Week Four

An Introduction to Data Ethics

Philip Leftwich

18.10.2021

Data Ethics

[Today we will investigate where our data comes from, and the ethics of collection and use]

Iris flower dataset

R has a number of 'built-in' datasets

- mtcars: Motor Trend Car Road Tests
- ToothGrowth
- PlantGrowth
- USArrests
- iris

```
head(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
```

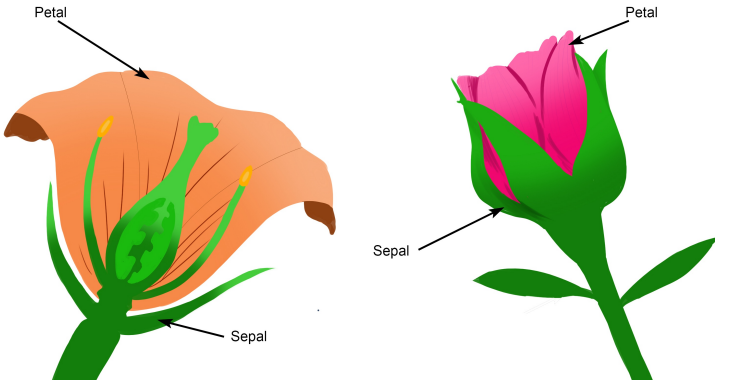
The Iris dataset

The iris dataset includes four continuous variables (measurements):

- Sepal length
- Sepal width
- Petal length
- Petal width

And one categorical variable:

- Species



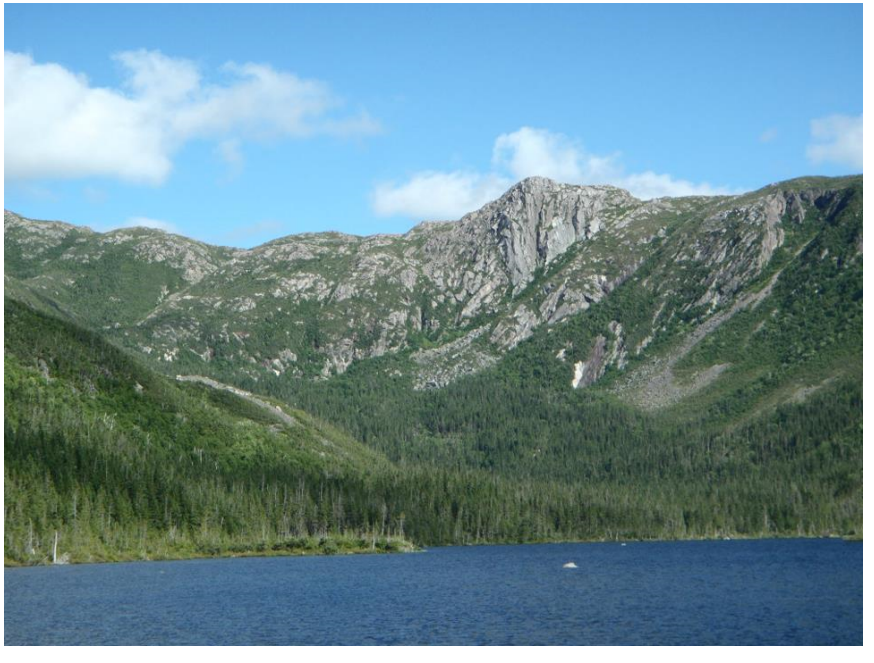
Data science concepts

```
iris %>%
 ggplot(aes(x=Petal.Length,
        y=Petal.Width,
        color=Species))+
 geom_point(aes(shape=Species))+
 theme_void()
```

- Machine learning
- Linear discriminant analysis

Where does the data come from?

Collected by the botanist Dr. Edgar Anderson, who collected most of the data from the Gaspé Peninsula, in Canada



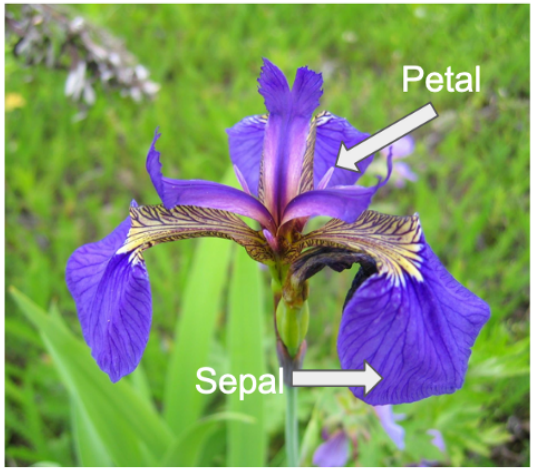
Where does the data come from?

Collected by the botanist Dr. Edgar Anderson, who collected most of the data from the Gaspé Peninsula, in Canada

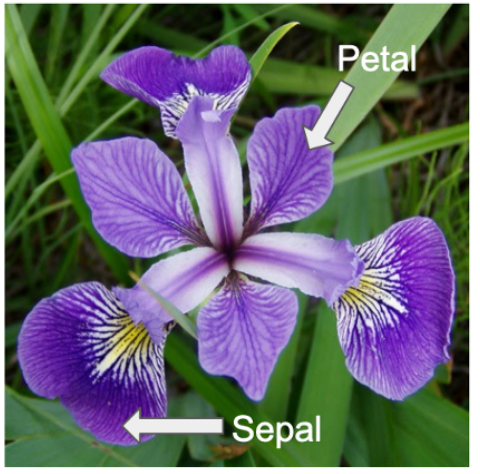
Where does the data come from?

Most of the data collected on a single day in 1935

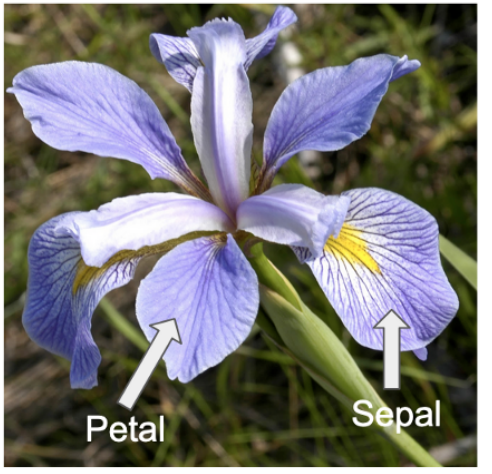
Iris setosa



Iris versicolor



Iris virginica



Ronald Fisher (1890-1962)

Edgar had accepted a fellowship in 1929 to work in Britain with several other scientists, among them Sir Ronald Fisher

Through this collaboration Fisher obtained permission to use the data in his own research paper

Who was Fisher?

"Single-handedly created the foundations for modern statistical science"
Hald (1988)

Statistics

- Fisher's exact Test
- ANOVA
- Student's t-distribution
- Null hypothesis testing

Genetics

- Gene linkage
- Fisherian selection
- Reproductive value
- Mimicry
- Heterozygotic advantage

Publication

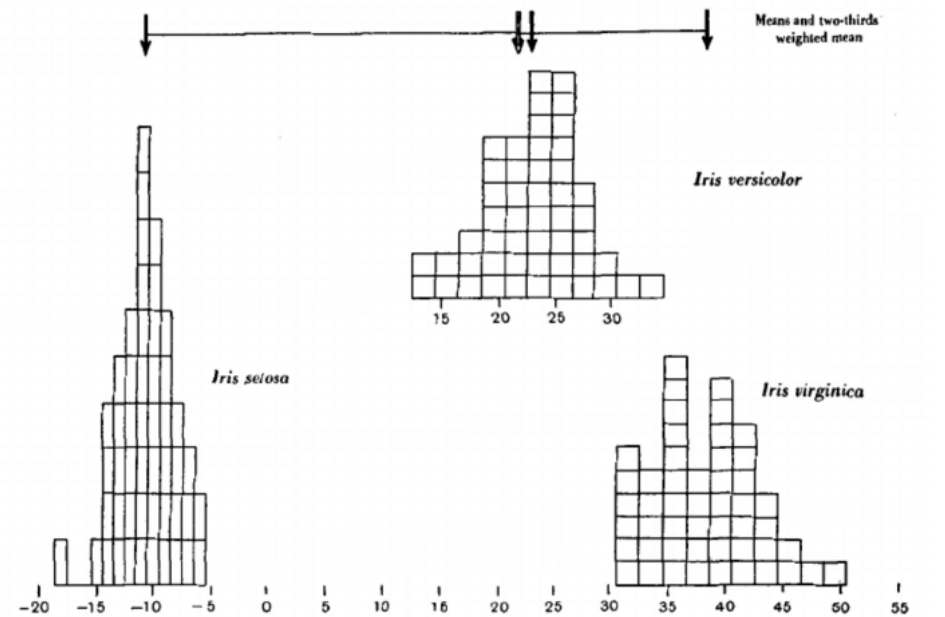


Fig. 1. Frequency histograms of the discriminating linear function, for three species of *Iris*.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179–188.

Eugenics

Fisher was an unrepentant Eugenicist

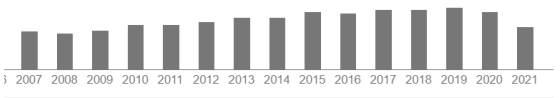
- 1911 -Founding chairman of the University of Cambridge Eugenics Society
- Attributed the fall of civilisations to a reduction in fertility of the upper classes
- Genes over environment on many social issues e.g. smoking
- Supported known associates of the Nazi party before and after WWII

Understand the origin of your data

Ronald A. Fisher

The use of multiple measurements in taxonomic problems

Authors	Ronald A Fisher
Publication date	1936/9
Journal	Annals of eugenics
Volume	7
Issue	2
Pages	179-188
Publisher	Blackwell Publishing Ltd
Description	The articles published by the Annals of Eugenics (1925–1954) have been made available online as an historical archive intended for scholarly use. The work of eugenicists was often pervaded by prejudice against racial, ethnic and disabled groups. The online publication of this material for scholarly research purposes is not an endorsement of those views nor a promotion of eugenics in any way.
Total citations	Cited by 19160



[HTML] AutoML: A Survey of the State-of-the-Art
X He, K Zhao, X Chu - Knowledge-Based Systems, 2021 - Elsevier
Deep learning (DL) techniques have obtained remarkable achievements on various tasks, such as image recognition, object detection, and language modeling. However, building a high-quality DL system for a specific task highly relies on human expertise, hindering its …
☆ Cited by 278 Related articles All 6 versions

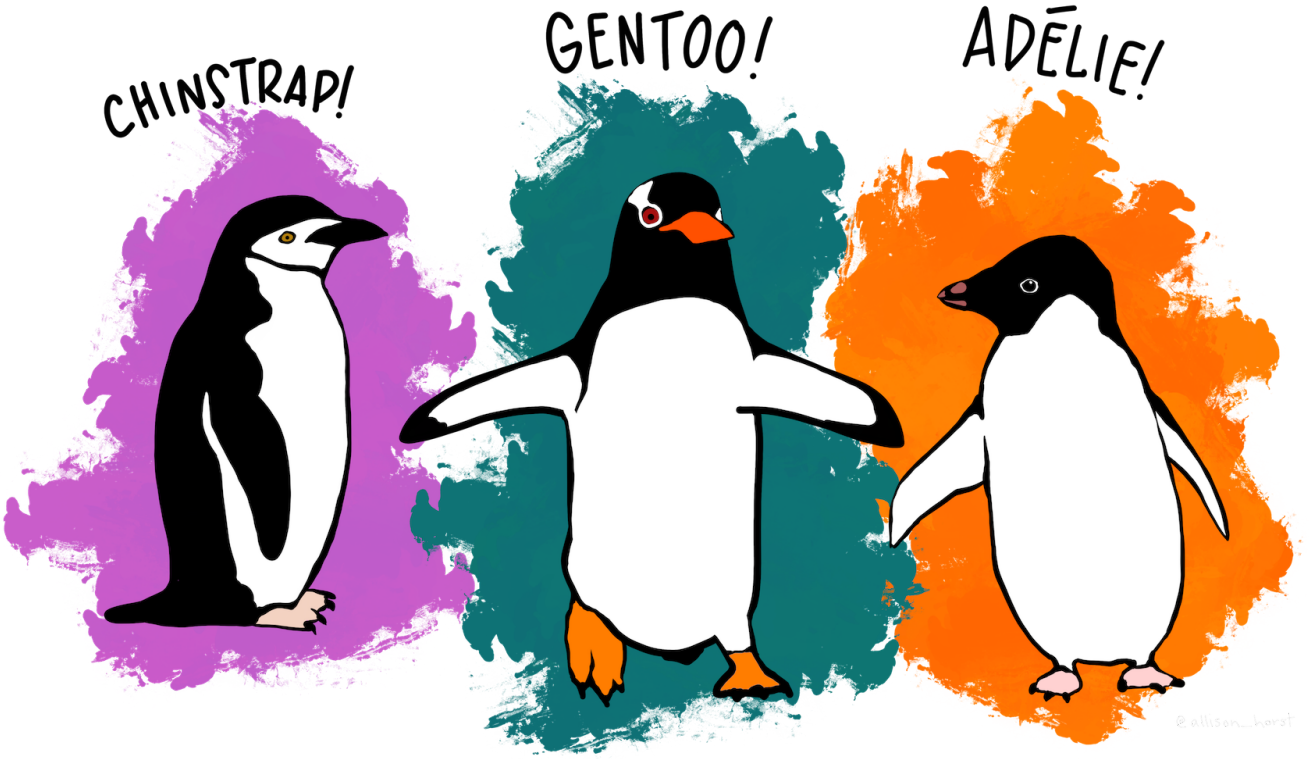
Reservoir computing with biocompatible organic electrochemical networks for brain-inspired biosignal classification
M Cucchi, C Gruener, L Petrauskas, P Steiner… - Science …, 2021 - science.org
Early detection of malign patterns in patients' biological signals can save millions of lives. Despite the steady improvement of artificial intelligence–based techniques, the practical clinical application of these methods is mostly constrained to an offline evaluation of the …
☆ Cited by 1 All 8 versions

[HTML] Deep learning in mining biological data
M Mahmud, MS Kaiser, TM McGinnity, A Hussain - Cognitive Computation, 2021 - Springer
Recent technological advancements in data acquisition tools allowed life scientists to acquire multimodal data from different biological application domains. Categorized in three broad types (ie images, signals, and sequences), these data are huge in amount and …
☆ Cited by 115 Related articles All 12 versions

[HTML] Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters
M Bourel, AM Segura, C Crisci, G López… - Water Research, 2021 - Elsevier
Predicting water contamination by statistical models is a useful tool to manage health risk in recreational beaches. Extreme contamination events, ie those exceeding normative are

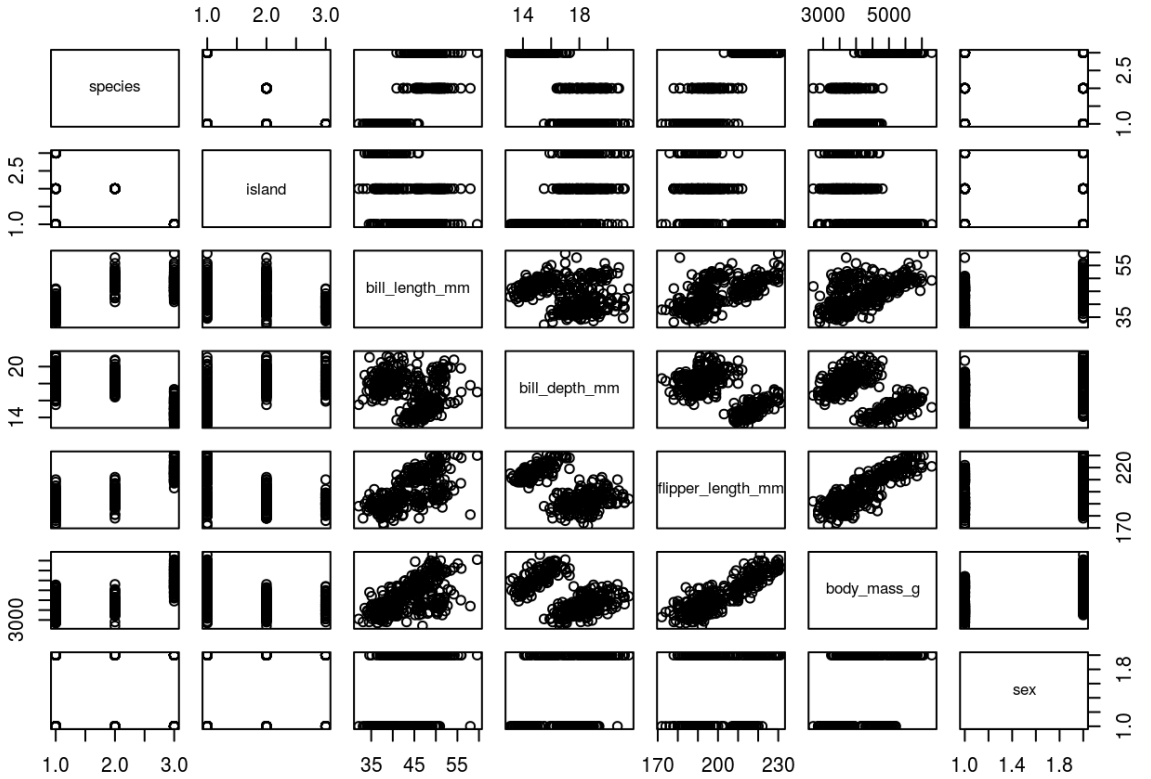
What can we do?

Stop using Iris?



What can we do?

Stop using Iris!



Data Ethics

- All data has a source and background
- Data Reproducibility
- Misrepresentation
- Bias
- Impact

Thank you!

Questions?