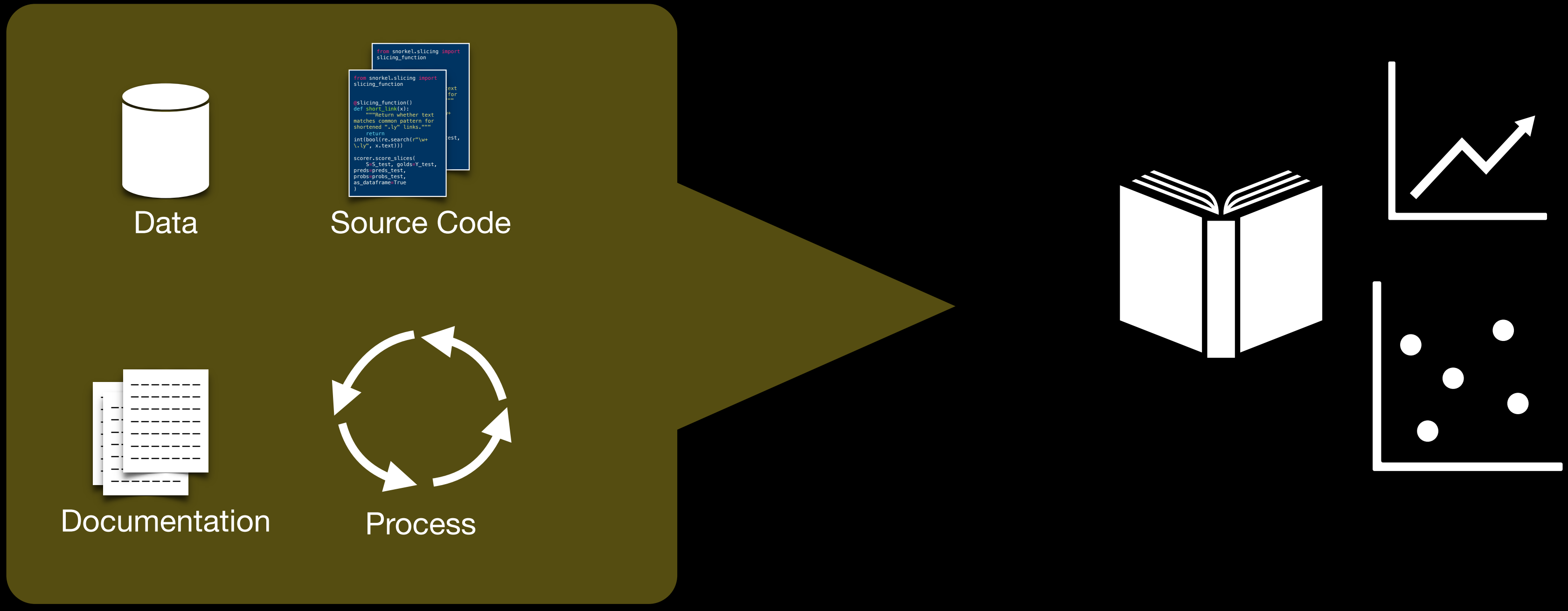
Transparency and Explainability

Jin Guo **19th Nov**

Transparency: AI systems should be designed and implemented in such a way that oversight of their operations are possible.



Data

Source Code

```
from snorkel.slicing import
slicing_function


@slicing_function()
def short_link(x):
    """Return whether text
matches common pattern for
shortened ".ly" links."""
    return
int(bool(re.search(r"\w+
\.ly", x.text)))

scorer.score_slices(
    S=S_test, golds=Y_test,
preds=preds_test,
probs=probs_test,
as_dataframe=True
)
```

Documentation

Process

Explainability: translate the technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation.

Activity

- Consider the system you chose for your assignments:
    - When and where do the stakeholders of the system have the need to understand the AI? Why? (consider only one type of stakeholder in this activity)
    - How the need can be addressed?
    - List three concrete questions the stakeholder might ask?
    - What should the explanation look like to the questions you gathered?

Is the question from

Data Scientist? ML Engineers?

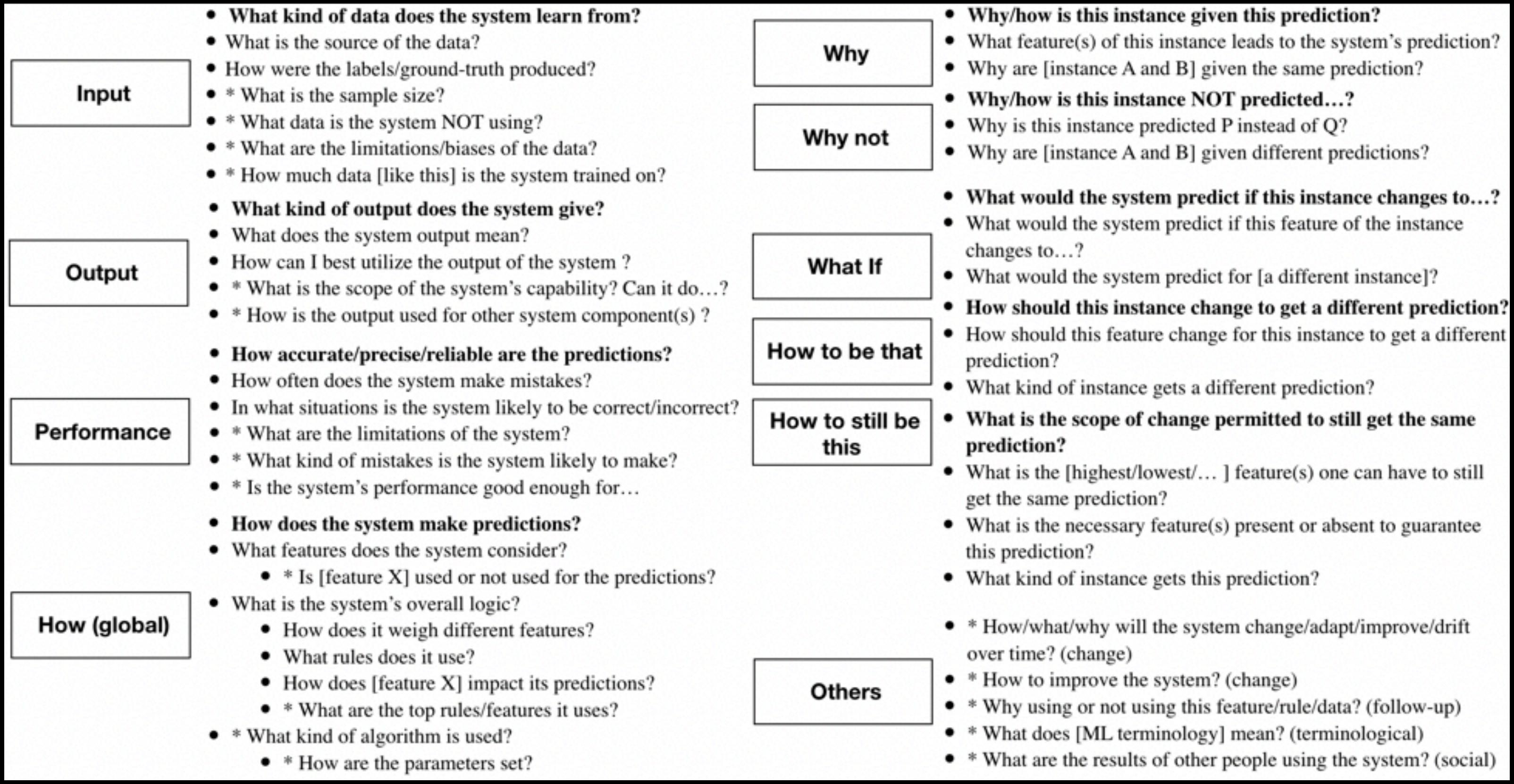
End Users? Product Managers? Auditors?

Is the question asking

How What if Why/Why not

How to be that / How to still be that

XAI Question Bank



Input

- **What kind of data does the system learn from?**
- What is the source of the data?
- How were the labels/ground-truth produced?
- * What is the sample size?
- * What data is the system NOT using?
- * What are the limitations/biases of the data?
- * How much data [like this] is the system trained on?

Output

- **What kind of output does the system give?**
- What does the system output mean?
- How can I best utilize the output of the system ?
- * What is the scope of the system's capability? Can it do…?
- * How is the output used for other system component(s) ?

Performance

- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/incorrect?
- * What are the limitations of the system?
- * What kind of mistakes is the system likely to make?
- * Is the system's performance good enough for…

How (global)

- **How does the system make predictions?**
- What features does the system consider?
    - * Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
    - How does it weigh different features?
    - What rules does it use?
    - How does [feature X] impact its predictions?
    - * What are the top rules/features it uses?
- * What kind of algorithm is used?
    - * How are the parameters set?

Why

- **Why/how is this instance given this prediction?**
- What feature(s) of this instance leads to the system's prediction?
- Why are [instance A and B] given the same prediction?

Why not

- **Why/how is this instance NOT predicted…?**
- Why is this instance predicted P instead of Q?
- Why are [instance A and B] given different predictions?

What If

- **What would the system predict if this instance changes to…?**
- What would the system predict if this feature of the instance changes to…?
- What would the system predict for [a different instance]?

How to be that

- **How should this instance change to get a different prediction?**
- How should this feature change for this instance to get a different prediction?
- What kind of instance gets a different prediction?

How to still be this

- **What is the scope of change permitted to still get the same prediction?**
- What is the [highest/lowest/…] feature(s) one can have to still get the same prediction?
- What is the necessary feature(s) present or absent to guarantee this prediction?
- What kind of instance gets this prediction?

Others

- * How/what/why will the system change/adapt/improve/drift over time? (change)
- * How to improve the system? (change)
- * Why using or not using this feature/rule/data? (follow-up)
- * What does [ML terminology] mean? (terminological)
- * What are the results of other people using the system? (social)

Liao, Q. Vera, Daniel Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-15. 2020.

What makes a good explanation?

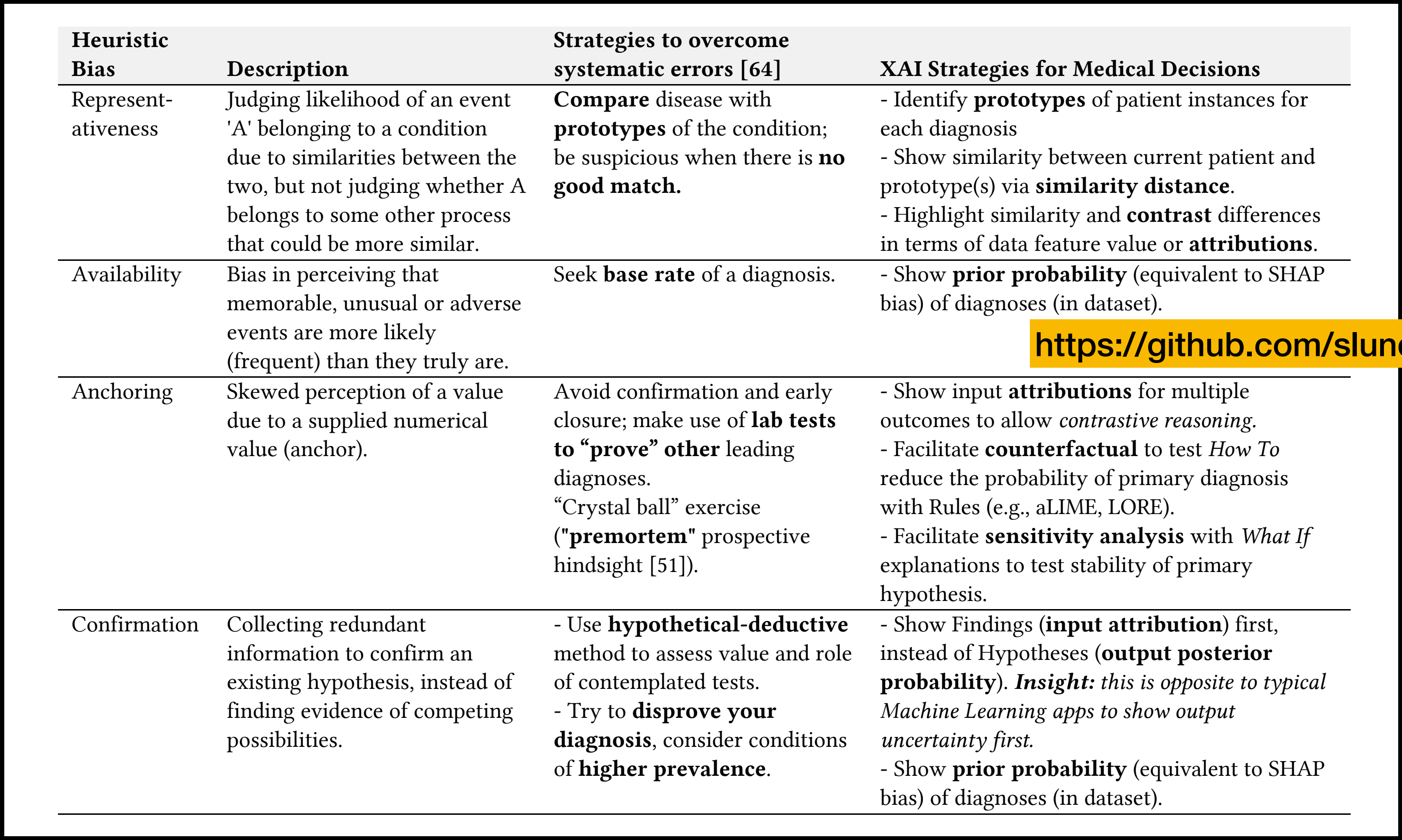
Human Explanation

- Contrastive
- Selective
- Interactive

To what extent are they applicable to the questions you listed in the activity? Why

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI."
In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279-288. 2019.

Support Human Reasoning



Heuristic Bias	Description	Strategies to overcome systematic errors [64]	XAI Strategies for Medical Decisions
Representativeness	Judging likelihood of an event 'A' belonging to a condition due to similarities between the two, but not judging whether A belongs to some other process that could be more similar.	**Compare** disease with **prototypes** of the condition; be suspicious when there is **no good match.**	- Identify **prototypes** of patient instances for each diagnosis - Show similarity between current patient and prototype(s) via **similarity distance**. - Highlight similarity and **contrast** differences in terms of data feature value or **attributions**.
Availability	Bias in perceiving that memorable, unusual or adverse events are more likely (frequent) than they truly are.	Seek **base rate** of a diagnosis.	- Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).
Anchoring	Skewed perception of a value due to a supplied numerical value (anchor).	Avoid confirmation and early closure; make use of **lab tests to "prove" other** leading diagnoses. "Crystal ball" exercise (**"premortem"** prospective hindsight [51]).	- Show input **attributions** for multiple outcomes to allow *contrastive reasoning*. - Facilitate **counterfactual** to test *How To* reduce the probability of primary diagnosis with Rules (e.g., aLIME, LORE). - Facilitate **sensitivity analysis** with *What If* explanations to test stability of primary hypothesis.
Confirmation	Collecting redundant information to confirm an existing hypothesis, instead of finding evidence of competing possibilities.	- Use **hypothetical-deductive** method to assess value and role of contemplated tests. - Try to **disprove your diagnosis**, consider conditions of **higher prevalence**.	- Show Findings (**input attribution**) first, instead of Hypotheses (**output posterior probability**). ***Insight:*** *this is opposite to typical Machine Learning apps to show output uncertainty first.* - Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).

https://github.com/slundberg/shap

Wang, Danding, Qian Yang, Ashraf Abdul, and Brian Y. Lim. "Designing theory-driven user-centric explainable AI." In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1-15. 2019.

"However, in the healthcare context, the urge to transcend the black box is confounded by the fact that in some cases “the human body is a black box,” in the words of a Sepsis Watch team member. The preeminent focus on machine learning model explainability or interpretability as a means to provide transparency and accountability in healthcare should be interrogated.

First, front-line clinicians may not want to be oriented towards technology and away from patients.

Second, the current practice of professional clinicians often includes the utilization of information that isn’t comprehensively understood.

Third, causal relationships are not always necessary for application in clinical decision making.

Finally, as scholars have begun to point out, explainability or interpretability is poorly defined and cannot be an end in and of itself without further specification.”

Sendak, Mark, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien.
"" The human body is a black box" supporting clinical decision-making with deep learning."
In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 99-109. 2020.