Extracting Information from Data

DAT-2.A.1

Information is the collection of facts and patterns extracted from data.

DAT-2.A.2

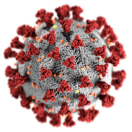
Data provide opportunities for identifying trends, making connections, and addressing problems.

DAT-2.A.3

Digitally processed data may show correlation between variables. A correlation found in data does not necessarily indicate that a causal relationship exists. Additional research is needed to understand the exact nature of the relationship.

DAT-2.A.4

Often, a single source does not contain the data needed to draw a conclusion. It may be necessary to combine data from a variety of sources to formulate a conclusion.



Contact tracing is an important tool for addressing the problem of how to limit the spread of deadly illnesses like covid-19.

Google tracks yearly trends by examining what people google: https://trends.google.com/trends/?geo=US

• • •

DAT-2.B.1

Metadata are data about data. For example, the piece of *data* may be an image, while the *metadata* may include the date of creation or the file size of the image.

DAT-2.B.2

Changes and deletions made to metadata do not change the primary data.

DAT-2.B.3

Metadata are used for finding, organizing, and managing information.

DAT-2.B.4

Metadata can increase the effective use of data or data sets by providing additional information.

DAT-2.B.5

Metadata allow data to be structured and organized.

Piece of Data (image)

Additional metadata
(posted on website where I downloaded picture)

PHOTO INFORMATION

Hotrod Die-cast Model on Board

Uploaded at September 15, 2018

Lens	**50.0mm ƒ/6.3 ISO 200**
Size	**6.19 MB**
Resolution	**5460px x 3116px**
Camera	**ILCE-6000**
Software	**PhotoScape**
Taken at	**July 15, 2018 5:37 am**
Aspect Ratio	**1365:779**

Metadata of Image
(included as part of image file)

▼ General:

Kind:	JPEG image
Size:	354,453 bytes (356 KB on disk)
Where:	Macintosh HD ▸ Users ▸ kevinmadden ▸ Downloads
Created:	February 3, 2021 at 11:01 AM
Modified:	February 3, 2021 at 11:01 AM

☐ Stationery pad
☐ Locked

▼ More Info:

Where from:	https://www.pexels.com/ https://images.pexels.com/photos/1422673/pexels-photo-1422673.jpeg?auto=compress&cs=tinysrgb&dpr=2&h=750&w=1260
Dimensions:	2520×1438
Color space:	RGB
Color profile:	sRGB IEC61966-2.1
Alpha channel:	No

DAT-2.C.1

The ability to process data depends on the capabilities of the users and their tools.

DAT-2.C.2

Data sets pose challenges regardless of size, such as:

- the need to clean data
- incomplete data
- invalid data
- the need to combine data sources

DAT-2.C.3

Depending on how data were collected, they may not be uniform. For example, if users enter data into an open field, the way they choose to abbreviate, spell, or capitalize something may vary from user to user.

DAT-2.C.4

Cleaning data is a process that makes the data uniform without changing their meaning (e.g., replacing all equivalent abbreviations, spellings, and capitalizations with the same word).

DAT-2.C.5

Problems of bias are often created by the type or source of data being collected. Bias is not eliminated by simply collecting more data.

DAT-2.C.6

The size of a data set affects the amount of information that can be extracted from it.

DAT-2.C.7

Large data sets are difficult to process using a single computer and may require parallel systems.

DAT-2.C.8

Scalability of systems is an important consideration when working with data sets, as the computational capacity of a system affects how data sets can be processed and stored.

• • •

Suppose you have a massive database. Part of the database lists first name/last name along with the person's cell phone number.

A person needs a single list of all phone numbers in the same format. Unfortunately, there is no consistent notation. Some numbers include dashes (### - ### - ####), and others do not (##########).

A person enters a cell # into the database but forgets to list the person's last name.

A person only types out 12 digits for a US phone number, which is impossible.

Oftentimes, data needs cleaned, completed, and/or validated prior to combining two or more data source. For example, you would want all phone numbers to either have dashes or not have dashes when combining data sources

Note: A first and last name generally isn't a unique identifier. This is generally why people are given id numbers when placed in a system.

Assigning ID #s would make it easier to tell the difference between multiple people with the same name.
1234 Kevin Madden ##########
9876 Kevin Madden ##########

1234 2020-02-03 French

Data may be considered incomplete if the data collected was intended to be representative of the population but due to the source you used to obtain the data, you end up with a sample of people that is not representative of the whole (for example, if you were collecting last names and phone #'s by making spam calls to people and tricking them into giving you such info, then there would be a high level of bias in your data because not everyone would fall for such a trick).

Multiple covid-19 tests were given to see if there is a widespread outbreak of covid-19 in the school. Upon discovering that 3 out of the 3 students tested were positive for covid-19, the conclusion is that every student in the school building has covid-19. Duh! 😎

This was the #1 problem we were always grappling with at my old programming job because our business model was largely based on processing humongous batches of data.

It is hard to predict what issues will arise if you are dealing with large amounts of data. Just because you can process 1 byte (8 bits) of data successfully does not mean you will be able to process 1 terabyte of data successfully.

The online self-certification application that the district requires people to fill out crashed on the first day it was used on a mass scale.