The Ethics of “Open” Data

Road Map

1. **Ethical** issues for a data set
2. **Epistemic** issues for a data set

   “epistemology” = the study of knowledge

   How do we form beliefs? *What* do we know?
   What are we ignorant of?

3. **What should this mean for our use of open source data?**

Open Data = publicly available data

Welcome to the UC Irvine Machine Learning Repository

We currently maintain 612 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!

VIEW DATASETS | CONTRIBUTE A DATASET

kaggle

Create

Home
Competitions
Datasets
Code
Discussions
Learn
More

Datasets

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.

+ New Dataset

Search datasets

All datasets | Computer Science | Education | Classification | Computer Vision | NLP | Data Visualization

Trending Datasets

French bakery daily sales
Matthieu Gimbert · Updated a day ago
Usability **9.1** · 2 MB
1 File (CSV)

World Carbon Pricing Dataset
Michael Bryant · Updated 14 days ago
Usability 7.6 · 11 MB

World Series 2022 Baseball - Phillies vs...
Matt OP · Updated 11 days ago
Usability **9.7** · 4 kB

72486: Iris
11942: Adult
85085: Dry Bean Dataset
40148: Wine
07176: Heart Disease
98799: Wine Quality
53694: Rice (Cammeo and Osmancik)
76179: Bank Marketing
43569: Breast Cancer Wisconsin (Diagnostic)
52442: Car Evaluation
39365: Raisin Dataset
1343502: Abalone

5,239 Datasets

COVID -19 Coronavirus Pandemic Dataset
Aman Chauhan · Updated a month ago
Usability **10.0** · 1 File (CSV) · 11 kB

Credit Card Customers Prediction
Aman Chauhan · Updated 8 days ago
Usability **10.0** · 1 File (CSV) · 388 kB

Students Performance in Exams
Aman Chauhan · Updated 2 months ago
Usability **10.0** · 1 File (CSV) · 9 kB

Data for Admission in the University
Akshay Dattatray Khare · Updated 11 days ago
Usability **10.0** · 1 File (CSV) · 4 kB

Udemy Courses
The Devastator · Updated 22 days ago
Usability **10.0** · 5 Files (CSV) · 439 kB

There is a lot!

Big data projects

the incentivisation of open data

Interface used by Amazon Turk Workers to label pictures in ImageNet

Consent

Pima Indians Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: From National Institute of Diabetes and Digestive and Kidney Diseases; Includes cost data (donated by Peter Turney)

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	1990-05-09
Associated Tasks:	Classification	**Mis**	Yes	**Number of Web Hits:**	96886

Figure 1.

Source:

Original Owners:

National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database:

Vincent Sigillito (vgs '@' aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

"a model organism for machine learning"

At the core of this history are questions about the origins, ownership, and reuse of the personal and bodily data that fuels information economies. The story of how Indigenous participants in the National Institutes of Health's longitudinal research on diabetes at Gila River became understood as donors of data used to study diabetes and later, how that data was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data

Pima Indians Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: From National Institute of Diabetes and Digestive and Kidney Diseases; Includes cost data (donated by Peter Turney)

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	1990-05-09
Associated Tasks:	Classification	**Mis**	Yes	**Number of Web Hits:**	96886

Figure 1.

Source:

Original Owners:

National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database:

Vincent Sigillito (vgs '@' aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

"a model organism for machine learning"

Consent

to *what*, exactly?

At the core of this history are questions about the origins, ownership, and reuse of the personal and bodily data that fuels information economies. The story of how Indigenous participants in the National Institutes of Health's longitudinal research on diabetes at Gila River became understood as donors of data used to study diabetes and later, how that data was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data

Ethical Issues

When data is about *people,* we should ask the following questions:

- “Did these people consent to being represented in this data set?”
- “What, exactly, did they consent to?”

Ethical Issues

When data is about *people,* we should ask the following questions:

- “Did these people consent to being represented in this data set?”
- “What, exactly, did they consent to?”

Consent = giving of genuine permission.

Needs to be *autonomous*, *informed*, and *specific*.

Ethical Issues

When data is about *people,* we should ask the following questions:

- “Did these people consent to being represented in this data set?”
- “What, exactly, did they consent to?”

Consent = giving of genuine permission.

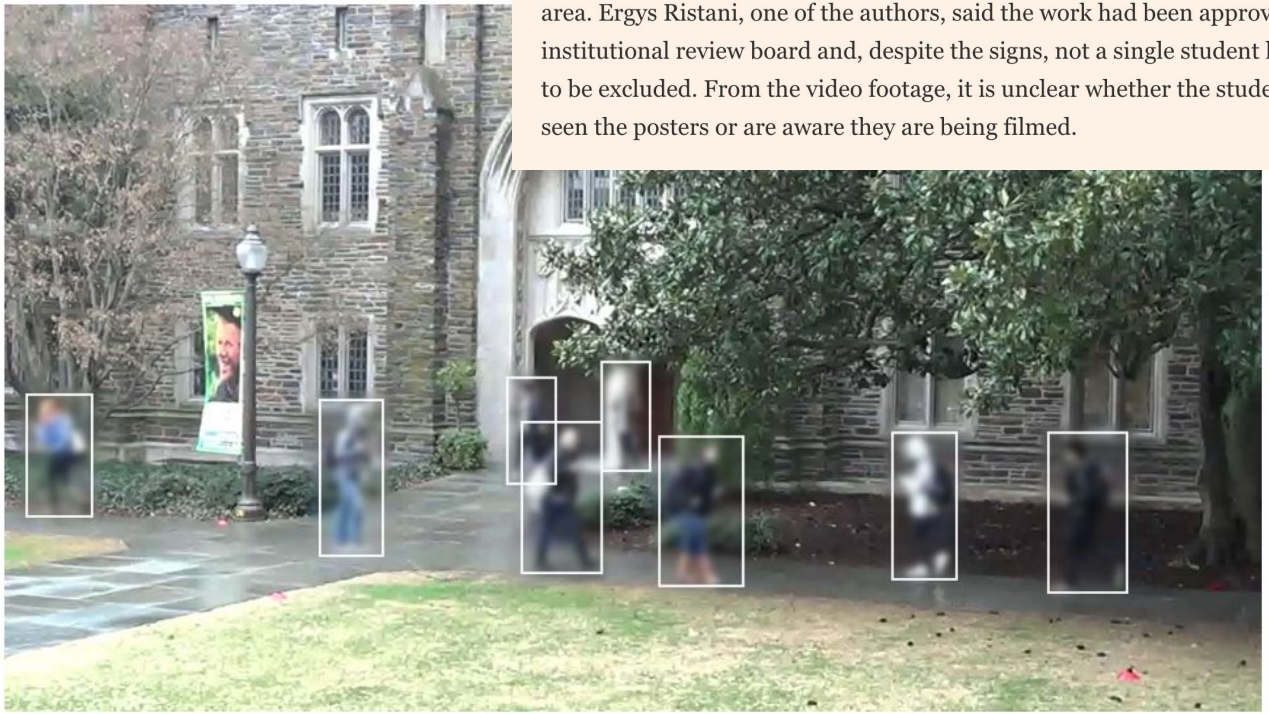
Needs to be *autonomous*, *informed*, and *specific*.

Seeking consent is about respecting the agency of others, and not just treating them as a means to the production of knowledge

Was there consent here?

If so, to *what*?

Researchers used eight cameras to film students walking on campus, and notified them through posters placed around the perimeter of the surveilled area. Ergys Ristani, one of the authors, said the work had been approved by an institutional review board and, despite the signs, not a single student had asked to be excluded. From the video footage, it is unclear whether the students have seen the posters or are aware they are being filmed.



A still frame from the Duke MTMC (Multi-Target-Multi-Camera) CCTV dataset captured on Duke University campus in 2014. The dataset has now been terminated by the author in response to this report.

DUKE MTMC

Can you consent to your data being *open*?

DUKE MTMC

Duke MTMC (Multi-Target, Multi-Camera) is a dataset of surveillance video footage taken on Duke University's campus in 2014 and is used for research and development of video tracking systems, person re-identification, and low-resolution facial recognition.

The dataset contains over 14 hours of synchronized surveillance video from 8 cameras at 1080p and 60 FPS, with over 2 million frames of 2,000 students walking to and from classes. The 8 surveillance cameras deployed on campus were specifically setup to capture students "during periods between lectures, when pedestrian traffic is heavy". [1]

IMAGES
2,000,000

IDENTITIES
2,700

PURPOSE
Person re-identification, multi-camera tracking

YEAR
2016

WEBSITE
duke.edu

For this analysis of the Duke MTMC dataset over 100 publicly available research papers that used the dataset were analyzed to find out who's using the dataset and where it's being used. The results show that the Duke MTMC dataset has spread far beyond its origins and intentions in academic research projects at Duke University. Since its publication in 2016, more than twice as many research citations originated in China as in the United States. Among these citations were papers links to the Chinese military and several of the companies known to provide Chinese authorities with the oppressive surveillance technology used to monitor millions of Uighur Muslims.

Retractions

In response to the backlash, the author of DukeMTMC **issued an apology** and took down the dataset. It is one of several datasets that has been removed or modified due to ethical concerns. But the story doesn't end here. In the case of DukeMTMC, the data had already been copied over into other derived datasets, which use data from the original with some modifications. These include **DukeMTMC-SI-Tracklet**, **DukeMTMC-VideoReID**, and **DukeMTMC-ReID**. Although some of these derived datasets were also taken down, others, like DukeMTMC-ReID, remain freely available.

- Dataset retraction has a limited effect on mitigating harms (Section 3). Our analysis shows that even after DukeMTMC and MS-Celeb-1M were retracted, their underlying data remained widely available and continued to be used in research papers. Because of such "runaway data," retractions are unlikely to cut off data access; moreover, without a clear indication of the underlying intention, retractions may have limited normative influence.

Epistemic Issues

Data sets render some things visible, and others invisible.

Sometimes, *invisibility* carries risks

Trans and nonbinary have been rendered invisible in HIV prevention data

BACKGROUND

The stakes are high: Efforts to bring an end to the global HIV pandemic will fail if effective, accessible HIV prevention interventions do not reach TGD people. Yet our people—who often face crushing stigma, marginalization, criminalization and violence, along with a disproportionate burden of HIV—continue to be overlooked, subsumed, underrepresented and/or excluded altogether from HIV prevention research. Though there has been a recent uptick of TGD-inclusive scientific literature and trans-led research in the field, it's far from adequate. It's important to rectify this troubling omission. As Jerome Singh, an ethicist from the University of Toronto and the University of KwaZulu-Natal noted in 2016:

"To date, there have been no HIV-endpoint trials that specifically focus on transgender individuals . . . Given their disproportionate burden of HIV, their historic and ongoing marginalization, and the knowledge gap related to HIV prevention specific to the transgender community, conducting focused HIV prevention research on transgender persons is an ethical imperative."[1]

The construction of categories

How many evictions were there in San Francisco in 2014?

EVICTION LAB

1,440

Eviction Lab is a team based out of Princeton University which gathers and publishes nationwide eviction data. The team is led by Matthew Desmond, author of the Pulitzer Prize winning book, *Evicted*. Eviction Lab gathers formal eviction records from 48 states and DC and combines those records with deomgraphic information from the Census. Their datasets also included state-reported, county-level statistics from landlord-tenant cases in 27 states. They collected this information via bulk reports of cases from courts, record collection from online portals and purchasing datasets of public eviction records from Lexis Nexis Risk Solutions and American Information Research Services. Their data is cleaned so that each observation represents a household, and they use imputation* to fill in missing data.

Anti-Eviction Mapping Project

3,310

The Anti-Eviction Mapping Project (AEMP) engages in community-led data collection. Their data is the result of grass-roots organizing and local knowledge. Their datasets include records from county-court cases, city rent boards, survery data from eviction clinics and narrative and qualatitive interviews with tenants. Groups like AEMP argue their data captures evictions which do not appear in court records. For example, evictions might look different in gentrifying cities where month-to-month leases might be ended with no recourse, rent increases price out tenants (economic evictions). Additionally in some states, (such as Oregon), no cause evictions do not require any formal court filing and as a result, will not be captured in court record-based datasets.

**imputation is a method of replacing missing data with estimates based on other available data. For example, with Eviction Lab, "when only one year of data was missing within a county between two years of valid data, the case volume was imputed using the average of the preceding and following years."*

What goals and assumptions are baked into our understanding of categories?

The construction of categories

How many evictions were there in San Francisco in 2014?

EVICTION LAB

1,440

Eviction Lab is a team based out of Princeton University which gathers and publishes nationwide eviction data. The team is led by Matthew Desmond, author of the Pulitzer Prize winning book, *Evicted*. Eviction Lab gathers formal eviction records from 48 states and DC and combines those records with deomgraphic information from the Census. Their datasets also included state-reported, county-level statistics from landlord-tenant cases in 27 states. They collected this information via bulk reports of cases from courts, record collection from online portals and purchasing datasets of public eviction records from Lexis Nexis Risk Solutions and American Information Research Services. Their data is cleaned so that each observation represents a household, and they use imputation* to fill in missing data.

Anti-Eviction Mapping Project

3,310

The Anti-Eviction Mapping Project (AEMP) engages in community-led data collection. Their data is the result of grass-roots organizing and local knowledge. Their datasets include records from county-court cases, city rent boards, survery data from eviction clinics and narrative and qualatitive interviews with tenants. Groups like AEMP argue their data captures evictions which do not appear in court records. For example, evictions might look different in gentrifying cities where month-to-month leases m ended with no recourse, rent increases price out (economic evictions). Additionally in some states, Oregon), no cause evictions do not require any fc filing and as a result, will not be captured in court record-based datasets.

*imputation is a method of replacing missing data with estimates based on other available data. For example, with Eviction Lab, "when only one year o data was missing within a county between two years of valid data, the case volume was imputed using the average of the preceding and following yec

What goals and assumptions are baked into our understanding of categories?

It's not always easy or possible, to answer that question

Browse Through: **622** Da

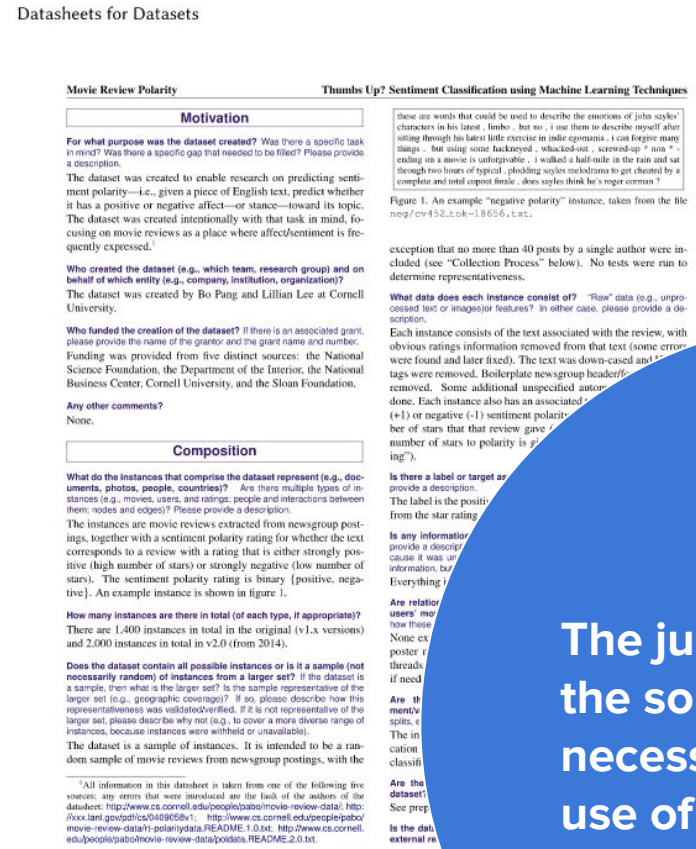
Default Task
Classification (466)
Regression (151)
Clustering (121)
Other (56)
Attribute Type
Categorical (38)
Numerical (422)
Mixed (55)
Data Type
Multivariate (480)
Univariate (30)
Sequential (59)
Time-Series (126)
Text (69)
Domain-Theory (23)
Other (21)
Area
Life Sciences (147)
Physical Sciences (57)
CS / Engineering (234)
Social Sciences (41)

Name
Abalon
Adult
Annealing
Anonymous Microsoft Web Data
Arrhythmia
Aa

So what?

Data histories in computer science

Professor Timnit Gebru



Datasheets for Datasets 15

Movie Review Polarity Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the

[1]All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file neg/cv452_tok-18656.txt.

exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and tags were removed. Boilerplate newsgroup header/f removed. Some additional unspecified autom done. Each instance also has an associated (+1) or negative (-1) sentiment polarit ber of stars that that review gave number of stars to polarity is g ing").

Is there a label or target as provide a description.

The label is the positi from the star rating

Is any informatio provide a descrip cause it was un information, bu Everything

Are relatio users' mo how these None ex poster threads if need

Are th ment/v splits, The in cation classifi

Are the dataset See prep

Is the dat external re

The jumping off point for the sort of critical inquiry necessary for the ethical use of open data

In Your Groups

Data Sheets for Data Strangers DATASET:

THE BASICS

WHO	WHAT
Who is this data about?	What is the data about?
Who collected this data?	Which headers or data are ambiguous?
Who is this dataset for? Who is it intended to benefit or serve?	Are there any headers for which an explicit definition is provided?

WHEN What timespan is represented in the data? | When was the data collected?

WHERE Where is the data from? *(i.e. does it pertain to residents of a specific area? a particular geographic location? etc.)*

collected? What is its intended use?

?

last class

Data Sheets for Data Strangers DATASET:

CONSENT

Did the data subjects consent to the data being collected initially? | Did the data subj re-used?

What specifically did they consent to? | Under what conc

DIGGING INTO DEFINITIONS

What headers in the data might need explicit definitions? List them below. | Are there multiple ways to What are examples?

Do we know how the data collectors have defined these headers?

Will some of these definitions include/exclude different people?

Do any of the headers in the data represent the subjects in a way that

RIPPLE EFFECTS

How might the dataset reflect the assumptions, motivations, and intere

BENEFITS & HARMS

Is the data about people who might be especially vulnerable? ☐ Yes ☐ No ☐ Not sure

If yes:
Could simply being represented in data put them at increased risk of harm? | Could the data be *used* in a way that might harm them?

Who is **missing** from the data? Is anyone harmed by not being included?

Can the data be put to use in a way that might **benefit the data subjects**? | Can the data be put to use in a way that might **benefit *other* people**?

Can the data be put to use in a way that might **harm the data subjects**? | Can the data be put to **harm *other* people**?

in class