ST2304 - Statistical Modelling for Biologists/Biotechnologists

Bob O'Hara

bob.ohara@ntnu.no

Administration Matters

(we will deal with these in more detail later)

- Reference Group
- Blackboard
- web page:
  `https://www.math.ntnu.no/emner/ST2304/2019v/`

Assessment

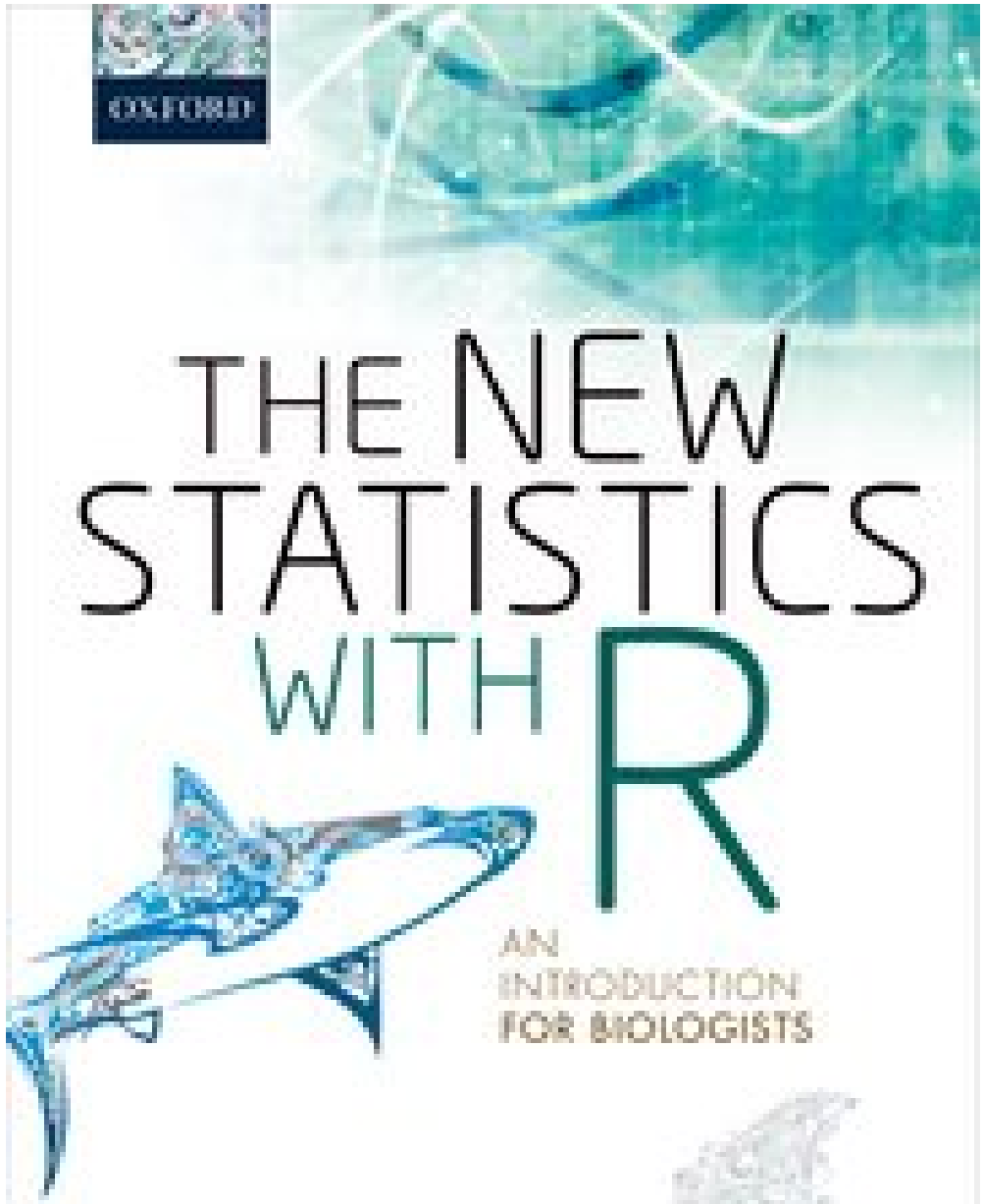
Complete 6 exercise sets (of about 10)

- do in groups
- pass/fail
- first couple of weeks won't count

An Examination

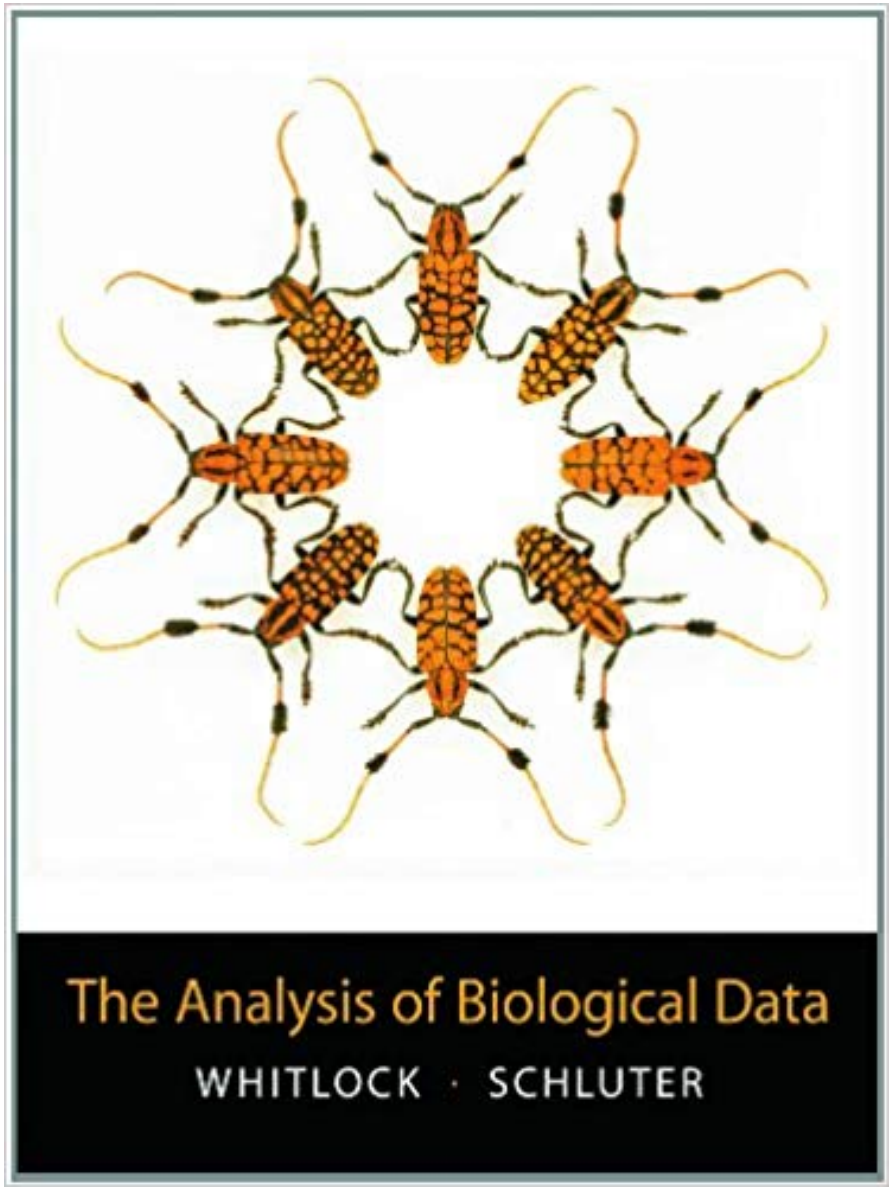
Text Books

New Statistics with R - Andy Hector



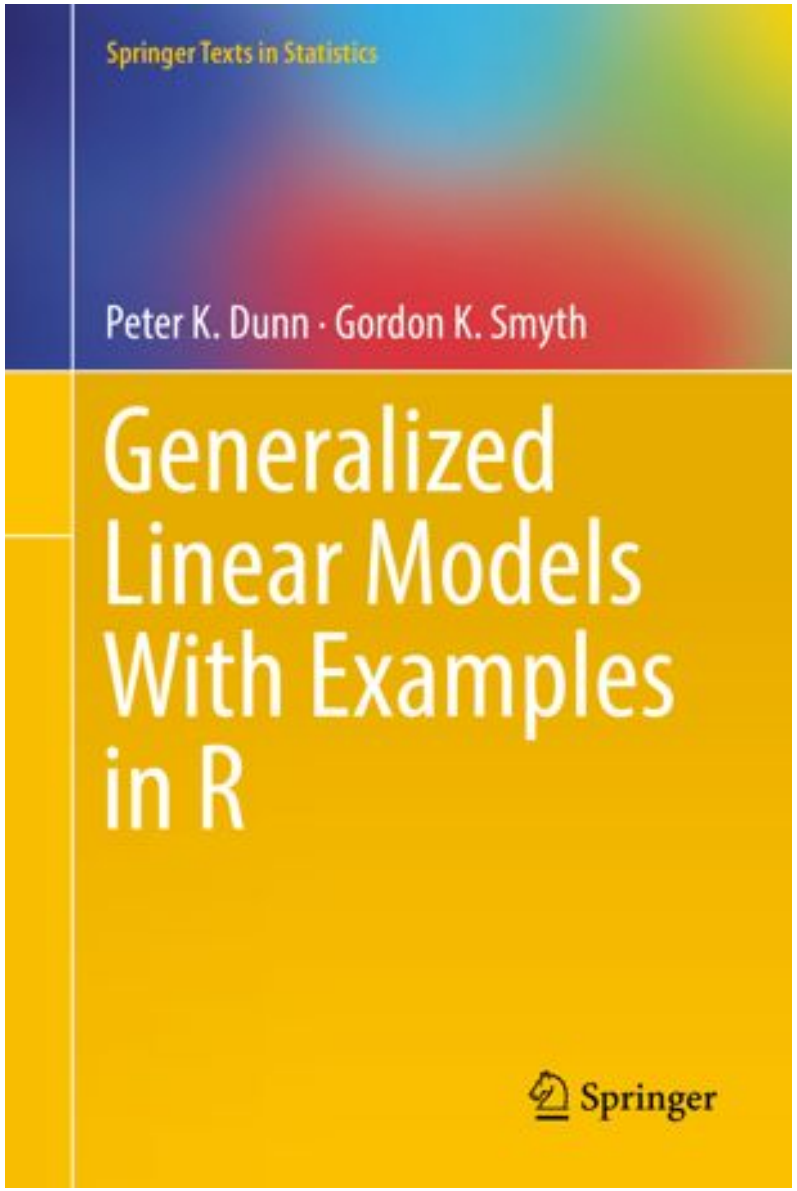
Text Books

The Analysis of Biological Data - Whitlock & Schluter



Text Books

Generalized Linear Models With Examples in R - Dunn & Smyth



Why do we need statistical modelling?

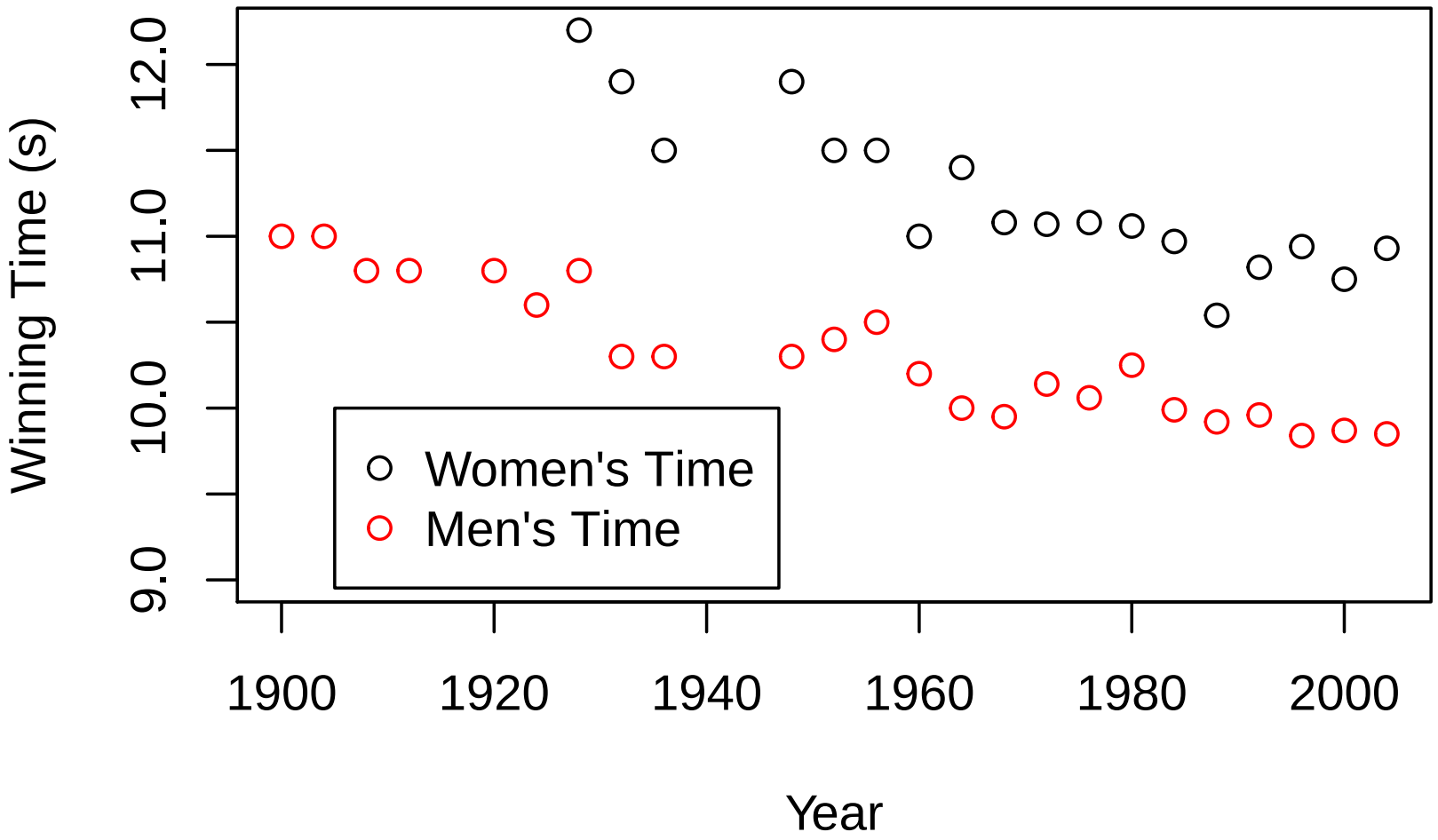
Because the world is complex!

Exercise one: Sorting yourselves out

Get into groups

Exercise one: Assessing data

Data on the winning times in the mens & womens Olympic 100m



- What can you tell about 100m times from the data?
- What conclusions could you draw?

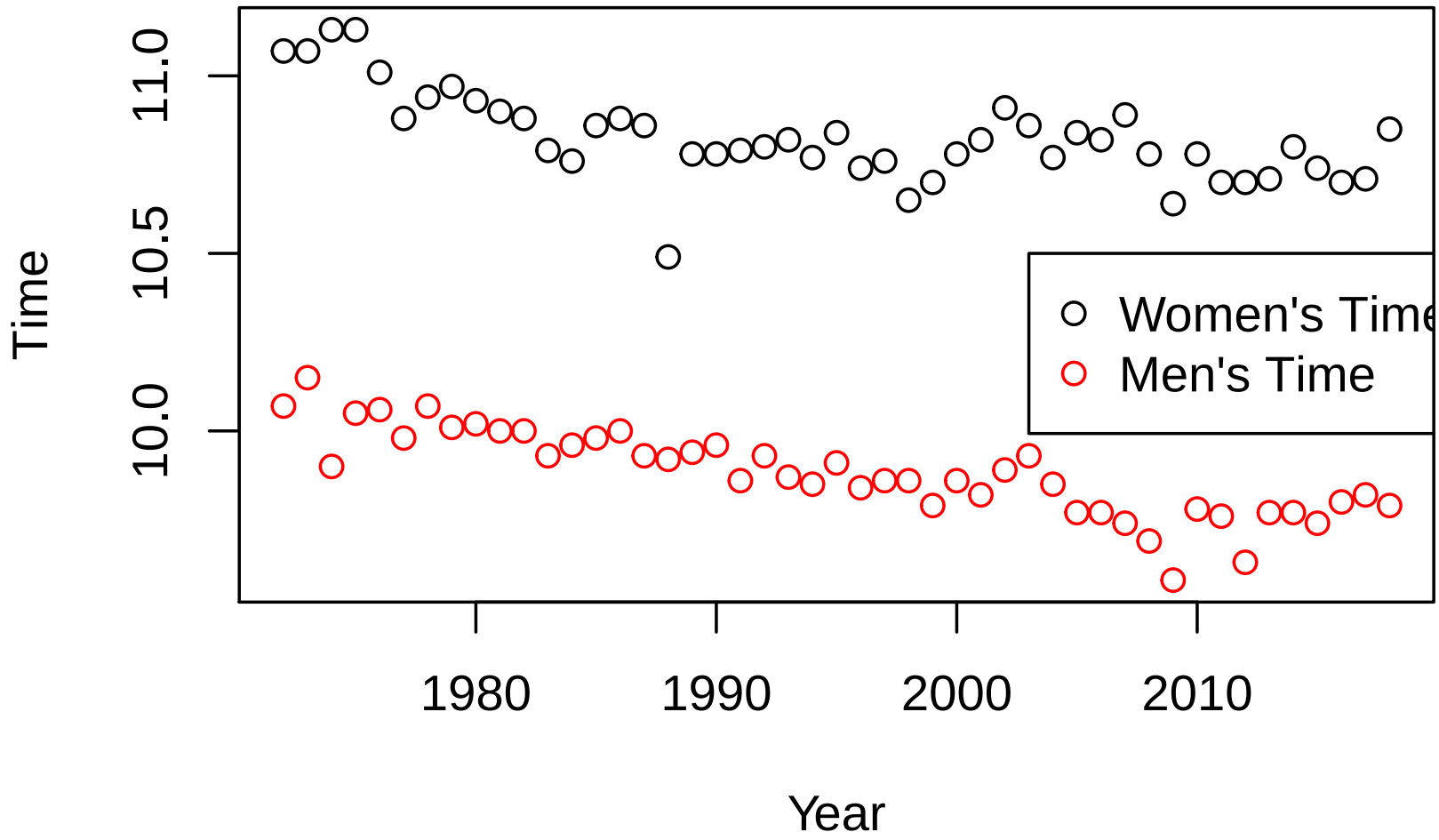
Exercise one: Assessing data

You have some conclusions you would like to draw, but are you certain about them?

What could you do to assess this?

Exercise one: Assessing data

These are the official fastest times for each year for the 100m



- What can you tell about 100m times from the data?
- What conclusions would you draw?
- Are the conclusions different looking at the Olympic data?

Data Analysis and Biology (or geography or chemistry or medicine or...)

When we collect data (or decide to collect data), what are we going to do with it?

What does the data tell us?

A Quick Overview of R

A statistics programme

The language (S) designed for statistics

Object oriented

Interpreted language

Basic Syntax: Assignment

```
x <- 2
x
```

```
## [1] 2
```

```
(y <- 1:3)
```

```
## [1] 1 2 3
```

```
(z <- c(4.5,6.7,-3))
```

```
## [1]  4.5  6.7 -3.0
```

A Calculator

```
x + y
```

```
## [1] 3 4 5
```

```
exp(z)
```

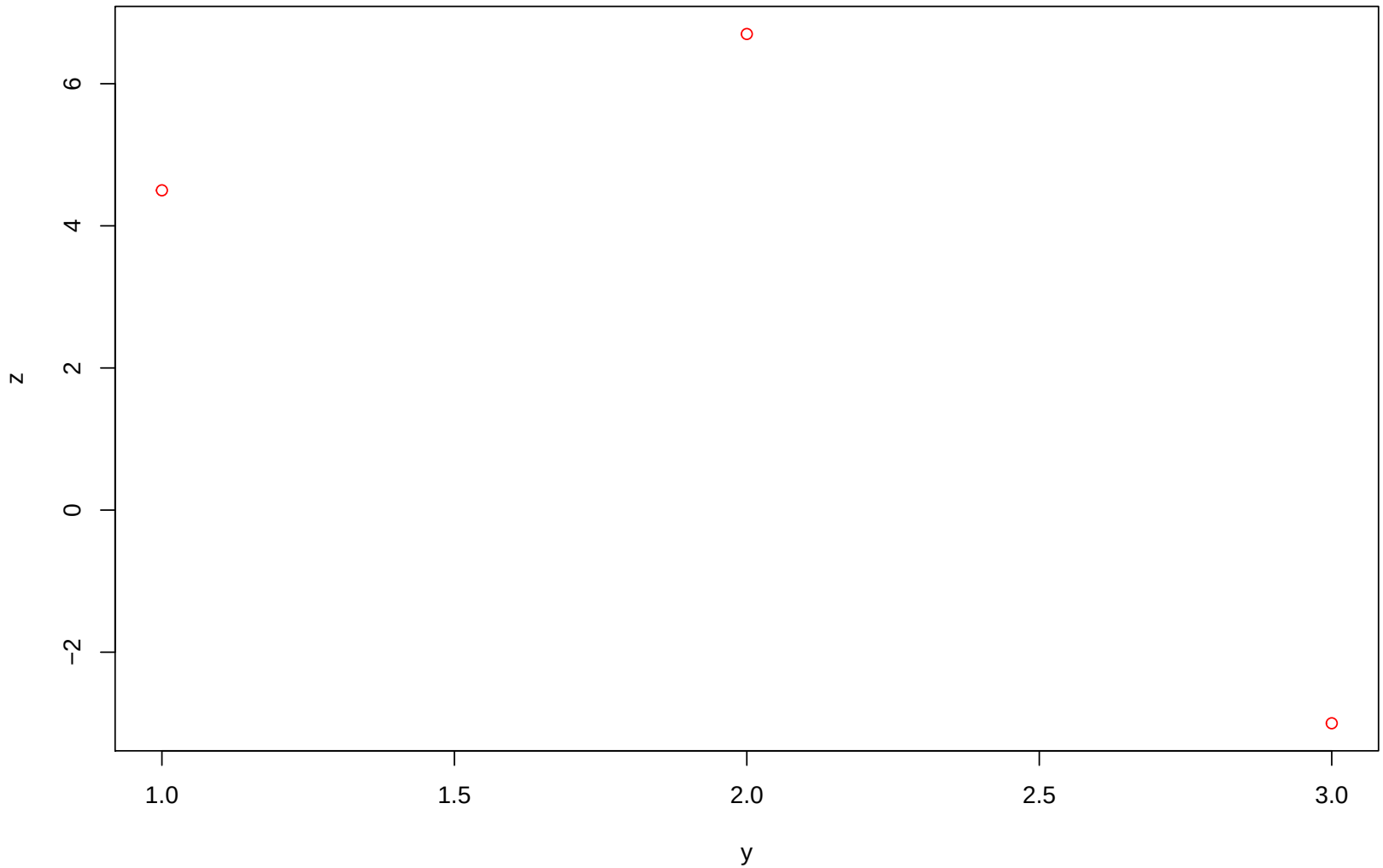
```
## [1]  90.01713130 812.40582517   0.04978707
```

```
z/y # = z[1]/y[1] z[2]/y[2] z[3]/y[3]
```

```
## [1]  4.50  3.35 -1.00
```

Simple Plotting

```
plot(y, z, col=2)
```



Functions & Objects

exp(z) & plot(y, z) are functions

y & z are *arguments*: they are given to the function, which does something with them

The function returns an object: - exp() returns a vector - plot() returns NULL (i.e. nothing: the plot is a side-effect)

Data Types

Our basic data can be of different types, e.g. integer, real number, text, factor, logical

```
(txt <- c("thing", "stuff", "thing")) # text
```

```
## [1] "thing" "stuff" "thing"
```

```
(f <- factor(c("thing", "stuff", "thing"))) # factor
```

```
## [1] thing stuff thing
## Levels: stuff thing
```

```
(lg <- c(TRUE, FALSE)) # logical
```

```
## [1]  TRUE FALSE
```

Functions

We can write our own functions. . . .

```
HelloWorld <- function(str) {
  if(!is.character((str)))
    stop("str should be a character string you idiot")
  cat(paste0("Hello World, ", str, "!\n"))
}
HelloWorld("Orpheus")
```

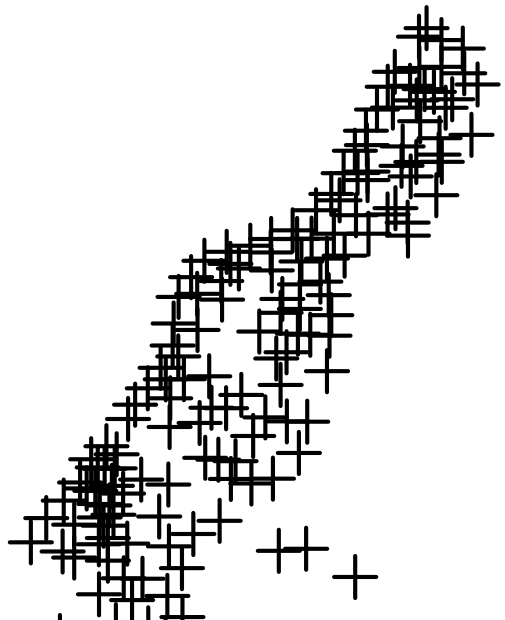
```
## Hello World, Orpheus!
```

Packages

Packages put together a lot of functions (and objects etc.)

A lot of these are on CRAN

```
library(sp)
data(meuse)
coordinates(meuse) <- c("x", "y")
par(mar=c(2,2,1,1))
plot(meuse)
```



More complex Stuff

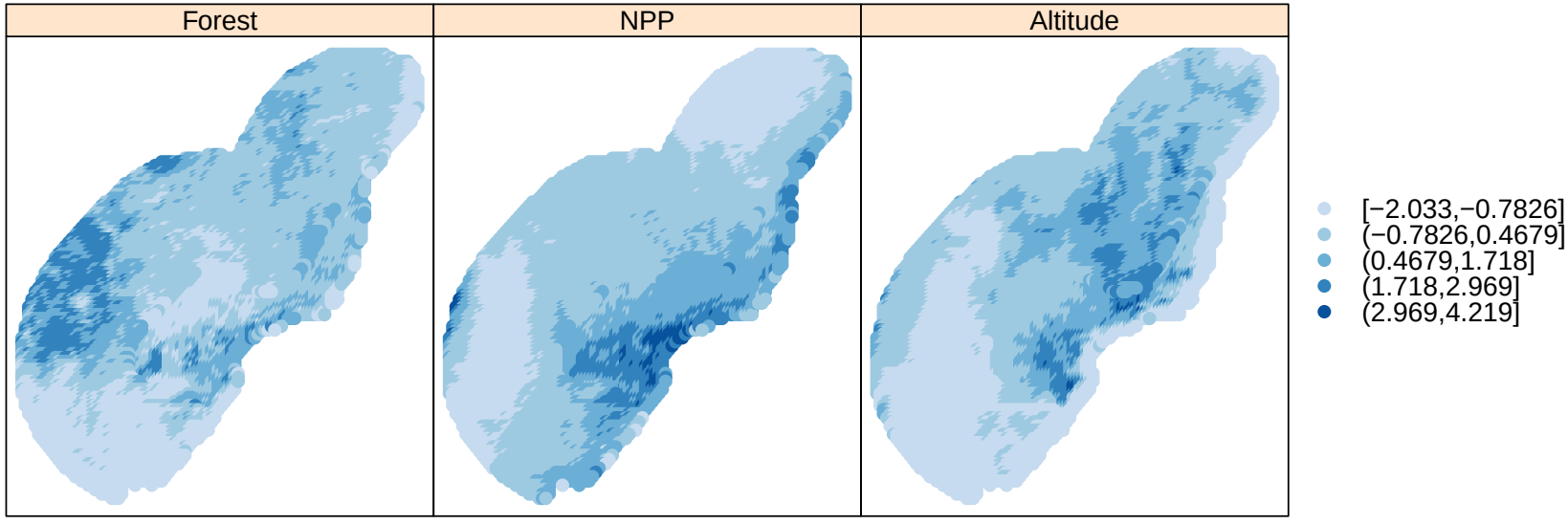
Real data analysis is messy: her is something really messy

```
library(PointedSDMs)
library(mapview)
library(sp)
library(spatstat)
library(RColorBrewer)
data("SolTin_covariates")
Projection <- CRS("+proj=longlat +ellps=WGS84")

data("SolTin_ebird")
ebird <- SpatialPoints(SolTin_ebird[,c("X","Y")], proj4stri
data("SolTin_gbif")
gbif <- SpatialPoints(SolTin_gbif[,c("X","Y")], proj4string
data("SolTin_parks")
Parks <- SpatialPointsDataFrame(SolTin_parks[,c("X","Y")],
data("SolTin_range")
Pgon.range <- Polygons(list(region=Polygon(coords=SolTin_ra
```

Plot the Environment

```
spplot(Covariates, layout=c(3,1), col.regions=brewer.pal(6,
```



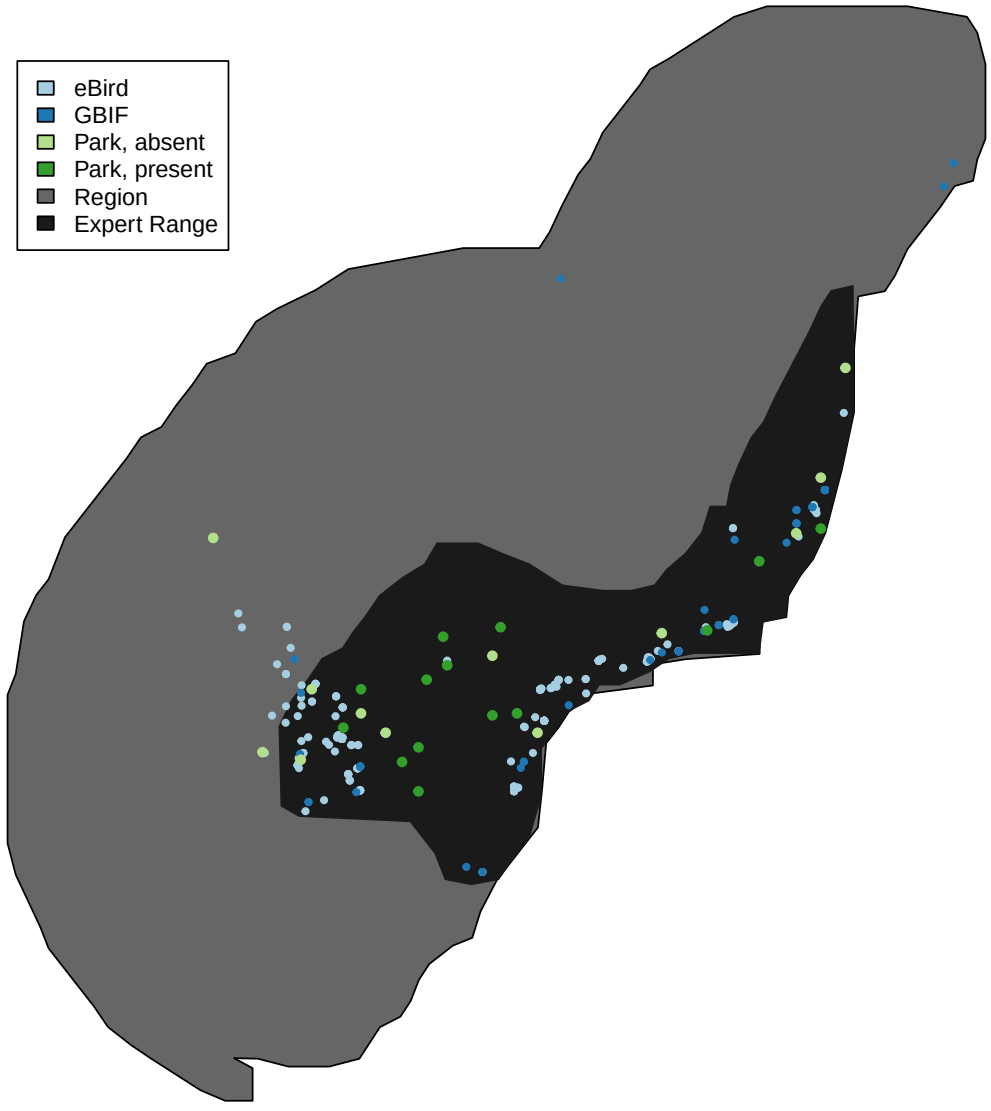
More complex Stuff

We have several data sets we want to analyse together, so there is a lot of formatting

```
region.mask=as.owin(cbind(SolTin_covariates[,c("X","Y")], I
region.mask$m[is.na(region.mask$m)] <- FALSE
Region.poly <- simplify.owin(as.polygonal(region.mask), dmi
PolyPoints <- cbind(Region.poly$bdry[[1]]$x[c(1,length(Regi
Pgon <- Polygons(list(region=Polygon(coords=PolyPoints)), I
region.polygon=SpatialPolygons(list(Pgon), proj4string = Pr

MapCols <- c(brewer.pal(4, "Paired"), grey(c(0.4,0.1)))
names(MapCols) <- c("eBird", "GBIF", "Park, absent", "Park,
```

Plot the data



Fitting a Model

There is more formatting, after which we fit the model & get some output

```
##                    mean     sd
## Forest          -0.0097 0.0250
## NPP             -0.0157 0.0299
## Altitude        -0.0012 0.0260
## int.ebird        1.3817 0.0939
## DistToPoly1      0.1514 0.0525
## int.gbif         1.4042 0.1668
## Intercept       -0.2125 0.1367
## X                0.0020 0.0036
## Y                0.0006 0.0051
## int.parks       -0.2892 0.1814
```

Getting and Using R

You can download R from CRAN: `http://www.r-project.org/`

A lot of people use RStudio: `http://rstudio.org/`

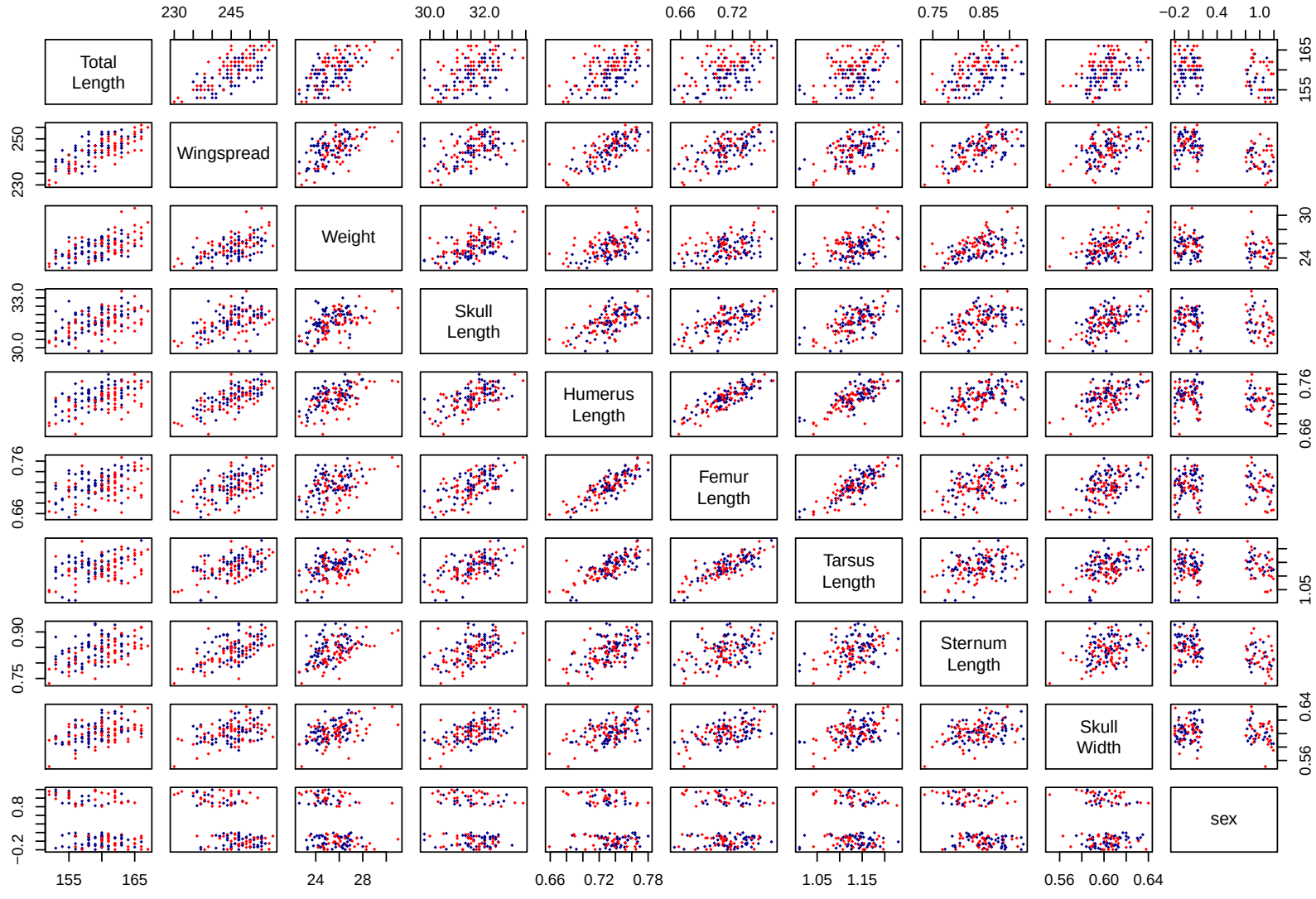
Bumpus' Sparrows

In 1898 Bumpus (or, probably, one of his technicians) collected some sparrows that had been blown out of their trees during a snow storm.

The aim of this was to look at which survived

They measured body size (different aspects), sex and whether the birds survived

Bumpus' Sparrows Data (red=dead)



Bumpus' Sparrows

What sort of biological questions can we ask about the sparrows?

- about survival
- about other aspects