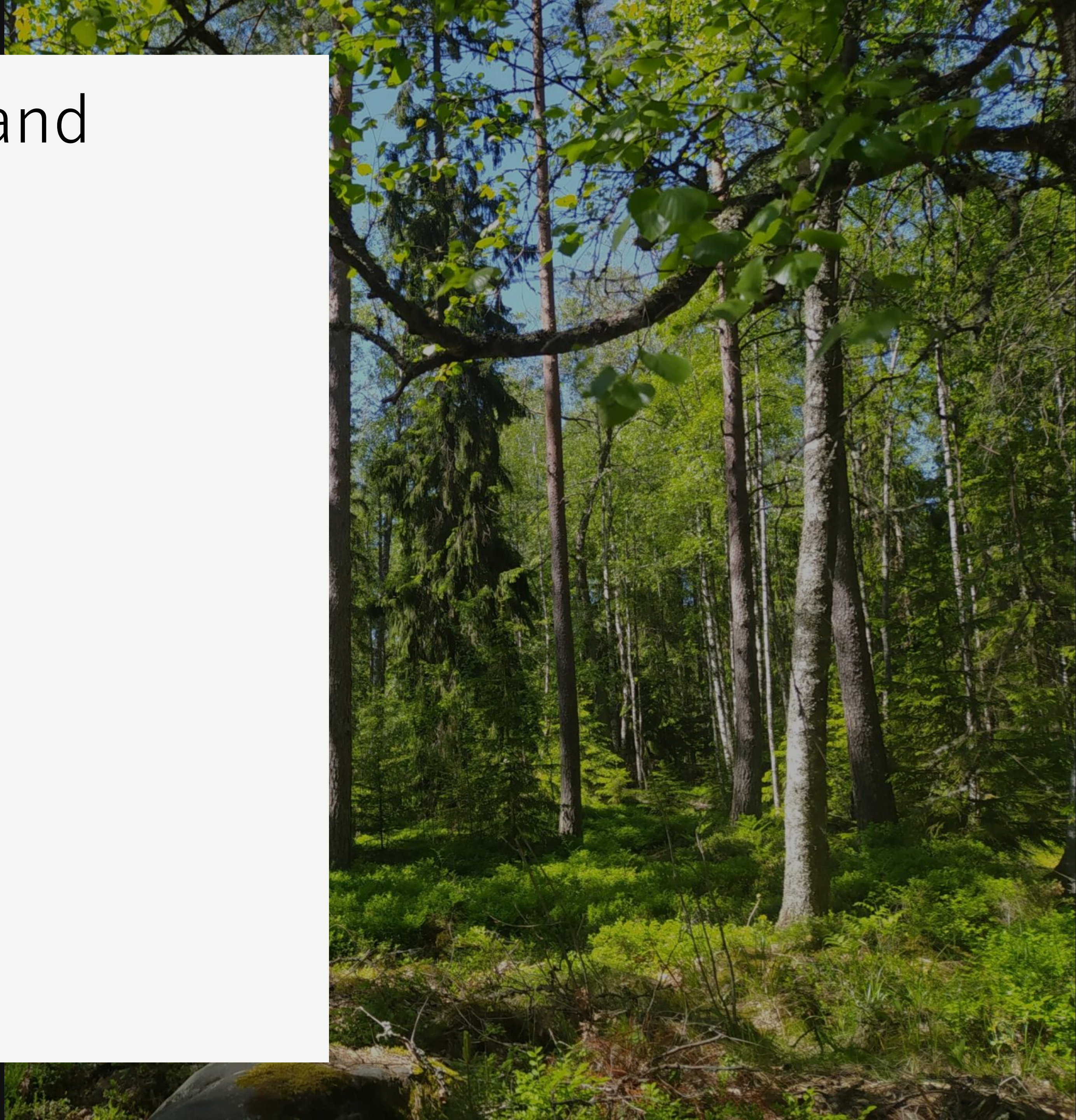
FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

Lecture 1: Data Screening and Exploration

Tuesday, October 1, 2024

Lecture 1: Data Screening and Exploration

- Data Screening
- Exploratory Analysis

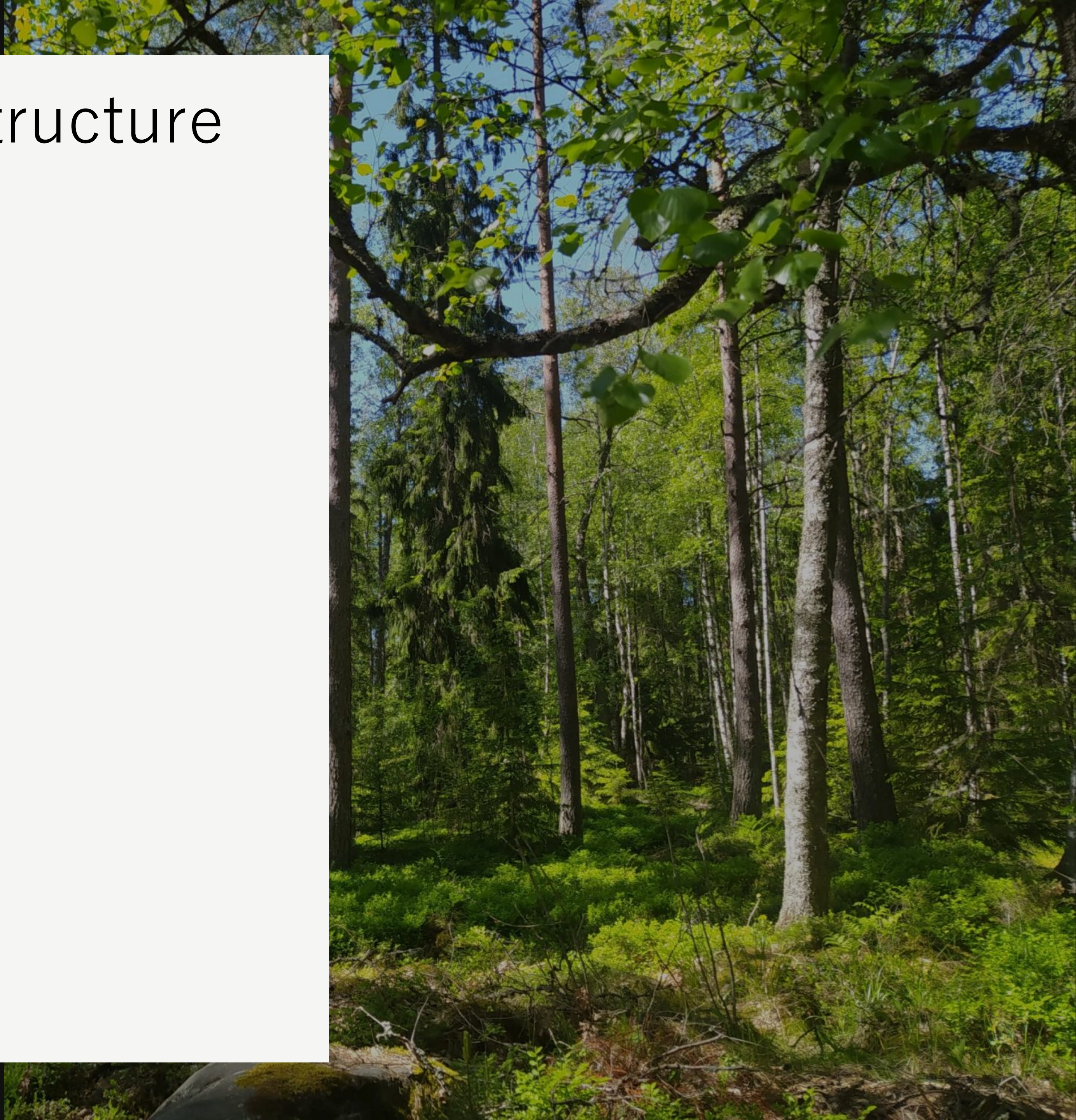


Lecture 2: Data Screening and Exploration

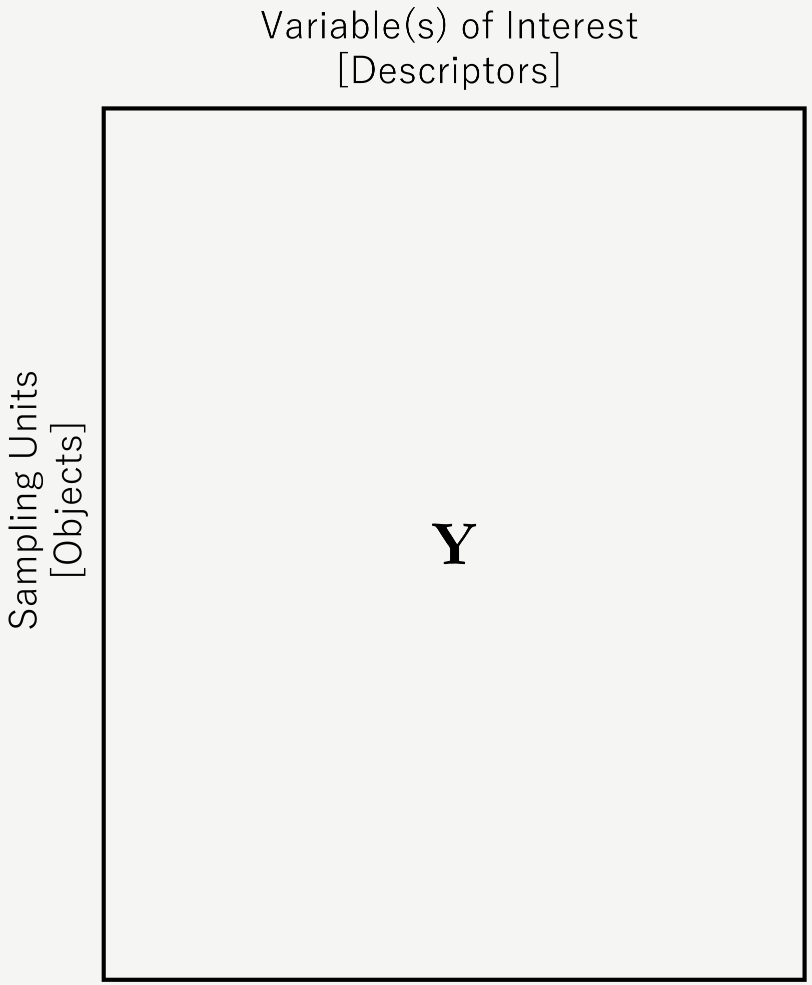
- Data Screening
- Exploratory Analysis

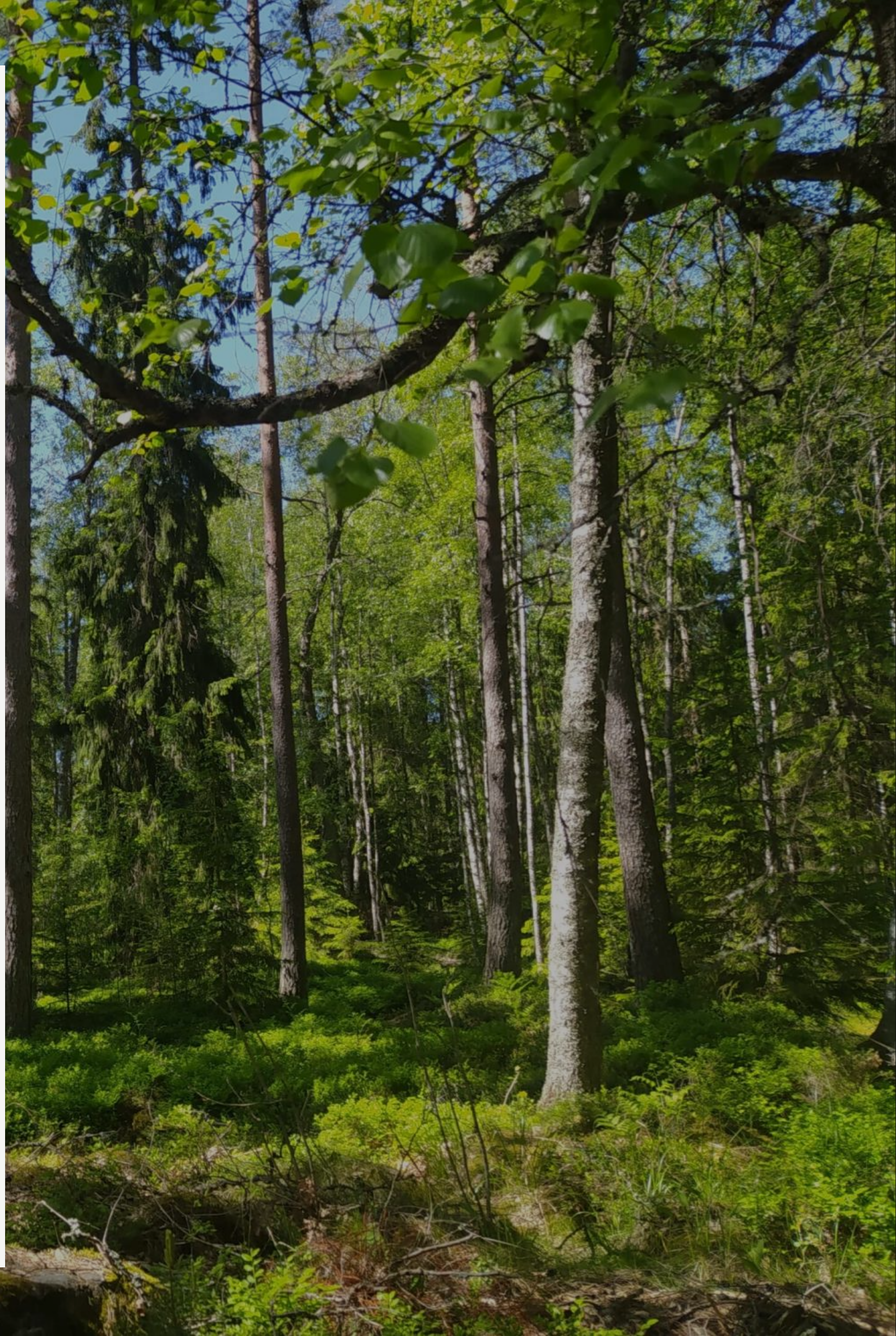
Sorry, you're taking a stats class!!!

Recap: Multivariate Data Structure

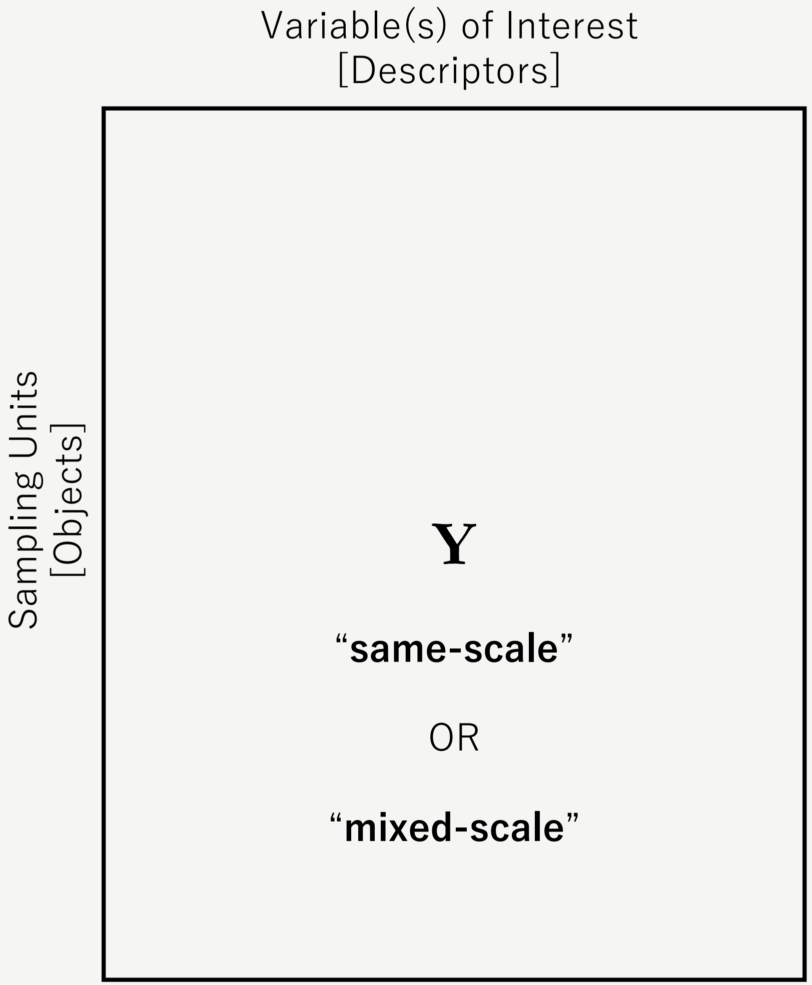


Recap: Multivariate Data Structure

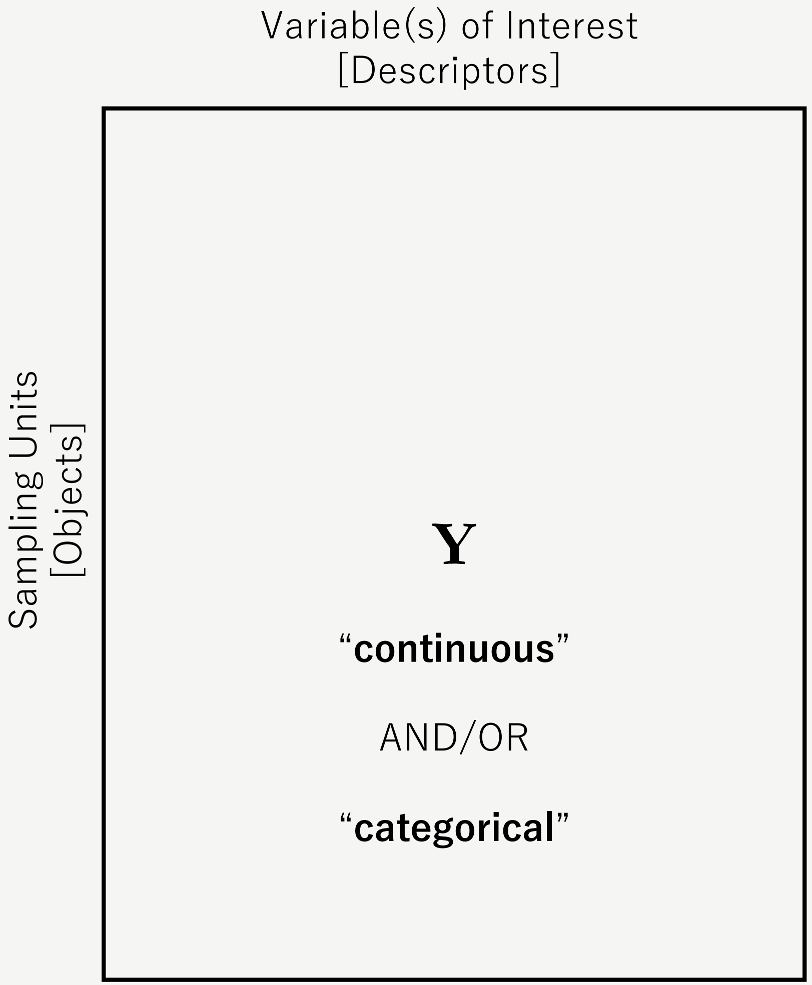




Recap: Multivariate Data Structure



Recap: Multivariate Data Structure



Variable(s) of Interest
[Descriptors]

Sampling Units
[Objects]

Y

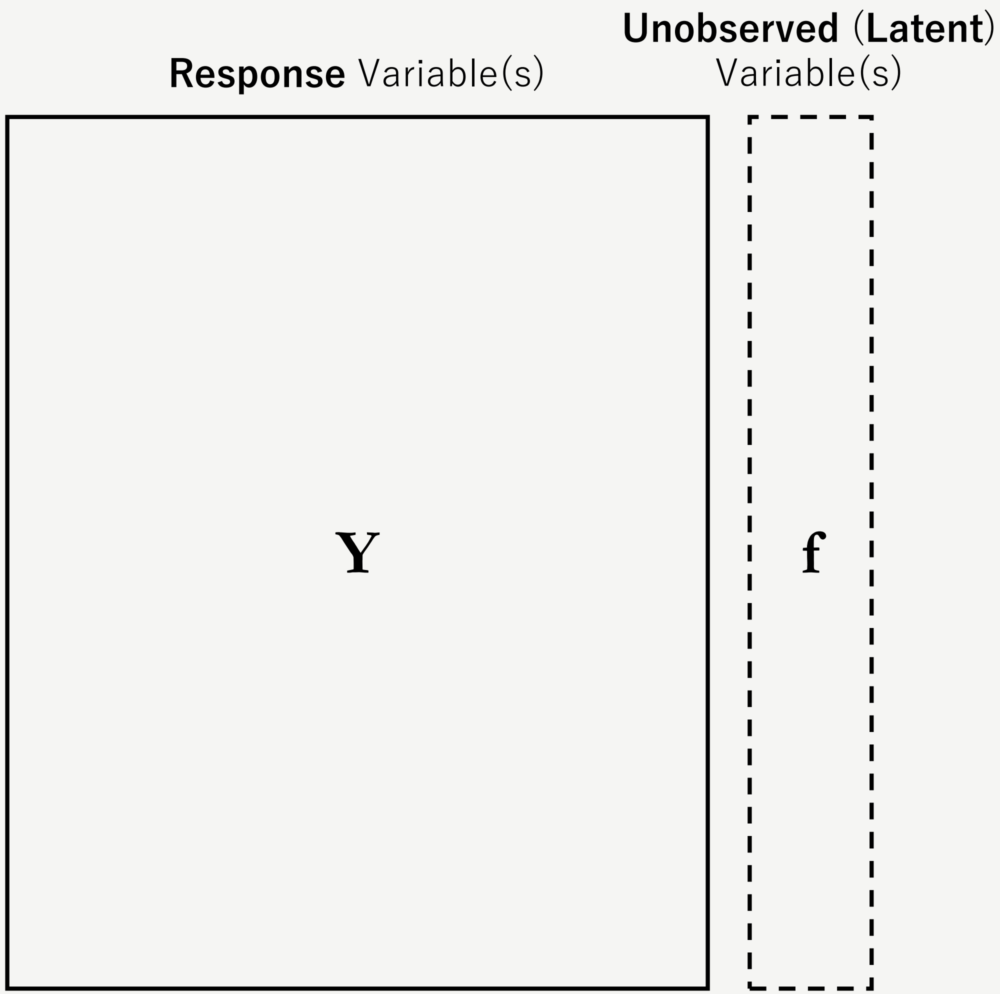
“**continuous**”

AND/OR

“**categorical**”

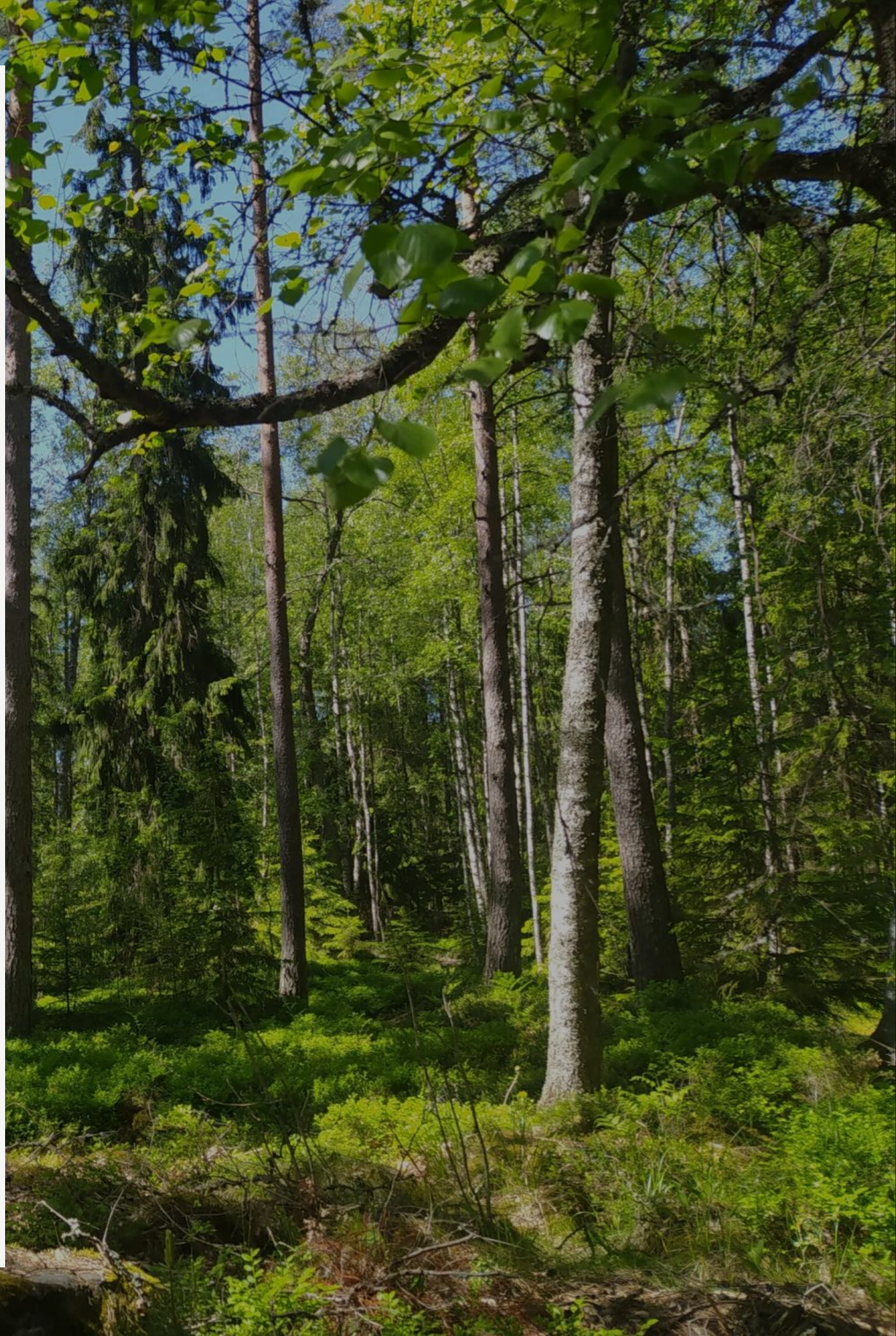
Recap: Multivariate Data Structure

Structural Methods: look for structure underlying the data matrix **Y**.



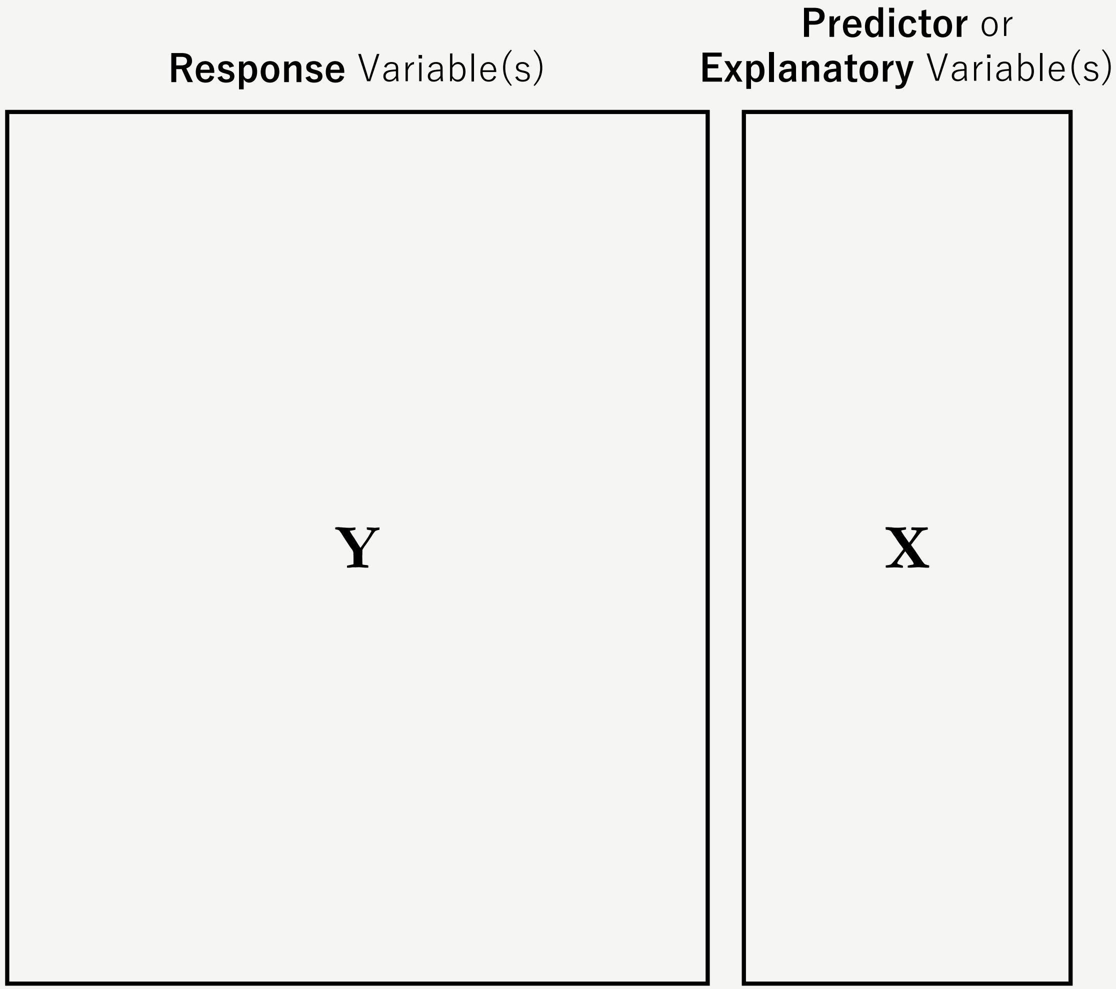
Response Variable(s)

Unobserved (**Latent**) Variable(s)



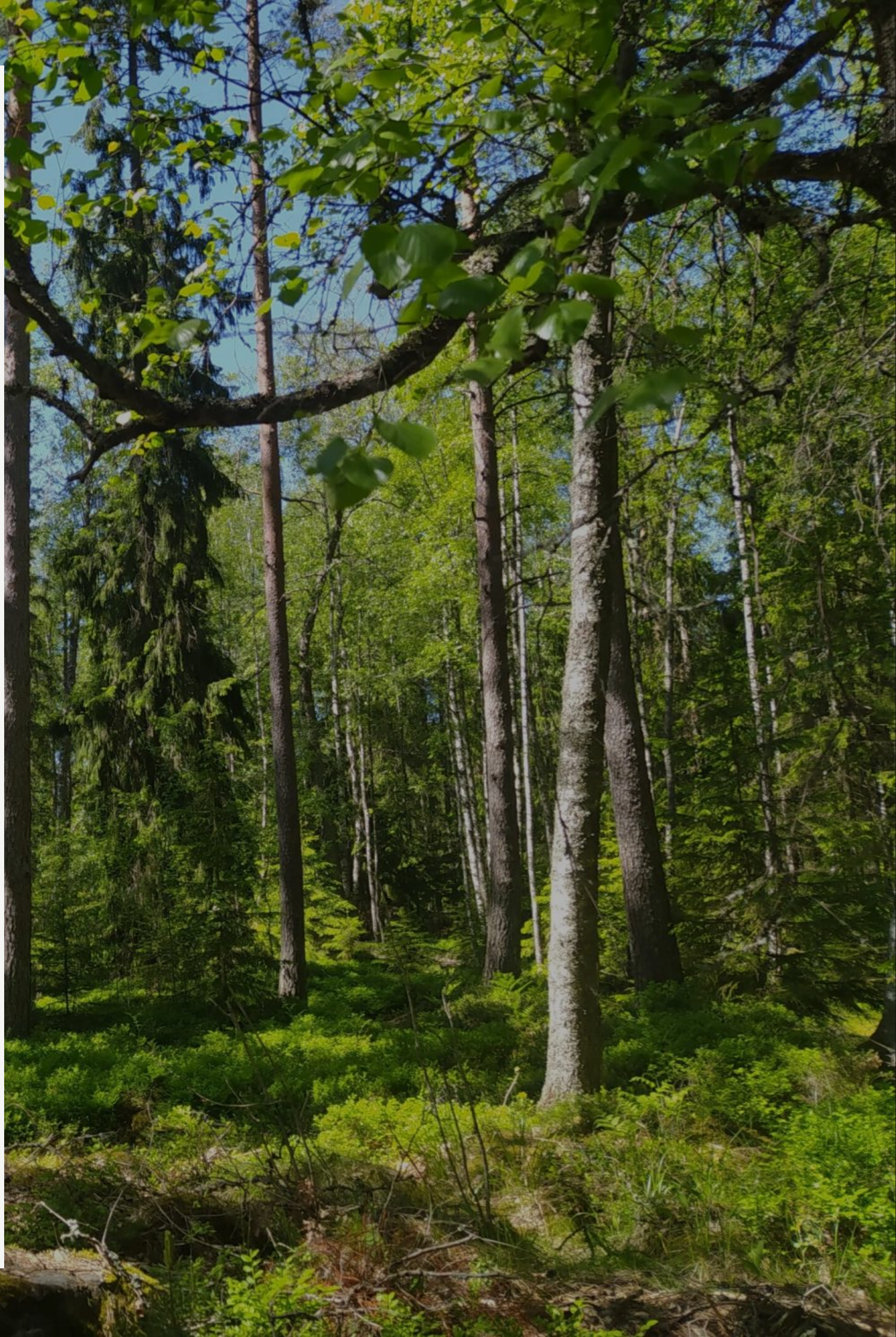
Recap: Multivariate Data Structure

Functional Methods: relate the response variable(s) **Y** as a function of the predictor variable(s) **X**.

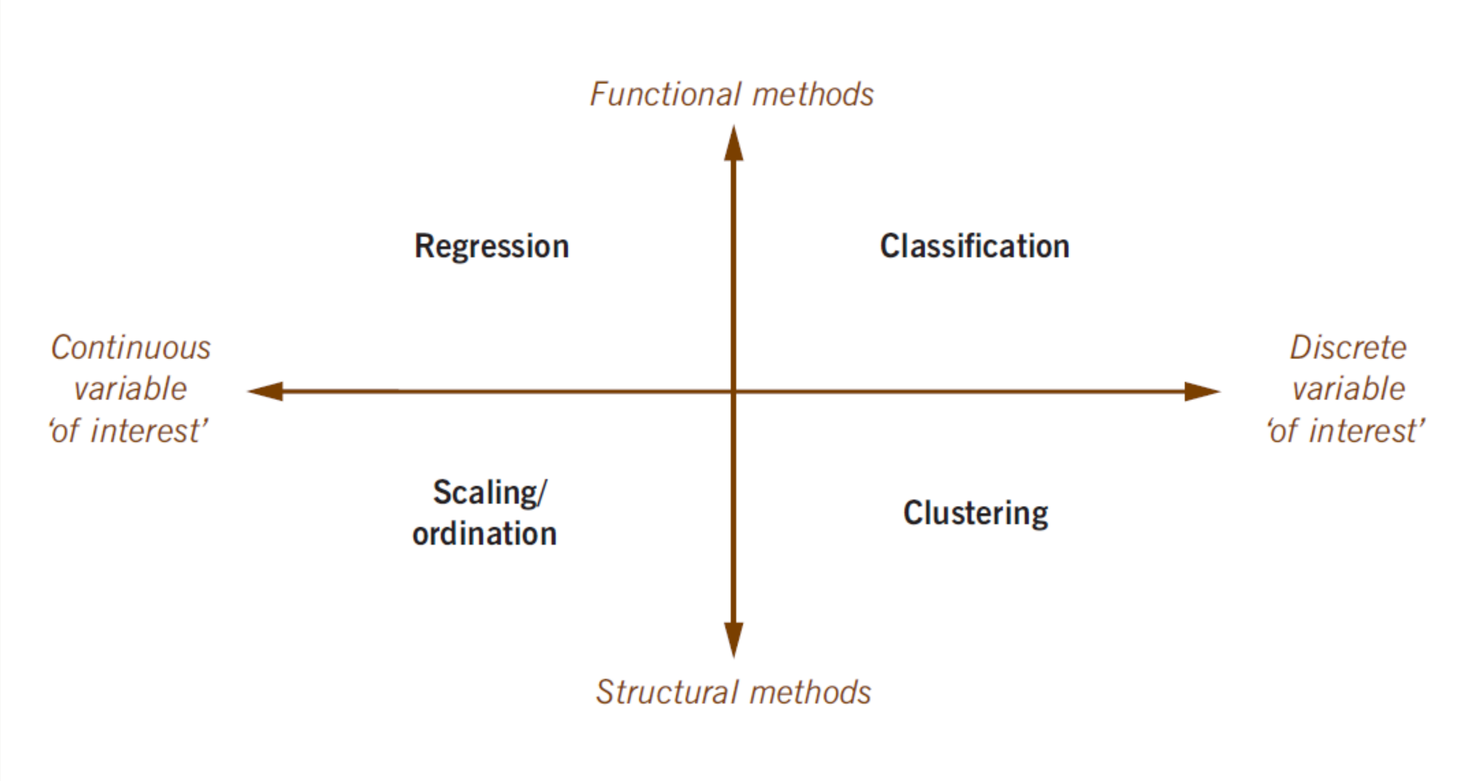


Response Variable(s)

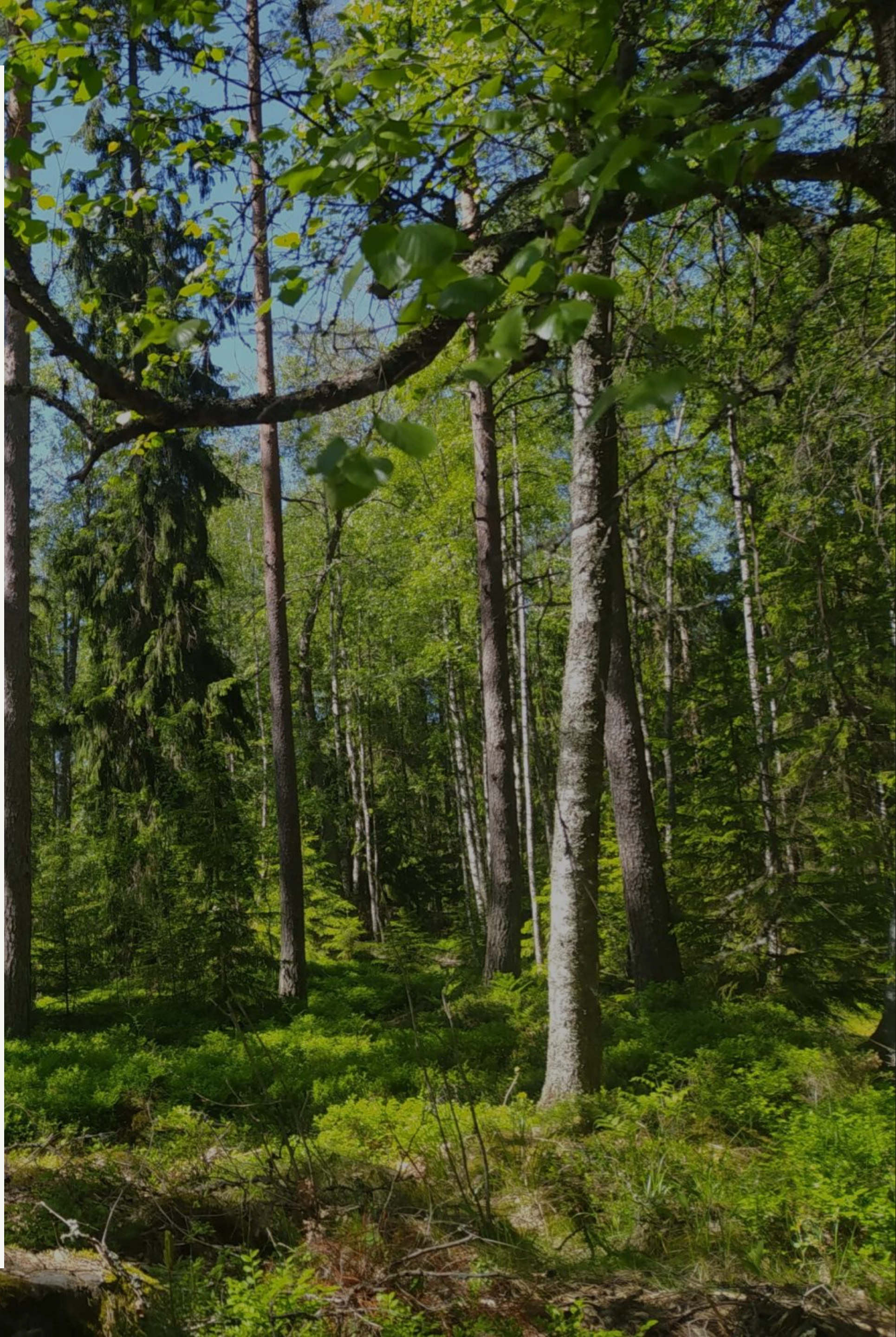
Predictor or **Explanatory** Variable(s)



Recap: Multivariate Data Structure

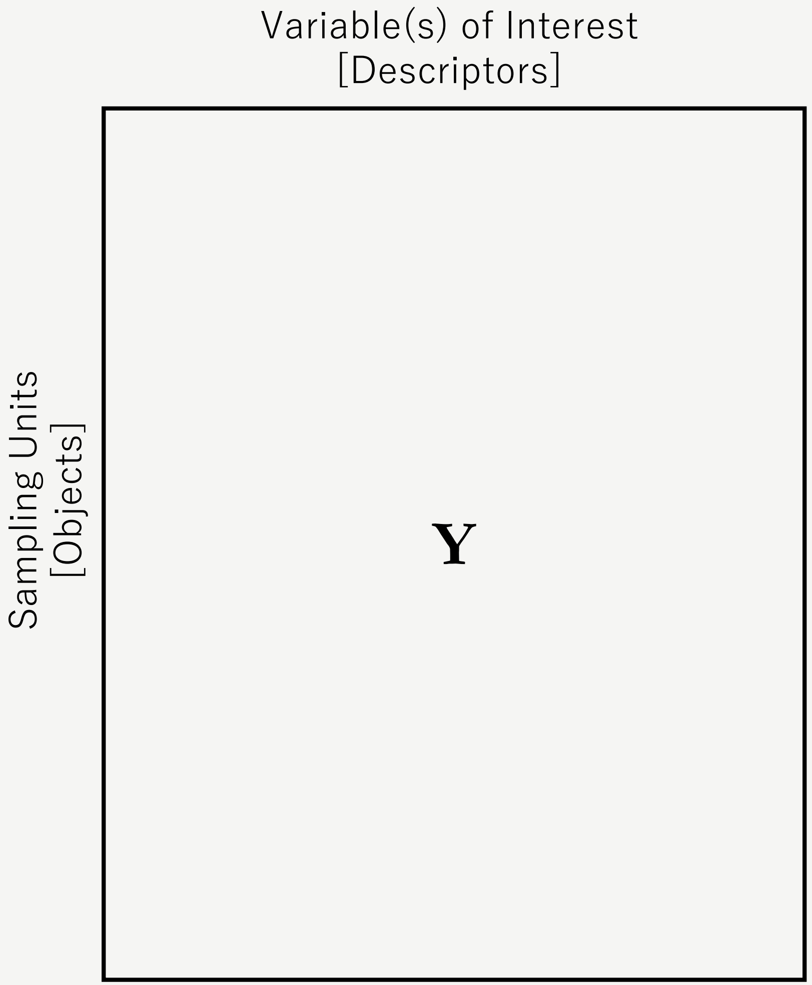


Greenacre and Primicerio



Data Screening

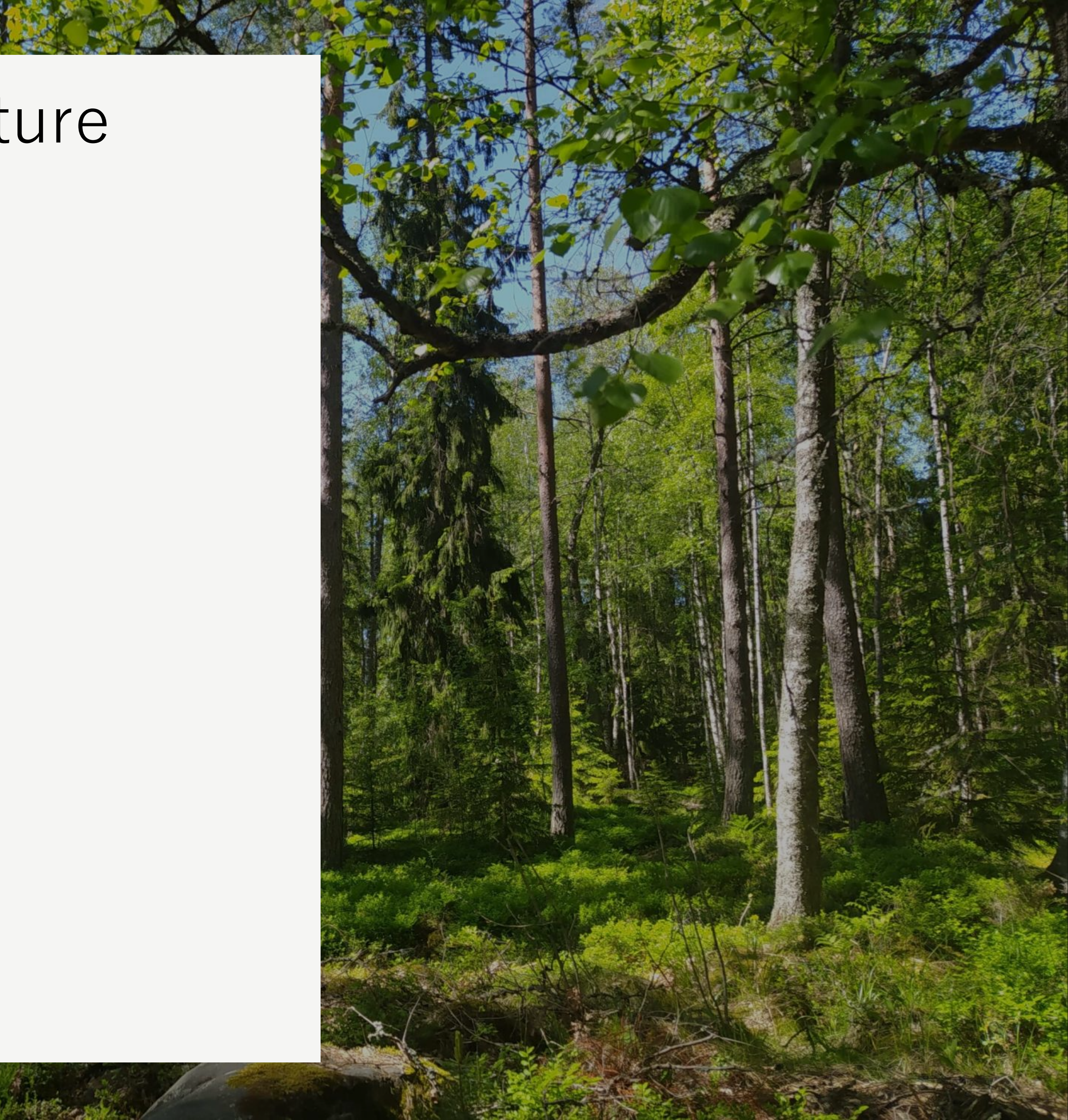
Data Screening: Data Structure



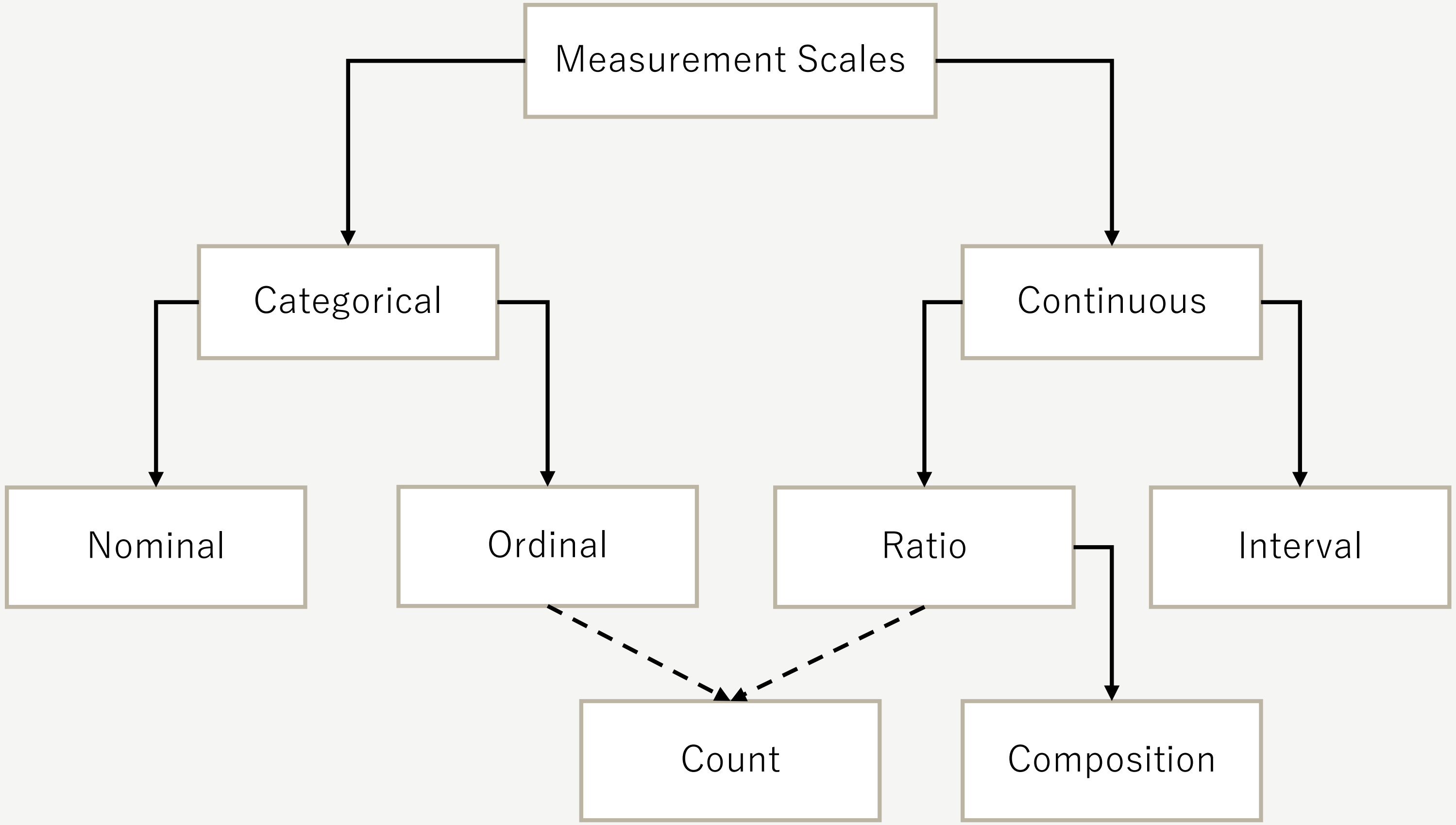
Variable(s) of Interest
[Descriptors]

Sampling Units
[Objects]

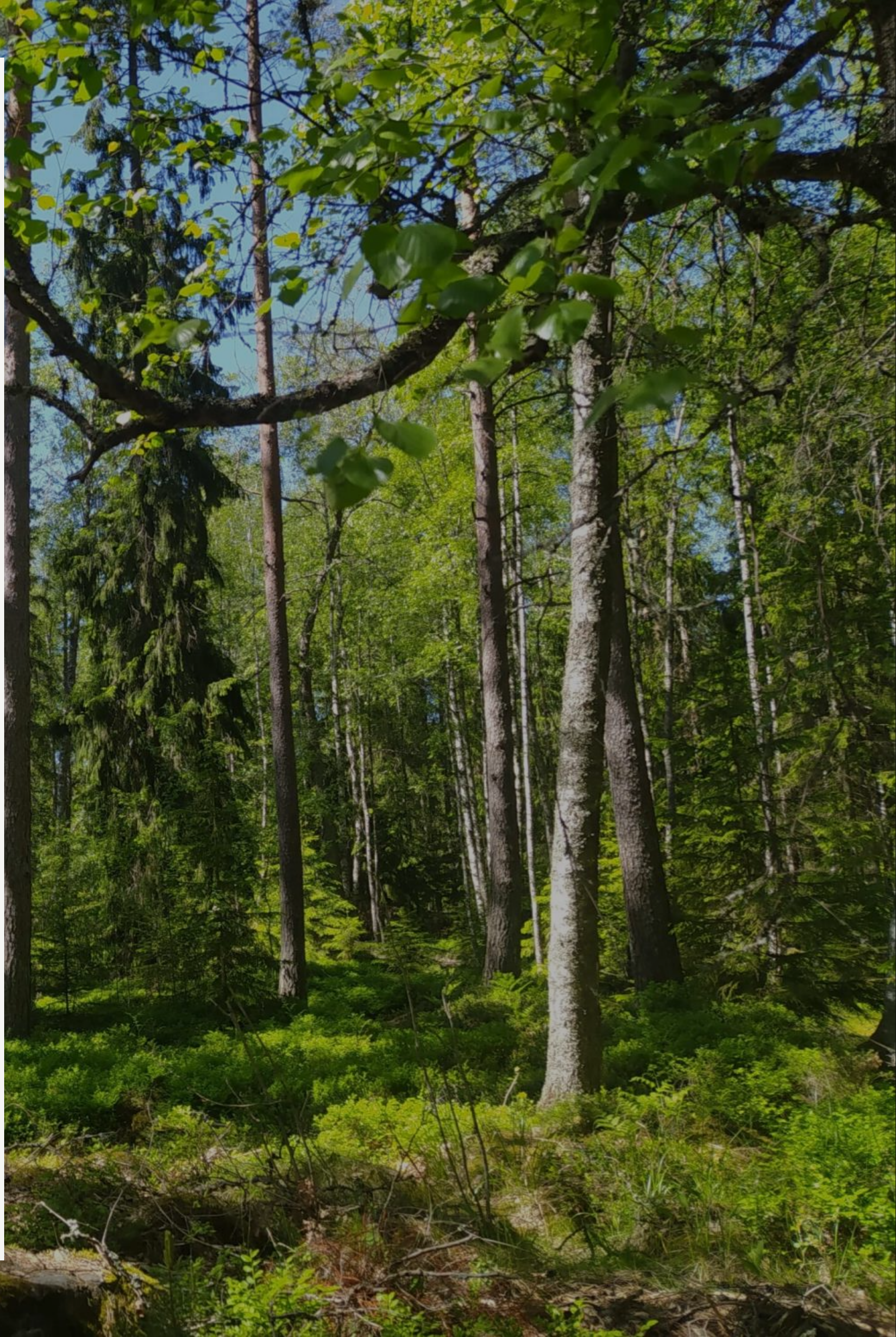
$\mathbf{Y}$



Data Screening: Data Structure

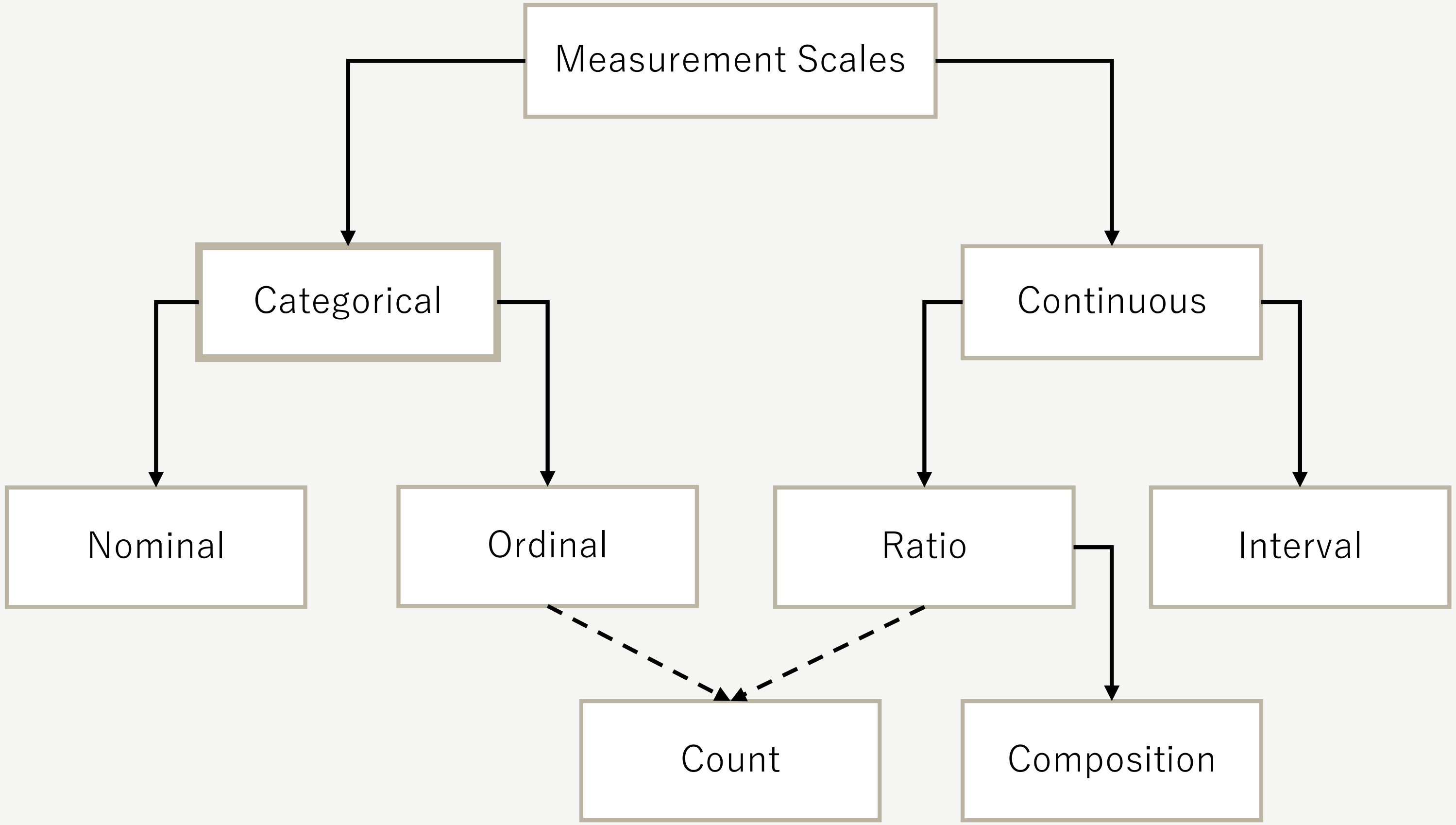


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2



Data Screening: Data Structure

Categorical Data: Have been *discretized* or divided into groups



Measurement Scales

Categorical

Continuous

Nominal

Ordinal

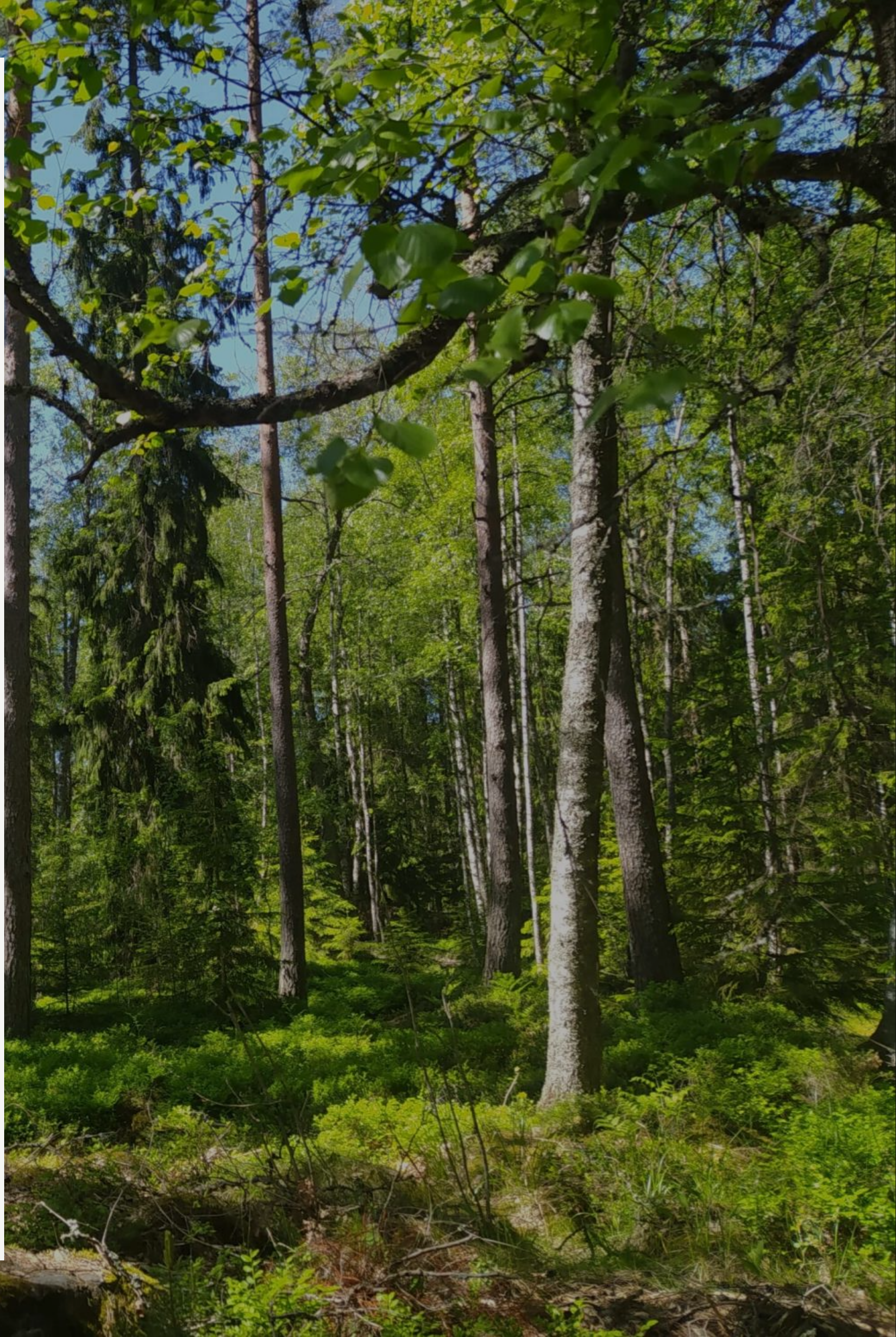
Ratio

Interval

Count

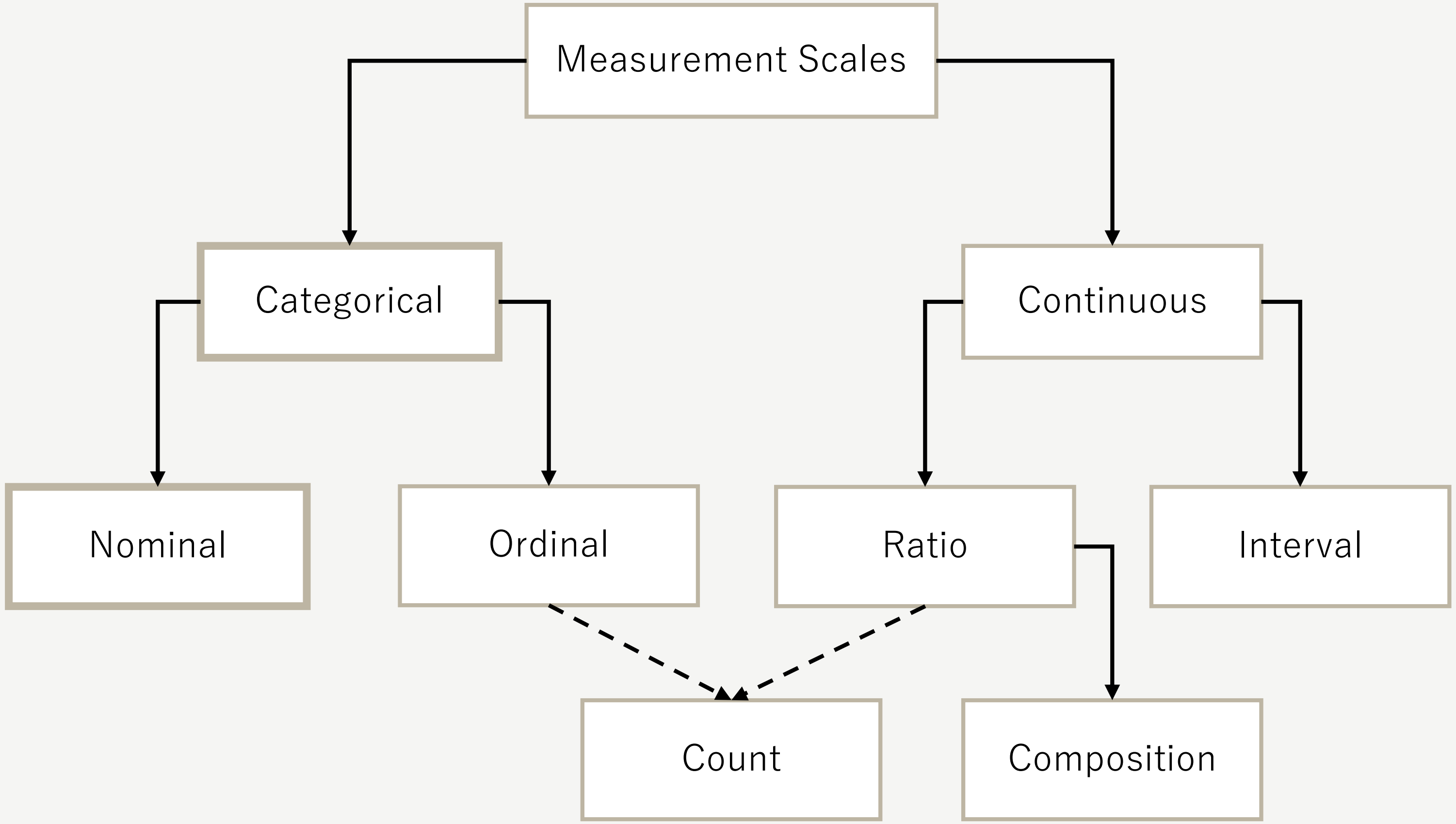
Composition

Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

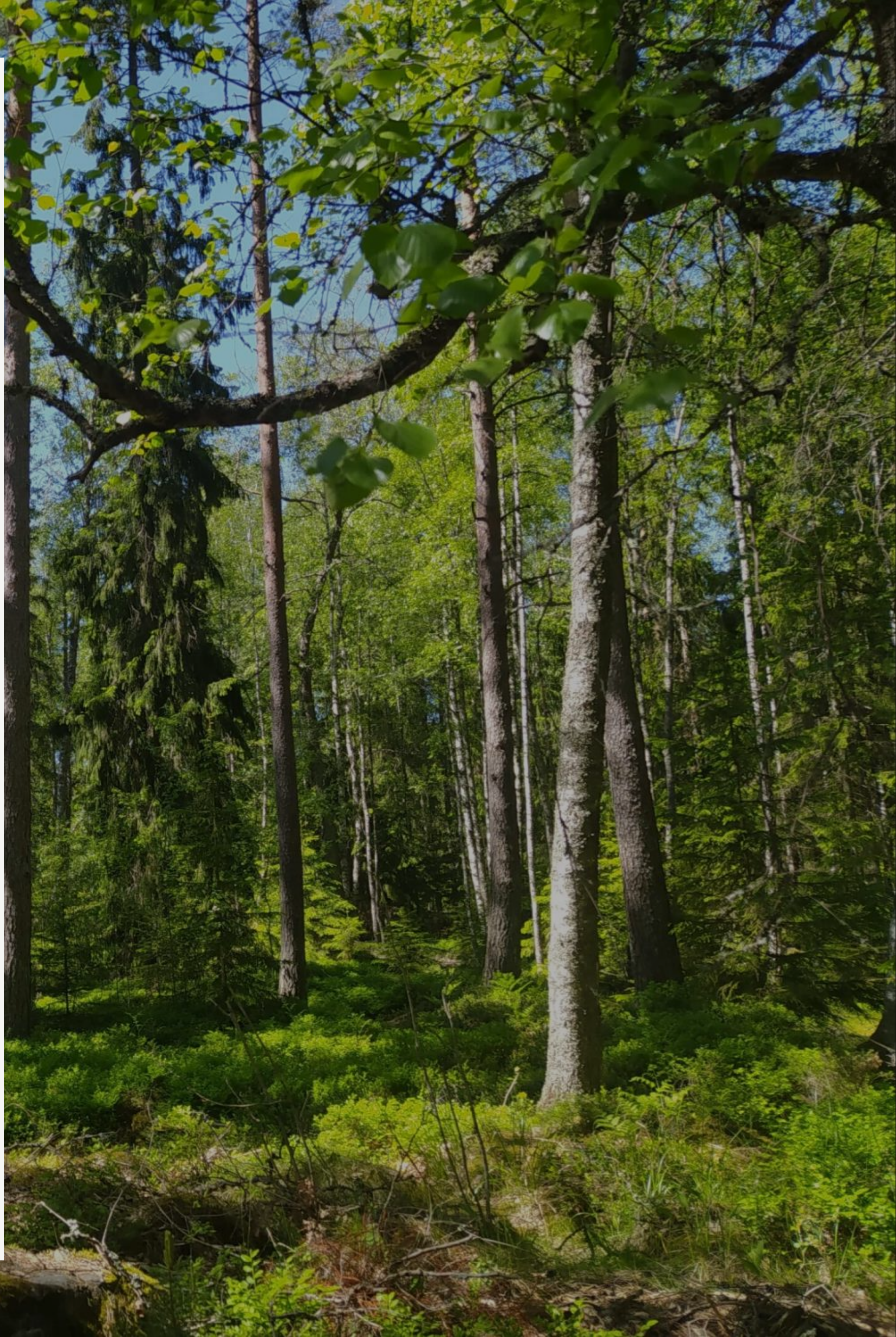


Data Screening: Data Structure

Nominal Categories: Have no ordering; e.g., region or habitat type

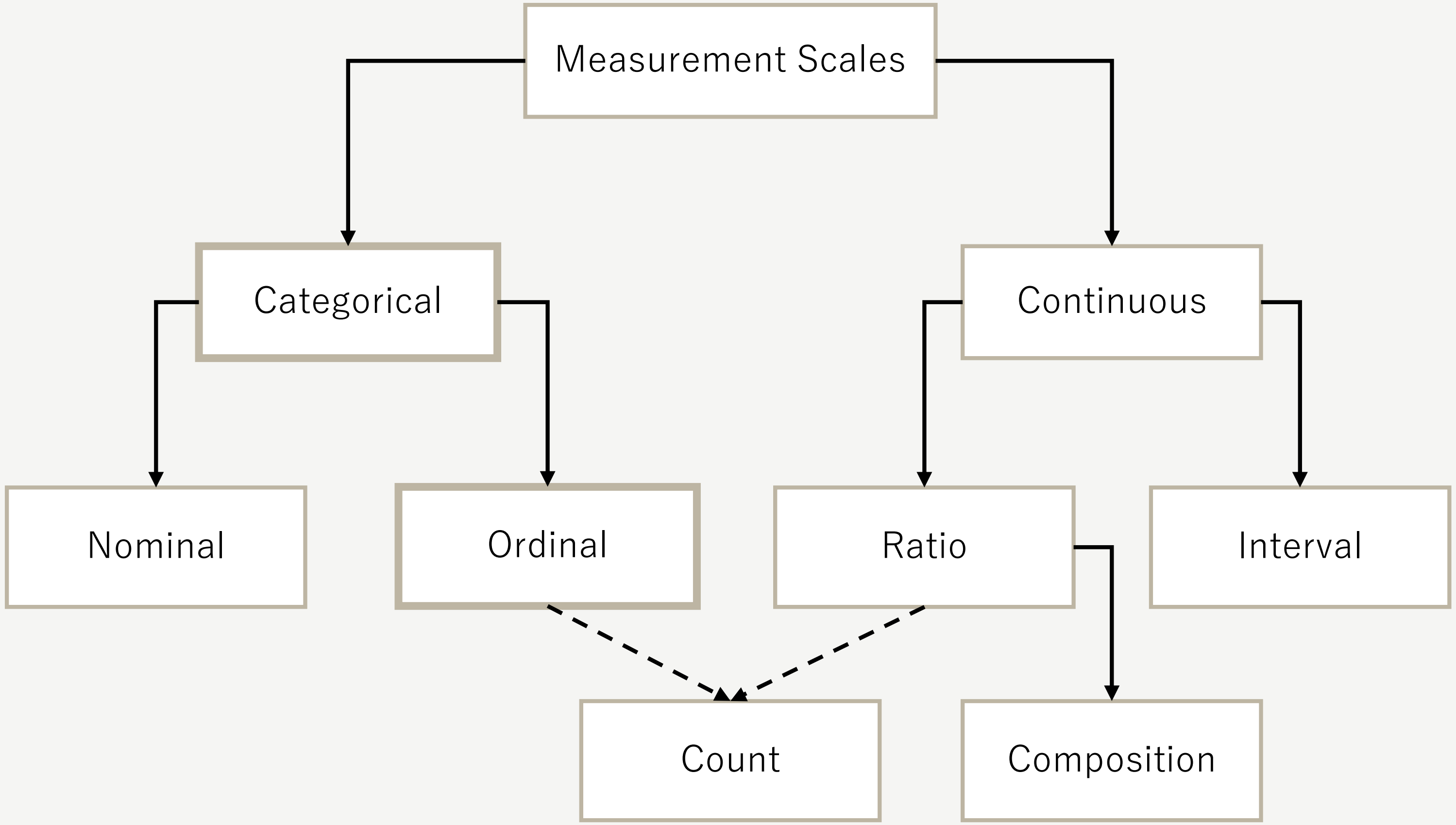


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

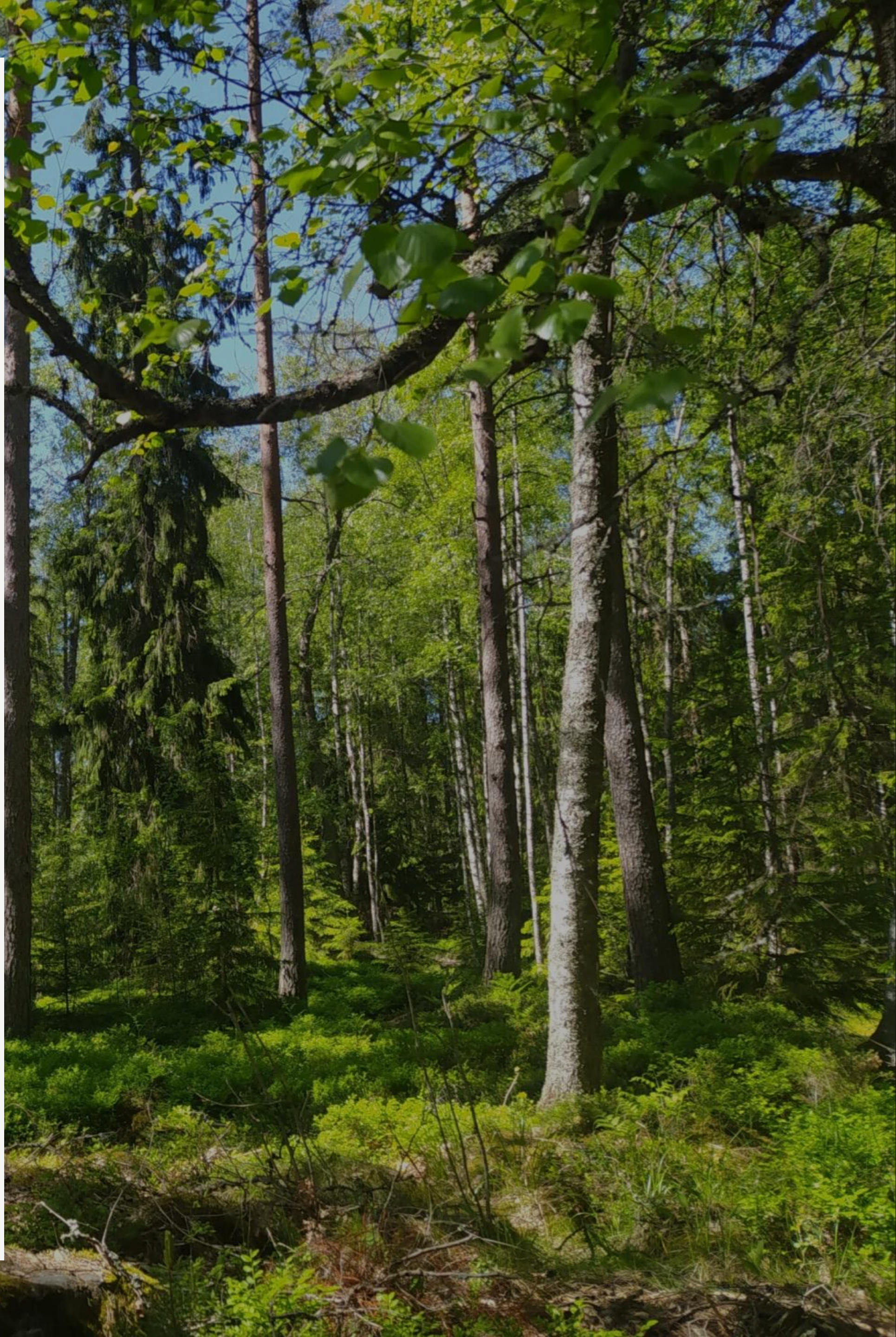


Data Screening: Data Structure

Ordinal Categories: Have an order; e.g., month



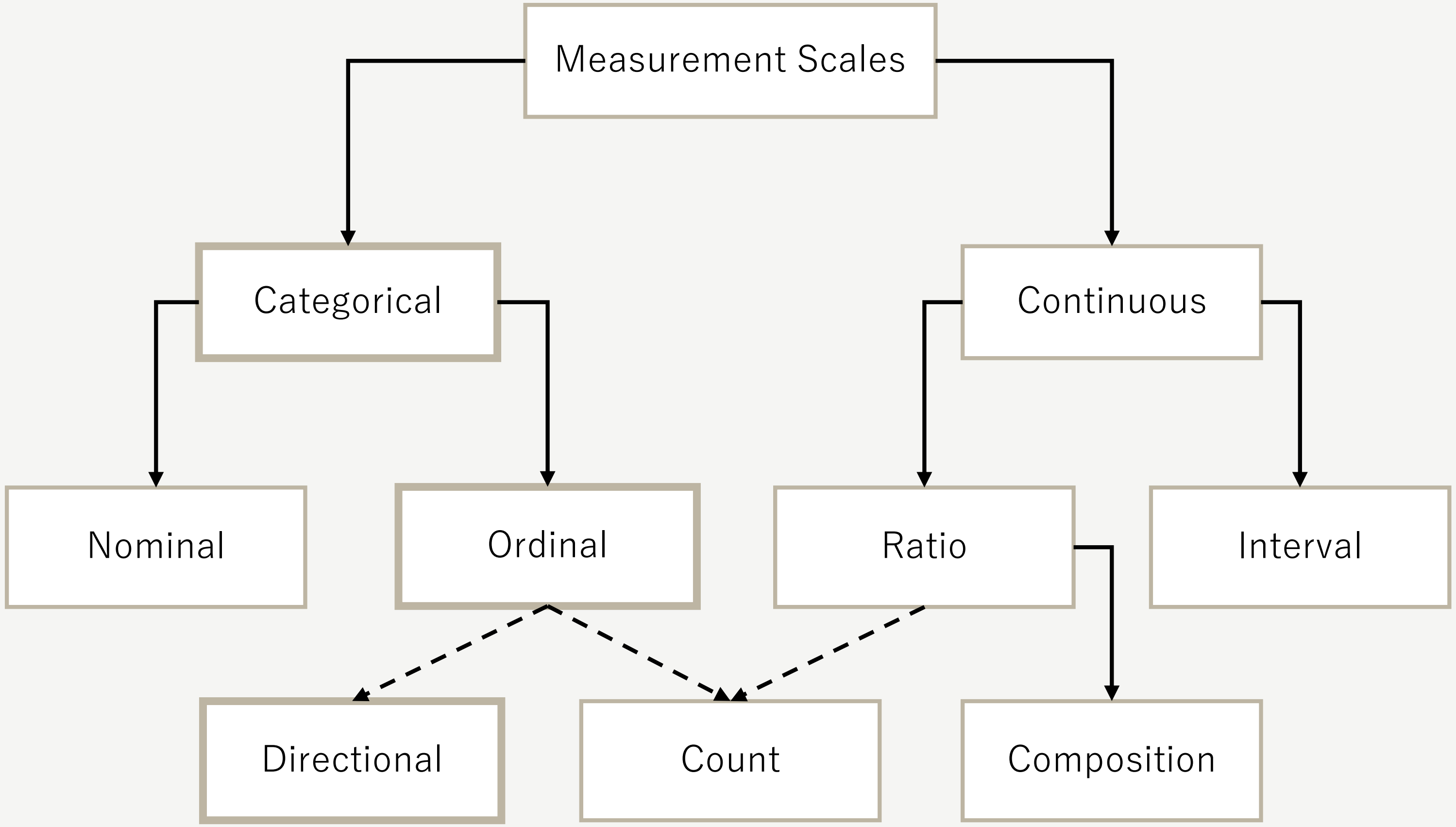
Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2



Data Screening: Data Structure

Ordinal Categories: Have an order; e.g., month

Directional data are a special case, “circular” ordering of categories



Measurement Scales

Categorical

Continuous

Nominal

Ordinal

Ratio

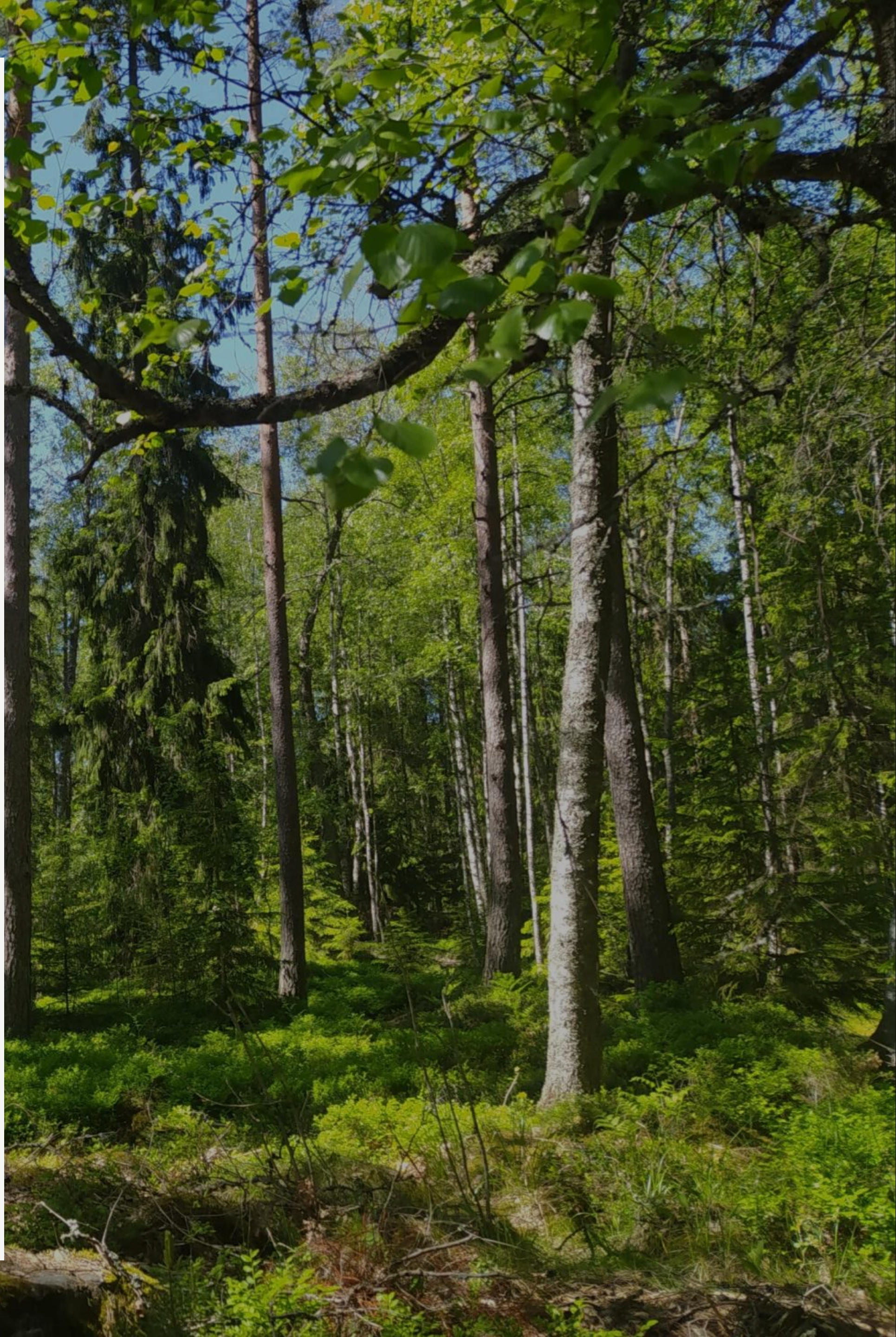
Interval

Directional

Count

Composition

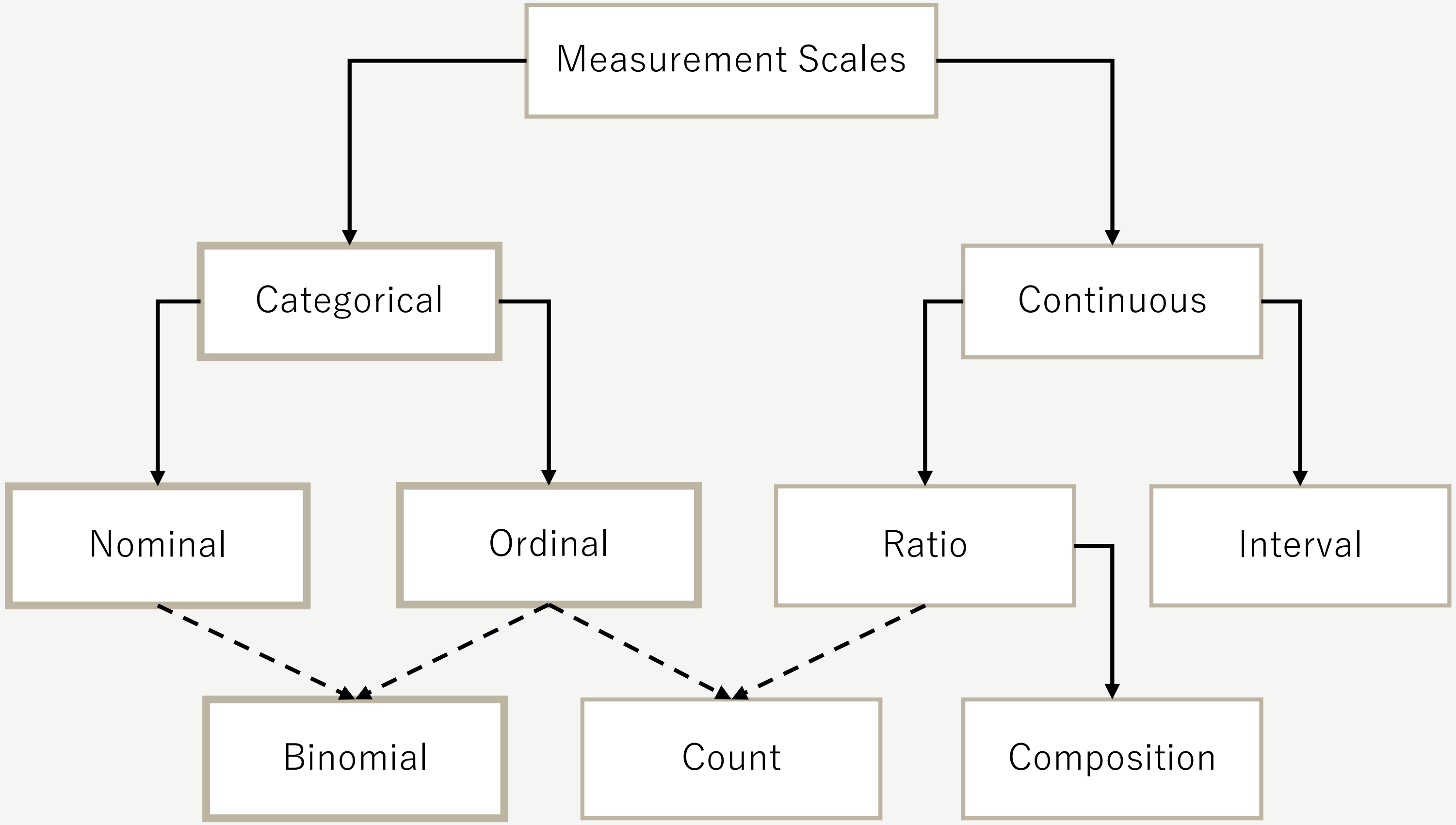
Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2



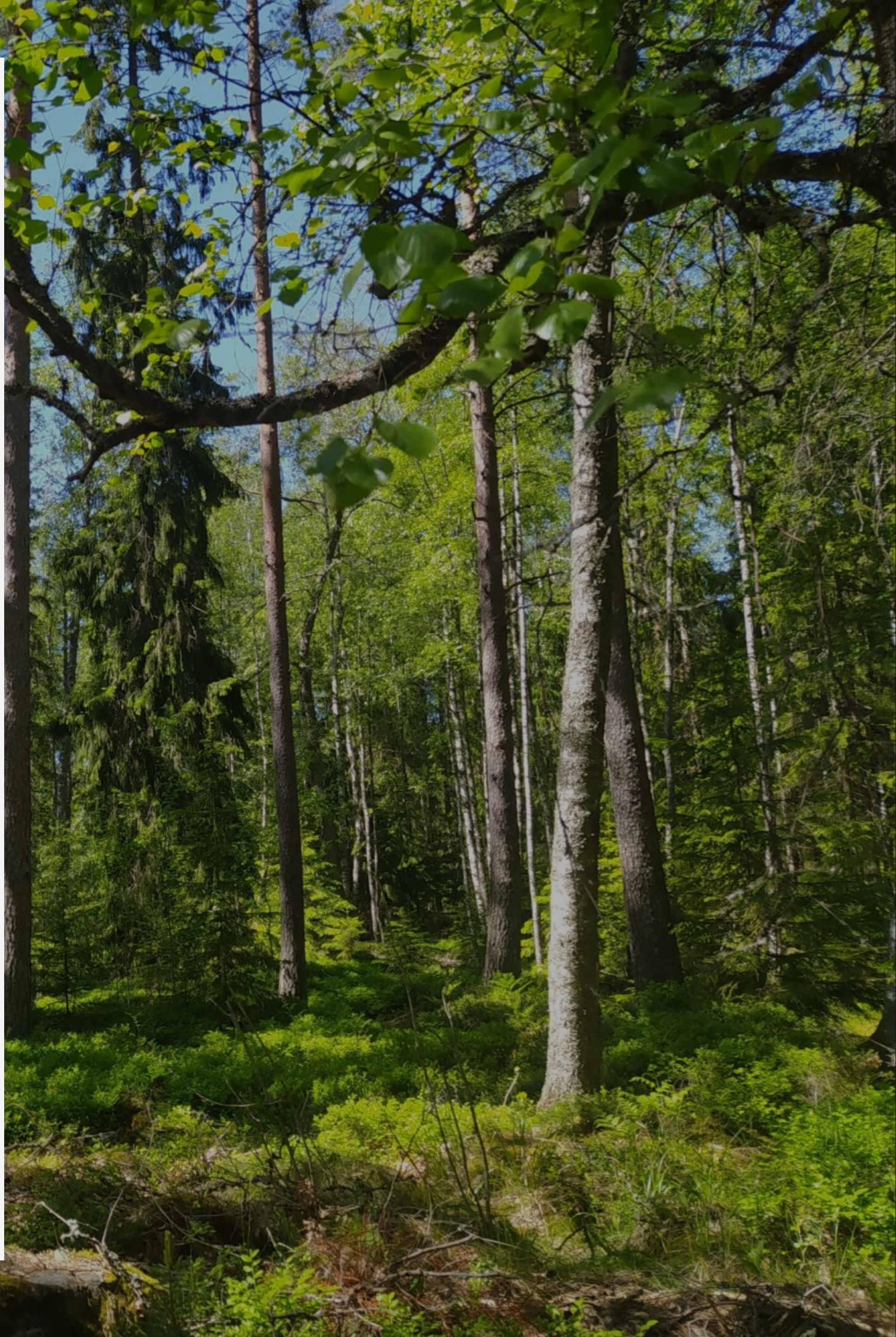
Data Screening: Data Structure

Ordinal Categories: Have an order; e.g., month

Binomial data can be nominal or ordinal depending on variable

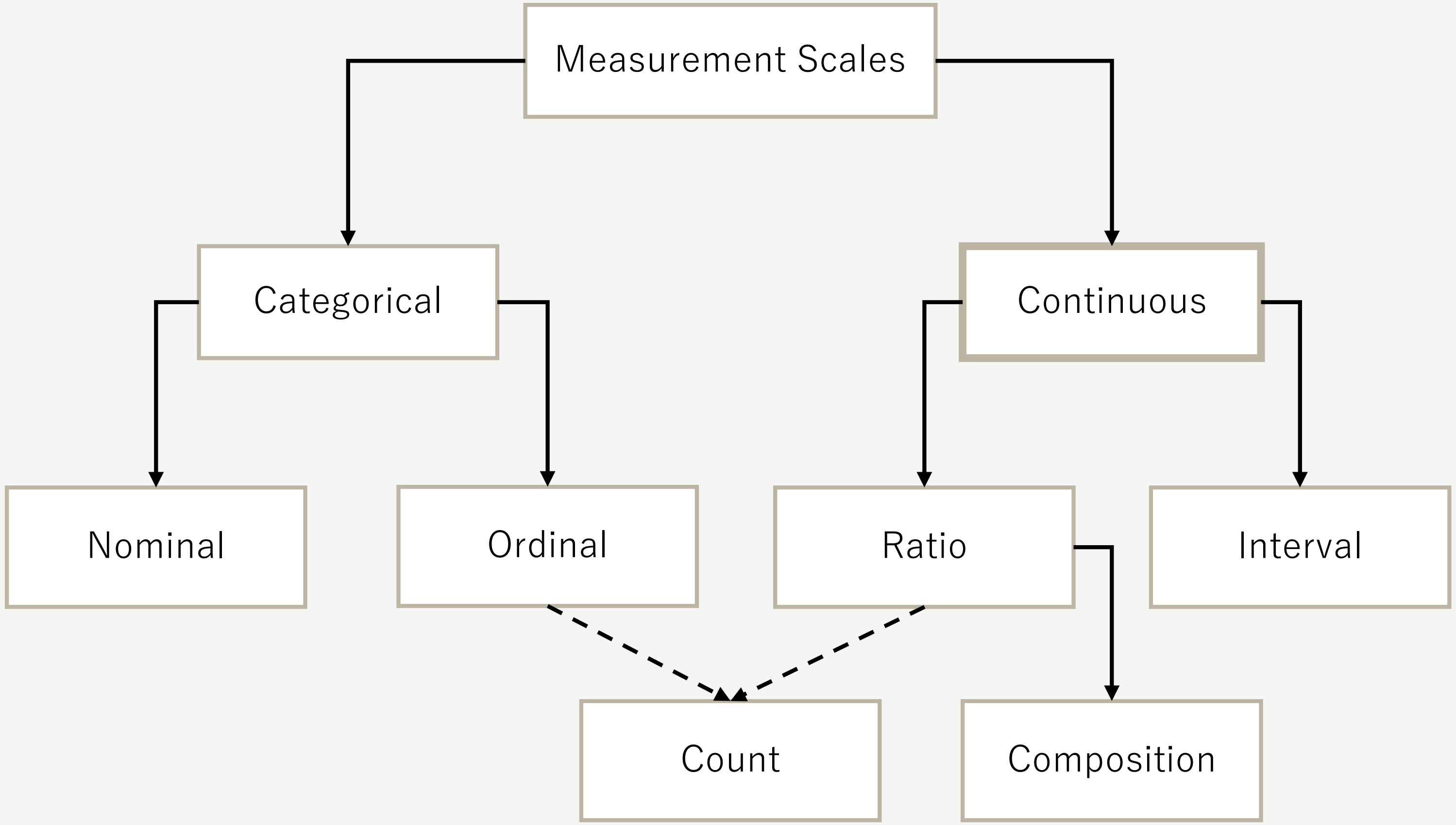


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

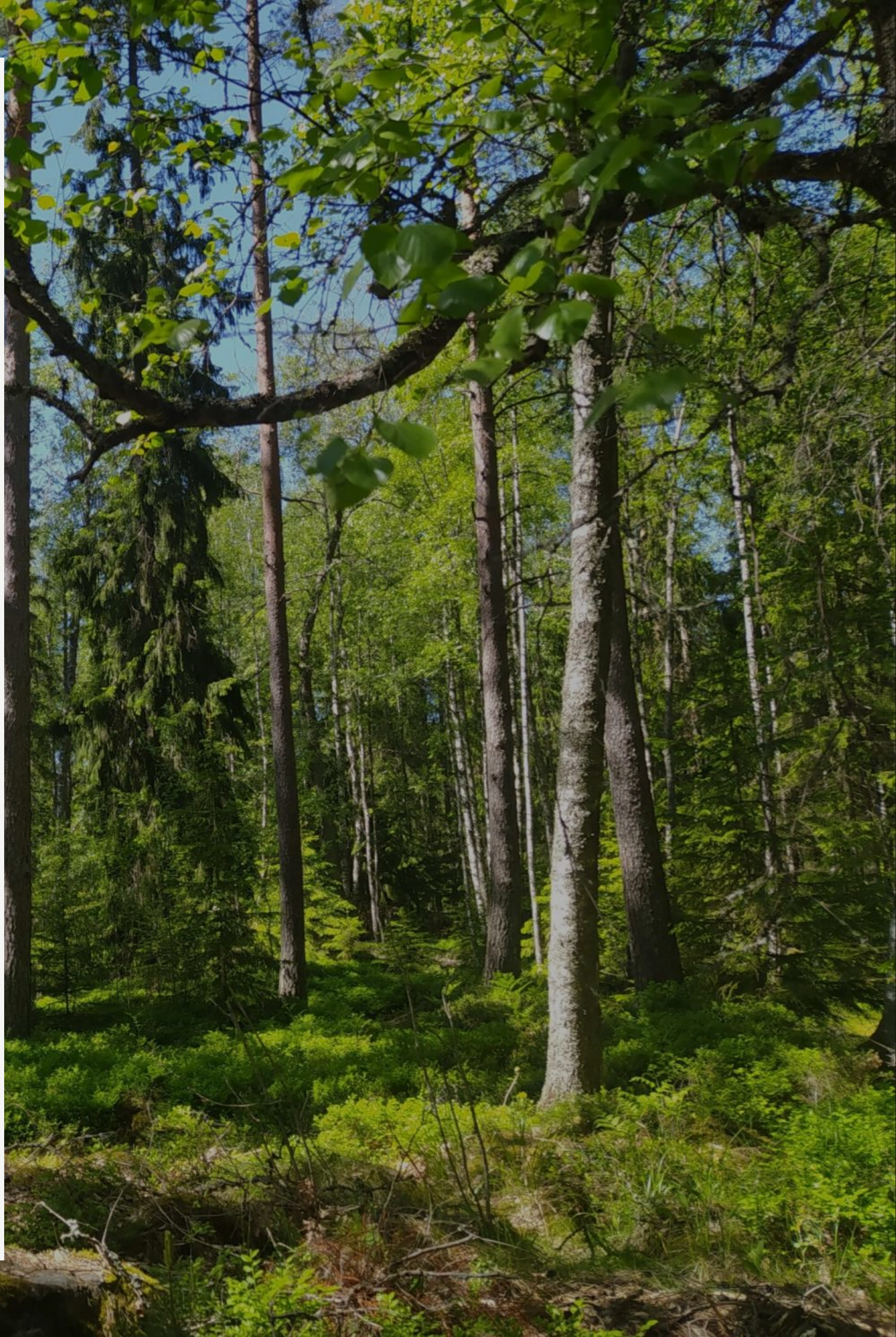


Data Screening: Data Structure

Continuous or Metric Data: Data with values that aren't fixed and can take on an unlimited number. *Can be plotted on a continuous axis of real numbers*.

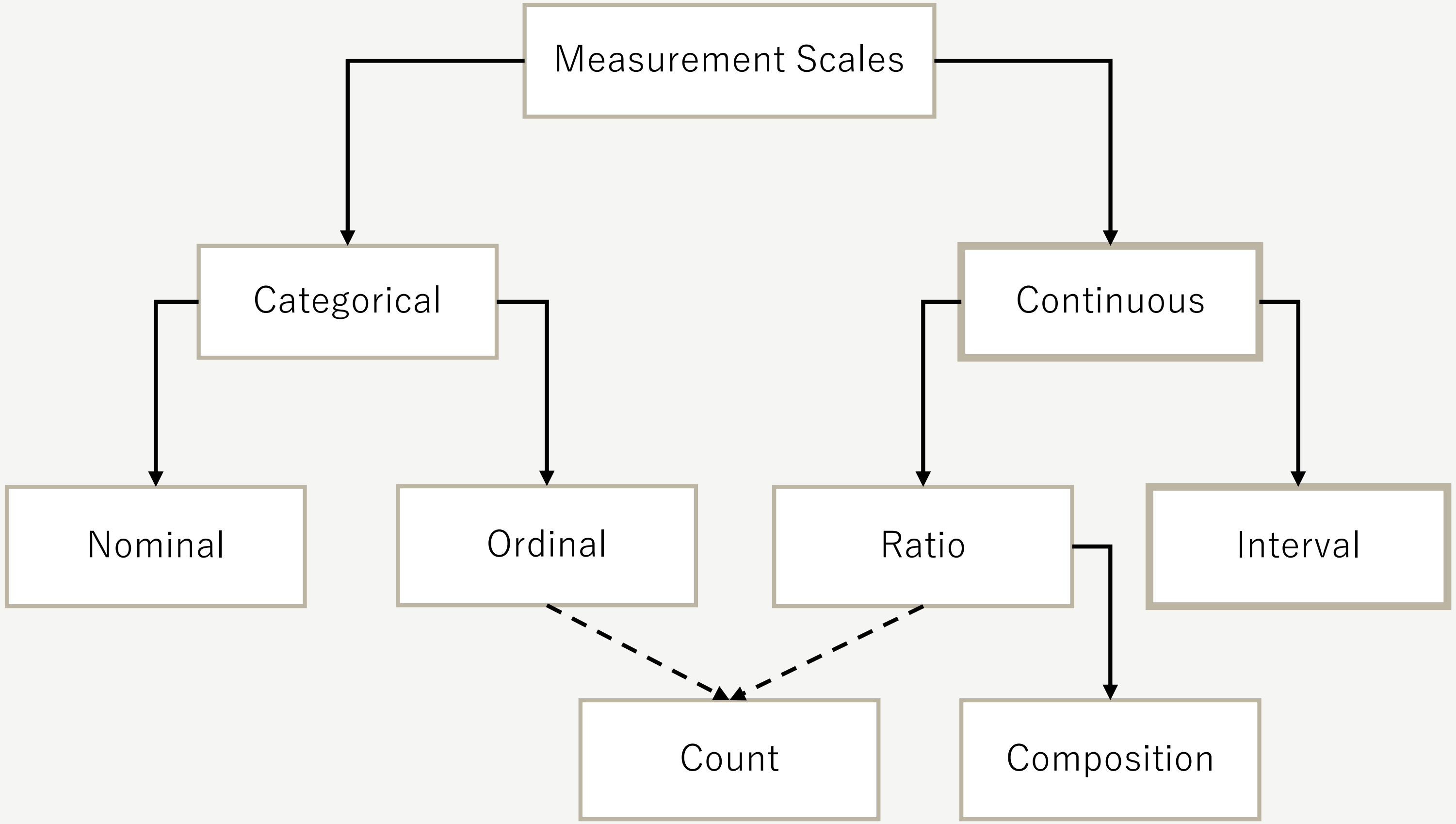


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

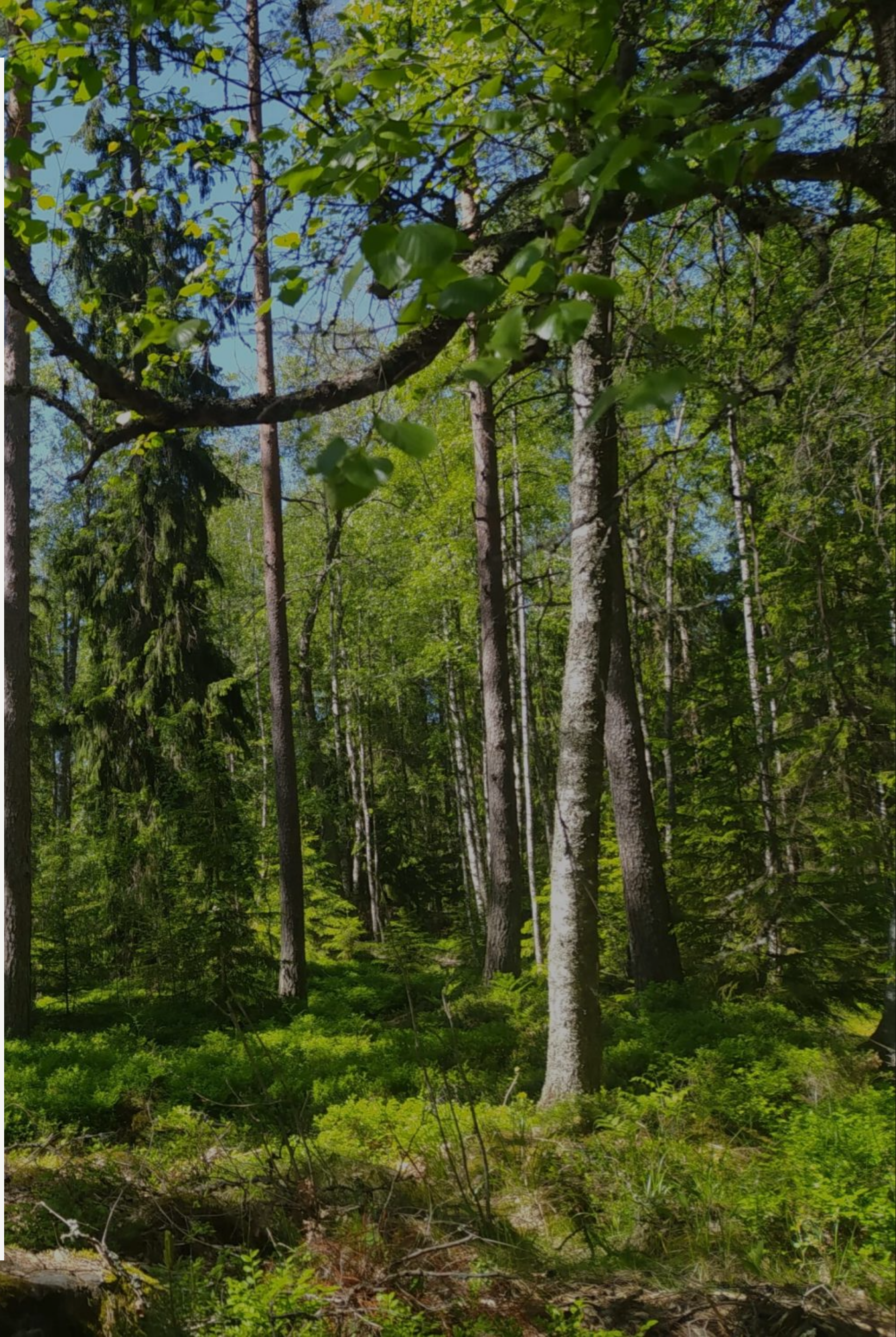


Data Screening: Data Structure

Interval Data: Values are compared *additively* and zero is chosen arbitrarily; e.g., time, temperature

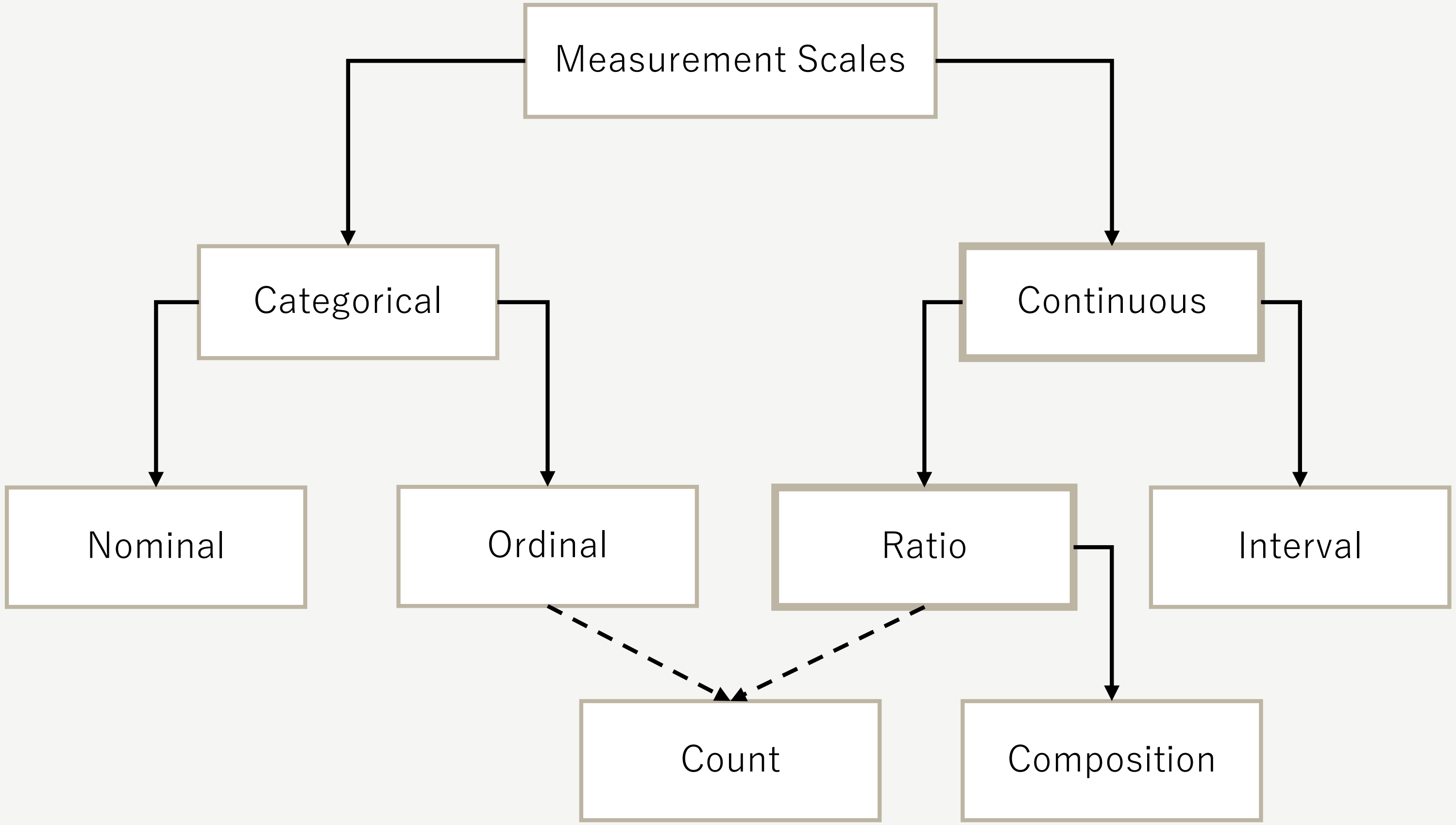


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

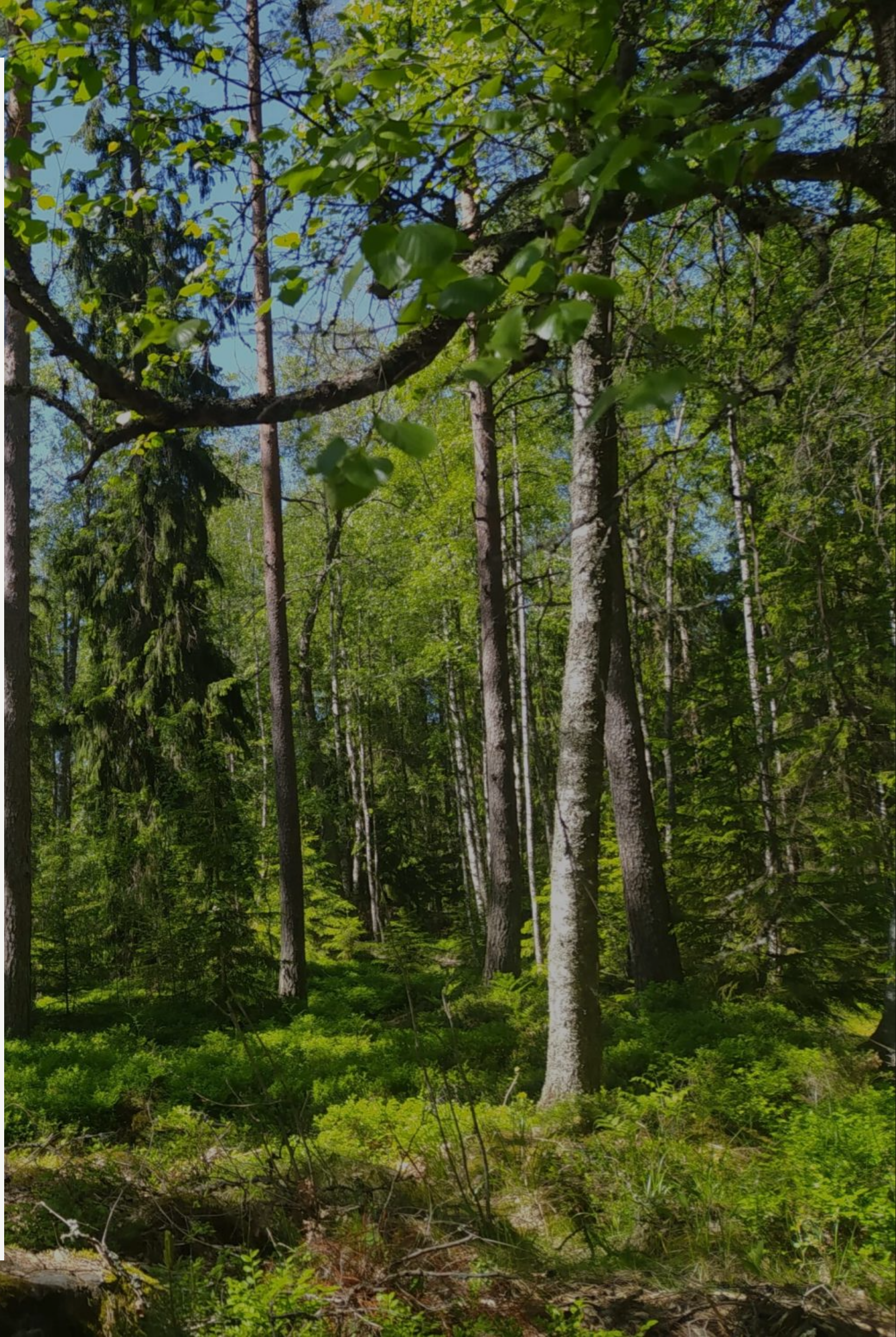


Data Screening: Data Structure

Ratio Data: Values are compared *multiplicatively*; e.g., length, weight, concentration, biomass. Zero means an absence of the characteristic, so they are *almost always non-negative*.

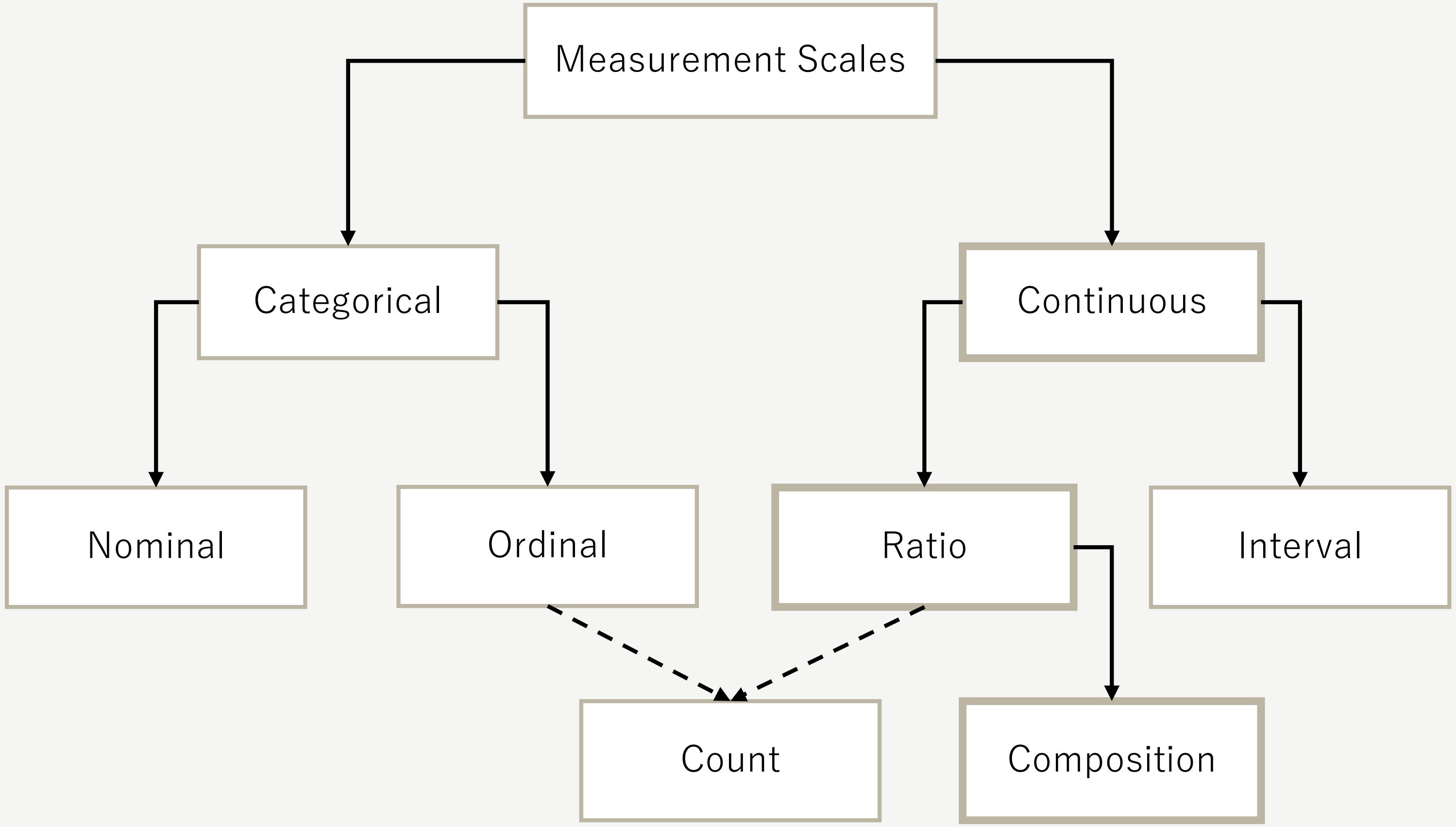


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2



Data Screening: Data Structure

Ratio Data: Values are compared *multiplicatively*, e.g., length, weight, concentration, biomass. ***Includes proportional (composition) data***



Measurement Scales

Categorical

Continuous

Nominal

Ordinal

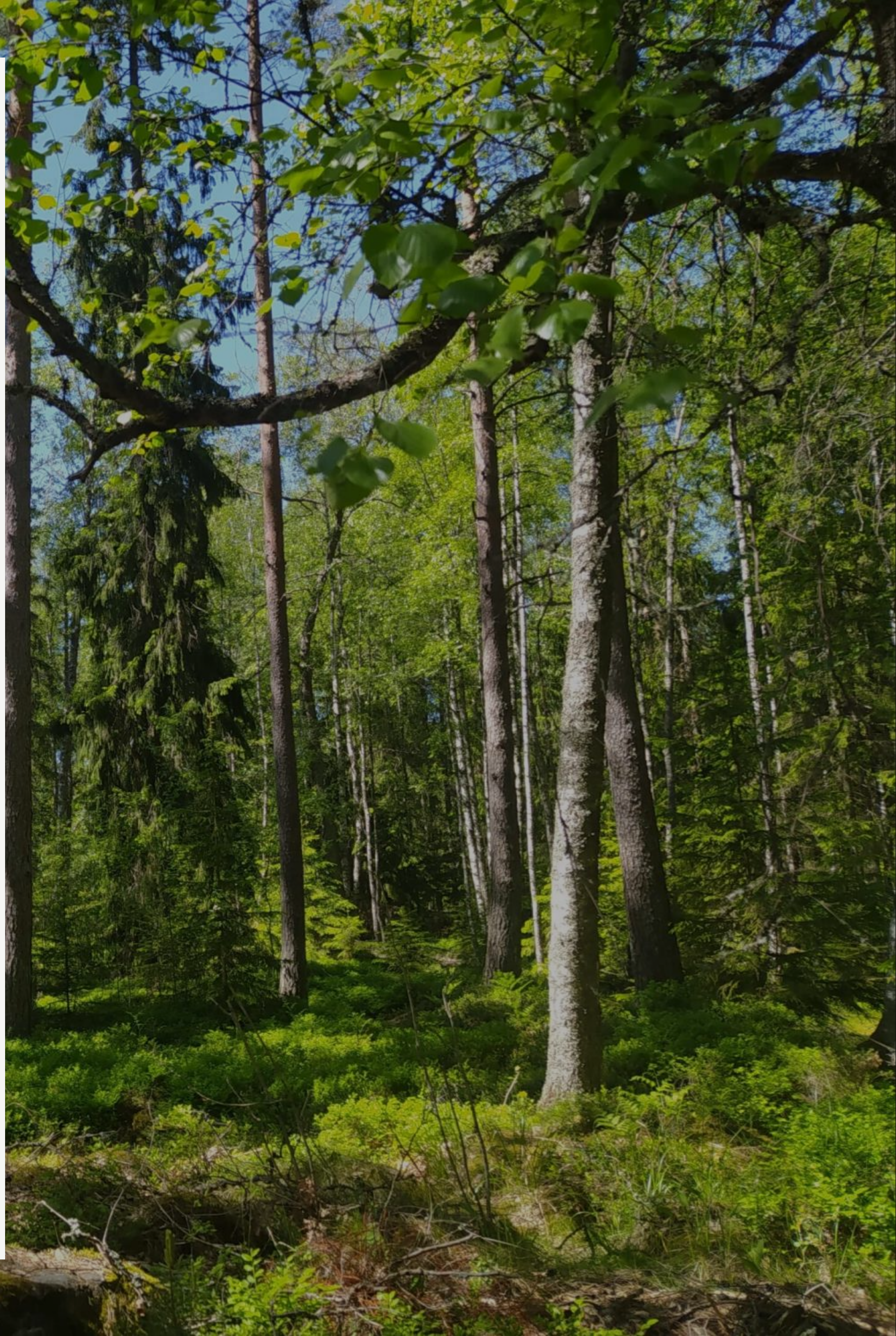
Ratio

Interval

Count

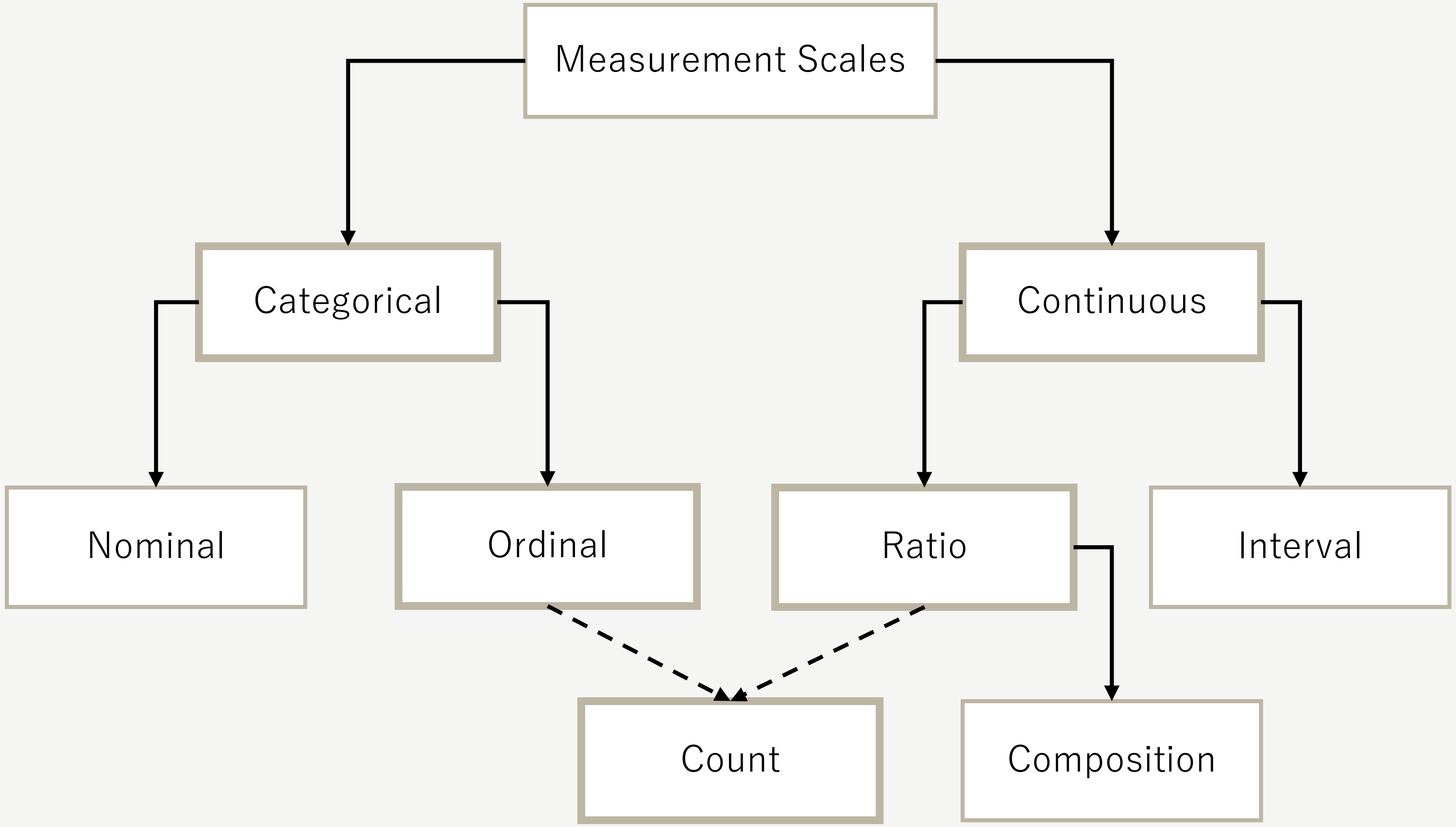
Composition

Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2

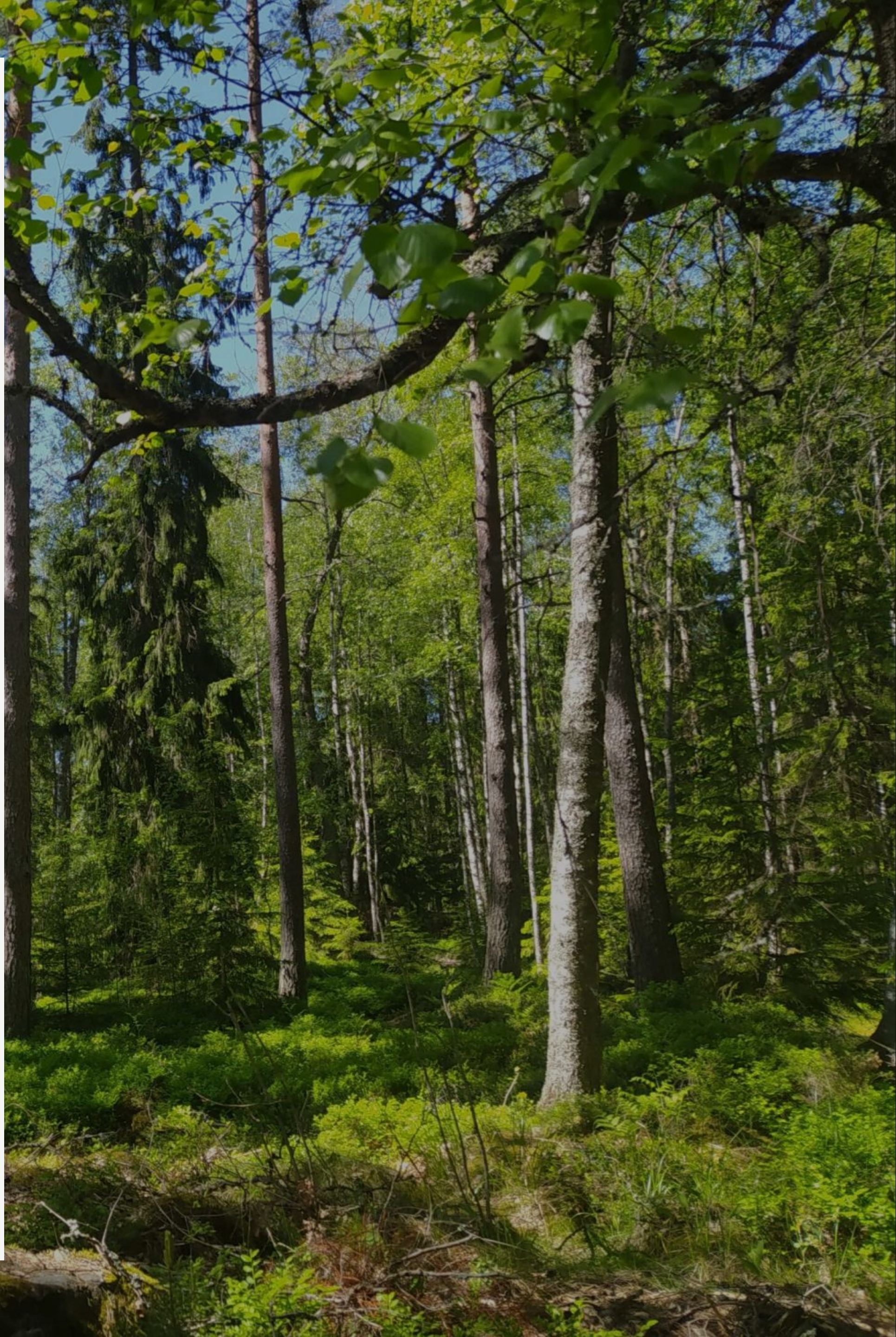


Data Screening: Data Structure

Count or Meristic Data: Can be **Ordinal** or **Ratio**. Often re-calculated as an average.

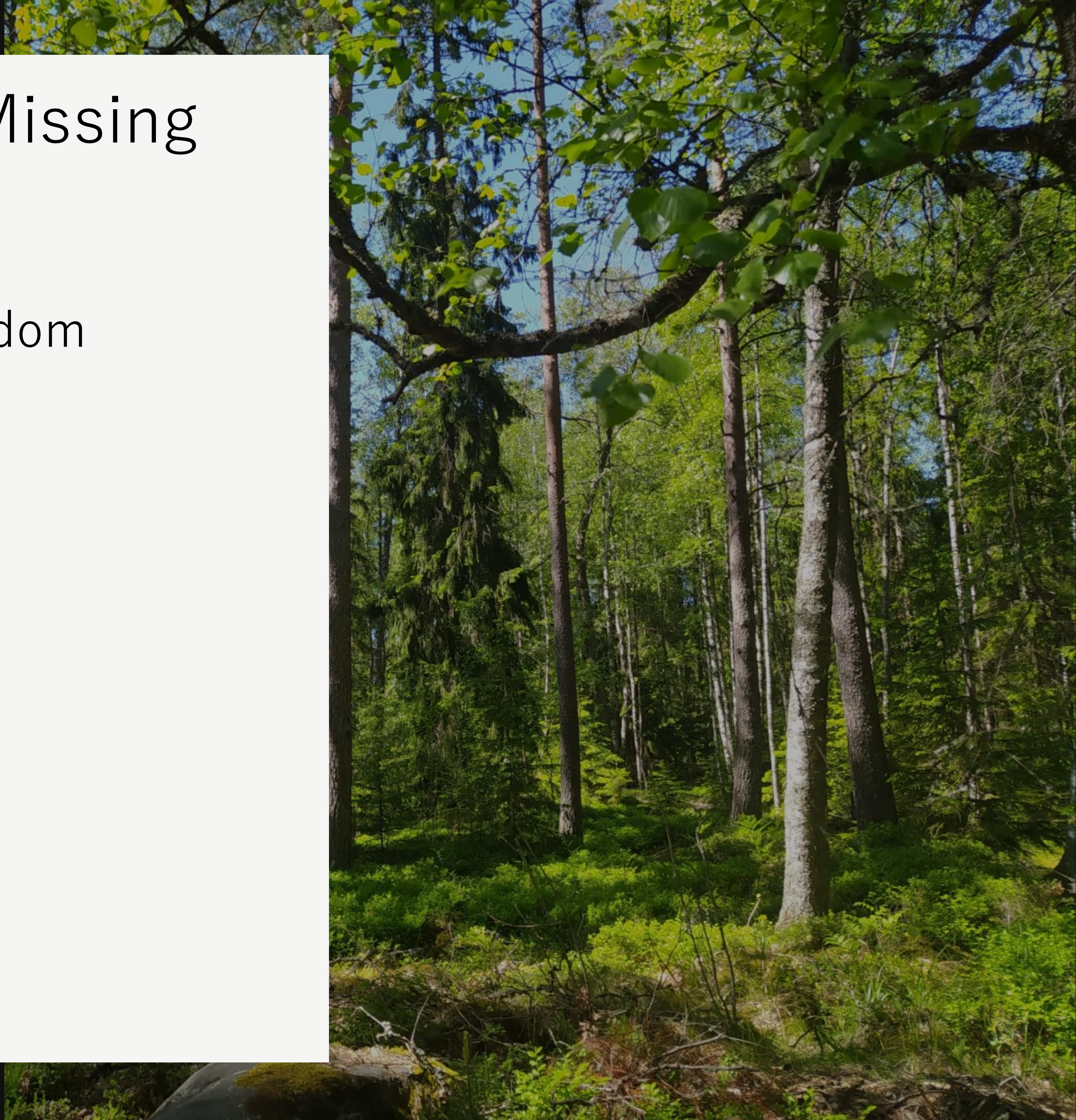


Greenacre and Primicerio: Fig. 3.1
Legendre & Legendre: Table 1.2



Data Screening: Handling Missing Data

- **MCAR:** Missing completely at random
- **MAR:** Missing at random
- **MNAR:** Missing not at random



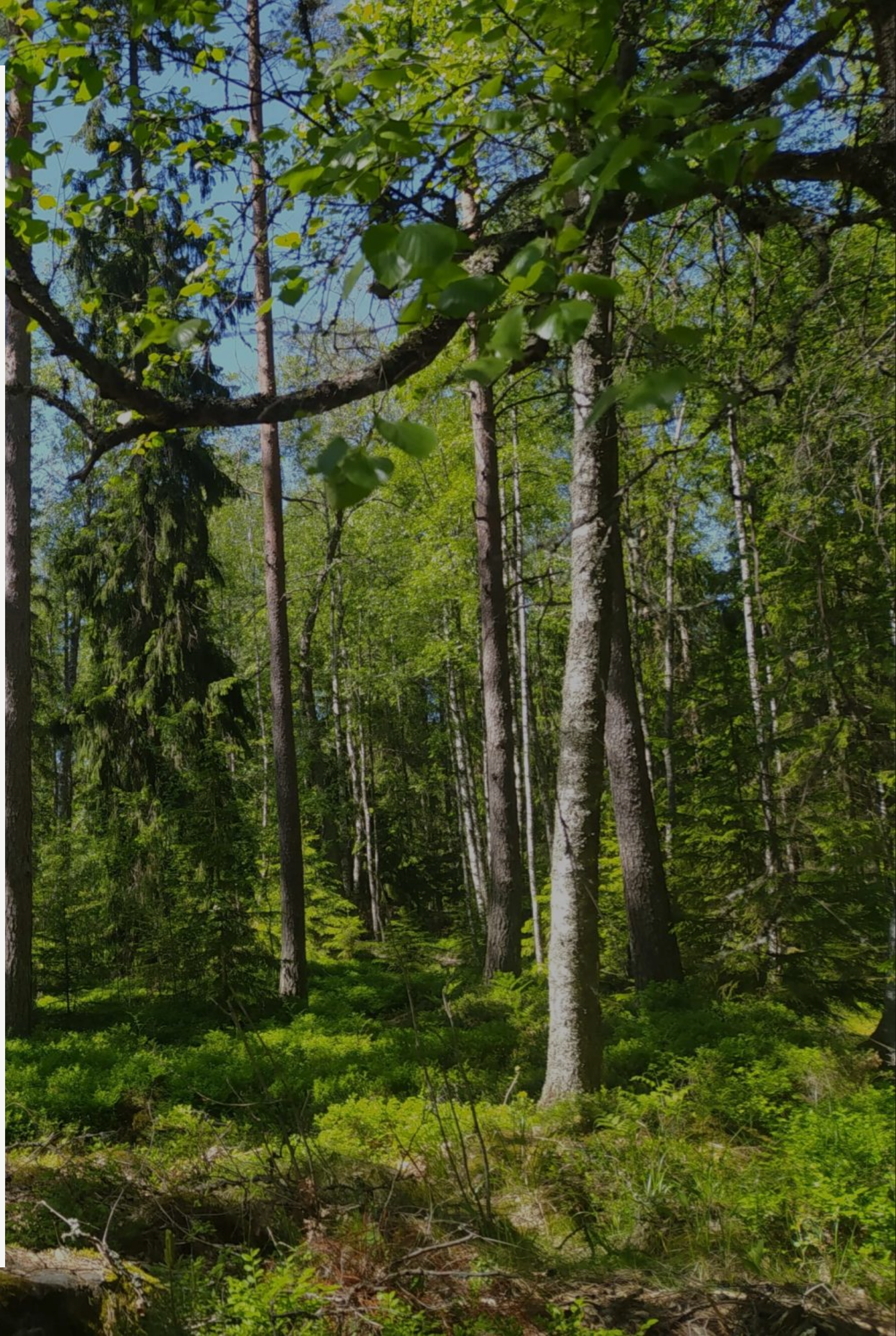
Data Screening: Handling Missing Data

Is it a **zero** or an **NA**?

“True” missing data implies non-measurement while a zero value is a measurement of absence. **Do not code missing data as zeros!**

Data Screening: Handling Missing Data

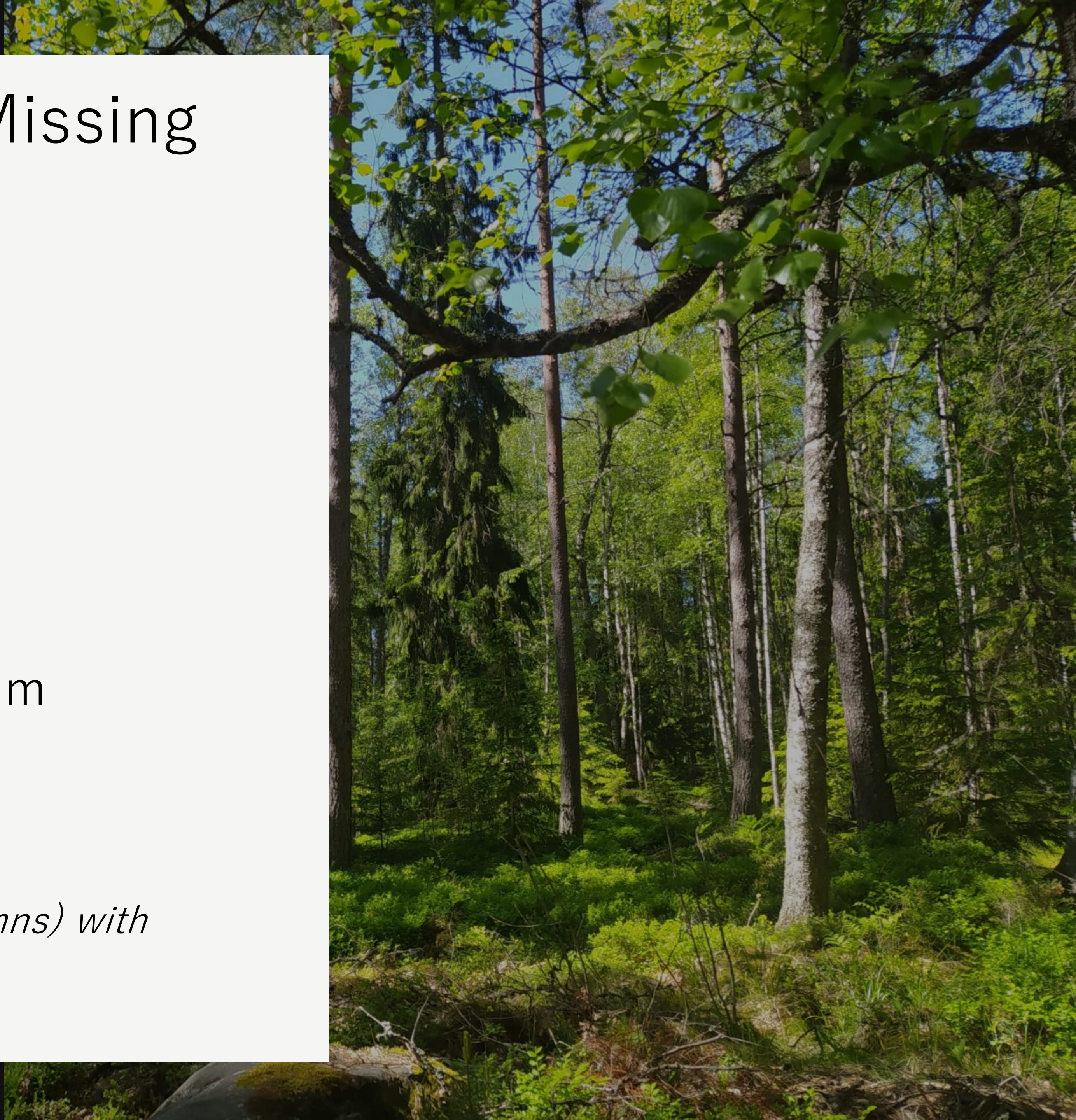
- Listwise Deletion
- Mean Imputation
- Regression Imputation
- Expectation-Maximization Algorithm
- Machine Learning Methods



Data Screening: Handling Missing Data

- **Listwise Deletion**
- Mean Imputation
- Regression Imputation
- Expectation-Maximization Algorithm
- Machine Learning Methods

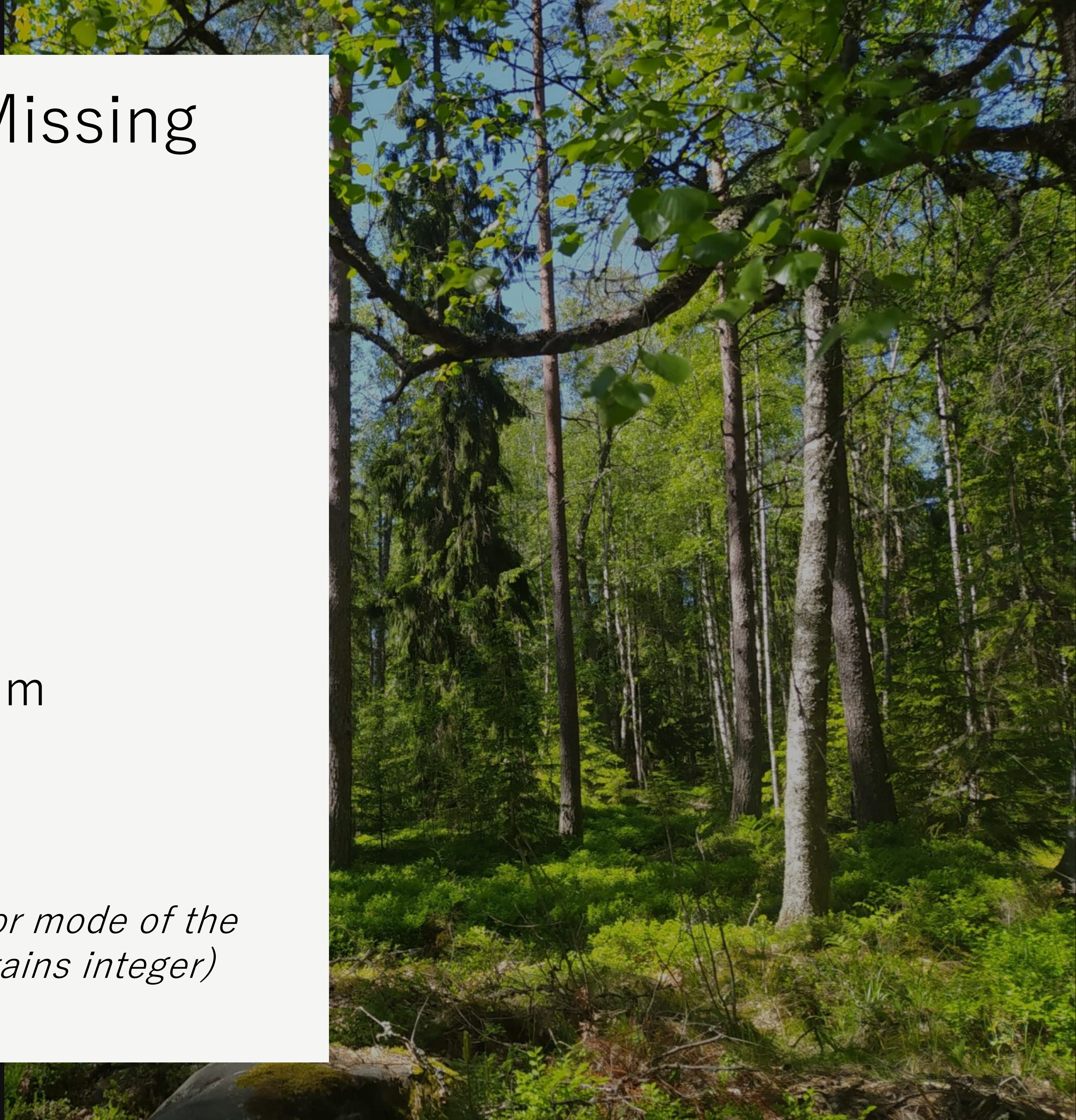
Exclude sites (rows), species, or variables (columns) with chronically missing data



Data Screening: Handling Missing Data

- Listwise Deletion
- **Mean Imputation**
- Regression Imputation
- Expectation-Maximization Algorithm
- Machine Learning Methods

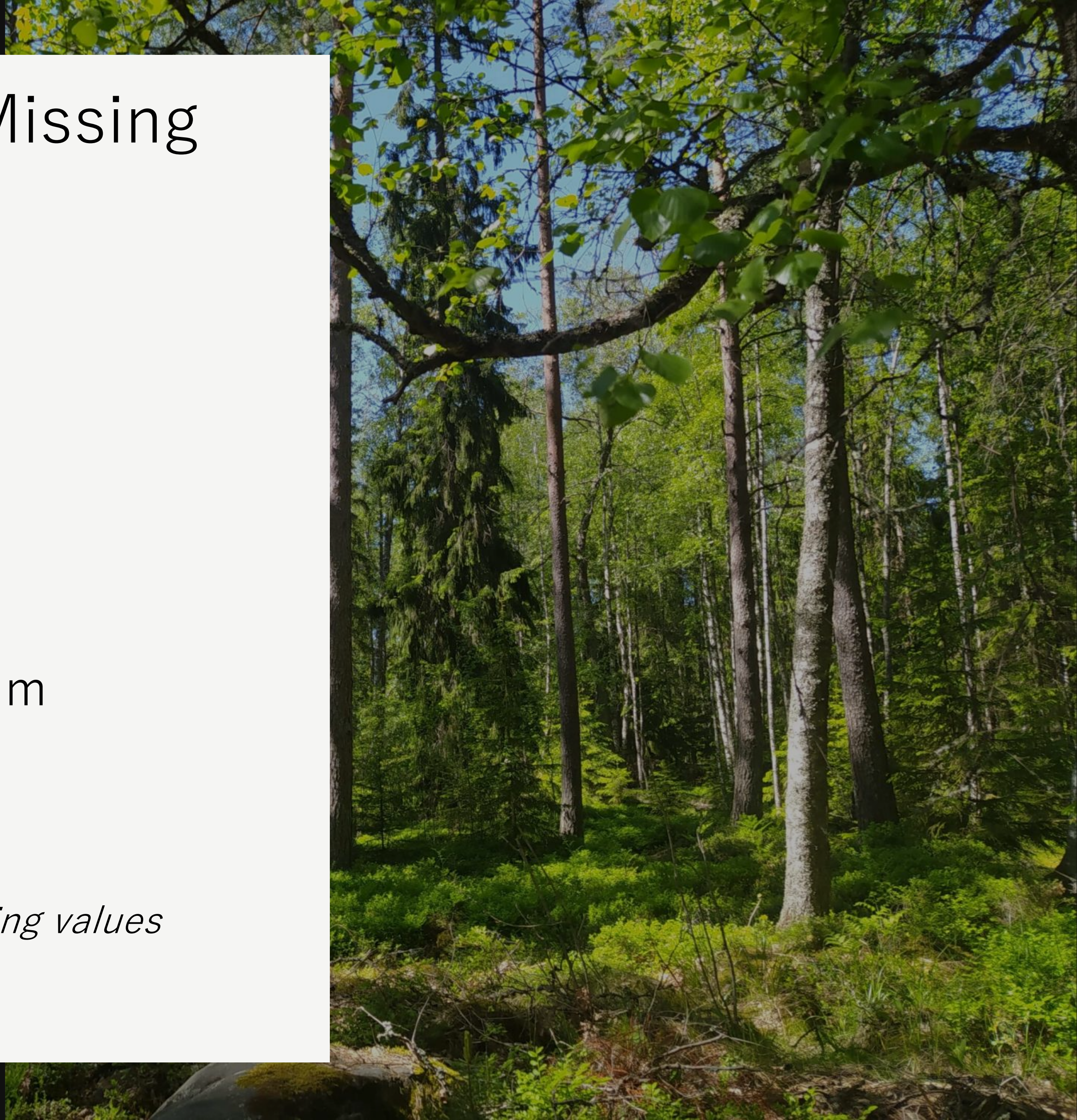
Replace missing values with the mean, median, or mode of the variable. Median is optimal for count data (maintains integer)



Data Screening: Handling Missing Data

- Listwise Deletion
- Mean Imputation
- **Regression Imputation**
- Expectation-Maximization Algorithm
- Machine Learning Methods

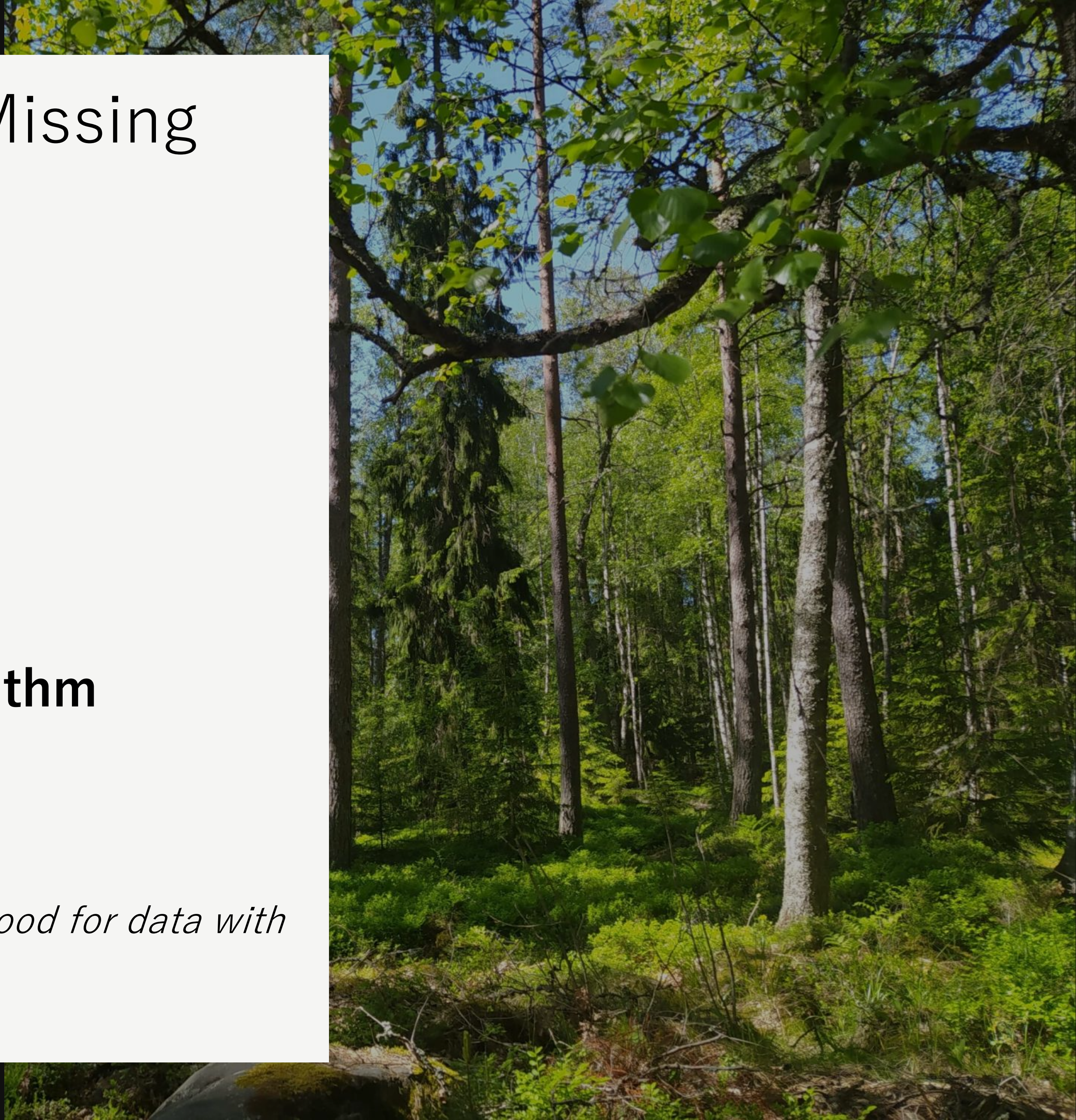
Use regression models to predict and fill in missing values



Data Screening: Handling Missing Data

- Listwise Deletion
- Mean Imputation
- Regression Imputation
- **Expectation-Maximization Algorithm**
- Machine Learning Methods

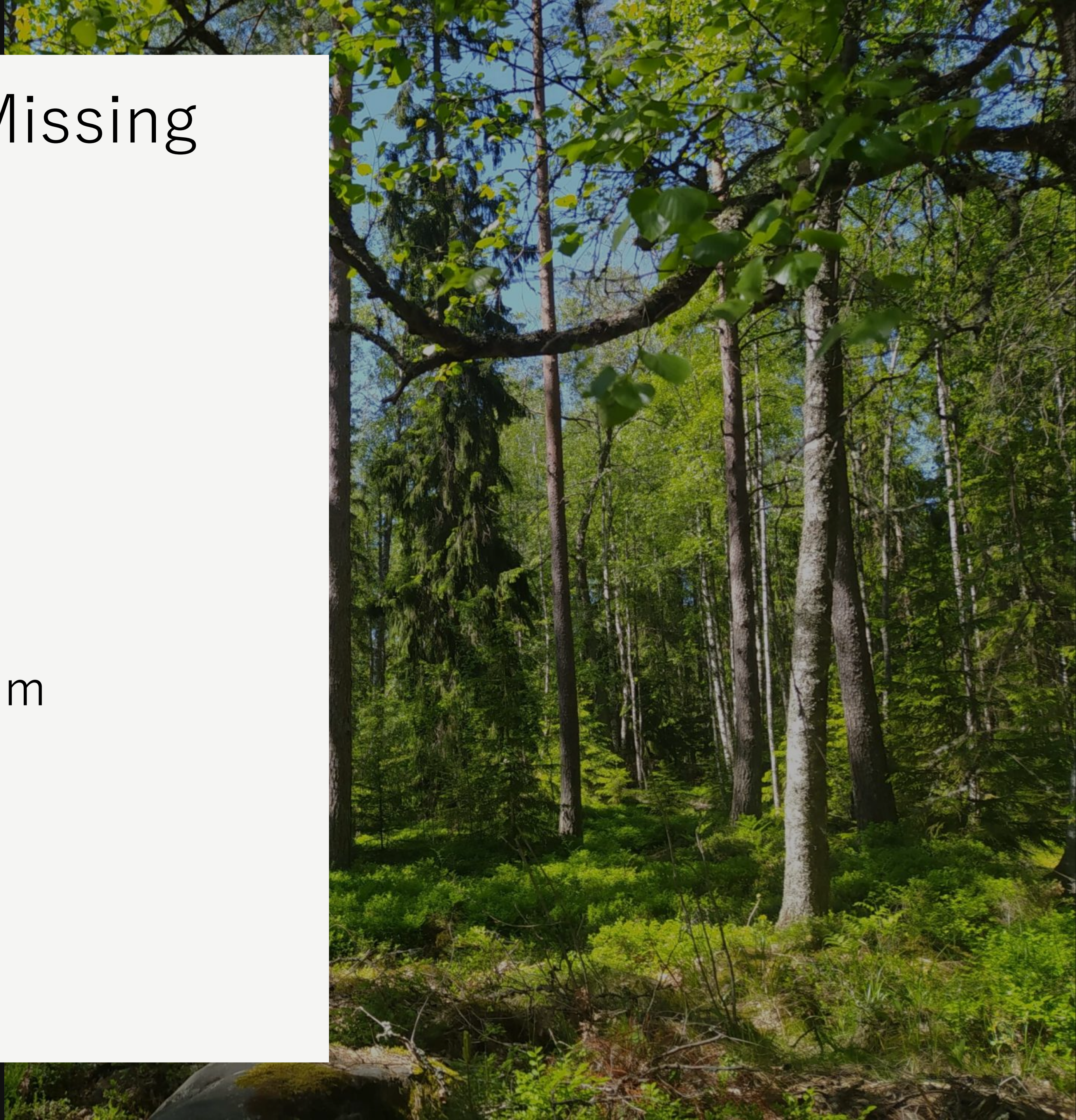
Estimates missing data by maximizing the likelihood for data with a well-defined distribution



Data Screening: Handling Missing Data

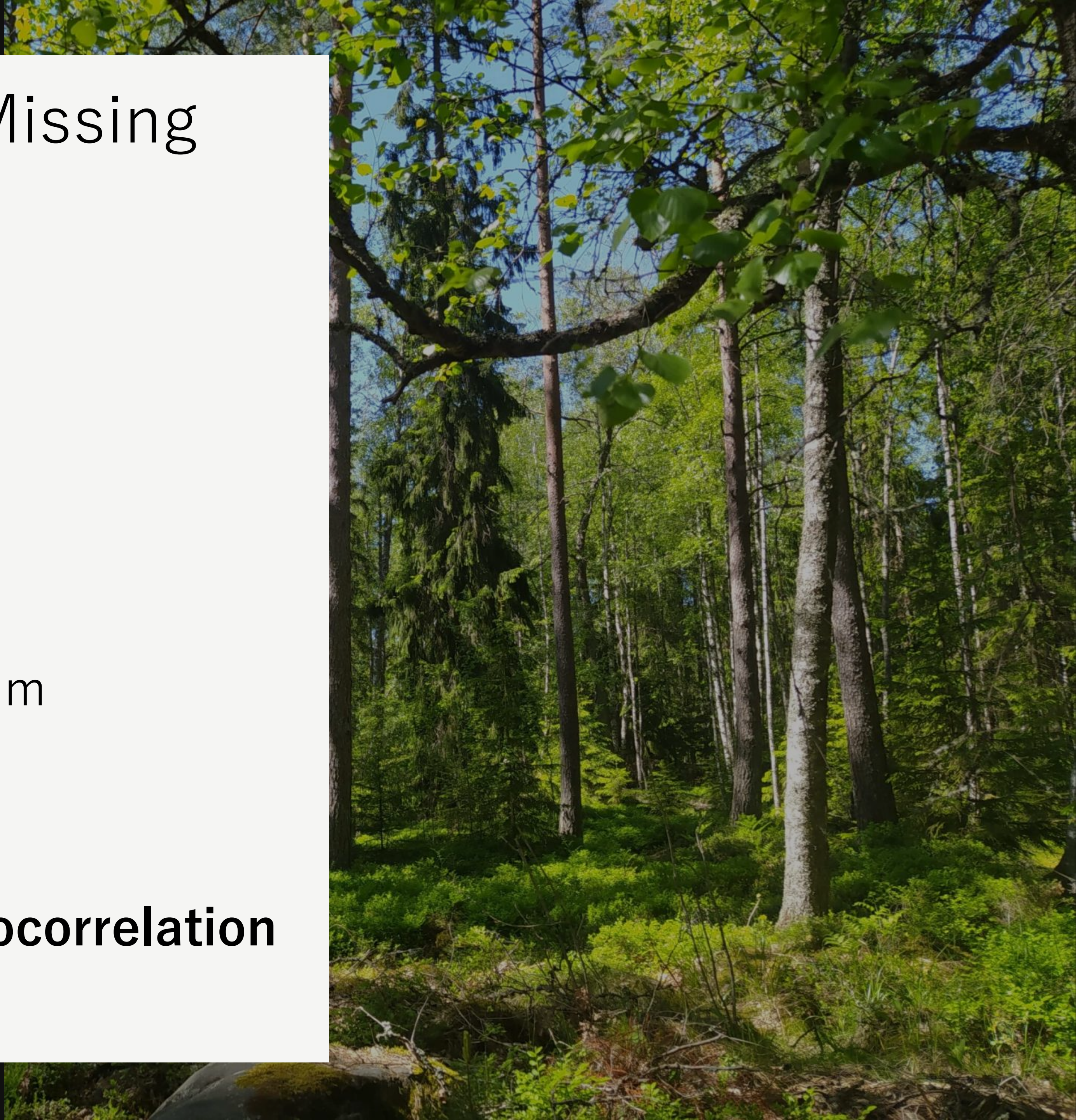
- Listwise Deletion
- Mean Imputation
- Regression Imputation
- Expectation-Maximization Algorithm
- **Machine Learning Methods**

Uses neural networks for imputation



Data Screening: Handling Missing Data

- Listwise Deletion
- Mean Imputation
- Regression Imputation
- Expectation-Maximization Algorithm
- Machine Learning Methods
- **Interpolate based on spatial autocorrelation**



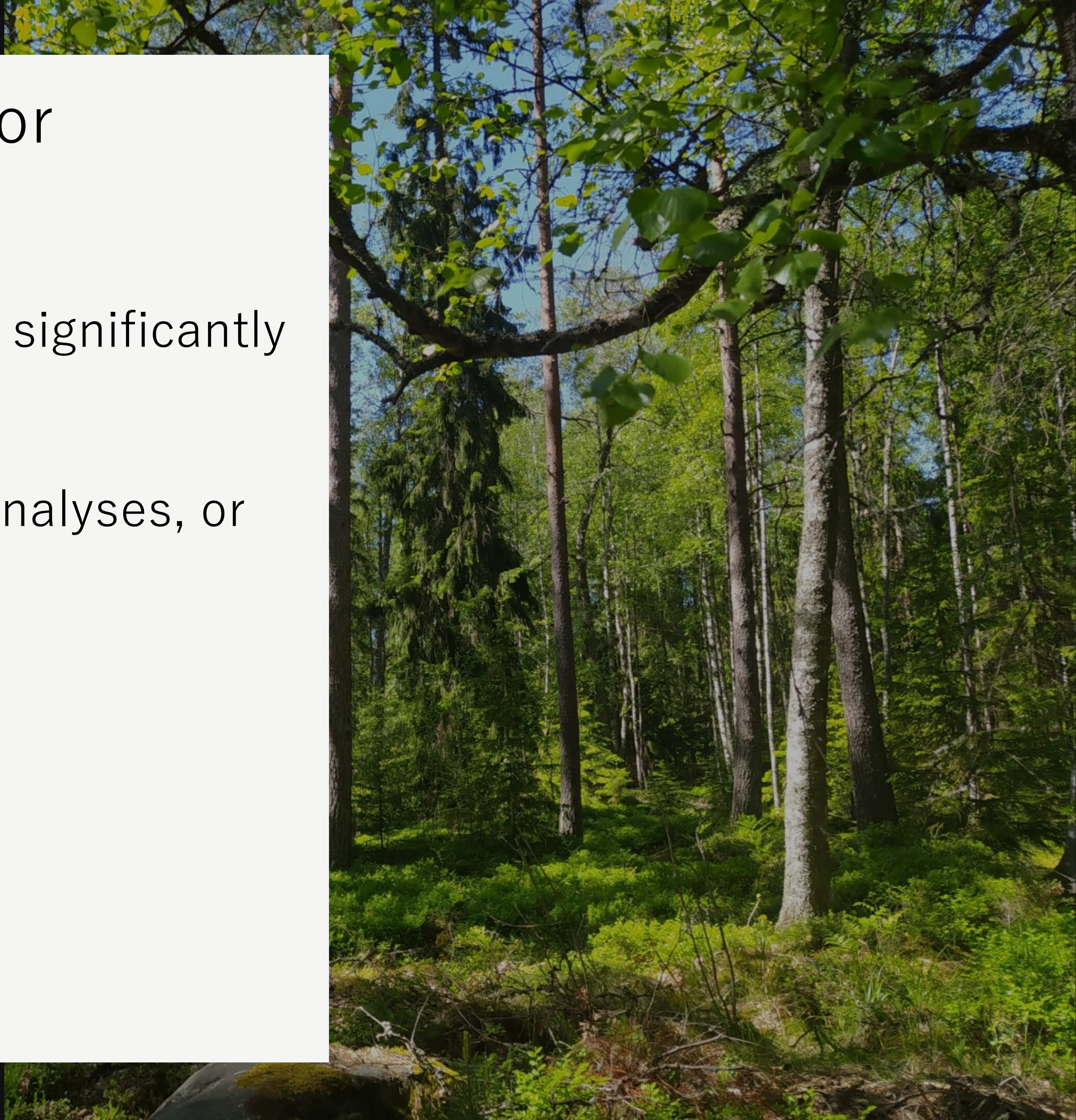
Data Screening: Checking for Outliers

Outliers are data points that deviate significantly from the majority of the dataset

May skew results, affect statistical analyses, or indicate data quality issues

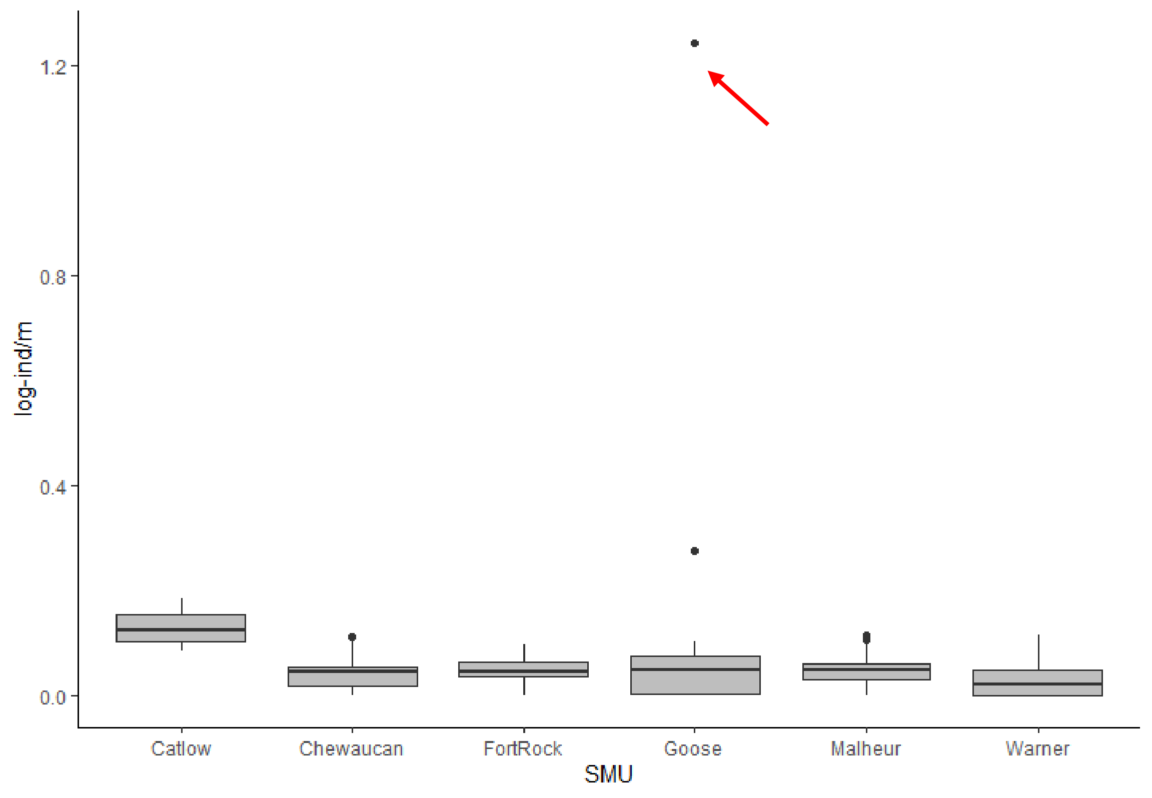
Sources:

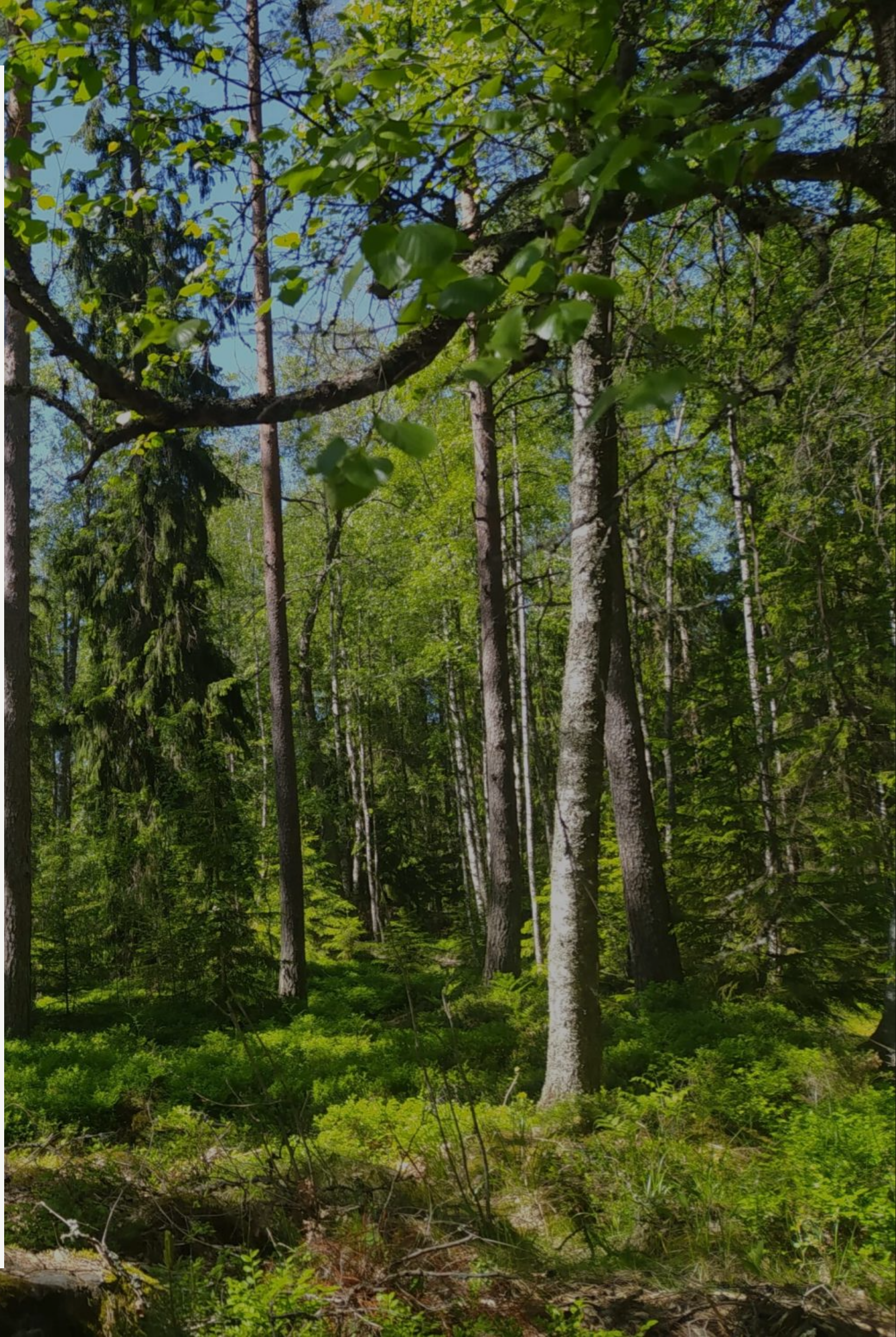
- Data entry errors
- Measurement errors
- Variability in the data



Data Screening: Checking for Outliers

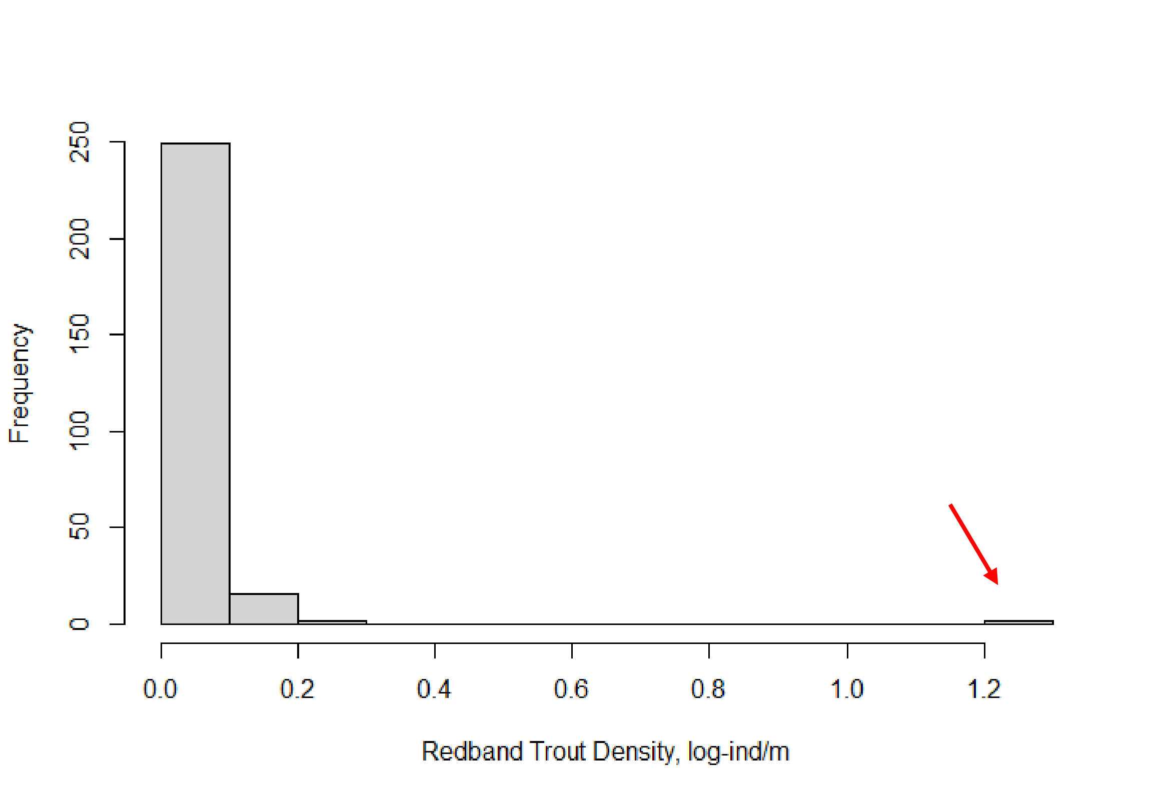
Box Plot: Points outside the whiskers are potential outliers

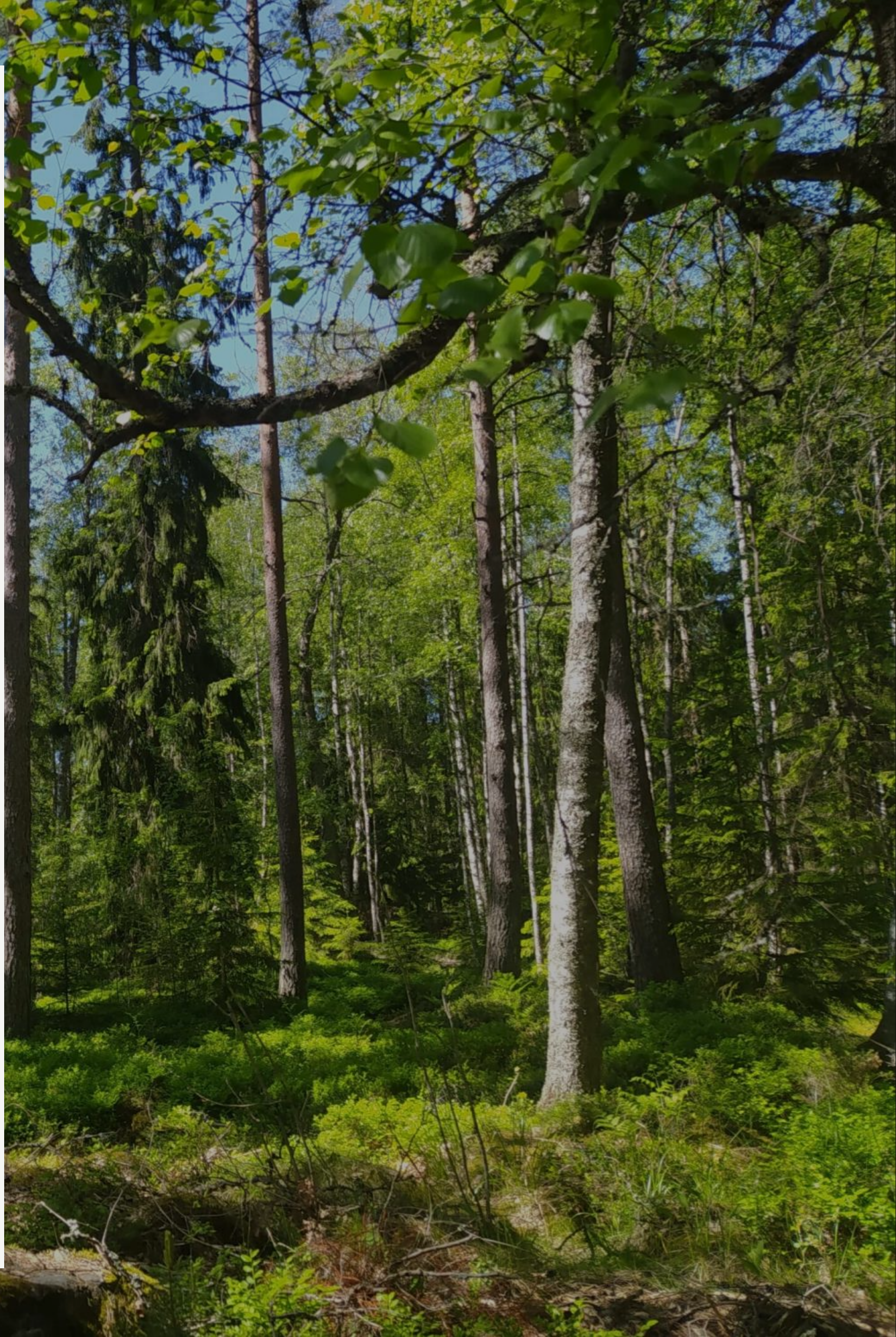




Data Screening: Checking for Outliers

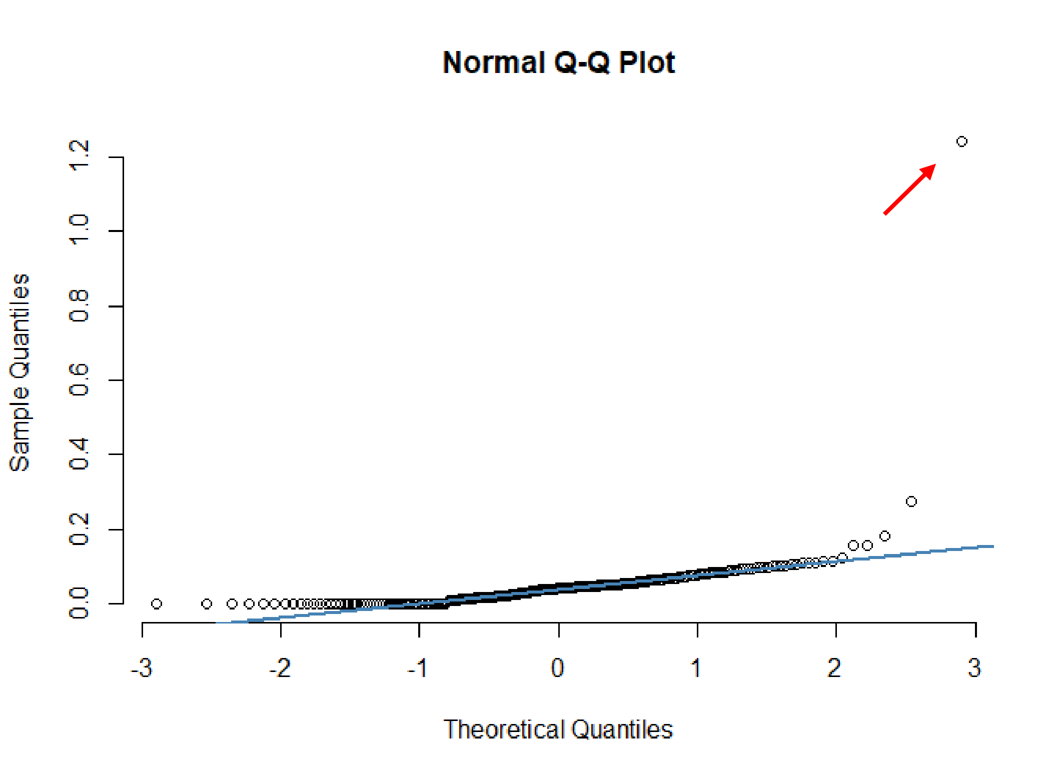
Histogram: Outliers appear as isolated bars or extreme tails

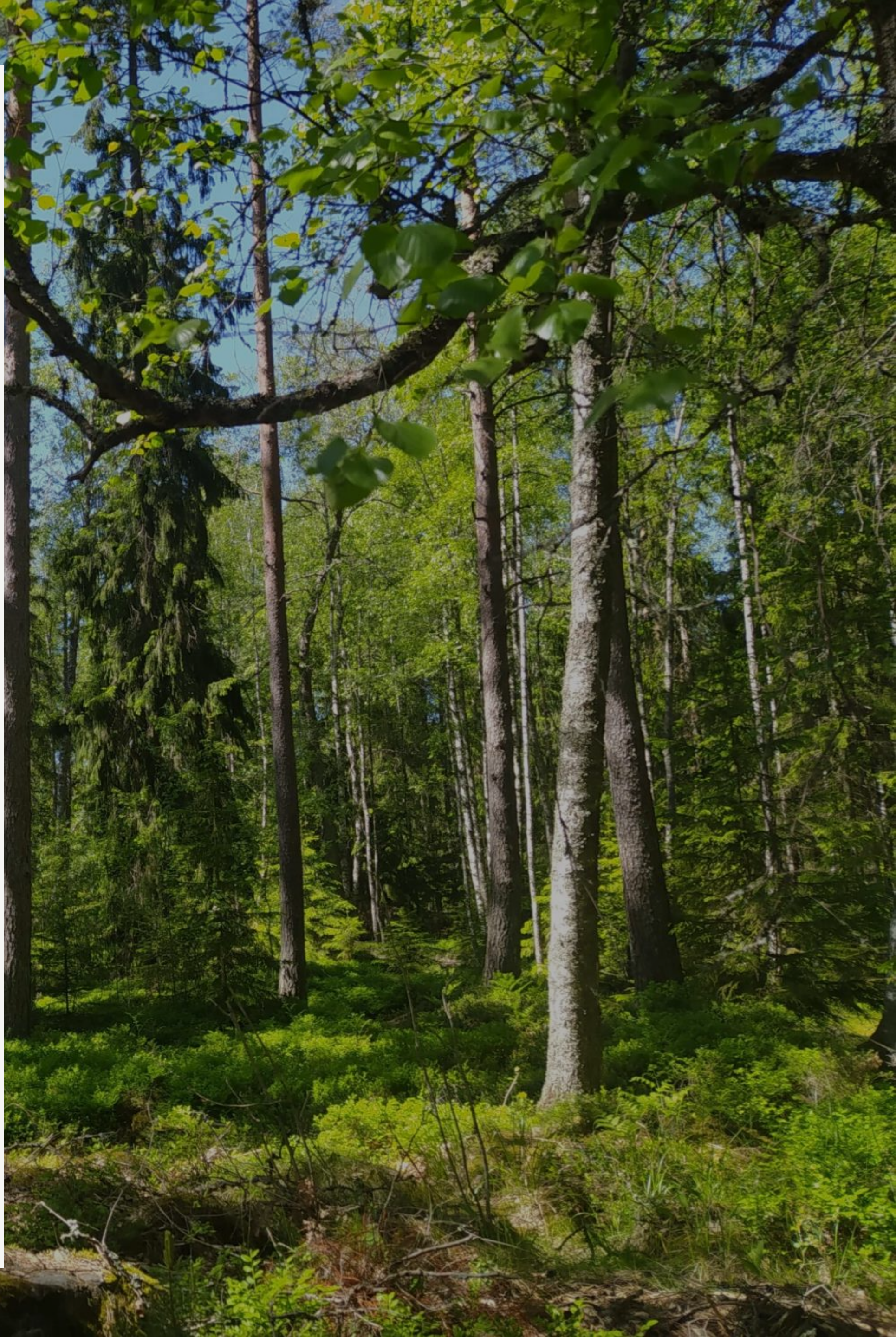




Data Screening: Checking for Outliers

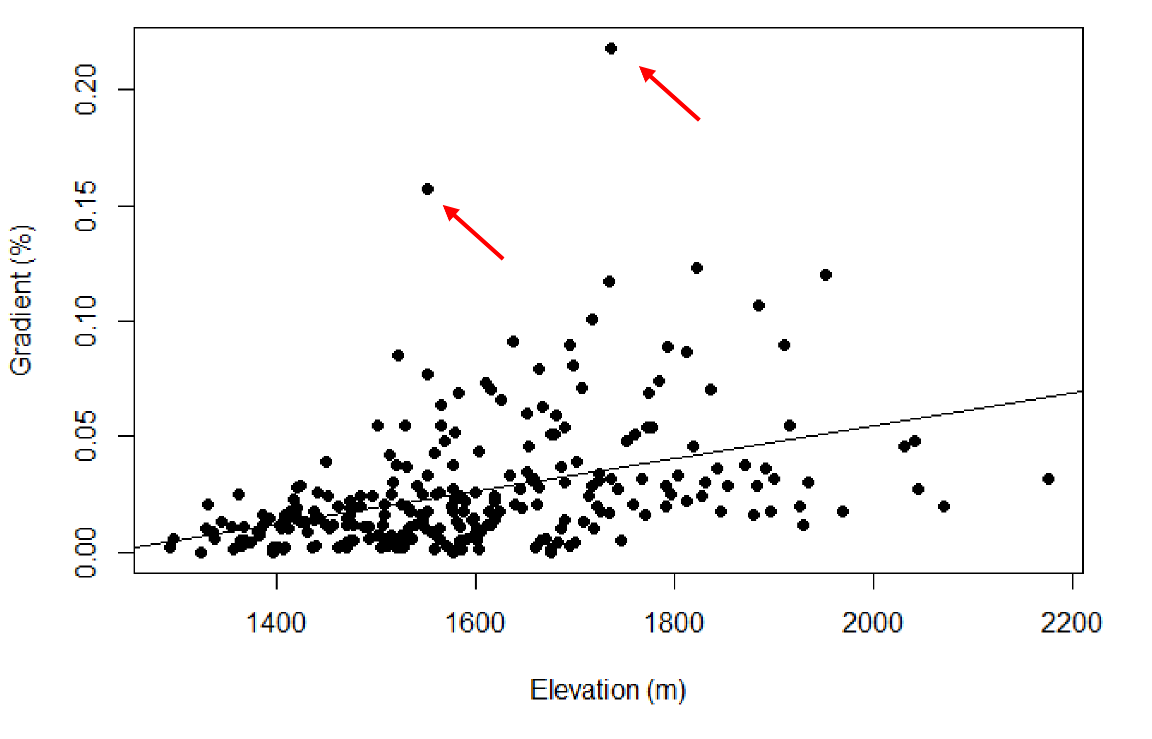
QQ Plot: Compares the distribution of the data to a theoretical (e.g., normal) distribution

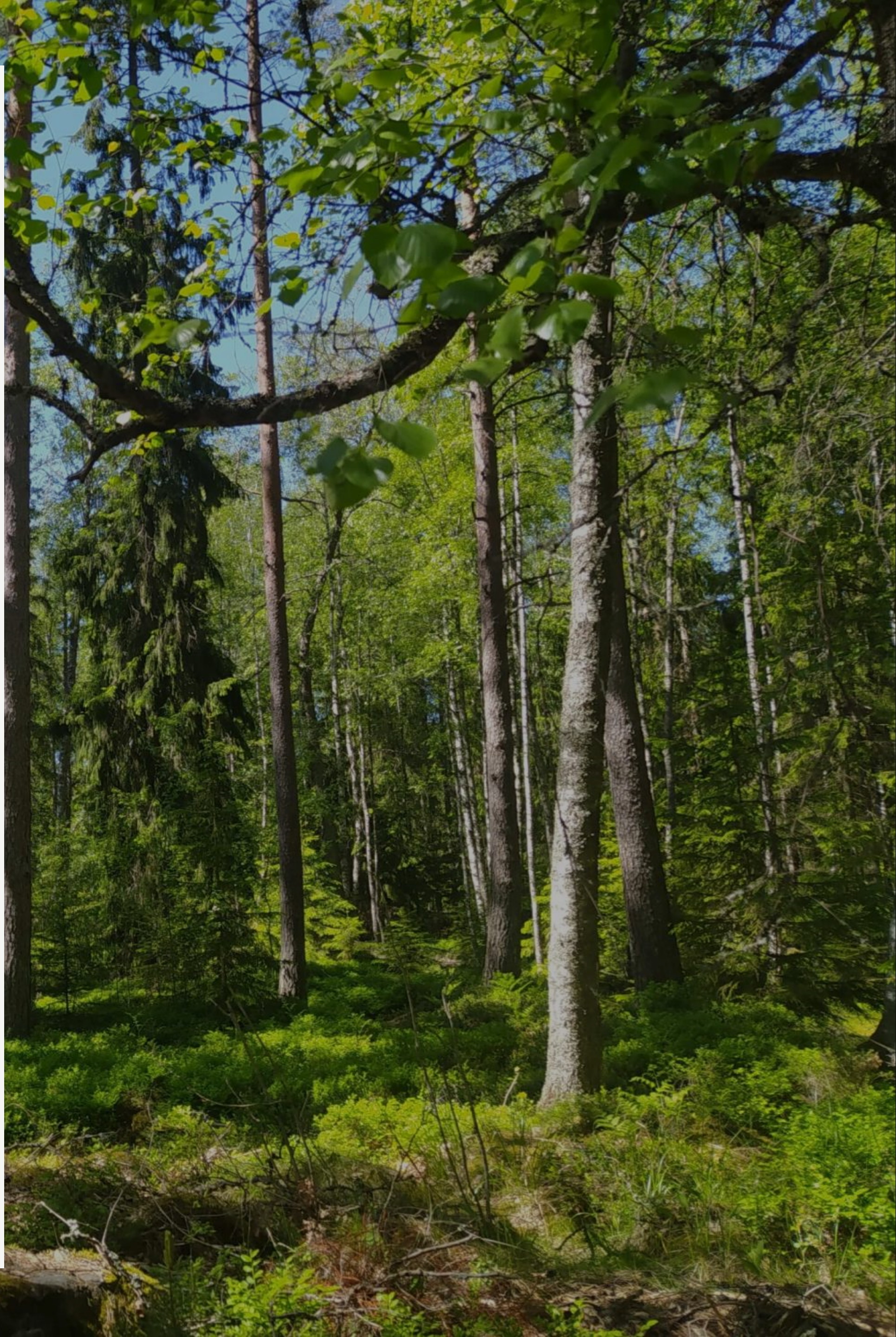




Data Screening: Checking for Outliers

Scatter Plot: Identifies outliers in relationship between two variables



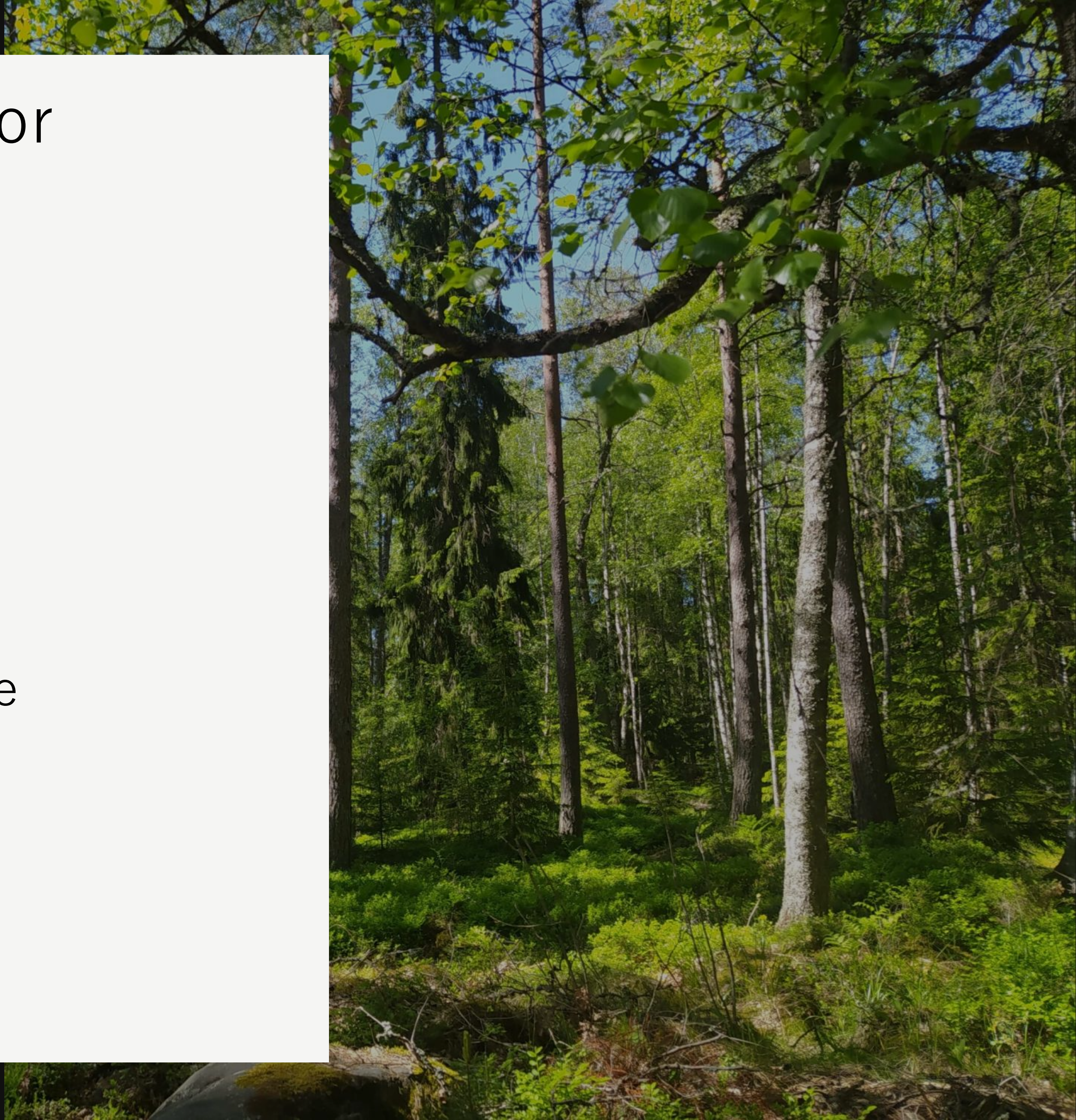


Data Screening: Checking for Outliers

Multivariate Approaches:

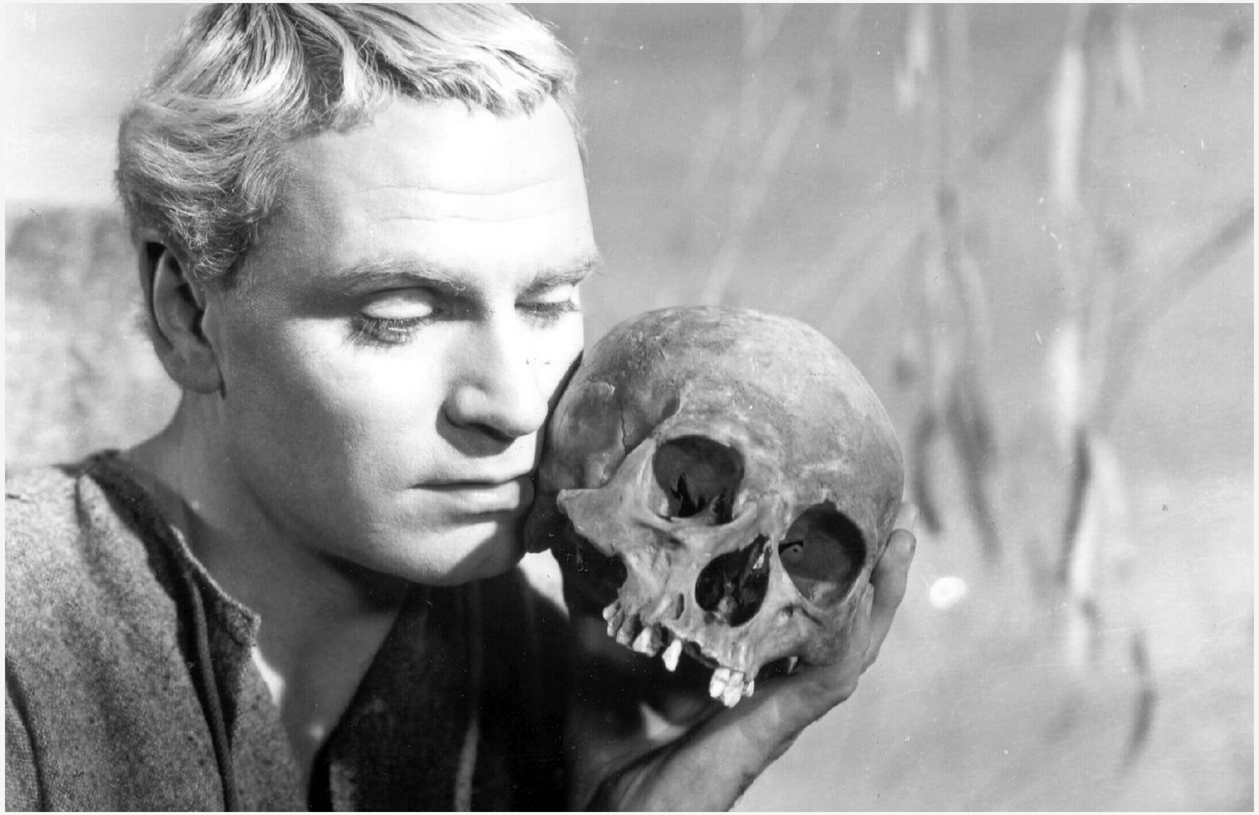
- Mahalanobis distance
- Minimum volume ellipsoid
- Elliptical symmetry robust distance
- Minimum covariance determinant

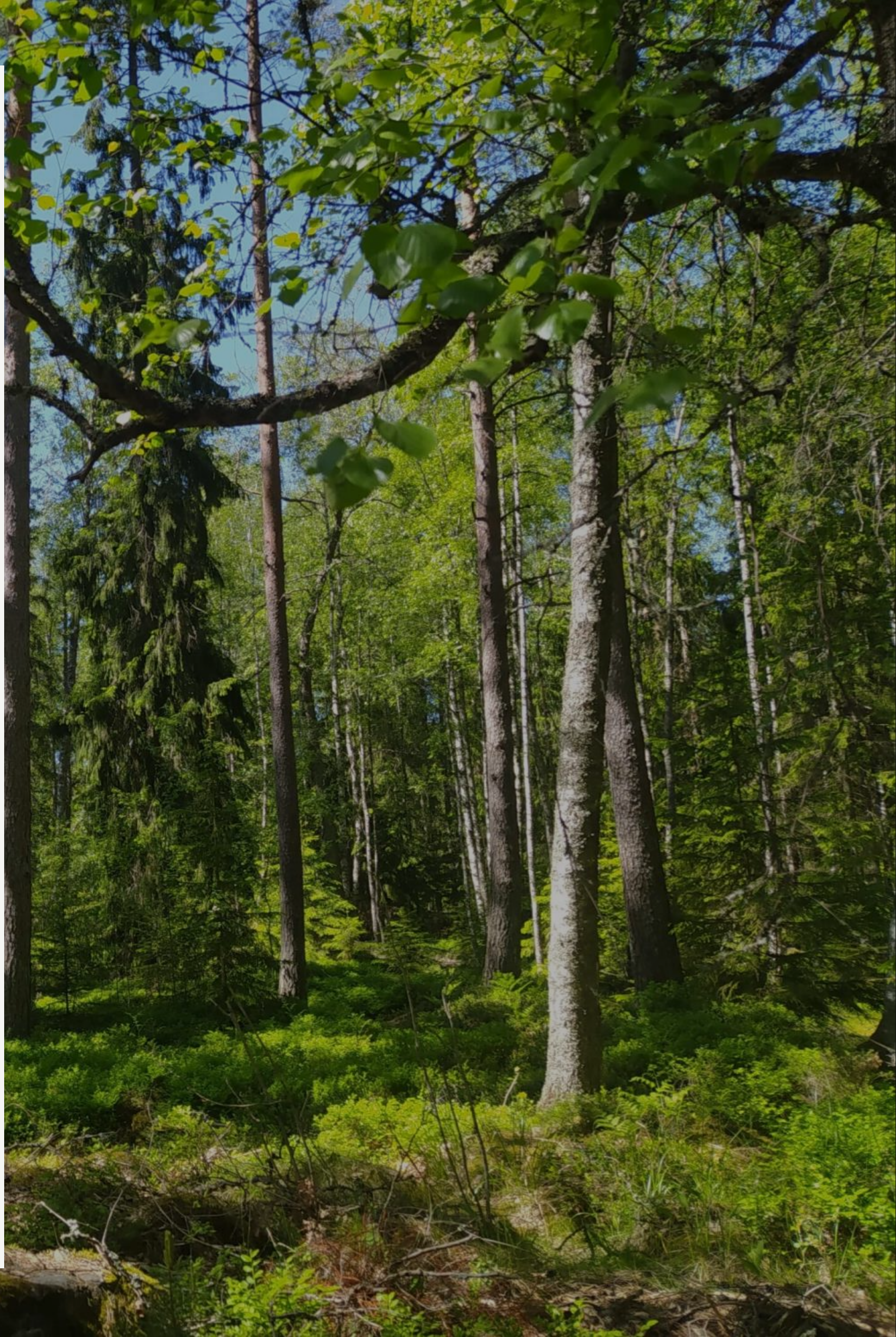
See Alameddine et al. 2010 for more discussion



Data Screening: Checking for Outliers

To remove or not to remove?





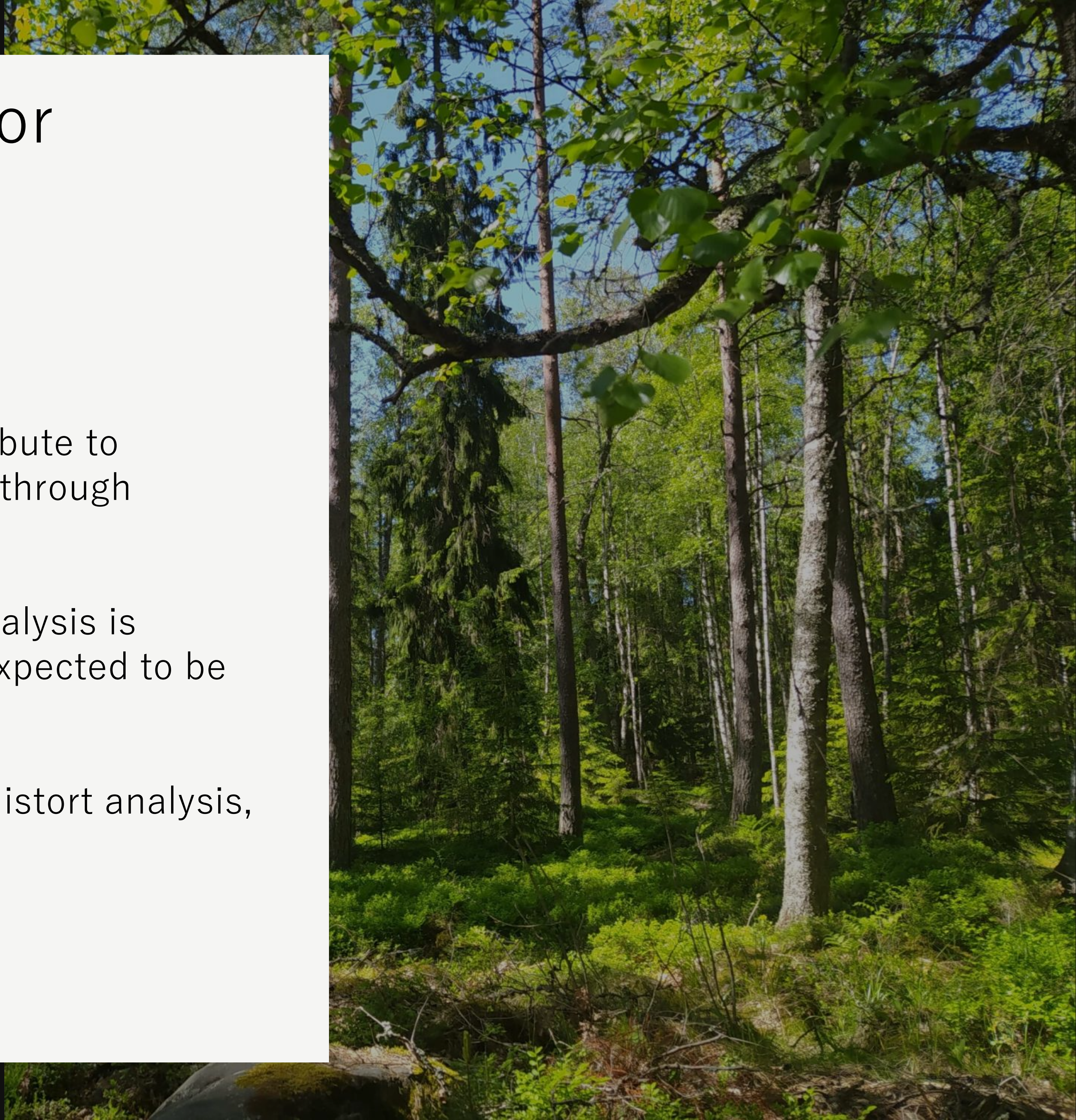
Data Screening: Checking for Outliers

To remove or not to remove?

- **Treat:** When they are valid data points, contribute to understanding variability, or can be managed through transformation or standardization

- **Ignore:** When the sample size is small, the analysis is exploratory, or the impact on conclusions is expected to be minimal

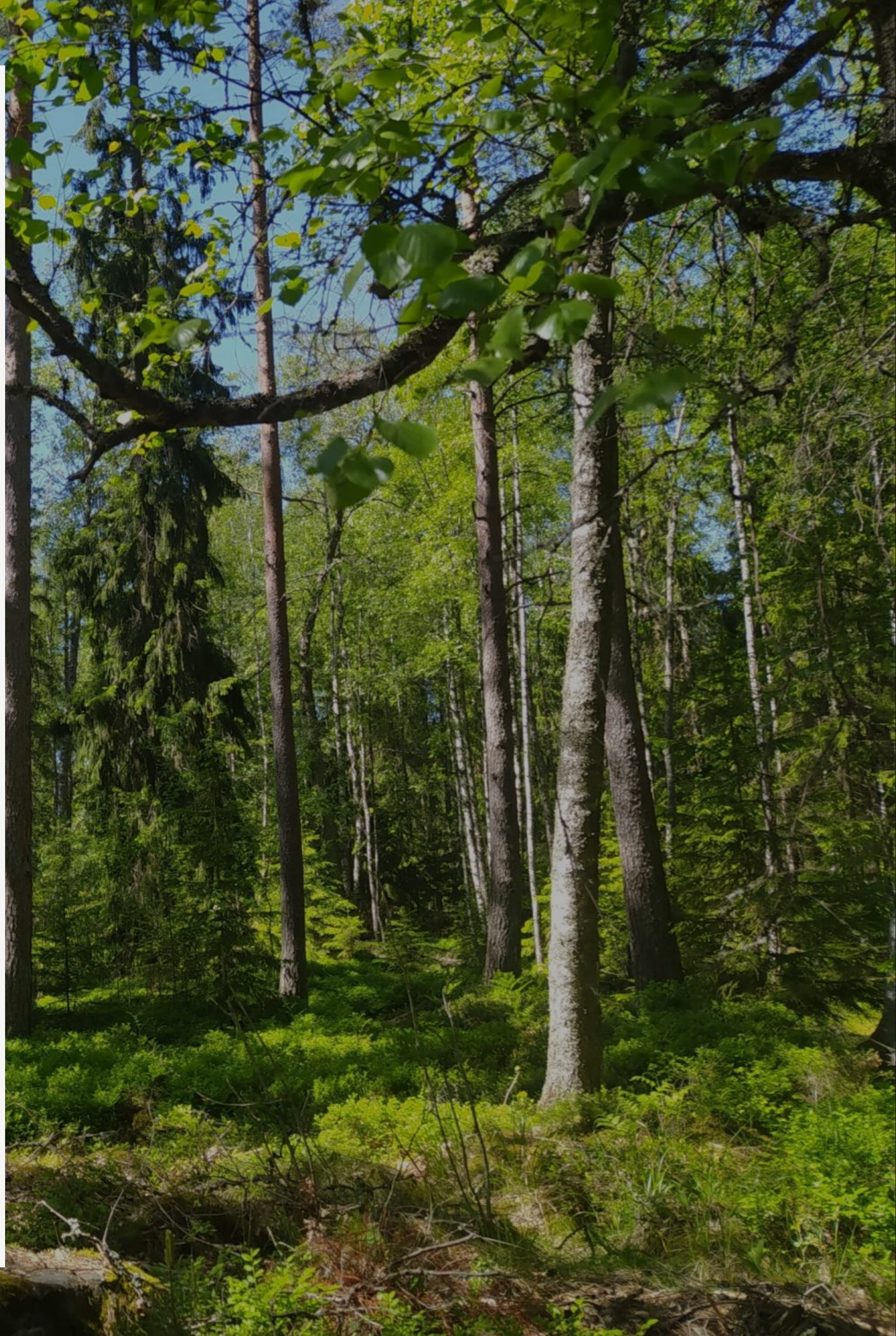
- **Remove:** When they are errors, significantly distort analysis, or do not represent the population of interest



Exploratory Analysis

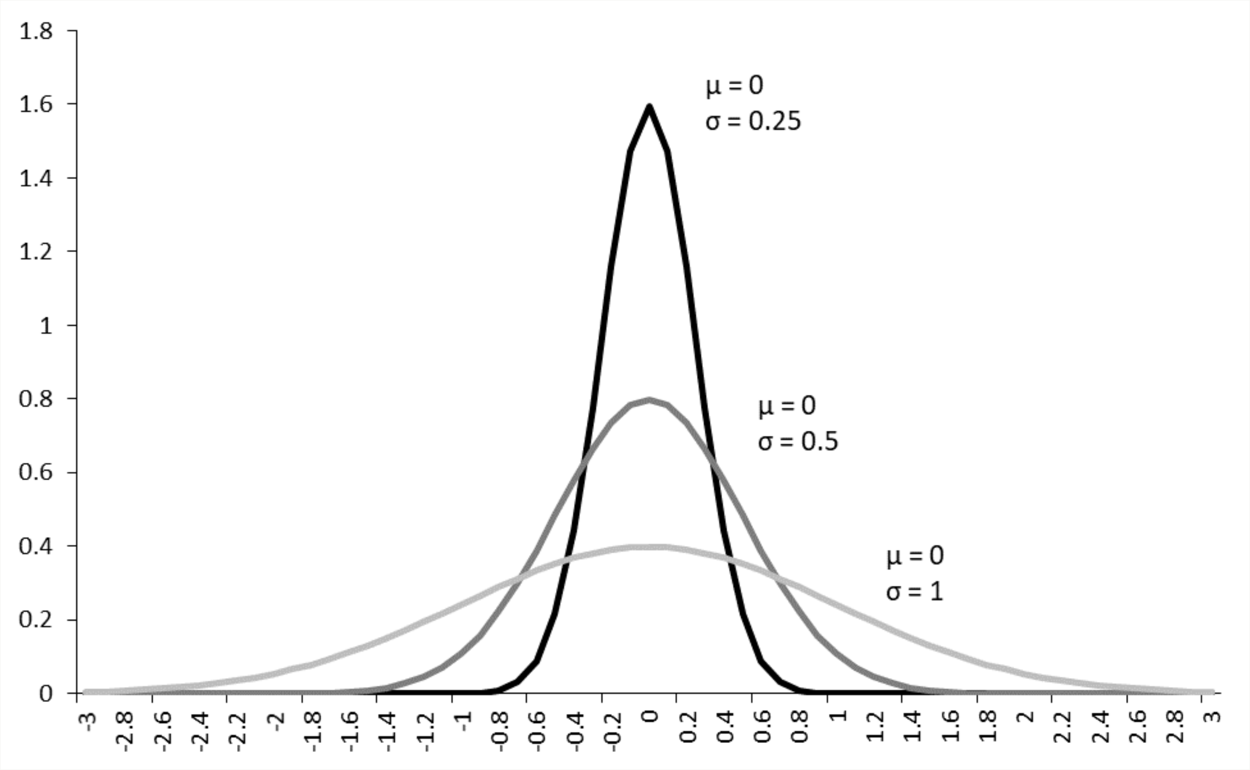
Exploratory Analysis: Data Distributions

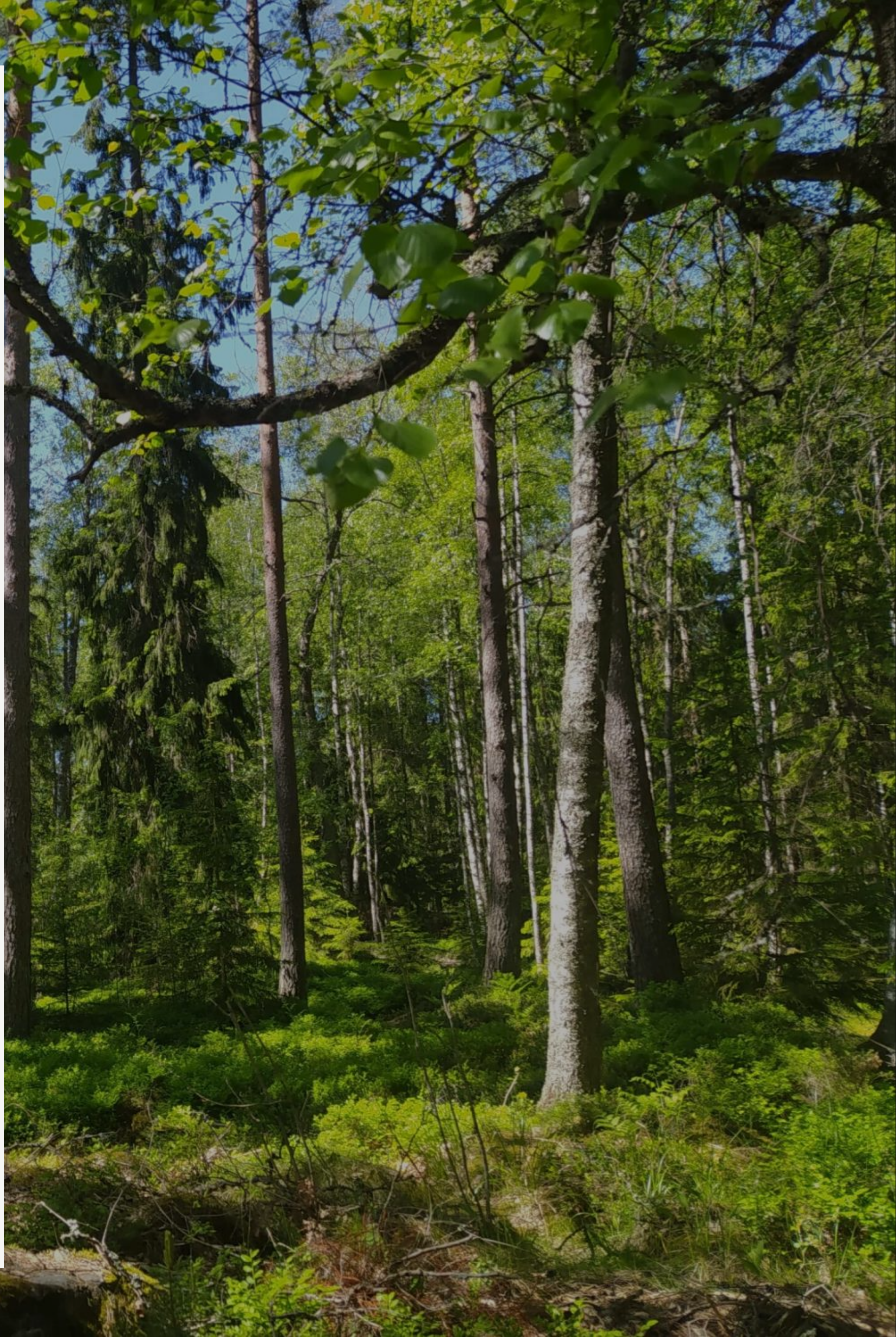
Distribution	Characteristics	Suited For...
Normal	Symmetrical, bell-shaped	Environmental variables, trait measurements
Poisson	Right-skewed, mean = variance	Integer/count data
Binomial	Can be symmetric or skewed	Presence/absence
Negative Binomial	Right-skewed, over-dispersed counts	Aggregated counts (i.e, N per unit)
Log-Normal	Right-skewed, log-transformed normal	Species abundance
Gamma	Right-skewed, flexible shape	Environmental variables
Beta	Flexible shapes, bounded on [0,1]	Proportional data
Uniform	Constant probability over interval	Indicative of complete randomness



Exploratory Analysis: Data Distributions

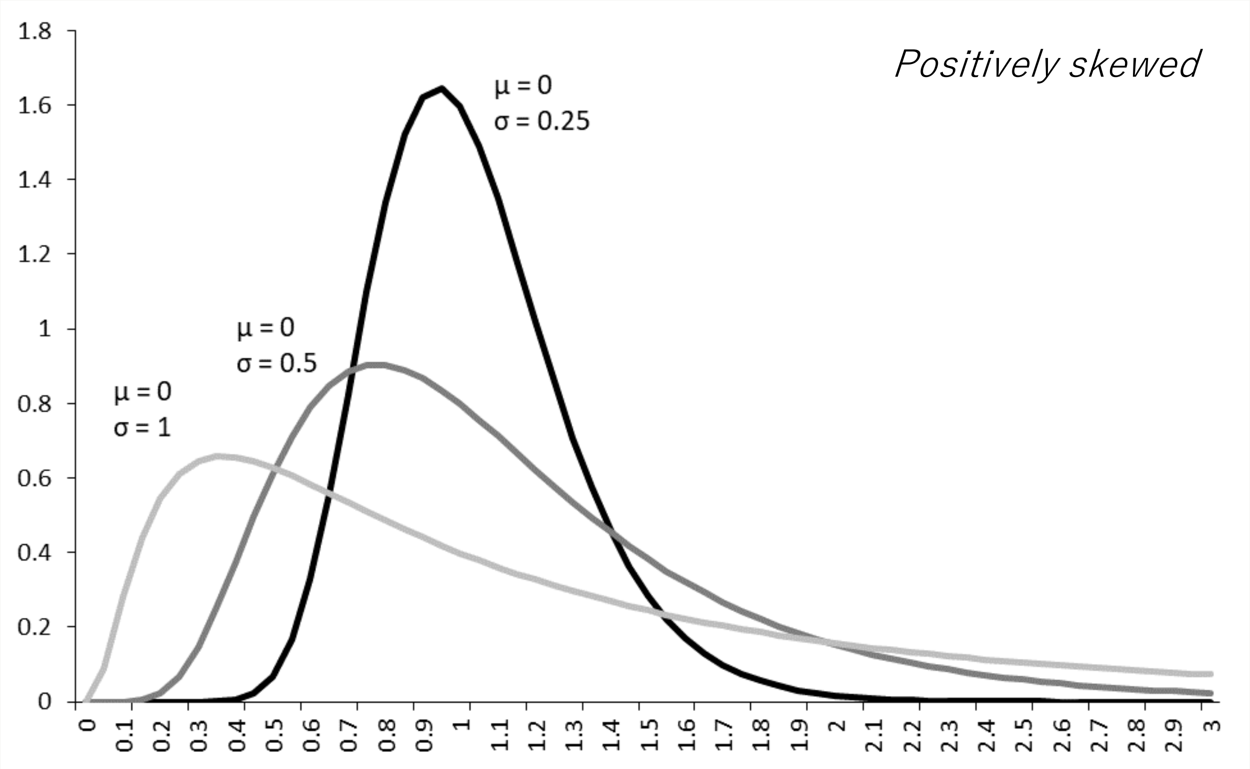
Normal/Log-Normal

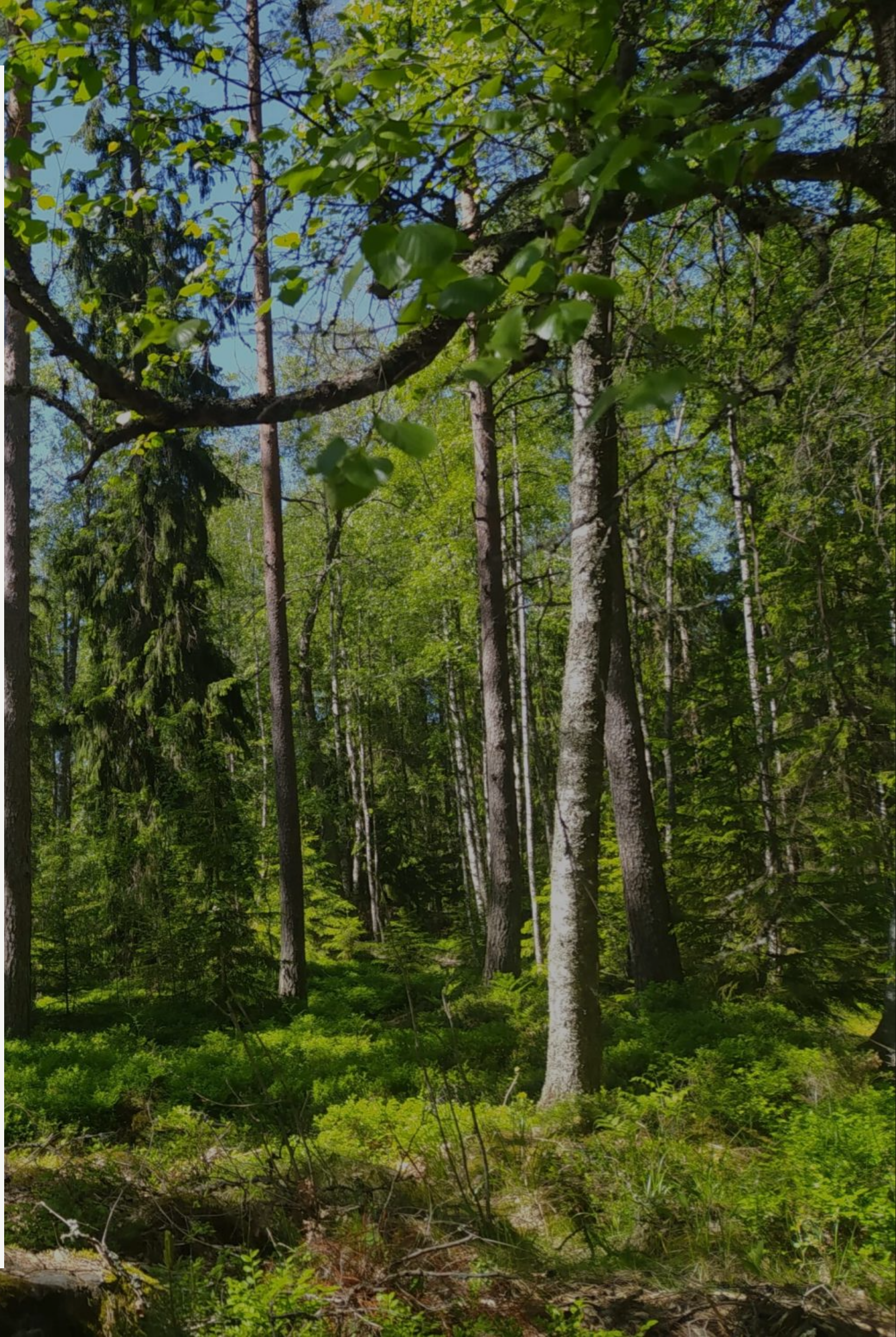




Exploratory Analysis: Data Distributions

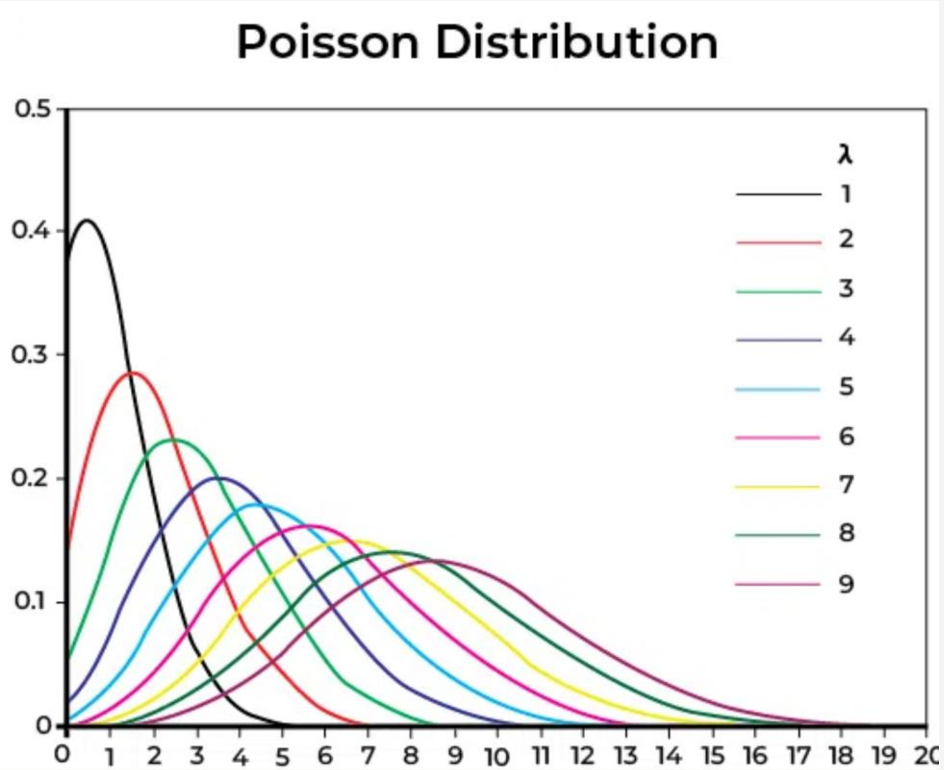
Normal/Log-Normal

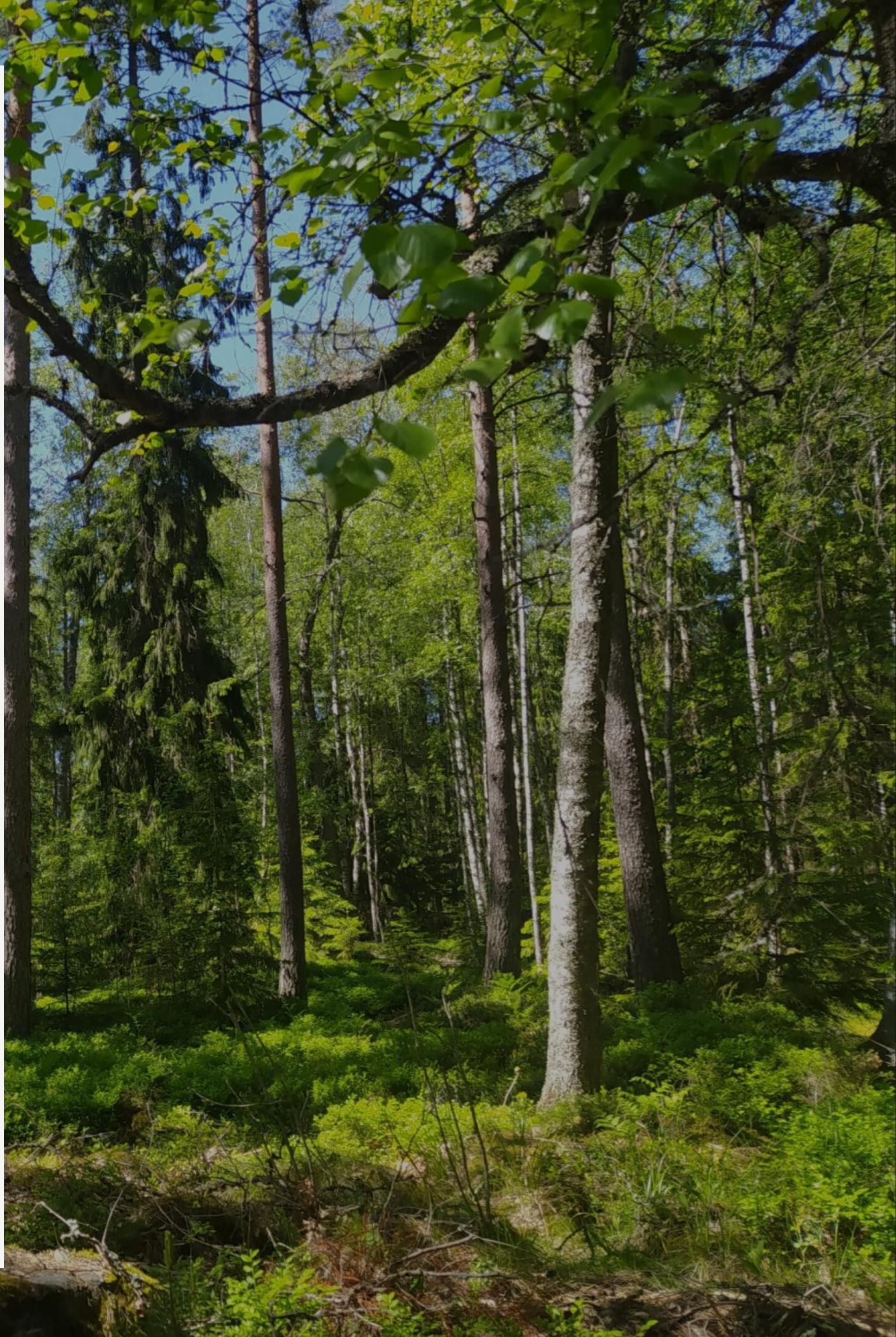




Exploratory Analysis: Data Distributions

Poisson – count data. *How many times is an event likely to occur over a given period/area?*





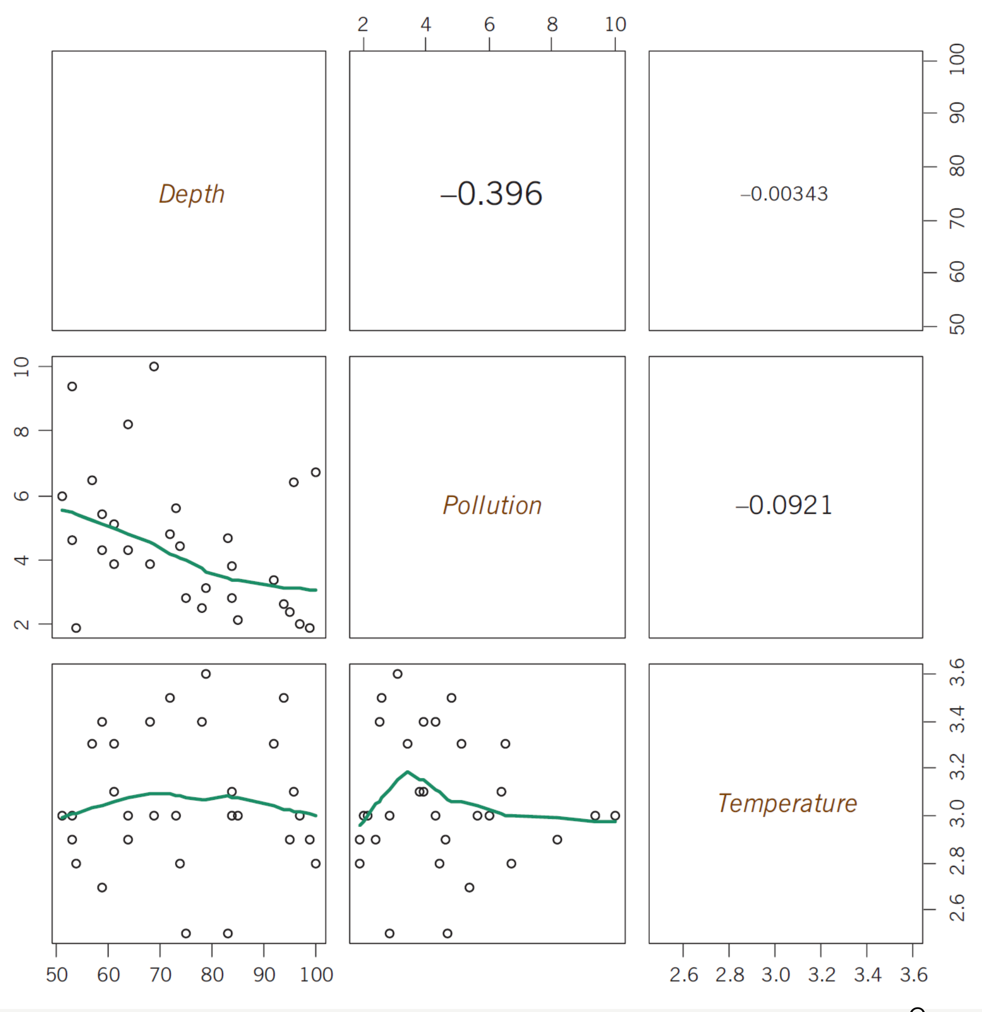
Exploratory Analysis: Homogeneity of Variance

Homoscedasticity means that the variances of the error terms are equal for all observations

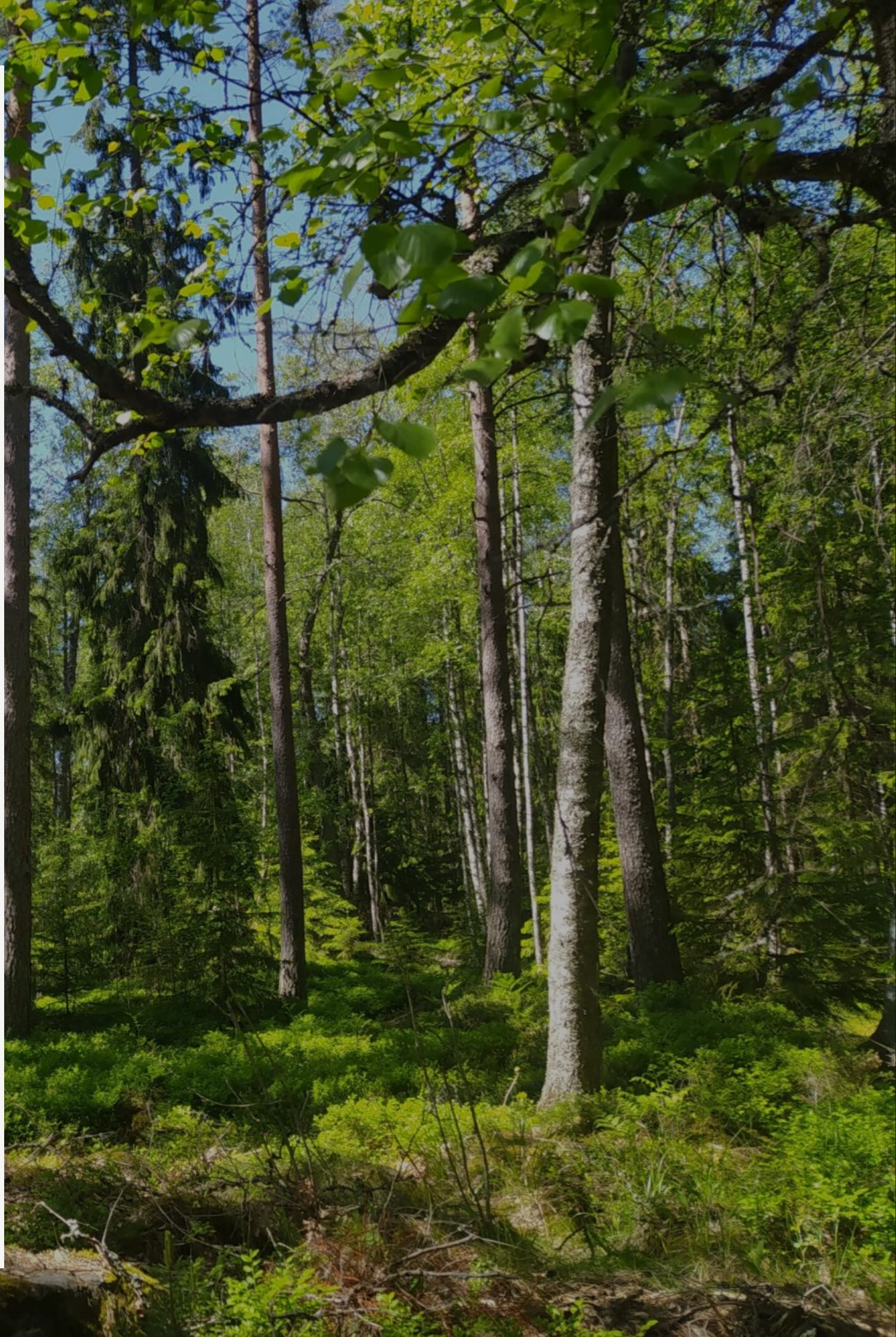
The variance in a sampled population remains the same, regardless of the mean

This means that Poisson distributed data are not homoscedastic!

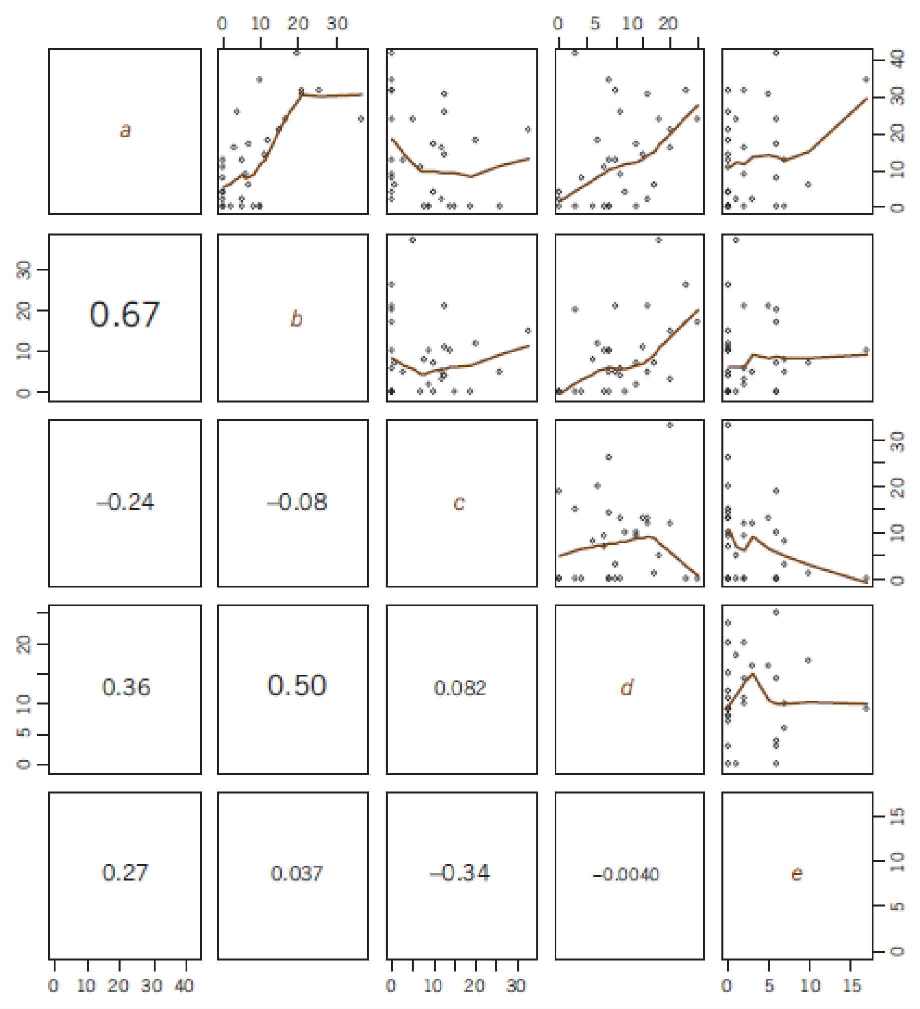
Exploratory Analysis: Multicollinearity



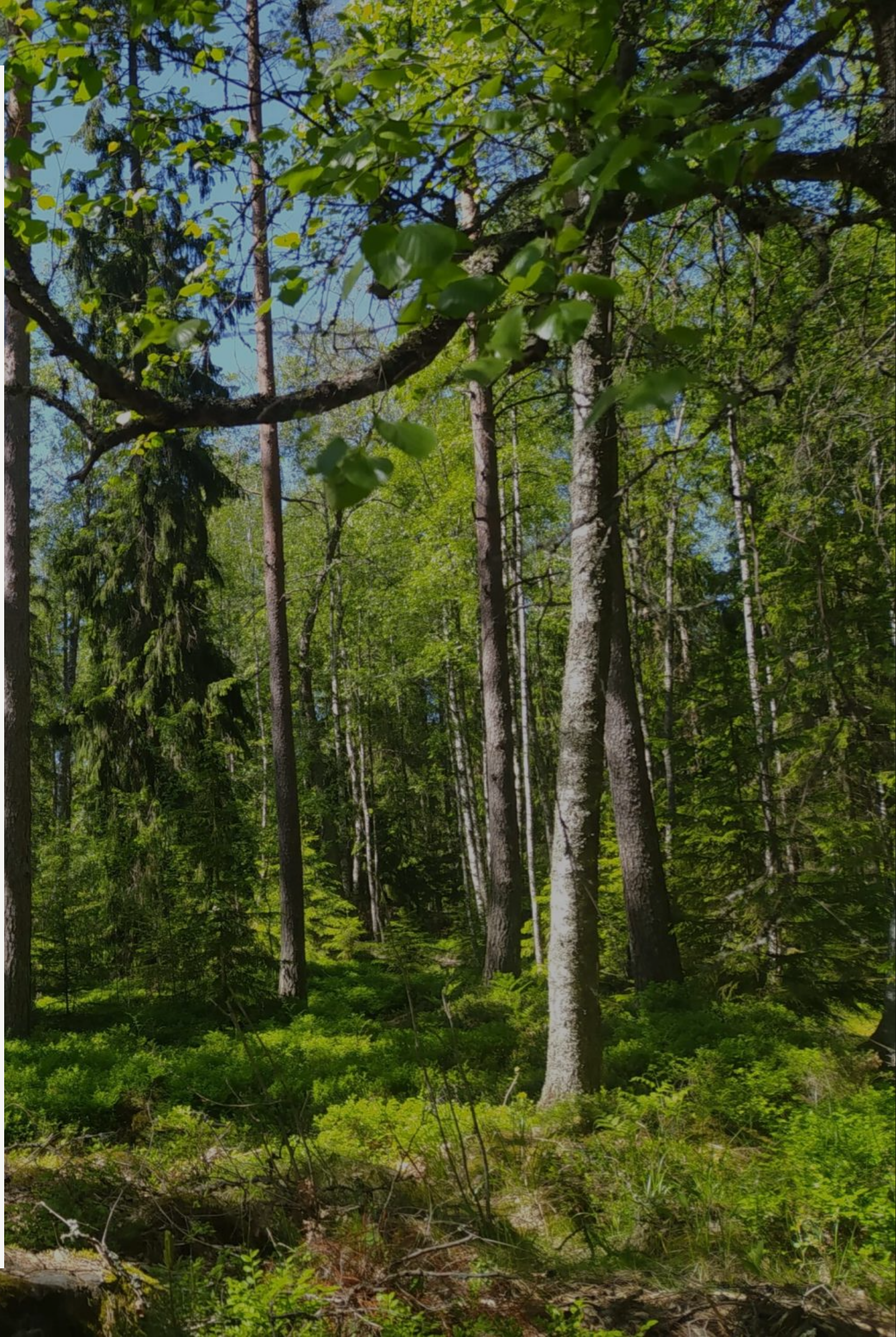
Greenacre and Primicerio



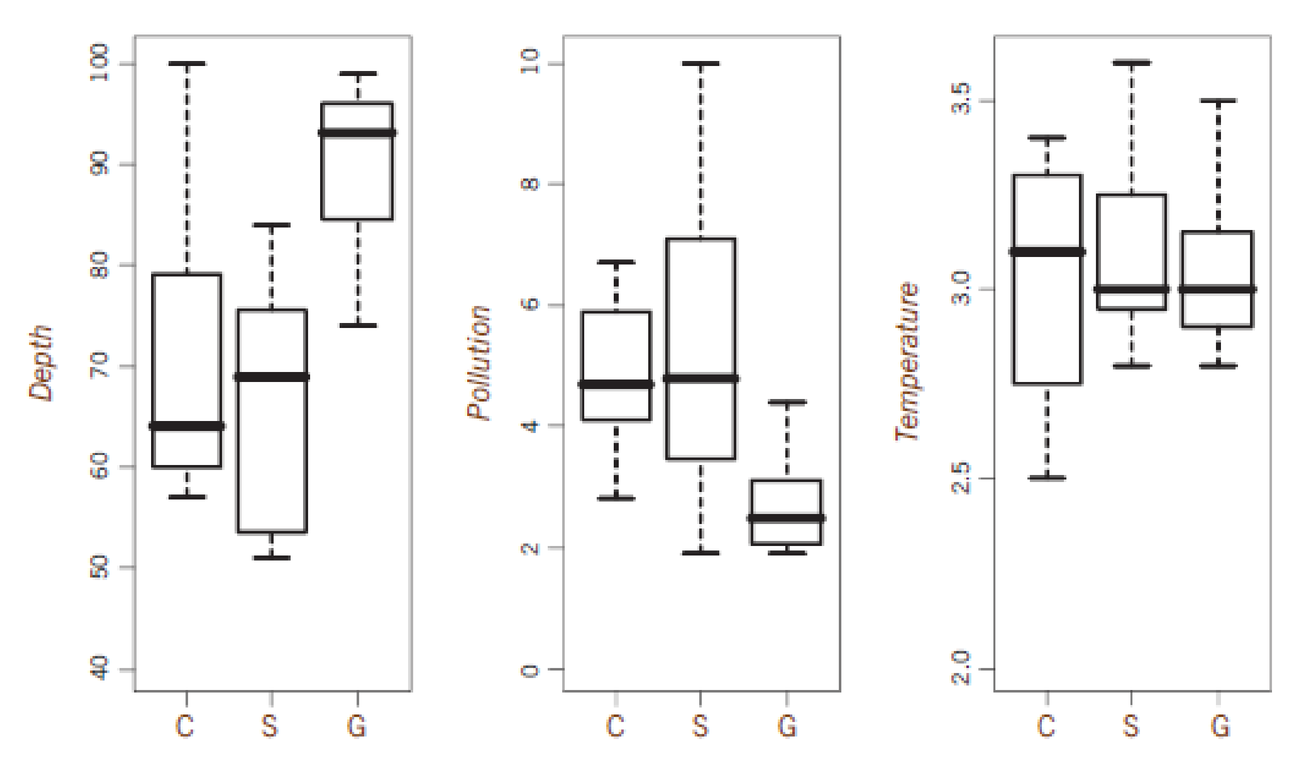
Exploratory Analysis: Multicollinearity



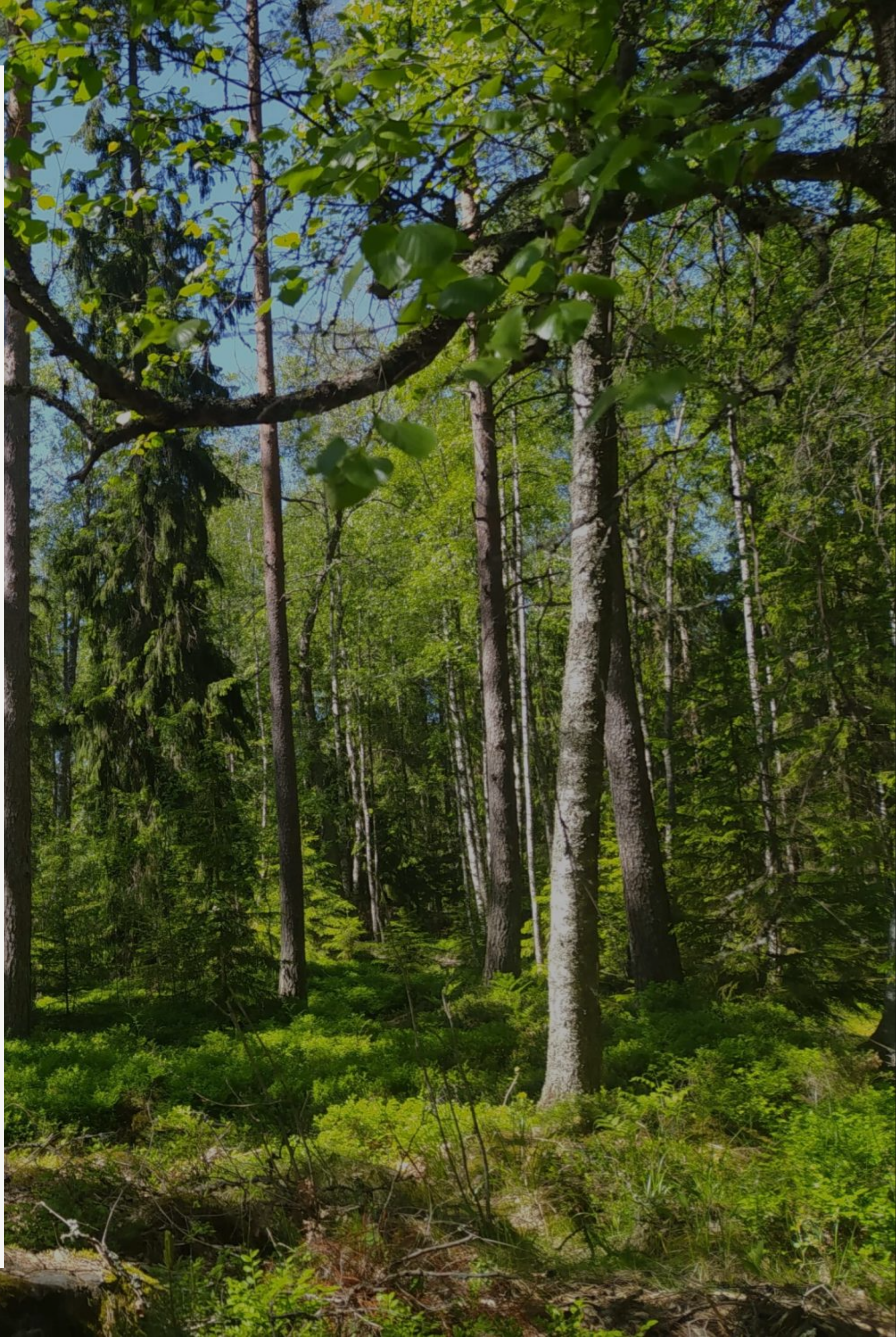
Greenacre and Primicerio



Exploratory Analysis: Multicollinearity

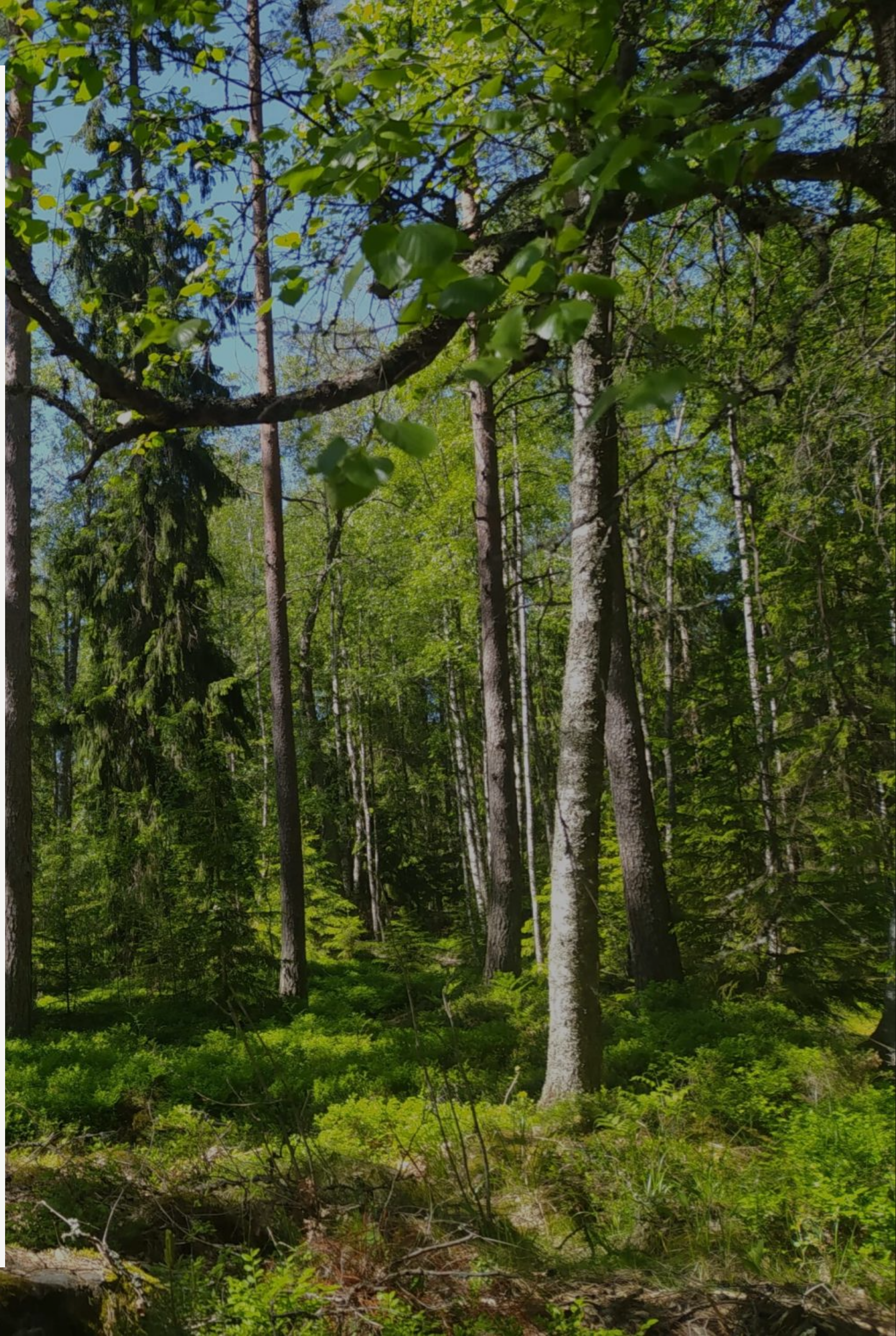


Greenacre and Primicerio



Conclusion: Summary of Key Points

- Evaluate **structure** of response and predictor variables
- Check for and treat **missing data** and **outliers** as needed
- Check for **multicollinearity** (especially for predictor variables)



Questions?