OLIVER WYMAN

MODEL RISK AND MACHINE LEARNING

HAZARDS RELATED TO AI AND ITS DEPLOYMENT

13 October 2020

David Waller

A ZOOMED-OUT VIEW

Security

Information falls into the wrong hands

Data

Data we use causes causes new harms or worsens others

Models

AI/ML model-based decisions fail in surprising and new ways

Systems

Connected systems of models can become brittle

People

Human-machine interactions can create failures

A PERIODIC TABLE OF AI MODEL RISK

	Security	Data		Models		Systems	People
Theory	**Pr** Privacy Leakage	**Hd** High Dimensionality	**Bi** Bias in Datasets	**Nc** Non-convex Functions	**Xp** Lack of Explainability	**En** Entanglement of Models	**Sk** Skill Gaps
	Cy Cyber Security	**Ds** Distribution Shift	**Fa** Fairness in Decisions	**Rd** Randomness in Algorithms	**Sf** Silent Failure	**Fb** Feedback Loops	**Hm** Human-Machine Interfaces
Applied		**Ad** Adversarial Attacks	**Df** Deep Fakes			**Sc** Scale of Errors	**Wf** Workforce Dislocation

Hd

High
Dimensionality

IN HIGH DIMENSIONS, THERE ARE NO “NEAR NEIGHBORS”

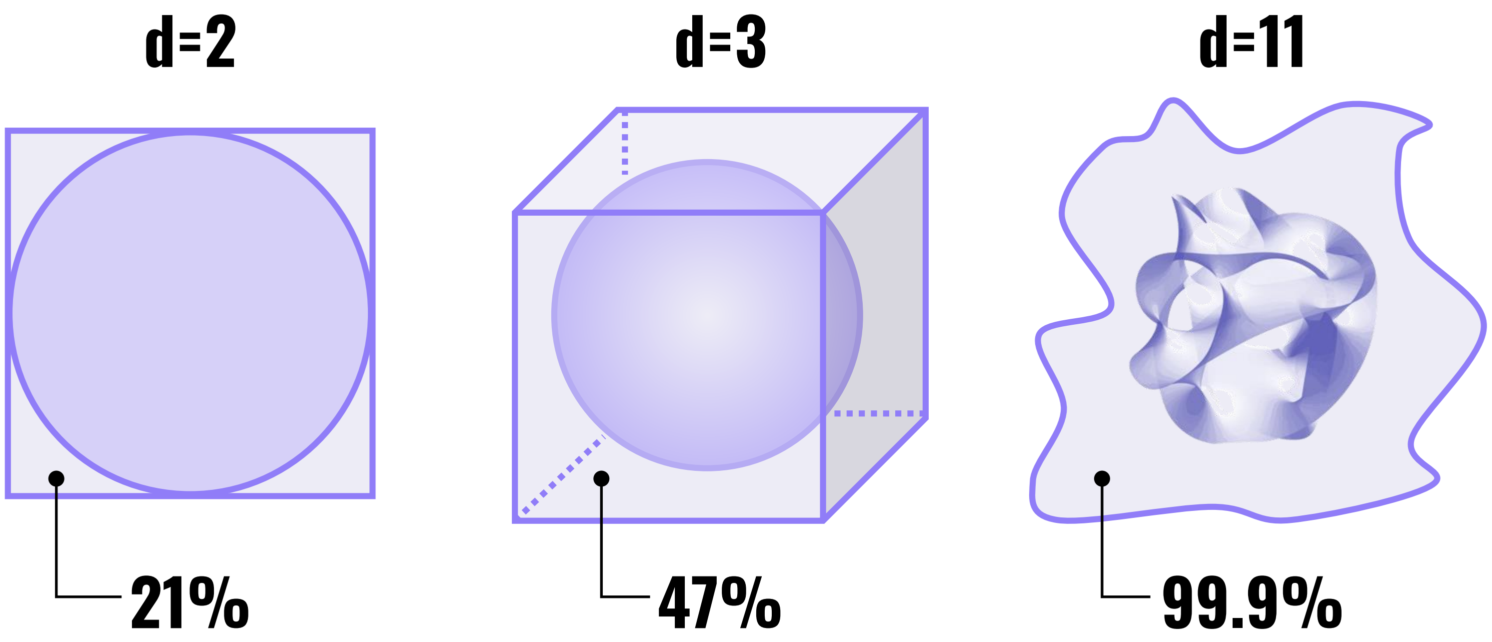
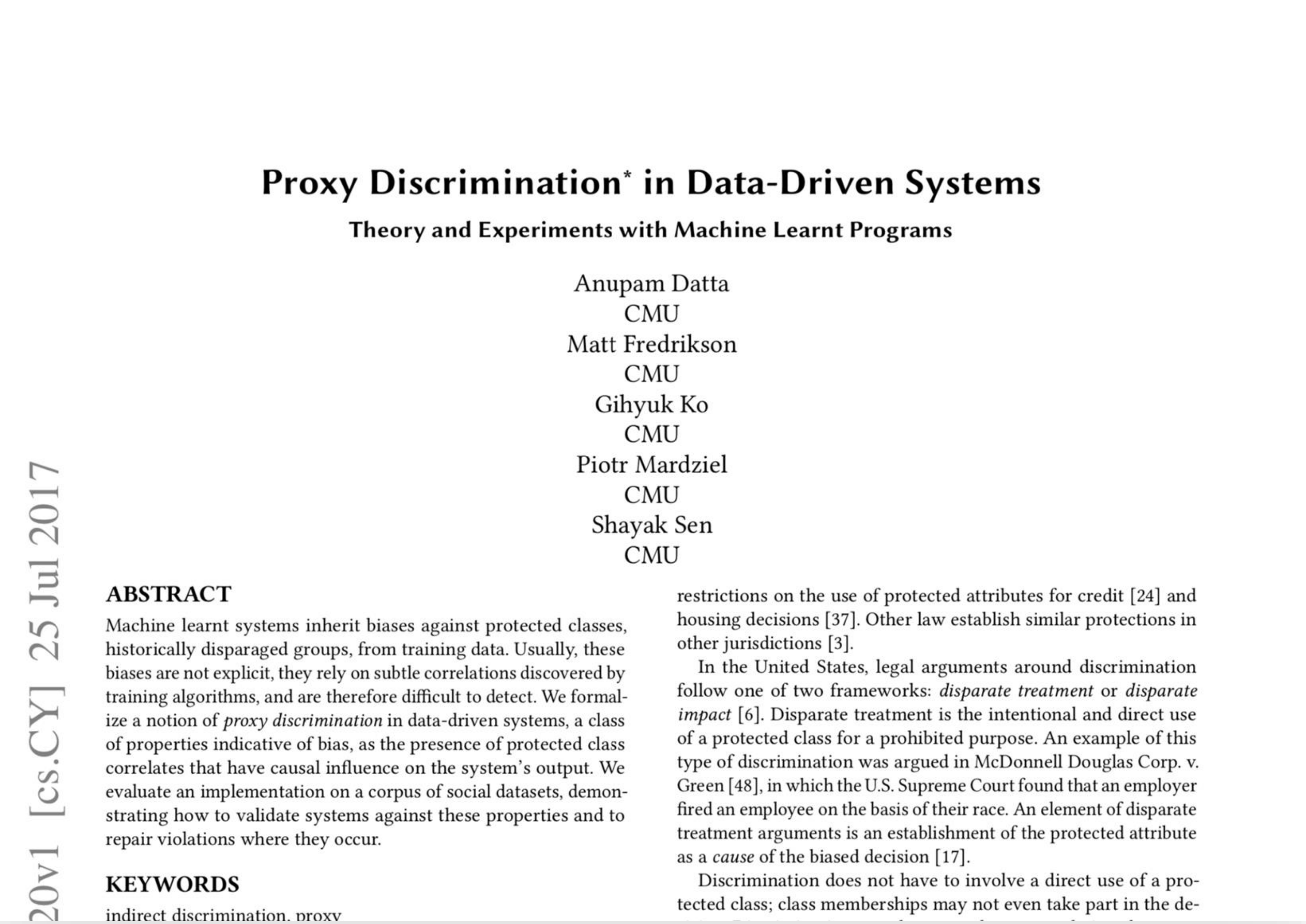


Image credit: https://en.wikipedia.org/wiki/File:Calabi-Yau.png

Fa

Fairness in Decisions

WITH MANY FEATURES, PROXIES FOR PROTECTED CLASSES MAY EXIST OR EMERGE



Proxy Discrimination* in Data-Driven Systems

Theory and Experiments with Machine Learnt Programs

Anupam Datta
CMU
Matt Fredrikson
CMU
Gihyuk Ko
CMU
Piotr Mardziel
CMU
Shayak Sen
CMU

ABSTRACT

Machine learnt systems inherit biases against protected classes, historically disparaged groups, from training data. Usually, these biases are not explicit, they rely on subtle correlations discovered by training algorithms, and are therefore difficult to detect. We formalize a notion of *proxy discrimination* in data-driven systems, a class of properties indicative of bias, as the presence of protected class correlates that have causal influence on the system's output. We evaluate an implementation on a corpus of social datasets, demonstrating how to validate systems against these properties and to repair violations where they occur.

KEYWORDS

indirect discrimination, proxy

restrictions on the use of protected attributes for credit [24] and housing decisions [37]. Other law establish similar protections in other jurisdictions [3].

In the United States, legal arguments around discrimination follow one of two frameworks: *disparate treatment* or *disparate impact* [6]. Disparate treatment is the intentional and direct use of a protected class for a prohibited purpose. An example of this type of discrimination was argued in McDonnell Douglas Corp. v. Green [48], in which the U.S. Supreme Court found that an employer fired an employee on the basis of their race. An element of disparate treatment arguments is an establishment of the protected attribute as a *cause* of the biased decision [17].

Discrimination does not have to involve a direct use of a protected class; class memberships may not even take part in the de-

Credit: Datta et al, arXiv:1707.08120

Xp Lack of Explainability

DATA IN HIGH DIMENSIONS IS INHERENTLY HARDER TO “EXPLAIN”

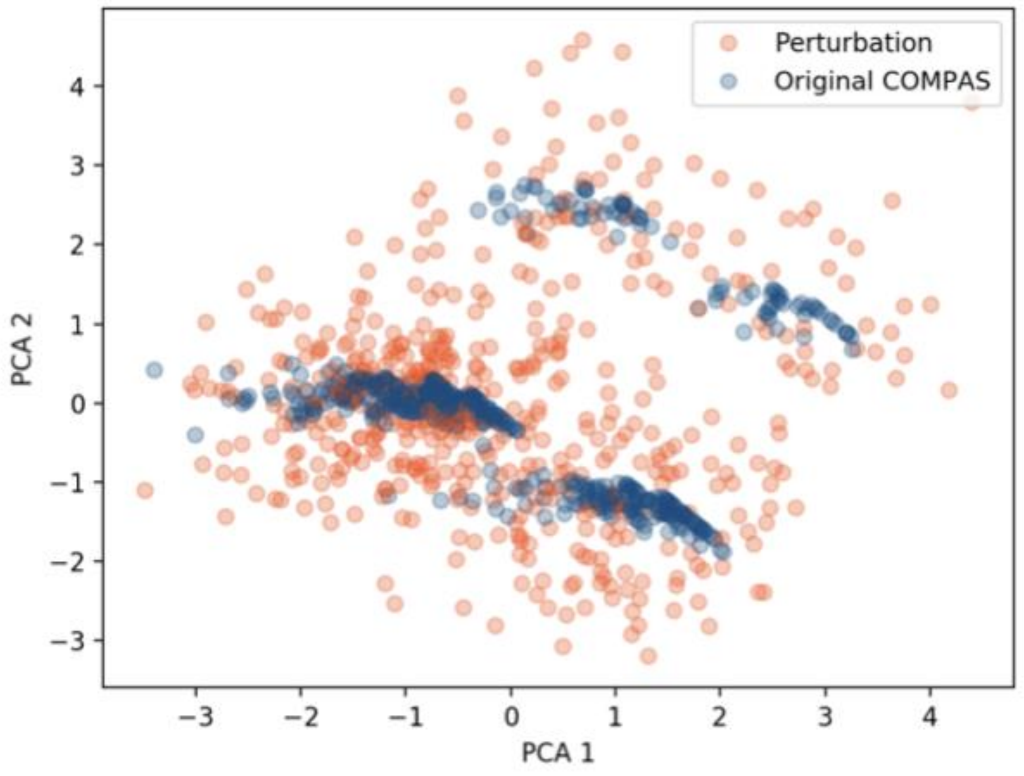


Figure 1: PCA applied to the COMPAS dataset (blue) as well as its LIME style perturbations (red).

Credit: Slack et al., 2020, Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Bias in Datasets

DISPARITIES IN DATA USED TO TRAIN MODELS CAN CREATE DIFFERENCES IN OUTCOMES

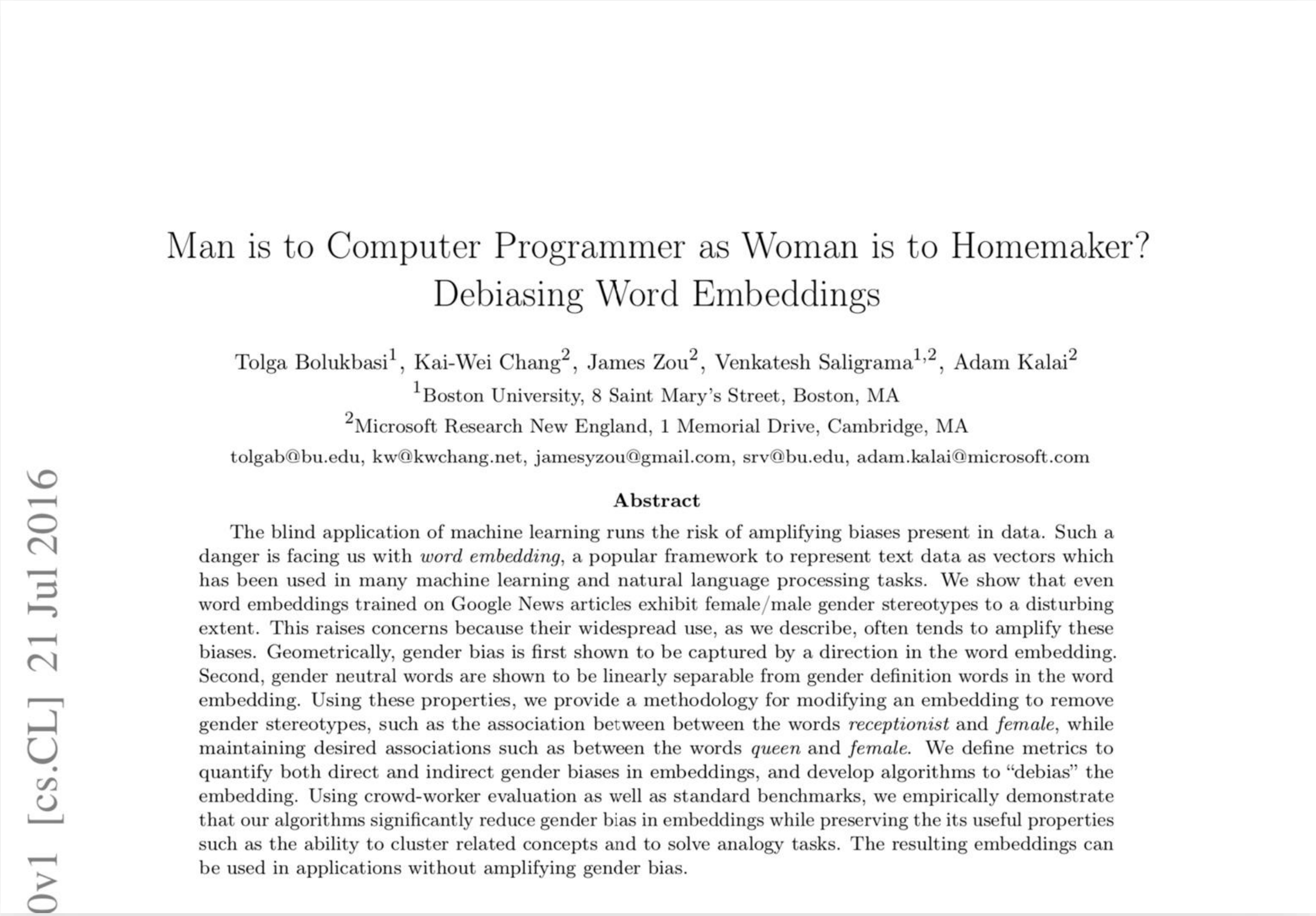
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Product A	94.0%	79.2%	100%	98.3%	20.8%
Product B	99.3%	65.5%	99.2%	94.0%	33.8%
Product C	88.0%	65.3%	99.7%	92.9%	34.4%

Credit: Joy Boulamwini, Gender Shades project, MIT Media Lab, https://www.media.mit.edu/projects/gender-shades/overview. Product names anonymized.

Bi

Bias in Datasets

MODELS CAN ENCODE AND REPLICATE SOCIETAL BIASES AND STEREOTYPES


0v1 [cs.CL] 21 Jul 2016

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

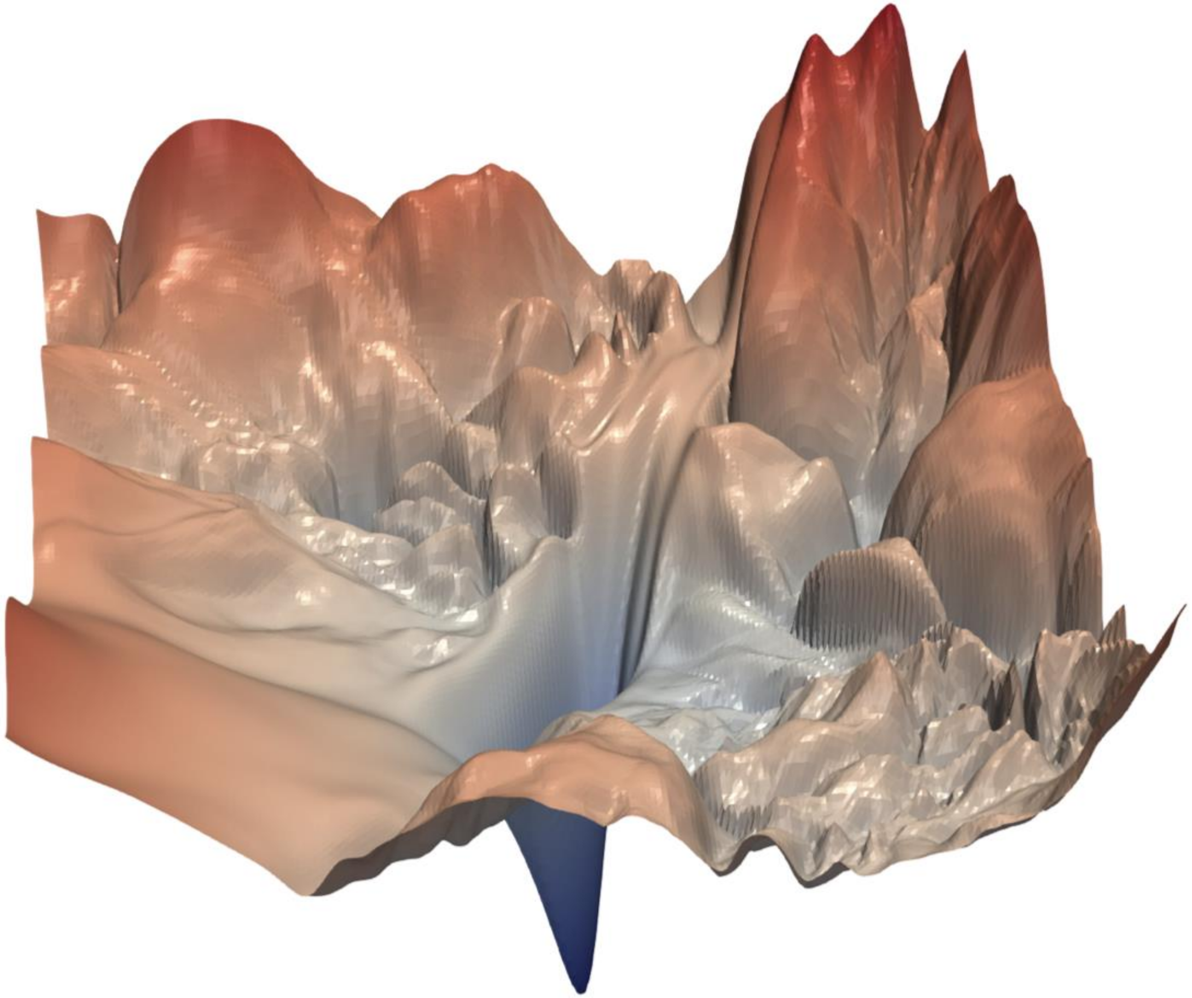
The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Credit: Bolukbasi et al., 2016, arXiv:1607.06520

Nc

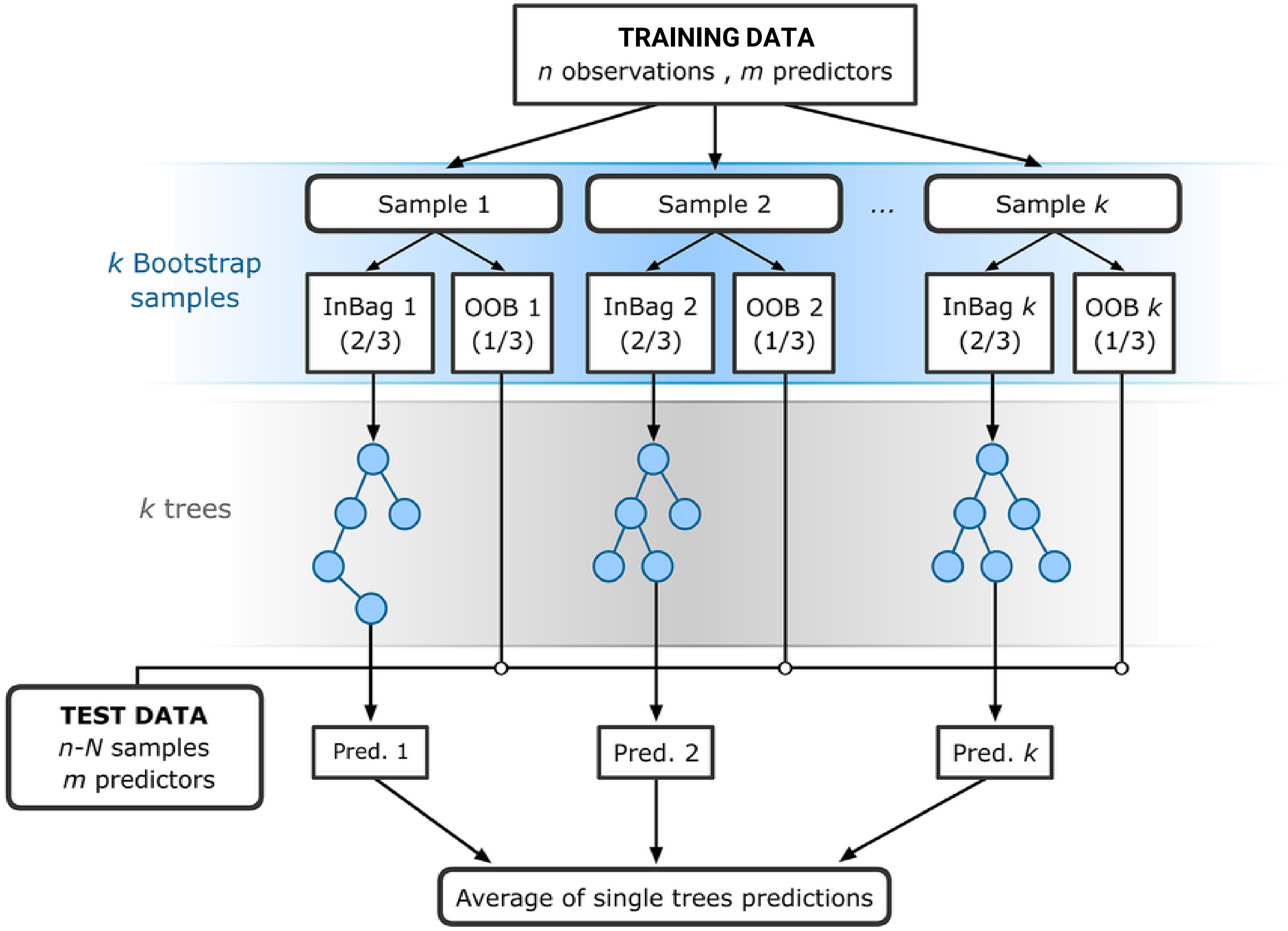
Non-convex
Functions

WITH HIGHLY NON-CONVEX FUNCTIONS, IT'S HARD TO FIND THE GLOBAL MINIMUM



Credit: Li et al, 2018, NeurIPS, Visualizing the Loss Landscape of Neural Nets

BY EMBEDDING RANDOMNESS IN ALGORITHMS, ML RESULTS CAN BE HARD TO REPRODUCE



Credit: Rodriguez-Galiano, 2016, Modelling interannual variation in the spring and autumn land surface phenology of the European forest, Biogeosciences. Text slightly edited.

En

Entanglement
of Models

WHEN MODELS DEPEND ON EACH OTHER, THE RESULTING SYSTEM CAN BE BRITTLE

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

1 Introduction

As the machine learning (ML) community continues to accumulate years of experience with live systems, a wide-spread and uncomfortable trend has emerged: developing and deploying ML systems is relatively fast and cheap, but maintaining them over time is difficult and expensive.

Credit: Sculley et al., 2015, Hidden Technical Debt in Machine Learning Systems, NeurIPS

FOR REFERENCE: A PERIODIC TABLE OF AI MODEL RISK

	Security	Data		Models		Systems	People
Theory	**Pr** Privacy Leakage	**Hd** High Dimensionality	**Bi** Bias in Datasets	**Nc** Non-convex Functions	**Xp** Lack of Explainability	**En** Entanglement of Models	**Sk** Skill Gaps
	Cy Cyber Security	**Ds** Distribution Shift	**Fa** Fairness in Decisions	**Rd** Randomness in Algorithms	**Sf** Silent Failure	**Fb** Feedback Loops	**Hm** Human-Machine Interfaces
Applied		**Ad** Adversarial Attacks	**Df** Deep Fakes			**Sc** Scale of Errors	**Wf** Workforce Dislocation

OLIVER WYMAN