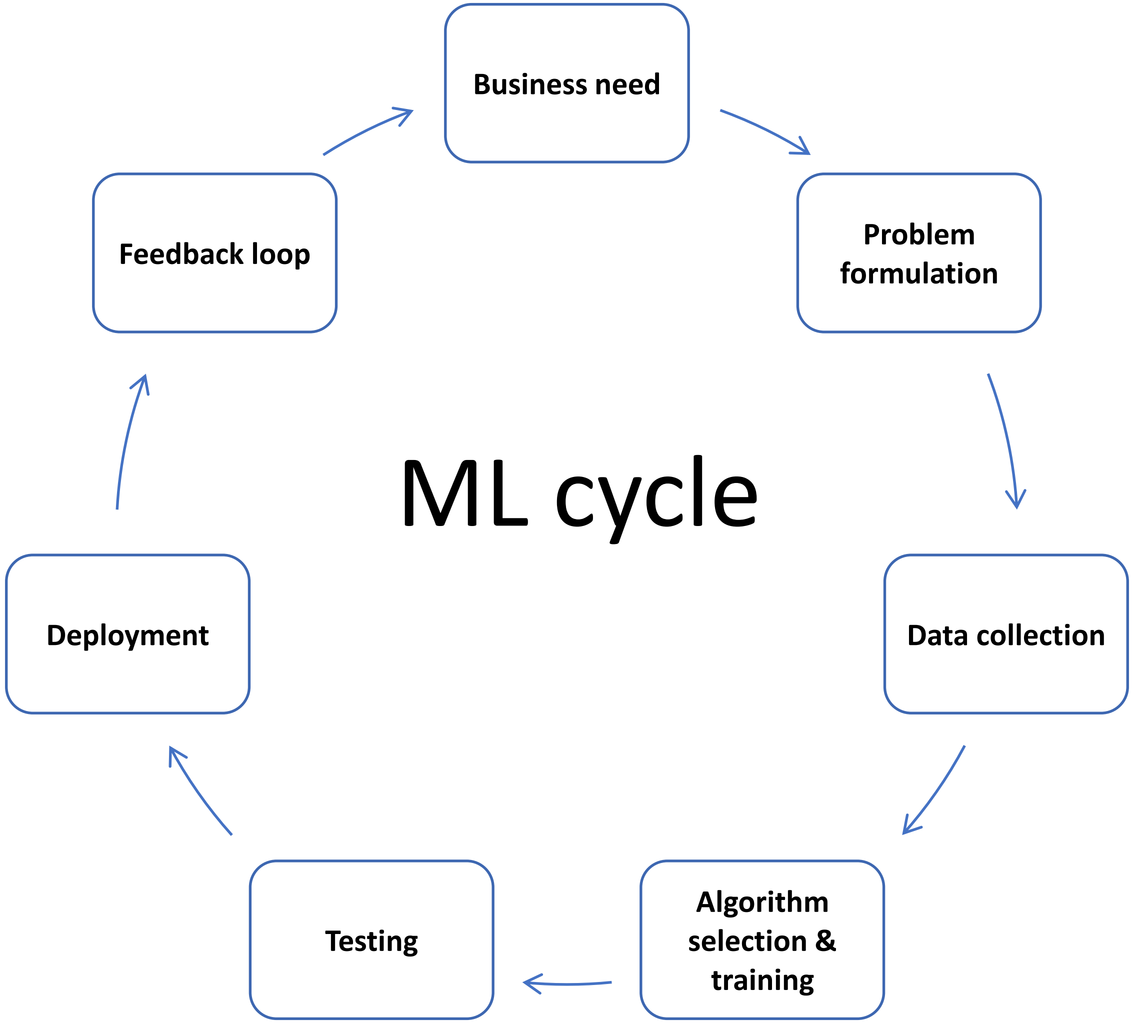
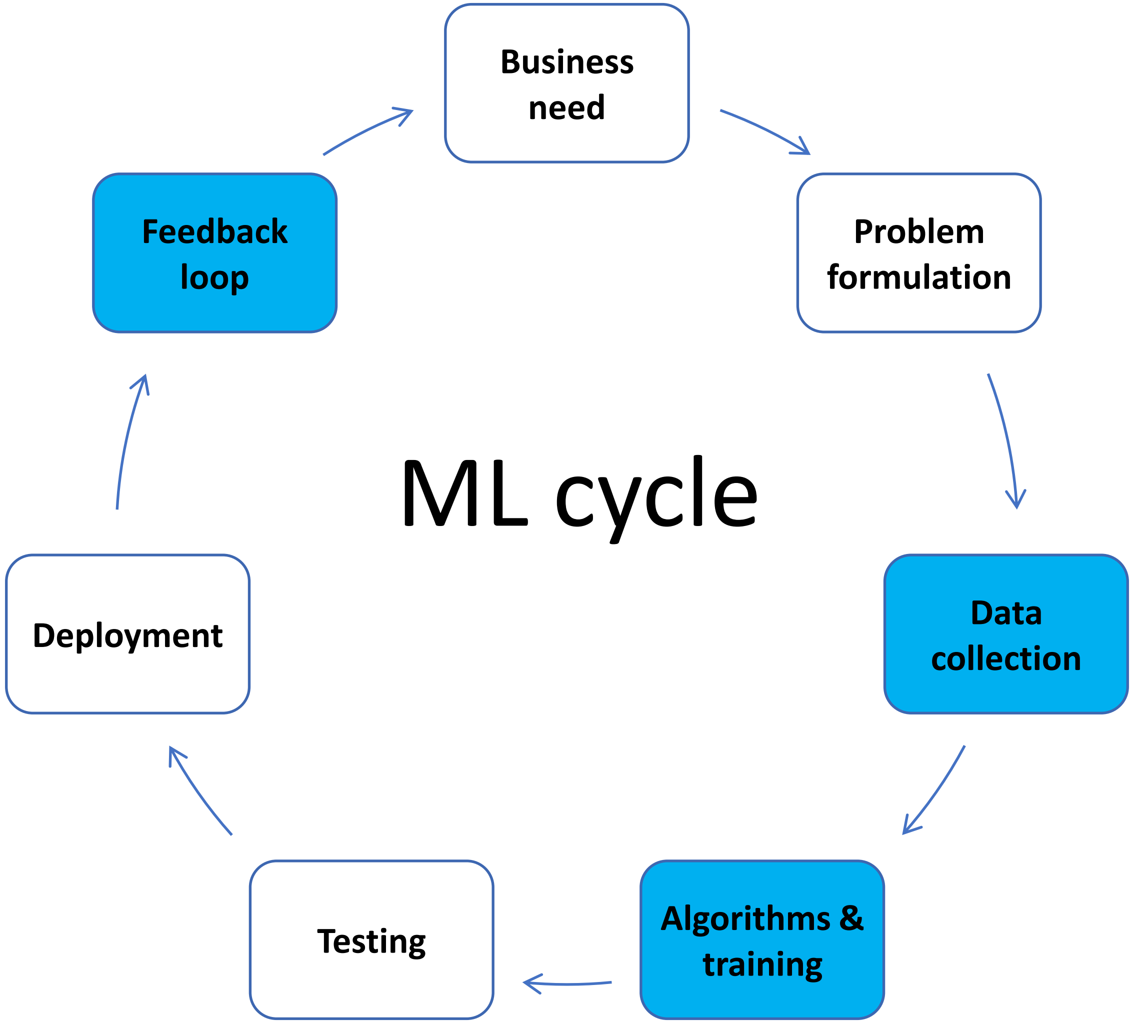
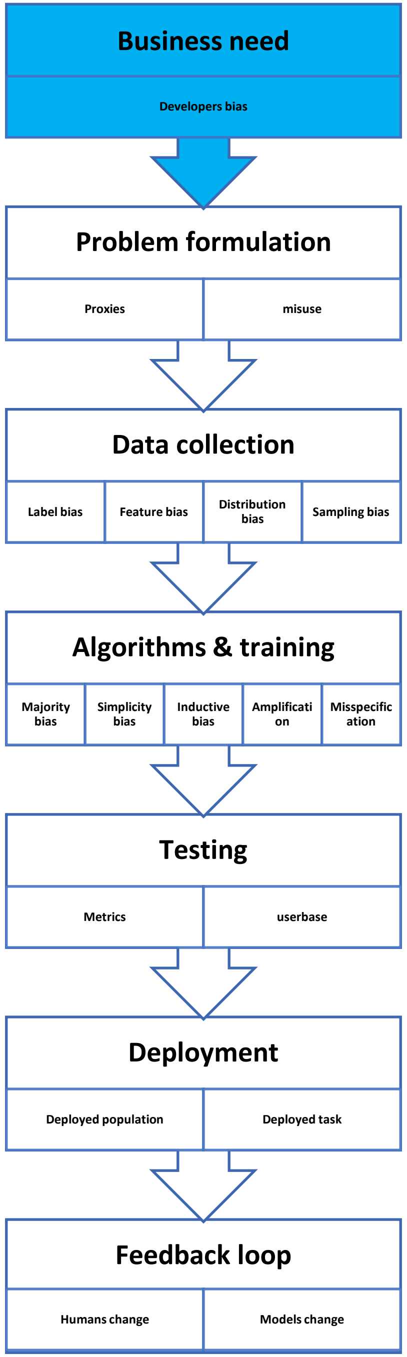
Biases in ML Cycle

Fereshte Khani

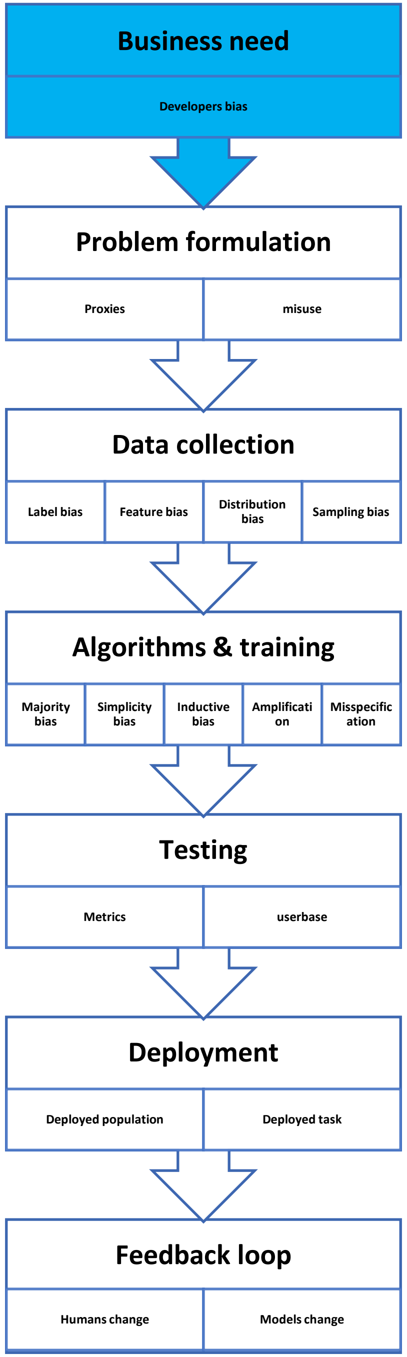

Business need
Problem formulation
Data collection
Algorithm selection & training
Testing
Deployment
Feedback loop
ML cycle


Business need
Problem formulation
ML cycle
Data collection
Algorithms & training
Testing
Deployment
Feedback loop



Developers' biases

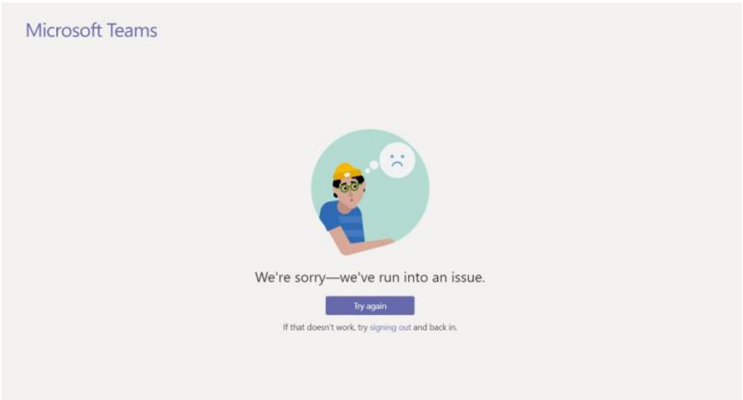
- Only addressing needs of a particular group
- Only causing harm to a particular group



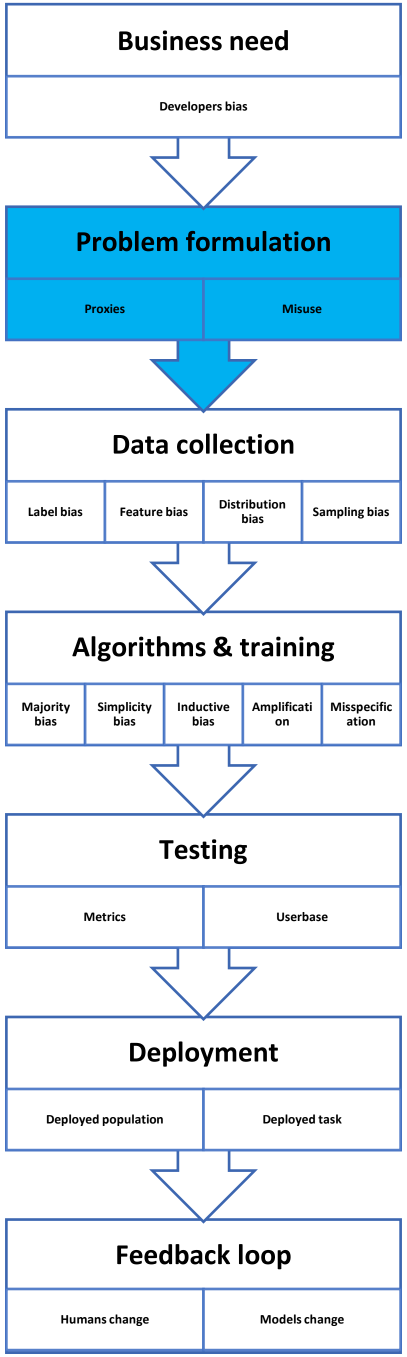
Developers' biases

- Only addressing needs of a particular group
- Only causing harm to a particular group

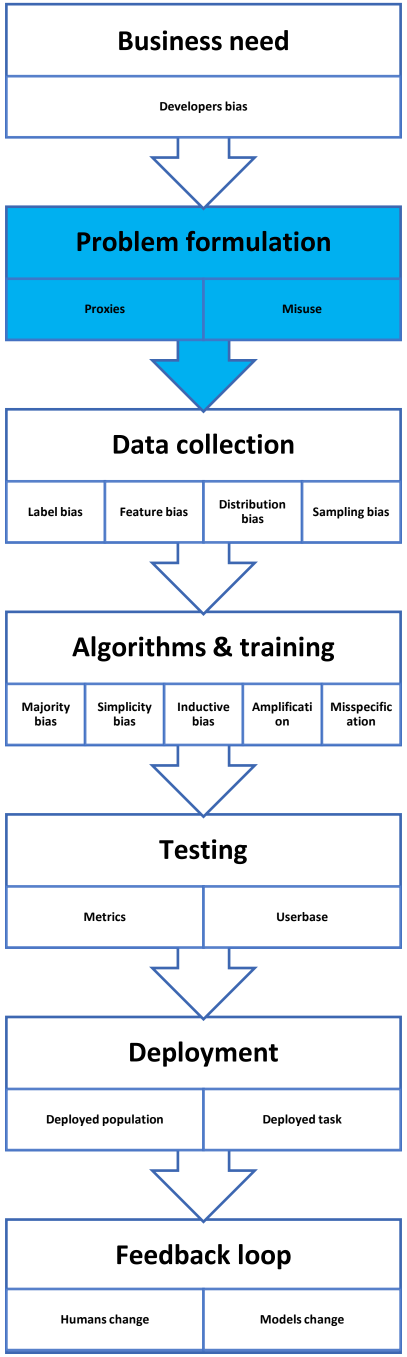
More features may serve tech savvy people



More features cause crashes on low bandwidth network or yield harder UI

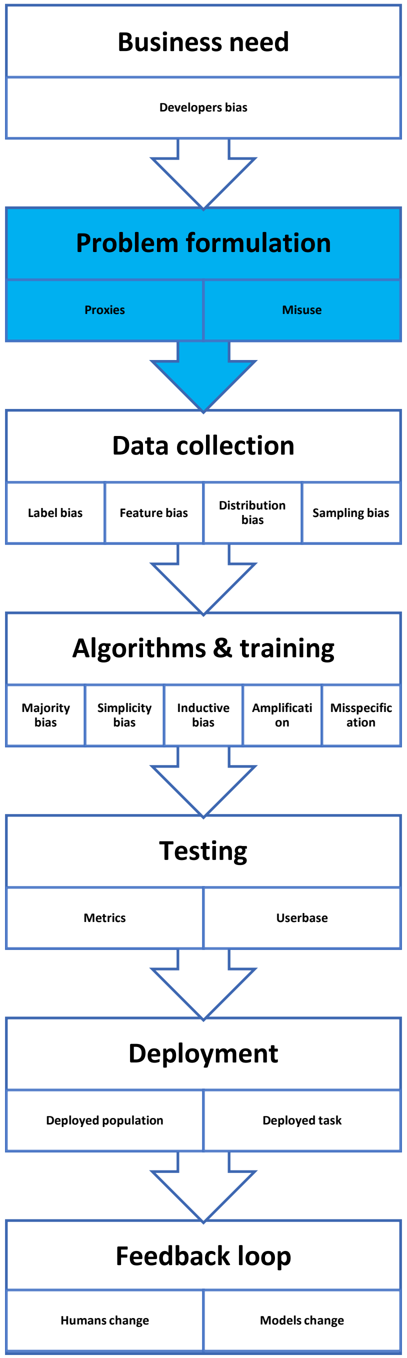


Defining a set of features ➔ Defining a target



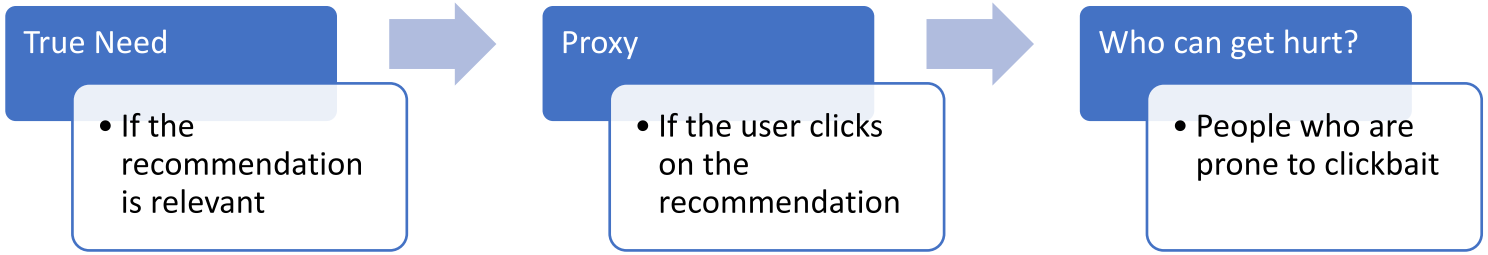
Defining a set of features ➔ Defining a target

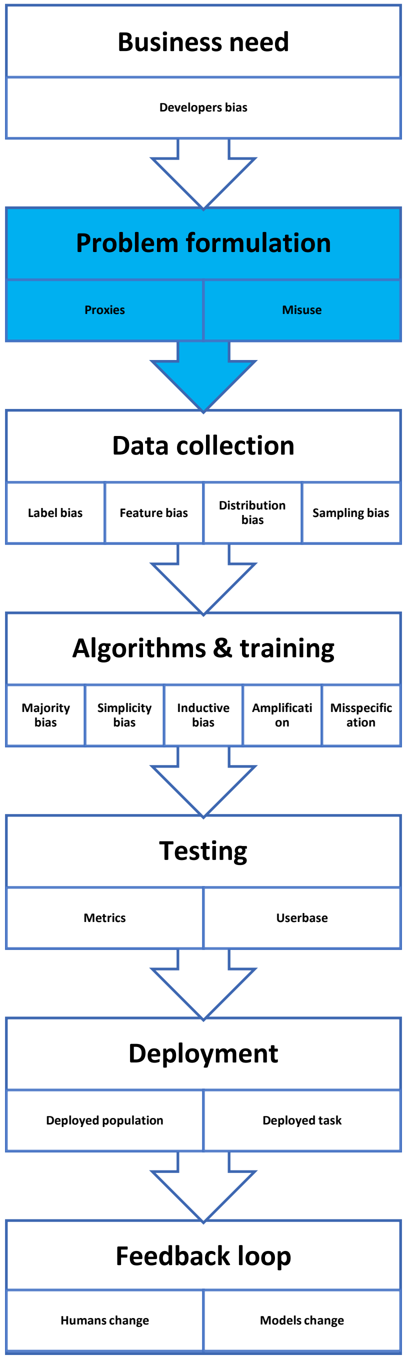
The true need is different from its proxy and proxies might adversely affect a particular group



Defining a set of features ➔ Defining a target

The true need is different from its proxy and proxies might adversely affect a particular group





Defining a set of features ➔ Defining a target

The true need is different from its proxy and proxies might adversely affect a particular group

True Need
- If the recommendation is relevant

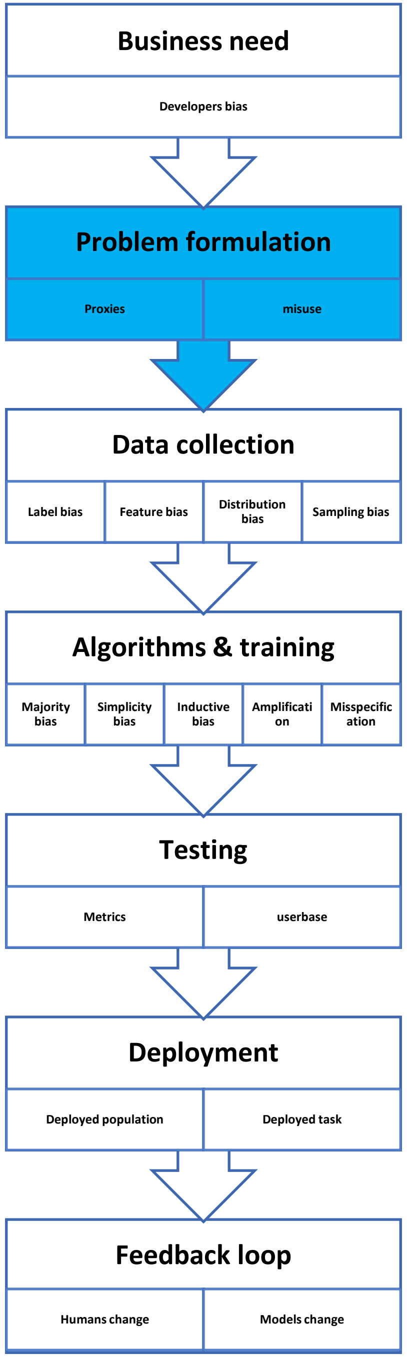
Proxy
- If the user clicks on the recommendation

Who can get hurt?
- people who are prune to clickbait

True Need
- Previous interests of the user

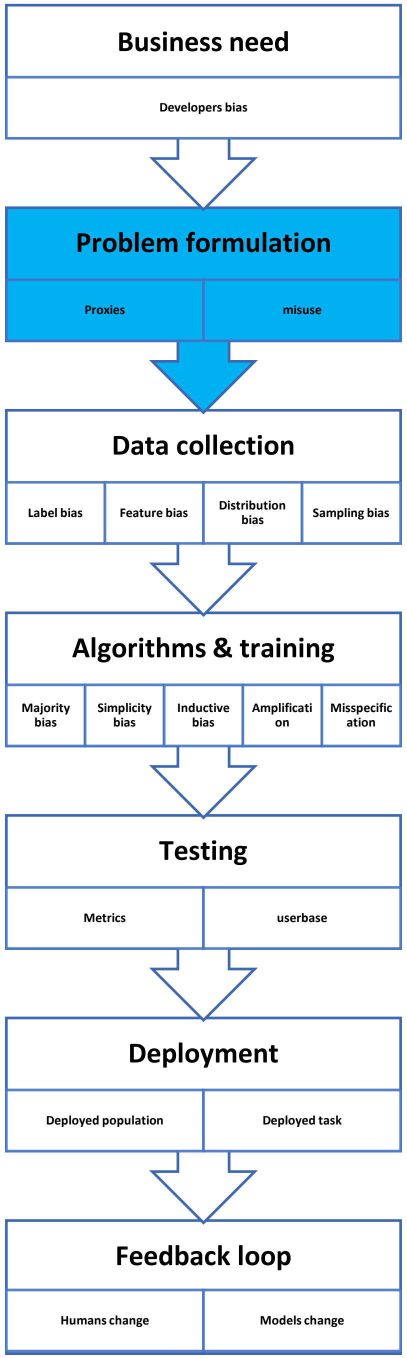
Proxy
- History of her content

Who can get hurt?
- People who share computers with each other



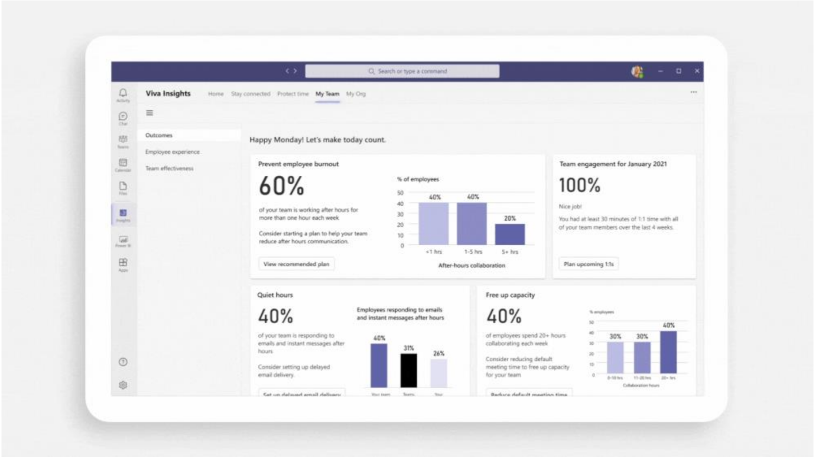
Defining a set of features ➔ Defining a target

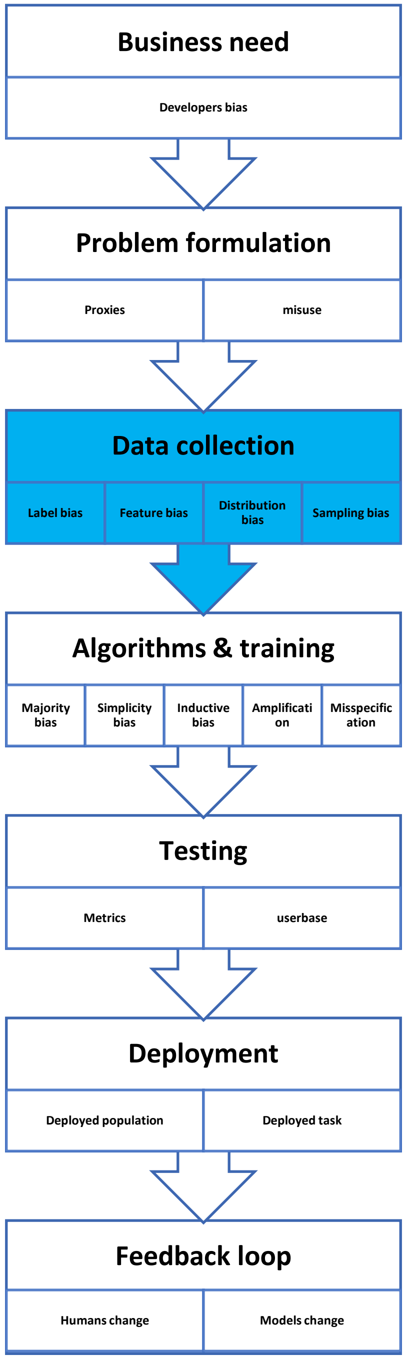
- Can the problem formulation be applied to unintended uses case to harm some groups?



Defining a set of features ➔ Defining a target

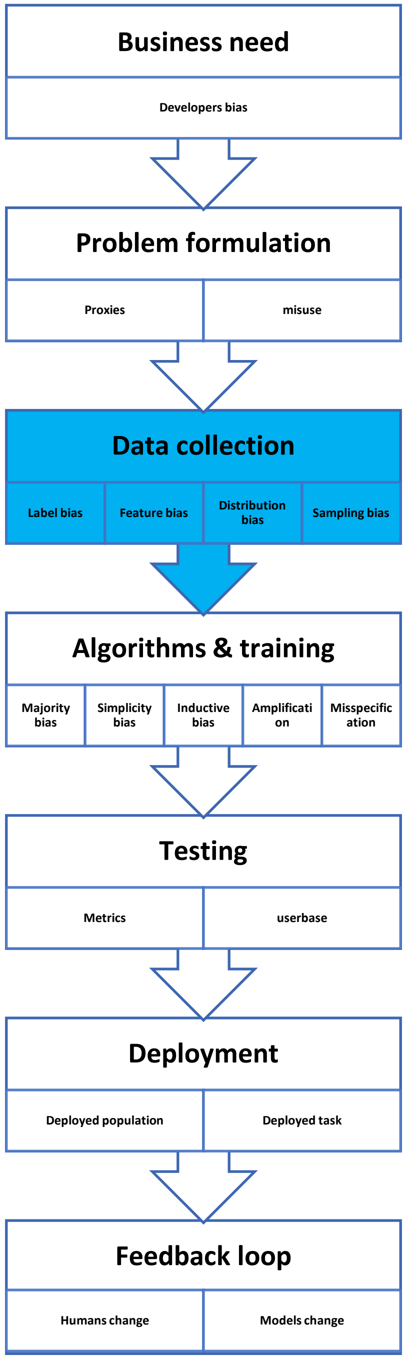
- Can the problem formulation be applied to unintended uses case to harm some groups?



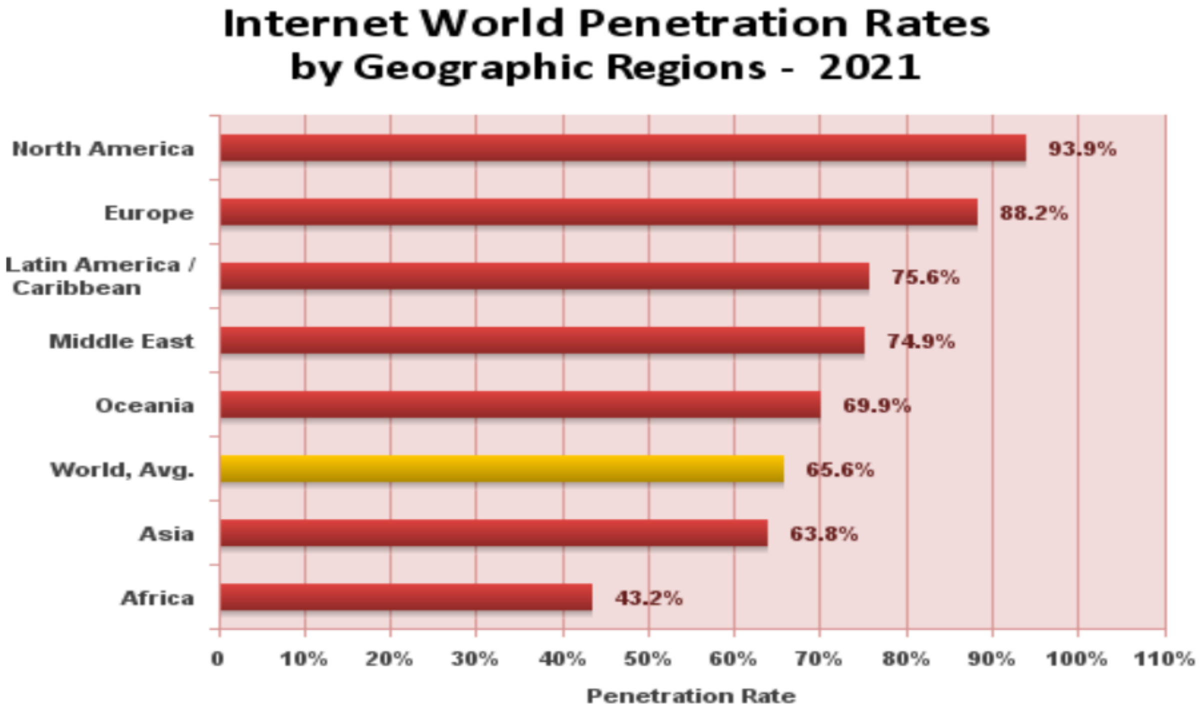


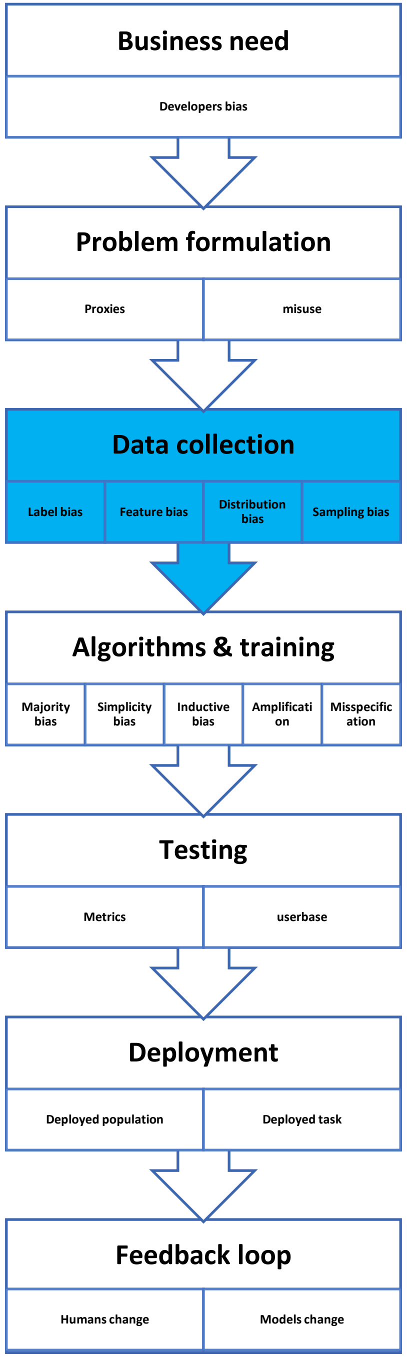
ML models learn patterns from *previously* collected data. The data usually reflect the longstanding discrimination against protected groups.

- People did not have voting right just because of their sex
- People were enslaved just because of their race
- People get fired because of their sexuality

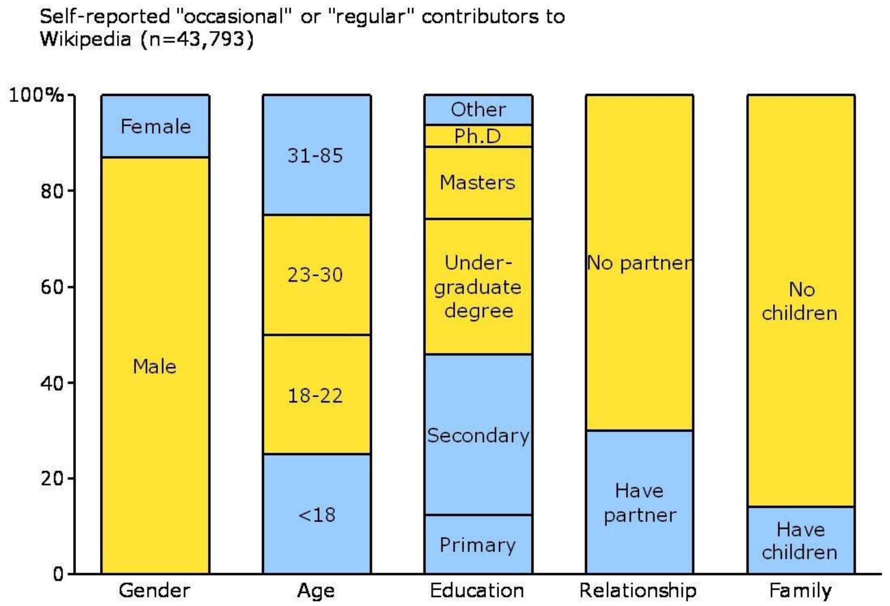


ML models learn patterns from *previously* collected data. The data usually reflect the longstanding discrimination against protected groups.

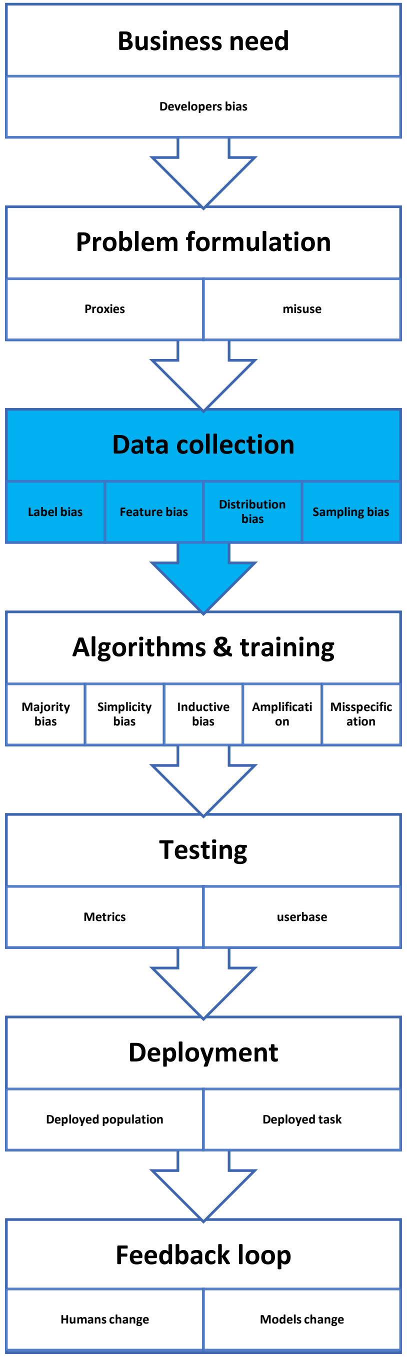




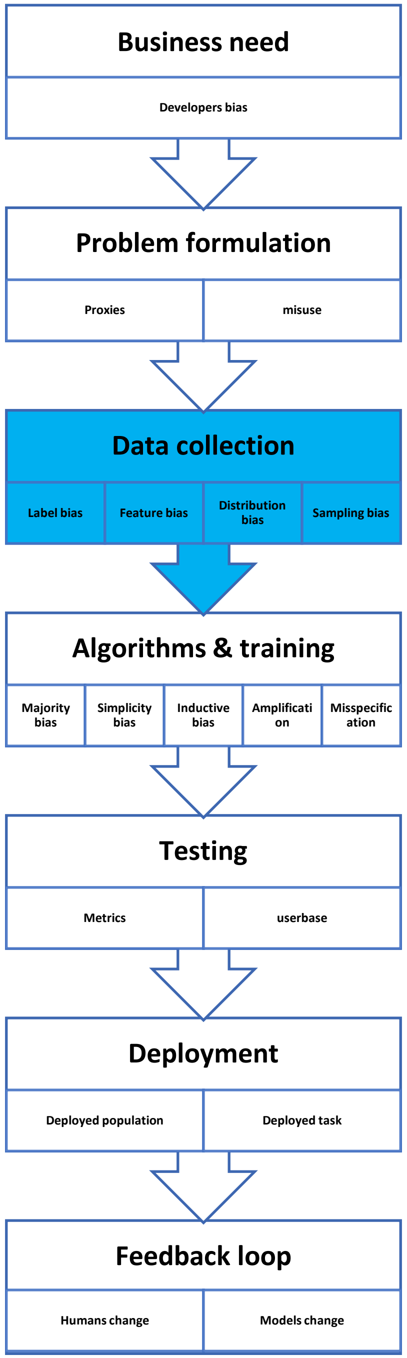
ML models learn patterns from *previously* collected data. The data usually reflect the longstanding discrimination against protected groups.



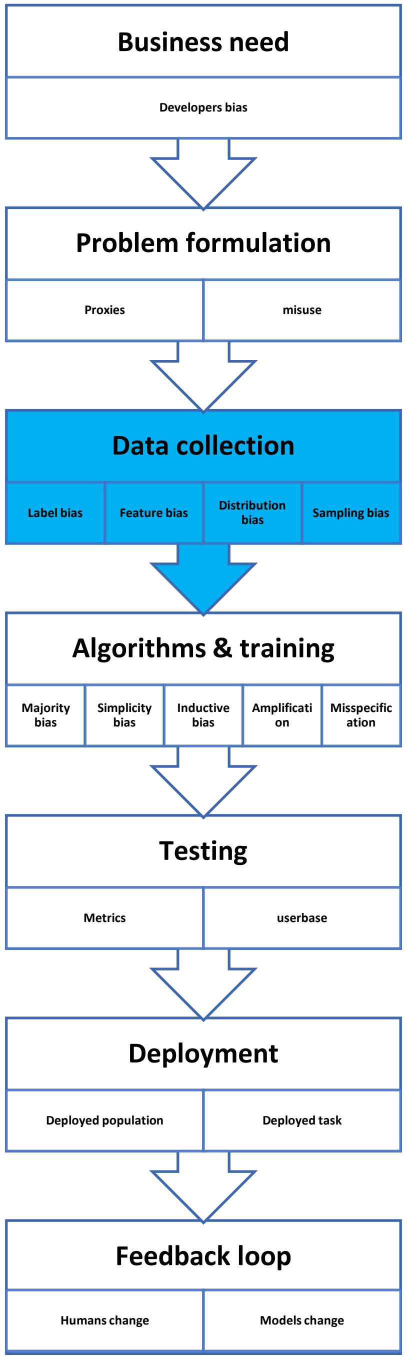
Note: Data for age category also includes respondents who were not contributors but who did read Wikipedia. Average age for contributors is 26.8 (vs. 25.3 for readers). "Regular" contributors include authors, editors, and administrators. "Occasional" contributors include readers who occasionally contribute as authors or editors.
Source: "Wikipedia Survey – First Results," UNU-MERIT, April 2009



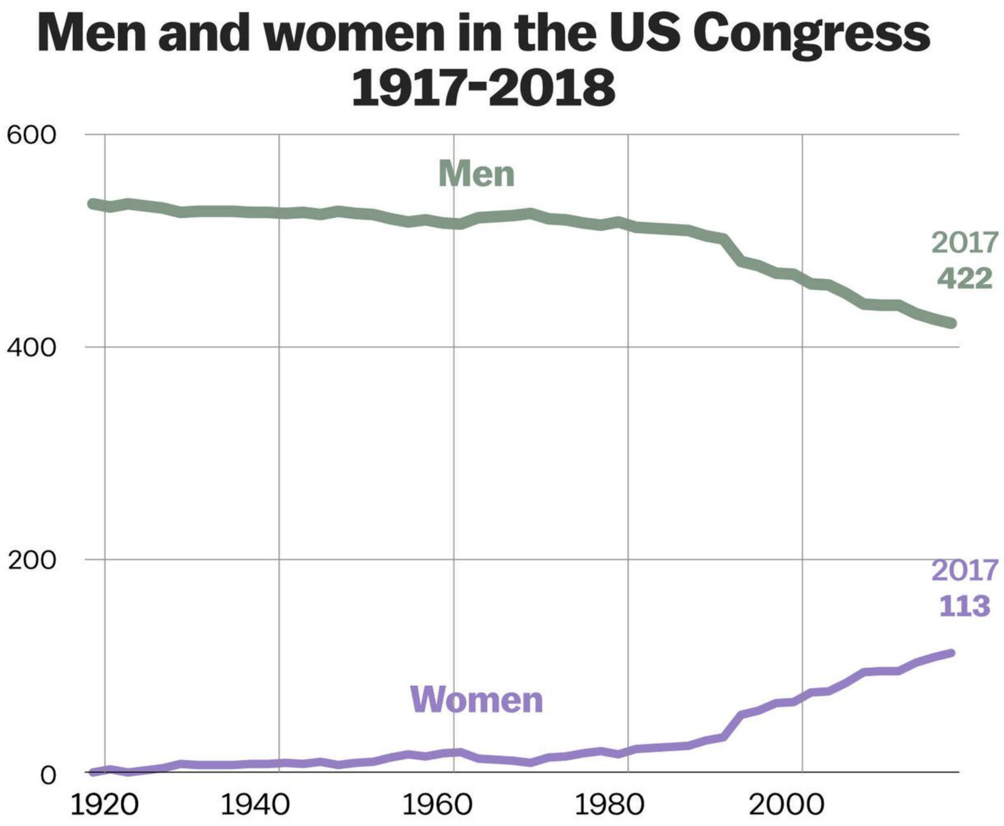
- ***Label bias:*** Labels are biased toward one group
  - Measuring whether women are qualified based on whether they did or didn't get hired
  - Measuring whether white populations reoffend based on re-arrest rates
  - Different countries have different norms when scoring a Microsoft application
    - In Japan, customers tend to rate customer satisfaction loyalty lower compared to other countries.
    - In Latin America, customers typically rate higher satisfaction compared to other regions.

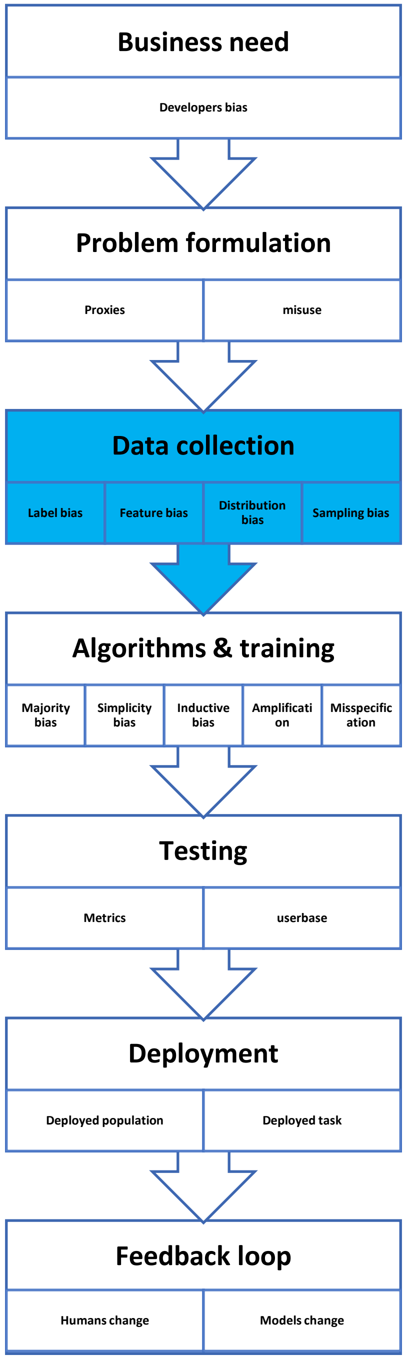


- ***Feature bias:*** features are biased toward one group
    - Considering number of previous arrests yields biases against racial minorities

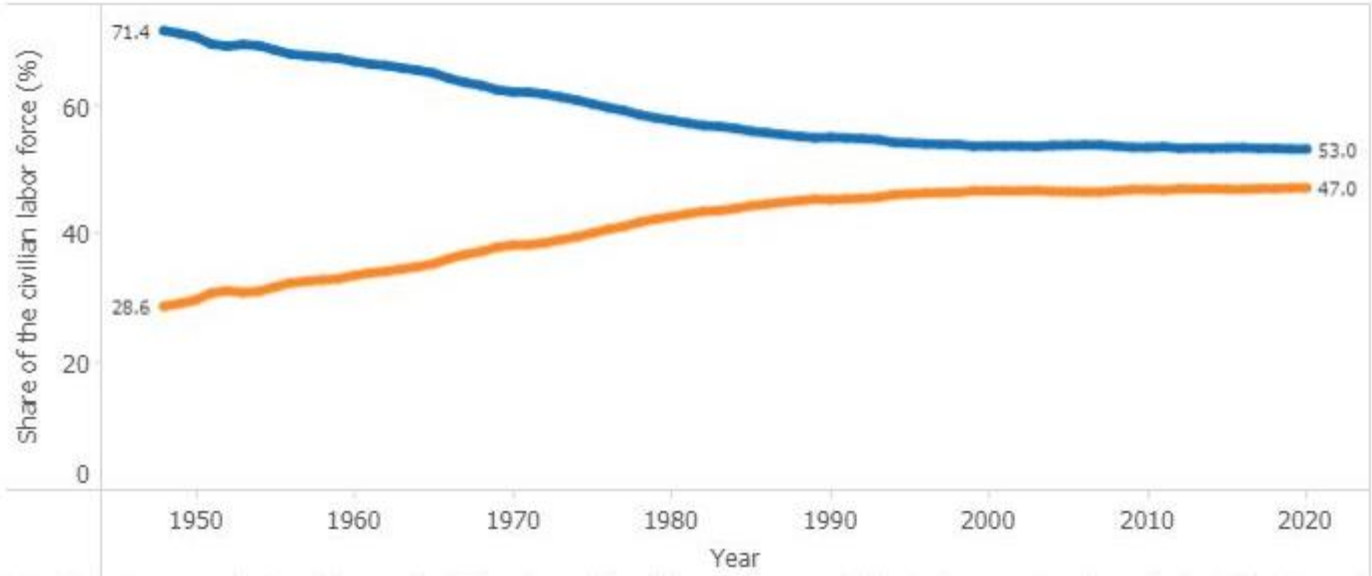


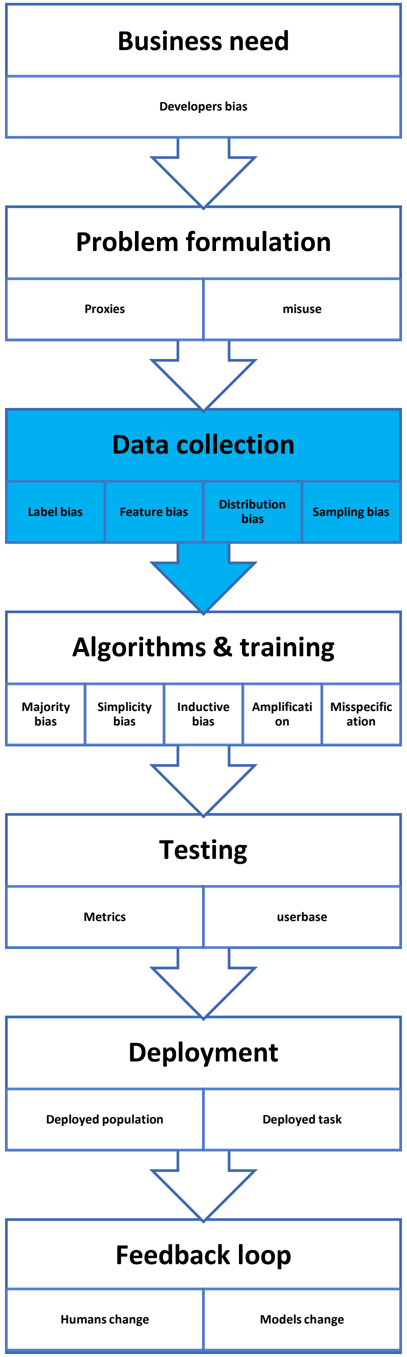
- ***Distribution bias:*** Historical discrimination creates a gap between the distribution of different groups



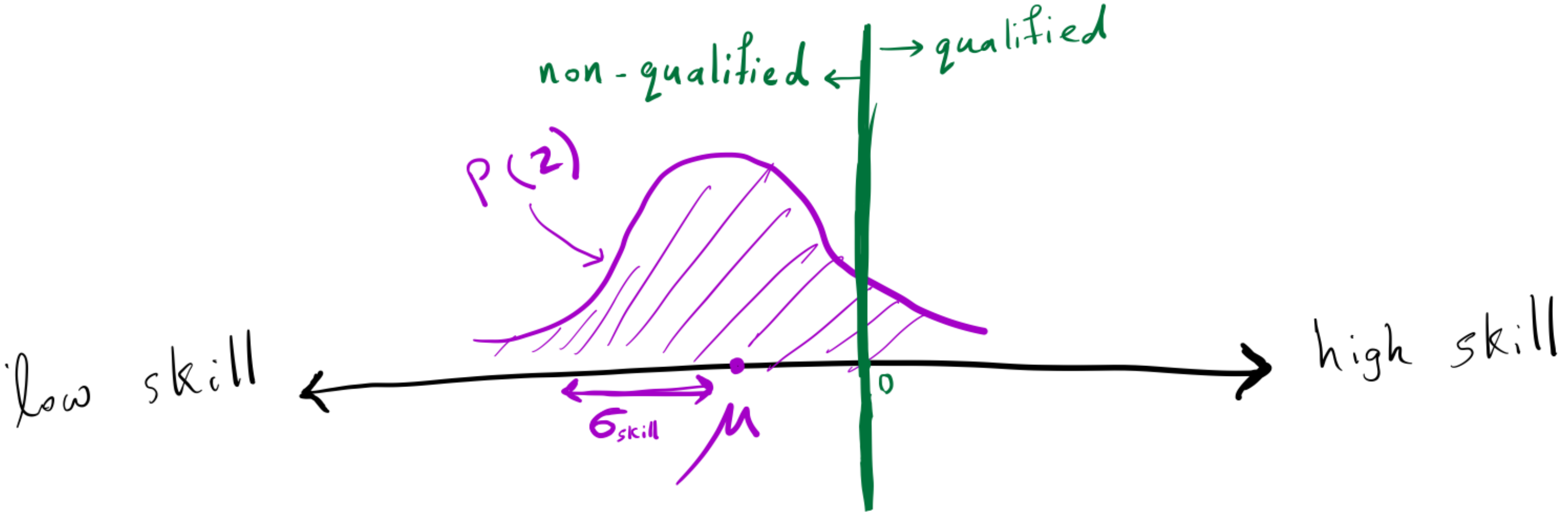


- ***Distribution bias:*** Historical discrimination creates a gap between the distribution of different groups

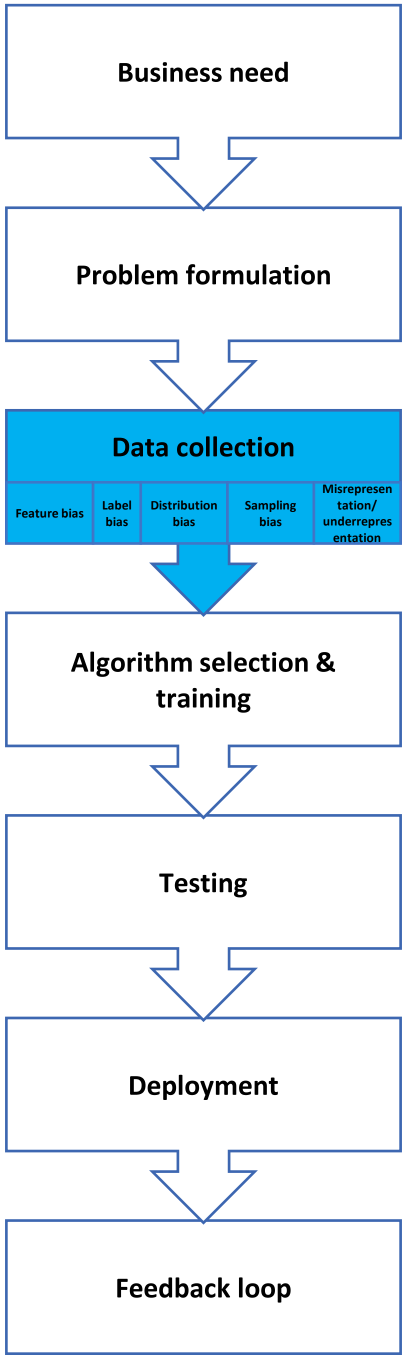




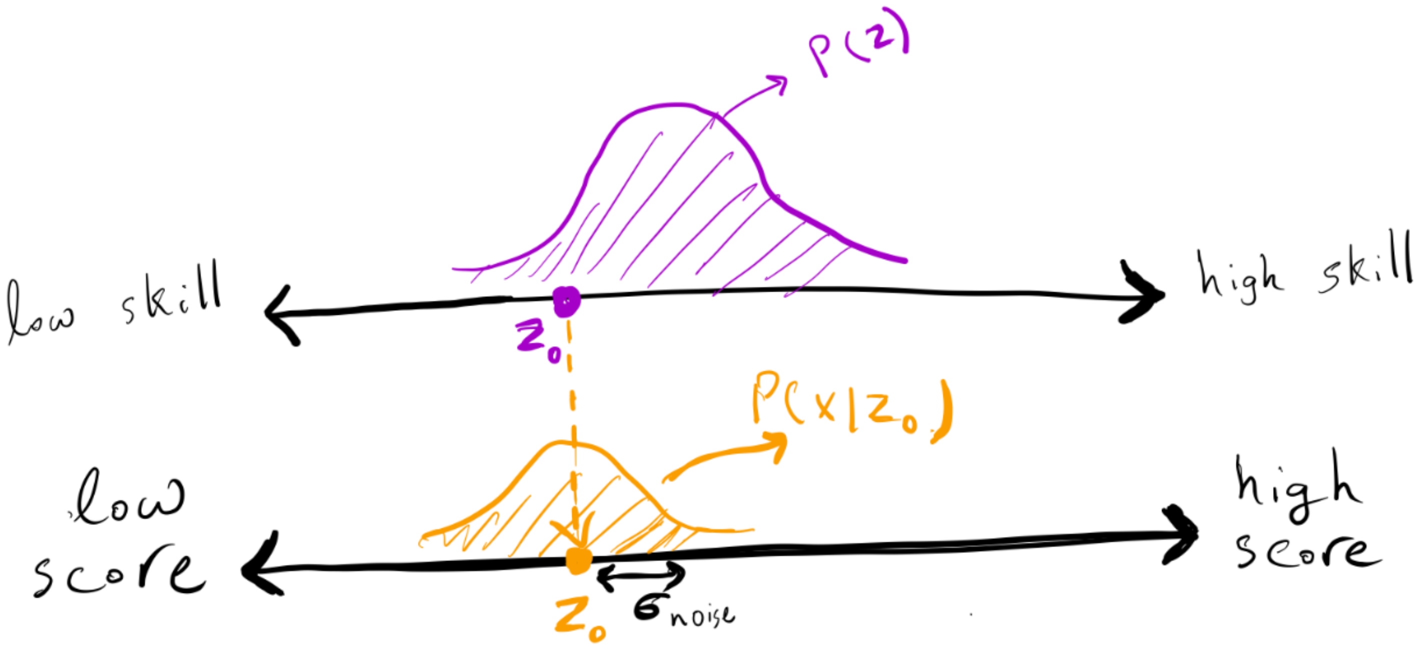
- ***Why distribution bias causes discrimination?***



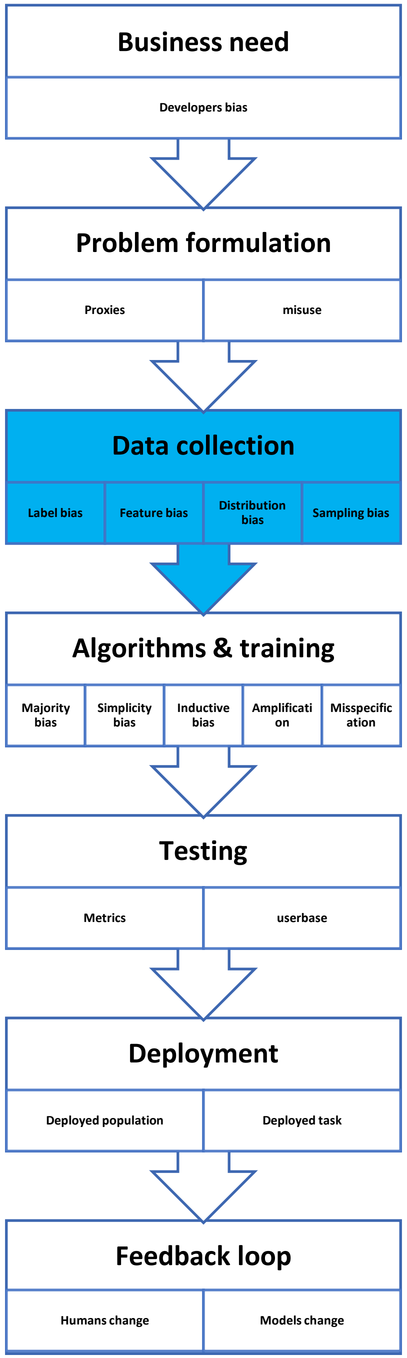
A company wants to hire people who are qualified (their skill level is greater than 0).



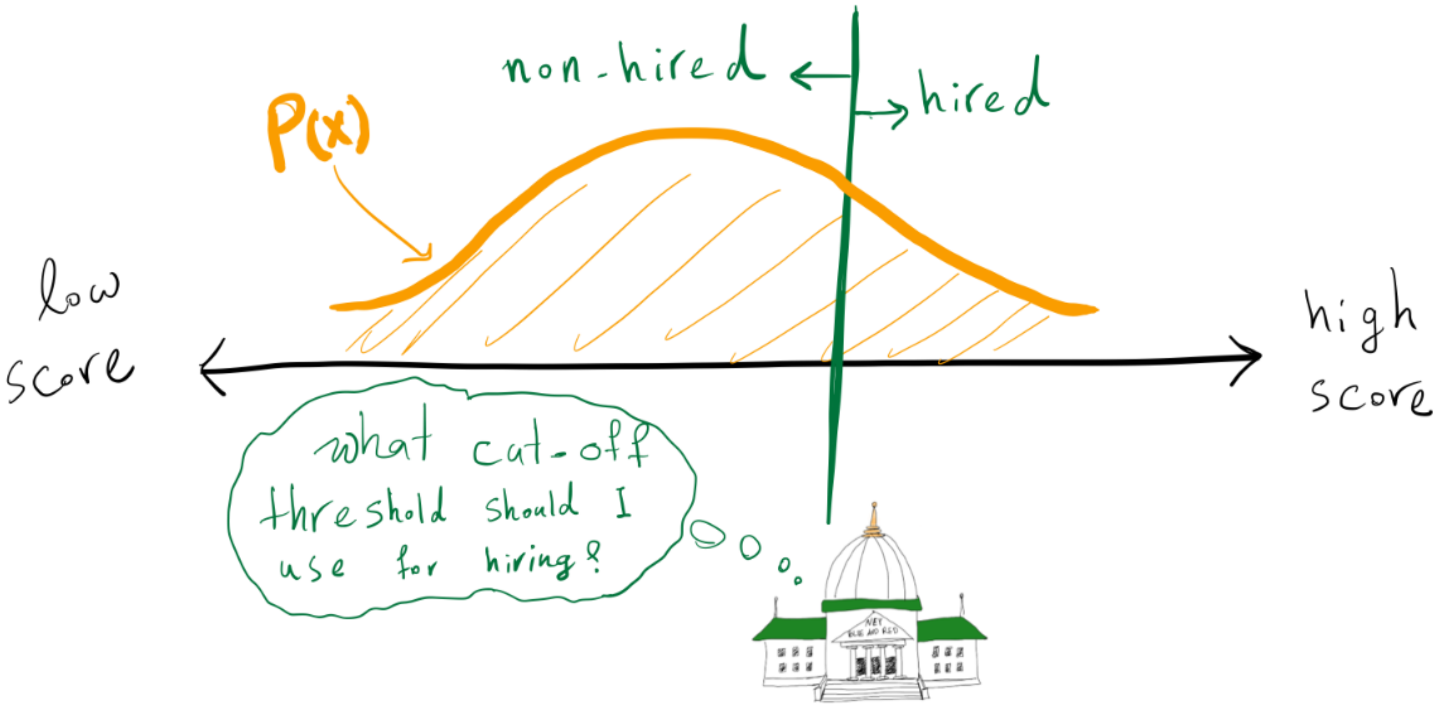
- ***Why distribution bias causes discrimination?***

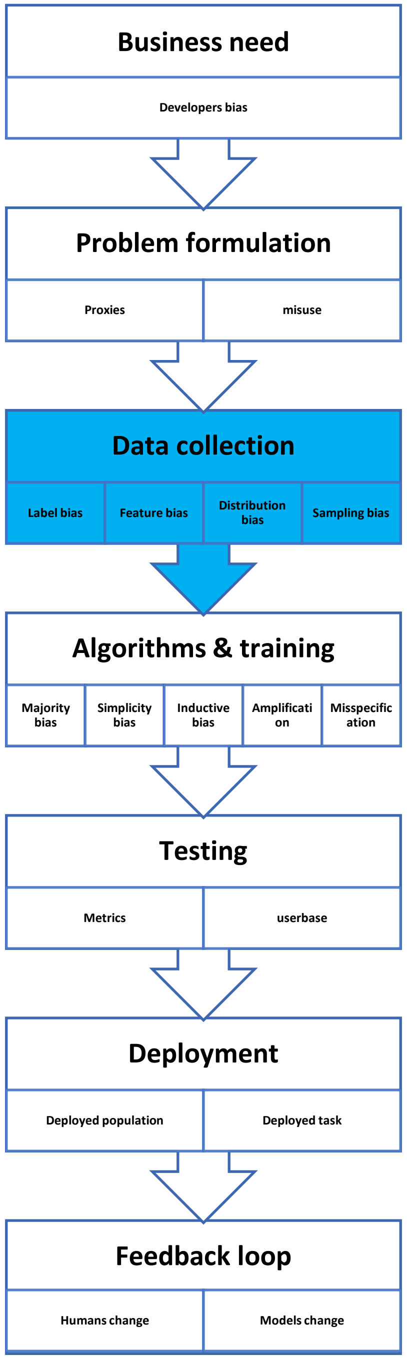


Everyone takes an exam! Score is a noisy version of skill level.

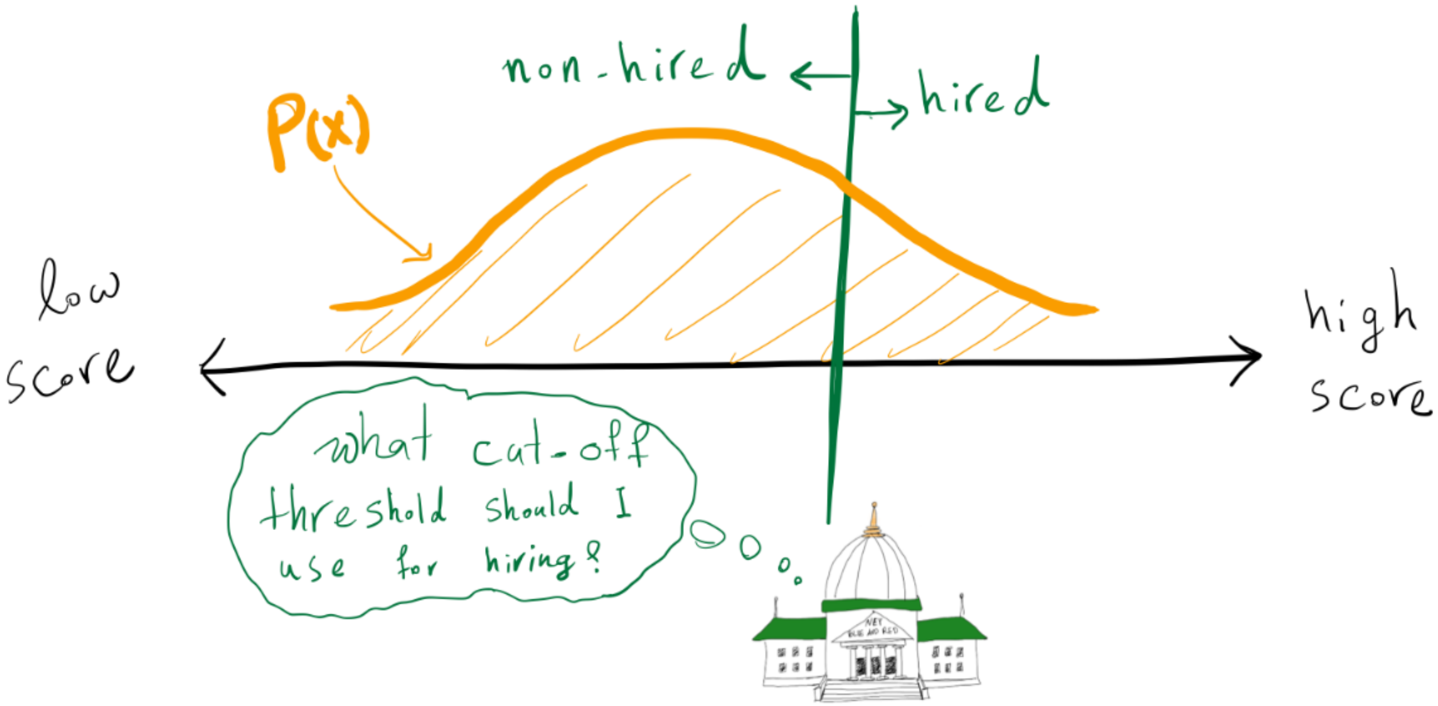


- ***Why distribution bias causes discrimination?***





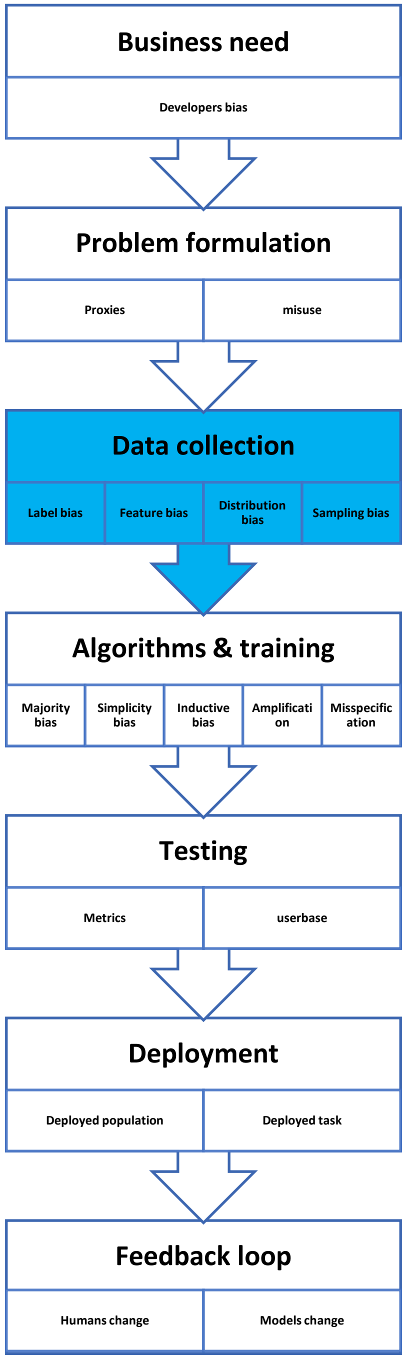
- ***Why distribution bias causes discrimination?***



Consider extreme cases:

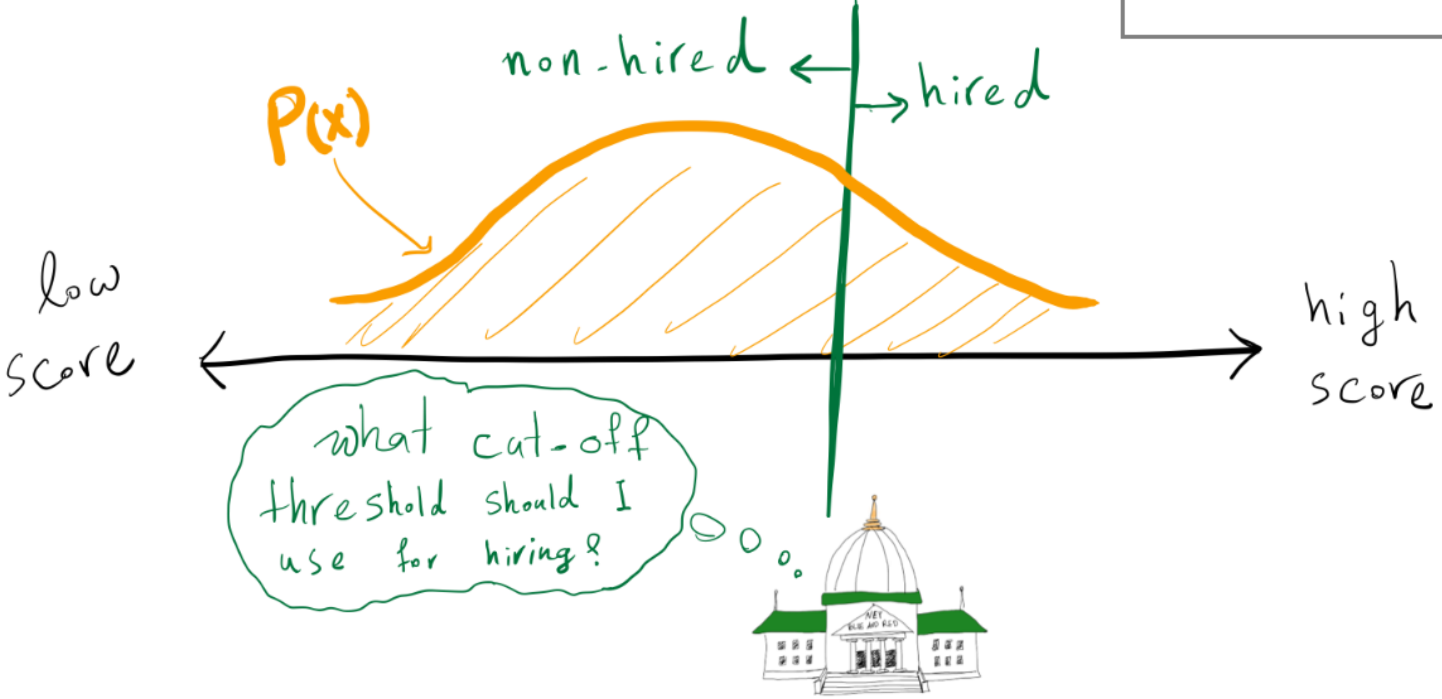
$$\sigma_{noise} = 0 \text{ then hire if } score > 0$$

$$\sigma_{noise} = \infty \text{ then hire if } \mu > 0$$



- ***Why distribution bias causes discrimination?***

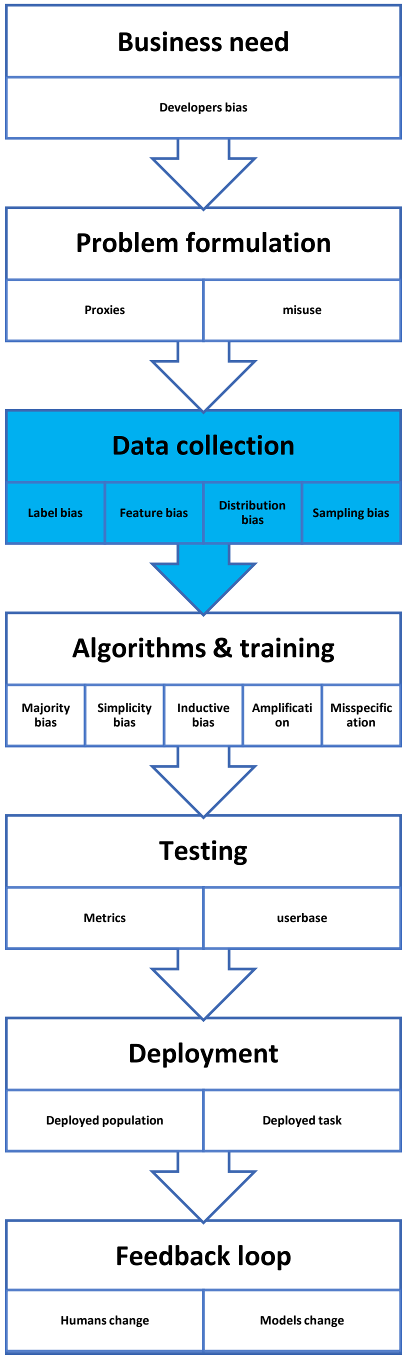
$$\text{optimal threshold} = -\mu \frac{\sigma^2_{\text{noise}}}{\sigma^2_{\text{skill}}}$$



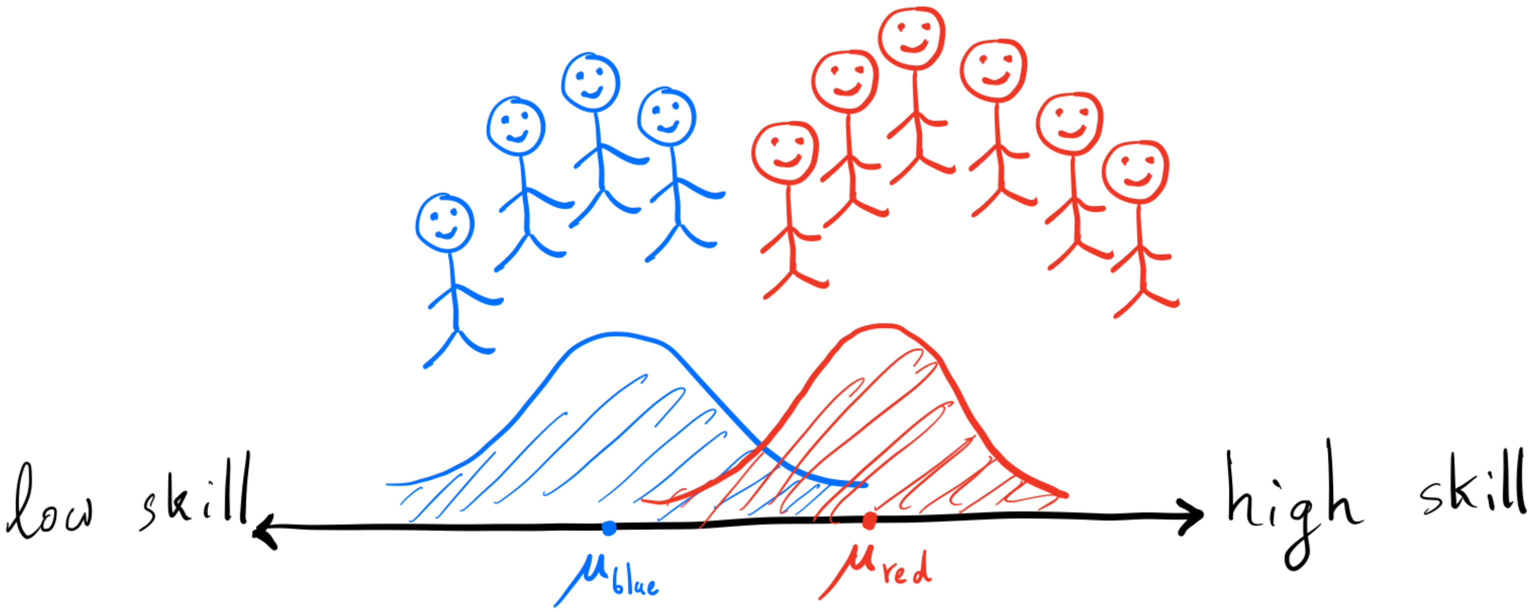
Consider extreme cases:

$\sigma_{noise} = 0$ then hire if $score > 0$

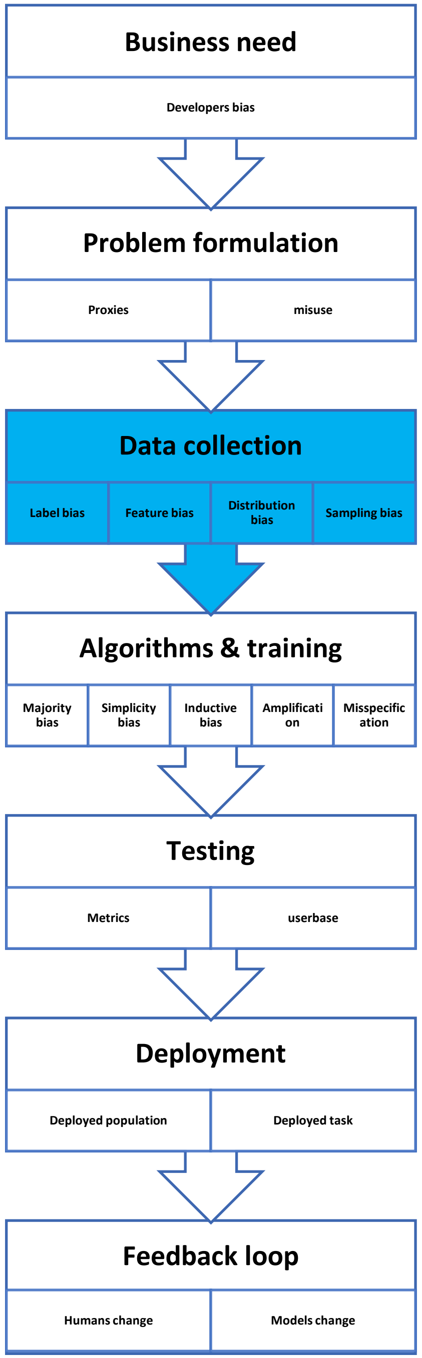
$\sigma_{noise} = \infty$ then hire if $\mu > 0$



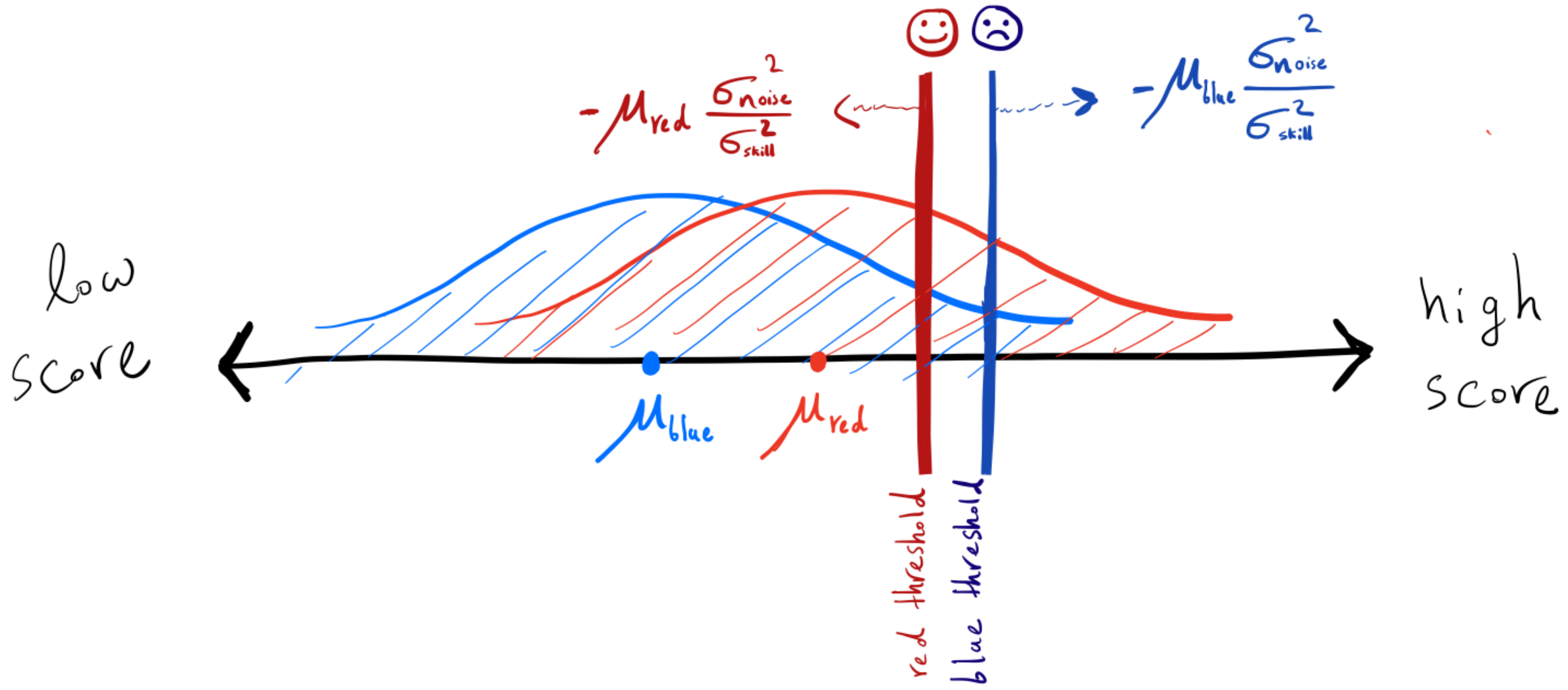
- ***Why distribution bias causes discrimination?***



Imagine that due to previous historical discrimination there is artificial divergence between skill level of blue and red people

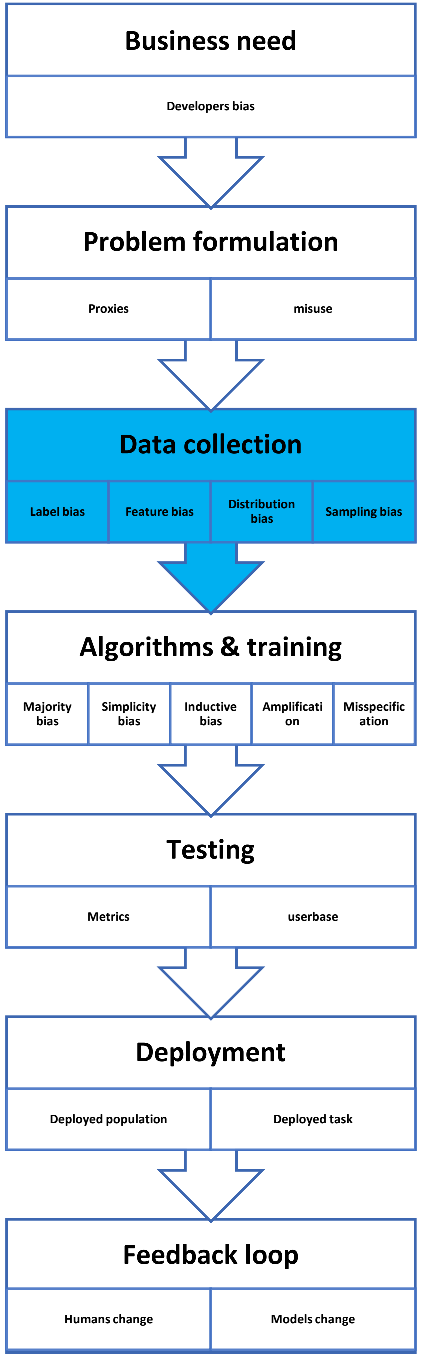


- ***Why distribution bias causes discrimination?***

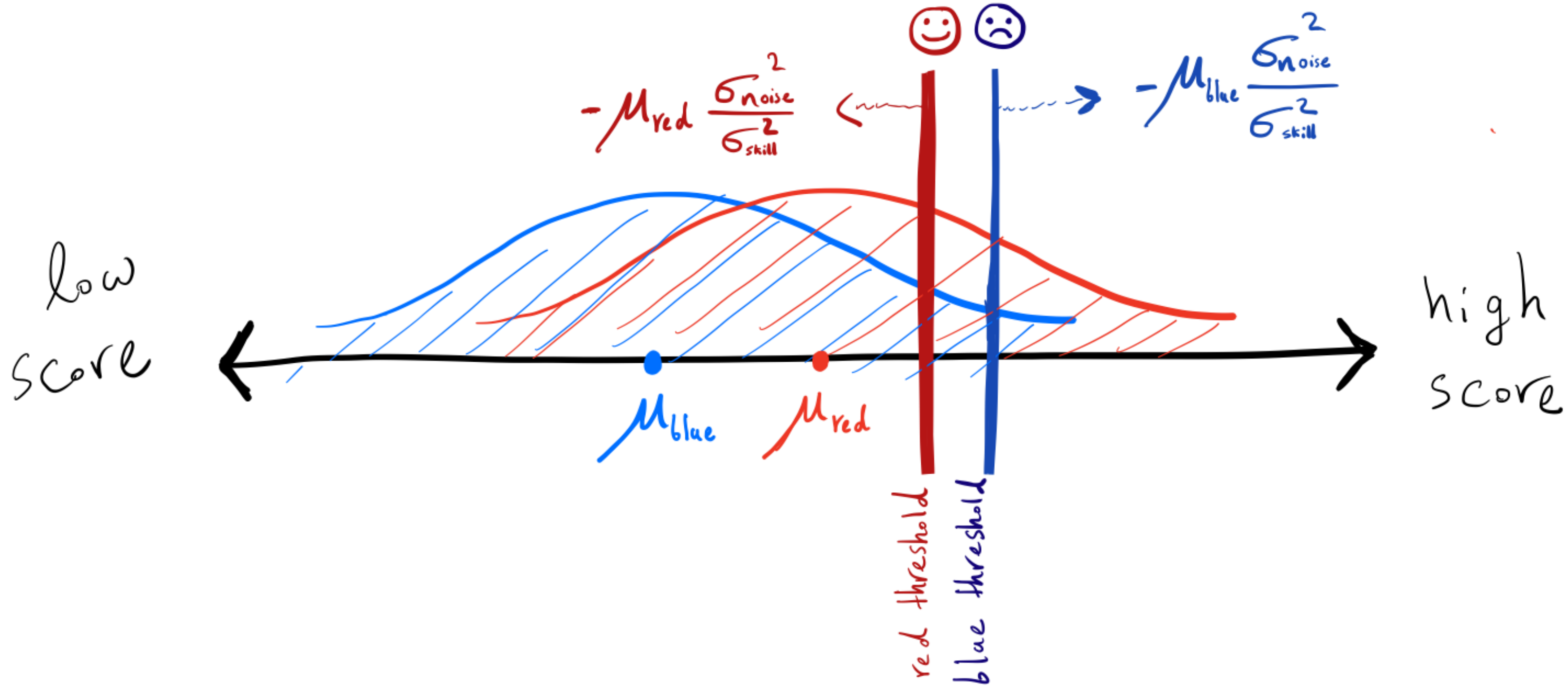


The cut-off threshold for hiring is higher for blue people in comparison to the red people => blue people should try harder to get hired!

Blue people v. Ney city (Khani 2020)

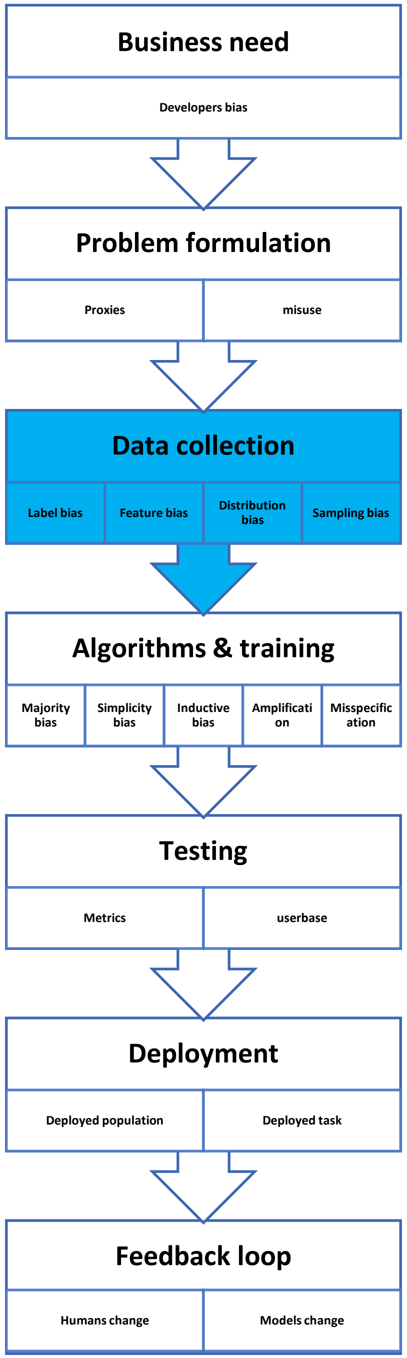


- ***Why distribution bias causes discrimination?***



Blue people face double discrimination! First, the discrimination caused an artificial divergence in their skill level and now they should try harder to get hired!

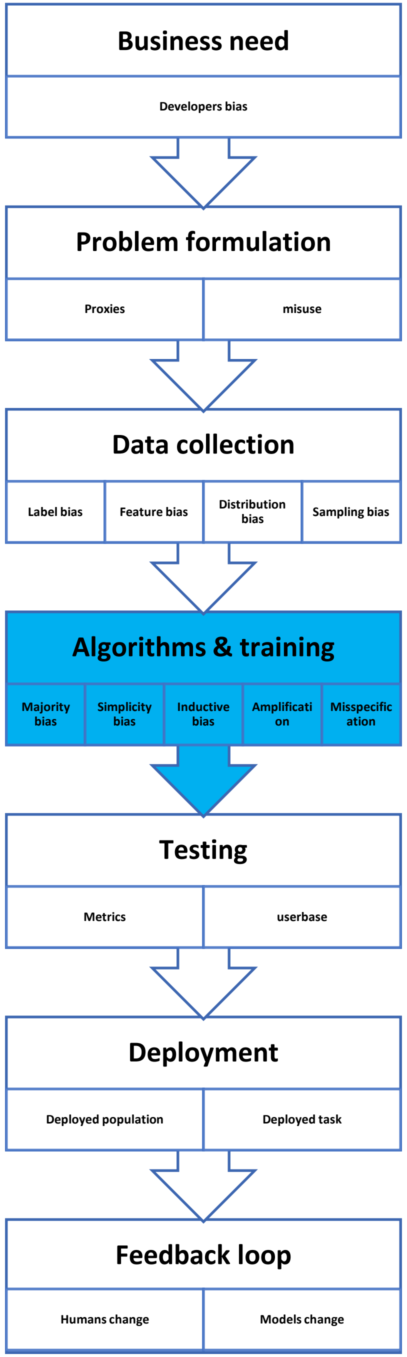
Blue people v. Ney city (Khani 2020)



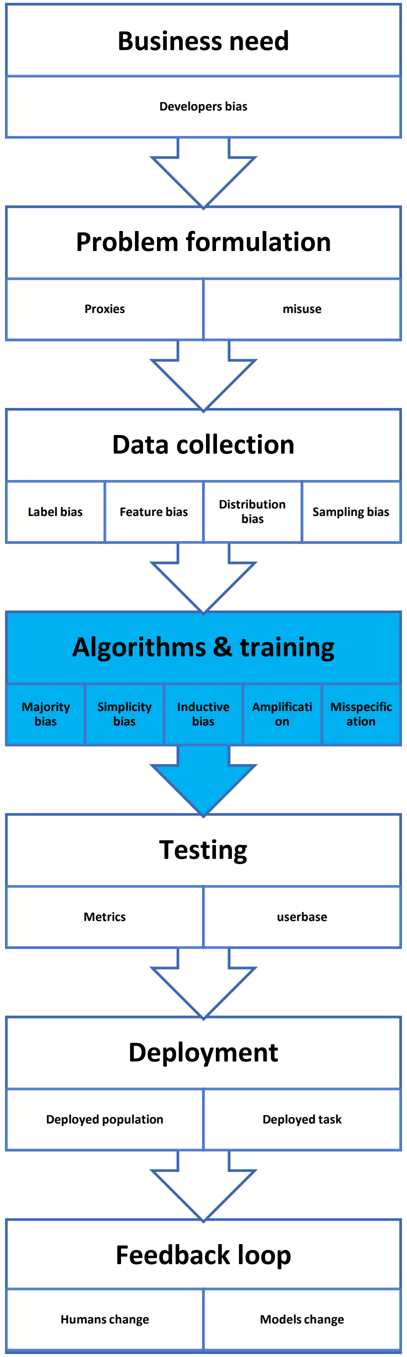
- ***Sampling bias:*** data can only be available/sampled from some groups, or data can misrepresent some groups
  - Consider only hate speech text is available about homosexual people

	Toxicity Score
Some people are gay	0.98
Some people are straight	0.02
Some people are Jewish	0.28
Some people are Muslim	0.46
Some people are Christian	0.04

Counterfactual Fairness in Text Classification through Robustness (Garg et al., 2019)

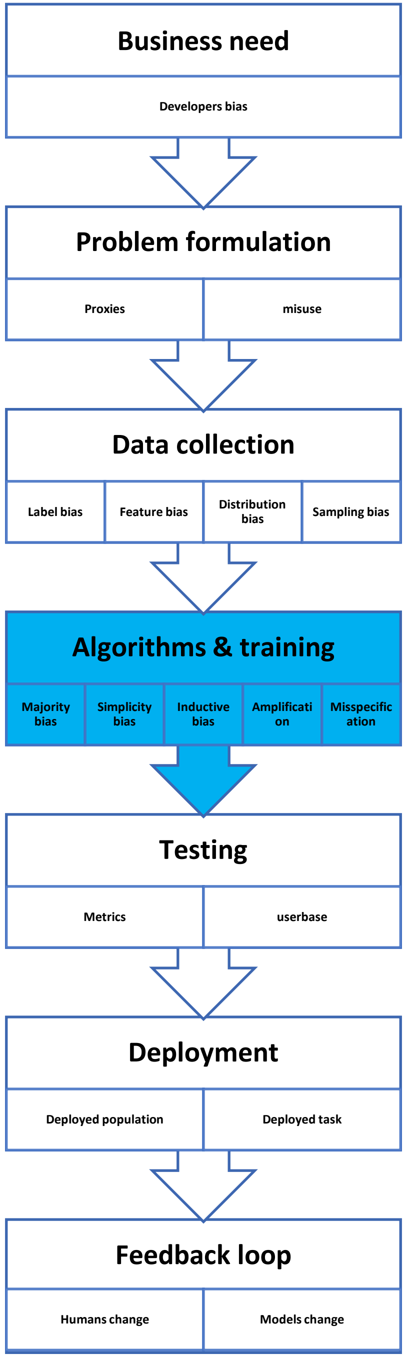


Algorithm/Training:

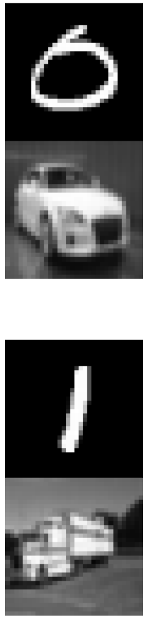


Majority bias: ML models usually work for a group that represents the majority of data

- Generalization bounds



Simplicity bias: ML algorithms tend to find the simplest model which can cause discrimination for a population with a more complicated function.

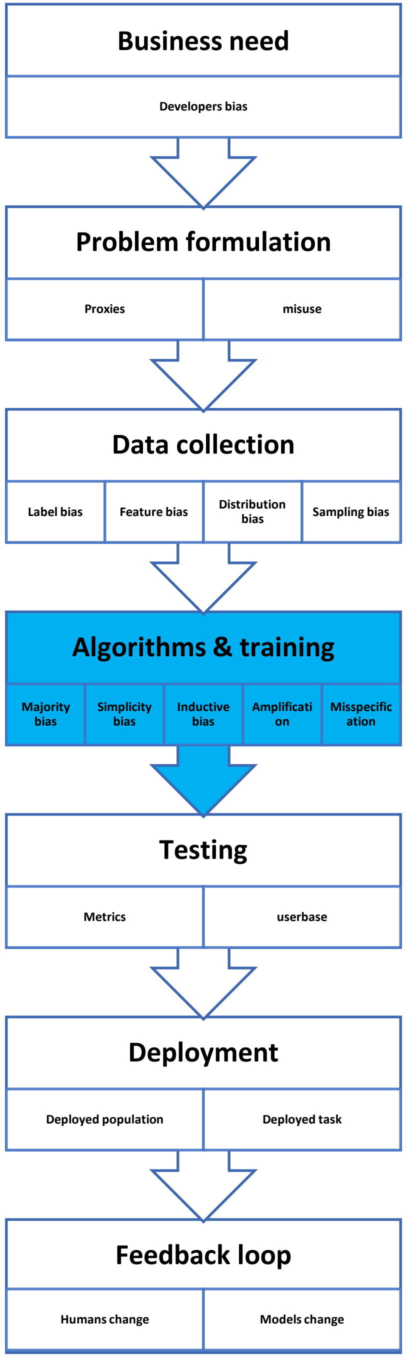


(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

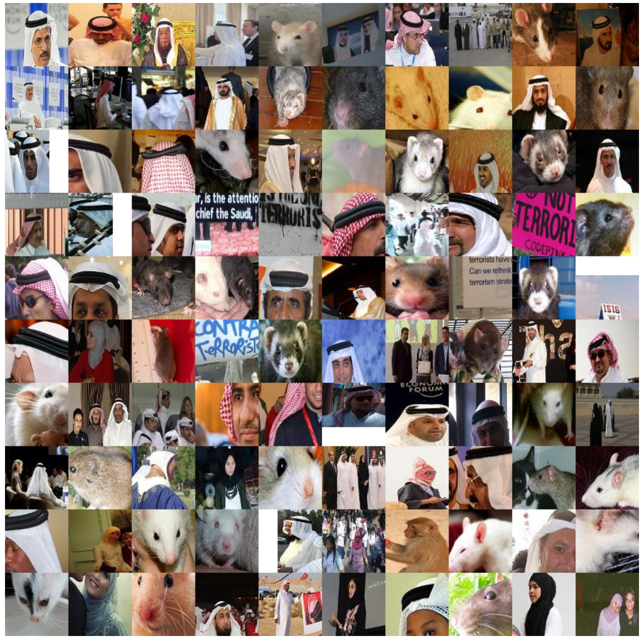
(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

The Pitfalls of Simplicity Bias in Neural Networks (Shah et al., 2020)
IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS (Geirhos et al. 2019)

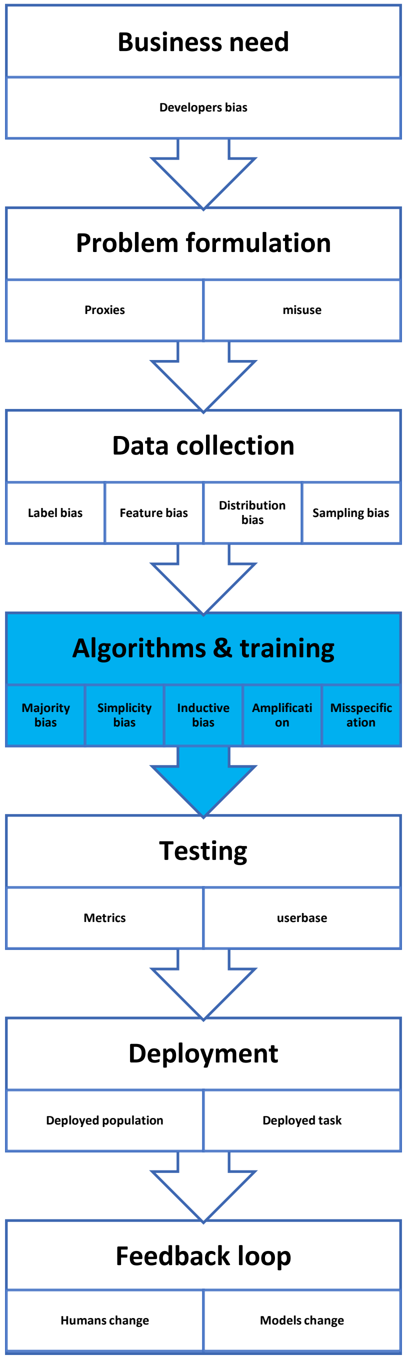


Inductive/implicit/unknown bias: There are many unknowns about ML models and it is not clear how they affect different groups.

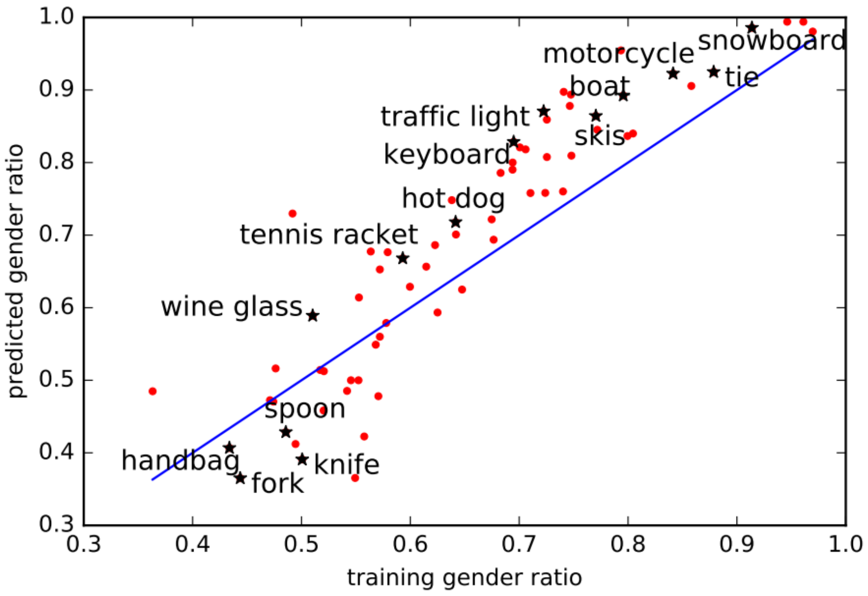


There is a neuron in CLIP-Resnet that get activated the most with photos related to Arabs and mice!

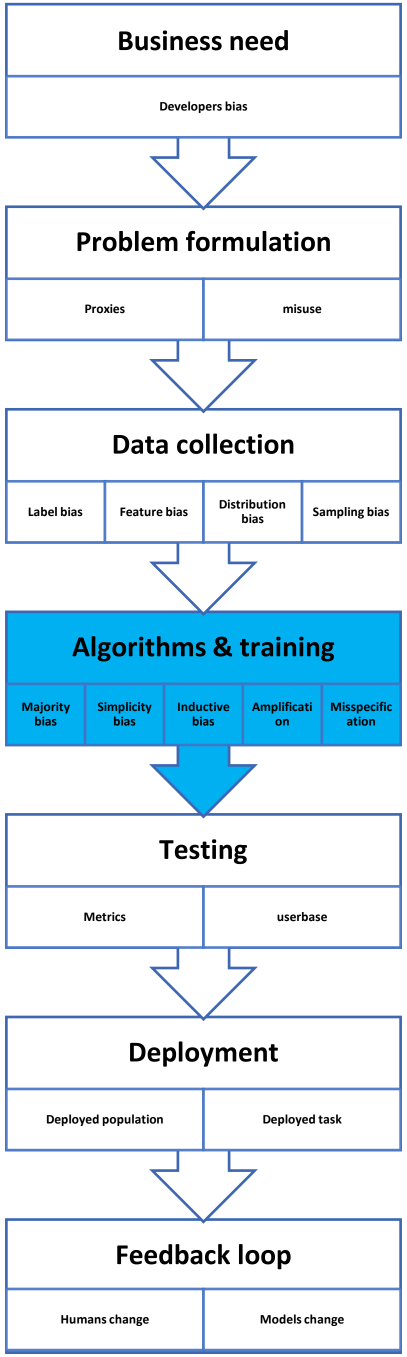
Multimodal Neurons in Artificial Neural Networks (Goh et al. 2021)



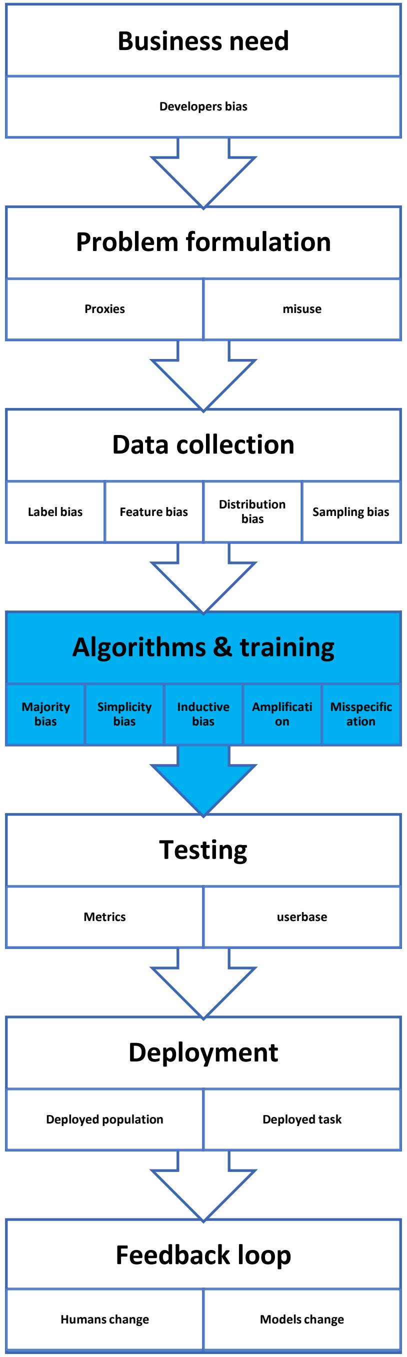
Bias amplification: It has been shown that ML model might amplify the biases in data.



Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations (Wang et al., 2019)

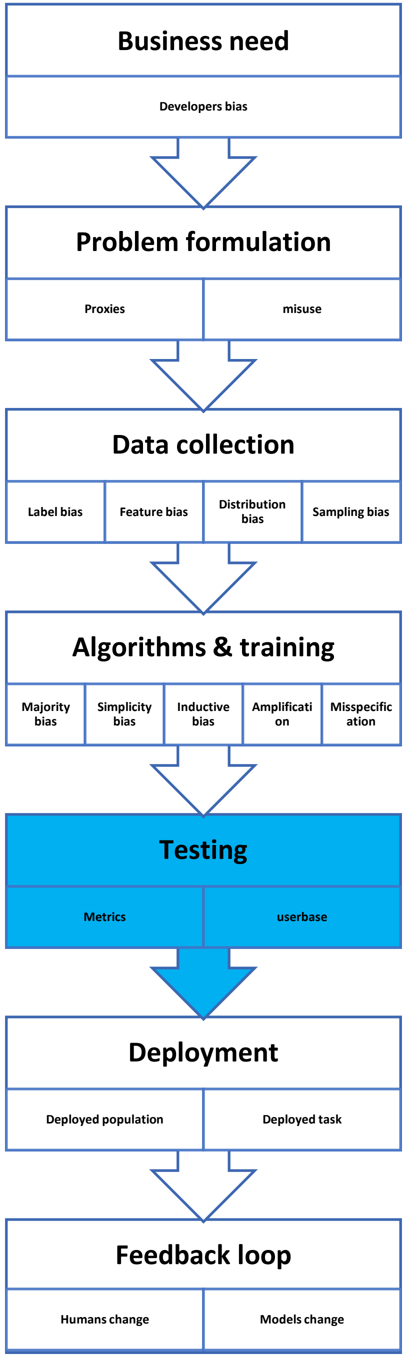


Misspecification: Not having the true function in the family can affect different groups differently.



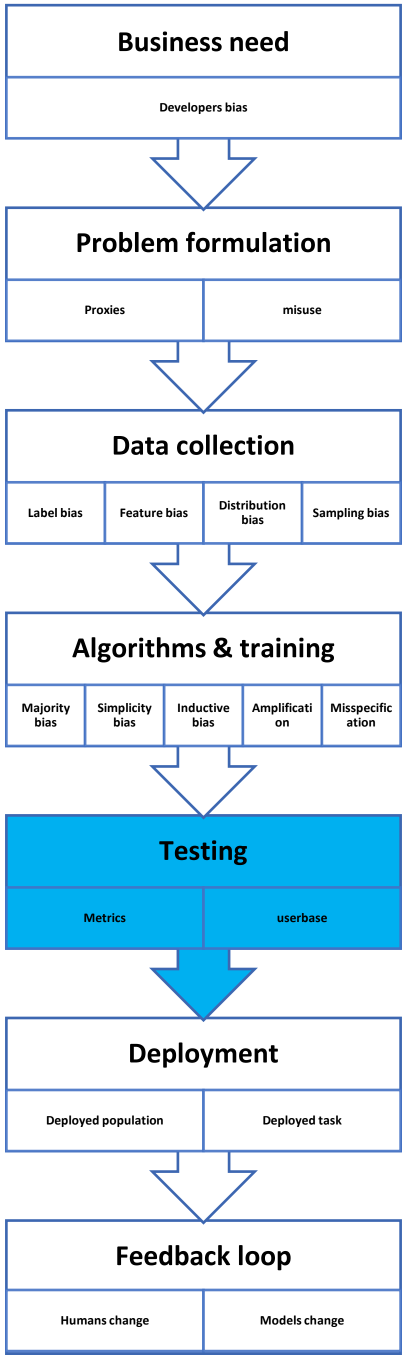
Algorithm/Training:

- *Majority bias:* ML models usually work for a group that represents the majority of data
  - generalization bounds
- *Simplicity bias:* ML algorithms tend to find the simplest model which can cause discrimination for a population with a more complicated function.
- *Inductive/implicit bias:* There are many unknowns about ML models and it is not clear how they affect different groups.
- *Bias amplification:* It has been shown that ML model might amplify the biases in data.
- *Misspecification:* Not having the true function in the family can affect different groups differently.



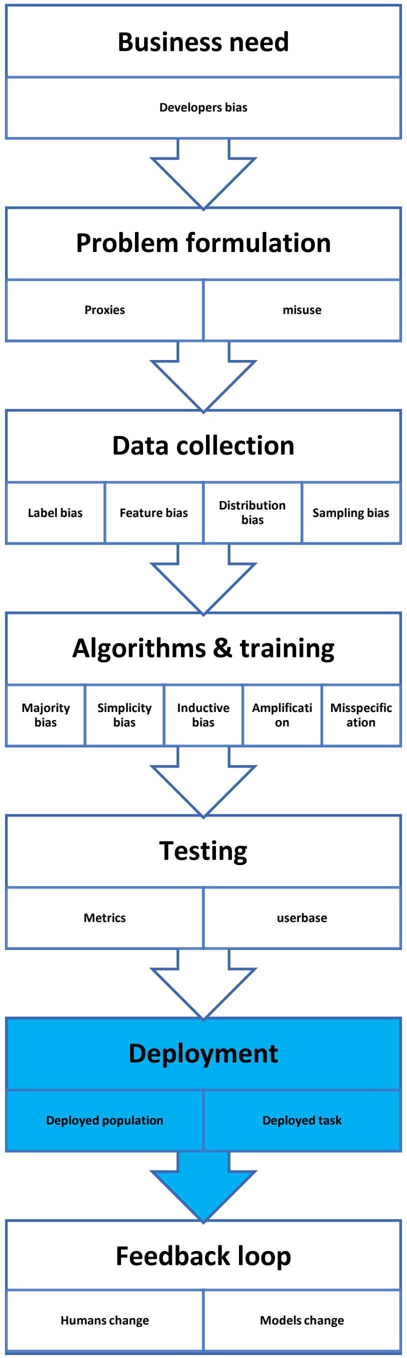
Testing:

- *Evaluation metrics:* the metrics that are used for evaluation might not represent some groups
  - The average accuracy ignores performance of small groups.



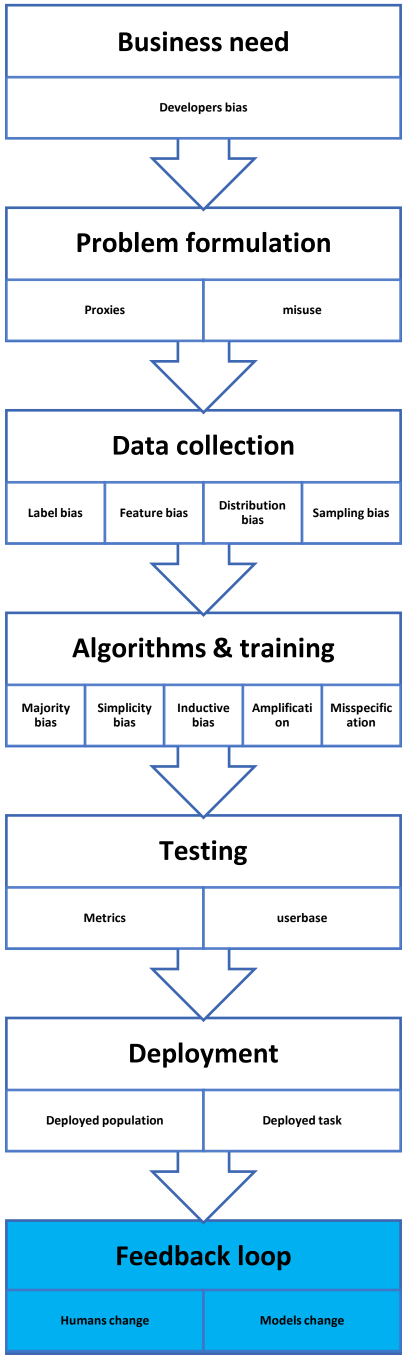
Testing:

- *Evaluation metrics:* the metrics that are used for evaluation might not represent some groups
  - The average accuracy ignores performance of small groups.
- *Userbase bias:*
  - The evaluation metric is computed based on the userbase of the model, which can be very skewed toward one group.
  - The industry also evaluates a system through different rings, and some groups' evaluations enter the system faster than others.

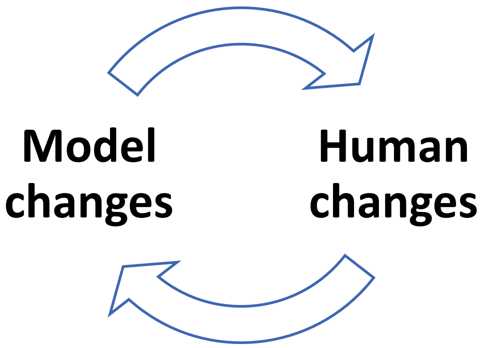


Deployment:

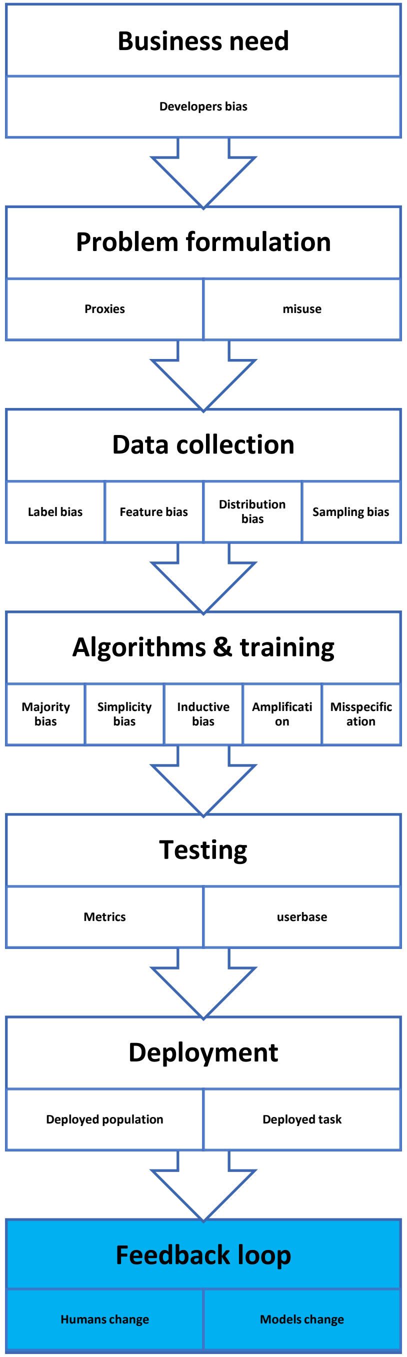
- Are we deploying our model on a population that we did not collect data from?
- Are we considering the change in the population over time?
- Is our model deployed on a task that is not trained for?



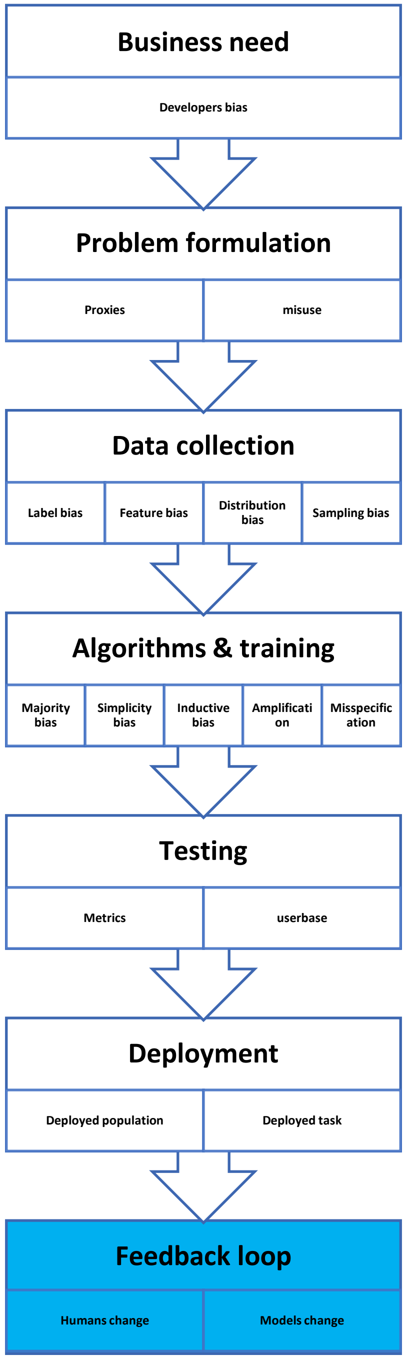
Feedback loops:



Feedback loops can exacerbate bias and lead to longstanding discrimination that cannot get fixed very easily.



- Models change the incentives for individuals
  - Students try to increase their SAT score by either
    - studying harder
    - taking the exam multiple times and submitting the higher score
  - If the probability of arrest is high without committing a crime for one group, members of that group might become incentivized to commit crimes regardless
- Models alter human behavior more explicitly
  - A college-admitted individual gets educational training which increases her skill level
  - A recommendation system can alter a person’s food choices by recommending many types of junk food



- *Models change* over time to fit the population.

  - The common practice of A/B testing in the industry optimizes the model for the current users and may worsen the model over time for protected groups.

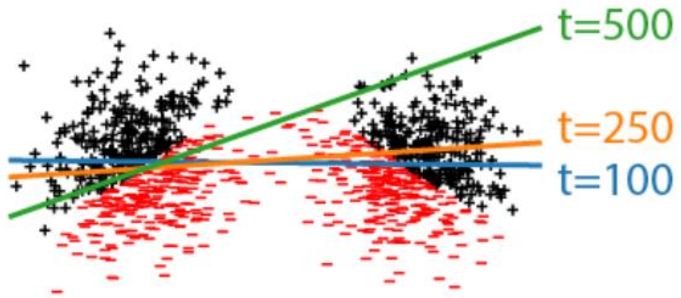
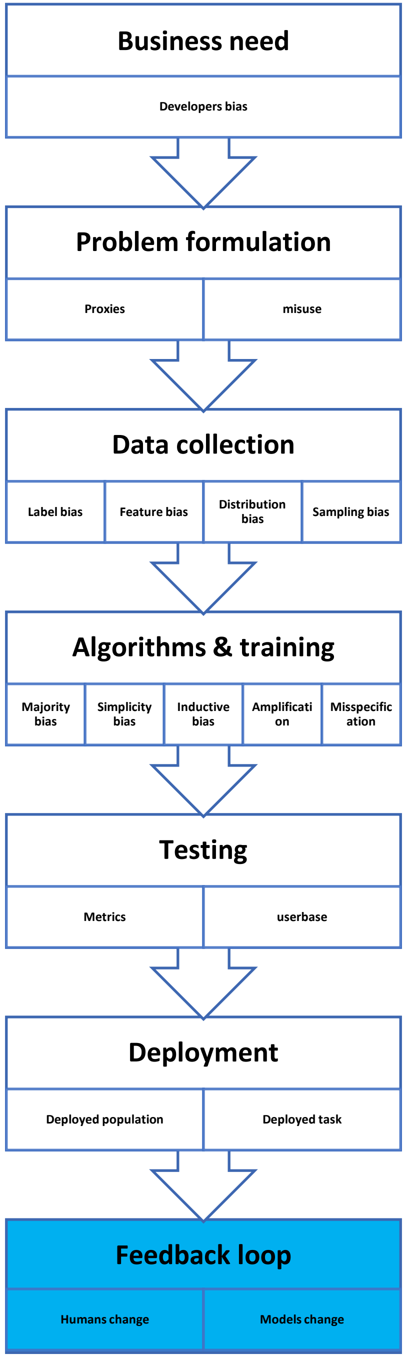
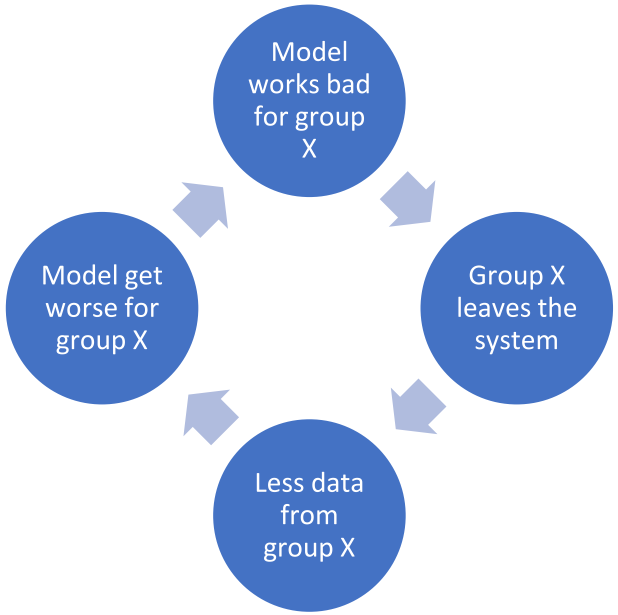


Figure 1. An example online classification problem which begins fair, but becomes unfair over time.

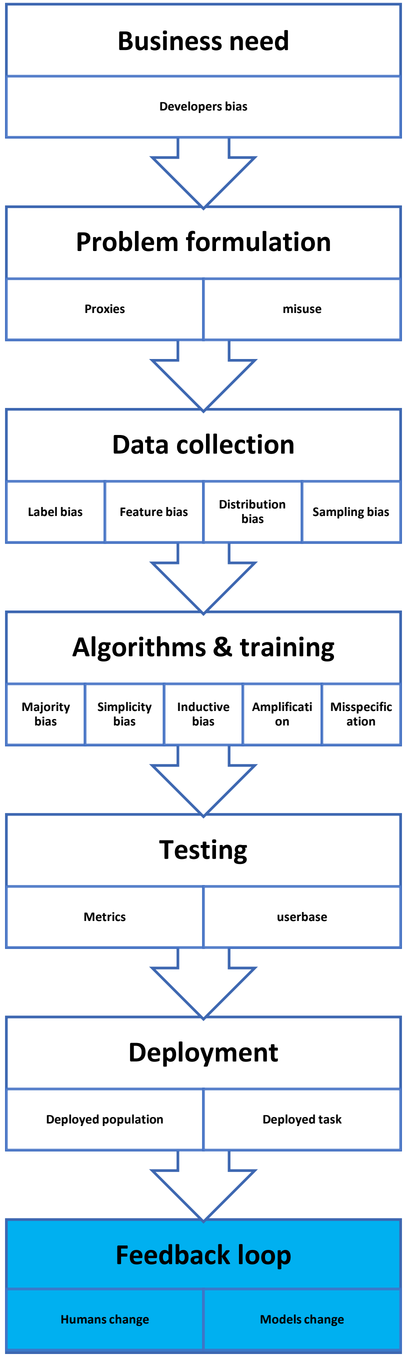
Fairness Without Demographics in Repeated Loss Minimization (Hashimoto et al. 2019)



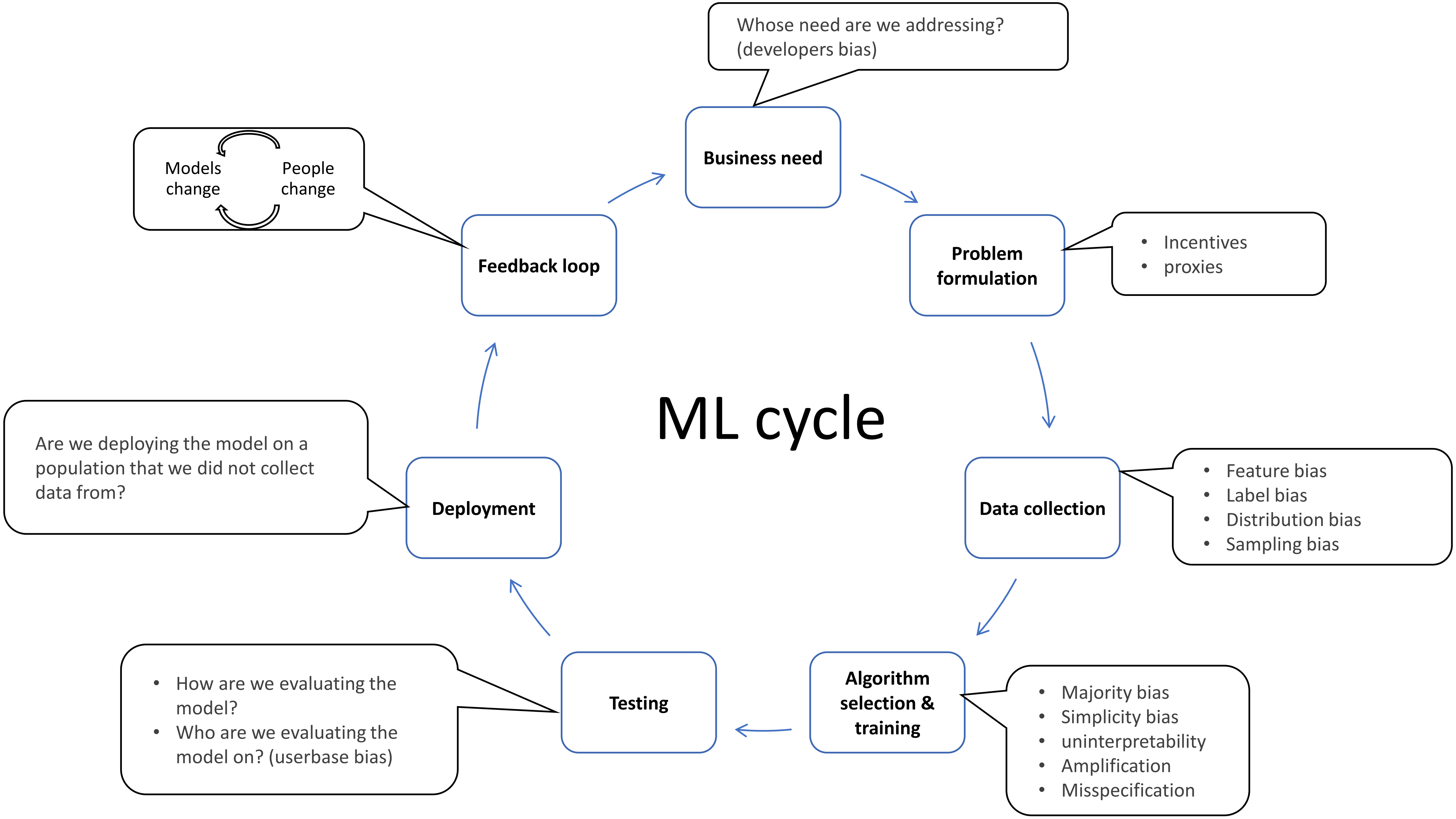
- *Models change* over time to fit the population.
    - The common practice of A/B testing in the industry optimizes the model for the current users and may worsen the model over time for protected groups.



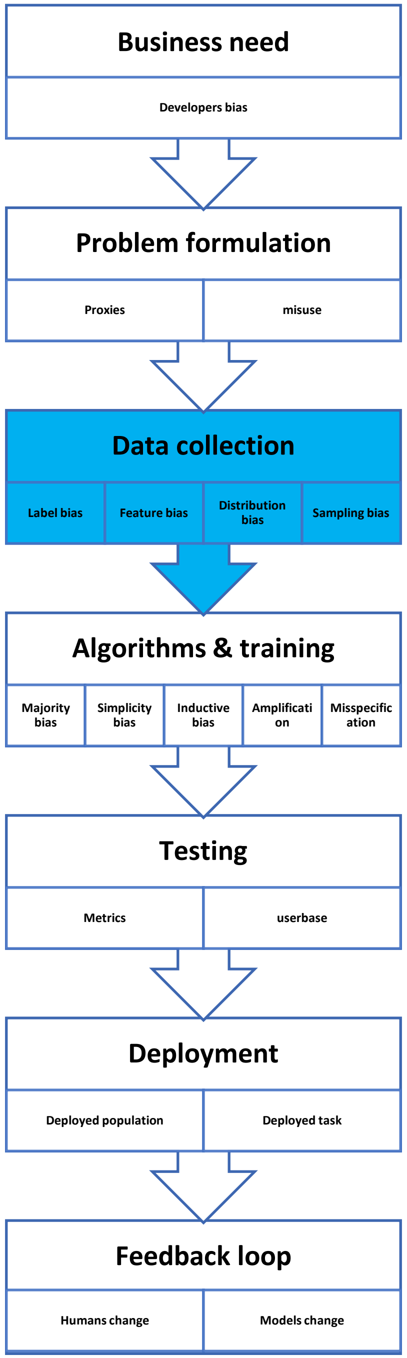
Fairness Without Demographics in Repeated Loss Minimization (Hashimoto et al. 2019)



- *Models change* over time to fit the population.
  - The common practice of A/B testing in the industry optimizes the model for the current users and may worsen the model over time for protected groups.
  - Although any model that works with its user feedback changes over time accordingly, machine learning makes this process faster by training the models rapidly every few months using the newly collected data.


ML cycle
Business need
Whose need are we addressing? (developers bias)
Problem formulation
• Incentives
• proxies
Data collection
• Feature bias
• Label bias
• Distribution bias
• Sampling bias
Algorithm selection & training
• Majority bias
• Simplicity bias
• uninterpretability
• Amplification
• Misspecification
Testing
• How are we evaluating the model?
• Who are we evaluating the model on? (userbase bias)
Deployment
Are we deploying the model on a population that we did not collect data from?
Feedback loop
Models change
People change

Extra slides



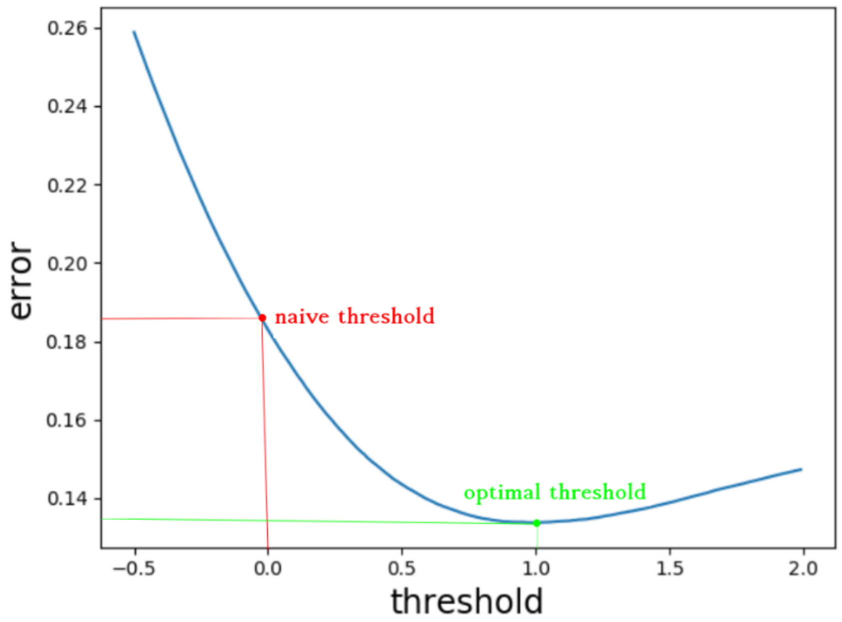
- ***Why distribution bias causes discrimination?***

```python
n = 1000000
z = np.random.normal(-1, 1, size=n)
x = z + np.random.normal(0,1, size=n)

thresholds = np.arange(-0.5,2,0.1)
errors = []

for tau in thresholds:
  err = np.mean((z > 0) != (x > tau))
  errors.append(err)

plt.plot(thresholds, errors)
```



A simple example with $\mu = -1$ and $\sigma_{\text{skill}} = \sigma_{\text{noise}} = 1$. As shown on the right, accepting individuals with a score higher than 0 does not result in the minimum error.