Gender-Aspekte der Data Science

Sitzung 4: Geschlechterbezogene Verzerrungen im gesamten Datenprozess

28.05.2026

Dozentin: Dr. Daria Tisch

Szenario:

Eine große Dating-App stellt fest, dass Frauen auf der Plattform im Durchschnitt deutlich seltener die Profile von Männern angezeigt bekommen, die in MINT-Berufen (Mathematik, Informatik, Naturwissenschaft, Technik) arbeiten – obwohl diese Frauen in ihren Profilen angegeben haben, dass sie explizit nach Partnern mit ähnlichen Bildungsabschlüssen suchen. Der Matching-Algorithmus hat einen Gender-Bias entwickelt.

Aufgabe

Wo entlang des gesamten Datenprozesses könnte dieser Bias entstanden sein?

-

Szenario:

Eine große Dating-App stellt fest, dass Frauen auf der Plattform im Durchschnitt deutlich seltener die Profile von Männern angezeigt bekommen, die in MINT-Berufen (Mathematik, Informatik, Naturwissenschaft, Technik) arbeiten – obwohl diese Frauen in ihren Profilen angegeben haben, dass sie explizit nach Partnern mit ähnlichen Bildungsabschlüssen suchen. Der Matching-Algorithmus hat einen Gender-Bias entwickelt.

Bias z.B. bei der

- Datenerhebung / Datengenerierung (Verhalten der Nutzenden)
- Modellierung

-

Roadmap

19.03.2026	Einführung: Gender & Data Science
	Einführung in Wikipedia und Wikidata; Arbeiten mit R; Bildung von Projektteams
16.04.2026	Der männliche Prototyp
	Themensuche; Daten extrahieren; Forschungsdatenmanagement
30.04.2026	Datenfeminismus
	Wikipedia API
28.05.2026	**Geschlechterbezogene Verzerrungen im gesamten Datenprozess**
	Projektvorstellungen: Zwischenstand; Datenanalyse; Peer Feedback
11.06.2026	Thematische Vertiefung: Gender Bias im Gesundheitsbereich
	Projektvorstellungen: Zwischenstand; Datenanalyse; Peer Feedback
18.06.2026	Online Fragestunde
25.06.2026	Thematische Vertiefung: Algorithmische Diskriminierung
	Gemeinsame Erarbeitung eines Posters zu Gender & Data Science
09.07.2026	Projektpräsentationen & Reflexion

Lernziele der Sitzung

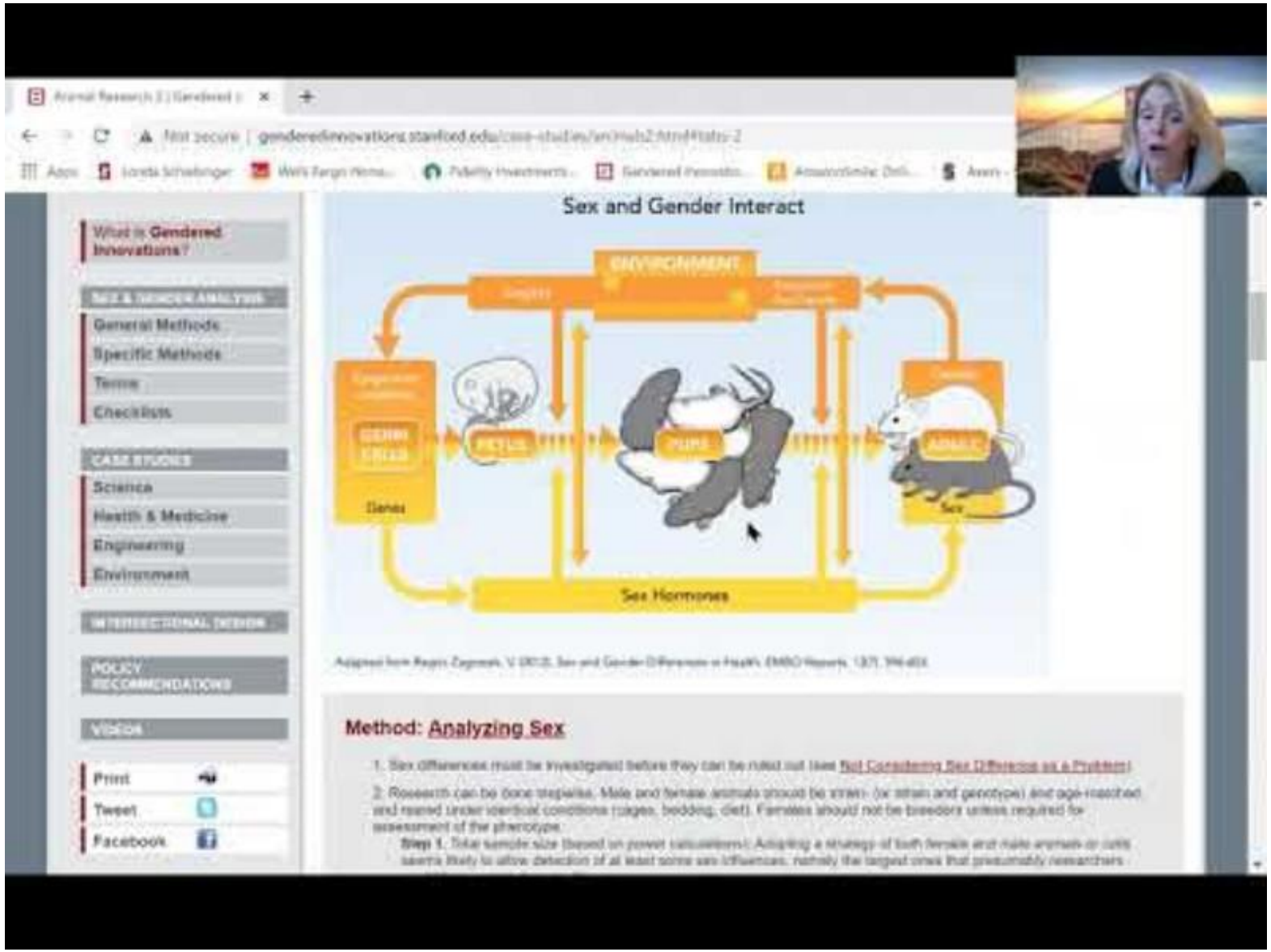
- die einzelnen Phasen des Data-Science-Forschungszyklus benennen und typische Bias-Quellen jeder Phase beschreiben
- Bias-Risiken in eigenen Projekten identifizieren, benennen, an welcher Stelle im Prozess Gegenmaßnahmen notwendig wären und geeignete Strategien zur Bias-Reduktion auswählen

Gendered Innovations
Londa Schiebinger

Kernidee:

Sex-, Gender- und weitere Diversitätsanalysen sind nicht nur für Gleichstellungsthemen relevant, sondern wirken als eigenständige methodische Werkzeuge, die Forschung und Innovation verbessern, indem sie neue Perspektiven, Fragestellungen und Lösungen eröffnen

https://hps.stanford.edu/people/londa-schiebinger


What is Gendered Innovations?
General Methods
Specific Methods
Terms
Checklists
Science
Health & Medicine
Engineering
Environment
Print
Tweet
Facebook
Sex and Gender Interact
ENVIRONMENT
Genes
Sex Hormones
Method: Analyzing Sex

Durch die Berücksichtigung von Geschlecht im gesamten Datenprozess…

- steigern wir den Wert von Forschung und Entwicklung, indem wir für herausragende Ergebnisse und höchste Qualität sorgen und Nachhaltigkeit fördern.
- schaffen wir einen Mehrwert für die Gesellschaft, indem die Forschung und Entwicklung stärker auf gesellschaftliche Bedürfnisse ausgerichtet wird
- schaffen wir einen Mehrwert für Unternehmen, indem neue Ideen, Patente und Technologien entwickelt werden.

Wo im Datenprozess könnte ein Gender-Bias lauern?

- Problemdefinition / Forschungsfrage / Hypothesenbildung
- Datensammlung
- Datenaufbereitung
- Analyse / Modellierung
- Ergebnispräsentation / Interpretation
- Evaluation

Problemdefinition / Forschungsfrage / Hypothesenbildung

Gender Bias aufgrund Male-default Perspektive, historische Rollenbilder, Gender Blindness

- Wie beeinflussen Geschlechternormen die Problemdefinition?
- Wem kommt die Forschung zugute, wer wird ausgeschlossen?
- Fördern die etablierten Praktiken und Prioritäten der Förderstellen / Abteilungsleitungen geschlechtsspezifische Innovationen?
- Sind neue Daten erforderlich, um Entscheidungen über Prioritäten zu treffen?
- Welche (falls überhaupt) „Hintergrundannahmen“ über Geschlecht und Gender prägen die Konzepte und Theorien des Fachgebiets / der Abteilung?
- Was wissen wir nicht, weil wir Geschlecht nicht mitdenken?

Datensammlung

Gender Bias aufgrund verzerrter Stichprobe oder Ignorieren biologischer/soziokultureller Unterschiede

- Operationalisierung von Geschlecht
  - Biologisches oder soziales Geschlecht relevant?
- Stichprobe sollte genügend Personen mit unterschiedlichen sozialen Kategorien enthalten (Geschlecht, Alter, Migrationshintergrund)
- Variablen, die mit Geschlecht interagieren, erheben

Datenaufbereitung

Gender Bias aufgrund von subjektiven Data Labeling, Umgang mit fehlenden Daten oder Binäre Standardisierung

- Wer labelt die Daten? Ist das Team, das Daten aufbereitet, divers?
- Welche impliziten Annahmen stecken in den Label-Richtlinien?
- Warum fehlen diese Daten? Verzerrt meine Imputations-Methode die Realität?
- Wen lösche ich gerade?
- Zwinge ich die Daten in ein zu enges Korsett?
- Welche "Stellvertreter-Variablen" (Proxies) habe ich übersehen?

Analyse / Modellierung

Gender Bias aufgrund von fehlender Differenzierung nach Geschlecht oder der Maskierung von Subgruppenfehlern durch globale Modellmetriken

- Forschungsdesign: Geschlecht als erklärender Faktor oder als Moderator?
- Unterscheidet das bestehende Modell zwischen Frauen, Männern und geschlechtsdiversen Menschen? Wie schneidet das Modell in den Subgruppen ab?
- Berücksichtigt das Modell neben geschlechtsspezifischen Unterschieden auch geschlechtsspezifische Faktoren bei Frauen (wie Schwangerschaft), Männern (wie die Anfälligkeit für Prostatakrebs) und geschlechtsdiversen Menschen (wie Hormontherapie)?
- Berücksichtigt das bestehende Modell die Unterschiede in den Einstellungen, Bedürfnissen und Interessen von Menschen mit unterschiedlichem Geschlecht?
- Haben wir auf Proxy-Variablen geprüft?
- Integrieren wir multiplikative Effekte verschiedener, aber miteinander verknüpfter Kategorien? (intersektionale Analyse)

Ergebnispräsentation / Interpretation

Gender Bias aufgrund von unreflektierter Verallgemeinerung, der visuellen oder sprachlichen Verstärkung von Stereotypen, oder dem Verschweigen von geschlechtsspezifischen Ergebnissen und schlechteren Modellleistungen in Subgruppen

- Sind die verwendete Sprache und die Abbildungen geschlechtergerecht?
- Werden Grafiken, Diagramme oder Bilder, die zur Veranschaulichung abstrakter Konzepte dienen, unbeabsichtigt geschlechtsspezifisch dargestellt?
- Verallgemeinern wir unzulässig?
- Welche Referenzmodelle werden genutzt?
- Legen wir Schwachstellen offen?

Evaluation

Gender Bias aufgrund der Überprüfung anhand verzerrter Testdaten oder dem Ignorieren geschlechtsspezifischer Fehlerraten

- Haben wir die Fehlerraten bei Gruppen mit unterschiedlichem Geschlecht berechnet oder tarnt die globale Genauigkeit das Versagen bei einer Subgruppe?
- Wie sensitiv reagiert das Modell auf Geschlechts-Proxies?
- Sind die Validierungsdaten aktuell genug?

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Aufgabe:

Analysiert möglichen Gender Bias entlang des gesamten Datenlebenszyklus:

- Problemdefinition / Forschungsfrage / Hypothesenbildung
- Datensammlung
- Datenaufbereitung
- Analyse / Modellierung
- Ergebnispräsentation / Interpretation
- Evaluation

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Problemdefinition / Forschungsfrage / Hypothesenbildung

- Hypothesen könnten auf veralteten Annahmen basieren; zum Beispiel, dass Burnout vor allem "leistungsorientierte Führungskräfte in Vollzeit" (oft männlich) betrifft.
- Folge: Symptome, die eher bei Frauen oder in Teilzeit/Care-Arbeit-Leistenden auftreten (z. B. emotionale Erschöpfung durch Doppelbelastung), werden in der Forschungsfrage gar nicht erst berücksichtigt.

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Datensammlung

- Unternehmen sammelt die Trainingsdaten vor allem in den Abteilungen, in denen historisch gesehen am meisten Überstunden erfasst werden
- Problem: Diese Abteilungen sind oft männerdominiert.
- Bias: Frauen im Datensatz unterrepäsentiert
- Folge: Modell lernt Burnout-Symptome von einer Gruppe, die überwiegend aus Männern besteht

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Datenaufbereitung

- Bias: "Lücken" im Lebenslauf wegen Elternzeit oder Teilzeitarbeit könnten unreflektiert als "Instabilität" codiert werden.
- Folge: Strukturelle Benachteiligungen (die statistisch häufiger Frauen betreffen) werden als individuelle Risikofaktoren für Burnout fehlinterpretiert

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Analyse / Modellierung

- Bias: Ein unkorrigiertes Modell optimiert die Vorhersage für die Mehrheitsgruppe.
- Wenn das Feature "Überstunden" als Hauptindikator gelernt wird, fallen Personen, die aufgrund von Care-Arbeit pünktlich gehen müssen, aber unter extremem Stress stehen, durch das Raster

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Ergebnispräsentation / Interpretation

- Bias: Wenn das Modell am Ende ausgibt: "Männer haben ein höheres Burnout-Risiko" (weil das Modell auf deren Symptome trainiert wurde), könnte die fehlerhafte Interpretation lauten: "Frauen sind stressresistenter"

Szenario:

Ein Team von Data Scientist entwickelt ein Modell, das Personen mit hohem Burnout-Risiko identifiziert

Evaluation

- Bias: Wenn die Genauigkeit des Modells nur über die Gesamtpopulation gemessen wird, fällt nicht auf, dass es bei Männern zu 90% richtig liegt, bei Frauen aber nur zu 50%.

E-learning Portal

Der Einschreibeschlüssel lautet “Gender2026” (ohne die Anführungszeichen)

Kurswebsite

https://dariatisch.github.io/gender_data_science_2026