Towards a Green AI

Evolutionary solutions for an ecologically-viable artificial intelligence

Nayat Sánchez-Pi and **Luis Martí**
Inria Chile Research Center, Santiago, Chile.

GECCO '21 Companion, July 10–14, 2021, Lille, France © 2021 Association for Computing Machinery. ACM ISBN 978-1-4503-8351-6/21/07 $15.00 https://doi.org/10.1145/3449726.3461428

Instructors

Nayat Sánchez-Pi is currently the Director and CEO of the Inria Chile Research Center, created in 2012 by Inria, the French National Research Institute for Digital Sciences to facilitate scientific and industrial cooperation between France, Chile, and Latin America. Before that, she was a professor of Artificial Intelligence and Human-Computer Interaction at the Department of Informatics and Computer Science of the Institute of Mathematics and Statistics of the Rio de Janeiro State University. Prof. Sánchez-Pi's research interests have broadened over the years and span topics that range from artificial intelligence, machine learning, and data mining to ambient intelligence, ubiquitous computing, and multi-agent systems. She received a degree in Computer Science in 2000 from the University of Havana and a Ph.D. degree in Computer Science in 2011 from the Universidad Carlos III de Madrid. She has led numerous research projects applying evolutionary computation, machine learning, and other artificial intelligence methods.

Luis Martí is currently the scientific director of Inria Chile, the Chilean Center of Inria, the French National Institute for Computational Sciences. Before that, he was a senior researcher of the TAU team at Inria Saclay since 2015. He was also an Adjunct Professor (tenured) at the Institute of Computing of the Universidade Federal Fluminense. Previous to that, Luis was a CNPq Young Talent of Science Fellow at the Applied Robotics and Intelligence Lab of the Department of Electrical Engineering of the Pontifícia Universidade Católica do Rio de Janeiro, Brazil. Luis did his Ph.D. at the Group of Applied Artificial Intelligence of the Department of Informatics of the Universidad Carlos III de Madrid, Madrid, Spain, and got his Computer Science degree from the University of Havana. He is mainly interested in artificial intelligence, and, in particular, machine learning, neural networks, evolutionary computation, optimization, machine learning, hybrid systems, and all that.

Inria in Chile

Since 2012

Proyecto apoyado por CORFO

Computing (AI) for efficient green *stuff*

OPTIMIZATION
Plant and export layouts and robustness.

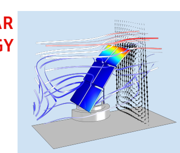
MODELING
Wind and wave modeling for solar, wind and tidal energies.

PREDICTION
Being able to predict the demand and production leads to efficient production.

Green energies and resilience

- ➜ Modeling and prediction of turbulent ocean flows.
  application *Power generation with tides.*
- ➜ High resolution wind prediction
  application *Offshore wind generators.*
  application *Impact of wind in solar generation.*
- ➜ Ecological impact of industries.
- ➜ Understand catastrophes and their impact.

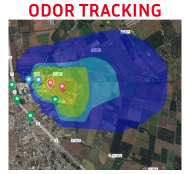
SOLAR ENERGY



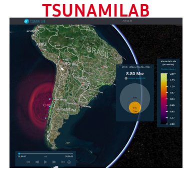
TIDAL ENERGY

OFFSHORE WIND

ODOR TRACKING

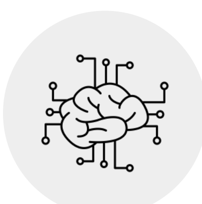


TSUNAMILAB

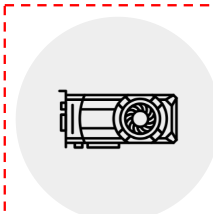


Ok, computing can help us reduce the environmental footprint...

...but, what about the impact of computing itself?



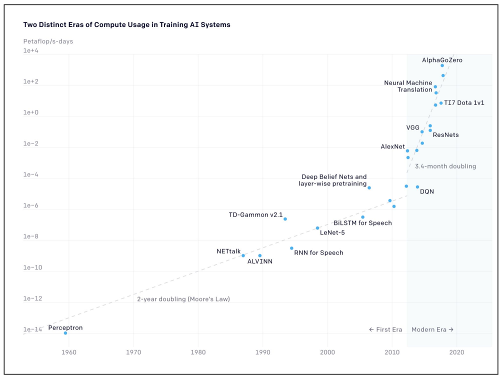
BETTER MODELS
Deep, reinforcement, transfer, active learning, etc.

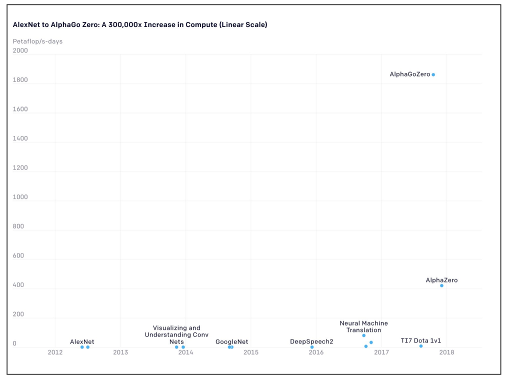


BETTER HARDWARE
GPUs were developed for rendering graphics but are the best hardware for machine learning.

BETTER DATASETS
Important effort to create challenges datasets: vision, natural language, etc.

Progress in ML has come at the expense of exponential needs of computing power





Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2018, May 16). *AI and Compute*. OpenAI. https://openai.com/blog/ai-and-compute/

Inria

Common carbon footprint benchmarks

in lbs of CO2 equivalent

Roundtrip flight b/w NY and SF (1 passenger)	1,984
Human life (avg. 1 year)	11,023
American life (avg. 1 year)	36,156
US car including fuel (avg. 1 lifetime)	126,000
Transformer (213M parameters) w/ neural architecture search	626,155

E. Strubell, A. Ganesh and A. McCallum. **Energy and Policy Considerations for Deep Learning in NLP.** Annual Meeting of the Association for Computational Linguistics (ACL short). Florence, Italy. July 2019.
https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Inria

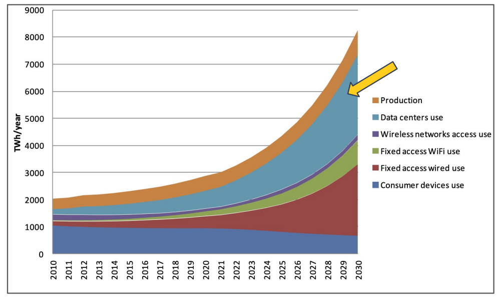
AI FOR HUMANITY

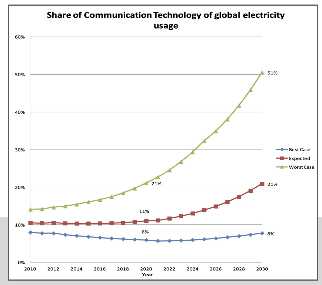
L'INTELLIGENCE ARTIFICIELLE
AU SERVICE DE L'HUMAIN.

Energy consumption by the digital sector could increase tenfold by 2030, accounting for between

20 and 50%

of global electricity use.





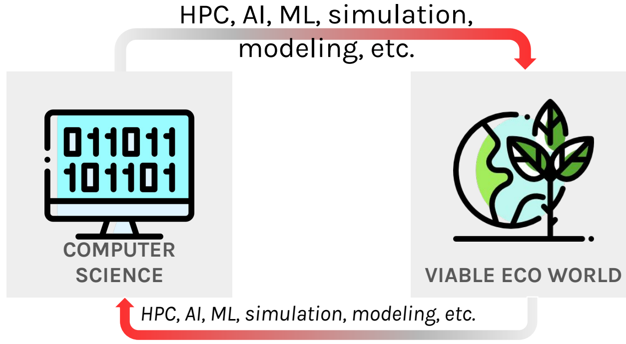
1. Cédric Villani, Marc Schoenauer, Yann Bonnet, Charly Berthet, Anne-Charlotte Cornut, François Levin, Bertrand Rondepierre (2018) **AI for Humanity.** https://www.aiforhumanity.fr
2. Andrae, A., & Edler, T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges*, 6(1), 117–157. https://doi.org/10.3390/challe6010117

The solution

Nature has provided a lot of inspiration for the AI/ML area:
- ↳ neural networks,
- ↳ learning theories,
- ↳ evolutionary computing, etc.

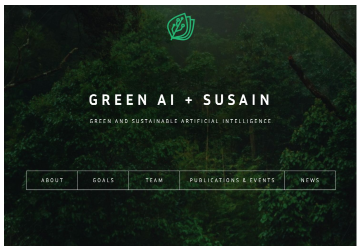
It is about time to return the favor!

HPC, AI, ML, simulation, modeling, etc.

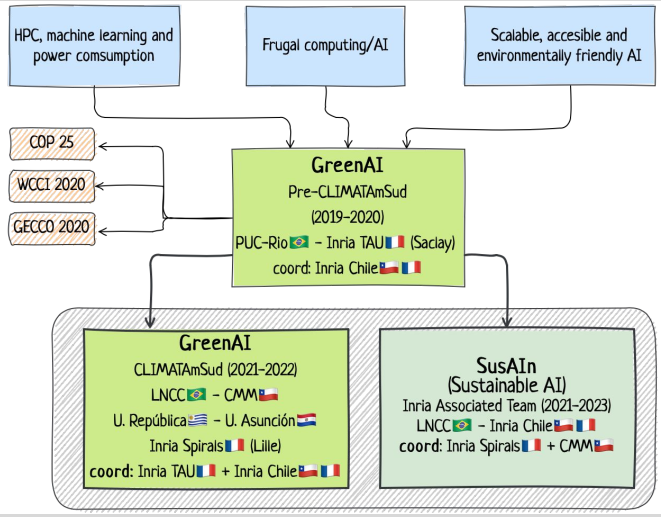


HPC, AI, ML, simulation, modeling, etc.

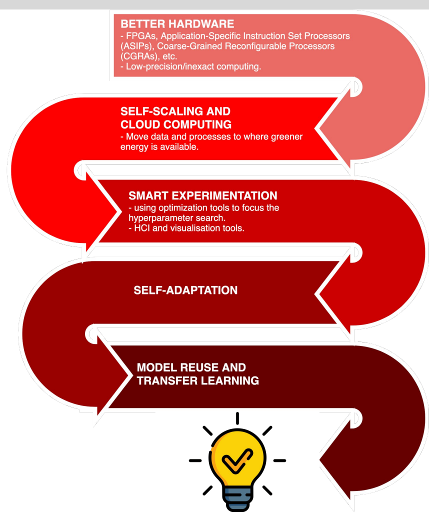
Energy-conscious computing



Web: https://greenai.inria.cl



The solution: a multifaceted approach



Two distinctive scenarios that require different solutions

Training

- **High demand of resources:**
  - Big masses of data, and particularly,
  - Power-hungry GPUs.
- **Aggravated by the need of hyperparameter tuning.**
  - e.g. impossible to determine optimal network topology.
  - Current "expensive" AutoML.
- **Limits access to technology to a few powerful companies.**
  - They are starting to monetize it directly: i.e. OpenAI's GPT3.
- **The good news: we do it once, use many times.**

Inference

- Current state of the art models (i.e. transformers) do not scale *as-is*.
- To take these results into edge/mobile computing and IoT scenarios:
  - i.e. phones, security cameras, etc.

Each scenario requires a different approach. There are actually many more scenarios.

Better Hardware

It is unlikely that we get right of GPUs (or TPUs) at training time.

There are **hardware alternatives** at **use time**:

➔ Field Programmable Gate Arrays (FPGAs), Application-Specific Instruction-set Processors (ASIPs), etc.

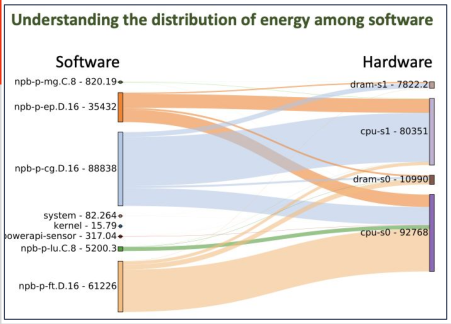
We should also keep exploring the use of ***low-precision computing:***

Reducing the quality (and therefore length) of the **floating-point representation of numbers.**

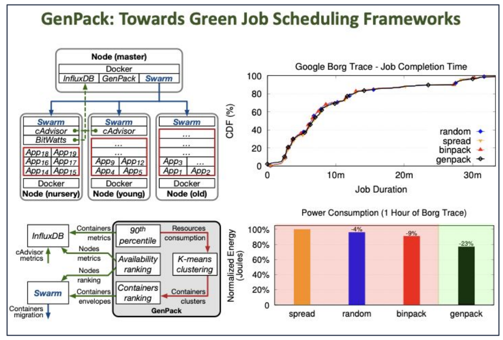
*« These results show that these programmers lacked knowledge of how to **accurately measure software energy consumption**. »*

PowerAPI

Enabling power monitoring of software systems



Everything starts with how we code, how we schedule, and organize resources.



Inria SPIRALS Université de Lille institut universitaire de France

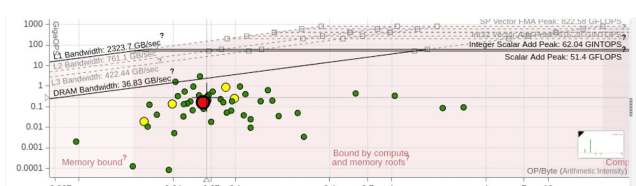
Romain Rouvoy (2020). **AI for Green Computing vs. Greening AI.** ACM GECCO 2020 Workshop "Green AI: Evolutionary and machine learning solutions in environment, renewable and ecologically-aware scenarios"

HPC efficiency optimization and auto-tuning

Roofline model for evaluating AI applications

Observed characteristics DT:

↳ Balanced in memory and processing requirement
↳ The bottleneck is DRAM memory
↳ They have low arithmetic intensity and GFLOPs



Vitor de Sá, V., Klôh, V., Schulze, B., and Ferro, M. (2020). **Análise de desempenho e de requisitos computacionais utilizando o modelo Roofline: Um estudo para aplicações de inteligência artificial e do NAS-HPC**. In WSCAD 2020 –WIC

Cloud computing for AI?

AI and ML pipelines are very computationally intensive but not continuously run:

Training: when we fit a model to the data available, very costly but run infrequently -or maybe once.

- ↳ Punctual use of expensive computing equipment.
- ↳ GPUs are both expensive to acquire and expensive to operate (high energy demands, cooling, etc.).

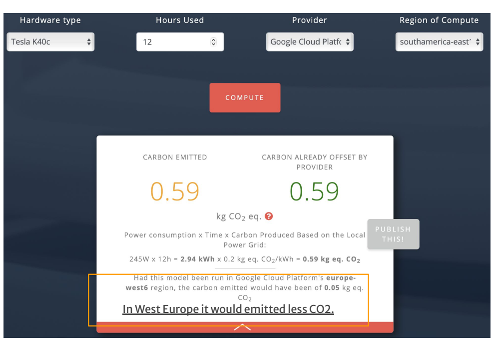
Prediction: using the model to make decisions, low computing requirements and used very frequently.

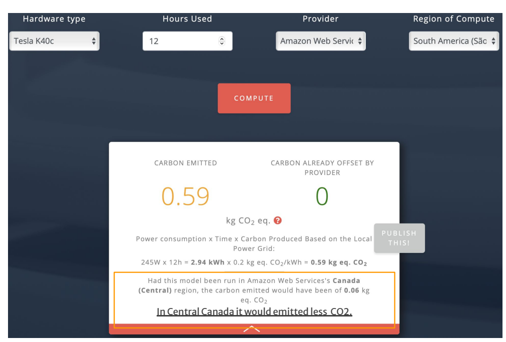
- ↳ High-availability low cost computing.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). **Carbon Emissions and Large Neural Network Training**. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2104.10350

Not all cloud locations are the same

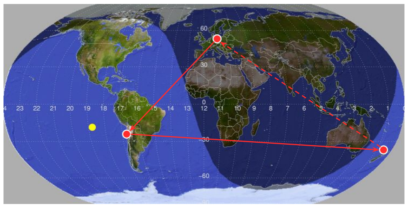
For example, different cloud locations and their CO_2 impact (indirectly, cost)





https://mlco2.github.io/

Self-scaling and cloud computing



Self-scaling computing facilities make available a pool of shared resources.

Optimally schedule computing time.

Cloud computing allows **to pick the location** where programs will be run.

Code is mobile!

We can, for example, "track the sun" and ensure that the AI/ML processes use renewable sources.

"the choice of DNN, datacenter, and processor can reduce the carbon footprint up to ~100x - 1000x."[1]

1. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). **Carbon Emissions and Large Neural Network Training.** In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2104.10350

Smarter experimentation: Better AutoML

Finding the **right configuration** of the **hyperparameters** probably where more energy is consumed.

➜ This is an optimization problem => NP-Hard problem

➜ Neural architecture search, AutoML, AutoDL, etc.

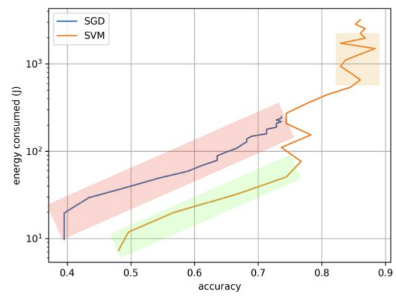
➜ **Go beyond 'regular' hyperparameter search**

➜ However, better approaches like **evolutionary computing** is here to help!
...but they need populations of individuals, hence more energy.

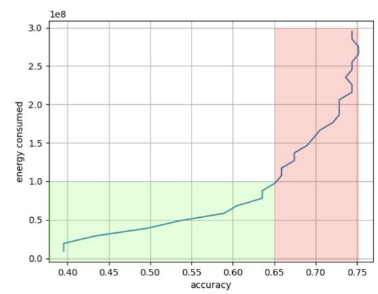
This is a multi-objective optimization problem!

High performance comes at high cost

Comparing SGD *Vs.* SVM efficiencies



Cost of training a SVM classifier



Romain Rouvoy (2020). **AI for Green Computing vs. Greening AI**. ACM GECCO 2020 Workshop "Green AI: Evolutionary and machine learning solutions in environment, renewable and ecologically-aware scenarios"

Measuring AI's footprint: *proxies*

- ➔ Carbon emissions generated.
- ➔ Electric energy consumed.
- ➔ Floating point operations.
- ➔ Elapsed time used for execution.
- ➔ Number of parameters (weights) and hyperparameters.

MLCommons Power Working Group:
Create power measurement techniques for various MLPerf benchmarks that enable reporting and comparing energy consumption, performance and power of benchmarks run on the submission systems.

https://mlcommons.org/en/groups/best-practices-power/

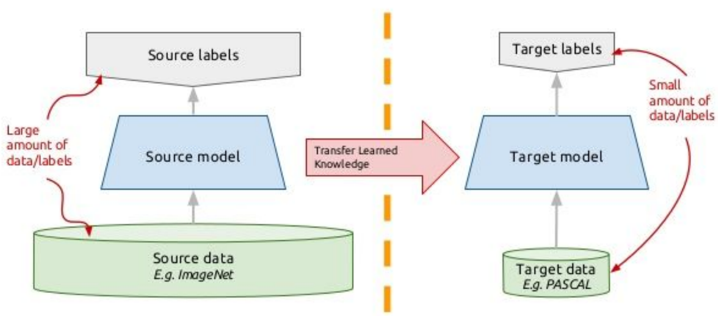
Self adaptation

To look for **methods** that **adapt** their complexity **automatically** to the **complexity** of the **problem** being solved.

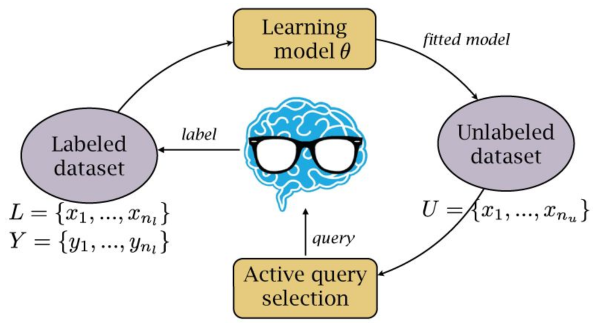
Neural networks based on **adaptive resonance theory (ART)** and **growing neural gas (GNG)** have rules to adapt themselves to the complexity of the problem.

This self-adaptation is best profited when using cloud-based infrastructure.

Model reuse and transfer learning



Active learning and sample efficiency



https://www.epfl.ch/labs/cvlab/research/machine-learning/page-123199-en-html/

The combination of mitigation factors

…(c/s)hould be posed as an optimization problem

Inria Challenge

OcéanIA

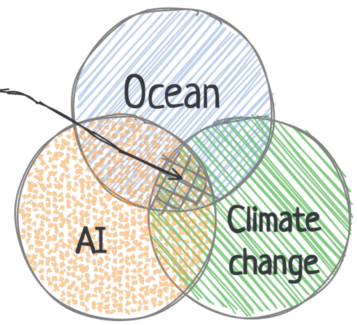
Artificial Intelligence and **Modelling** for **Understanding Climate Change**

OcéanIA

Nature has provided a lot of inspiration for the AI/ML area:
- ↳ neural networks,
- ↳ learning theories,
- ↳ evolutionary computing,

etc.

It is about time that we return the favor!

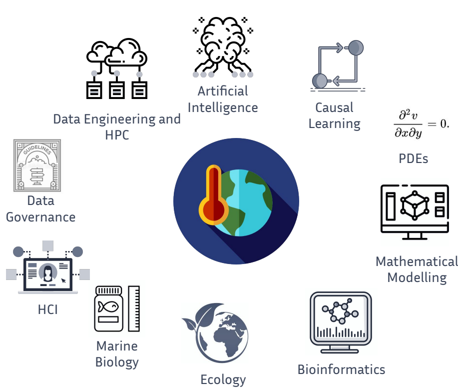


How the ocean helps to mitigate climate change?	How to protect this mitigation effect?

https://oceania.inria.cl/

Inria

The OcéanIA team



Inria Challenge

LEADED BY Inria Chile

In France
- Inria: ANGE, TAU, BIOCORE
- CNRS GO-SEE Federation
- LS2N COMBI, U.Nantes

In Chile
- Universidad de Chile
- CMM Center for Mathematical Modeling
- Pontificia Universidad Católica de Chile

Industries

Others
- TARA OCEANS
- Copernicus

Additional Support
- Cooperación Regional Francesa para América del Sur

NEURAL INFORMATION PROCESSING SYSTEMS

LXAI

Inria

EMISTRAL

Environmental **M**onitoring and
Inspection **S**ailboat via
Transfer, **R**einforcement and
Autonomous **L**earning

Source: Beckler, J. S., et al.. (2019). **Coastal Harmful Algae Bloom Monitoring via a Sustainable, Sail-Powered Mobile Platform.** Frontiers in Marine Science, 6 (October), 1–14.

STIC AmSud

UFRN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Inria

Thank you! **Obrigados!**
Merci ! - *¡Gracias!*

Questions?

Find more in greenai.inria.cl

Inria