# AI Biases: What they are and how to identify and mitigate them

*Natalia Modjeska*
*Research Director, DnA*
*Info-Tech Research Group*



INFO~TECH
RESEARCH GROUP

# Do we have the necessary skills?

If not yet, can we acquire or grow them over time?


FOR
RENT

# AWS Machine Learning Services Stack

## AI SERVICES

| Vision | | Speech | | | Language | | Chatbots | Forecasting | Recommendations |
|---|---|---|---|---|---|---|---|---|---|
| REKOGNITION IMAGE | REKOGNITION VIDEO | TEXTRACT | POLLY | TRANSCRIBE | TRANSLATE | COMPREHEND | LEX | FORECAST | PERSONALIZE |

## ML SERVICES

AMAZON SAGEMAKER

| BUILD | TRAIN | DEPLOY |
|---|---|---|
| Pre-built algorithms & notebooks | One-click model training & tuning | One-click deployment & hosting |
| Data labeling (GROUND TRUTH) | Optimization (NEO) | |
| Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING) | Reinforcement learning | |

## ML FRAMEWORKS & INFRASTRUCTURE

| Frameworks | Interfaces | Infrastructure | | | | |
|---|---|---|---|---|---|---|
| TensorFlow | GLUON | EC2 P3 & P3N | EC2 C5 | FPGAs | GREENGRASS | ELASTIC INFERENCE |
| mxnet | Keras | | | | | |
| PYTORCH | | | | | | |

AWS re:Invent

Info-Tech Research Group



Image source: AWS

aws

# All ML is biased
# All AI is biased

# And bias is only going to get worse

"Like relational databases, AI is going to get into every important piece of software."

– Benedict Evans, 2018

Image: Lawrence Livermore National Laboratory

“People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.”

— Pedro Domingos,
”The Master Algorithm”, 2015

# We can’t see bias until the problem is big


**AI biases can harm your organization's reputation, ability to deliver services, and more!**

# And for you personally…

# New vendors are riskier than established ones

- Higher ability to solve problems, agility and innovation, but…
- Less dollars
- Higher risk tolerance
- Less diversity on teams
- Less governance
- Privacy is not a cultural priority it?

# AI/Algorithmic bias

Systematic and repeatable errors in a computer system that create ***unfair*** outcomes, such as privileging one arbitrary group of users over others

Definition: Wikipedia

# “Apple Card investigated after gender discrimination complaints”

(The New York Times)

# “Racial Bias Found in a Major Health Care Risk Algorithm”

(Scientific American)


# “COVID-19 vaccine distribution algorithms may cement health care inequalities”

(VentureBeat)

“UK ditches exam results generated by biased algorithm after student protests

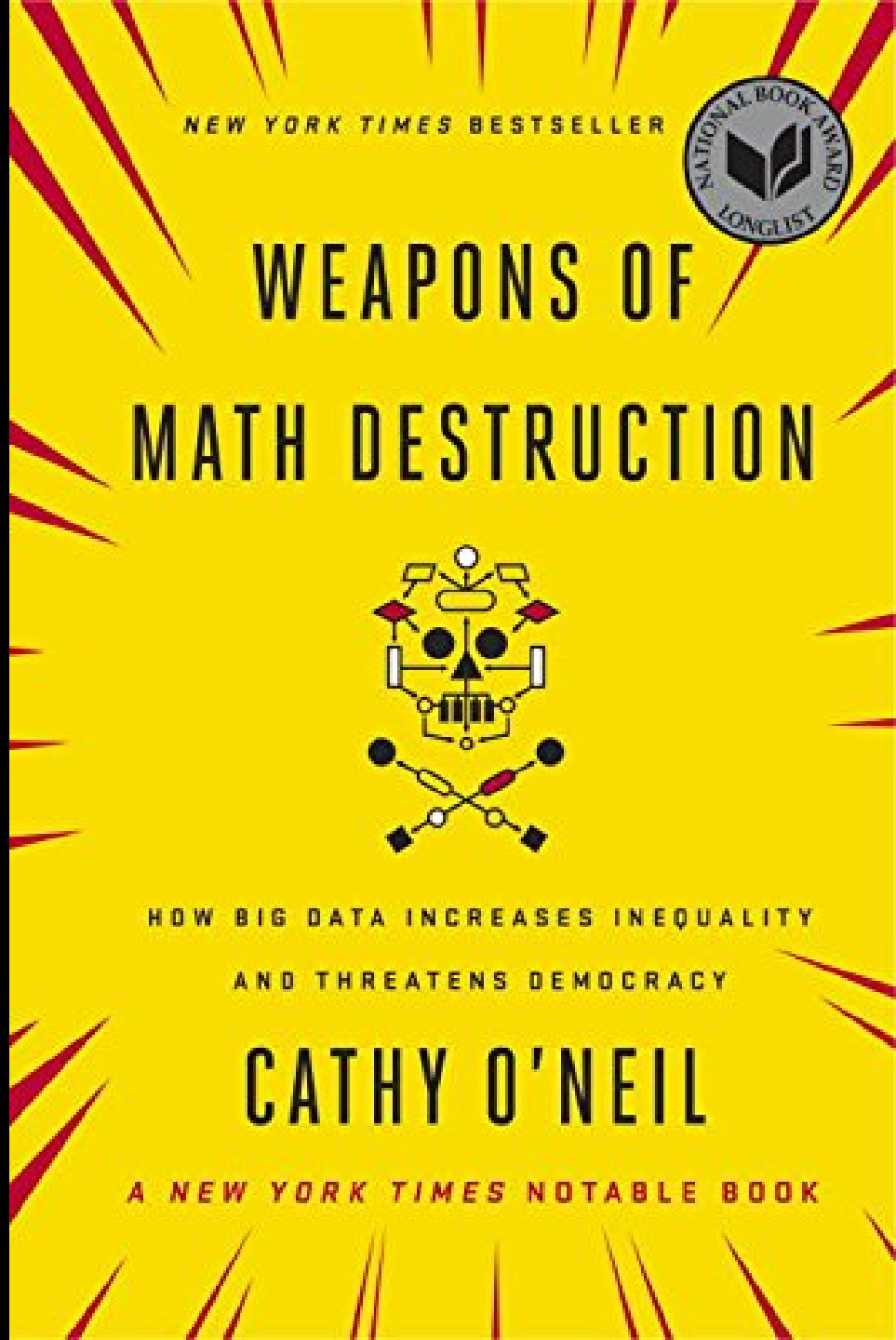
(The Verge)

# “Dutch court prohibits government’s use of AI software to detect welfare fraud”

(The Guardian)

# Biased AI systems are discriminatory

- Title VII of the Civil Rights Act (1964)
- Equal Pay Act (1963)
- Age Discrimination in Employment Act (1967)
- Rehabilitation Act (1973)
- Equal Credit Opportunity Act (1974)
- The Civil Rights Act (1991)
- Fair Housing Act  (1968)
- Genetic Information Nondiscrimination Act (2008)
- GDPR, CCPA…
- Illinois AI Video Interview Act…



# Biased AI will harm state residents


# How do AI/ML systems get biased?

Chef = Data Scientist

Ingredients = Data

Recipe = Algorithm

# AI as a service is a DIY meal kit


# Biases can be introduced at any step of the ML process and they propagate through it

# Data biases, aka ingredients

# Data selection bias

COMPAS Risk Assessment questionnaire (137 questions)

- Was your father [...] ever arrested [...]?
- How many of your friends/ acquaintances have served time in jail or prison?
- How many of your friends/acquaintance are gang members?
- Did a parent [...] have a drug or alcohol problem?
- [...] have some of you fiends or family been crime victims?
- How often do you have barely enough money to get by?

Source: Documentcloud.org

# Data capture bias


Image: Street Bump

***"While massive datasets may feel abstract, they are intricately linked to physical place and human culture."***

– Kate Crawford, 2013

# “Data is destiny” (Joy Buolamwini)

# Remember Amazon's sexist recruiting tool?

**60%** Percentage males in workforce

**74%** Percentage male managers

# Data is a social construct

One in four children will experience some form of abuse or neglect in their lifetimes

Child abuse became academic discipline in the US in the 1970s

**“If we allowed a model to be used for college admissions in 1870, we’d still have 0.7% of women going to college. Thank goodness we didn’t have big data back then!”**

– Cathy O’Neil, 2014

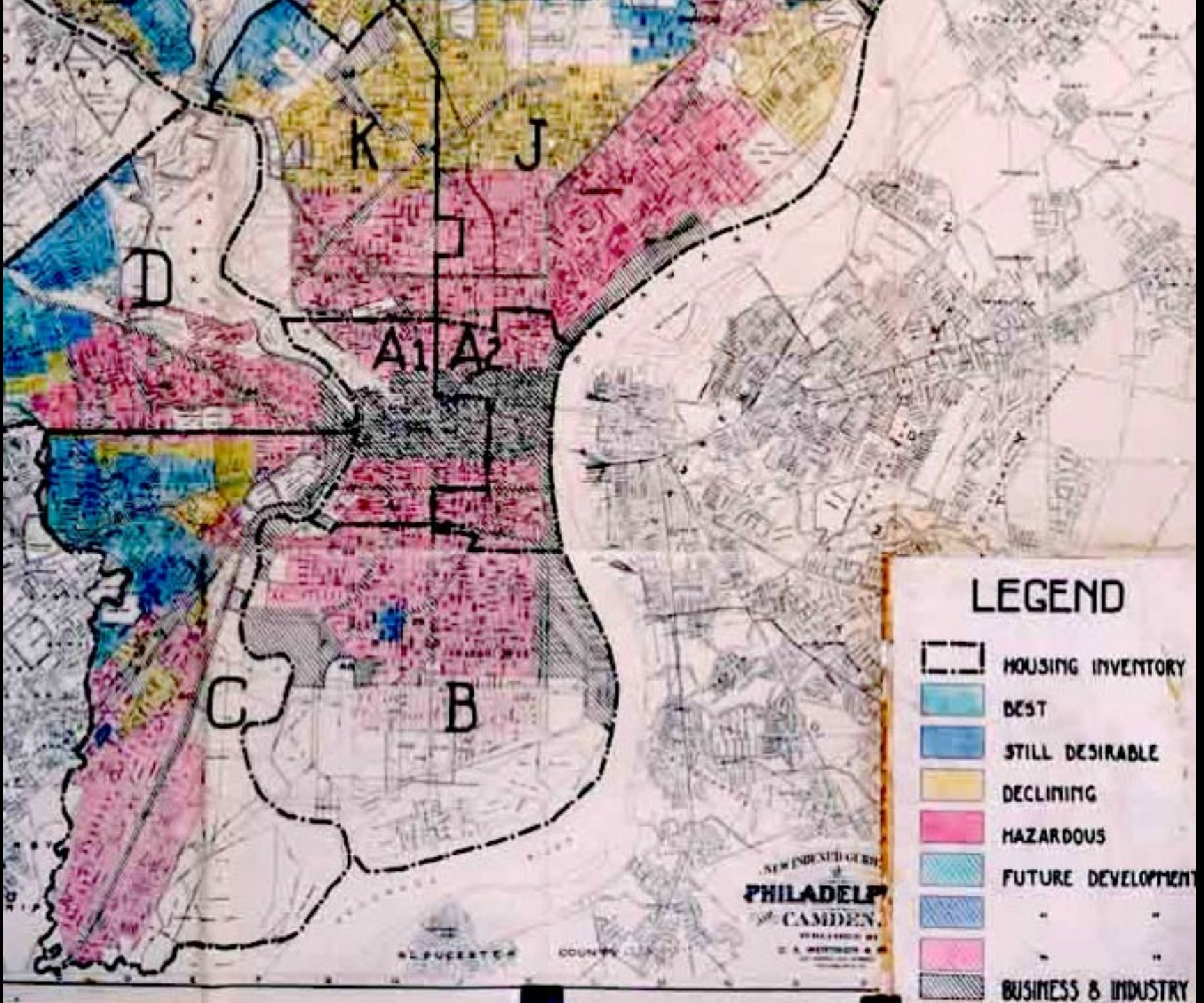
# Are there gaps in your data?

**“We definitely oversample the poor […] All of the data systems we have are biased. We still think this data can be helpful in protecting kids.”**

Erin Dalton, director of Allegheny County’s Office of Data Analysis, Research and Evaluation

***“No algorithm focused on human behavior is neutral. Anything which is trained on historical human behavior embeds and codifies historical and cultural practices.”***

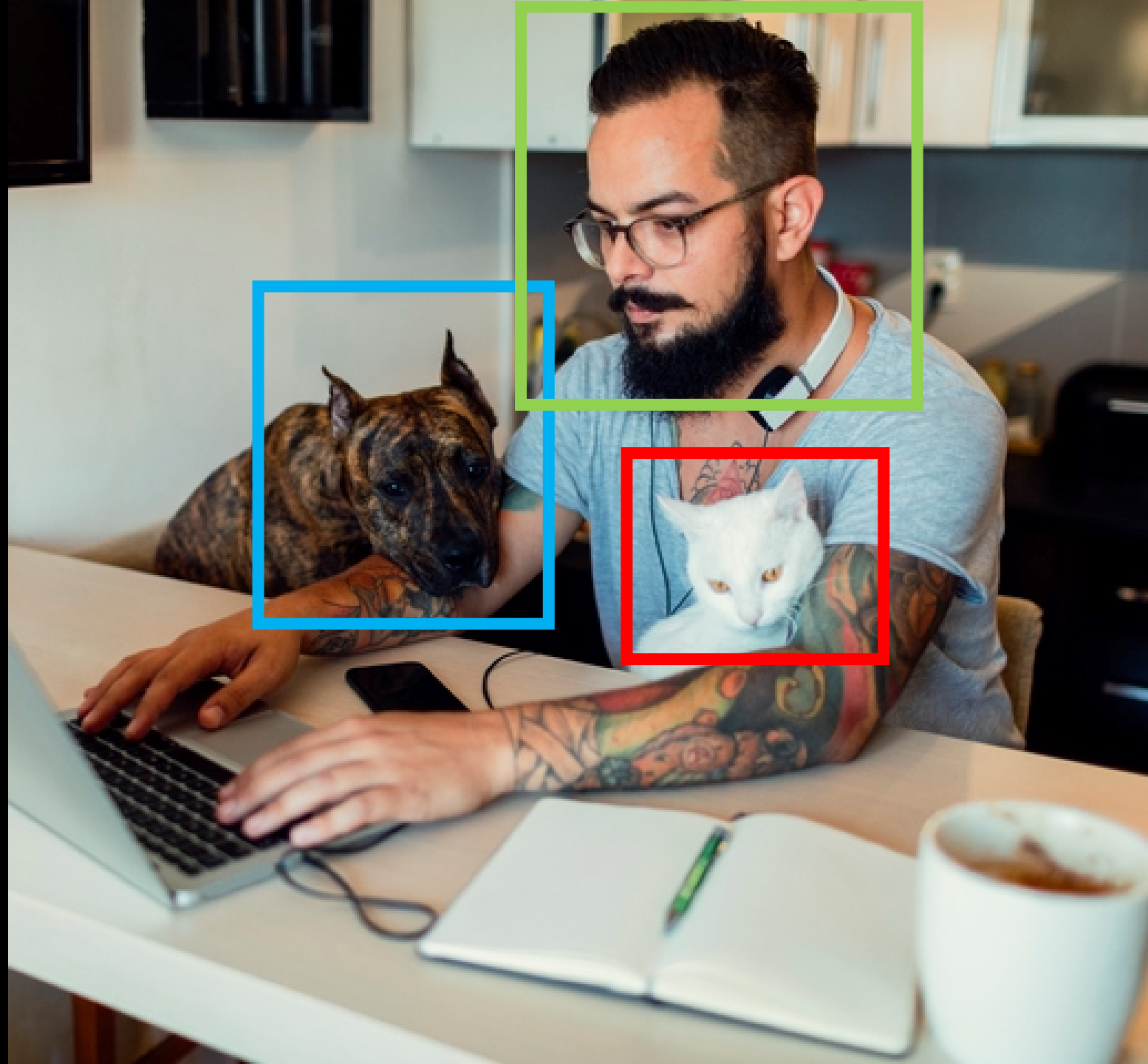
– Cathy O’Neil, 2014




# Machine biases replicate, amplify and systematize societal biases

Image:
“A HOLC 1936 security map of Philadelphia showing redlining of lower income neighborhoods,”
Wikimedia Commons

# Data (ingredients) are labeled



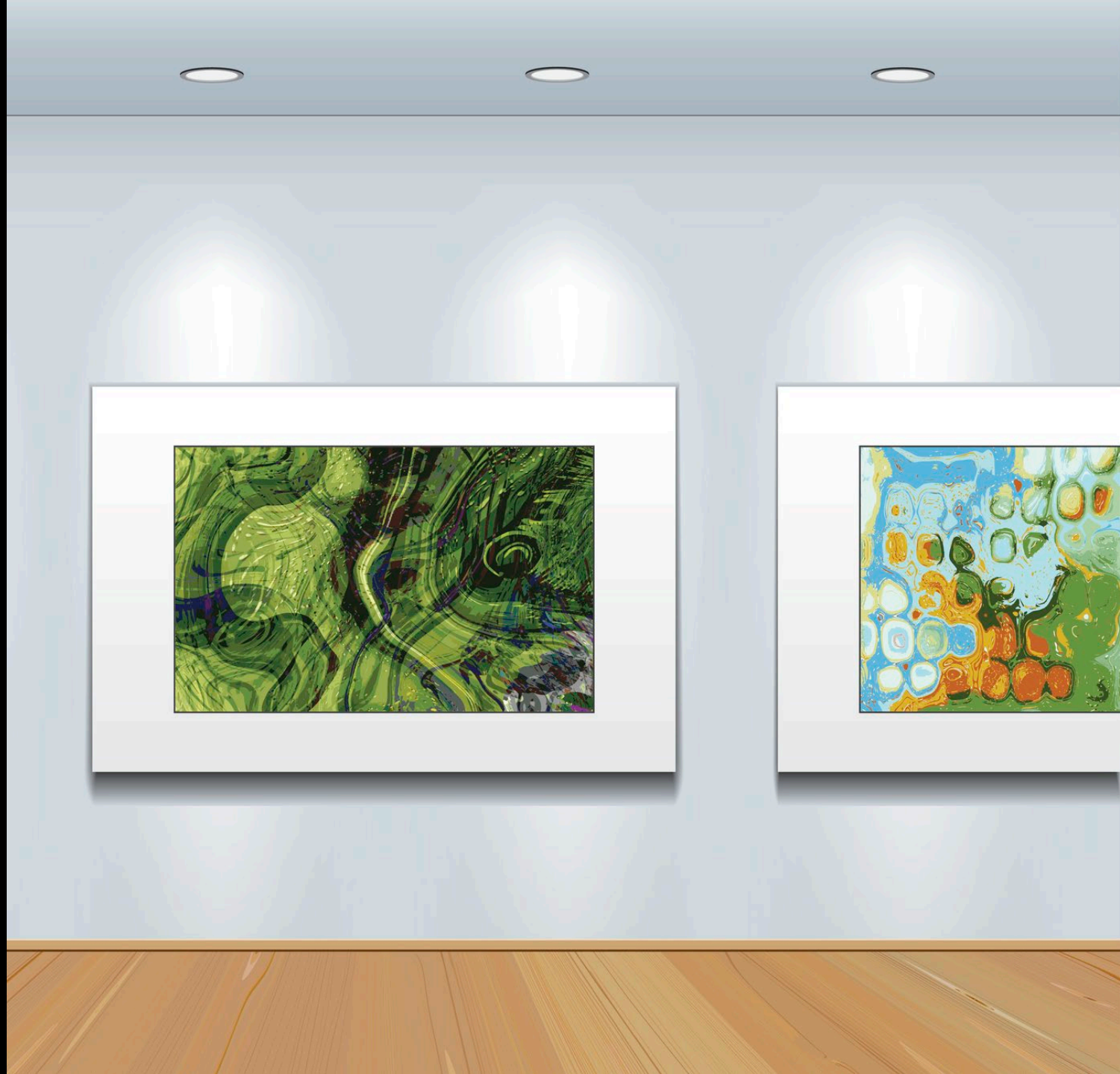
Dog, Man, Cat

# Data label bias

Are these pictures
or paintings?

# Data label bias

What about these?



# Data itself may contain a lot of surprises

# Sensitive attributes may be redundantly encoded in data

- Music tastes > age
- Shopping patterns > gender
- Zip codes > race, income
- Family status > gender
- Education > race
- Height, weight > gender

# What else is hiding in your data?

## FEMALES


## MALES


Figure 3. Words, phrases, and topics most highly distinguishing females and males.

Schwartz, H. Andrew, et al. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PLOS ONE* 8(9), 25 Sept. 2013. Web.

PLOS | ONE

# Questions to ask of your AI vendor

- Which data sets was the service trained on?
  - Public/Open-source? Built-for-purpose? Adapted?
  - Transformation of data?
  - Quality assurance process?
- Was any of the data synthetic?
- Who labeled the training data? (Which country?)
- Was the data set checked for bias?
  - Which biases?
  - Detection methods
- Was any remediation performed?
  - Techniques used
  - Results: before and after

# Algorithm and proxy biases, aka what can go wrong with the recipe

# Algorithm is a recipe for how to convert data into predictions

- Boil
- Scramble
- Poach
- Bake
- Omelet,
- Fry, etc..

# Inductive bias is assumptions about the future

**K-nearest neighbor classifier**

A new swan will be the same color as the most common color of the five nearest previous swan sightings.

**Naive Bayes**

Each swan color has a normal distribution about a certain latitude.

**Neural networks**

The swan color distribution can be represented using 20 properly scaled and rotated sigmoidal functions over location

Source: InductiveBias, 2013

# “Racial Bias Found in a Major Health Care Risk Algorithm”

(Scientific American)

# What are we really predicting?

# of prior arrests > committing a new crime? Or probability of being arrested again?

# Questions to ask of your AI vendor

- Which algorithms were used and why?
- Which proxies?
- How is performance (accuracy) measured?
  - False positive
  - False negatives
  - Intersectionally
  - Which definition of fairness (21-70 mathematical definitions out there)
- Trade-offs between accuracy and fairness?

# Design biases, aka the chef

# Many AI biases are a product o human cogniti biases


## If X, then Y

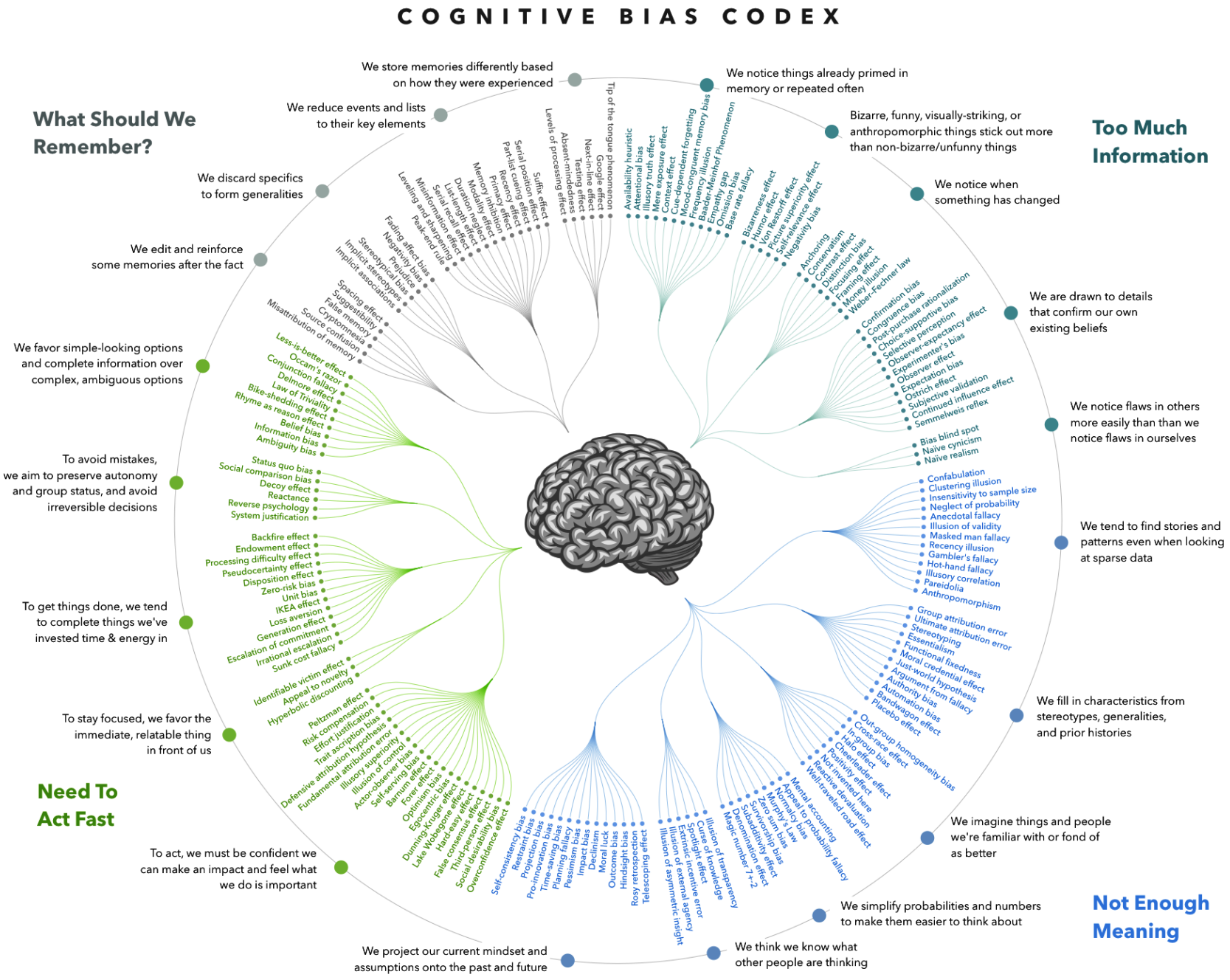
# Human cognitive biases are mental shortcuts

# Built-in mechanisms to understand the world and make decisions quickly

# Just how biased are we?

Almost 200 human cognitive biases identified

Image: Wikipedia, “Cognitive Bias Codex”



***"Applied machine learning is basically feature engineering."***

– Andrew Ng, Stanford University,
quoted in Google Cloud

- Data cleansing
- Partitioning: train, validate, test
- Tuning: outliers, missing values, etc.
- Transformation: numerical to categorical
- Feature extraction: text to word vectors
- Feature selection, removing redundancy
- New feature creation

***“[Machine learning] models are opinions embedded in mathematics.”***

– Cathy O’Neill,

“Weapons of Math Destruction”

# Which of these data sets is most representative of what a car is?

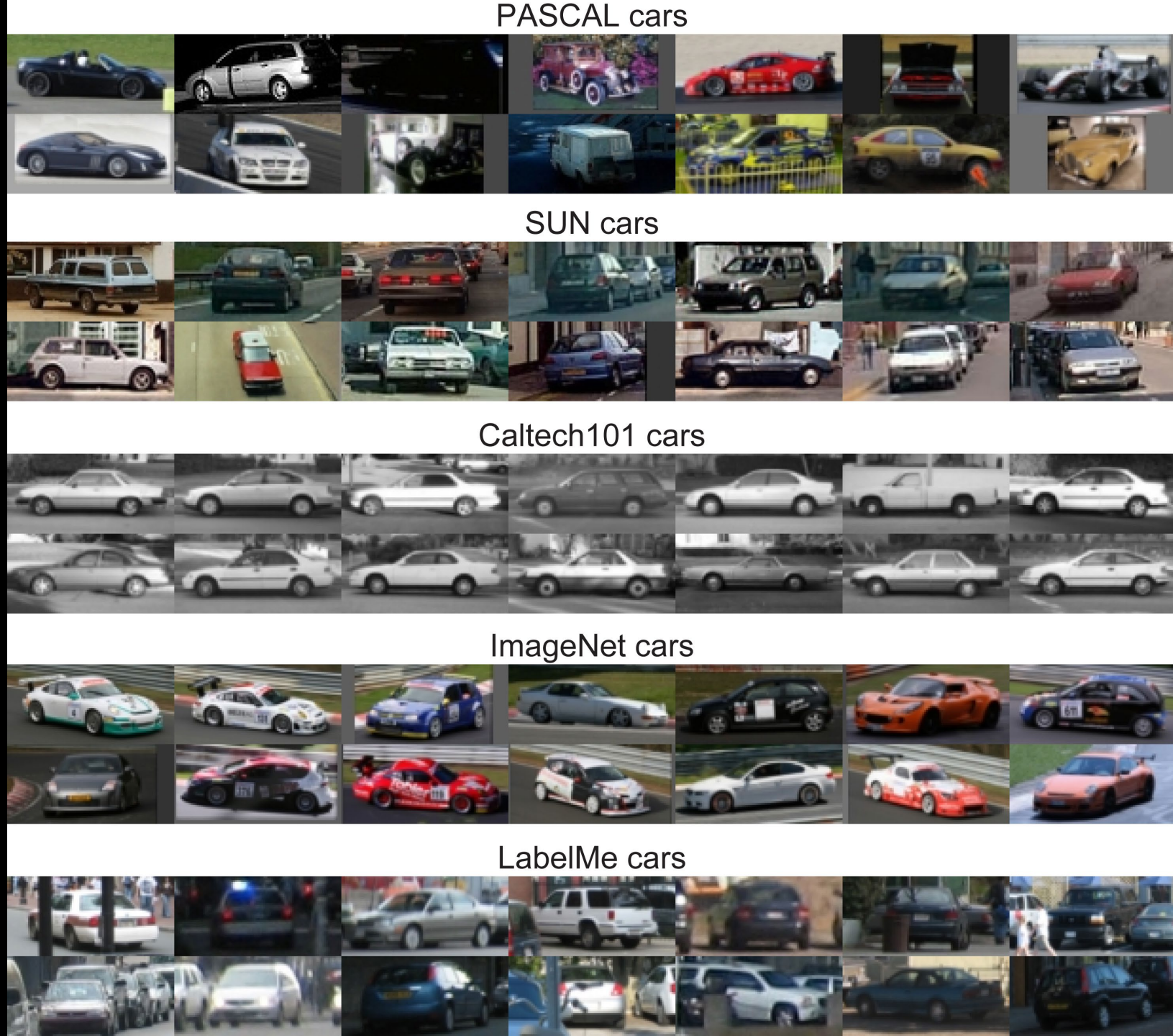
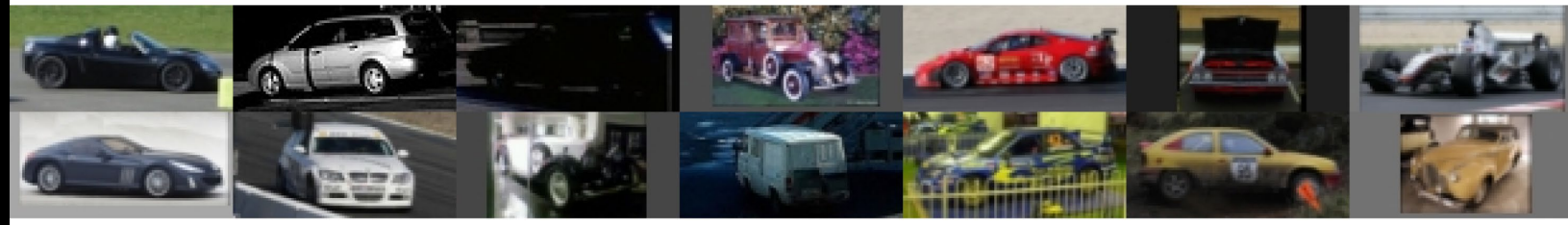



Image: “Unbiased Look at Dataset Bias,” Torralba & Efros

**Non-canonical views**

PASCAL cars



SUN cars

**Side views**

Caltech101 cars

**Race cars**

ImageNet cars

**Occluded by small objects**

LabelMe cars

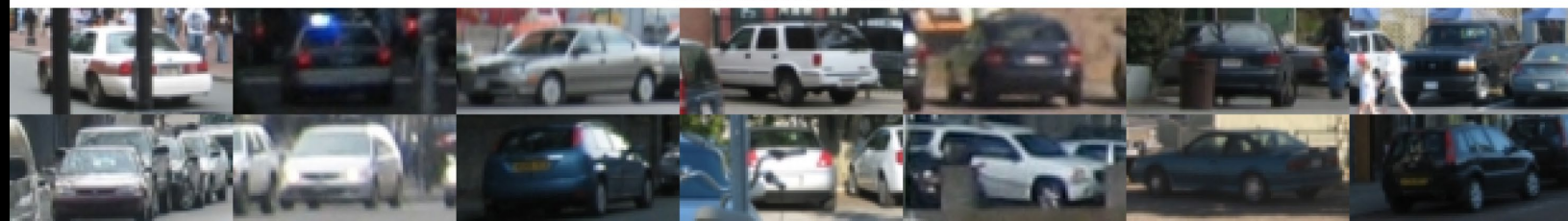




**Image: Torralba & Efros**

# Team diversity is the best mechanism to mitigate AI biases

- **18% –** female data scientists
- **18% –** female authors at leading AI conferences
- **80% –** male AI professors
- **15% –** female AI researchers at Facebook; **10%** at Google
- **2.5% –** black workforce at Google; **4%** at Facebook and Microsoft each

# Questions to ask of your AI vendor

- How diverse is the development team?
- Were domain experts involved?
- What are the intended use cases? (And out-of-scope)
- How and by whom was the service tested? (Metrics)
  - Third party?
- Were the potential sources of biases analyzed?
  - Do they arise from data? Feature engineering? Algorithms used? Assumptions? Etc.
  - How were they addressed?
- Are the service outputs explainable?

# But… ultimately, you, not the vendor, are responsible

- What is your team composition like?
- What data will you be using?
- Quality?
- How well does it align with the vendor’s training data?

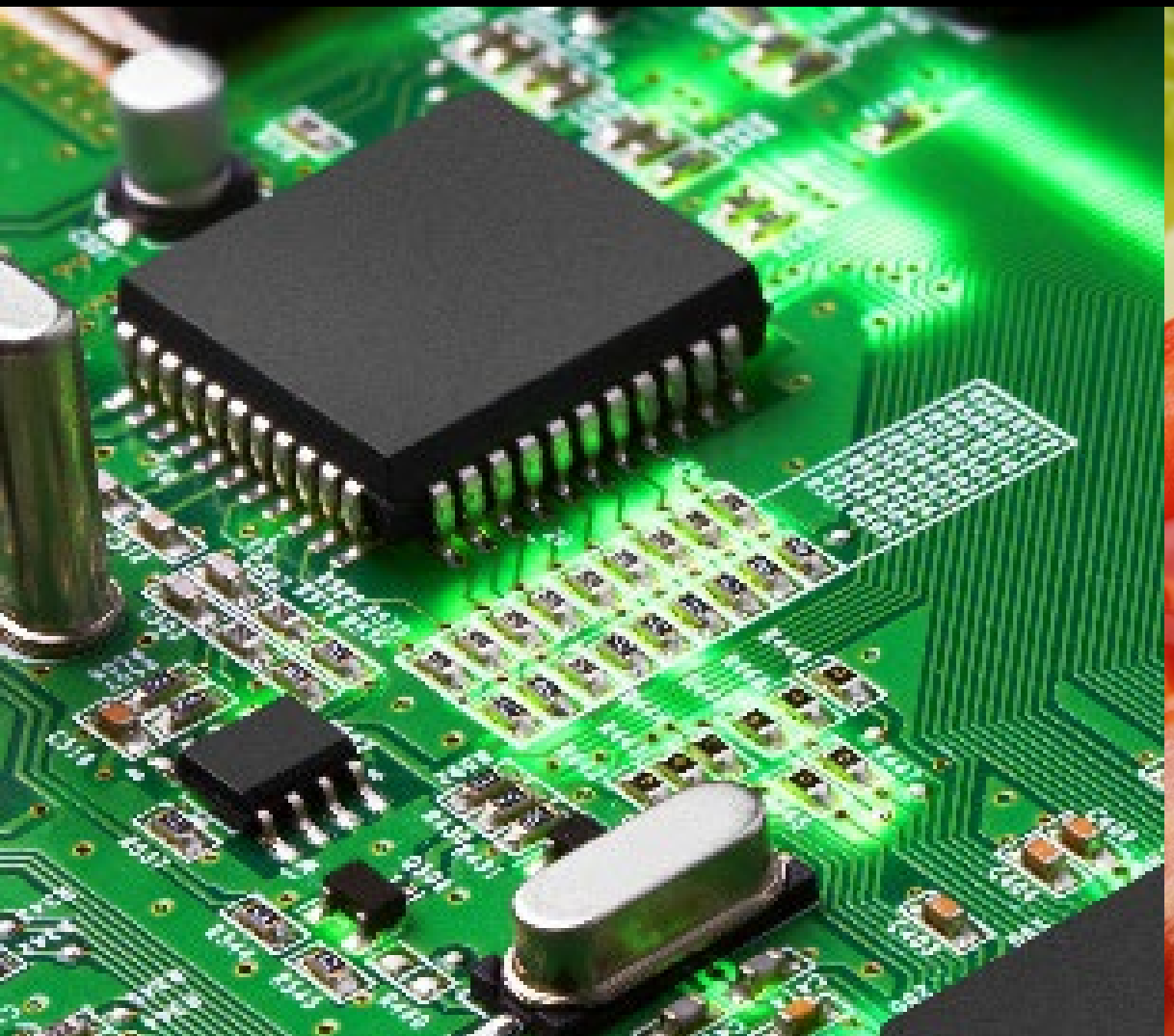
# Use tools and humans to identify biases

- AI Fairness 360 Open Source Toolkit (IBM)
- What-If, Facets, Fairness Indicators (Google)
- FairLearn (Microsoft)
- SageMaker Clarify (Amazon)
- Themis (UMASS)
- FairTest (Columbia)
- FairML (GitHub)
- Cortex Certifai (CognitiveScale)

# Leave no stone unturned and no assumption unexamined

# Frameworks to mitigate AI biases

**Datasheets for Datasets**



**Model Cards for Model Reporting**

# FactSheets: Increasing Trust in AI Services Through Supplier's Declaration of Conformity



Info-Tech Research Group



## FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1] R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1] J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

### Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

### 1 Introduction

Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

# AI Registers: A tool to create transparency and accountability around AI/ML applications in government

## Algorithmic systems of Amsterdam

Learn about the use cases where we currently utilise algorithmic systems

Economic Services Departments

### Automated parking control

In Amsterdam, the number of cars allowed to park in the city is limited, keeping the city liveable and accessible. The municipality checks whether a parked car has the right to be parked, for example, because parking fees have been paid via a parking meter or app, or because the owner has...

> Read more

Economic Services

### Holiday rental hou fraud...

Amsterdam has limited living s both for citizens and visitors. If citizen wants to rent out their h houseboat to tourists, they nee meet certain requirements. Fo example, they can do so for a maximum of 30 nights per yea maximum of 4 people at a tim must...

> Read more

Source: Algorithmic Systems of Amsterdam

Info-Tech Research Group

City management

Overview

## Reporting issues in public space

Tags: Reports | Complaints | Natural language processing

When someone encounters rubbish or a maintenance issue on the street or in a park, they can report this to the municipality via an online reporting system. A dangerous traffic situation or disturbance from people or cafés can also be reported.

This system used to be a collection of drop-down menus, from which the user would pick the category that best suited their report. The department responsible for a certain category would then take care of the report. However, as the municipality is a complex organisation, there are countless categories. Many times the wrong category would be chosen, resulting in delays. Now, an algorithm recognizes certain keywords, for example, 'waste' and 'sidewalk'. From these keywords, it determines which category it belongs to, and ultimately, which department within the municipality should examine the case.

As a result, there are fewer administrative steps for the person reporting on the issue. Also, the report can be processed much faster, because it arrives at the right department more quickly.

Link to service

### Contact Information

Department
Research, Information & Statistics (OIS)

Contact person for inquiries
Adviseur R&D (Adviseur Onderzoek en ontwikkeling)

External vendors
Developed in-house

Contact email
CIO-office@amsterdam.nl

Contact phone
14020

## More detailed information on the system

Here you can get acquainted with the information used by the system, the operating logic, and its governance in the areas that interest you.

**Datasets** — Show More

**Data processing** — Show Less

The operational logic of the automatic data processing and reasoning performed by the system and the models used.

**Model architecture**

The text of the report is broken down into single words. The model has been trained to recognize the weight of each word by using 'TF-IDF' or term frequency-inverse document frequency'. This representation will create weights for words that show how unique they are for the specific citizen report compared to the overall collection. A word such as 'the' will get a low weight, and a word such as 'garbage' will get a higher weight. This makes it perfect for classes that have very specific words describing them. It also helps with bigrams or unigrams (Like: "thank you", "please") occurring in all documents not to affect the classification too much.

A logistic regression (a machine-learning technique) of that combination of words is then used to determine which category is most likely to fit, and therefore which department within the municipality needs to act on the report.

Link to source code

| Content | Attachment |
|---|---|
| Model architecture | Reporting issues in public space architecture image |

**Performance**

This algorithm can detect very accurately which category a combination of words belongs to: the algorithm has a score of 0.88 (macro-weighted F1 score). Other methods have also been implemented (W2V, CNN + LSTM, BERT), but have been found to perform less. More information: https://medium.com/maarten-sukel/how-to-use-machine-learning-for-the-classification-of-citizen-service-requests-b71137fa55f56

**Non-discrimination** — Show More

**Human oversight** — Show More

**Risks** — Show More

Was this information useful?

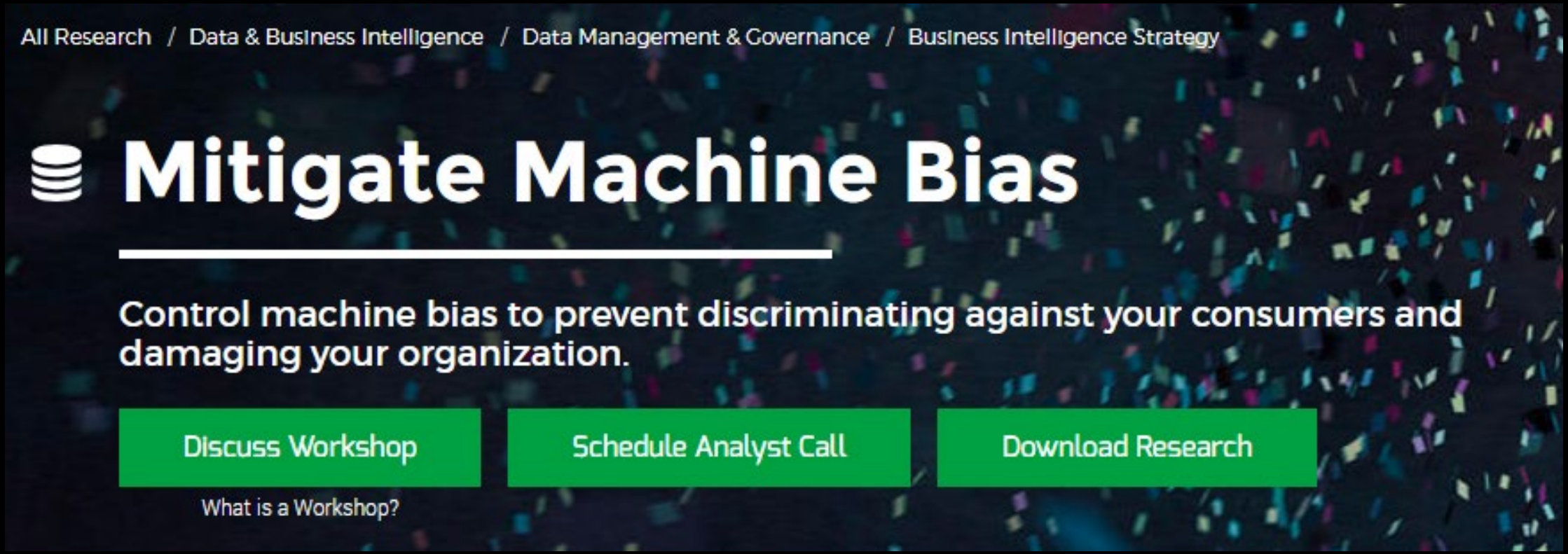
Yes, it was | Partially | Not really

Source: https://algoritmeregister.amsterdam.nl/en/reporting-issues-in-public-space/

***“Without trust, there is no use for AI.”***

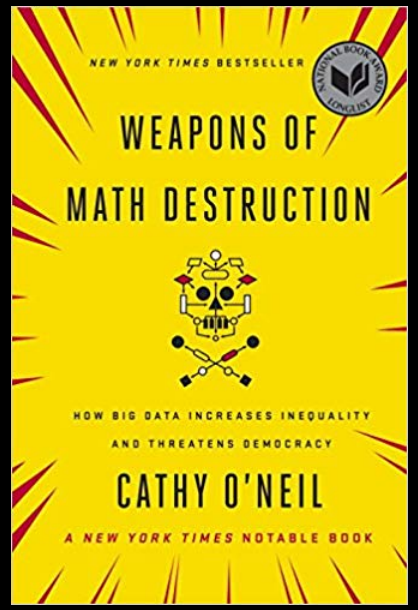
– Mikko Rusama, City of Helsinki Chief Digital Officer

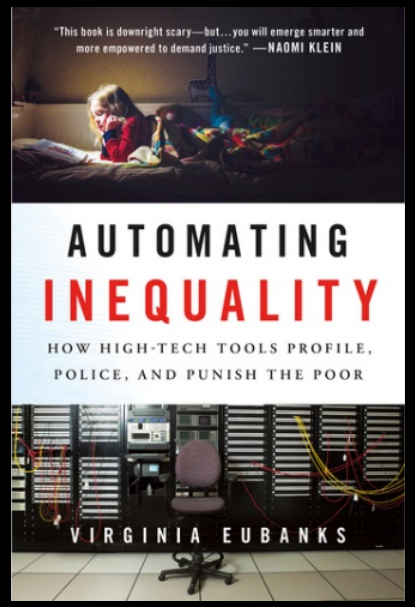
# For more information, consult blueprint

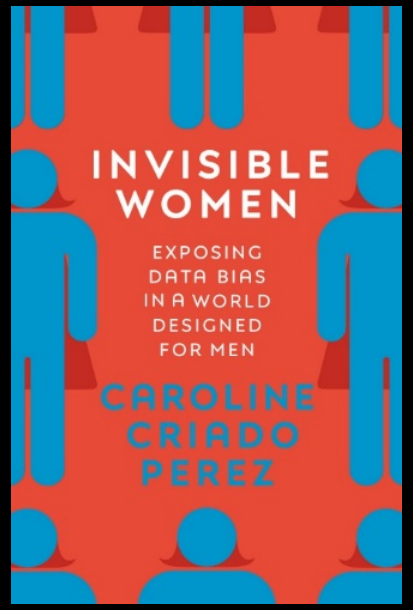


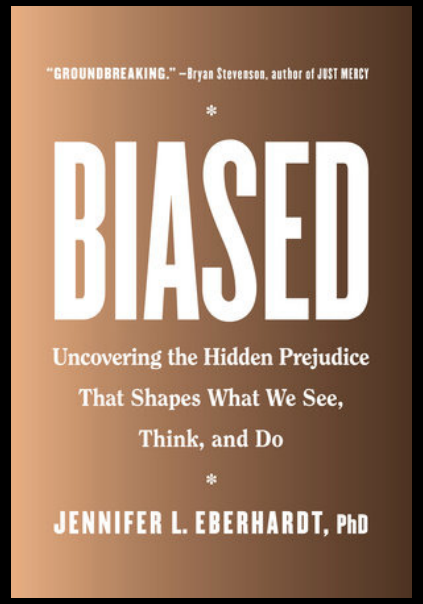
# Recommended resources

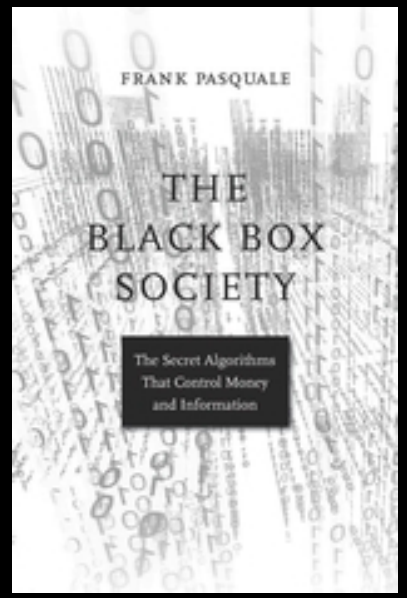
## Books




## Videos

"Scottish Voice Recognition Elevator – ELEVEN!" (YouTube)

"The Trouble with Bias"
NIPS 2017 Keynote by Kate Crawford, cofounder of AINow Institute at NYU, principal researcher at Microsoft, and distinguished research professor at NYU

"Translation Tutorial: 21 Fairness Definitions And Their Politics"
by Arvind Narayanan, Professor of Computer Science at Princeton University

# Thank you!

# Questions?