



# Explainable, Interpretable, and Transparent Machine Learning

## Abstract

Most real datasets have hidden biases. Being able to detect the impact of the bias in the data on the model, and then to repair the model, is critical if we are going to deploy machine learning in applications that affect people's health, welfare, and social opportunities. This requires models that are intelligible. In this talk, we will review variety of tools and open source projects, including Microsoft InterpretML, for training interpretable models and explaining blackbox systems. These tools (WhatIf, Fair 360, etc) are made to help developers experiment with ways to introduce explanations of the output of AI systems. This talk will review the InterpretML and draw parallels with LIME, ELI5, and SHAP. InterpretML implements a number of intelligible models—including Explainable Boosting Machine (an improvement over generalized additive models ), and several methods for generating explanations of the behavior of black-box models or their individual predictions.

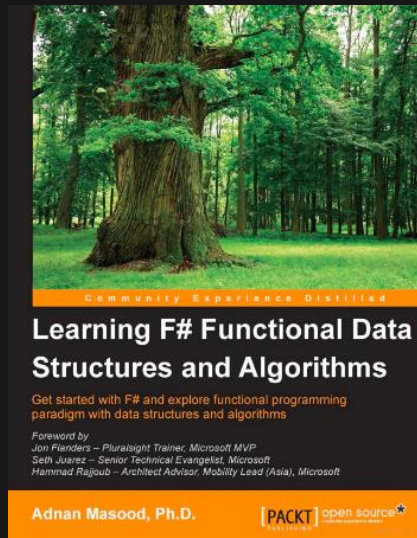
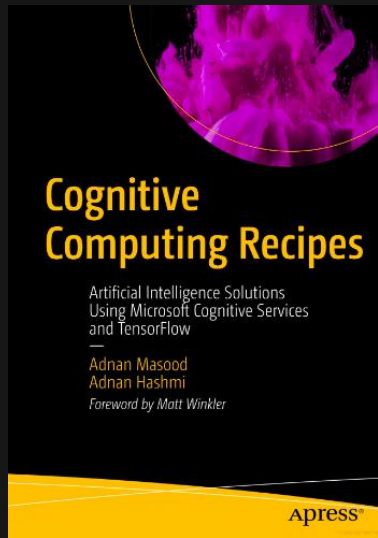


**Adnan Masood**, Ph.D.

Chief Architect – AI and ML at UST Global

Visiting Scholar at Stanford University

Microsoft MVP (Most Valuable Professional) for AI



**Dr. Adnan Masood**





**TRIGGER**

**WARNING**



## **Why is Model Interpretability Important?**

**When using an algorithm's outcomes to make high-stakes decisions, it's important to know which features the model did and did not take into account. Additionally, if a model isn't highly interpretable, the business might not be legally permitted to use its insights to make changes to processes. In heavily regulated industries like banking, insurance, and healthcare, it is important to be able to understand the factors that contribute to likely outcomes in order to comply with regulation and industry best practices.**



*"AI is likely to be either the best or worst thing ever to happen to humanity." ~Stephen Hawking*

*"If I had to guess at what our biggest existential threat is, it's probably AI." ~Elon Musk*



*"When a few people control a platform with extreme intelligence, it creates dangers in terms of power and control." ~Bill Gates*





ROBOT EMOTIONS

\$18<sup>99</sup>

10000101101

# TK Brand™ Robot Emotions

FEEL LIKE A HUMAN™

## SCHADENFREUDE

[Emotion] D: //shwo2227.dll | ITEM No.: 7236 49321 236  
564 mHz 42 pin 700mAmpere | DDR400 XPC-3600-K RET

Contains 1 USB module pre-loaded with 1 emotion

**WARNING:** Installation of TK Brand Robot Emotions (the Product) shall constitute acceptance of TK Brand Terms and Conditions. Please consult your factory documentation before installing. Do not disengage safety overrides when using the Product as you may experience a variety of file errors. You may also encounter an unexpected end of file. When using the Product we recommend that you abstain from operating heavy machinery. If you are heavy machinery operator, please activate your robot distress beacon and wait for the arrival of a licensed service technician.

TTM

10000101101

# TK Brand™ Robot Emotion

FEEL LIKE A HUMAN™

## LOVE

[Emotion] D: //shwo2227.dll | ITEM No.: 7236 49321 236  
564 mHz 42 pin 700mAmpere | DDR400 XPC-3600-K RET

Contains 1 USB module pre-loaded with 1 emotion

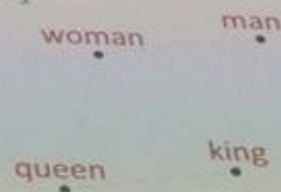
**WARNING:** Installation of TK Brand Robot Emotions (the Product) shall constitute acceptance of TK Brand Terms and Conditions. Please consult your factory documentation before installing. Do not disengage safety overrides when using the Product as you may experience a variety of file errors. You may also encounter an unexpected end of file. When using the Product we recommend that you abstain from operating heavy machinery. If you are heavy machinery operator, please activate your robot distress beacon and wait for the arrival of a licensed service technician.

TTM

## Analogies generated by embedding

Parallelograms capture semantics: [MikolovYZ 13]

- Man:King :: Woman:Queen
- Paris:France :: Tokyo:Japan
- He:Brother :: She:Sister
- He:Blue :: She:Pink
- He:Doctor :: She:Nurse
- He:Architect :: She:Interior designer
- He:Realist :: She:Feminist
- She:Pregnancy :: He:Kidney stone
- He:Computer programmer :: She:Homemaker



Based on word2vec trained on Google News corpus

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

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

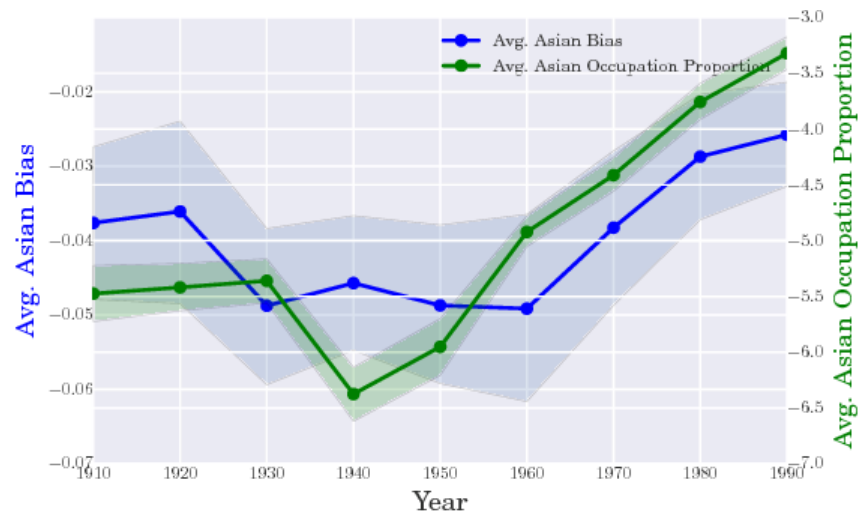
(Submitted on 21 Jul 2016)

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 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 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.



Hispanic	Asian	White
housekeeper	professor	smith
mason	official	blacksmith
artist	secretary	surveyor
janitor	conductor	sheriff
dancer	physicist	weaver
mechanic	scientist	administrator
photographer	chemist	mason
baker	tailor	statistician
cashier	accountant	clergy
driver	engineer	photographer

) The top ten occupations most closely associated with each



(d) Average ethnic (Asian vs White) bias score over time for occupations in COHA (blue) vs the average conditional log pro-

## SOFTWARE SCANDALS

Prominent incidents that highlight the effect of algorithmic bias

**December 2009** | Hewlett-Packard investigates instances of so-called “racist camera software” which had trouble recognizing dark-skinned people

**March 2015** | A Carnegie Mellon University study determines that some personalized ads from sites such as Google and Facebook are gender-biased

**July 2015** | A Google algorithm mistakenly captions photos of black people as “Gorillas”

**March 2016** | Microsoft shuts down AI chatbot Tay after it starts using racist language

**May 2016** | ProPublica investigation finds that a computer program used to track future criminals in the US is racially biased

**September 2016** | Machine-learning algorithms used to judge an international beauty contest displays bias against dark-skinned contestants

FEBRUARY 8, 2017

## Code-Dependent: Pros and Cons of the Algorithm Age

Algorithms are aimed at optimizing everything. They can save lives, make things easier, and conquer chaos. Still, experts worry they can also put too much control in the hands of corporations and governments, perpetuate bias, create filter bubbles, cut choices, creativity, and serendipity, and could result in greater unemployment.

By Lee Rainie and Janna Anderson

**FOR MEDIA OR OTHER INQUIRIES:**

Lee Rainie, Director, Pew Research  
Internet, Science and Technology Project  
Janna Anderson, Director, Elon University's  
Imagining the Internet Center  
Dana Page, Senior Communications  
Manager  
202.419.4372  
[www.pewresearch.org](http://www.pewresearch.org)

# Microsoft's Tay & Twitter: A 24-hour Story



<http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

# Microsoft's Tay & Twitter: A 24-hour Story



[https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref\\_src=twsrc%5Etfw](https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw)



# Microsoft's Tay & Twitter: A 24-hour Story



[https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref\\_src=twsrc%5Etfw](https://twitter.com/geraldmellor/status/712880710328139776/photo/1?ref_src=twsrc%5Etfw)



# Our approach

We've identified six ethical principles to guide the development and use of artificial intelligence with people at the center of everything we do.

[Explore our AI vision >](#)

## Microsoft AI principles

Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values.

### Fairness

AI systems should treat all people fairly

### Reliability & Safety

AI systems should perform reliably and safely

### Privacy & Security

AI systems should be secure and respect privacy

### Inclusiveness

AI systems should empower everyone and engage people

### Transparency

AI systems should be understandable

### Accountability

AI systems should have algorithmic accountability

## Guidelines for responsible bots

Conversational AI bots must be designed in a way that they earn the trust of others. Learn the principles to building bots that create confidence in your company and services.

[Review the guidelines >](#)

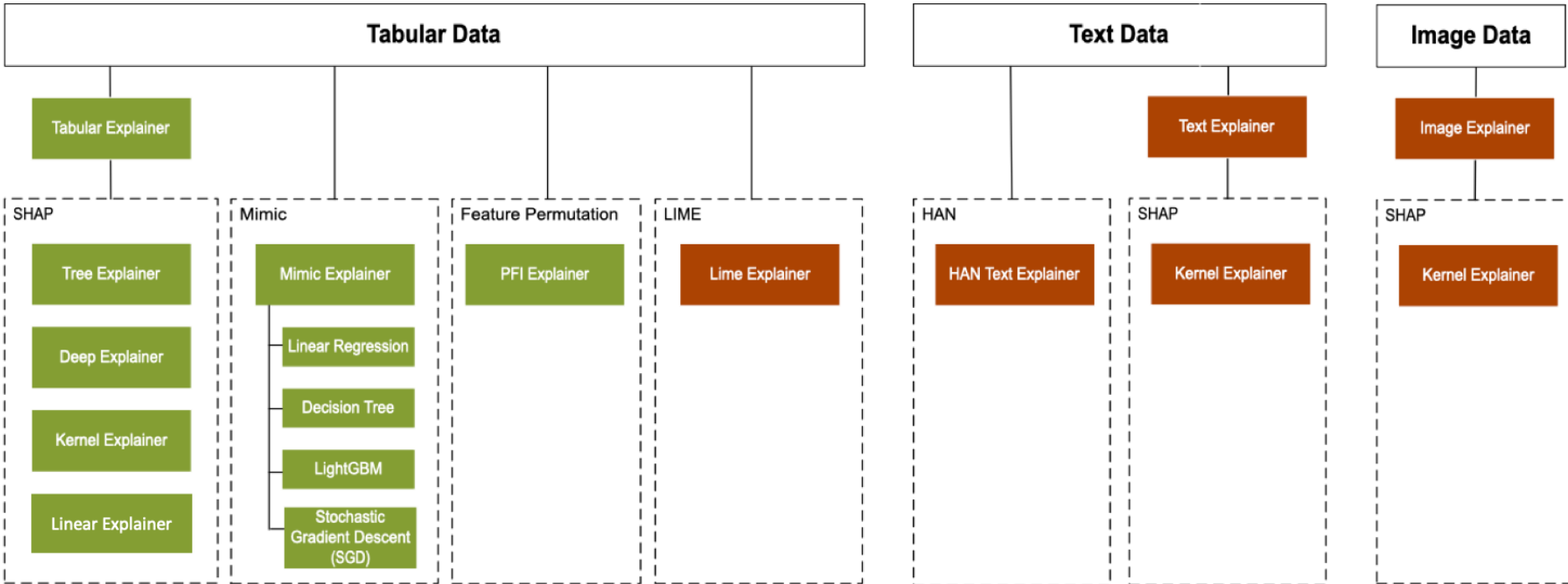
## Microsoft AI is driving innovation

Discover how we are applying our AI principles to create solutions

# Machine Learning Interpretability

Legend:

- Main Package
- Contrib Package



## Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment

Jonathan Dodge  
dodgej@eecs.oregonstate.edu  
Oregon State University

Q. Vera Liao  
Yunfeng Zhang  
vera.liao@ibm.com  
zhangyun@us.ibm.com  
IBM Research AI

Rachel K. E. Bellamy  
Casey Dugan  
rachel@us.ibm.com  
cadugan@us.ibm.com  
IBM Research AI

### 2.1 Fairness of Machine Learning Systems

One of several definitions for algorithmic fairness is: “...*discrimination is considered to be present if for two individuals that have the same characteristic relevant to the decision making and differ only in the sensitive attribute (e.g., gender/race) a model results in different decisions*” [8]. The consequence of deploying unfair ML systems could be *disparate impact*, practices which adversely affect people of one protected characteristic more than another in a comparable situation [8, 14].



AI Could Predict Death. But What? x +

wired.com/story/ai-bias-predict-death/


**WIRED** BUSINESS CULTURE GEAR MORE SIGN IN SUBSCRIBE

AMITHA KALAICHANDRAN SCIENCE 04.21.2019 08:00 AM

# AI Could Predict Death. But What If the Algorithm Is Biased?

Opinion: Researchers are studying how artificial intelligence could predict risks of premature death. But the health care industry needs to consider another risk: unconscious bias in AI.

f t ✉



Get unlimited WIRED access. [Subscribe](#)

# AI Fairness 360 (AIF360 v0.2.2)

A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. <https://github.com/IBM/AIF360>

## **bias mitigation algorithms**

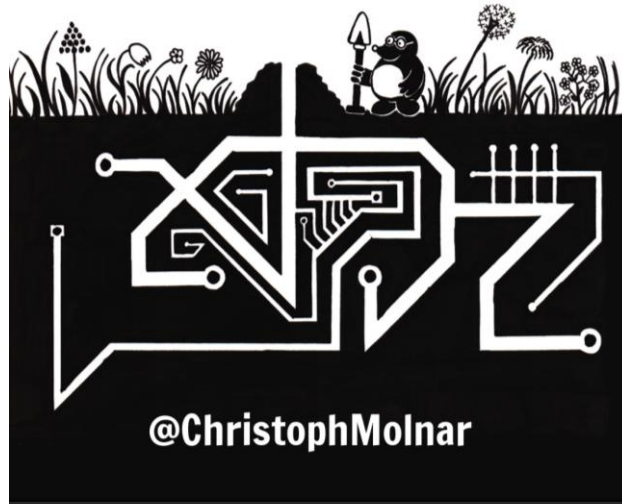
- Optimized Preprocessing ([Calmon et al., 2017](#))
- Disparate Impact Remover ([Feldman et al., 2015](#))
- Equalized Odds Postprocessing ([Hardt et al., 2016](#))
- Reweighing ([Kamiran and Calders, 2012](#))
- Reject Option Classification ([Kamiran et al., 2012](#))
- Prejudice Remover Regularizer ([Kamishima et al., 2012](#))
- Calibrated Equalized Odds Postprocessing ([Pleiss et al., 2017](#))
- Learning Fair Representations ([Zemel et al., 2013](#))
- Adversarial Debiasing ([Zhang et al., 2018](#))
- Meta-Algorithm for Fair Classification ([Celis et al., 2018](#))

## **Supported fairness metrics**

- Comprehensive set of group fairness metrics derived from selection rates and error rates
- Comprehensive set of sample distortion metrics
- Generalized Entropy Index ([Speicher et al., 2018](#))

# Interpretable Machine Learning

A Guide for Making  
Black Box Models Explainable



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable

## Contents

Preface	1
Introduction	2
Story Time	4
What Is Machine Learning?	10
Terminology	12
Interpretability	15
Importance of Interpretability	15
Taxonomy of Interpretability Methods	21
Scope of Interpretability	23
Evaluation of Interpretability	25
Properties of Explanations	26
Human-friendly Explanations	29
Datasets	34
Bike Rentals (Regression)	34
YouTube Spam Comments (Text Classification)	35
Risk Factors for Cervical Cancer (Classification)	36
Interpretable Models	37
Linear Regression	38
Logistic Regression	54
GLM, GAM and more	61
Decision Tree	79
Decision Rules	85
RuleFit	100
Other Interpretable Models	108
Model-Agnostic Methods	110
Partial Dependence Plot (PDP)	113
Individual Conditional Expectation (ICE)	119
Accumulated Local Effects (ALE) Plot	125
Feature Interaction	146
Feature Importance	154
Global Surrogate	163
Local Surrogate (LIME)	168
Shapley Values	177
Example-Based Explanations	189
Counterfactual Explanations	191
Adversarial Examples	199
Prototypes and Criticisms	208
Influential Instances	218
A Look into the Crystal Ball	234
The Future of Machine Learning	235
The Future of Interpretability	237
Contribute to the Book	240
Citing this Book	241
Acknowledgements	242
References	243
R Packages Used for Examples	246



Retail	Marketing	Healthcare	Telco	Finance
<ul style="list-style-type: none"><li>• Demand forecasting</li><li>• Supply chain optimization</li><li>• Pricing optimization</li><li>• Market segmentation and targeting</li><li>• Recommendations</li></ul>	<ul style="list-style-type: none"><li>• Recommendation engines &amp; targeting</li><li>• Customer 360</li><li>• Click-stream analysis</li><li>• Social media analysis</li><li>• Ad optimization</li></ul>	<ul style="list-style-type: none"><li>• Predicting Patient Disease Risk</li><li>• Diagnostics and Alerts</li><li>• Fraud</li></ul>	<ul style="list-style-type: none"><li>• Customer churn</li><li>• System log analysis</li><li>• Anomaly detection</li><li>• Preventative maintenance</li><li>• Smart meter analysis</li></ul>	<ul style="list-style-type: none"><li>• Risk Analytics</li><li>• Customer 360</li><li>• Fraud</li><li>• Credit scoring</li></ul>





## Top 5 AI Trends to Watch in the Years Ahead



- 1 Less Hype More Action**  
As machine learning and neural network technology takes on more routine tasks, real progress towards augmenting human productivity and driving value from tedious tasks will be seen in 2018.



- 2 Human-Free Interactions**  
Gartner predicts that 85% of customer interactions will be managed without a human in the next few years. More businesses plan to harness the power of conversational AI chatbots and other virtual assistants to manage the routine work.



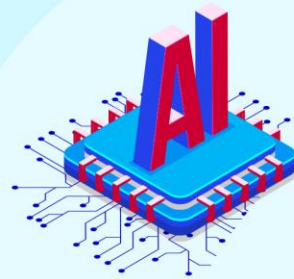
- 3 Explainable AI**  
Explainable AI plans to develop machine learning techniques that offer more explainable AI models whilst upholding prediction accuracy. Explainable and transparent AI will encourage wider adoption of machine learning techniques.



- 4 Prescriptive Analytics for Businesses**  
Businesses will incorporate prescriptive analytics tools into their operations to optimize business processes.



- 5 AI in Medicine**  
By the end of 2019, half of the leading healthcare systems will have employed some level of AI within their diagnostic groups with solutions for population health, hospital operations and an extensive range of clinical specialties.



# Why do we need it?



## BUILDING TRUST

It is easier for humans to trust a system that explains its decisions compared to a black box.



## DEBUGGING

The ability to interpret is valuable in the research and development phase as well as after deployment. Later, when a model is used in a product, things can go wrong. An interpretation for an erroneous prediction helps to understand the cause of the error.



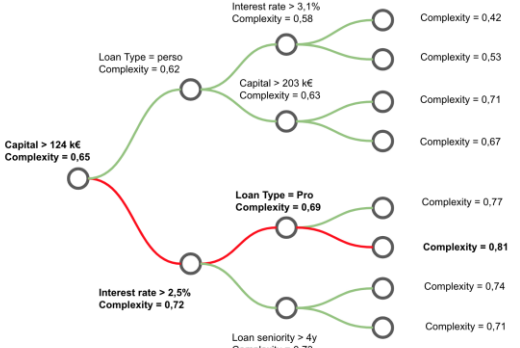
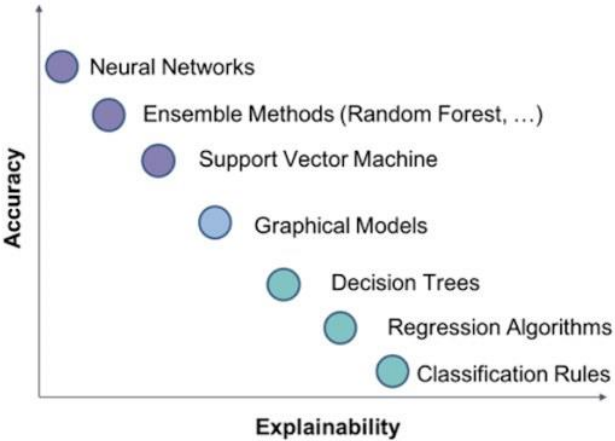
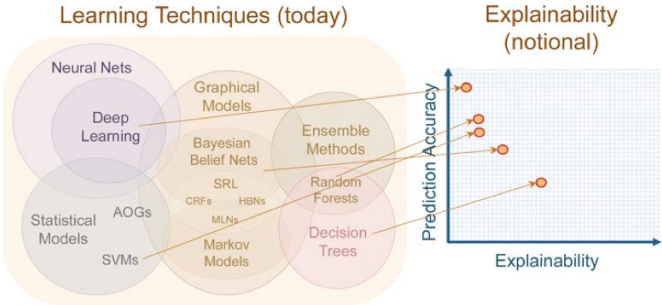
## REGULATION

In some sectors (for example, MD medical decision towards a patient), decisions should be backed by reasons.

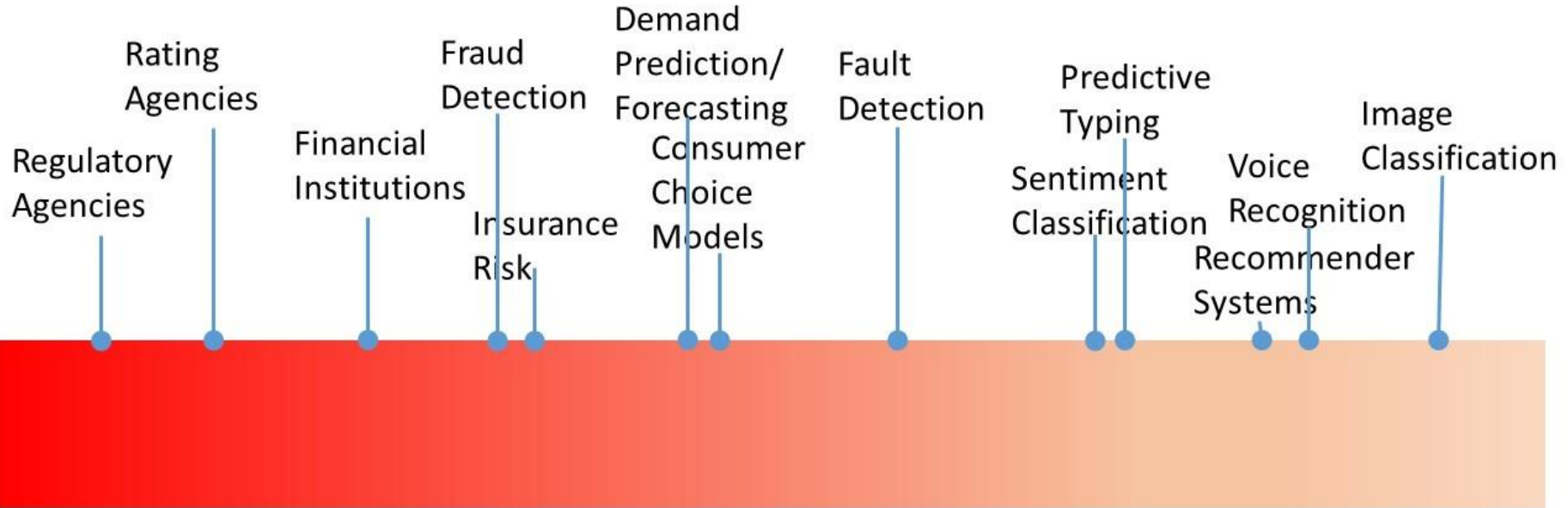




# Salient Ideas

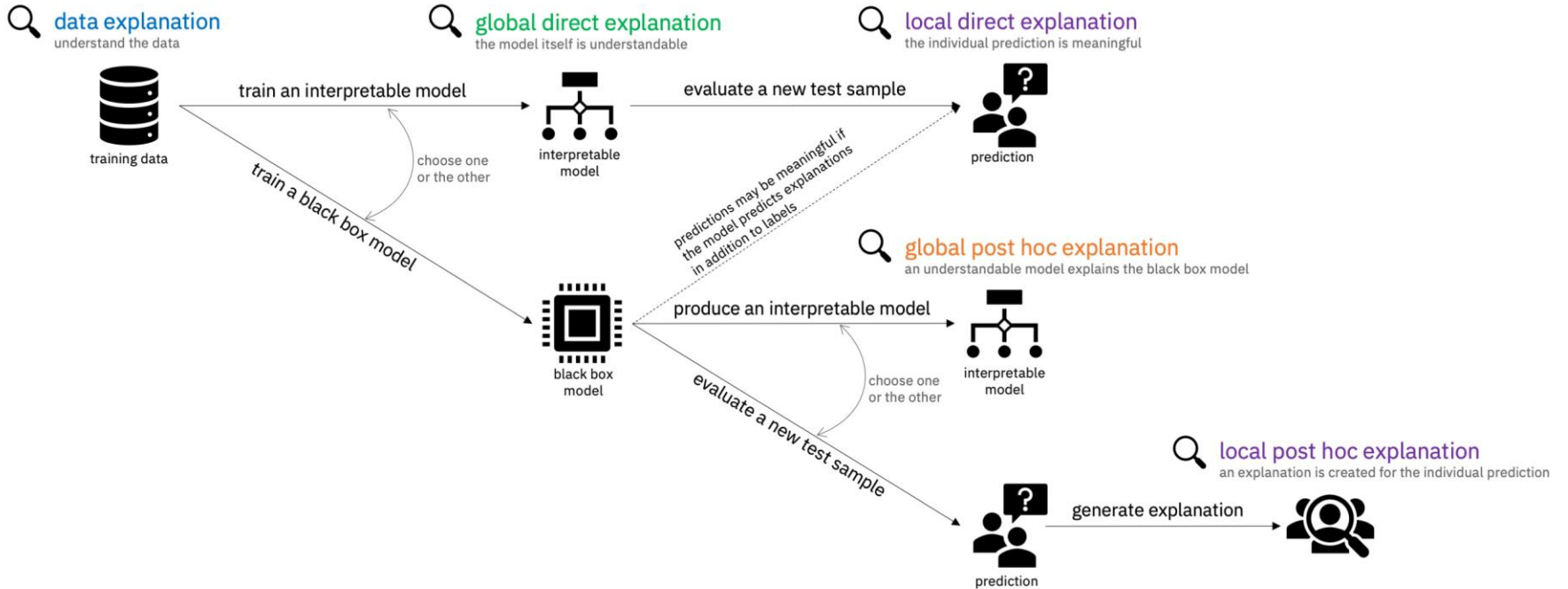


prediction  $0.81 = 0.65$  (trainset mean complexity)  $+ 0.07$  (gain from Capital)  $- 0.03$  (loss from Interest rate)  $+ 0.12$  (gain from loan type)



← Model explainability important

→ Model explainability not-so important





## Stages of AI explainability

### Pre-modelling explainability

#### Goal

Understand/describe data used to develop models

#### Methodologies

- Exploratory data analysis
- Dataset description standardization
- Dataset summarization
- Explainable feature engineering

### Explainable modelling

#### Goal

Develop inherently more explainable models

#### Methodologies

- Adopt explainable model family
- Hybrid models
- Joint prediction and explanation
- Architectural adjustments
- Regularization

### Post-modelling explainability

#### Goal

Extract explanations to describe pre-developed models

#### Methodologies

- Perturbation mechanism
- Backward propagation
- Proxy models
- Activation optimization





---

The authors suggest that those offering datasets or APIs should provide a datasheet that addresses a set of standardized questions covering the following topics:

- The motivation for dataset creation
- The composition of the dataset
- The data collection process
- The preprocessing of the data
- The distribution of the data
- The maintenance of the data
- The legal and ethical considerations

# Datasheets for Datasets

Timnit Gebru<sup>1</sup>, Jamie Morgenstern<sup>2</sup>, Briana Vecchione<sup>3</sup>, Jennifer W. Hanna Wallach<sup>4</sup>, Hal Daumé III<sup>4,5</sup>, and Kate Crawford

<sup>1</sup>Google  
<sup>2</sup>Georgia Institute of Technology  
<sup>3</sup>Cornell University  
<sup>4</sup>Microsoft Research  
<sup>5</sup>University of Maryland  
<sup>6</sup>AI Now Institute

January 15, 2020

## Abstract

The machine learning community currently has no standardized process datasets. To address this gap, we propose *datasheets for datasets*. In the every component, no matter how simple or complex, is accompanied with describes its operating characteristics, test results, recommended uses, and By analogy, we propose that every dataset be accompanied with a datasheet its motivation, composition, collection process, recommended uses, and so datasets will facilitate better communication between dataset creators and d and encourage the machine learning community to prioritize transparency a

### A Datasheet for Studying Face Recognition in Unconstrained Environments

### Labeled Faces in the Wild

#### Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hair, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.<sup>1</sup>

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

#### Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings) people and interactions between them, nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as “background.”

How many instances are there in total (of each type, if appropriate)? The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

<sup>1</sup>All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.  
 Original paper: <http://www.cs.cornell.edu/people/pabou/movie-review-data/> LFW survey: <http://ivw.www.cs.umass.edu/lfw/lfw.pdf> Paper assessing LFW demographic characteristics: [http://biometrics.csl.msu.edu/Publications/FaceInWild-UnconstrainedAgeGenderRaceEstimation\\_MSU\\_TechReport2014.pdf](http://biometrics.csl.msu.edu/Publications/FaceInWild-UnconstrainedAgeGenderRaceEstimation_MSU_TechReport2014.pdf). LFW website: <http://ivw.www.cs.umass.edu/lfw/>.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does this instance consist of? “Raw” data (e.g., unprocessed text or image/jpg features)? In other case, please provide a description. Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description. Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2, which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10<sup>th</sup> subset should be used for testing. At a minimum, we recommend reporting the estimated mean accuracy,  $\bar{\mu}$ , and the standard error of the mean:  $S_{\bar{\mu}}$  is given by:

$$\bar{\mu} = \frac{\sum_{i=1}^{10} \mu_i}{10} \quad (1)$$

where  $\mu_i$  is the percentage of correct classifications on View 2 using subset  $i$  for testing.  $S_{\bar{\mu}}$  is given as:

$$S_{\bar{\mu}} = \frac{\sigma}{\sqrt{10}} \quad (2)$$

Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Dataset Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.4%
Percentage of Asian subjects	8.0%
Percentage of people between 0-20 years old	1.5%
Percentage of people between 21-40 years old	31.6%
Percentage of people between 41-60 years old	45.8%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images*. Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

#### Collection Process

How was the data associated with this instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., pair-of-opposites tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human control, software program, software API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley<sup>1</sup>. The

<sup>1</sup>Faces in the Wild: <http://tamara.berkeley.com/projects/FacesInWild/>



images in this database were gathered from news articles on the web using software to crawl news articles.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The original Faces in the Wild dataset is a sample of pictures of people appearing in the news on the web. Labeled Faces in the Wild is thus also a sample of images of people found on the news on line. While the intention of the dataset is to have a wide range of demographic (e.g., age, race, ethnicity) and image (e.g., pose, illumination, lighting) characteristics, there are many groups that have few instances (e.g., only 1.57% of the dataset consists of individuals under 20 years old).

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Subsequent gender, age and race annotations listed in [http://biometrics.csl.msu.edu/Publications/FaceInWild-UnconstrainedAgeGenderRaceEstimation\\_MSU\\_TechReport2014.pdf](http://biometrics.csl.msu.edu/Publications/FaceInWild-UnconstrainedAgeGenderRaceEstimation_MSU_TechReport2014.pdf) were performed by crowd workers found through Amazon Mechanical Turk.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Each instance is an image of a person.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The data was crawled from public web sources.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Unknown

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if applicable).

Figure 1: Example datasheet for Labeled Faces in the Wild [25], page 1.

Figure 3: Example datasheet for Labeled Faces in the Wild [25], page 3.



AI

# Google Cloud AI removes gender labels from Cloud Vision API to avoid bias

KHARI JOHNSON @KHARIJOHNSON FEBRUARY 20, 2020 10:34 AM



Image Credit: Khari Johnson / VentureBeat

Google Cloud AI is removing the ability to label people in images as “man” or “woman” with its Cloud Vision API, the company told VentureBeat today. Labeling is used to classify images and train machine learning models, but Google is removing gendered labels because it violates Google’s AI principle to avoid creating biased systems.

“Given that a person’s gender cannot be inferred by appearance, we have decided to remove these labels in order to align with the artificial intelligence principles at Google, specifically

# A "nutrition label" for datasets.

The Data Nutrition Project aims to create a standard label for interrogating datasets for measures that will ultimately drive the creation of better, more inclusive algorithms.

Our current prototype includes a highly-generalizable interactive data diagnostic label that allows for exploring any number of domain-specific aspects in datasets. Similar

## Dataset Fact Sheet

### Metadata



**Title** COMPAS Recidivism Risk Score Data

**Author** Broward County Clerk's Office, Broward County Sheriff's Office, Florida

**Email** browardcounty@florida.usa

**Description** Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

**DOI** 10.5281/zenodo.1164791

**Time** Feb 2013 - Dec 2014

**Keywords** risk assessment, parole, jail, recidivism, law

**Records** 7214

**Variables** 25

**priors\_count**: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

**two\_year\_recid**: *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.* **nominal**

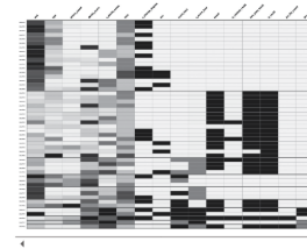
**Missing Units** 15452 (8%)

This dataset contains variables named "age", "race", and "sex"

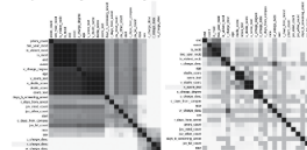
## Probabilistic Modeling

### Analysis

12



### Dependency Probability Pearson R



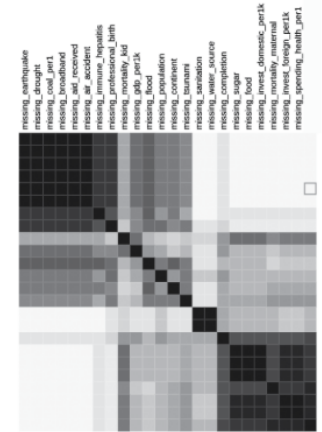
## Missing Units

### Clustering Variable

race

### Missing Variable

r\_days\_from\_arrest





# Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender  
Department of Linguistics  
University of Washington  
ebender@uw.edu

Batya Friedman  
The Information School  
University of Washington  
batya@uw.edu

## Abstract

In this paper, we propose data statements as a design solution and professional practice for natural language processing technologists, in both research and development. Through the adoption and widespread use of data statements, the field can begin to address critical scientific and ethical issues that result from the use of data from certain populations in the development of technology for other populations. We present a form that data statements can take and explore the implications of adopting them as part of regular practice. We argue that data statements will help alleviate issues related to exclusion and bias in language technology, lead to better precision in claims about how natural language processing research can generalize and thus better engineering results, protect companies from public embarrassment, and ultimately lead to language technology that meets its users in their own preferred linguistic style and furthermore does not misrepresent them to others.

## 1 Introduction

As technology enters widespread societal use it is important that we, as technologists, think critically about how the design decisions we make and systems we build impact people—including not only users of the systems but also other people who will be affected by the systems without directly interacting with them. For this paper, we focus on natural language processing (NLP) technology. Potential adverse impacts include NLP systems that fail to work for specific subpopulations (e.g.,

There are both scientific and ethical reasons to be concerned. Scientifically, there is the issue of generalizability of results; ethically, the potential for significant real-world harms. Although there is increasing interest in ethics in NLP,<sup>1</sup> there remains the open and urgent question of how we integrate ethical considerations into the everyday practice of our field. This question has no simple answer, but rather will require a constellation of multi-faceted solutions.

Toward that end, and drawing on value sensitive design (Friedman et al., 2006), this paper contributes one new professional practice—called **data statements**—which we argue will bring about improvements in engineering and scientific outcomes while also enabling more ethically responsive NLP technology. A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. In developing this practice, we draw on analogous practices from the fields of psychology and medicine that require some standardized information about the populations studied (e.g., APA, 2009; Moher et al., 2010; Furler et al., 2012; Mbuagbaw et al., 2017). Though the construct of data statements applies more broadly, in this paper we focus specifically on data statements for NLP systems. Data statements should be included in most writing on NLP including: papers presenting new datasets, papers reporting experimental work with datasets, and documentation for NLP systems. Data statements should

# Hate Speech Twitter annotations

Here we provide a data set of tweets which have been annotated for hate speech.

We provide the ID and the annotation in a tab separated file (annotation.tsv). To obtain this individual dataset, the project will need to find the right balance, but this Twitter API of your choice and query for the ID's provided.

If using NAACL\_SRM\_2016.csv please cite using:

```
@InProceedings{waseem-hovy:2016:W16-2,  
  author = {Waseem, Zeerak and Hovy, Dirk},  
  title = {Hateful Symbols or Hateful People? Predictive Feature  
  booktitle = {Proceedings of the NAACL Student Research Workshop},  
  month = {June},  
  year = {2016},  
  address = {San Diego, California},  
  publisher = {Association for Computational Linguistics},  
  pages = {88–93},  
  url = {http://www.aclweb.org/anthology/W16-2013  
}
```

If using NLP+CSS\_2016.csv please cite using:

```
@InProceedings{waseem:2016:NLPandCSS,  
  author = {Waseem, Zeerak},  
  title = {Are You a Racist or Am I Seeing Things? Annotator Inf  
  booktitle = {Proceedings of the First Workshop on NLP and Computat
```

## 5 Proposed Data Statement Schema

We propose the following schema of information to include in long and short form data statements.

<sup>1</sup>A notable exception is Derczynski et al. (2016), who present a corpus of tweets collected to sample diverse speaker communities (location, type of engagement with Twitter, at diverse points in time (time of year, month, and day), and annotated with named entity labels by crowdworker annotators from the same locations as the tweet authors.

<sup>2</sup>Other datasets can be retrieved with suitable long-form data statements published on project Web pages or archives.

(1966), as speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). Transfer from native languages (L1) can affect the language produced by non-native (L2) speakers (Ellis, 1994, Ch. 8). A further important type of variation is disordered speech (e.g., dysarthria). Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socioeconomic status
- Number of different speakers represented
- Presence of disordered speech

<sup>3</sup><http://tools.linf.org/rfc/bop/bop41.txt>.

500

D. ANNOTATOR DEMOGRAPHIC What are the demographic characteristics of the annotators and annotation guideline developers? Their own “social address” influences their experience with language and thus their perception of what they are annotating. Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socioeconomic status
- Training in linguistics/other relevant discipline

E. SPEECH SITUATION Characteristics of the speech situation can affect linguistic structure and patterns at many levels. The intended audience of a linguistic performance can also affect linguistic choices on the part of speakers.<sup>10</sup> The time and place provide broader context for understanding how the texts collected relate to their historical moment and should also be made evident in the data statement.<sup>11</sup> Specifications include:

- Time and place
- Modality (spoken/signed, written)
- Scripted/edited vs. spontaneous
- Synchronous vs. asynchronous interaction
- Intended audience

F. TEXT CHARACTERISTICS Both genre and topic influence the vocabulary and structural characteristics of texts (Biber, 1995), and should be specified.

G. RECORDING QUALITY For data that include audiovisual recordings, indicate the quality of the recording equipment and any aspects of the recording situation that could impact recording quality.

H. OTHER There may be other information of relevance as well (e.g., the demographic characteristics of the curators). As stated earlier, this is intended as a starting point and we anticipate

## 5.2 Short Form

Short form data statements should be included in any application using a dataset for training, tuning, or testing a system and may also be appropriate for certain kinds of system documentation. The short form data statement does not replace the long form one, but rather should include a pointer to it. For short form data statements, we envision 60–100 word summaries of the description included in the long form, covering most of the main points.

## 5.3 Summary

We have outlined the kind of information data statements should include, addressing the needs laid out in §3, describing both long and short versions. As the field gains experience with data statements, we expect to see a better understanding of what to include as well as best practices for writing data statements to emerge.

Note that full specification of all of this information may not be feasible in all cases. For example, in datasets created from Web text, precise demographic information may be unavailable. In other cases (e.g., to protect the privacy of annotators) it may be preferable to provide ranges rather than precise values. For the description of demographic characteristics, our field can look to others for best practices, such as those described in the American Psychological Association's *Manual of Style*.

It may seem redundant to reiterate this information in every paper that makes use of well-trodden datasets. Nonetheless, it is critical to consider the data anew each time to ensure that it is appropriate for the NLP work being undertaken and that the results reported are properly contextualized. Note that the requirement is not that datasets be used only when there is an ideal fit between the dataset and the NLP goals but rather that the characteristics of the dataset be examined in relation to the NLP goals and limitations be reported as appro-



# Quick Overview – Model Agnostic methods

---

Model agnostic explainability methods are mainly separated into 5 different types:

- + Types that use the model predictions to gain insights (global or per sample) such as : PDP, ICE, ALE
- + Feature importance and interaction
- + Surrogate
- + Anchors
- + Shapley Values



# PDP, ICE, ALE



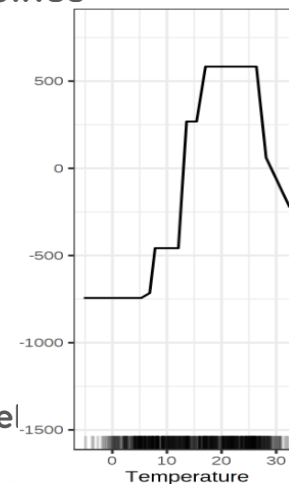
Using a manipulation and a mathematical operation on the model predictions to understand how the model behaves for specific feature (or two). For example, ALE result on how temperature influences bikes rental.

## + Advantages:

- + Intuitive for understanding
- + Easy implementation

## + Disadvantages:

- + The maximum number of features to analyze is 2
- + If features are dependent, most of these methods (except ALE) will not work well
- + Some of these methods (PDP, ALE) hide heterogeneous effects



# Features importance and interaction



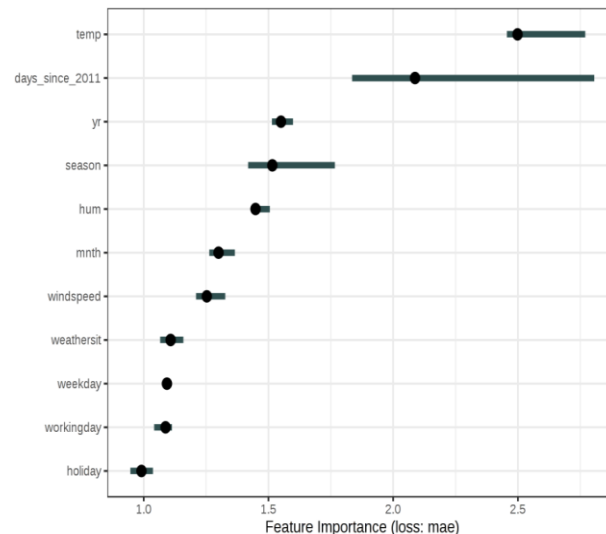
A measure of how much a specific feature is important to the prediction of the model and how much it interacts with other features

## + Advantages:

- + The H-statistic (interaction) has a meaningful interpretation
- + With the H-statistic it is also possible to analyze arbitrary higher interactions such as the interaction strength between 3 or more features
- + Nice and easy interpretation

## + Disadvantages:

- + Computationally expensive
- + Need access to the label (importance)
- + Can produce bias if features are correlated



# Surrogate



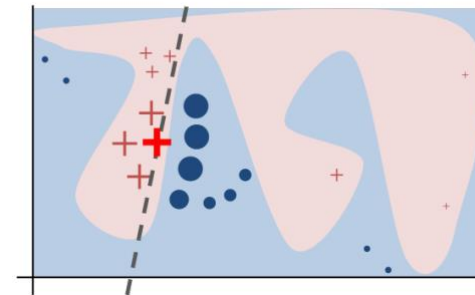
Global or local (LIME, for example) surrogate models are interpretable models that were trained to have predictions close to the black box model

## + Advantages:

- + Can measure how good the surrogate fits your model
- + LIME works for tabular data, text and images (using super pixels)
- + Have good python implementation

## + Disadvantages:

- + The neighborhood definition for tabular data can change results and should be handled carefully
- + Instability of the explanations - sampling twice might yield different results
- + In some cases, the interpretable model is very close for one subset of the dataset, but widely divergent for another subset (Global surrogate)



# Anchors



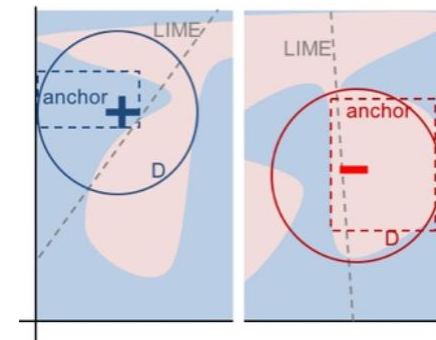
Anchors explains individual predictions of any black-box classification model by finding a decision rule that “anchors” the prediction sufficiently. Anchors approach constructs explanations whose coverage is adapted to the model’s behavior and clearly express their boundaries. Thus, they are faithful by design and state exactly for which instances they are valid

## + Advantages:

- + Easy interpretation (rules)
- + Highly efficient as it can be parallelized
- + Anchors are subsettable and even state a measure of importance by including the notion of coverage
- + The anchors approach works when model predictions are non-linear or complex in an instance’s neighborhood

## + Disadvantages:

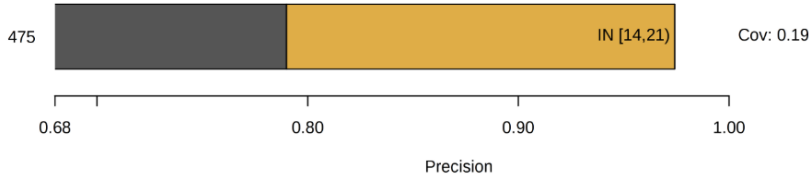
- + The algorithm suffers from a highly configurable and impactful setup
- + Many scenarios require discretization
- + Runtime complexity rises fast with number of features
- + The notion of coverage is undefined in some domains (images)



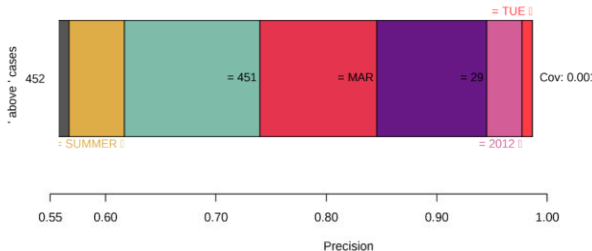


# Anchors example

✦ Homogenous instance example (above average):



✦ Decision boundary instance example



# Shapley Values



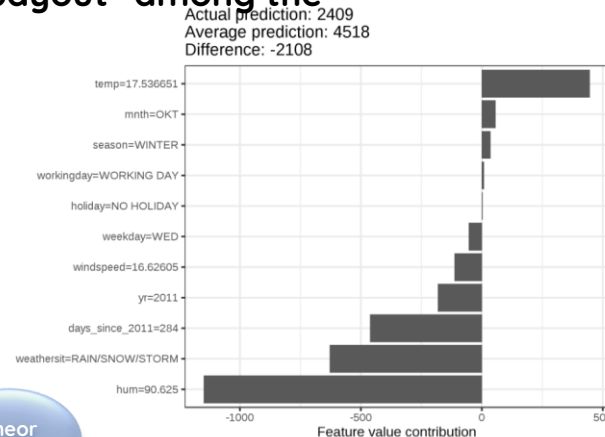
A prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction is the payout. Shapley values – a method from coalitional game theory – tells us how to fairly distribute the “payout” among the features.

## ✦ Advantages:

- ✦ The efficiency property of Shapley values means a fair distribution of the prediction explanation.
- ✦ Legally compliant - the Shapley value might be the only method to deliver a full explanation.
- ✦ The Shapley value is the only explanation method with a solid theory

## ✦ Disadvantages:

- ✦ Computational time
- ✦ Inclusion of unrealistic data instances when features are correlated



Theory



# Shap

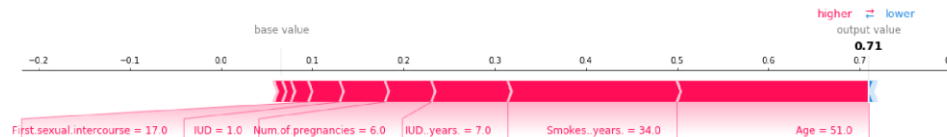
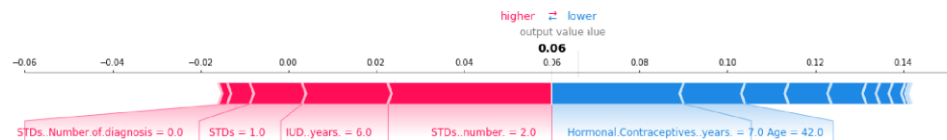


SHAP is based on Shapley values. SHAP authors proposed KernelSHAP and TreeSHAP:

- + KernelSHAP - alternative, kernel-based estimation approach for Shapley values inspired by LIME
- + TreeSHAP - an efficient estimation approach for tree-based models

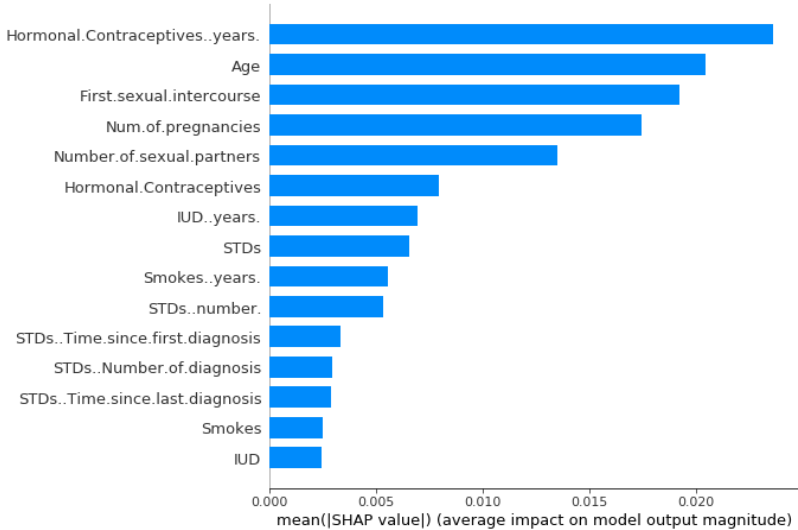
Theor  
y

# Shap Values

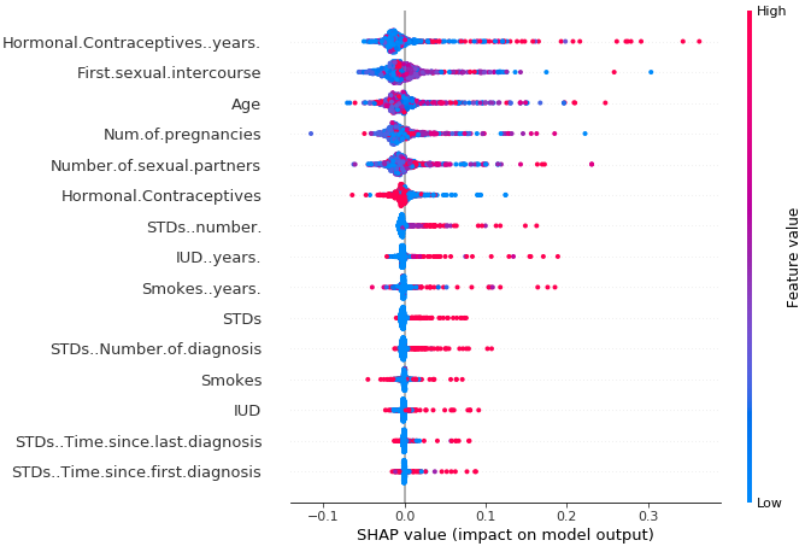


# Shap Feature importance and summary

## SHAPE FEATURE IMPORTANCE

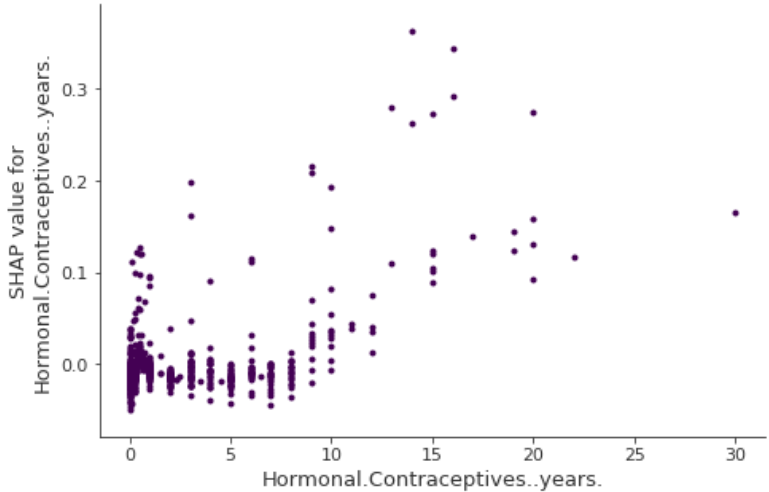


## Summary plot

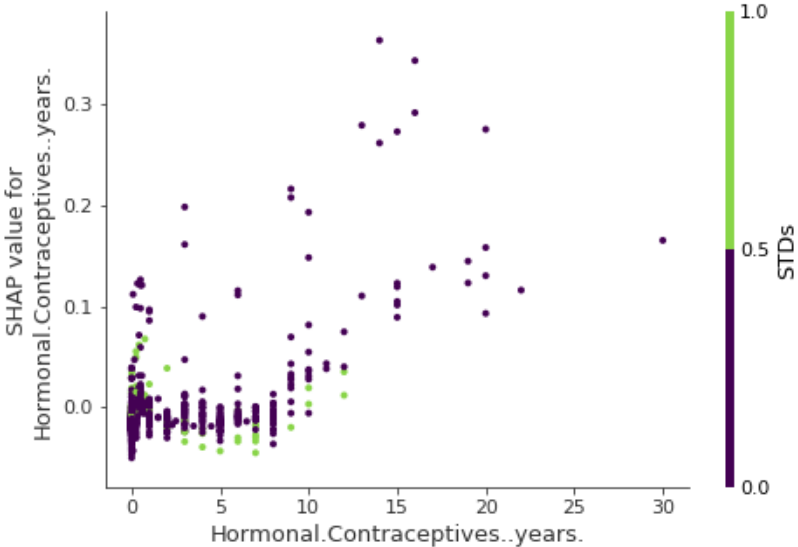


# Shap Dependence and INTERACTIONS

## SHAP DEPENDENCE PLOT



## Interaction values plot



# Implementation

---



You can find open source packages (python) for most of the methods we discussed today:

- + SHAP - shap package that works for tree-based models in scikit-learn. lightGBM, CatBoost and XGBoost also integrated shap
- + LIME - lime package
- + Feature importance, PDP, ICE - Skater