High Performance Language Technologies

Understanding and measuring the environmental impact of Natural Language Processing

Aurélie Névéol

aurelie.neveol@lisn.upsaclay.fr

HPLT Winter School, February 6, 2024

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
université PARIS-SACLAY
cnrs

Where am I speaking from?

- Natural Language Processing aplied to the biomedical domain
  - Focus on ethical practices
  - Needs to scale up
  - Needs to be resource-aware

- I share an office with Anne-Laure Ligozat 

TLDR;

On the path to greener AI:
involving all stakeholders
measuring impact
informing decisions

Natural Language Processing has enabled powerful advances

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature communications > articles > article

Article | Open access | Published: 01 September 2020

Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský

Nature Communications **11**, Article number: 4381 (2020) | Cite this article

Natural Language Processing has enabled powerful advances

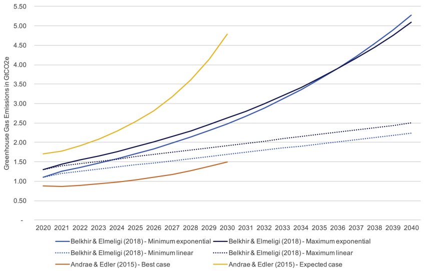
E46 E86 F03 G200

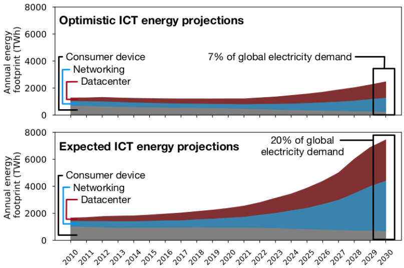
DENUTRITION DESHYDRATATION
DEMENCE MIXTE EVOLUEE (stade sévère)
Maladie de Parkinson idiopathique
Angioedème membres sup récent non exploré par TDM (a priori pas de cause médicamenteuse)

Robert A, Baghdadi Y, Zweigenbaum P, Morgand C, Grouin C, Lavergne T, Névéol A, Fouillet A, Rey G. Développement et application de méthodes de traitement automatique des langues sur les causes médicales de décès pour la santé publique. Bull Epidémiol Hebd. 2019;(29-30):603-9.

... at a cost

ICT emissions estimated at 1.8%-2.8% of global GHG emissions (aviation, globally: 1.9%)





Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., Friday, A. (2021). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. Patterns, 2(9).
Gupta, U., et al. "Chasing carbon: The elusive environmental footprint of computing." 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2021.

What is the impact of NLP?

- A growing insterest since 2019
- … or not - see NeurIPS author guideline change 2021 vs. 2023

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Strubell E, Ganesh A and McCallum A. Energy and Policy Considerations for Deep Learning in NLP. Proc Annual Meeting of the Association for Computational Linguistics (ACL):3645-3650 (2019).

Ten simple rules to make your computing more environmentally sustainable

- Rule 1: Calculate the carbon footprint of your work
- Rule 2: Include the carbon footprint in your cost–benefit analysis
- Rule 3: Keep, repair, and reuse devices to minimise electronic waste
- Rule 9: Be aware of unanticipated consequences of improved software efficiency

Lannelongue L, Grealey J, Bateman A, Inouye M (2021) Ten simple rules to make your computing more environmentally sustainable. PLoS Comput Biol 17(9): e1009324.

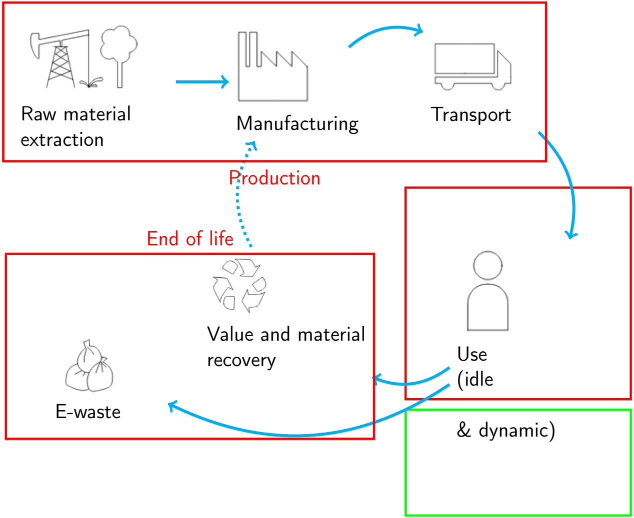
Why measure the impact of experiments?

- Need for sustainable research
- Need for a comprehensive approach to evaluation, beyond leaderboards

Image credit: C. Morand.
Ethayarajh K and Jurafsky D Utility is in the Eye of the User: A Critique of NLP Leaderboards. Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 4846-53. (2020).

How can we measure the impact of experiments?

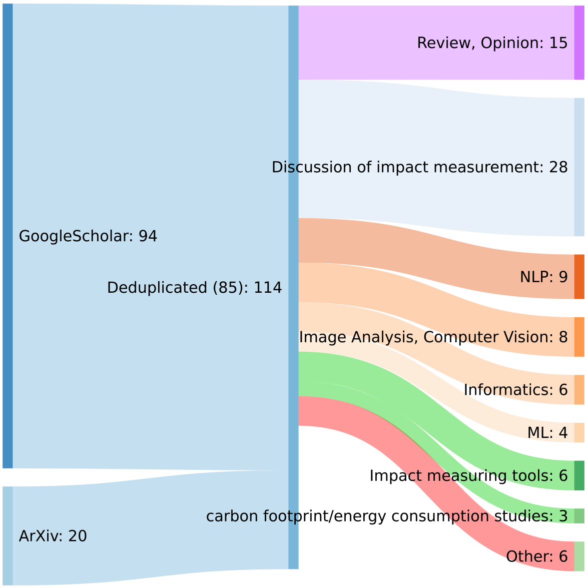
Sources of CO_2 emissions include:



Can a tool provide CO2 impact measurement?

- Literature search:
  - Seed tools: Experiment Impact Tracker, Pyjoules, Carbon Tracker
  - Snowballing in Google Scholar + ArXiv "related papers"
- Selection criteria:
  - Freely available
  - usable in linux/mac OS
  - documented in a scientific publication
  - suitable to measure the impact of NLP experiments
  - CO2 equivalent measure

Literature survey



85 publications reviewed lead to identification of 6 tools providing C02 impact measurement

- Online tools
  1. Green Algorithms
  2. ML CO2 Impact
- Python toolkits
  3. Energy Usage
  4. Experiment Impact Tracker
  5. Carbon Tracker
  6. Cumulator

  2'. Code Carbon

Sources :
Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. ACL Workshop SustainNLP 2021:11-21
Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram and Anne-Cécile Orgerie. An experimental comparison of software-based power meters: focus on CPU and GPU. CCGrid 2023:1-13

Criteria for characterizing tools

- 3 publication criteria
  1. Publication year
  2. Citations (overall, user studies)
- 7 technical criteria
  1. Availability, ease of installation
  2. Documentation, version
- 5 configuration criteria
  1. Source of carbon intensity and power usage effectiveness values
  2. Equipment covered by the measurements
- 2 functional criteria
  1. Sources of emissions targetted
  2. Type of hardware

Features of measurement tools

Feature	online (Green Algorithms)	toolkit (Code Carbon)
direct measure	X	✓
estimation	✓	X
asynchronous	✓	X
comparison on same hardware	$\sim$	✓
easy to install	✓	$\sim$

Modeling the dynamic impacts of computation

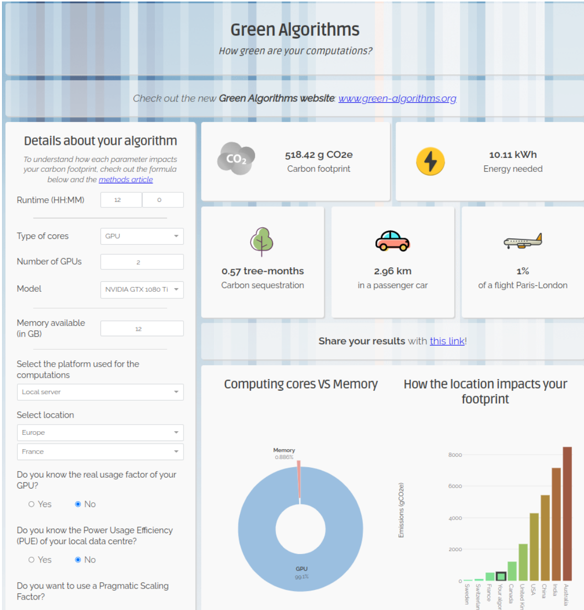
$$E = t * (P_c + P_m) * PUE$$

where:

- t = Running time (h)
- $$P_c$$
  = Power draw of processing cores (W)
- $$P_m$$
  = Power draw from memory (W)
- PUE (Power Usage Effectiveness) = efficiency of data center

Green Algorithm online tool

http://calculator.green-algorithms.org/



Code Carbon python package

https://github.com/mlco2/codecarbon

```
import tensorflow as tf

from codecarbon import EmissionsTracker

mnist = tf.keras.datasets.mnist

(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0


model = tf.keras.models.Sequential([tf.keras.layers.Flatten(input_shape=(28, 28)),
        tf.keras.layers.Dense(128, activation="relu"), tf.keras.layers.Dropout(0.2),
        tf.keras.layers.Dense(10),])

loss_fn = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
model.compile(optimizer="adam", loss=loss_fn, metrics=["accuracy"])

tracker = EmissionsTracker()
tracker.start()
model.fit(x_train, y_train, epochs=10)
emissions: float = tracker.stop()
print(emissions)
```

Code Carbon python package

https://github.com/mlco2/codecarbon

```
import tensorflow as tf

from codecarbon import track_emissions
@track_emissions(project_name="mnist")
def train_model():
    mnist = tf.keras.datasets.mnist
    (x_train, y_train), (x_test, y_test) = mnist.load_data()
    x_train, x_test = x_train / 255.0, x_test / 255.0
    model = tf.keras.models.Sequential([tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(128, activation="relu"), tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(10)])
    loss_fn = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
    model.compile(optimizer="adam", loss=loss_fn, metrics=["accuracy"])
    model.fit(x_train, y_train, epochs=10)
    return model


if __name__ == "__main__":
    model = train_model()
```

Code Carbon python package

Results

```
[codecarbon INFO @ 11:15:30] Energy consumed for RAM : 0.000018 kWh.
                             RAM Power : 5.737926006317139 W
[codecarbon INFO @ 11:15:30] Energy consumed for all CPUs : 0.000044 kWh.
                             Total CPU Power : 14.0 W
[codecarbon INFO @ 11:15:30] 0.000061 kWh of electricity used since the beginning.
3.56889761642147e-06
```

Adding this info in your research paper

https://doi.org/10.1016/j.jbi.2022.104073

Table 2
Overall results on test corpus.

	Precision	Recall	F-Measure	CO_2 equivalent (g.)
Private Model (*MERLOT, teacher model*)	0.852	0.862	0.857	123
Public Model (*DEFT*)	0.592	0.383	0.465	22
Dictionary-based Model (*JDM*)	0.153	0.062	0.089	–
Dictionary-based Model (*UMLS*)	0.246	0.168	0.200	–
Privacy-Preserving Mimic Model (*DEFT, student model*)	0.604	0.743	0.666	30
Privacy-Preserving Mimic Model (*CAS, student model*)	**0.628**	**0.806**	**0.706**	169
Privacy-Preserving Mimic Model (*CépiDc, student model*)	0.580	0.710	0.638	394

Adding this info in your research paper

This algorithm runs in 12h on 2 GPUs NVIDIA GTX 1080 Ti and 12 CPUs Xeon E5-2683 v4, and draws 35.74 kWh. Based in France, and ran 3 times in total, this has a carbon footprint of 1.83 kg CO2e, which is equivalent to 2.00 tree-months (calculated using green-algorithms.org v2.2 [1]).

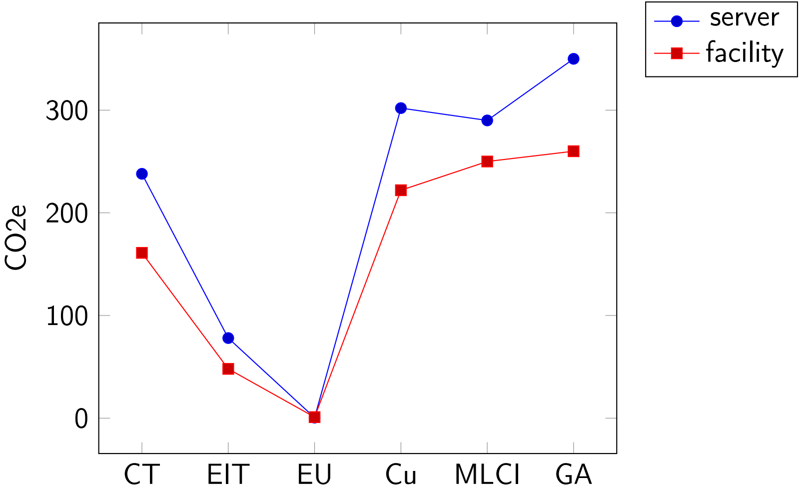
[1] Lannelongue, L., Grealey, J., Inouye, M., Green Algorithms: Quantifying the Carbon Footprint of Computation. Adv. Sci. 2021, 2100707.

Application to a named entity recognition task

- 2 NER tools
  - one that addresses flat entity recognition [Ma and Hovy, 2016]
  - one that addresses both flat and nested entity recognition, introduced by [Yu et al., 2020]
- 2 setups
  - GTX 1080 Ti GPUs used on a server
  - Tesla V100 GPUs used on a computing facility
- 2 datasets
  - QUAERO Broadcast News Extended Named Entity dataset [Galibert et al., 2010] (French press)
  - QUAERO French Med dataset [Névéol et al., 2014]
- 2 measures
  - energy consumption
  - carbon footprint

Results

for [Yu et al., 2020] on the French Press corpus



Why are the results so heterogeneous?

- Carbon intensity varies: CT used the average carbon intensity for EU-28 in 2017 (294.21 gCO2eq/kWh), while electricityMap gives around 30 to 40 gCO2eq/kWh
- Hardware options may not be available
- Tools not adapted to a multi-user setting
- Direct measures vs estimations

Carbon intensity variations throughout the day

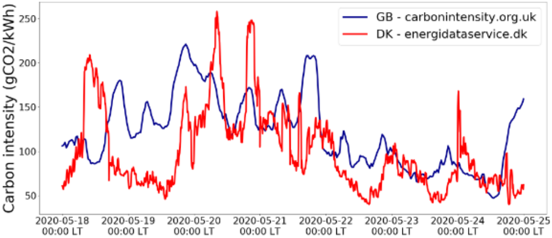
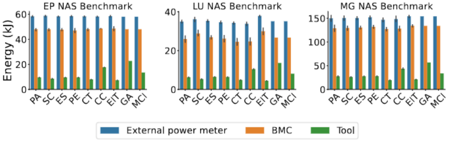


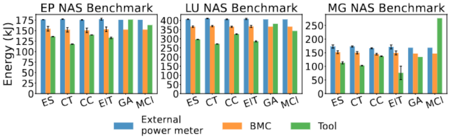
Figure 5. Real-time carbon intensity (gCO_2eq/kWh) for Denmark (DK) and Great Britain (GB) from 2020-05-18 to 2020-05-25 shown in local time. The data is collected using the APIs supported by *carbontracker*. The carbon intensities are volatile to changes in energy demand and depend on the energy sources available.

Anthony, L. F. W., Kanding, B., Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems", 2020

Comparing tool measurements to electricity measurement



(a) [CPU Benchmarks]



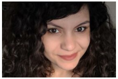
(b) [GPU Benchmarks]

Fig. 2: Total energy consumed by the benchmarks as reported by the power meters. Tools: PowerAPI (PA), Scaphandre (SC), Energy Scope (ES), Perf (PE), Carbon Tracker (CT), Code Carbon (CC), Experiment Impact Tracker (EIT), Green Algorithm (GA), ML CO2 Impact (MCI)

Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram and Anne-Cécile Orgerie. An experimental comparison of software-based power meters: focus on CPU and GPU. CCGrid 2023:1-13

What did we learn about measuring CO2 impact in NLP?

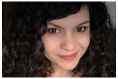
- Server seems more carbon intensive than computing facility
- Tools provide different measures for the same experiments
  - direct measure vs. estimation of computation
  - values of Carbon Intensity, Power Usage Effectiveness (PUE)
  - some tools are not sensitive enough to capture small impact



Measurements conducted as part of N. Bannour's PhD work.

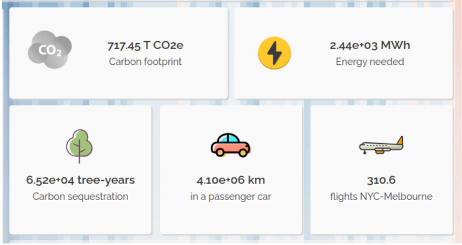
What did we learn about measuring CO_2 impact in NLP?

- Replicability and comparability over time and hardware set-ups
  - check for tool versioning (and ability to select version)
  - check for tool parameters (may vary between set-ups)
- Tools only account for dynamic use of hardware (1 in 4 sources of carbon emission)


Measurements conducted as part of N. Bannour's PhD work.

What is the impact of chatGPT? - Training

- Data is hard to find!
  - It was suggested that OpenAI required 3,617 of NVIDIA's HGX A100 servers to train for [90-100] days on Azure cloud



What is the impact of chatGPT? - Use

- Based on public OpenAI sources, **chatGPT query impact estimated at 4.32 g. CO2**
  - Based on a 2009 Google report, **search query impact estimated at 0.2 g. CO2**
  - The impact of a chatGPT query is **22 times higher** that of a traditional search query

Wong V. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. Piktochart blog, November 2023.

What does this impact mean?

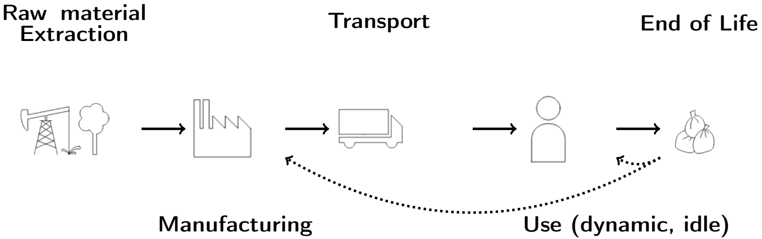
- 16 queries is equivalent to boiling a kettle
- 20 queries per day for a year will get you to Berlin (and back)
  - 7,300 queries is equivalent to the impact of a return train trip from Paris to Berlin.
  - Also equivalent to a flight from Orly to Charles de Gaules

Reports indicate that OpenAI uses **30,000 NVidia A100 GPUs** to keep the generative AI running

How can we account for more impact sources?

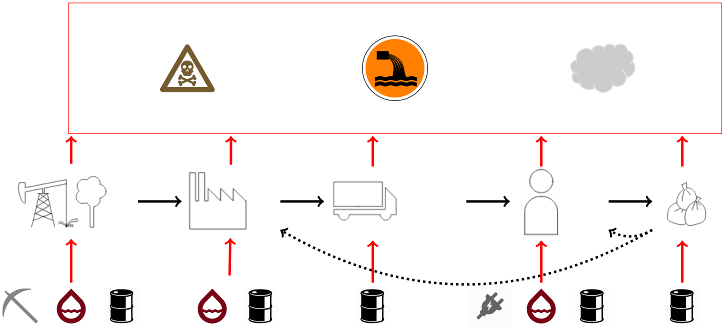
- Work by Clément Morand

Phases of hardware Life Cycle



Impacts differ per phase

Pollution (soil, water, air)



Natural resources

Tools for evaluating the impact of computation

Outil	Phase du cycle de vie considérée						évaluation multi-critères	estimation de la consommation	support des GPU
	Ext.	Fab.	Dis.	Uti.		FdV.			
				Infra.	Dyn.				
Green Algorithms	✗	✗	✗	✓	✓	✗	✗	✓	✓
ML CO_2 Impact	✗	✗	✗	✗	✓	✗	✗	✓	✓
CarbonTracker	✗	✗	✗	✓	✓	✗	✗	✗	✓
CodeCarbon	✗	✗	✗	✓	✓	✗	✗	✗	✓
Boavizta	✓	✓	✗	✗	✗	✗	✓	-	✗

Tools for evaluating the impact of computation

Outil	Phase du cycle de vie considérée						évaluation multi-critères	estimation de la consommation	support des GPU
	Ext.	Fab.	Dis.	Uti. Infra.	Uti. Dyn.	FdV.			
Green Algorithms	✗	✗	✗	✓	✓	✗	✗	✓	✓
ML CO_2 Impact	✗	✗	✗	✗	✓	✗	✗	✓	✓
CarbonTracker	✗	✗	✗	✓	✓	✗	✗	✗	✓
CodeCarbon	✗	✗	✗	✓	✓	✗	✗	✗	✓
Boavizta	✓	✓	✗	✗	✗	✗	✓	-	✗

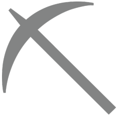
Impacts measures

- Abiotic Depletion Potential (ADP), measured in kgSbeq [van Oers et al., 2020, Bruijn et al., 2002]
- Primary Energy (PE), measured in MJ [Frischknecht et al., 2015]
- Global Warming Potential (GWP) , mesuré en gCO_2e

Proposed tool: MLCA

	ADP	GWP	PE	Toxicité humaine	Consommation d'eau	...
Extraction	✓	✓	✓	✗	✗	✗
Fabrication	✓	✓	✓	✗	✗	✗
Distribution	✗	✗	✗	✗	✗	✗
Utilisation	✓	✓	✓	✗	✗	✗
Fin de Vie	✗	✗	✗	✗	✗	✗

GPU modeling



Attribution of production impacts

Infrastructure energy use

What do these impacts mean?

Modeling the manufacturing impacts of Graphics Cards

Graphics card = GPU + Memory + Base

- GPU modeled by die size
- memory modeled by memory size
- base impacts computed from [Loubet et al., 2023]

$$\begin{aligned}\text{Graphics card}_{impact} = \quad & \text{die}_{size} * die_{impact_{per-cm^2}} \\ & + \text{memory}_{size} * memory_{impact_{perGB}} \\ & + \text{base}_{impact}\end{aligned}$$

where impact $\in$ {ADP, PE, GWP}.

Manufacturing impacts Attribution

- linear attribution
- [?] on the Jean Zay cluster

$$\texttt{embodied}_{impact} = \texttt{manufacturing}_{impact} \frac{\texttt{hours usage}}{\texttt{total available hours}}$$

impact $\in$ {ADP, PE, GWP}

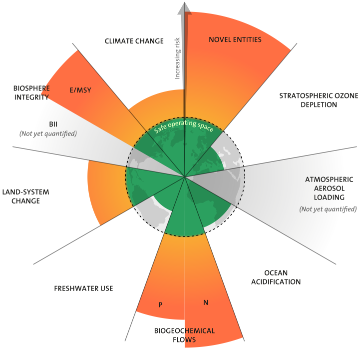
Some perspective on impacts

COP21 · CMP11
PARIS 2015
UN CLIMATE CHANGE CONFERENCE

French National Low Carbon Strategy: 2 tCO_2 e/person/year

https://indicateurs-snbc.developpement-durable.gouv.fr/empreinte-carbone-des-francais-a27.html

Some perspective on impacts

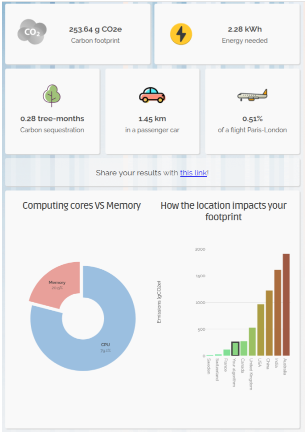


Limites planétaires [Sala et al., 2020]

- PB_{GWP} = 985 kgCO_2 e/personne/an
- PB_{ADP} = 3.17E-02 kgSbeq/personne/an

picture: an analysis by Persson et al 2022 and Steffen et al 2015

Some perspective on impacts



http://calculator.green-algorithms.org/

Evaluation protocol

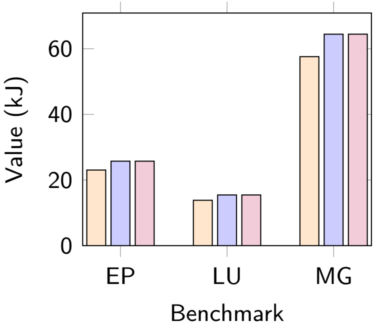
Replication of impact studies of experiments
5 studies

Comparison with production phase assessment studies
Dell Server LCA

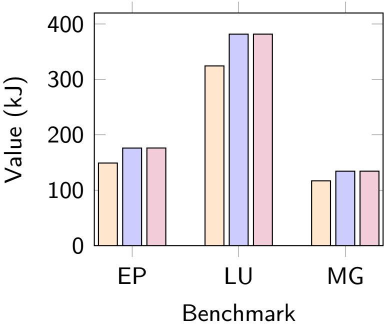
Assessment over the life cycle
Reproducing Bloom study

Replicating results from Jay et al.

Difference between real TDP of the GPU and the TDP used in Green Algorithms



(a) CPU benchmarks



(b) GPU benchmarks

Figure: Value Real (orange ▯), Value Match (blue ▯), Expected (purple ▯)

Sphera for Dell on the R6515, R6525 servers

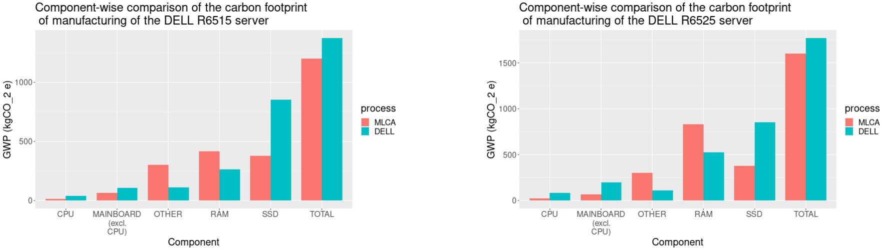
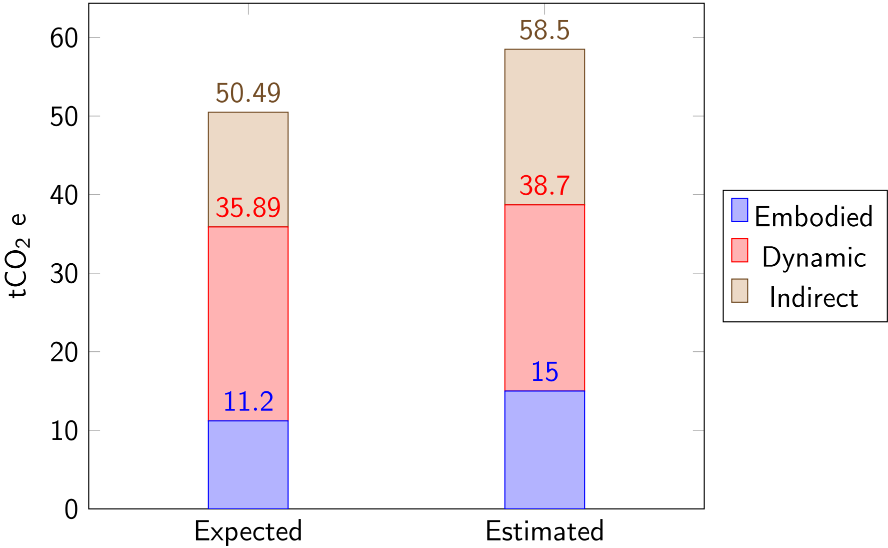


Figure: Component-wise comparison of the GWP of manufacturing for the Dell R6515 (left) and R6525 servers (right)

Evaluation of the whole tool: reproducing results from Luccioni et al.



Implications of the integration of LCA in impact measurement

- Estimation quality is good
  - Attested by replication studies
  - Impact of LCA is significant (half the total impact for BLOOM!)
- Collaboration with Green Algorithms team
- Available for new impact studies

A broad approch to "green computing"

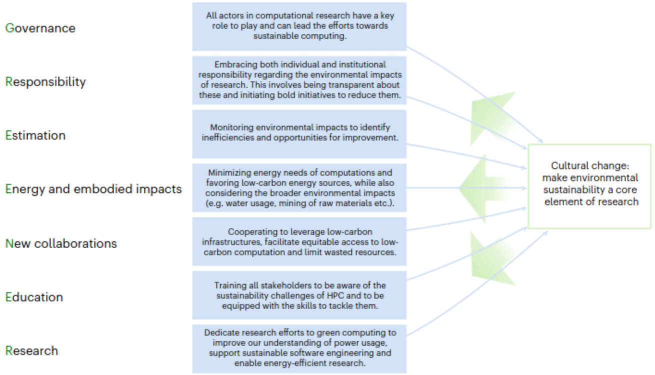


Fig. 1 | GREENER principles for ESCS. The GREENER principles enable cultural change (blue arrows), which in turn facilitates their implementation (green arrows) and triggers a virtuous circle.

Lannelongue, L., Aronson, HE.G., Bateman, A. et al. GREENER principles for environmentally sustainable computational science. Nat Comput Sci 3, 514–521 (2023).

Impact of NLP for health in France

- Attempt at a broad view
  - Identifying impact generating NLP activities
  - Understanding stakeholders knowledge and opinion
- Focus on Clinical Data Warehouses
- Unstructured interviews of 7 participants

Interview scenario

Presentation of the study followed by questions:

1. Please describe your background and your current position.
2. Which AI or NLP tools do you or your collaborators use in your professional practice ?
3. Which infrastructure is needed to support the use of these tools?
4. What are your thoughts on the environnemental impact of these tools?
5. What are your thoughts on the ethical impact of these tools?

Study participant profiles

Training	Hospital staff	NLP researcher	Government employee	CDW management	Location
CS & MD	✓	✓	✗	✗	Paris
CS & MD	✓	✓	✗	✗	Paris
CS	✓	✓	✗	✗	Paris
CS	✓	✓	✗	✓	Paris
CS & MD	✓	✓	✗	✓	Rouen
MBA	✗	✗	✓	✗	Paris
CS & MD	✓	✓	✓	✗	Paris

Interview findings on ICT use & infrastructure

Use

- Developement of CDW : research, care, operations
- Need for new automation tools

Infrastructure

- System duplication
- Need for compute power
- At a standpoint for major infrastructure mutation

Interview findings on ethics and environemental impact

Environemental policy

- Some isolated efforts, but no overall policy
- Carbon footprint of French hospital information systems is 190,000 tCO_2e [?]
- Health department is working on an eco scale

Ethics

- Privacy is of utmost importance
- Risk of impact on patient-clinician rapport
- Who is responsible for the impact of AI recommendations?
- How does AI impact physician training?
- Need for cyber-attack management and mitigation of digital dependency

Overall findings

1. Health in France is increasingly digital (hospital operations, medical practice, public health research)
2. The development of Clinical Data Wharehouses is a major change in the healthcare system and strongly impacts the use of digital health solutions
3. Digital health sustainability is not fully part of the decision process

Acknowledgements

- Colleagues
  - Nesrine Bannour, Sahar Ghannay, Anne-Laure Ligozat, Clément Morand at LISN/STL
  - LISN-DD
  - Loïc Lannelongue at University of Cambridge
- Funding
  - ANR-23-IAS1-0004 InExtenso

Summary:

On the path to greener AI:

involving all stakeholders

measuring impact

informing decisions

References

Bruijn, H., Duin, R., Huijbregts, M. A. J., Guinee, J. B., Gorree, M., Heijungs, R., Huppes, G., Kleijn, R., Koning, A., van Oers, L., Sleeswijk, A. W., Suh, S., and de Haes, H. A. U. (2002).
Handbook on Life Cycle Assessment - Operational Guide to the ISO Standards.
Springer Dordrecht.

Frischknecht, R., Wyss, F., Knöpfel, S. B., Lützkendorf, T., and Balouktsi, M. (2015).
Cumulative energy demand in lca: the energy harvested approach.
International Journal of Life Cycle Assessment 20, page 957–969.

Galibert, O., Quintard, L., Rosset, S., Zweigenbaum, P., Nédellec, C., Aubin, S., Gillard, L., Raysz, J.-P., Pois, D., Tannier, X., Deléger, L., and Laurent, D. (2010).
Named and specific entity detection in varied data: The quæro named entity baseline evaluation.
In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Loubet, P., Vincent, A., Collin, A., Dejous, C., Ghiotto, A., and Jego, C. (2023).
Life cycle assessment of ict in higher education: a comparison between desktop and single-board computers.
The International Journal of Life Cycle Assessment, pages 1–19.

Ma, X. and Hovy, E. (2016).
End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014).
The QUAERO French medical corpus: A resource for medical entity recognition and normalization.
In Proc. BioTextM.

Sala, S., Crenna, E., Secchi, M., and Sanyé-Mengual, E. (2020).
Environmental sustainability of european production and consumption assessed against planetary boundaries