

# Dataset Cheat Sheet

## BASICS

☐ Definition:

A dataset is a structured collection of data, typically organized in a table or database format.

☐ Importance:

Datasets are crucial for training, validating, and testing AI models, enabling data-driven decision-making.

☐ Examples of personal data:

<https://rgpd.com/basics/200examples-of-personal-data/>

## PROHIBITED DATASETS

☐ Sensitive Data:

Datasets containing personal data without consent or that violate privacy laws (e.g., GDPR).

☐ Biased Datasets:

Datasets that reinforce discrimination or bias against individuals or groups.

☐ Fake or Misleading Data:

Datasets that include fabricated or misleading information.

## HIGH RISK DATA SETS

☐ Definition:

Datasets used in high-stakes applications (e.g., healthcare, criminal justice) that may impact individuals' rights or freedoms.

## KEY REQUIREMENTS FOR HIGH RISK DATA SETS

☐ Data Quality:

Ensure accuracy, completeness, and reliability of data.

☐ Documentation:

Maintain clear records of data sources, collection methods, and processing.

☐ Bias Mitigation:

Implement strategies to identify and reduce bias in datasets.

☐ Transparency:

Provide clear information about data usage and limitations.

## GENERAL PURPOSE DATASETS

☐ Definition:

Datasets that can be used across various applications and domains.

☐ Examples:

MNIST, CIFAR-10, the Iris dataset, the Titanic dataset, IMDB reviews, and collections from the UCI Machine Learning Repository.

☐ Best Practices:

Ensure interoperability and standardization. Maintain clear licensing and usage rights.

## PENALTIES AND ENFORCEMENT

☐ Non-Compliance:

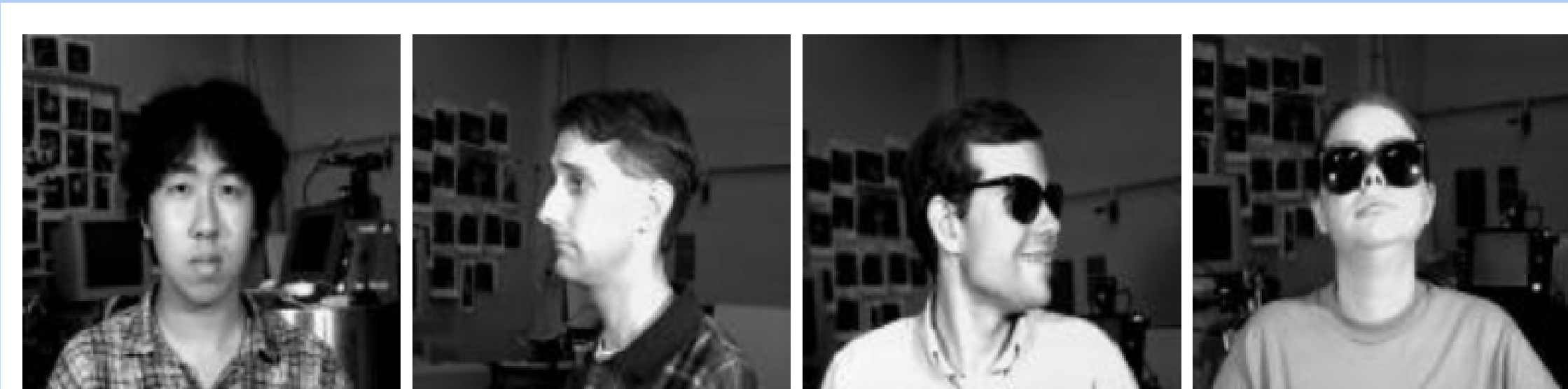
Organizations may face fines or sanctions for using prohibited datasets or failing to meet requirements.

☐ Regulatory Oversight:

Authorities monitor compliance with data protection laws and ethical guidelines

# Dataset Examples

## Pictorial Data



| User ID  | Pose     | Expression | Eyes       | Scale |
|----------|----------|------------|------------|-------|
| an2i     | straight | neutral    | open       | 1     |
| at33     | left     | happy      | sunglasses | 2     |
| boland   | right    | sad        | open       | 4     |
| bpm      | up       | angry      | sunglasses | 1     |
| ch4f     | straight | happy      | open       | 2     |
| cheyer   | left     | neutral    | sunglasses | 4     |
| choon    | right    | sad        | open       | 1     |
| danieln  | up       | happy      | sunglasses | 2     |
| glickman | straight | angry      | open       | 4     |
| karyadi  | left     | neutral    | sunglasses | 1     |

### EXPLANATION

- ☐ User ID: Identifies the person in the image
- ☐ Pose: Describes the head position
- ☐ Expression: Indicates the person's facial expression
- ☐ Eyes: Shows the state of the person's eyes
- ☐ Scale: Represents the resolution of the image

# Dataset Examples

## Time Series Data

| Date time        | Site name | County         | Aqi | Pollutant | Status   | SO <sub>2</sub> | NO <sub>2</sub> | No <sub>x</sub> | NO  | Wind speed | Wind direc | co_8hr | Pm 2.5_avg | pm10_avg | SO <sub>2</sub> _avg | Longitude | Latitude | Site ID |
|------------------|-----------|----------------|-----|-----------|----------|-----------------|-----------------|-----------------|-----|------------|------------|--------|------------|----------|----------------------|-----------|----------|---------|
| 31/08/2024 23:00 | Hukou     | Hsinchu County | 62  | PM2.5     | Moderate | 0.9             | 2.3             | 2.6             | 0.3 | 2.3        | 225        | 0.2    | 20.1       | 26       | 1                    | 121.0389  | 24.9001  | 22      |
| 31/08/2024 23:00 | Zhongming | Taichung City  | 50  |           | Good     | 1.6             | 7.6             | 9.3             | 1.6 | 1.1        | 184        | 0.2    | 15.3       | 23       | 1                    | 120.6411  | 24.15196 | 31      |
| 31/08/2024 23:00 | Zhudong   | Hsinchu County | 45  |           | Good     | 0.4             | 2.9             | 4.1             | 1.1 | 0.4        | 210        | 0.2    | 13.8       | 24       | 0                    | 121.089   | 24.74091 | 23      |
| 31/08/2024 23:00 | Hsinchu   | Hsinchu City   | 42  |           | Good     | 0.8             | 4               | 4.8             | 0.7 | 1.9        | 239        | 0.2    | 13         | 26       | 1                    | 120.9724  | 24.80564 | 24      |
| 31/08/2024 23:00 | Toufen    | Miaoli County  | 50  |           | Good     | 1               | 1.8             | 3.1             | 1.2 | 1.8        | 259        | 0.1    | 15.3       | 28       | 1                    | 120.8987  | 24.69691 | 25      |
| 31/08/2024 23:00 | Miaoli    | Miaoli County  | 40  |           | Good     | 1.1             | 4               | 5.1             | 1.1 | 1.4        | 235        | 0.1    | 12.2       | 17       | 1                    | 120.8201  | 24.56499 | 26      |
| 31/08/2024 23:00 | Sanyi     | Miaoli County  | 39  |           | Good     | 0.9             | 2.4             | 3.1             | 0.7 | 0.6        | 203        | 0.1    | 11.4       | 16       | 0                    | 120.7596  | 24.38248 | 27      |
| 31/08/2024 23:00 | Fengyuan  | Taichung City  | 44  |           | Good     | 1.3             | 6.8             | 7.3             | 0.5 | 1.2        | 38         | 0.2    | 13.5       | 21       | 1                    | 120.7425  | 24.257   | 28      |
| 31/08/2024 23:00 | Shalu     | Taichung City  | 46  |           | Good     | 2.5             | 7.3             | 7.7             | 0.3 | 0.5        | 8          | 0.1    | 14.2       | 26       | 1                    | 120.5688  | 24.22563 | 29      |
| 31/08/2024 23:00 | Dali      | Taichung City  | 49  |           | Good     | 0.7             | 5.6             | 6.3             | 0.7 | 0.9        | 97         | 0.2    | 15.1       | 21       | 0                    | 120.6784  | 24.09961 | 30      |

# Dataset Examples

## Personal Data

| Passenger (P)<br>Id | P<br>class | Name                                         | Sex    | Age  | SibSp | Parch | Ticket  | Fare    | Embark<br>-ed |
|---------------------|------------|----------------------------------------------|--------|------|-------|-------|---------|---------|---------------|
| 892                 | 3          | Kelly, Mr. James                             | Male   | 34.5 | 0     | 0     | 330911  | 7.8292  | Q             |
| 893                 | 3          | Wilkes, Mrs. James (Ellen Needs)             | Female | 47   | 1     | 0     | 363272  | 7       | S             |
| 894                 | 2          | Myles, Mr. Thomas Francis                    | Male   | 62   | 0     | 0     | 240276  | 9.6875  | Q             |
| 895                 | 3          | Wirz, Mr. Albert                             | Male   | 27   | 0     | 0     | 315154  | 8.6625  | S             |
| 896                 | 3          | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | Female | 22   | 1     | 1     | 3101298 | 12.2875 | S             |
| 897                 | 3          | Svensson, Mr. Johan Cervin                   | Male   | 14   | 0     | 0     | 7538    | 9.225   | S             |
| 898                 | 3          | Connolly, Miss. Kate                         | Female | 30   | 0     | 0     | 330972  | 7.6292  | Q             |
| 899                 | 2          | Caldwell, Mr. Albert Francis                 | Male   | 26   | 1     | 1     | 248738  | 29      | S             |
| 900                 | 3          | Abraham, Mrs. Joseph (Sophie Halaut Easu)    | Female | 18   | 0     | 0     | 2657    | 7.2292  | C             |

## Variables Description

☐ survival: Survival (0 = No, 1 = Yes)

☐ P class: Ticket class (1 = 1<sup>st</sup>, 2 = 2<sup>nd</sup>, 3 = 3<sup>rd</sup>)

☐ sex: Sex

☐ Age: Age in years

☐ sibsp: # of siblings/spouses aboard the Titanic

☐ parch: # of parents/children aboard the Titanic

☐ ticket: Ticket number

☐ fare: Passenger fare

☐ cabin: Cabin number

☐ embarked: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)