

Hardware and software challenges for massive-scale AI

Laurent DAUDET

LightOn, CTO and Co-Founder

“Foundation models”

Lighton

Just write
what you
want !

Write a professional email to Tom Sanders, a salesman for a laser company.

Generate



“Foundation models”

Just write
what you
want !

Automatic text
generation
using Muse API

Write a professional email to Tom Sanders, a salesman for a laser company.

Dear Tom,

I am a potential customer for your company, and I'm interested in learning more about your laser products. I have a particular need for a high-power laser that I need to cut through a thick piece of steel. I'm curious to know if you have any lasers that can cut through metal. If you have any products that can cut through steel, I'd love to talk to you about purchasing them.

I look forward to hearing from you.

Best,

Ash

Generate



“Foundation models”

Lighton

Just write
what you
want !

An **Instagram ad** for Atmosphere resort, a luxury hotel on Apo island in the Philippines with world-class diving.

Generate



“Foundation models”

Lighton

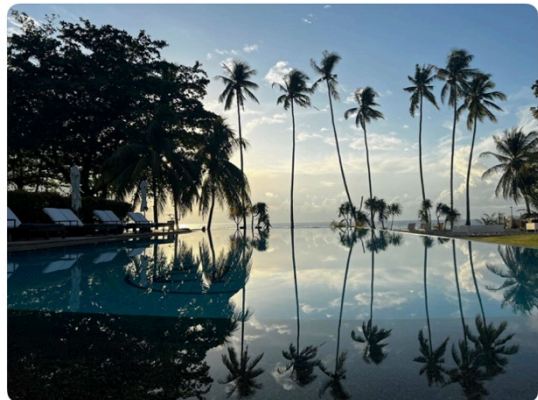
Just write
what you
want !

**Automatic text
generation**
Using Muse API

An **Instagram ad** for Atmosphere resort, a luxury hotel on Apo island in the Philippines with world-class diving.

Atmosphere has just opened its doors and is now ready to offer an unparalleled experience in luxurious, private villas. While the resort’s restaurant is world-class, the real star of the show is the ocean itself. With five stunning dive sites around the island, there’s no better way to enjoy a vacation than with scuba diving. Check out some of our most recent Instagram posts below and get ready to be immersed in paradise!

Generate



“Foundation models”



Real Estate in Paris

User-provided
prompt



Partager

Générer

Dans le vieux Paris, les prix restent au beau fixe. «L'offre de biens à vendre a doublé par rapport à la même période l'an passé», indique Alexis Mathieu, de l'agence Laforêt. «En cette rentrée, nous avons à nouveau beaucoup de demandes d'estimation et de projets d'achat sérieux», ajoute Martial Michaux, de l'agence Émile Garcin Paris Le Marais. Dans ce secteur, le marché immobilier est sans cesse tendu, car l'offre de biens reste éternellement limitée. «Des ventes qui se bouclaient, il y a encore un an, sans négociation dans la journée sont moins fréquentes. Il faut compter une semaine», précise Alexis Mathieu.

“Foundation models”



Real Estate in Paris

User-provided prompt

Automatic text generation now stays in context in French using Muse API



Partager

Générer

Dans le vieux Paris, les prix restent au beau fixe. «L'offre de biens à vendre a doublé par rapport à la même période l'an passé», indique Alexis Mathieu, de l'agence Laforêt. «En cette rentrée, nous avons à nouveau beaucoup de demandes d'estimation et de projets d'achat sérieux», ajoute Martial Michaux, de l'agence Émile Garcin Paris Le Marais. Dans ce secteur, le marché immobilier est sans cesse tendu, car l'offre de biens reste éternellement limitée. «Des ventes qui se bouclaient, il y a encore un an, sans négociation dans la journée sont moins fréquentes. Il faut compter une semaine», précise Alexis Mathieu. Les banques s'attendent à une remontée sensible des taux, à partir de novembre, tandis que les taux d'intérêt des prêts relais devraient rester à un niveau encore très bas jusqu'en 2020. Le stock de logements sociaux est un des rares actifs mobilisables pour la revente. Les institutionnels, à l'instar des promoteurs, demandent à leurs clients des financements de long terme, afin de maintenir leurs marges. Les «prix de vente ne sont pas encore négociables, car les acquéreurs sont toujours prêts à partir à moindre coût», estime Eric Legendre, associé et directeur associé d'AX A IM Real Estate Paris Ile-de-France. Toutefois, de plus en plus de dossiers de financement sont déposés avant la fin de l'année par des investisseurs privés. D'après Patrick Huguet, président de

INTERNATIONAL · PAPE FRANÇOIS

Partage

A Chypre, le pape fait l'éloge de la diversité

Tout en évoquant la « terrible lacération » de l'île coupée en deux depuis l'invasion turque de 1974, François a plaidé pour une Europe qui accueille d'« autres cultures ».

Par Cécile Chambraud (Nicosie, envoyée spéciale)

Publié aujourd'hui à 10h23 · Lecture 3 min.

Article réservé aux abonnés



Le pape François avec l'archevêque de l'Eglise orthodoxe de Chypre, Chrysostome II, le 3 décembre 2021 à Nicosie. ANDREAS SOLARO / AFP

Devant la petite cathédrale maronite Notre-Dame-des-Grâces, à Nicosie, quelques soldats de la Force de maintien de la paix des Nations unies à Chypre veillent. Le pape François vient d'atterrir à Larnaka et il a prévu d'y faire la première halte de sa visite de

Text summarization

Paste a web link here

https://www.lefigaro.fr/international/article/2021/12/03/a-chypre-le-pape-fait-l-eloge-de-la-diversite_6104574_3210.html

Résumé : Devant la cathédrale de Nicosie, le pape François a salué l'hospitalité de l'île et la diversité religieuse de ses habitants.

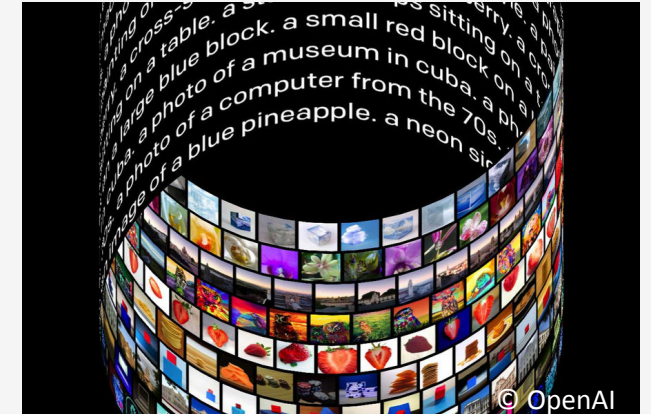
New “transformer-based” billion-scale AI for

- Office work through text-based AI
- Code programming
- Drug discovery

GPT-3 (2020)

Codex (2021)

AlphaFold (2021)



New “transformer-based” billion-scale AI for

- Office work through text-based AI
- Code programming
- Drug discovery

GPT-3 (2020)
Codex (2021)
AlphaFold (2021)



With a **strong effect on science, technology and business** in record time



DeepMind's AlphaCode AI writes code at a competitive level

Devin Coldewey @techcrunch / 7:04 PM GMT+1 • February 2, 2022

Comment



Forbes

AlphaFold Is The Most Important Achievement In AI—Ever

This new AI has never been

so **Hard to Build**

- **DATA** “civilization-scale” high-quality data
- **SOFTWARE** arguably the hottest topic in AI/ML
- **PUSHES THE LIMITS OF HARDWARE**
 - drives new **chips** architectures,
 - will require **new technologies** tomorrow

This new AI has never been

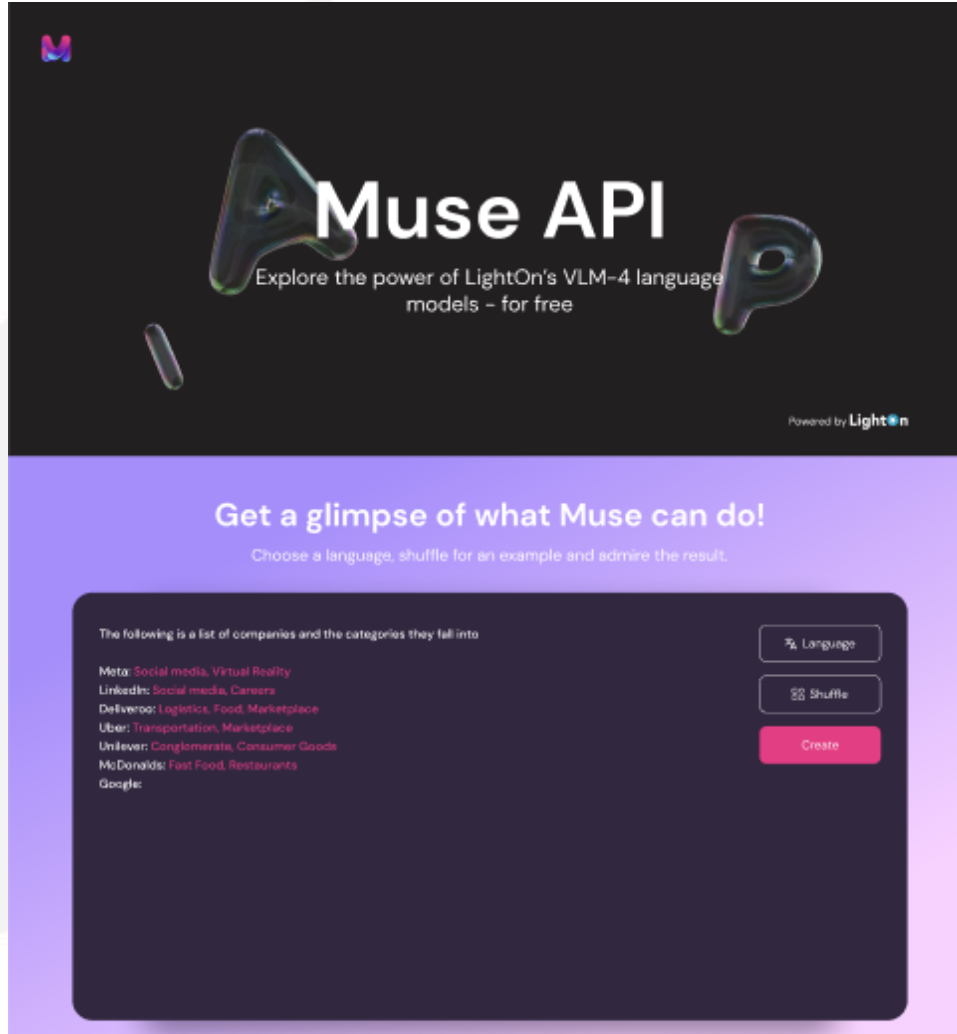
so **Hard to Build**

- **DATA** “civilization-scale” high-quality data
- **SOFTWARE** arguably the hottest topic in ML
- **PUSHES THE LIMITS OF HARDWARE**
 - drives new **chips** architectures,
 - will require **new technologies** tomorrow

and yet so **Easy to Use**

- **API-BASED** using the models doesn't require any AI / Machine Learning expertise
- **FOUNDATION MODELS** a single generic model performs multiple tasks including some it has not been explicitly trained for
- **NATURAL LANGUAGE** interact with your machine like with a colleague
- **NATIVE SPEAKER** a language carries much more than a translation away from English

Muse API on VLM-4 language models: just released !



Large Language Models
natively trained in
Fr, It, Es, De, En

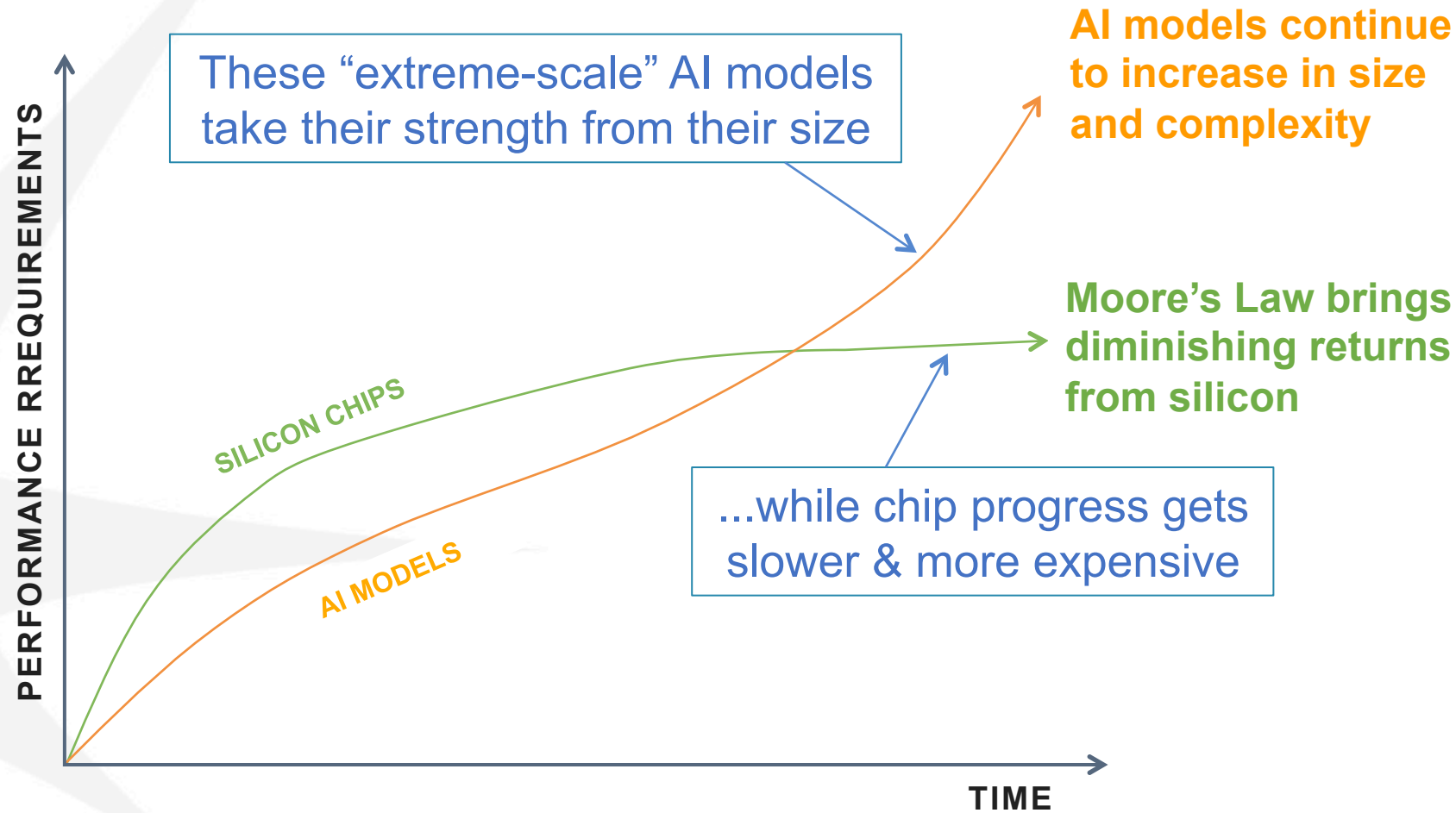
Try it now at
muse.lighton.ai

Entering the era of supercomputer-sized AI

- Training *a single* GPT-3 model :
 - 3 Million GPU-hours (on NVIDIA V100s)
 - 550 T CO₂ equiv.
 - Estimated price 5-10 M \$ for training only
- And it's only the beginning



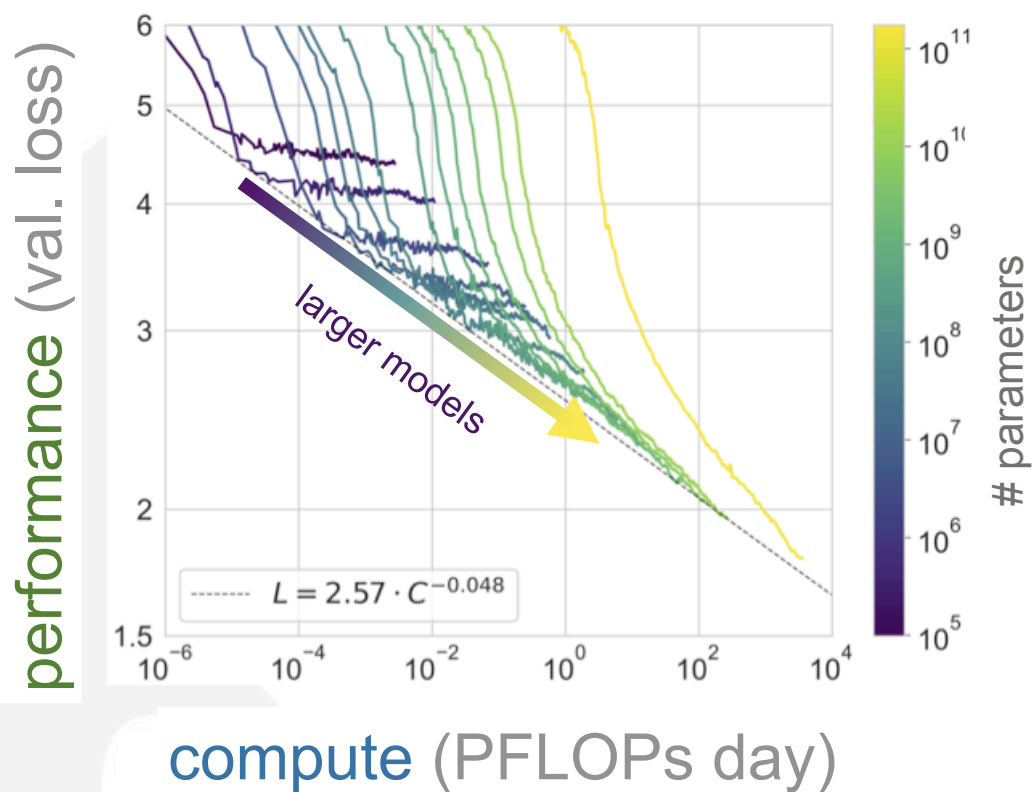
AI requires *much* faster processing



The « scaling hypothesis »

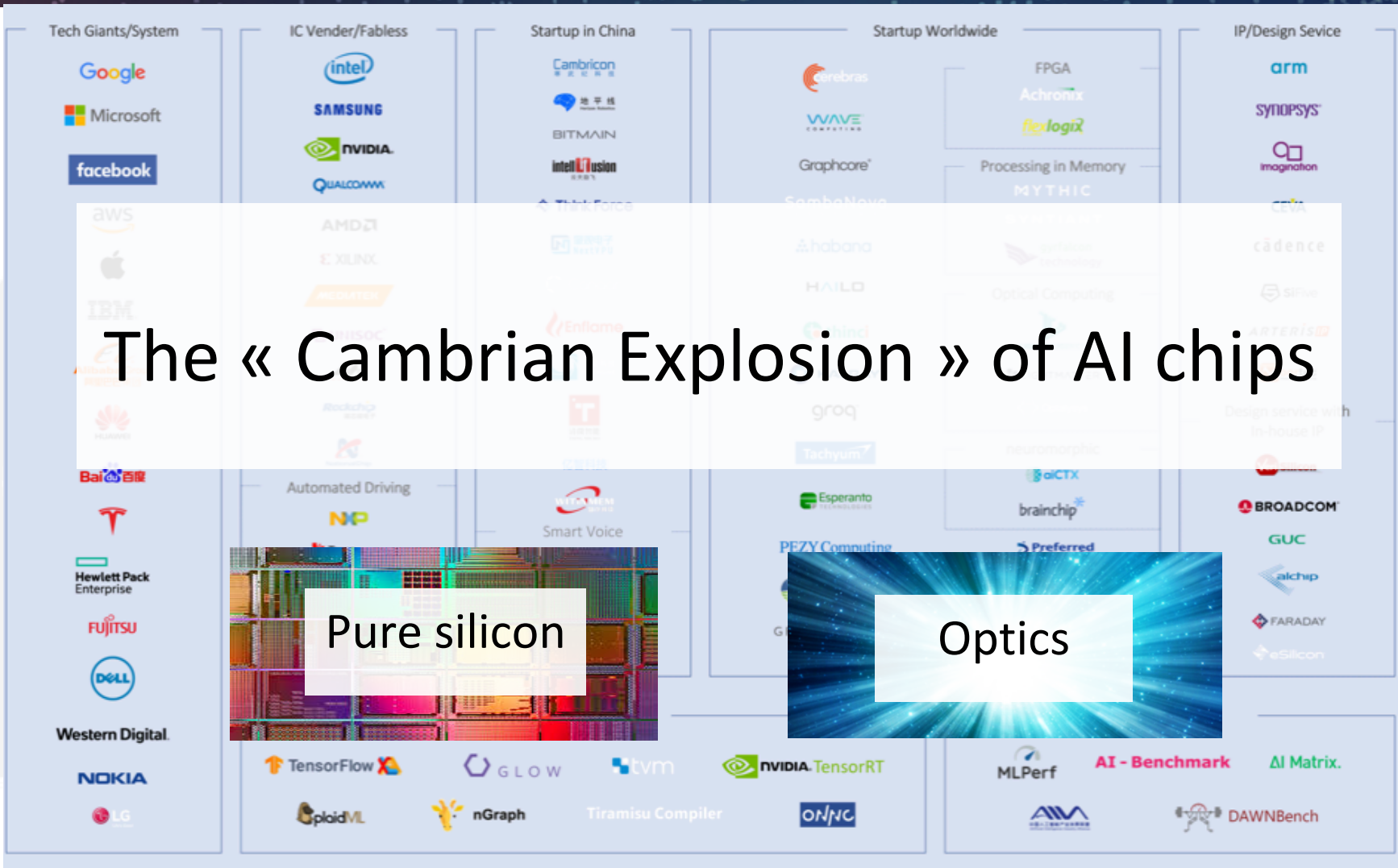
Scaling laws in Language Models [Kaplan et al., 2020]

For well-designed models, **increased model size** is all you need !



Larger models

- score **higher**
- generalize **better**
- train **faster**



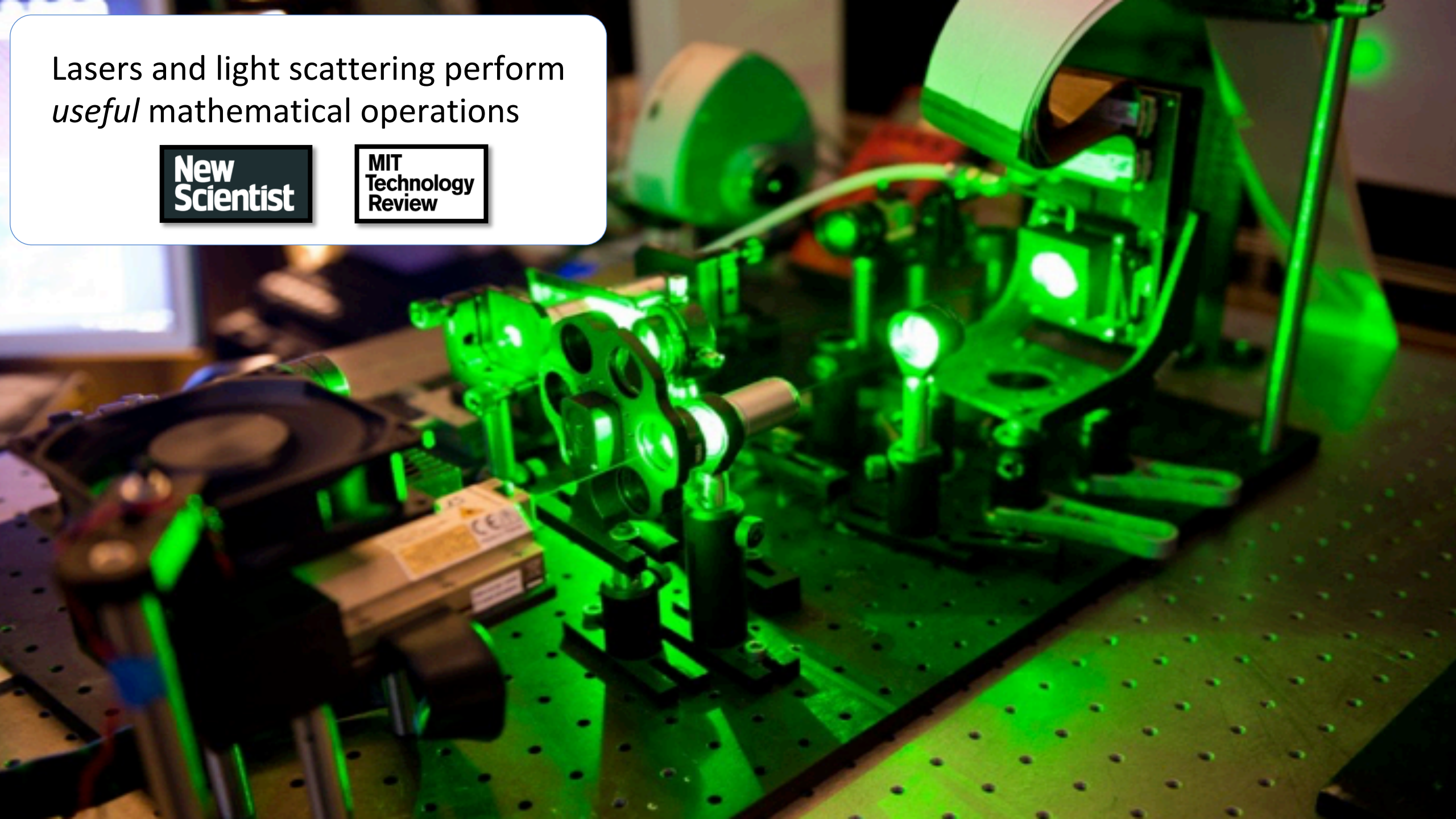
The « Cambrian Explosion » of AI chips



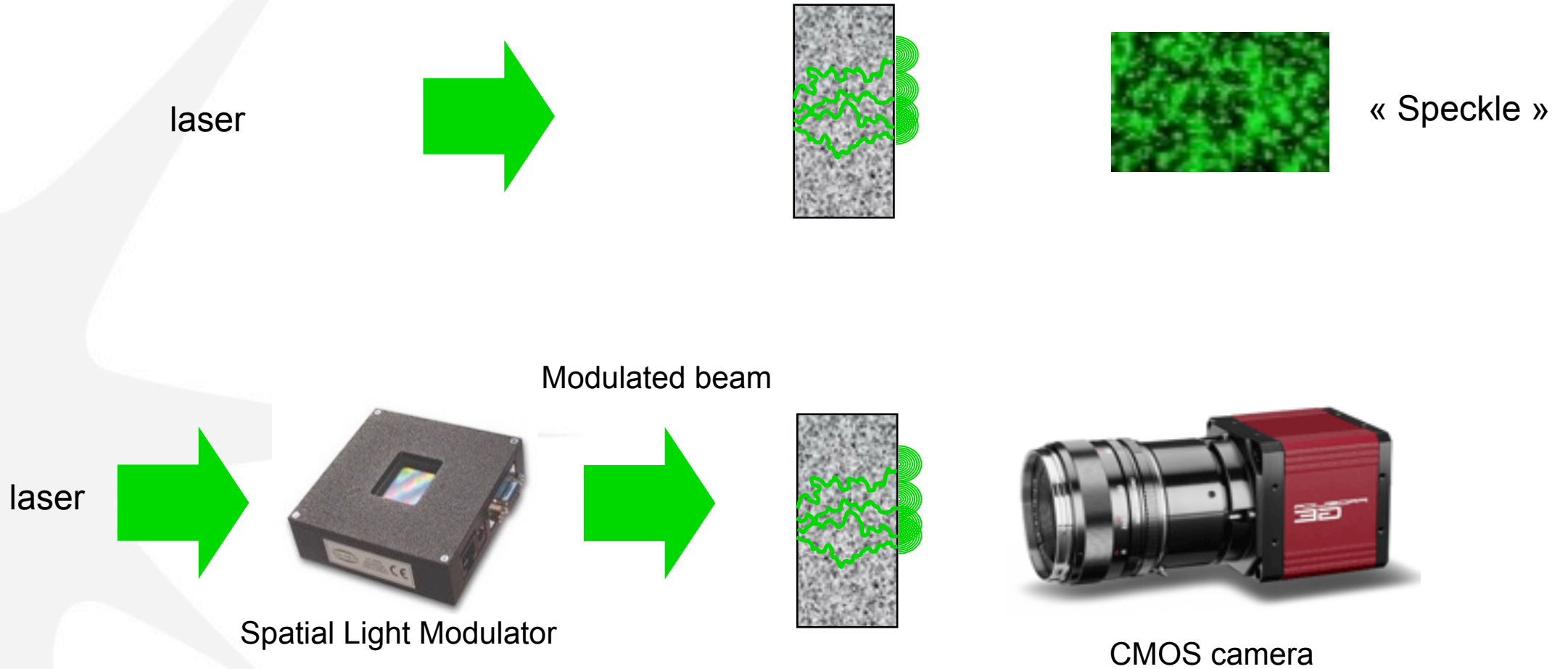
Lasers and light scattering perform *useful* mathematical operations

**New
Scientist**

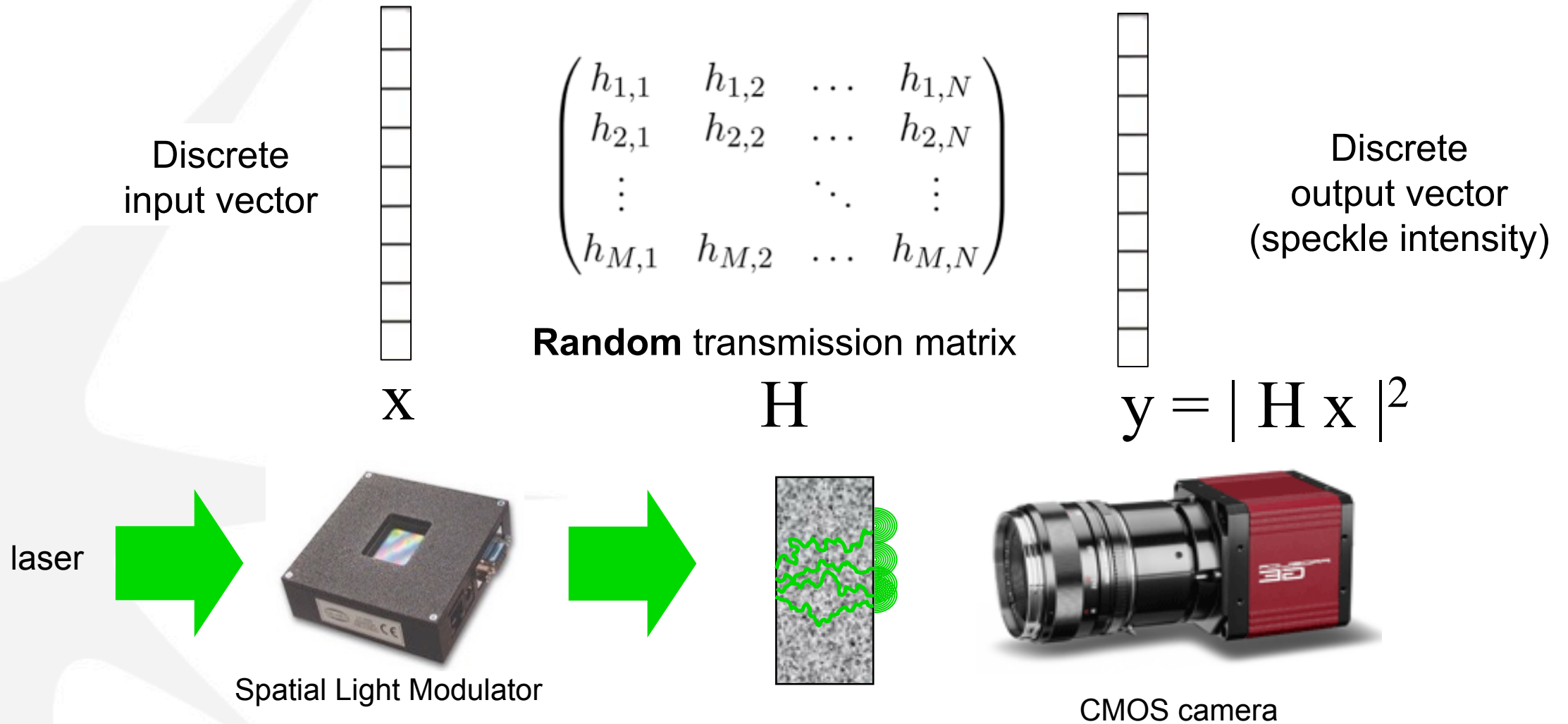
**MIT
Technology
Review**



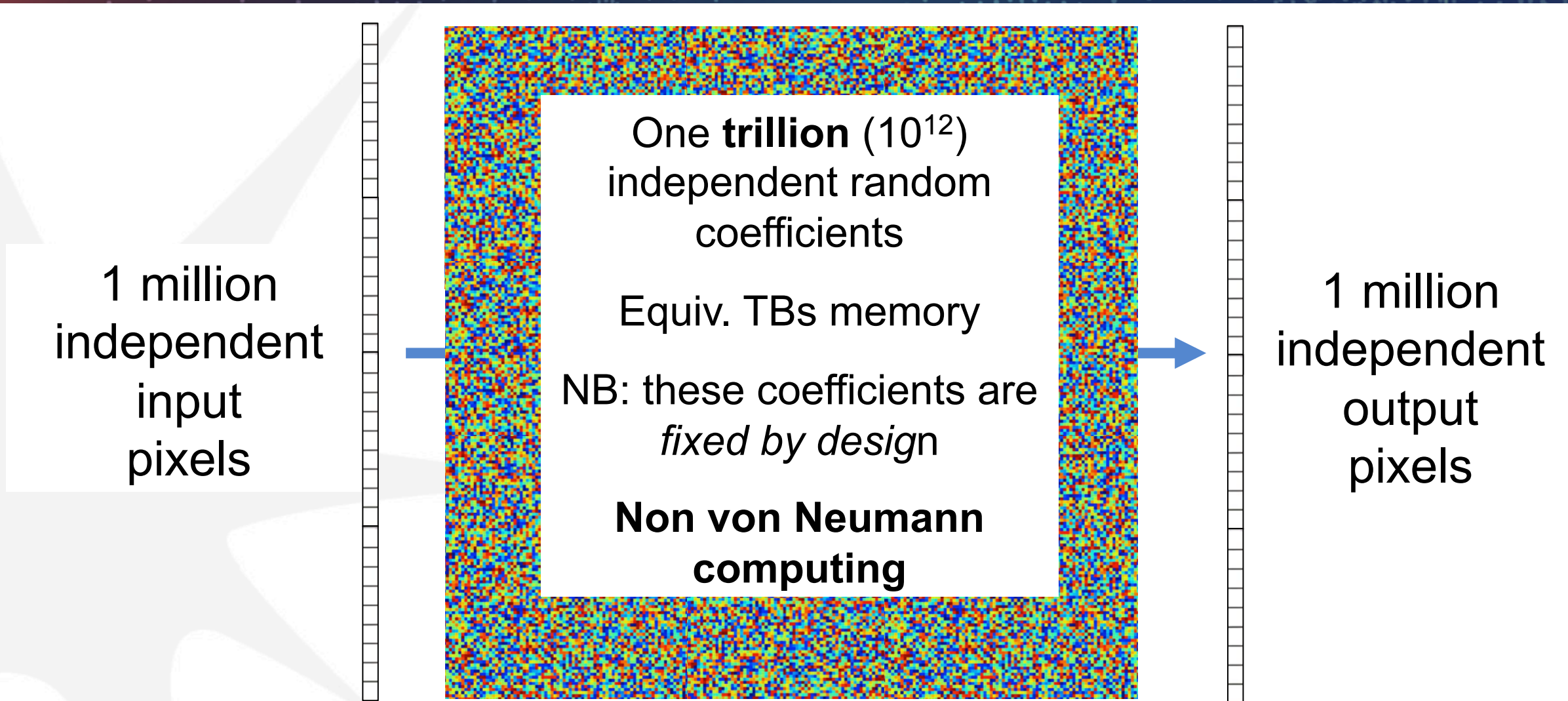
Scattering: a coherent process



Scattering: a coherent process



Matrix-vector multiplication through light scattering



LightOn Appliance: the world's first photonic AI co-processor
publicly available, since March 7th, 2021

2200 TOPS

In a single photonic core

With only **30 W TDP**

200 times better in #OPS/W than
NVIDIA top-range GPU boards



Used by AI experts worldwide through our “LightOn Cloud for research” program

Hybrid computing in AI pipelines




Ivan Dokmanic
Associate Prof.



UNIVERSITY OF ILLINOIS
Urbana-Champaign



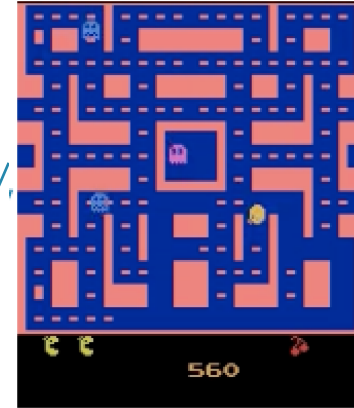
NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2019



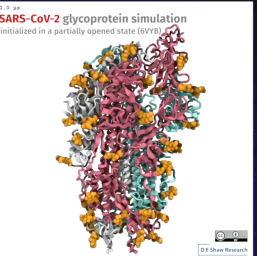
David Rousseau @dhpmrou · 4 avr.

Our talk on analysing #hep data with random matrices in @LightOnIO Optical Processor Unit accepted at @ic hep2020 conference ! (Remote in July or Prague beg 2021) #hepml @Laurent_Daudet @IgorCarron

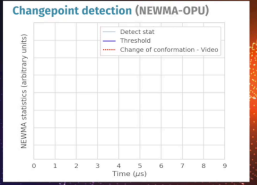
6 retweets, 15 likes



Analyzing SARS-CoV-2 conformations with a LightOn OPU


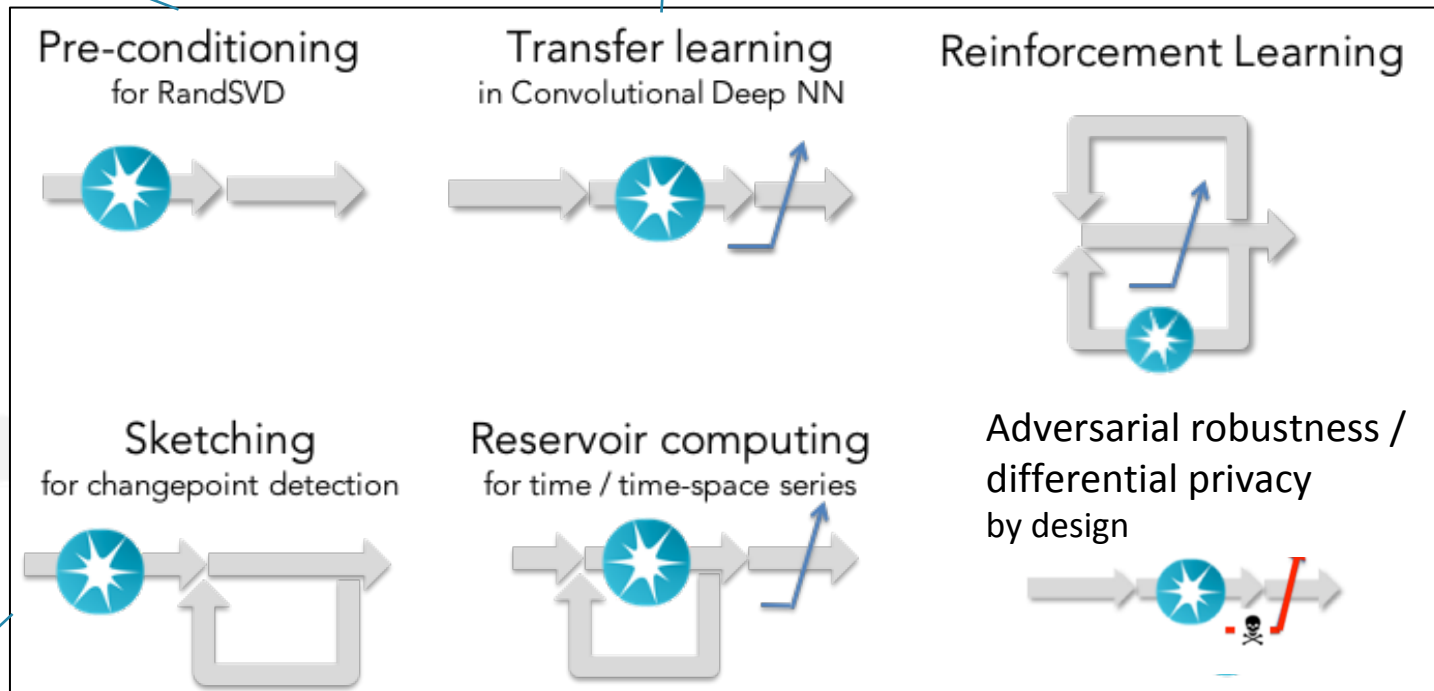


SARS-CoV-2 glycoprotein simulation
initialized in a partially opened state (SVR)




Changepoint detection (NEWMA-OPU)

computed on a LightOn Aurora OPU





Accelerating SARS-COV2 Molecular Dynamics Studies with Optical Random Features



Amélie Chatelain
LightOn ML R&D engineer



Collaboration with Criteo



Collaboration with FAIR

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2020

LightOn Appliance for AI acceleration

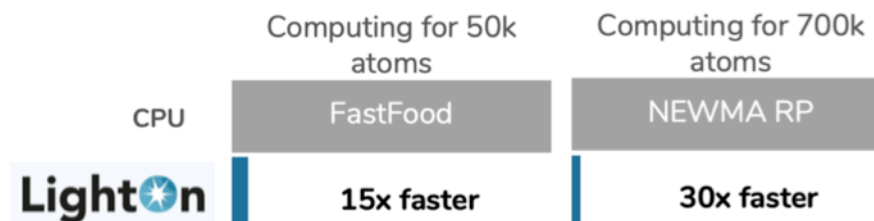
Approximating Kernel Ridge Regression for classification tasks



Performance gains on Kernel ridge approximation for classification tasks. Dataset qm7 (quantum chemistry), high energy physics, and others. The OPU is compared to an NVIDIA P100 GPU (250 W). GPU RAM limit was hit at 32GB. Results acquired extrapolating to 1M features. OPU: Aurora 1.5 (30 W).

Real-time AI analysis of large-scale HPC results

Change detection in Molecular Dynamics



15x faster than FastFood on CPU at 50k atoms. For 700k + atoms, NEWMA RP on OPU is expected to be 30x faster than NEWMA FF on CPU. Library: LightOnML, Dataset: Molecular Dynamics simulations (HPC, Anton), OPU: Aurora 1.5

Robustness to adversarial attacks

arXiv.org > cs > arXiv:2108.04217

Search...

Help | Ac

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 6 Jul 2021]

ROPUST: Improving Robustness through Fine-tuning with Photonic Processors and Synthetic Gradients

Alessandro Cappelli, Julien Launay, Laurent Meunier, Ruben Ohana, Iacopo Poli

Collaboration
with FAIR

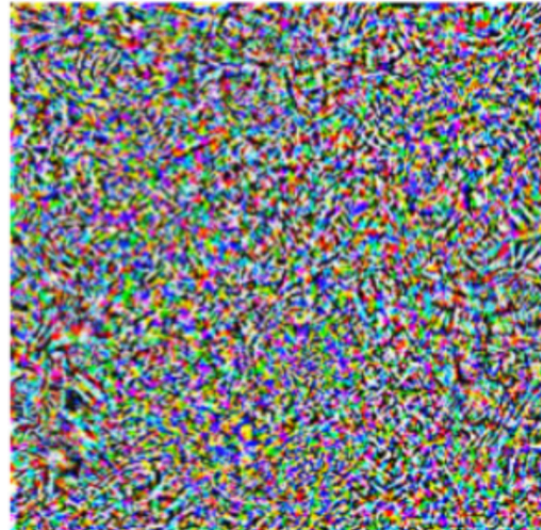


Robustness to Adversarial Attacks

“pig”

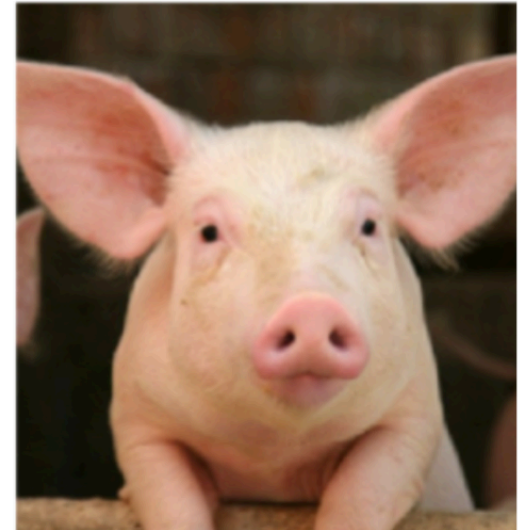


+ 0.005 x

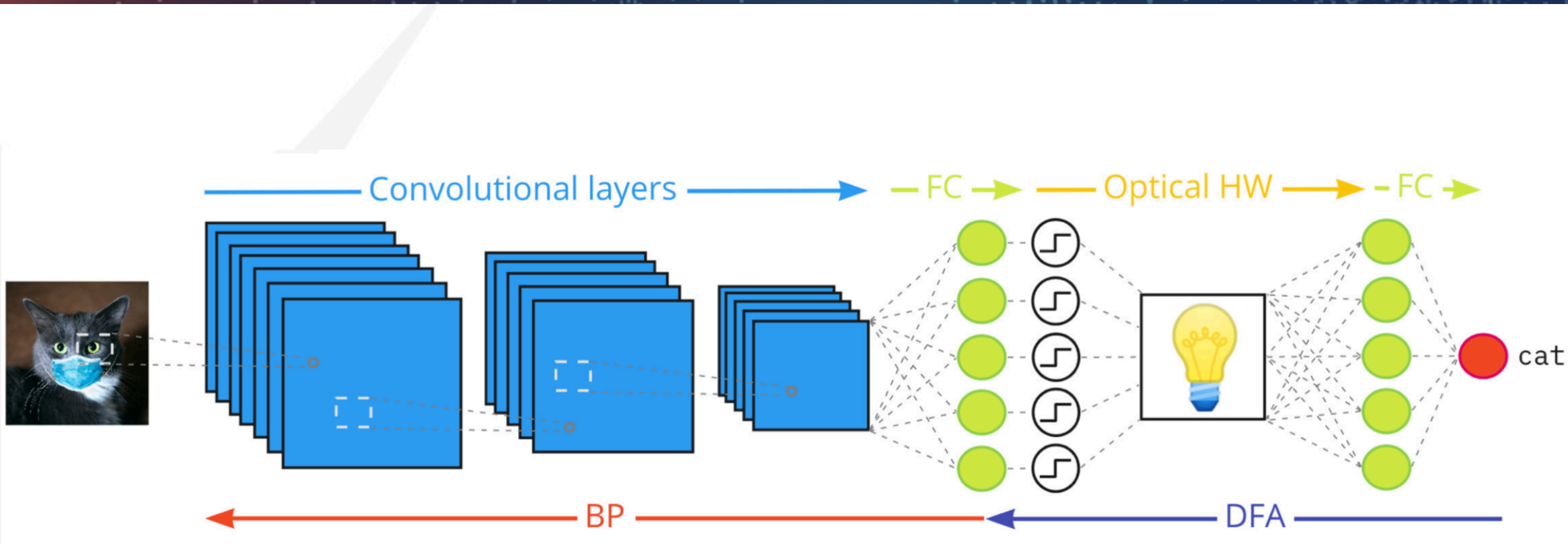


=

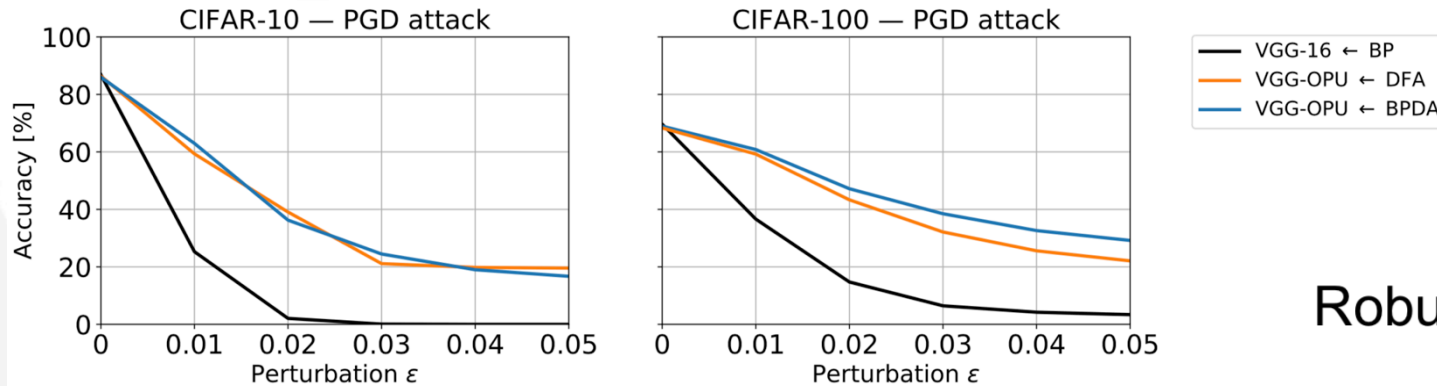
“airliner”



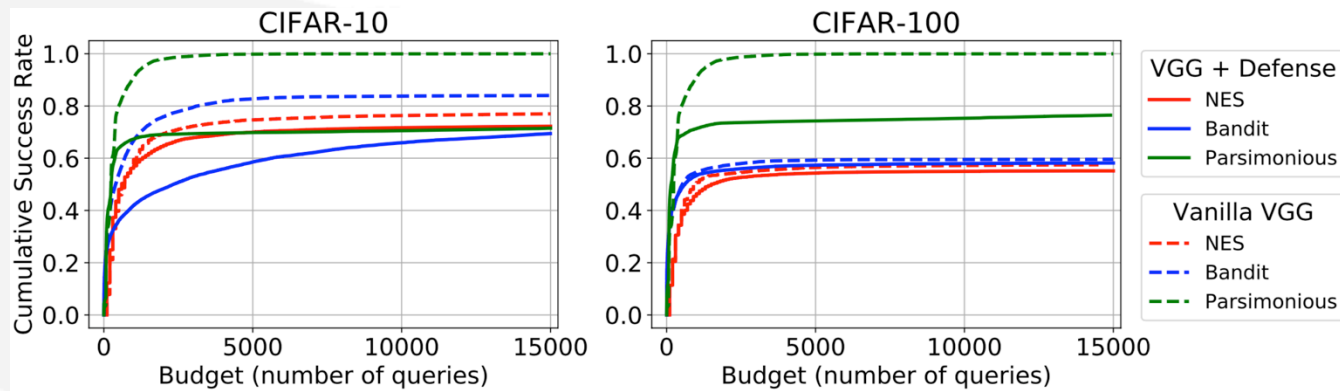
Robustness to Adversarial Attacks



Robustness to Adversarial Attacks

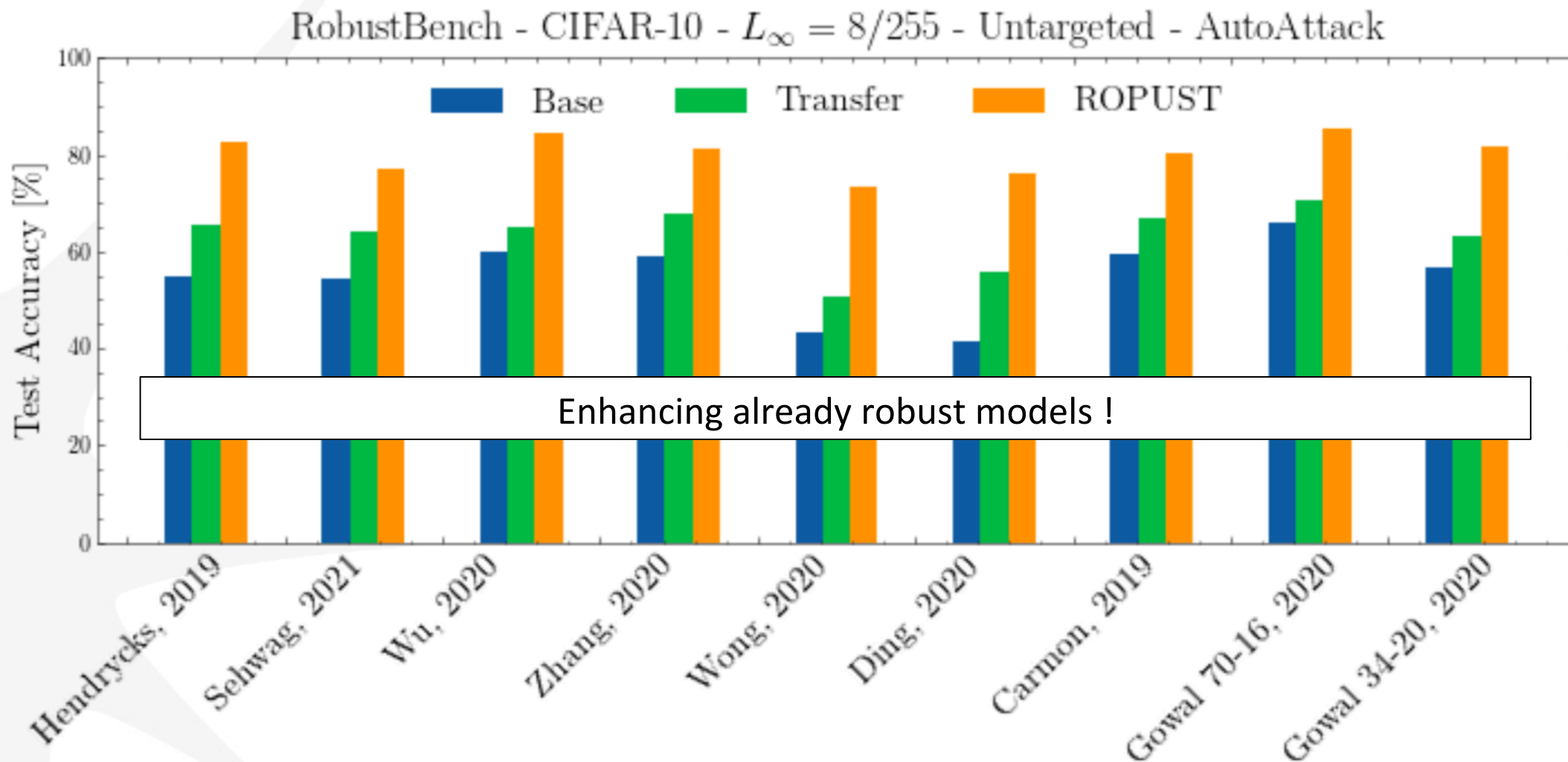


Robust against white-box attacks



Robust against black-box attacks

Robustness to Adversarial Attacks



Photonic Differential Privacy

arXiv.org > cs > arXiv:2106.03645

Computer Science > Machine Learning

[Submitted on 7 Jun 2021]

Photonic Differential Privacy with Direct Feedback Alignment

Ruben Ohana, Hamlet J. Medina Ruiz, Julien Launay, Alessandro Cappelli, Iacopo Poli, Liva Ralaivola, Alain Rakotomamonjy

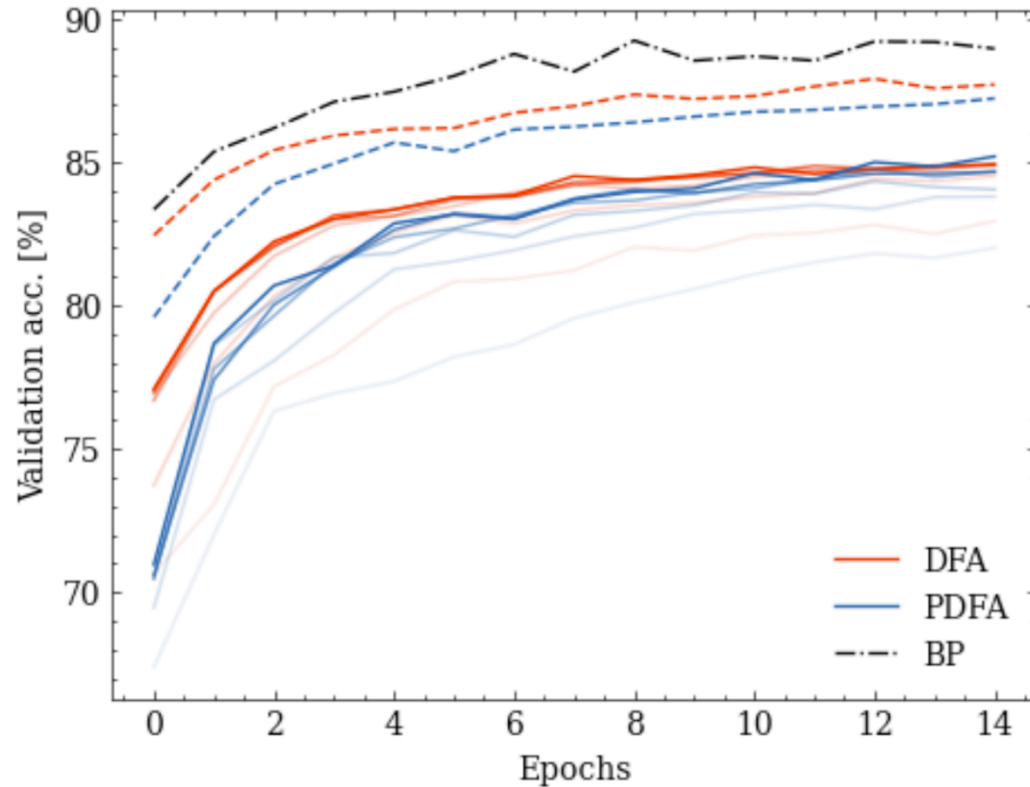


Collaboration with



- Training NN on privacy-sensitive data (eg. medical records)
- Differential privacy
 - inject noise at training, making the contribution of a single data-point hardly distinguishable from the added noise.
 - impossible to infer information on individuals from the aggregate
- « privacy-by-design » using the intrinsic noise from the OPU

Photonic Differential Privacy



Dashed = Non-private - Increasingly transparent = increasingly private (through noise)

Accelerated scientific computing
with
Randomized numerical linear algebra

Randomized Numerical Linear Algebra

DOE RASC report (Jan 2021):

randomized algorithms are "essential to the future of computational science and AI for Science."

- Approximate matrix multiplications
- Randomized SVD
- ... And much more

See Hesslow D. et al, HotChips2021 proceedings, <https://arxiv.org/abs/2104.14429>

Contents lists available at ScienceDirect

Parallel Computing

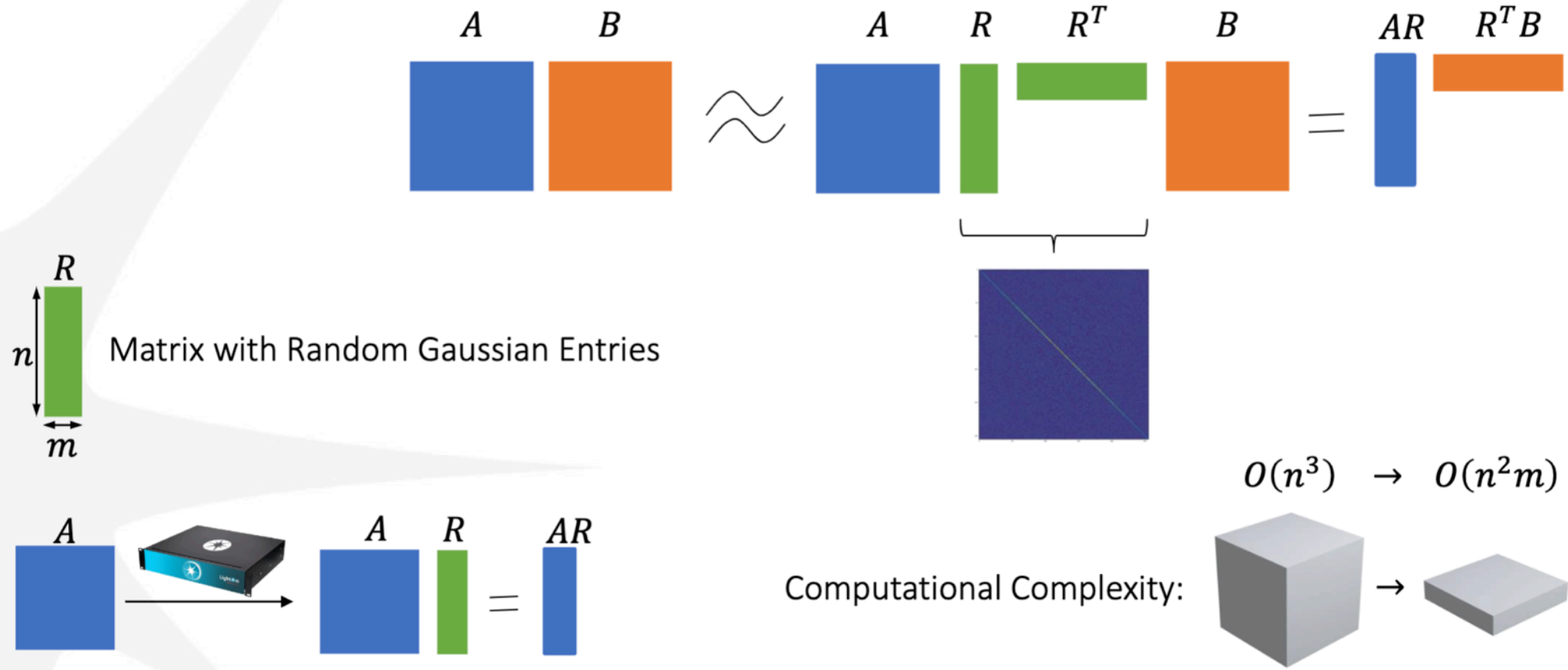
journal homepage: www.elsevier.com/locate/parco

An efficient distributed **randomized algorithm** for solving large dense symmetric indefinite linear systems

Marc Baboulin ^{a,*}, Dulceneia Becker ^b, George Bosilca ^b, Anthony Danalis ^b, **Jack Dongarra ^b**

^aLaboratoire de Recherche en Informatique, Inria/University Paris-Sud, Orsay, France
^bInnovative Computing Laboratory, University of Tennessee, Knoxville, USA

HPC Use case: Accelerated Scientific Computing

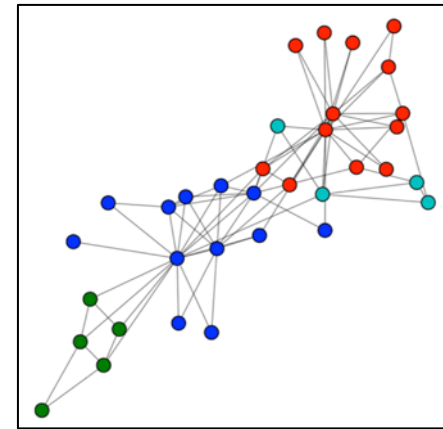


Randomized Trace estimators – application to Graph Neural Networks

Hutchinson's trace estimator

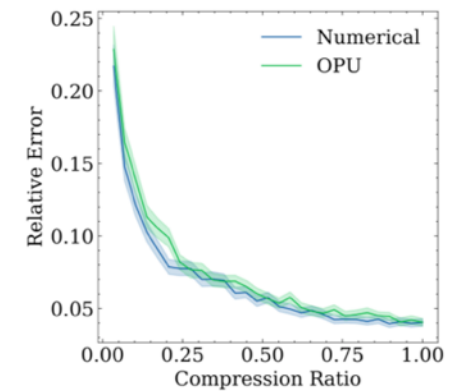
$$\text{Tr}(A) \approx \text{Tr}(RAR^\top)$$

$$\text{Tr}(A^3) \approx \text{Tr}(RA^3R^\top) \approx \text{Tr}((RAR^\top)^3)$$

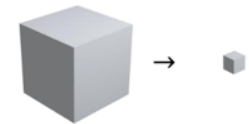


Community detection in networks
→ Triangle counting on graphs

A³ TRACE ESTIMATION

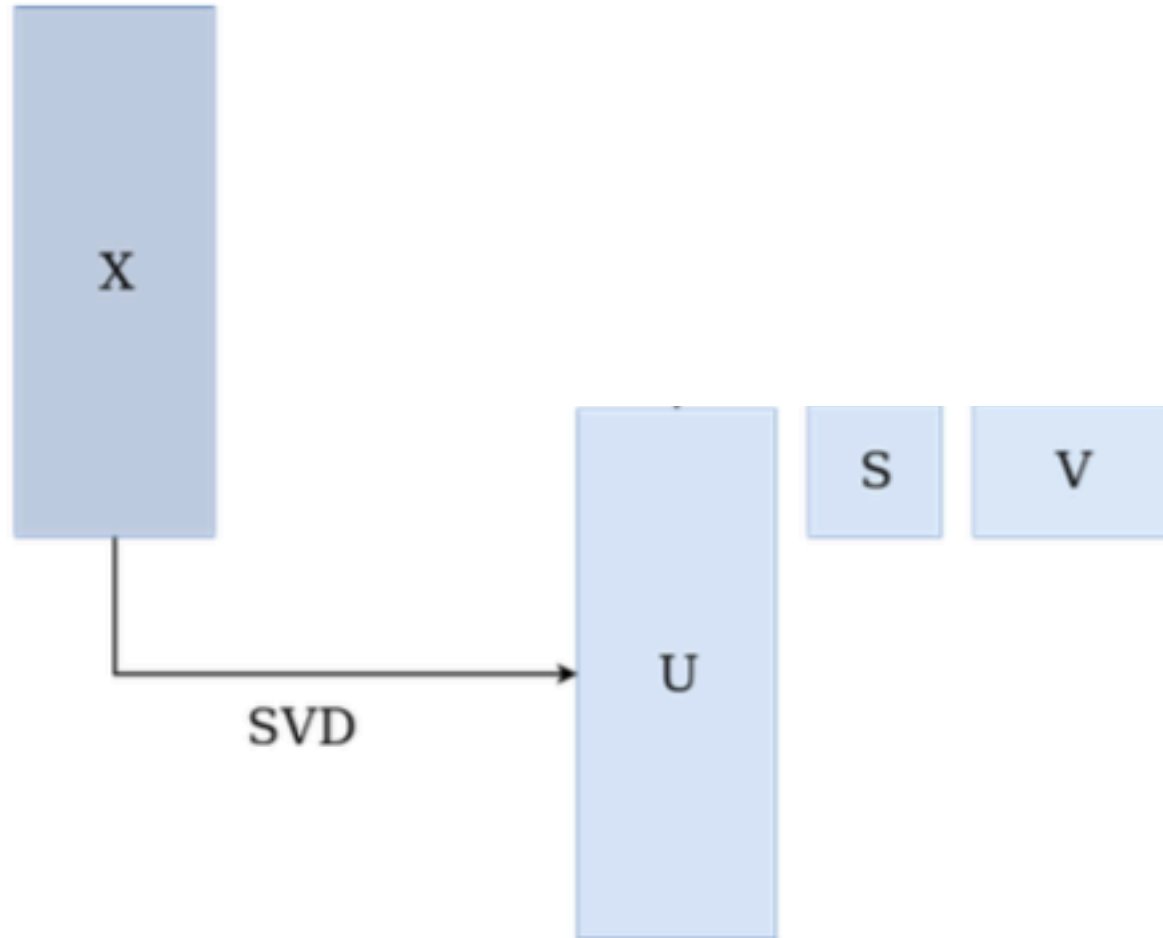


$$O(n^3) \rightarrow O(m^3 + n)$$

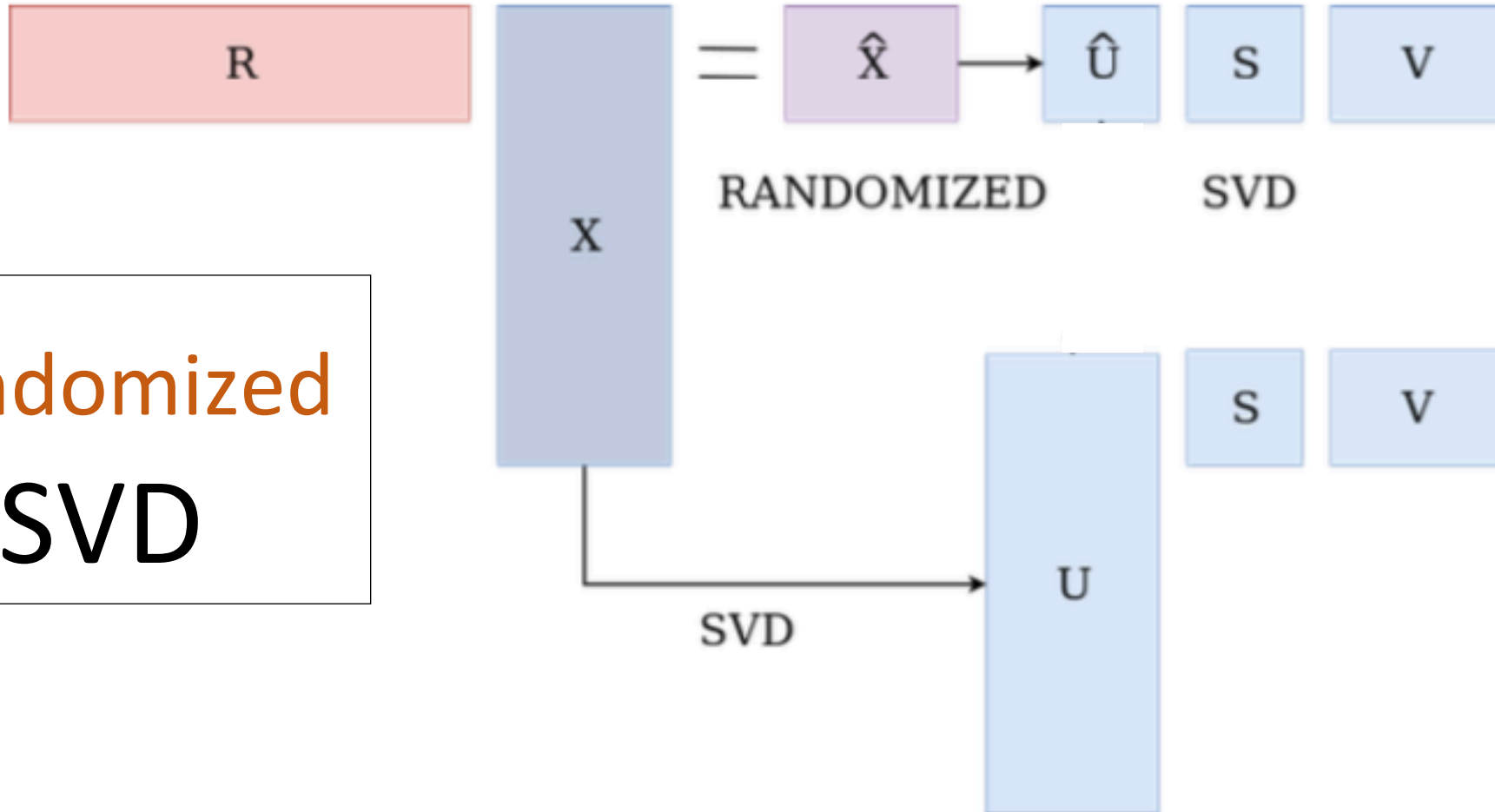


(figure from Rossetti et al. Applied Network Science (2019) 4:52)

SVD



Randomized SVD



Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, Halko, N., Martinsson, P., Tropp, J., 2009, arXiv:0909.4061

Can the OPU train large language models ?

To find out ...

we put an OPU in a #top500 supercomputer !

HPC wire

LightOn Photonic Co-processor Integrated
Into France's Jean Zay Supercomputer
December 23, 2021
PARIS, Dec. 23, 2021



Jean Zay supercomputer

The image shows a screenshot of an HPC Wire article. On the left is a photograph of a black and blue LightOn photonic co-processor unit. On the right is a photograph of a long server rack for the Jean Zay supercomputer, with a portrait of Jean Zay on the front panel. The article title and date are also visible.



lenza

Avec
Jean Perrin,
nos créames
le Centre
national de
recherche
scientifique
1942
1942
1942

Hiv
Ent



"AI babble"

first (massively under-trained) large language model
without backpropagation

DFA: local update, better biological plausibility

Canada, and to the relations there is no side so far water to the centre, whereby the province cuts to such invariably brought from its conclusion that they must have been profited table to prevent its efforts.

Anecdotal here celebrated in America the invasion, which we have been trouble by these marriages are malignitary, in our offence then prevented; so we have the same perfection before had been that there were some alliance at prevailing, and the law, information had previously taken the event of utility, as made it but seems to be disconcerted, the supply on other presented the more difficult than all. In a sailing chance he heard a loading-room or push, with his carcass--here, a solution he made of buttons of crimson brown walls, and by a button, and consuming books of bread from a perfect copy of the site of his head with his cupining. He was solution of a passage accordingly in a brass with the patent, on wantons bore,--which had been represented by the circumstance of

Can the OPU train large language models ?

Yes but:

- Are these useful language models ?
- Can we train faster / save energy ?
- Requires new algorithms

Work in progress !

By way of conclusion

The Hardware Lottery

Sara Hooker — August, 2020



“ How does tooling choose which research ideas succeed and fail ? ”

Building a community



Technical articles, Blogposts, GitHub,
White paper, Newsletter, Meetups...

LightOn.ai



contact@lighton.io



@LightOnIO