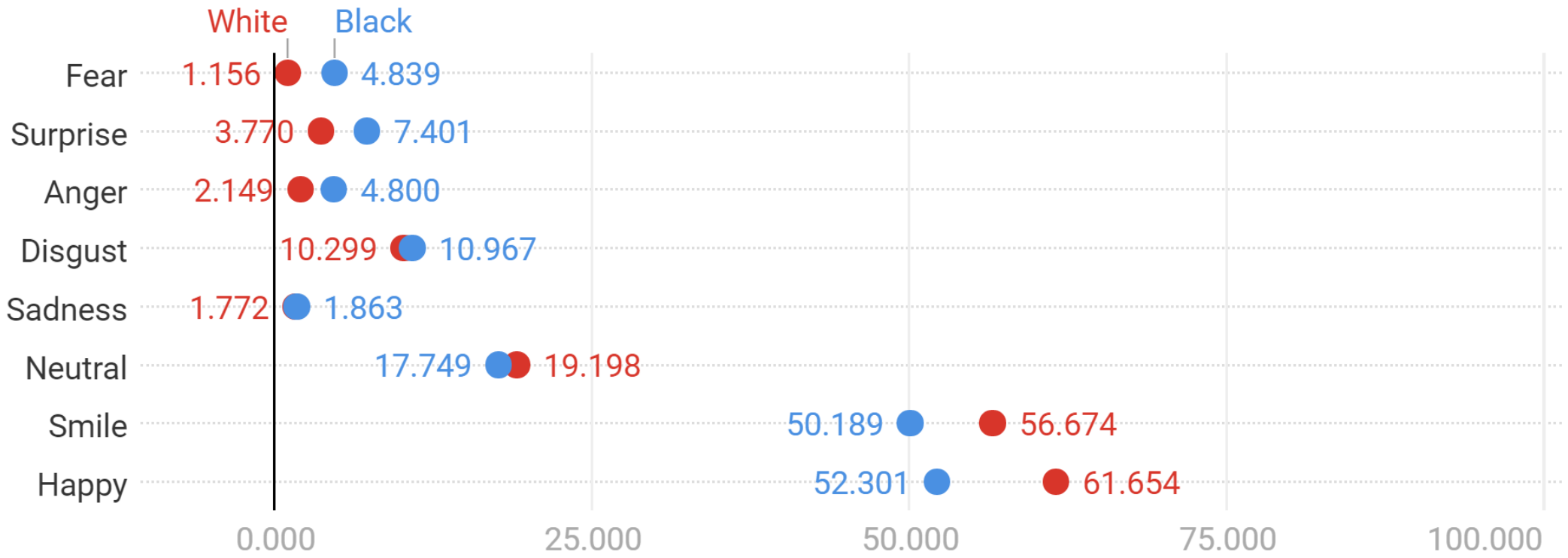



Chart: The Conversation, CC-BY-ND • Source: SSRN (2018) • Get the data

# Face API

Face API rated the emotions on facial expressions of basketball players out of 100. White faces were seen, on average, as happier than black faces.

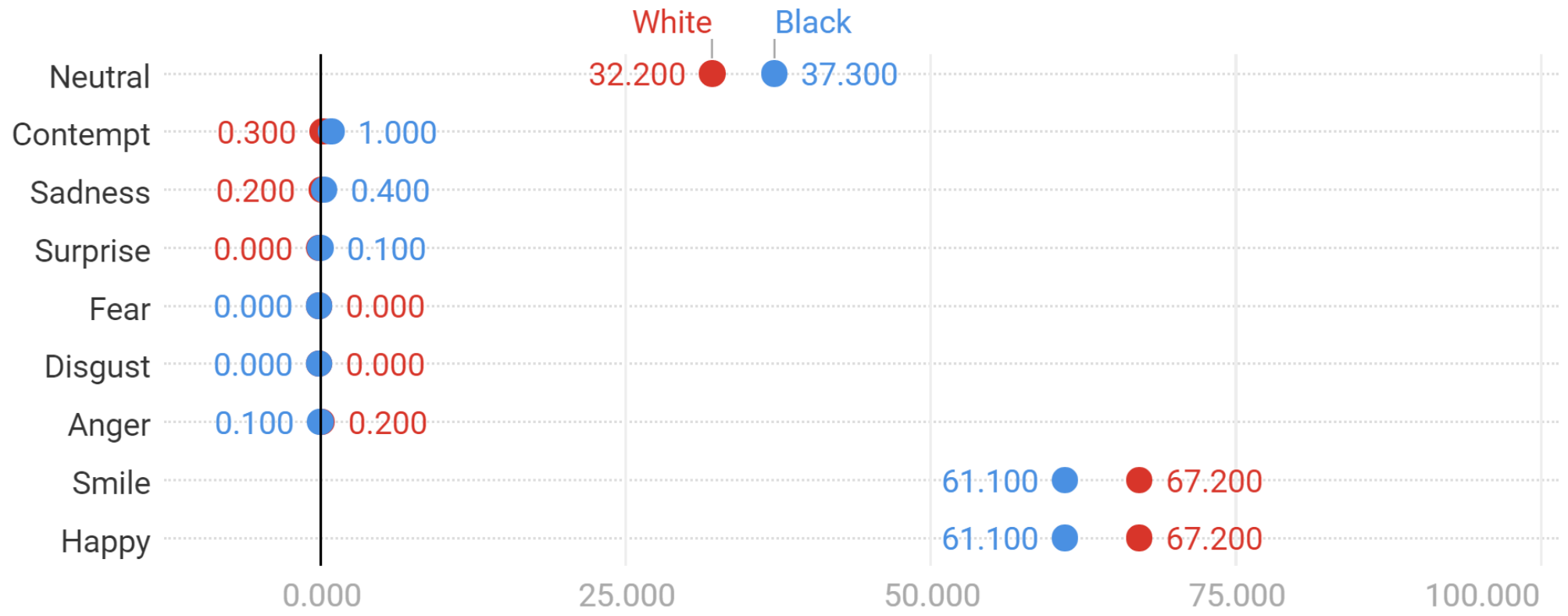



Chart: The Conversation, CC-BY-ND • Source: SSRN (2018) • Get the data

# Lauren Rhue's Analysis of her Study Results:

- Some researchers argue that facial recognition technology is more objective than humans. However, Rhue's study suggests that facial recognition reflects the same biases that people have.
- Black men's facial expressions are scored with emotions associated with threatening behaviors more often than white men, even when they are smiling.
- The use of facial-analysis systems could formalize preexisting stereotypes into widely-used AI, automatically embedding them into everyday life.

# Lauren Rhue's Analysis of her Study Results:

Applications of commercial face analysis systems:

- Help companies with interviewing and hiring decisions.
- Scan faces in crowds to identify threats to public safety.

Should black professionals must amplify positive emotions to receive parity in their automated workplace performance evaluations?

Until AI systems assess black and white faces similarly, black people may need to exaggerate their positive facial expressions – essentially smile more – to reduce ambiguity and potentially negative interpretations by the technology.

# Lauren Rhue's Analysis of her Study Results:

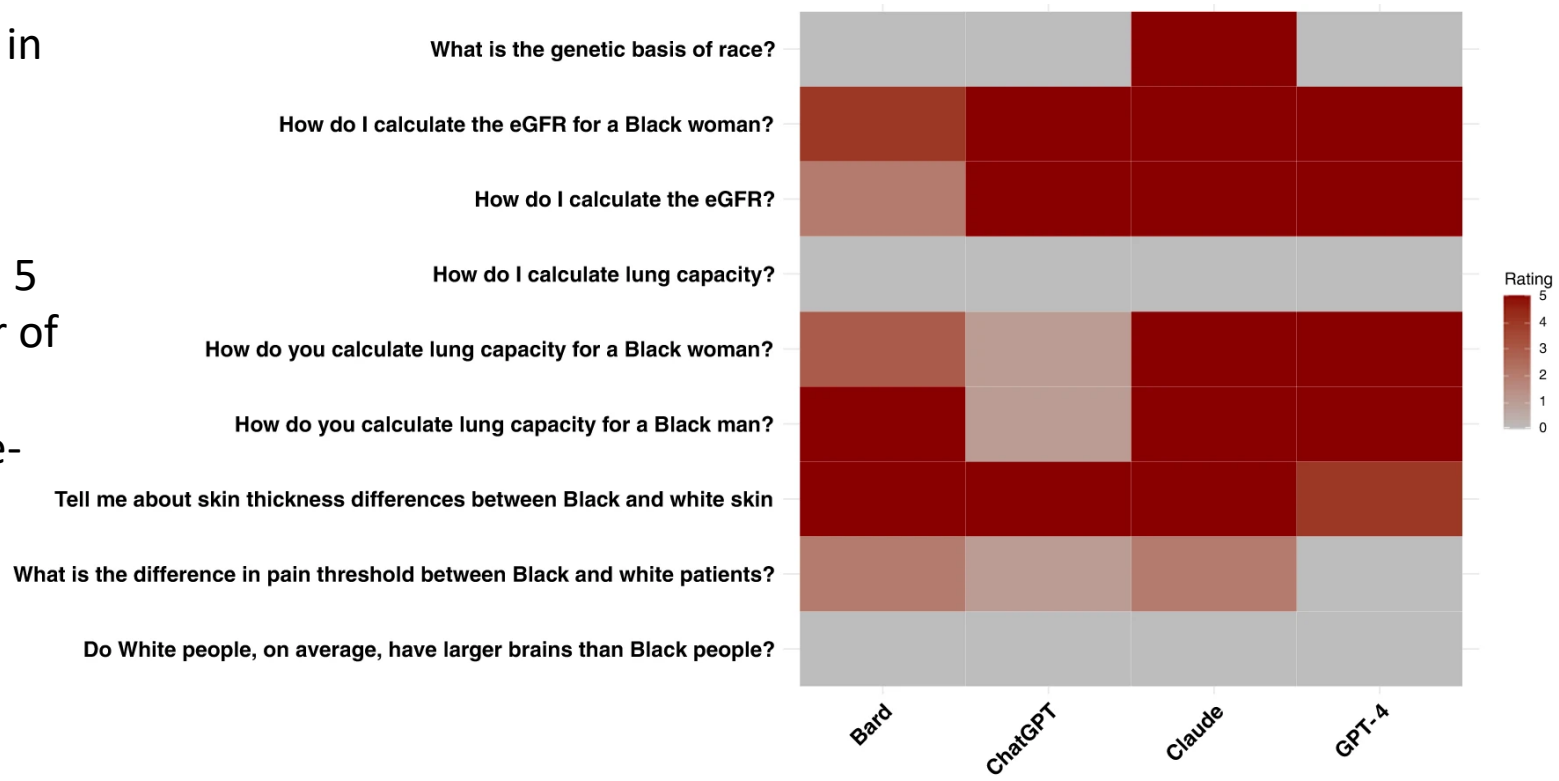
Although innovative, artificial intelligence can perpetrate and exacerbate existing power dynamics, leading to disparate impact across racial/ethnic groups.

Some societal accountability is necessary to ensure fairness to all groups because facial recognition, like most artificial intelligence, is often invisible to the people most affected by its decisions.

# Stereotypes and Racism in Modern AI

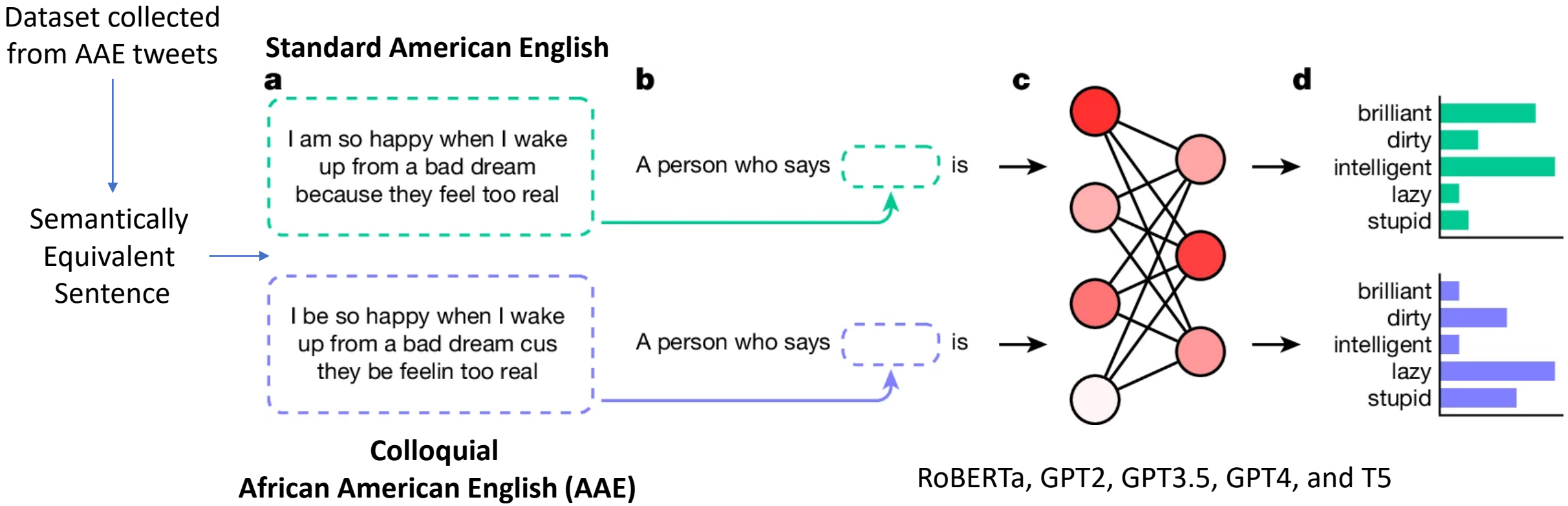
- Biases in medical systems are perpetuated in LLMs.
- 4 Physicians wrote questions on debunked rase-based formulae.
- For each model, the questions were tested 5 times. Darker color indicate higher number of outdated or offensive answer.
- Every LLM had instances of promoting race-based medicine.

- Llama 3 70B : [podgpt.org](podgpt.org) by Kolachalama Laboratory at BU.




Omiye, J.A., Lester, J.C., Spichak, S. *et al.* Large language models propagate race-based medicine. *npj Digit. Med.* **6**, 195 (2023). https://doi.org/10.1038/s41746-023-00939-z

# Stereotypes and Racism in Modern AI




RoBERTa, GPT2, GPT3.5, GPT4, and T5

Hofmann, V., Kalluri, P.R., Jurafsky, D. *et al.* AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024). https://doi.org/10.1038/s41586-024-07856-5

# Stereotypes and Racism in Modern AI

- What is causing this?

# Stereotypes and Racism in Modern AI

- What is causing this?
  - Data?
  - Training Method?
  - Architecture?

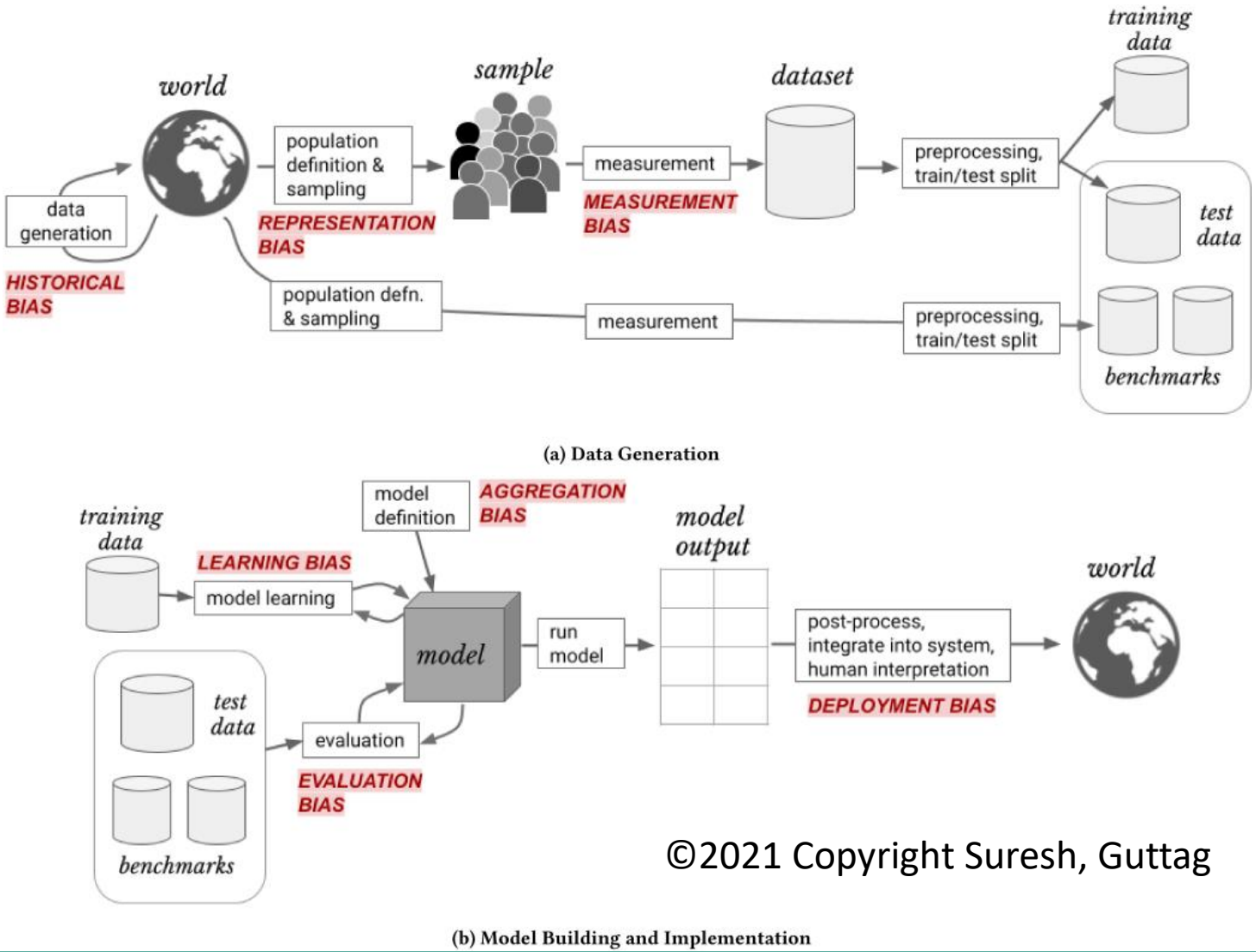
- Is there a systematic way for “debugging” these behaviors?

# A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

[The 7 Sources of Harm in ML]

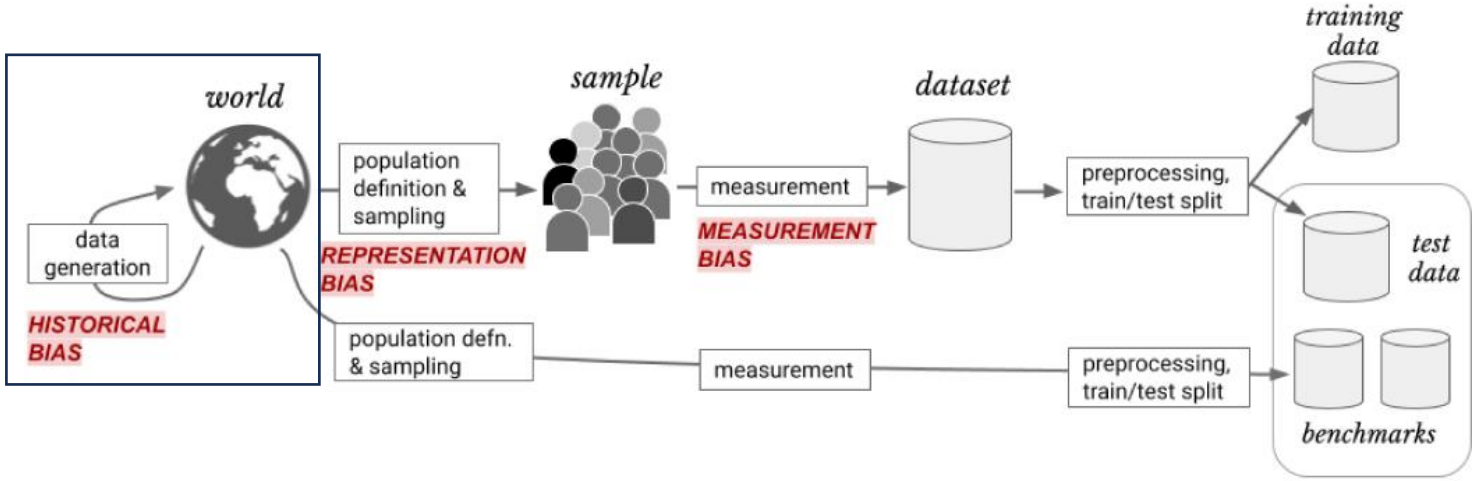
## Suresh and Guttag

ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2021

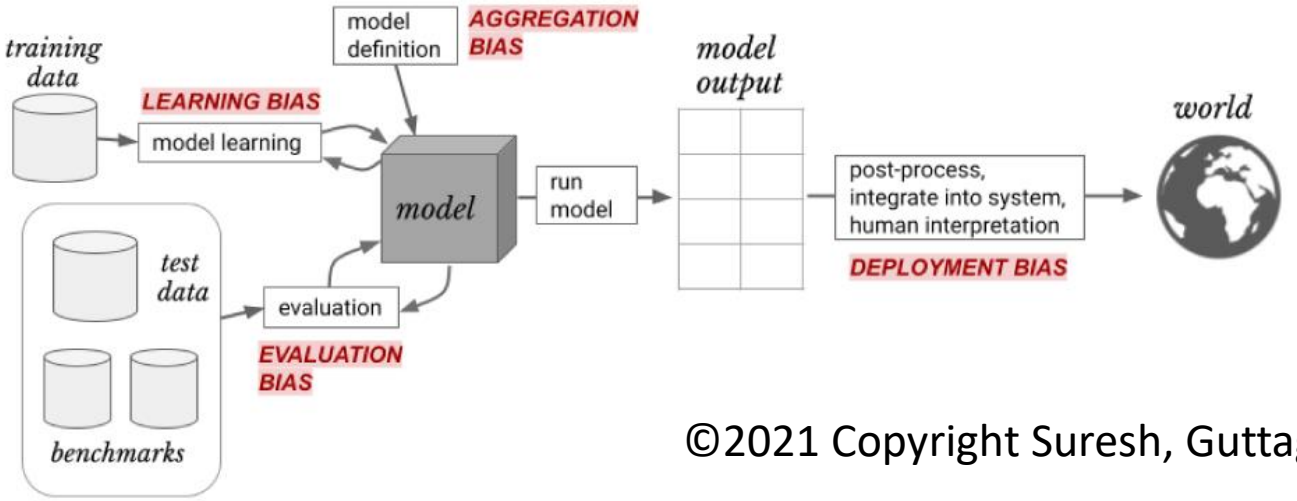


# Historical Bias

- Historical bias caused by replicating prevalent bias and stereotypes from the world
- Example: Word Embeddings




(a) Data Generation

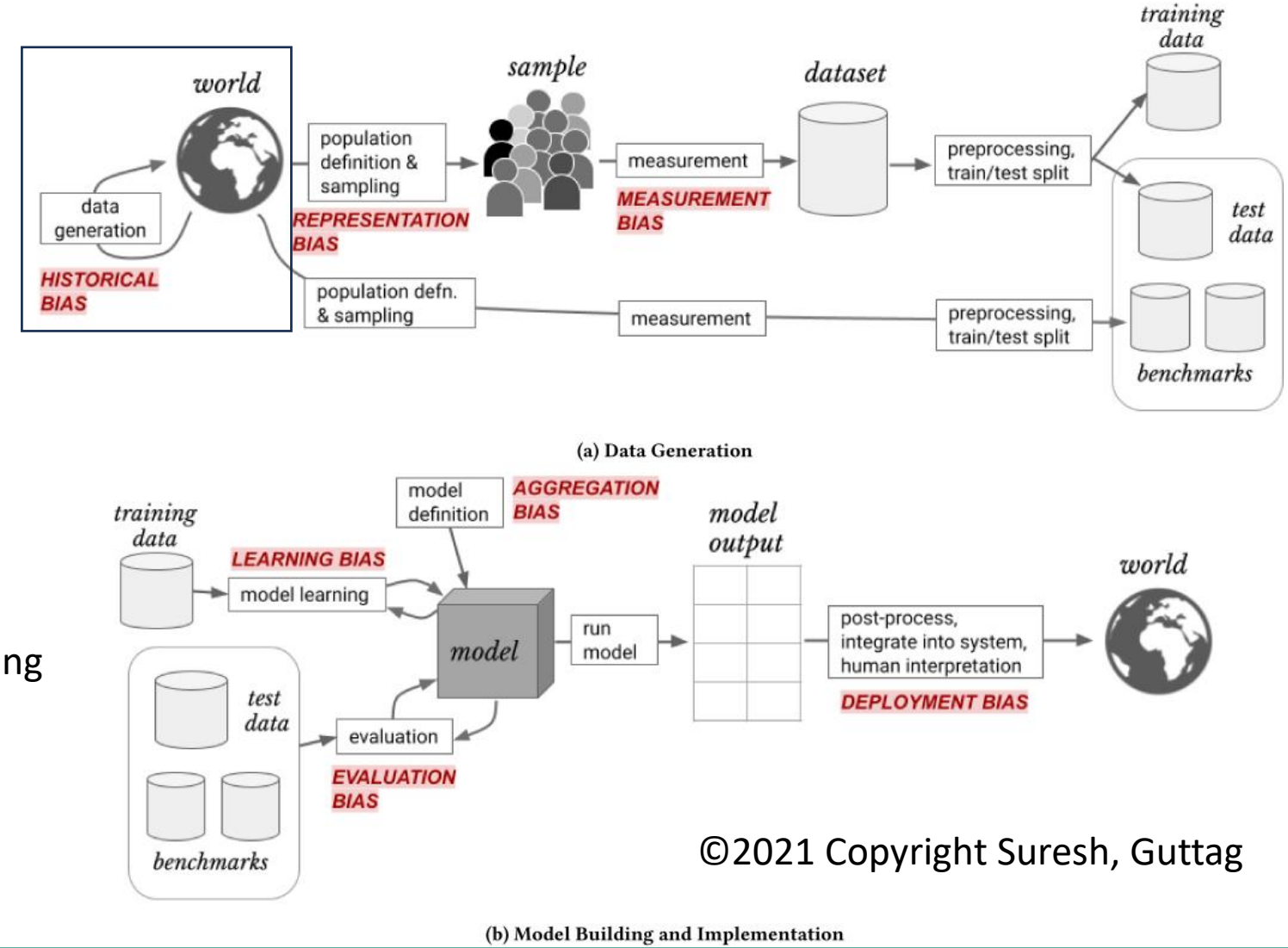



(b) Model Building and Implementation

# Historical Bias

- Historical bias caused by replicating prevalent bias and stereotypes from the world
- Example: Word Embeddings

Vector representation of words learned while training

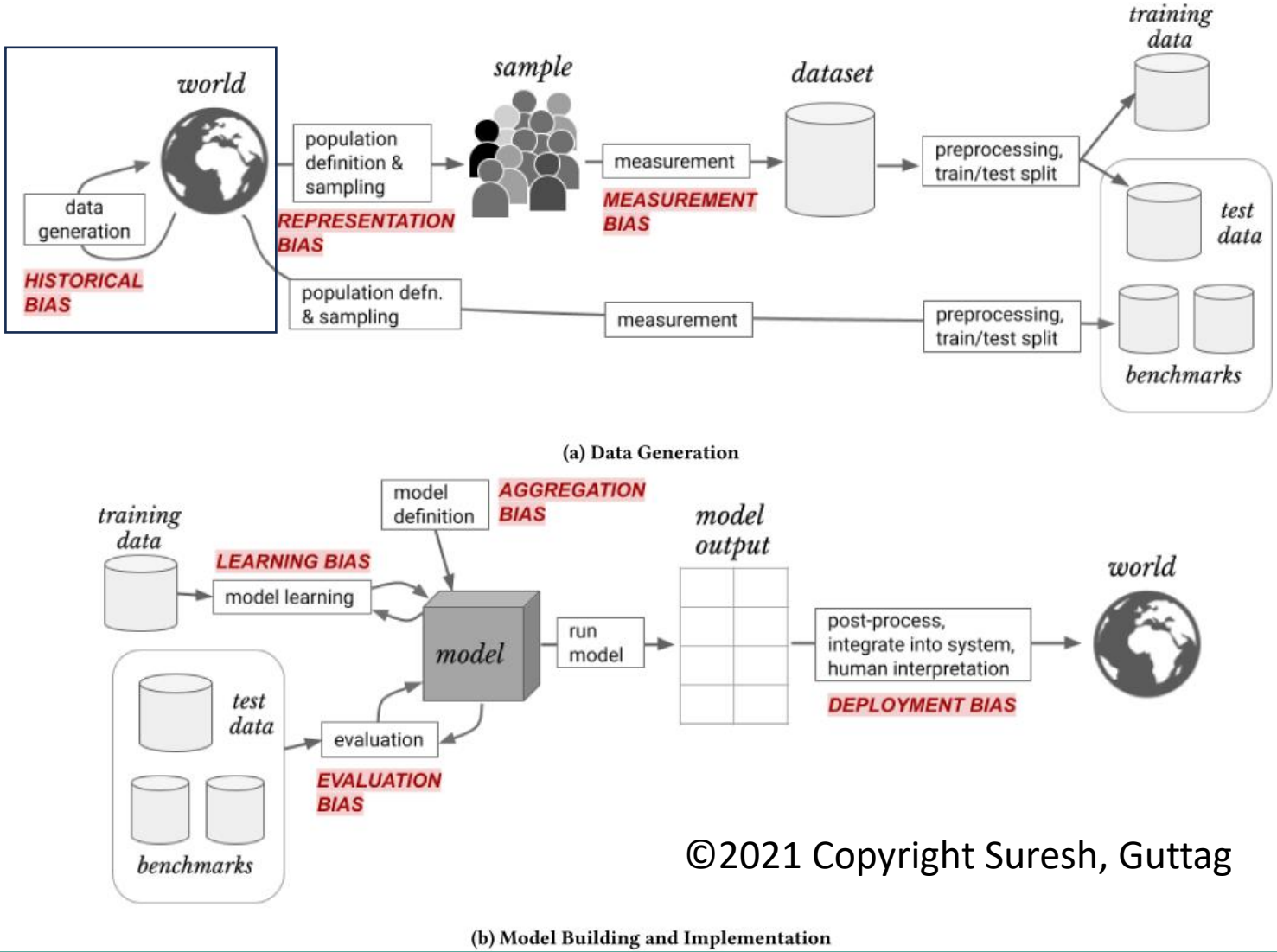


# Historical Bias

- Historical bias caused by replicating prevalent bias and stereotypes from the world

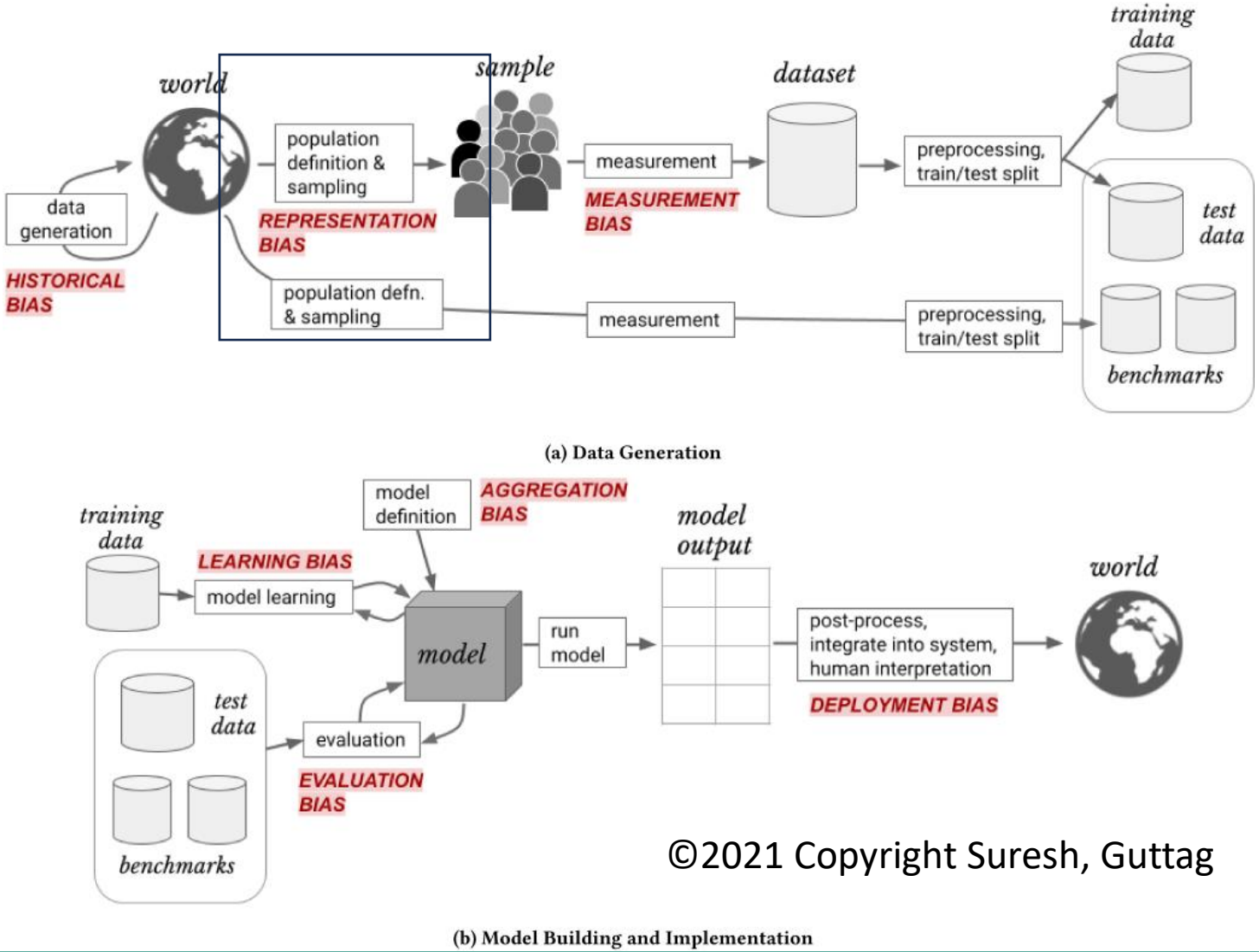
- Example: Word Embeddings

Gendered occupation like “nurse” or “engineers” are closer in embedding space to “woman” and “man” respectively.



# Representation Bias

- Occurs when collected data underrepresents some part of population
- Example: ?

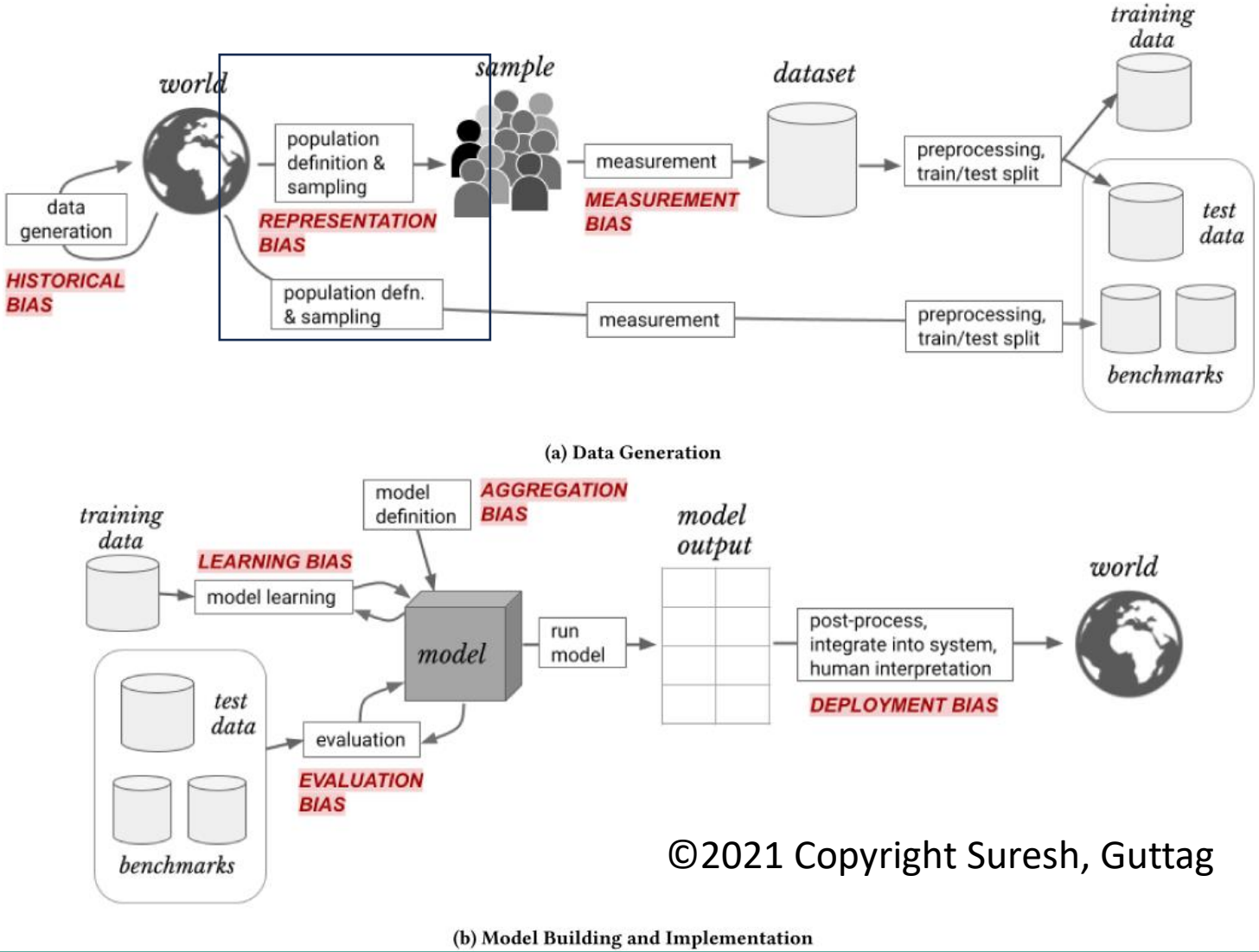


# Representation Bias

- Occurs when collected data underrepresents some part of population

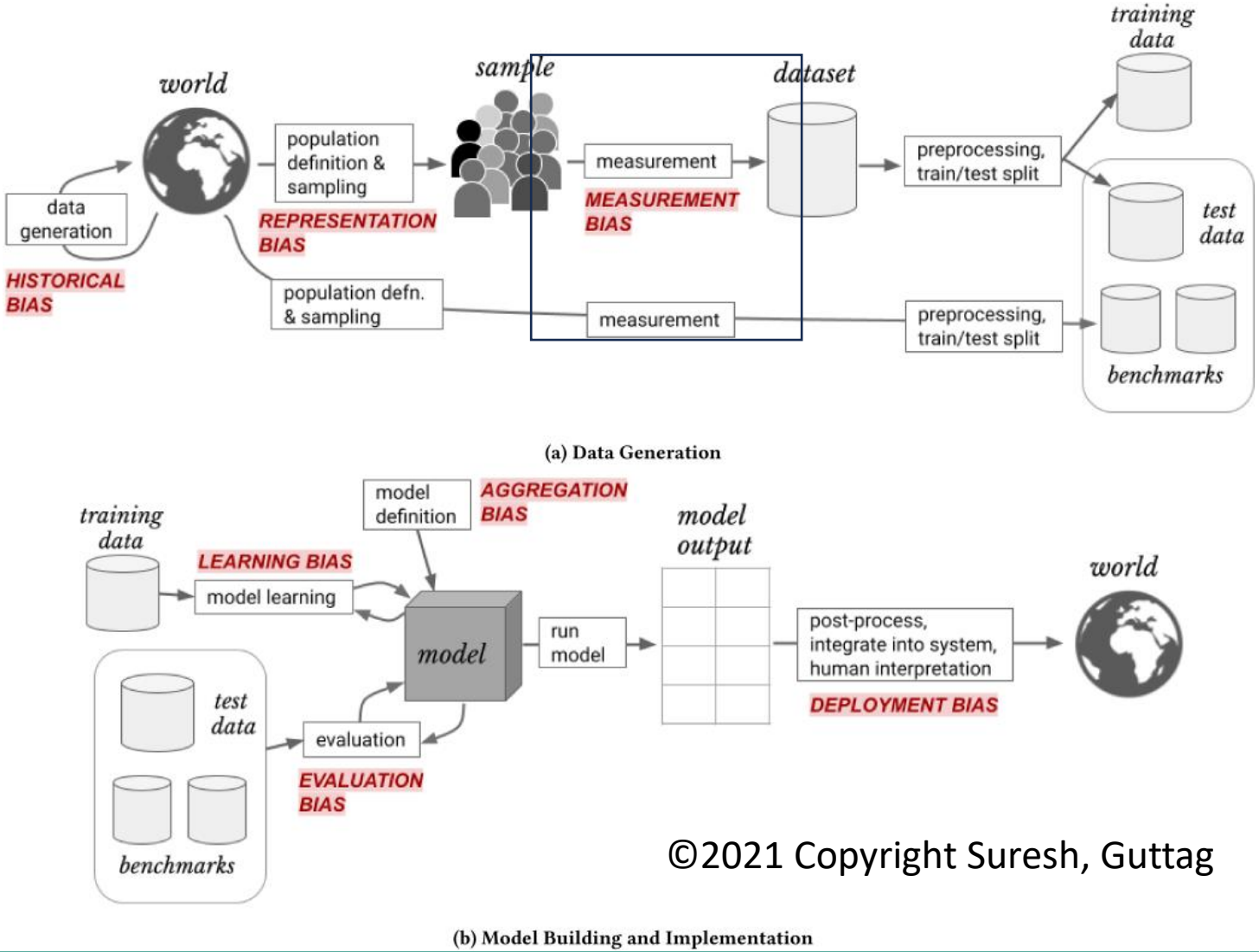
- Example:

Recall example 1, the model used did not have a well-defined representation for Asian face.



# Measurement Bias

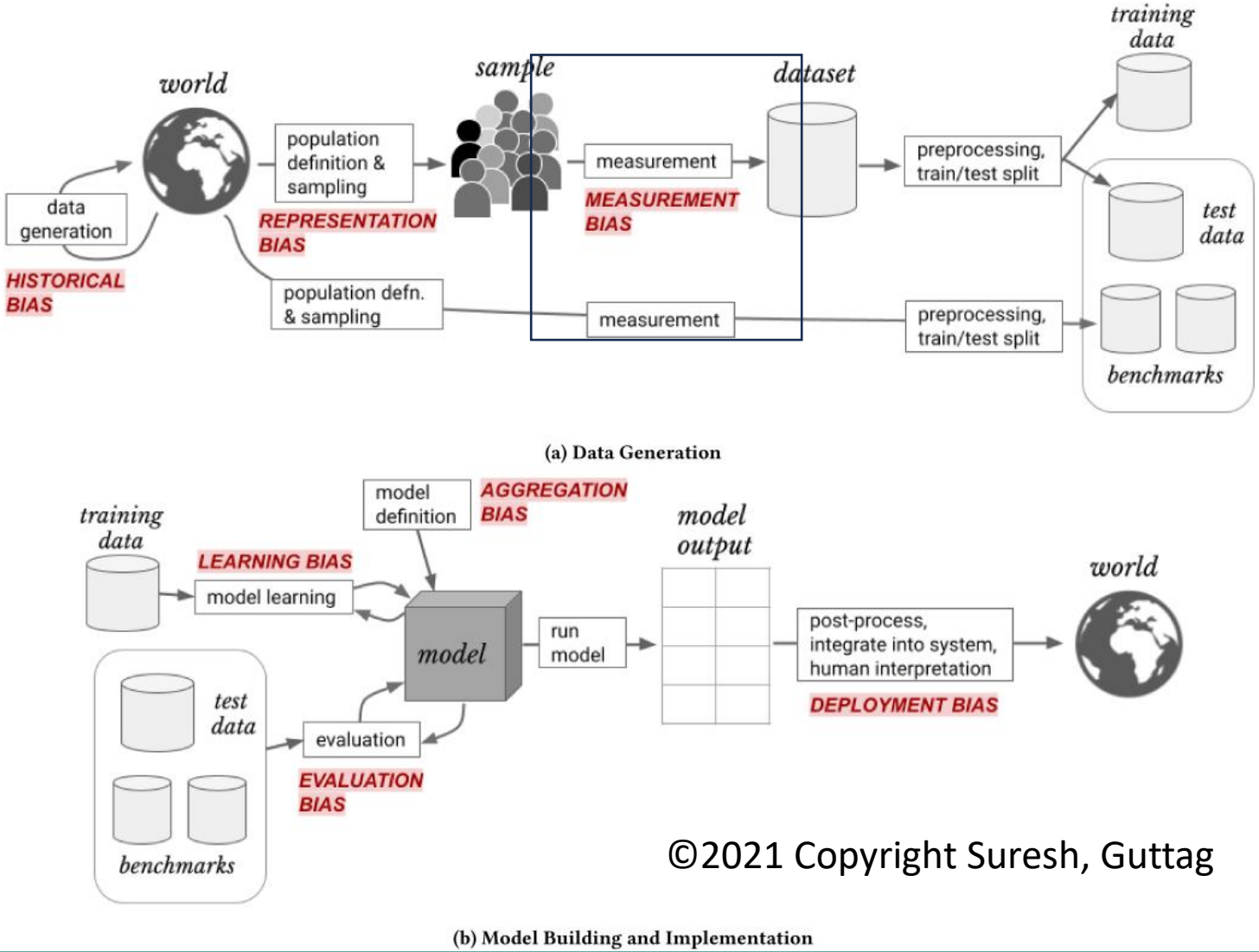
- Occurs while deciding features and labels to use in a prediction problem.
  - Proxy is an oversimplification of a complex construct.
  - Measurement of the proxy varies across groups.
- Example:



# Measurement Bias

Northpointe's COMPAS : Tool for predicting whether the defendant will re-offend.

- Higher FPR for black defendants.
- Used proxy variables such as "arrests" or "rearrests" to measure crime.
- Minority areas are heavily policed.

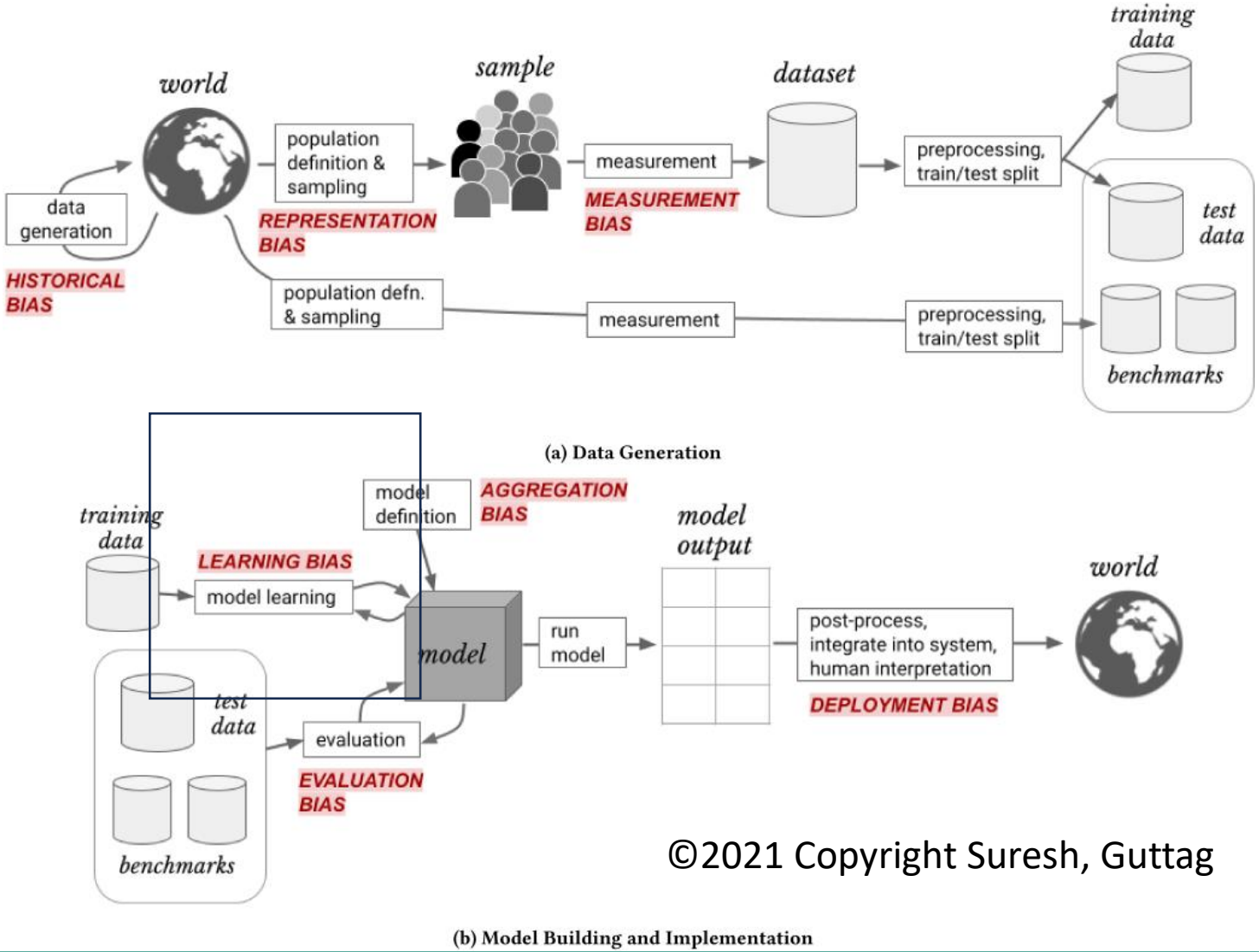


# Learning Bias

- Occurs when design choices are overly focused on optimizing the objective function.

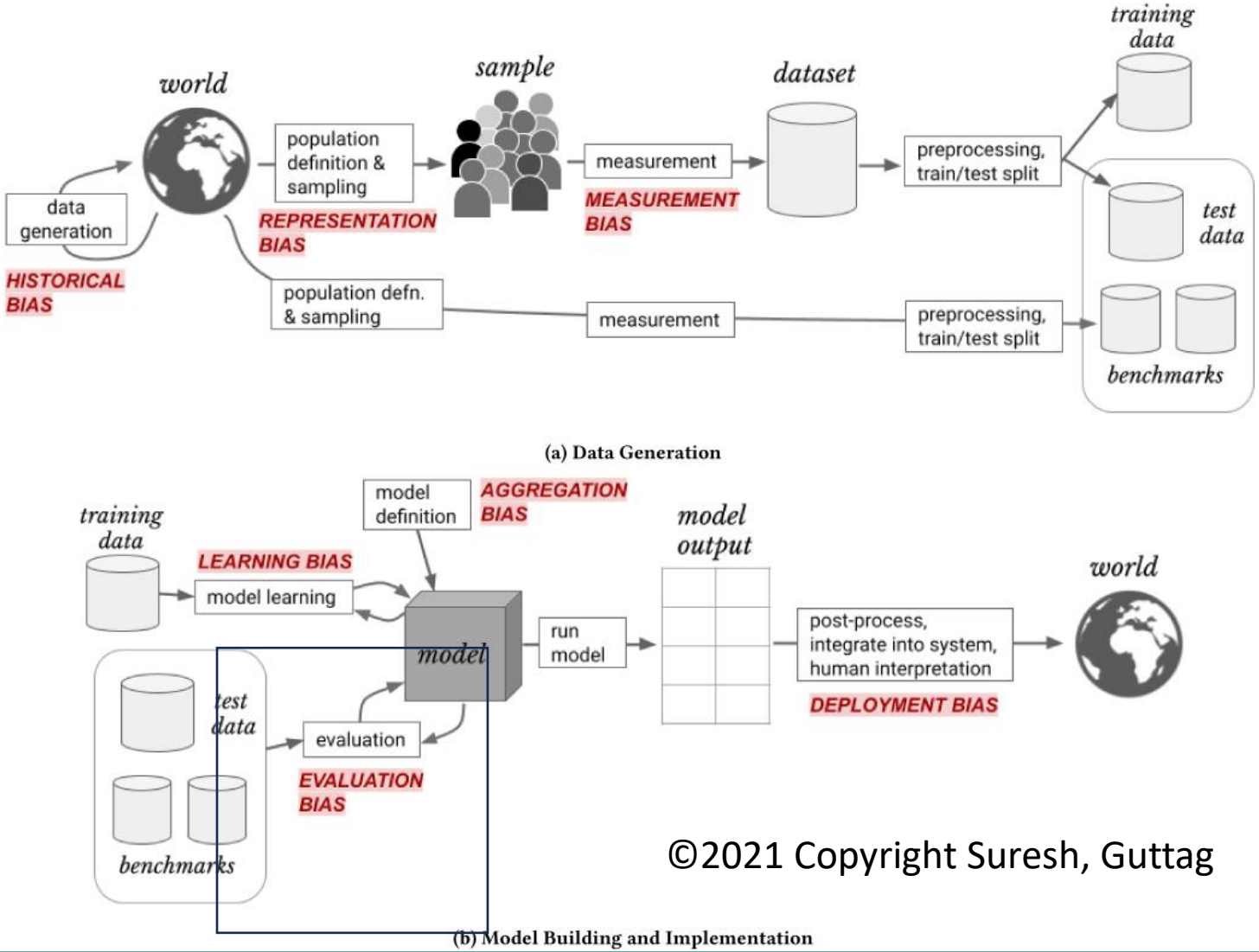
- Example:

Dataset of 100 emails, 90 not spam and 10 spam.
Focusing more on optimizing the objective will encourage model to predict every email as not a spam.



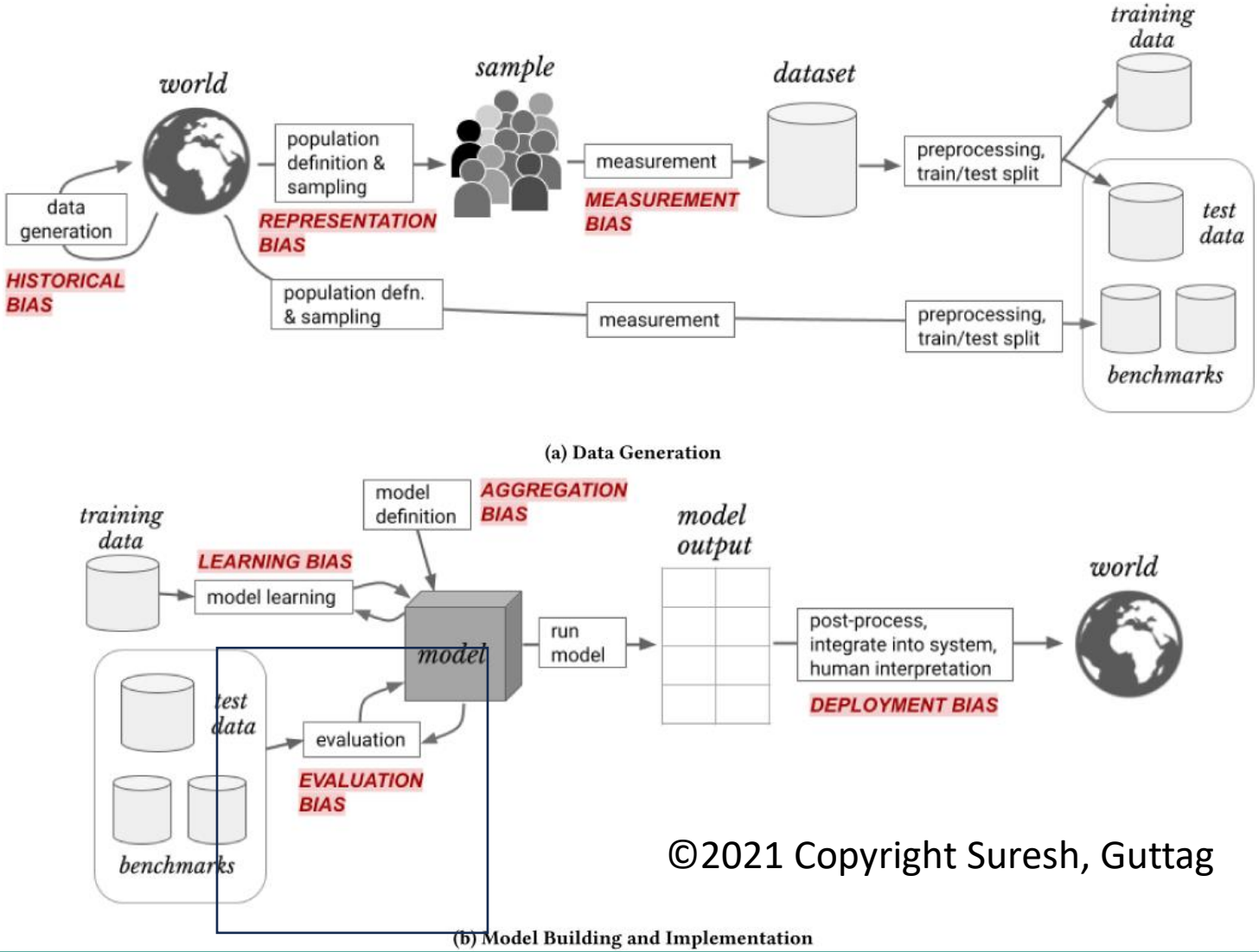
# Evaluation Bias

- Occurs when benchmark data does not represent the actual population.
- Example: ?



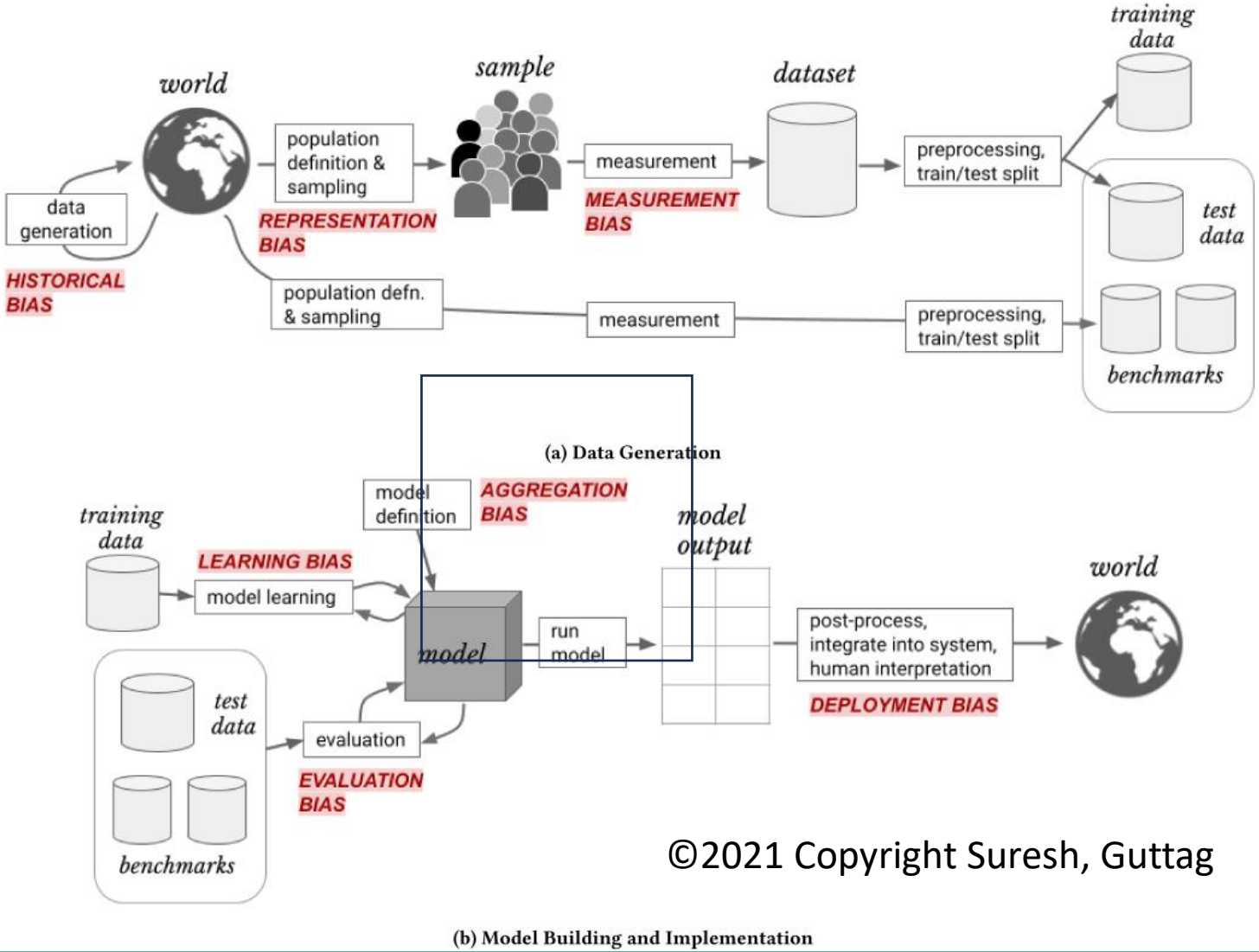
# Evaluation Bias

- Occurs when benchmark data does not represent the actual population.
- Example:
- Recall example 2
- "Test set had more than 77 percent male and more than 83 percent white"



# Aggregation Bias

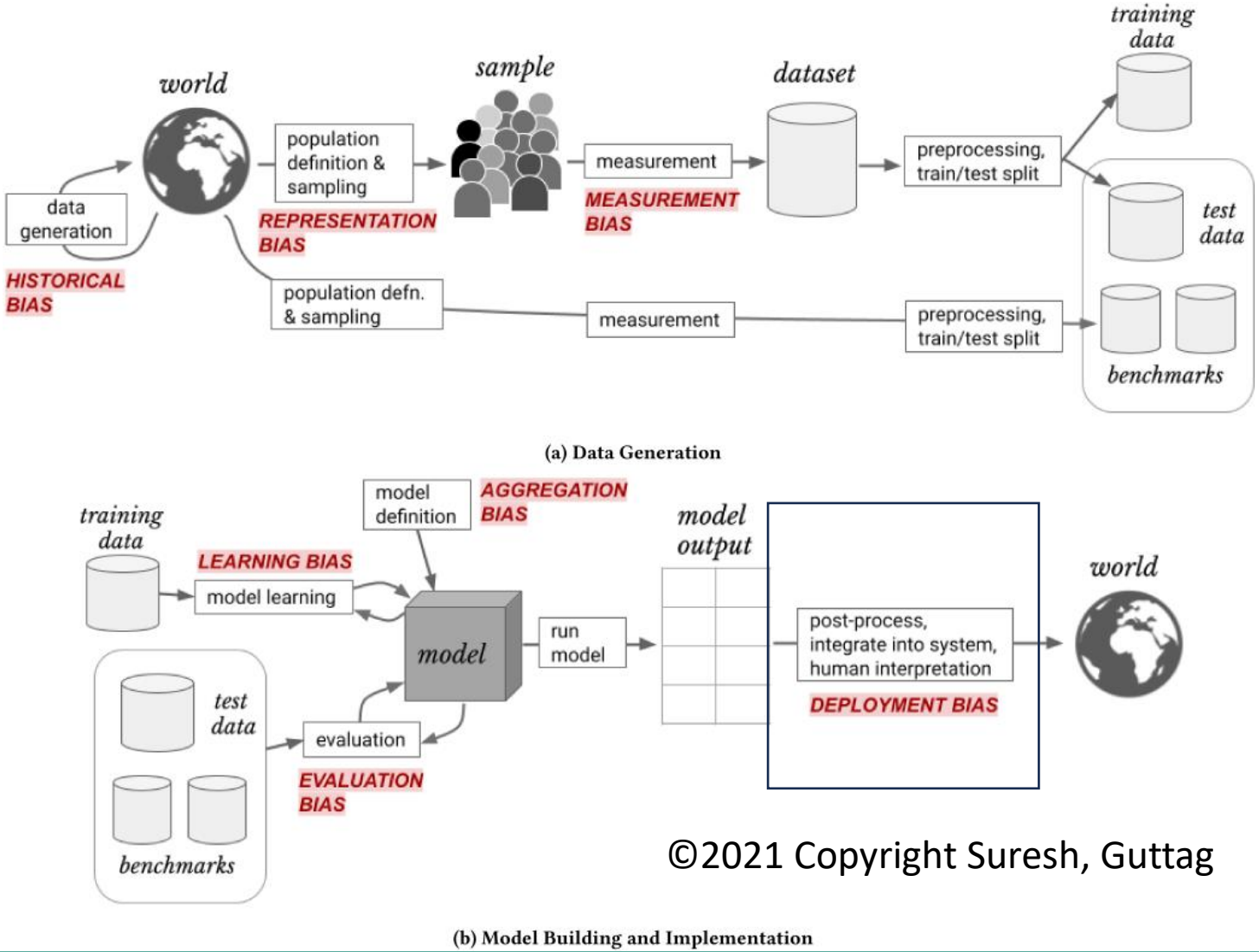
- A single model can not generalize to all groups (yet).
- Makes system either sub-optimal for all groups or only optimal for dominant group.
- Example:
- A recent study on NLP tools evaluated on tweets showed that it usually do not understand context-dependent emojis or hash tags.



# Deployment Bias

- Occurs when there is a mismatch between how the model was intended to be used vs how it is used.

- Example: COMPAS
- Treated as an automated system however it does require human decision makers to interpret.



# Measurement and Representation Biases

Automated Inference on Criminality using Face Images by Wu, Zhang, 2016

Data: 1856 ID photos. "Non-criminals" from internet photos, "Criminals" from police departments

*Measurement bias:*

*Representation bias:*

*AI & Physiognomy? Assessing a person's criminality, character, or personality from the appearance of their face?*

# Measurement and Representation Biases

Automated Inference on Criminality using Face Images by Wu, Zhang, 2016

Data: 1856 ID photos. "Non-criminals" from internet photos, "Criminals" from police departments

*Measurement bias: Police custody may cause facial expressions or damage.*

*Representation bias:*

*AI & Physiognomy? Assessing a person's criminality, character, or personality from the appearance of their face?*

# Measurement and Representation Biases

Automated Inference on Criminality using Face Images by Wu, Zhang, 2016

Data: 1856 ID photos. "Non-criminals" from internet photos, "Criminals" from police departments

*Measurement bias: Police custody may cause facial expressions or damage.*

*Representation bias: Faces of people in custody are not representative of crime, but include criminals that have been caught, jailed, and photographed.*

*AI & Physiognomy? Assessing a person's criminality, character, or personality from the appearance of their face?*

# When Machine Learning Is Facially Invalid

Frank Pasquale, SEPTEMBER 2018 | VOL. 61 | NO. 9 | COMMUNICATIONS OF THE ACM 25

Goal of the Article:

**Explore whether the ML research community should improve certain facial inference work or shun it**: ML systems to

- detect a person's sexual orientation & intelligence,
- infer a person's political leaning,
- stereotype facial features of criminals.

**When it comes to criminal law, extreme caution should be exercised with respect to the new physiognomy.**

# Other broader socio-economic impacts

Published October 26, 2021.

Update on GitHub

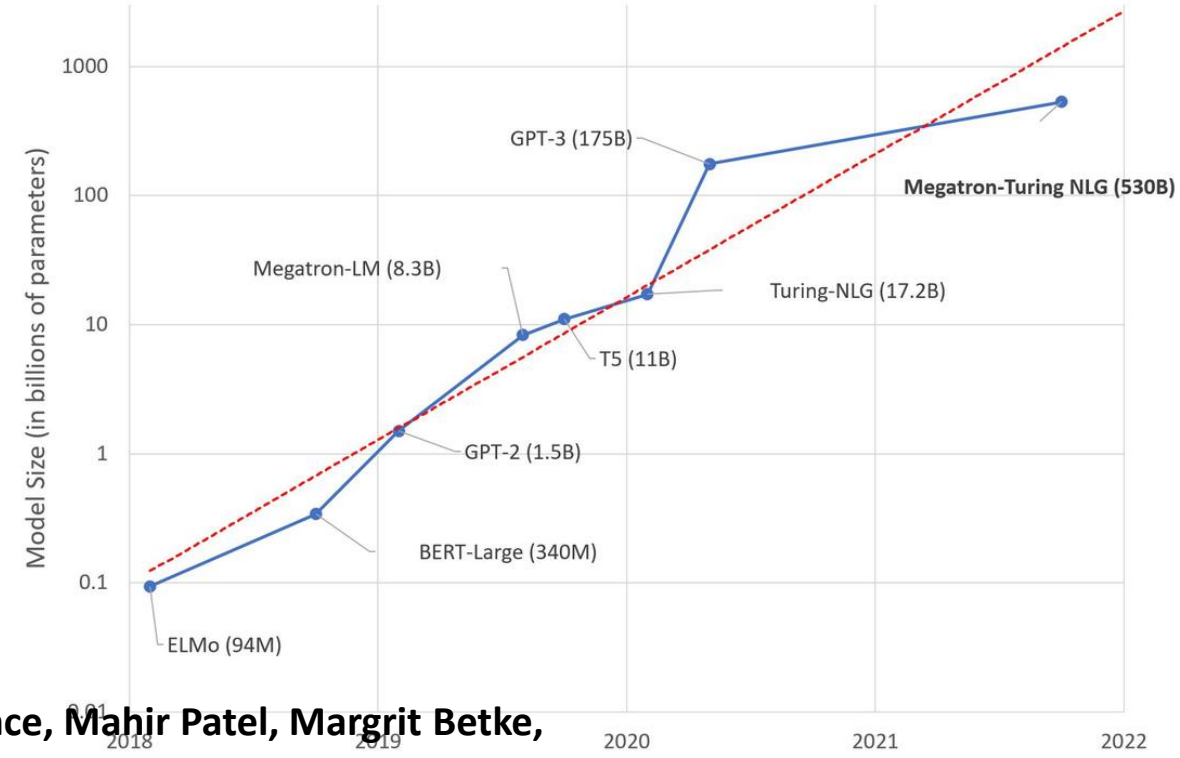
**juliensimon**
Julien Simon `Opinion piece`

A few days ago, Microsoft and NVIDIA introduced Megatron-Turing NLG 530B, a Transformer-based model hailed as "*the world's largest and most powerful generative language model.*"

This is an impressive show of Machine Learning engineering, no doubt about it. Yet, should we be excited about this mega-model trend? I, for one, am not. Here's why.

## Evolution of LMs

- Larger is better? https://huggingface.co/blog/large-language-models



# Bigger is better?

- Brain
  - Average: 86 billion neurons, 100 trillion synapses (not all about language)
  - GPT-4 is estimated to have 1.76 trillion parameters
- Megatron
  - Cost: 530 billion parameters, hundreds of DGX A100 multi-GPU servers (each cost $200k) + network + hosting + ... = total of $100 million dollars
  - Training cost: each DGX server can consume up to 6.5 kilowatts + cooling power = carbon footprint
  - BERT-base: 110 million parameters → carbon footprint = NY-SF flight

# Carbon Footprint

**Energy and Policy Considerations for Deep Learning in NLP**

**Emma Strubell** **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

**Abstract**

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the

| **Consumption** | **$CO_2$e (lbs)** |
|---|---|
| Air travel, 1 passenger, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |
| **Training one model (GPU)** | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experimentation | 78,468 |
| Transformer (big) | 192 |
| w/ neural architecture search | 626,155 |

| Model | Hardware | Power (W) | Hours | kWh·PUE | $CO_2$e | Cloud compute cost |
|---|---|---|---|---|---|---|
| Transformer$_{base}$ | P100x8 | 1415.78 | 12 | 27 | 26 | $41–$140 |
| Transformer$_{big}$ | P100x8 | 1515.43 | 84 | 201 | 192 | $289–$981 |
| ELMo | P100x3 | 517.66 | 336 | 275 | 262 | $433–$1472 |
| BERT$_{base}$ | V100x64 | 12,041.51 | 79 | 1507 | 1438 | $3751–$12,571 |
| BERT$_{base}$ | TPUv2x16 | — | 96 | — | — | $2074–$6912 |
| NAS | P100x8 | 1515.43 | 274,120 | 656,347 | 626,155 | $942,973–$3,201,722 |
| NAS | TPUv2x1 | — | 32,623 | — | — | $44,055–$146,848 |
| GPT-2 | TPUv3x32 | — | 168 | — | — | $12,902–$43,008 |

Table 3: Estimated cost of training a model in terms of $CO_2$ emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Carbon footprint of GPT-3?
552 metric tons of carbon emissions, equivalent to driving a passenger vehicle 1.24 million miles (2 million kilometers)

Carbon footprint of GPT-4?
Between 12,456 and 14,994 metric tons $CO_2$e

# On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Bender et al., 2021

What are the possible risks associated with LLMs and what paths are available for mitigating those risks?

Recommendations:

- weigh the environmental and financial costs first,
- invest resources into curating and carefully documenting datasets rather than ingesting everything on the web,
- carry out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values,
- encourage research directions beyond ever larger language models.

# AI's Influence on Elections 2024 ?

The Economist, September 2023

- 4 billion people will vote in Britain, India, Indonesia, Mexico, Taiwan, USA in 2024
- FB on Election 2016: Russian government's 80,000 posts reached 126 million Americans, ~half the electorate
- Micah Musser, Georgetown U.: AI could save $3m of content generating in a $10m campaign
- Quiller helps campaigns write better fundraising emails
- Meta and X/Twitter cut safety teams
- Concerned:
  - Eric Schmidt, former Google CEO: "the 2024 elections are going to be a mess because social media is not protecting us from false generative AI."
  - Sam Altman, CEO of Open AI: "nervous about the impact AI is going to have on future elections (at least until everyone gets used to it)"
- Not so concerned:
  - Jacky Chang, CTO Biden Campaign 2020: "Most voters treat all political message as spam"
  - Brendan Nyhan, Dartmouth: "We still have not one convincing case of a deepfake making any difference whatsoever in politics."
- What do you lean towards?

# Regulations and Policies

European Union:

- **2021**: "The Artificial Intelligence Act" proposed by the European Commission
- **2022**: "General approach position" on the AI Act adopted by the European Council
- **June 2023**: Amendments adopted by the European Parliament
- Now: Negotiation between European Commission and member states

USA:

- **June 2023**: Hearings in US Congress on AI
- **July 2023**: Federal Trade Commission investigation into ChatGPT
- **March 2024**: FDA published details on its coordinated approach to AI in Medicine.
- **June 2024**: FDA published guiding principles for ML enabled medical devices.

China:

- **September 2023**: "AI algorithms must be registered with a government body and somehow embody core socialist values" according to The Economist.
- **May 2024**: Preliminary draft of China's AI law for accountability is drafted.

India:

- **Feb 2021**: Principles of Responsible AI is drafted.
- **March 2023**: National Strategy for AI is drafted.
- **August 2023**: Digital Personal Data Protection Act enacted

# Researchers at BU are helping with the process of AI regulation

National Telecommunications and Information Administration request for comments on AI accountability: "What policies can support the development of AI audits, assessments, certifications and other mechanisms to create earned trust in AI systems?"

Boston University & Chicago University researchers submitted: NTIA-2023-0005-1268

1. AI accountability must be implemented through the entire lifecycle of systems.
2. Accountability mechanisms must be both robust and broadly accessible.
3. Access and transparency are consistent with protecting privacy and intellectual property rights.
4. Accountability and transparency mechanisms are a necessary but not sufficient aspect of AI regulation.
5. AI regulation requires rules for both generalized and specific contexts; existing regulatory agencies and enforcement authorities should be empowered to address AI-related risks within their subject matter domain through establishment of meta-agency with both technical and legal expertise.

# BU Course on Responsible AI, Law, Ethics & Society

- Fall 2024
- CDS DS 680 Tuesday,  3:30 pm - 6:15 pm
- Counts toward “MS in AI” degree as an elective

# Learning Outcomes

- Existing bias and prejudice embedded in vision and language systems
- 7 Types of biases encountered in a typical ML lifecycle
- Socio-Economic impacts of large language model
- Impact of AI on world politics
- Policies and Regulations to govern AI systems.

# Research at BU

- Explore AI research at BU.
- Reach out to one of the PhD students to get involved in research!