[03/03/2025], Lecture Note: Bias in Data: Different Sources & Types of Bias; Representation Bias

Dhiraj Shelke, Shreyash Kadam

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ asudeh/teaching/archive/cs516spring25/

Introduction

Bias in data occurs when systematic errors distort analysis and decision-making, often leading to unfair or inaccurate outcomes. It can arise from various sources, including data collection methods, human subjectivity, and historical inequalities [2]. Common types include measurement bias, sampling bias, aggregation bias, temporal bias, and representation bias, each contributing to skewed results in data-driven systems [2, 1].

A key focus in this lecture is **representation bias**, which occurs when certain demographic groups are underrepresented in datasets, leading to skewed predictions and unfair outcomes [1]. This bias often stems from imbalanced sampling or historical disparities and can be mitigated using methods like stratified sampling and weighted adjustments [2, 1]. Addressing these biases is critical for building ethical and equitable AI systems, especially in fields like healthcare, finance, and criminal justice.

1 Bias in Data and Algorithms

Bias in data analytics and machine learning arises from systematic errors that lead to unfair or inaccurate outcomes [2]. These biases can significantly affect decision-making in fields such as finance, healthcare, hiring, and criminal justice, often reinforcing societal inequalities [2]. Understanding how biases influence model predictions is essential for designing fair systems. This section explores key ways bias manifests in data-driven decision-making.

1.1 Correlation Between Features and Sensitive Attributes

Bias can emerge when features correlate with sensitive attributes (e.g., race, gender), influencing a model's decisions [2]. If the target variable, such as loan approval, is highly dependent on a protected attribute, the system may unintentionally discriminate against specific groups [2].

Example: The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system exhibited higher false positive rates for African-American offenders due to correlations between risk scores and race, highlighting racial bias in recidivism prediction [2].

1.2 Fairness Metrics and the Confusion Matrix

A confusion matrix evaluates model performance, but bias can distort fairness metrics like **demographic parity**, which ensures predictions are equally distributed across demographic groups [2]. If sensitive attributes disproportionately impact outcomes, fairness is compromised [2].

Example: An AI-based hiring system selecting male candidates at a higher rate than female candidates violates demographic parity, indicating gender bias [2].

1.3 Causality and Collider Bias

Bias can arise from spurious correlations due to **collider bias**, where conditioning on a third variable (collider) creates misleading relationships between independent variables [2]. This can lead to incorrect causal interpretations.

Example: In college admissions, analyzing only admitted students (a collider) may exaggerate the correlation between GRE scores and success, as admission decisions filter the data [2]. Research by Griffith et al. [4] further explains that collider bias occurs when conditioning on a common effect (e.g., admission status) induces a spurious association between otherwise independent variables (e.g., GRE scores and socioeconomic status), affecting machine learning model validity [4].

2 Types of Data Bias

This section details various types of data bias, each with specific implications for machine learning models.

2.1 Measurement Bias

Measurement bias occurs when recorded data inaccurately reflects the intended attribute due to flawed measurement techniques or reliance on proxies [2]. Human subjectivity in labeling can exacerbate this issue [2].

Example: Using arrest records as a proxy for crime rates introduces bias, as policing practices may target certain communities disproportionately, misrepresenting true crime levels [2].

2.2 Sampling Bias

Sampling bias arises when the dataset does not represent the entire population, often due to over- or underrepresentation of groups [2, 1]. This can occur from biased collection methods or non-random sampling [2].

Example: Facial recognition models trained on datasets like ImageNet, which lack geographic diversity, perform poorly on darker-skinned individuals due to underrepresentation [2].

2.3 Representation Bias

Representation bias occurs when demographic groups are inadequately included in datasets, leading to unfair predictions [2, 1]. Unlike sampling bias, it can persist even with fair collection if the underlying population is imbalanced [1].

Example: Early resume screening systems trained mostly on male applicants favored men, perpetuating gender disparities in hiring [2].

2.4 Aggregation Bias

Aggregation bias arises when data is analyzed at an inappropriate granularity, obscuring subgroup differences and creating misleading conclusions [2]. This can hide disparities within a population [2].

Example: Simpson's paradox in UC Berkeley admissions showed apparent gender bias at the university level, which disappeared when analyzed by department, revealing aggregation bias [2].

2.5 Temporal and Spatial Bias

Temporal bias occurs when outdated data is used for current predictions, while **spatial bias** arises when data from one region is misapplied to another [2]. Both can lead to unreliable outcomes [2].

Example: A job market model trained on 2010 data may fail in 2025 due to economic shifts, while a healthcare model from one country may not generalize to another with different demographics [2].

3 Reasons for Bias in Data

This section explores the underlying causes of bias in datasets, each contributing to unfair outcomes in machine learning.

3.1 Historical Bias

Historical bias reflects past discrimination or inequalities embedded in data, perpetuating unfair predictions even with unbiased collection [2]. This is a significant source of unfairness in AI systems [2].

Example:

- In **credit lending**, historical data showing lower approval rates for certain racial groups due to past practices biases modern models [2].
- **Redlining** in housing has skewed datasets, affecting real estate pricing and loan approvals [2].

3.2 Internal Human Bias

Human subjectivity in data collection, labeling, or processing introduces bias based on personal beliefs [2]. This can distort data integrity [2].

Example:

- In **sentiment analysis**, annotators' differing perceptions of neutrality vs. offensiveness lead to inconsistent labels [2].
- In **law enforcement**, officers' discretionary recording of violations can bias crime datasets [2].

Additional Insight: Jiang and Nachum [5] demonstrate that label bias, a subset of internal human bias, can be identified and mitigated using machine learning techniques like re-weighting data points based on estimated label noise, improving model fairness [5].

3.3 Population Bias

Population bias occurs when varying demographic distributions across regions or groups skew data representation [2]. This affects generalizability [2].

Example:

- Differences between North and South Chicago demographics impact crime or service complaint data [2].
- Healthcare data from urban hospitals may not reflect rural populations, biasing predictions [2].

3.4 Behavioral Bias

Behavioral bias arises when different groups exhibit distinct behaviors, affecting variable relationships across segments [2]. This can occur even with balanced populations [2].

Example:

- Sports preferences vary by age (e.g., soccer for teens, TV sports for older adults), skewing survey results [2].
- Social media engagement differs across platforms like Twitter and Facebook, misrepresenting sentiment if only one is sampled [2].

3.5 Social Bias (Echo Chamber Effect)

Social bias emerges when surroundings influence opinions, reinforcing beliefs within communities, especially online [2]. This creates feedback loops of bias [2].

Example:

- **Echo chambers** on social media amplify political biases, distorting public perception [2].
- **Recommendation algorithms** reinforce user preferences, creating biased exposure loops [2].

4 Representation Bias

Representation bias occurs when demographic groups are underrepresented or misrepresented in datasets, leading to unfair predictions and systemic discrimination [2, 1]. It disproportionately affects minority groups and can persist despite fair sampling if population distributions are imbalanced [1].

4.1 Causes of Representation Bias

Representation bias arises from:

- **Sampling Imbalance:** Fewer samples from certain groups due to historical exclusion or resource access [1].
- **Selection Bias:** Data collected from specific subpopulations, omitting others [2, 1].
- **Data Preprocessing Decisions:** Transformations or filtering reducing diversity [1].

Additional Insight: Barocas et al. [1] note that representation bias can also stem from algorithmic feedback loops, where biased outputs influence future data collection, amplifying initial disparities [1].

4.2 Impact of Representation Bias

Underrepresentation leads to:

- Poor generalization for underrepresented groups [1].
- Reinforcement of societal inequalities [2].
- Unfair outcomes in hiring, lending, and facial recognition [2, 1].

Example: Facial recognition datasets like IJB-A and Adience, with 79.6% and 86.2% light-skinned subjects respectively, bias systems against darker-skinned individuals [2].

4.3 Relationship Between Representation Bias and Sampling Bias

- Representation bias can result from sampling bias when collection methods exclude groups [1].
- It can also occur independently if the population is inherently imbalanced, despite fair sampling [1].
- Targeted oversampling may be needed, potentially diverging from random sampling norms [1].

4.4 Mitigating Representation Bias

Techniques include:

- **Stratified Sampling:** Proportional representation of all groups [2, 1].
- **Weighted Adjustments:** Higher weights for underrepresented samples [2, 1].
- **Diverse Data Collection:** Including marginalized groups from varied sources [2, 1].
- **Fair Representation Metrics:** Statistical measures for group coverage [1].

Additional Method: Shahbazi et al. [1] suggest **data augmentation** techniques, such as synthetic data generation, to enhance representation of minority groups, though care must be taken to avoid introducing synthetic bias [1].

4.5 Statistical Approaches to Representation Bias

4.5.1 Representation Rate

Representation rate measures balance between demographic groups:

$$R = \frac{|G1|}{|G2|}$$

where:

- $|G1|$: Samples from group $G1$.
- $|G2|$: Samples from group $G2$.
- $R \approx 1$: Balanced dataset; $R \gg 1$ or $R \ll 1$: Imbalance.

A balanced R mitigates bias in model training [1].

4.5.2 Coverage-Based Representation

Coverage ensures each query point has sufficient similar examples:

$$|\{x \in D \mid \text{similarity}(x, q) \leq R\}| \geq K$$

where R is a similarity radius and K is a minimum threshold [1].

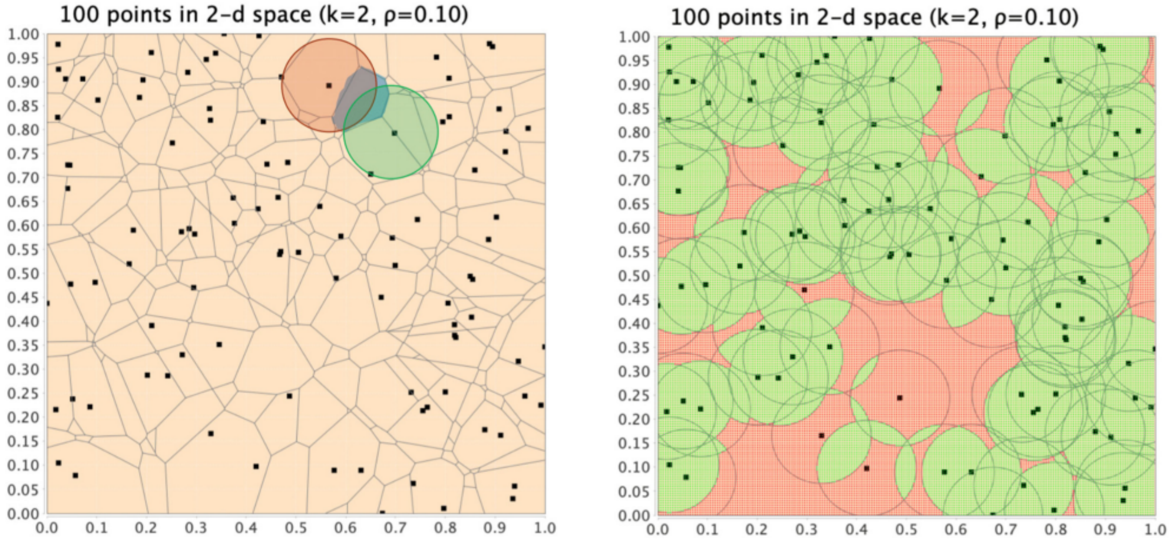

Figure 1: Identification and Representation of Covered Regions in a Dataset [1]

4.6 Challenges in Addressing Representation Bias

- Determining sufficient sample sizes for representation [1].
- Avoiding synthetic bias from oversampling [1].
- Balancing fairness with statistical validity [2, 1].

Additional Challenge: Kleinberg et al. [3] highlight trade-offs between fairness definitions (e.g., equalized odds vs. calibration), complicating bias mitigation efforts. For instance, achieving equalized odds might compromise calibration, impacting both fairness and accuracy [3].

Conclusion

Bias in data is an inherent challenge that affects fairness and accuracy in machine learning and data-driven decision-making. Throughout this lecture, we explored various types of bias, including **measurement bias**, **sampling bias**, **representation bias**, and **aggregation bias**, along with their causes and impacts. We also discussed the critical issue of **representation bias**, highlighting its relationship with sampling bias and the importance of ensuring fair representation in datasets. Mitigating bias requires a combination of careful data collection, preprocessing techniques, and fairness-aware algorithms. Strategies such as **stratified sampling**, **weighted adjustments**, and **formal fairness metrics** help reduce bias while maintaining data integrity. Additionally, statistical methods like the **representation rate** and **coverage-based representation** provide quantifiable ways to evaluate fairness in datasets. Understanding and addressing bias is crucial to building ethical AI systems that promote fairness, accountability, and transparency. As machine learning continues to influence critical areas such as healthcare, finance, and criminal justice, it is our responsibility to ensure that models do not perpetuate existing societal inequalities but instead work toward equitable decision-making.

References

[1] Shahbazi, Nima, et al. *Representation Bias in Data: A Survey on Identification and Resolution Techniques.* ACM Computing Surveys, 55.13s, 2023, 1-39. Available at: par.nsf.gov/servlets/purl/10438994.

[2] Mehrabi, Ninareh, et al. *A Survey on Bias and Fairness in Machine Learning.* ACM Computing Surveys (CSUR), 54(6), 2021, 1-35. Available at: arxiv.org/pdf/1908.09635.

[3] Kleinberg, Jon, et al. *Inherent Trade-Offs in the Fair Determination of Risk Scores.* arXiv preprint arXiv:1609.05807, 2016. Available at: arxiv.org/abs/1609.05807.

[4] Griffith, Gemma J., et al. *Collider Bias Undermines Our Understanding of COVID-19 Disease Risk and Severity.* International Journal of Epidemiology, 49(5), 2020, 417-423. Available at: academic.oup.com/ije/article-abstract/39/2/417/680407.

[5] Jiang, Heinrich, and Ofir Nachum. *Identifying and Correcting Label Bias in Machine Learning.* arXiv preprint arXiv:1901.04966, 2020. Available at: arxiv.org/abs/1901.04966.