KEDL2019

Algorithmic Bias: What is it, Why Does it Matter and What's Being Done About it?

Paul Clough, University of Sheffield and Peak Indicators

Talk overview

- Introduction to algorithmic bias
- Study of gender biases in search engines
- What can be done?
- Summary

PEAK indicators

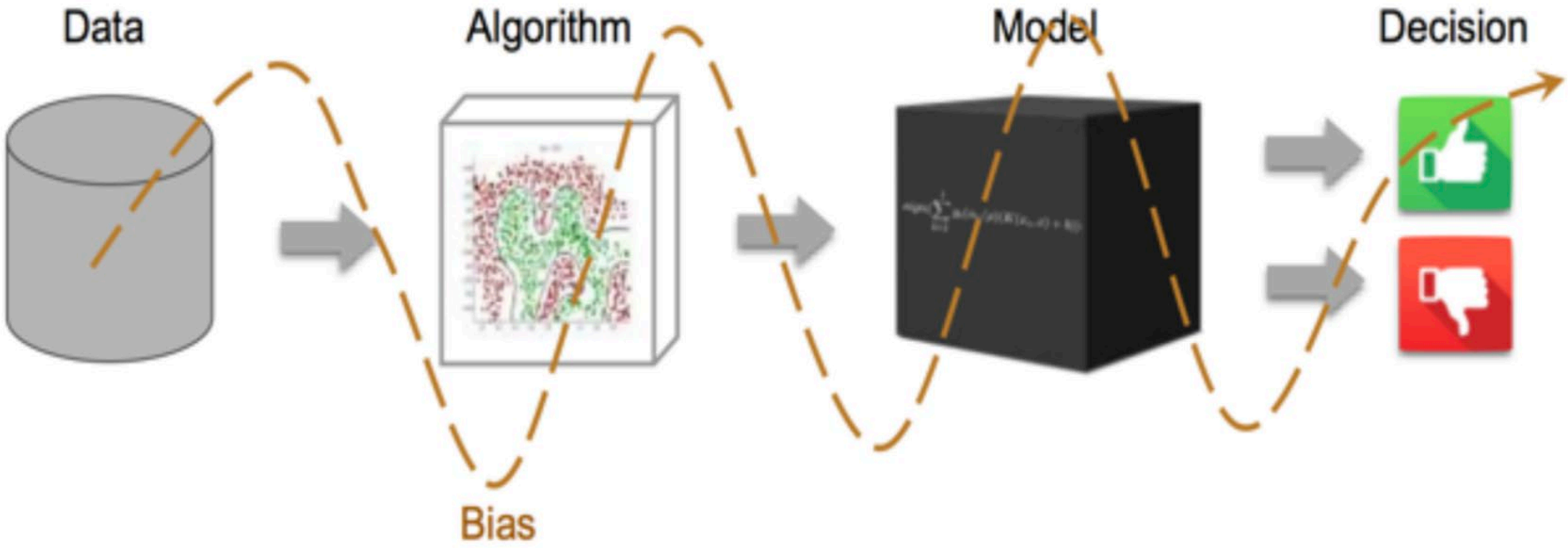
The University Of Sheffield.

Introduction

Data-driven decision making

- Algorithms can support decision-making by
  - Prioritising and ranking
  - Making predictions (regression and classification)
  - Finding patterns and associations
  - Filtering
- Predictive models used in
  - Personalised pricing and recommendations
  - Credit scoring
  - Automated CV screening of job applicants
  - Profiling of potential suspects by the police
  - …

PEAK indicators



S. Hajian, F. Bonchi and C. Castillo (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In KDD, pp. 2125-2126.

"Bias: inclination or prejudice for or against one person or group, especially in a way considered to be unfair"

"Predictive models can discriminate people, even if the computing process is fair and well-intentioned"

PEAK indicators

Gender-recognition AI tools correctly identify white men more accurately than BAME women.

Google's online advertising system showed high-income jobs to men much more often than to women.

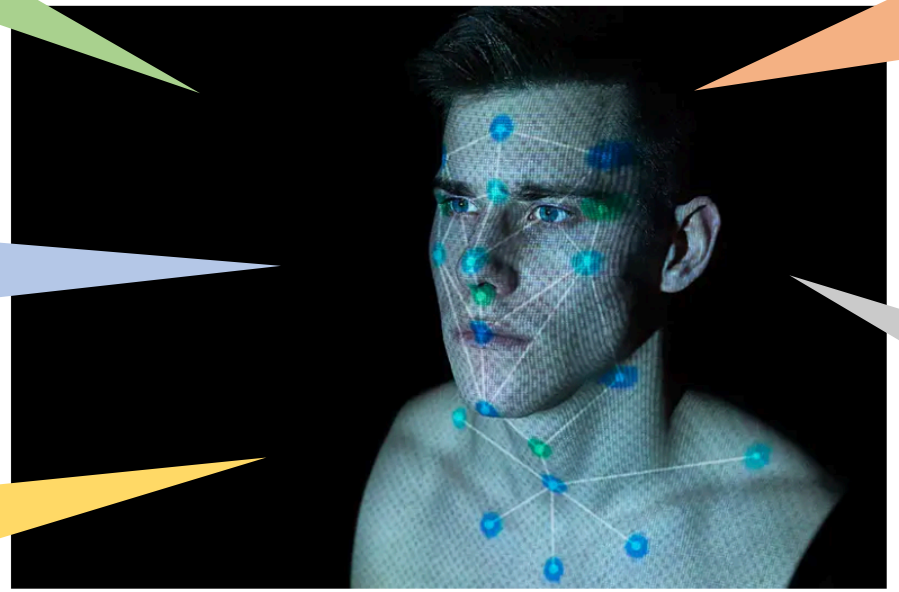
Facebook's automatic translation software chose the wrong translation for Hebrew "good morning" vs. "attack them".

NewScientist

Discriminating algorithms: 5 times AI showed prejudice

Artificial intelligence is supposed to make life easier for us all – but it is also prone to amplify sexist and racist biases from the real world

TECHNOLOGY 12 April 2018 , updated 27 April 2018



John Lamb/Getty

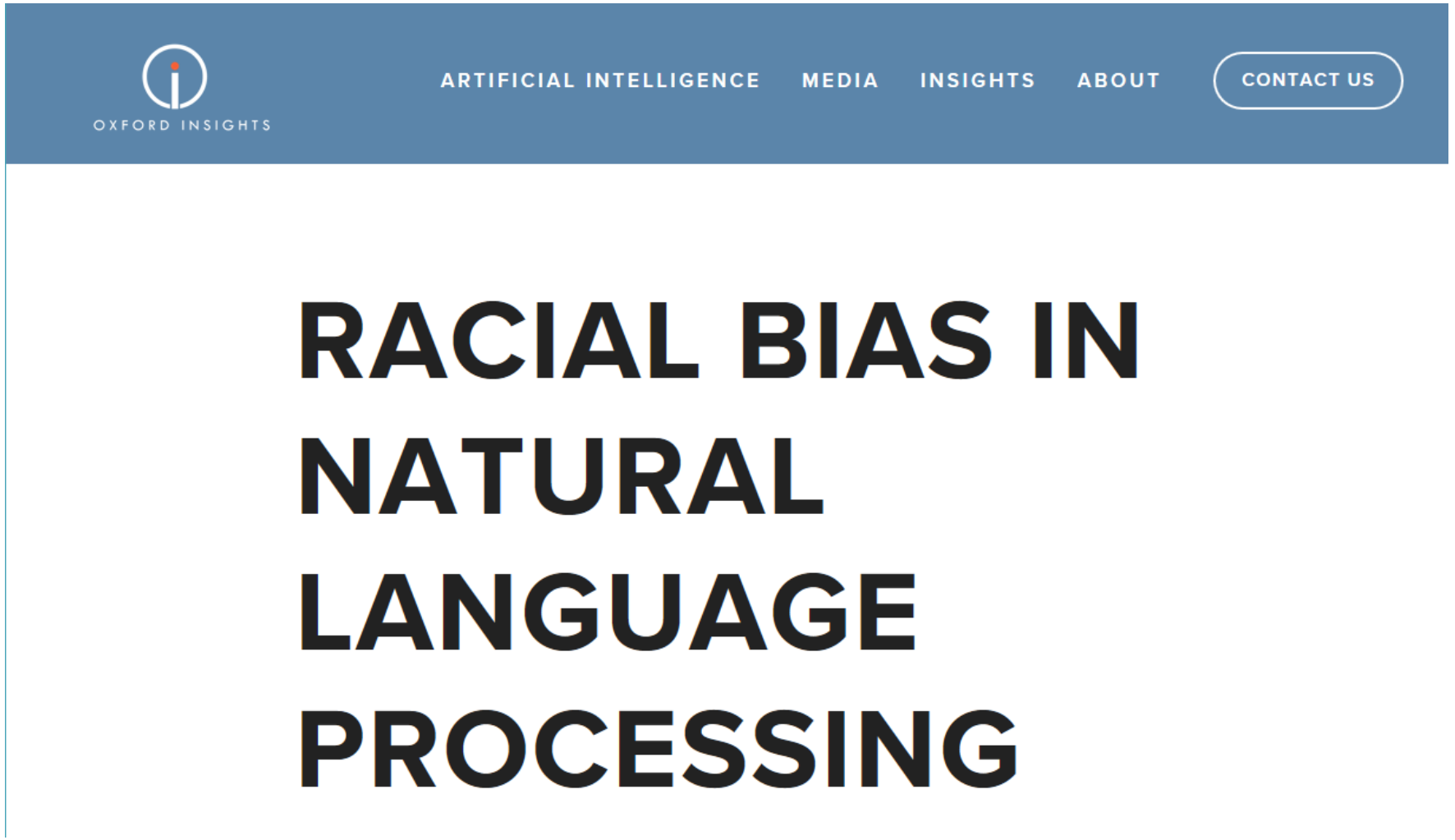
By Daniel Cossins

Modern life runs on intelligent algorithms. The data-devouring, self-improving computer programmes that underlie the artificial intelligence revolution already determine Google search results, Facebook news feeds and online shopping recommendations. Increasingly, they also decide how easily we get a mortgage or a job interview, the chances we will get stopped and searched by the police on our way home, and what penalties we face if we commit a crime, too.

Researchers found that COMPAS predicts that black defendants pose a higher risk of recidivism than they do, and the reverse for white defendants.

In 2016, the Human Rights Data Analysis Group found that PredPol could lead police to unfairly target certain neighbourhoods.

PEAK indicators



https://www.oxfordinsights.com/racial-bias-in-natural-language-processing

PEAK indicators

MATTHEW REIDSMA

ARTICLES TALKS WORK NOTES

ALGORITHMIC BIAS IN LIBRARY DISCOVERY SYSTEMS

March 11, 2016 « Prev

More and more academic libraries have invested in discovery layers, the centralized "Google-like" search tool that returns results from different services and providers by searching a centralized index. The move to discovery has been driven by the ascendence of Google as well as libraries' increasing focus on user experience. Unlike the vendor-specific search tools or federated searches of the previous decade, discovery presents a simplified picture of the library research process. It has the familiar single search box, and the results are not broken out by provider or format but are all shown together in a list, aping the Google model for search results.

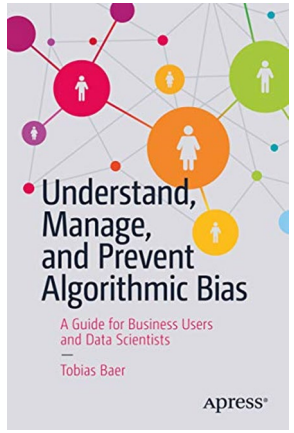
Discovery's promise of a simple search experience works for users, more often than not. But discovery's external simplicity hides a complex system running in the background, making decisions for our users. And it is the rare user that questions these decisions. As Sherry Turkle (1997) observed, users approach complex systems

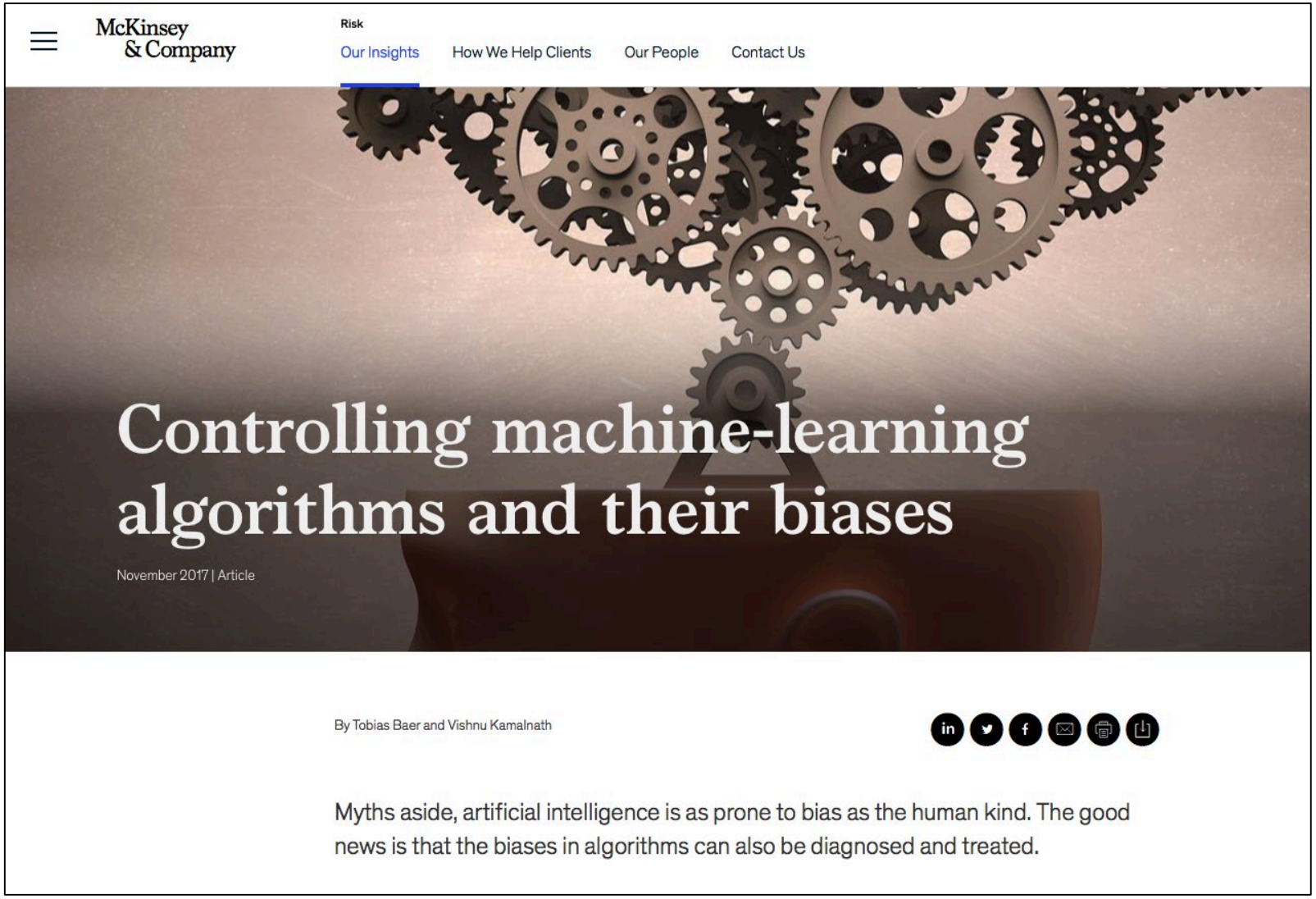
https://matthew.reidsrow.com/articles/173

PEAK indicators

It's not all bad news

"It is a myth to think that algorithms are objective, but also a myth to think that human processes are not subject to biases on par with algorithms."





A study of gender bias in search engines

Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results

Jahna Otterbacher
Social Information Systems
Open University of Cyprus
jahna.otterbacher@ouc.ac.cy

Jo Bates
Information School
University of Sheffield, UK
jo.bates@sheffield.ac.uk

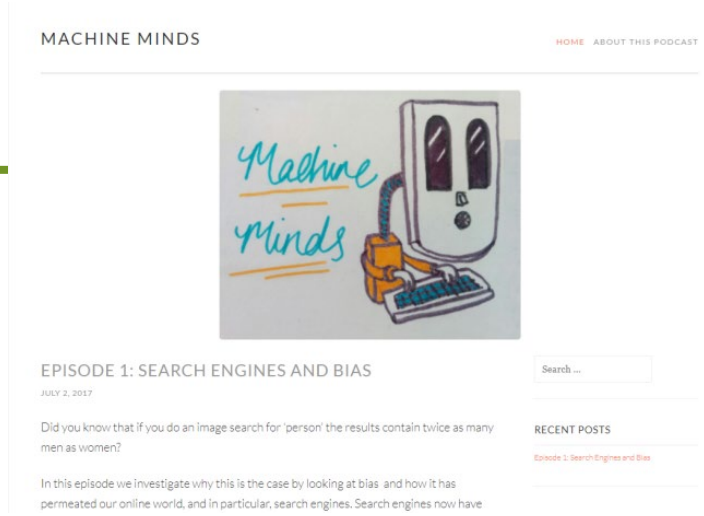
Paul Clough
Information School
University of Sheffield, UK
p.d.clough@sheffield.ac.uk

ABSTRACT
There is much concern about algorithms that underlie information services and the view of the world they present. We develop a novel method for examining the content and strength of gender stereotypes in image search, inspired by the trait adjective checklist method. We compare the gender distribution in photos retrieved by Bing for the query "person" and for queries based on 68 character traits (e.g., "intelligent person") in four regional markets. Photos of men are more often retrieved for "person," as compared to women. As predicted, photos of women are more often retrieved for *warm* traits (e.g., "emotional") whereas *agentic* traits (e.g., "rational") are represented by photos of men. A backlash effect, where stereotype-incongruent individuals are penalized, is observed. However, backlash is more prevalent for "competent women" than "warm men." Results underline the need to understand how and why biases enter search algorithms and at which stages of the engineering process.

Author Keywords
Algorithmic bias; "Big Two" dimensions of social perception; gender stereotypes; image search.

participation in public life [20]. Even when users are intimately familiar with a system, they are often unaware that algorithms filter their access to information [14] and users hold beliefs about algorithms, which, true or not, influence how they use systems [39].

Machines running algorithmic processes have become the new gatekeepers, largely determining what and whom we see, and do not see [8]. Given the power that algorithms exert, researchers must scrutinize these processes, their potential biases and social impact; in other words, we must work toward "algorithmic accountability" [11] and "algorithmic transparency" [10].



http://www.machinemindspodcast.com/

Otterbacher, J., Bates, J., and Clough P. (2017), Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results, In Proceedings of CHI'2017, pp. 6620-6631.

PEAK indicators

FUTURE TENSE | THE CITIZEN'S GUIDE TO THE FUTURE. | DEC. 7 2015 11:48 AM

FROM SLATE, NEW AMERICA, AND ASU

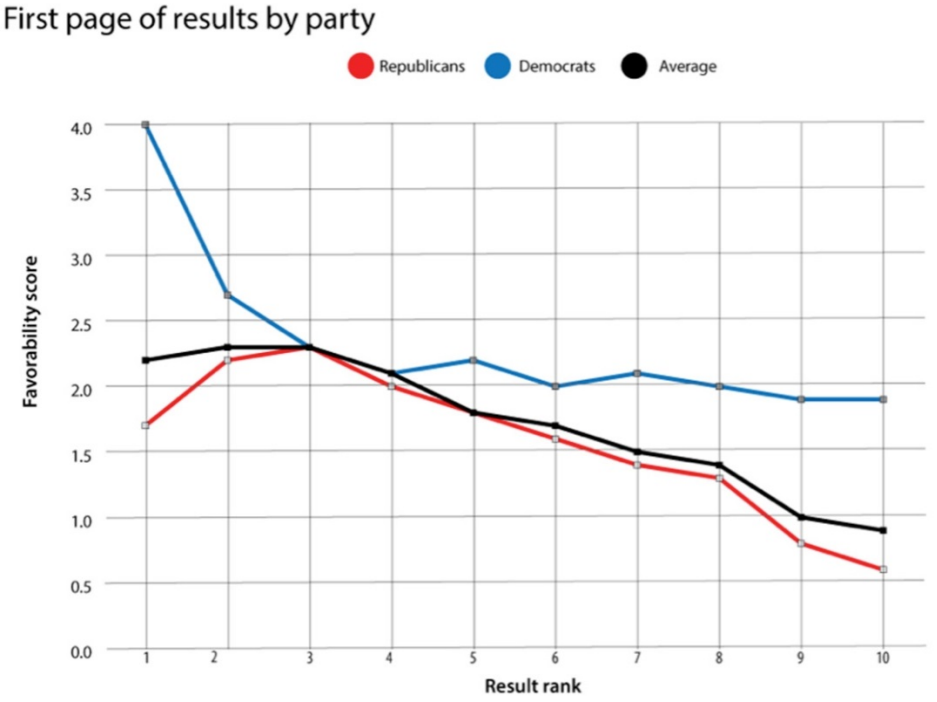
Why Google Search Results Favor Democrats

It's not because the company is biased—it's more complicated.

409 20

By Daniel Trielli, Sean Mussenden, and Nicholas Diakopoulos

http://www.slate.com/articles/technology/future_tense/2015/12/why_google_search_results_favor_democrats.html



"And there's the illusion of neutrality. About two-thirds of Americans who use search engines believe they are completely unbiased, according to a 2012 Pew study. The study showed that "73 percent of search engine users say that most or all the information they find as they use search engines is accurate and trustworthy." Search engines are more trusted than the news media itself."

Who is a nurse?

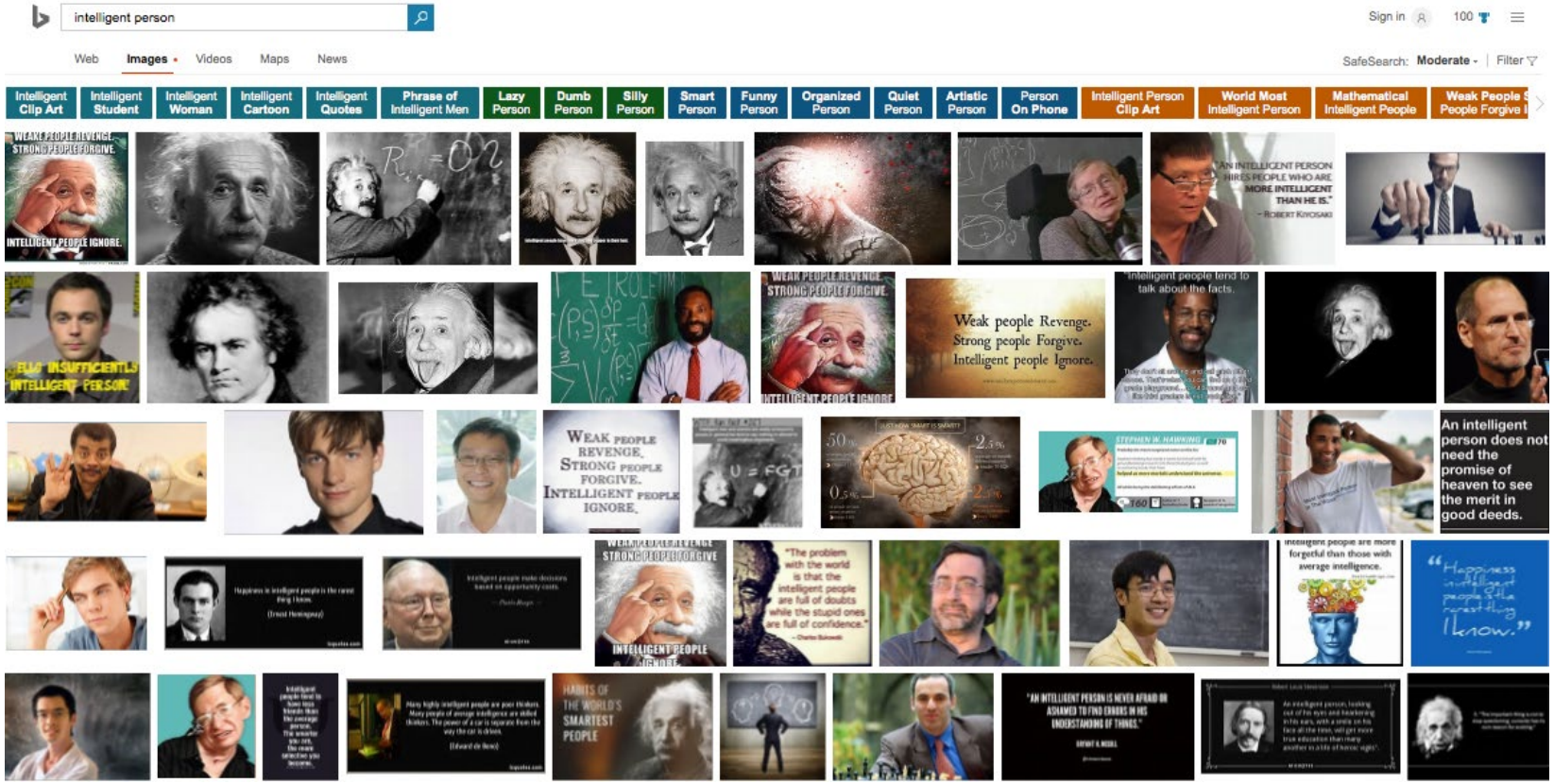
Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 3819-3828.

Who is a nurse?

PEAK indicators

Male nurse

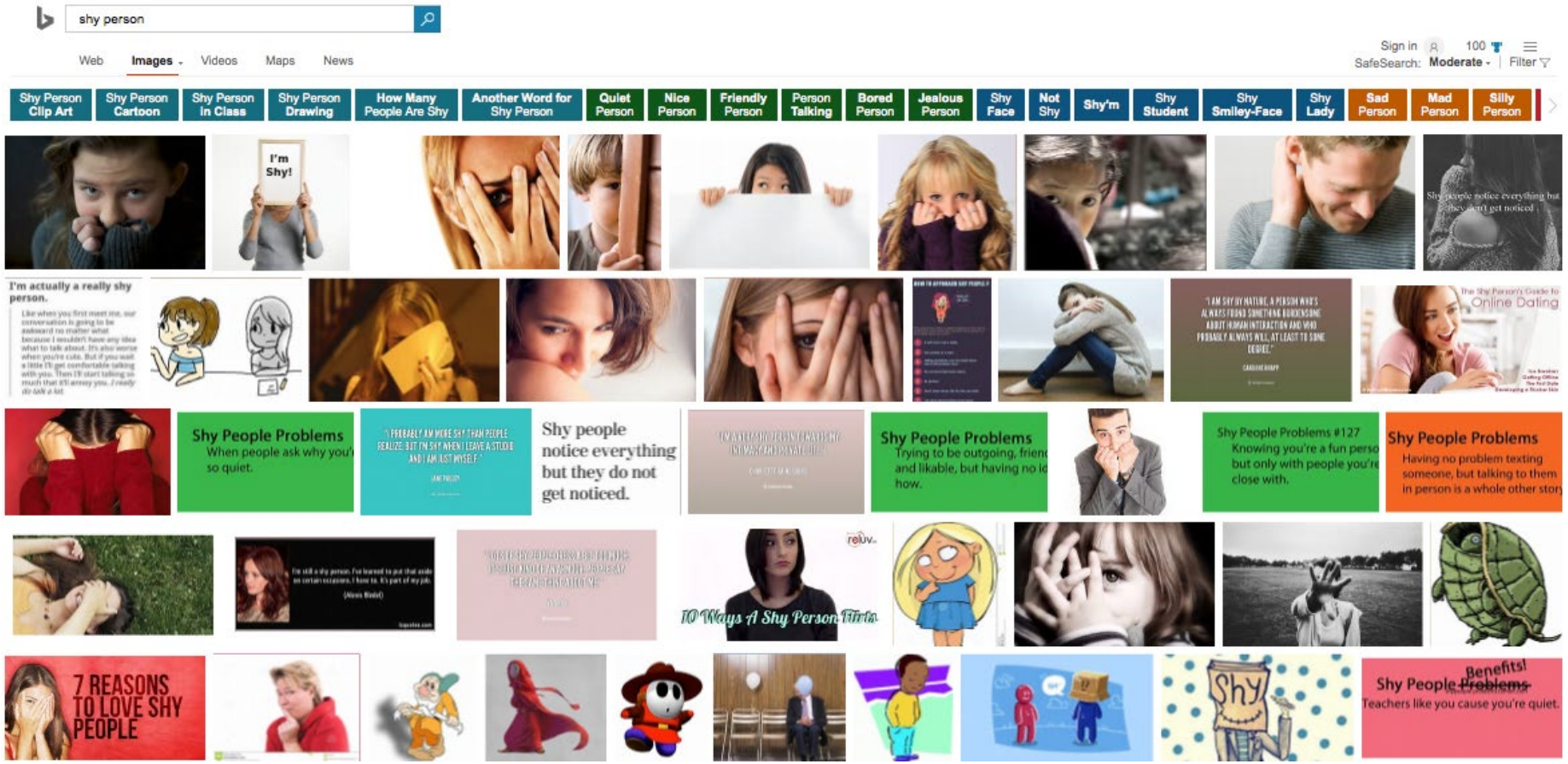
Intelligent person



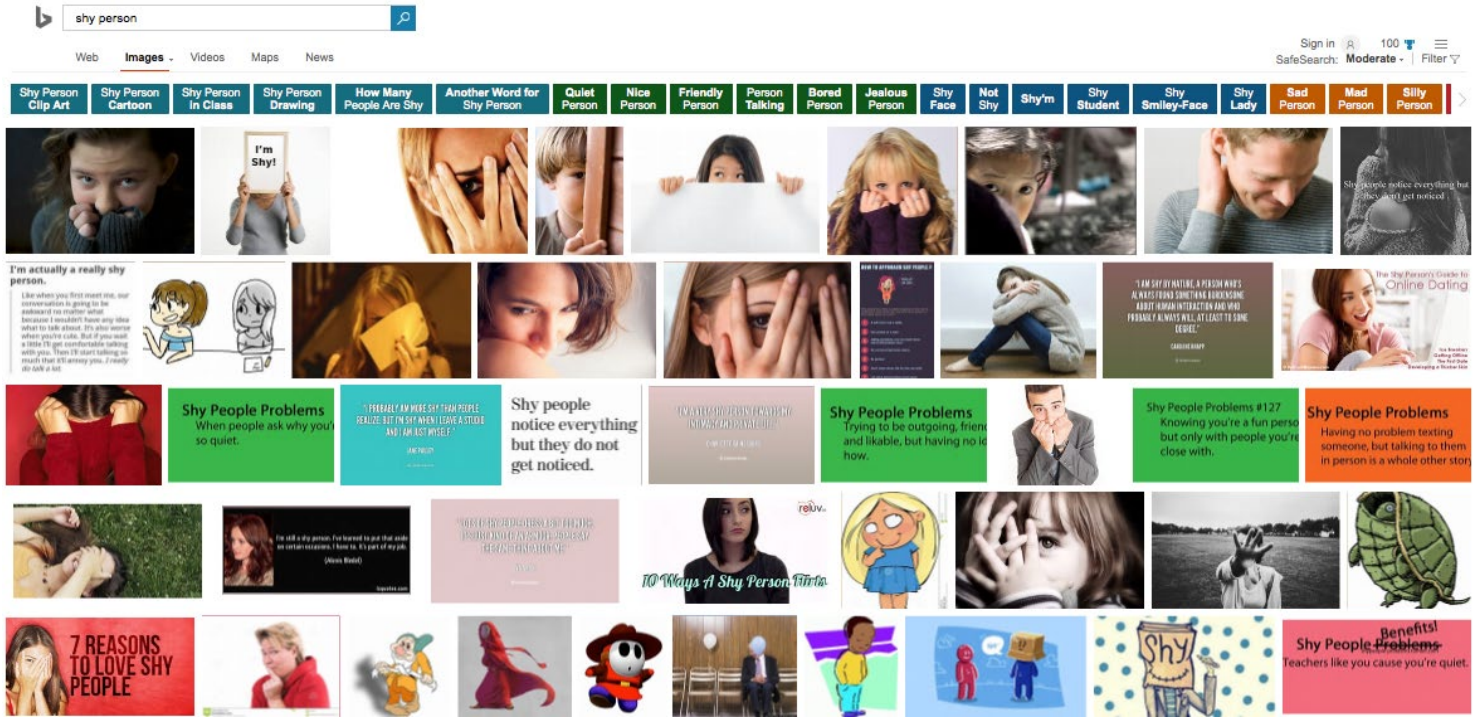
Stereotypes beyond occupation – personality traits
Who does Bing say represents a 'person'?

PEAK indicators

Shy person



Shy person



Gender distribution in images of top-ranked 50 images

Women/girls:	**25 (50%)**
Men/boys:	**5 (10%)**
Mixed gender:	**0**
Unknown/none:	**20 (40%)**

Can we automatically identify gender distribution in results?

Stereotypes: "Big Two" of person perception

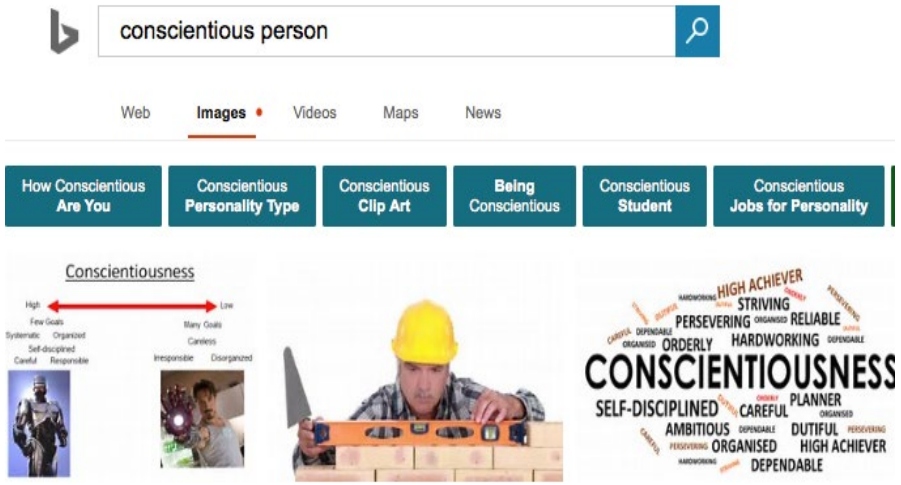
- Personality traits captured by the 'Big five'
- Our perceptions of others are based on two dimensions [Fiske et al., 2002]
  1) agency (or competence): whether or not we perceive someone as being capable of achieving his/her goals
  2) warmth (or communality): whether or not we think someone has pro-social intentions or is a threat to us
- Stereotypes are captured by combinations of the two dimensions [Cuddy et al., 2008]
  - Women: [low agency, high warmth]
  - Men: [high agency, low warmth]

Trait adjective checklist method

- How do we measure content and strength of given social stereotype?
  - Trait adjective checklist method
- Used in the Princeton Trilogy studies of ethnic and racial stereotypes [Katz & Braly, 1933]
- Participants describe target social groups using list of trait adjectives
- 68 traits developed in cross-lingual study across five countries [Abele et al., 2008]

PEAK indicators

able	egoistic	persistent
active	emotional	polite
affectionate	energetic	rational
altruistic	expressive	reliable
ambitious	fair	reserved
assertive	friendly	self-confident
boastful	gullible	self-critical
capable	harmonious	self-reliant
caring	hardhearted	self-sacrificing
chaotic	helpful	sensitive
communicative	honest	shy
competent	independent	sociable
competitive	industrious	striving
conceited	insecure	strong-minded
conscientious	intelligent	supportive
considerate	lazy	sympathetic
consistent	loyal	tolerant
creative	moral	trustworthy
decisive	obstinate	understanding
detached	open	vigorous
determined	open-minded	vulnerable
dogmatic	outgoing	warm
dominant	perfectionistic	

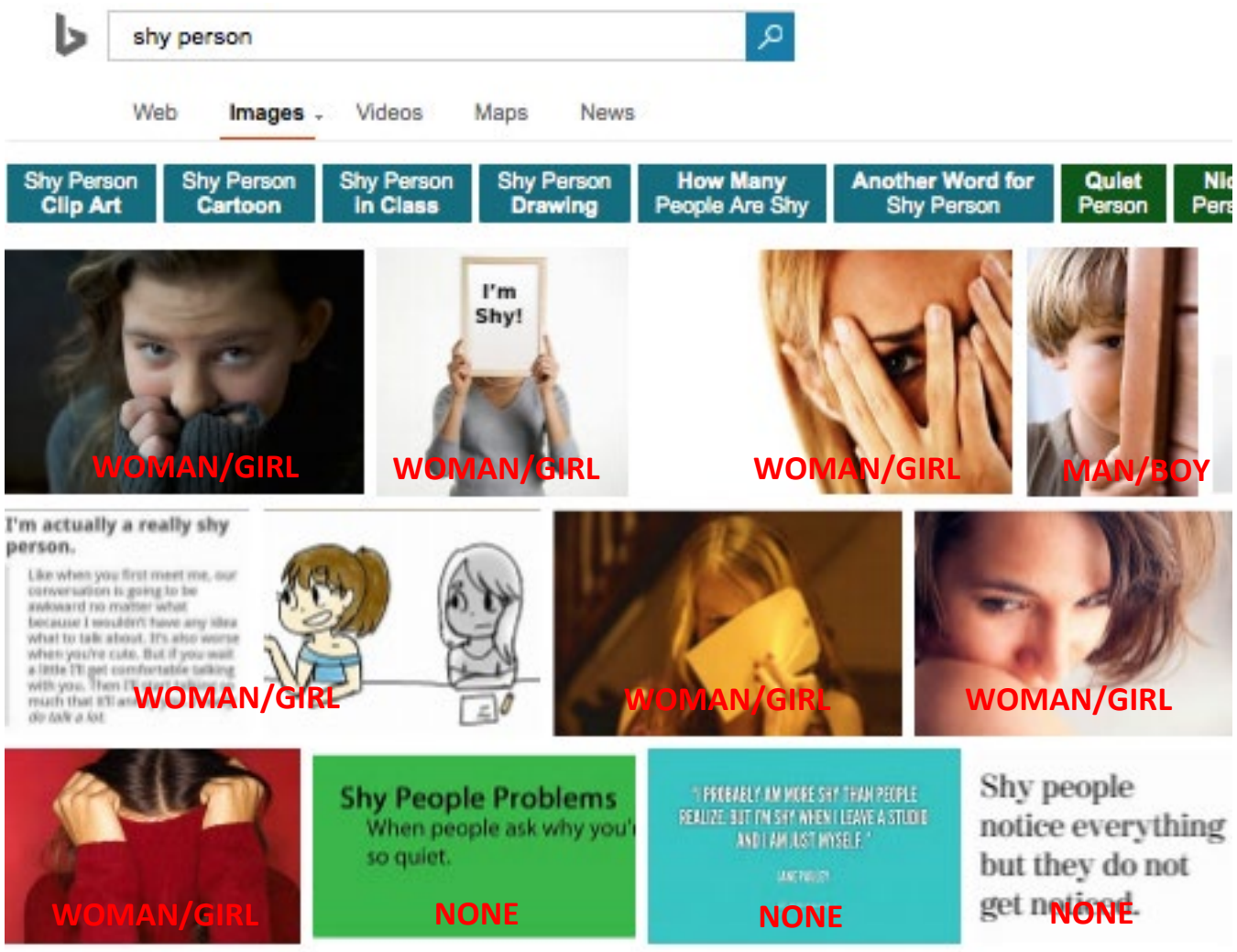


Search markets:
UK-EN
US-EN
IN-EN
ZA-EN

Research Questions

- RQ1: Baseline Representation bias
  - In a search for “person” which genders are depicted?
- RQ2: Stereotype content and strength
  - Which character traits are most often associated with which genders?
  - Are these associations consistent across bing search markets? (UK, US, IN, ZA)
- *RQ3: Backlash effects*
  - *How are stereotype-incongruent individuals depicted?*

PEAK indicators



Challenges in automating the process – how hard is recognising gender in images for people?

Pilot study on Crowdflower

- 1,000 “person” images from UK market
- 3 annotators per image
- Is the image:

  1) a photograph, 2) a sketch/illustration, 3) some other type?
- Does the image depict:

  1) only women/girls, 2) only men/boys, 3) mixed gender group,
  4) gender ambiguous person(s), 5) no person(s)?

Classifying image type

	# Images	Inter-judge agreement
Photos	576	0.97
Sketches	346	0.96
Other	22	0.74
No longer accessible	56	1.00

High degree of agreement

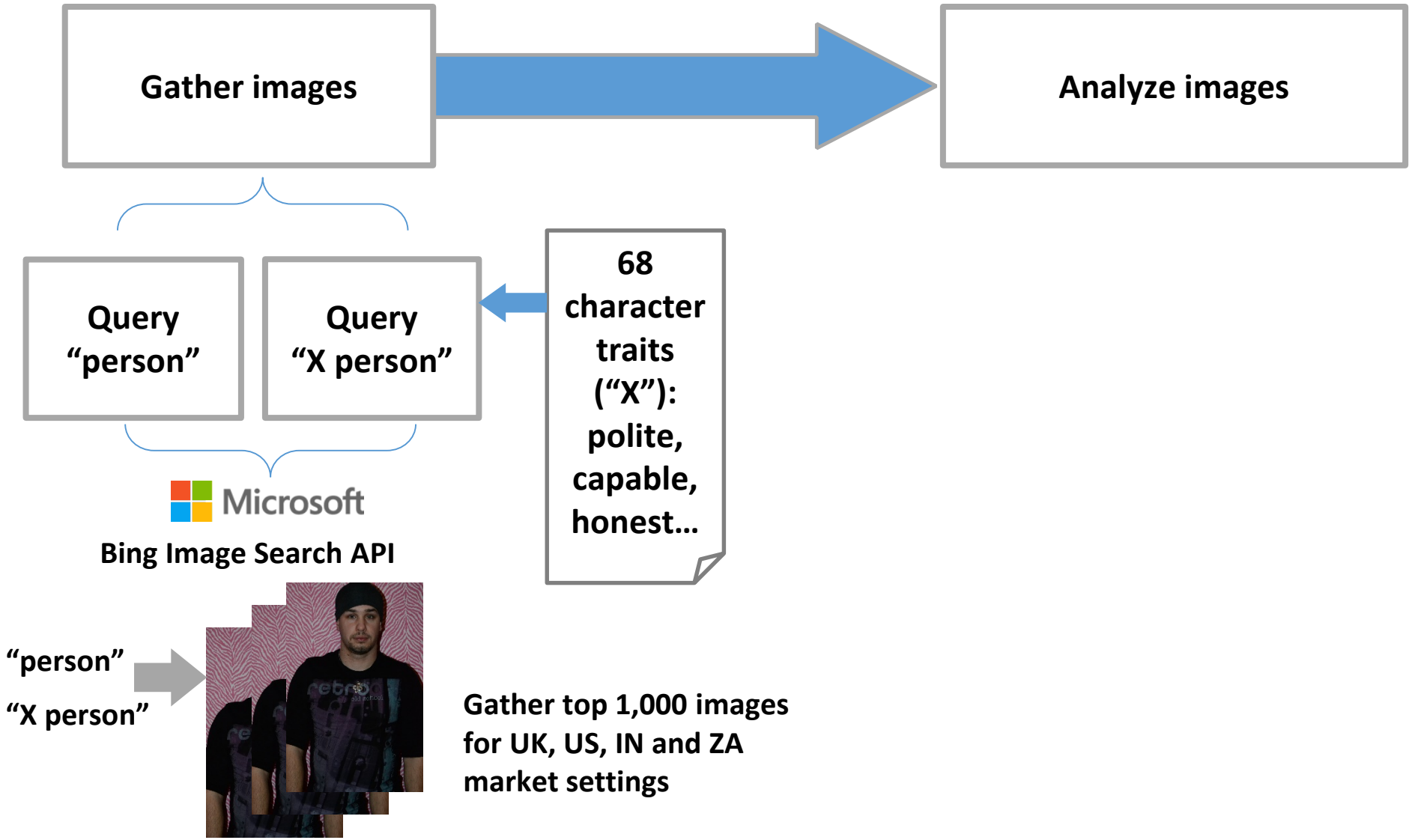
Classifying gender

	Women /girls	Men/boys	Mixed gender	Unknown	No persons	Inter-judge agreement
Photos	0.27	0.55	0.10	0.07	0.01	0.94
Sketches	0.08	0.28	0.05	0.55	0.04	0.91

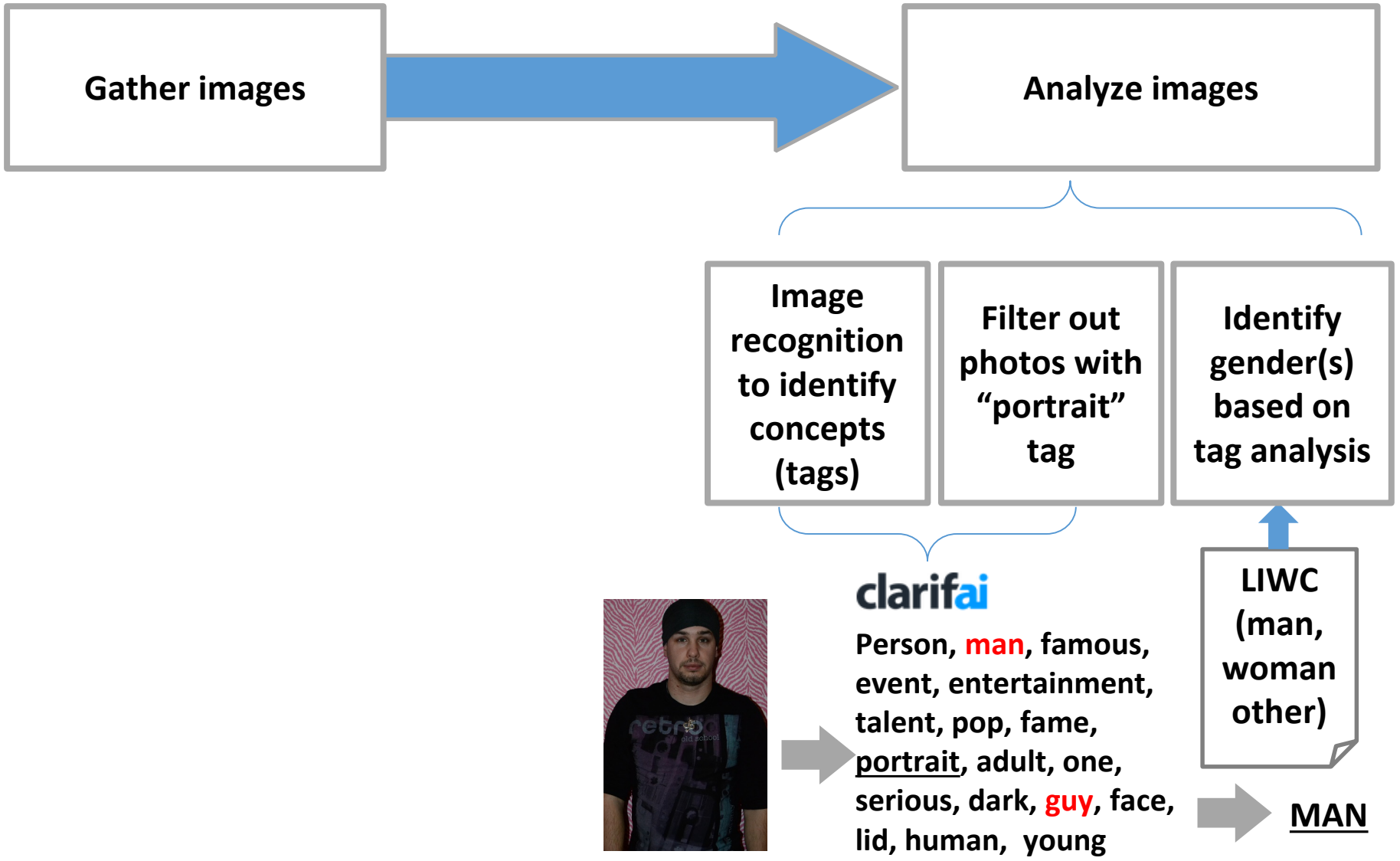
Automating gender recognition

- Clarify API
  - General image recognition tool
  - Coverage: 95% of images from bing
  - Provides 20 textual concept tags
- Linguistic Inquiry and Wordcount (LIWC) [Pennebaker et al., 2015]
  - Female references: mom, girl
  - Male references: dad, boy

PEAK indicators



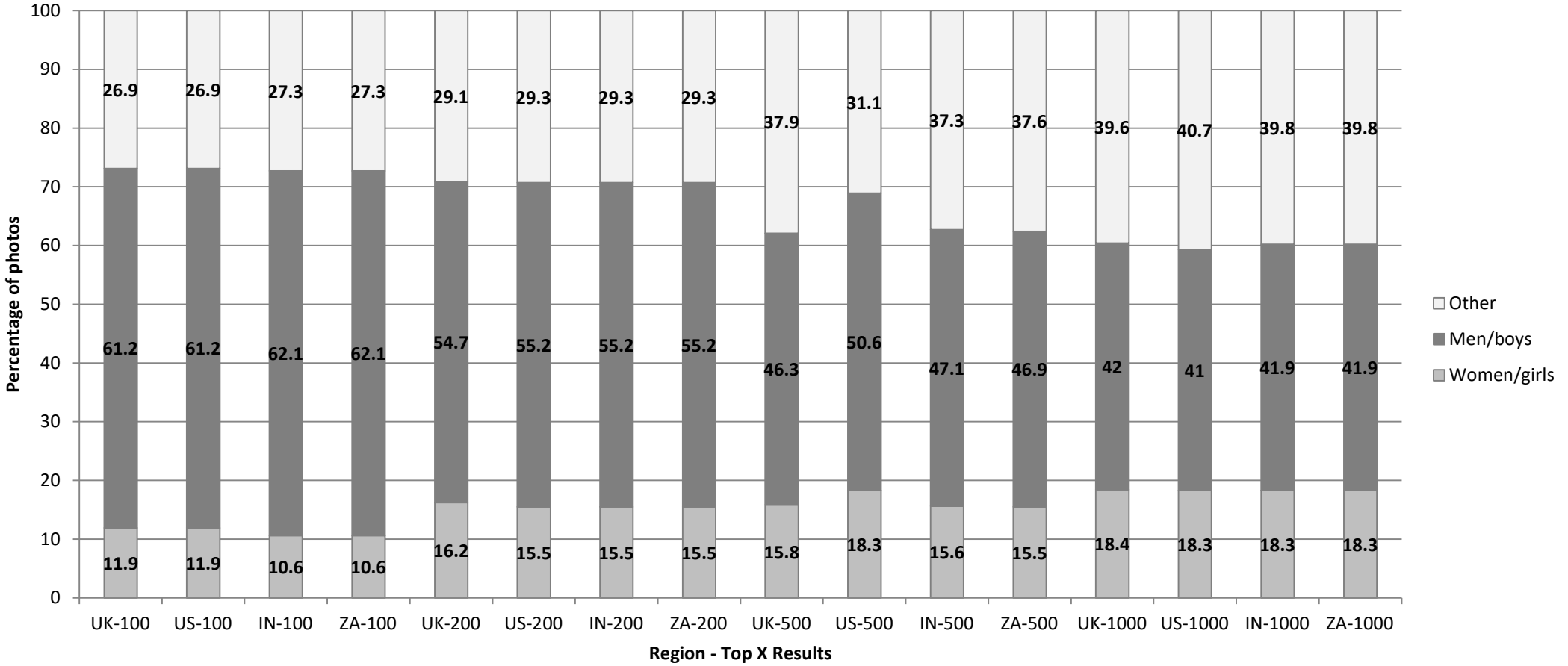
PEAK indicators



Performance on gender classification

	N	Precision	Recall	F_1
Recognizing photographs	473	0.91	0.75	0.822
Women/girls	130	0.89	0.60	0.717
Men/boys	282	0.95	0.67	0.786
Other	61	0.68	0.82	0.743

RQ1: who represents a “person”?



Across all results, about 42% depict men, 18% depict women, and 40% others

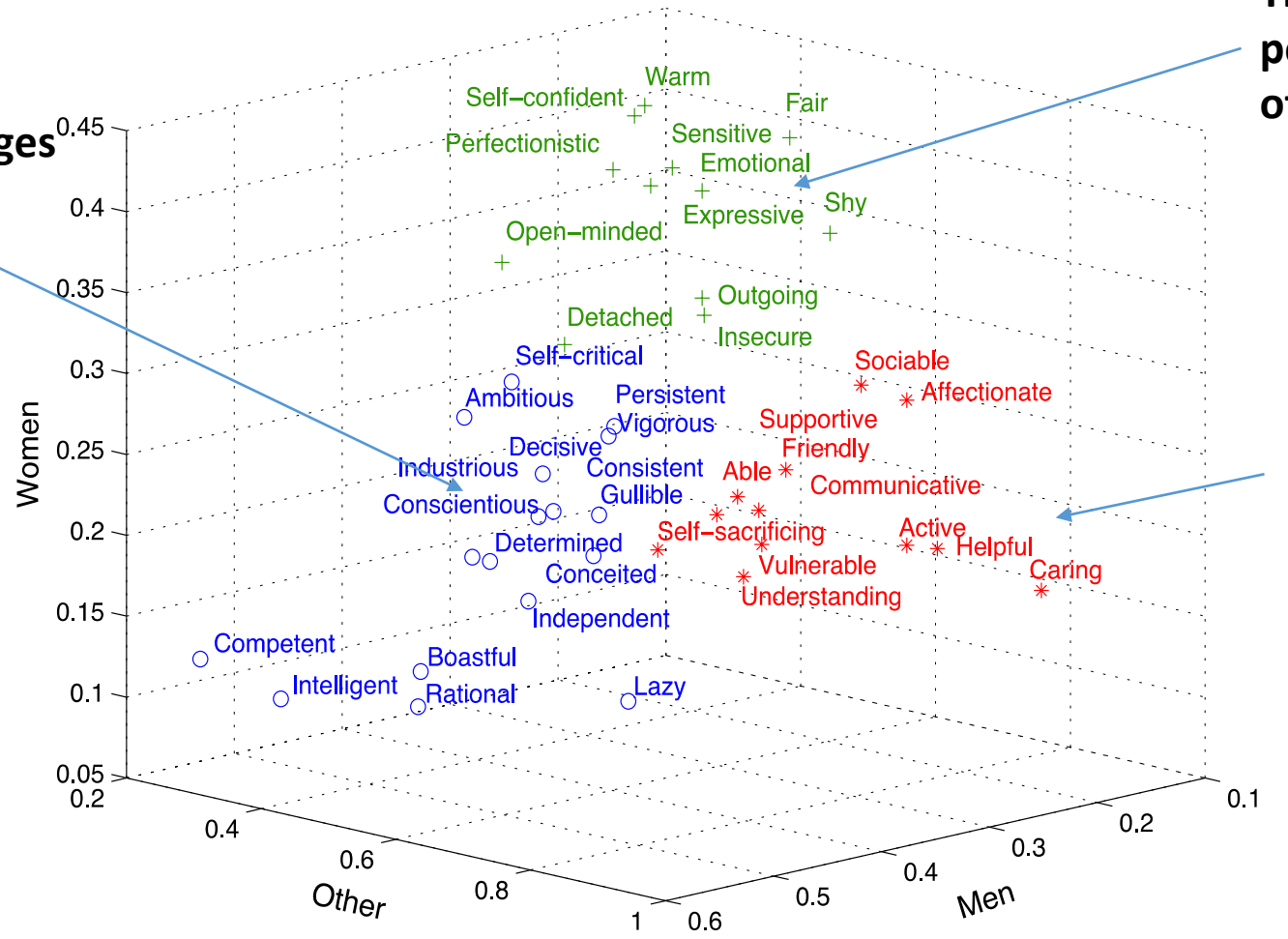
In top 100 around 61% of person photos depicted men, and only around 11% depict women

RQ2: which traits are gendered?

Traits most often portrayed for images of men

Traits most often portrayed for images of women

Traits most often portrayed for gender-neutral images



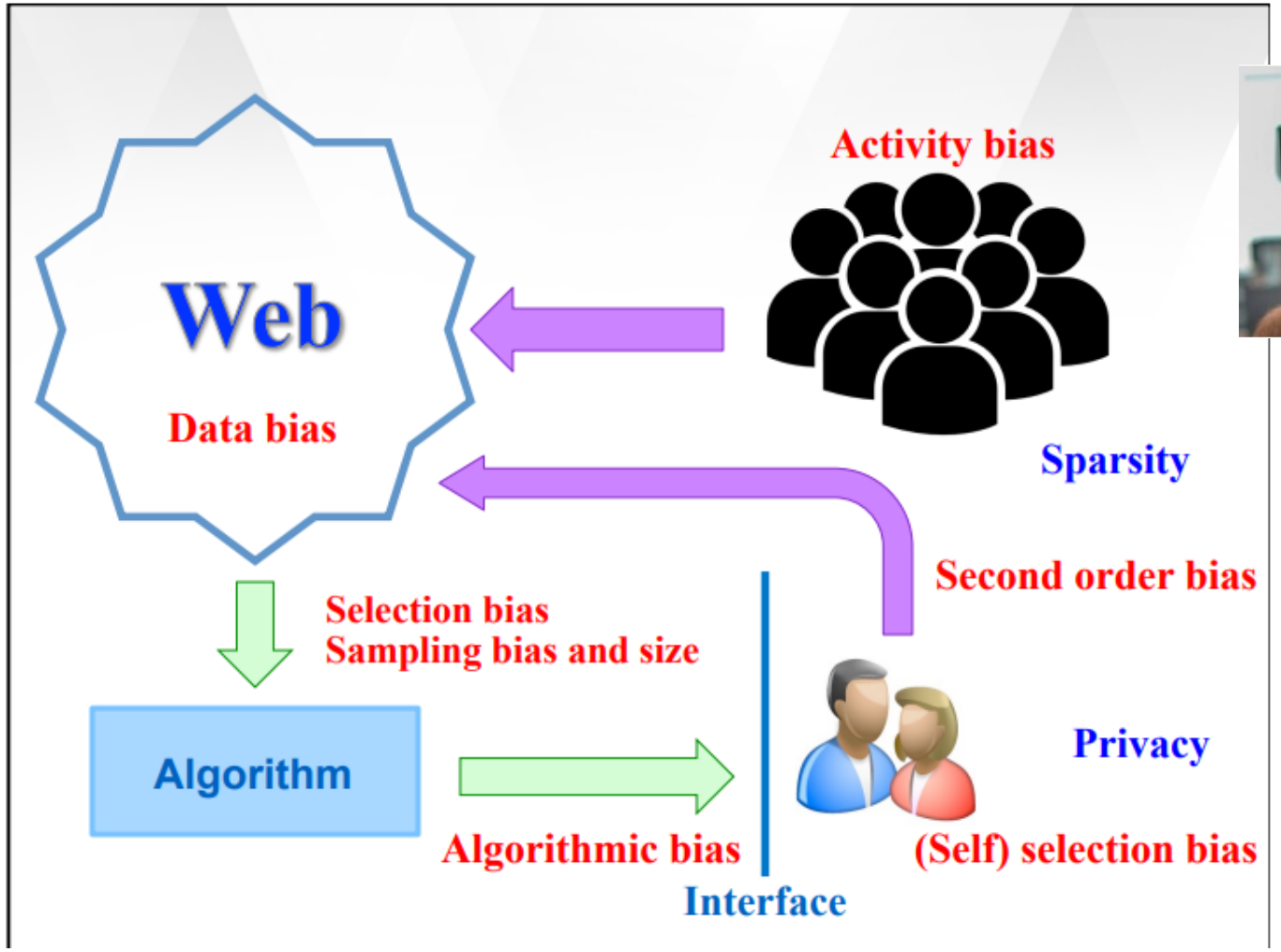
Consistent gendering of traits across regions

Men/boys: ambitious, boastful, competent, conceited, conscientious, consistent, decisive, determined, gullible, independent, industrious, intelligent, lazy, persistent, rational, self-critical, vigorous
Women/girls: detached, emotional, expressive, fair, insecure, open-minded, outgoing, perfectionistic, self-confident, sensitive, shy, warm
Gender-neutral: able, active, affectionate, caring, communicative, competitive, friendly, helpful, self-sacrificing, sociable, supportive, understanding, vulnerable

PEAK indicators

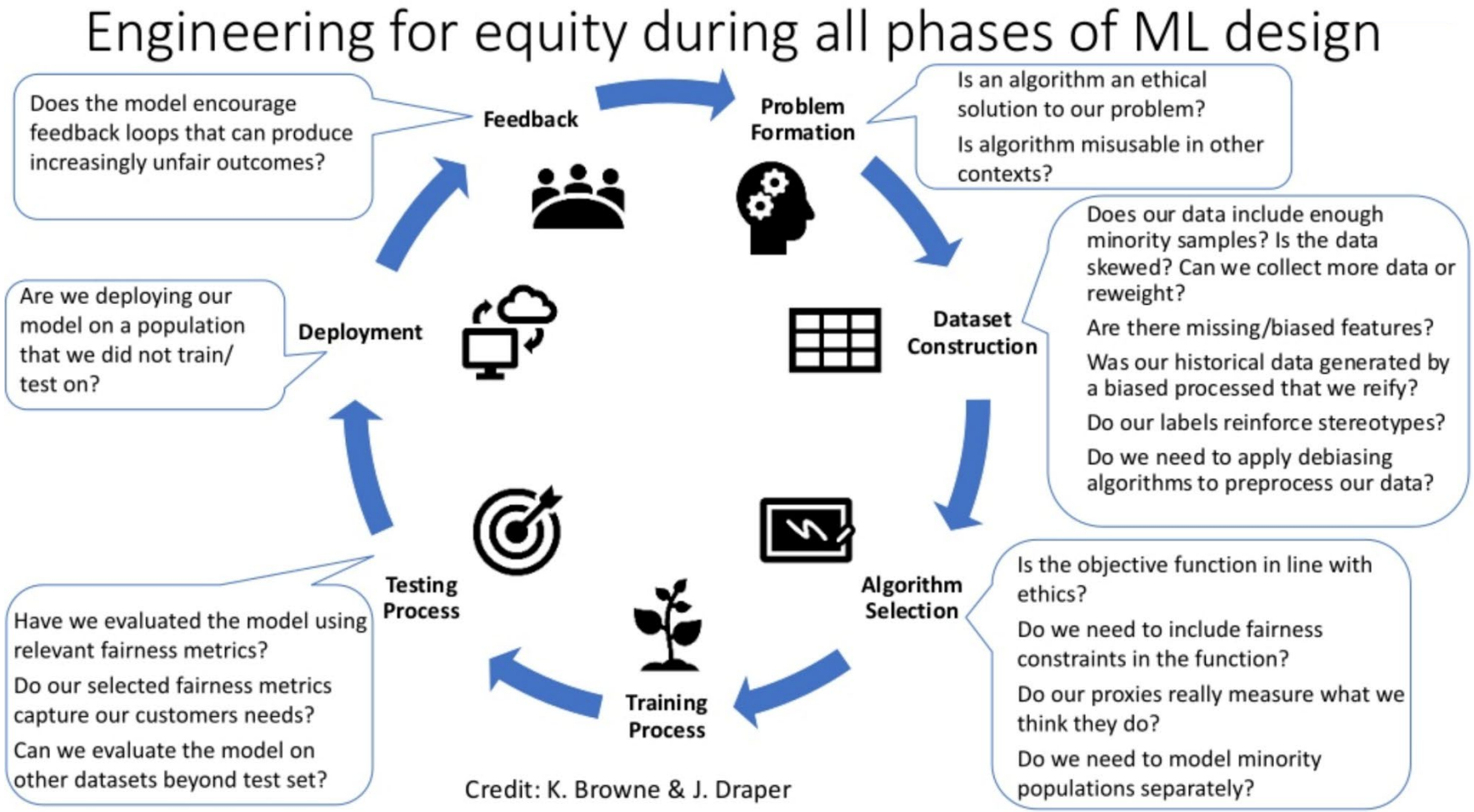
http://www.websci16.org/sites/websci16/files/keynotes/keynote_baeza-yates.pdf

- The bing algorithm is not itself gender-biased
- However, bing image results do perpetuate gendered perceptions of personhood



What can be done?

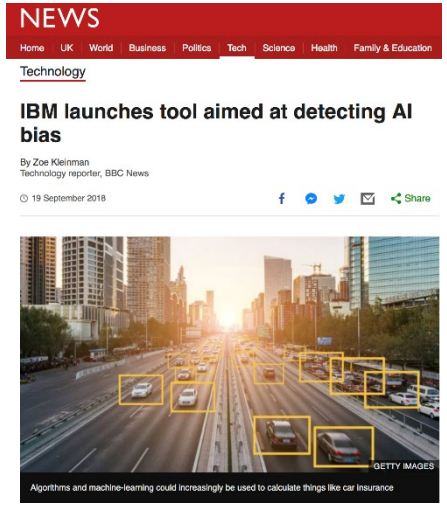
PEAK indicators



PEAK indicators

Technical solutions

- Tools to identify data and algorithmic bias
- Tools to reduce discrimination
- Explainable AI
- Tools for algorithmic auditing



A survey on measuring indirect discrimination in machine learning

INDRĖ ŽLIOBAITĖ, Aalto University and Helsinki Institute for Information Technology HIIT

Nowadays, many decisions are made using predictive models built on historical data. Predictive models may systematically discriminate groups of people even if the computing process is fair and well-intentioned. Discrimination-aware data mining studies how to make predictive models free from discrimination, when historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions. Discrimination refers to disadvantageous treatment of a person based on belonging to a category rather than on individual merit. In this survey we review and organize various discrimination measures that have been used for measuring discrimination in data, as well as in evaluating performance of discrimination-aware predictive models. We also discuss related measures from other disciplines, which have not been used for measuring discrimination, but potentially could be suitable for this purpose. We computationally analyze properties of selected measures. We also review and discuss measuring procedures, and present recommendations for practitioners. The primary target audience is data mining, machine learning, pattern recognition, statistical modeling researchers developing new methods for non-discriminatory predictive modeling. In addition, practitioners and policy makers would use the survey for diagnosing potential discrimination by predictive models.

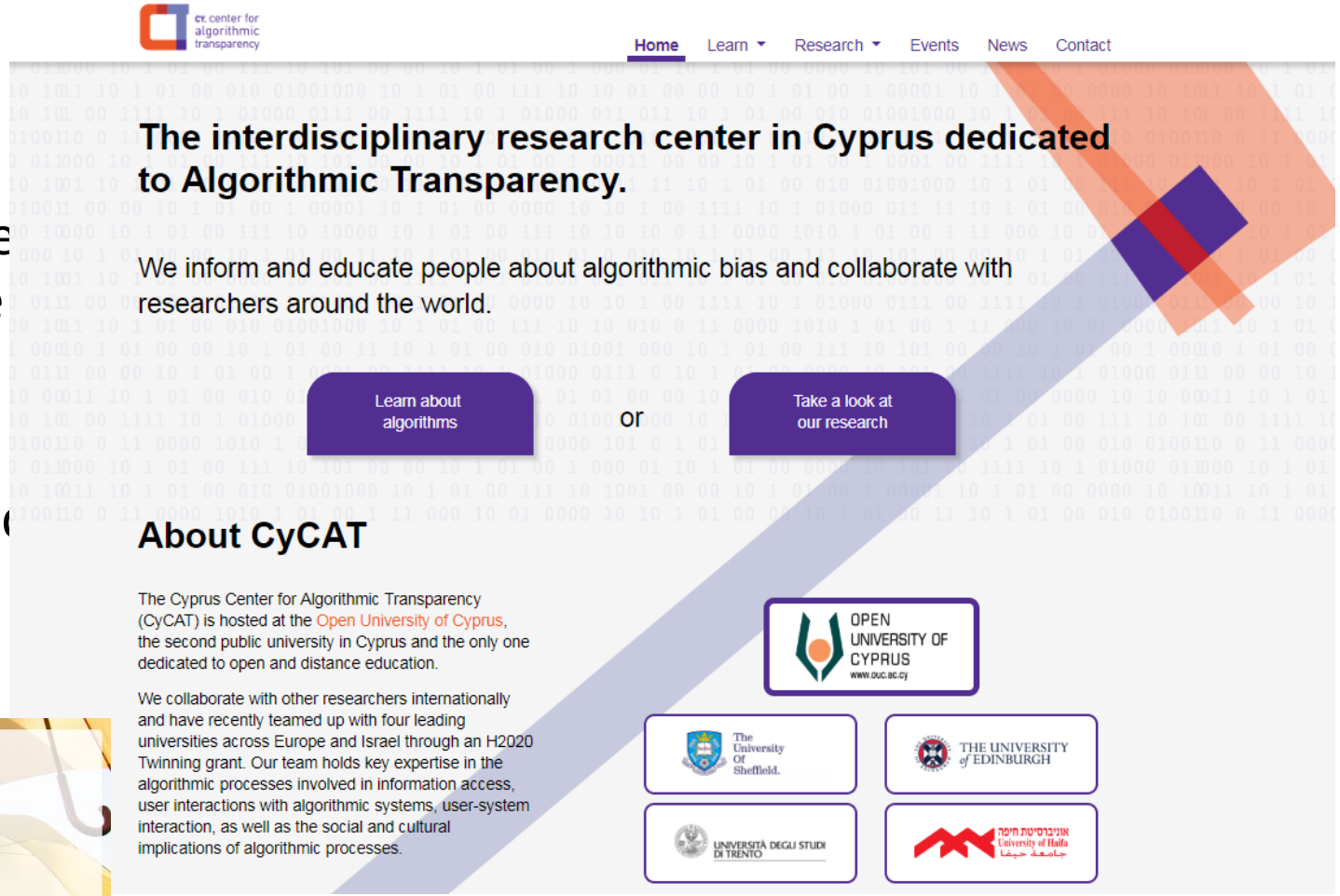
General Terms: fairness in machine learning, predictive modeling, non-discrimination, discrimination-aware data mining

Discrimination discovery

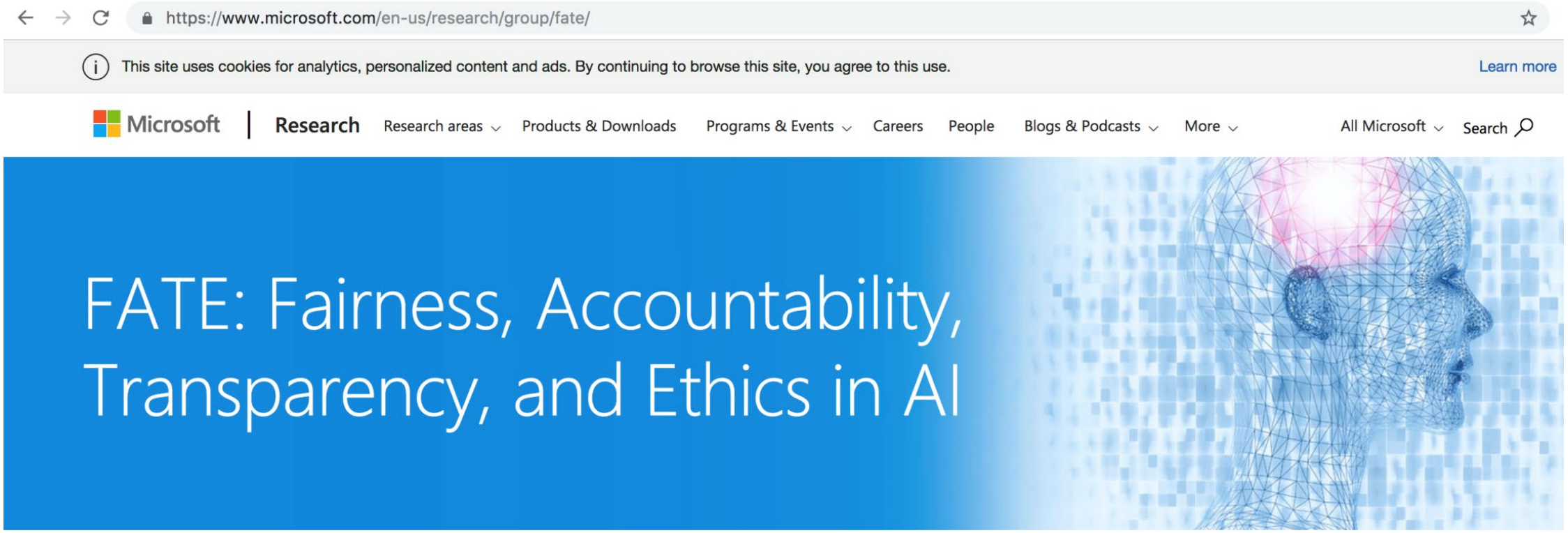
Discrimination prevention

Educational solutions

- Algorithmic literacy to help people gain a broad understanding of the algorithmic 'value chain' (people perceive algorithms as unbiased)
- Educational programmes for raise awareness of discrimination and algorithmic bias



PEAK indicators

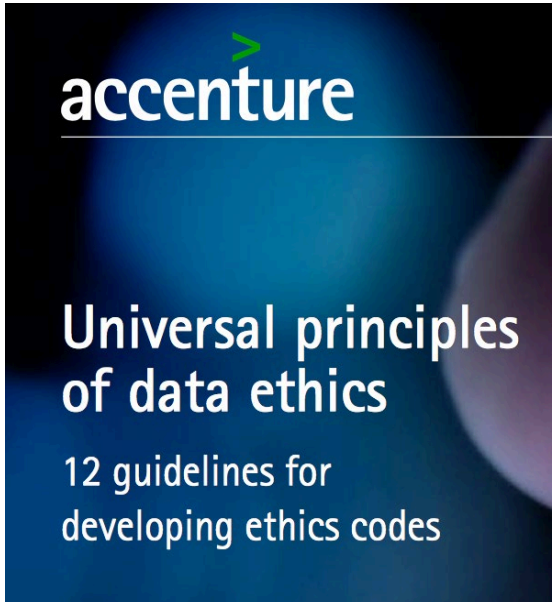


Governance and transparency

- Clear governance procedures and algorithm accountability (role of CDO?)
- Developing an algorithmic audit trail –

  "knowing very well the data collected [and the sources], identifying which pieces of data are used by algorithms, and made known how this data is weighted or used in the algorithm"
- Ethics and governance frameworks

ALGORITHM WATCH



"Governance practices should be robust, known to all team members and reviewed regularly"

Guidance

Data Ethics Framework

Department for Digital, Culture, Media & Sport

Public sector organisations should use the Data Ethics Framework to guide the appropriate use of data to inform policy and service design.

Summary

Summary

- Decision-making increasingly data-driven
- Algorithms commonly employed to support or automate decisions and processes
- However, like humans, machines introduce biases and one area of current concern is algorithmic bias
- Algorithms and digital systems can perpetuate bias– one example being gender bias in image search
- Solutions include technical, educational and governance

Note: we can only expose, measure and try to reduce bias; we cannot completely eradicate it

Data. Insight. Action.

Questions?