Sources of Bias in data

Abuses **Bias** and Blessings of Data

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, *"A Survey on Bias and Fairness in Machine Learning,"* 2019.

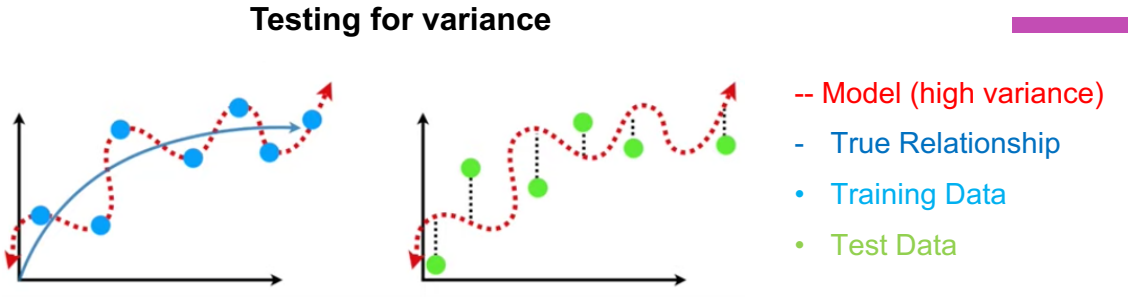
Definition of Bias?

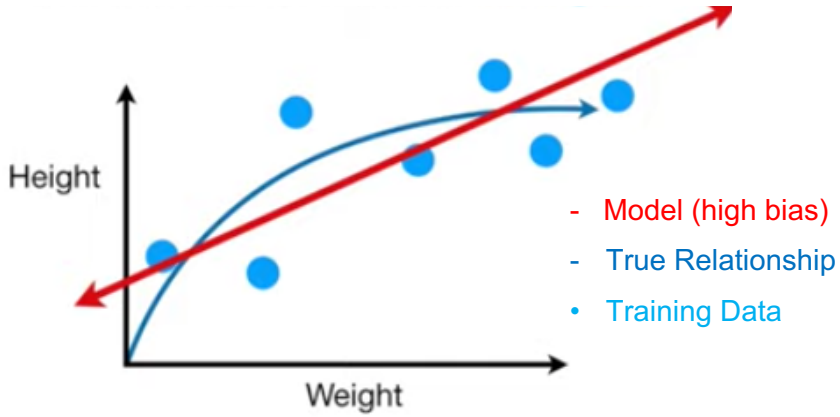
Cambridge Dictionary

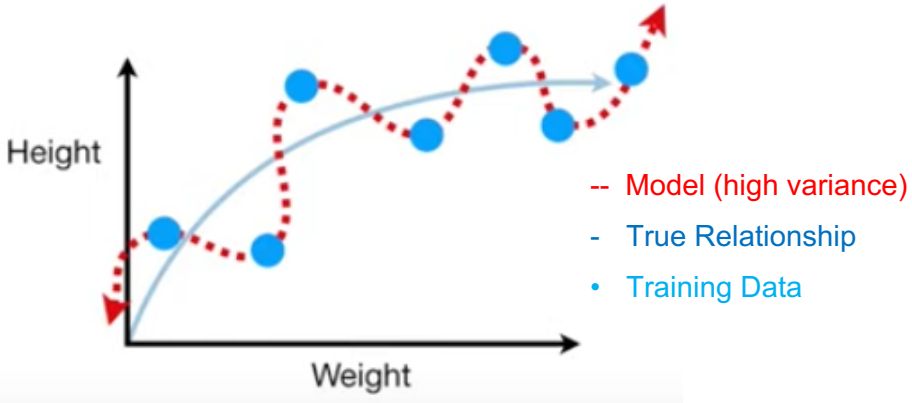
- Noun
  - inclination or prejudice for or against one person or group, especially in a way considered to be unfair.
    - "there was evidence of **bias against** foreign applicants"
- Verb
  - cause to feel or show inclination or prejudice for or against someone or something.
    - "all too often, our recruitment processes are **biased towards** younger candidates"

- the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment:
  - The senator has accused the media of bias.
  - Reporters must be impartial and not show political bias.
  - There was clear evidence of a strong bias against her.
  - There has always been a slight bias in favour of/towards employing liberal arts graduates in the company.
  - Unconscious bias (= that the person with the bias is not aware of) can influence decisions in recruitment, promotion, and

Not to be confused with Statistical Bias







Statistical Bias (and Variance)

- statistical bias – there is a statistical bias if a predictor is consistently over shooting or undershooting the target.
- Says nothing about errors or distribution of errors
- Data Bias are inevitable we must design algorithms that account for them
- Reframe – not about mathematical correctness: make algorithmic systems that support human values

Hypothetical Example

- study how body mass index (BMI) changes as a function of daily pasta intake

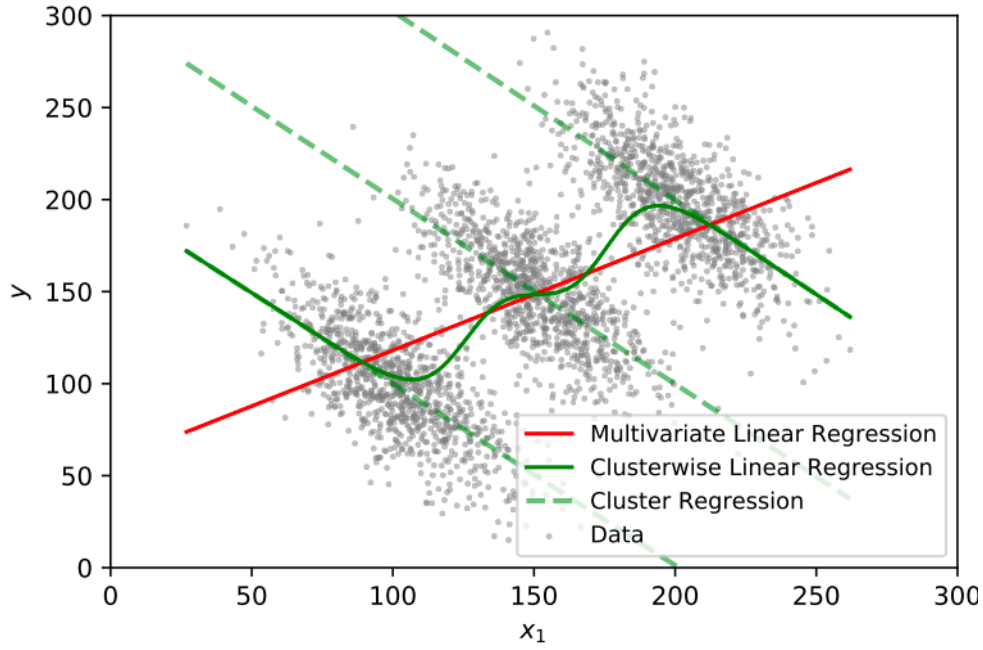
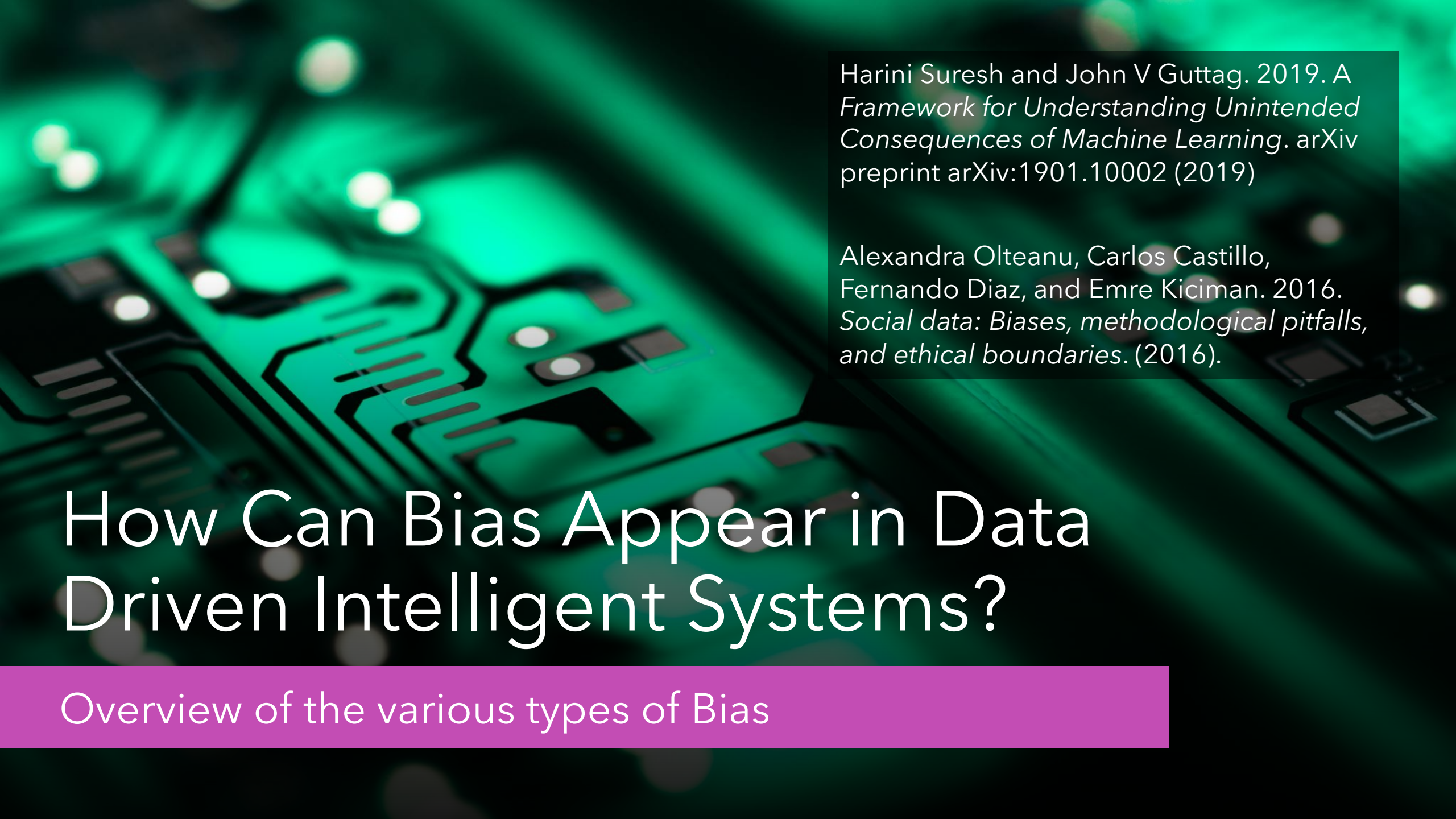


Illustration of biases in data. Red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (Credit: Nazanin Alipourfard)



Harini Suresh and John V Guttag. 2019. A *Framework for Understanding Unintended Consequences of Machine Learning*. arXiv preprint arXiv:1901.10002 (2019)

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. *Social data: Biases, methodological pitfalls, and ethical boundaries*. (2016).

How Can Bias Appear in Data Driven Intelligent Systems?

Overview of the various types of Bias

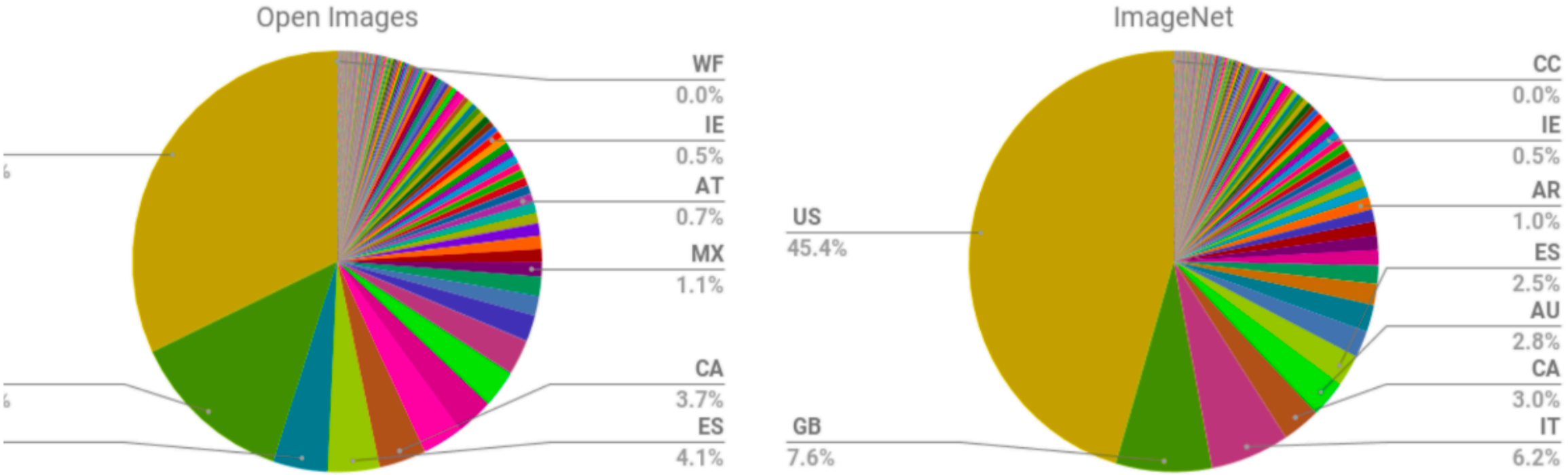
Historical Bias

Google image search for “CEO”

- *Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection.*
  - Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

Representation Bias

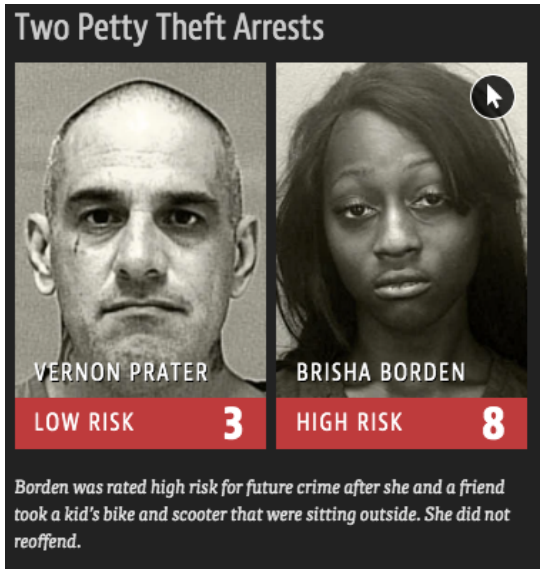
Representation bias happens from the way we define and sample from a population

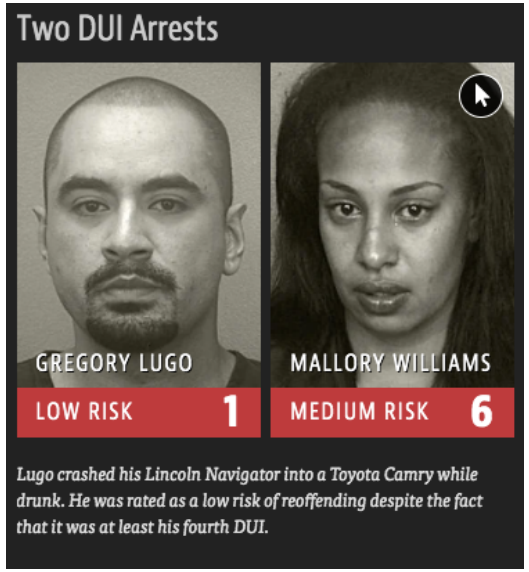


Fraction of each country, represented by their two-letter ISO codes, in Open Images and ImageNet image datasets. In both datasets, US and Great Britain represent the top locations, from:
Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. stat 1050 (2017), 22.

Measurement Bias

- *Measurement bias happens from the way we choose, utilize, and measure a particular feature*
  - Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).
- **Example:** friend/family arrests were used as proxy variables to measure level of "riskiness" or "crime" in COMPAS
  - https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

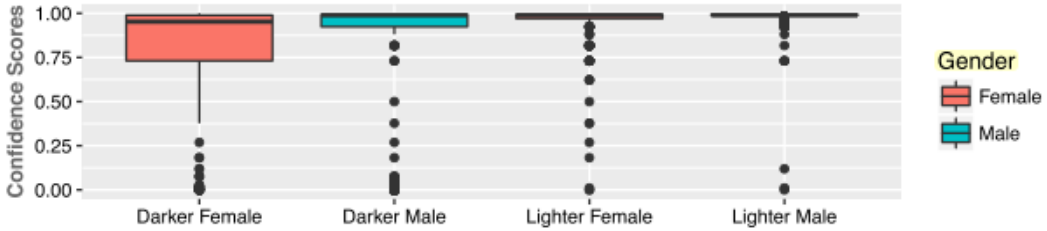




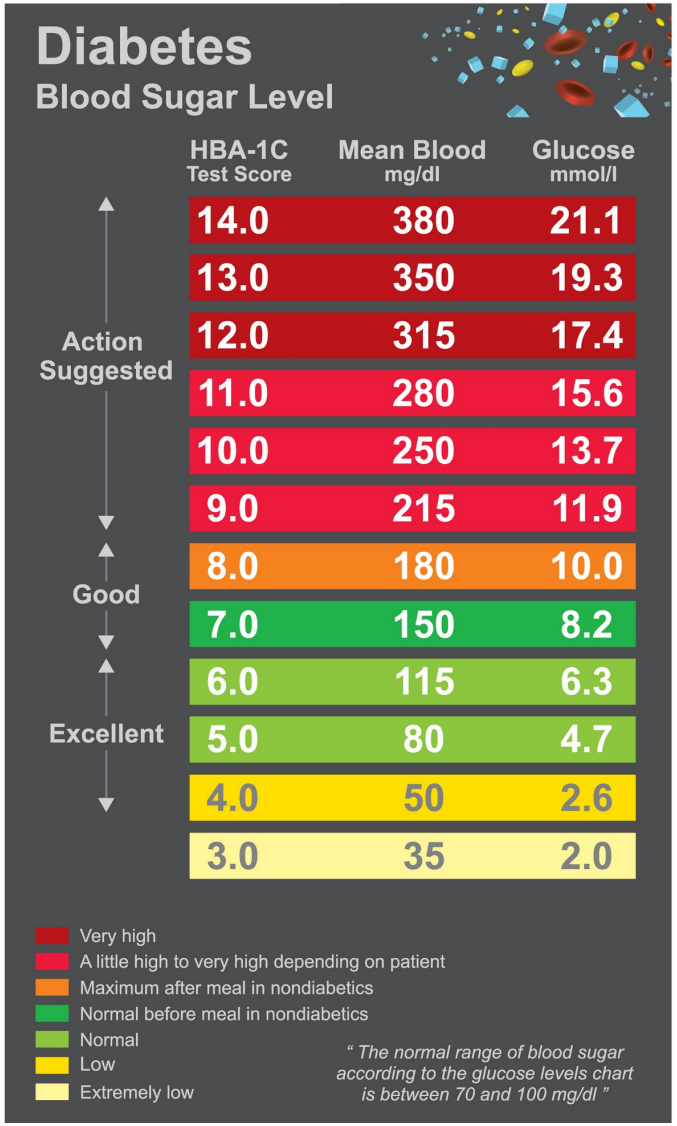
Evaluation Bias

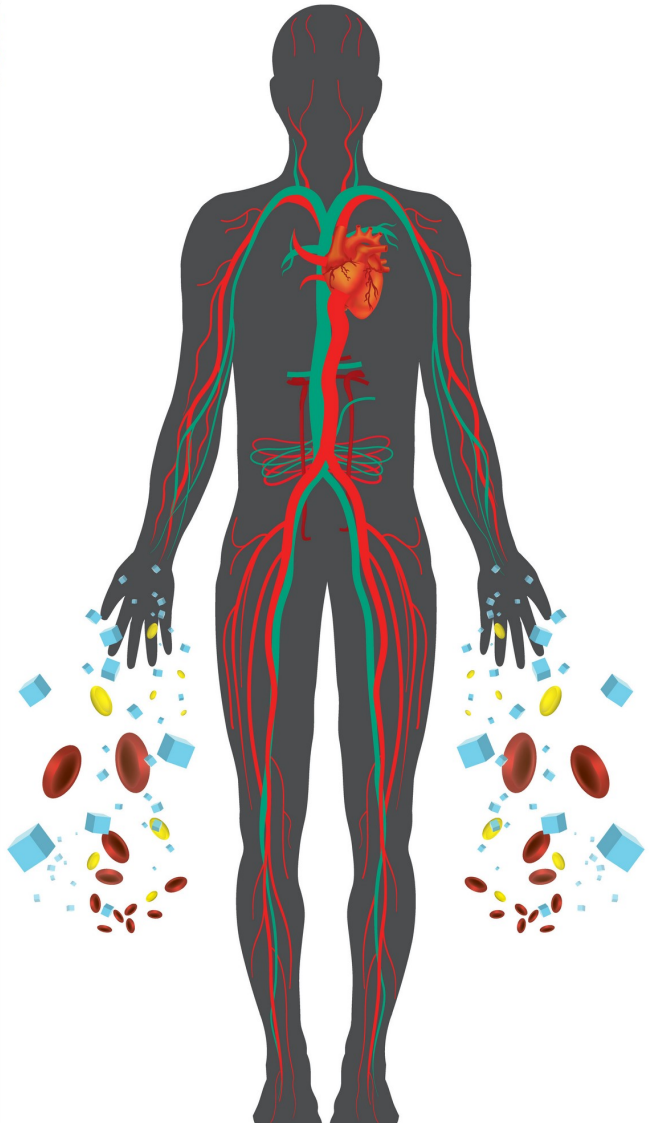
- *Evaluation bias happens during model evaluation*
  - Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).
- Use of inappropriate and disproportionate benchmarks for evaluation of applications
  - Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91.

The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB



Gender classification confidence scores from IBM (IBM). Scores are near 1 for lighter male and female subjects while they range from ~ 0.75 – 1 for darker females.





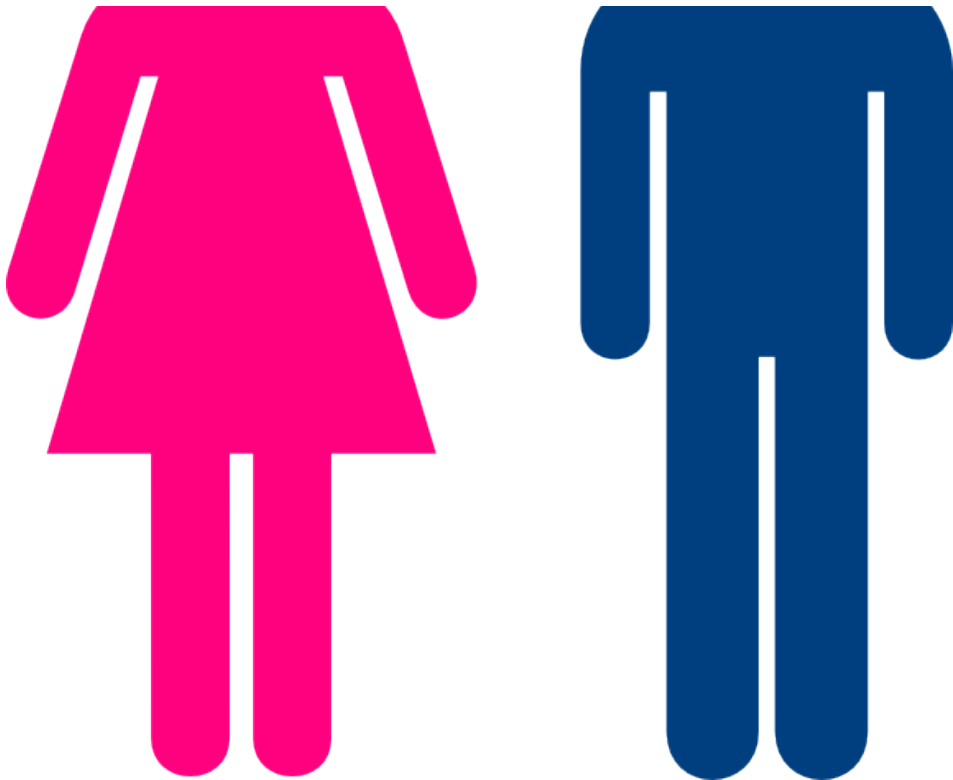
Aggregation Bias

- *Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition.*
  - Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

Image source: https://diabeticmuscleandfitness.com/how-to-convert-average-blood-glucose-into-hba1c/

Population Bias

- *Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population.*
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).
- different user demographics on different social platforms
  - Eszter Hargittai. 2007. Whose Space? Differences among Users and Non-Users of Social Network Sites. Journal ofComputer-Mediated Communication 13, 1 (10 2007), 276–297. https://doi.org/10.1111/j.1083-6101.2007.00396.x

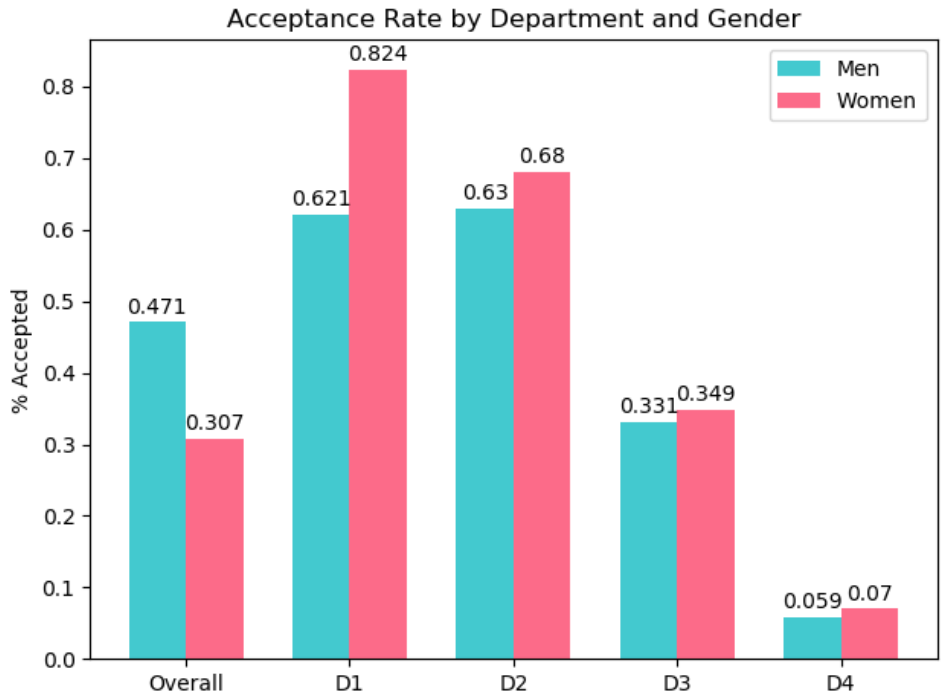


Simpson's Paradox

- *According to Simpson's paradox, a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated*

Fact 1: *In 1973, at the University of California, Berkeley, the overall acceptance rate in four departments for female applicants was roughly 30%. At the same time, the overall acceptance rate for male applicants was roughly 47%.*

Fact 2: *In each of the departments, the acceptance rate for female applicants was higher than the acceptance rate for male applicants*



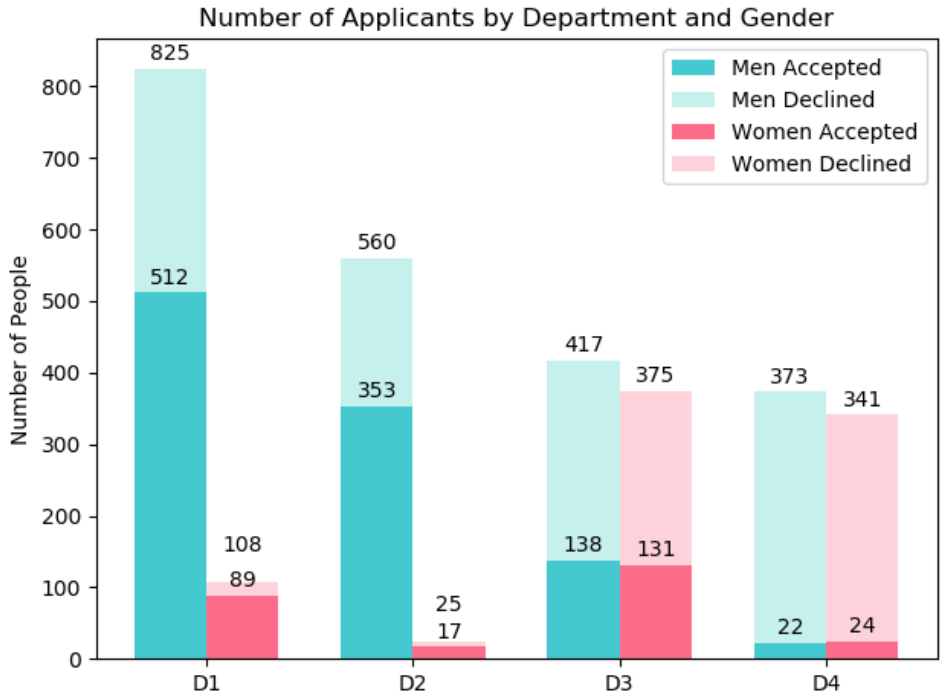


Image Source: https://towardsdatascience.com/gender-bias-in-admission-statistics-the-simpson-paradox-cd381d994b16

Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. Science 187, 4175 (1975), 398–404.

Longitudinal Data Fallacy

- *Observational studies often treat cross-sectional data as fit were longitudinal, which may create biases due to Simpson's paradox*
- Reddit example
  - Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar-Jr. 2016. Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior. (April 2016), 829–841.

Sampling Bias

Sampling bias arises due to non-random sampling of subgroups

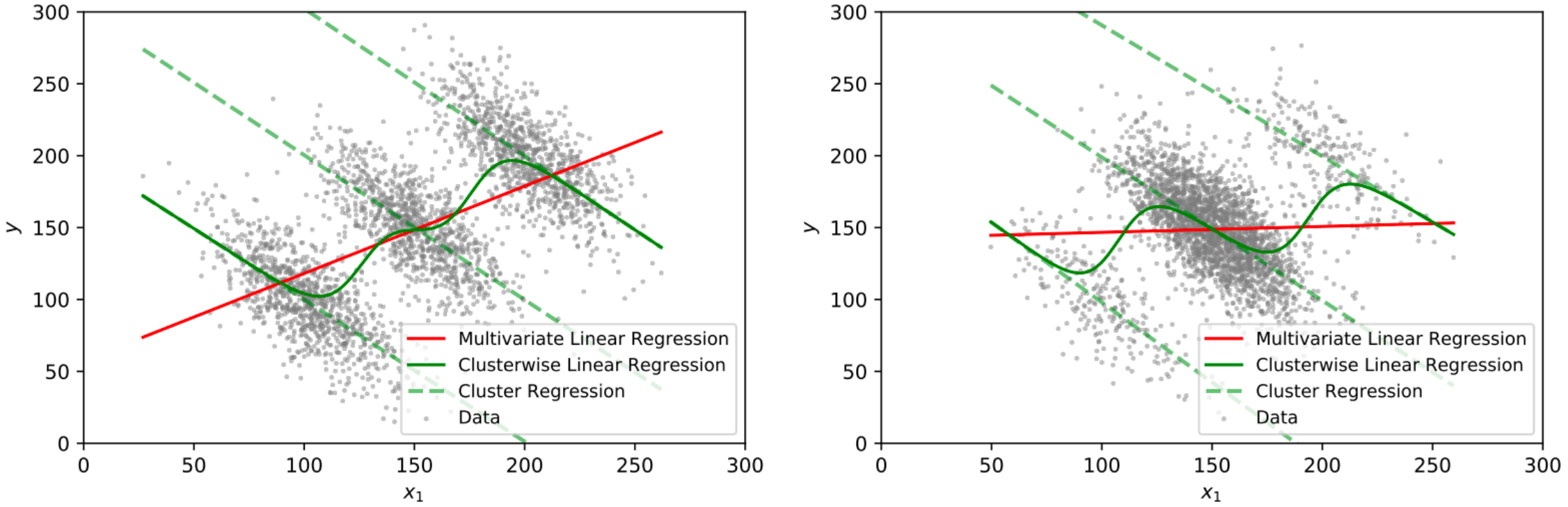
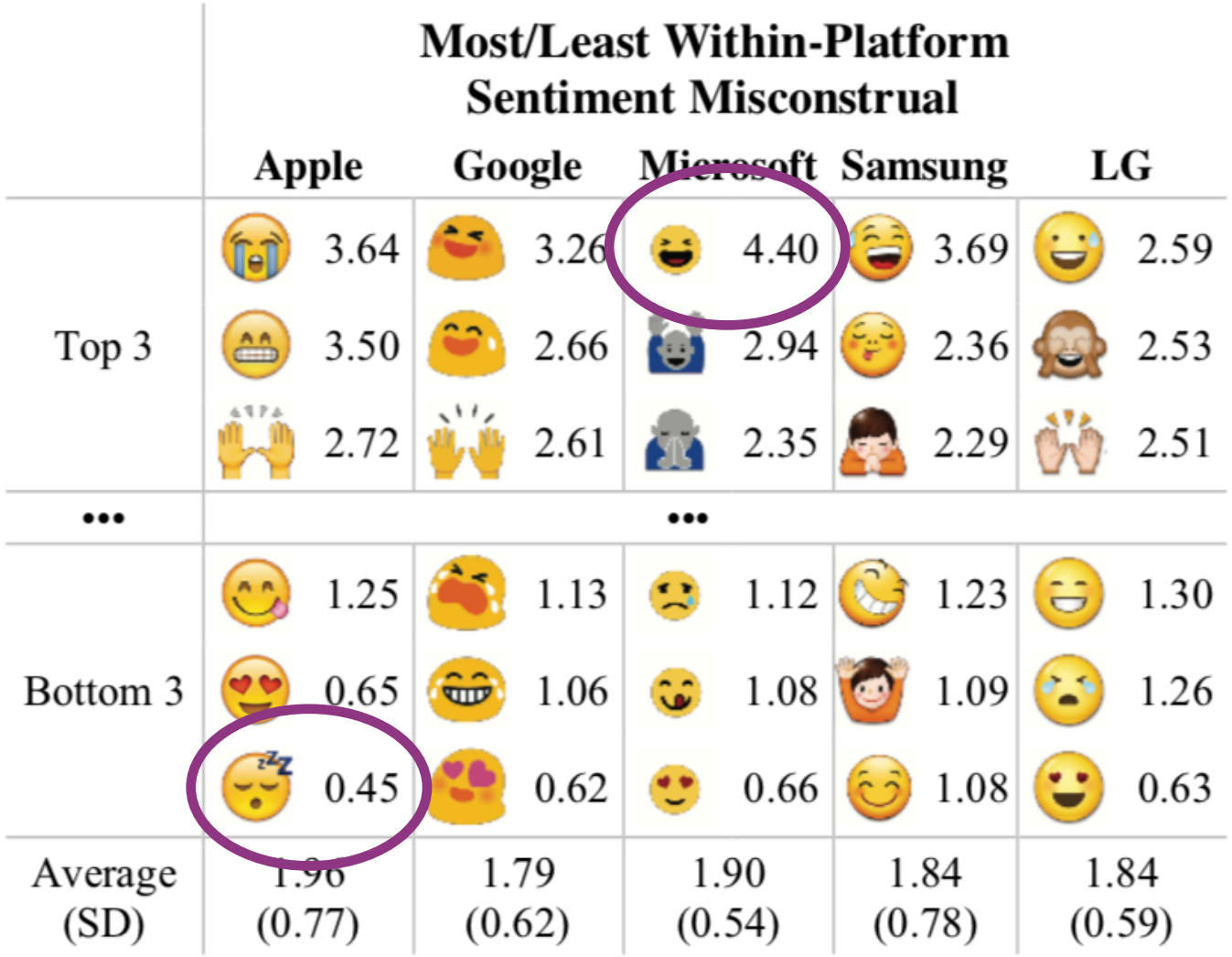


Illustration of biases in data. Red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between the outcome and the independent variable. (b) Regression shows almost no relationship in less balanced data. The relationships between variables within each subgroup, however, remain the same. (Credit: Nazanin Alipourfard)

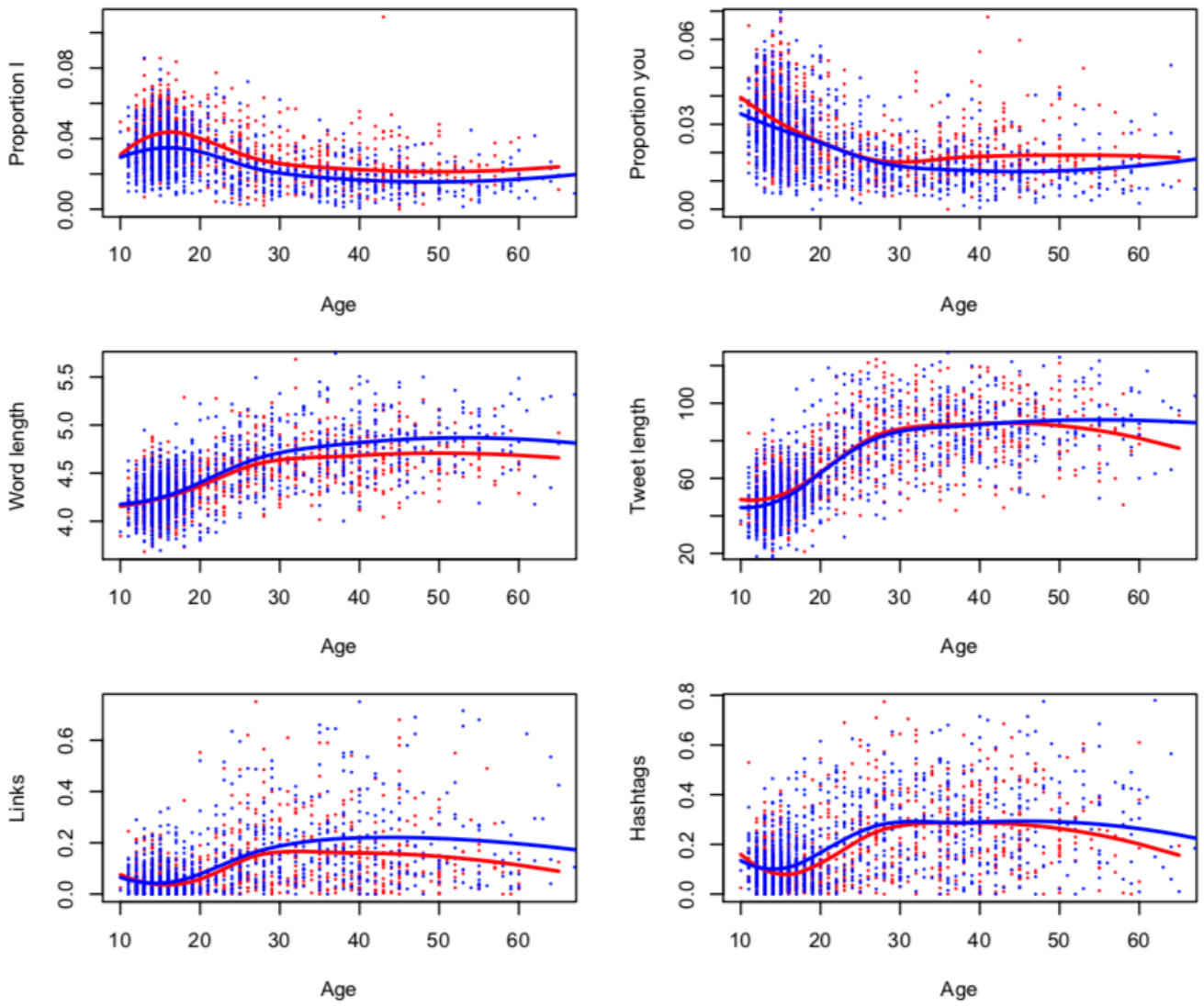
Behavioral Bias

- *Behavioral bias arises from different user behavior across platforms, contexts, or different datasets.*
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016)
- Interpretations of Emojis 😂 Example
  - Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji. In Tenth International AAAI Conference on Web and Social Media



	Most/Least Within-Platform Sentiment Misconstrual				
	Apple	Google	Microsoft	Samsung	LG
Top 3	3.64	3.26	4.40	3.69	2.59
	3.50	2.66	2.94	2.36	2.53
	2.72	2.61	2.35	2.29	2.51
•••			•••		
Bottom 3	1.25	1.13	1.12	1.23	1.30
	0.65	1.06	1.08	1.09	1.26
	0.45	0.62	0.66	1.08	0.63
Average (SD)	1.96 (0.77)	1.79 (0.62)	1.90 (0.54)	1.84 (0.78)	1.84 (0.59)

Top-3 and bottom-3 most different in terms of sentiment.
Higher values indicate greater response variation.



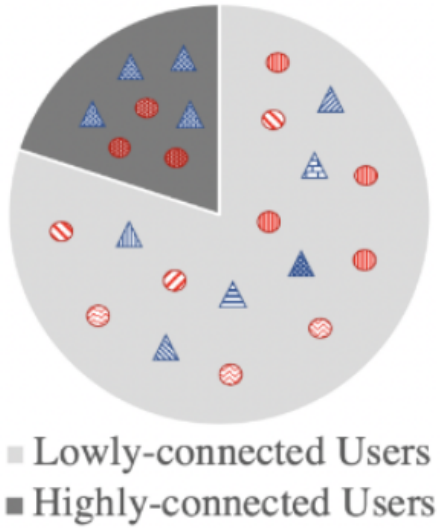
Plots of Variables as they change with age. Blue: males, Red: females

Content Production Bias

- *Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users.*
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).
- Example: differences in use of language across gender and age
  - Dong-Phuong Nguyen, Rilana Gravel, Rudolf Berend Trieschnigg, and Theo Meder. 2013. "How old do you think I am?": A study of language and age in Twitter. In Proceedings ofthe Seventh International AAAI Conference on Weblogs and Social Media, ICWSM 2013. AAAI Press, 439–448. eemcs-eprint-23604.

Linking Bias

- *Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users*
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).
- social networks analysis biased toward low-degree nodes when only considering the links in the network
  - Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Debiasing Community Detection: The Importance of Lowly-Connected Nodes. arXiv preprint arXiv:1903.08136 (2019).



Demonstration of versatility of relevant low-degree people in social networks who get ignored by biased community detection methods. Shapes (circle, triangle) represent opinion, and texture represents variations of the opinion.

The Gamergate retweet network colored based on the network structure is shown on the left hand side, and the net- work colored by the ground truth labels is shown on the right hand side. The callout zooms one of the components, showing the disagreement between the two labeling approaches. Purple nodes represent Gamergate opposers and green nodes represent Gamergate supporters.

Temporal Bias

- *Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation*
  - Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

Popularity Bias

- *Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation*
- for example, by fake reviews or social bots
  - Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. 2017. How algorithmic popularity bias hinders or promotes quality. arXiv preprint arXiv:1707.00574 (2017).

Quality

Faithfulness

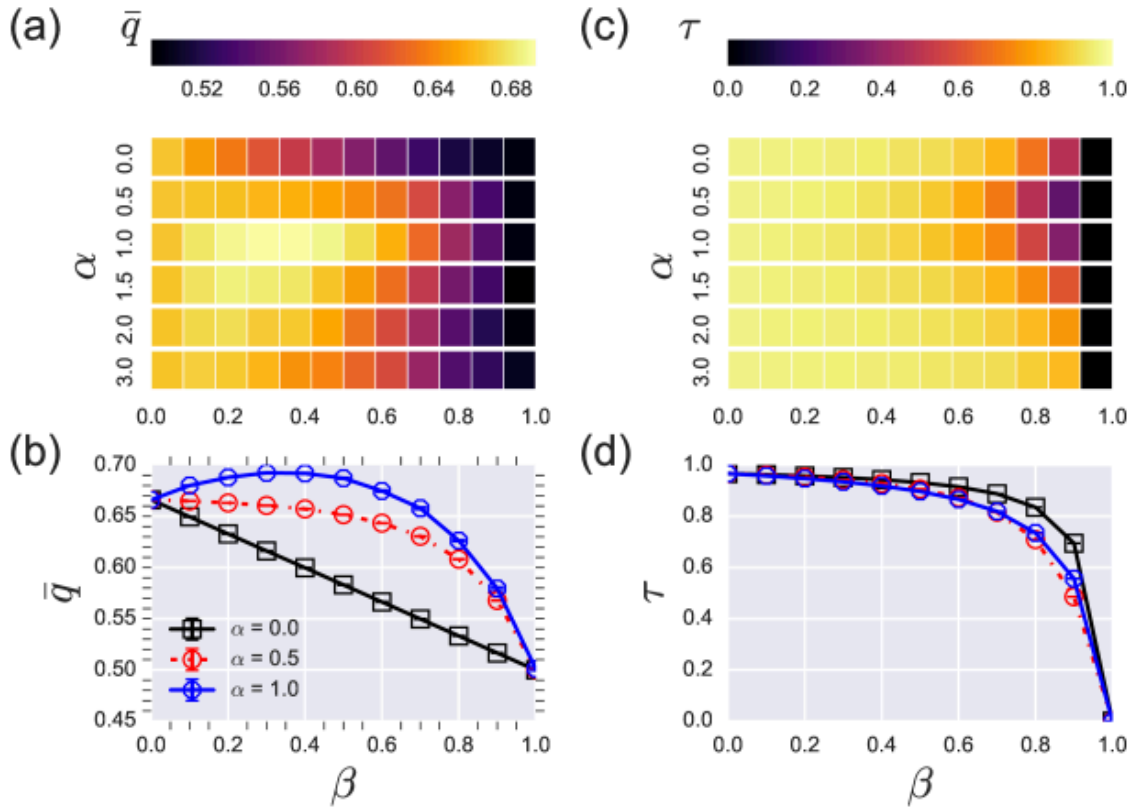
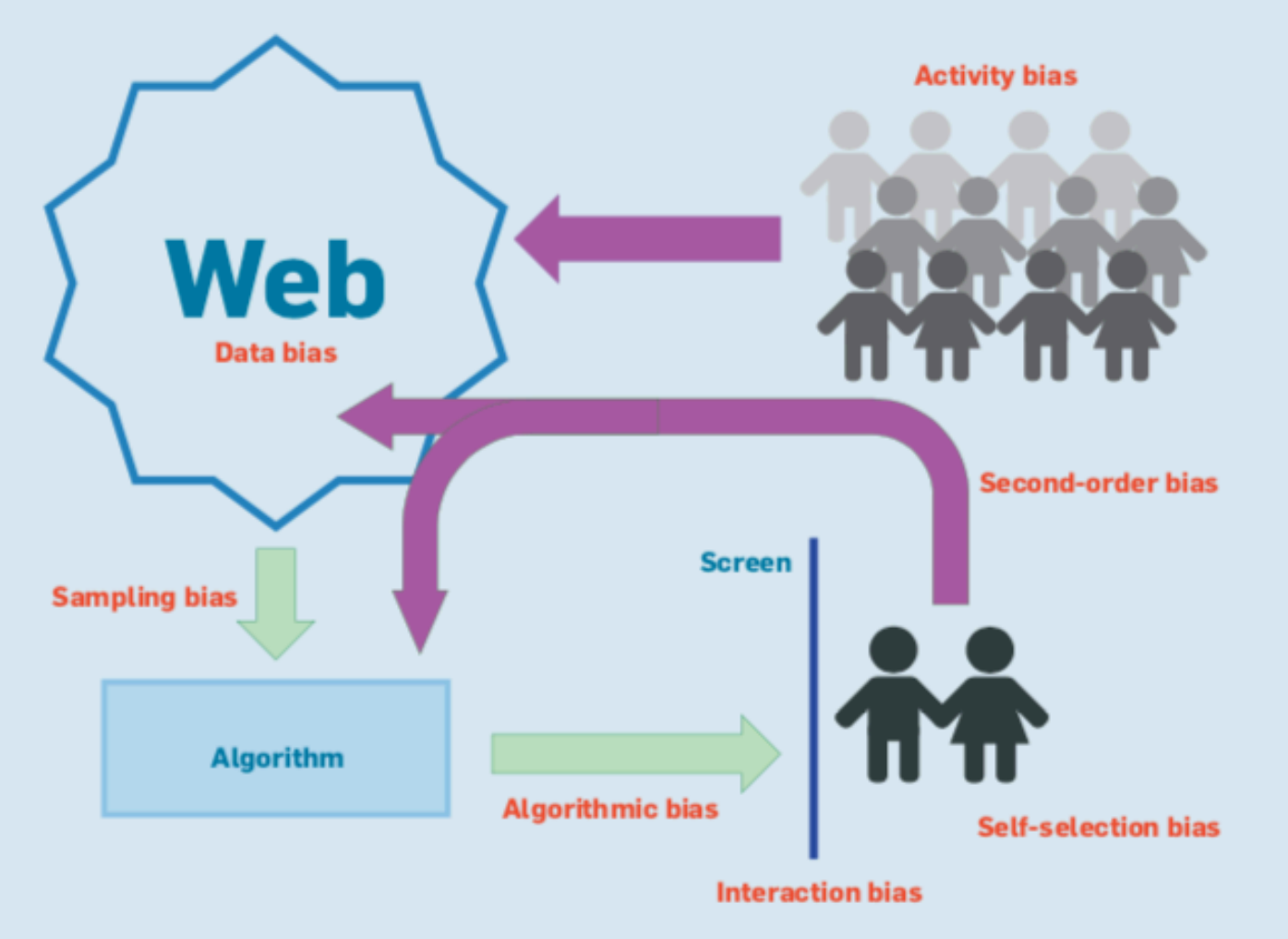


Figure 1: ***Effects of popularity bias on average quality and faithfulness.****. (a) Heatmap of average quality $\bar{q}$ as a function of α and β, showing that $\bar{q}$ reaches a maximum for $\alpha = 1$ and $\beta \approx 0.4$, while for $\alpha = 3$ the maximum is attained for a lower β. (b) The location of the maximum $\bar{q}$ as a function of β depends on α, here shown for $\alpha = 0, 0.5, 1.0$. (c) Faithfulness τ of the algorithm as a function of α and β. (d) τ as a function of β for the same three values of α. Standard errors are shown in panels (b,d) and are smaller than the markers.*

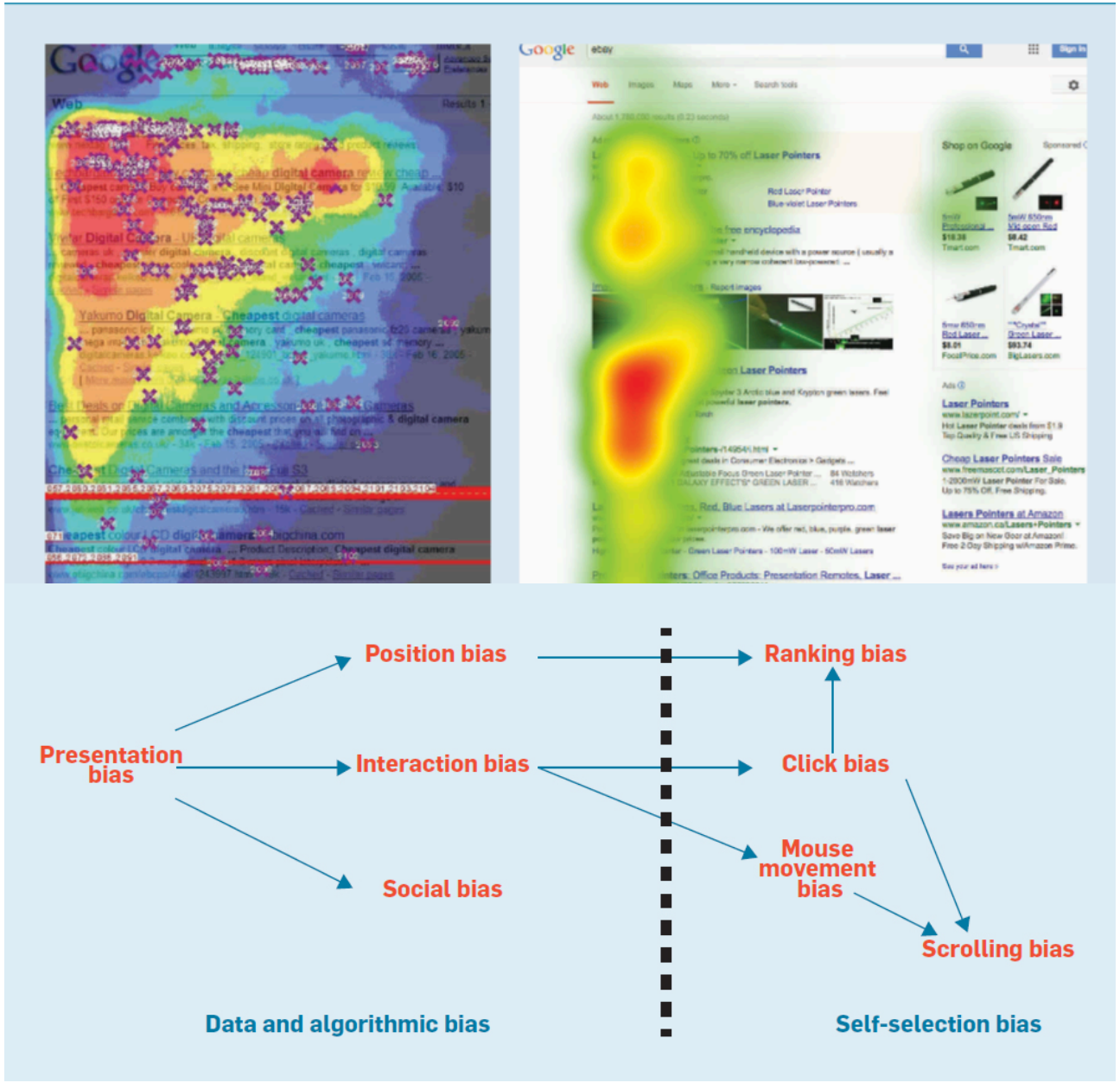
Algorithmic Bias

- Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm
  - Ricardo Baeza-Yates. 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581



User Interaction Bias

- *User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources–the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction*
- This type of bias can be influenced by other types and subtypes, such as Presentation and Ranking biases.
- **Presentation Bias**
  - Presentation bias is a result of how information is presented. For example, on the Web users can only click on content that they see, so the seen content gets clicks, while everything else gets no click. And it could be the case that the user does not see all the information on the Web.
- **Ranking Bias**
  - The idea that top-ranked results are the most relevant and important will result in attraction of more clicks than others. This bias affects search engines and crowdsourcing applications



Ricardo Baeza-Yates. 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

Social Bias

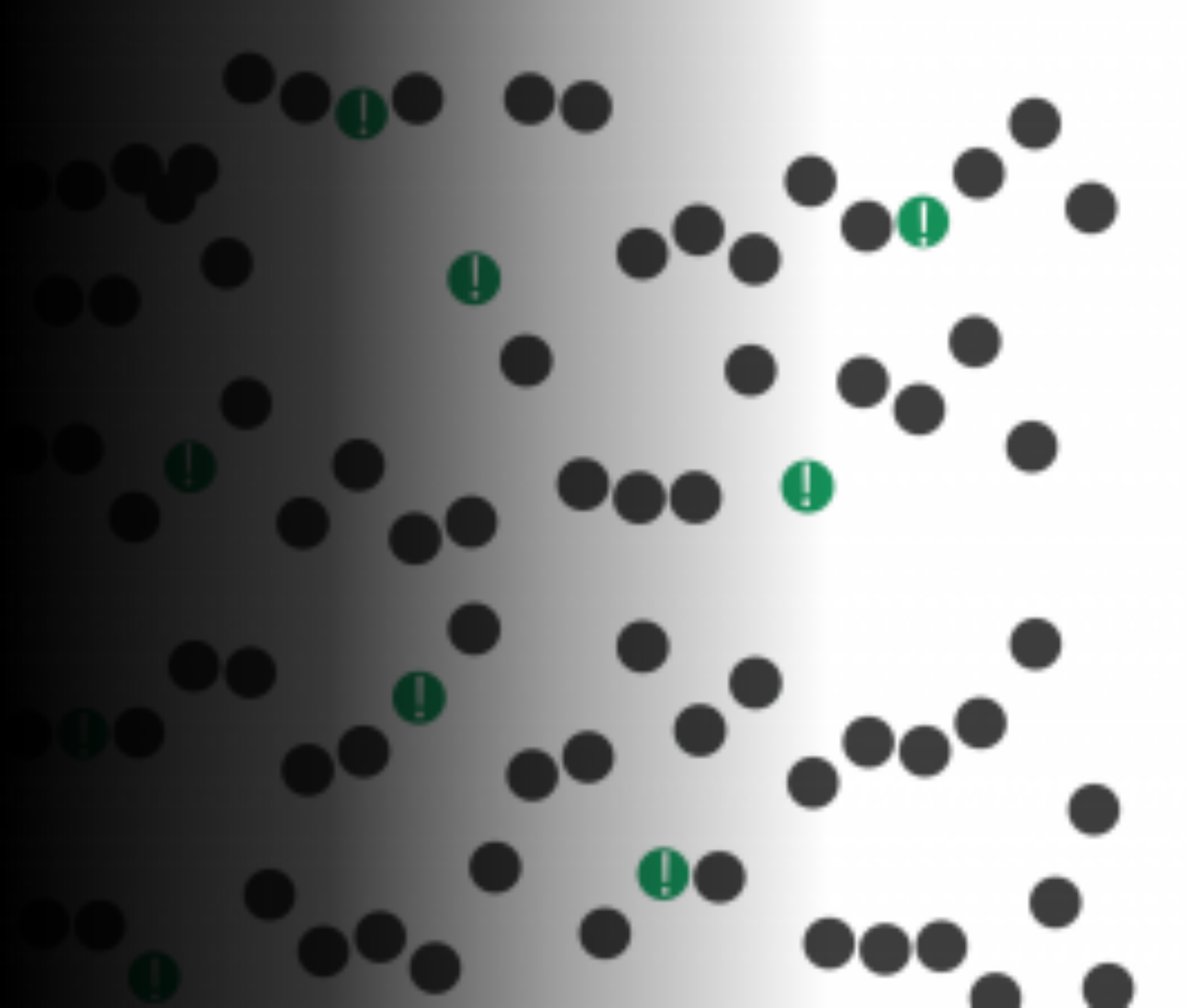
- Social bias happens when other people's actions or content coming from them affect our judgment.
  - Ricardo Baeza-Yates. 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

Emergent Bias

- *Emergent bias happens as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design.*
  - Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

Self-Selection Bias

- *Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves*
  - *https://data36.com/statistical-bias-types-explained/*



Omitted Variable Bias

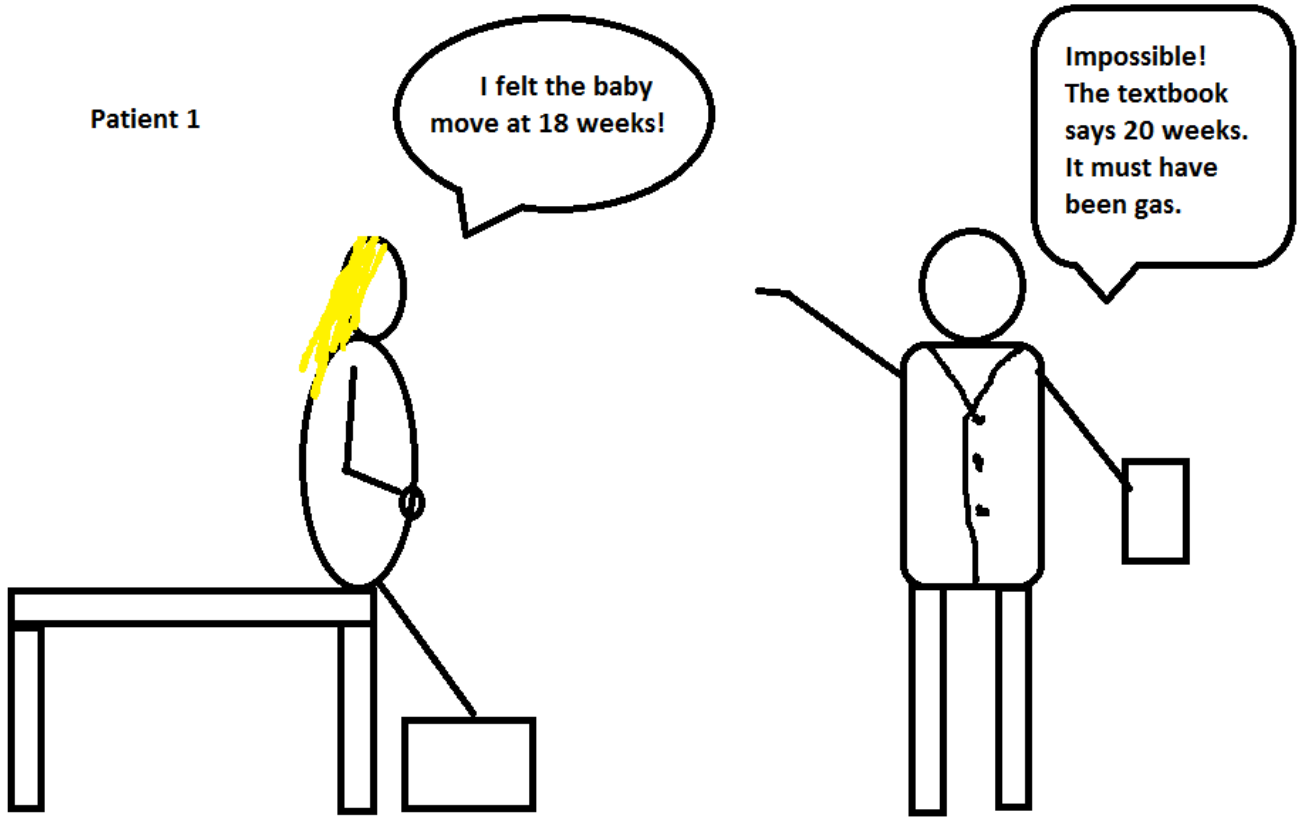
- *Omitted variable bias occurs when one or more important variables are left out of the model.*
  - https://data36.com/statistical-bias-types-explained/

Cause-Effect Bias

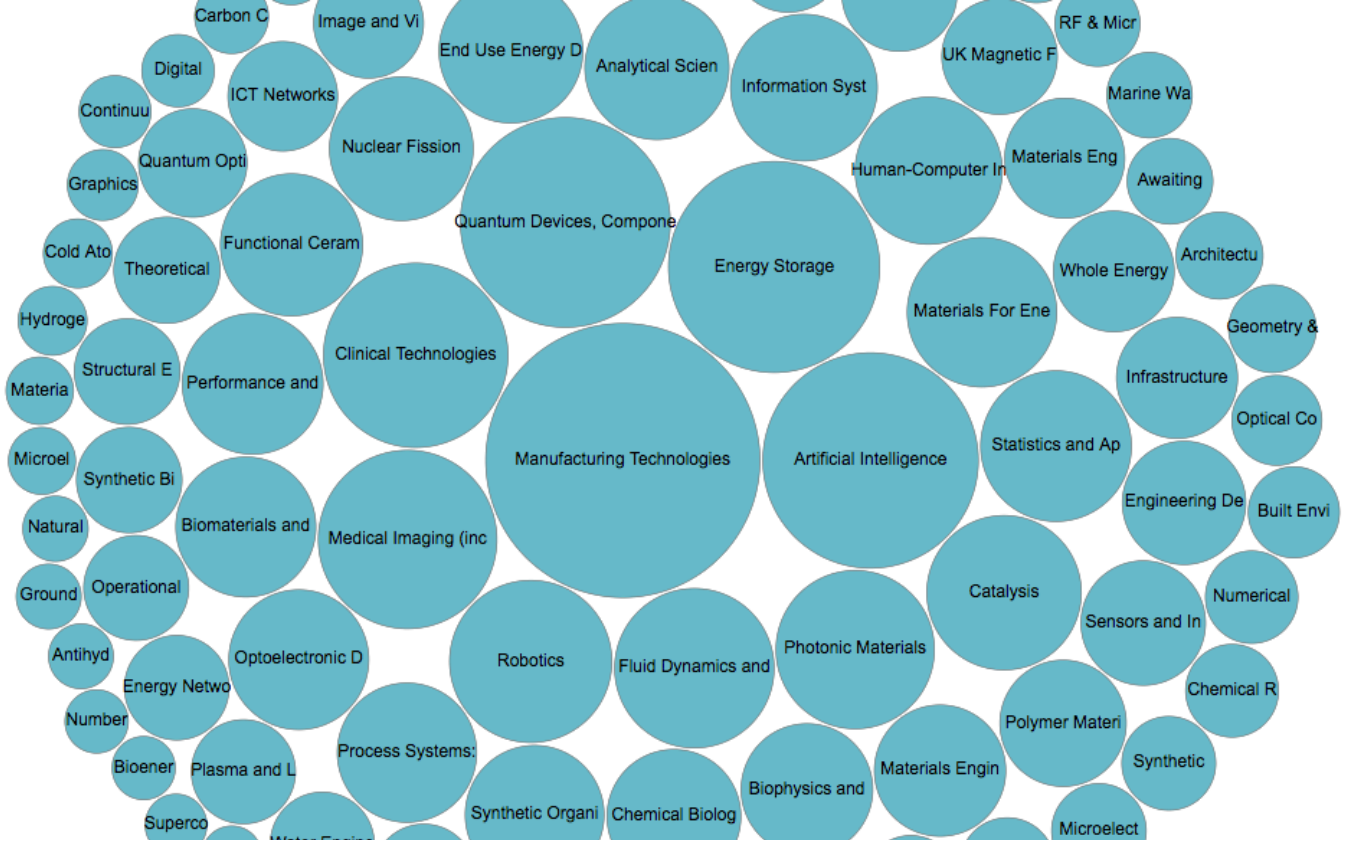
- *Cause-effect bias can happen as a result of the fallacy that correlation implies causation.*
  - https://data36.com/statistical-bias-types-explained/

Observer Bias

- *Observer bias happens when researchers subconsciously project their expectations onto the research*
  - https://data36.com/statistical-bias-types-explained/



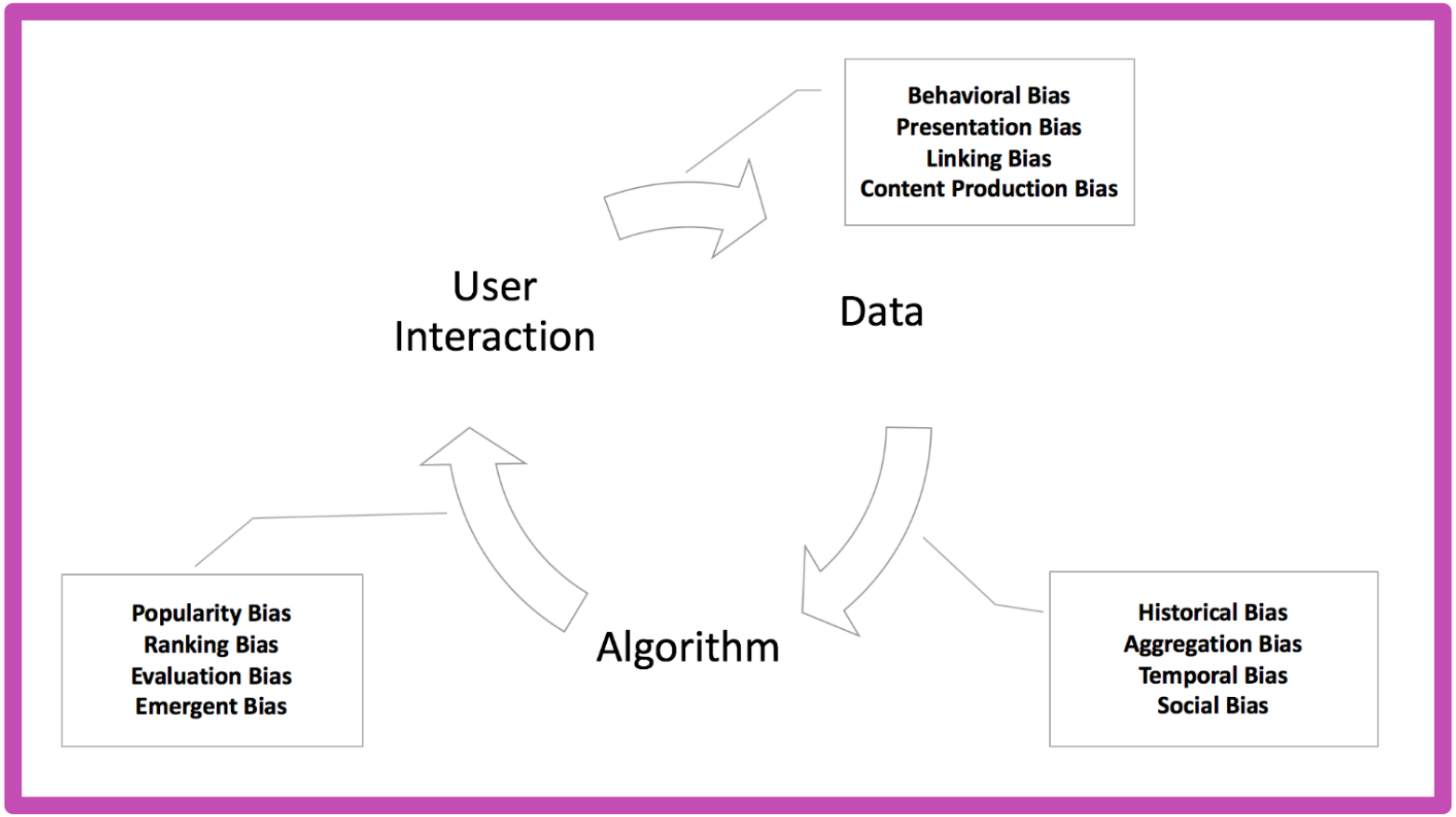
Engineering and Physical Sciences Research Council



Funding Bias

- Funding bias arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study
  - https://data36.com/statistical-bias-types-explained/

Bias definitions in the data, algorithm, and user interaction feedback loop



What to do about bias?

- Pre-processing.
  - Pre-processing techniques try to transform the data so that the underlying discrimination is removed
- In-processing.
  - In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process.
- Post-processing.
  - Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model.