Automating the Status Quo

How Machine Learning Algorithms Get Biased

Celeste Tretto | Data Scientist, Splunk UBA
Sarah Moir | Senior Technical Writer, Security

September 26, 2017 | Washington, DC

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance 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 could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. 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 functionality described or to include any such feature or functionality in a future release.

Welcome

Who Are We?

Celeste Tretto

- Data Scientist @ Splunk
  - Writes algorithms on the daily
- Twitter: @_HappyC
- Likes climbing, music, running, yoga, cooking

Sarah Moir

- Senior Technical Writer @ Splunk
  - Writes about technology on the daily
- Twitter: @smorewithface
- Likes climbing, music, writing, hiking, livetweeting events

What is this about?

Automating the Status Quo

What are we talking about?

- What is machine learning?
- How does bias get introduced in machine learning models?
- Why does it matter?
- How can we not do that?

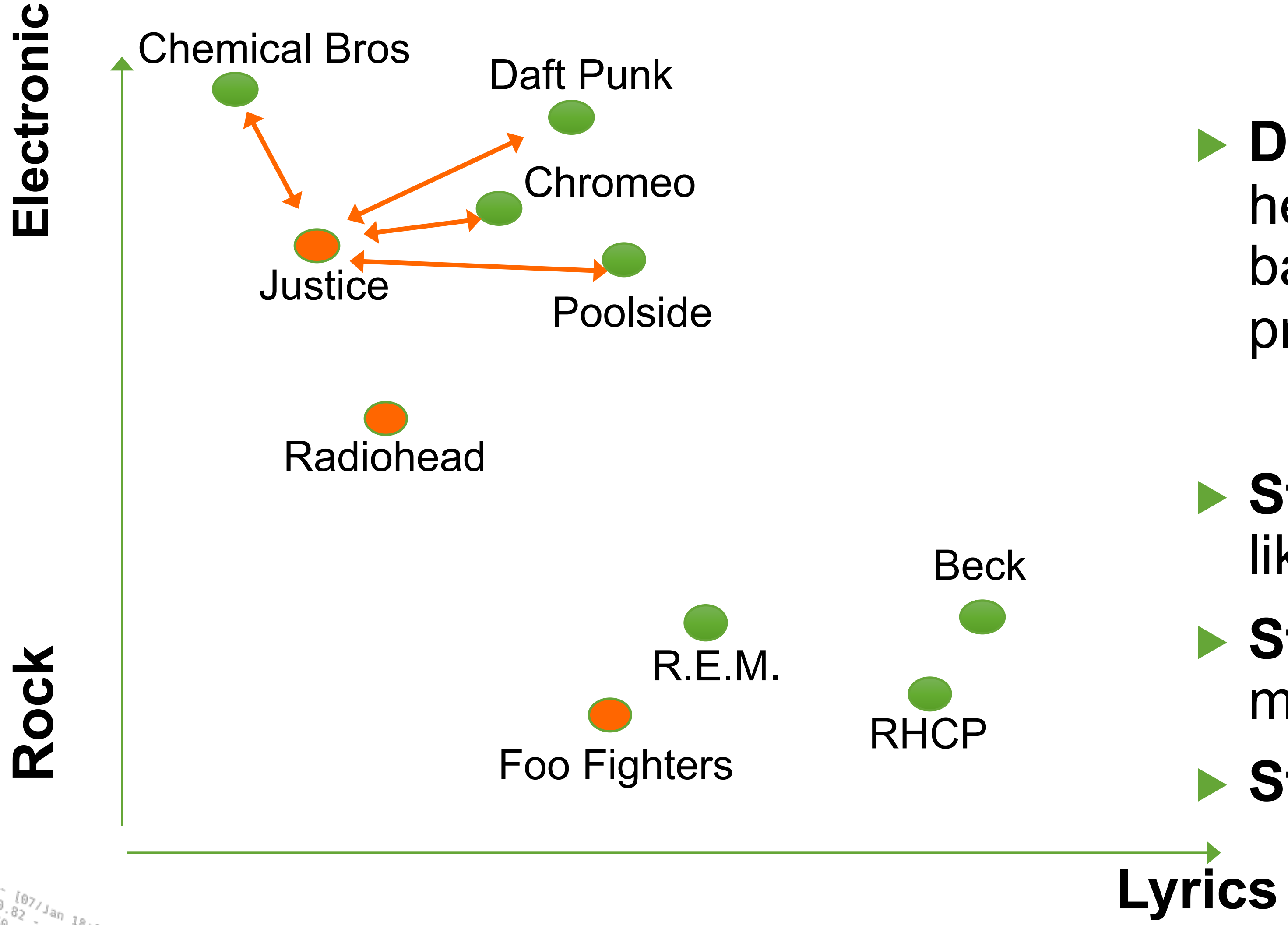
Wait, what's machine learning actually?

Machine learning is the **abstraction** of a **decision process** into an **algorithm**.

For example…

Music Recommendation Model

Sarah recommends a band that Celeste might like

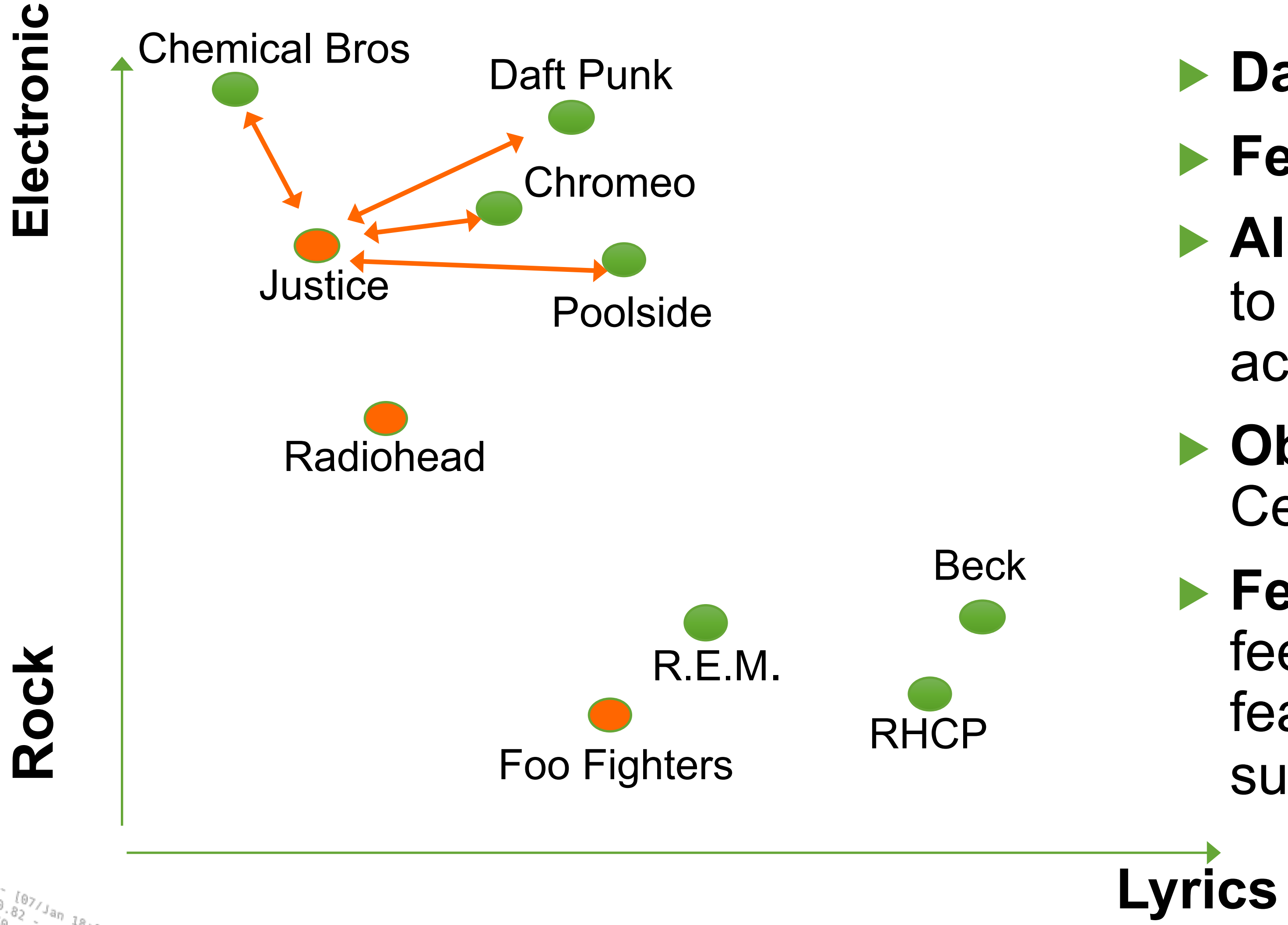


- **Decision process**: Sarah uses her knowledge about music to find bands that are close to Celeste's preferences

- **Step 1**: build a map of bands you like
- **Step 2**: find a friend, build their map
- **Step 3**: match preferences

Music Recommendation Model

Breaking down a model into components



- **Data:** two users, a few dozen bands
- **Features:** genre, lyrics
- **Algorithm:** find band that is closest to Celeste's current preferences, according to features
- **Objective function:** maximize Celeste's musical enjoyment
- **Feedback loop**: use Celeste's feedback to model to help with feature selection for future suggestions

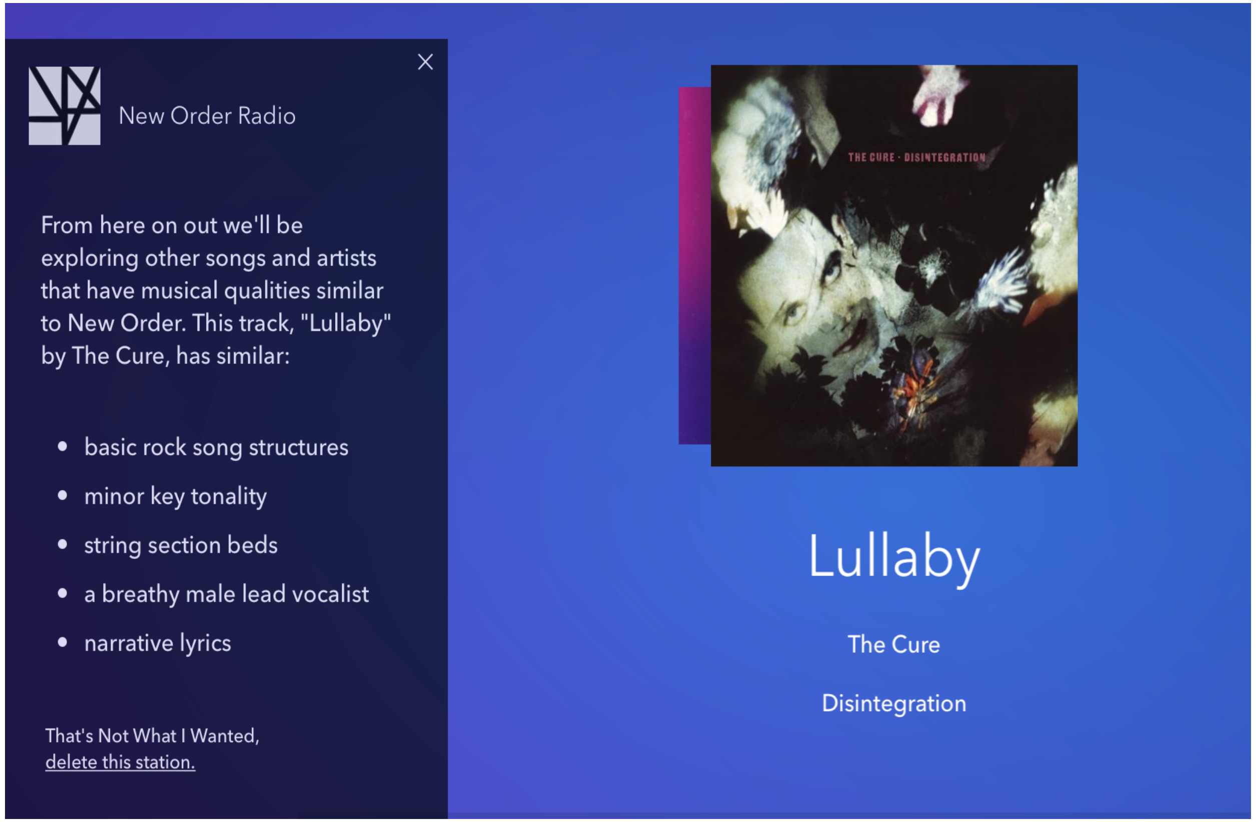
Let's do that again, but bigger!

Scaling models

Big Music Radio Company Recommendations

Same decision process, but at scale

- More users, more bands, more data
- More features
- More complex algorithms
- More complex objective function
- Feedback loop at scale



Machine Learning

A gray box, not a black box

- Now we have a common understanding of what machine learning is designed for….
- Based on that, how does bias get introduced into machine learning?

So what about bias?

Let's have a more detailed look

What is bias?

Bias introduces disparity

- Prejudice or discrimination against something, someone, or some group.
- With machine learning, algorithms can introduce bias.
- Discriminatory bias is created when data-driven decisions have unbalanced outcomes

Why?

- Machine learning is a tool
- Machine learning is not going to solve discrimination, unless we specifically build a model that does that
- As data practitioners, we are responsible for educating ourselves on how machine learning decision are affecting our society

Big data and bias

Data driven decisions are subject to human biases

- Finding sources of bias and flawed decisions is not a scientific process that we can automate
- Bias is usually involuntary
- We are dealing with complex questions
- Decisions are subject to constraints (budget, timeline, regulations, …)
- Decisions might be based on poor quality data

We are all in this together

It's Easy to Introduce Bias

Feature engineering

Subjective features
Incomplete features
Wrong features

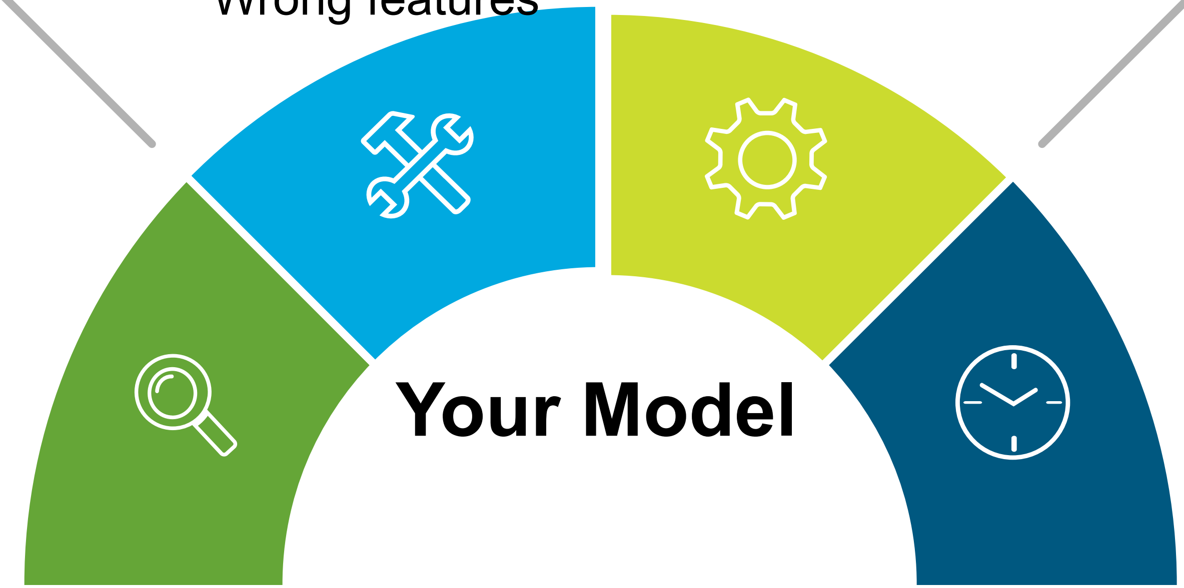
Algorithm / Objective function

Opaque "black-boxes"

Standardized preferences

Data collection

Incomplete data
Biased data

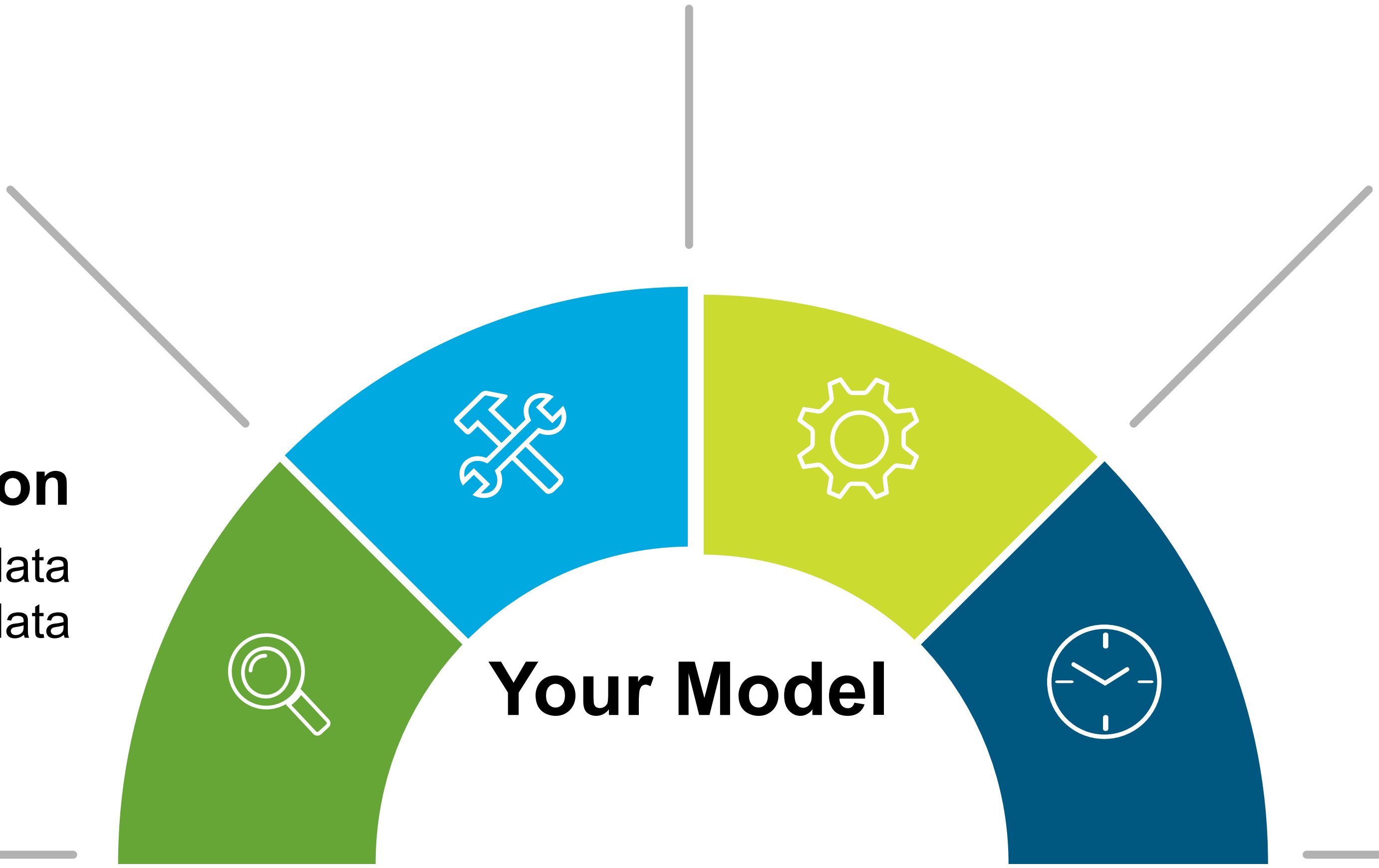


Decision-making / Feedback loops

Model outcomes have global impact

Data selection bias

We’re here



Data collection

Incomplete data
Biased data

Representative Data

Data should include complete information about the problem to solve

Facial-Recognition Software Might Have a Racial Bias Problem

Depending on how algorithms are trained, they could be significantly more accurate when identifying white faces than African American ones.

CLARE GARVIE AND JONATHAN FRANKLE | APR 7, 2016 | TECHNOLOGY

Learning from biased data

Word embedding

- Word embedding transforms text into vectors of words
- Characterize the “meaning” of a word using the words that are close by:
  - Paris : France = Tokyo : x
  - x = Japan

From Google News articles:

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

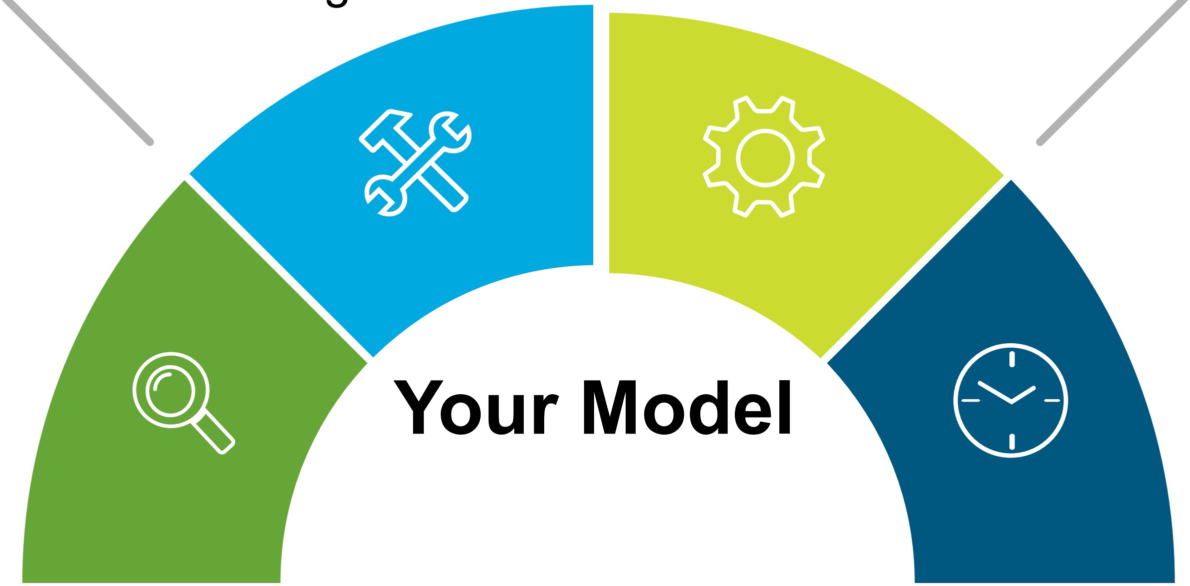
Source: Bolukbasi, Chang, Zou, Saligrama, Kalai (2016)

Feature selection bias

We’re here

Feature engineering

Subjective features
Incomplete features
Wrong features

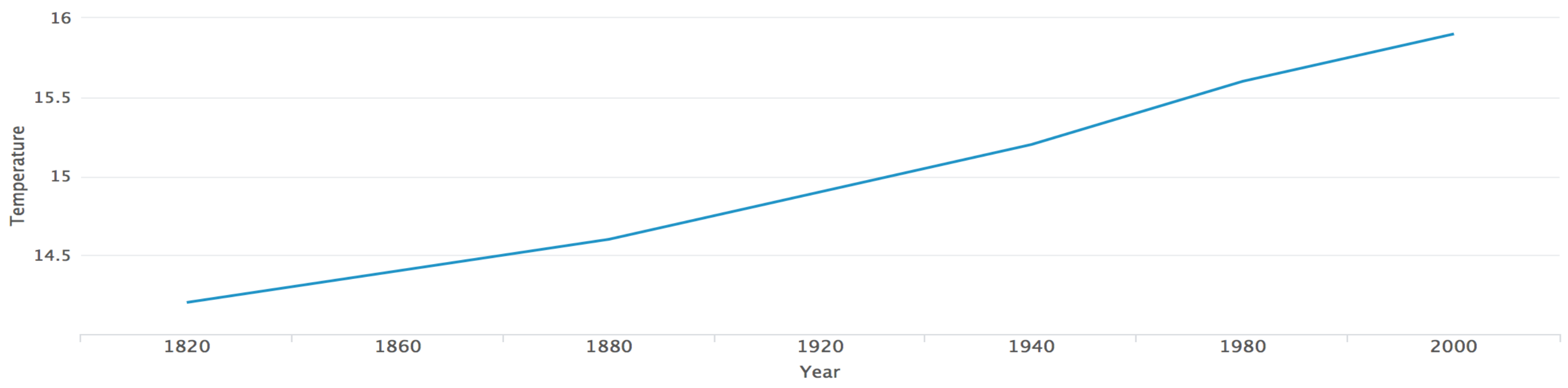


Pirates will fix global warming!

Can they?

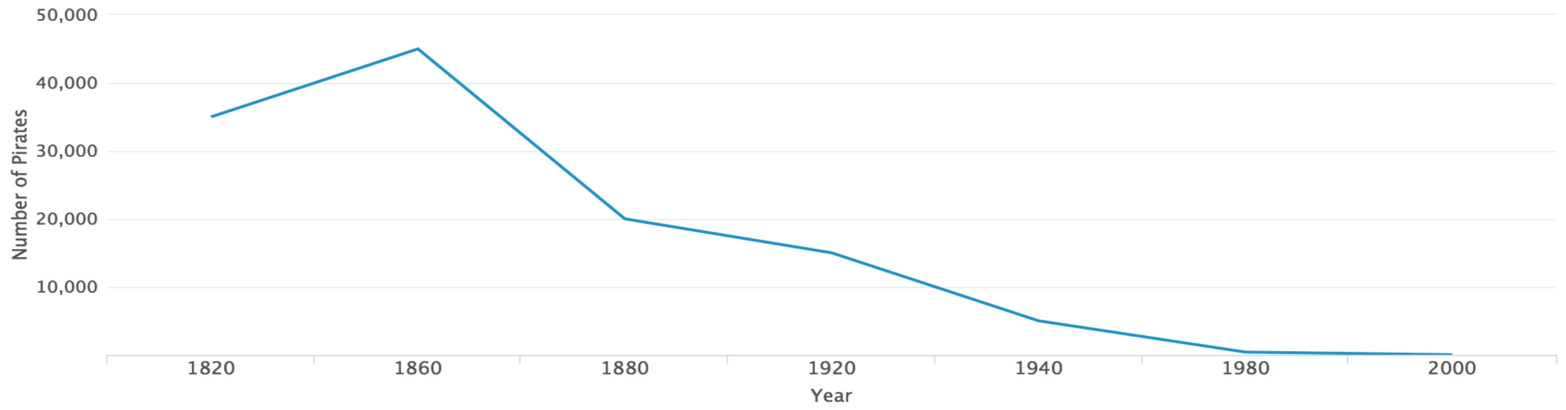
Average temperatures

1820 - 2000



Number of Pirates

1820 - 2000



Risk assessments for recidivism

Are features representative of reality?

- Courts use "risk scores" at various stages of the criminal justice system
- What is an indicator of recidivism?
  - Poverty / homelessness?
  - Petty crimes?
  - Causal or correlated features? Signal or noise?
- Success at forecasting crime:
  - 60% of "High Risk" individuals were arrested within two years
  - 20% of "Violent Crimes High Risk" individuals were arrested for violent crimes within two years

Noisy features! Number of crimes increases with higher levels of poverty. But, being poor doesn't automatically make somebody a criminal.

Feature validation

Are features representative of reality?

- How often is my model giving me a true answer (precision)?
  - "High Risk, Reoffend" vs "High Risk, Didn't Reoffend"
- Out of all true instances, how often do I get a true answer (recall)?
  - "High Risk, Reoffend" vs "Low Risk, Reoffend"

	White	African American
High Risk, Didn't Reoffend	24%	45%
Low Risk, Reoffend	48%	28%

Source:https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Feature validation

Are features representative of reality?

- How can we measure discrimination?

- Risk score should have the same effectiveness regardless of group membership:
  - "Well calibrated" (model probability reflects reality)
  - Balance for the positive class (precision)
  - Balance for the negative class (recall)

If the underlying data is unbalanced, the three conditions of fairness cannot be met simultaneously.

Source: Kleinberg, Mullainathan, Raghavan (2016)

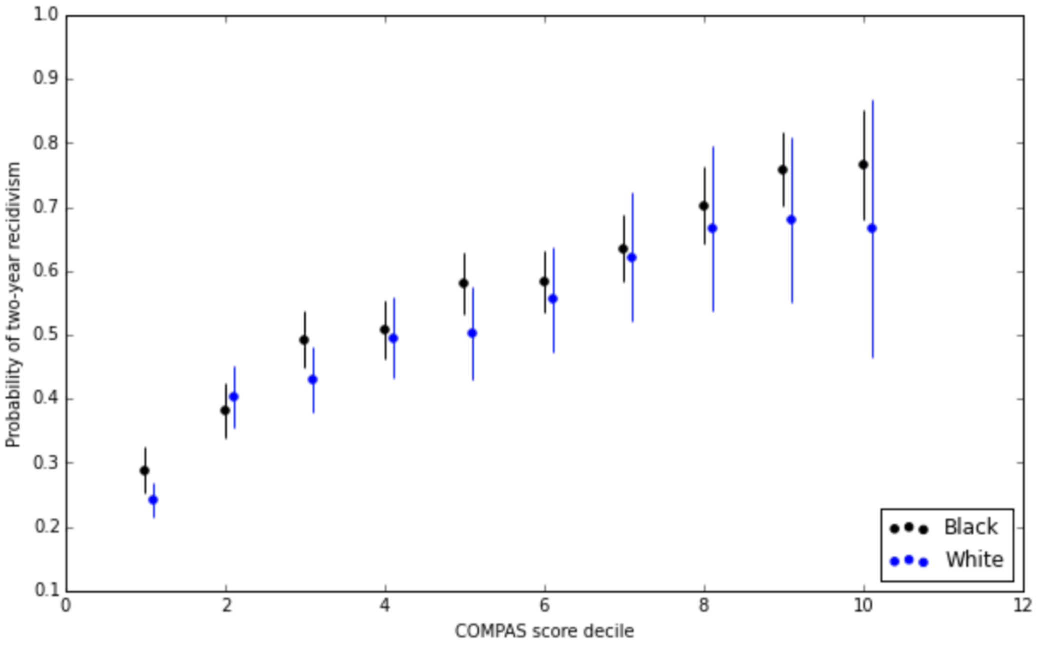
"Inherent Trade-Offs in the Fair Determination of Risk Scores"

Feature validation

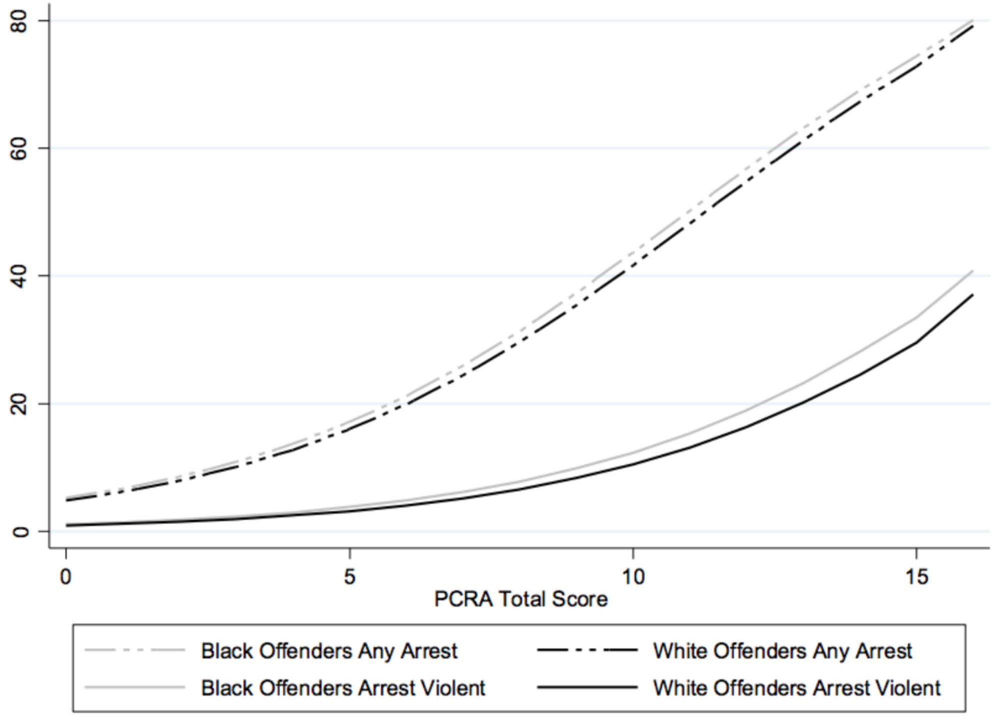
Are features representative of reality?

- When rates of arrest are unbalanced, the model is not well calibrated

Model: probability of recidivism



Real world: number of arrests

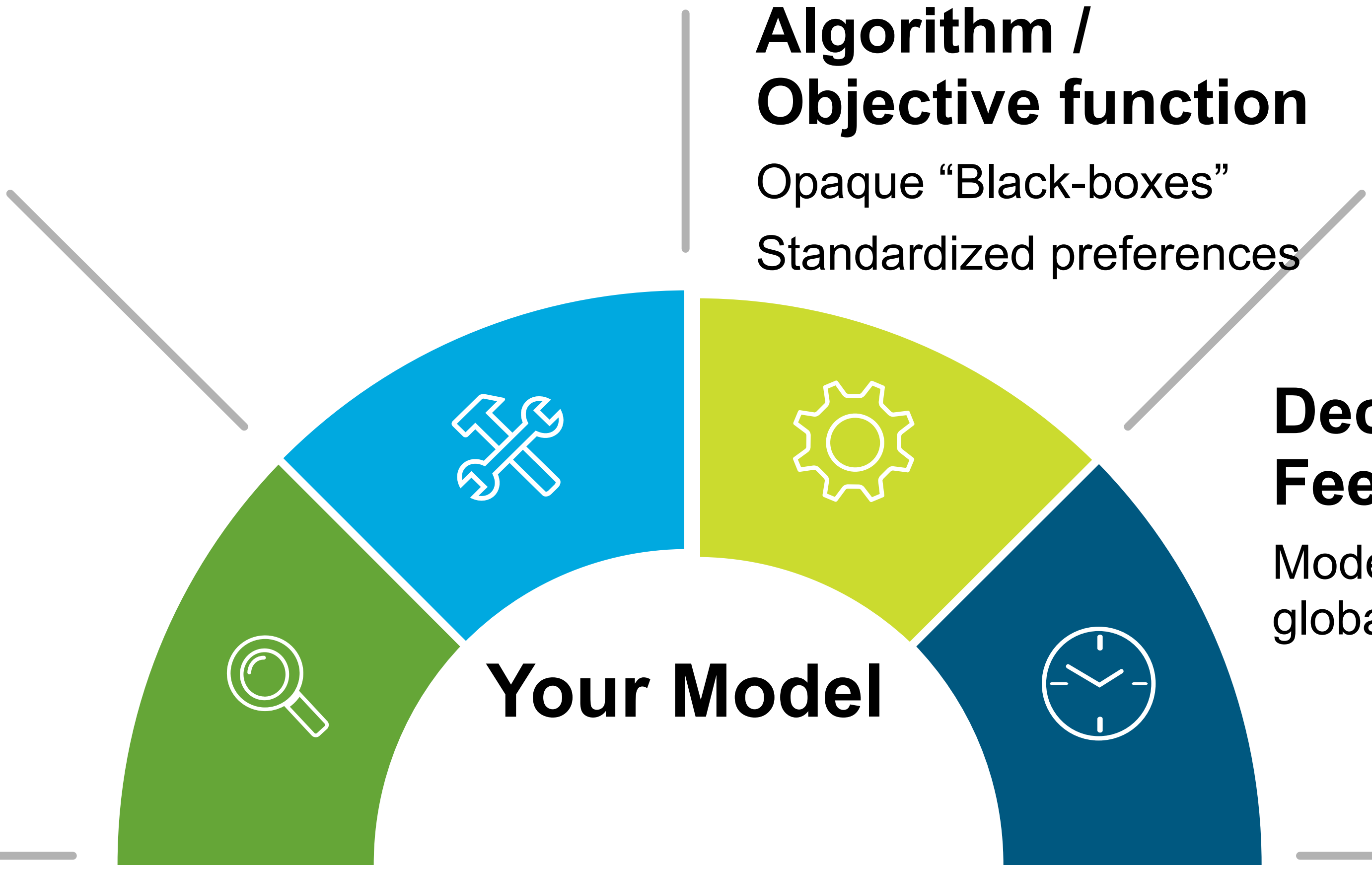


Source: Kleinberg, Mullainathan, Raghavan (2016)
“Inherent Trade-Offs in the Fair Determination of Risk Scores”
Plots: https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84

Model bias and feedback loops

Let's rank some universities

We’re here



Algorithm / Objective function

Opaque “Black-boxes”

Standardized preferences

Decision-making / Feedback loops

Model outcomes have global impact

Algorithm/Objective Function

- The objective function that you choose for the algorithm matters
- The algorithm takes a decision and translates it into math
- Biased objective function leads to a biased model

University rankings

It all started with the best of intentions

Ranking models are useful to help sort through and prioritizing a large amount of information.

In 1983, US News and World Report published their first University Ranking feature, to help students make more informed academic choices

Source: “Weapons of Math Destruction” by Cathy O’Neil

University rankings

A model is a simplified version of reality

- Raw data
  - N of professors / instructors
  - Research publications
  - Infrastructures
  - Classes
- Real metrics
  - Satisfaction
  - Personal growth
  - Career success
  - Happiness
- Features
  - Teacher / student ratio
  - SAT scores
  - Graduation rates
  - Employment rate
  - Reputation scores

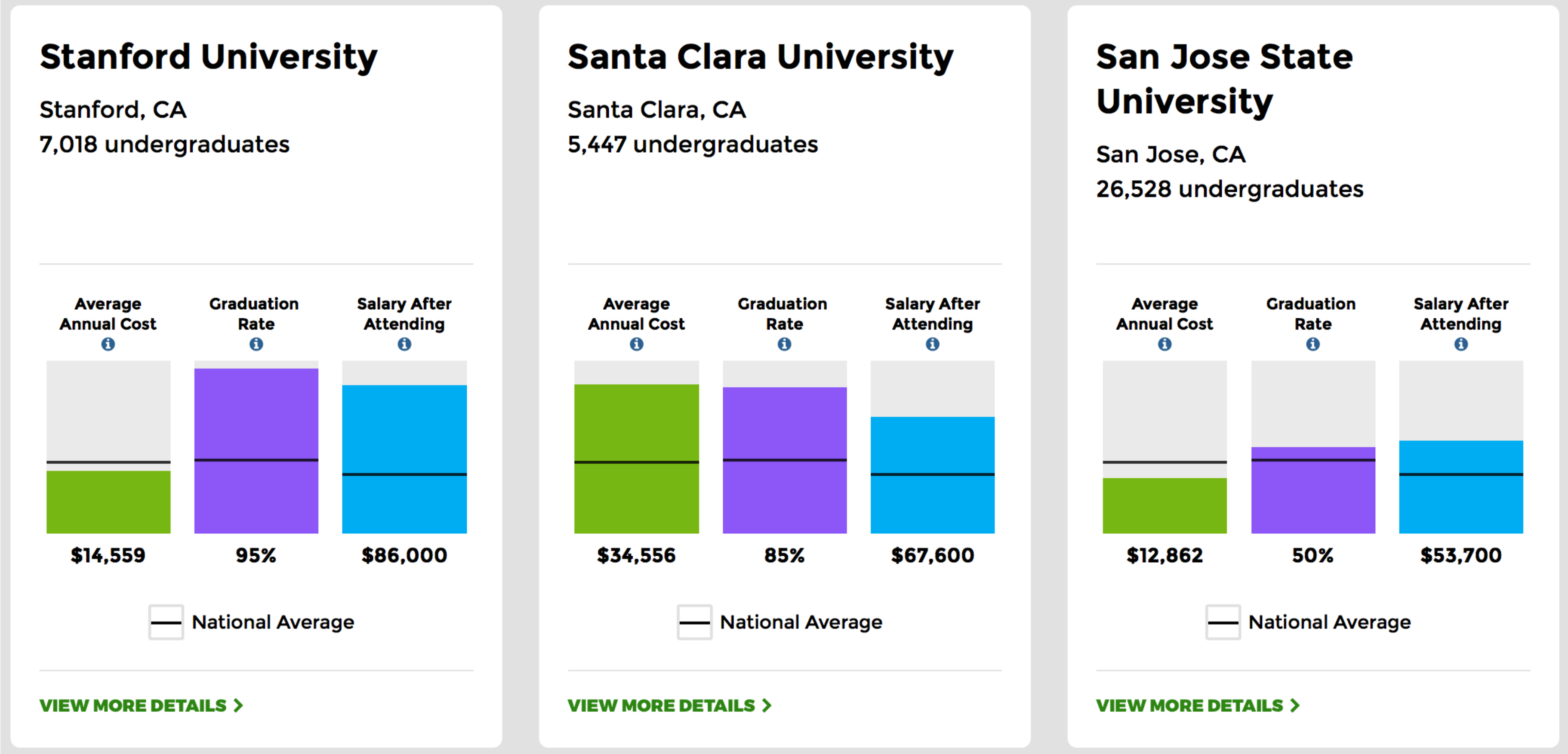
What went wrong?

Colleges were pushed to invest in research, infrastructure, student well-being. At the same time, disparities increased

- Universities focused resources on a few programs
- Tuition and fees were left out of the equation to start
- Features are easy to cheat and subjective
- Opaque model
- Objective function too broadly defined
- Self-reinforcing algorithm

College Scorecard

Available to the public – “build your own model”



https://collegescorecard.ed.gov

Bias-driven decisions

Data-driven decisions have global implications

- Data reflects the past, and the past is biased
- Model can “learn” from past data, persisting biases
- Model can create negative feedback loops, increasing biases

So what can we do?

Key Takeaways

Recognizing bias in data requires everybody's best effort

1. Ask if the data is **representative**.
2. Ask if the data is **biased**.
3. Ask if the features are **accurate proxies**.
4. Ask if the goal of the model is **unbiased**.
5. Ask about the **implications** of the model results.

Learn more

- Weapons of Math Destruction by Cathy O’Neil
  - Engaging book full of examples about machine learning and bias
- How to Lie with Statistics by Darrell Huff
  - Book with insights into statistical thinking
- Freakonomics Radio Podcast
  - Engaging stories about economic and social science research, real world biases, etc..
- Data & Society research institute
  - Blog, reports, and talks about big data and society
- ProPublica Machine Bias series
  - Investigative journalists focusing on machine bias

Thank You

Don't forget to rate this session in the .conf2017 mobile app

splunk> .conf2017