Data Handling: Import, Cleaning and Visualisation

Lecture 3:

A Brief Introduction to Data and Data Processing

Prof. Dr. Ulrich Matter

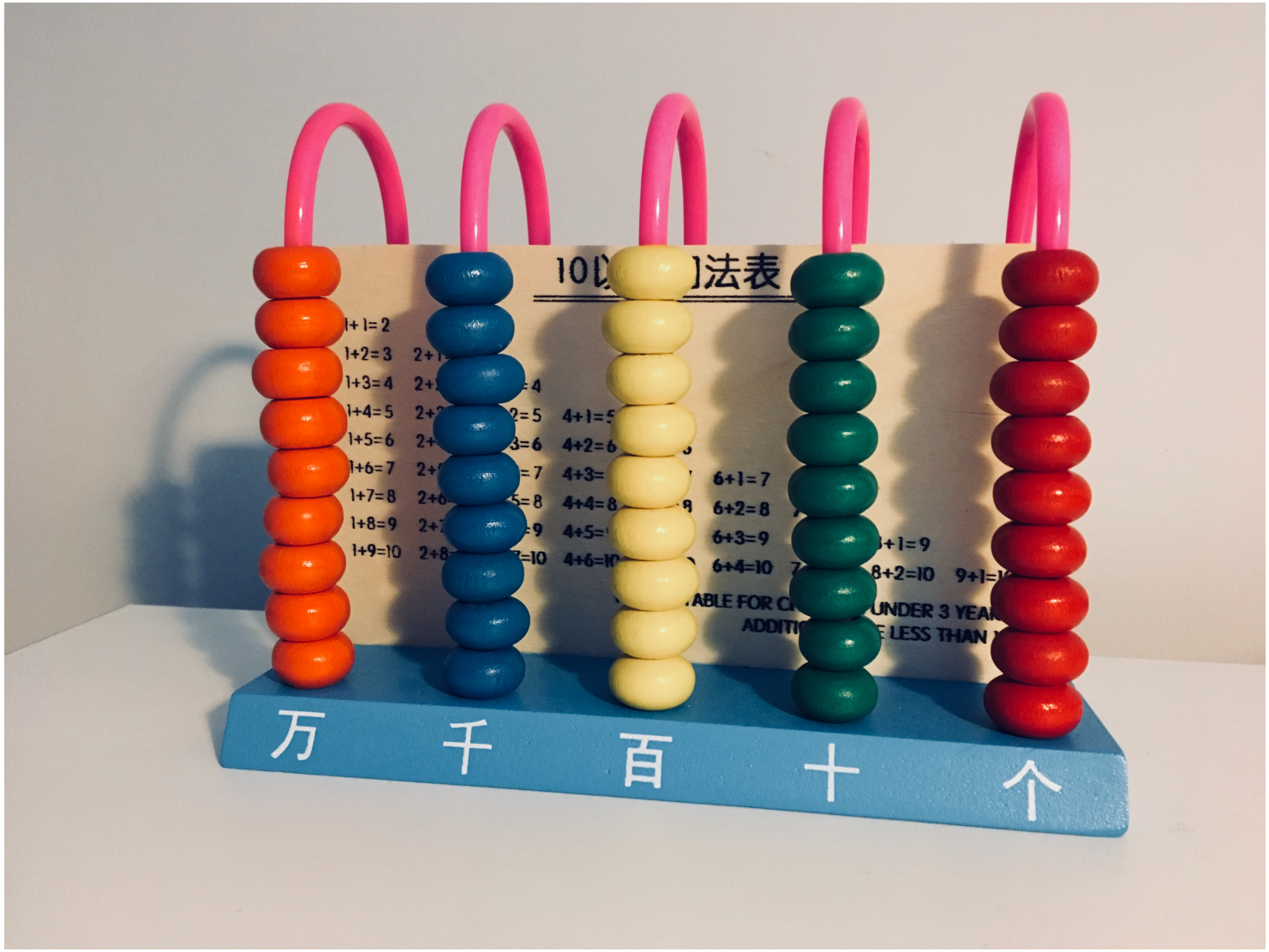
Recap

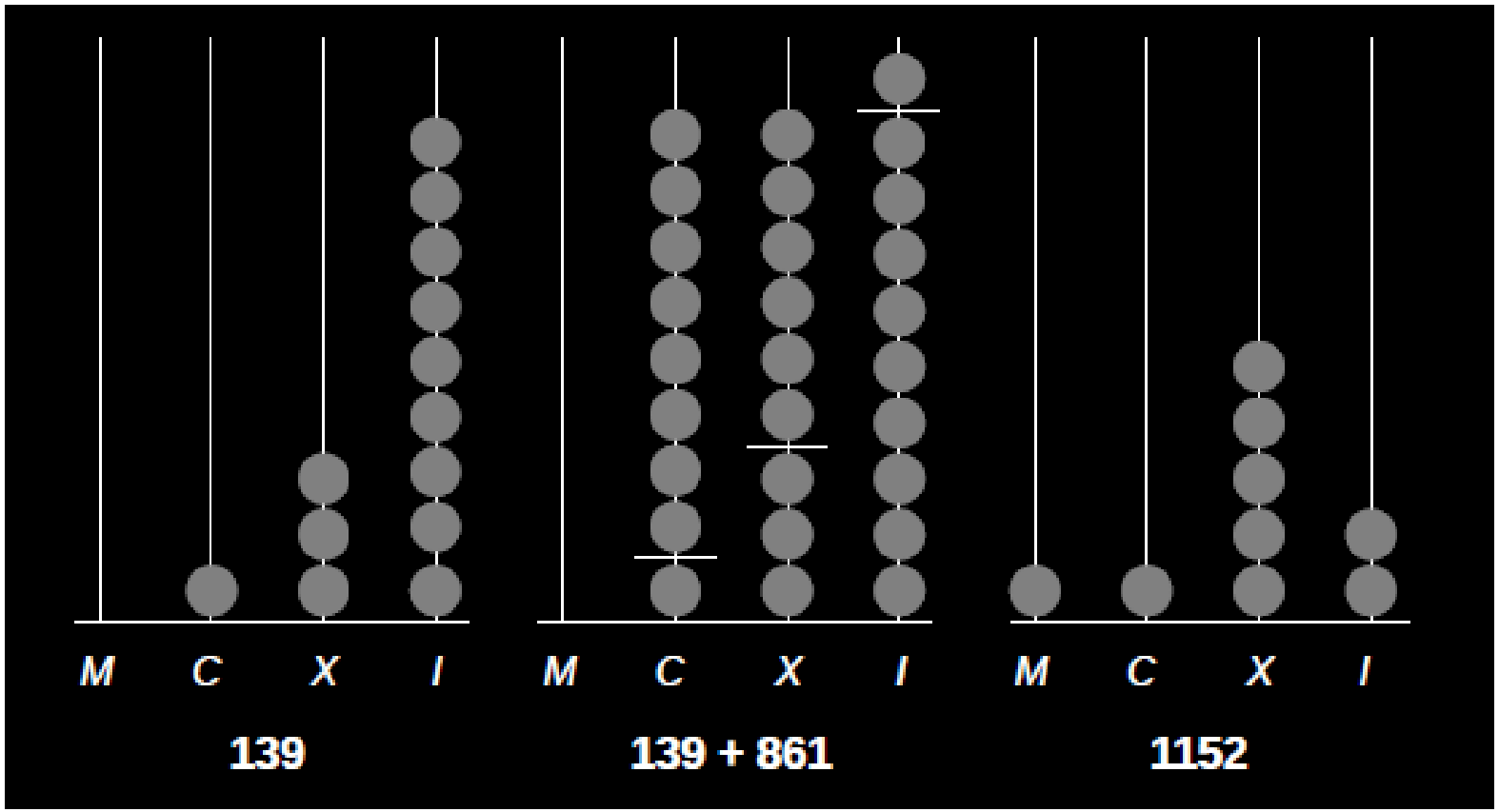
Basic programming concepts

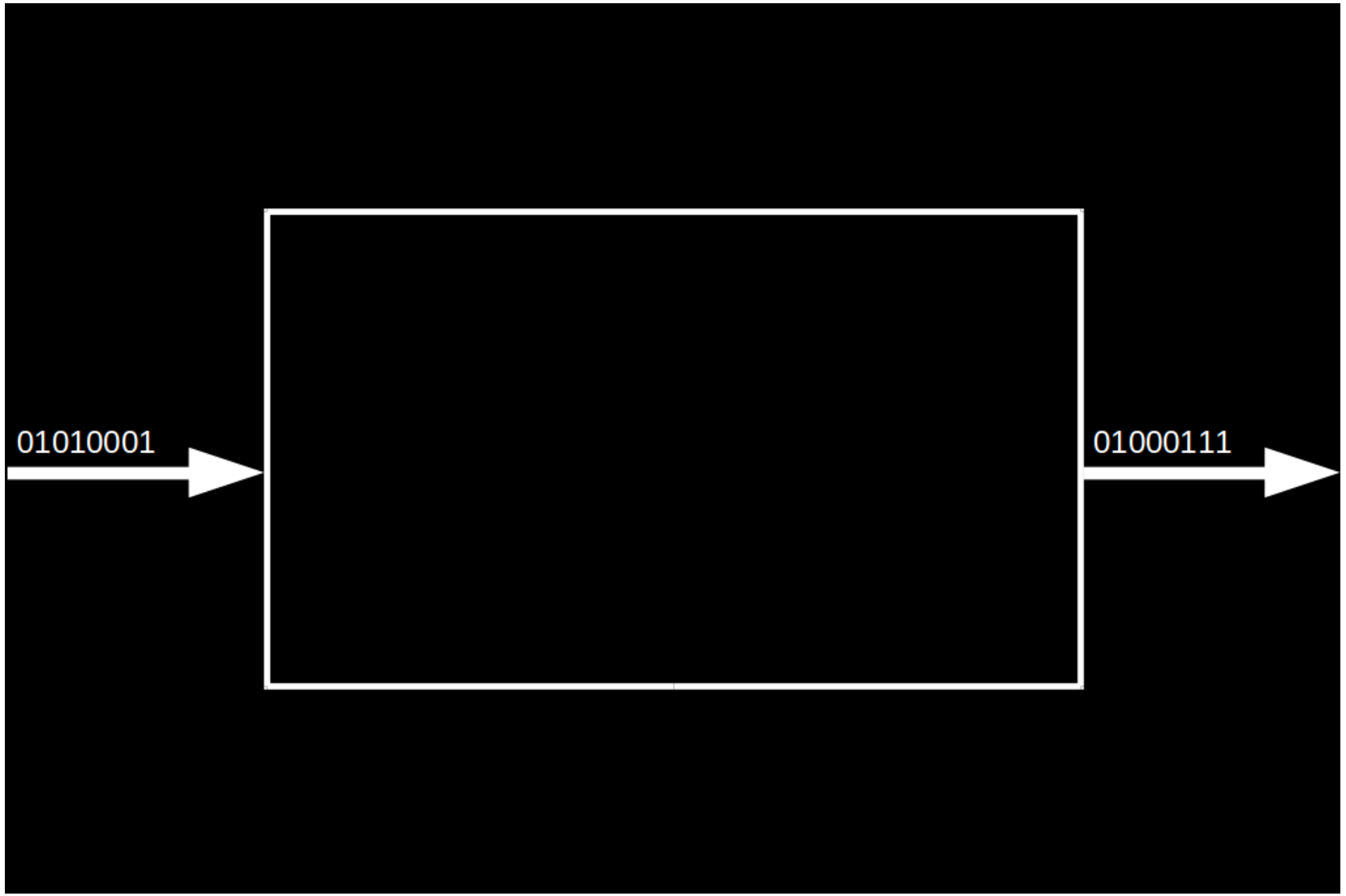
- Values, variables
- Vectors
- Loops
- Logical statements
- Control statements
- Functions

https://scratch.mit.edu/

Data Processing


万
千
百
十
个

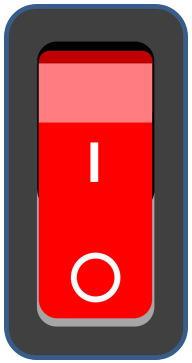

M C X I
M C X I
M C X I
139
139 + 861
1152


01010001
01000111

The binary system

Microprocessors can only represent two signs (states):

- ‘Off’ = 0
- ‘On’ = 1



The binary counting frame

- Only two signs: `0`, `1`.
- Base 2.
- Columns: $2^0 = 1$, $2^1 = 2$, $2^2 = 4$, and so forth.

The binary counting frame

- Sufficient to represent all **natural** numbers in the decimal system.

The binary counting frame

- Sufficient to represent all **natural** numbers in the decimal system.
- Representing fractions is tricky
  - e.g. $1/3 = 0.333..$ actually constitutes an infinite sequence of 0s and 1s.
  - Solution: ‘floating point numbers’ (not 100% accurate)

The binary counting frame

What is the decimal number 139 in the binary counting frame?

The binary counting frame

What is the decimal number 139 in the binary counting frame?

- Solution:

$$(1 \times 2^7) + (1 \times 2^3) + (1 \times 2^1) + (1 \times 2^0) = 139.$$

The binary counting frame

What is the decimal number 139 in the binary counting frame?

- Solution:

$$(1 \times 2^7) + (1 \times 2^3) + (1 \times 2^1) + (1 \times 2^0) = 139.$$

- More precisely:

$$(1 \times 2^7) + (0 \times 2^6) + (0 \times 2^5) + (0 \times 2^4) + (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0) = 139.$$

- That is, the number `139` in the decimal system corresponds to `10001011` in the binary system.

Decimal numbers in a computer

If computers only understand `0` and `1`, how can they express decimal numbers like 139?

Decimal numbers in a computer

If computers only understand `0` and `1`, how can they express decimal numbers like **139**?

- **Standards** define how symbols, colors, etc are shown on the screen.
- Facilitates interaction with a computer (our keyboards do not only consist of a `0`/`1` switch).

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1
0 =	0	0	0	0	0	0	0	0

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1
0 =	0	0	0	0	0	0	0	0
1 =	0	0	0	0	0	0	0	1

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1
0 =	0	0	0	0	0	0	0	0
1 =	0	0	0	0	0	0	0	1
2 =	0	0	0	0	0	0	1	0

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1
0 =	0	0	0	0	0	0	0	0
1 =	0	0	0	0	0	0	0	1
2 =	0	0	0	0	0	0	1	0
3 =	0	0	0	0	0	0	1	1

Conversion between binary and decimal

Number	128	64	32	16	8	4	2	1
0 =	0	0	0	0	0	0	0	0
1 =	0	0	0	0	0	0	0	1
2 =	0	0	0	0	0	0	1	0
3 =	0	0	0	0	0	0	1	1
...								
139 =	1	0	0	0	1	0	1	1

Character Encoding

Computers and text

How can a computer understand text if it only understands `0`s and `1`s?

A modified version of South Korean Dubeolsik (two-set type) for old hangul letters. (Illustration by Yes0song 2010, Creative Commons Attribution-Share Alike 3.0 Unported)

Computers and text

How can a computer understand text if it only understands `0`s and `1`s?

- **Standards** define how `0`s and `1`s correspond to specific letters/characters of different human languages.
- These standards are usually called **character encodings**.
- Coded character sets that map unique numbers (in the end in binary coded values) to each character in the set.

Computers and text

How can a computer understand text if it only understands `0`s and `1`s?

- **Standards** define how `0`s and `1`s correspond to specific letters/characters of different human languages.
- These standards are usually called **character encodings**.
- Coded character sets that map unique numbers (in the end in binary coded values) to each character in the set.
- For example, ASCII (American Standard Code for Information Interchange).

ASCII logo. (public domain).

ASCII Table

Binary	Hexadecimal	Decimal	Character
0011 1111	3F	63	?
0100 0001	41	65	A
0110 0010	62	98	b

Character encodings: why should we care?

Character encodings: why should we care?

- In practice, Data Science means handling digital data of all formats and shapes.
  - Diverse sources.
  - Different standards.
  - read/store data.
- At the lowest level, this means understanding/handling encodings.

Computer Code and Text-Files

Putting the pieces together…

Two core themes of this course:

1. How can **data** be **stored** digitally and be **read** by/imported to a computer?
2. How can we give instructions to a computer by writing **computer code**?

Putting the pieces together…

Two core themes of this course:

1. How can **data** be **stored** digitally and be **read** by/imported to a computer?
2. How can we give instructions to a computer by writing **computer code**?

In both of these domains we mainly work with one simple type of document: **text files**.

Text-files

- A **collection of characters** stored in a designated part of the computer memory/hard drive.
- An easy-to-read representation of the underlying information (`0`s and `1`s)!

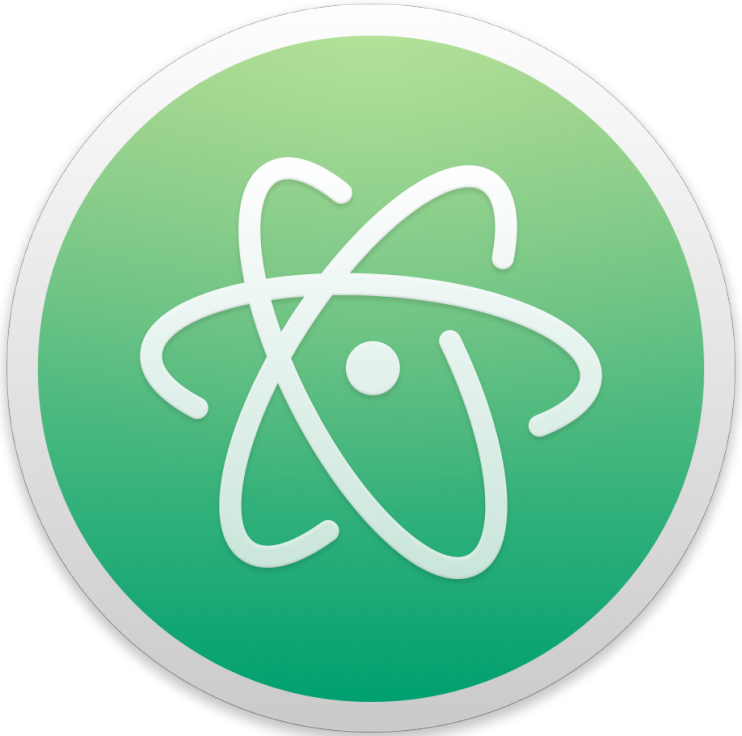
Text-files

- A **collection of characters** stored in a designated part of the computer memory/hard drive.
- An easy to read representation of the underlying information (`0`s and `1`s)!
- Common device to store data:
  - Structured data (tables)
  - Semi-structured data (websites)
  - Unstructured data (plain text)
- Typical device to store computer code.

Text-editors: RStudio

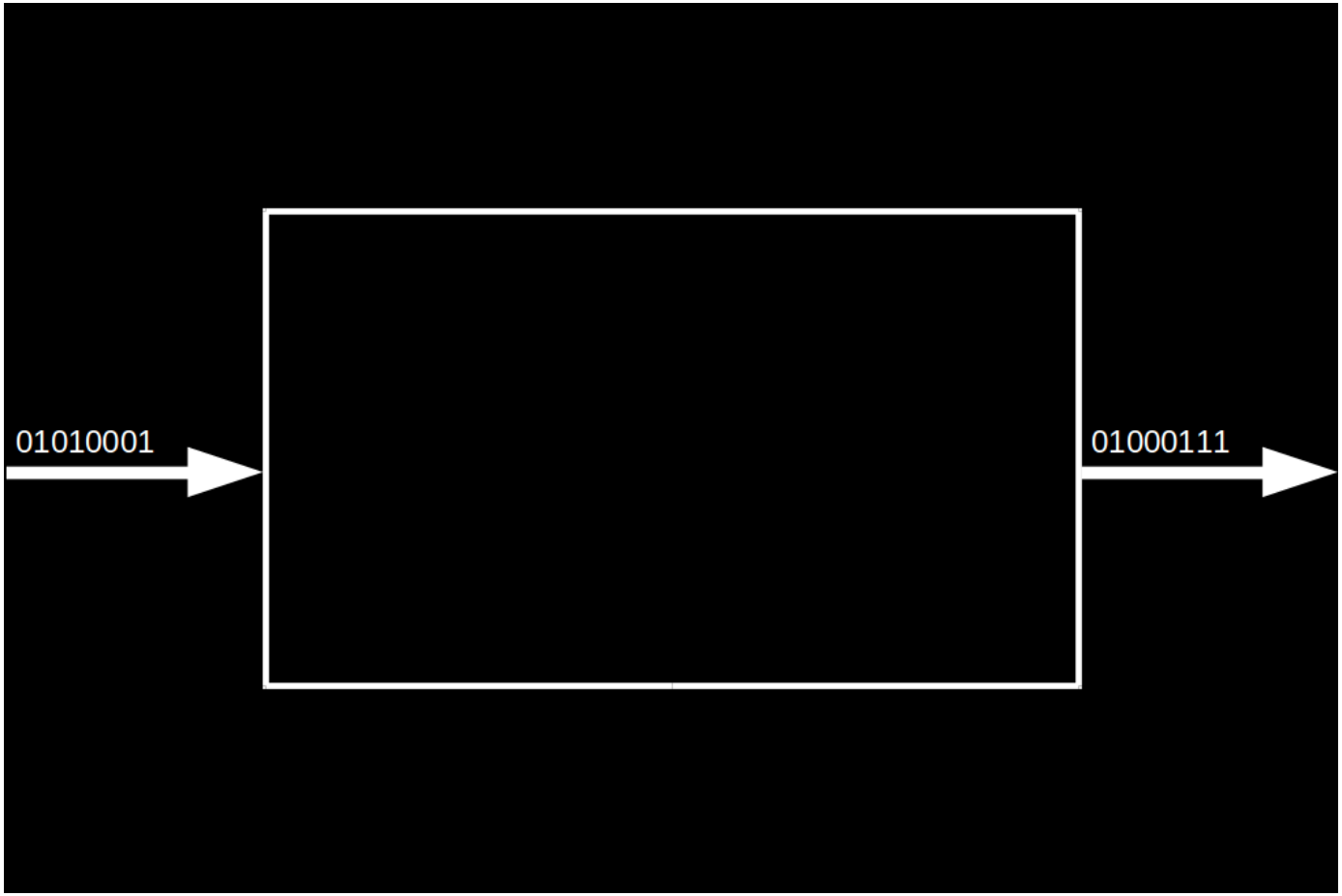
Install RStudio from here!

Text-editors: Atom



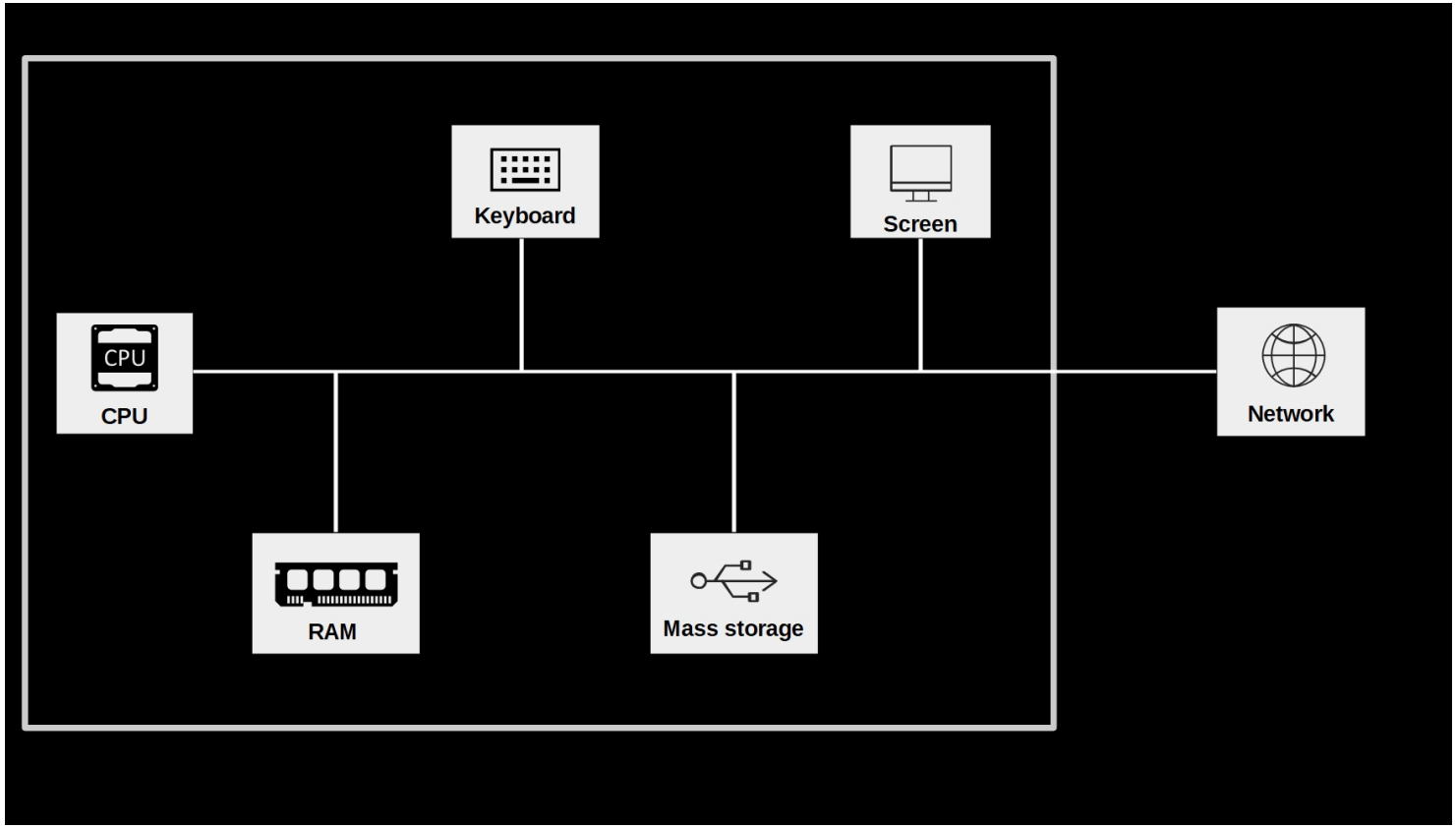
Install Atom from here!

Data Processing Basics



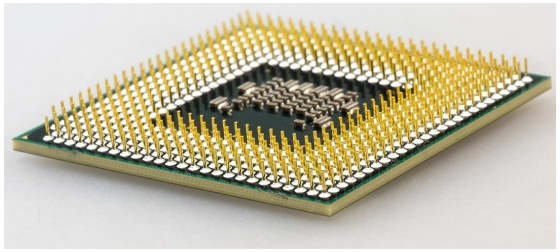
The 'blackbox' of data processing.

Components of a standard computing environment



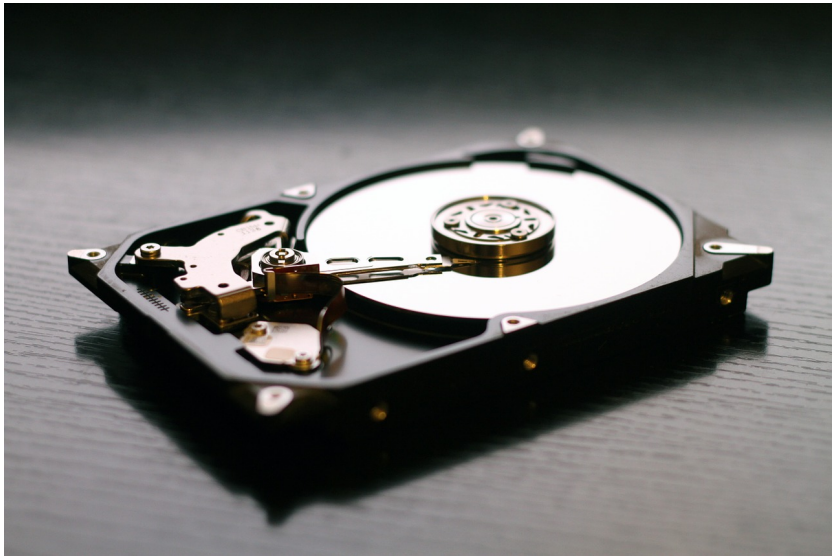
Basic components of a standard computing environment.

Central Processing Unit



Random Access Memory

Mass storage: hard drive



Network: Internet, cloud, etc.

Putting the pieces together…

Recall the initial example (survey) of this course.

1. Access a website (over the Internet), use keyboard to enter data into a website (a Google sheet in that case).

2. R program accesses the data of the Google sheet (again over the Internet), downloads the data, and loads it into RAM.

3. Data processing: produce output (in the form of statistics/plots), output on screen.

Q&A

References