CRITIQUE

# The Mythos of Model Interpretability

Zachary C. Lipton

# Critique 0

- This paper is hard to critique … at first glance

- **Good Points** vs **Cheap Arguments**

# Focus on “4. Discussion”

# Critique 1

- "Linear models are not strictly more interpretable than deep neural networks"
- **Common sense** vs **Extreme example**?
- Which one is easy to understand?
  - Difficult to compare...
- As for the post-hoc explanations for deep learning, how do we know these heuristic explanation methods are trustworthy?

# Critique 2

- ❏ “Claims about interpretability **MUST** be qualified”
  - ❏ Author calls for “a solid problem formulation”
- ➔ It is good, but what is it?
  - ◆ Interpretability is a subjective goal.
- ➔ Is it always possible?
  - ◆ Depends on how you define “solid”
  - ◆ Will it stifle the creativity of researchers?

# Critique 3

- "In some cases, transparency may be at odds with the broader objectives of AI"
- Contradiction?
  - One side: Interpretability is so important
  - Opposite side: Interpretability limits higher performance
- It is a trade-off problem.

# Critique 4

- "Post-hoc Models might provide potentially misleading information"
- Make an analogy to human brain.
- Lack a concrete example of post-hoc model that provides misleading explanations.

# More points...

- Critique the whole machine learning society!
- No clear definition of interpretability
- No answer to when interpretability is important and when is not
- Comparison to the next paper ...
- Difficult Words & Sentences!
    - E.g. Desiderata