Webinar 0

Terms, Concepts, Data Formats – A Tutorial for Background

Goals

Introduce the technical terminology that will be used throughout the webinars

Learning Objectives

- ✓ Understand the meaning of the technical terms used in the webinars

- ✓ **Participants** should be able to recognize the terms during the webinar series and understand their meaning to facilitate comprehension of the workflow.

Webinar organization

1. **Introducing the technical terminology.**

2. **Understanding the meaning of each term.**

MACHINES

Introducing the technical terminology

BiotaPhy

Host machine

The local physical machine on which the user will run Docker.

Image from WikiCommons

Remember these?

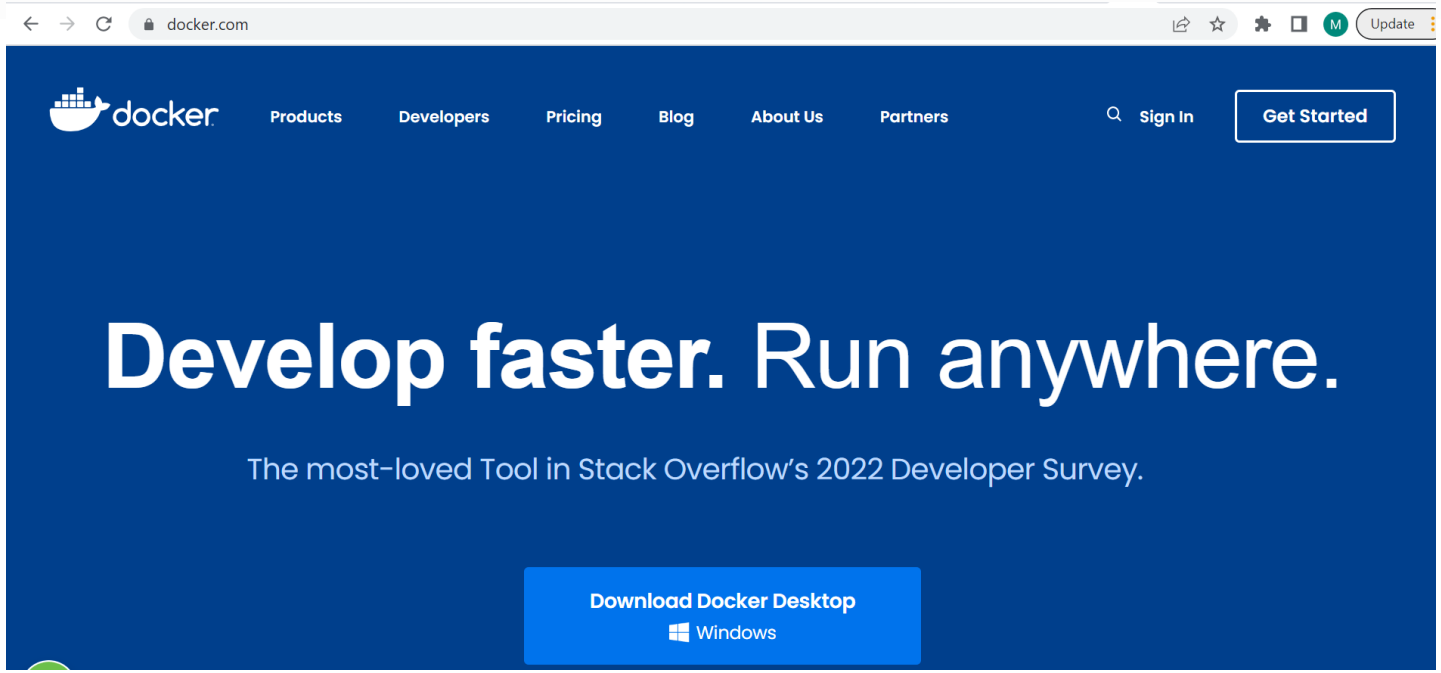
Introducing the technical terminology

BiotaPhy

Docker

Docker is an application which can run on Linux, MacOSX, or Windows. With a Docker-ized application, such as this tutorial, a user can run the application on their local machine in a controlled and sequestered environment, with a set of dependencies that may not be easy, allowed, or even available for their local machine.

docker.com



docker

Introducing the technical terminology

BiotaPhy

Docker image

A Docker-ized application, built into a single package, with all required software dependencies and files.

Image from WikiCommons

docker

Introducing the technical terminology

BiotaPhy

Container

A Docker instance which runs as an application on a Host machine. The Docker container contains all software dependencies required by the programs it will run.

docker

Images from WikiCommons

Introducing the technical terminology

Docker volume

Docker volumes are file systems mounted on a Docker Container to share data from the Host machine or preserve data generated by the running Container. The volumes are stored on the host, independent of the container life cycle allowing users to back up or share file systems between containers.

Volumes mounted in container

Images from WikiCommons

FILE TYPES

Introducing the technical terminology

CSV

CSV (Comma Separated Values) is a file format for records, in which fields are separated by a delimiter. Commas and tabs are common, but other characters may be used as delimiters.

This is what a .csv file looks like!

	species_name	x	y
2	Bensoniella oregona	-123.751	42.802
3	Bensoniella oregona	-123.7903	42.802
4	Bensoniella oregona	-123.7903	42.802
5	Bensoniella oregona	-123.7707	42.7873
6	Bensoniella oregona	-123.751	42.9927
7	Bensoniella oregona	-123.9646	42.7788
8	Bensoniella oregona	-123.7117	42.9047
9	Bensoniella oregona	-123.7117	42.9047
10	Bensoniella oregona	-123.8266667	42.625
11	Bensoniella oregona	-123.8038889	42.60861111
12	Bensoniella oregona	-123.9075	42.54222222

Introducing the technical terminology

DwCA

DwCA (Darwin Core Archive) is a packaged dataset of occurrence records in Darwin Core standard format, along with metadata about the contents.

According to Wikipedia:

Darwin Core Archive

From Wikipedia, the free encyclopedia

Darwin Core Archive (DwC-A) is a biodiversity informatics data standard that makes use of the Darwin Core terms to produce a single, self-contained dataset for species occurrence, checklist, sampling event or material sample data. Essentially it is a set of text (CSV) files with a simple descriptor (meta.xml) to inform others how your files are organized. The format is defined in the Darwin Core Text Guidelines.[1] It is the preferred format for publishing data to the GBIF network.

Introducing the technical terminology

JSON

JSON (pronounced “jason”) is a structured file format containing groups of keys with values, all enclosed in curly braces ({}). Values may be basic data types like strings (enclosed in double-quotes, “”), numbers (not quoted), booleans (true or false, in lowercase and not quoted) or other literals. Values may also be arrays (comma-delimited lists of basic data types, enclosed in square brackets, []), or another (nested) group of keys with values. More information about JSON is here and you can check your format with an online validator.

10 lines (10 sloc) | 211 Bytes

```
1  [
2      {
3          "wrangler_type": "MatchTreeMatrixWrangler",
4          "tree": "/volumes/data/input/heuchera.nex",
5          "species_axis": 1
6      },
7      {
8          "wrangler_type": "PurgeEmptySlicesWrangler"
9      }
10 ]
```

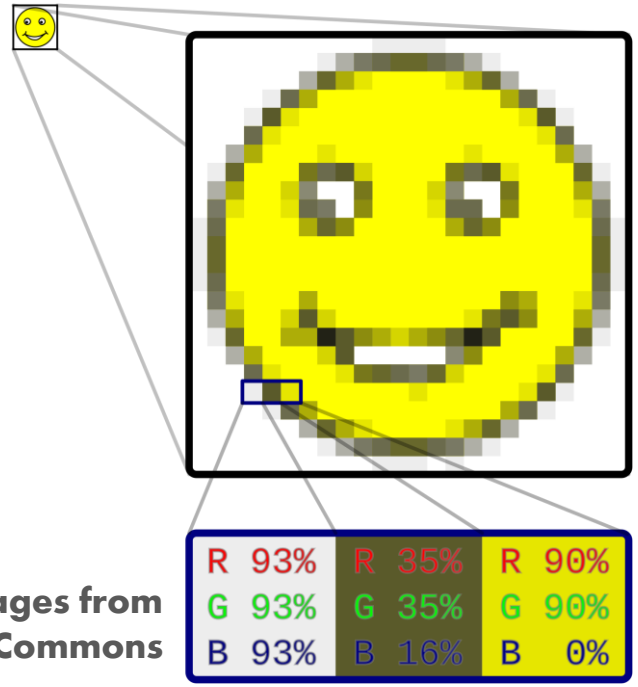
DATA TYPES

Introducing the technical terminology

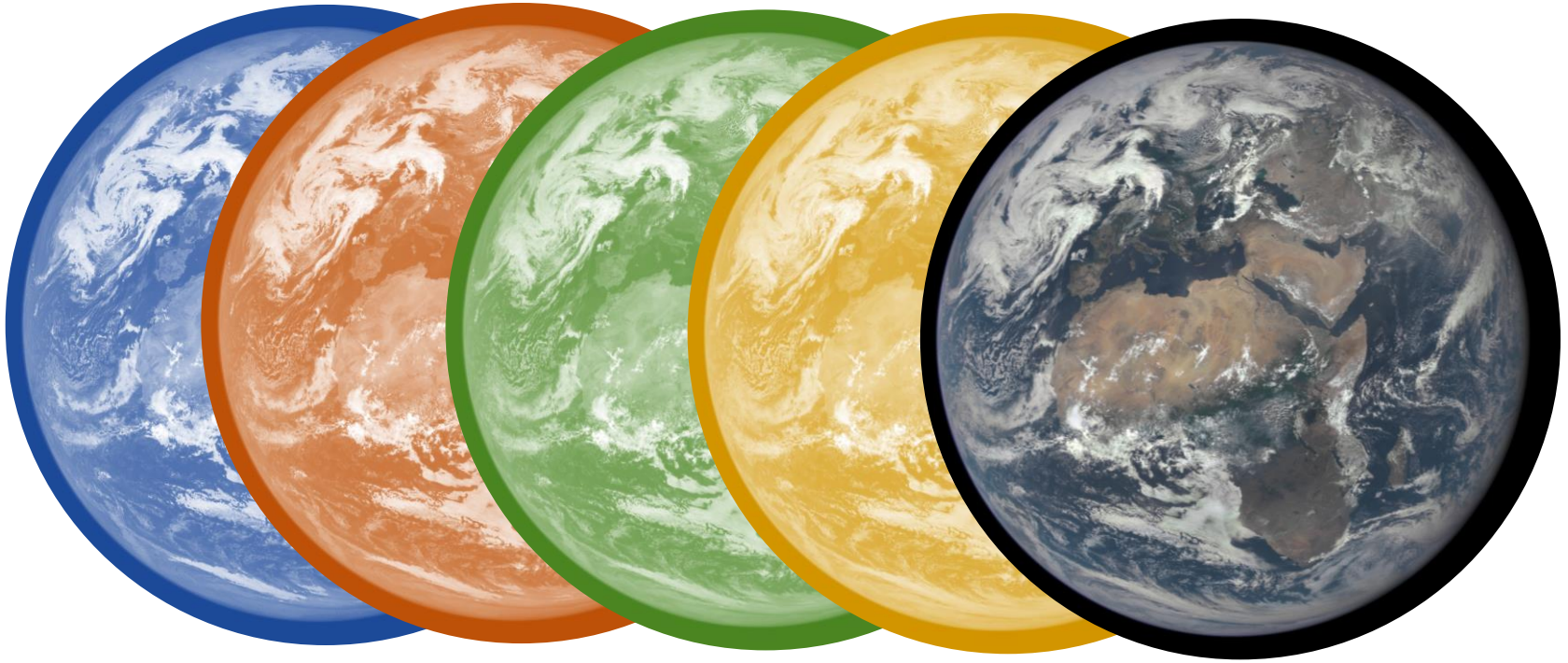
BiotaPhy

Environmental Layer

Raster data representing environmental values for cells in a map. Data may be numeric or categorical, with only one value per cell. Public data in BiotaPhy installations are in the 'Geographic' spatial reference system, latitude and longitude in decimal degrees.



Images from WikiCommons



Introducing the technical terminology

Geospatial data

Geospatial data are data with geographic location associated with it, i.e. map data. There are two kinds of spatial data, raster data and vector data. Each has properties that make it better for representing different information. Geospatial data are discussed in more detail on the Geospatial Data Terms page.

Raster (PNG)

Vector (SVG)

Images from WikiCommons

Introducing the technical terminology

Occurrence

An occurrence is a record of a specimen occurrence including metadata about the specimen and the spatial location where it was found.

Occurrence Data

Point data representing specimens collected for a single species or taxon. Each data point contains a location, x and y, in some known geographic spatial reference system. Public data in BiotaPhy installations are in the 'Geographic' spatial reference system, latitude and longitude in decimal degrees.

This .csv file contains occurrence data

	species_name	x	y
2	Bensoniella oregona	-123.751	42.802
3	Bensoniella oregona	-123.7903	42.802
4	Bensoniella oregona	-123.7903	42.802
5	Bensoniella oregona	-123.7707	42.7873
6	Bensoniella oregona	-123.751	42.9927

DATA FORMATTING

Introducing the technical terminology

BiotaPhy

Data Wrangler

Data wranglers are types of specify-lmpy data modification tools for filtering or editing data using specified criteria or methods.

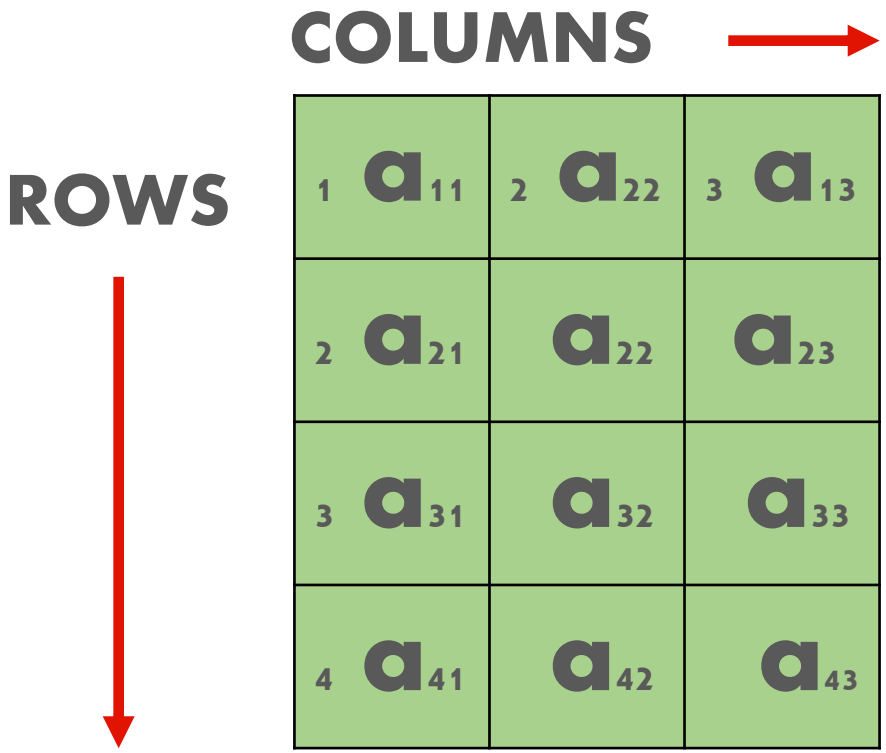
Image from WikiCommons

Introducing the technical terminology

BiotaPhy

Matrix

A Matrix is a multi-dimensional array of values.

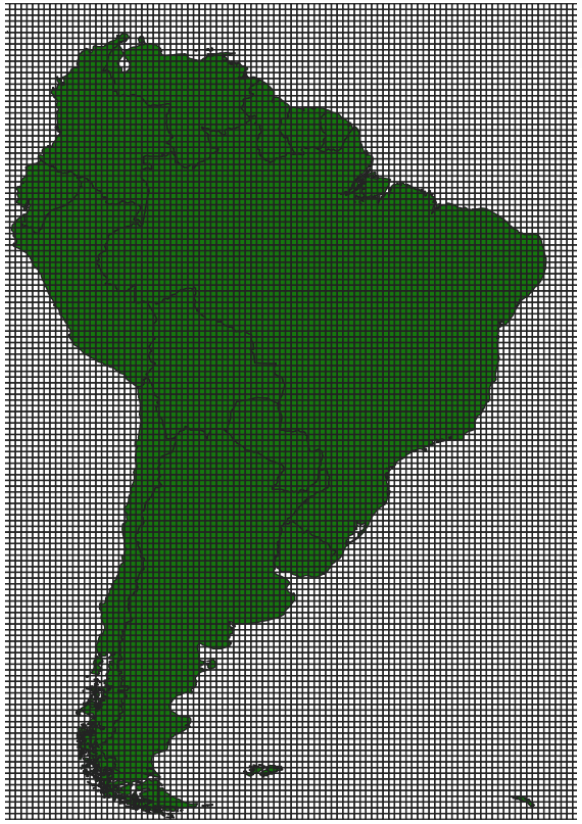


Introducing the technical terminology

Grid

A grid (in this context) is a geospatial region represented as a contiguous set of square polygons (cells) to be used for matrix creation. Grids are created as vector data, with one square polygon for every grid-cell, and stored in shapefile format.

Grid of South America!



Introducing the technical terminology

PAM

A Presence-Absence Matrix (PAM) is a 2-dimensional, binary Matrix of sites (rows) and species (columns). The matrix contains species distributions of 0 and 1 indicating presence or non-presence (absence) in each grid cell of a region. The matrix may be thought of as a three-dimensional cube of binary maps, with one layer per species. The 3-dimensional matrix is flattened into 2 dimensions, with rows representing sites with an x,y coordinate for the center of a gridcell on a map, and columns representing species.

	A	B	C	BNH	BNI	BNJ	BNK	BNL	BNM	BNN
1				Colobanthus apetalus	Colobanthus affinis	Rhagodia spinescens	Atriplex limbata	Einadia hastata	Atriplex incrassata	Atriplex nana
3006	3004	144.25	-27.75	0	0	1	1	0	0	0
3007	3005	144.75	-27.75	0	0	1	1	0	0	0
3008	3006	145.25	-27.75	0	0	1	1	0	0	0
3009	3007	145.75	-27.75	0	0	1	1	0	0	0
3010	3008	146.25	-27.75	0	0	1	0	0	0	0
3011	3009	146.75	-27.75	0	0	1	0	0	0	0
3012	3010	147.25	-27.75	0	0	1	0	0	0	0
3013	3011	147.75	-27.75	0	0	1	0	0	0	0
3014	3012	148.25	-27.75	0	0	1	0	0	0	0
3015	3013	148.75	-27.75	0	0	1	0	0	0	0
3016	3014	149.25	-27.75	0	0	1	0	0	0	0

TREES

Introducing the technical terminology

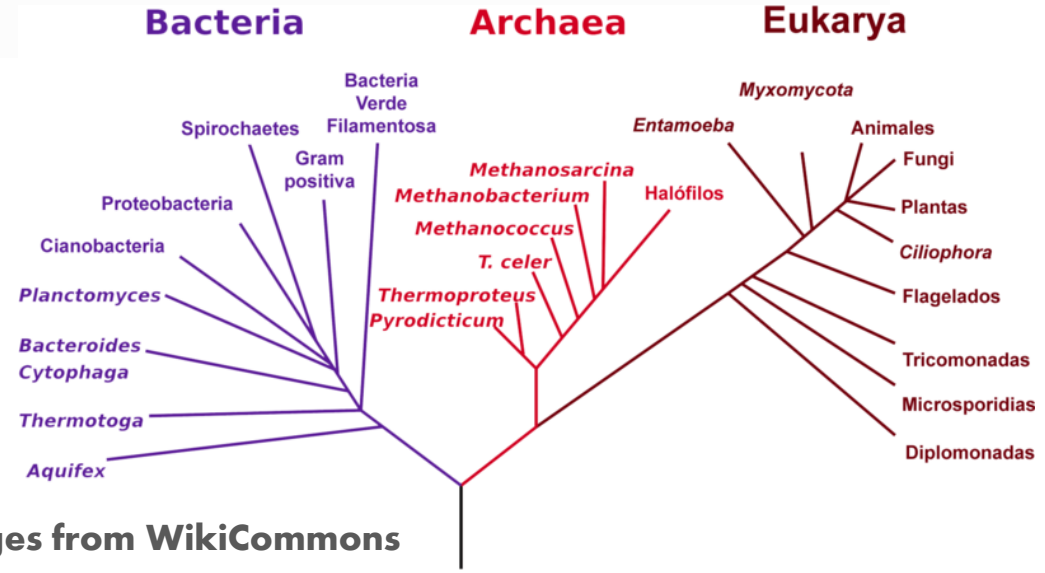
BiotaPhy

Tree

A Tree is a set of hierarchical data.

Phylogenetic Tree

A Phylogenetic Tree contains species names or identifiers for analyzing evolutionary patterns. BiotaPhy uses phylogenetic trees matching species data in a PAM to correlate evolutionary patterns with species distributions and landscape features. Trees are stored in Newick or Nexus format.

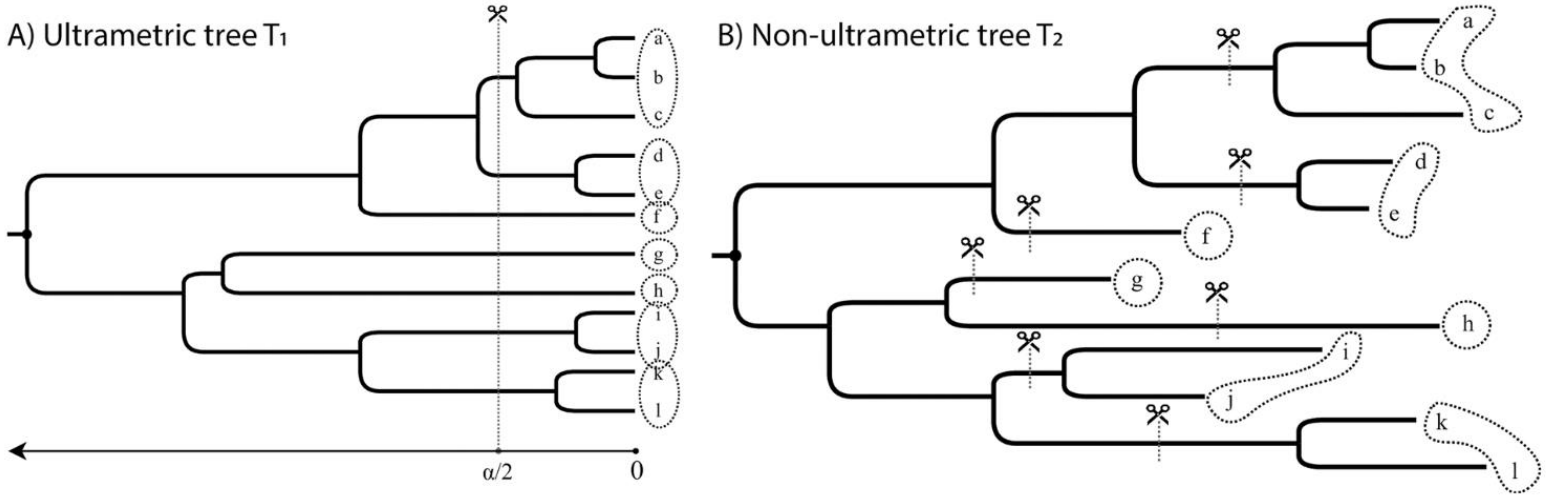


Images from WikiCommons

Introducing the technical terminology

Ultrametric Tree

A Phylogenetic Tree may contain numbers on the edges between species nodes corresponding to the hypothesized time between the evolution of one species node to the other. In an Ultrametric tree, the branch length from each tip in the tree up to the root, is equal to all other tip-to-root total lengths.



doi: https://doi.org/10.1371/journal.pone.0221068.g001

Balaban M, Moshiri N, Mai U, Jia X, Mirarab S (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS ONE 14 (8): e0221068

SDM

Introducing the technical terminology

SDM

Species Distribution Modeling (SDM) is also known by several other names, including environmental niche modeling, ecological niche modeling, and habitat modeling. SDM refers to the process of creating mathematical formulas (models) to predict the geographic distribution of species based on where they have been found and the environmental conditions in those locations.

Species Distribution Model

A species distribution model (SDM) is an estimation of potential habitat for a particular species.

Prunus geniculata **(scrub plum)**

SDM for *Prunus geniculata*