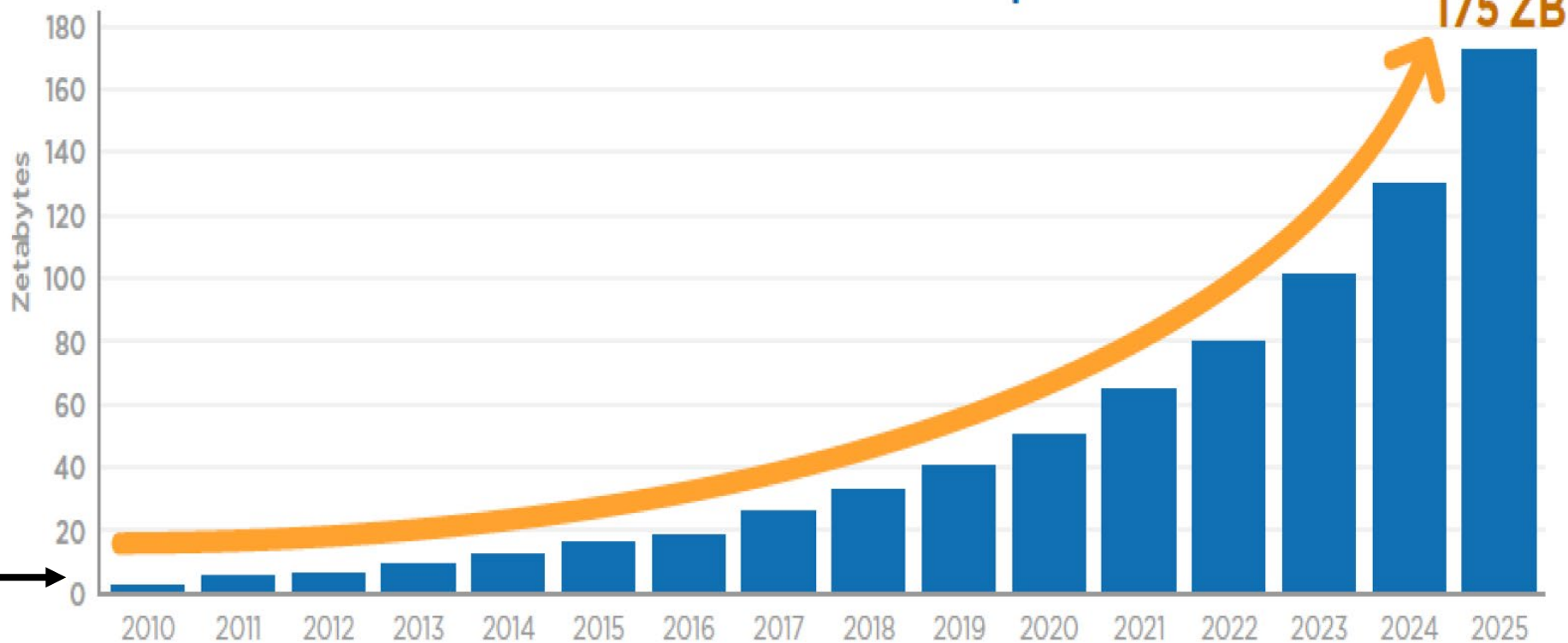


What are some types of data that you can think of?

Annual Size of the Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Sequencing every individual's genome in the world is only 1.6 zettabytes of data.

WHAT'S A ZETTABYTE?

1 kilobyte	1,000,000,000,000,000,000
1 megabyte	1,000,000,000,000,000,000,000
1 gigabyte	1,000,000,000,000,000,000,000,000
1 terabyte	1,000,000,000,000,000,000,000,000,000
1 petabyte	1,000,000,000,000,000,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000,000,000,000,000,000

“It Took Longer than I was Expecting:” Why is Dataset Search Still so Hard?

Madelon Hulsebos, Wenjing Lin, Shreya Shankar, Aditya G. Parameswaran
UC Berkeley

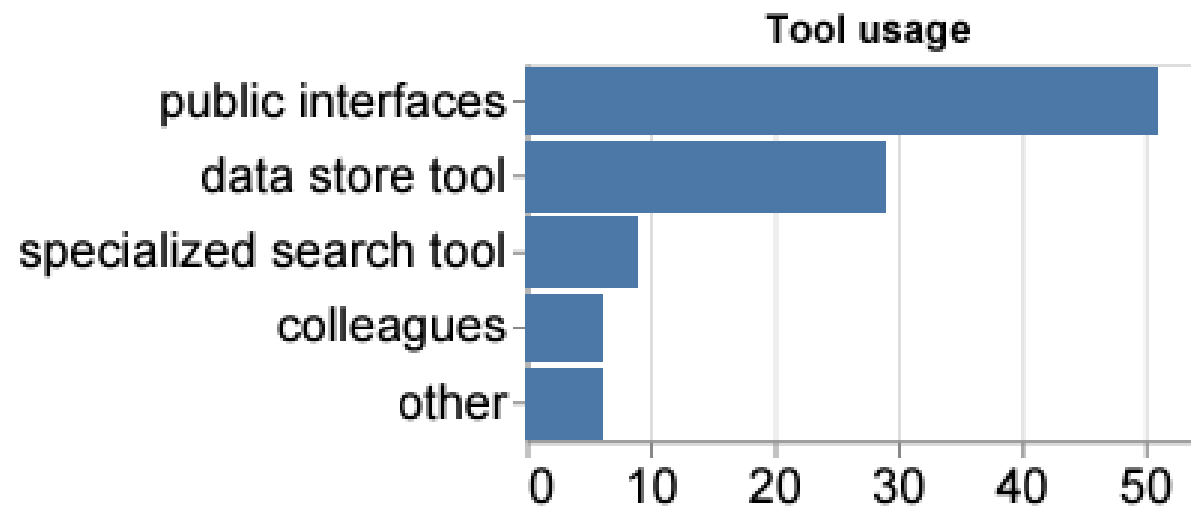


Figure 2: Tools used in practice to search for datasets.

Data Search

Data can come in a CSV (comma separated value) file format which can easily be read into a table.

- Each line is a row.
- Each column value is separated by a comma (,) but can also be separated by other values such as semi-colons (;) or tabs.

```
"Aruba","ABW","Agricultural methane emissions (% of
total)","EN.ATM.METH.AG.ZS","","","","","","","7.814
21030039674","7.66006558122809","7.51350582031861","7.3763881979
1023","7.25186175131417","7.1368728744082","6.95855804144777","6
.86024705076641","6.54509894939719","6.41305105257498","6.317023
51206149","6.18933535805329","6.07430067722162","5.9923838534923
4","5.86655468078772","5.73265496772068","5.78249982141771","5.5
9873522218585","5.50322899312953","5.37460693156412","","",""
,"","","","","","","","","","","","","","",""
,"","","","","","","","","","","","","",""
,"Africa Eastern and Southern","AFE","Agricultural methane
emissions (% of
total)","EN.ATM.METH.AG.ZS","","","","","","","70.019266
1831853","69.0702826025236","68.7439888324679","67.5572007298257
","66.9885636419988","67.0599176710941","66.632527316319","66.38
4898493342","67.3248356979876","67.0615813442033","66.6847847898
879","66.1468020706628","65.685686728042","64.8881810436463","64
.0554819352707","64.045986050421","62.4652146101028","63.0104454
439688","62.6985710972566","","","","",""
,"",""
```

Country N	Country C	Indicator I	Indicator	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Aruba	ABW	Agricultur	EN.ATM.M	7.81421	7.660066	7.513506	7.376388	7.251862	7.136873	6.958558	6.860247	6.545099	6.413051	6.317024	6.189335	6.074301
Africa Eas	AFE	Agricultur	EN.ATM.M	77.07323	76.64942	72.75751	74.01685	73.86948	74.79648	75.49685	74.42067	74.36701	74.11124	71.80097	71.74905	71.41616
Africa Wes	AFW	Agricultur	EN.ATM.M	71.90688	55.19772	57.09618	58.70903	53.72682	59.87467	57.20339	59.41951	60.58952	68.30273	62.74781	58.75897	62.46237
Angola	AGO	Agricultur	EN.ATM.M	55.97464	55.21655	53.55665	53.55947	54.30299	53.70927	54.66718	54.99007	55.27375	55.78043	56.41108	57.25735	57.4976
Albania	ALB	Agricultur	EN.ATM.M													
Andorra	AND	Agricultur	EN.ATM.M													
Arab Worl	ARB	Agricultur	EN.ATM.M													
United Arc	ARE	Agricultur	EN.ATM.M	1.219889	1.016195	0.928165	0.800697	0.726258	0.609301	0.3261	0.422949	0.450714	0.654558	0.746094	0.982156	1.157703
Argentina	ARG	Agricultur	EN.ATM.M	82.53948	84.25903	84.23633	85.35771	84.40681	84.32267	83.617	84.96823	83.1144	82.31711	81.55548	81.76058	80.72015
Armenia	ARM	Agricultur	EN.ATM.M	73.61899	73.45842	72.75043	71.88064	71.35427	71.17247	71.0905	71.7035	73.42678	74.97868	76.42723	77.26777	77.84686
American ASM	ASM	Agricultur	EN.ATM.M	47.38844	46.39361	45.63697	45.07248	44.50631	43.98328	43.54928	43.73352	43.071	42.31432	41.90651	41.8619	42.31321
Antigua ar	ATG	Agricultur	EN.ATM.M	51.90761	51.25905	51.40965	49.98851	50.33253	50.23436	51.17686	49.3468	55.80258	61.52256	64.127	64.35279	63.5808
Australia	AUS	Agricultur	EN.ATM.M	75.35098	75.96743	73.77147	72.79222	73.64035	76.08441	75.74539	73.61694	72.50029	71.37537	70.36638	67.58284	
Austria	AUT	Agricultur	EN.ATM.M	52.23143	53.61234	53.39494	52.90146	53.50957	53.49092	52.8418	52.76808	53.06089	53.12509	52.24483	51.34825	51.45589
Azerbaijan	AZE	Agricultur	EN.ATM.M	55.1172	55.59897	54.99308	53.80368	52.70321	51.68871	50.00396	48.51722	47.5224	46.35736	45.28903	44.03127	42.91236
Burundi	BDI	Agricultur	EN.ATM.M	41.14317	41.47436	42.44487	43.61734	43.67514	44.03798	43.71585	43.73005	43.09567	43.0837	37.12237	31.68165	31.00226
Belgium	BEL	Agricultur	EN.ATM.M	37.51703	39.1916	39.02119	39.5318	40.82071	41.33362	40.84205	40.83939	41.87618	42.42111	42.65138	43.02441	42.96421
Benin	BEN	Agricultur	EN.ATM.M	33.80207	34.17494	35.07825	35.73468	36.46271	37.11601	37.08866	36.27657	37.48401	37.09429	36.68196	38.04545	37.31701
Burkina F	BFA	Agricultur	EN.ATM.M	69.18267	68.51898	66.70292	65.45685	63.47759	64.12595	64.95724	65.89111	65.07357	66.35331	67.89251	66.91665	68.23264
Banglade	BGD	Agricultur	EN.ATM.M	89.34079	88.55977	88.43867	88.15226	87.8425	87.92103	87.2053	86.8659	86.74291	86.44222	85.19268	84.90584	84.40699
Bulgaria	BGR	Agricultur	EN.ATM.M	47.34679	47.68176	48.36183	47.54171	47.14222	47.21256	47.82142	46.97901	46.47655	45.84074	45.24025	44.06048	43.4713
Bahrain	BHR	Agricultur	EN.ATM.M	0.797745	0.823846	0.871168	0.798778	0.640858	0.684589	0.66419	0.622812	0.620705	0.652432	0.640934	0.590829	0.660206
Bahamas	BHS	Agricultur	EN.ATM.M	9.633619	8.862083	8.290197	7.558549	5.946782	5.533625	5.677441	4.794791	4.675594	4.486922	4.207793	3.988946	3.909462
Bosnia an	BIH	Agricultur	EN.ATM.M	54.04934	54.38432	53.63505	54.34635	53.70388	57.3106	56.69623	55.54067	53.96426	52.41722	49.17583	46.85653	42.85434
Belarus	BLR	Agricultur	EN.ATM.M	84.81345	84.99729	84.82806	84.51762	84.32675	84.14954	83.58491	83.10674	82.84307	82.71382	82.33057	81.94153	81.65416
Belize	BLZ	Agricultur	EN.ATM.M	63.32884	67.1709	67.11168	66.99096	68.78743	69.63229	70.08282	67.97482	66.23754	69.56481	68.08278	66.97048	69.45938
Bermuda	BMU	Agricultur	EN.ATM.M	9.080264	9.23123	8.337609	7.269248	5.863826	5.834402	6.176004	5.855913	6.248207	6.904611	7.421622	7.40575	7.023074
Bolivia	BOL	Agricultur	EN.ATM.M	46.26927	52.5399	46.68798	49.56755	50.16601	48.47314	46.64157	52.67861	50.89993	51.53996	51.17959	52.182	51.10256
Brazil	BRA	Agricultur	EN.ATM.M	60.85901	67.89232	64.59714	68.04933	66.90145	64.86558	64.67922	69.4817	65.88613	64.63659	64.74182	66.90407	66.20072
Barbados	BRB	Agricultur	EN.ATM.M	41.66594	41.32096	38.98815	38.26058	36.18671	32.98504	33.75042	33.62255	35.34683	36.2172	36.98468	37.17423	35.4938

Dataset Search



Try [coronavirus covid-19](#) or [water quality site:canada.ca](#).

[Learn more](#) about Dataset Search.

Data Collection



- Government Websites (Federal and States)
- Published data from articles
- Publicly available datasets
- Using APIs to scrape data

The screenshot displays the Data.gov website interface. At the top, there's a blue header with "DATA CATALOG" on the left and navigation links for "Datasets" and "Organizations" on the right. Below the header, a search bar is labeled "Search datasets...". To the right of the search bar, there's an "Order by:" dropdown menu set to "Popular".

On the left side, there's a "Filter by location" section with a "Clear" button and an "Enter location..." input field. Below this is a map of the United States with a zoom in (+) and zoom out (-) button. Under the map, there's a "Topics" section with a list of categories and their counts: Local Government (21199), Climate (525), Older Adults... (89), and Energy (21). Below the topics, there's a "Topic Categories" section with a list of categories and their counts: Arctic (133), Ecosystem Vulnerability (91), and Water (89).

The main content area shows "291,272 datasets found". Below this, there are three dataset listings:

- Electric Vehicle Population Data** (4023 recent views) - State of Washington — This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department... (CSV, RDF, JSON, XML)
- Lottery Powerball Winning Numbers: Beginning 2010** (3368 recent views) - State of New York — Go to <http://on.ny.gov/1GpWiHD> on the New York Lottery website for past Powerball results and payouts. (CSV, RDF, JSON, XML)
- Crime Data from 2020 to Present** (3262 recent views) - City of Los Angeles — Starting on March 7th, 2024, the Los Angeles Police Department (LAPD) will adopt a new Records Management System for reporting crimes and arrests. This new system is... (CSV, RDF, JSON, XML)

At the bottom, there's a listing for **FDIC Failed Bank List** (2369 recent views) - Federal Deposit Insurance Corporation — The FDIC is often appointed as receiver for failed

Data Collection

- Government Websites (Federal and States)
- Published data from articles
- Publicly available datasets
- Using APIs to scrape data

Biol Invasions (2022) 24:1895–1904
<https://doi.org/10.1007/s10530-021-02568-7>

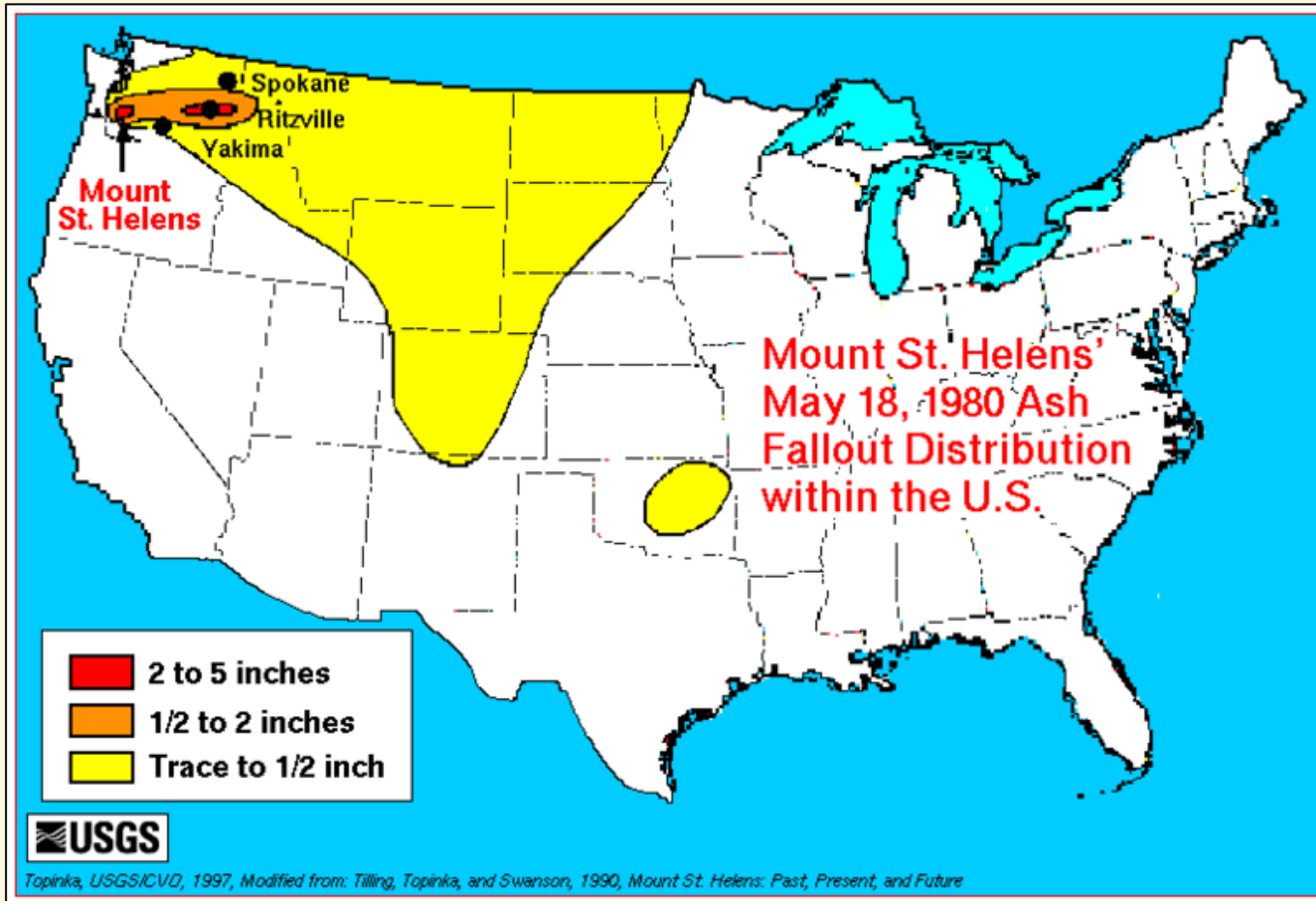
ORIGINAL PAPER

Are the “100 of the world’s worst” invasive species also the costliest?

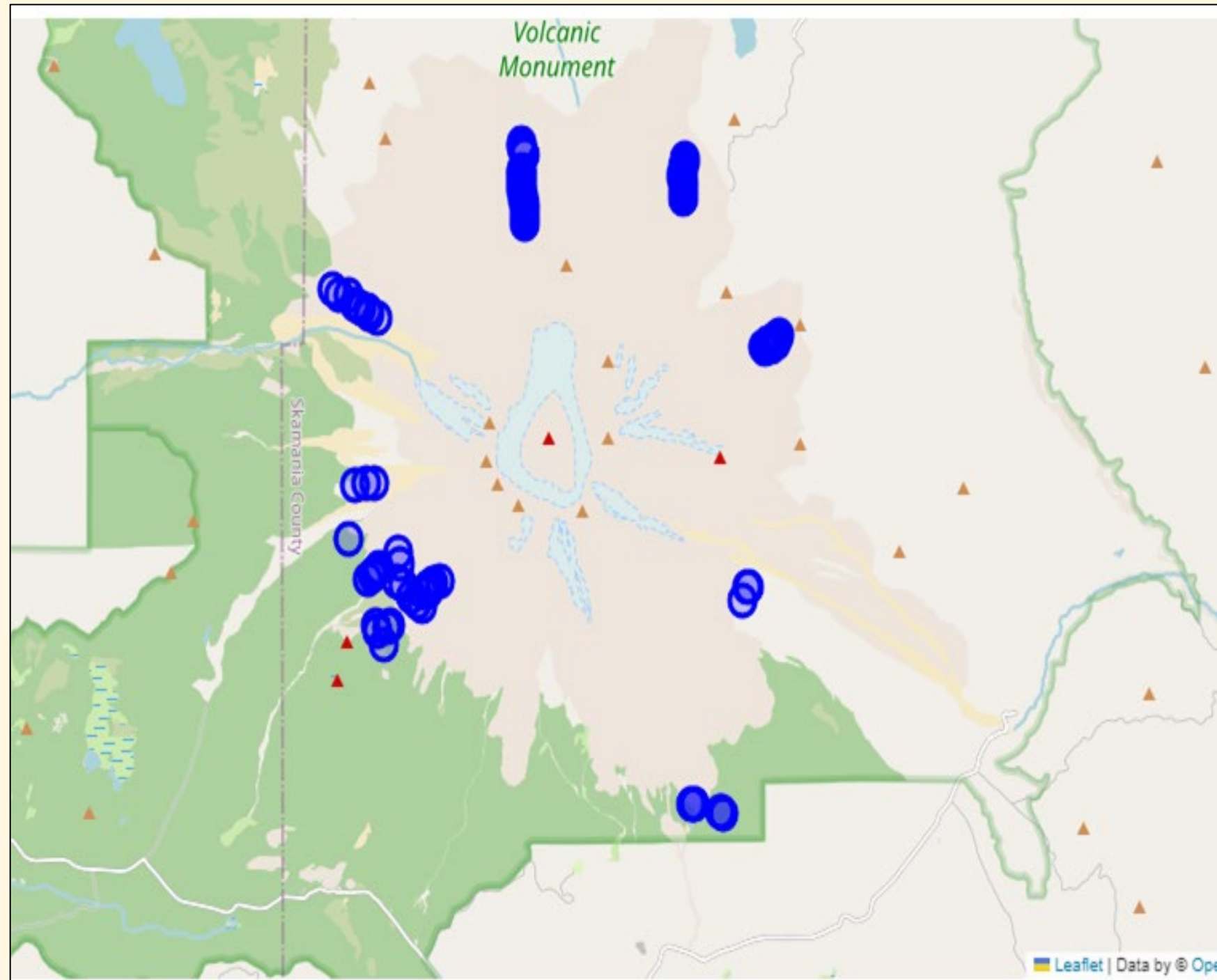
Ross N. Cuthbert · Christophe Diagne · Phillip J. Haubrock · Anna J. Turbelin · Franck Courchamp

Availability of data and material Underlying data are publicly available in Diagne et al. (2020b: accessible at <https://www.nature.com/articles/s41597-020-00586-z>) and in an online repository (<https://doi.org/10.6084/m9.figshare.12668570>). The final dataset used for analysis in this paper will be provided as Supplementary Material.

Ecological Recovery with Mount St. Helen's Data



Ecological Recovery with Mount St. Helen's Data



Data Collection



- Government Websites (Federal and States)
- Published data from articles
- **Other publicly available datasets**
- Using APIs to scrape data



Indicators		
Featured indicators	All indicators	<input type="text" value="Quick search"/>
Agriculture & Rural Development		Agriculture & Rural Development
Access to electricity, rural (% of rural population)		Aid Effectiveness
Agricultural irrigated land (% of total agricultural land)		Climate Change
Agricultural land (% of land area)		Economy & Growth
Agricultural land (sq. km)		Education
Agricultural machinery, tractors		Energy & Mining
Agricultural machinery, tractors per 100 sq. km of arable land		Environment
Agricultural methane emissions (% of total)		External Debt
Agricultural methane emissions (thousand metric tons of CO2 equivalent)		Financial Sector
Agricultural nitrous oxide emissions (% of total)		Gender
Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)		Health
Agricultural raw materials exports (% of merchandise exports)		Infrastructure
Agricultural raw materials imports (% of merchandise imports)		Poverty
Agriculture, forestry, and fishing, value added (% of GDP)		Private Sector
Agriculture, forestry, and fishing, value added (current US\$)		Public Sector
Annual freshwater withdrawals, agriculture (% of total freshwater withdrawal)		
Arable land (% of land area)		
Arable land (hectares per person)		
Arable land (hectares)		
Average precipitation in depth (mm per year)		
Cereal production (metric tons)		
Cereal yield (kg per hectare)		
Crop production index (2014-2016 = 100)		
Employment in agriculture (% of total employment) (modeled ILO estimate)		
Employment in agriculture, female (% of female employment) (modeled ILO estimate)		

Data Collection

- Government Websites (Federal and States)
- Published data from articles
- Publicly available datasets
- Using APIs to access data

Index

[Animals](#)
[Anime](#)
[Anti-Malware](#)
[Art & Design](#)
[Authentication & Authorization](#)
[Blockchain](#)
[Books](#)
[Business](#)
[Calendar](#)
[Cloud Storage & File Sharing](#)
[Continuous Integration](#)
[Cryptocurrency](#)
[Currency Exchange](#)
[Data Validation](#)
[Development](#)
[Dictionaries](#)
[Documents & Productivity](#)
[Email](#)
[Entertainment](#)
[Environment](#)
[Events](#)
[Finance](#)
[Food & Drink](#)
[Games & Comics](#)
[Geocoding](#)
[Government](#)
[Health](#)
[Jobs](#)
[Machine Learning](#)
[Music](#)
[News](#)
[Open Data](#)
[Open Source Projects](#)
[Patent](#)
[Personality](#)
[Phone](#)
[Photography](#)
[Programming](#)
[Science & Math](#)
[Security](#)
[Shopping](#)
[Social](#)
[Sports & Fitness](#)
[Test Data](#)
[Text Analysis](#)
[Tracking](#)
[Transportation](#)
[URL Shorteners](#)
[Vehicle](#)
[Video](#)
[Weather](#)

THIS IS YOUR MACHINE LEARNING SYSTEM?

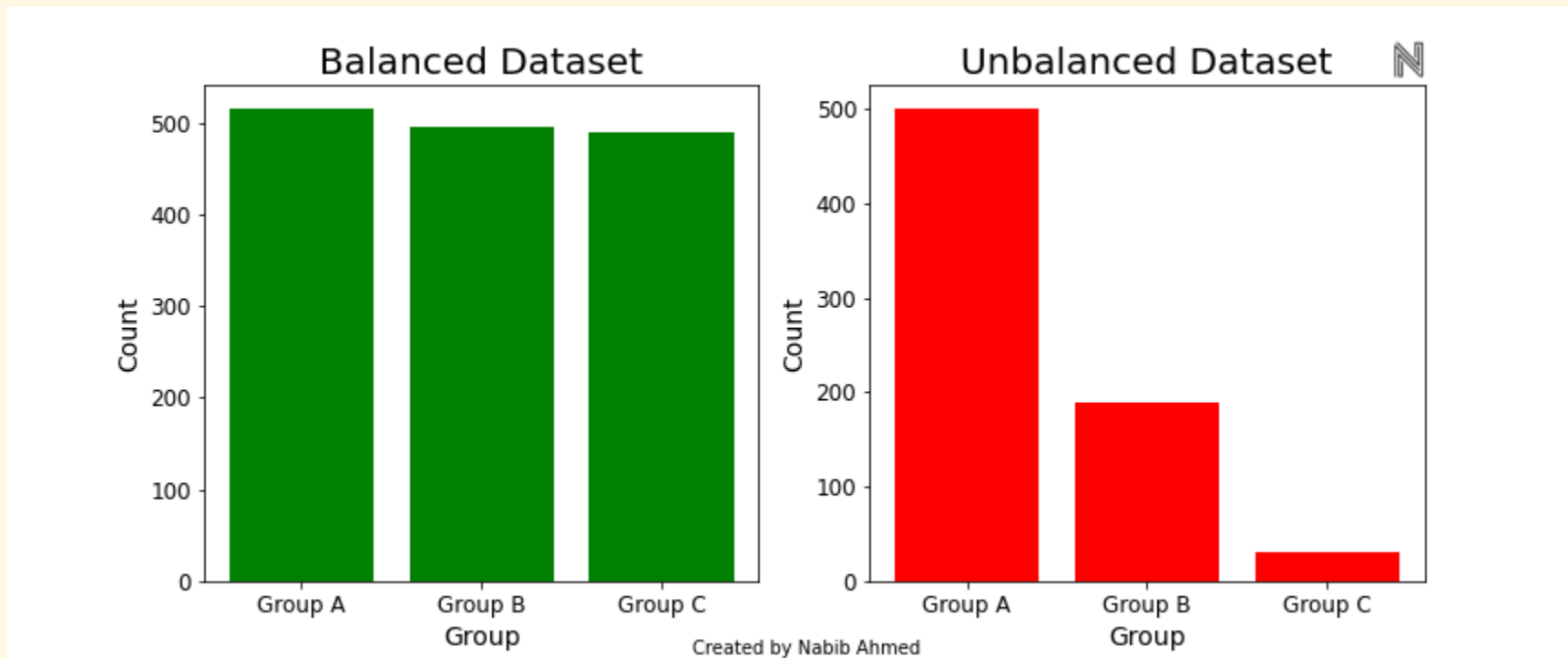
YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.



Sampling Bias



Example:

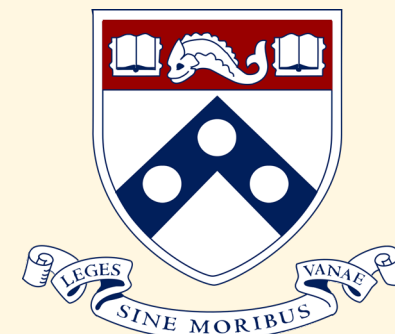
A company performed a survey of 20 of their total 100 workers about job satisfaction. Looking at the data, do we think they performed a good survey?

Things to note:

- The two largest departments are finance and sales.
- The average age of the workers at the company is about 37 years old.
- There's 47 male workers and 53 female workers at the company.
- The average time at the company is 8 years.

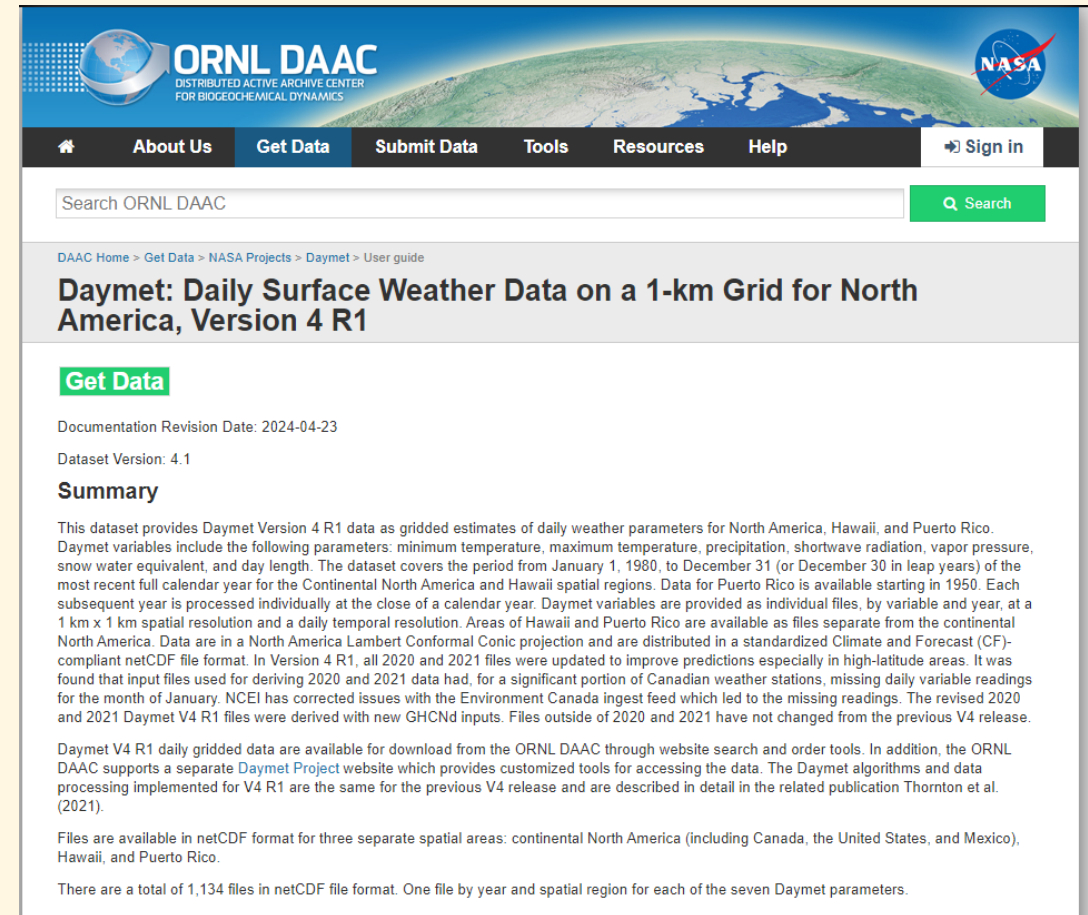
Okay, you've identified a dataset, now what?

- Who published it?



Okay, you've identified a dataset, now what?

- Who published it?
- Do they describe the data and it was collected?



The screenshot shows the ORNL DAAC (Distributed Active Archive Center for Biogeochemical Dynamics) website. The header includes the ORNL DAAC logo and the NASA logo. The navigation bar contains links for Home, About Us, Get Data, Submit Data, Tools, Resources, Help, and Sign in. A search bar is located below the navigation bar. The main content area displays the breadcrumb trail: DAAC Home > Get Data > NASA Projects > Daymet > User guide. The title of the dataset is "Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1". A green "Get Data" button is prominently displayed. Below this, the documentation revision date is 2024-04-23 and the dataset version is 4.1. The "Summary" section provides a detailed description of the dataset, including its coverage (North America, Hawaii, and Puerto Rico), temporal resolution (daily), and spatial resolution (1 km x 1 km). It also mentions that the data is available in netCDF format and that the ORNL DAAC supports a separate Daymet Project website for customized tools. The footer notes that there are a total of 1,134 files in netCDF format, one file by year and spatial region for each of the seven Daymet parameters.

ORNL DAAC
DISTRIBUTED ACTIVE ARCHIVE CENTER
FOR BIOGEOCHEMICAL DYNAMICS

NASA

Home About Us Get Data Submit Data Tools Resources Help Sign in

Search ORNL DAAC

DAAC Home > Get Data > NASA Projects > Daymet > User guide

Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1

[Get Data](#)

Documentation Revision Date: 2024-04-23

Dataset Version: 4.1

Summary

This dataset provides Daymet Version 4 R1 data as gridded estimates of daily weather parameters for North America, Hawaii, and Puerto Rico. Daymet variables include the following parameters: minimum temperature, maximum temperature, precipitation, shortwave radiation, vapor pressure, snow water equivalent, and day length. The dataset covers the period from January 1, 1980, to December 31 (or December 30 in leap years) of the most recent full calendar year for the Continental North America and Hawaii spatial regions. Data for Puerto Rico is available starting in 1950. Each subsequent year is processed individually at the close of a calendar year. Daymet variables are provided as individual files, by variable and year, at a 1 km x 1 km spatial resolution and a daily temporal resolution. Areas of Hawaii and Puerto Rico are available as files separate from the continental North America. Data are in a North America Lambert Conformal Conic projection and are distributed in a standardized Climate and Forecast (CF)-compliant netCDF file format. In Version 4 R1, all 2020 and 2021 files were updated to improve predictions especially in high-latitude areas. It was found that input files used for deriving 2020 and 2021 data had, for a significant portion of Canadian weather stations, missing daily variable readings for the month of January. NCEI has corrected issues with the Environment Canada ingest feed which led to the missing readings. The revised 2020 and 2021 Daymet V4 R1 files were derived with new GHCNd inputs. Files outside of 2020 and 2021 have not changed from the previous V4 release.

Daymet V4 R1 daily gridded data are available for download from the ORNL DAAC through website search and order tools. In addition, the ORNL DAAC supports a separate [Daymet Project](#) website which provides customized tools for accessing the data. The Daymet algorithms and data processing implemented for V4 R1 are the same for the previous V4 release and are described in detail in the related publication Thornton et al. (2021).

Files are available in netCDF format for three separate spatial areas: continental North America (including Canada, the United States, and Mexico), Hawaii, and Puerto Rico.

There are a total of 1,134 files in netCDF file format. One file by year and spatial region for each of the seven Daymet parameters.

Okay, you've identified a dataset, now what?

- Who published it?
- Do they describe the data and it was collected?
- Is the data up-to-date?

[DAAC Home](#) > [Get Data](#) > [NASA Projects](#) > [Daymet](#) > [User guide](#)

Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1

Get Data

Documentation Revision Date: 2024-04-23

Dataset Version: 4.1

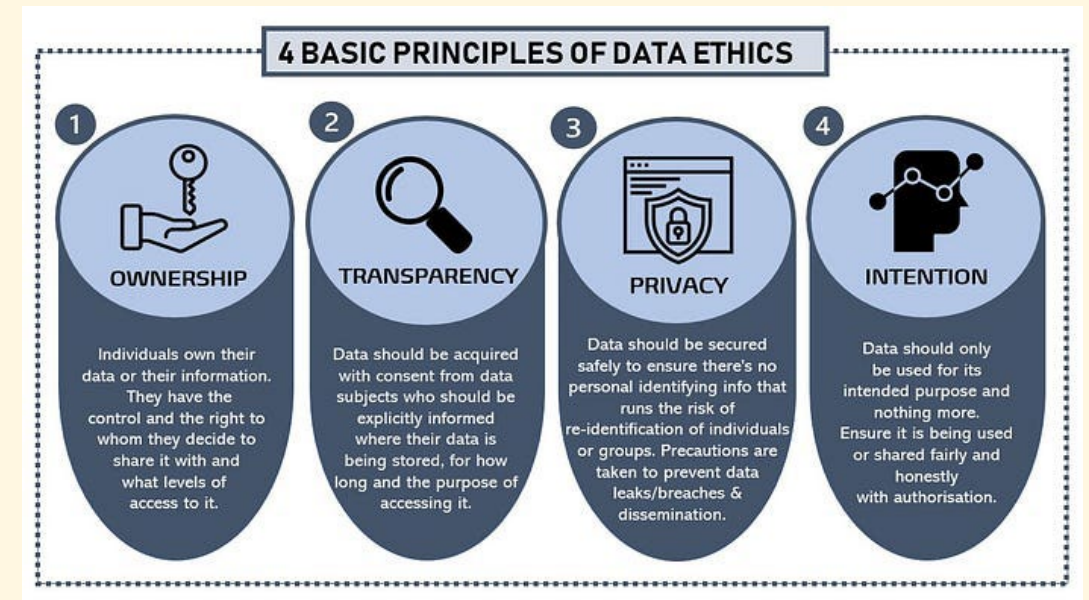
Okay, you've identified a dataset, now what?

- Who published it?
- Do they describe the data and it was collected?
- Is the data up-to-date?
- How complete is the data?

ID	Name	Age	Gender	Major	Study Hours per Week	GPA	Preferred Study Method	Graduation Year
1	Alice	20	F	Computer Science	15	3.8	Group Study	2023
2	Bob		M		10	3.2	Solo Study	
3	Charlie	21		Mathematics				2024
4	Dorothy		F	Physics	12	3.5	Group Study	2023
5	Eve	22		Computer Science		3.9	Online Resources	
6	Frank		M	Biology	8		Solo Study	2025
7	Grace	23	F		20	3.7		
8	Hector		M	Chemistry		3.4	Group Study	2024
9	Ivy	21	F	Mathematics	14		Online Resources	
10	Jack							2023

Okay, you've identified a dataset, now what?

- Who published it?
- Do they describe the data and it was collected?
- Is the data up-to-date?
- How complete is the data?
- How was the data collected and shared?



Okay, you've identified a dataset, now what?

- Who published it?
- Do they describe the data and it was collected?
- Is the data up-to-date?
- How complete is the data?
- How was the data collected and shared?
- How useable is it?

Format?

Cost?

Accessibility?

Let's try an example:



<https://www.kaggle.com/datasets/thedevastator/us-weather-history-12-months-of-record-setting-t>