Mind the gap! The what, why and how of data bias, why you should avoid it and how you can save money or lives if you do.

Dipock Das
Senior Director, PM | Splunk

splunk> .conf19

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results may differ materially. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only, and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionalities described or to include any such feature or functionality in a future release.

In this session ...

- What is data bias?
- Why you should avoid it
- Why and how does data bias occur?
- How can you mitigate bias?
- References and enlightened reading

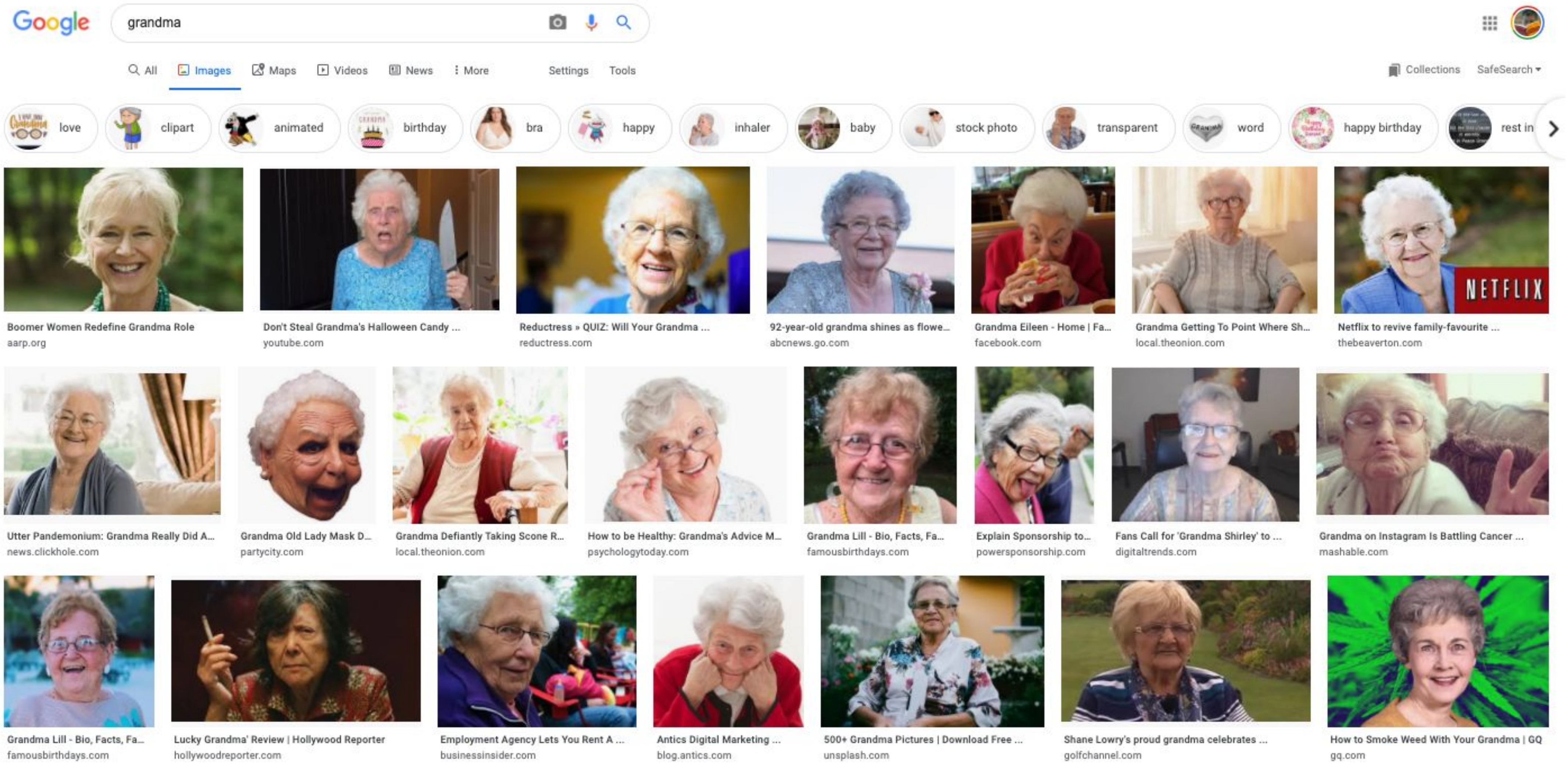
Disclaimer

- The objective of the presentation is to help you understand how biased data and decisions represent significant risks to your organization both monetarily and ethically and ultimately may impact your ability to achieve revenue goals and maintain brand reputation.
- This presentation uses real world examples to validate the points that some might find uncomfortable or disturbing
- There are many slides in this presentation, too much to cover in the time we have - so I plan to skip some of the detailed sections but have included for your benefit later
- This presentation is in no way trying to convey any specific gender, race, religious, social, economic or political point of view

What is data bias?

"No, no! The adventures first, explanations take such a dreadful time."
Lewis Carroll, Alice in Wonderland

“Data bias is all around us, Neo”



What is Data Bias?

Actually we're talking about two things

Statistical bias

- the tendency of a measurement process to over- or under-estimate the value of a population parameter
- the difference between the expected value of an estimator and its estimand
- or in plain English ..
  - something that causes your data to be non-representative of the true population data
  - results that are systematically off the mark

Algorithmic bias

- Algorithms, despite a reputation for impartiality, reflect statistical and cognitive bias resulting in "opinions" embedded in the maths
- Garbage in, garbage out. Just a lot more efficient ..

Bias doesn't come from AI algorithms

Cognitive bias

- a systematic error in thinking that affects the decisions and judgments that people make
- in many cases people are unaware of the biases they carry

Cognitive bias

Similarity/Confirmation Bias

- we select people that are more similar to us, as opposed to people who appear more different. We listen to people who confirm our beliefs more.

Expedience Bias

- make a decision based on the information that's most readily accessible in the brain (what comes to mind most quickly) instead of taking varied perspectives into account

Experience Bias

- when a fact is known, it is hard to imagine not knowing it. Recognise however there was a time when you did not know the fact - so don't assume others will

Distance Bias

- the tendency to favor people who are closer to us in time and space

Safety Bias

- tendency to avoid loss. Many studies have shown that we would prefer not to lose money even more than we'd prefer to gain money. In other words, bad is stronger than good
- influence any decision about the probability of risk or return, or the allocation of resources including money, time, and people

More biases (lots more …)

- Business Insider article here

How does Cognitive bias impact Statistical Bias?

(Disclaimer: I am not a psychologist, but as an old PM, I seen a few of these tricks..)

Selection bias

- when the data does not accurately represent the population
- feature extraction - variables selected do not reflect the nature of the problem
- undercoverage - when members of the population are inadequately represented
- nonresponse bias - respondents differ in meaningful ways from nonrespondents
- voluntary response bias - when sample members are self-selected volunteers. The resulting sample tends to over represent individuals who have strong opinions

9 out of 10

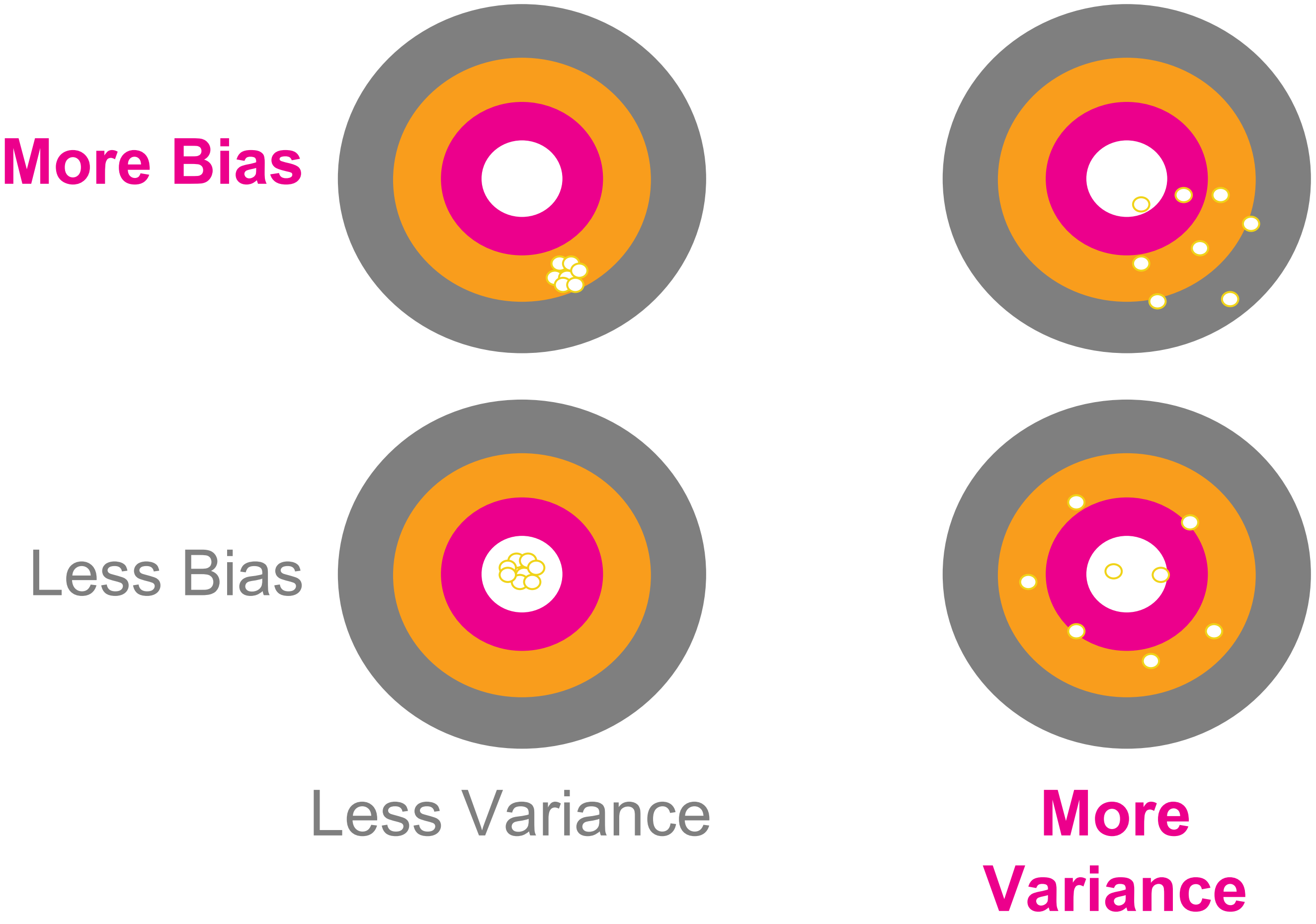
people were fed up with statistics that never published the population size.

Response bias

- a poor measurement process can also lead to bias.
- In survey research, the measurement process includes the environment in which the survey is conducted, the way that questions are asked, and the state of the survey respondent
- Leading questions. The wording of the question may be loaded in some way to unduly favor one response over another.
- Social desirability. Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes

What does Statistical Bias look like?

If Statistical bias was an archery competition



Why you should avoid bias...

“If you drink much from a bottle marked 'poison' it is certain to disagree with you sooner or later.”

Lewis Carroll, Alice in Wonderland

splunk> .conf19

The Working Class

Think of a working class person ...

1. Coal Industry: 53,420 jobs
2. Mostly male
3. Median salary: $59,380

The Real Working Class

Don’t presume what is male is universal

1. Housekeeping/Cleaner: 924,640 jobs
2. Mostly female
3. Median salary: $21,820

The Yorkshire Ripper

How bias lead to multiple mistakes and 7 preventable murders

1975

Begins attacking women.
All three survive and police do not link the attacks. Attacker used hammer and screwdriver in attacks.

One provides a photo-fit with the exact likeness of **Peter Sutcliffe.**

October 30th 1975
Wilma McCann
prostitute (P).

*Police determine the killer is targeting **prostitutes only.***

1976

January 20th
Emily Jackson

May 9th
Marcella Claxton, 20 (P)
Survived attack. Evidence ignored due to **racial** bias.

1977

February 5th
Irene Richardson
April 23rd
Patricia Atkinson
June 26th
Jayne MacDonald
Shop Assistant. *Gains attention of national press*
July 10th
Maureen Long,
Survived attack. A

October 1st
Jean Jordan
Dumps body with new £5 note. Serial number traced to the payroll of hauliers T & W H Clark, **Sutcliffe** *employers.*

November
Sutcliffe interviewed
Fail to check alibi. Fail to examine his car. *Tyres would have linked him* with the murder of Irene Richardson.

1978

January 21st
Yvonne Pearson

January 31st
Helen Rytka

May 16th
Vera Millward

1979

March 2nd
Ann Rooney, 22, Student

April 4th
Josephine Whitaker
Building Society clerk
June
Tape sent to police.
Despite all evidence to the contrary, head of investigation decides tape is genuine. Police stop looking for a local man and focus on a man with the accent on the tape.
July
5th Sutcliffe interview
Voice/handwriting do not match tape or letters.

September 2nd
Barbara Leach
University Student

1980

August 20th
Marguerite Walls
Civil Servant

September 24th
Dr. Upadhya Bandara, 34
Accurate description of Sutcliffe. Evidence dismissed due to lack of stab wounds and attempted use of ligature.
November 5th
Theresa Sykes,
Accurate description of Sutcliffe.
(matches Dr. Bandara)

November 17th
Jacqueline Hill,
University Student

1981

January 3rd
Police arrest **Sutcliffe** on a traffic violation. He is with a prostitute. Discover weapons in car.
Sutcliffe admits he is the Yorkshire Ripper

13 murder victims
28,687 statements
250,000 people questioned

Sutcliffe was questioned and released 9 times

Is snow clearing sexist?

Cars

Ownership

- Women are less likely than men to own cars or drive the car if in a family
- Women more likely to use public transport, walk, ride bikes
- Public transport passengers are women
  - The data is sparse (i.e. the data is just not collected)
  - France (60%), Chicago (62%) Philadelphia (64%)

Think about the journey ...

- Men usually make single trips
  - Drive to the office.
  - Drive home.
- Women make several small interconnected trips
  - “Trip chaining”
  - Dropping kids off at school or childcare before going to work
  - Taking elderly relative to doctor
  - Doing grocery shopping on way home from work

Accidents




- 3x people injured on footpaths than accidents on roads
- 69% women involved in single person accidents
- 60% injured pedestrians fallen in icy conditions
- 79% during winter months
- Increased hospital admissions
- Lost productivity

So is snow clearing sexist?

Karlskoga, Sweden, 2013

Old system

- key roads cleared first
- areas where men worked e.g. construction areas
- areas used by pedestrians and cyclists

Change in priorities

- emphasis on areas frequented by women
- primary schools, places of work
- change did not incur any extra cost
- “easier to drive a car through 3 inches of snow than push a buggy, bike or wheelchair”

Outcome

- Lower hospital admissions
  - 50% drop in accidents (in areas measured)
- Saved money
  - Kr 36m / £3.2m / $3.7m estimated cost of accidents
  - cost of pedestrian accidents = 2x- 3x cost of road maintenance


B

Sexual Harassment On Public Transport

A daily problem for women all over the world

90% French women have been victims of sexual harassment on public transport

Safety of passengers is a priority

- Dramatic changes made to handle terrorism (kind of obvious)
- Example: putting security cameras on buses

However ...

- No changes have been made to ensure the safety of women. Why?
- Example: not putting cameras at bus stops or providing live information on bus arrivals

Because

- Not enough data collected
- Where data is collected it is gender neutral - not disambiguated
- There is insufficient analysis of travel patterns.

When this has been addressed

- Late night buses will drop female passengers off en-route and ensure no one else gets off at the same spot.
- Simple zero cost measure can save lives, increase saftey/confidence and ridership

Let’s talk about Home Insurance !

Premiums calculated using these variables

- Location (exposure to hazards, e.g. storms, wildfires, burglaries)
- Value and Age of building
- Cost to rebuild if completely destroyed + Local construction costs (reflect materials availability and price, regulations)
- Risk exposure (e.g. swimming pool, trampoline, guest house or **aggressive** dog breed)
- **Neighborhood’s** fire protection rating, or, how close your home is to a fire station
- Personal and **neighborhood claims history**, as well as **the previous homeowner claim history**
- Your insurance score, **which is based, in part, on your credit score** (only two states don’t allow this – Maryland and Hawaii)

There’s just a few problems though ...

- Lack of transparency in the calculation
- How often is the model updated or retrained?
- Has your premium ever dropped - even when you have not made a claim?

Most expensive locations by Zip code

ZIP code	State	City	Average annual premium
33050	Florida	Conch Key	$11,702
70091	Louisiana	Venice	$11,151
39595	Mississippi	Pascagoula	$7,922
36561	Alabama	Gulf Shores	$7,850
77550	Texas	Galveston	$7,105

Least expensive locations by Zip code

ZIP code	State	City	Average annual premium
96814	Hawaii	Honolulu	$332
07920	New Jersey	Basking Ridge	$485
83729	Idaho	Boise	$498
05404	Vermont	Winsooki Burlington	$525
97003	Oregon	Beaverton	$532

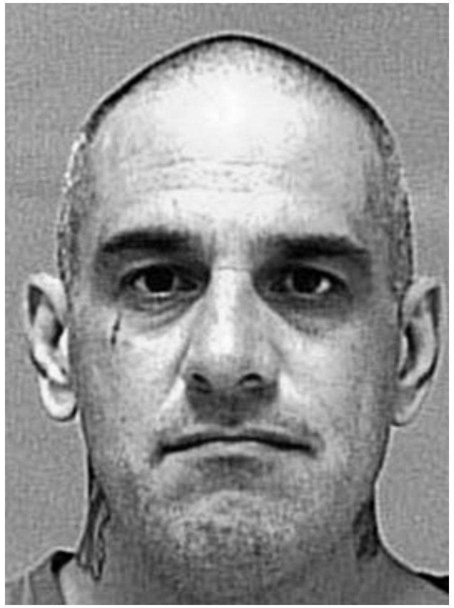
Bias in Risk Assessment For Criminal Sentencing

Propublica.org : software used to predict future criminals, biased against blacks

HIGH RISK: 8

Brisha Borden

- Prior Offenses
  - 4 juvenile misdemeanor
- **Subsequent Offenses**
  - **None**



LOW RISK: 3

Vernon Prater

Prior Offenses
- 2 armed robberies
- 1 attempted armed robbery

- **Subsequent Offenses**
  - **1 grand theft**

The Risk Assessment Questionnaire

4. Based on the screener's observations, Is this person a suspected or admitted gang member?
☐ No ☑ Yes

74. Were you ever suspended or expelled from school?
☐ No ☑ Yes

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

90. How often do you have barely enough money to get by?
☐ Often ☑ Sometimes ☐ Never

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

95. How often did you feel bored?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

55. How often have you moved in the last twelve months?
☐ Never ☑ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

111. "I have never felt sad about things in my life."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

65. Is there much crime in your neighborhood?
☑ No ☐ Yes

124. "If people make me angry or lose my temper, I can be dangerous."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Ag

67. In your neighborhood, have some of your friends or family been crime victims?
☐ No ☑ Yes

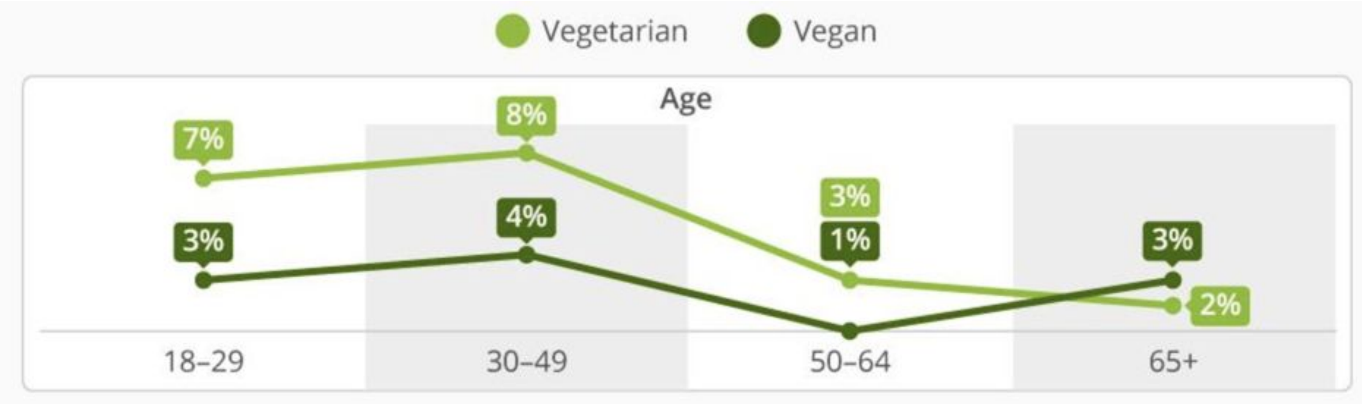
127. "A hungry person has a right to steal."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Similar problems - but this time the implications are serious

- Cognitive bias in the questions and dangerous use of proxies (using other stats)
- Lack of transparency in the model
- How often is the model updated or retrained?

Questionnaire online:
https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

Let’s Talk About Vegans



Gallup 2015 North American survey (1033 adults)

The Vegan Sausage Roll

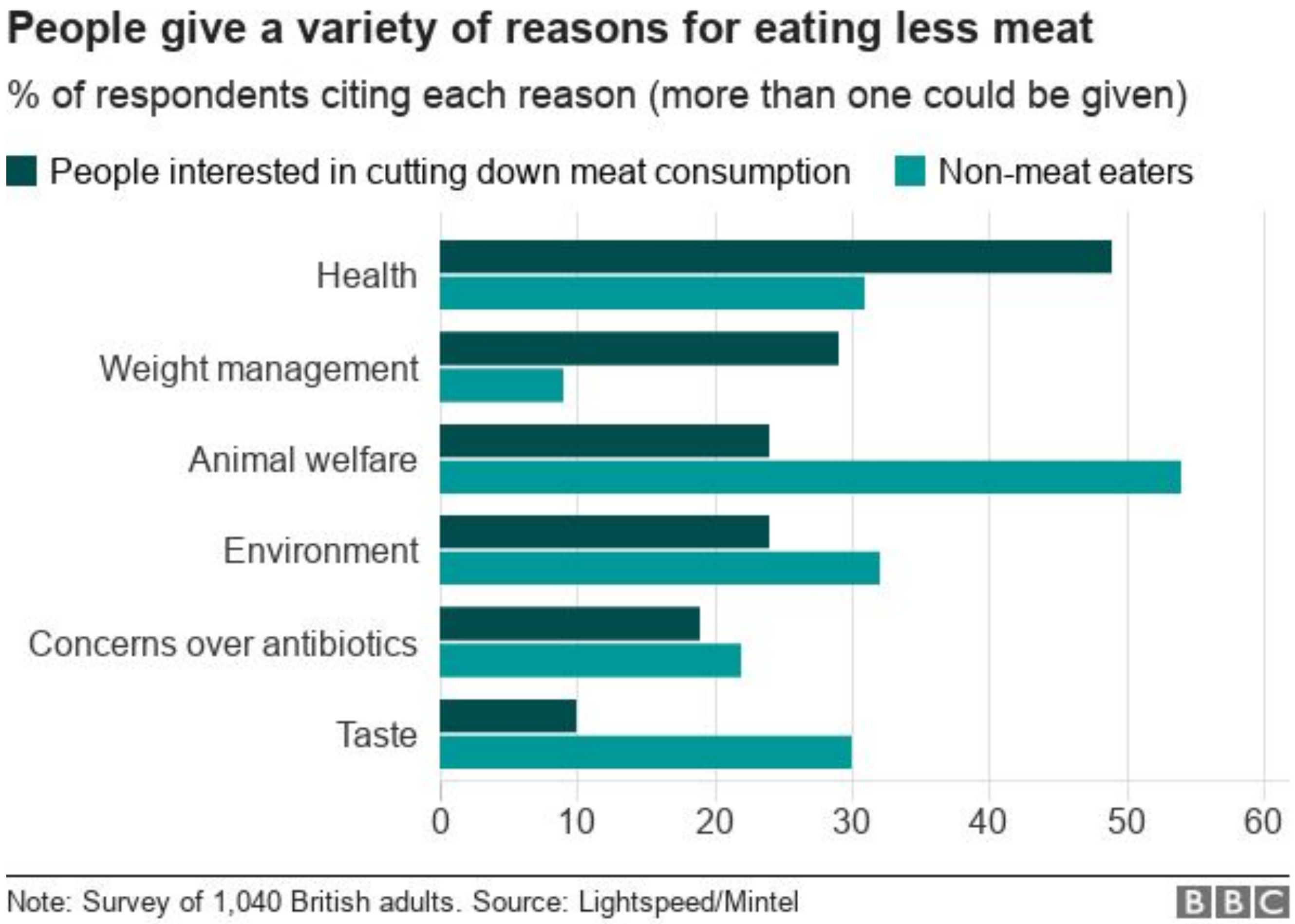
Developed after 20,000 people signed an online petition.
Spent a year innovating before releasing.
(There was even an iPhone-esque unboxing event)

“Nobody was waiting for a vegan bloody sausage, you PC-ravaged clowns”

Piers Morgan

Turns out, you don't have to be vegan ...

It's about asking the right question



Overcoming bias to capture an economic opportunity

- Surge in sales since launch - up 10% in seven weeks to Feb 2018
- Now one of 5 best sellers - but all sales increased
- Profit advice upgraded 3 times in 3 months
- Value of company up to all time high
- Distributed £35m special dividend to shareholders in July

GREGGS ANNOUNCES RELEASE OF VEGAN SAUSAGE ROLLS NATIONWIDE AND TWITTER DESCRIBES IT AS THE 'BEST NEWS EVER'

'I never thought in my life I'd be a regular customer of yours'

Increased sales across the board

- Mainstream adoption
- Uptick in sales
  - New and returning customers
  - Higher price point
  - Increased basket

The New York Times

What Sandwich War? KFC Sells Out of Plant-Based 'Chicken' in Atlanta

In April, Burger King began testing its plant-based Impossible Whopper in St. Louis and later in other markets around the United States.

2 weeks ago

Adweek

How Faux Meat Is Beefing Up Burger King, Subway and Dunkin' Menus

Most recently, Impossible Foods and Burger King announced the national ... Fanning out from beef substitutes, the brand is tackling more items, ...

3 weeks ago

Forbes

Religious Groups Are Giving Plant-Based Eating a Boost

Fast food chains are introducing more and more new plant-based products, including Burger King's Impossible Burger. Raised awareness of ...

2 hours ago

According to Barclays, the alternative meat industry can hit $140 billion over the next decade.

The Negative Impact of Bias

- Biased decisions, data and algorithms represent significant risks to your organization both monetarily and ethically and ultimately may impact your ability to achieve revenue goals and maintain brand reputation
- Disproportionately affects
  - Women
  - Minorities
  - Poor
- Algorithms are efficient, not fair
- Economic opportunities missed

How can you avoid data bias?

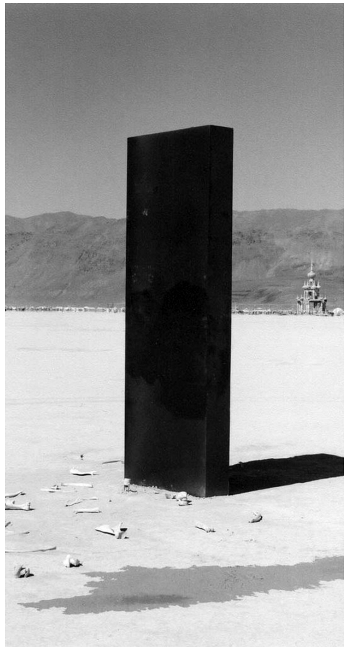
"She generally gave herself very good advice, (though she very seldom followed it)."

Lewis Carroll, Alice in Wonderland

Interpretability of the ML model

Training data
Used by the ML algorithm to configure an output function

New data
Exposed to output function to get prediction



"Black Box"
with output function based on ML algorithm

Prediction
Different types exist

Can I trust this black box?

Stages of Interpretability

Pre-Modelling

Understand/Describe data used to develop models

Explainable Model

Develop explainable model

Post-Modelling

Extract explanations to describe the model

Pre-Modellin g

- Data Collection
- Transparency
- Validation

Data Transparency

- Data disambiguation
- Feature extraction
- Feature selection
- Proxies
- Feedback Loop

Data Transparenc y

Data Disambiguation

- Are you collecting the right data?
- Population - do you disambiguate
  - Gender
  - Race
  - Age

Data Transparency - avoiding Statistical Bias

Avoiding systematic favoritism that is present in the data collection process, resulting in lopsided, misleading results.

Selection bias

- when the survey sample does not accurately represent the population i.e. is not representative

Survey Bias and Response Bias

- arises from problems in the measurement process
- Leading questions. The wording of the question may be loaded in some way to unduly favor one response over another.
- Social desirability. Most people like to present themselves in a favorable light, so they will be reluctant to admit to unsavory attitudes

Measurement error

- A poor measurement process can also lead to bias.
- In survey research, the measurement process includes the environment in which the survey is conducted, the way that questions are asked, and the state of the survey respondent

Sampling Error

- The variability among statistics from different samples is called sampling error.
- Increasing the sample size tends to reduce the sampling error; that is, it makes the sample statistic less variable

Data Transparency

Feature Extraction

- Transform raw data into features suitable for modelling
  - Text (ngrams, word2vec, tf-idf etc)
  - Image (CNN'S, texts, q&a)
  - Geospatial data (lat, long etc)
  - Date and time (day,month,week,year..)
  - Time series, web, etc.
  - Dimensional Reduction Techniques

- Feature transformation
  - Normalization and changing distribution (Scaling)
  - Interactions
  - Filling in the missing values (median filling)

Data Transparency

Feature Selection

- With a large number of variables X, in a model, there will be many that have little or no effect on Y
- Leaving these variables in the model makes it harder to see the "big picture", i.e., the effect of the "important variables"
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables
  - Statistical approaches
  - Selection by modeling
  - Grid search
  - Cross Validation

Data Transparency

Feature selection

- Be careful of the data you have
- And the data you do not have…
- Would gender and race be relevant in this model to project height?

Height (cm) at age 20	Height (cm) at age 10	-	Socks owned at age 10
182	80		20
170	120	…	10
…	…		…
166	145	…	8
152	100		24

Data Transparency

Subset selection and Shrinkage

- Subset Selection
  - Identify a subset of predictors that we believe to be related to the response, and then fitting the model using this subset (e.g. best subset selection and stepwise selection)
- Shrinkage
  - shrink the estimates coefficients towards zero
  - shrinkage reduces the variance
  - Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection
  - E.g. Ridge regression and the Lasso
- Dimension Reduction
  - project all predictors into an M-dimensional space where M < p, and then fitting linear regression model (e.g. Principle Components Regression)

Data Transparency

Proxies

- When a substitute or stand-in for data is used in a model
- Example: zipcode criminal sentencing
- The danger of proxies is that the correlations can be discriminatory or sometimes illegal

Data Transparency

Proxies

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

4.7%
18-24 year olds are Black/Latino males

40.6%
of all Stop and Search

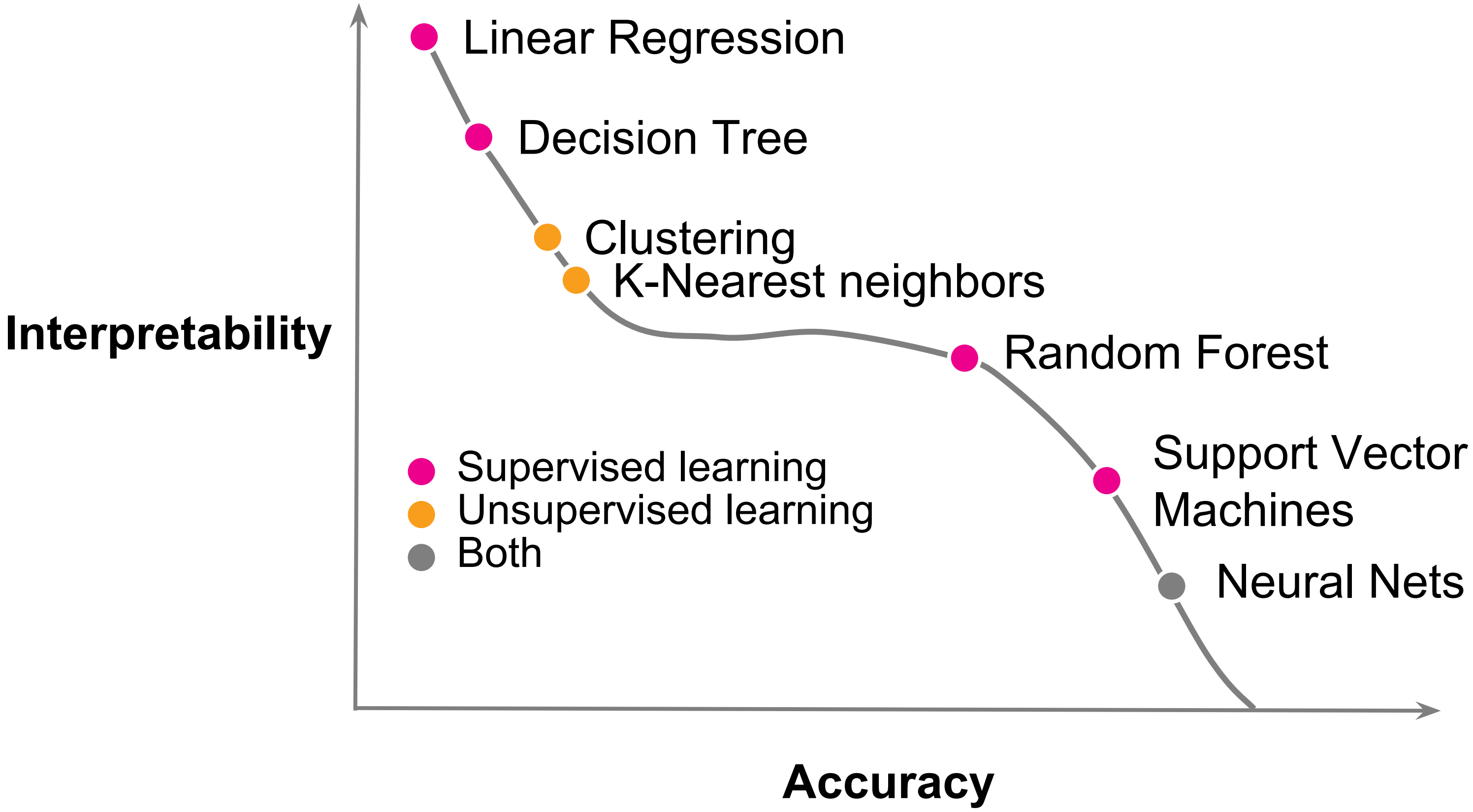
>90%
innocent

Still a fair question?

Model Transparenc y

1. Disclosure of data used
2. Goals of the model
3. What is the definition of success
4. Model Interpretability
5. Model feedback

Accuracy VS interpretabilit y



Interpreting a linear regression model is not as complicated as interpreting Random Forest, Support Vector Machine or Neural Net

Model Transparency

Accuracy VS Interpretability

1. Characteristics of highly interpretable models
   - Linear and Smooth
   - Well defined relationships
   - Easy to compute

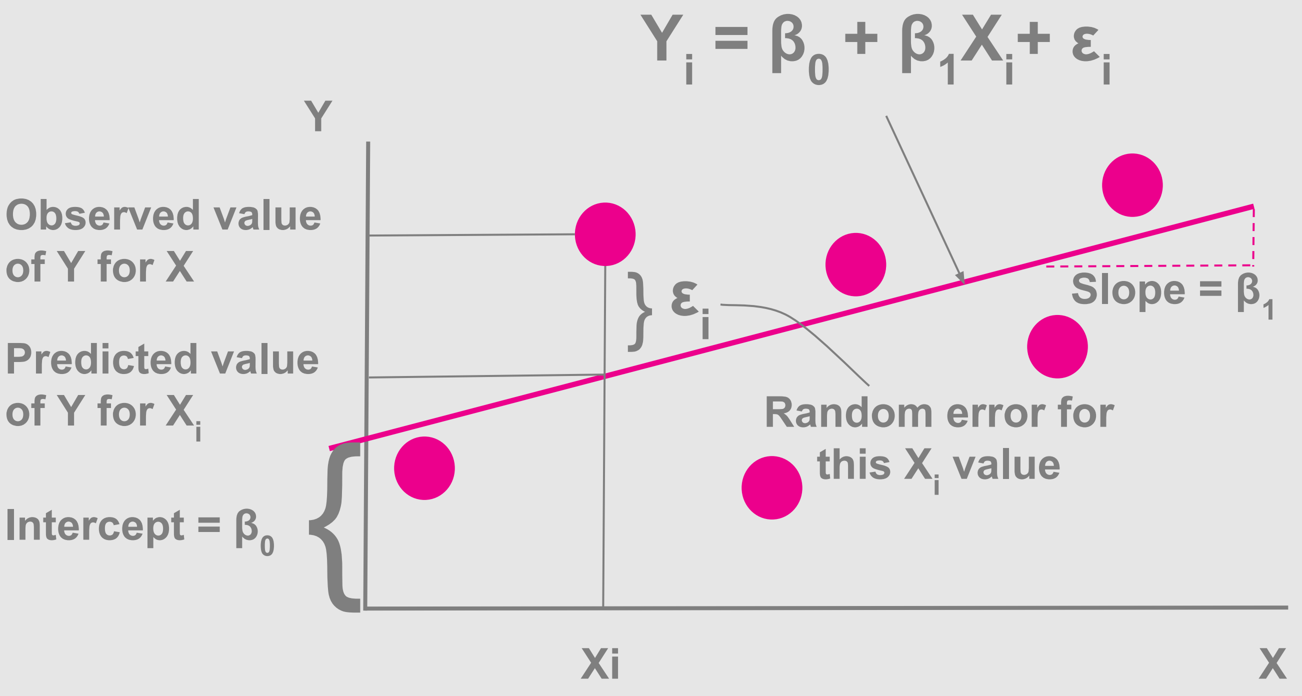
2. Characteristics of highly accurate models
   - Non-Linear relationships
   - Non-Smooth relationships
   - Long computation time

3. Decision trees
   - Provide good accuracy with high interpretability

Model Transparenc y

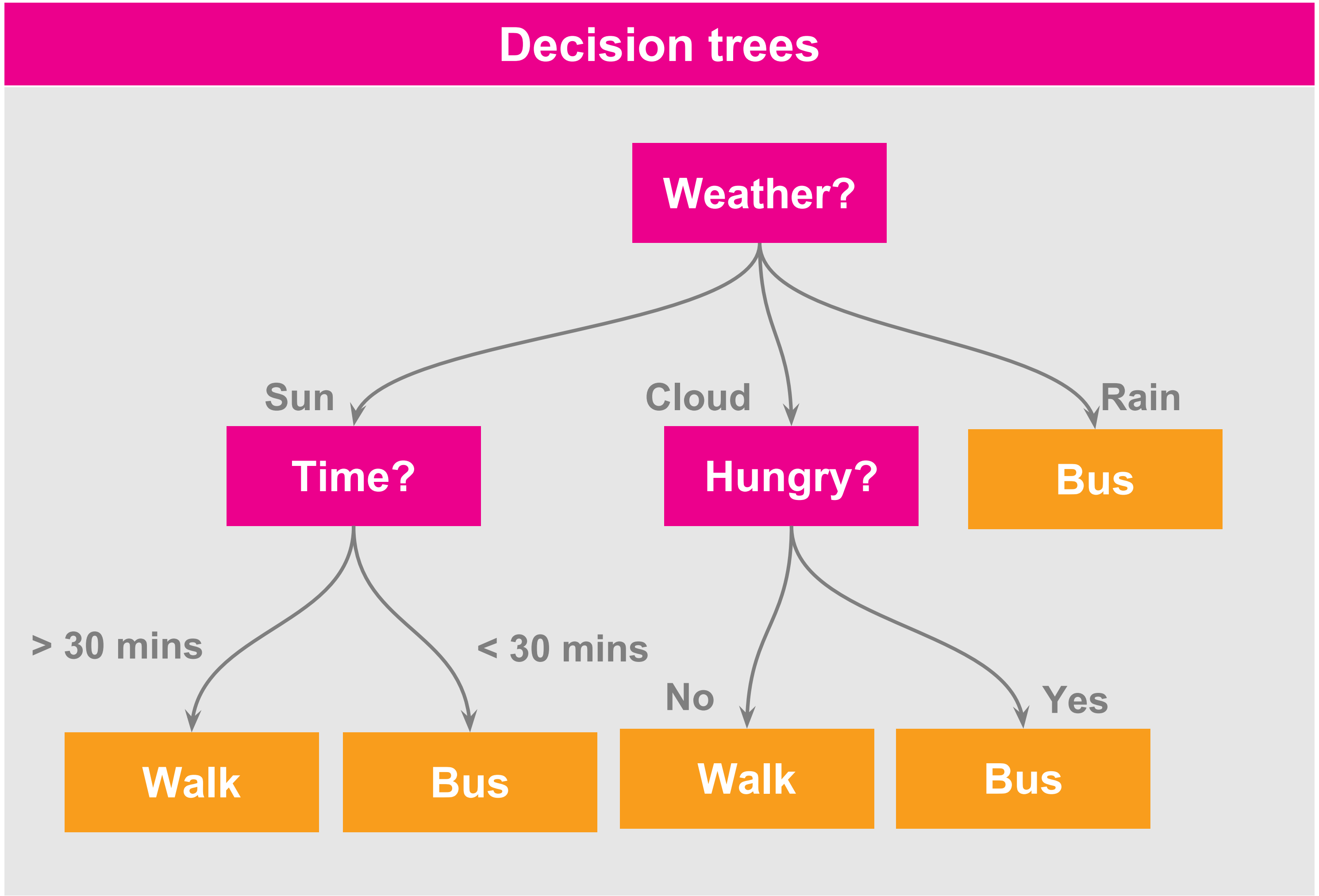
Explainable models

Linear regression



Model Transparency

Explainable models



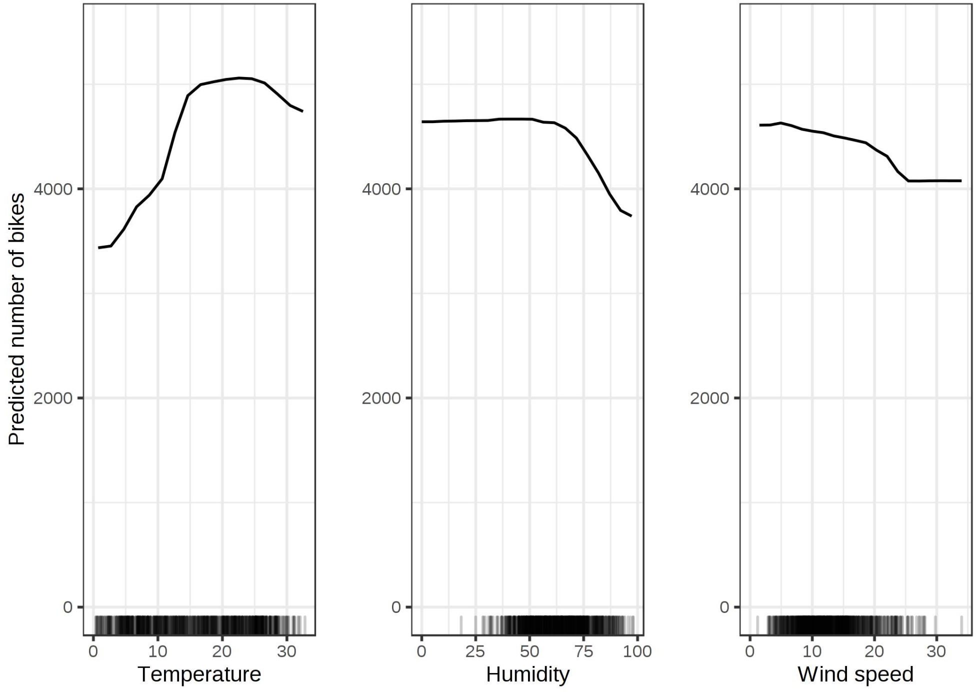
Post-model Transparenc y

Methods for model explainability

- Partial dependence plot (PDP)
- Individual Conditional Expectation
- Local Interpretable Model-agnostic Explanations (LIME)
- Shapley Values

Post-model Transparency

Partial Dependence Plot

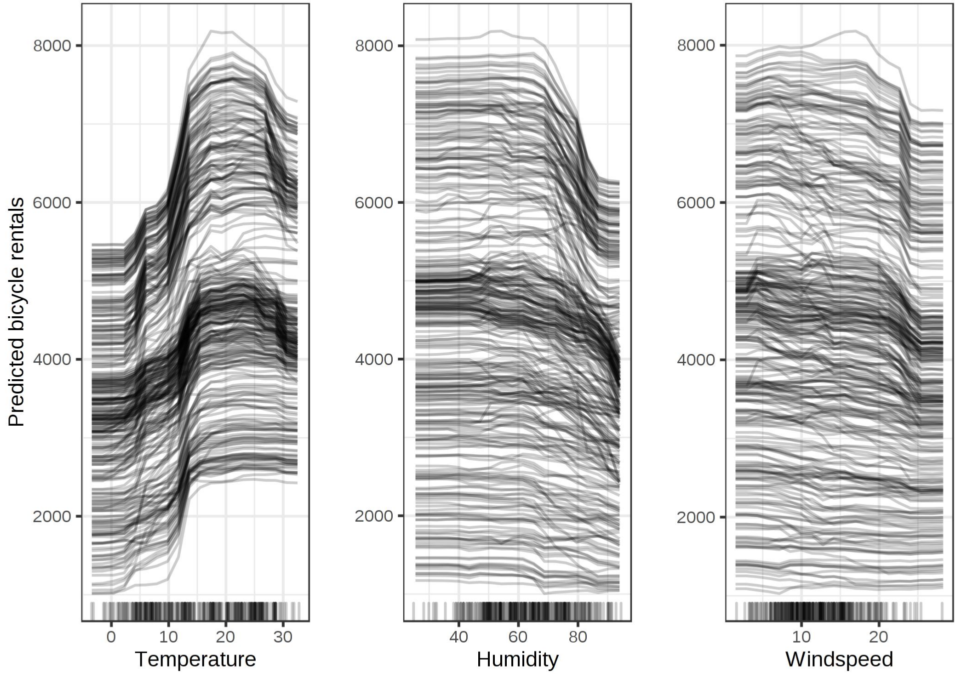


PDP plots of predicted bicycle rentals by weather conditions

- Shows how each variable affects the model's predictions.
- Good for understanding questions like
  - How much of the wage difference between men and women is due solely to gender, as opposed to differences in education backgrounds or work experience?
  - Are health differences between two groups due to differences in their diets, or due to other factors?

Post-model Transparency

Individual Conditional Expectation

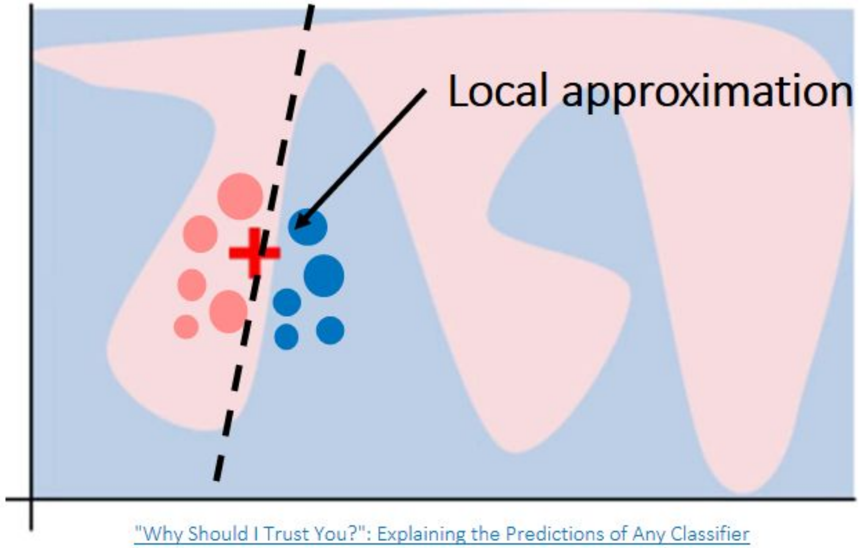


ICE plots of predicted bicycle rentals by weather conditions. The same effects can be observed as in the partial dependence plots.

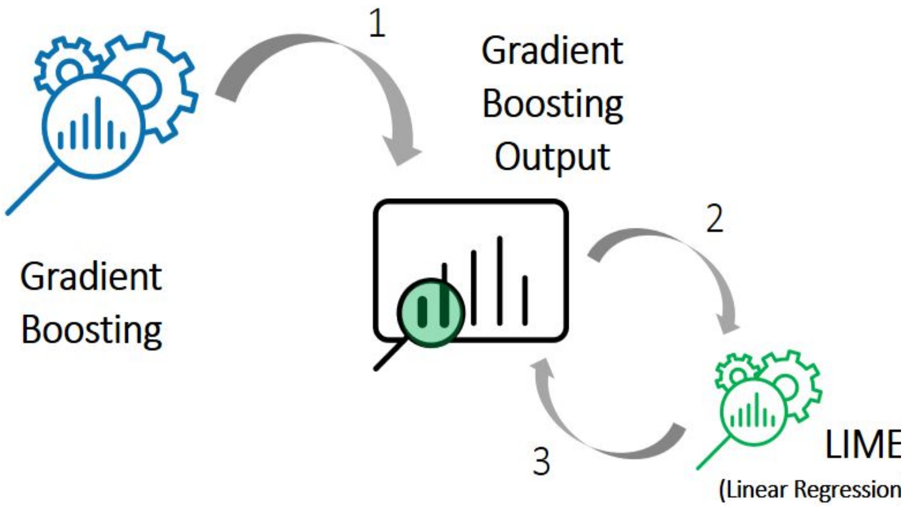
- Displays individual differences so you can identify subgroups and interactions between model inputs
  - Think of each ICE curve as a simulation that shows what would happen to the model's prediction if you varied one characteristic of a particular observation.
  - To avoid visualization overload, ICE plots only show one model variable at a time.

Local Interpretable Model-agnostic Explanations (LIME)

Builds an interpretable model of explanatory data samples at local areas in the analyzed data



"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, ACM SIGKDD, 2016
the red cross is the instance being predicted



Local
A cluster centroid is chosen, along with other data points within close proximity

Interpretable
Builds a linear regression model to fit the data

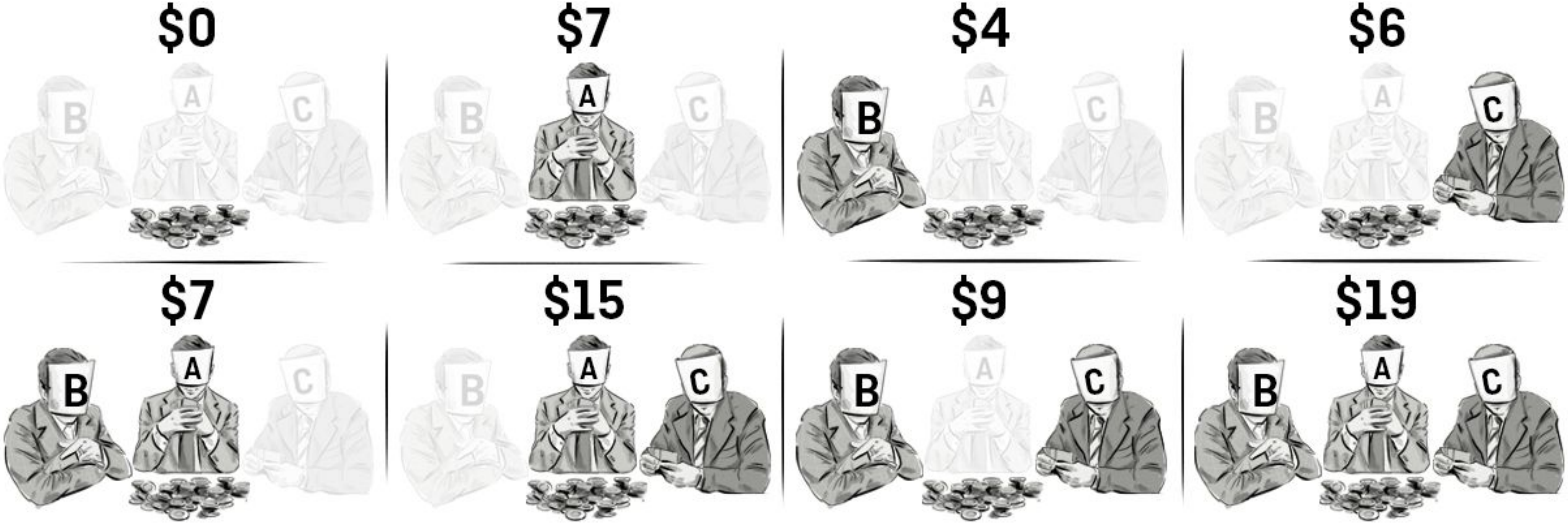
Model agnostic
Builds a linear regression around local points regardless of how the original prediction was generated

Post-model Transparenc y

- Originating from game theory
- In a coalition of multiple players what is the fairest way of dividing the payoff?
- Finding each player’s marginal contribution, averaged over every possible sequence in which the players could have been added to the group

Shapley Values

Find each marginal contribution over every possible sequence



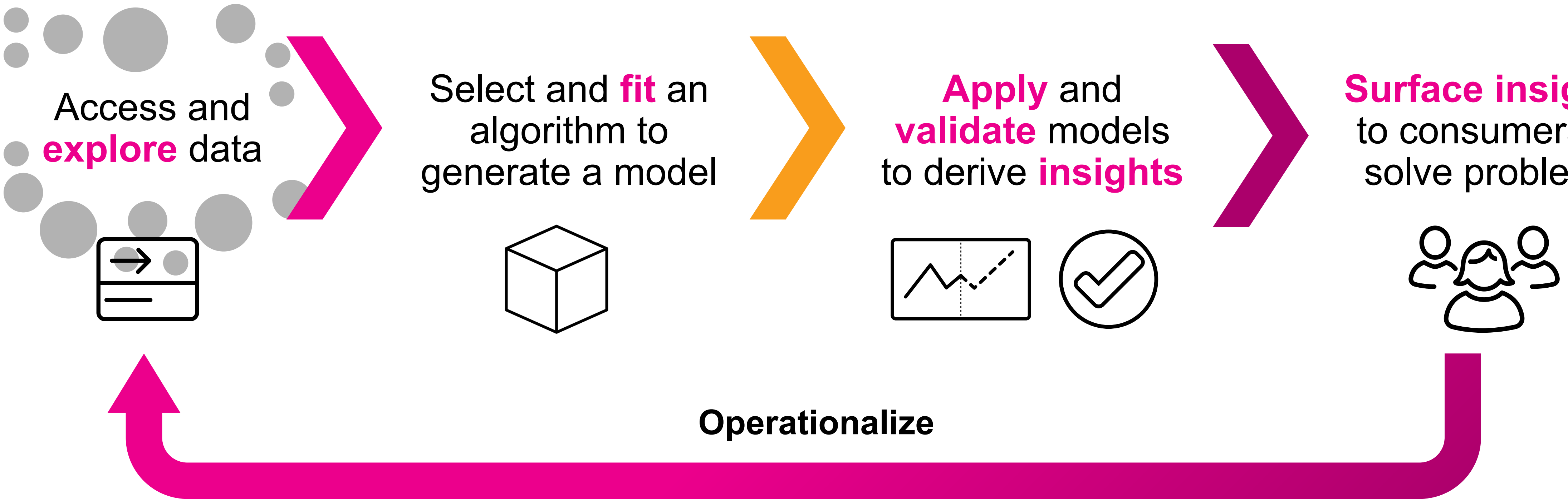
Post-model Transparenc y

Shapley Values

- Apply the same principle to the problem of features contributing to a model
- Instead of a payoff
  - we're evaluating the value of a model
  - with and without feature i added in to some subset of other features
  - and then weighting that subset based on how many sequences it represents
- When the model is linear, or the features are truly independent
  - the problem is a trivial one
  - no matter the values of other features, or the sequence in which features are added to the model,
  - the contribution of a given feature is the same

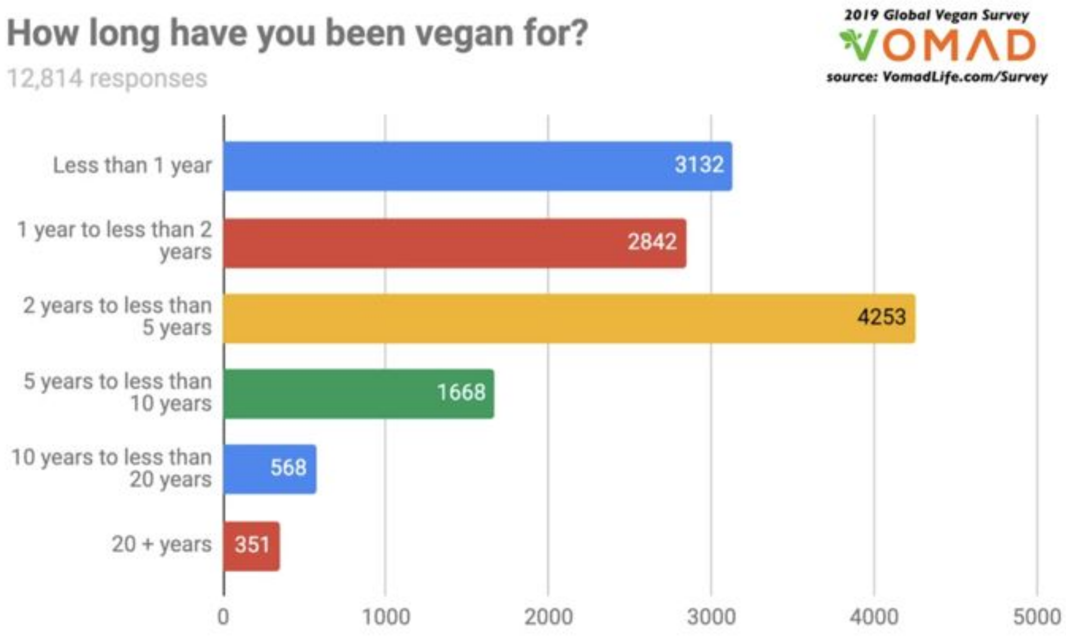
Model Transparency

Feedback loop to tune or retrain the model



Post Model

Transparency of Results



- **33.2%** or almost exactly a third of participants (4,253 people) had been vegan for between 2 and 5 years.
- **24.4%** or 3,132 people went vegan in the previous year.
- **79.8**% or 10,227 people went vegan in the last 5 years.
- **92.8%** or 11,895 people turned vegan in the last 10 years.
- **7.1%** or 919 people who took the survey have been vegan for more than 10 years.
- **2.7%** or 351 people have been vegan for more than 20 years.

Add additional data your charts

Textual description of the results. Data literacy is a problem.

Source: *Why People Go Vegan: 2019 Survey Results*

Validation

Do this throughout the process

1. Get validation from
   - Your department
   - Your organisation
   - Your customers, partners, suppliers
2. (don't forget similarity bias, distance bias etc …)

Data and Model Transparenc y

Interpretable machine learning helps

- Making business decisions
- Detecting Bias
- Debugging and Auditing
- Increasing Social Acceptance
- Legal and Ethical

Why is the women's queue always longer than the mens?

Male Based Design

- Same amount of floor space as dictated by building codes
- Male toilets = 
- Womens = 
- Men enjoy higher rate of relief per sq ft
- So *increase the floor space and allocate the same number of cubicles?*

Hold it ...

An equal number of stalls would not resolve the issue ...

Image credit: Broadway's bathroom problem

- Duration
  - women take twice as long to use a toilet as men
  - about 90 seconds for women, 40 seconds for men
- Think about the journey ...
  - Women make the majority of Elderly + Disabled (er.. and pregnant)
  - Women will take children to the toilet
  - 20%-25% of women of child bearing age on period - so have to change sanitary product

Splunk Machine Learning Toolkit

https://splunkbase.splunk.com/app/2890/

Assistants, Experiments and Showcases

- Guided model building, testing, and deployment for common objectives.
- Interactive examples for typical IT, security, business, and IoT use cases

Algorithms

- 80+ standard algorithms out of the box (supervised and unsupervised)
- Github Community: Share or import algorithms

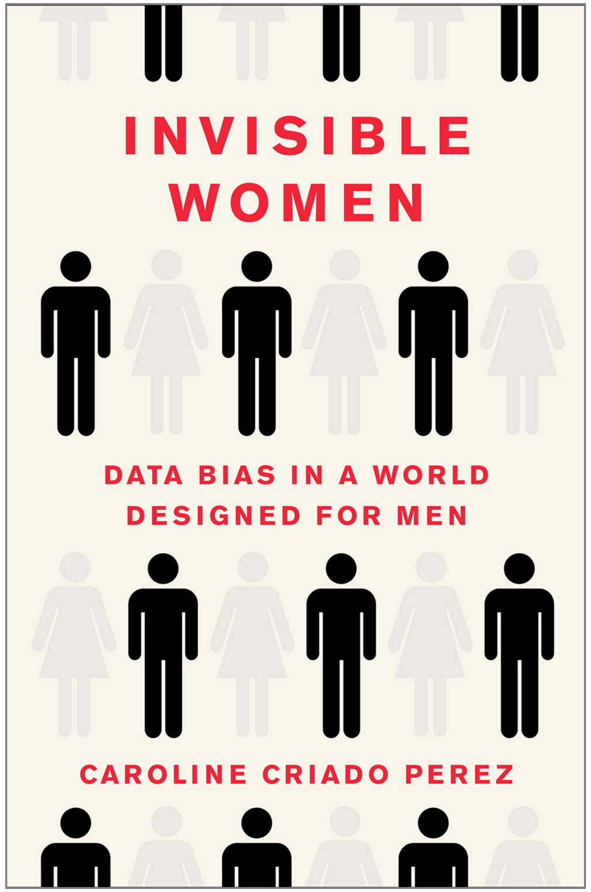
ML Commands and ML API

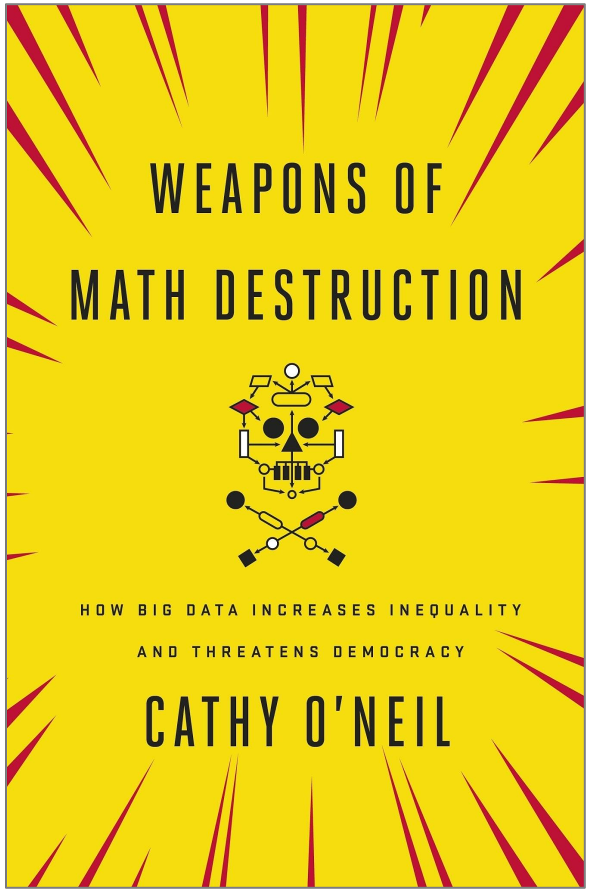
- New SPL commands to fit, test, score and operationalize models
- Extensibility to easily import any algorithm (proprietary / open source)

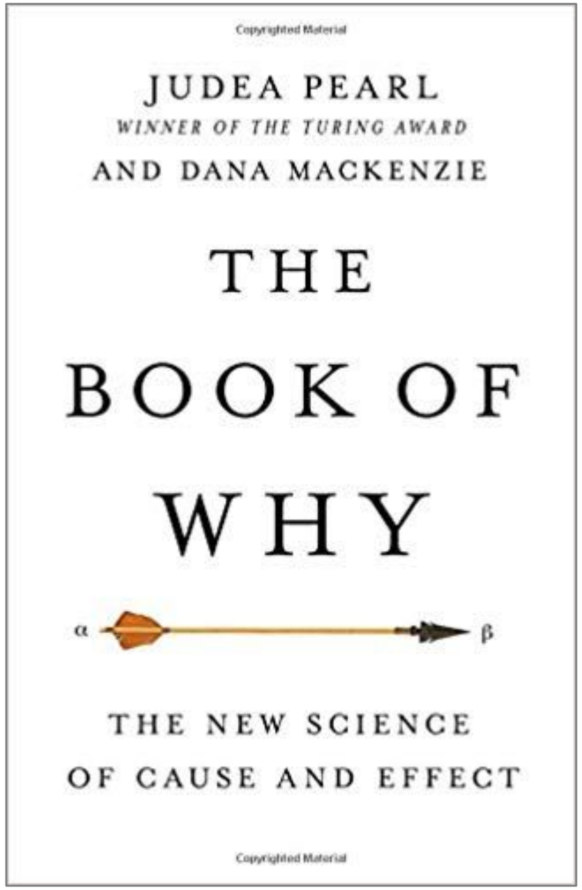
Support for libraries and frameworks

- Python for Scientific Computing Library. Access to 300+ open source algorithms
- Apache Spark MLLib. Support large scale model training via Spark Add-on for MLTK (LAR)
- Tensorflow Container. Supports NN and GPU accelerated ML

Enlightened Reading - Reference Books







References

[Individual Conditional Expectation]

[Partial dependence plots 1]

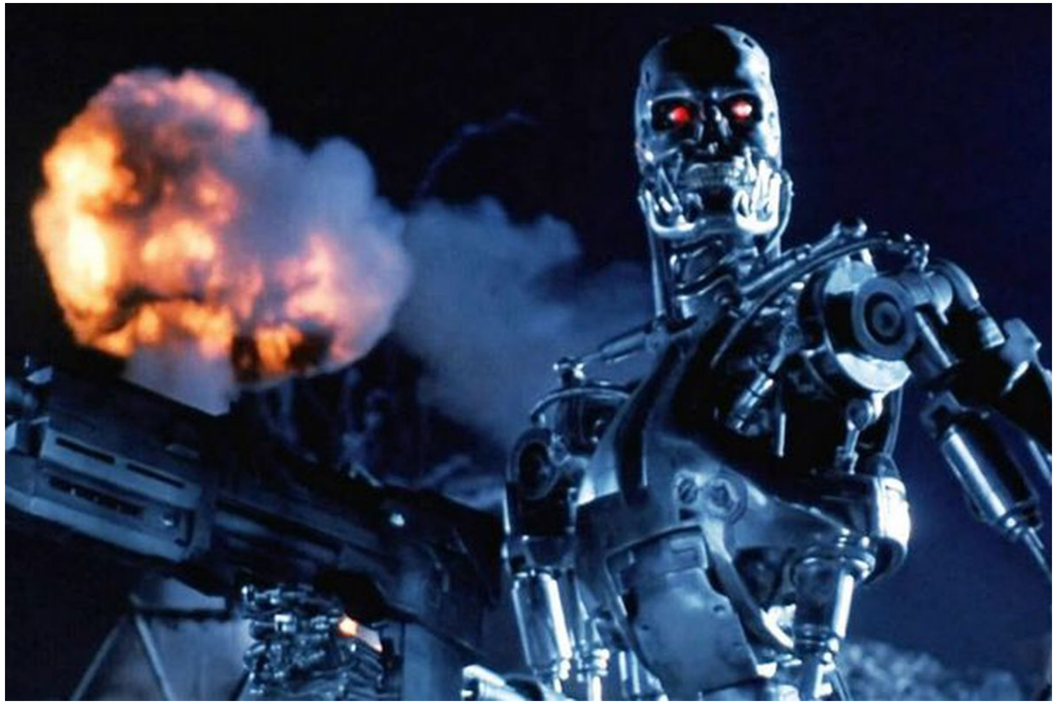
[Partial dependence plots 2]

[Shapley Values 1]

[Shapley Values 2]

Biased data comes from humans ..

- Represents significant financial and ethical risks to your organization
- Disproportionately affects Women, Minorities and the Poor
- Causes economic opportunities to be missed and may impact your ability to achieve revenue goals or maintain brand reputation



Don’t blame the algorithm ..

Dipock Das

Senior Director, PM, Splunk

Thank You!

Go to the .conf19 mobile app to

RATE THIS SESSION