Apply Machine Learning in research

- An abundance of data become ***accessible*** in recent years. Much of these data, however, including texts, images, sound recordings, and videos, exist in unstructured formats that are difficult for traditional algorithms to handle

- Many recent papers use ML algorithms to ***extract*** useful signals from mostly unstructured data (text, image, sound)

- Appropriateness of using ML algorithms depends on how well they work relative to alternative approaches (manual coding or simpler algorithms)
  - Trade-off between accuracy/efficiency and interpretability/explainability/transparency

Supervised versus unsupervised

- Supervised machine learning
  - *Assumption: I **know** the true construct, help me accurately and efficiently measure them*
  - Requires training sample that fits the research question
    - Manual annotation: e.g., sentiments
    - Naturally occurring annotations: e.g., stock returns, financial statement, ratings, event outcomes

- Unsupervised machine learning
  - Clustering and dimensionality reduction
  - Topic modeling is commonly used because topics can be intuitively interpreted
    - *Assumption: I know **the number of dimensions**, automatically cluster words or contents for me.*

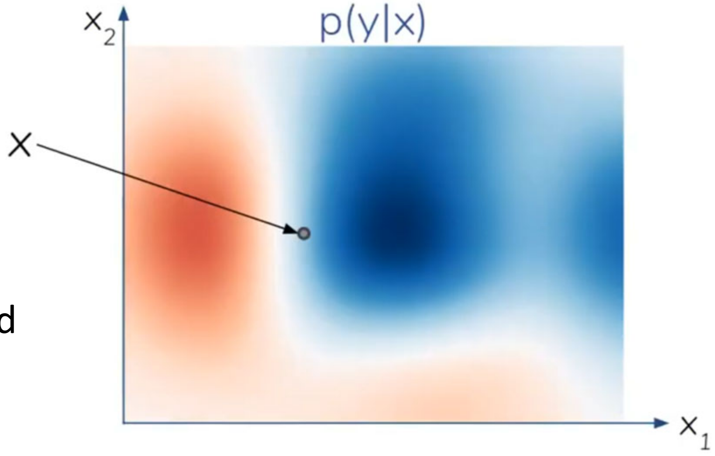
- Interpretable models usually only apply to ***supervised*** machine learning
  - For unsupervised ML, *interpreting* the model output is equivalent to checking if the output ***matches*** with the theoretical concept
    - E.g., *interpreting* topic modeling usually involves reading sentences or examining word clouds

Interpreting supervised ML models

- The ability to explain or to provide the meaning in understandable terms to a human

- How does interpretability help?
  - Detect overfitting; understandable models may work better out of sample
  - Find errors, bugs and undesirable behavior in ML models
  - Understand why models work so we (authors, reviewers, editors) can trust the model
  - Uncover patterns that can detect bias and improve human decisions
    - AlphaGo, insurance loss reserve estimate, Ding et al. 2021; loan approval, Liu 2022, judicial bail decisions, Kleinberg et al. 2018

Machine learning interpretation methods



- Interpretability can be interpreted in many ways

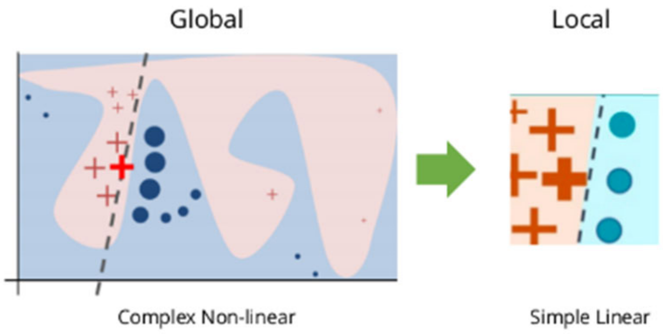
- What parts of **an input** leads to an output?
    - Saliency maps (using gradients or perturbations)
        - Relative importance of inputs
        - If you change or remove an input: Leave-one-out method

- Not very frequently used in business research
    - Which observations in **training example** drive model output?
        - Show where the model picked up certain patterns; actionable
            - If you remove a training point… (computationally expensive)
    - Input reduction/adversarial perturbations
    - Global decision rules
    - Probing internal representations (e.g., layers of neural network or Transformer)

Local Interpretable Model-Agnostic Explanations (LIME)

- “Why should I trust you?” Ribeiro et al. 2016, >10K citations
- Model-Agnostic: can be applied to models for text and image classification.

- Look at model’s prediction for nearby inputs



The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	**Positive** 63.4%
The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.	**Positive** 74.5%
The ~~movie~~ is mediocre, maybe even ~~bad~~.	**Negative** 97.9%

- Closer points are more important than further ones

The movie is mediocre, maybe even bad.

Problems with Interpretability Methods

- Generating interpretation is expensive (many calls to underlying model)
- May not accurately describe model because assume linear relation between models' input and outputs
  - This quarter's sales is ***not*** bad
  - This quarter's sales is ***not*** good