Data Sourcing

Daily Agenda

- Describe the data life cycle (5 min)
- Discuss data sourcing considerations (15 min)
- Explore publicly-available environmental datasets (5 min)
- Write quality metadata (5 min)
- Data validation exercise (60 min)

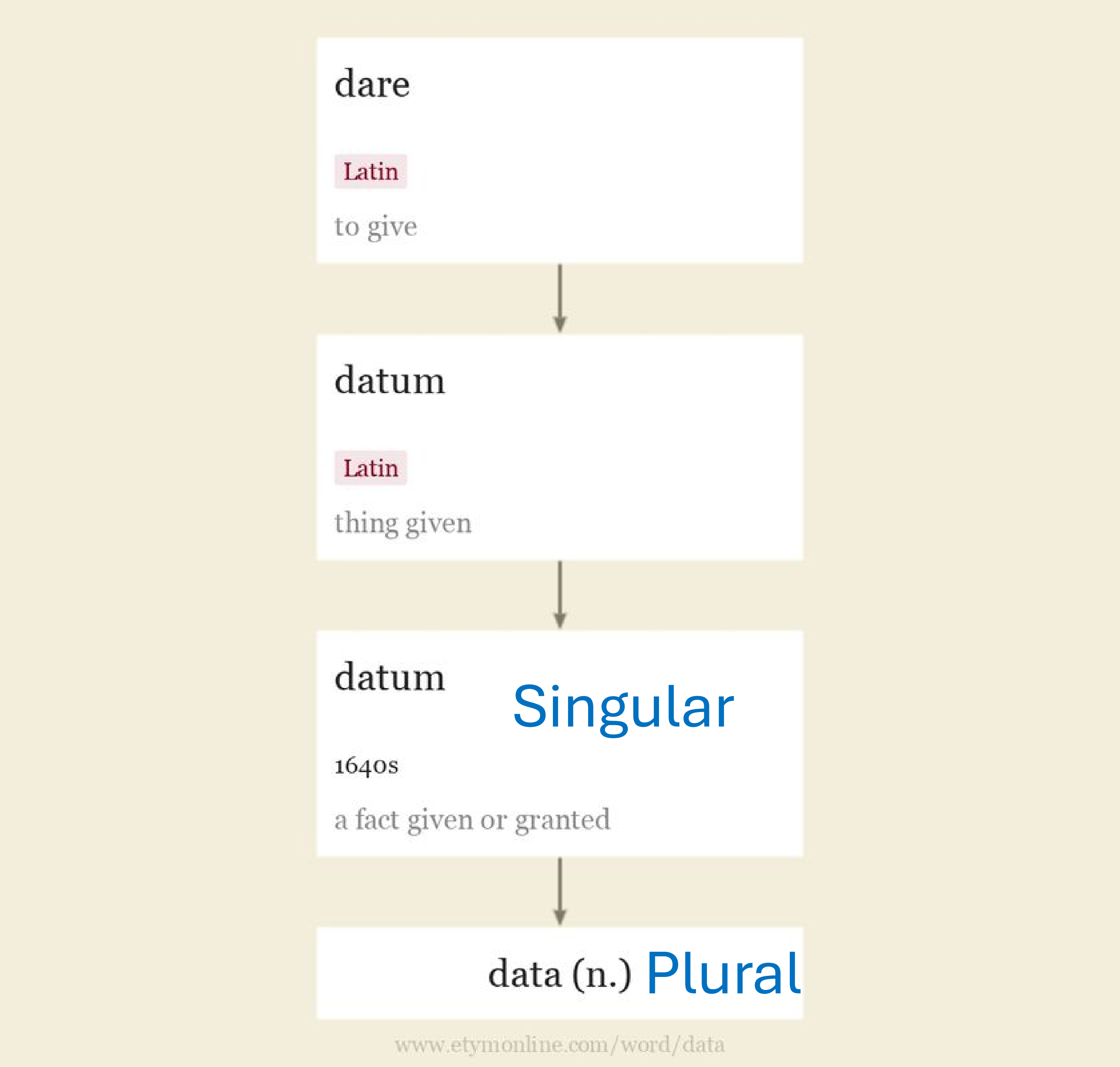
Week 1 Action Items

Due this week

- Nothing ☺

Due next week

- GitHub onboarding (1/21)
  - Create a GitHub account
  - Join the course GitHub repository
  - Download GitHub Desktop and clone repository



dare

Latin

to give

datum

Latin

thing given

datum

Singular

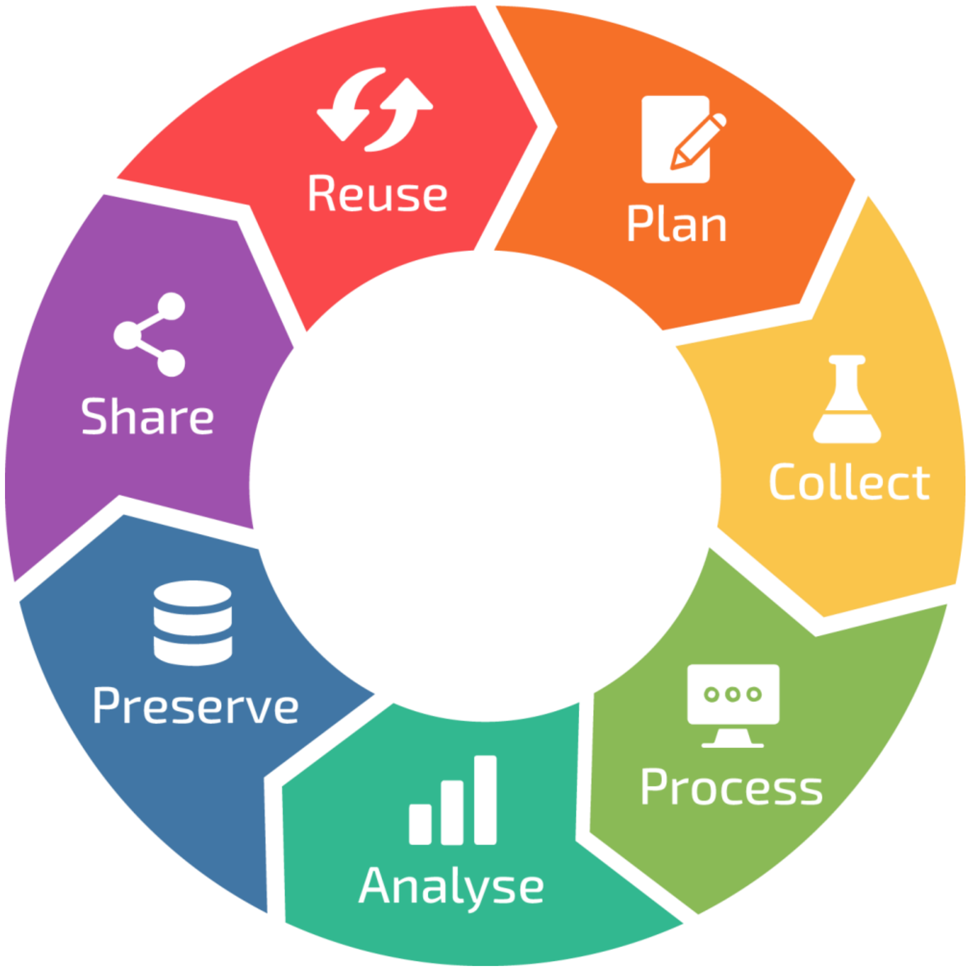
1640s

a fact given or granted

data (n.) Plural

www.etymonline.com/word/data

Data Life Cycle



Reuse

Plan

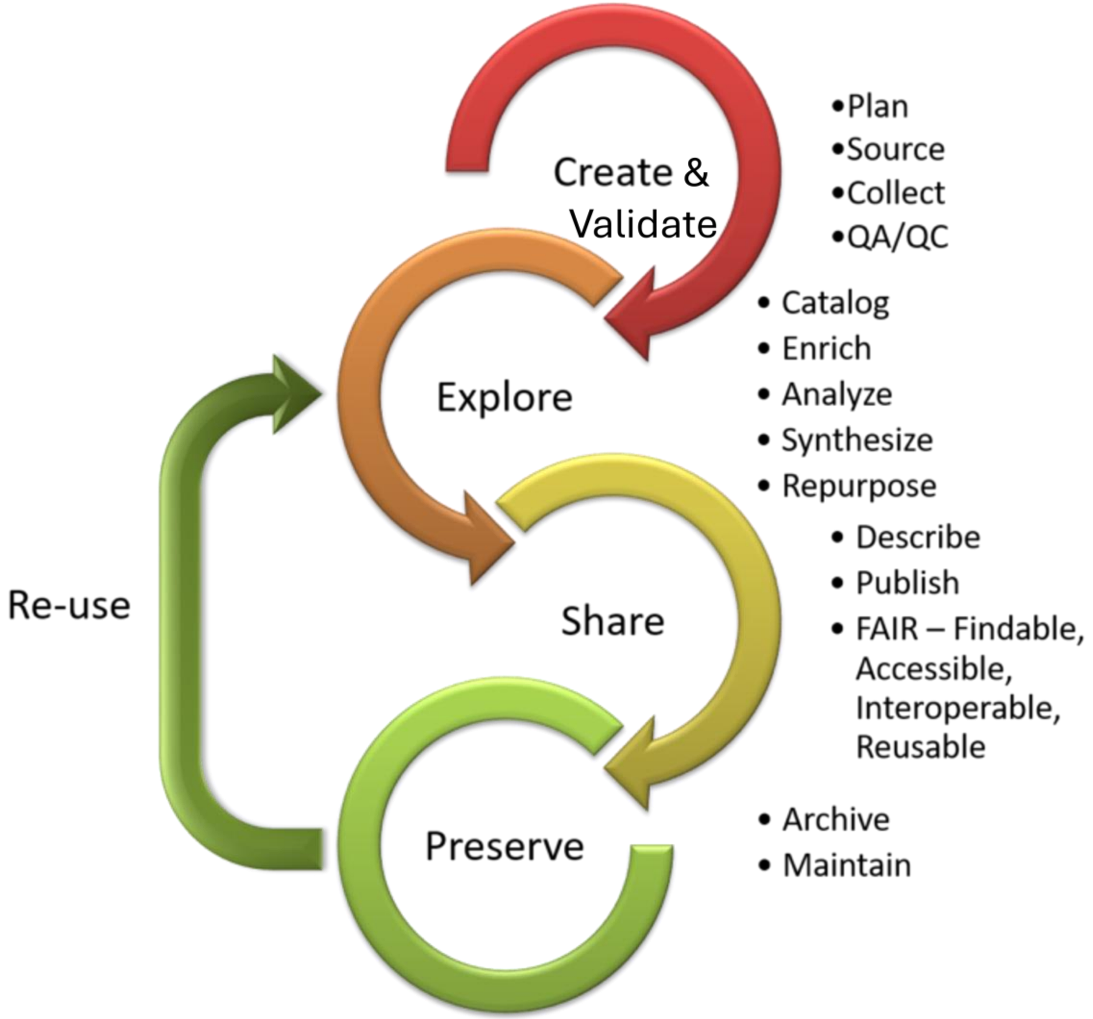
Collect

Process

Analyse

Preserve

Share



Create & Validate

- Plan
- Source
- Collect
- QA/QC

Explore

- Catalog
- Enrich
- Analyze
- Synthesize
- Repurpose

Share

- Describe
- Publish
- FAIR – Findable, Accessible, Interoperable, Reusable

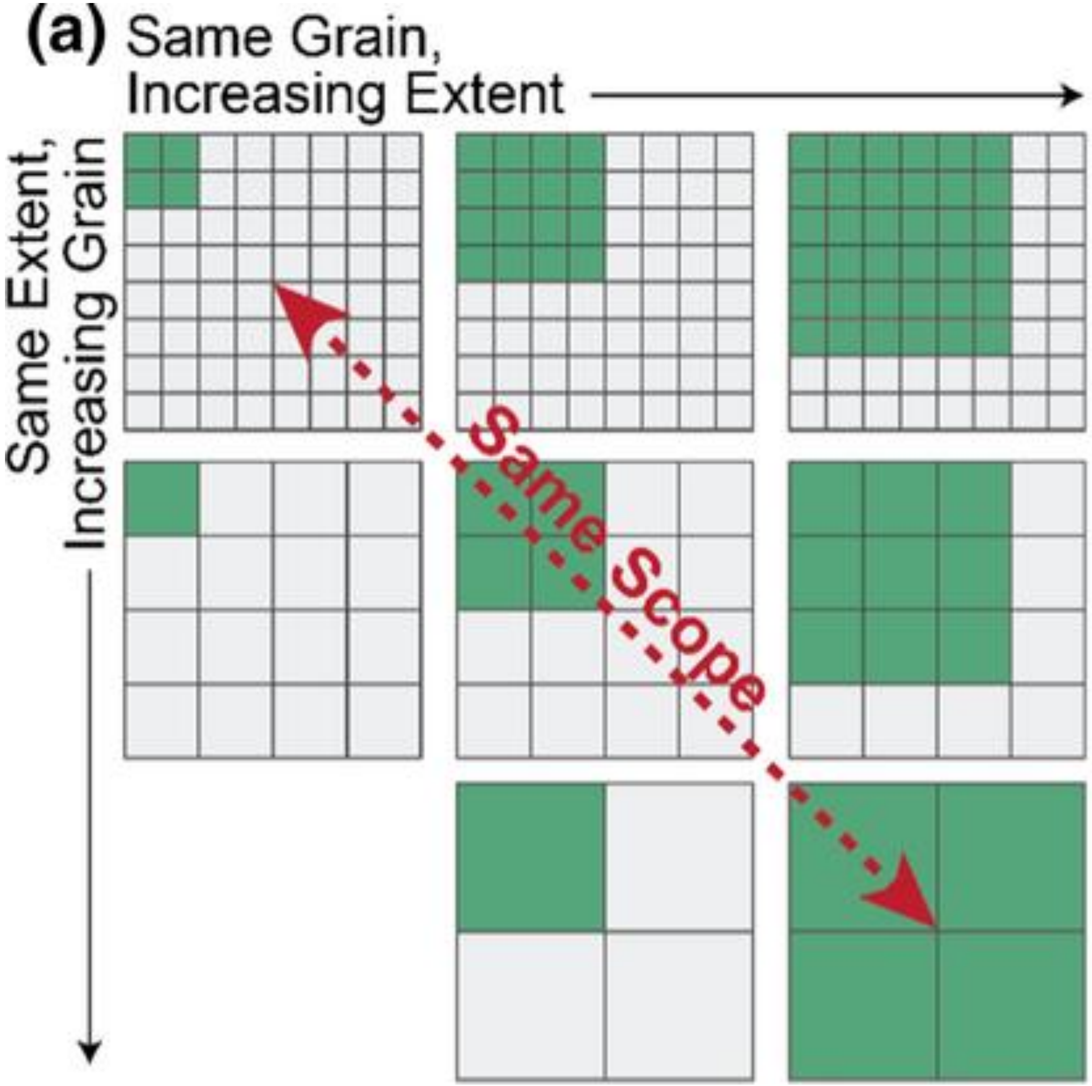
Preserve

- Archive
- Maintain

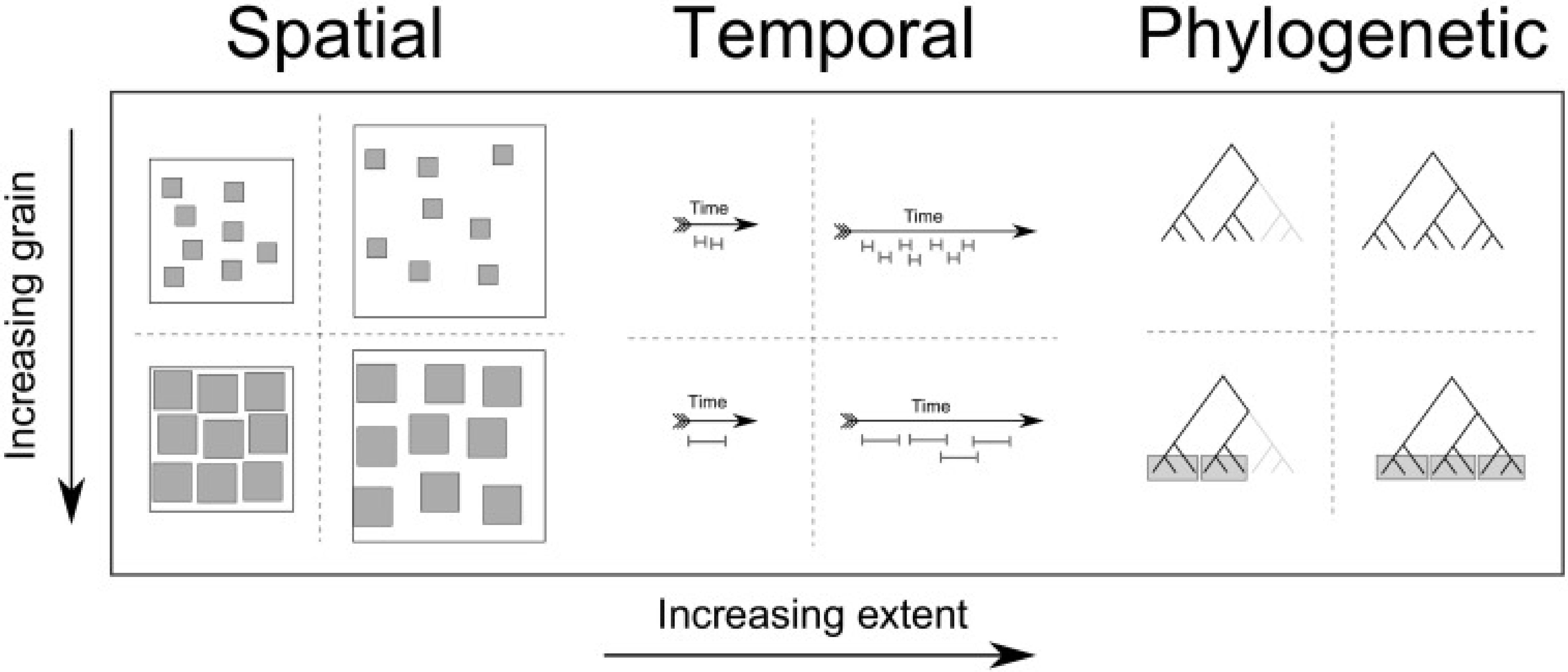
Re-use

Data Relevance

- Do the data directly address the analytical needs or purpose of the study?
- Are the variables sufficiently granular?
- Are the temporal and spatial scales appropriate?
- Do the data adequately represent the population, area, or system of interest?



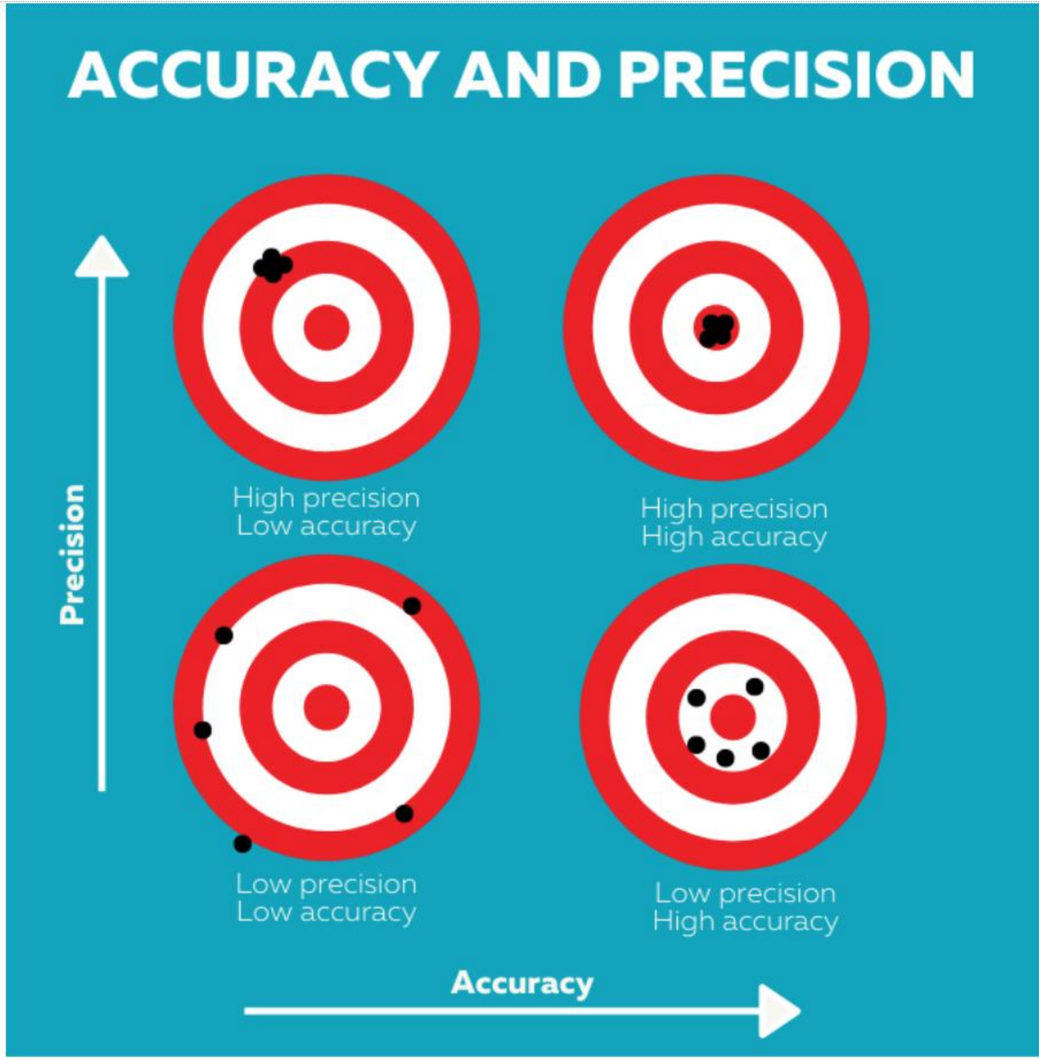
Frazier 2022 *Landscape Ecology*

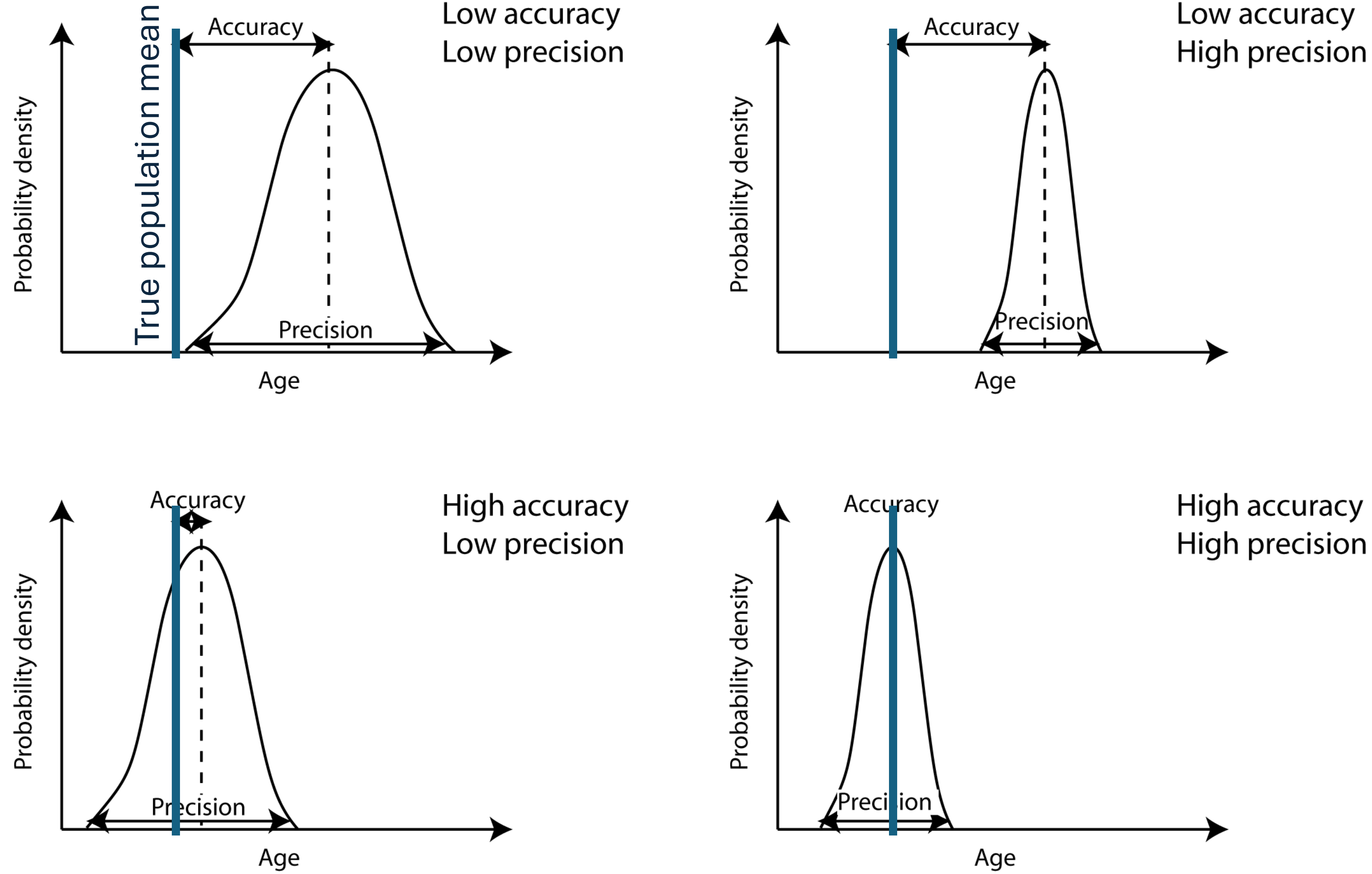


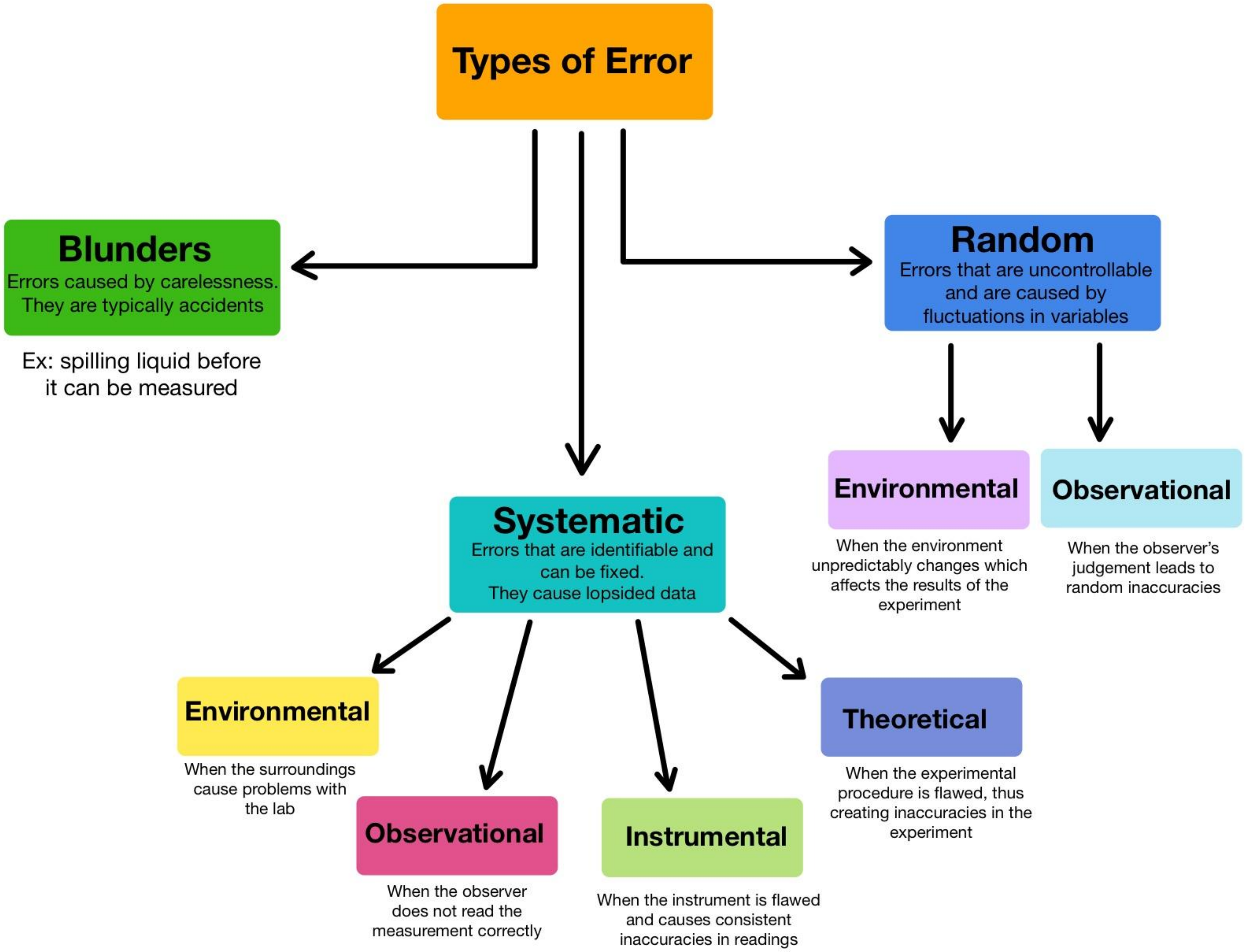
Trends in Microbiology

Data Quality and Reliability

- Are the data from a reputable organization or do they have a well-documented origin?
- Are the measurements collected precisely and accurately?
- Are there known sources of error?
- Are there missing values or large gaps in space and time?




True population mean
Accuracy
Low accuracy
Low precision
Probability density
Precision
Age
Accuracy
Low accuracy
High precision
Probability density
Precision
Age
Accuracy
High accuracy
Low precision
Probability density
Precision
Age
Accuracy
High accuracy
High precision
Probability density
Precision
Age



Types of Error

Blunders

Errors caused by carelessness. They are typically accidents

Ex: spilling liquid before it can be measured

Systematic

Errors that are identifiable and can be fixed.

They cause lopsided data

Environmental

When the surroundings cause problems with the lab

Observational

When the observer does not read the measurement correctly

Instrumental

When the instrument is flawed and causes consistent inaccuracies in readings

Theoretical

When the experimental procedure is flawed, thus creating inaccuracies in the experiment

Random

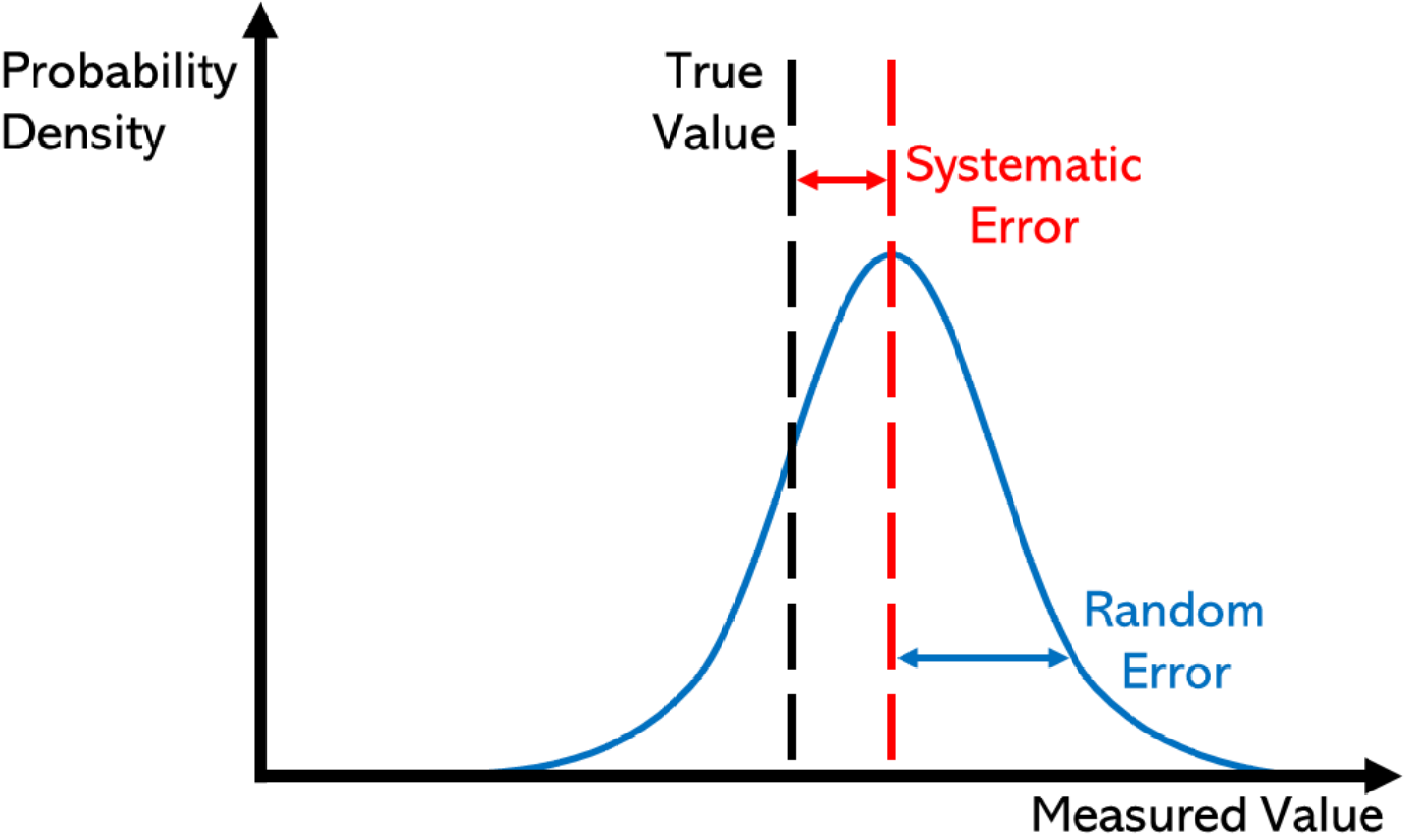
Errors that are uncontrollable and are caused by fluctuations in variables

Environmental

When the environment unpredictably changes which affects the results of the experiment

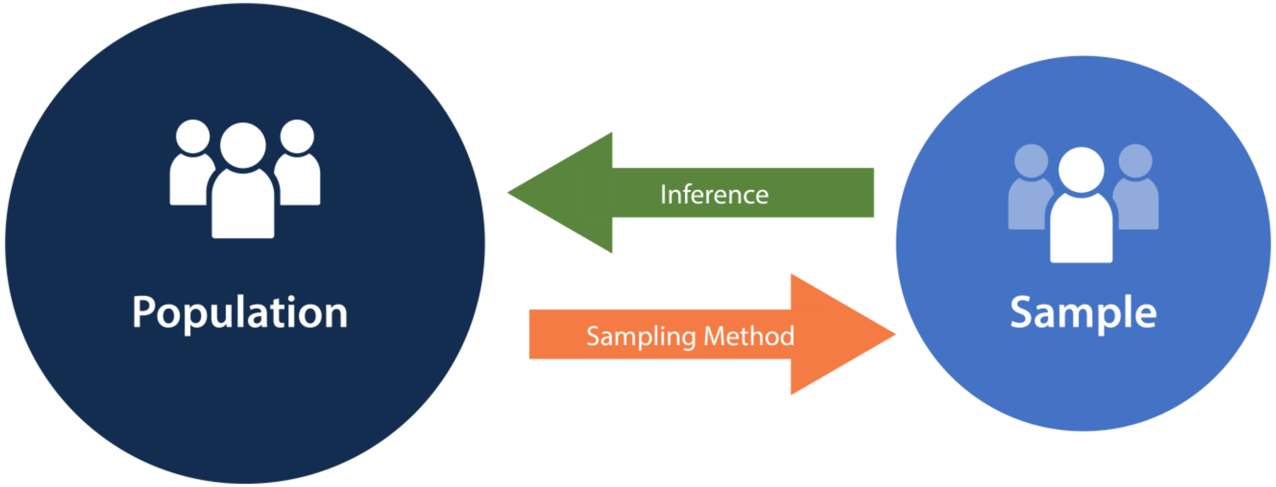
Observational

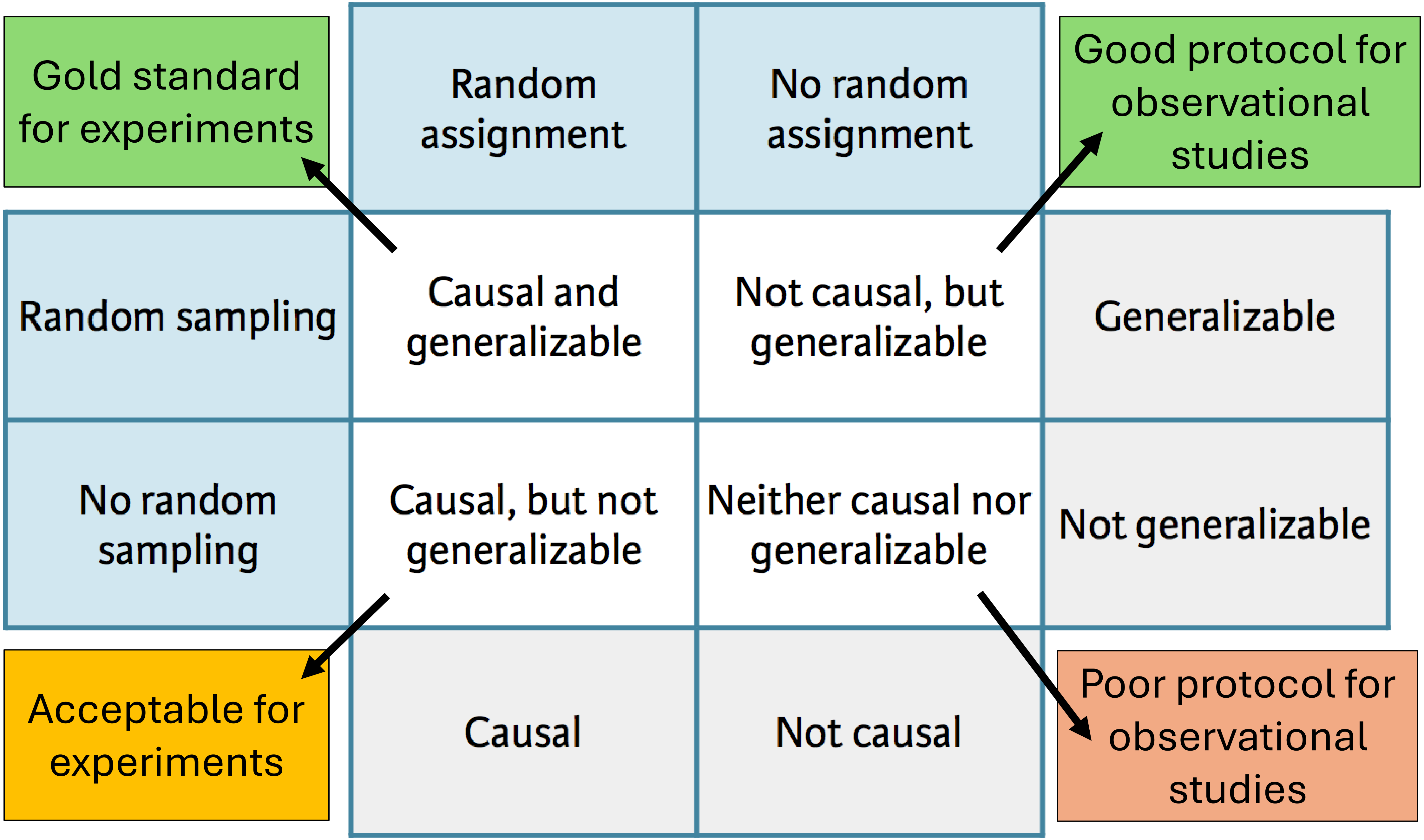
When the observer's judgement leads to random inaccuracies


Probability Density
True Value
Systematic Error
Random Error
Measured Value

Bias and Representativeness

- Is the sampling or treatment assignment strategy appropriate for the research questions?
- Do the data reflect a sample from a broader population?
- Were certain groups or locations more likely to be sampled; i.e., was there selection bias?





	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

Gold standard for experiments

Good protocol for observational studies

Acceptable for experiments

Poor protocol for observational studies

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

We established 260 experimental plots and assessed their community structure annually from 2006 to 2020. These plots were located in the low intertidal zone at 13 sites, which are nested on four capes in Oregon and northern California. At each site, we had 20 plots divided among 4 treatments in 5 replicate blocks. The treatments were control (no clearing), recovery (initial clearing), macrophyte-only (initial clearing, then repeated removal of sessile invertebrates) and invertebrate-only (initial clearing, then repeated removal of macrophytes).

Research sample

The samples were the surfgrasses, macroalgae, and sessile invertebrates present in the experimental plots per year. Plots are 25 x 25 cm and cover enough area to be representative of the low intertidal zone per site. The 13 sites are representative of rocky intertidal habitats in the northern California Current Large Marine Ecosystem.

Sampling strategy

We performed annual surveys of community structure in the experimental plots. A sample size of 20 plots per site is large enough to capture variation in the low intertidal zone community structure, and small enough to prevent excessive removal of organisms due to the experimental manipulations.

Data collection

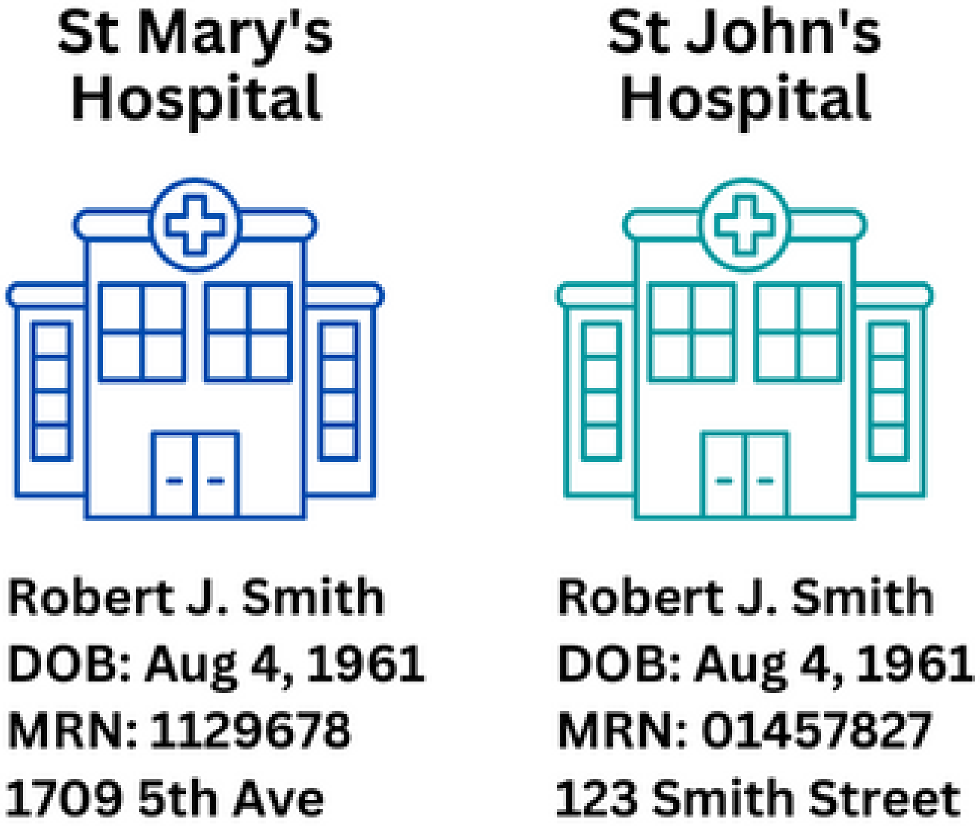
In our annual community surveys, we identified organisms to the lowest practical taxonomic rank (except coralline algae, usually species, but occasionally genus) and quantified their abundance in each plot. Abundances of sessile invertebrates, algal crusts, macrophytes, and substrate (bare rock and sand) were measured as percent cover. The authors conducted the vast majority of surveys analyzed in this article, with assistance from technicians and graduate students. We measured intertidal temperatures using temperature loggers anchored to the rock inside small stainless-steel cages. Finally, we assessed sea star wasting disease symptoms, counted Pisaster density, and measured size structure using belt transect surveys conducted by the researchers and their labs.

Timing and spatial scale

We conducted community surveys annually during the spring and summer from August 2006 to August 2020. Rocky intertidal species are relatively slow growing; thus annual surveys adequately captured community dynamics. We conducted 3,703 of 3,900 possible surveys (95%), with most of the missing surveys from the southernmost field site that was difficult to visit. Additionally, some surveys could not be conducted due to dangerous wave conditions and five plots were lost due to rock breaking off. Please see Supplementary Table 6 and Supplementary Figs. 1-16 for which surveys were missing. The field sites span 650 km of coastline and each site has 20 experimental plots spread over 25-100 m of the low intertidal zone. Sea star belt transects were conducted near the experimental plots one to three times annually during spring and summer 2006-2021.

Ethics and Permissions

- For subjects, were ethical standards met (e.g., IRB for humans, IACUC for vertebrates)?
- Are data anonymized if necessary?
- Are there restrictions on how the data can be used, shared, or published?



St Mary's Hospital

Robert J. Smith
DOB: Aug 4, 1961
MRN: 1129678
1709 5th Ave

St John's Hospital

Robert J. Smith
DOB: Aug 4, 1961
MRN: 01457827
123 Smith Street

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This information has not been collected because our research does not involve human participants.
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Publicly-available Environmental Datasets

- US Government Open Data (https://data.gov/)
  - Archive of Data.gov (https://source.coop/repositories/harvard-lil/gov-data/description)
- Environmental Data Initiative (https://edirepository.org/).
- Data Observation Network for Earth (https://www.dataone.org/)
- NASA Earth Science Data Systems (https://www.earthdata.nasa.gov/)
- Historical climate data from Canada (https://climate.weather.gc.ca/)

Writing Quality Metadata

- **Metadata** – data about data
- Good data documentation includes:
  - The context of why and how the data were collected
  - The structure of the data, including how files relate to each other
  - A data dictionary used to catalog and provide meaningful descriptions for individually named data objects
  - Quality assurance that data are complete and accurate
  - Information on data confidentiality, access, and use conditions
  - Identification and tracking of different versions of datasets

Metadata Examples

- Best practices for metadata creation from USGS

How can these metadata be improved?

- Title: Avian point count surveys on Steens Mountain
  - Method: Estimated birds up to 40 ms away for 5 mins
  - Location: Arroyo near Riddle Brothers Ranch
  - Time: April 2022
  - Species: Passerines including *L. ludovicianus*
  - Sampling strategy: Audio and visual detection
  - Sample size: 200