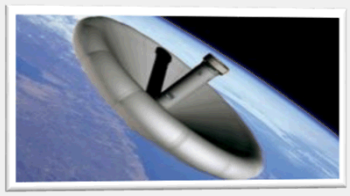
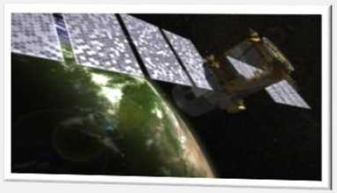
LANGLEY RESEARCH CENTER

Data Science

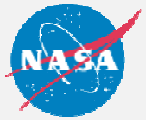



Machine Learning to Assess Pilots' Cognitive State

March 21, 2018

Tina Heinich
AST, Data Systems
Computer Engineer, OCIO Data Science Team



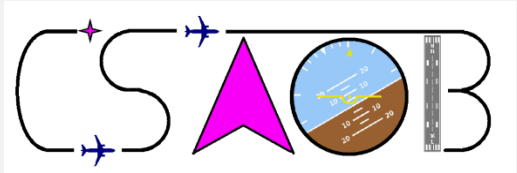
Summary

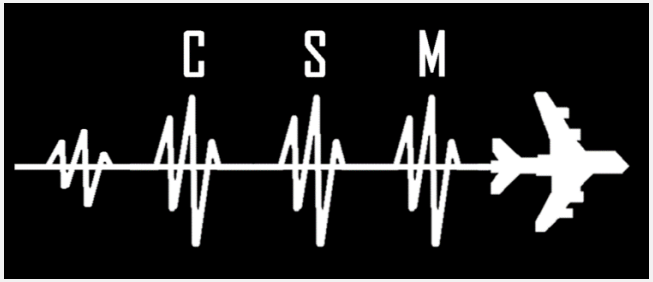
- Goal
- The Data
- Challenges
- Current Pipeline
- Future Work

The Team

Data Science

- Tina Heinich (NASA)
- Angela Harrivel (NASA)
- Chad Stephens (NASA)
- Robert Milletich (Booz Allen Hamilton)
- Kellie Kennedy (NASA)
- James Comstock (NASA)
- Alan Pope (NASA Distinguished Research Associate)
- Charles Liles (NASA)
- Mary Carolyn Last (Analytical Mechanics Assoc.)
- Nijo Abraham (NASA)
- Nick Napoli (UVA)
- And many more….





Goal of CSM

Data Science

- In 2001-2010, Commercial Aviation Safety Team (CAST) identified unsafe cognitive states as a key factor in almost all studied airplane accidents.
- Want to predict if a subject is currently in an unsafe cognitive state
- Create a machine learning model that can reliably predict the cognitive state a subject is in using various physiological sensors
- **Specific Goal: Using previously collected data in a non-flight scenario, can we predict cognitive states in an actual flight simulation?**

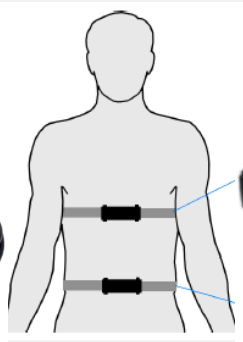
Cognitive States

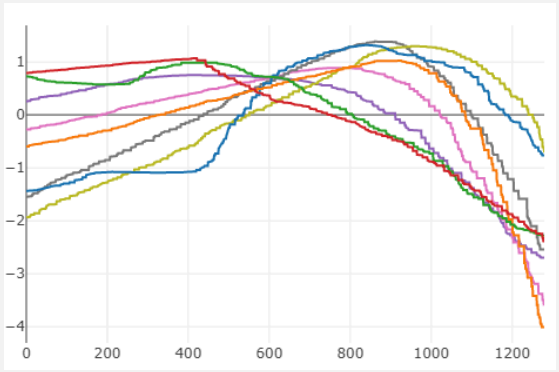
- Startle/Surprise (SS)
  - Includes the “onset” as well as the initial 13 seconds “recovery” afterwards
- Channelized Attention (CA)
  - Subject is focused on one particular task/input to the exclusion of all other tasks/inputs
- Low Workload (LW)
  - Demand requires minimal resources to complete
- “Other”
  - Not really a class, but whenever a pilot is NOT experiencing these states

Physiological Modalities

Data Science

- 20-channel Electroencephalography (EEG)
  - ABM B Alert X24
- Galvanic Skin Response (GSR)
  - NeXus-10
- Electrocardiogram (EKG)
  - NeXus-10
- Respiration (R)
  - NeXus-10

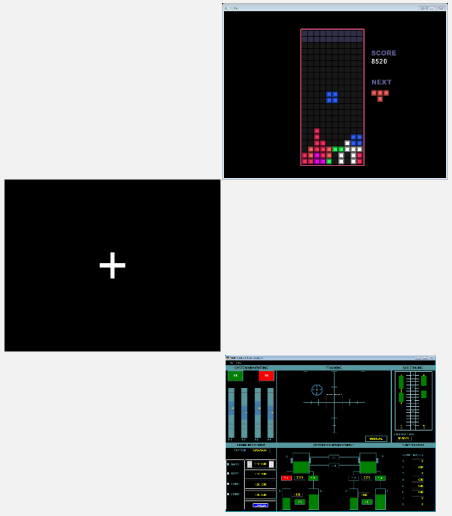




Tour of the Data

The Experiments

- Data was collected for 13 crews of pilot and copilot (26 total participants)
  - Because of various data collection issues, we concentrate primarily on 5 full crews (10 pilots)
- Benchmark Tasks
  - 6 minute tasks that induce various cognitive states
  - Tasks were done on two different days
    - Known as “Day 1 Benchmark” and “Day 2 Benchmark”
- Line-Oriented Flight Training (LOFT) Data
  - Flight scenario with full flight



Tour of the Data

The Features

- EEG
  - Focus on MVP Channels
  - Summary Statistics :
    - Quantiles: Q005, Q995, Q75 and Inter-Quantile Range
    - Skew and Kurtosis
    - Variance
  - 4 Power Band Features
    - Alpha Band Mean (4-8 Hz)
    - Beta Band Mean (8-13 Hz)
    - Theta Band Mean (13-22 Hz)
  - Engagement Index and Task Load Index
- ECG
  - Heart Rate Variability Mean and Variance
  - Summary Statistics
    - Quantiles: q005 and IQR
    - Skew and Kurtosis
    - Variance
- GSR
  - Avg Slope and Drop Score
    - “drop score” essentially counts how often the slope drops under a given threshold
  - Summary Statistics
    - Quantiles: q975
    - Variance
    - Skew
- Respiration
  - Respiration Rate
  - Summary Statistics
    - Quantiles q75, q975, and q025
    - Skew and Kurtosis
    - Variance

Data was divided into sliding 5 second windows with 1 second stride.
Within each window, features for each modality were calculated.

Tour of the Data

Various Notes

- There is a class imbalance between SS, CA, and LW
- Class labels are largely based on state induction (not actual cognitive state)
- We can't rely on the sensor data always being there
  - Sometimes a sensor falls off, or gets skewed a little
- There is often variance between the different datasets

Future Work Note: Currently have experts and flight instructor reviews of the data recordings so to improve the class labels/find other cognitive states in the data.

	Bench1	LOFT
CA Windows	2852	2932
LW Windows	2841	6669
SS Windows	241	282
Other Windows	3504	31173

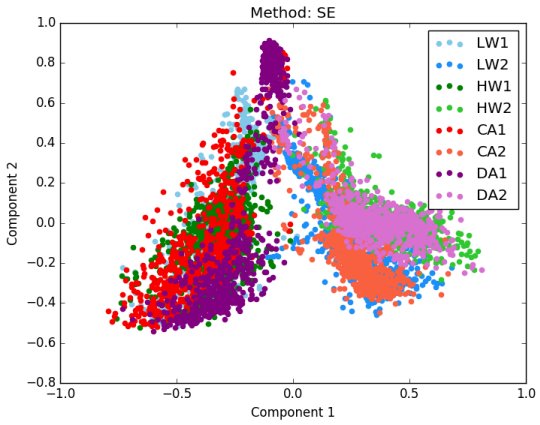
Variance Challenge

Differences between Days

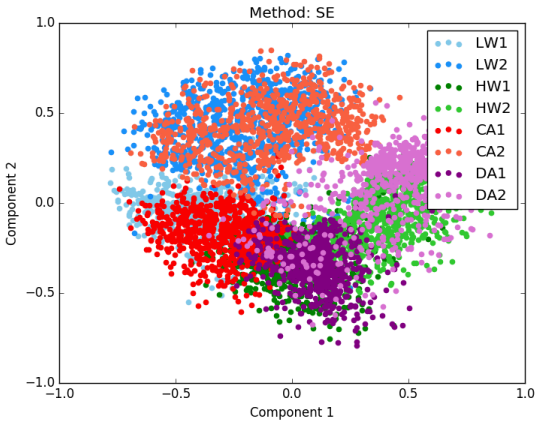
Data Science

- Found that a model built on Day 1 Bench won't necessarily work on Day 2 Bench

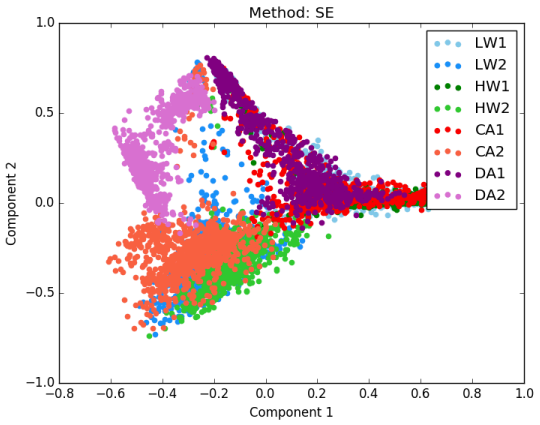
Crew 7



Crew 11



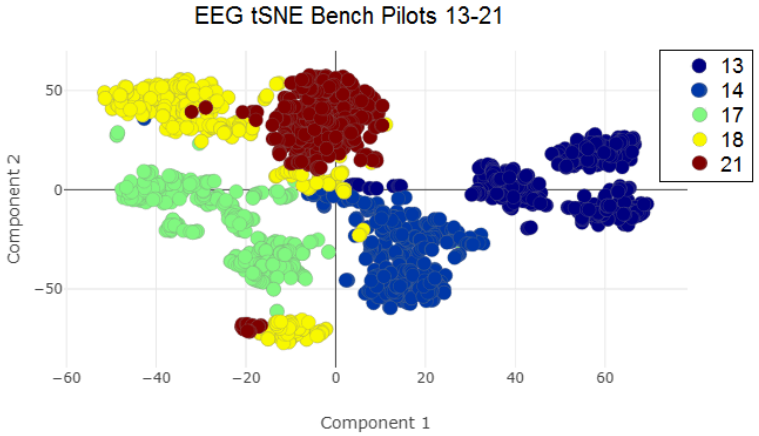
Crew 13

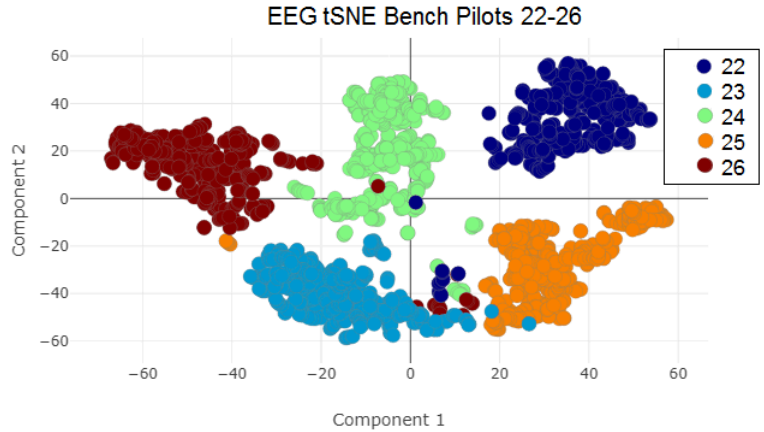


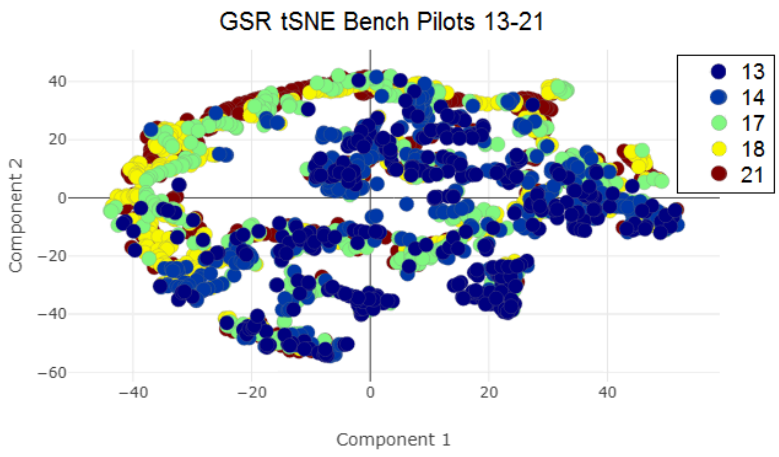
Variance Challenge

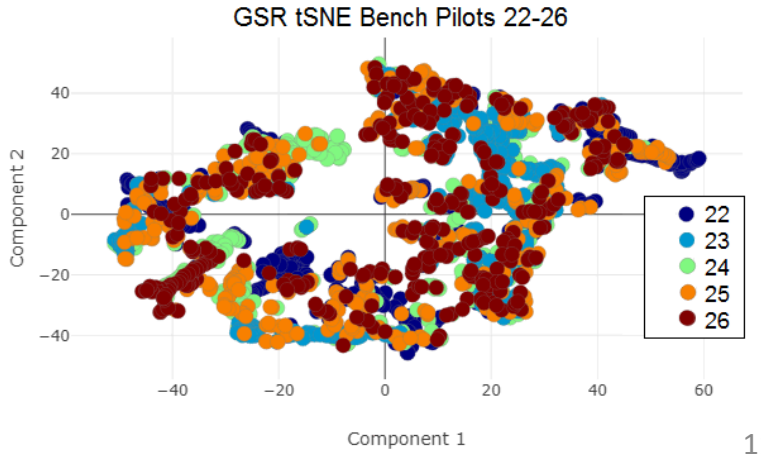
Differences between Pilots

- Hard to predict from pilot to pilot; variance is more pronounced in certain modalities

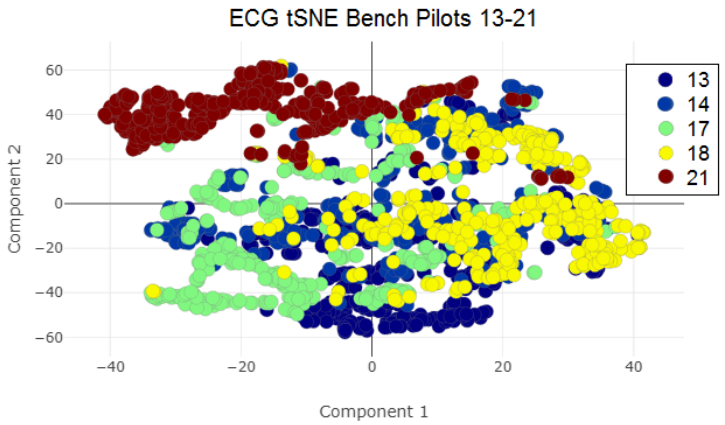


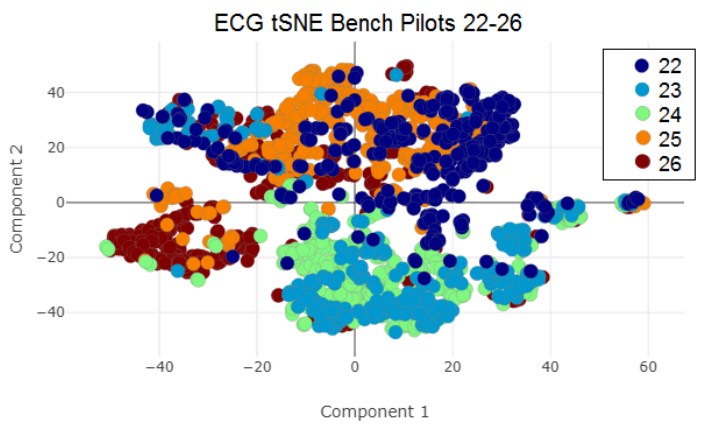


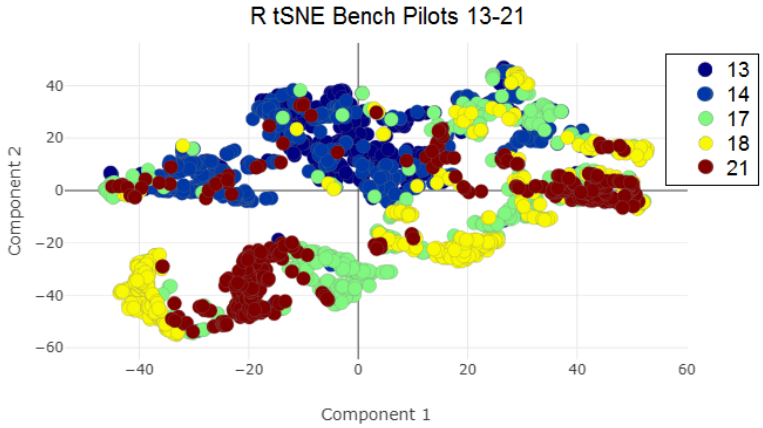


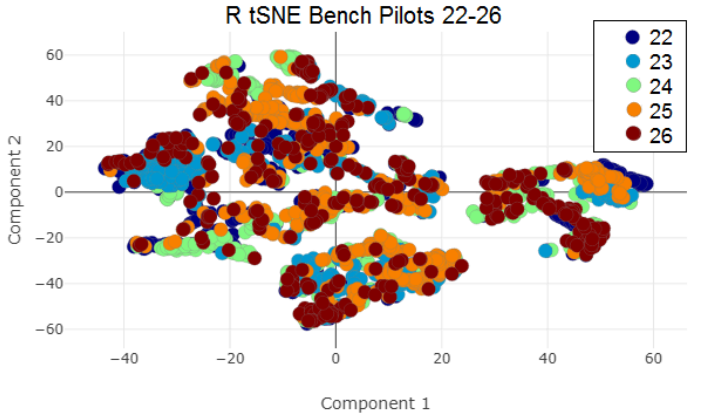


Variance Challenge



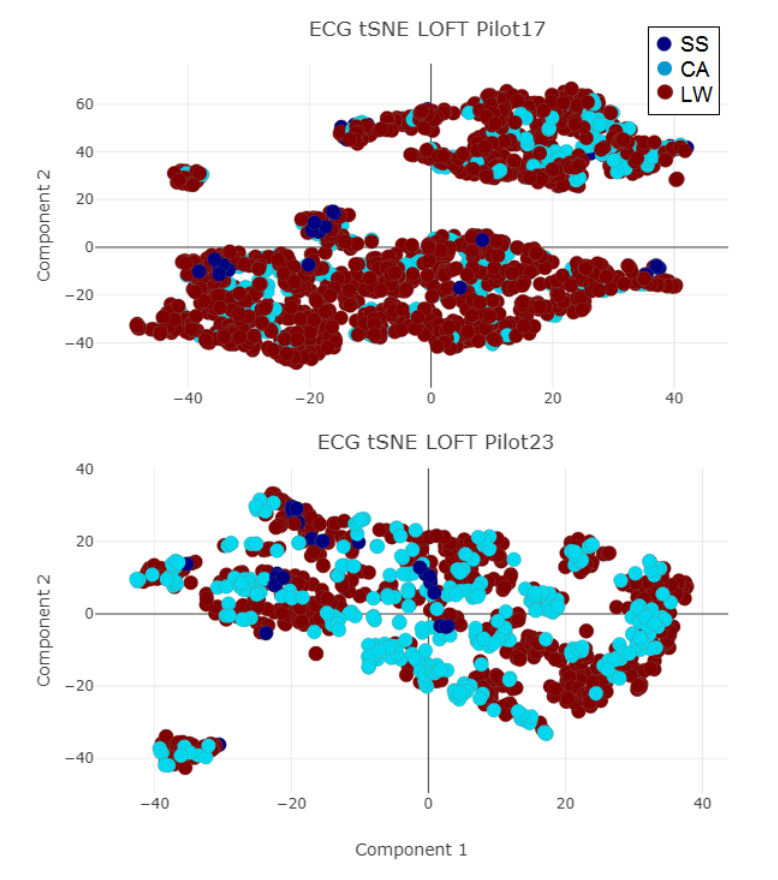


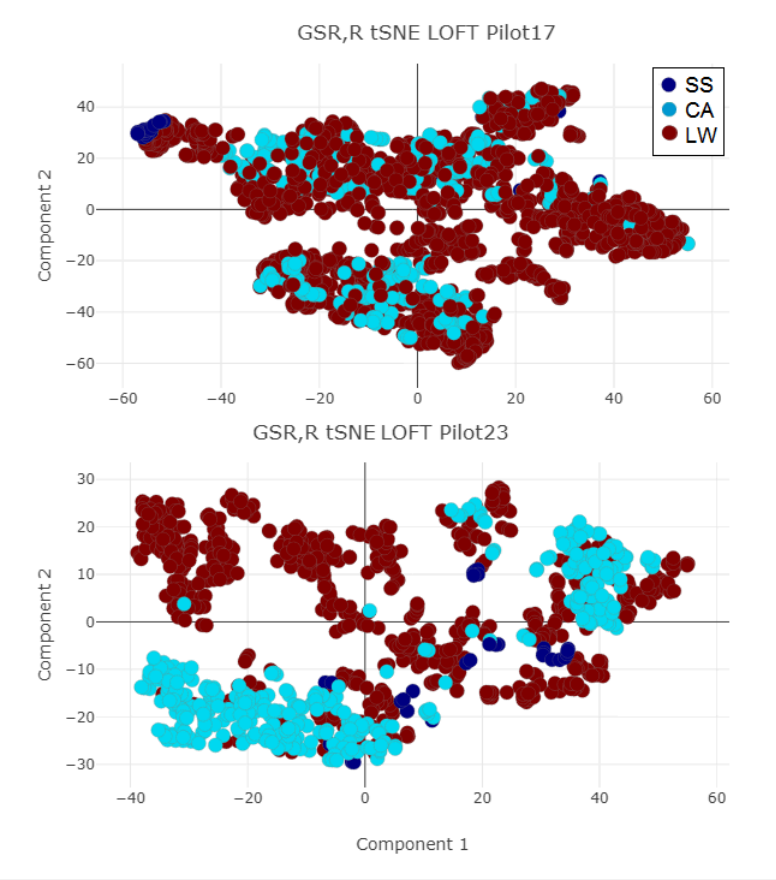




Variance Challenge

Preferred Modalities

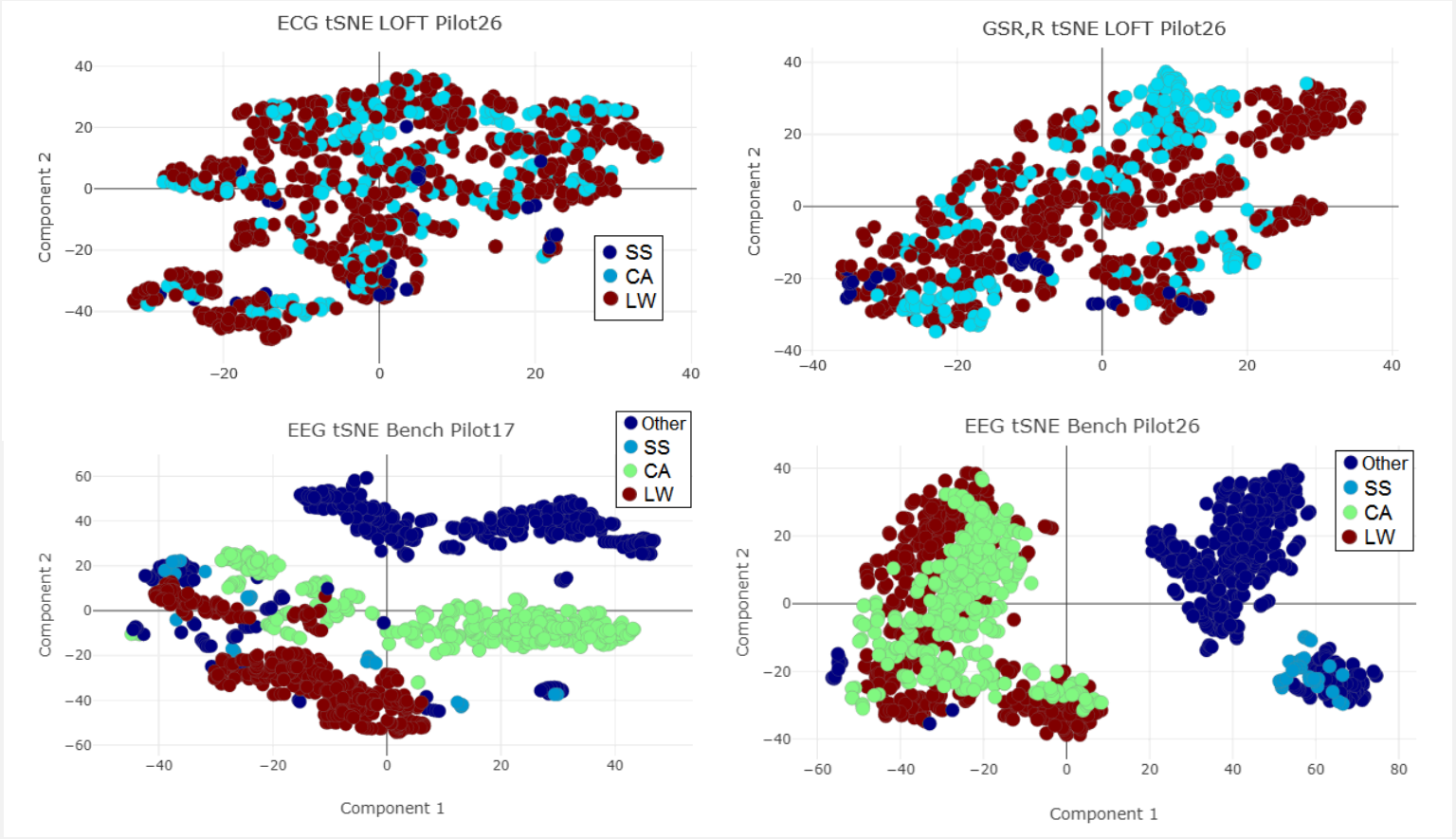




SS = Startle/Surprise
CA = Channelized Attention
LW = Low Workload

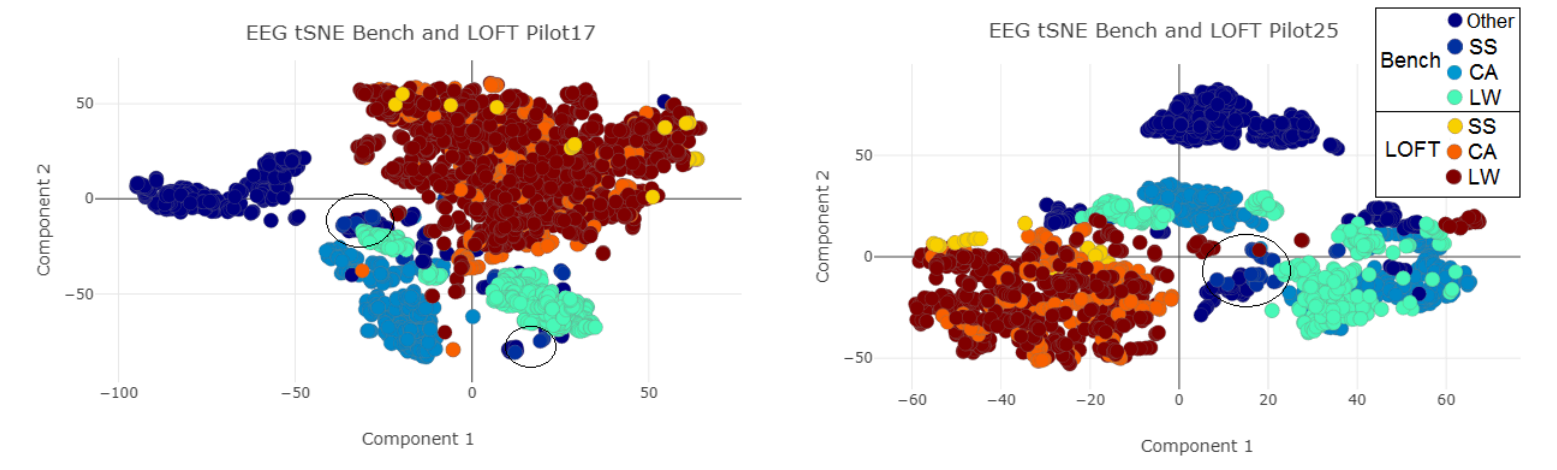
Variance Challenge

Preferred Modalities



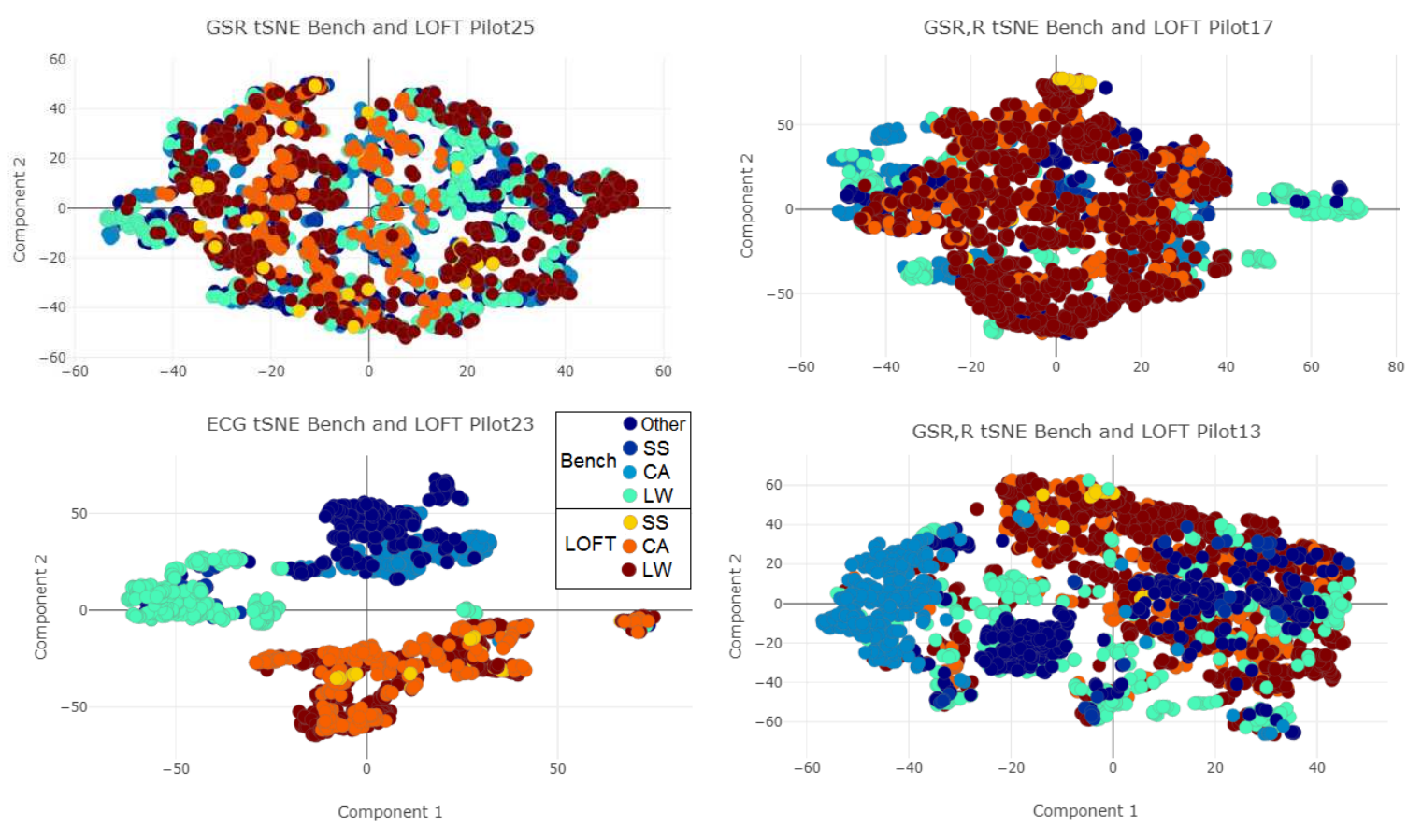
Variance Challenge

Difference Between LOFT and Benchmarks



- Often difficult to translate between benchmark and LOFT

Variance Challenge

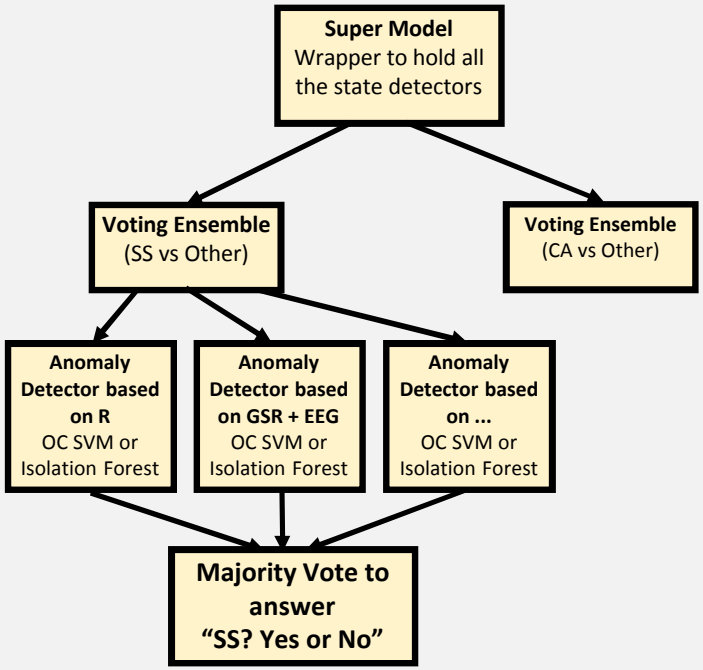


The Pipeline

Training a Pilot Dependent Model

Data Science

- Input raw sensor data
  - SS, LW, and CA Day 1 Benchmarks
  - 66% holdout of LOFT
- Generate features
- Create X State Detector for each state
- Each detector is comprised of multiple anomaly/novelty detectors (with the state being the anomaly) each built with a different modality combination
- Each anomaly detector is trained with benchmark data and individually tuned on the LOFT holdout

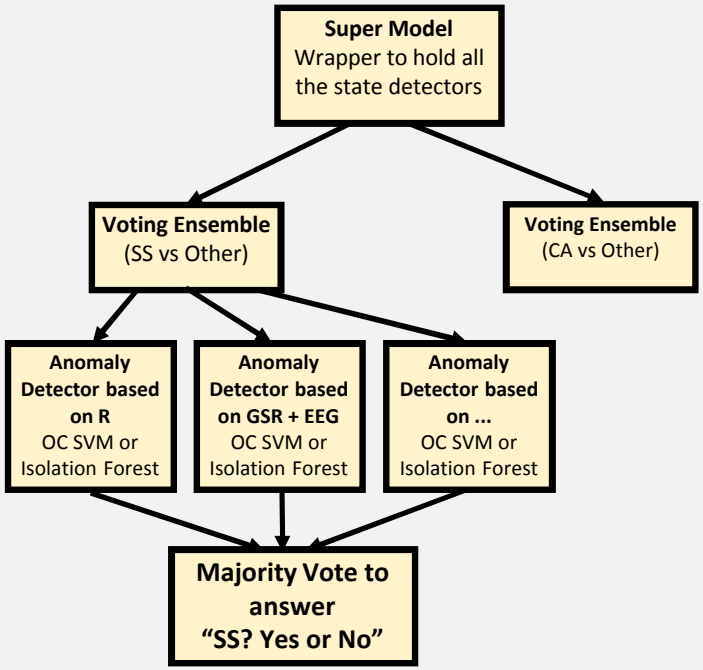


The Pipeline

Testing

- Deploy on remaining 34% of LOFT
- For each testing window:
  - Generate features
  - For each state detector:
    - Run data through each model in the ensemble and gather predictions
    - Hold a simple majority vote to choose the prediction
    - Alternative: return the probability:

$$\frac{\#\ number\ of\ yes\ votes}{\#\ number\ of\ models}$$



Results

Data Science

SS Detector Metrics

Pilot	AUC	ACC	SS ACC	Other
13	0.5073	0.9029	0.0909	0.9236
14	0.6309	0.8849	0.3636	0.8981
17	0.7504	0.5099	1	0.5008
18	0.5495	0.8162	0.2727	0.8263
21	0.8945	0.7929	1	0.7889
22	0.6772	0.7165	0.6364	0.718
23	**0.8752**	**0.9283**	**0.8182**	**0.9323**
24	0.662	0.6012	0.7273	0.5968
25	0.7034	0.6812	0.7273	0.6794
26	0.7384	0.6644	0.8182	0.6585
Avg	**0.6989**	**0.7499**	**0.6455**	**0.7523**
STD	**0.1179**	**0.1314**	**0.2916**	**0.1388**

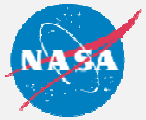
CA Detector Metrics

Pilot	AUC	ACC	CA ACC	Other
13	0.4991	0.5643	0.3554	0.6429
14	0.6321	0.754	0.3636	0.9006
17	0.4554	0.4189	0.51	0.4008
18	0.5291	0.3278	0.83	0.2282
21	0.6438	0.7606	0.404	0.8836
22	0.6388	0.5271	0.8676	0.41
23	0.4617	0.3531	0.8091	0.1143
24	0.5771	0.675	0.2636	0.8905
25	0.6354	0.5336	0.8378	0.433
26	0.4448	0.349	0.6351	0.2545
Avg	**0.5517**	**0.5263**	**0.5876**	**0.5158**
STD	**0.0789**	**0.1558**	**0.2236**	**0.2799**

AUC: Receiver Operating Characteristic Area Under the Curve
ACC: Straight Accuracy Metric
X ACC: How well did we predict the state?
Other: How well did we predict “Other” as “Other”?

Takeaways and Future Work

- Generalization is difficult (in general ☺)
  - Some modalities are easier to generalize than others
  - Some state detection is easier to generalize than others
  - Some machine learning tools can be used to mitigate some of this
- Work is continuing on many avenues:
  - Different Features
  - Different Modalities
  - Different Sensors
  - Concentrating on Startle/Surprise



Questions?

- Commercial Aviation Safety Team, "Airplane State Awareness Joint Safety Analysis Team Interim Report," URL: http://www.skybrary.aero/index.php/Commercial_Aviation_Safety_Team_%28CAST%29_Reports [cited 5 March 2017]
- Harrivel, Angela R., et al. "Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing." *AIAA Information Systems-AIAA Infotech @ Aerospace*, May 2017, doi:10.2514/6.2017-1135.
- Christensen, J. C., Estepp, J. R., Wilson, G. F., Russel, C. A.,"The effects of day-to-day variability of physiological data on operator functional state classification," NeuroImage, Vol.59,2012, pp.57–63.
- Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., Craven, P. L., "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks," Aviation, Space and Environmental Medicine, Vol. 78, No. 5, 2007, Section II.
- Li, F., "Improving Engagement Assessment by Model Individualization and Deep Learning," Dissertation, Old Dominion University, 2015.
- Gross, J. J. and Levenson, R. W. Emotion elicitation using films. Cognition & Emotion, Vol. 9, 1995, pp. 87-108.
- Pope, A. T., Bogart, E. H., Bartolome, E. S., "Biocybernetic system evaluates indices of operator engagement in automated task," Biological Psychology, Vol. 40, 1995, pp. 187-195.
- Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock, J. R. (2011). The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide. NASA, Langley Research Center. Hampton: NASA/TM-2011-217164, L-20031, NF1676L-12800.
- "Scikit-Learn." *Scikit-Learn: Machine Learning in Python*, 0.19.0, scikit-learn.org/stable/index.html.
- Bao, Forrest Sheng. "PyEEG." *PyEEG Reference Guide*, 0.02 r1, pyeeg.sourceforge.net/.
- Python Core Team. "Python." *Python Software Foundation,* 3.6.3, https://www.python.org
- A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, M. Hämäläinen, MNE software for processing MEG and EEG data, NeuroImage, Volume 86, 1 February 2014, Pages 446-460, ISSN 1053-8119
- A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, M. Hämäläinen, MEG and EEG data analysis with MNE-Python, Frontiers in Neuroscience, Volume 7, 2013, ISSN 1662-453X