# What is data?

20 Jan 2022

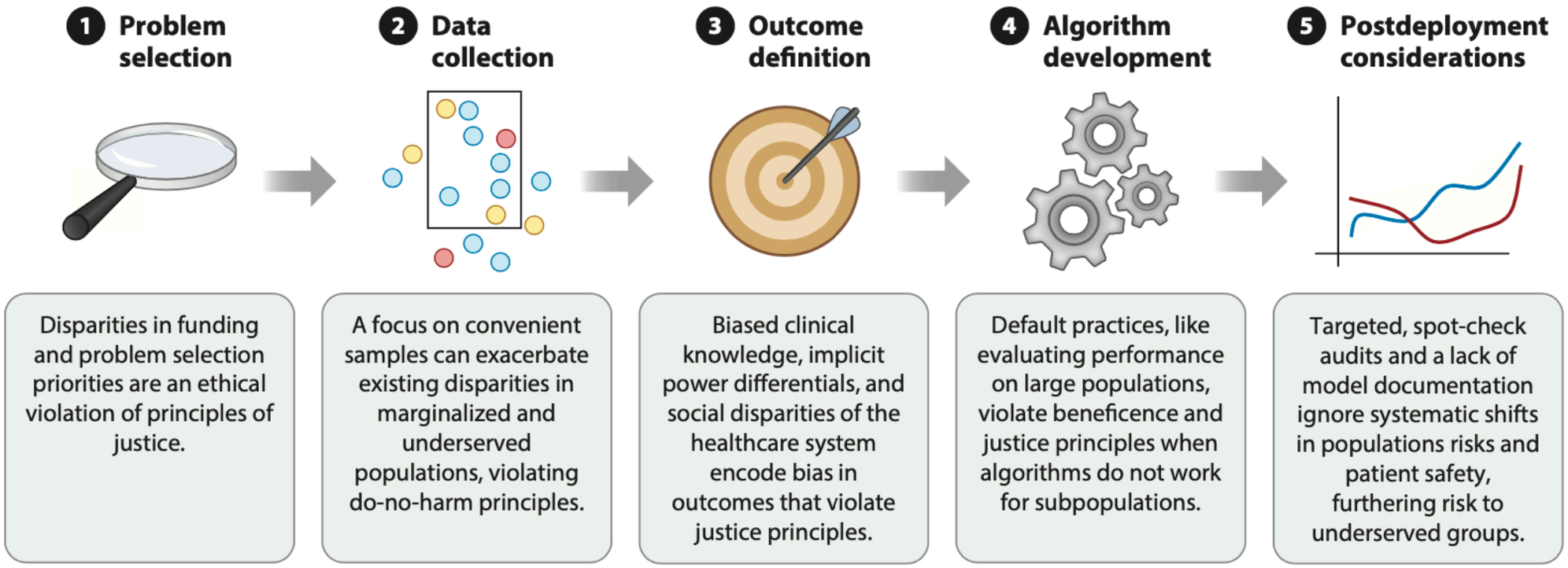



**Figure 1**

We motivate the five steps in the ethical pipeline for healthcare model development. Each stage contains considerations for machine learning where ignoring technical challenges violates the bioethical principle of justice, either by exacerbating existing social injustices or by creating the potential for new injustices between groups. Although this review's ethical focus is on social justice, the challenges that we highlight may also violate ethical principles such as justice and beneficence. We highlight a few in this illustration.

Chen et al. "Ethical Machine Learning in Healthcare"

# *What is data?*

# *What is data?*

Broadly speaking, data is/are **information**.

To be precise, we define data to be information that is **recorded/observed/created?** about an object or set of objects.

For a job applicant, data could mean prior job experience, high school GPA, college diploma.

In practice, data is always **limited**.

# *What is data?*

Data do not exist in a vacuum.

Who or what creates data?

Who "gets to" create data?

What is good data? Bad data? Big data?

# *Why data?*

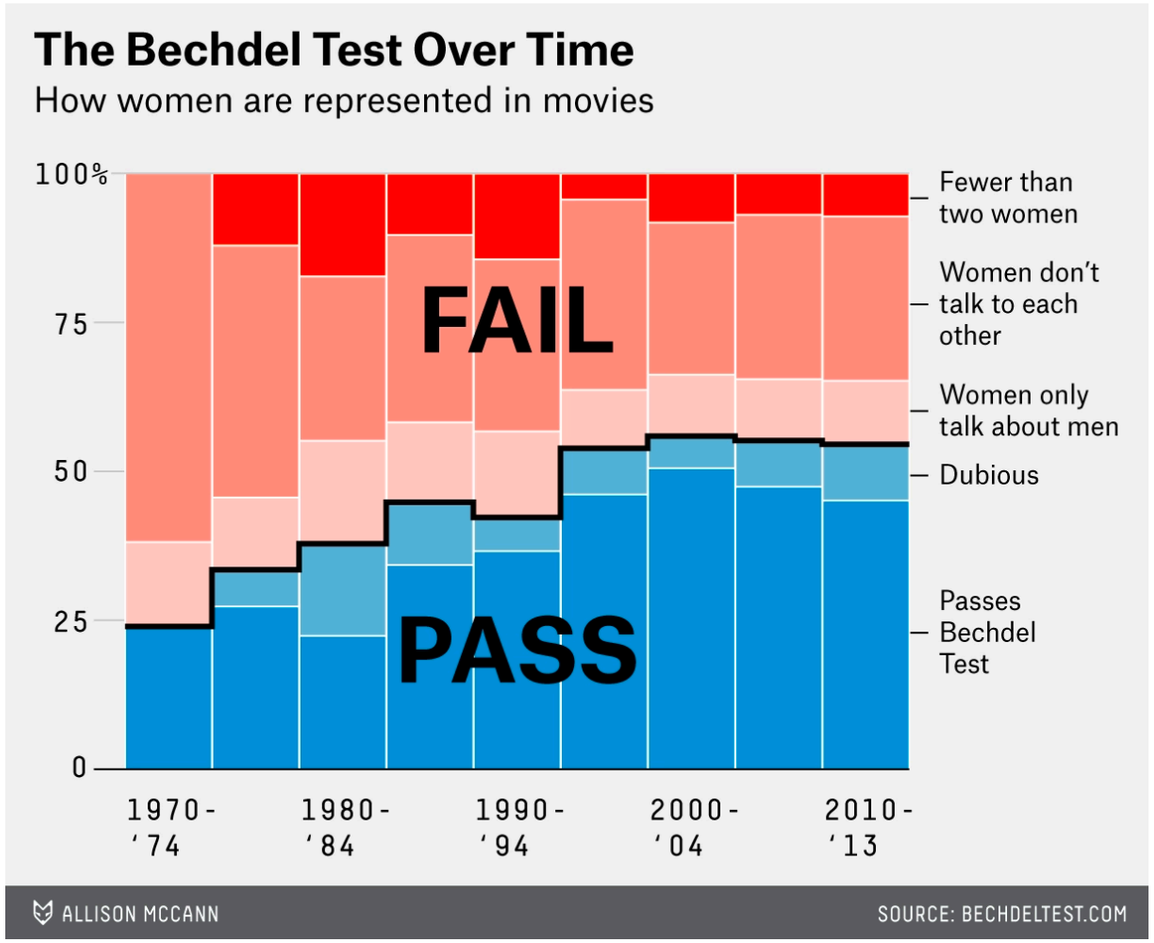
# *Why data?*

Data gives us one way to learn and answer questions about the world.

We can contrast a "datafied" perspective with a more instinctual, holistic perspective.

**‘Staggering’: The Ten Largest, Youth, 1907.** Photograph: Albin Dahlström/Courtesy of Stiftelsen Hilma af Klints Verk


HUGH BONNEVILLE
SALLY HAWKINS
BRENDAN GLEESON
JULIE WALTERS
JIM BROADBENT
PETER CAPALDI
WITH HUGH GRANT
AND BEN WHISHAW
AS THE VOICE OF PADDINGTON
PADDINGTON 2
IT TAKES A BEAR TO CATCH A THIEF
IN CINEMAS NOVEMBER 10
/PADDINGTONBEAR
#PADDINGTON2


**Alison Bechdel**

**FiveThirtyEight**

# Discussion Questions

1. Can you think of any examples of times when you have relied on data to make a decision or form an opinion?

2. Are there settings where you think it is really important to use quantitative data for decision making? Are there settings in which quantitative data are less valuable?

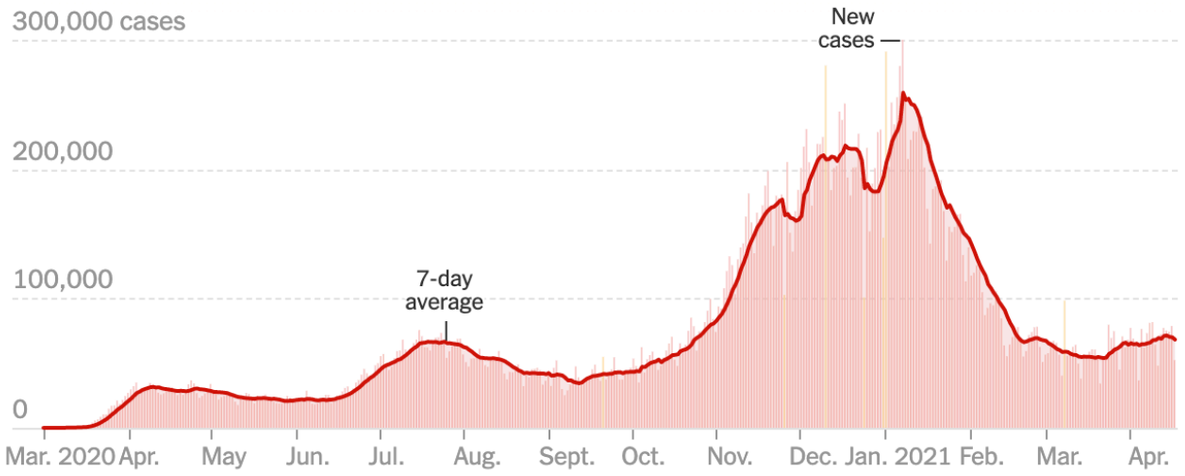
# Uses of data

# Data as description

## Coronavirus in the U.S.: Latest Map and Case Count

Updated April 18, 2021, 12:13 A.M. E.T.

Leer en español




| | TOTAL REPORTED | ON APRIL 17 | 14-DAY CHANGE |
|---|---|---|---|
| **Cases** | 31.6 million+ | 52,391 | +5% → |
| **Deaths** | 566,452 | 674 | –12% → |
| **Hospitalized** | | 45,497 | +9% → |

Day with reporting anomaly. Hospitalization data from the U.S. Department of Health and Human Services; 14-day change trends use 7-day averages.

https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

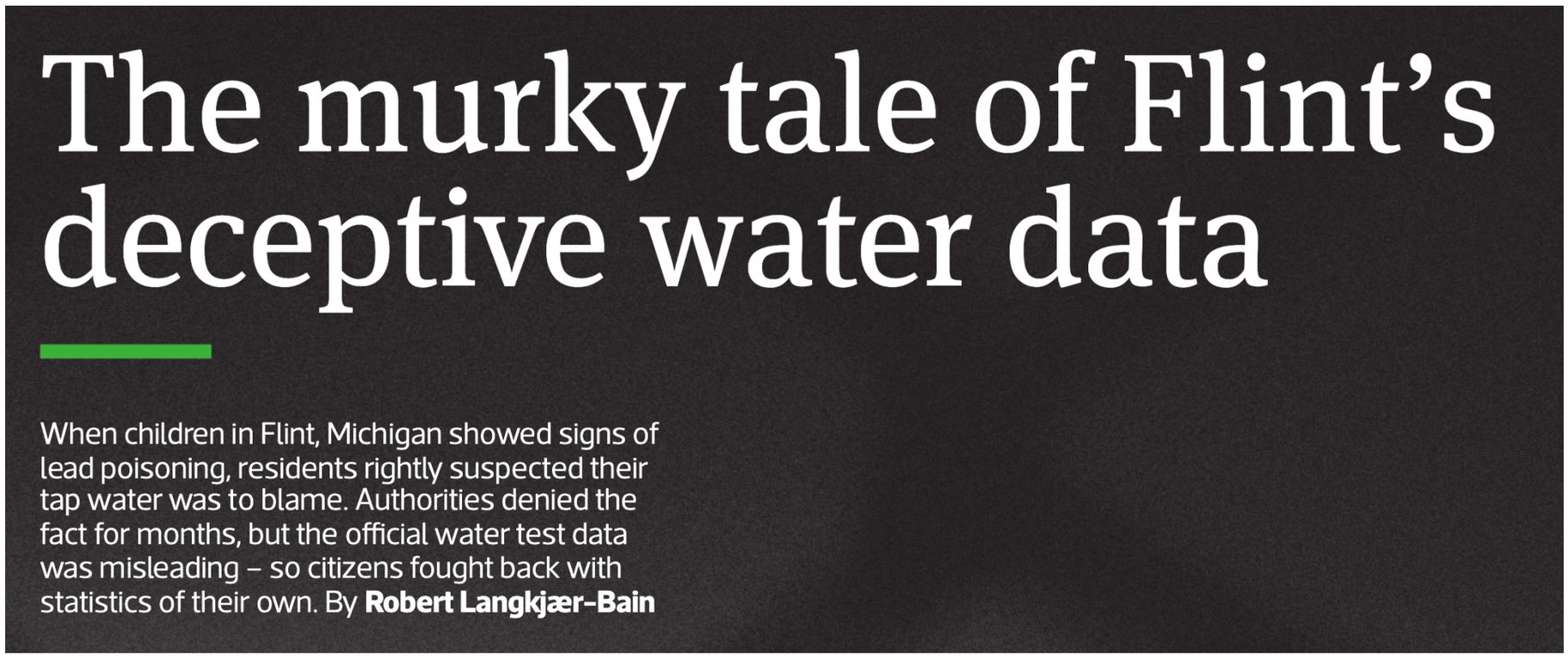
# Data for prediction

**The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever**

By Dan Jackson Published on 7/7/2017 at 2:58 PM

https://www.thrillist.com/entertainment/nation/the-netflix-prize

# Data as evidence



The murky tale of Flint's deceptive water data

When children in Flint, Michigan showed signs of lead poisoning, residents rightly suspected their tap water was to blame. Authorities denied the fact for months, but the official water test data was misleading – so citizens fought back with statistics of their own. By **Robert Langkjær-Bain**

https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2017.01016.x

# *Why data?*

If we agree that data can accurately (objectively?) describe the world, data can help give us a common toolbox/language to talk about issues and answer questions.

# Why data?

Or maybe not so unpredictable. In the early eighteen-thirties, a Belgian astronomer and mathematician named Adolphe Quetelet analyzed the numbers and discovered a remarkable pattern. The crime records were startlingly consistent. Year after year, irrespective of the actions of courts and prisons, the number of murders, rapes, and robberies reached almost exactly the same total. There is a "terrifying exactitude with which crimes reproduce themselves," Quetelet said. "We know in advance how many individuals will dirty their hands with the blood of others. How many will be forgers, how many poisoners."

To Quetelet, the evidence suggested that there was something deeper to discover. He developed the idea of a "Social Physics," and began to explore the possibility that human lives, like planets, had an underlying mechanistic trajectory. There's something unsettling in the idea that, amid the vagaries of choice, chance, and circumstance, mathematics can tell us something about what it is to be human. Yet Quetelet's overarching findings still stand: at some level, human life can be quantified and predicted. We can now forecast, with remarkable accuracy, the number of women in Germany who will choose to have a baby each year, the number of car accidents in Canada, the number of plane crashes across the Southern Hemisphere, even the number of people who will visit a New York City emergency room on a Friday evening.

# Discussion Questions

1. Can you think of any examples where collecting data makes something more “real”?

# Histories of quantification

# Why have a census?

"Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence Plantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three."

**Article I, Section II, U.S. Constitution**

# *Why do we count?*

The initial purpose of the U.S. census was simply to allocate representation and tax burdens among the states.

Even today, the census is limited to eleven questions, mostly about basic demographic information.

The American Community Survey provide researchers, government officials, and businesses with additional information about income, education, employment, and more.

"Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, ***three fifths of all other Persons***. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence Plantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three."

**Article I, Section II, U.S. Constitution**

# *Privacy concerns*

US & Canada

## US Supreme Court blocks census citizenship question for now

28 June 2019

https://www.bbc.com/news/world-us-canada-48791272

# *Privacy concerns*

"In 2017 and 2018, the Census Bureau found that a data scientist who had access to commercial and public databases could match that information up with census statistics in a way that could identify as many as 17% of Americans who had completed the 2010 census."

**https://theconversation.com/census-2020-will-protect-your-privacy-more-than-ever-but-at-the-price-of-accuracy-130116**

# *Privacy versus legibility*

Ideally, the census allows the government to allocate resources and representation in a fair way.

In this way, the government takes away some of its subjects' privacy in exchange for (hopefully) achieving a fair allocation.

Today, we are often faced with trade-offs between **privacy** and "**legibility**"—governments and businesses often argue that users/subjects give up privacy in exchange for convenience/fairness/other benefits.

# *Who benefits from the census?*

What benefits do individuals get from the census?

Businesses?

The government?

Do we all benefit from the census equally?

"In the meantime the highly anticipated census of 1939 was expected to produce a new, definitive list of Jews in Germany. In addition to the regular schedule, a so-called supplementary card was delivered to each household. This card focused on just two additional topics: occupation (for military mobilization) and ancestry. The latter operationalized the Nuremberg race definitions, asking for each person in the household, "Were or are any of the grandparents full-Jewish by race?" with a column for each grandparent. The card includes example rows to make clear how households should respond. The fate of the fictional, perfectly Aryan Schmitz family (nein—nein—nein—nein) would be very different from that of their imaginary neighbor Solly Cohn (ja—ja—ja—ja). Once completed, each card was to be placed in a sealed envelope, a show of confidentiality that belied the murky reality of statistics in the Third Reich."

**Excerpt From: Andrew Whitby. "The Sum of the People." Apple Books.**

# *The development of statistics*

Especially in Europe, the development of statistics was often motivated by the desire to study and manage populations.

LE

SUICIDE

ÉTUDE DE SOCIOLOGIE

PAR

ÉMILE DURKHEIM

Professeur de Sociologie à la Faculté des Lettres
de l'Université de Bordeaux

PARIS

ANCIENNE LIBRAIRIE GERMER BAILLIÈRE ET Cie

FÉLIX ALCAN, ÉDITEUR

108, BOULEVARD SAINT-GERMAIN, 108

1897

"This ideology and the Nazi "race science" that developed to provide its pseudoscientific backing were not an entirely German phenomenon. Of course, Hitler's views were seen as extreme and repugnant by many at the time, both inside and outside Germany. But anti-Semitism was casually accepted, to varying degrees, in much of the world then. Meanwhile, the idea of the population as a body whose health could be measured and manipulated was also promoted in Britain, the United States, and beyond, under the guise of eugenics."

**Excerpt From: Andrew Whitby. "The Sum of the People." Apple Books.**

"Eugenics is the now-controversial science of manipulating human populations to improve their genetic "quality." The term was coined in 1883 by Francis Galton, an English polymath and another classifier of people. As well as inventing eugenics, Galton published the seminal text on fingerprints, building the scientific foundation for their use in forensics and identification. But he is best known as one of the founders of modern statistics. Because of the influence of Galton and his intellectual descendants, the history of statistics is inescapably entwined with that of eugenics."

**Excerpt From: Andrew Whitby. "The Sum of the People." Apple Books.**

# Discussion Questions

1. Are you worried about privacy and unethical uses of your data by the government? By private companies?

# When does data fail?

# When Proof Is Not Enough

Throughout history, evidence of racism has failed to effect change.

By Mimi Onuoha

Filed under 2020 Police Protests

Published Jul. 1, 2020

https://fivethirtyeight.com/features/when-proof-is-not-enough/

# Discussion Questions

1. Why does Onuoha say data/quantitative evidence may not be enough?
2. If data is not enough, what is "enough"?
3. Besides histories of police brutality, can you think of any other examples in which data/quantitative evidence may not be sufficient or necessary?

# Histories of algorithmic fairness

# Discussion Questions

1. What is Ochigame writing about and why?

2. How does Ochigame critique today's definition of "algorithmic fairness?"

> Risk classification soon surfaced controversies over racial discrimination: in 1881, life insurance corporations started to charge differential rates on the basis of race. Unlike cooperative insurers, whose policyholders paid the same rates regardless of age or health or race, the corporate insurance firms Prudential and Metropolitan imposed penalties on African American policyholders. When civil rights activists challenged this policy, the corporations claimed differences in average mortality rates across races as justification. According to historian Dan Bouk, in 1884, Massachusetts state representative Julius C. Chappelle—an African American man born in antebellum South Carolina—challenged the fairness of the policy and proposed a bill to forbid it. The bill's opponents invoked statistics of deaths, but Chappelle and his allies reframed the issue in terms of the future prospects of African Americans, emphasizing their potential for achieving equality.
>
> **Rodrigo Ochigame, "The Long History of Algorithmic Fairness"**