- **Software and computing tools**

ANACONDA
Powered by Continuum Analytics®

Coding: Python Anaconda package to run codes locally:

https://www.anaconda.com/distribution/

Python 3.X version is recommended.

Document production: MS Word, Pages, OpenOffice, Latex (e.g., on overleaf.com)

Reference management: EndNote (may not be free), Zotero, Mendeley

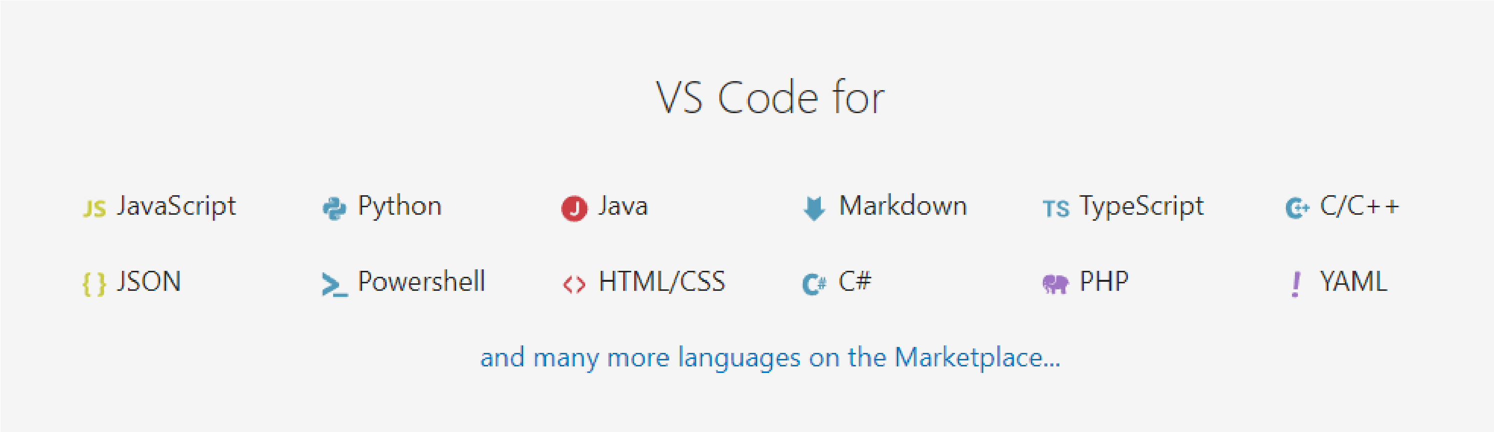
Other supporting software: Excel, Mathematica, MatLab

Version control software: to help keep track of source code modifications; they track changes to files, which allows mistakes to be easily undone. They also provide a way to synchronize the contents of multiple computers and serve as backup tools. Examples: svn (Subversion) and Git (https://git-scm.com/).

• Software and computing tools

Another coding platform that is very popular at the moment is Visual Studio Code: https://code.visualstudio.com/

It has built-in Git and other features, but I never used it myself.



VS Code for

JavaScript	Python	Java	Markdown	TypeScript	C/C++
JSON	Powershell	HTML/CSS	C#	PHP	YAML

and many more languages on the Marketplace...

Python programming

We can use the Jupyter Notebook platform at https://ucalgary.syzygy.ca/

NO INSTALLATION REQUIRED! But note that it can crash during busy usage times!

Alternatively, one can install Python Anaconda package to run codes locally in the embedded Jupyter Notebook from

https://www.anaconda.com/distribution/

Python 3.X version is recommended.

Python programming

This term we will also be testing a new Jupyter platform within the UCalgary domain at http://talc.ucalgary.ca/. **But note that this is still in testing mode!** Some of the libraries may not be yet set up and further testing is still needed.

NO INSTALLATION REQUIRED! But note that it can crash during busy usage times!

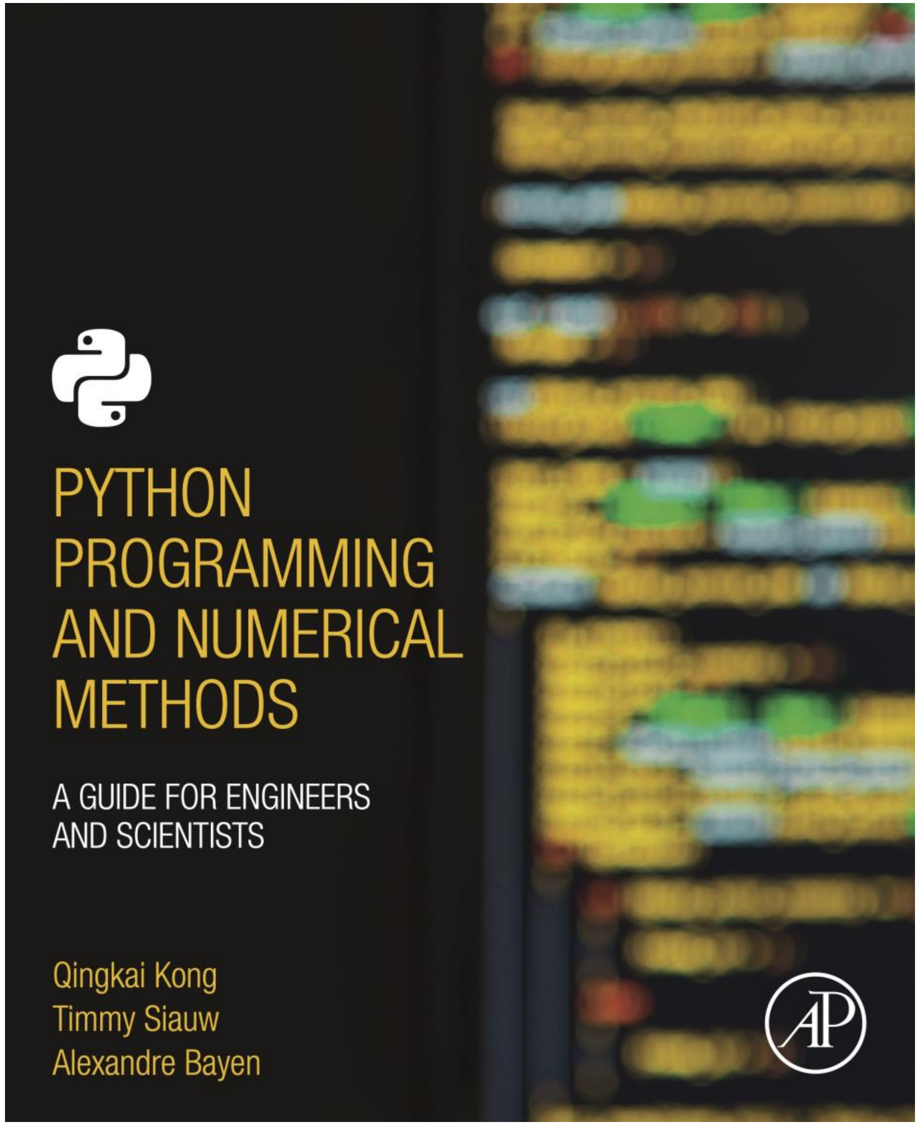
More information on https://rcs.ucalgary.ca/TALC_Cluster

ONLY PYTHON NOTEBOOK CODES (i.e., scripts written in .ipynb) WILL BE ACCEPTED!

Only standard Python libraries will be accepted in the homeworks:
`Python/iPython built-ins, Numpy, Scipy, Matplotlib, Plotly, Pandas, Networkx, Seaborn, Scikit-learn, Statsmodels, PyTorch, TensorFlow, Keras, Python Image Library, OpenCV.`

Other programming languages may be considered but <u>only for the final project</u>.

https://pythonnumericalmethods.berkeley.edu/notebooks/Index.html



PYTHON PROGRAMMING AND NUMERICAL METHODS

A GUIDE FOR ENGINEERS AND SCIENTISTS

Qingkai Kong
Timmy Siauw
Alexandre Bayen

- **Respect and inclusion in PHYS 605**

The university seeks to create and maintain a positive and productive learning, working and living environment; an environment in which there is:

✓ respect for the dignity of all
✓ fair treatment of individuals
✓ respect for academic freedom
✓ respect for university resources and the property of individuals

Btw, I am also the PHAS Associate Head EDI!

https://www.ucalgary.ca/pubs/calendar/current/k.html

- **Academic misconduct**

Academic integrity is the foundation of the development and acquisition of knowledge and is based on values of honesty, trust, responsibility, and respect.

See section 12.d of the course outline.

(See Section 12.i of the course outline)

All course materials (including those posted on the course D2L site, a course website, or used in any teaching activity such as (but not limited to) examinations, quizzes, assignments, laboratory manuals, lecture slides or lecture materials and other course notes) are for the sole use of students registered in this course and must not be redistributed.

DO NOT SHARE ANY COURSE MATERIALS with anyone outside the class and DO NOT post any course material in forums, public websites, and public repositories on the internet. Moreover, DO NOT share any course materials in so-called online learning (notes-sharing) platforms such as Chegg, Course Hero, or OneClass.

- **Recommendations**

- Attend class and check our course D2L constantly!
- Plan the elaboration and writing of your assignments: they may take longer than one thinks!
- There are tons of information online about scientific coding and programming. Spend some time with off-class tutorials and manuals especially those dedicated to data science programming.

Data Analysis (for scientific research)

datum

1. A single piece of information.

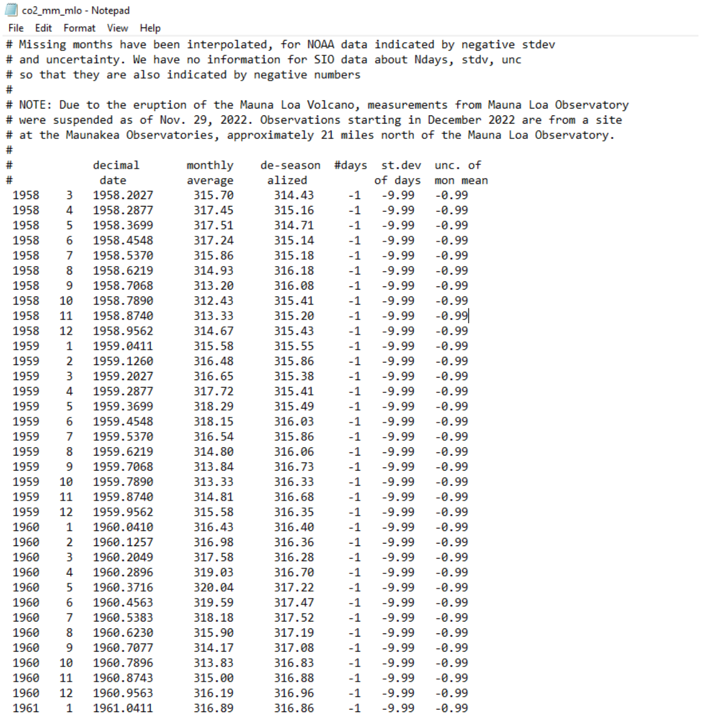
data

1. Plural form of datum: pieces of information.
2. (uncountable, collectively) information.

-logy

1. a branch of learning; a study of a particular subject.

Examples: biology, geology, genealogy



```
# Missing months have been interpolated, for NOAA data indicated by negative stdev
# and uncertainty. We have no information for SIO data about Ndays, stdv, unc
# so that they are also indicated by negative numbers
#
# NOTE: Due to the eruption of the Mauna Loa Volcano, measurements from Mauna Loa Observatory
# were suspended as of Nov. 29, 2022. Observations starting in December 2022 are from a site
# at the Maunakea Observatories, approximately 21 miles north of the Mauna Loa Observatory.
#
#            decimal       monthly    de-season  #days  st.dev  unc. of
#             date         average     alized           of days  mon mean
 1958    3   1958.2027      315.70      314.43     -1   -9.99   -0.99
 1958    4   1958.2877      317.45      315.16     -1   -9.99   -0.99
 1958    5   1958.3699      317.51      314.71     -1   -9.99   -0.99
 1958    6   1958.4548      317.24      315.14     -1   -9.99   -0.99
 1958    7   1958.5370      315.86      315.18     -1   -9.99   -0.99
 1958    8   1958.6219      314.93      316.18     -1   -9.99   -0.99
 1958    9   1958.7068      313.20      316.08     -1   -9.99   -0.99
 1958   10   1958.7890      312.43      315.41     -1   -9.99   -0.99
 1958   11   1958.8740      313.33      315.20     -1   -9.99   -0.99
 1958   12   1958.9562      314.67      315.43     -1   -9.99   -0.99
 1959    1   1959.0411      315.58      315.55     -1   -9.99   -0.99
 1959    2   1959.1260      316.48      315.86     -1   -9.99   -0.99
 1959    3   1959.2027      316.65      315.38     -1   -9.99   -0.99
 1959    4   1959.2877      317.72      315.41     -1   -9.99   -0.99
 1959    5   1959.3699      318.29      315.49     -1   -9.99   -0.99
 1959    6   1959.4548      318.15      316.03     -1   -9.99   -0.99
 1959    7   1959.5370      316.54      315.86     -1   -9.99   -0.99
 1959    8   1959.6219      314.80      316.06     -1   -9.99   -0.99
 1959    9   1959.7068      313.84      316.73     -1   -9.99   -0.99
 1959   10   1959.7890      313.33      316.33     -1   -9.99   -0.99
 1959   11   1959.8740      314.81      316.68     -1   -9.99   -0.99
 1959   12   1959.9562      315.58      316.35     -1   -9.99   -0.99
 1960    1   1960.0410      316.43      316.40     -1   -9.99   -0.99
 1960    2   1960.1257      316.98      316.36     -1   -9.99   -0.99
 1960    3   1960.2049      317.58      316.28     -1   -9.99   -0.99
 1960    4   1960.2896      319.03      316.70     -1   -9.99   -0.99
 1960    5   1960.3716      320.04      317.22     -1   -9.99   -0.99
 1960    6   1960.4563      319.59      317.47     -1   -9.99   -0.99
 1960    7   1960.5383      318.18      317.52     -1   -9.99   -0.99
 1960    8   1960.6230      315.90      317.19     -1   -9.99   -0.99
 1960    9   1960.7077      314.17      317.08     -1   -9.99   -0.99
 1960   10   1960.7896      313.83      316.83     -1   -9.99   -0.99
 1960   11   1960.8743      315.00      316.88     -1   -9.99   -0.99
 1960   12   1960.9563      316.19      316.96     -1   -9.99   -0.99
 1961    1   1961.0411      316.89      316.86     -1   -9.99   -0.99
```

The word "data" seemingly first appeared in 1640 in the English language when physicists realized that certain systems were better understood when described in a small number of parameters whose values would be known as "data".

The term "datalogy" (as in the notes' title of Dr. Jackel) has been introduced by Naur [1966] to denote "the science of the nature and use of data". Since then, it has become synonymous with the field of computer science in Denmark according to Sveinsdottir and Frø kjær [1988]. However, this terminology has not been widely adopted worldwide.

Data Analysis (for scientific research)

Data Science

Data science is the study of data to extract meaningful insights for business (and scientific research). It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.
Based Source: Amazon Web Services

Data science is used in physics to develop data-driven models that can make predictions and generate insights.

"big data"

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.
Source: oracle.com

Data Analysis (for scientific research)

Data Mining

Data mining is the process of “digging” through vast raw data to discover new and previously unknown information. The process may include data extraction, data cleaning, data fusion, data reduction, and feature creation, known as preprocessing and postprocessing steps such as pattern and model analysis, confirmation, and generation of hypotheses.

Data mining refers to discovering new patterns from a wealth of data in databases by focusing on algorithms to extract useful knowledge.

Source: I. A. Rauf, Physics of Data Science and Machine Learning

Data Analysis (for scientific research)

APS Advancing Physics

Data Science in Industry

A physicist in a data science job will spend most of their time analyzing data and designing and developing models to predict how something will behave based on data of how it has behaved in the past.

https://www.aps.org/careers/physicists/data.cfm

IOP Institute of Physics

The rise and rise of data science

https://www.iop.org/rise-and-rise-data-science#gref

Analysis

The separation of any material or abstract entity into its constituent elements.
Source: The Random House College Dictionary

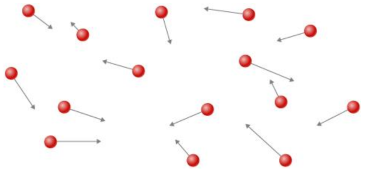
- For our analysis to be meaningful, it is implicit that the data being analyzed contain some "signal" representing the phenomenon of interest (or some aspect of it).
- Attempt to separate the signal (or feature) from the noise present in the data.
- A signal can be characterized in terms of its robust features and, in the case of complex phenomena, separated into further constituents, each of which may provide additional insights regarding the character, and behaviour of multiple processes contributing to the phenomenon of interest. *"divide and conquer"*

Based Source: D. G. Martinson, Quantitative Methods of Data Analysis for the Physical Sciences and Engineering.

Types of Data

Data can represent measurements of quantities or variables. The latter can be classified as:

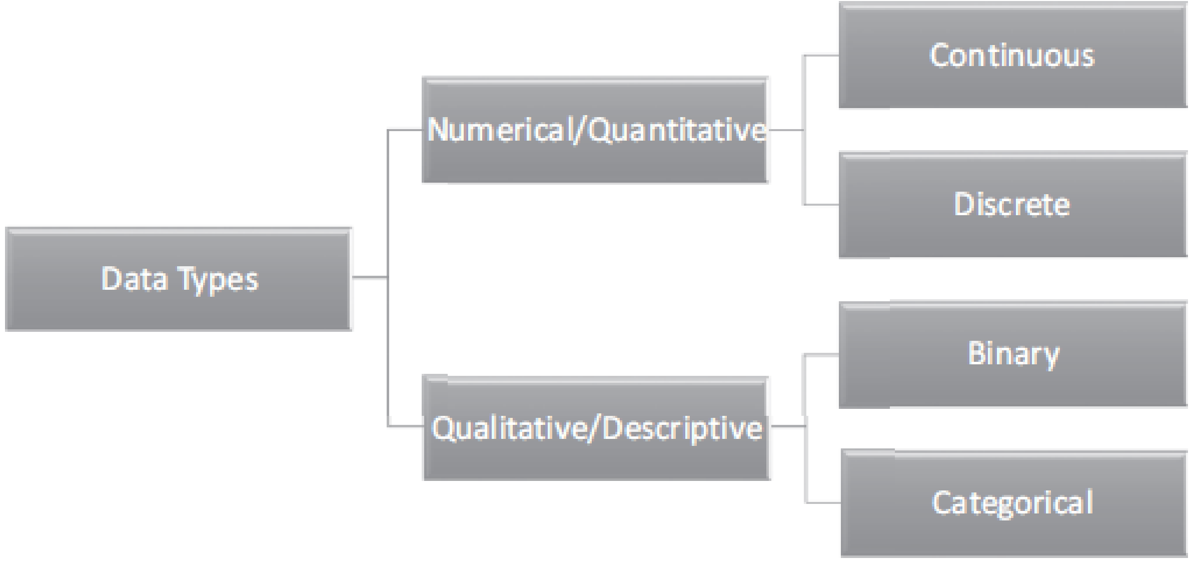
- **Discrete variables:** having discontinuous or individually distinct possible outcomes. E.g., flipping of a coin or rolling of dice.
- **Continuous variables:** having an uninterrupted range of possible outcomes. E.g., position and velocity of a free particle.



The data are also classified according to how they are recorded:

- **Analog data:** have been recorded "continuously" though, technically, purely continuous data is not realistic given that instrumentation does not necessarily respond instantaneously.
- **Discrete (or digital) data:** have been recorded at discrete (well-defined) intervals. All data, when represented on digital computers, are discrete.

Types of Data



Positive/Negative, Pass/Fail, etc.

Attributes such as colour, nationality, movie genre, etc.

Source: I. A. Rauf, Physics of Data Science and Machine Learning

Types of Data: sequential vs. non-sequential

A **sequential data series** consists of measurements of a quantity that vary as a function of another (control) quantity, and the order in which the measurements occur is important. In this case, the variable being measured is typically referred to as the **dependent variable**, while the control variable is the **independent variable**.

Most physical phenomena are described by sequential data!

Non-sequential data sets have no order. E.g., recording groceries purchased, Facebook friends, and information depicted in histograms.

Data representation

Multidimensional data: refers to data that has more than one dimension, where each dimension represents a distinct attribute or variable. In general, data can be organized and visualized across multiple dimensions or axes and a function/model can be used to map their relation. This means high-dimensional hypersurfaces can be analyzed! Examples:

- Pressure of gas collected in terms of pressure, volume, and density. Or kinetic energy as a function of (v_x, v_y, v_z).
- Information about galaxies in the universe being collected in multi-dimensional frame such as spatial coordinates (x, y, z), redshift, luminosity, distance from Earth, etc.
- In particle physics, experiments at accelerators involve collecting data from particle collisions. The multidimensional data might include variables such as energy, momentum, charge of the particles, spatial trajectories, etc.

Data representation

Multivariate data: refers to data sets where each observation (or data point) is described by multiple variables or features. In other words, each entry in the data set contains measurements or values for multiple characteristics. Multivariate data analysis involves exploring relationships and patterns among these variables within the same observation.

This may be data that may be very difficult to characterize in various axes as in the multidimensional representation.

For example, if one is analyzing a dataset of people’s medical records which can include heights, weights, and ages, each person’s data point would contain values for all three variables. Or student performance could include test scores, homework averages, and participation scores as multivariate data.

Multivariate statistical analysis proposes to study the joint distribution of all attributes, in which the distribution of any single variable is analyzed in terms of the other attributes’ distributions or through correlation/relationships between them.

Multidimensional vs. Multivariate

The key distinction is that “multidimensional data” refers to data that spans multiple dimensions or axes, while “multivariate data” refers to data sets where each observation is characterized by multiple variables.

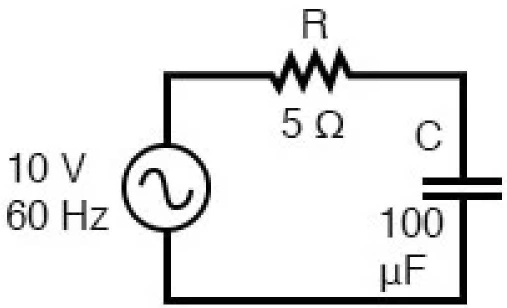
Multidimensional data can encompass multiple variables within each dimension, but it is the concept of “dimensionality” that is highlighted. Multidimensional data interpretation also usually applies to quantities that have a theory or equation describing their relations. Multivariate data focuses on the idea of multiple variables/features within individual data points in which there may not necessarily exist an expression describing their dependency.

Multidimensional data often allows the use of linear or tensorial algebra techniques to manage the data, whereas multivariate data may require the use of multivariate statistical analysis techniques, some being covered in this course.

Real vs. Complex (number) Data

“Real” data are what we deal with in the real world but treating them mathematically as complex numbers affords us the ability to conveniently consider rotations (or phases) of a quantity and to avail from complex mathematical methods. A complex data item consists of two parts—a real part and an imaginary part. But note that certain real data can be organized as if they are complex quantities.

For instance, in an AC circuit with a resistor and a capacitor in series, the impedance of the circuit can be represented as a complex number:



$$Z = R + jX$$

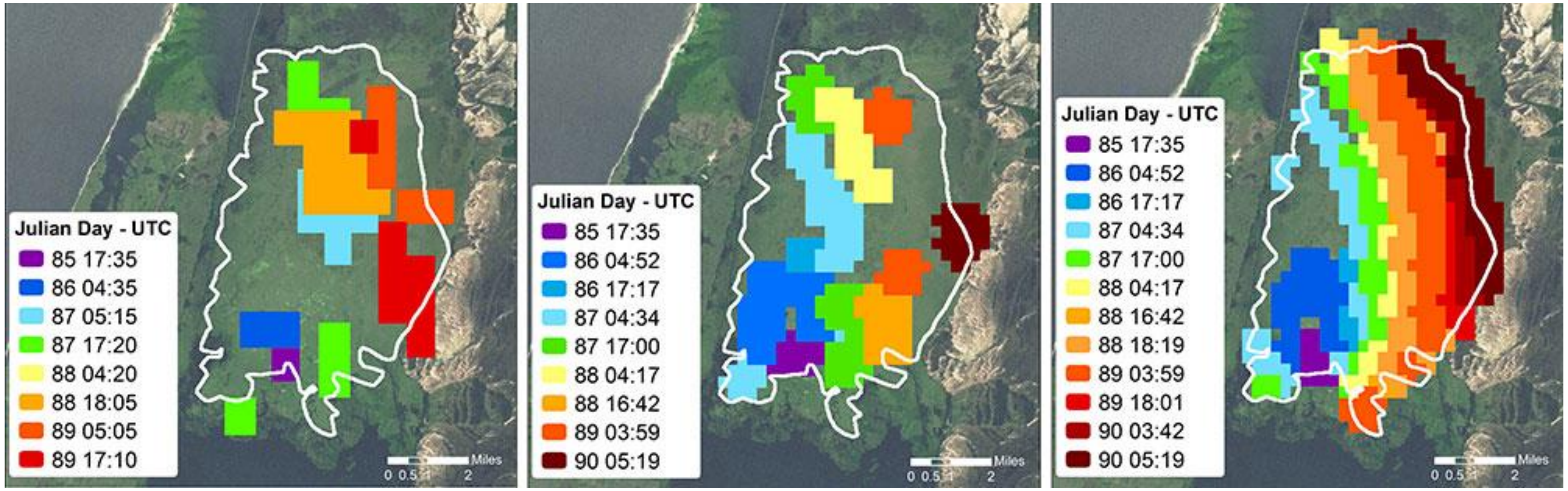
Z is the impedance (complex number).
R is the resistance (real part).
j is the imaginary unit.
X is the reactance (imaginary part).

The imaginary part X captures the phase difference between the voltage and the current due to the reactance of the capacitor. The real part, R, accounts for the resistance in the circuit.

Reading and Storing Data

VIIRS I-Band 375 m Active Fire Data

The Visible Infrared Imaging Radiometer Suite (VIIRS) 375 m thermal anomalies/active fire product provides data from the VIIRS sensor aboard the joint NASA/NOAA Suomi National Polar-orbiting Partnership (Suomi NPP) and NOAA-20 satellites.



A comparison of daily fire spread mapped by 1 km Aqua/MODIS (left), 750m VIIRS (center) and 375m VIIRS (right) data at the Taim Ecological Reserve in southern Brazil (March 26-31, 2013).

https://doi.org/10.1016/j.rse.2013.12.008

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