# Introduction

This document describes the features of the compiled and cleaned version of the [Crowd Counting Consortium](https://ash.harvard.edu/programs/crowd-counting-consortium/) (CCC) data on U.S. protest events from the project’s second phase, which ran from January 2021 through December 2024.

In late 2020, Harvard Kennedy School’s [Nonviolent Action Lab](https://ash.harvard.edu/programs/nonviolent-action-lab/), one of the consortium’s two partner organizations, hired a full-time research program manager, most of whose time was devoted to work on CCC. This hire allowed the project to broaden the scope of its collection efforts, to increase the amount of information it collected on each protest event, and to supervise more closely the work of paid undergraduate and graduate research assistants (RAs) who assisted with the data collection. As part of that expansion, the project also developed a more extensive and explicit set of coding guidelines, which can be found [here](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?usp=sharing). In late 2023, the Lab hired a second full-time staff member who also devotes the bulk of their time to this project, including but not limited to training and reviewing the work of the project’s rotating team of RAs.

During this four-year phase, the project made some changes to the collection process that should be kept in mind when summarizing and analyzing the data.

* *Expanded and continually evolving social media tracking*. During CCC’s second phase, the project significantly expanded its use of social media to collect information about U.S. protests, and it continually adapted its collection streams to try to keep up with rapid and significant changes in the social-media landscape. The chief innovation during this time was the creation and routine use of a Twitter list comprising organizers, activist researchers, and independent journalists who regularly covered protest activity. As organizers fled Twitter/X in 2023 and 2024, we gradually but significantly expanded our routine collection from Instagram with snowballing additions of accounts to follow. In 2022, the project also began researching protest activity via Telegram, mostly focusing there on actions by right-wing and far-right groups that were not on Twitter or Instagram at the time.
* *Multiple records for multi-day events*. Throughout the project’s first phase, events that spanned multiple days (e.g., an extended occupation of a city park or a campus building) were recorded in a single record, with the event’s start date as the date. In Phase 2, we began creating separate records for each day of these events. This allows us to capture more complete information about variation in the event over time (e.g., daily information about how many people were present or what police did), and we think it more accurately represents the mobilizational energy involved.
* *Verbatim claim text collection*. In 2021, the project began experimenting with capturing verbatim the text of protesters’ signs, banners, and chants in an effort to represent more accurately and completely what each protest was about. In spring 2022, this expansion became routine and was added to the coding guidelines. The text of this claims rhetoric was captured in the existing ‘claims’ field alongside the coder’s own summary of participants’ demands, but coders were instructed to use different syntax for phrases associated with those two sources (coder vs. verbatim) so they could be distinguished from each other for purposes of analysis.
* *Cessation of labor strike coverage*. Throughout CCC’s first phase (2017-2020) and into its second (2021-2024), the project included labor strikes in the set of events it sought to track. In April 2023, however, CCC stopped doing that. Labor strikes were always a bit of a conceptual stretch of CCC’s idea of political crowds; the project had limited resources; and, sincelate 2020, the [ILR Labor Action Tracker](https://striketracker.ilr.cornell.edu/) project at Cornell University has been doing an excellent job making structured data on U.S. labor strikes. So, in early 2023, we decided to stop duplicating that project’s work. CCC continues to collect on rallies, protests, and similar actions in support of striking workers, but not on the strikes themselves. In light of this change, researchers doing longitudinal comparisons of gross event counts and the like may want to filter out records with “strike” in the ‘type’ field. Note, though, that some waves of student-led demonstrations in the late 2010s calling for action on climate change were assigned “strike” as the event type at the time. So, if you are attempting gross comparisons that include that period, you’ll probably want to include a carveout in your filtering rules that excludes events with a valence of 1 and “environment” as an issue tag.

The compiled version of the data includes numerous features that have been added algorithmically after the monthly tables have been merged. If you want to see or work with the data in its raw form, please use the original Google Sheets, which you can access from the [CCC website](https://ash.harvard.edu/programs/crowd-counting-consortium/).

# Variable Descriptions

## date

Date of event in YYYY-MM-DD format.

* When an event spans multiple days, each of those days is captured in a separate record.

## locality

Name of the locality (i.e., city or town) in which the event took place.

* When the event location is only given as a county, the county seat is recorded as the locality.
* When an event is held online, it is associated with a specific locality if organizers give one. If it is described as statewide, it is associated with the state capital. If it is national or international, its location is recorded as Washington, DC. In all of these cases, the ‘location\_detail’ (see below) is recorded as “online” and the descriptor “virtual” is used in ‘event\_type’ (e.g., “virtual rally”).

## state

Two-letter U.S. postal abbreviation for the state or U.S. territory in which the event took place.

## location\_detail

Where available, text giving additional details on the location(s) within the city or town where the action took place, usually in the form of a street address, an intersection (e.g., “9th St and Broad St”), a landmark (e.g., “Akron City Hall”, “Colorado State Capitol”), or the name of a school or institution (e.g., “University of North Carolina at Chapel Hill”).

## online

Binary indicator for online-only events. 1 = online, 0 = in-person. Generated from location and event type information in the source data.

## type

Type(s) of protest action (e.g. march, protest, demonstration, strike, counter-protest, sit-in), separated with semicolons or commas when more than one. See [Coding Guidelines](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?tab=t.0#heading=h.86aoh6qprspd) for more information on the array of labels used.

## title

Title of action when one is given, usually on a flyer announcing the event.

## macroevent

A unique id that associates a counter-protest with the protest event it countered. These strings are composed of a date, a location, and something about the nature of the event, all separated by hyphens (e.g., "20220624-phoenix-abortion"). In most cases, these ids will uniquely identify pairs of events. In cases where the counter-protest is itself countered, however—e.g., a community defense action in response to a protest targeting an LGBTQ+ pride festival—the additional events are given the same id, so these clusters will sometimes include three or more events.

## organizations

Semicolon-separated names of organization(s) that participated in the event, including but not limited to organizers and endorsers (.e.g, “Fridays for Future New York; Extinction Rebellion New York; Greenfaith”).

## participants

Semicolon-separated text descriptors of participants in the event, usually from press reports, sometimes recorded by coders based on photos or video (e.g., “students; faculty; staff; community members”).

## claims

Comma-separated text phrases describing what the event was about. As noted in the Introduction, in 2022, we began capturing verbatim the text of protesters’ claims from signs, banners, and chants seen or heard at the event. See the [Coding Guidelines](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?tab=t.0) for details on how these were handled. The key point here is that all coder-generated summary phrases begin with “for”, “against”, or “in [action word] [of/for/with]”, so researchers particularly interested in the verbatim rhetoric can use a regular expression to identify and remove the coder-summary phrases from a vectorized version of the comma-separated claims phrases.

## claims\_summary

Text phrase or comma-separated text phrases giving coder’s summary of the action’s main claims or demands. Each phrase should begin with either “for”, “against”, or “in [action word] [of/with/for]” (e.g., “in commemoration of…”, “in solidarity with…”, “in remembrance of…”). Generated programmatically from *claims* field, so there will be occasional errors of omission and commission due to typos and such.

## claims\_verbatim

Comma-separated verbatim text captures of claims made by event participants on signs, banners, t-shirts, and flags or as chants or shouts. Generated programmatically from *claims* field, so there will be occasional errors of omission and commission due to typos and such.

A few notes on the syntax within these phrases.

* Phrases that end with an exclamation point (e.g., “no justice no peace!”) are chants or shouts.
* Text inside square brackets (e.g., “[wire hanger ]”, “[watermelon]”) is a coder’s description of an image or symbol seen on a sign or banner.
* Flags are recorded as phrases ending in “flag” (e.g., “Confederate flag”, Gadsden flag”). National flags other than the U.S. flag are recorded as the name of the country followed by “flag” (e.g., “Israel flag”, “Mexico flag”). The American flag is usually recorded as “American flag”, sometimes as “U.S. flag”, “US flag”, or “USA flag”.
* When the same claim is seen more than once, only one instance of it is recorded here.

## issue\_tags\_summary

String of semicolon-separated tags identifying political issues (or themes) associated with the event, based only on the coder-generated summary descriptions of participants’ demands in *claims\_summary*. These are generated after data compilation by running a series of regular expressions over the claims description text. The list of issue tags and the regular expressions used to generate them can be found [here](https://github.com/nonviolent-action-lab/crowd-counting-consortium/blob/master/data_compilation/ccc_issue_regex_list.R).

## issues\_tags\_verbatim

String of semicolon-separated tags identifying political issues (or themes) associated with the event, based only on the verbatim text captures of protester claims or demands found in *claims\_verbatim*.The list of issue tags and the regular expressions used to generate them can be found [here](https://github.com/nonviolent-action-lab/crowd-counting-consortium/blob/master/data_compilation/ccc_issue_regex_list.R).

## issue\_tags

String of semicolon-separated tags identifying political issues (or themes) associated with the event, as generated from either the coder summary or verbatim capture—i.e., the concatenation of *issue\_tags\_summary* and *issue\_tags\_verbatim*. The list of issue tags and the regular expressions used to generate them can be found [here](https://github.com/nonviolent-action-lab/crowd-counting-consortium/blob/master/data_compilation/ccc_issue_regex_list.R).

## valence

Political valence of the event. See the [Coding Guidelines](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?tab=t.0) for details on how these judgments were made.

2 = right-wing

1 = left-wing

0 = other or neither

## size\_text

Words or phrases that journalists or eyewitnesses used to describe the size of the crowd at the protest event (e.g., “more than 100”, “dozens”, “about 50”)—or, in cases where no text descriptions were found but other sources gave information about crowd size, the alternative source used: “count pic” for photograph(s), “count vid” for video(s), “count FB” for Facebook, or “eyewitness” for eyewitness.

## size\_low

Lowest crowd size reported in, or estimated from, *size\_text*.

To convert vague text counts to numbers, the following three rules are always used for both *size\_low* and *size\_high* (see below).

1. For phrases with fudge words (e.g., “about”, “nearly” “approximately”, “maybe”), ignore the fudge word and treat the number that follows it as the count.
2. When unspecific multiples are given, assume the multiple is 2 (e.g., “hundreds” becomes 200, “thousands” becomes 2,000).
3. Assume “several” means 3 (e.g., “several dozen” becomes 36, “several thousand” becomes 3,000).

## size\_high

Highest crowd size reported in, or estimated from, *size\_text*.

When vague text counts are converted to numbers, the high and low estimates are assumed to be equivalent. For example, if the only information available about crowd size is a single report describing it as “hundreds”, 200 would be assigned to both *size\_low* and *size\_high*. If, however, one source described the crowd size as “hundreds” and another as “about 500”, then *size\_low* would still be 200, but *size\_high* would be 500.

## size\_mean

The mathematical average of *size\_low* and *size\_high*, rounded up to the nearest integer.

## size\_cat

Ordered categorical indicator of crowd size, representing orders of magnitude and derived from *size\_mean*.

0 = unknown

1 = 1-99 (tens)

2 = 100-999 (hundreds)

3 = 1,000-9,999 (thousands)

4 = 10,000+ (tens of thousands)

## arrests

Text, sometimes specifying the count of protesters reportedly arrested (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”).

* In cases where multiple numbers are reported, we use the latest one given by police (as seen in our press or social-media sources).
* If police do not report a number, we use the latest one given by organizers.
* If neither police nor organizers report a count, we use the highest count reported.

## arrests\_any

Binary indicator for whether or not any arrests occurred, derived from *arrests*. 1 = yes, 0 = no.

## injuries\_crowd

Text, sometimes giving a count of protesters reportedly injured (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”). This includes protesters who suffer accidents or illness (e.g., heat exhaustion, heart attack) during a protest event, as well as ones injured by police, counter-protesters, or other protesters.

## injuries\_crowd\_any

Binary indicator for whether or not any protesters were reportedly injured, derived from *injuries\_crowd*. 1 = yes, 0 = no.

## injuries\_police

Text, sometimes giving a count of police officers reportedly injured (e.g., “5”), sometimes a phrase indicating ambiguity about that count (e.g., "more than 5", "unclear", “unspecified”). This includes police officers who suffer accidents or illness (e.g., heat exhaustion, heart attack) during a protest event, as well as ones injured by protesters, counter-protesters, or other police.

## injuries\_police\_any

Binary indicator for whether or not any police officers were reportedly injured, derived from *injuries\_police*. 1 = yes, 0 = no.

## property\_damage

Text briefly describing the nature of any property damage caused by protesters (e.g., “graffiti”, “toppled statue”)..

## property\_damage\_any

Binary indicator for whether or not protesters reportedly caused any property damage, derived from *property\_damage*. 1 = yes, 0 = no.

## chemical\_agents

Binary indicator for whether police or other state security forces used tear gas or other chemical irritants, such as pepper spray or pepper balls, during the protest event. 1 = yes, 0 = no.

## participant\_measures

Text field describing notable actions taken by protest participants in the course of the event, including but not limited to interactions with police and/or counter-protesters. See the [Coding Guidelines](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?tab=t.0#heading=h.cdhngnt893ka) for details on what is recorded and how.

## police\_measures

Text field describing presence of, and notable actions taken by, police in response to the protest event, including interactions with protesters and/or counter-protesters. See the [Coding Guidelines](https://docs.google.com/document/d/1oaOf9s72FQnzQA8sbE8h0PwMIZLP6p0EDUV2ya065is/edit?tab=t.0#heading=h.dx26jlm0cpmj) for details on what is recorded and how.

## participant\_deaths

Reported count of protest participants who died as a result of their participation in the protest. This includes deaths by natural causes during a protest event as well as deaths reportedly caused by the actions of police or counter-protesters.

## police\_deaths

Reported count of police officers who died while policing the protest. This includes deaths by natural causes during a protest event as well as deaths reportedly caused by the actions of protesters or counter-protesters.

## source\_n

URL of nth source, or description where the source is not a web page (e.g., “eyewitness”, “correspondence”).

## notes

Any additional information about the event that the coder chose to capture.

## lat

Latitude of locality in which the event took place, as resolved by Google Maps Geocoding API. Note that this is not based on address or landmark-level information where that is given, only on the name of the city or town.

## lon

Longitude of locality in which the event took place, as resolved by Google Maps Geocoding API. Note that this is not based on address or landmark-level information where that is given, only on the name of the city or town.

## resolved\_locality

Name of the locality in which the event occurred, as resolved by running the city or town name and state abbreviation through the Google Maps Geocoding API.

## resolved\_state

Postal abbreviation of the state or territory in which the event occurred, as resolved by the Google Maps Geocoding API.

## resolved\_county

Name of the county in which the event occurred, as resolved by running *resolved\_locality* and *resolved\_state* through the Google Maps Geocoding API..

## fips\_code

Five-digit FIPS code for the county (or LA parish or AK borough or independent city or DC or U.S. territory) given in *resolved\_county*. See 'data-compilation/fips\_for\_county\_function.r' for details on how these are generated using the 'tigris' package and some custom code to handle various exceptions.

NOTE: When you load the data from the stored .csv, you will probably need to add leading zeros back to FIPS codes that have them, because your software will probably read that column as integers instead of strings. In R, you could do this with `ifelse(nchar(fips\_code) == 4, paste0("0", fips\_code), fips\_code)`.