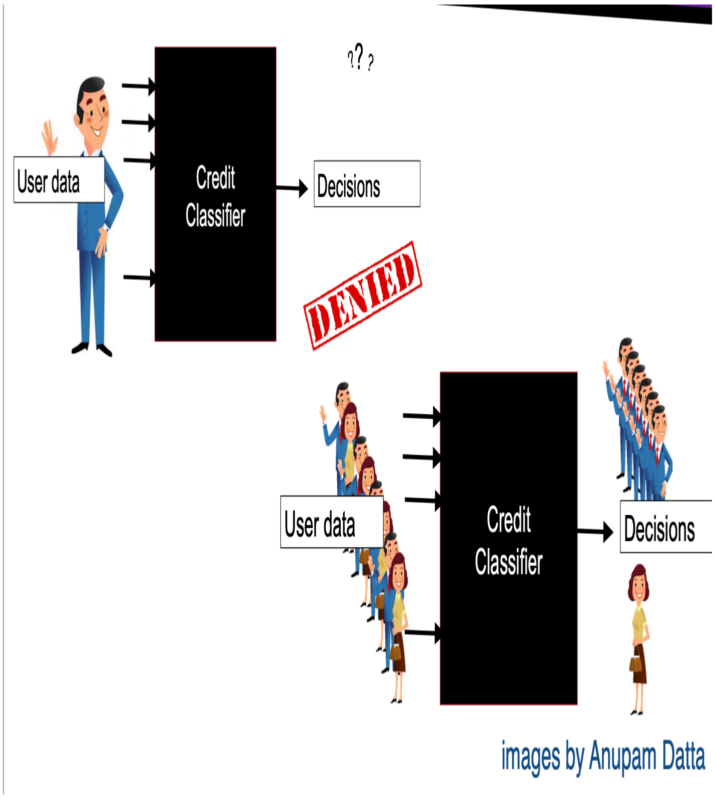
Transparency and Interpretability in Data Science

Introduction **Beyond the Blackbox** **Explanatory Methods**

Introduction

In an era where algorithmic systems increasingly contribute to making decisions such as loan approval, predictive policing, and online content personalization, it is an undeniable fact that machine learning models are permeating people's daily lives on a large scale. While applying models with high accuracy does serve the purpose of increasing efficiency and taking precautions, we must consider the benefits and drawbacks for different stakeholders and allow them to interpret how the decisions are made in this process, ensuring that as we harness the power of machine learning, we do so responsibly.



images by Anupam Datta

Tradeoff between Accuracy and Interpretability

There is an inherent trade-off in machine learning models between accuracy and interpretability. Simpler models such as linear regression compromise accuracy for the ease of explanability. In comparison, highly accurate models, like deep neural networks and ensembles, are often complex and operate as "black boxes", making them less interpretable. As the figure shows, imaging you are seeking mortgage loan and have filled all the information needed for application. The credit classifier will decide whether it is approved by running its algorithm with your input. This process remains as a blackbox as you, as the applicant, only knows the result. We need explanatory methods to peel back the layers to reveal the decision-making process, to ensure transparency and equitable process.

This website briefly shows my understanding in the course DS-UA 202 Responsible Data Science