Data Science

Artificial Intelligence

Tinglong Dai, PhD

Bernard T. Ferrari Professor

Johns Hopkins Carey Business School

Joy Buolamwini

https://bit.ly/voicingbias

Agenda

1 **Week 1: Foundations**

2 **Week 2: Tools & Techniques**

Machine learning, deep learning, Keras, Python, and no-code tools.

3 **Week 3: CNNs & Business Models**

Convolutional neural networks, using AI to create human value. Assignment #1 due.

4 **Week 4: MLOps, RNNs, & Transformers**

MLOps, RNNs, transformers, evaluation of AI. AI Lab proposal due.

5 **Week 5: LLMs & GenAI**

Large language models, generative AI, using AI to create business value. Assignment #2 due.

6 **Week 6: Reinforcement Learning**

Reinforcement learning, human-AI interaction, system design.

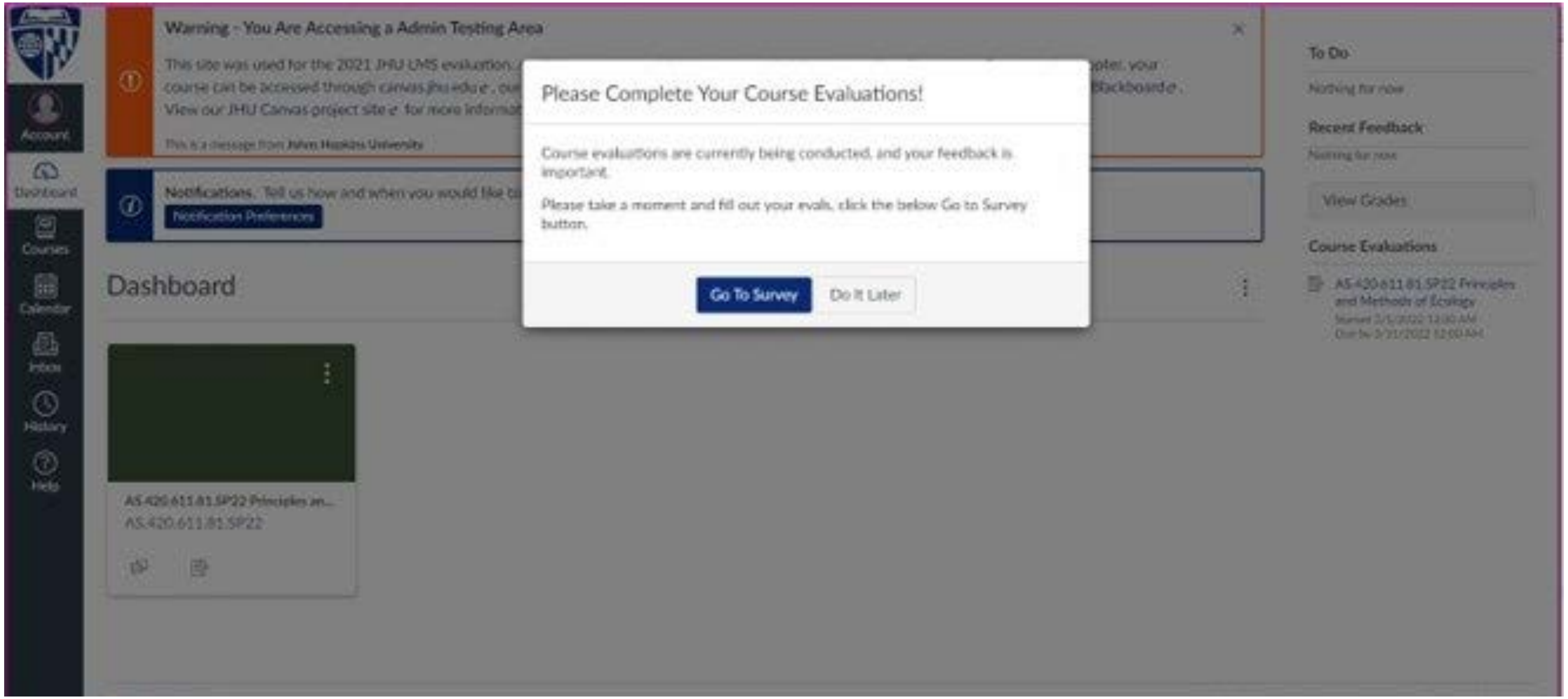
7 **Week 7: Responsible AI**

Responsible AI, future of AI. Lab final deliverable due.

8 **Week 8: Individual Final Exam; Peer Reviews of AI Lab**

Coure Evaluation (10 minutes)

Login to Canvas (https://canvas.jhu.edu) using your JHED ID and password



Final Review

Study Aid for Final Exam

Posted on Canvas

- AI Capabilities
- Machine Learning Basics
- Deep Learning Basics
- Convolutional Neural Networks
- MLOps
- Recurrent Neural Networks and Transformer
- Reinforcement Learning
- AI and Work

Q1. AI Capabilities

A deep learning model can be a type of

A. Supervised learning

B. Unsupervised learning

C. Reinforcement learning

D. Any of the above

Q2. Machine Learning Basics

A machine learning model achieves an 86% accuracy on the training set and 92% accuracy on the test set. The model suffers from

A. Overfitting

B. Underfitting

C. Feature leakage

D. None of the above

Q3. Machine Learning Basics

A machine learning model achieves an 98% accuracy on the training set and 97% accuracy on the test set after extensive feature selection efforts. Yet, when tested on a dataset that the model has never seen before, it achieves a 65% accuracy. The model most probably suffers from

A. Overfitting

B. Underfitting

C. Feature leakage

D. None of the above

Q4. Deep Learning Basics

The most commonly used activation function for a convolutional neural network model is _______; the most commonly used activation function for a recurrent neural network model is _________.

A. ReLU

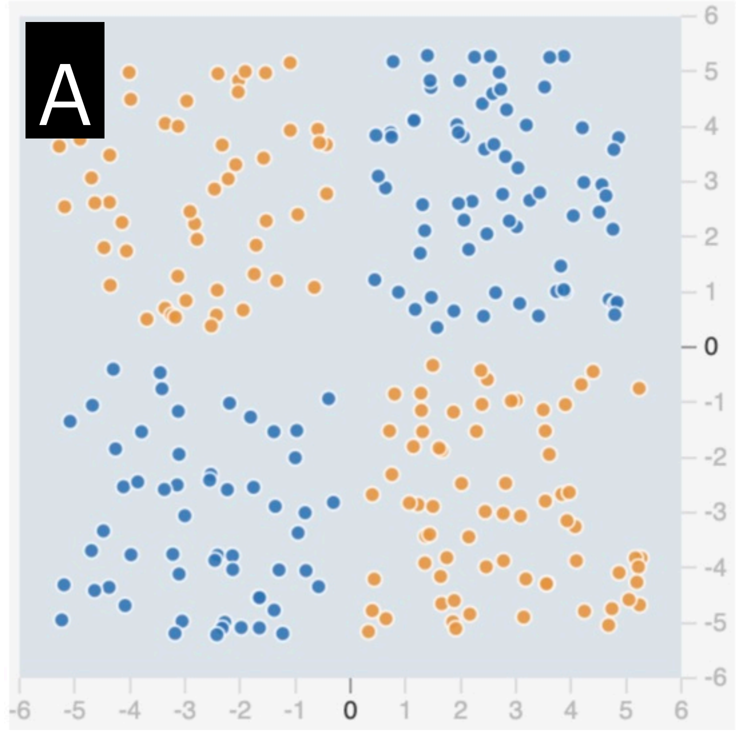
B. Sigmoid

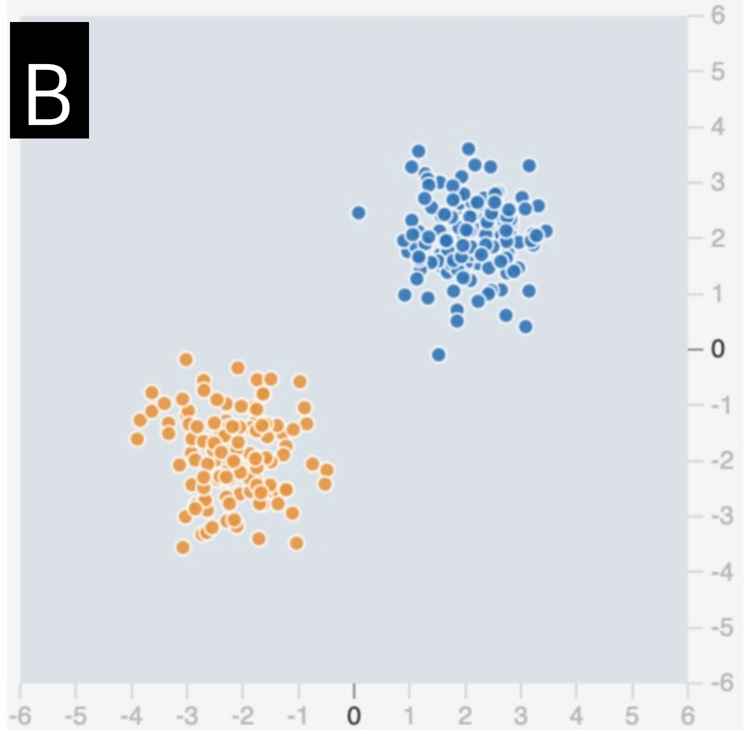
C. Tanh

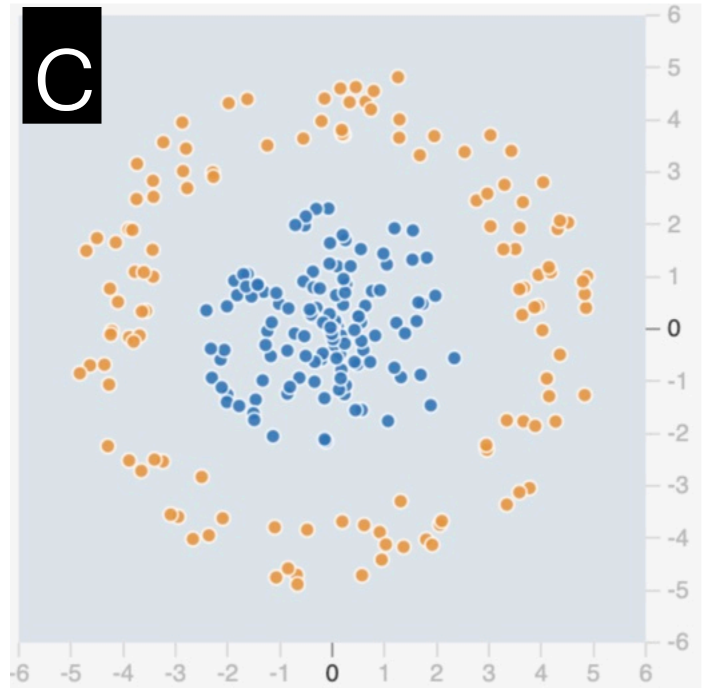
D. Sigmoid or Tanh

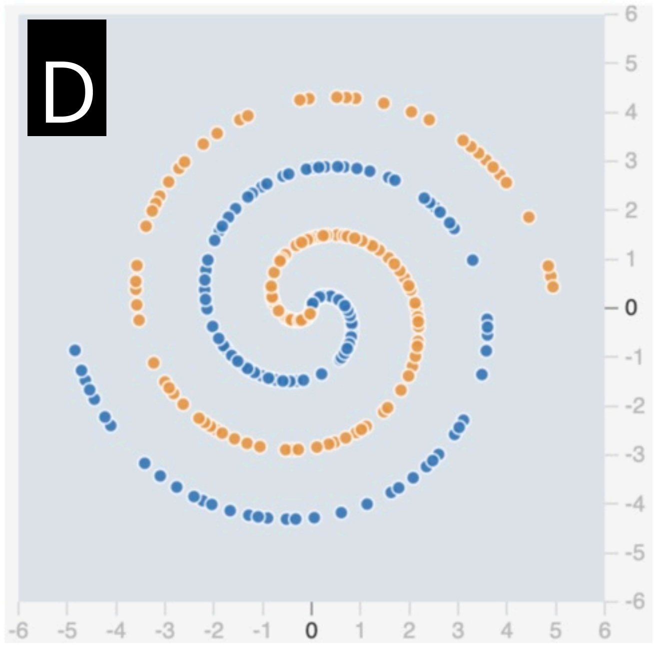
Q5. Deep Learning Basics

Which of the following four data patterns are linearly separable?









For the above linearly separable data pattern, if we build an artificial neural network model, the most suitable activation function for the model is:

A. ReLU

B. Sigmoid

C. Tanh

D. Linear

Q6. Deep Learning Basics

An artificial neural network model suffers from _______ under an excessively small learning rate and suffers from ______ under an excessively large learning rate

A. Excessively slow training; inability to converge to an accurate model

B. Inability to converge to an accurate model; excessively slow training

C. Underfitting; overfitting

D. Overfitting; underfitting

Q7. Convolutional Neural Networks

Convolution

Convolutional filter

0	1
1	0

Input feature map

0	1	1	1	0
0	-1	0	-1	1
0	0	-1	0	1
0	-1	0	-1	1
0	0	0	0	0

Output feature map

1	0	1	-1
-1	0	-2	1
0	-2	0	0
-1	0	-1	1

Q8. Convolutional Neural Networks

ReLU

Input feature map

1	0	1	-1
-1	0	-2	1
0	-2	0	0
-1	0	-1	1

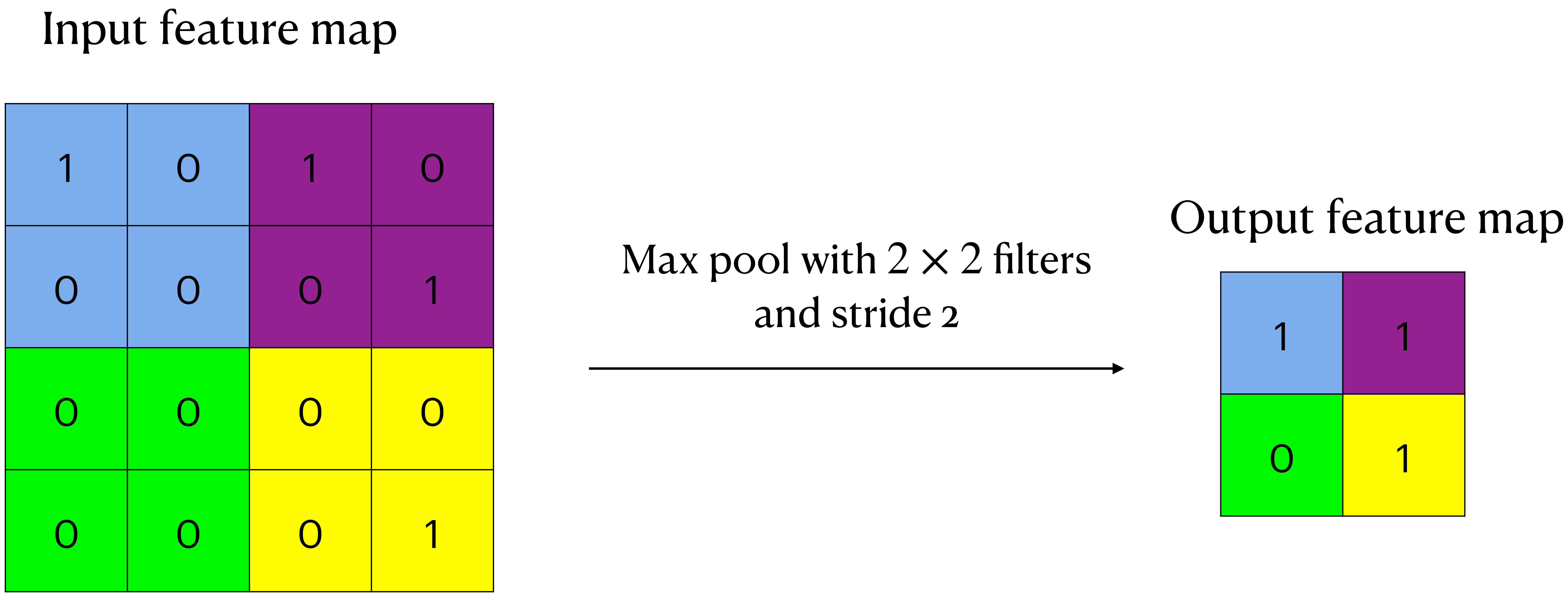
$$\xrightarrow{\text{ReLU } = \max\{x, 0\}}$$

Output feature map

1	0	1	0
0	0	0	1
0	0	0	0
0	0	0	1

Q9. Convolutional Neural Networks

Max Pool



Q10. Interpreting Keras Codes for CNN

```
from tensorflow import keras

from tensorflow.keras import layers

inputs = keras.Input(shape=(28, 28, 1))

x = layers.Conv2D(filters=32, kernel_size=3, activation="relu")(inputs)

x = layers.MaxPooling2D(pool_size=2)(x)

x = layers.Conv2D(filters=64, kernel_size=3, activation="relu")(x)

x = layers.MaxPooling2D(pool_size=2)(x)

x = layers.Conv2D(filters=128, kernel_size=3, activation="relu")(x)
```

A convolutional layer with 32 3×3 filters, followed by a ReLU activation function.

The output of the previous layer is maxpooled with 2×2 filters and a stride of 2

A second convolutional layer with 64 3×3 filters, followed by a ReLU activation function.

The output of the previous layer is maxpooled with 2×2 filters and a stride of 2

A third convolutional layer with 128 3×3 filters, followed by a ReLU activation function.

Q11. MLOps

Suppose you create a feedforward neural network model with three hidden layers, each with 512 neurons. After 30 epochs of training, the model achieved an accuracy level of 95% on the training set and 90% on the test set. Training the model for an additional 20 epochs results in an even larger accuracy gap. Which of the following next steps is the best?

A. Increasing the number of neurons per layer to 1024

B. Increasing the number of neurons per layer to 1024 and add dropout

C. Decreasing the number of neurons per layer to 256 and add dropout

D. Randomly dropping out 50% of the training set

Q12. Recurrent Neural Networks

LSTM RNNs stands for _________ recurrent neural networks, and they rely on ____ units to complement short memory provided by simple RNNs.

A. Linear Short Term Memory; notebook

B. Long Short Term Memory; carry

C. Learning Sequential Time Memory; diary

D. Linear Supervised Training Model; history

Q13. Reinforcement Learning

In a Q-Learning model, the most important tradeoff is between _____ and _____. This tradeoff is reflected through the ______ parameter.

A. Exploration; exploitation; α (learning rate)

B. Optimization; generalization; α (learning rate)

C. Exploration; exploitation; ϵ (exploration rate)

D. Overfitting; underfitting; γ (discount rate)

Q14. AI and Work

Which of the following AI systems (or robots) has the highest level of autonomy:

A. Neglect tolerance = 12 hours; interaction time = 90 minutes

B. Neglect tolerance = 48 hours; interaction time = 30 minutes

C. Neglect tolerance = 24 hours; interaction time = 45 minutes

D. Neglect tolerance = 24 hours; interaction time = 60 minutes

AI Lab & Final Exam

- **AI Lab due on December 12th** (Friday) **12pm**; see Canvas for detailed instructions
  - Please ensure your names do not appear in your submission

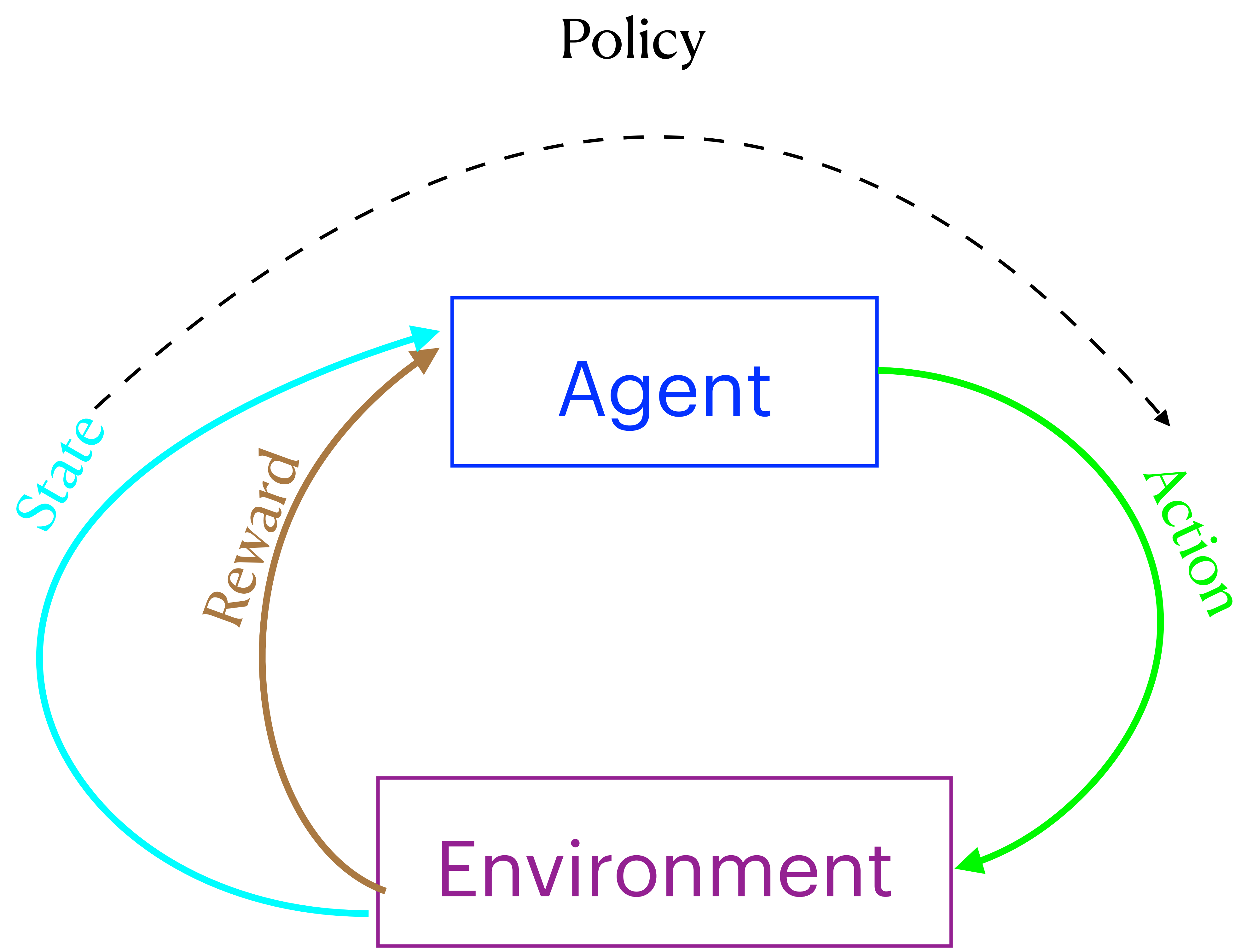
- **Final Exam: December 18th** (Thursday), regular class time, via Canvas
  - ~35 multiple choice questions + 3 short answer questions
  - This is a closed book, closed notes exam; the test will be made available on Canvas a few minutes before the next class
  - We expect everyone to be **in person** throughout the exam
  - More details will be announced later this week via email

Reinforcement Learning; Human-AI Interaction

Reinforcement Learning

- A reinforcement learning model **learns by doing**
  - By contrast, a supervised or unsupervised learning model learns from data

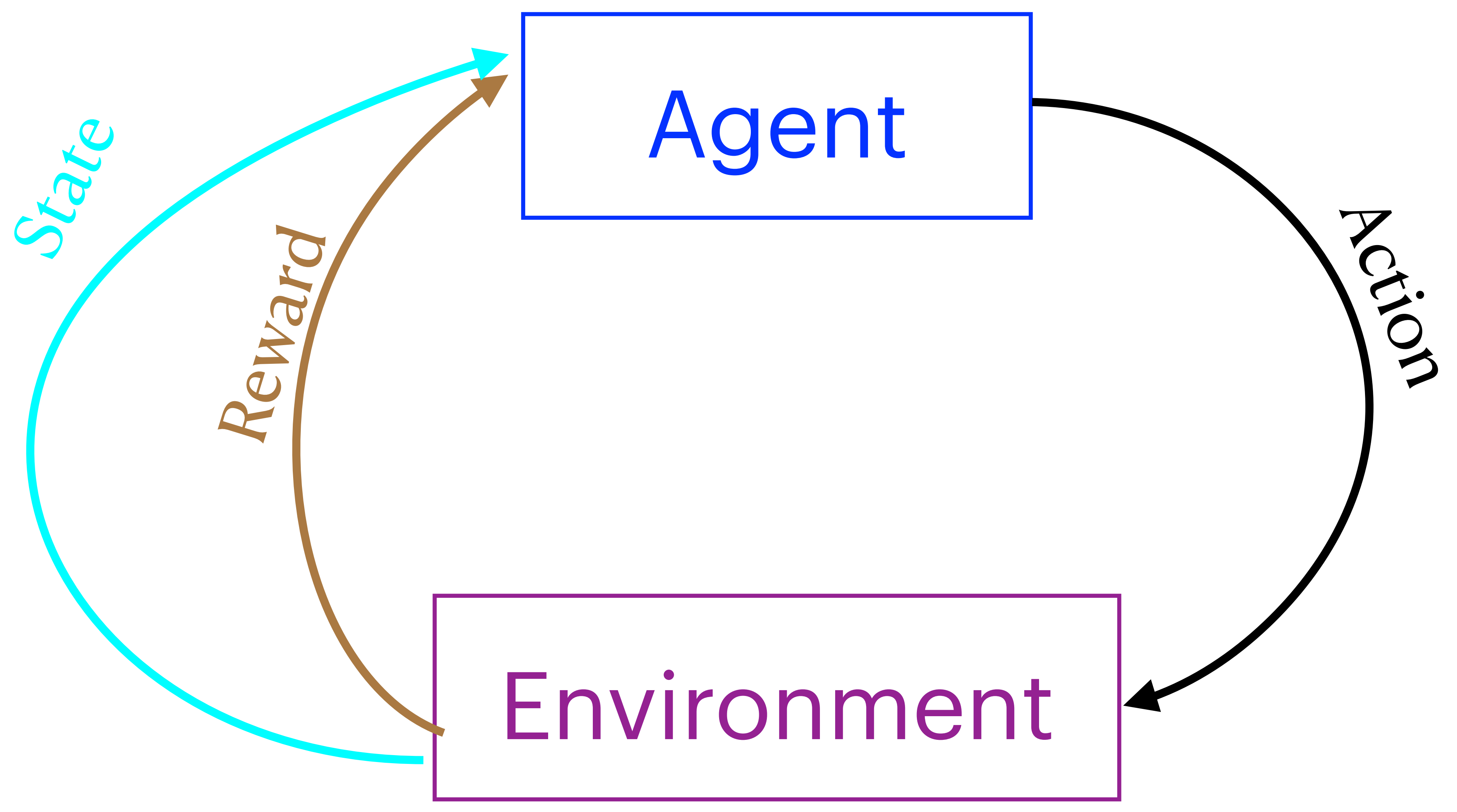
- A reinforcement learning model is an **agent** who takes a sequence of **actions** within some **environment** and receives direct feedback (**reward** and **state**) on the actions it takes



Elements of Reinforcement Learning

The Case of Roomba

- **Actions**: Moving direction and speed
- **Environment** (the floor) returns information back to the agent
  - **State**: Where the Roomba is
  - **Reward**: The agent's score (floor cleaning without incidents)
- **Policy**: The algorithm that the software agent uses to determine its actions (i.e., the policy tells the agent what to do in each state)



Q-Learning

Q = Quality

- We define the quality of an action by how rewarding it is, which consists of
  - instant reward
  - indication of future reward
- Quality = instant reward + indication of future reward
  $Q(s, a) = r + \gamma \cdot \max Q(s', a')$
  - $Q(s, a)$ is the quality value we are calculating for given state s and action a
  - r is the instance reward we achieve by performing the action a in the current state s
  - $\max Q(s', a')$ is the maximum quality of the resulting state s'
  - γ is the discount factor, i.e., how much importance to give to future rewards versus immediate rewards. A high value prioritizes future rewards; a low value prioritizes short-term outcomes

Q-Learning Algorithm

Four Steps

- **Step 1.** For each episode, record the entire sequence of moves and the outcome
- **Step 2.** Assign a reward to the last move the machine makes as its quality value
- **Step 3.** For each previous move, update its quality value using the Bellman equation: $Q(s, a) = Q(s, a) + \alpha \cdot (r + \gamma \cdot \max Q(s', a') - Q(s, a))$
  - α is the learning rate to ensure the algorithm learns the optimal policy gradually
  - If α is too big, the algorithm may learn wrong lessons
- **Step 4.** Using the learned model, the agent chooses an action a to maximize $Q(s, a)$

Exploration vs. Exploitation

- Now, let's rethink **Step 4,** what could possibly go wrong?
  - Using the value tables $Q(s, a)$, the agent chooses an action a to maximize $Q(s, a)$
- Problem: The learned model $Q(s, a)$ is not the same as the true environment, so acting upon the learned model may lead to suboptimal results in the long run
- The agent must make a tradeoff between **exploitation** to maximize its rewards based on the current model and **exploration** to maximize its long-run being
- To ensure the agent explores sufficiently, we often add an **exploration rate** ϵ, which is the probability that the agent takes a random action

Three Key Parameters

- α: learning rate, which determines to what extent newly acquired information overrides old information

- ϵ: exploration rate, which determines the probability of taking a random action, rather than action that gives a maximum value of Q

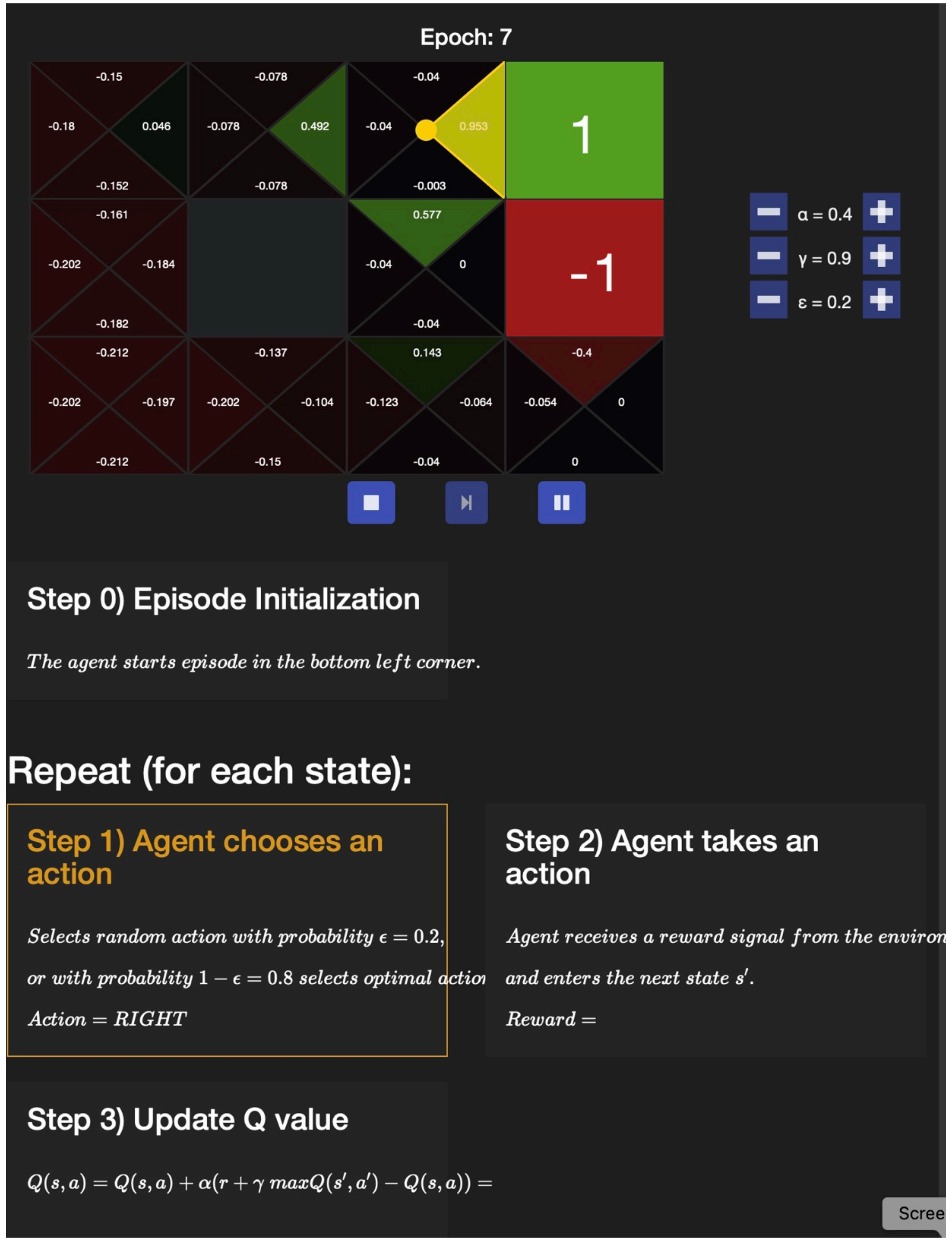
- γ: discount rate, which determines the importance of future rewards vs. instant rewards

Let's explore a simulator to make sense of Q-learning:

https://bit.ly/qsimulator

Q-Learning Simulator

- In this 4×3 maze, Q-learning agent learns by trial and error from interactions with the environment
- Agent starts the episode in the bottom left corner
- The action that is optimal for each state is the action that has the highest long-term reward
- Episode terminates when the agent reaches $+1$ or -1 state; in all other states, the agent receives an instant reward of -0.1
- Essence of the problem: Figuring out a “policy” to get to the green destination with the fewest moves



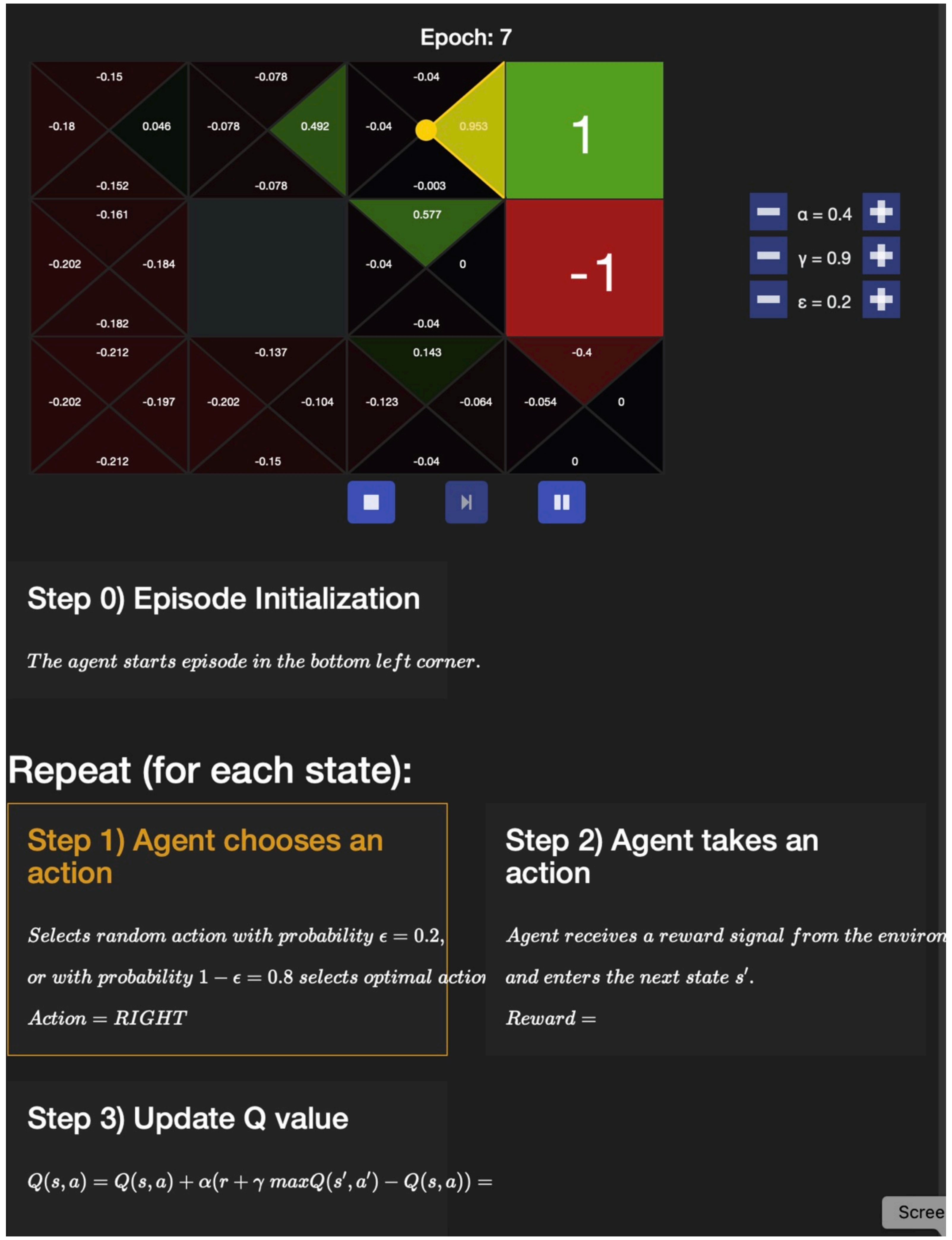
Experiment the Following Settings (10 minutes)

1. Default parameters: $\alpha = 0.1, \gamma = 0.9, \epsilon = 0.3$

2. A high learning rate: $\alpha = 0.9, \gamma = 0.9, \epsilon = 0.3$
   - What's the effect of the learning rate?

3. A low discount rate: $\alpha = 0.9, \gamma = 0.1, \epsilon = 0.3$
   - What's the effect of the discount rate?

4. A high exploration parameter: $\alpha = 0.9, \gamma = 0.1, \epsilon = 0.9$
   - What's the effect of the exploration parameter?

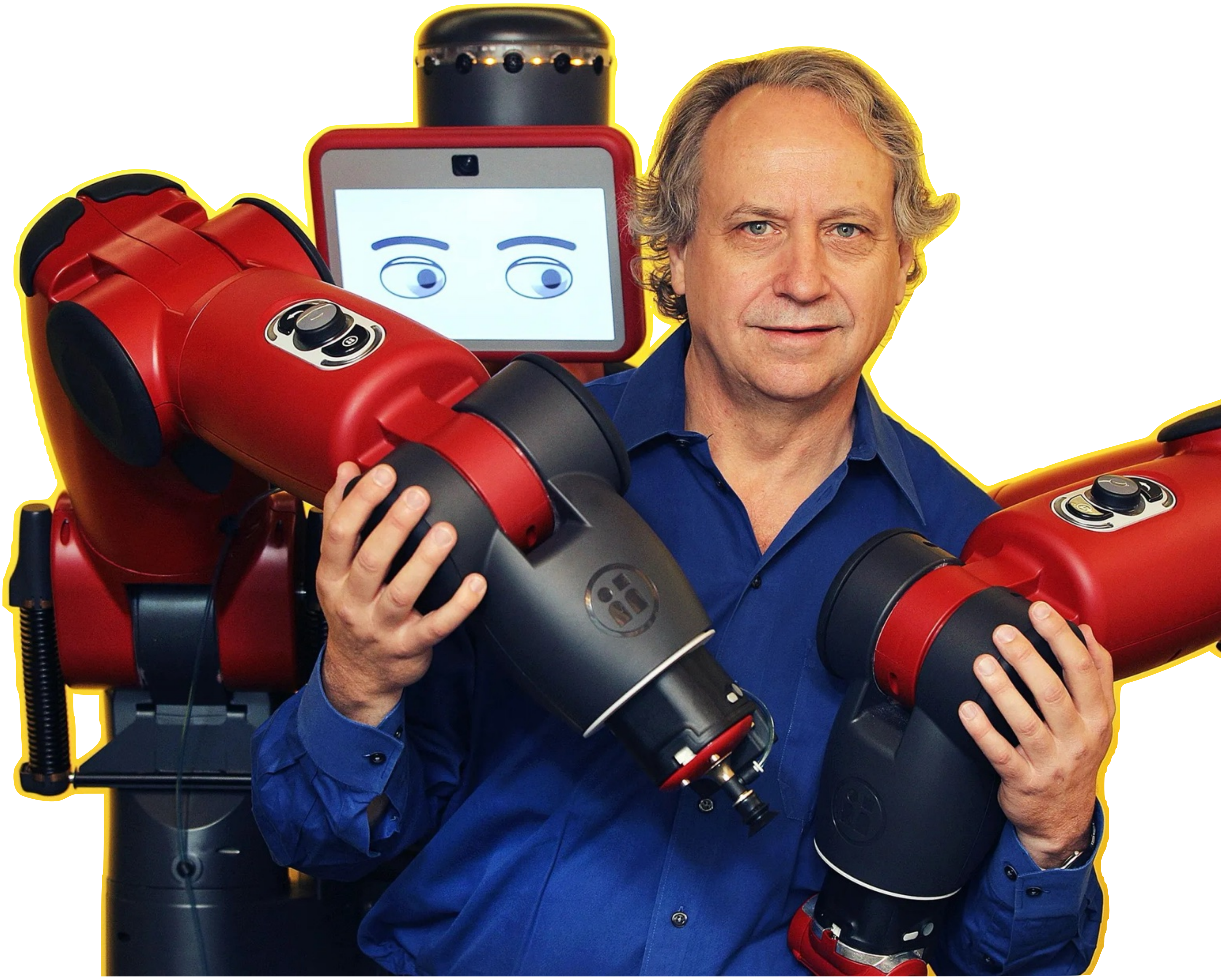


Rodney Brooks' Three Laws of AI

1. When an AI system performs a task, human observers immediately estimate its general competence in areas that seem related. Usually that estimate is wildly overinflated

2. Most successful AI deployments have a human somewhere in the loop (perhaps the person they are helping) and their intelligence smooths the edges

3. Without carefully boxing in how an AI system is deployed there is always a long tail of special cases that take decades to discover and fix. Paradoxically all those fixes are AI-complete themselves



Human-AI Interaction: Autonomy

- An important measure of autonomy for human-AI interaction is the amount of time that a robot can be neglected, that is, the **neglect tolerance** of the AI system (e.g., robot)

- Related measure: **interaction time** = the average amount of time it takes for human operators to interact with the AI system

- An AI system with a high level of autonomy is one that can be neglected for a long period of time without interaction

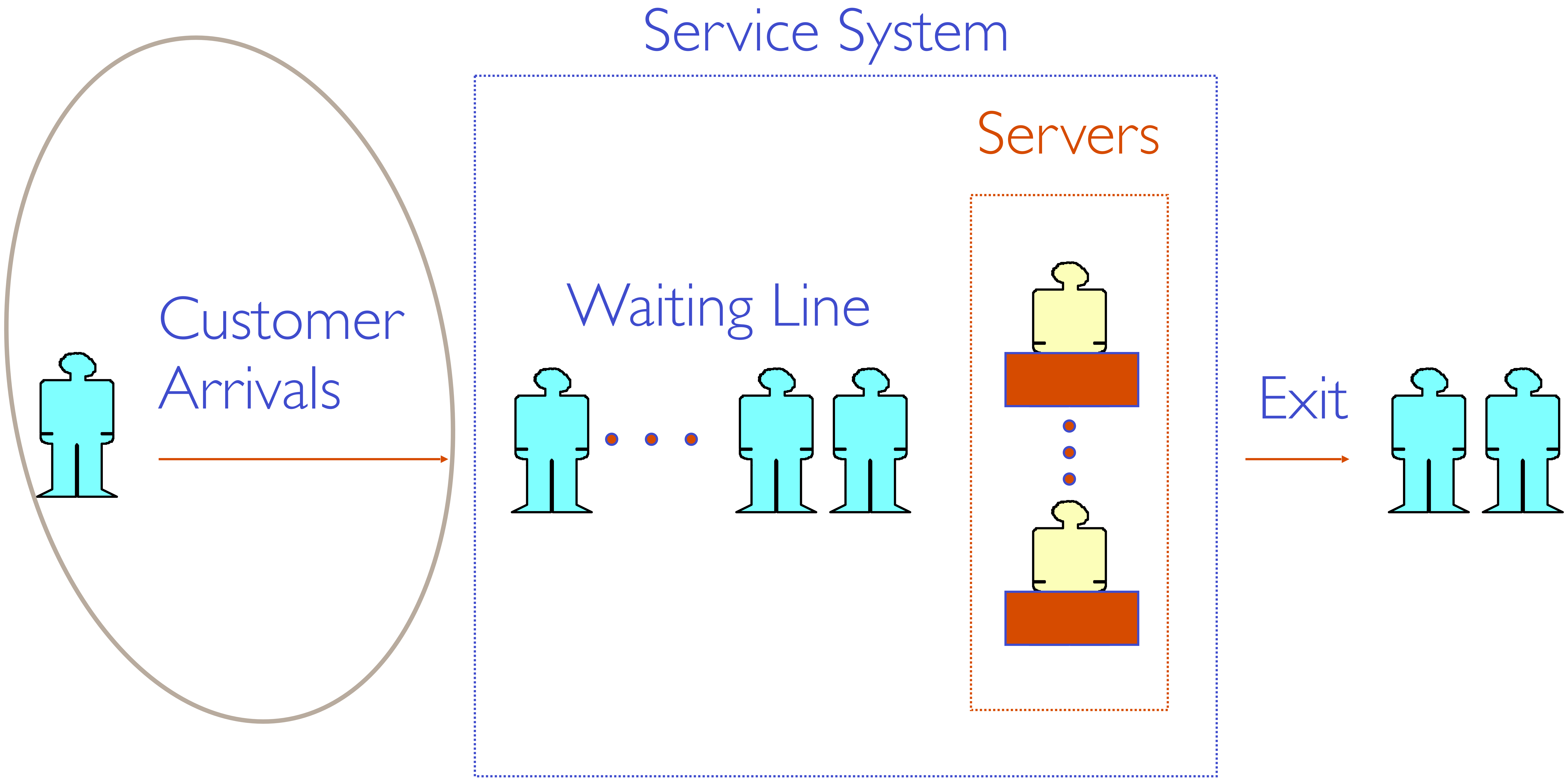
Human-AI Interaction: Teams

- Another important question in human-AI interaction: How many AI systems (e.g., remote robots) a single human can manage?

- **Fan-out**: an upper bound on the number of AI systems (e.g., robots) that a single human operator can manage
  - The fan-out depends on neglect tolerance and interaction time

- A more practical question: How many human operators does it take to efficiently manage a fixed number of AI systems (e.g., robots), allowing for the possibility of adaptable autonomy and dynamic handoffs between humans?

Components of a Queueing System



Queuing Theory

- *a* = average inter-arrival time (**neglect tolerance**)
- *p* = average processing/service time (**interaction time**)
- *m* = number of servers (**number of human operators**)

- *u* = system utilization

  = arrival rate / aggregate service rate

  $= (1/a) / m \cdot (1/p)$

  $= p / m \cdot a$

Applying Queuing Theory to Analyze Human-AI Interaction

- u = system utilization

  = arrival rate / aggregate service rate

$$= \frac{\frac{1}{a}}{m \cdot \frac{1}{p}}$$

$$= \frac{p}{m \cdot a}$$

We can use the utilization rate to determine the required number of human operators

Applying Queuing Theory to Analyze Human-AI Interaction

- Consider a robot team consisting of 100 robots, each with a neglect tolerance of 10 hours and an interaction time of 30 minutes
- To ensure each human operator's utilization level is no higher than 80%, what is the minimal number of human operators?

Applying Queuing Theory to Analyze Human-AI Interaction

- Consider a robot team consisting of 100 robots, each with a neglect tolerance of 10 hours and an interaction time of 30 minutes. To ensure each human operator's utilization level is no higher than 80%, what is the minimal number of human operators?

- Solution:
  - a = 10/100 = 0. 1 hour
  - p = 0.5 hour
  - Utilization $= \dfrac{p}{m \cdot a} \leq 80\,\%$ gives $m \geq \dfrac{5}{0.8} = 6.25$
  - So we need at least 7 human operators

Responsible AI: Managing AI Bias and Opacity

AI Bias in the News

The New York Times

We Teach A.I. Systems Everything, Including Our Biases

Researchers say computer systems are learning from lots and lots of digitized books and news articles that could bake old attitudes into new technology.

Can We Make Our Robots Less Biased Than We Are?

A.I. developers are committing to end the injustices in how their technology is often made and used.

Using A.I. to Find Bias in A.I.

The problem of bias in artificial intelligence is facing increasing scrutiny from regulators and is a growing business for start-ups and tech stalwarts.

THE WALL STREET JOURNAL.

The AI world is making a strong push to root out bias in AI systems, but it faces some significant obstacles. KEITH A. WEBB AND IMAGES FROM ISTOCK

How to Make Artificial Intelligence Less Biased

AI systems can unfairly penalize certain segments of the population—especially women and minorities. Researchers and tech companies are figuring out how to address that.

How Adobe's Ethics Committee Helps Manage AI Bias

A diverse selection of voices can help companies spot potential problems

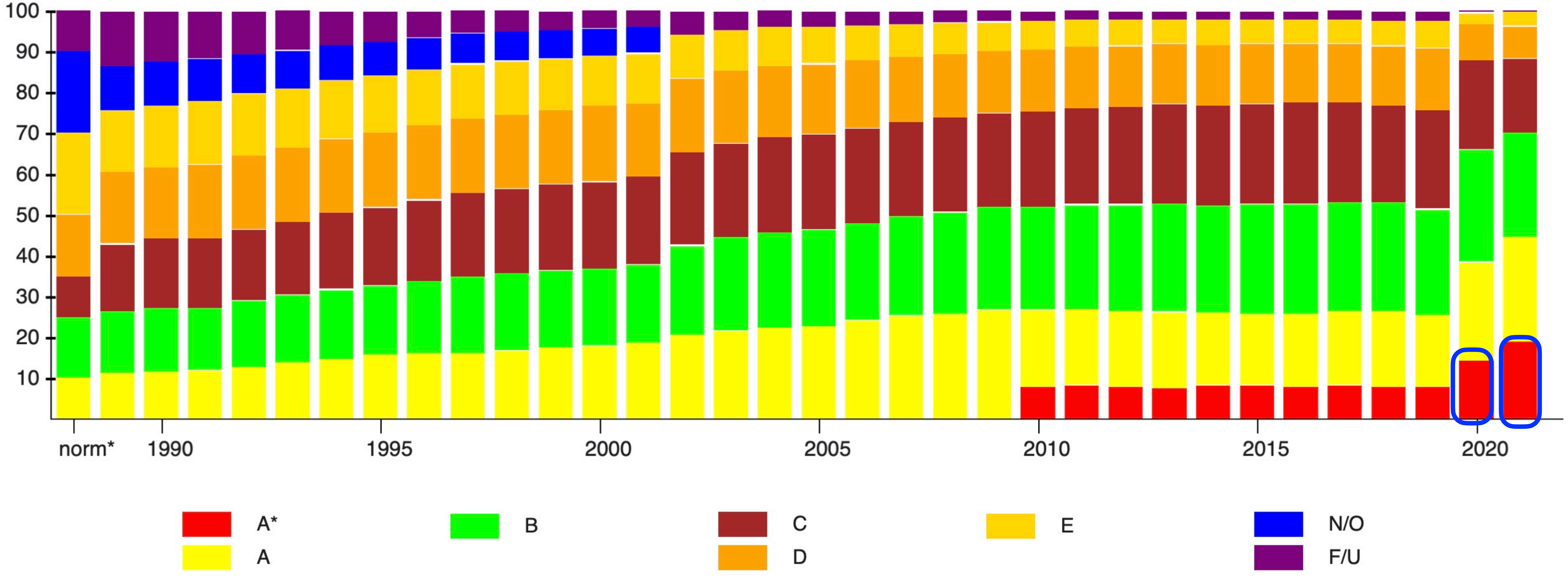
Facebook Dataset Addresses Algorithmic Bias

The dataset, called Casual Conversations, consists of some 45,000 videos of 3,000 participants of various skin tones sharing their age and gender

The Case of UK A-Level Exams

- A levels are U.K subject-based qualifications for students aged between 16 and 19
- Recognized by universities as the standard for admissions
- Canceled in 2020 and 2021 due to the COVID-19 pandemic
  - In lieu of exams, teachers estimate each student's grade based on past performances
- The UK government used an algorithm to moderate teacher predictions

UK A-Level classifications from June 1989 to 2018



https://bit.ly/ftalevel

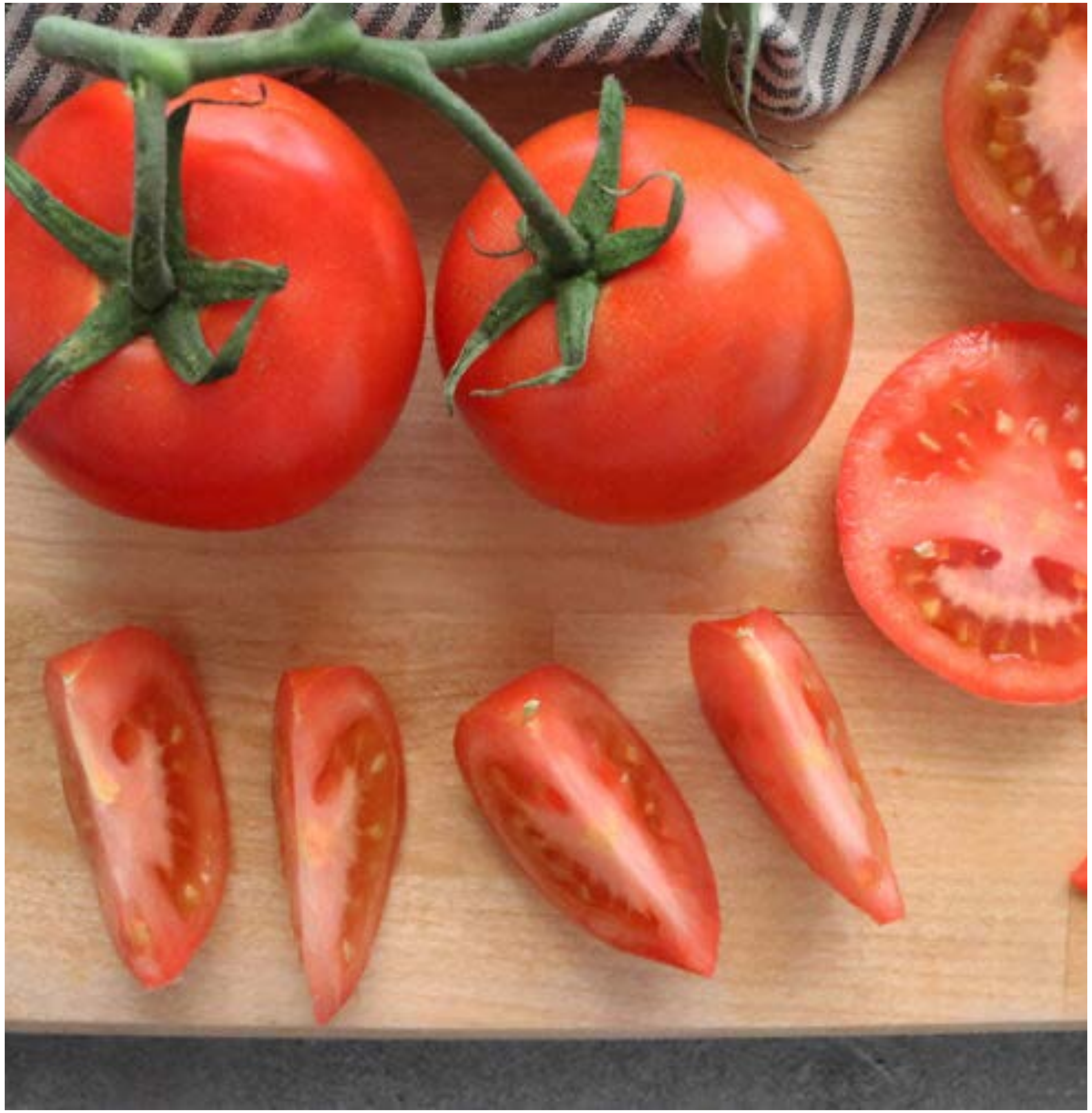
Discussion:
How to design the algorithm differently in view of such biases?

A more fundamental question:
What does **bias** mean?

What Do You See?

- Tomatoes
- Tomato slices
- Tomatoes with stems
- Fresh tomatoes
- A Fruit
- Tomatoes ready for cooking...

What about “red tomatoes”?



What Do You See?

- A **<u>yellow</u>** tomato
- A slice of **<u>yellow</u>** tomato
- A fresh **<u>yellow</u>** tomato
- **<u>Yellow</u>** tomatoes ready for cooking...

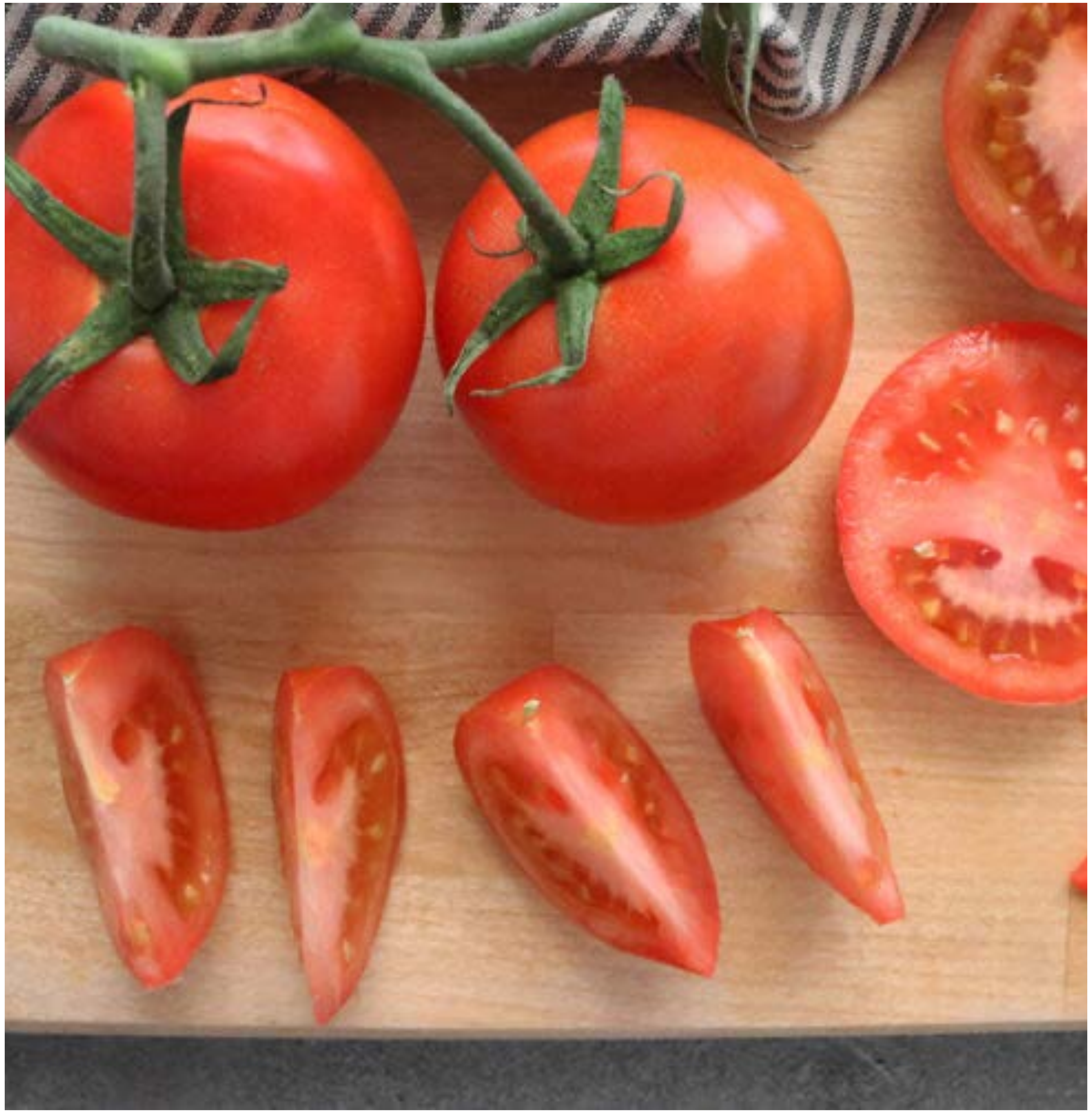
...

What Do You See?

What about “**<u>red tomatoes</u>**”?

We tend not to think of the contents of the image as red tomatoes.

<u>Red</u> is the <u>prototypical</u> color for tomatoes



Labeling, Prototyping, and Stereotyping

- We **label** and **categorize** the world to reduce complex sensory inputs into **simplified** groups that are easier to work with

- **Prototypes** are “typical” representations of a concept or object

- We tend to notice and talk about things that are **atypical**

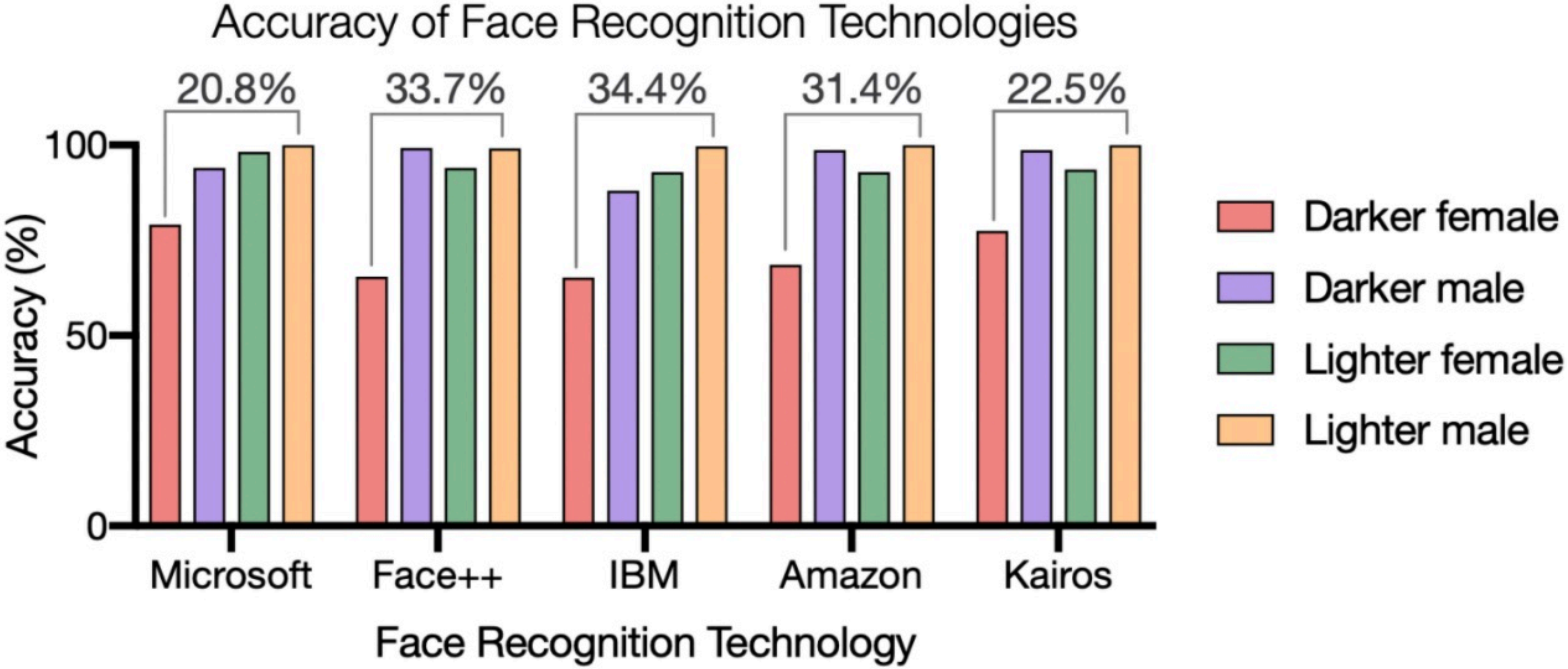
- **Biases** and **stereotypes** arise when particular labels and features **confound decisions** — whether human or artificial

Bias in Facial Recognition

Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

http://gendershades.org/

Bias in Facial Recognition



https://bit.ly/biashsph

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Photos were selected from among those used in Joy Buolamwini's study.

Source: Joy Buolamwini, M.I.T. Media Lab

https://bit.ly/nyt18jb

Bias in Object Recognition

Ground Truth: Soap
Country of Origin: UK

CNN

Predicted object:

Soap

https://bit.ly/fbbias

Bias in Object Recognition

Ground Truth: Soap
Country of Origin: Nepal

→ CNN →

Predicted object:

Food

https://bit.ly/fbbias

Bias in Object Recognition

Ground Truth: Spices
Country of Origin: USA

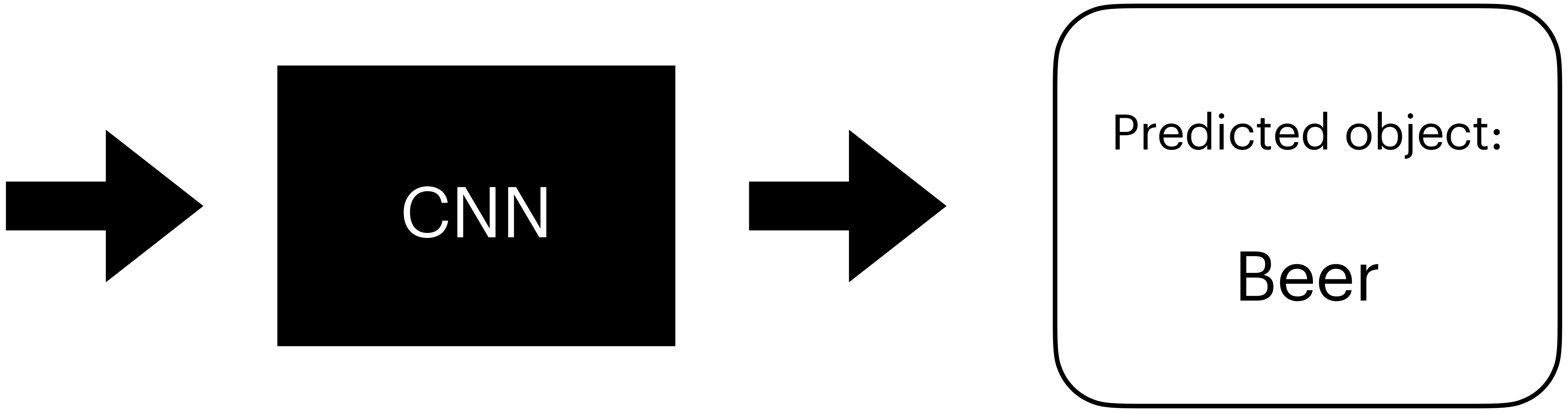
→ CNN →

Predicted object:

Spices

https://bit.ly/fbbias

Bias in Object Recognition



Ground Truth: Spices
Country of Origin: Philippines

https://bit.ly/fbbias

Bias in Object Recognition

Ground Truth: Toothpaste
Country of Origin: USA

CNN

Predicted object:

Toothpaste

https://bit.ly/fbbias

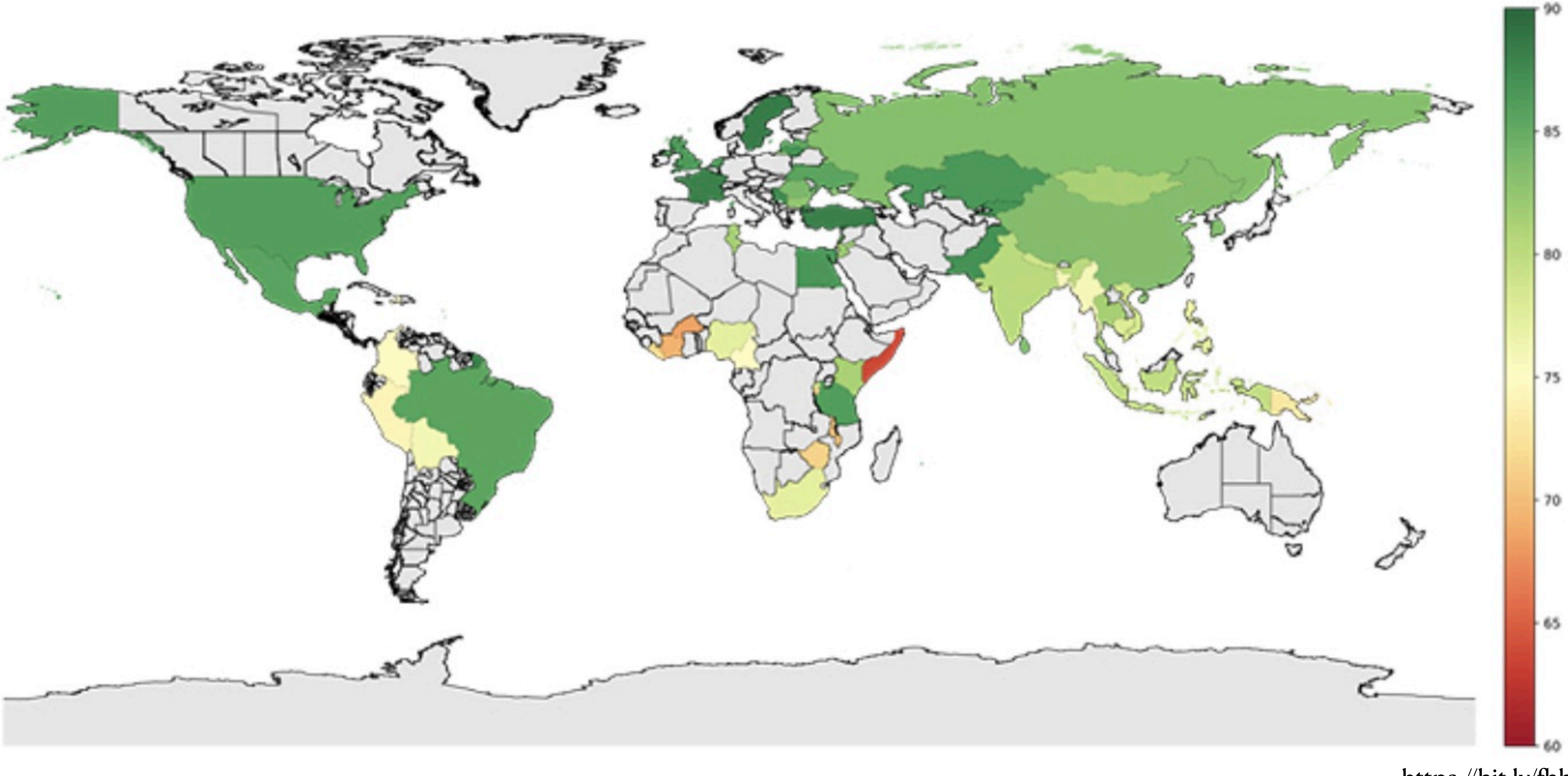
Bias in Object Recognition

Ground Truth: Toothpaste
Country of Origin: Burundi

→ CNN → Predicted object: Wood

https://bit.ly/fbbias

Performance of Six Widely Used Object-Recognition Systems



https://bit.ly/fbbias

Bias at All Stages of AI Development Cycle

- **Data**: imbalances with respect to class labels, features, and input structure
- **Model**: lack of unified uncertainty, interpretability, and performance metrics
- **Training and deployment**: feedback loops that enhance biases
- **Evaluation**: done in bulk (esp. under the Waterfall Model), lack of systematic analysis with respect to data subgroups
- **Interpretation**: human errors and biases distort meaning of results

Common Biases

Data-Driven

Selection bias

Data selection does not reflect randomization (e.g., **class imbalance**)

Reporting bias

What is shared does not reflect real likelihood (e.g., news coverage)

Sampling bias

Particular data instances are more frequently sampled (e.g., hair, skin tone)

Interpretation-Driven

Correlation fallacy

Correlation $\neq$ Causation

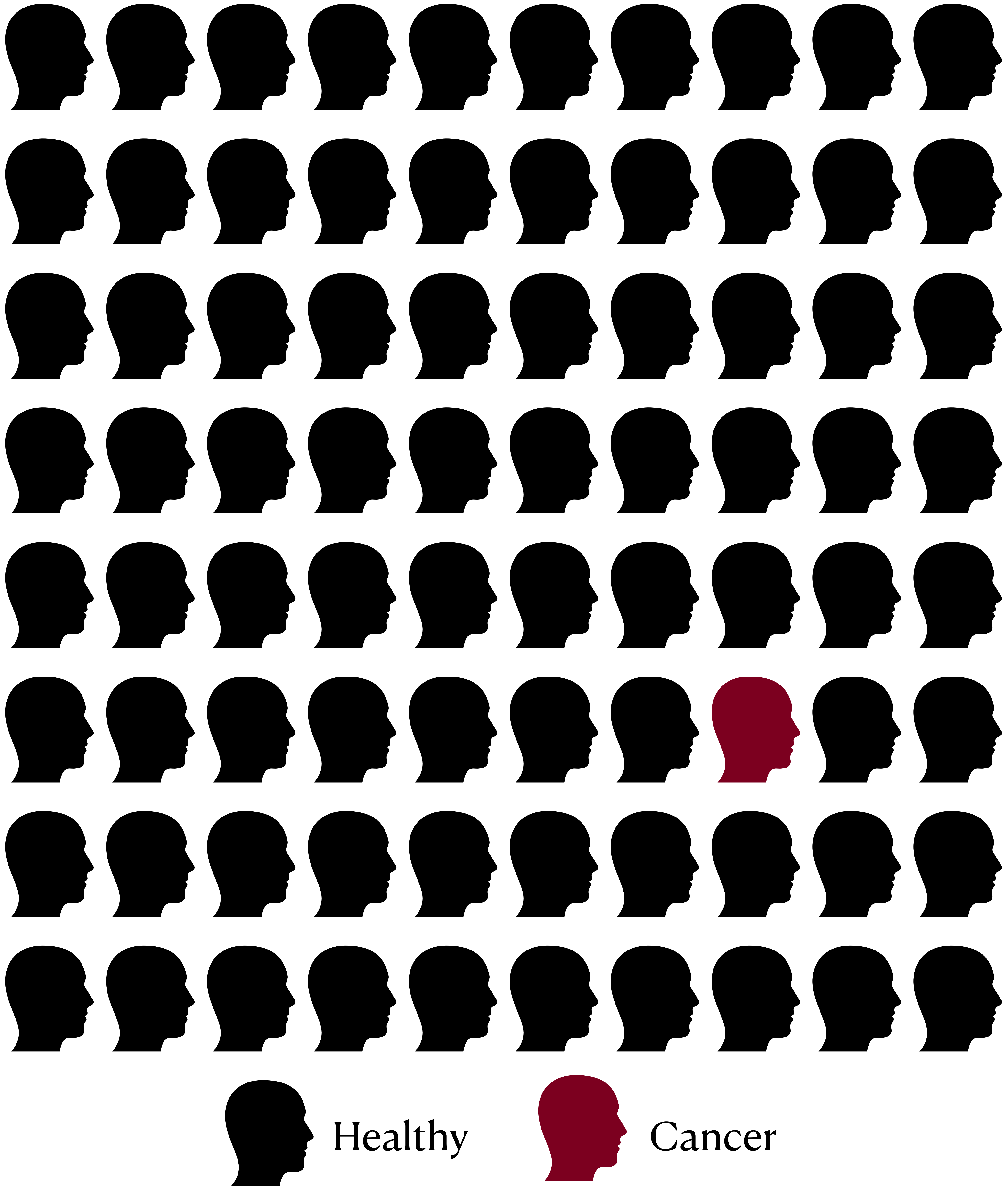
Overgeneralization

"General" conclusions from limited data

Automation bias

AI-generated decisions are favored over human-generated decisions

Class Imbalance

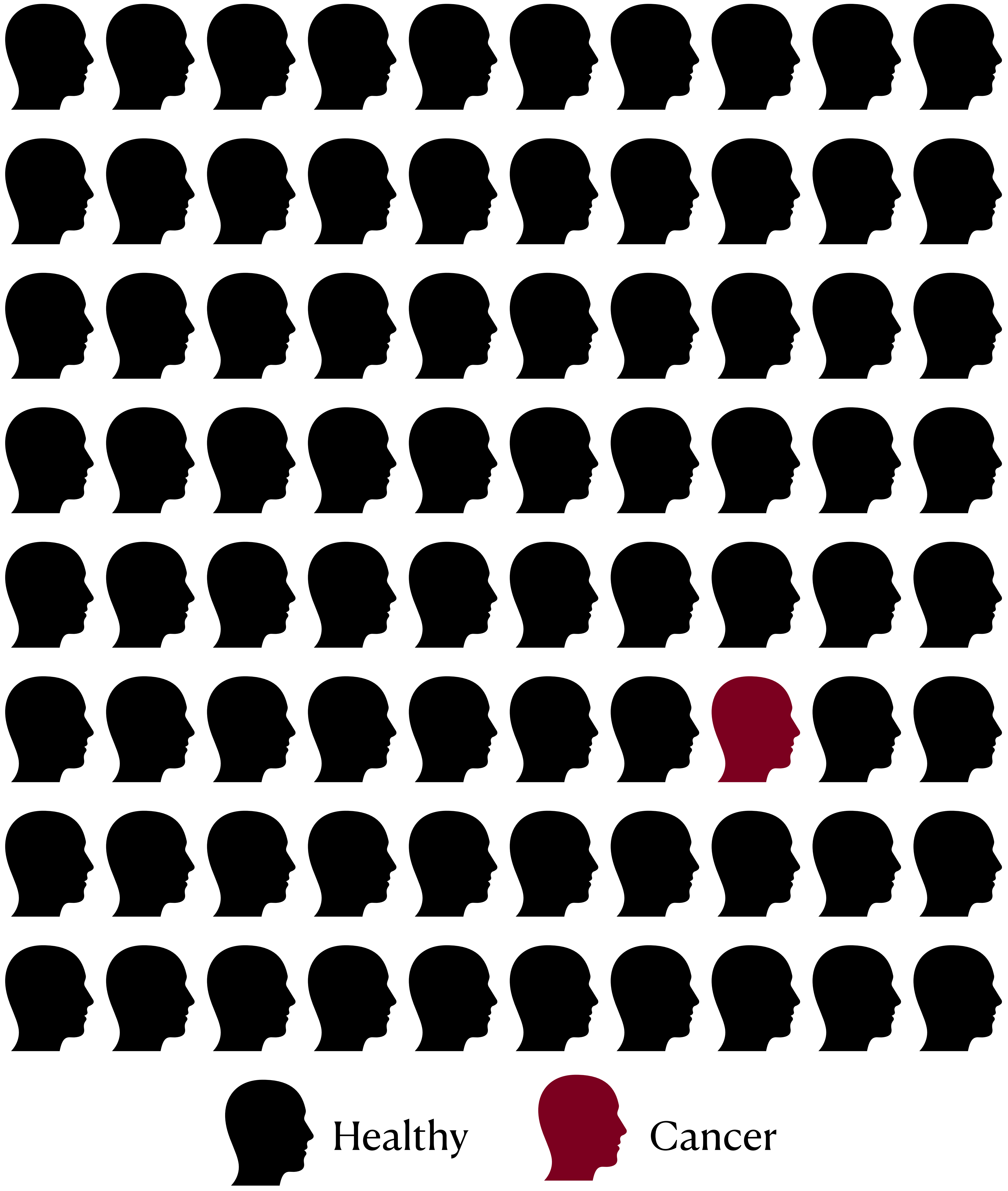


Cancer detection from medical images

- Glioblastoma (GBM): most aggressive and deadliest brain tumor
- GBM incidence in the US: 3 per 100,000 individuals
- Task: Train a CNN model from MRI scan of the brain
- Suppose that class incidence in dataset reflected real-world incidence

- What could possibly go wrong?

Class Imbalance



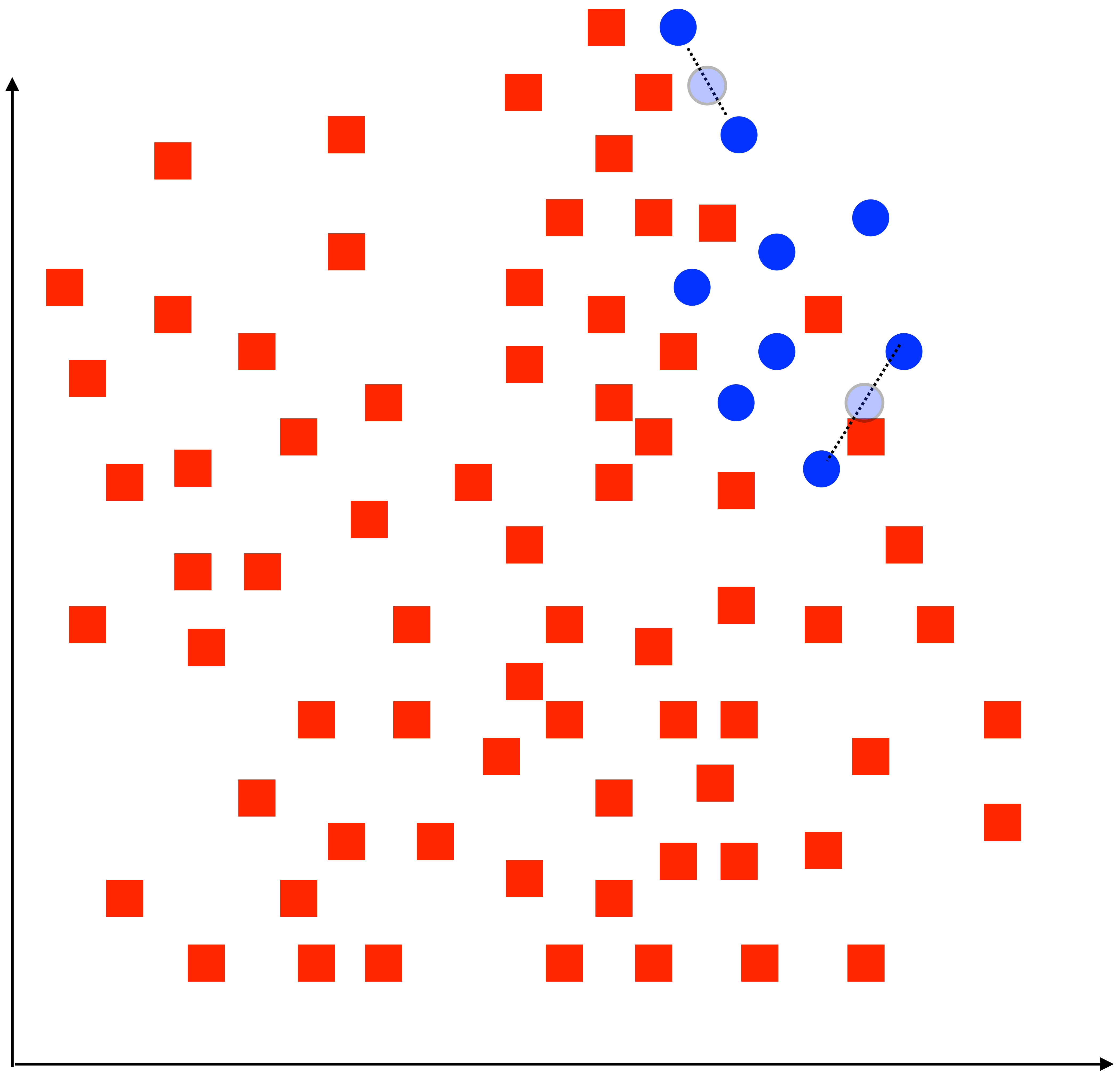
Cancer detection from medical images

- Glioblastoma (GBM): most aggressive and deadliest brain tumor
- GBM incidence in the US: 3 per 100,000 individuals
- Task: Train a CNN model from MRI scan of the brain
- Suppose that class incidence in dataset reflected real-world incidence

By diagnosing everyone as negative, a classifier can achieve a 99.997% accuracy!

Mitigating Strategies for Class Imbalance

- Collecting more data
- Using different performance metrics (e.g., costs of misdiagnosis)
- **Resampling** the dataset
  - Oversampling: Add copies of instances from the under-represented class
  - Undersampling: Delete instances from the over-represented class
- Generate **synthetic samples**
  - SMOTE (Synthetic Minority Over-sampling Technique) algorithm selects two or more similar instances and creates one instance that is randomly combined using those two instances



Example: European Credit Card Database

- Dataset: credit cards transactions in Sept. 2013 by European cardholders
- 492 frauds out of 284,807 transactions
  - Unbalanced because positive class (frauds) account for 0.17% of all transactions
- In Python, SMOTE can be implemented using the following code:

```
from imblearn.over_sampling import SMOTE
oversampler = SMOTE(random_state=0)
X_train ,y_train = oversampler.fit_sample(X_train, y_train)
```

https://bit.ly/ccfd-smote

Lack of transparency complicates debiasing efforts in AI development

Key to improve transparency:
Ensuring subject matter experts can understand AI's decisions and confirms the system is reliable

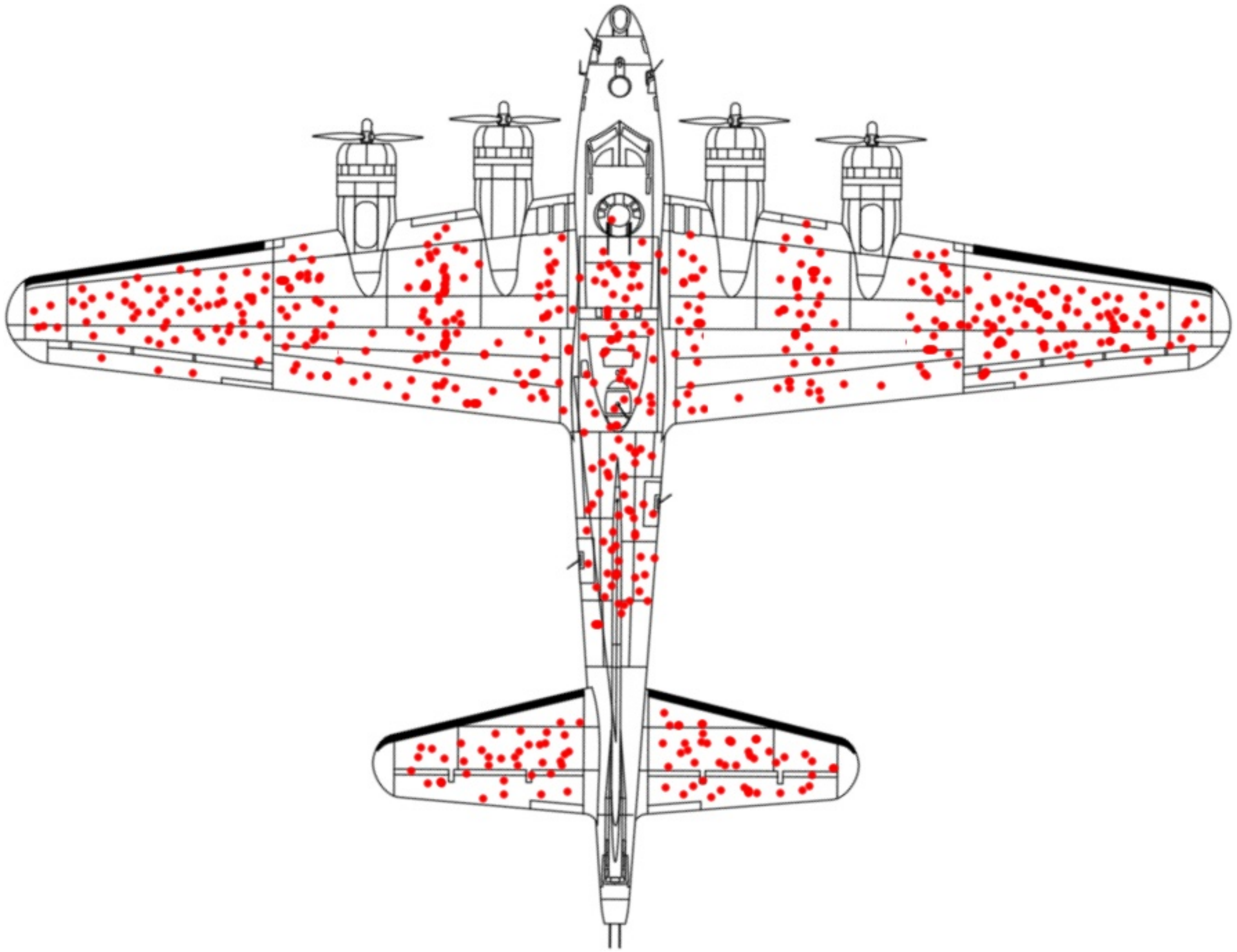
The Importance of Domain Knowledge

An Example from World War II

- Problem: Deciding **where** to provide additional protection to minimize bomber losses to enemy fire

- Observation: Most bullet holes are in wings, tailplanes, and core bodies

- Recommendation: Reinforcing each bomber's wings, tailplanes, and core bodies



What could possibly go wrong?

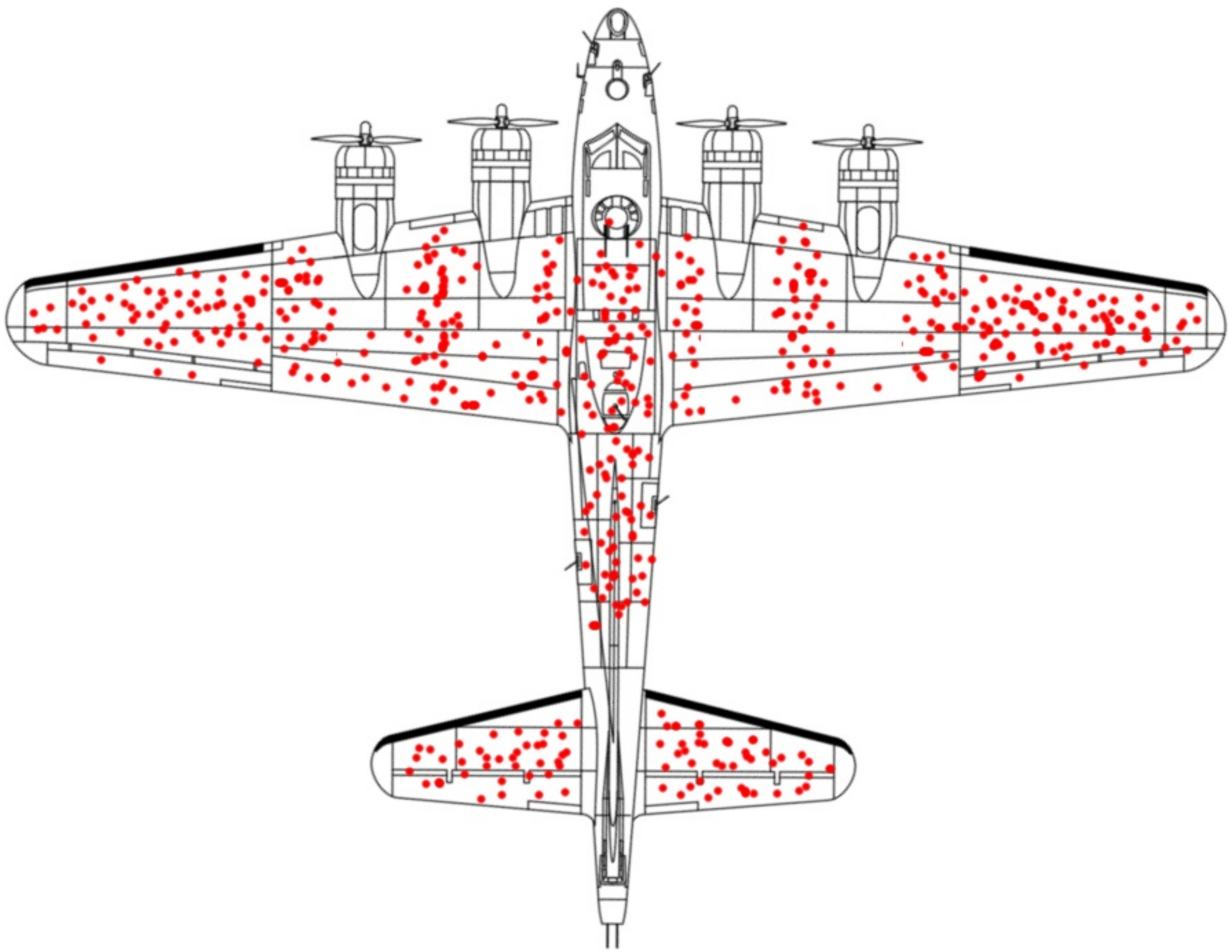
Carmona (2019)

The Importance of Domain Knowledge

An Example from World War II

- Fallacy: The observed dataset only contains bombers that survived

- Correct recommendation: Areas less impacted by bullets should be reinforced



Carmona (2019)

The Importance of Domain Knowledge

An Example from Healthcare

- A Microsoft-funded project developed an AI model to predict the risk for patients with pneumonia admitted to the hospital

- Model prediction: Patients with asthma were at lower risk than the general population, and therefore recommended that those patients not be hospitalized

- What could possibly go wrong?

Carmona (2019)

The Importance of Domain Knowledge

An Example from Healthcare

- Model prediction: Patients with asthma were at lower risk than the general population, and therefore recommended that those patients not be hospitalized

- The fallacy: The dataset was biased because any asthmatic patient admitted with pneumonia went directly to the ICU, lowering the risk for those patients dramatically

- How to prevent such faulty models from being developed?
  - Involving subject-matter experts throughout AI development (MLOps)

Carmona (2019)

“The real risk with AI isn’t malice but competence. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren’t aligned with ours, we’re in trouble. You’re probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green energy project and there’s an anthill in the region to be flooded, too bad for the ants. Let’s not place humanity in the position of those ants.”

Stephen Hawking (1942–2018)

~~We’re Doomed…~~

We’ll Be Just Fine

Some Parting Words…

JAMA Network | Open

Original Investigation | Health Informatics

Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages

Patricia Garcia, MD; Stephen P. Ma, MD, PhD; Shreya Shah, MD; Margaret Smith, MBA; Yejin Jeong, BA; Anna Devon-Sand, MPH; Ming Tai-Seale, PhD, MPH; Kevin Takazawa, BBA; Danyelle Clutter, MBA; Kyle Vogt, BA; Carlene Lugtu, MCiM; Matthew Rojo, MS; Steven Lin, MD; Tait Shanafelt, MD; Michael A. Pfeffer, MD; Christopher Sharp, MD

Abstract

IMPORTANCE The emergence and promise of generative artificial intelligence (AI) represent a turning point for health care. Rigorous evaluation of generative AI deployment in clinical practice is needed to inform strategic decision-making.

OBJECTIVE To evaluate the implementation of a large language model used to draft responses to patient messages in the electronic inbox.

DESIGN, SETTING, AND PARTICIPANTS A 5-week, prospective, single-group quality improvement study was conducted from July 10 through August 13, 2023, at a single academic medical center (Stanford Health Care). All attending physicians, advanced practice practitioners, clinic nurses, and clinical pharmacists from the Divisions of Primary Care and Gastroenterology and Hepatology were enrolled in the pilot.

INTERVENTION Draft replies to patient portal messages generated by a Health Insurance Portability and Accountability Act–compliant electronic health record–integrated large language model.

MAIN OUTCOMES AND MEASURES The primary outcome was AI-generated draft reply utilization as a percentage of total patient message replies. Secondary outcomes included changes in time measures and clinician experience as assessed by survey.

RESULTS A total of 197 clinicians were enrolled in the pilot; 35 clinicians who were prepilot beta users, out of office, or not tied to a specific ambulatory clinic were excluded, leaving 162 clinicians included in the analysis. The survey analysis cohort consisted of 73 participants (45.1%) who completed both the presurvey and postsurvey. In gastroenterology and hepatology, there were 58 physicians and APPs and 10 nurses. In primary care, there were 83 physicians and APPs, 4 nurses, and 8 clinical pharmacists. The mean AI-generated draft response utilization rate across clinicians was 20%. There was no change in reply action time, write time, or read time between the prepilot and pilot periods. There were statistically significant reductions in the 4-item physician task load score derivative (mean [SD], 61.31 [17.23] presurvey vs 47.26 [17.11] postsurvey; paired difference, −13.87; 95% CI, −17.38 to −9.50; *P* < .001) and work exhaustion scores (mean [SD], 1.95 [0.79] presurvey vs 1.62 [0.68] postsurvey; paired difference, −0.33; 95% CI, −0.50 to −0.17; *P* < .001).

CONCLUSIONS AND RELEVANCE In this quality improvement study of an early implementation of generative AI, there was notable adoption, usability, and improvement in assessments of burden and burnout. There was no improvement in time. Further code-to-bedside testing is needed to guide future development and organizational strategy.

Key Points

Question What is the adoption of and clinician experience with clinical practice deployment of a large language model used to draft responses to patient inbox messages?

Findings In this 5-week, single-group, quality improvement study of 162 clinicians, the mean draft utilization rate was 20%, there were statistically significant reductions in burden and burnout score derivatives, and there was no change in time.

Meaning These findings suggest that the use of large language models in clinical workflows was spontaneously adopted, usable, and associated with improvement in clinician well-being.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

JAMA Network Open. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

Finding: No time savings, but significant drops in cognitive task load and work exhaustion.

We Live in the Era of Vibe Coding,
Vibe Researching,
Vibe Building.

We Live in the Golden Era of Builders.

AI AS CATALYST: LOWERING ACTIVATION ENERGY

AI shortens the path from ideation to execution, reducing the startup time that fuels procrastination.

THE BUILDER FLYWHEEL


LIFEHACKER

Home → Tech → AI

Google's Co-Founder Says AI Performs Best When You Threaten It

Key principles: Start ugly • Iterate fast • Learn from feedback, especially from others

Let’s face it: Listening to feedback is hard,
especially when it’s from others.

https://www.deathbulge.com/comics/155

AI AS THE EVALUATION SHIELD

Summarize Themes First

Ask AI to extract patterns and high-level feedback before showing individual comments

Delay Verbatims

Read specific critiques only after you understand the overall themes

Act, Then Read

Make improvements based on themes, then review details periodically for trends

This protocol helps us process feedback constructively: ingest data, extract themes, sample representative verbatims, take action, then review trends.

THE COGNITIVE LOAD FUNNEL

Human decision-making involves three stages, each demanding cognitive resources:

Parse

Gathering and organizing raw information

Frame

Structuring the problem and identifying options

Decide

Evaluating trade-offs and choosing a path

AI can absorb much of the Parse and Frame work, leaving us with fewer micro-choices and clearer options when us reach the Decide stage.

TRUST IS DESIGNED

- **Calibrate Help**
  Match AI assistance to task complexity and user expertise
- **Shield Emotion**
  Create buffers between raw AI output and human reaction
- **Attribute Fairly**
  Design systems that reward outcomes, not tool stigma

What are we really talking about
when we talk about AI?

THREE PLACEMENTS FOR HUMANS

Human-In-the-Loop

Assist and verify during the process. Humans provide real-time guidance and corrections as AI works.

Human-Over-the-Loop

Audit and override after completion. Humans review outputs and can reverse AI decisions.

Human-After-the-Loop

Review and learn from patterns. Humans analyze aggregate outcomes to improve the system.

The right placement depends on our domain, risk tolerance, and the stakes of the decision. Use domain-specific examples to determine which model fits your context.

"Let others praise ancient times; I am glad I was born in these."

–OVID (43 BC – 17/18 AD)

Two hikers walked through the woods,
when suddenly, a bear appeared–
chasing them.

They ran,
one stopped, quickly putting on running shoes.

The other hiker, confused, asked,
"Why are you putting on running shoes?
You can't outrun a bear!"

The first hiker smiled,
"I don't need to outrun the bear–
I just need to outrun you!"

WHAT MACHINES CANNOT REPLICATE (YET)

- Machines don't know love, feel joy, or experience the profound satisfaction of overcoming failure through persistence
- Machines cannot grapple with ethical dilemmas or develop the moral urgency essential for leadership and decision-making
- Only humans can exercise true discernment, care about truth and beauty, and take genuine responsibility for their choices

Information is not knowledge.

Knowledge is not wisdom.

Wisdom is not truth.

Truth is not beauty.

Beauty is not love.

LOVE IS ALL YOU NEED.

PARAPHRASED FROM FRANK ZAPPA

May AI Be With You