*Advanced Pattern Recognition workshop, Oct 17th, 2019*

# Optical random features for large-scale machine learning

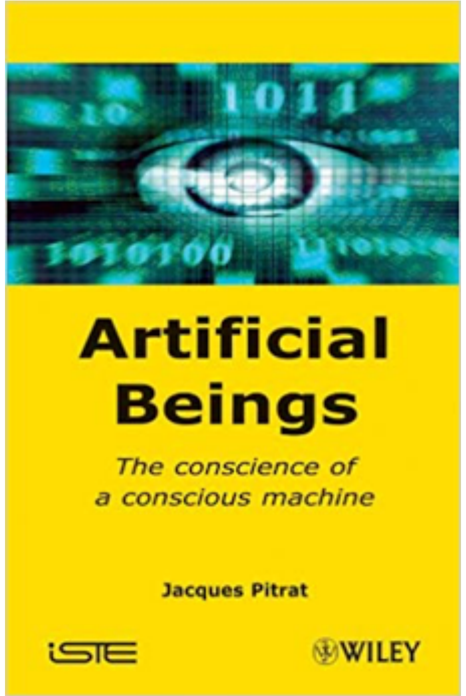

## Laurent Daudet

Paris Diderot University

Co-founder & CTO at LightOn

laurent@lighton.ai

# In memoriam Jacques Pitrat (1934-2019)




« We must not blindly imitate human behavior because computers are tools which work differently from our brain. »

LightOn

Confidential & Proprietary

# The team

**Sylvain Gigan**

LKB (UPMC / ENS / CdF)

**Florent Krzakala**

LPS (UPMC / ENS)

**Igor Carron**

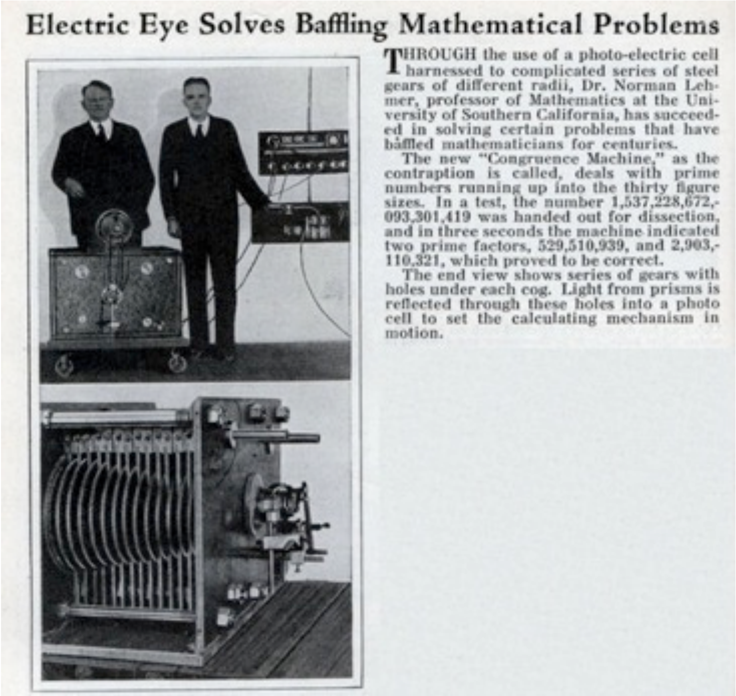
Nuit Blanche / LightOn

And *many* others in these labs and at LightOn

# A short history of Optical Processing of Information



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

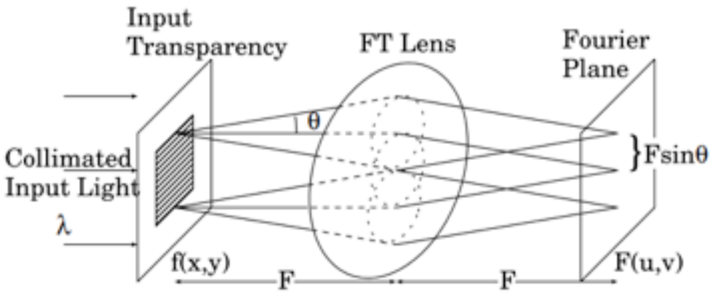


Electric Eye Solves Baffling Mathematical Problems

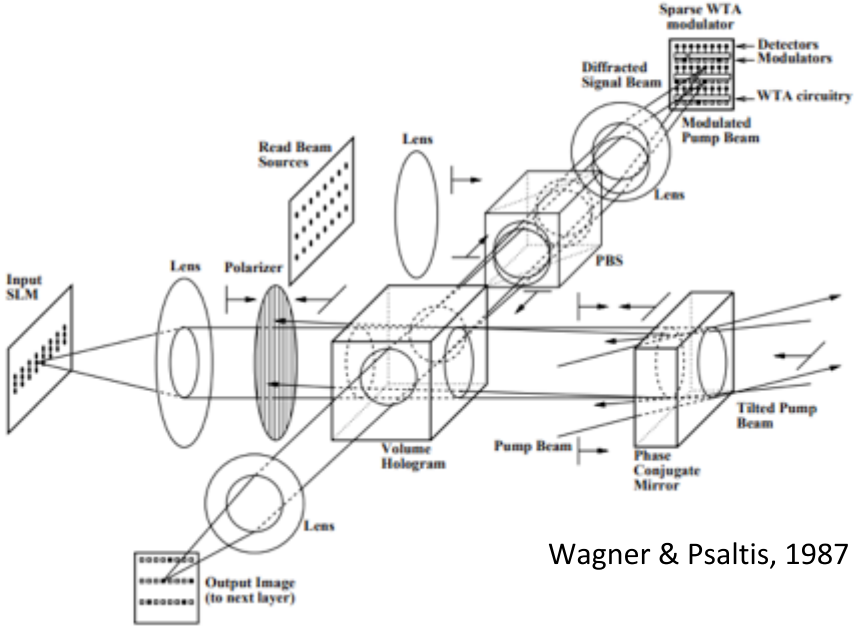
THROUGH the use of a photo-electric cell harnessed to complicated series of steel gears of different radii, Dr. Norman Lehmer, professor of Mathematics at the University of Southern California, has succeeded in solving certain problems that have baffled mathematicians for centuries.

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

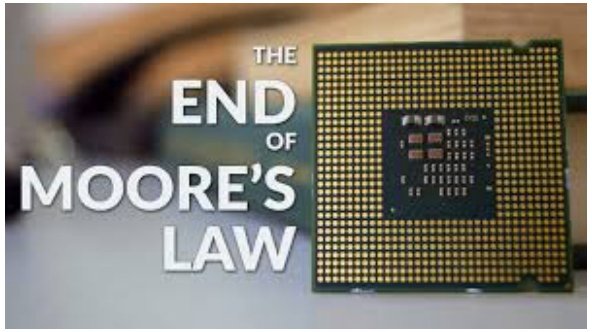
# Then Came Winter

https://imgflip.com/memegenerator/18552174/Winter-Is-Coming

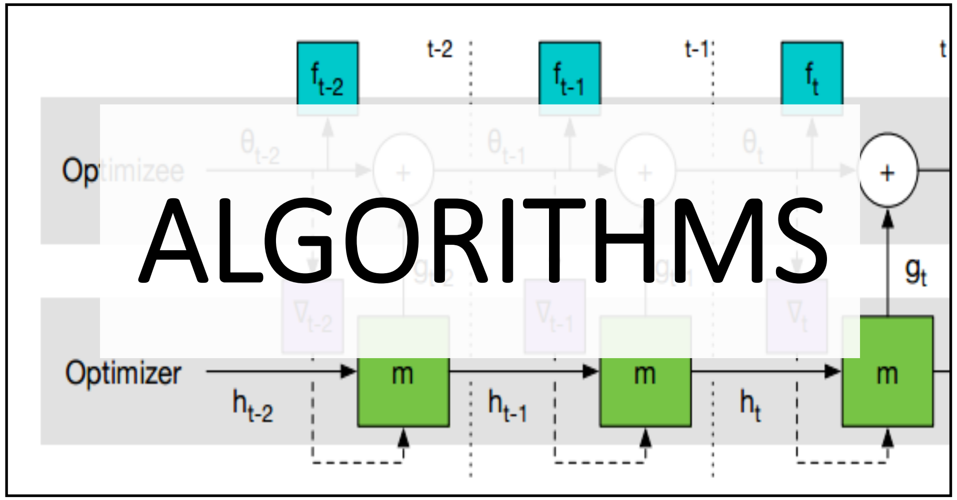
# Rebooting Optical Computing: the AI era




**SCALABLE ?**




https://www.youtube.com/watch?v=Ak7HPuuJ1Ow




**SUSTAINABLE ?**


Sport Culture Lifestyle More

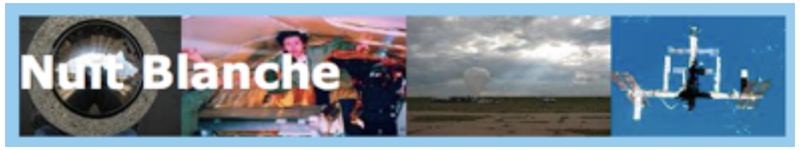
The Guardian

'Tsunami of data' could consume one fifth of global electricity by 2025

*11 Dec 2017*

Information theory
Compressive Sensing

Igor

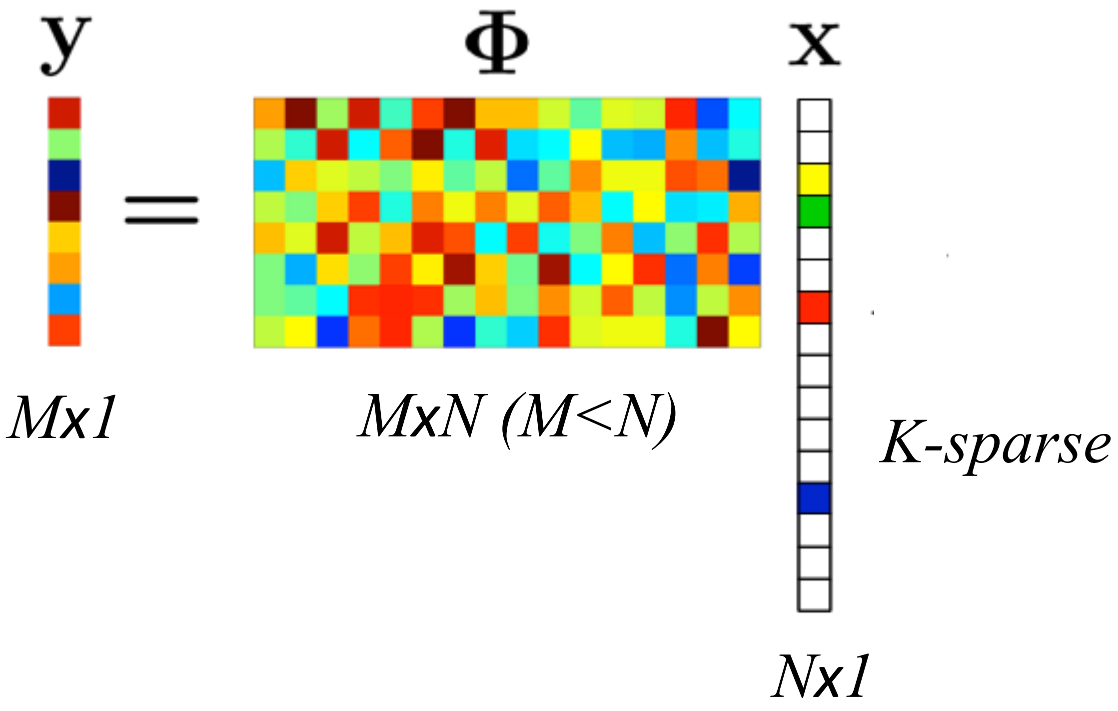



http://nuit-blanche.blogspot.com

Laurent

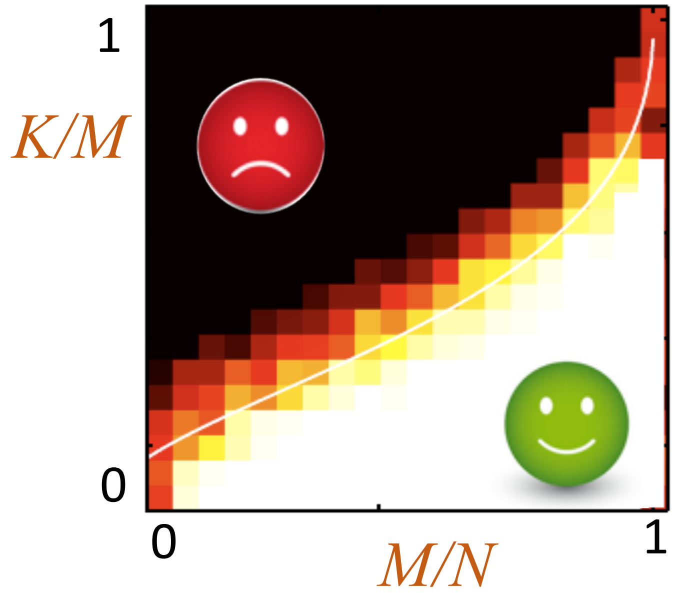
# Compressive Sensing






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
- Strong theoretical background and large empirical evidence

LightOn

Information theory
Compressive Sensing

**Igor**

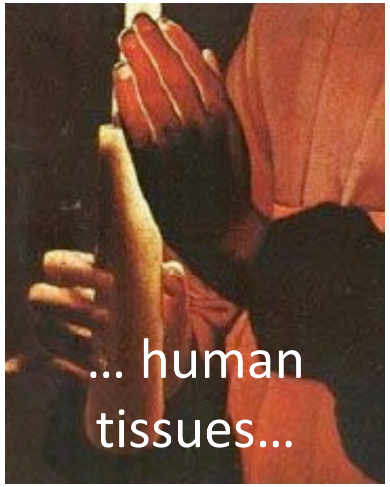
**Laurent**

Light Transport
in Diffusive
Media

**Sylvain**

# Light transport in diffusive media





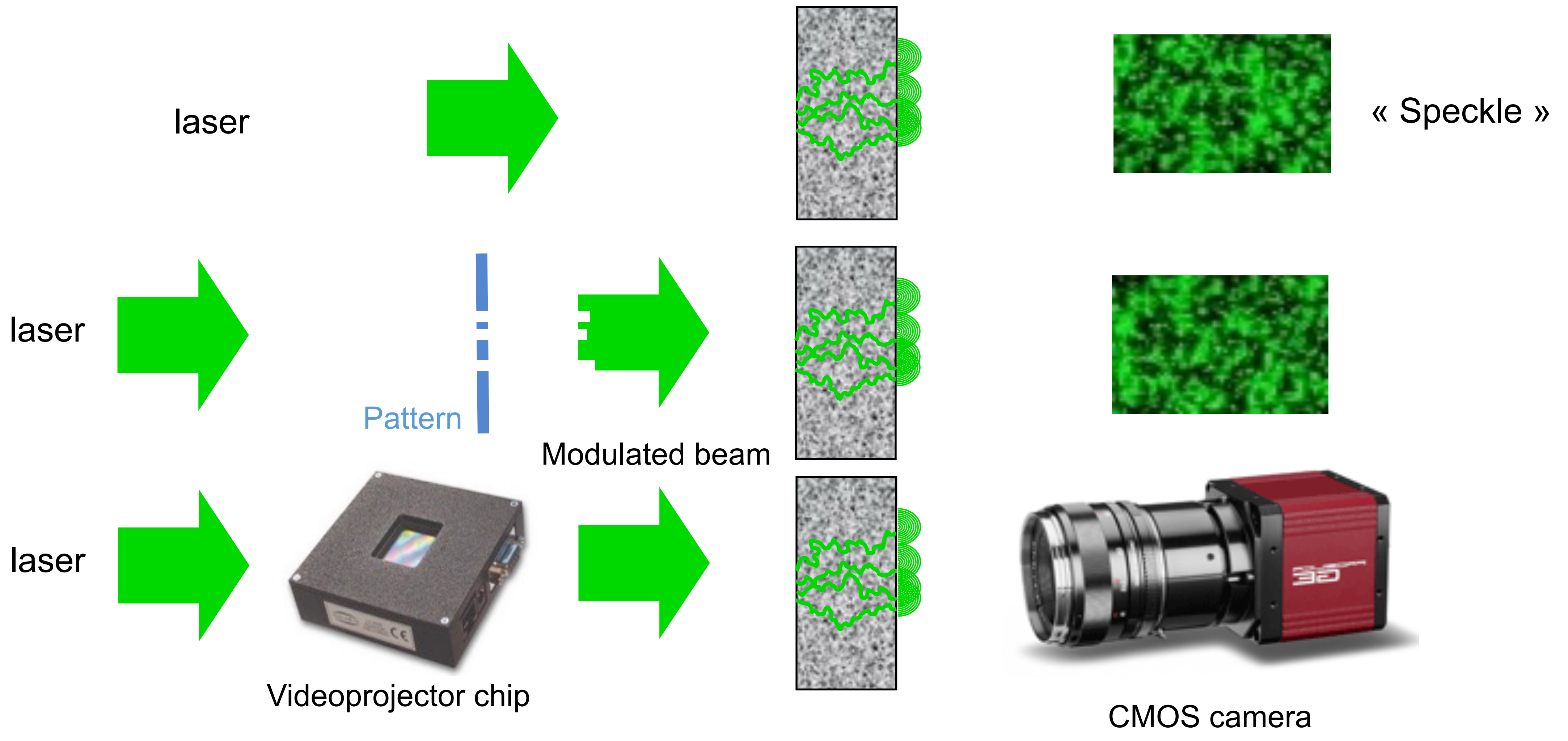

How deep can one see ?

Not much

or

?

# Scattering: a coherent process



# Scattering: a coherent process

Discrete input vector

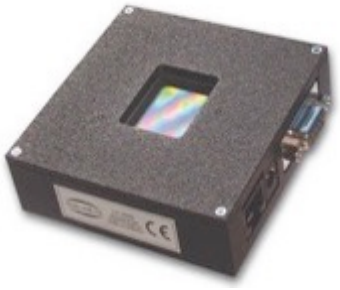
$$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N} \\ \vdots & & \ddots & \vdots \\ h_{M,1} & h_{M,2} & \cdots & h_{M,N} \end{pmatrix}$$

**Random** transmission matrix

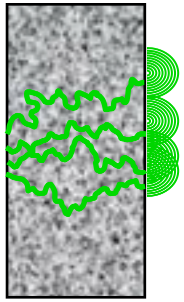
Discrete output vector (speckle intensity)

$\mathrm{x}$ $\qquad$ $\mathrm{H}$ $\qquad$ $\mathrm{y} = |\,\mathrm{H}\,\mathrm{x}\,|^2$

laser



Videoprojector chip



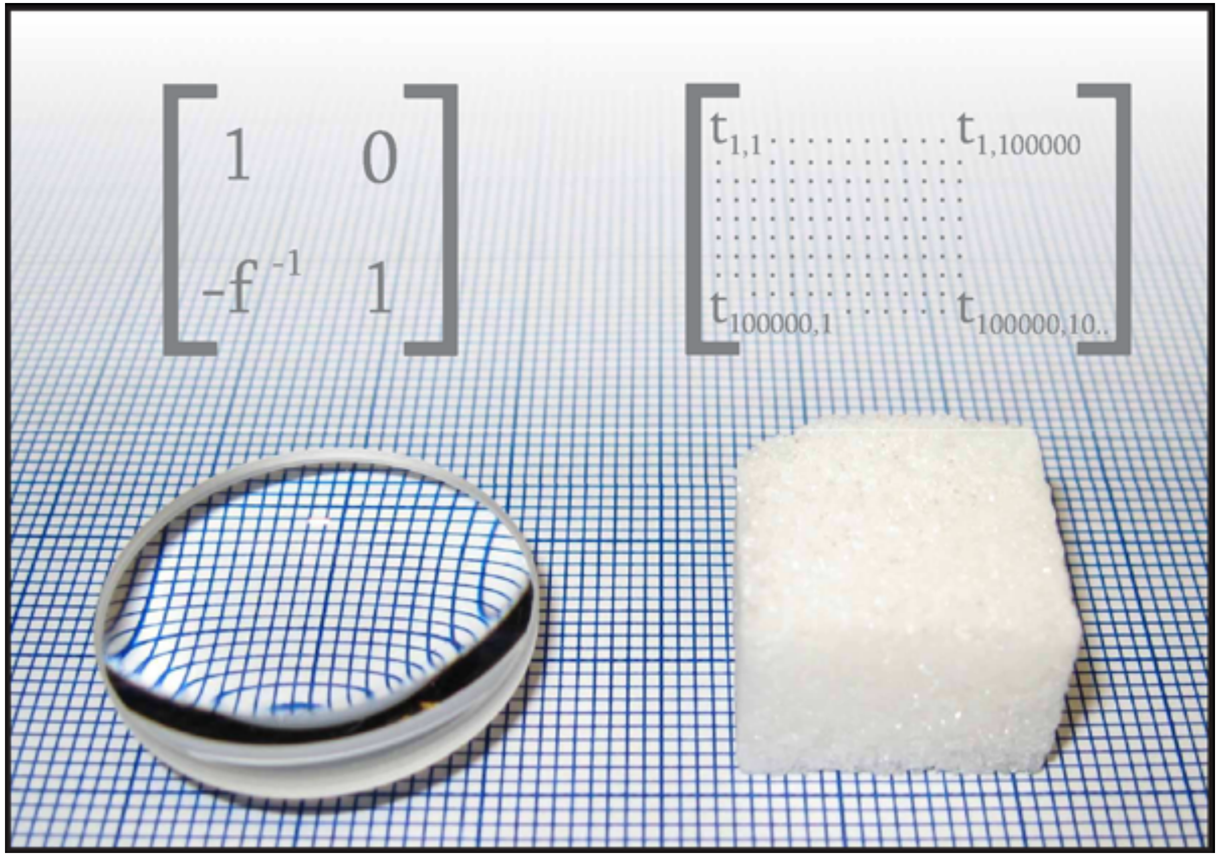
CMOS camera

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

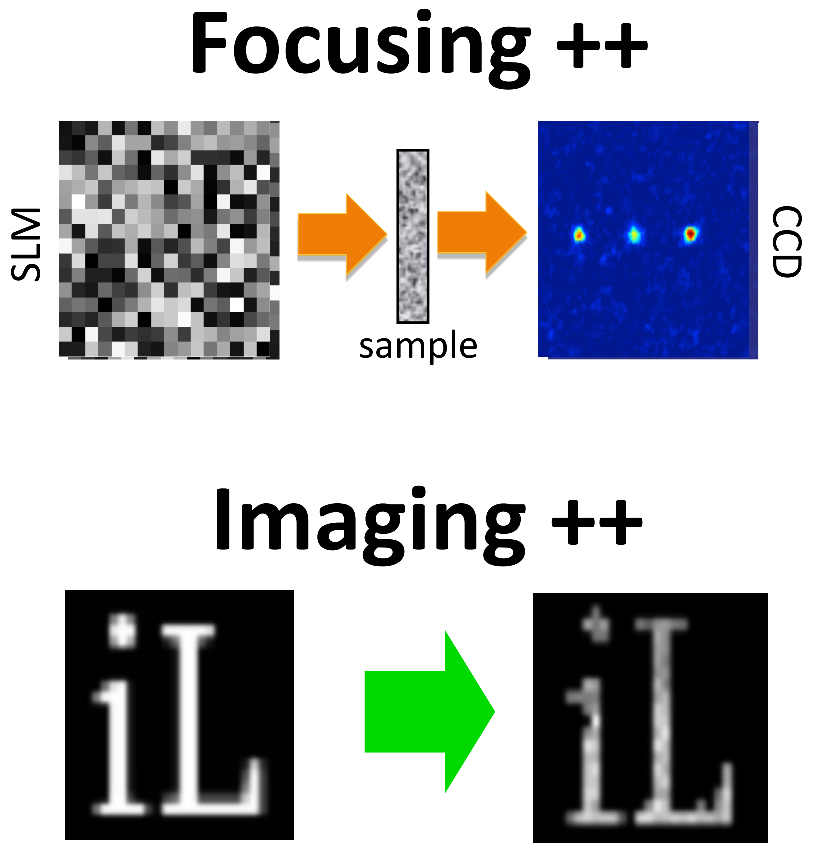
# The transmission matrix

Confidential & Proprietary

Scattering materials are « super-lenses »




Popoff et al. Phys. Rev. Lett. 104,100601 (2010)




Liutkus et al., Scientific Reports 4, 5552 (2014)

# 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: sounds familiar ?

Information theory
Compressive Sensing

**Igor**

Machine Learning

**Laurent**

Light Transport in Diffusive Media

**Sylvain**

**Florent**

LightOn

Confidential & Proprietary

# Random Projections in Machine Learning

- Random Projections act as distance-preserving point cloud embeddings

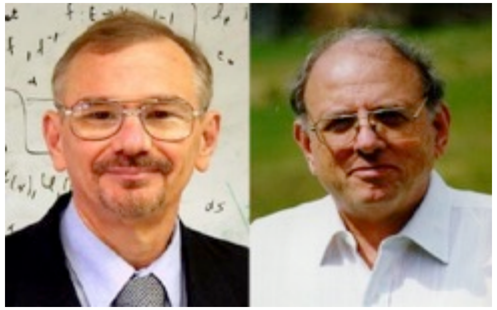
**Johnson-Lindenstrauss Lemma** (1984)

**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 $f : \Re^d \rightarrow \Re^k$ such that for all $x_i, x_j \in A$

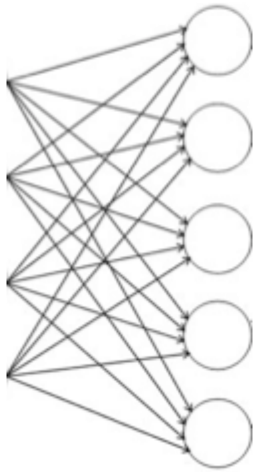
$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2$$



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

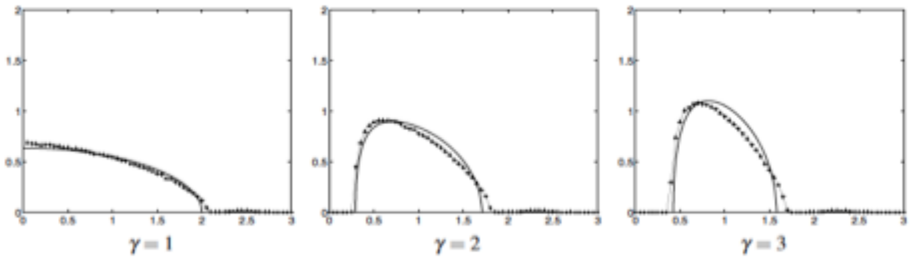
# Random Projections in Machine Learning





- A matrix-vector multiplication followed by a non-linearity: a fully connected layer of a Neural Network

- Fixed dense random weights - you can guarantee their distribution (Gaussian iid complex)




Marčenko-Pastur law on singular values

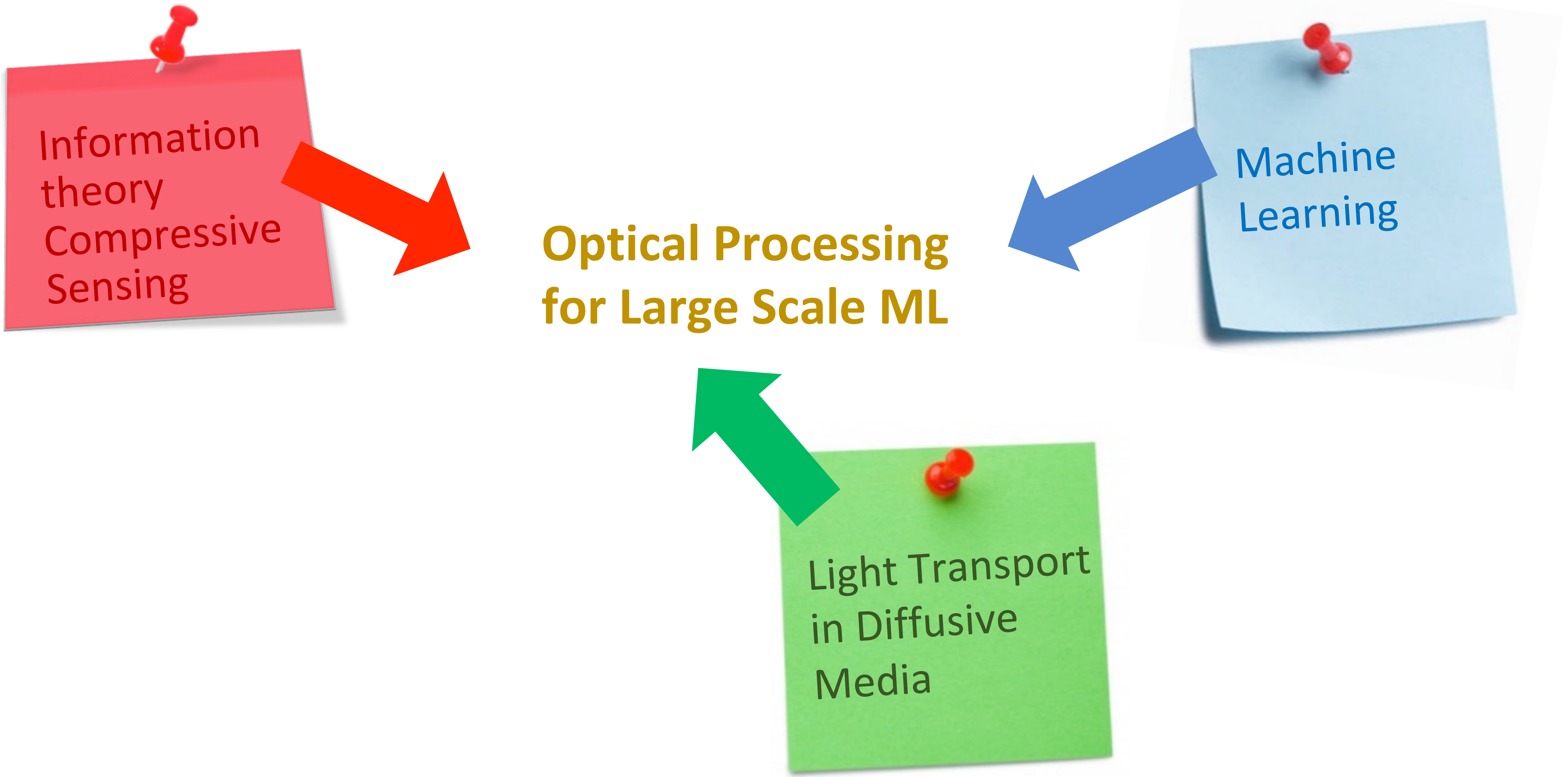
- Random projections made $O(n^2) \rightarrow O(1)$

# Lessons from Random Projections in Machine Learning

- Random projections act as dimensionality reduction or expansion
- Supervised or unsupervised
- Can also be seen as a fixed <u>dense</u> layer in a Deep Learning model

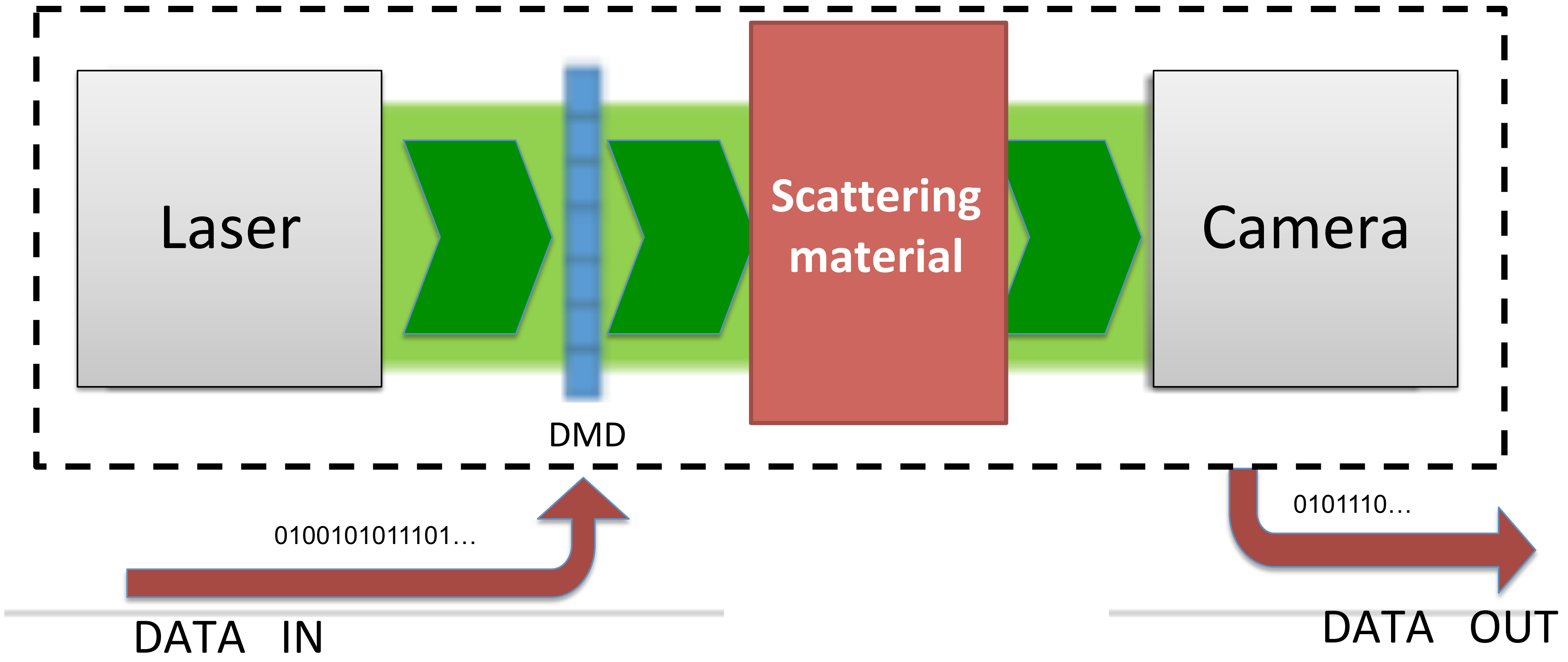
« Ask not what AI can do for Physics
– ask what Physics can do for AI »

# The Convergence



# Optical Processing for Large Scale ML






Optical Processing Unit (OPU)

# Optical Processing for Large Scale ML

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

$$y = |Hx|^2$$

with H a complex random iid matrix

EXTRA-LARGE & SUPER-FAST

H of size higher than $10^6 \times 10^6$ (TBs of memory)

kHz operation → $10^3$ such multiplies / s

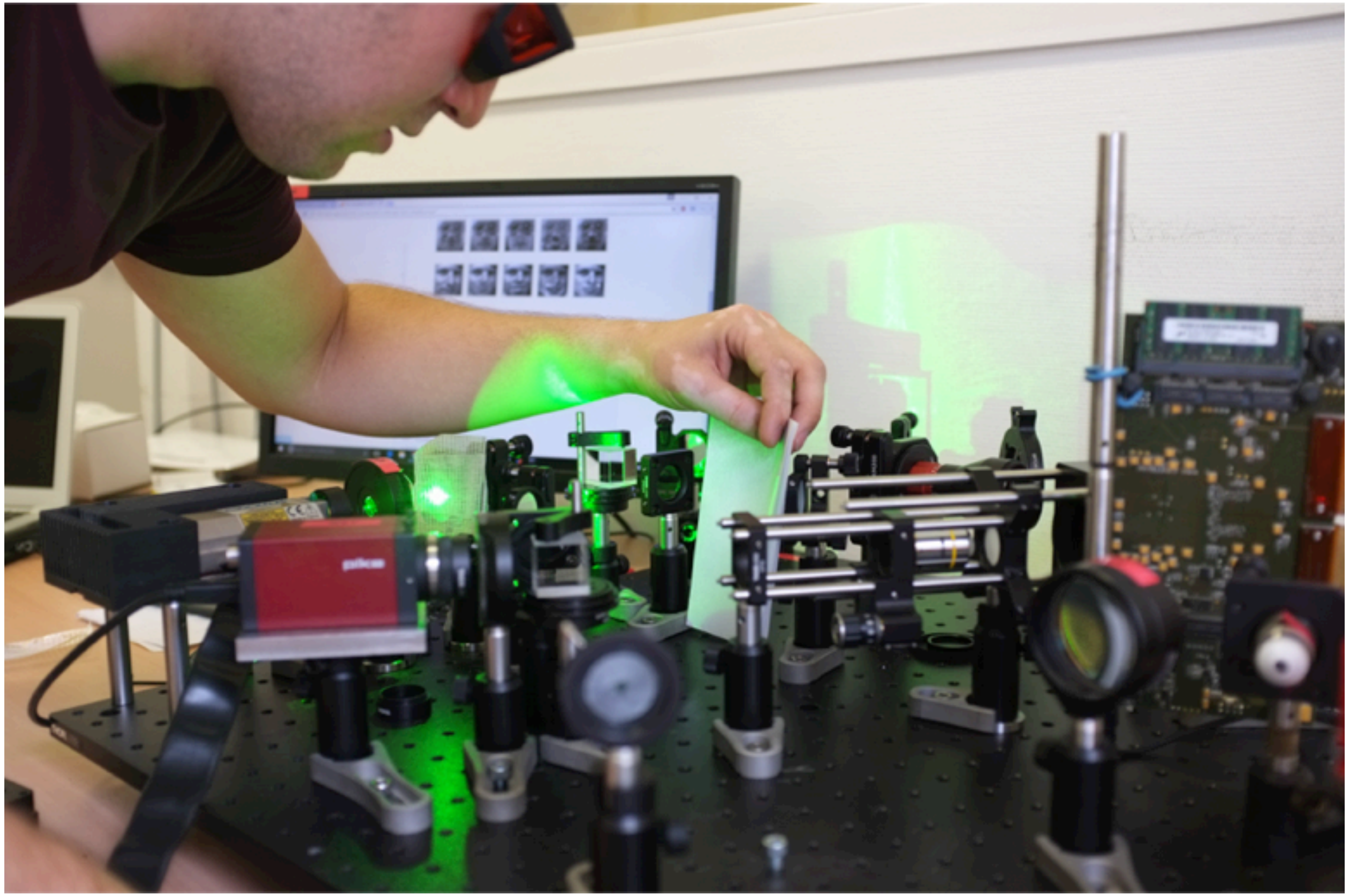
Equivalent $10^{15}$ operations* / s … for a few W

* Analog « operations » not directly comparable to flops

# Optical Processing for Large Scale ML

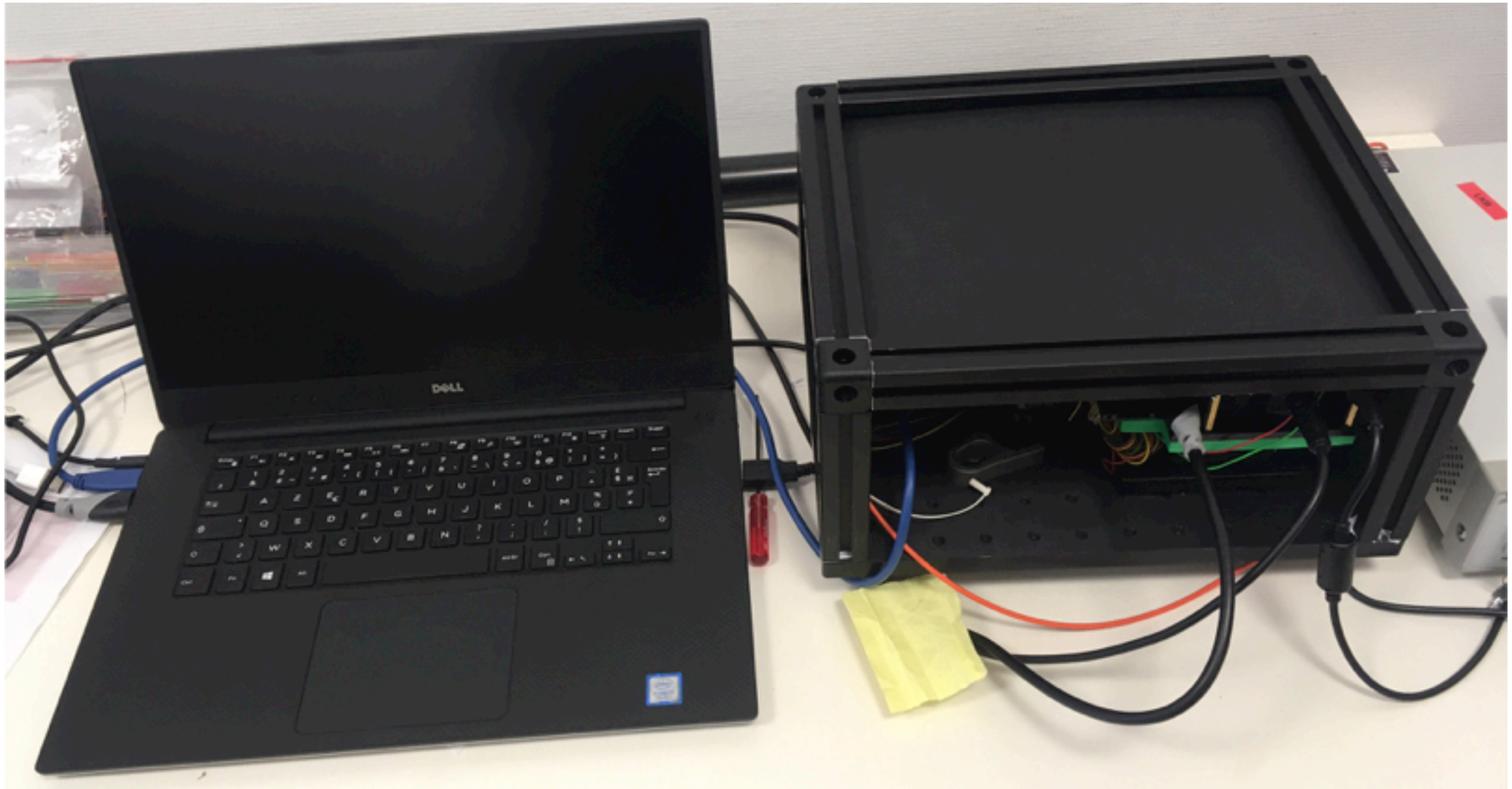
LightOn

Confidential & Proprietary



# Optical Processing for Large Scale ML

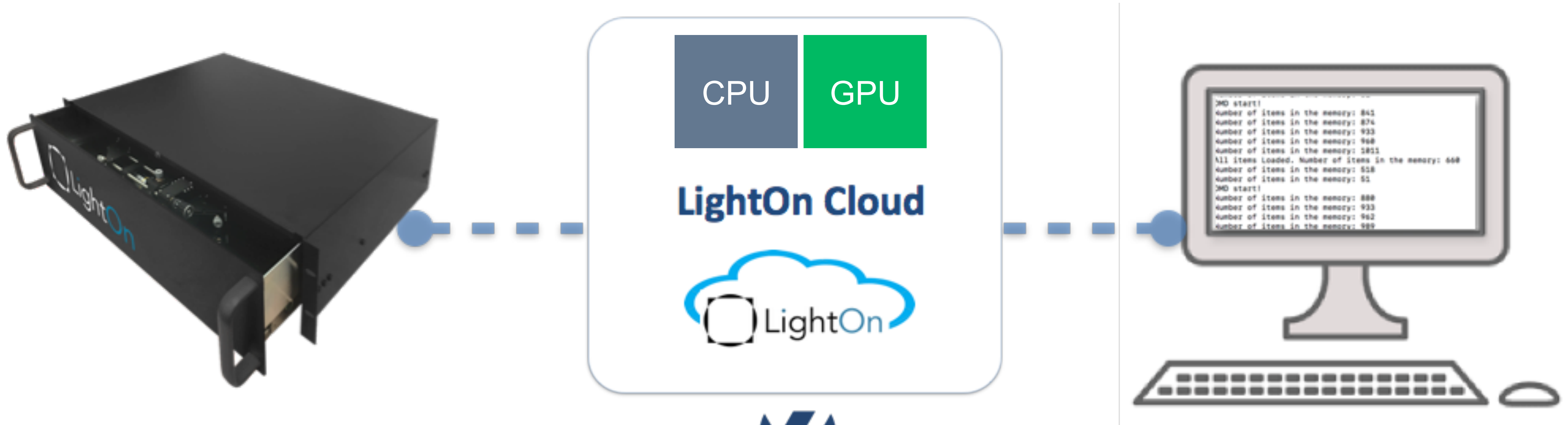





Spring 2017 - The first « OPU »: Optical Processing Unit

# Optical Processing for Large Scale ML



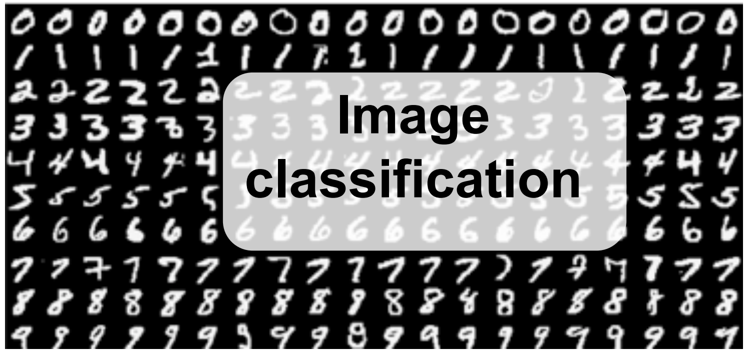



« Zeus » OPU prototype
30 W

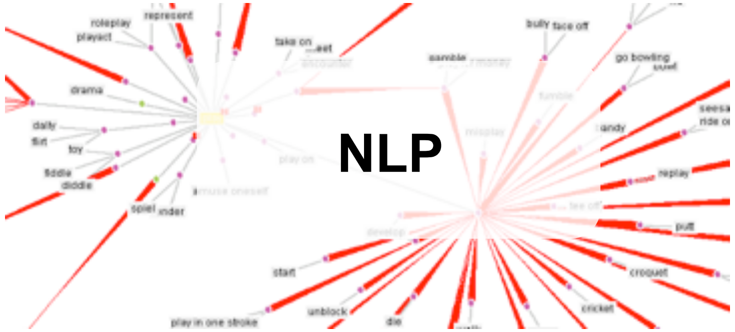
OVH.com

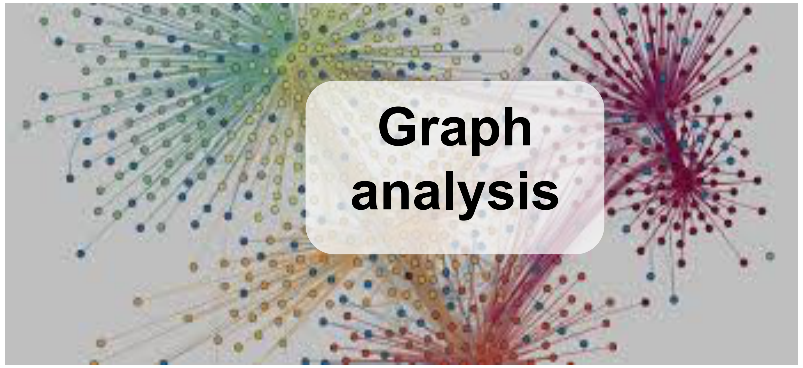
python
scikit learn
jupyter

# Some use cases









# Case study 1: classification with kernel ridge regression

training

U : data Y: labels

$$\underset{\beta \in \mathbb{R}^{p \times q}}{\operatorname{argmin}} ||\mathbf{U}\beta - \mathbf{Y}||_2^2 + \gamma ||\beta||_2^2$$

Example : classifying the MNIST database

training set of 60000 training pictures (28x28) of handwritten digits

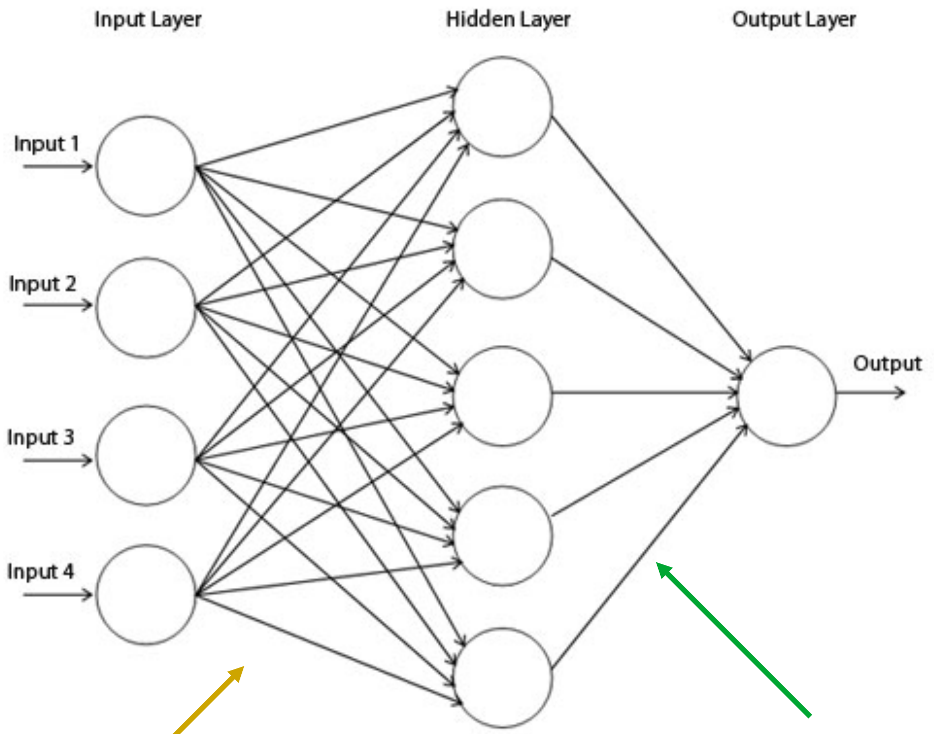
test set of 10000 digits

# Case study 1: classification with kernel ridge regression



## Using random features

[in the spirit of Rahimi-Recht]




Fully connected layer with fixed random weights

Hidden layer with ***n*** random features

Ridge regression on output

# Case study 1: classification with kernel ridge regression

## Kernel ridge regression

As n → ∞, inner products tend towards a kernel that can be computed explicitly

$$k(\mathbf{U}_i, \mathbf{U}_j) = \sqrt{\mathbf{U}_i^T\, \mathbf{U}_i\, \mathbf{U}_j^T\, \mathbf{U}_j}\, \{2\, \mathcal{E}_E[\cos^2\theta] - \sin^2\theta\, \mathcal{E}_K[\cos^2\theta]\}$$

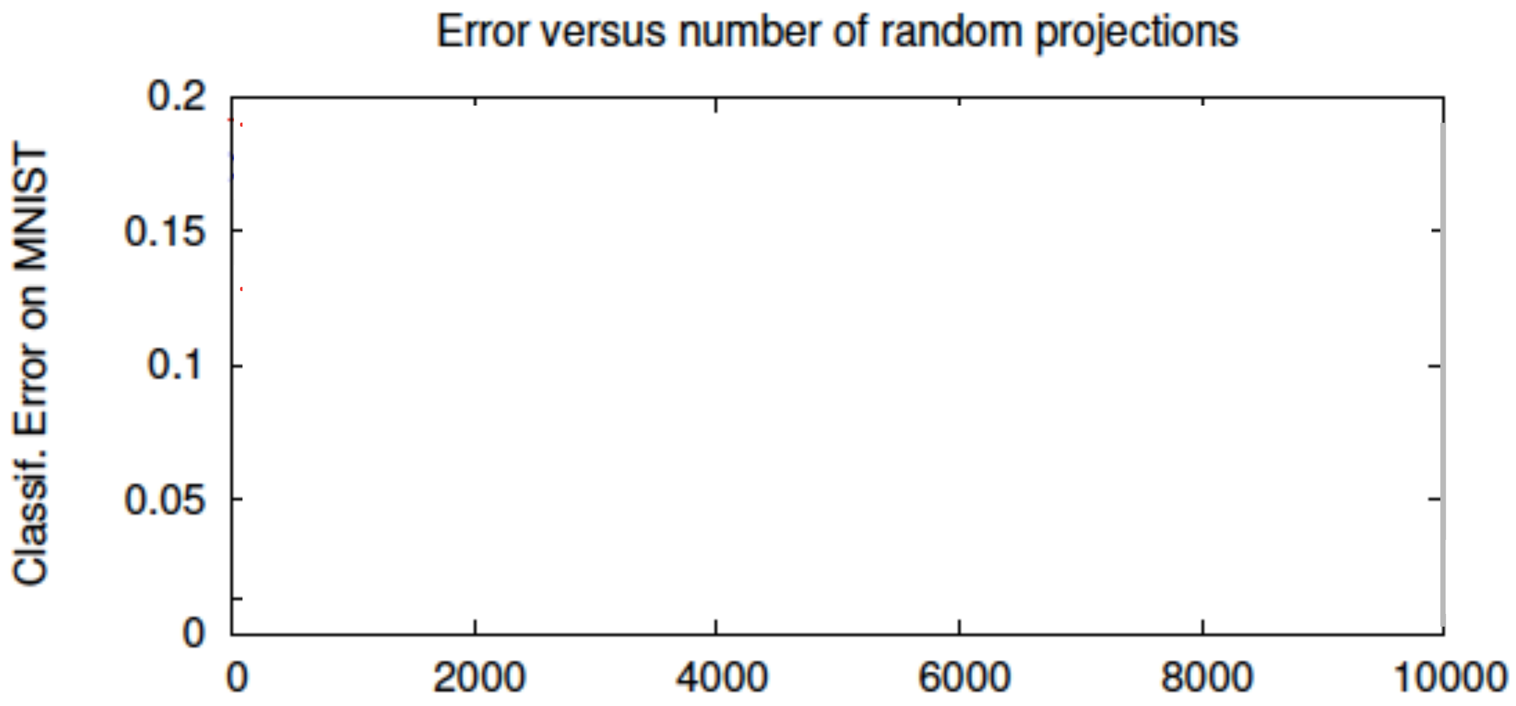
$\mathcal{E}_K[.]$ and $\mathcal{E}_E[.]$ are the complete elliptic integrals of the first / second kind

$\theta$ is the angle between $\mathrm{U}i$ and $\mathrm{U}j$

This kernel *numerically* provides a 1.31 % error rate on MNIST

# Case study 1: classification with kernel ridge regression

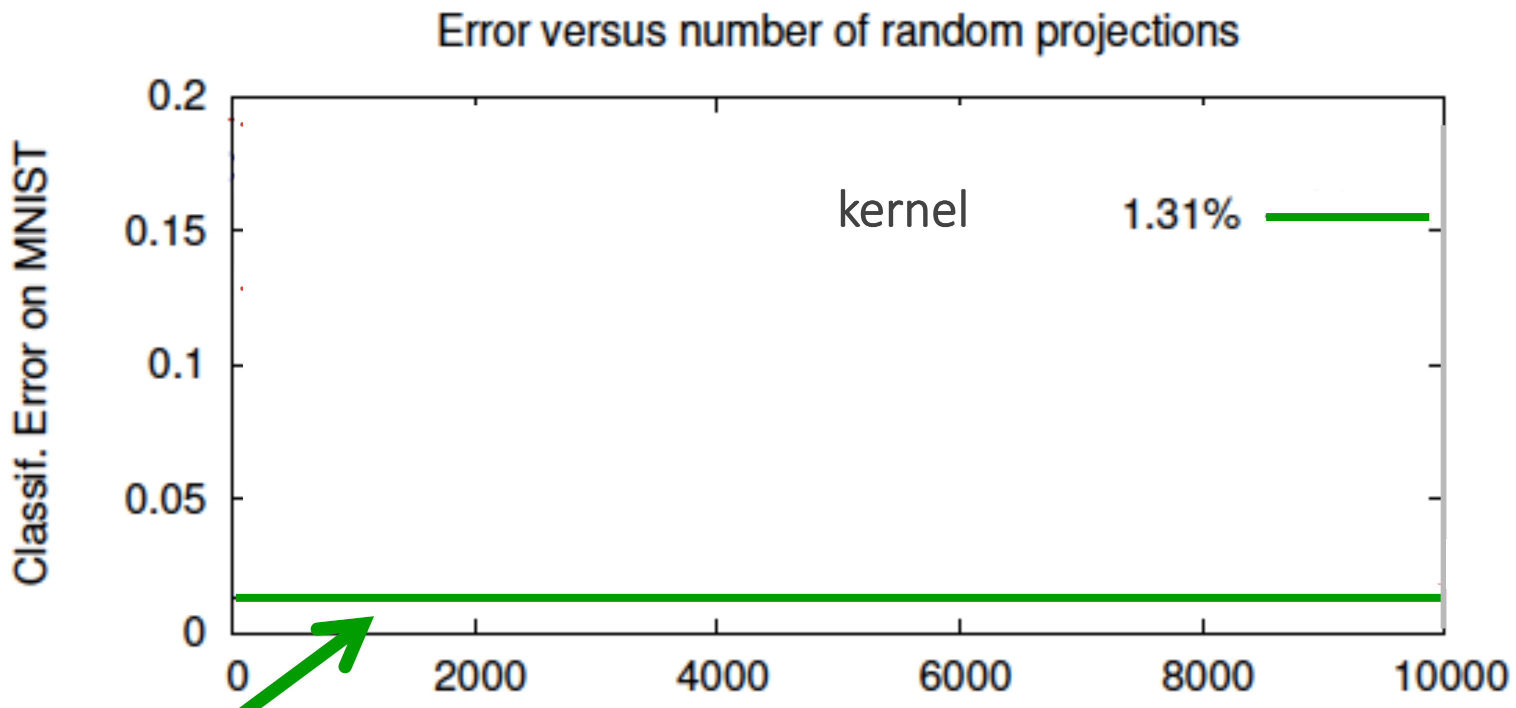
Confidential & Proprietary




n number of random features

# Case study 1: classification with kernel ridge regression

Confidential & Proprietary




n number of random features

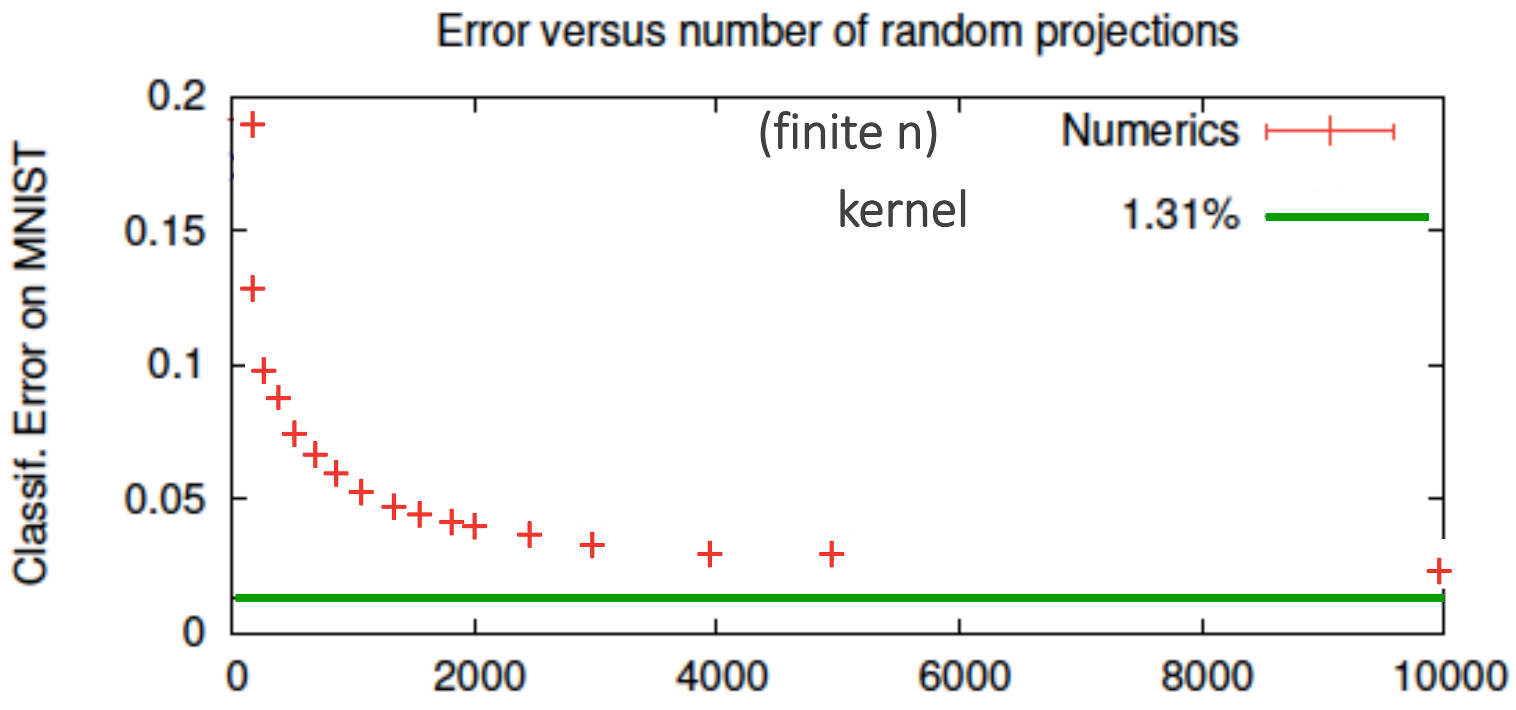
**mathematics**

kernel: asymptotic behavior as $n \to \infty$

# Case study 1: classification with kernel ridge regression

Confidential & Proprietary

**numerical simulations**




n number of random features

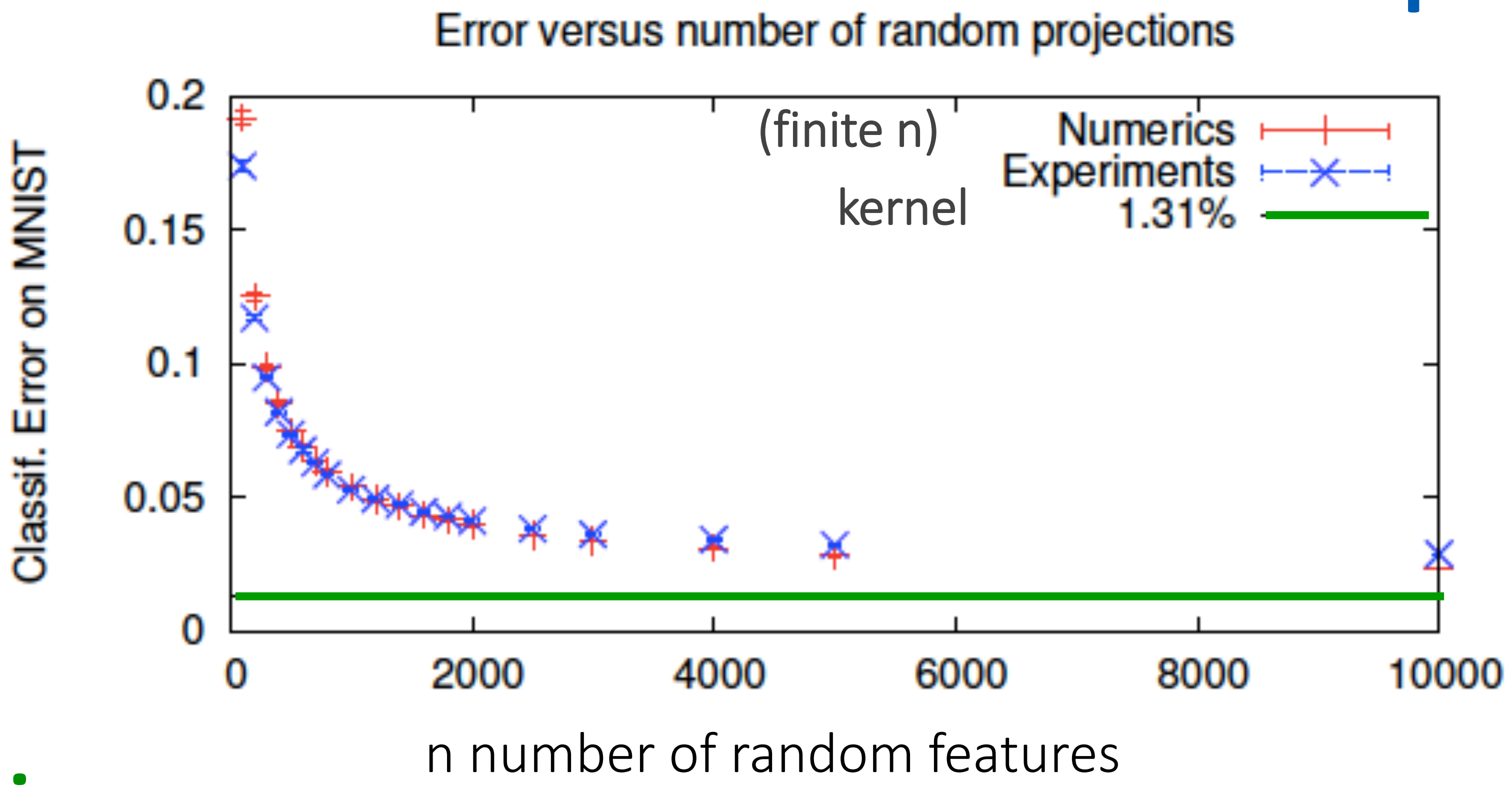
**mathematics**

# Case study 1: classification with kernel ridge regression

Confidential & Proprietary

**numerical simulations**

**optics experiment**




**mathematics**

LightOn

Confidential & Proprietary

# Biological motivation for dimensionality expansion



**Science**

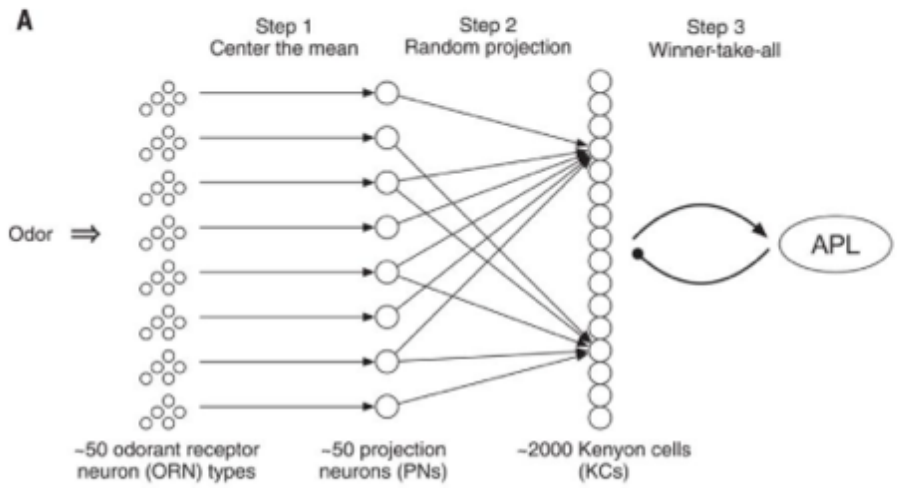
Vol 358, Issue 6364

10 November 2017


## Fly brain inspires computing algorithm

Flies use an algorithmic neuronal strategy to sense and categorize odors. Dasgupta *et al.* applied insights from the fly system to come up with a solution to a computer science problem. On the basis of the algorithm that flies use to tag an odor and categorize similar ones, the authors generated a new solution to the nearest-neighbor search problem that underlies tasks such as searching for similar images on the web.


Muhammad M. Karim, GDFL 1.2




## A neural algorithm for a fundamental computing problem


Sanjoy Dasgupta[1], Charles F. Stevens[2,3], Saket Navlakha[4,*]

+ See all authors and affiliations



*Science* 10 Nov 2017:

Vol. 358, Issue 6364, pp. 793-796

DOI: 10.1126/science.aam9868

# Case Study 2: Fast Transfer Learning

## AI systems « learn » by labelled examples

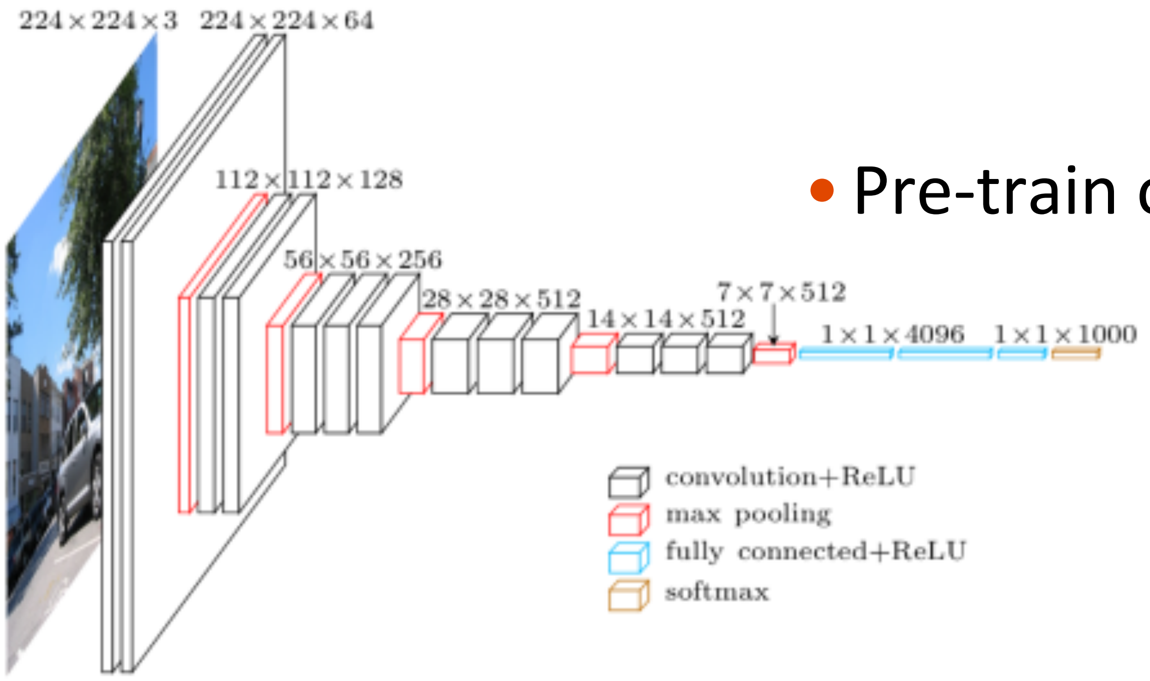
« cat »

« horse »

- It is an extremely inefficient process: huge amount of data & compute
- « Transfer learning »:
  - pre-train network on large amount of generic data → slow but done once
  - Slightly adapt network to small amount of specific data → fast

# Case Study 2: Fast Transfer Learning

- Start with a standard VGG16 [Simonyan & Zisserman '14] architecture



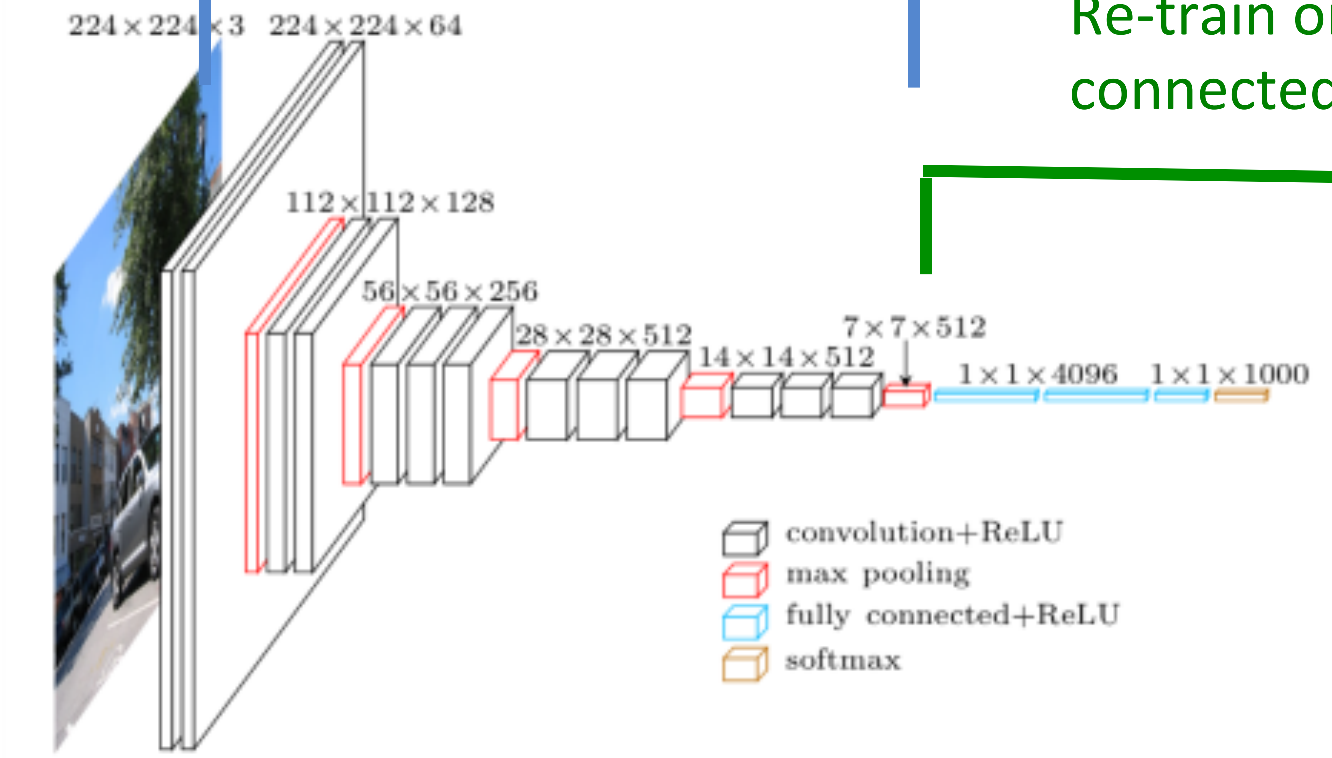

- Pre-train on generic data with a GPU

# Case Study 2: Fast Transfer Learning

- Now comes a second dataset : STL10

Keep unchanged

Re-train only fully connected layers



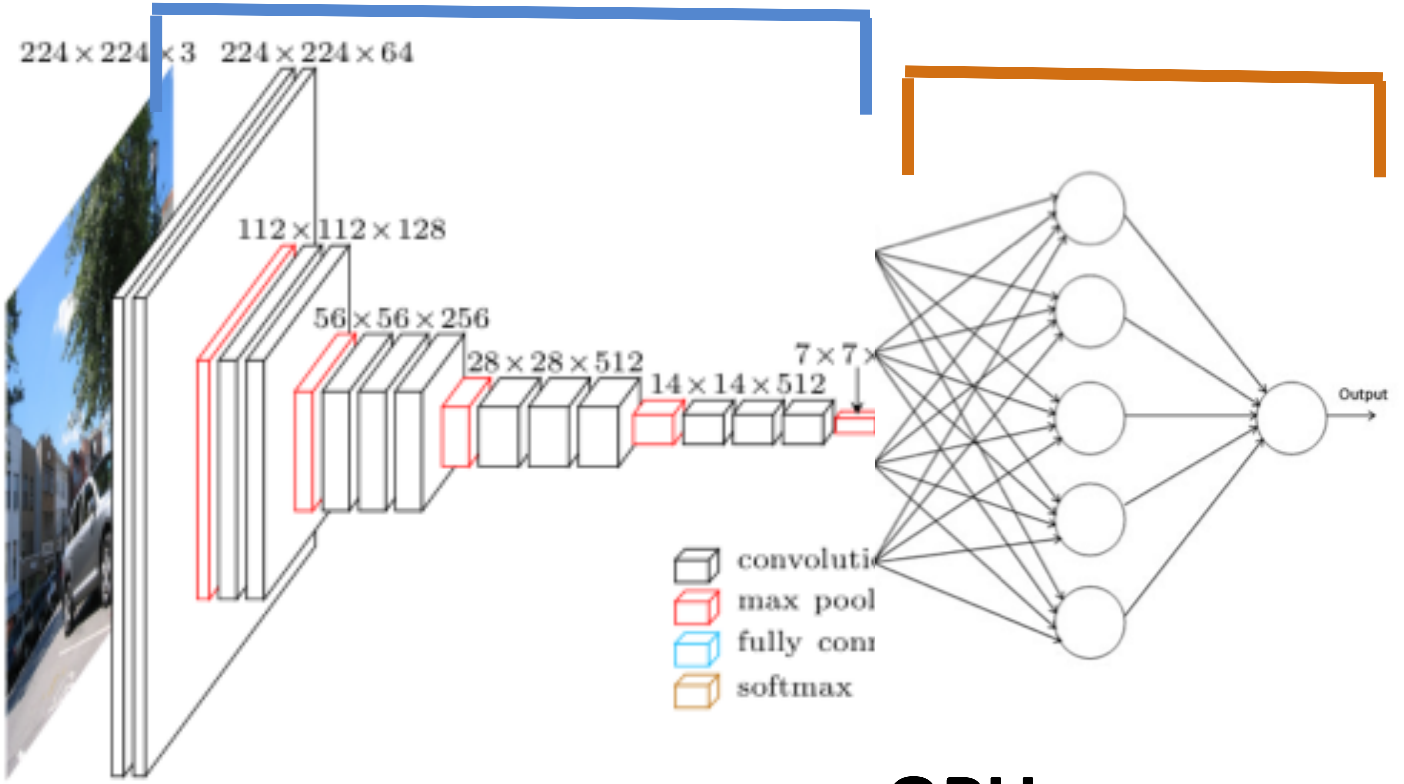

25’

# Case Study 2: Fast Transfer Learning

- Alternative approach

Keep unchanged

Dense random layer
linear reg. for output

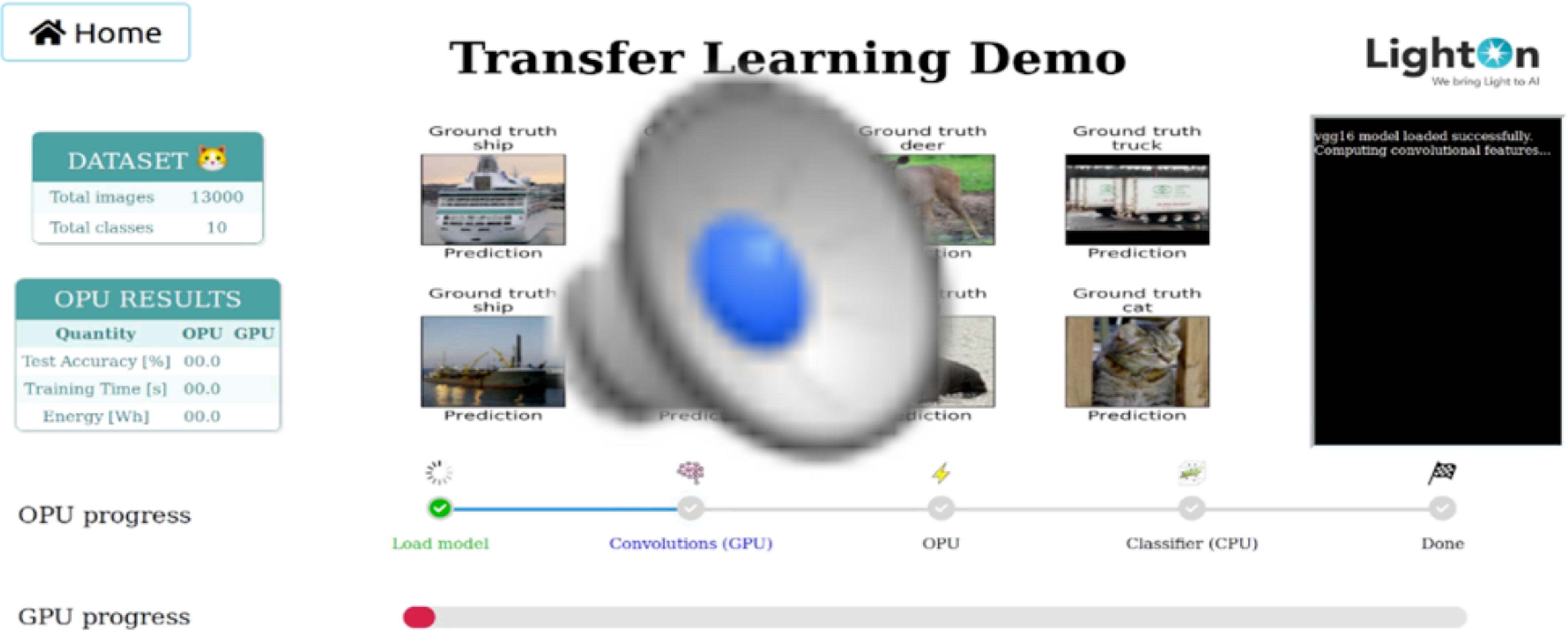



GPU — OPU — CPU

1’40’’

→ Hybrid computing

# Case Study 2: Fast Transfer Learning

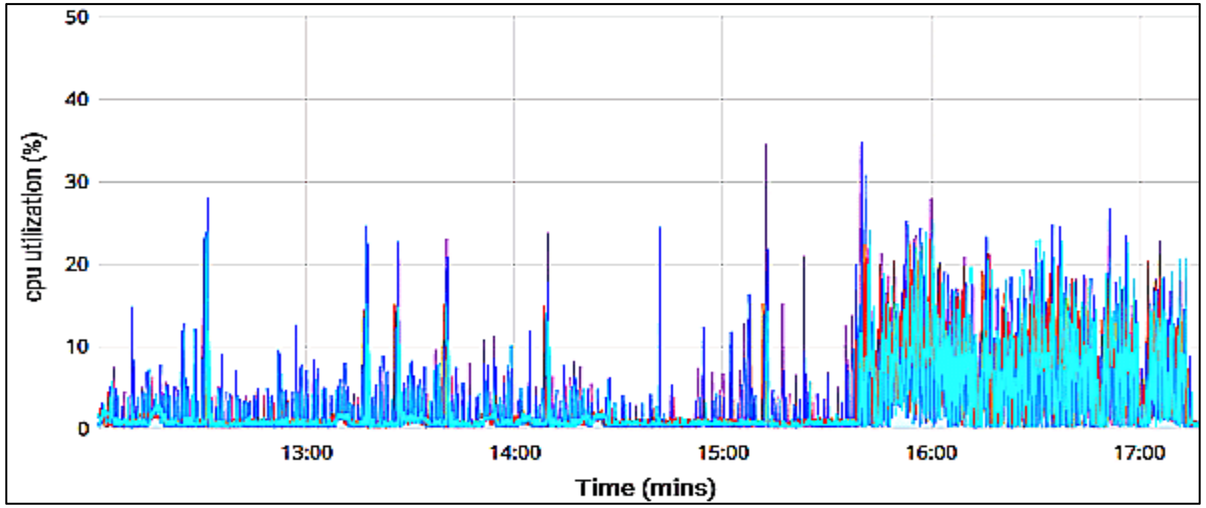



NB : x2 video playback speedup

# Case Study 3: Sketching the Change-point

How to detect changes / anomalies on the 1000s of signals monitoring a complex system (factory, datacenter, airplane engine…) ?

Example : monitoring a datacenter



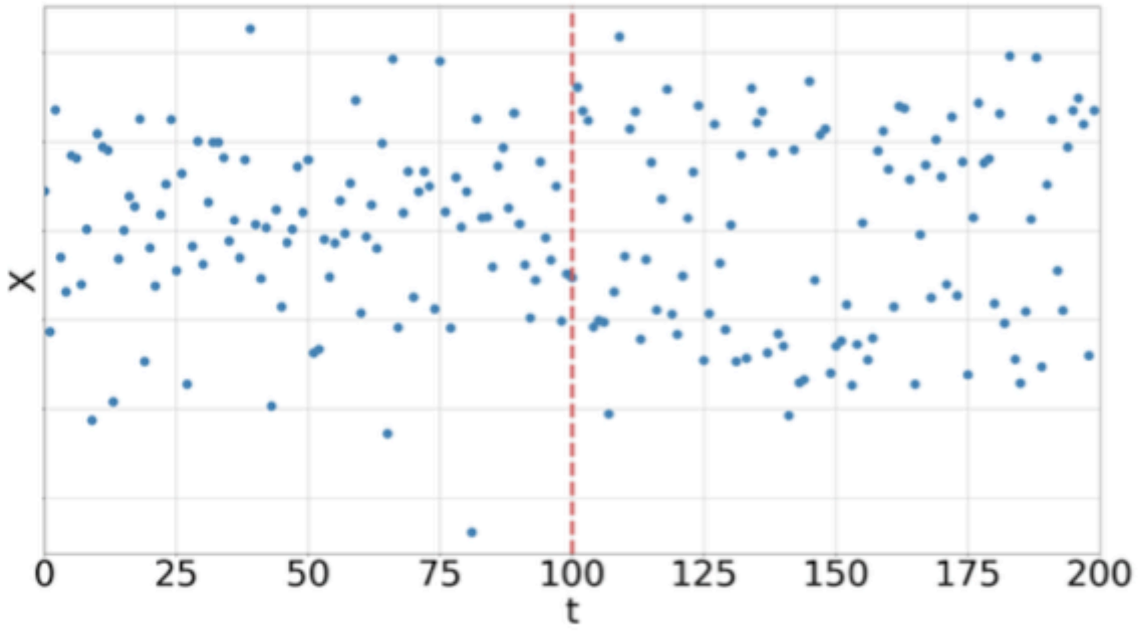

Valsamas *et al*, 2018

# Case Study 3: Sketching the Change-point



***NEWMA: a new method for scalable model-free online change-point detection*, Keriven, N., Garreau, D., Poli, I., arXiv:1805.08061**

| | |
|---|---|
| Data stream | $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ with $x_i \in \chi$ |
| Detect *abrupt* change in | $\mathbb{E}\left[\Psi\left(\mathbf{x}\right)\right]$ with $\Psi : \chi \rightarrow \mathcal{H}$ |

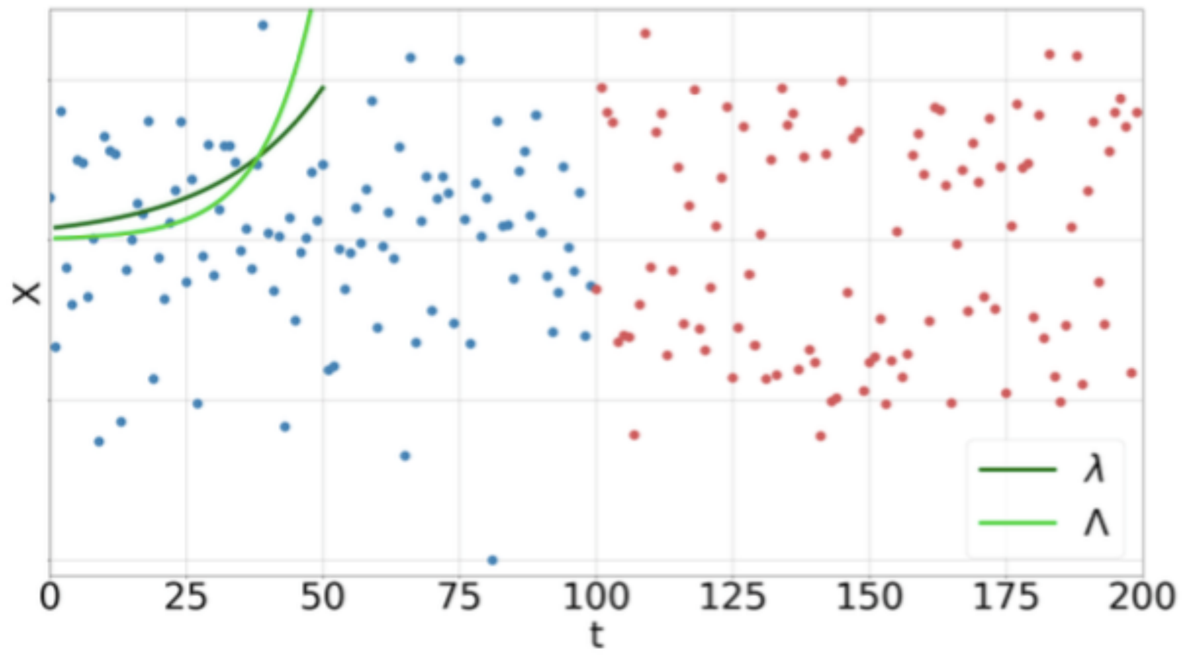


# Case Study 3: Sketching the Change-point



**NEWMA equations:**

$$\mathbf{z}_1^t = (1-\lambda)\,\mathbf{z}_1^{t-1} + \lambda\Psi\left(\mathbf{x}^t\right)$$
$$\mathbf{z}_2^t = (1-\Lambda)\,\mathbf{z}_2^{t-1} + \Lambda\Psi\left(\mathbf{x}^t\right)$$
$$0 < \lambda < \Lambda < 1$$



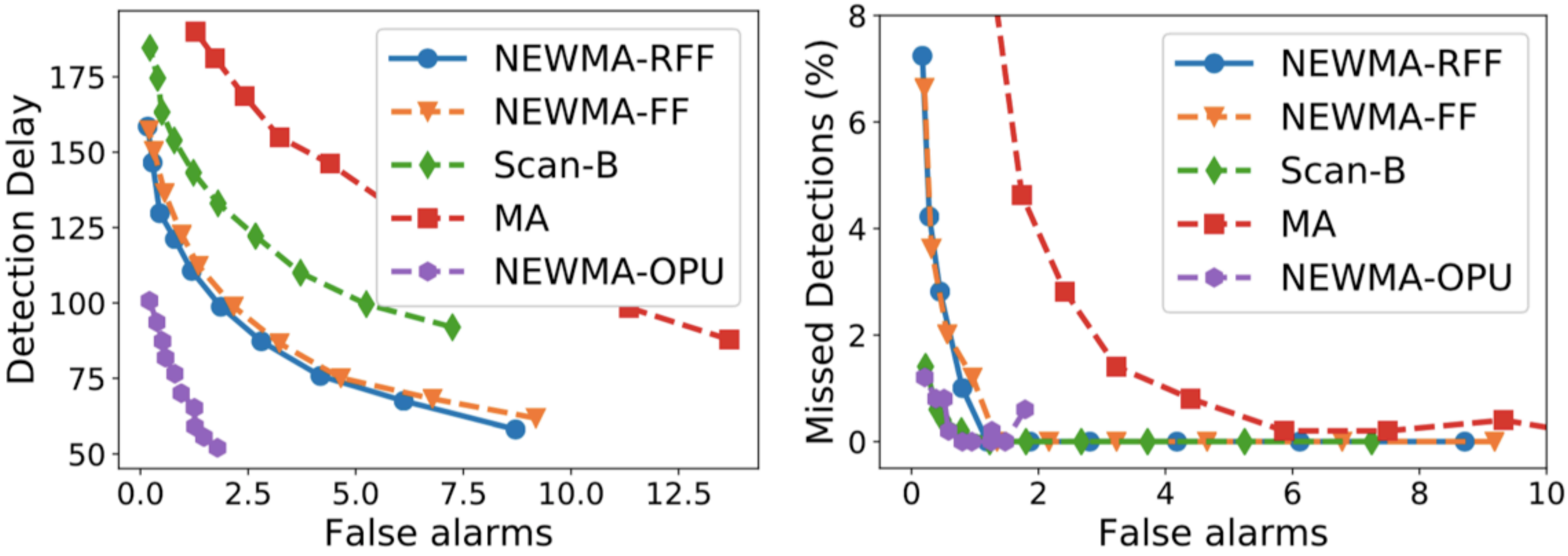

- Compute « sketches » with RPs
- Bounds can be derived
- No need to store samples in memory

# Case Study 3: Sketching the Change-point

Confidential & Proprietary

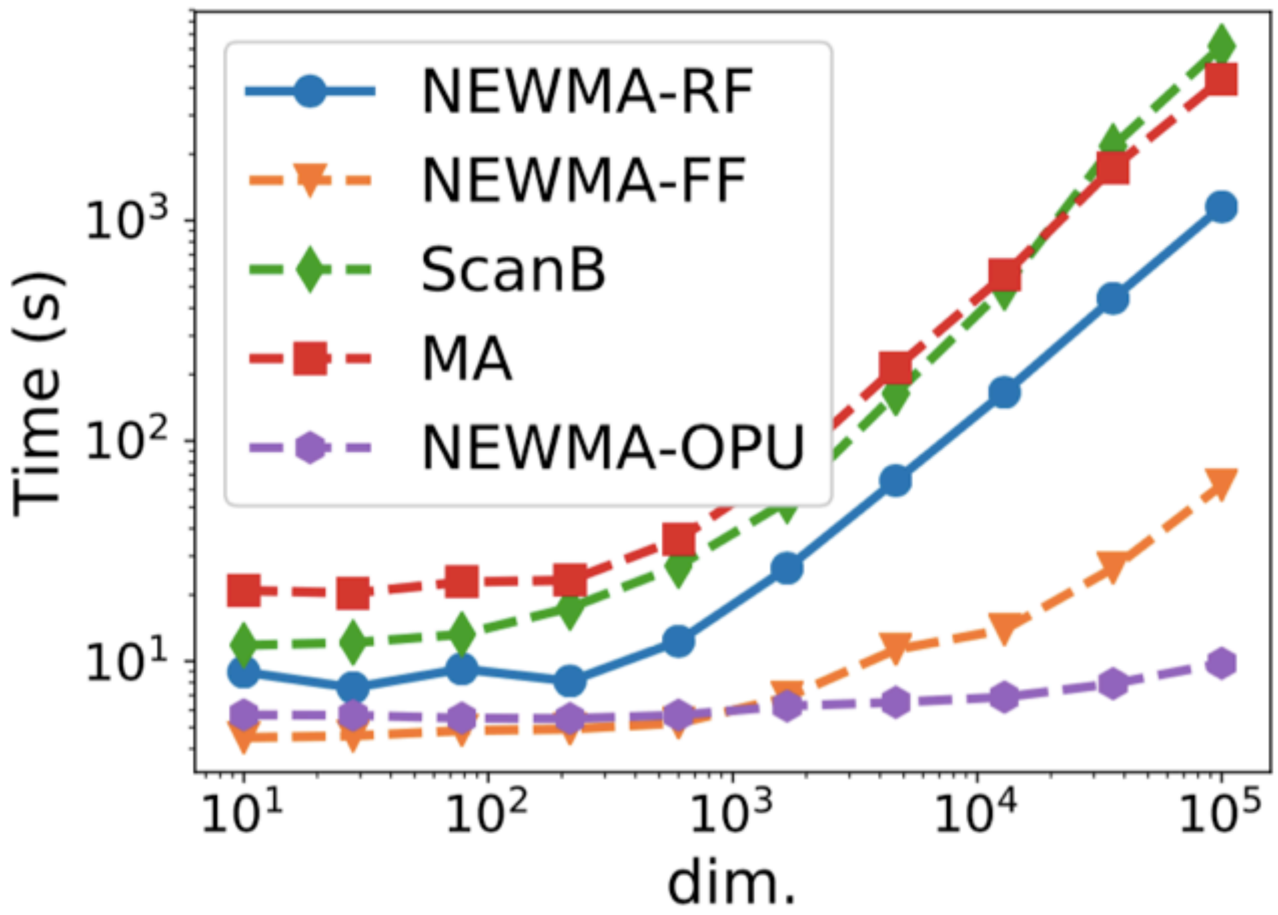
# Case Study 3: Sketching the Change-point





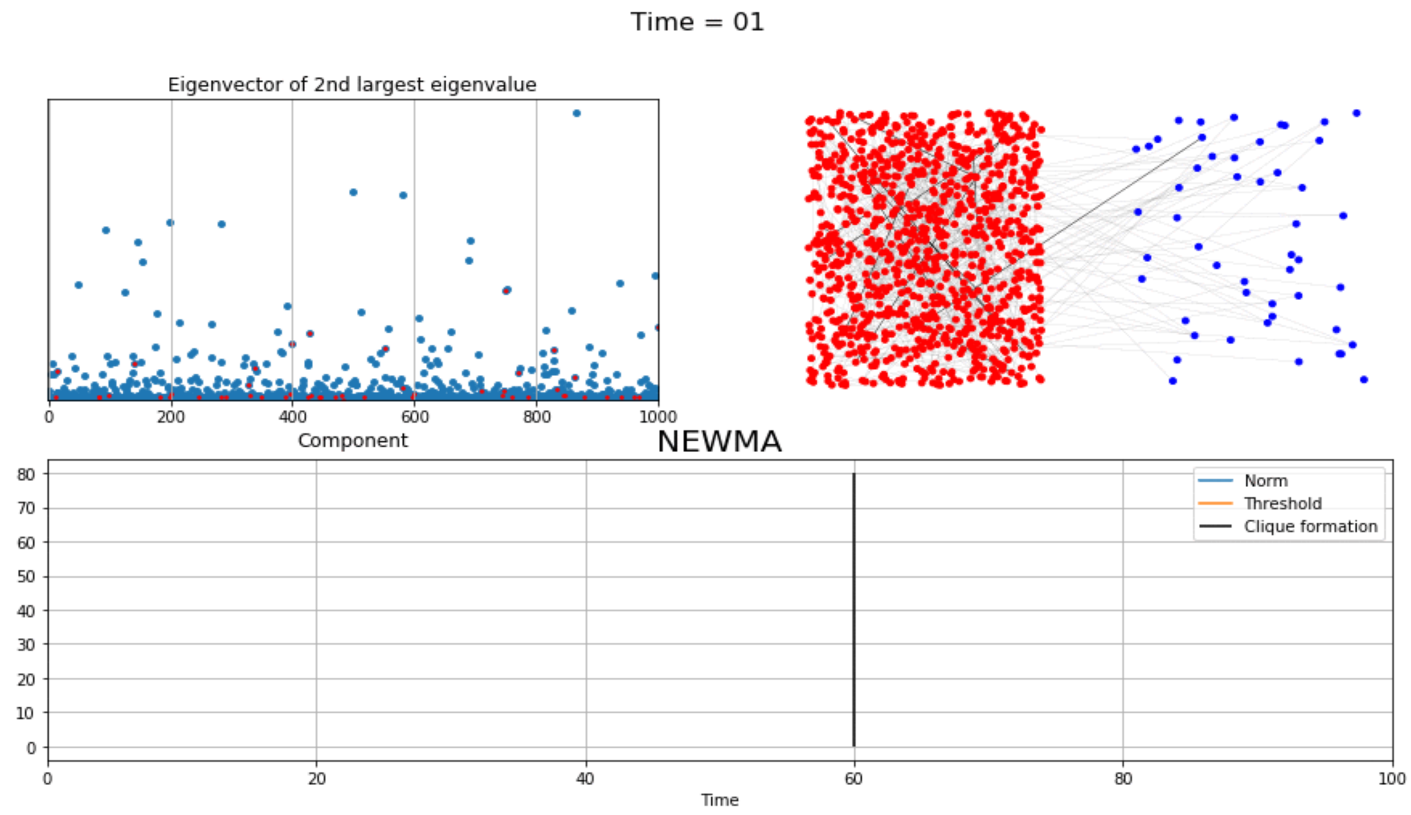
# Case Study 3: Sketching the Change-point





# Case Study 3: Sketching the Change-point

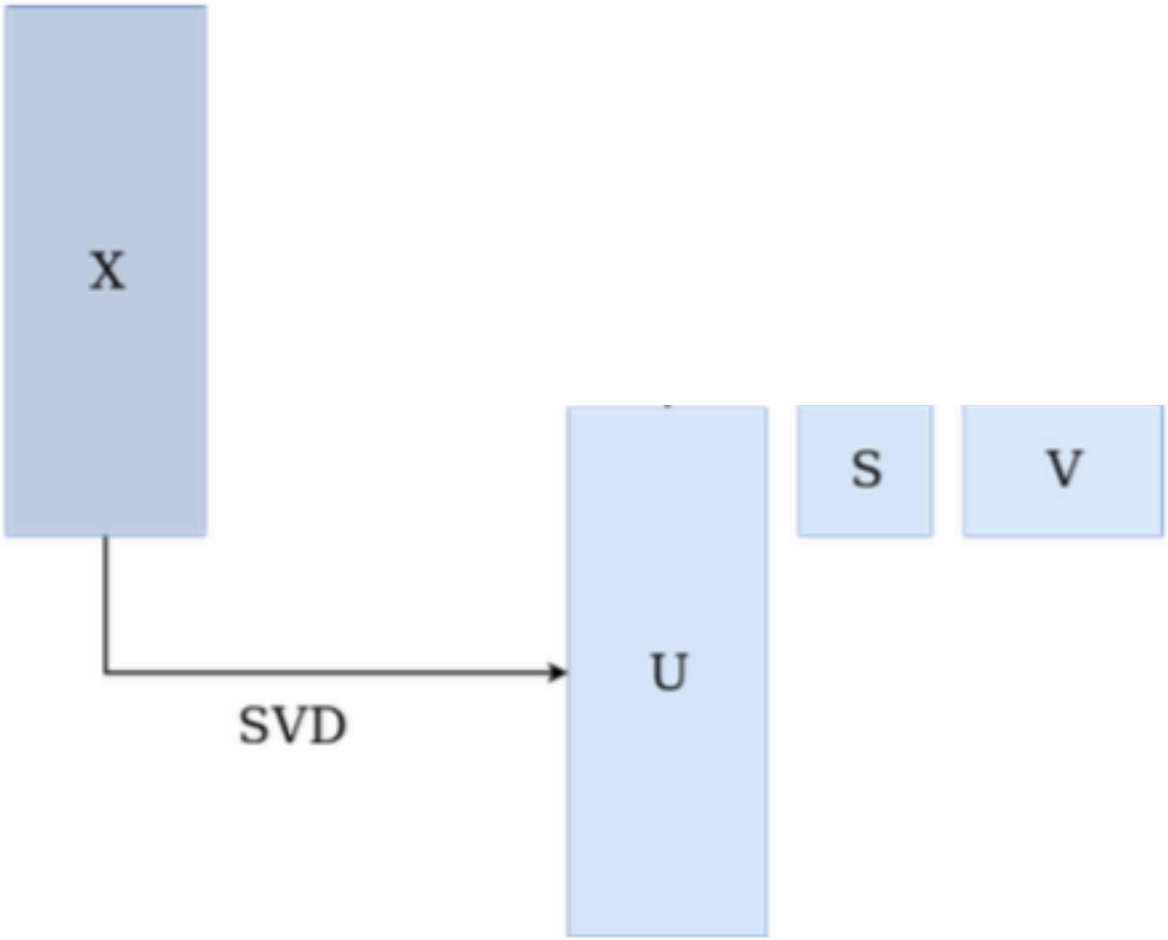
## Clique detection in graphs




Credits : L. Tommasone

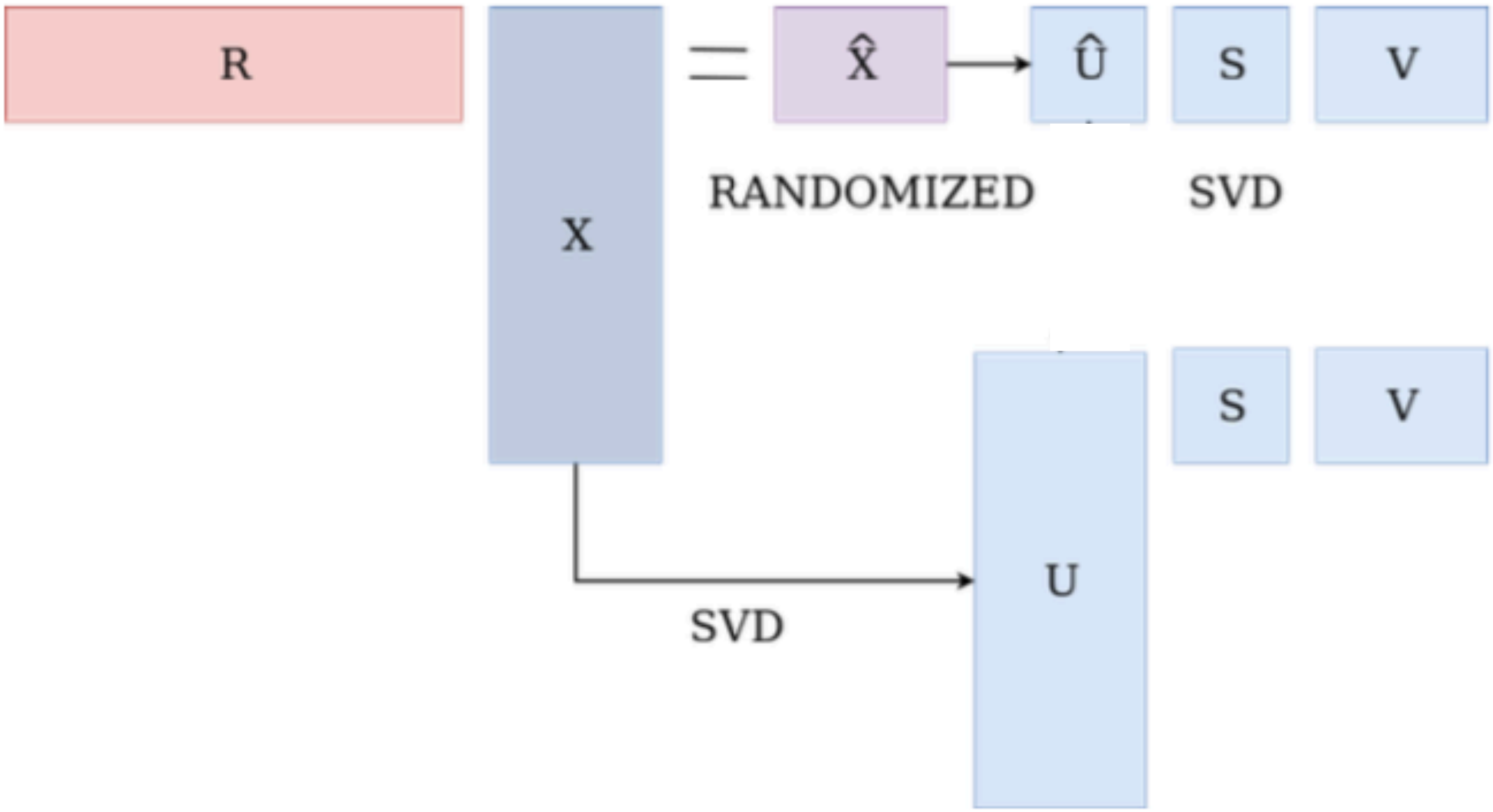
# Case Study 4: Randomized SVD





# Case Study 4: Randomized SVD






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

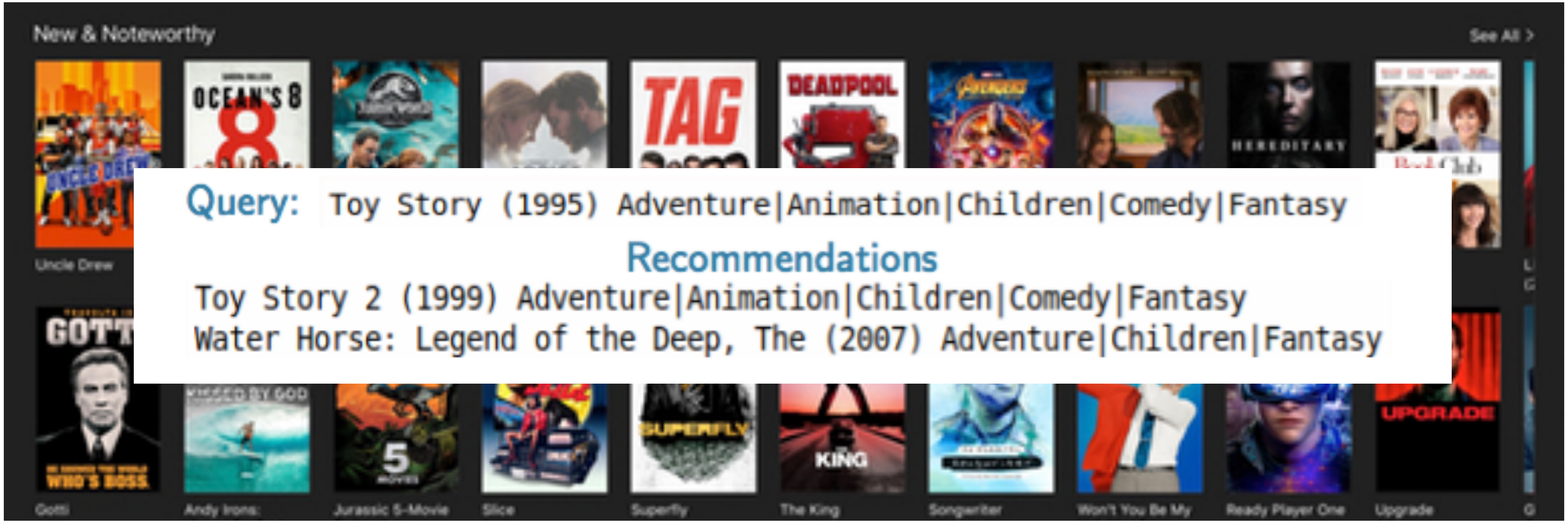
# Case Study 4: Randomized SVD

LightOn

Confidential & Proprietary

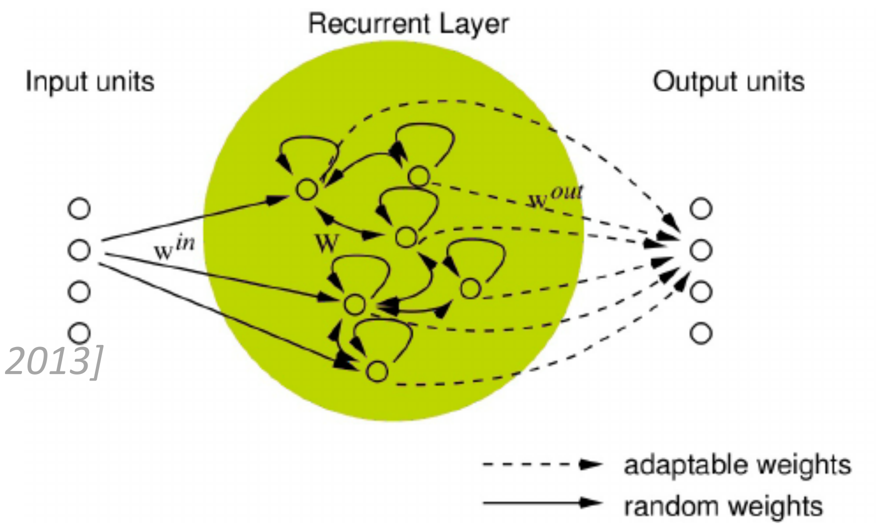
## Recommender system based on RandSVD

MovieLens 20M database: 27.000 movies x 138.000 users, with 0.5% non-zero entries

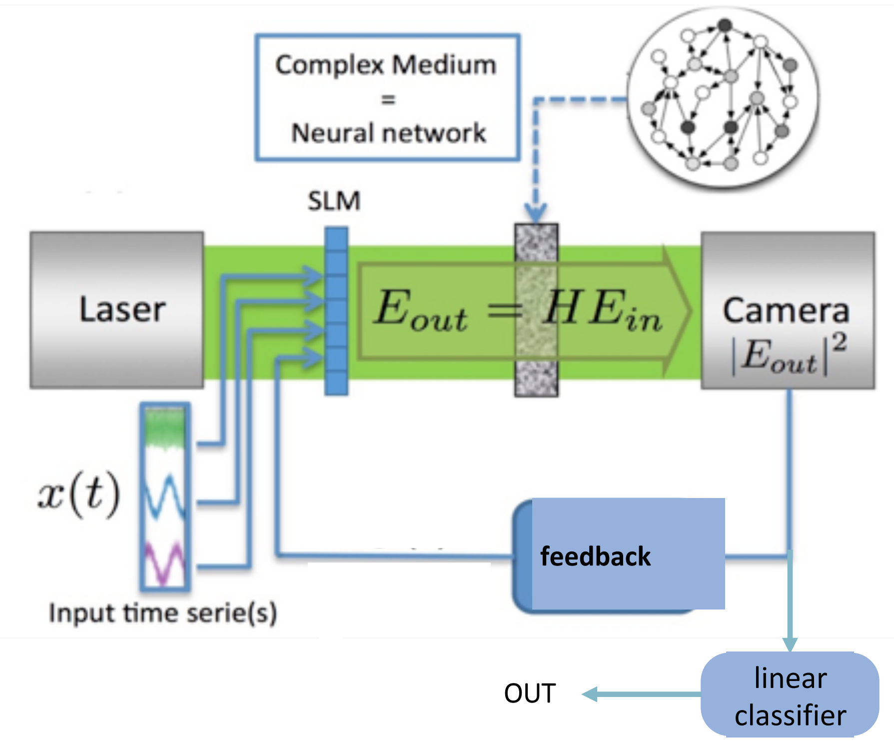


# Case Study 5: Optical Echo-State Network

## A physical implementation of large-scale echo-state networks

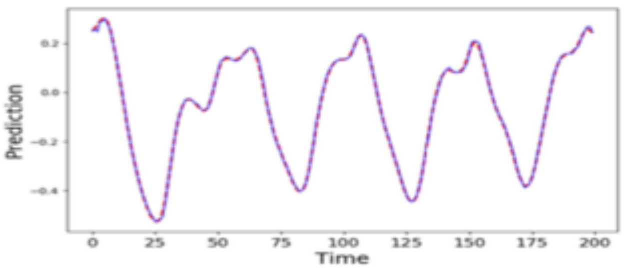



*[from Obst et al. 2013]*




regression on complex time series

Ex: predict dynamics of Mackey-Glass eqs. *(Dong. et al)*

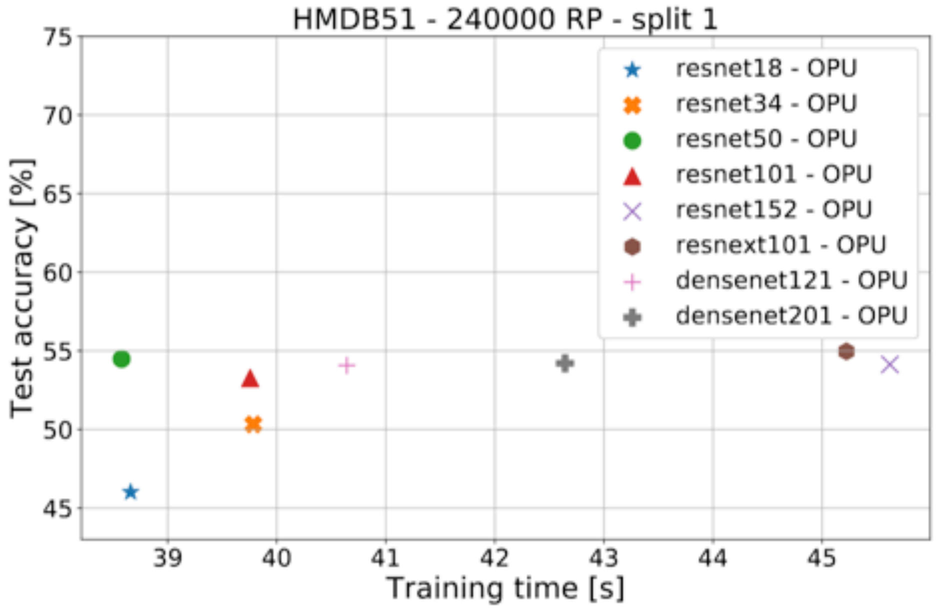



2 orders of magnitude larger than standard PCs

# Case Study 6: Video classification

Confidential & Proprietary

Transfer Learning on videos [1] using the Kinetics dataset




| Method | HMDB-51 |
|---|---|
| ResNet-18 (scratch) | 17.1 |
| ResNet-18 | 56.4 |
| ResNet-34 | 59.1 |
| ResNet-50 | 61.0 |
| ResNet-101 | 61.7 |
| ResNet-152 | 62.4 |
| ResNet-200 | 63.5 |
| DenseNet-121 | 59.6 |
| ResNeXt-101 | **63.8** |

**Left:** Accuracy versus training time for the HMDB51 dataset on different pretrained Kinetics models. **Right:** Accuracy with backprop.

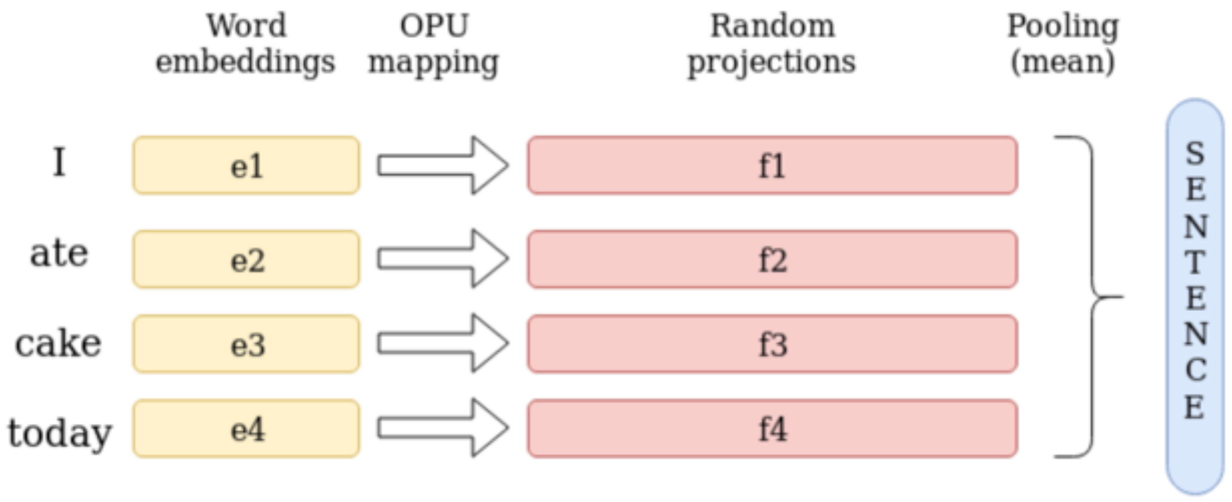
Standard training times for 3D-CNN:

- 150/200 epochs with 8 GPUs [2]
- 5k epochs “using just 16 GPUs” [3]


[1] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." 2018
[2] DIBA, Ali, et al. Temporal 3d convnets: New architecture and transfer learning for video classification. arXiv preprint arXiv:1711.08200, 2017.
[3] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." 2017.

# Case Study 7: Natural Language Processing





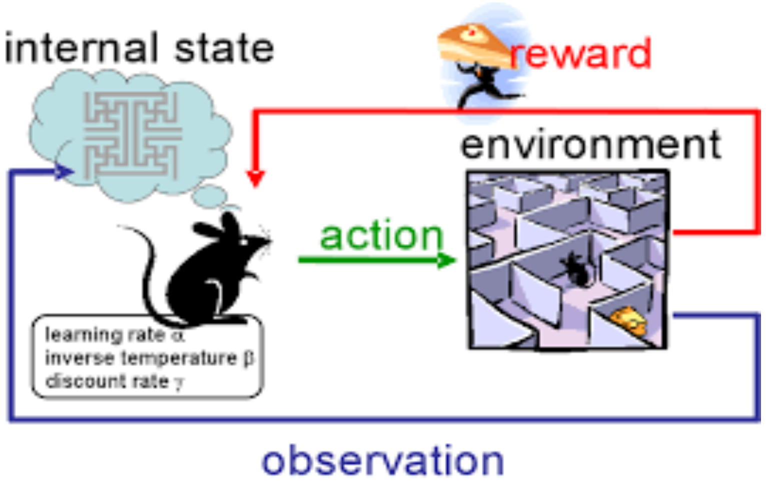

| | Ours | SkipThought | InferSent |
|---|---|---|---|
| MR | 79.4 | 79.4 | **81.1** |
| CR | 81.6 | 83.1 | **86.3** |
| MPQA | 87.7 | 89.3 | **90.2** |
| SST2 | 82.7 | 82.9 | **84.6** |
| SUBJ | 92.9 | **93.7** | 92.4 |
| TREC | **89.2** | 88.4 | 88.2 |
| SICK-E | 83.5 | 79.5 | **86.3** |
| SICK-R | 84.3 | 85.8 | **88.3** |
| MRPC | 75.2 | 73.2 | **76.2** |
| STSB | 70.6 | 68.9 | **75.6** |

Table 1: Results on downstream tasks of the SentEval toolkit [2]. Our model (15,000 random projections) is inspired from [1] and gets comparable results to much more sophisticated works. SkipThought [3] took a month to train and InferSent [2] requires annotated data that is thus far available only in English.

See LightOn's blogpost « Shedding Light On "Grand Débat" » on *Medium*, April 11th, 2019

# Case study 8: Reinforcement learning

How to learn complex tasks through
feedback ?



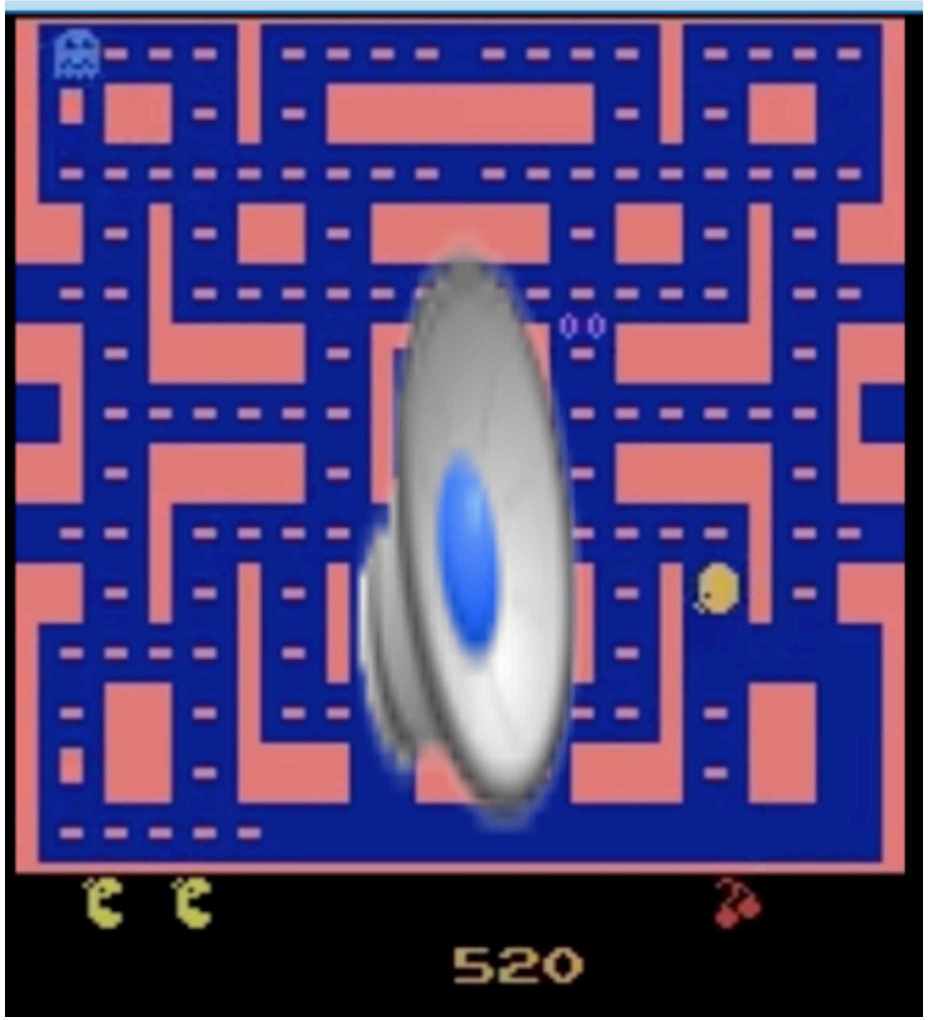


Becominghuman.ai

LightOn

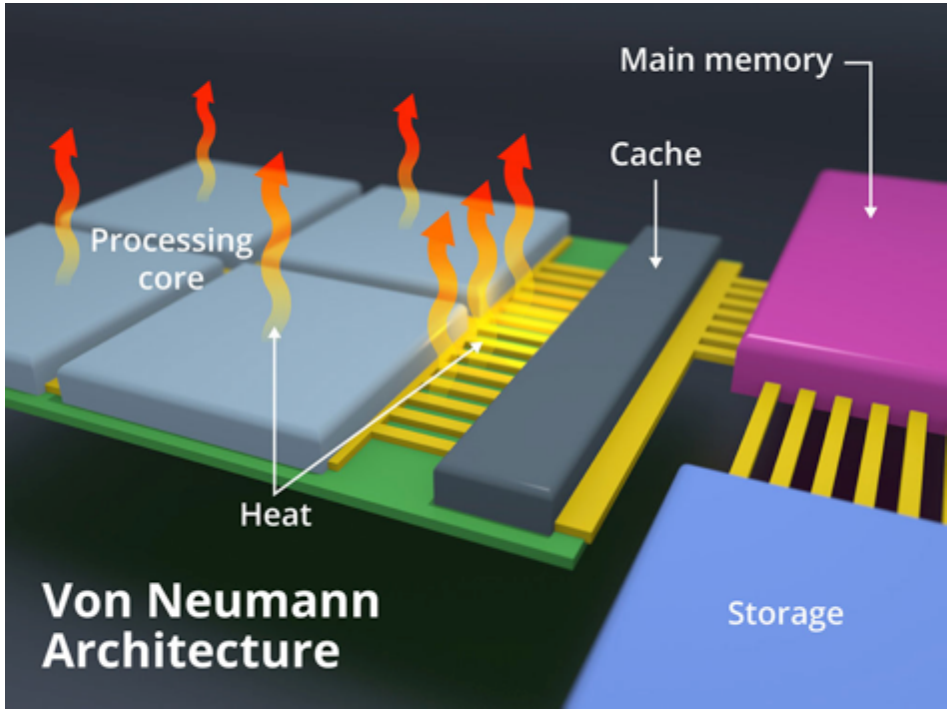
# Case study 8: Reinforcement learning

This system uses the OPU operation to search in past sequences a similar position



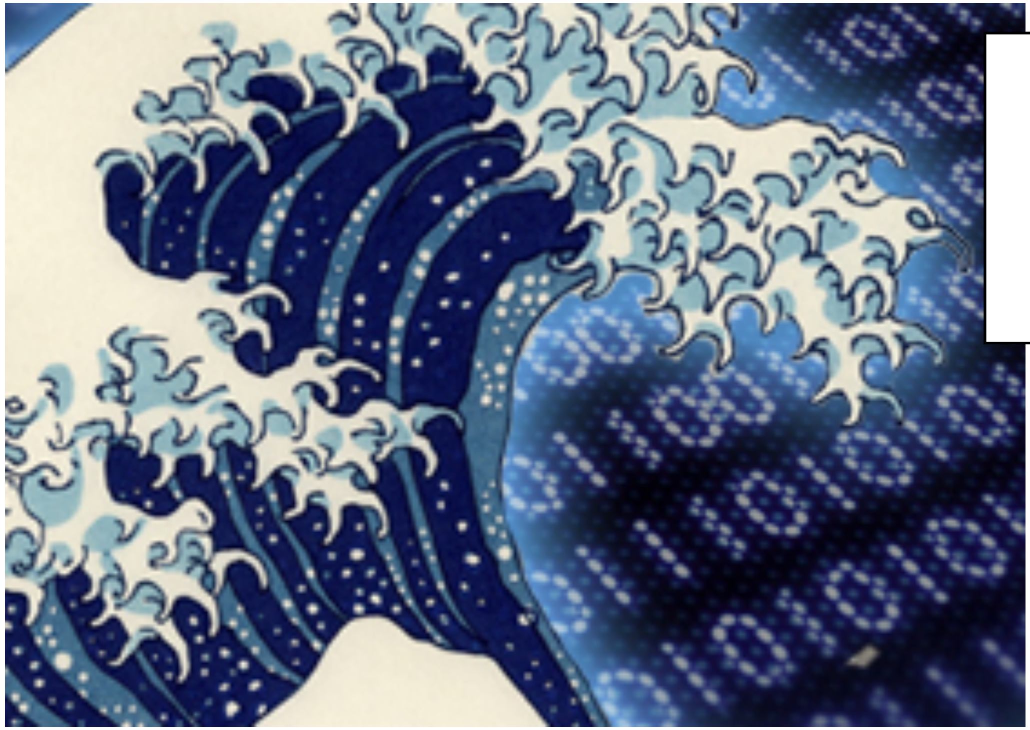

LightOn

Will Von Neumann architectures stay prevalent in the AI era ?




http://www.rochester.edu/newscenter/microprocessors-computing-architecture-304252/vonneumann-architecture/

# A conclusion and an invite



https://blog.apterainc.com/the-data-tsunami-is-coming.-is-your-company-ready

Optical Co-processing already allows some ML computations **at scale**

**Fast ML prototyping**

- Kernel Ridge Regression
- Fast Transfer Learning
- Echo-State Networks
- Change-point detection in streams and graphs
- Recommender systems with RandSVD
- …

**What's your case study ?**

laurent@LightOn.io

Confidential & Proprietary

# Selected references


- "Imaging With Nature: Compressive Imaging Using a Multiply Scattering Medium”, A. Liutkus et al., Scientific Reports 4 (july 2014)
- "Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques", A. Drémeau et al., Optics Express 23(9), 2015
- "Random Projections through multiple optical scattering: Approximating kernels at the speed of light", A. Saade et al., Proc. ICASSP (2016)
- "Scaling up Echo-State Networks with multiple light scattering", J. Dong et al., arXiv:1609.05204
- "NEWMA: a new method for scalable model-free online change-point detection", Nicolas Keriven, Damien Garreau, Iacopo Poli, arXiv:1805.08061
- "Machine Learning and the Physical Sciences », G. Carleo et al., arXiv 1903.10563
- ”Principled training of Neural Networks with Direct Feedback Alignment”, J. Launay et al., arXiv 1906.04554