# The Mythos of Model Interpretability

Zachary C. Lipton

Presented by:

(Ethan) Yuqiang Heng, Dave Van Veen

# Interpretability: What, Why and How

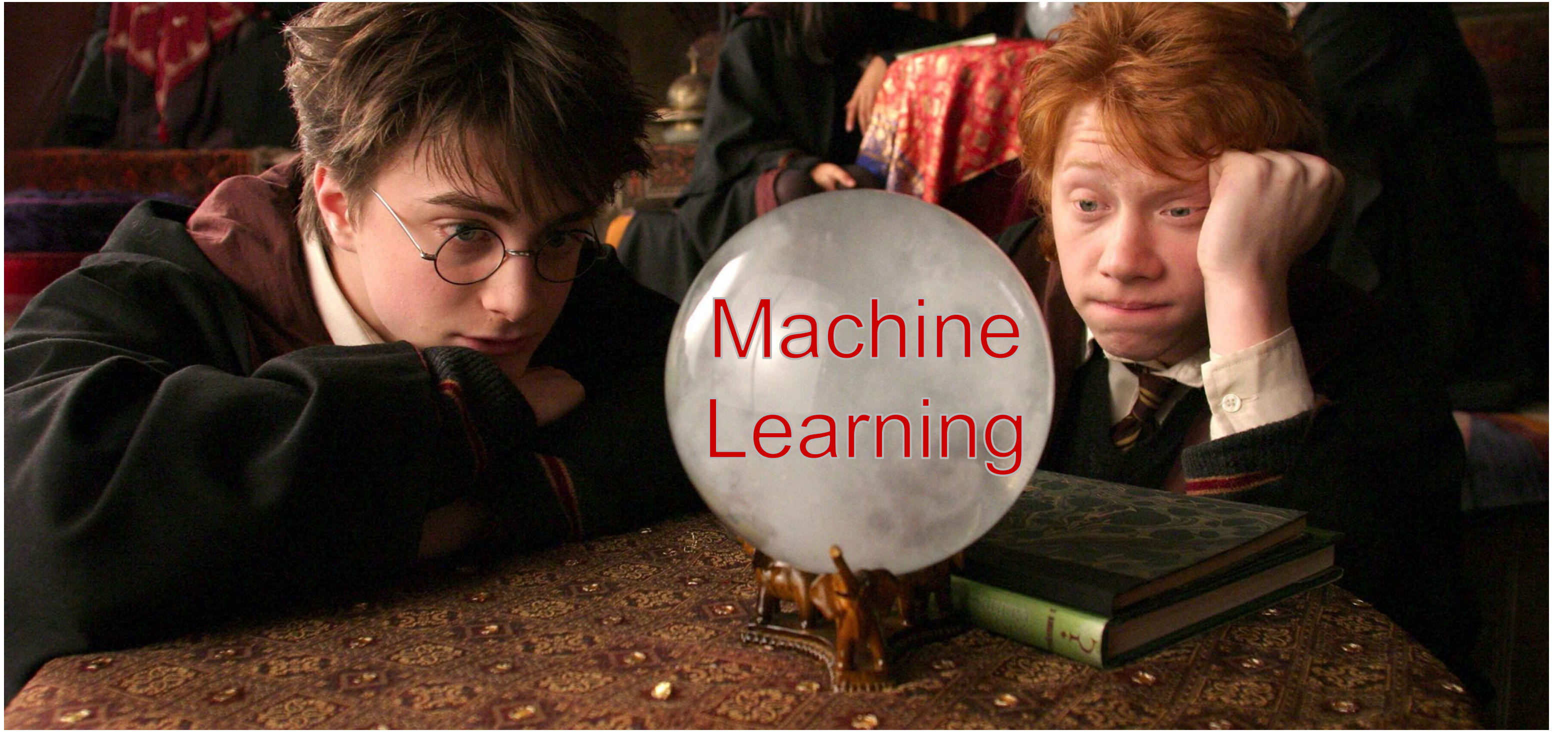


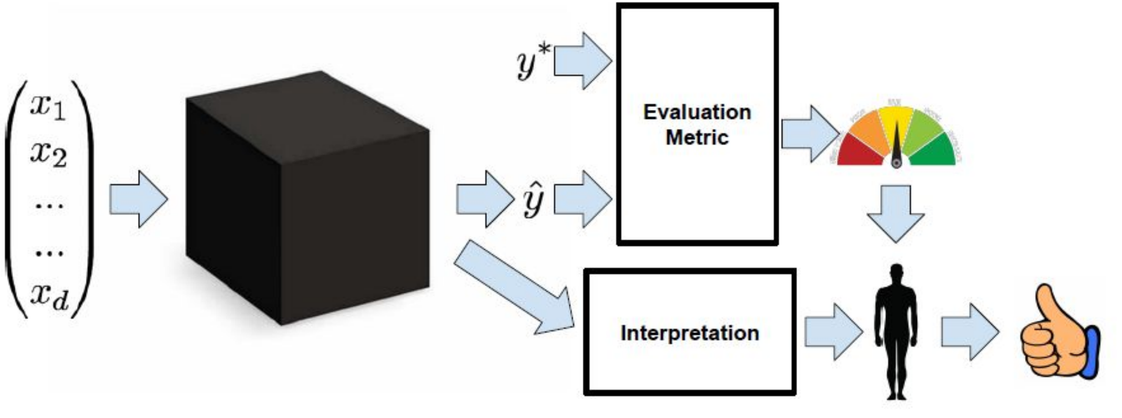

Image from: Harry Potter and the Prisoner of Azkaban (film)

# What is Interpretability?

Interpretable, explainable, intelligible, transparent, understandable?

- Inconsistent notions sometimes used interchangeably

Mismatch between formal objectives of models and real-world costs of deployment settings → Something more than performance !



# Why Interpretability? - Trust

- Trust ≠ Simple confidence that model will perform well
- Trust is subjective, specific requirements depend on person & domain
- Model performs well w.r.t. real objectives and scenarios when training and deployment objectives diverge
- Are we comfortable relinquishing control to model?
    - Does the model make the same mistakes as humans do?

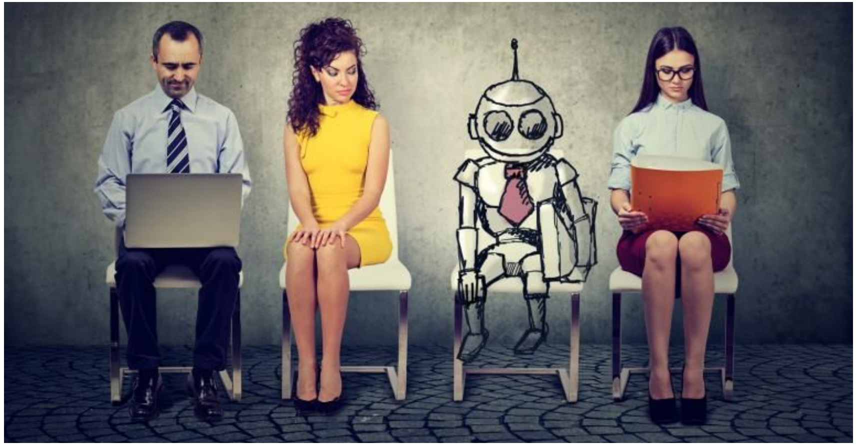
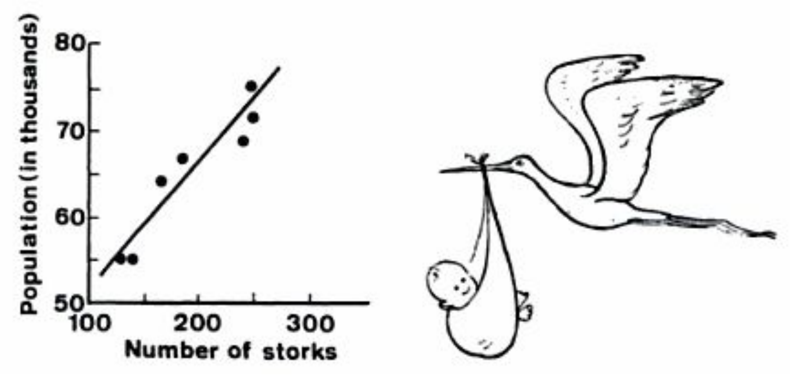


Image from:
www.electronicdesign.com/sites/electronicdesign.com/files/styles/article_featured_retina/public/Imagination_AI_Promo.jpg?itok=Hc1hs2vj

# Why Interpretability? - Causality

- Supervised models learn association
  - Association (correlation) ≠ Causality!
  - Exist unobserved causes for associated variables
- But we hope to infer causal relationships from observational data
  - By interpreting some models (regression trees, Bayesian neural nets) then form hypotheses that can be further tested experimentally




From Richard F. Mould's Introductory Medical Statistics

# Why Interpretability? - Transferability

- Training scenarios and deployment scenarios often diverge

  → **We need models to generalize**

- Often judge generalization error by performance gap between training and testing data
  - But they are often randomly partition examples **from the same distribution**
- Humans have far greater ability to generalize by transferring learned skills to unfamiliar scenarios
- Problems: non-stationary environments; models might even alter the environment

# Transferability Examples

- Pneumonia mortality model assigns less risk to asthma patients [Caruana et al. 2015]
    - Asthma patients receive more aggressive treatments
    - But if the model were deployed to aid in triage, asthma patients would then receive less aggressive treatment, invalidating the model
- FICO trains creit models using logistic regression [Fair Isaac Corporation, 2011]
    - Attributes susceptible to manipulation (debt ratio, total number of accounts, etc)
    - FICO themselves provide guides to improve credit rating that do not fundamentally change one's ability to pay off debt → **invalidate predictive power**
- Adversarial Attacks: CNNs are sensitive to human-imperceptible perturbation
    - We want models not to make mistakes that humans won't make

# Why Interpretability? - Informativeness

- Models are meant for prediction, but we often use outputs to take actions
  - Need to provide information to assist human decision makers
  - ML objective to reduce error while real-world purpose is to provide useful information
- Does not always need complete understanding of models' inner workings
  - Locally explainable (LIME, Anchors, Pertinent Negatives)
  - Prototypes that point to similar cases in support of diagnostic decisions (DL for Case-Based Reasoning through Prototypes)


Image from [Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance]

# Why Interpretability?
# - Fair & Ethical Decision-Making

- Models mediates more aspects of our lives: credit, curate news, filtering job applicants, recidivism etc.

  → **Need interpretability to assess ethicality of algorithmic decisions**

- Conventional metrics (accuracy, AUC) do not guarantee against discrimination

# Fair & Ethical Decision-Making

EU proposes:

1. Right to explanation for people affected by algorithmic decision

- What form of explanation? How such explanation should be proven correct?

2. Algorithmic decisions should be contestable

- Present clear reasoning based on falsifiable propositions
- Offer natural way of contesting propositions and modifying decisions appropriately (Actionable Recourse in Linear Classification)

# Properties of Interpretable Models

Two categories

Transparency: *How does the model work?*

Post-hoc Explanations: *What else can this model tell me?*

**Properties**

- **Transparency**
  - Simulatability
  - Decomposability
  - Algorithmic
- **Explanations**
  - Text
  - Visualization
  - Local
  - By example

# Transparency

At the level of the...

... entire model (simulatability)

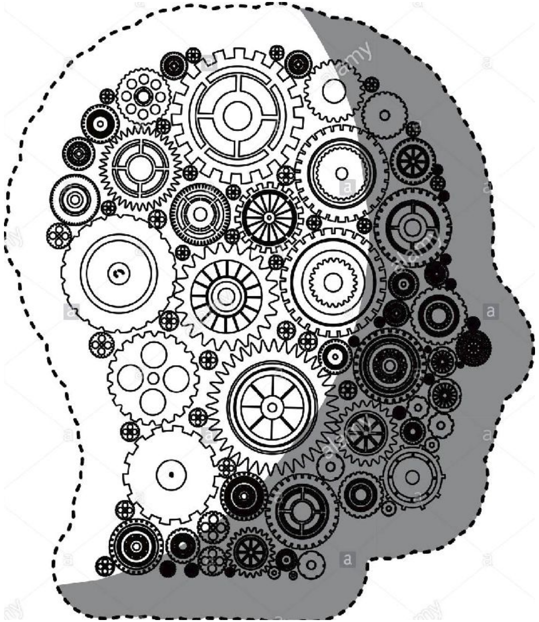
... individual components (decomposability)

... training algorithm (algorithmic transparency)

# Simulatability

**Properties**
- **Transparency**
  - **Simulatability**
  - Decomposability
  - Algorithmic
- Explanations

- Can a person comprehend the entire model?
- Examples:
  - For linear models, sparse > dense
  - LIME: model complexity low enough to be interpretable
- NNs might be more interpretable than…
  - High-dimensional linear models
  - Complex rule lists
  - Deep decision trees



# Decomposability

Properties
- **Transparency**
  - Simulatability
  - **Decomposability**
  - Algorithmic
- Explanations

Idea: Each component admits an intuitive explanation

Examples:

Generative additive models [Lou et al., 2012]

Each node of decision tree

Regression parameters**

Note: Disqualifies feature engineering

**Properties**
- **Transparency**
  - Simulatability
  - Decomposability
  - **Algorithmic**
- Explanations

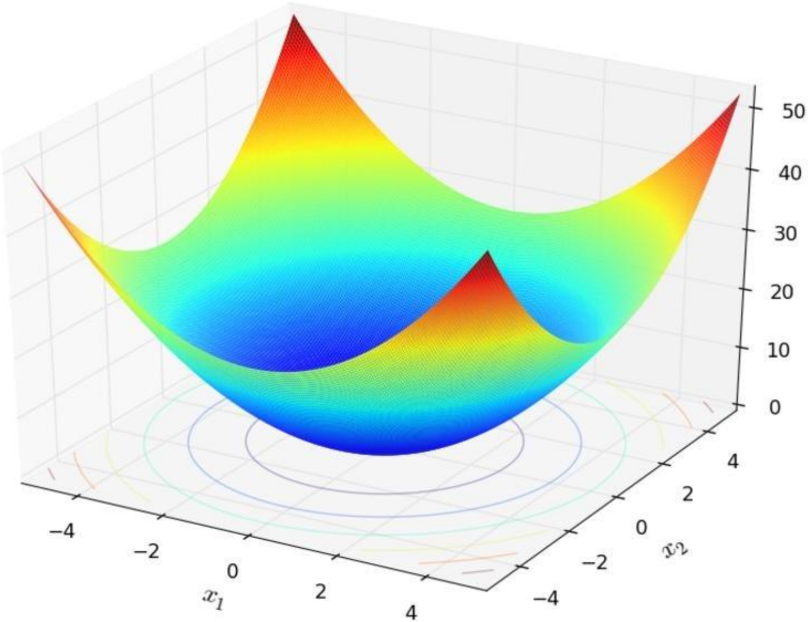
# Algorithmic Transparency

Linear models - shape of error surface is understood

NNs

Error surface - ??

Optimization procedures - ??




Note: By these notions, humans aren't transparent!

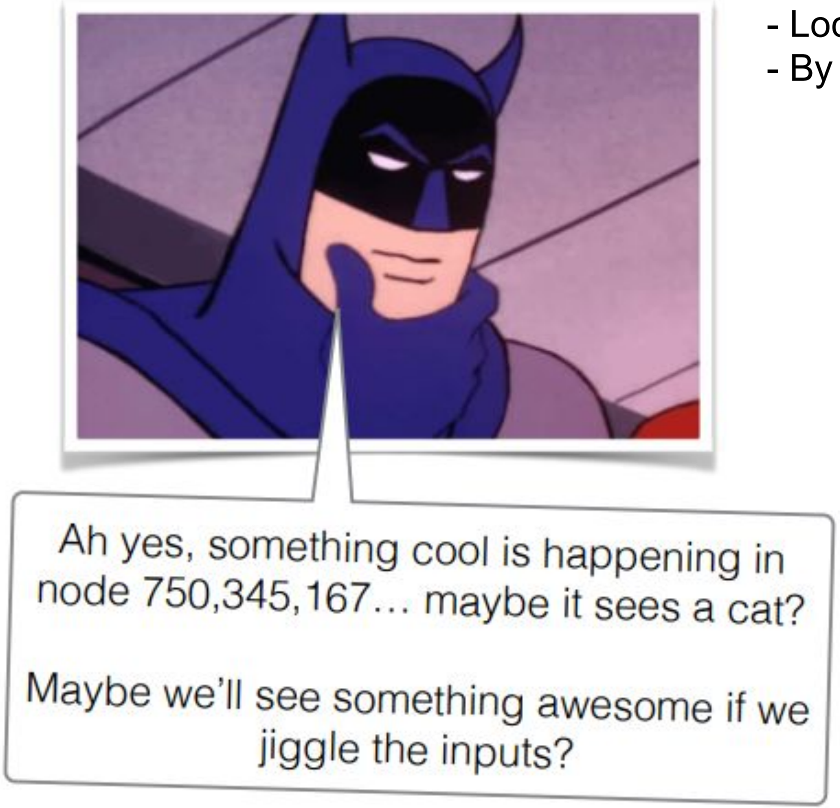
# Post-hoc Explanations

**Properties**
- Transparency
- **Explanations**
  - Text
  - Visualization
  - Local
  - By example

Extracting useful information

Potentially from a black box

Human interpretability

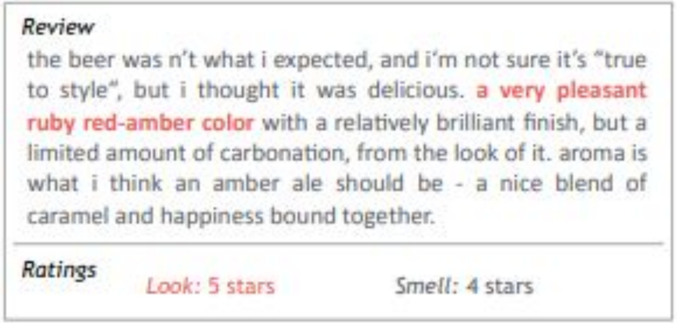


# Text Explanations

**Properties**

- Transparency
- **Explanations**
  - **Text**
  - Visualization
  - Local
  - By example

<u>Idea</u>: Train another model to provide explanations via text




<u>Examples</u>:

Beer ratings [Tao et al., 2016]

Latent factors for product recommendations [McAuley et al., 2013]

Text (product reviews) serve as labels for the latent dimension

Will a user enjoy Harry Potter?

# Visualization

**Properties**
- Transparency
- **Explanations**
  - Text
  - **Visualization**
  - Local
  - By example

Idea: Determine visually what a model has learned

Examples:

t-SNE: project high-dimensional data onto 2D or 3D [van der Maaten, 2008]

Perturb various inputs, compare output images [Mordvintsev, 2015]

Recover original image from CNN representation [Mahendran, 2015]

# Local Explanations

**Properties**
- Transparency
- **Explanations**
  - Text
  - Visualization
  - **Local**
  - By example

Example: LIME explains decision near a point

Notes:

Local region small → explanations don't transfer

Even one pixel - saliency maps [Mahendran et al., 2015]

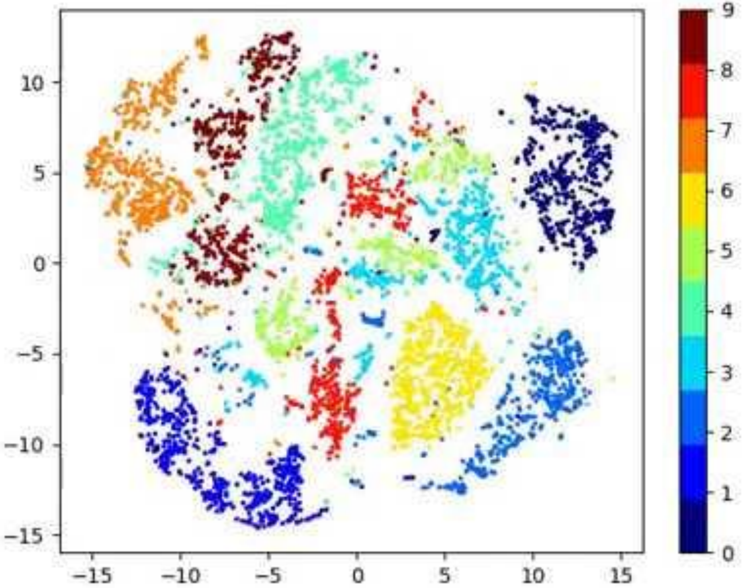
Linear models - global relationship

# Explanation by Example

**Properties**
- Transparency
- **Explanations**
  - Text
  - Visualization
  - Local
  - **By example**

Idea: Along with prediction, report similar training examples

Example: Perform k-NN in latent space
[Caruana et al., 1999 + others]



# Takeaways

- Linear models not strictly more interpretable than NNs
  - Depends on which notion of transparency
  - Features often heavily engineered
- Interpretability claims require specific definitions - be precise!
- Be careful giving up predictive power
- Be careful with post-hoc explanations
  - Humans aren't great at this

ML researchers: consider impact of your work