Structured vs. Unstructured Data: What's the Difference?

trifacta.com/blog/structured-unstructured-data/

Structured and unstructured data are both used extensively in data analysis but operate quite differently. Let's take a closer look at these two data formats to understand just how different structured data and unstructured data are.

Structured data vs unstructured data

Searchability is often used to differentiate between structured vs unstructured data. Structured data typically contains data types that are combined in a way to make them easy to search for in their data set. Unstructured data, on the other hand, makes a searching capability much more difficult.

Structured data is easily detectable via search because it is highly organized information. It uploads neatly into a relational database (think traditional row database structures) and lives in fixed fields. It's the data that most of us are used to working with in order to analyze largely quantitative problems—think "how many products have been sold this quarter" or "how many customers have subscribed to the monthly newsletter," for example.

Examples of structured data include:

- Dates
- Phone numbers
- ZIP codes
- Customer names

|	A	B	C	D	E	F	G	H	I
1	Sales Representative	Location	Region	Customer	Order Date	Item	Quantity	Price	Total Sale Amount
2	Sara Snyder	New York	East	Phyllis Johnston	2016-10-30	Things	1	17.83	17.83
3	Sara Snyder	New York	East	Kimberly Little	2016-05-23	Junk	3	12.42	37.26
4	Frances Warren	Massachusetts	East	Justin Dixon	2016-09-27	Widgets	4	53.35	213.4
5	Sara Snyder	Massachusetts	East	Shirley Rivera	2016-02-12	Junk	5	12.42	62.1
6	Diane Gonzalez	Oregon	West	Marilyn Franklin	2016-02-14	Things	8	17.83	142.64
7	Patrick Graham	Washington	West	Henry Sanders	2016-04-11	Widgets	4	53.35	213.4
8	Sara Snyder	Connecticut	East	Benjamin Phillips	2016-09-02	Junk	4	12.42	49.68
9	Frances Warren	New Jersey	East	Theresa Torres	2016-11-26	Junk	4	12.42	49.68
10	Patrick Graham	Oregon	West	Roger Bell	2016-07-13	Junk	10	12.42	124.2
11	Sara Snyder	New Jersey	East	Harold Matthews	2016-06-02	Junk	3	12.42	37.26
12	Frances Warren	New York	East	Roy Young	2016-06-02	Widgets	8	53.35	426.8
13	Sara Snyder	New York	East	Debra Allen	2016-02-20	Things	1	17.83	17.83
14	Randy Watson	Connecticut	East	Alan Dean	2016-06-07	Junk	7	12.42	86.94
15	Randy Watson	Massachusetts	East	Robin Matthews	2016-10-31	Stuff	5	16.32	81.6
16	Randy Watson	New York	East	Randy Burton	2016-03-13	Stuff	4	16.32	65.28
17	Patrick Graham	Washington	West	Terry Nguyen	2016-02-10	Widgets	10	53.35	533.5

So what is unstructured data? Unstructured data may have its own internal structure but it does not conform neatly into a spreadsheet or database. It includes everything outside the bounds of structured data. It may be generated from a human or a machine; it can be text or images.

While unruly in nature, it is also incredibly valuable—unstructured data has the potential to depict a complex web of information that offers strong clues about future outcomes. Unstructured data analysis is a crucial part of the data analytics process. Think of customer web chats, for example, a platform where customers commonly air out their complaints and troubleshooting questions. Analyzed as a whole, this web chat data can help guide companies on what to prioritize resolving or what aspect of the product is driving the most interest. Or social media data, which can signal customer buying trends before they even start searching for a product. If structured data has historically been a company's backbone, unstructured data is its competitive edge.

Examples of unstructured data include:

- Web logs
- Multimedia content
- Email
- Text files
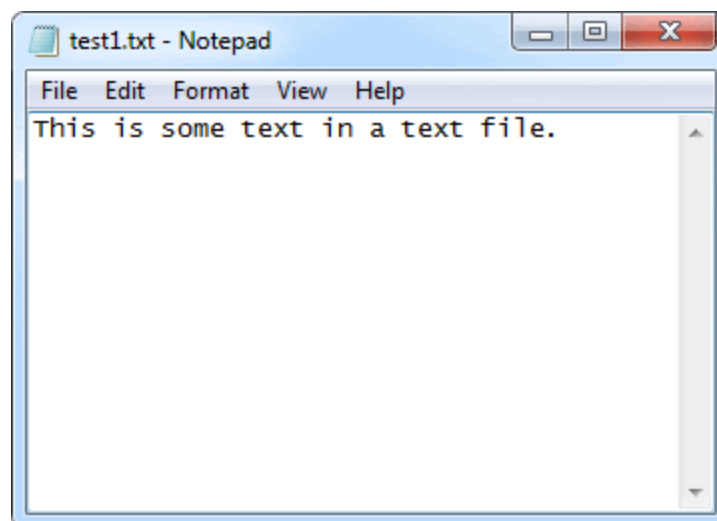
Semi-structured data: neither structured nor unstructured

Semi-structured data is often left out of the structured vs unstructured data conversation, but it's worth mentioning. At first glance, semi-structured data seems very messy, which might prompt you to ask, if this is semi-structured data, what is unstructured data?

```json
{
    "orders": [
        {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 AM",
            "trackingno": "TN0039291",
            "custid": "11045",
            "customer": [
                {
                    "custid": "11045",
                    "fname": "Sue",
                    "lname": "Hatfield",
                    "address": "1409 Silver Street",
                    "city": "Ashland",
                    "state": "NE",
                    "zip": "68003"
                }
            ]
        }
    ]
}
```

In reality, semi-structured data has characteristics of both structured and unstructured data—it doesn't conform to the structure associated with typical relational databases as structured data does, but it also has some structure in the form of semantic markup, which enforce hierarchies of records and fields within the data.

Examples of semi-structured data include:

- XML
- JSON

Storing and analyzing structured and unstructured data

Structured data is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type (e.g. alpha, numeric, date, currency), and as a result, is often restricted by character numbers or specific terminology. Analysts typically use simple or more complex VLOOKUP queries in Excel spreadsheets or Structured Query Language (SQL) to perform queries on structured data within relational databases.

On the other hand, developments in preparing and analyzing unstructured data are fairly recent. New data storage systems, such as data lakes, have allowed organizations to make great strides in capturing and storing unstructured data since it allows data to be stored in its raw format. However, the fundamental challenge of unstructured data sources is that they are difficult for non-technical business users and data analysts alike to unbox, understand, and prepare unstructured data for analytic use.

Though there's a lot of talk about the difficulty in managing today's volume of data—which is certainly a challenge—leveraging a reasonable amount of highly unstructured data can be equally trying.

The future of data

As data shifts to machine learning, big data, and the cloud, the future of data structures will also evolve. By 2025, 80% of all data will be unstructured, and many organizations have reached that ratio already. There is undoubtedly a huge opportunity ahead with unstructured data sources, yet it poses the greatest challenge to organizations in terms of being able to access and analyze that data.

What's more, organizations likely won't be just using unstructured data, but some combination of structured, unstructured or semi-structured data. Take the use case we mentioned earlier about the web chat data, for example. It's worthwhile to analyze customer web chat text, but the analysis would be made much more valuable should the company be able to tie that text data to structured customer information stored neatly in a CRM.

The challenge is accessing, preparing, and combining this data in order to make sense of it —especially among business analysts who weren't trained in computer engineering techniques.