**Sampling bias:** a sample that isn’t representative of the population as a whole

**Observer bias** (experimenter bias/ research bias): the tendency for different people to observe things differently

**Interpretation bias:** the tendency to always interpret (dien giai) ambiguous situations in a positive or negative ways

**Confirmation bias**: the tendency to search for or interpret information in a way that confirms pre-existing beliefs

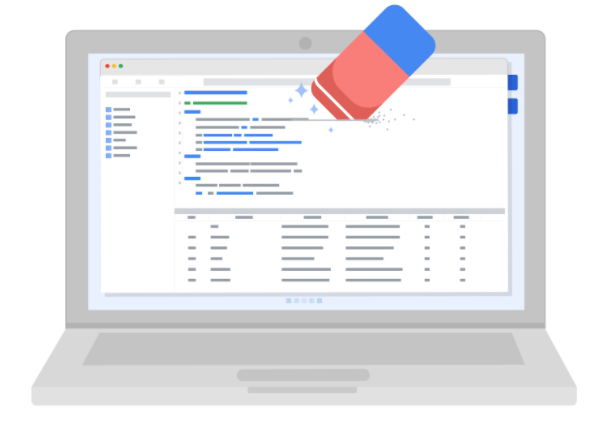
**The dataset is R.O.C.C.C**

* Reliable
* Original
* Comprehensive
* Current
* Cited

**Aspects of data ethics**

* Ownerships
* Transaction transparency
* Consent: Now let's talk about another aspect of data ethics, consent. This is an individual's right to know explicit details about how and why their data will be used before agreeing to provide it. They should know answers to questions like why is the data being collected? How will it be used? How long will it be stored? The best way to give consent is probably a conversation between the person providing the data and the person requesting it.
* Currency:
* Privacy
* Openess

Healthcare and financial data are two of the most sensitive types of data. These industries rely a lot on data anonymization techniques. After all, the stakes are very high. That’s why data in these two industries usually goes through **de-identification**, which is **a process used to wipe data clean of all personally identifying information**.

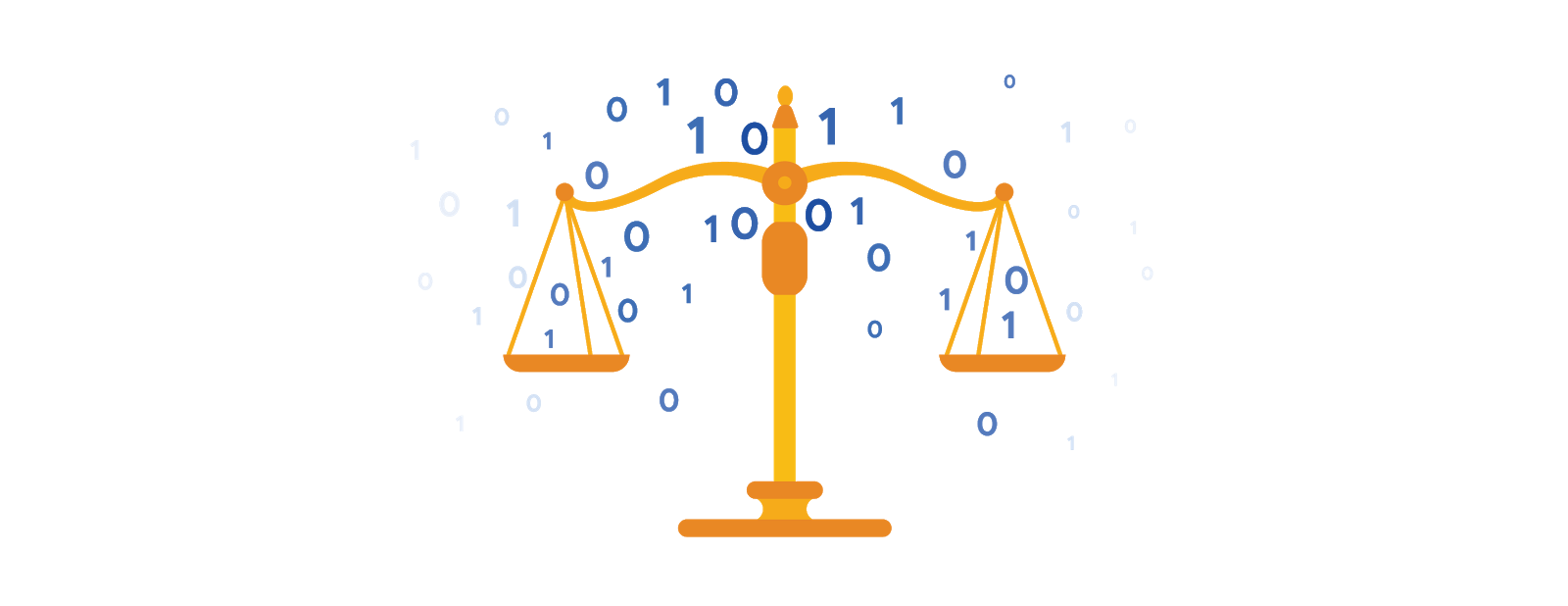


Data anonymization is used in just about every industry. That is why it is so important for data analysts to understand the basics. Here is a list of data that is often anonymized:

* Telephone numbers
* Names
* License plates and license numbers
* Social security numbers
* IP addresses
* Medical records
* Email addresses
* Photographs
* Account numbers

**The open data debate**

Just like data privacy, open data is a widely debated topic in today’s world. Data analysts think a lot about open data, and as a future data analyst, you need to understand the basics to be successful in your new role.



**What is open data?**

In data analytics, **open data** is part of **data ethics,** which has to do with using data ethically. **Openness** refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

* Be available and accessible to the public as a complete dataset
* Be provided under terms that allow it to be reused and redistributed
* Allow universal participation so that anyone can use, reuse, and redistribute the data

Data can only be considered open when it meets all three of these standards.

**The open data debate: What data should be publicly available?**

One of the biggest benefits of open data is that credible databases can be used more widely. Basically, this means that all of that good data can be leveraged, shared, and combined with other data. This could have a huge impact on scientific collaboration, research advances, analytical capacity, and decision-making. But it is important to think about the individuals being represented by the public, open data, too.

**Third-party data** is collected by an entity that doesn’t have a direct relationship with the data. You might remember learning about this type of data earlier. For example, third parties might collect information about visitors to a certain website. Doing this lets these third parties create audience profiles, which helps them better understand user behavior and target them with more effective advertising.

**Personal identifiable information (PII)** is data that is reasonably likely to identify a person and make information known about them. It is important to keep this data safe***.*** PII can include a person’s address, credit card information, social security number, medical records, and more.

Everyone wants to keep personal information about themselves private. Because third-party data is readily available, it is important to balance the openness of data with the privacy of individuals.

**Resources for open data**

Luckily for data analysts, there are lots of trustworthy resources available for open data. It is important to remember that even reputable data needs to be constantly evaluated, but these websites are a useful starting point:

1. [**U.S. government data site**](https://www.data.gov/): Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
2. [**U.S. Census Bureau**](https://www.census.gov/data.html): This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
3. [**Open Data Network**](https://www.opendatanetwork.com/): This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
4. [**Google Cloud Public Datasets**](https://cloud.google.com/datasets): There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
5. [**Dataset Search**](https://datasetsearch.research.google.com/)**:** The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

**Glossary terms from module 2**

**Terms and definitions for Course 3, Module 2**

**Bad data source:** A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

**Bias:** A conscious or subconscious preference in favor of or against a person, group of people, or thing

**Confirmation bias:** The tendency to search for or interpret information in a way that confirms pre-existing beliefs

**Consent:** The aspect of data ethics that presumes an individual’s right to know how and why their personal data will be used before agreeing to provide it

**Cookie:** A small file stored on a computer that contains information about its users

**Currency:** The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

**Data anonymization:** The process of protecting people's private or sensitive data by eliminating identifying information

**Data bias:** When a preference in favor of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

**Data ethics:** Well-founded standards of right and wrong that dictate how data is collected, shared, and used

**Data interoperability:** A key factor leading to the successful use of open data among companies and governments

**Data privacy:** Preserving a data subject’s information any time a data transaction occurs

**Ethics:** Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

**Experimenter bias:** The tendency for different people to observe things differently (also called observer bias)

**Fairness:** A quality of data analysis that does not create or reinforce bias

**First-party data:** Data collected by an individual or group using their own resources

**General Data Protection Regulation of the European Union (GDPR):** Policy-making body in the European Union created to help protect people and their data

**Good data source:** A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

**Interpretation bias:** The tendency to interpret ambiguous situations in a positive or negative way

**Observer bias:** The tendency for different people to observe things differently (also called experimenter bias)

**Open data:** Data that is available to the public

**Openness:** The aspect of data ethics that promotes the free access, usage, and sharing of data

**Sampling bias:** Overrepresenting or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

**Transaction transparency:** The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

**Unbiased sampling:** When the sample of the population being measured is representative of the population as a whole