WHAT IS DARK DATA? THE BASICS & THE CHALLENGES

Dark data and unstructured data are about the same thing. The difference lies in *to whom* the term is directed. Unstructured data tends to be a word directed at engineers. It refers to the structural qualities of the data, signaling to the engineer how they'll have to go about refining the data to make any use of it.

Unstructured data is unrefined data, requiring more work to make it usable; structured data is already refined data where the data's purpose is already determined. Unstructured data is the yin to structured data's yang, but, mostly, unstructured data comes from an engineering-centric point of view.

What is dark data?

Dark data, however, emerges from the user-centric point of view. Where structured data refers to the structural qualities of the data, dark data refers to the visible qualities of the data. There is data the user can see, like Instagram photos, profile names, hashtags, but then there is data the user cannot see. The Dark Data.

On a social media platform like Instagram, the dark data would be:

- How many login instances does the user have?
- Does their user activity cluster around certain times of the day?
- How many people liked the post who have large networks of users? (To measure a user's clout.)

- From where was the photo taken?
- Where was the person when they posted the photo?

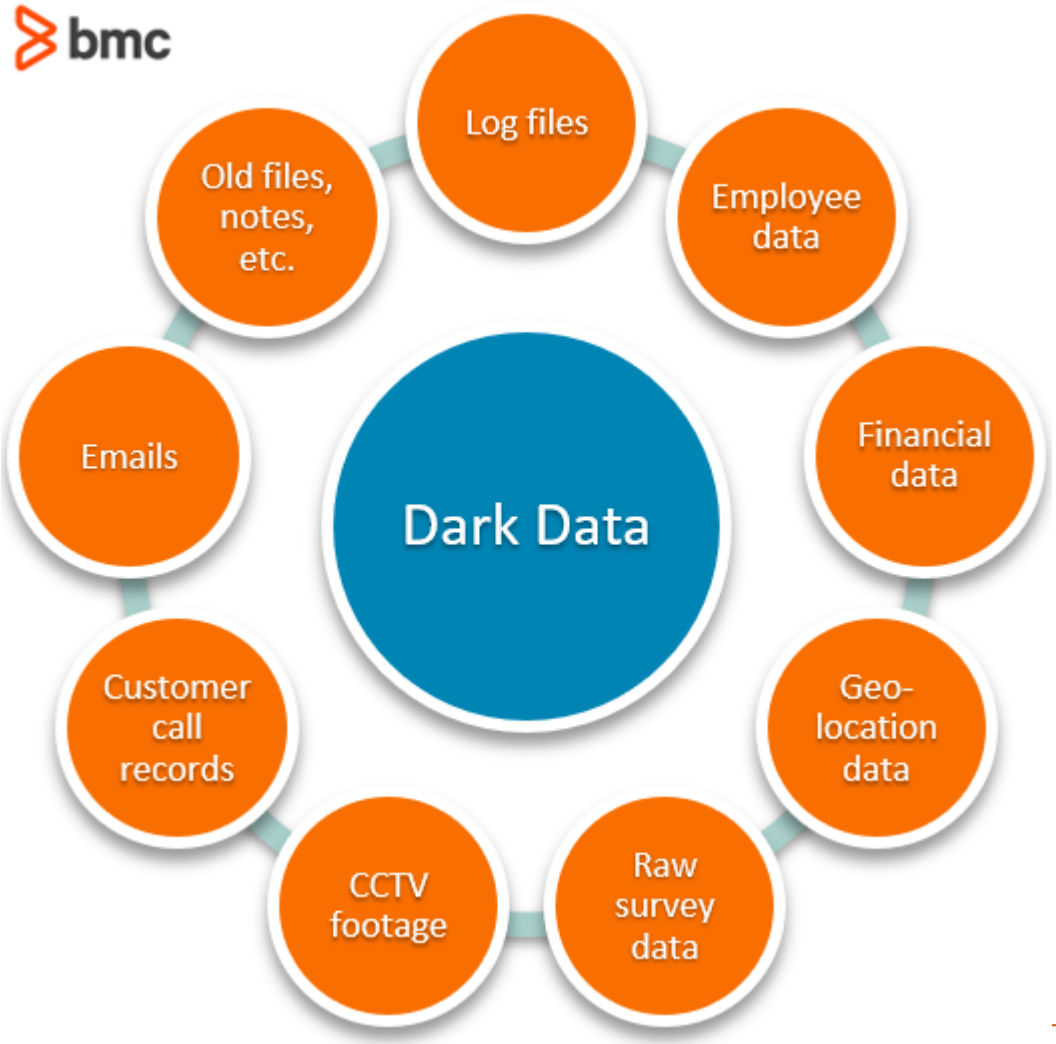
People can get overwhelmed by seeing so much data. Standard design practice says Keep It Simple Stupid (KISS) and holds white space as its central virtue. Instagram even decreased the amount of data it showed by generalizing the number of likes a photo would get from a very specific 134,392 to simply saying, "Thousands".

When the users are the engineers, dark data will refer to unstructured data that does not get analyzed. It's the data stored through various network processes on servers and in data lakes that ends up sitting around to satisfy the industry's statute of limitations or is kept because data storage can be so cheap.

Types of dark data

The types of dark data that exist are industry specific. Background weather data might be collected in a running app, and browser history might be collected in a shopping app.

Basically, anything that is sent over the internet has potential to be, and create dark data. Packages are sent from point A to point B. While those packages can be encrypted and those looking in can have a hard time seeing what is in the package itself, there are other known entities in the process.



Types of dark data include:

- Log files (servers, systems, architecture, etc.)

- Previous employee data
- Financial statements
- Geolocation data
- Raw survey data
- Surveillance video footage
- Customer call records
- Email correspondences
- Notes, presentations, or old documents

Most data is dark data

 In order to make software services work, some data must be collected. IP address must be known to get data from somewhere else on the network and return it to a user somewhere else on the network. Artificial Intelligence-backed services are showing how the more data a company has on a user, the better the service they can provide.

The IDC estimates that 90% of data is unstructured data. A.I. is helping make more use of this unstructured data, which should decrease the numbers, but it is so much easier to collect unstructured data than it is to build Machine Learning models to actually do something with it, that, likely, that percentage will increase greatly. In just a few years, dark data could comprise 95-97% of the total percentage of data. If the trend continues, reasonably, Dark Data could comprise 99%+ of all data.

The number is neither good nor bad. Having 99.9999% of all data in the world be dark data means little. It just means there sits a lot of unused data. If anything, that number should signal there might be great opportunity to turn data into something no one else has.

Privacy with dark data

People are creating their technological footprint with data. This is fine when people don't mind if others know where they've been walking, but, sometimes there's other items—medical queries, Google searches, less savory sites, and even information you need to hide from a partner or relative—that individuals don't want others to see.

When it comes to data, security is very challenging.

Challenge 1: Anonymous data

People often think the first step to securing data is to anonymize the data. This means that all the data points can exist, but they'll remove any account numbers, names, email addresses, etc., from the person's data so it can't identify them directly. That method worked in elementary school, when a name was removed from an assignment someone turned in, and it could work for someone like Frank Abagnale as he put new names on checks and diplomas to parade around the country as an airline pilot, doctor, and lawyer.

But data in the technological world works differently. Any set of data points is an identifier. Five data points linked to one person, regardless of the name being given, are an identifier. If someone is known to wake in the morning, go for a walk, sneeze, yawn, kick a rock, go back to sleep, that is an impression of a single identity stamped upon the world.

Challenge 2: Intersections of data

There is so much data out there, that a person's name can exist among another set of data. Then, when these datasets have data points that intersect, the two sets are cross-referenced, it's possible to place an identity upon the anonymized data. Creating a Venn diagram of different data sources and finding which ones overlap is a simple option, and statistics invites more complex methods to deanonymize data.

There's the story of a legal case where an old lady was hit by a car and the car drove off. The woman was able to say the car was yellow (she didn't know the make), and the driver was a brown man with dark hair. That is not a lot to go off, but a few more dark data points add the time of day of the accident, and the location of the accident. From these four data points, in a town of about 120,000, the investigators were able to narrow down their search, from what seemed to be impossible odds, to having only a few suspects who could have hit the woman.

Similarly, from the technology world, the 7scientists research team presented a similar case at Defcon (see below for video clip). They purchased anonymous browsing data, which is easy to purchase, and showed they were able to identify the user from it based on just five data points.

The graph illustrates how many possible users the browsing data could belong to after each known data point was added.

Open source data privacy

Open Mined is an open-source research group working to make data more privacy-preserving. In a world with more and more dark data, their work benefits the general population to make data more anonymous and ensure that identities are made private even in the increasing amounts of available data.

Specifically, machine learning models are trained upon data. Machine learning models can both offer high value and work with sensitive data. While all data can be considered sensitive, and can be treated equally, legal conditions put medical records among the most sensitive.

Thus, training machine learning models on people's medical histories is very difficult in nature because of how sensitive the industry has treated the records in the past. Challenges include: not enough data, data being isolated to different locations for security purposes, having to jump through many extra hoops to meet "best safety practices" created by regulatory institutions.

The goal of Open Mined is two-fold: to create a framework where people get paid for their data and to truly anonymize data when passed through ML models. To that end, the open-source group currently offers three major software solutions:

- Encrypted Machine Learning as a Service
- Privacy Preserving Data Science Platform
- Federated Learning

Security isn't privacy

There is a lot of dark data out there, and there will likely be more. Security practices, as they are, do not preserve privacy with all the dark data points, but research groups are out there successfully improving the data landscape improving people's privacy, and advocating for people to get paid for the data they create.

Additional resources

For more on this topic, explore the BMC Machine Learning & Big Data Blog or browse these articles:

- Data is the New Electricity
- What Is a Data Pipeline?
- Data Management vs Data Governance: Main differences
- How to Create a Machine Learning Pipeline
- How to Apply Machine Learning to Cybersecurity
- The Role of Machine Learning in Datacenter Network Security