Chi-Hsun (Eric) Chang

Statistics
Data-driven modeling
Correlation
Regression
Classification
Clustering
Natural language processing
Image processing
Visualization
Experiment/Research
Hypothesis testing

Contact:

✉ chchang8305@gmail.com

in linkedin.com/in/chi-hsun-eric-chang/

⬤ github.com/EricCHChang

Brief summary:

Hello, thank you for taking time to read through my portfolio. It contains a couple of projects I am working on recently and several research work I conducted during graduate school. I hope it shows my interests and experience in using a variety of data analytical techniques to tackle research and real-world problems, and my ability to present results and actionable insights as a clear, succinct story.

Table of Content

Reddit Post Recommendation & Topic Prediction Models

web scrapping, NLP, classification, visualization, SQL

1. Traditionally, scientific findings are published in scientific journals by domain experts and usually not accessible to the general public

2. With the development of social media platforms, almost anyone can post and discuss science on the Internet

3. A trade-off to this increased accessibility of scientific articles is the rise of misinformation. Hence, it is important that scientists have access to these articles on social media and potentially help counter fake science.
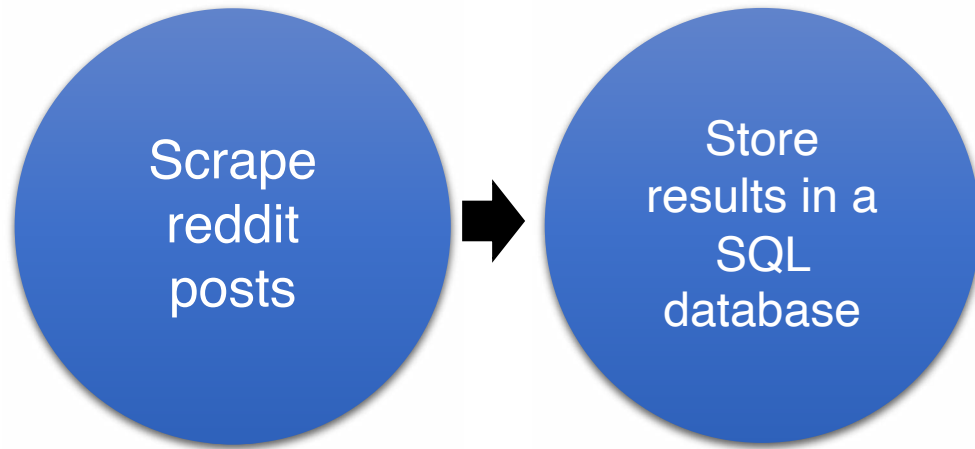
Goals:
- Scrape reddit posts and store in a SQL database
- Build a content-based reddit post recommendation and a topic prediction model

This is part of the Brainfeed project

Pipeline

Data collection

Data preparation/ preprocessing

Post recommend -ation model

Modeling

Scrape reddit posts

Store results in a SQL database

Extract data from the database

Preprocess text features

Post topic prediction model
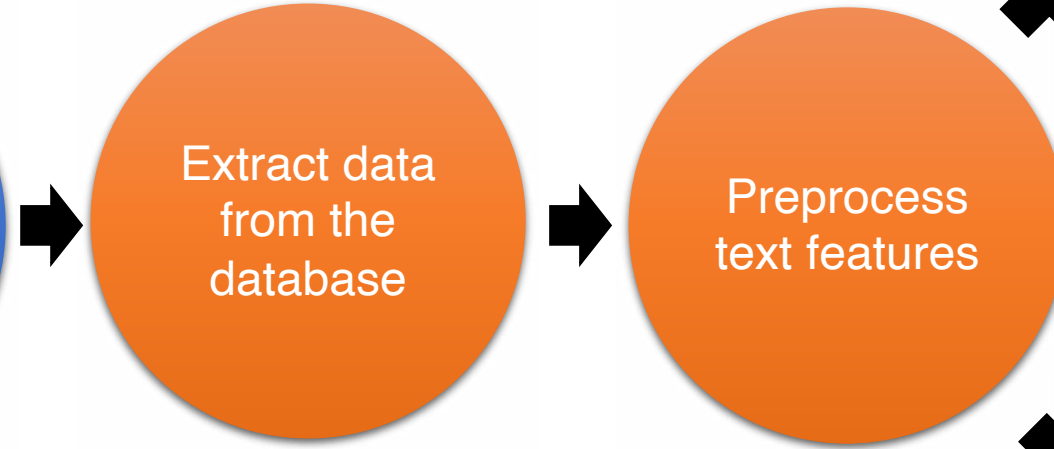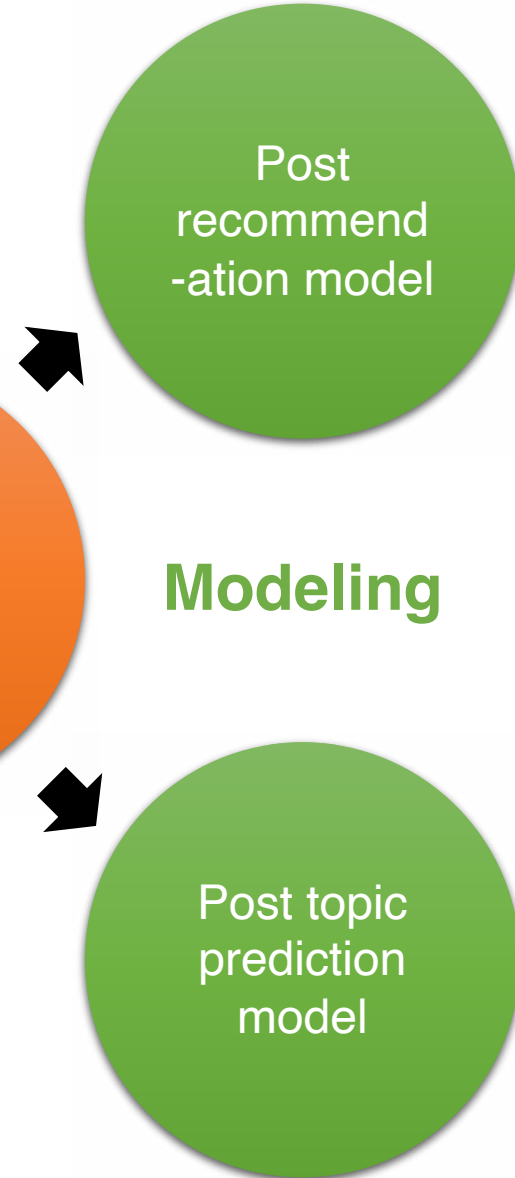
Web application*: Searching Reddit Posts

- Search reddit posts based on an user's inputs
- Return search results

Overview of the Data

- Health is the top 1 topic, followed by Psychology, Environment and Biology
- Most posts have 1 comment

Post Recommender: Recommend posts similar to a given post

Scenario: An user reads a post about COVID-19 and wants to find other similar posts

	id	author	title	subreddit	link_flair_text
0	ujvpgv	Additional-Two-7312	New study finds boosting against COVID-19 in a...	science	Biology

Recommender model: Return other posts that have similar content (title). The top 5 similar posts are displayed here and their all have COVID-19-related content (e.g., vaccines)

	id	author	title	subreddit	link_flair_text	cosine_sim
0	uj9ql9	Additional-Two-7312	New study finds vaccines induce multiple immun...	science	Biology	0.853618
1	ty3gmt	Additional-Two-7312	Third dose of COVID-19 vaccine significantly i...	science	Biology	0.848778
2	u1a5pg	MistWeaver80	Fully vaccinated people had a reduced infectio...	science	Medicine	0.842615
3	uinqcf	Additional-Two-7312	New study suggests that vaccine effectiveness ...	science	Health	0.841503
4	txbhk5	Hrmbee	Protection against infection offered by fourth...	science	Medicine	0.838720

Similarity is computed based on SpaCy's word vectors

Overview of the Data for Post Topic Prediction

Demonstration example: Posts with topic label of Biology or Environment

Project text features (TF-IDF) into a 2D space via Truncated SVD
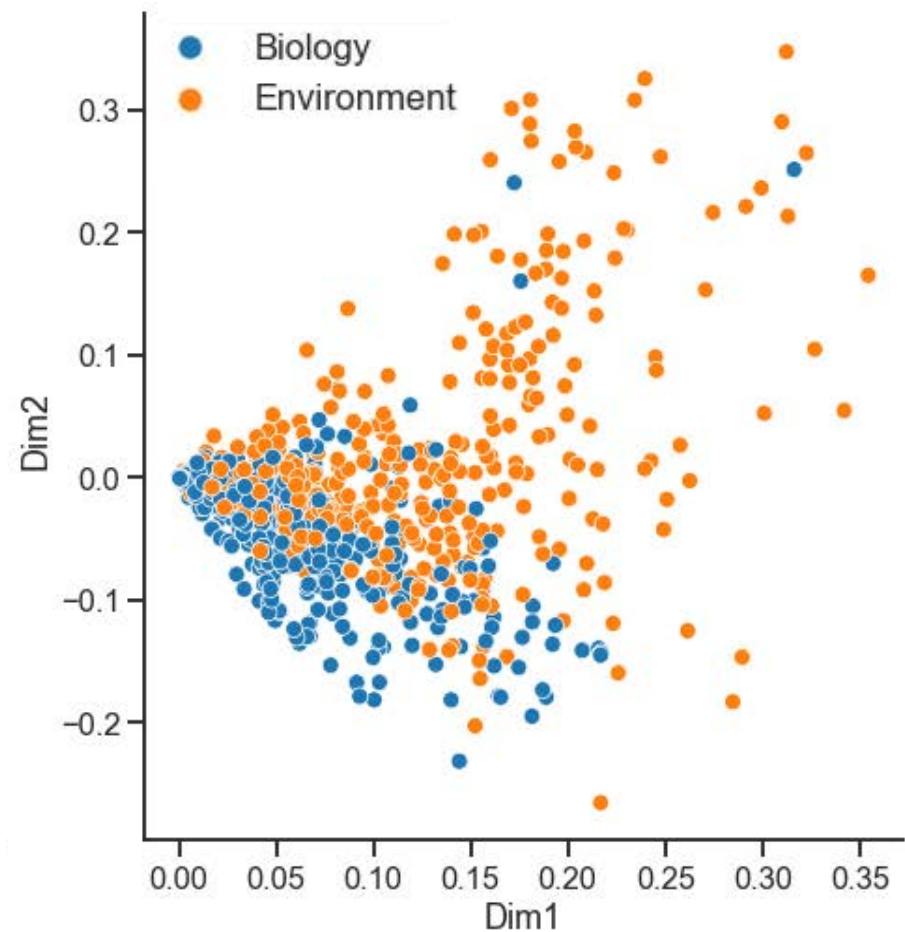
Identify the most common words via a Latent Dirichlet Allocation analysis

Topic	Top 5 common words
Biology	human, cell, find, time, new
Environment	climate, use, study, change, year

Number of posts for each topic group
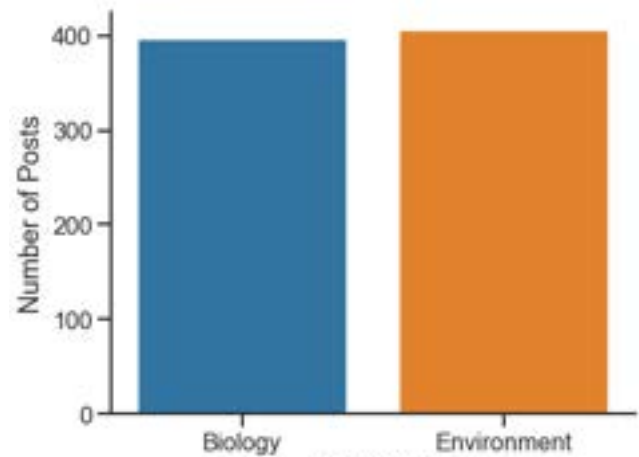
Post Topic Prediction: Predict a post's topic based on its content

Demonstration example: Predict whether a post's topic is Biology or Environment

1. Iteratively fine-tuned parameters for the TF-IDF vectorizer and the Naïve Bayes classifier based on the training data (70% of the total sample)
2. Applied the final trained model to the left-out testing data (30% of the total sample)

Prediction accuracy: 82%   Area under ROC: 0.93
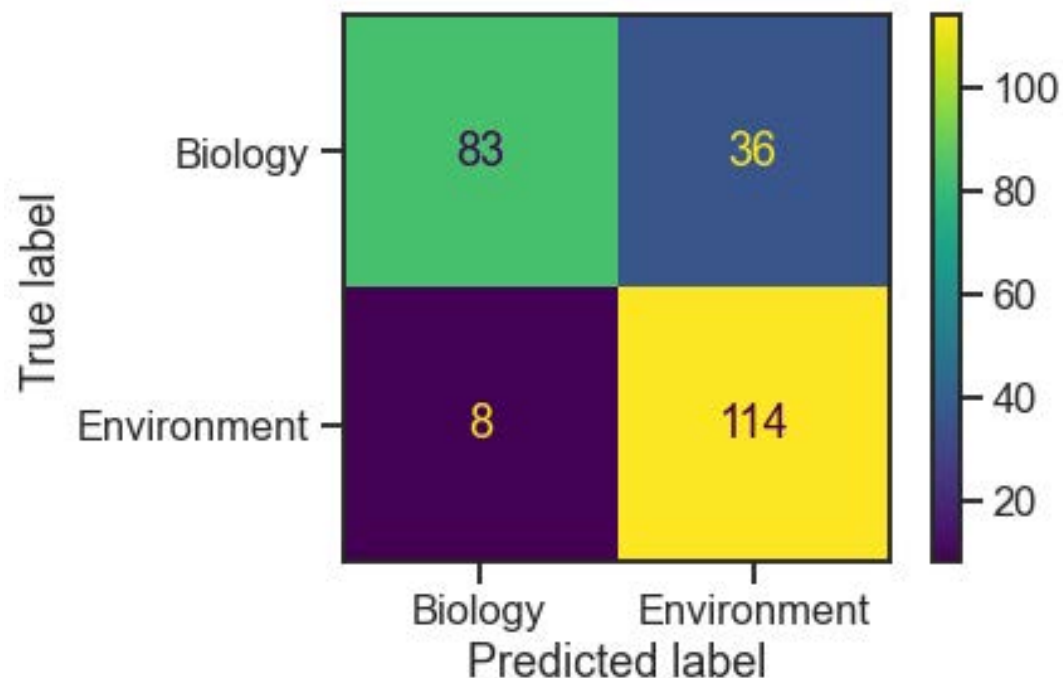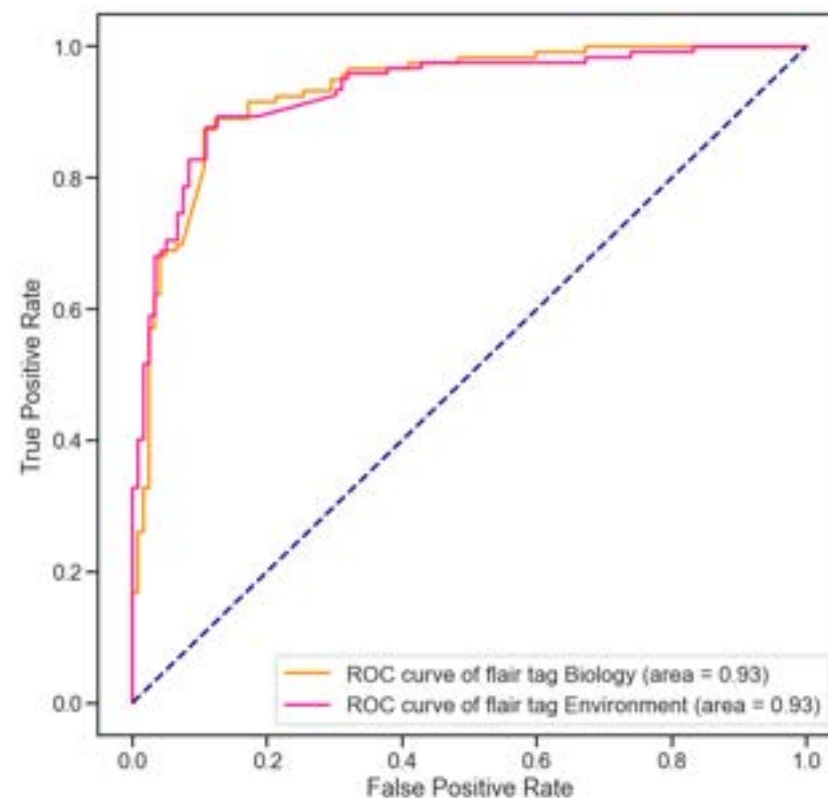
Reddit Recommendation & Topic Prediction Conclusions

- The recommender model can help users find posts similar to a specific reddit post
  - Beneficial for users since they can read posts with similar content without the need to manually search from the database

- The prediction model successfully predicts posts' topics
  - Can be implemented as a automated topic labeling system when posts are stored into the database (if the topic label is missing)

- These models can be applied to other social media platforms (e.g., Twitter)

Python codes and more analyses are available on GitHub

Exploring an IMDb Korean TV Dataset: Foundation for a K-Drama Recommendation App

statistics, correlation, regression, NLP,
visualization

1. During the pandemic, I have more time staying at home, so I got a new hobby – watching Korean dramas (K-dramas)!

2. When a drama is finished, one thing that sometimes bothers me: What should I watch next? Even if I find one, how do I know if it is a good drama?

3. It would be great if there is a K-Drama recommendation app!
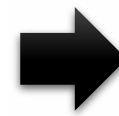
Goals:
- A starting point for such an app – Identify information predictive of K-dramas/shows with high ratings

An IMDb K-Drama/Shows Dataset

- After data cleaning, 1803 Korean dramas/shows were included in the present analysis

- After feature engineering, features consist of:
  - Title
  - Rating
  - Release year
  - Broadcast length (years)
  - Number of votes
  - Episode length
  - Genre
  - Cast
  - Synopsis

kaggle

CHANON CHARUCHINDA · UPDATED 6 MONTHS AGO

IMDb Korean TV Series

Korean TV series listed on IMDb

- The raw data contains 1989 dramas
- Data was scrapped on December 11, 2021

Exploratory Analyses of Release Year & Broadcast Length
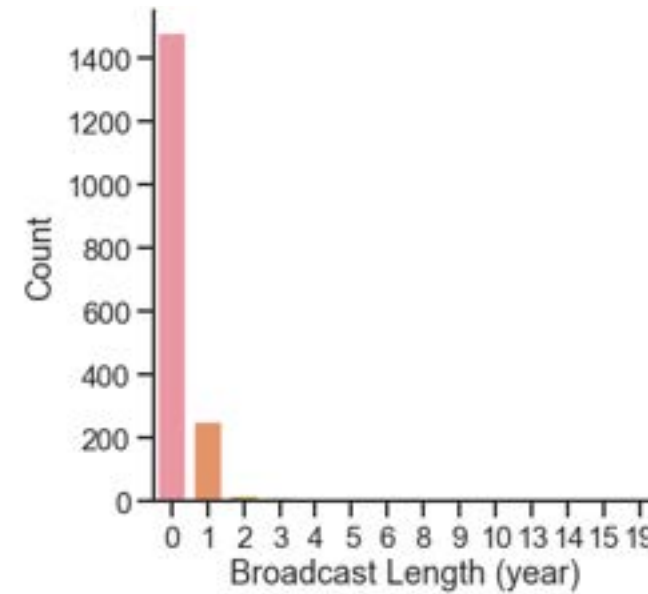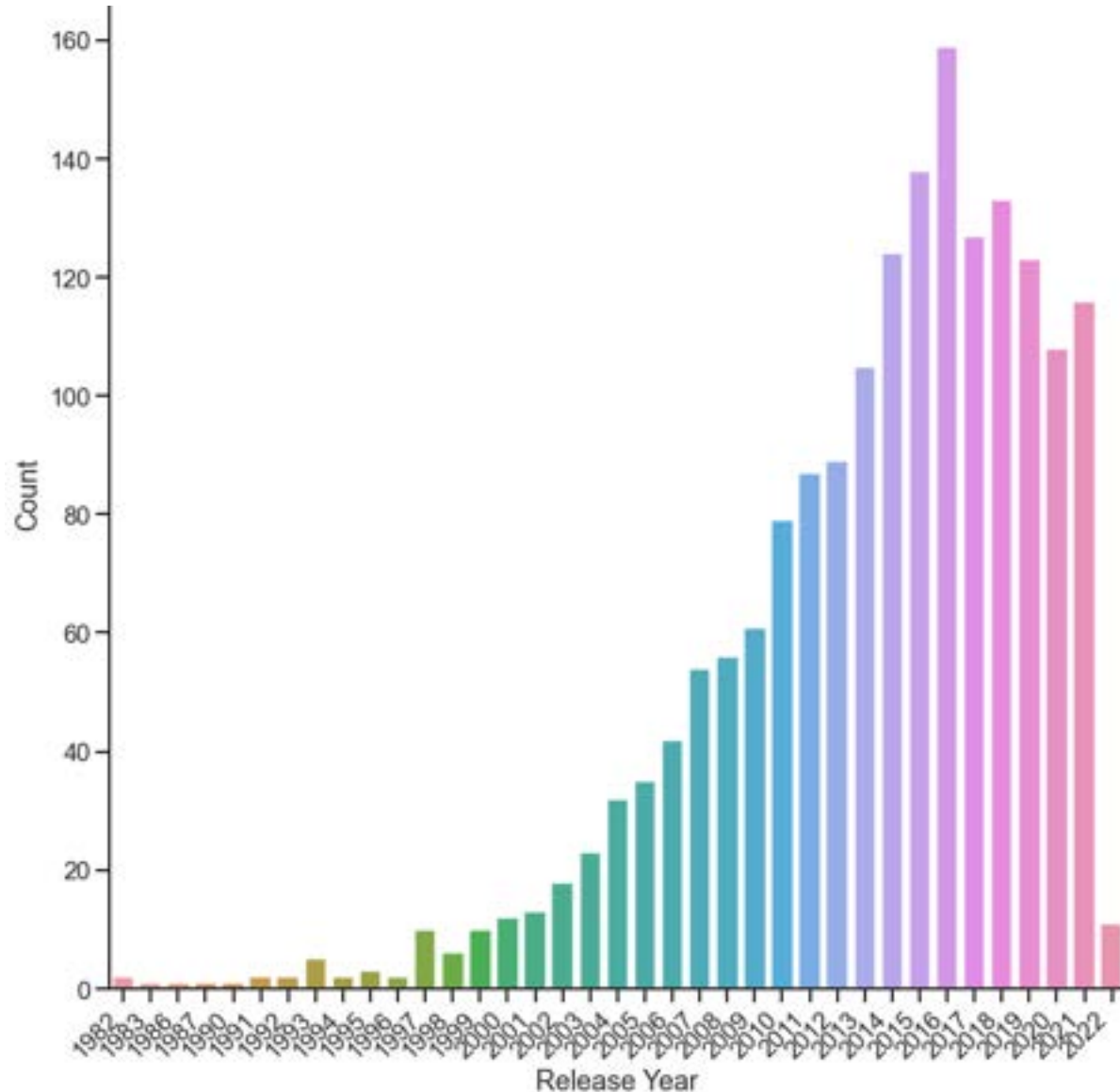
- The number of K-dramas/shows released increases each year and peaks at 2016

- The majority of dramas/shows are finished within a year, which is reasonable given that most K-dramas have 12 to 20 episodes.

Exploratory Analyses of Genre & Cast
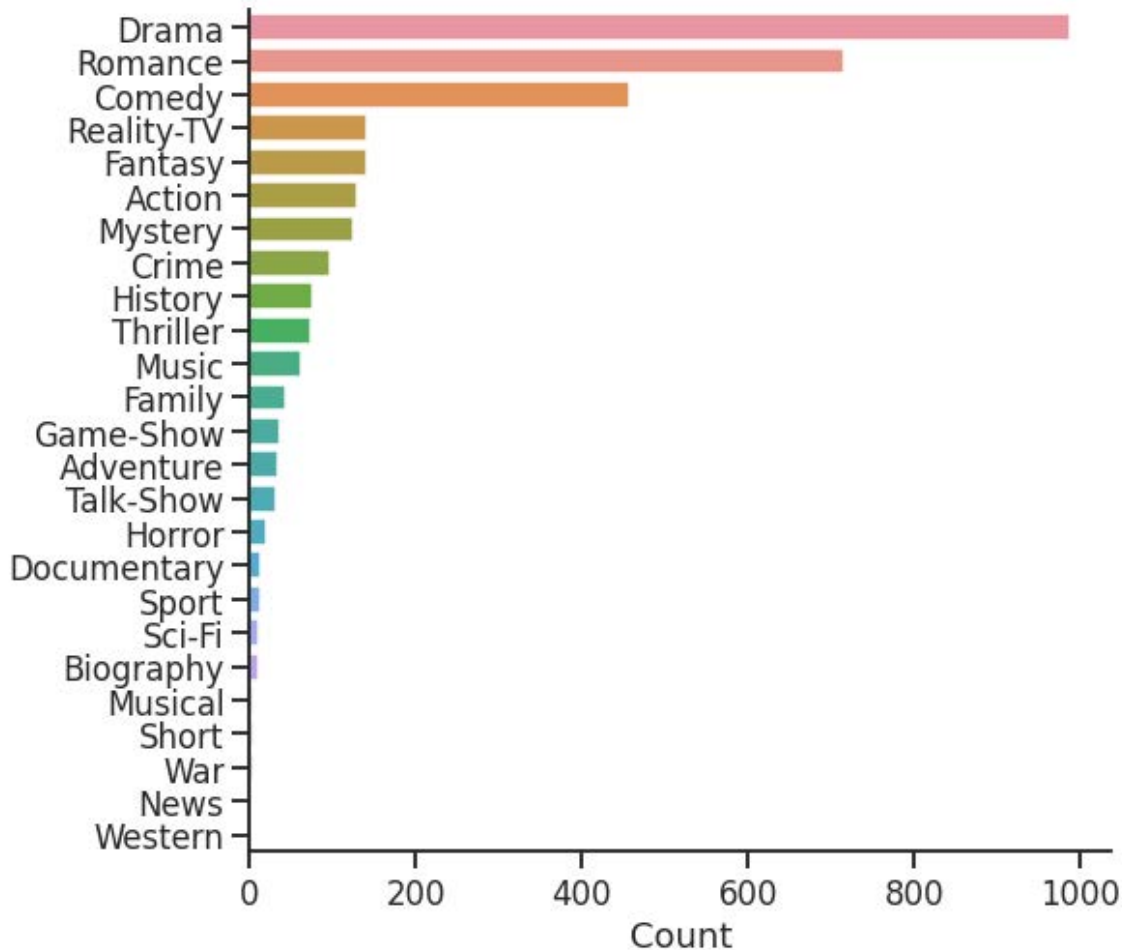
- Drama is the most common genre in this dataset, followed by romance and comedy.
  - Note: each drama/show can have more than 1 genre labels.
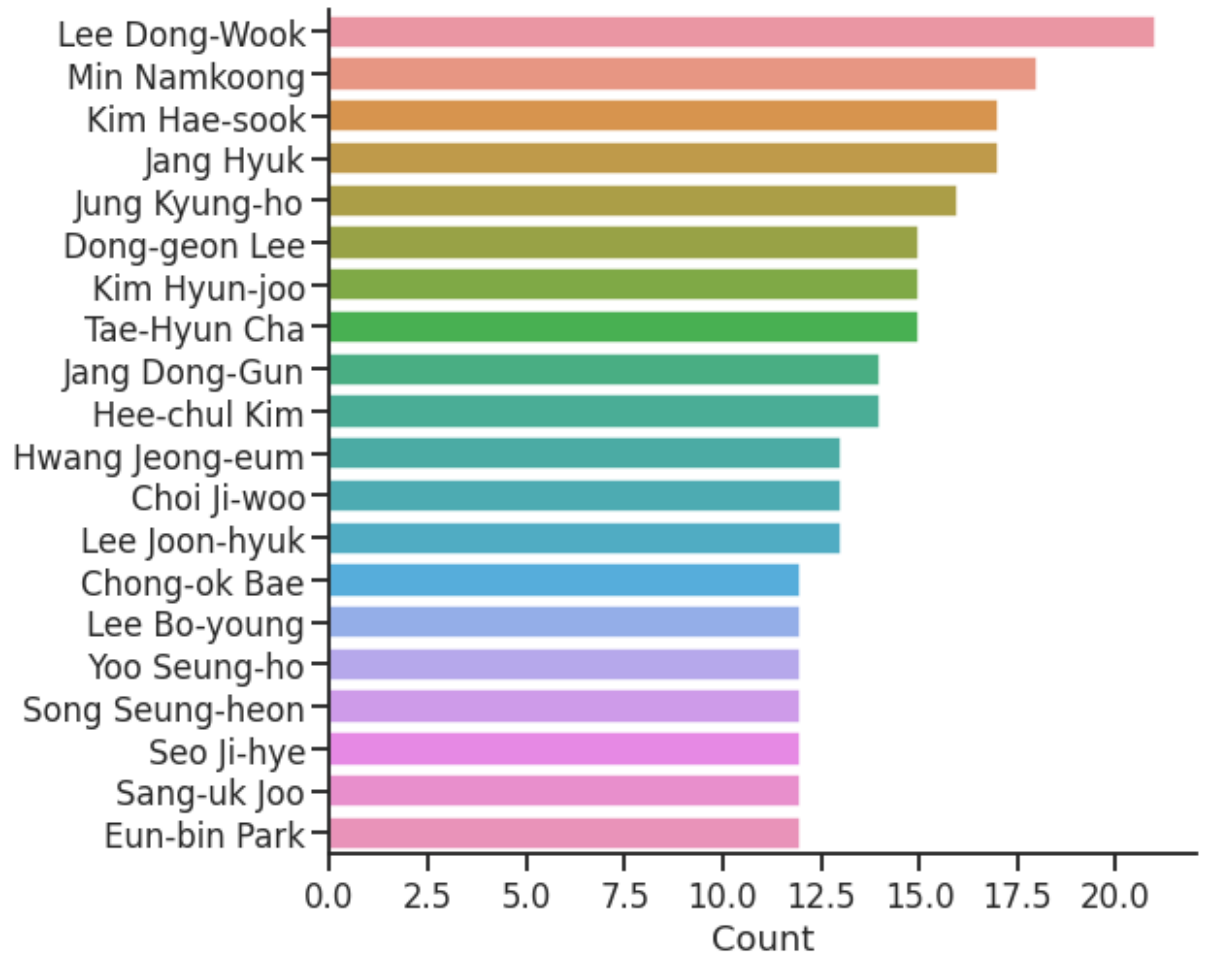
- Lee Dong-Wook is the actors/actresses who has starred in the most K-dramas/shows in this dataset

K-Dramas/Shows Ratings
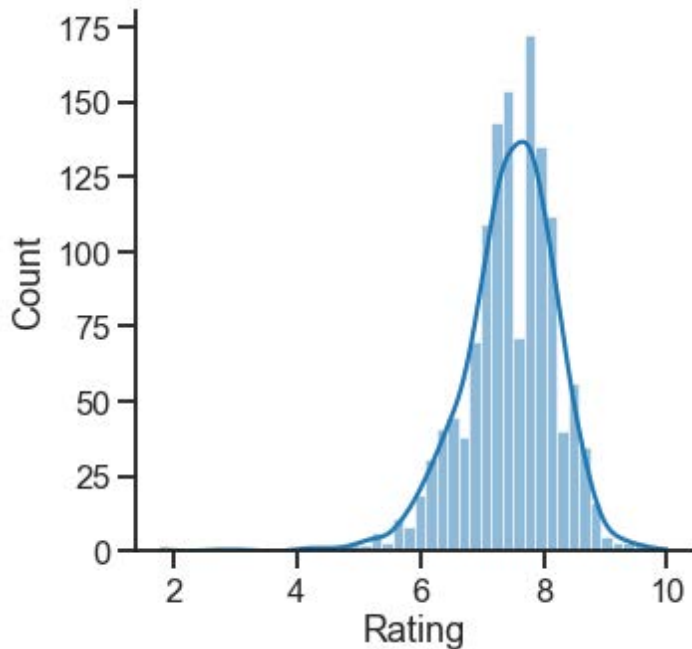
- Most K-dramas and shows have ratings approximately between 7 and 8

- Three dramas/shows have extremely high ratings relative to others. A quick inspection reveals that two of them are about BTS, a super popular K-pop group and was the top-selling singers across the globe in 2021[1].

Title	YearRelease	nYears	Rating	Votes	Time	Genre
Snowdrop	2021.0	0.0	10.0	140.0	70.0	Drama, Romance
BTS: Bon Voyage	2016.0	0.0	9.6	908.0	NaN	Reality-TV
Run BTS!	2015.0	0.0	9.7	1013.0	NaN	Game-Show
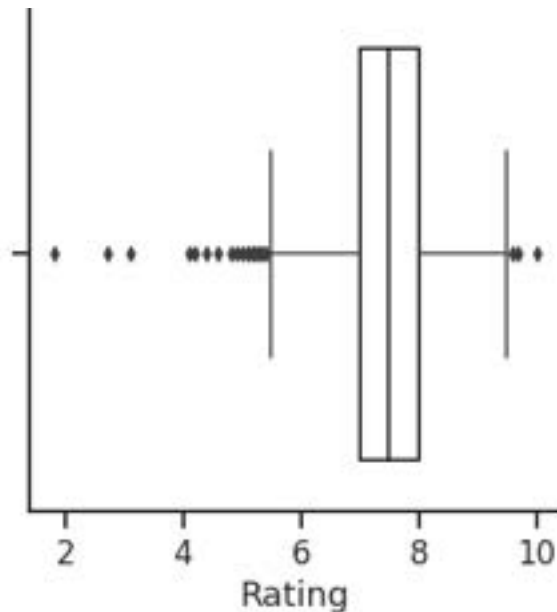
- In contrast, some have extremely low ratings – as low as 1.8 out of 10!

Title	YearRelease	nYears	Rating	Votes	Time	Genre
Racket Boys	2021.0	0.0	1.8	27231.0	80.0	Comedy, Drama, Sport
Unknown Woman	2017.0	0.0	2.7	45.0	NaN	Drama
Strangers 6	2012.0	0.0	3.1	14.0	NaN	Drama

Are K-dramas/shows' ratings related to their number of votes?

All dramas
(n = 1343)

r = .05[§]

Remove
Outliers

(n = 1143)   §: *p* = .06;
                    ****p* < .001

r = .12[***]
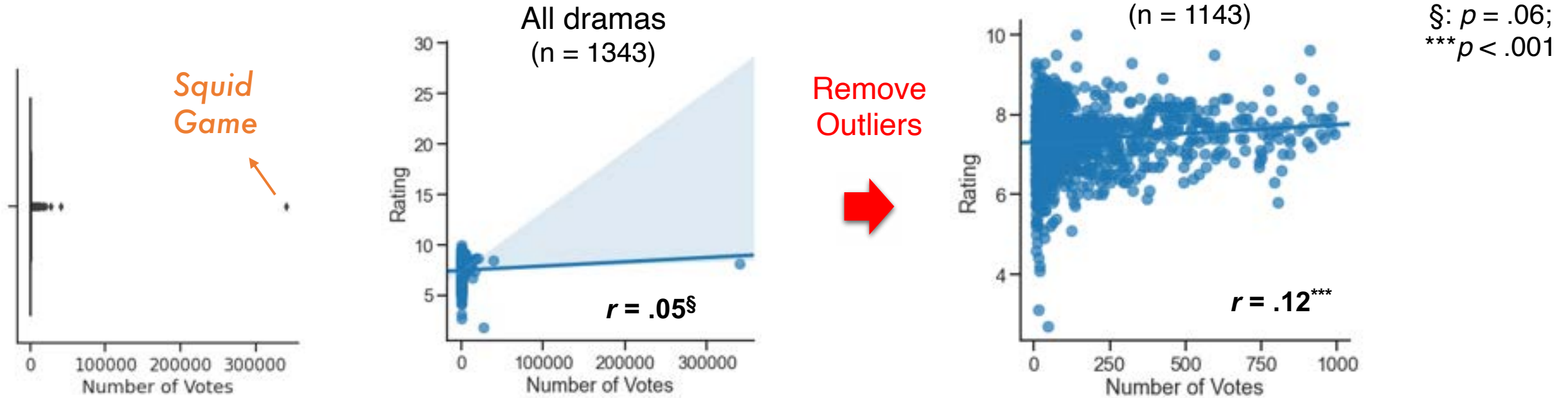
- One potential feature predictive of ratings is the number of votes (i.e., how popular it is).
- The dataset clearly contains some shows with many more votes than others (i.e., outliers). The one with more than 300,000 votes is Squid Game – the most popular show on Netflix ever[2].
- The correlation between ratings and number of votes is biased by "Squid Game"

- To address this issue, I removed outliers that have a value larger than $Q_3 + 1.5*IQR$
- The result shows that the larger the number of votes (more popular), the higher the rating a drama/show is.

- **Yes! The ratings of K-dramas/shows are related to the number of votes**

Are K-dramas/shows' ratings related to their episode length?

Might not be valid information

***$p < .001$

Based on background knowledge

Remove Outliers

Based on 1.5*IQR

(n = 762)

(n = 749)

r = .21***

r = .26***
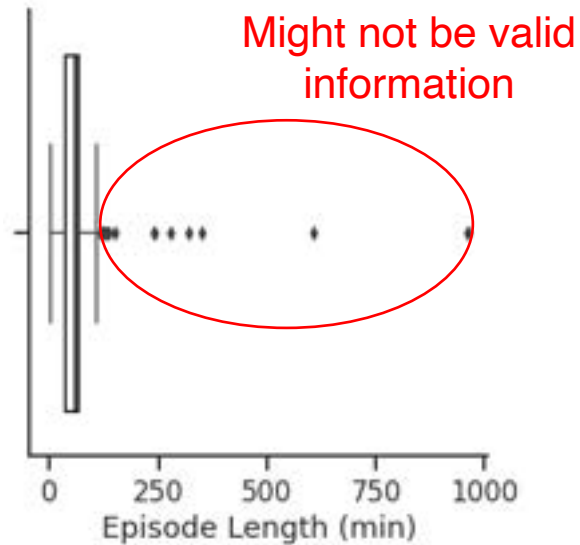
- Another feature that could be predictive of ratings is the episode length.
- Some dramas/shows have unreasonably long episode length (e.g., roughly 1,000 minutes for just one episode). This information might not be valid and should be removed before analyses.
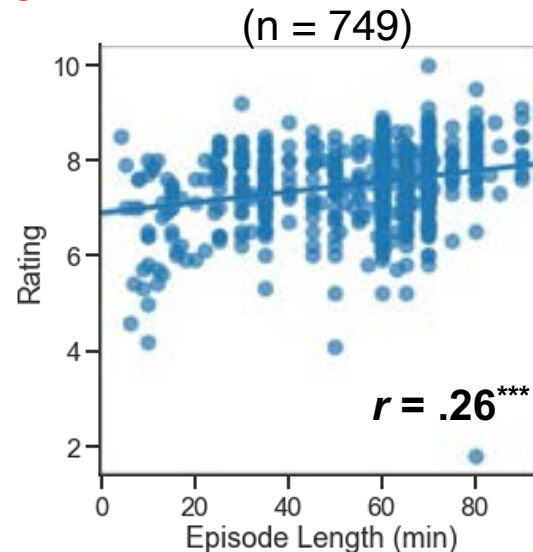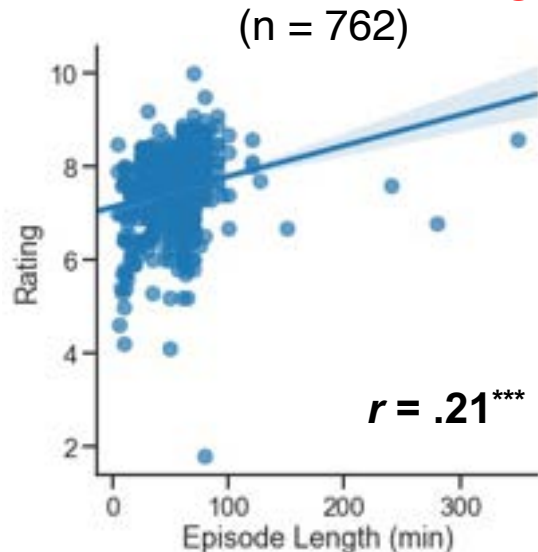
- The dramas/shows with unreasonable episode length (i.e., outliers) can be removed based on (a) background knowledge: remove dramas with episode length longer than 2 hours but keep music shows with 3 to 4-hour episode length, or (b) an objective threshold: *Q3 + 1.5*IQR*.
- Both approaches yield similar results: the longer the episode length, the higher the rating.

- **Yes! The ratings of K-dramas/shows are related to the episode length**

Linear Regression Model

Construct a regression model to predict ratings of K-drama/shows based on their number of votes and episode length

Dep. Variable:		Rating	**R-squared:**			**0.154**
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.8859	0.078	88.748	0.000	6.734	7.038
Votes	0.0001	1.48e-05	8.175	0.000	9.16e-05	0.000
Time	0.0092	0.001	6.906	0.000	0.007	0.012

(n = 749)

- Both features, number of votes (*"votes"*) and episode length (*"Time"*) are significant predictors of ratings ($p < .001$).
- Yet, the R^2 of this model is only 0.15, suggesting that a huge amount of variance (85%) is not explained by these 2 features (i.e., there may be a better model)
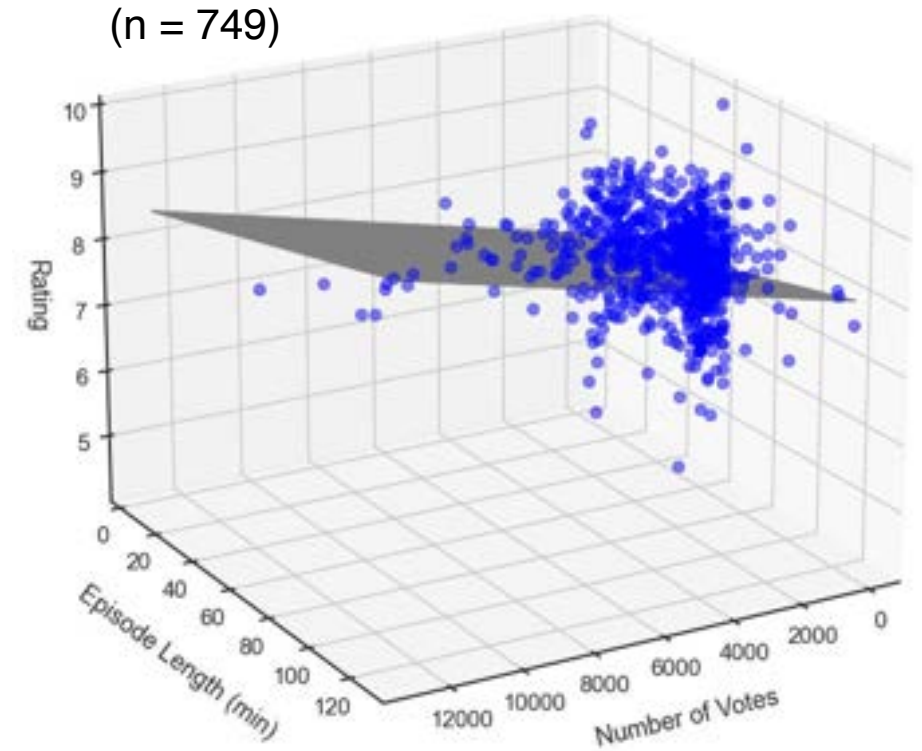
- Blue point: a drama/show's number of votes, episode length, and rating
- Gray plane: regression plane derived from the regression model

Who has starred in good dramas?

- Look for good K-dramas/shows by seeking for actors/actresses who have starred in dramas with high ratings before

- Calculate average ratings across dramas an actor/actress has starred in

- To get an accurate estimate, I only included actors/actresses who has starred in more than 10 dramas (n = 37)
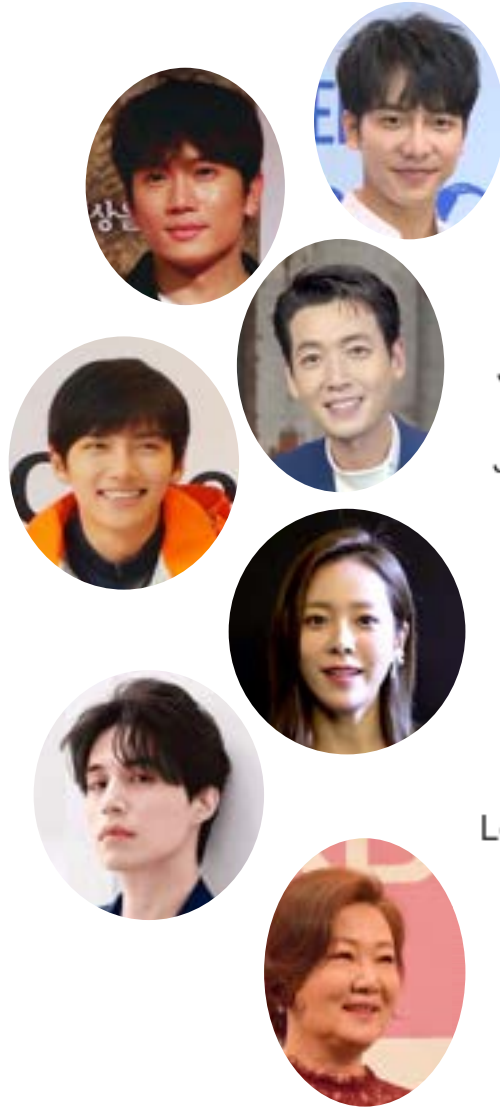
Top 10 Highest Average Rating

What words appear more frequently in the synopses of dramas with high ratings?

Generate a wordcloud showing the frequent words in high-rating dramas/shows across all genres or within a specific genre (note: some stopwords were removed, such as "South Korea", "Korean", etc.)

Across All Genres

Romance

Reality-TV

Based on 341 dramas with ratings equal to or larger than 8
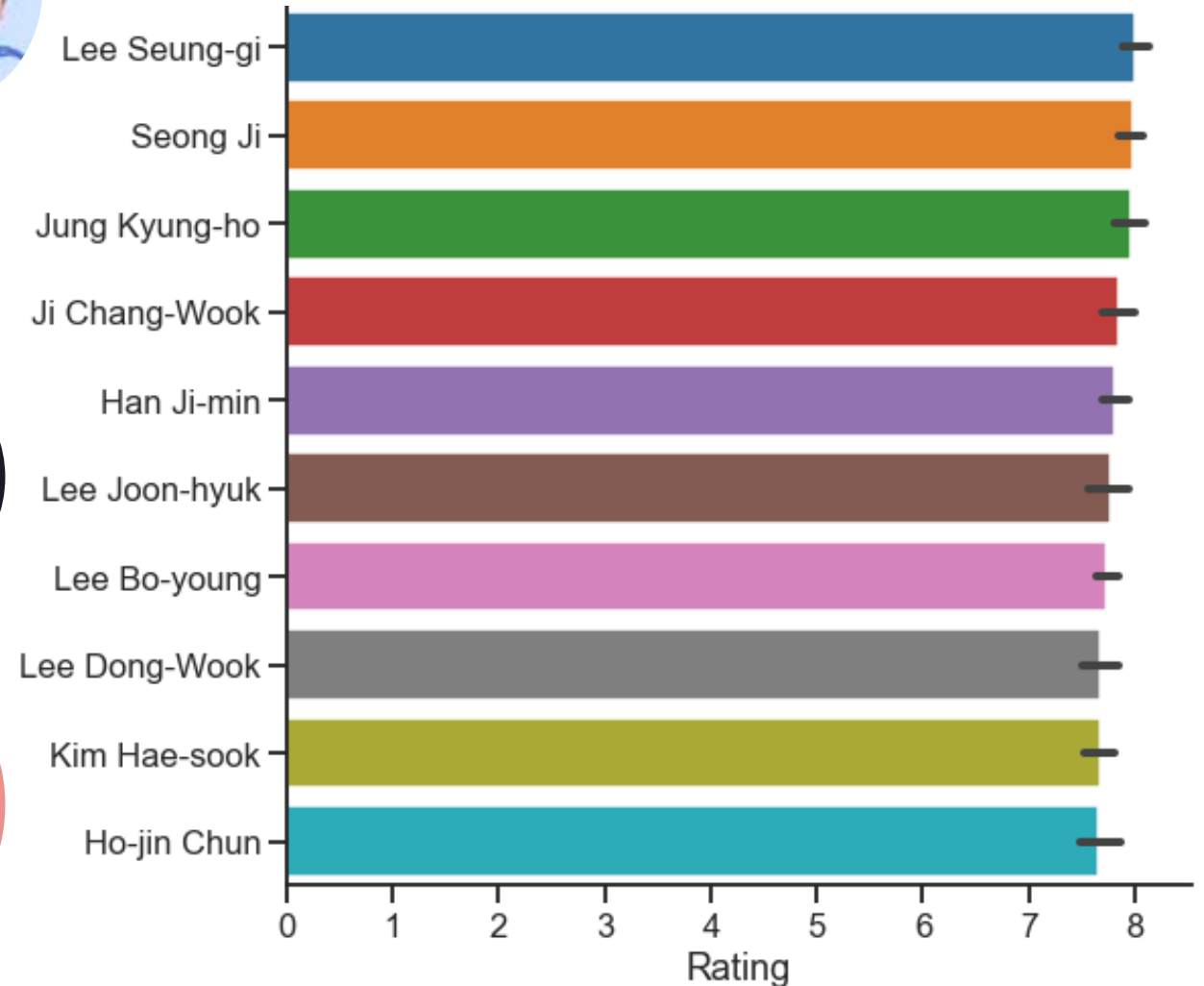
Looking for Good K-Dramas/Shows?

✓ **Choose one that has a longer episode length and more votes!**

✓ **Look for these actors/actresses!**

✓ **Look for these keywords in the synopsis!**

Python codes and more analyses are available on GitHub

Reference for K-Drama/Show Project

1. https://www.bbc.com/news/entertainment-arts-60505910

2. https://www.forbes.com/sites/paultassi/2021/10/13/squid-game-is-now-netflixs-most-popular-show-ever-and-its-not-even-close/?sh=25786fb16c42

3. Images of Korean actors/actresses:

- PLAYIN (2019). *Image of Lee Seung-gi* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:VGB_LeeSeungGi.jpg

- 뉴스인스타 (2018). *Image of Seong Ji* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:180212_지성.png

- 셀러게이트 (2021). *Image of Jung Kyung-ho* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Jung_Kyung-ho,_in_2021.png

- Rei Hyo (2014). *Image of Ji Chang-wook* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Jichangwook.jpg

- Marie Claire Korea (2019). *Image of Han Ji-min* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Han_Ji-min_at_Asian_Star_Awards_2018.jpg

- Leedongwook (2021). *Image of Lee Dong-wook* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:2473CD0.webp

- Kinocine PARKJAEHWAN (2019). *Image of Kim Hae-sook* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Kim_Hae-sook.2019.jpg

My M.A. Project – Face Image Reconstruction: Foundation for an Automated Sketch Artist

statistics, data-driven modeling, image processing, visualization, experiment/research, software development

Background

Human faces are unique visual stimuli we interact with almost everyday and the ability to recognize faces is crucial for social interaction.

However, what exactly we perceive and remember remains largely unclear.

Individual experience may influence how we perceive and remember a person's appearance

A data-driven image reconstruction technique that can recreate the appearance of faces from perception and memory based on a single human participant will have both theoretical and practical values

Data Acquisition Software – Face Similarity Rating Task

The first step in the pipeline of image reconstruction is to acquire data from human participants.

I designed a face similarity rating task, where human observers rated the visual similarity of pairs of faces using a 7-point scale (1: very different; 7: very similar).

There are two versions of this task: a perception-based and a memory-based version. Both of them have been used to collect more than 200 human participants' similarity ratings of facial identities and expressions in the laboratory and online.

Perception-based
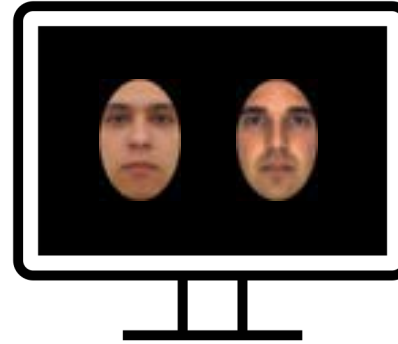
Human observers rate the visual similarity of pairs of faces presented on the screen by pressing a corresponding number key or moving a slider to select a rating.

Memory-based
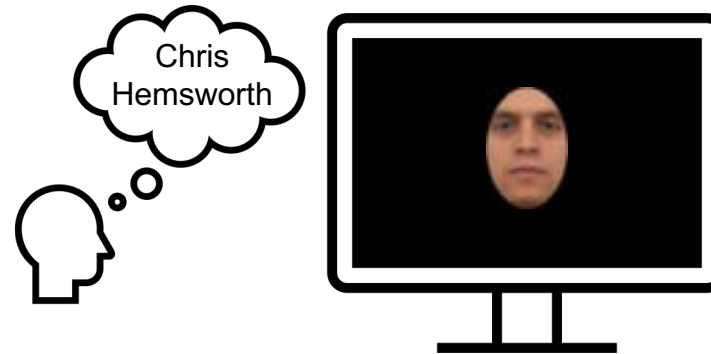
Human observers are instructed to recall a familiar face (e.g., a celebrity) and hold it in memory. Then, they rate the visual similarity between the remembered face and a series of faces presented on the screen by pressing a corresponding number key or moving a slider to select a rating.

Codes for the face similarity rating task are available on GitHub

Image Reconstruction Pipeline

1 **Face Similarity Rating**　　**2** **Face Space**　　**3** **Reconstruct a Face Image**
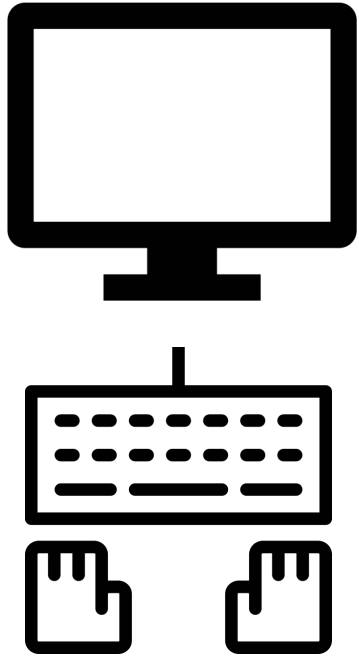
Human participants rate visual similarity of pairs of faces or pairs consisting of one face recalled from memory and one presented on the screen
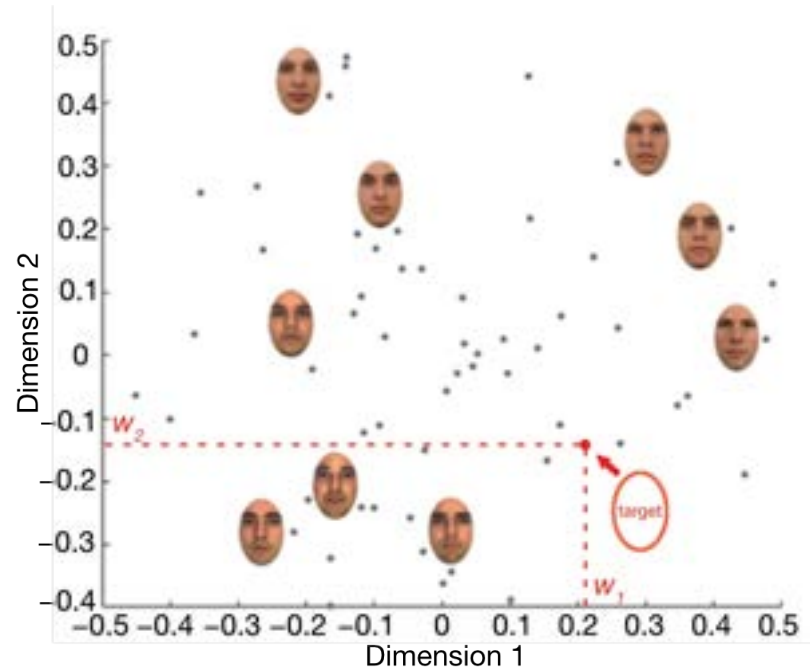
A multidimensional face space is constructed based on pairwise similarity ratings via multidimensional scaling, and the coordinates of the target face is estimated within the space
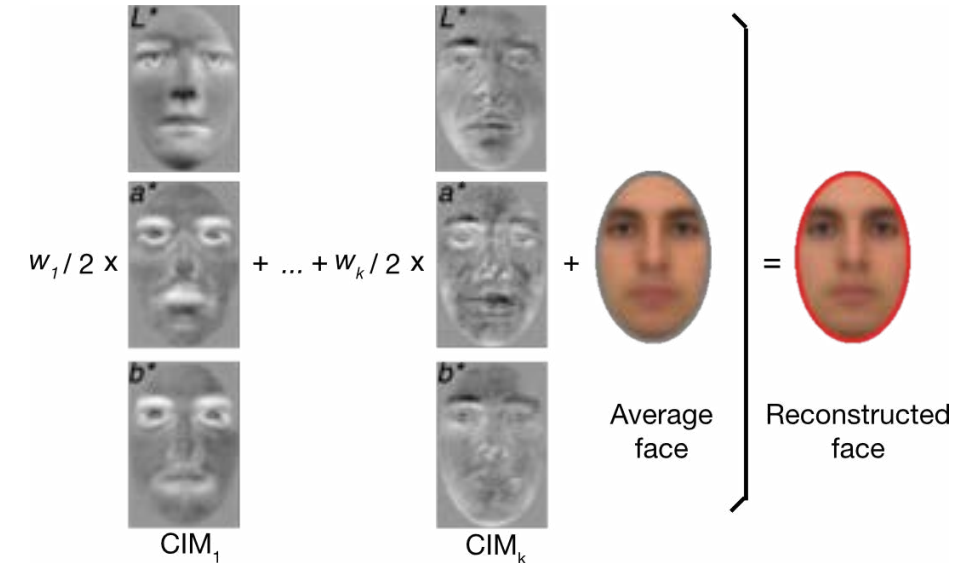
Significant visual features are derived for each dimension of the face space and are linearly combined to estimate the appearance of the target face (i.e., reconstructed face)

Evaluation of Reconstructions

Objective/Image-based

D_{target}

D_{other1}

$D_{other56}$

Original stimulus

Other alternative original stimuli
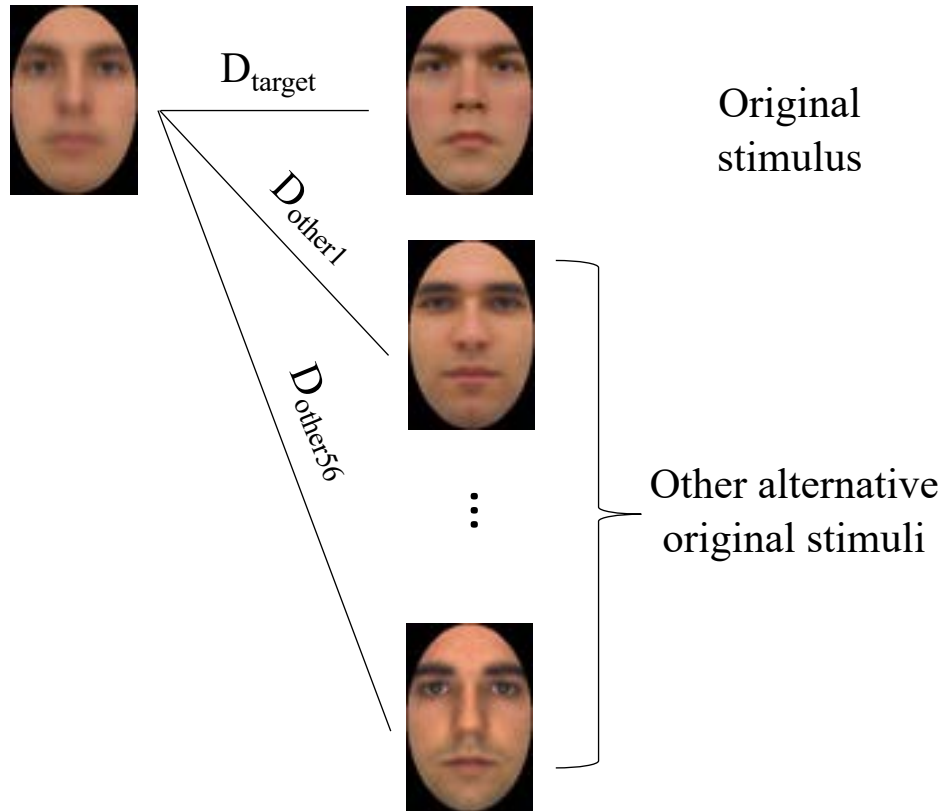
Objective accuracy is estimated as the proportion of instances for which the reconstruction is more similar, via Euclidean pixelwise similarity, to its original stimulus than any other image

Subjective/Experiment-based

Face Matching

Face-Name Matching

Brad Pitt   Johan Hill
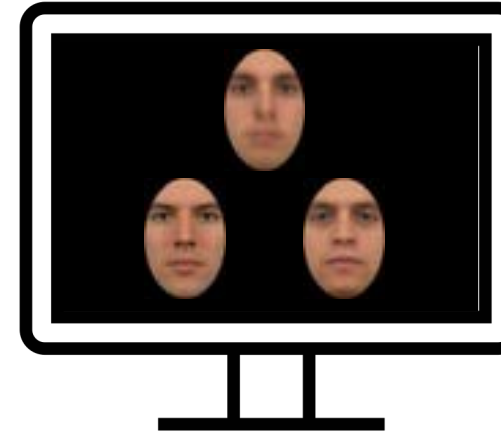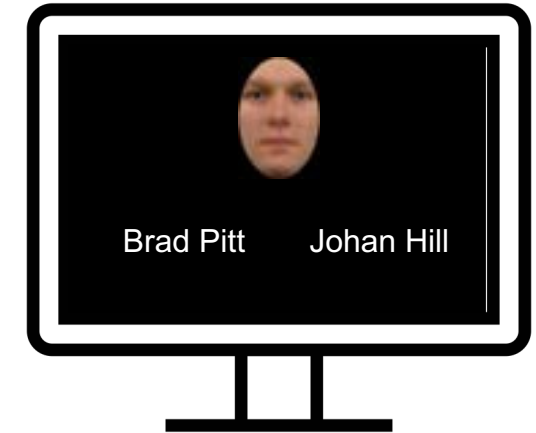
Face matching: The same participant and a group of naïve observers evaluate the similarity between a reconstruction (top) and two original stimuli (bottom; one of which is the target and the other is a foil).

Face-name matching: Participants determine which of the two named individuals is closest in appearance to the reconstruction

Reconstructed Face Images

Reconstruction   Stimulus   Reconstruction   Name

Jonah Hill

Chris Hemsworth

Perception-based reconstructions of unfamiliar faces

Memory-based reconstructions of famous faces
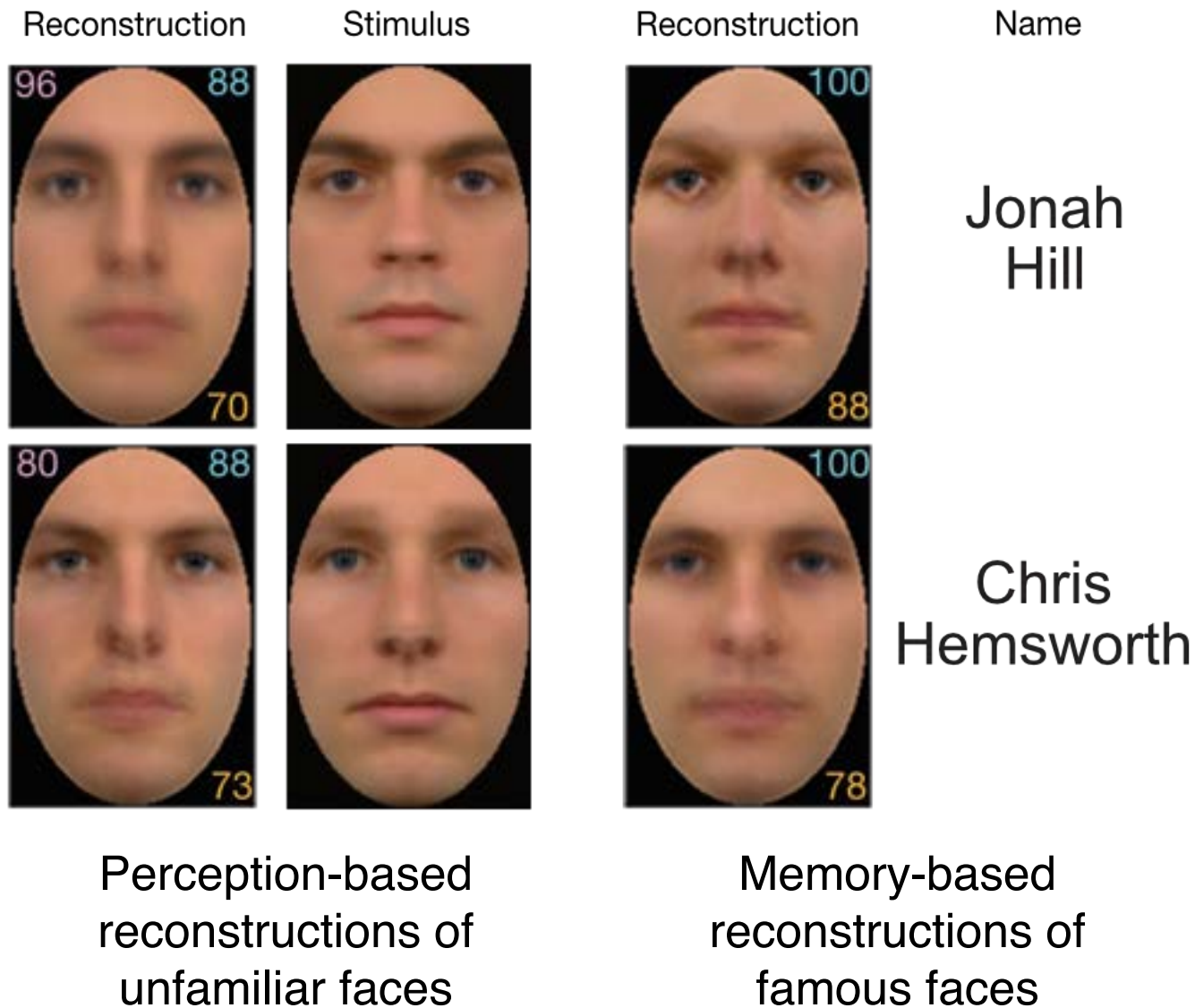
- Examples of reconstructed face images from one participant

- Accuracy is computed via (a) objective pixelwise similarity (top left), (b) subjective evaluation of the same person (top right), and (c) subjective assessment of a group of naïve observers (bottom right).

- Note: Objective accuracy cannot be estimated for famous faces reconstructed from memory since no corresponding stimuli are presented to participants

Average Reconstruction Accuracy
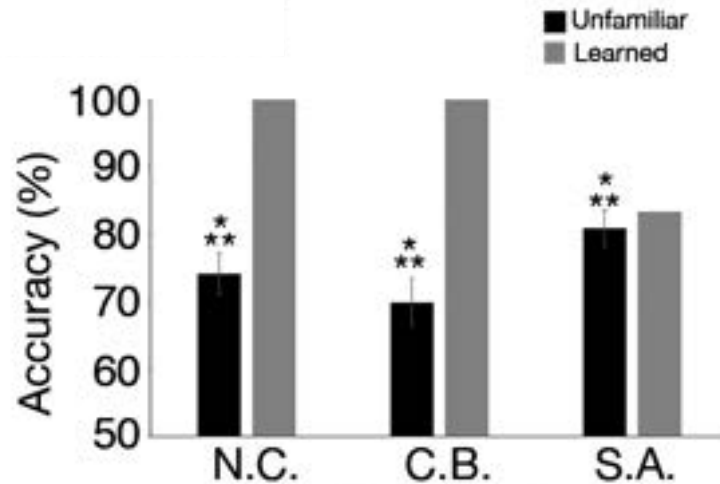
Objective

Subjective (Same Participant)

Subjective (Naïve Observers)
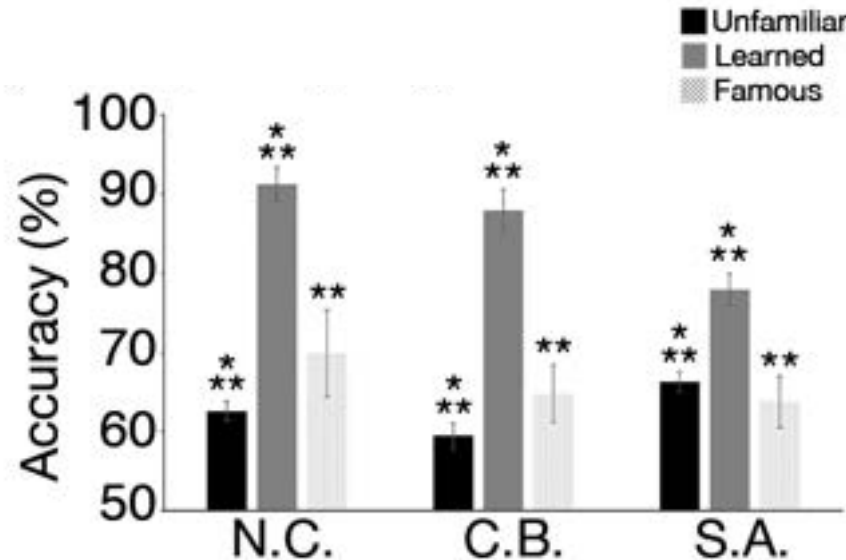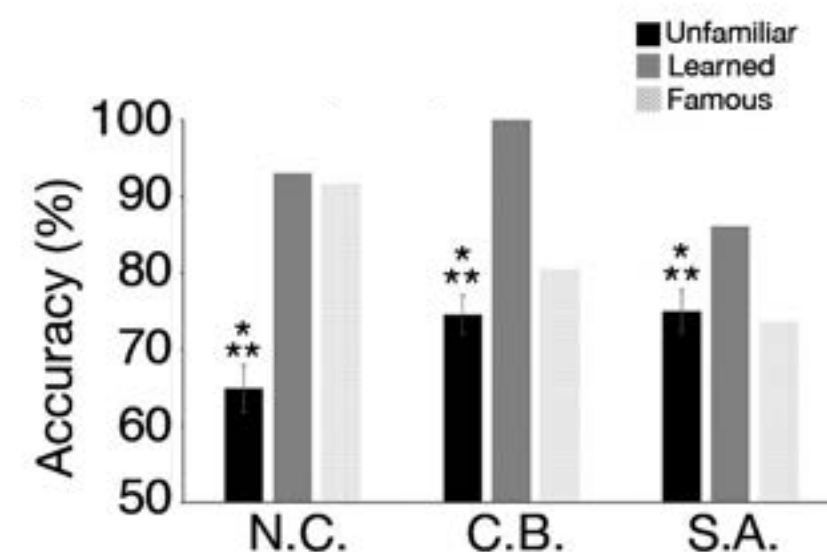
- Results of three participants

- Average reconstruction accuracy across all images is computed separately for (a) perceived unfamiliar faces, (b) unfamiliar faces learned by participants and recalled from memory, and (c) famous faces recollected from memory

- Accuracy is significantly above chance (50%) for all reconstructions recovered from perception and memory

one-sample t-test;
**p < .01; ***p < .001;
Error bar: ±1SE

Conclusions for Face Image Reconstruction Project

- Image reconstruction technique, coupled with a simple, intuitive task, can successfully recreate the appearance of faces human participants perceive and recall from memory

- The success of this methodological framework opens paths to novel applications, such as an automated sketch artist
  - Forensic application: depict and visualize the face of a person of interest without verbal descriptions

- In my PhD work, the applicability and utility of the present methodology are tested across a diverse population, including individuals varying in age and cognitive functions.

Reference for Face Image Reconstruction Project

- Figures are adapted from my publication: Chang, C. H., Nemrodov, D., Lee, A. C., & Nestor, A. (2017). Memory and perception-based facial image reconstruction. *Scientific reports*, *7*(1), 1-9.

- Face images used in this study are obtained from four databases: Radbound[1], AR[2], FEI[3], and FERET[4,5]

1. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and emotion*, *24*(8), 1377-1388.

2. Martinez, A., & Benavente, R. (1998). The AR face database: *CVC Technical Report*, 24

3. Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and vision computing*, *28*(6), 902-913

4. Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, *16*(5), 295-306

5. Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(10), 1090-1104

My PhD – Image Reconstruction Visualizes the Vulnerability of Face Perception and Memory

statistics, data-driven modeling, PCA,
classification, correlation, image processing,
visualization, experiment/research,
hypothesis testing

Abstract

Extensive research has shown that face processing can be disrupted by various factors; however, the extent of the impact that these factors have upon face representations remains unclear. My PhD work capitalizes on the behavioural-based image reconstruction technique developed in my M.A. project to examine individual differences in visual representations underlying face perception and memory and how they are impacted by aging, brain damage, and personality disorder traits. I found that all of them negatively affected a wealth of visual information crucial for recognition of facial identity and expression (e.g., eye shape, skin colour, expressiveness). My findings not only confirm hypotheses and theories in the field, but also demonstrate the applicability and utility of image reconstruction across a broader population varying in age, brain health and features related to personality disorders.

Reconstructions From Younger and Older Adults

Stimuli   YA   OA
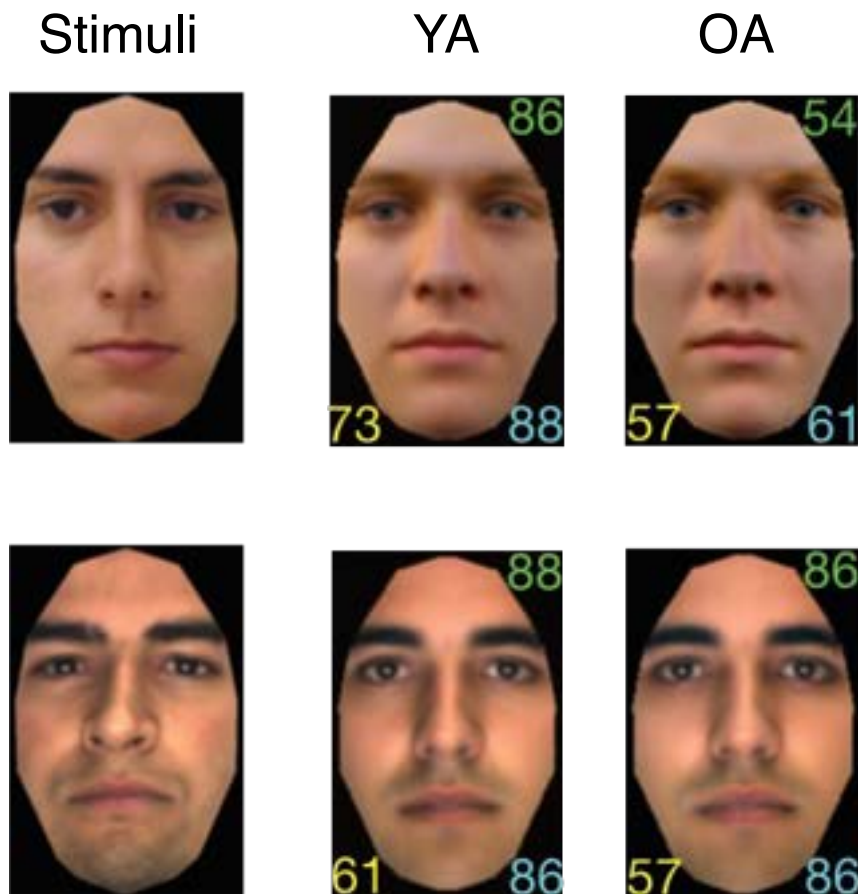
- Image reconstruction was applied to 21 younger (YA) and 24 older adults (OA) to investigate the nature of age-related deficits in face perception

- Examples of original stimuli and corresponding reconstructed face images from one YA and one OA are shown.

- Accuracy (%) is separately provided for reconstructed shape (bottom left), surface (bottom right) and the overall appearance (top right).
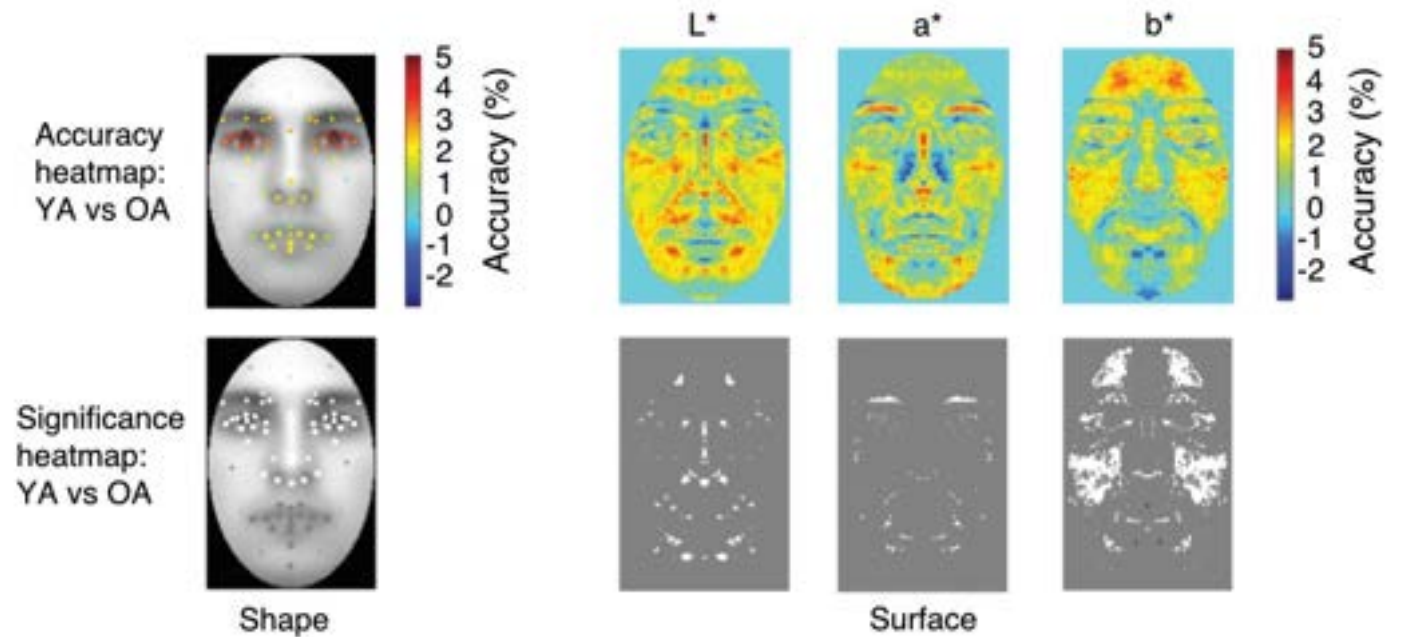
# Age-Related Difference in Reconstruction Accuracy
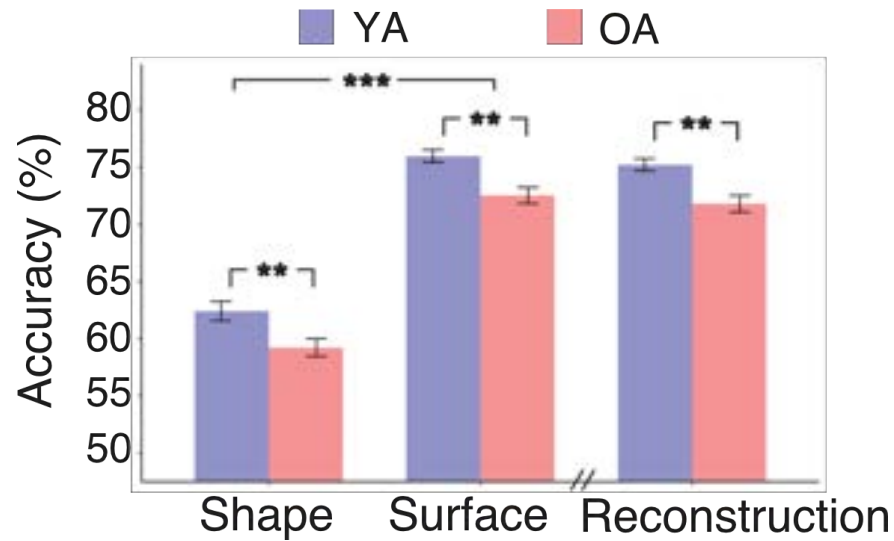
- Objective average reconstruction accuracies are lower for older adults (OA) than younger adults (YA) for all three types of reconstructions (assessed via a two-way ANOVA and an independent-sample t-test).

**p < .01; ***p < .001; Error bar: ±1SE

- A heatmap depicting differences in accuracy between the two groups is separately computed for shape, for each facial marker, and for surface, for each pixel. The significance of the difference is evaluated via a t-test.
- White colour in the significance heatmap indicates information is more accurately represented by YA than OA. The visual representations of the shape of eyes and nose, as well as those of skin colour are compromised in OA.

Individual Variability in Face Representations
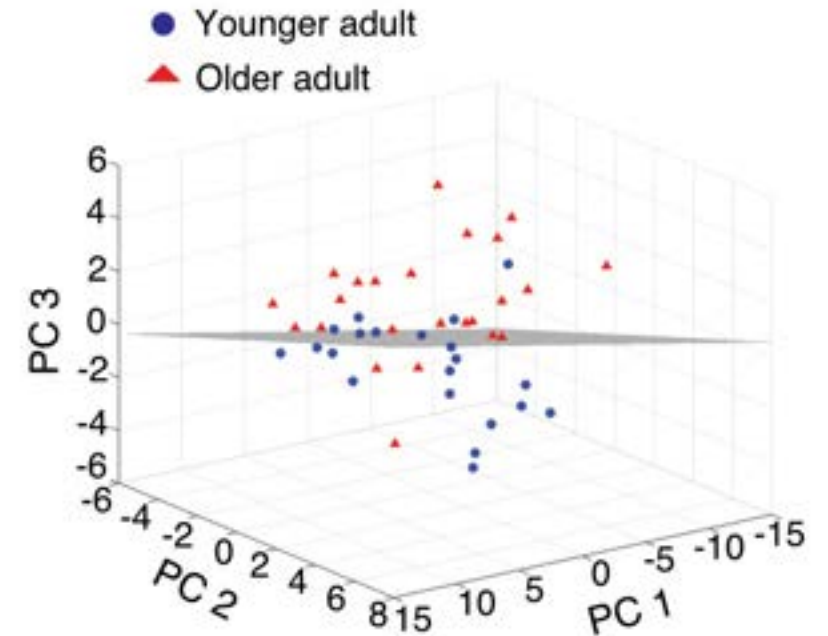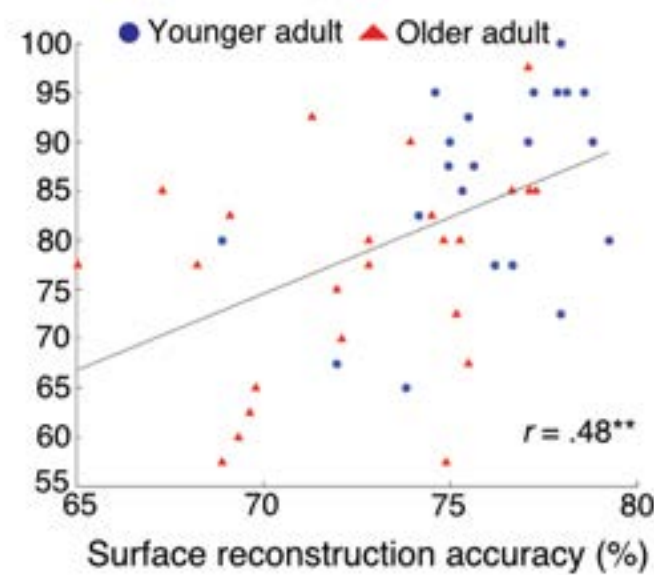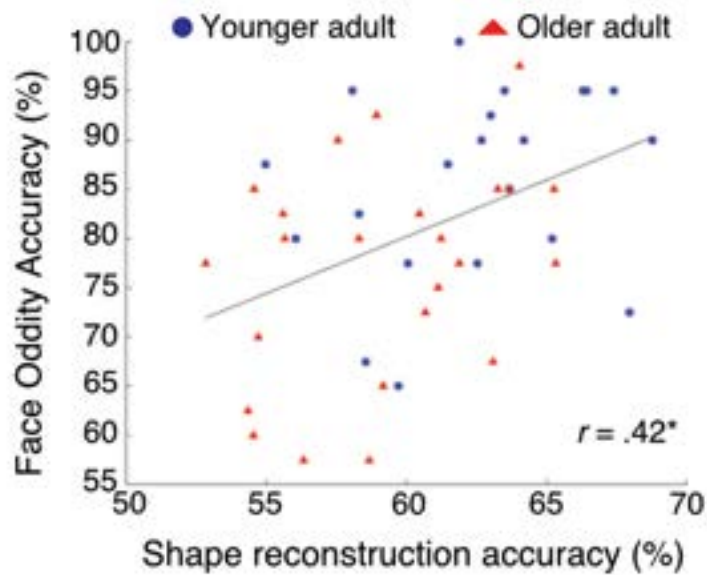
- Individual variability in representational accuracy of shape and surface across younger and older adults are associated with differences in the ability to process feature conjunction, measured by a different-view face oddity task.

*p < .05; **p < .01*

- A participant pace is constructed by applying principal component analysis (PCA) to pairwise face similarity ratings across people in both age groups.
- A linear discriminant analysis (LDA) classifier predicts participant's age group at 86% accuracy.

Reconstructions From An Amnesic Patient BL
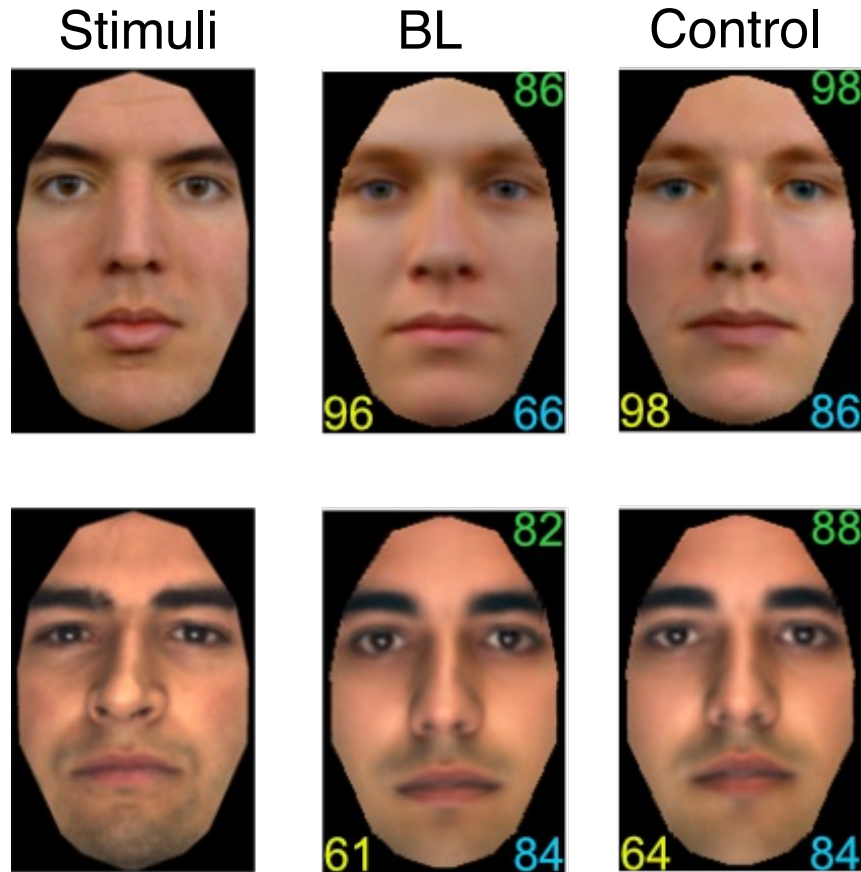
Stimuli   BL   Control

- Image reconstruction was applied to 2 amnesic patients and 33 neurologically healthy controls to examine how damage to brain regions crucial for memory impacts face perception.

- Examples of original stimuli and corresponding reconstructed face images from the amnesic patient BL, who has damage to his hippocampus, and one healthy control are displayed.

- Accuracy (%) is separately provided for reconstructed shape (bottom left), surface (bottom right) and the overall appearance (top right).

Impact of Brain Damage on Face Representations
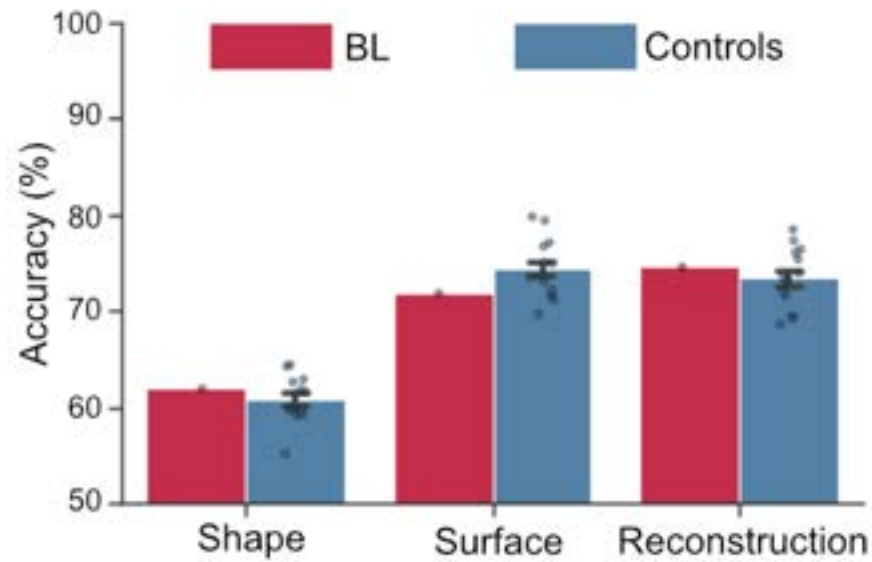
- Accuracy of reconstructed images is not different between BL and the healthy control participants.
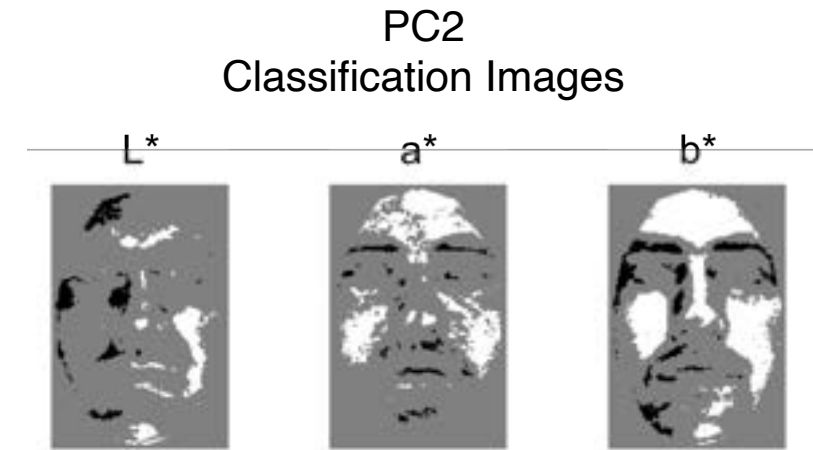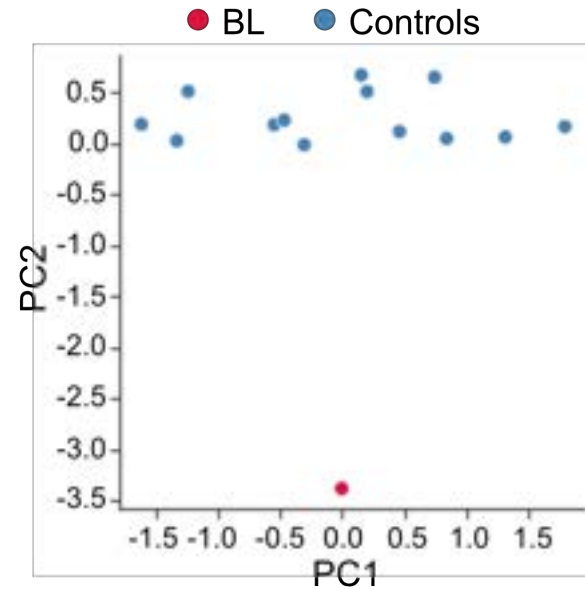
- However, a separation between BL and the controls in the participant space, constructed by applying PCA to pairwise face similarity ratings, suggests that BL's face representation differs from controls.

- Classification images show that PC2 in this space captures differences in pixel intensity values, particularly in the colour channels (a*, b*). This indicates that BL may have less access to colour information relative to controls in face perception.

Reconstructed Facial Expressions From Individuals Varying in Personality Disorder Traits

Stimuli | Perception-based | Memory-based
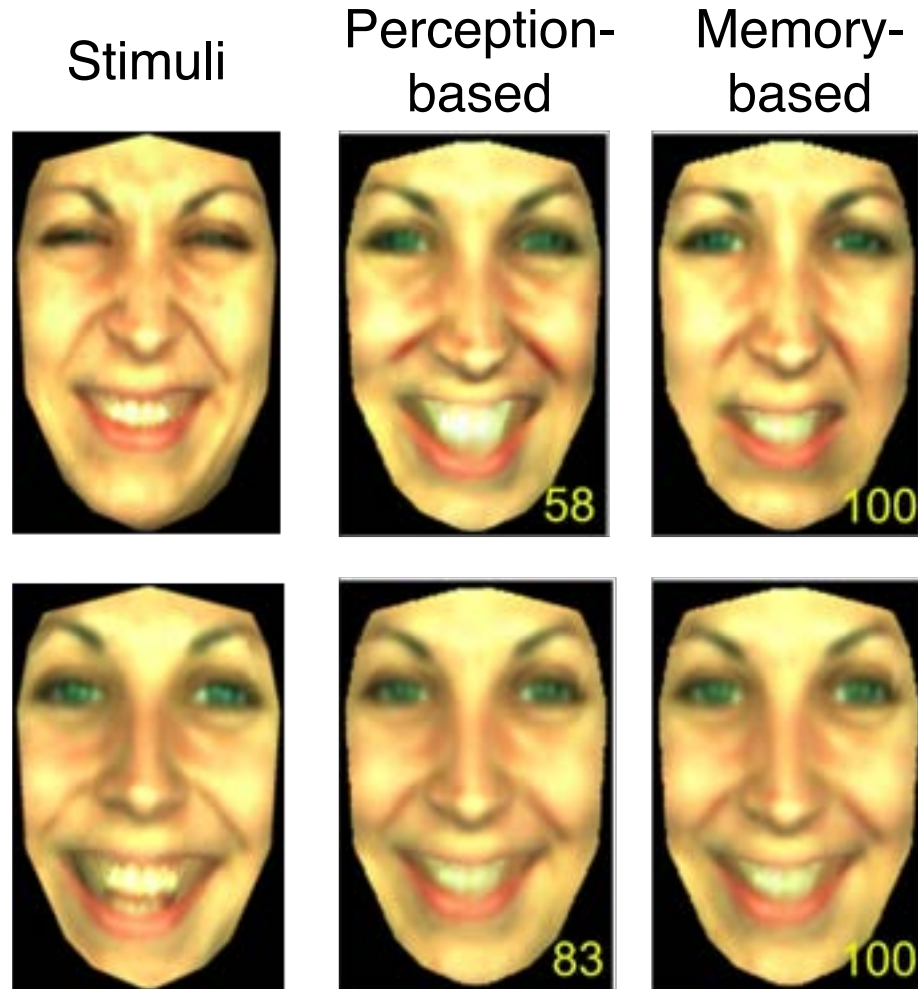
- Image reconstruction was applied to data of 89 community-dwelling human participants to study the association between personality disorder traits and visual representations of expression perception and memory.

- Examples of original stimuli of facial expressions (top row: disgust; bottom row: happy laughing) and corresponding reconstructions from perception and memory are shown.

- Numbers represent accuracy (%) computed based on the overall appearance.

Visual Characteristics Associated With Personality Disorder Traits

Perception PC1

$r = -.32*$

Self-Direction Score

Stimuli   Reconstructions from people with

Greater impairments   Less impairments
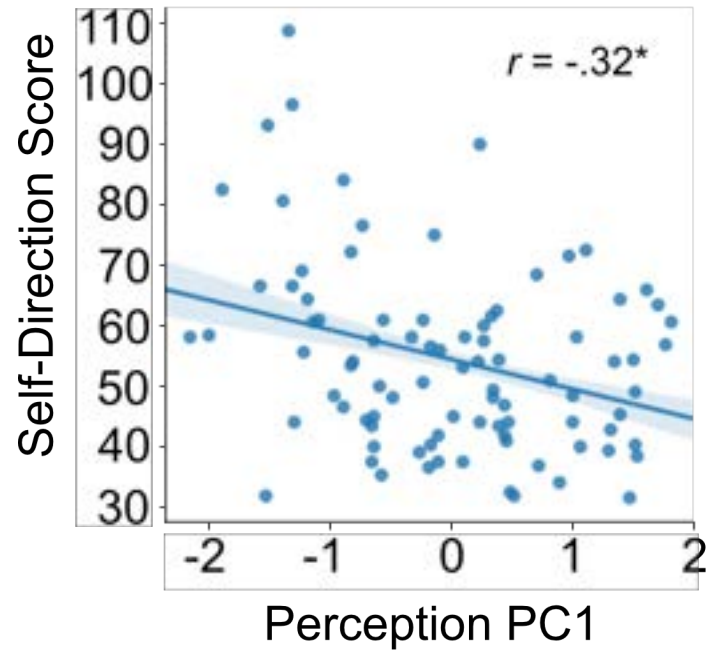
- Differences in visual representations of expressions across participants are associated with individual variability in the dysfunction of a self-related trait – self-direction (impulsivity; the ability to regulate behaviours).
  - The higher the self-direction score, the greater the severity (impairment)

- Average reconstructed expressions (top row: confused; bottom row: smiling yeah right) from individuals with greater and less impairments in self-direction. Those from people with more severe self-direction dysfunction are less expressive than their counterparts (validated by 81 naïve human observers; $p < .01$).

My PhD Work Conclusions

- Prove the applicability and utility of the image reconstruction technique to recover the appearance of facial identities and expressions from perception and memory, from a large, diverse group of individuals varying in age, brain health, and personality disorder traits.

- The findings elucidate the impact of various conditions on face processing, and provide actionable insights into how we can help individuals who are vulnerable to these risk factors
  - Improve older adults' ability to recognize faces by training them to focus on diagnostic features (e.g., eyes, skin colour)

Codes for the shape and surface image reconstruction pipeline are available on GitHub

References for My PhD Work

- Figures for age-related decline in face perception are adapted from my publication: Chang, C. H., Nemrodov, D., Drobotenko, N., Sorkhou, M., Nestor, A., & Lee, A. C. (2021). Image reconstruction reveals the impact of aging on face perception. *Journal of Experimental Psychology: Human Perception and Performance*.

- Images of facial identities are obtained from four databases: Radbound[1], AR[2], FEI[3], and FERET[4,5]

   1. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and emotion, 24*(8), 1377-1388.
   2. Martinez, A., & Benavente, R. (1998). The AR face database: *CVC Technical Report*, 24
   3. Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and vision computing, 28*(6), 902-913
   4. Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing, 16*(5), 295-306
   5. Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence, 22*(10), 1090-1104

- Images of facial expressions are obtained from the MPI facial expression database[6]

   6. Kaulard, K., Cunningham, D. W., Bülthoff, H. H., & Wallraven, C. (2012). The MPI facial expression database - a validated database of emotional and conversational facial expressions. *PLoS ONE, 7*(3).