
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM341-R

# Transforming responsible AI from theory into practice

Dr. Nashlie Sephus (she/her)

Tech Evangelist
AWS AI

Mia Mayer (she/her)

Data Scientist
AWS Machine Learning University

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

What is responsible AI?

Dimensions of responsible AI

Group fairness vs. individual fairness

Fairness criteria and metrics

Responsible AI throughout the machine learning lifecycle

Q&A

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# The reach of AI and ML is growing

Global spending on AI-based technologies will reach **$204 billion by 2025**

57% of executives said that AI would **transform their organization** in the next 3 years

AI plays a role in nearly **every part of an organization** from chatbots to personalized recommendations

With the growth of AI comes the recognition that we must all use it **responsibly**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

www.idc.com

# Our commitment

**Our commitment to develop AI and machine learning in a responsible way is integral to our approach**



Teach practical code implementations and best practices



Nurture and educate a more diverse generation of leaders in ML



Integrate responsible AI into the end-to-end ML lifecycle



Advance the science behind responsible AI

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Responsible AI:
# AI that is innovative and trustworthy and respects human rights and democratic values

**Source: OECD**

https://oecd.ai/en/ai-principles

→ Dimensions of responsible AI

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Dimensions of responsible AI

| Dimension | Example metric |
| --- | --- |
| Privacy & security | Is data used in accordance with privacy and legal considerations and protected from theft and exposure? |
| Fairness | Are there harmful disparities in system performance across subpopulations? |
| Explainability | Does the system offer a clear rationale for its decisions? |
| Robustness | How hard is it to confuse or fool the system, for example, with “adversarial” examples? |
| Transparency | Are users enabled to make informed choices about their use of the system? |
| Governance | How do you enforce and ensure these responsible AI practices are being carried out among all stakeholders? |

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Trade-offs in responsible AI

1. Dimensions of responsible AI (e.g. privacy preservation) can be at odds with primary objective to have best possible model performance
2. It is challenging to maximize all dimensions of responsible AI simultaneously

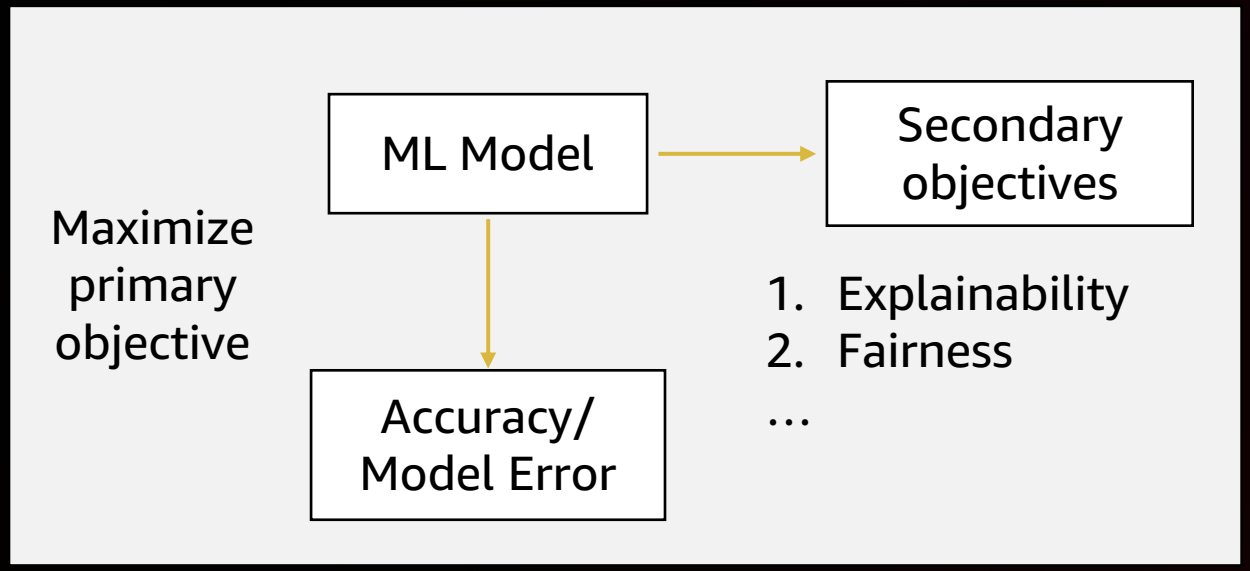
What could be a potential issue when making the model and all training data completely transparent?

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Trade-offs in responsible AI

1. Dimensions of responsible AI can be at odds with primary objective to have best possible model performance




2. It is challenging to maximize all dimensions of responsible AI simultaneously
   - Example: Making model results and components more transparent → potential security and privacy risk
   - Example: Preserving privacy makes data more coarse → potentially degrades ability to explain model behavior

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Dimensions of responsible AI

| Dimension | Example metric |
| --- | --- |
| Privacy & security | Is data used in accordance with privacy and legal considerations and protected from theft and exposure? |
| Fairness | Are there harmful disparities in system performance across subpopulations? |
| Explainability | Does the system offer a clear rationale for its decisions? |
| Robustness | How hard is it to confuse or fool the system, for example, with “adversarial” examples? |
| Transparency | Are users enabled to make informed choices about their use of the system? |
| Governance | How do you enforce and ensure these responsible AI practices are being carried out among all stakeholders? |

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Unwanted bias

Bias is a term used in different domains, with different meanings:

| | |
|---|---|
| **Model Evaluation** | Unexplained error stemming from model selection & assumptions. |
| **Neural Networks** | Value that helps control quality of fit during training of NN. |
| **Social Context** | Systematic differences; disproportionate favor for or against different subpopulations*. |
| **Responsible AI** | Harmful disparities in system behavior on subpopulations*; also “unwanted bias”. |

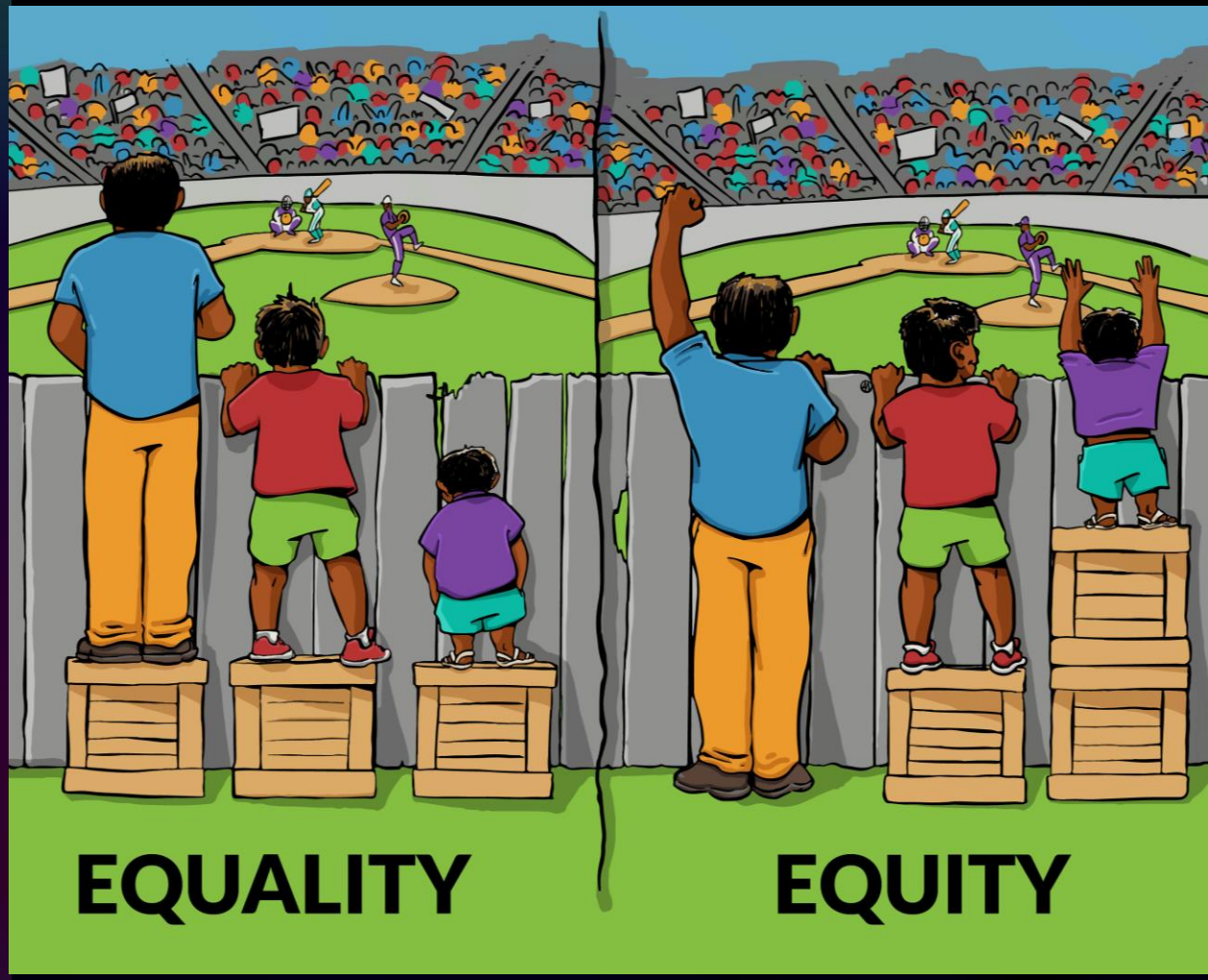
aws © 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Subpopulations: groups with shared values of demographic variables.

# Fairness criteria and metrics

To operationalize theoretical constructs such as fairness, we need mathematical constructs:

- Equality of opportunity
- Demographic parity
- …




Source: https://interactioninstitute.org/illustrating-equality-vs-equity/

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Fairness types

## Group fairness

Equalize impact/treatment across groups

- Strong theory and algorithms and practical implementations

## Individual fairness

Ensures that similar individuals are treated similarly

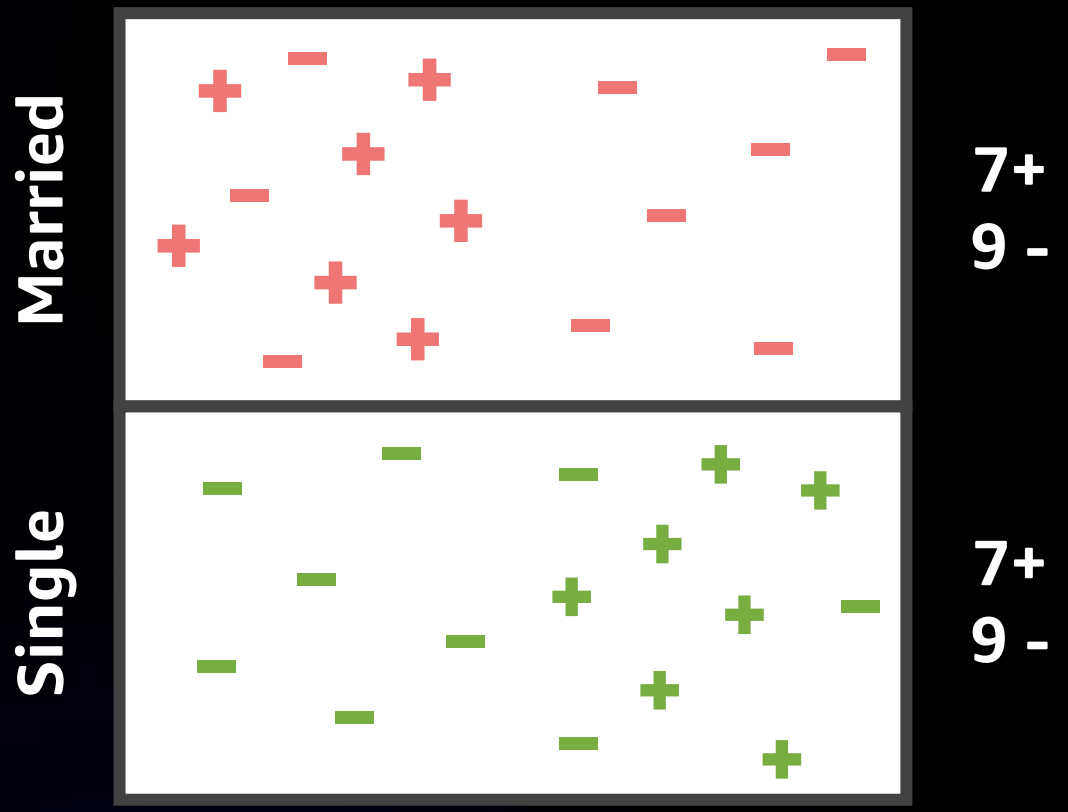
- Strong (nonstatistical) assumptions prevent practical implementations

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How to define a group?

- People can exhibit multiple demographic attributes
- If multiple attributes intersect, new subgroups emerge



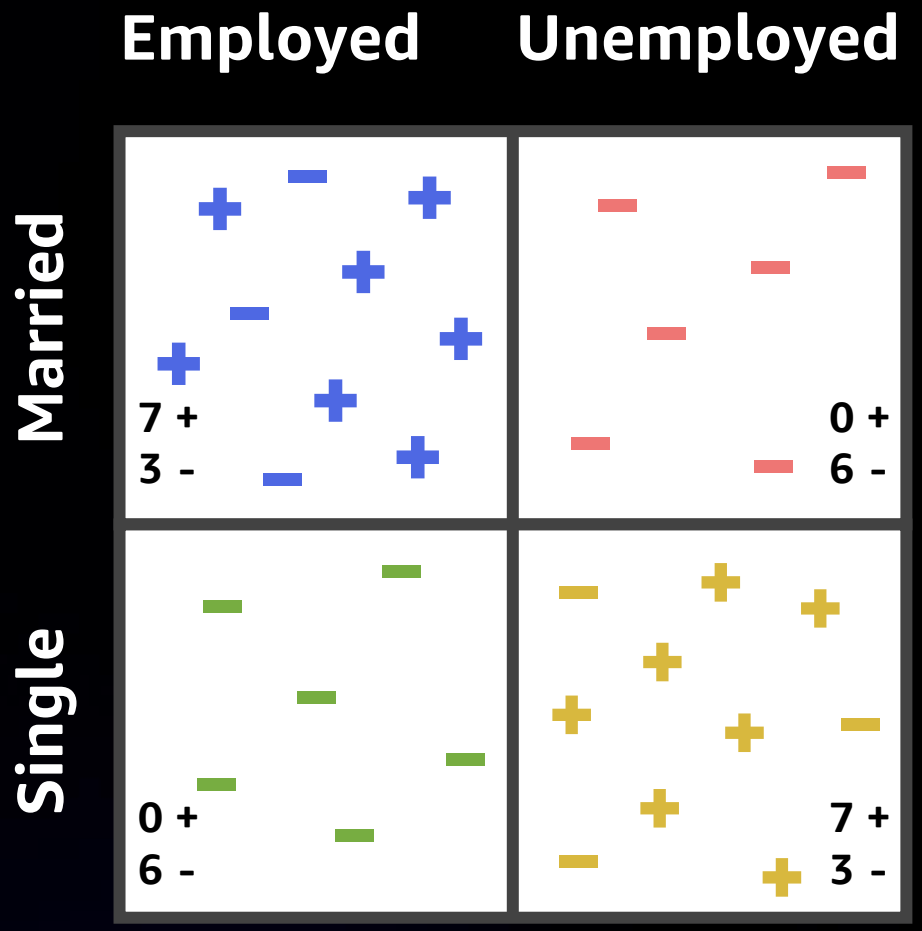

Example Admission: Accepted/Rejected

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Intersectional fairness

Subgroups may experience additional disparity (see positive/negative outcomes per subgroup)

How granular should we define the groups?




Example Admission: Approved/Denied

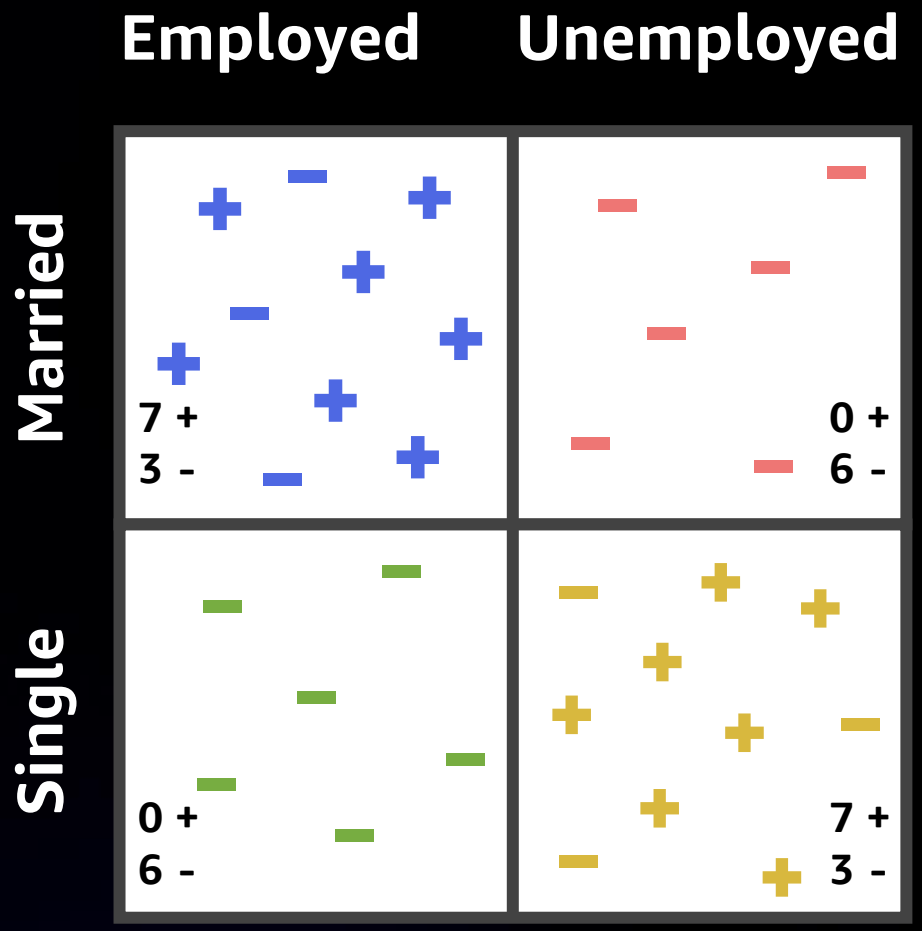
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Intersectional fairness

Subgroups may experience additional disparity (see positive/negative outcomes per subgroup)

Challenge

- Can’t build models for separate groups, this constitutes disparate treatment (in the US)!




Example Hiring: Approved/Denied

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Why is responsible AI complex? How can we succeed?

| | | |
|---|---|---|
| Success is **use case** specific | → | Simplify defining and measuring success; include bias metrics in evaluation |
| Rapid changes in **AI landscape** (users, technologies, companies, societies; e.g. AI Act in EU) | → | Improve continuously over the long term; adopt best practices now |
| Technical **complexity** (for example, root cause analysis of bias) | → | Make responsible AI part of all DevOps/ML Lifecycle steps (next topic) |

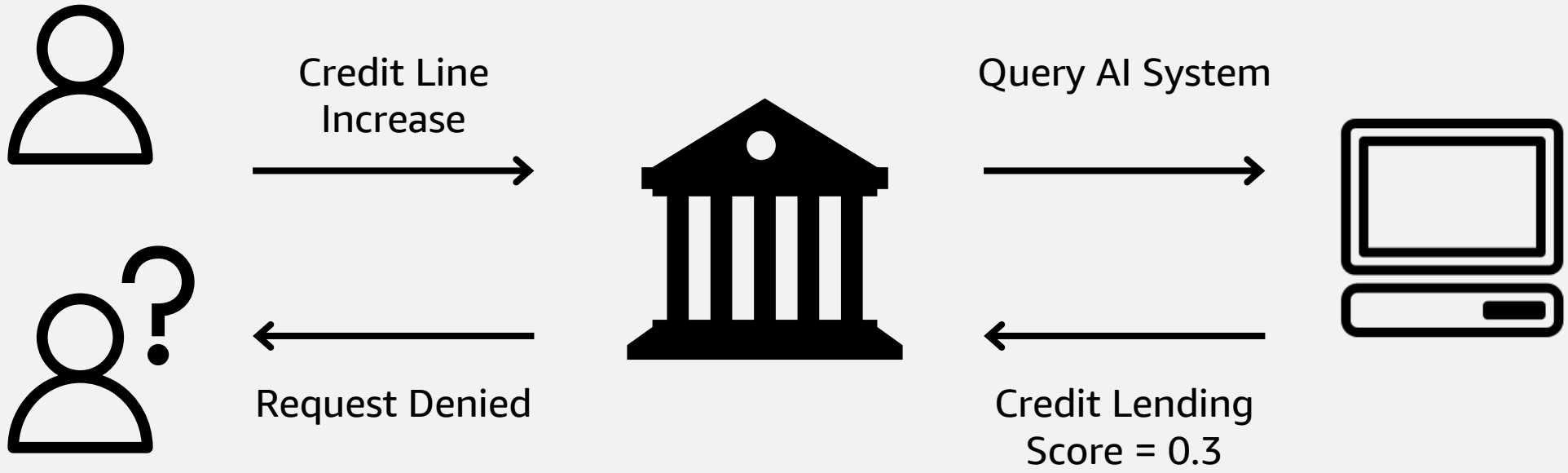
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Fairness throughout the ML lifecycle

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

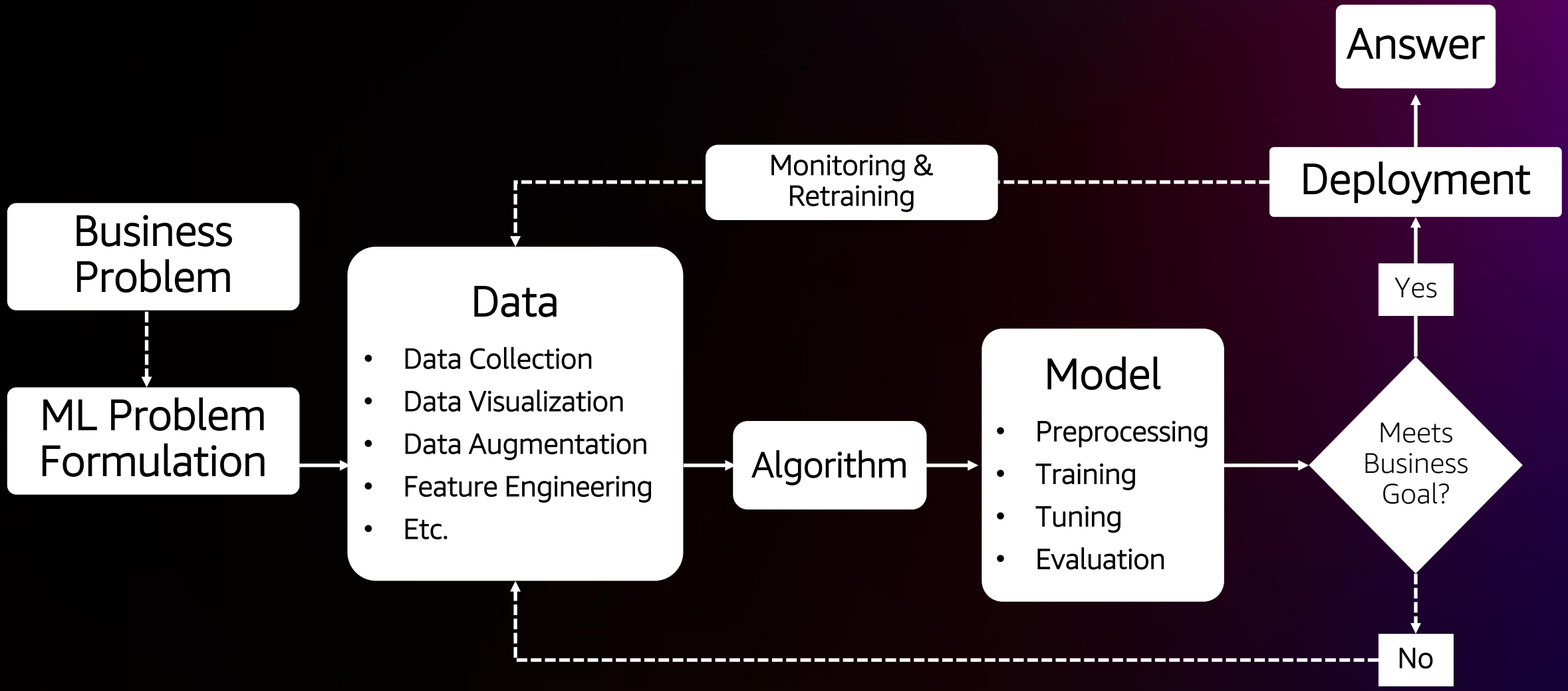
# Credit line increase example




Fair lending laws [ECOA, FCRA] require credit decision to be explainable

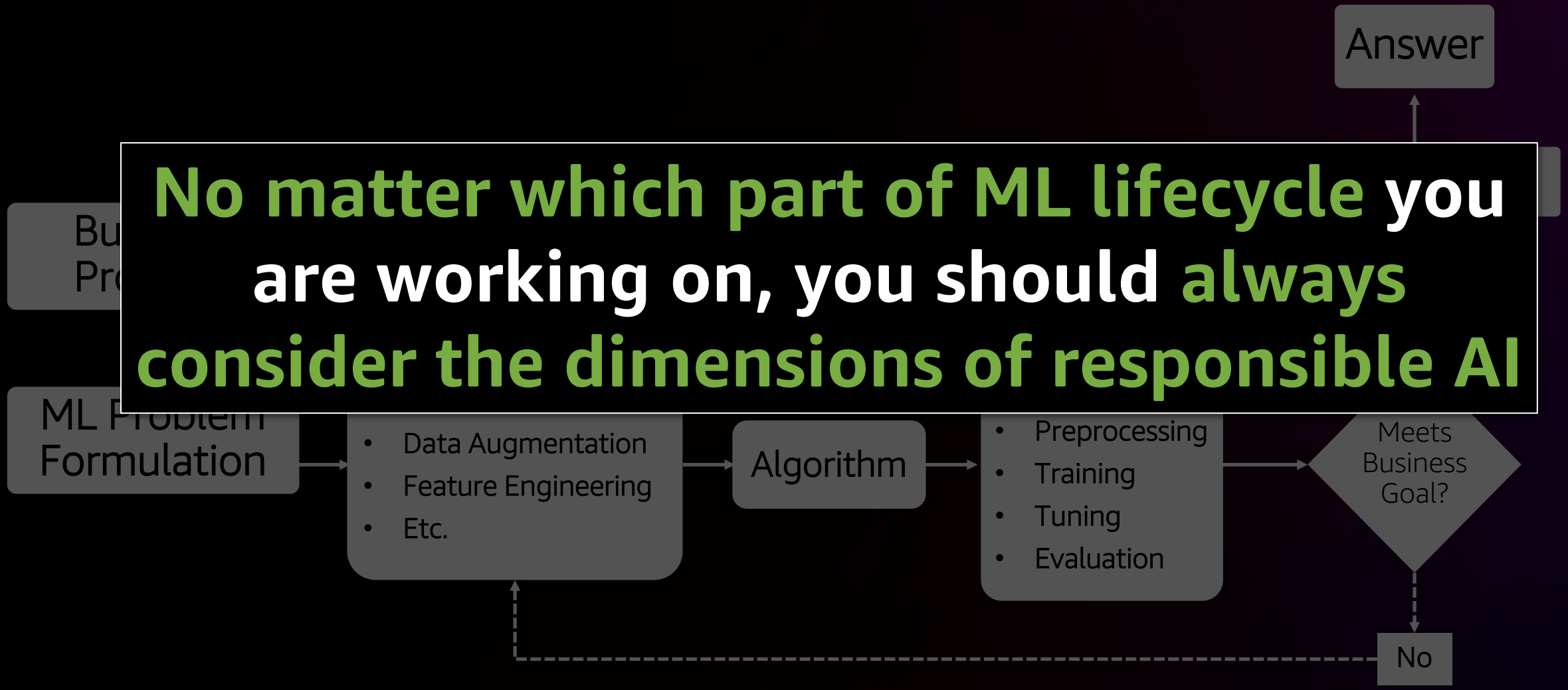
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Where to consider dimensions of RAI?



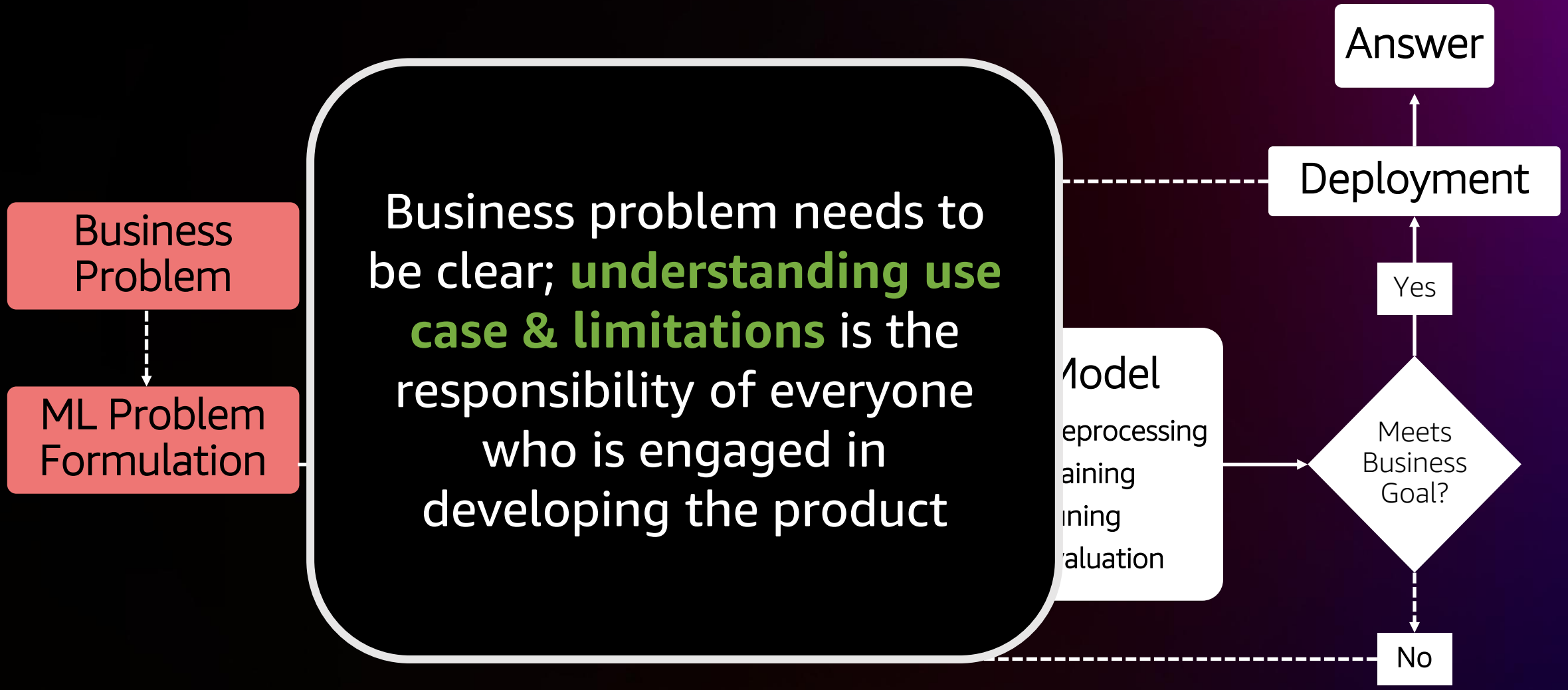

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Everywhere ☺



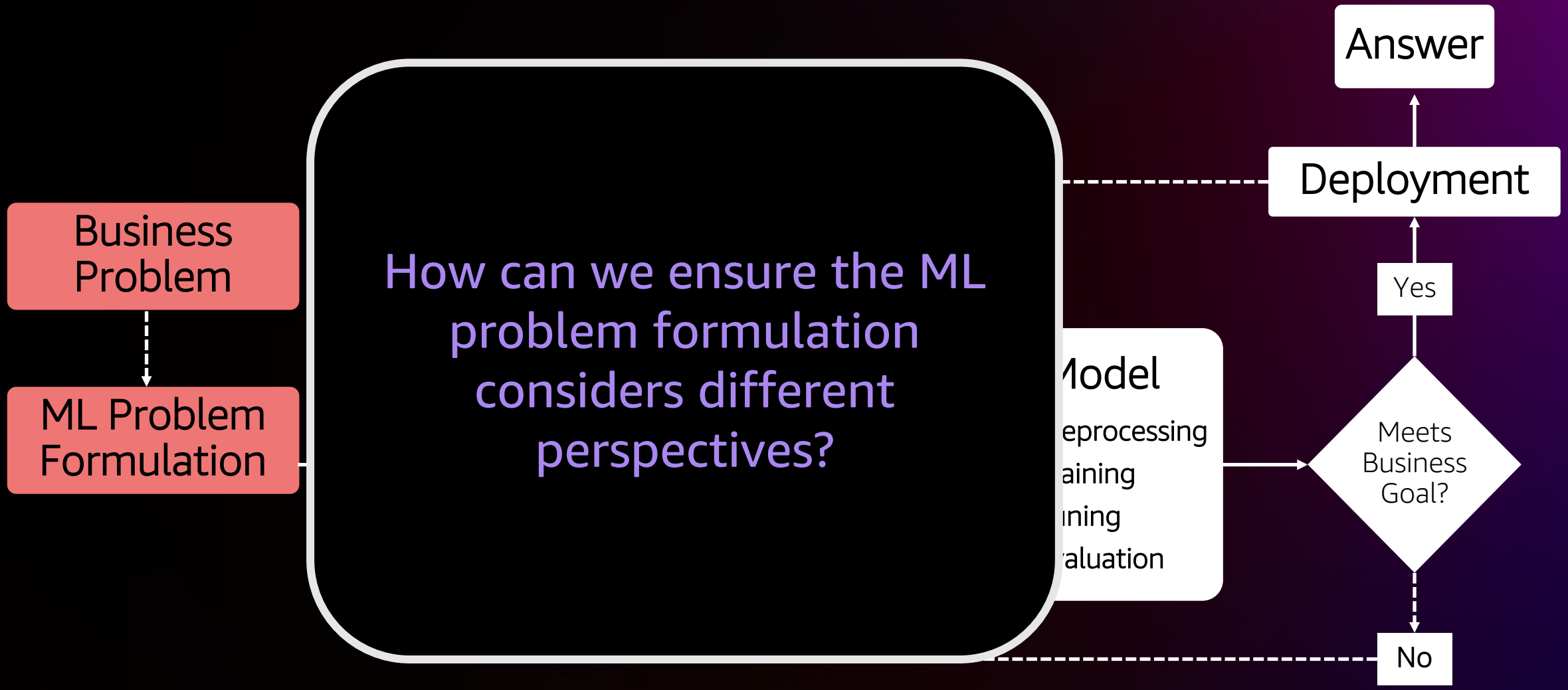

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Business problem and ML problem formulation




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Business problem and ML problem formulation



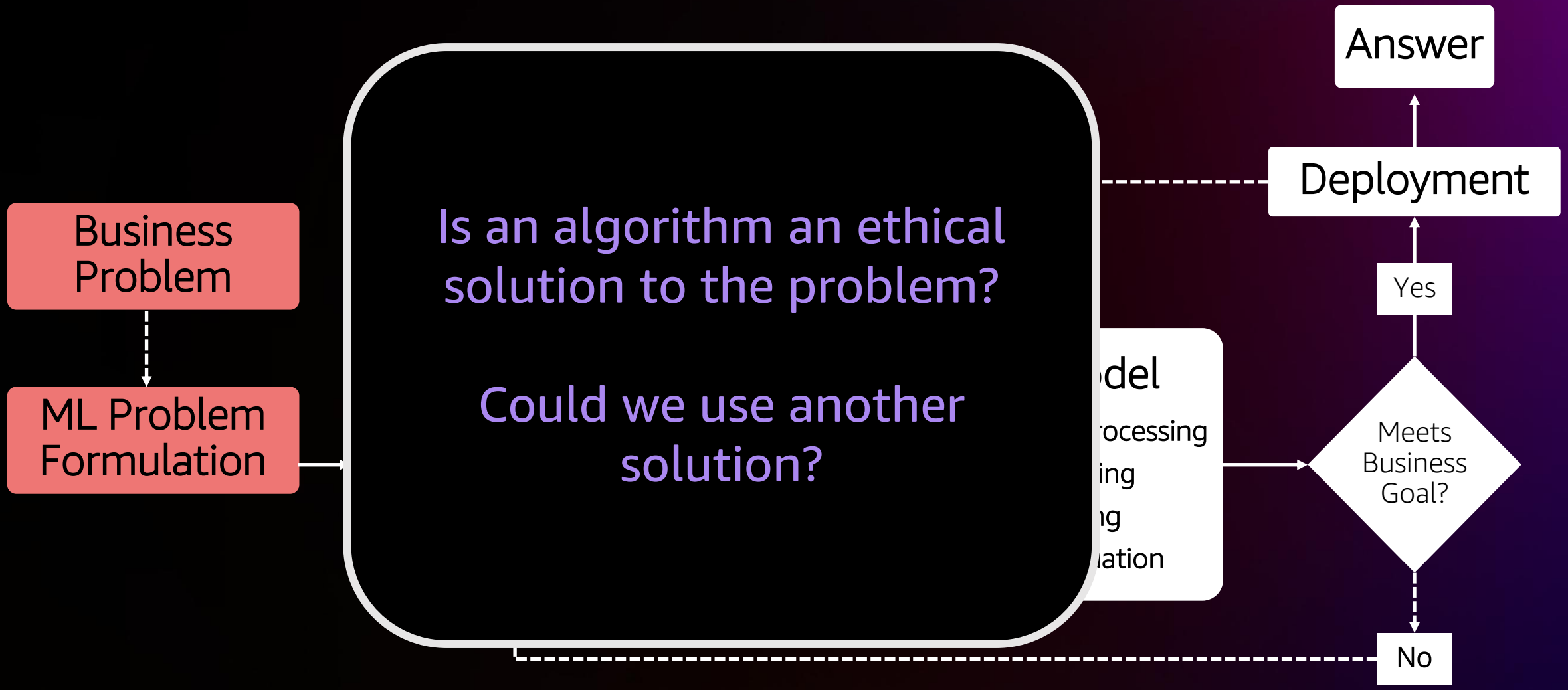

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Business problem and ML problem formulation



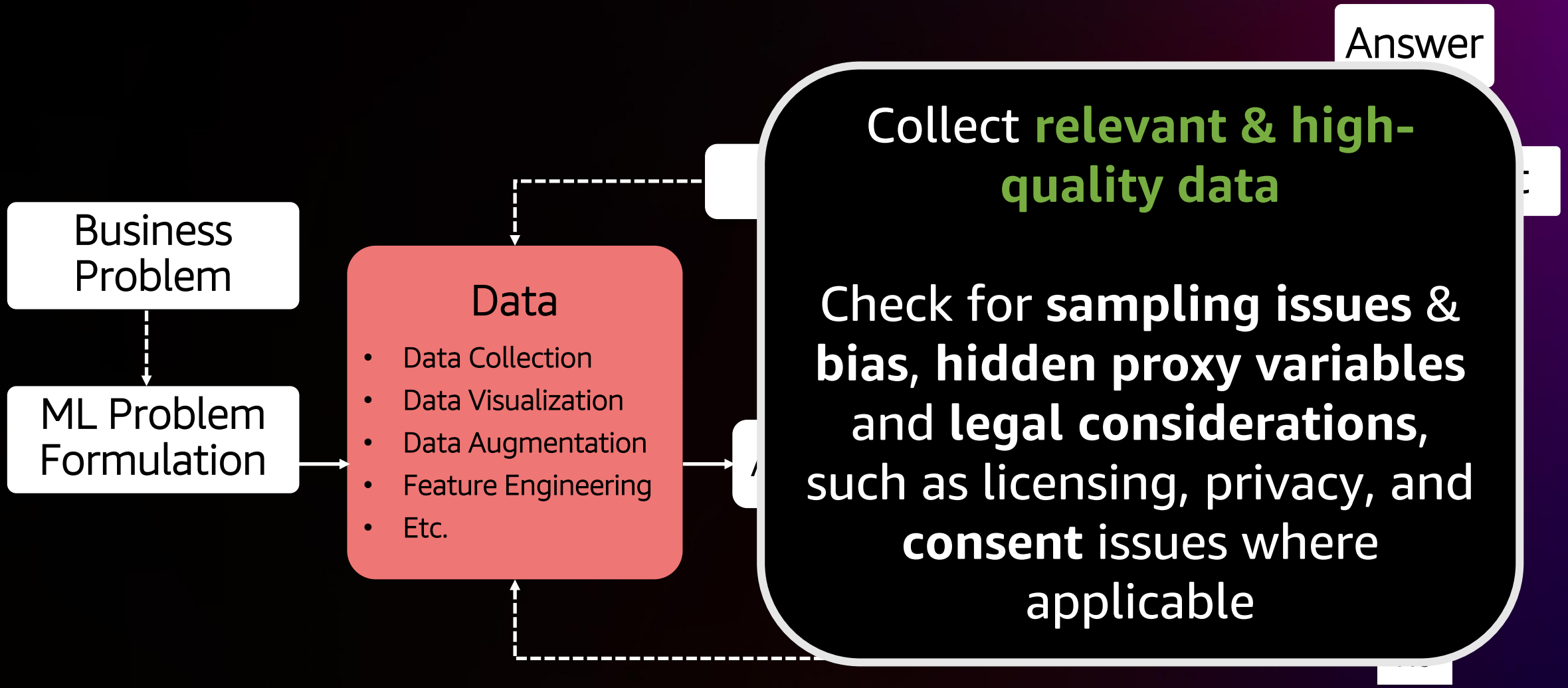

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Business problem and ML problem formulation



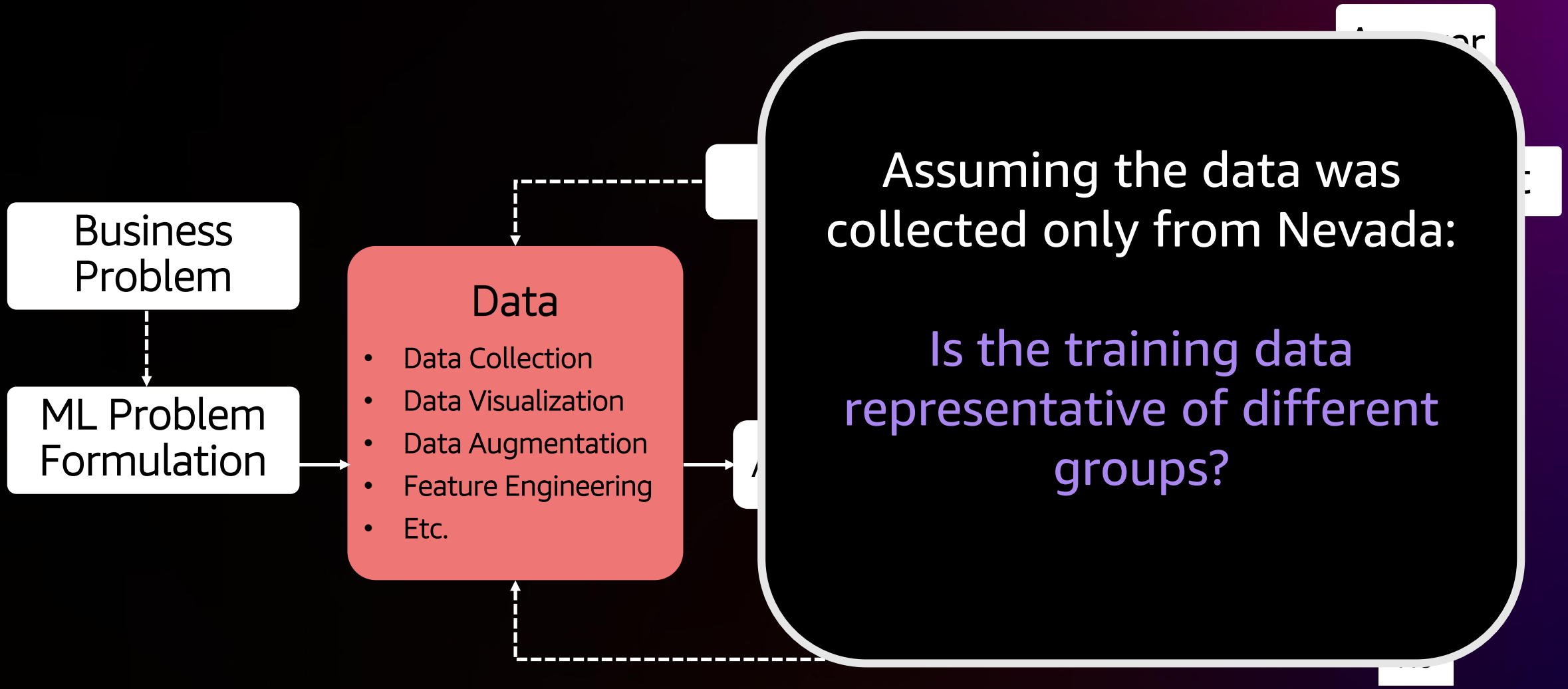

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Data collection and feature engineering



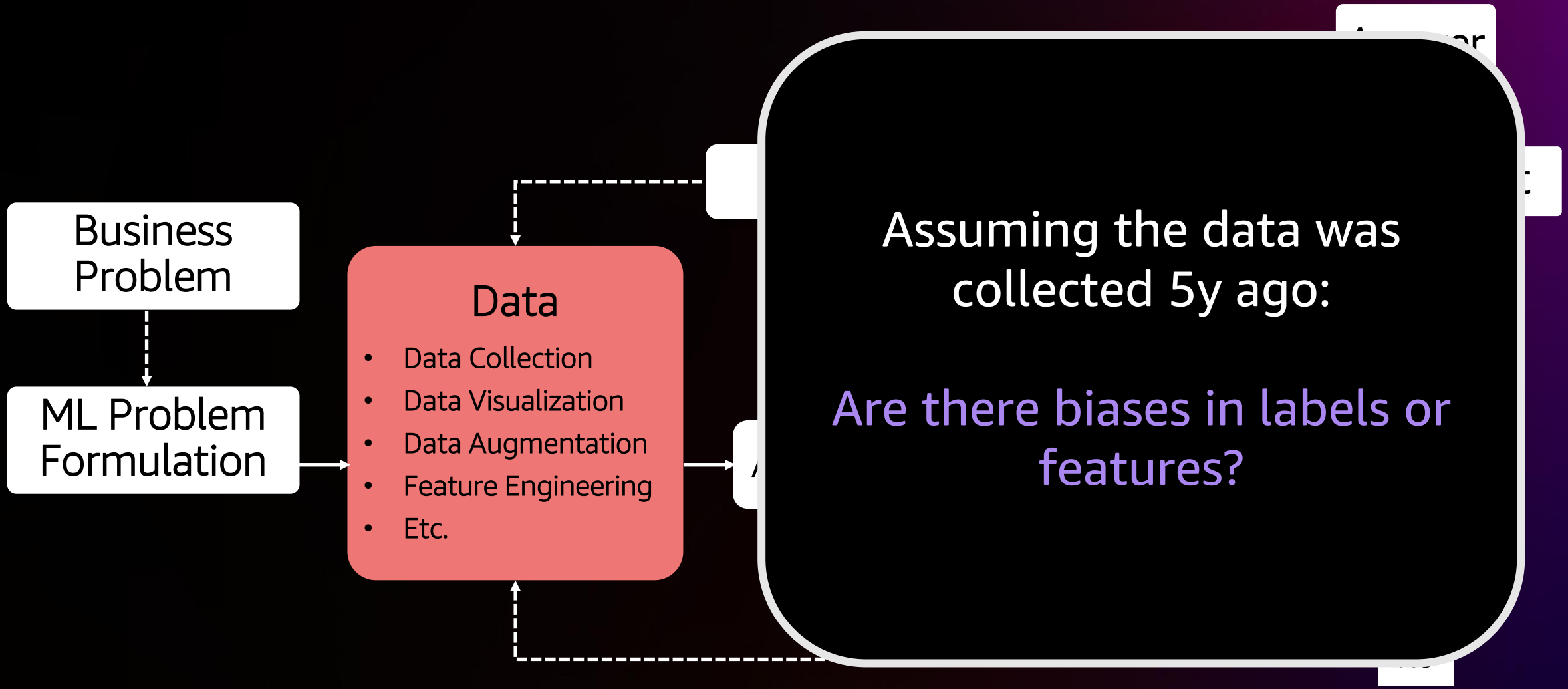

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Data collection and feature engineering




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Data collection and feature engineering




aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Biased datasets

Datasets can be biased:

- Historical or systematic bias is often present in data collected from “the wild”
- Geographical representation is not equal for many of the most widely used (computer vision) datasets
- …

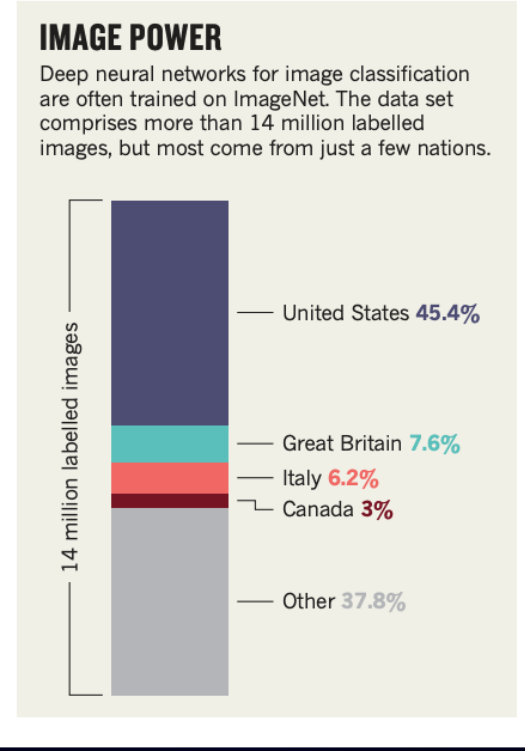
Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary’s Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

**Abstract**

The blind application of machine learning runs the risk of amplifying b
danger is facing us with *word embedding*, a popular framework to represe
has been used in many machine learning and natural language processin
word embeddings trained on Google News articles exhibit female/male gend
extent. This raises concerns because their widespread use, as we describe,
biases. Geometrically, gender bias is first shown to be captured by a direc
Second, gender neutral words are shown to be linearly separable from gende
embedding. Using these properties, we provide a methodology for modify
gender stereotypes, such as the association between between the words *r*
maintaining desired associations such as between the words *queen* and *f*
quantify both direct and indirect gender biases in embeddings, and develo
embedding. Using crowd-worker evaluation as well as standard benchmarks
that our algorithms significantly reduce gender bias in embeddings while pres
such as the ability to cluster related concepts and to solve analogy tasks. T
be used in applications without amplifying gender bias.

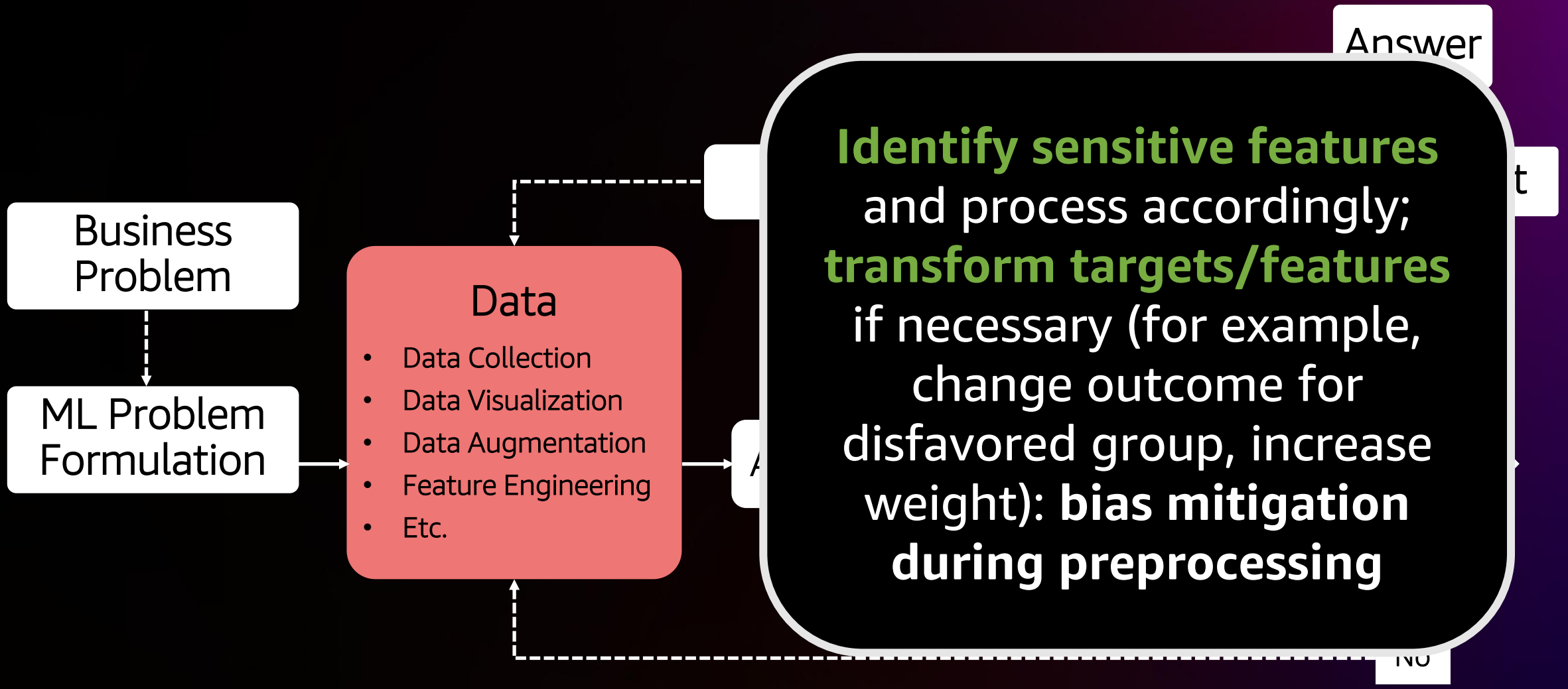
https://arxiv.org/abs/1607.06520




https://www.nature.com/nature/volumes/559/issues/7713

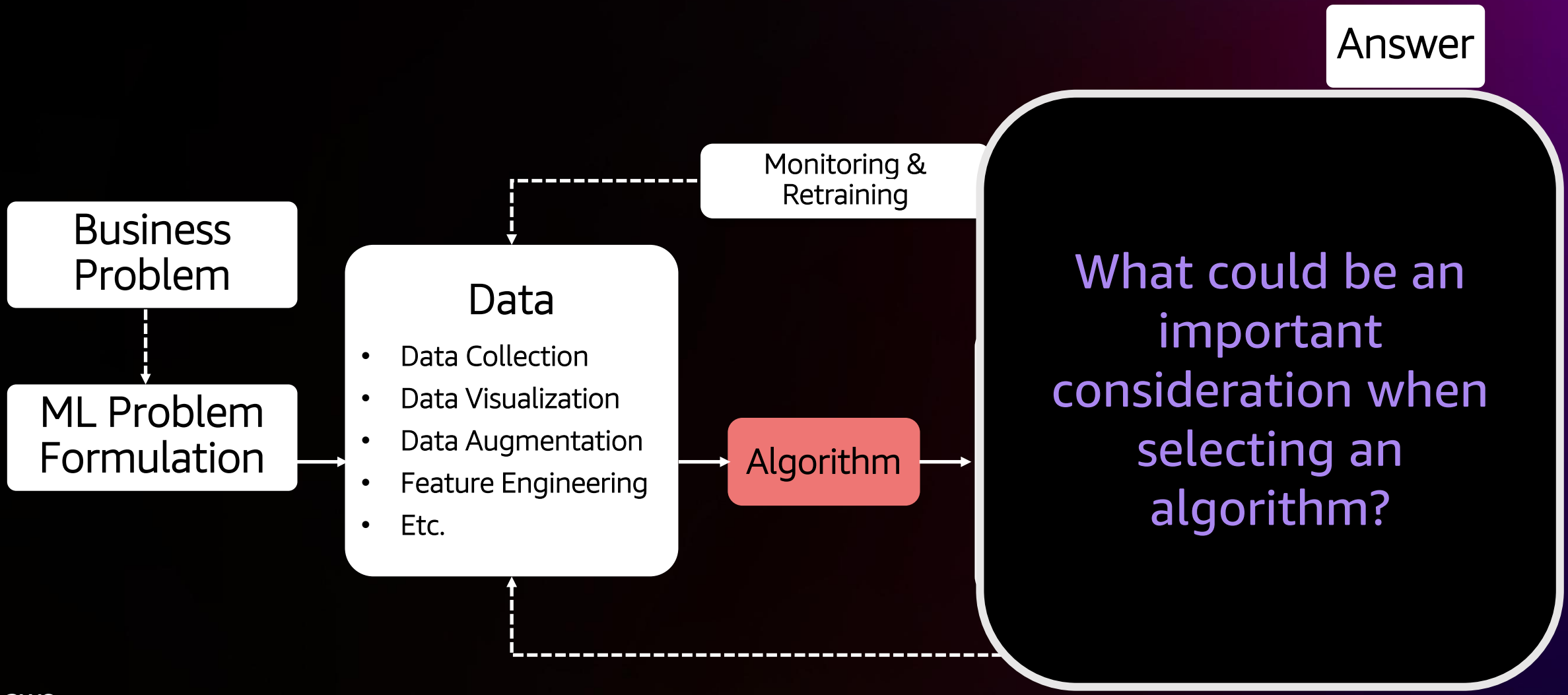
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Data collection and feature engineering



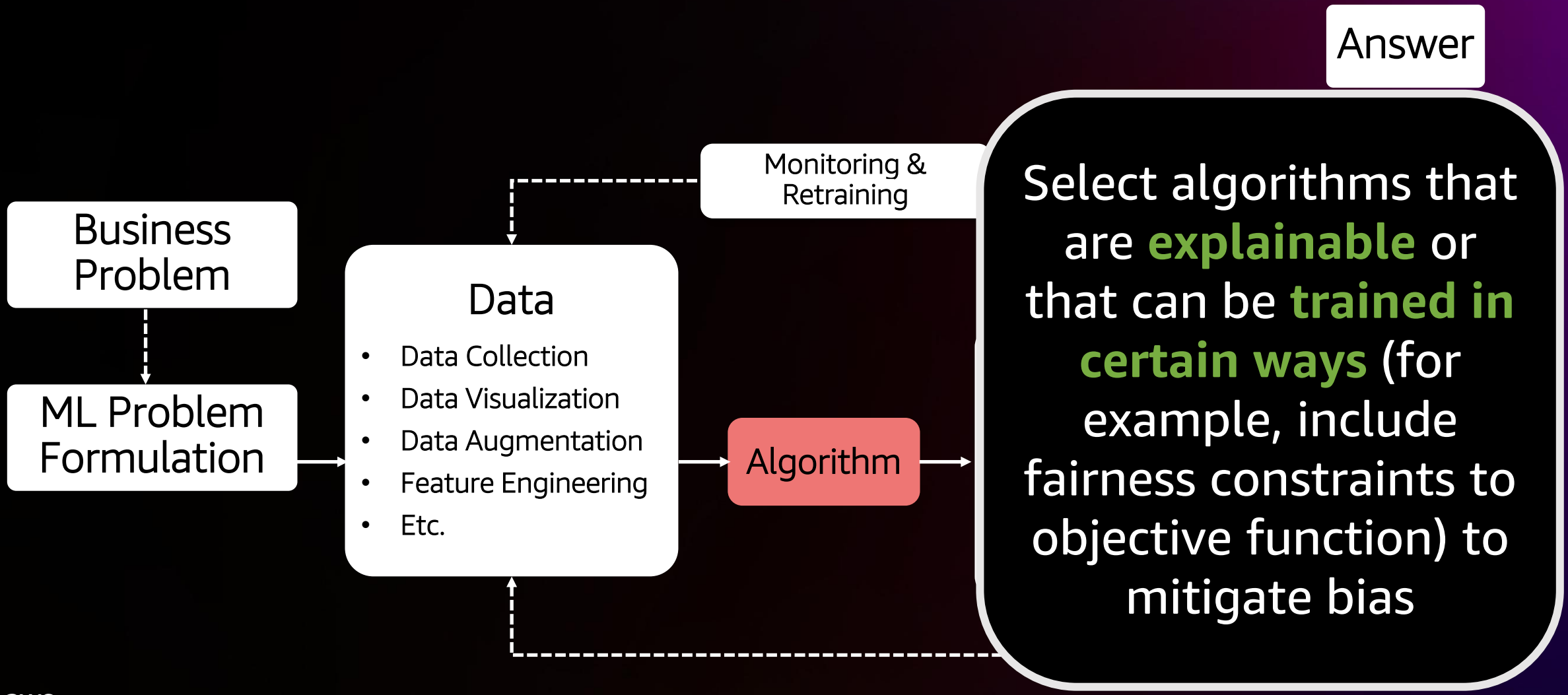

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Algorithm selection



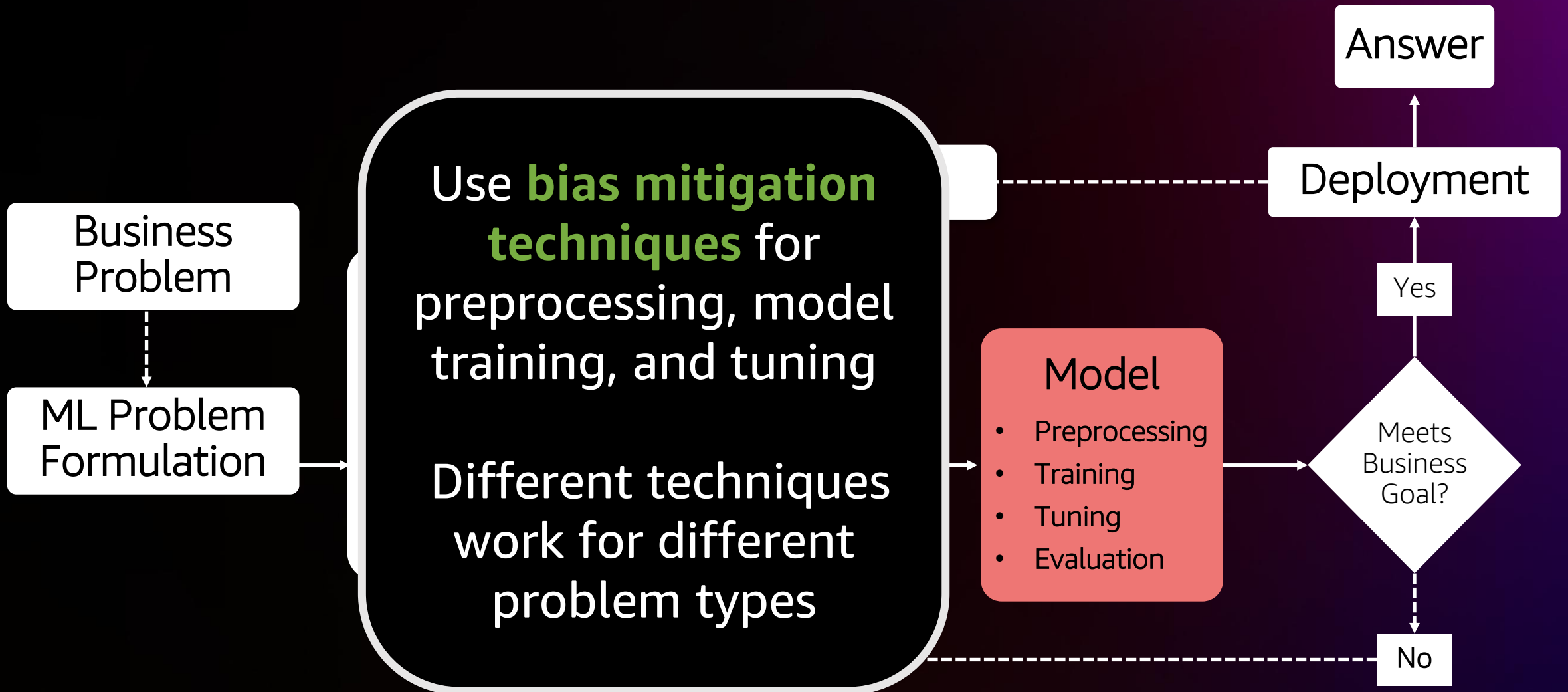

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Algorithm selection



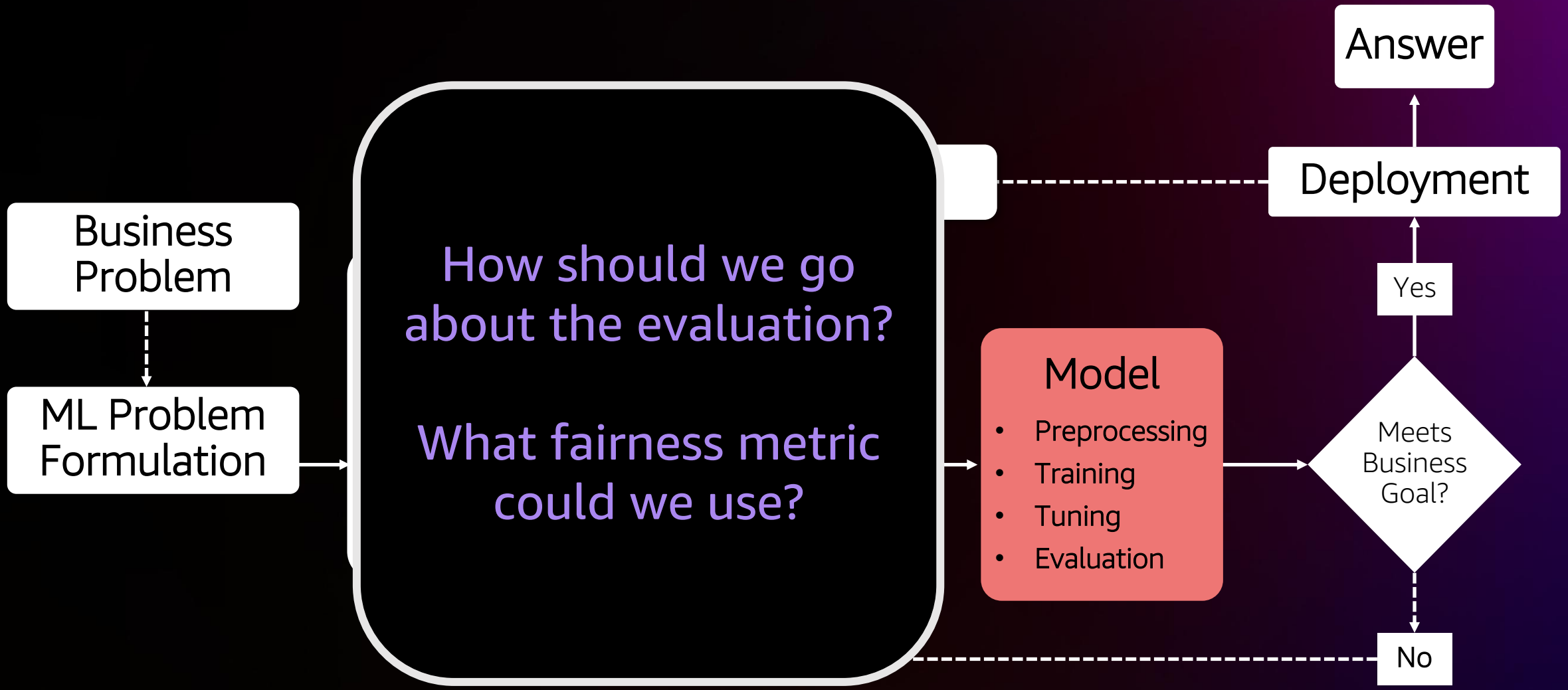

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Bias mitigation



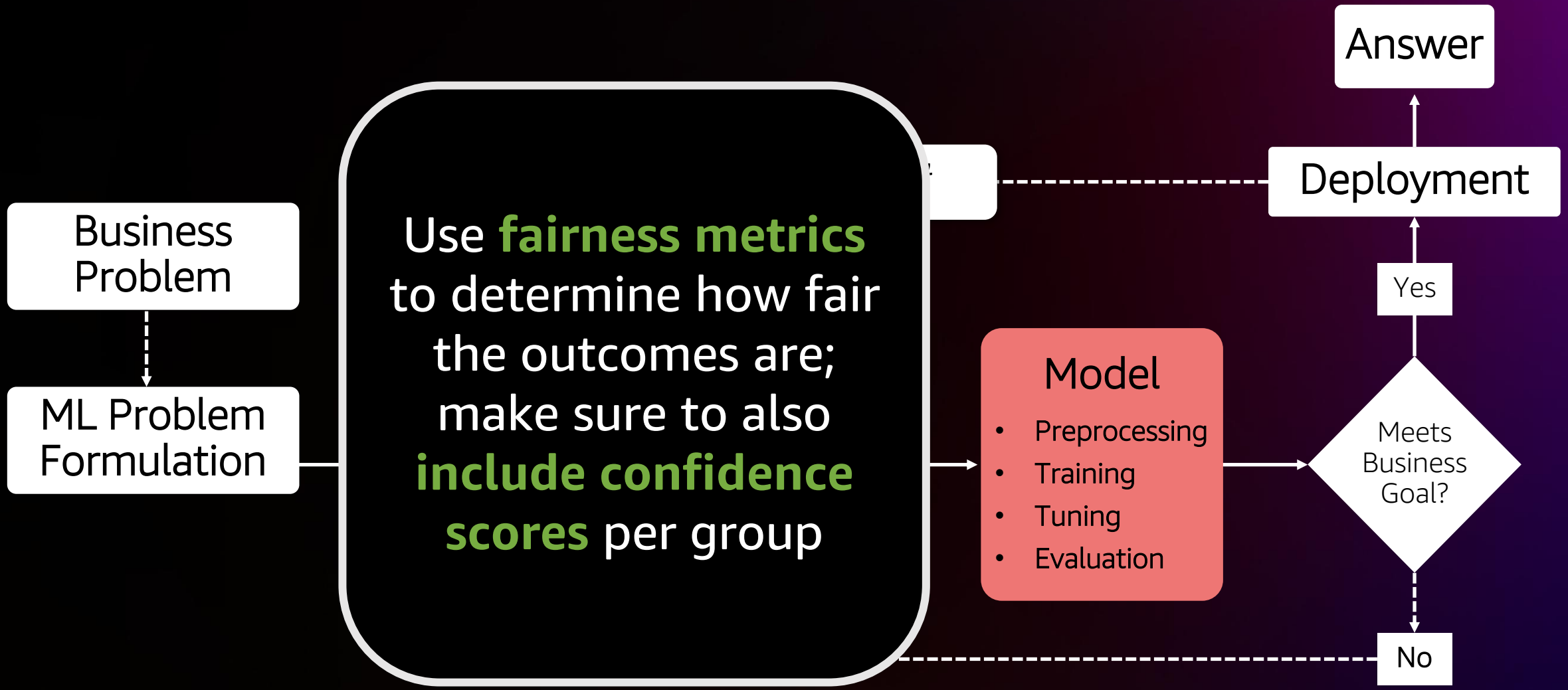

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Model evaluation



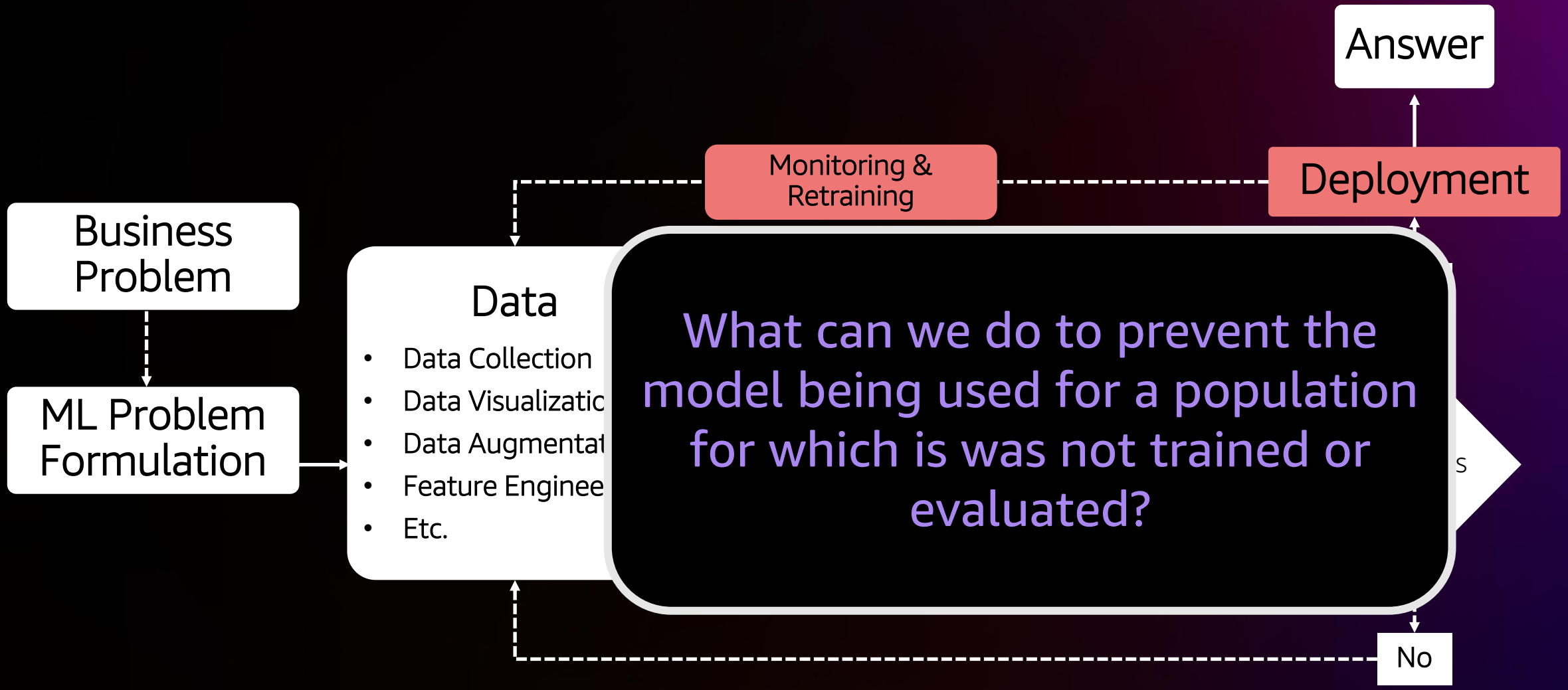

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Model evaluation



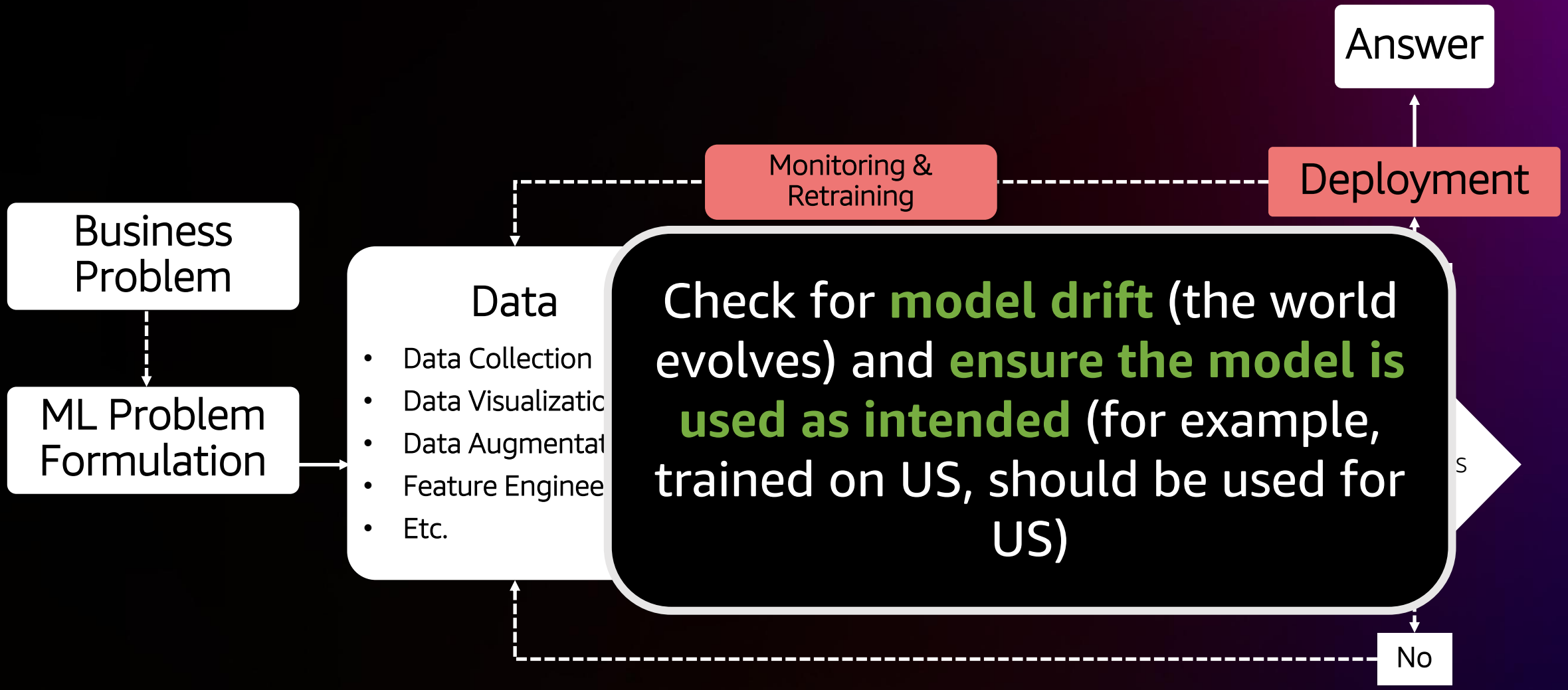

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deployment and retraining



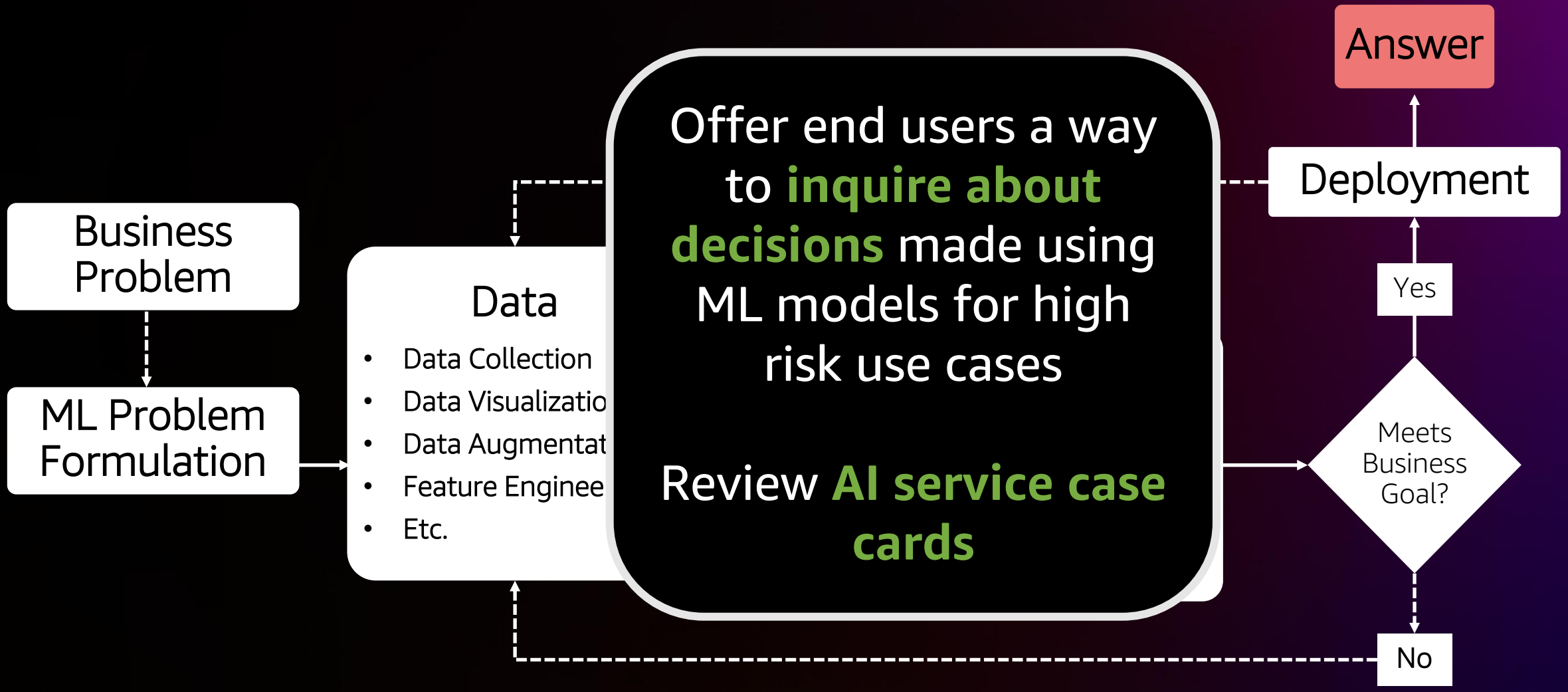

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Deployment and retraining




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Model predictions




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

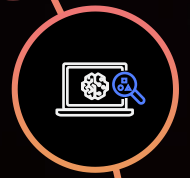
# Amazon SageMaker Clarify

Detect bias in ML models and understand model predictions



## Identify imbalances in data

Detect bias during data preparation



## Check your trained model for bias

Evaluate the degree to which various types of bias are present in your model



## Explain overall model behavior

Understand the relative importance of each feature to your model's behavior



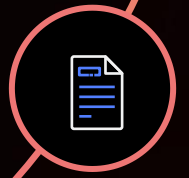
## Explain individual predictions

Understand the relative importance of each feature for individual inferences



## Detect drift in bias and model behavior over time

Provide alerts and detect drift over time due to changing real-world conditions



## Generate automated reports

Produce reports on bias and explanations to support internal presentations

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS AI Service Cards

NEW!

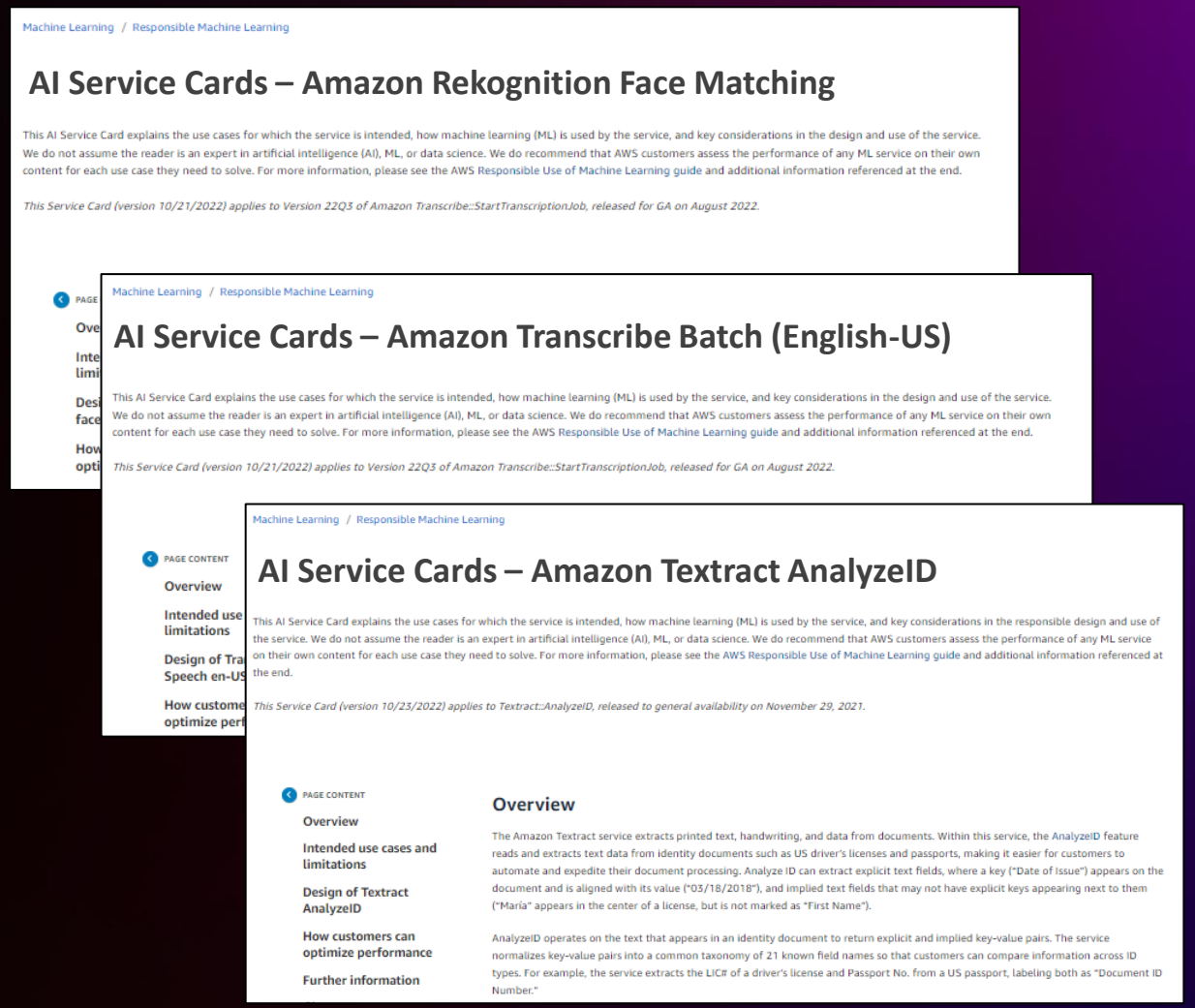
## New transparency resource to advance responsible AI

- Documents the intended use cases and fairness considerations of our AWS AI services
- Reflects our comprehensive development process
- Three new AI Service Cards published

**Amazon Rekognition Face Matching**

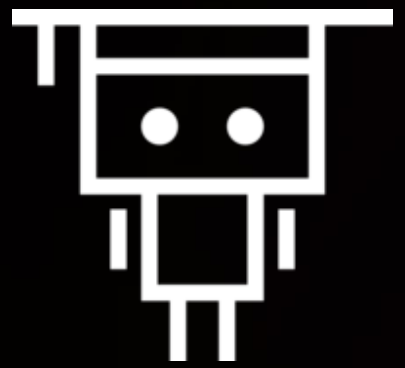
**Amazon Textract AnalyzeID**

**Amazon Transcribe – Batch (English-US)**




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# New fairness and bias mitigation course



**Machine Learning University**

**N E W !**

Free, public course on fairness criteria and bias mitigation in the ML lifecycle

Taught by Amazon data scientists who train AWS employees on ML

9+ hours of lectures and hands-on exercises

**>>> Get started today**

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# What next?

- Investigate tools that can help measure & mitigate bias
- Share responsible AI principles in your team/organization
- Nominate individuals that can implement responsible AI principles in your team/organization

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Q&A

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Check out other sessions on responsible AI at re:Invent

| | | | |
|---|---|---|---|
| DEI201-L | **The power of responsibility in AI/ML**<br>Leadership session | Monday, 11/28 | 2:30–3:30 PM |
| AIM328 | **Explain your ML models with Amazon SageMaker Clarify**<br>Chalk talk | Tuesday, 11/29 | 2:00–3:00 PM |
| DEI205 | **A Geographic Perspective on Responsible AI**<br>Breakout session | Tuesday, 11/29 | 5:45–6:45 PM |
| AIM213 | **Building more inclusive, responsible AI**<br>Breakout session (panel) | Wednesday, 11/30 | 8:30–9:30 AM |
| AIM342 | **Advancing responsible AI: Bias assessment and transparency**<br>Workshop | Wednesday, 11/30 | 2:30–4:30 PM |

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

Nashlie Sephus
nashlies@amazon.com

Mia Mayer
mimayer@amazon.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.