

# Teach the tidyverse to beginners

David Robinson  
DataCamp

# Teach the Tidyverse

# Teach the Tidyverse



## Data Science in the Tidyverse

*Charlotte Wickham*

This is a two-day hands on workshop based on the book [R for Data Science](#). This workshop is designed for people who are familiar with R and want to learn how to achieve their data analysis goals the “tidy” way. You will learn how to visualize, transform, and model data in R and work with date-times, character strings, and untidy data formats. Along the way, you will learn and use many packages from the tidyverse including ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, lubridate, andforcats.

## The “what”

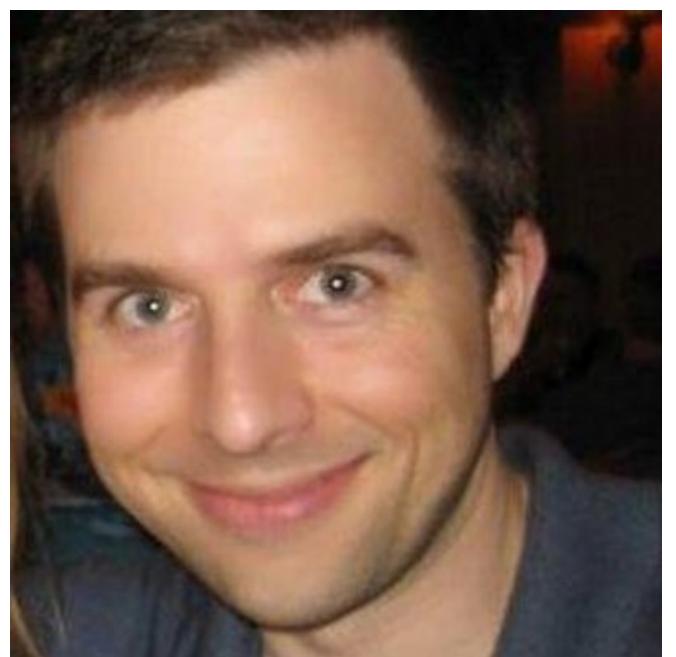
# Teach the Tidyverse



## Data Science in the Tidyverse

*Charlotte Wickham*

This is a two-day hands on workshop based on the book [R for Data Science](#). This workshop is designed for people who are familiar with R and want to learn how to achieve their data analysis goals the “tidy” way. You will learn how to visualize, transform, and model data in R and work with date-times, character strings, and untidy data formats. Along the way, you will learn and use many packages from the tidyverse including ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, lubridate, andforcats.



## Tidyverse Train-the-Trainer

*Garrett Grolemund*

This two day workshop will equip you to teach R effectively. We will draw on RStudio’s experience teaching R to recommend tips for designing, teaching, and supporting short R courses. You will learn practical activities that you can use immediately to improve your presentation style, learning outcomes, and student engagement. Classroom examples will focus on teaching Master the Tidyverse, a two day workshop developed by RStudio; as well as on using RStudio Cloud and its curriculum of tutorials to jump-start your own lessons. You will leave the class with a cognitive model of learning that you can use to develop your own effective workshops or courses within your organization. Participants will receive the course materials for Master the Tidyverse as well as a certificate of completion.

The “what”

The “how”

# Why Teach the Tidyverse?



## Data Science in the Tidyverse

*Charlotte Wickham*

This is a two-day hands on workshop based on the book [R for Data Science](#). This workshop is designed for people who are familiar with R and want to learn how to achieve their data analysis goals the “tidy” way. You will learn how to visualize, transform, and model data in R and work with date-times, character strings, and untidy data formats. Along the way, you will learn and use many packages from the tidyverse including ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, lubridate, andforcats.



## Tidyverse Train-the-Trainer

*Garrett Grolemund*

This two day workshop will equip you to teach R effectively. We will draw on RStudio’s experience teaching R to recommend tips for designing, teaching, and supporting short R courses. You will learn practical activities that you can use immediately to improve your presentation style, learning outcomes, and student engagement. Classroom examples will focus on teaching Master the Tidyverse, a two day workshop developed by RStudio; as well as on using RStudio Cloud and its curriculum of tutorials to jump-start your own lessons. You will leave the class with a cognitive model of learning that you can use to develop your own effective workshops or courses within your organization. Participants will receive the course materials for Master the Tidyverse as well as a certificate of completion.

The “what”

The “how”

# Posts on Variance Explained

DECEMBER 16, 2014

**Don't teach built-in plotting to beginners  
(teach ggplot2)**

FEBRUARY 12, 2016

**Why I use ggplot2**

JULY 5, 2017

**Teach the tidyverse to beginners**

SEPTEMBER 21, 2017

**Don't teach students the hard way first**

How do I recommend  
teaching R?

Have goals for what you want your students to do,  
and start them doing it as early as possible.

Have goals for what you want your students to do,  
and start them doing it as early as possible.

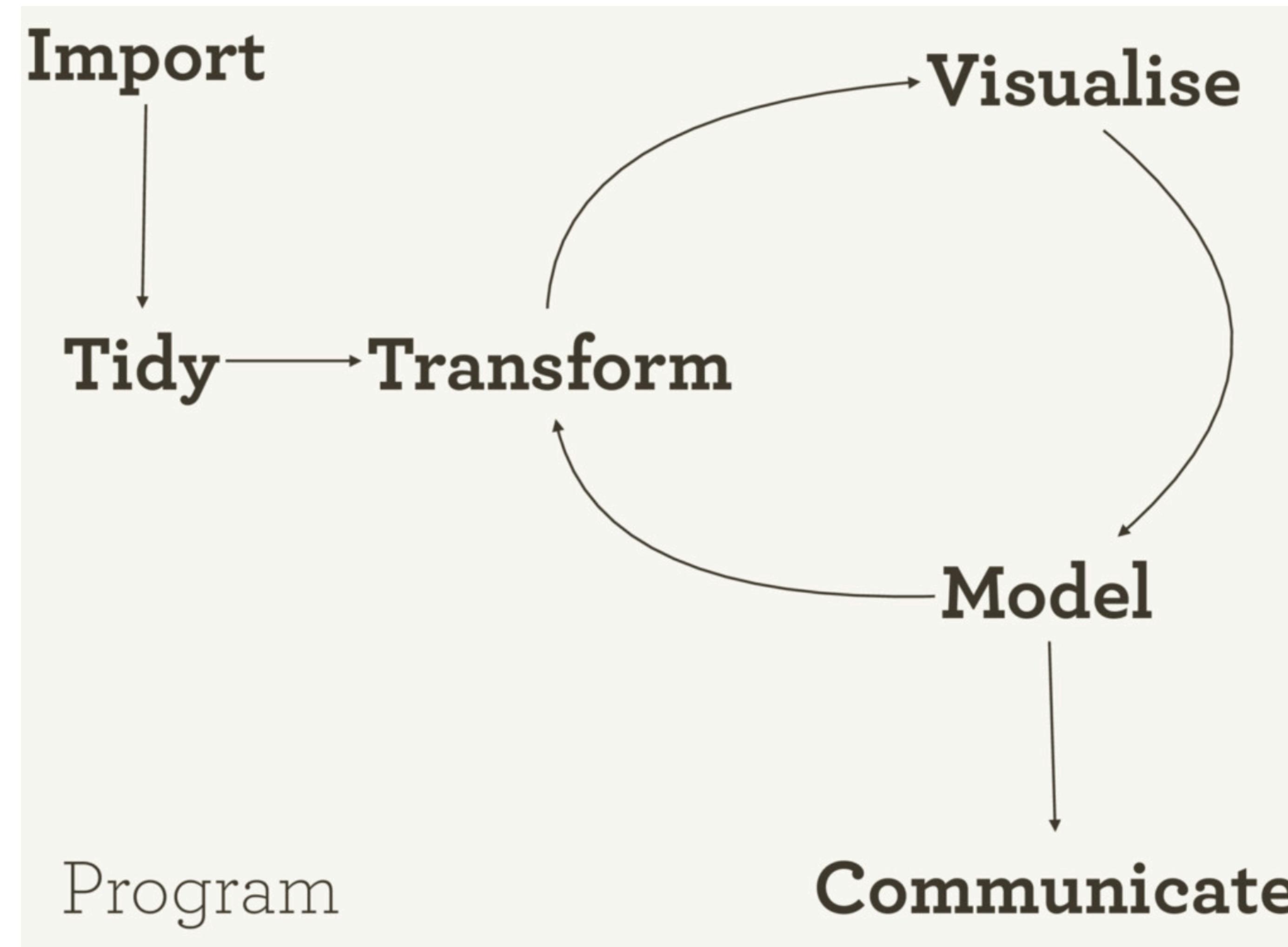
Have goals for what you want your students to do,  
and start them doing it as early as possible.

- Order matters

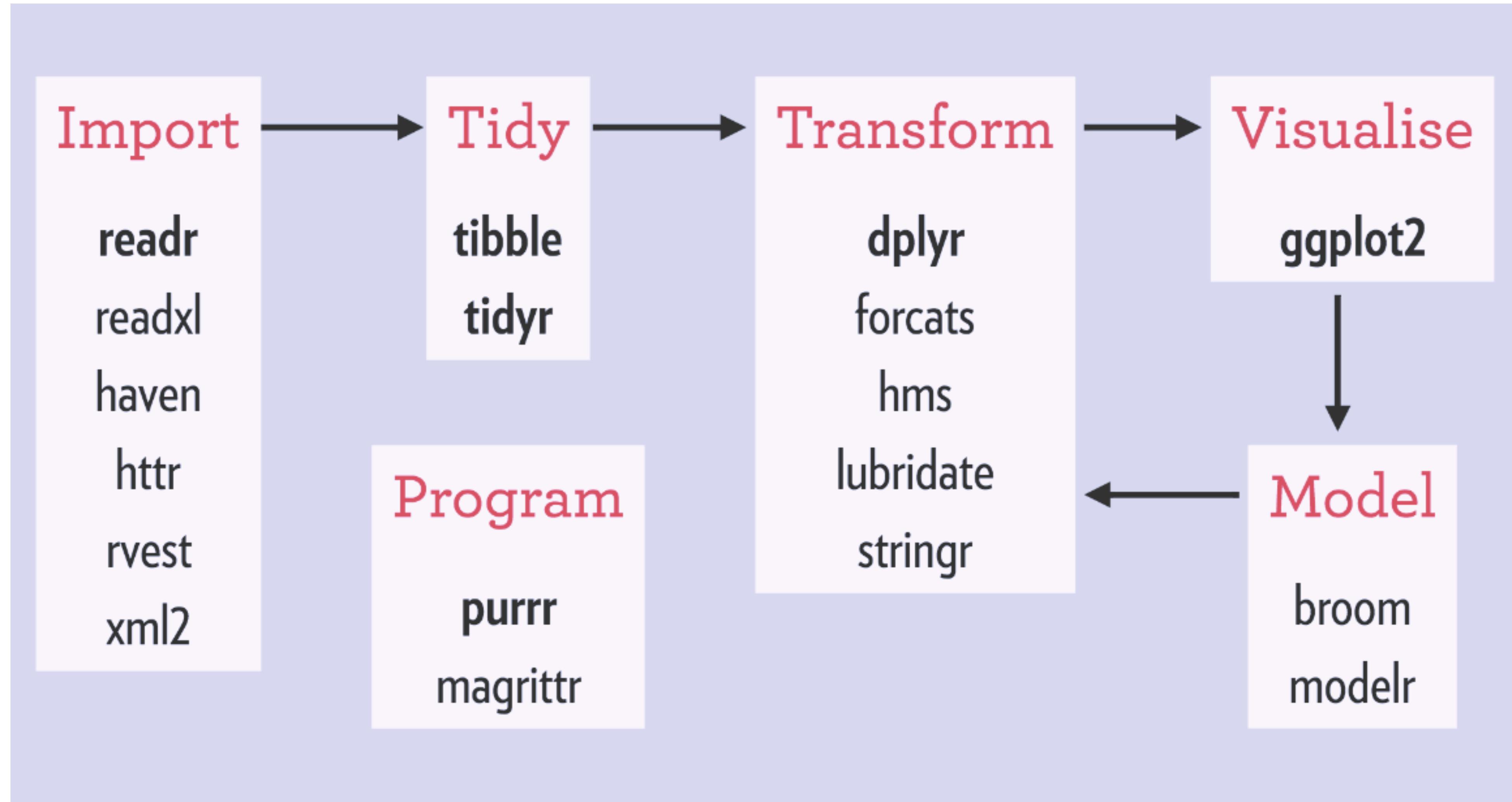
Have goals for what you want your students to do,  
and start them doing it as early as possible.

- Order matters
- When you teach something, show why it's useful

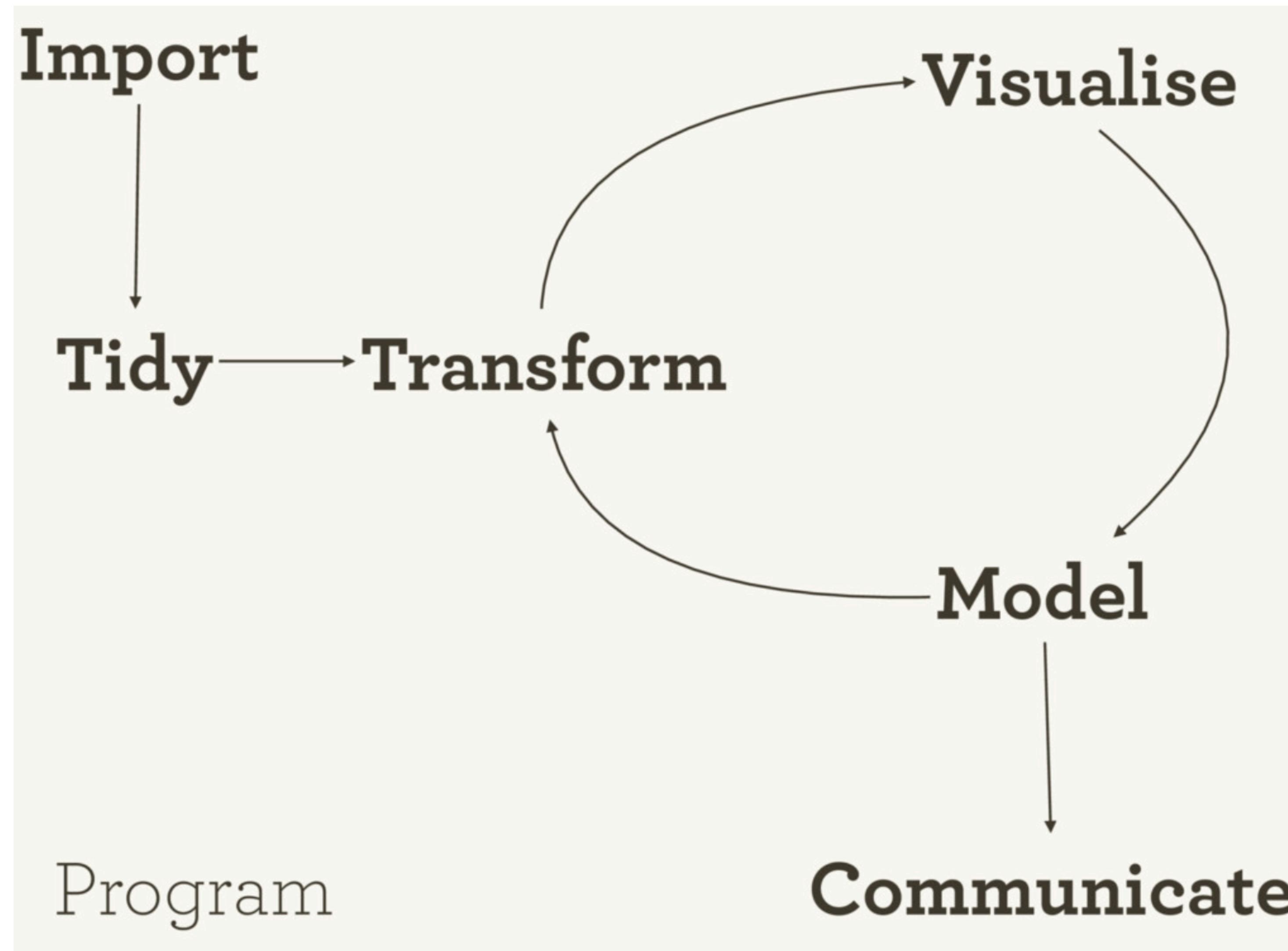
# I want students to understand their data



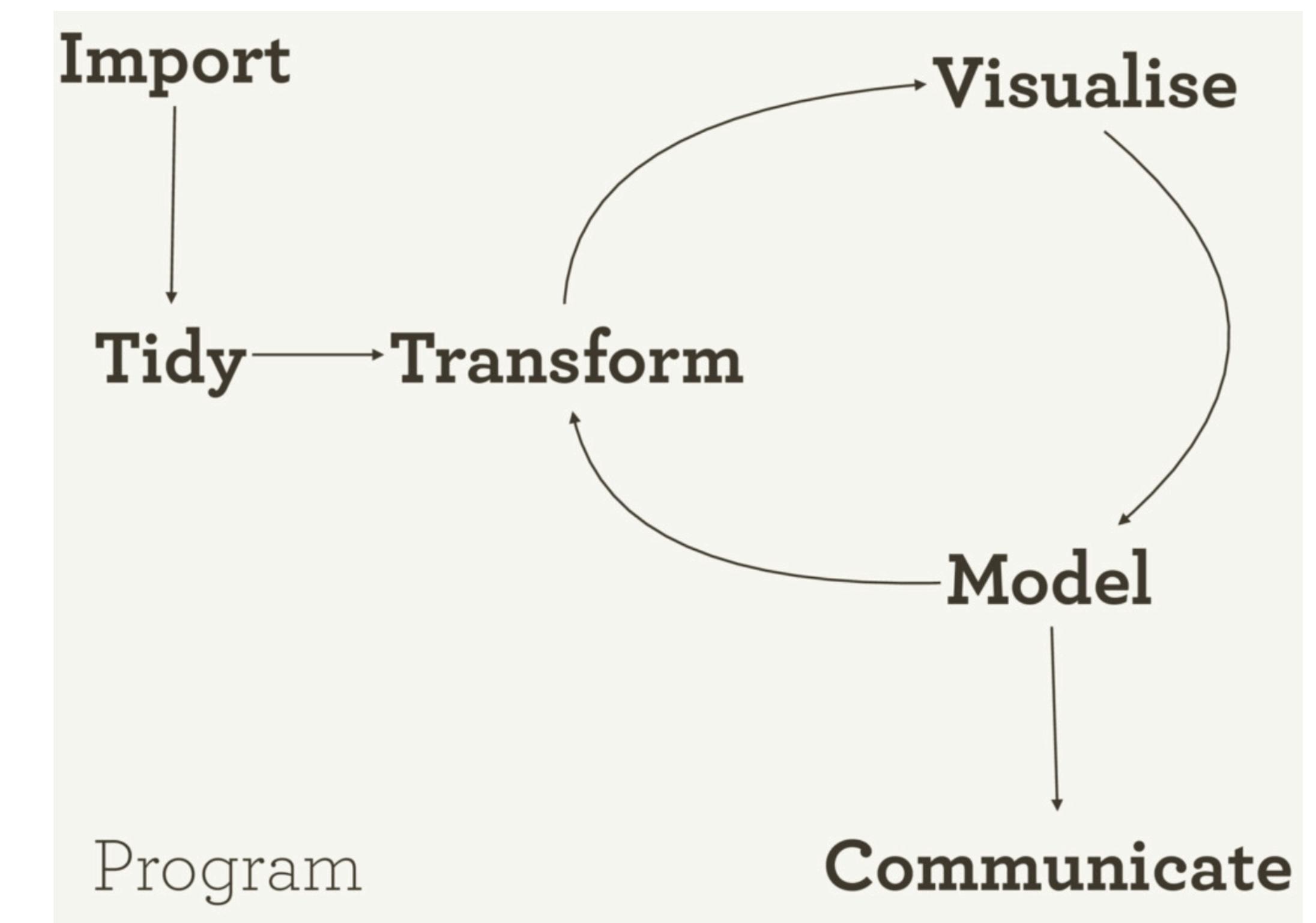
# I want students to understand their data



# I want students to understand their data



# I want students to understand their data



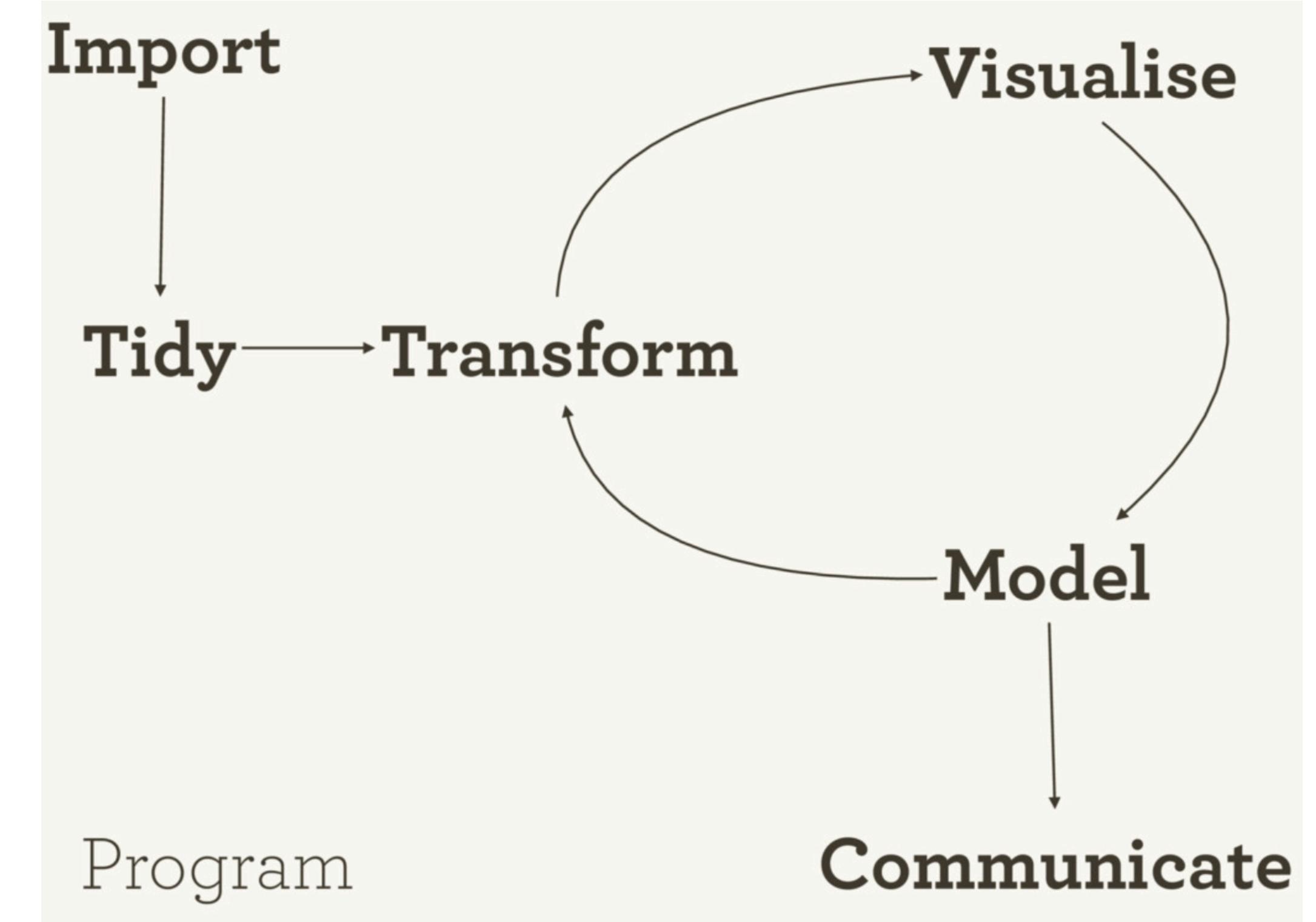
# I want students to understand their data

A screenshot of a DataCamp course page. The top navigation bar includes 'Learn', 'Groups', 'About', '22,140 XP', and user icons. The main title is 'PAID COURSE Introduction to the Tidyverse'. Below it is a yellow 'Continue Course' button. To the right is a circular course icon labeled 'INTRODUCTION TO THE TIDYVERSE' with a stylized R and Q logo. At the bottom, course details are listed: '4 hours', '16 Videos', '50 Exercises', '7,003 Participants', and '4,150 XP'.

## Course Description

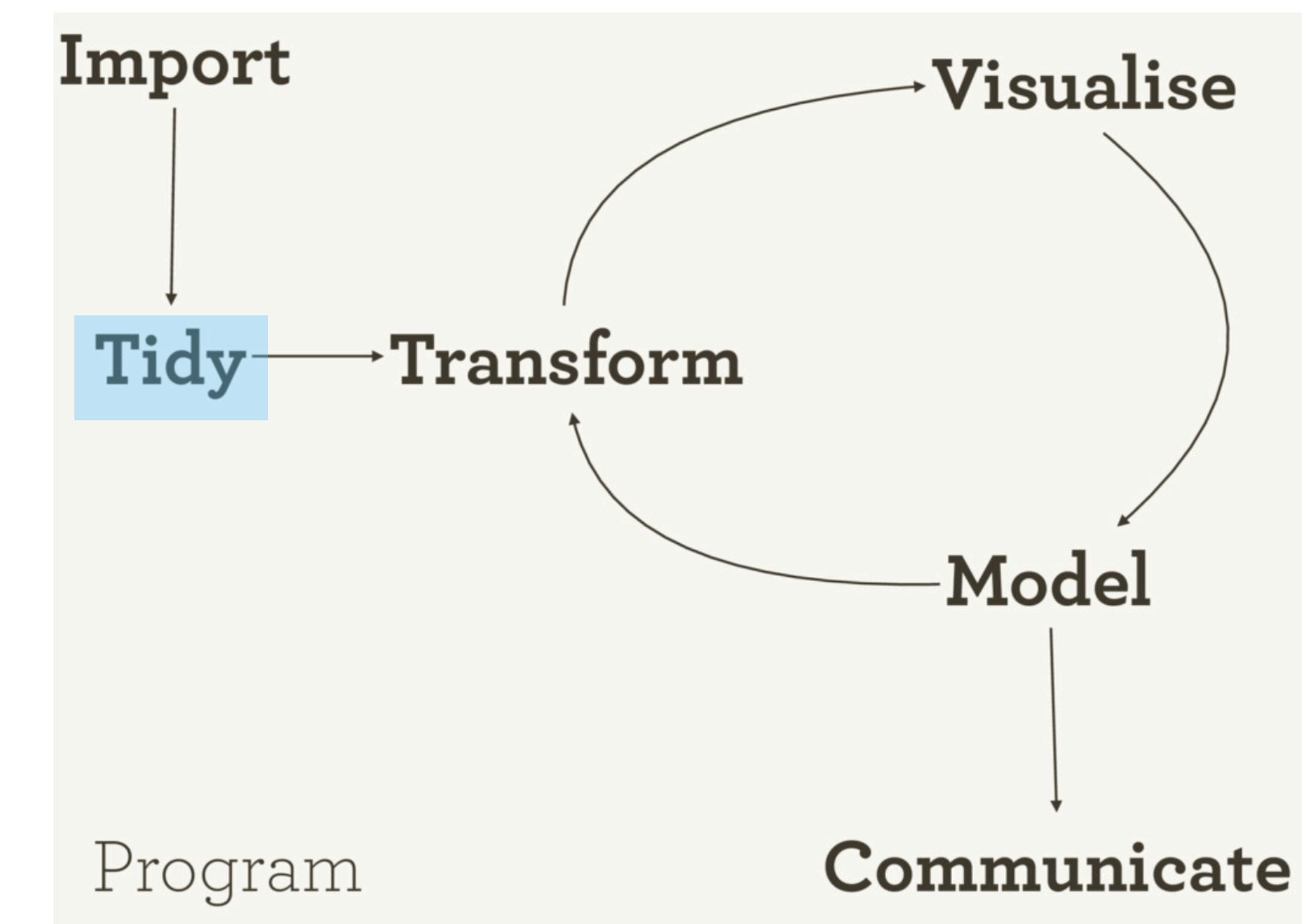
This is an introduction to the programming language R, focused on a powerful set of tools known as the "tidyverse". In the course you'll learn the intertwined processes of data manipulation and visualization through the tools dplyr and ggplot2. You'll learn to manipulate data by filtering, sorting and summarizing a real dataset of historical country data in order to answer exploratory questions. You'll then learn to turn this processed data into informative line plots, bar plots, histograms, and more with the ggplot2 package. This gives a taste both of the value of exploratory data analysis and the power of tidyverse tools. This is a suitable introduction for people who have no previous experience in R and are interested in learning to perform data analysis.

A profile box featuring a portrait of David Robinson, Chief Data Scientist at DataCamp. His title and company name are listed below his photo. A short bio notes that he works on the data science behind DataCamp's product.



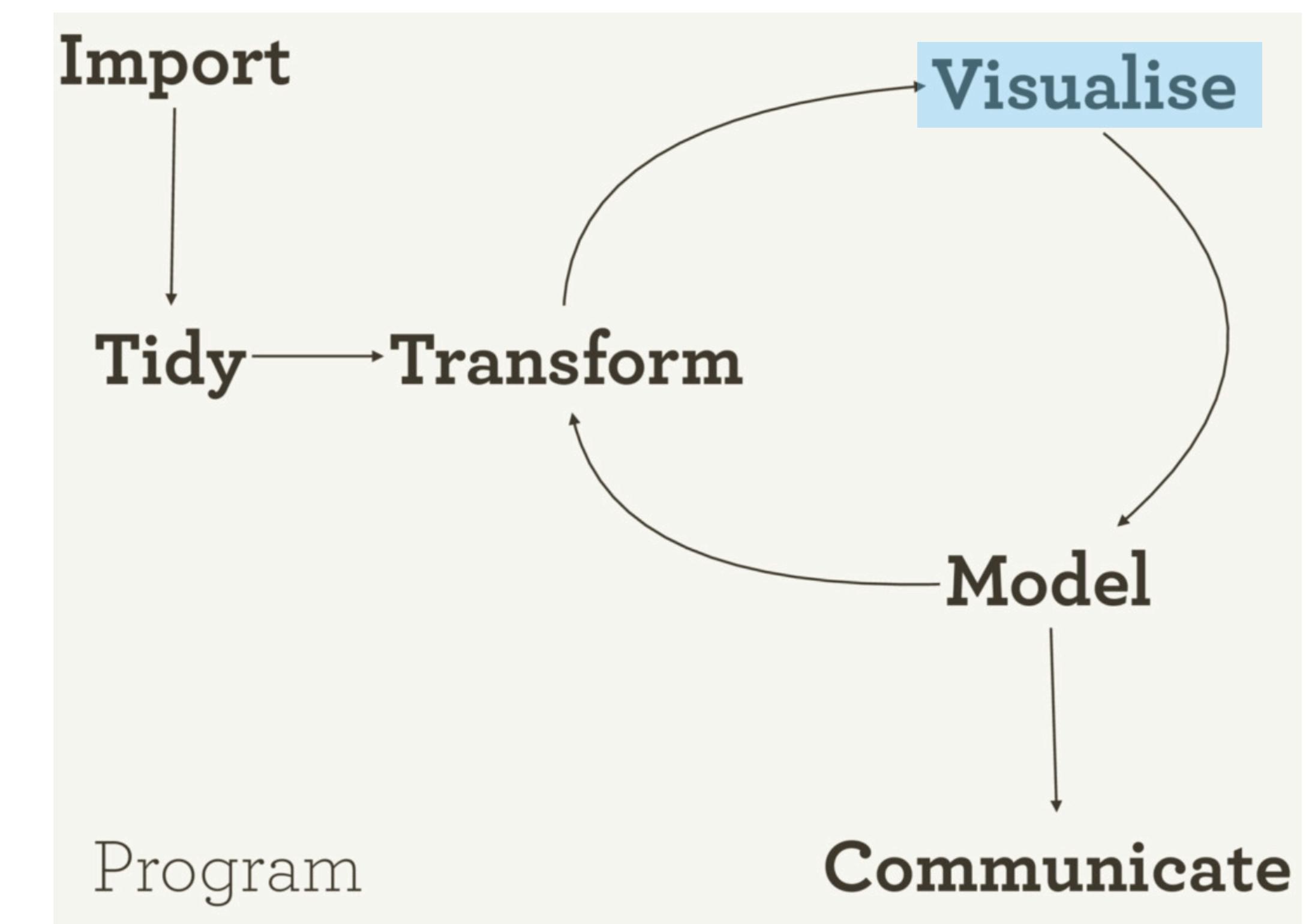
# I want students to understand their data

## 1 Data wrangling



# I want students to understand their data

- 1 Data wrangling
- 2 Data visualization

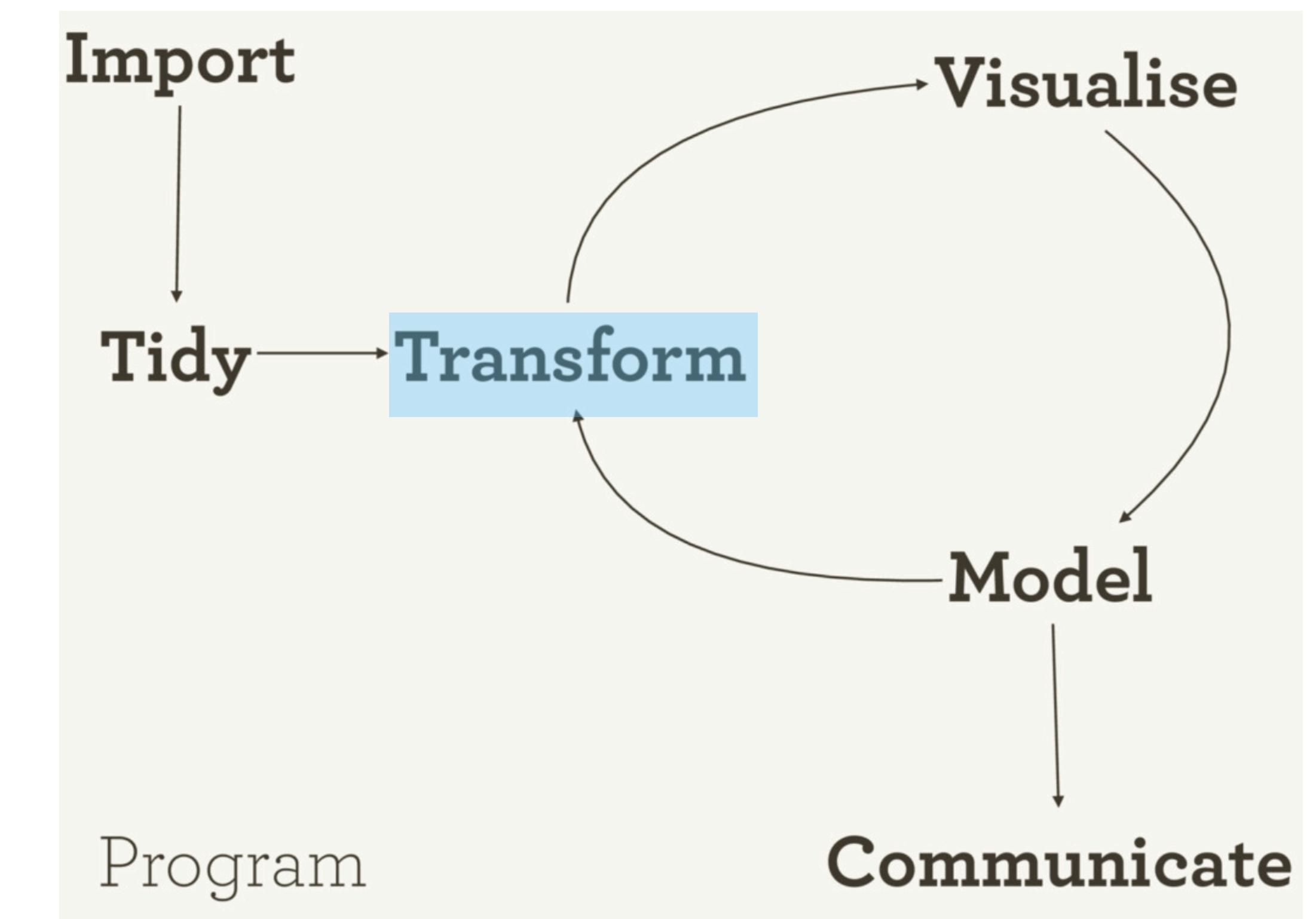


# I want students to understand their data

1 Data wrangling

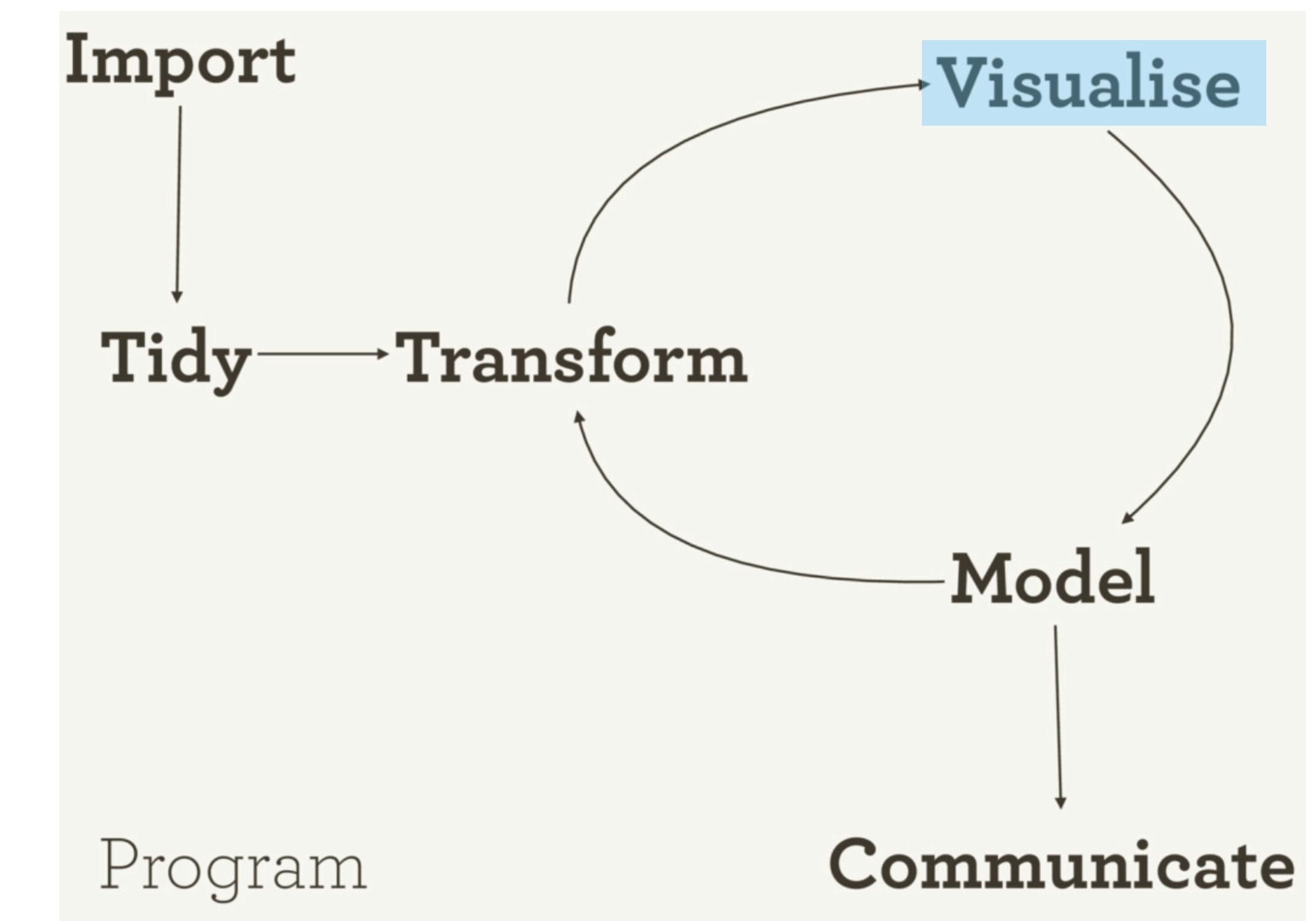
2 Data visualization

3 Grouping and summarizing

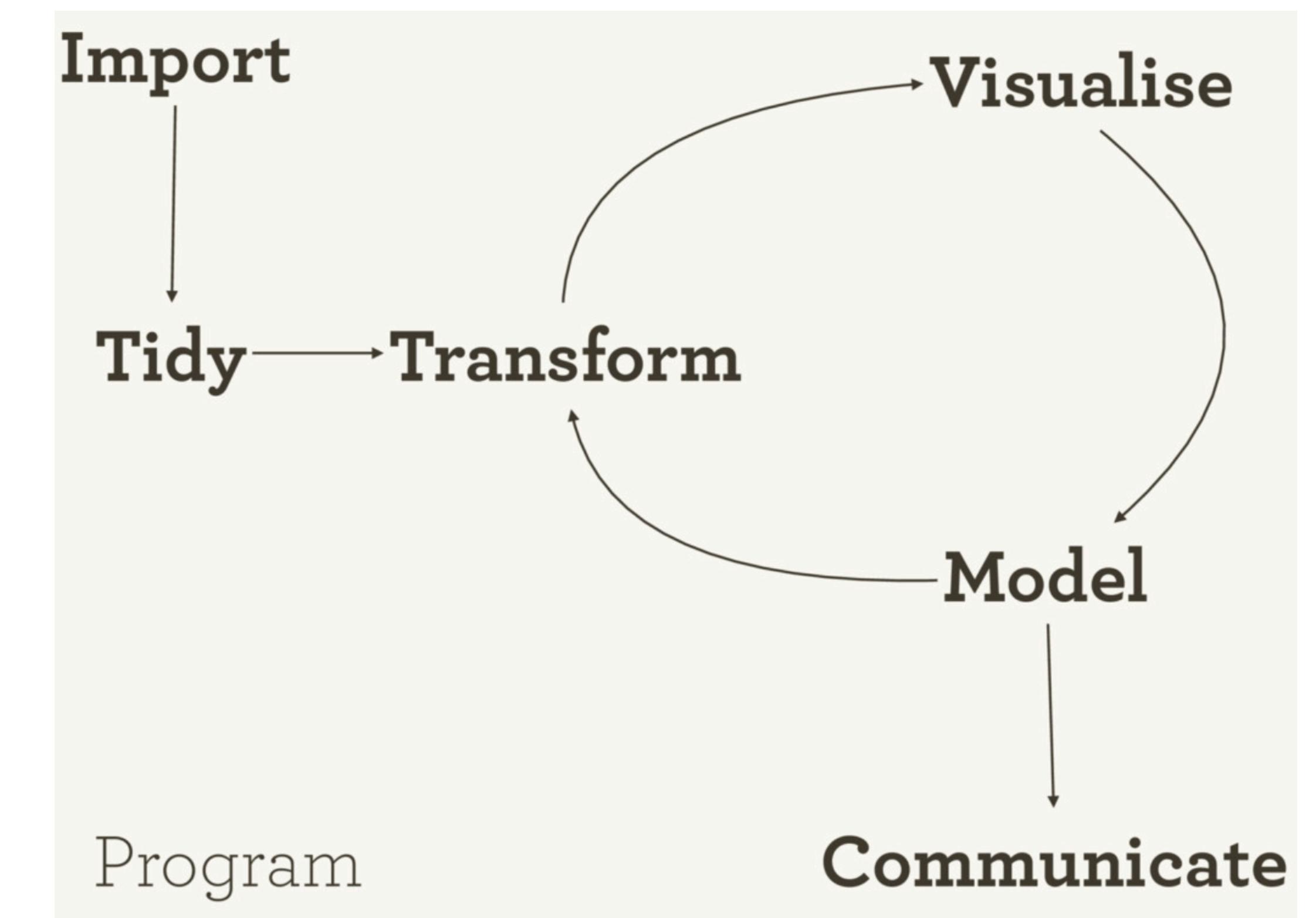


# I want students to understand their data

- 1 Data wrangling
- 2 Data visualization
- 3 Grouping and summarizing
- 4 Types of visualizations



# I want students to understand their data

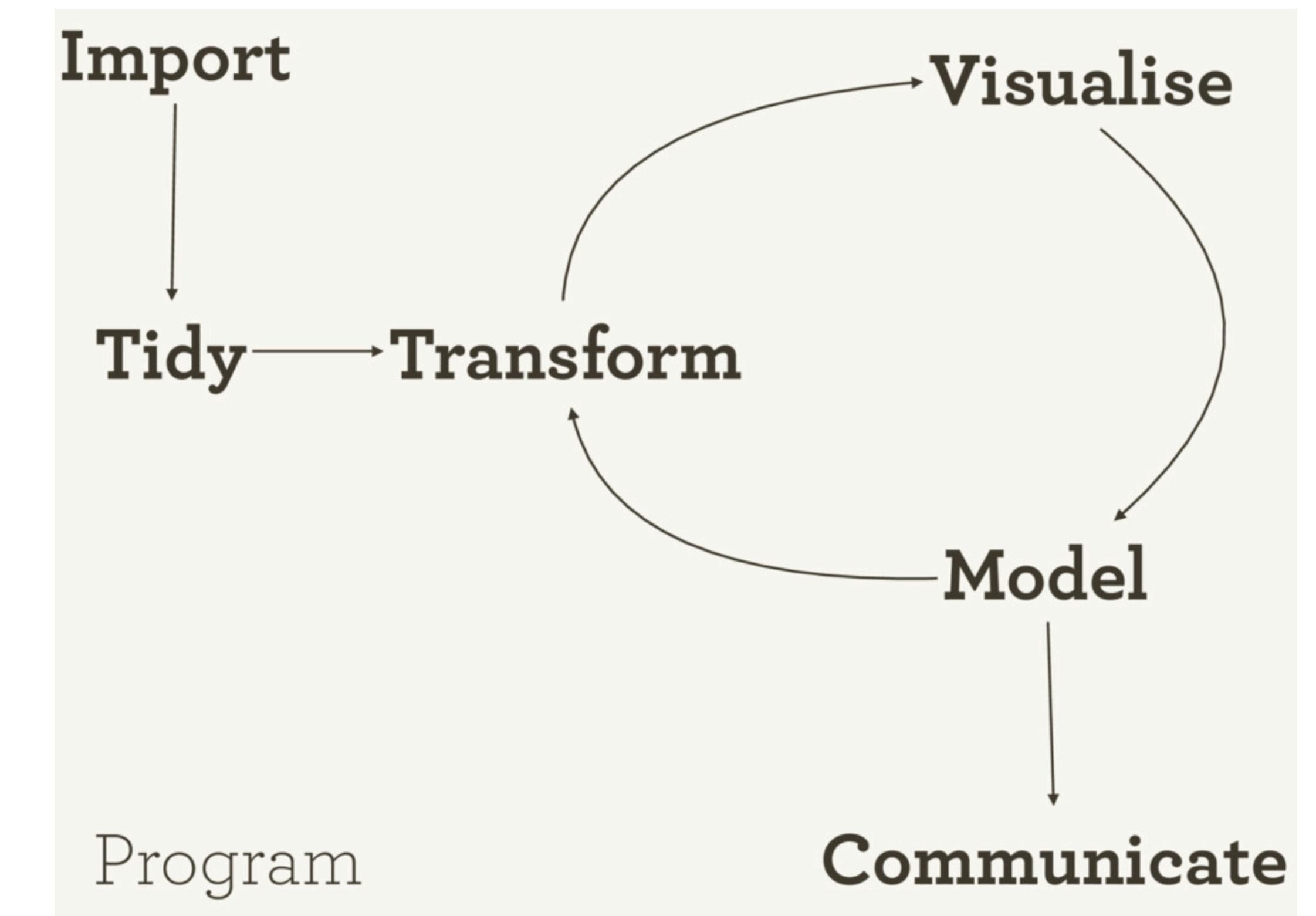
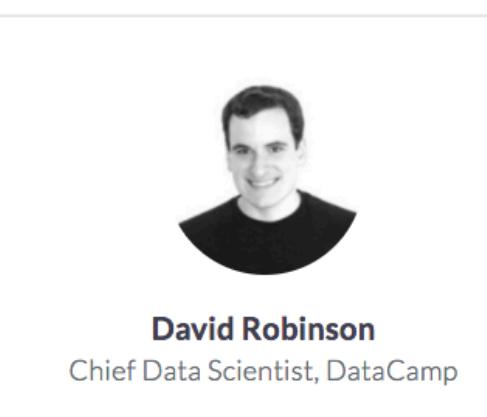


# I want students to understand their data

The screenshot shows the DataCamp platform interface. At the top, there's a navigation bar with 'Learn', 'Groups', 'About', and user stats ('22,140 XP'). Below the header, the course title 'Exploratory Data Analysis in R: Case Study' is displayed in large white text on a blue background. A yellow button labeled 'Continue Course' is visible. To the right of the title is a circular icon containing a 3D model of a server or data storage unit, labeled 'EDA: CASE STUDY'. At the bottom, course details are listed: '4 hours', '15 Videos', '58 Exercises', '9,348 Participants', and '4,800 XP'.

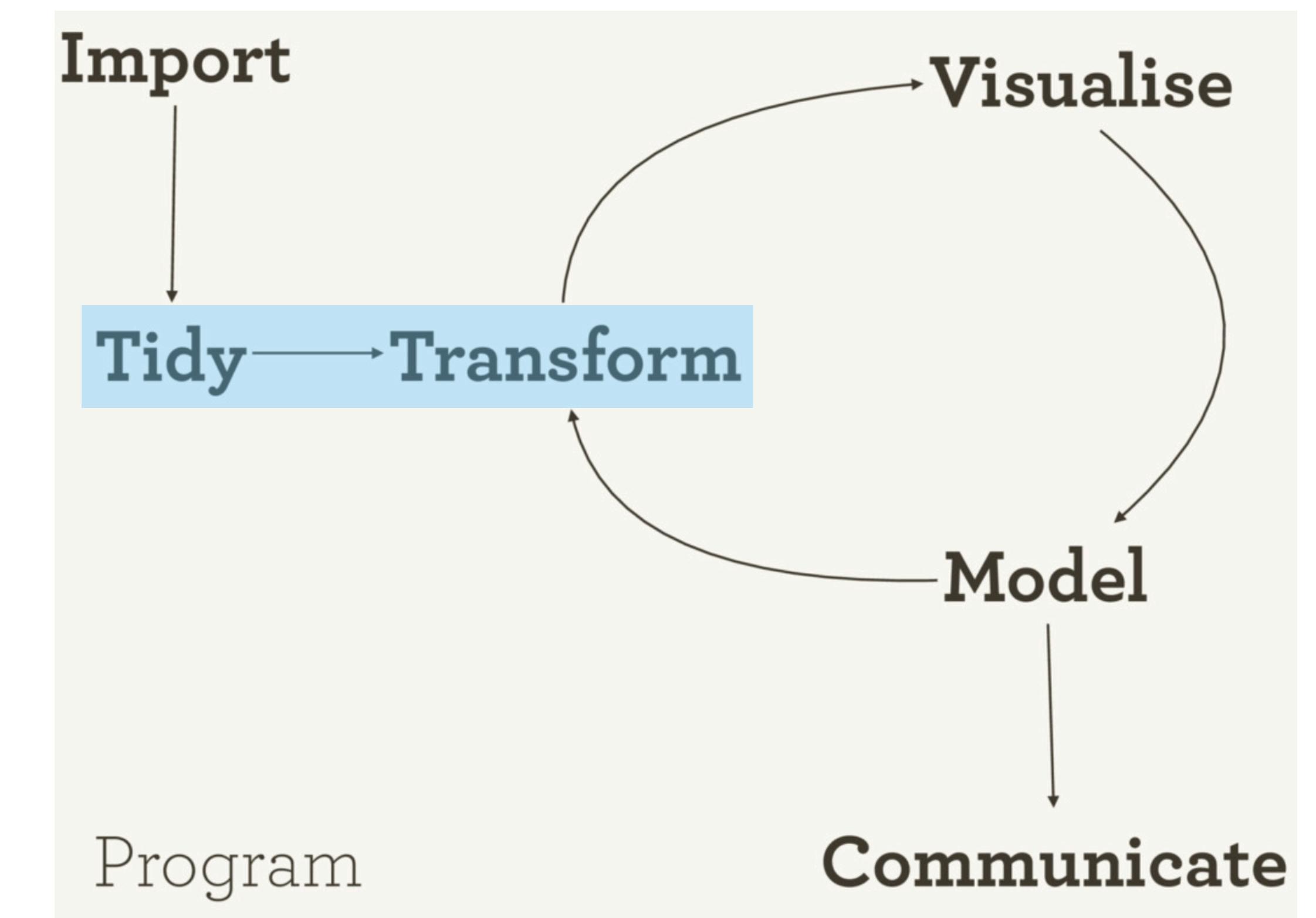
## Course Description

Once you've started learning tools for data manipulation and visualization like dplyr and ggplot2, this course gives you a chance to use them in action on a real dataset. You'll explore the historical voting of the United Nations General Assembly, including analyzing differences in voting between countries, across time, and among international issues. In the process you'll gain more practice with the dplyr and ggplot2 packages, learn about the broom package for tidying model output, and experience the kind of start-to-finish exploratory analysis common in data science.



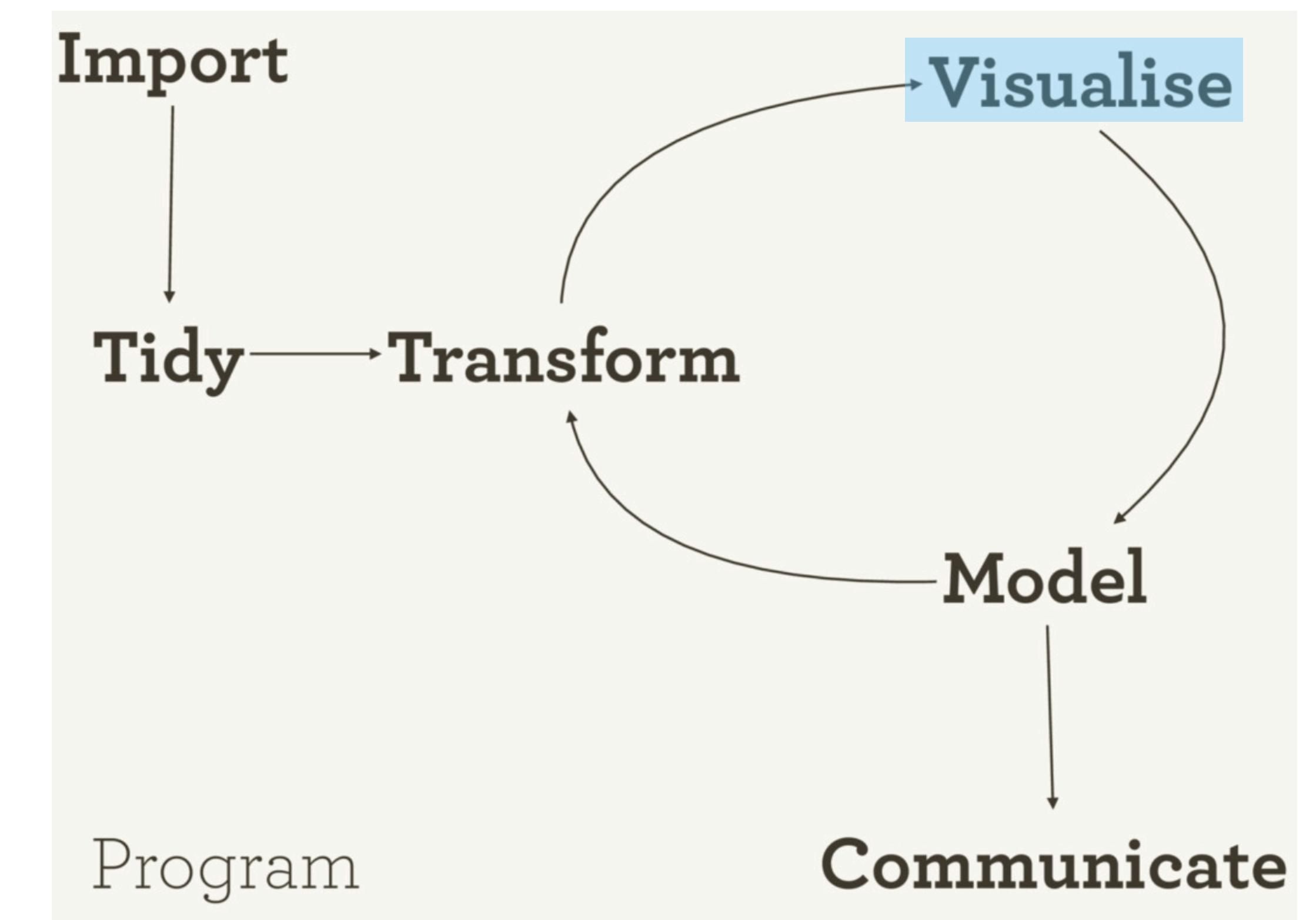
# I want students to understand their data

## ① Data cleaning and summarizing with dplyr



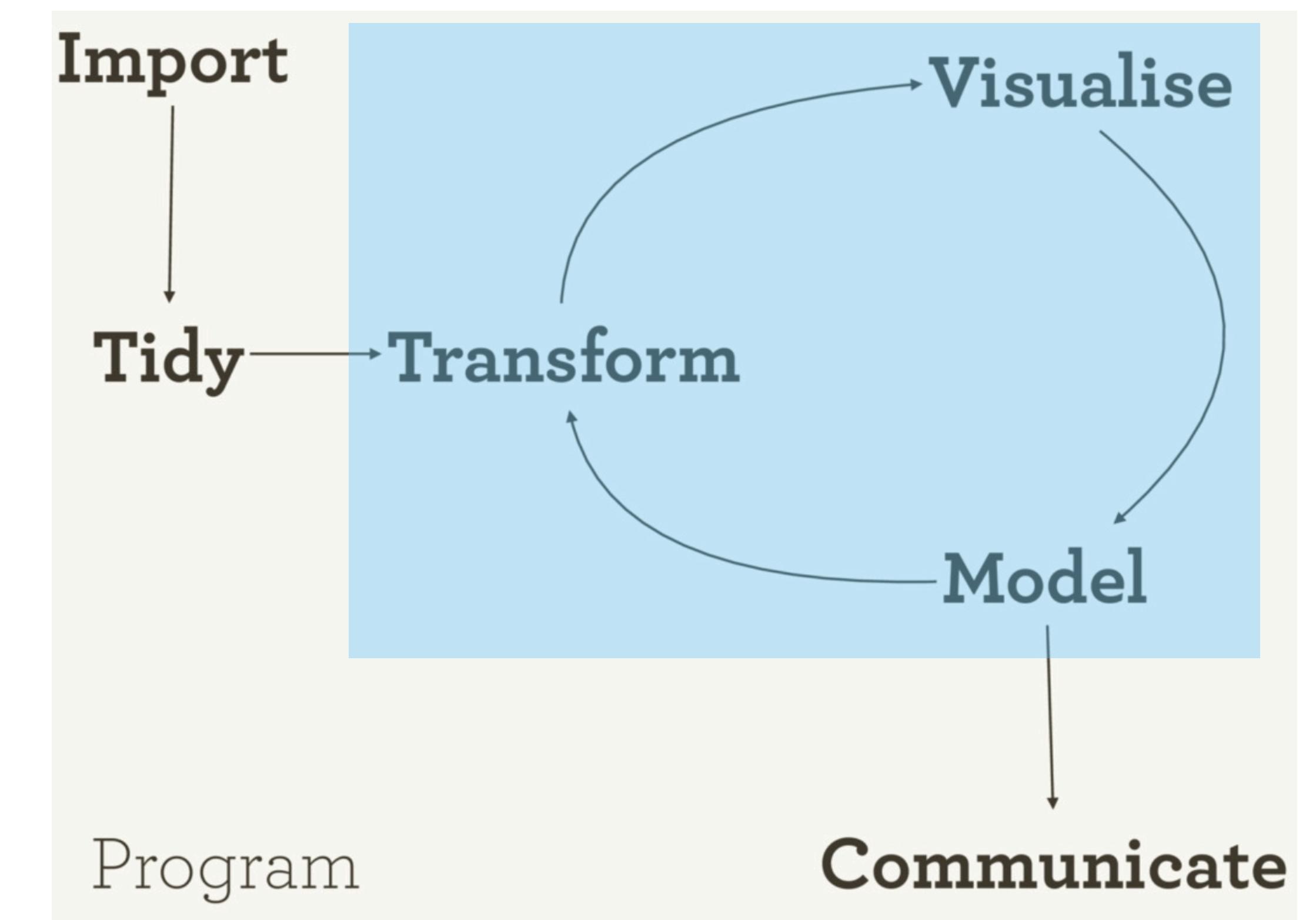
# I want students to understand their data

- 1 Data cleaning and summarizing with dplyr
- 2 Data visualization with ggplot2



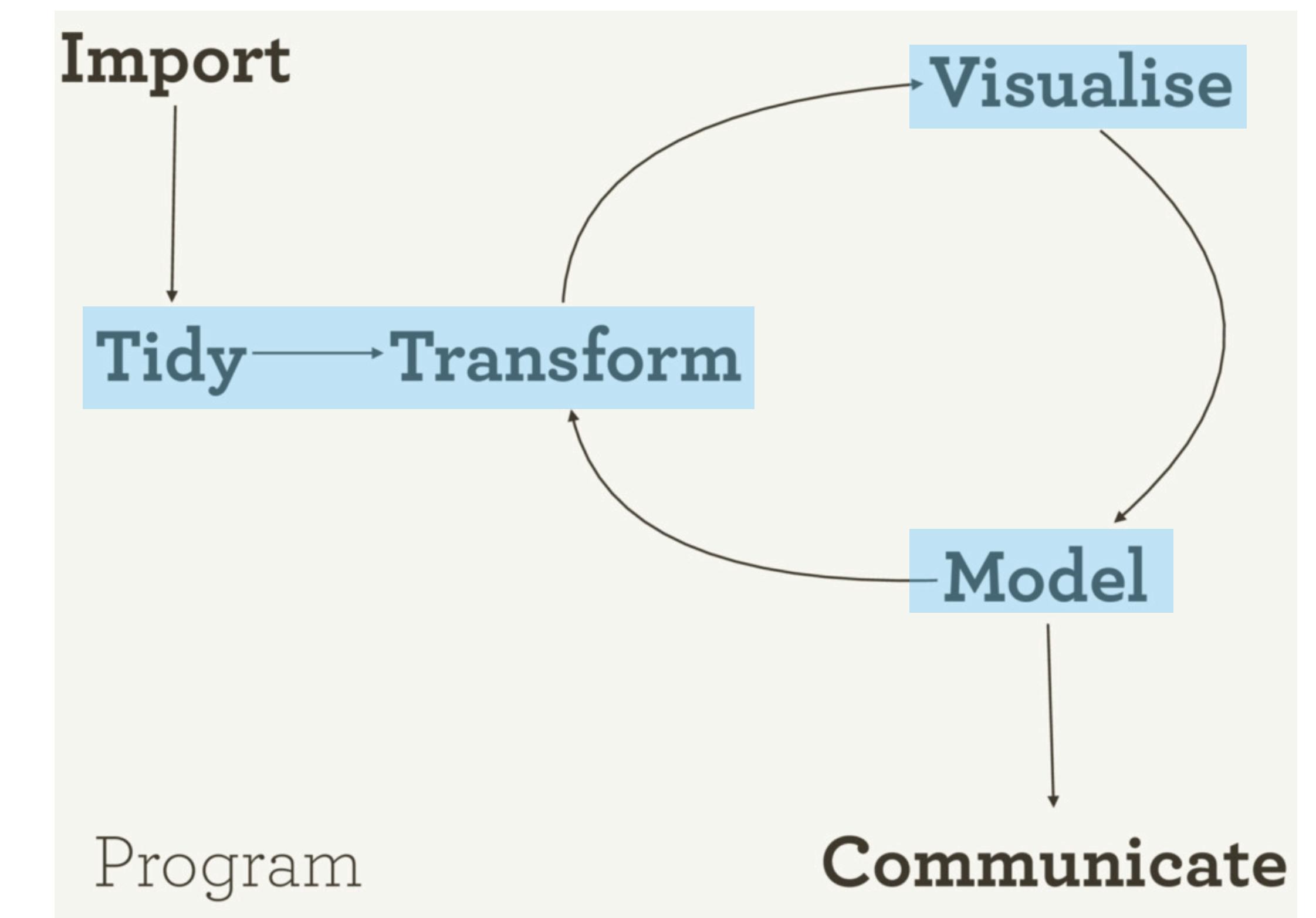
# I want students to understand their data

- 1 Data cleaning and summarizing with dplyr
- 2 Data visualization with ggplot2
- 3 Tidy modeling with broom



# I want students to understand their data

- 1 Data cleaning and summarizing with dplyr
- 2 Data visualization with ggplot2
- 3 Tidy modeling with broom
- 4 Joining and tidying

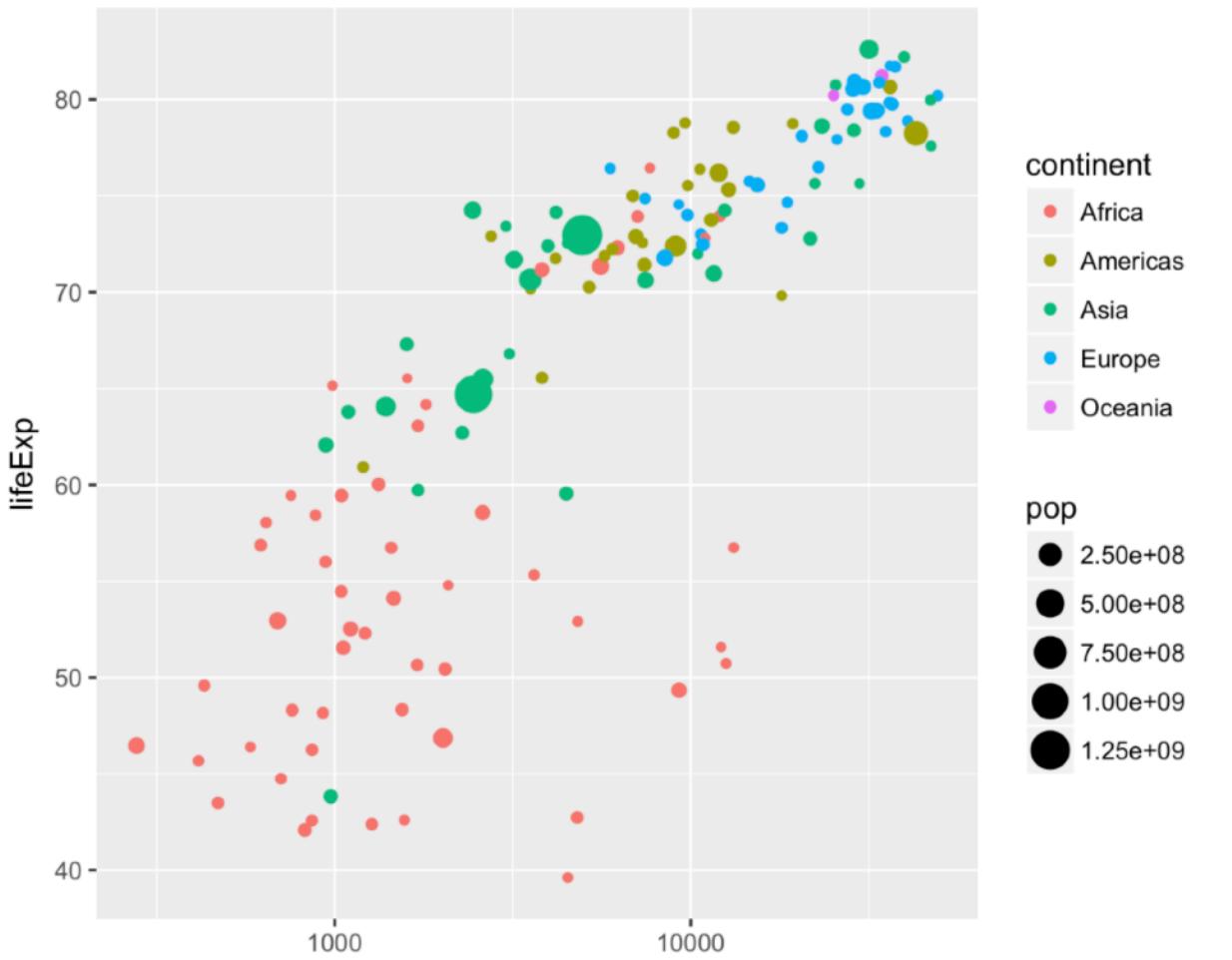


# Why is the tidyverse well suited for beginners?

# Why is the tidyverse well suited for beginners?

```
gapminder %>%
  mutate(gdp = gdpPercap * pop) %>%
  filter(year == 2007) %>%
  arrange(desc(gdp))
```

```
# A tibble: 142 x 7
  country continent year lifeExp      pop gdpPercap      gdp
  <fctr>   <fctr> <int>   <dbl>    <dbl>    <dbl>
1 United States Americas 2007 78.242 301139947 42951.653 1.293446e+13
2 China Asia 2007 72.961 1318683096 4959.115 6.539501e+12
3 Japan Asia 2007 82.603 127467972 31656.068 4.035135e+12
4 India Asia 2007 64.698 1110396331 2452.210 2.722925e+12
5 Germany Europe 2007 79.406 82400996 32170.374 2.650871e+12
6 United Kingdom Europe 2007 79.425 60776238 33203.261 2.017969e+12
7 France Europe 2007 80.657 61083916 30470.017 1.861228e+12
8 Brazil Americas 2007 72.390 190010647 9065.801 1.722599e+12
9 Italy Europe 2007 80.546 58147733 28569.720 1.661264e+12
10 Mexico Americas 2007 76.195 108700891 11977.575 1.301973e+12
# ... with 132 more rows
```



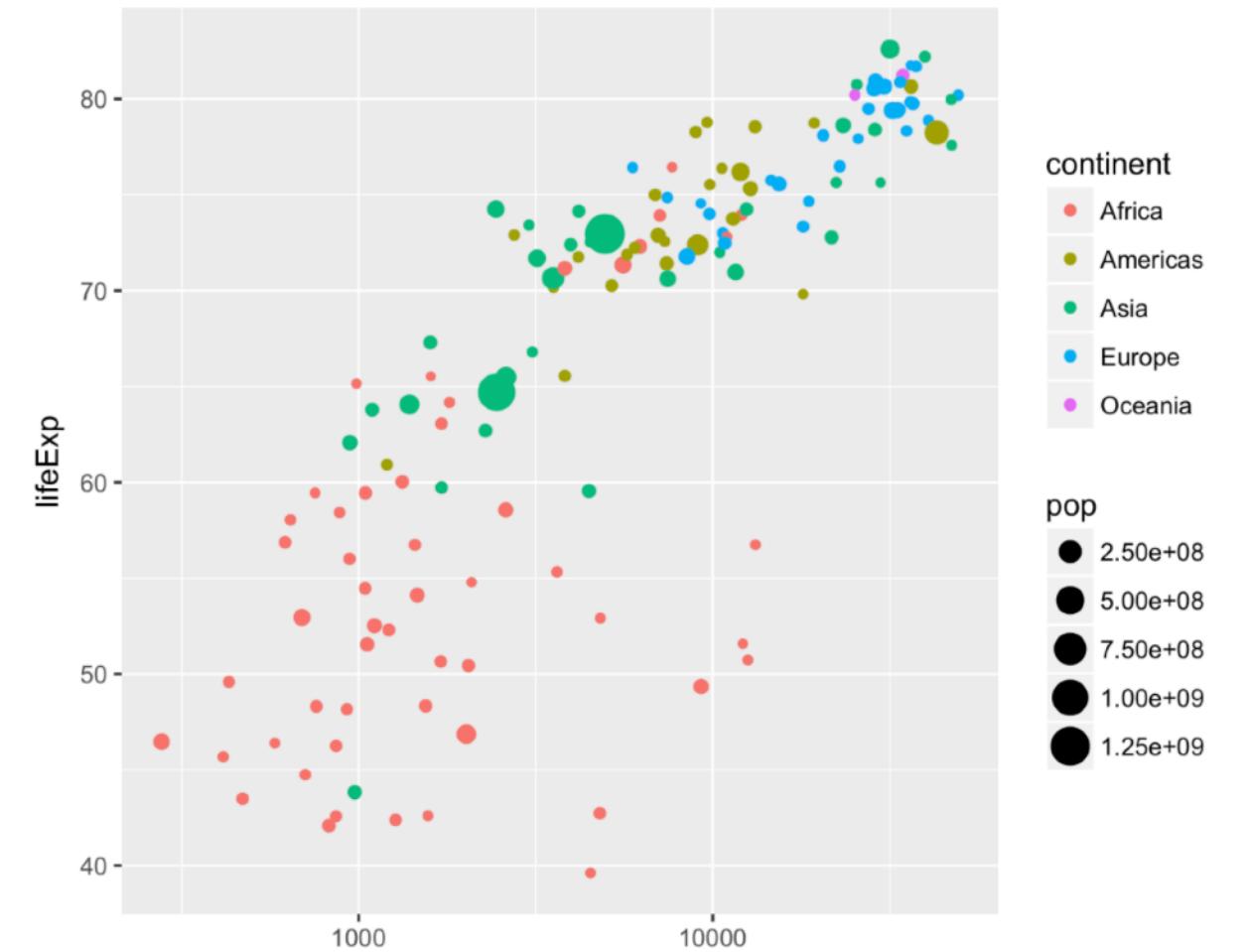
Powerful

# Why is the tidyverse well suited for beginners?

```
gapminder %>%  
  mutate(gdp = gdpPercap * pop) %>%  
  filter(year == 2007) %>%  
  arrange(desc(gdp))
```

```
# A tibble: 142 x 7  
# ... with 132 more rows
```

	country	continent	year	lifeExp	pop	gdpPercap	gdp
	<fctr>	<fctr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	United States	Americas	2007	78.242	301139947	42951.653	1.293446e+13
2	China	Asia	2007	72.961	1318683096	4959.115	6.539501e+12
3	Japan	Asia	2007	82.603	127467972	31656.068	4.035135e+12
4	India	Asia	2007	64.698	1110396331	2452.210	2.722925e+12
5	Germany	Europe	2007	79.406	82400996	32170.374	2.650871e+12
6	United Kingdom	Europe	2007	79.425	60776238	33203.261	2.017969e+12
7	France	Europe	2007	80.657	61083916	30470.017	1.861228e+12
8	Brazil	Americas	2007	72.390	190010647	9065.801	1.722599e+12
9	Italy	Europe	2007	80.546	58147733	28569.720	1.661264e+12
10	Mexico	Americas	2007	76.195	108700891	11977.575	1.301973e+12



Powerful

Consistent

fs package

Naming convention. **fs** functions use a consistent naming convention. Because base R's functions were gradually added over time there are a number of different conventions used (e.g. `path.expand()` vs `normalizePath()` ; `Sys.chmod()` vs `file.access()` ).

# **What is left until later?**

# What is left until later?

## Data Structures

```
factor(c("a", "b"), levels = c("b", "a"))
```

```
matrix(1:12, nrow = 3)
```

```
list(b = 2, c = 8)
```

# What is left until later?

## Data Structures

```
factor(c("a", "b"), levels = c("b", "a"))
matrix(1:12, nrow = 3)
list(b = 2, c = 8)
```

## Loops

```
for (x in 1:10) {
  y <- x + 2
  print(y)
}
```

# What is left until later?

## Data Structures

```
factor(c("a", "b"), levels = c("b", "a"))
matrix(1:12, nrow = 3)
list(b = 2, c = 8)
```

## Loops

```
for (x in 1:10) {
  y <- x + 2
  print(y)
}
```

## Conditionals

```
x <- 2
if (x > 4) {
  print("x is greater than 4")
} else {
  print("x is not greater than 4")
}
```

# What is left until later?

## Data Structures

```
factor(c("a", "b"), levels = c("b", "a"))
matrix(1:12, nrow = 3)
list(b = 2, c = 8)
```

## Loops

```
for (x in 1:10) {
  y <- x + 2
  print(y)
}
```

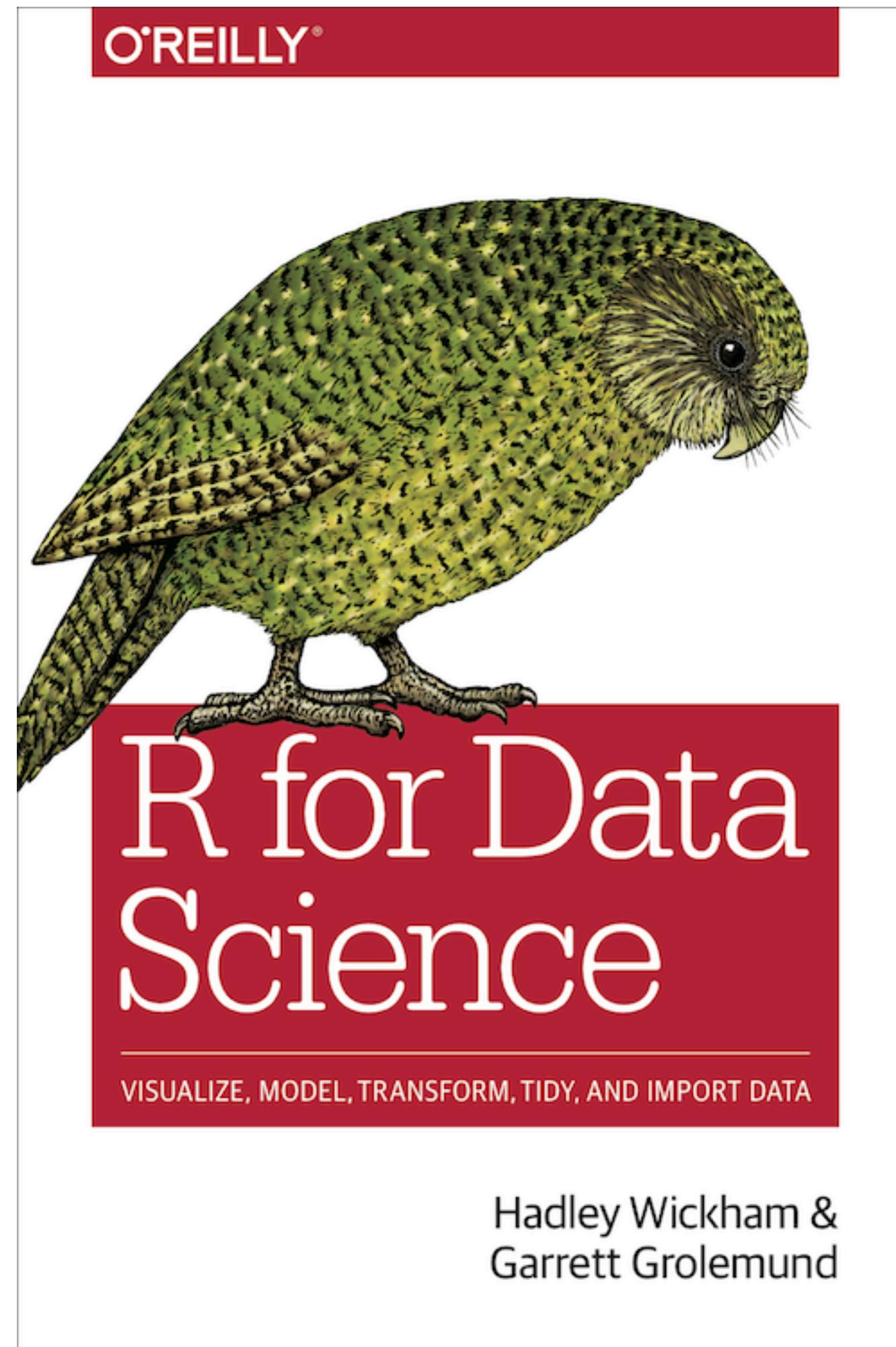
## Conditionals

```
x <- 2
if (x > 4) {
  print("x is greater than 4")
} else {
  print("x is not greater than 4")
}
```

## User-defined functions

```
add_three <- function(x) {
  x + 3
}
add_three(8)
```

# What is left until later?



[R for Data Science](#)

# What is left until later?

**1** Introduction

**I** Explore

**2** Introduction

**3** Data visualisation

**4** Workflow: basics

**5** Data transformation

**6** Workflow: scripts

**7** Exploratory Data Analysis

**8** Workflow: projects

**II Wrangle**

**9** Introduction

**10** Tibbles

**11** Data import

**12** Tidy data

**13** Relational data

**14** Strings

**15** Factors

**16** Dates and times

**III Program**

**17** Introduction

**18** Pipes

**19** Functions

**20** Vectors

**21** Iteration

**IV Model**

**22** Introduction

**23** Model basics

**24** Model building

**25** Many models

**V Communicate**

**26** Introduction

**27** R Markdown

**28** Graphics for communication

**29** R Markdown formats

**30** R Markdown workflow

# What is left until later?

**1** Introduction

**I** Explore

**2** Introduction

**3** Data visualisation

**4** Workflow: basics

**5** Data transformation

**6** Workflow: scripts

**7** Exploratory Data Analysis

**8** Workflow: projects

**II Wrangle**

**9** Introduction

**10** Tibbles

**11** Data import

**12** Tidy data

**13** Relational data

**14** Strings

**15** Factors

**16** Dates and times

**III Program**

**17** Introduction

**18** Pipes

**19** Functions

**20** Vectors

**21** Iteration

**IV Model**

**22** Introduction

**23** Model basics

**24** Model building

**25** Many models

**V Communicate**

**26** Introduction

**27** R Markdown

**28** Graphics for communication

**29** R Markdown formats

**30** R Markdown workflow

# What is left until later?

**1** Introduction

**I** Explore

**2** Introduction

**3** Data visualisation

**4** Workflow: basics

**5** Data transformation

**6** Workflow: scripts

**7** Exploratory Data Analysis

**8** Workflow: projects

**II Wrangle**

**9** Introduction

**10** Tibbles

**11** Data import

**12** Tidy data

**13** Relational data

**14** Strings

**15** Factors

**16** Dates and times

**III Program**

**17** Introduction

**18** Pipes

**19** Functions

**20** Vectors

**21** Iteration

**IV Model**

**22** Introduction

**23** Model basics

**24** Model building

**25** Many models

**V Communicate**

**26** Introduction

**27** R Markdown

**28** Graphics for communication

**29** R Markdown formats

**30** R Markdown workflow





*Wax on, wax off.*





```
x <- 1:10  
  
for (item in x) {  
  print(x)  
}
```



```
x <- 1:10  
  
for (item in x) {  
  print(x)  
}
```



```
s <- numeric(length(x) + 1)  
for (i in seq_along(s)) {  
  if (i == 1) {  
    s[i] <- x[i]  
  } else {  
    s[i] <- alpha * x[i] + (1 - alpha) * s[i - 1]  
  }  
}
```

Have goals for what you want your students to do,  
and start them doing it as early as possible.

Have goals for what you want your students to do,  
and start them doing it as **early** as possible.

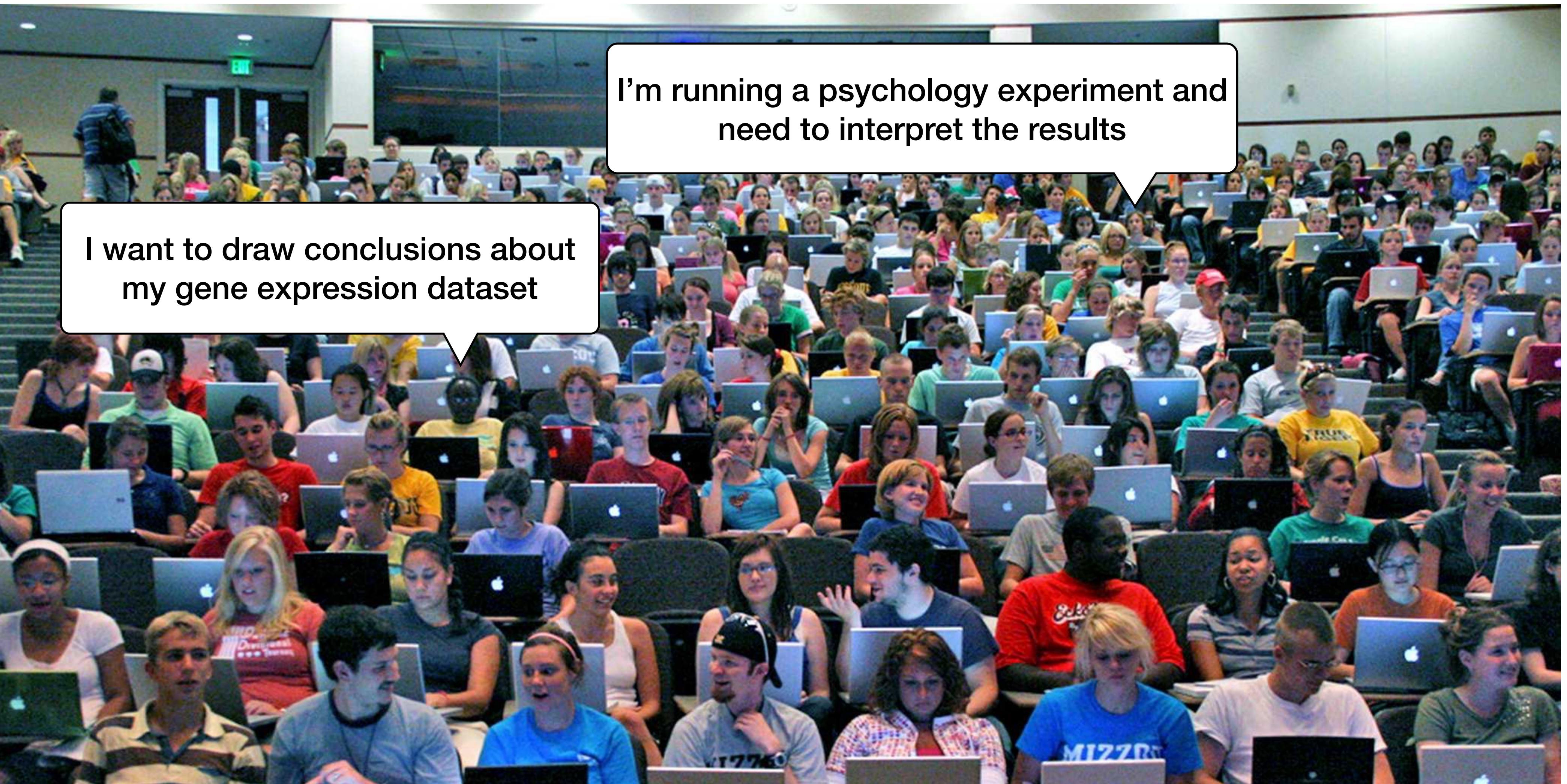
# What do beginners want to learn?



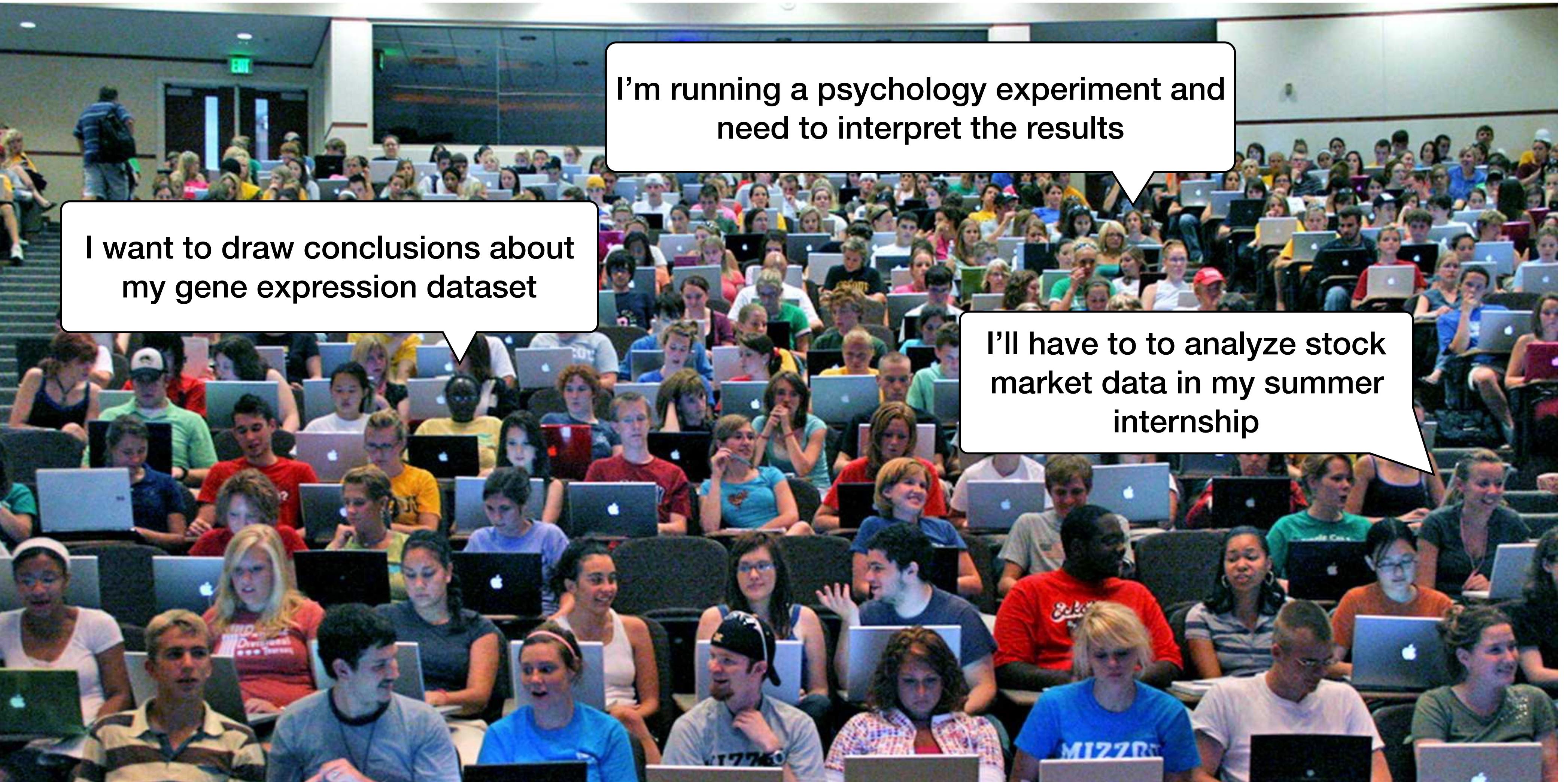
# What do beginners want to learn?



# What do beginners want to learn?



# What do beginners want to learn?



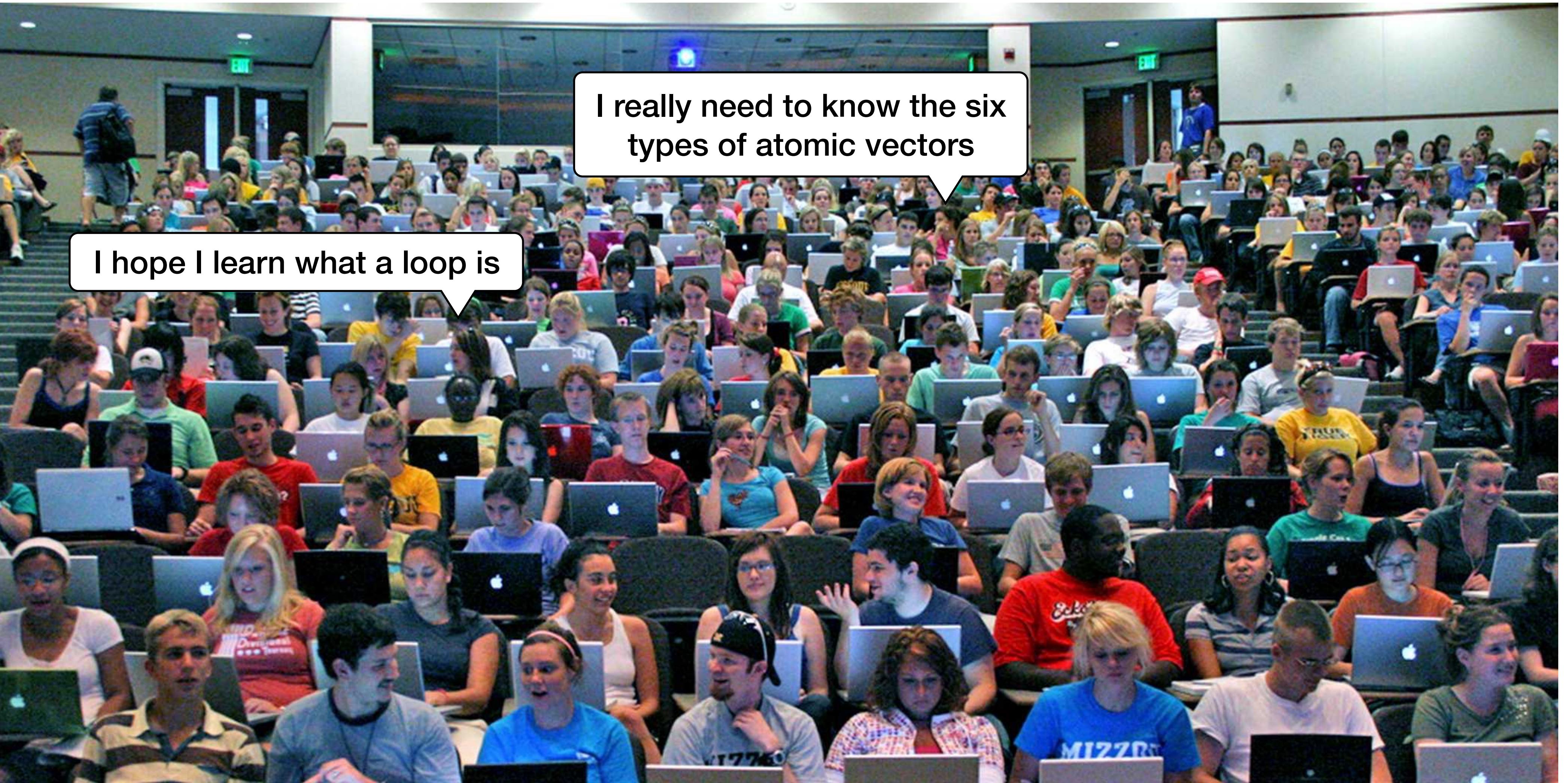
# What do beginners want to learn?



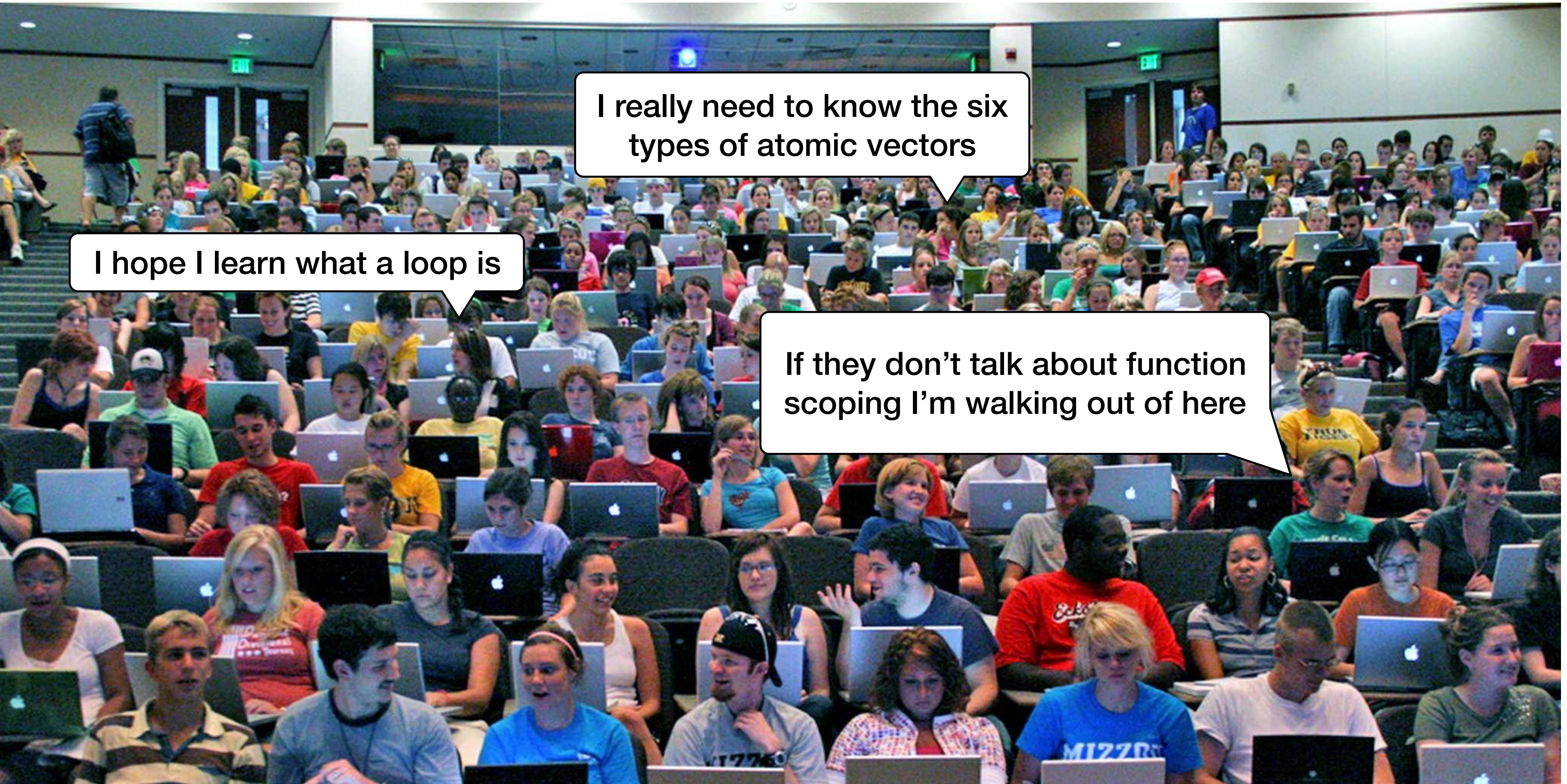
# What do beginners want to learn?

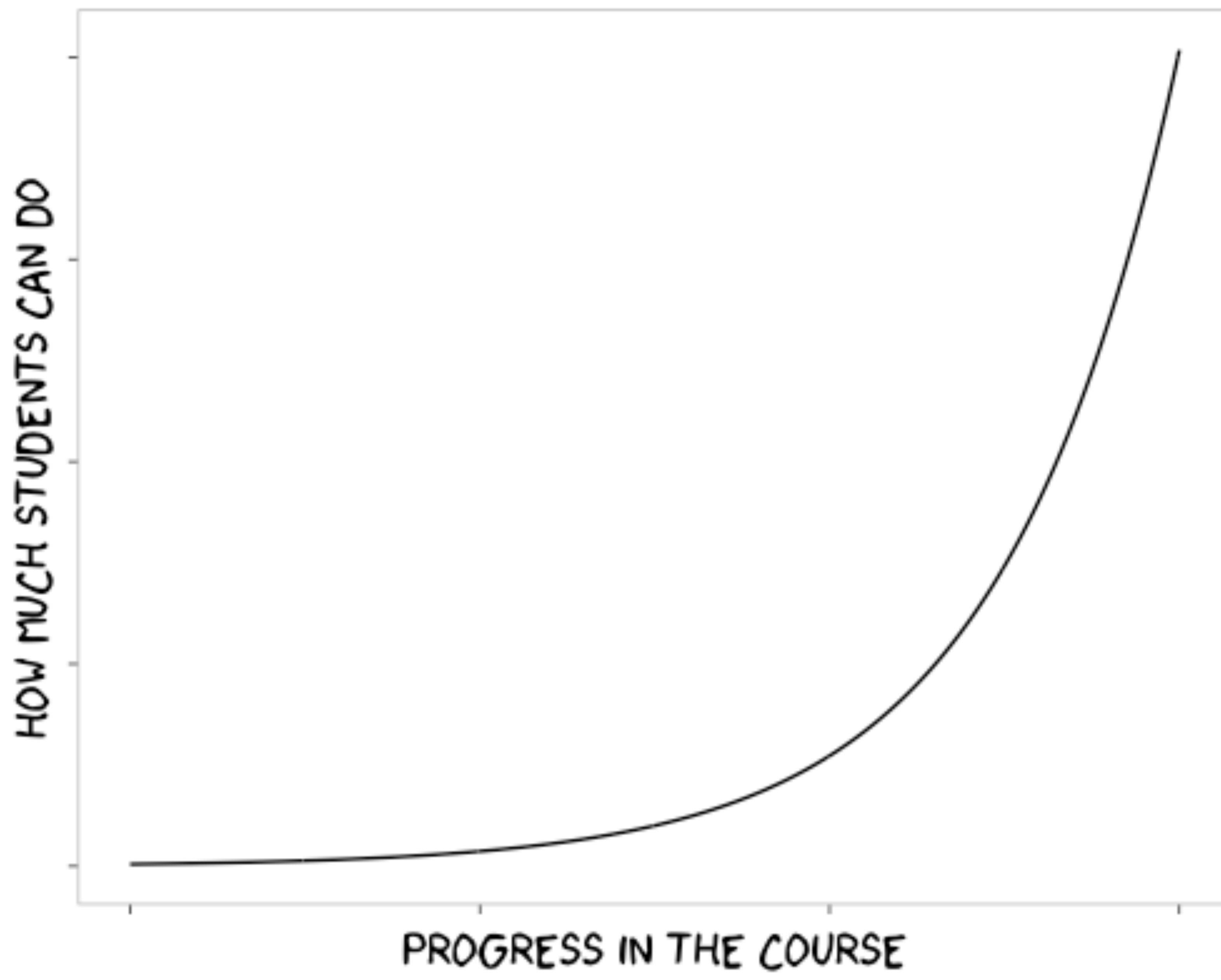


# What do beginners want to learn?



# What do beginners want to learn?





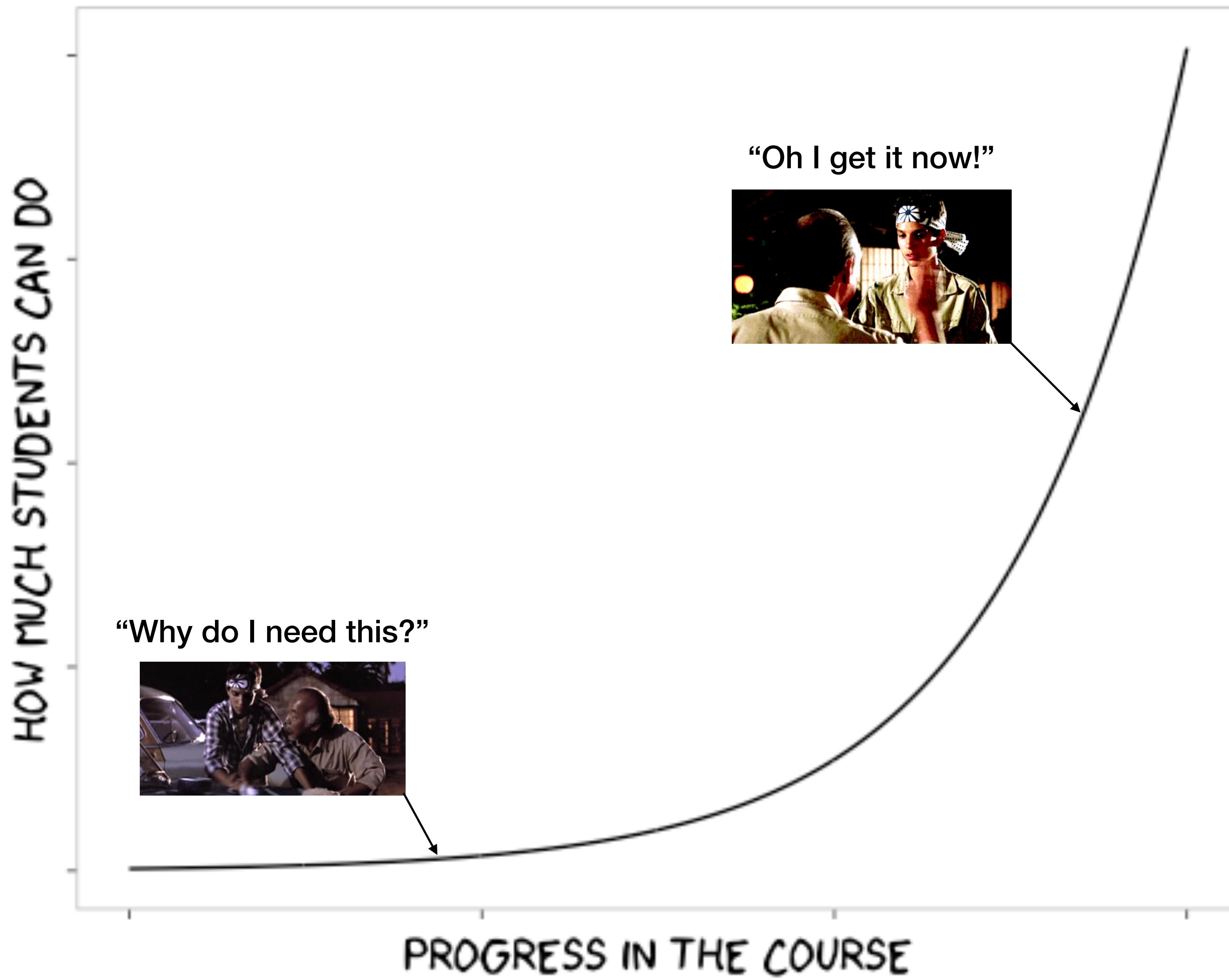
HOW MUCH STUDENTS CAN DO

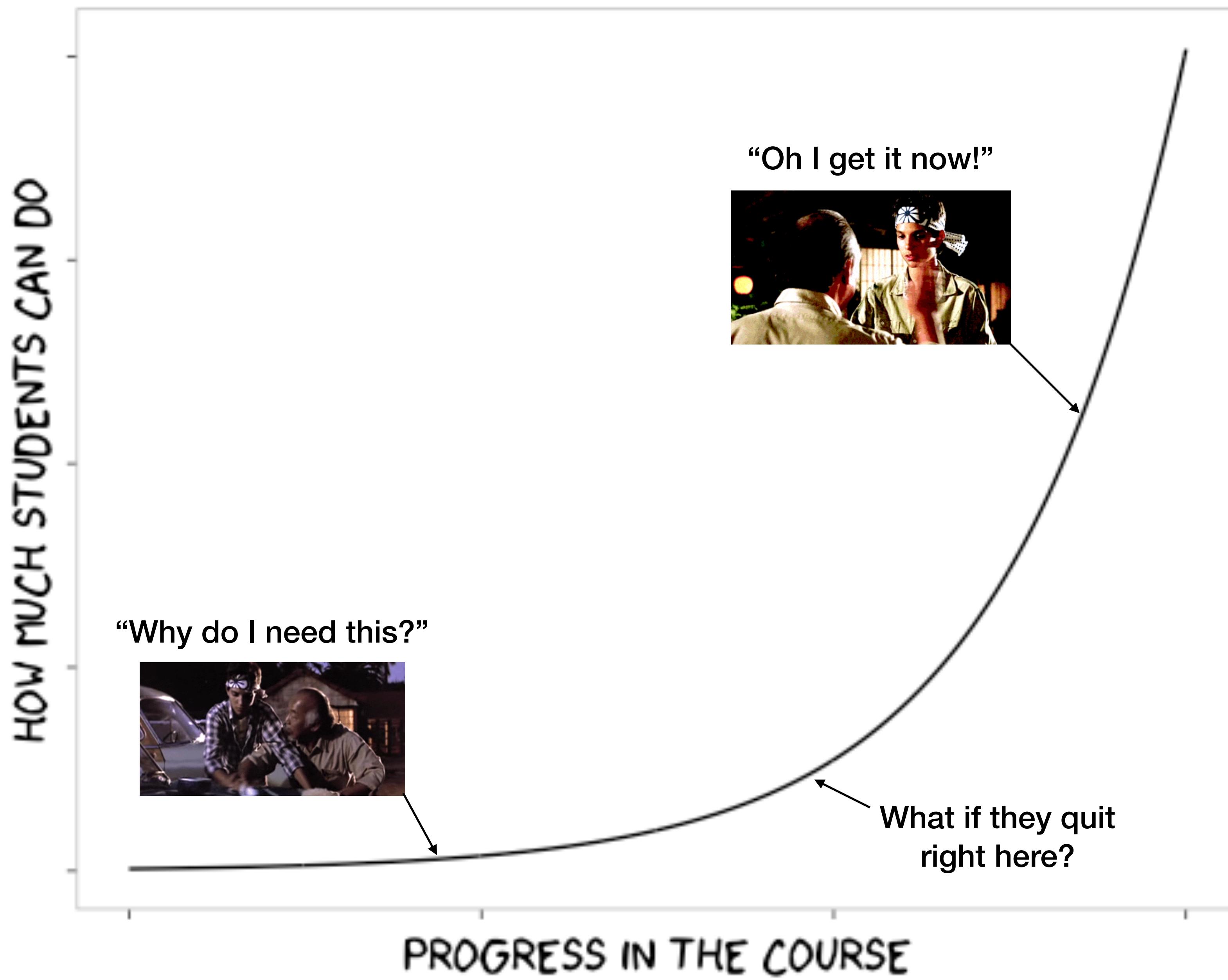
“Why do I need this?”

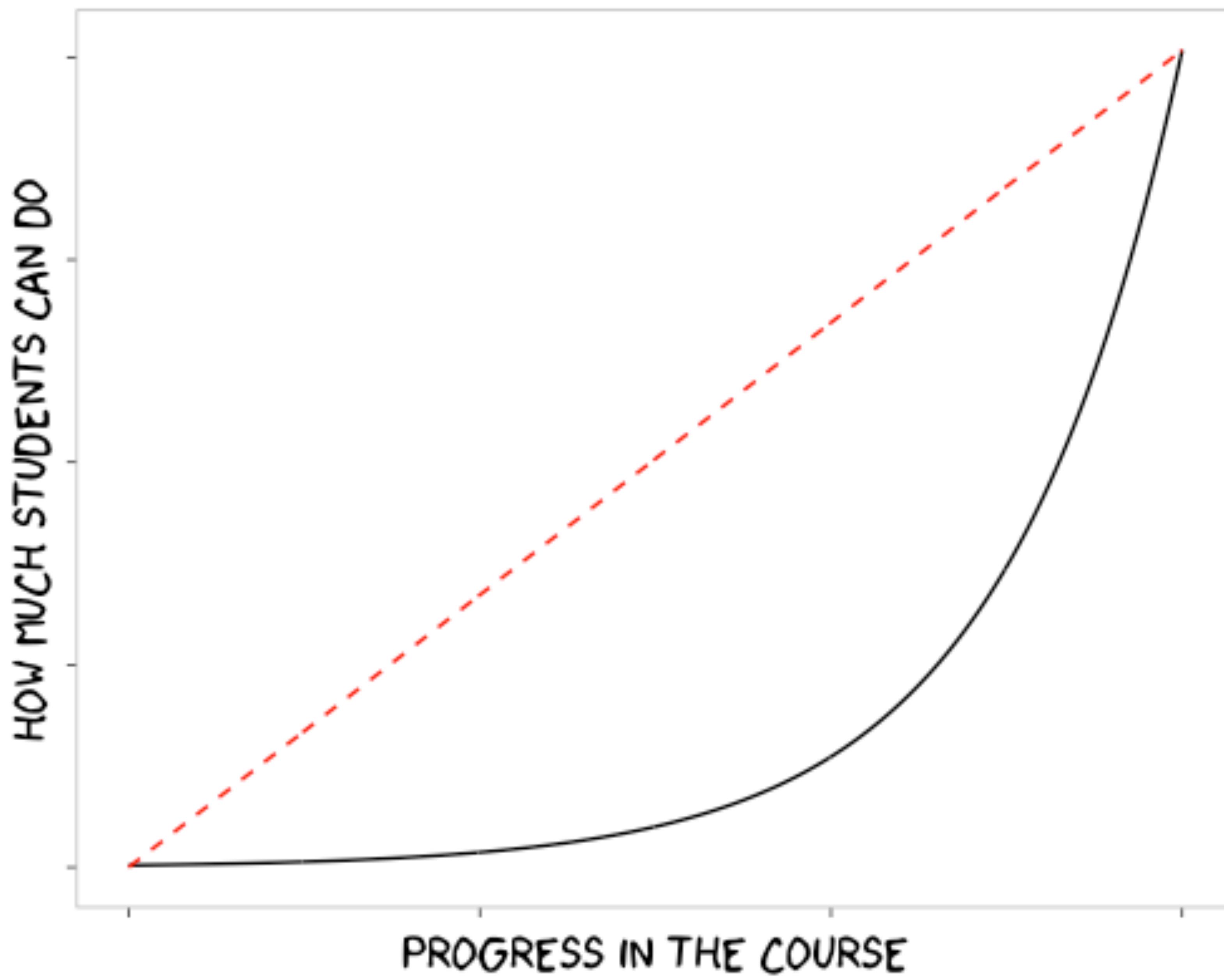


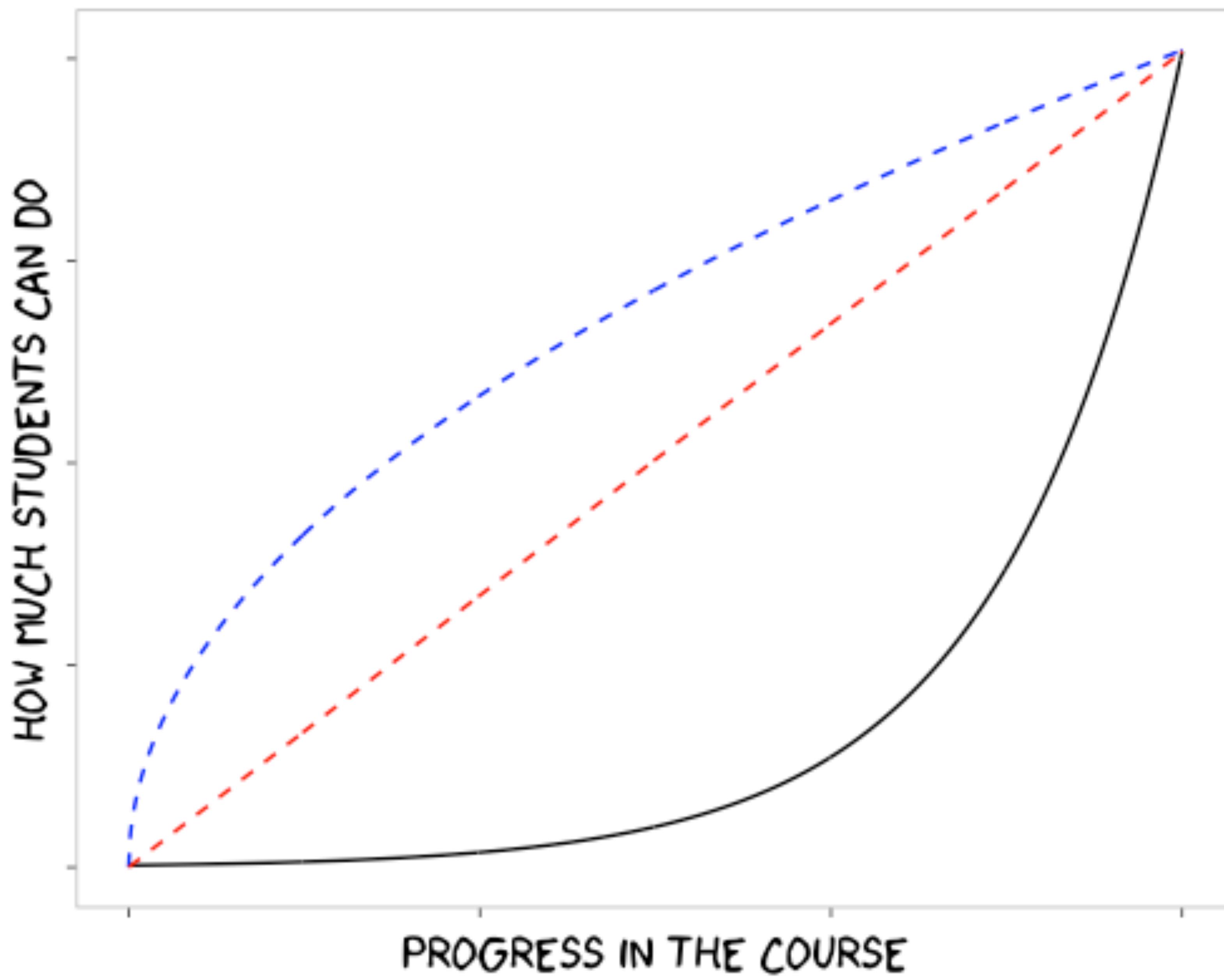
PROGRESS IN THE COURSE

1



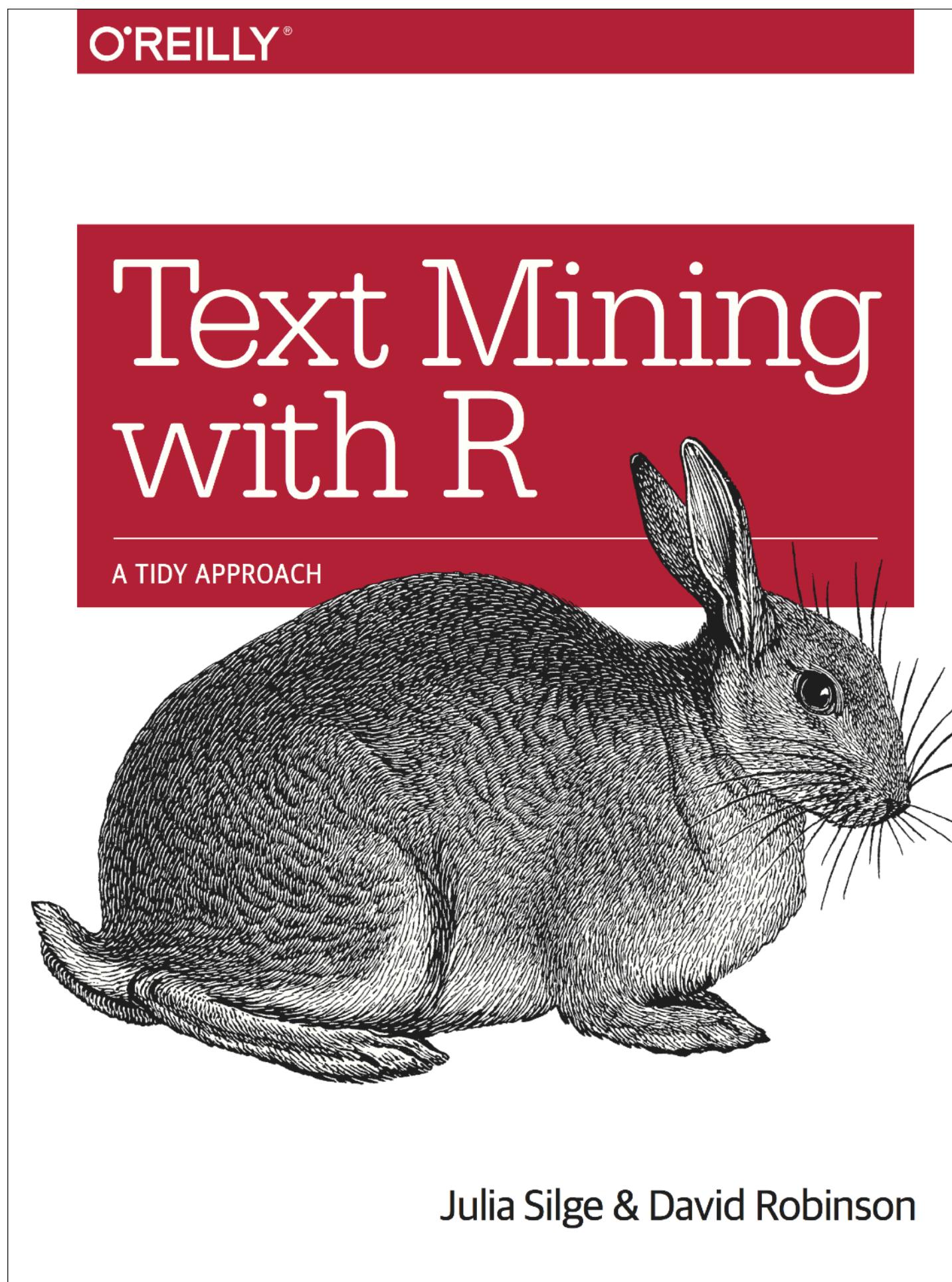




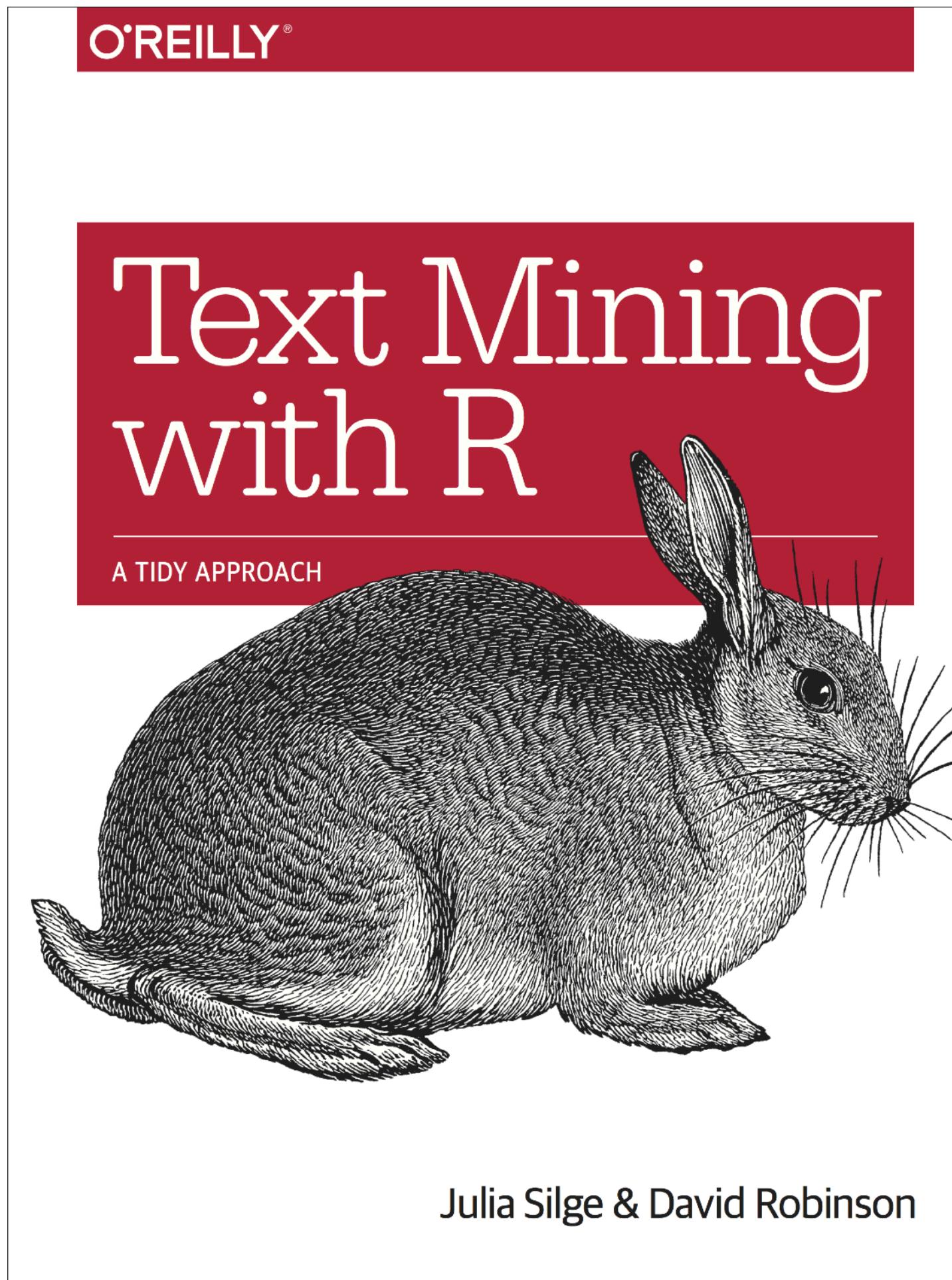


**“Students won’t get very far with  
just the tidyverse”**

**“Students won’t get very far with  
just the tidyverse”**



**“Students won’t get very far with  
just the tidyverse”**



**Not a single loop or conditional**

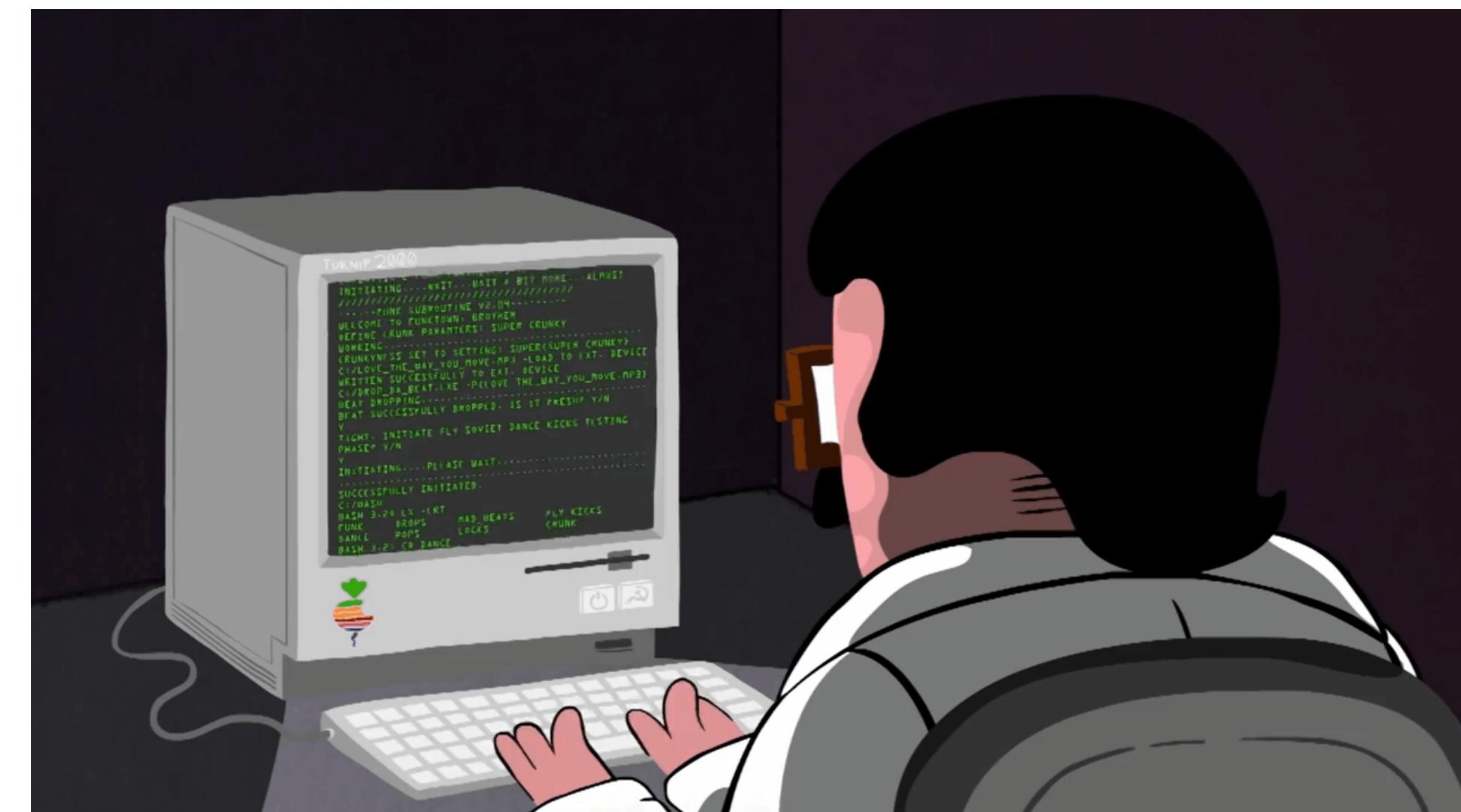
**When *shouldn't* you teach the  
tidyverse first?**

Have goals for what you want your students to do,  
and start them doing it as early as possible.

Have **goals** for what you want your students to do,  
and start them doing it as early as possible.

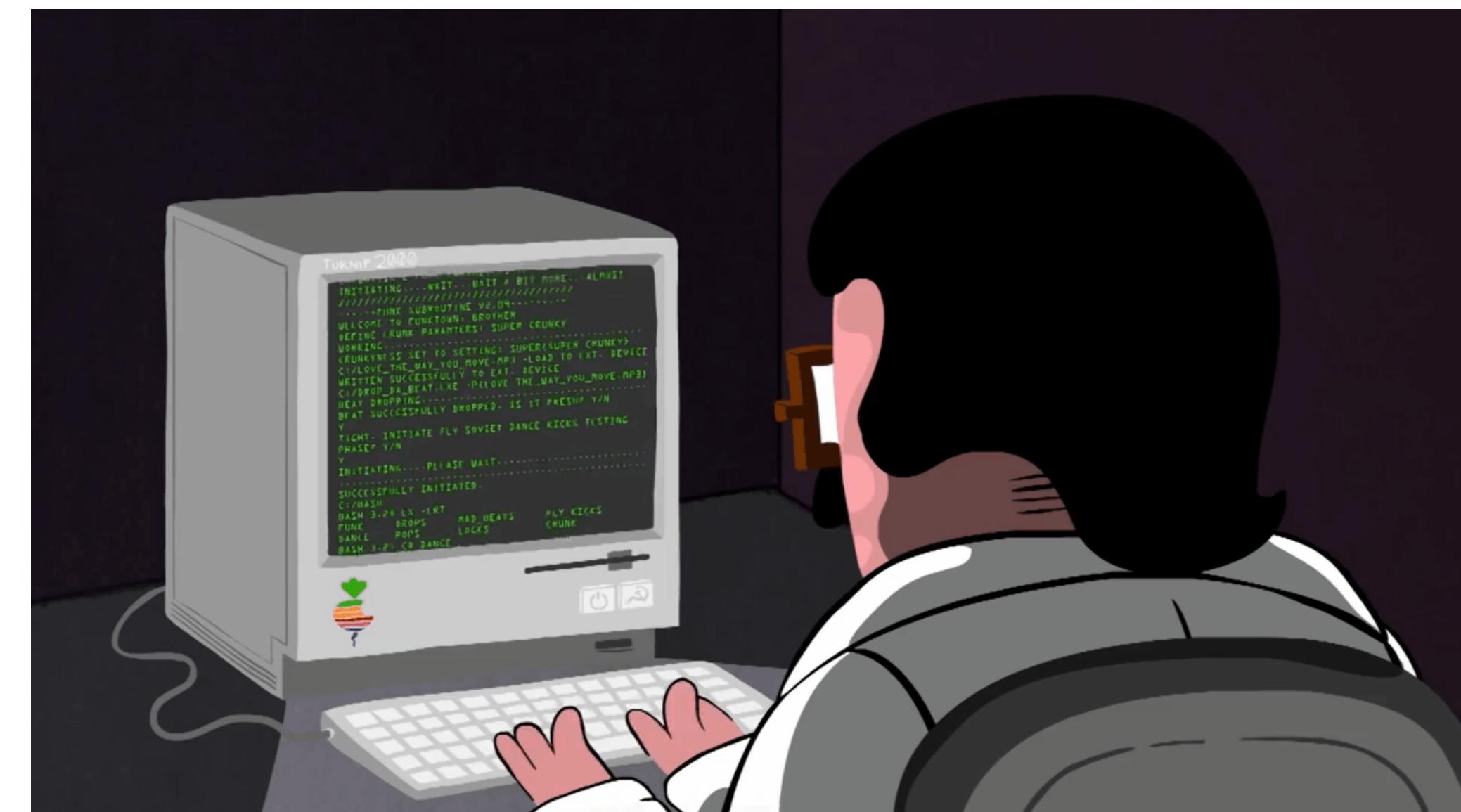
What goals could students have besides understanding data?

# Programming



What goals could students have besides understanding data?

# Programming



# What goals could students have besides understanding data?

## Programming

- Why choose R?



# What goals could students have besides understanding data?

## Programming

- Why choose R?
- Are students as interested in writing packages as you think?



# What goals could students have besides understanding data?

Mathematics

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$

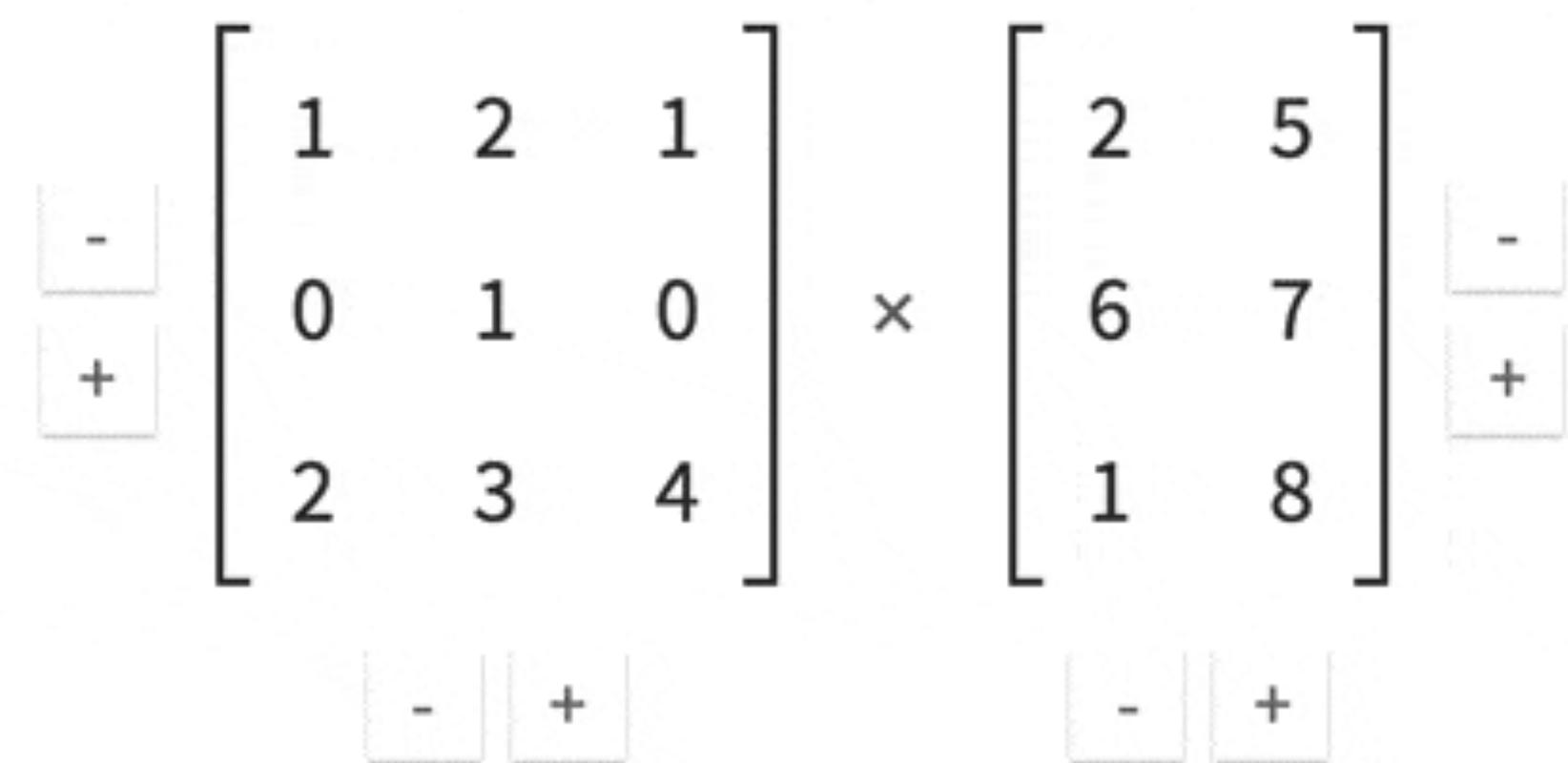
# What goals could students have besides understanding data?

Mathematics

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$

# What goals could students have besides understanding data?

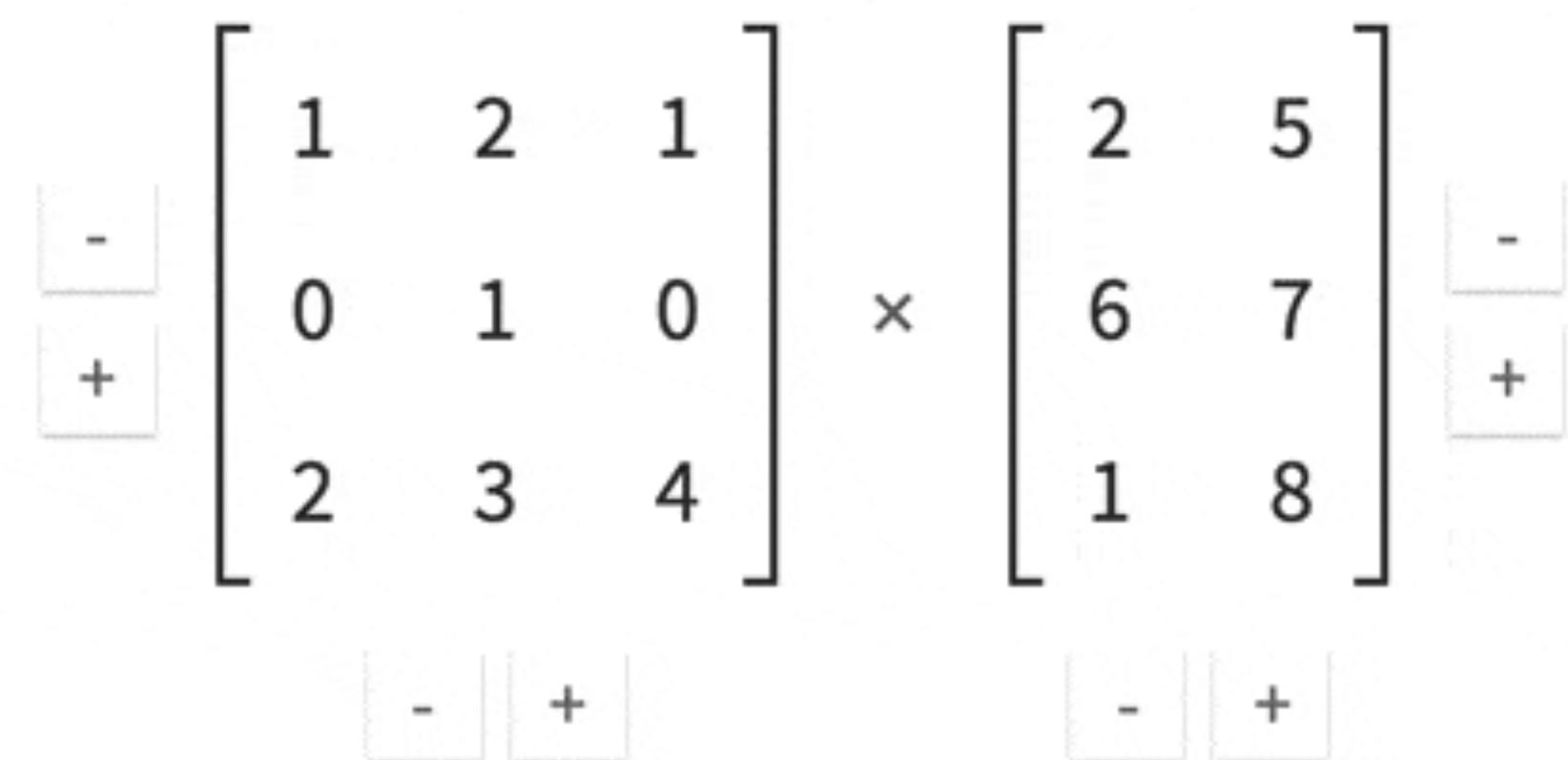
Mathematics

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$


- Teach matrices before data frames

# What goals could students have besides understanding data?

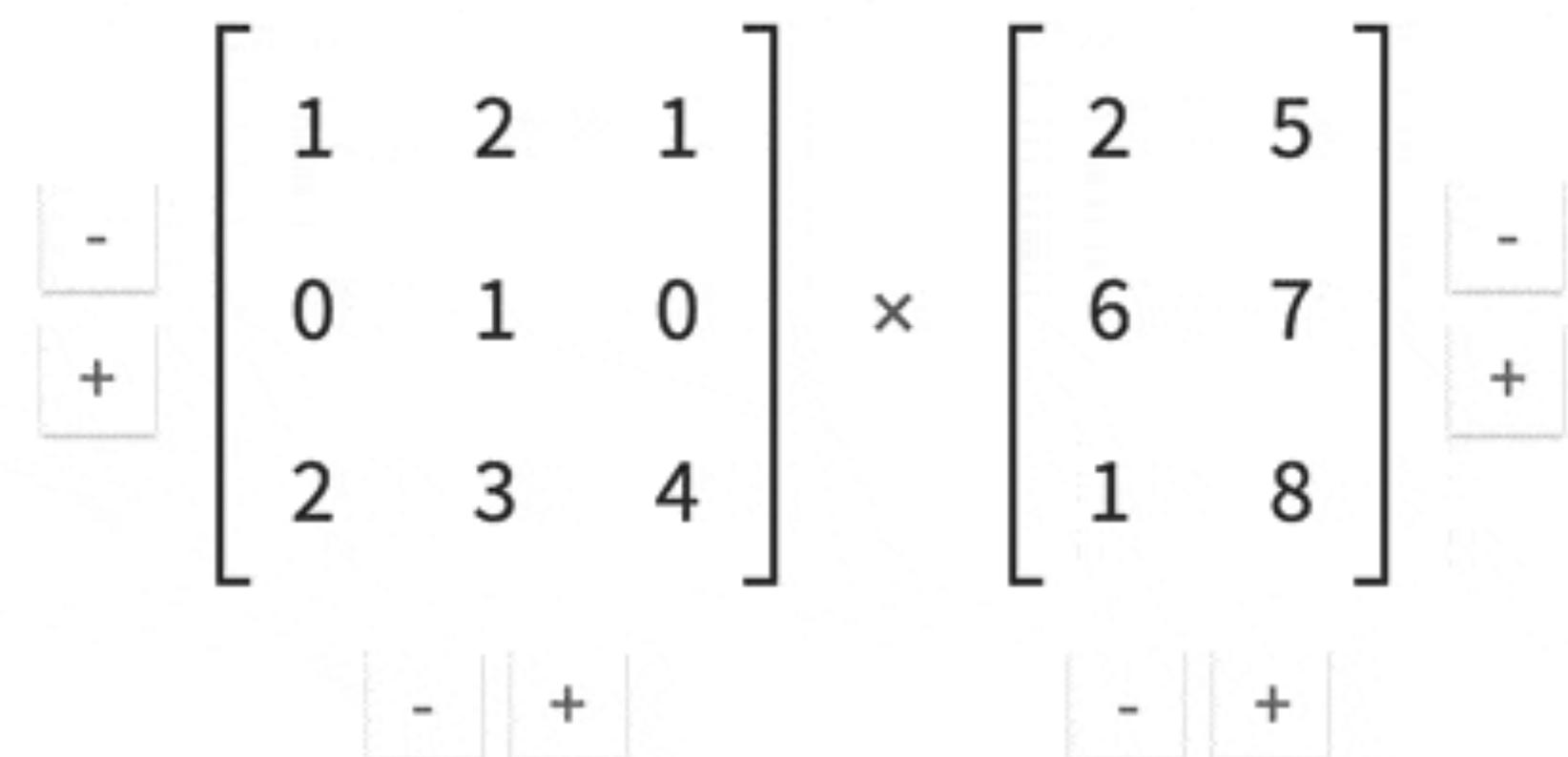
## Mathematics

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$


- Teach matrices before data frames
- Teach linear algebra before relational algebra (SQL)

# What goals could students have besides understanding data?

## Mathematics

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$


- Teach matrices before data frames
- Teach linear algebra before relational algebra (SQL)
- Teach the abstract before the concrete

# What's next?



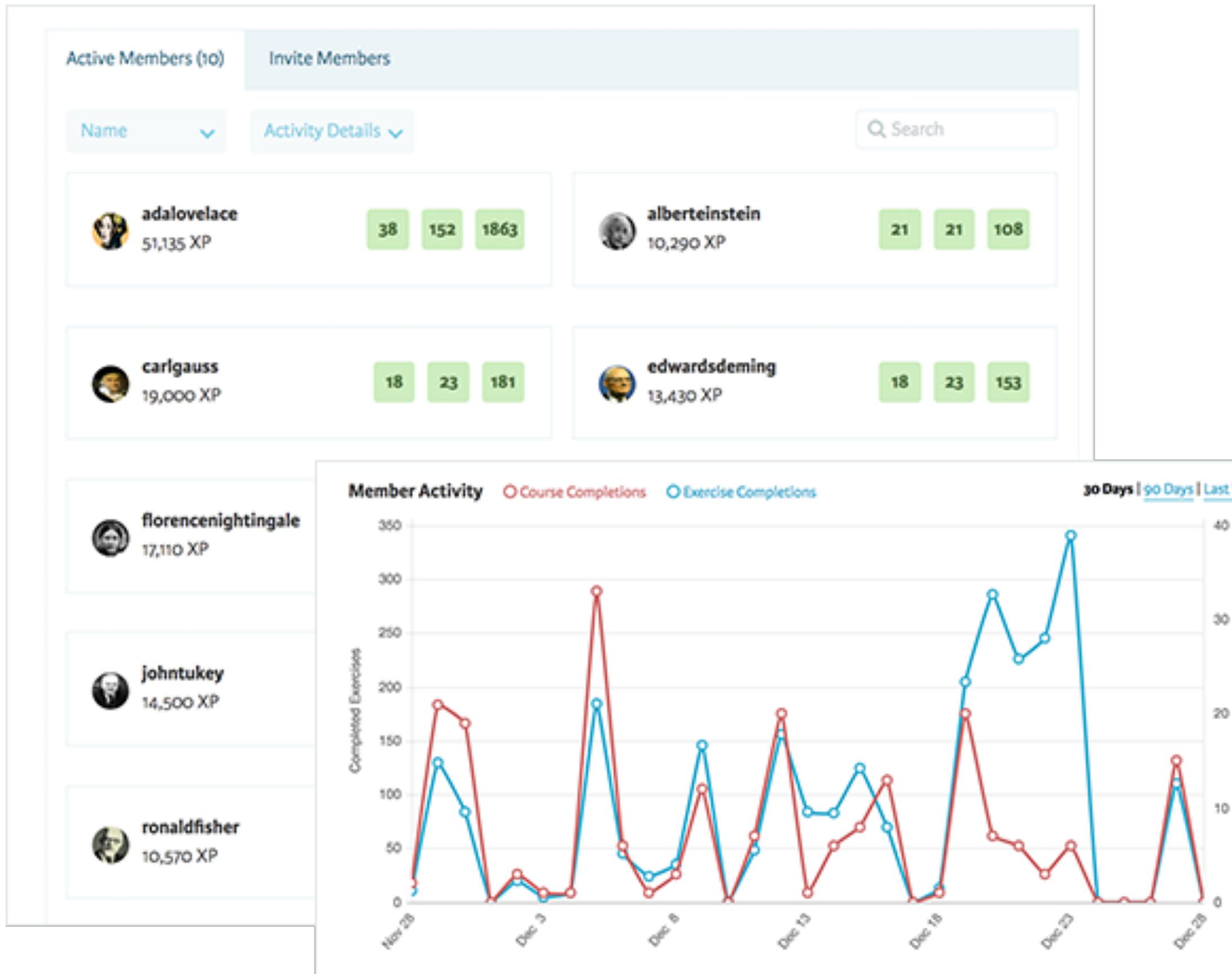
DataCamp

# What's next?



**Data-driven curriculum development**

# Want your team to learn the tidyverse?



## DataCamp for Business

# Want to teach data science to a huge audience?

## Apply to become an instructor

Your First Name\*

Your Last Name\*

Your Email\*

What are you interested in teaching?\*

[datacamp.com/create](https://www.datacamp.com/create)



LinkedIn Profile



Twitter Handle



GitHub Account



Personal Website

# Thank you



@drob

[www.varianceexplained.org](http://www.varianceexplained.org)

- RStudio
  - Hadley Wickham
  - Joe Cheng
  - Garrett Grolemund
  - Anne Carome
  - Julia Silge
  - Chester Ismay

VARIANCE EXPLAINED

ABOUT ME   POSTS   R COURSE   INTRODUCTION TO EMPIRICAL BAYES



David Robinson

*Data Scientist at Stack Overflow, works in R and Python.*

Email  
 Twitter  
 Github  
 Stack Overflow

Subscribe

This is the homepage and blog of David Robinson, a Data Scientist at Stack Overflow. For more about me, [see here](#).

## Recent Posts

---

### Teach the tidyverse to beginners

*July 05, 2017*

An argument for teaching R packages like dplyr and tidyr as the first part of a data science course.

### Two years as a Data Scientist at Stack Overflow

*June 22, 2017*

Looking back at my second year at the first job I've had outside academia.

### Words growing or shrinking in Hacker News titles: a tidy analysis

*June 08, 2017*

An analysis of one million Hacker News titles, and what topics and technologies are changing in frequency over time.

### Slides, videos, and tweets from the 2017 New York R Conference

*May 22, 2017*

Some notes from the 2017 New York R Conference, and slides and video from my talk.