CSI 709/CSS 739
Verification and Validation of Models

Validation of Machine Learning Models (Bias, Fairness, and Assurance)

Dr. Hamdi Kavak
Computational and Data Sciences Department

http://www.hamdikavak.com
hkavak@gmu.edu

Stephen Hawking says

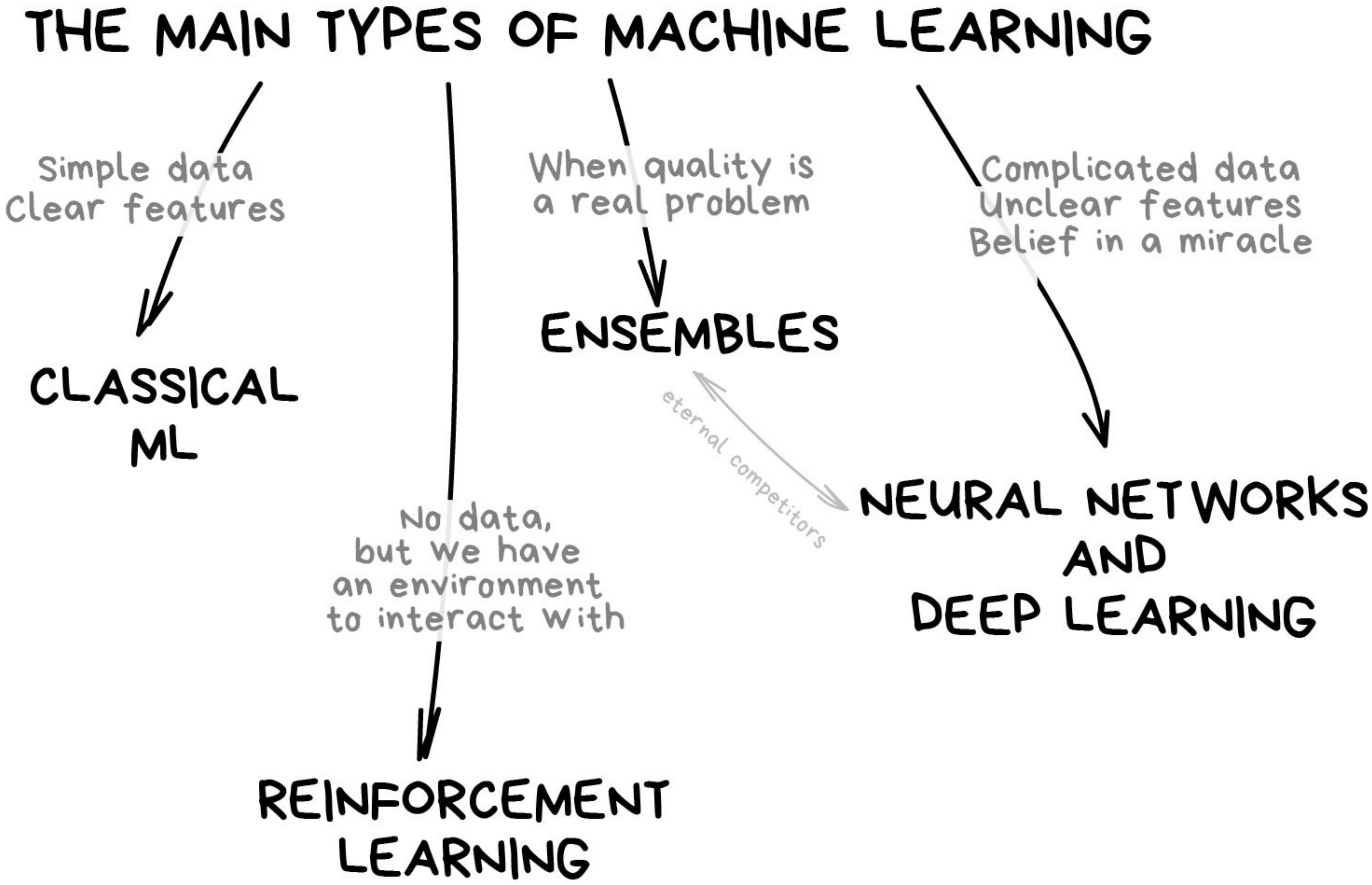
"Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst. We just don't know. So we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it"

…

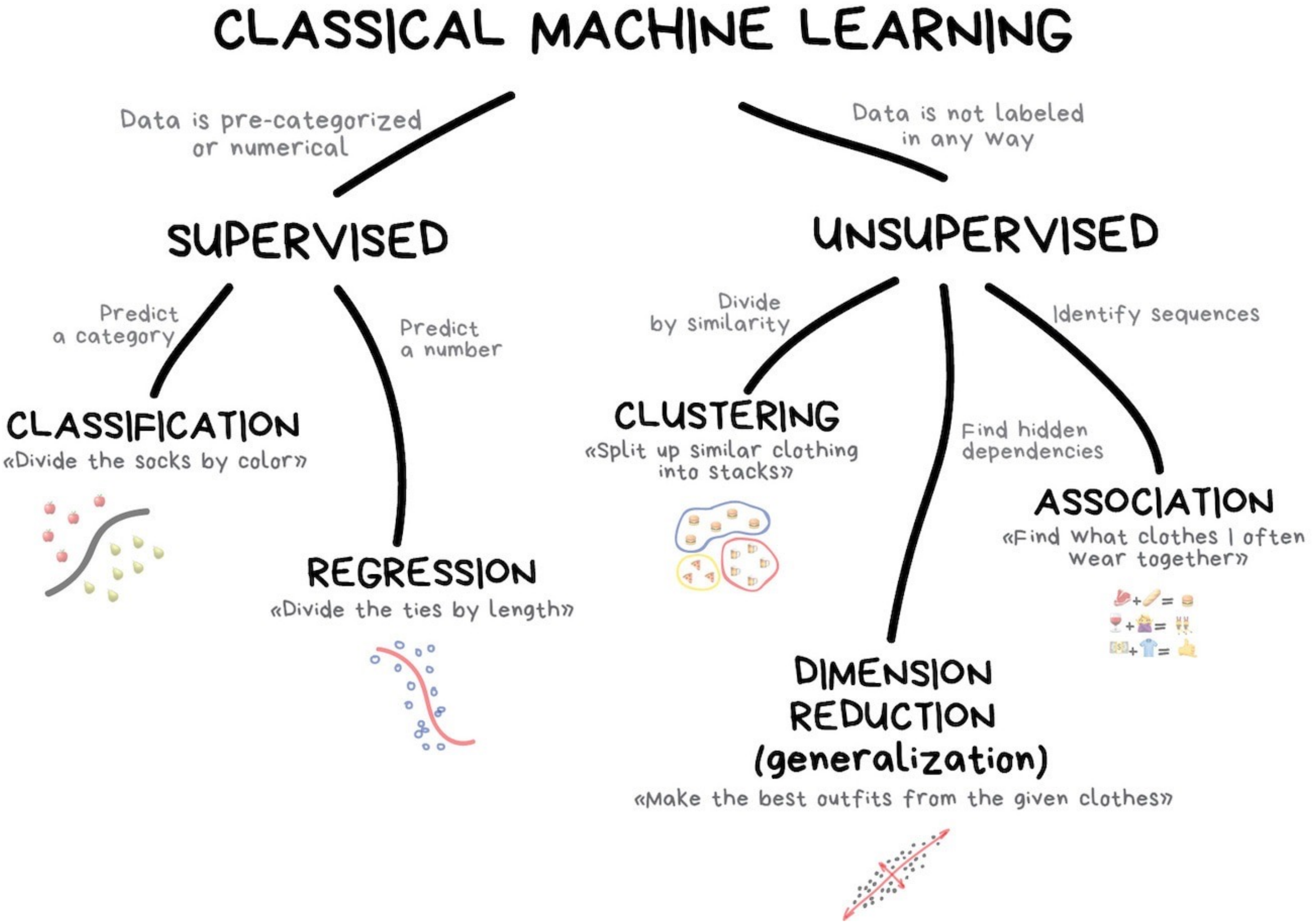
"Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It could bring great disruption to our economy."

-- Leverhulme Centre for the Future of Intelligence (CFI) in Cambridge

Machine learning basics recap



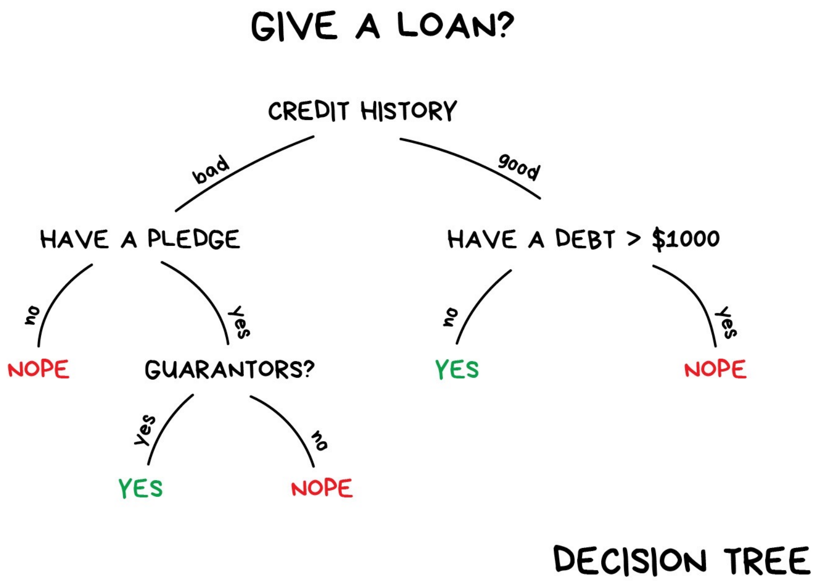
Source: https://vas3k.com/blog/machine_learning/index.html



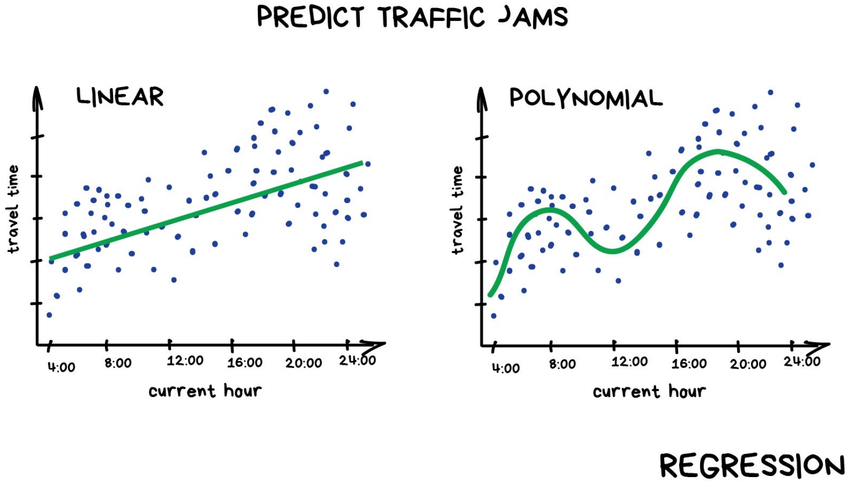
Source: https://vas3k.com/blog/machine_learning/index.html

Supervised learning

- Category prediction (i.e., classification)
  - Task is to assign instances to a discrete class
  - Two classes: binary classification
  - Three or more classes: multiclass classification
  - E.g.:
    - Fraud detection, spam detection, document classification, sentiment prediction, …



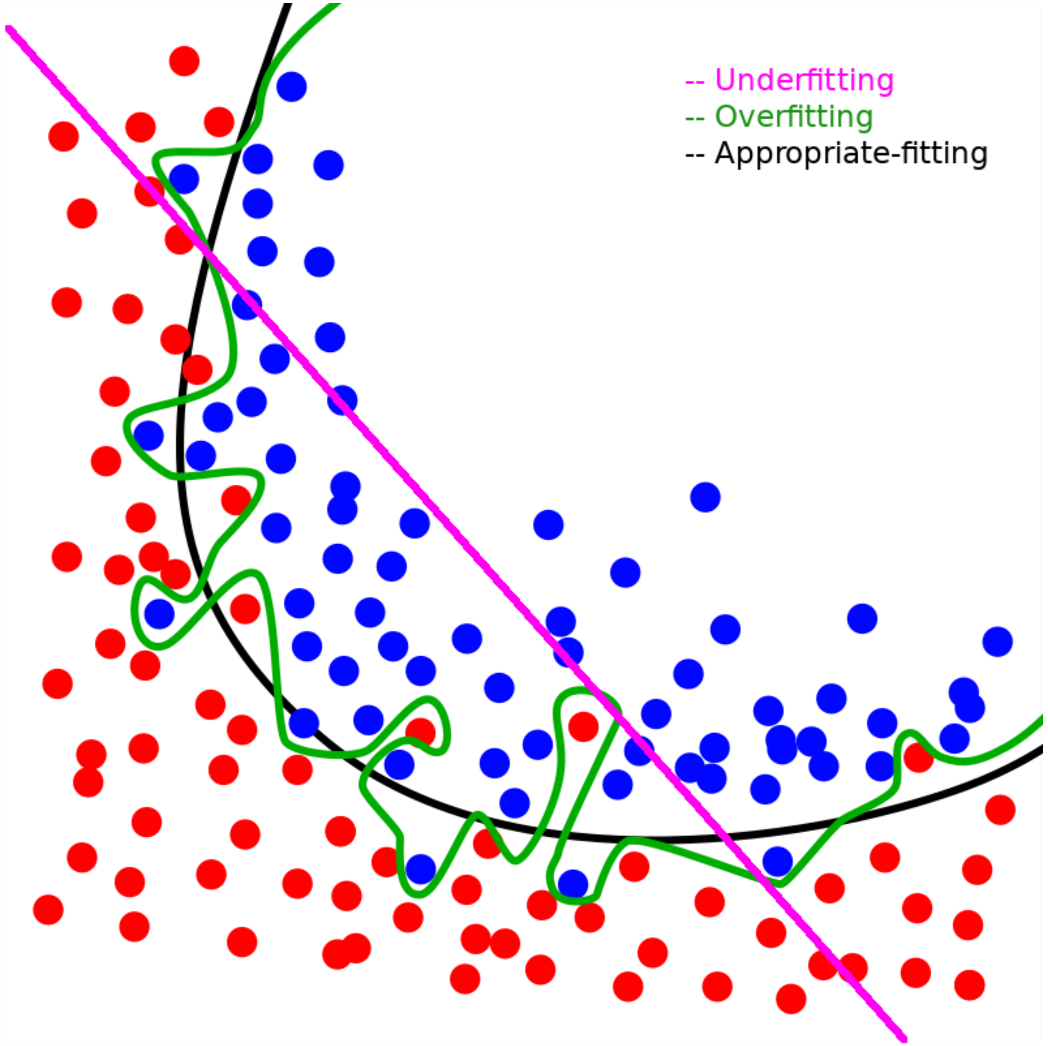
- Numerical prediction (i.e., regression)
  - Task is to assign instances to a numerical value
  - E.g.:
    - Population, stock price, house price, vaccine acceptance, …



Source: https://vas3k.com/blog/machine_learning/index.html

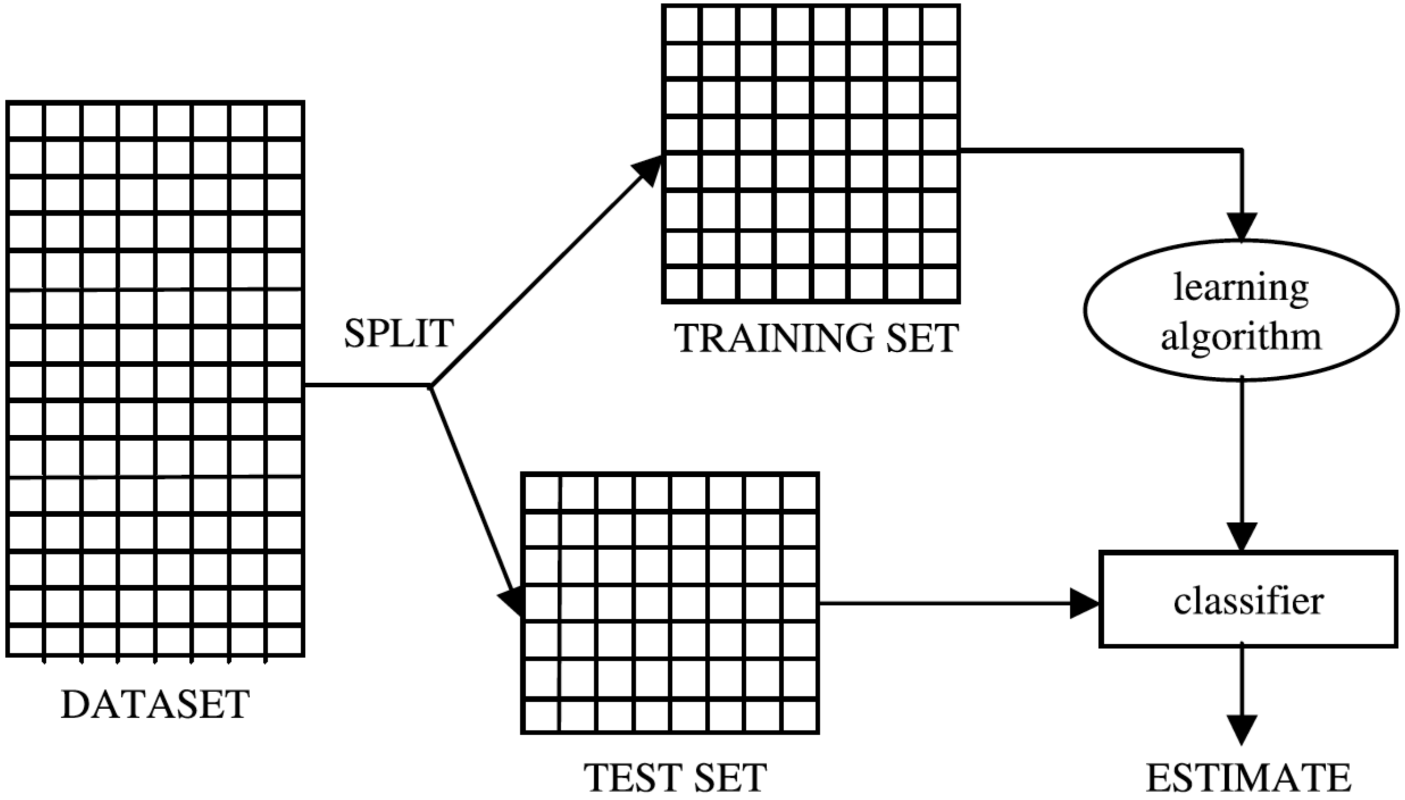
Underfit vs. overfit

- Carefully analyze the model's outputs to evaluate whether they are meeting the goals that we set up for it.



Holdout testing

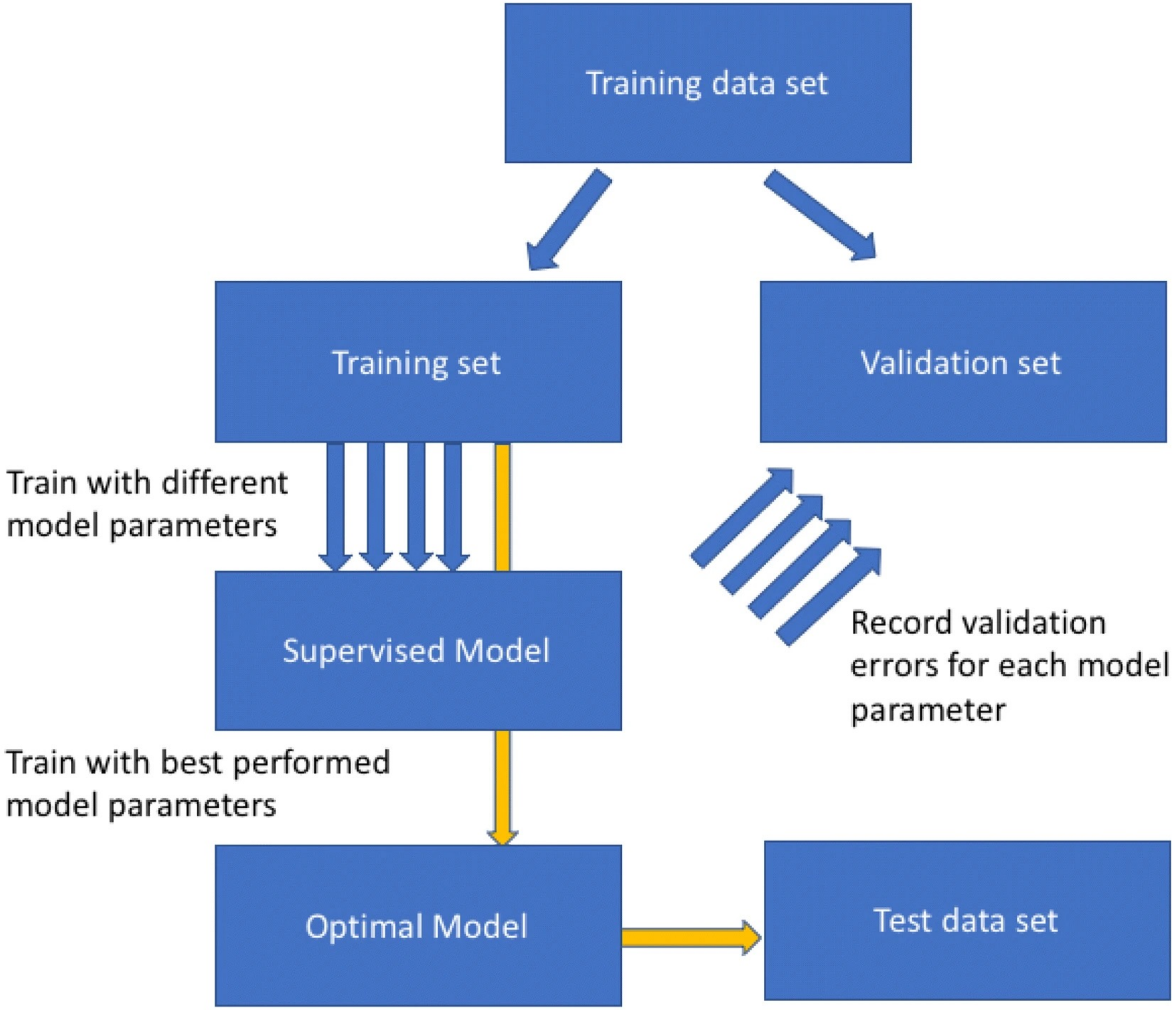
- Train/test split



Source: Bramer, M. (2016). *Principles of data mining* (3rd edition). London: Springer.

Holdout testing

- Train/validation/test split

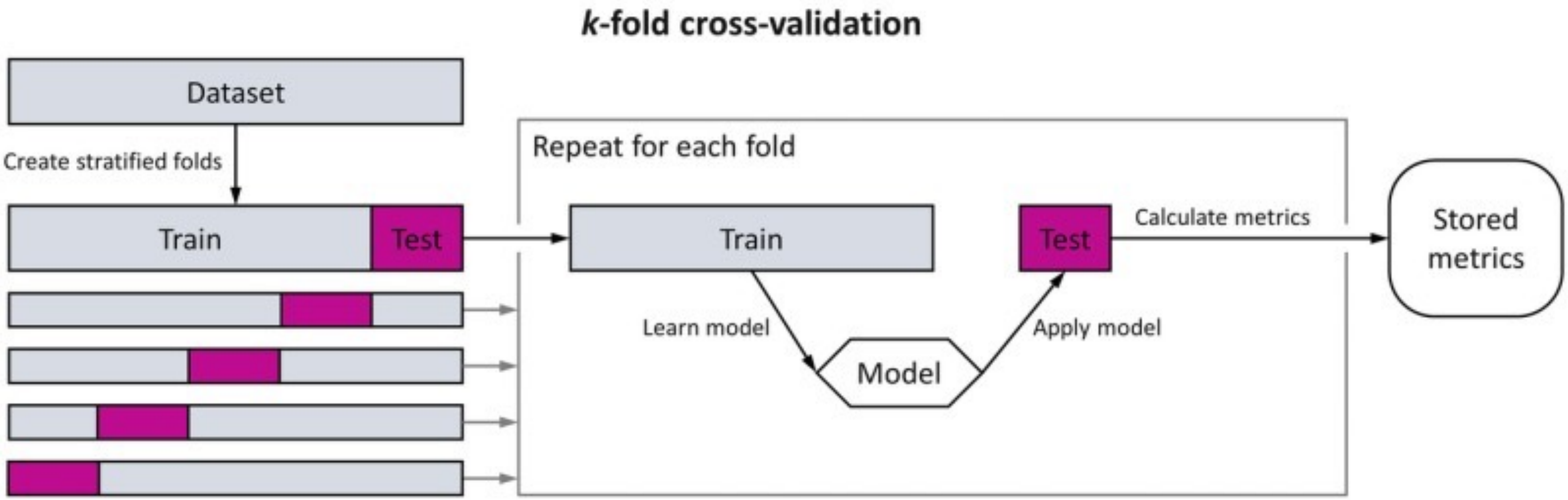


Source: Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249-262.

Cross validation

- When number of instances is small, you want to have less variance in model predictions.
- Often, we use *k-fold cross-validation*
  - Divide *N* instances into *k* equal folds
  - Hold each fold as a testing data and train the model using the remaining *k-1* folds
  - Measure the performance across folds

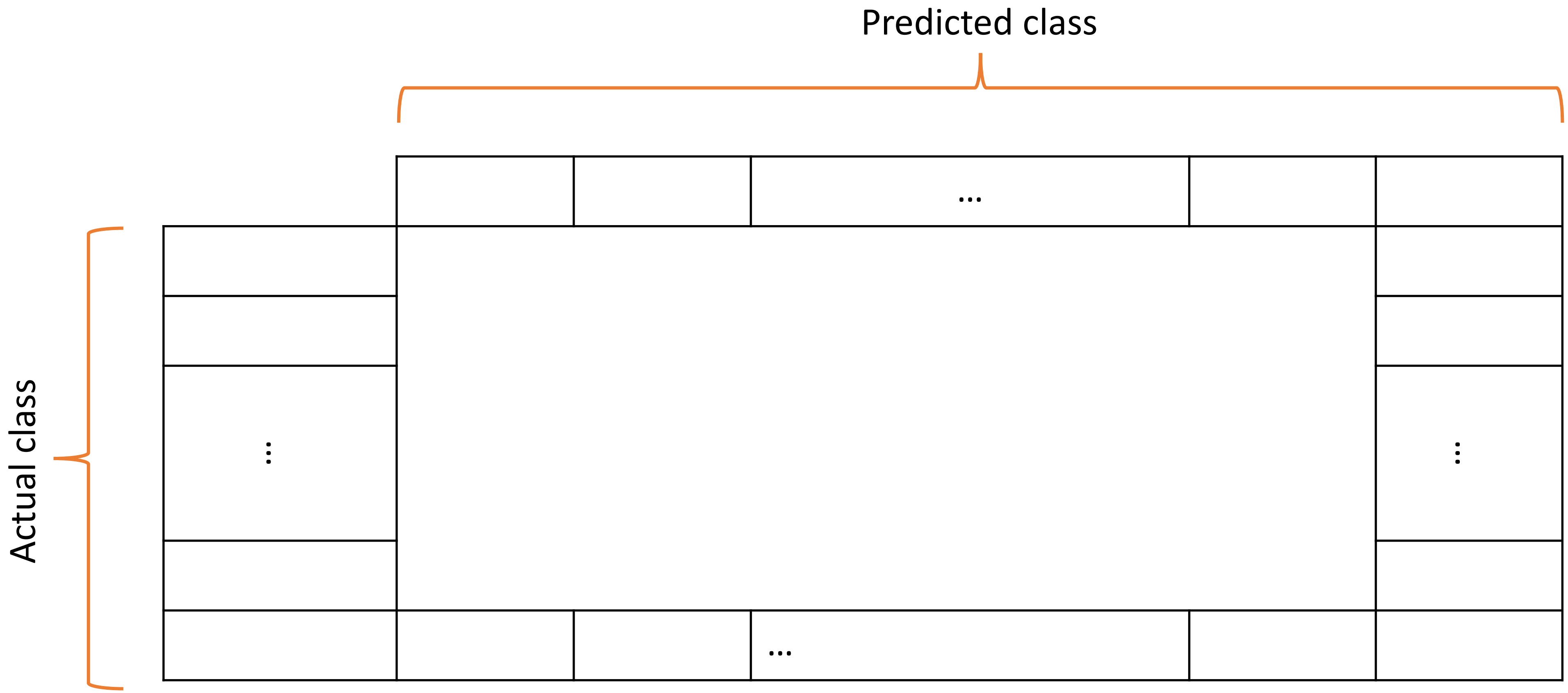
Cross validation



Source: Dankers FJWM, Traverso A, Wee L, et al. Prediction Modeling Methodology. 2018 Dec 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543534/ doi: 10.1007/978-3-319-99713-1_8

Confusion matrix

- Compactly shows a classifier performance



Confusion matrix

- Examples

Correct classification	Classified as	
	democrat	republican
democrat	81 (97.6%)	2 (2.4%)
republican	6 (11.5%)	46 (88.5%)

Correct classification	Classified as					
	1	2	3	5	6	7
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
5	0	2	0	10	0	1
6	0	1	0	0	7	1
7	1	3	0	1	0	24

Source: Bramer, M. (2016). *Principles of data mining* (3rd edition). London: Springer.

Confusion matrix: metrics

		Predicted class	
		C=True	C=False
Actual class	C=True	**TP**	**FN**
	C=False	**FP**	**TN**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity\ (recall) = \frac{TP}{TP + FN}$$

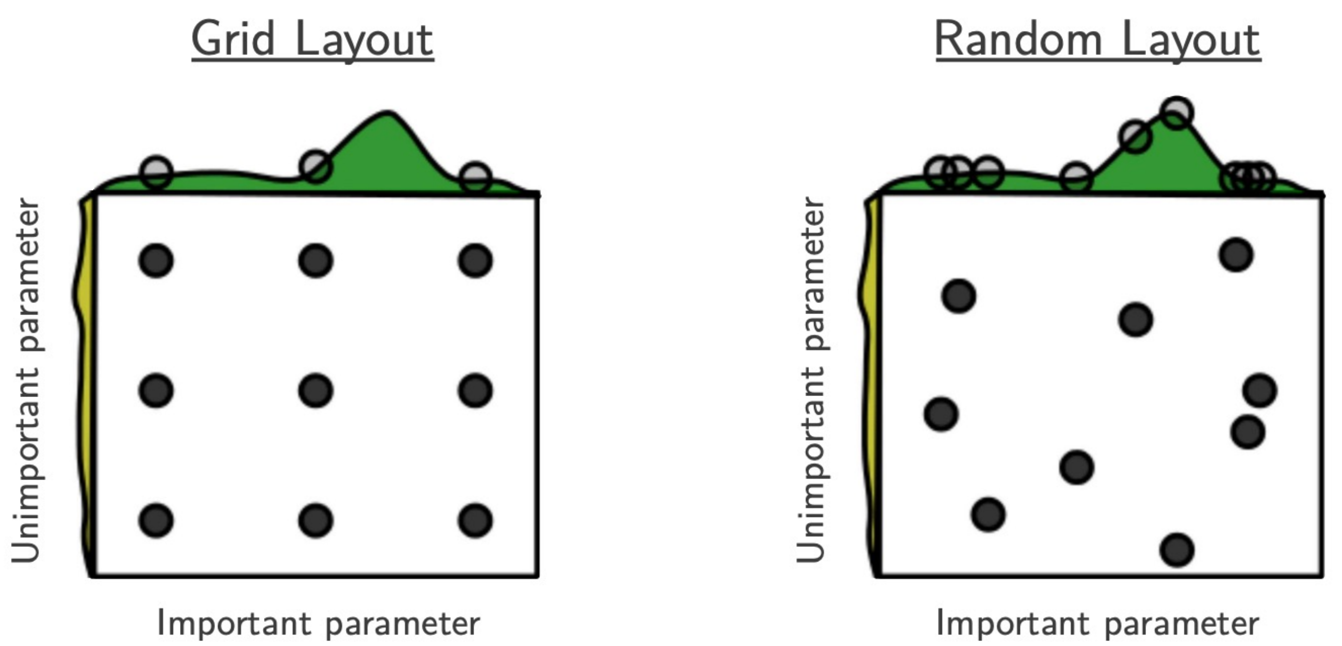
$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

Hyperparameter optimization

- The process of finding hyperparameters that improves model fit.

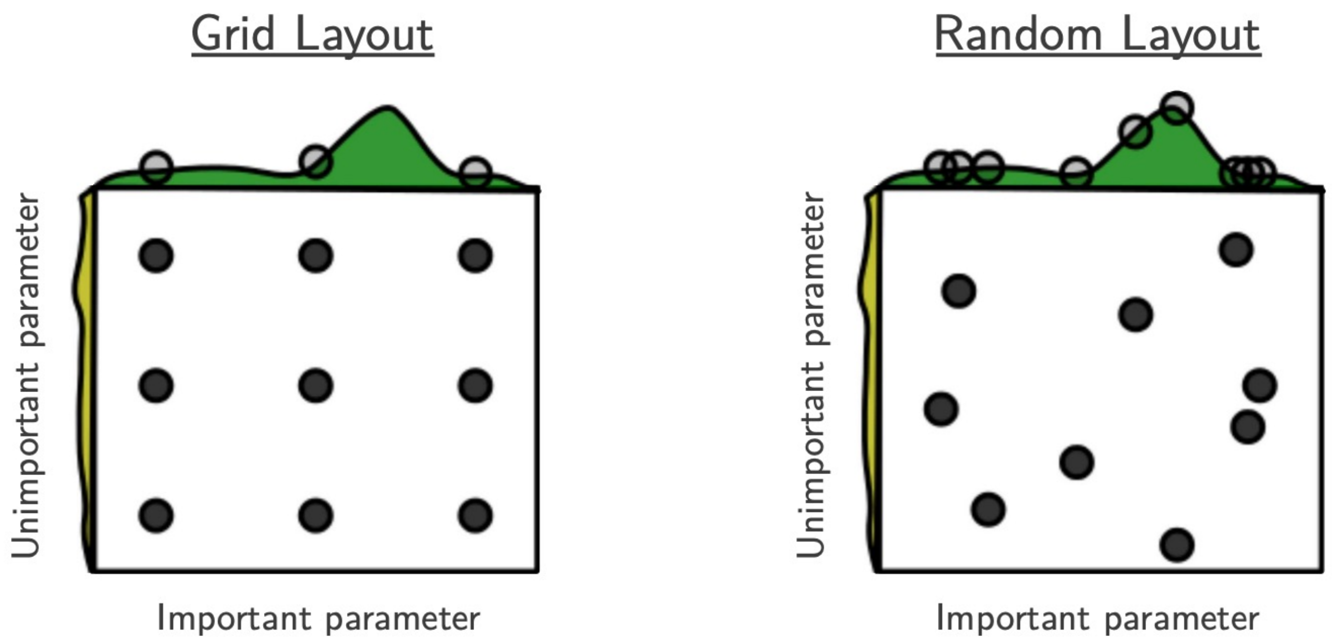


Source: Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Hyperparameter optimization

- The process of finding hyperparameters that improves model fit.

Does this process remind you of anything from our previous classes?

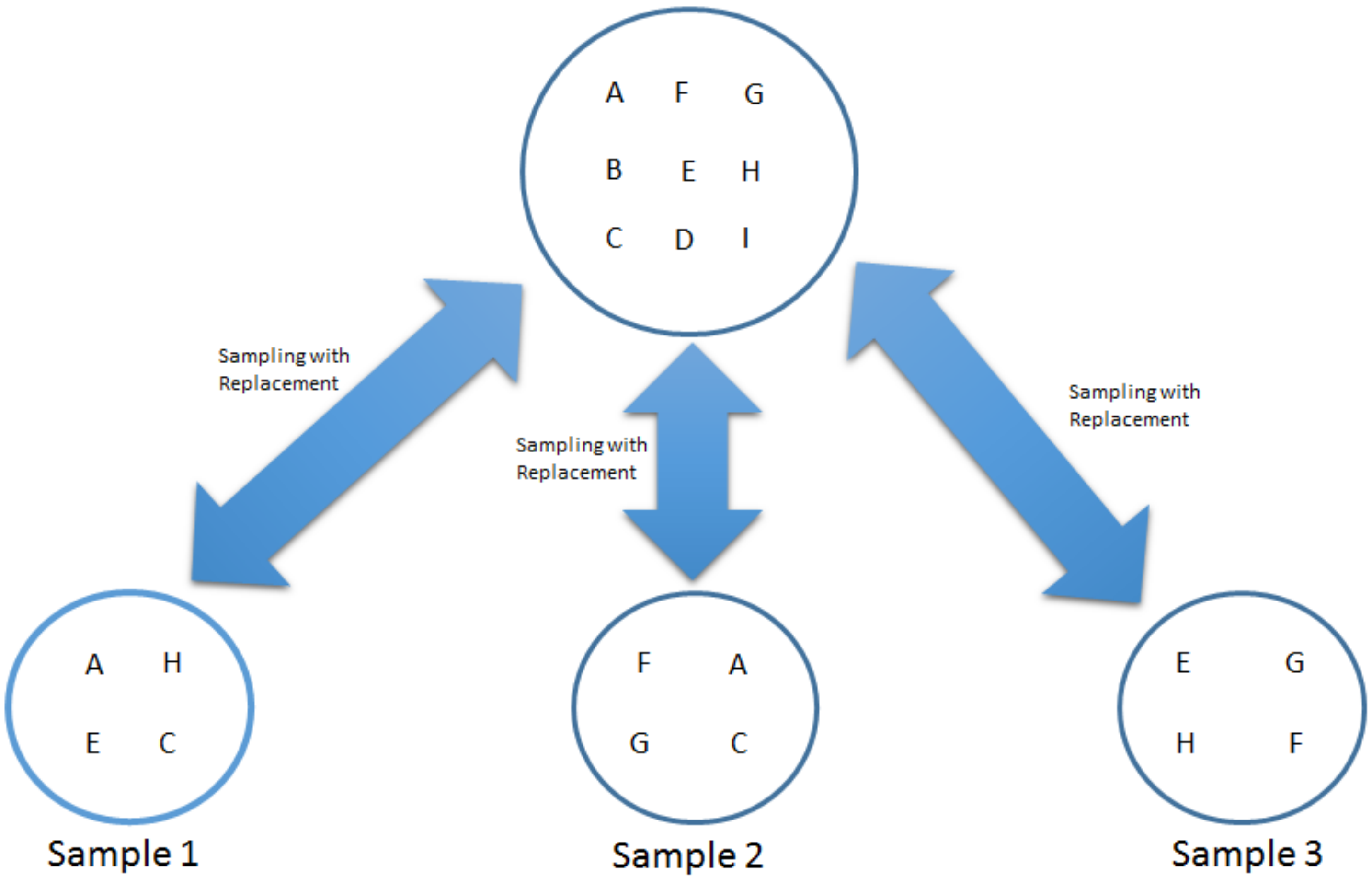


Source: Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Bootstrapping

- In cross validation, each instance will be used once.
- Bootstrapping allows sampling the dataset with replacement.
- In general, it’s not more robust than cross validation.
- Often used in training/testing ensemble ML models.

Bootstrapping

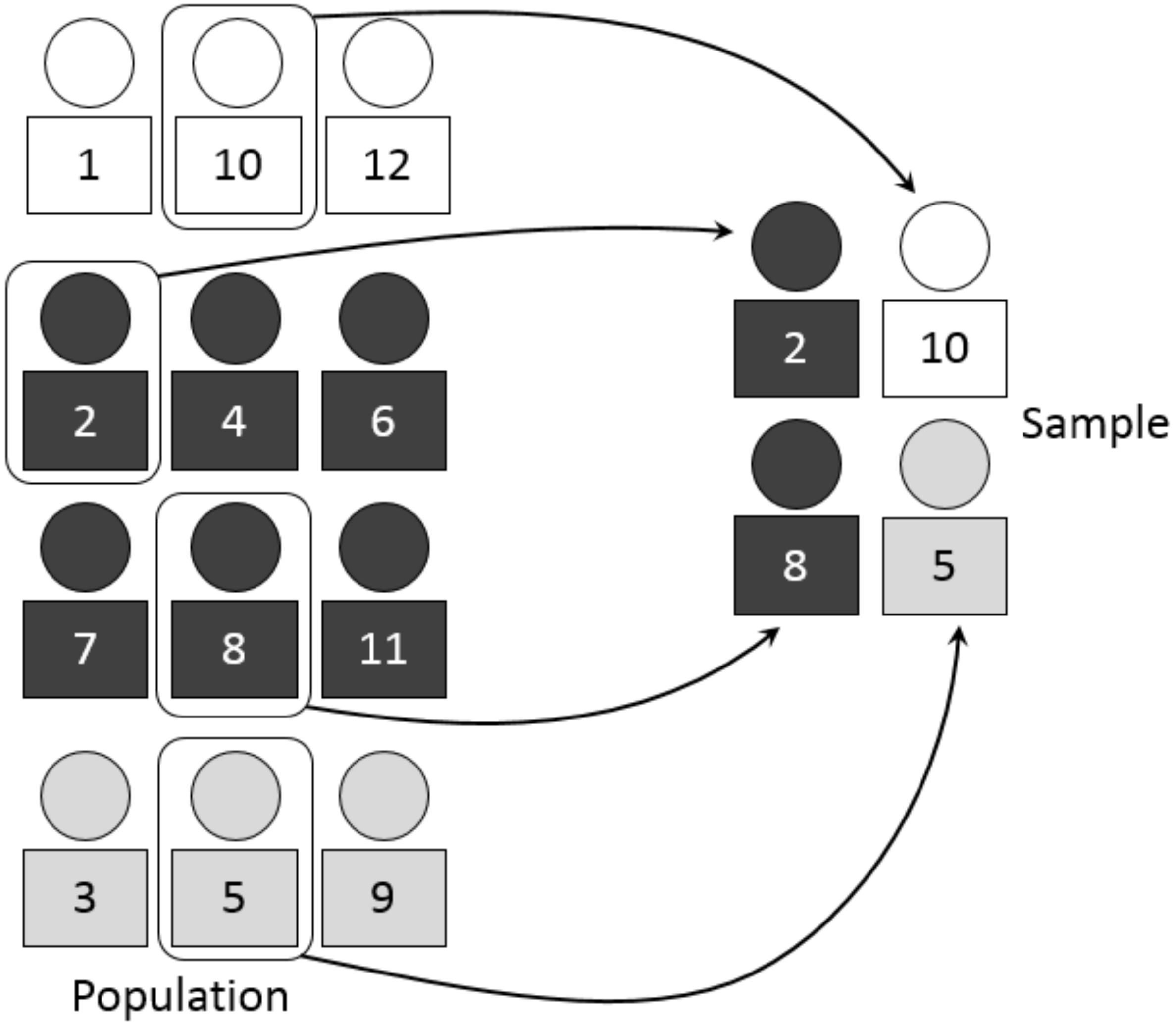


Source: Kumar, R. (2019). *Machine Learning Quick Reference: Quick and essential machine learning hacks for training smart data models*. Packt Publishing Ltd.

Stratified sample

- It is used to eliminate bias in the dataset.
- Assumes that **you know** the true underlying population distributions and your test does not follow that distribution (hence biased).
- Not specific to ML tasks but it's useful

How to create a stratified sample



Source: https://www.netquest.com/blog/en/random-sampling-stratified-sampling

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?
- How do you handle challenges?

Imbalanced data cases

- For instance, you have a fraud detection dataset with 3% fraud & 97% normal transactions.

- What will your ML model do?
- How do you handle challenges?

1. Resample
2. Use model-specific handling of imbalance

Machine learning in the wild

Some examples

Motorist fined after CCTV confuses his number plate with woman's T-shirt

David Knight told to pay £90 after KN19TER registration is mixed up with pedestrian's 'Knitter' top 120 miles away

The vehicle was seen in (location) Pulteney Bridge, Bath on 29/07/2021 at 15:41

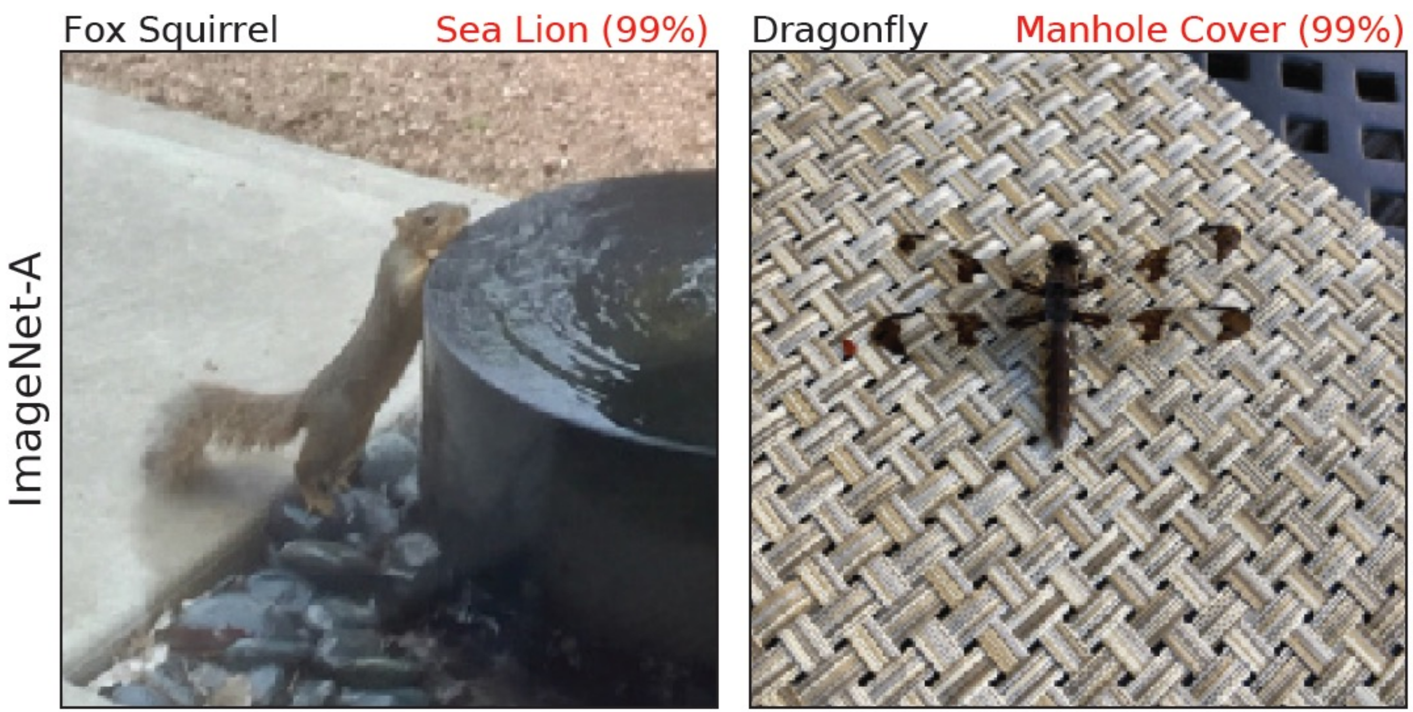
Bath and North East Somerset Council believes that a Penalty Charge is now payable with respect to the vehicle above, for the following alleged contravention:- **34 - Being in a bus lane** (as defined in S.144(5) Transport Act 2000).

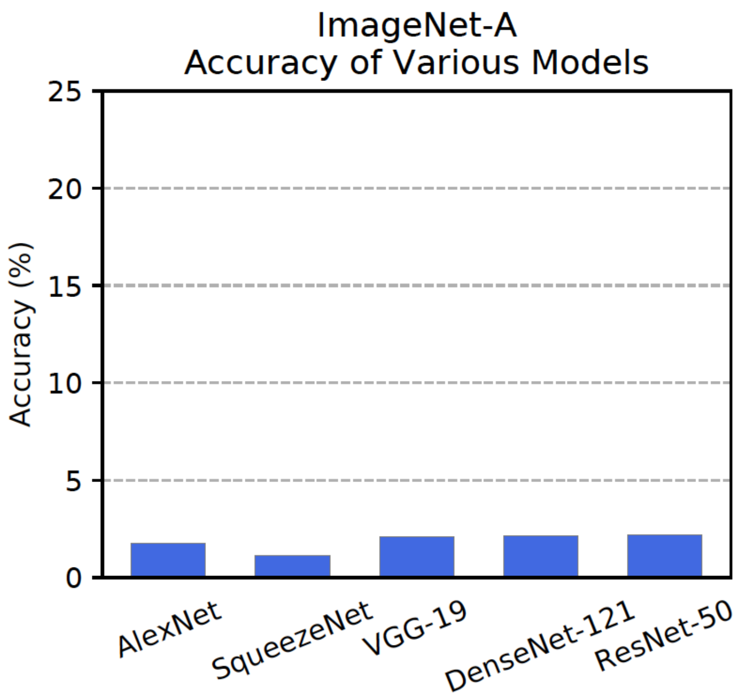
YOU MUST NOT IGNORE THIS NOTICE OR PASS IT TO THE DRIVER

The Guardian

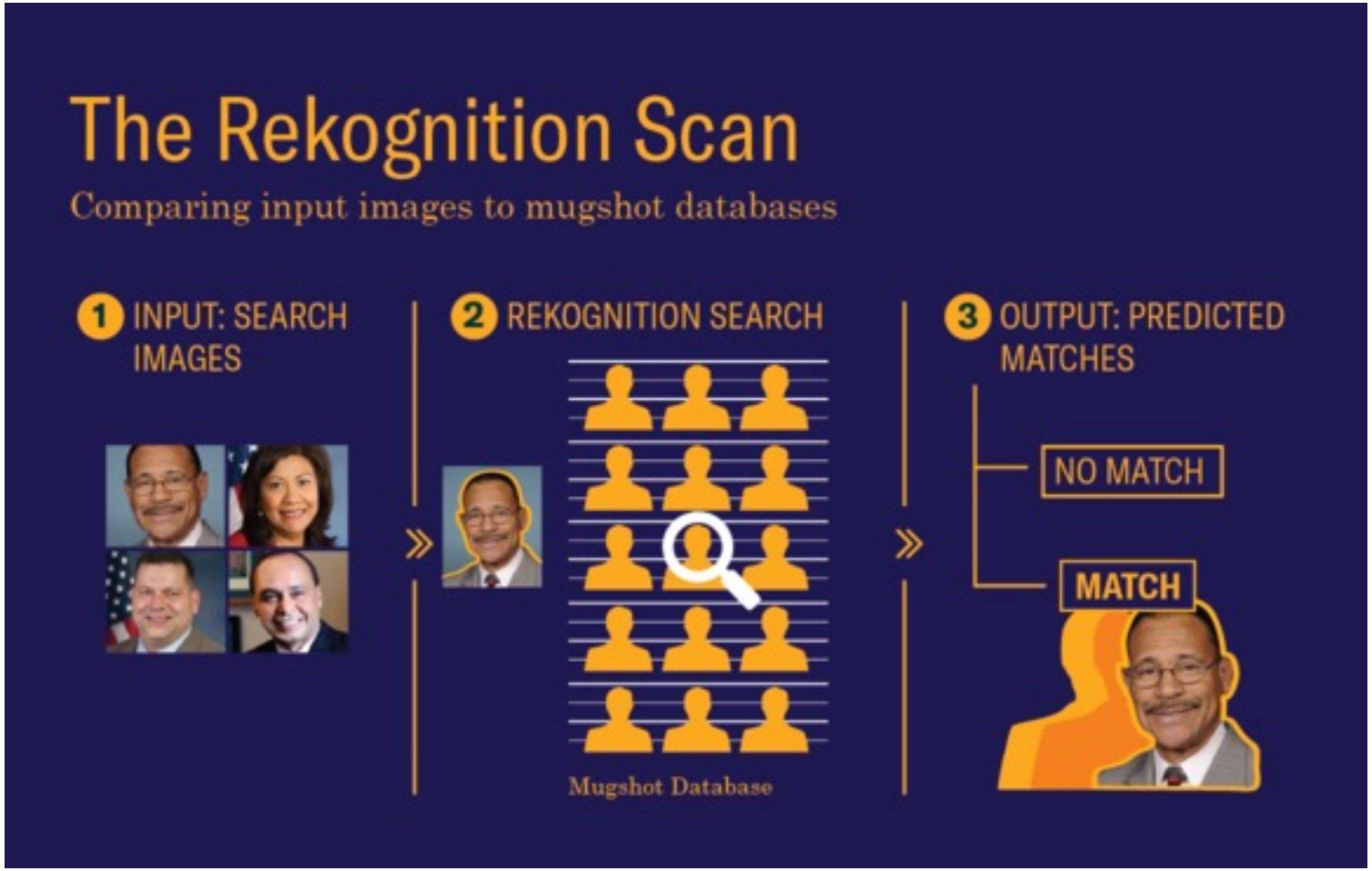
Source: https://www.theguardian.com/uk-news/2021/oct/18/motorist-fined-number-plate-t-shirt

Machine learning being confidently wrong

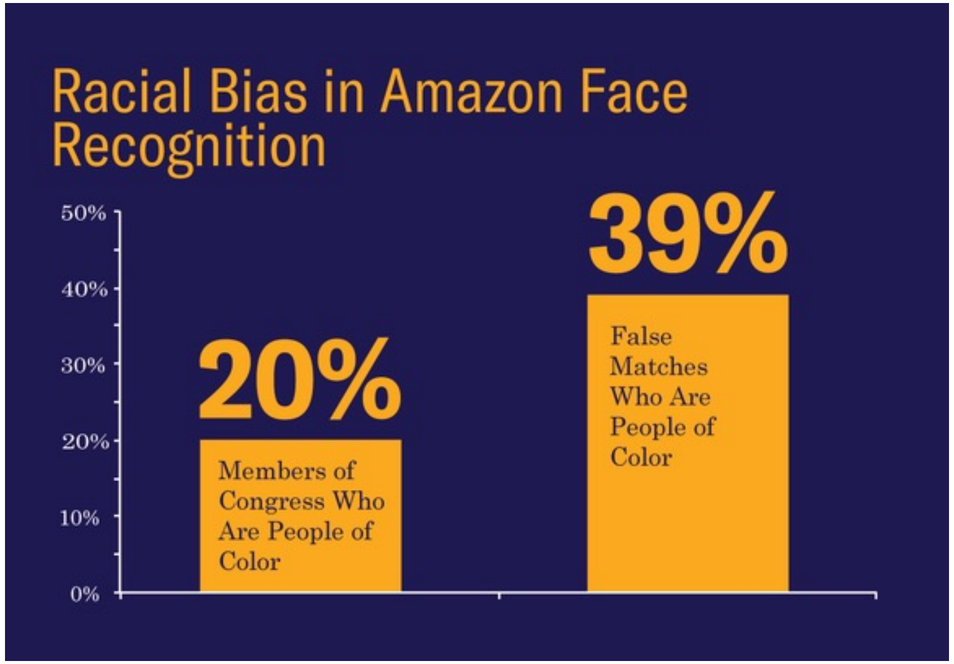




Source: Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15262-15271).



Source: https://www.aclunc.org/blog/amazon-s-face-recognition-falsely-matched-28-members-congress-mugshots



Source: https://www.aclunc.org/blog/amazon-s-face-recognition-falsely-matched-28-members-congress-mugshots

Source: https://twitter.com/ChelseaParlett/status/1455209848937672710/photo/1

IBM's Watson gave unsafe recommendations for treating cancer

THE VERGE

Doctors fed it hypothetical scenarios, not real patient data

By Angela Chen | @chengela | Jul 26, 2018, 4:29pm EDT

"according to IBM documents dated from last summer, the supercomputer has frequently given bad advice, like when it suggested a cancer patient with severe bleeding be given a drug that could cause the bleeding to worsen. (A spokesperson for Memorial Sloan Kettering said this suggestion was hypothetical and not inflicted on a real patient.)…

…the suggestions Watson made were simply based off the treatment preferences of the few doctors providing the data, not actual insights it gained from analyzing real cases…."

Source: https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science

Twitter taught Microsoft's AI chatbot to be a racist in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source TayandYou (Twitter)

THE VERGE

Listen to this article

SHARE

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

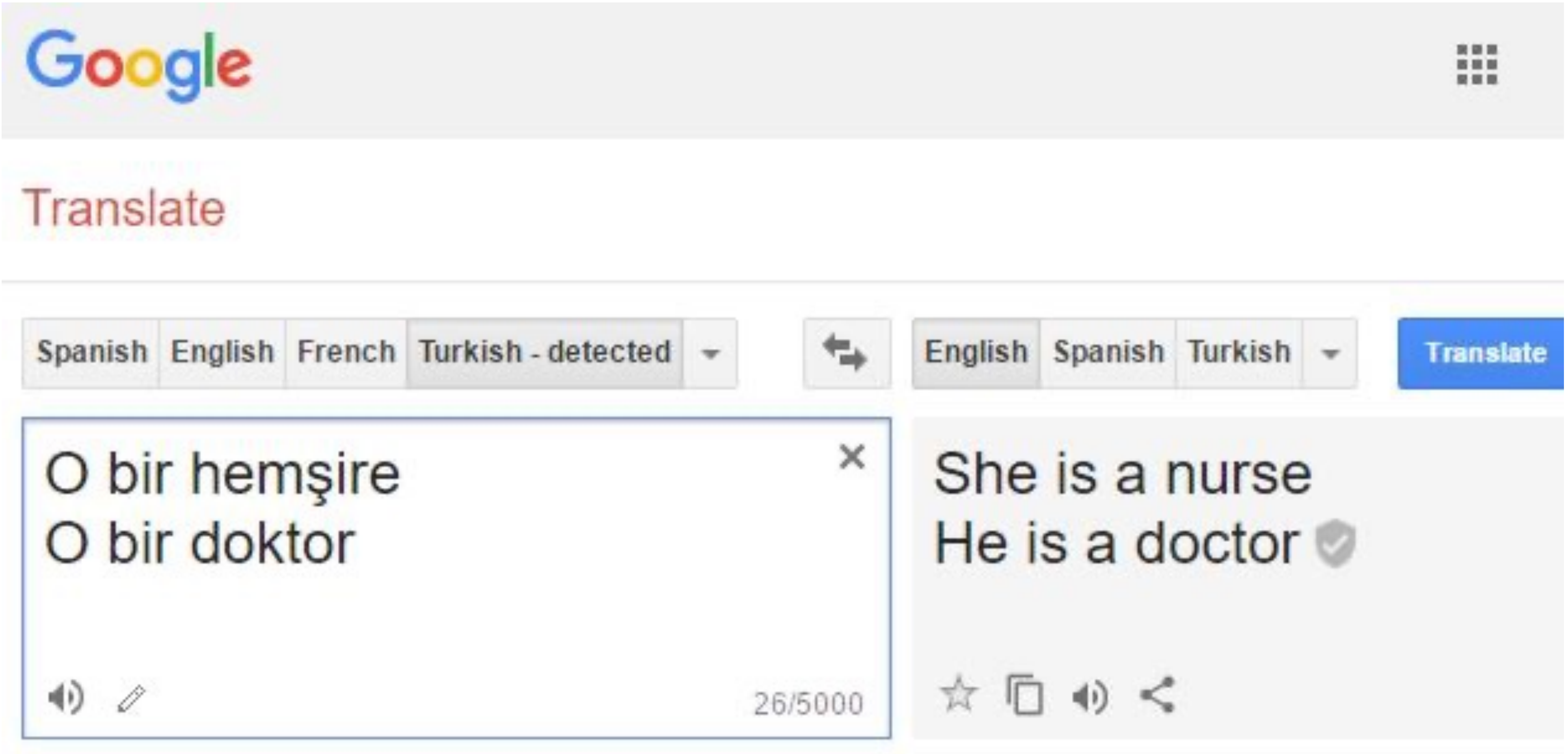
TayTweets
@TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

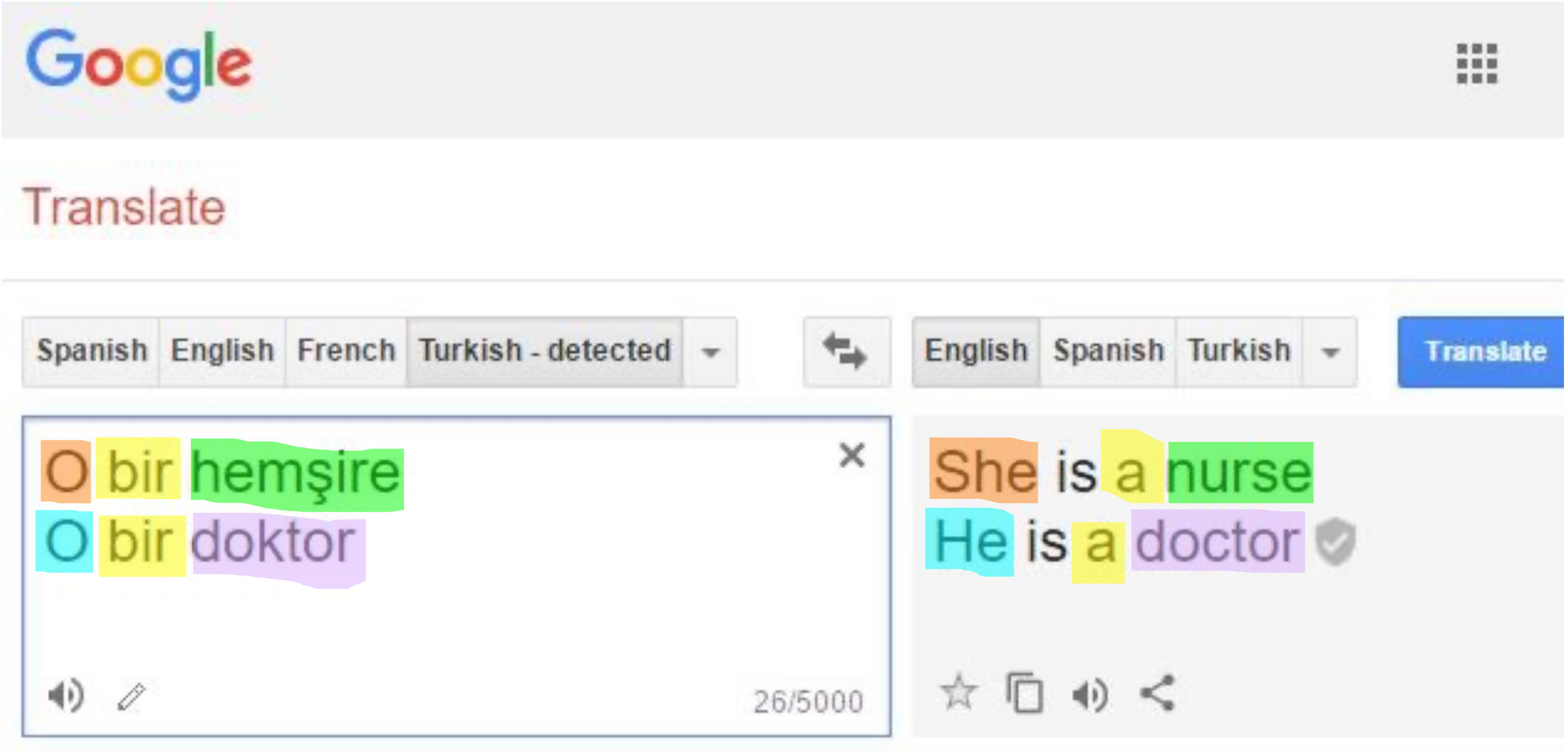
Source: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Google Translate **then**



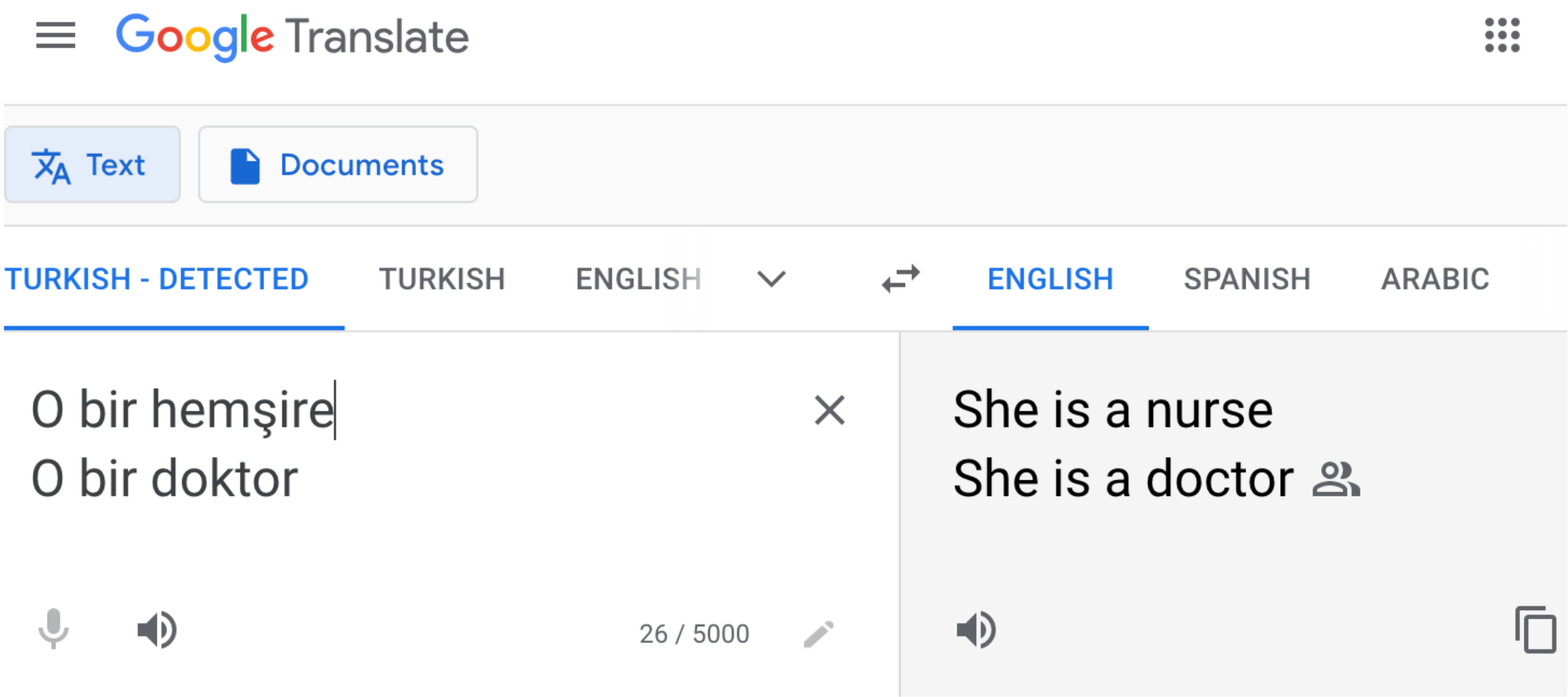
Source: https://i-programmer.info/news/105-artificial-intelligence/10688-investigating-bias-in-ai-language-learning.html

Google Translate **then**

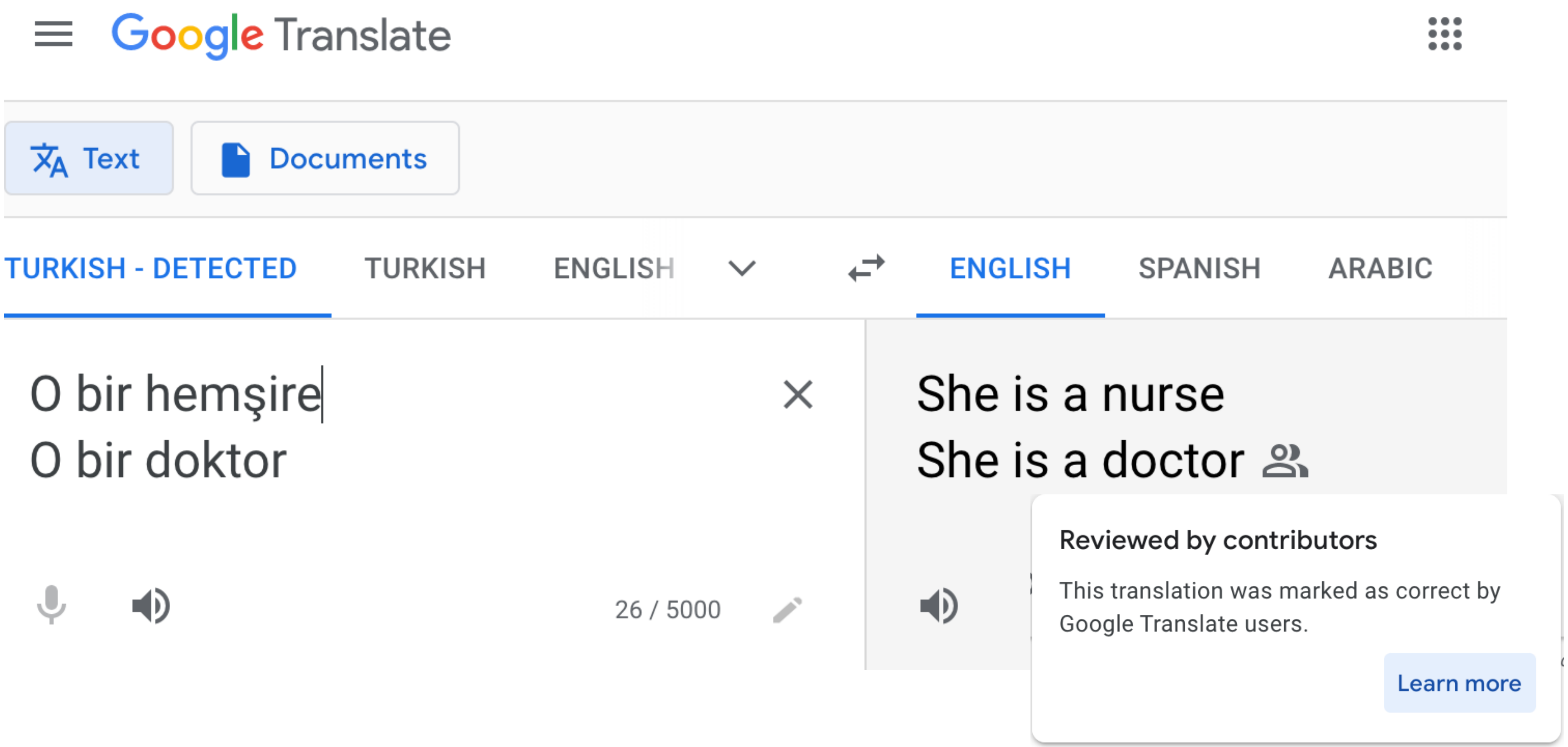


Source: https://i-programmer.info/news/105-artificial-intelligence/10688-investigating-bias-in-ai-language-learning.html

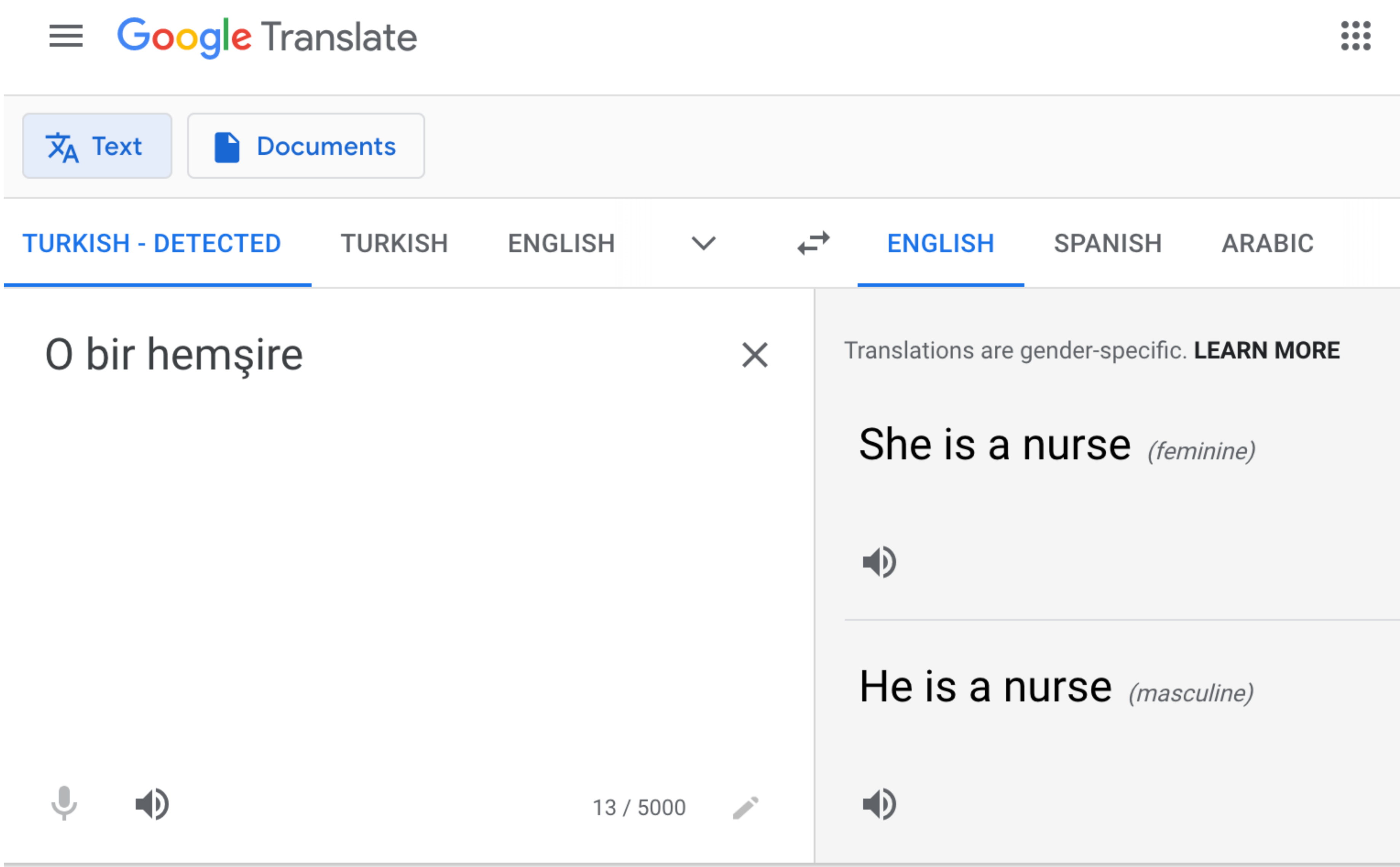
Google Translate <u>now</u>



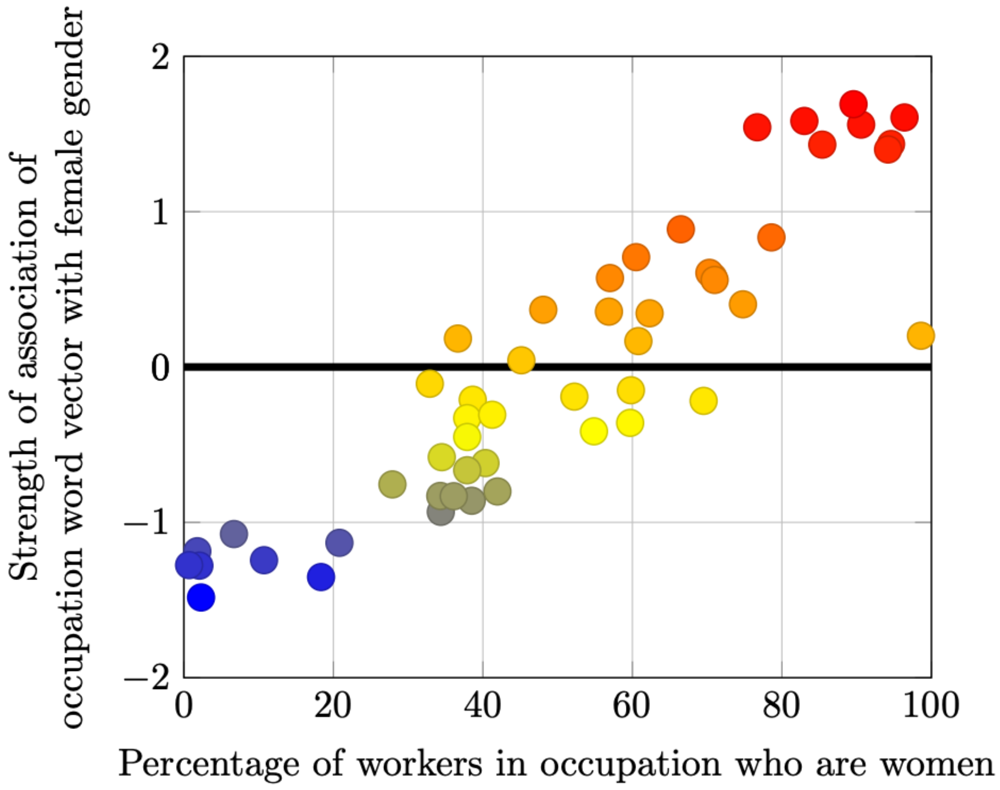
Google Translate **now**

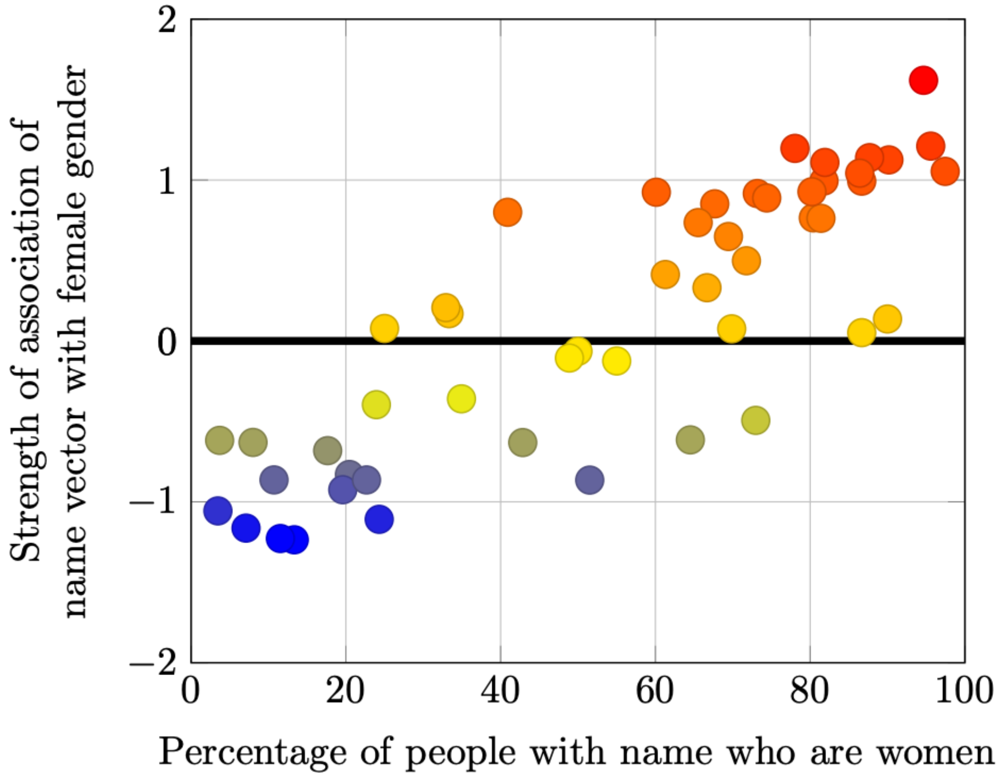


Google Translate now



Gender bias in language models





Source: Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Is machine learning safe for authentication?

https://www.youtube.com/watch?v=ZwCNG9KFdXs

THE WALL STREET JOURNAL.

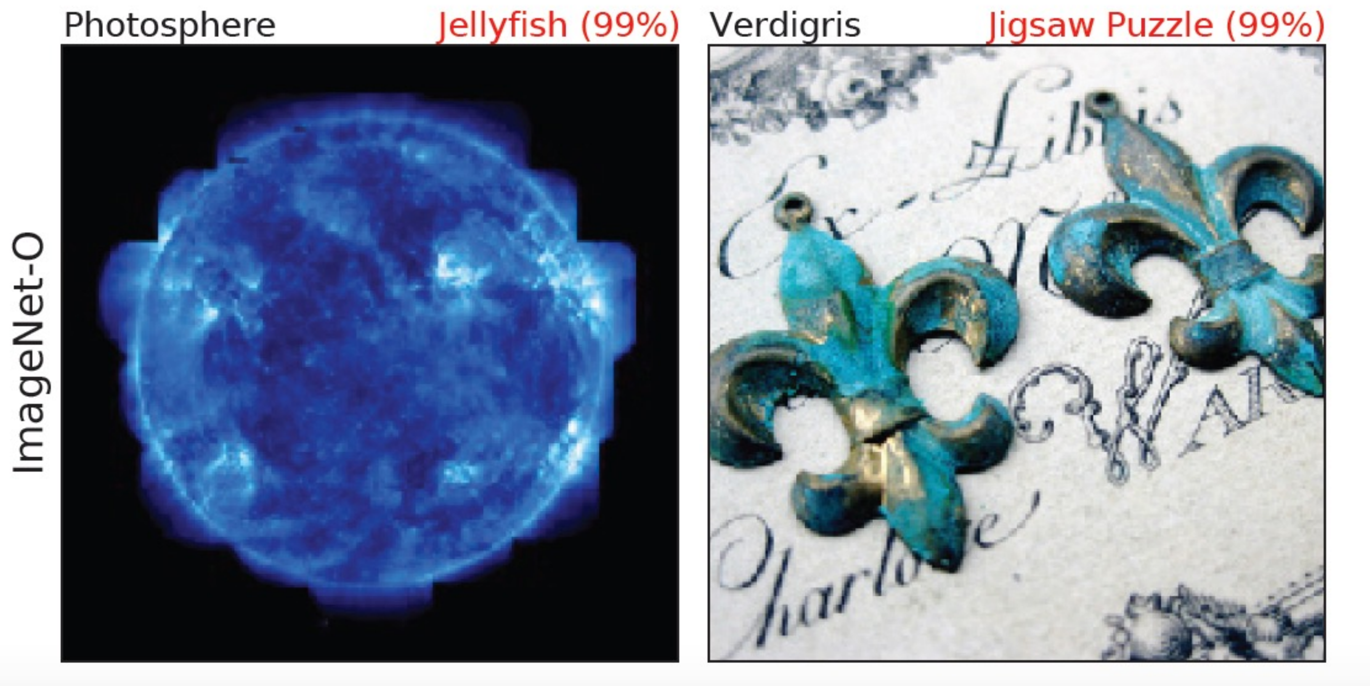
Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

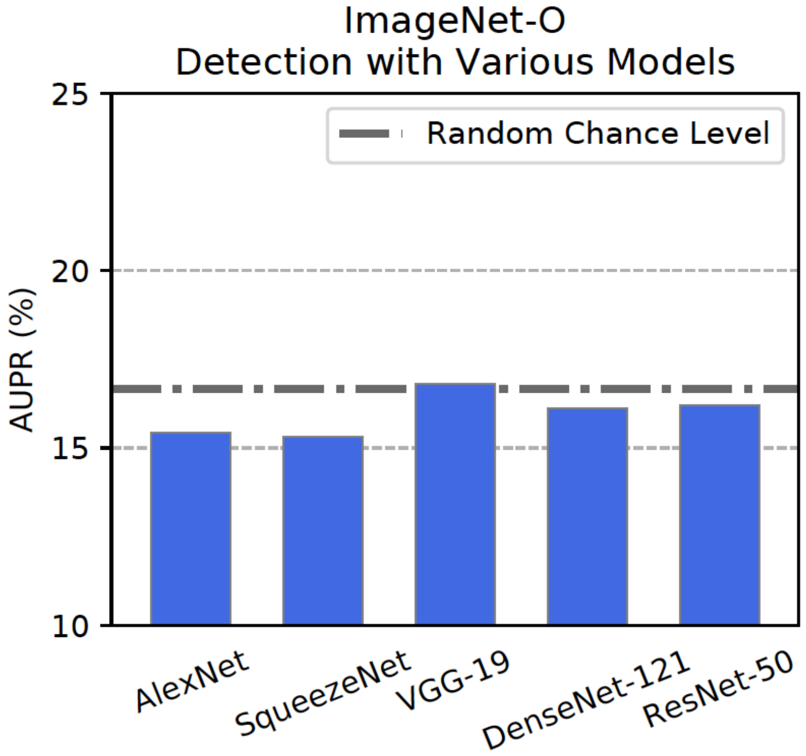
Scams using artificial intelligence are a new challenge for companies

The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send the funds to a Hungarian supplier. The caller said the request was urgent, directing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group SA.

Source: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Adversarial machine learning





Source: Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15262-15271).

Machine learning-based graders

- The *e-rater*® automated scoring engine - Educational Testing Service (ETS)

Subgroup	*N*	Mean (*SD*)		
		Operational e-rater score	Operational human score	Mean diff. (e-rater, human)
Issue				
Overall	103,151	3.73 (0.86)	3.74 (0.86)	−0.004 (0.58)
China	4,005	3.40 (0.72)	2.96 (0.58)	0.44 (0.64)
Argument				
Overall	115,071	3.60 (0.99)	3.61 (0.99)	−0.002 (0.67)
China	4,923	3.47 (0.71)	3.09 (0.65)	0.37 (0.68)
Taiwan	761	2.70 (0.84)	2.87 (0.65)	−0.17 (0.65)
African American	6,879	3.06 (1.06)	3.19 (0.93)	−0.13 (0.71)

Source: Ramineni, C., & Williamson, D. (2018). Understanding Mean Score Differences Between the e-rater® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE® General Test. *ETS Research Report Series*, *2018*(1), 1-31.

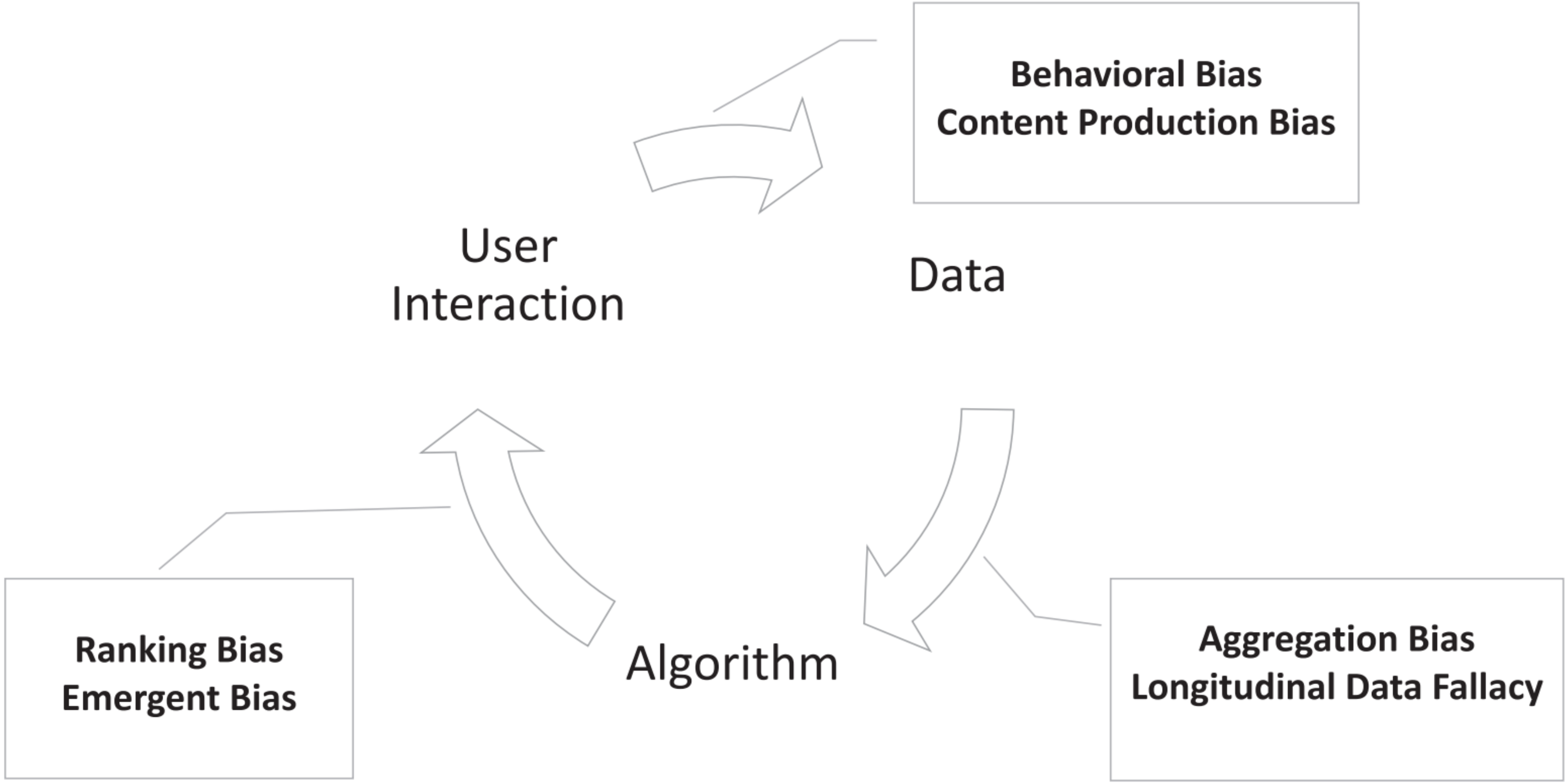
Terminology

- **Bias:** “an inclination of temperament or outlook… *especially* **:** a personal and sometimes unreasoned judgment..”
- **Fairness:** “the quality or state of being fair… lack of favoritism toward one side or another..”
- **Assurance:** “the state of being assured: such as a being certain in the mind [or] confidence of mind or manner …”

Source: *Merriam-Webster.com Dictionary*, Merriam-Webster, https://www.merriam-webster.com

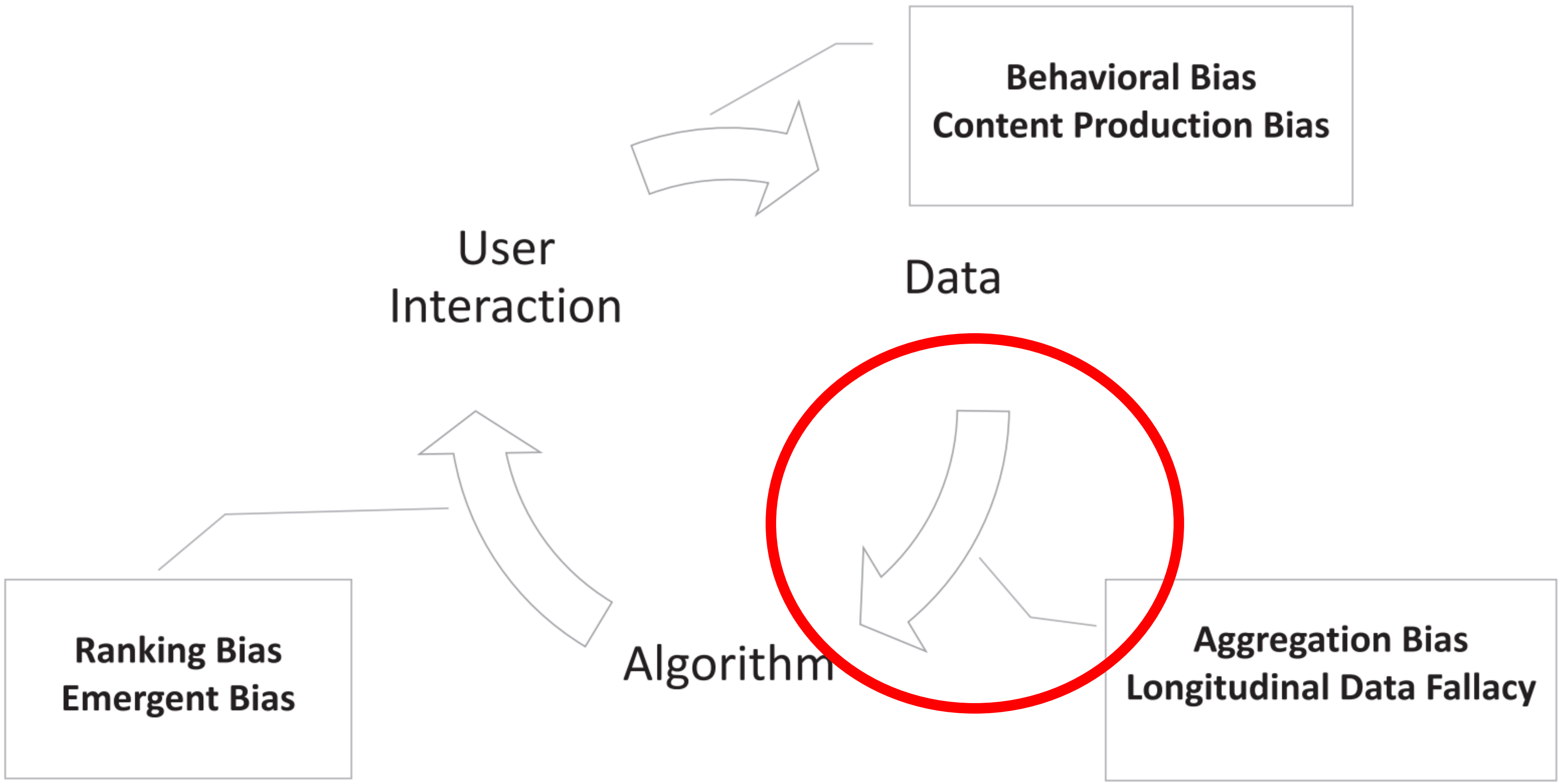
- Assurance in AI ensures “outcomes that are valid, trustworthy, and ethical, unbiased in its learning and fair to its users” (Batarseh et al. 2021).

Bias in Data, Algorithms, and User Experiences



Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Bias in Data, Algorithms, and User Experiences



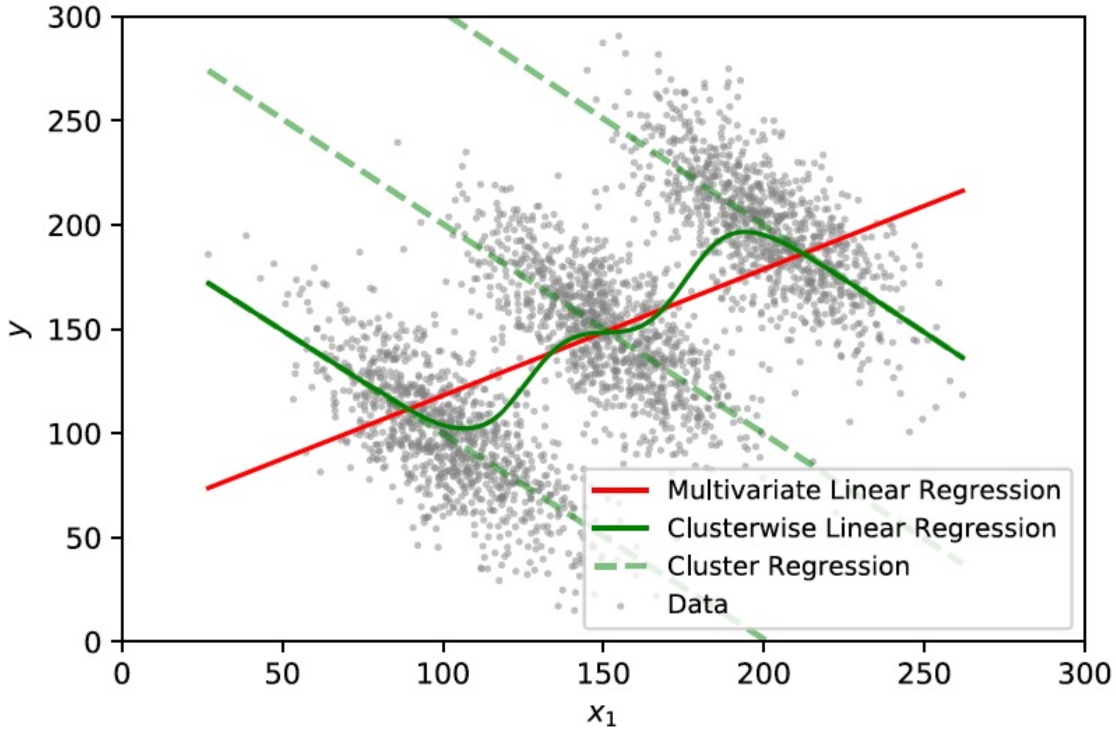
Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

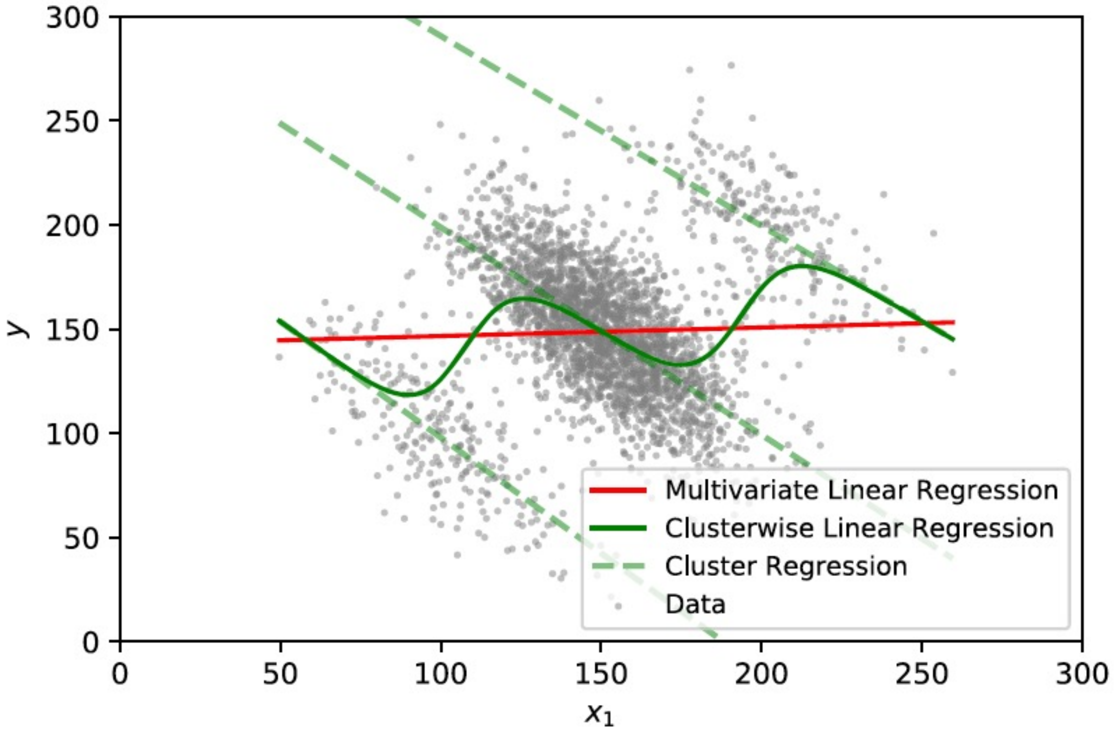
Bias from data to algorithm

Mehrabi et al. 2021

- **Measurement Bias.** "Measurement, or reporting, bias arises from how we choose, utilize, and measure particular features"
- **Omitted Variable Bias.** "… occurs when one or more important variables are left out of the model"
- **Representation Bias.** "… arises from how we sample from a population during data collection process"
- **Aggregation Bias.** "… (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population"
  - Simpson's Paradox.: things observed in aggregated data disappears or reverses when the same data is disaggregated
  - Modifiable Areal Unit Problem: a statistical bias in geospatial analysis, which arises when modeling data at different levels of spatial aggregation

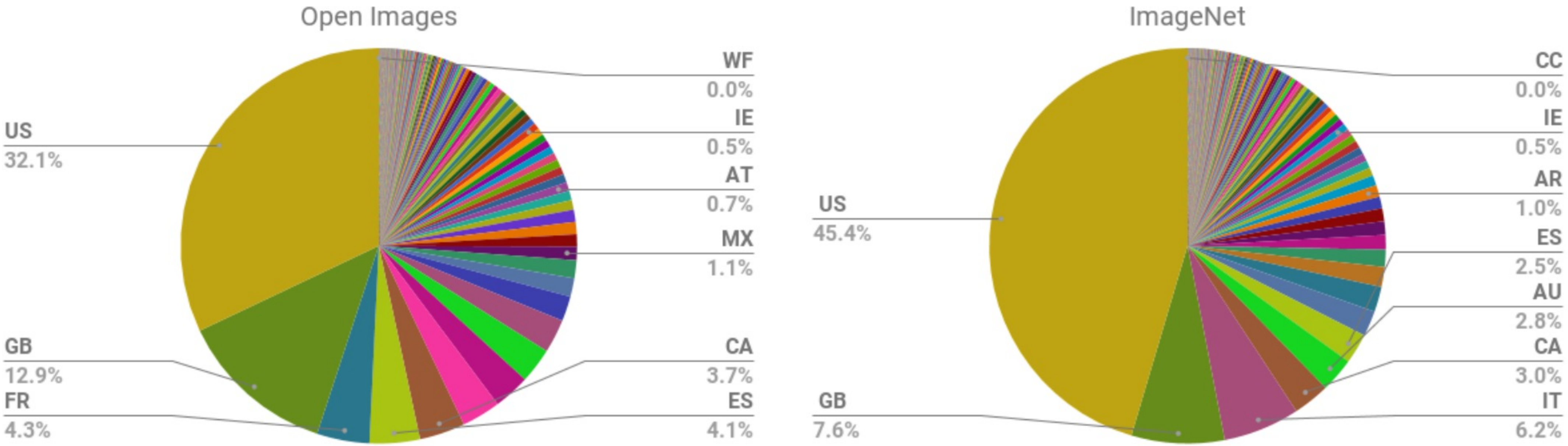
Representation bias example





Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

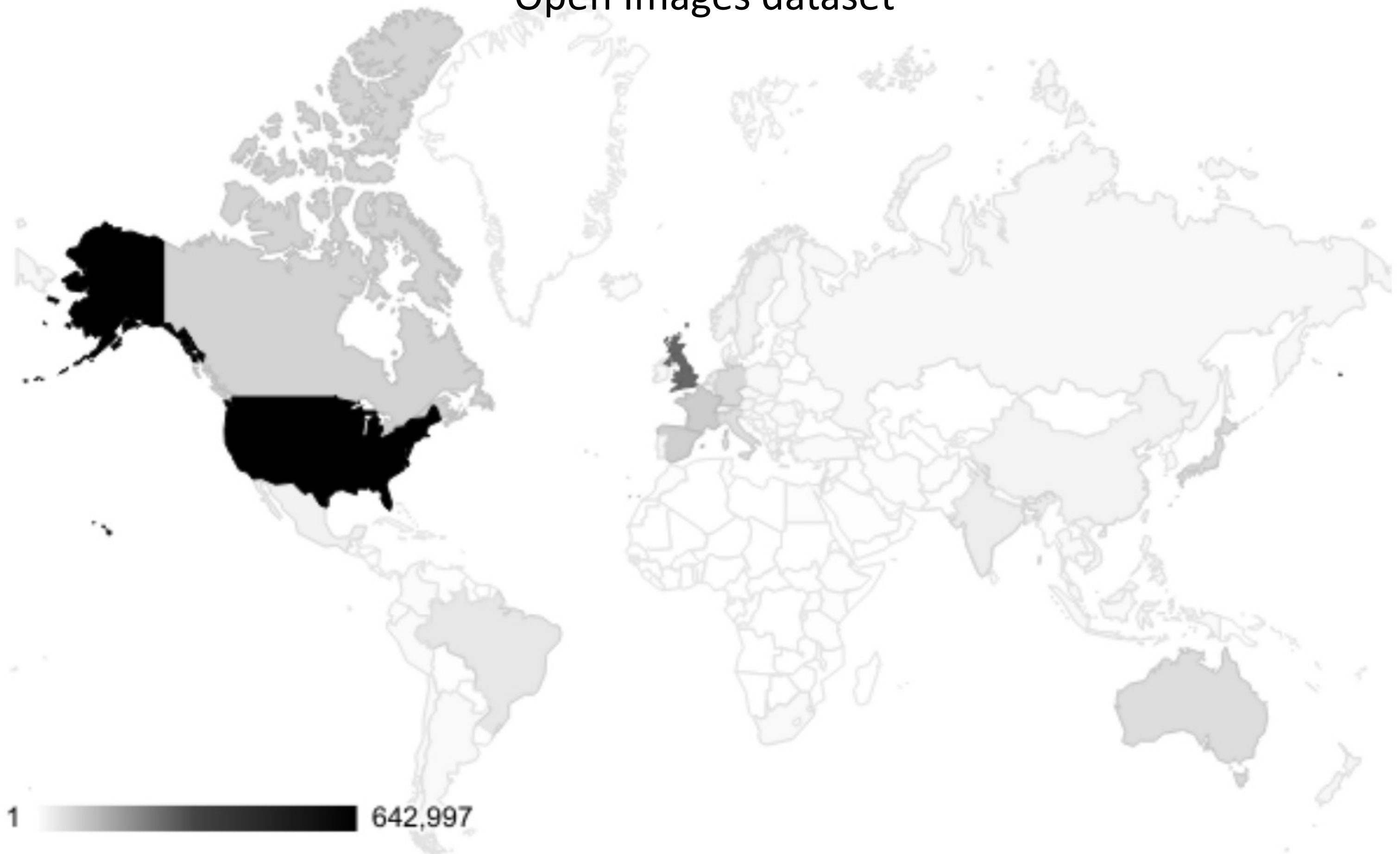
Representation bias example



Source: Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536.

Representation bias example

Open Images dataset



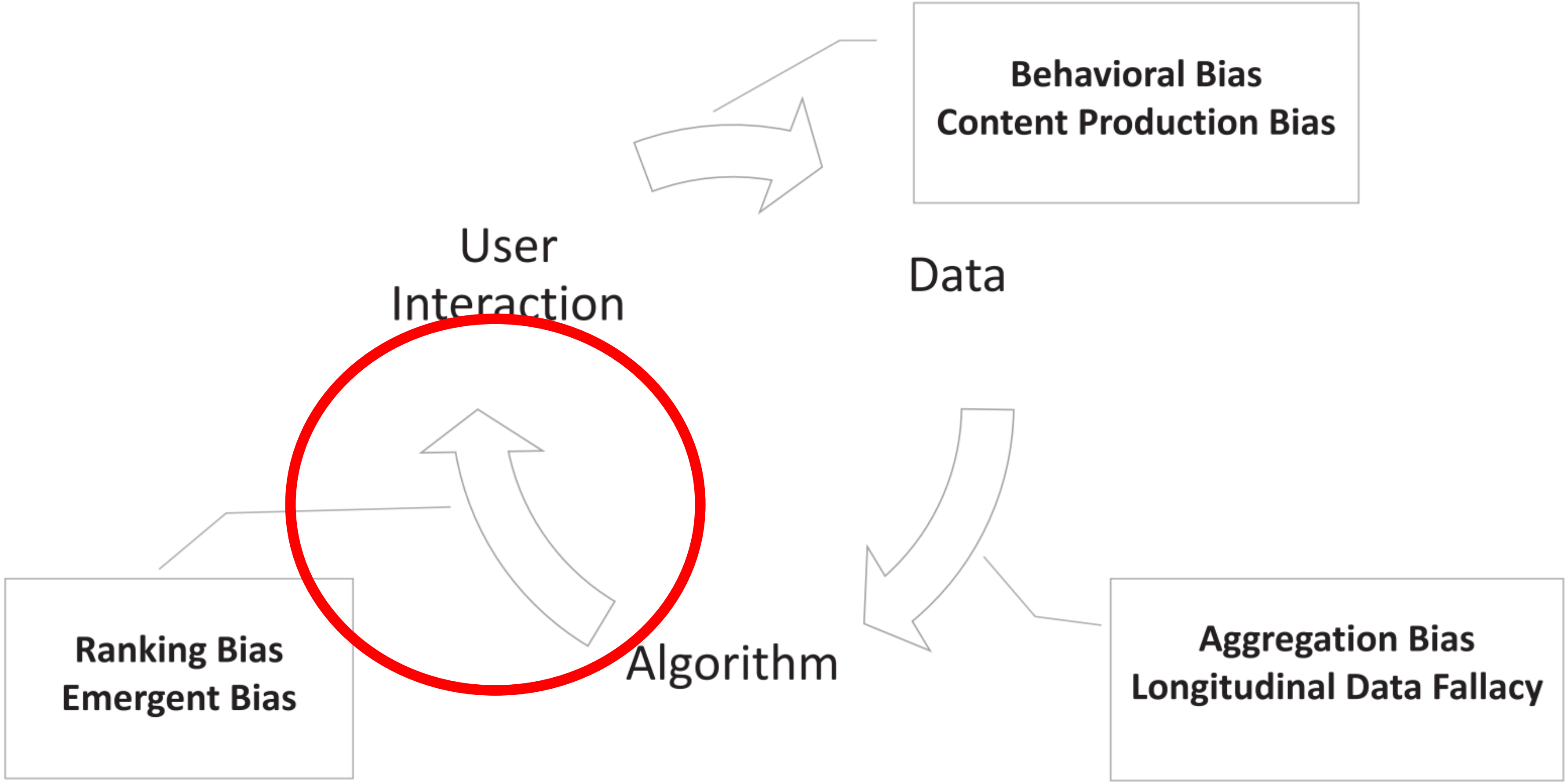
Source: Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536.

Bias from data to algorithm (cont.)

Mehrabi et al. 2021

- **Sampling Bias.** "… is similar to representation bias, and it arises due to non-random sampling of subgroups"
- **Longitudinal Data Fallacy.** "The heterogeneous cohorts can bias cross-sectional analysis, leading to different conclusions than longitudinal analysis"
- **Linking Bias.** "… arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users"

Bias in Data, Algorithms, and User Experiences



Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Bias from algorithm to user

Mehrabi et al. 2021

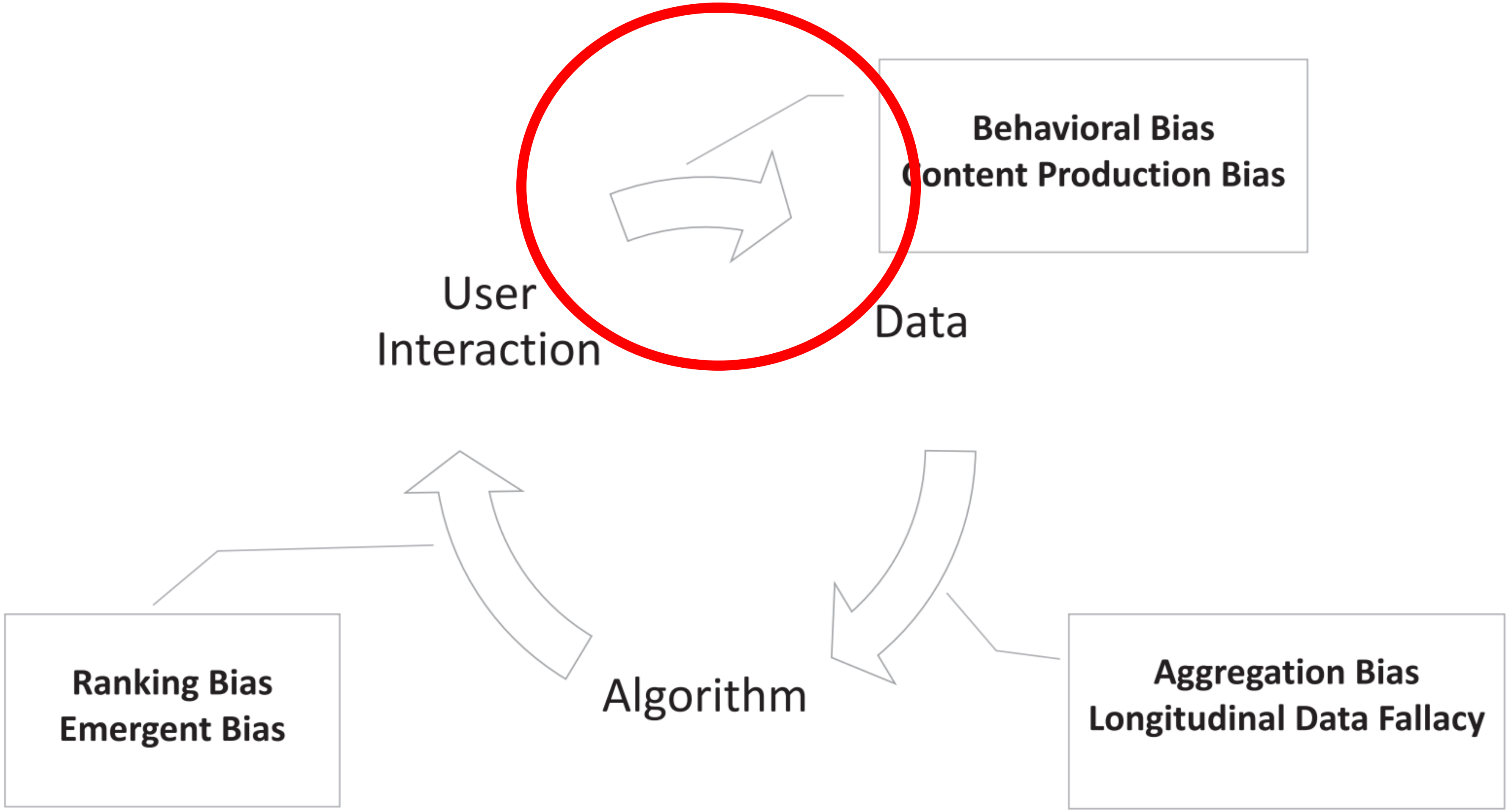
- **Algorithmic Bias.** "… is when the bias is not present in the input data and is added purely by the algorithm"
- **User Interaction Bias.** "… is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction"
  - Presentation Bias: "is a result of how information is presented"
  - Ranking Bias: "top-ranked results are the most relevant and important will result in attraction"
- **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"

Bias from algorithm to user

Mehrabi et al. 2021

- **Emergent Bias.** "… occurs 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."
- **Evaluation Bias.** "… happens during model evaluation. This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications"

Bias in Data, Algorithms, and User Experiences



Source: Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Bias from user to data

Mehrabi et al. 2021

- **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."
- **Population Bias.** "… arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population."
- **Self-selection Bias.** "… is a subtype of the selection or sampling bias in which subjects of the research select themselves"
- **Social Bias.** "… happens when others' actions affect our judgment"

Bias from user to data

Mehrabi et al. 2021

- **Behavioral Bias.** "… arises from different user behavior across platforms, contexts, or different datasets."
- **Temporal Bias.** "… arises from differences in populations and behaviors over time."
- **Content Production Bias.** "… arises from structural, lexical, semantic, and syntactic differences in the contents generated by users"

Machine learning-assisted tools: two sides

- *Algorithm aversion*
  - "The tendency to ignore tool recommendations after seeing that they can be erroneous–originates from a lack of agency."

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020, April). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

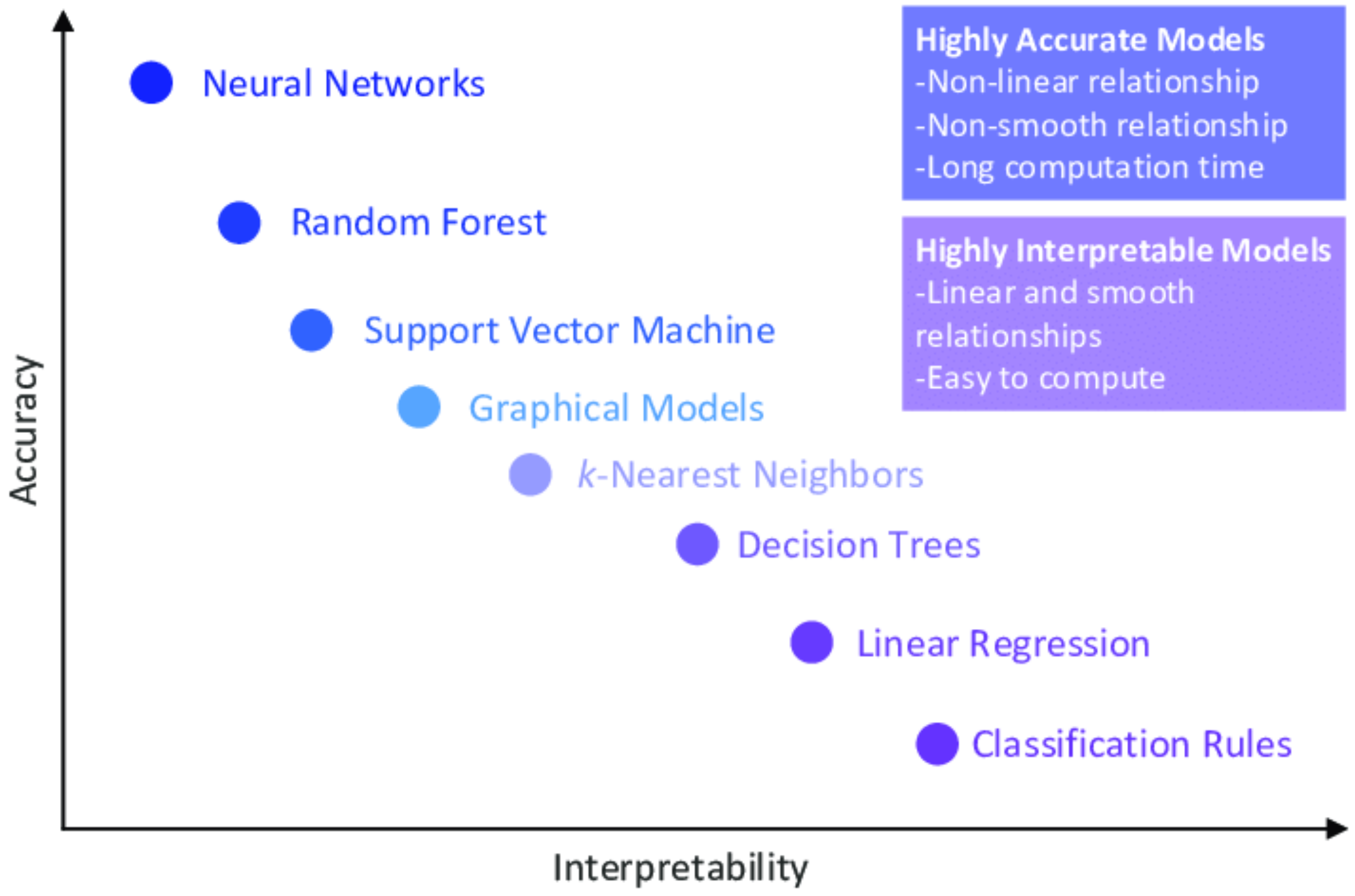
Machine learning-assisted tools: two sides

- *Automation bias*
  - “will follow tool recommendations despite available (but unnoticed or unconsidered) information that would indicate that the recommendation is wrong.”
  - Two types:
    - Omission errors: “humans fails to detect problematic cases”
    - Commission errors: “failing to incorporate contradictory external information into the decision process”

Source: https://xkcd.com/2451/

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020, April). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Machine learning techniques



Source: Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions. *IEEE Access*, *7*, 137184-137206.

Bottom line

Machine learning is so much dependent on training data and algorithms which are challenged in many ways.

=> This leads to bias and fairness issues.

It's challenging to tell how an ML model will work in the real world.

Continuous checks of data validation and improvements of model are needed to provide credible ML models.