Promises and challenges of massive-scale AI

The case of large language models

Laurent DAUDET

LightOn, CTO and Co-Founder / Université de Paris

diiP seminar series, November 3rd, 2021

A new type of “foundation” ML models

LightOn

LightOn

User-provided prompt

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

Automatic text generation

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

A new type of “foundation” ML models

User-provided prompt

Automatic text generation

Write a creative ad for the following product to run on Facebook:

"""

Airee is a line of skin-care products for young women with delicate skin. The ingredients are all-natural.

"""

This is the ad I wrote for Facebook aimed at teenage girls:

"""

Airee is a line of natural skin-care products for young women with delicate skin. The ingredients are all-natural. From the day you're born to the day you're a senior in high school, all your skin needs is love. Airee moisturizes your skin naturally so you feel comfortable wearing beautiful clothing. Also, your skin is protected against bacteria and dirt so you stay healthy and your skin stays supple.

"""

Generate

A new type of “foundation” ML models

Real Estate in Paris

Partager | Générer

User-provided prompt

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

Automatic text generation now **stays in context in French**

The new AI generation : Transformative AI

GPT-3 released in May 2020 triggered **a new AI generation**

- “Understand” text with unprecedented accuracy
- Generate text in the right context and style
- Raw, unlabelled training data at “civilization scale”
- Perform tasks they have not been trained for
- Beyond text : images, video, computer code ...

Across businesses, this is a gamechanger, with “**Transformative AI**” disrupting text-based businesses, content creation, human-computer interaction

→ A worldwide race led by giants

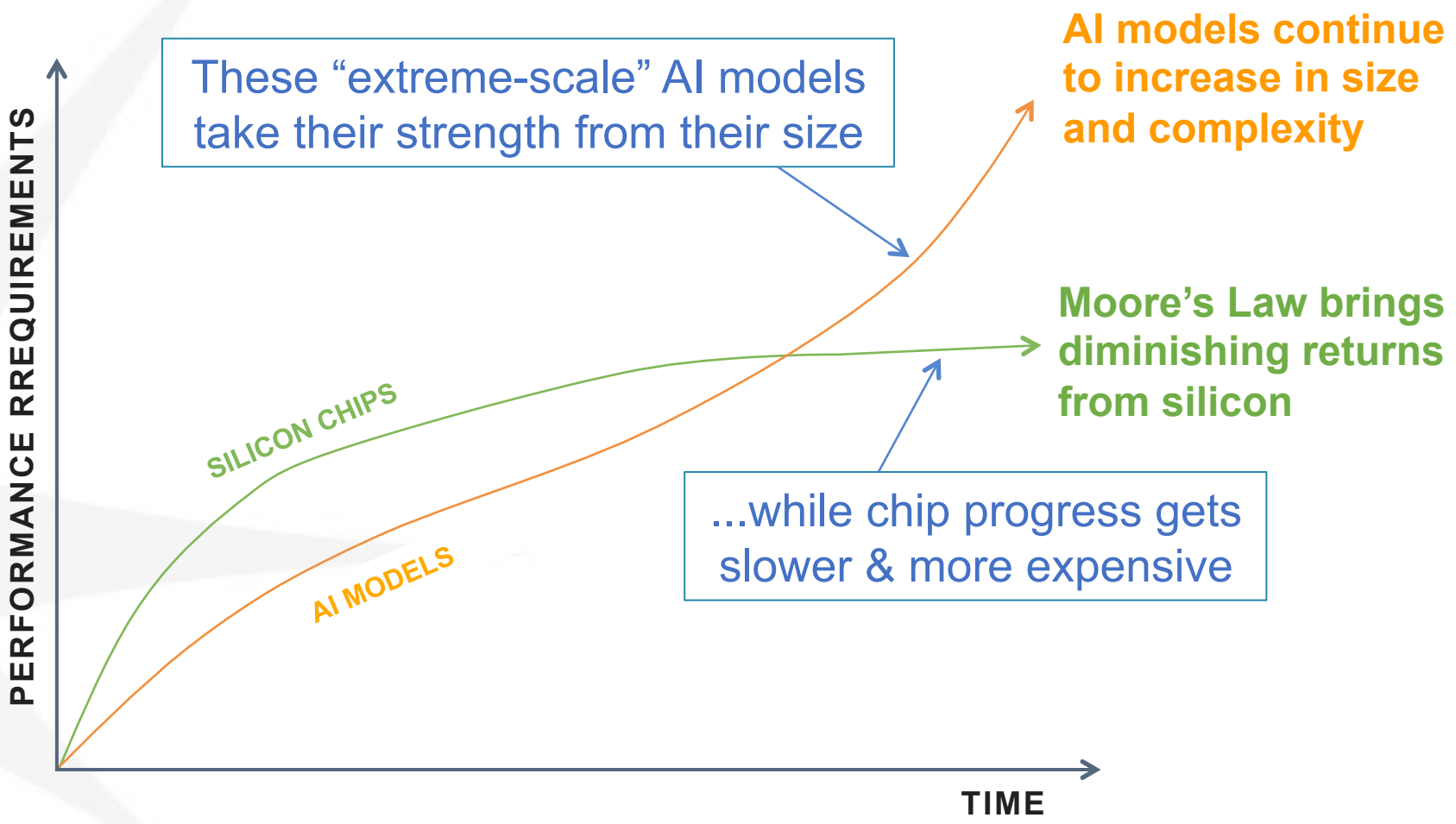
Microsoft Alibaba Google HUAWEI NVIDIA

The hidden ingredient

The barrier to entry: compute

- Entering the era of supercomputer-sized AI
- Training GPT-3 :
  - 3 Million GPU–hours (V100)
  - Estimated price 5-10 M $ for training a single model
- And it’s only the beginning

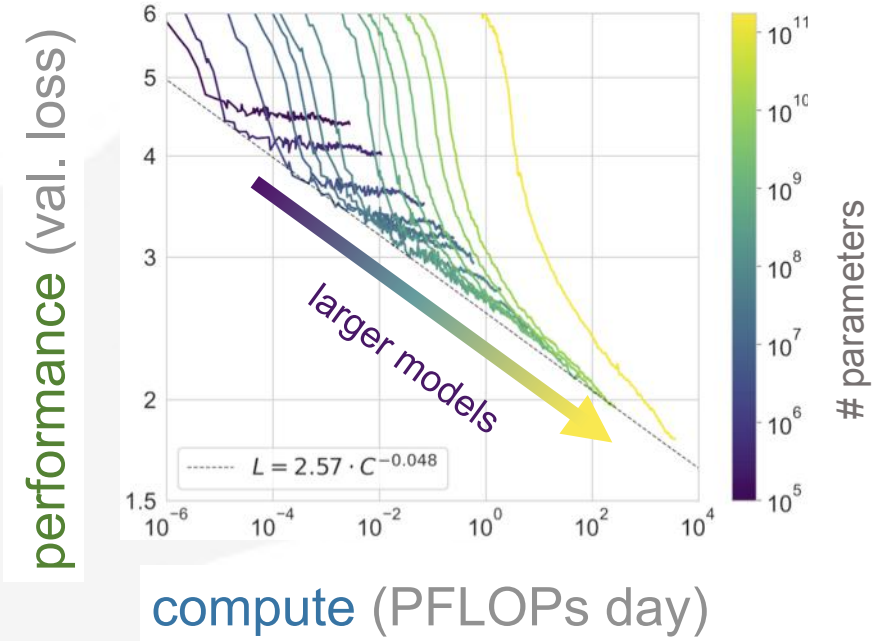
Artificial Intelligence requires *much* faster processing



The AI scaling hypothesis

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

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

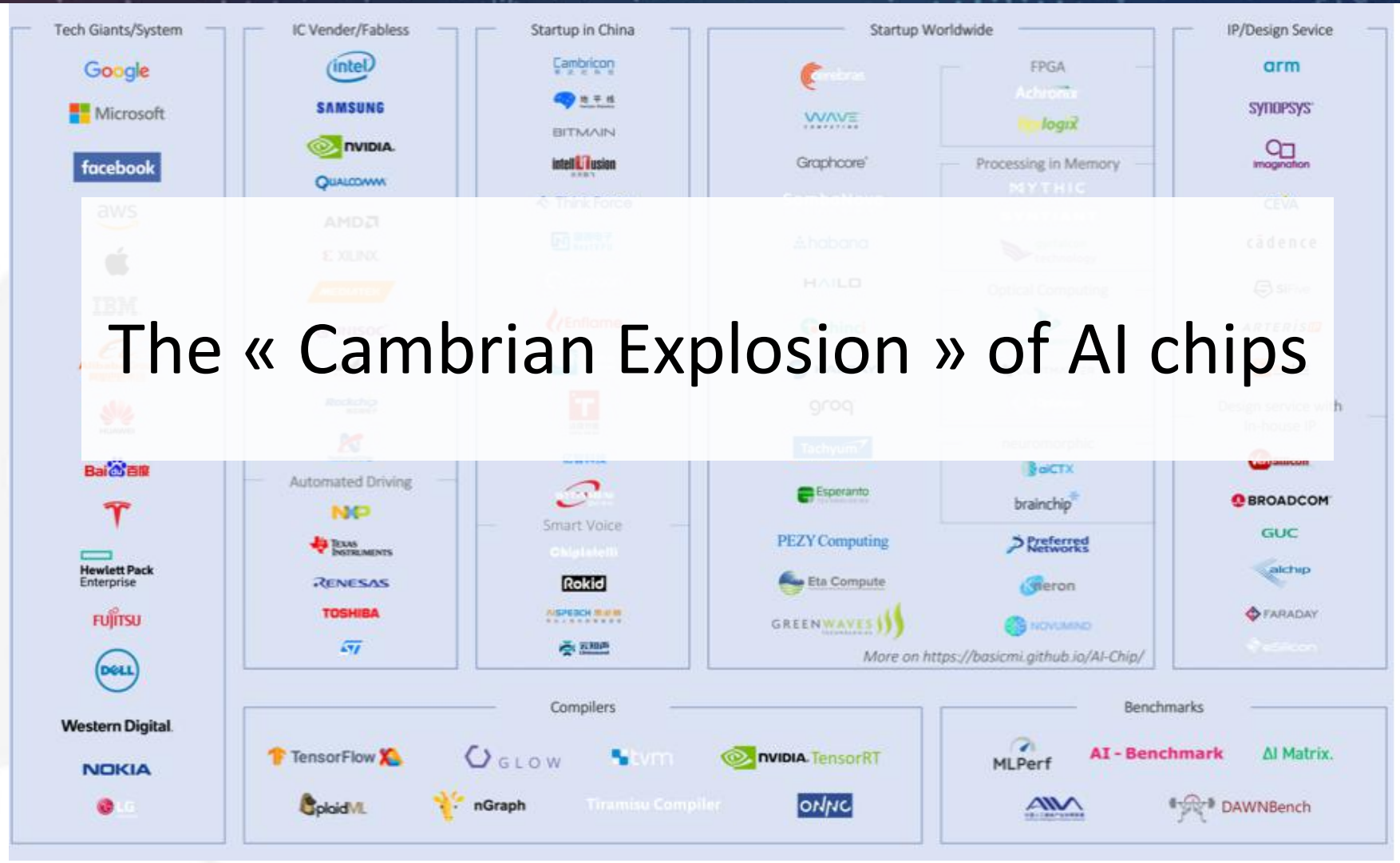


Larger models score higher, generalize better, train faster

> The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

Rich Sutton, The Bitter Lesson.

The « Cambrian Explosion » of AI chips

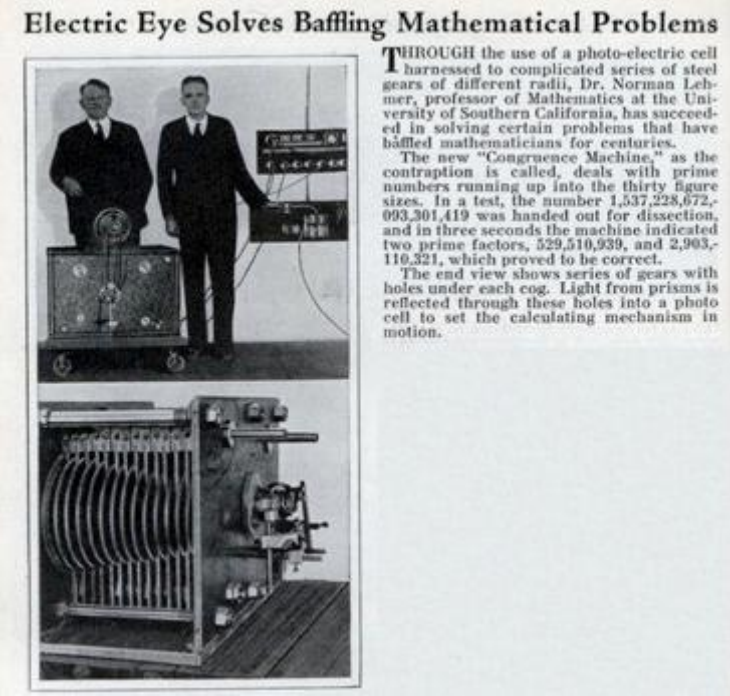


There is light beyond pure silicon

Quantum
Classical Optics

A short history of Optical Processing of Information

From Sieves … to Fourier Transforms … all the way to Neural Networks

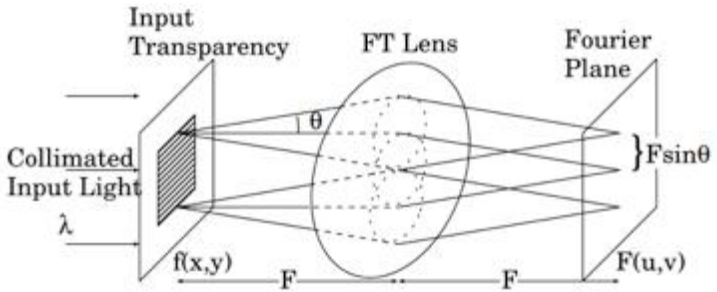


Electric Eye Solves Baffling Mathematical Problems

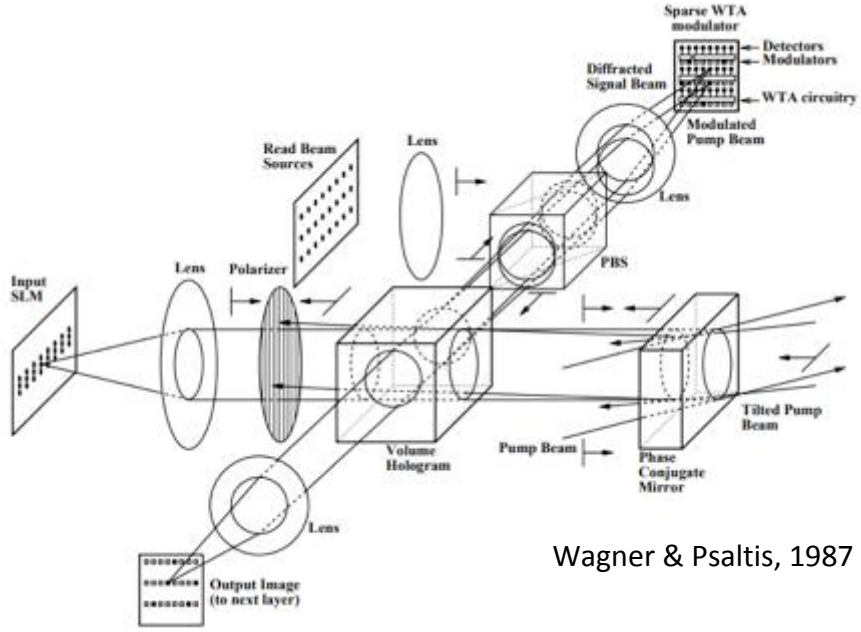
THROUGH the use of a photo-electric cell harnessed to complicated series of steel gears of different radii, Dr. Norman Lehmer, professor of Mathematics at the University of Southern California, has succeeded in solving certain problems that have baffled mathematicians for centuries.

The new "Congruence Machine," as the contraption is called, deals with prime numbers running up into the thirty figure sizes. In a test, the number 1,537,228,672,093,301,419 was handed out for dissection, and in three seconds the machine indicated two prime factors, 529,510,939, and 2,903,110,321, which proved to be correct.

The end view shows series of gears with holes under each cog. Light from prisms is reflected through these holes into a photo cell to set the calculating mechanism in motion.



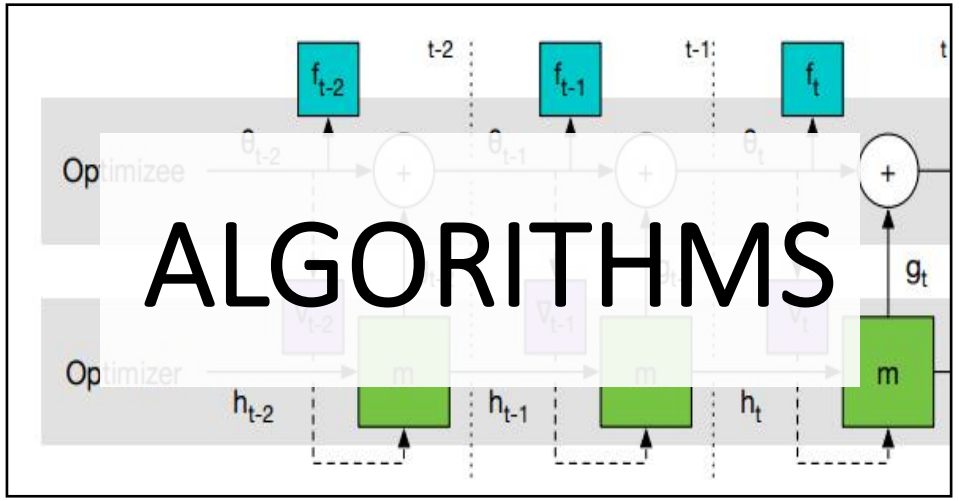
$$F(u,v) = \int\int f(x,y) e^{i\frac{2\pi}{\lambda F}(xx'+yy')} dxdy$$



Wagner & Psaltis, 1987

1930's → 1950's → 1980's

Rebooting Optical Computing: the AI era



How LightOn *actually* started

Procrastination

Coffee

It all started with a blog

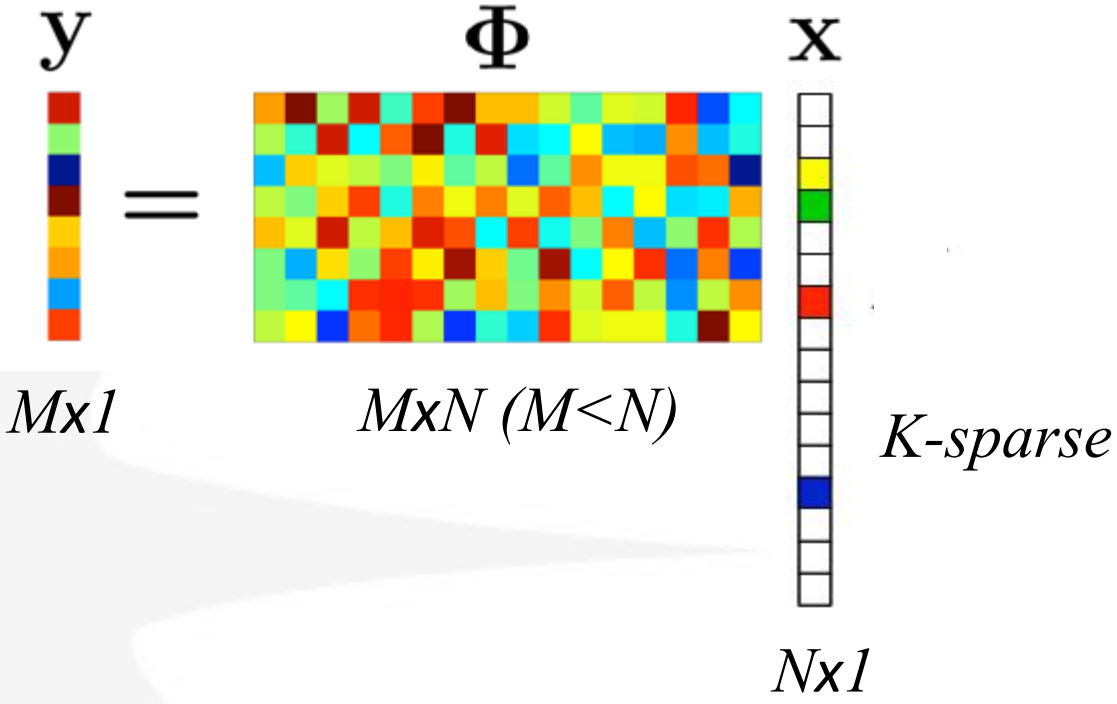
Information theory Compressive Sensing

Igor

http://nuit-blanche.blogspot.com

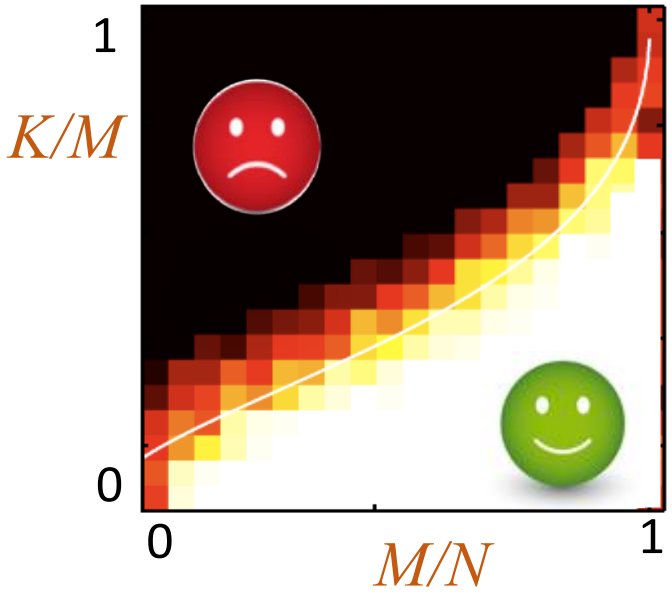
Laurent

Compressive Sensing



Can one recover x from y ?

YES with tractable algorithms for right values of N, M, K



Lessons from Compressive Sensing

- Signals can be sampled at the level of their information content
- Random Projections are very good for sensing at low data rate

Information theory Compressive Sensing

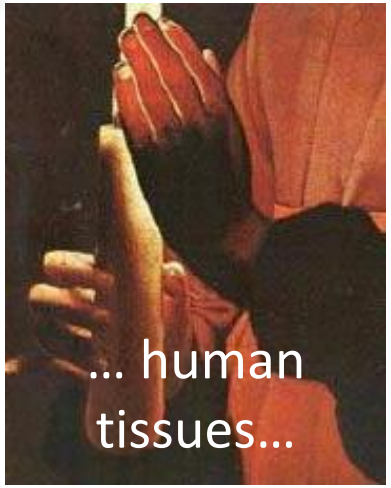
Igor

Laurent

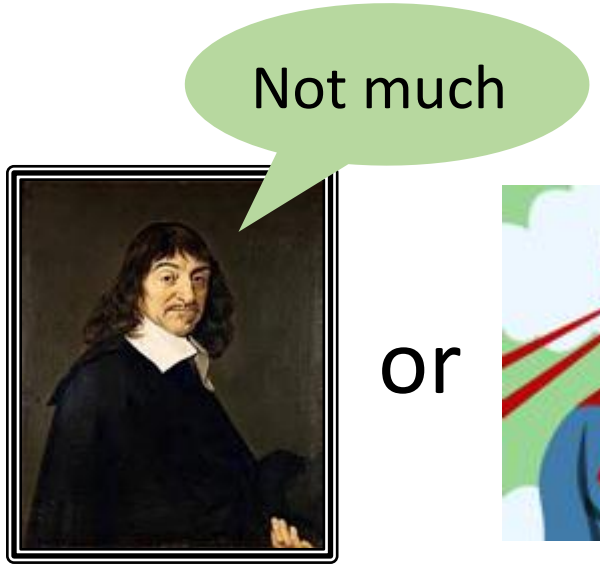
Light Transport in Diffusive Media

Sylvain

Light transport in diffusive media



How deep can one see ?

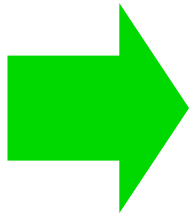


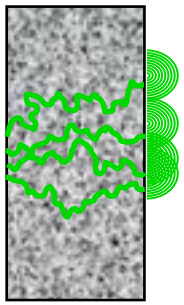
or

?

Scattering: a coherent process

laser

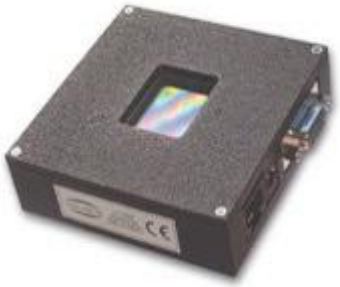




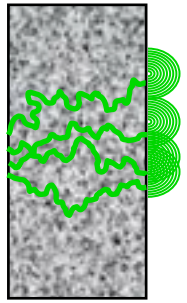
« Speckle »

Modulated beam

laser



Spatial Light Modulator



CMOS camera

Scattering: a coherent process

Discrete input vector

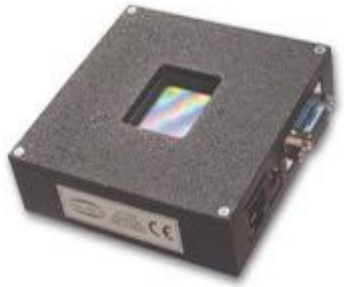
$$\begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,N} \\ h_{2,1} & h_{2,2} & \dots & h_{2,N} \\ \vdots & & \ddots & \vdots \\ h_{M,1} & h_{M,2} & \dots & h_{M,N} \end{pmatrix}$$

Random transmission matrix

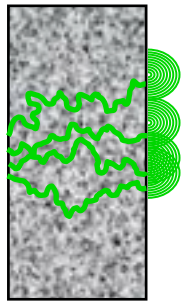
Discrete output vector (speckle intensity)

x H $y = | H x |^2$

laser



Spatial Light Modulator



CMOS camera

Popoff et al. *Nat. Commun.* 1:81 doi: 10.1038/ncomms1078 (2010)

Lessons from Light Transport in Diffusing Media

- **Scattering preserves the information content**
- **Scattering *optimally scrambles* information**
  - just like a Random Projection
  - just like in Compressive Sensing
- **Matrix-vector multiplication, followed by non-linearity**

  → ubiquitous in Machine Learning !

Information theory Compressive Sensing

Igor

Laurent

Light Transport in Diffusive Media

Sylvain

Machine Learning

Florent

Random Projections in Machine Learning

- Random Projections act as distance-preserving point cloud embeddings

Johnson-Lindenstrauss Lemma (1984)

Lemma For any $0 < \epsilon < 1$ and any interger n let k be a po sitive interger such that

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n$$

then for any set A of n points $\in \Re^d$ there exists a map $f : \Re^d \to \Re^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2$$

- NeurIPS 2017 Test of Time Award
  "Random Features for Large-scale Kernel Machines", Rahimi, Recht, 2008

Lessons from Machine Learning

A **Random Projection** is an elementary computing building block that is well matched to the statistical nature of Machine Learning

- an optimal mixer / preconditioner of information
- changes the dimension of a set of vectors without changing distances / angles
- for data compression *or* expansion
- useful in supervised or unsupervised settings
- can be seen as a dense (fully connected) layer of a DNN

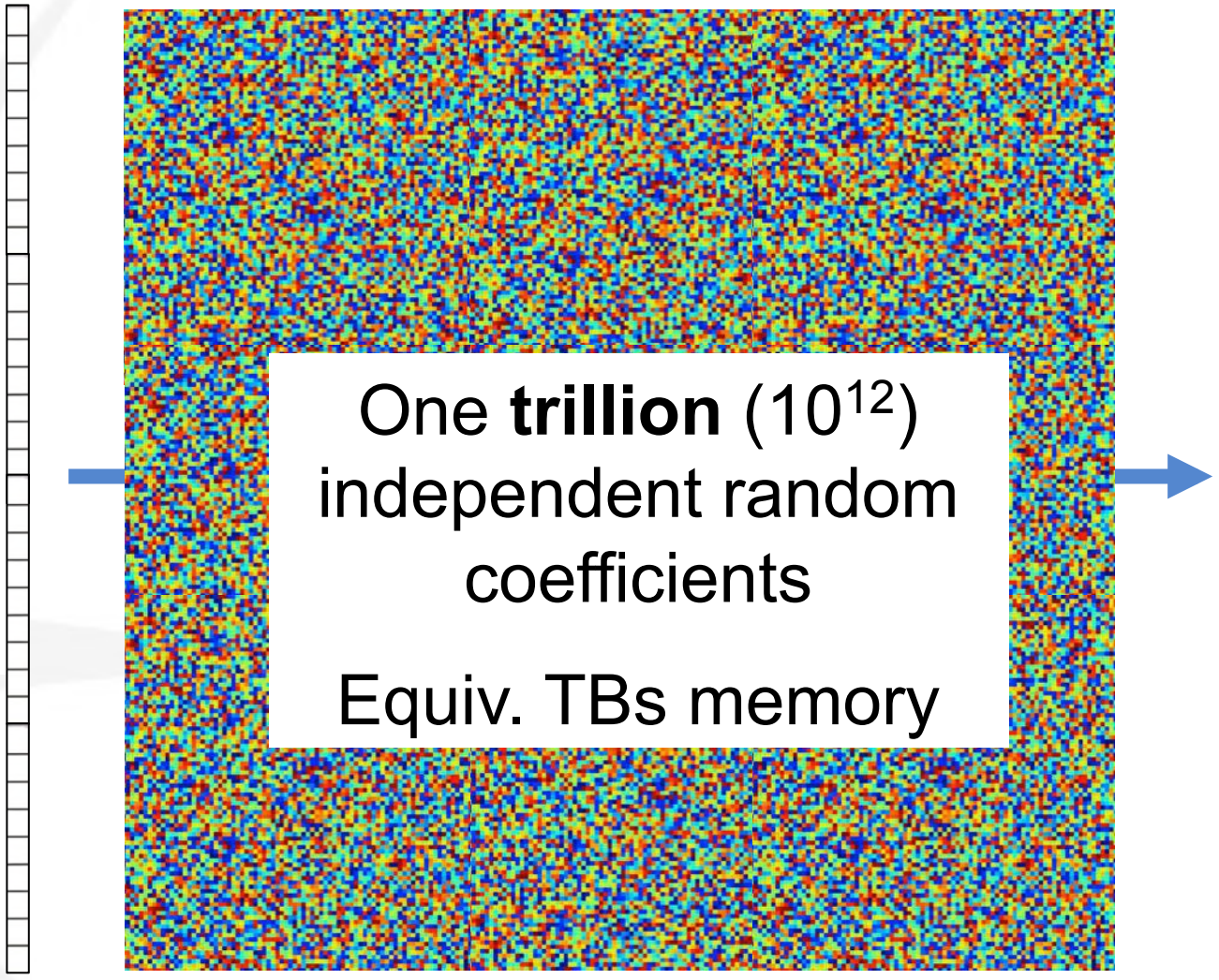
Lessons from Machine Learning

Random projections made $O(n^2) \rightarrow O(1)$

- How does it change computing pipelines ?
- Can we engineer it so that it has economic value ?

Matrix-vector multiplication through light scattering

1 million independent input pixels



One **trillion** (10^{12}) independent random coefficients

Equiv. TBs memory

1 million independent output pixels

Designed for Large Scale Machine Learning

The OPU performs **Random Projections** in the analog domain

input vector x → *output vector* y = Hx or $y = |Hx|^2$

with H a *fixed by design* random matrix

LARGE	&	FAST
H of size higher than 10^6 x 10^6 (TBs of memory)		kHz operation →10^3 such multiplies / s

Equivalent 10^{15} OPS … for a few W

* Analog non-programmable - non Von Neumann - OPS not directly comparable to Flops

LightOn Appliance

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

Ultra-fast: 1500 TOPS

In a single photonic core

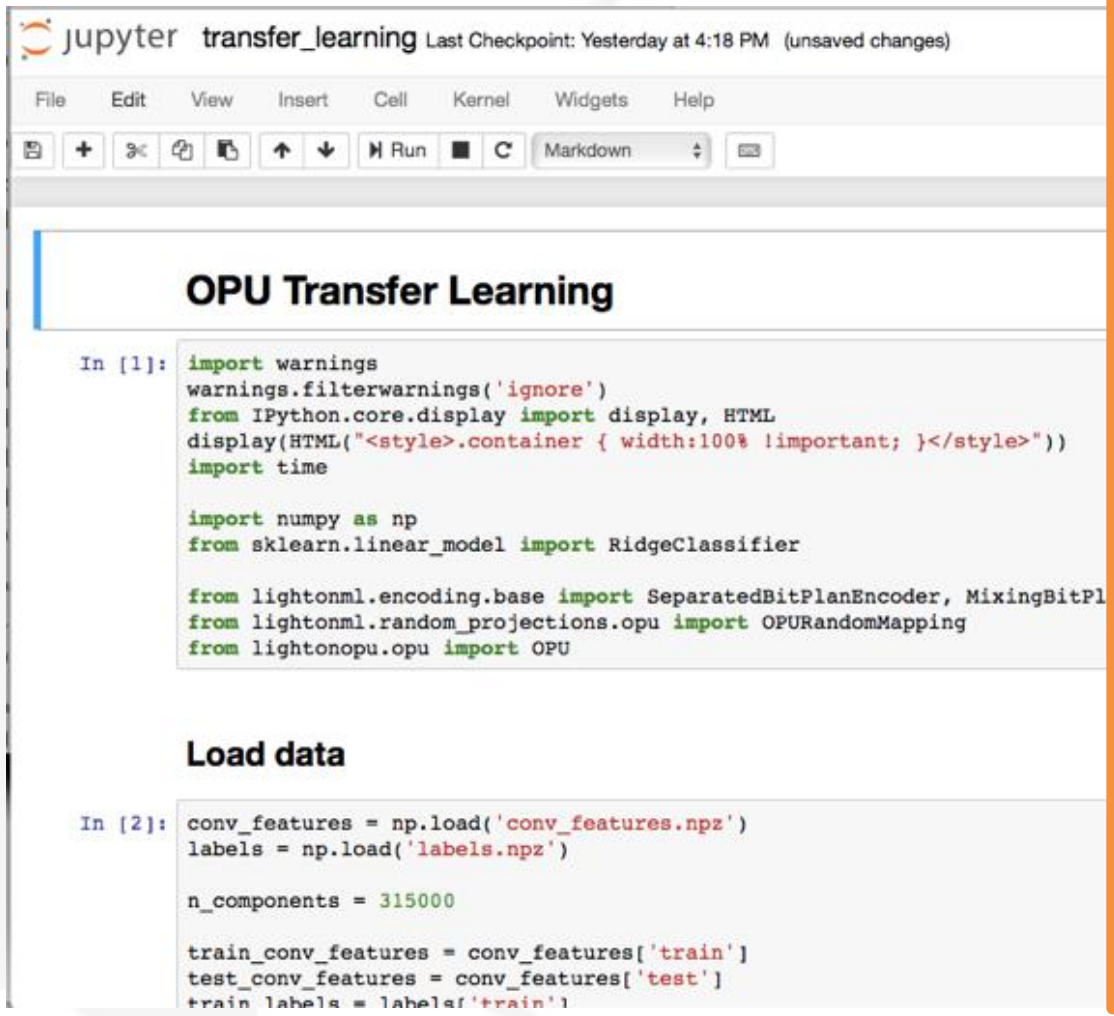
With only 30 W TDP

200 times better in #OPS/W than NVIDIA latest GPU boards

Reduces the energy impact of AI

Provides enhanced data privacy

User interface (Jupyter)



LightOn Cloud : remote access to OPU, free for academic research

Access to OPU through Python library

Compliant with popular dev and Machine Learning environments

python PyTorch learn jupyter ANACONDA.

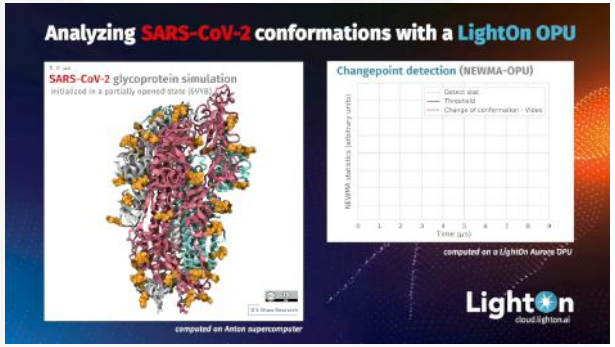
Hybrid computing in AI pipelines

Ivan Dokmanic
Associate Prof.
UNIVERSITY OF ILLINOIS

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 @ichep2020 conference ! (Remote in July or Prague beg 2021)#hepml @Laurent_Daudet @IgorCarron

6 15



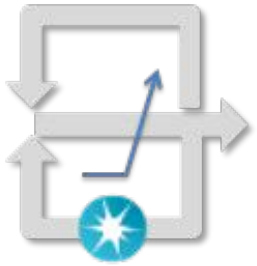
Accelerating SARS-COv2 Molecular Dynamics Studies with Optical Random Features

Amélie Chatelain
LightOn ML R&D engineer

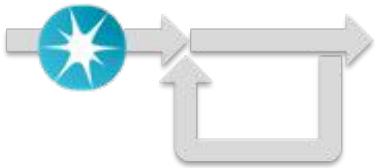
Pre-conditioning
for RandSVD

Transfer learning
in Convolutional Deep NN

Reinforcement Learning



Sketching
for changepoint detection



Reservoir computing
for time / time-space series

Adversarial robustness / differential privacy
by design

Collaboration with Criteo

criteo

Collaboration with FAIR

Facebook AI Research

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2020

Big Science: time series

Nicolas Keriven, CNRS researcher with the CICS team (Communication and Information in Complex Systems) in the GIPSA laboratory, Grenoble, France.

How to detect changes / anomalies within the 1000s of signals monitoring a complex system (factory, airplane engine, power plant, stock market…) ?

Example : monitoring a datacenter

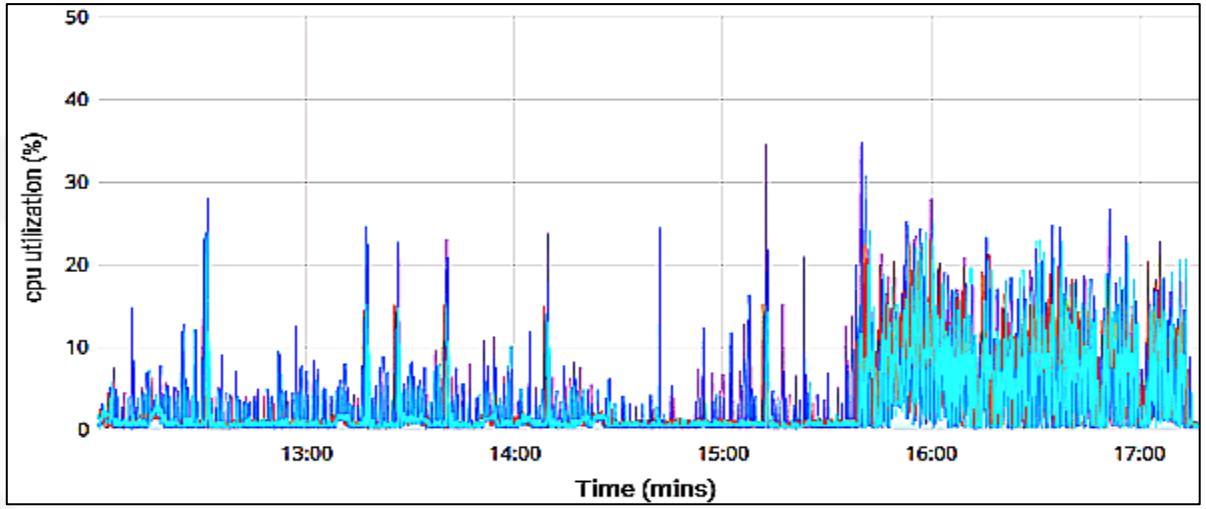


Image from Valsamas *et al*, 2018

Big Science: time series

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 68, 2020 3515

NEWMA: A New Method for Scalable Model-Free Online Change-Point Detection

Nicolas Keriven, Damien Garreau, and Iacopo Poli

***Abstract*—We consider the problem of detecting abrupt changes in the distribution of a multi-dimensional time series, with limited computing power and memory. In this paper, we propose a new, simple method for model-free online change-point detection that relies only on fast and light recursive statistics, inspired by the classical Exponential Weighted Moving Average algorithm (EWMA). The proposed idea is to compute *two* EWMA statistics on the stream of data with different forgetting factors, and to compare them. By doing so, we show that we implicitly compare recent samples with older ones, without the need to explicitly store them. Additionally, we leverage Random Features (RFs) to efficiently use the Maximum Mean Discrepancy as a distance between distributions, furthermore exploiting recent optical hardware to compute high-dimensional RFs in near constant time. We show that our method is significantly faster than usual non-parametric methods for a given accuracy.**

***Index Terms*—Change detection algorithms, Hilbert space, method of moments, optical computing.**

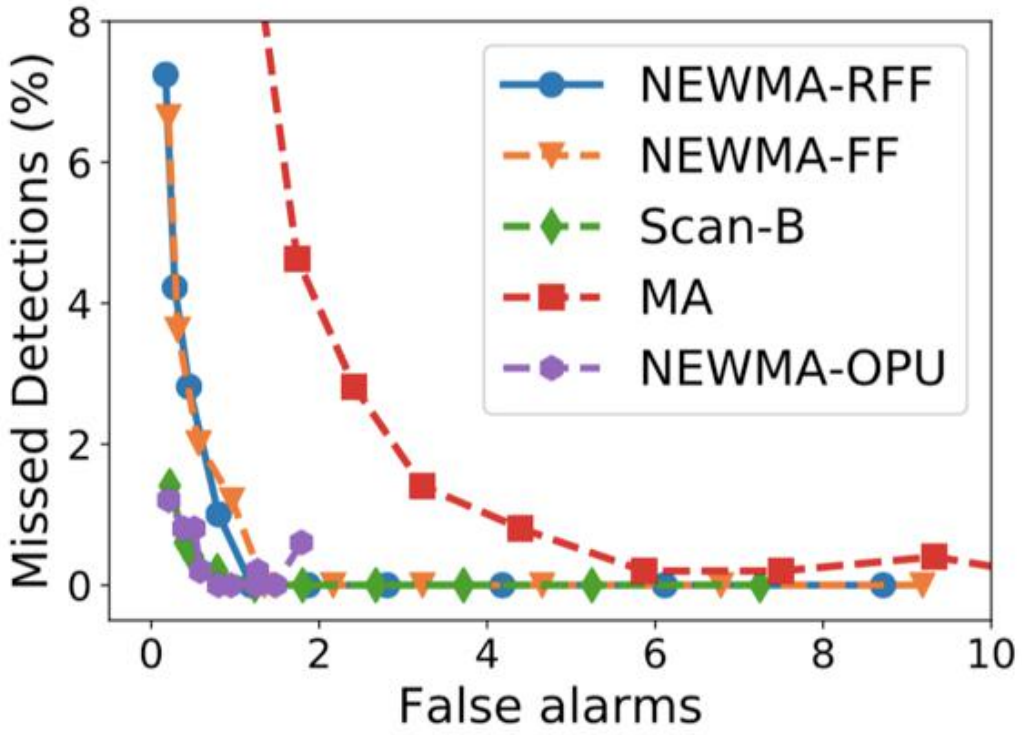
In this paper, we propose a new approach for online, non-parametric change-point detection, whose main advantage is that it does not require to store any raw data in memory, but only appropriate smoothed quantities. It is inspired by: a) the classical Exponentially-Weighted Moving Average (EWMA), but requires less prior knowledge about the in-control distribution of the data, and b) a simple Sliding Window (SW) strategy in its model-free version, but is more efficient in memory and preserves data privacy.

A. *Framework: Model-Free Methods and Generalized Moments*

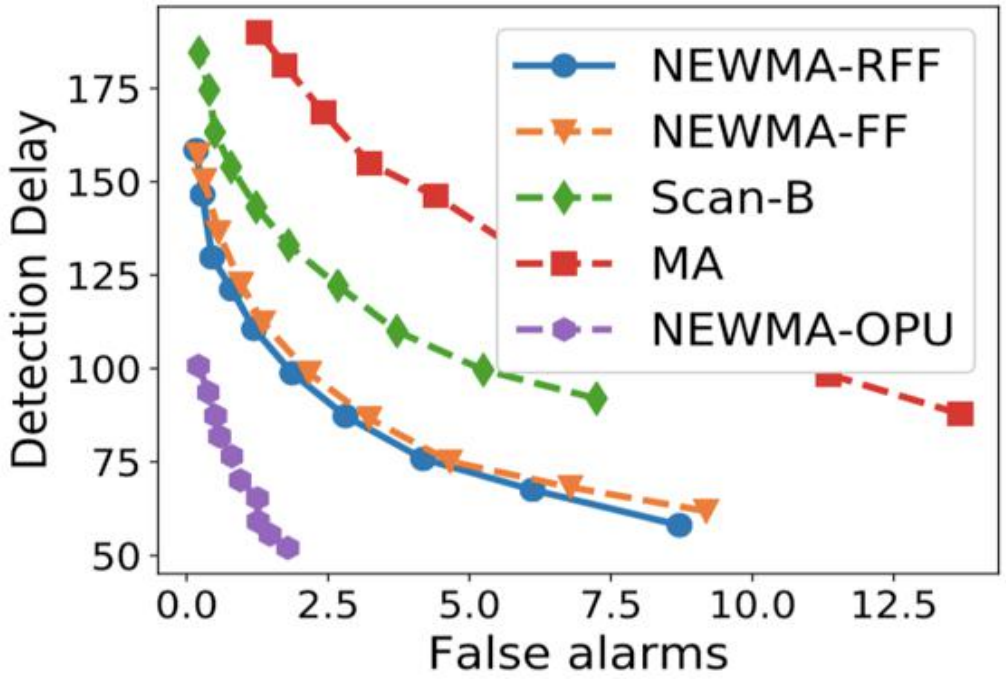
We consider a stream of samples $(x_t)_{t\in\mathbb{N}}$ with values in $\mathbb{R}^d$ with potentially large d. The goal of online change-point detection is to detect changes in the distribution of the samples x_t in a sequential manner. We assume that the samples are inde-

Big Science: time series

Detection accuracy

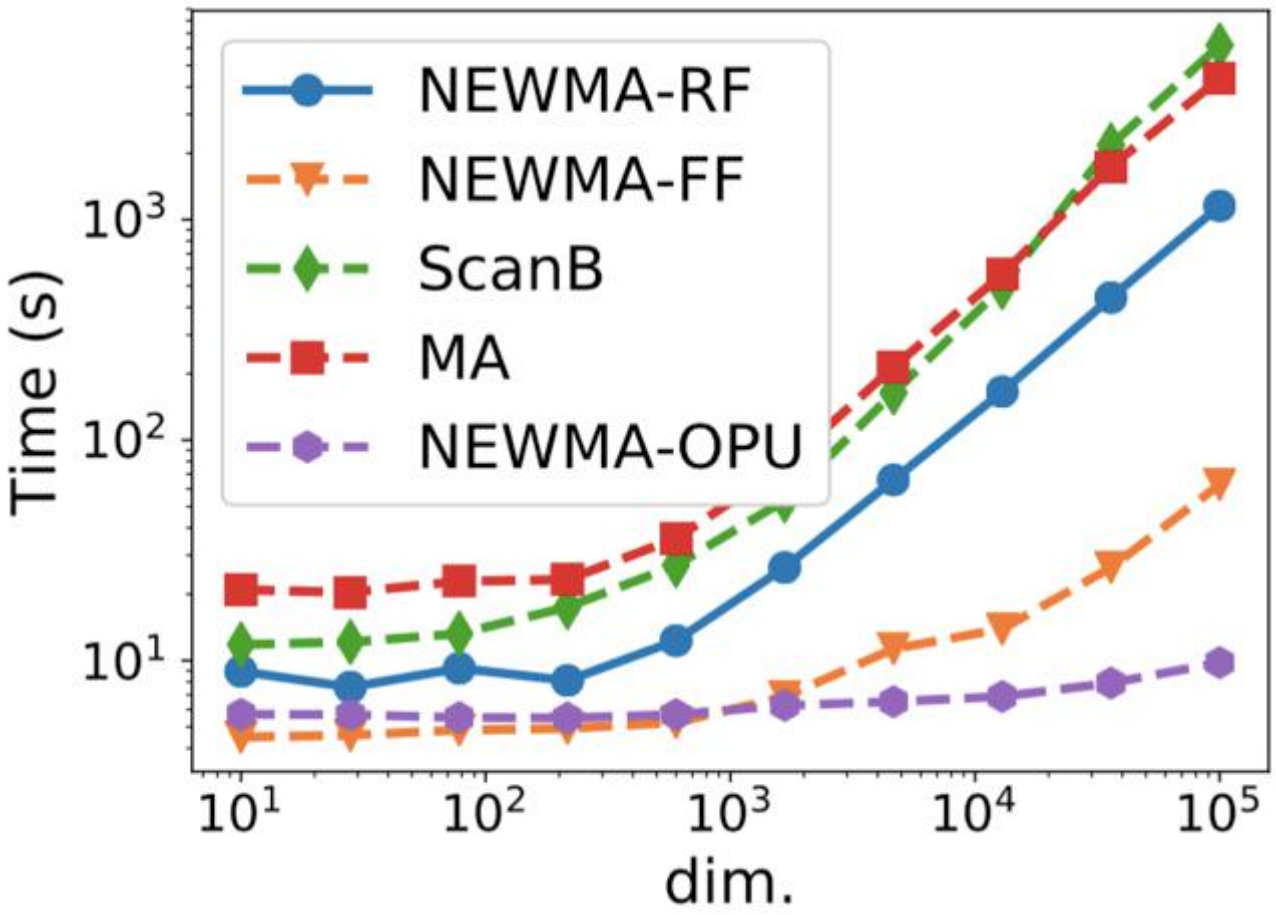


Detection delay



Big Science: time series

Computing time



Several orders of magnitude faster than state-of-the-art

OPU compute time is essentially **idenpendent of dimension !**

HPC Use case: Accelerated Scientific Computing

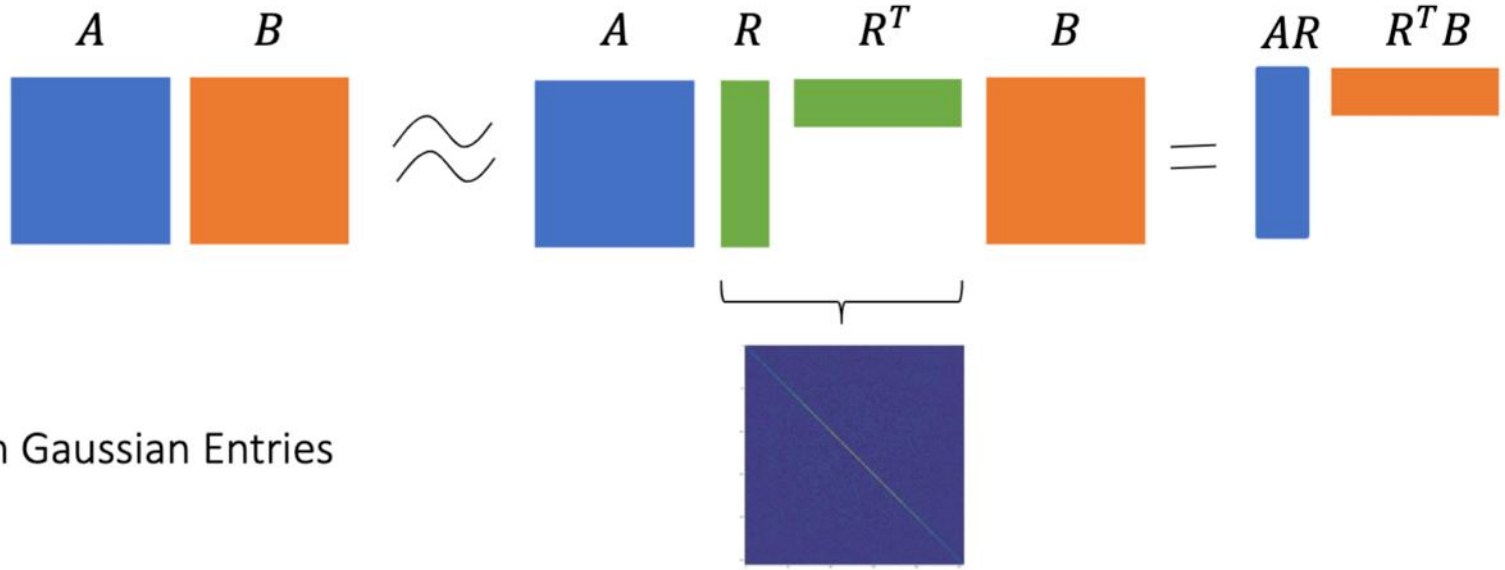
Randomized Numerical Linear Algebra

DOE RASC report (Jan 2021):
randomized algorithms are "essential to the future of computational science and AI for Science."

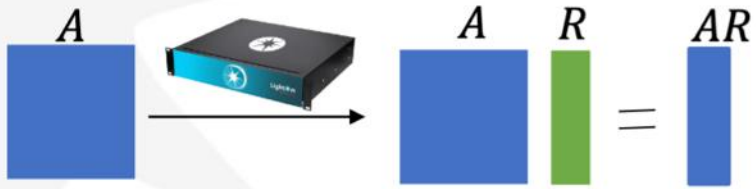
- Approximate matrix multiplications
- Randomized SVD → recommender systems
- … And much more

More info: https://arxiv.org/abs/2104.14429

HPC Use case: Accelerated Scientific Computing

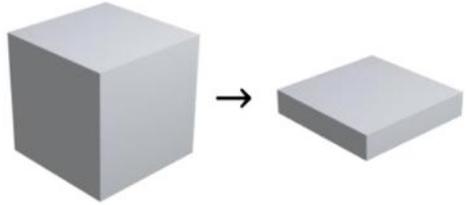


Matrix with Random Gaussian Entries



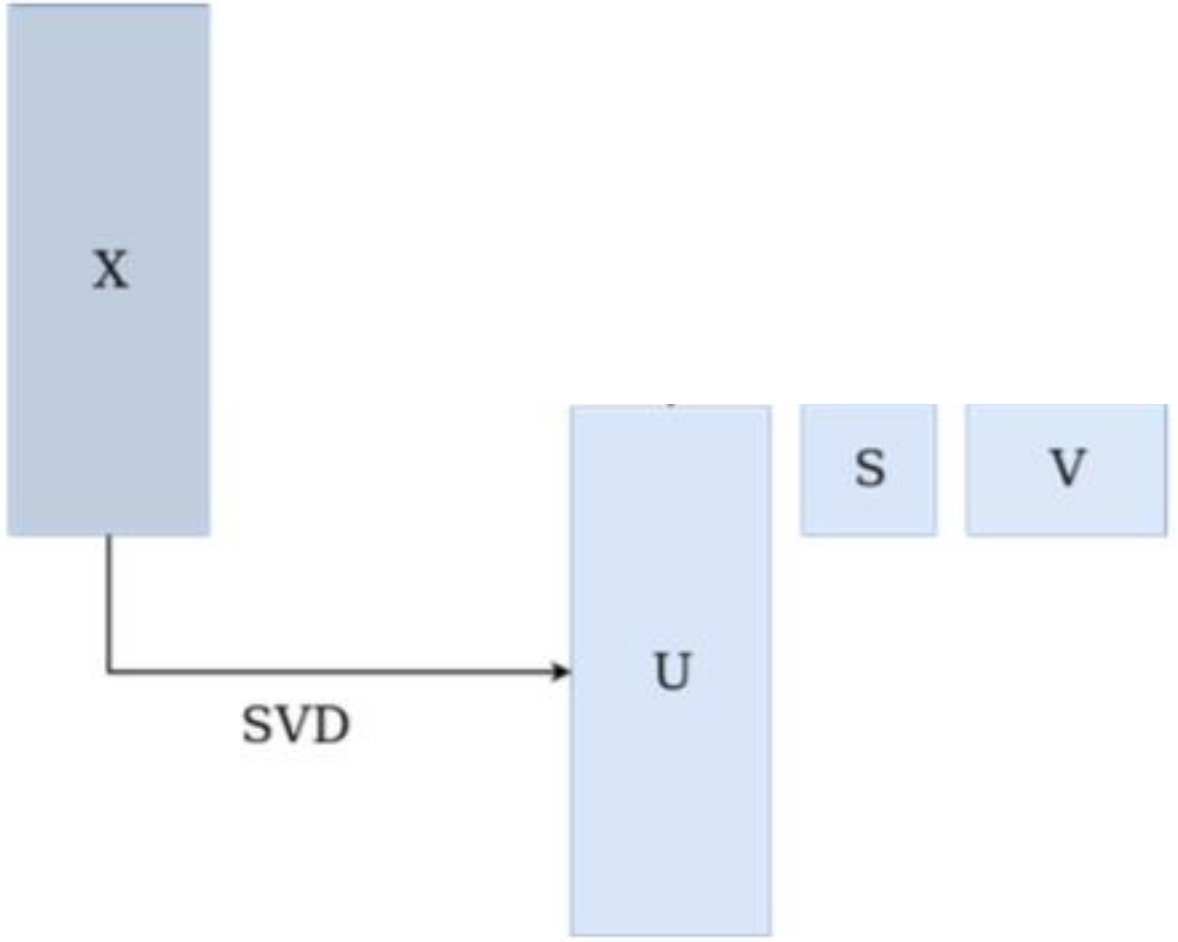
$O(n^3) \rightarrow O(n^2 m)$

Computational Complexity:



HPC Use case: Accelerated Scientific Computing

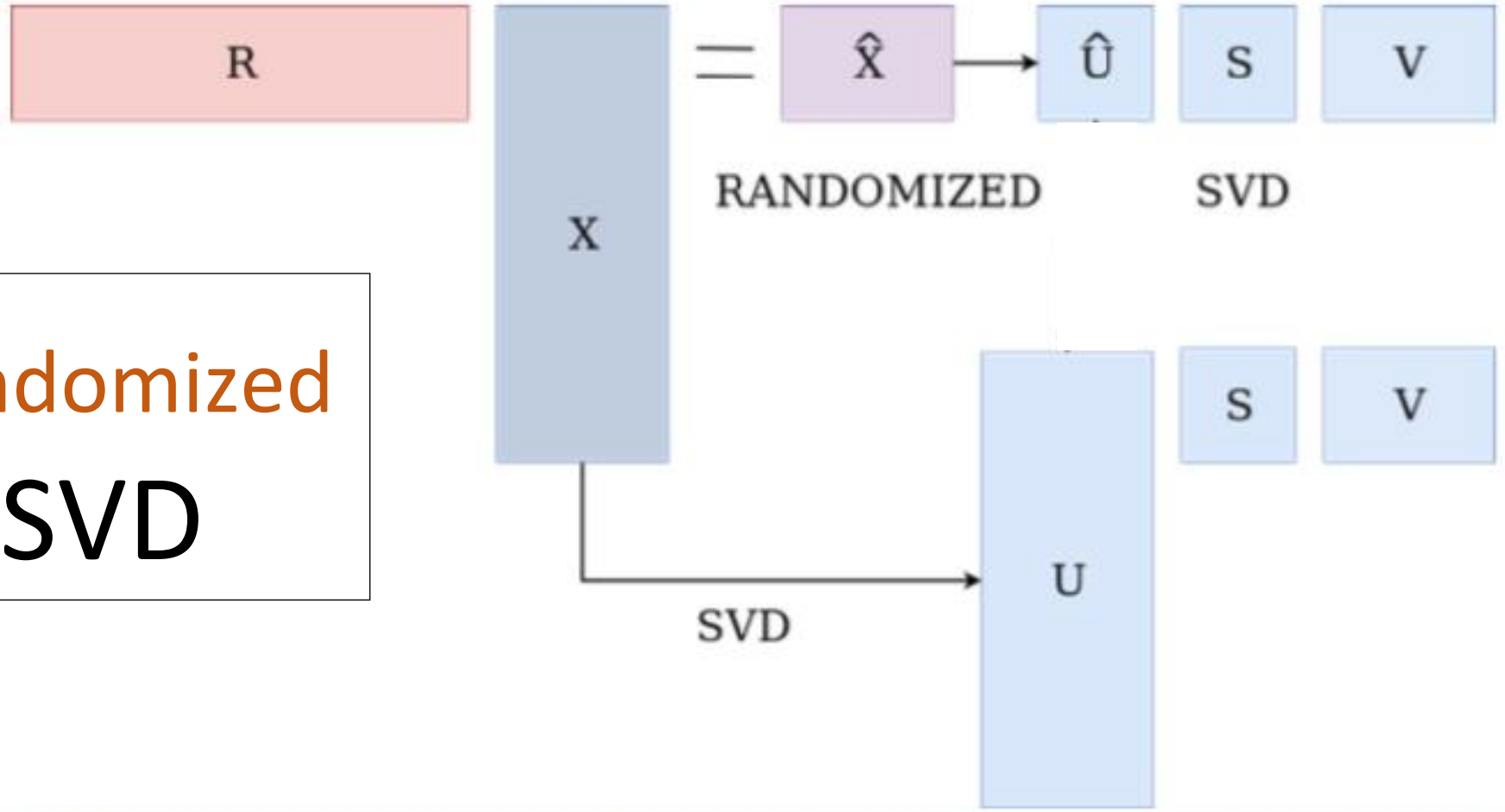
SVD



HPC Use case: Accelerated Scientific Computing

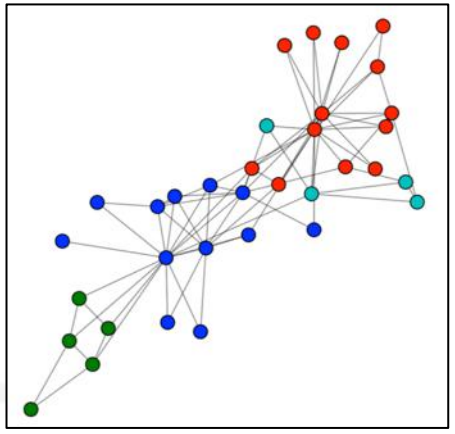
Randomized

SVD



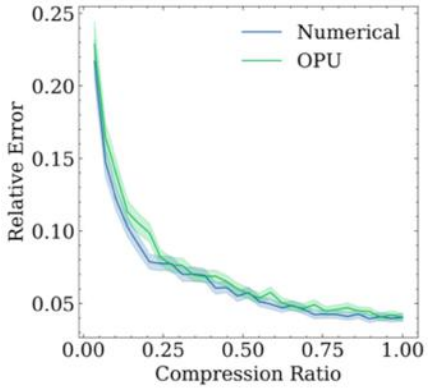
Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, Halko, N., Martinsson, P., Tropp, J., 2009, arXiv:0909.4061

Randomized Trace estimators

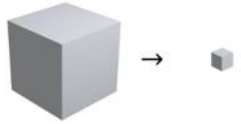


Community detection in networks
→ Triangle counting on graphs

A^3 TRACE ESTIMATION



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

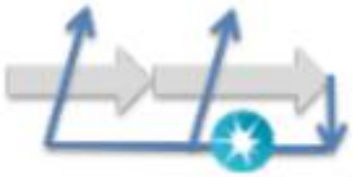


(figure from Rossetti et al. Applied Network Science (2019) 4:52)

Can the OPU help for GPT-like computing ?

A new paradigm for AI training

BEYOND BACKPROPAGATION: A NEW DISTRIBUTED TRAINING PARADIGM



NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2020

Direct Feedback Alignment Scales to Modern Deep Learning Tasks and Architectures

Julien Launay[1,2] Iacopo Poli[1] François Boniface[1] Florent Krzakala[1,2,3]
[1]LightOn [2]LPENS, École Normale Supérieure [3] IdePhics, EPFL

- **Architecture agnostic**: scales to modern deep learning architectures
  neural view synthesis, NLP, recommender systems ...
- First **optical training** demonstrated on **graph neural networks**
  Oral presentation at NeurIPS 2020 “Beyond backprop” workshop
- Inference on silicon → model portability
- Currently restricted to “small” models on LightOn’s infrastructure
  → how to scale up ?

At NeurIPS 2020, researchers proposed faster, more efficient alternatives to backpropagation

Kyle Wiggers
@Kyle_L_Wiggers
December 16, 2020 11:18 AM

VentureBeat, Dec 2020

Optical computing meets Supercomputing

Nov 2021: First photonic AI co-processor in a #top100 supercomputer

As a matter 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, access to LightOn Cloud ...

LightOn.ai

contact@lighton.io

@LightOnIO