HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 1 — Guidance and Introduction

Jorge N. Tendeiro

Hiroshima University

Today

- Course guidance.
- Data Science.
- Statistics.
- R, RStudio, Posit Cloud.

Course guidance

Hardware, software

Hardware
Please make sure your PC satisfies the following requirements:
https://www.hiroshima-u.ac.jp/en/about/initiatives/jyoho_ka/hikkei_pc

Software
We will use software to help us with running basic data analytic methods.

- Excel:
  Provided by HU.
  https://www.media.hiroshima-u.ac.jp/services/ms-ees/
  $\longrightarrow$ Make sure you can open and run Excel!

- R, RStudio:
  Open software, for free.
  https://www.r-project.org/
  $\longrightarrow$ Installing R and RStudio in computers with Asian language can be difficult. I therefore advise you to use a *free* web service called Posit Cloud (https://posit.cloud/).

Course guidance

Requirements
Little statistical knowledge is required as a prerequisite.
This course is designed for beginners.

- Some parts of the lectures require *high school* mathematics (i.e., data analysis, probability/statistics).
- However, **no** complex formulas (e.g. calculus) will be used in the course.
- We do use the four arithmetic operations ($+, -, \times, \div$) and the square root ($\sqrt{\cdot}$).

Goal
We focus on the use of statistics using software.
We do **not** focus on difficult mathematical computations or derivations.

Lecture format

The lectures will be given on-demand via *Moodle*.

I will upload lecture slides and videos under the "Fundamental Data Science" course in *Moodle*.

- Each lecture video consists of one or more video clips.
- Since lectures are offered on-demand, you can watch them whenever you want during the scheduled period.
- Other lecture materials (e.g. data to be analyzed) will be announced in the lecture videos, so please **pay attention**!!

Important:

> ***DO NOT USE*** *lecture slides, videos, and handouts for purposes other than this course.*

Dates for releasing the course materials

Lecture	Date	Topic
1	Nov 29 (08:30 AM)	Guidance and Introduction
2	Dec 04 (08:30 AM)	Data acquisition and open data, data science ethics
3	Dec 06 (08:30 AM)	Types of data and descriptive statistics
4	Dec 11 (08:30 AM)	Descriptive statistics
5	Dec 13 (08:30 AM)	Visualize data in R
6	Dec 18 (08:30 AM)	Correlation and regression
7	**Dec 20 (08:30 AM)**	Simple regression analysis in Excel
8	**Dec 20 (08:30 AM)**	Principal Component Analysis, Cluster Analysis in R
9	Jan 08 (08:30 AM)	Probability
10	Jan 10 (08:30 AM)	Random variables and probability distributions
11	**Jan 15 (08:30 AM)**	Basic probability distributions
12	**Jan 15 (08:30 AM)**	Bivariate probability distributions
13	Jan 22 (08:30 AM)	Methods for data collection
14	Jan 24 (08:30 AM)	Point estimate and interval estimation
15	Jan 29 (08:30 AM)	Interval estimation

Course grade

Your course grade will be completely determined by a series of check tests.

- Given via *Moodle*.
- One check test per lecture (thus, 15 in total), available after each lecture.
- You should take each check test **after** going through the corresponding lecture slides and video!
- You can only submit your answer **ONCE**.
  (Once uploaded to *Moodle*, there is **no way** to delete and resubmit!).
- Pay attention to the **deadline** for each check test (see next page).
  You cannot take a test after the deadline, in principle.
- Taking a check test $=$ "Attended" the lecture.

Grade evaluation requires **at least 10** attendances.
Attend 9 or less times $\Rightarrow$ Your grade will **NOT** be evaluated!

Deadlines for the check tests

Lecture	From...	To...
1	Nov 29 (08:30)	Dec 05 (18:00)
2	Dec 04 (08:30)	Dec 10 (18:00)
3	Dec 06 (08:30)	Dec 12 (18:00)
4	Dec 11 (08:30)	Dec 17 (18:00)
5	Dec 13 (08:30)	Dec 19 (18:00)
6	Dec 18 (08:30)	Jan 07 (18:00)
7, 8	**Dec 20 (08:30)**	**Jan 09 (18:00)**
9	Jan 08 (08:30)	Jan 14 (18:00)
10	Jan 10 (08:30)	Jan 16 (18:00)
11, 12	**Jan 15 (08:30)**	**Jan 23 (18:00)**
13	Jan 22 (08:30)	Jan 28 (18:00)
14	Jan 24 (08:30)	Jan 30 (18:00)
15	Jan 29 (08:30)	Feb 04 (18:00)

There is about one week available for each check test, from the day the corresponding lecture is released.

Meeting the check tests' deadlines

Extending a deadline
We may allow to extend a deadline in cases of **force majeure only**.

- Deadlines may be postponed for 3 days (including weekends/holidays).
- In this case, please send me (Jorge Tendeiro) an email **BEFORE** the deadline.
- Also, please hand in a proof of your special reason
  (in case of sickness, receipt of visiting hospital or medical certificate).

Technical problems
We do **not** accept any PC problem as a special reason to extend a deadline.

Furthermore, inability to play the lecture videos is **not** a valid reason to ask for a deadline extension.
In case you have problems to watch the lecture videos in *Moodle*:
Check here: http://support.vle.hiroshima-u.ac.jp/bb9:video-emberr.

Meeting the check tests' deadlines

Extension of an extension
In principle, we do **not** accept ANY extra extension beyond 3 extra days after the original deadline.

Submitting after the deadline

- Once one student's deadline is extended, *all* students will be able to see their check tests even after the deadline.
- However, students with no granted extension **cannot** submit after the deadline.
  In such cases, submission after the deadline = **fail** (no grade, no attendance).

Important:

> *Please submit your check test **as soon as possible**. Manage your time **wisely**!!*

Other remarks

Questions related to course content

- Please post **all** your questions or comments about the lectures or their contents on *Moodle*'s forum.
  Do **not**, under any circumstance, email me or the teaching assistant to ask content-related questions:

  > *Such emails will not be served.*

- Questions posted in *Moodle*'s forum will be handled *as soon as possible* by the teaching assistant, Ms. Alolabi, or myself (Jorge Tendeiro).

Questions related to personal matters

If you have something personal to discuss with me, you can send an email to discuss about it.

Furthermore,

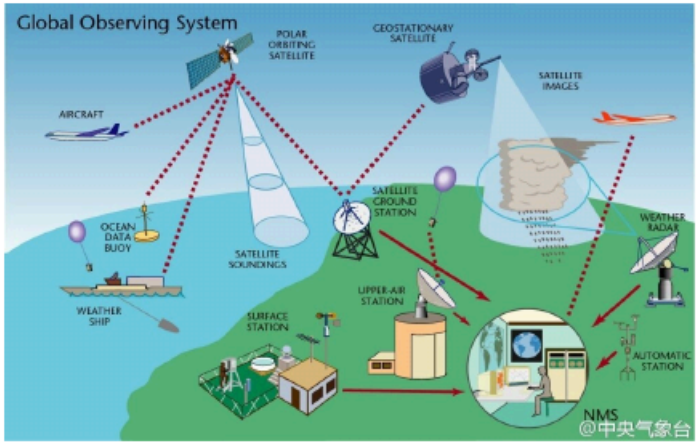
- Besides the information given in this lecture, please pay close attention to all announcements in *Moodle*.
- Lectures will **not** be rescheduled in case the university closes (e.g., due to bad weather).

Data Science

What is Data Science

> *Data Science allows retrieving meaningful information from data.*

Data Science is used for example to do weather forecast.



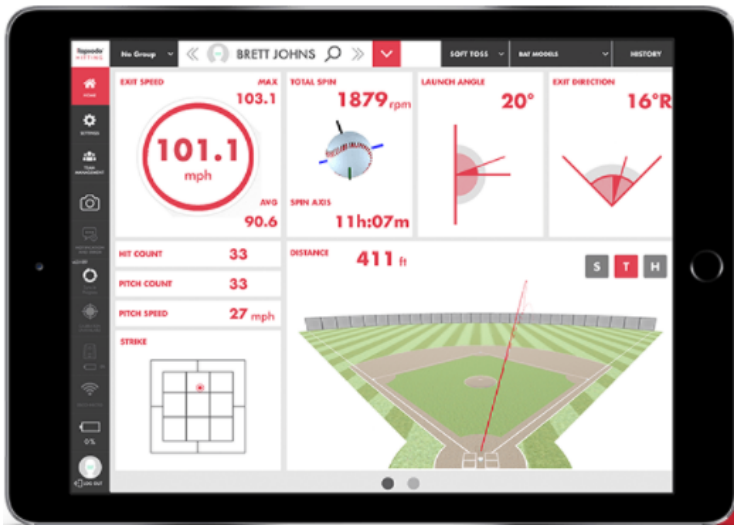
Prediction!

Source: photo, photo.

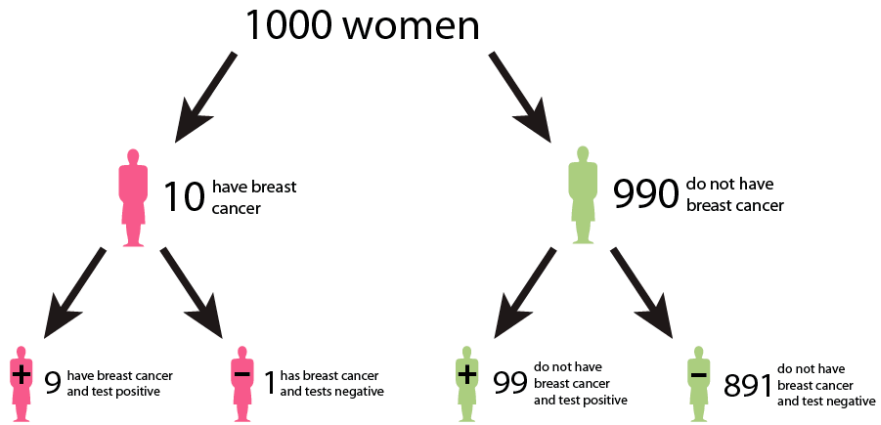
What is Data Science

Other examples:

Baseball analytics



Disease diagnostic



Source: photo, photo.

Can we *avoid* Data Science

If you do not use data science...

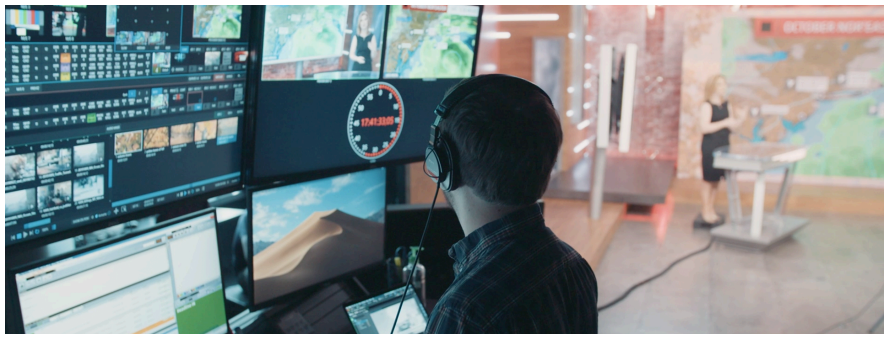
... you will fully rely on intuition/experience.

Decisions are **poorly** informed.

If you do use data science...

... besides intuition/experience, you also learn from data.

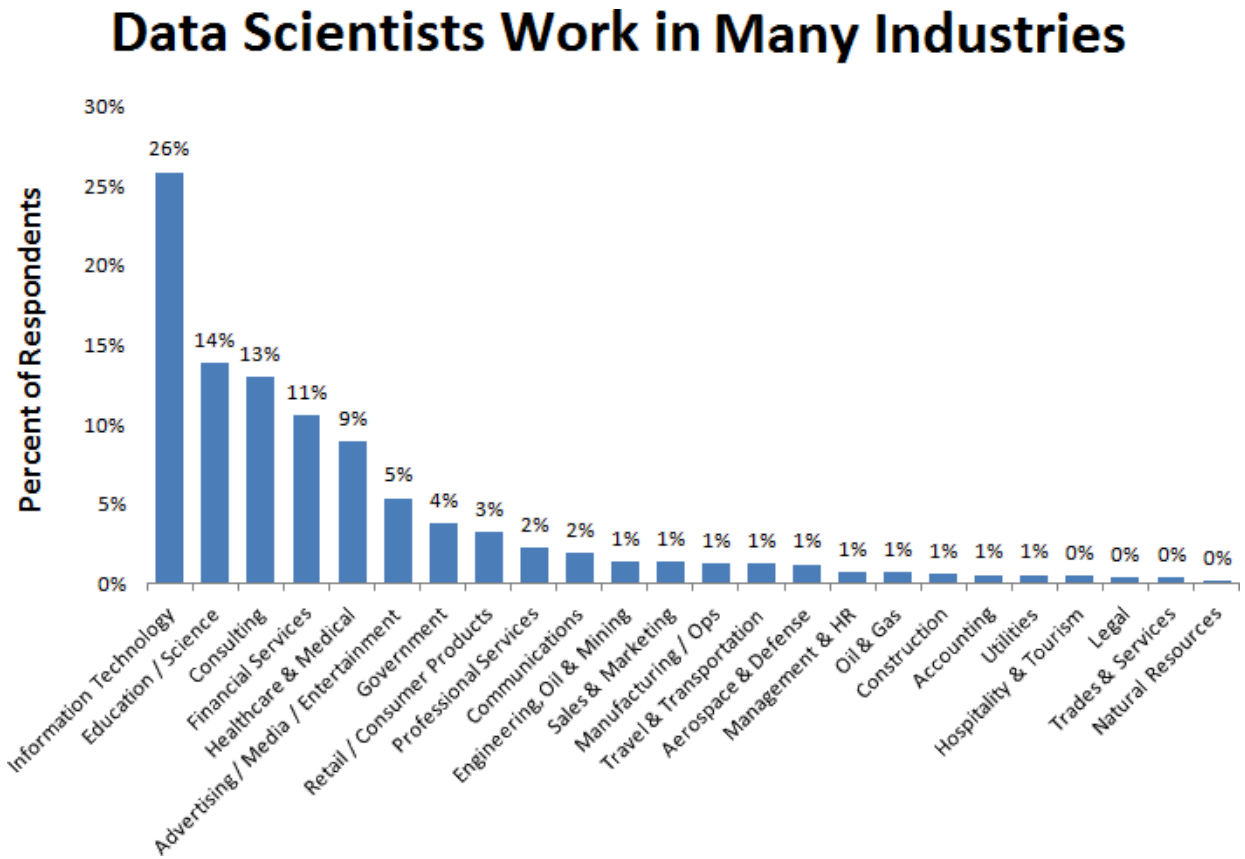
Decisions are **better** informed.



Source: photo, photo.

Data Scientists

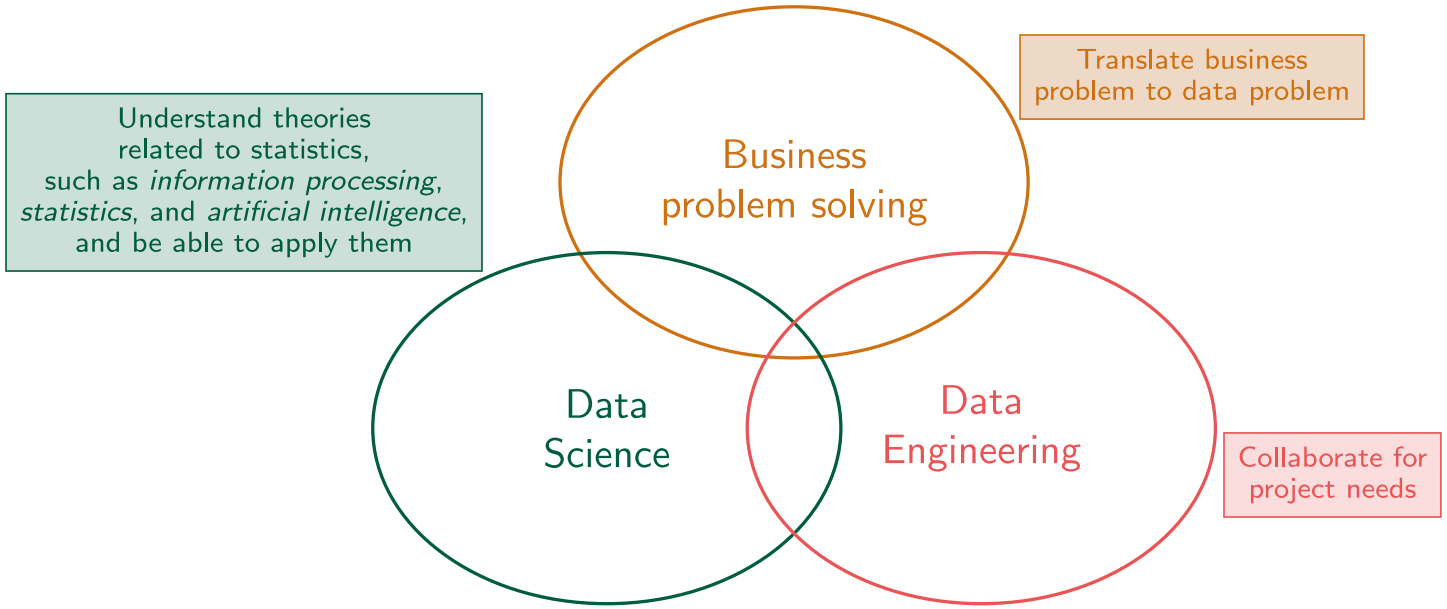
> *Data Scientists:* *Researchers or practitioners working on data science methods.*



Source: photo.

Data Scientists

Data scientists are required to master various skills.
In this course we will focus on a few, statistics-oriented, skills.
Below are the three skill sets required of data scientists, according to the Japan Data Scientist Society.



Translated from: https://prtimes.jp/main/html/rd/p/000000005.000007312.html.

The data problem-solving cycle (PPDAC)

Statistics

Statistics

Definition

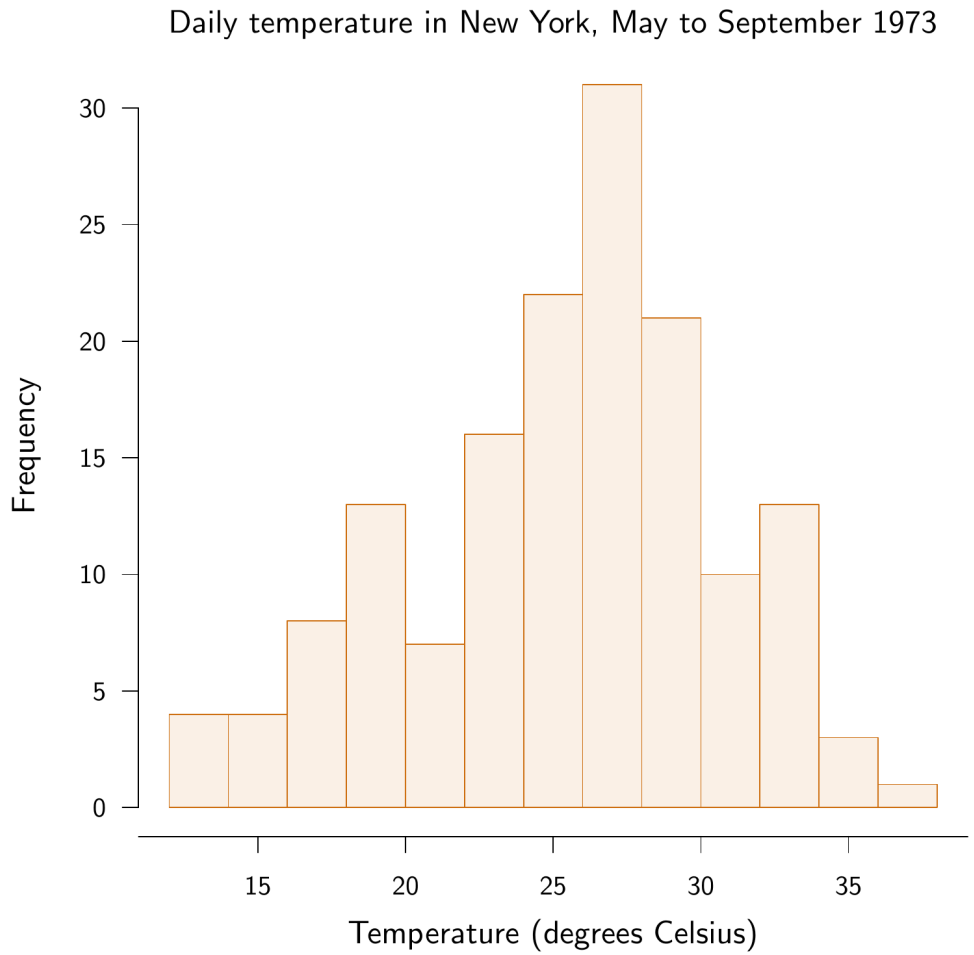
> *Statistics: branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.*
>
> *(Merrian-Webster)*

Two types of statistics

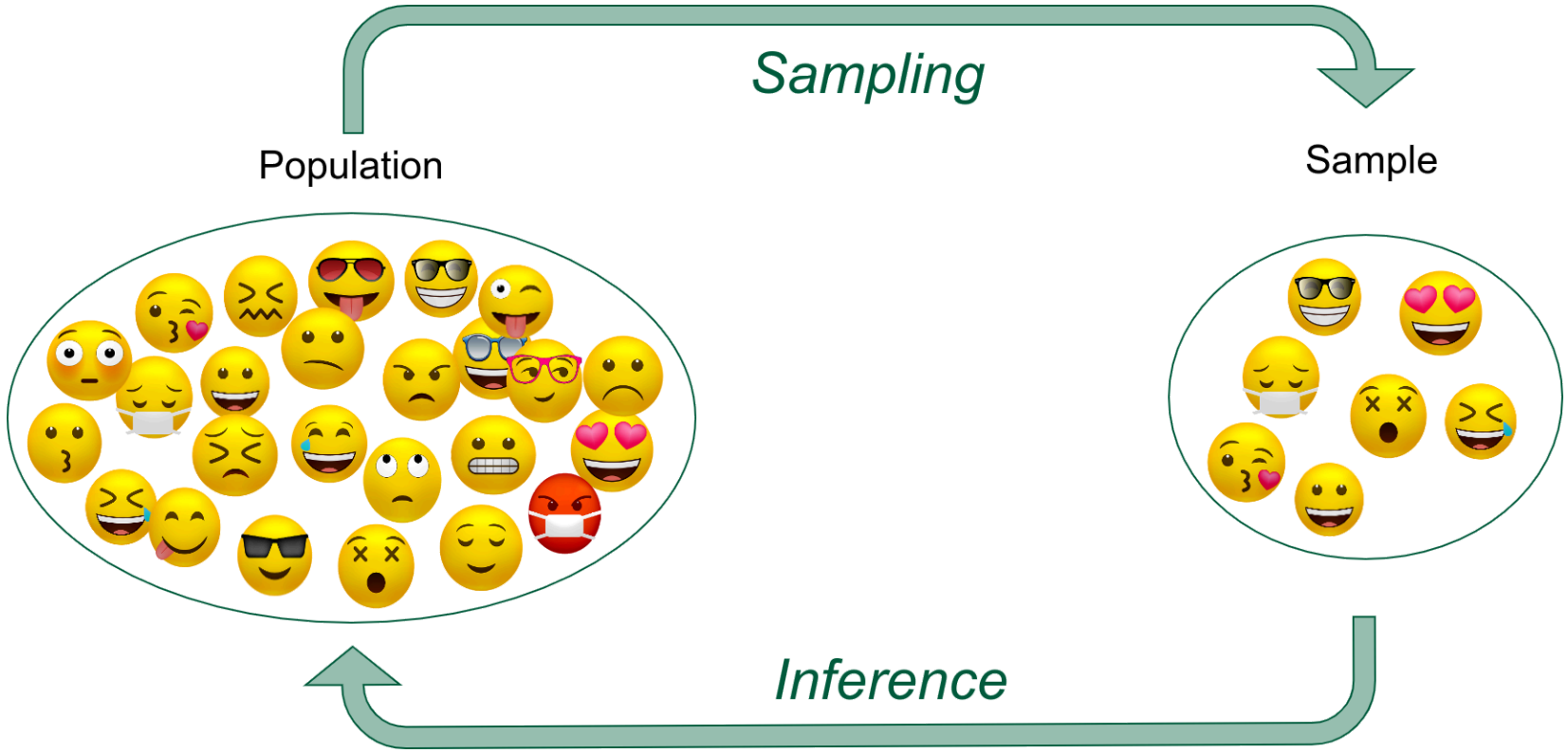
- **Descriptive** statistics:
  Describe the observed characteristics of collected data in a simple form (e.g., tables, graphs).
- **Inferential** Statistics:
  Generalize from the sample to the larger population from which the sample came from.

Examples of descriptive statistics

Count	Mean	Median	Min	Max	Range	Std Dev
153	25.49	26.11	13.33	36.11	22.78	5.26



Sample and population, sampling and inference



Emoticons from Pixabay, license, adapted.

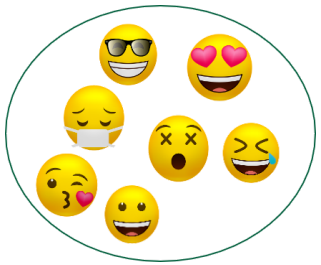
Example of inferential statistics

Consider the following research question of a researcher studying the topic of *attitude toward guns*:

> *What is the average attitude toward guns in Japan?*

Population = All people in Japan.
Unfortunately for the researcher, they can not collect data from each person in Japan...

Instead, the researcher may collect data from a sample.



- The sample is a (much) smaller collection of people.
- The sample should be representative of the entire population.

Inferential statistics uses the laws of probability to generalize the findings from the sample toward the entire population!

Lecture schedule revisited

Lecture	Topic
1	Guidance and Introduction
2	Data acquisition and open data, data science ethics
3	Types of data and descriptive statistics
4	Descriptive statistics
5	Visualize data in R
6	Correlation and regression
7	Simple regression analysis in Excel
8	Principal Component Analysis, Cluster Analysis in R
9	Probability
10	Random variables and probability distributions
11	Basic probability distributions
12	Bivariate probability distributions
13	Methods for data collection
14	Point estimate and interval estimation
15	Interval estimation

Descriptive statistics

Data analysis methods

Probability

Inferential statistics

R, RStudio, Posit Cloud

What is R?

Besides being the 18th letter of the English alphabet, I mean. 😄

From R's homepage:

> *R is a free software environment for statistical computing and graphics.*

It is essentially a *very* fancy calculator...

Input

```
2 + 3
```

```
2 * 3
```

```
5 * (4 + 2/3)
```

```
sqrt((2/3) * 4 - 1)
```

```
2^3
```

Output

```
[1] 5
```

```
[1] 6
```

```
[1] 23.33333
```

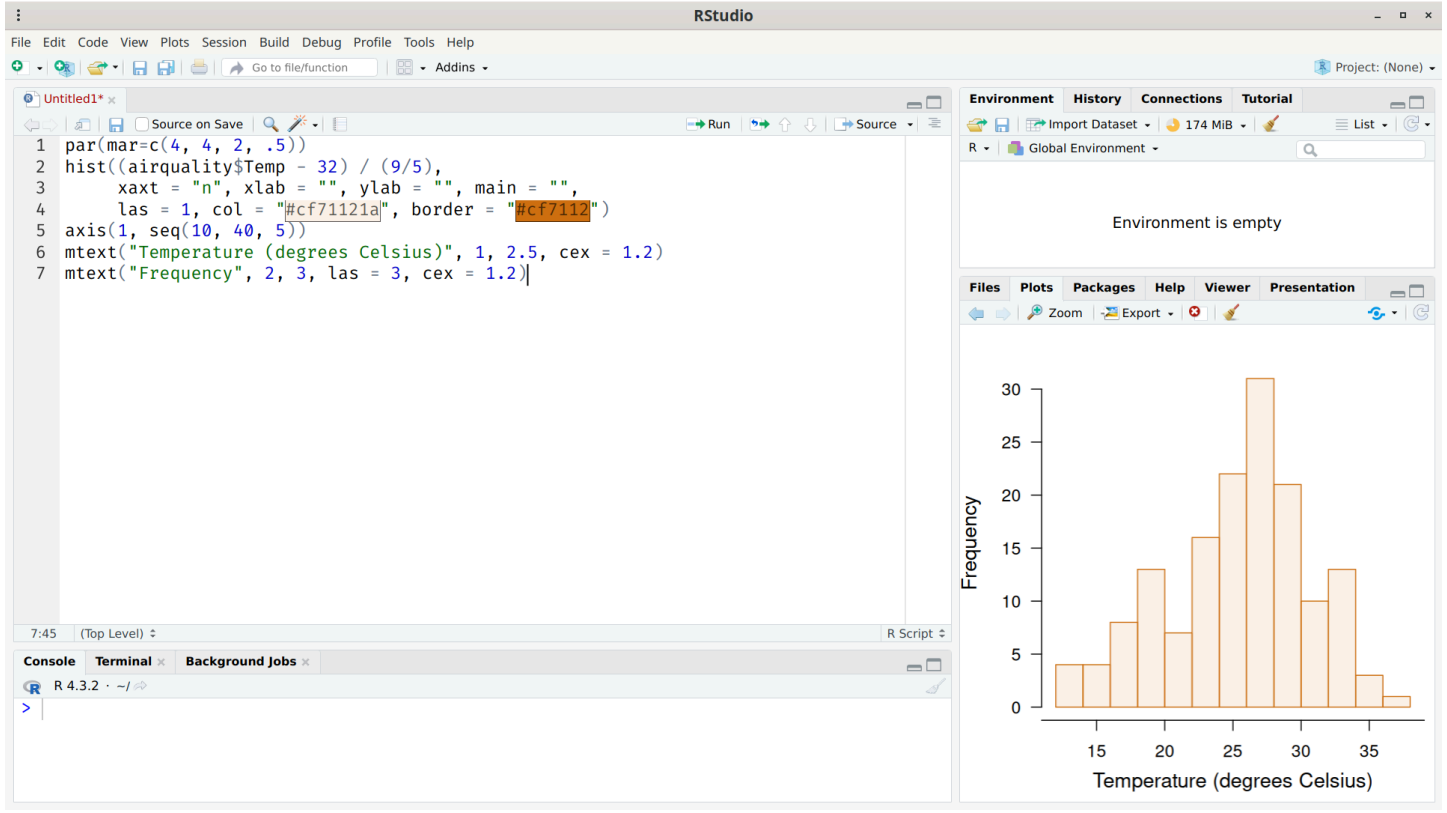
```
[1] 1.290994
```

```
[1] 8
```

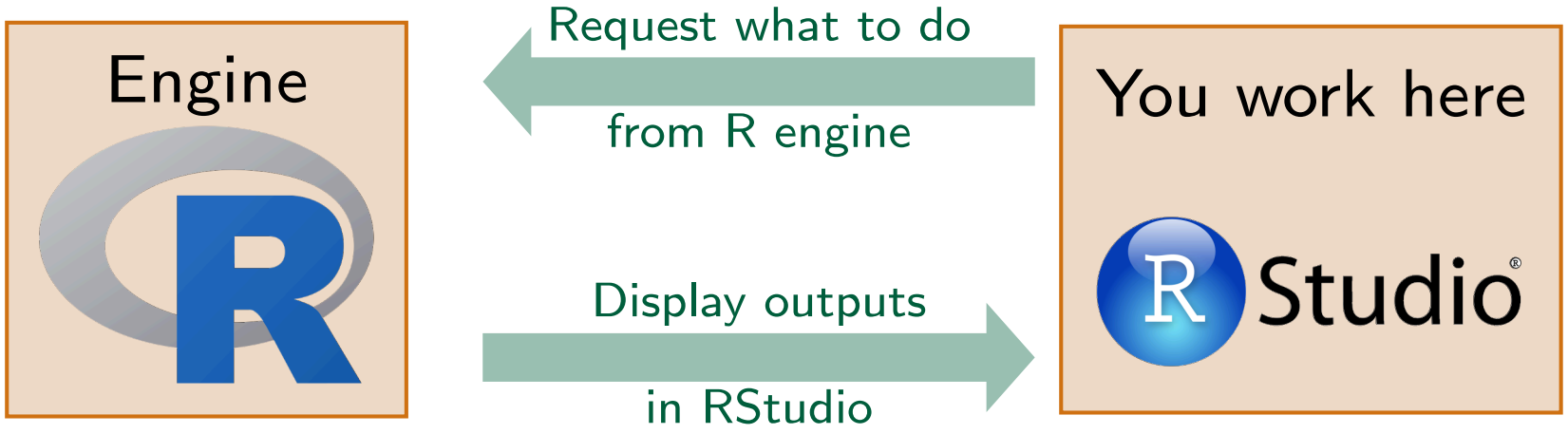
... but it can do so so much more!!

What is RStudio?

RStudio is an IDE (Integrated Development Environment) for R.
It is the program that we use to type and execute R code.
It is entirely free and *extremely* powerful.

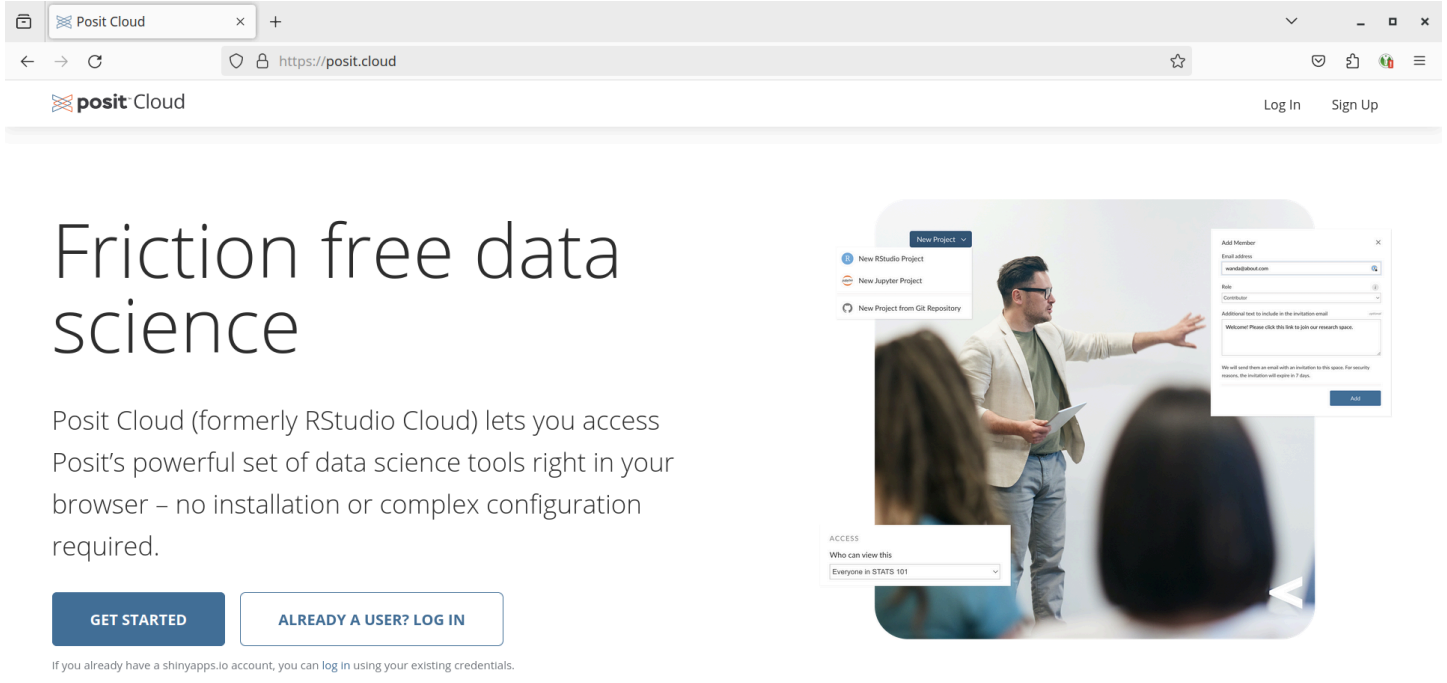


R and RStudio



What is Posit Cloud?

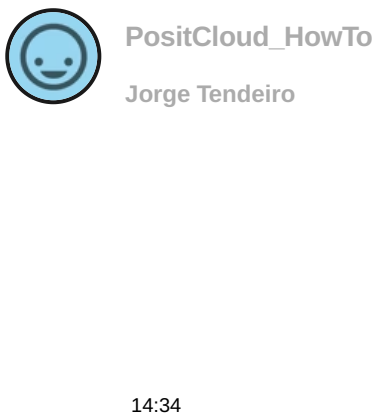
With Posit Cloud we can use R and RStudio from the browser.
There is **no** need to install R or RStudio.
The advised solution for this course is to use the free version of Posit Cloud.



Posit Cloud

The advised solution for this course is to use the free version of Posit Cloud.

I created a video that shows how you can create a free account and use Posit Cloud: https://vimeo.com/manage/videos/913207949.



Posit Cloud

Important

- Posit Cloud is a cloud-based service.
  An internet connection is required to use it!
- As much as possible, type in English.
  (Typing in Japanese can lead to some issues.)
- Pay attention to the number of free hours per month
  (see the video above at the 1:24 and 13:10 marks).
  We will use 3 hours in lectures 5 and 8 in the course.
- Do notice that the time counter in Posit Cloud is running as long as the webpage is open.
  Therefore **log out** from Posit Cloud when you are not working!
- Make sure to create a Posit Cloud free account **before** Lecture 5!

In case you have any questions: Please use *Moodle*'s forum.

OPTIONAL: Installing R and RStudio Desktop locally

Students who want to install R and RStudio locally on theirs PCs are free to do so, of course.

You can follow the videos below for guidance (Windows and Mac).
(Linux users: I'm positive you know what to do 😄. Just follow the instructions for your distro.)

Tip: Your username should be in *halfsize alphabet*.

Windows:

RandRStudio_Win

Jorge Tendeiro

08:10

Mac:

RandRStudio_MacOS

Jorge Tendeiro

06:30