Thinking as a "Data Scientist"

Saghir Bashir
www.ilustat.com

Outline

Questions to Decisions

Data Processing

Analysis

Communication

Summary

Objectives

My objectives are to encourage you to:

> Adapt good working practices.

> Challenge your thinking.

> **Build trust in your work.**

> Enjoy your work.

What about R?

This presentation applies to data science independent of the software you use.

> I will give examples and references from R.

Questions to Decisions

Weather Example

Questions

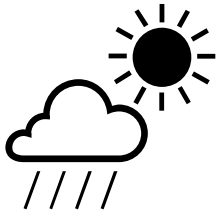
Will it rain today?

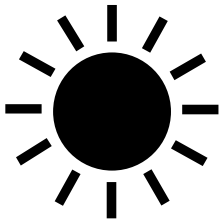
Decisions

Take an umbrella.

Don’t take an umbrella.

Don’t go out.

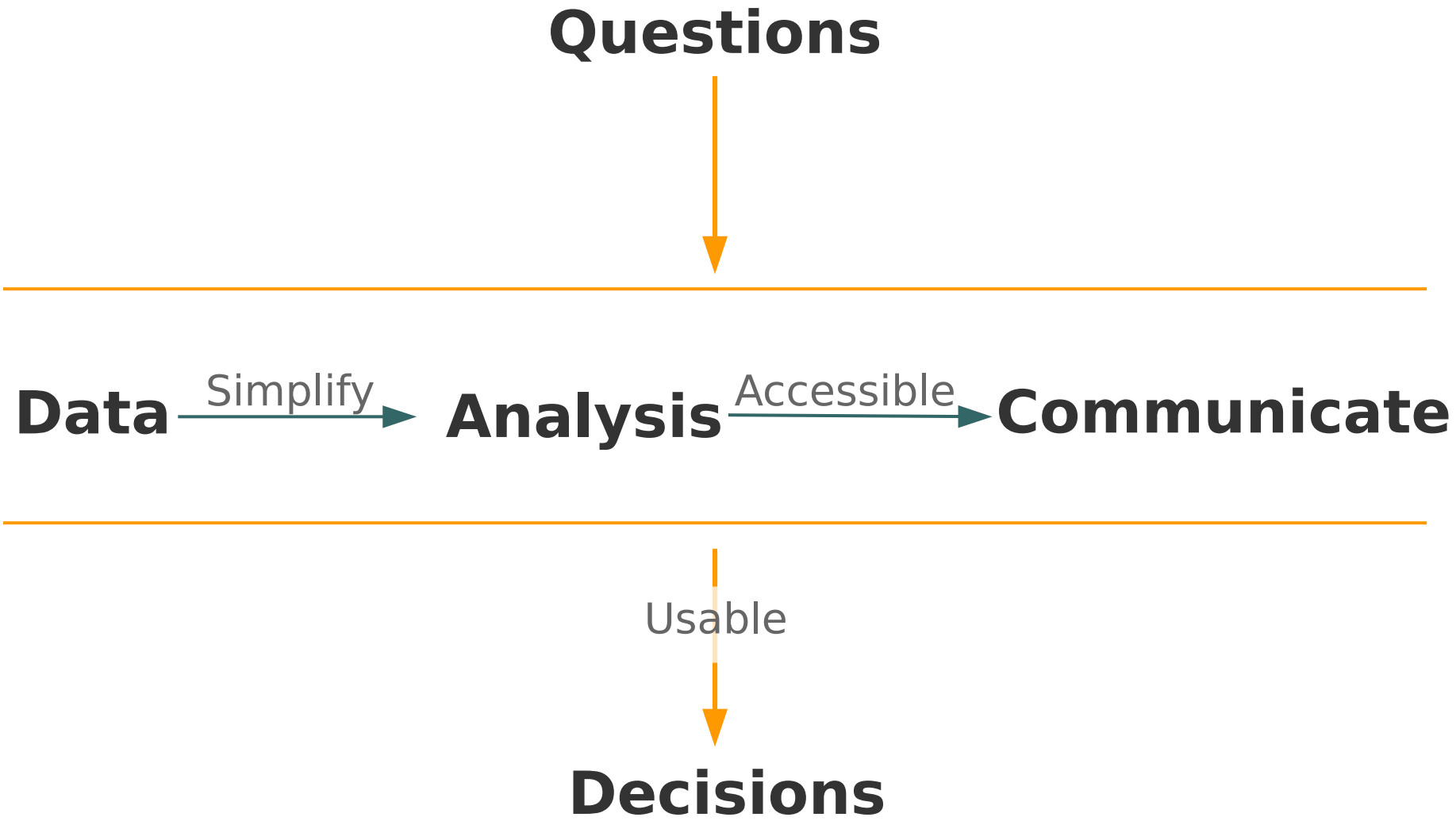




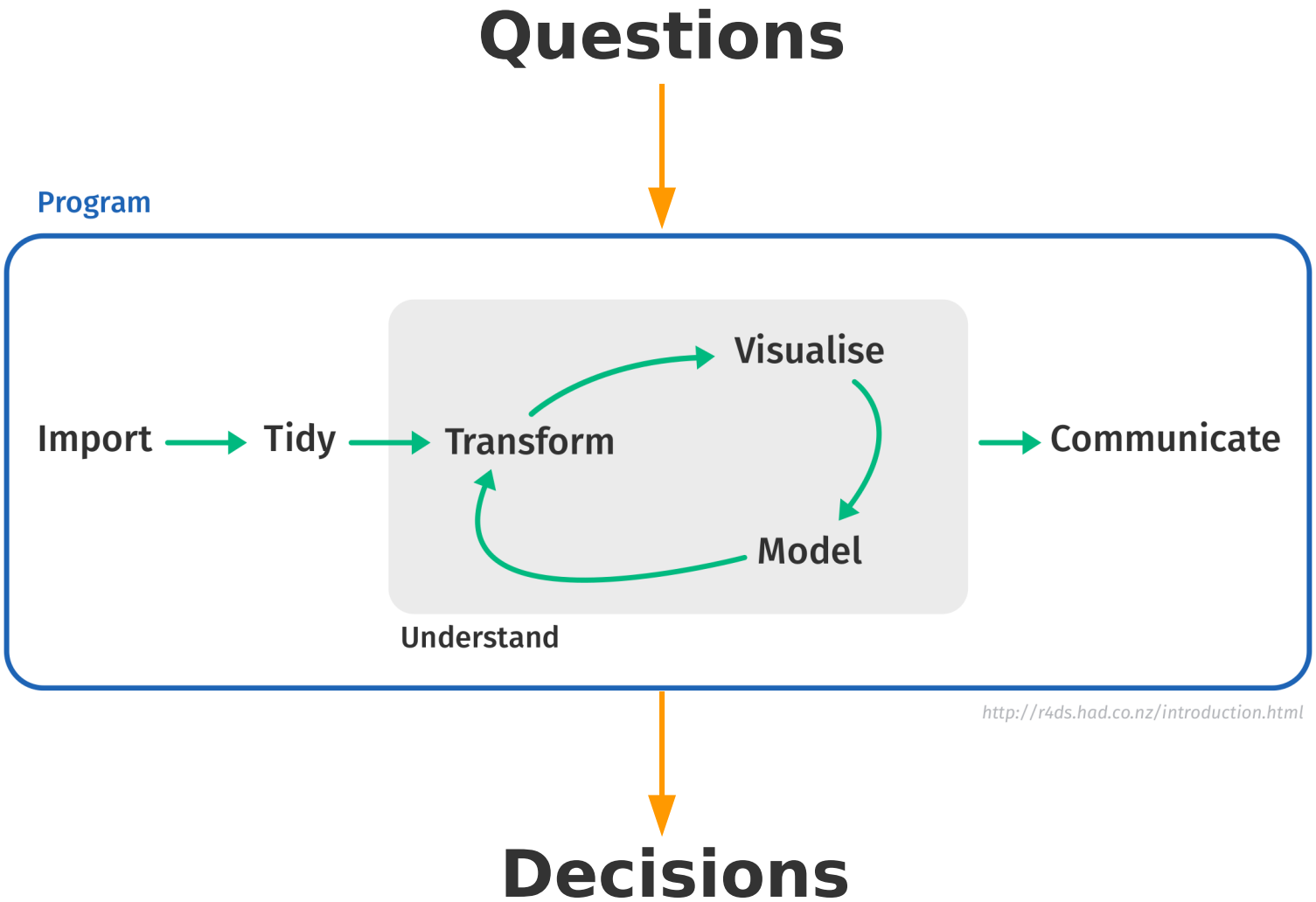
Weather Example

Yesterday	08:00	Forecast	Decision
⛈	☁	🌧	☂
🌥	⛈	🌧	☂
🌦	☀	🌧	☂

“Data Science” Thinking



“Data Science” Doing



Definition: “Data Science”

Generally accepted definition

Does not exist.

This is a discussion for another day.

Presentation definition

“Using **data, statistics** and **programming,** in a given **context,** to support **decision** making.”

Questions to Decisions

Define unbiased and clear questions

Will it rain today? / What is the weather forecast today?

Do free gifts increase sales? / What factors impact sales?

Decisions

Understand the **decisions** that could be taken.

Very useful for data science **thinking** and **planning**.

Weather Example

Questions

What is the weather forecast for today?

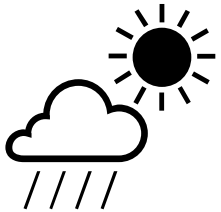
Key interest is in ***going to work and returning.***

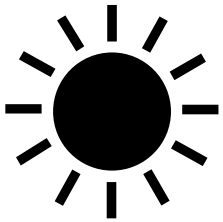
Decisions

Take an umbrella.

Don’t take an umbrella.

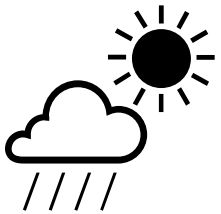
Work from home.

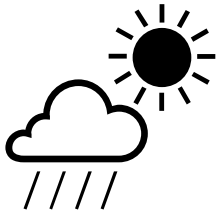


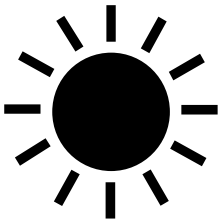


Weather Example – Original

Yesterday	08:00	Forecast	Decision
⛈	☁	🌧	☂
⛅	⛈	🌧	☂
🌦	☀	🌧	☂







Weather Example – Updated

Yesterday	08:00	12:00	18:00	Decision
⛈	☁	🌧	⛅	?
⛅	⛈	🌧	🌦	🏠
🌦	☀	🌧	🌧	☂

Making Decisions

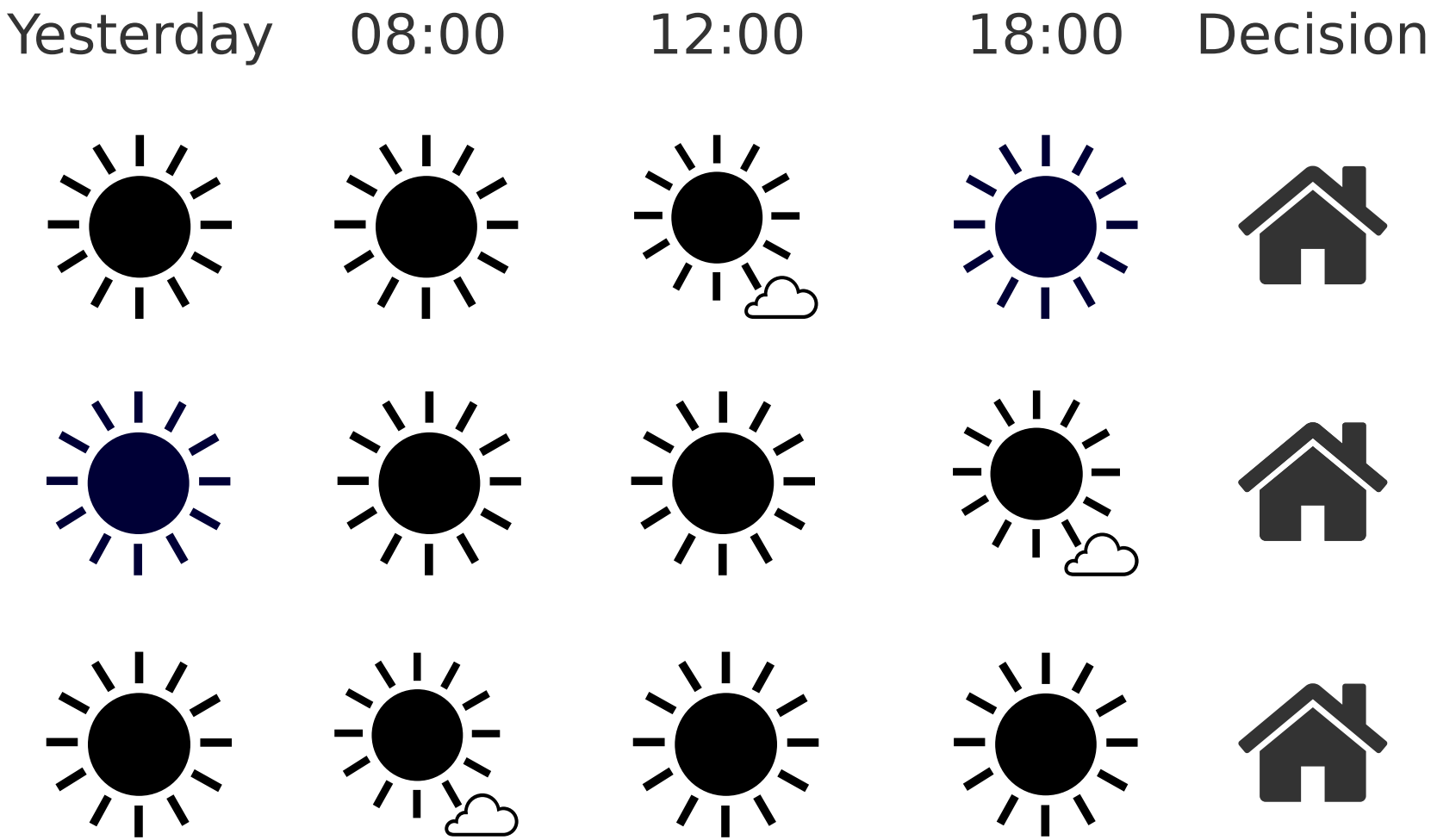
Data Science supports decision making, which involves:

> **Balancing information**
  → Data science is often one part of a bigger picture.

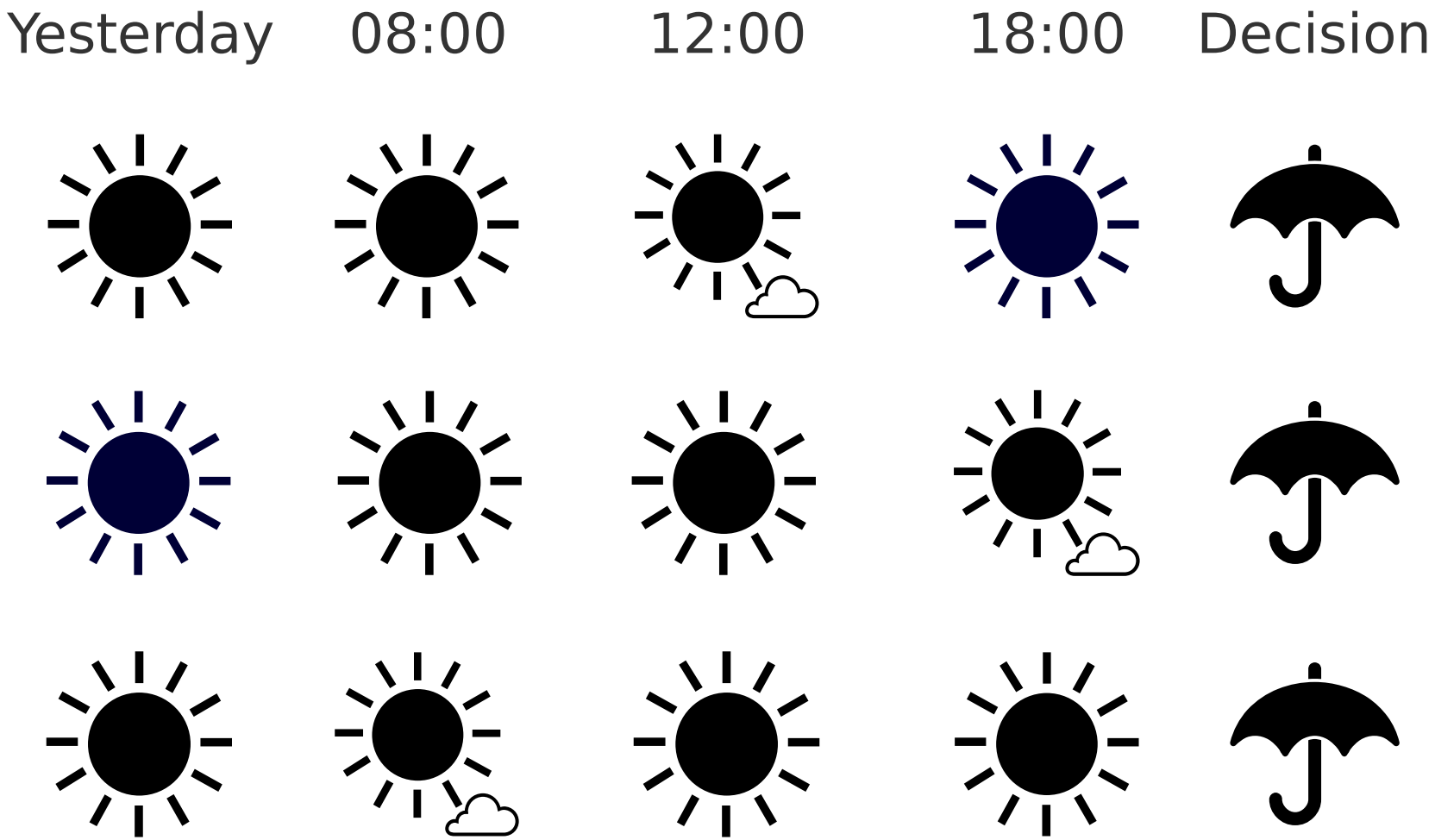
> **Personal experience**
  → Different decisions can be taken using the same information.

> **Risk taking**
  → Varies by person and situation.

Valid Decisions – Skin Sensitivity



Valid Decisions – British



Data Processing

The Data

Key Points

> **Accessibility** – *Format & legal restrictions*

> **Appropriateness & Validity** – *Generalisability*

> **Quality** – *Garbage in, garbage out (GIGO)*

Understand The Data

Before doing analysis or programming, ask:

> **How** and **when** was the data collected?

> Who collected it? Who owns it?

> Was it **quality controlled**? How?

> Are there **confidentiality** or **privacy** issues?

> What information (e.g. variables) do you have?

> **Can the data answer the questions of interest?**

Tidy Data

> **Wrangle your data into tidy data* where:**
  - → Each variable is in a column.
  - → Each observation is a row.
  - → Each value is a cell.

Var 1	Var 2	Var 3
#	#	#
#	#	#
#	#	#

> **Will most likely take a majority of the time.**

> **R makes this easier with tidyverse packages.**
  - → *See www.tidyverse.org

Data Processing in R

> Importing Data

- → **From Files** – readr & readxl
- → **SAS, Stata & SPSS** – haven
- → **Web –** rvest, xml2, httr & jsonlite

> Tidy and Transform

- → **Tidy** – tibble & tidyr
- → **Transform** – dplyr, stringr, lubridate, hms & forcats
- → **Pipes** – Use %>% (magrittr)

Analysis

Analysis Objectives

Your answers should be:

- **Unbiased**
- **Robust**
- **Generalisable**

Analysis

Key Point – Simplify The Data

- **Data Summaries**
- **Visualisation**
- **Modelling**

Basic Statistics and Plots

Start simple

> Understand the **raw** data.

> ***Summary statistics*** are your friends.

> ***Data visualisations*** can teach you a lot.

> *These might be* enough to answer the questions.

> ***Very useful to understand further analysis.***

Modelling

Specify and justify all models fully:

- Data used
- Model variables
- Model equations, formulas and/or algorithms
- Model **ASSUMPTIONS**

This applies to machine learning too!

Health Warning

Modelling (analysis) is STATISTICS!

> The ***laws of gravity*** apply to Data Scientists too!

> You must ***understand*** the models you use.

> *All models have **strengths** and **weaknesses**.*
  - → *Understand them.*
  - → *Be open and transparent about them.*

Useful Quotes

"*Essentially, all models are wrong, but some are useful*"

George E.P. Box (1987)

"*Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.*"

George E.P. Box (1987)

"*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*"

John W. Tukey (1962)

Analysis in R

> **Basic Statistics and Visualisation**

- → **Summary Statistics** – dplyr (summarise)
- → **Visualisation** – ggplot2 & plotly

> **Modelling**

- → **Tidy modelling** – broom & modelr
- → **Statistical models** – lm, glm, anova, nlm, ...
- → **Machine learning** – caret, rpart, randomForest, ...

> **Reproducibility**

- → **Code, results & commentary** – Rmarkdown

Communication

Communication – Key Points

- **Objectives**
  - → Questions
- **Data**
  - → Source, collection methodology (e.g. survey), representativeness, quality and validity
- **Analysis**
  - → Summary statistics/graphs
  - → Analysis – assumptions, methods?
  - → Results – graphical / quantitative
- **Conclusions**
- **Subject matter expert input needed throughout**

Communication

> Understand Your Audience

- → **Need full details** – full report or publication.
- → **Summary details** – article or blog.
- → **Executive summary** – presentation.

> Openess and Transparency

- → Share and link programs, data and full report.
- → Make sure your work is reproducible.

> Communication Style

- → Understandable, relevant and interesting.
- → Keep it simple, clear and concise.

Communication Via R

> **Outputs & Presentations**

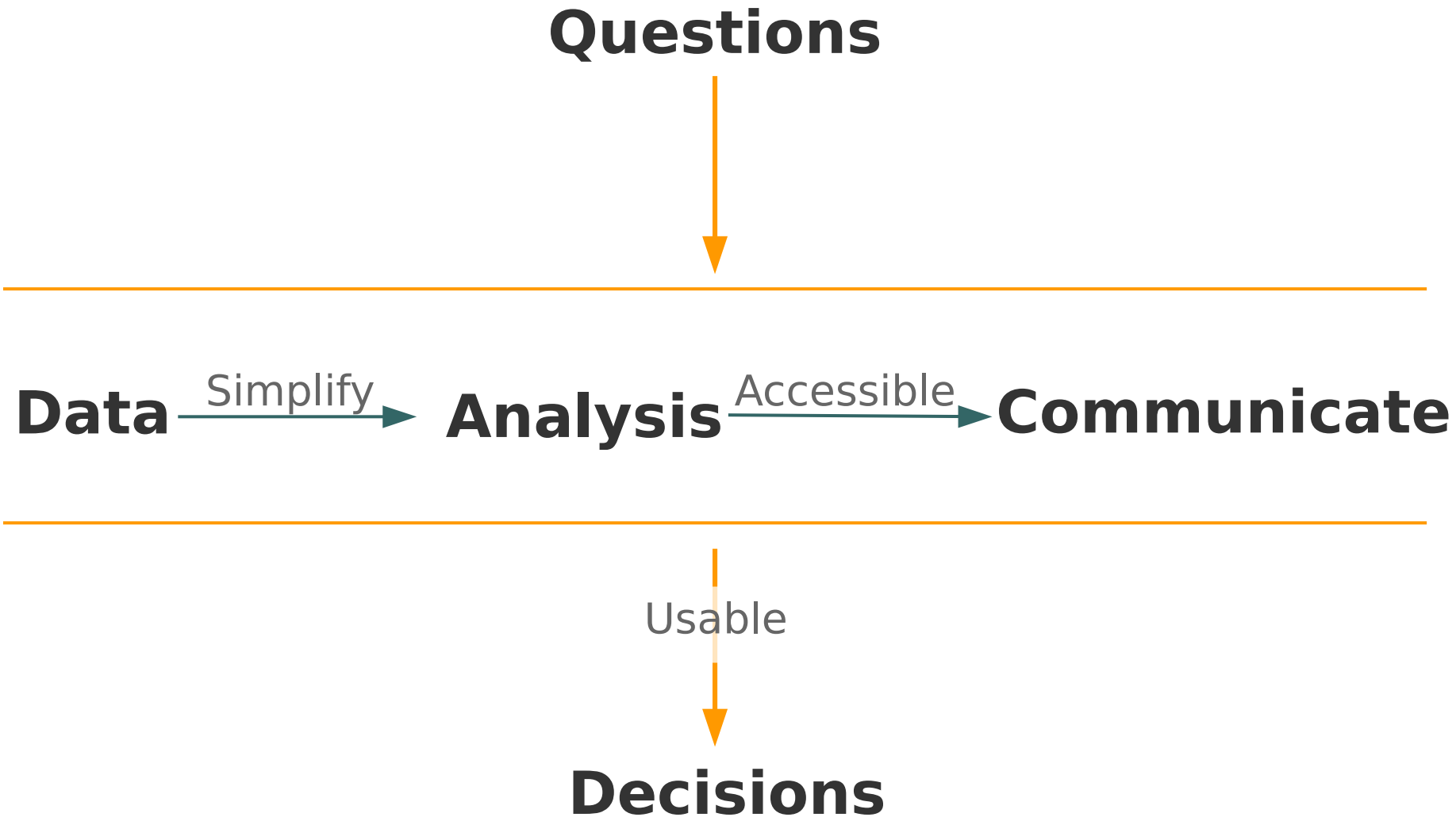
→ **PDF, HTML & DOCX** – Rmarkdown

> **Sharing data and results**

→ **Web applications** – shiny, opencpu & htmlwidgets
→ **Interactive maps** – leaflet & rmaps

Summary

“Data Science” Thinking



Summary

- **Focus on answering the questions with data**
  - Understand the decisions that could be taken.
  - Don’t answer the wrong question.
- **Try to keep everything simple**
  - Easier for you to understand and explain.
  - Communicate clearly and concisely.
  - Make your work reproducible.
- **Work closely with your collaborators**
  - Subject area experts, programmers, statisticians, ...
  - Data Science & R user communities.

References

- *R Project – www.r-project.org*
- *Tidyverse packages – www.tidyverse.org*
- *Hans Rosling's 200 Countries, 200 Years (4 minutes); The Joy of Stats - BBC Four: https://www.youtube.com/watch?v=jbkSRLYSojo*
- *Cambridge Ideas – Professor Risk (6 minutes) https://youtu.be/a1PtQ67urG4*
- *Box, George E. P. & Norman R. Draper* (1987). "Empirical Model-Building and Response Surfaces", Wiley.
- *John W. Tukey* (1962). *"The future of data analysis",* Annals of Mathematical Statistics 33: 1-67
- Images: https://commons.wikimedia.org/wiki/Main_Page