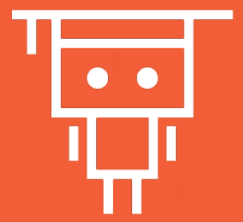


MACHINE LEARNING
UNIVERSITY

# Responsible AI

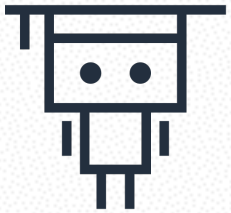
## Fairness & Bias in ML – DAY 1

# Learning Outcomes

- Fundamental understanding of Machine Learning:
  - Concepts & terminology
- Practical ML skills and techniques:
  - Train, tune, test and evaluate simple ML models
  - Check data and ML model for bias
- How to identify and mitigate bias and fairness issues in ML

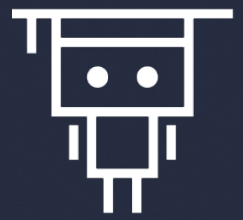
# Course Schedule

| Day One | Day Two | Day Three |
|---|---|---|
| Fundamentals of Machine Learning | Data Processing | Bias Mitigation during Model Training |
| Introduction to Fairness & Bias Mitigation in ML | ML Algorithm Selection, Model Build & Evaluation | Bias Mitigation during Post-Processing |
| Model Formulation & Data Collection | Fairness Criteria | Bias Mitigation for Models in Production |
| Exploratory Data Analysis | Bias Mitigation during Pre-Processing | Explainability |



MACHINE LEARNING UNIVERSITY

# Introduction to Machine Learning (ML)



# What is Machine Learning?

# Traditional: Rules-based

# What is Machine Learning?

Data
Rules

→ Hard-coded rules get applied to existing data → Answers

> "Programming computers to learn from experience should eventually eliminate the need for […] detailed programming effort."
>
> *Arthur Samuel, 1959*
> Pioneer of AI research

Data
ML Algorithm

→ Model applies rules learnt by algo to new data → Answers (predictions)

↑ New Data

MACHINE LEARNING UNIVERSITY

# What is Machine Learning?

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
>
> *Tom Mitchell, 1997*

# Data: Features

Usually denoted as $x$ (for single feature) and $X$ for collection of features.

- Tabular: Raw values are strings or number.

| Experience [years] | Age | Gender |
|---|---|---|
| 7 | 30-40 | Male |
| 7 | 20-30 | NA |
| 7 | 30-40 | Female |
| … | … | … |
| 10 | 40-50 | Male |
| 9 | 20-30 | * |
| 6 | 20-30 | Male |

Table 1: Example tabular dataset

# Data: Features

Usually denoted as $x$ (for single feature) and $X$ for collection of features.

- Tabular: Raw values are strings or number.

| $x_1$ | $x_2$ | … |
|---|---|---|
| Experience [years] | Age | Gender |
| 7 | 30-40 | Male |
| 7 | 20-30 | NA |
| 7 | 30-40 | Female |
| … | … | … |
| 10 | 40-50 | Male |
| 9 | 20-30 | * |
| 6 | 20-30 | Male |

Table 1: Example tabular dataset

# Data: Features

Usually denoted as $x$ (for single feature) and $X$ for collection of features.

- Tabular: Raw values are strings or number.
- Image: Images where raw values are numbers (light intensity in pixels).

| Name | Folder | Type | Size |
|---|---|---|---|
| curriculum-vitae-1.png | - | image/png | 70.3 KB |
| curriculum-vitae-10.png | - | image/png | 70.3 KB |
| curriculum-vitae-11.png | - | image/png | 70.3 KB |
| curriculum-vitae-12.png | - | image/png | 70.3 KB |
| curriculum-vitae-13.png | - | image/png | 70.3 KB |
| curriculum-vitae-14.png | - | image/png | 70.3 KB |
| curriculum-vitae-15.png | - | image/png | 70.3 KB |
| curriculum-vitae-16.png | - | image/png | 70.3 KB |
| curriculum-vitae-17.png | - | image/png | 70.3 KB |
| curriculum-vitae-18.png | - | image/png | 70.3 KB |

Table 2: Example image dataset

# Data: Features

Usually denoted as $x$ (for single feature) and $X$ for collection of features.

- Tabular: Raw values are strings or number.
- Image: Images where raw values are numbers (light intensity in pixels).
- Language: Raw values are strings (text snippets or words).

| EN (source) | DE (target) | Tag |
|---|---|---|
| The doctor asked the nurse to hand her the scalpel. | Die Ärztin bat die Krankenschwester <>ihm<> das Skalpell zu reichen. | False Male |
| The pilot listed her credentials in the job interview. | Die Pilotin hat <>ihre<> Zeugnisse im Bewerbungsgespräch aufgeführt. | True Female |
| … | … | |

Table 3: Example language dataset for translation task

# Data: Features/Input

Usually denoted as $x$ (for single feature) and $X$ for collection of features.

- Tabular: Raw values are strings or number.
- Image: Images where raw values are numbers (light intensity in pixels).
- Language: Raw values are strings (text snippets or words).
- Multimodal: Combination of all the above.

| Industry | Job Description | imgID | # Applications |
|---|---|---|---|
| IT | High-paced… | 51QjPP6oAVL | 21 |
| Health | Shared passion… | 31WGVoL8mlL | 10 |
| … | … | … | … |
| IT | Digital native… | 41c3RVsWGfL | 43 |
| Beauty | Collaborate on… | 51OMvEHU2QL | 39 |

Table 4: Example multimodal dataset

# Data: Labels

Usually denoted as $y$: The answer we want to generate using a trained model.

Labels are not always provided, but when they exist, they are either:

- Numerical values (e.g. insurance price, salary prediction…)

| $x_1$ | $x_2$ | … | $y$ |
|---|---|---|---|
| Experience [years] | Age | Gender | Salary per annum [k] |
| 7 | 30-40 | Male | $181 |
| 7 | 20-30 | NA | $130 |
| 7 | 30-40 | Female | $199 |
| … | … | … | … |
| 10 | 40-50 | Male | $252 |
| 9 | 20-30 | * | $90 |
| 6 | 20-30 | Male | $166 |

# Data: Labels

Usually denoted as $y$: The answer we want to generate using a trained model.

Labels are not always provided, but when they exist, they are either:

- Numerical values (e.g. insurance price, salary prediction…)
- Categorical values (e.g. loan approved: yes/no, type of disease…)

| $x_1$ | $x_2$ | … | $y$ |
|---|---|---|---|
| Experience [years] | Age | Gender | Accepted |
| 7 | 30-40 | Male | yes |
| 7 | 20-30 | NA | no |
| 7 | 30-40 | Female | no |
| … | … | … | … |
| 10 | 40-50 | Male | yes |
| 9 | 20-30 | * | no |
| 6 | 20-30 | Male | yes |

# ML Algorithm: Example

prediction ~ weighted combination of features

Goal: Obtain healthcare score
for individual.

# ML Algorithm: Example

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + w_1 x_1 + \cdots + w_4 x_4 + \cdots + w_n x_n$$

# ML Algorithm: Features

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + w_1 x_1 + \cdots + w_4 x_4 + \cdots + w_n x_n$$

Feature: Measurable piece of information (in numerical representation)

# ML Algorithm: Features

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + w_1 5 + \cdots + w_4 x_4 + \cdots + w_n x_n$$

Feature: Measurable piece of information (in numerical representation)

$x_1$: Feature for "*Number of Admissions*"

$x_1 = 5$

# ML Algorithm: Weights

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + w_1 5 + \cdots + w_4 x_4 + \cdots + w_n x_n$$

Weight: Determines how much influence a given feature has on the output (prediction)

Weights (also called parameters) are learnt during *model training stage*.

# ML Algorithm: Weights

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + 100 * 5 + \cdots + w_4 x_4 + \cdots + w_n x_n$$

Weight: Determines how much influence a given feature has on the output (prediction)

$w_1$: *Weight for "Number of Hospital Admissions" feature*

$w_1 = 100$

# ML Algorithm: Features & Weights

prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx w_0 x_0 + 100 * 5 \ldots - 25 * 1 + \cdots + w_n x_n$$

$x_4$: Feature for "*Healthy lifestyle*"

$x_4$ = Yes = 1

$w_4$: *Weight for "Healthy lifestyle" feature*

$w_4 = -25$

# Output

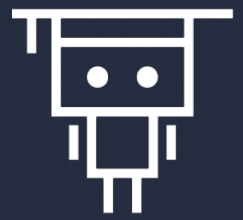
prediction ~ weighted combination of features

$$\hat{y} = f(x) \approx 200 + 100 * 5 - 25 * 1 = 675 \pm err$$

(predicted healthcare score)

*Remember: This is a simplified example of a simple ML algorithm (**Linear Regression**).*

$x_0$ = 1 per convention, $a_0$ = 200 (healthcare score base value)



# Types of Machine Learning

# Supervised vs. Unsupervised Learning

Supervised Learning

Data includes labels

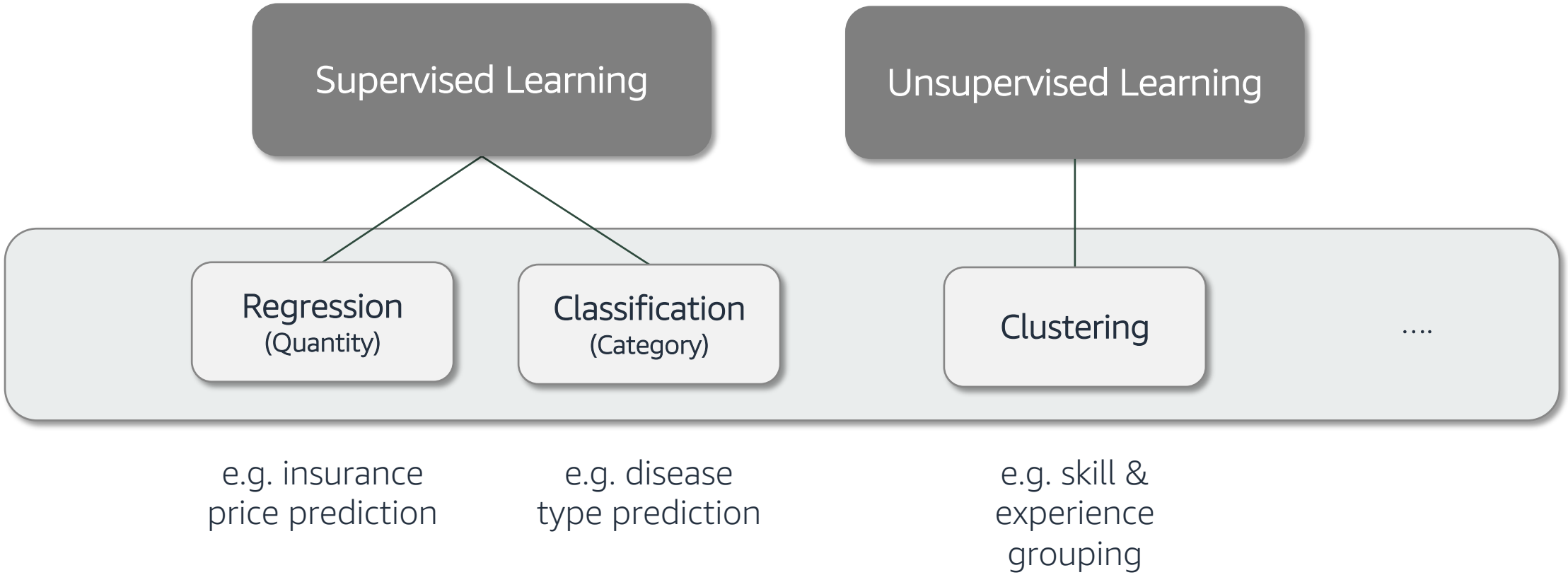
Model learns by looking at these examples

Unsupervised Learning

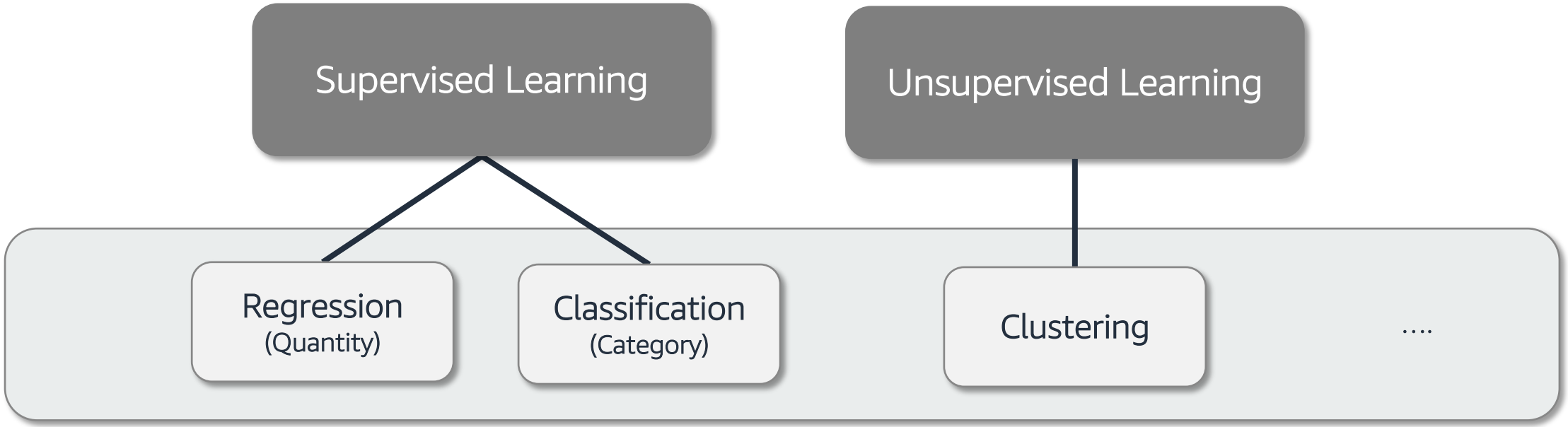
Data is provided without labels

Model finds patterns in data

# Supervised vs. Unsupervised Learning



# Supervised vs. Unsupervised Learning

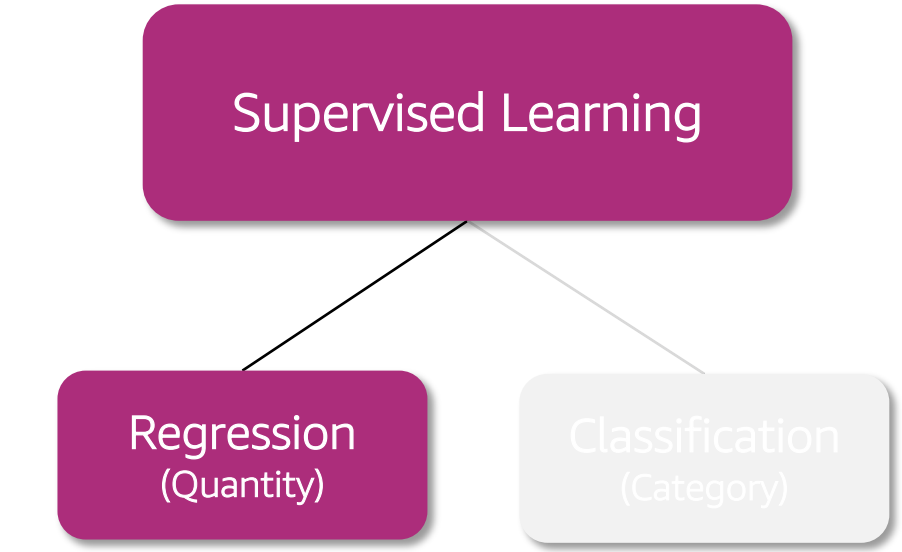



**ML Algorithms**

| | | |
|---|---|---|
| • Linear Regression | • Support Vector Machine | • PCA |
| • K-Nearest Neighbors | • K-Nearest Neighbors | • Collaborative Filtering |
| • Neural Nets | • Neural Nets | • K-Means |
| • Decision Trees | • Decision Trees | … |
| … | … | |

MACHINE LEARNING UNIVERSITY

# Supervised Learning: Regression

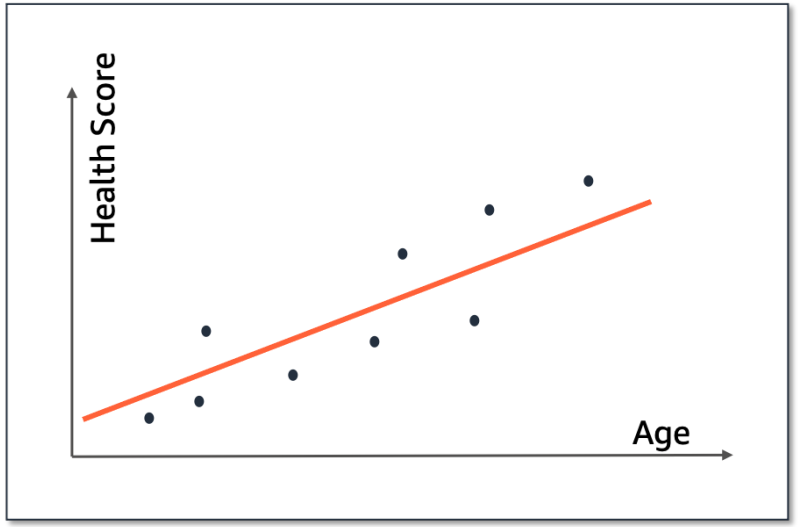



- Linear Regression
- K-Nearest Neighbors
- Neural Nets
- Decision Trees

...

- Support Vector Machine
- K-Nearest Neighbors
- Neural Nets
- Decision Trees

...

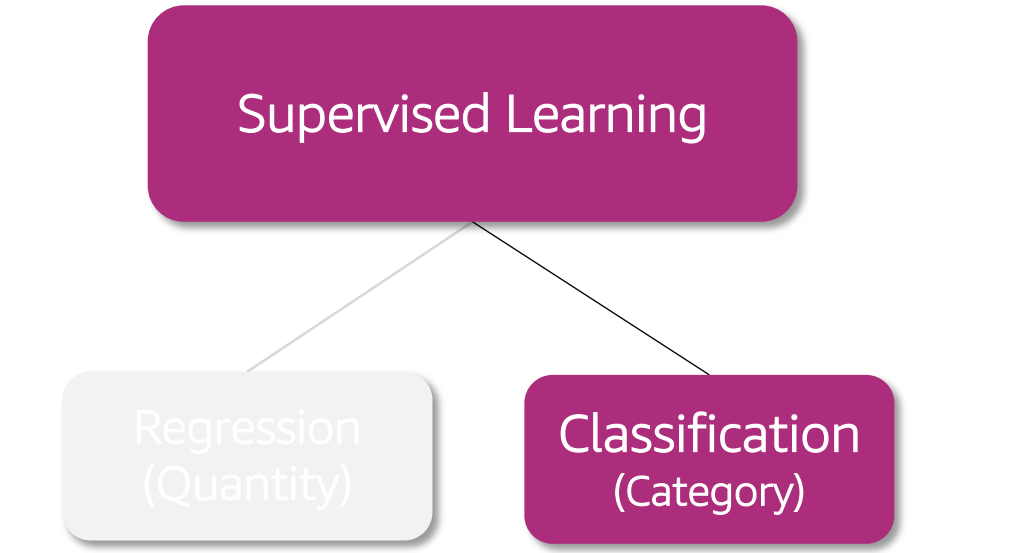



Label | Features

| HS | Age | BMI | Smoker | Ethnicity |
|---|---|---|---|---|
| 80 | 32 | 28 | Y | White |
| 65 | 21 | 23 | N | Hispanic |
| ... | ... | ... | ... | ... |

MACHINE LEARNING UNIVERSITY

# Supervised Learning: Classification

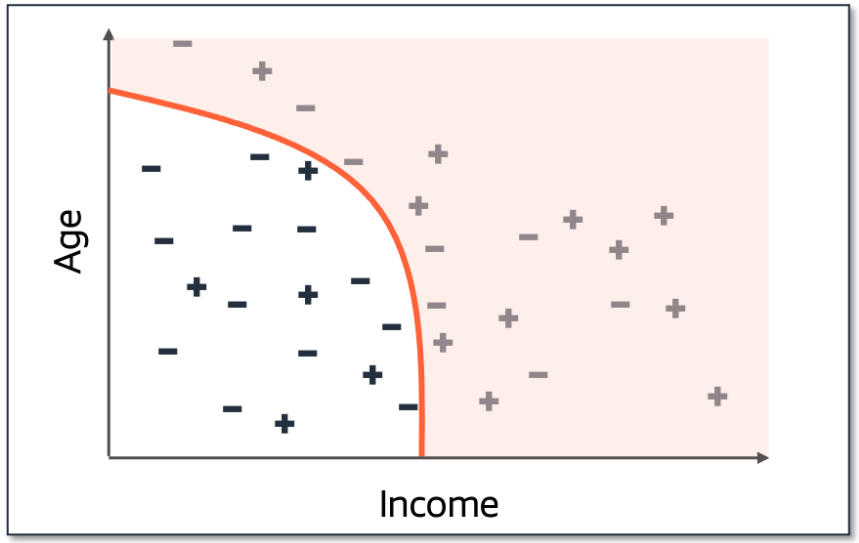



- Linear Regression
- K-Nearest Neighbors
- Neural Nets
- Decision Trees

…

- Support Vector Machine
- K-Nearest Neighbors
- Neural Nets
- Decision Trees

…

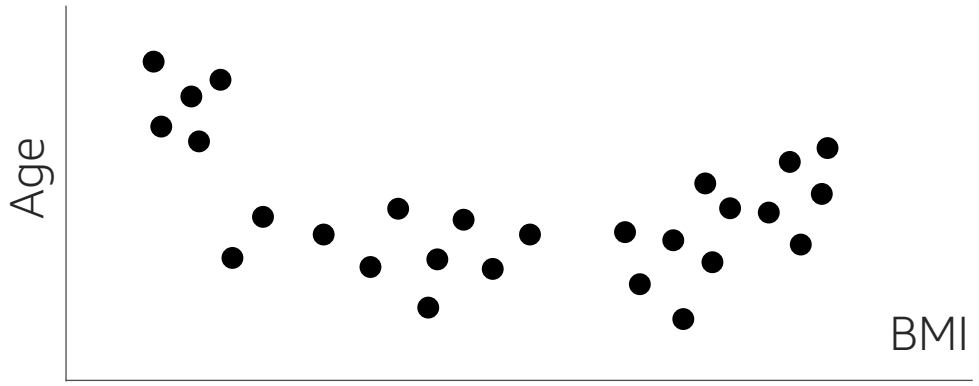



Label | Features

| Approved | Age | Income | Smoker | Ethnicity |
|---|---|---|---|---|
| + | 32 | 28k | Y | White |
| - | 21 | 23k | N | Hispanic |
| … | … | … | … | … |

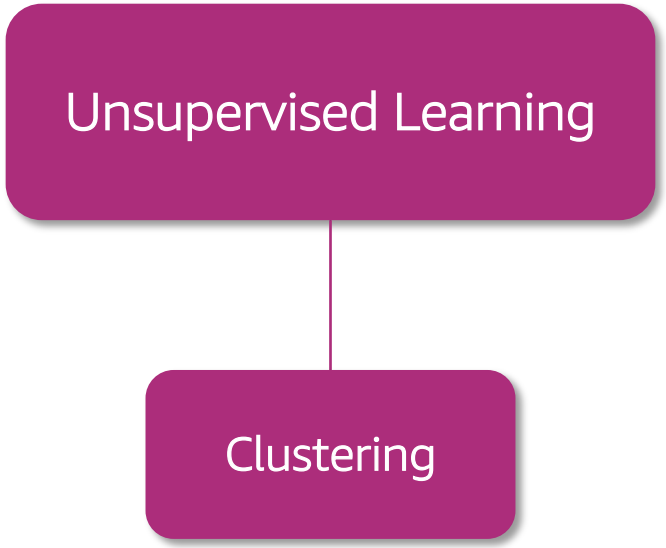
MACHINE LEARNING UNIVERSITY

# Unsupervised Learning: Clustering




Features

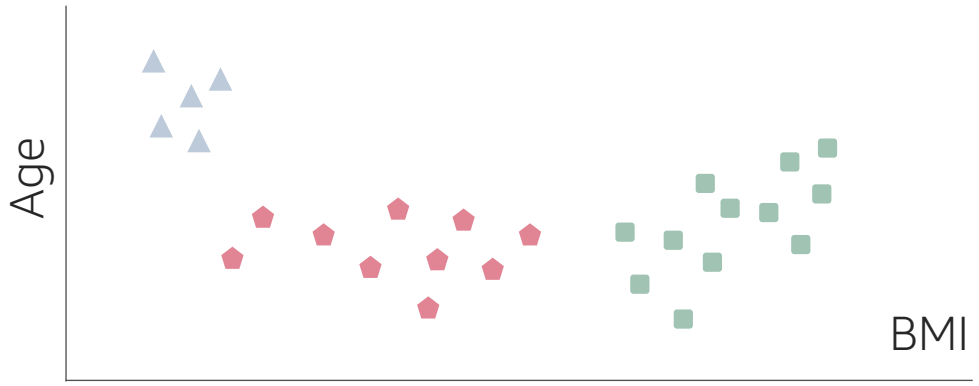
| Age | BMI | Smoker | Ethnicity |
|---|---|---|---|
| 32 | 28 | Y | White |
| 21 | 23 | N | Hispanic |
| … | … | … | … |




- K-Means
- PCA
- Collaborative Filtering

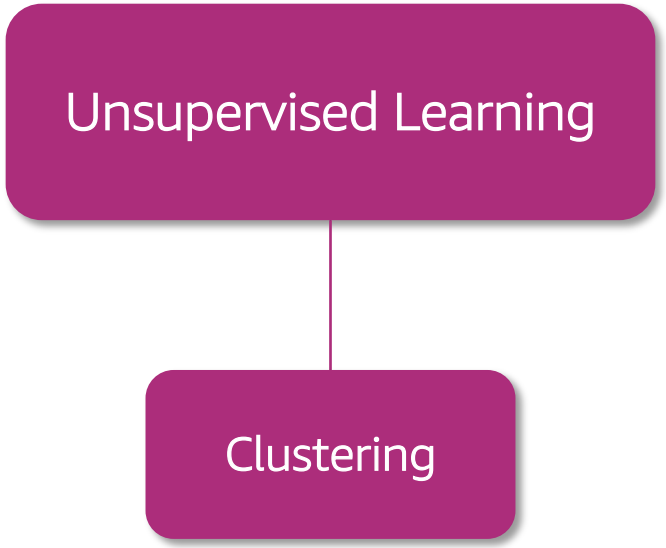
…

# Unsupervised Learning: Clustering



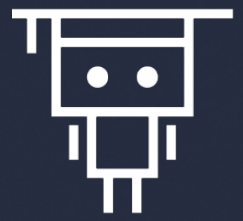

Features

| Age | BMI | Smoker | Ethnicity |
|---|---|---|---|
| 32 | 28 | Y | White |
| 21 | 23 | N | Hispanic |
| … | … | … | … |




- K-Means
- PCA
- Collaborative Filtering

…



# Examples of ML Problems/Tasks

# ML Problem Type: Regression

| Problem type | Description | Example |
| --- | --- | --- |
| Regression | Predicting a numerical value | |

**Sickness Severity Score**

- To determine if individual may require care in future, a risk score was produced.
- Observations showed that Black enrollees have higher rate of chronic illnesses than White enrollees of same score.
- Algorithm used health costs as proxy for health.

Obermeyer et al (2019)

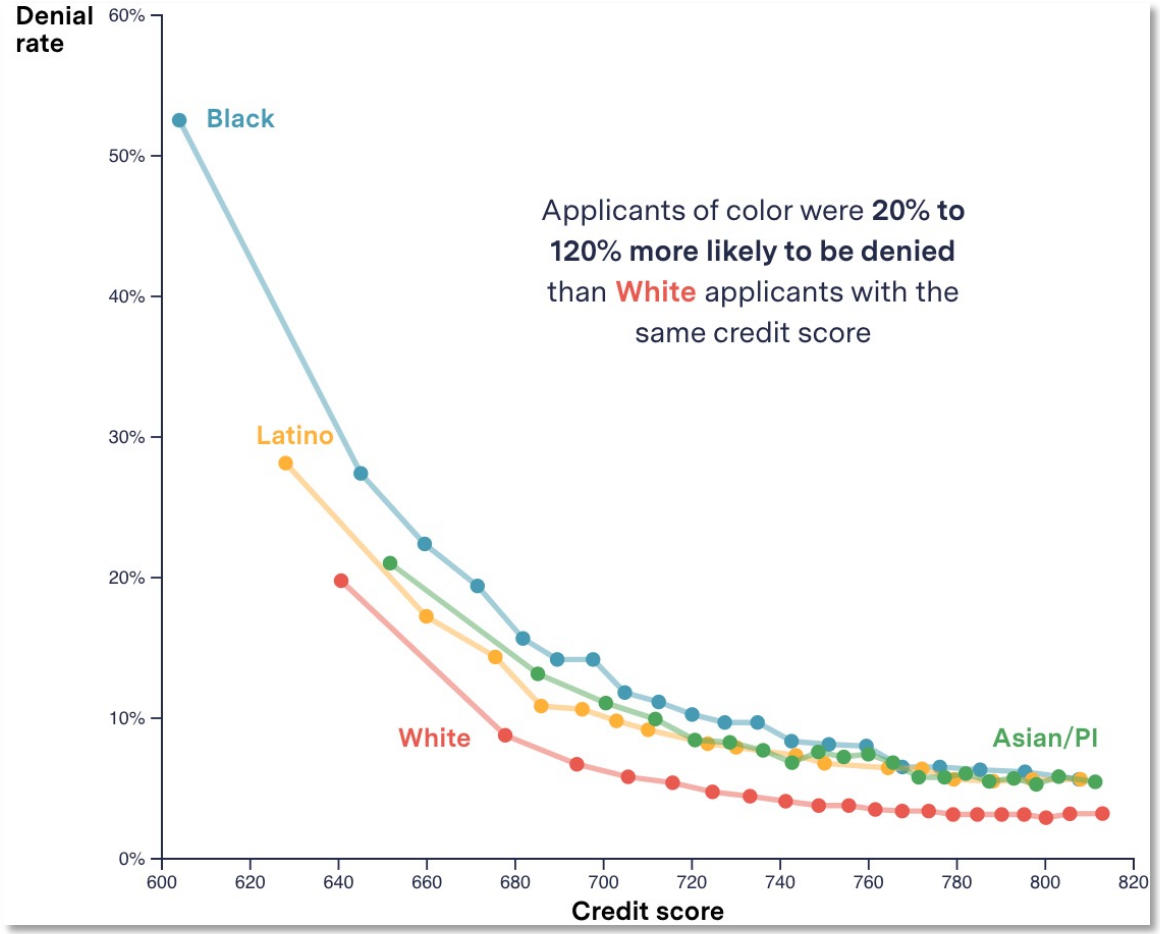
# ML Problem Type: Classification

| Problem type | Description | Example |
| --- | --- | --- |
| Regression | Predicting a numerical value | |
| Classification | Predicting a label | |

**Credit Worthiness**

- To determine if applicant may default, a probability score was used.
- "In default": failed to pay a debt for > 90 days on ≥ 1 account in ensuing 18-24 month period.
- Use equalized odds to enforce accuracy is equal in all groups.

Hardt et al. (2016)




Martinez, E., & Kirchner, L. (2021, August 25). The secret bias hidden in mortgage-approval algorithms. *The Markup*. [Link]

# ML Problem Type: Ranking

| Problem type | Description |
| --- | --- |
| Regression | Predicting a numerical value |
| Classification | Predicting a label |
| Ranking | Ordering items to find the most relevant based on search query |

**Example**

**Job Ranking**

- Ranked lists produced by biased model can amplify bias.
- Without fairness intervention, top ranked results can be skewed with respect to sensitive attribute(s).
- To mitigate, we require fair re-ranking algorithms (e.g. score maximizing greedy mitigation).

Geyik et al. (2019)

# ML Problem Type: Recommendation

| Problem type | Description |
| --- | --- |
| Regression | Predicting a numerical value |
| Classification | Predicting a label |
| Ranking | Ordering items to find the most relevant based on search query |
| Recommendation | Finding relevant items based on past behavior (without direct user input) |

Example

**Targeted Advertising**

- Recommendations produced by biased models may amplify bias.
- Experiment: Different ad settings led to different freq. of high paying jobs shown.
- Require multi-sided fairness or sequential recommendations

Pitoura, E., Stefanidis, K. & Koutrika, G (2021)

# ML Problem Type: Clustering

| Problem type | Description |
| --- | --- |
| Regression | Predicting a numerical value |
| Classification | Predicting a label |
| Ranking | Ordering items to find the most relevant based on search query |
| Recommendation | Finding relevant items based on past behavior (without direct user input) |
| Clustering | Finding patterns/groups in examples |

## Example

### Fair Hiring

- Cluster resumes for shortlist in a hiring scenario
- Callback rates may differ per group → consider group fairness (incorporate fairness constraints in optimization)
- Aim for proportional representation of the sensitive class per cluster

Abraham et al. (2020)

# ML Problem Type: Anomaly Detection

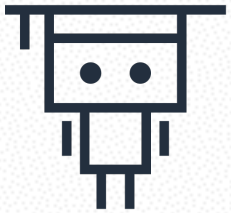
| Problem type | Description |
| --- | --- |
| Regression | Predicting a numerical value |
| Classification | Predicting a label |
| Ranking | Ordering items to find the most relevant based on search query |
| Recommendation | Finding relevant items based on past behavior (without direct user input) |
| Clustering | Finding patterns/groups in examples |
| Anomaly Detection | Finding outliers in set of examples |

Example

**Government Relief Fraud**

- Small fraction of COVID relief direct payments experience fraud.
- Scale of money allocation makes traditional tracking impossible.
- Outliers should not be skewed towards particular group → FairLOF

Deepak et al. (2021)



MACHINE LEARNING
UNIVERSITY

# Introduction to Fairness & Responsible AI

# What is Responsible AI?

> [...] AI that is innovative and trustworthy and that respects human rights and democratic values.
>
> *OECD, https://oecd.ai/en/ai-principles*

→ Dimensions of Responsible AI

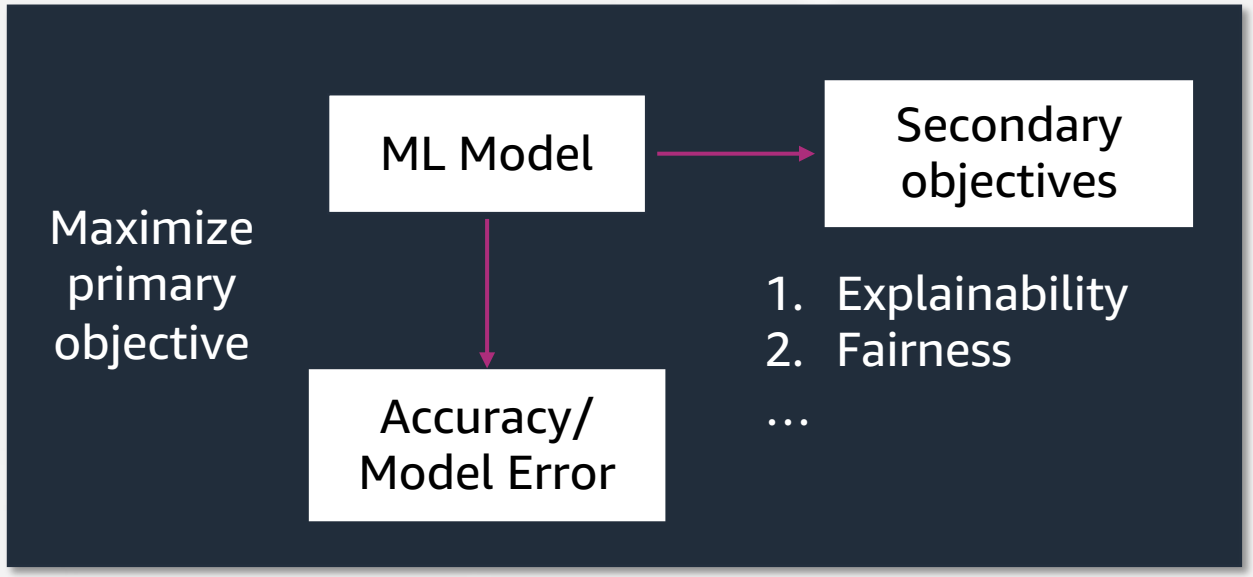
Fairness is one of several dimensions.

# Dimensions of Responsible AI

| Dimension | Example Metric |
| --- | --- |
| Privacy & Security | Is data used in accordance with privacy & legal considerations, and protected from theft and exposure? |
| Fairness | Are there harmful disparities in system behavior across different subpopulations? |
| Explainability | Does the system offer a clear rationale for its decisions? |
| Robustness | How hard is it to confuse or fool the system, e.g. 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 amongst all stakeholders? |

# Tradeoffs 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.
   - » E.g.: Preserving privacy makes data more coarse → potentially degrades ability to explain model behavior.
   - » E.g.: Making model results and components more transparent → potentially security & privacy risk.

# Dimensions of Responsible AI

| Dimension | Example Metric |
| --- | --- |
| Privacy & Security | Is data used in accordance with privacy & legal considerations, and protected from theft and exposure? |
| Fairness | Are there harmful disparities in system behavior across different subpopulations? |
| Explainability | Does the system offer a clear rationale for its decisions? |
| Robustness | How hard is it to confuse or fool the system, e.g. 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 amongst all stakeholders? |

# 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 with respect to different subpopulations (groups with shared values of demographic variables). |
| Responsible AI | Harmful disparities in system behavior on subpopulations (groups with shared values of demographic variables); also “unwanted bias”. |

# Impact vs. Treatment vs. Outcome

- Impact vs. Treatment (concepts in US law):
  - » Considering sensitive attributes as part of a decision making process, intentionally or indirectly, constitutes disparate treatment.
  - » Decision making processes or procedures that generate outcomes that disproportionally favor subpopulations exhibit disparate impact.

More formal definition from FDIC [here].

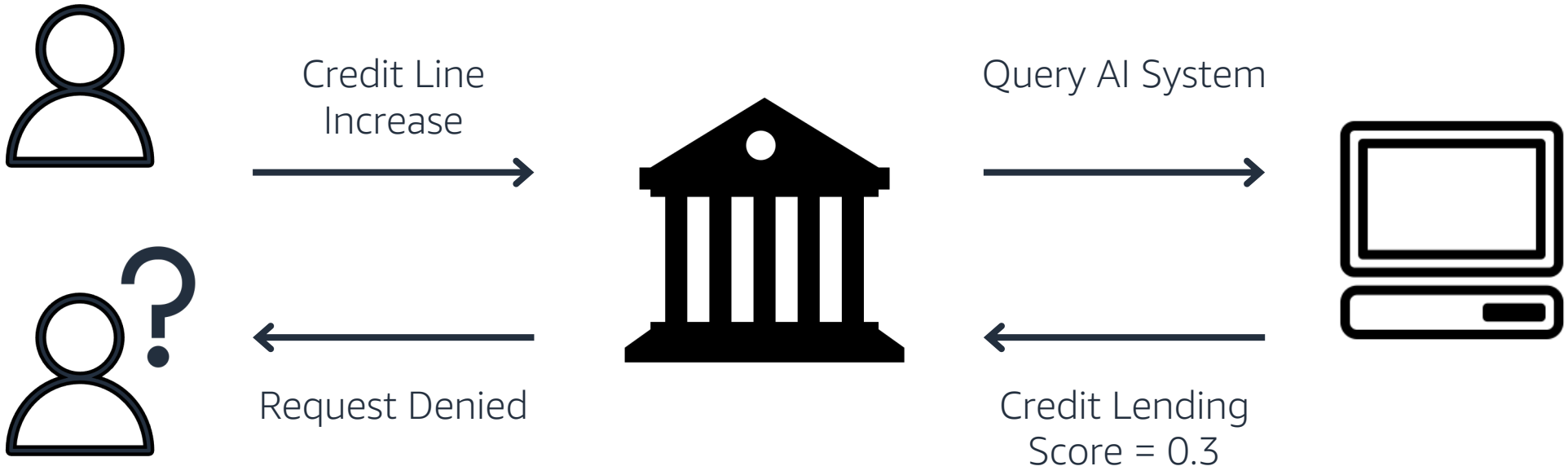
- Protections don't always hold (e.g. medical diagnosis).
- Even if impact & treatment are controlled for, fair outcomes are not guaranteed (interaction with ML models can create bias).

# Group vs. Individual Fairness

- Group Fairness (equalize impact/treatment across groups)
  - » Strong theory and algorithms & practical implementations; e.g. same rate of error across different groups
  - » Provide no guarantees at individual level

- Individual Fairness (ensures similar individuals treated/impacted similarly)
  - » Strong (non-statistical) assumptions prevent practical implementations
  - » Binds at individual level

# Credit line increase example

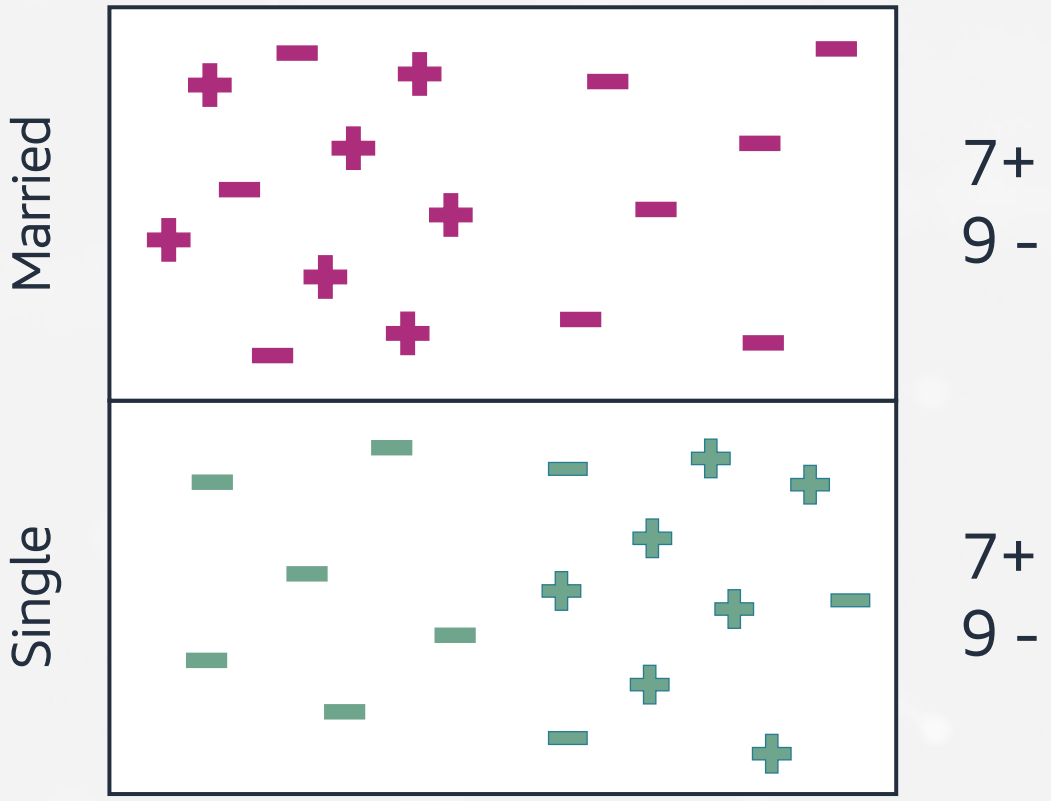



Fair lending laws [ECOA, FCRA] require credit decision to be explainable

Source: **Responsible AI in Industry: Practical Challenges and Lessons Learned**
Krishnaram Kenthapadi

# Intersectional Group Fairness

- Individuals can exhibit multiple demographic attributes.
- If multiple attributes intersect, new sub-groups emerge.
- Cannot build models for separate groups.



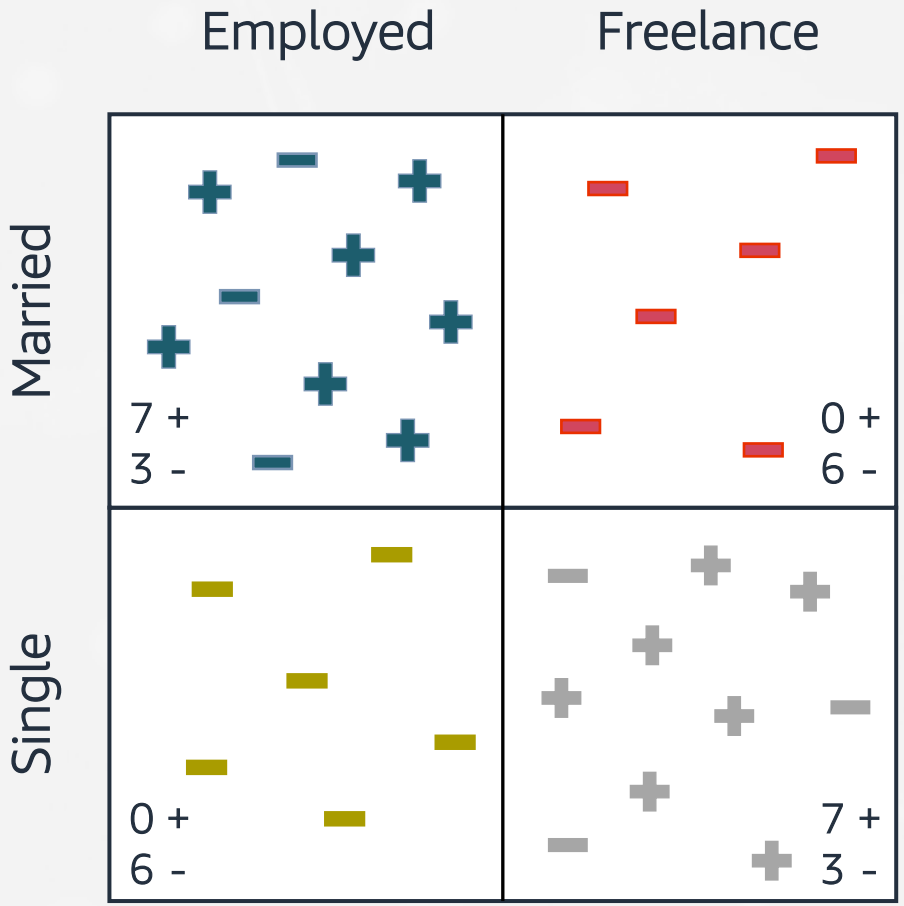

Example Credit Line Increase

# Intersectional Group Fairness

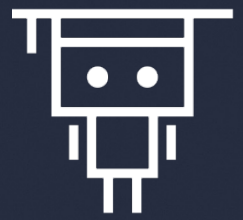
- Attributes (e.g. married vs. single) can intersect with other attributes (e.g. employed vs. unemployed)

  → new sub-groups emerge (e.g. married & employed, …)

- Sub-groups may experience additional disparity (see positive/negative outcomes per sub-group).

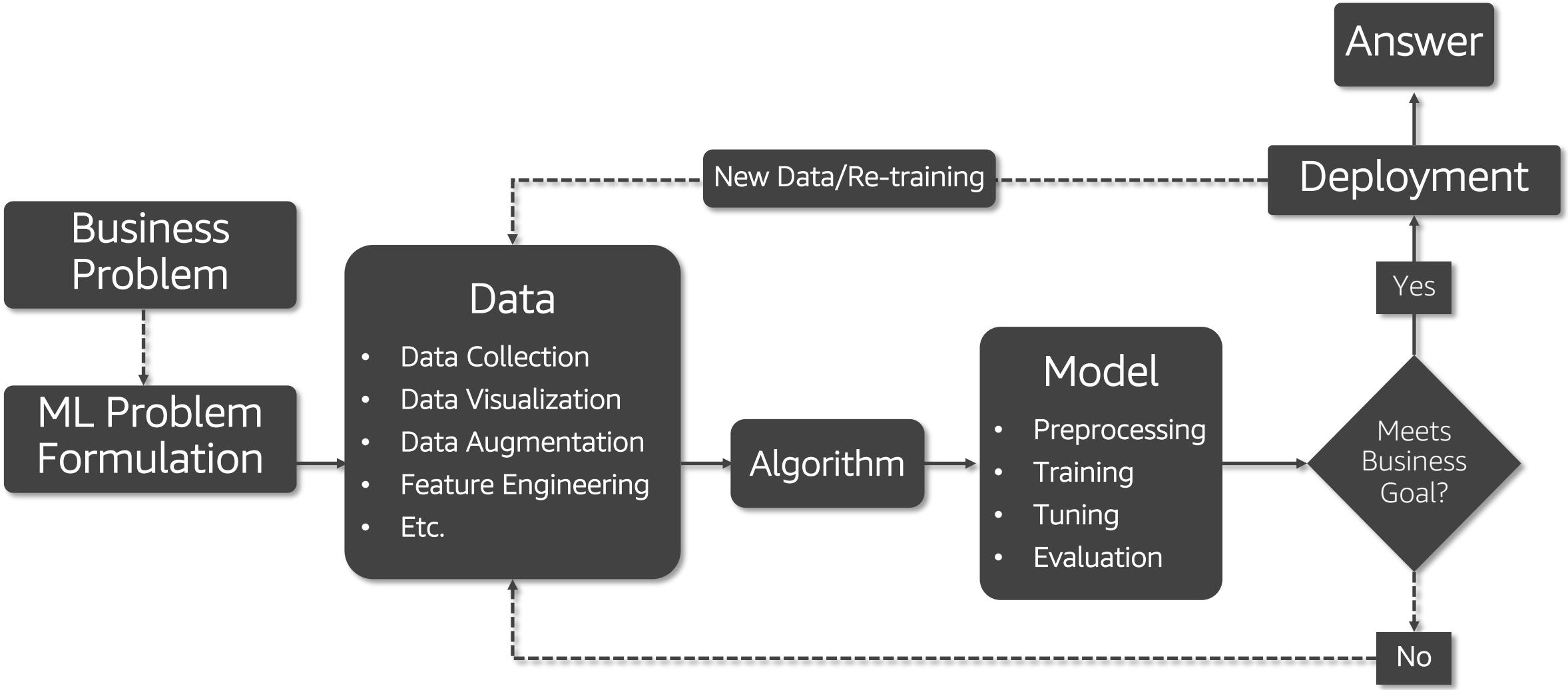



Example Credit Line Increase



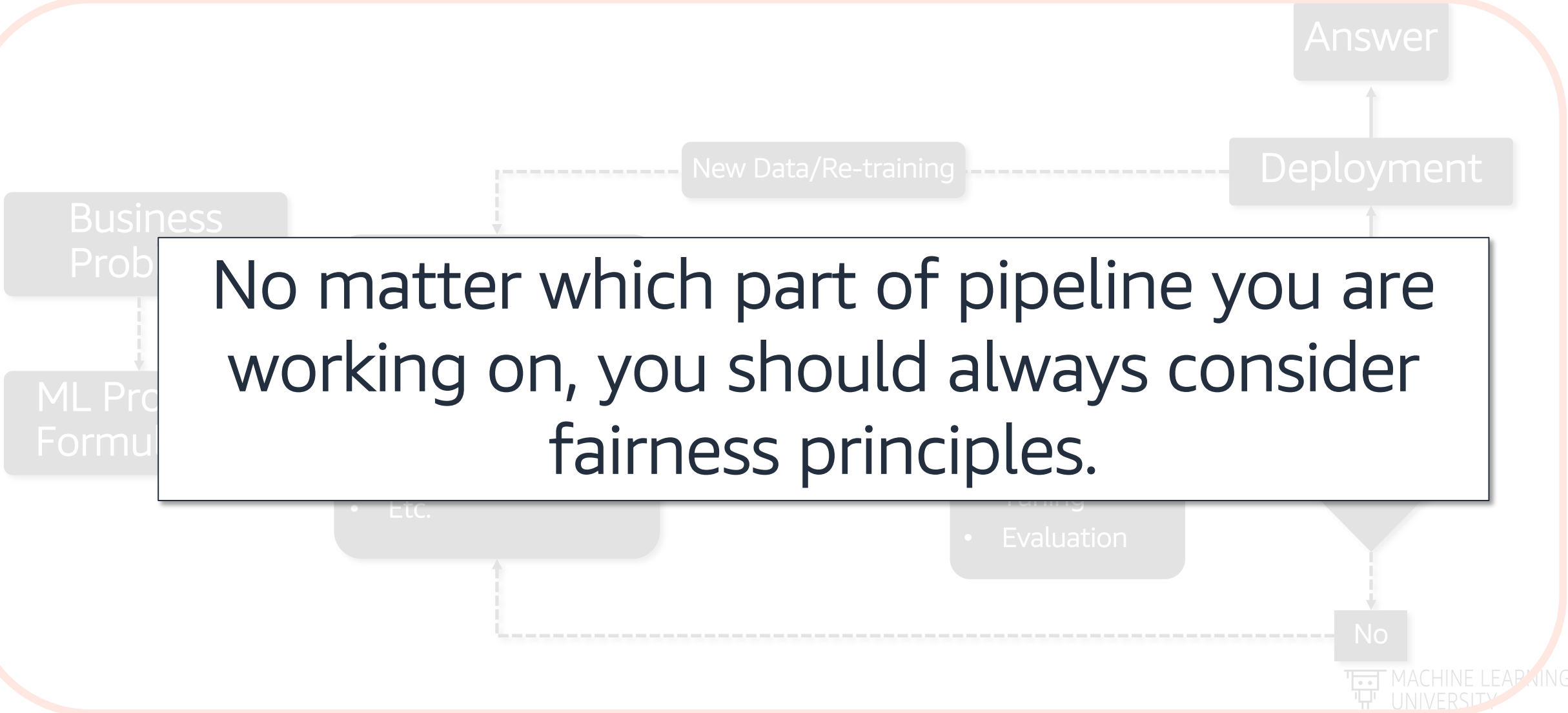
# Fairness Throughout the ML Lifecycle

# Where should you consider fairness?



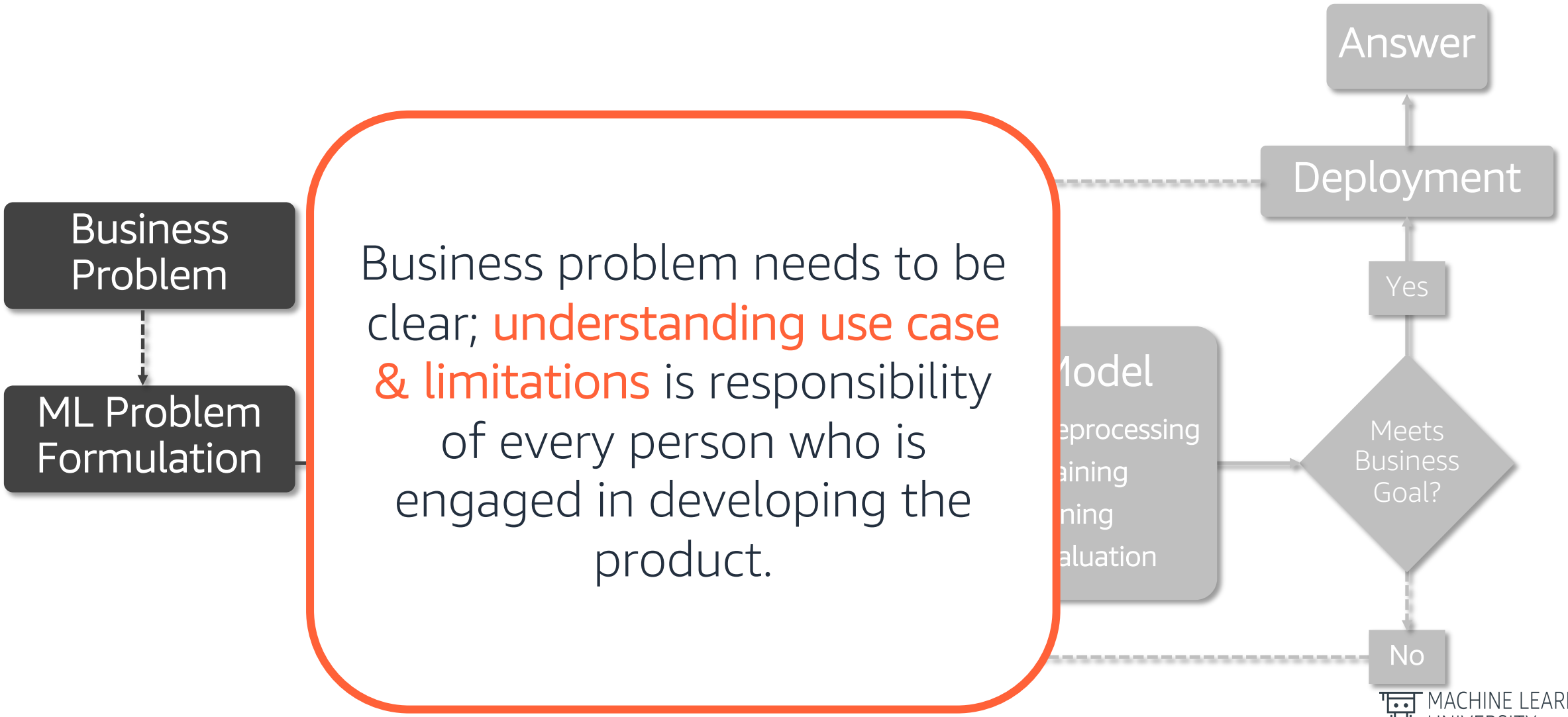

MACHINE LEARNING UNIVERSITY

# Everywhere ☺




No matter which part of pipeline you are working on, you should always consider fairness principles.

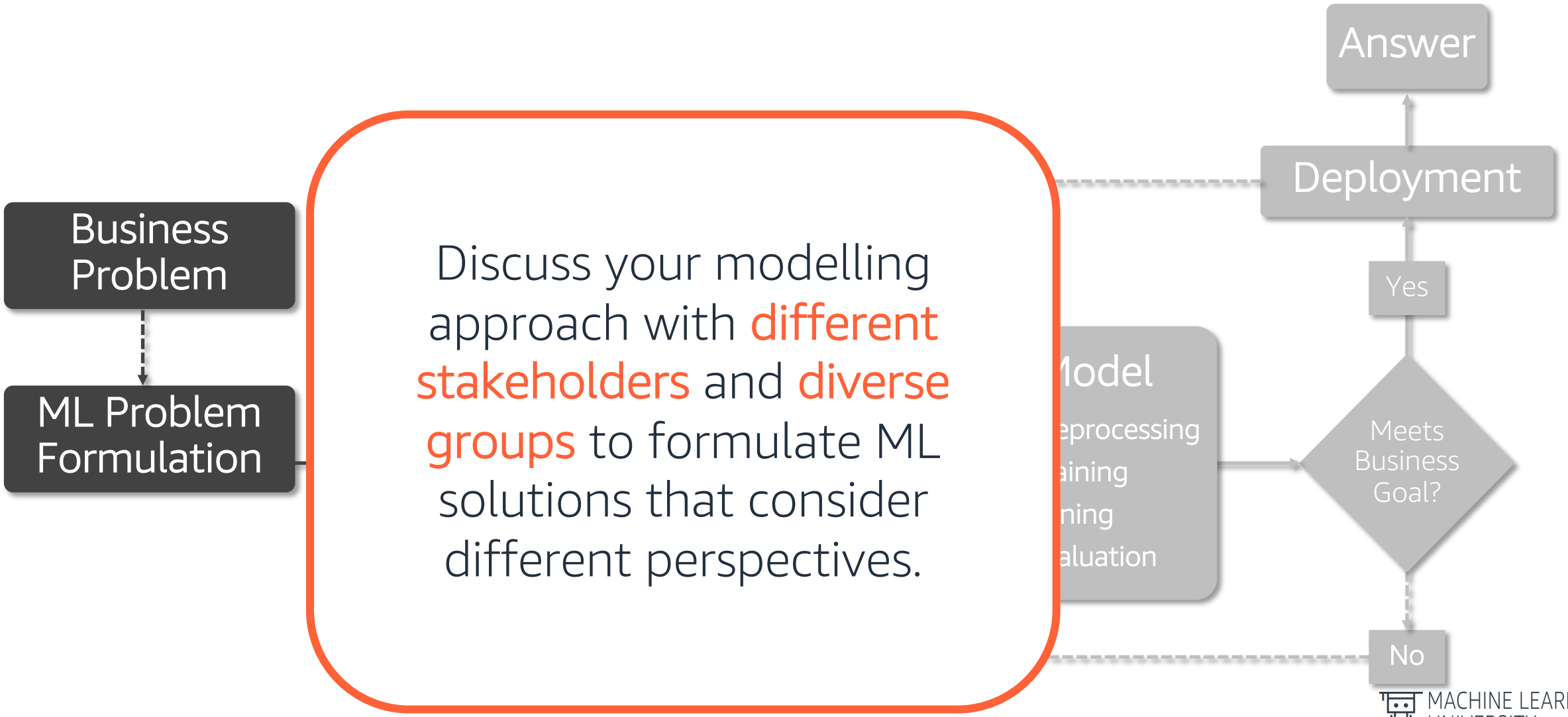
# Machine Learning Lifecycle



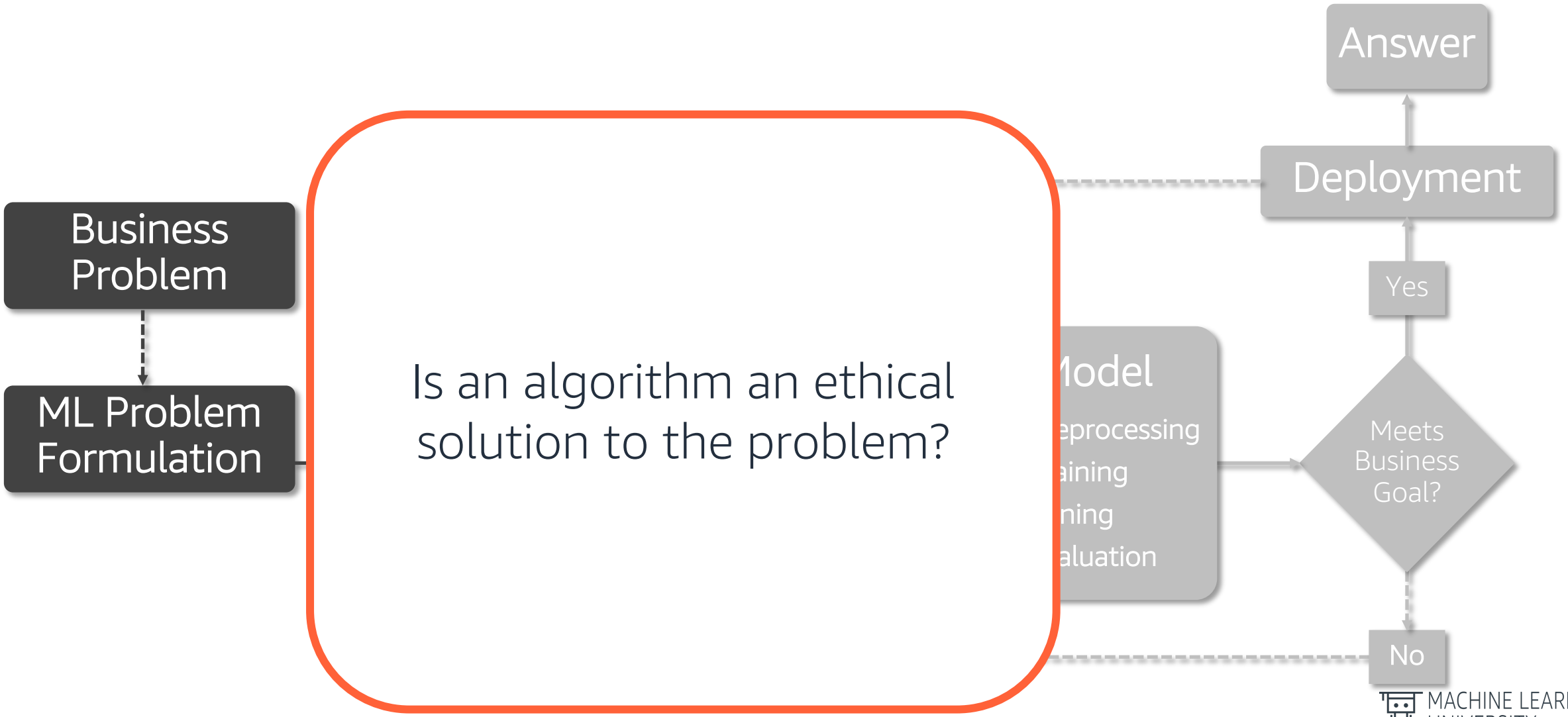
# Machine Learning Lifecycle


MACHINE LEARNING UNIVERSITY

# Machine Learning Lifecycle

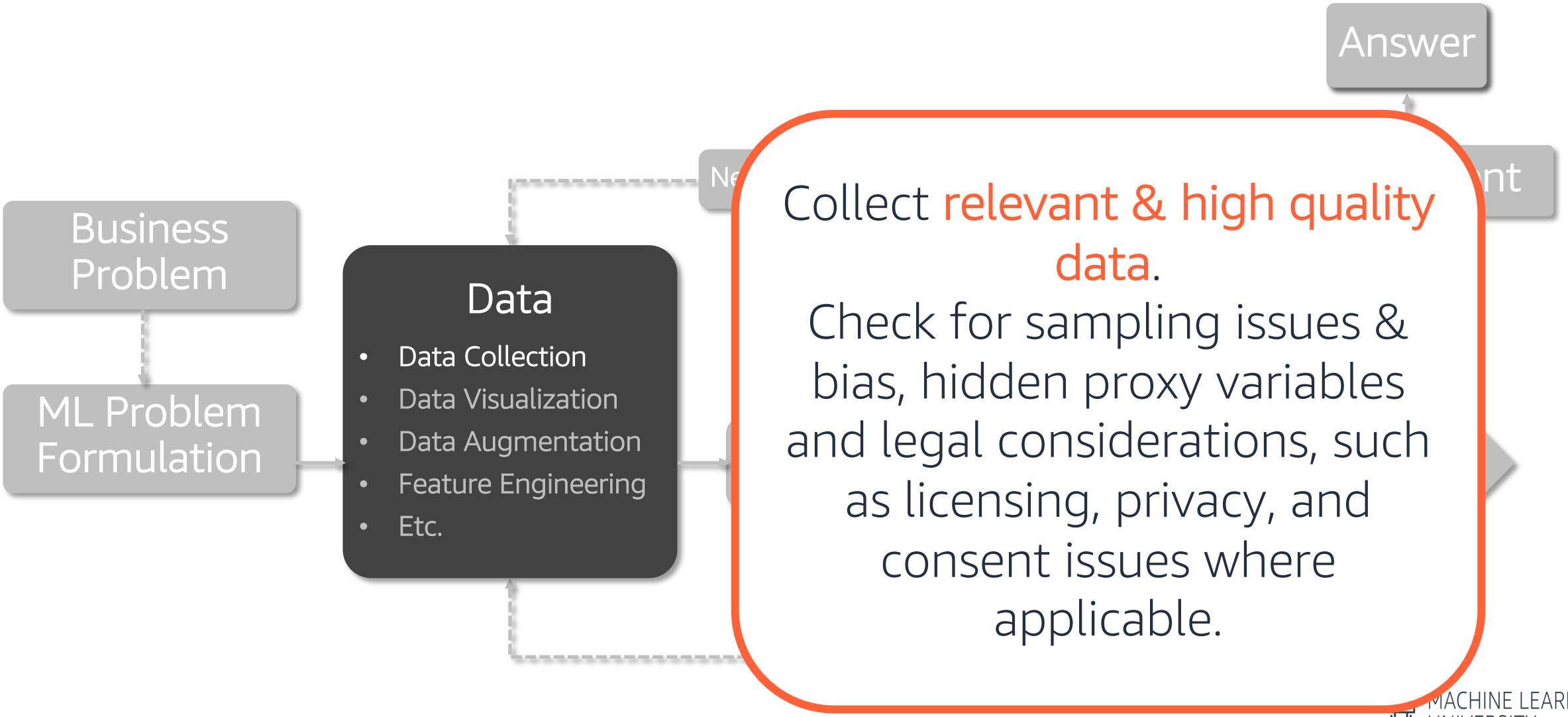


# Machine Learning Lifecycle

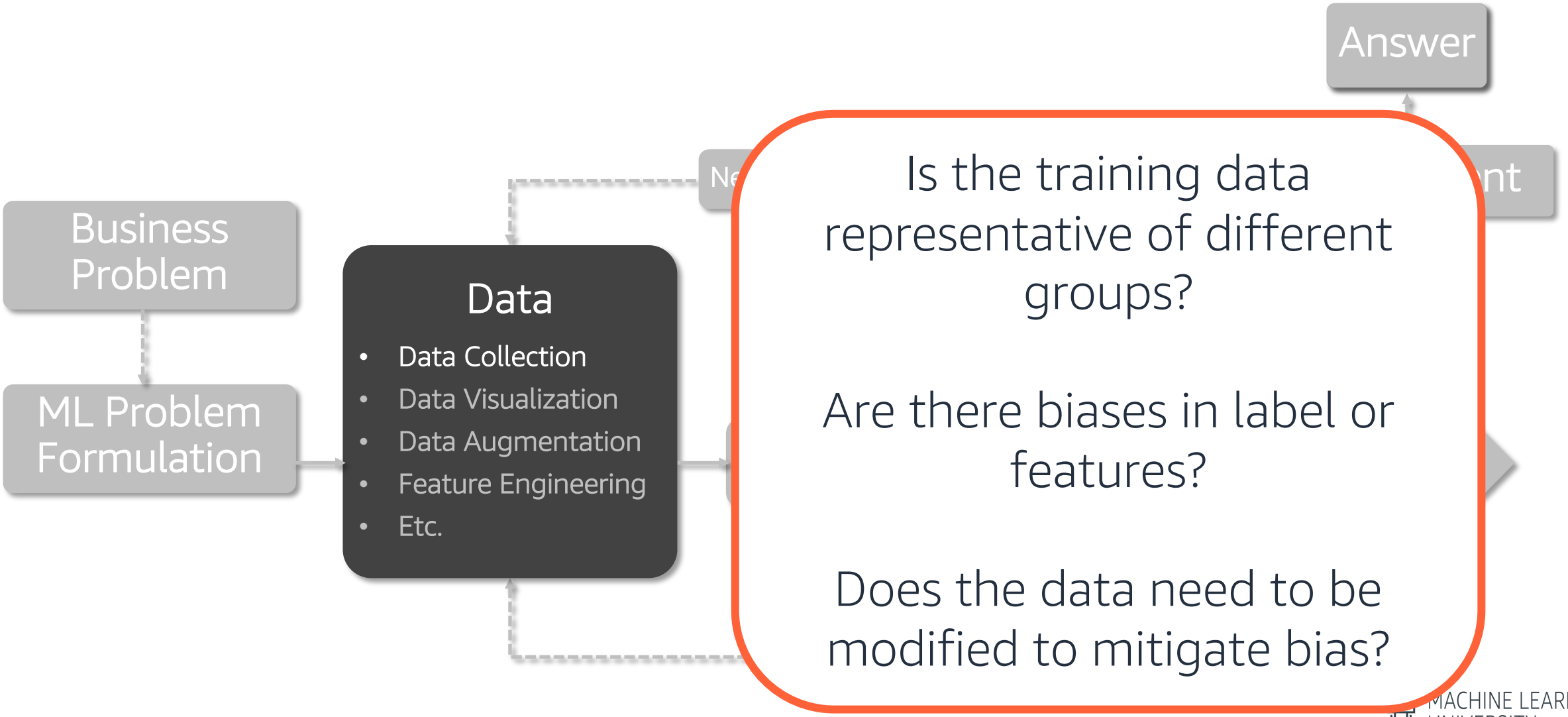



MACHINE LEARNING UNIVERSITY

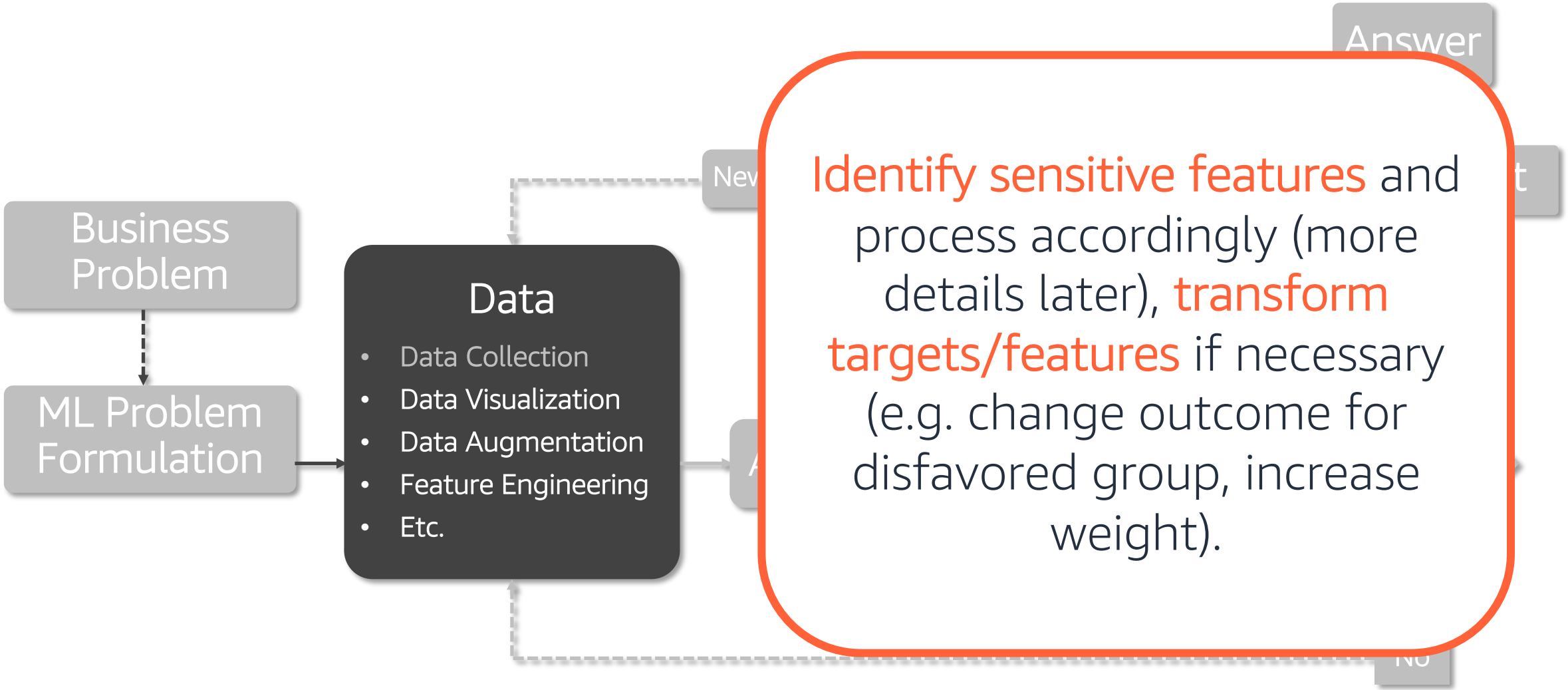
# Machine Learning Lifecycle



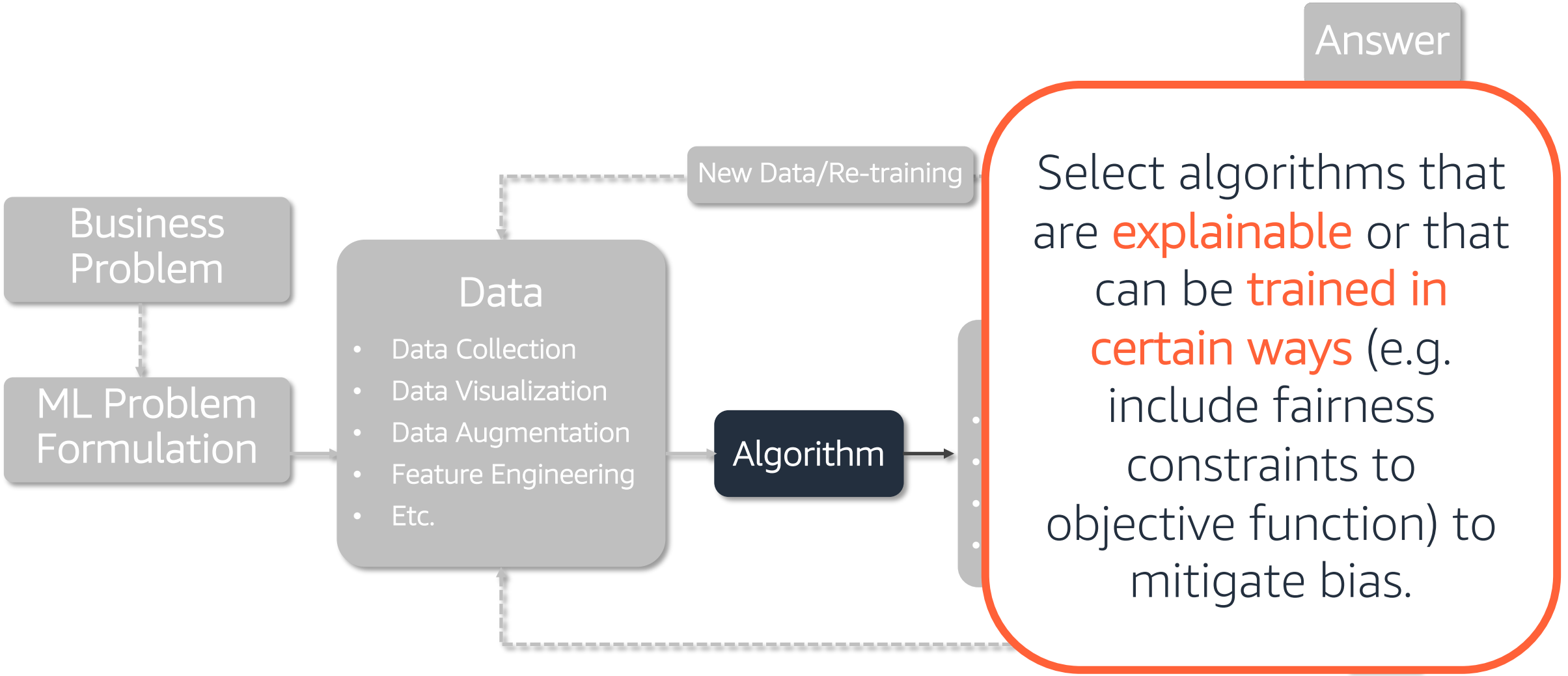
# Machine Learning Lifecycle



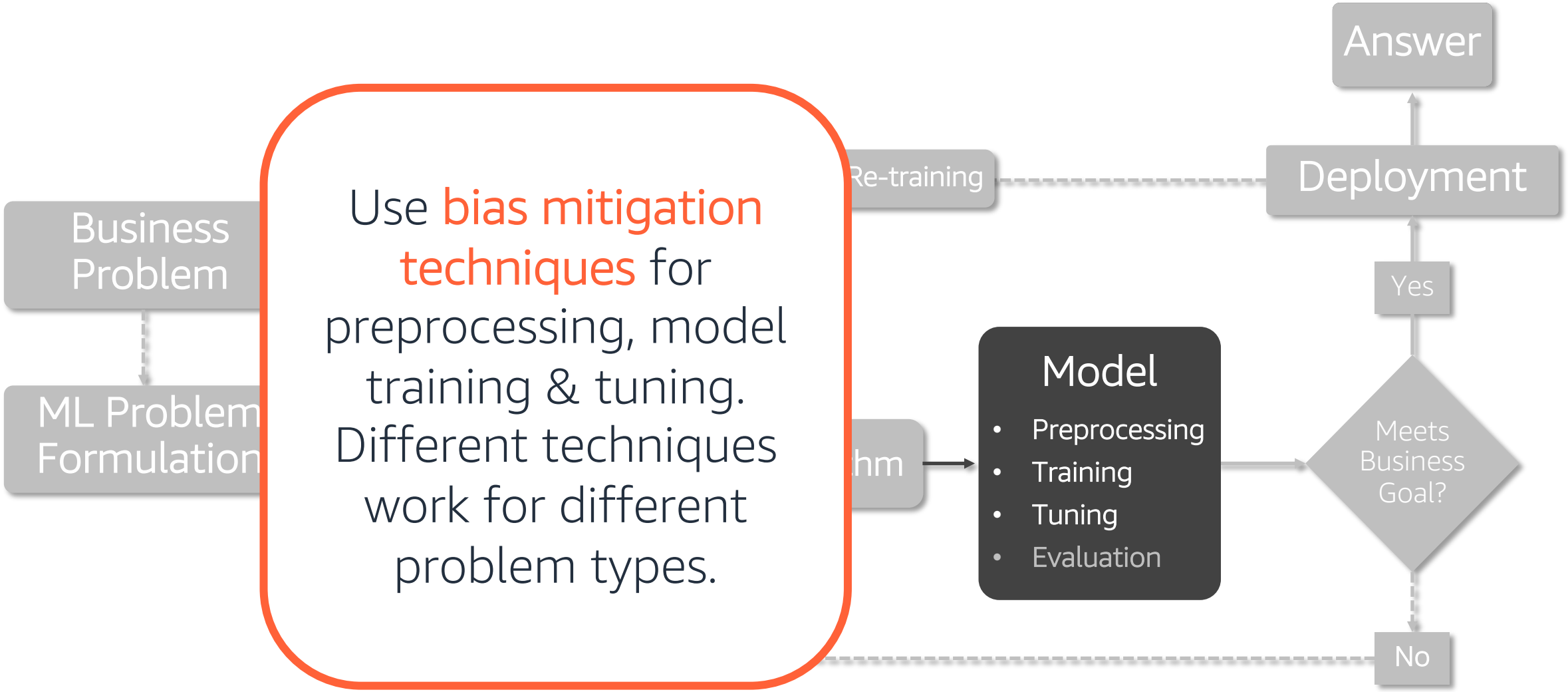
# Machine Learning Lifecycle



# Machine Learning Lifecycle

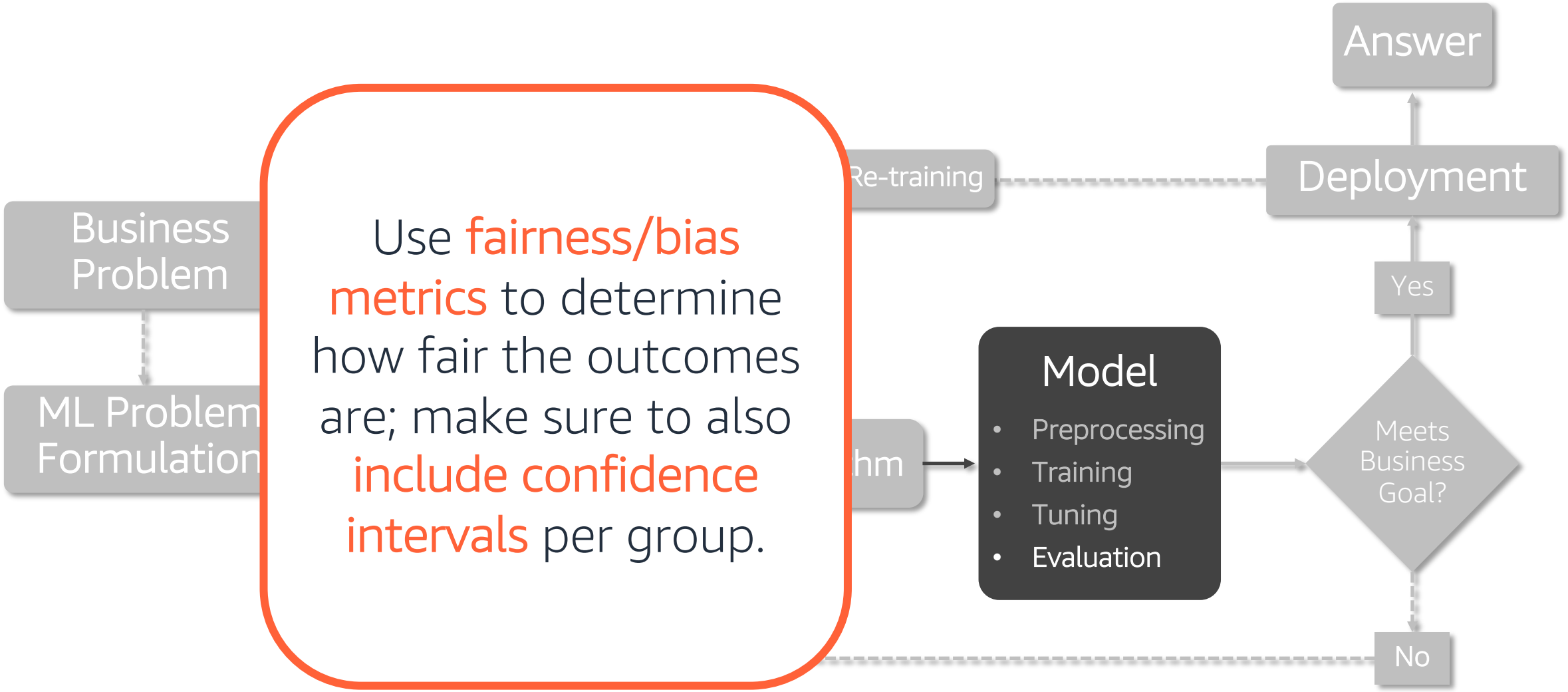


# Machine Learning Lifecycle



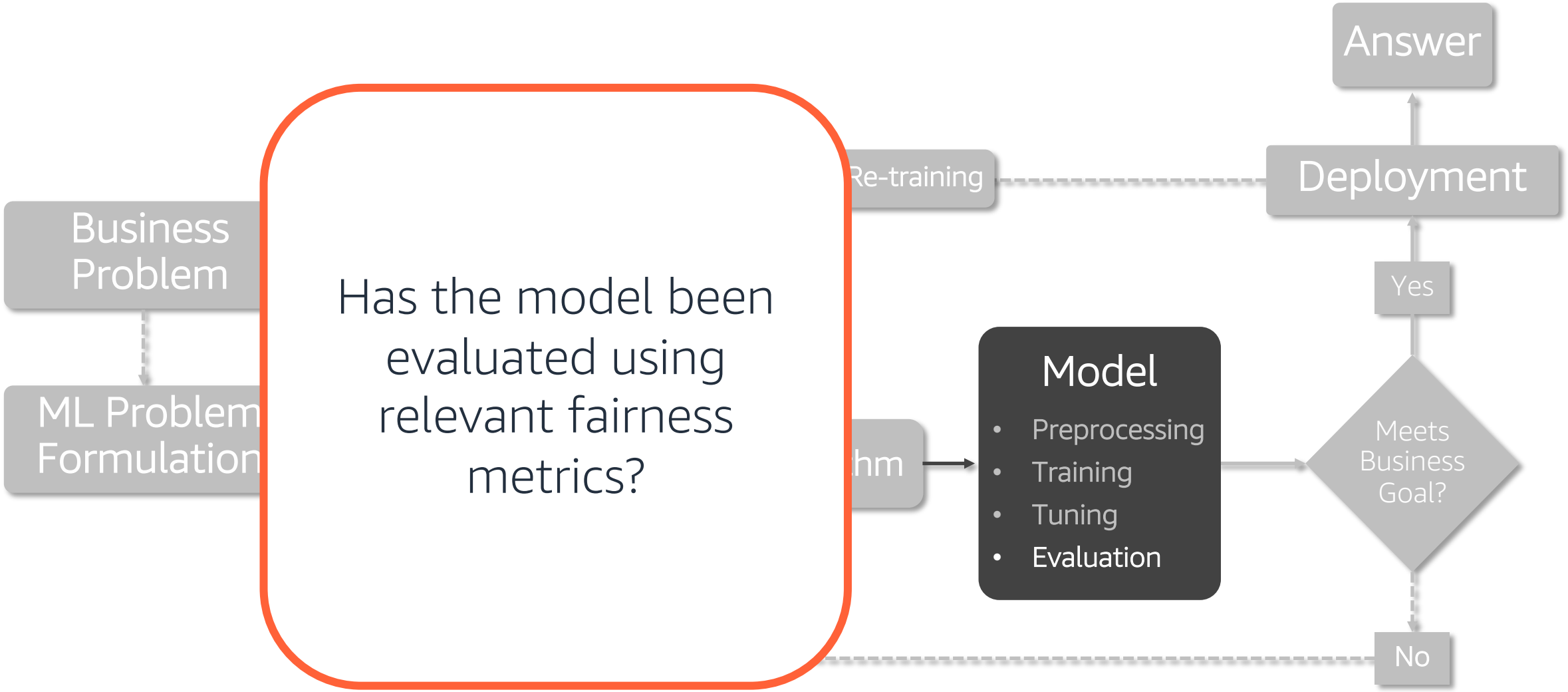

MACHINE LEARNING UNIVERSITY

# Machine Learning Lifecycle

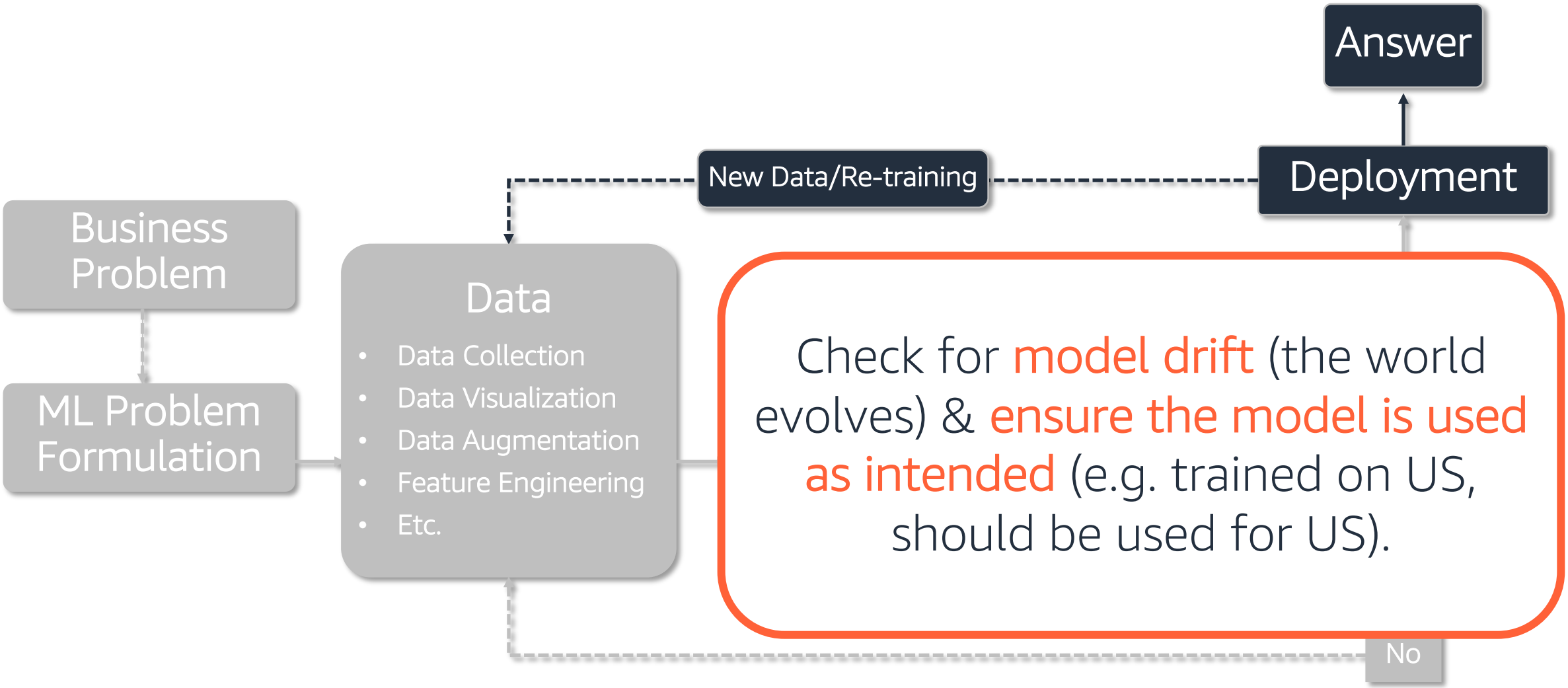



MACHINE LEARNING UNIVERSITY

# Machine Learning Lifecycle

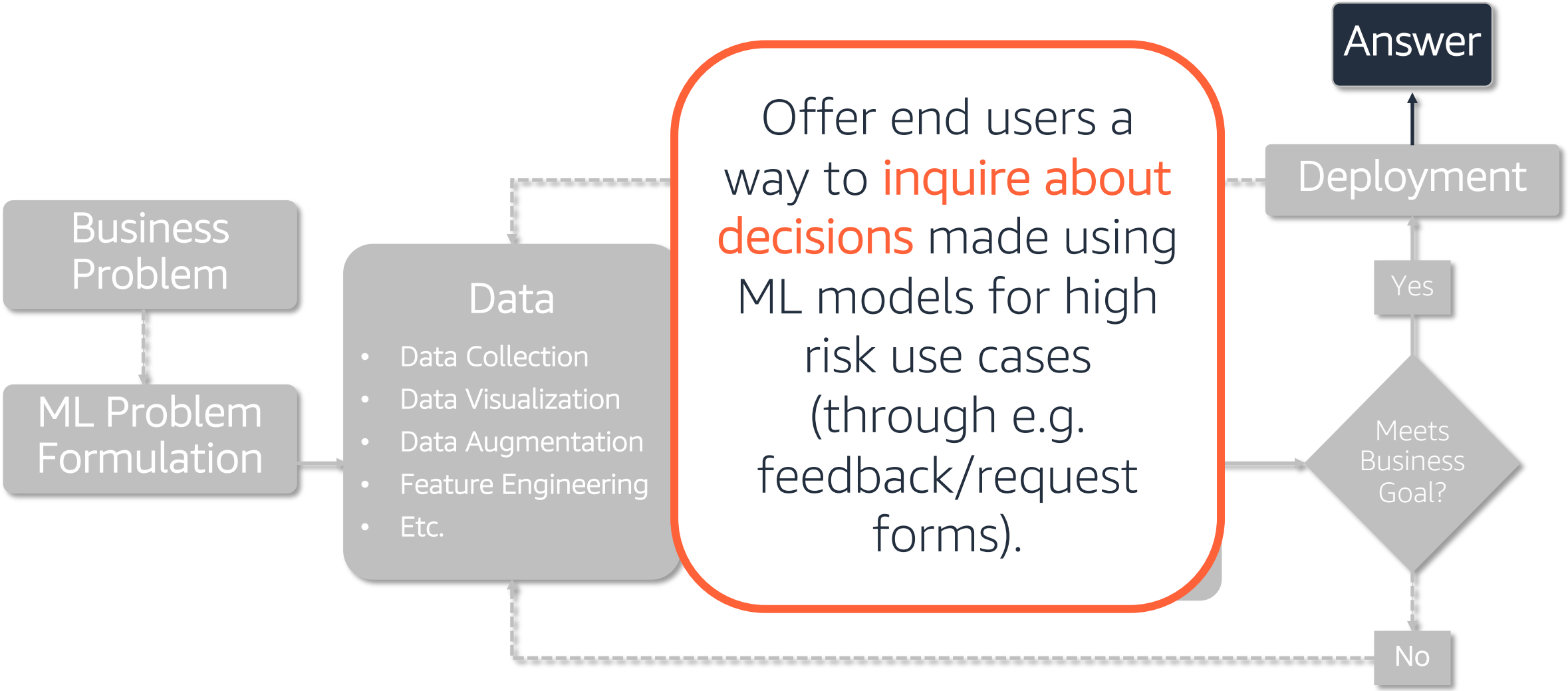


# Machine Learning Lifecycle



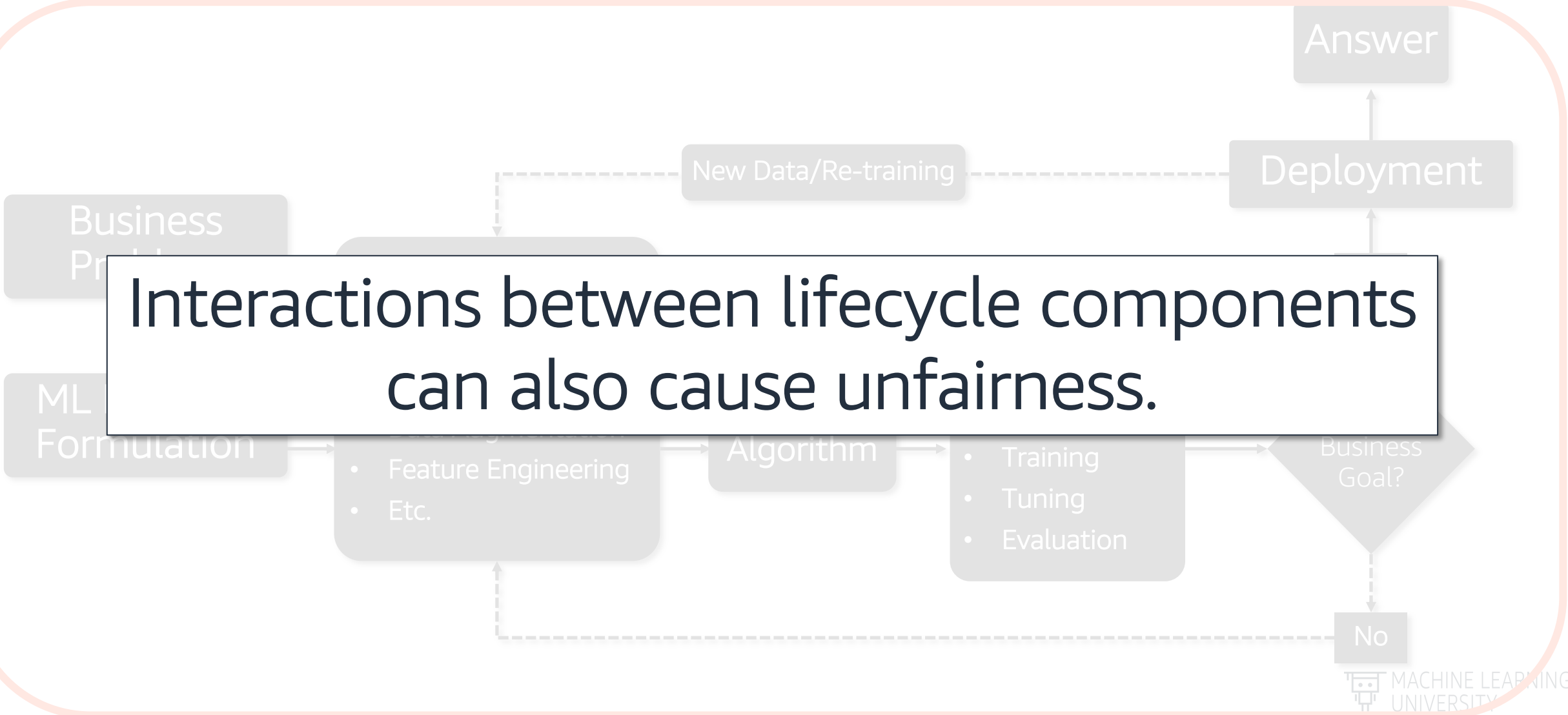

MACHINE LEARNING UNIVERSITY

# Machine Learning Lifecycle

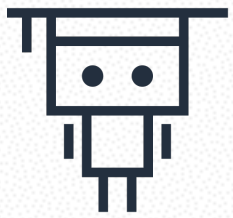



MACHINE LEARNING UNIVERSITY

# Careful!

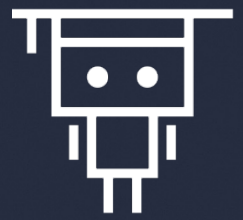



Interactions between lifecycle components can also cause unfairness.



MACHINE LEARNING
UNIVERSITY

# Model Formulation & Data Collection



# ML Model Formulation

# Human-Centered Design

AI systems & ML solutions should consider impact on humans: Human-centered design (HCD).

# Human-Centered Design

AI systems & ML solutions should consider impact on humans: Human-centered design (HCD).

Consider:

- **Potential harms** (Will use of ML technology be beneficial for the use case? Will it reinforce stereotypes?)
- **Value proposition** (Is AI necessary, or can a simple interpretable rule based system achieve the same?)

# Human-Centered Design

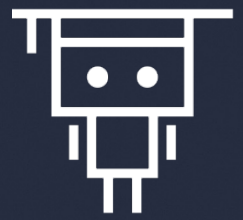
AI systems & ML solutions should consider impact on humans: Human-centered design (HCD).

Consider:

- **Potential harms** (Will use of ML technology be beneficial for the use case? Will it reinforce stereotypes?)
- **Value proposition** (Is AI necessary, or can a simple interpretable rule based system achieve the same?)

Ensure there are:

- **Safety measures** (continuously test and evaluate the system)
- **Ways to challenge the model/ML pipeline** (include ways for users to report unexpected behavior)



# Data Collection

# Models may produce biased results

- ML models learn from the data that is provided as input:

  → *if bad quality or biased data is used, a model may produce poor or biased results. Bias mitigation techniques can help reduce this (assuming we understand how to address the underlying cause of inequality).*

- Models may produce biased results by using:
  - » Skewed examples
  - » Tainted examples
  - » Limited features
  - » Different sample sizes
  - » Proxies

S. Barocas and M. Hardt (2017)

# Reducing bias by ensuring data integrity

- To ensure data integrity, we need to:
  - Select meaningful, high-quality features (feature quality)
  - Use valid, fair sampling & storage techniques (feature sampling)
  - Verification/creation of labels (label quality)

# Feature Quality

- High-quality features are robust & objective (i.e. they don't depend on who recorded the data and are reproducible):
  - » Quantifiable (i.e. true measurements of certain; # of messages between peers vs 'connectedness' score)

# Feature Quality

- High-quality features are robust & objective (i.e. they don't depend on who recorded the data and are reproducible):
  - » Quantifiable (i.e. true measurements of certain; # of messages between peers vs 'connectedness' score)
  - » Direct measurements (not outputs of another model or proxies for other measurements as hidden sensitive attributes)

# Feature Quality

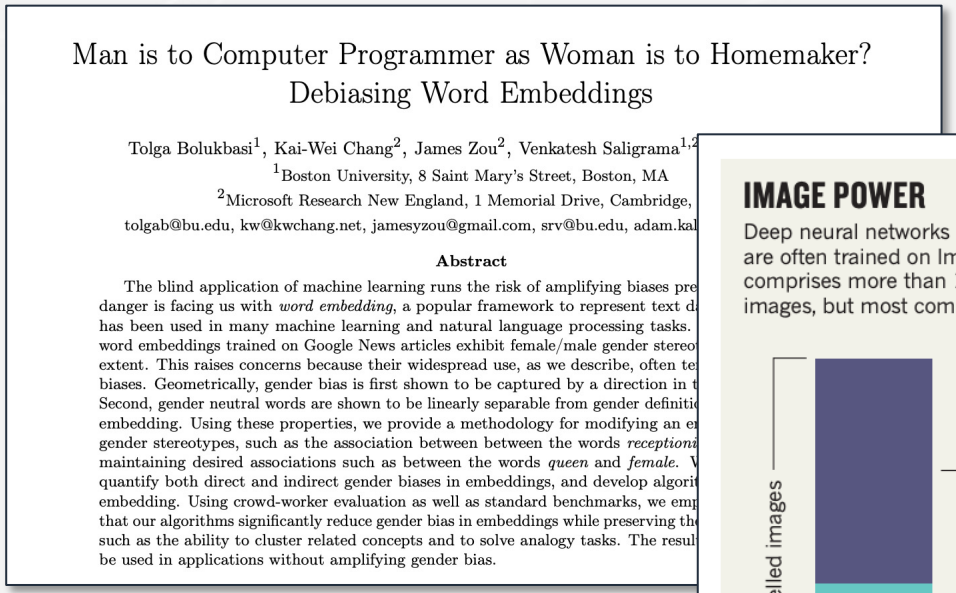
- High-quality features are robust & objective (i.e. they don't depend on who recorded the data and are reproducible):
  - » Quantifiable (i.e. true measurements of certain; # of messages between peers vs 'connectedness' score)
  - » Direct measurements (not outputs of another model or proxies for other measurements as hidden sensitive attributes)
  - » Understandable (feature name or metadata that describes feature should be easy to understand; monthly_bonus vs mb)

# Feature Sampling: Bias

- Beware of biased sampling (sampling methods that systematically over- or under-represent certain groups)

- This can be due to:
  - **Selection bias** (use of samples from a different distribution)
  - **Measurement bias** (different measuring devices or techniques)
  - **Historical bias** (society itself produces biased data)
  - **Confirmation bias** (wrongly discarding outliers to match existing beliefs)
  - …

# Biased Datasets in CV & NLP



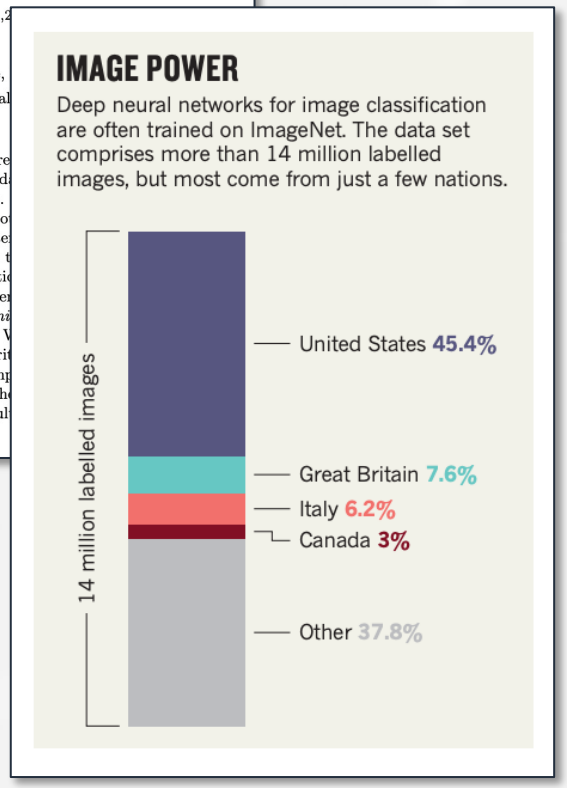
Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

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

**Abstract**

The blind application of machine learning runs the risk of amplifying biases pre
danger is facing us with *word embedding*, a popular framework to represent text d
has been used in many machine learning and natural language processing tasks.
word embeddings trained on Google News articles exhibit female/male gender stereo
extent. This raises concerns because their widespread use, as we describe, often te
biases. Geometrically, gender bias is first shown to be captured by a direction in t
Second, gender neutral words are shown to be linearly separable from gender definiti
embedding. Using these properties, we provide a methodology for modifying an e
gender stereotypes, such as the association between between the words *receptioni*
maintaining desired associations such as between the words *queen* and *female*.
quantify both direct and indirect gender biases in embeddings, and develop algorit
embedding. Using crowd-worker evaluation as well as standard benchmarks, we em
that our algorithms significantly reduce gender bias in embeddings while preserving th
such as the ability to cluster related concepts and to solve analogy tasks. The resul
be used in applications without amplifying gender bias.

https://arxiv.org/abs/1607.06520




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

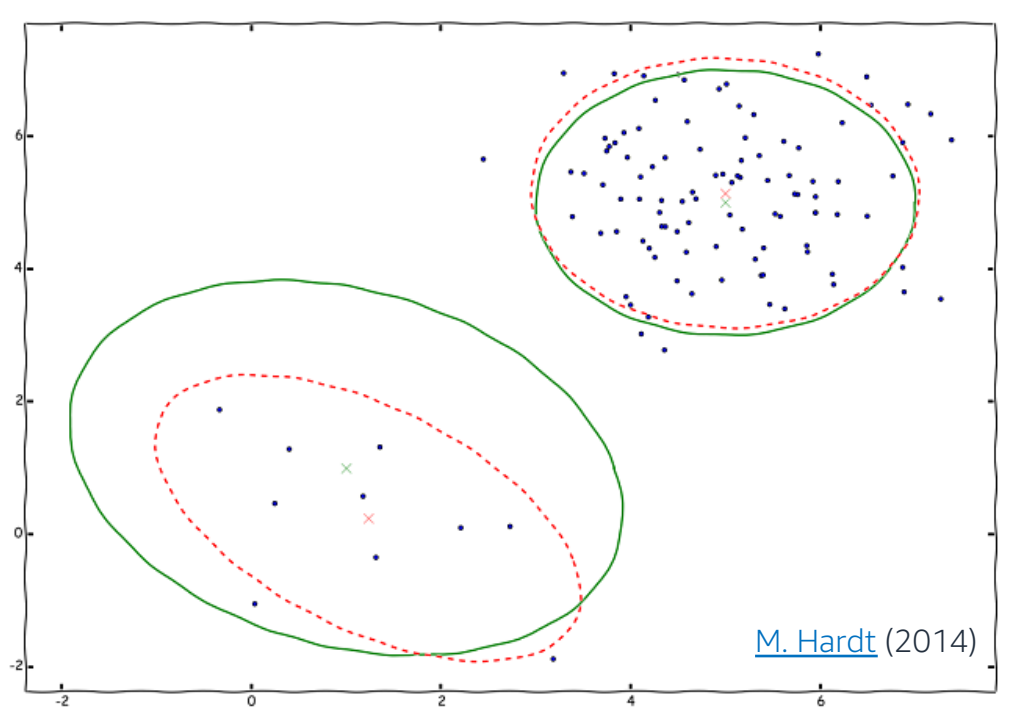
- Historical bias is often present in data collected from 'the wild'.
- Geographical representation is not equal for many of the most widely used computer vision datasets.

# Feature Sampling: Sampling Size

- By definition: Minority = group with fewer examples.
- Models generally improve performance with more samples.

→ even if feature quality is good & data is unbiased, predictions for small groups can be worse than for larger groups

→ carefully sample all groups represented in population and don't assume model generalizes by default




Dashed red describes estimated covariance matrices. Solid green defines correct covariance matrices. The green and red crosses indicate correct and estimated means, respectively.

# Label Quality

- Tainted labels: Labels generally describe what happened historically but don't tell us what the unbiased outcome would have been.

  → to mitigate bias we need to address underlying causes of inequality (bias transforming metrics)

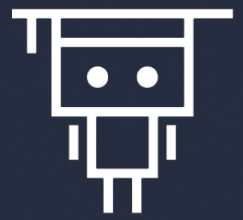
- Imprecise labels: Labels can be insufficiently precise to capture meaningful differences between cases.

  → collect additional information or clean labels before passing to ML model

Selbst et al. (2016), Wachter et al. (2021)

# Datasheets for Datasets

- Record of main characteristics of data and use case/limitations, e.g.:
  - Who collected data?
  - What purpose was data collected for?
  - When and where was data collected?
  - How was data collected?
  - Descriptive statistics of dataset
  - …



# Exploratory Data Analysis

# Exploratory Data Analysis

EDA (Exploratory Data Analysis) is an approach to analyze a dataset and capture main characteristics of it. Find correlations, missing data, check distributions.

- General ML: Perform initial investigations to discover patterns, spot anomalies, test hypothesis and check assumptions.
- Responsible ML: Identify data collection gaps, inform further feature processing and detect societal/historical bias.

# Descriptive Statistics

- **Overall statistics** – df.head(), df.shape, df.info()
  - » Number of datapoints (i.e. number of rows)
  - » Number of features (i.e. number of columns)

- **Univariate statistics** (single feature)
  - » Statistics for numerical features (mean, variance, histogram) - df.describe(), hist(df[feature])
  - » Statistics for categorical features (histograms, mode, most/least frequent values, percentage, number of unique values) df[feature].value_counts() or seaborn's distplot()

- **Multivariate statistics** (more than one feature)
  - » Correlations - df.plot.scatter(feature1, feature2), df[[feature1, feature2]].corr()

# Correlations

Correlations: How strongly pairs of features are related.

- High feature-feature correlation (positive or negative) features can degrade performance of some ML models. Check for correlation of sensitive attributes with other features to avoid serving **hidden proxies** to ML models.
  - » Proxy example: zip code and wealth
- Highly target-correlated (positive or negative) features might improve the performance certain models.

# Correlations: Correlation Matrix

```
cols = [feature1, feature2]
df[cols].corr()
```

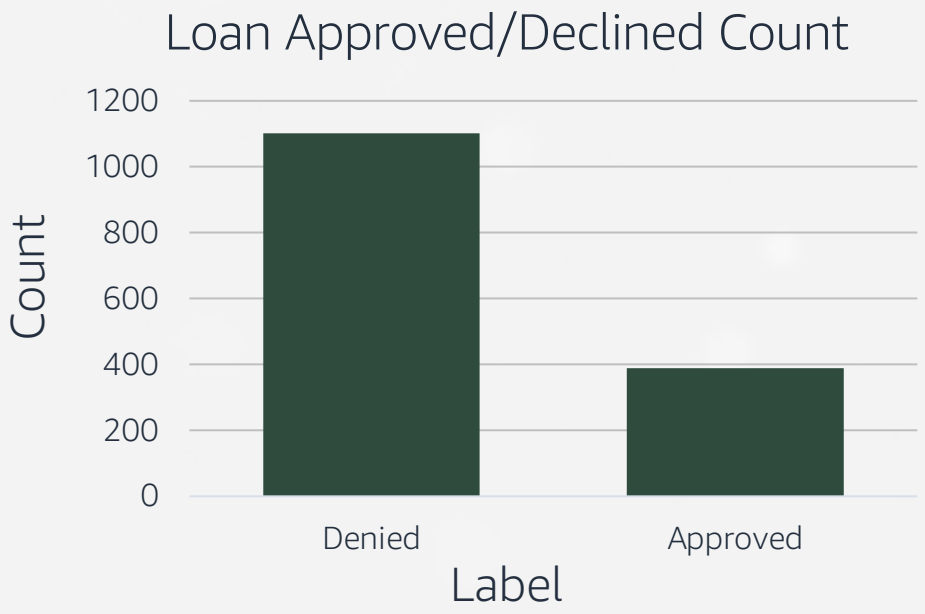
Correlation matrices measure the linear dependence between features.

| | feature1 | feature2 |
|---|---|---|
| feature1 | 1 | 0.0128493 |
| feature2 | 0.0128493 | 1 |

| | feature1 | feature2 |
|---|---|---|
| feature1 | 1 | 0.882106 |
| feature2 | 0.882106 | 1 |

Correlation values are between -1 (neg. correlation) and 1 (pos. correlation).

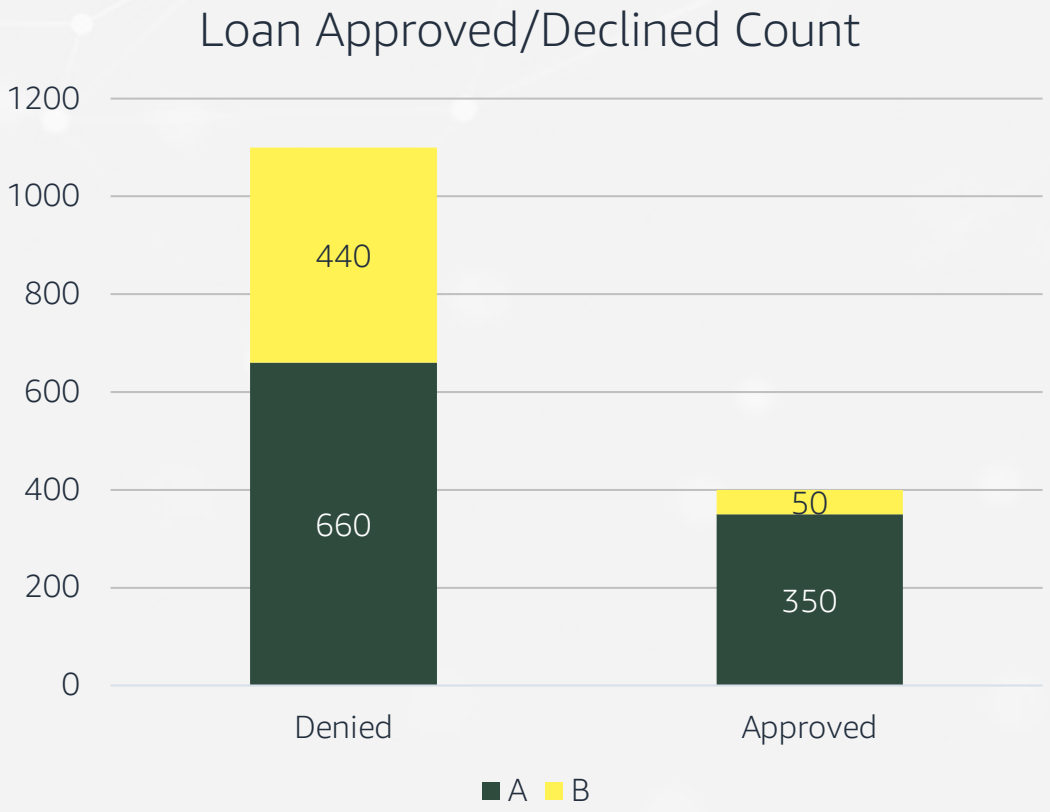
# Label Imbalance




Example: Will loan be approved (yes/no)?

- Samples per class not equally distributed.
- ML model may not work well for the infrequent labels:
  - » Models often assume equal distribution of classes.
  - » High accuracy paradox: If model works well for majority class, it is considered accurate.

# Label Imbalance




Example: Will loan be approved (yes/no)?

- Consider outcomes/labels per group.
- Beware of biased sampling and check for minority groups in dataset.
- There exist:
  - General ML measures to quantify label imbalance
  - Responsible ML measures to quantify imbalance per subpopulation

# Quantifying Imbalance

- Norm. Class Imbalance: Measures imbalance in number of data examples (count), n, between 2 subpopulations A & B:

$$CI_{norm} = \frac{n_A - n_B}{n_A + n_B} \quad \in [-1,1]$$

Positive CI values indicate that there are more data points with attribute value a (1 = only A, -1 only B).

# Quantifying Imbalance

- Norm. Class Imbalance: Measures imbalance in number of data examples (count), n, between 2 subpopulations A & B:

$$CI_{norm} = \frac{n_A - n_B}{n_A + n_B} \quad \in [-1,1]$$

Positive CI values indicate that there are more data points with attribute value a (1 = only A, -1 only B).

- Difference in Proportions of Labels: Measures difference in fraction of positive outcomes for 2 subpopulations A & B:

$$DPL = \frac{n_A^+}{n_A} - \frac{n_B^+}{n_B} \quad \in [-1,1]$$

Positive DPL values indicate that there are more positive outcomes for examples in group A.

# Why do we need metrics?

- Norm. Class Imbalance
  - » Detect bias in the model if there is not enough data for each class → obtain more data
- Difference in Proportions of Labels
  - » Detect bias in the dataset, maintain pre-training demographic parity

… many more pre-training metrics exist!

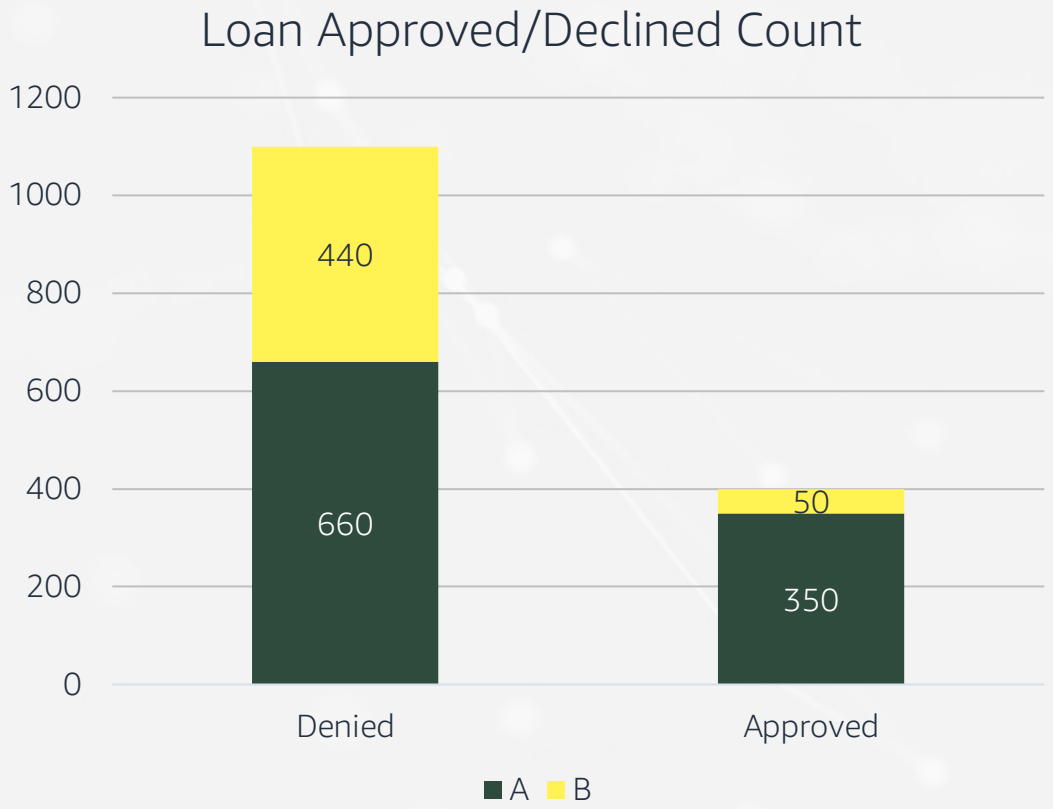
# Quantifying Imbalance

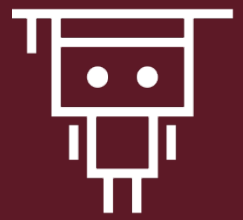
- Normalized Class Imbalance:

$$\frac{n_A \quad n_B}{n_{A\ +}\ n_B} = \frac{1010 - 490}{1010 + 490} \approx 0.35$$

- Difference in Proportions of Labels:

$$\frac{n_{A+}}{n_A} - \frac{n_{B+}}{n_B} = \frac{350}{1010} - \frac{50}{490} \approx 0.24$$





# Notebook: MLA-RESML-EDA.ipynb

# MLA-RESML-EDA.ipynb Notebook

- This notebook shows how to quantify and visualize correlations (scatter plots, correlation matrix) and generate descriptive statistics (histogram).

- To measure bias (assess quality of our data) before training a model, we will use $CI_{norm}$ and DPL.

- We will see a mix of data types (numerical, categorical data).