Datasets

Datasets -- exactly what the name conveys -- a collection of data.

Datasets -- exactly what the name conveys -- a collection of data.

We didn't say "set" of data here since mathematically a set does not contain duplicate elements, but from now on we use "set" and "collection" interchangeably unless we need to be concerned about duplicate elements.

Data

Data

Mirriam-Webster:

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.

Data

Mirriam-Webster:

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.
2. Information in "digital form" [whatever that means] that can be transmitted or processed.

Data

Mirriam-Webster:

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.
2. Information in "digital form" [whatever that means] that can be transmitted or processed.

Is *"data"* singular or plural?

Data

Mirriam-Webster:

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.
2. Information in "digital form" [whatever that means] that can be transmitted or processed.

Is *"data"* singular or plural? Once more, from Mirriam-Webster

Data

Mirriam-Webster:

1. Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.
2. Information in "digital form" [whatever that means] that can be transmitted or processed.

Is *"data"* singular or plural? Once more, from Mirriam-Webster

Data leads a life of its own quite independent of *datum*, of which it was originally the plural. It occurs in two constructions: as a plural noun (like *earnings*), taking a plural verb and plural modifiers (such as *these, many, a few*) but not cardinal numbers, and serving as a referent for plural pronouns (such as *they, them*); and as an abstract mass noun (like *information*), taking a singular verb and singular modifiers (such as *this, much, little*), and being referred to by a singular pronoun (*it*). Both constructions are standard. The plural construction is more common in print, evidently because the house style of several publishers mandates it.

Digital Form

Digital Form

To be used on a computer all data must ultimately be represented as a collection of sequences of 0s and 1s, where each of these represents the status of an electronic component such as a transistor ("off" or "on"), voltage level ("low" or "high"), capacitor ("not charged" or "charged"), etc.

Digital Form

To be used on a computer all data must ultimately be represented as a collection of sequences of 0s and 1s, where each of these represents the status of an electronic component such as a transistor ("off" or "on"), voltage level ("low" or "high"), capacitor ("not charged" or "charged"), etc.

Over time various "standard formats" have emerged for representing various types of data.

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers:

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

- Typically "0" means non-negative and "1" means negative.
- To be useful for transmitting and processing the total number of bits being used to represent integers is fixed: 8-bits, 16-bits, 32-bits, 64-bit, 128-bits, etc.
- The leftmost bit is normally the sign bit.
- The remaining bits determine magnitude of the number. Here too there are different ways of determining this: signed magnitude, 1s complement, **2s complement**.

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

- Typically "0" means non-negative and "1" means negative.
- To be useful for transmitting and processing the total number of bits being used to represent integers is fixed: 8-bits, 16-bits, 32-bits, 64-bit, 128-bits, etc.
- The leftmost bit is normally the sign bit.
- The remaining bits determine magnitude of the number. Here too there are different ways of determining this: signed magnitude, 1s complement, **2s complement**.

Character Data

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

- Typically "0" means non-negative and "1" means negative.
- To be useful for transmitting and processing the total number of bits being used to represent integers is fixed: 8-bits, 16-bits, 32-bits, 64-bit, 128-bits, etc.
- The leftmost bit is normally the sign bit.
- The remaining bits determine magnitude of the number. Here too there are different ways of determining this: signed magnitude, 1s complement, **2s complement**.

Character Data - done using "codes"

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

- Typically "0" means non-negative and "1" means negative.
- To be useful for transmitting and processing the total number of bits being used to represent integers is fixed: 8-bits, 16-bits, 32-bits, 64-bit, 128-bits, etc.
- The leftmost bit is normally the sign bit.
- The remaining bits determine magnitude of the number. Here too there are different ways of determining this: signed magnitude, 1s complement, **2s complement**.

Character Data - done using "codes"

- ASCII code ("American Standard Code for Information Interchange")

Non-negative integers: binary numbers

$$27_{10} = 2\times10 + 7 = 2\times10^1 + 7\times10^0 = 1\times2^4 + 1\times2^3+0\times2^2+ 1\times2^1+ 1\times2^0 =11011_2$$

Non-negative and Negative Integers: Requires designating a bit as a "sign bit"

- Typically "0" means non-negative and "1" means negative.
- To be useful for transmitting and processing the total number of bits being used to represent integers is fixed: 8-bits, 16-bits, 32-bits, 64-bit, 128-bits, etc.
- The leftmost bit is normally the sign bit.
- The remaining bits determine magnitude of the number. Here too there are different ways of determining this: signed magnitude, 1s complement, **2s complement**.

Character Data - done using "codes"

- ASCII code ("American Standard Code for Information Interchange")
- Unicode

Other examples of data representation

Other examples of data representation

Images:

Other examples of data representation

Images: tiff, bmp, jpeg, etc.

Other examples of data representation

Images: tiff, bmp, jpeg, etc.

Audio Files

Other examples of data representation

Images: tiff, bmp, jpeg, etc.

Audio Files: WAV, MP3, AAC, FLAC, etc.

Other examples of data representation

Images: tiff, bmp, jpeg, etc.

Audio Files: WAV, MP3, AAC, FLAC, etc.

Video Files:

Other examples of data representation

Images: tiff, bmp, jpeg, etc.

Audio Files: WAV, MP3, AAC, FLAC, etc.

Video Files: MP4, MOV, WMV, etc.

Datasets may also be classified according to their overall structure

Datasets may also be classified according to their overall structure

- Structured datasets

Datasets may also be classified according to their overall structure

- Structured datasets
- Semi-structured dataset

Datasets may also be classified according to their overall structure

- Structured datasets
- Semi-structured datasets
- Unstructured datasets

Structured datasets

Structured datasets

A structured dataset, as its name implies, is one whose data is required to conform to a strict, predetermined structure known as a *schema*.

Structured datasets

A structured dataset, as its name implies, is one whose data is required to conform to a strict, predetermined structure known as a *schema*.

A general characteristic of a structured dataset is that overall it is organized as a table with each cell of the table representing one datum of the dataset.

Structured datasets

A structured dataset, as its name implies, is one whose data is required to conform to a strict, predetermined structure known as a *schema*.

A general characteristic of a structured dataset is that overall it is organized as a table with each cell of the table representing one datum of the dataset.

Spreadsheets and relational databases are the most common examples of structured datasets.

Structured datasets

A structured dataset, as its name implies, is one whose data is required to conform to a strict, predetermined structure known as a *schema*.

A general characteristic of a structured dataset is that overall it is organized as a table with each cell of the table representing one datum of the dataset.

Spreadsheets and relational databases are the most common examples of structured datasets.

Example of a structured dataset: Given in class

Semi-structured datasets

Semi-structured datasets

A semi-structured datasets is one whose data has some structure to its organization, but this structure is not as rigid, consistent, or complete as those of a structured data type.

Semi-structured datasets

A semi-structured datasets is one whose data has some structure to its organization, but this structure is not as rigid, consistent, or complete as those of a structured data type.

Rather than conform to a structure that is specified independent of the data itself, in a semi-structured dataset the data does not reside in fixed fields or records, but does contain elements that can separate the data into various hierarchies.

Semi-structured datasets

A semi-structured datasets is one whose data has some structure to its organization, but this structure is not as rigid, consistent, or complete as those of a structured data type.

Rather than conform to a structure that is specified independent of the data itself, in a semi-structured dataset the data does not reside in fixed fields or records, but does contain elements that can separate the data into various hierarchies.

Semi-structured datasets gained importance as a means for data exchange between disparate and independently controlled Web sites.

Semi-structured datasets

A semi-structured datasets is one whose data has some structure to its organization, but this structure is not as rigid, consistent, or complete as those of a structured data type.

Rather than conform to a structure that is specified independent of the data itself, in a semi-structured dataset the data does not reside in fixed fields or records, but does contain elements that can separate the data into various hierarchies.

Semi-structured datasets gained importance as a means for data exchange between disparate and independently controlled Web sites.

Example of a semi-structured dataset: Given in class

Unstructured datasets

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.
- Email files.

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.
- Email files.
- An audio file.

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.
- Email files.
- An audio file.
- An image file.

Unstructured datasets

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.
- Email files.
- An audio file.
- An image file.
- A file containing sensor data.

To illustrate one of

An unstructured dataset is exactly what its name implies -- it is a dataset that is not organized in a predefined way. The data is stored in its native format.

Examples of unstructured data:

- The content of a text or word processing file.
- Email files.
- An audio file.
- An image file.
- A file containing sensor data.