Data wrangling & manipulation in R

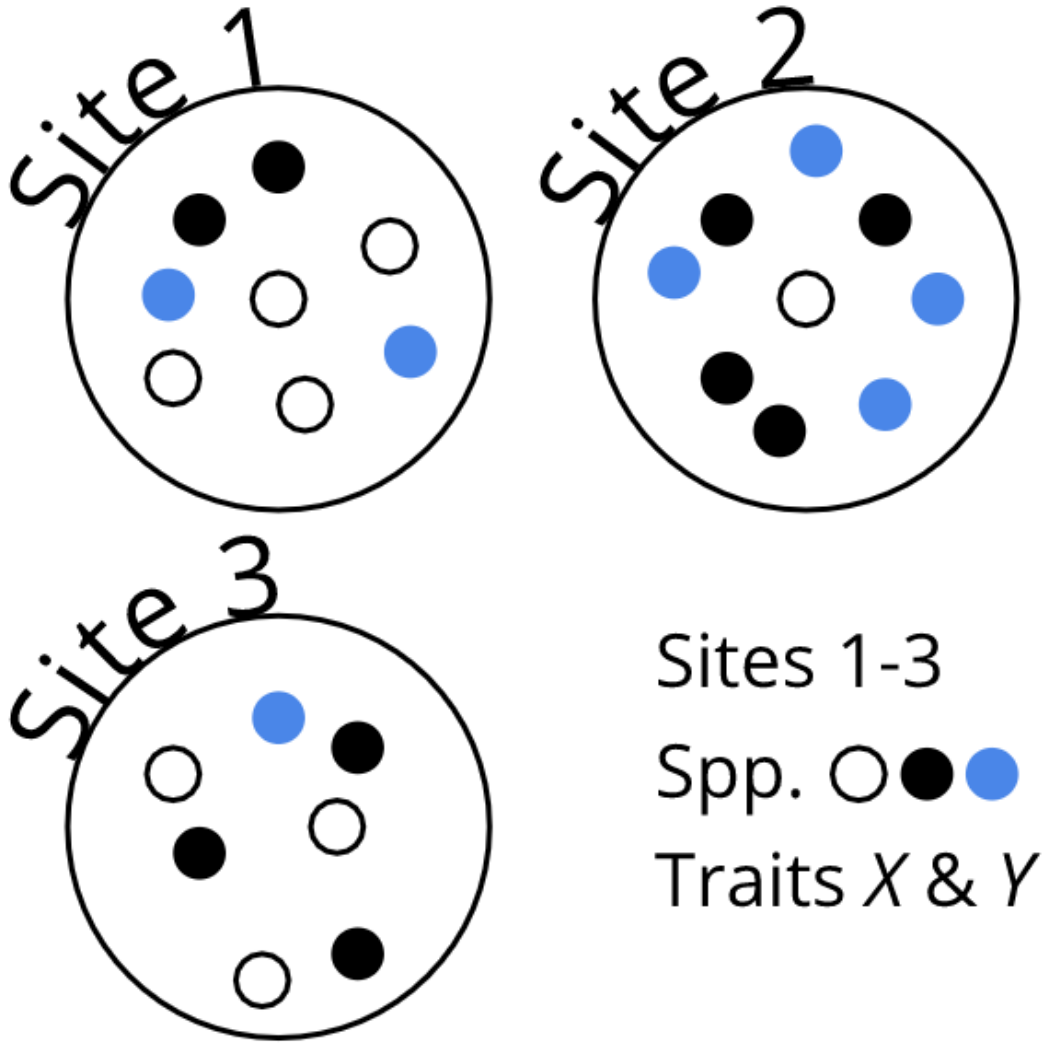
Dept. Biological Sciences Postgraduate Workshop

Ruan van Mazijk, MSc candidate

2019-05-13

Motivation

Motivation



An example data-collection scenario in biology

Site 1					Site 1					
Sp. 1		Sp. 2		Sp. 3		Sp. 1		Sp. 2		Sp. 3	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y

One way to lay out your collected data...

Site 1			Site 2		
Sp.	X	Y	Sp.	X	Y

Another way...

Site	Sp.	X	Y

The “best” way. Will make your life easiest in the long-term.

Workshop outline

- Embracing the rectangle

Workshop outline

- Embracing the rectangle
- **Making** your data rectangular

Workshop outline

- Embracing the rectangle
- **Making** your data rectangular
- Things to see & do in rectangle land

Workshop outline

- Embracing the rectangle
- **Making** your data rectangular
- Things to see & do in rectangle land
- `mutate()` & friends—How to extend your raw dataset

Workshop outline

- Embracing the rectangle
- **Making** your data rectangular
- Things to see & do in rectangle land
- `mutate()` & friends—How to extend your raw dataset
- ~~Complicated~~ Exotic problems

Embracing the rectangle

Embracing the rectangle

Long vs wide data

Remember this?

Site 1						Site 1					
Sp. 1		Sp. 2		Sp. 3		Sp. 1		Sp. 2		Sp. 3	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y

Embracing the rectangle

Long vs wide data

Remember this?

Site 1						Site 1					
Sp. 1		Sp. 2		Sp. 3		Sp. 1		Sp. 2		Sp. 3	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y

This is *wide-form* data. Let's move away from that…

Using the `iris` dataset built into R!

Wide-form data

Wide-form data

```
## $setosa
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
##
## $versicolor
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          7.0         3.2          4.7         1.4
## 2          6.4         3.2          4.5         1.5
##
## $virginica
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          6.3         3.3          6.0         2.5
## 2          5.8         2.7          5.1         1.9
```

Classic long-form data

Classic long-form data

```
##     Species Sepal.Length Sepal.Width Petal.Length Petal.W
## 1    setosa          5.1         3.5          1.4       0.2
## 2    setosa          4.9         3.0          1.4       0.2
## 3    setosa          4.7         3.2          1.3       0.2
## 4 versicolor          7.0         3.2          4.7       1.4
## 5 versicolor          6.4         3.2          4.5       1.5
## 6 versicolor          6.9         3.1          4.9       1.5
## 7  virginica          6.3         3.3          6.0       2.5
## 8  virginica          5.8         2.7          5.1       1.9
## 9  virginica          7.1         3.0          5.9       2.1
```

Site 1			Site 2		
Sp.	X	Y	Sp.	X	Y

We can get longer…

We can get longer...

```
##       Species        trait trait_value
## 1      setosa Sepal.Length         5.1
## 2  versicolor Sepal.Length         7.0
## 3   virginica Sepal.Length         6.3
## 4      setosa  Sepal.Width         3.5
## 5  versicolor  Sepal.Width         3.2
## 6   virginica  Sepal.Width         3.3
## 7      setosa Petal.Length         1.4
## 8  versicolor Petal.Length         4.7
## 9   virginica Petal.Length         6.0
## 10     setosa  Petal.Width         0.2
## 11 versicolor  Petal.Width         1.4
## 12  virginica  Petal.Width         2.5
```

The advantages of long data

- Machine-readable

The advantages of long data

- Machine-readable
- The standard for most software/R-functions (e.g. `lm()`, `plot()`, `ggplot()`)

The advantages of long data

- Machine-readable
- The standard for most software/R-functions
  (e.g. `lm()`, `plot()`, `ggplot()`)
- How most statistical methods treat data mathematically

The advantages of long data

- Machine-readable
- The standard for most software/R-functions
  (e.g. `lm()`, `plot()`, `ggplot()`)
- How most statistical methods treat data mathematically
- Easier to subset & wrangle further!

Making your data rectangular

Making your data rectangular

What are your options?

Easiest to lay it out like that from the start…

Tools to follow assume your data is nice & *tidy*

Things to see & do in rectangle land

`mutate()` & friends

How to extend your raw dataset

~~Complicated~~ Exotic problems

dplyr::

- select()
- filter()
- group_by()
- summarise()
- arrange()
- join()
- mutate()

tidyr::

- gather()
- spread()
- separate()
- unite()