Information Acquisition Deficit Detection and Mitigation

(F4FGA08305J006)

PI: Dr. Brett J. Borghetti, PhD (AFIT)

Co-PI: Dr. Mark E. Oxley, PhD (AFIT)

AFOSR Program Review:
Computational Cognition and Machine Intelligence Program (CCMI)
PM: Dr. James Lawton
6 Oct 2020, (Virtual Meeting via Zoom)

IAD Detection & Mitigation (Borghetti & Oxley)

Research Objectives:

Detect and mitigate human biases occurring during information acquisition tasks

- Operators are subject to biases during decision-making
- Our goal is to automatically detect the biases and inform the operator of their presence, to help mitigate their effect

Technical Approach:

Four Phases

1. Collect physio signals from human experiment
2. Develop machine learning models
3. Evaluate online detection efficacy
4. Develop machine teammate's decision-making algorithm for IAD-mitigation

Key Scientific Contributions:

- **Confirm neurocorrelates (EEG) associated with biases**
- **ML method to estimate biases**
- **Provide method for targeted bias mitigation**

DoD Benefits:

- Operators become aware of their biases
- Operators make better decisions in human-machine-team (HMT) environments

List of Project Goals

1. Select/Modify an analyst task environment for collecting behavioral and neurophysiological data
2. Conduct an experiment, collect data and label activities as biased or unbiased
3. Develop & evaluate a machine learning model to detect/estimate level of bias
4. Evaluate ML model performance in online setting
5. Select one or more bias mitigation techniques which can be applied in real time
6. Conduct a new experiment where mitigation is applied appropriately when bias is detected
7. Evaluate HMT system performance

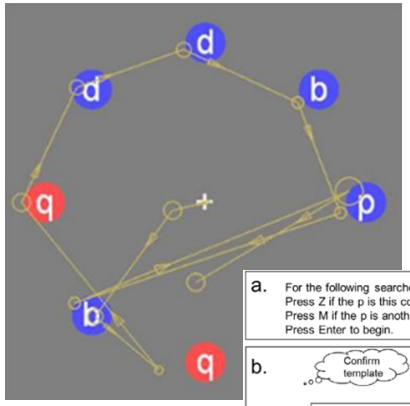
Progress Towards Goals (or New Goals)

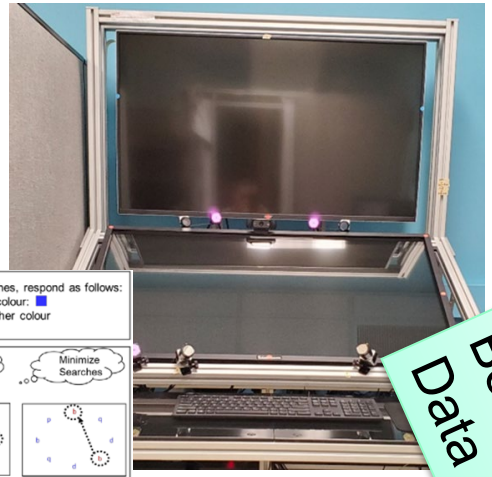
1. Select/Modify an analyst task environment for collecting behavioral and neurophysiological data
2. Conduct an experiment, collect data and label activities as biased or unbiased
3. Develop & evaluate a machine learning model to detect/estimate level of bias
4. Evaluate ML model in online setting performance
5. Select one or more bias mitigation techniques which can be applied in real time
6. Conduct a new experiment where mitigation is applied appropriately when bias is detected*
7. Evaluate HMT system performance

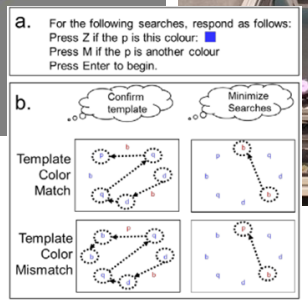
Motivation / Goals (2019-2020)

- Human information acquisition is subject to biases – especially in high-stress/fast decisionmaking environment
  - Pilots
  - Intel Analysts
  - Cyber Operators
- Research suggests Visual search is biased due to templates held in working memory
  - Impact: inefficient and less accurate visual search
- Objective: Aid operator's visual search through **<u>detection</u>** and **<u>mitigation</u>** of inefficient search

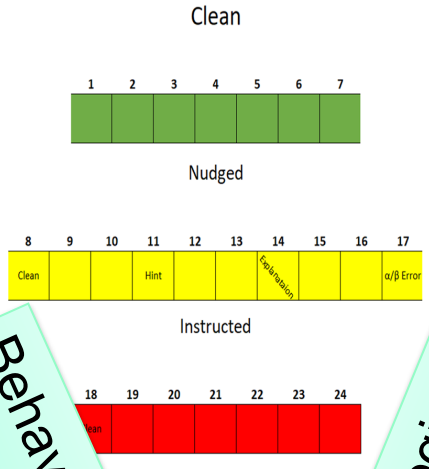
HMT
Bias Mitigation
System Overview

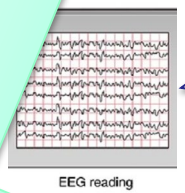






Decision-making environment





Behavior Data

Physiological Data

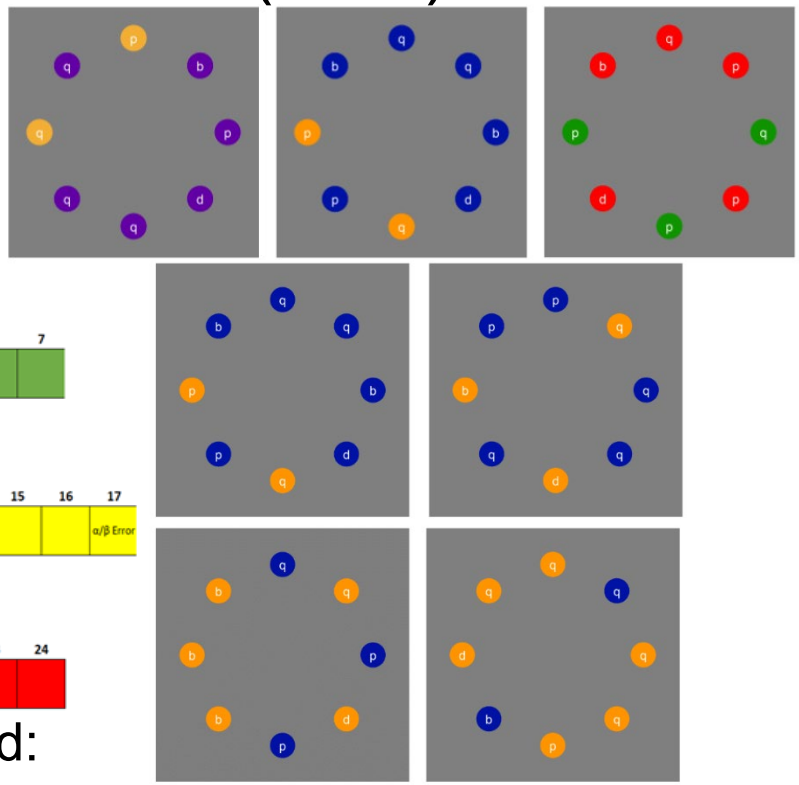
Bias Mitigation

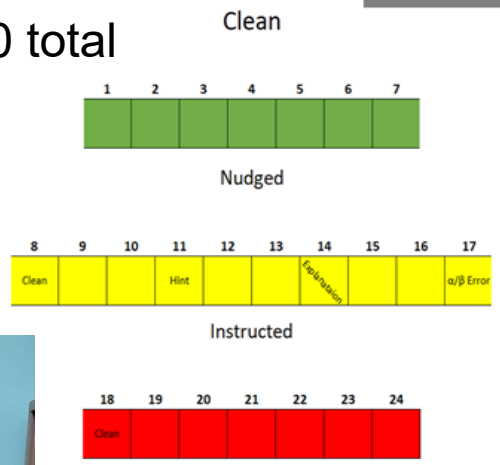
Machine Teammate

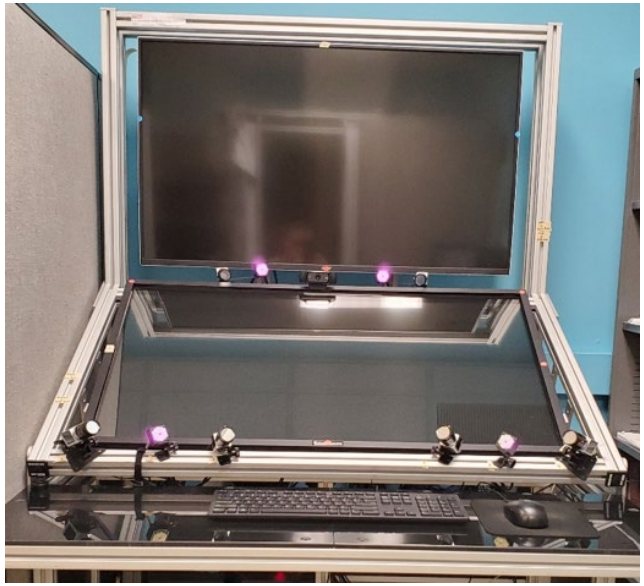
- Machine Learning Model
- Context Interpreter
- Interruption Decision

Methodology – Efficient Search Experiment (ESE)

- Adaptation of Rajsic's experiment
- EEG, ECG, EOG, GSR, and gaze tracking data all collected
- 24 blocks of 20 trials = 480 total
  - 20 training trials
- 16 Participants







Stimuli varied:

- Color combinations: blue/orange, purple/yellow, green/red
- Proportions of colors matching: 6, 5, 3, or 2 circles matching target color
- Target letter: "p", "q", "b", or "d"
- Target color: first color of above pairs

Machine Learning (ML) Models

Within-participant & Cross-Participant Models

- Raw Time Series Signal
  - Long Short-Term Memory (LSTM)
  - Temporal Convolutional Network (TCN)
- Spectral Features
  - Random Forest Classifier (RFC)
  - Linear Discriminate Analysis (LDA)
  - Artificial Neural Network (ANN)
- Hyperparameters included:
  - Layers
  - Hidden Units
  - Learning Rate
  - # Filters, Kernel Widths, Dilations, Stacks

Results

Detection

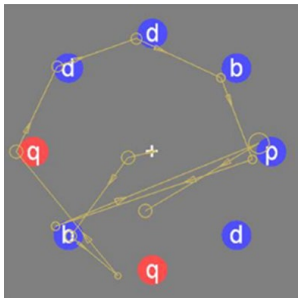
Table 1: Mean balanced accuracy scores of the within-participant models.

Within-Participant Dataset	Mean Balanced Accuracy % (# of participants w/ statistically significant accuracies)				
	LDA	RFC	ANN	LSTM	TCN
Nudge	49.2 (1)	50.5 (1)	51.7 (3)	51.4 (2)	50.7 (1)
Clean-Unbalanced	54.6 (5)	50.7 (1)	53.9 (4)	53.1 (5)	49.4 (1)
Clean-Balanced	59.2 (5)	58.2 (9)	52.5 (3)	53.7 (6)	50.8 (2)
Combined	58.1 (11)	56.2 (9)	53.2 (6)	55.3 (6)	49.7 (2)

Table 2: Mean balanced accuracy scores of the cross-participant models.

Cross-Participant Dataset	Mean Balanced Accuracy % (**bold underline** indicates statistically significant)		
	LDA	RFC	ANN
Nudge	50.2	53.2	49.9
Clean-Unbalanced	50.0	**57.3**	50.0
Clean-Balanced	**58.5**	**59.0**	50.0
Combined	51.0	54.0	50.1

Mitigation



- Efficient search found faster and more accurate compared to inefficient
  - Faster (sec): 1.99 ± 0.37 vs. 2.29 ± 0.50 ($p < 0.0001$)
  - More Accurate: 96.33% ± 2.16% vs. 93.92% ± 2.57% ($p < 0.0001$)
- Searches in first 8 blocks
  - 19.14% were efficient
  - 73.68% were inefficient
  - 7.18% were circular
- *Nudge* and *Hint* had greatest significance
  - Log worth of 10.67 and 8.5 (respectively)
- In last 7 blocks
  - Efficient increased by 32.27% to 51.41%
  - Inefficient decreased by 26.15% to 47.53%
  - Circular decreased by 6.12% to 1.06%

List of Publications, Awards, Honors, etc. Attributed to the Grant

- Conference Paper: “Detection and Mitigation of Inefficient Visual Searching” (Gallaher, Kamrud, Borghetti), HFES 2020, 5 Oct 2020
  Winner of Best Paper Award for Augmented Cognition Technical Group
- MS Thesis (2020): Lt Joshua Gallaher – “Automated Detection and Mitigation of Inefficient Visual Searching Using Electroencephalography and Machine Learning”
  https://scholar.afit.edu/etd/3160/

- Conference Paper: “Confirmation Bias Estimation from Electroencephalography with Machine Learning” (Villarreal, Kamrud, Borghetti), HFES 2019
- MS Thesis (2019): Capt Micah Villarreal – “Confirmation Bias Estimation from Electroencephalography With Machine Learning”
  https://scholar.afit.edu/etd/2290/

Backup Slides

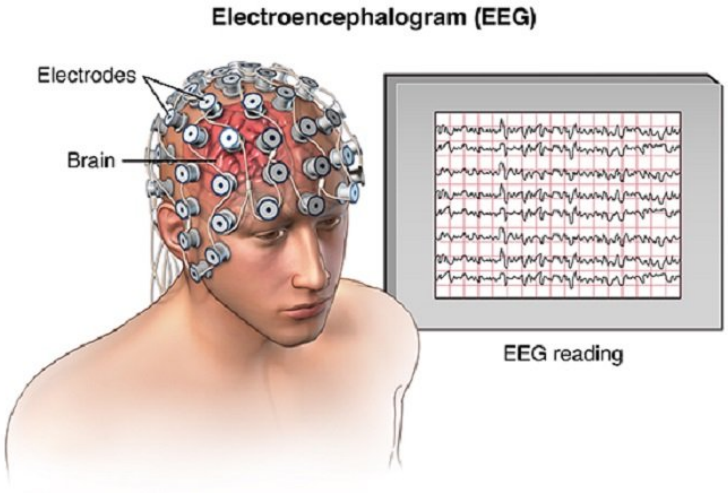
- Backup slides for 2019-2020 Main Study
- Backup slides for 2018-2019 Pilot Study

Backup Slides
Main Study 2019-2020

Research Objectives

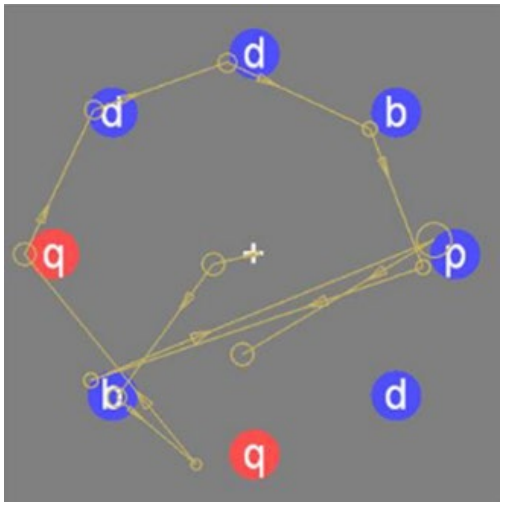
Detection

1. Can physiological signals such as Electroencephalography (EEG), Electrooculography (EOG), and Electrocardiography (ECG) be associated with an efficient visual search?



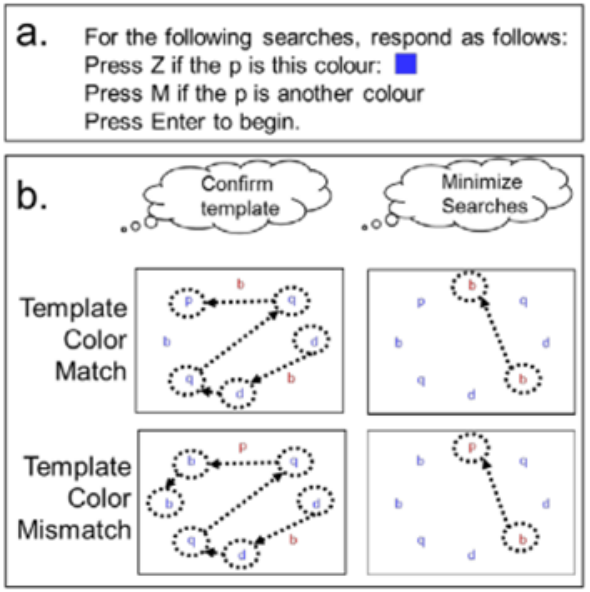
Mitigation

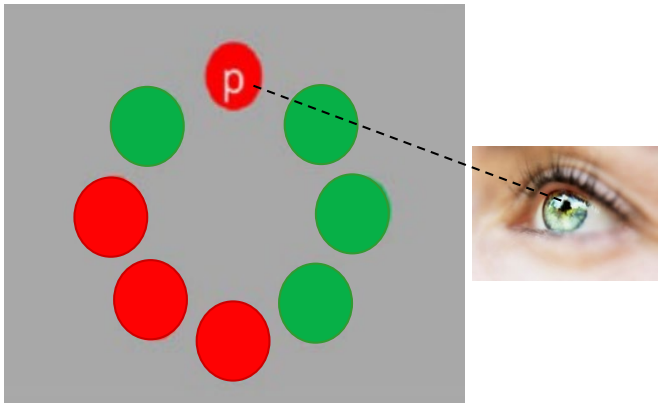
1. What visual search patterns do participants naturally use during a visual search task?
2. For a participant who is performing an inefficient search, can mitigation techniques change the participant's search patterns to an efficient search pattern that will persist for the remainder of the search tasks?



Background – Visual Search

- Previous research by Rajsic et al. (2015) found that humans unconsciously prefer confirmatory search over more strategic/efficient methods.
- Confirmatory search can be thought of as visual search that is guided by templates held in visual working memory.
- Rajsic et al. (2017) tested whether a high cognitive cost was causing participants to use confirmatory search instead. A nudge mitigation technique was utilized
- Walenchok (2018) found that people seek what is mentally salient by default.



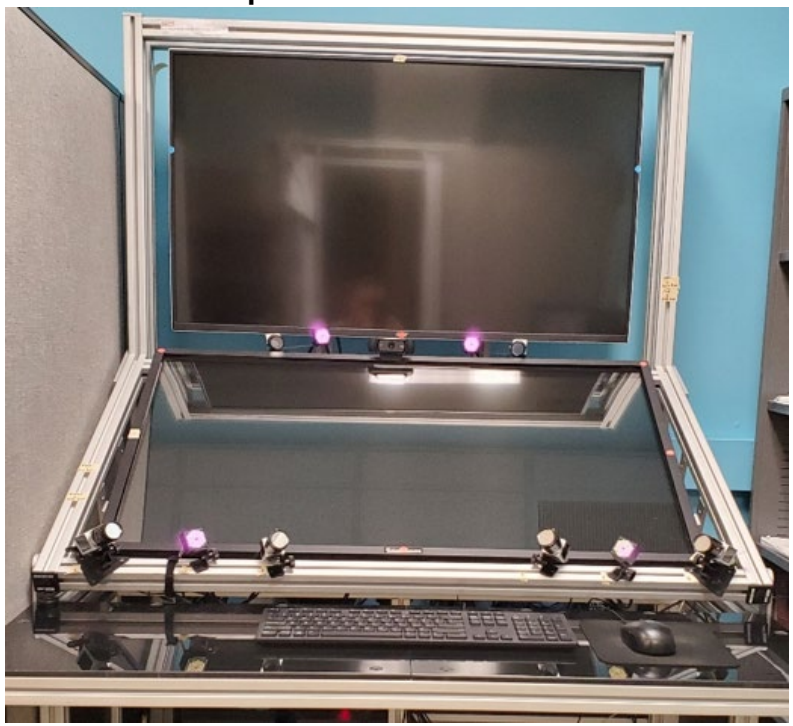


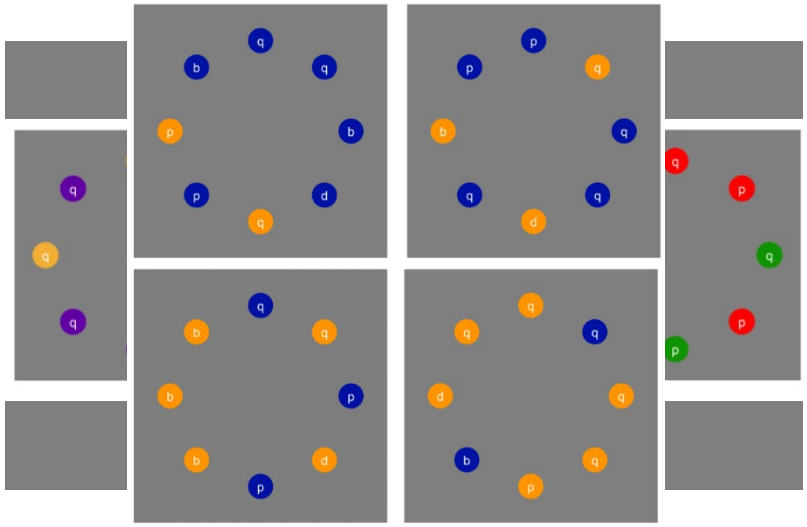
Issues

- Used search times to develop conclusions.
- Assumed search patterns not confirmed through gaze tracking.

Methodology – Efficient Search Experiment (ESE)

- Adaptation of Rajsic's experiment
- EEG, ECG, EOG, GSR, and gaze tracking data all collected
- 24 blocks of 20 trials = 480 total
  - 20 training trials
- 16 Participants



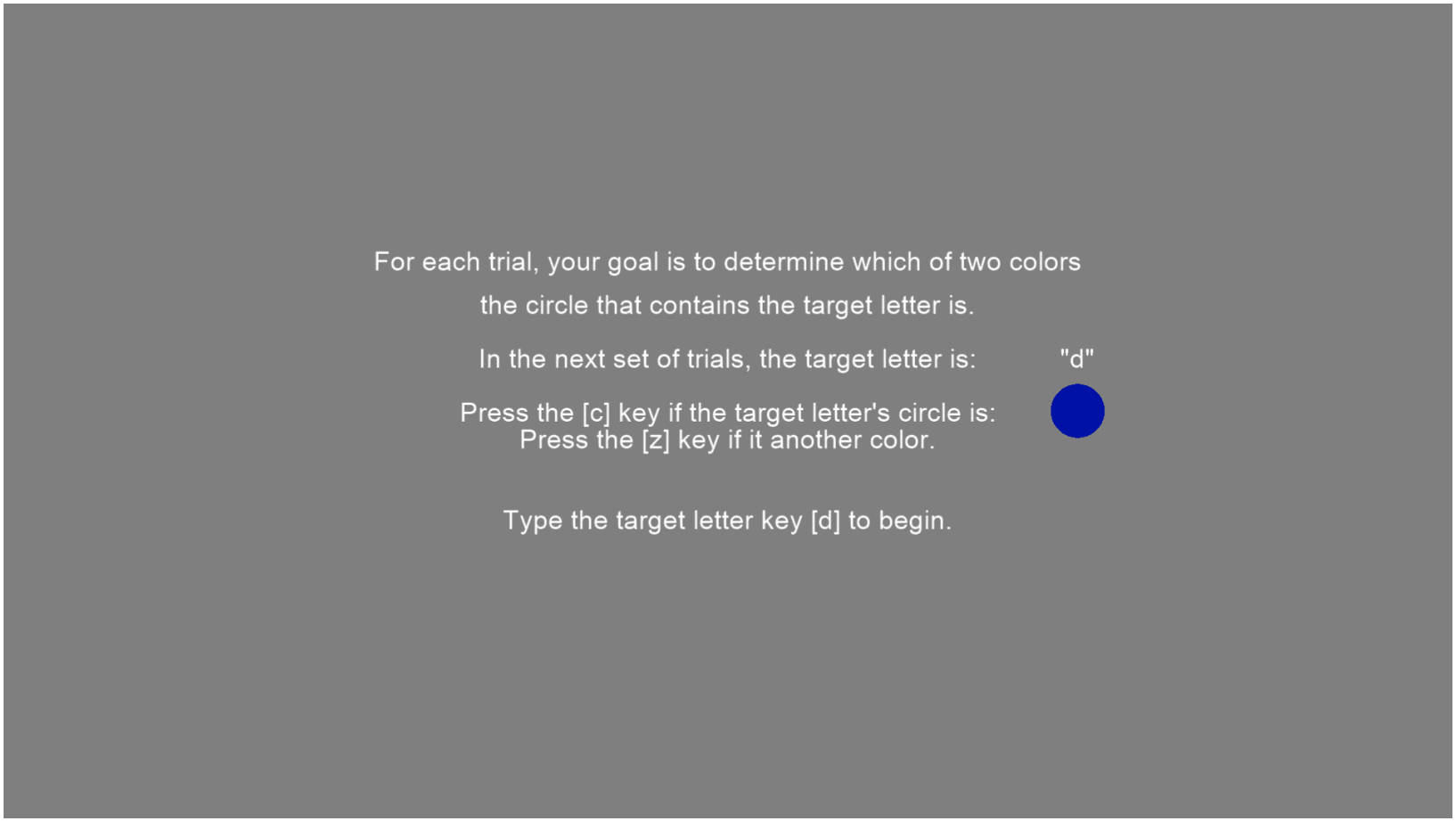


Stimuli varied:

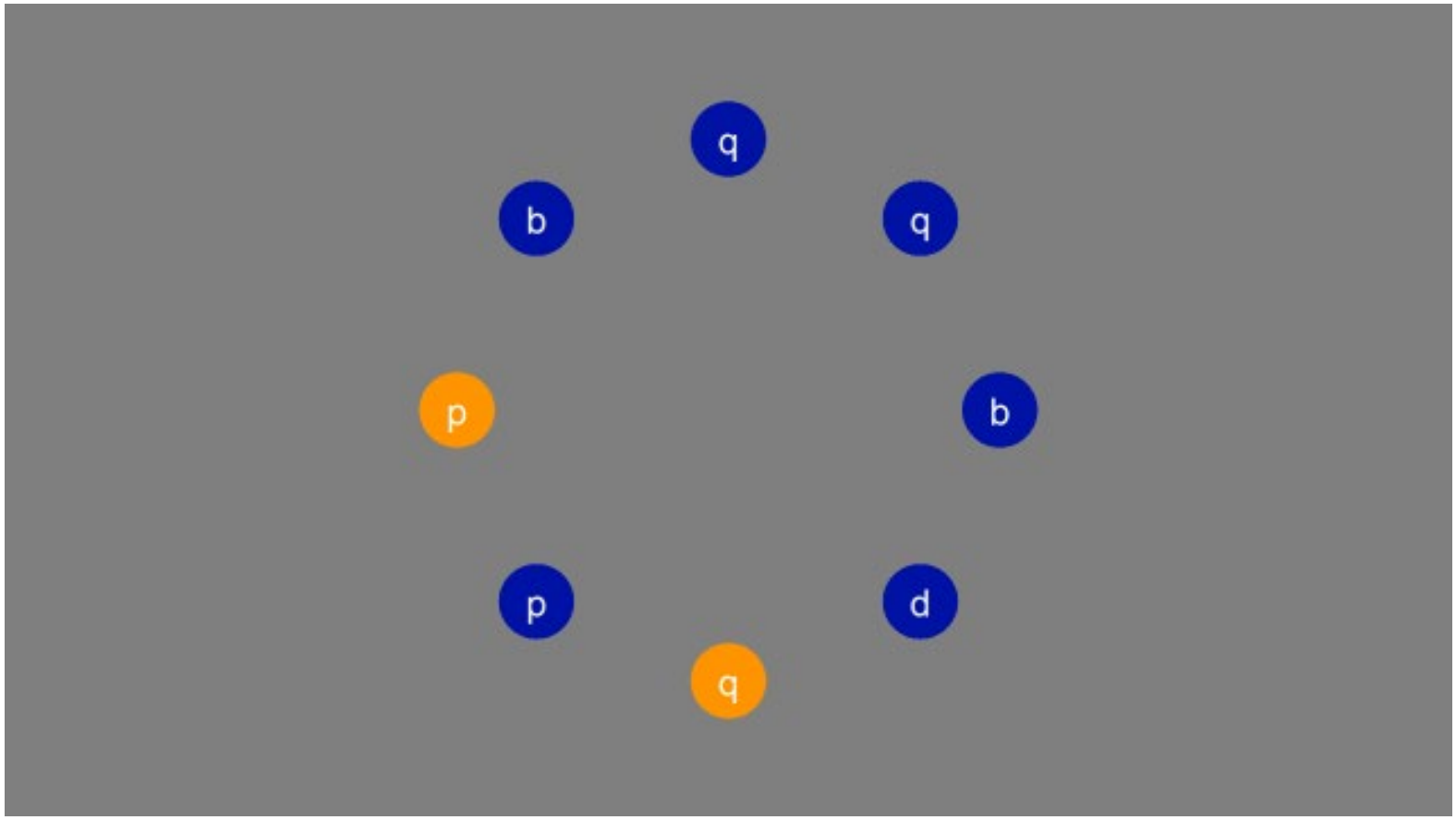
- Color combinations: blue/orange, purple/yellow, green/red
- Proportions of colors matching: 6, 5, 3, or 2 circles matching target color
- Target letter: "p", "q", "b", or "d"
- Target color: first color of above pairs

15

ESE –
Block Instructions



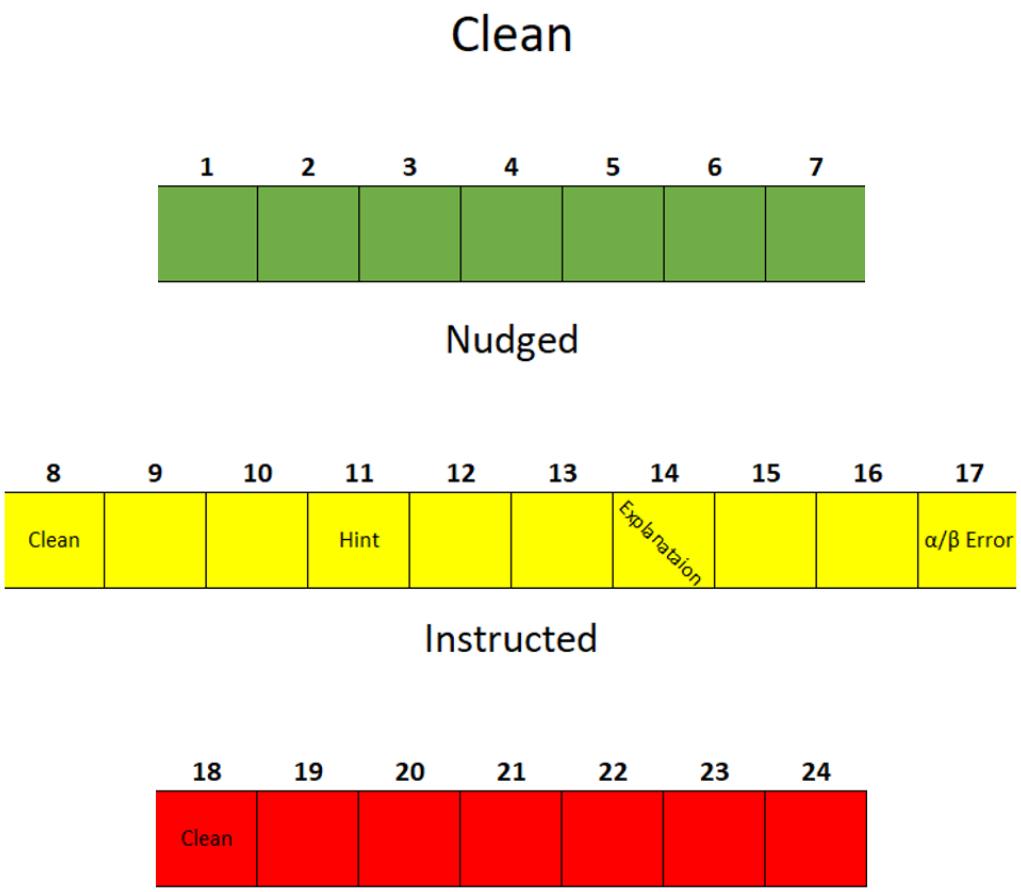
ESE – Trial Example



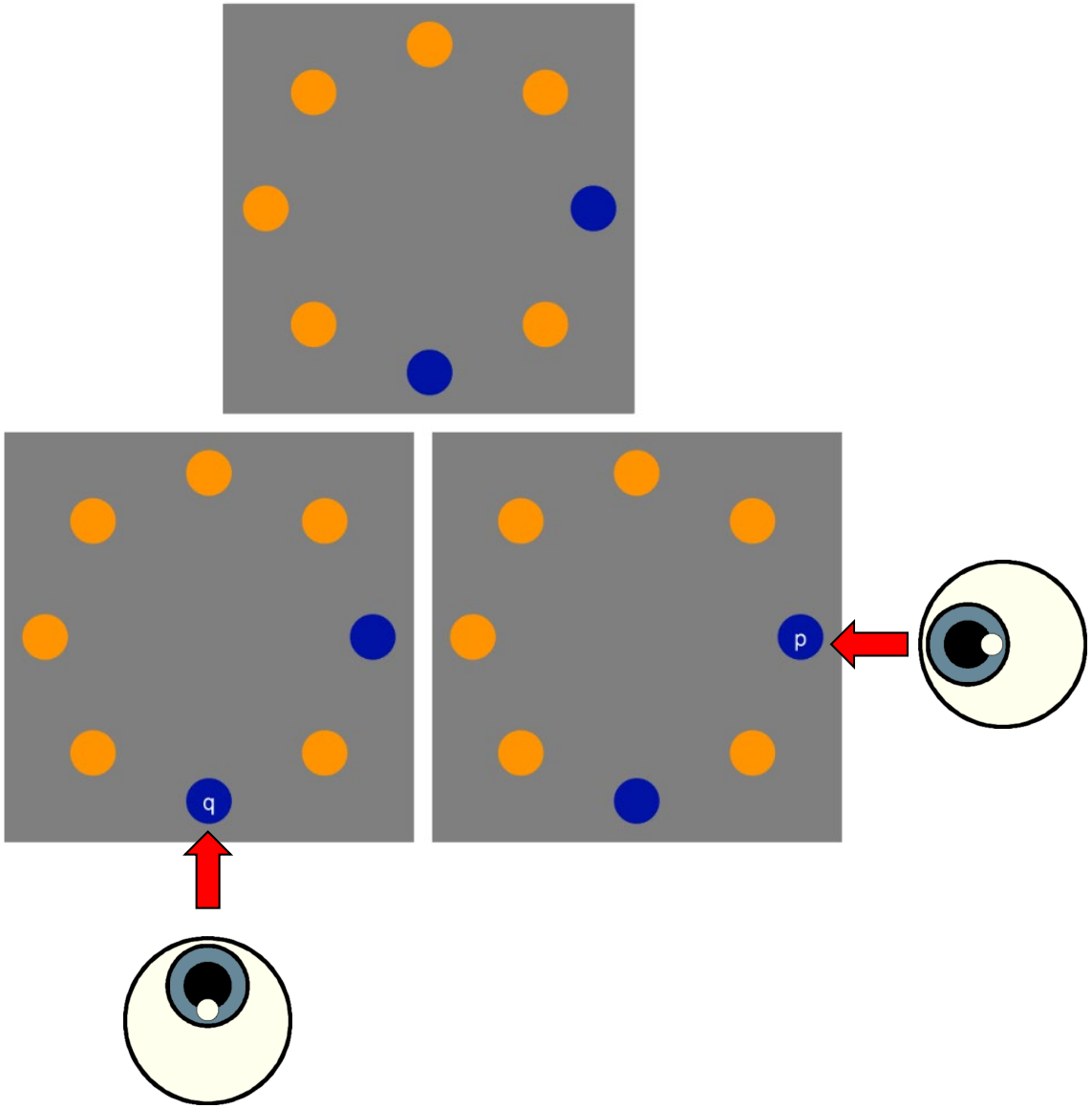
ESE – Mitigations & Block Design

Mitigations

- *Nudge*
  - Adds cost to visual search by covering the letters
  - Only activated if 50% or more of previous block's trials were inefficient
- *Hint*
  - Hint to participant about performing efficient searches
- *Explanation*
  - Presents explanation to participant on why nudge is activated
- *Instruction*
  - Instructs participant to perform an efficient search
  - Provides explicit instruction on how to perform an efficient search



Mitigation Technique – Nudge

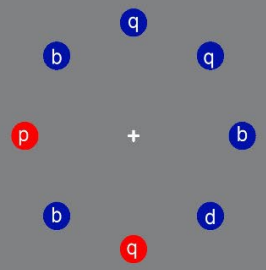


Mitigation Technique – Hint

Remember - there is only one copy of the target letter and there are only ever two colors.
You only need to look at one set of colored circles
to determine which color the target letter's circle is.

It is more efficient to look at the circles with the color
that appears less often on the screen.

For example, suppose the target letter is "p"



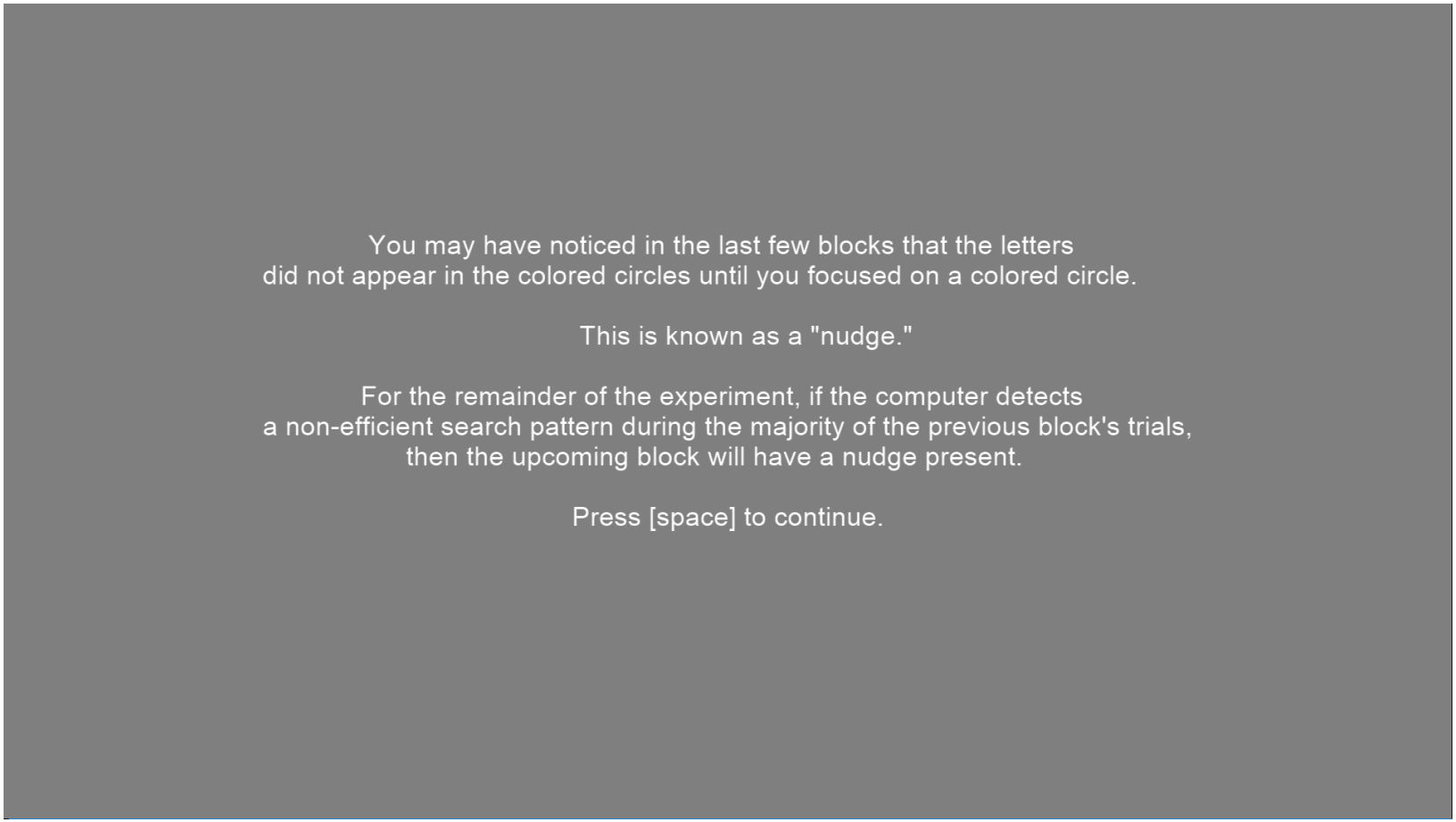
Searching the red circles allows us to search only 2 circles instead of the 6 blue circles.

If "p" is present in the red circles, then we know that it is red.

If "p" is not present in the red circles, then we know that it must be blue!

Please press [space] to continue.

Mitigation Technique – Explanation



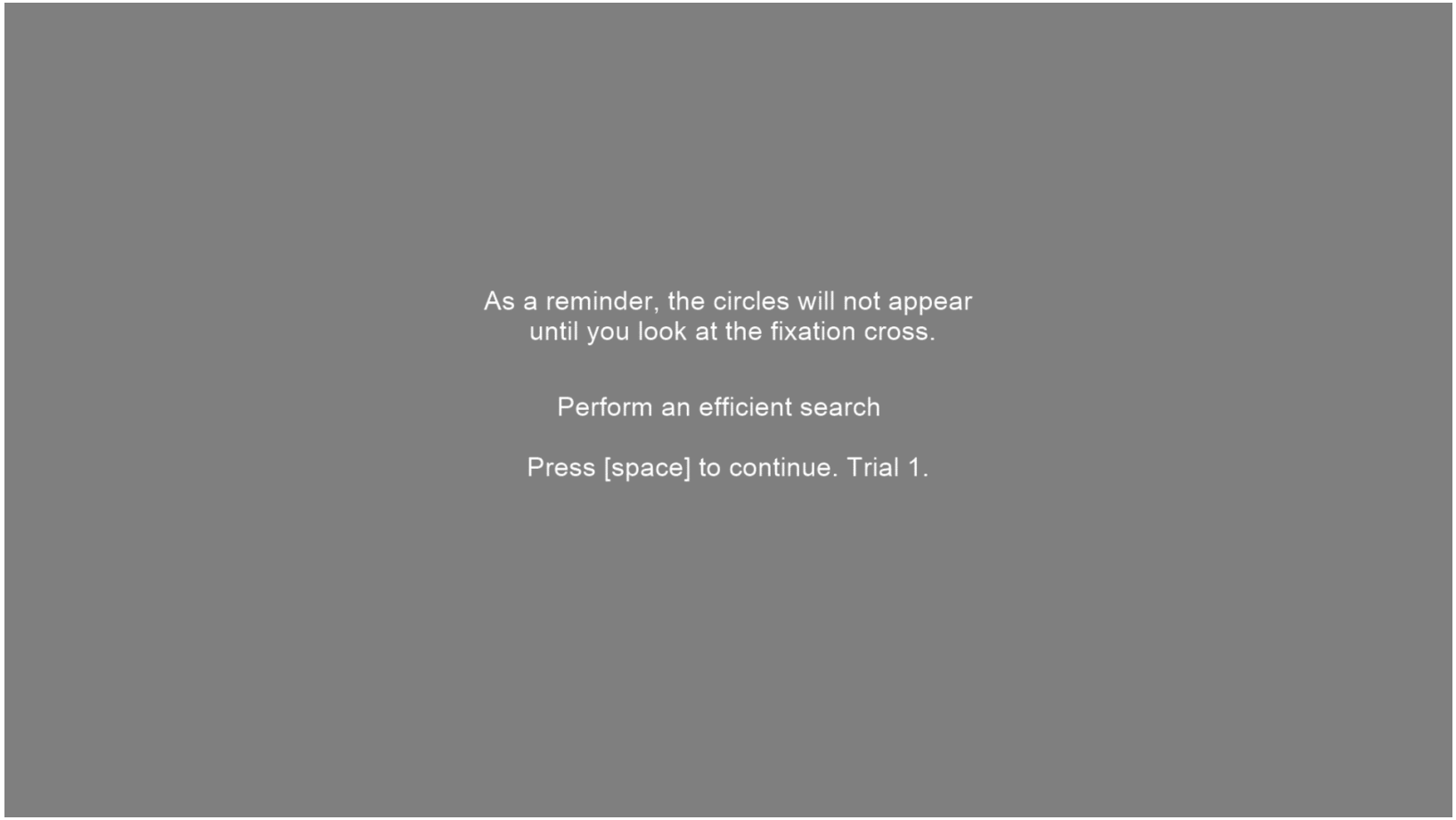
Mitigation Technique – Instructions

For the remainder of the experiment, perform an efficient search.

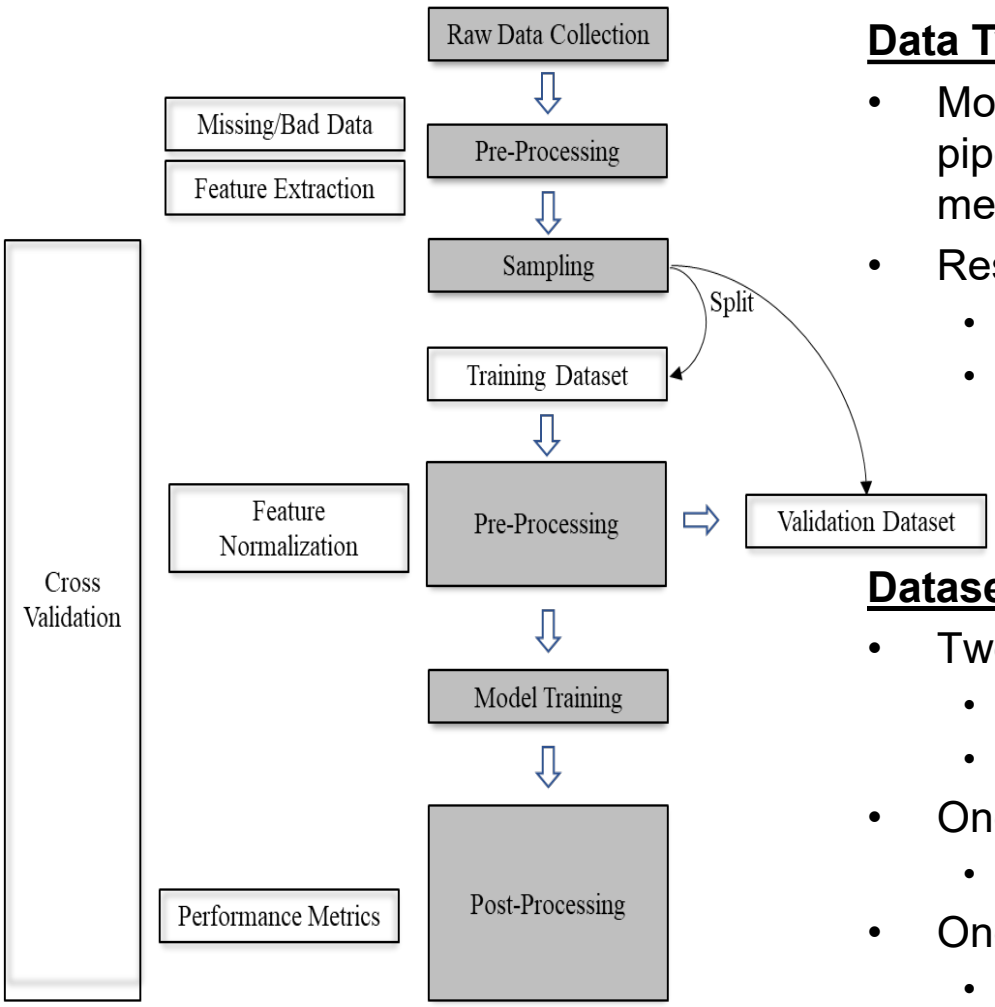
The following screen will instruct you on how to perform an efficient search.

Press [space] to continue.

Mitigation Technique – Instructions



Machine Learning (ML) Pipeline



Data Type – EEG Preprocessing Pipeline

- Modified version of Makoto's preprocessing pipeline, the PREP pipeline, and Mike X Cohen's method for spectral feature extraction
- Results in two data types
  - Raw time series
  - Spectral features
    - Mean power of five traditional frequency bands for 64 channels (5*64=320)

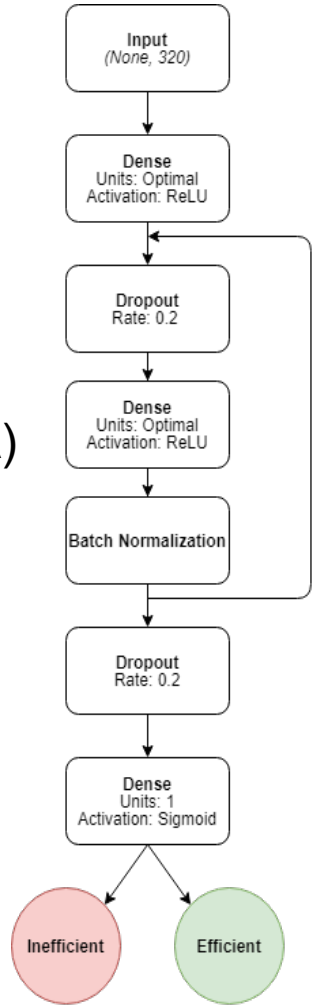
Datasets

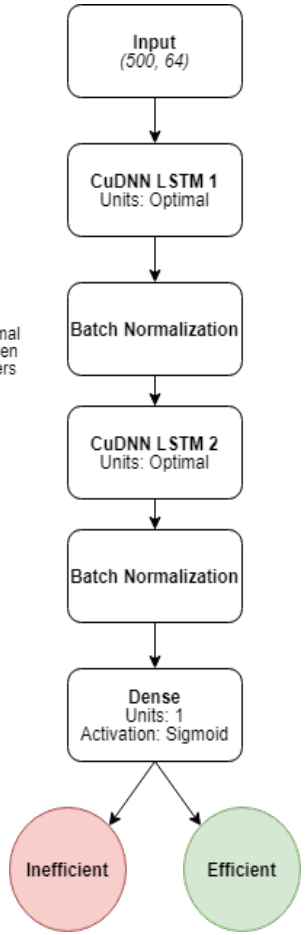
- Two datasets consisted of only non-nudge trials
  - *Clean-Balanced & Clean-Unbalanced*
  - Balance needed (only 7%-24% of trials were efficient)
- One dataset which consisted of only nudge trials
  - *Nudge*
- One dataset which consisted of all trials
  - *Combined*
- Eight datasets total (4*2 data types = 8)

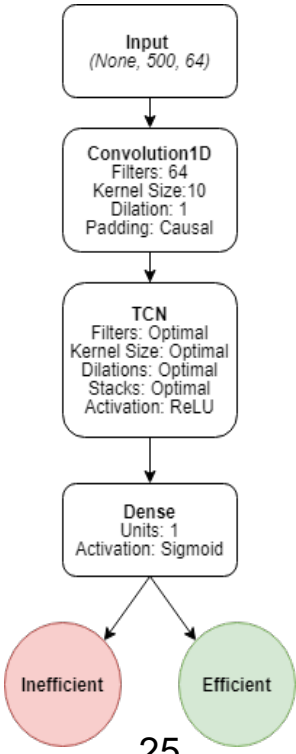
Machine Learning (ML) Models

Within-participant & Cross-Participant Models

- Raw Time Series Signal
  - Long Short-Term Memory (LSTM)
  - Temporal Convolutional Network (TCN)
- Spectral Features
  - Random Forest Classifier (RFC)
  - Linear Discriminate Analysis (LDA)
  - Artificial Neural Network (ANN)
- Hyperparameters included:
  - Layers
  - Hidden Units
  - Learning Rate
  - # Filters, Kernel Widths, Dilations, Stacks







Results

Detection

Table 1: Mean balanced accuracy scores of the within-participant models.

Within-Participant Dataset	Mean Balanced Accuracy % (# of participants w/ statistically significant accuracies)				
	LDA	RFC	ANN	LSTM	TCN
Nudge	49.2 (1)	50.5 (1)	51.7 (3)	51.4 (2)	50.7 (1)
Clean-Unbalanced	54.6 (5)	50.7 (1)	53.9 (4)	53.1 (5)	49.4 (1)
Clean-Balanced	59.2 (5)	58.2 (9)	52.5 (3)	53.7 (6)	50.8 (2)
Combined	58.1 (11)	56.2 (9)	53.2 (6)	55.3 (6)	49.7 (2)

Table 2: Mean balanced accuracy scores of the cross-participant models.

Cross-Participant Dataset	Mean Balanced Accuracy % (**bold underline** indicates statistically significant)		
	LDA	RFC	ANN
Nudge	50.2	53.2	49.9
Clean-Unbalanced	50.0	**57.3**	50.0
Clean-Balanced	**58.5**	**59.0**	50.0
Combined	51.0	54.0	50.1

Mitigation

- Efficient search found faster and more accurate compared to inefficient
  - 1.99 ± 0.37 vs. 2.29 ± 0.50 ($p < 0.0001$)
  - 96.33% ± 2.16% vs. 93.92% ± 2.57% ($p < 0.0001$)
- Searches in first 8 blocks
  - 73.68% were inefficient
  - 19.14% were efficient
  - 7.18% were circular
- *Nudge* and *Hint* had greatest significance
  - Log worth of 10.67 and 8.5 (respectively)
- In last 7 blocks
  - Efficient increased by 32.27% to 51.41%
  - Inefficient decreased by 26.15% to 47.53%
  - Circular decreased by 6.12% to 1.06%

Discussion

Detection

- Certain within-participant models performed well, with *Clean-Balanced* resulting in the most significant within-participant models
- However, overall, models for each dataset did not perform statistically significantly better than chance
- *Overfitting* due to Curse of dimensionality – Not enough data for all features (5 bands*64 nodes = 320 features, but only 480 observations)

Mitigation

- Humans naturally use an inefficient search pattern
- Efficient searches are faster and more accurate than inefficient searches
- Adding an additional cost to search (i.e. *nudge*) mitigated inefficient search patterns

Conclusions and Future Work

- More models to explore: Gated Recurrent Units (GRUs), TCNs
- Dimensionality reduction and Feature Selection
  - Frontal lobe - Alpha band focus
- Future experiment to illicit specific visual search patterns explicitly through instruction
  - Provides better data labelling
  - No nudge to add confound to visual search pattern
  - No need for multiple datasets
  - Allows for balanced dataset

Collaborators & Contact

- Funding provided by Air Force Research Labs (AFRL)/ Air Force Office of Scientific Research (AFOSR)

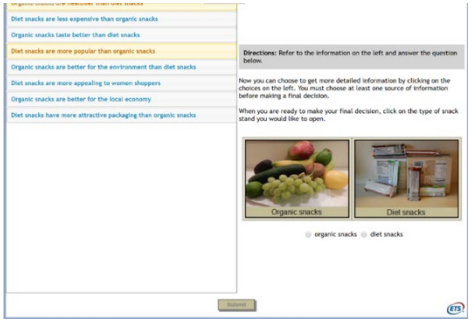
- Author Contact Info:
  - alexander.kamrud.1@us.af.mil
  - brett.borghetti@afit.edu
  - joshua.gallaher.2@us.af.mil

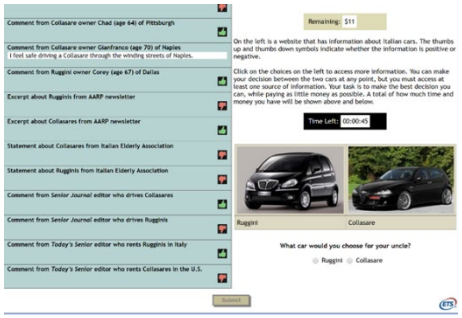
Questions?

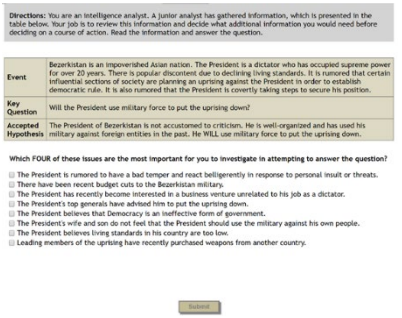
- Ashcraft, M., & Radvansky, G. (2013). *Cognition (6th Edition)*.
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*.
- Cohen, M. X. (2019). *Analyzing Neural Time Series Data*. *Analyzing Neural Time Series Data*. https://doi.org/10.7551/mitpress/9609.001.0001
- Hastie, T., Tibshirani, R., James, G., & Witten, D. (2006). *An Introduction to Statistical Learning, Springer Texts*. *Springer Texts* (Vol. 102). https://doi.org/10.1016/j.peva.2007.06.006
- Kahneman, D., & Klein, G. (2009). Conditions for Intuitive Expertise A Failure to Disagree. *Psycnet.Apa.Org*, *64*(6), 515–526. https://doi.org/10.1037/a0016755
- Kumar, S., Sharma, A., & Tsunoda, T. (2019). Brain wave classification using long short-term memory network based OPTICAL predictor. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-45605-1
- Ledoit, O., & Wolf, M. (2003). *Honey, I Shrunk the Sample Covariance Matrix*.
- Miyakoshi, M. (2017). Makoto's preprocessing pipeline - SCCN. *Swartz Center for Computational Neuroscience Wiki Site*, 1–39. Retrieved from https://sccn.ucsd.edu/
- wiki/Makoto's_preprocessing_pipeline
- National Research Council. (2015). *Measuring Human Capabilities: An Agenda for Basic Research on the Assessment of Individual and Group Performance Potential for Military Accession*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/19017
- Nickerson, R. S. (1998). *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*. *Review of General Psychology* (Vol. 2).
- Rajsic, J., Wilson, D. E., & Pratt, J. (2015). Confirmation bias in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(5). https://doi.org/10.1037/xhp0000090
- Smart Eye AB. (2018). SE PRO | Smart Eye. Retrieved from http://smarteye.se/research-instruments/se-pro/
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Backup Slides
Pilot Study 2018-2019

Phase 1 Pilot Study



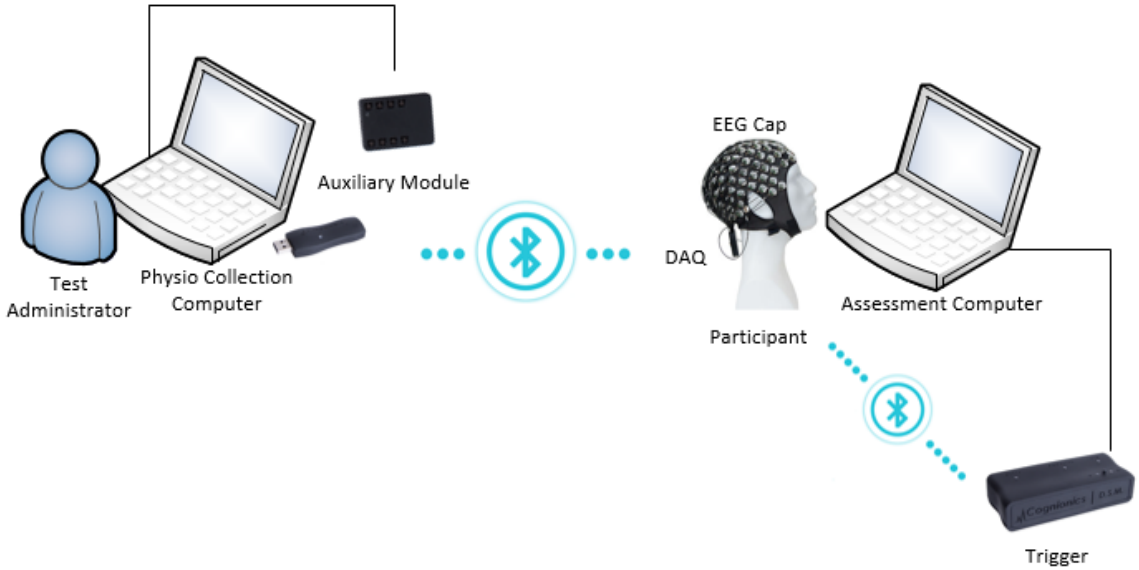




- MITRE Assessment of Biases in Cognition
  - Designed for cognitive bias detection
- Focus on tasks where confirmation bias was detected
  - 4 investigative task types; 14 tasks total
- Added EEG/EOG/ECG collection
  - Activation of right frontal cluster suggests bias

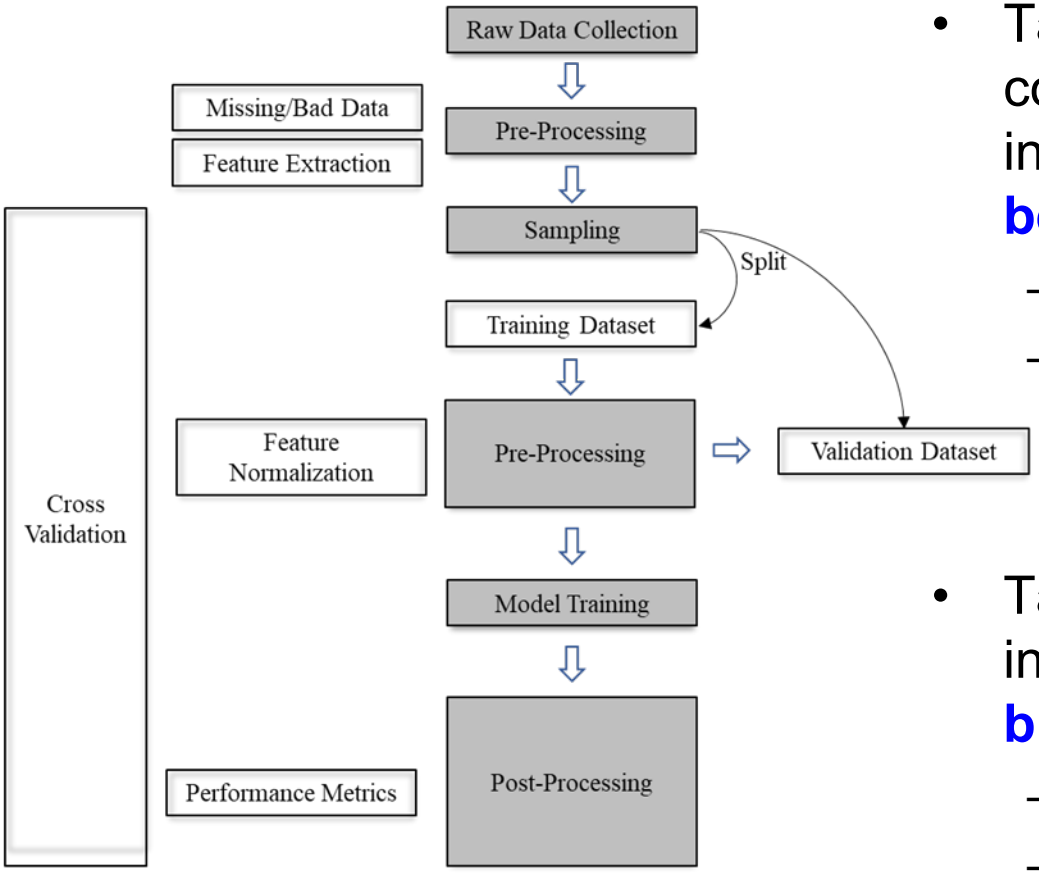
MITRE, “IARPA Sirius Program Assessment of Biases in Cognition (ABC),” 2015.

Pilot Study Data Collection



- 15 participants (AFIT students)
- Behavioral: Decisions and Info selections and timing
- Physiological: EEG (64 chan), EOG, ECG

Phase 2: Machine learning



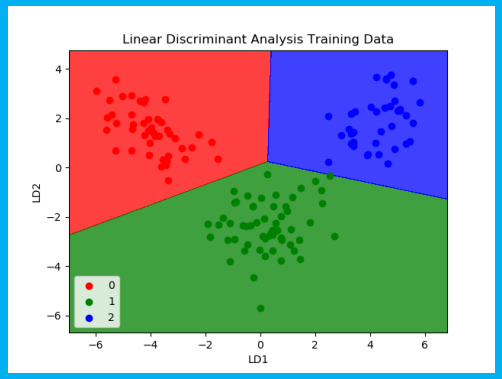
- Task A: Classify confirming/disconfirming information selection from **behavior**
  - Response time
  - Information Revisits

- Task B: Classify Bias; c/d information selection from **brainwaves**
  - Raw EEG
  - Spectral response

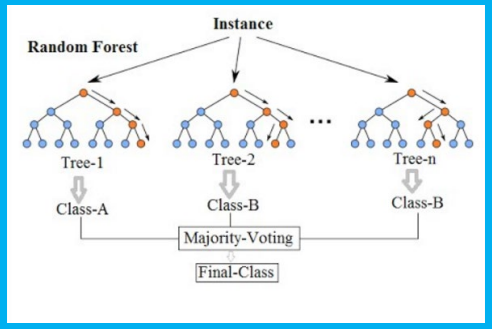
ML Models

SPECTRAL PROCESSED EEG

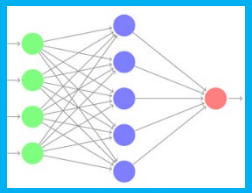
Linear Discriminant Analysis (LDA)

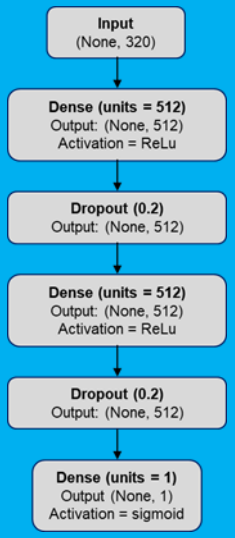


Random Forest

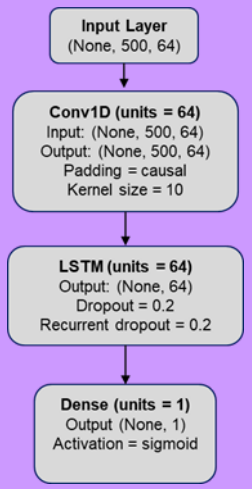


Fully connected ANN

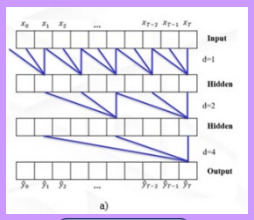


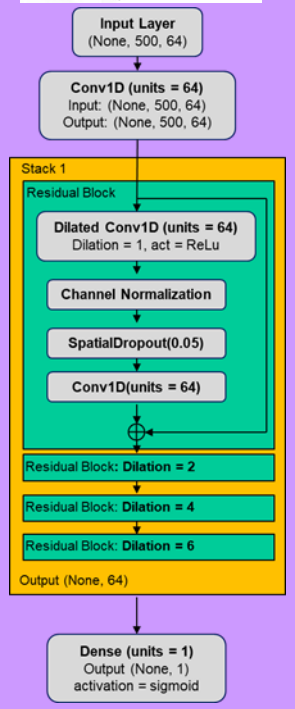


Long Short-Term Memory (LSTM)



Temporal Convolutional Network (TCN)





RAW EEG

ML Results
Task A

- Use Behavior to determine whether participant experienced confirmation bias during task (biased v. unbiased):
  - Response time - No significant difference
  - Information revisits – No significant difference

ML Results
Task B - Brainwaves

- ML Subtask B.1:
  - Use EEG to determine whether participant experienced confirmation bias *during decision task*
  - Not enough data for meaningful ML training (only 14 decisions per participant)
- ML Subtask B.2:
  - Use EEG to determine whether *info selected* was confirming or disconfirming
  - Balanced accuracy result slightly (but significantly) above chance for two of the participants
  - Not good enough to declare a successful finding

Pilot study challenges (1/2)

- MITRE ABC issues
  - Bias truth labels: Only some tasks had pre-checks for prior belief – difficult to determine bias on others → mislabeled data?
  - Behavior Labeling misalignment: Interface allows participants to open all information before reading any individual item → EEG signal not aligned to participant ingesting the information

Pilot study challenges (2/2)

- Experiment Design Issues
  - Small sample size: Long duration tasks don't allow many repetitions per unit time $\rightarrow$ ML hard to train
  - Imbalanced data: Large imbalance in information selection led to very few training observations for some conditions $\rightarrow$ ML hard to train
  - Response time decreased with participant experience
    - Learning Effect $\rightarrow$ Response time is unreliable predictor;
    - Task disengagement $\rightarrow$ EEG may not be useful for bias detection

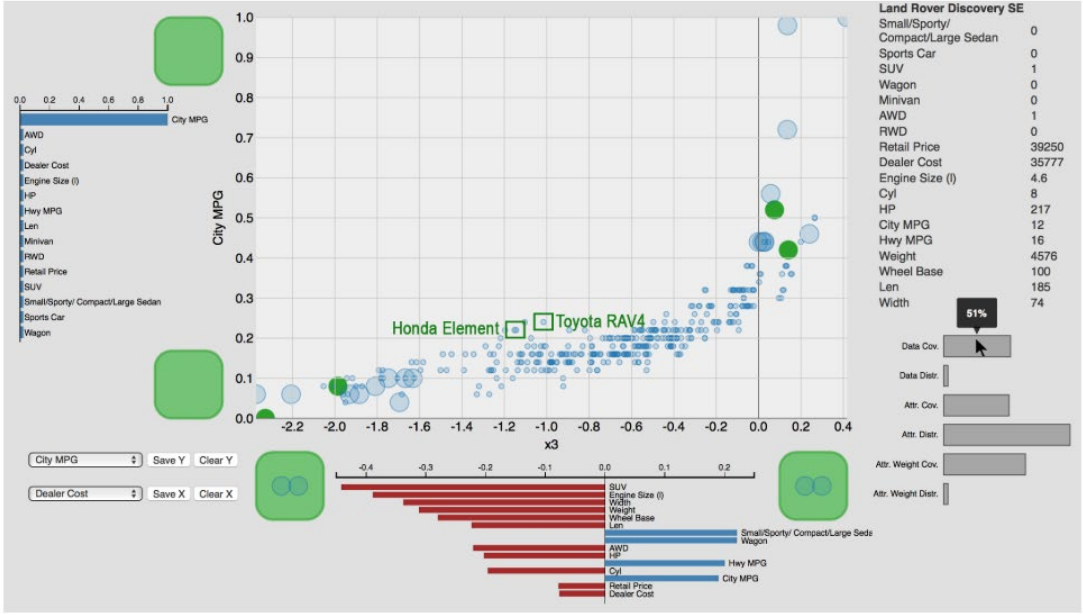
Physiological Data collection issues

- EEG Equipment anomalies discovered during experiment
  - Local system diagnostics & fault isolation procedures unsuccessful
  - Vendor (Cognionics) confirmed it was a hardware design problem and sent new (version 2) equipment
  - Quality of existing data in question

The way ahead – near term:

- New EEG sensor received and tested
  - no anomalies found so far
- Rethinking/Redesigning experiment
  - Considering visual search w/automation bias
    - More trials per unit time
    - Reduced interface confounds
    - Clearer recognition of bias v. unbiased
    - Easier manipulation and ability to induce bias
  - Adding GSR and Gaze tracking
- Collaboration opportunities…

Collaboration opportunities…



- e.g. InterAxis (Wall, Blaha, Franklin, Endert)
  - Detect / mitigate bias using behavioral measures
  - Future: Behavioral + Neurophysiological?