Accessing data repositories

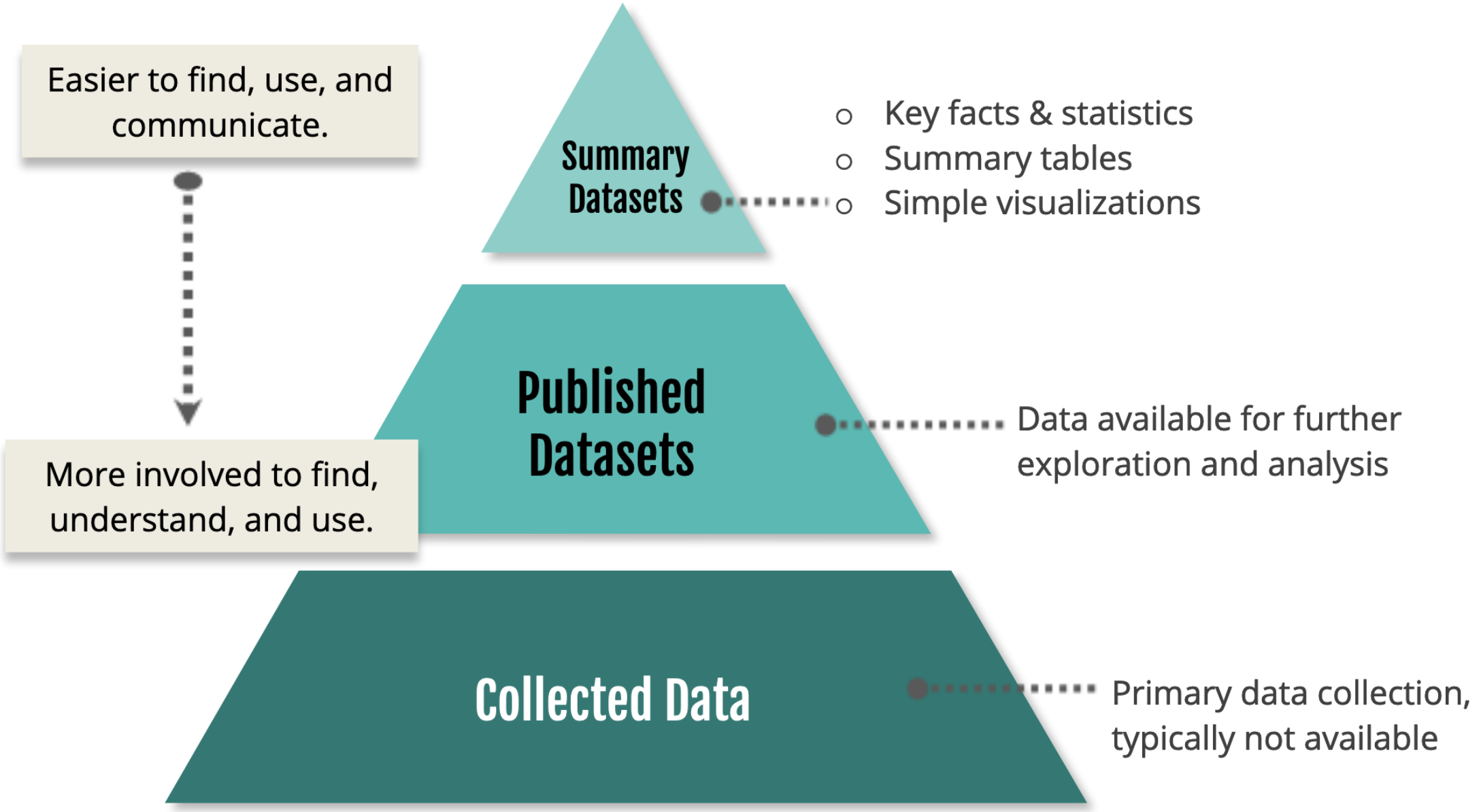
Oregon Data Science Collaborative

Spring 2022

Finding data

Data Production is a Process

Image credit Cal Poly Libraries



How to find the data you need?

Who collects the data?

Who aggregates or processes the data?

How do others access the data and report on the data?

Possible data sources

- Government data portals
- Research data repositories
- Other secondary data portals
- Library databases
- News reports and visualizations

Match search practices with data needs

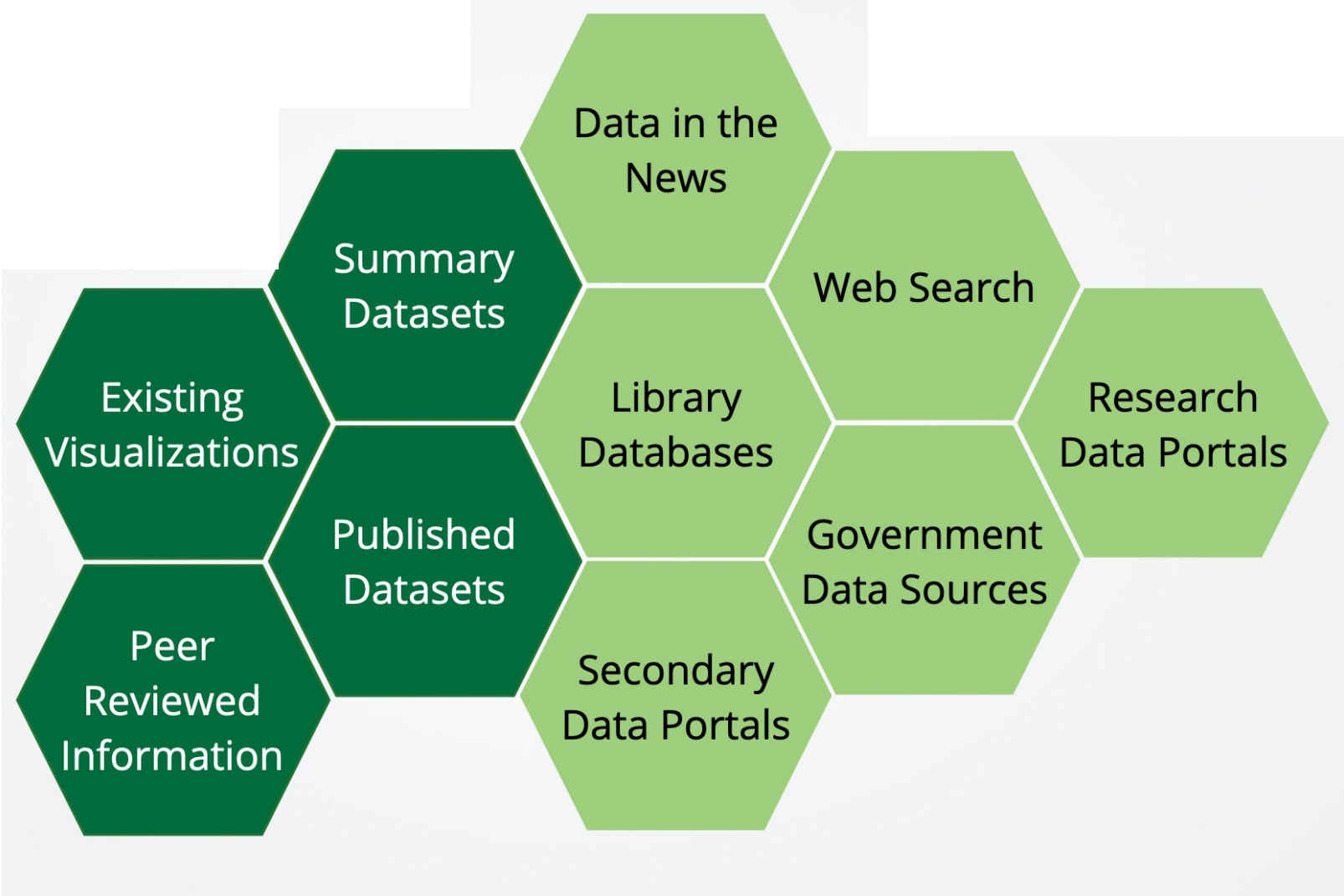


Image credit Cal Poly Libraries

What are your data needs?

How do you find the data you need?

What repositories have you used in the past?

Research data repositories

Landscape of data repositories

Collections of repositories

DataOne

Re3

Project specific

UN's IPCC data center

Institutional

UO Scholars Bank

General

Zenodo

Figshare

Dryad

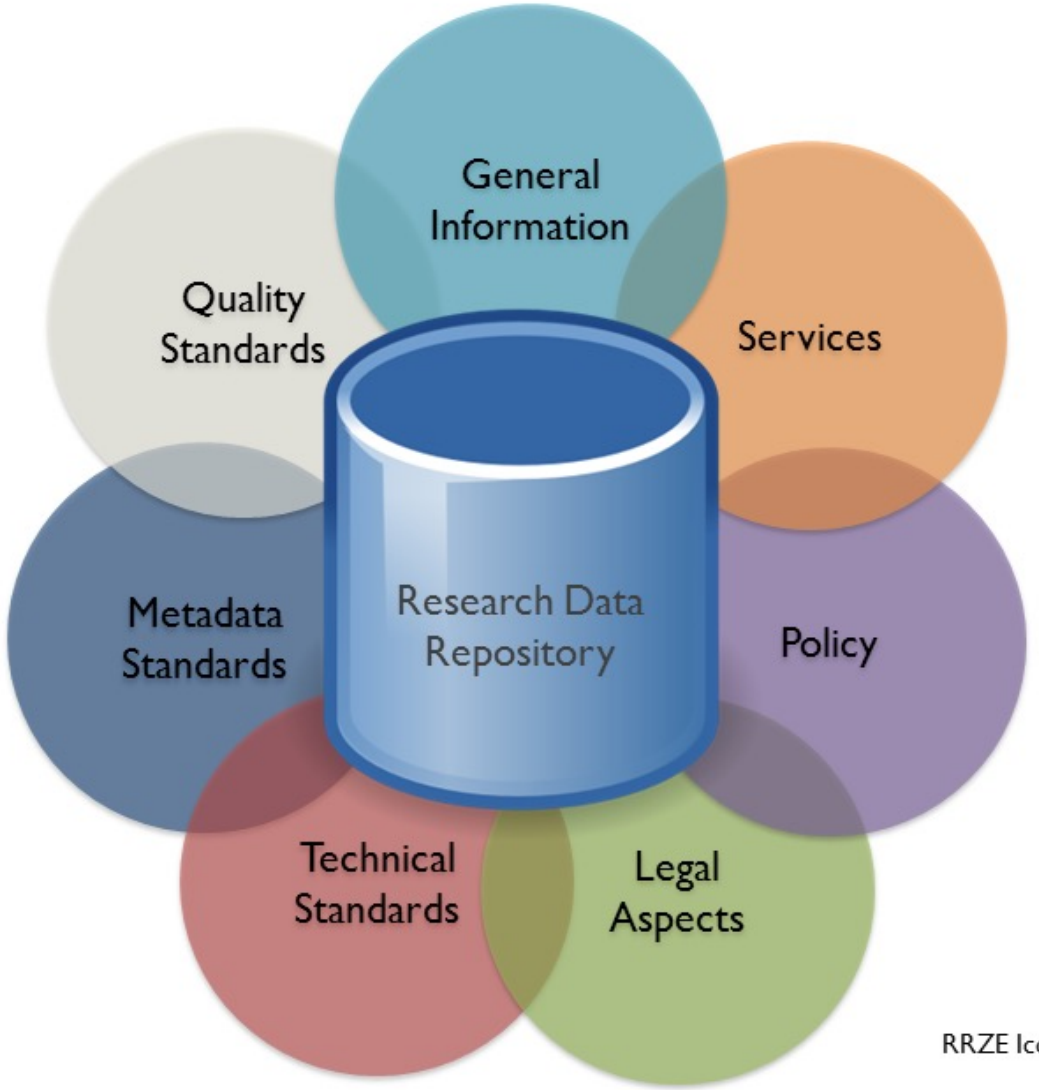
Discipline-specific

Environmental Data Initiative

Funder-specific

NSF's BCO-DMO

Different repositories have different features



RRZE Icon Set (CC: BY-SA)

Some data repos house specific scientific products

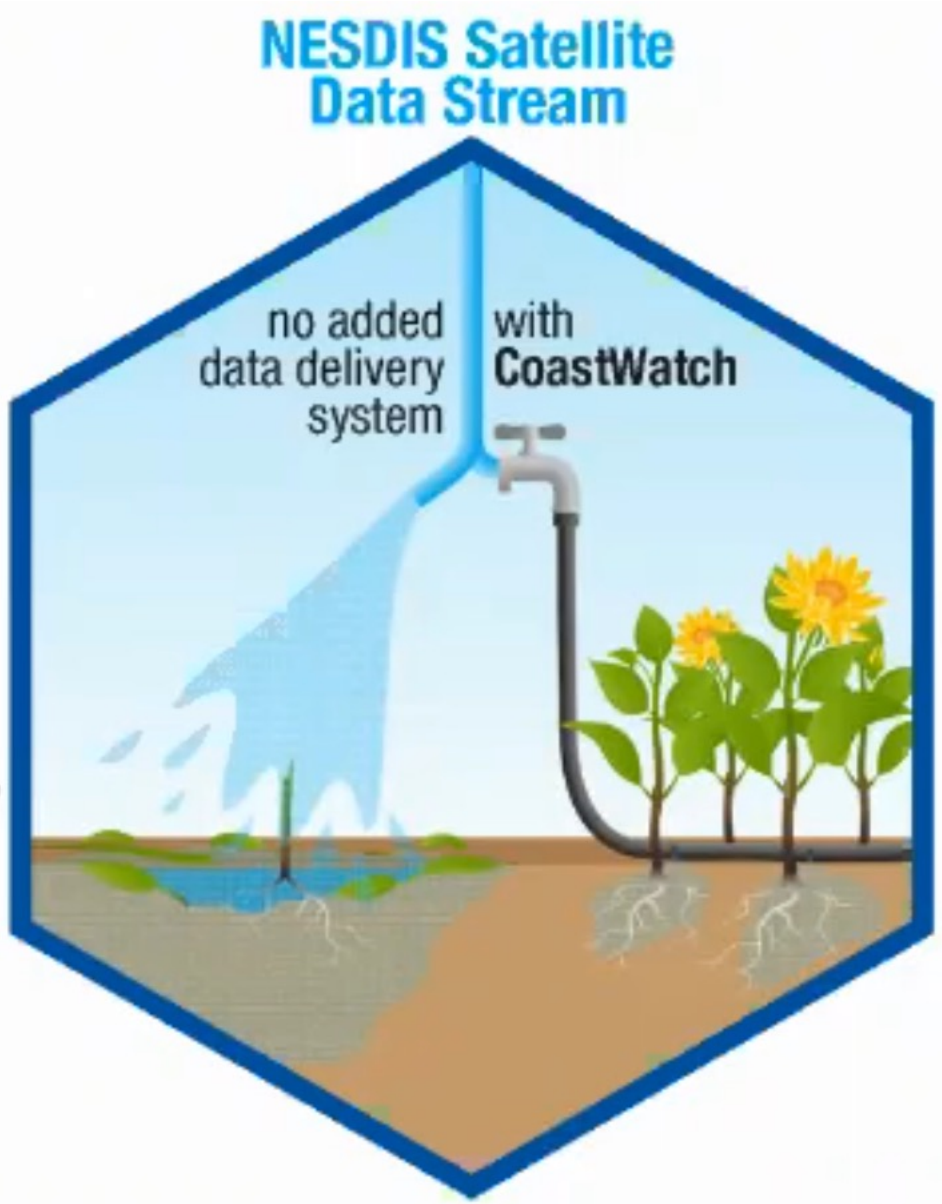
- Datasets attached to a particular paper or project
- Protocols and other related material can also be housed in repositories

Open Science Framework

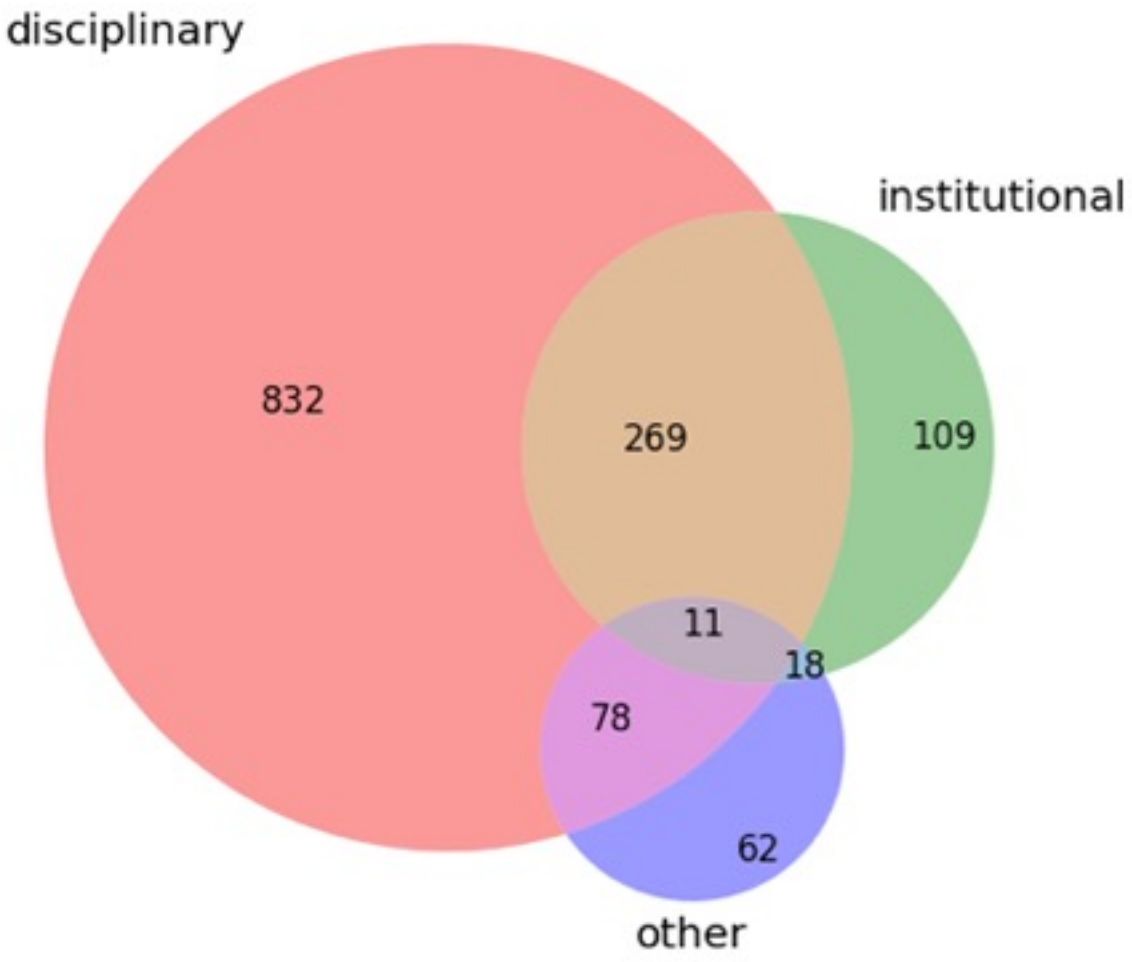
https://osf.io/

Some data repos provide data streams

- Especially pertinent for big data
- Repositories can manage, curate, and process raw data to provide the most useful outputs to users
- Example – CoastWatch Data Portal



Repositories serve multiple functions



The Landscape of Research Data Repositories in 2015

Finding datasets from a published paper

Ancestry-inclusive dog genomics challenges popular breed stereotypes

Accessing data

Mechanics of data access

Identify dataset of interest

Where is the dataset located?

What format is available?

How can you access the dataset?

How can you make data retrieval reproducible?

A few notes on data formats

Flat vs structured data

Subsetting is often critical

Specific data formats will have specific tools – e.g. netCDF, fastaq

Data access methods

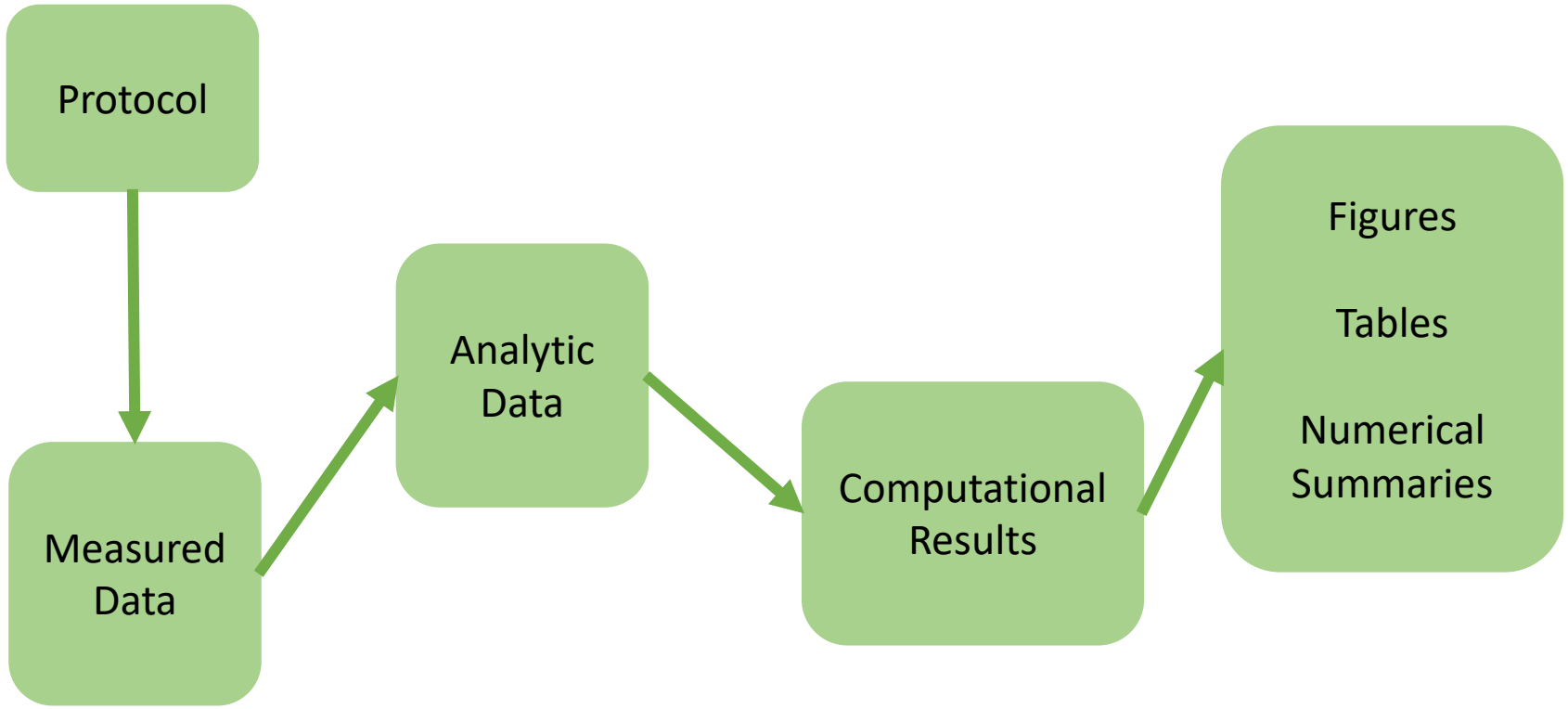
Direct download

- Point-and-click
- APIs
- Various tools to automate downloading

Cloud computing

Reproducibility + collaboration

How does data access fit into a reproducible workflow?



How can data access improve collaboration?

Resources for accessing data

- Colleagues
- Specialists – librarians, repository managers, data scientists

- Prior work – look for details in published papers and datasets
- Metadata

- Open science tools