## **Summary:**

**Data Ethics:**

Data ethics refers to well- founded standards of right and wrong that dictate how data is collected, shared, and used. Since the ability to collect, share and use data in such large quantities is relatively new, the rules that regulate and govern the process are still evolving.

The concept of data ethics and issues related to transparency and privacy are part of the process. Data ethics tries to get to the root of the accountability companies have in protecting and responsibly using the data they collect.

**According to GoogleDAC, the six data ethics are as follows:->**

There are lots of different aspects of data ethics but we'll cover six: ownership, transaction transparency, consent, currency, privacy, and openness.

**1.Ownership**: This answers the question: who owns data? It isn't the organization that invested time and money collecting, storing, processing, and analyzing it. It's individuals who own the raw data they provide, and they have primary control over its usage, how it's processed and how it's shared.

**2.Transaction Transparency**: which is the idea that all data processing activities and algorithms should be completely explainable and understood by the individual who provides their data. This is in response to concerns over data bias, which we discussed earlier, is a type of error that systematically skews results in a certain direction. Biased outcomes can lead to negative consequences. To avoid them, it's helpful to provide transparent analysis especially to the people who share their data.

**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

Consent is important because it prevents all populations from being unfairly targeted which is a very big deal for marginalized groups who are often disproportionately misrepresented by biased dat

But with so much activity happening online these days, consent usually just looks like a terms and conditions checkbox with links to more details.

**4.Currency**: Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions. If your data is helping to fund a company's efforts, you should know what those efforts are all about and be given the opportunity to opt out. The last two aspects of data ethics, privacy and openness, deserve their own spotlight on this data stage. Coming up, you'll see why.

**5.Privacy:** When talking about data, privacy means preserving a data subject's information and activity any time a data transaction occurs. This is sometimes called information privacy or data protection. It's all about access, use, and collection of data. It also covers a person's legal right to their data.

It also covers a person's legal right to their data. This means someone like you or me should have protection from unauthorized access to our private data, freedom from inappropriate use of our data, the right to inspect, update, or correct our data, ability to give consent to use our data, and legal right to access our data

**6.Openness:** The idea of openness, free access, usage, and sharing of 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.

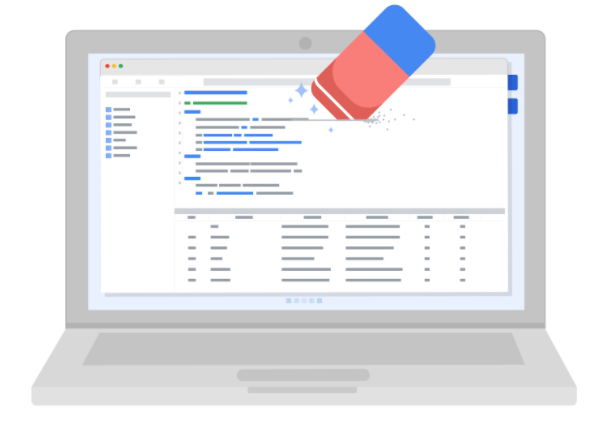
**Personal identifiable information (PII)** is data that is reasonably likely to identify a person and make information known about them. It is important to keep this data safe***.*** PII can include a person’s address, credit card information, social security number, medical records, and more.

**Data Anonymization:**

Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values.

## **What types of data should be anonymized?**

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

For some people, it just makes sense that this type of data should be anonymized. For others, we have to be very specific about what needs to be anonymized. Imagine a world where we all had access to each other’s addresses, account numbers, and other identifiable information. That would invade a lot of people’s privacy and make the world less safe. Data anonymization is one of the ways we can keep data private and secure!

# Sites and resources for open data

Luckily for data analysts, there are lots of trustworthy sites and 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/public-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.

Question 2

So far, you’ve learned a lot about how to use Kaggle to explore datasets. During this activity, you used this knowledge to find datasets you’re interested in. Keep those datasets in mind and in the text box below, write 2-3 sentences (40-60 words) in response to each of the following questions:

* What type(s) of data are in this dataset? Is this dataset biased or unbiased? How do you know?
* Based on what you’ve explored so far, how might you use Kaggle’s Datasets feature to help develop your data analysis skills?

**Answer:** The datasets consist of all sorts of qualitative and quantitave data which are in structured form. To know if the data is biased or not we'll have to plot the data and then we can see if there are certain outliers.

Kaggle consists of a humongous amount of open source datasets which can be used for all sorts of purposes. In near future I'll be using kaggle to fetch me datasets to carry out my analysis projects on.