Towards A Rigorous Science of Interpretable Machine Learning

A paper by **Finale Doshi-Velez and Been Kim**

March 17 2021

Agenda

1. Motivation
2. Problem Statement
3. Contribution
4. Critique

Authors

Finale Doshi-Velez
Harvard University, MA

Been Kim
Google Brain, CA

Photo (left): http://finale.seas.harvard.edu
Photo (right): https://slideslive.com/38918114

Motivation

1

Up next:

2.Problem Statement

3.Contribution

4.Critique

Optimized … for what?

Unquantifiable criteria

- ML systems outperform humans.
- Expected Performance is not always entirely defined. What about…
  - **Safety?**
  - **Nondiscrimination?**
  - **Avoiding technical debt?**
  - **Providing the “right to explanation”?**

- ML systems guide us to make decisions or make decisions directly
- **Problem: these criteria can't be completely quantified.**
  - Is my set of unit tests sufficient?
  - What if I deny someone credit due to poor training breadth?
- Want to understand and verify the criteria used.

The Right to Explanation

(Aside)

- Not only a wish, but also a requirement!
  - e.g. EU requirement since 2018 in the GDPR
  - Also relevant for Swiss companies doing business with EU parties
- Reference: https://www2.deloitte.com/ch/en/pages/risk/articles/gdpr-consequences-for-swiss-businesses.html

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

(OJ L 119, 4.5.2016, p. 1)

Article 22

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

A lack of Consensus

How to evaluate interpretability?

- **Current interpretability evaluation:**
  - Category 1: in the context of an application (useful → somehow interpretable)
  - Category 2: via a quantifiable proxy (optimize within a class).
- "*You'll know [a good explanation] when you see it*": lack of rigor.
- **Pros and cons of the current way:**
  - \+ Somewhat reasonable: we have face-validity (it "looks like" it does the job to humans)
  - \- Many questions can NOT be answered!
- Need to make these notions evidence-based and formal.

https://images.app.goo.gl/vqxLxEfB9o1mVG9d9

Problem Statement

2

Up next:

3.Contribution

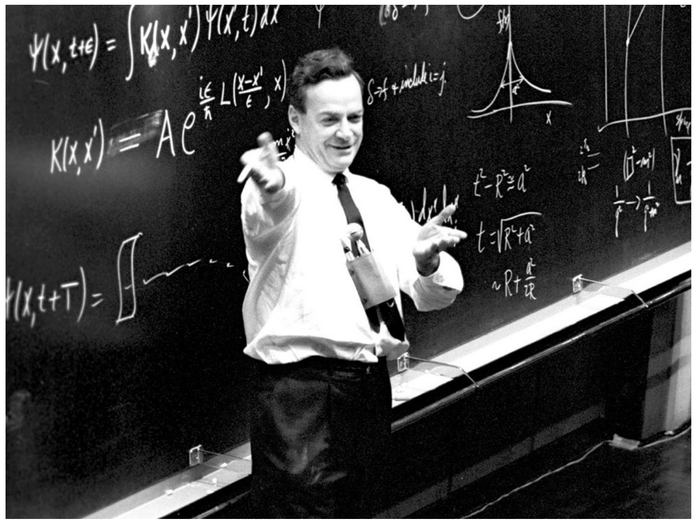
4.Critique

What is Interpretability?

It's *not* reliability or fairness (1/2)

p	q	~p	~p v q
T	T	F	T
T	F	F	F
F	T	T	T
F	F	T	T

- Merriam-Webster: "Able to be explained or presented in understandable terms"
- **Authors: "Ability to explain or present in understandable terms *to a human*"**
- "Explanations are **the currency in which we exchanged beliefs**" [Lombrozo, 2006]
  - **Explicit** explanation: "deductive-nomological" (logical proofs)
  - **Implicit** explanation: provide some sense of mechanism (broader definition)
- Need *data-driven* ways to evaluate and define explanations (and thus, interpretability)



https://images.app.goo.gl/fFpECMtYzUwr41da7
https://images.app.goo.gl/Hdwoo3jouNH7v3SG7

What is Interpretability?

It's *not* reliability or fairness (2/2)

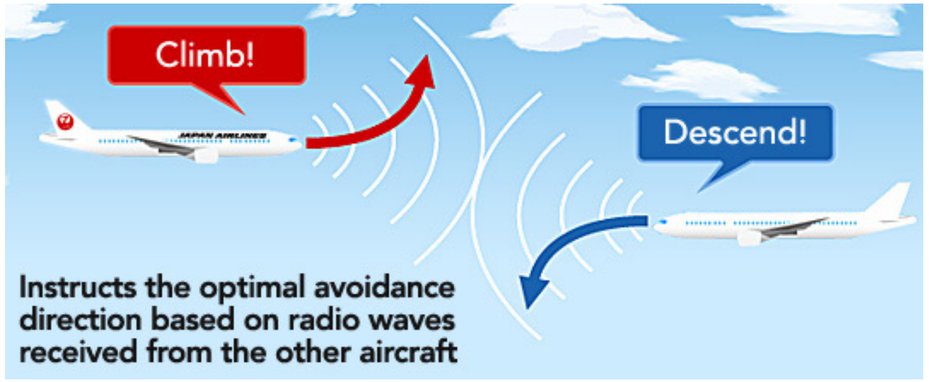
- Confirming other important desiderata of ML systems (i.e., many other criteria we want to optimize).
  - **Fairness, Unbiasedness**
  - **Privacy**
  - **Reliability, robustness**
  - **Causality**
  - **Usability**
  - **Trustworthiness**

- Definitions remain elusive
  - Highlighting an incompleteness
  - Not all explanations are good! Need to be exigent enough.

https://images.app.goo.gl/qFcgPuBuYcbkVg4Y8

Why do we need Interpretability?

Various scenarios

- **Not all ML systems require it** (e.g. aircraft collision avoidance systems…)
- 

- Incompleteness in the problem formalization → need for interpretability
  - ≠ Uncertainty!
  - Unknowns ≠ Incompleteness
- 5 scenarios:
  - **Scientific understanding**
  - **Safety**
  - **Ethics**
  - **Mismatched objectives**
    - e.g. medication adherence
  - **Multi-objective trade-offs**
- Explanations allow us to see the effects of formalization gaps

https://images.app.goo.gl/ejxWrdvLiWpV7DE6A

Contribution

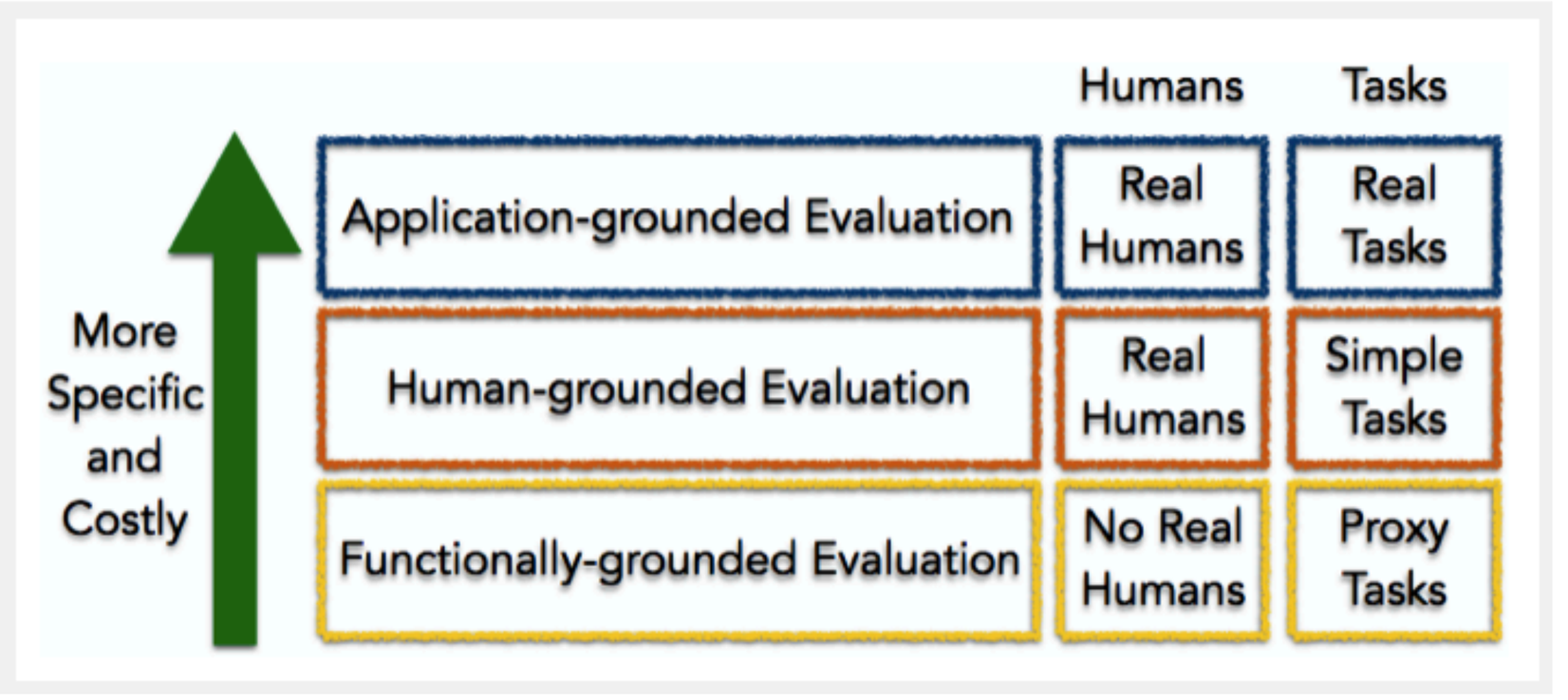
Up next:

4.Critique

Achieving Interpretability

A Taxonomy of Interpretability Evaluation

- The evaluation should match the claimed contribution.
  - **Applied work**: evaluation should demonstrate success (e.g. game-playing agent should win)
  - **Core methods**: should demonstrate generalizability via careful evaluation
    - Synthetic + standard benchmarks
- Human evaluation is essential, but not easy
  - Confounding factors, time, resources, …


Humans
Tasks
More Specific and Costly
Application-grounded Evaluation
Real Humans
Real Tasks
Human-grounded Evaluation
Real Humans
Simple Tasks
Functionally-grounded Evaluation
No Real Humans
Proxy Tasks

Authors

Application-Grounded Evaluation

Real humans, real tasks

- **Human experiments within a real application.**
- User studies similar to UI evaluations.
- Evaluate **w.r.t. the end-task** with domain experts.
- Task could be exact or partial.
- Time constraints should be considered to adapt task.
- Baseline: how humans assist other humans. Direct testing (+). But high standards of experimental design (-).

https://www.swissinfo.ch/eng/medical-training_what-can-be-done-to-boost-doctors-numbers/42339434

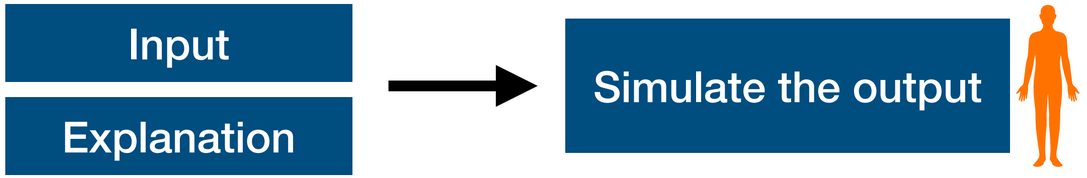
Human-Grounded Evaluation

Real humans, simplified tasks

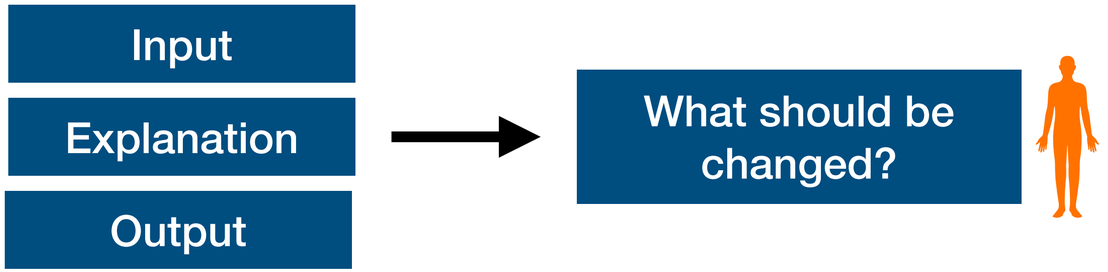
- **Make human experiments simpler, but maintain the essence of the target.**
- Easier to attract participants. Invite *lay users* for less financial compensation.
- Test more general notions of the quality of an explanation.
- *Only* evaluate the explanation quality. Nothing else.

- Binary forced choice

- Forward simulation/prediction



- Counterfactual simulation

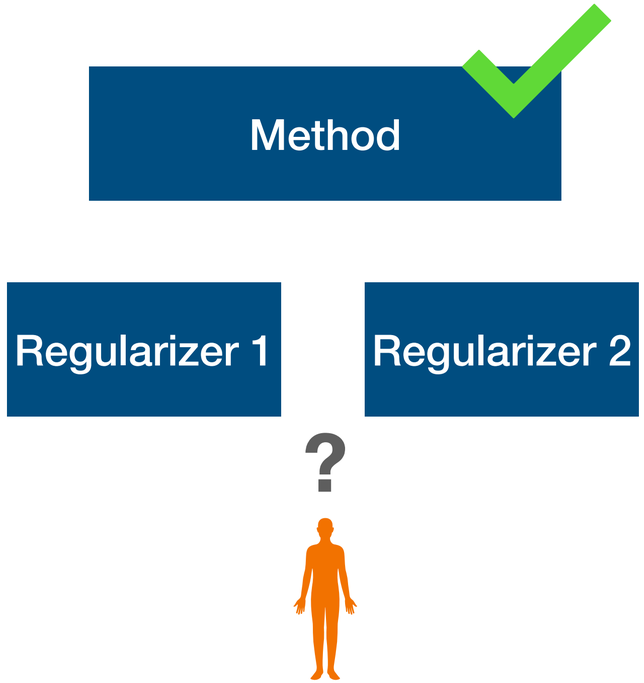


Functionally-Grounded Evaluation

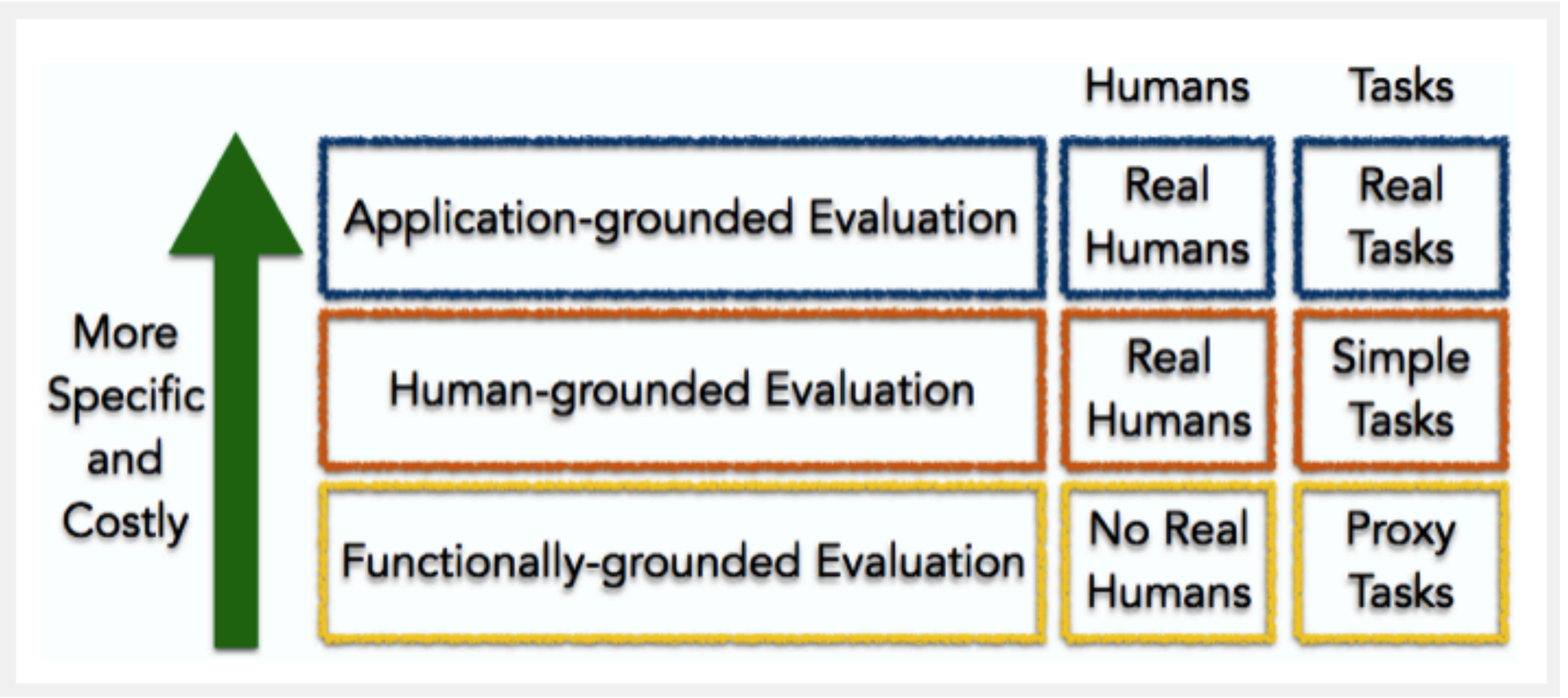
No humans, proxy tasks

- Use a formal definition of interpretability as a *proxy* for expl. quality.
- Appealing: low cost, fast, no approval required.
- Appropriate for models that have *already* been validated or when a method is not yet mature.
- **Challenge**: *what proxies* should one use?
  - Then, just optimize.

Agreed upon fact: "Linear models are interpretable"



Which of LASSO or Ridge brings the most/best sparsity?

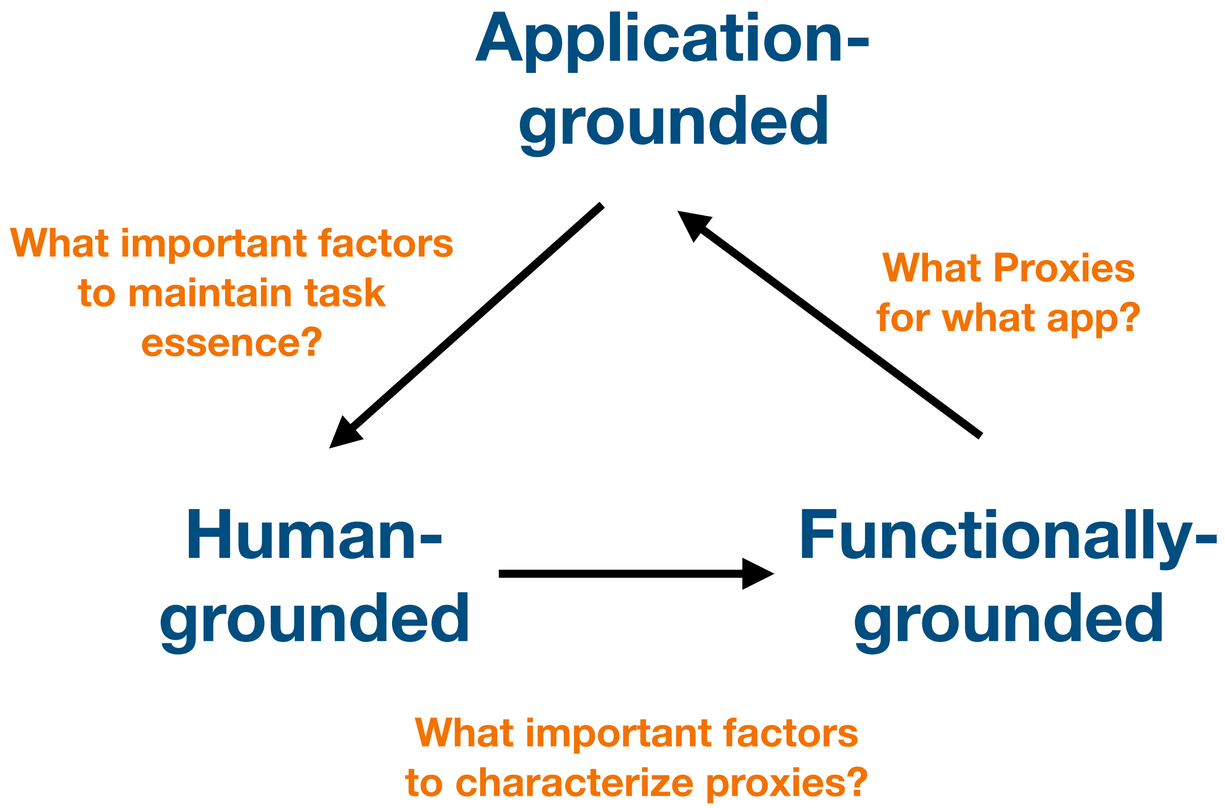

Humans
Tasks
More Specific and Costly
Application-grounded Evaluation
Real Humans
Real Tasks
Human-grounded Evaluation
Real Humans
Simple Tasks
Functionally-grounded Evaluation
No Real Humans
Proxy Tasks

Authors

Open Problems

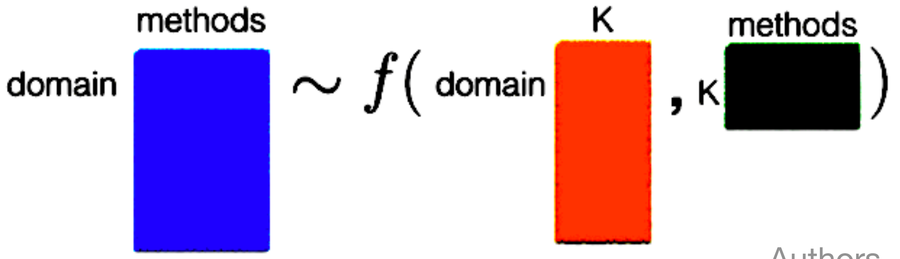
Evaluation types inform each other

- Factors that capture essential real world task needs should inform about what kinds of simplified tasks we perform.
- Performance w.r.t proxies should reflect their performance in real-world settings.



Data-Driven Approach

Discovering factors of Interpretability



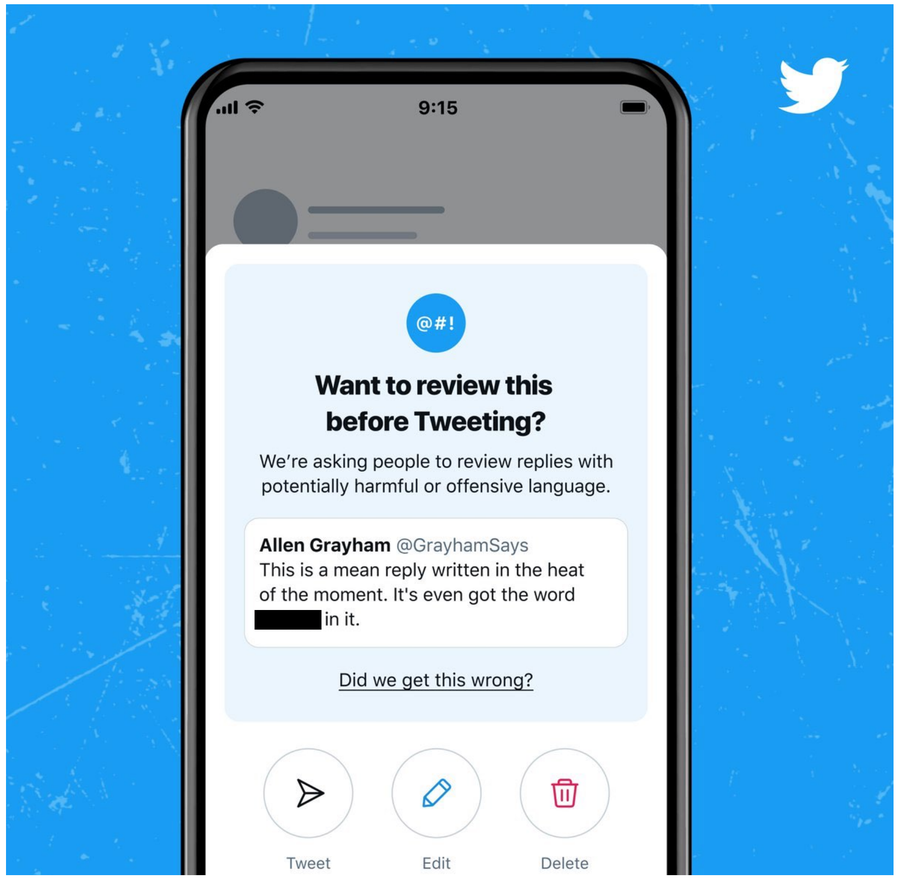
Authors

- Imagine a matrix: entries are the performance of the method on the end task.
- Identify latent dimensions that represent factors that are important to interpretability.
  - e.g. matrix factorization to embed tasks and methods.
- **Challenge**: creating this matrix! Need to quantify performance with domain experts who have agreed. Each cell takes hours to come up with.
- Lot of open repositories for ML problems: why not a repository for tasks with human input? (crowdsourcing)
- **What could latent dimensions correspond to? (Authors have their theory …)**

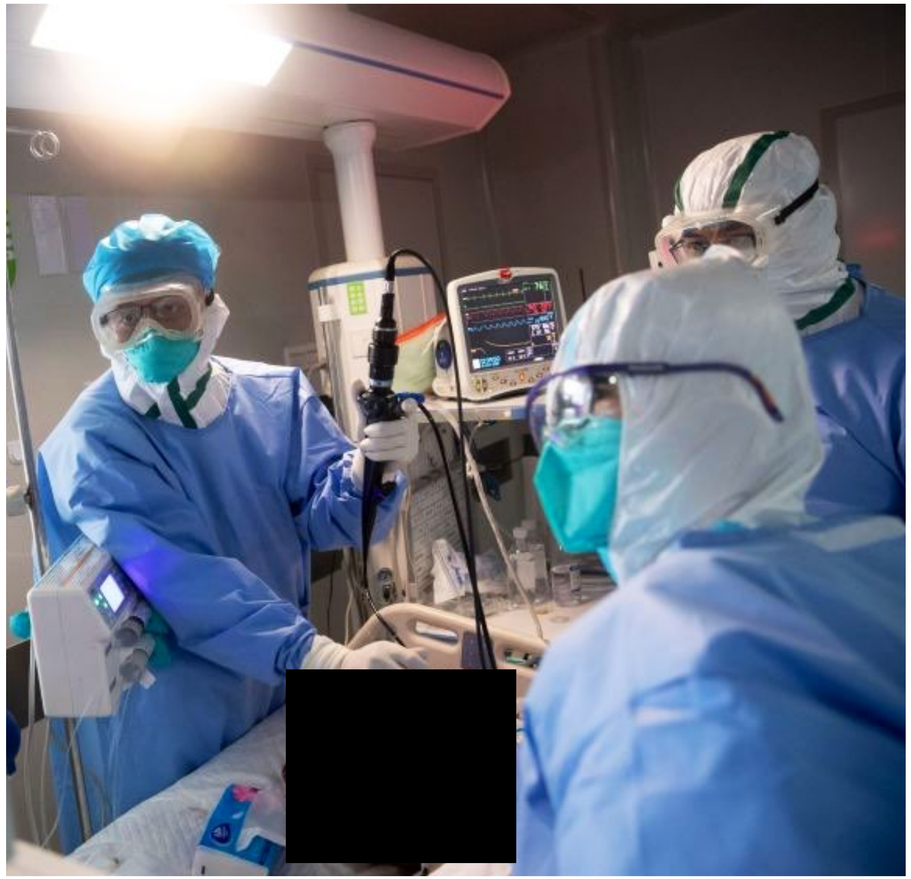
Task-related latent dimensions

A hypothesis (1/2)

- Disparate-seeming apps may share common categories.
- Example: both make a decision fast about a specific case.



Source: Official twitter account



Source: https://images.app.goo.gl/By2B9uFvUwC35hGK8

Task-related latent dimensions

A hypothesis (2/2)

- What might make ***tasks*** similar in their explanation needs?

Global interp. (Deep understanding) >< **Local interp.** (Reasons for a decision)

Area / Severity of Incompleteness

Time Constraints (bedside decision or time-sensitive decision)

Nature of user expertise (what kind of “cognitive chunks” they have)

Method-related latent dimensions

A hypothesis

- Disparate ***methods*** may also share common qualities that correlate to their utility as explanation.
- What might make ***methods*** similar in their explanation needs?
- **Cognitive chunk** :=
  “A basic unit of explanation”

Form of cognitive chunks (what are they? Derived features? Prototypes?)

Number of cognitive chunks (How many? How to handle prototypes vs features?)

Monotonicity and interactions between chunks (What is natural?)

Level of Compositionality (Are the chunks organized? Rules? Hierarchies?)

Uncertainty and stochasticity (How well do people understand them?)

Author Recommendations

Three main takeaways (1/2)

- Groundwork was laid for a process to rigorously define and evaluate interpretability.
- Many open questions in creating formal links, human understanding etc…
- **Takeaway 1: The claim of the research should match the type of the evaluation.**
- **Takeaway 2: We should categorize our applications and methods with a common taxonomy. Create a shared language!**

Author Recommendations

Three main takeaways (2/2)

- **Takeaway 3: Every publication should contain the following elements.**
  - How is the problem formulation incomplete?
  - At what level is the evaluation being performed?
  - What are the task-related relevant factors?
  - What are the method-related relevant factors being explored?

Critique

4

Critique

My personal opinion on this paper

- Paper was well-structured and well-motivated/situated.
- A good list of recommendations for inexperienced explanation designers.
- Paper paves the way for a taxonomy for applications and methods.
- Asks all the right questions, but provides few examples.
  - Focuses on facts, definitions etc. and builds upon them in a constructive way, with creativity.
  - Leaves a lot of the work to researchers and app designers, but provides them with a more structured way to do that work.

Thank you for your attention