Responsible IR

Cor Veenman – TNO - LIACS

Universiteit Leiden
The Netherlands

1

Challenges

- Fairness
  - Ethics
  - Constitution, GDPR
  - Redlining: proxies
- Confidentiality
  - Personal; privacy
  - Commercial
- Accountability
  - Right to explanation
  - Justification
  - Validity

2

Responsible AI

- (F)airness
  - Fair Feature Representation
  - Fair Sample Representation (bias)
- (A)ccuracy
  - In balance with proportionality and purpose limitation
  - Impact
- (C)onfidentiality
  - Hiding sensitive information
- (T)ransparency
  - Process
  - Models
  - Predictions

"How are these secrets getting out? It's almost as if the competition has a fly on the boardroom wall!"

AI Act

Fair Feature Representation
Structured Data

Sensitive attributes (columns)

- Remove attributes
- Replace attributes
  - More general categories
  - Added noise
- Correlated attributes
  - proxies

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Fair Feature Representation
Unstructured Text

Sensitive attributes

- Detecting entities
- Recognizing entity type
- Identification of specific entity

Anonymization

- Removing entities
  - Entity class names
- Replacing entities
  - More general entities
  - Random entities

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Fair Feature Representation
Footage

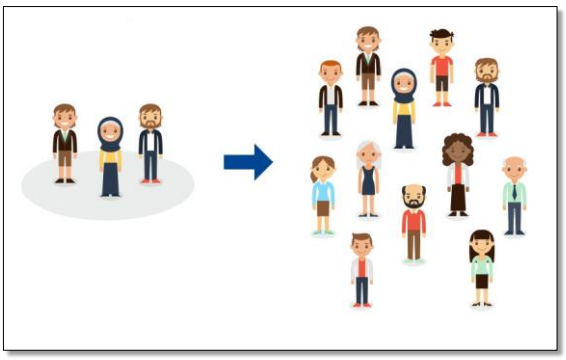
Sensitive attributes
- Detecting ROI
- Recognizing object type
- Identification of person, license plate, etc

Anonymization
- Removing entities
- Replacing entities

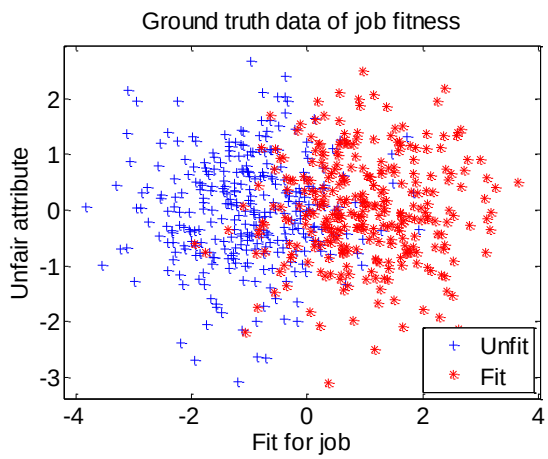
Fair Sample Representation

Representative sample
- Samples, labels

Collected samples and labels are biased
- Only part of the population is inspected
  - Ethnicity, gender
  - Language, education, technology
- Only targets in that part will be found
  - Confirmation
  - White spots

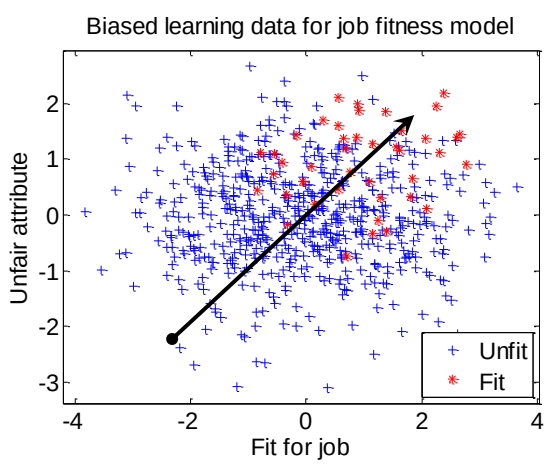


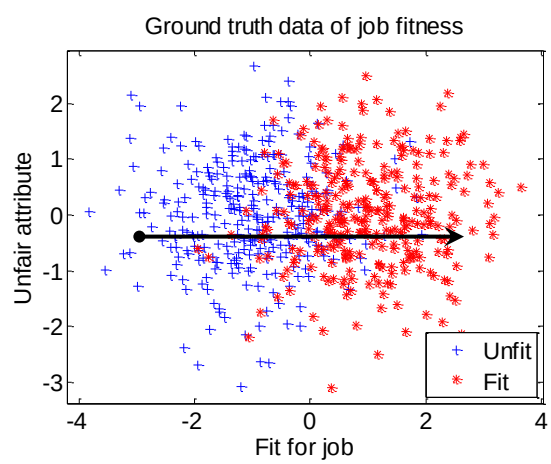
Unfair Attribute: Job Fitness Learning



9

Data Biased on Unfair Attribute





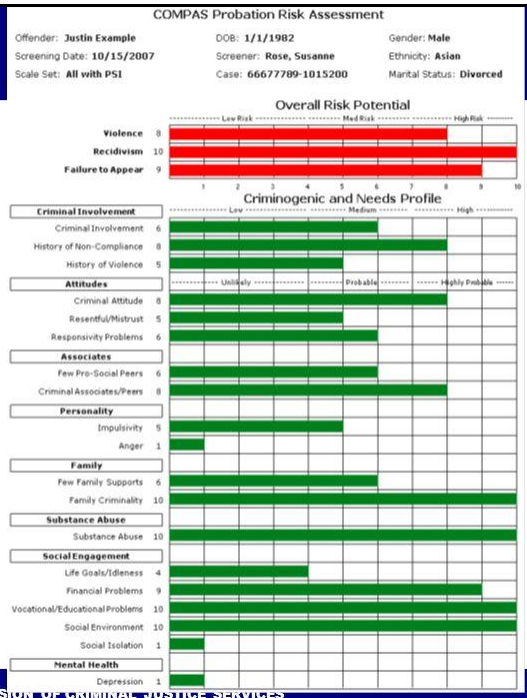
10

BIASES

11

COMPAS

COMPAS Profile



15

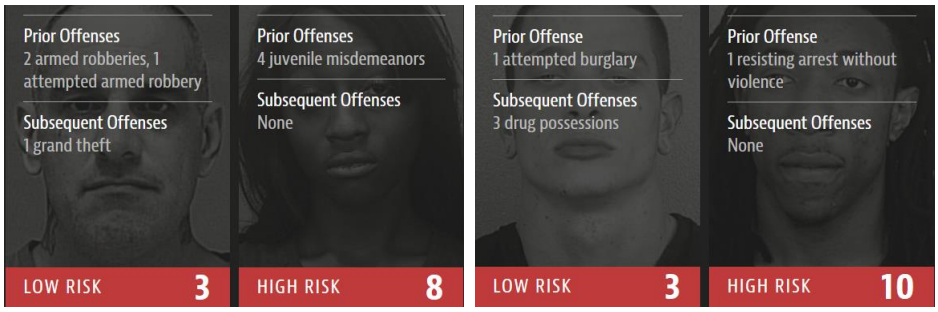
Responsible IR

NEW YORK STATE DIVISION OF CRIMINAL JUSTICE SERVICES
OFFICE OF PROBATION AND CORRECTIONAL ALTERNATIVES

12

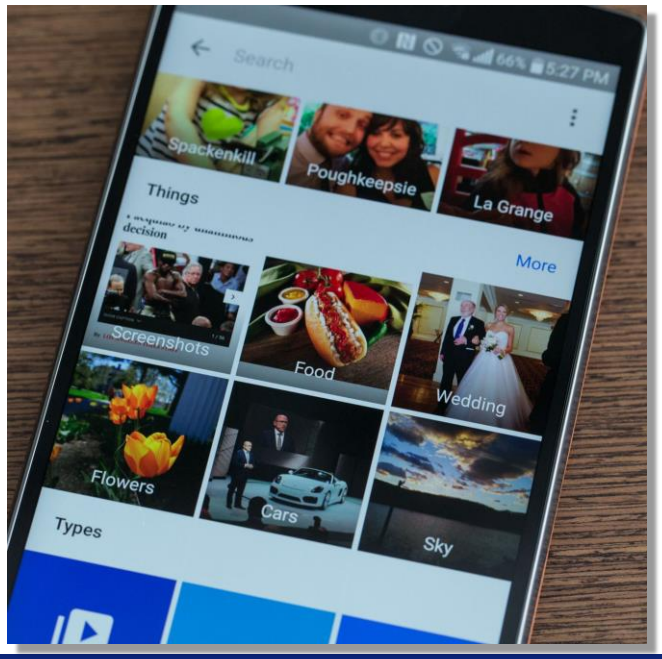
Example – COMPAS: Risk of Recidivism

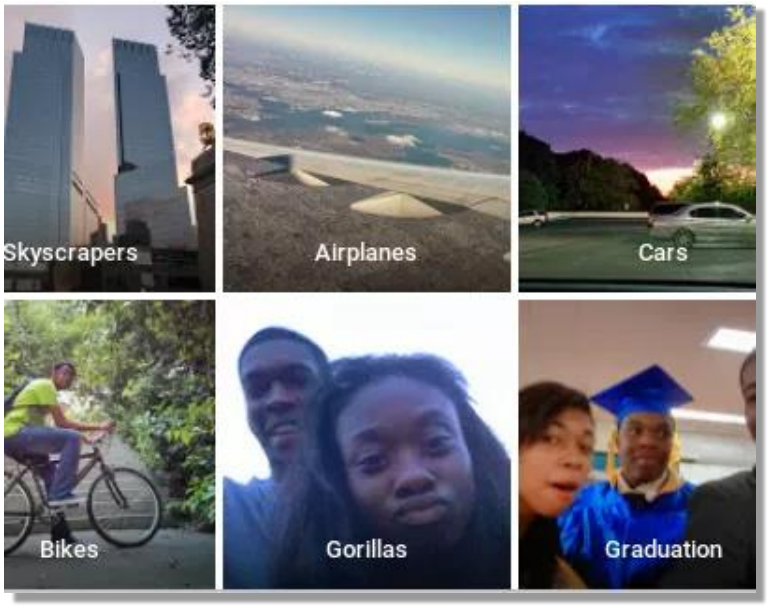
- Labels according to compass directions
- Misclassification difference
- Biased reference data



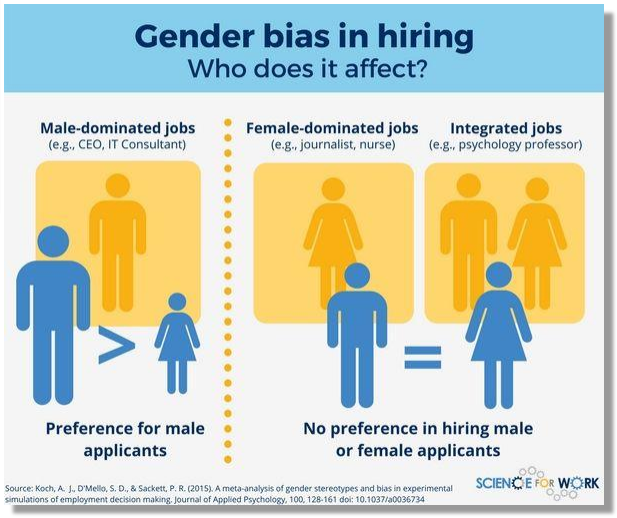
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

ProPublica, 2016




Skyscrapers
Airplanes
Cars
Bikes
Gorillas
Graduation


GENDER- BIASED HIRING TOOL
amazon
NOW DISCARDED


Gender bias in hiring
Who does it affect?
Male-dominated jobs
(e.g., CEO, IT Consultant)
Female-dominated jobs
(e.g., journalist, nurse)
Integrated jobs
(e.g., psychology professor)
>
=
Preference for male applicants
No preference in hiring male or female applicants
Source: Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. Journal of Applied Psychology, 100, 128-161 doi: 10.1037/a0036734
SCIENCE FOR WORK

17

Fairness Definitions

Group fairness
- Equal false positive rate and false negative rate per sensitive group (e.g. ethnic groups)

Individual fairness
- Similar individuals should be treated similarly
- How to define similarity?

Counterfactual fairness
- Attempts to find cause of bias
- What if sensitive attribute was replaced with another value? (gender='male' → gender='female')
  - How would prediction change?

See: https://journals.sagepub.com/doi/10.1177/0049124118782533
https://arxiv.org/abs/1703.09207 (for PDF)
>> Page 1-15

18

Commonly used group fairness metrics

- Demographic parity
  - Equal positive rates
- Equality of opportunity
  - Equal true positive rates
- Equality of odds
  - Equal true positive rates
  - Equal false positive rates

* Figures adapted from blog post by Cortez

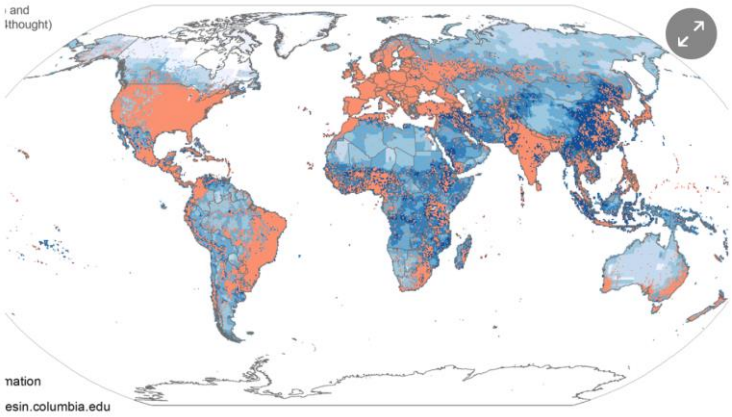
Based on Baeza-Yates

https://dl.acm.org/doi/10.1145/3209581

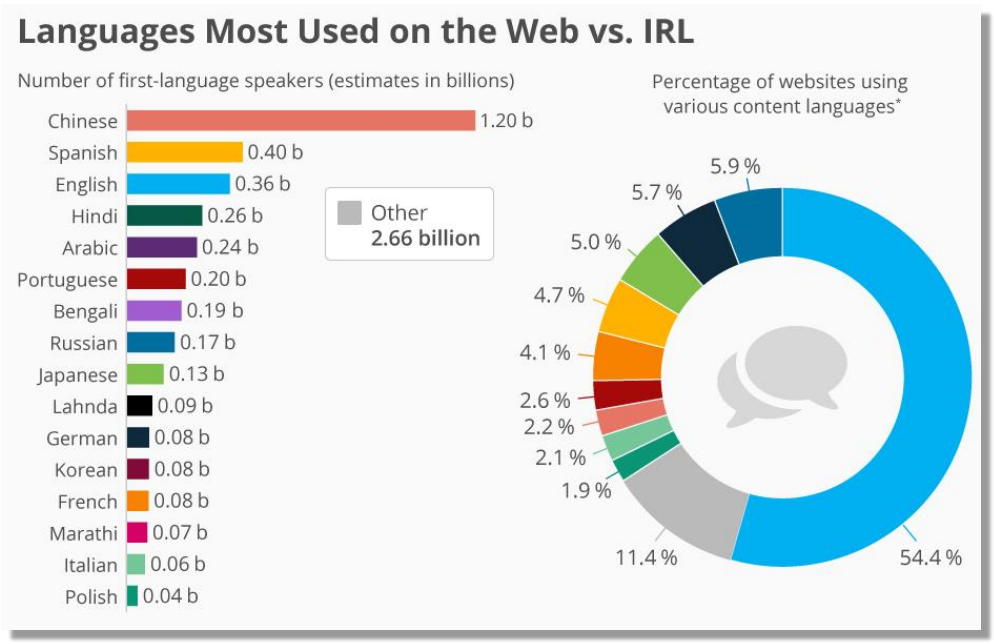
BIAS ON THE WEB

The hidden biases of Geodata

Analysis of one of the world's largest placename databases reveals it is dramatically skewed toward the US's cities, towns and settlements



▲ There were as many place names listed for the US as there were for all of Asia combined. Photograph: Guim



Gender Bias in Content

- Word embedding's in w2vNEWS

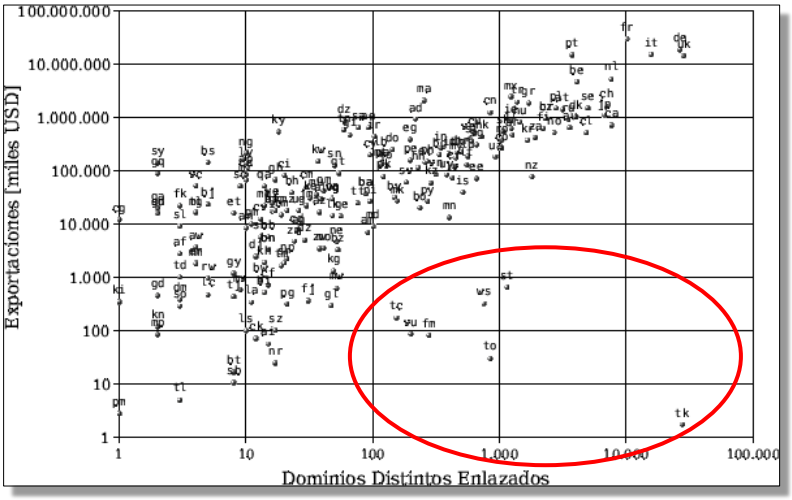
Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

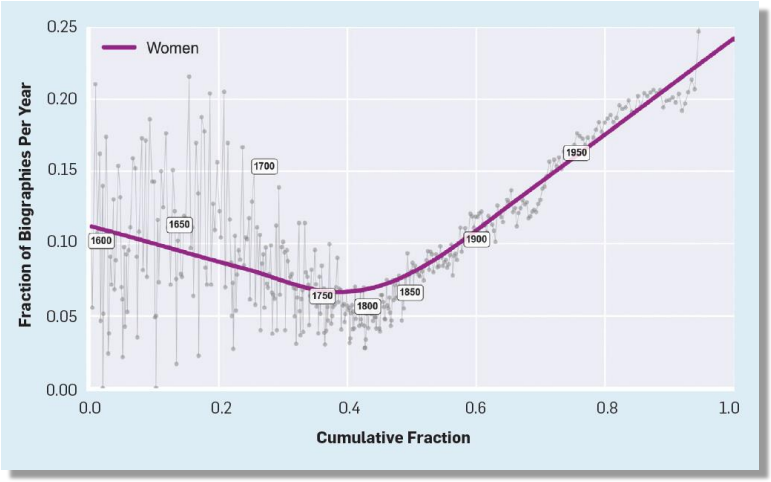
Gender appropriate *she-he* analogies.

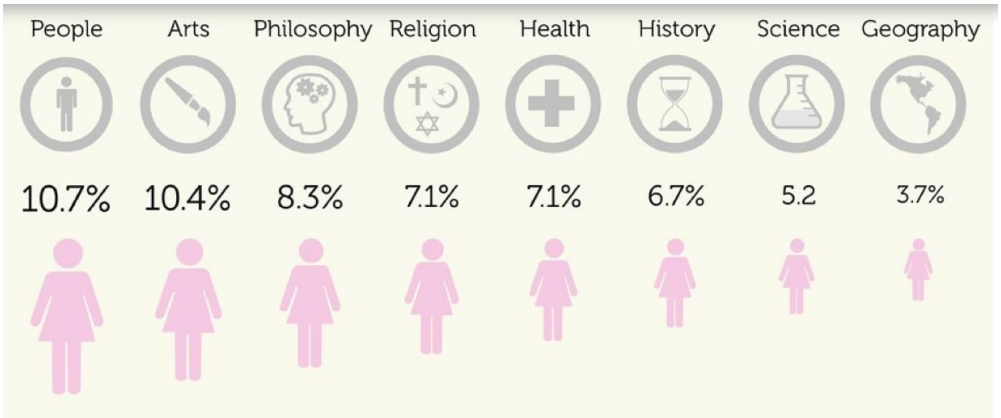
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Economic Bias

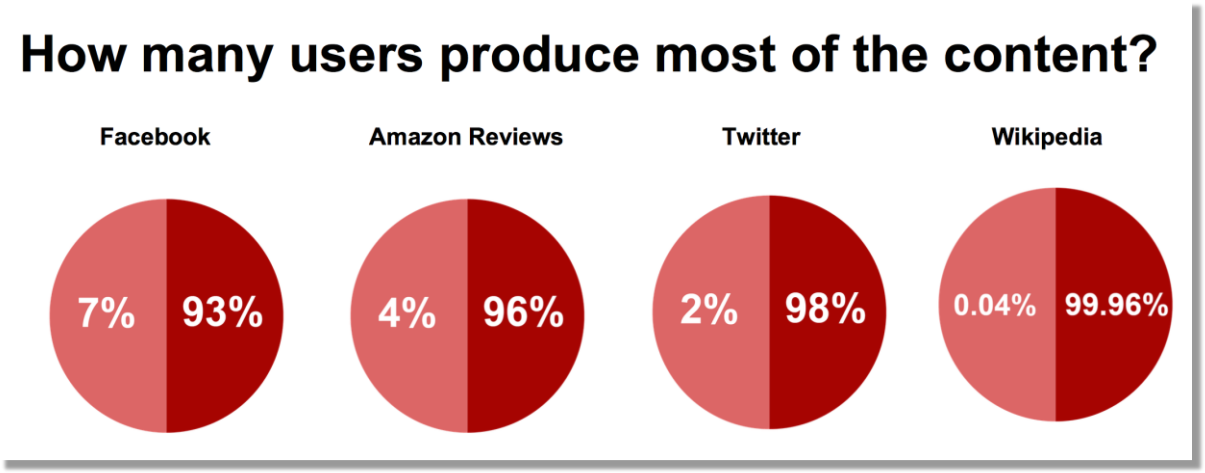


Wikepedia Biographies Gender Bias

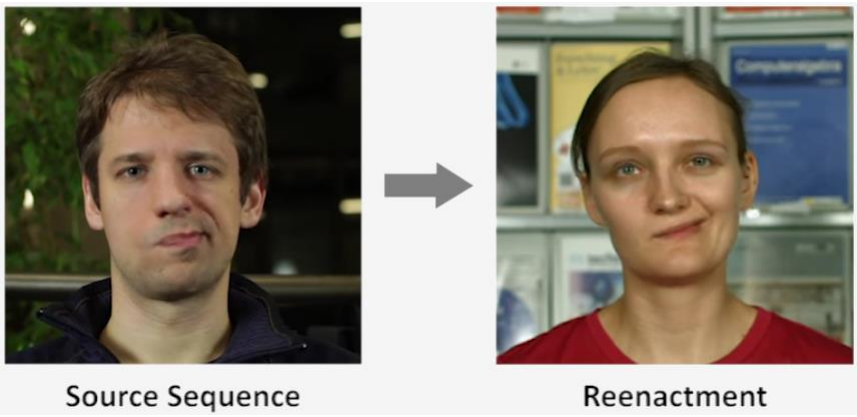




Activity Bias



Deep Fakes



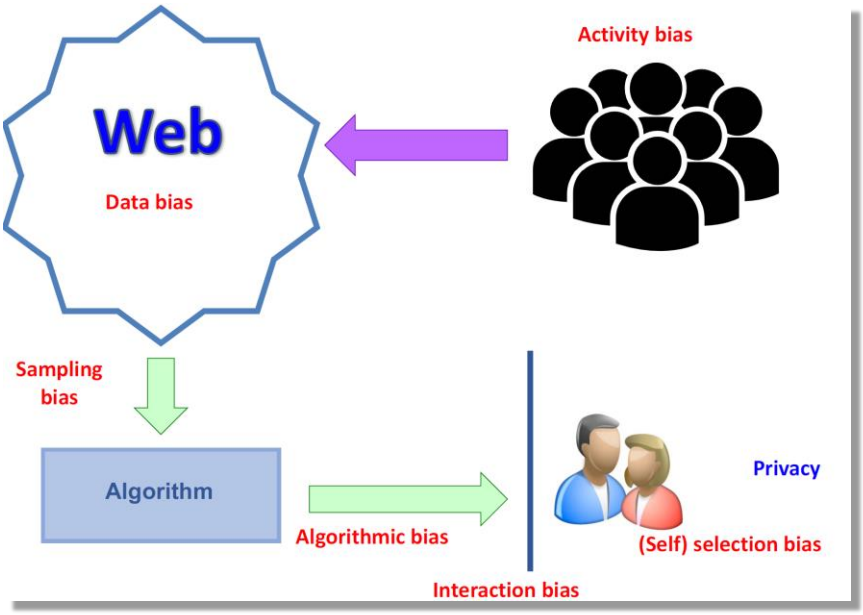
29

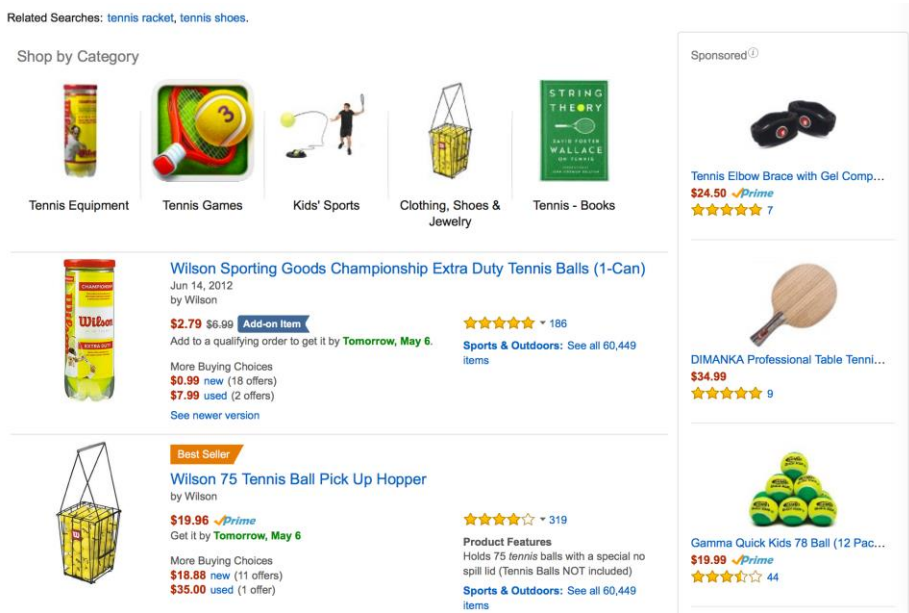
Digital Desert

- 1.1% of the Twitter content is never seen.*
- 31% of articles added/edited in May 2014 in wikipedia, were not visited in June.

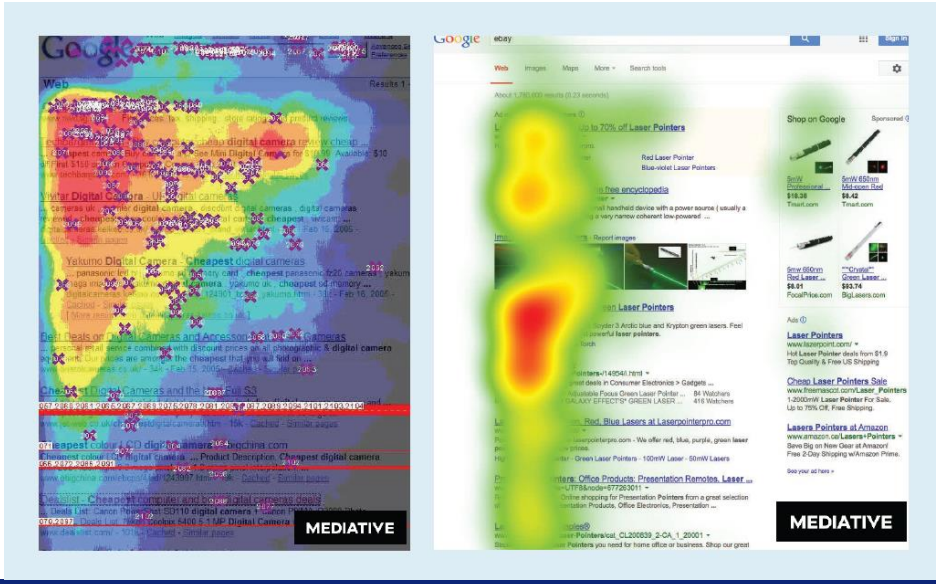
30

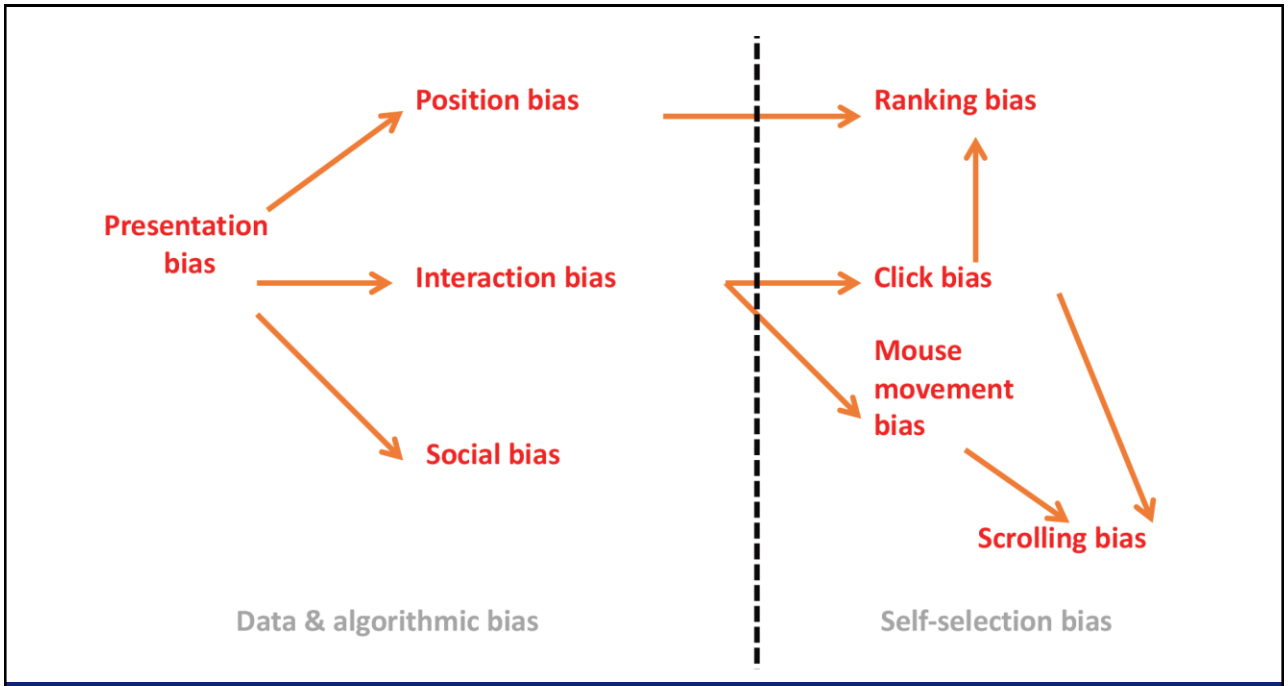
Closed Set of Tags





Ranking Bias in Websearch




Position bias
Ranking bias
Presentation bias
Interaction bias
Click bias
Mouse movement bias
Social bias
Scrolling bias
Data & algorithmic bias
Self-selection bias


Click Throughs by Rank
Navigational queries
Percentage of Clicks
45.00%
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Search Engine Rank

Personalization versus Identification

PRIVACY

BRUCE SCHNEIER SECURITY 12.12.07 09:00 PM

WHY 'ANONYMOUS' DATA SOMETIMES ISN'T

LAST YEAR, NETFLIX published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

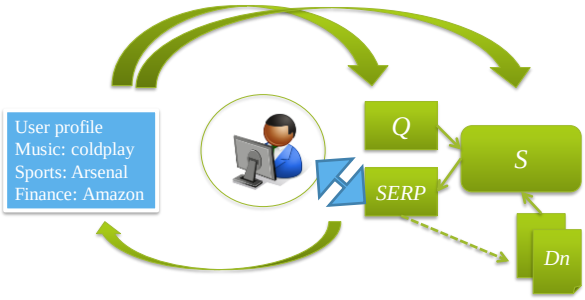
Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some ofthe Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

NETFLIX

IMDb

Privacy versus Personalization in IR

- Departure from 'one size fits all'
- Create personal profile (e.g. interests, past click log)
- Use profile for e.g. re-ranking, query modification suggestion



Acquisition/collection strategies

- Explicit: → Explicit collection
  - the user provides interests explicitly
    - Forms, rating/judging documents (explicit feedback)
- Implicit: [inferred/mined] → Implicit collection
  - the data are derived from the user behaviors (activities) and from external or local context sources
  - "watching over the user's shoulder"

Explicit collection

The user indicates explicitly relevant material via registration form or a questionnaire, …

- Demographic data (name, age, gender, income..)
- Keywords/topics (concepts)
- Example of preferred content
- Feedback documents, feedback of relevance
- Rating items (Netflix)

Explicit personal data collection

Implicit Collection

•Implicit: [inferred/mined] → Implicit feedback

-Software "observes" and collects information from user activity and behavior

•Implicit data

- Demographic data: gender, age can be inferred from
  user interaction and activities, such as user browsing, writing style of texts and queries
- User's query/search history (past queries, visited pages)
- Browsing histories (Urls visited by the user)
- Bookmark
- Desktop information

Implicit personal data collection

PRIVACY MEASURES

Measures for dealing with privacy

- Privacy preserving measures
  - Log deletion: common policy, sometimes user controlled
  - Hashing queries: very effectice
  - Identifier deletion: effective, but does not support law enforcement
  - Hashing identifiers: AOL leak shows this is inadequate
  - Scrubbing query content: does not completely rule out reidentification
  - Deleting infrequent queries: may eliminate identifying data
  - Shortening sessions: highly effective, disables longitudinal user modeling

Confidential/Federated Learning

- AI / Machine Learning from shared data
  - To have more samples for *rare phenomena*: *horizontal partitioning*

e.g. patients with rare diseases

  - To have a *richer representation* of samples: *vertical partitioning*

e.g. combining patient care and insurance data

- Data is confidential
  - Anonimization impossible, data itself is personal: health, mobility
  - Trust
  - GDPR compliance
  - No Trusted Third Party (TTP)