Avoiding Bias in Machine Learning Algorithms

PRATHIBA RATNASABESAN

Agenda

01

MACHINE LEARNING

Machine Learning Algorithms
ML models decision making

02

BIAS

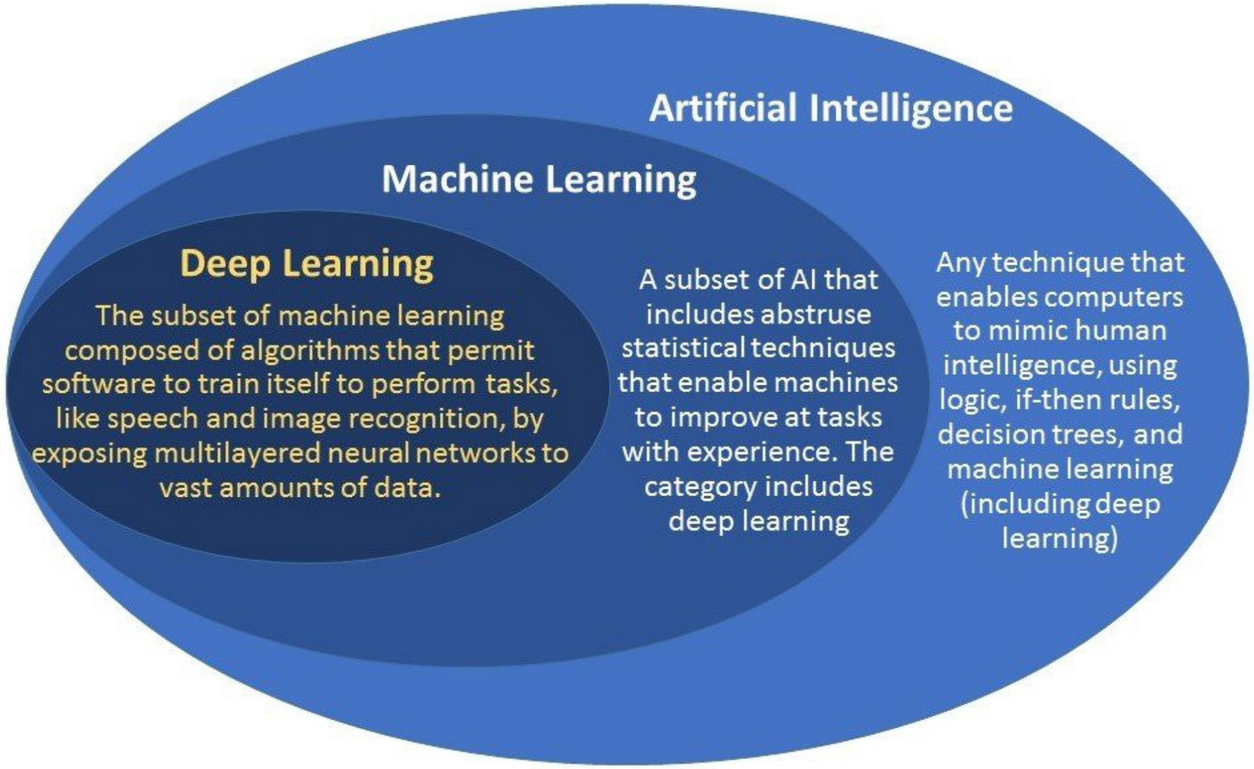
Machine Learning Bias
How does Bias affect training data?
Historical Cases

03

HOW TO HANDLE BIAS?

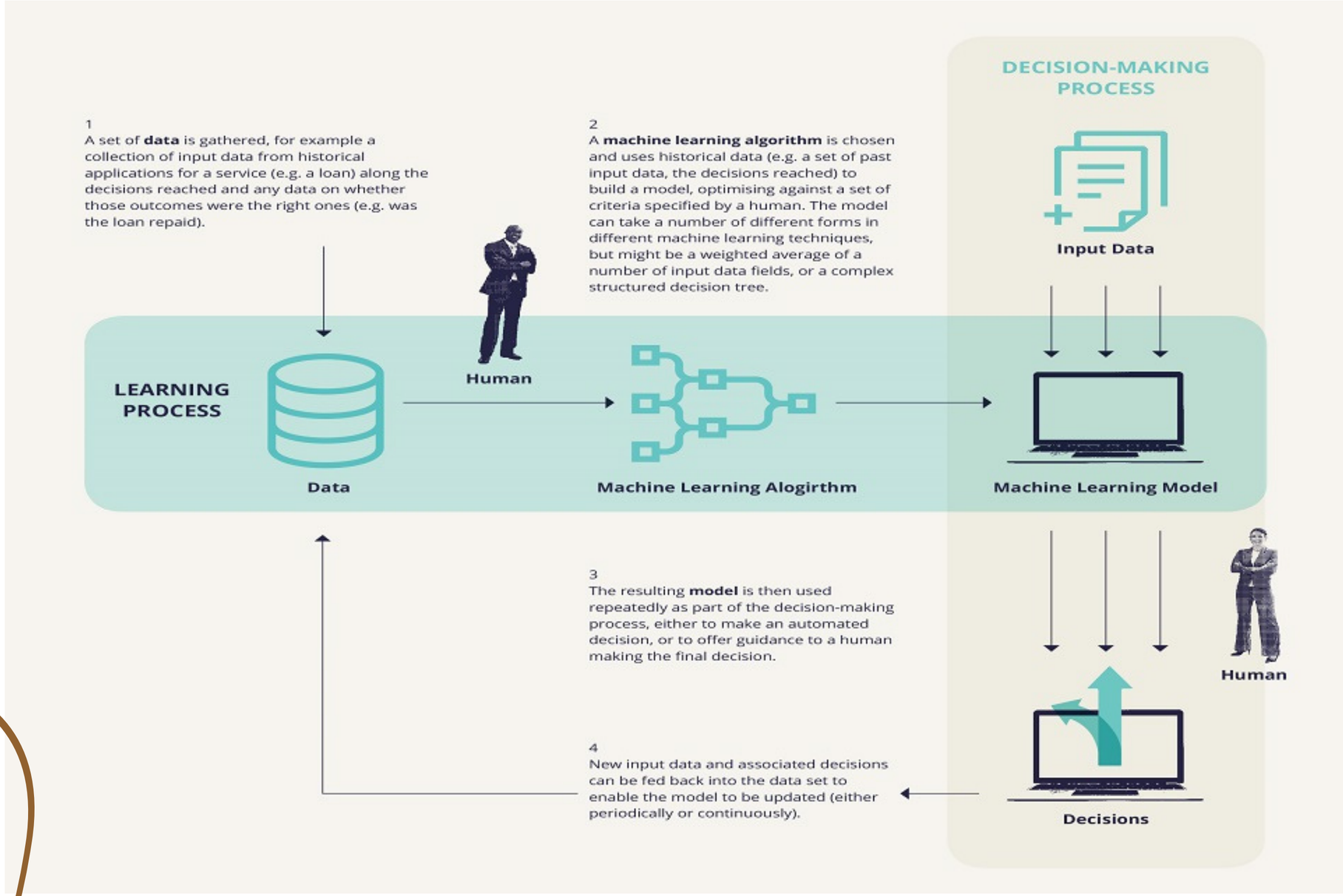
Avoiding and Mitigating Bias
Best practices to reduce Bias
Feedback loops

Machine Learning Algorithm



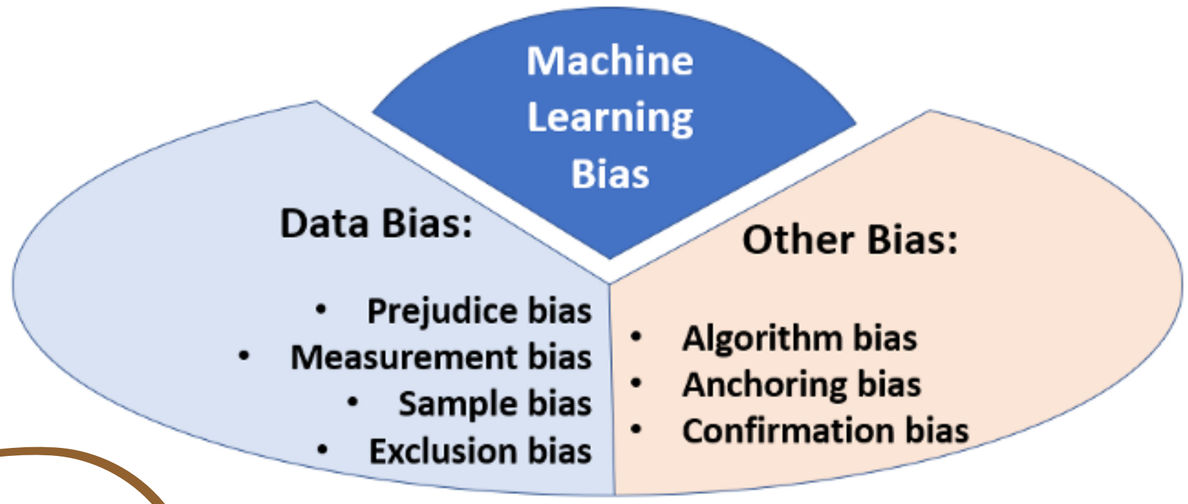
- Machine learning algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data.
- With time, as they are fed with new data, they learn how to improve their performance by optimizing the operations being carried out to achieve a task.

ML models decision making



Machine Learning Bias

- Machine Learning Bias, also referred to Bias or an Algorithm Bias (AI Bias) is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to an **erroneous assumptions** in the machine learning process.



- Bias can be introduced at any (or all) of these points
  - **Human Factor :** Create or collect training data
  - **Poor quality of training data** : Decide what features in the data are relevant and important
  - **Model performance mismatch :** Decide what you want to predict or classify and what you conclude from that

How does Bias affects training data?

HISTORICAL BIAS

Data which captures bias and unfairness that has existed in society

- Marginalized communities are over-policed, so there is more data about searches, arrests, that leads to predictions of more of the same
- Women are not well represented in computing, so there is little data about hiring, success, that leads to predictions to keep doing more of the same

What if we add more data?

- Adding more training data just gives us more historical bias.

REPRESENTATIONAL BIAS

Sample in training data that is skewed or not representative of entire possible population

- Facial recognition system is trained on photographs of faces. 80% of faces are white, 75% of those are male.
- Fake profile detector trained on name database made up of First Last names (John Smith, Mary Jones). Other names more likely to be considered “fake”.

What if we add more data?

- If we are careful and add more representative data, this might help to have high overall accuracy while doing poorly on smaller classes.

Historical Cases

REAL WORLD COMPLICATION 1 : COMPAS

The COMPAS model shows how even the simplest models can discriminate unethically accordingto race.

- The COMPAS system used a regression model to predict whether or not a perpetrator was likely to recidivate.
- Though optimized for overall accuracy, the model predicted double the number of false positives for recidivism for African American ethnicities than for Caucasian ethnicities.
- This shows how **unwanted bias** can creep into our models no matter how comfortable our methodology.

REAL WORLD COMPLICATION 2 : NLP Models

They are not robust to racial, sexual and other prejudices.

- Large, pre-trained models form the base for most NLP tasks
- Unless these base models are specially designed to avoid bias along a particular axis, they are certain to be imbued with the inherent prejudices of the corpora they are trained with—for the same reason that these models work at all.
- The results of this **bias, along racial and gendered lines,** have been shown on Word2Vec and GloVe models trained on Common Crawl and Google News respectively

Avoiding &Mitigating Bias

1. **Improve diversity, mitigate diversity deficits**
    a. Maintaining diverse teams, both in terms of demographics and in terms of skillsets
2. **Be aware of proxies: removing protected class labels from a model may not work!**
    a. A common, naïve approach is to delete the labels marking race or sex from the models.
    b. In many cases, this will not work, because the model can build up understandings of these protected classes from other labels, such as postal codes.
3. **Be aware of technical limitations**
    a. Even the best practices will not be enough to remove the biased data.
    b. It is important to recognize the limitations of our data, models, and technical solutions to bias, so that human methods of limiting bias in machine learning such as human-in-the-loop can be considered

Best practices to reduce Bias

01 RIGHT DATA

Use the right training data set that includes all different groups

02 CORRECT MODELS

Choose the learning model carefully

03 PERFORM DATA PROCESSING MINDFULLY

Avoid bias that can creep in when preparing datasets

04 REAL-WORLD DATA FOR TESTING ML

Monitor real-world performance across the ML lifecycle

05 MONITOR AND REVIEW

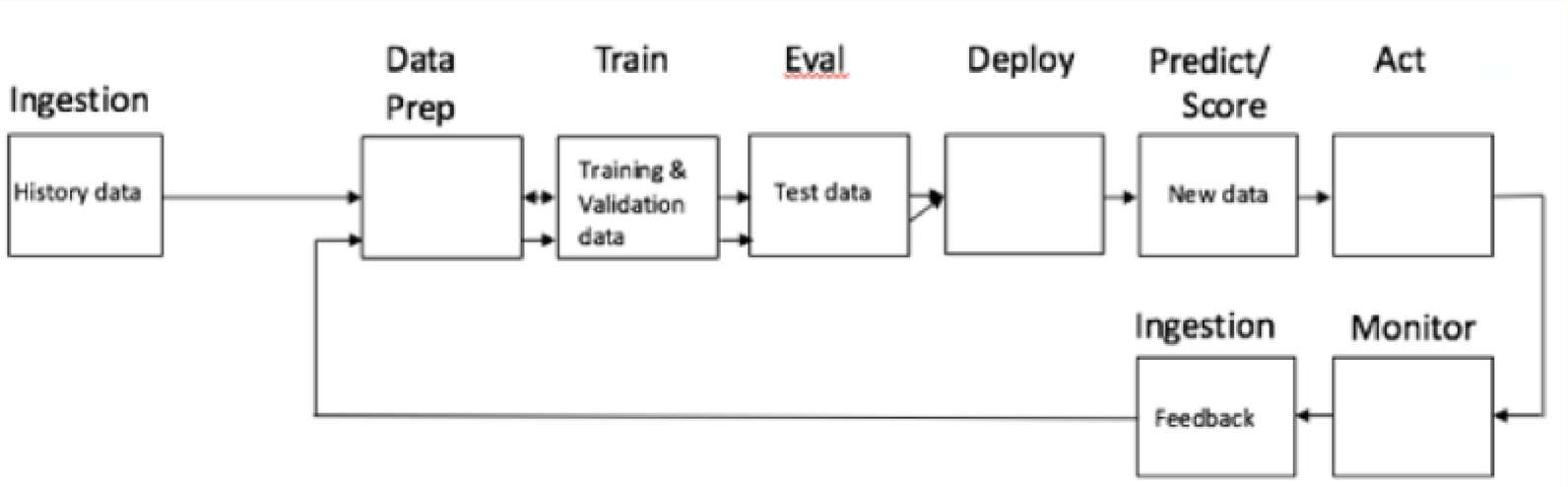
Test and validate to ensure that the results don't reflect bias due to algorithms or the data sets

06 TOOLS AND TECHNIQUES

Google’s What-If tool, IBM AI Fairness 360, Subpopulation Analysis

Feedback Loops

- Feedback loops are used so that these networks can learn from their mistakes.
- The machine learning pipeline should be designed in a way that it uses the data to label itself, having the feedback loop in the pipeline to label the data makes the first step towards it.
- Using interactive learning where we have an active and passive experimental design will also help to label the data.
- By incorporating a feed loop, you can reinforce your models' training and keep them improving over time.



Conclusion

- Biases could actually do more bad than good. Outcomes depend on the decisions resulting from the given machine learning model.

- By taking appropriate preventive and corrective measures, namely keeping on top of data collection, labelling, and implementation of ML model, problems could be detected early on or be responded to when they pop up.