EmbRACE- An AI IMAGE RECOGNITION MODEL TO OVERCOME UNCONSCIOUS BIAS IN SOCIAL NETWORKS

RESEARCH QUESTION

How can we overcome unconscious bias by improving diverse representation in our social networks?

HYPOTHESIS

Artificial Intelligence can be leveraged to prevent, detect and remediate unconscious bias through improved diverse representation in our social networks.

RESEARCH

Unconscious, implicit, or hidden bias is a social stereotype assumption that people hold about a category of people outside of their own conscious awareness. People naturally categorize others, quickly and instinctively based on visible attributes of identity such as race, ethnicity, gender and age. People make assumptions without realizing that they can have unconscious bias that either favors those who match their own identity, or the identity that is most culturally privileged. These stereotypes are influenced by personal experiences and cultural forces that can cause people to act in ways that undermines their personal values and goals. According to a research conducted by The Undefeated and the Kaiser Family Foundation (KFF) in October 2020, 7 in 10 African Americans feel that they are treated unfairly based on race or ethnicity when they seek medical care. It is a feeling born of unequal history and intensified by the coronavirus pandemic, which has disproportionately ravaged people of color, both physically and economically. Additionally, 58% of African Americans reported that they experienced discrimination in just the past year. One in 4 said they were discriminated against dealing with police in traffic and other incidents. 28% said they experienced racial bias on the job, and 40% said they were treated unfairly while shopping. A large majority of responders cited implicit bias as a major factor, both in preventing racial equality and as an obstacle in their own lives. It is common to hold unconscious beliefs about various races and ethnicity. Based on substantial research conducted over the past several decades, unconscious bias begins to develop in childhood. People are unconsciously primed from a young age to form biases for one racial group and against another racial group. Unconscious bias is absorbed and learned over time. It is automatic and unless attention is paid, one can enact racial biases towards another group without even consciously being aware of doing it [14]. Social media has a huge impact on unconscious bias. According to a research published by Pew Research Center in July 2020, About two-thirds of Americans (64%) believe that social media has a negative impact on the society due to misinformation, spreading hate and harassment. 55% of Americans get their news from social media. Additionally, 23% of social media users in America indicated that they changed their view about a political or social issue because of something they saw on social media. Too often social media is also allowing an agenda to be promoted, which can be divisive and promote racism. The average user has an account on more than 8 different social media platforms with 150 or more followers on each platform and spends an average of 2 hours and 29 minutes using social media each day. Constant feeding of negative rhetoric and hate can not only enhance unconscious bias, but it can also harm psychological and physical well-being. On the other hand, being surrounded by a diverse group of influencers across all ethnicities can help develop a balanced view. Humans consciously and unconsciously retain information that later influence instantaneous, automatic decision-making, which is a critical cognitive function of the brain [13]. Humans are a prey to unconscious bias which can inadvertently affect their thinking and decision-making. The importance of having a diverse group of friends and influencers, motivates the need to focus this project on the design and development of an intelligent and supervised learning model, EmbRACE to help overcome unconscious bias in social networks. With AI, racial bias in social networks can be identified and addressed. EmbRACE is based on the rationale that when people are made aware of the lack of diversity in their social networks, they can work towards becoming more inclusive. Doing so will help promote an environment that is diverse and inclusive, based on mutual respect and appreciation for all.

ENGINEERING GOAL

Design and develop a supervised AI model using Convolutional Neural Networks (CNN) architecture to perform image recognition and classification tasks. Use transfer learning to create a classifier to classify faces based on ethnicity and race.

HARDWARE AND SOFTWARE REQUIREMENTS

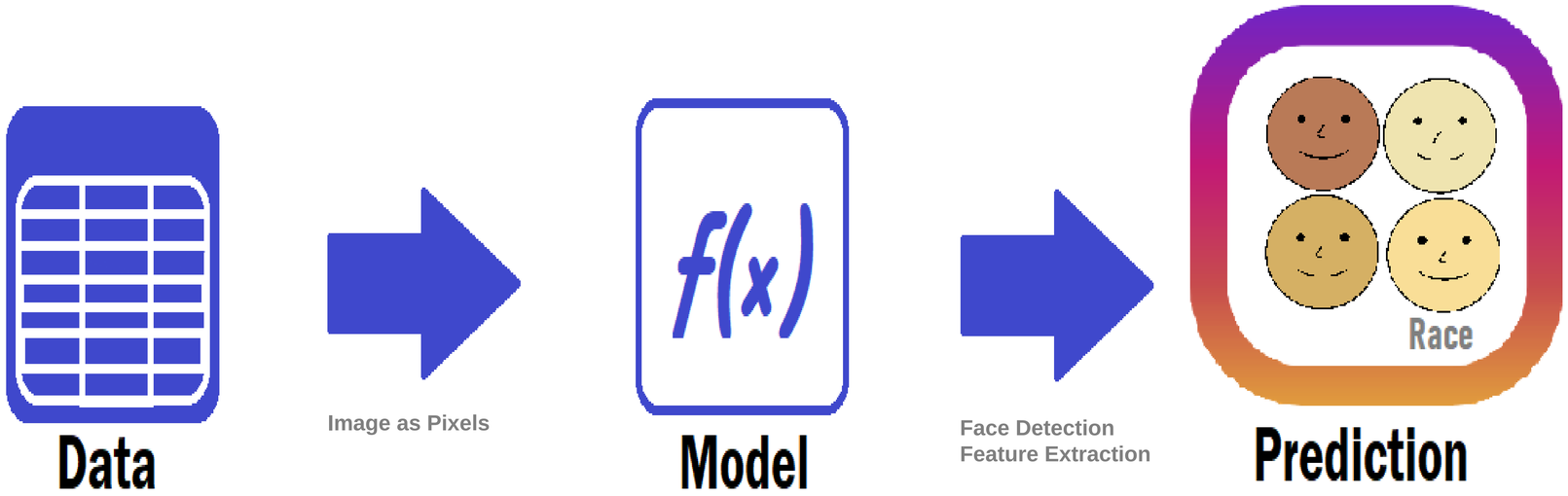
Hardware	16GB Memory (RAM)	1TB (1000 GB) Hard disk	Intel Core i7 Processor (CPU)	NVIDIA GeForce MX250 GPU
	Windows 10 Operating System		Workstation	
Software	Anaconda	Git Hub	Google Chrome	Google Colab
	Jupyter Notebook	Microsoft Access	Microsoft Excel	Python
	Python libraries - Keras, Numpy, Matplot		TensorFlow	

PROCEDURE

To develop a deep learning model for image recognition and race classification the first step was to gather publicly available datasets for facial images and labelled data. After a thorough research of open-source datasets with race and ethnicity labeled facial images, FairFace turned out to be the best publicly available dataset. FairFace is a large dataset with 97,698 race-labelled images spread across all six races in scope for this project. Images from FairFace were sampled for quality and the corresponding race information in the Comma-delimited (CSV) file was reviewed. The quality of the images was high and just a couple of images had to be deleted and replaced with better quality images. The corresponding race labels had to be hand-annotated. A data analysis of percentage distribution of samples by race was performed to ensure adequate coverage across all races. All races were fairly represented with at least 14 percent of the dataset aligned to each category. The next step was to split the dataset up between training and testing sets. 86,744 images were assigned for training and 10,954 for testing. A 'Facial Recognition' workspace was created in the workstation and the images and Comma-delimited (CSV) files with the race labels for both the training and testing dataset were stored separately. The next step was to research a compatible and reliable Convolutional Neural Network (CNN) based base model to use as the framework for building the race classification model. Seven state-of-the-art face recognition models including VGG-Face , Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace and Dlib were evaluated and VGG-Face came across as the most promising. VGG-Face is a face recognition model developed by Visual Geometry Group (VGG) at the University of Oxford and pre-trained with millions of images. VGG model expects 224x224x3 sized input images and are represented as a 2622-dimensional vector. The next step was to setup the computing environment by installing all the required software listed under materials on a high-performance laptop. Some of the packages had to be updated to the most current version of Python libraries. VGG-Face Application Programming Interface (APIs) was installed as well. A Jupyter notebook was created for the model and code was written to import Keras, Pandas, NumPy, MatplotLib, and tqdm libraries. Additional functions were created to explore the dataset. Data analysis was performed by evaluating the total counts of actual images in the folders and labelled race information with the image names in the CSV file. After loading the data, additional checks were performed by picking a few image samples and looking to see if the race was annotated as expected. Images had to be transformed to an array pixel format for the model to read. These image pixels were stored in a NumPy array along with the image name and pre-labelled race. Additional data validation was preformed on a few random samples from the NumPy array to ensure conversion of images to pixels. Furthermore, the race labels were converted to numerical codes and get_dummies Panda function one-hot coding was applied to ensure a binary value for each feature. The data was now ready for training the model. Transfer learning was applied to the VGG-Face model to leverage the pre-trained layers on millions of images and retrain the model for custom face recognition with additional training data. Since this pre-trained base model was capable of performing some level of facial recognition, all but the last seven layers of the neural network were locked and the remain were left unlocked to learn through training. The next step was to load the VGG-Face model and fine-tune hyperparameters including batch size and epochs based on model performance which was evaluated by the loss function.

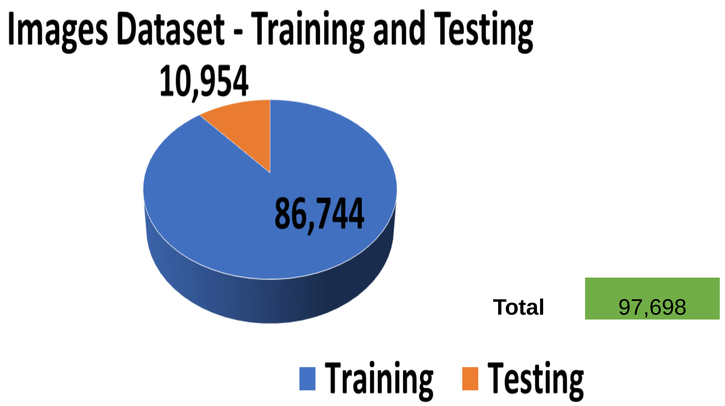
DEEP LEARNING FRAMEWORK

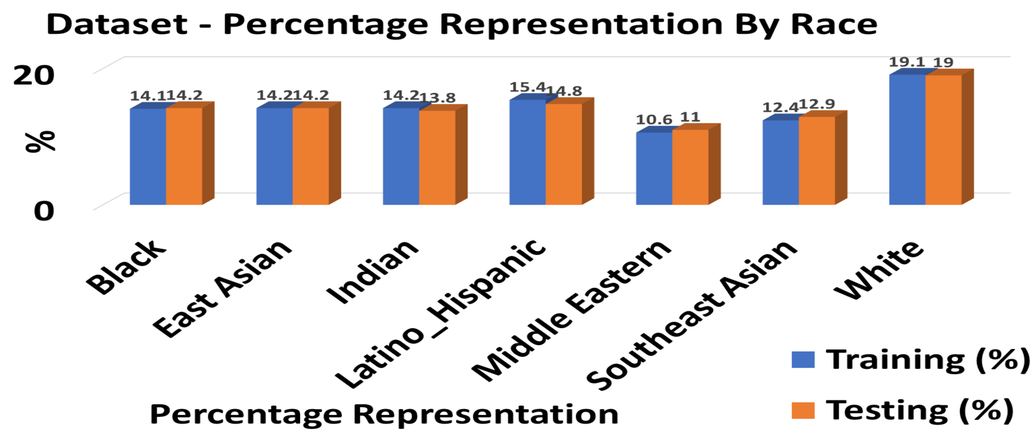
Gather Data | Explore Data | Prepare Data | Build, Train, and Evaluate



SOURCE: Original Illustration Designed by Researcher

DATA ANALYSIS





SOURCE: Original Data Analysis Performed by Researcher

COVID-19 SAFETY PRECAUTIONS

The following safety precautions were taken to reduce the spread of COVID:

- ✓The project was completed at home
- ✓All materials were purchased online
- ✓Hand sanitizer was used when printing paperwork
- ✓Masks and social distancing protocols were followed

The total number of files is 12146 The total number of words in the files is 25195 The average number of words in the files is 2.074345463527087

RESULTS

As expected with a CNN model, trying to feed the 86,744 images to the model left to memory issues. The batch was Start will a large batch size and reduce batch size if model runs into memory issues which are common for CNN due to hierarchical structure which allows network to concentrate on low-level features in the first hidden layer and then assemble them into higher level layers resulting in heavy usage of memory. The training was performed in mini batches with 20,000 images for 29 epochs with each epoch selecting 16,000 random instances. After 29 epochs the training loss started to decreased while validation loss increased. Once the model was fully trained, the test images were passed as input to the model to make predictions. The predictions were 91% accurate for “White”, “Black”, “Asian” and “Indian” race. The accuracy was lower, 79% for “Latino-Hispanic” and “Middle-Eastern” race. Additionally, bi-racial images were often mis-classified due to insufficient data. Future considerations include enhancing the dataset with more bi-racial images and building a web-interface to enable social media users to see a statistical analysis of their follower/influencer base so they can be more deliberate in being inclusive.

AmanEmbRACEModel.ipynb

Code

```
[4]: train_df = pd.read_csv("fairface_label_train.csv")
     test_df = pd.read_csv("fairface_label_val.csv")
     print("Training set consists of ",train_df.shape)
     print("Testing set consist of ",test_df.shape)
     test_df = test_df[['file', 'race']]
     train_df['file'] = 'Downloads/FairFace/'+train_df['file']
     train_df.head()
```

```
Training set consists of  (86744, 5)
Testing set consist of  (10954, 5)
```

[4]:

	file	race
0	Downloads/FairFace/train/1.jpg	East Asian
1	Downloads/FairFace/train/2.jpg	Indian
2	Downloads/FairFace/train/3.jpg	Black
3	Downloads/FairFace/train/4.jpg	Indian
4	Downloads/FairFace/train/5.jpg	Indian

AmanEmbRACEModel.ipynb

Code Python 3

```
[9]: # There are 2 races as East and Southeast Asian. Group them in a single Asian Race
     idx = train_df[(train_df['race'] == 'East Asian') | (train_df['race'] == 'Southeast Asian')].index
     train_df.loc[idx, 'race'] = 'Asian'

     idx = test_df[(test_df['race'] == 'East Asian') | (test_df['race'] == 'Southeast Asian')].index
     test_df.loc[idx, 'race'] = 'Asian'

     100*train_df.groupby(['race']).count()[['file']]/train_df.groupby(['race']).count()[['file']].sum()
```

[9]:

race	file
Asian	26.609339
Black	14.102475
Indian	14.201559
Latino_Hispanic	15.409710
Middle Eastern	10.624366
White	19.052615

AmanEmbRACEModel.ipynb

Code

```
[24]: train_df.head()
```

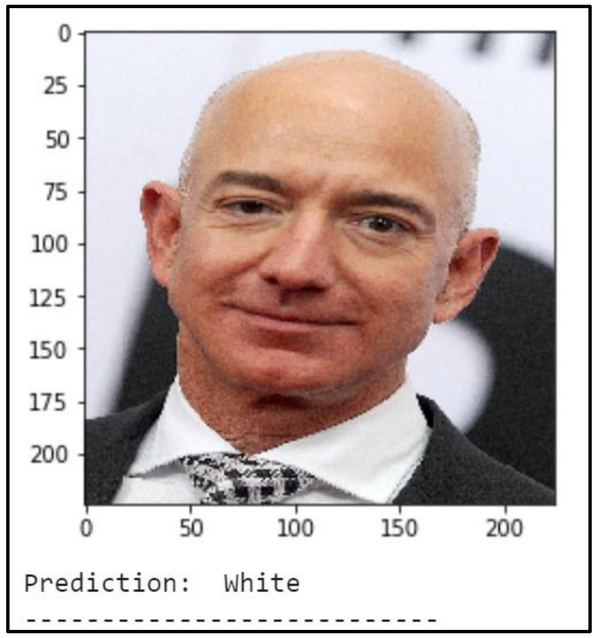
[24]:

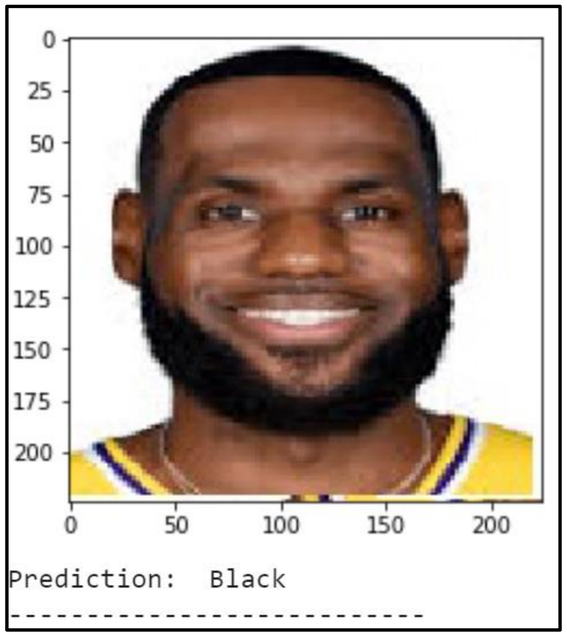
	file	race	pixels
0	Downloads/FairFace/train/1.jpg	Asian	[8.0, 8.0, 10.0, 9.0, 9.0, 11.0, 10.0, 8.0, 11...
1	Downloads/FairFace/train/2.jpg	Indian	[129.0, 127.0, 104.0, 127.0, 125.0, 102.0, 123...
2	Downloads/FairFace/train/3.jpg	Black	[216.0, 171.0, 174.0, 212.0, 167.0, 170.0, 206...
3	Downloads/FairFace/train/4.jpg	Indian	[42.0, 47.0, 50.0, 42.0, 47.0, 50.0, 41.0, 46....
4	Downloads/FairFace/train/5.jpg	Indian	[44.0, 39.0, 35.0, 44.0, 39.0, 35.0, 43.0, 40....

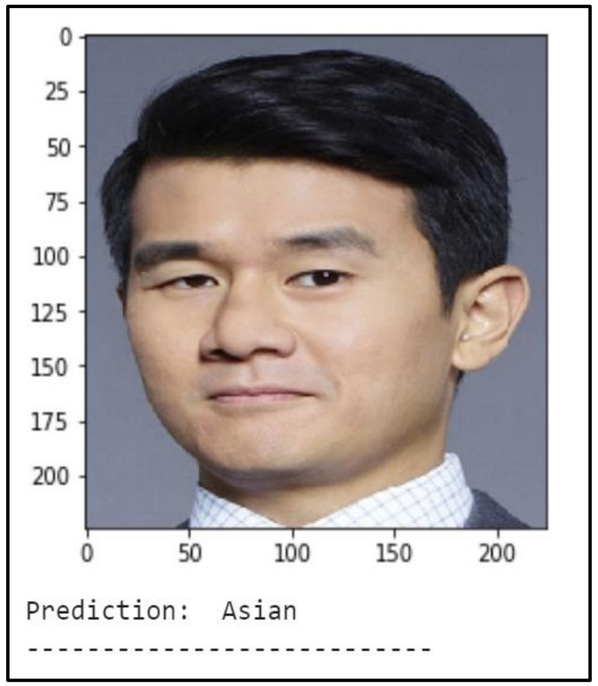
	file	race	pixels
10101	Downloads/FairFace/train/10102.jpg	Black	[204.0, 235.0, 255.0, 204.0, 235.0, 255.0, 204...
10102	Downloads/FairFace/train/10103.jpg	Latino_Hispanic	[22.0, 12.0, 11.0, 24.0, 14.0, 13.0, 27.0, 17....
10103	Downloads/FairFace/train/10104.jpg	Asian	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
10104	Downloads/FairFace/train/10105.jpg	Middle Eastern	[102.0, 102.0, 100.0, 100.0, 100.0, 98.0, 97.0...
10105	Downloads/FairFace/train/10106.jpg	White	[65.0, 44.0, 43.0, 67.0, 43.0, 43.0, 67.0, 43....

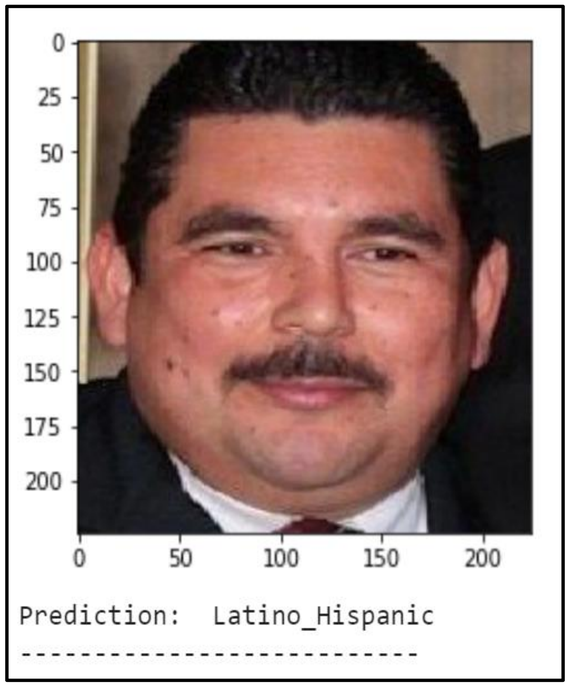
SOURCE: Original Model Code Developed by Researcher

PREDICTIONS MADE BY THE MODEL





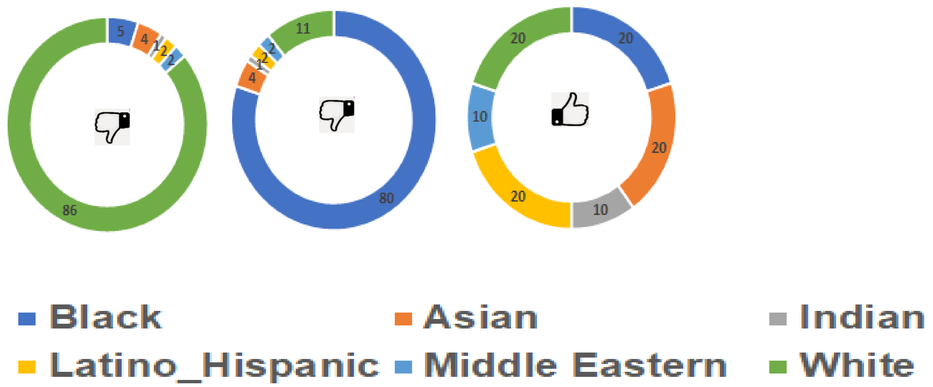




SOURCE: Original Predictions made using the Model Developed by Researcher. Public Stock Photos sourced from Getty Images

FUTURE STUDIES AND APPLICATION

Future considerations include enhancing the dataset with more Latino-Hispanic and bi-racial images and building a web-interface to enable social media users to see a statistical analysis of their follower/influencer base so they can be more deliberate in being inclusive. Being surrounded by a diverse group of influencers across all ethnicities can help develop a balanced view.



SOURCE: Designed by Researcher

CONCLUSION

The model proved that through continued focus on improving datasets for bi-racial images, race classification will have a promising application in overcoming unconscious bias in social networks.

BIBLIOGRAPHY

1. "Exploring Unconscious Bias in Disparities Research and Medical Education", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4169280/#R3
2. "GitHub", https://github.com/
3. "Google Face Detection Concepts", https://developers.google.com/vision/face-detection-concepts
4. Himanshu Singh, Practical Machine Learning and Image Processing: For Facial Recognition, Object Detection, and Pattern Recognition Using Python, O'Reilly, 2019.
5. Jon Flanders; Implementing Image Recognition Systems with TensorFlow, Pluralsight, 2019.
6. "Kaggle", https://www.kaggle.com/datasets
7. "Neural Networks and Deep Learning", Coursera, https://www.coursera.org/learn/neural-networksdeeplearning/home/welcome
8. Paul J Deitel; Harvey M Deitel. Python for programmers with introductory AI case studies, Pearson, 2019.
9. "Pew Research Center", https://www.pewresearch.org
10. Sefik Ilkin Serengil, Race and Ethnicity Prediction in Keras, https://sefiks.com/2019/11/11/race-and-ethnicityprediction-in-keras/.
11. "Social Media Statistics and Trends", https://www.omnicoreagency.com/social-media-statistics
12. "TensorFlow", https://www.tensorflow.org/
13. "Unconscious Bias in the Classroom: Evidence and Opportunities", https://cepa.stanford.edu/content/unconscious-bias-classroom-evidence-and-opportunities
14. "Unconscious Bias in Schools", https://www.gse.harvard.edu/news/19/11/harvard-edcast-unconscious-biasschools