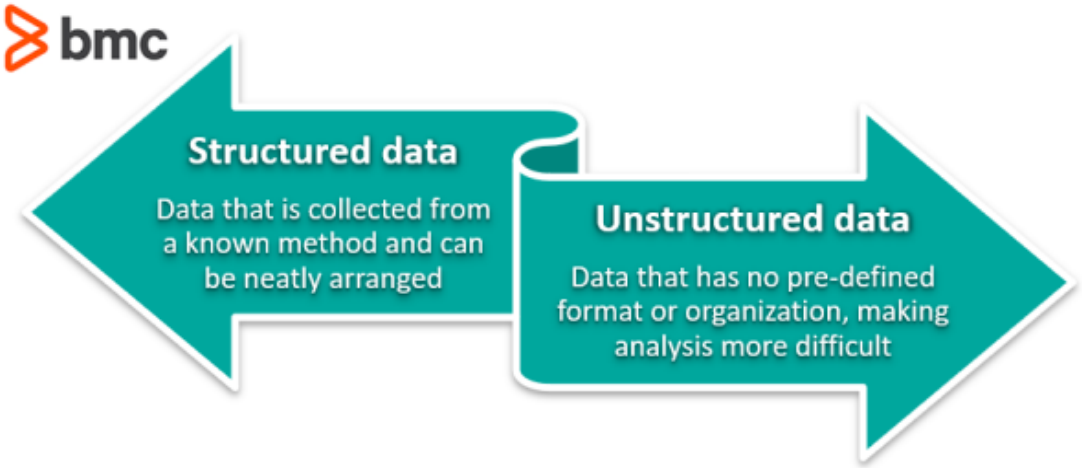
STRUCTURED VS UNSTRUCTURED DATA: A SHIFT IN PRIVACY



Data comes in many forms. The first step when looking at data is to ask, is this data structured or unstructured? Another way to make this distinction: Did this data arrive in response to a question, or has this data been collected with no intention?



What is structured data?

Structured data is data that can be neatly arranged. In the past, companies only had the tools and resources to explore structured data. Analyzing unstructured data was either impossible or extremely costly, so companies had to settle for structured data.

Collecting structured data

Structured data is hard to collect. By nature, the quantity of structured data that can be collected is limited.

Questionnaires are a common way of collecting structured data, though these have limitations. Any questionnaire requires a person to be attentive and willing to engage. There, a person's attention only lasts so long, so the questionnaire must have an appropriate length. Surveys have only 10-30 data points. Even a long survey of 100 data points does not compare to the thousands of data points that can be extracted from unstructured data.

If the structured data is not collected via questionnaire, the data is typed out, or coded, and people interact with it in a built system. A common task is to predict the churn rate of an organization's memberships or subscriptions. The company will collect data such as:

- Times per week the individual interacts with the organization
- Hours of the day the individual interacts
- Email info
- Age
- And so on...

Structured data is also collected every time you swipe a card or whenever you open an HTML page. It is collected by monitoring interactions with other users, such as reporting who talks to whom. This data is coded into the system and tracked at each instance of the occurrence. Each piece of structured data serves a purpose.

Large amounts of structured data could be collected by having large numbers of people submitting data, effectively using a distributed network to collect data. In this scenario, 1,000,000 people answer a survey of 10 questions to produce 10,000,000 data points. Or 1,000,000 people tag a few photos of sidewalks on a CAPTCHA. (But, this method of collection is available to unstructured data, too. The same method can be used just the same on unstructured data, so 1,000,000 people provide thousands of data points instead of 10.)

Given the limits of structured data, companies had an economy of space problem. Statisticians were hired to determine which questions, or data points, were most effective. In the beginning of the computer age, data storage could be costly. At any point in time, getting a person to sit down and answer a survey is costly. Statisticians were hired to maximize the effectiveness of a 10-question survey by determining which questions were the most important. They had to determine how related two data points might be, such as income and education. If the two were similar, asking both questions proved redundant, and the questionnaire could save space by asking one, and inferring the other.

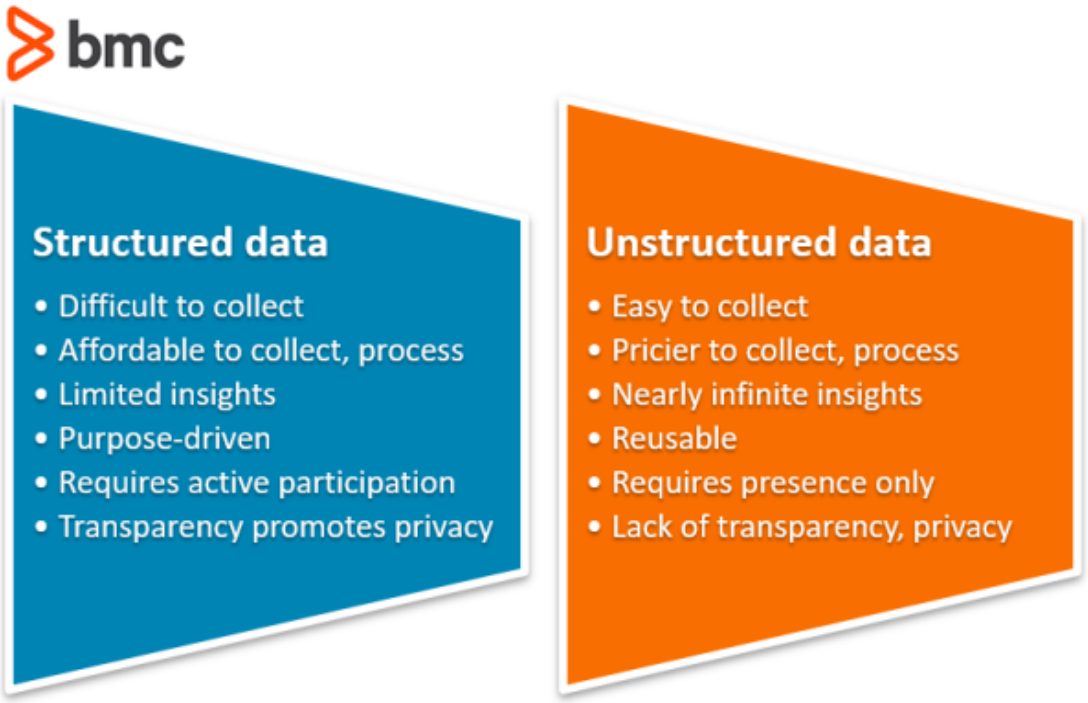
Transparency of structured data offers privacy

Though collecting structured data can be difficult and untimely, it has its advantages to communities. The efforts people had to go through to collect structured data naturally made the motives of both parties transparent. People on both sides could tell what information was wanted from them, and what information they were giving the receiving party responsibility to know. Boundaries were set.

The nature of structured data allowed any person to see what the data collector was asking of them. They could see and understand the kinds of things the analyst wanted to know.

If Walmart was conducting a structured survey that asked about a person's marital life, and how often they got upset with their children, and how the individual disciplined their child at their household, a person could decide those questions were not the responsibility of Walmart's to know, nor critique. The person could decide not to complete that form, and Walmart would not get that data.

The boundaries between two parties are more explicit within these structured forms.



What is unstructured data?

Compare our Walmart example—a choice on whether to respond to a survey—with this story. Target famously got burned for sending pregnancy advertisements to the household of a teenage girl. The father was outraged and called the Target manager for being irresponsible, only to discover that his daughter really was pregnant.

Target pieced together unstructured data, like what and when a person buys, going beyond the lines of its responsibilities. There is no way companies are no longer doing this—they have just learned to be more tactful.

Unstructured data, then, is any type of data that lacks pre-defined formats and organizations. A good way of thinking: structured data is collected from a known method or instance, and unstructured data is everything else. Unstructured data can comprise:

- Audio
- Video
- Text content, often unformatted for structured data collection, including text messages, emails, social media, business documents and news

Gathering unstructured data

The ability to decode unstructured data poses new options for collecting data. Data storage is cheap. Machine learning models can be created to extract information. Data collection isn't explicit. A company can say, "Talk. We'll get the information we want later." Raw audio, full video can be used as a data source.

There doesn't have to be a specially designed questionnaire. There just has to be a user's presence—a user's attention was too costly and time limited. Unstructured data doesn't require a person's immediate attention. Unstructured data can be collected on a user purely by their existence. And the user doesn't fully know what they are consenting to when they say they'll participate.

The unstructured data they supply today, can be used for entirely different purposes down the road. It is very easy for a company to say one thing and do another.

Unstructured data is reusable

One piece of unstructured data can be examined many, many ways. The person collecting the data does not have to define ahead of time what information is to be collected from it. This is great for lazy analysts, who can now start randomly collecting audio files with no purpose and figure out what they will use it for later.

Privacy concerns

The advent of unstructured data poses privacy concerns for the user. Data can be used to satisfy multiple intentions.

The face detector on the iPhone can provide 3D visualizations of a user's face. The incoming face data is unstructured data. Its first utilitarian use case—its stated purpose—is to securely lock and unlock a phone. But that same data can be used slightly differently. It can be used as a novelty to allow a user to create an avatar on their phone that displays their smile or frown in real-time.

Apple could also use the unstructured data in a different way. They could take that very same facial data and sell it to inform advertisers how their ads were being received by a user. That is a much more personal use case. The ability to analyze unstructured data has opened data to be used for multiple use cases.

With the ability to sift through unstructured data, a gym can forgo using card memberships to allow a person in their gym and start using facial recognition software. They collect unstructured video data. Instead of getting only time-entered and time-exited data, the gym can collect lots more information about its gym members: when they came and went, who they came with, what they were wearing, what mood they were in when they arrived and when they left, their athletic physique, and more. The gym member may have agreed for the gym to collect data on them "to provide better services".

New agreements for data collection in 2020

Unstructured data can be used for multiple use cases. Today, a person has to add a condition when they agree to someone using it. They go from saying, "Yes, you can collect this." to saying, "Yes, you

can collect this, and I permit you to use it for said use-case only. Using this data for any other reason is a violation of the boundaries between us."

Walmart and Amazon are responsible for providing low cost goods to a consumer. They are not responsible for being a marriage counselor. Though they might be able to make an educated statement about the state of a marriage given all the data they collect, any statement they may make on the subject goes outside the boundaries of their responsibilities.

Once upon a time, Facebook's only responsibility was to allow people to network and share information with one another. But in 2016, after the U.S. Presidential election, people felt like they were stung, and they permitted Facebook to have the responsibility to grade the content shared between users as good or bad. Facebook was granted the power to filter information, for better or worse, and now Facebook can read into unstructured data any way it pleases on the basis that it is trying to protect the people on its platform—2.6 billion monthly active users (on just one of its platforms. It owns WhatsApp and Instagram, too). That's roughly one-third of the global population, a larger jurisdiction than any single government.

Now, weirdly, Facebook's hand is actually forced to lie, because their duty is to read the unstructured data in all the ways they can in order to help protect their users. But its users will also tell Facebook it is not responsible to know the state of their mental health, or their relationships, or who is having an affair. So, to achieve its purpose and police bad, or harmful, content and to please the most immediate privacy concerns of the users, Facebook must say it is not collecting, or making decisions, on any personal information. Facebook has matured into a two-faced citizen, dancing about in the upper echelons of society taking cues from Jane Austen novels.

It is not fair to say the deceit is totally Facebook's fault. The error is also in the contradiction within the people's desires. They wish to have their cake and eat it too. Facebook must read into unstructured data in ways people probably don't want, if it is to achieve policing the posts of all its users.

People face a decision, either:

- Being uncomfortable knowing that Facebook has private and personal information about their lives.
- Knowing that Facebook lies to make them less uncomfortable.

The latter poses a threat to privacy and disguises a true concern.

Data collection best practices

It is most important for us to speak honestly about the data companies collect. While it may be uncomfortable, the focus should shift to speaking how much that private information is valued by each person, so companies, like Facebook, should be very, very careful in protecting the information and not letting it get into the hands of bad actors.

It would be better to know that Facebook uncomfortably collects personal information and that they do a lot of work to keep it safe, than to be lied to and find out only through a data breach that they were collecting uncomfortable information which is now in the hands of an unknown third-party. Pressure should be placed on privacy and security and not on unknowing.

Most of the world's data is in the form of unstructured data, and the data requires a different kind of

privacy agreement made between those providing the data, and those collecting the data.

Additional resources

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

- Big Data vs Analytics vs Data Science: What's The Difference?
- Big Data Security Issues in the Enterprise
- Enabling the Citizen Data Scientists
- Data Management vs Data Governance: Main differences
- Data Architecture vs Information Architecture: What's The Difference?
- 4 Reasons to Automate the Ingestion of Data