Reading and Storing Data

VIIRS I-Band 375 m Active Fire Data

In [2]:
```
# Display the first 5 records from our dataset
df.head()
```

Out[2]:

	latitude	longitude	bright_ti4	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_ti5	frp	daynight
0	0.05836	29.59085	295.64	0.38	0.59	2023-07-12	3	N	VIIRS	n	2.0NRT	275.15	0.83	N
1	0.48765	31.50760	296.73	0.51	0.66	2023-07-12	3	N	VIIRS	n	2.0NRT	275.15	0.56	N
2	2.15227	13.94524	305.26	0.51	0.49	2023-07-12	3	N	VIIRS	n	2.0NRT	287.94	1.08	N
3	2.15681	13.94618	319.05	0.51	0.49	2023-07-12	3	N	VIIRS	n	2.0NRT	288.77	1.81	N
4	2.15754	13.94131	301.13	0.51	0.50	2023-07-12	3	N	VIIRS	n	2.0NRT	288.17	1.81	N

In [3]:
```
# Display the last 5 records from our dataset
df.tail()
```

Out[3]:

	latitude	longitude	bright_ti4	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_ti5	frp	daynight
74600	61.42408	-110.40578	350.48	0.4	0.4	2023-07-12	1950	N	VIIRS	n	2.0URT	309.39	16.01	D
74601	61.42510	-110.39867	336.03	0.4	0.4	2023-07-12	1950	N	VIIRS	l	2.0URT	308.08	32.98	D
74602	61.42733	-110.40780	328.53	0.4	0.4	2023-07-12	1950	N	VIIRS	n	2.0URT	298.15	16.01	D
74603	61.42834	-110.40069	338.45	0.4	0.4	2023-07-12	1950	N	VIIRS	n	2.0URT	302.81	32.98	D
74604	61.42936	-110.39356	339.52	0.4	0.4	2023-07-12	1950	N	VIIRS	n	2.0URT	306.58	32.98	D

For data of sufficiently moderate sizes, the whole information contained in the table can be stored as a matrix object in a coding platform such as Python.

Or the data can be directly treated as a spreadsheet and be manipulated in data analysis and visualization software such as Excel.

https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/viirs-i-band-375-m-active-fire-data

Data Limits

Domain: represents the spread or extent of the independent variable (control variable) over which the quantity being measured varies. It is usually given as the maximum value of the independent variable minus the minimum value, or it can be **an interval window**.

Range: represents the spread or extent over which the dependent variable (i.e., the quantity being measured) can take on values. It is usually given as the maximum value of the dependent variable minus the minimum value, or it can be **an interval window**. Note that the lower limit on sufficiently small (or large) quantities can be set by the noise level (or capabilities) of the measuring instrument.

Reading and Storing Data (Formats)

However, the table information has to be in a readable file for whatever data analysis platform to be able to interpret it.

Common file formats are CSV, XLSX (Open XML), TXT (ASCII files or plain text files), JSON.

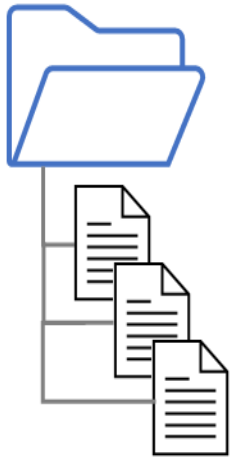
We will be mostly working with textual and tabular data. But note that data can be packed in multiple forms so other file formats can come into play:

Textual data: XML, TXT, HTML, PDF/A (Archival PDF)
Tabular data (including spreadsheets): CSV
Databases: XML, CSV, JSON
Images: TIFF, PNG, JPEG
Audio: WAV, MP3

CSV (Comma-separated values)
XML (Extensible Markup Language)
JSON (JavaScript Object Notation)

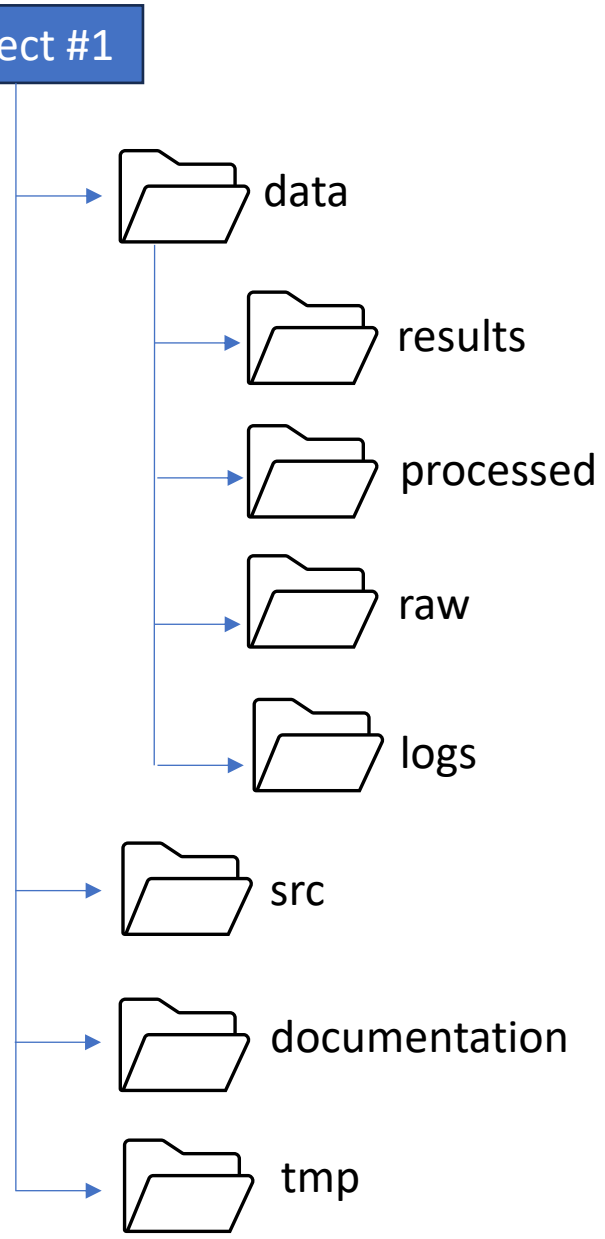
Reading and Storing Data

Data (processed or unprocessed) can also be stored in multiple files and folders from which data manager tools or scripting can be used for reading/importing the desired information prior to analysis.



There is not necessarily a rule for folder structure. The hierarchy of the folders should be consistent and logical. Go from a general, high-level folder (starting with a single folder for the project, using its name or acronym) to more specific lower-level folders. The structure should not be too deep or too shallow. Depending on the size of the project, this could mean 3-4 levels, but it could be more or less for small or very large projects.

On the suggested structure on the side, there may be multiple data folders, e.g., data → sample_01, sample_02, sample_03, etc.



Data, Dataset, Database

- **Data** are observations or measurements (unprocessed or processed) represented as text, numbers, or multimedia.

- A **dataset** is a structured collection of data generally associated with a unique body of work. A dataset can be as simple as a spreadsheet containing rows and columns of data, or it can be more complex, involving multiple tables or files.

- A **database** is an organized collection of data stored as multiple datasets. Database is a broader system for storing, managing, and retrieving data efficiently, often catering to larger volumes of data with various functionalities for data manipulation and retrieval.

Based Source: U.S. Geological Survey

Note that some sources may use dataset and database intermittently.

Popular Databases

MNIST (Modified National Institute of Standards and Technology) database -- Images of handwritten digits commonly used to test classification, clustering, and image processing algorithms.

European Data Portal -- a single access point for open data published by EU Institutions, national portals of EU Member states and non-member states, as well as international organisations of predominantly European scope. https://data.europa.eu/en

National Geologic Map Database -- an archive of geoscience maps (including geology maps), reports, and stratigraphic information for the United States.
https://www.usgs.gov/faqs/what-national-geologic-map-database

Global Monitoring Laboratory (GML) -- acquire, evaluate, and make available accurate, long-term records of atmospheric gases, aerosol particles, clouds, and surface radiation in a manner that allows the causes and consequences of change to be understood. https://gml.noaa.gov/

Popular Databases

NASA database -- https://www.nasa.gov/centers/hq/library/find/databases
https://data.nasa.gov/

Molecular database -- database of experimentally determined molecular structures and other relevant data for molecular modelling.
https://www.ncbi.nlm.nih.gov/Structure/MMDB/docs/mmdb_help.html
https://next-gen.materialsproject.org/
https://www.rcsb.org/
https://c2db.fysik.dtu.dk/

Open Government Portal, Government of Canada -- Information resources and data sets published by government institutions. https://search.open.canada.ca/opendata/

Links to open datasets, government data and information portals:
https://science.gc.ca/site/science/en/open-science-helping-make-science-accessible-all-canadians/datasets-and-portals

Popular Databases

Other databases cited in class:

CERN

Allen Brain Map

Canadian Astronomy Data Centre
https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/

University of Calgary's Rothney Astrophysical Observatory (RAO)
https://science.ucalgary.ca/rothney-observatory

Contacts: Dr. Phil Langill and Jennifer Howse
rao@phas.ucalgary.ca

The Human Brain Project
https://www.humanbrainproject.eu/en/

OpenNeuro
https://openneuro.org/

Data Errors

Errors can enter data through experimental design, measurement and collection techniques, assumptions about the data, discretization and computational or analysis procedures. Therefore, it is important to estimate the uncertainty contained in the measurements and their influence on the interpretations.

Instrument Errors

Errors associated with the capabilities of the recording device.

Experimental/Observational Errors

Errors associated with the experimental design, sampling or observational methods.

Analysis of uncertainty (Measurement System Analysis – MSA)

The process of assessing the uncertainty associated with a measurement. We usually believe that there is a particular **exact or actual** value when we measure something. This "perfect" value can never be known with absolute certainty, but **we can estimate ranges** of values that most likely will include the correct value.

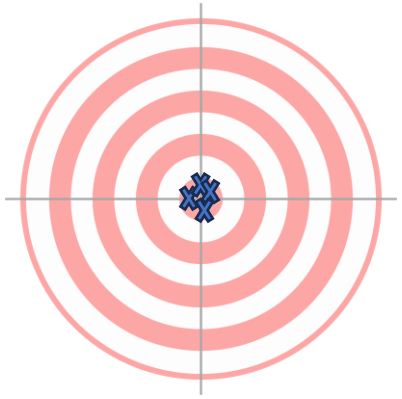
Precision: refers to a measurement system's ability on average to reproduce a measured value over and over (reproducibility of the measurement system).

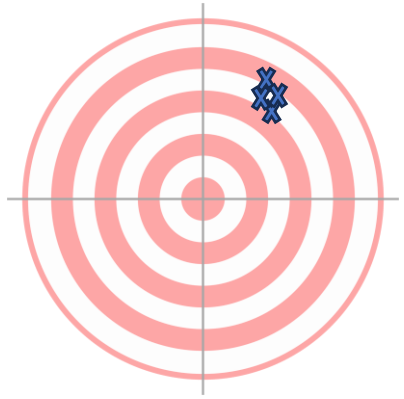
Accuracy: refers to a measurement system's ability to produce an average value that matches the actual accepted value (closeness to actual value).

Process Gage R&R (repeatability and reproducibility)

Analysis of uncertainty (Measurement System Analysis – MSA)

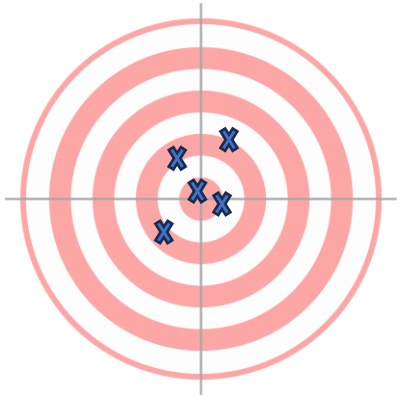
High-P
High-A

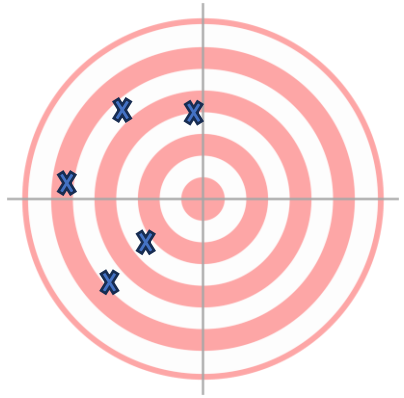




High-P
Low-A

Low-P
High-A





Low-P
Low-A

Analysis of uncertainty (Measurement System Analysis – MSA)

Exact or actual values may come from:

- An abstract value that is measured based on universal values and theoretical principles and can be taken as an "ideal" value to estimate the accuracy of our result.
- From various measurements being performed across the globe, establishing hence a "standard" value for the *population*.
- Manufacturer's precision requirements or instrumentation manuals.

Types of Errors

Random errors: statistical deviations (in any direction) in the measured data due to measuring device precision limitations and uncontrollable experimental factors; e.g., undesired equipment movements, humidity changes, fluctuation in temperatures, etc.

Suppose you measure **the mass of a metal disk 4 times** using the same balance: 5.62 g, 5.58 g, 5.61 g, 5.57 g. Random errors can be reduced by taking more data and averaging over a significantly large number of observations.

Systematic errors: an error that occurs continuously in just one direction every time the experiment is conducted (measured value increasingly being shifted too high or too low from true value); e.g., measurements from uncalibrated instruments, imperfect modelling, experiments operated manually by different people (different readings).

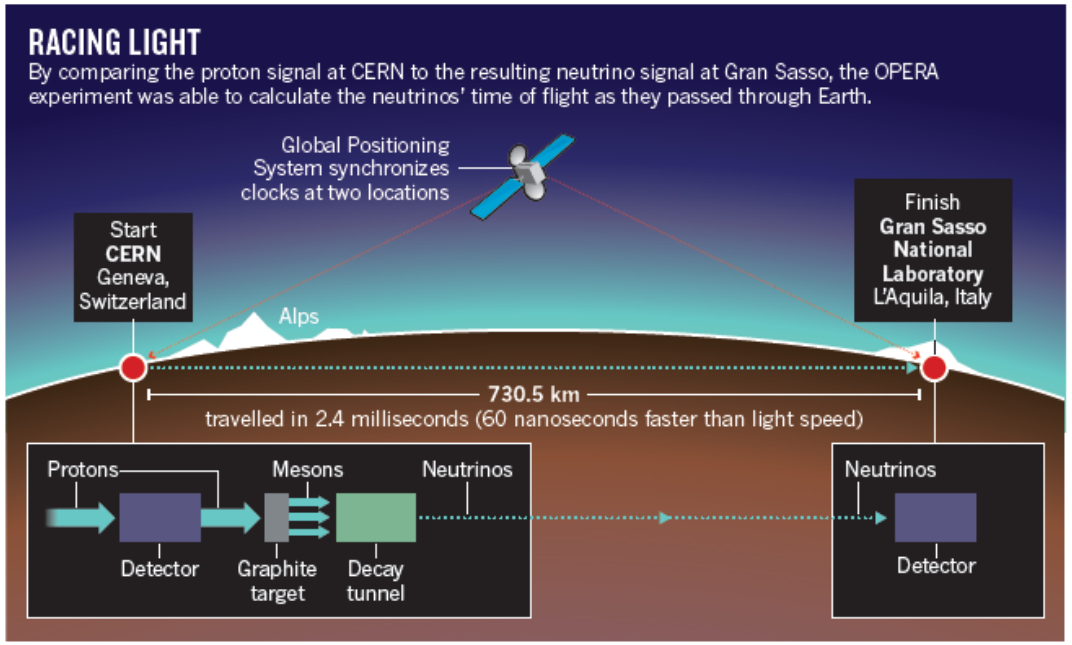
Types of Errors

While *random errors (statistical uncertainty)* are Gaussian and scale $\sim 1/\sqrt{N}$ with N being the number of samples, *systematic errors* do not necessarily follow this rule.

Random errors improve with **more data**.

Samuel Reich, E. Speedy neutrinos challenge physicists. Nature (2011). https://doi.org/10.1038/477520a

Systematic errors improve with **more understanding**.
(cleaning/treating data based on well-founded criteria and arguments can also help!)



PARTICLE PHYSICS

Speedy neutrinos challenge physicists

Experiment under scrutiny as teams prepare to test claim that particles can beat light speed.

BY EUGENIE SAMUEL REICH

The joke begins with the barman saying: "I'm sorry, we don't serve neutrinos." Then the punch line: a neutrino walks into a bar.

Such causality-bending humour has been rife on the Internet in the past week, following the news that an experiment at the Gran Sasso National Laboratory near L'Aquila, Italy, has apparently clocked neutrinos exceeding the speed of light as they travelled 730 kilometres from their source at CERN, Europe's particle-physics laboratory near Geneva, Switzerland.

The finding by the OPERA (Oscillation Project with Emulsion-tracking Apparatus) collaboration, released on 22 September, has the media abuzz with talk of a century's worth of physics upended, starting with Albert Einstein's special theory of relativity. This sets the velocity of light as the inviolable and unattainable limit for matter in motion, and links it to deeper aspects of reality, such as causality.

Physicists, for the most part, suspect that an unknown systematic error lies behind OPERA's startling result. But nothing obvious has emerged, and many see the experiment as a tour de force because of its high precision. "It is quite a complicated experiment but they did a professional job," says Rob Plunkett, co-spokesman for the MINOS (Main Injector Neutrino Oscillation Search) experiment at Fermilab in Batavia, Illinois, which is likely to investigate the claim.

OPERA was switched on in 2006 to study the peculiar ability of the fleeting, nearly massless