Methods for Interpretable Machine Learning

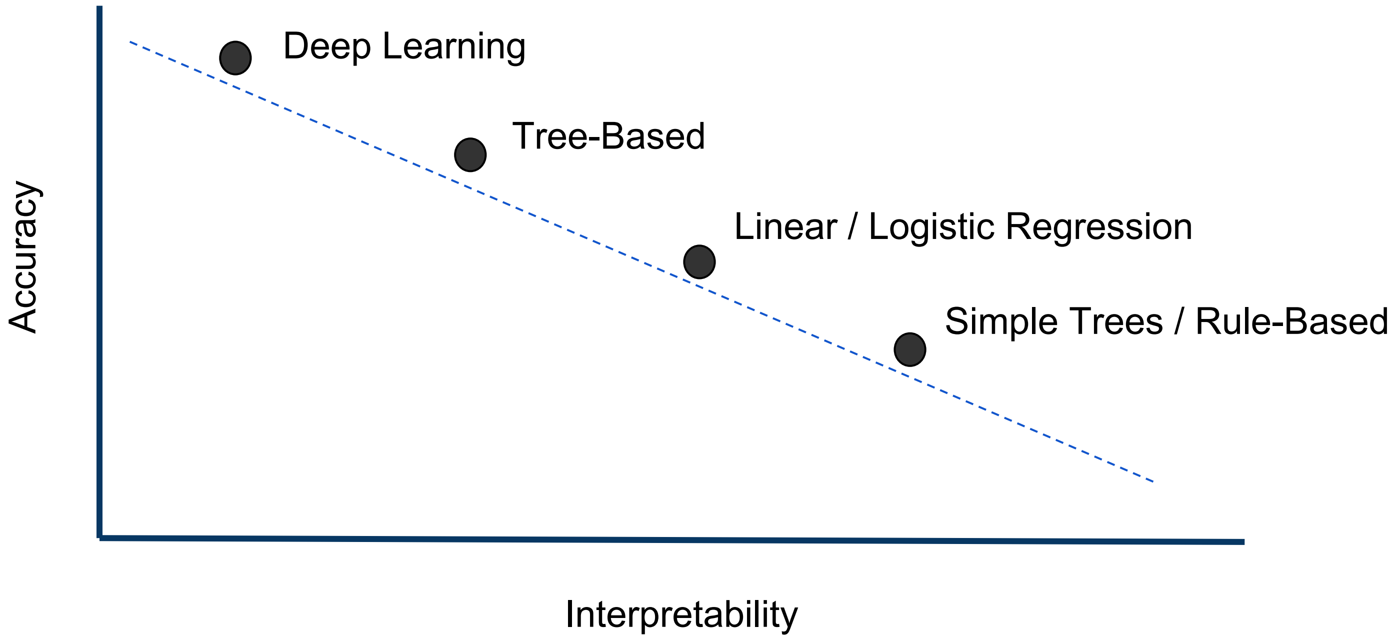
PyData - Dec 4, 2018

About JOOL

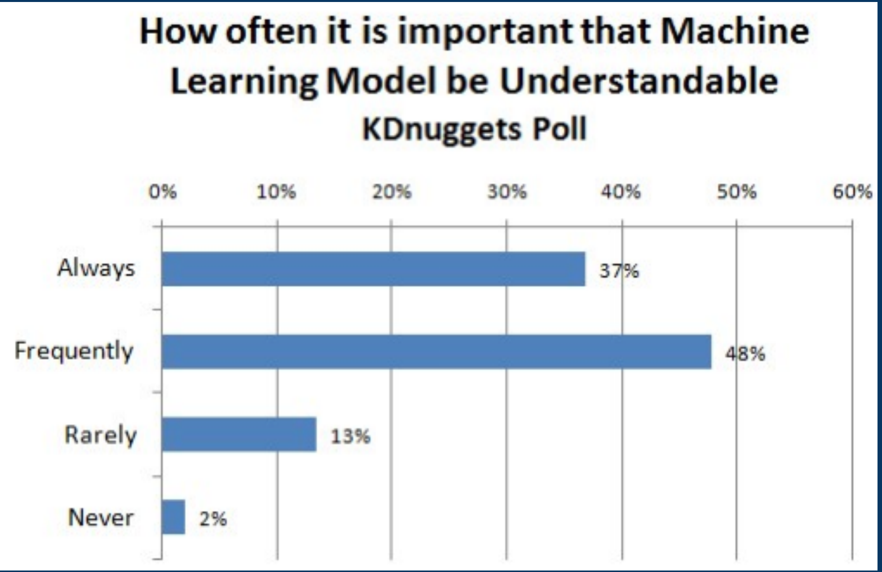
What do you mean "interpretable"...

- Human-interpretable?
- Point Estimates, a linear equation?
- Feature Importance?
- Stability / Reproducibility?

Accuracy vs. Interpretability Trade-off



Why do we care?



https://www.kdnuggets.com/2018/11/machine-learning-model-understandable-poll-results.html

Generalization Problems

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

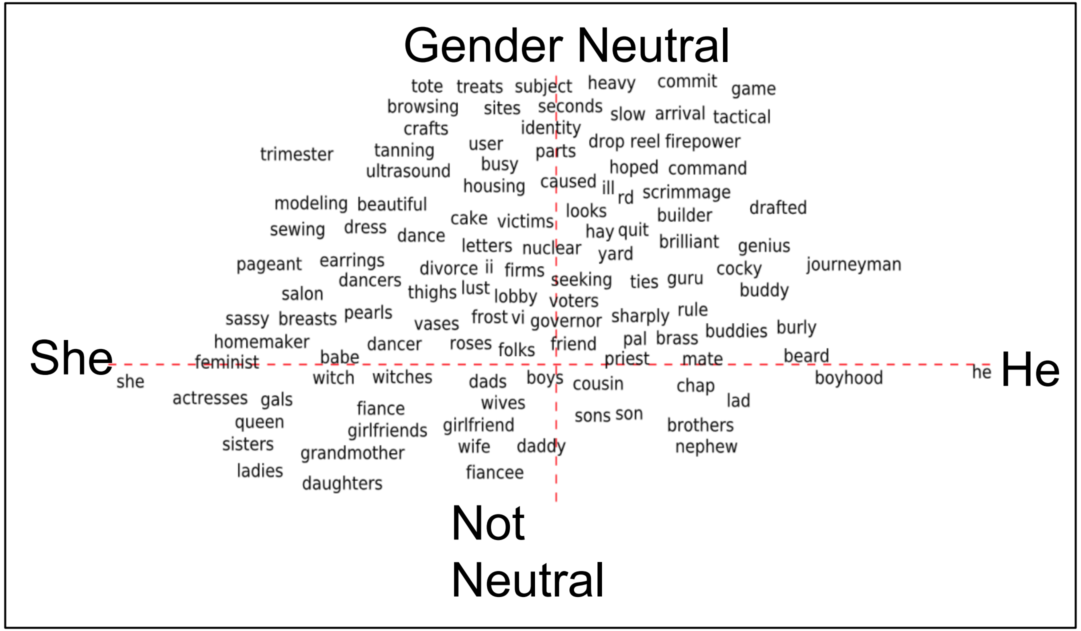
	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

- A model can learn elements from data that aren't core to the problem being solved
  - Over-fitting
  - Spurious correlations (E.g. wolves are more likely to be found in snow than Huskies)

Tulio Ribeiro, Marco & Singh, Sameer & Guestrin, Carlos. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97-101. 10.18653/v1/N16-3020.

Models Inherit Bias in the Data



Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS*.
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Audit a Model

- Backed by domain knowledge
  - effect size or direction grossly different from expectations
  - latent variables
- Verify safety/limitations
  - local areas of poor accuracy

User Buy-in

“The weather doesn’t influence my sleep! This doesn’t make sense!”

“You can’t possibly predict my eating tomorrow.”

“This recommendation doesn’t apply to me at all!”

“Show me the data.”

Regulatory Requirements

- **Finance**: Fair Credit Reporting Act requires that companies notify a consumer if consumer report information is used to deny credit
- **FDA/Healthcare**: Audit/explain the decision process
- **GDPR**: "Where personal data relating to a data subject are collected from the data subject, the controller shall...provide the data subject with…(f) existence of automated decision-making, including profiling...meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." - *Article 13*

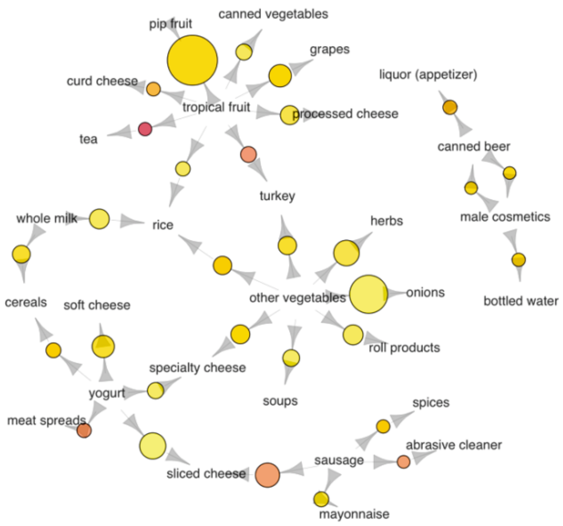
“Perfectly” interpretable approaches

MLxtend.frequent_patters.apriori

Rule-based (Assoc. Rules)

$$\{A, B\} \rightarrow \{C\}$$

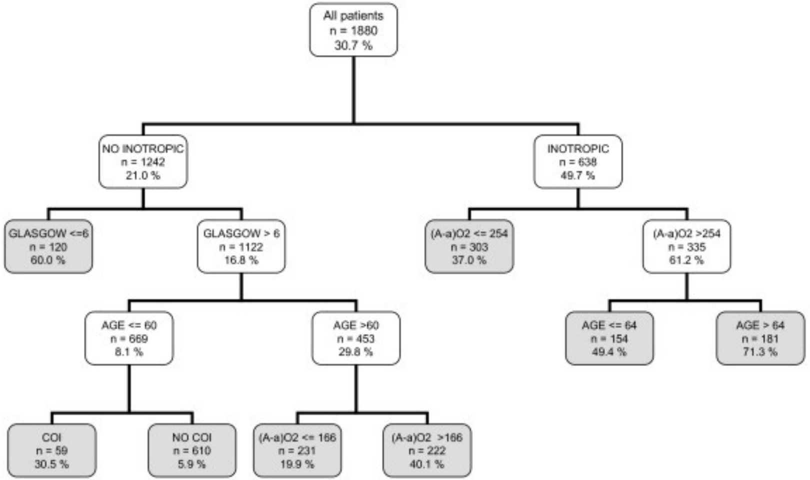
- Apriori, Eclat, FP-Growth
- “If A and B occur, C occurs X% of the time”



sklearn.tree.DecisionTreeClassifier

Simple Decision Trees

- Variables have easy to follow split-points that segment outcomes
- Shorter path length is more interpretable



Trujillano J, Badia M, Serviá L, March J, Rodriguez-Pozo A. Stratification of the severity of critically ill patients with classification trees. *BMC Med Res Methodol*. 2009;9:83. Published 2009 Dec 9. doi:10.1186/1471-2288-9-83

sklearn.linear_model

Linear/Logistic Regression

- Point Estimates
- P-Values
- Odds Ratios

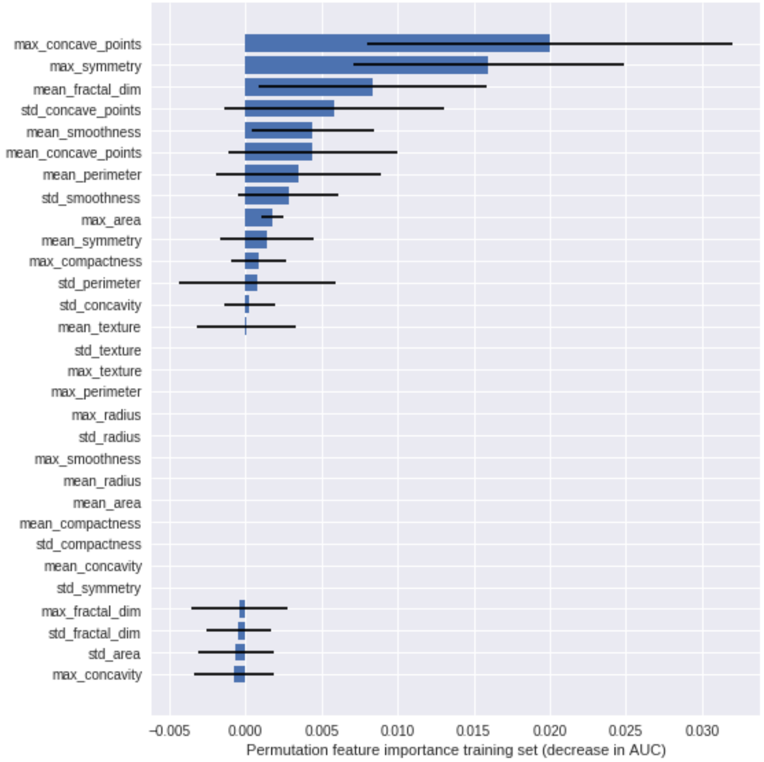
$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1}\overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Semi-interpretable approaches

Variable Importance

sklearn.ensemble

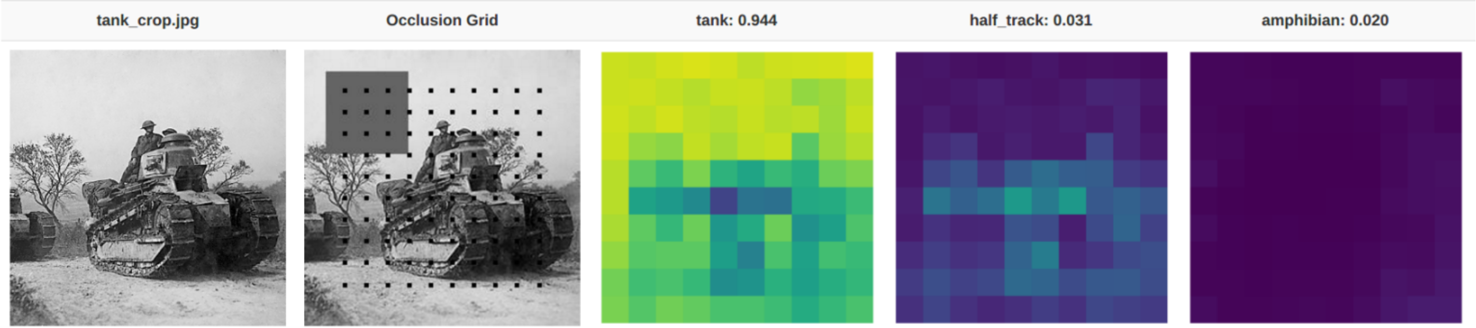
- Random Forest, GBM
- Variables are included/excluded in various model iterations
- Measure importance by decrease in accuracy or node purity

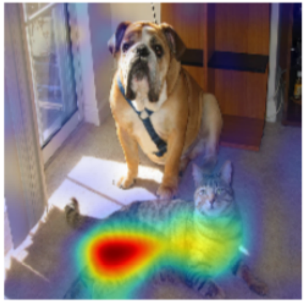


Tensorflow, PyTorch, Keras

Neural Network Approaches

- Gradient-Based Methods (Saliency)
  - Partial diff of output w.r.t input
  - Encoder - Decoder Network
  - Use Gradients of last CNN Layer (Grad-CAM)





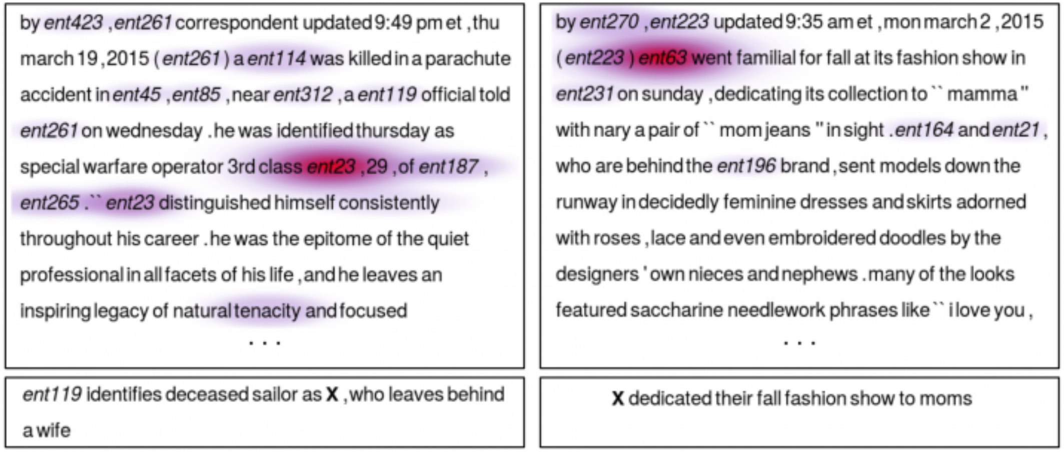
(c) Grad-CAM 'Cat'

(i) Grad-CAM 'Dog'

Tensorflow, PyTorch, Keras

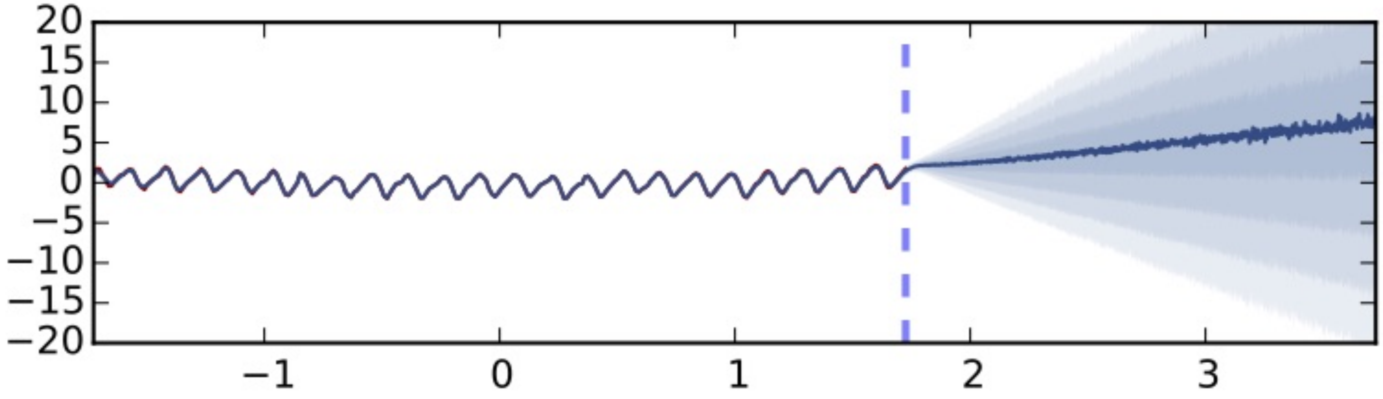
Neural Network Approaches

- Attention Methods (Memory Networks)
  - Visualize Attention Matrix
  - Commonly used with LSTM and CNN architectures



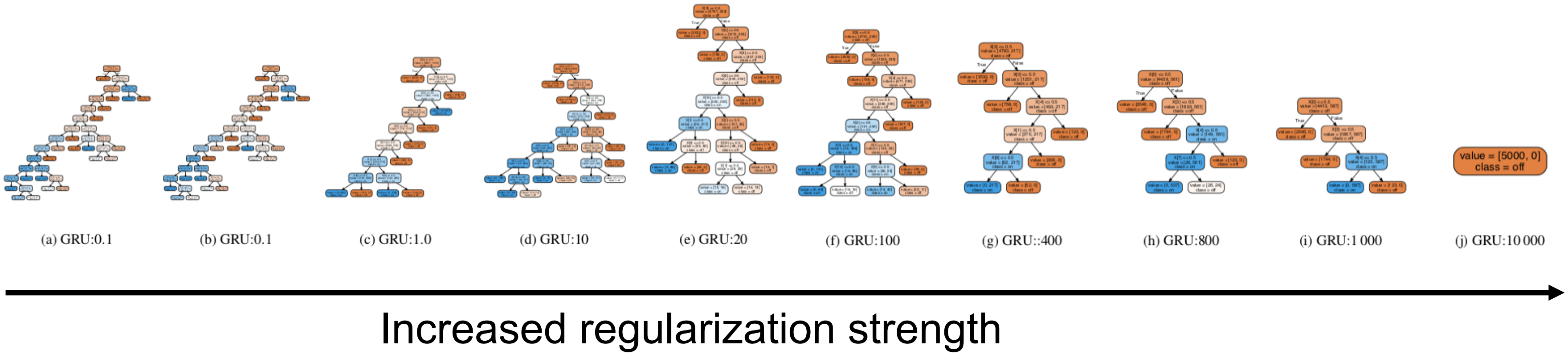
Neural Network Approaches

- Apply Dropout on Inference
  - Requires many additional predictions
  - Returns something similar to a Bayesian Posterior*



Neural Network Approaches

- Regularize on the depth of an approximate decision tree
  - Able to produce a decision tree that approximates the complex learned relationships
  - Results in networks that have less complexity given any level of accuracy
  - No work yet on problems with non-interpretable data points (images)



Mike Wu, Michael C. Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, Finale Doshi-Velez. Beyond Sparsity: Tree Regularization of Deep Models for Interpretability. https://arxiv.org/abs/1711.06178

Model agnostic approaches

Vary Inputs -> Measure Output

Pros:

- Works for any model
- X change in input yields an expected Y change in output

Cons:

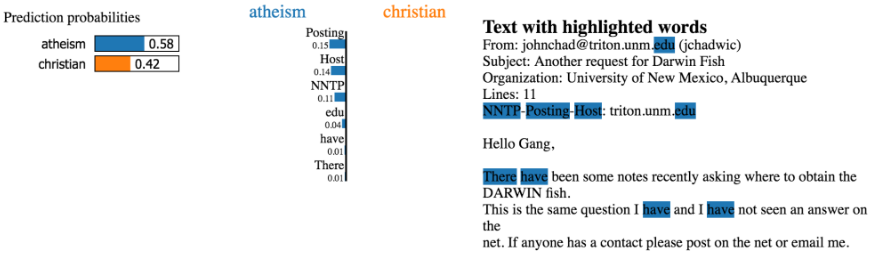
- Requires careful planning and understanding of the problem / data
- Requires multiple predictions on same observation
- May want to maintain feature covariance, depending on the model

Local Interpretable Model-Agnostic Explanations (LIME)

- Vary input data by zeroing out features in the chosen observation
- For images, create “super pixels”
- Weight points by similarity to original
- Fit a simplified linear model on the perturbed observations
- Interpret the linear model

Original Image

Interpretable Components



Takeaways

Model Interpretability Summary

Difficult to Interpret	Semi-Interpretable	"Perfectly" Interpretable
- Neural Networks - Multi-model ensembles	- GBM/XGBoost - Random Forest - Large Decision Trees - Engineered Features	- Association Rules - Simple Decision Trees - Linear/Logistic Regression

Good Practices

1. Interpretably can be more important than accuracy
2. Use more interpretable models when possible
3. Keep the audience in mind
4. Consider limitations and biases in the data
5. Several methods exist for interpreting “Black Box” models

Good Reads

- The Mythos of Model Interpretability - Zachary Lipton
- Introduction to LIME
- What My Deep Learning Model Doesn't Know... - Yarin Gal
- Teaching Models to Read and Comprehend
- Beyond Sparsity: Tree Regularization of Deep Models for Interpretability

Questions?

Brandon Stange
Data Scientist, JOOL Health
Brandon.Stange@gmail.com

Haitham Maya
Data Scientist, JOOL Health
haitham@mmaya.me