# Data definitions and formats

A short lecture by Akshay Mehra and Marine Denolle

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

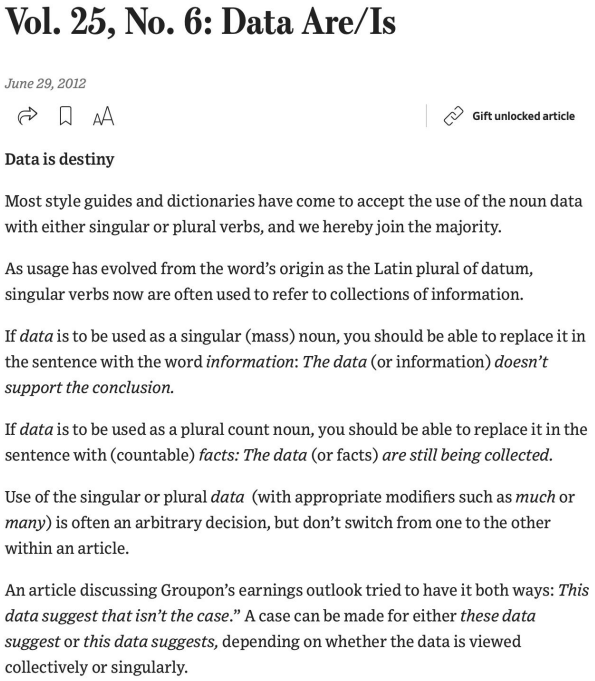
# What sort of data do (or might) you work with?

Another way to ask this question: what is the ***modality*** of your data?

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# An aside: The term “data” is the plural form of “datum”

The phrases “the data show” and “the data shows” both are valid.



## Vol. 25, No. 6: Data Are/Is

*June 29, 2012*

Gift unlocked article

**Data is destiny**

Most style guides and dictionaries have come to accept the use of the noun data with either singular or plural verbs, and we hereby join the majority.

As usage has evolved from the word’s origin as the Latin plural of datum, singular verbs now are often used to refer to collections of information.

If *data* is to be used as a singular (mass) noun, you should be able to replace it in the sentence with the word *information*: *The data* (or information) *doesn’t support the conclusion.*

If *data* is to be used as a plural count noun, you should be able to replace it in the sentence with (countable) *facts: The data* (or facts) *are still being collected.*

Use of the singular or plural *data* (with appropriate modifiers such as *much* or *many*) is often an arbitrary decision, but don’t switch from one to the other within an article.

An article discussing Groupon’s earnings outlook tried to have it both ways: *This data suggest that isn’t the case.*” A case can be made for either *these data suggest* or *this data suggests,* depending on whether the data is viewed collectively or singularly.

https://www.wsj.com/articles/vol-25-no-6-data-are-is-01557859235

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# “Big data”

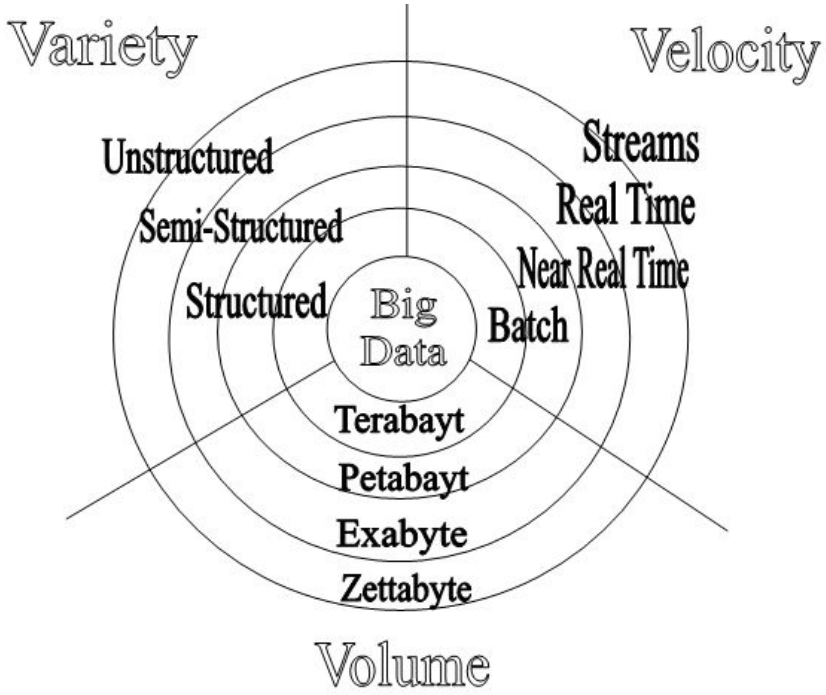
“The problem today is that the ever-increasing deluge of information—terabytes to petabytes to exabytes—threatens to swamp us in a gusher of unfiltered, unstructured, unprocessed, and seemingly unmanageable information.”

Big data have three characteristics: **variety**, **volume**, and **velocity**.

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

Gerhardt et al., 2012

# The characteristics of “big data”



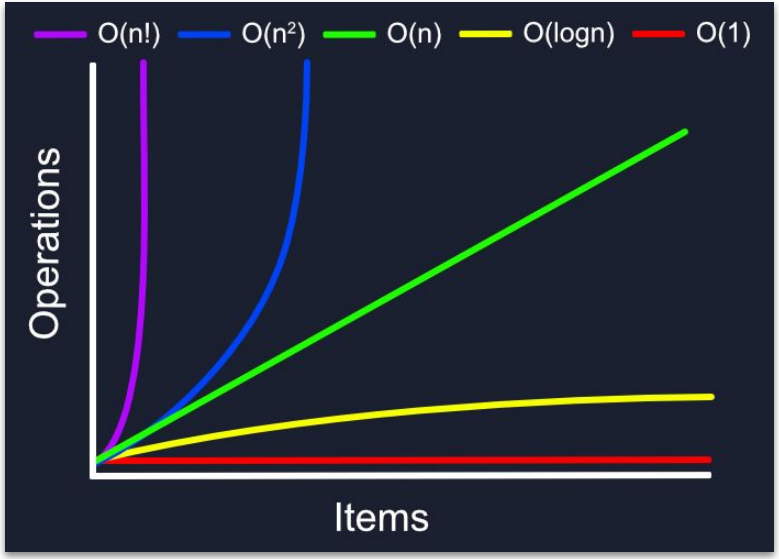

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

Sagiroglu and Sinanc, 2013

# Working with “big data”

One consequence of big data analyses is that computation, while easily described (e.g., “big O notation”), may be non-intuitive.

Seemingly straightforward tasks can take longer or use more resources than expected. For example, just moving data around can take frustratingly long.




Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

https://www.digitalocean.com/community/tutorials/js-big-o-notation

# Structured vs. unstructured data

**Structured data**: Organized in a specific format. Think: position information, geochemical measurements, morphological characteristics.

**Unstructured data**: Do not follow a specific format. Think: papers, photographs, audio, video.

Referred to as “quantitative” and “qualitative”, respectively.

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# Data dimensions

The dimensionality of data refers to the number of properties or categories within a dataset.

Each property is associated with multiple measurements.

Consider: Satellite images can be considered two-dimensional, three-dimensional, or even four-dimensional. How?

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# Data “types”

Each value in your dataset has a *data type*.

In the context of programming, “type” refers to an attribute which defines various properties of the data, such as range of values and valid operations.

Integers, floats, and strings are examples of data types.

*Categorical data* (qualitative or  nominal data) is a type of data that represents categories or labels and cannot be measured on a numerical scale.

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# Data characteristics in geosciences

Tabular Data

Time series data

Geospatial data

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# Implicit vs. explicit types

Data types may be implicit (i.e., the compiler figures them out, *dynamically*, at runtime) or can be explicitly declared (by the programmer).

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# What data types are you working with?

Two points to consider:

1. When dealing with large amounts of data (or many or complex calculations), the choice of data type may have very big implications for performance.
2. An incorrect data type (e.g., using integers instead of floats) can lead to calculation errors.

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu

# Data formats

Ultimately, data must be stored in a *file* with some sort of pre-defined *format*.

File format specifications tell you how the data are structured within a format.

Without proper documentation, formats easily can become obsolete!

Marine Denolle mdenolle@uw.edu
Akshay Mehra akmehra@uw.edu