AI: Strategy + Marketing (MGT 853)

The AI $\Longleftrightarrow$ Human Interface (Session 5)

Vineet Kumar

Yale School of Management

Spring 2025

Agenda for Today's Session

- Driving as an ML Problem

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Autonomous Vehicles


80 HYSTER
24

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s) → **Expert Systems** Second-wave used computer programming (1980s to early 2000s)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

If condition X, Then do Y

- Could go to 1000s or 100K lines of code
- Need to add code for each new condition and reprogram system

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s) → **Expert Systems** Second-wave used computer programming (1980s to early 2000s) → **Prediction Models** Third-wave uses (2007 – current)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

If condition X, Then do Y

- Could go to 1000s or 100K lines of code
- Need to add code for each new condition and reprogram system

Predictive Model

- AI system learns and builds the model and delivers better (more accurate prediction) as more data is generated

Converting to Prediction Problem (In class exercise)

Autonomous Vehicles

- Consider the role of prediction in autonomous driving
- Let's walk through the AI Decision Framework

Questions to Ponder

0) What sources of data should the system use?
1) What are the possible predictive problems one might encounter?
2) How should we measure performance?
3) What are appropriate ML algorithms in our toolbox to solve them?
4) What role does judgment play in this problem?

ML Pipeline

Where do humans interface?

ML Pipeline

0. Data Sources
   - First Party
   - Data Broker
   - External Survey
   - Sensors
   - Modality
     Video Text ...

1. Pre-processing
   - Collect & Connect
   - Cleaning & Filtering
   - Outliers
   - Standardization
   - Format → Analysis

2. Visualization
   - Bar Charts
   - Initial insights
   - Correlations
   - Distribution

3. Feature Engineering
   - No new data
   - New feature

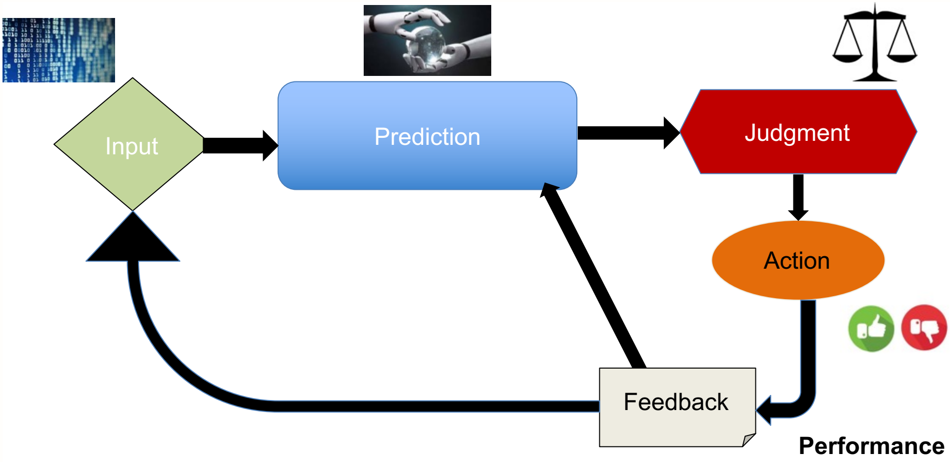
4. Data Splitting
   Train | Test
   Validation

5. Model
   - Accuracy
   - Dependent Variable
   - Explainable
   - Gen or Predict
   - Cost / Time
   - Data

6. Hyperparameter
   - Human designer selects

7. Learning (Training) 8. Validation 9. Testing

AI Decision-Making Framework



Does Feedback also inform Judgment?

Why worry about Black Box?

What if we get very high accuracy?

- 95

Why worry about Black Box?

What if we get very high accuracy?

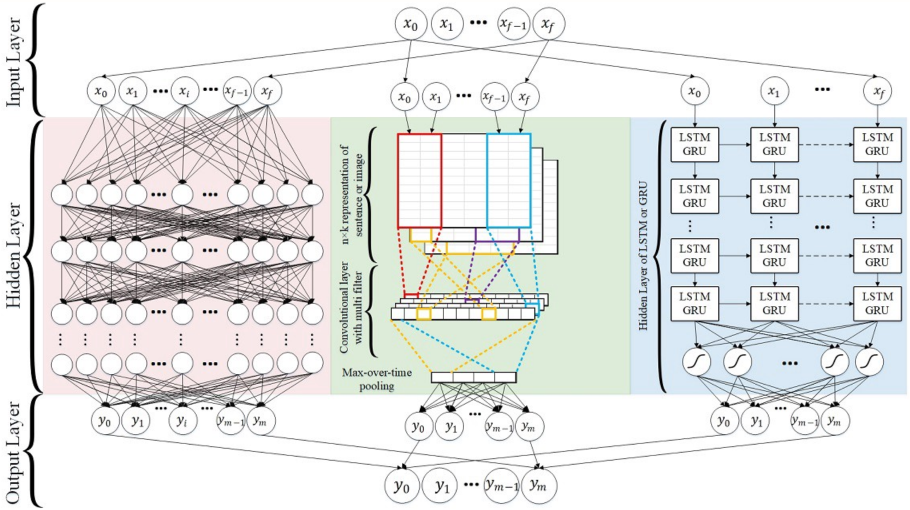
- 95
- 99.x?

Why worry about Black Box?

What if we get very high accuracy?

- 95
- 99.x?
- 100

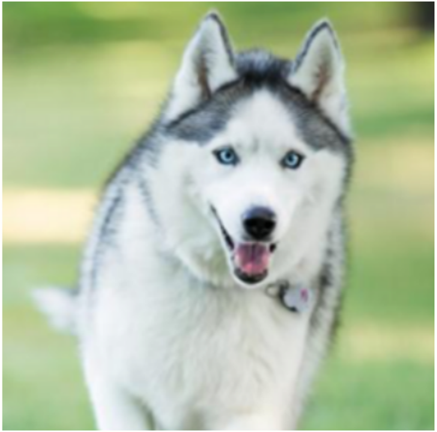
Can we understand this?



Why worry about Black Box?

Why worry about Black Box?

Wolf or Husky?



Source: Why Should I Trust You? Explaining the predictions of any classifier

Even 100% accuracy may not be enough
Need to understand why the model works

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)
- Model produces constructs with meanings known to humans

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

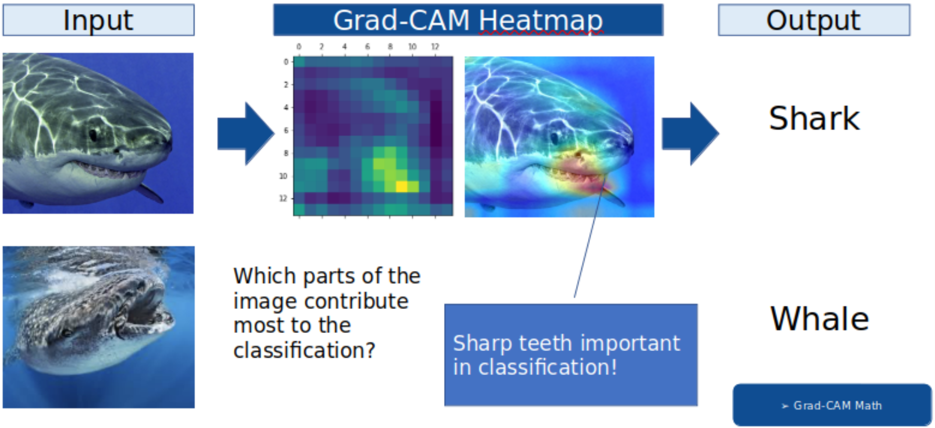
Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)
- Model produces constructs with meanings known to humans
- May not be easy for all models

Explainability in Complex Models

Shark or Whale?

Explainability in Complex Models



Source: Understand your Algorithm with Grad-CAM

Grad-CAM

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable

Explainability $\Longrightarrow$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$

Explainability $\Longrightarrow$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

nature > nature machine intelligence > perspectives > article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

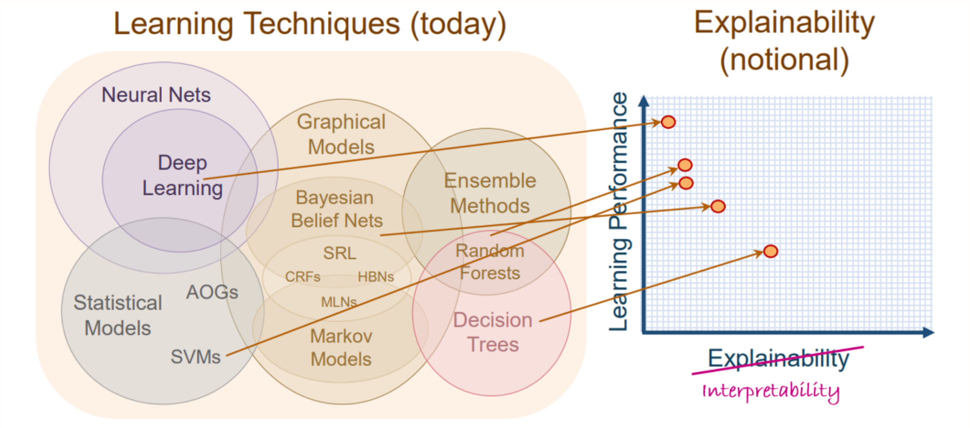
Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently

Performance $\Longleftrightarrow$ Transparency Tradeoff?



https://www.researchgate.net/figure/
Current-Machine-Learning-Techniques-and-Notional-Explainability-Source-1

Big Picture Takeaways

- Accuracy is not enough (even if close to 100%)

Big Picture Takeaways

- Accuracy is not enough (even if close to 100%)
- Transparency, Explainability and Interpretability can be very important for application

Big Picture Takeaways

- Accuracy is not enough (even if close to 100%)
- Transparency, Explainability and Interpretability can be very important for application
  - Often critically important to helping humans understand why AI makes the decisions it does

Big Picture Takeaways

- Accuracy is not enough (even if close to 100%)
- Transparency, Explainability and Interpretability can be very important for application
  - Often critically important to helping humans understand why AI makes the decisions it does
- Without that, we're guessing at how a "well performing" black box is doing its job

Big Picture Takeaways

- Accuracy is not enough (even if close to 100%)
- Transparency, Explainability and Interpretability can be very important for application
    - Often critically important to helping humans understand why AI makes the decisions it does
- Without that, we're guessing at how a "well performing" black box is doing its job
- Broadly, many applications of interpretability, e.g. with cars or watches, why are some products visually appealing?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?