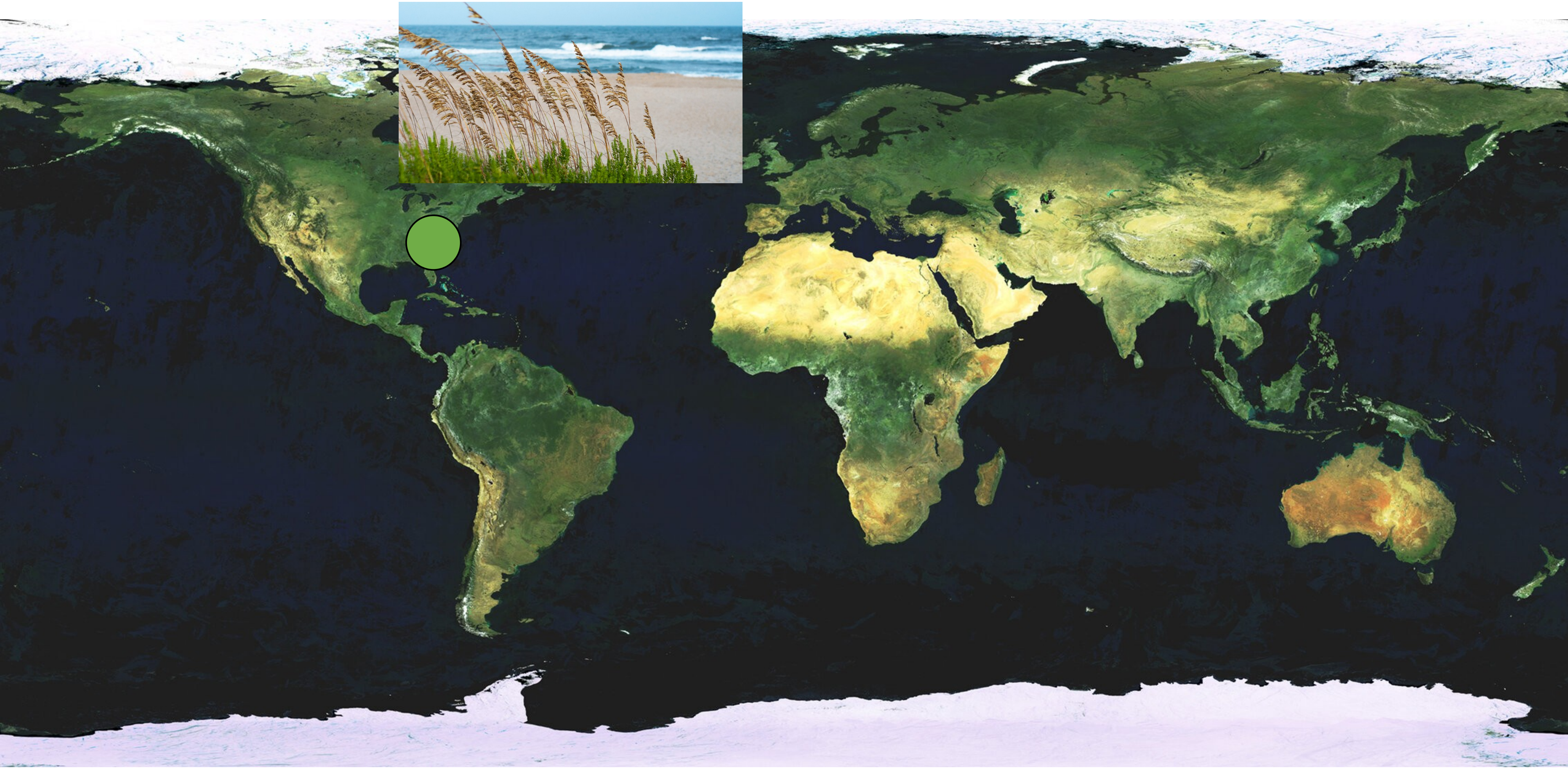
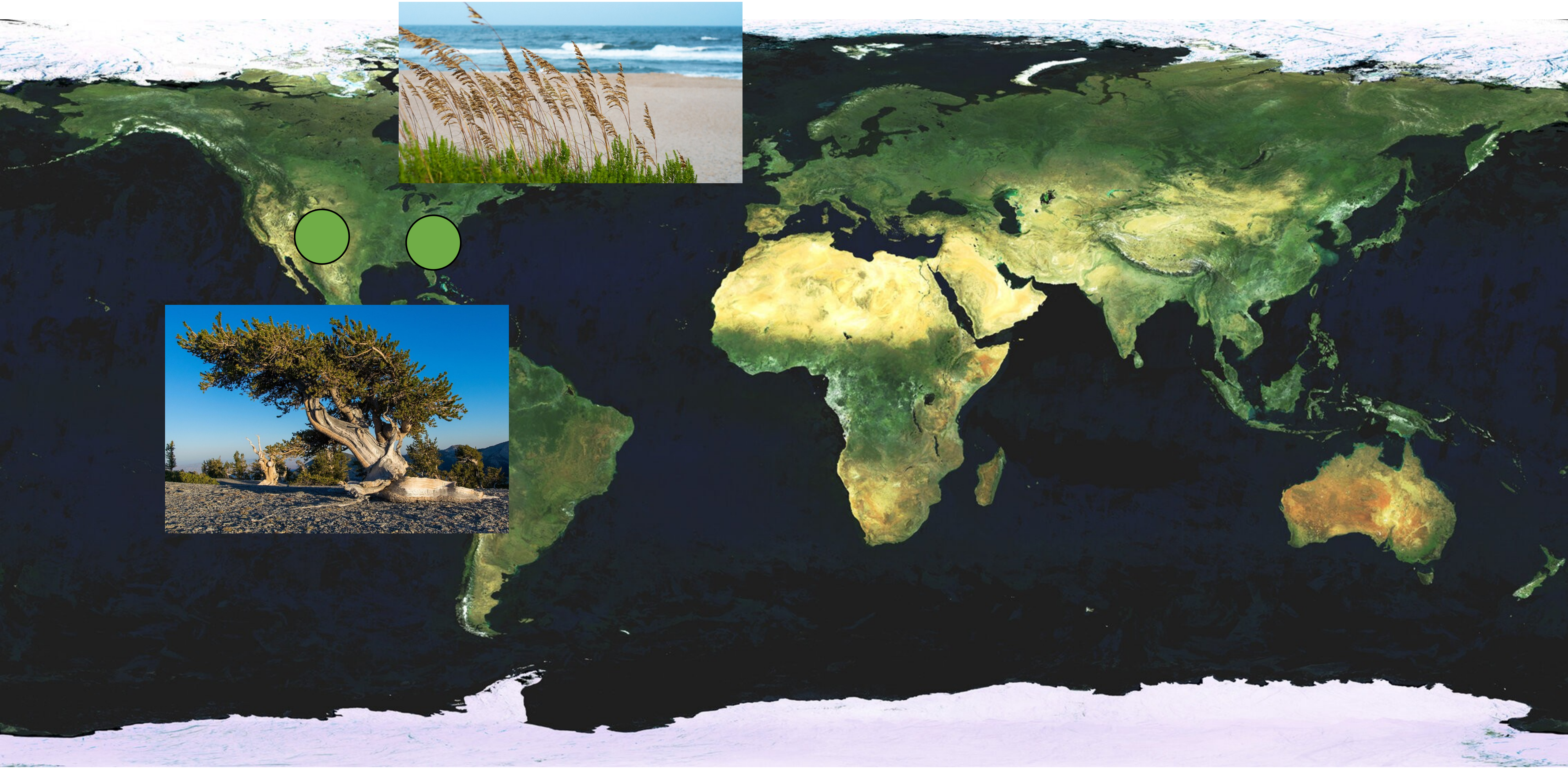
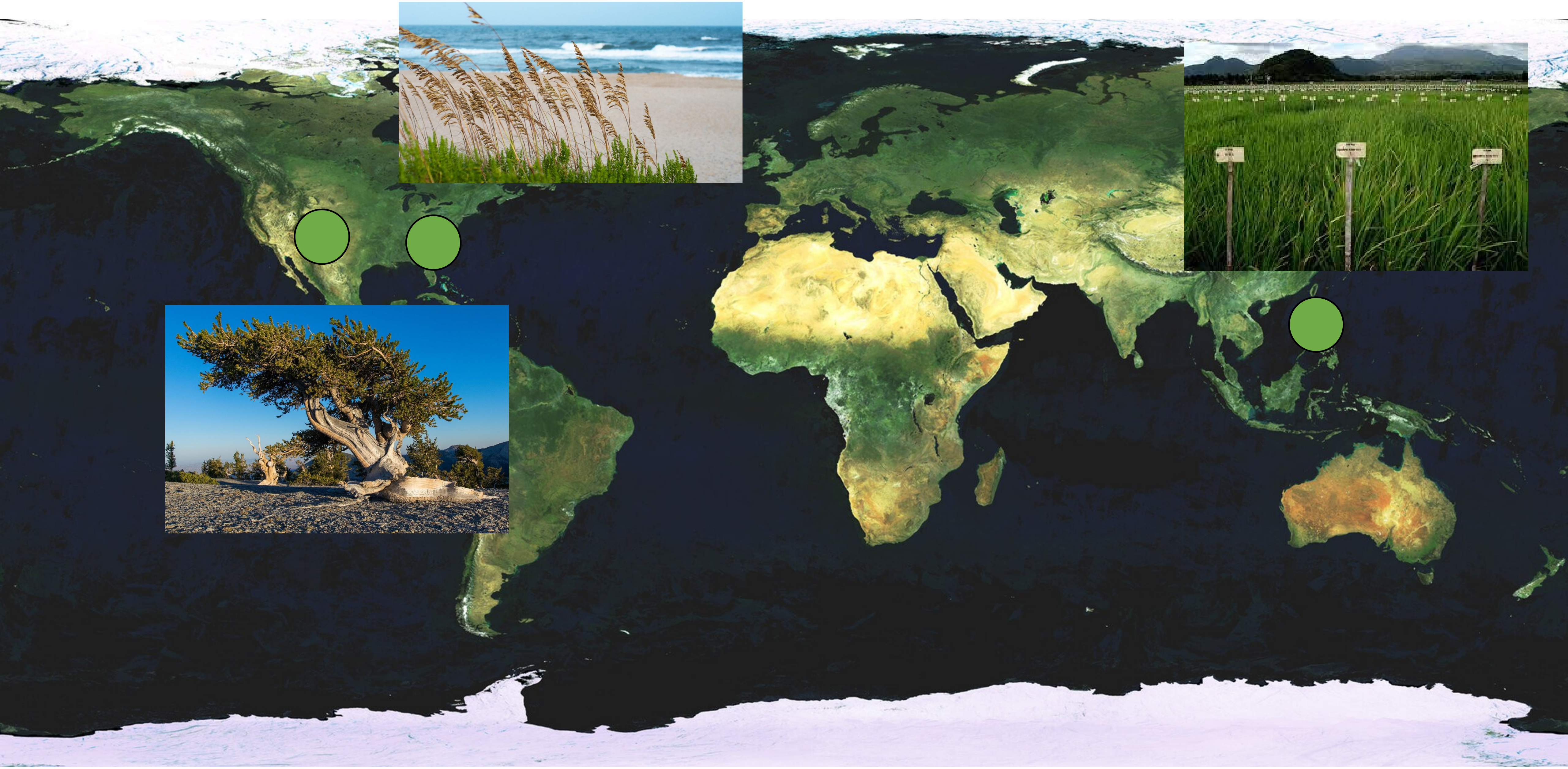
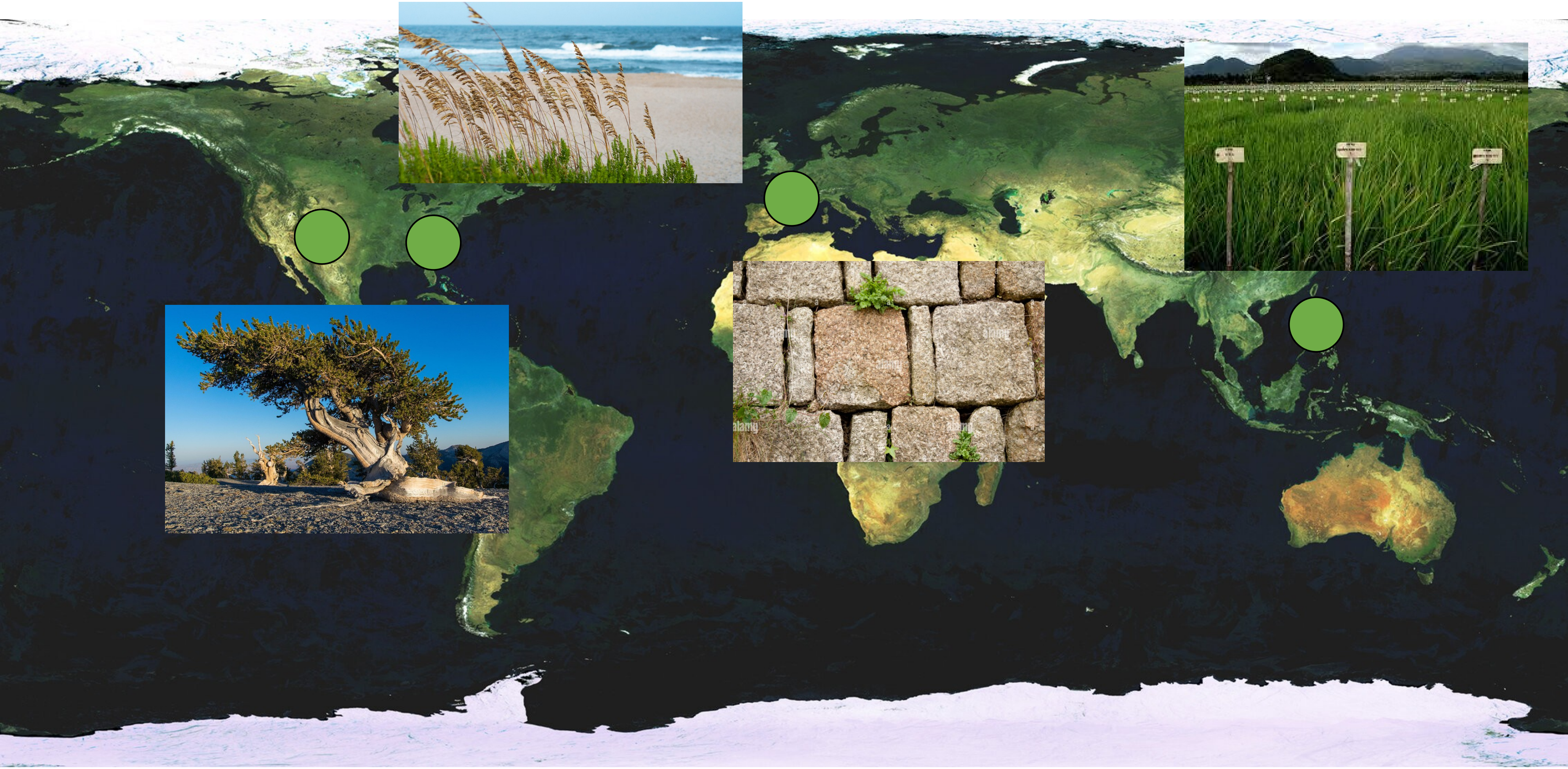
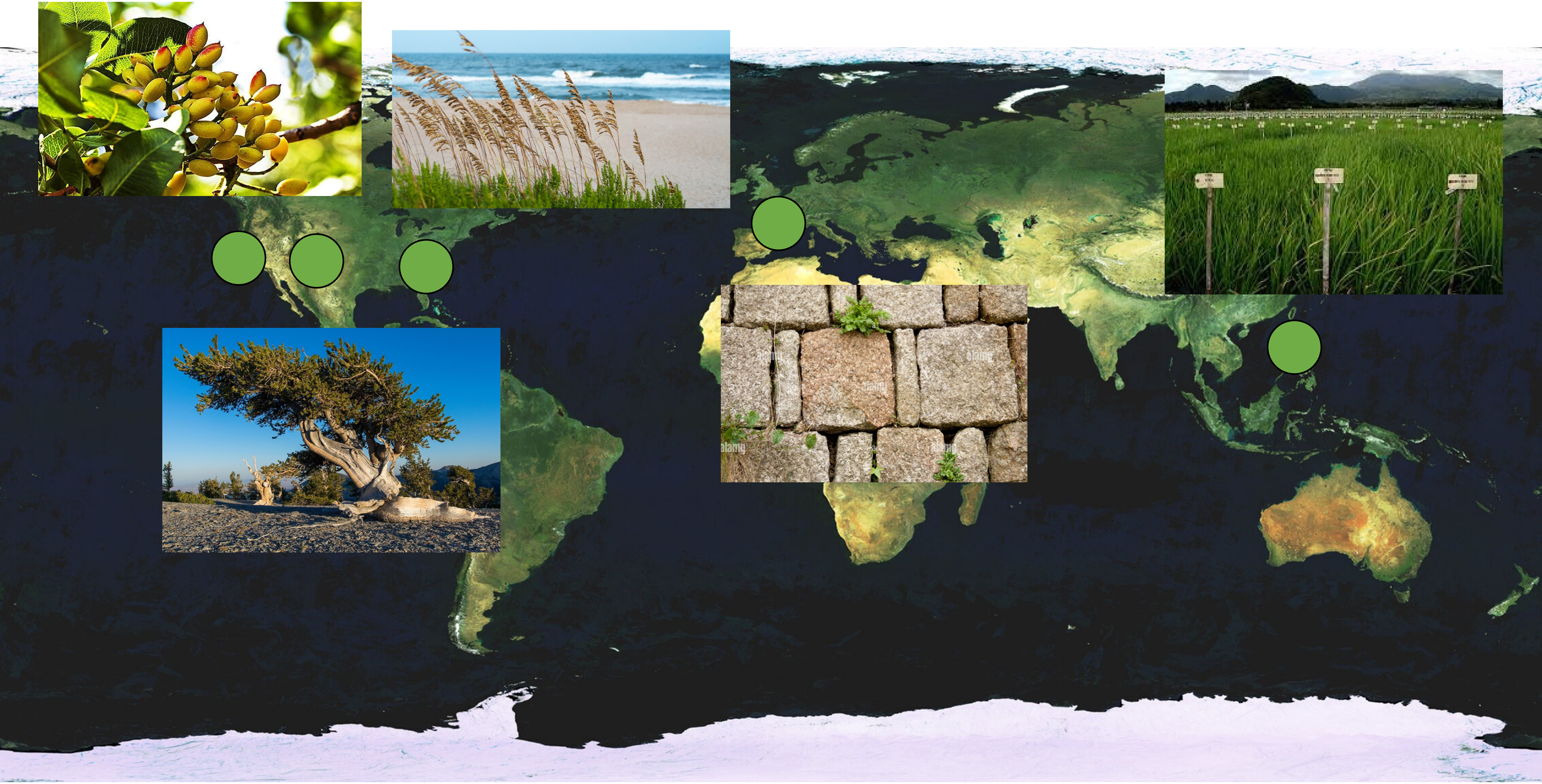
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**



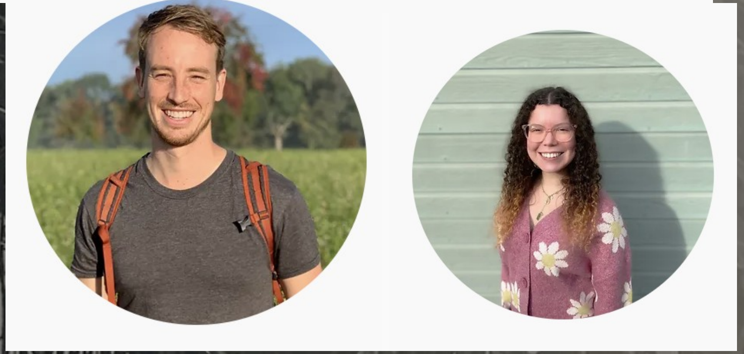






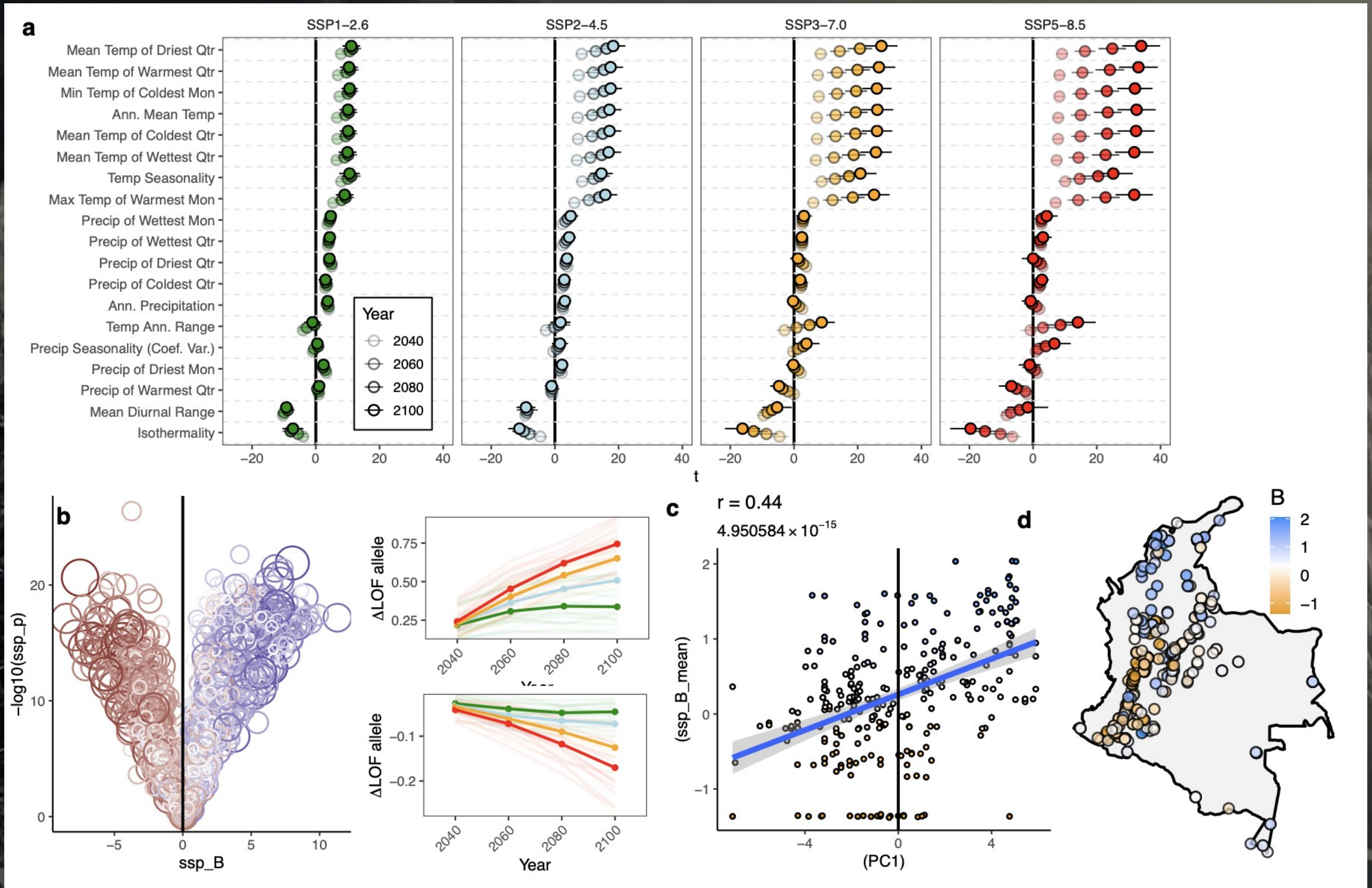


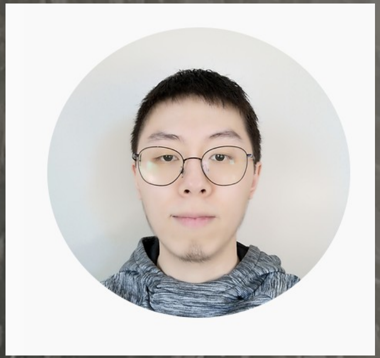
Monroe Lab @ UC Davis



Plant genomics on a changing planet

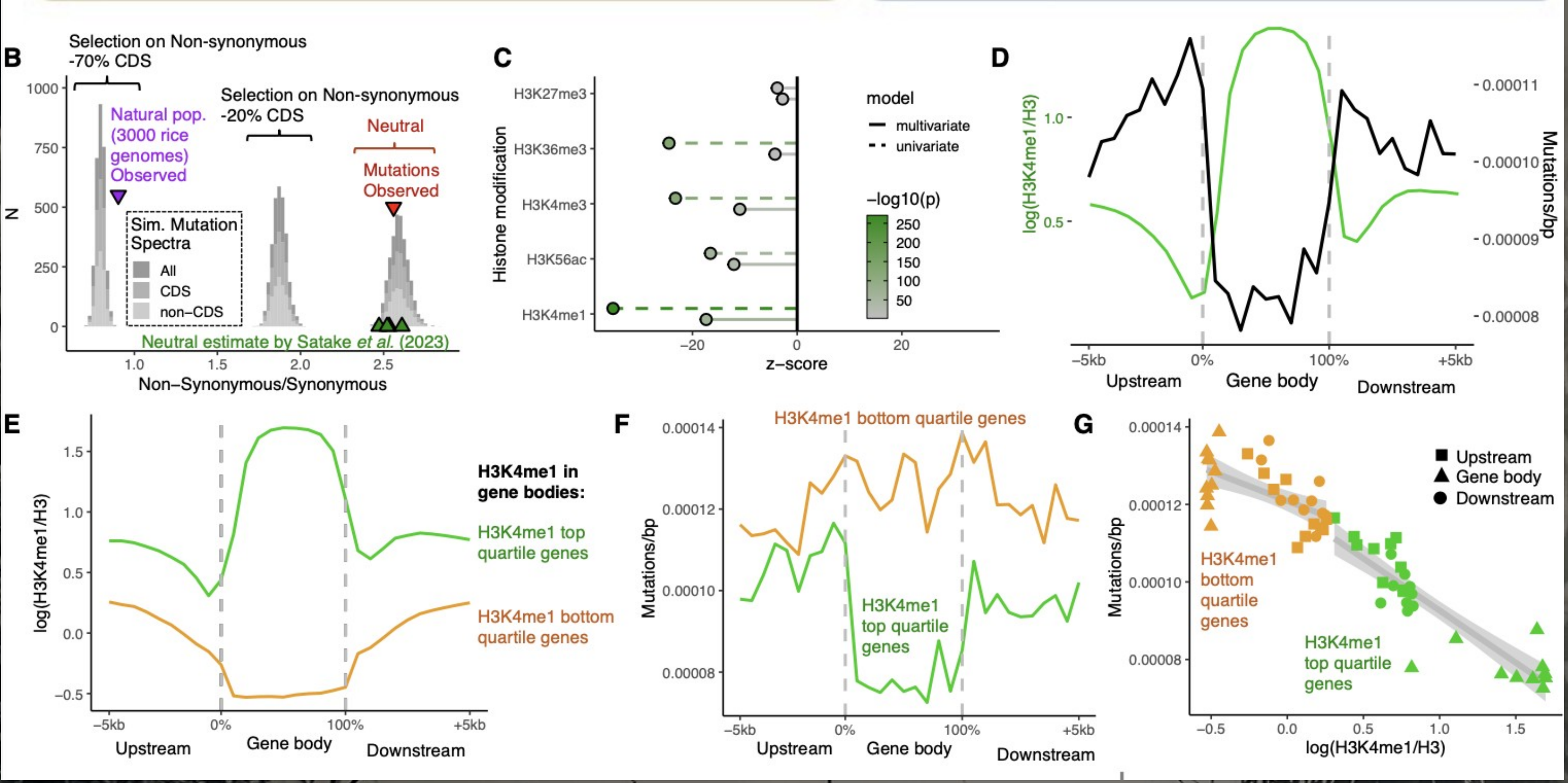
We study genetic variation in plants at the intersection of evolutionary theory, functional population genomics, and molecular biology.

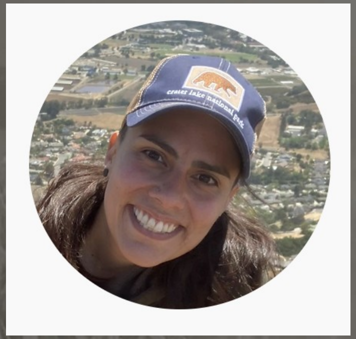




Plant genomics on a changing planet

We study genetic variation in plants at the intersection of evolutionary theory, functional population genomics, and molecular biology.





Plant genomics on a changing planet

We study genetic variation in plants at the intersection of evolutionary theory, functional population genomics, and molecular biology.

slido

Please download and install the Slido app on all computers you use

What graduate group are you in?

ⓘ Start presenting to display the poll results on this slide.

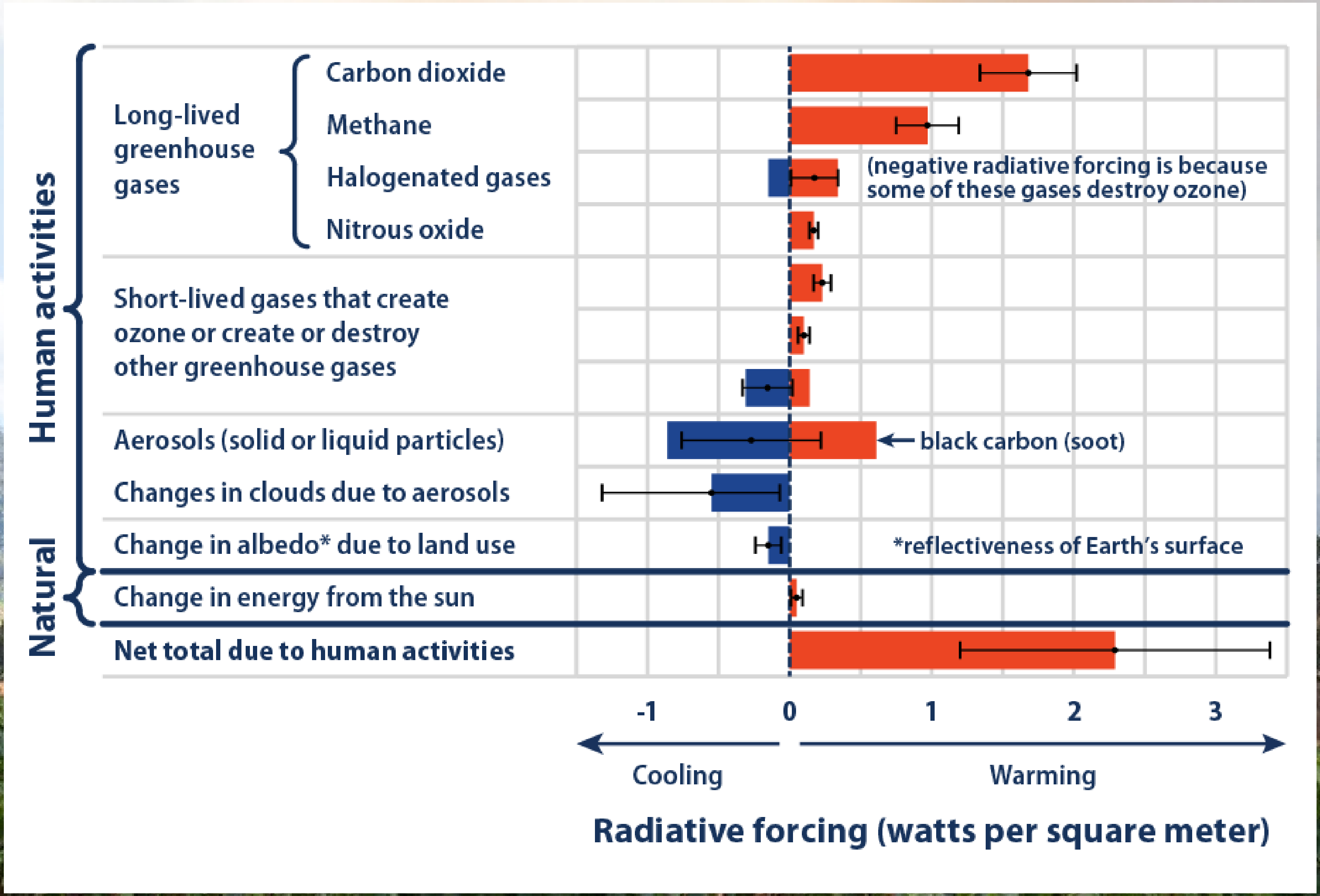
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Today:

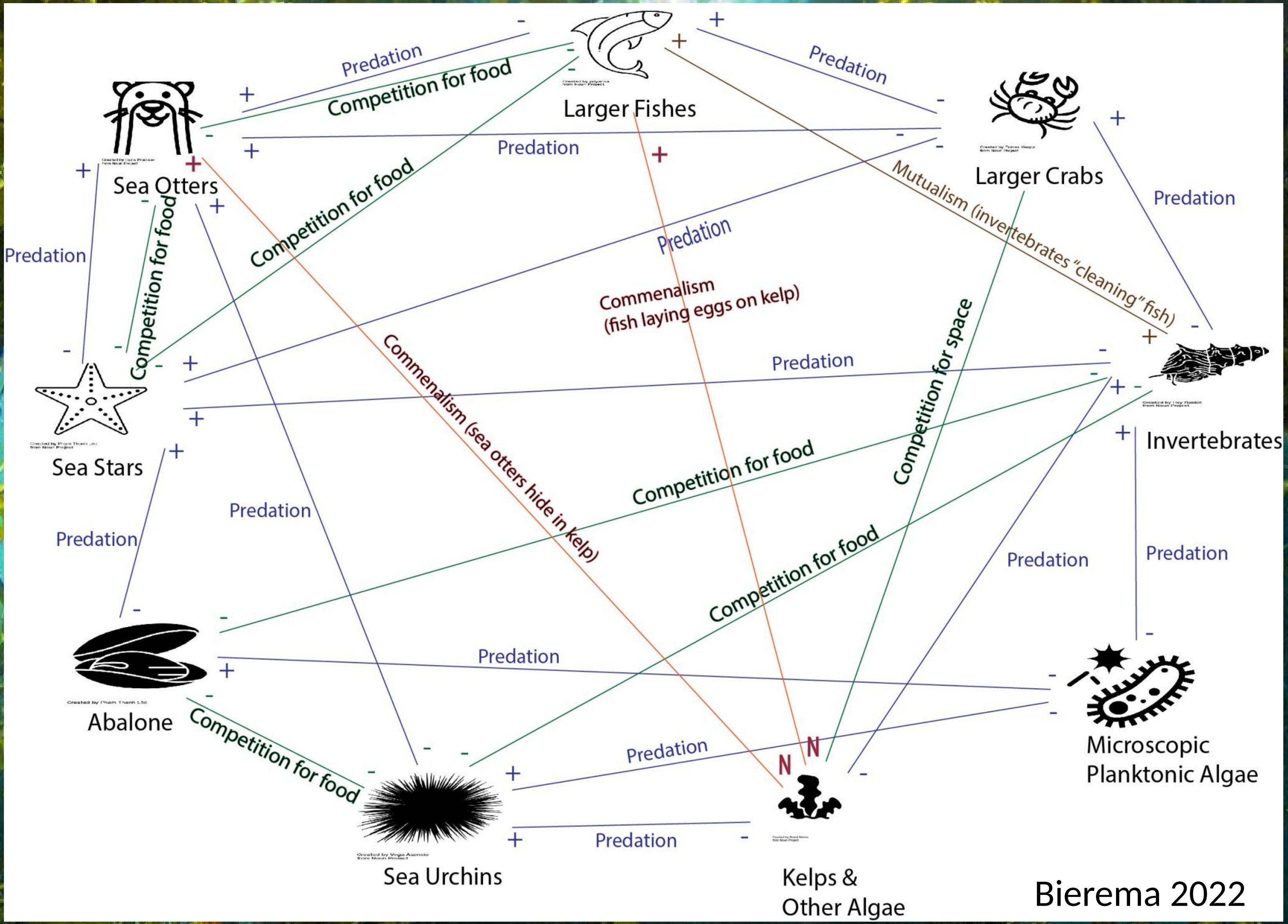
Why multivariate statistics matters

Intro to the course

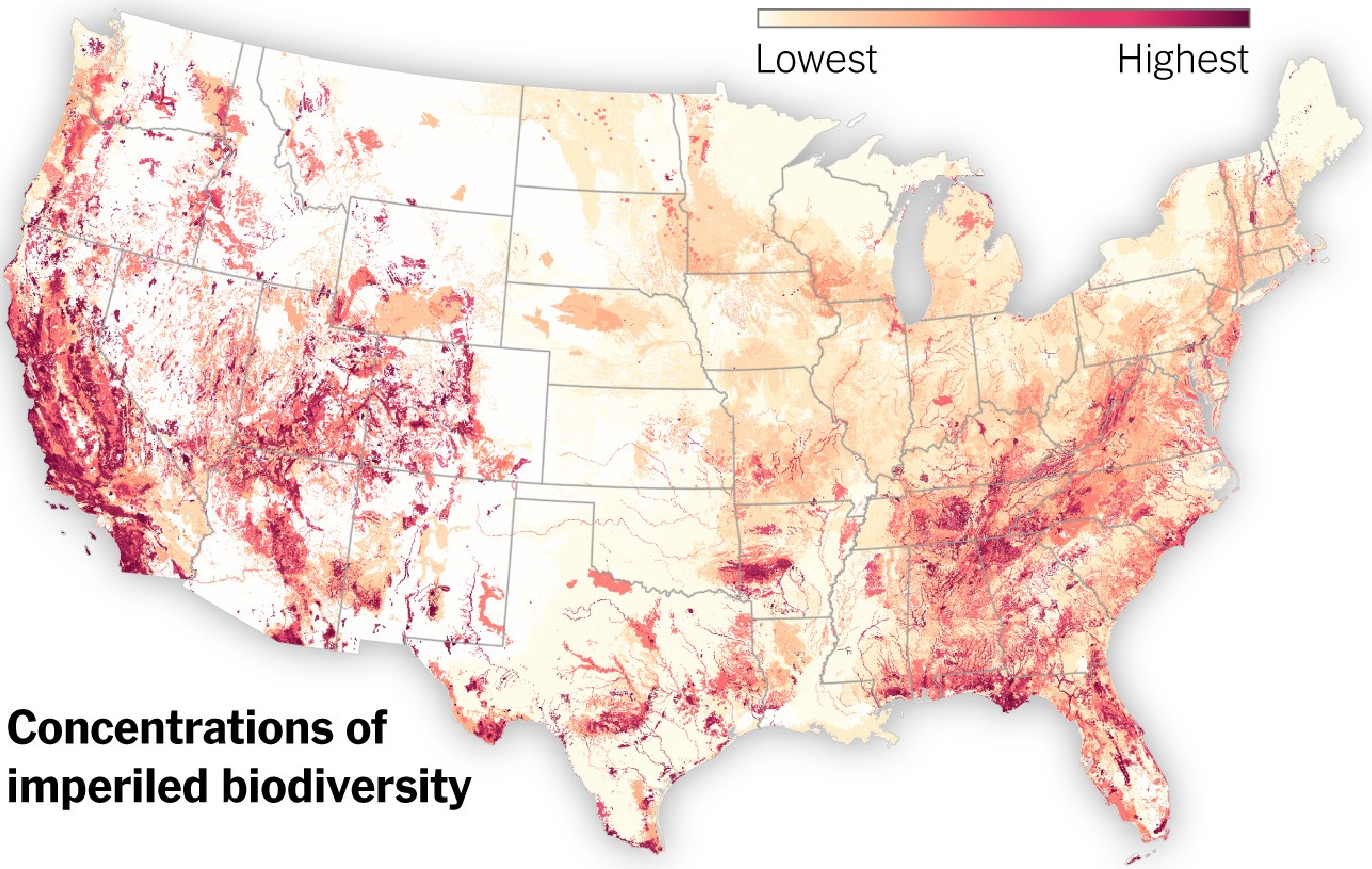

Src: CalCAN

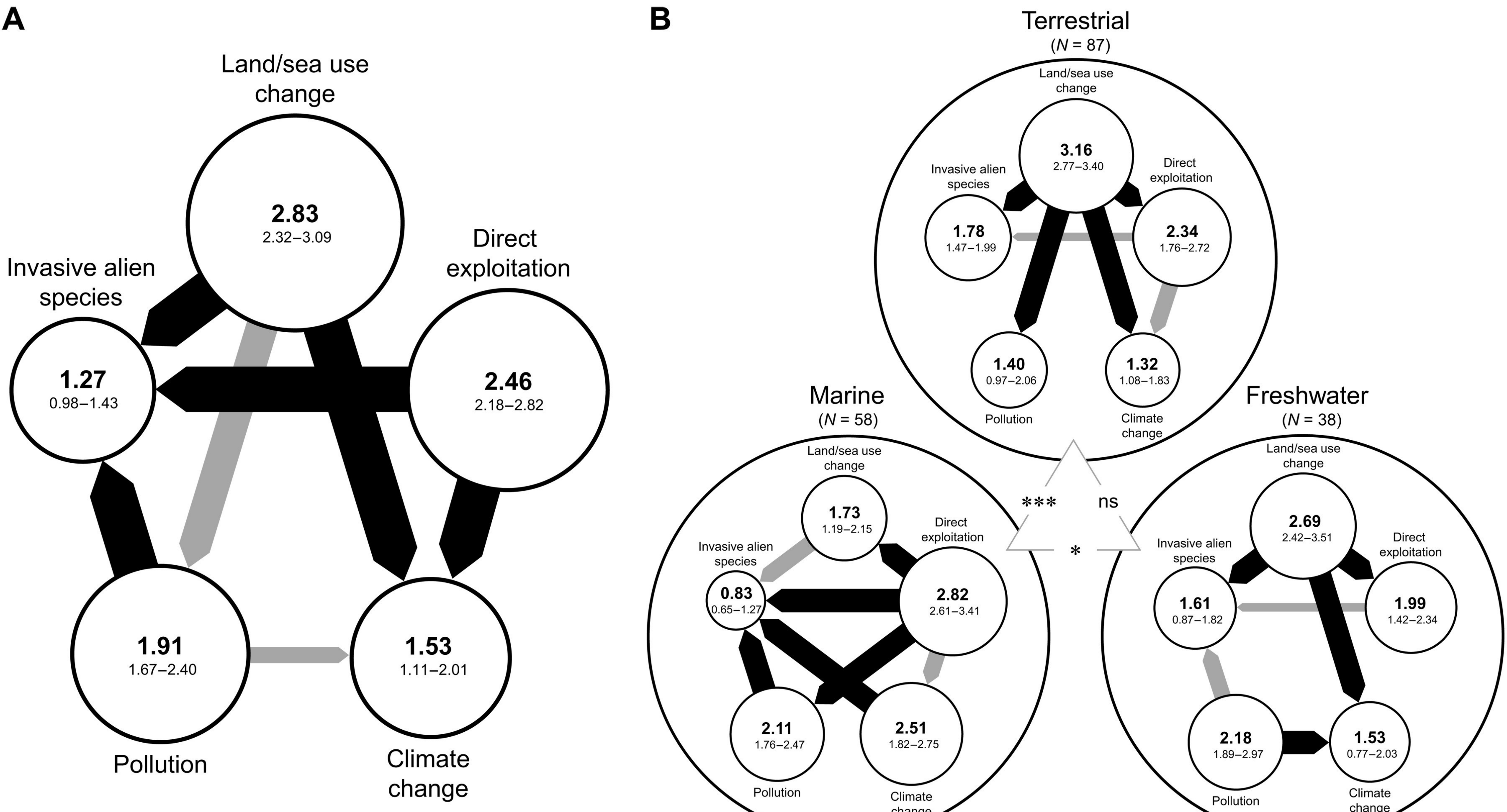

Human activities
Natural
Long-lived greenhouse gases
Carbon dioxide
Methane
Halogenated gases
Nitrous oxide
(negative radiative forcing is because some of these gases destroy ozone)
Short-lived gases that create ozone or create or destroy other greenhouse gases
Aerosols (solid or liquid particles)
black carbon (soot)
Changes in clouds due to aerosols
Change in albedo* due to land use
*reflectiveness of Earth's surface
Change in energy from the sun
Net total due to human activities
-1
0
1
2
3
Cooling
Warming
Radiative forcing (watts per square meter)

Src: CalCAN



Bierema 2022

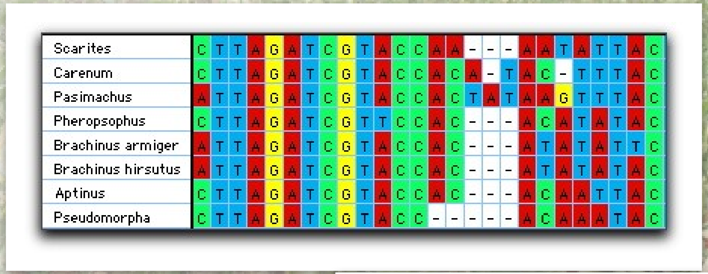

Lowest
Highest
Concentrations of imperiled biodiversity

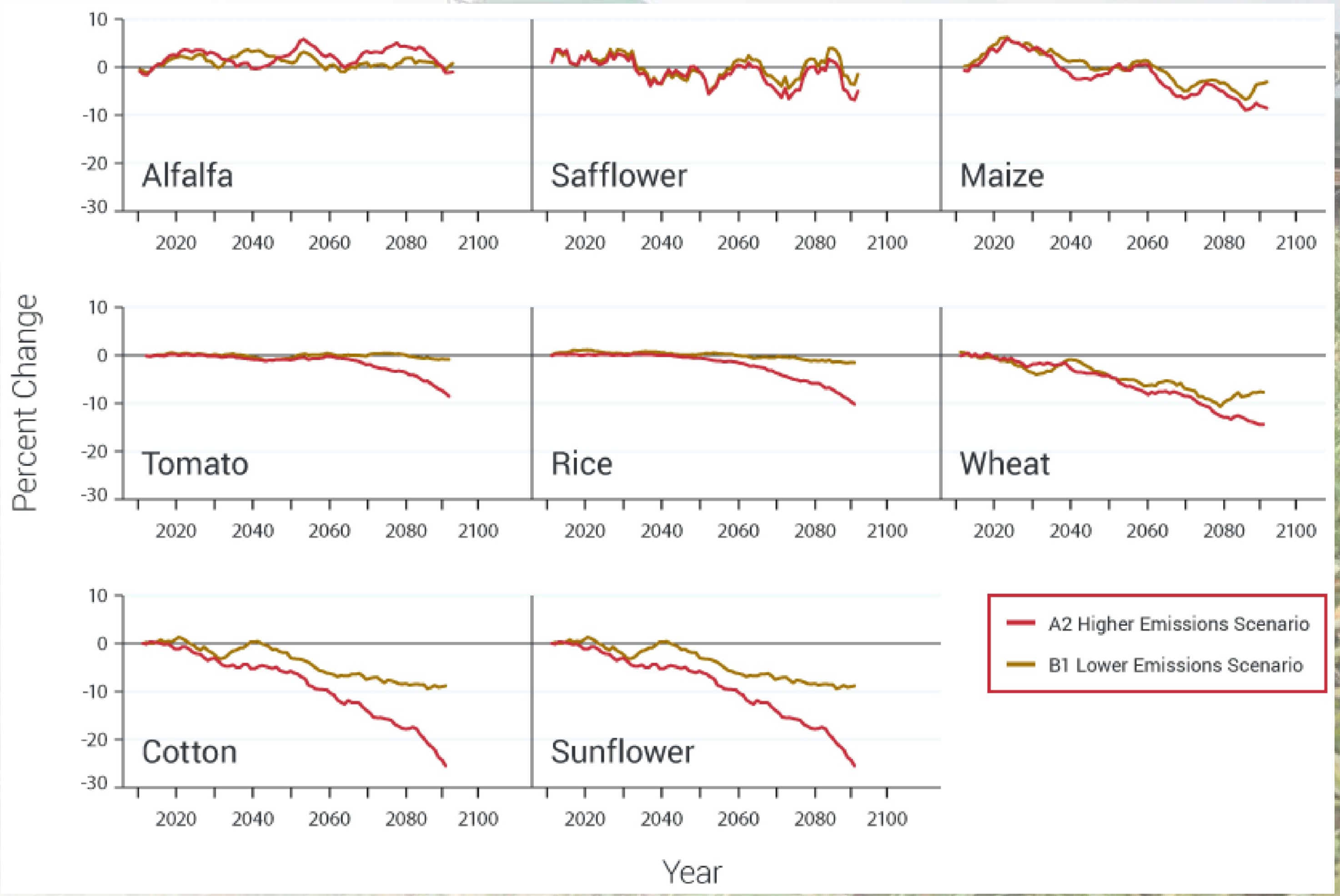


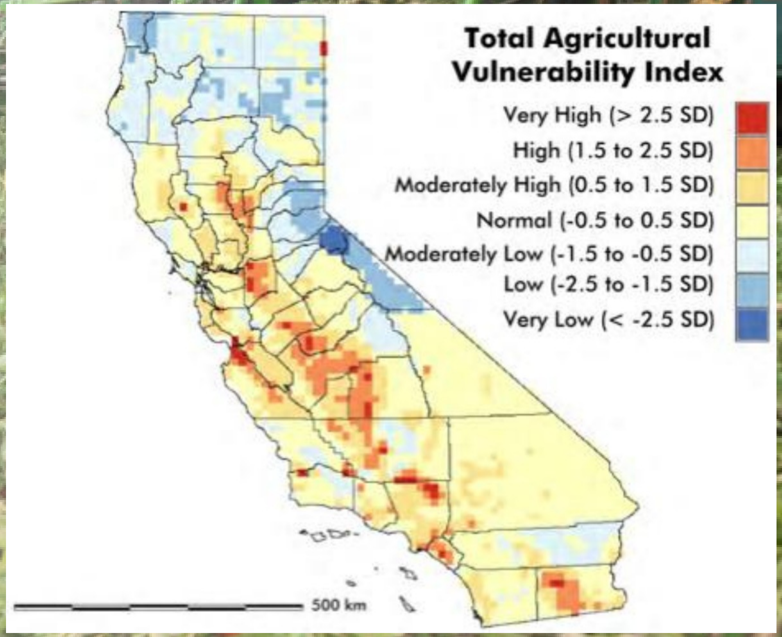
Jaureguiberry et al, 2022


agraz
670

Crop yield = **genotype** x **environment** x **management**

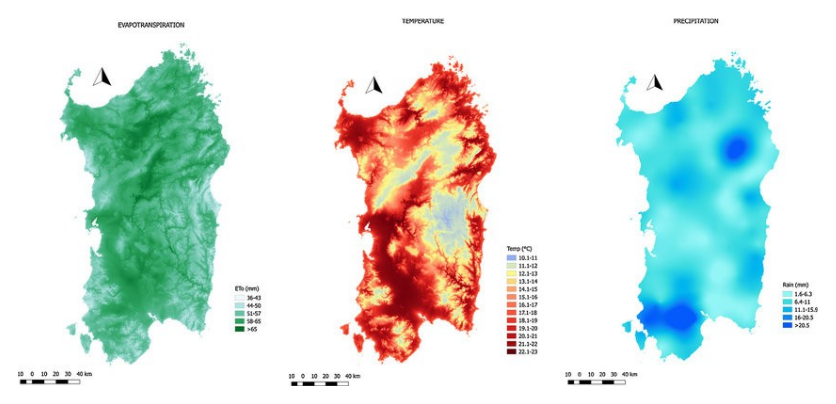



Alfalfa
Safflower
Maize
Tomato
Rice
Wheat
Cotton
Sunflower
Percent Change
Year
10
0
-10
-20
-30
2020
2040
2060
2080
2100
A2 Higher Emissions Scenario
B1 Lower Emissions Scenario

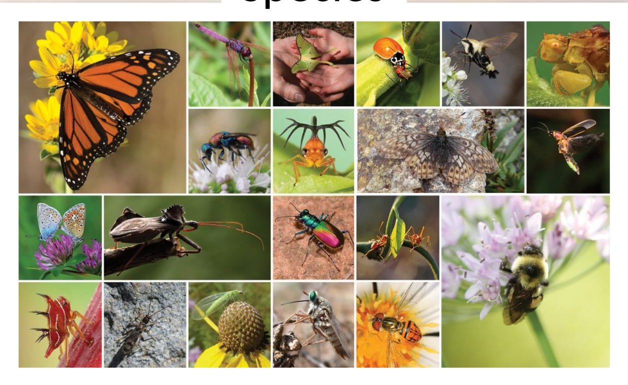

Total Agricultural Vulnerability Index
Very High (> 2.5 SD)
High (1.5 to 2.5 SD)
Moderately High (0.5 to 1.5 SD)
Normal (-0.5 to 0.5 SD)
Moderately Low (-1.5 to -0.5 SD)
Low (-2.5 to -1.5 SD)
Very Low (< -2.5 SD)
500 km

Multivariate data in agricultural and environmental sciences

Climate

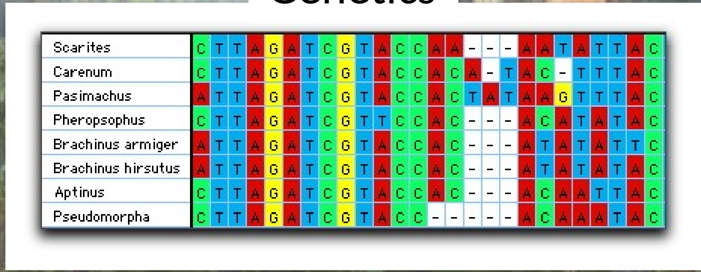


Species

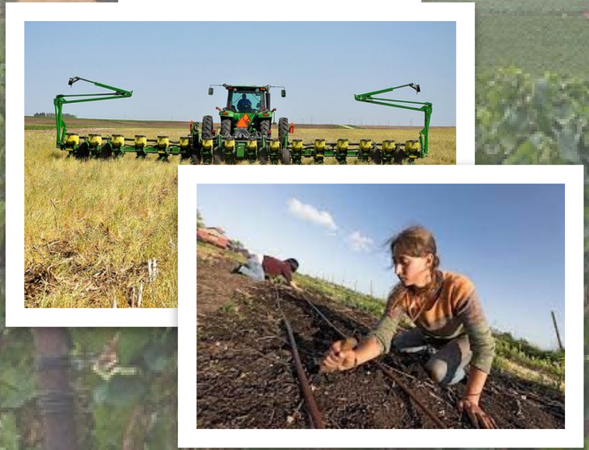


Environment

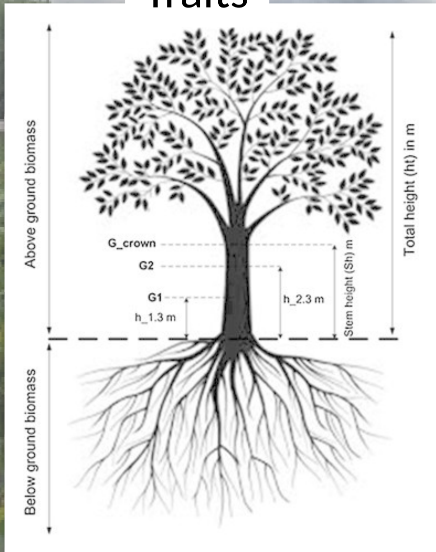
Genetics



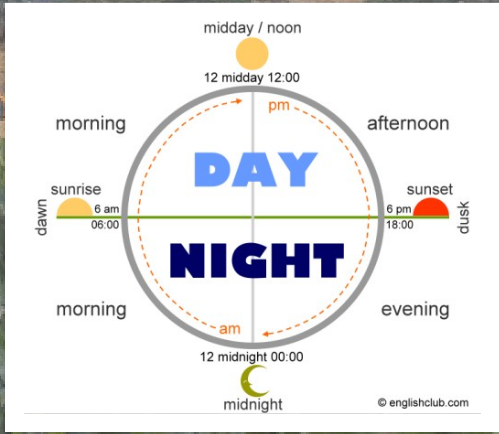
Management



Traits



Timing



slido

Please download and install the Slido app on all computers you use

What kind of data will your research focus on?

ⓘ Start presenting to display the poll results on this slide.

Source: Statista

THE SOLAR SYSTEM

WITH REAL-TIME SPEED OF LIGHT

— Distances to scale —

— All bodies x1000 —

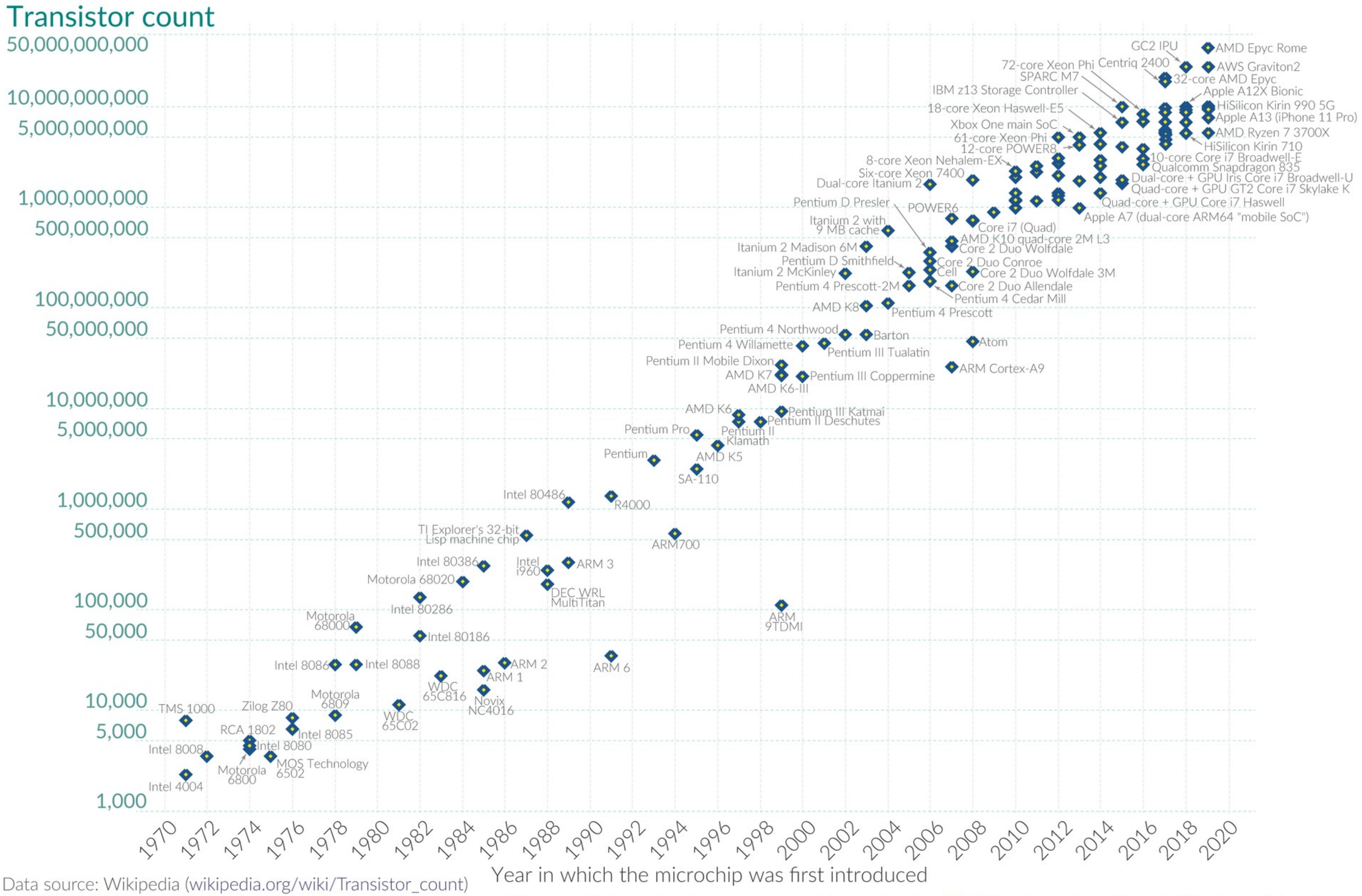
1 Astronomical Unit (1AU) = 149.6 Million km

James O'Donoghue / NASA imagery — T: @physicsJ

Moore's Law: The number of transistors on microchips doubles every two years

Our World in Data

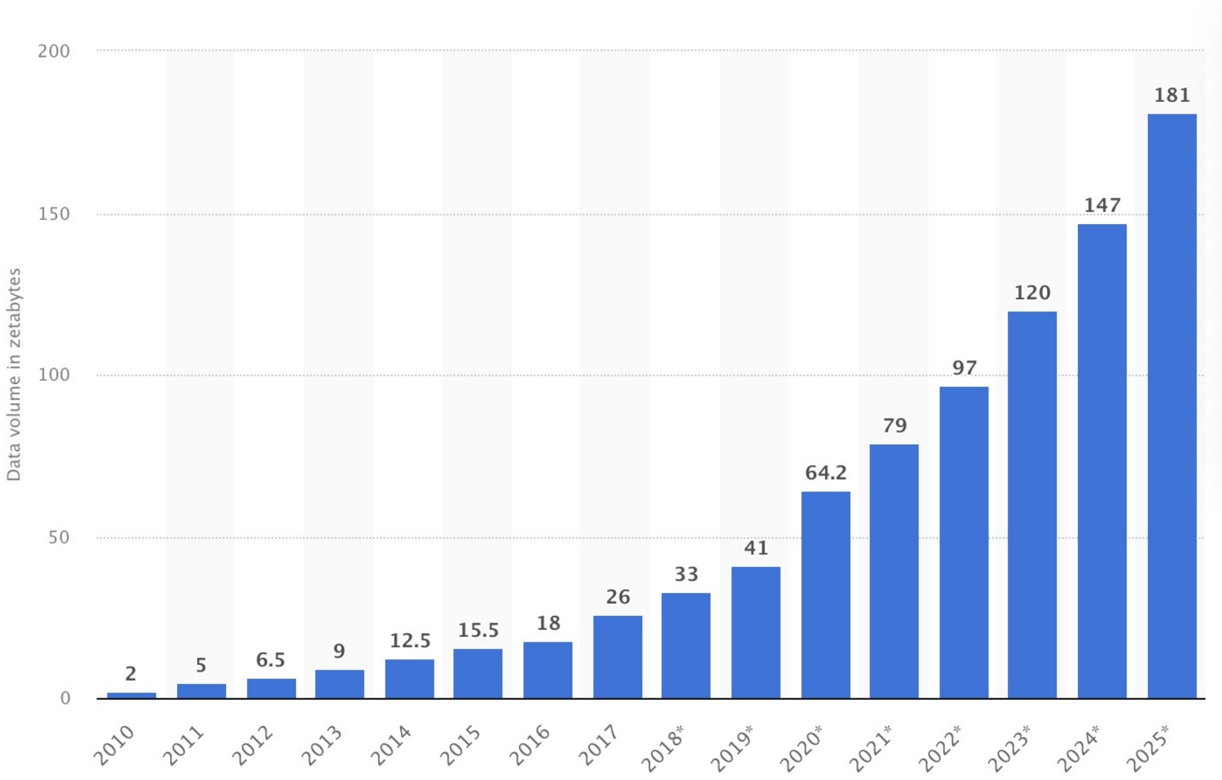
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

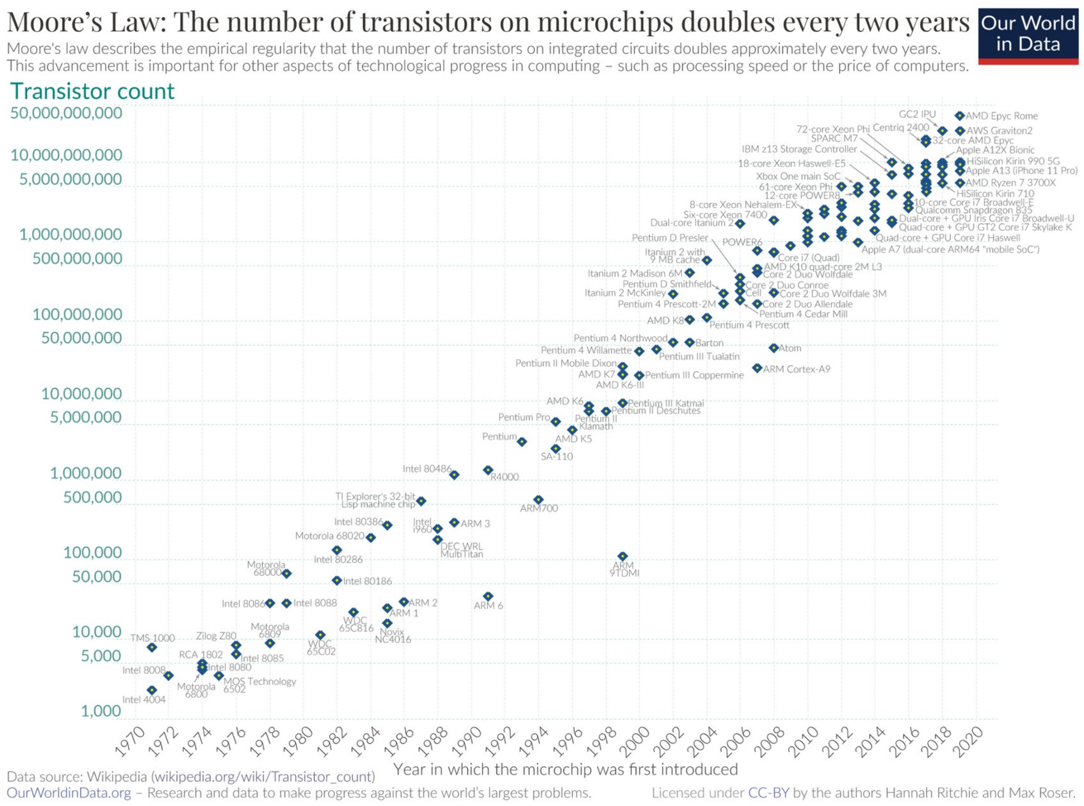


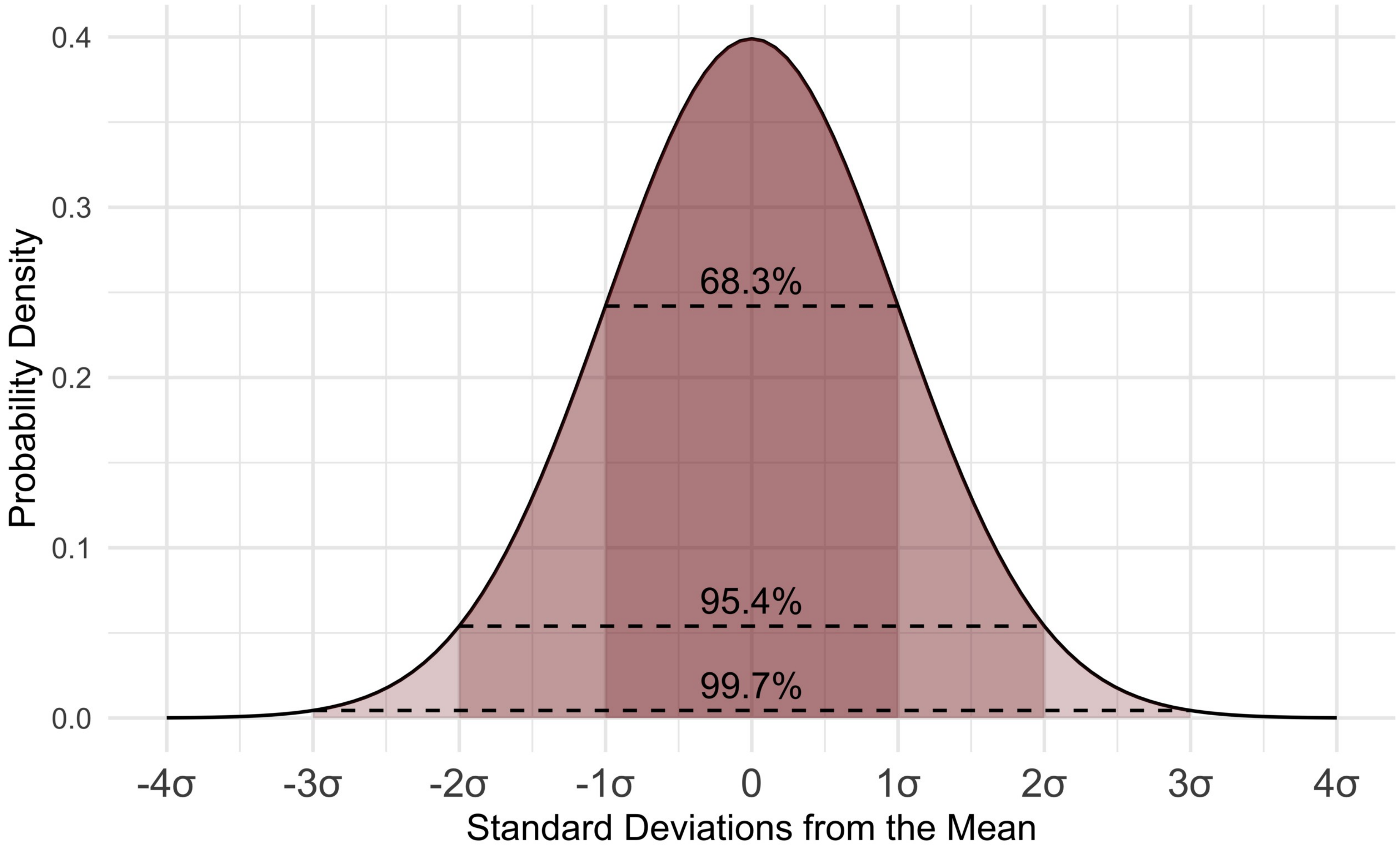
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

How do we use data and computational power as scientists?

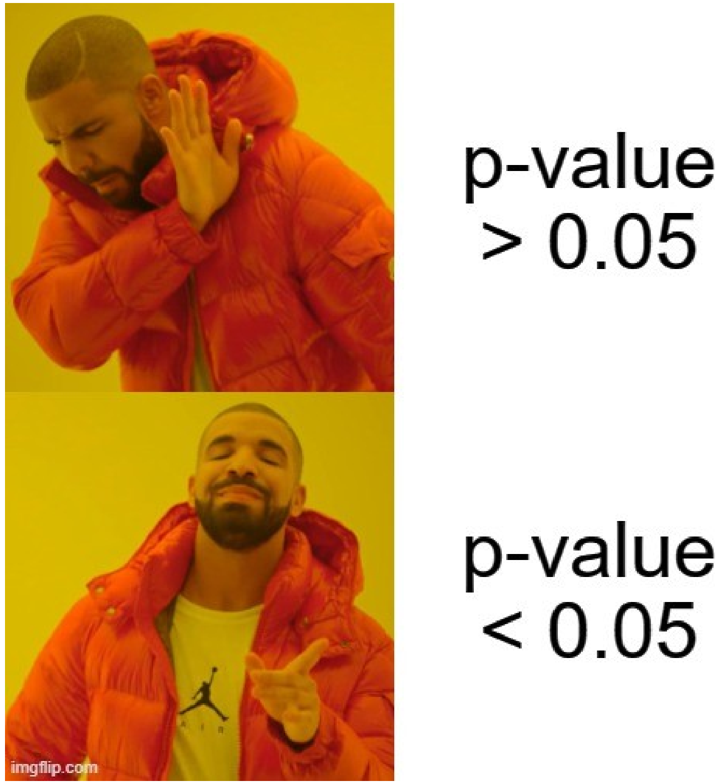


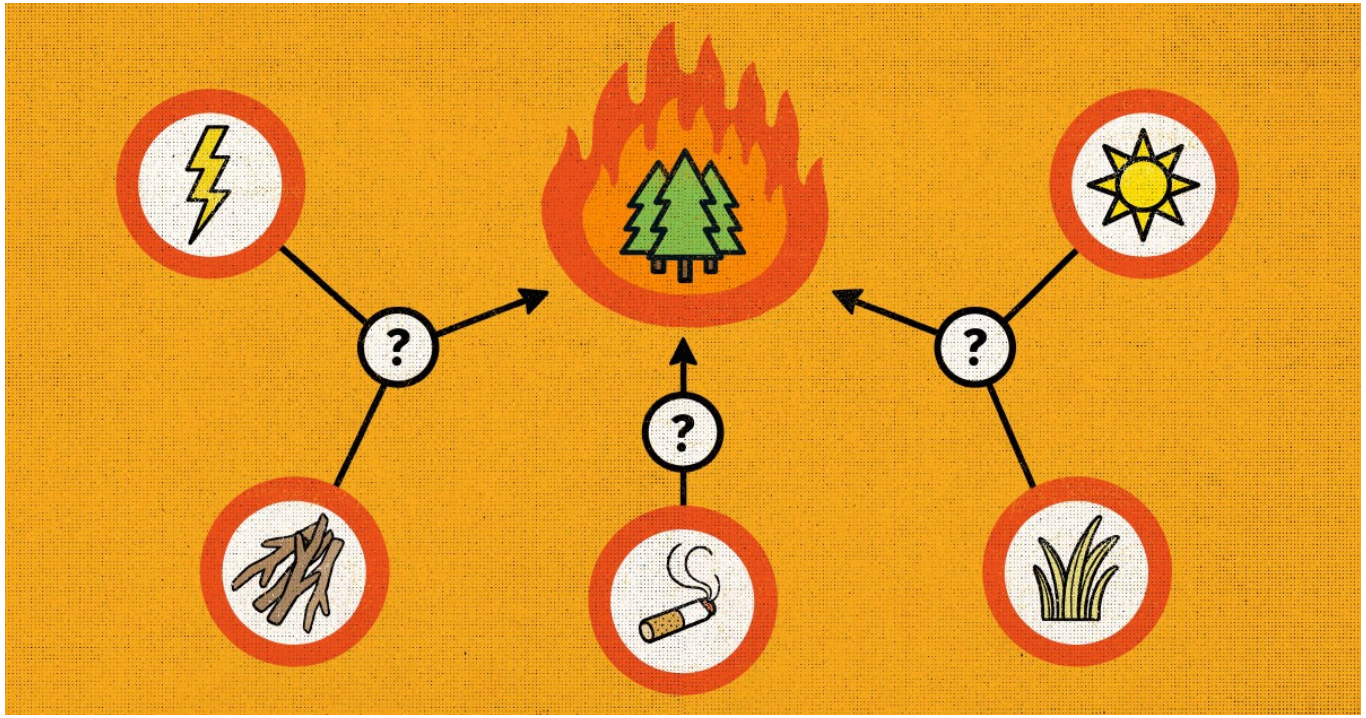




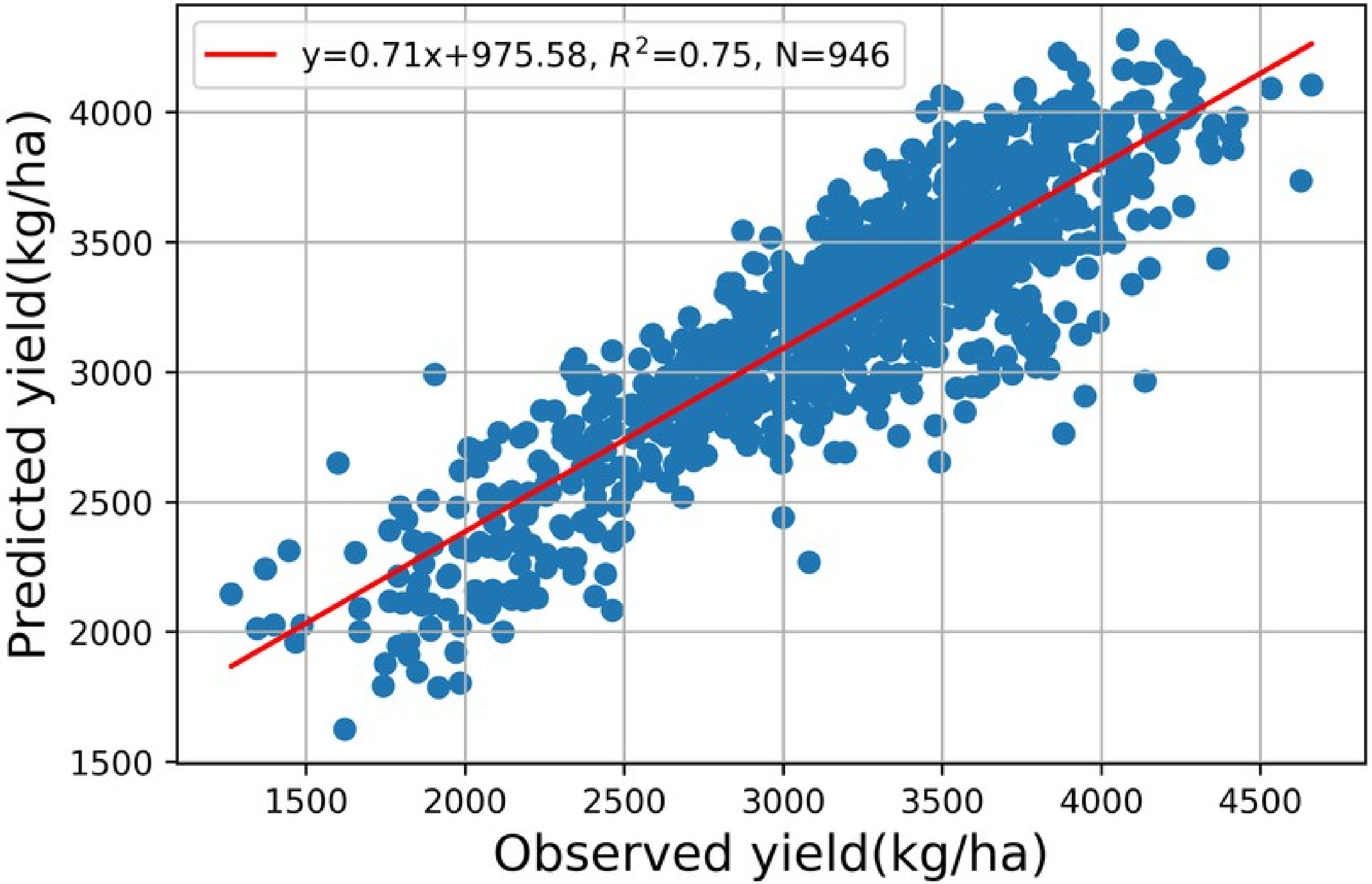
0.4
0.3
0.2
0.1
0.0
Probability Density
68.3%
95.4%
99.7%
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
Standard Deviations from the Mean

Hypothesis testing: making inferences about causality



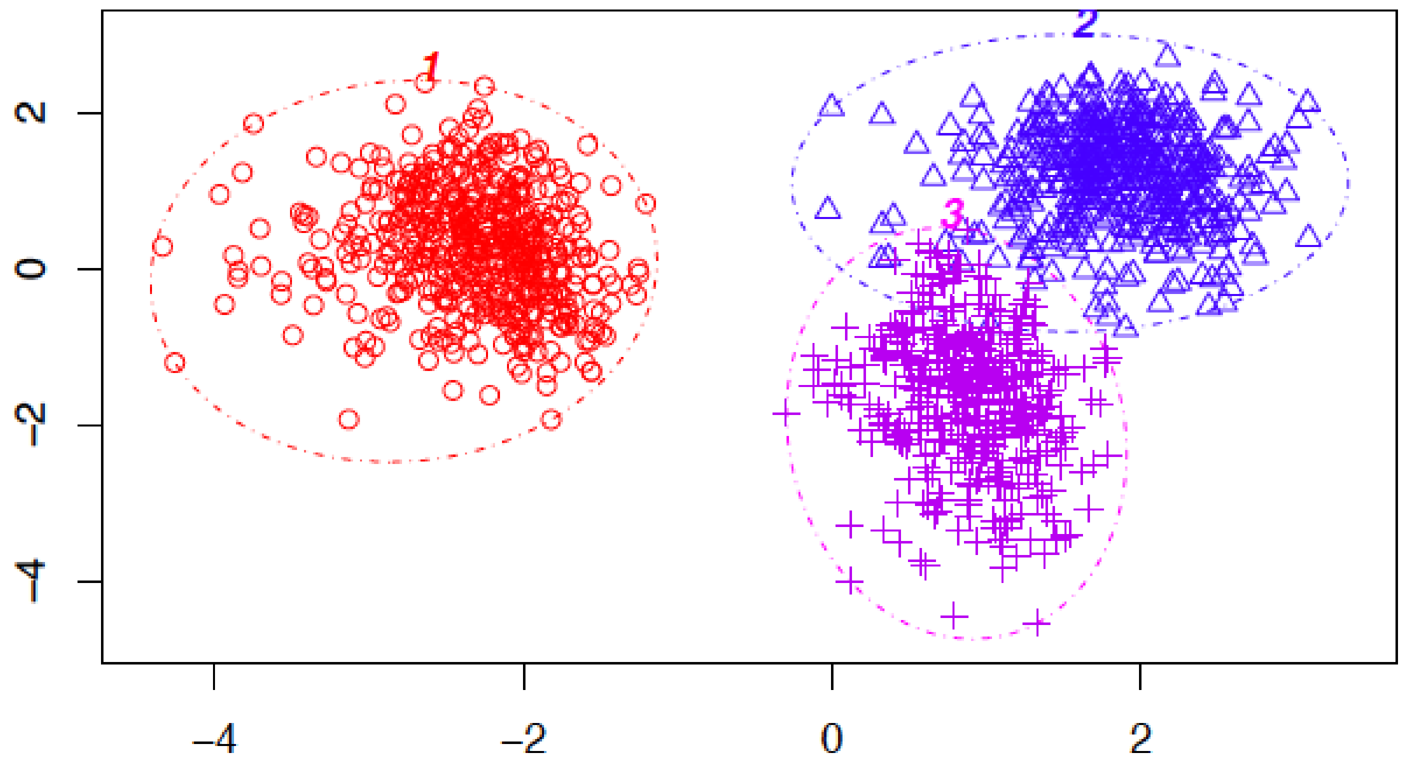


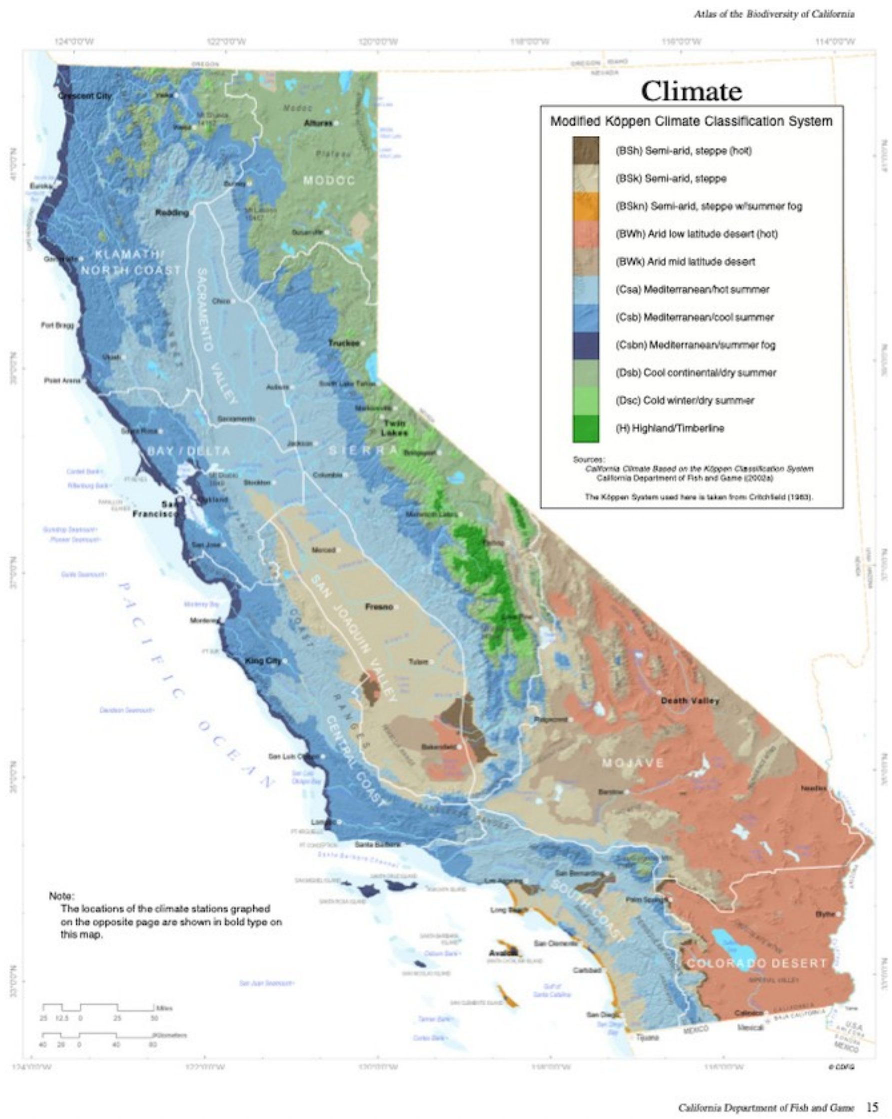
Prediction: making better decisions

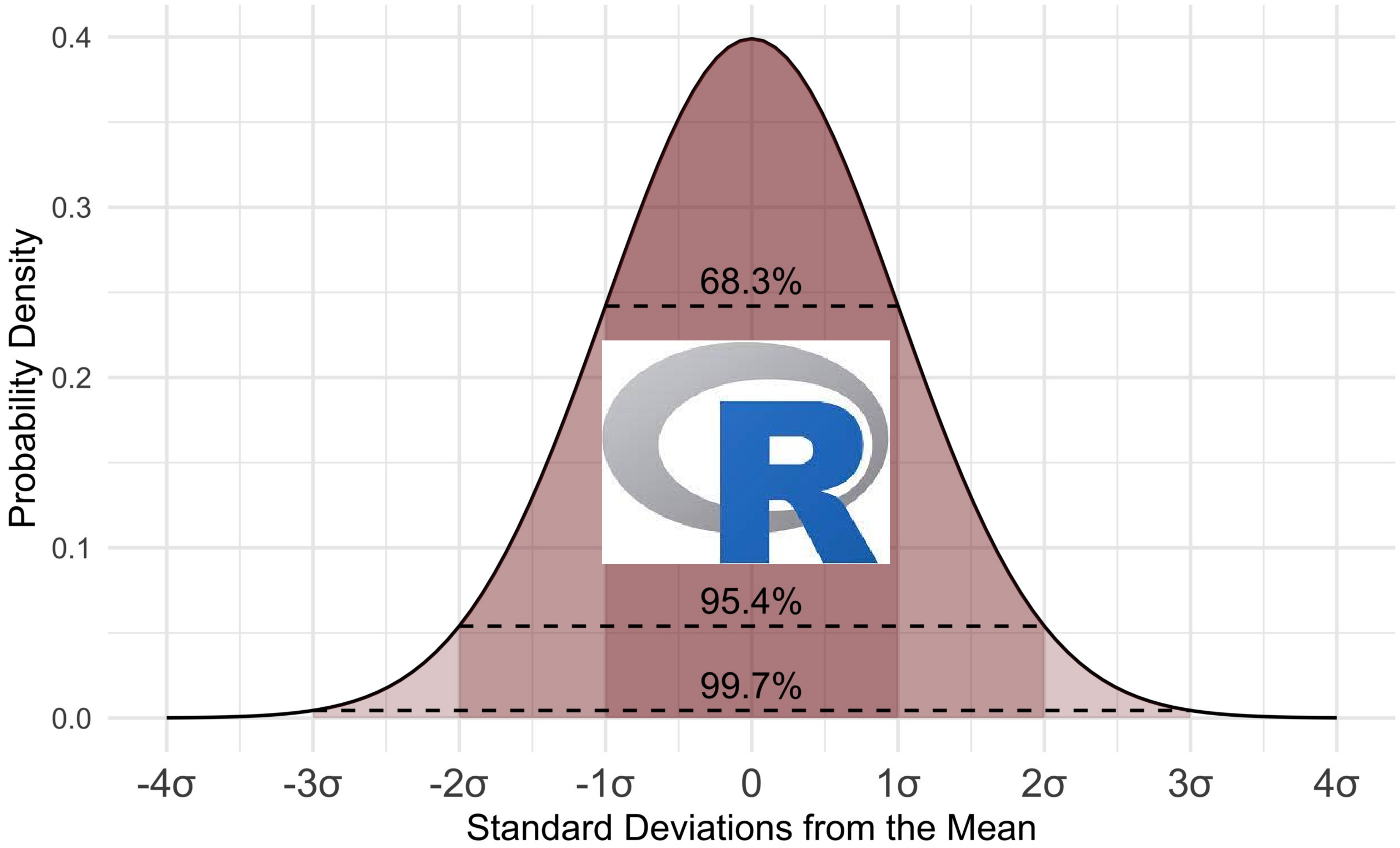


Discovering patterns: gaining new insights

Example: clustering







0.4
0.3
0.2
0.1
0.0
Probability Density
68.3%
95.4%
99.7%
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
Standard Deviations from the Mean

slido

Please download and install the Slido app on all computers you use

How would you rate your experience with R programming?

ⓘ Start presenting to display the poll results on this slide.

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Goal of PLS 206:

Leave with a fundamental understanding of the statistics and the <u>tools</u> to study multivariate data

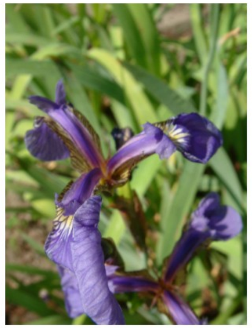
Basic understanding of statistical nomenclature

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n} \end{pmatrix}$$

Data frame!

Bird_ID	Sex	Lifestage	Wingchord	Mass	Bill	Tarsus
A2409	F	Fledgling	145.3	52.5	6.2	15.6
C1168	M	Hatchling	124.9	48.0	4.7	12.9
A0883	M	Adult	162.0	72.0	8.5	19.1
B9715	F	Adult	167.1	76.5	8.6	19.5
C5701	F	Hatchling	119.8	44.0	5.0	12.3

Ronald Fisher



Iris Setosa

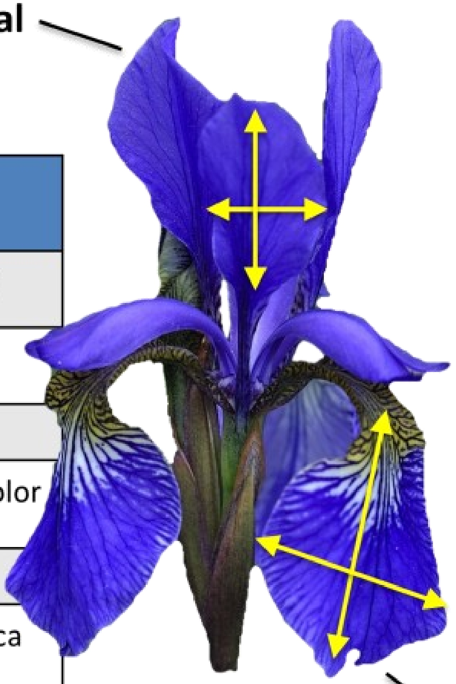
Iris Versicolor

Iris Virginica

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Variables
(attributes, measurements)

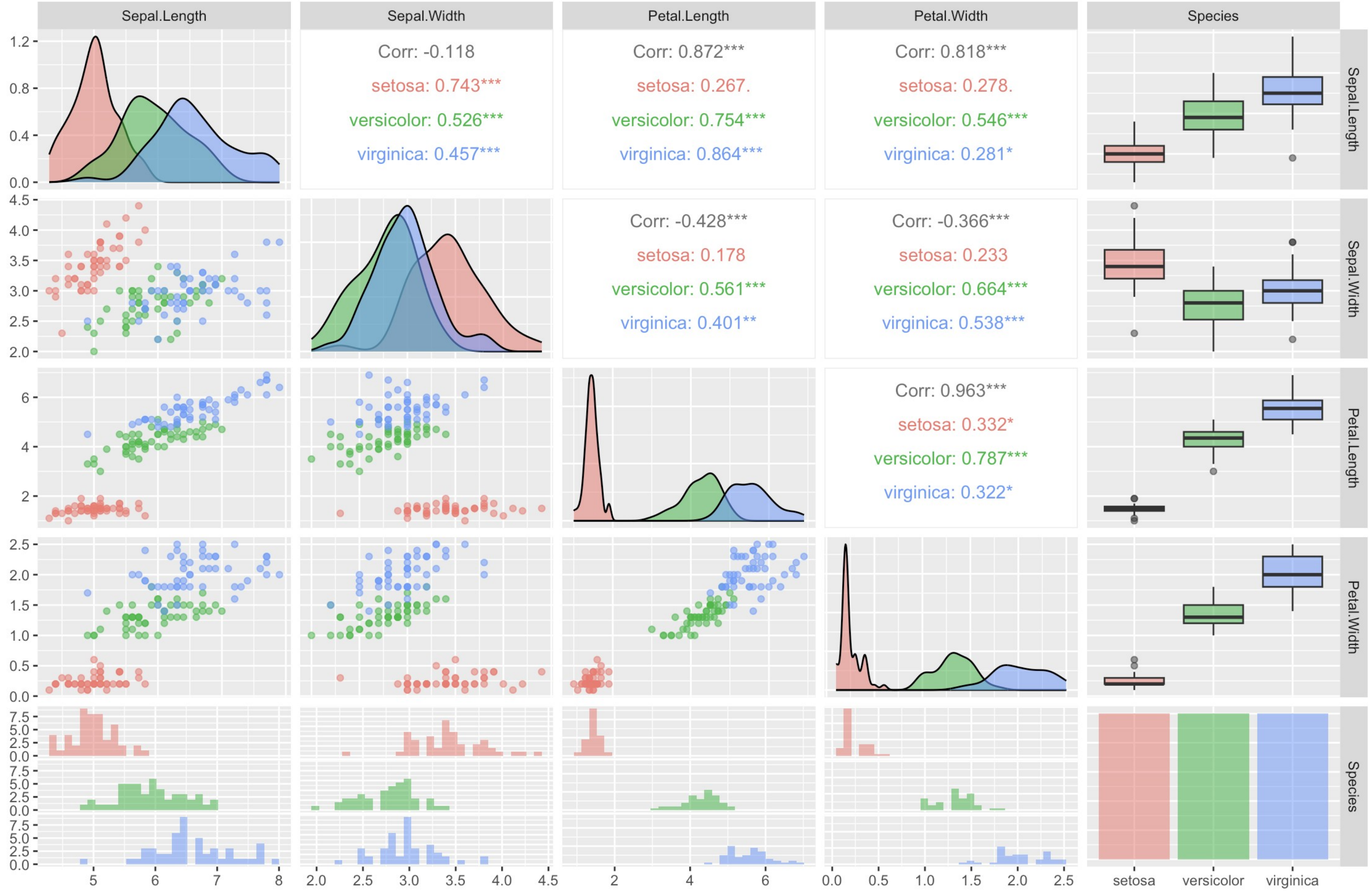


“Iris dataset”

```
iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa

```r
install.packages("GGally")
library(GGally)
ggpairs(iris, aes(col=Species,  alpha=0.5))
```



Interpreting model results in R

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.1379	0.1418	78.559	< 2e-16	***
Nlow	-1.0451	0.2005	-5.213	7.84e-06	***
Waterdrought	-2.9932	0.2005	-14.929	< 2e-16	***
Nlow:Waterdrought	0.9270	0.2836	3.269	0.00238	**

The mean yield under of high N, control

The difference between low N and high N under control

The difference between control and drought under high N

The difference in the effect of N between control and drought

Crash course on R programming and essential stats

Base R
Cheat Sheet

Getting Help

Accessing the help files

`?mean`
Get help of a particular function.

`help.search('weighted mean')`
Search the help files for a word or phrase.

`help(package = 'dplyr')`
Find help for a package.

More about an object

`str(iris)`
Get a summary of an object's structure.

`class(iris)`
Find the class an object belongs to.

Using Packages

`install.packages('dplyr')`
Download and install a package from CRAN.

`library(dplyr)`
Load the package into the session, making all its functions available to use.

`dplyr::select`
Use a particular function from a package.

`data(iris)`
Load a built-in dataset into the environment.

Working Directory

`getwd()`
Find the current working directory (where inputs are found and outputs are sent).

`setwd('C://file/path')`
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors

Creating Vectors

`c(2, 4, 6)`	2 4 6	Join elements into a vector
`2:6`	2 3 4 5 6	An integer sequence
`seq(2, 3, by=0.5)`	2.0 2.5 3.0	A complex sequence
`rep(1:2, times=3)`	1 2 1 2 1 2	Repeat a vector
`rep(1:2, each=3)`	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions

`sort(x)`
Return x sorted.

`rev(x)`
Return x reversed.

`table(x)`
See counts of values.

`unique(x)`
See unique values.

Selecting Vector Elements

By Position

`x[4]`	The fourth element.
`x[-4]`	All but the fourth.
`x[2:4]`	Elements two to four.
`x[-(2:4)]`	All elements except two to four.
`x[c(1, 5)]`	Elements one and five.

By Value

`x[x == 10]`	Elements which are equal to 10.
`x[x < 0]`	All elements less than zero.
`x[x %in% c(1, 2, 5)]`	Elements in the set 1, 2, 5.

Named Vectors

`x['apple']`	Element with name 'apple'.

Programming

For Loop

```
for (variable in sequence){
  Do something
}
```

Example

```
for (i in 1:4){
  j <- i + 10
  print(j)
}
```

While Loop

```
while (condition){
  Do something
}
```

Example

```
while (i < 5){
  print(i)
  i <- i + 1
}
```

If Statements

```
if (condition){
  Do something
} else {
  Do something different
}
```

Example

```
if (i > 3){
  print('Yes')
} else {
  print('No')
}
```

Functions

```
function_name <- function(var){
  Do something
  return(new_variable)
}
```

Example

```
square <- function(x){
  squared <- x*x
  return(squared)
}
```

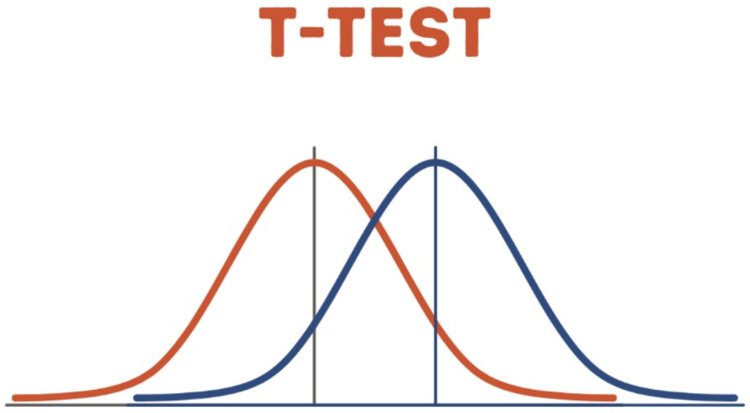
Reading and Writing Data

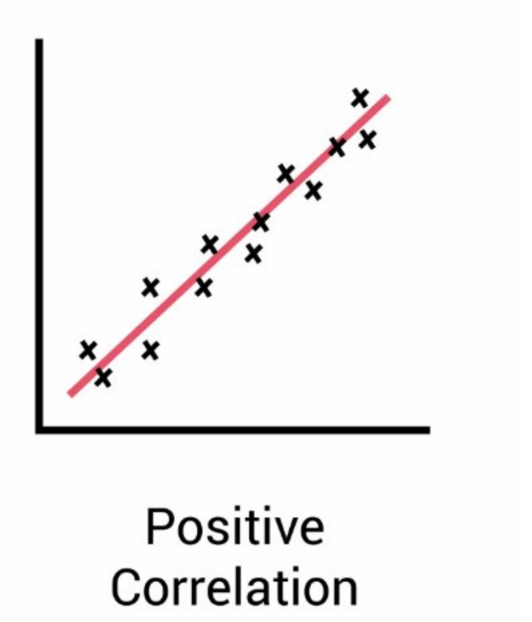
Also see the **readr** package.

Input	Ouput	Description
`df <- read.table('file.txt')`	`write.table(df, 'file.txt')`	Read and write a delimited text file.
`df <- read.csv('file.csv')`	`write.csv(df, 'file.csv')`	Read and write a comma separated value file. This is a special case of read.table/write.table.
`load('file.RData')`	`save(df, file = 'file.Rdata')`	Read and write an R data file, a file type special for R.

Conditions

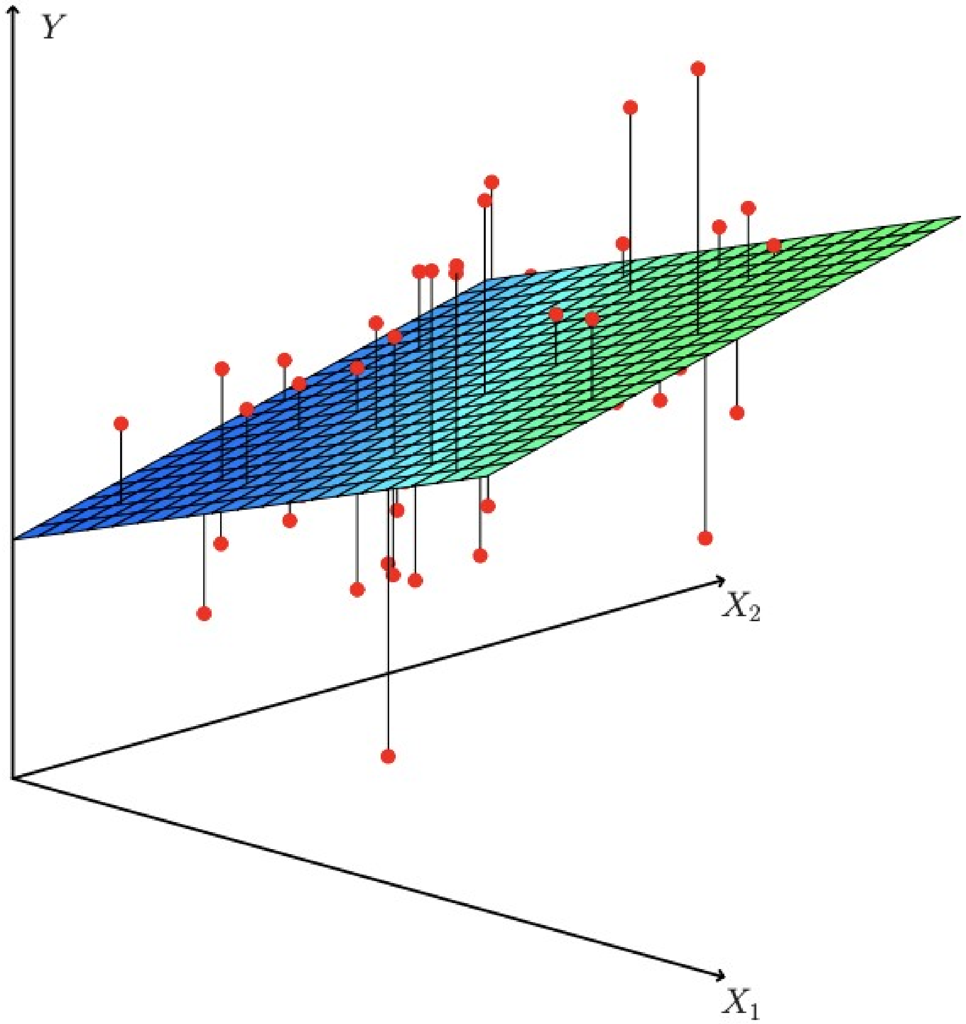
`a == b`	Are equal	`a > b`	Greater than	`a >= b`	Greater than or equal to	`is.na(a)`	Is missing
`a != b`	Not equal	`a < b`	Less than	`a <= b`	Less than or equal to	`is.null(a)`	Is null





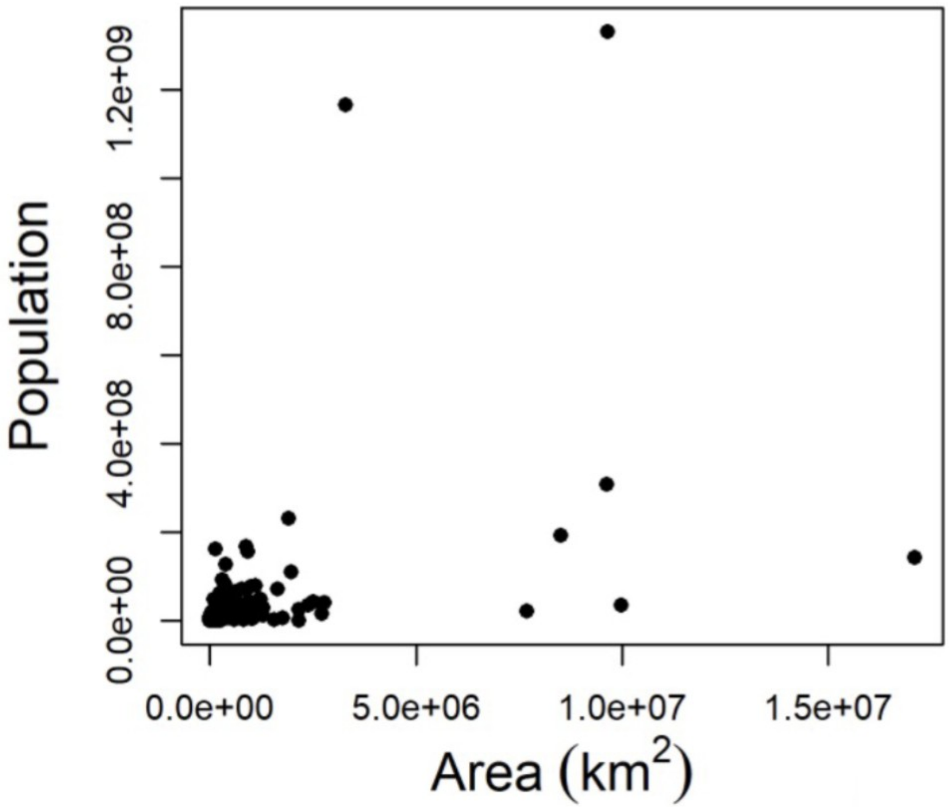
Linear regression and its many derivatives/relatives

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

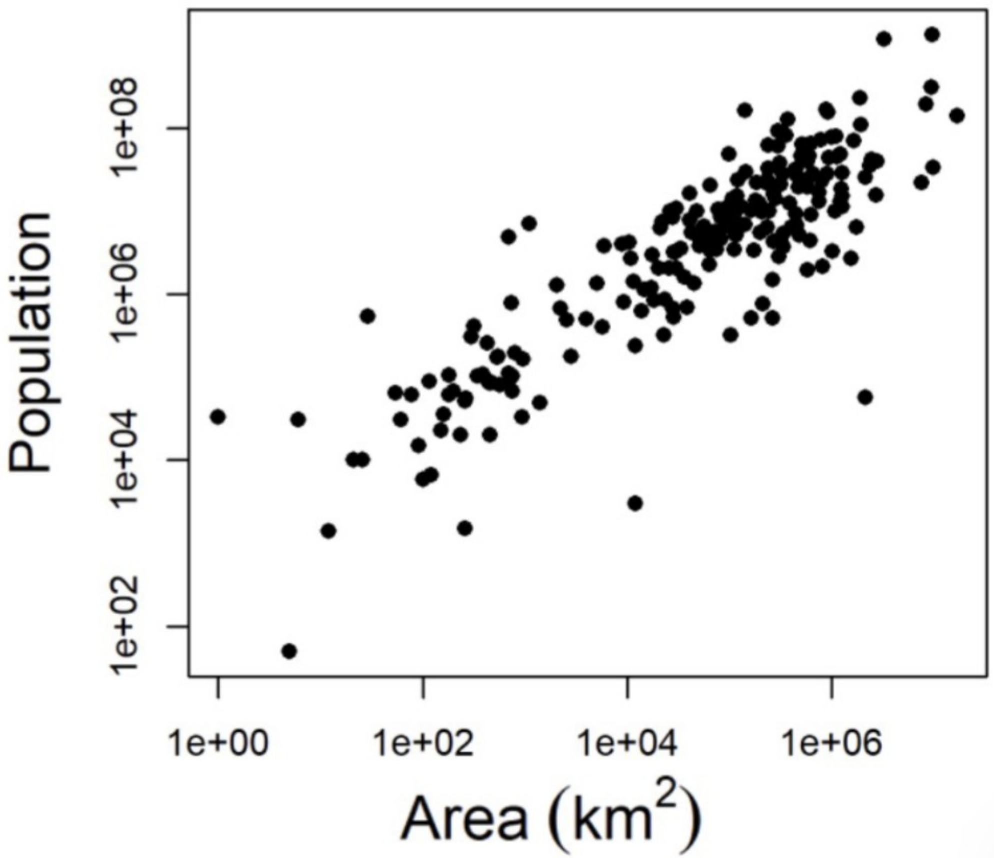


Data transformations

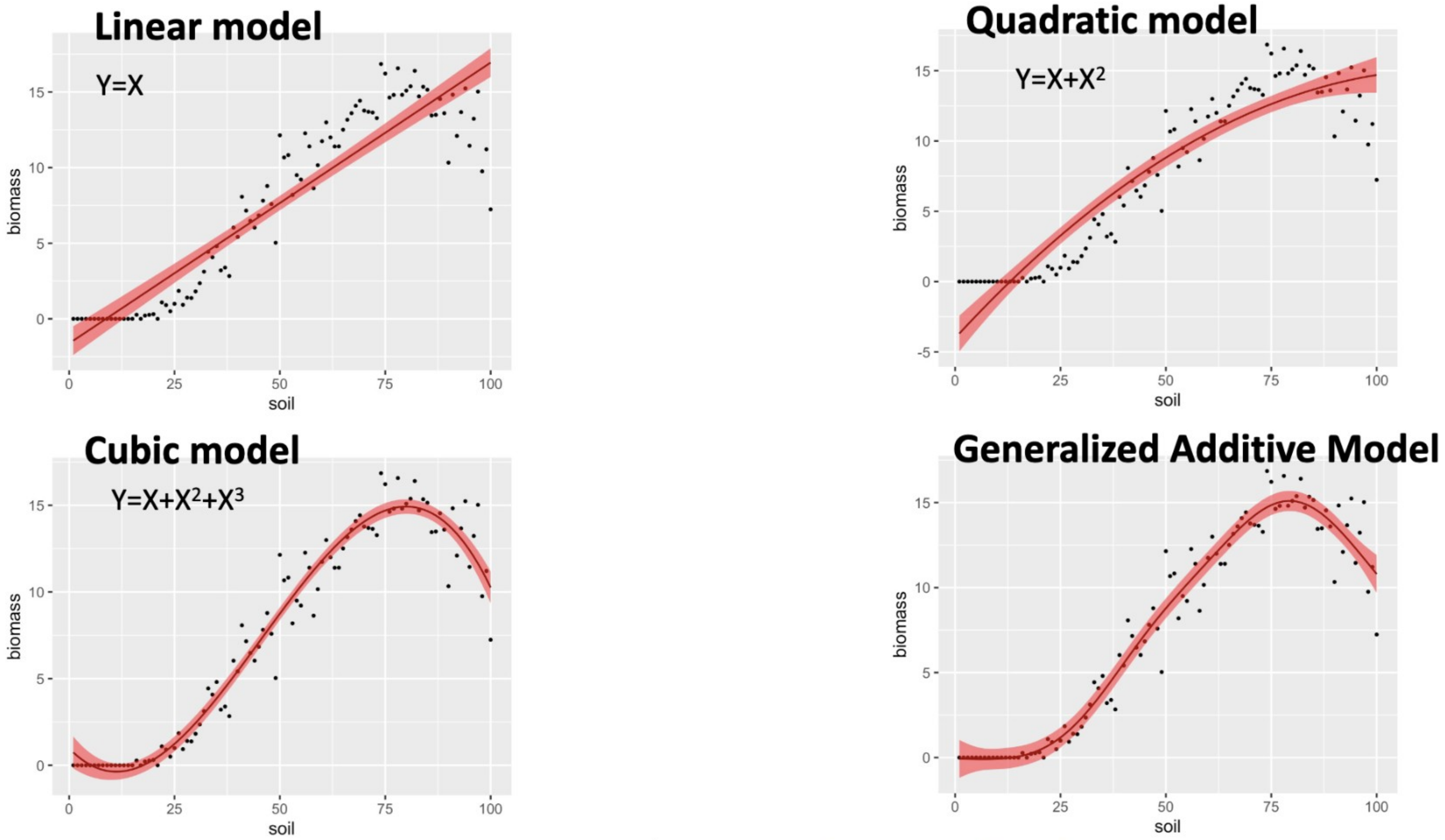
Raw data



Log transformed



Non-linearity




```
model<-gam(biomass~s(soil), data=species1, method="REML")
```

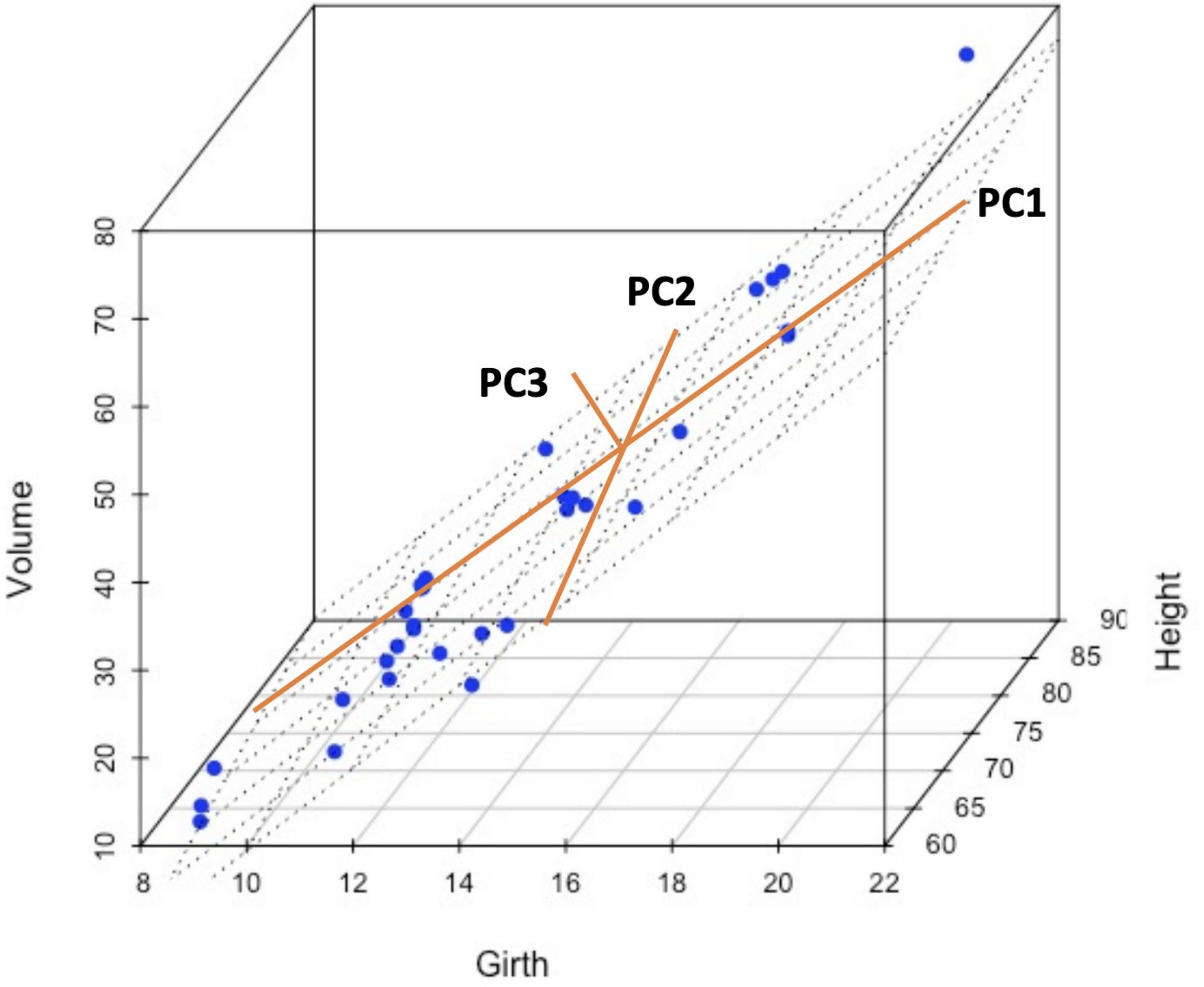
Model selection

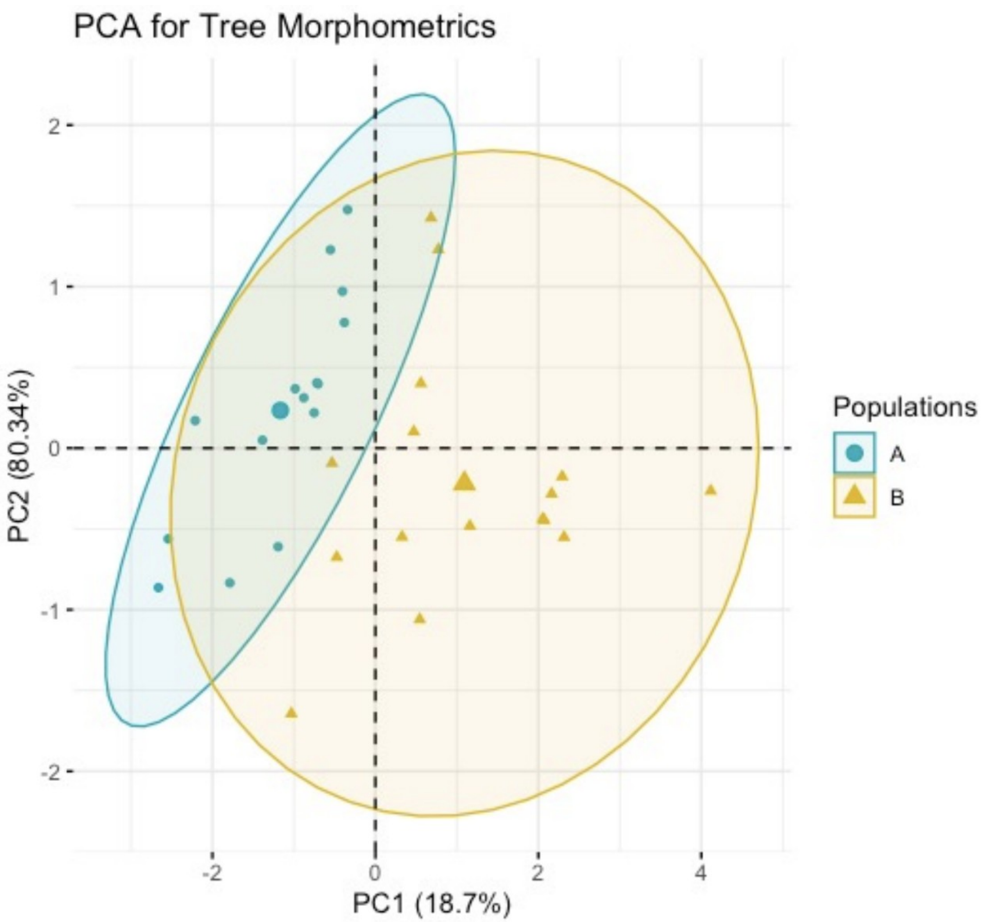
Underfit model
Y~X

Well-fit model
Y~X+Z

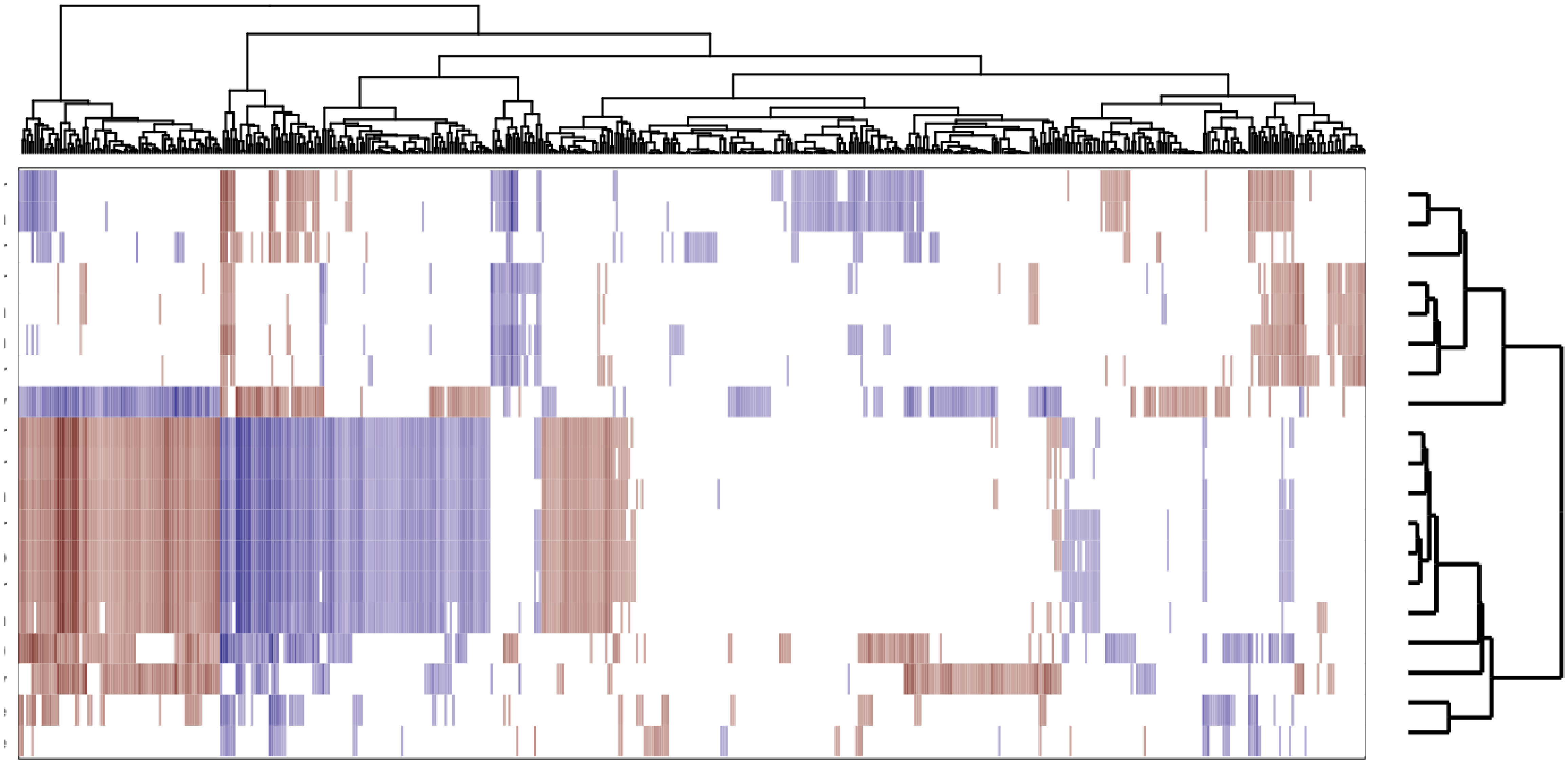
Overfit-fit model
$Y \sim X+Z+X^2+Z^2+A+B+C$

Dimension reduction (e.g., Principal component analysis)

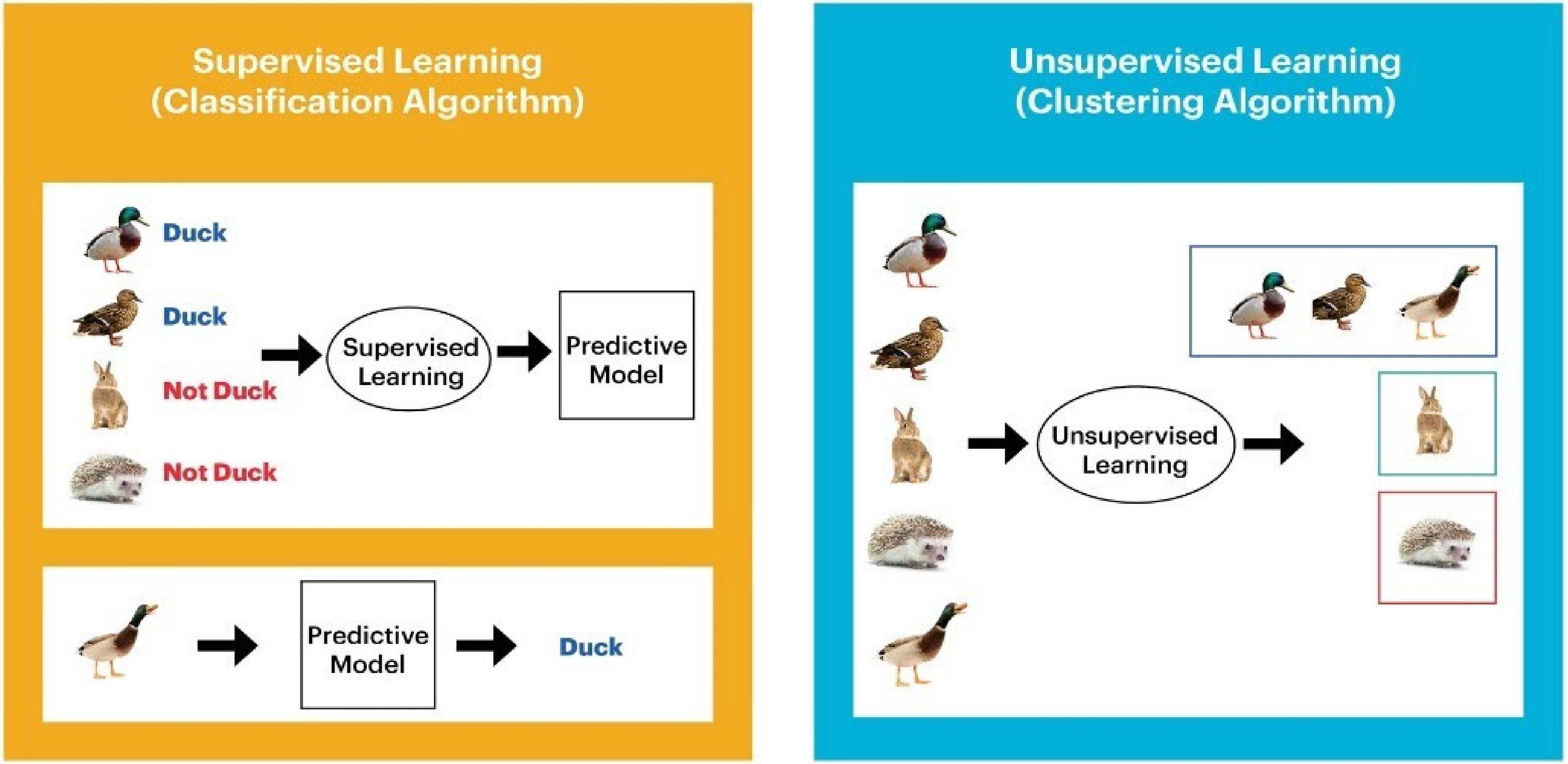




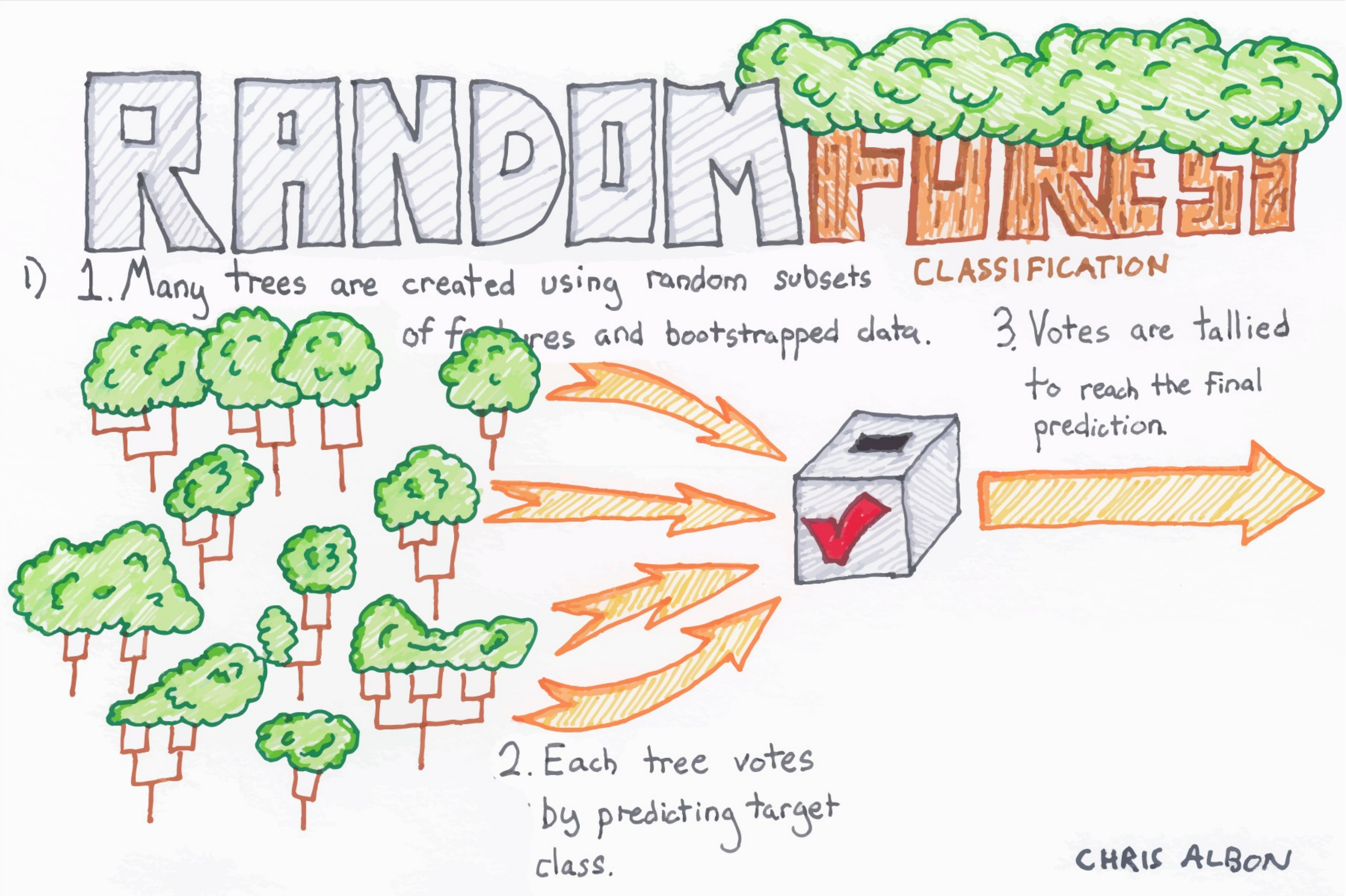
Clustering methods



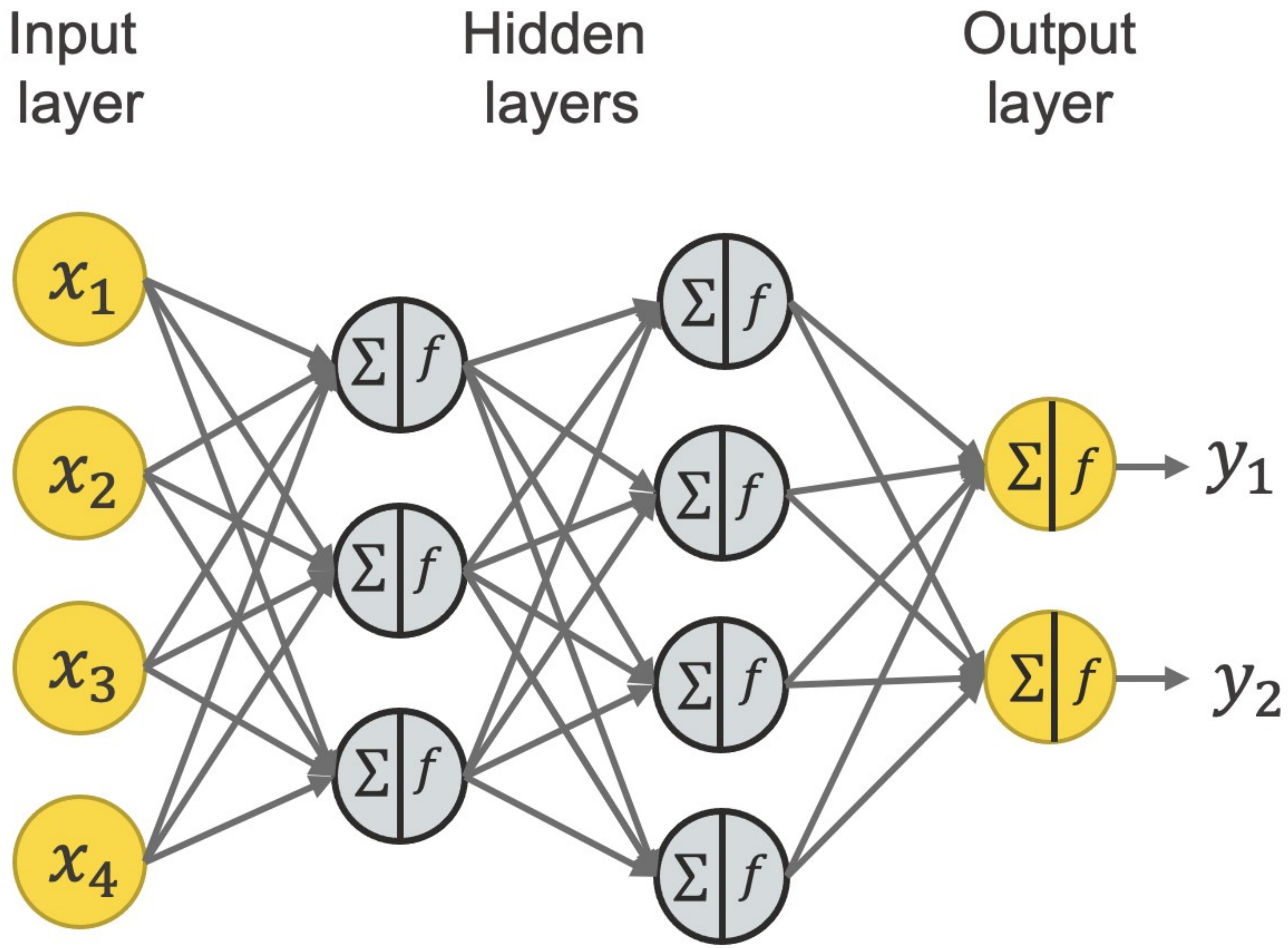
Supervised vs unsupervised learning



Intro to tree-based analyses

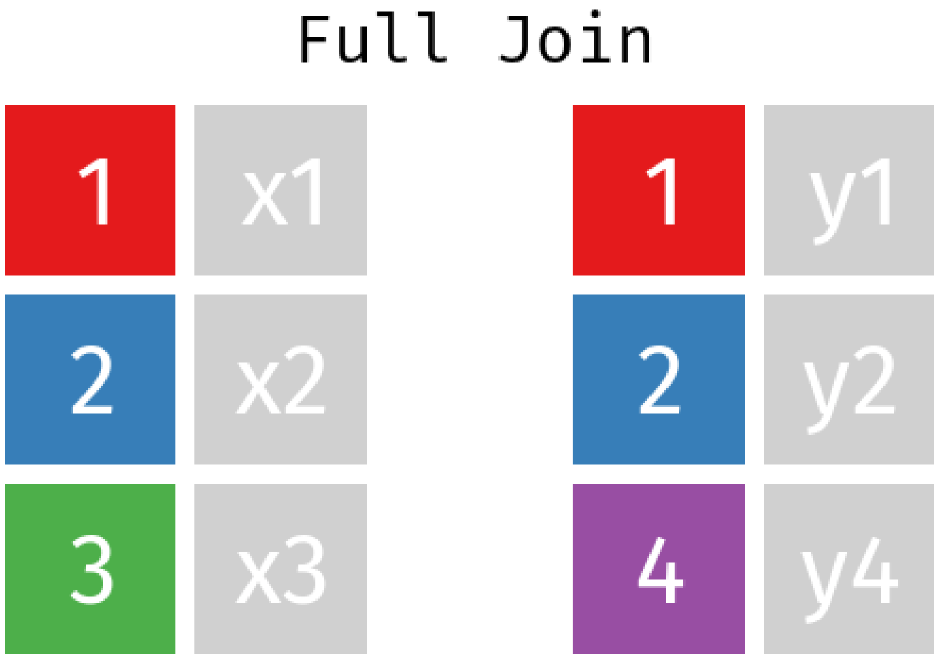


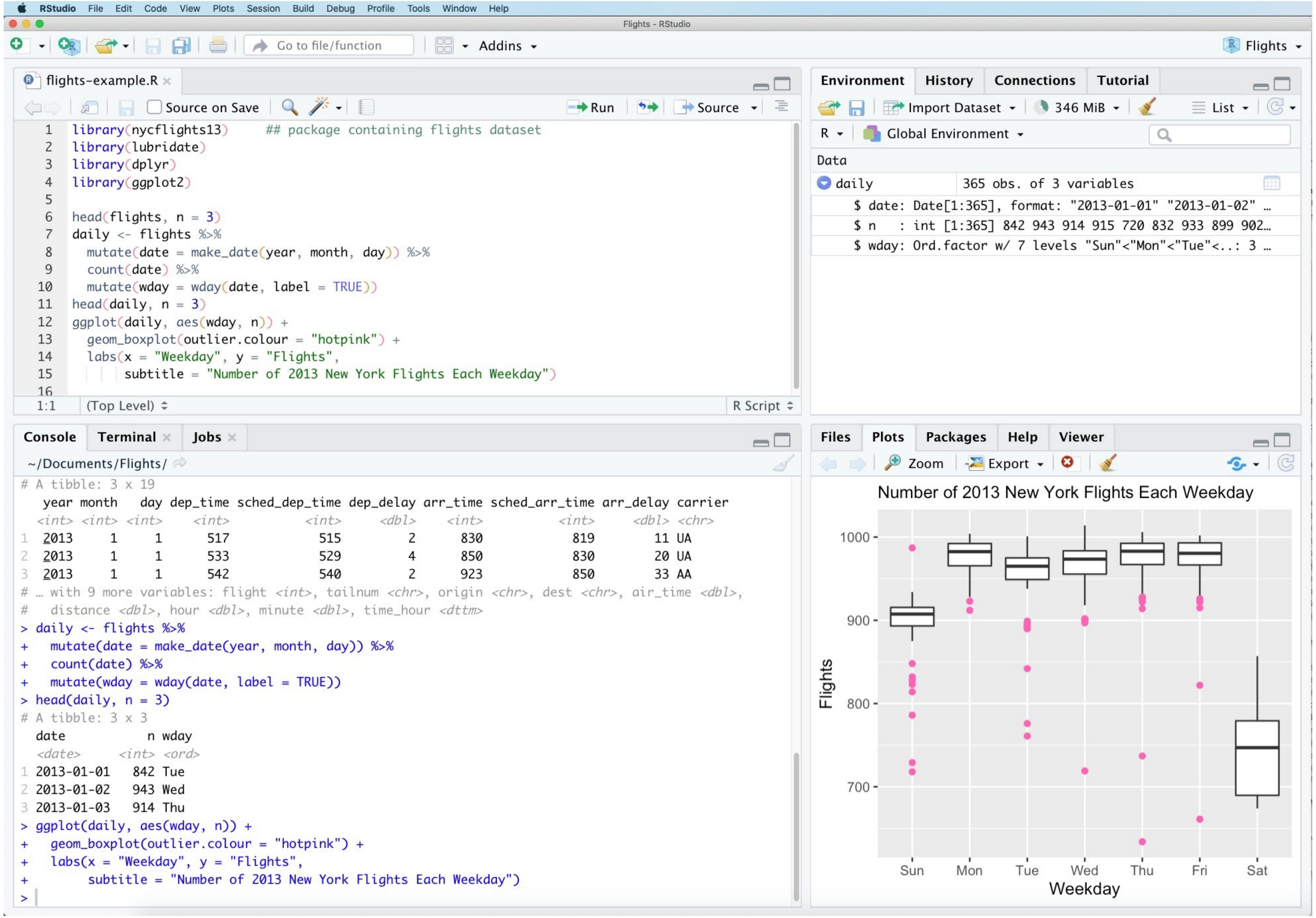
Intro to neural networks



Mastering R

- Data wrangling and packages to help
- Writing your own functions
- Using loops and other fundamental programming skills
- Creating publishable figures




RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Flights - RStudio
Go to file/function
Addins
Flights
flights-example.R
Source on Save
Run
Source
library(nycflights13)   ## package containing flights dataset
library(lubridate)
library(dplyr)
library(ggplot2)
head(flights, n = 3)
daily <- flights %>%
mutate(date = make_date(year, month, day)) %>%
count(date) %>%
mutate(wday = wday(date, label = TRUE))
head(daily, n = 3)
ggplot(daily, aes(wday, n)) +
geom_boxplot(outlier.colour = "hotpink") +
labs(x = "Weekday", y = "Flights",
subtitle = "Number of 2013 New York Flights Each Weekday")
1:1
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
346 MiB
List
R
Global Environment
Data
daily
365 obs. of 3 variables
$ date: Date[1:365], format: "2013-01-01" "2013-01-02" …
$ n : int [1:365] 842 943 914 915 720 832 933 899 902…
$ wday: Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 3 …
Console Terminal Jobs
~/Documents/Flights/
A tibble: 3 x 19
year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
<int> <int> <int> <int> <int> <dbl> <int> <int> <dbl> <chr>
1 2013 1 1 517 515 2 830 819 11 UA
2 2013 1 1 533 529 4 850 830 20 UA
3 2013 1 1 542 540 2 923 850 33 AA
… with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
> daily <- flights %>%
+ mutate(date = make_date(year, month, day)) %>%
+ count(date) %>%
+ mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
A tibble: 3 x 3
date n wday
<date> <int> <ord>
1 2013-01-01 842 Tue
2 2013-01-02 943 Wed
3 2013-01-03 914 Thu
> ggplot(daily, aes(wday, n)) +
+ geom_boxplot(outlier.colour = "hotpink") +
+ labs(x = "Weekday", y = "Flights",
+ subtitle = "Number of 2013 New York Flights Each Weekday")
Files Plots Packages Help Viewer
Zoom
Export
Number of 2013 New York Flights Each Weekday
Flights
1000
900
800
700
Sun Mon Tue Wed Thu Fri Sat
Weekday


I AM SMILING BECAUSE
I AM USING R
imgflip.com

Data wrangling

Parsing R objects

Making figures

Programming skills

Statistical analyses

Linear regression

Dimension reduction

Machine learning

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Course Schedule: Overview

10 weeks — Monday/Wednesday lecture + Friday discussion, practice, bonus topics

Stats/R refresher → regression → model selection → PCA/clustering → machine learning → multivariate methods

Weekly problem sets due Mondays; exact dates & topics kept current on the website:

https://greymonroe.github.io/PLS_206/schedule.html

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Details about PLS 206:

Monday/Wednesday Lecture + R live coding demonstration

Friday discussion –Questions on Problem sets and Bonus material

This week (week 0): Friday discussion is R crash course, part 1

Zoom option and all recorded

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Grade:

- Weekly problem sets graded for **completion due Monday before class (none for next Monday)**

2 drops for problem set grade

- Final week will be an individual project (cant be dropped)
- Daily reflection/feedback (Mon/Wed) due before next class (10% of grade)

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Use of AI and Outside Resources

- AI tools, Google, Stack Overflow, classmates — always fair game; AI is just another resource
- Problem sets are graded on reasoning, not a single right answer: use AI to understand why, not just to get unstuck
- Submit your own understanding — write your own answers, and cite AI use the way you'd cite a classmate or a Stack Overflow post
- Final paper: every data point, statistic, and citation must be real — never submit fabricated or AI-hallucinated results

Before Friday (intro to R):

1. Download and Install R and R studio
https://posit.co/download/rstudio-desktop/

This Friday

R crash course, part 1 (stats refresher continues next week)