# Interpreting Models with Shapley Values

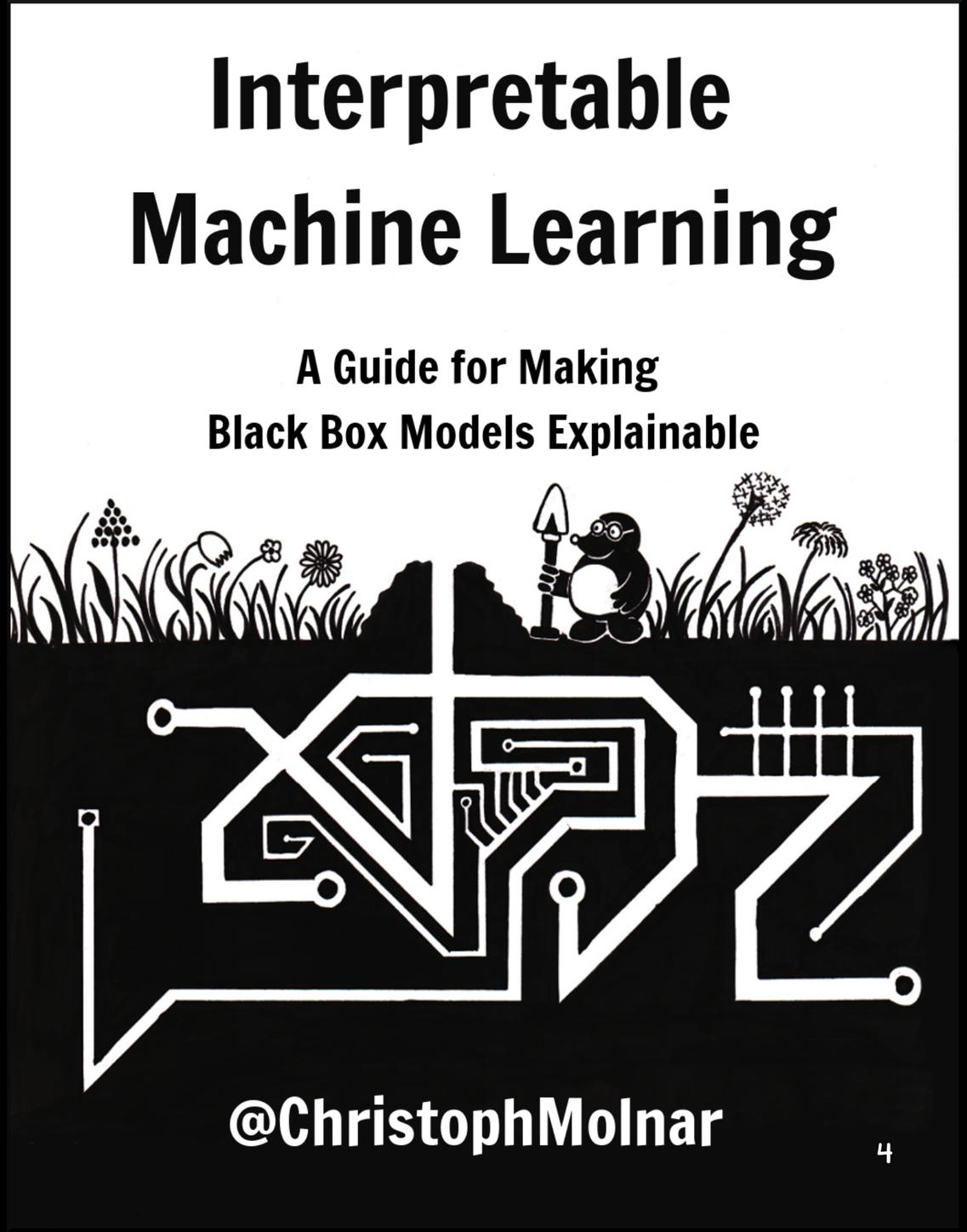
# Agenda

1. Intro to interpretability 🕵️
2. What are shapley values ❓
3. Strengths 💪
4. Weaknesses 💩
5. Our experience with it 👨‍🔬👨‍🔬
6. Situations where it shines 🌟

# Intro to interpretability? 

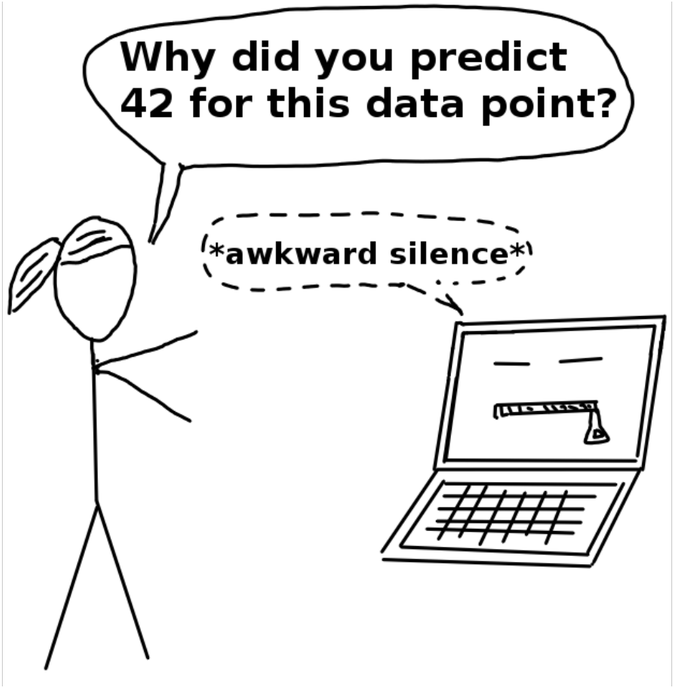
# Credits to Christoph Molnar[1]

[1] Credits: Christoph Molnar's Book on Interpretable Machine Learning



# Black Box Algorithm

"In machine learning, 'black box' describes models that cannot be understood by looking at their parameters"
-- Christoph Molnar



# Interpretability

- ”...the degree to which a human can understand the cause of a decision.”[2]

[2] Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017).

# Explanation

"...usually relates the feature values of an instance to its model prediction in a humanly understandable way."
-- Christoph Molnar

# Why interpretability?

- Learning/Curiousity
- Safety/Bias/Debugging
- Social acceptance

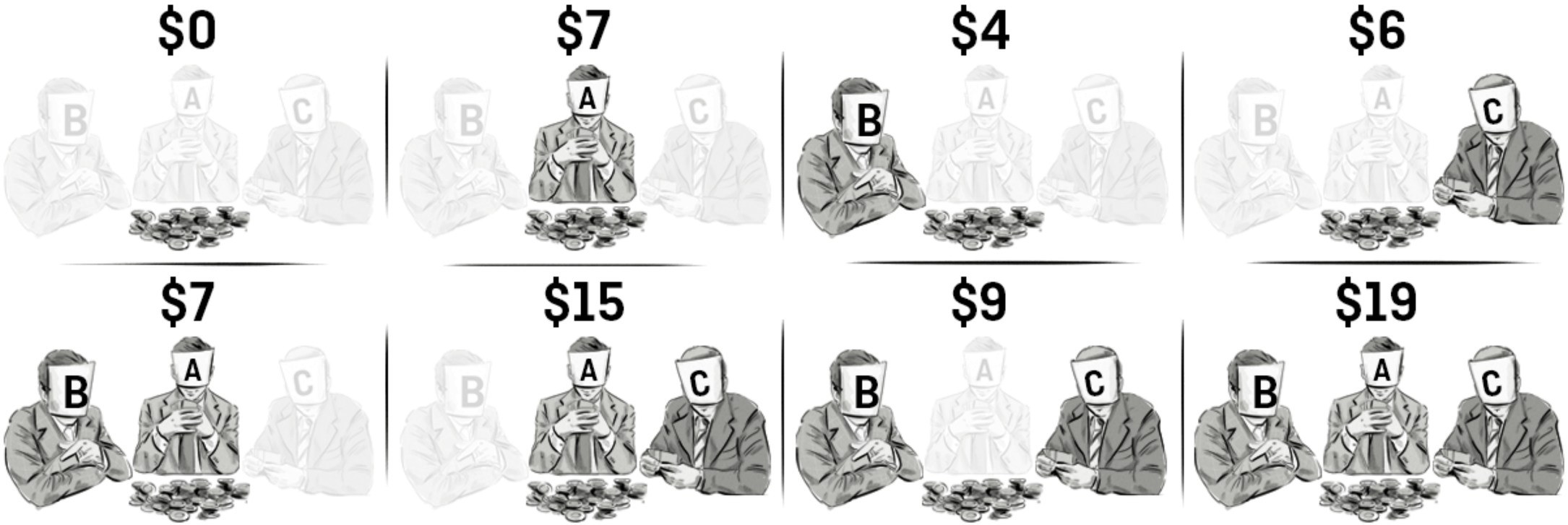
Example by Ribeiro et al (2016)

# What are Shapley Values ?

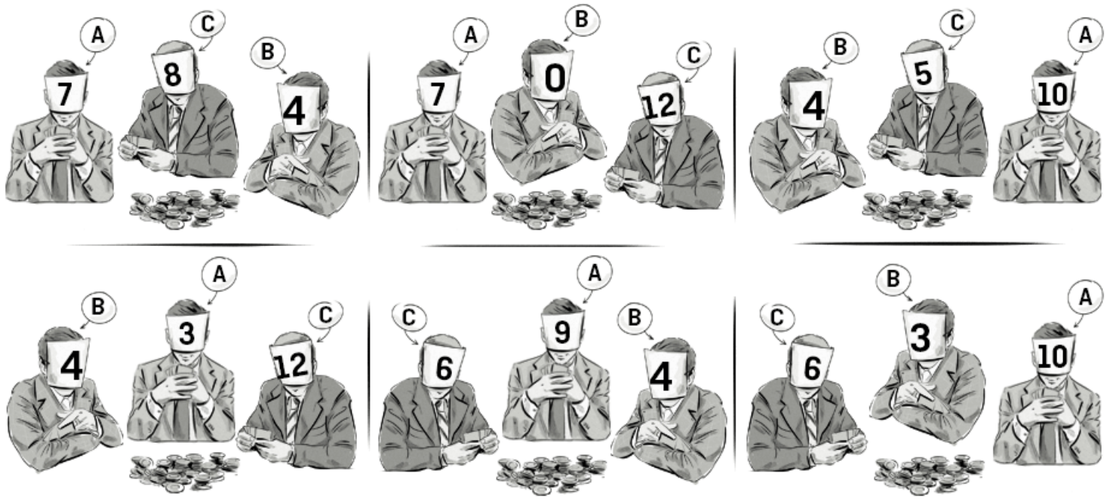
# Three pickpockets dividing a $19 loot[3]


[3] Image Credits: Michael Sweeney, "Game Theory Attribution: The Model You Probably Never Heard of" (clearcode.cc)

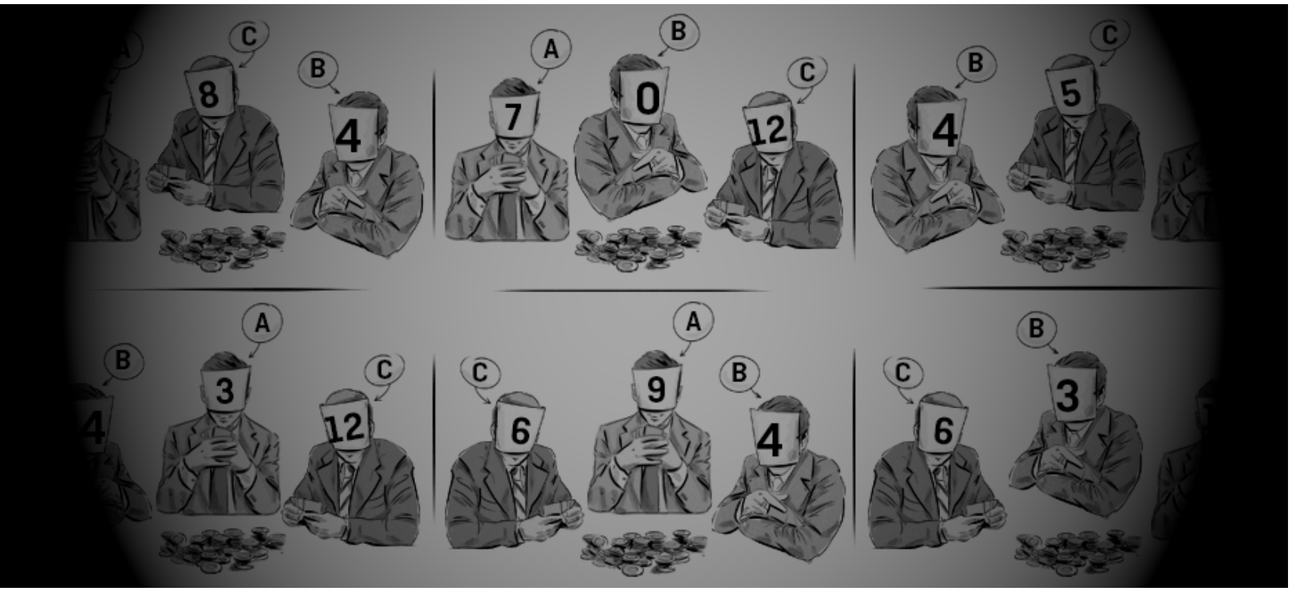
# Possible loot with subset of pickpockets



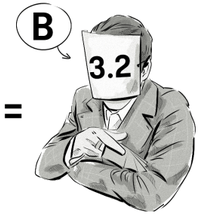
# Order Affects Marginal Contributions

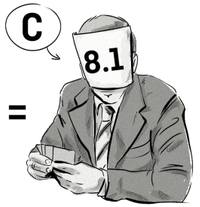



# Average Marginal Contribution




(7+7+10+3+9+10) / 6 = 


(4+0+4+4+4+3) / 6 = 


(8+12+5+12+6+6) / 6 = 

# The Shapley value is...

...the average marginal contribution of a pickpocket to the loot over all possible coalitions.

...**NOT** the difference in loot when we would remove the pickpocket from the full coalition.

# Pickpockets?

# Shapley Values @ Predictions

- pickpockets = features
- loot = target
- shapley value = feature importance
- one day of pickpocketing = an observation

# Properties of Shapley Values

- Efficiency
- Additivity
- Symmetry
- Dummy

Shapley = proven to be the only consistent linearly additive feature attribution

# Strengths 

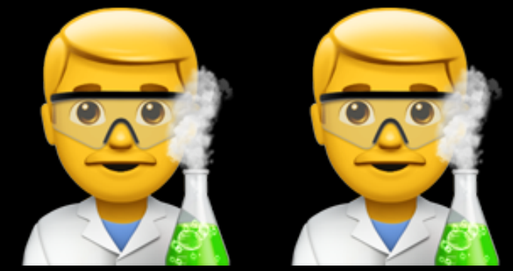
# What we like about Shapley

1. Game-theoretic foundations
2. Model agnostic
3. Shap package visualizations

# Weaknesses 

# What we don't like about Shapley

1. Shap's approximation is a blackbox to me.
2. It doesn't solve the problem of correlated features
3. Slow on non-tree models

# Our Experience 

# Shapley @ ML: SHAP

- Read the papers[4] and tried the python package[5]
- Compared with other techniques[6]
- Implemented shapley from scratch on a toy problem[6]
- Applied it for feature selection[6]

---

[4] Lundberg & Lee (2017), Lundberg, Erion & Lee (2018)

[5] https://github.com/slundberg/shap

[6] https://github.com/rensdimmendaal/shapley-exploration

# Where Shapley Shines 

# Shapley is a good candidate when you need

1. To explain a model's predictions, but really cannot use a simpler model
2. Feature selection
3. Feature engineering

# Lazy Modeling Loop

1. Linear model --> underfit
2. Complex model --> overfit
3. Regularize complex model --> get it just right

# Extended Modeling Loop

1. Use shapley --> find relevant features
2. Dependence plots --> find shape & interaction
3. Fit a linear model on subset of engineered features

--> high performance with low model complexity


#CBBTV

# Conclusion

1. Don't blindly trust blackbox models
2. That includes blackbox interpretability models
3. That doesn't mean they're useless

# Additional Material

# Shap

- 2017 paper
- 2018 paper
- github

# Christoph Molnar

- Interpretable Machine Learning
- Twitter

# Terrence Parr

- Intro to permutation importance
- Stratified Partial Dependence Plots (19-Jul-2019)
  - Article
  - Github


T. HANKS