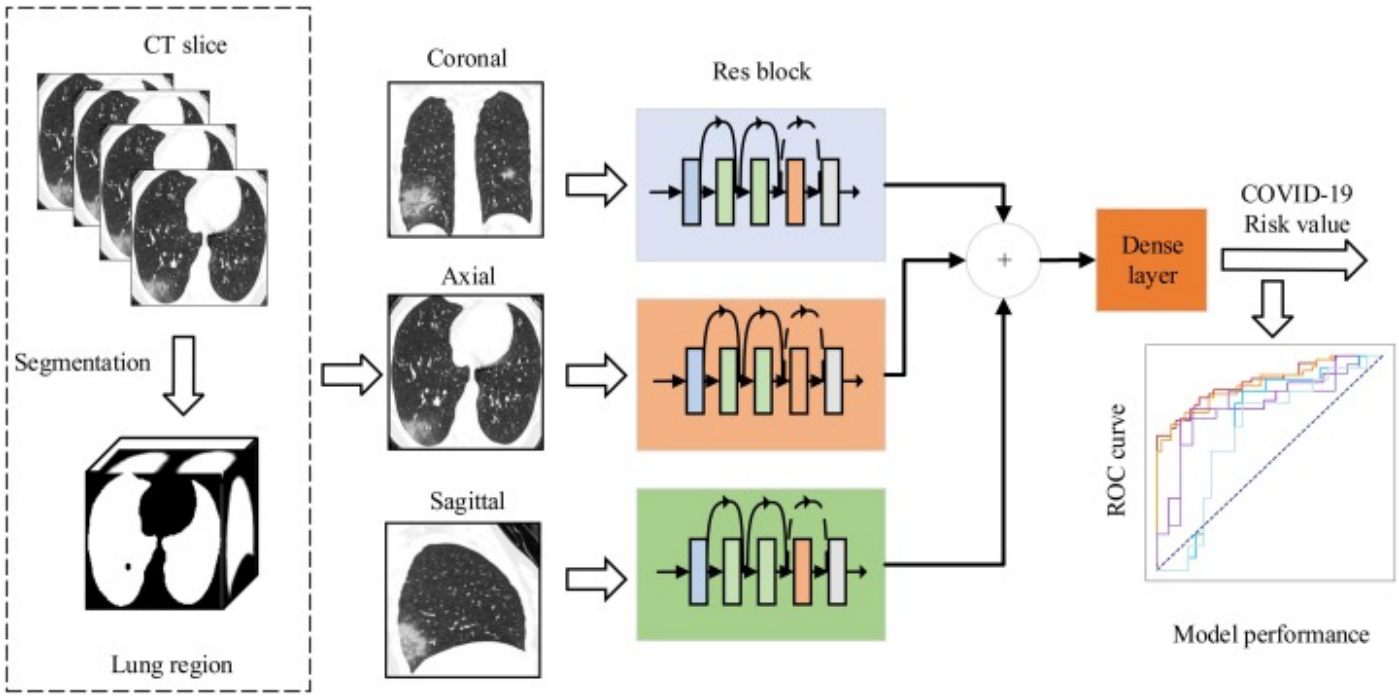
Explainability and Interpretability in ML systems

Sékou-Oumar Kaba

COMP-599 : Design and Build Intelligent Systems

ML-based systems are increasingly used

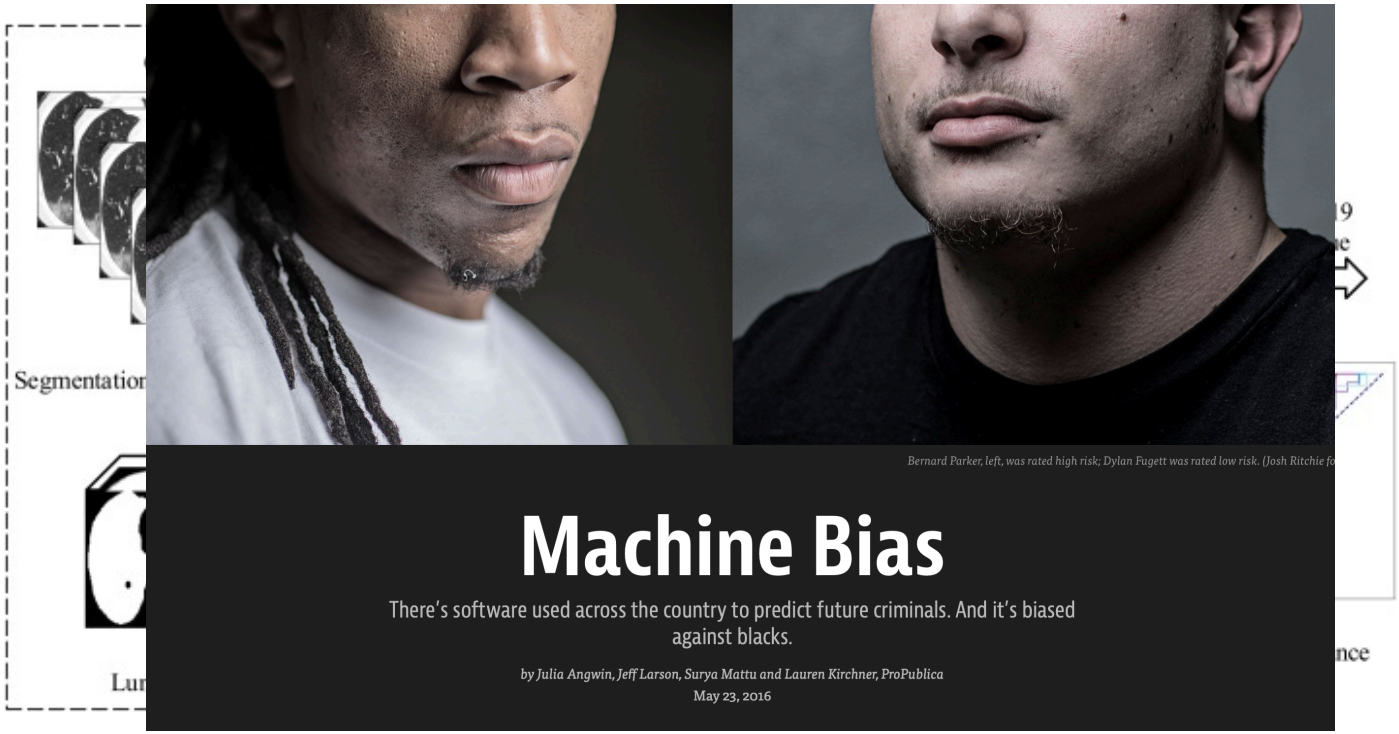
Diagnosis



https://www.ejradiology.com/article/S0720-048X(20)30230-8/fulltext

ML-based systems are increasingly used

Justice and law enforcement



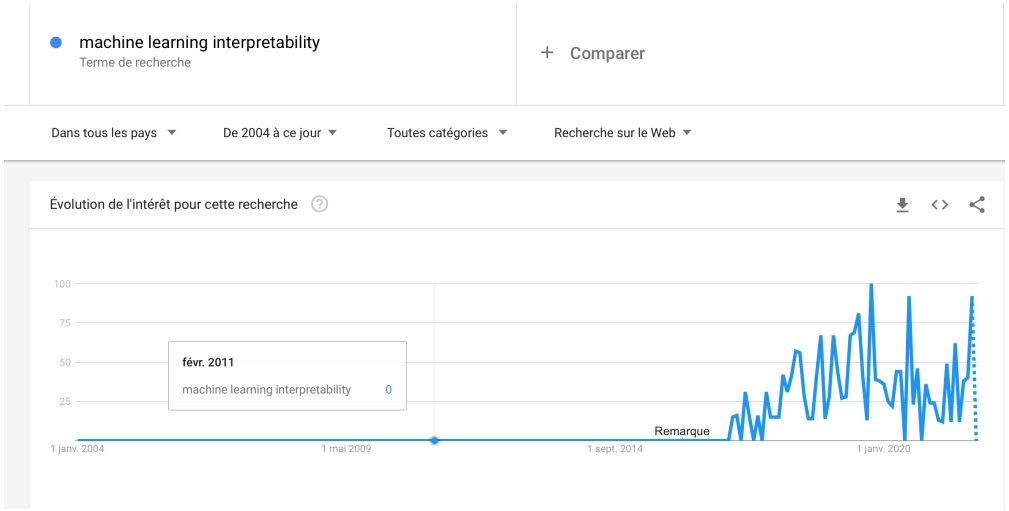
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ML-based systems are increasingly used

Self-driving cars

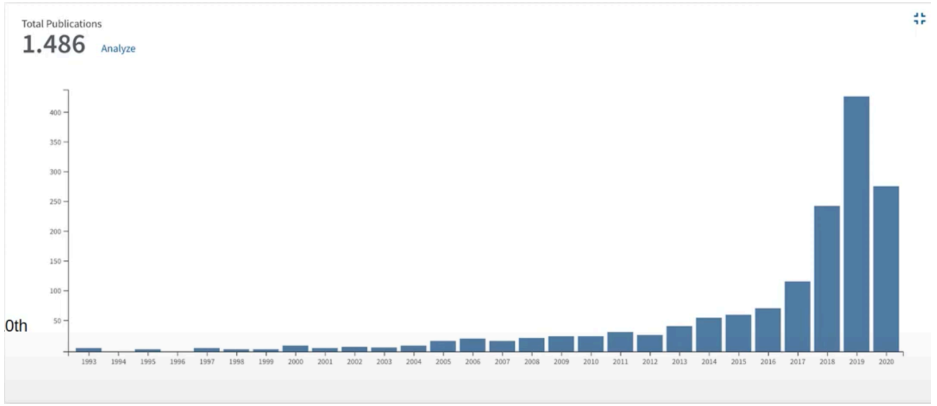
https://www.abc15.com/news/region-southeast-valley/chandler/waymo-car-involved-in-chandler-arizona-crash

Growing discussion on interpretability



Google Trends

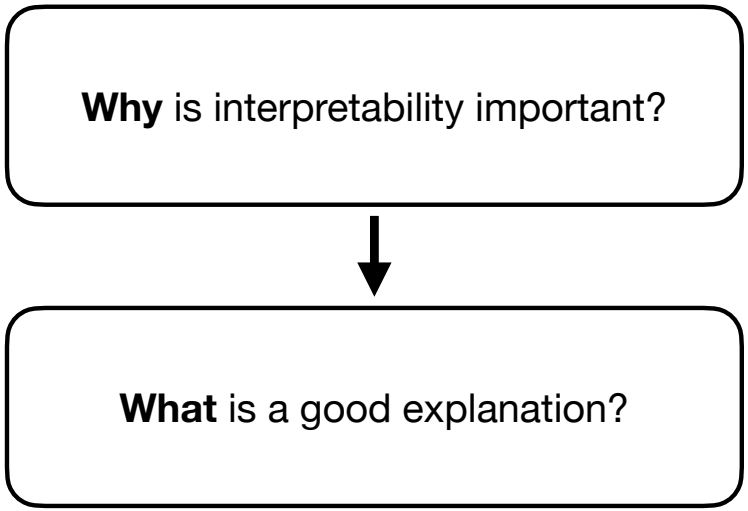
Number of academic papers according to Web of Science



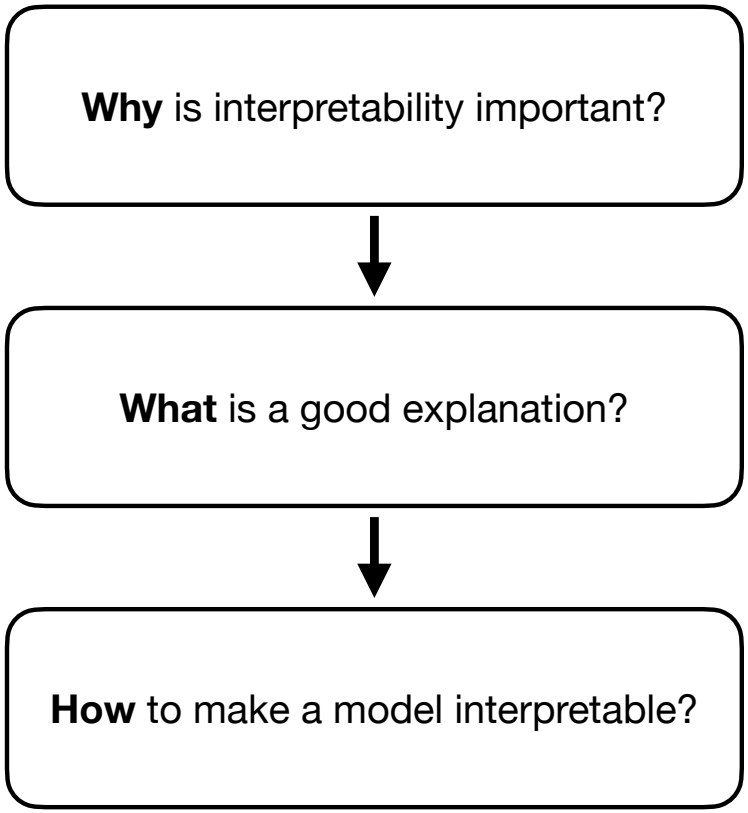
Outline of the presentation

Why is interpretability important?

Outline of the presentation



Outline of the presentation



Why is interpretability important?

Need

- Produce knowledge not just predictions
- Facilitate collaboration with humans
- Debug models
- Audit models
- Increase confidence in models

Why is interpretability important?

Need

- Produce knowledge not just predictions
- Facilitate collaboration with humans
- Debug models
- Audit models
- Increase confidence in models

Stakeholder

- Designers and engineers
- Executives
- End users
- Other stakeholders

Why is interpretability important?

Need

- Produce knowledge not just predictions
- Facilitate collaboration with humans
- Debug models
- Audit models
- Increase confidence in models
- Unfavourable tradeoff
- Security or protection of system

Stakeholder

- Designers and engineers
- Executives
- End users
- Other stakeholders
- Adversaries

What is a good explanation?

Why was I diagnosed cancer? Because :

- You have a genetic predisposition to it
- You did not eat enough berries, which contain antioxidants that naturally prevent cancer
- A mass of tumorous cells that is growing at significant speed was detected in your liver
- An above 10 ppm concentration of the carcinogenic Helicobacter pylori metabolites was detected in your falciform ligament

Why is a high increase in the price of Tesla stock predicted? Because :

- There is acutally no reason, the stock market is inherently chaotic
- Madam Johan Smith bough 12$ worth of Tesla story on July 23th, which triggered a chain reaction among investors
- There is a strong increase in the demand for electric vehicles due to growing environmental concerns and Tesla promises to fulfil that demand
- Tesla's stock is positively correlated with JPMorgan Chase stock and negatively correlated with Boeing stock which are increasing and decreasing respectively

What is a good explanation?

A satisfying explanation is often

- Contrastive
- Selected
- Social
- Focus on the abnormal
- In line with priors
- Counterfactual

What is a good explanation?

A satisfying explanation is often

- Contrastive
- Selected
- Social
- Focus on the abnormal
- In line with priors
- Counterfactual

Should acknowledge human biases

- Representativeness
- Availability
- Anchoring
- Confirmation
- Overconfidence

What is a good explanation?

An ML model can be made interpretable at different levels

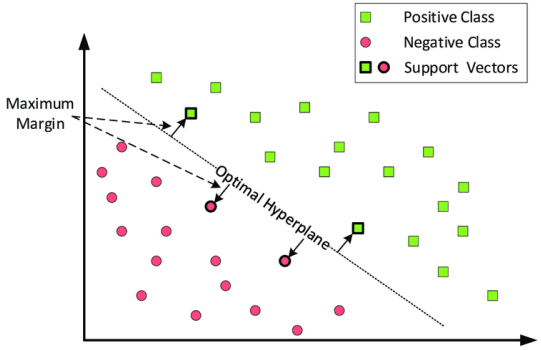
- Openness : Is information about the model readily available?

What is a good explanation?

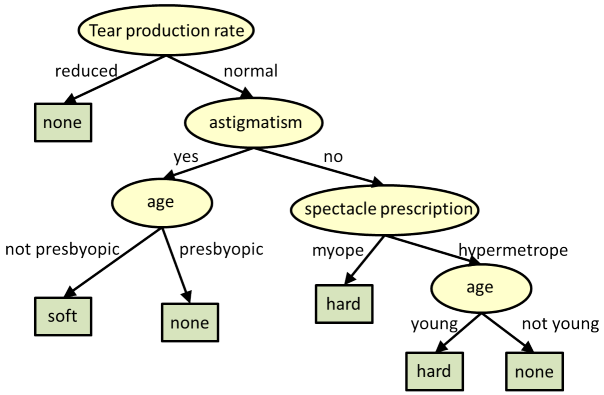
An ML model can be made interpretable at different levels

- Openness : Is information about the model readily available?
- Transparency : Can a human understand how the model produces its outputs?
  - Simulatability
  - Decomposability
  - Algorithmic transparency

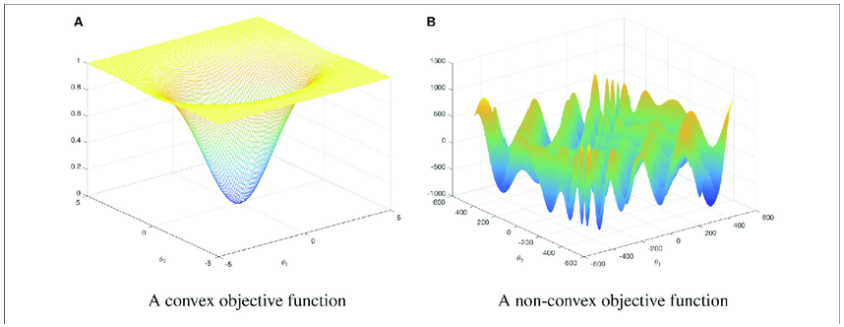
What is a good explanation?



Simulatability



Decomposability



Algorithmic transparency

What is a good explanation?

An ML model can be made interpretable at different levels

- Openness : Is information about the model readily available?
- Transparency : Can a human understand how the model produces its outputs?
  - Simulatability
  - Decomposability
  - Algorithmic transparency
- Post-hoc interpretability : Can explanation for the outputs be obtained?
  - Visualizations
  - Local explanations
  - Examples or prototypes

How to design interpretable models?

- Decide for a definition satisfying level of interpretability before starting the analysis

How to design interpretable models?

- Decide for a definition satisfying level of interpretability before starting the analysis
- Use of simple models
  - Linear (sparse) models
  - Rule-based models
  - Decision trees
  - Deep models

How to design interpretable models?

- Decide for a definition satisfying level of interpretability before starting the analysis
- Use of simple models
  - Linear (sparse) models
  - Rule-based models
  - Decision trees
  - Deep models
- Model-level decisions
  - Feature importance, Visualization, Disentangled representations

How to design interpretable models?

- Decide for a definition satisfying level of interpretability before starting the analysis
- Use of simple models
  - Linear (sparse) models
  - Rule-based models
  - Decision trees
  - Deep models
- Model-level decisions
  - Feature importance, Visualization, Disentangled representations
- Model-agnostic methods
  - Sensitivity analysis, Influential samples, Prototypes

How to design interpretable models?

Evaluation of interpretability

- There should a different explanation for each type of stakeholder
- Expert-level evaluation
- User-level evaluation
- Beware some pitfalls
  - Simpler models are not necessarily more interpretable
  - Post-hoc explanations can be msileading

Take home message

- There should a different explanation for each type of stakeholder
- Explanations should be adapted to user needs but also help them overcome their biases
- There are many ways to make a model more interpretable
  - Openness
  - Transparency
  - Post-hoc explanations
- In practice
  - Use simple models as much as performance requirements permits
  - Look to understand feature importance or parts of the model