Lancaster University

Quantitative Research Methods

Matthew Ivory

matthew.ivory@Lancaster.ac.uk

1. Introduction to the course

- Session 1: Introduction to quantitative research methods using R
- **Session 2: Data management and data wrangling**
- Session 3: Exploratory data analysis
- Session 4: Data visualization
- Session 5: Mid-term assignment
- Session 6: Significance tests for continuous variables
- Session 7: Tests for discrete variables: Analysing contingency tables
- Session 8: Correlation and linear regression. Tests for categorical variables.
- Session 9: ANOVA and tests for N groups
- Session 10: Multiple regression

Session Outline

- 1. Last session refresher
  - The take home tasks
  - The mid-term assignment
- 2. Quantitative Research
  - The PPDAC cycle
  - The data science workflow
- 3. Getting to know Rstudio

1. Refresher – take home tasks

- The solutions to the tasks and all the other R related content in last week's worksheet is available (The worksheet looks the same, but it will have the output underneath each section)

- This table displays the scores of students in two foreign language exams, one administered at the beginning of term, the other at the end of term.

Student ID	Exam 1	Exam 2
Elin	93	98
Spencer	89	96
Crystal	75	94
Arun	52	65
Lina	34	50
Maximilian	50	68
Leyton	46	58
Alexandra	62	77
Valentina	84	95
Lola	68	86
Garfield	74	89
Lucy	51	70
Shania	84	90
Arnold	34	50
Julie	57	67
Michaela	25	37
Nicholas	72	90

1. What are the mean scores for exam 1 and exam 2?

```
mean(language_exams$exam_1)
```

```
[1] 61.76471
```

```
mean(language_exams$exam_2)
```

```
[1] 75.29412
```

2. What is the difference between the two means?

```
mean(language_exams$exam_2) - mean(language_exams$exam_1)
```

```
[1] 13.52941
```

3. What are the mean scores for the two exams if you remove extreme values (the top and bottom 20%) from each?

```
mean(language_exams$exam_1, trim=0.2)
```

```
[1] 62.81818
```

```
mean(language_exams$exam_2, trim=0.2)
```

```
[1] 77.63636
```

4. Based on the previous step (with outliers removed): What is the difference between the two means now? Please round the value before reporting the result.

```
round(mean(language_exams$exam_2, trim=0.2) - mean(language_exams$exam_1, trim=0.2))
```

```
[1] 15
```

5. Can you do steps 3 and 4 in a single command?

```
#the short version
round(mean(language_exams$exam_2, trim=0.2) - mean(language_exams$exam_1, trim=0.2))
```

```
[1] 15
```

Mid-term Assignment

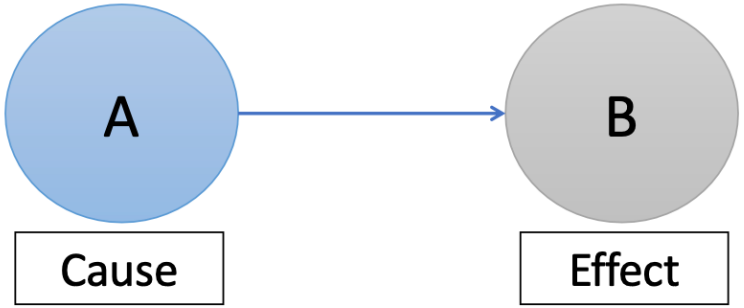
- Information is on Moodle
  - See “Assessment” for the coversheet and submission criteria
  - See “Mid-term assignment, due Monday, Feb 19, 2024, 12pm” for assessment files
- Asked on content learnt in weeks 2, 3, and 4.
  - Reading in data
  - Descriptive analysis
  - Producing plots
  - Successful interpretation and reasoning on information in plots
- 1,500 words: an upper limit only
- Submit your answers to questions (with coversheet) and R script
- R script is not assessed, but I will give feedback on structure/layout etc.

Two basic strategies in quantitative research

1. Observational research
   - Simply observe and describe behaviour in natural situations
   - Example: Systematic reviews, corpus analyses, correlational research, etc.
2. Experimental research
   - Systematically manipulate variables of interest and see what effect our manipulation has in the world
   - Example: Experimental studies in the lab or in the wild

Experimental Research

- In experimental research, we try to discover causal relationships between variables
- We do this by manipulating certain variables to test the outcome on other variables
- We randomly assign participants to groups where they experience different manipulations or their environment
- Example: higher levels of glucose improves cognitive function following an overnight fast, but lower doses are better following shorter fasts (Owen et al., 2011)



The statistical method: PPDAC (MacKay & Oldford, 2000)

- Statistical method
- “Elements and procedures common to all statistical investigations”
- Can be represented as a series of five interconnected stages:

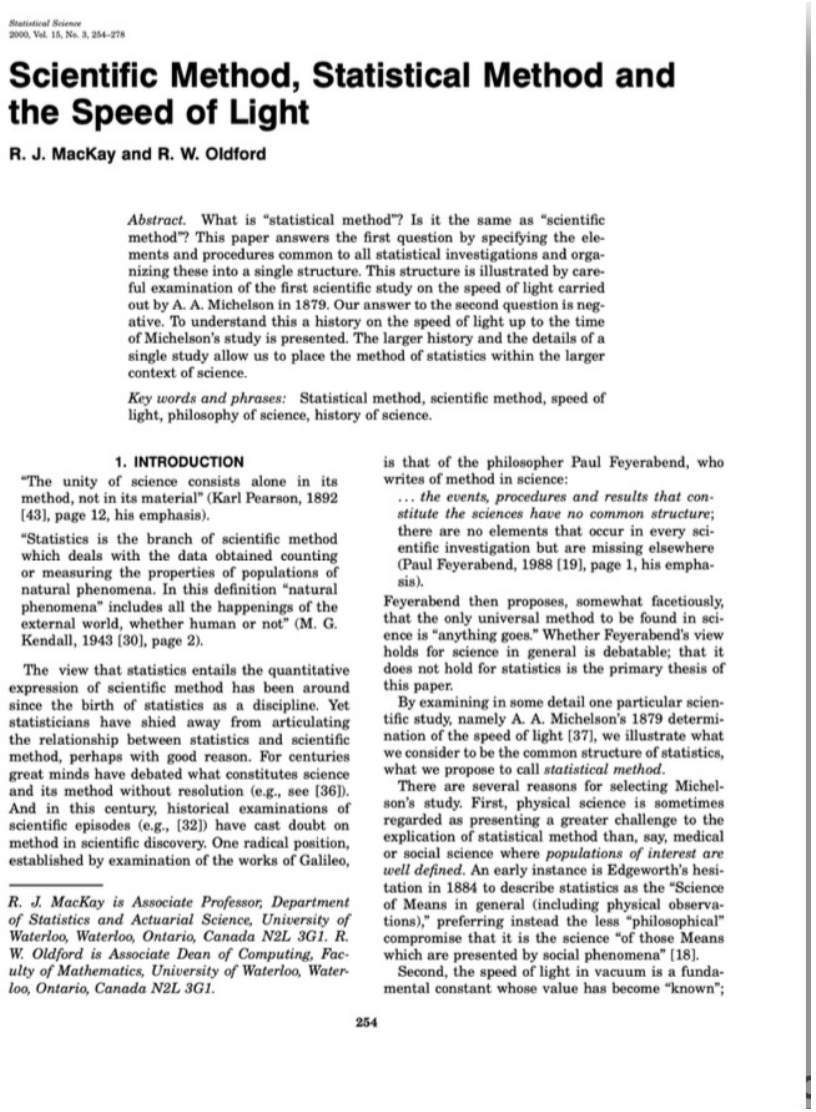
Problem

Plan

Data

Analysis

Conclusion



Statistical Science
2000, Vol. 15, No. 3, 254–278

Scientific Method, Statistical Method and the Speed of Light

R. J. MacKay and R. W. Oldford

Abstract. What is “statistical method”? Is it the same as “scientific method”? This paper answers the first question by specifying the elements and procedures common to all statistical investigations and organizing these into a single structure. This structure is illustrated by careful examination of the first scientific study on the speed of light carried out by A. A. Michelson in 1879. Our answer to the second question is negative. To understand this a history on the speed of light up to the time of Michelson’s study is presented. The larger history and the details of a single study allow us to place the method of statistics within the larger context of science.

Key words and phrases: Statistical method, scientific method, speed of light, philosophy of science, history of science.

1. INTRODUCTION

“The unity of science consists alone in its method, not in its material” (Karl Pearson, 1892 [43], page 12, his emphasis).

“Statistics is the branch of scientific method which deals with the data obtained counting or measuring the properties of populations of natural phenomena. In this definition “natural phenomena” includes all the happenings of the external world, whether human or not” (M. G. Kendall, 1943 [30], page 2).

The view that statistics entails the quantitative expression of scientific method has been around since the birth of statistics as a discipline. Yet statisticians have shied away from articulating the relationship between statistics and scientific method, perhaps with good reason. For centuries great minds have debated what constitutes science and its method without resolution (e.g., see [36]). And in this century, historical examinations of scientific episodes (e.g., [32]) have cast doubt on method in scientific discovery. One radical position, established by examination of the works of Galileo, is that of the philosopher Paul Feyerabend, who writes of method in science:

> *... the events, procedures and results that constitute the sciences have no common structure*; there are no elements that occur in every scientific investigation but are missing elsewhere (Paul Feyerabend, 1988 [19], page 1, his emphasis).

Feyerabend then proposes, somewhat facetiously, that the only universal method to be found in science is “anything goes.” Whether Feyerabend’s view holds for science in general is debatable; that it does not hold for statistics is the primary thesis of this paper.

By examining in some detail one particular scientific study, namely A. A. Michelson’s 1879 determination of the speed of light [37], we illustrate what we consider to be the common structure of statistics, what we propose to call *statistical method*.

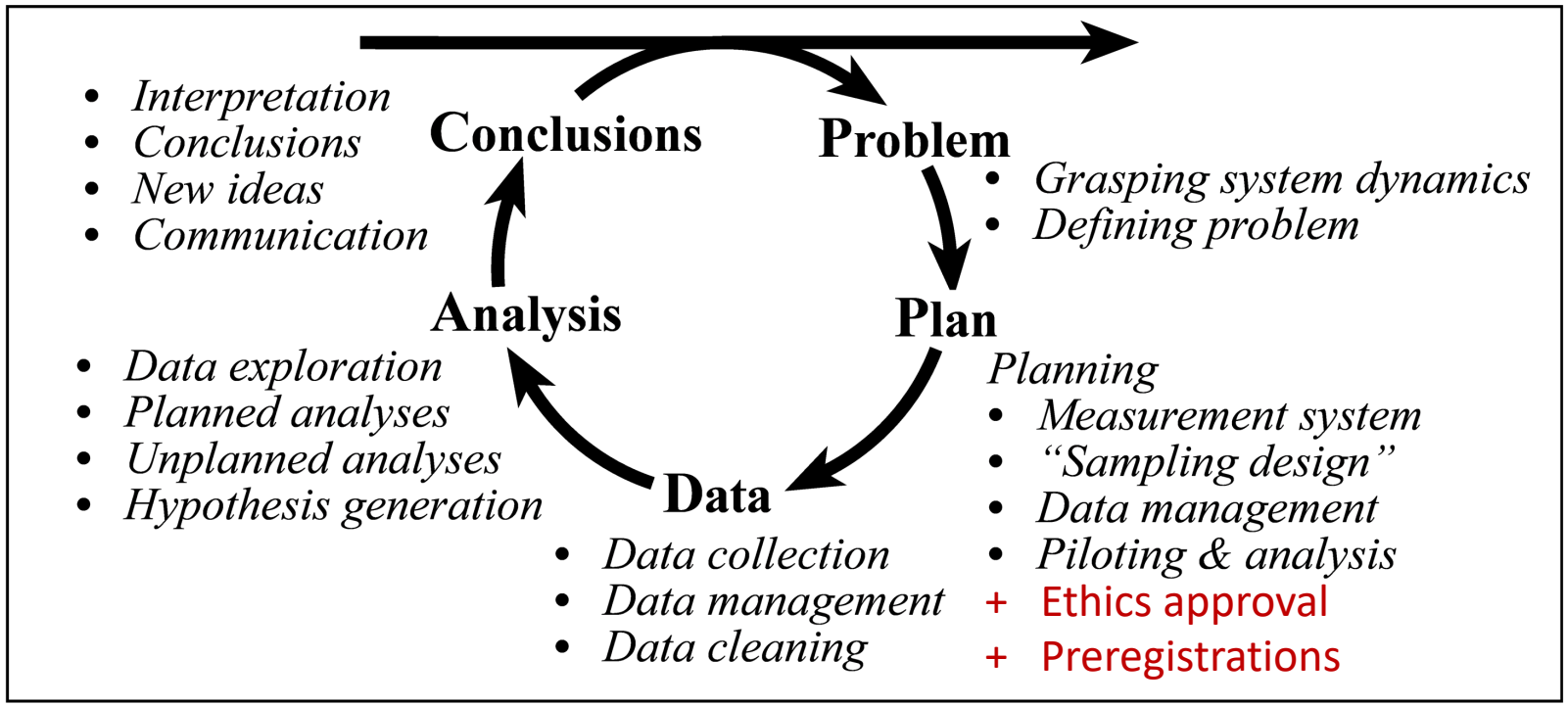
There are several reasons for selecting Michelson’s study. First, physical science is sometimes regarded as presenting a greater challenge to the explication of statistical method than, say, medical or social science where *populations of interest are well defined*. An early instance is Edgeworth’s hesitation in 1884 to describe statistics as the “Science of Means in general (including physical observations),” preferring instead the less “philosophical” compromise that it is the science “of those Means which are presented by social phenomena” [18].

Second, the speed of light in vacuum is a fundamental constant whose value has become “known”;

R. J. MacKay is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. R. W. Oldford is Associate Dean of Computing, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

(a) DIMENSION 1 : THE INVESTIGATIVE CYCLE

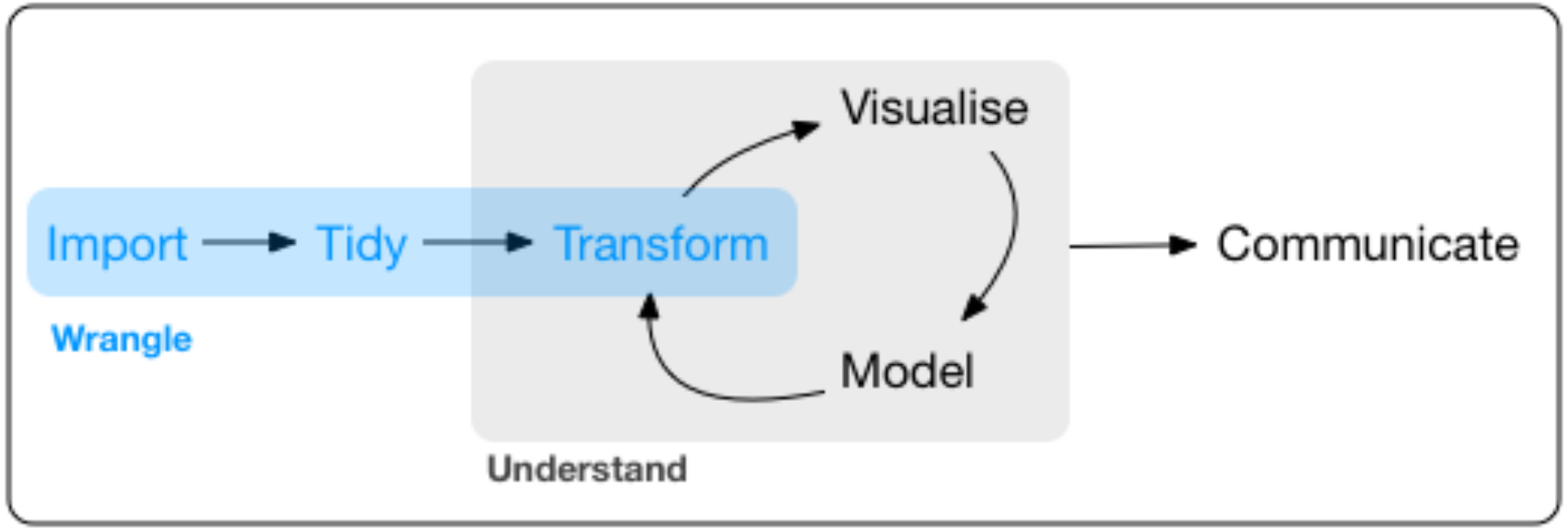
(PPDAC)



Chris Wild

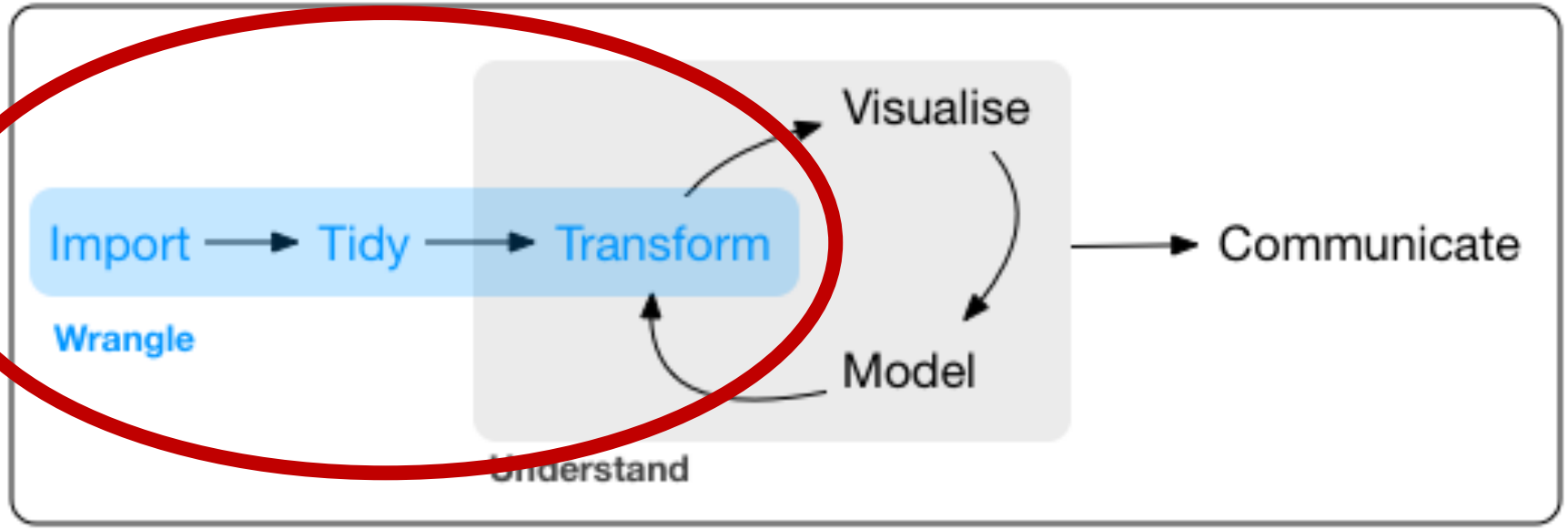
The Data Science workflow

- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.



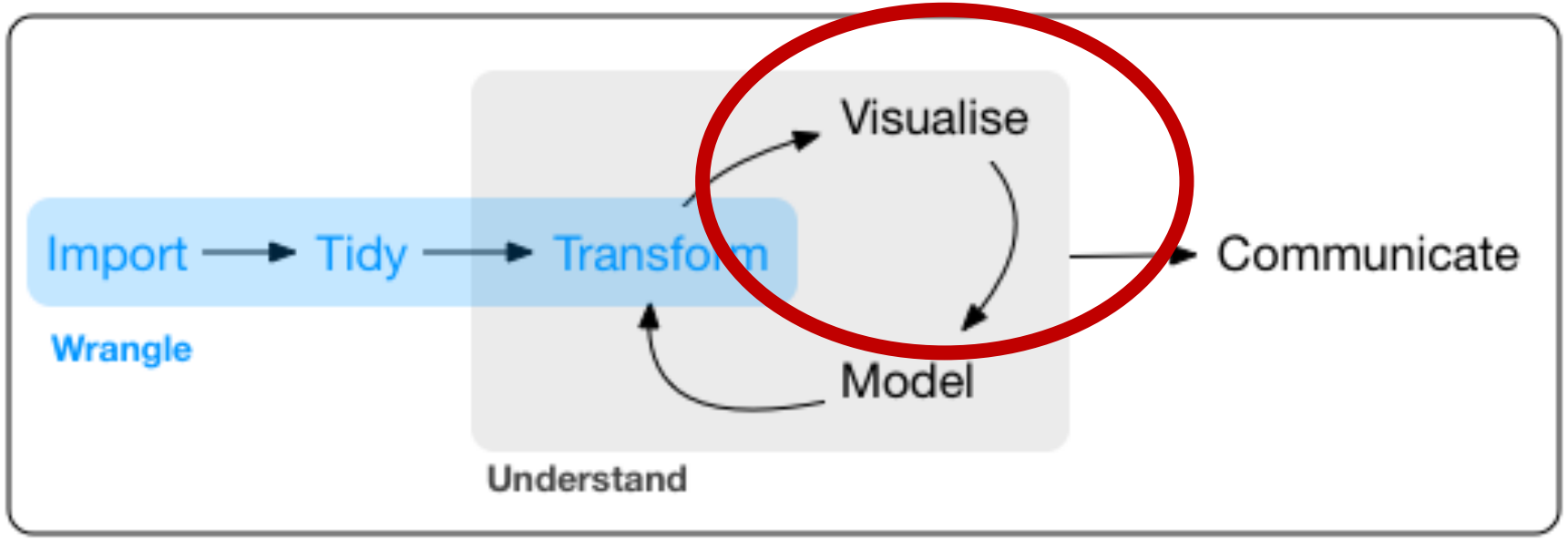
Step 1: Wrangling

- Data can be messy (e.g., there might be missing values). Before analyzing the data, we need to ”tidy” it.
- Data wrangling: The process of preparing raw data for further analysis.



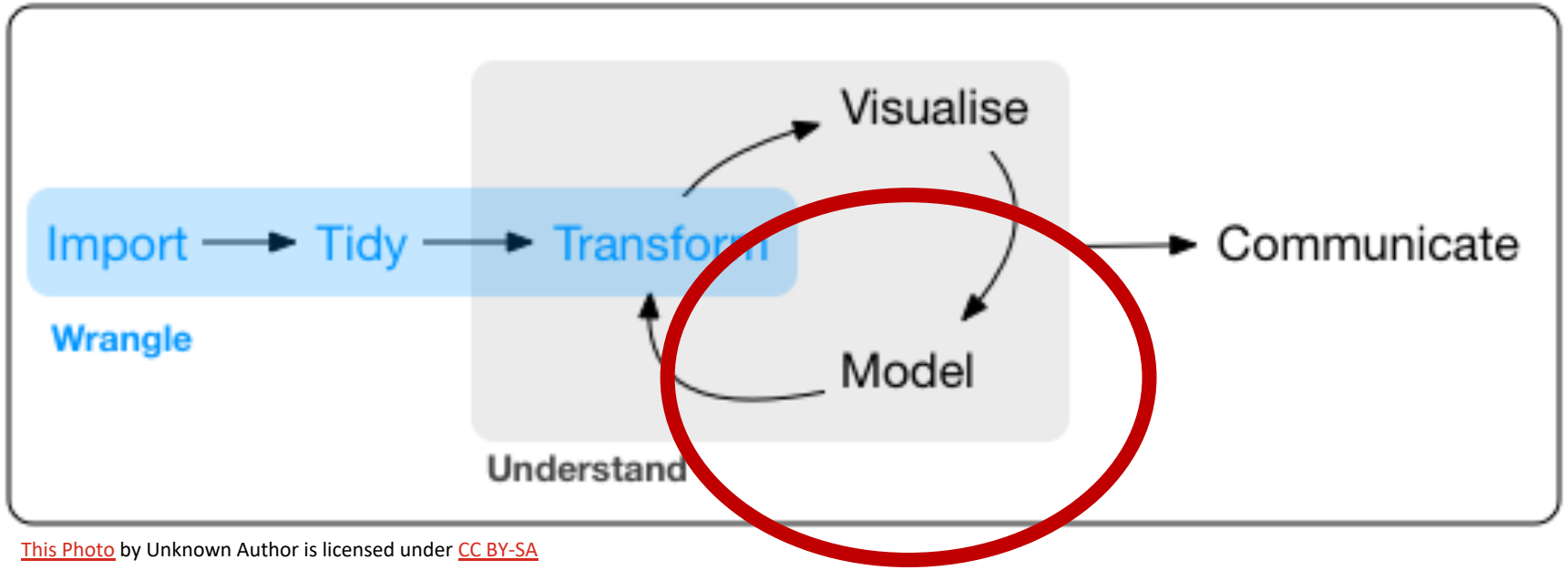
Step 2: Visualise

- Once data is tidy, we use computational tools and statistical methods for data exploration and visualization.
- The aim is to discover potentially interesting patterns and behaviours in the data.
- We will focus on this in sessions 3 and 4.



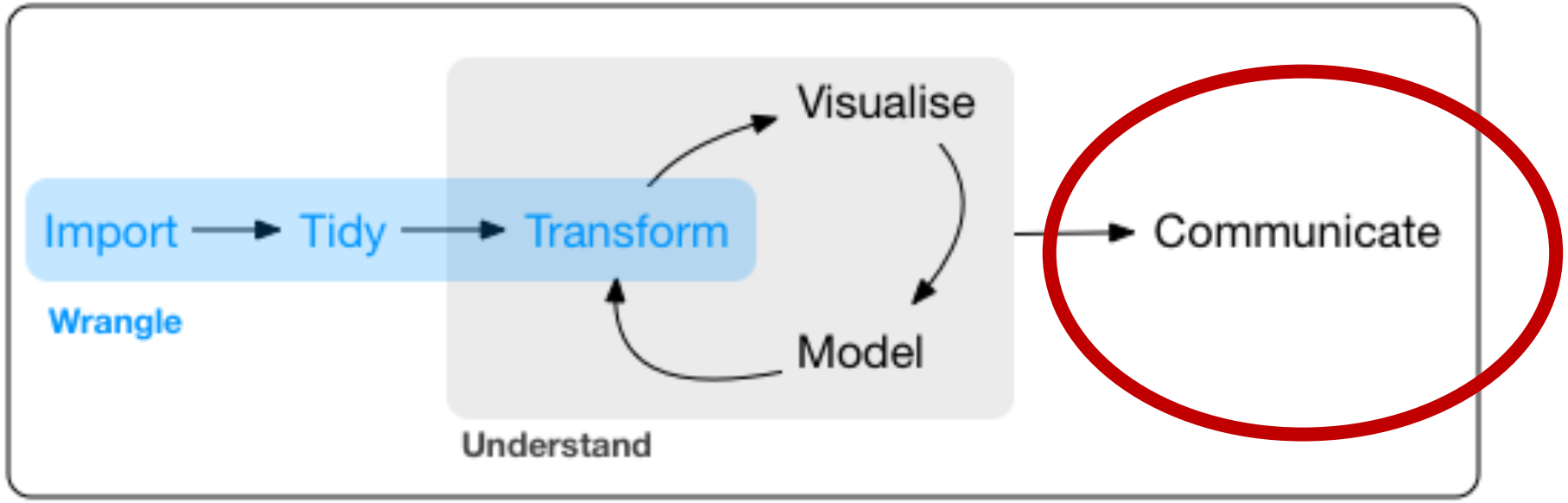
Step 3: Model

- The visualisation and exploratory analysis (step 2) leads us to a probabilistic model of the data.
- This is a model of the phenomenon that generated the data.
- Modelling involves statistical inference and model evaluation.



Step 4: Communicate

- We disseminate our results (e.g., presentations, publications, corpora).
- • But: Science should be **open, transparent and reproducible**, so we should also share materials, data and scripts via freely available platforms (e.g., OSF).
- • This way, others can replicate our studies and build on our research.



Getting to know RStudio

- In last week's worksheet, I said two things
  - "We will discuss how to import files next week", and
  - **"I don't like** nesting functions and neither should you"
- The first comment we will look at in today's worksheet,
- The second comment I will address now, as it will be used implicitly throughout the rest of the course

Nested functions are bad

- Why?
  - They are hard to read and quickly become very dense

```
select(filter(mutate(rowwise(data_GDMS), mean = mean(c(Rational:Spontaneous)),
                     SD = sd(c(GDMS_1_Rational:GDMS_25_Spontaneous))),
              mean == 1 | mean == 2 | mean == 3 | mean == 4 | mean == 5),
       mean, SD)
```

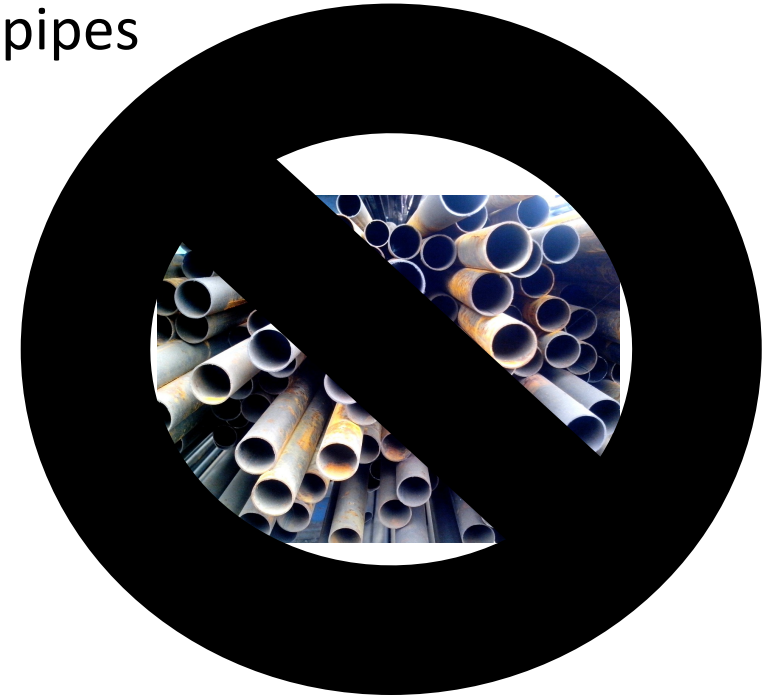
- How do we fix this?
  - Using tidy functions and pipes

Packages and Libraries

- In R, we are only limited in terms of what we know and what our computer can handle
- R comes with certain functions pre-packaged (often called base R)
  - Like mean() or sd()
- We can add in additional functions through packages or libraries (which are collections of packages)
- A package is a collection of functions that are bundled together, these can be downloaded from CRAN and used locally on your computer

The Tidyverse library

- Tidyverse is a collection of packages that are aligned through their approach to data science and share an underlying design philosophy, grammar, and data structures
- It includes many functions that we will be using throughout the course
- One thing that comes with tidyverse and is a large part of writing in the “tidy style” are pipes



%>% or **|>**

Pipes Primer

```
iris %>%
  arrange(Species)

arrange(iris, Species)
```

Lancaster University

- Tidy pipes allow us to format our code so that it runs across multiple lines and enhances readability
- Use %>% to emphasise a sequence of actions, rather than the object that the actions are being performed on.
- Pipes take the data object (such as a dataframe) and 'pipes' it into the RHS function
  - Dataframe %>% function() %>% function %>%...

- Pipes replace nested functions and are easier to follow and read – they are also easier to troubleshoot
- Do not worry too much about pipes right now, they **will** be introduced, and you will get a chance to get familiar with them

This week's worksheet

1. Scripts
2. Installing and loading packages
3. Working directories and clean workspaces
4. Loading data (various formats)
5. Examining datasets
6. Closing your R session

Questions?

- I will be walking around while you work through the worksheet