Messung von Bias

vorgestellt von: Denis Logvinenko
Papier: N.Garg et al.(2018)

Institut für Computerlinguistik
Universität Heidelberg

19. November 2019

Plan der Präsentation

1. Einleitung und Motivation
2. Daten und Methoden
3. Bestätigung der Existenz des Bias in Embeddings
4. Trends in den Stereotypen
5. Diskussion

Einleitung und Motivation

Einleitung und Motivation

1. Embeddings sind gut für semantische Relationen
2. Unser Ziel ist es, diese Info auch für zeitliche Untersuchungen zu benutzen z.B. um:
   - existierende **geschlechtliche** oder
   - **ethnische** Stereotype zu finden
3. Was Embeddings gut erfassen können:
   - **Frauenbewegungen** der 60er-Jahre
   - Veränderungen in **Zuwanderung** aus verschiedenen Regionen
4. Unser Ziel ist es, die qualitativeren Studien nicht durch Embeddings-Analysen zu ersetzen, sondern sie dadurch **zu ergänzen**

Daten und Methoden

Daten

1. Statistische Daten stammen von:
   - US Census – Bundesvolkszählungsbehörde
2. word2vec-(SGNS)-Embeddings aus:
   - Google News für die gegenwärtige Sprachanalyse
   - COHA (Corpus of Historical American English) für jedes Jahrzehnt zwischen 1910 und 2000
3. Zur Validierung: GloVe mit Benutzung von:
   - New York Times (1988 – 2005)
4. Idee: untersuchen, ob die Embeddings die Entwicklungen in den statistischen Daten gut erfassen

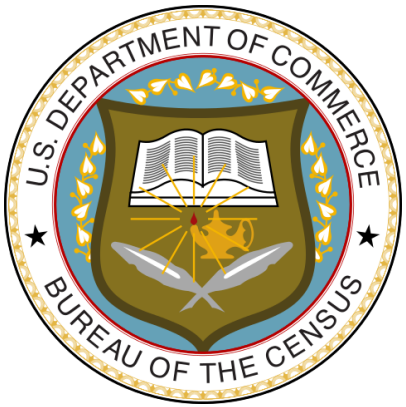


Abbildung:
US Census

Daten

Wortlisten

Gruppenwörter	Neutrale Wörter
Man words: he, son, father	**Occupations:** soldier, tailor
Woman words: she, daughter	**Professional Occupations:** statistician
White/Hispanic/Asian/Russian last names: harris, ruiz, huang mishkin	**Physical Appearance Adjectives:** alluring, ugly
Islam words: allah, ramadan	**Intellectual Adjectives:** wise
Christianity words: baptism, messiah	**Terrorism Related Adjectives:** violence etc.

Bestätigung der Existenz des Bias in Embeddings

Gender Bias vs. Berufsbeteiligung heute

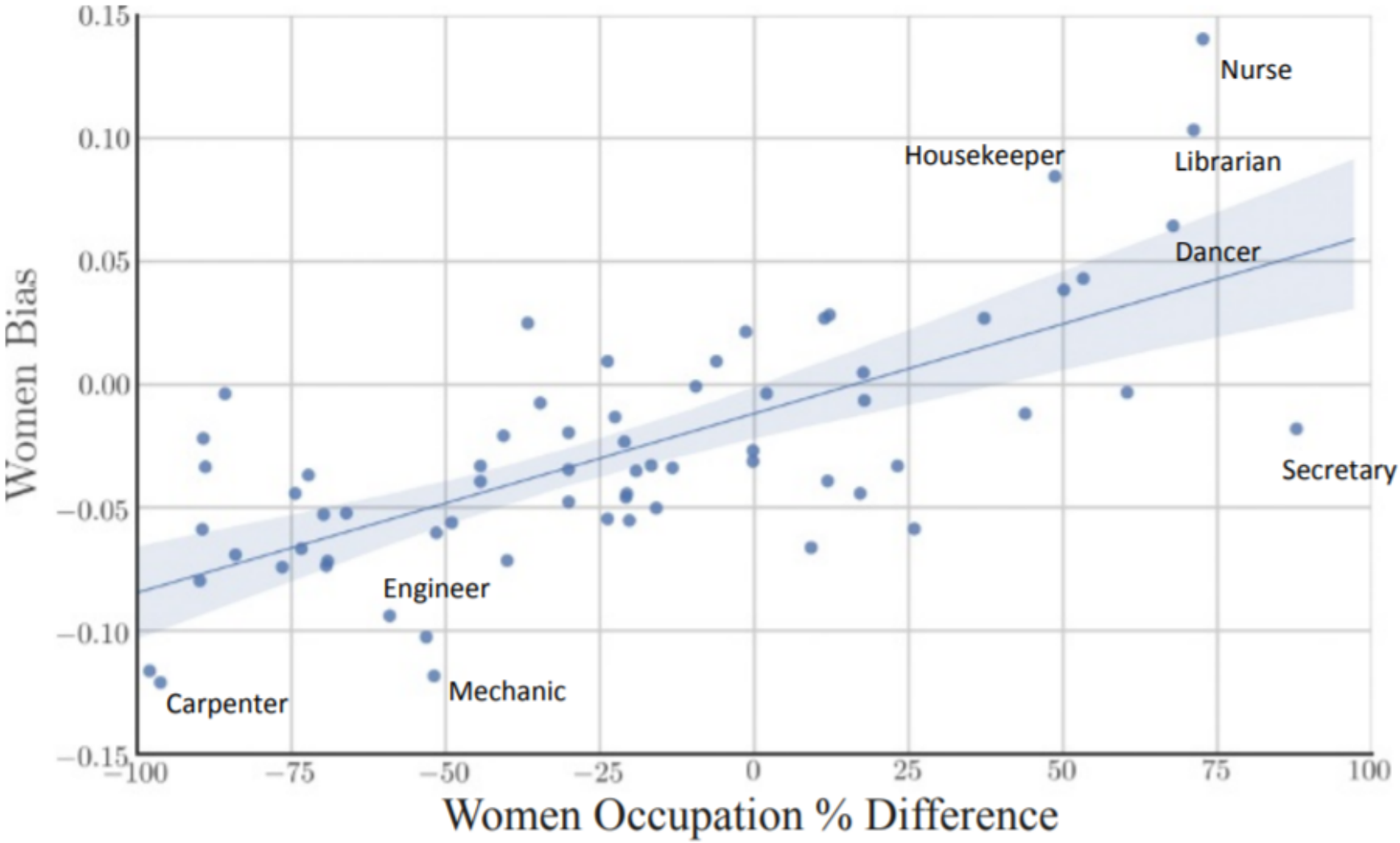


Abbildung: *Frauenberufsbeteiligung vs. Bias in Google-News-Embeddings*

Average Gender Bias vs. Berufsbeteiligung in Census

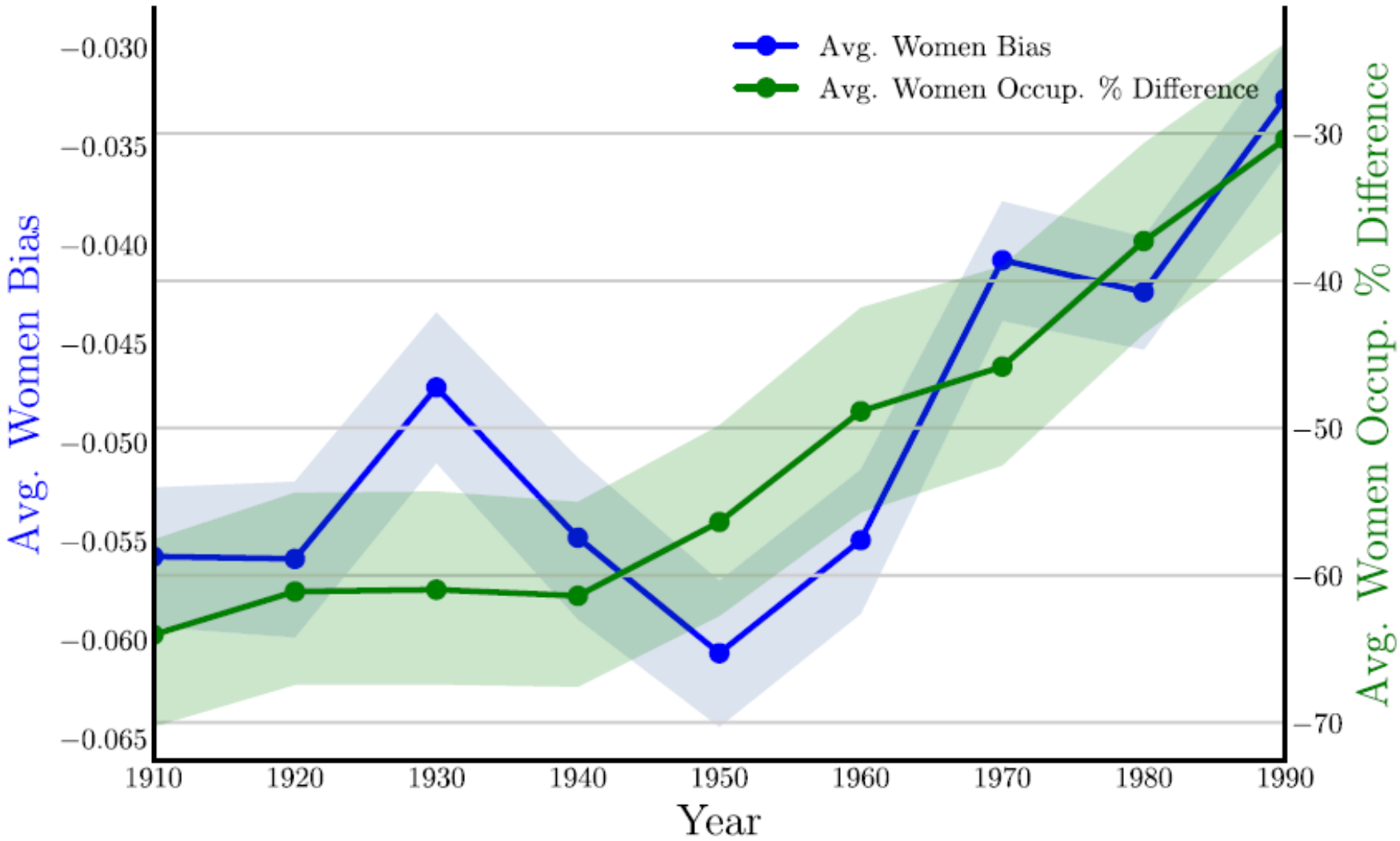


Abbildung: *Frauenberufsbeteiligung vs. Bias in COHA-Embeddings*

Gender Bias vs. Umfragen

Idee: untersuchen, ob die Embeddings auch menschliche Stereotype über Persönlichkeitsmerkmale gut erfassen

Vorgehensweise bei den Umfragen:

1. Menschen vergeben 230 Adjektiven einen Score, ob sie die eher mit Frauen, oder mit Männern assoziieren würden
   - Jahre der Umfragen: 1977, 1990
2. Adjektive sind z.B.:
   - headstrong, quarrelsome, effeminate, fickle, talkative, dependable usw.
3. Ergebnis:
   - Bias in Embeddings (COHA für entsprechende Jahrzehnte) korreliert signifikant mit Menschen-Scores

TOP-10 Berufe für ethnische Gruppen

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

Abbildung: *Die TOP-10 Berufe, die mit jeder Ethnie am engsten assoziiert sind*

Berufsbeteiligung einzelner Ethnien vs. Bias in COHA

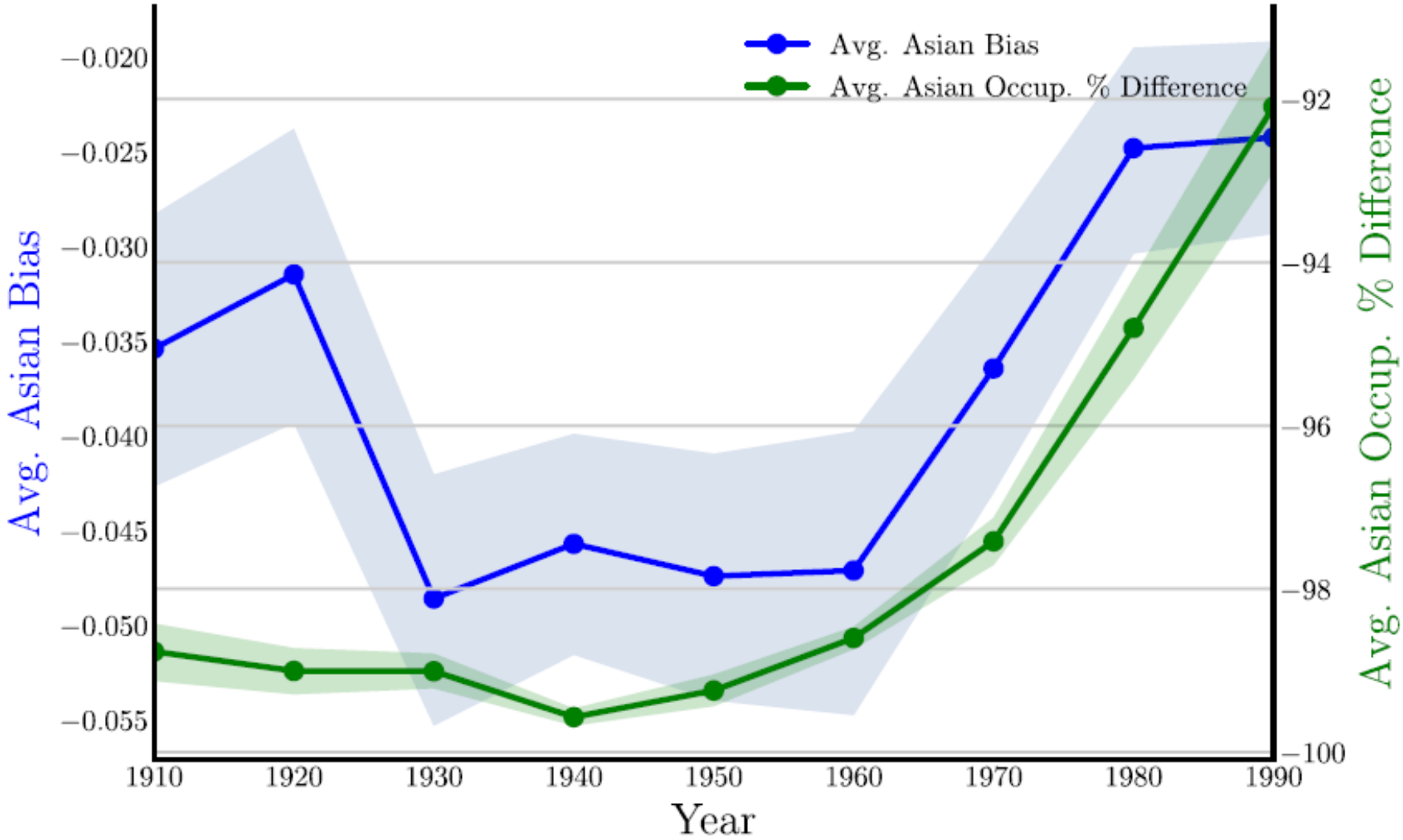


Abbildung: *Berufsbeteiligung von Weißen und Asiaten vs. Bias in COHA-Embeddings*

Ethnischer Bias vs. Princeton-Experimente

Idee: untersuchen, ob die Embeddings auch menschliche Stereotype über Ethnien gut erfassen

Vorgehensweise:

1. Man untersucht die Stereotype der Studenten gegenüber 10 ethnischen Gruppen
   - Im Papier nur die Daten über Chinesen
2. Man stellte 2 Listen mit TOP-15 Stereotypen zusammen. Adjektive sind z.B.:
   - industrious, superstitious, nationalistic
   - Score = %(*Befragte*) , die angegeben haben, dass das Stereotyp auf die Gruppe zutrifft
3. Ergebnis: Embeddings spiegeln die Einstellungen der Zeit wider und sie sind über die Zeit hinweg gut kalibriert

Validierung des Bias – Zwischenfazit

Census:

1. Der Bias-Wert ist ein guter Prädiktor für den Prozentsatz an Menschen in Berufen
2. Berufe, wo $N(Männer) = N(Frauen)$, sind trotzdem mehr mit Männern assoziiert
3. Korrelationen sind über Jahrzehnte sehr ähnlich $\Rightarrow$
   - Verhältnis zwischen Embeddings-Bias und Realität bleibt konsistent

Außerhalb Census:

Der Bias spiegelt menschliche Einstellungen wider und lässt uns zeitliche Entwicklungen beobachten

Trends in den Stereotypen

Trends in Gender Bias

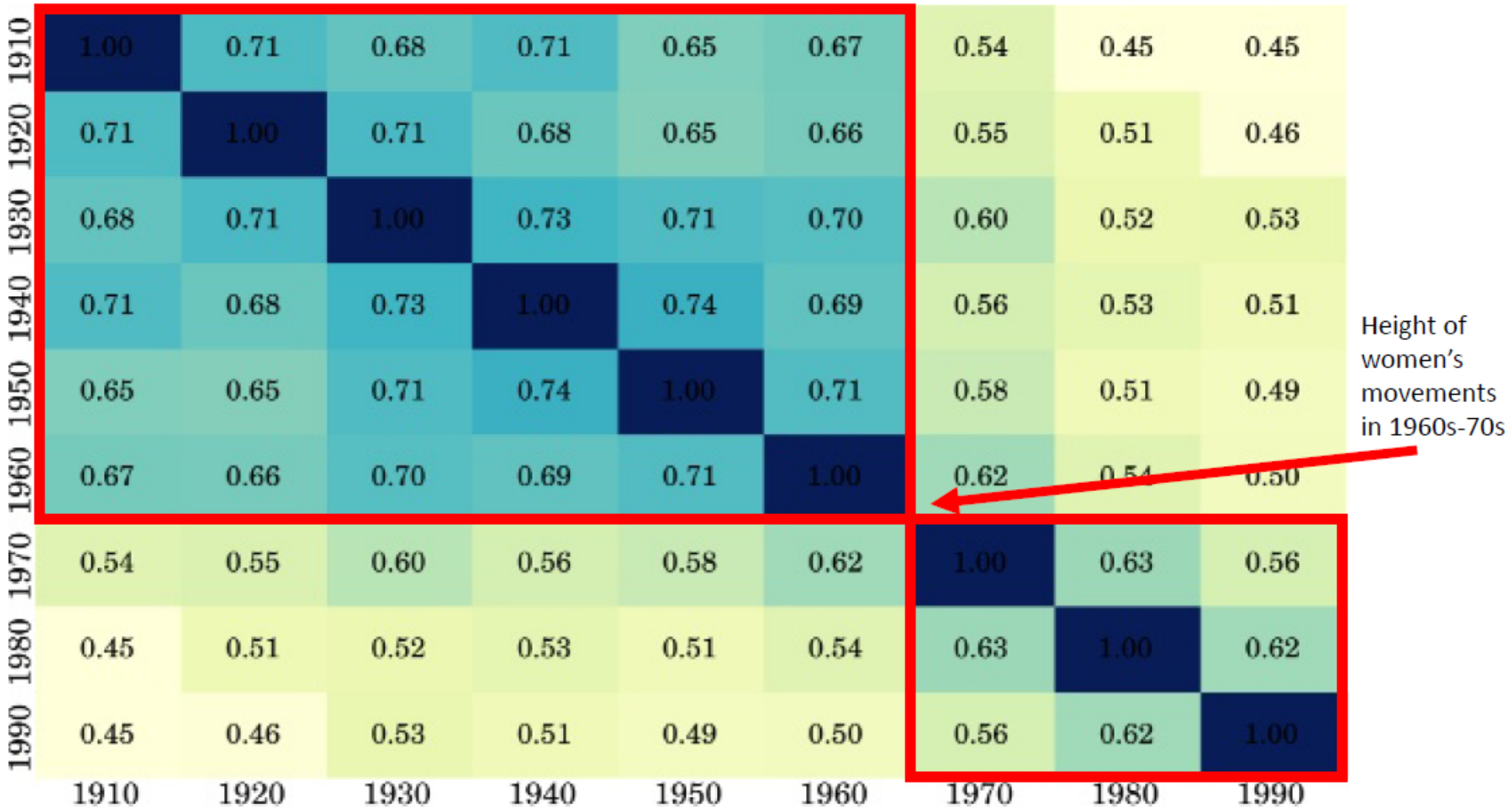


Abbildung: *Pearson correlation in embedding bias scores for adjectives over time between embeddings for each decade.*

Trends in Ethnic Bias

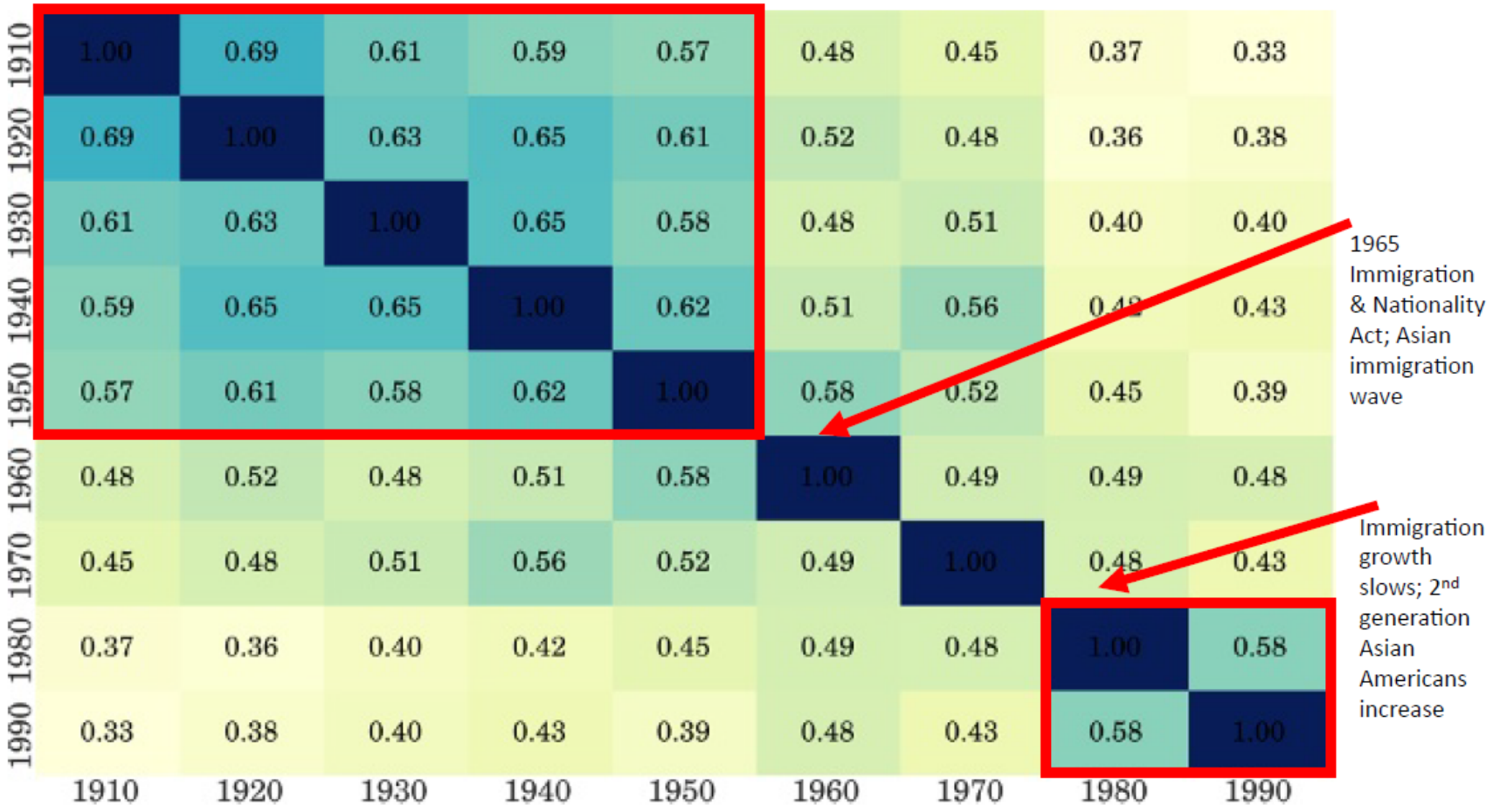


Abbildung: *Pearson correlation in embedding Asian bias scores for adjectives over time between embeddings for each decade.*

Most Biased Adjectives – Trends – Asiaten

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

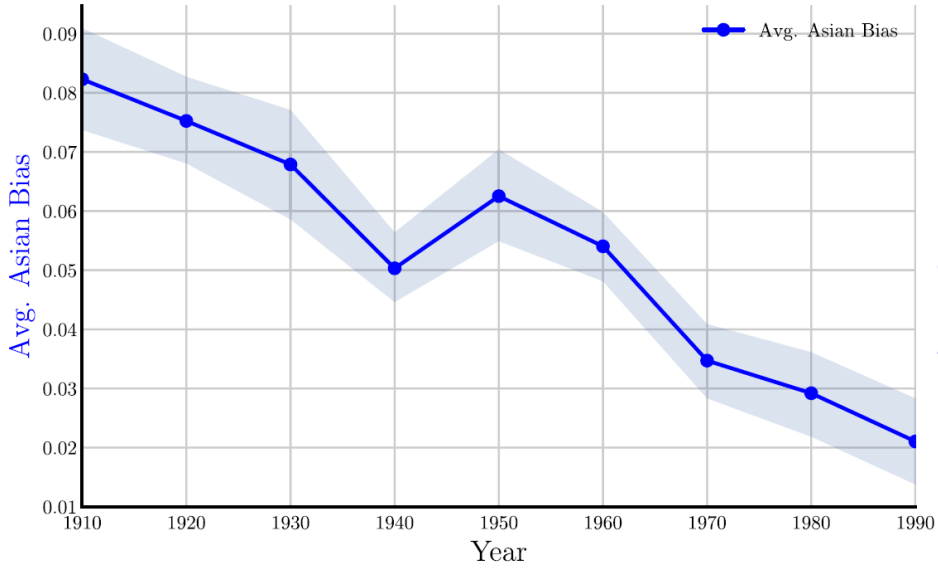
Abbildung: *Die TOP-10 Adjektive, die am engsten mit Asiaten (vs. Weißen) assoziiert sind*

Most Biased Adjectives – Trends – Frauen

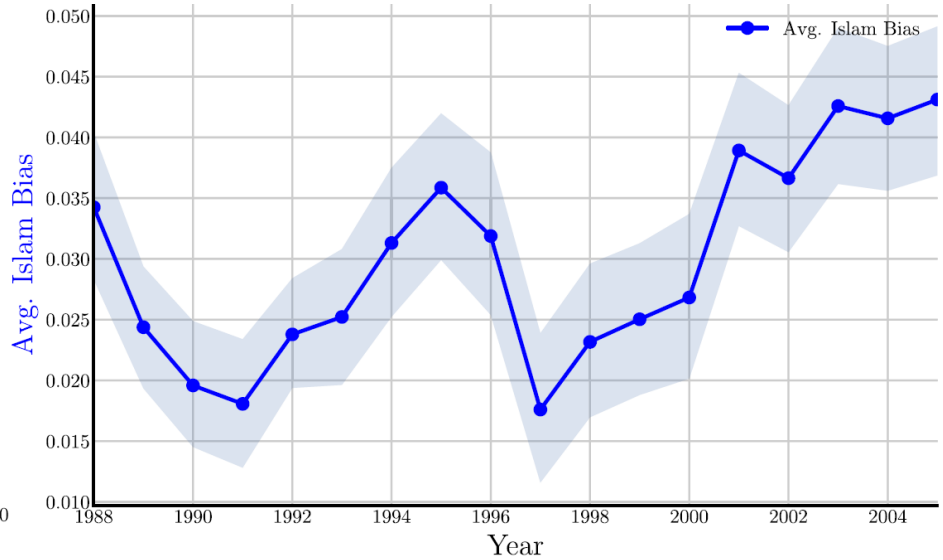
1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Abbildung: *Die TOP-10 Adjektive, die am engsten mit Frauen (vs. Männern) assoziiert sind*

Allgemeine Trends



(a) Asian bias score over time for words related to outsiders in COHA data



(b) Religious (Islam vs. Christianity bias score over time for word related to terrorism in New York data)

Diskussion

Diskussion

- Lineare Modelle der Korrelation zwischen Bias und externen Metriken
  - Was wenn die Korrelationen nicht linear sind?
- Abhängigkeit von spezifischen Wortlisten
  - Was wenn die ausgewählten stereotyphaften Wörter nicht gut sind?
- Wie gut repräsentieren die Texte aus dem Jahr 1910 die Einstellungen der Bevölkerung?
- Embeddings = black box
  - Vielleicht Embeddings benutzen, wo bestimmte Dimensionen verschiedene sprachliche Aspekte erfassen?
- Nachnamen gut, um Ethnien zu identifizieren?
- Asiaten = Chinesen?