Strategies to increase trust in machine learning models

Based on experiments with wearable sensor data

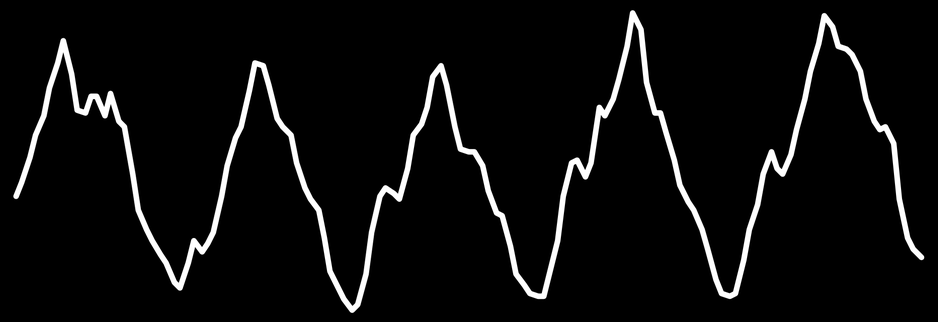
Sagar Sen, 30.04.2021

Outline

- Physiological signals - the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Physiological signals

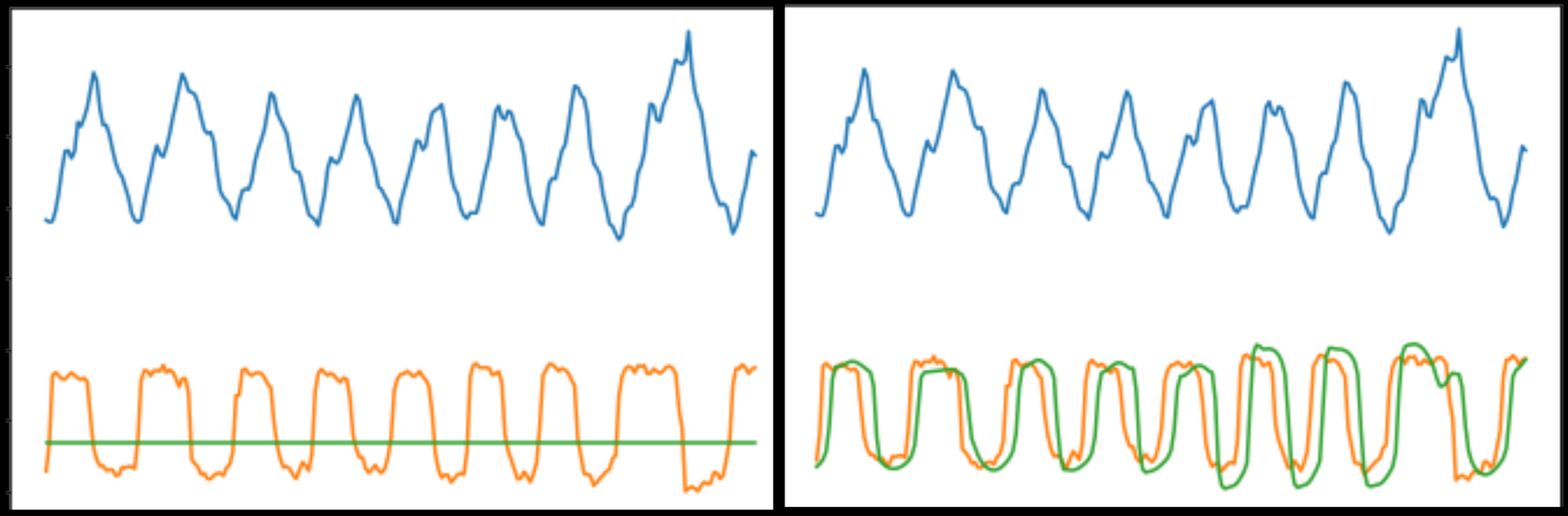
Interpreting breathing patterns



Physiological signals

Interpreting breathing patterns to *predict* effort

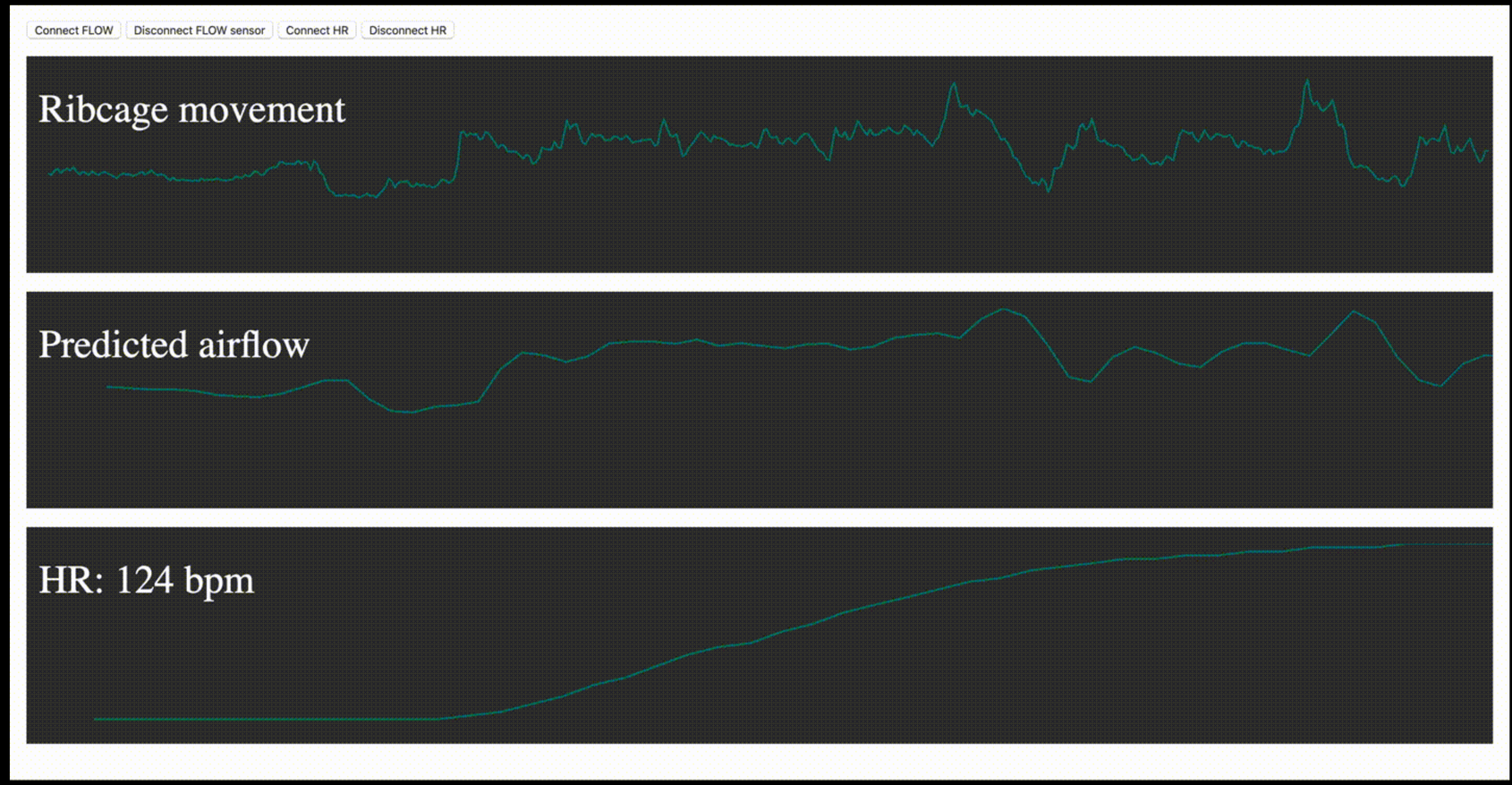
input
output
estimated



In one afternoon's work: 95% *accuracy*

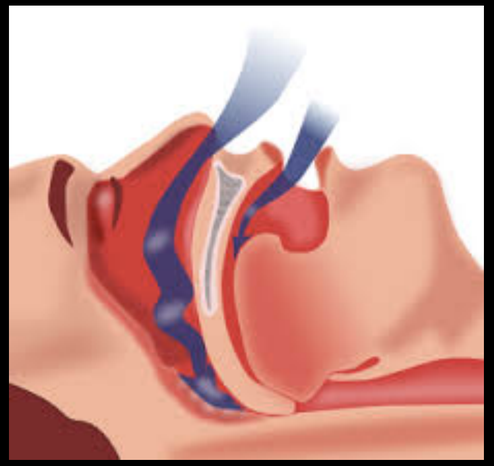
Physiological signals

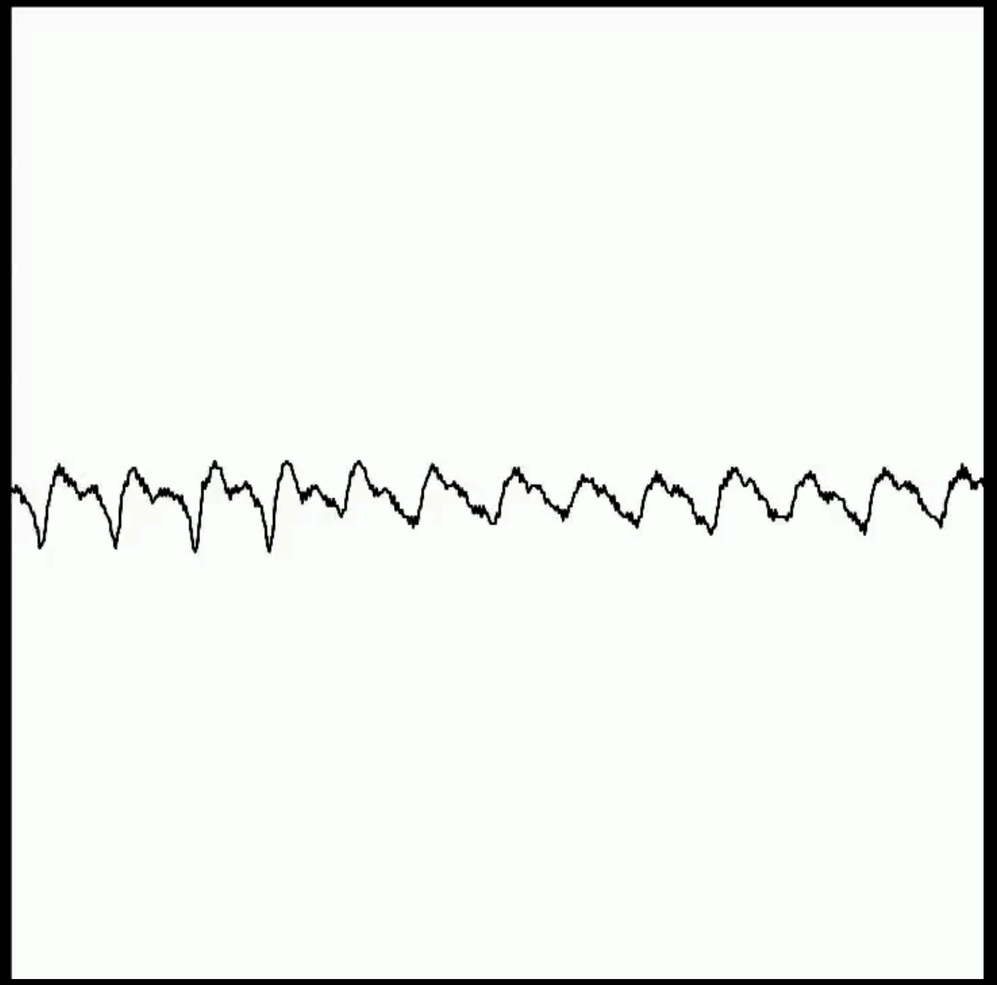
Interpreting breathing patterns to *predict* effort



Physiological signals

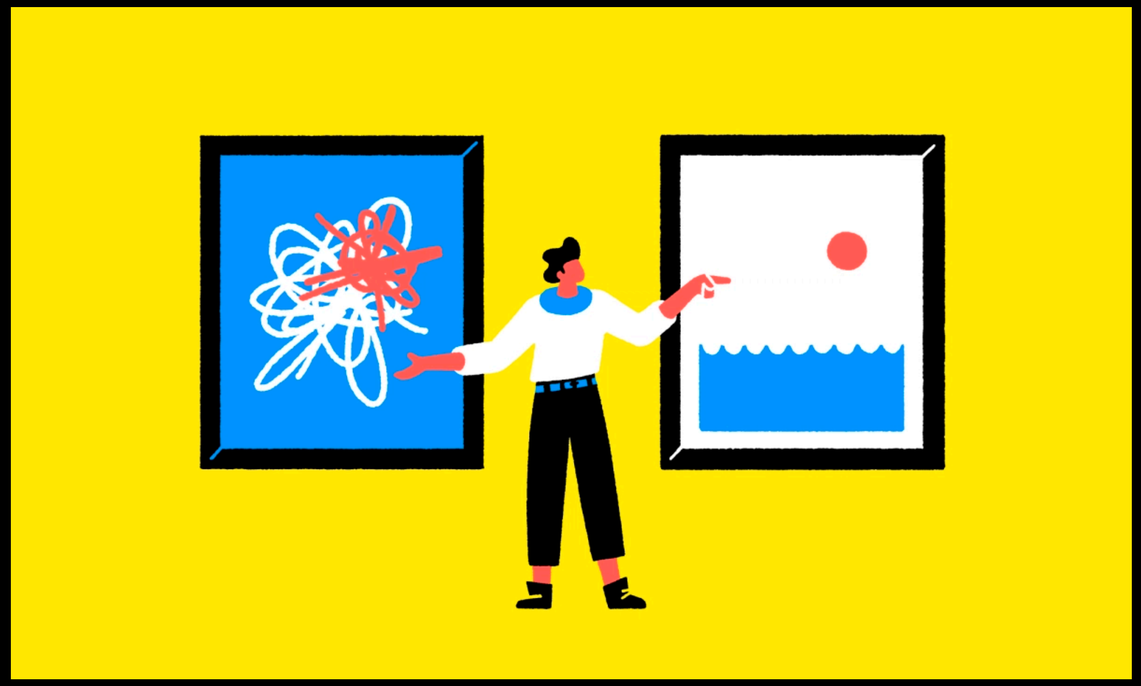
Interpreting breathing patterns to *classify* apnea





Physiological signals

Can I explain how these predictions happen?



Mapping to an Interpretable Domain

Diagrams

Natural Language

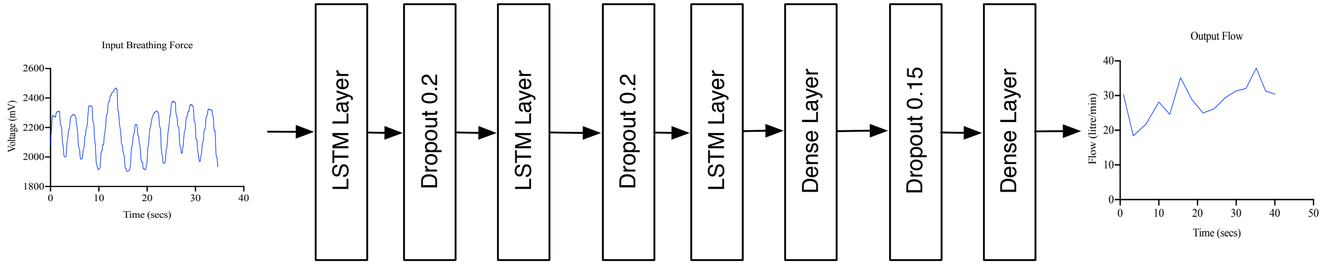
Interpretable Input

Outline

- Physiological signals - the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Machine learning models

Exploring explanations with LIME for TIME

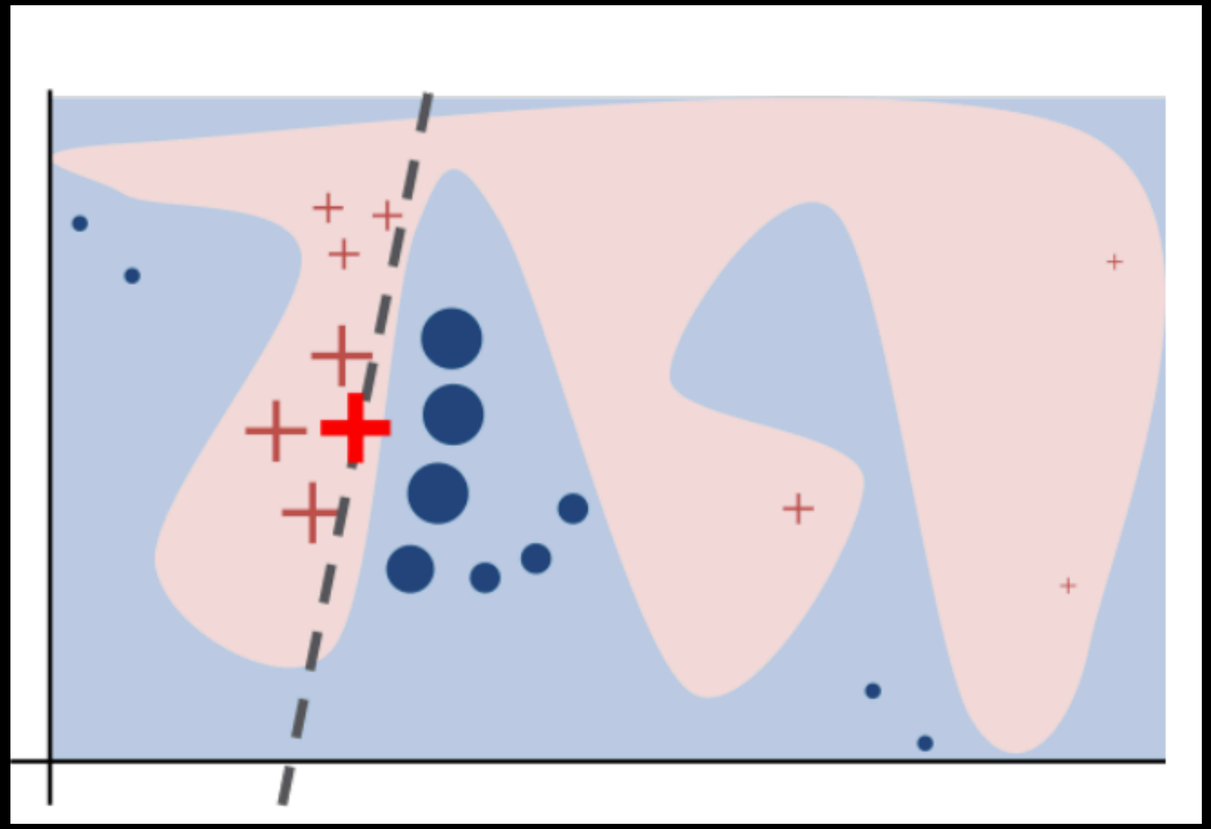


LSTM layer
Historic size: 100 points
Hidden neurons: 50

Architectures

- Bi-directional LSTM
- Variational Auto-encoder
- Transformer models
- Dense feedforward neural network

Sparse Linear Explanations in LIME

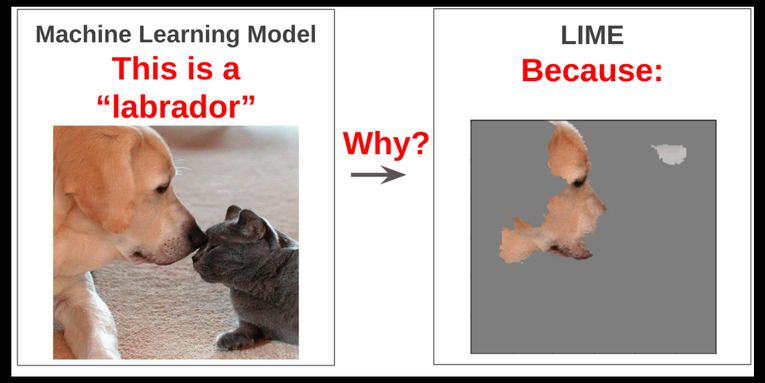


$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

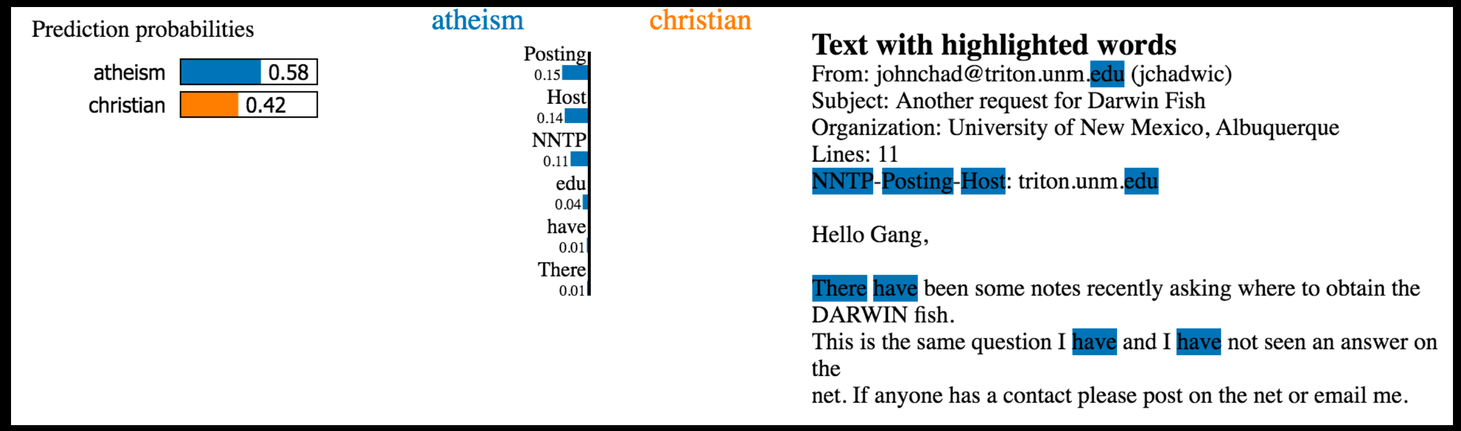
$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z') \right)^2 \qquad (2)$$

Interpretable data representations

Super-pixel - Presence or absence of a continuous patch of similar pixels essentially a tensor with three colour channels per pixel.

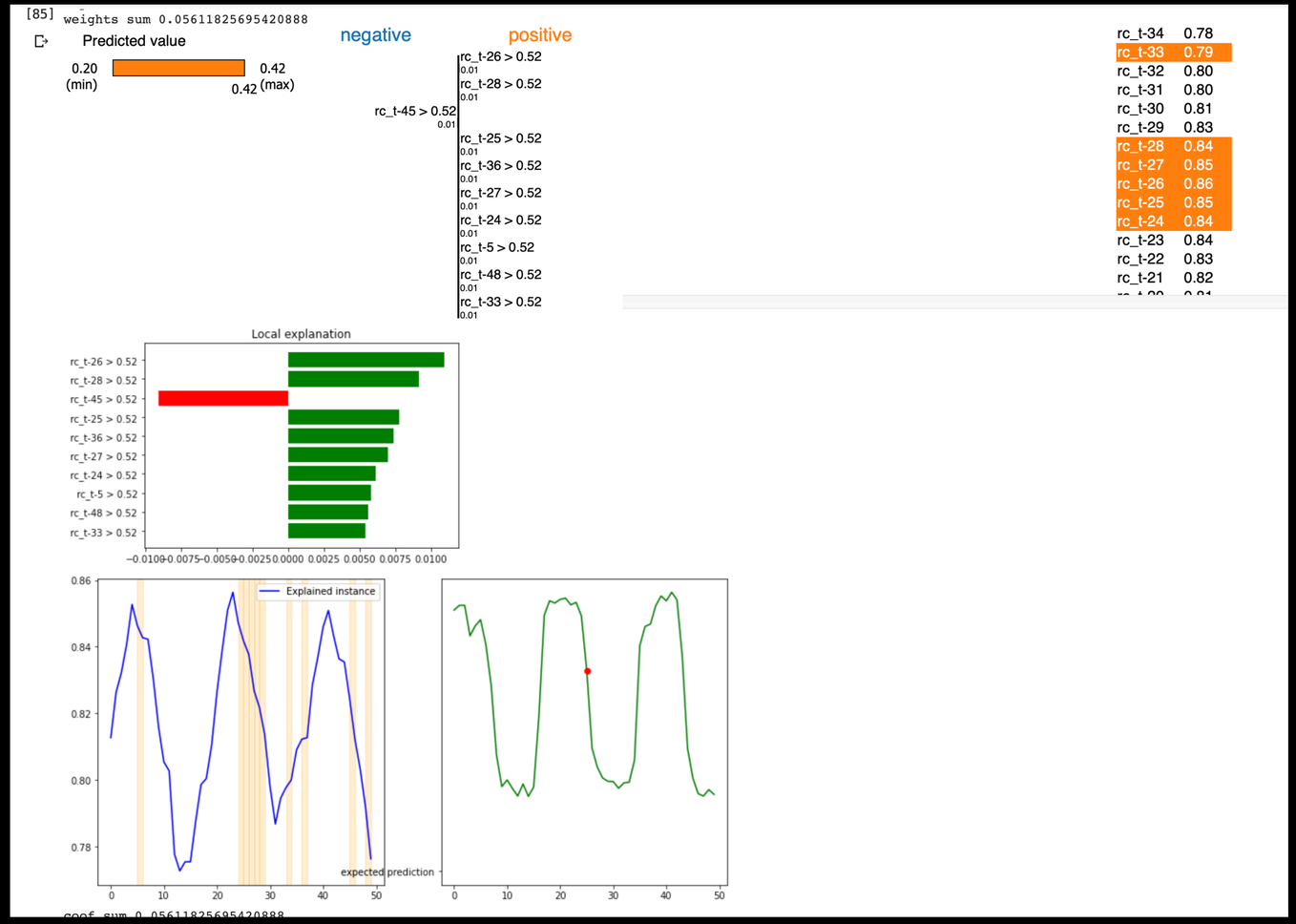


Binary vector indicating presence or absence of words - even though some classifiers use word embeddings such as bag of words



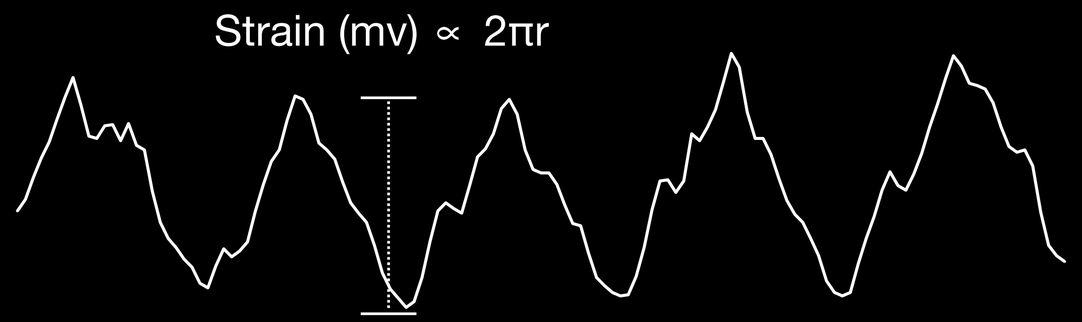
Raw input data in time domain with LSTM

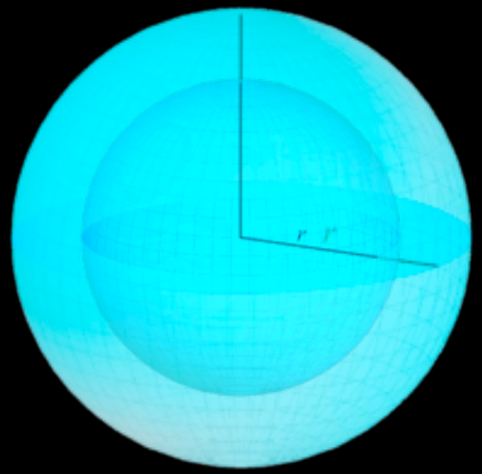
Explaining the prediction of one point

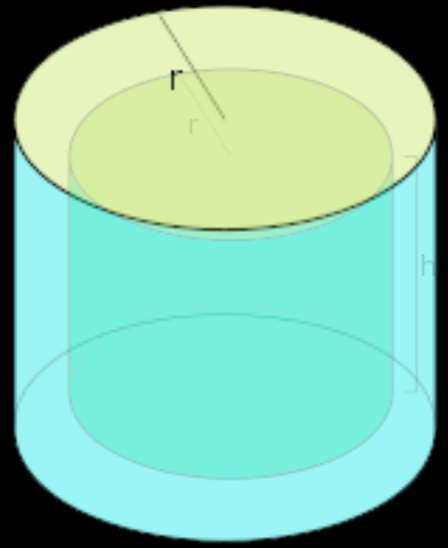


Work done with intern Gutama Ibrahim this summer

Can machine learning models match our perception of reality?

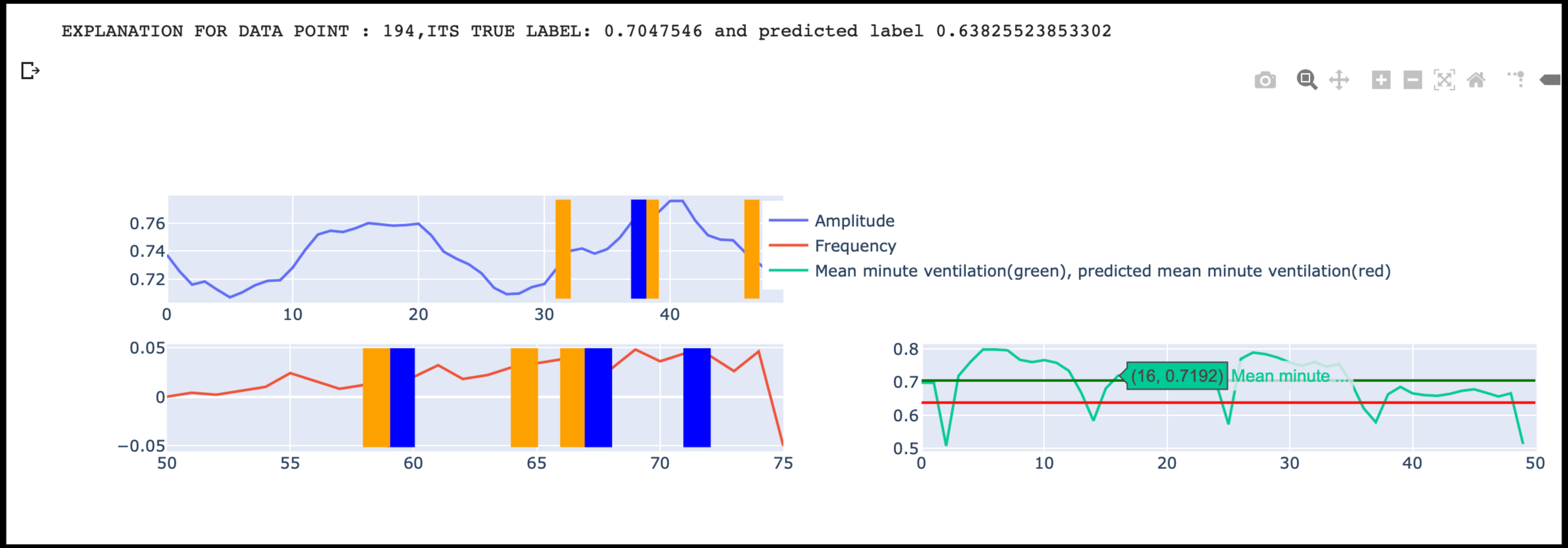






With more input feature engineering

Explaining prediction of mean from frequency and amplitude



Looking at mean output instead of point in a sequence

Outline

- Physiological signals - the context
- Exploring explanations for models that interpret physiological signals
- Conclusion

Conclusion

1. Current state of the art in explainable AI works better for classification
2. All machine learning models had very high training and test accuracy but are the explanations palatable?
3. Input and output features can be modified before training such that the explanation is closer to our perception of reality. (e.g. a cylinder)
4. Machine learning models need to be explored such that they use meaningful input features to produce physically meaningful output features. (e.g. amplitude and frequency)
5. Machine learning models that use seemingly meaningless inputs to get very high accuracy need to be discarded.

A flowchart for trustworthy AI models