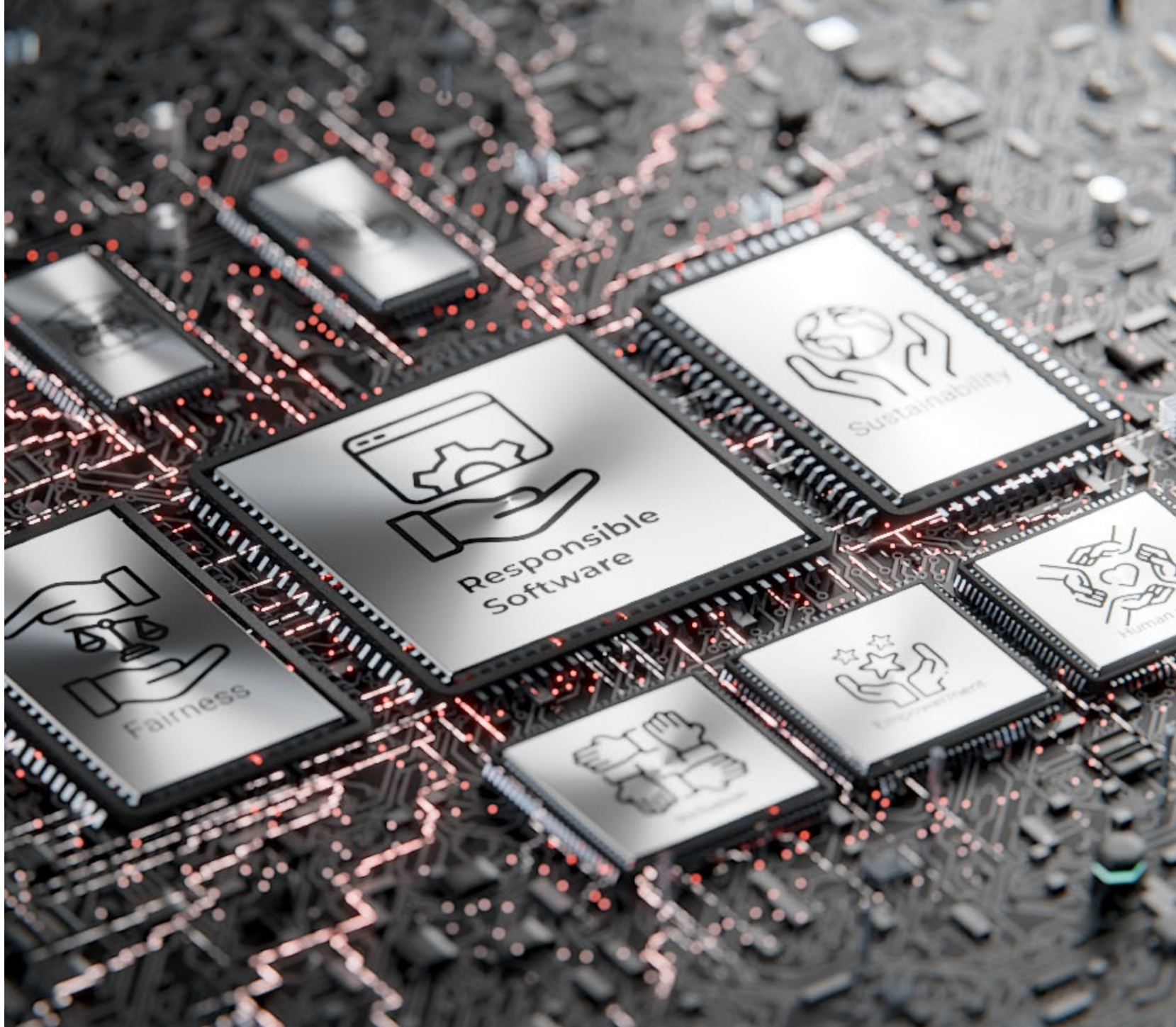
# Sustainability 2
# Review & Case studies
## 18 nov.

Cécile Hardebolle

## Responsible Software



# Agenda for today

1. Information for the Graded Assignment 2

2. Interactive review questions on Sustainability 2

3. Case studies:
   a) Causal loop diagram
   b) Ethical decision making

# Graded Assignment 2

# Graded assignment 2: logistics

You will be **assigned the same seat** as for Graded 1
The seating plan will be communicated this week on moodle

- If you see an issue with the seat you are assigned, please contact me!
- Make sure to display your camipro card on your table

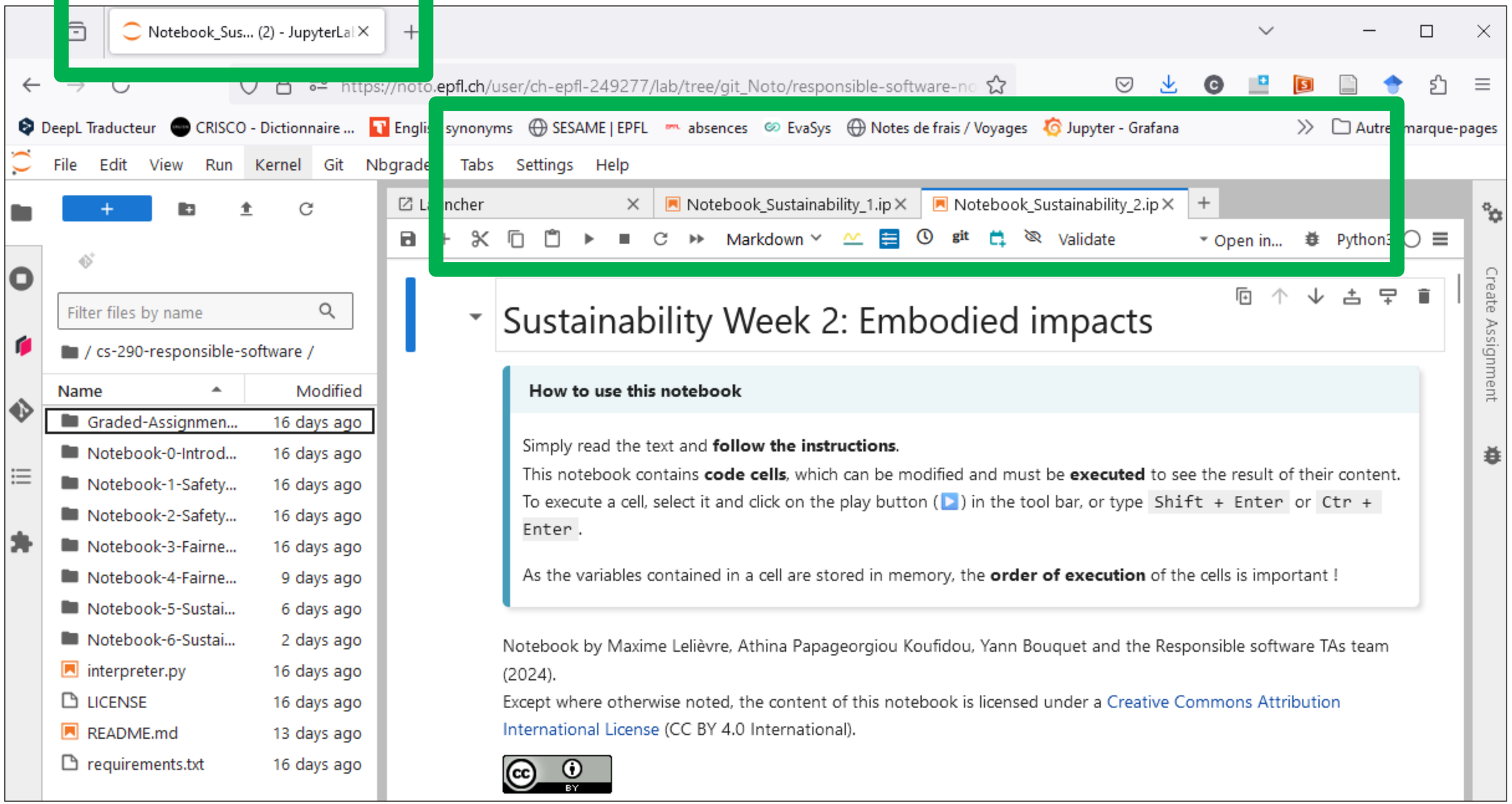
The notebook will be **shorter** than Graded 1.

- **Distribution of the assignment:**
  you will find it directly into your noto workspace
- **Submission of the assignment:**
  you will use moodle as last time

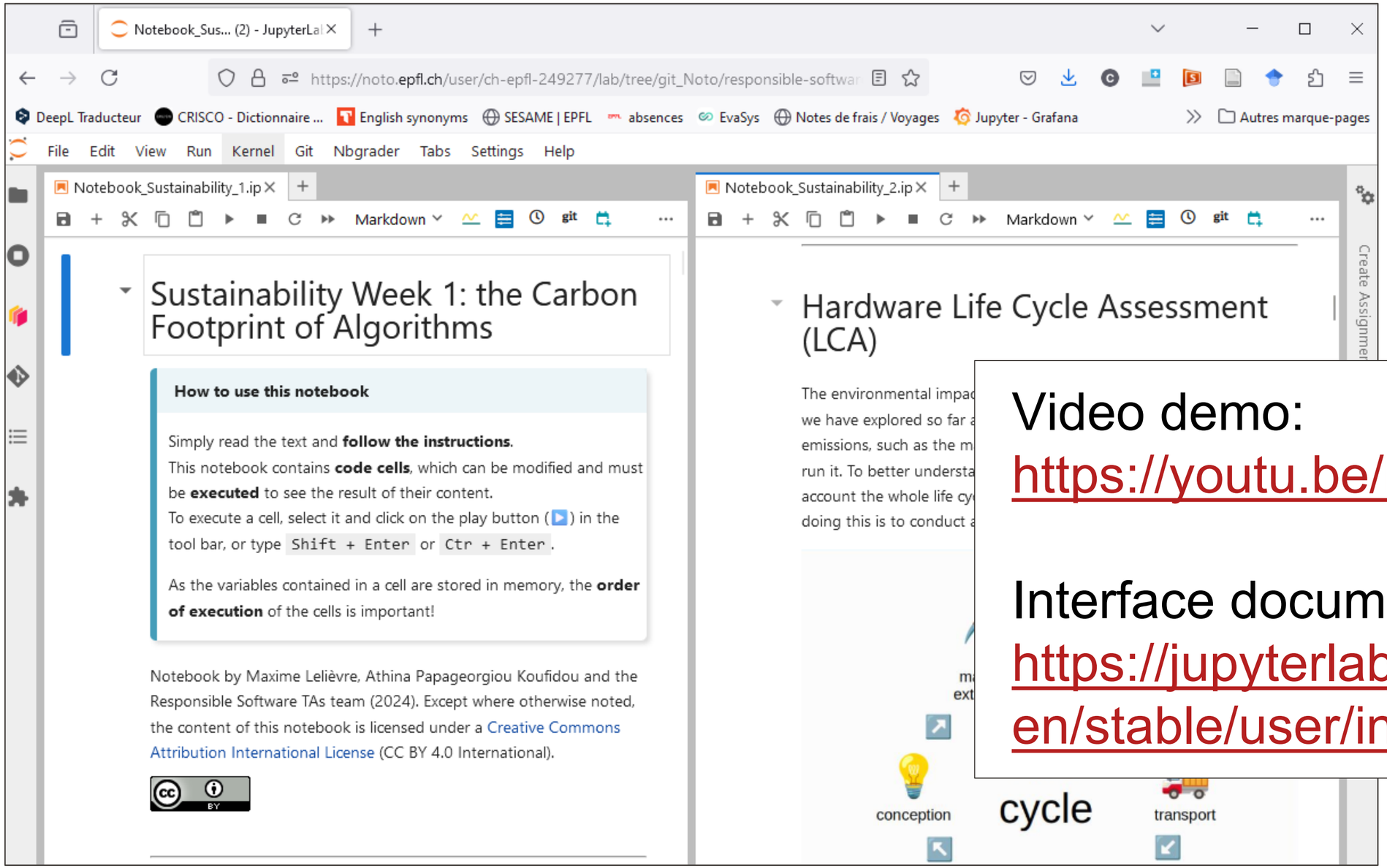
# How to use noto

1. Check that noto works with your **web browser** beforehand
   - Use Chrome / Chromium or Firefox (Safari or Edge not recommended)
   - ⚠️ Some extensions or plugins can prevent noto from working!
2. Open **only 1 window/tab**
   - If you want to see several notebooks, open them in that single window
   - Do not open too many notebooks (max 5)
3. It is normal if the server is a little bit slow at the beginning
   **DO NOT REFRESH the page** or **open a new tab/window** on noto – it can make it worse!
   👉 Wait 1 minute without clicking anything
   👉 If it is still not loading, report to an assistant

# How to open several notebooks? Option 1



# How to open several notebooks? Option 2




Video demo:
https://youtu.be/Ka8qS7CO1XQ

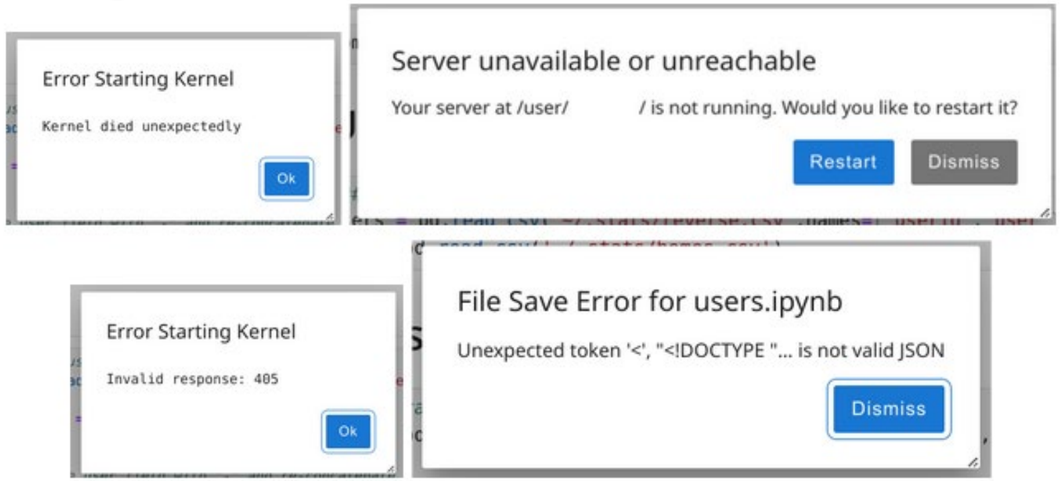
Interface documentation:
https://jupyterlab.readthedocs.io/en/stable/user/interface.html

# In case of technical issue with noto

Make sure to **report it to an assistant** <u>as soon as possible</u>

- DO NOT REFRESH the page or open a new tab/window on noto – it can make the problem worse
  - 👉 Follow the instructions of the assistant

- If you see a popup with an error message:
  - 👉 **DO NOT CLOSE the popup**
  - 👉 **<u>Stop working!</u> Your work is NOT SAVED anymore!**
  - 👉 Report to an assistant

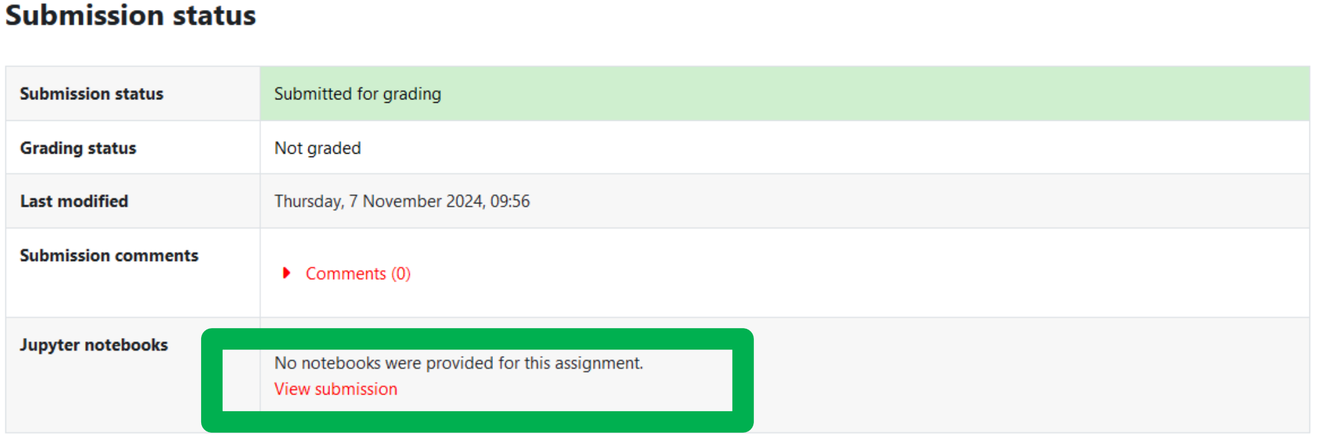


# Checking what you have submitted

1. Go to moodle:
   - Find the assignment “Test the submission process” [**Fairness 1 section**]
   - Click on it 👉 View submission
   - Select **where to save this copy of your submission** on noto

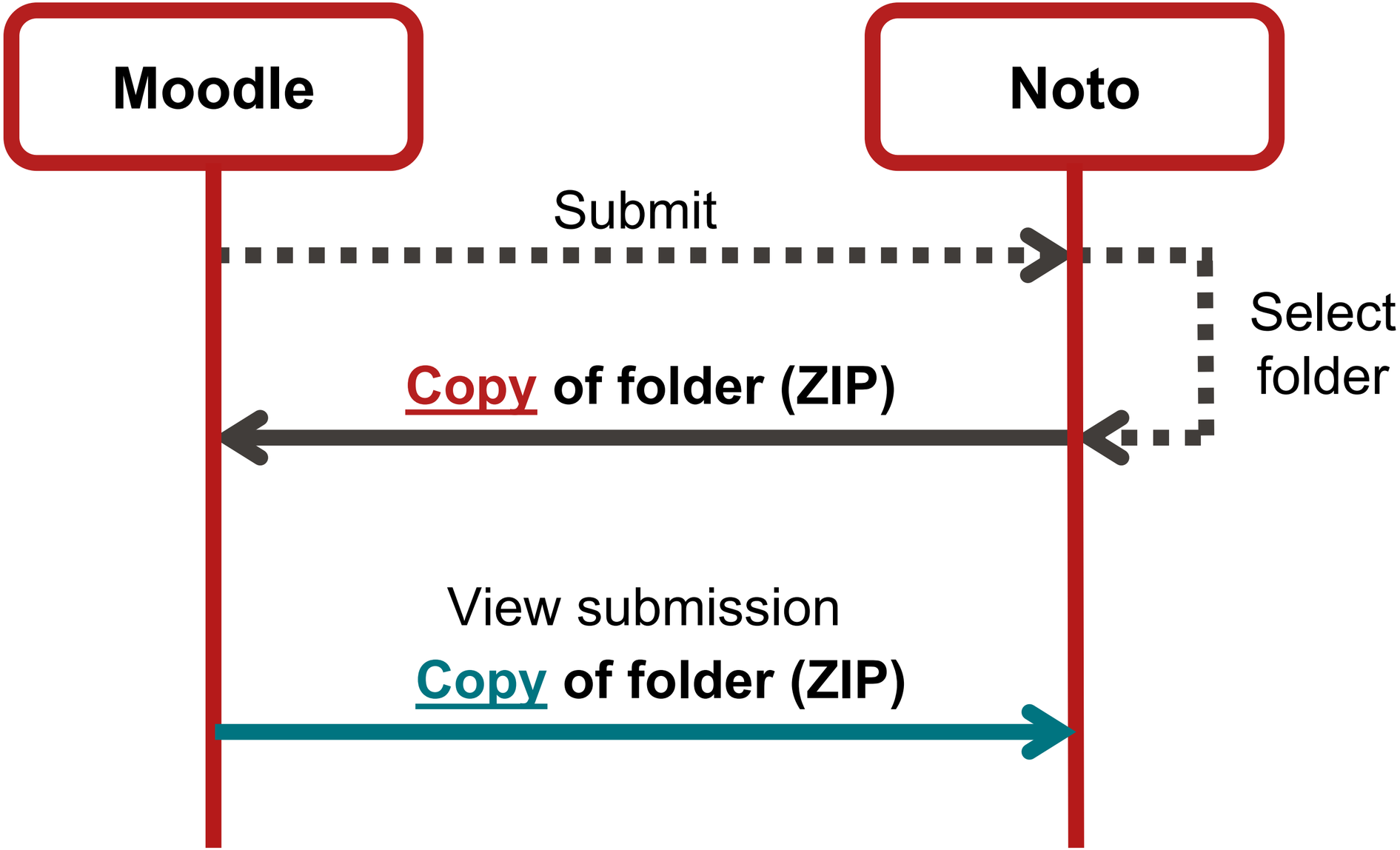
     ⚠️ Choose a folder that is DIFFERENT from "Notebook-3-Fairness-1"!

     i.e., do not put this copy back into its origin folder…




2. Go to noto:
   - Find the folder where you have put the copy of your submission
   - You should see a new folder "Notebook-3-Fairness-1" with your notebook

# How does it work?

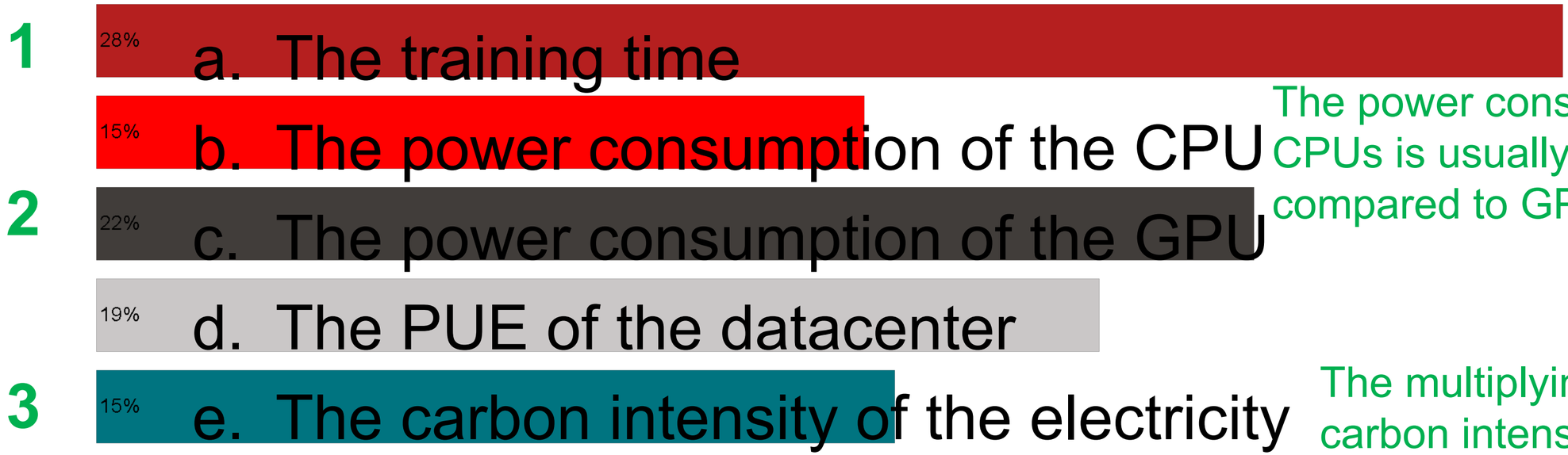


# Review questions Sustainability 2

# The footprint of training - 1

URL: ttpoll.eu
Session ID: cs290

What are the 3 most important elements in the carbon footprint of ML training?
Rank them **by decreasing impact** (i.e. most impactful first) :



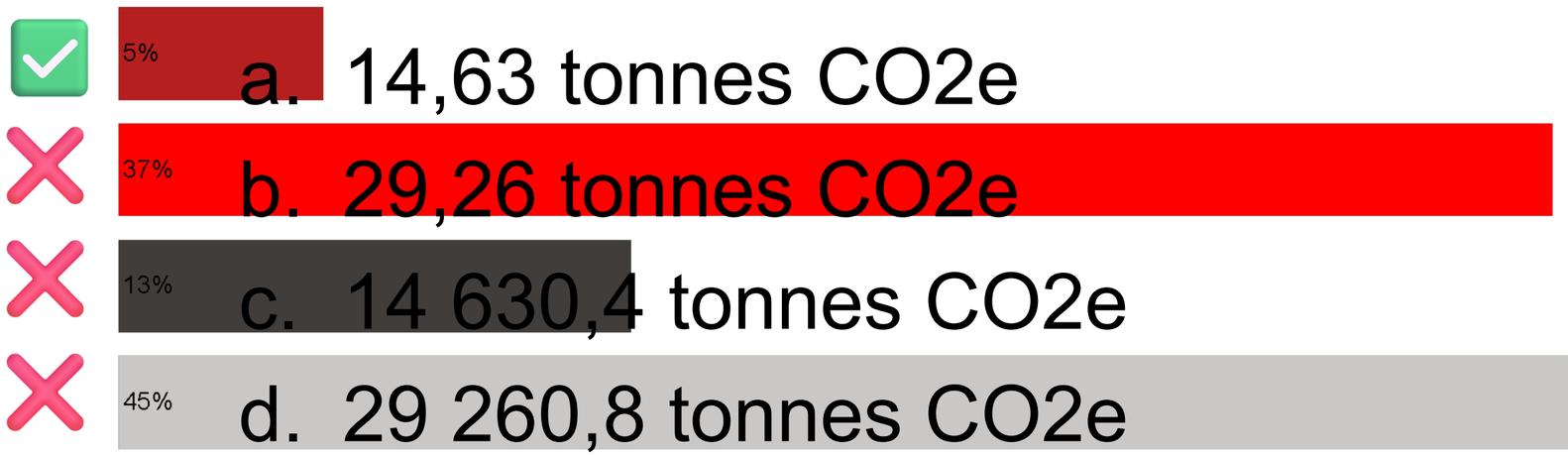

The power consumed by CPUs is usually negligible compared to GPUs

The multiplying factor from carbon intensity is usually higher than that of the PUE

# The footprint of training - 2

URL: ttpoll.eu
Session ID: cs290

Let’s consider the training of the model SupChat-7B. The computing node has 2 GPUs of the model Nvidia A100 80GB, which consume 400W each. Our datacenter, which has a PUE of 1.2, is located in Germany (carbon intensity: 381g CO2e / kWh).
The training time is 80 000 hours of total GPU computation time.
What is the carbon footprint for the training of SUPMOD-7B?




The GPU time already accounts for the number of GPUs

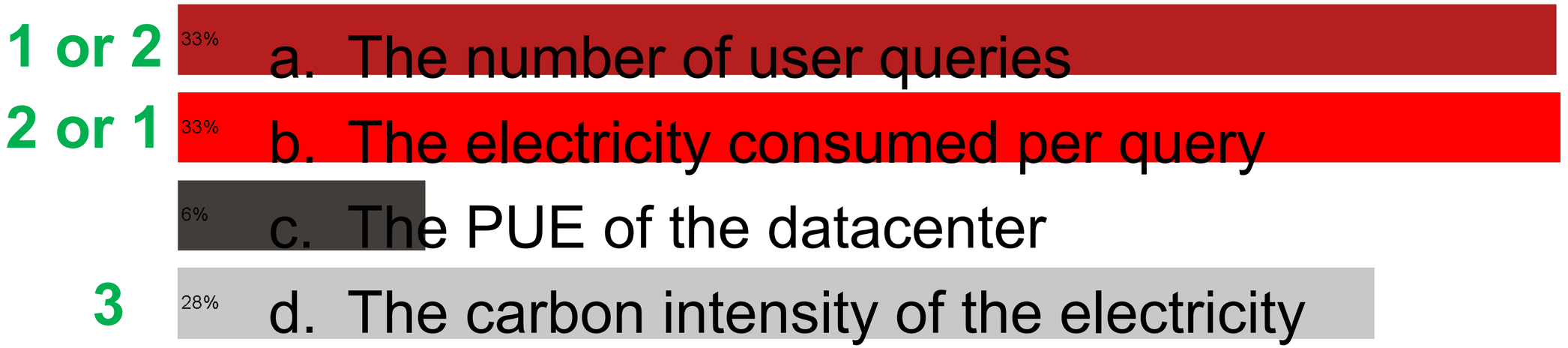
- 400W converted to 0,400 kW beforehand since the carbon intensity is in g CO2 / kWh
- 80000 x 0,400 x 1,2 x 381
- then conversion from grams to tonnes (/ 1000000)

⚠️ units!

# The footprint of inference - 1

URL: ttpoll.eu
Session ID: cs290

What are the 3 most important elements in the carbon footprint of ML inference?
Rank them **by decreasing impact** (i.e. most impactful first) :



# The footprint of inference - 2

The model SupChat-7B is now deployed in production. It is hosted on the same computing node with 2 GPUs of the model Nvidia A100 80GB, which consume 400W each. Our datacenter, which has a PUE of 1.2, is located in Germany (carbon intensity: 381g CO2e / kWh). Our model is able to serve 120 token per second of computation time. It has an average of 5000 users daily and generates an average of 5000 tokens per user per day.

1. What is the total GPU computation time used over 1 day (in h)?
2. What is the power consumed by the model for inference (in W)?
3. What is the total electricity consumed over 1 day (in kWh)?
4. What is the carbon footprint over 1 day (in kg CO2e)?

# The footprint of inference - 2

1. The total computation time used over 1 day can be obtained from the speed of the model and the total number of tokens served per day + you need to convert from seconds to hours

$$GPUtime_{perday} = \frac{nbusers_{perday} \times nbtokens_{peruser_{perday}}}{modelspeed} \times \frac{1}{3600}$$

$$GPUtime_{perday} = 57{,}87\ h$$

NB: ⚠️ this number of hours is not realistic for 1 day, I made up the numbers and didn't realize it could not work out - sorry for that…

# The footprint of inference - 2

2. The power consumed by the model can be obtained from the number of GPUs, the power per GPU and the PUE to account for cooling:

$$Power_{inference} = nbGPUs \times power_{GPU} \times PUE$$

$$Power_{inference} = 960\ W$$

NB: ⚠️ here we hypothesize that the 2 GPUs are used by the model at the same time, which means the total time we have computed at step 1 is for 2 GPUs, so we account for that in the power consumption (watch out not to account for it twice…)

# The footprint of inference - 2

3. To get the electricity consumed we multiply equation 1 with equation 2, and we scale to kW (i.e. divide by 1000):

$$Electricity_{inference} = \frac{57{,}87 \times 960}{1000}$$
$$Electricity_{inference} = 55{,}56\ kWh$$

3. To get the carbon footprint we multiply equation 3 by the carbon intensity of the electricity, then we scale to kg (i.e. divide by 1000)

$$Footprint_{inference} = \frac{55{,}56 \times 381}{1000}$$
$$Footprint_{inference} = 21{,}16\ kg\ CO_2e$$

# Total carbon footprint

URL: ttpoll.eu
Session ID: cs290

We have obtained the carbon footprint of SupChat-7B at training and at inference time. What is its total carbon footprint?

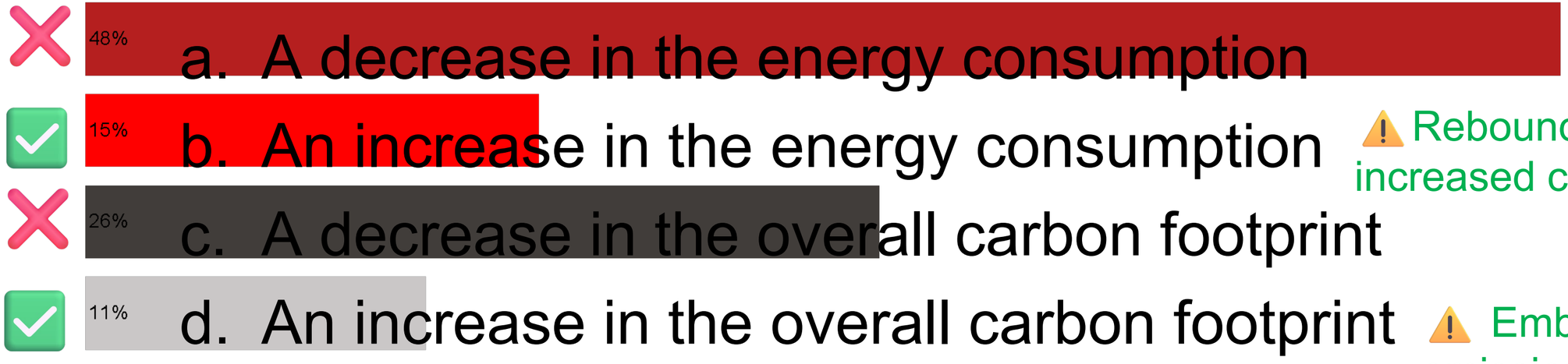
- ❌ a. Training
- ❌ b. Inference
- ❌ c. Training x Inference
- ❌ d. Inference – Training
- ✅ e. Training + Inference
- ✅ f. Other

We also need to add the footprint associated with embodied emissions (i.e. hardware manufacturing mainly)

# Hardware renewal

URL: ttpoll.eu
Session ID: cs290

We want to optimize the energy consumption of SupChat-7B at inference time. We decide to upgrade our hardware platform and to replace our A100 GPUS with H100 GPUs. The H100 are 4 times more performant than the A100 in terms of computation speed. Their power consumption is 700W at maximum use.
What effect(s) are we likely to observe (select all that apply)?



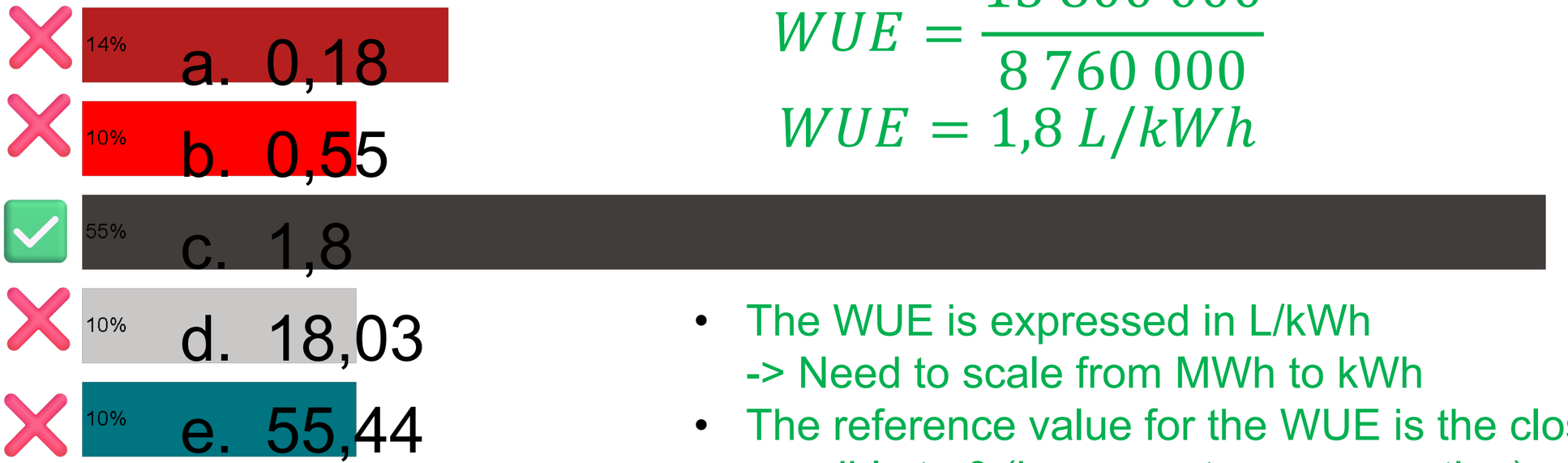

⚠️ Rebound effect (+ increased cooling needs)

⚠️ Embodied emissions!

# Water Usage Effectiveness

URL: ttpoll.eu
Session ID: cs290

The datacenter hosting SupChat-7B consumes an average of 1 MW.
This means annually a total of 8 760 MWh of electricity.
It consumes approximately 15.8 million liters of water each year.
What is the WUE of the datacenter?




$$WUE = \frac{15\,800\,000}{8\,760\,000}$$

$$WUE = 1{,}8\ L/kWh$$

- The WUE is expressed in L/kWh
  -> Need to scale from MWh to kWh
- The reference value for the WUE is the closest possible to 0 (i.e. no water consumption)

https://dgtlinfra.com/data-center-water-usage/

# Case studies

# Where to find the cases?

1. Go to **moodle**

2. Find the **link to the case studies** for today: **Sustainability 2**

3. Download:
   - The **instruction sheet**
   - 1 cheatsheet: Ethical Decision Making

**+ From previous chapters**, you will need:
   - Causal Loop Diagram (2 - Safety 2)

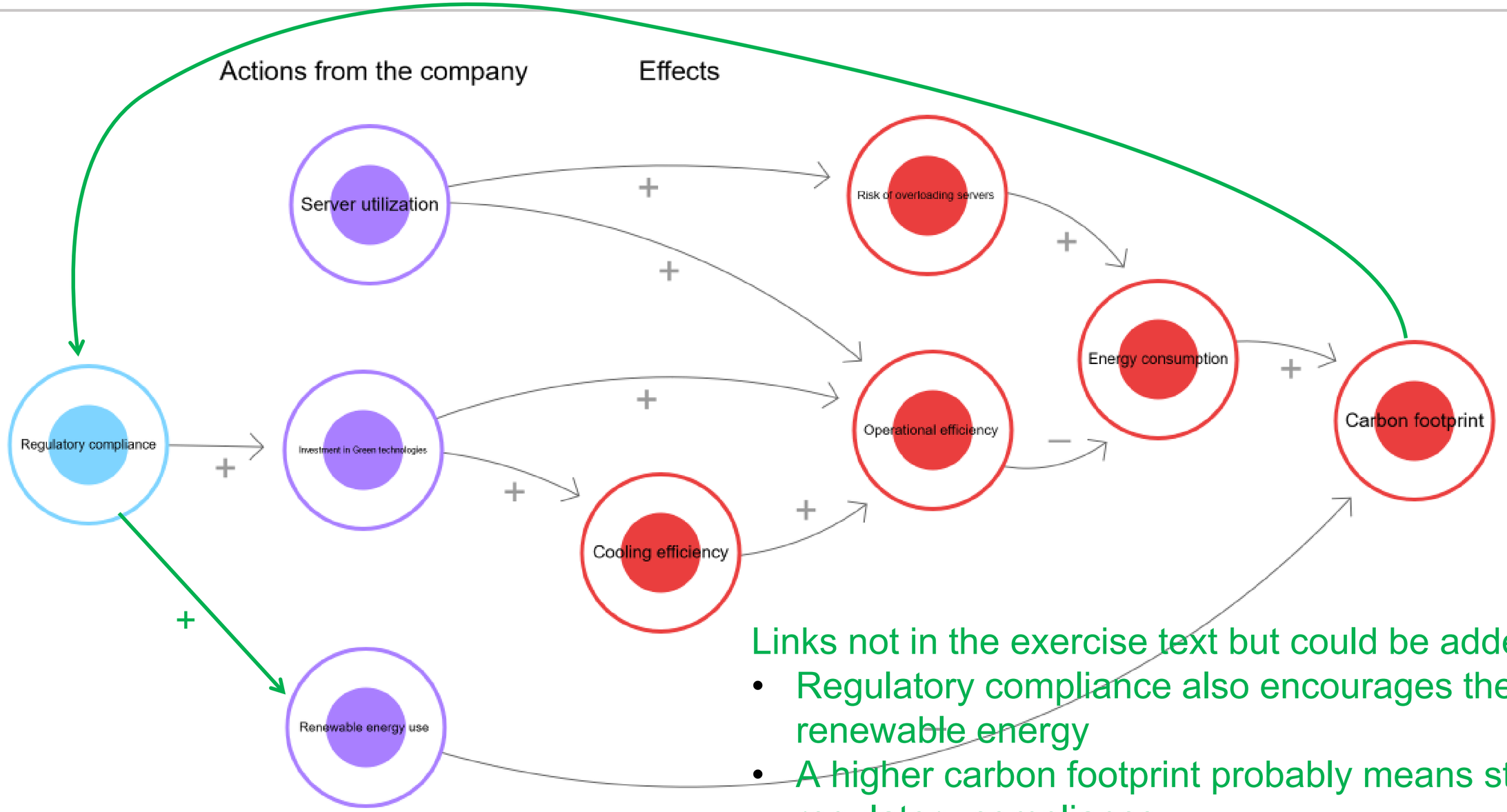
# **Causal Loop Diagram**
(review from Safety 2)

# Instructions – Creating the diagram

- Read the context description
- Build a causal loop diagram with the following variables (you need to add the causal links):
  1. carbon footprint
  2. energy consumption
  3. renewable energy use
  4. investment in green technologies
  5. regulatory compliance
  6. operational efficiency
  7. cooling efficiency
  8. server utilization
  9. risk of overloading servers

# Resulting diagram

https://go.epfl.ch/su2-case1




Links not in the exercise text but could be added:

- Regulatory compliance also encourages the use of renewable energy
- A higher carbon footprint probably means stricter regulatory compliance

# Instructions – Using the diagram

Using the diagram, answer the following questions:

1. How would a significant increase in **investment in green technologies** affect the overall system ?
2. What happens to **operational efficiency** if the server utilization decreases ?
3. What happens to the **energy consumption** if the carbon footprint increases ?
4. Which actions taken could **decrease** the overall carbon footprint?
5. What is the effect of stricter regulatory compliance on investment in green technologies and renewable energy use?

# Ethical Decision Making

# Instructions

- Read the context description
- Fill the table

| **Ethical lens** | **Justification** | **Option chosen** |
|---|---|---|
| Rights | | |
| Justice | | |
| Utilitarian | | |
| Common good | | |

# Ethical lens: Rights

**Which option best respects the rights of all who have a stake?**

👉 1 post =

- India / Switzerland
- Justification according to this lens

See posts on SpeakUp

**Post your ideas:**
https://speakup.epfl.ch
Room key: **68415**

# Ethical lens: Justice

**Which option treats people fairly, giving them each what they are due?**

👉 1 post =

- India / Switzerland
- Justification according to this lens

See posts on SpeakUp

**Post your ideas:**
https://speakup.epfl.ch
Room key: **61431**

# Ethical lens: Utilitarian

**Which option will produce the most good and do the least harm for as many stakeholders as possible?**

👉 1 post =

- India / Switzerland
- Justification according to this lens

See posts on SpeakUp

Key issue = how to count, number of people and/or severity of harm

**Post your ideas:**
https://speakup.epfl.ch
Room key: **44366**

# Ethical lens: Common good

**Which option best serves the community as a whole? And the most vulnerable?**

👉 1 post =

- India / Switzerland
- Justification according to this lens

  See posts on SpeakUp

  Key issue = which community to consider (e.g. India, the social media users, the world…)

**Post your ideas:**
https://speakup.epfl.ch
Room key: **22633**

# What’s next?

# We start Empowerment 1!

Tomorrow, Tuesday 19: notebook on automation bias

By Monday 25:

- Watch **videos 7.1 to 7.4** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)