Introduction: Water data, Python and You

Waterhackweek 2020

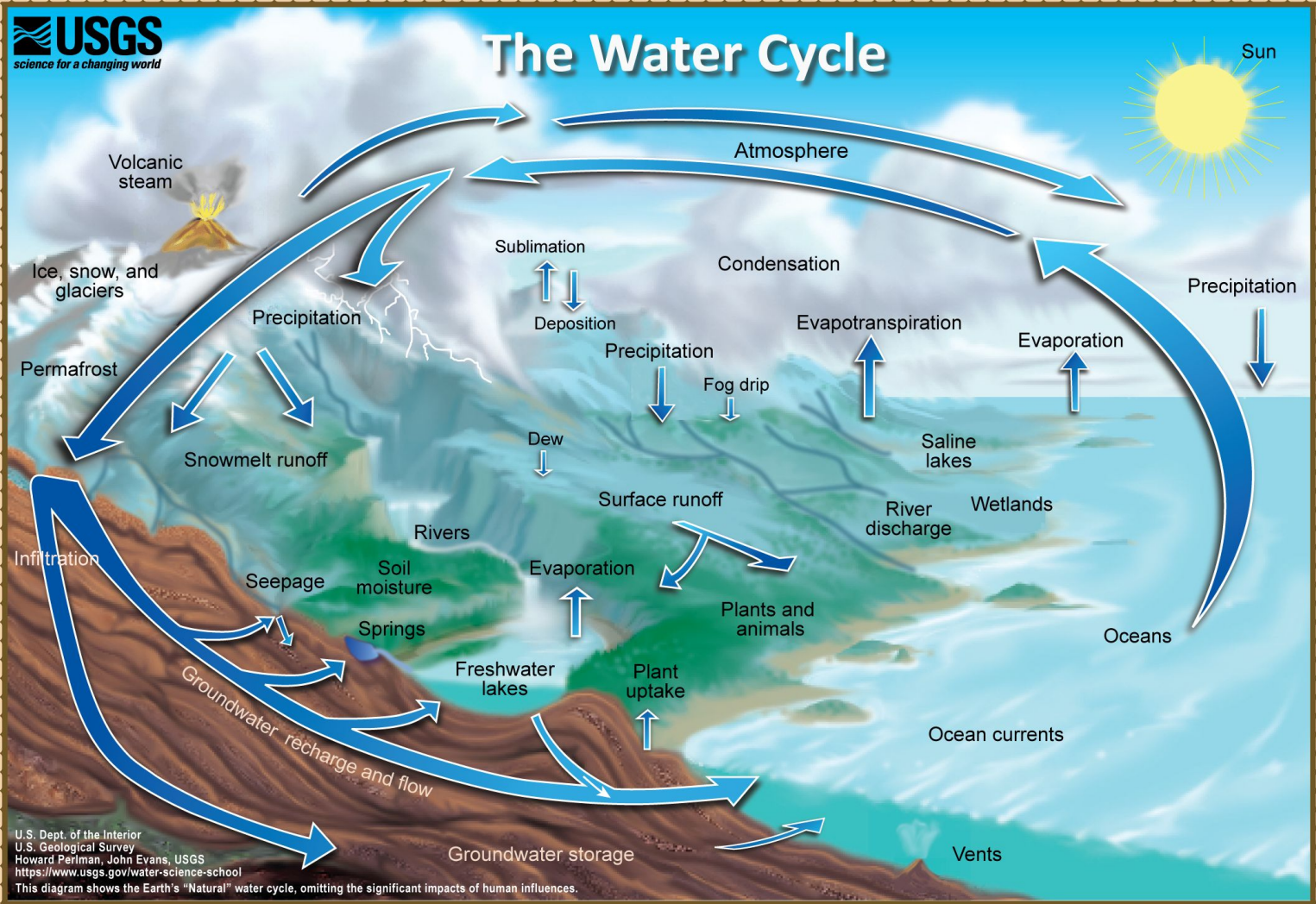
Bart Nijssen

Presentation: 30 minutes

Tuesday Schedule

Post questions on slack in #tutorials

08:00 PDT	**Project Updates**
08:30 PDT	**Introduction: Water data, Python and You** Bart Nijssen
09:00 PDT	**Break**
09:15 PDT	**Access and analyze point time series data** Yifan Cheng
10:00 PDT	**Break**
10:15 PDT	**Access and analyze raster and multi-dimensional gridded data** Steven Pestana
11:00 PDT	**Break**
11:15 PDT	**Water data mash up** Emilio Mayorga
12:00 PDT	**Break: Qiqochat Zoom rooms & YoTribe Open** Mary Dumas



Source:
The USGS Water Science School - The Water Cycle

Data comes in many forms from many sources ...

Credit: John Lee 05/19/2003
https://ameriflux.lbl.gov/wp-content/uploads/2015/05/Howland_main2-768x1024.jpg

Some examples:

Point time series of environmental observations

Advantage: Measurement in real world

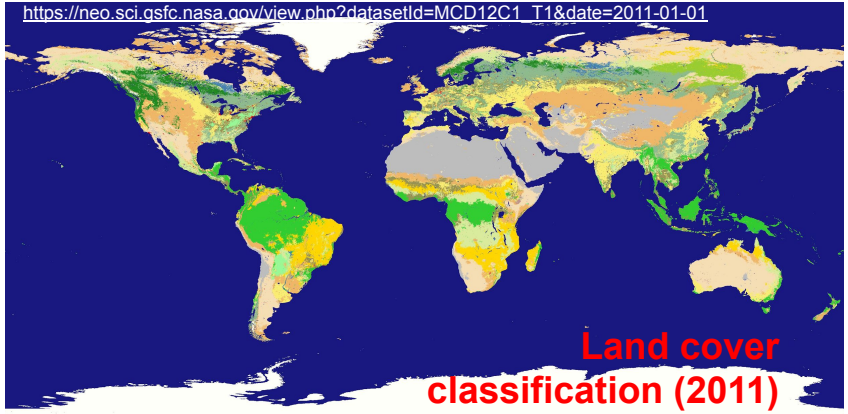
https://waterdata.usgs.gov/usa/nwis/uv?01581960

https://www.usgs.gov/media/images/rain-gage

Challenges:
- Representativeness
- Measurement error
- Missing data
- Changes in technique and location
- Record length
- Data format
- etc.

https://www.zooniverse.org/projects/mozerm/snow-spotter

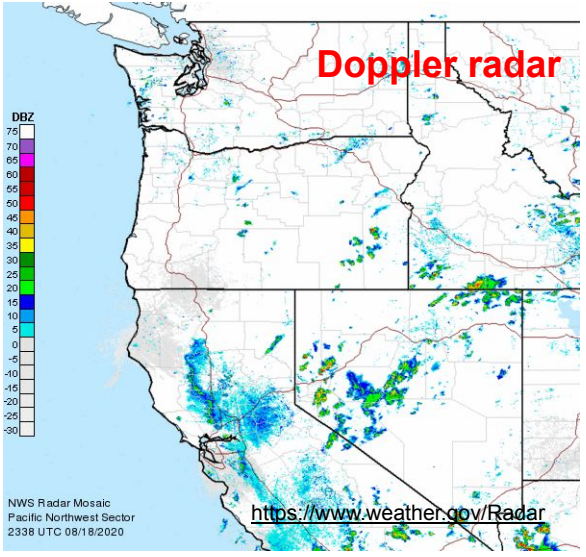
Data comes in many forms from many sources ...



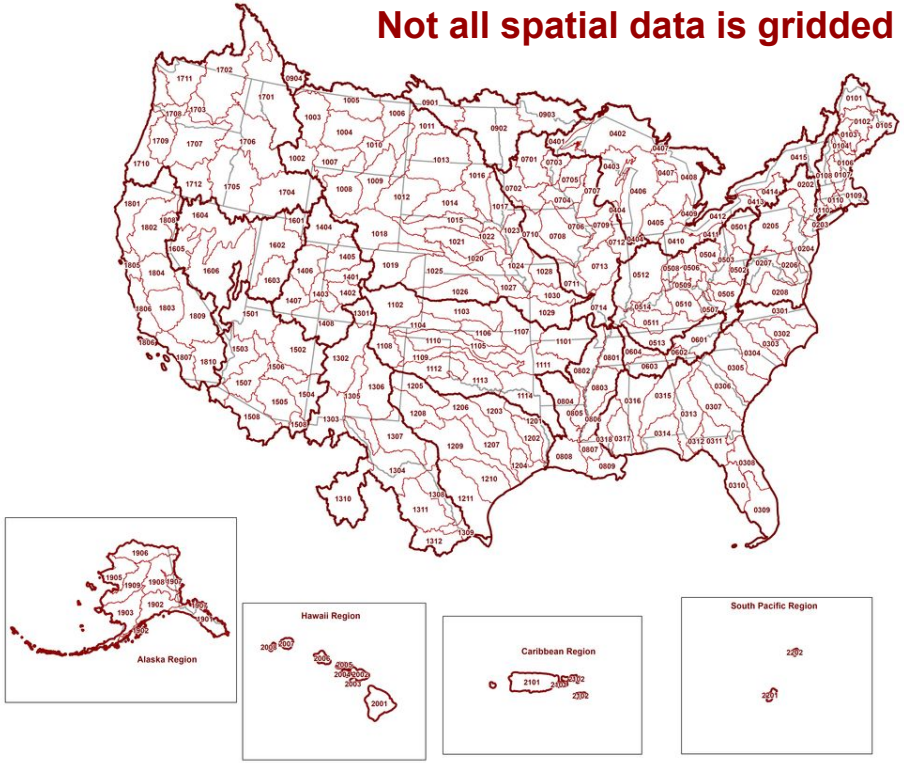
Spatial data from remote sensing platforms

Challenges:

- Measurement error
- Missing data
- Changes in technique
- Record length
- Data format
- Data size
- Projection
- Many different spatial arrangements
- etc.



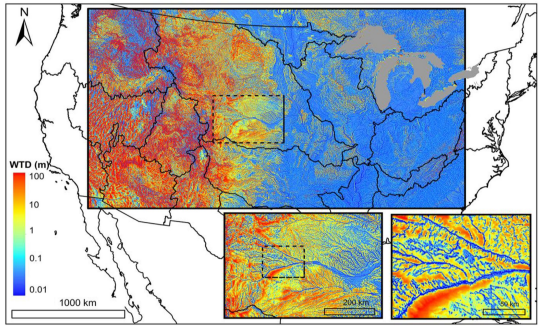
Data comes in many forms from many sources ...



USGS Watershed Boundary Dataset

https://www.usgs.gov/core-science-systems/ngp/national-hydrography/watershed-boundary-dataset

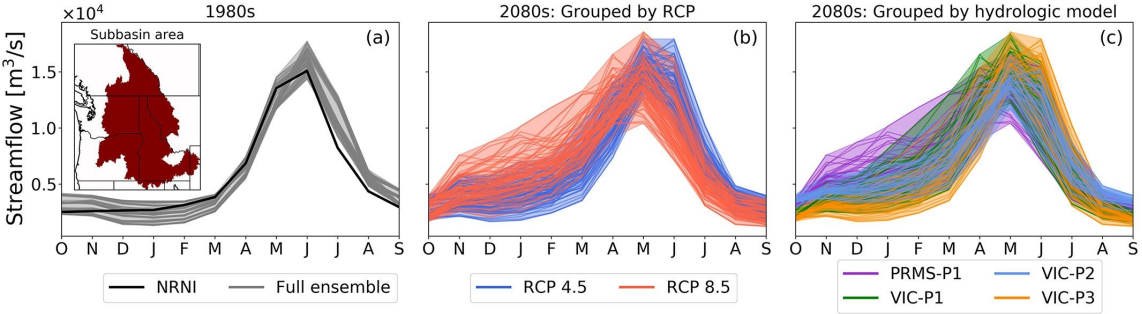
Data comes in many forms from many sources ...



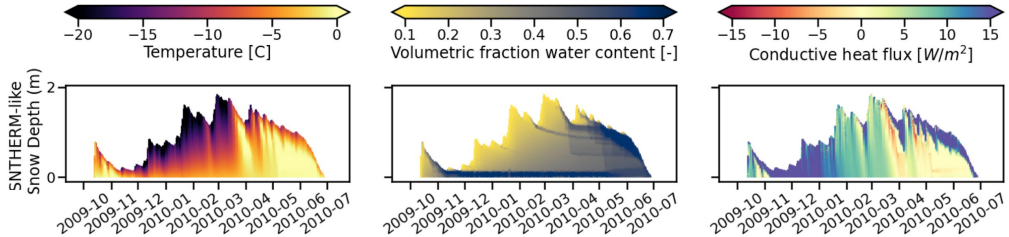
ParFlow: Modeled water table depth
Kuffour et al., 2020:
https://doi.org/10.5194/gmd-13-1373-2020

Climate change projections for streamflow at the Dalles along the Columbia River

Chegwidden et al., 2019: https://doi.org/10.1029/2018EF001047



Model output



SUMMA snow model output

Advantages:

- Typically no missing values
- Can provide insight into variables that cannot be measured easily (or at all)
- Can provide insight into conditions that do not exist: scenarios, forecasting, climate change, land use change, etc.: **what-if**

Challenges:

- Model error
- Model uncertainty
- Data format
- Data size
- Many different spatial arrangements
- etc.

Data discovery

Data tends to be scattered across many sites, repositories, etc. and may be difficult to locate at times.

Data is also made available under many different licenses and with many different restrictions and is easier to access in some geographic regions than in others.

A number of initiatives exist to make data discovery easier, but it often remains a significant effort.

Where do I find the data I need?

No single catalog will meet all your needs for search, discovery, and convenient access to data, equally well, in all domains of water research!

Possible starting points:

- **Google Search**
  Overwhelming. Many irrelevant results (but often the first place to start)

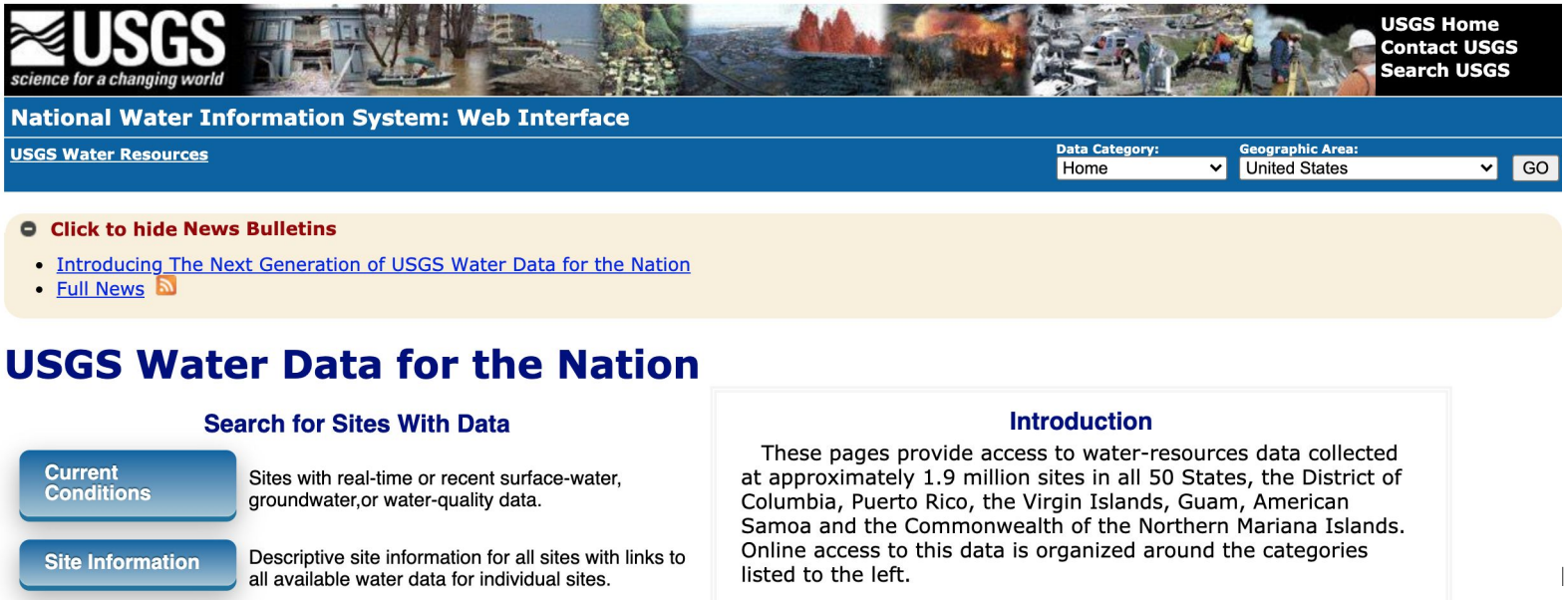
- **Google Earth Engine**: https://earthengine.google.com/
  Fantastic for remote sensing data and processing. Also holds some gridded model products. Not yet there for site time series, site data

- **Google Dataset Search**: https://toolbox.google.com/datasetsearch
  New, promising, but still rather messy

Where do I find the data I need?

Many topical or organizational repos, e.g. in U.S.:

- USGS: https://waterdata.usgs.gov/nwis



Where do I find the data I need?

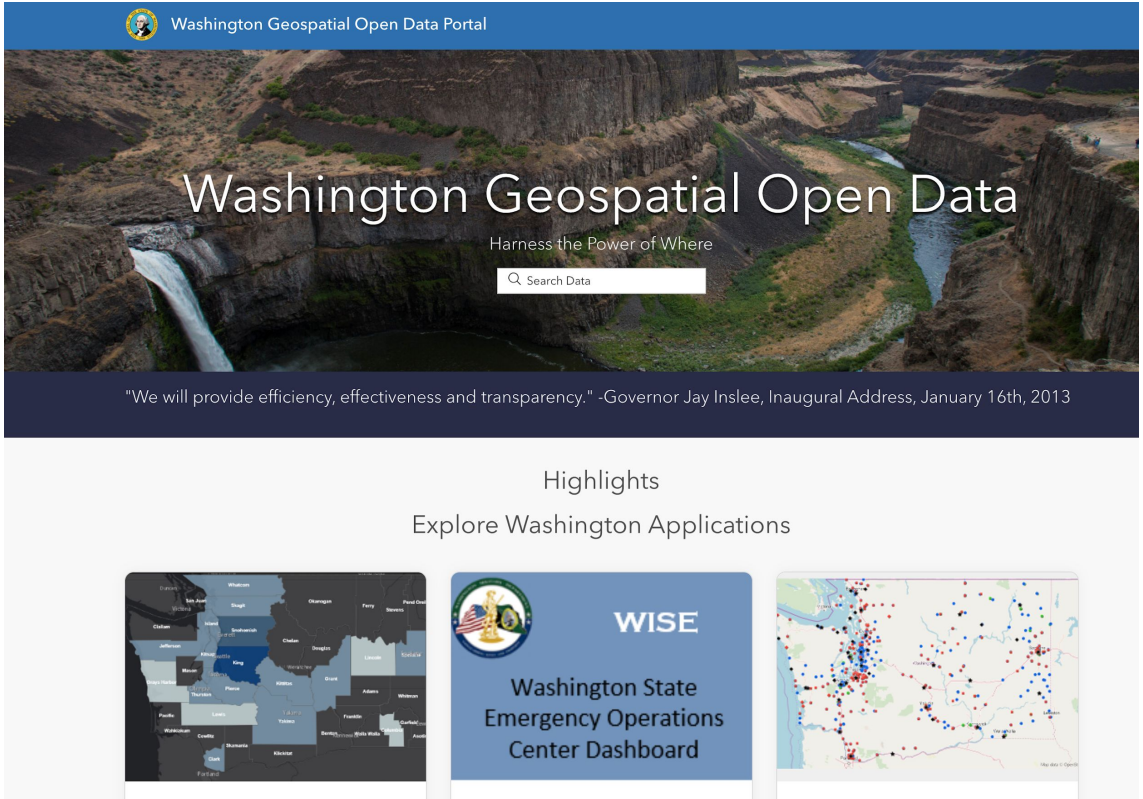
Many topical or organizational repos, e.g. in U.S.:

- NOAA NCDC: https://www.ncdc.noaa.gov/

Where do I find the data I need?

Many topical or organizational repos, e.g. in U.S.:

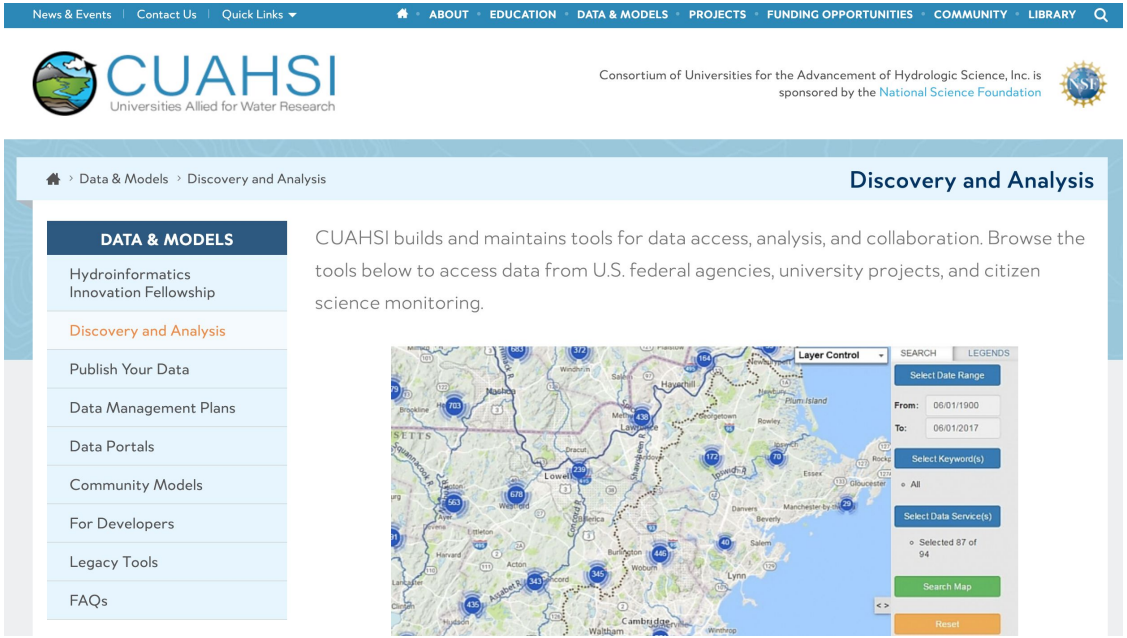
- Washington State: https://geo.wa.gov



Where do I find the data I need?

CUAHSI also provides data discovery tools that are relevant for hydrology:

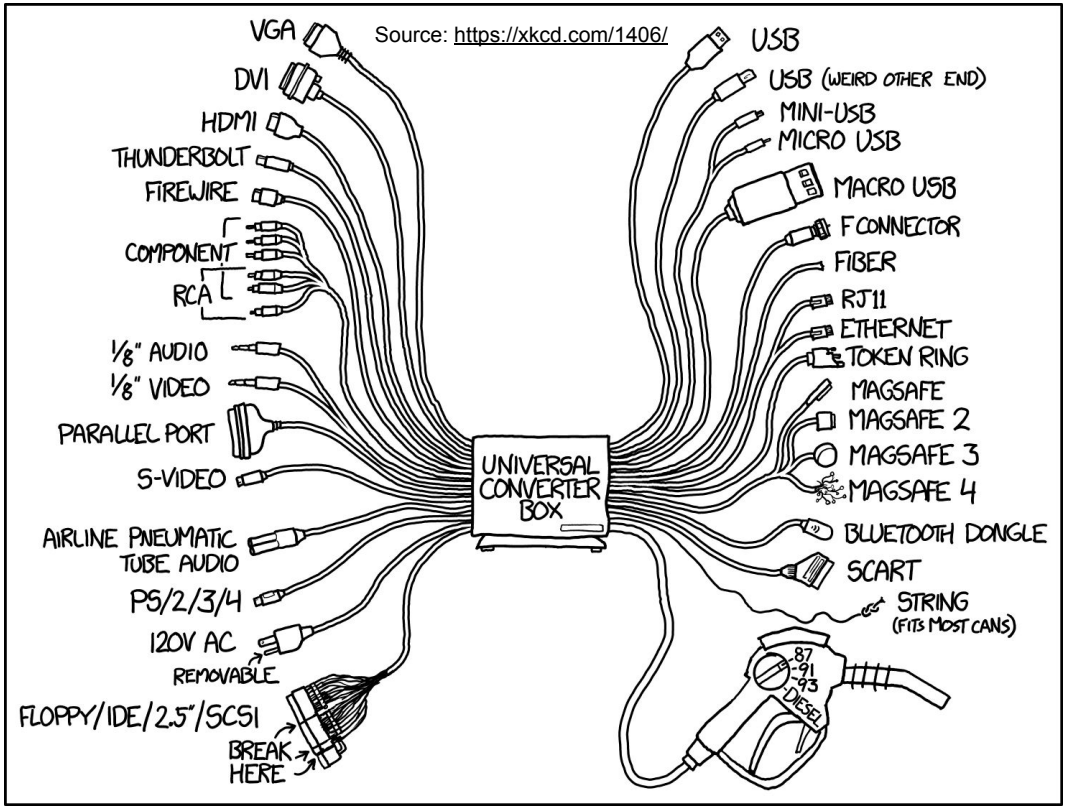
- CUAHSI: https://www.cuahsi.org/data-models/discovery-and-analysis



Data access, analysis, and export

Once you find the right data, accessing it, manipulating it, and exporting it in a usable form can be a significant task.

Fortunately there is a large set of tools that make this process easier. We'll see some of these later today.



Data access and ingest

Common approaches and tools

- Manual browsing, downloads, and reading local files (but issues of reproducibility, efficiency, thorougness)
- requests Python package (and wget, curl): generic remote access through the web.
- Pandas read_csv function. Not just local files, but also remote files.
- Custom web APIs (often called "REST" APIs) from the data provider (eg, NEON). Often fairly easy to use, but highly variable across systems.
- Standards-based resources:
    - APIs: OPeNDAP, Open Geospatial Consortium (OGC) Web Services (WFS, SOS, etc), **CUAHSI WaterOneFlow**
    - Formats: WaterML (**CUAHSI WaterML 1.x** vs OGC WaterML 2.0), NetCDF (3 "classic" vs 4), Metadata standards
    - See this old but still very useful descriptions of CUAHSI "HIS" standards
    - Standards enable reusability across multiple data sources, systems
- ulmo. Water and climate data. Wraps a lot of the underlying complexity into simpler, more user-friendly Python APIs.

Data storage

Data volumes can become very large, very quickly.

Examples:

- Remote sensing time series
- Model output

Storing data on a local hard-drive without backups is (more than) risky.

Downloading data locally can be very challenging when the volumes are large.

For very large data sets: Take computing to the data

- Examples: Google Earth Engine, Pangeo, Hydroshare.

Data publishing - why?

AGU DATA POLICY

AGU affirmed in its 2015 position statement that "***Earth and space science data should be widely accessible in multiple formats and long-term preservation of data is an integral responsibility of scientists and sponsoring institutions.***" Following this statement and to advance scientific exploration and discovery, and allow a full assessment of results presented in AGU's journals, ***all data necessary to understand, evaluate, replicate, and build upon the reported research must be made available and accessible whenever possible***.

AGU encourages authors to identify and archive their data in approved data centers.

AGU requires an explicit statement in the "Acknowledgments" section of a paper that clarifies how users can access the data from a paper (via supplements, repositories, other sources, etc.) and states any restrictions on access.

Data publishing - FAIR

FAIR - Findable, Accessible, Interoperable, Reusable

The **FAIR principles** were designed with data-driven and machine-assisted open science in mind. The final aim of following FAIR principles is that machines as well as people can Find, Access, Interoperate and thus Reuse each other's research objects.

FAIR is **not a standard**, although the acronym is frequently used in that context. The GO FAIR view is that standards are needed for the **Internet of FAIR Data and Services** and that ideally, standards, API's and protocols are developed 'following FAIR guiding principles'.

FAIR is **not equivalent to open** (and open is not equivalent to 'free').

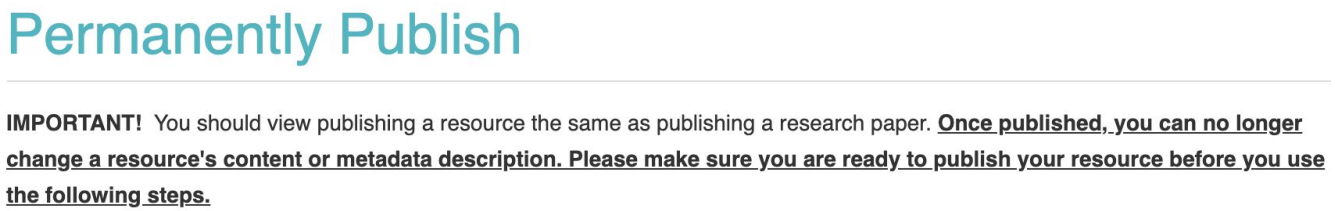
FAIR principles do not, in themselves, cover the crucial aspects of **intrinsic data quality or ethics**.

Source: GO FAIR https://www.go-fair.org/

Data publishing - where?

Sites that offer the opportunity to mint a DOI (Digital Object Identifier) for data sets:

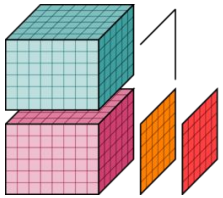
- **HydroShare:** https://help.hydroshare.org/introduction-to-hydroshare/getting-started/permanently-publish-a-resource/



- **Zenodo:** https://zenodo.org/

Python ecosystem

ulmo



xarray

Rasterio: access to geospatial raster data

cartopy

pandas

bokeh

NumPy

matplotlib

More information

- https://waterhackweek.github.io/learning-resources/
- https://www.cuahsi.org/education/cyberseminars/waterhackweek-cyberseminar-series/
- https://github.com/waterhackweek