15 Case Study in Human Contexts and Ethics

Code

> **Learning Outcomes**
>
> - Learn about the ethical dilemmas that data scientists face.
> - Know how critique models using contextual knowledge about data.

> **Disclaimer**: The following chapter discusses issues of structural racism. Some of the items in the chapter may be sensitive and may or may not be the opinions, ideas, and beliefs of the students who collected the materials. The Data 100 course staff tries its best to only present information that is relevant for teaching the lessons at hand.

Note: Given the nuanced nature of some of the arguments made in the lecture, it is highly recommended that you view the lecture recording in order to fully engage and understand the material. The course notes will have the same broader structure but are by no means comprehensive.

Let's immerse ourselves in the real-world story of data scientists working for an organization called the Cook County Assessor's Office (CCAO). Their job is to **estimate the values of houses** in order to **assign property taxes**. This is because the tax burden in this area is determined by the estimated **value** of a house, which is different from its price. Since values change over time and there are no obvious indicators of value, they created a **model** to estimate the values of houses. In this chapter, we will dig deep into what problems biased the models, the consequences to human lives, and how we can learn from this example to do better.
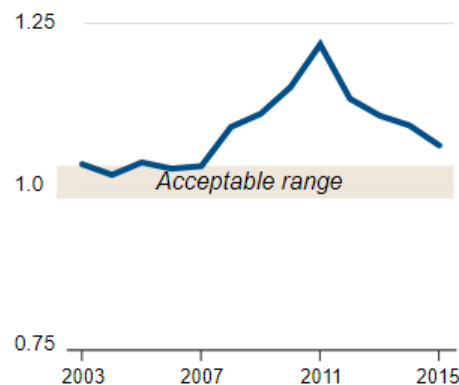
15.1 The Problem

A report by the Chicago Tribune uncovered a major scandal: the team showed that the model perpetuated a highly regressive tax system that disproportionately burdened African-American and Latinx homeowners in Cook County. How did they know?
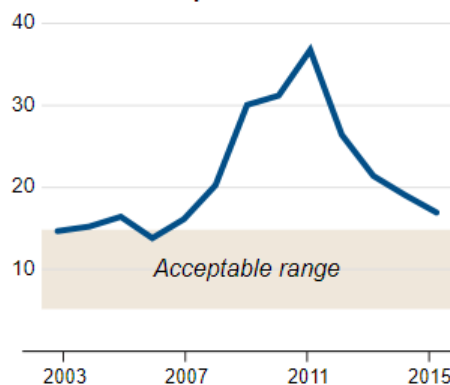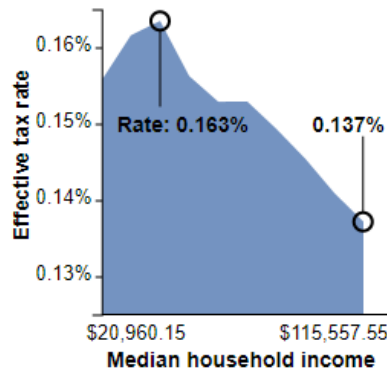

In the field of housing assessment, there are standard metrics that assessors use across the world to estimate the fairness of assessments: coefficient of dispersion and price-related differential. These metrics have been rigorously tested by experts in the field and are out of scope for our class. Calculating these metrics for the Cook County prices revealed that the pricing created by the CCAO did not fall in acceptable ranges (see figure above). This by itself is **not the end** of the story, but a good indicator that **something fishy was going on**.

## As income level rises, effective tax rates decline

These rates represent the percentage of homes' value that owners pay toward two taxing districts that cover all of Cook County. They are a small fraction of the overall tax rate but allow for comparisons between communities with widely differing tax bases.

This prompted them to investigate if the model itself was producing fair tax rates. Evidently, when accounting for the home owner's income, they found that the model actually produced a **regressive** tax rate (see figure above). A tax rate is **regressive** if the percentage tax rate is higher for individuals with lower net income. A tax rate is **progressive** if the percentage tax rate is higher for individuals with higher net income.

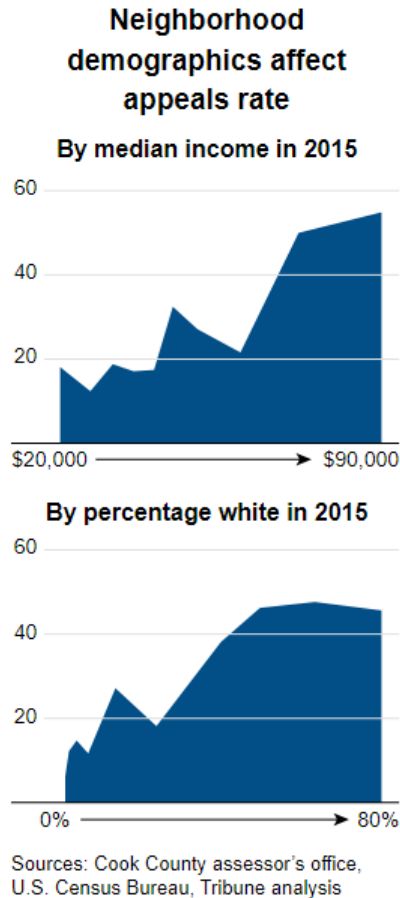Chart 2: Average Percent Over/Under Assessment by Percent White, 2011-2015
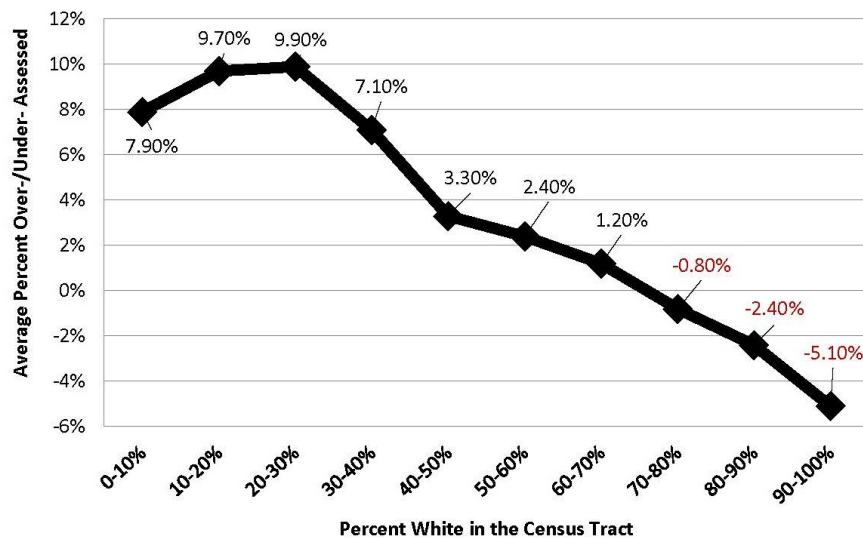
Further digging suggests that not only was the system unfair to people across the axis of income, it was also unfair across the axis of race (see figure above). The likelihood of a property being under- or over-assessed was highly dependent on the owner's race, and that did not sit well with many homeowners.
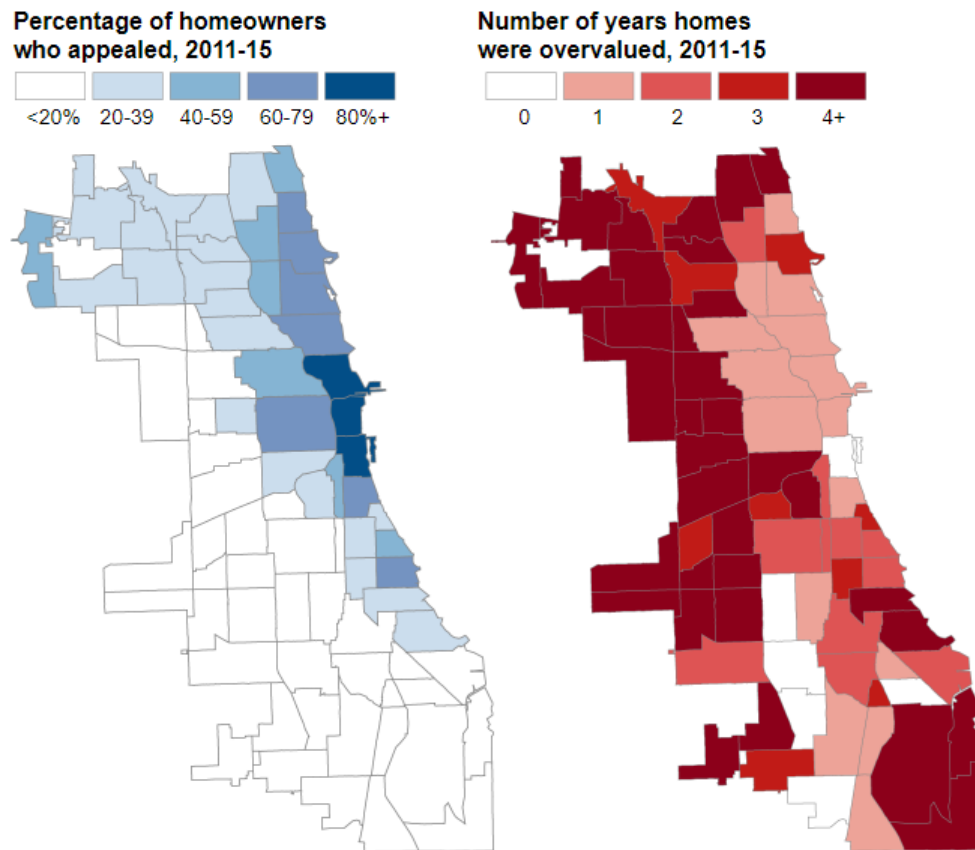
15.1.1 Spotlight: Appeals

What actually caused this to come about? A comprehensive answer goes beyond just models. At the end of the day, these are real systems that have a lot of moving parts. One of which was the **appeals system**. Homeowners are mailed the value their home assessed by CCAO, and the homeowner can choose to appeal to a board of elected officials to try and change the listed value of their home and thus how much they are taxed. In theory, this sounds like a very fair system: someone oversees the final pricing of houses rather than just an algorithm. However, it ended up exacerbating the problems.

> "Appeals are a good thing," Thomas Jaconetty, deputy assessor for valuation and appeals, said in an interview. "The goal here is fairness. We made the numbers. We can change them."

Neighborhood demographics affect appeals rate

By median income in 2015

$20,000 ────────→ $90,000

By percentage white in 2015

0% ────────→ 80%

Sources: Cook County assessor's office, U.S. Census Bureau, Tribune analysis

Here we can borrow lessons from Critical Race Theory. On the surface, everyone having the legal right to try and appeal is undeniable. However, not everyone has an equal ability to do so. Those who have the money to hire tax lawyers to appeal for them have a drastically higher chance of trying and succeeding (see above figure). This model is part of a deeper institutional pattern rife with potential corruption.

Percentage of homeowners who appealed, 2011-15
<20% | 20-39 | 40-59 | 60-79 | 80%+

Number of years homes were overvalued, 2011-15
0 | 1 | 2 | 3 | 4+

Sources: Cook County assessor's office, Illinois Department of Revenue, Tribune analysis

Homeowners who appealed were generally under-assessed relative to homeowners who did not (see above figure). Those with higher incomes pay less in property tax, tax lawyers are able to grow their business due to their role in appeals, and politicians are commonly socially connected to the aforementioned tax lawyers and wealthy homeowners. All these stakeholders have reasons to advertise the model as an integral part of a fair system. Here lies the value in asking questions: a system that seems fair on the surface may in actuality be unfair upon taking a closer look.

15.1.2 Human Impacts

In 2011 Braxton-Williams learned the assessor's office had valued the house at $147,550. "I love my house, but I know it's not worth that much," she said. "And they know it's not worth that much." (Terrence Antonio James/Chicago Tribune)

The impact of the housing model extends beyond the realm of home ownership and taxation. Discriminatory practices have a long history within the United States, and the model served to perpetuate this fact. To this day, Chicago is one of the most segregated cities in the United States (source). These factors are central to informing us, as data scientists, about what is at stake.

15.1.3 Spotlight: Intersection of Real Estate and Race

Housing has been a persistent source of racial inequality throughout US history, amongst other factors. It is one of the main areas where inequalities are created and reproduced. In the beginning, Jim Crow laws were explicit in forbidding people of color from schools, public utilities, etc.
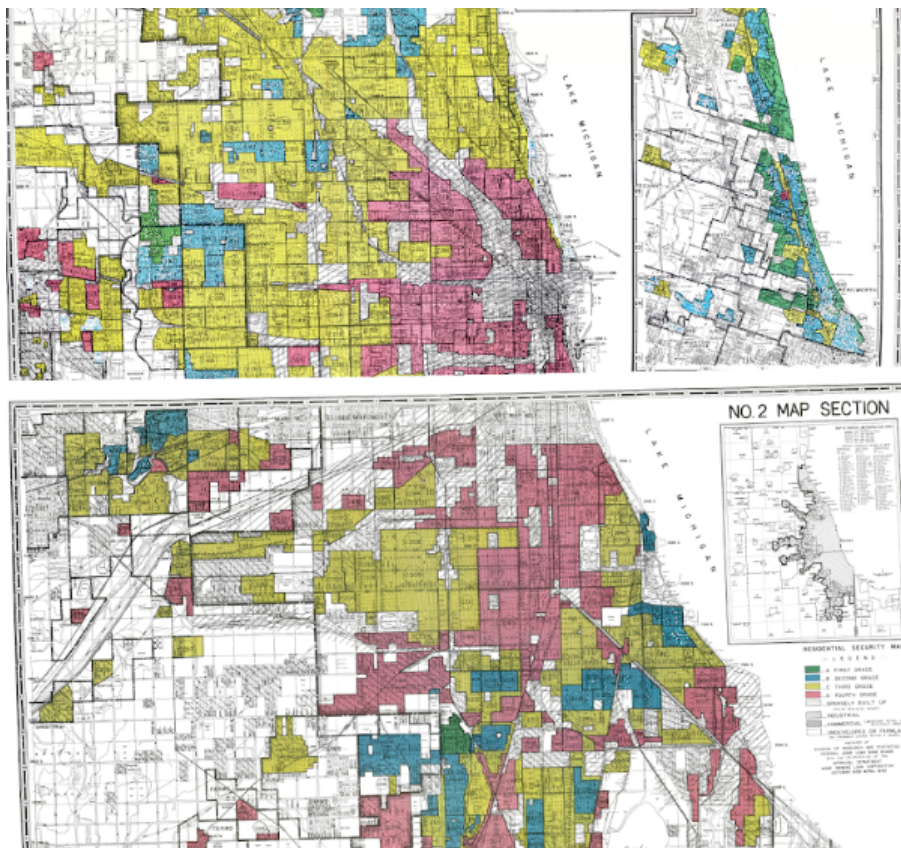
Today, while advancements in Civil Rights have been made, the spirit of the laws are alive in many parts of the US. The real estate industry was "professionalized" in the 1920's and 1930's by aspiring to become a science guided by strict methods and principles outlined below:

- Redlining: making it difficult or impossible to get a federally-backed mortgage to buy a house in specific neighborhoods coded as "risky" (red).
  - What made them "risky" according to the makers of these was racial composition.
  - Segregation was not only a result of federal policy, but developed by real estate professionals.
- The methods centered on creating objective rating systems (information technologies) for the appraisal of property values which encoded **race** as a factor of valuation (see figure below),
  - This, in turn, influenced federal policy and practice.

ADDITIONS	Per Cent
Site	15
Type of neighborhood and social factors	20
View and climate	15
Public utilities and schools	5
Streets and alleys; distance to work in city	10
Contour and soil	5
Physical environment	20
Restrictions and planning	10

A table of common deductions reducing the above percentages of value follows:

DEDUCTIONS	Per Cent
Noise and dirt, up to	25
Racial and foreign neighbors, up to	60
Adjacent vacancy, up to	20
Poor architecture, up to	20
Obsolescence, up to	70
Distances from city, work, schools, etc., up to	100
Nuisances (funerals, freight, trucks, etc.), up to	100
Dead-end streets, up to	15

Source: Colin Koopman, How We Became Our Data (2019) p. 137

15.2 The Response: Cook County Open Data Initiative

The response started in politics. A new assessor, Fritz Kaegi, was elected and created a new mandate with two goals:

1. Distributional equity in property taxation, meaning that properties of same value treated alike during assessments.
2. Creating a new Office of Data Science.

15.2.1 Question/Problem Formulation

Driving Questions

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?

The new Office of Data Science started by redefining their goals.

1. Accurately, uniformly, and impartially assess the value of a home by
   - Following international standards (coefficient of dispersion)
   - Predicting value of all homes with as little total error as possible

2. Create a robust pipeline that accurately assesses property values at scale and is fair by
   - Disrupts the circuit of corruption (Board of Review appeals process)
   - Eliminates regressivity
   - Engenders trust in the system among all stakeholders

Definitions: Fairness and Transparency

The definitions, as given by the Cook County Assessor's Office, are given below:

- Fairness: The ability of our pipeline to accurately assess property values, accounting for disparities in geography, information, etc.
- Transparency: The ability of the data science department to share and explain pipeline results and decisions to both internal and external stakeholders

Note how the Office defines "fairness" in terms of accuracy. Thus, the problem - make the system more fair - was already framed in terms amenable to a data scientist: make the assessments more accurate.

The idea here is that if the model is more accurate it will also (perhaps necessarily) become more fair, which is a big assumption. There are, in a sense, two different problems - make accurate assessments, and make a fair system.

The way the goals are defined lead us to ask the question: what does it actually mean to accurately assess property values, and what role does "scale" play?

1. What is an assessment of a home's value?
2. What makes one assessment more accurate than another?
3. What makes one batch of assessments more accurate than another batch?

Each of the above questions leads to a slew of more questions. Considering just the first question, one answer could be that an assessment is an estimate of the value of a home. This leads to more inquiries: what is the value of a home? What determines it? How do we know? For this class, we take it to be the house's market value.

15.2.2 Data Acquisition and Cleaning

Driving Questions
- What data do we have, and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?

The data scientists also critically examined their original sales data:

Feature Name	Category	Type
Age	Characteristic	numeric
Central Air Conditioning	Characteristic	categorical
Apartments	Characteristic	categorical
Attic Finish	Characteristic	categorical
Attic Type	Characteristic	categorical
Bedrooms	Characteristic	numeric
Building Square Feet	Characteristic	numeric
Basement	Characteristic	categorical
Basement Finish	Characteristic	categorical
Wall Material	Characteristic	categorical
Full Baths	Characteristic	numeric
Fireplaces	Characteristic	numeric
Garage 1 Area	Characteristic	categorical
Garage 1 Attached	Characteristic	categorical
Garage 1 Material	Characteristic	categorical

and asked the questions:

1. How was this data collected?
2. When was this data collected?
3. Who collected this data?
4. For what purposes was the data collected?
5. How and why were particular categories created?

For example, attributes can have different likelihoods of appearing in the data, and housing data in the floodplains geographic region of Chicago were less represented than other regions.

The features can even be reported at different rates. Improvements in homes, which tend to increase property value, were unlikely to be reported by the homeowners.

Additionally, they found that there was simply more missing data in lower income neighborhoods.

15.2.3 Exploratory Data Analysis

Driving Questions

- How is our data organized, and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

Before the modeling step, they investigated a multitude of crucial questions:

1. Which attributes are most predictive of sales price?
2. Is the data uniformly distributed?

3. Do all neighborhoods have up to date data? Do all neighborhoods have the same granularity?
4. Do some neighborhoods have missing or outdated data?

Firstly, they found that the impact of certain features, such as bedroom number, were much more impactful in determining house value inside certain neighborhoods more than others. This informed them that different models should be used depending on the neighborhood.

They also noticed that low income neighborhoods had disproportionately spottier data. This informed them that they needed to develop new data collection practices - including finding new sources of data.

15.2.4 Prediction and Inference

> **Driving Questions**
> - What does the data say about the world?
> - Does it answer our questions or accurately solve the problem?
> - How robust are our conclusions and can we trust the predictions?

Rather than using a singular model to predict sale prices ("fair market value") of unsold properties, the CCAO fit machine learning models that discover patterns using known sale prices and characteristics of **similar and nearby properties**. It uses different model weights for each township.

Compared to traditional mass appraisal, the CCAO's new approach is more granular and more sensitive to neighborhood variations.

Here, we might ask why should any particular individual believe that the model is accurate for their property?

This leads us to recognize that the CCAO counts on its performance of "transparency" (putting data, models, pipeline onto GitLab) to foster public trust, which would help it equate the production of "accurate assessments" with "fairness".

There's a lot more to be said here on the relationship between accuracy, fairness, and metrics we tend to use when evaluating our models. Given the nuanced nature of the argument, it is recommended you view the corresponding lecture as the course notes are not as comprehensive for this portion of lecture.

15.2.5 Reports Decisions, and Conclusions

> **Driving Questions**
> - How successful is the system for each goal?
>   - Accuracy/uniformity of the model
>   - Fairness and transparency that eliminates regressivity and engenders trust
> - How do you know?

The model is not the end of the road. The new Office still sends homeowners their house evaluations, but now the data that they get sent back from the homeowners is taken into account. More detailed reports are being written by the Office itself to democratize the information. Town halls and other public facing outreach helps involves the whole community in the process of housing evaluations, rather than limiting participation to a select few.

15.3 Key Takeaways

1. Accuracy is a necessary, but not sufficient, condition of a fair system.

2. Fairness and transparency are context-dependent and sociotechnical concepts.

3. Learn to work with contexts, and consider how your data analysis will reshape them.

4. Keep in mind the power, and limits, of data analysis.

15.4 Lessons for Data Science Practice

1. Question/Problem Formulation

   - Who is responsible for framing the problem?
   - Who are the stakeholders? How are they involved in the problem framing?
   - What do you bring to the table? How does your positionality affect your understanding of the problem?
   - What are the narratives that you're tapping into?

2. Data Acquisition and Cleaning

- Where does the data come from?
- Who collected it? For what purpose?
- What kinds of collecting and recording systems and techniques were used?
- How has this data been used in the past?
- What restrictions are there on access to the data, and what enables you to have access?

3. Exploratory Data Analysis & Visualization

- What kind of personal or group identities have become salient in this data?
- Which variables became salient, and what kinds of relationship obtain between them?
- Do any of the relationships made visible lend themselves to arguments that might be potentially harmful to a particular community?

4. Prediction and Inference

- What does the prediction or inference do in the world?
- Are the results useful for the intended purposes?
- Are there benchmarks to compare the results?
- How are your predictions and inferences dependent upon the larger system in which your model works?

5. Reports, Decisions, and Solutions

- How do we know if we have accomplished our goals?
- How does your work fit in the broader literature?
- Where does your work agree or disagree with the status quo?
- Do your conclusions make sense?