# Chapter 1 - Introduction

## 1 Introduction

This book describes how to make supervised machine learning models interpretable. Basically, you can use interpretable models or model-agnostic methods (can give local or global explanations).

## 2 Story Time

Three fictional stories that are supposed to motivate why interpretability is important. Skip.

## 3 What is Machine Learning?

We focus on supervised learning (not unsupervised or reinforcement learning). Machine learning often beats hand designed algorithms, but is hard to debug.

## 4 Terminology

An algorithm is a set of rules. Machine learning is a set of techniques for learning from data and making predictions. A machine learning algorithm defines how to create a machine learning model from data. A machine learning model maps inputs to outputs. A black box (in contrast to white box) model is one whose internal mechanisms are not decipherable (e.g. a complicated neural net is a black box model, a simple linear regression is white box).

Interpretable machine learning is a set of techniques that aims to make a model's predictions understandable to humans.

A dataset is made of (features, target) pairs. We assume features are interpretable. A prediction is a model's estimate of the target.