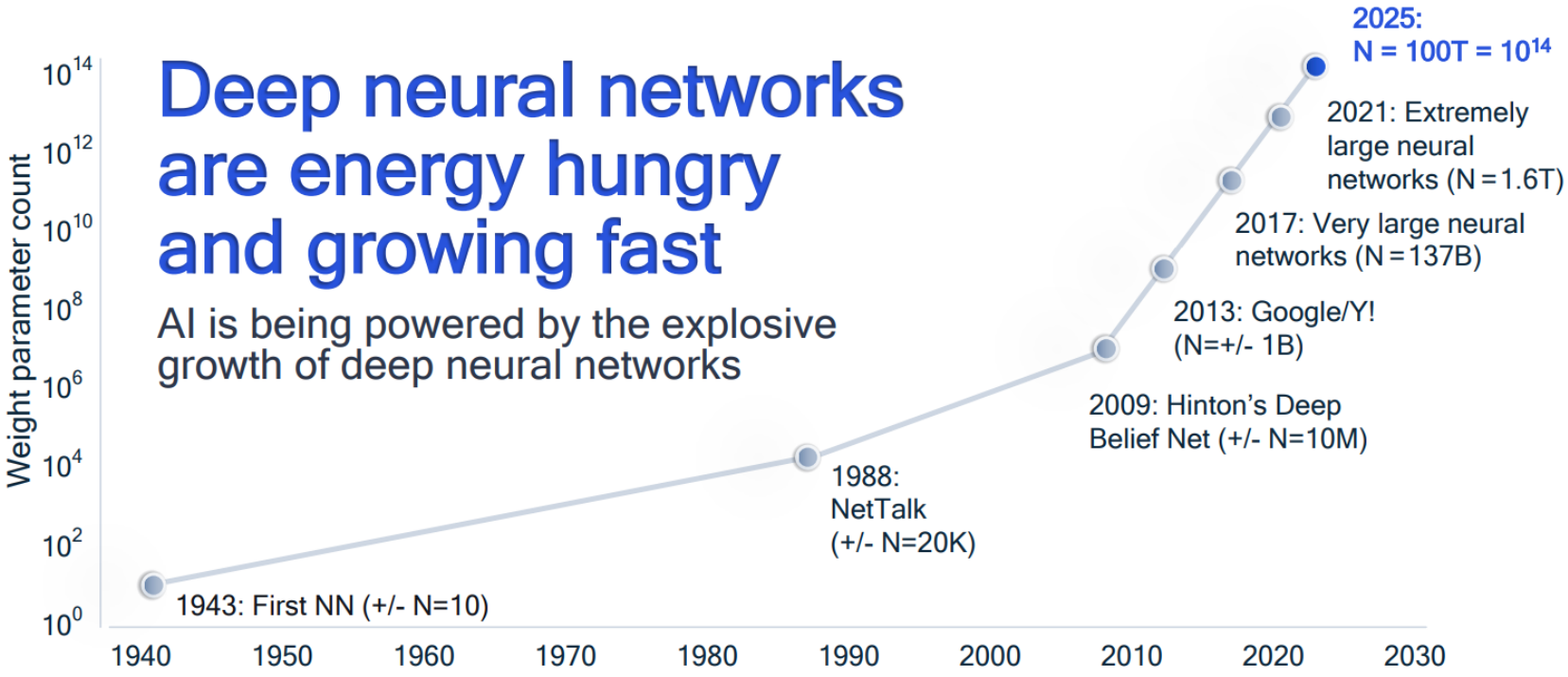
Integrating environmental impact of AI in a data center

Paul Gay, Anne-Laure Ligozat, Éric Bilinski

LISN

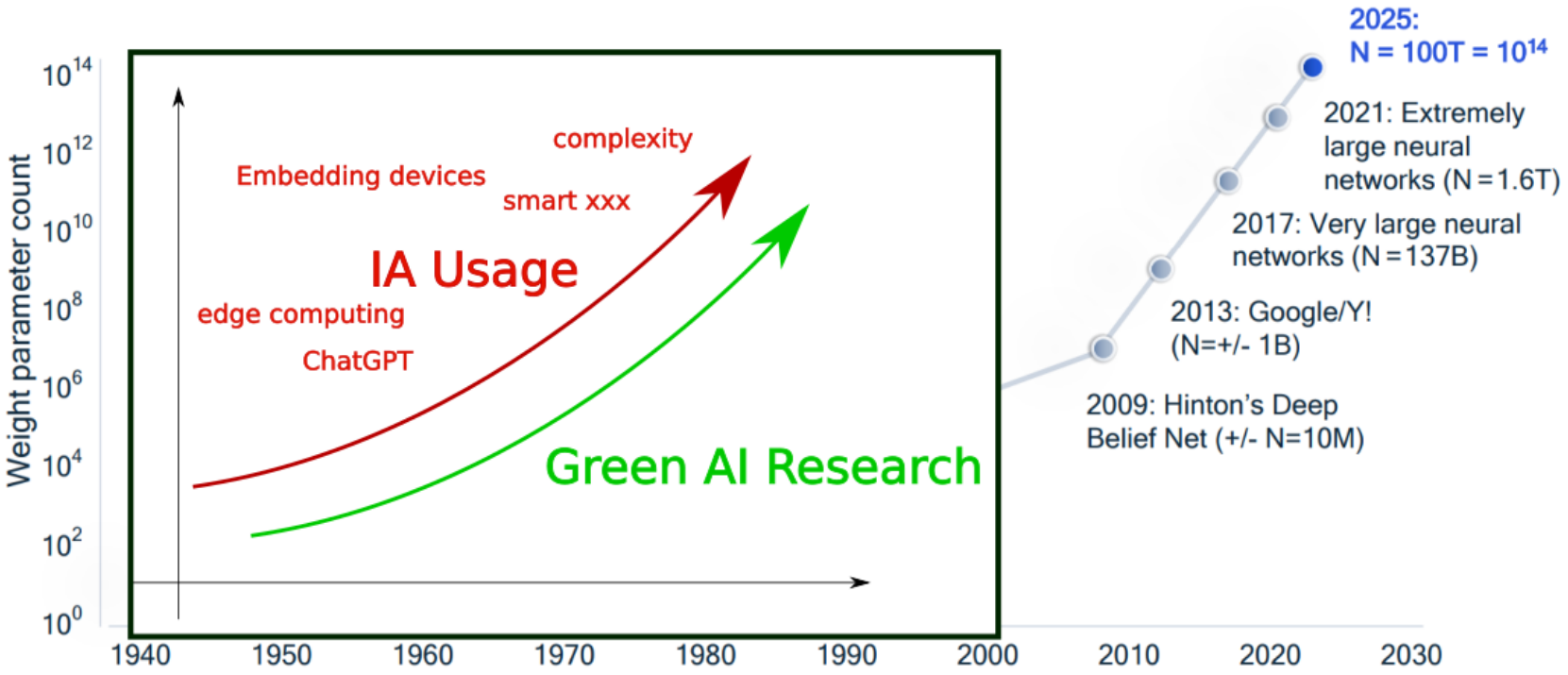
June 2023

AI impact



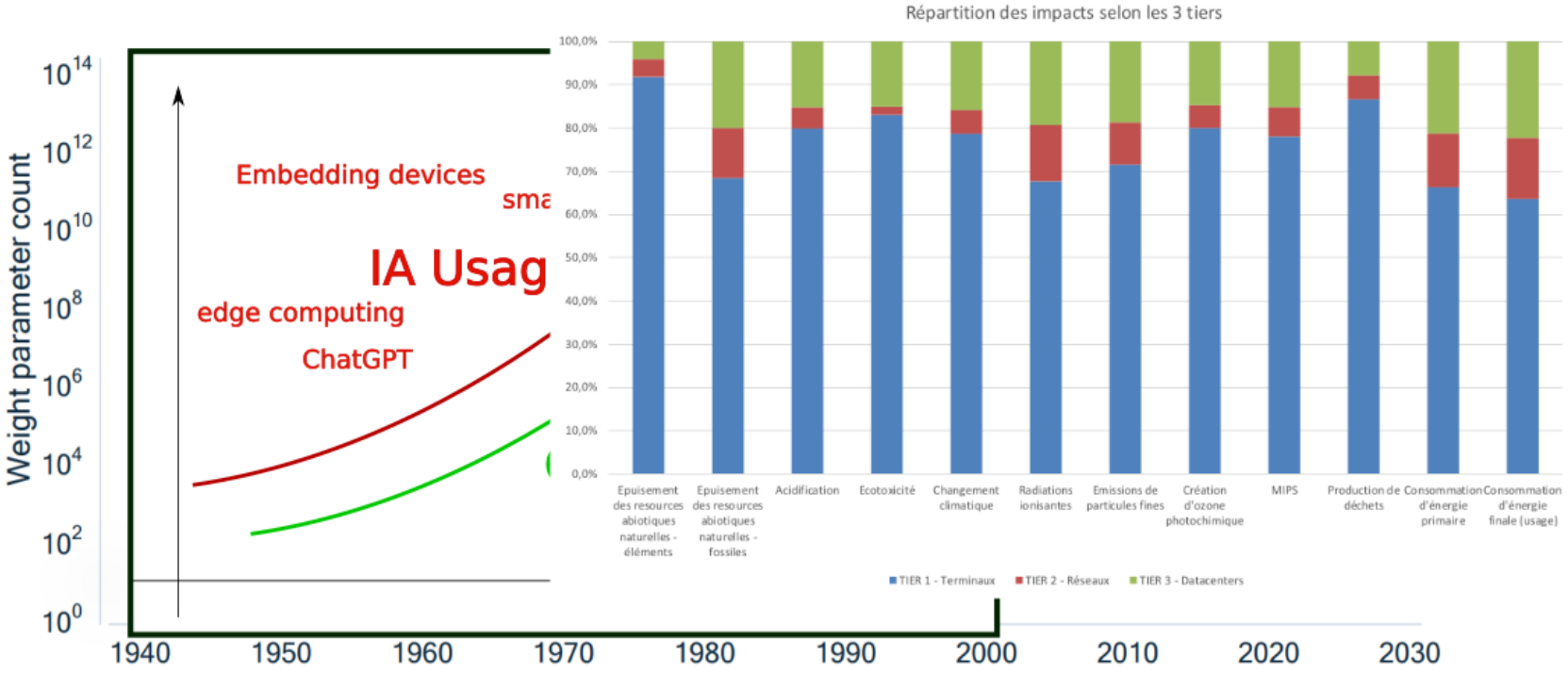
- Exponential growth of Deep Learning models [Fournarakis, 2021]

AI impact



- Exponential growth of Deep Learning models [Fournarakis, 2021]
- Jevons paradox

AI impact



- Exponential growth of Deep Learning models [Fournarakis, 2021]
- Jevons paradox
- Direct cost mainly for manufacturing and usage of IOT (Arcep/Ademe 2021)

Coca4AI Motivation

one year project funded by RFSI

We plan to build multi-scale models
to measure carbon footprint
and evaluate a compromise between environmental impact
and algorithm accuracy
for deep learning algorithms
at the scale of a (small) data center

Applications to

- Data scientists behavior
- Students (IAPau, universities)
- Summer School UPPA 2022, 22/06!

Our cluster

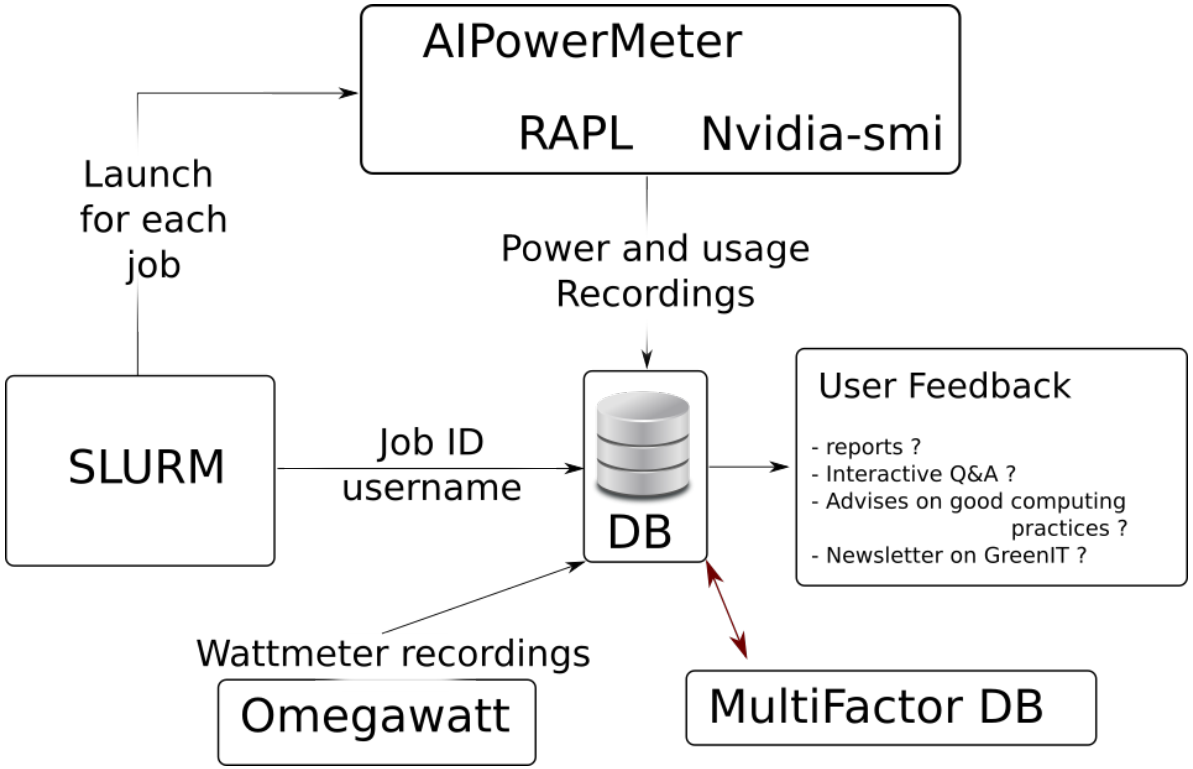
The LabIA cluster

- 12 GPUs equipped nodes dedicated to AI
- Slurm based job management
- over a year:
  - 99 users from 5 laboratories: LTCI, LISN, SAMOVAR, L2S, SATIE
  - 20k jobs
- Software tools: Zabbix (network), Idrac (gross power estimation), Graphana

Objectives

- Data collection: Statistics per job and per user
  - GPU and CPU Power Consumption for each job
  - + PowerMeter on each node
  - Source code of each job
- Identifying behaviors
  - How to trace them? From source code?
- Insights for small and large scale datacenters (Jean-Zay)

Setup for the Lab-IA Cluster



Inference phase

Object Detector	CNN	Vision Transformer	Text Transformer
Yolov5s	Resnet	VIT_B_16	Bert
0.61	0.27	0.94	0.07

Table 1: GPU joule consumption for one inference (check Aipowermeter doc for experimental details).

- Easy monitoring of order of magnitude
  - 10^{12} yolo inferences $\approx$ 50 km by car

Data provided to users

User Feedback?
Under development ...

- Updated summary in text format on terminal
- Dashboard resuming the consumption usage over the last month
- LCA, Tree of consequences

Inference phase

Many joules wasted in data scientists practices (Khan 2019)

- Job crashing
- Brute force optimization to earn a few percents
- Hidden knobs and bad use of the GPUs

Spread good practices

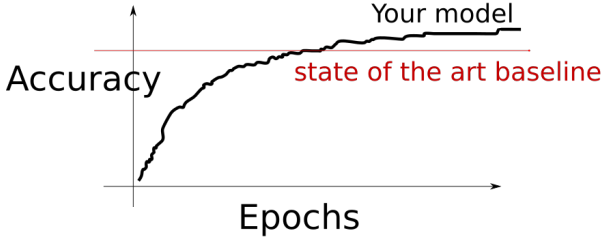
- Avoid GPU bottlenecks
- Normalize your loss to avoid cuda runtime error
- Normalize your layers also
- Most of the time spent by building auxiliary code (evaluation metrics, data formatting, rect or square inference)
  - Training is slow: multiple days
  - You obtain most of the clues with small experiments and unitary tests
- Detect inefficient use of GPU

Engagement

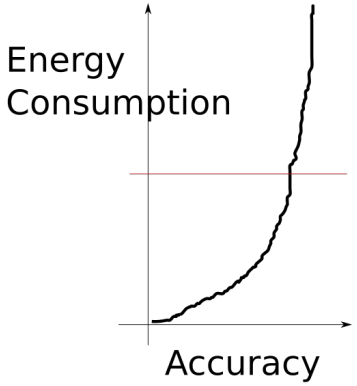
- How to create engagement ?
  - How to be legimitate ?
  - How to answer the politic question ?
  - How to avoid the "I am technical" rejection ?

Engagement

- How to create engagement ?
  - How to be legimitate ?
  - How to answer the politic question ?
  - How to avoid the "I am technical" rejection ?
- Providing metrics which matter from an environmental point of view



Common view of machine learning



Change of focus

- A playground to create dialog (about rebounce effect)
- Destroy the hype of ICT ?

Discuss consequences of IT

- Carbon footprint
- Abiotic resources depletion
- Toxicity
- Water depletion
- Energy used

What about carbon footprint ?

Advantages

- **Global** change
- Hype and Popular
- More extensively studied

Let's skip for once and for today

What about mineral resource depletion?

- France 2022 : 9.48E+05 Kg Sb eq (Source Arcep 2022)
- Kg Sb eq $\rightarrow$ Kilogs équivalent antimony
- Antimony is a mineral often used to make lead
- World wide production of antimony: 100k tons (copper is 20M), mainly China and Russia (Statista, 2023)

Relation with the dozens metals used in ICT is not trivial

Which metal are in ICT?

- Coppper : 170K tons en 2017 for the French electricity network, 30K tons for offshore windpower. **28Mt world wide production**
- Tantale : capacitors : 1000 condensers in an iphone10. **1800 tons**
- Indium : screen covers.
- Gallium and Germanium : power amplifiers. **320 tons and 106 tons**

Metals are vitamins : small quantities, more efficient products.

Source : Gaetan Lefevre La consommation croissante en matières premières du numérique : l'urgence d'une prise de conscience. 2019

Mineral Depletion ?

- The constraint comes from the market and not from the resources
- Most of the metal vitamins are subproduct of larger industries
- Inegality among the countries
- Competition with other usages. Gallium is used to build solar panels.

What about toxicity ?

- France 2022 : 2.63 E+11 CTUe (Source Arcep 2022)
- CTU : Comparative Toxic Units

Various effects summarised in one number

- Mercure dams, Acid mine drainage
- Issue of the after mining period
  - Desertion and state intervention
- Recylcing
- Others

However, there are local consequences...

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- France : 2.7 millions m^3
- World wide : 75 millions m^3

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- France : 2.7 millions m^3
- World wide : 75 millions m^3

Some orders of magnitude

- In France : 4,1 billions m^3 per year
- 2.3 billions m^3 for Agriculture
- 8 millions m^3 for artificial snow in Savoie in 2021
- GLENCORE : 331 millions m^3

again, these consequences are local

Source : Prélevée ou consommée : comment compter (sur) l'eau ? last cheked the 29th/032023 | Commissariat général au développement durable. online article.

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- France : 2.7 millions m^3
- World wide : 75 millions m^3

Some orders of magnitude

- In France : 4,1 billions m^3 per year
- 2.3 billions m^3 for Agriculture
- 8 millions m^3 for artificial snow in Savoie in 2021
- GLENCORE : 331 millions m^3

again, these consequences are local

Source : Prélevée ou consommée : comment compter (sur) l'eau ? last cheked the 29th/032023 | Commissariat général au développement durable. online article.

How to attribute water consumption from mining ?

What about rebound effect?

The scope of Arcep study does not include the digitalisation of the other sectors

Contradictory results from other studies

GSMA reports and the 1/10 ratio

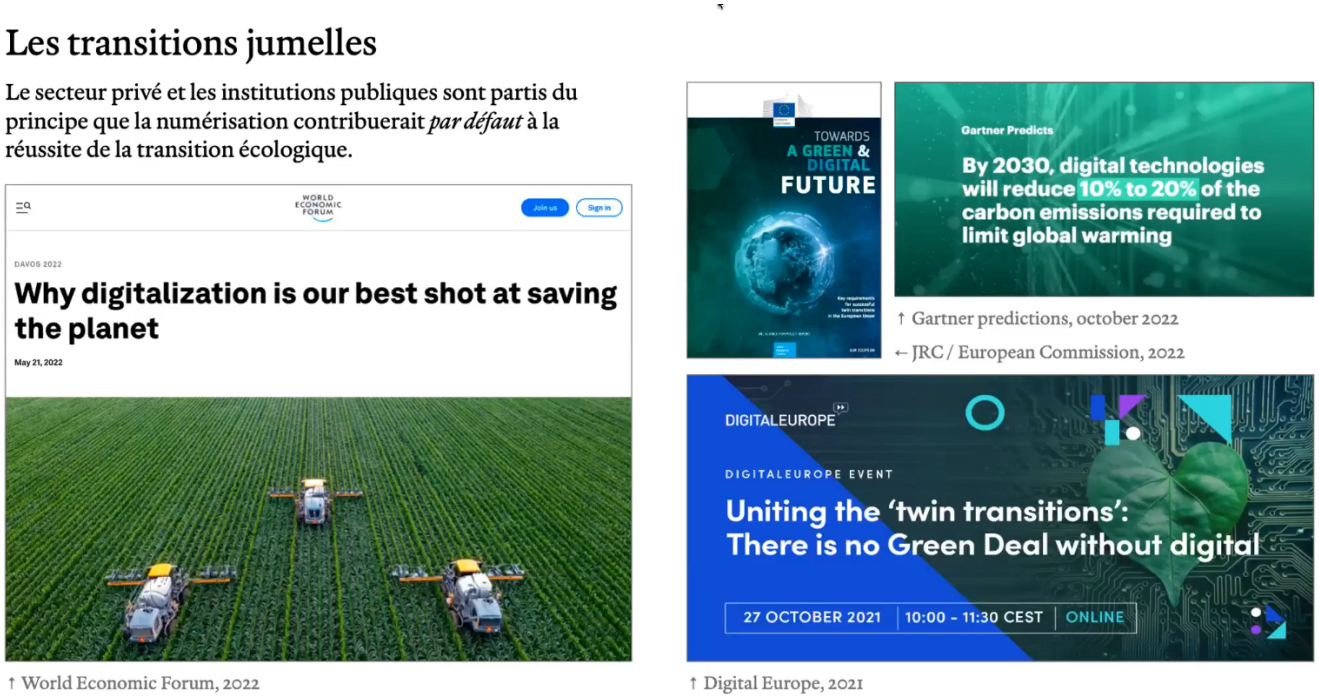


Figure 2: Slide from Gauthier Roussilhe, *La numérisation aide-t'elle la transition écologique* https://labos1point5.org/les-seminaires

- smart building (-40%), smart agriculture (-65%), remote medicine, airbnb,...

Problems of estimating rebounce effect [Rasoldier et al., 2022]

- Researchers tend to avoid world wide prospectives, because it is not reliable
- Comparison with a worst case scenario
- Only one environmental factor is considered
- End of cycle is difficult to take into account

Impact assessment of AI project

[Lefèvre et al., 2023]

- Estimate the impact of your research
- including:
  - Devices for initial training
  - Devices while in production
  - Life cycle assessment
  - Resilience, quality of service

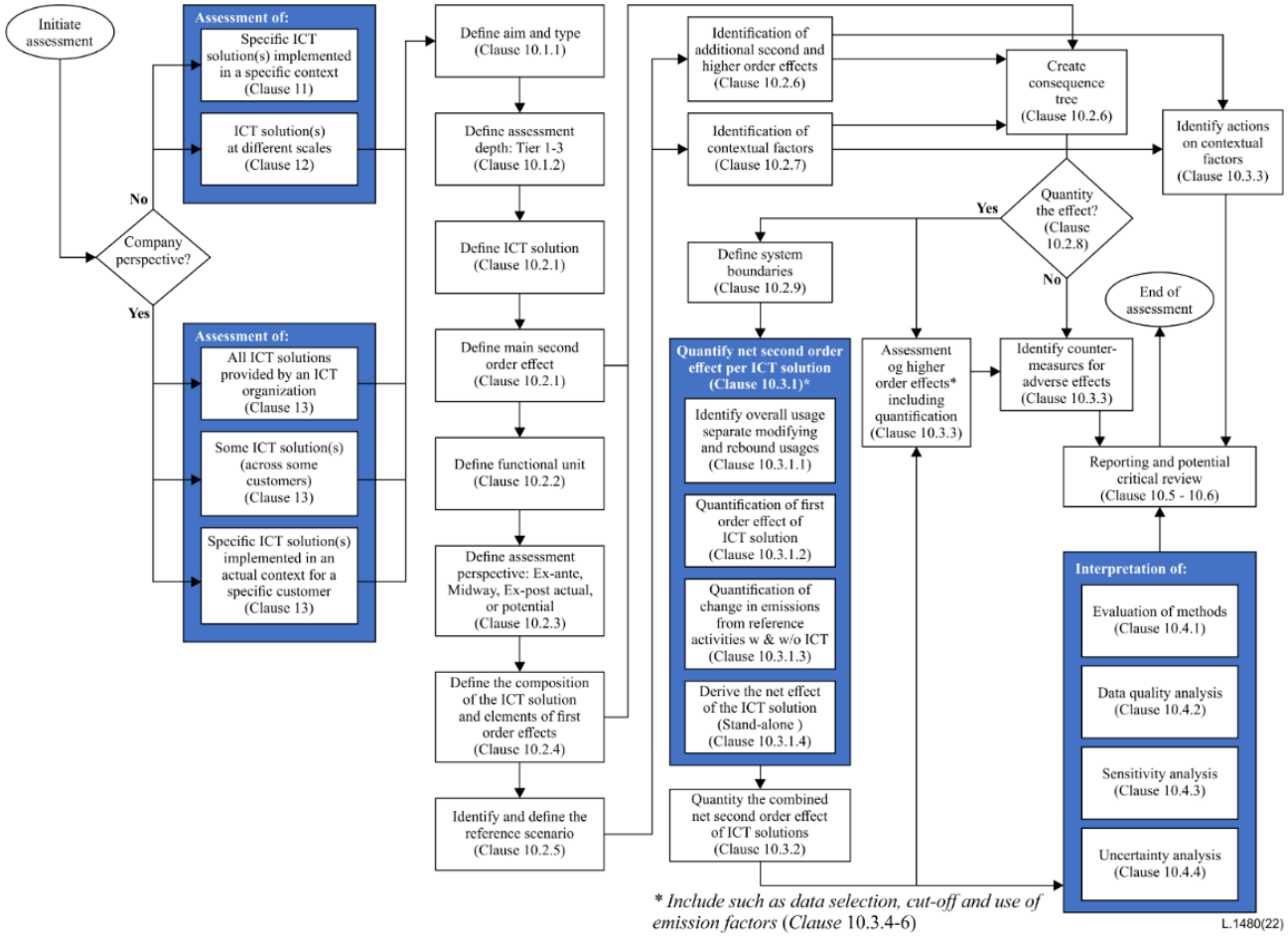


Figure 3: Recommendation ITU-T L.1480. 2022

Prospective exercice : imagine the future

Thanks for your attention
Questions?

References I

[Fournarakis, 2021] Fournarakis, M. (2021).
A practical guide to neural network quantization.

[Lefèvre et al., 2023] Lefèvre, L., Ligozat, A.-L., Trystram, D., Bouveret, S., Bugeau, A., Combaz, J., Frenoux, E., Guennebaud, G., Lefèvre, J., Nicolaï, J.-P., and Dassas, K. (2023).
Environmental assessment of projects involving AI methods.
working paper or preprint.

[Rasoldier et al., 2022] Rasoldier, A., Combaz, J., Girault, A., Marquet, K., and Quinton, S. (2022).
How realistic are claims about the benefits of using digital technologies for ghg emissions mitigation?