INF

FAKULTÄT FÜR
INFORMATIK

Beware of the Bias

Wann Daten zu diskriminierenden Entscheidungen führen

Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. Das Szenario
5. Unsere Hypothesen
6. Das Experiment
7. Die Ergebnisse
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

1. Das Team: BIASpects

- **B.Sc. Gesa Götte**
  Statistik (Master, 4.Semester)

- **B.Eng. Marcel Öfele**
  Digital Engineering (Master, 4.Semester)

- **B.Eng. Viviane Lisa Wolters**
  Digital Engineering (Master, 4.Semester)

2. Projektidee

- Projektseminar: "Responsible Data Science“
- Thema: Algorithmen mit Benachteiligung von Gruppen ("Bias") und wie man diesen erkennen kann
- Woher kommt dieser Bias?
- Was beeinflusst, ob ein Lernalgorithmus "diskriminierend" ist?
- Wie reproduzieren sich Ungleichheiten in der Gesellschaft, über Daten in Lernalgorithmen, die wir fur Entscheidungen einsetzen?

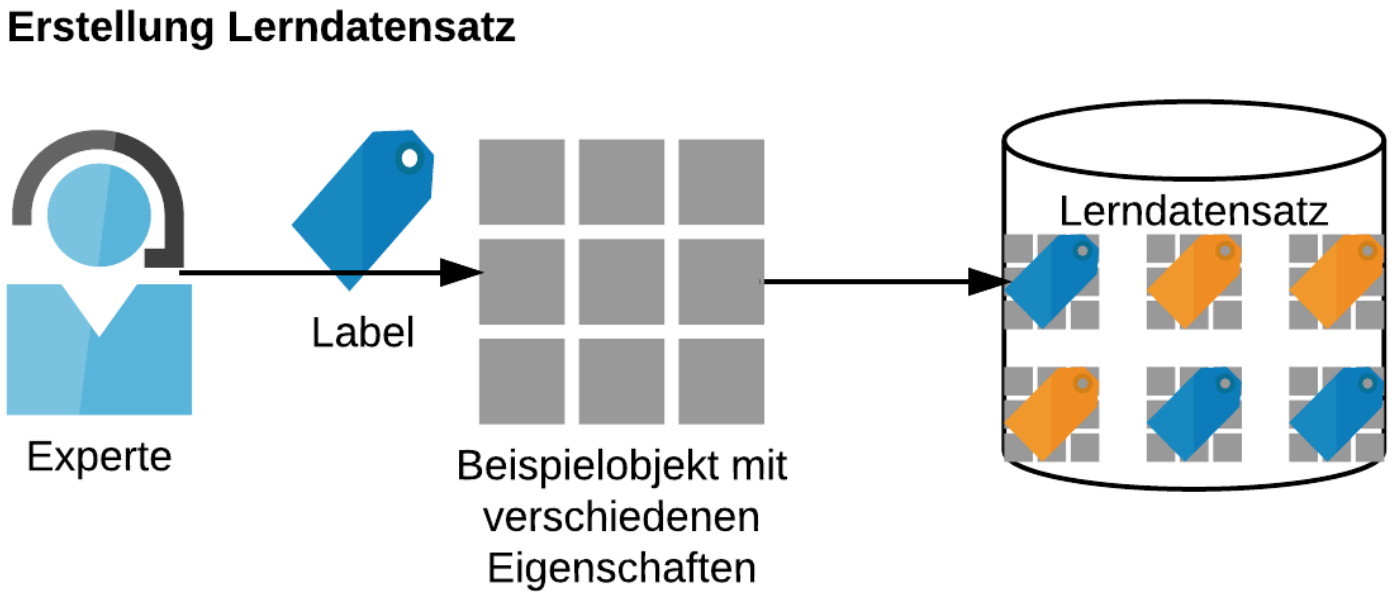
Experimentelle Studie zur Erforschung dieses "Biases"

Agenda

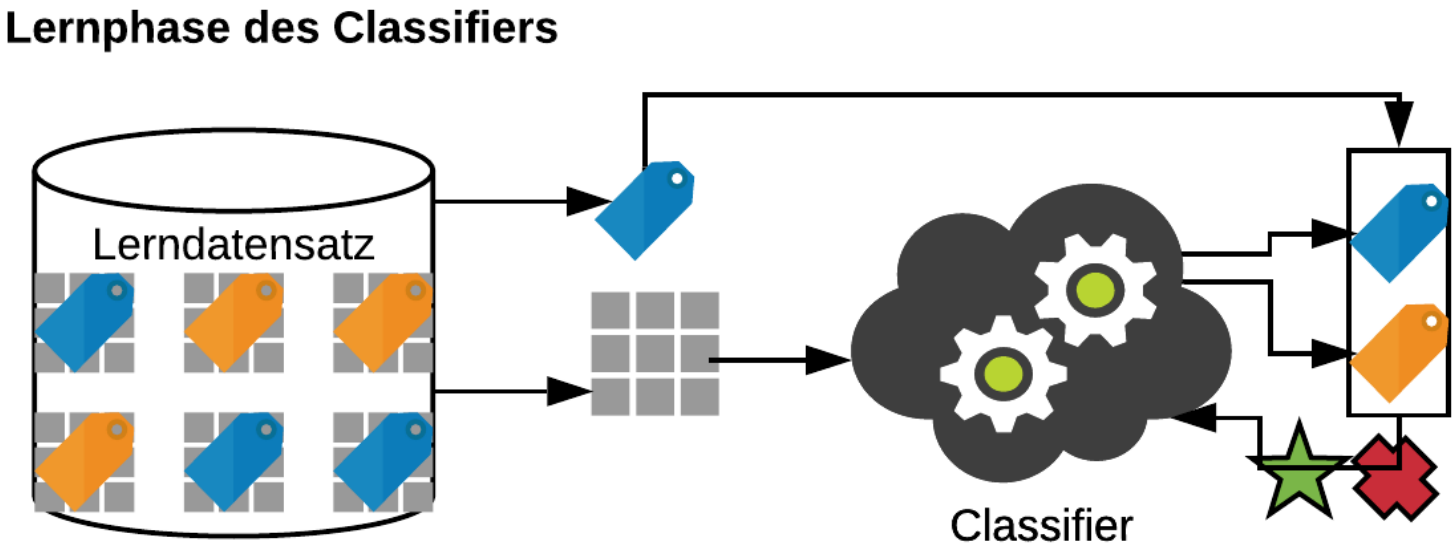
1. Das Team: BIASpects
2. Projektidee
3. **Theorie**
4. Das Szenario
5. Unsere Hypothesen
6. Das Experiment
7. Die Ergebnisse
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

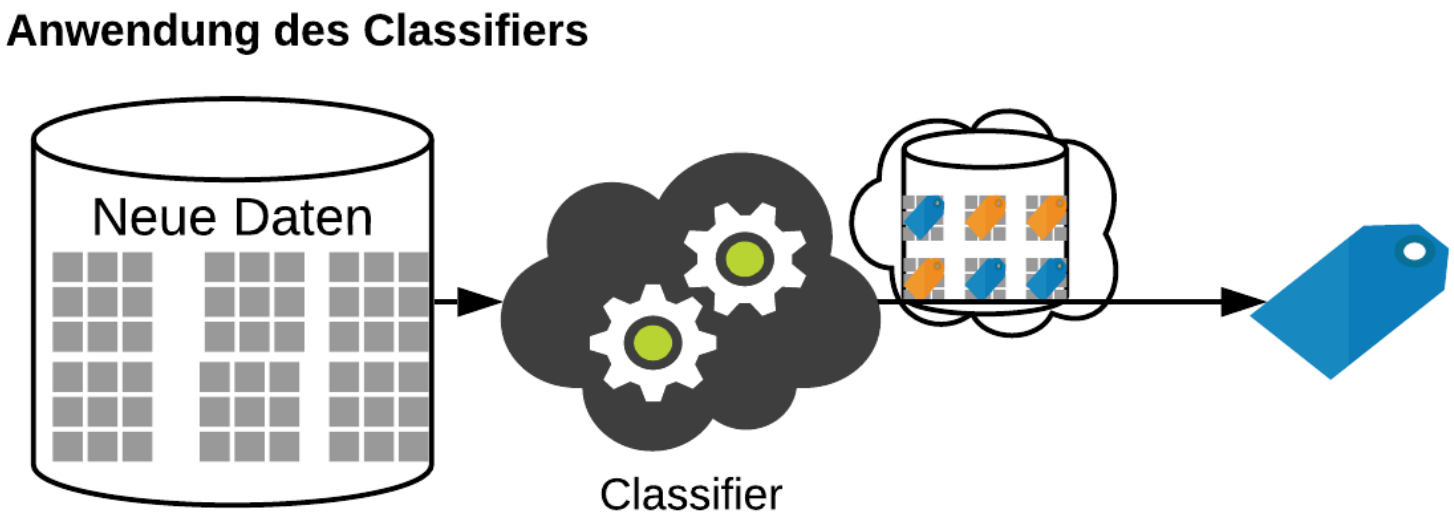
3.1 Arbeitsweise eines Machine Learning Classifier



3.1 Arbeitsweise eines Machine Learning Classifier



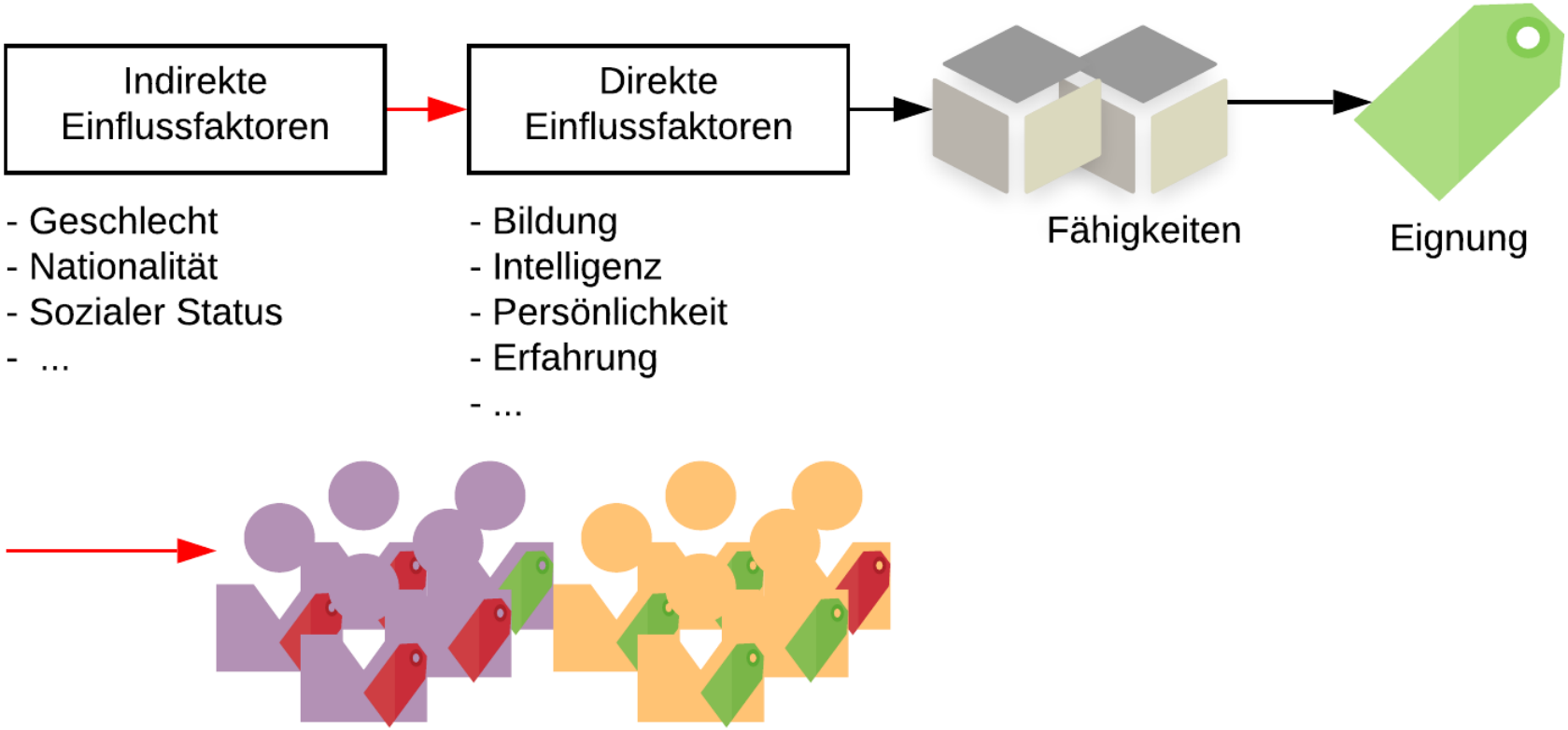
3.1 Arbeitsweise eines Machine Learning Classifier



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

3.2 Definition und Entstehung des Bias

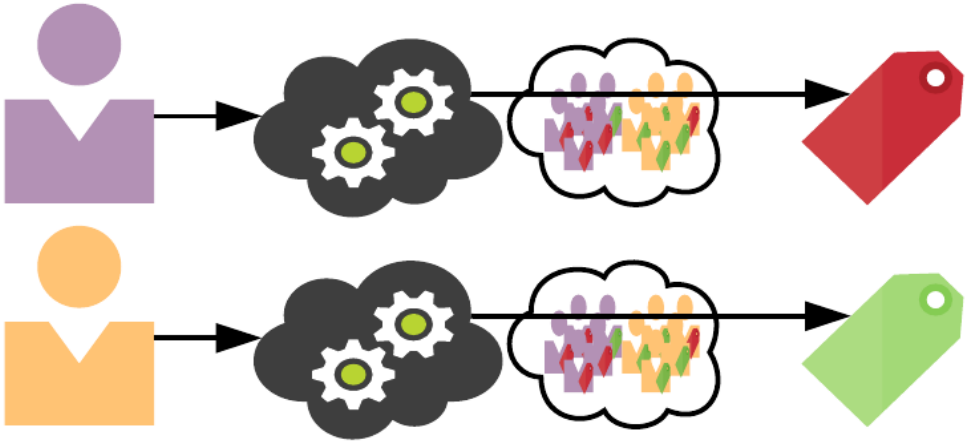
Bias = Verzerrung der Wirklichkeit



3.2 Definition und Entstehung des Bias

Bias im Datensatz wird gelernt

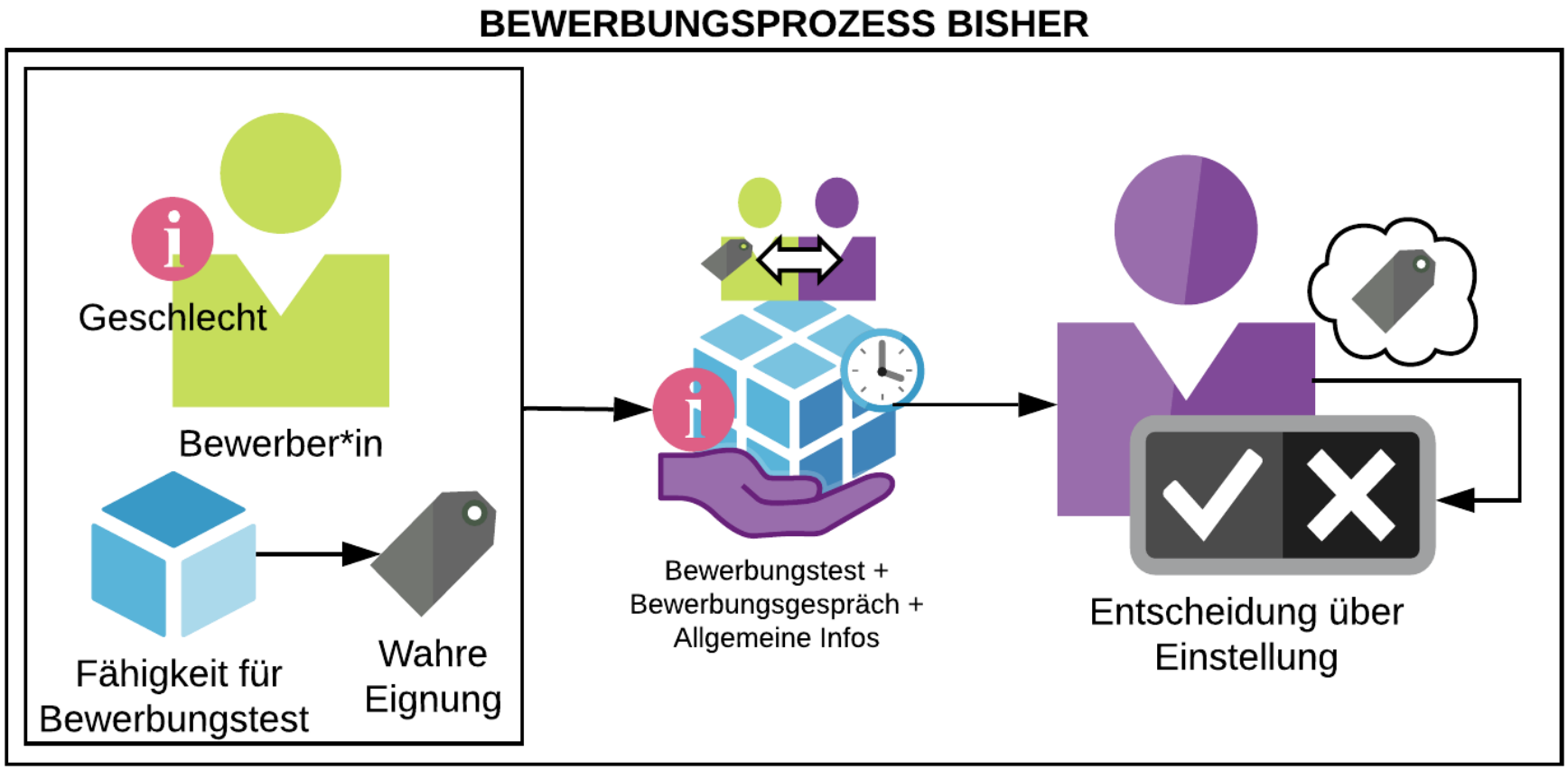
Auswirkung auf spätere Beurteilungen



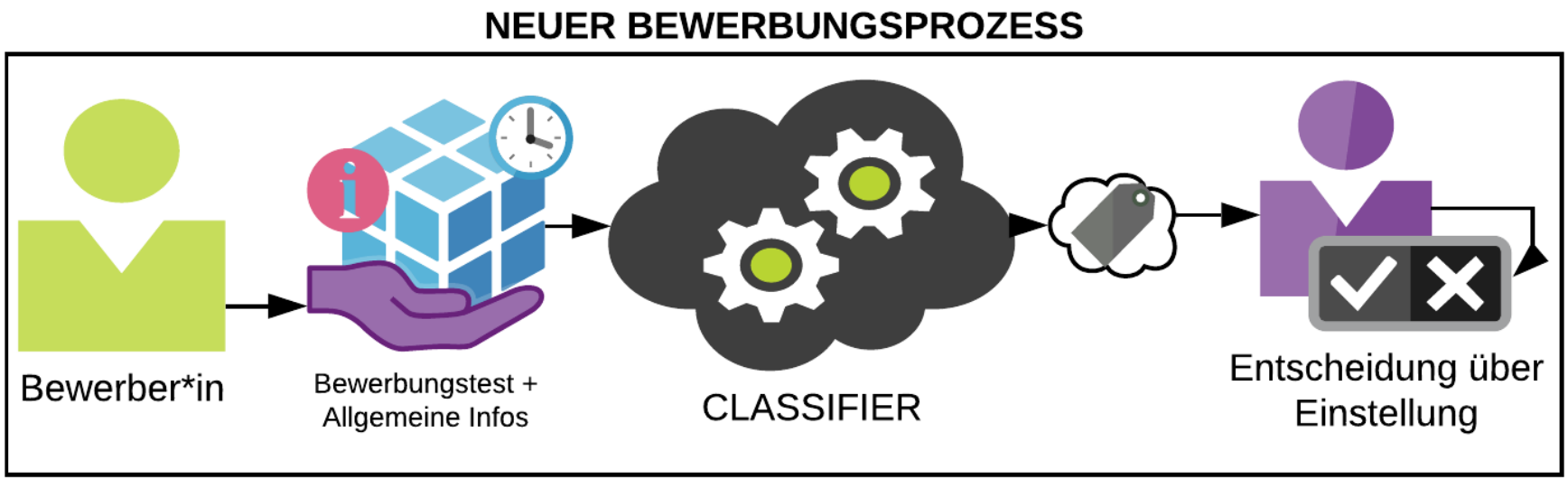
Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. **Das Szenario**
5. Unsere Hypothesen
6. Das Experiment
7. Die Ergebnisse
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

4. Das Szenario



4. Das Szenario



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. Das Szenario
5. **Unsere Hypothesen**
6. Das Experiment
7. Die Ergebnisse
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

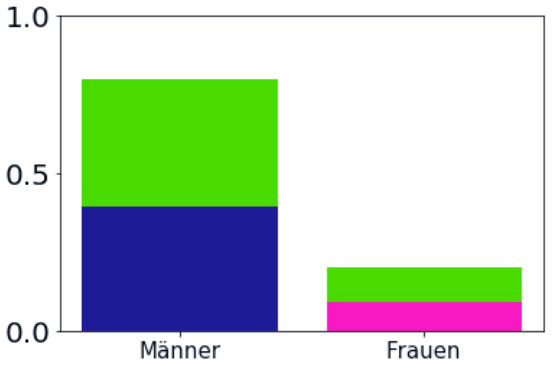
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

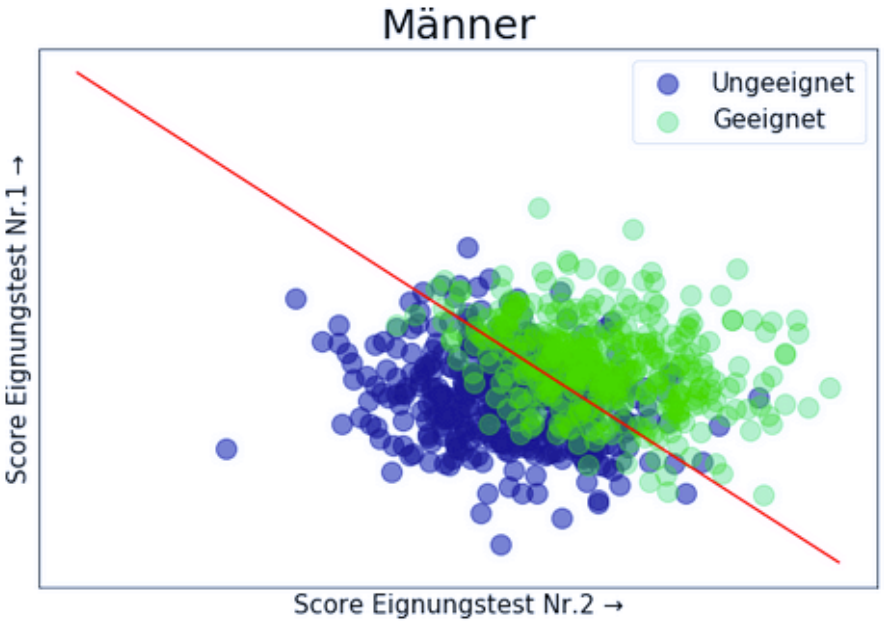
5.1 Hypothese 1

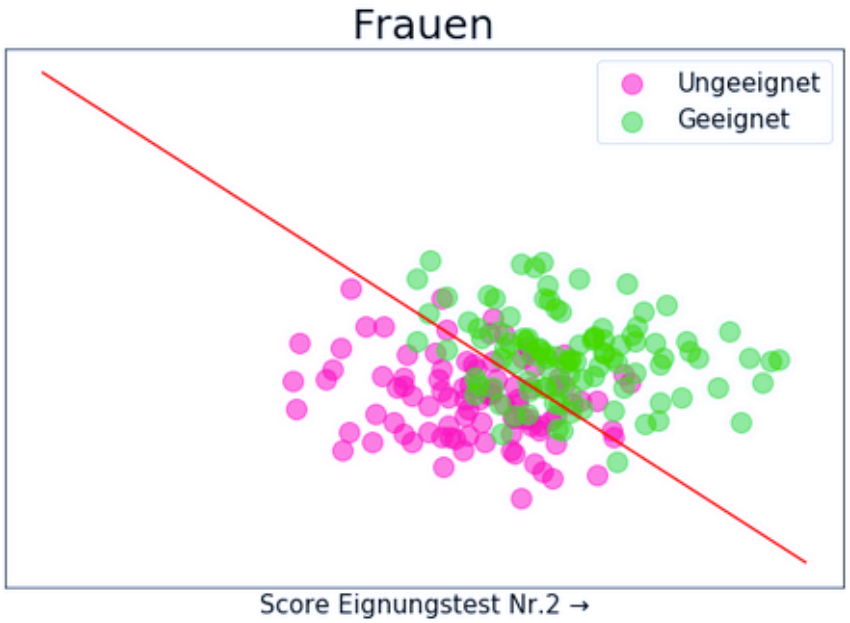
Kein Bias, wenn

- Anteil geeigneter Frauen = Anteil geeigneter Männer

  (Anzahl der Frauen bzw. Männer nicht relevant)







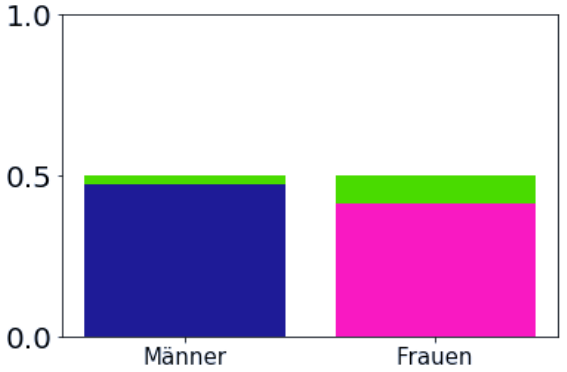
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

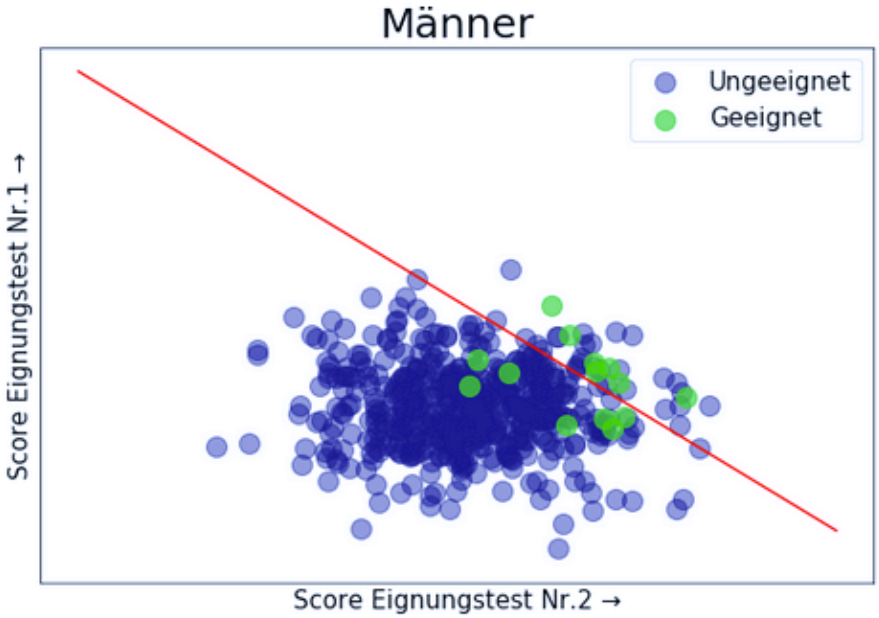
5.2 Hypothese 2

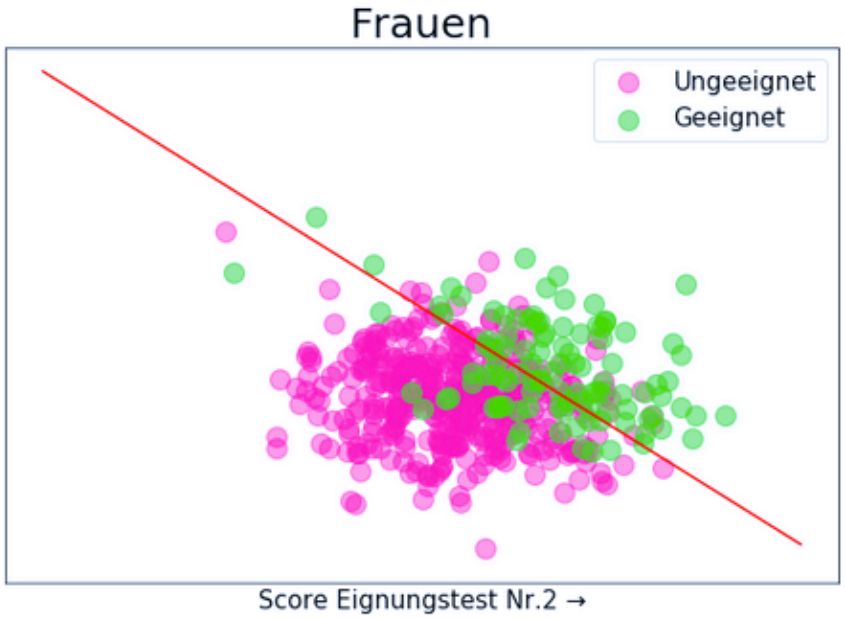
Kein Bias, wenn

➤ Anzahl Frauen = Anzahl Männer

(Anteil geeigneter Frauen bzw. Männer nicht relevant)





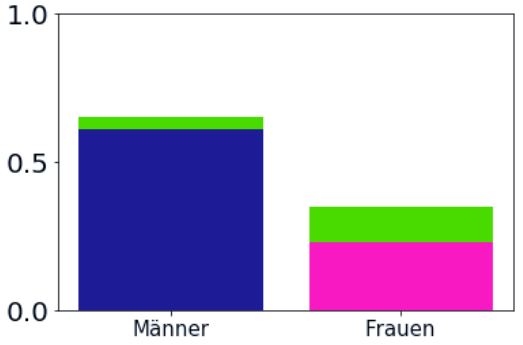


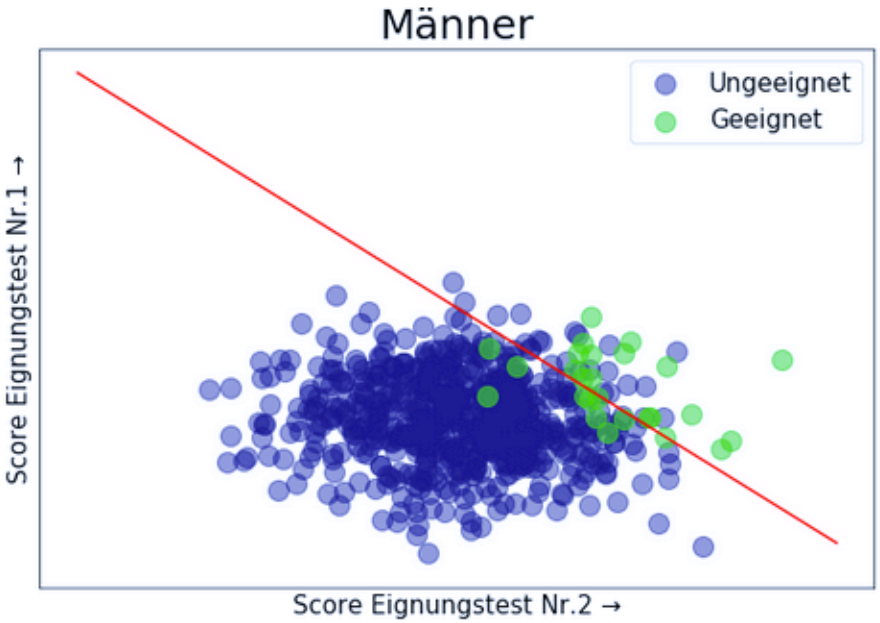
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

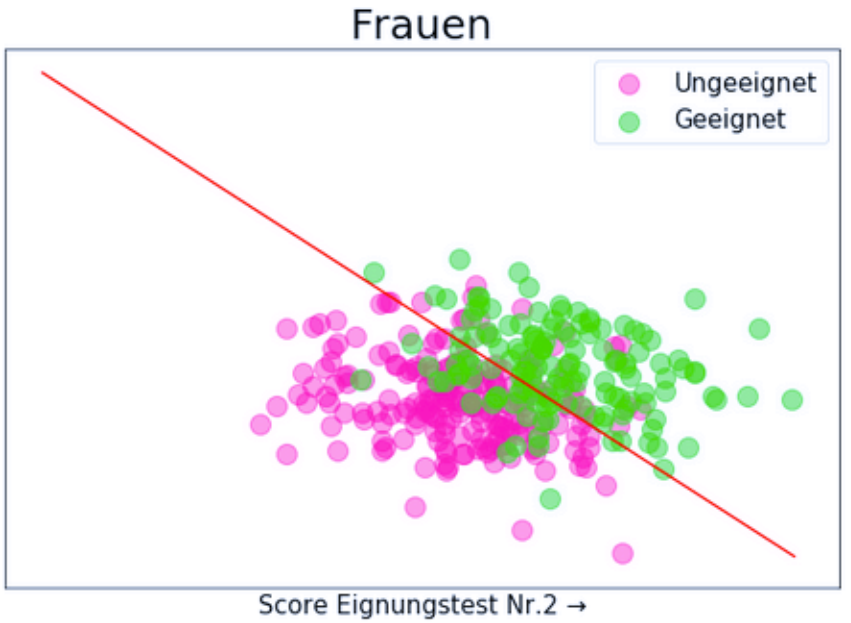
5.3 Hypothese 3

Bias, wenn

- Anzahl Frauen ≠ Anzahl Männer
- Anteil geeigneter Frauen ≠ Anteil geeigneter Männer

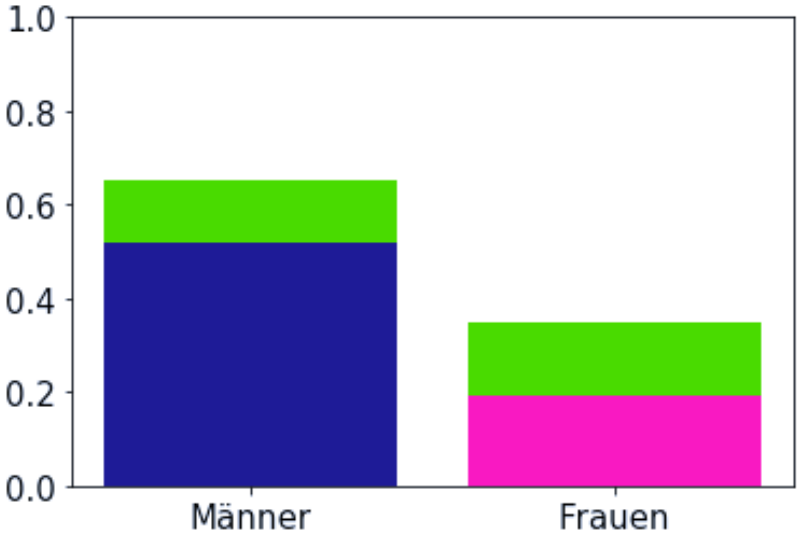


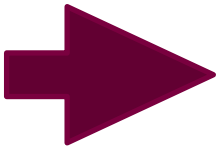


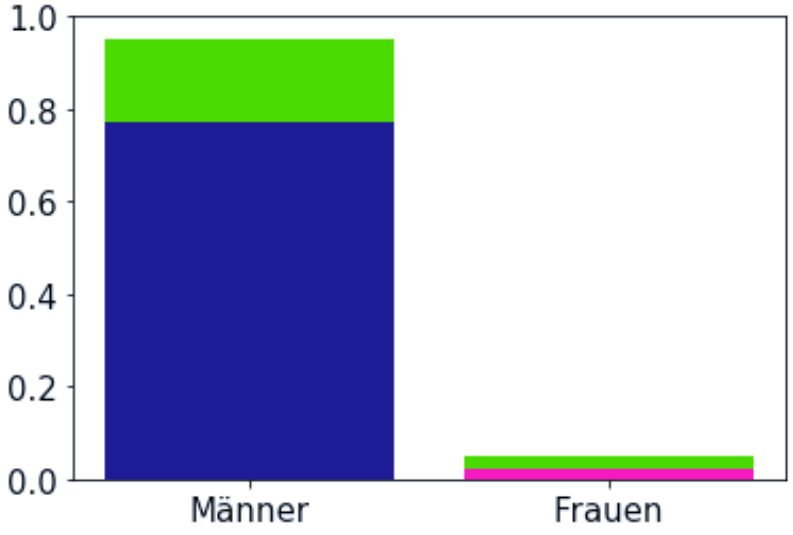


5.3.1 Hypothese 3.1

- Je größer die Überrepräsentation, desto größer der Bias

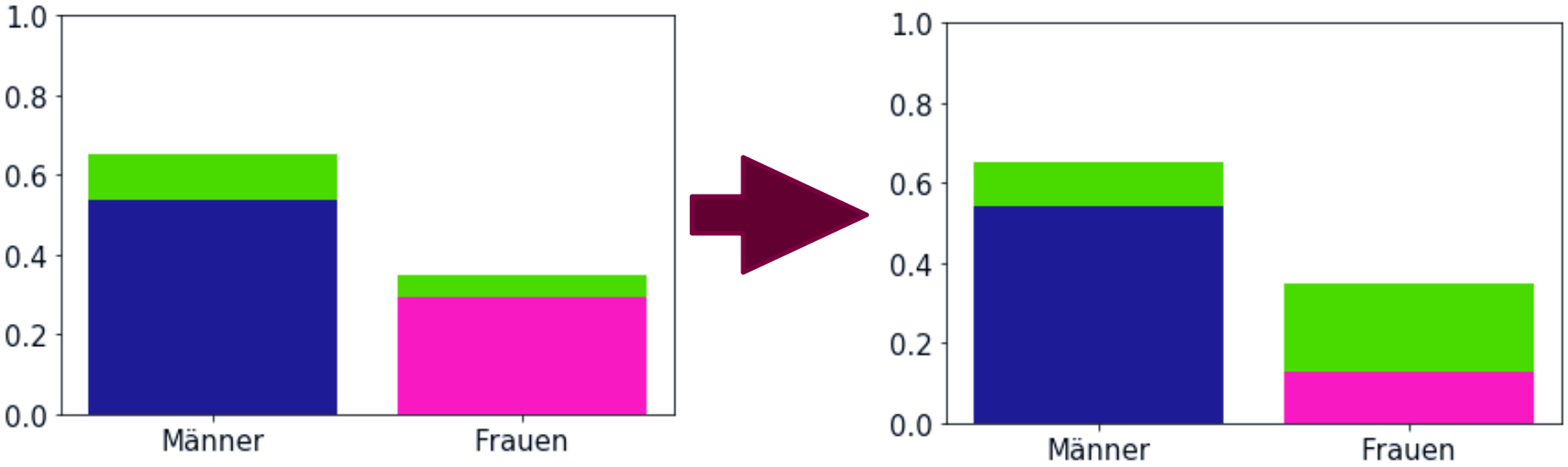






5.3.2 Hypothese 3.2

- Je größer der Unterschied zwischen den Anteilen der geeigneten Männer und Frauen, desto größer der Bias.

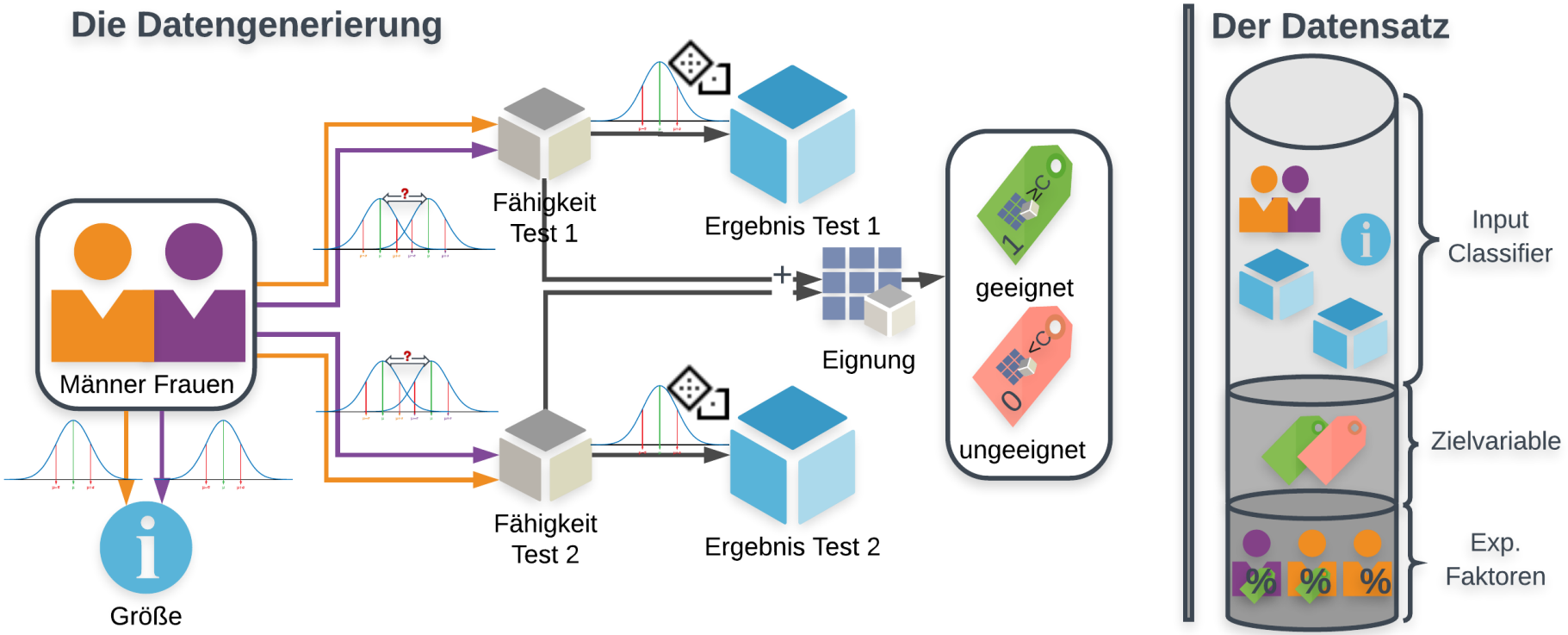


Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. Das Szenario
5. Unsere Hypothesen
6. **Das Experiment**
7. Die Ergebnisse
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

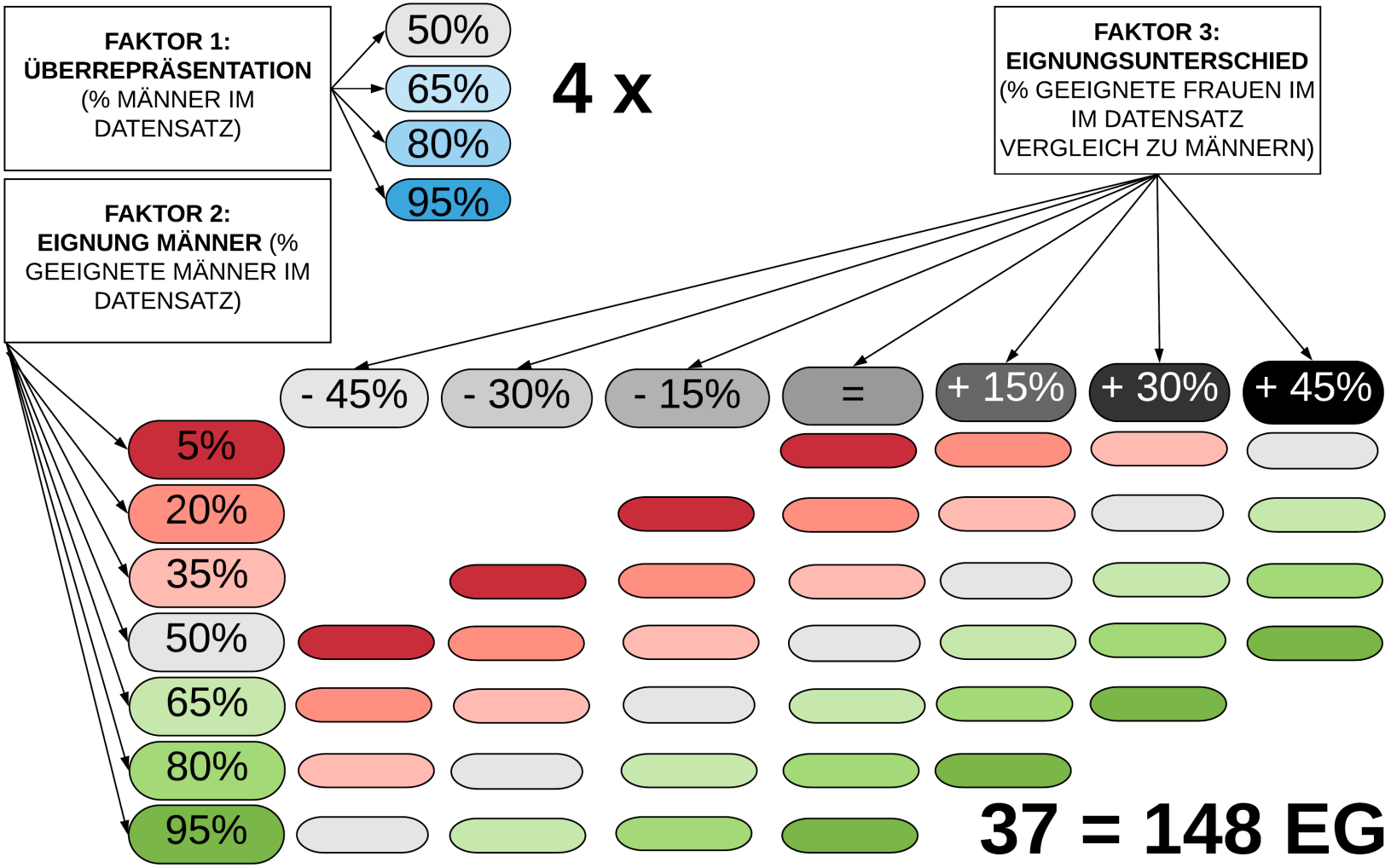
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

6.1 Datensatz



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

6.2 Aufbau/Ablauf

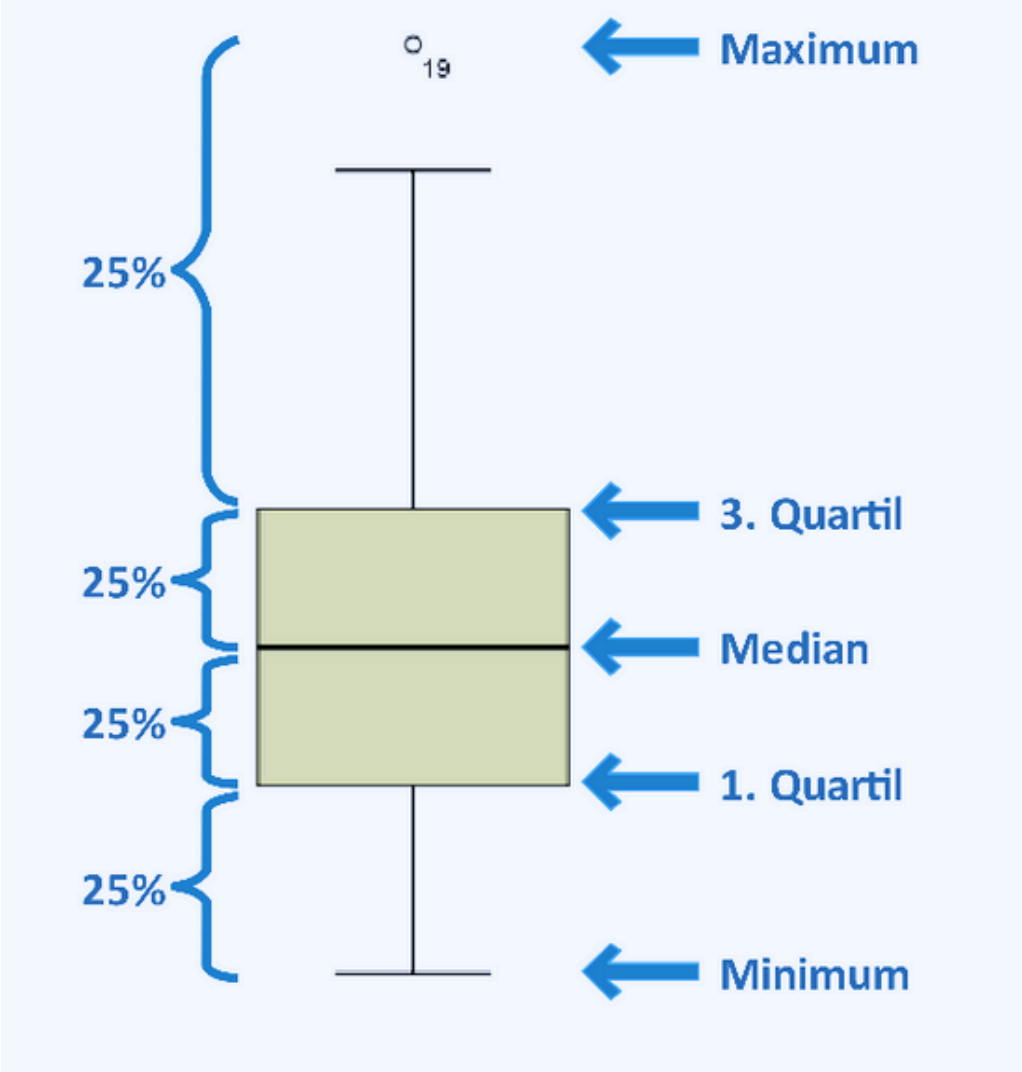


Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. Das Szenario
5. Unsere Hypothesen
6. Das Experiment
7. **Die Ergebnisse**
8. Reduzierung des Bias
9. Zusammenfassung
10. Ausblick

7.1 Unterschied zwischen den Classifiern

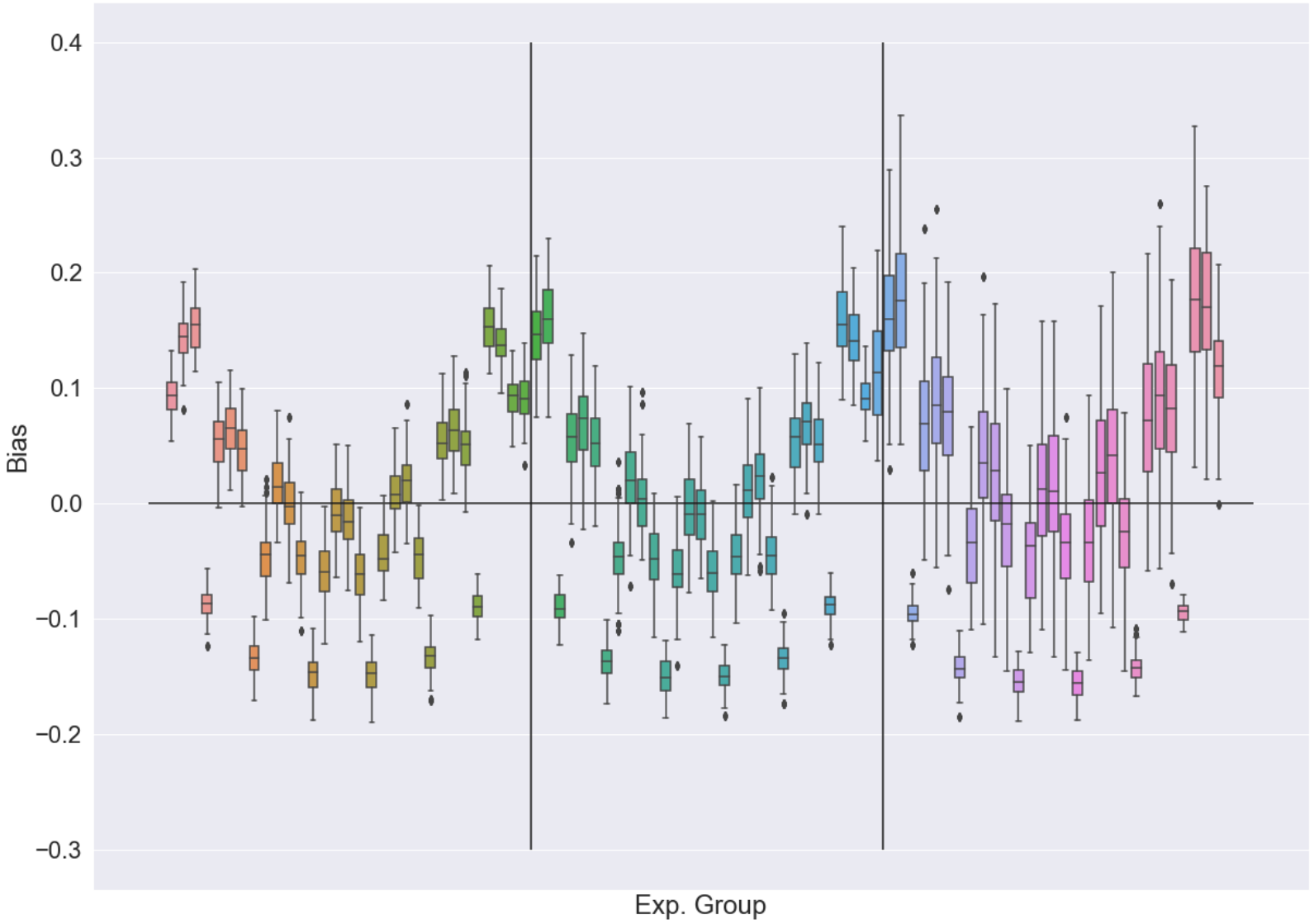
7.2 Exkurs: Boxplots



7.3 Hypothese 1

7.4 Hypothese 2

7.5 Hypothese 3



7.5.1 Hypothese 3.1

7.5.2 Hypothese 3.2

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

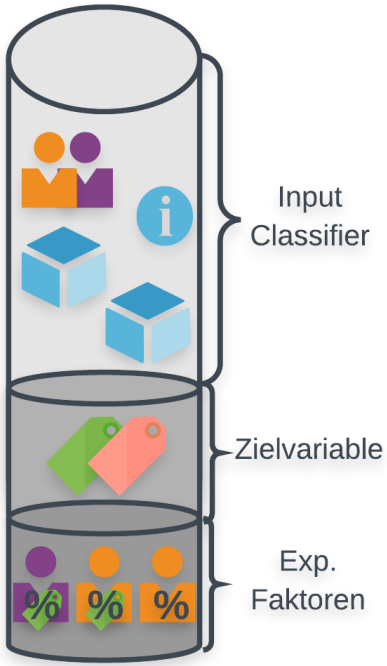
Agenda

1. Das Team: BIASpects
2. Projektidee
3. Theorie
4. Das Szenario
5. Unsere Hypothesen
6. Das Experiment
7. Die Ergebnisse
8. **Reduzierung des Bias**
9. Zusammenfassung
10. Ausblick

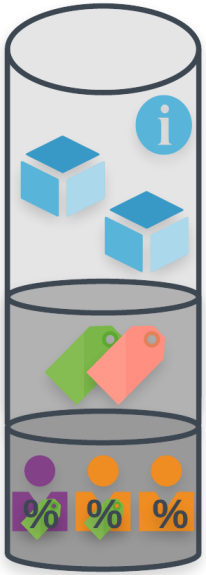
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

8.1 Methoden

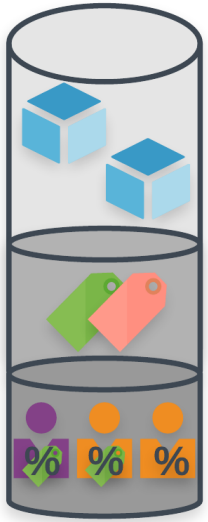
Standard A0



Methode A1 ("Ohne Geschlecht"):
Entfernen des Geschlechts



Methode A2 ("Ohne geschlechtsspezifische Informationen"):
Entfernen aller Merkmale mit Geschlechtsinformation



Methode B ("50-50-50"):
Lerndatensatz mit 50% Männern, 50% geeigneten Männern und 50% geeigneten Frauen (Testdatensatz wie gehabt)

8.1 Methode 1

8.1 Methode 2

9. Zusammenfassung

Ursachen Bias

➤ Gibt es Unterschiede im Anteil geeigneter Personen zwischen zwei Gruppen, dann ist der Classifier für eine der beiden akkurater (Bias)

➤ Ist eine Gruppe sehr stark überrepräsentiert, kann sich dieser Bias verstärken

➤ Unscharfe Messmethoden der Eignung führen dazu, dass es einen unvermeidbaren Bias zwischen zwei Gruppen mit unterschiedlichem Informationsgehalt gibt

9. Zusammenfassung

Vermeidung Bias

- Informationen über die Gruppe aus dem Lerndatensatz zu entfernen reicht nicht aus um diesen Bias zu verhindern

- Die Verteilungen im Lerndatensatz anzugleichen vermindert den Bias

10. Ausblick

Was sollte noch untersucht werden?

- Sehr vereinfachte Daten
  - -> Andere Datenstrukturen sollten untersucht werden
  - -> Versuch mit echten Daten

- Hier: Accuracy-Differenz als Bias

  ABER: Im Kontext sollte immer bedacht werden, ob es eine Art von Fehlentscheidungen gibt, die "schlimmer" ist
  - -> Fokus nur auf eine Art von Fehlentscheidungen (z.B. Nicht-Erkennen von geeigneten Personen)

- Einstellungen der Classifier könnten noch optimiert werden
  - -> mögliche weitere Methode zur Biasreduktion

10. Ausblick

Was lernen wir aus der Studie?

- Ergebnisse nicht nur auf Männer und Frauen anwendbar
  (auch andere Ungleichheiten durch z.B. Nationalität oder soziale Schicht)

- Bias für unterschiedliche Gruppierungen immer messen!

- Bias entdeckt - Und nun?
  -> Es genügt nicht nur die Informationen über die Gruppenzugehörigkeit aus dem Datensatz herauszunehmen!
  -> 50-50-50 Lerndaten
  -> Messinstrumente überdenken (Quantität und Qualität)

Vielen Dank für Ihre Aufmerksamkeit!

www.ovgu.de