The Speed Dating Experiment

Jose Reyes

02/01/2018

Introduction

The Modern-day Dating

- There's a plethora of web-based dating apps!

- Various dating apps/services:

    1. Personality-based online dating: Match.com and eHarmony.com

    2. Location-based apps: Happn, OkCupid, and Tinder

    3. Online/Real-life Hybrid: Match.com/Speed-Dating Events.

Modern-Day Dating (con't)

- Why such apps exist:
  - Convenient way to meet people.
  - Plenty of options!
  - Search potential mates based on like-minded preferences.
- Dating App Short-comings:
  - Profiles are often inactive.
  - Strictly a numbers game!
  - Other attractive attributes not able to be communicated in a profile can be overlooked.

The Solution: Speed-Dating!

- If we can analyze participant ratings/preferences in a speed-dating event we can :

    1. Identify the key factors that makes a man/woman attractive.

    2. Provide guidance how to improve attractive factors for dating success

    3. Predict whether a speed-dating match!

- Using Python and the Speed Dating dataset from the Kaggle website, we will aim to predict whether a female speed-dating partner would like to see the male partner again!

The Dataset

Dataset Inspection

- Dataset contains 195 columns and 8,378 observations
- Collection of ratings from 4-minute dating encounters over twenty-one speed dating events, held between Oct. 2002 and April. 2004.
- Participants are a mix of graduate and undergraduate students/faculty from Columbia University.
- 551 participants: 274 men & 277 women.

Data Wrangling

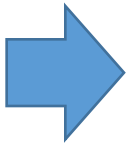
Initial Data Quality Findings

- Relatively clean dataset with minor exceptions:
  - Missing values in columns of interest.
  - Duplicate data: same speed-date encounter from two perspectives, male and female.
  - Binary data not binary in columns of interest. No explanation provided.

Data Wrangling

Original dataframe

- 195 columns and 8,378 rows.
- Missing data in key columns.
- Incorrect Binary values

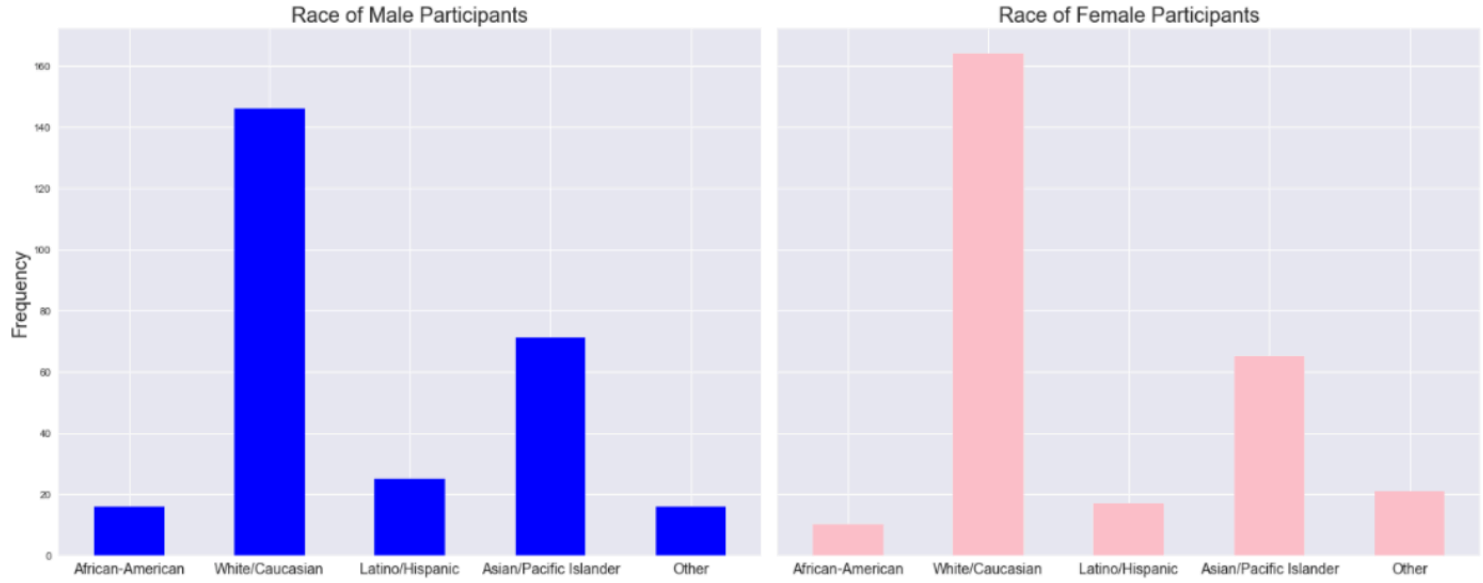


Revised dataframe

- 35 columns and 4,184 rows.
- Imputed with most freq. value.
- Corrected to true binary values.

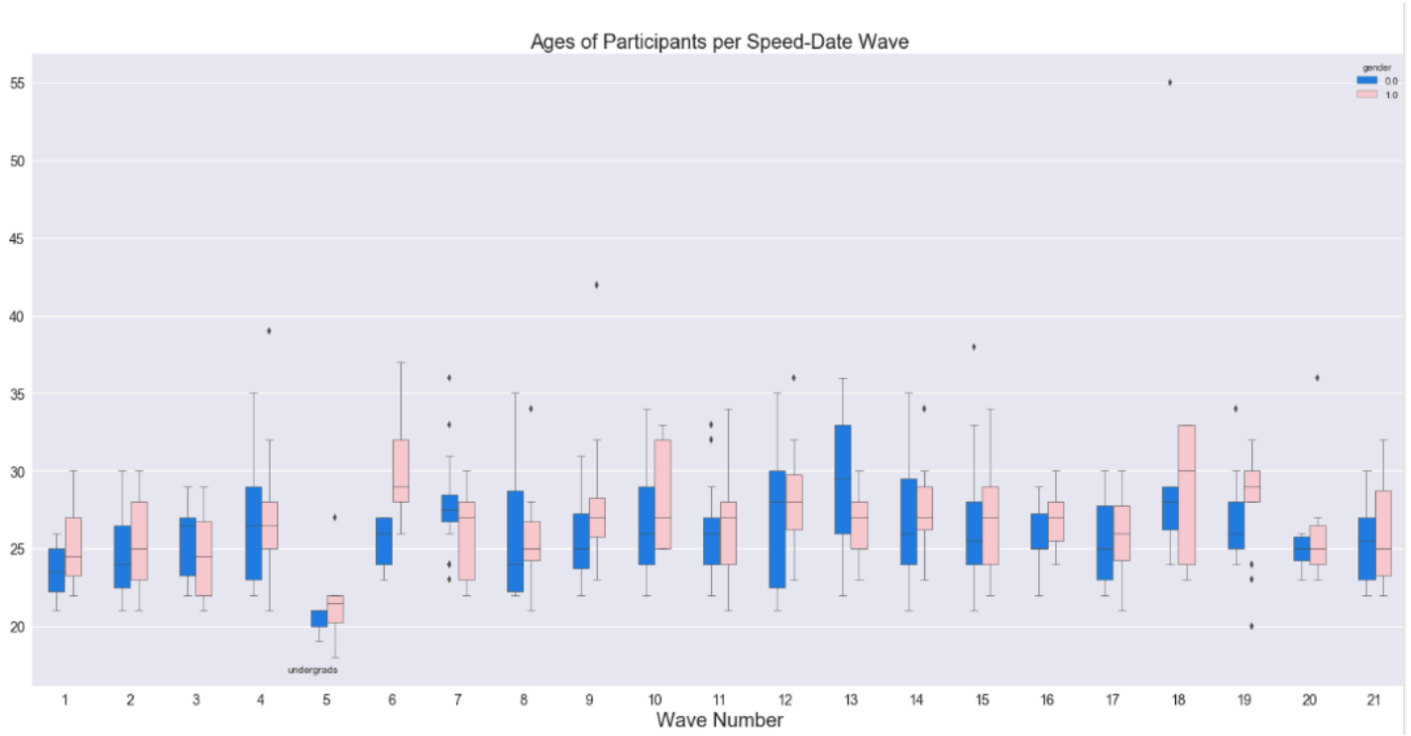
The Participants

Participant Demographics

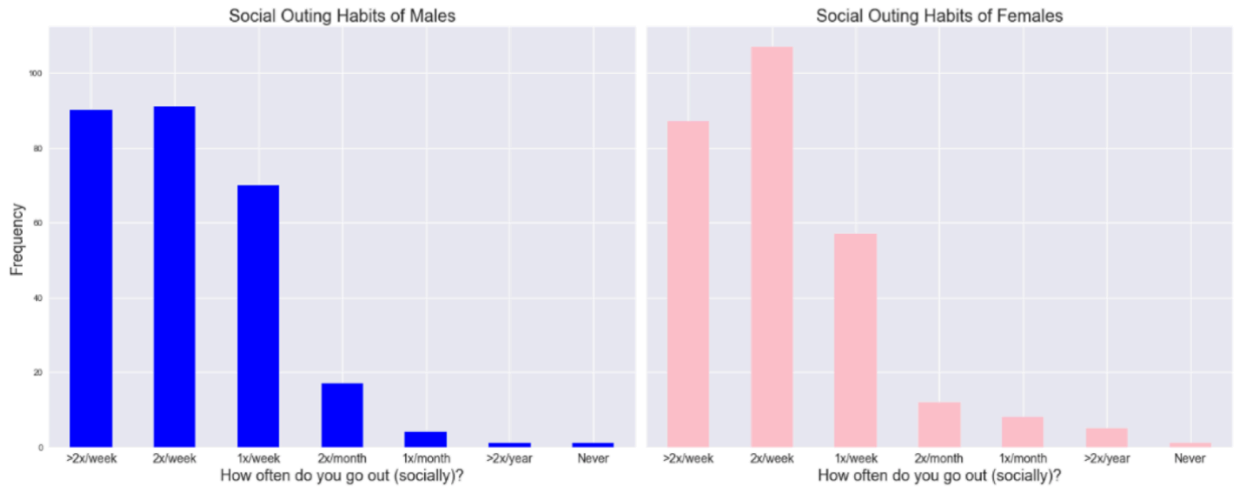


	African-American	White/Caucasian	Latino/Hispanic	Asian/Pacific Islander	Other
Male	16	146	25	71	16
Female	10	164	17	65	21

Participant Demographics (con't)

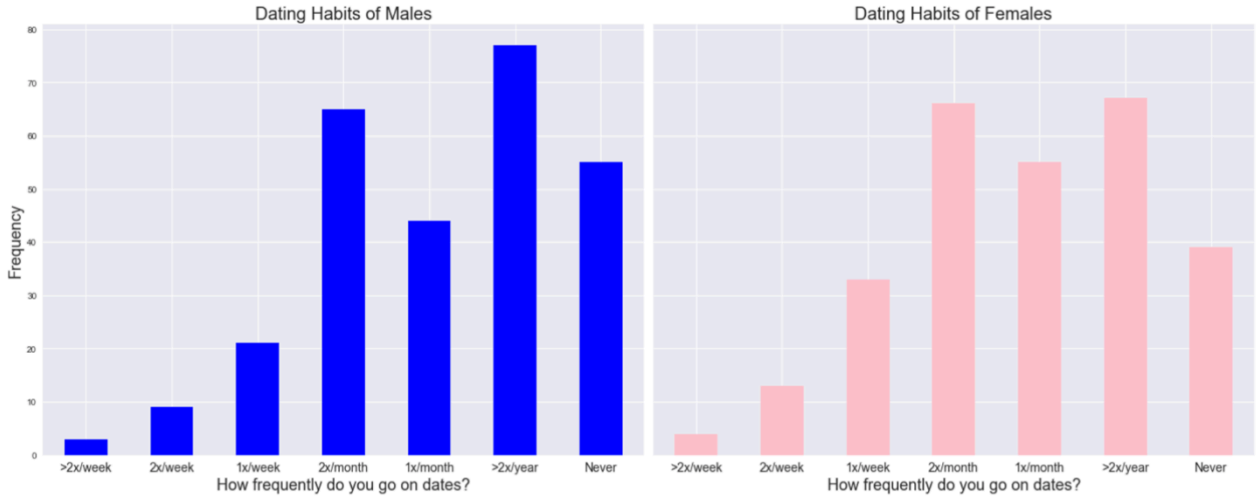


Participant Social/Dating Habits (con't)



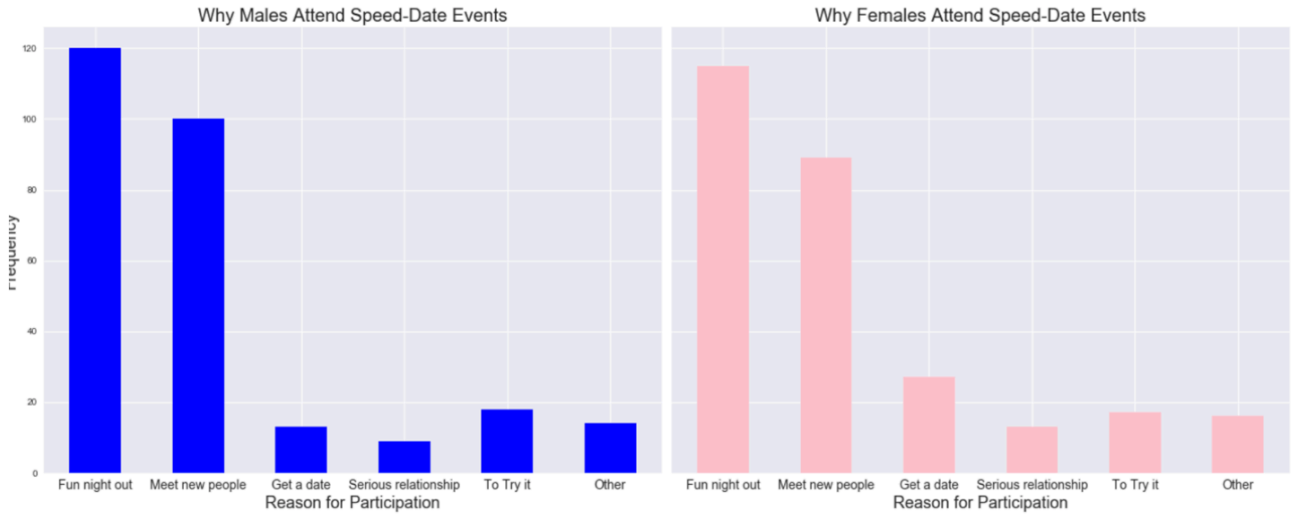
	>2x/week	2x/week	1x/week	2x/month	1x/month	>2x/year	Never
Male	90	91	70	17	4	1	1
Female	87	107	57	12	8	5	1

Participant Social/Dating Habits (con't)



	>2x/week	2x/week	1x/week	2x/month	1x/month	>2x/year	Never
Male	3	9	21	65	44	77	55
Female	4	13	33	66	55	67	39

Participant Participation Purpose



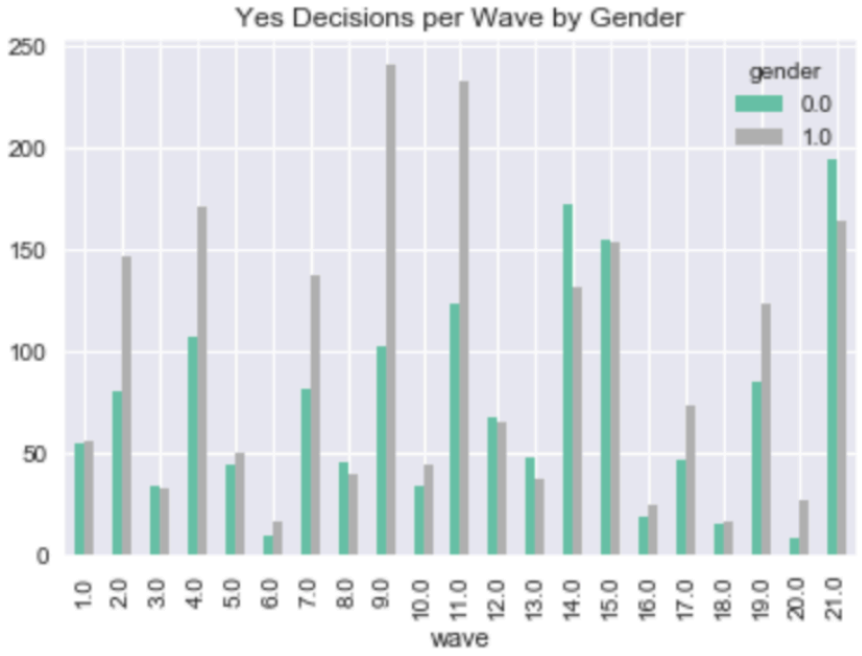
	Fun night out	Meet new people	Get a date	Serious relationship	To Try it	Other
Male	120	100	13	9	18	14
Female	115	89	27	13	17	16

Summary: The typical participant

- 26-years old
- Caucasian
- Inversely proportional dating/social habits: goes out often, infrequently dates.
- Participation in speed-dating event is purely recreational.

Inferential Statistics

Participant “Yes” responses per event

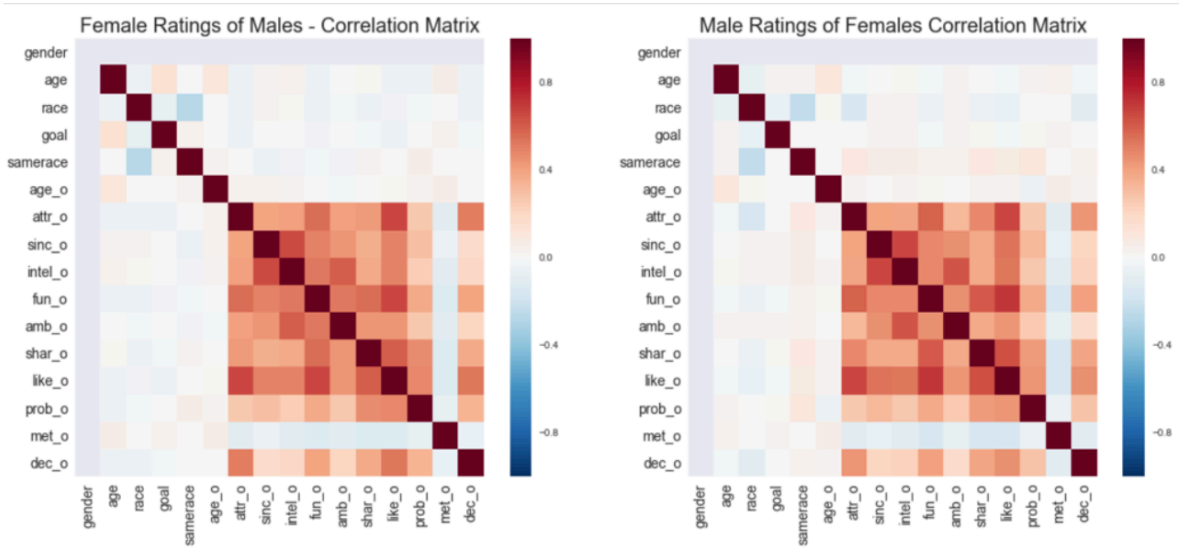


“Yes” responses by gender

wave	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Male	55	80	34	107	45	9	82	46	103	34
Female	56	147	33	171	50	16	138	40	241	45

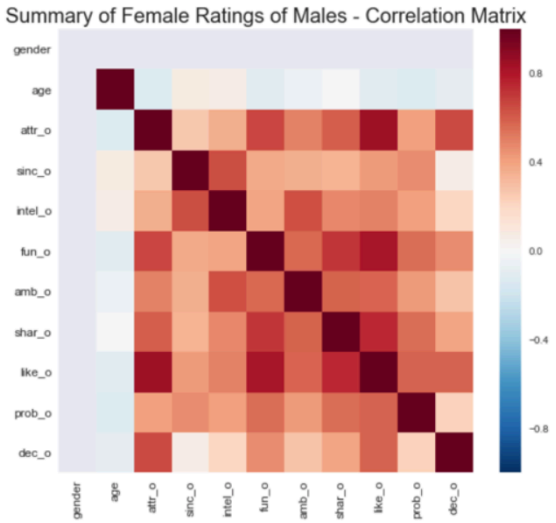
wave	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0
Male	123	68	48	172	155	19	47	15	85	8	194
Female	233	65	38	132	154	25	74	17	123	27	164

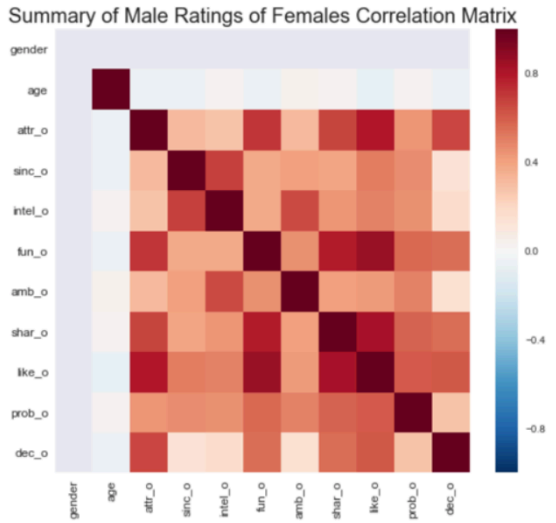
Correlation Studies



	age_o	attr_o	sinc_o	intel_o	fun_o	amb_o	shar_o	like_o	prob_o	met_o	dec_o
Ratings_F_of_M	-0.006218	0.514399	0.177262	0.200603	0.394551	0.205370	0.369519	0.527882	0.336398	-0.067802	1.0
Ratings_M_of_F	0.018818	0.440536	0.208706	0.222494	0.399644	0.167687	0.389185	0.458861	0.282691	-0.096556	1.0

Correlation Studies (con't)



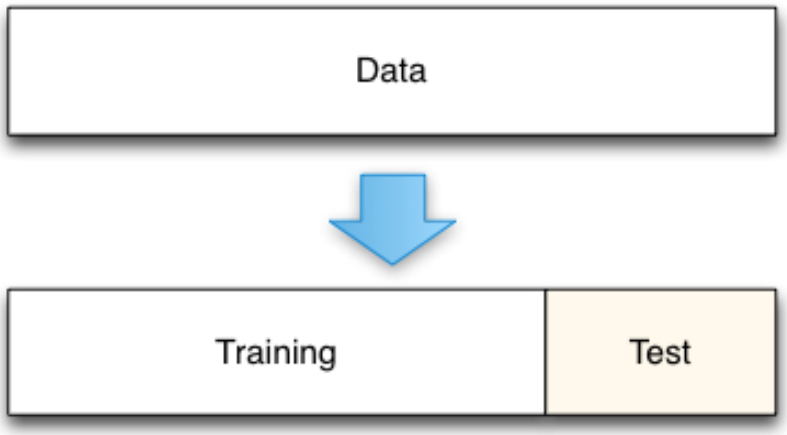


	gender	age	attr_o	sinc_o	intel_o	fun_o	amb_o	shar_o	like_o	prob_o	dec_o
Summary_Ratings_F_of_M	NaN	-0.089238	0.651337	0.058880	0.210556	0.464848	0.283609	0.386909	0.585656	0.233987	1.0
Summary_Ratings_M_of_F	NaN	-0.048572	0.662869	0.126484	0.156710	0.547985	0.147924	0.557225	0.610379	0.286299	1.0

Building the Algorithm

Training/Testing Set Creation

- Utilized Python's Scikit-Learn machine learning library.
- Dataset was split into striated training and testing sets:
  - Training set = 80% of data
  - Testing set = 20% of data
- Target variable: "yes" partner response.
- Features: Speed-Date Questionnaire ratings.



Results

- Each algorithm was trained using the training set.
- Each algorithm was fitted to the testing set to assess model performance.
- Each algorithm was fine-tuned to improve performance.

Algorithm	Best Test Score
Decision Tree	78.14%
Gradient Boosted Decision Tree	**82.32%***
Random Forest	82.20%
Logistic Regression	80.41% (AUC = .8039)
Logistic Regression w/ Standard Scaler	80.53% (AUC = .8053)
Logistics Regression w/ Standard Scaler & less features	80.65% (AUC = .8064)
Logistics Regression w/ Standard Scaler & top 3 features.	76.70% (AUC = .7675)

***Best Score**

Discussion/Recommendations

Discussion

- The best performing algorithm was the Gradient Boosted Decision Tree. It performed the best in predicting a “yes” response from a female participant.

- A score of 82.32% is decent, but preferable if 90% or above.

Recommendations

- Use additional data found in the dataset, namely hobbies/interests. Do people with shared interests correlate with a positive match?

- Collect additional data, namely a transcript of each 4-minute conversation for Natural Language Processing. Can language predict a match?

- Collect non-invasive thermal imaging data to assess the genuiness of a “yes” response. Can biological indicators predict a “yes” response?

- Utilize more sophisticated predictive techniques (Deep Learning/Neural Networks).

Questions?