RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

cea cnrs Inria

FRANCE 2030 PROGRAMME DE RECHERCHE *INTELLIGENCE ARTIFICIELLE*

SHARP

Sharp Theoretical and Algorithmic Principles for frugal ML

Rémi Gribonval, LIP, Inria & ENS de Lyon, coordinator

anr agence nationale de la recherche

"Frugal learning" : an oxymoron ?

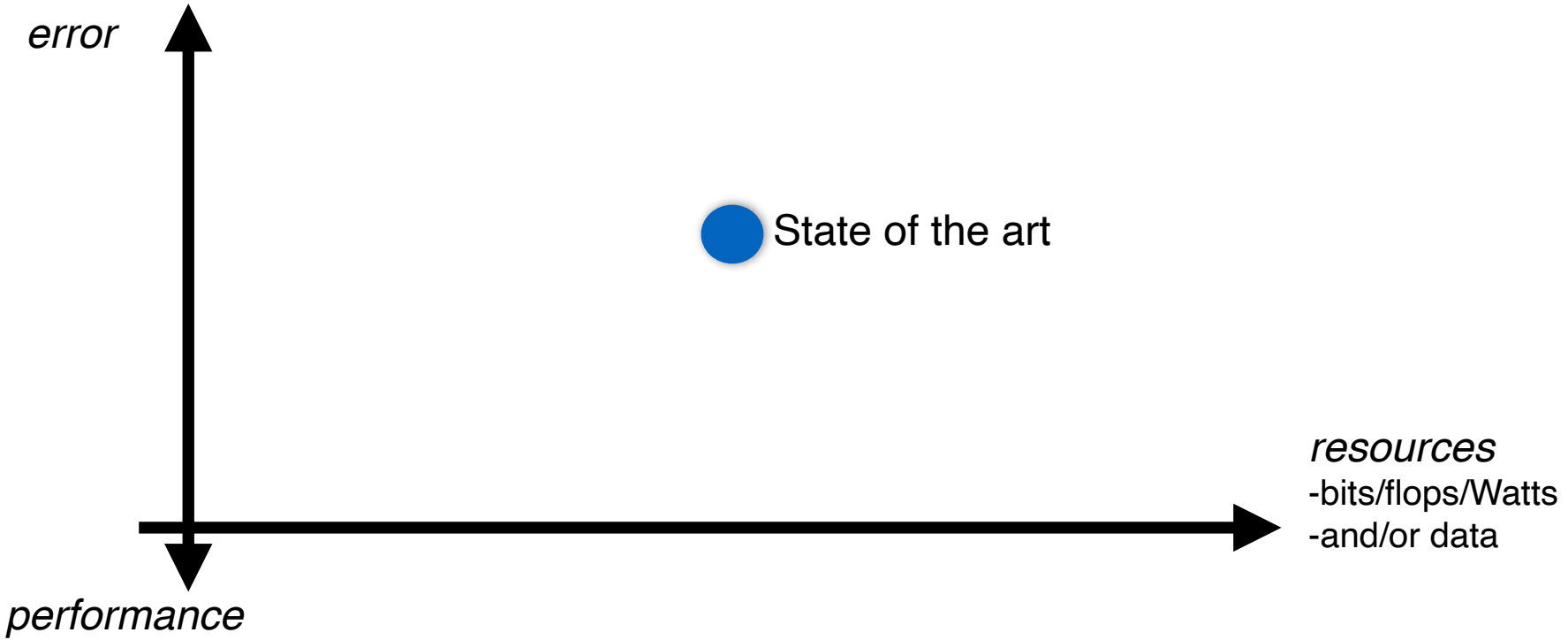
ML = evermore ?

- Economic competition
- Maximizing performance
- Race towards giant models
- Unrestrained consumption

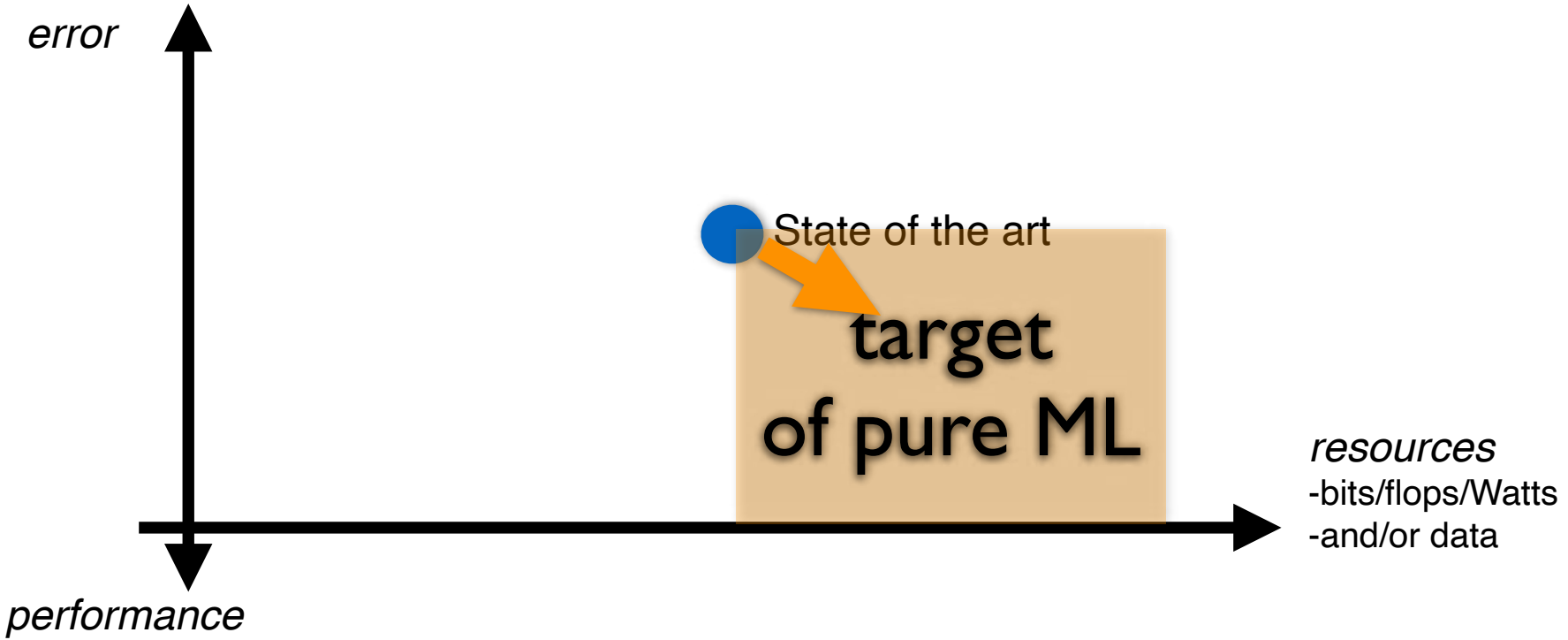
Frugality = sobriety ?

where to set the cursor ?

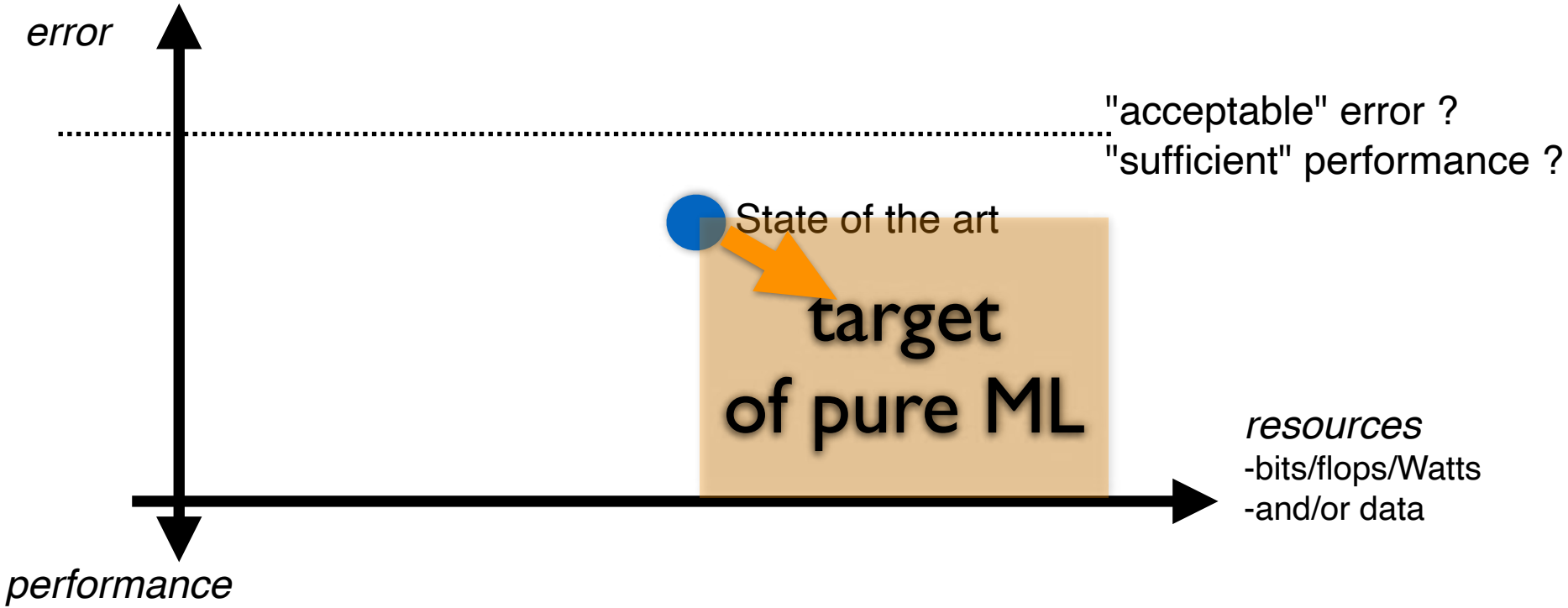
Frugality: what are we talking about ?



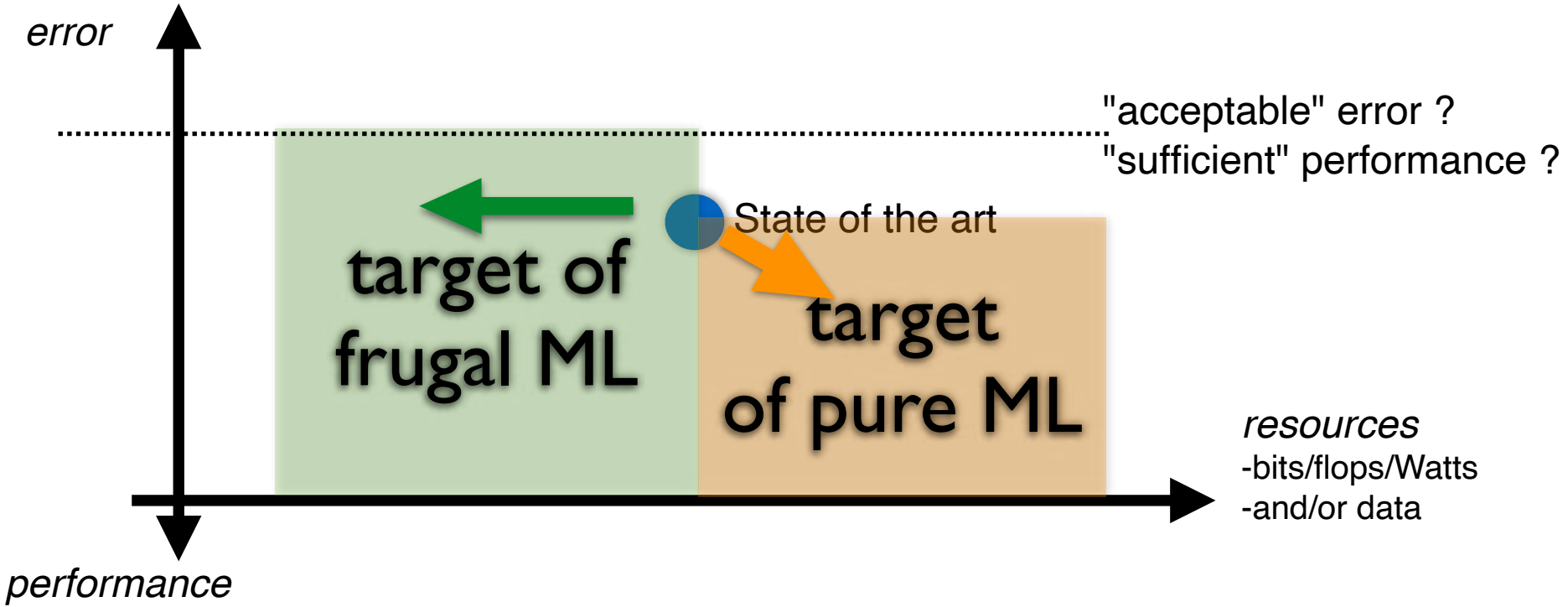
Frugality: what are we talking about ?



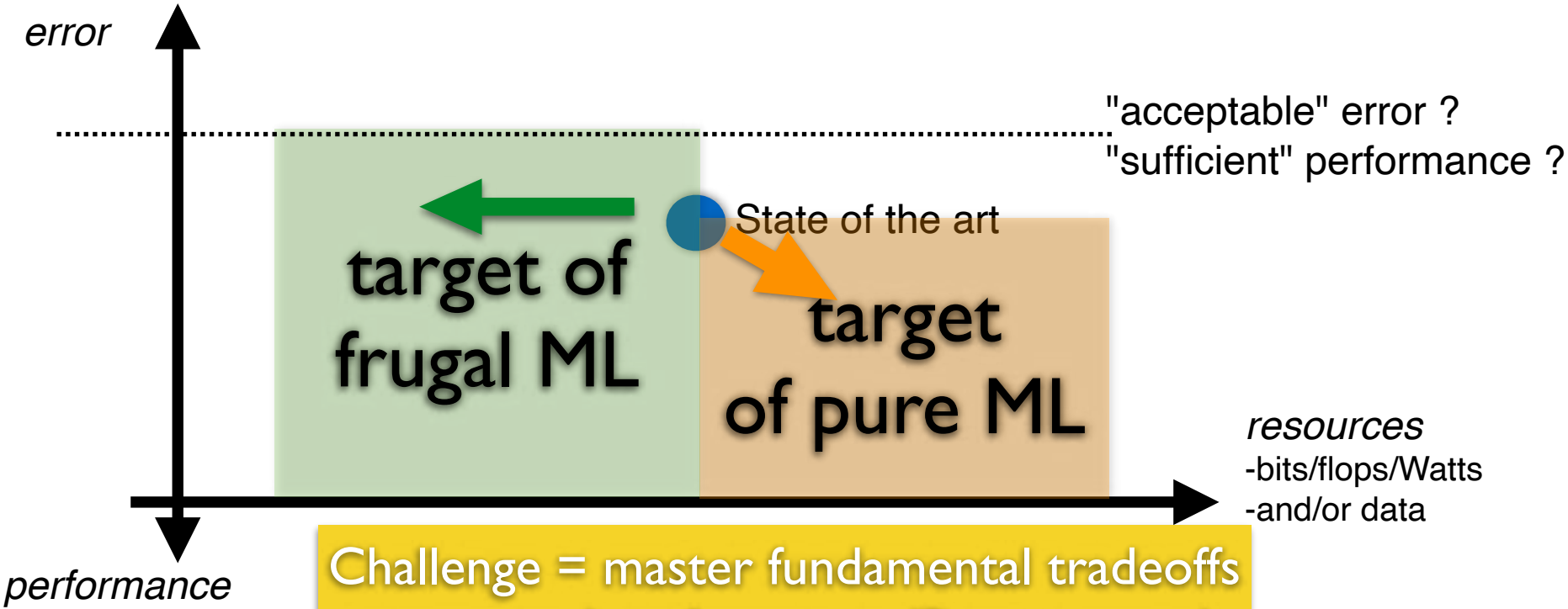
Frugality: what are we talking about ?



Frugality: what are we talking about ?

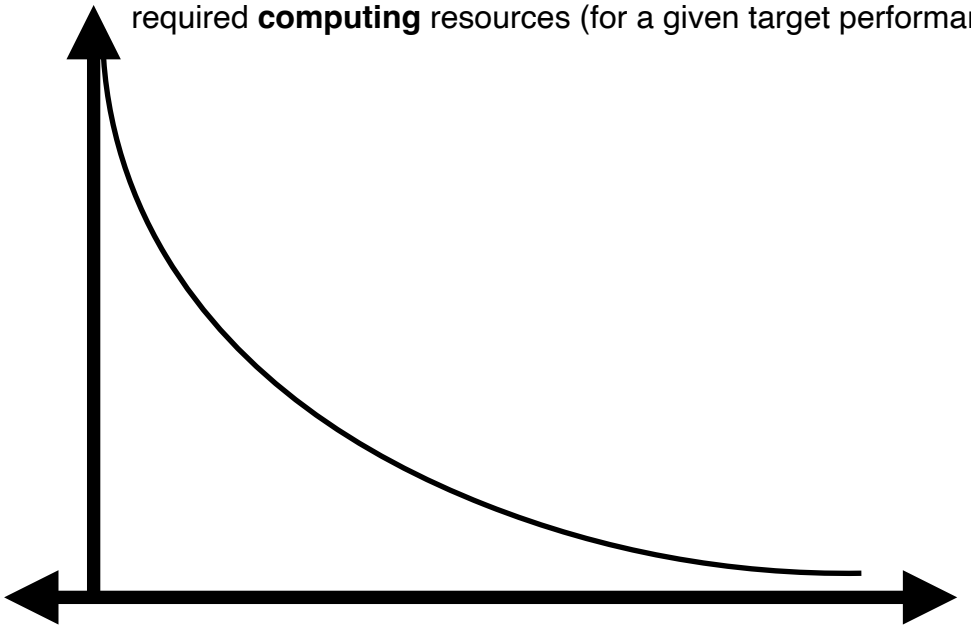


Frugality: what are we talking about ?

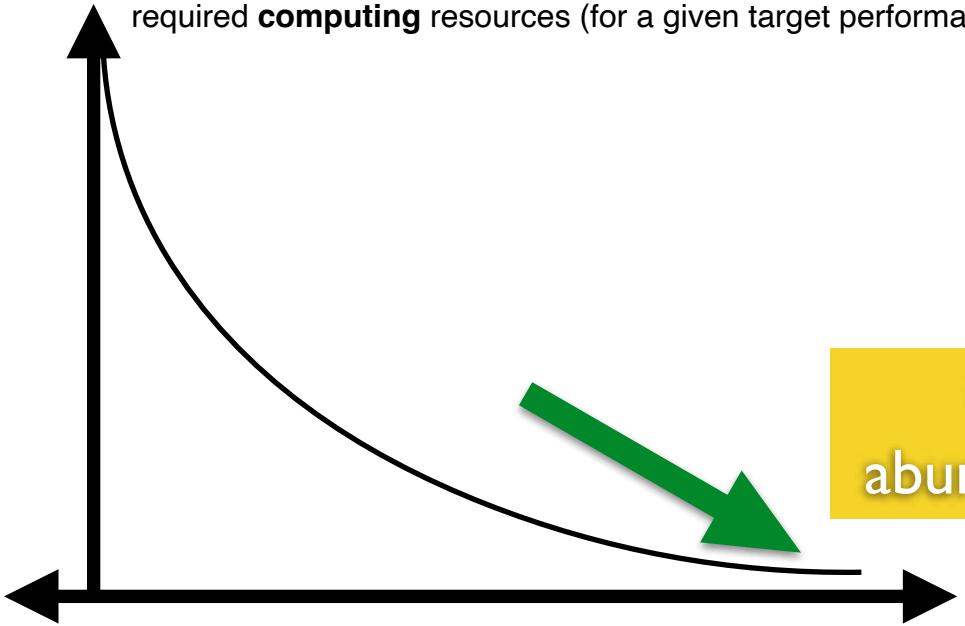


Challenge = master fundamental tradeoffs resources / performance (Pareto curve)

Data / computation tradeoffs ?

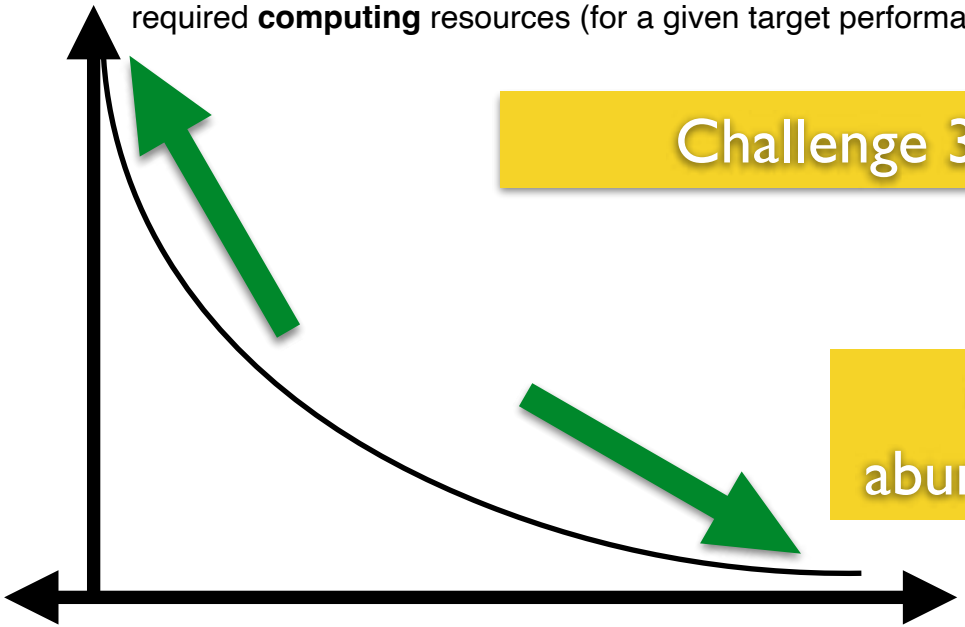


Data / computation tradeoffs ?



Challenge 2 = exploit data abundancy to *reduce* computations

Data / computation tradeoffs ?



Challenge 3 = learn *despite data scarcity*

Challenge 2 = exploit data abundancy to *reduce* computations

3 Main Axes

Axis 1

Frugal
Architectures

Axis 2

Frugal
Principles

Axis 3

"Small" and "Raw"
Data

Showcase : "leap forward" in frugal vision & language models

Organization

WP6 – *Dissemination* – lead LAMSADE

WP1 – ISIR
Ensuring frugal training via dedicated taylored architectures

(Axis 1)

3PhD, 3 years postdoc

WP2 – IRISA
Efficient and Powerful architectures for structured data

(Axes 1 and 2)

3PhD, 8 years postdoc

WP3 – LIGM
Incorporating and leveraging prior knowledge

(Axis 2)

6PhD, 1.5 years postdoc

WP4 – GENESIS
Lifelong and reusable foundation models

(Axes 2 and 3)

5PhD, 7 years postdoc

WP5 – List
Frugal and Unbiased Generative Models for Multimedia Understanding

(Axis 3)

3PhD, 2 years postdoc

WP0 – *Management* – lead LIP

Consortium

Integrated pipeline from theoretical foundations to flagship AI applications in vision & NLP

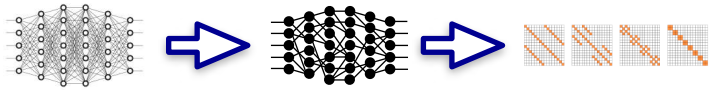
- **LIP (ENS de Lyon, Univ. Claude Bernard Lyon 1, CNRS, Inria), Rémi Gribonval, coordinator**
  Sparsity, sketching, statistical learning theory, information theory, mathematics of deep learning
- **LAMSADE** (Paris-Dauphine and PSL University, CNRS), **Alexandre Allauzen**, partner
  Deep learning for natural language and speech processing
- **LIGM** (École des Ponts ParisTech, Univ Gustave Eiffel, CNRS), **Loïc Landrieu**, partner
  Representation learning for computer vision
- **GENESIS (**Inria & University College London), **Benjamin Guedj**, partner
  Statistical learning theory, PAC-Bayes, mathematics of deep learning, representation learning
- **IRISA** (CNRS, Univ Rennes, Inria, INSA), **Nicolas Keriven,** partner
  Sketched learning, graph neural networks, statistical learning on graphs, optimal transport
- **List** (Université Paris Saclay, CEA), **Hervé Le Borgne**, partner
  Visual and textual content analysis, zero-shot learning, control of generative models
- **ISIR** (Sorbonne University, CNRS), **Nicolas Thome**, partner
  Statistical signal processing applied to distributed learning and deep learning

* uniquement tutelles contractualisantes pour chaque site

Axis 1– Frugal architectures

Endowed with intrinsic tractability guarantees, both for training and inference.

- mathematical foundations of **sparse (& structured) networks**
  - to reduce bits / flops / watts
- optimal **distributed** sparse learning
  - beyond the bottlenecks of back-propagation
- architectures with **PAC-Bayes guarantees**
  - trustworthy models with non-vacuous statistical bounds
- **dimension reduction** via randomized *sketching*
  - of gradients, vectors, datasets ...
- sound **quantization** principles
  - guided by numerical linear algebra & information theory
  - binarized architectures



Axis 2 – Frugal learning principles

To drastically reduce the cost of learning by exploiting prior knowledge.

- limit the waste of current methods
  - avoid end-to-end learning
  - avoid data-augmentation
- learn from the past by incorporating knowledge
  - symmetries & invariances
  - physical models, known or parameterized
  - graphs & (PAC)-Bayesians priors
- showcases : "leap forward" in
  - frugal vision models
  - frugal language models

Axe 3 – "Small" and "raw" data

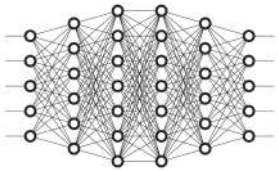
Leveraging the information from as limited and unrefined data as possible.

- minimalistic foundation models
  - via architectures & principles of Axes 1+2
  - *dissection* of large foundation models
  - *small hidden dimension* of latent spaces, implicit bias, conservation laws, ...
- handling scarce and/or raw data
  - uncover the *structure* & the *relative importance* of data through graphs
  - *relax* textbook *i.i.d. assumptions* on training dataset
  - frugal *lifelong learning* from tiny and diverse chunks of data
  - explain & exploit *compressibility* of state of the art models

Spotlight: sparsity as a frugality enabler

Natural idea: fewer parameters = fewer bits / flops / watts ...

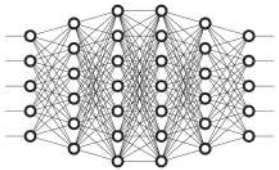
Dense



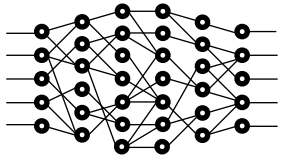
Spotlight: sparsity as a frugality enabler

Natural idea: fewer parameters = fewer bits / flops / watts ...

Dense



Sparse



Devil in the details

-not hardware friendly

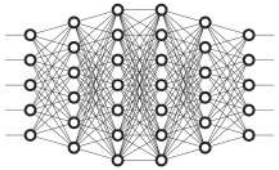
-can make optim ill-posed !!!

Where to put zeroes ???

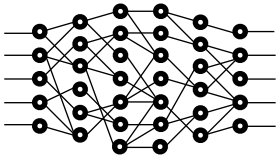
Spotlight: sparsity as a frugality enabler

Natural idea: fewer parameters = fewer bits / flops / watts ...

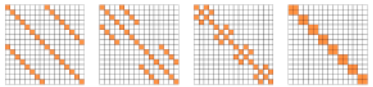
Dense



Sparse



Structured ?



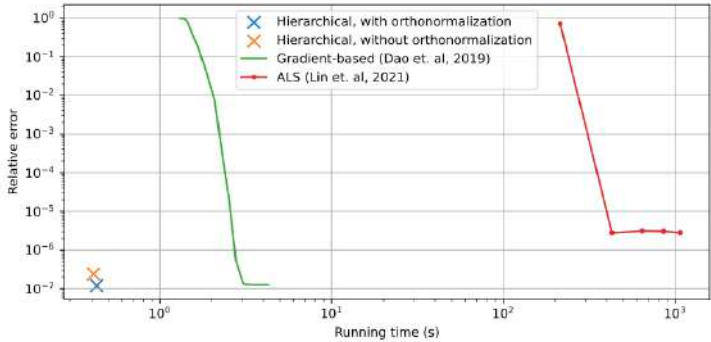
Devil in the details

-not hardware friendly

-can make optim ill-posed !!!

Where to put zeroes ???

Beyond backprop with "butterflies"



Summary

Axis 1

Frugal
Architectures

Axis 2

Frugal
Principles

Axis 3

"Small" and "Raw"
Data

Showcase : "leap forward" in frugal vision & language models

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria