ECO 602
Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

McGarigal Chapter 3

1. sections 1 – 4
2. sections 4 – 6
3. **sections 6 - 9**

Today's Agenda

1. Associations recap
2. Data dimensionality
3. Quiz and discussion
4. Missing data
5. Variable sufficiency
6. Transformations and extreme values
7. Assignment 2

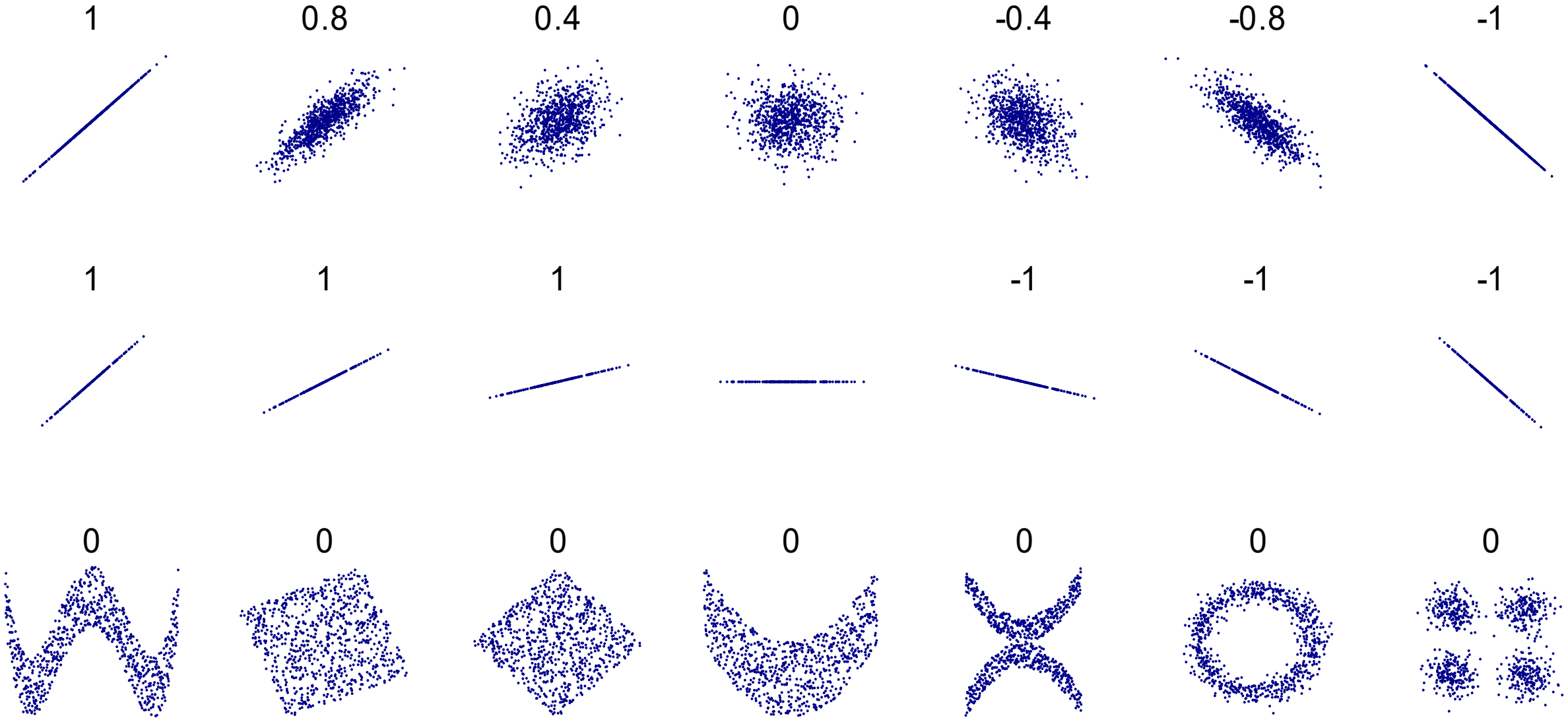
Associations recap

‘Association’ is a value-neutral term. It is useful when you don’t want to imply causality, or any specific form of a relationship.

Association is a general concept in modeling.

Our goal is to quantify the strength of associations

Our tools include Pearson’s correlation for _____ associations, and Spearman’s correlation for _____ associations.



Public domain image: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

Data dimensionality

What do I mean by data dimensionality?

What does it mean in terms of model thinking?

What is the relationship between sampling units and data dimensionality?

Can you ever reduce data dimensionality?

Data dimensionality

How many variables were there in your Question Set 1 datasets?

How have we visualized:

- 1D data (single variable)
- 2D data (two variable)
- 3D data
- 4D + data

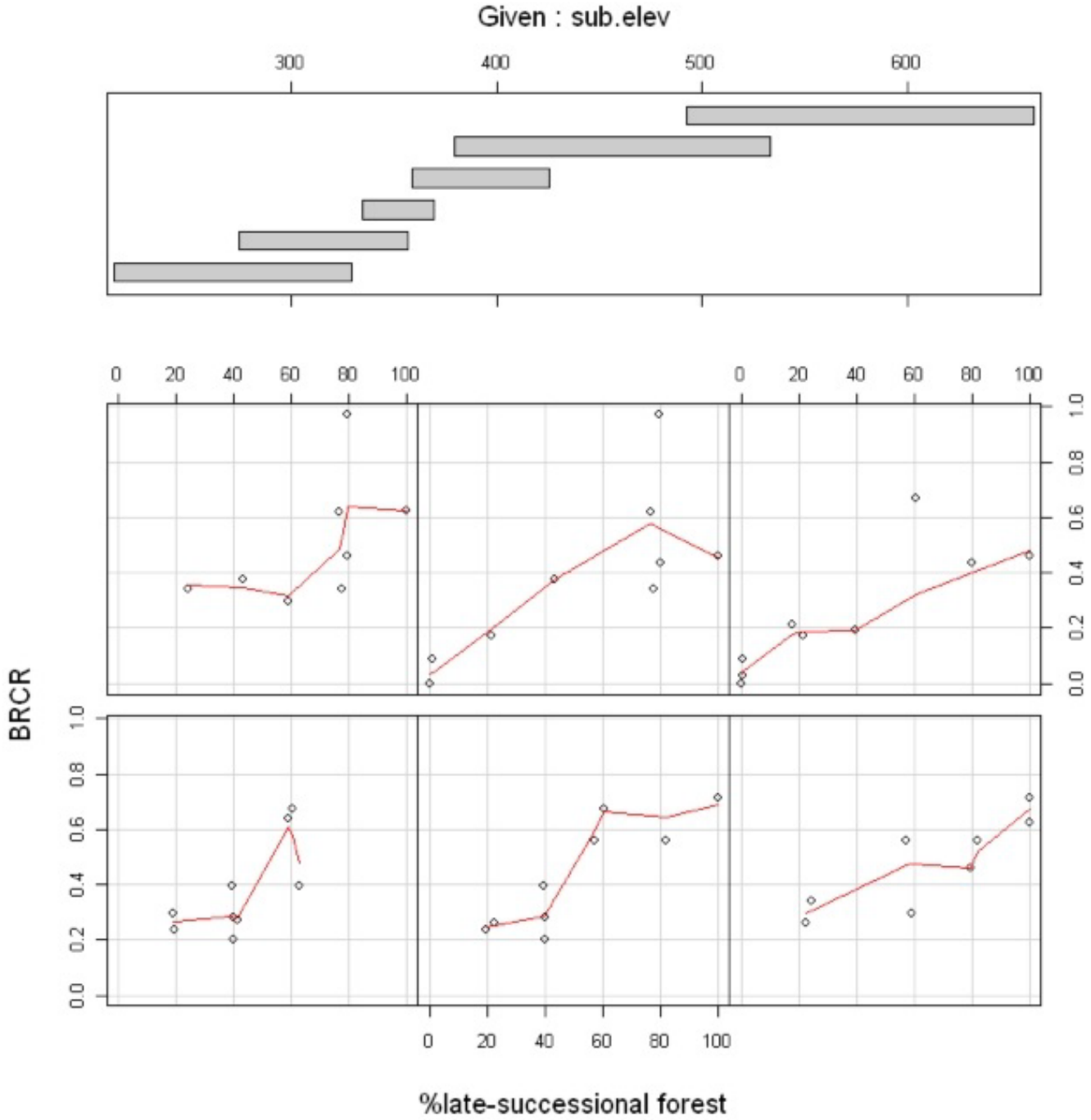
Coplots

A way to visualize 3D data using 2D slices.

The data points are plotted on the x-y plane.

The z axis is divided into bins*

Each z-bin is flattened and the data presented in 2D


Given : sub.elev
300
400
500
600
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
BRCR
%late-successional forest

Scatterplot matrix

Scatterplot of each possible pairwise combination of variables.

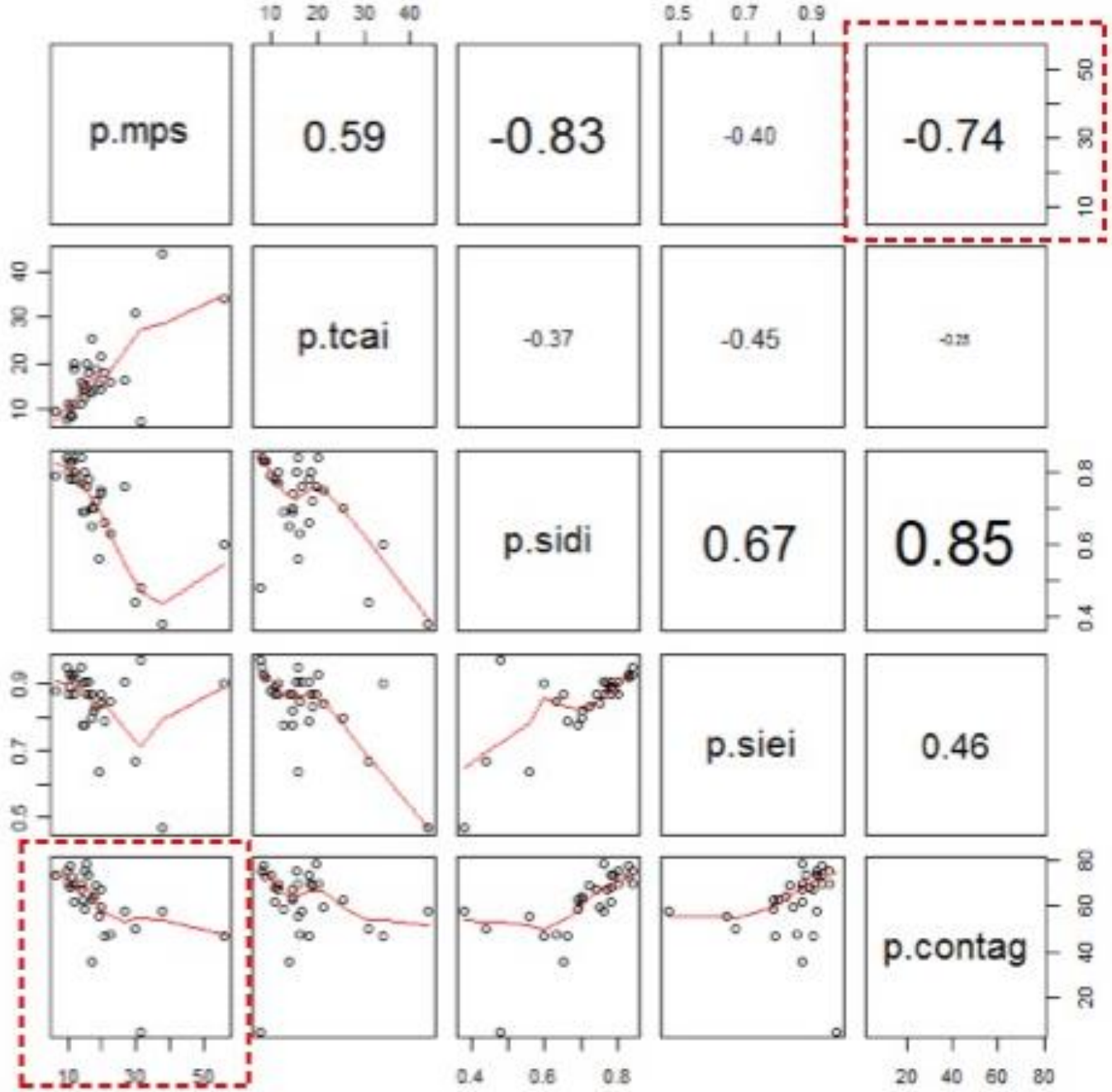
May include correlation values

A way to visualize multidimensional data.

What are the lurking variables in each 2D plot?

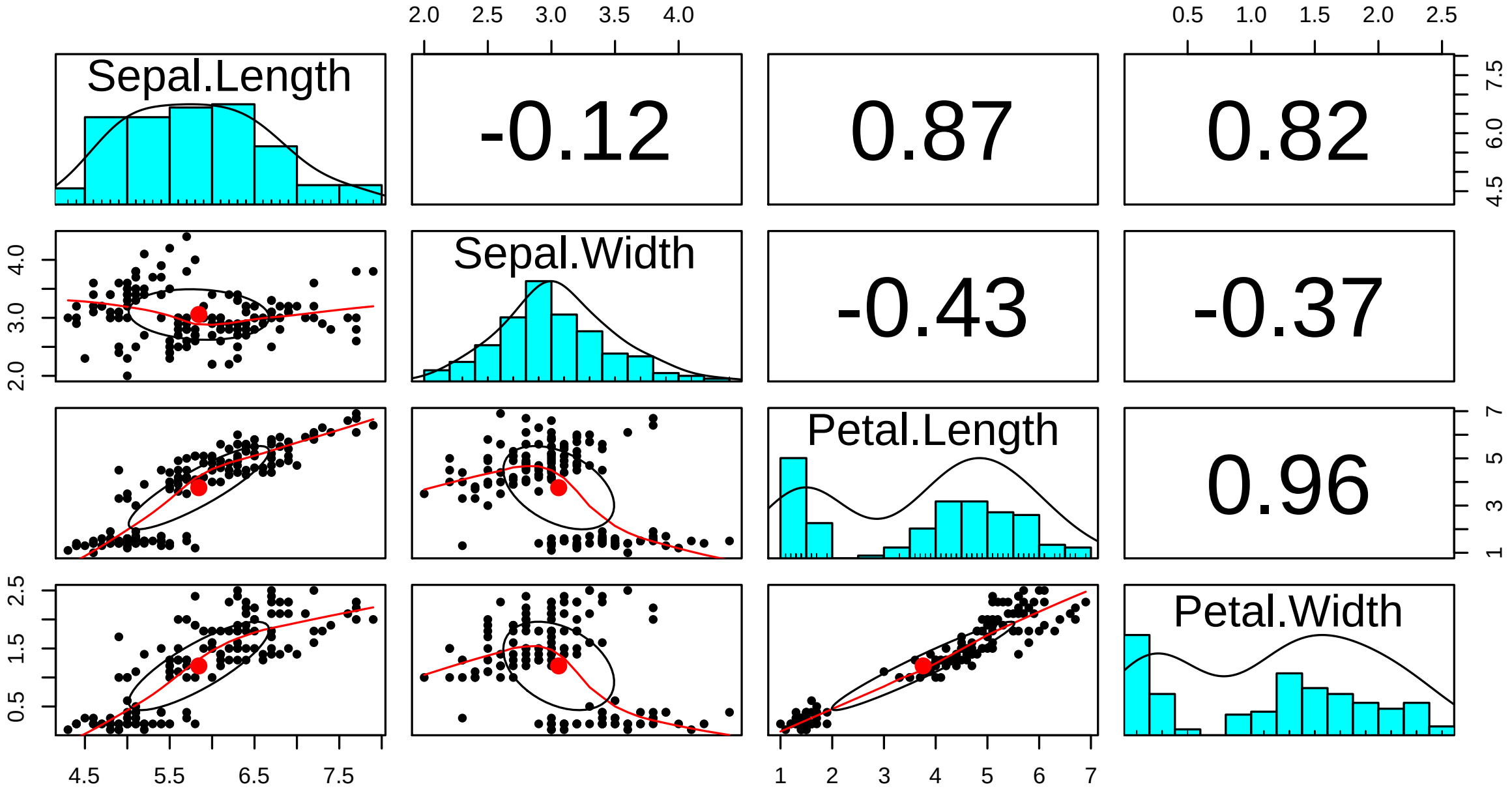
- Matrix of scatterplots for every combination of variables (and the corresponding correlation coefficients)

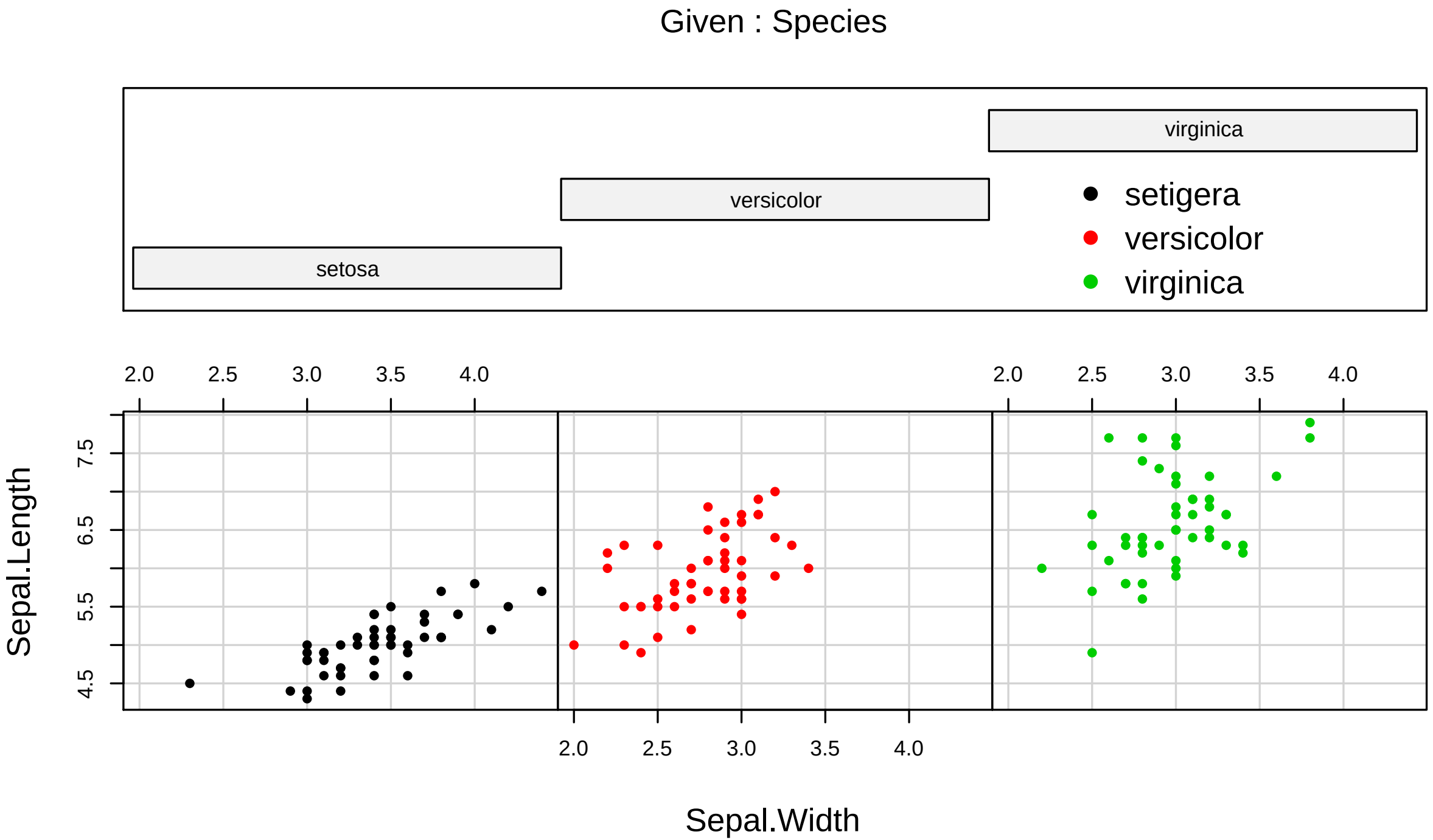
	p.mps	p.tcai	p.sidi	p.siei	p.contag
1	17.19	25.59	0.70	0.80	62.97
2	22.54	16.01	0.63	0.85	47.32
3	10.04	11.41	0.80	0.87	69.69
4	17.93	18.66	0.72	0.83	69.42
5	11.80	20.05	0.84	0.93	70.15
6	11.91	18.54	0.80	0.87	69.74
7	26.85	16.29	0.76	0.91	58.12
8	19.53	21.32	0.75	0.84	59.29
9	37.65	43.83	0.38	0.47	57.88
10	29.88	31.07	0.44	0.67	50.12

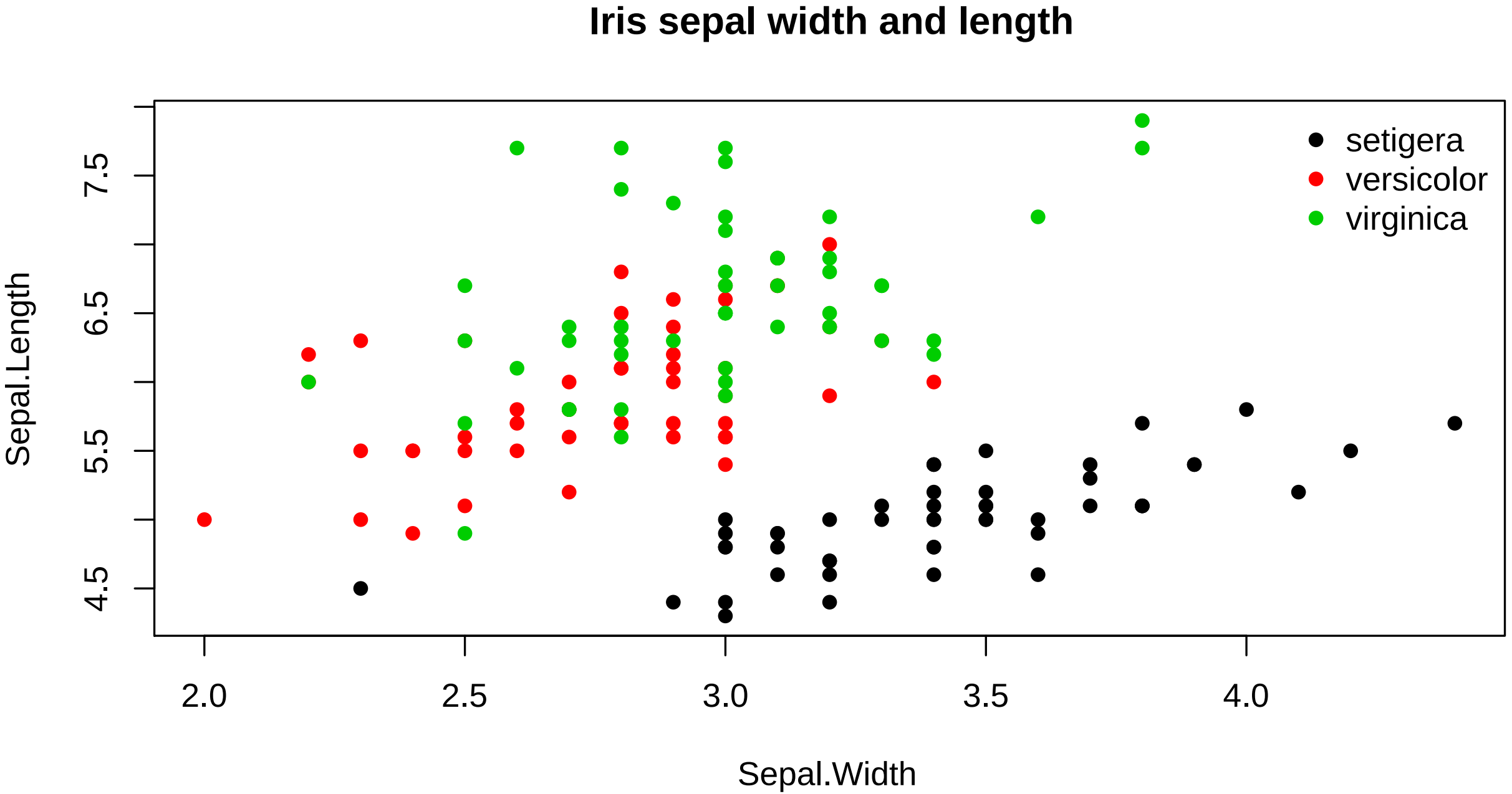


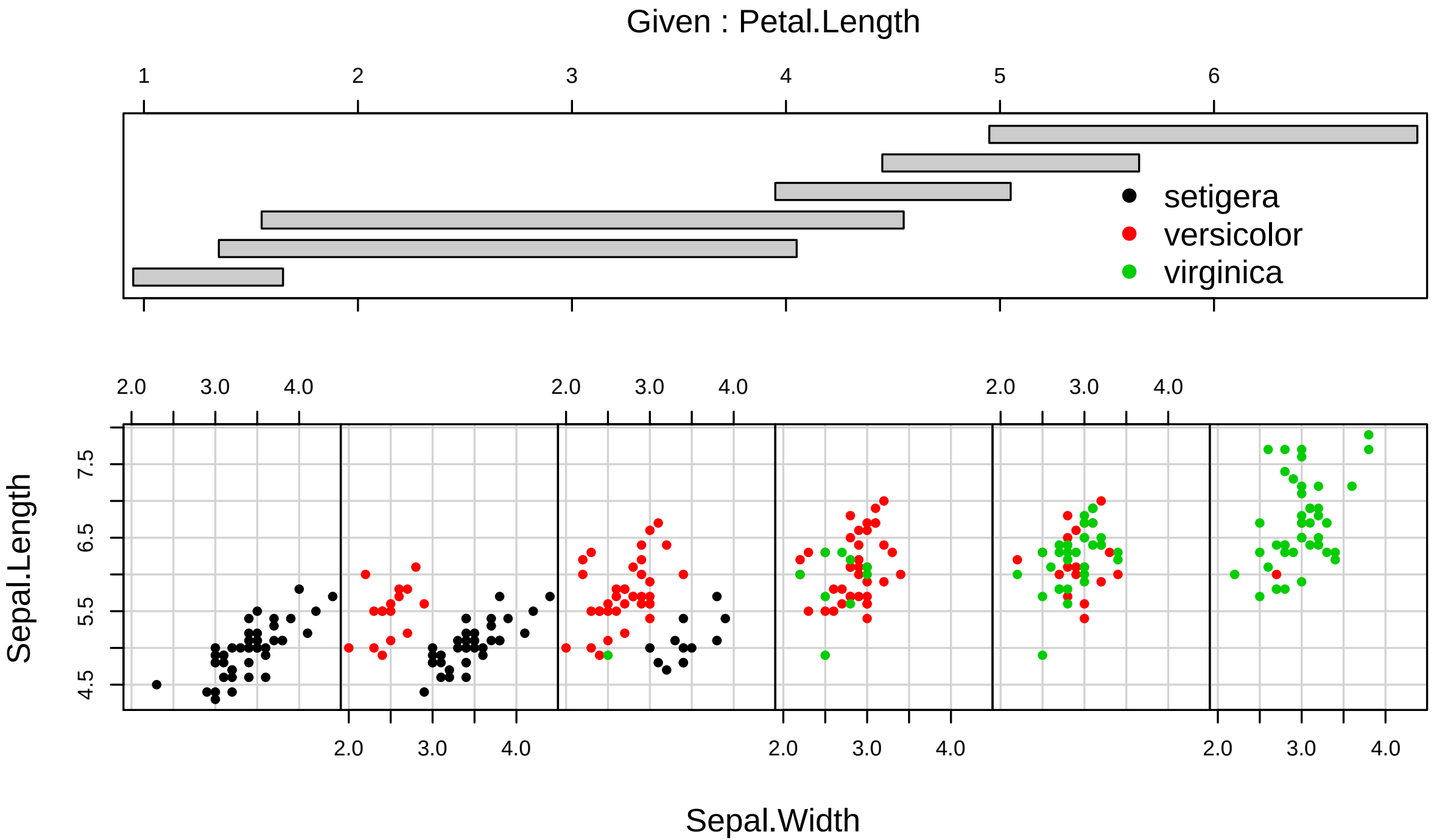
Iris example data in R

	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


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
-0.12
0.87
0.82
-0.43
-0.37
0.96
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 6.0 7.5
2.0 3.0 4.0
1 3 5 7
0.5 1.5 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7


Given : Species
virginica
versicolor
setosa
setigera
versicolor
virginica
2.0
2.5
3.0
3.5
4.0
4.5
5.5
6.5
7.5
Sepal.Length
Sepal.Width


Iris sepal width and length
Sepal.Length
Sepal.Width
4.5
5.5
6.5
7.5
2.0
2.5
3.0
3.5
4.0
setigera
versicolor
virginica


Given : Petal.Length
1
2
3
4
5
6
setigera
versicolor
virginica
2.0
3.0
4.0
7.5
6.5
5.5
4.5
Sepal.Length
Sepal.Width

3D plot in RStudio

Example with iris dataset and package rgl

Associations quiz

10 minutes for quiz and stretch break.

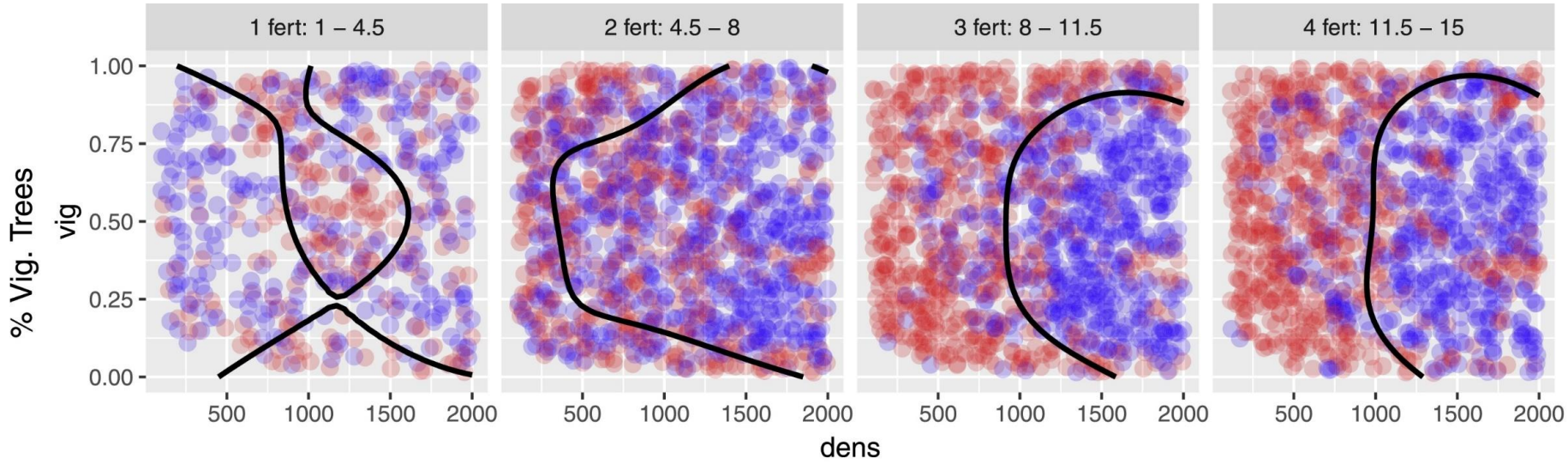
4-diemensional slice plots

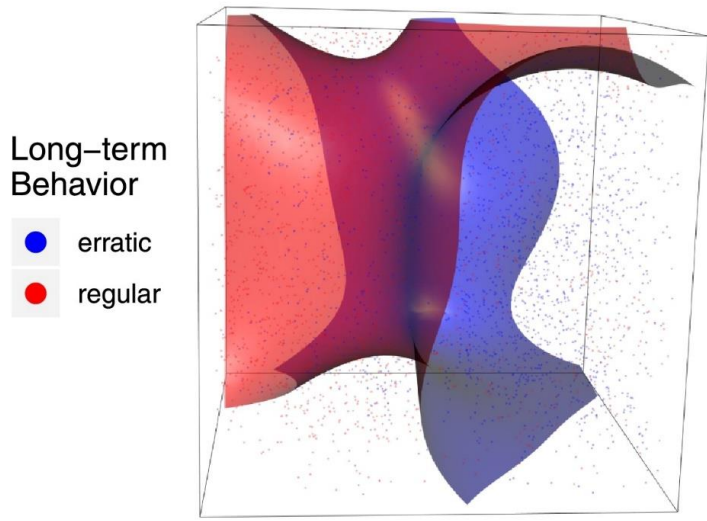
3 model continuous parameters: predictors on x, y, z axes

- Tree vigor
- Beetle fertility
- Tree density

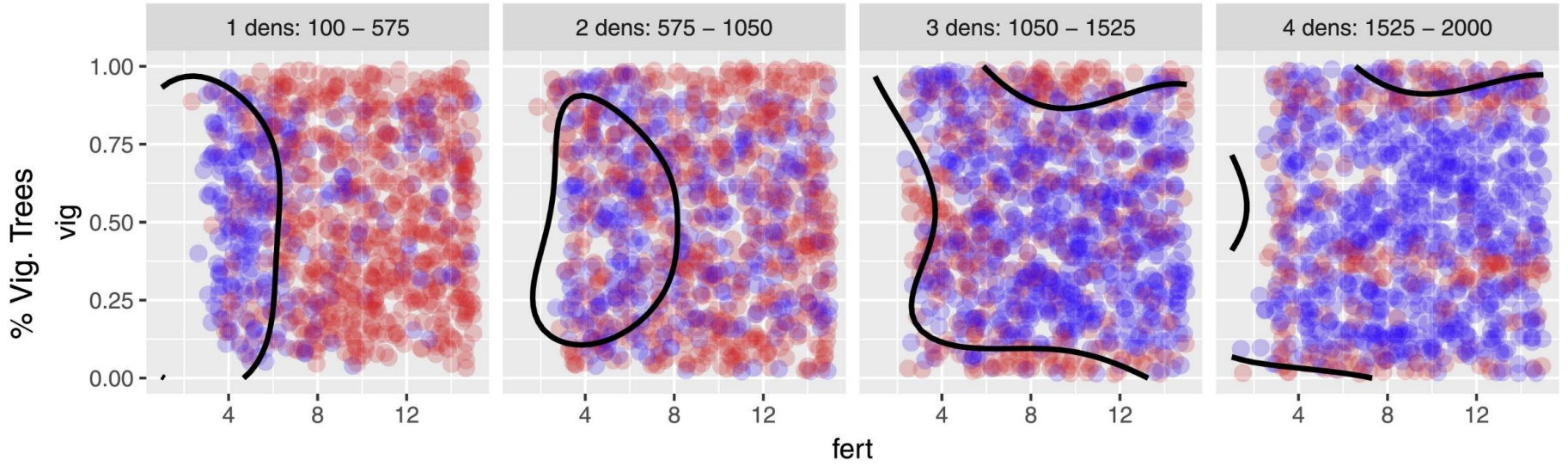
1 binary response variable: Epidemic return interval

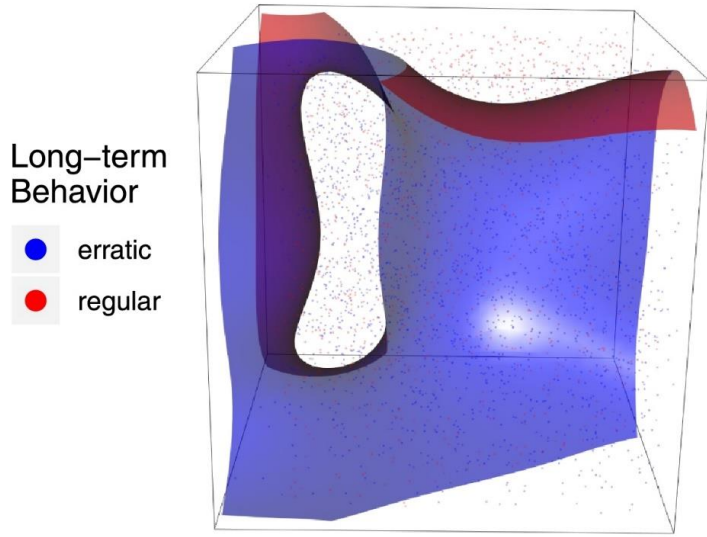
- Red = regular intervals, ca. 80 years
- Blue = erratic return intervals





Max trees per ha.





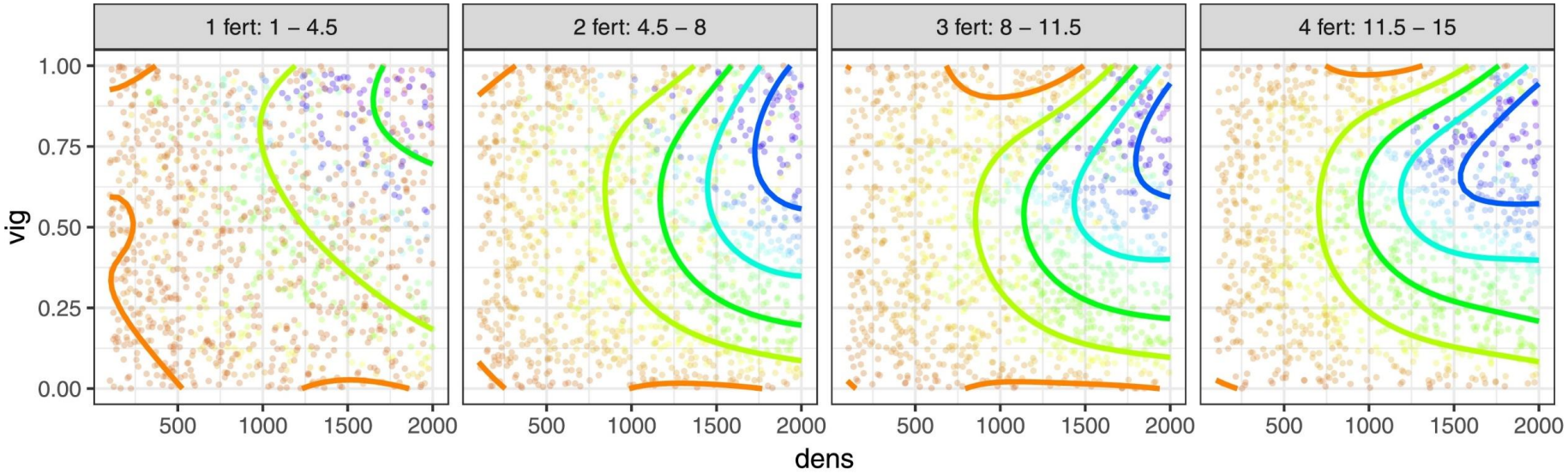
4-diemensional slice plots

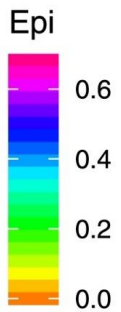
3 model continuous parameters: predictors on x, y, z axes

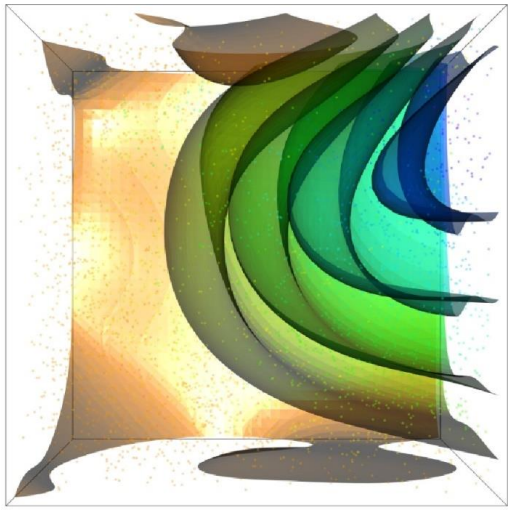
- Tree vigor
- Beetle fertility
- Tree density

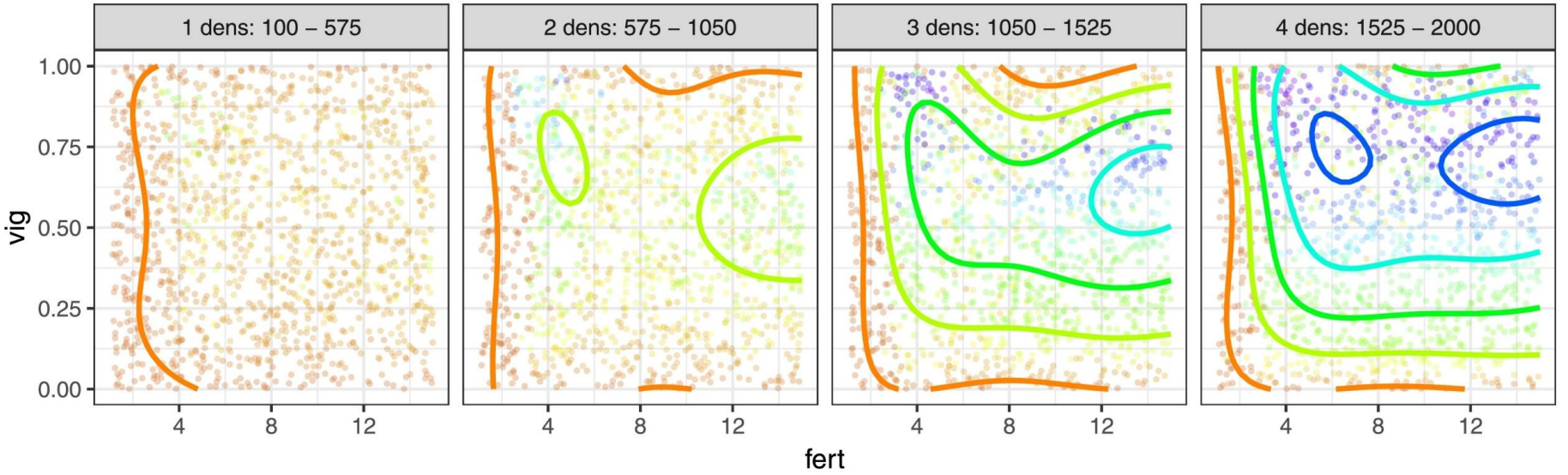
1 continuous response variable: Epidemic index

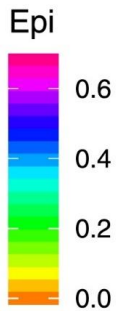
- Unitless, 0 – 1 range
- Approximately the proportion of area in epidemic state through time

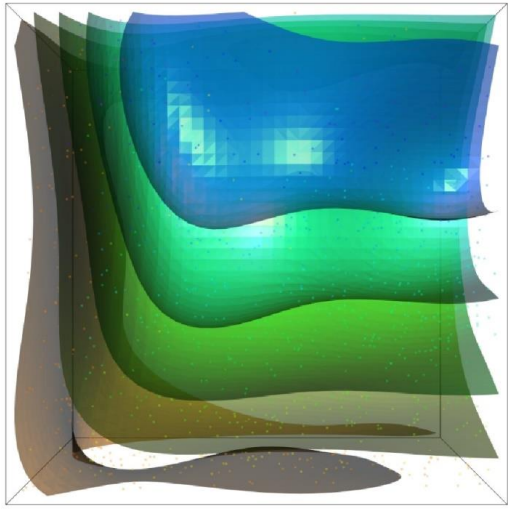

1 fert: 1 – 4.5
2 fert: 4.5 – 8
3 fert: 8 – 11.5
4 fert: 11.5 – 15
vig
1.00
0.75
0.50
0.25
0.00
500
1000
1500
2000
dens


Epi
0.6
0.4
0.2
0.0




1 dens: 100 – 575
2 dens: 575 – 1050
3 dens: 1050 – 1525
4 dens: 1525 – 2000
vig
1.00
0.75
0.50
0.25
0.00
4
8
12
fert


Epi
0.6
0.4
0.2
0.0



Scatterplot matrix vs. coplots

Scatterplot matrices and coplots are different ways of looking at multi dimensional data.

Coplots show 3D data, scatterplot matrices can show n-dimensional data.

Coplots explicitly show a 3rd, lurking variable by plotting slices through 3D space.

Scatterplot matrices hide effects of lurking variables in individual panes but can summarize more than 3 dimensions.

Assignment 1: peer review forms

1. 30% of your grade
2. Forms will be available on Moodle for 5 days following the due date (Sep 29th, midnight)
3. Constructive comments

Assignment 1: group time

Many of you have sent me drafts.

Let’s take 10 – 15 minutes to discuss my comments in your groups and ask me questions.

Missing data

How to deal with missing data depends on data dimensionality.

Different analyses have different ways to deal with (or not) missing data.

2D data: doesn't make sense to include sampling units with missing data for one variable.

Missing data

N-dimensional data:

- Excluding sampling units with missing value in 1 variable means losing potentially informative data. But… there are lots of pitfalls (in scare-quotes below).

Imputation

- Taking average is a ‘neutral’ option
- Statistical model techniques might find ‘better’ values for missing data

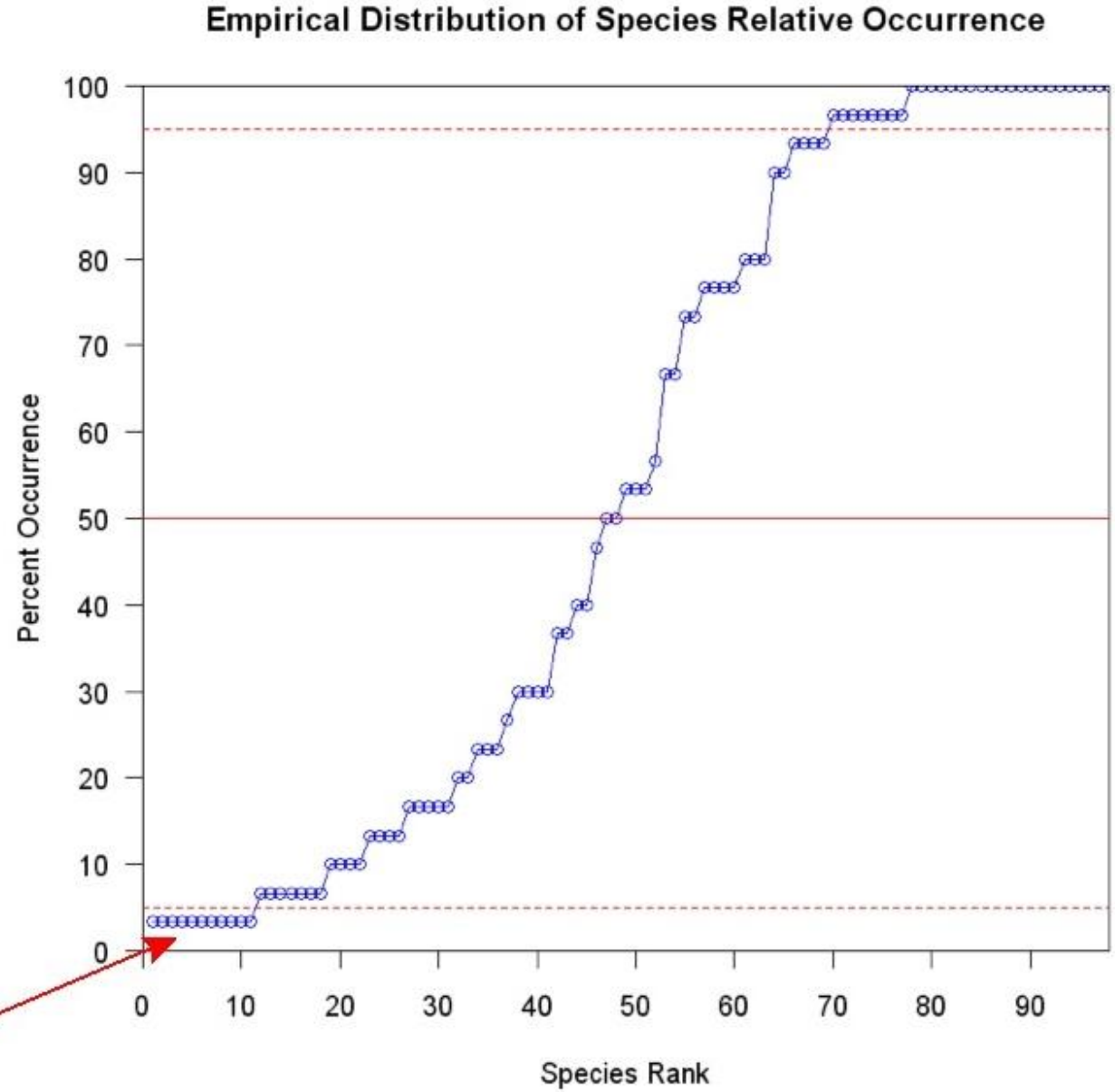
Variable sufficiency

For categorical variables: if very few sampling units fall into a category, that category may not be sufficiently represented in the sample.

Making inferences about low-frequency events is tricky.

Problems with detection limits, sampling error, etc.

- Does the variable hold sufficient information to be deemed reliable and meaningful?
- Example: rare species in community data sets are not likely to be accurately placed in ecological space



Data transformations

Usually we want to make a nonlinear relationship linear.

Variance stabilization (more later)

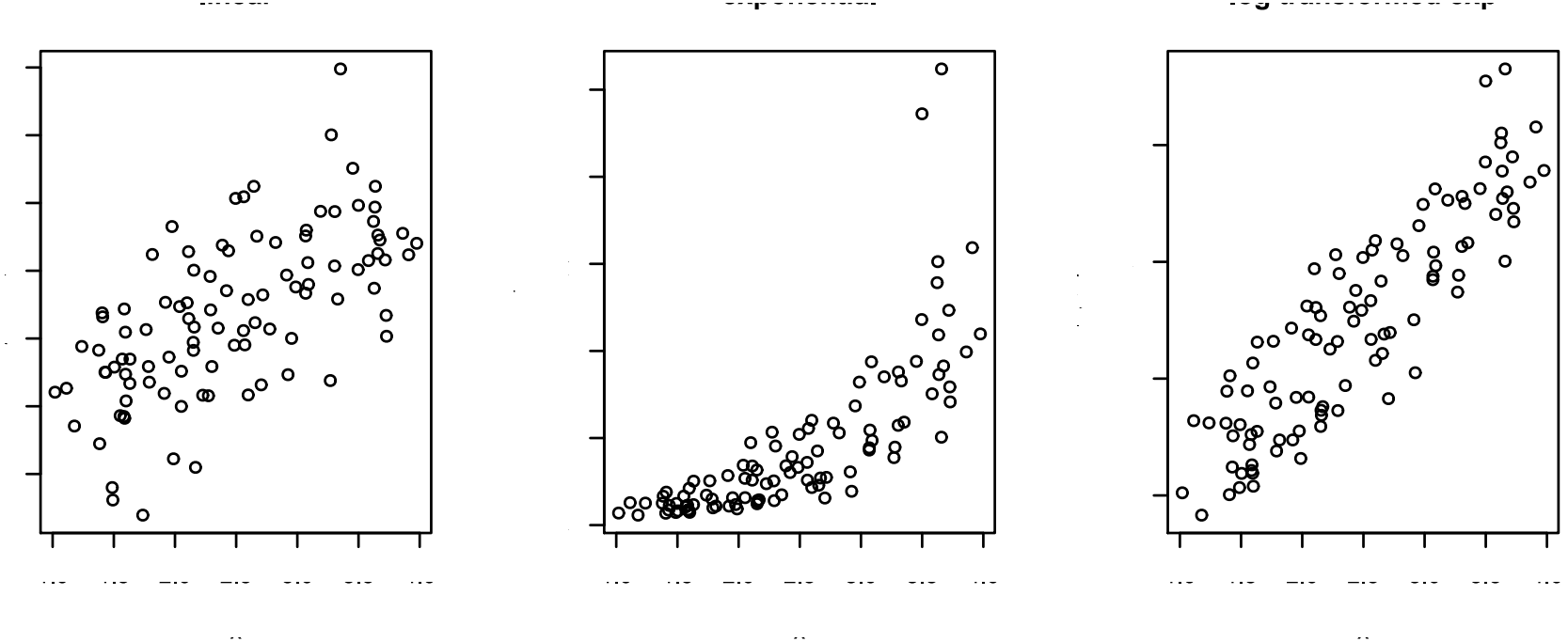
Linear relationships are analytically simple.

Linear relationships are easy to interpret.

Interpretation of transformed variables can be difficult.

We'll discuss standardizations later.

Log transformations



Extreme values

Extreme values may be due to error:

- Measurement
- Data entry
- Transcription

Extreme values may be real

Extreme values can violate inference assumptions.

We'll revisit extreme values many times.

Assignment 2

In groups, choose a paper from the abstract examples from last Thursday's in-class activity.

Assignment instructions and papers are on Moodle.

Model Thinking

How do we describe a linear association verbally?

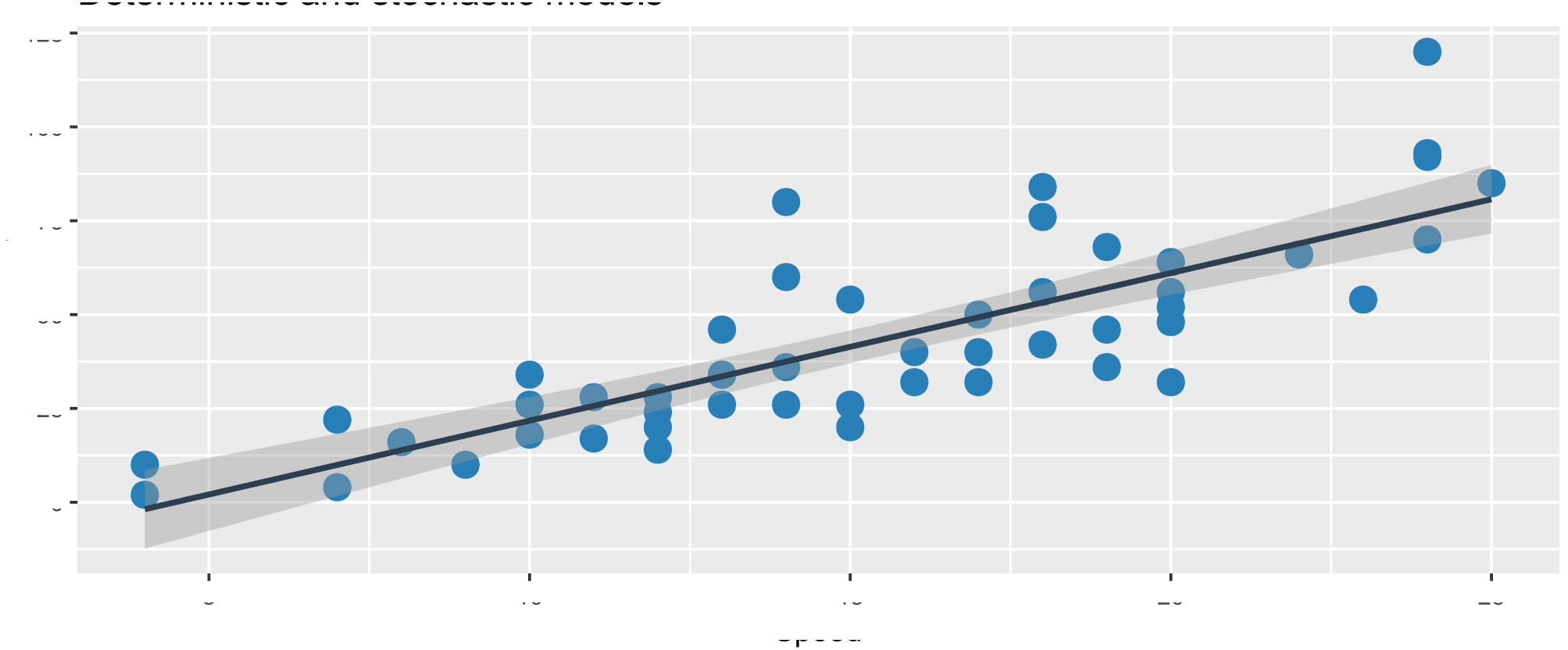
What are the essential components of a conceptual model of a linear association?

Model Thinking: 2 models

Deterministic model

Stochastic model

Model Thinking: 2 models



Model Thinking: 2 models

Deterministic functions

Probability distributions

For next time:

McGarigal Chapter 4: don’t try to memorize details about all of the specific deterministic functions!

We’ll discuss the functions in terms of broader categories.

We’ll go into detail about only a few, for now.