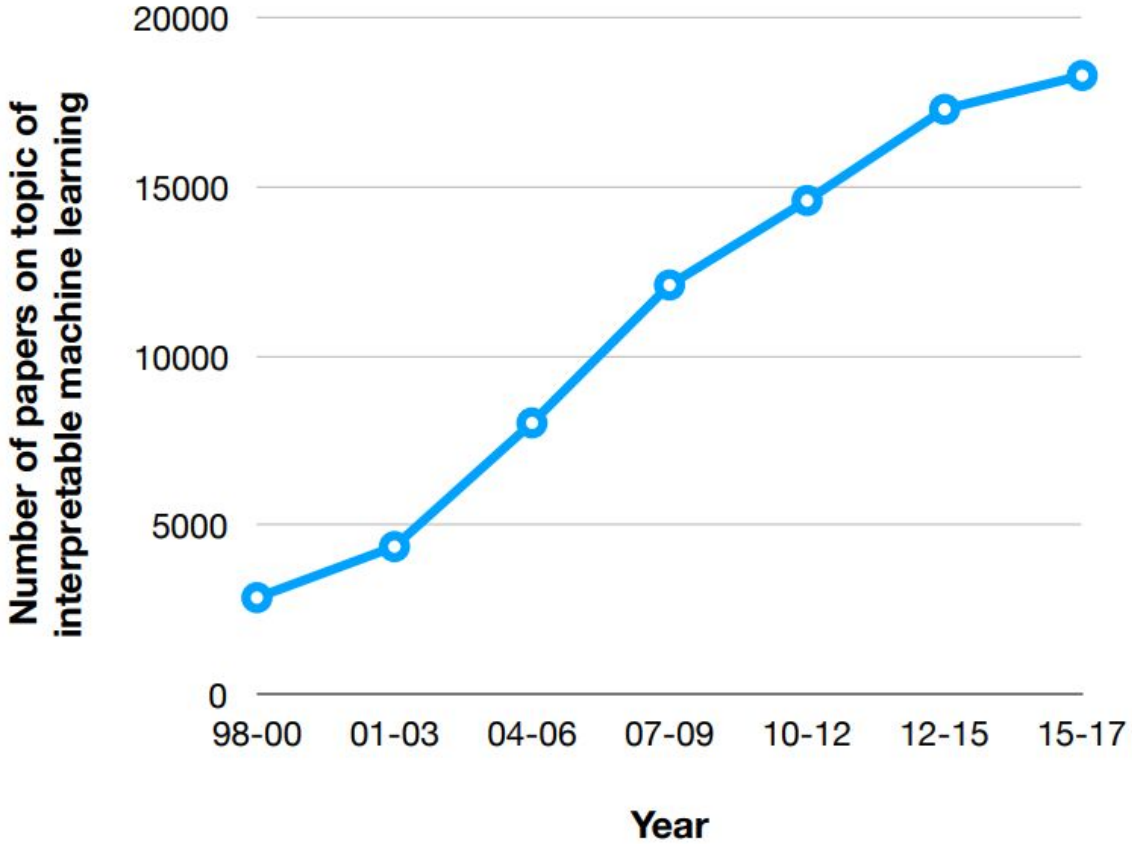
Towards A Rigorous Science of Interpretable Machine Learning

[Defining and Evaluating Interpretability]

By: Finale Doshi-Velez and Been Kim

Presented by: Harrison Keane, Kurtis David, Jun Min Noh

ML community is responding



Why Should We Evaluate Interpretability?

- Multiple/ambiguous definitions of interpretability-

  “You know it when you see it”

- Interpretability is not quantifiable like other performance metrics

- EU 2018 mandate on algorithms

Outline

Purpose: to define and set rigorous evaluation of interpretability

1. What is interpretability?
2. Need for interpretability
3. Taxonomy for evaluation of interpretability
4. Approaches to answer open problems of interpretability

Defining Interpretability

ability to explain or to present in understandable terms to a human

Which can confirm:

- Fairness, unbiasedness
- Privacy
- Reliability, robustness
- Causality
- Safety
- Trusted/Usable

Need For Interpretability

- Need for interpretability rises from fundamental incompleteness

  Incompleteness ≠ uncertainty

  Incompleteness from

  - Scientific understanding
  - Ethics
  - Mismatched objectives
  - Safety

Taxonomy of Evaluating Interpretability

1. Application Grounded

2. Human Grounded

3. Functionally Grounded

Application Grounded Evaluation

- Context:
  - Real applications (e.g. diagnosing patients)
  - Assisting domain experts

- Baseline Experiment:
  - How do model explanations compare to human-produced explanations?

Human Grounded Evaluation

- Context:
  - Simplified task for explainability
  - Assisting lay humans

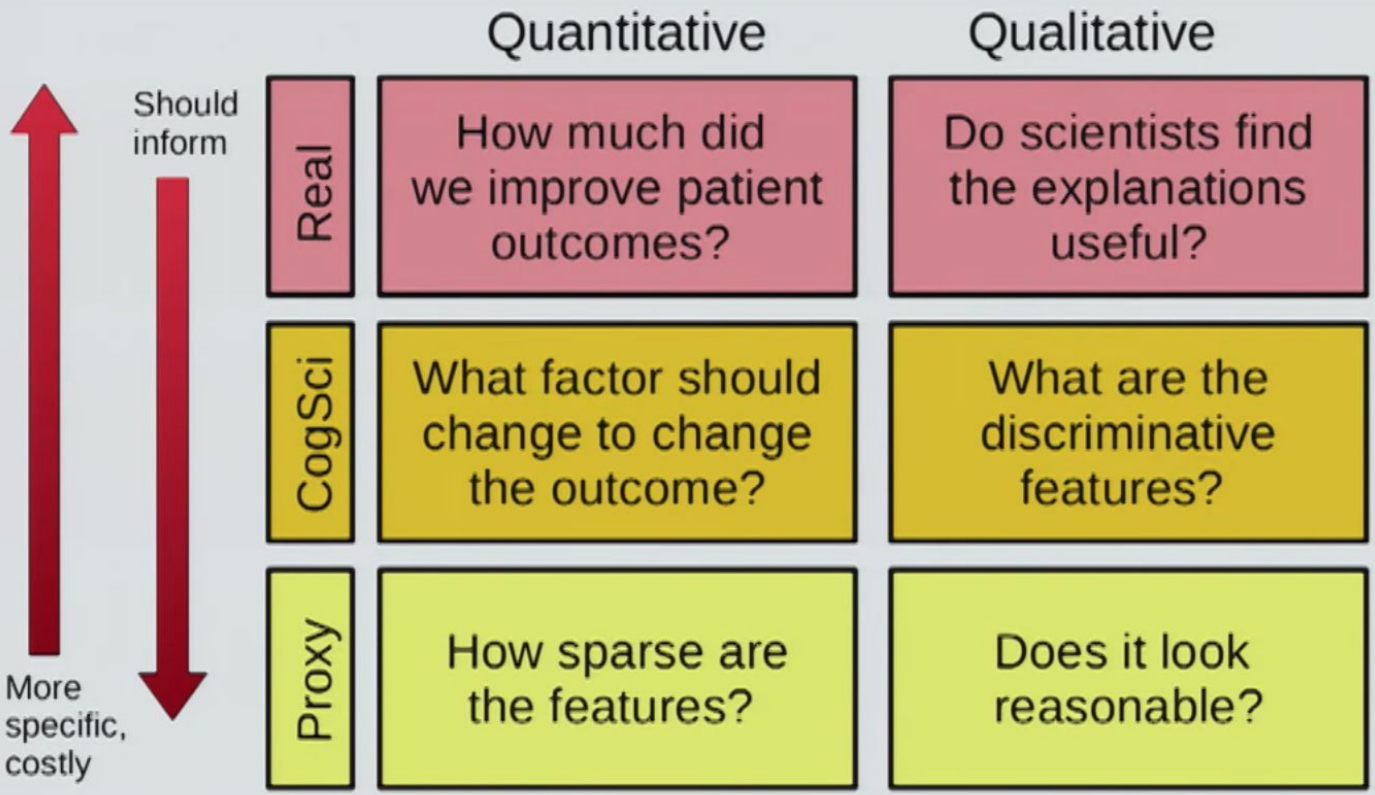
- Possible experiments:
  - Which explanation is better? (binary forced choice)
  - Given input and explanation, simulate model output (forward simulation)
  - What input should be changed to change output (counterfactual)

Functionally Grounded Evaluation

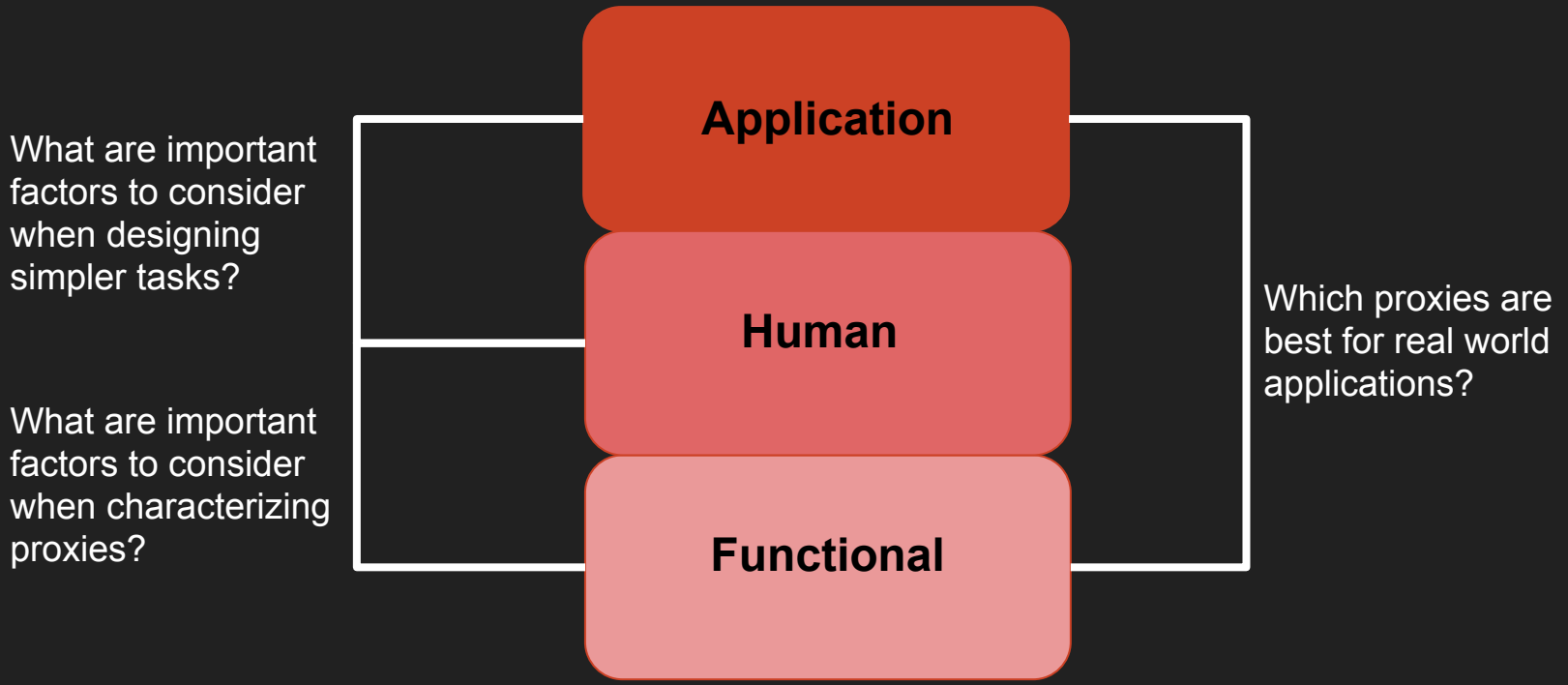
- Context:
  - Proxy for explainability
  - No humans, comparisons through formal definitions

- Possible Experiments:
  - Which model is more sparse?
  - Which *interpretable* model has better performance?

A Spectrum for Evaluation



Evaluations Should Inform Each Other



How to Approach Open Problems in Interpretability

Mass effort of data collection to create matrix:

COLUMNS:
Specific Methods
(i.e. Decision Tree of Depth < 4)

ROWS
Specific real-word tasks
(i.e. “assisting doctors in identifying pneumonia patients under 30 in US”)

NEED: *open repositories that contain problems corresponding the real world tasks in which human input is required*


METHOD FACTORS
(Hypothesis #2)
METHODS
TASKS
TASK FACTORS
(Hypothesis #1)

Hypothesis 1: Task Related Latent Dimensions

Scope: global/local interpretability

Area/Severity of Incompleteness

Time Constraint

User Preference/Expertise

Hypothesis 2: Method Related Latent Dimensions

cognitive chunk: basic unit for explanation

Miller's Law: 7 ± 2 chunks in working memory

Factors involving cognitive chunks:

- How many chunks in an explanation?
- Structure/compositionality of cognitive chunks (i.e. A->B->C)
- Monotonicity/Interactions between cognitive chunks
- Human understanding of uncertainty and stochasticity

NEED FURTHER WORK IN COGNITIVE SCIENCE TO IMPROVE INTERPRETABILITY

Summary/Recommendation to Researchers

To push the field further, works of research should describe:

- The incompleteness of their problem that triggers need for explanation
- Which levels of evaluation are being explored
- Any task-related factors (scope, area, time budget)
- Any method-related factors (cognitive chunks, expertise)