SMU
School of
Computing and
Information Systems

2021-22 Term 1

IS457: Fairness in Socio-technical Systems

Week 5 - Bias in data and machine learning models (I)

KWAK Haewoon

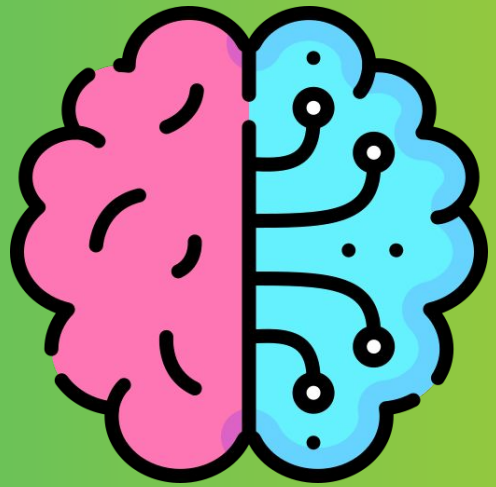


Image from Flaticon.com

Study questions

How biased (or unbiased) is our everyday decision-making?

What are two cultures of statistical modeling?

What are five mechanisms of occurring biases in data mining?

ARTICLES

Judgment under Uncertainty: Heuristics and Biases

Amos Tversky[1], Daniel Kahneman[1]

+ See all authors and affiliations

Science 27 Sep 1974:
Vol. 185, Issue 4157, pp. 1124-1131
DOI: 10.1126/science.185.4157.1124

https://science.sciencemag.org/content/185/4157/1124

Human decision making and uncertain events

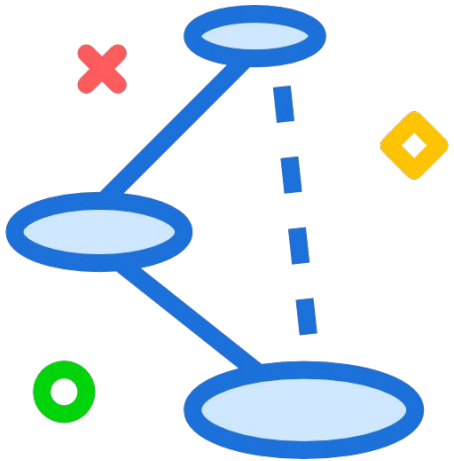
Many decisions are based on beliefs concerning the likelihood of uncertain events.

How do people assess the probability of an uncertain event or the value of an uncertain quantity?

People rely on heuristic principles

People transform the complex tasks of assessing probabilities to simpler judgemental operations.

These heuristics are quite useful but sometimes lead to severe and systematic errors.



Errors in subjective judgments

Someone thinks that the distance of an object can be estimated by its clarity.

- The more sharply the object is seen, the closer it appears to be.
- When does this estimation work poorly?

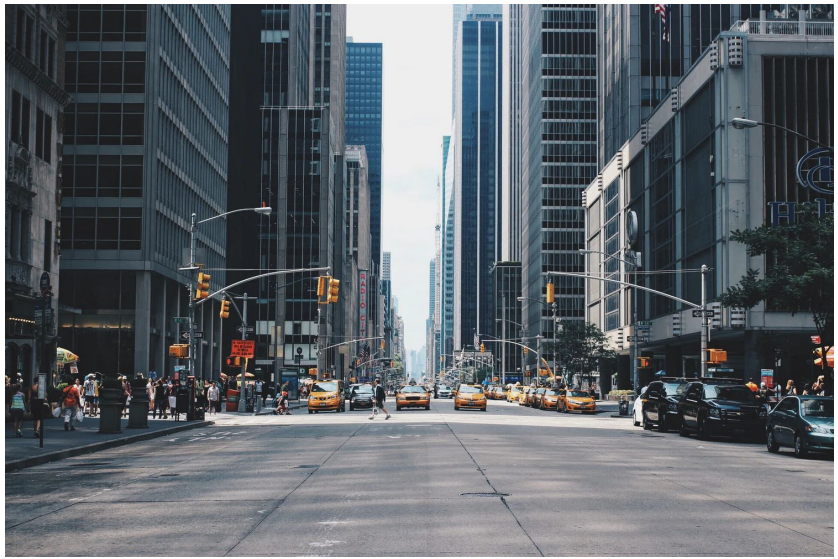


Image from https://unsplash.com/photos/nV7GJmSq3zc

Three common heuristic techniques

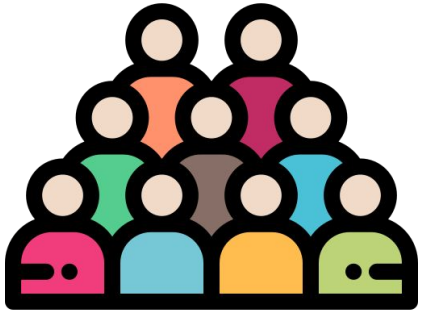
Representativeness

Availability

Adjustment and anchoring

An example scenario

A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, brief descriptions of the 30 engineers and 70 lawyers have been written.



Kahneman, Daniel, and Amos Tversky. "On the psychology of prediction." Psychological review 80.4 (1973): 237.

WooClap #Q1

Suppose now that you are given no information whatsoever about an individual chosen at random from the sample. The probability that this man is one of the 30 engineers in the sample of 100 is ___%.

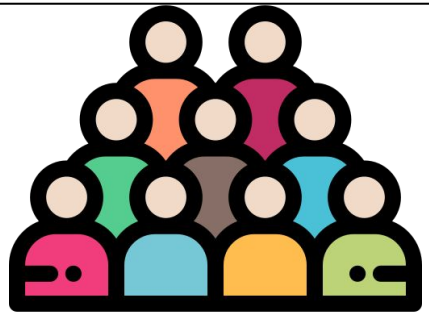
Kahneman, Daniel, and Amos Tversky. "On the psychology of prediction." Psychological review 80.4 (1973): 237.

WooClap #Q2

Suppose now you have a description like below:

James is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

The probability that James is one of the 30 engineers in the sample of 100 is ___%.



Kahneman, Daniel, and Amos Tversky. "On the psychology of prediction." Psychological review 80.4 (1973): 237.

WooClap #Q2: Explained

The description about James was deliberately constructed to be totally neutral (uninformative) with regard to an engineer or a lawyer.

Thus, the answer to Q2 should be the same as that to Q1.

In the study (n=171):

- When no evidence is given (=Q1), prior probabilities are properly utilized.
- When worthless specific evidence is given (=Q2), prior probabilities are ignored (They answered **50%-50%**!)

Heuristics #1. Representativeness

What is the probability that object **A** belongs to class **B**?

What is the probability that event **A** originates from process **B**?

What is the probability that process **B** will generate event **A**?

To answer these Qs, people typically rely on the representativeness heuristic, in which probabilities are evaluated by the degree to which A is representative of B, that is, by the degree to which A resembles B.

Consider another example

> Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Rank a list of occupations [farmer, salesman, airline pilot, librarian, physician] based on the probability that Steve will be engaged in.

How do people order these occupations from most to least likely?

Judgment by representativeness

People assess the probability that Steve becomes a librarian by the degree to which he is representative of (or similar to) the stereotype of a librarian.

But, is it a reasonable estimate?

Origin of errors: Insensitivity to prior probability

One of the factors that have no effect on representativeness but have a major effect on probability is the **prior probability**, or base-rate frequency, of the outcomes.

Back to the case of Steve:

The fact that there are many more farmers than librarians in the population should enter into any reasonable estimate of the probability that Steve is a librarian rather than a farmer.

WooClap #Q3

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

- The larger hospital
- The smaller hospital
- About the same (that is, within 5 percent of each other)

Kahneman, Daniel, and Amos Tversky. "Subjective probability: A judgment of representativeness." Cognitive psychology 3.3 (1972): 430-454.

WooClap #Q3: Explained

In the study (n=95),

- The larger hospital (21)
- The smaller hospital (21)
- About the same (that is, within 5 percent of each other) (53)

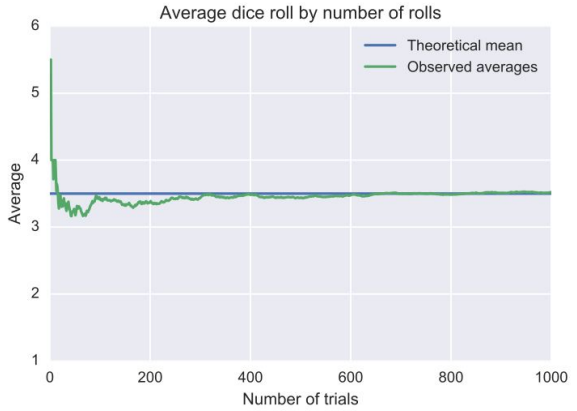
Since these events are described by the same statistic (50% of all babies are boys), people think that both hospitals are equally representative of the general population.

Kahneman, Daniel, and Amos Tversky. "Subjective probability: A judgment of representativeness." Cognitive psychology 3.3 (1972): 430-454.

WooClap #Q3: Explained

Law of large numbers:

The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.



Thus, the expected number of days on which more than 60 percent of the babies are boys is much greater in the small hospital than in the large one.

https://en.wikipedia.org/wiki/Law_of_large_numbers

Origin of errors: Insensitivity to sample size

One of the factors that have no effect on representativeness but have a major effect on probability is the **sample size**.

WooClap #Q4

In considering tosses of a coin for heads or tails, which sequence is more likely to happen?

- `H-T-H-T-T-H`
- `H-H-H-T-T-T`
- `H-H-H-H-T-H`
- Above 3 have the same probability.

WooClap #Q4: Explained

People expect that a sequence of events generated by a random process (e.g., coin toss) will represent the essential characteristics of that process even when the sequence is short.

Thus, people regard the sequence `H-T-H-T-T-H` to be more likely than the sequence `H-H-H-T-T-T` and `H-H-H-H-T-H`, which do not appear random.

Heuristics based on representativeness lead to errors

In summary, prior probability, sample size, and chance should be considered in judgments of probability, but they do not influence representativeness (similarity).

Thus, heuristics based on representativeness lead to serious errors.

WooClap #Q5

Consider the letter R. Is R more likely to appear in:

- the first position of a word?
- the third position?

(My estimate for the ratio of these two values is ___ : 1)

WooClap #Q5: Explained

In the study (n=152),

- 105 judged the first position to be more likely for a majority of the letters (K, L, N, R, V).
- The median estimated ratio was 2:1.

It is certainly easier to think of words that start with R than of words where R is in the third position. Thus, words that start with R should be judged more frequent.

(In fact, all letters were more frequent in the third position.)

WooClap #Q6

Consider the two structures, A and B, which are shown below.

(A) 8 cols x 3 rows

```
X X X X X X X X
X X X X X X X X
X X X X X X X X
```

(B) 2 cols x 9 rows

```
X X
X X
X X
X X
X X
X X
X X
X X
X X
```

A path in a structure is a line that connects an element in the top row to an element in the bottom row, and passes through one and only one element in each row. In which of the two structures are there more paths?

WooClap #Q6: Explained

In the study (n=54), 46 saw more paths in A than in B.

The paths in A more available than those in B:

- The most immediately available paths are the columns of the structures. There are 8 columns in A and only 2 in B.
- Among the paths that cross columns, those of A are generally more distinctive and less confusable than those in B.
- The paths in A are shorter and hence easier to visualize than those in B.

Heuristics #2. Availability

People assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind.

- One may assess the risk of heart attack among middle-aged people by recalling such occurrences among one's acquaintances.

Availability can be a useful clue for assessing frequency or probability because instances of large classes are usually recalled better and faster than instances of less frequent classes.

Heuristics based on availability lead to errors

Availability can be affected by retrievability, imaginability, or some other factors rather than frequency or probability.

Thus, the reliance on availability leads to predictable biases.

Heuristics #3. Adjustment and anchoring

In many situations, people make estimates by starting from an initial value that is adjusted to yield the final answer.

Different starting points yield different estimates, which are biased toward the initial values.

Errors in adjustment and anchoring

Estimate the product in 5 seconds by high-school students.

8x7x6x5x4x3x2x1 (median estimation: 2,250)

1x2x3x4x5x6x7x8 (median estimation: 512)

Experts also rely on heuristics

The reliance on heuristics and the prevalence of biases are not restricted to laymen.

Experienced researchers are also prone to the same biases — when they think intuitively.

More cognitive biases (1)

Confirmation Bias: tendency to listen more often to information that confirms our existing beliefs.

Hindsight Bias: tendency to see events, even random ones, as more predictable than they are.

Misinformation Effect: tendency for memories to be heavily influenced by things that happened after the actual event itself.

Actor-Observer Bias: tendency to attribute our actions to external influences and other people's actions to internal ones.

https://www.verywellmind.com/cognitive-biases-distort-thinking-2794763

More cognitive biases (2)

False Consensus Effect: tendency people have to overestimate how much other people agree with their own beliefs, behaviors, attitudes, and values.

Halo Effect: tendency for an initial impression of a person to influence what we think of them overall.

Self-Serving Bias: tendency for people tend to give themselves credit for successes but lay the blame for failures on outside causes.

Optimism Bias: tendency to overestimate the likelihood that good things will happen to us while underestimating the probability that negative events will impact our lives.

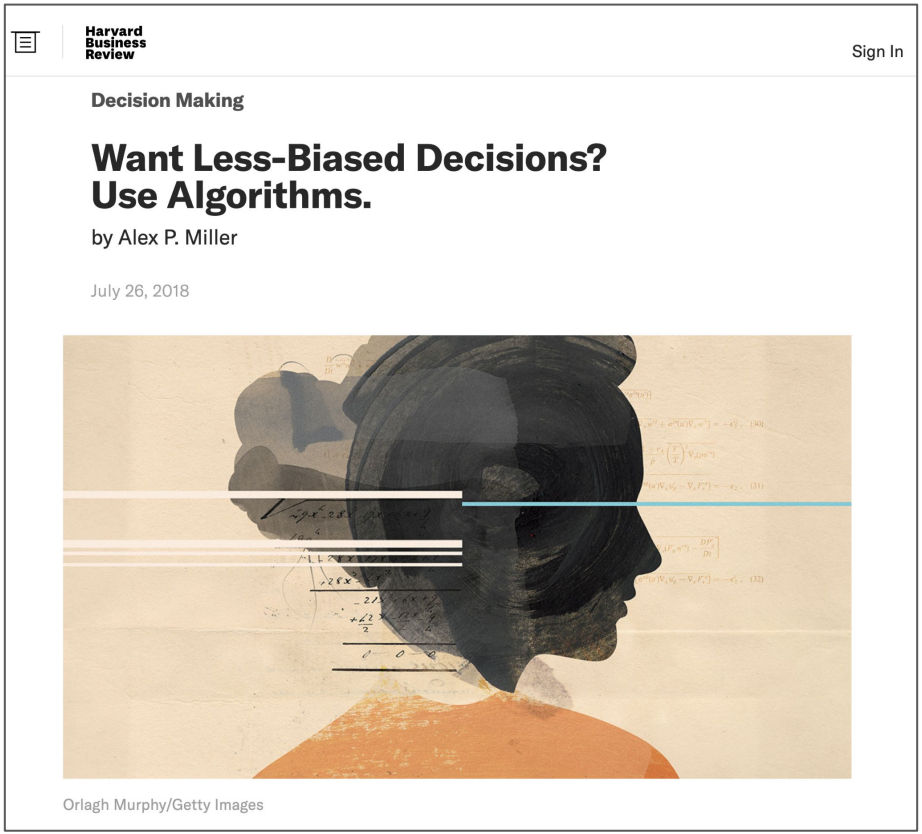
https://www.verywellmind.com/cognitive-biases-distort-thinking-2794763

Group activity

For each bias, please share your experience in your everyday life and write 2 stories on the spreadsheet.

https://docs.google.com/spreadsheets/d/11xGpTqhApo5YdleZBQbEwAhhqsrs6pulOmCeF0gM6eE/edit?usp=sharing

Cognitive biases in decision making



https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Leo Breiman

Leo Breiman in 2003

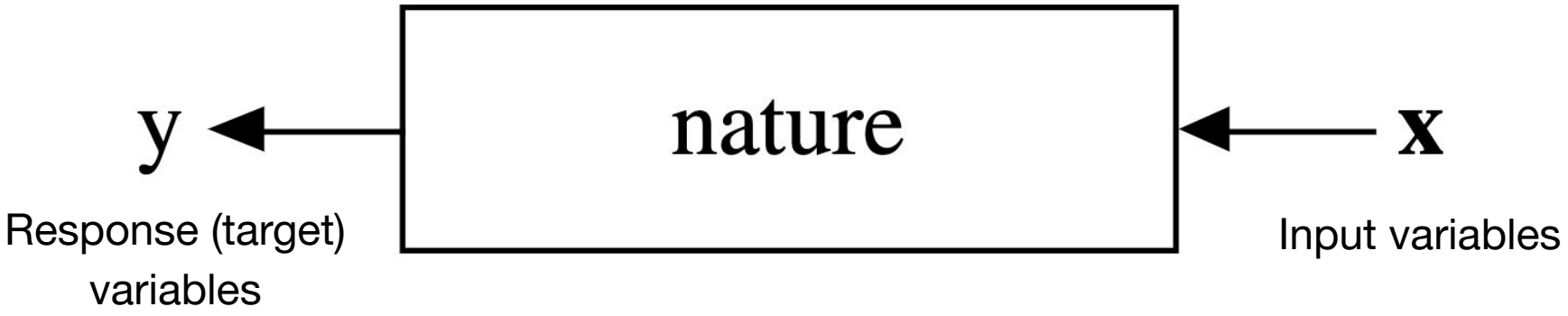
Born	January 27, 1928 New York City, United States
Died	July 5, 2005 (aged 77) Berkeley, California, United States
Nationality	American
Alma mater	University of California, Berkeley
Known for	CART, Bagging, Random forest

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.

Two goals of statistical modeling

Prediction: what the responses are going to be to future input variables.

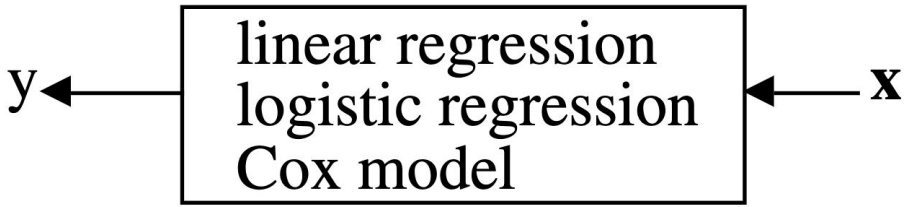
Information: how nature is associating the response variables to the input variables.



Two modeling approaches #1: Data modeling

Assumes a model for the inside of the black box.

Estimate the values of the parameters of the assumed model from the data.

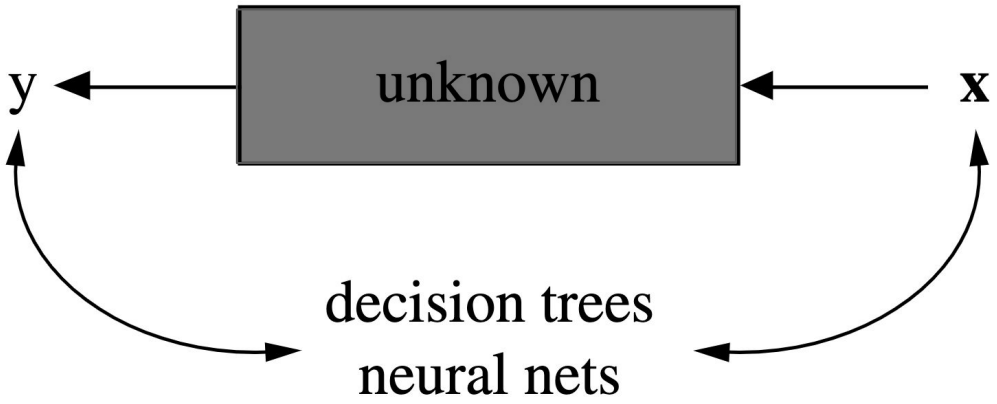


$$\text{response variables} = f(\text{predictor variables}, \text{random noise}, \text{parameters})$$

Two modeling approaches #2: Algorithmic modeling

Considers the **inside of the box complex and unknown**.

Find a function f(***x***)—an algorithm that operates on ***x*** to predict the responses ***y***.



Simplicity vs. accuracy

Accuracy generally requires more complex prediction methods.

Simple and interpretable functions do not make the most accurate predictors.

In general, algorithmic models can give better predictive accuracy than data models.

In practice, using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why.

Multiplicity of good models

Suppose there are 30 variables and we want to find the best five variable linear regressions (${}_nC_r = {}_{30}C_5 = 142{,}506$ possible subsets). Usually we pick the one with the lowest residual sum-of-squares (RSS), or, if there is a test set, the lowest test error. But there may be (and generally are) many five-variable equations that have RSS within 1.0% of the lowest RSS.

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

Which one is better?
Each one tells a completely different story.

Breiman, Leo. "Heuristics of instability and stabilization in model selection." The annals of statistics 24.6 (1996): 2350-2383.

Five mechanisms where biases might occur

1. Defining the target variable and class labels
2. Collecting the training data
3. Feature selection
4. Proxies
5. Masking

Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." Calif. L. Rev. 104 (2016): 671.

Data mining and a model

Data mining

- Finds statistical relationships in a dataset.
- Automates the process of discovering useful patterns.

Model

- A set of discovered relationships.
- Can be employed for classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes.

Fayyad, Usama. "The digital physics of data mining." Communications of the ACM 44.3 (2001): 62-65
Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.

1. Target variables and class labels

The algorithm “learns” which related attributes or activities can serve as potential predictors for those qualities or outcomes of interest (p39).

Target variables: The outcomes of interest.

Class labels: All possible values of the target variable into mutually exclusive categories.

Subjective process of specifying target variables

The proper specification of the target variable is frequently not obvious.

Data miners understand the project objectives and requirements, and convert this knowledge into a data mining problem definition.

Through this necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes.

E.g., Definition of “good” employees in hiring

Consider an employer who wants to develop ways of improving and automating their search for good employees.

How can good employees be defined in the first place?

- “Good” must be defined in ways that correspond to measurable outcomes.

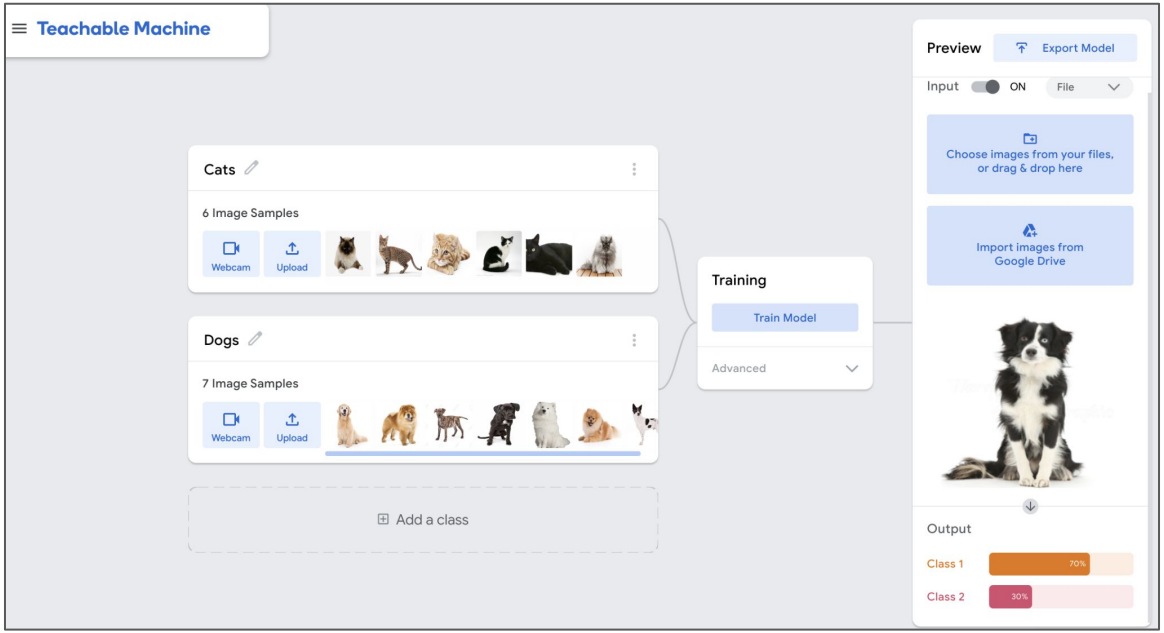
Danger at this stage has been neglected

While critics of data mining have tended to focus on inaccurate classifications, the same amount of danger also resides in the definition of the class label.

The different choices may have a greater adverse impact on protected classes.

2. Training data

The algorithm learns through the examples to which it has been exposed. These examples are called “training data.”



Labeling “examples”

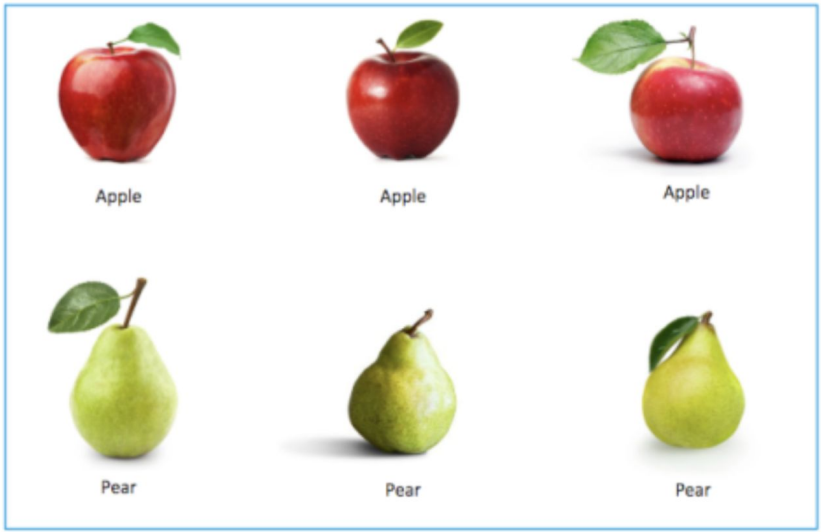
The process by which the training data is manually assigned class labels.

When prelabeled examples are available: spam mails, reviewed performance, etc.

When no prelabeled examples exist: data miners have to label examples. This can be a laborious process, and sometimes it is hard and can be biased.

Representative data collection

As data mining finds statistical relationships in a dataset, building a representative dataset is crucial.



Training Dataset

Test Photo

Is this an Apple or a Pear?

Image from https://developer.salesforce.com/blogs/2018/06/why-representative-datasets-are-important-for-computer-vision-models.html

Origin of biases in training data

Improper labeling of examples: if data mining treats cases in which prejudice has played some role as valid examples, the model may simply reproduce the prejudice.

Non-representative data collections: if data mining learns through a non-representative sample of the population, any decision based on these inferences may systematically disadvantage those who are under- or overrepresented in the dataset.

When past decisions are swayed by prejudice

St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants (1980s).

Training data: Previous admission decisions

Result: Those admissions decisions had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants.

As editors at the British Medical Journal noted at the time, "[T]he program was not introducing new bias but merely reflecting that already in the system."

https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias

When past decisions are swayed by prejudice

St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants (1980s).

Training data: Previous admission decisions

Result: Tho**If their names were non-Caucasian, the selection process was weighted against them. In fact, simply having a non-European name could automatically take 15 points off an applicant's score.**al minorities ar nts.

As editors at the British Medical Journal noted at the time, "[T]he program was not introducing new bias but merely reflecting that already in the system."

https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias

What Amazon's tool learned

School of Computing and Information Systems

In Week 3

Amazon's system taught itself that male candidates were preferable.

"Women's"

"Women's chess club captain"

Downgraded graduates of two all-women's colleges.

Masculine language

"Executed"

"Captured"

Risks of a non-representative data collection

The representativeness of records might correlate with class membership. In other words, certain classes of people might have systematically less accurate, timely, and complete records because they are:

- less involved in the formal economy and data-generating activities.
- relatively less fluent in the technology necessary to engage online.
- less profitable customers and thus less interesting as targets of observation.

Example: Street Bump (2012)

Street Bump, an application for Boston residents that takes advantage of accelerometers built into smartphones to detect when drivers ride over potholes. What is the potential risk of this app?

3. Feature selection

A process about choosing a set of attributes out of many for further analyses.

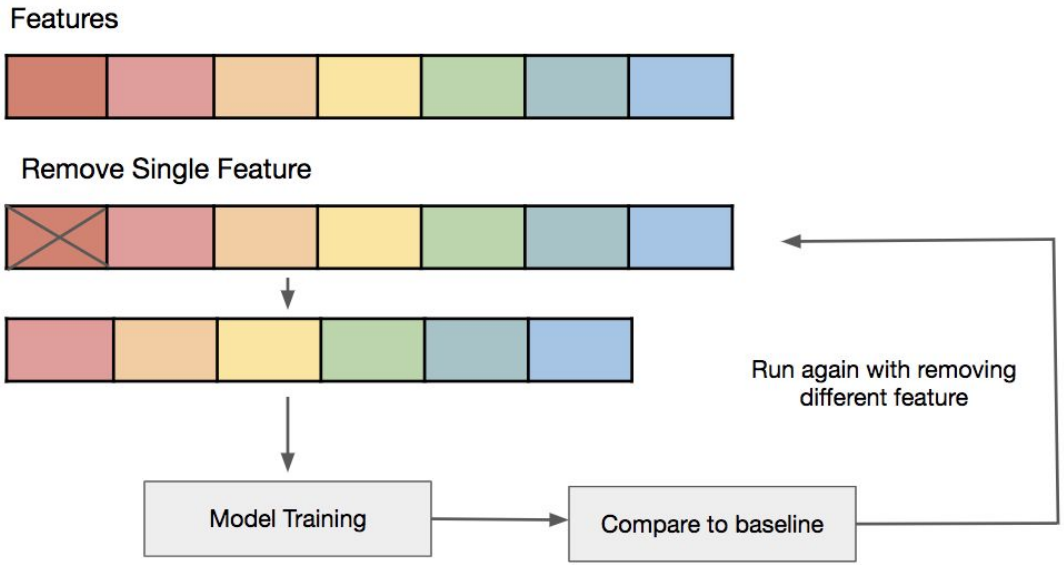


Image from https://www.kdnuggets.com/2019/10/feature-selection-beyond-feature-importance.html

Risks of feature selection

Selected features may fail to capture enough detail.

- Members of protected classes may find that they are subject to systematically less accurate classifications or predictions.

Unfair feature selection in history

Redlining

In the late 1930s, the Home Owners' Loan Corporation “graded” neighborhoods into four categories, based in large part on their racial makeup. Neighborhoods with minority occupants were marked in red — hence “redlining” — and considered high-risk for mortgage lenders.

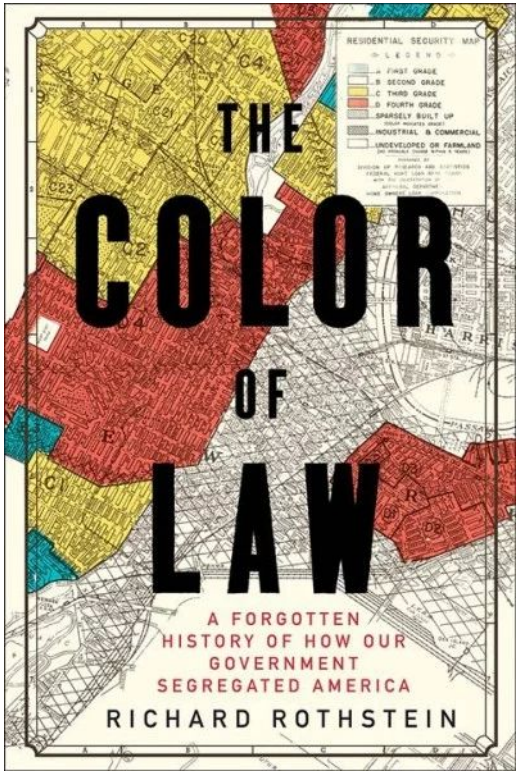
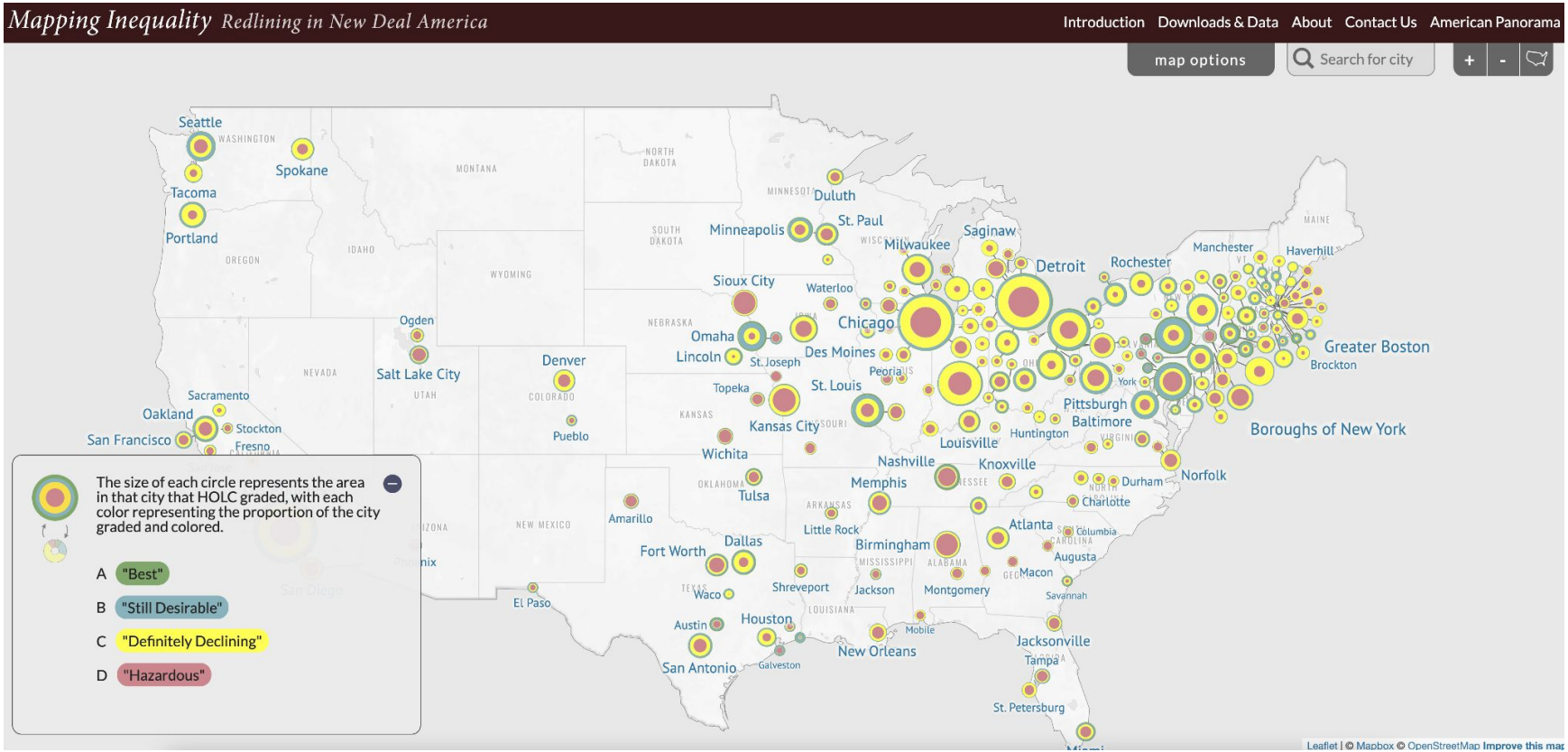


Image from https://www.npr.org/2017/05/03/526655831/a-forgotten-history-of-how-the-u-s-government-segregated-america

Redlining in USA



https://dsl.richmond.edu/panorama/redlining/

Redlining in USA

Boston, MA

D8 D9 D10

Show Full | Show Scan

5. Clarifying Remarks

Negro heavily concentrated north of Ruggles St. on the west side of Washington. Jewish centered near Columbus Square. A large territory with some streets showing better experiences than the balance of the section.

2. Inhabitants

e. Infiltration of foreign - negro

c. Foreign-born families 50%%; mixturepredominating

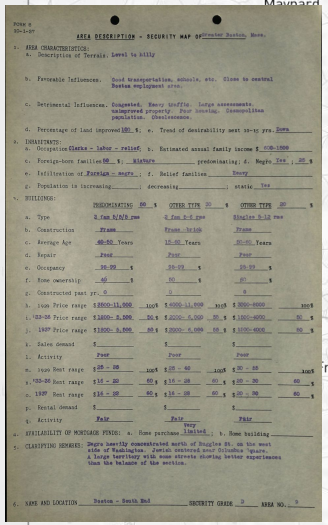
d. Negro yes%; 25%predominating

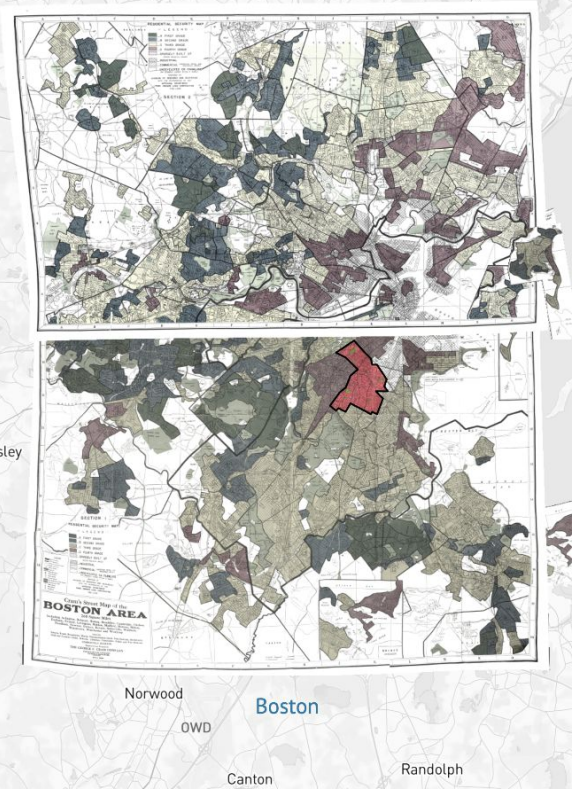
f. Relief families heavy

a. Occupation clerks - labor - relief

b. Estimated Annual Family Income $600-$1,500

1. Area Characteristics





https://dsl.richmond.edu/panorama/redlining/

E.g., Graduate school in hiring decisions

Hiring decisions tend to assign enormous weight to the college or university from which an applicant has graduated along with other features.

If equally competent members of protected classes happen to graduate from these colleges or universities at disproportionately low rates, decisions will systematically discount these individuals.

4. Proxies

Proxy variables (proxies) can be used when a certain characteristic is hard to measure directly.

Examples:

- Quality of life → Per-capita GDP as a proxy
- True body fat percentage → Body Mass Index (BMI) as a proxy
- Cognitive ability → GPA, School of graduation as a proxy

Potential biases in proxies

Membership in a protected class happens to be encoded in other data, called “redundant encodings.”

- Race and neighborhood in the U.S.
- Gender and time spending on housework
- And many more.

If redundant encodings exist, data mining can result in discriminatory models even when it only aims to ensure the high accuracy of the model.

https://www.bls.gov/charts/american-time-use/activity-by-sex.htm#

5. Masking

Decision makers with prejudicial views can “mask” their intentions by exploiting each of the mechanisms 1-4.

Data mining can **infer** (unseen) individuals’ membership in protected classes with high accuracy, and discount, penalize, or exclude such people accordingly.

Reflection

https://smu.sg/IS457r5