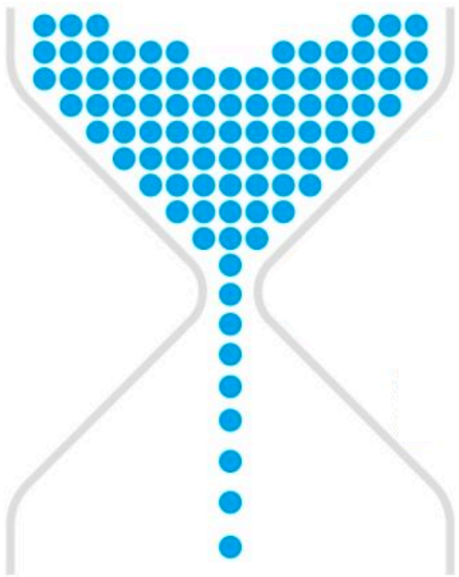
Basic R Programming

Mustafa Anil Kocak March 4th 2024

Why learn statistical programming?



Data is the new commodity

- ☐ We have better access to data than most of the world
- ☐ Not using it efficiently is leaving money on the table!

Analysis/analysts are typical bottlenecks

- ☐ Analysts are working one experiment at a time
- ☐ Understanding the experiment, meta-data and questions takes time

Most of the analyses are …

- ☐ straightforward once you got the experimental thinking
- ☐ can be performed with a limited vocabulary!

Most analyses are doable with limited vocabulary

- ☐ **Visualization (ggplot)**
- ☐ **Vectors, lists and data tables (tibbles)**
- ☐ **Libraries and how to get help**
- ☐ **Loading/Saving datasets:** data.table::fread, readRDS(), write_csv()
- ☐ **Taking a peek:** head(), tail(), summary(), names(),View(), nrow(), table(), dim(), summarize(), describe()
- ☐ **Data manipulation:** select(), distinct(), filter(), group_by(), ungroup(), summarize(), summarize_all(), arrange(), count()
- ☐ **Joins:** left/right/inner/full/anti/semi_join()
- ☐ **Extra:** top_n(), is.finite(), is.na(), %in% , paste0(), max(), min(), mean(), median(), sd(), mad(), quantile(), is.finite(), toupper(), word(), n(), %>%, write_csv(), bind_rows(), c(), psych::pairs.plots(), pheatmap(), reorder(), *taiga (?)*

What is our objective?

How to draw an owl

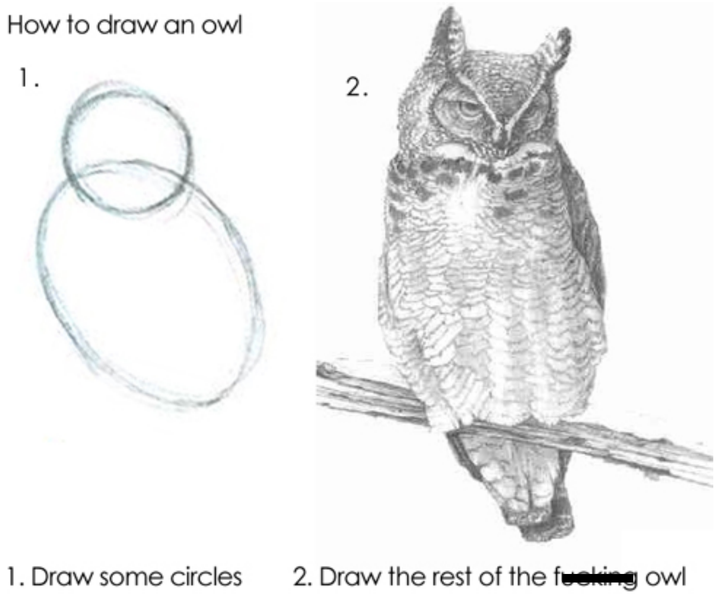
1.

1. Draw some circles

2.

2. Draw the rest of the ~~fucking~~ owl

What is our objective?

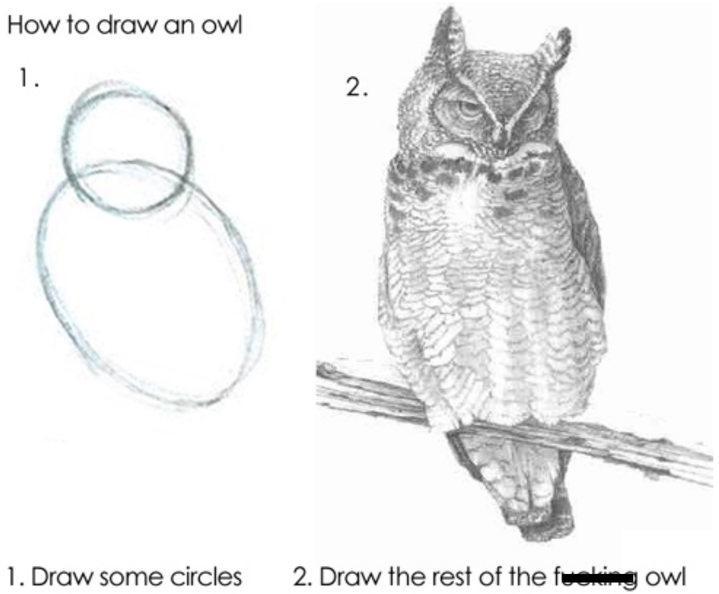


Knowledge

Skill

Attitude

What is our objective?



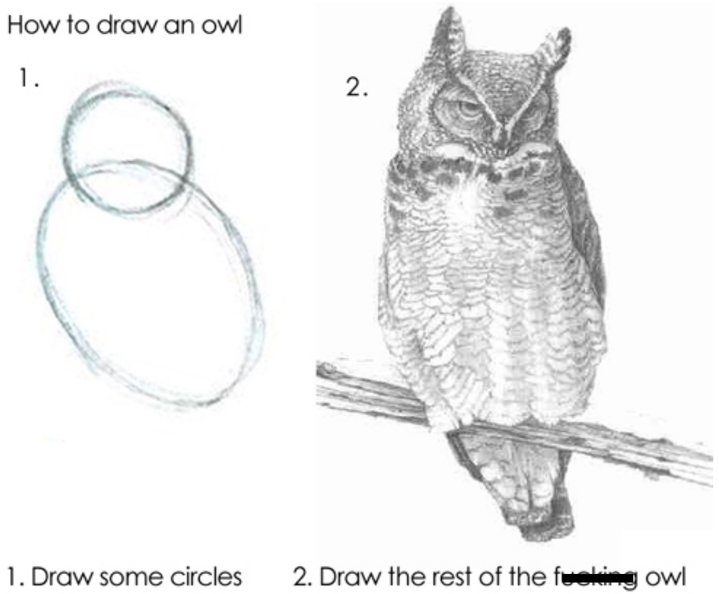
Knowledge

- ☐ What R is good for; tutorials, resources, lectures

Skill

Attitude

What is our objective?



Knowledge

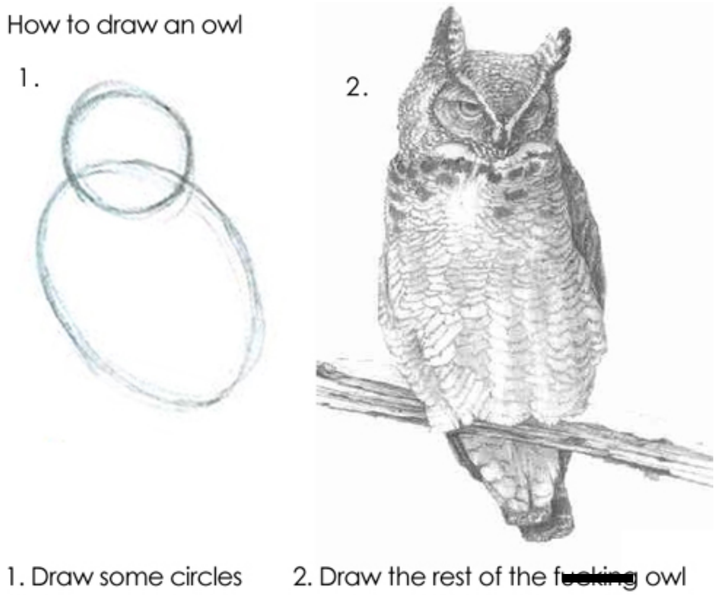
- ☐ What R is good for; tutorials, resources, lectures

Skill

- ☐ Using what you learned in your projects next week

Attitude

What is our objective?



Knowledge

- ☐ What R is good for; tutorials, resources, lectures

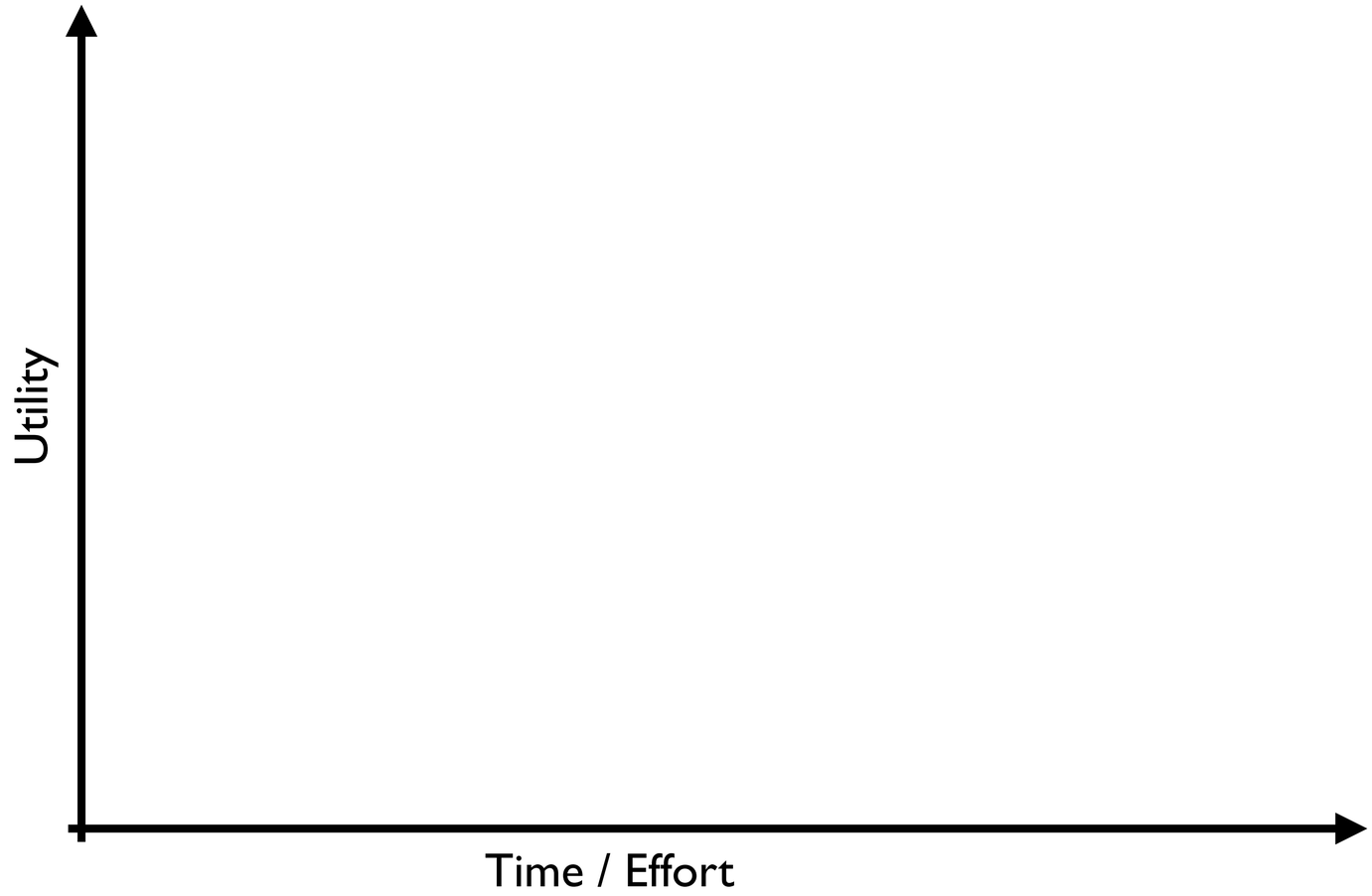
Skill

- ☐ Using what you learned in your projects next week

Attitude

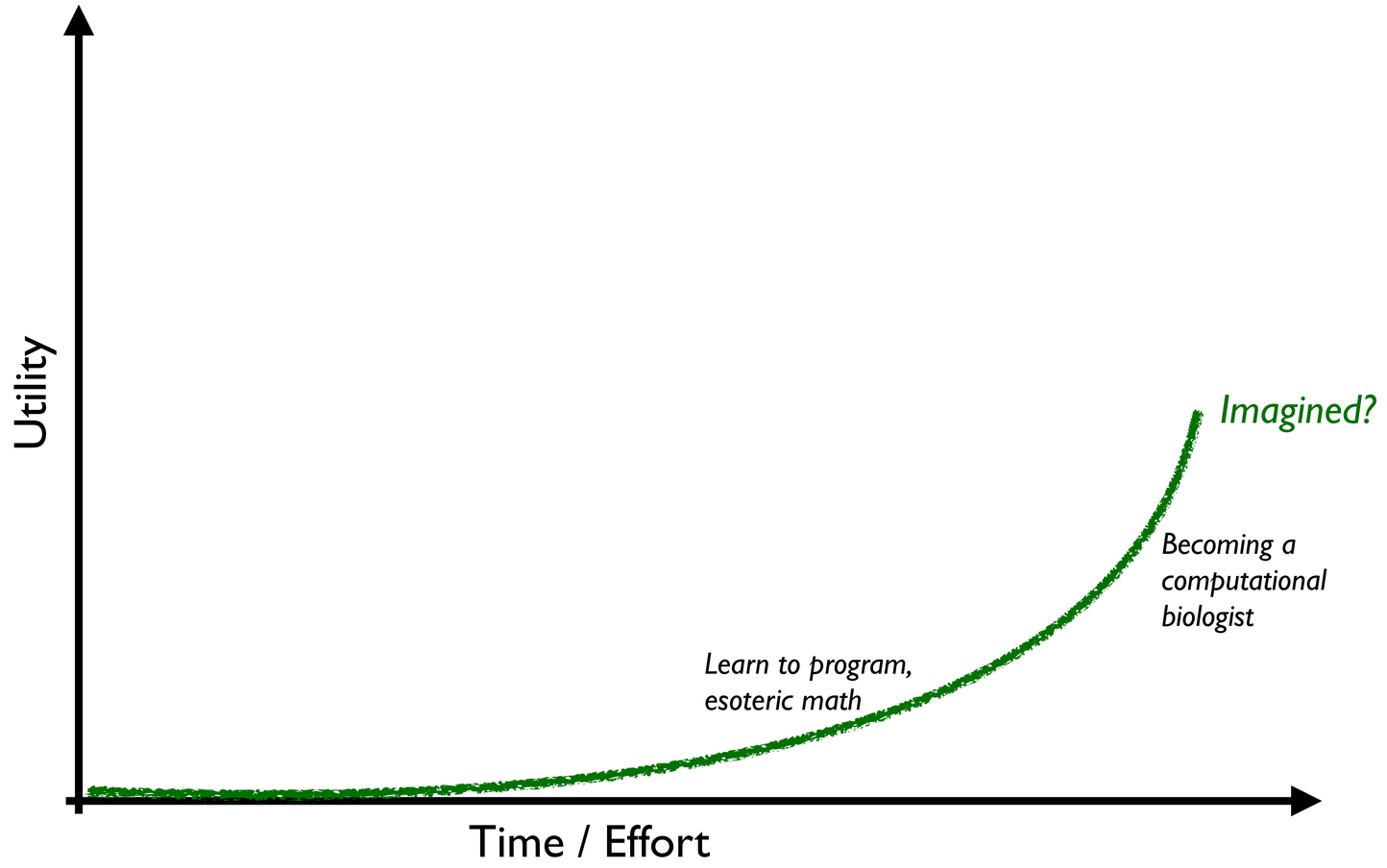
- ☐ Keep using what you learned in the bootcamp
- ☐ Having joy through playing with data!

Our View of the Learning Curve



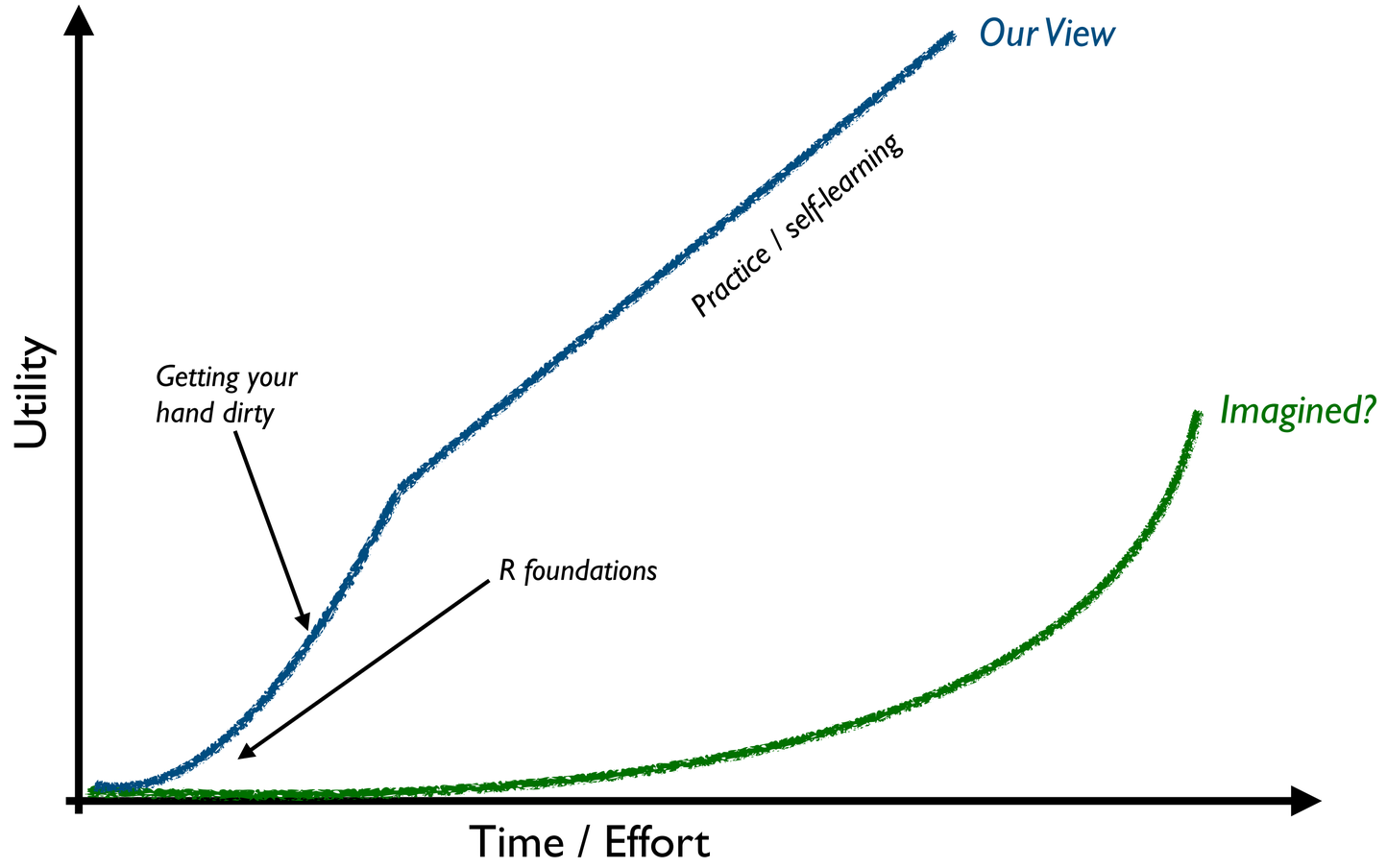
Credit: James McFarland

Our View of the Learning Curve



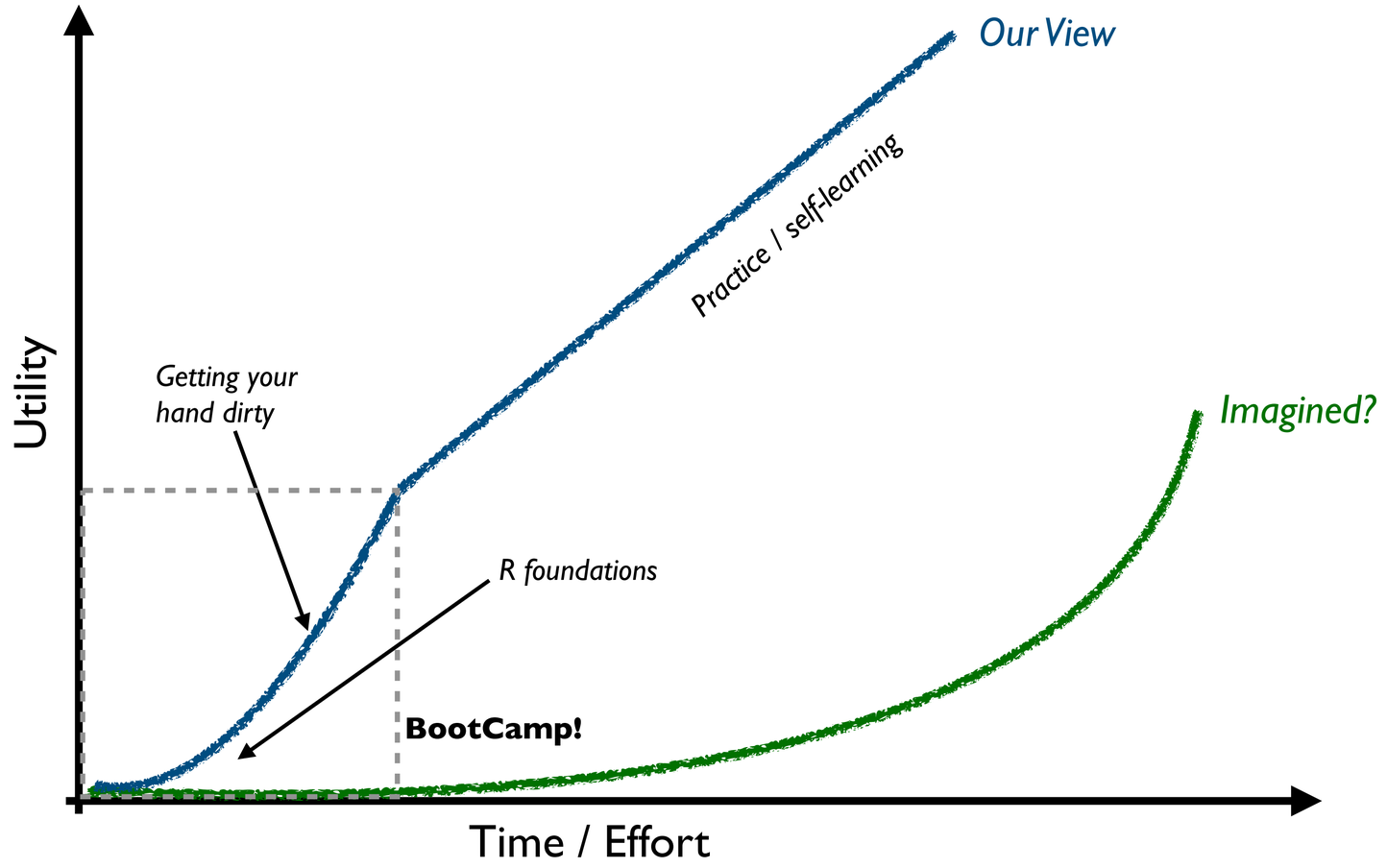
Credit: James McFarland

Our View of the Learning Curve



Credit: James McFarland

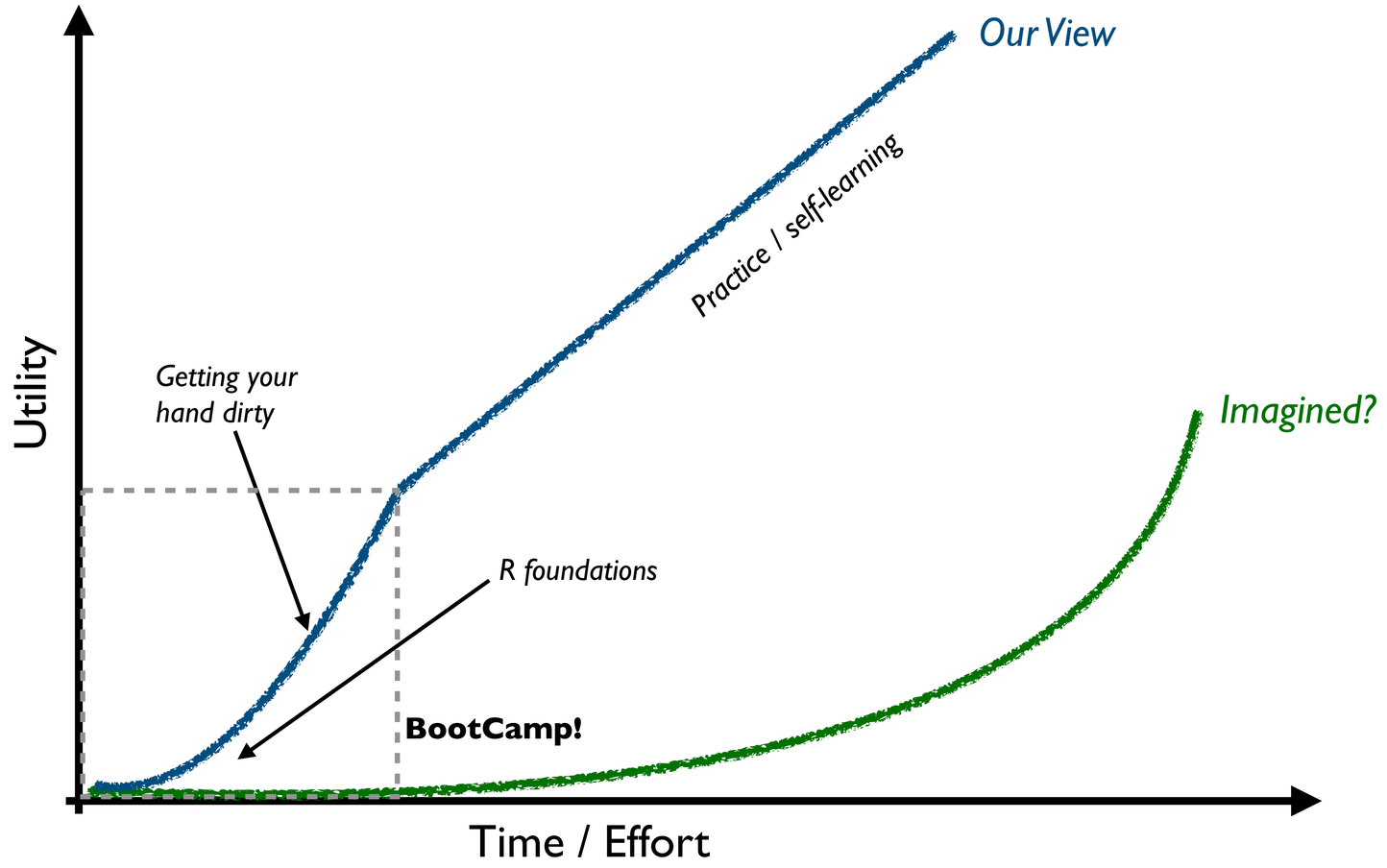
Our View of the Learning Curve



Credit: James McFarland

Our View of the Learning Curve

I need your help!

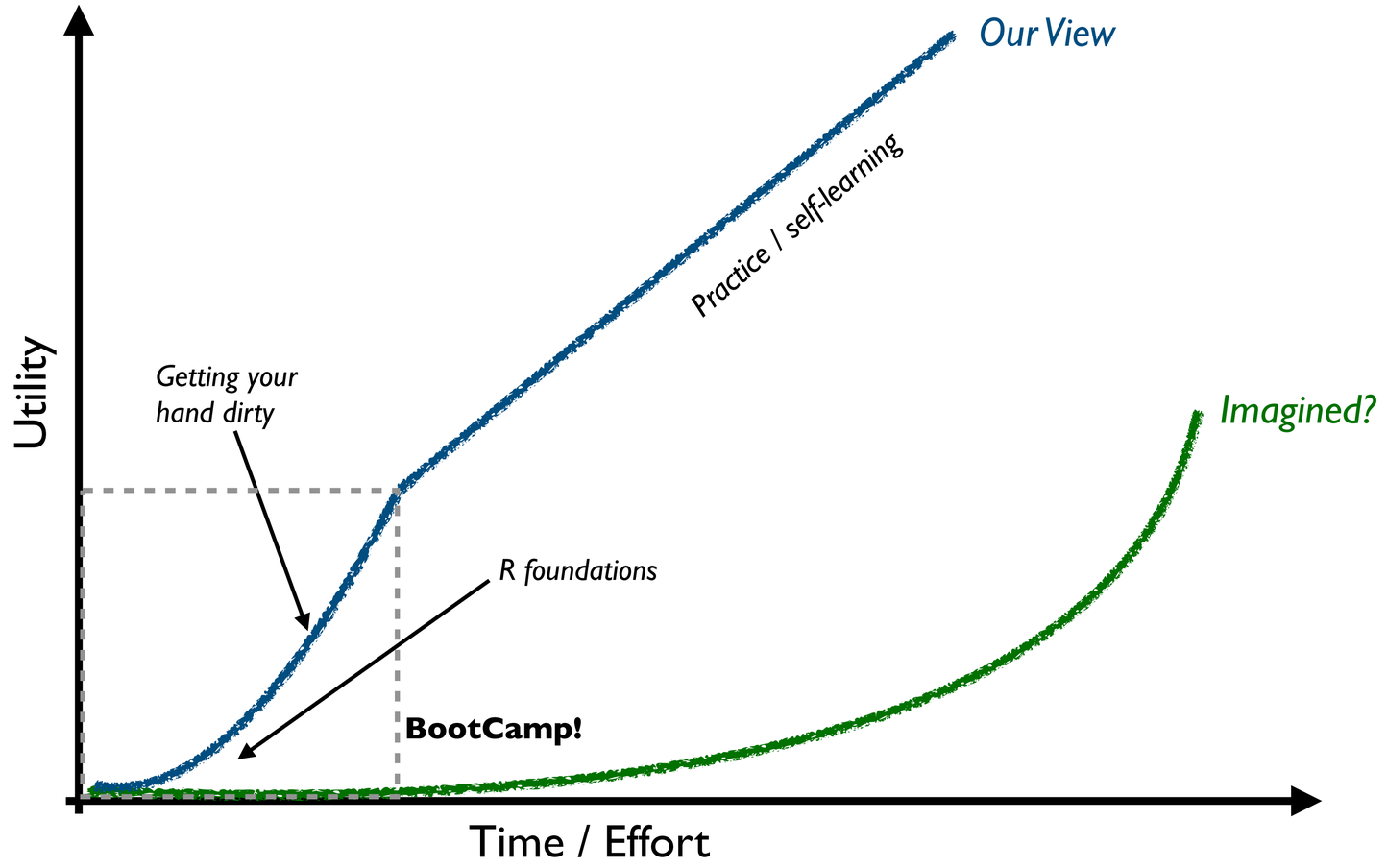


Credit: James McFarland

Our View of the Learning Curve

I need your help!

- [ ] This is my ~~first~~ second bootcamp

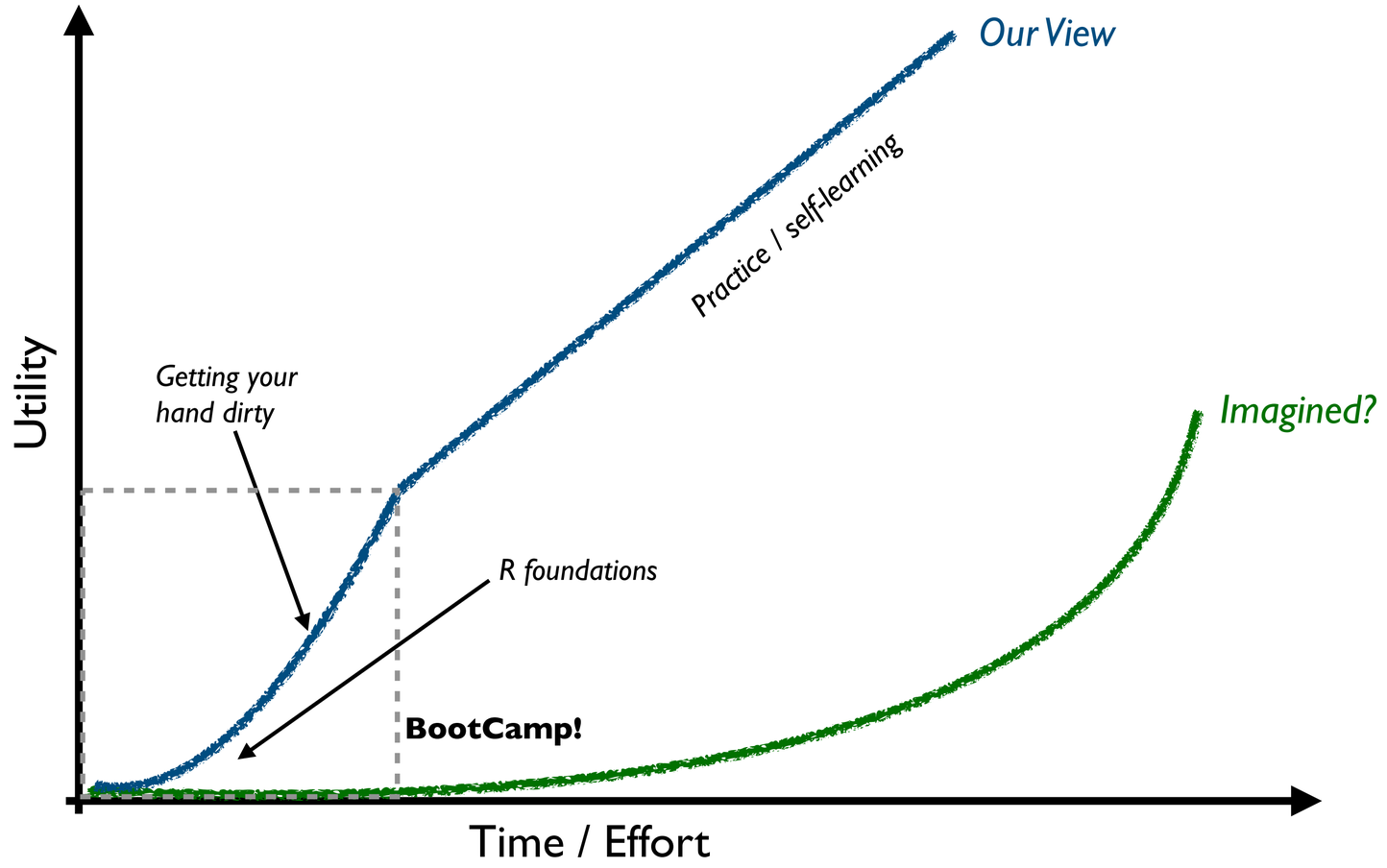


Credit: James McFarland

Our View of the Learning Curve

I need your help!

- ☐ This is my ~~first~~ second bootcamp
  - ☐ And I am nervous :)

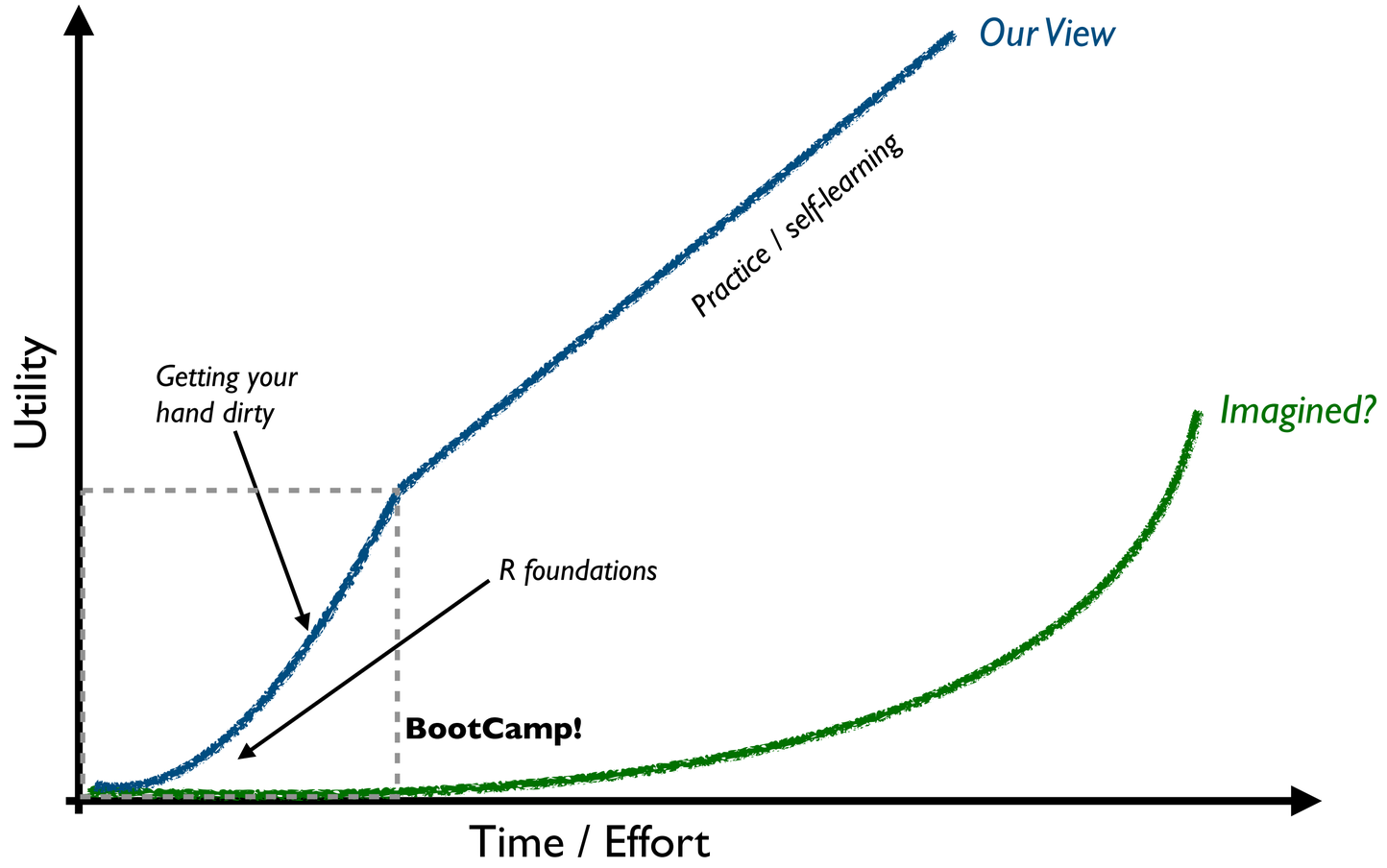


Credit: James McFarland

Our View of the Learning Curve

I need your help!

- ☐ This is my ~~first~~ second bootcamp
  - ☐ And I am nervous :)
- ☐ We will figure it out together

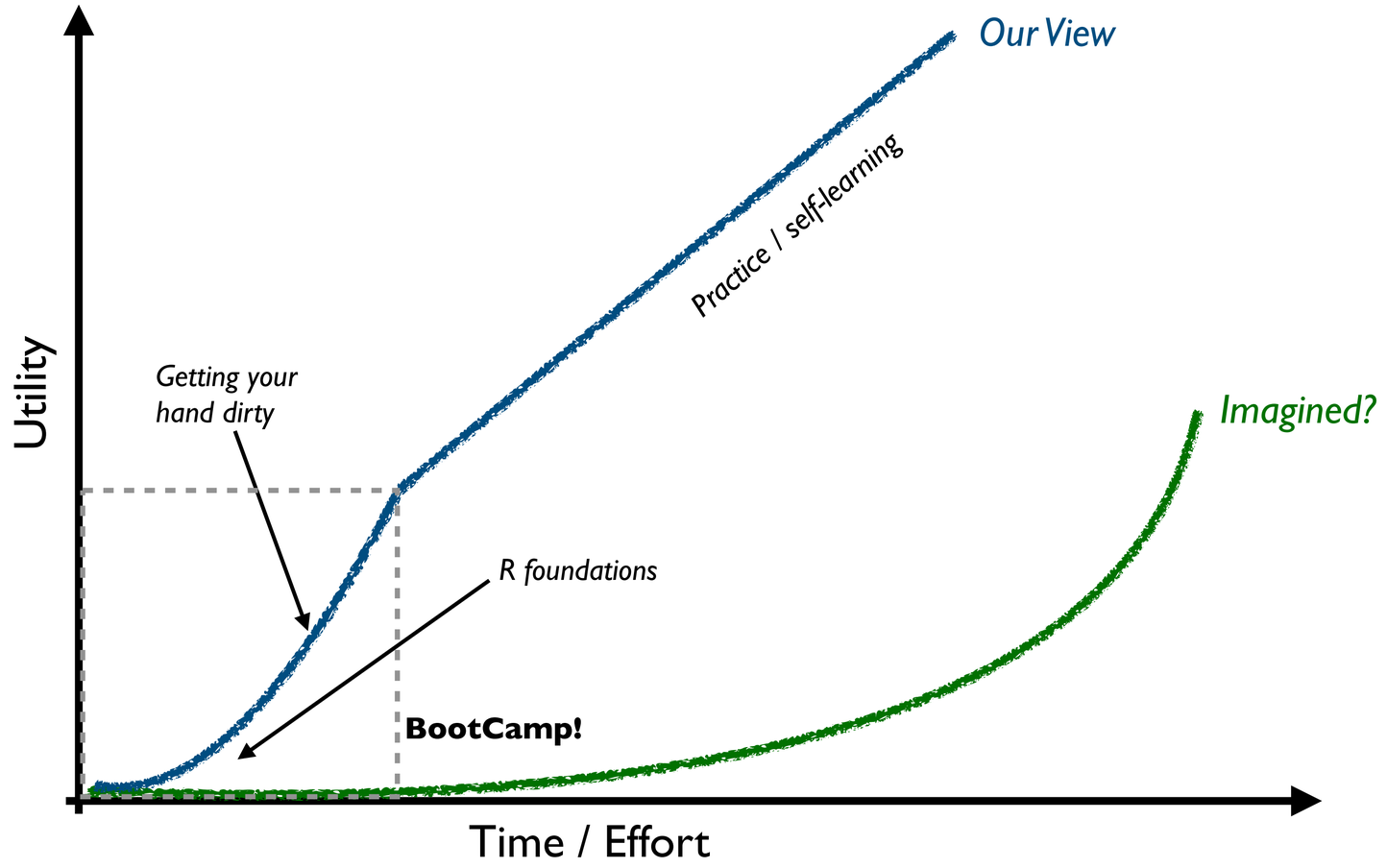


Credit: James McFarland

Our View of the Learning Curve

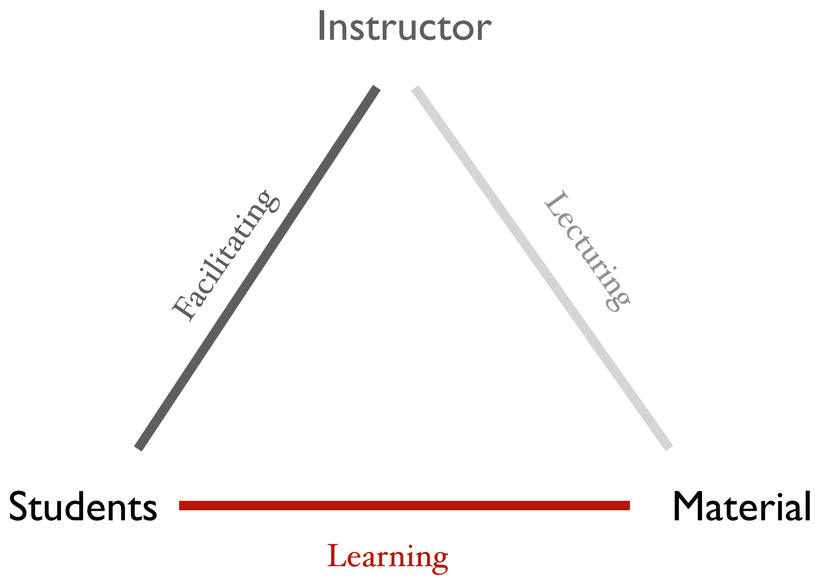
I need your help!

- ☐ This is my ~~first~~ second bootcamp
  - ☐ And I am nervous :)
- ☐ We will figure it out together
- ☐ Please ask questions and give feedback



Credit: James McFarland

Approach

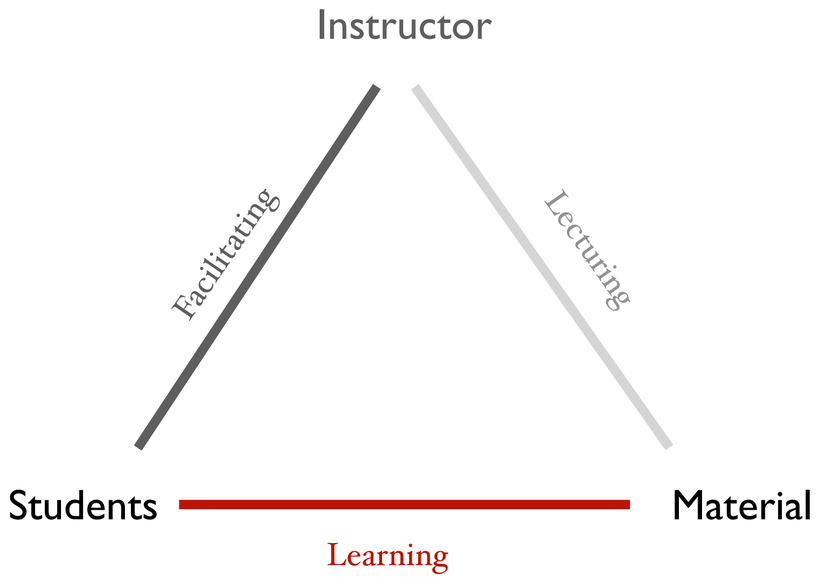


Credit: Jean-Luc Dumont

Approach

Emphasis on the practical skills

- ☐ Basics, visualization, wrangling, …



Credit: Jean-Luc Dumont

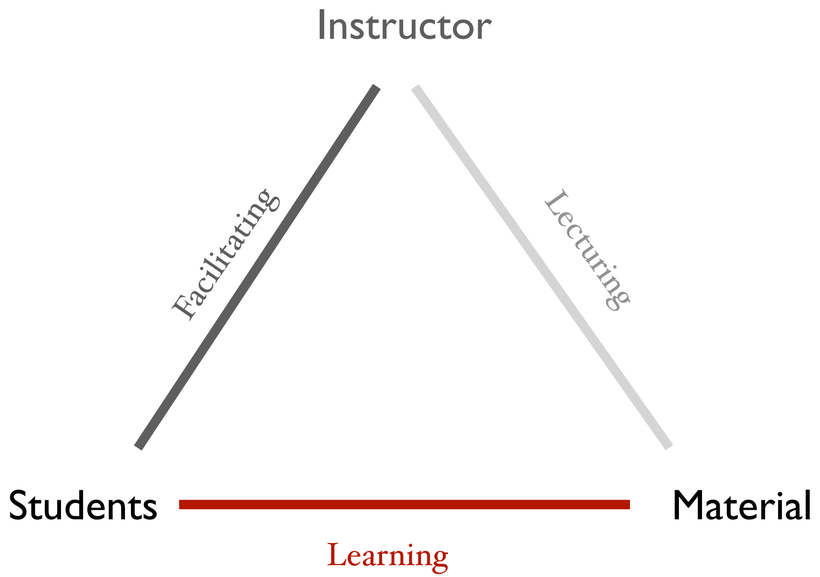
Approach

Emphasis on the practical skills

- ☐ Basics, visualization, wrangling, …

Pointing out resources but mostly giving the 'need-to-know'

- ☐ Think about what else would you like to do with R and data?



Credit: Jean-Luc Dumont

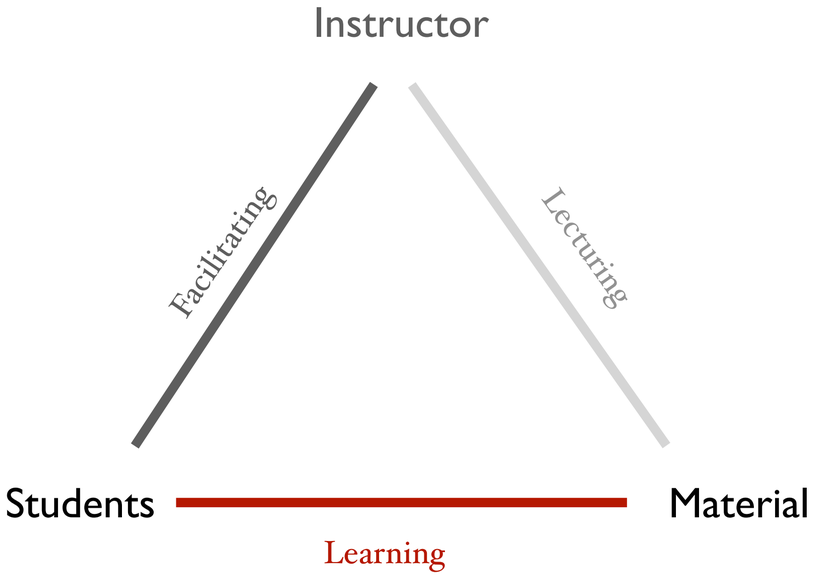
Approach

Emphasis on the practical skills

- ☐ Basics, visualization, wrangling, …

Pointing out resources but mostly giving the 'need-to-know'

- ☐ Think about what else would you like to do with R and data?



Credit: Jean-Luc Dumont

Approach

Emphasis on the practical skills

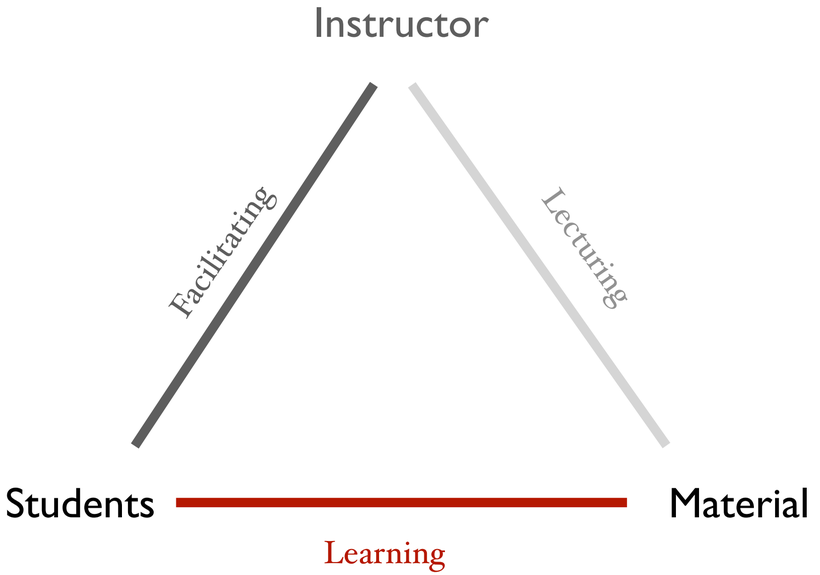
- ☐ Basics, visualization, wrangling, …

Pointing out resources but mostly giving the 'need-to-know'

- ☐ Think about what else would you like to do with R and data?

Slides are reminders, we will write code during the sessions.

- ☐ Learn from your experience and peers!



Credit: Jean-Luc Dumont

Course Content

First Encounter with R Programming!

- [ ] R studio, scripts, markdowns
- [ ] Functions, packages, help pages
- [ ] R Objects, notation & modifying values
- [ ] Data structures, programming

Introduction to Tidyverse

- [ ] Visualization
- [ ] Wrangling
- [ ] Exploratory Data Analysis

Miscellaneous

- [ ] Practice and case studies
- [ ] Advanced data visualization
- [ ] Machine learning examples

Course Content

First Encounter with R Programming!

- ☐ **R studio, scripts, markdowns**
- ☐ **Functions, packages, help pages**
- ☐ **R Objects, notation & modifying values**
- ☐ **Data structures, programming**

Introduction to Tidyverse

- ☐ Visualization
- ☐ Wrangling
- ☐ Exploratory Data Analysis

Miscellaneous

- ☐ Practice and case studies
- ☐ Advanced data visualization
- ☐ Machine learning examples

A hands-on introduction to basic R

The Very Basics

R Objects & Notation

Basic Programming

First Encounter with with R Programming!

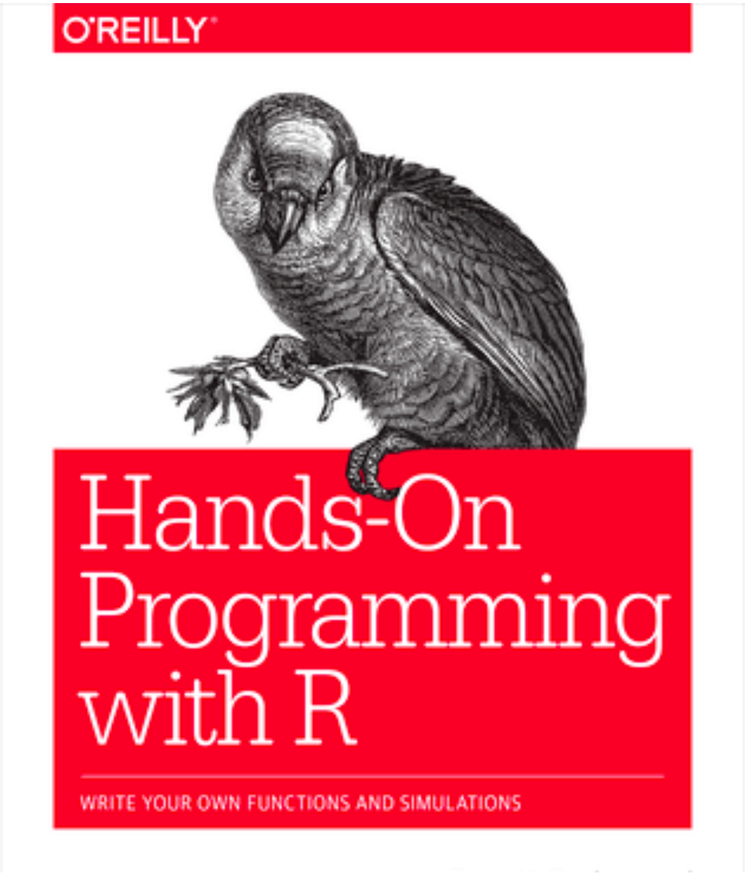
Freely available online:

- [ ] https://rstudio-education.github.io/hopr/

Little pedestrian, but great introduction.

We will skip chapters 8 and 10.

- [ ] Recommended for self-study
- [ ] Great basis, but we will significantly refine in the following sessions.



Garrett Grolemund

The Very Basics

User Interface

Objects

Functions

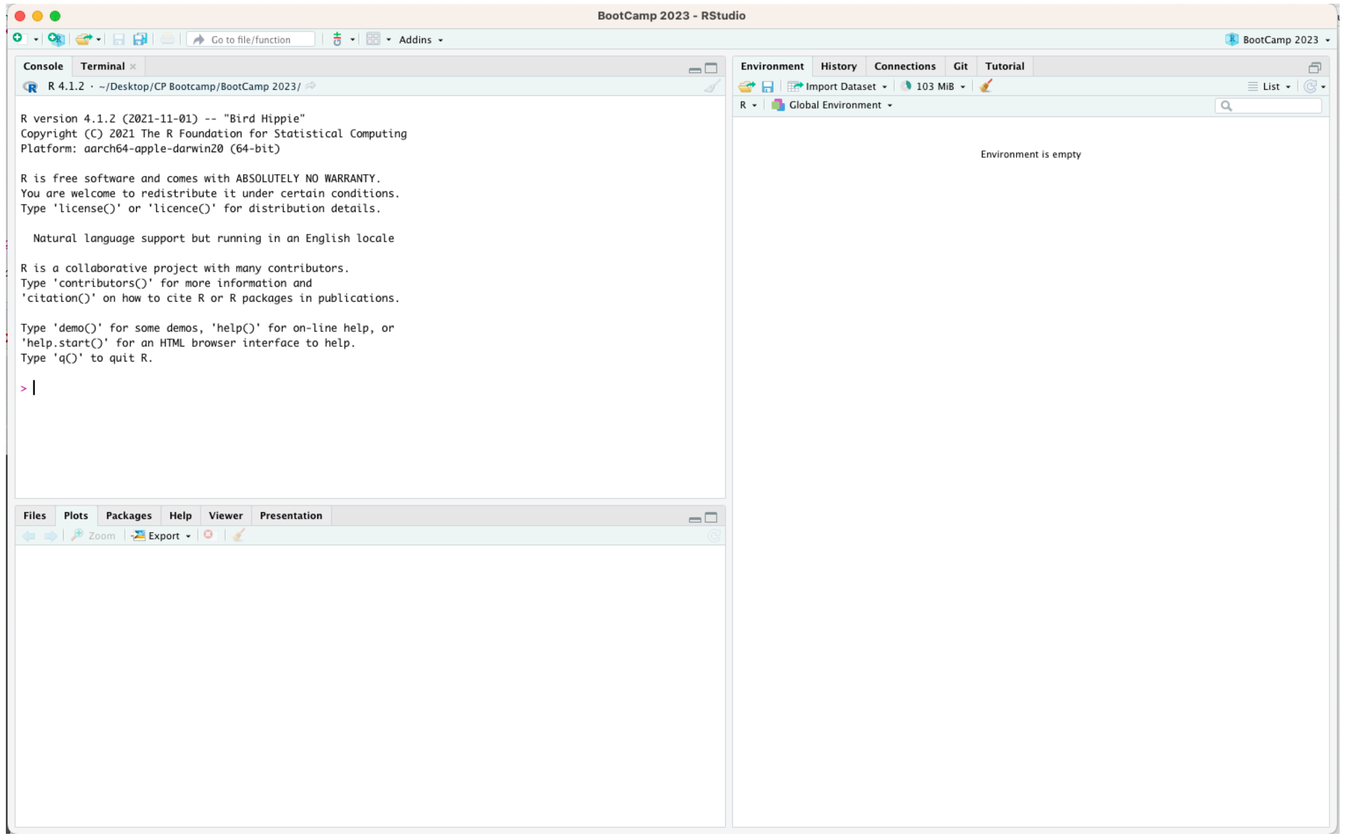
Writing your Own Functions

Packages

Scripts / Mark-downs

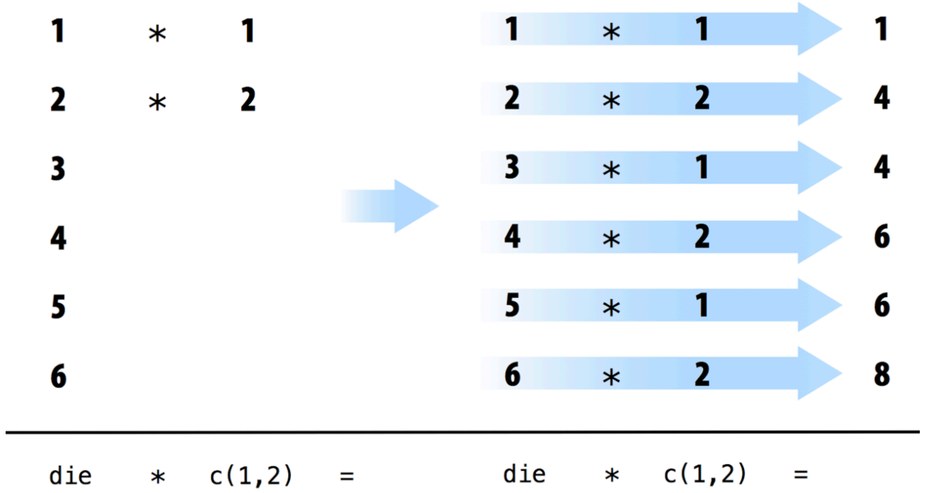
R Studio Interface

- [ ] Interface
- [ ] Projects
- [ ] R as a calculator
- [ ] Scripts and mark-downs
- [ ] Arithmetic operators



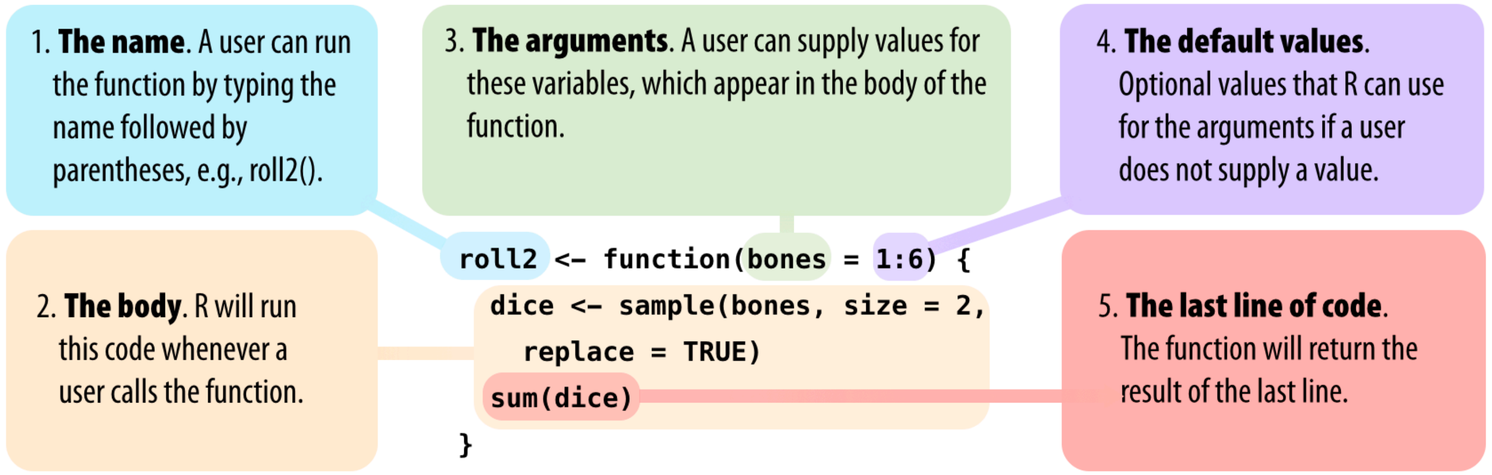
R Objects

- [ ] Variables
- [ ] Vectors
- [ ] Naming convention
- [ ] Vector recycling



Functions

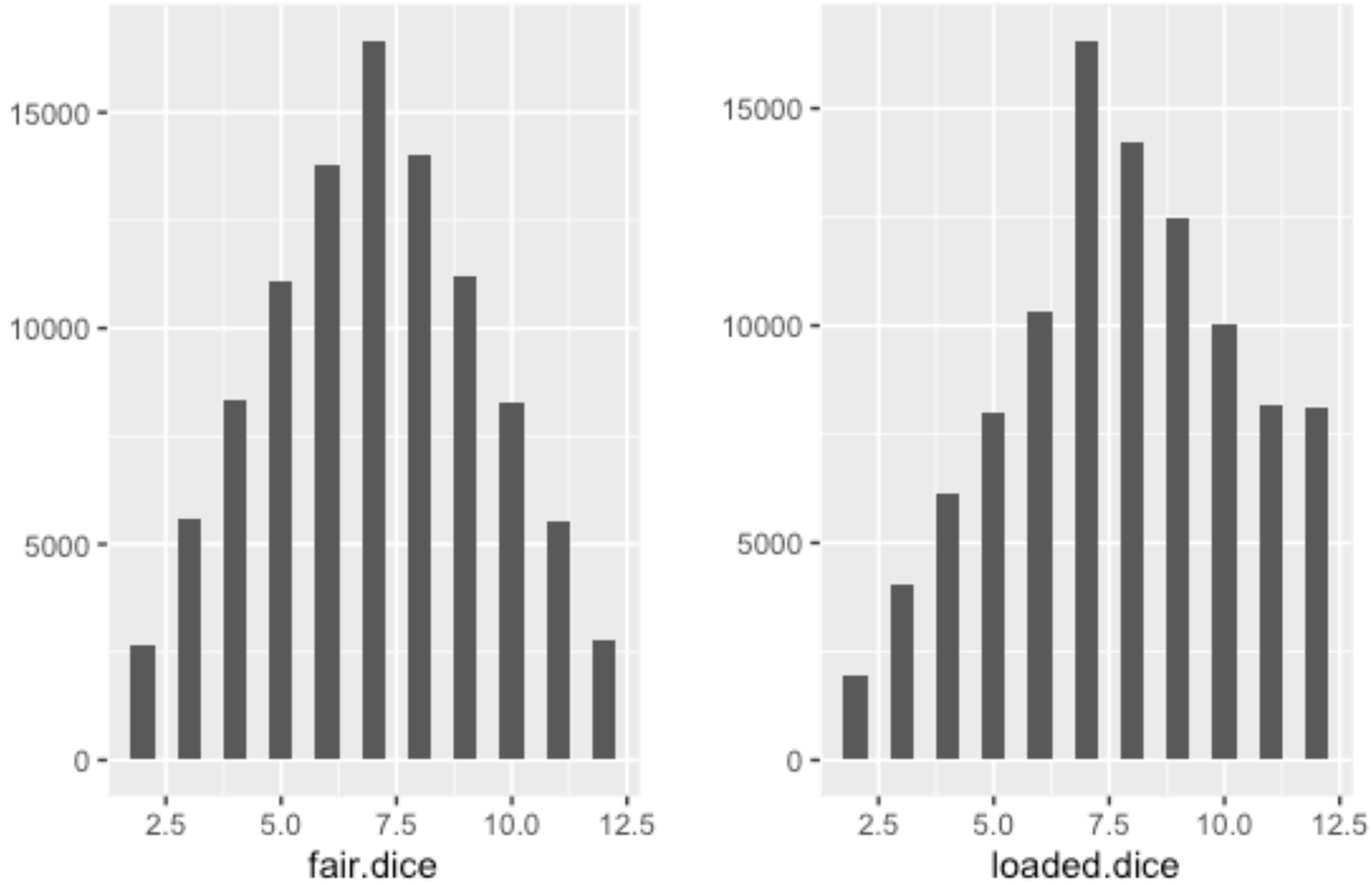
- [ ] Calling a function
- [ ] Arguments
- [ ] Help pages
- [ ] Writing a function
- [ ] Extract function



Packages

- [ ] Installing packages
- [ ] Loading packages
- [ ] Functions from a package
- [ ] Help pages re-visitied
- [ ] Misc:

  c / replicate / qplot / ctrl + c / pipe / plot_grid



R Objects & Notation

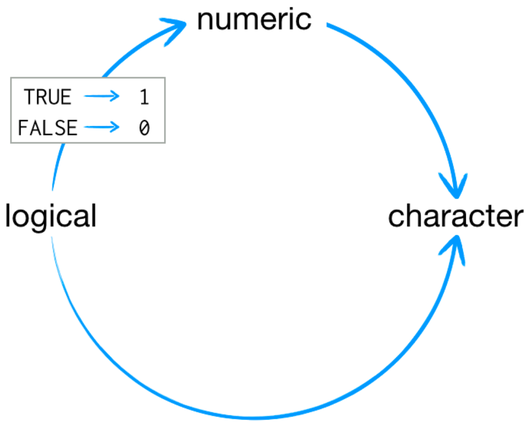
R Objects

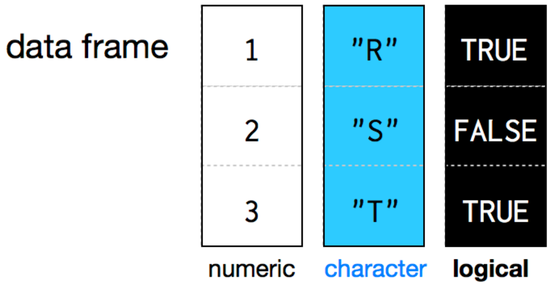
Slicing

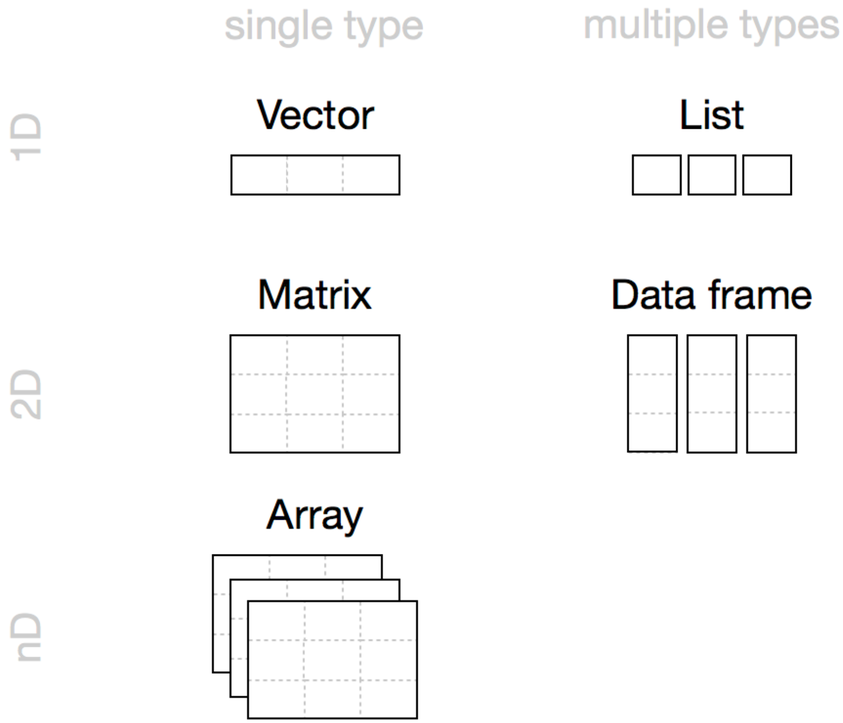
Modifying Values

R Objects

- [ ] Atomic vectors
- [ ] Basic types: **doubles**, integers, **characters**, **logicals**, complex, raw
- [ ] Coercion
- [ ] Matrices & arrays
- [ ] Attributes & class
- [ ] Lists & data frames
- [ ] Loading & saving data

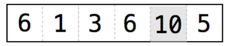






Slicing and Modifying Values

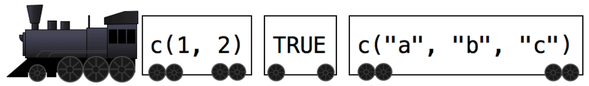
- [ ] Slicing datasets: [ ]
- [ ] $ and [[ ]] to slice lists
- [ ] Modifying values
  - [ ] Changing in-place
  - [ ] Logical subsetting
  - [ ] Missing values



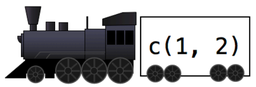
vec[5]

John	1940	guitar
Paul	1941	bass
George	1943	guitar
Ringo	1940	drums

df[2, c(2,3)]



lst



lst[1]

c(1, 2)

lst[[1]]

x > 2 & x < 9

TRUE & TRUE

TRUE

x > 2 & < 9

TRUE & Error!

Error!

Operator	Syntax	Tests
`>`	`a > b`	Is a greater than b?
`>=`	`a >= b`	Is a greater than or equal to b?
`<`	`a < b`	Is a less than b?
`<=`	`a <= b`	Is a less than or equal to b?
`==`	`a == b`	Is a equal to b?
`!=`	`a != b`	Is a not equal to b?
`%in%`	`a %in% c(a, b, c)`	Is a in the group c(a, b, c)?

Operator	Syntax	Tests
`&`	`cond1 & cond2`	Are both `cond1` and `cond2` true?
`|`	`cond1 | cond2`	Is one or more of `cond1` and `cond2` true?
`xor`	`xor(cond1, cond2)`	Is exactly one of `cond1` and `cond2` true?
`!`	`!cond1`	Is `cond1` false? (e.g., `!` flips the results of a logical test)
`any`	`any(cond1, cond2, cond3, ...)`	Are any of the conditions true?
`all`	`all(cond1, cond2, cond3, ...)`	Are all of the conditions true?

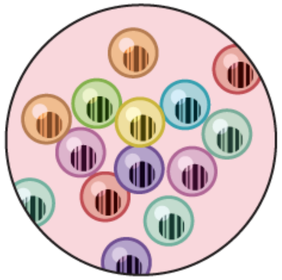
Basic Programming

Slot Machines (Conditionals)

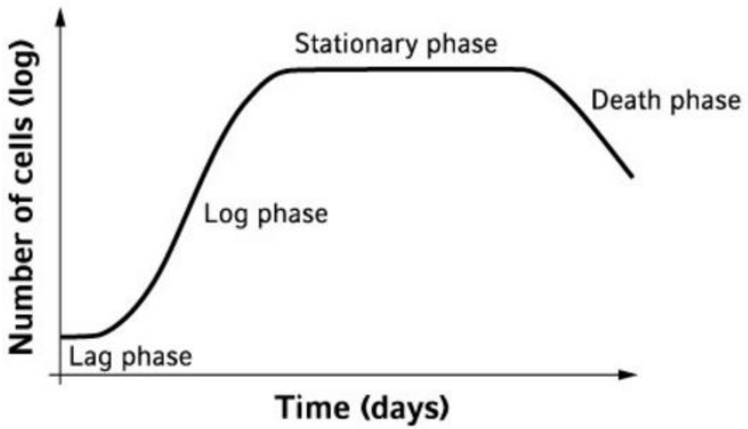
Expected Returns (Loops)

Speed

Pooled Cell Lines Example



Assay-ready pools



- ☐ Cell Lines are characterized by
  - ☐ Growth rates (doubling per day)
  - ☐ Initial cell count (cells)
  - ☐ Duration of lag-phase (days)
- ☐ Assay is characterized by
  - ☐ End-point (days)
  - ☐ PCR bottleneck (cells)
  - ☐ Sequencing depth (counts)
- ☐ # of cell lines with more than *400* counts.

Pooled Cell Lines and Conditionals

- Strategy:
  - Break complex tasks into simple subtasks.
  - Use concrete examples.
  - Describe your solutions in English, then convert them to R.
- If / else statements

```
if (condition){
  # This will run only when
  # the condition is TRUE
} else {
  # This will run only when
  # the condition is FALSE
}
```

```
# First attempt (planning)
simulate_experiment <- function(){

  # calculate cell counts at the end-point
  lysate_cell_counts <- simulate_growth()

  # simulate the pcr bottleneck
  pcr_counts <- simulate_pcr()

  # simulate sequencing
  sequencing_counts <- simulate_sequencing()

  # visualize final population
  hist(sequencing_counts, 100)

  # count the detected cell lines
  detected_lines <- count_detected()
}
```

Loops

- [ ] Expected values
- [ ] expand.grid
- [ ] for loops
- [ ] while loops
- [ ] repeat loops

$$E(x) = \sum_{i=1}^{n} (x_i \cdot P(x_i))$$

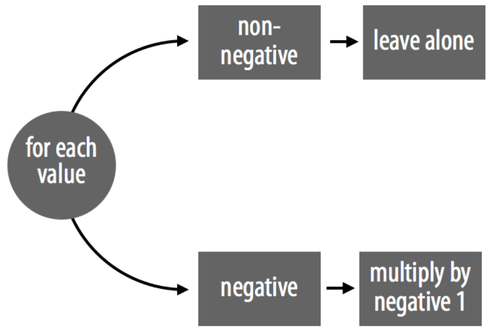
$$\begin{aligned} E(\text{die}) &= \sum_{i=1}^{n} (\text{die}_i \cdot P(\text{die}_i)) \\ &= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} \\ &= 3.5 \end{aligned}$$

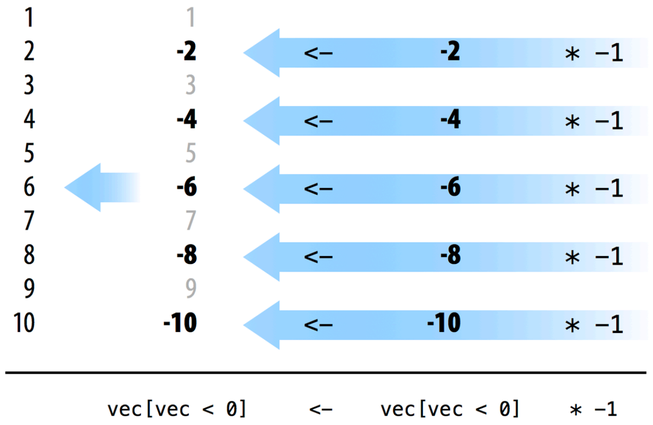
```
for (value in that) {
  this
}
```

```
while (condition) {
  code
}
```

Speed

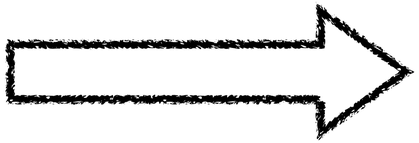
- [ ] Vectorized code
  - logical tests
  - subsetting
  - element-wise execution
- [ ] Faster for loops
- [ ] apply() and its friends
- [ ] Monte Carlo simulations

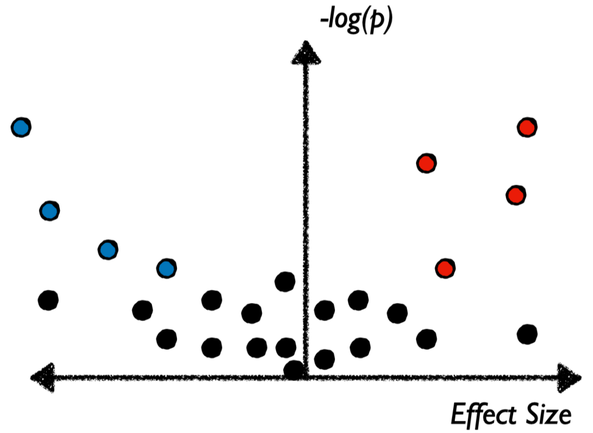




Practice: Correlations and Volcano Plots

volcano(X , y)





- ☐ Are rows matching?
- ☐ What if there are missing values in X or y?
- ☐ Multiple hypothesis correction — q values (FDR)
- ☐ Bonus: Replace y with a matrix Y

$$z = r\sqrt{\frac{n-2}{1-r^2}} \quad \sim \quad t_{n-2}$$

Homework

1. From 2019 HW's: 1 & 2
2. Review all the code we wrote and Mustafa sent.

   Check out chapters 7,9,11, and 12 (optional) from HOPR
3. Videos from Roger Peng

   Data Types: I / II / III

   Subsetting: I / II / III / IV / V

   Control Structures: I / II / III / IV / V


BROAD
INSTITUTE

depmap

PRISM

CANCER
DATA SCIENCE