LightOn
We bring Light to AI

# Building Innovative Hardware in an Era of Artificial Intelligence

Igor Carron, CEO and Co-Founder

LightOn.ai

# Why Build Innovative Hardware?

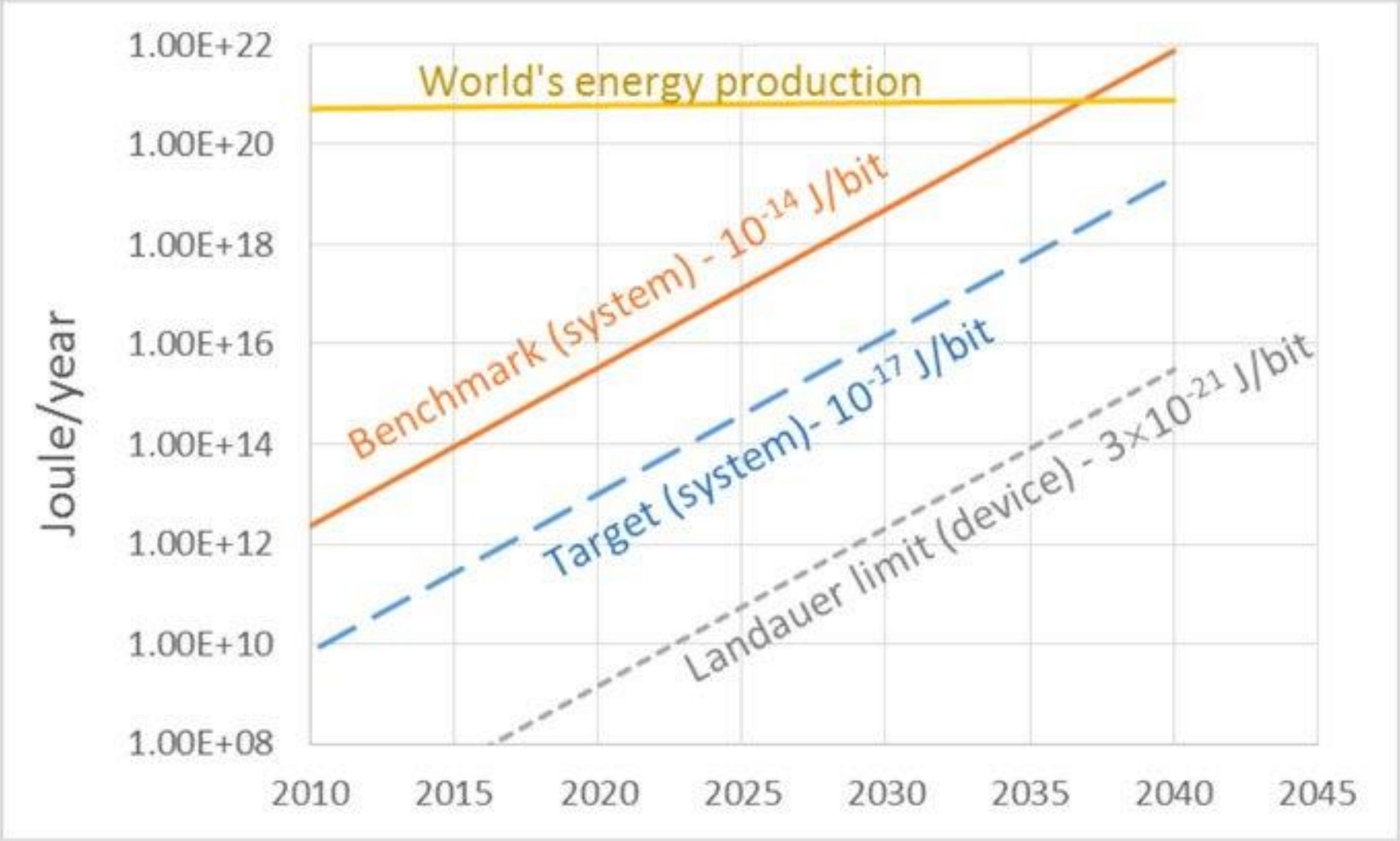
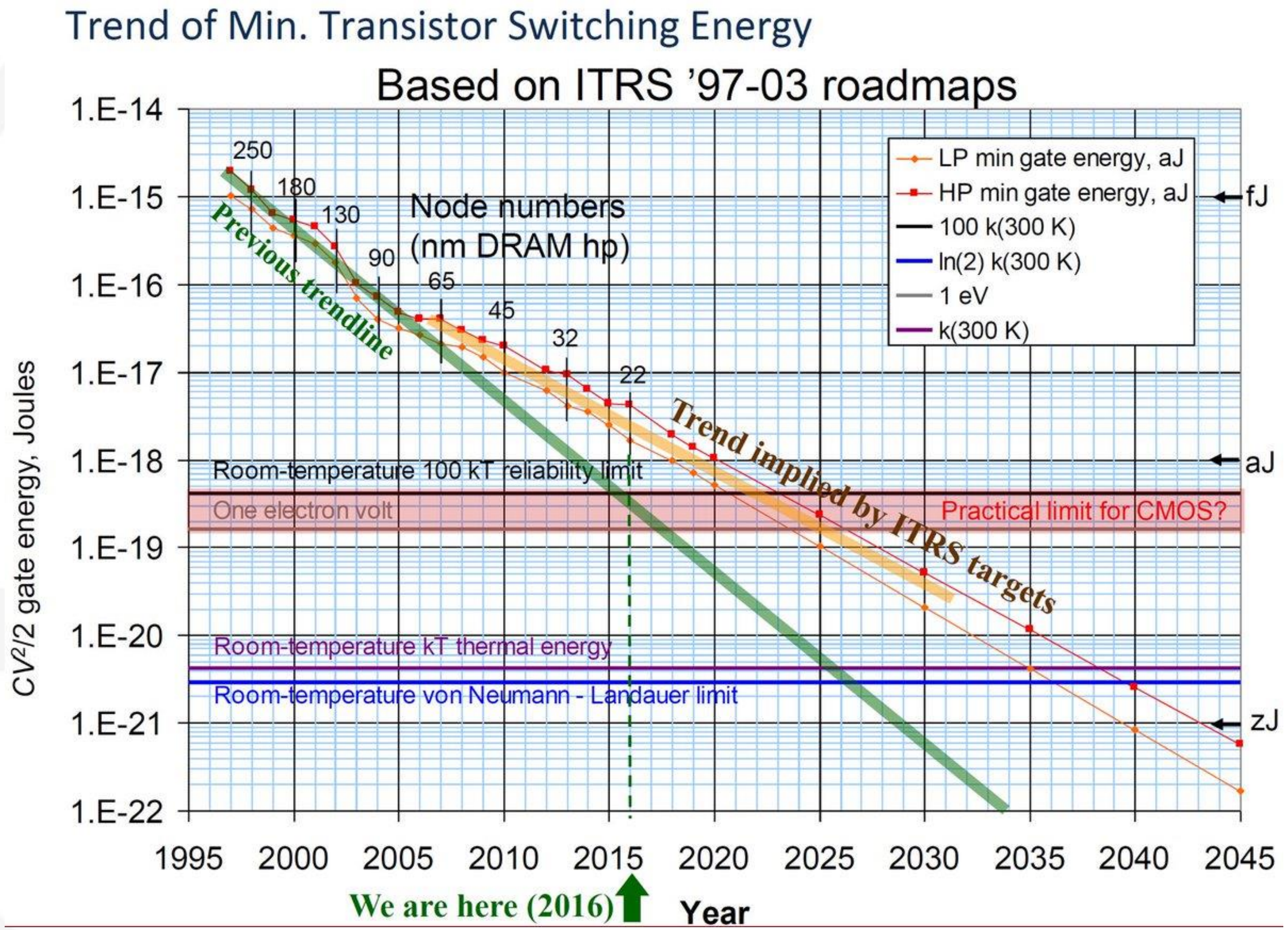



**Fig. A8. Total energy of computing.**

Source: Rebooting the IT Revolution: A Call to Action, SIA, SRC, 2015




Source: Mike P. Frank, SNL

# It’s worse!

blame it on the FWAlcohol
@nullhund
Suivre

crypto rush in one image

Traduire le Tweet


16:32 - 12 janv. 2018

## Bitcoin price (BTC)

€7,295.28 +€7,291.02 (71.3K%)

1H 24H 1W 1M 1Y ALL


| Market Cap | Volume (24 hours) | Circulating Supply | All-Time High |
|---|---|---|---|
| $129.5B | $16.6B | 17.8M BTC | $17,810.43 |

Source: Twitter, Coinbase

# Microsoft boss: World needs more computing power

By Joe Miller
BBC News, Davos

23 January 2018

Share

Davos 2018

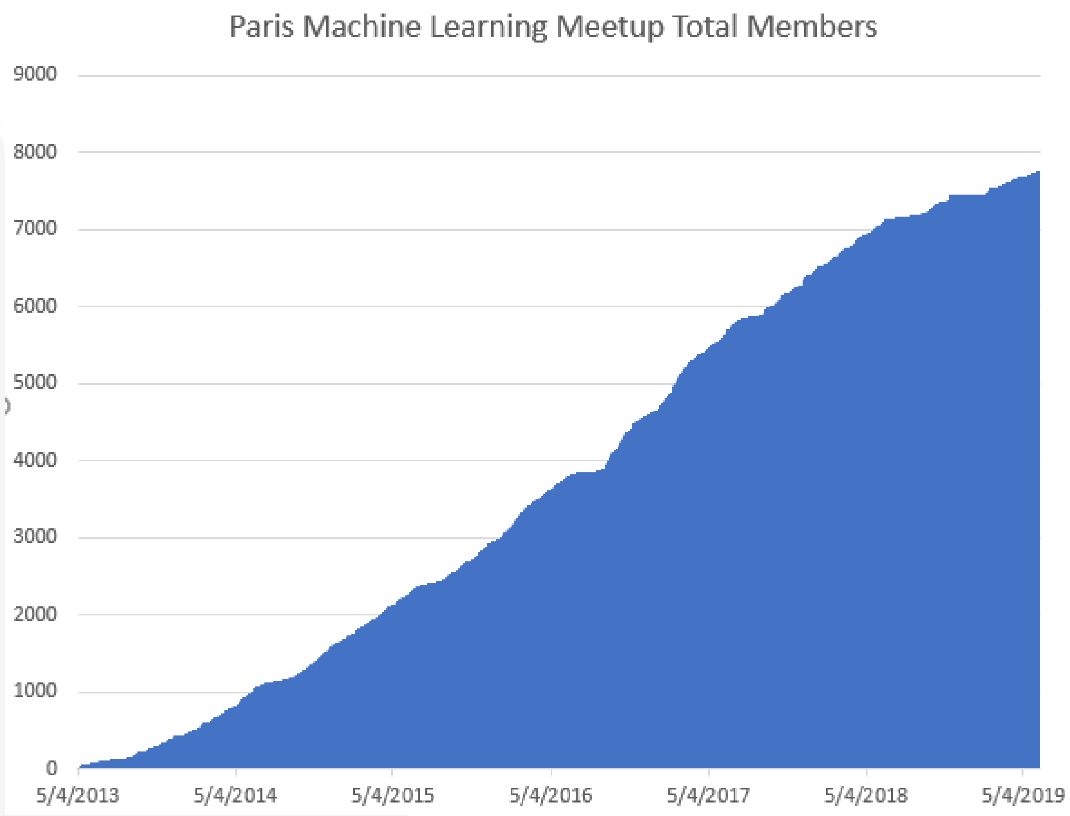

**The world is rapidly "running out of computing capacity", the head of tech giant Microsoft has warned.**

Source: BBC

# It’s much worse!

# The Footprint of Machine Learning


Chart: MIT Technology Review • Source: Strubell et al.

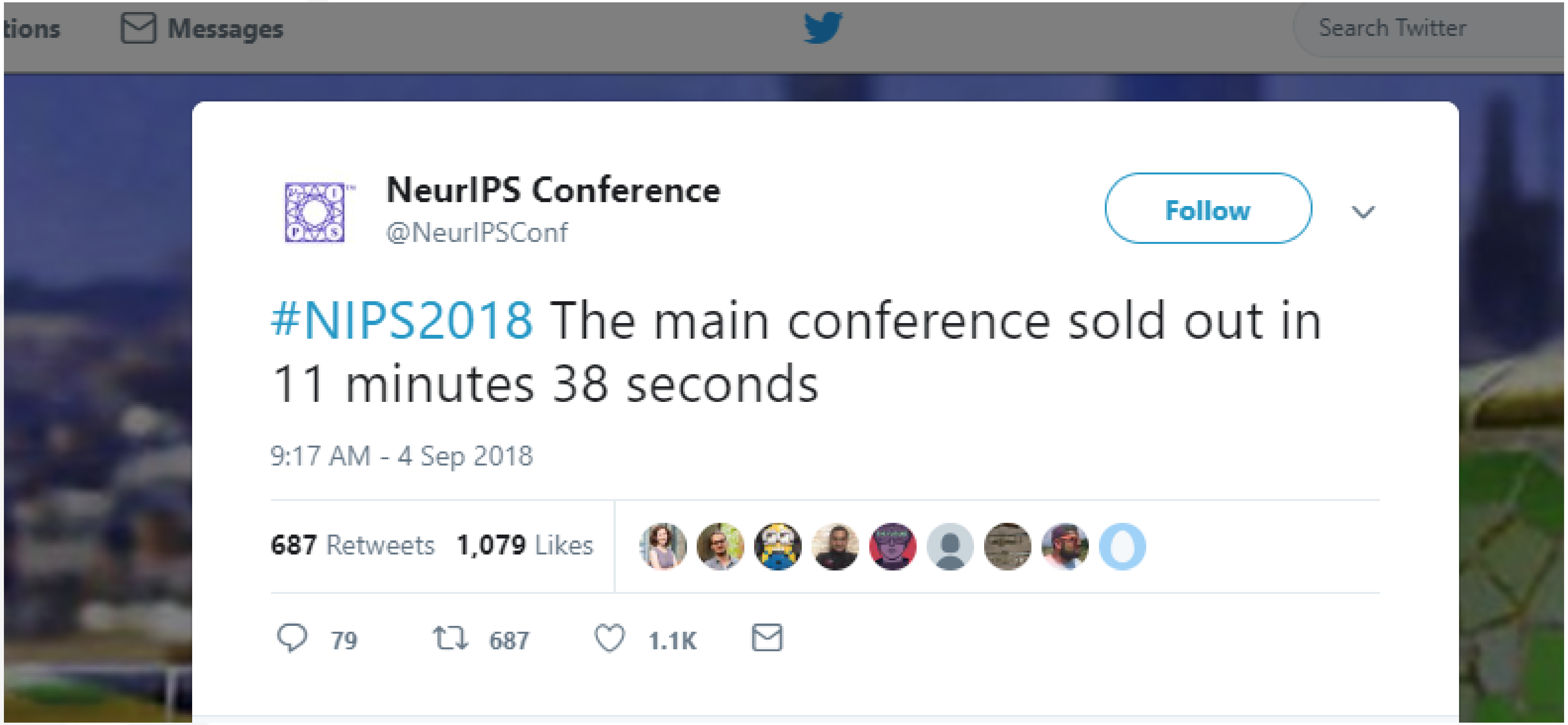




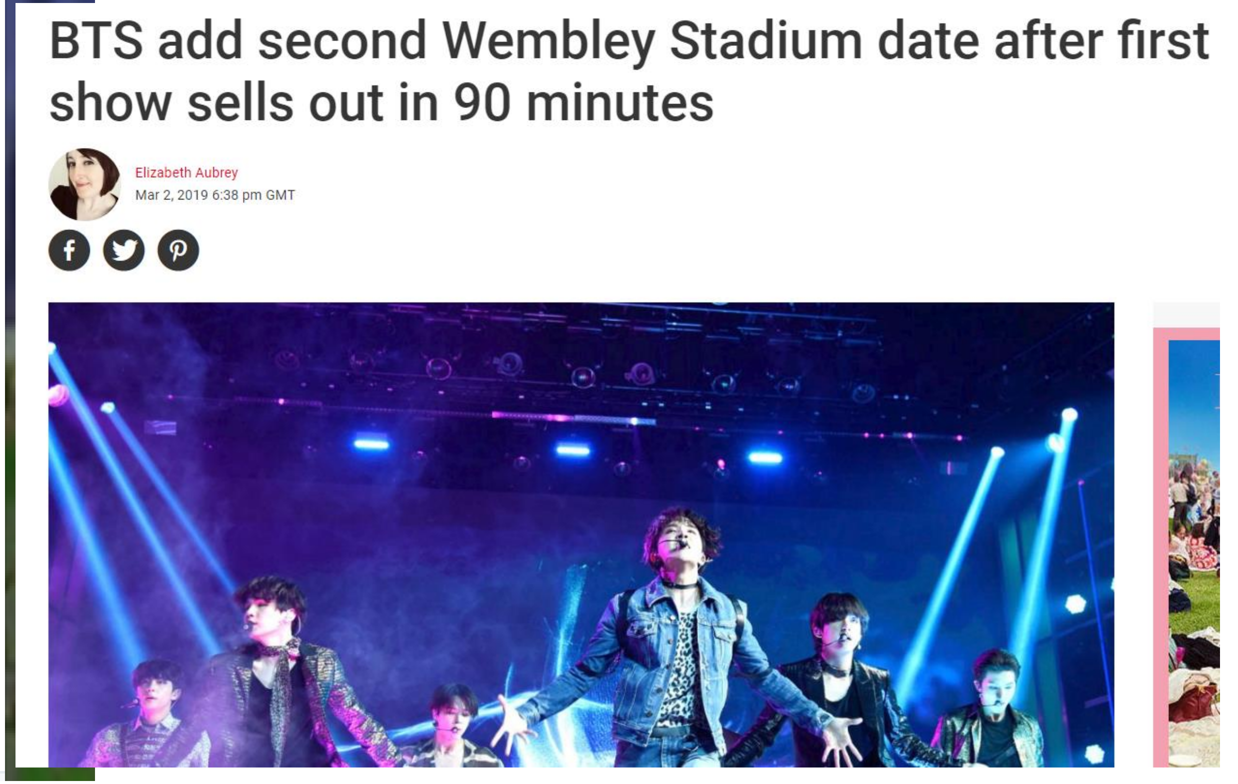

MachineLearning

join 682,597 readers

1,139 users here now


tions
Messages
Search Twitter
NeurIPS Conference
@NeurIPSConf
Follow
#NIPS2018 The main conference sold out in 11 minutes 38 seconds
9:17 AM - 4 Sep 2018
687 Retweets 1,079 Likes
79
687
1.1K


Messages
Search Twitter
NeurIPS Conference
@NeurIPSConf
Follow
#NIPS2018 The main conference sold out in 11 minutes 38 seconds
9:17 AM - 4 Sep 2018
687 Retweets 1,079 Likes
79
687
1.1K


BTS add second Wembley Stadium date after first show sells out in 90 minutes
Elizabeth Aubrey
Mar 2, 2019 6:38 pm GMT

# Scaling up a new technology from the ground up....fast

# Scaling up a Technology fast

European Commission

## Technology Readiness Levels

**TRL 0:** **Idea**. Unproven concept, no testing has been performed.

**TRL 1:** **Basic research.** Principles postulated and observed but no experimental proof available.

**TRL 2:** **Technology formulation**. Concept and application have been formulated.

**TRL 3:** **Applied research**. First laboratory tests completed; proof of concept.

**TRL 4:** **Small scale prototype** built in a laboratory environment ("ugly" prototype).

**TRL 5:** **Large scale prototype** tested in intended environment.

**TRL 6:** **Prototype system** tested in intended environment close to expected performance.

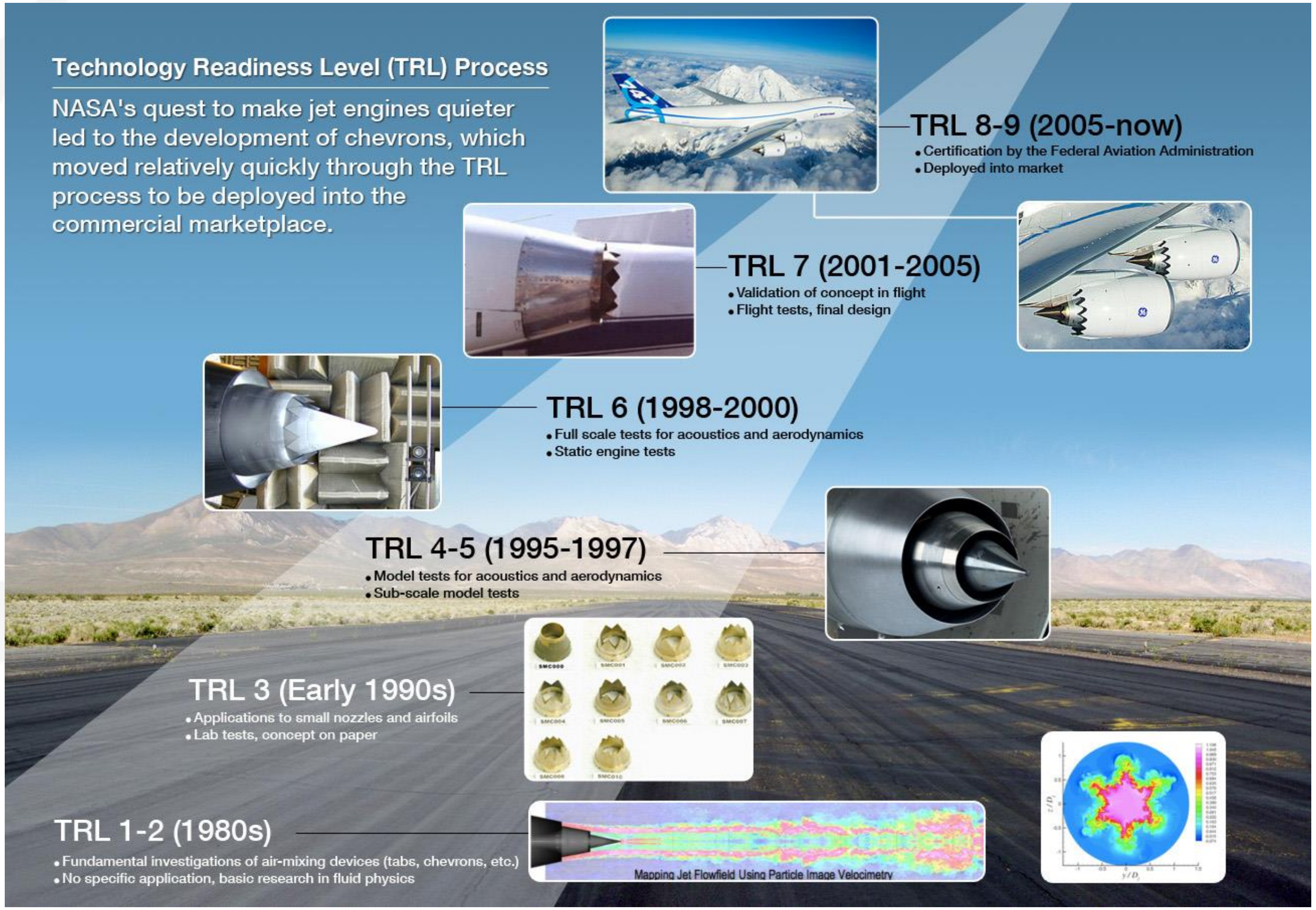
**TRL 7:** **Demonstration system** operating in operational environment at pre-commercial scale.

**TRL 8:** **First of a kind commercial system**. Manufacturing issues solved.

**TRL 9:** **Full commercial application**, technology available for consumers.

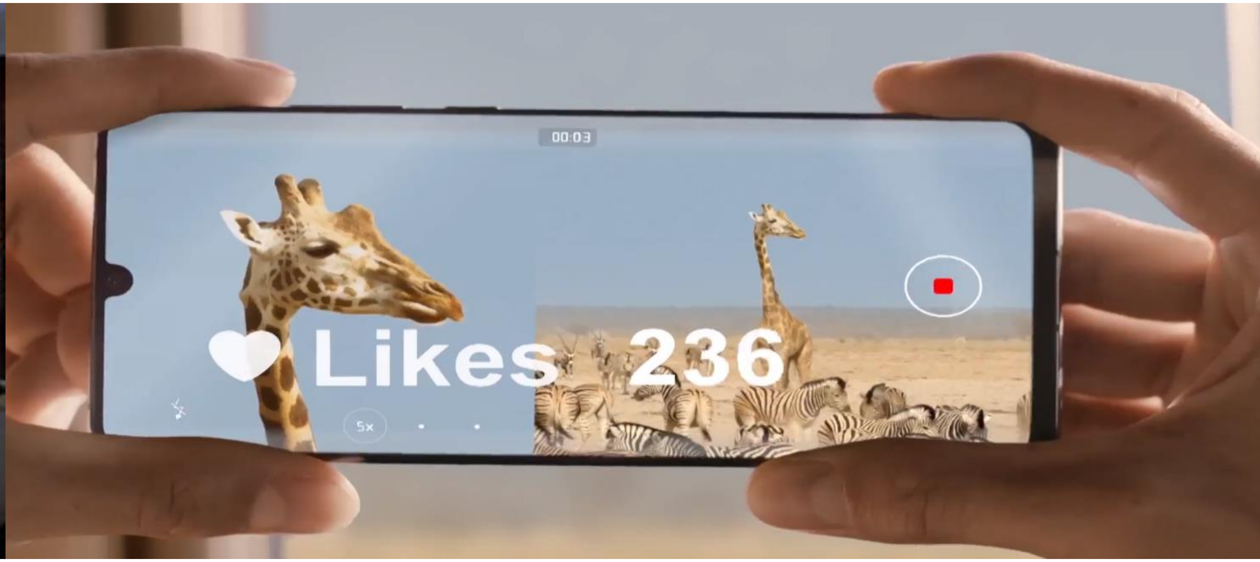
Source: European Commission

# Unlike NASA/FAA, we don't have 15+ years



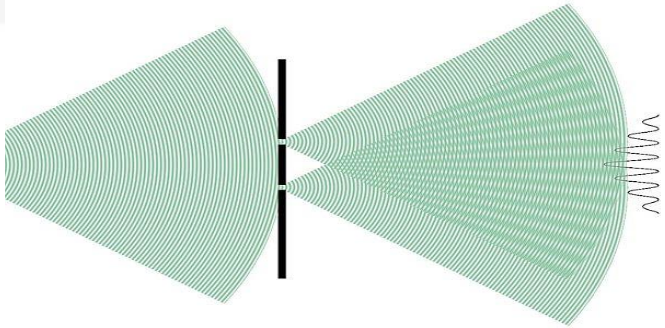

Source: NASA

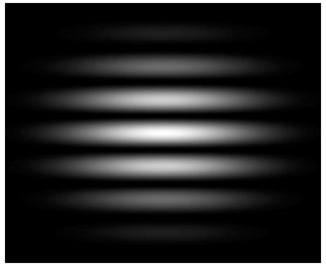
# Surfing on Moore's Law




Source: TI, Huawei

# Surfing Moore's Law and Using Nature





# A Fast and Large Random Projection

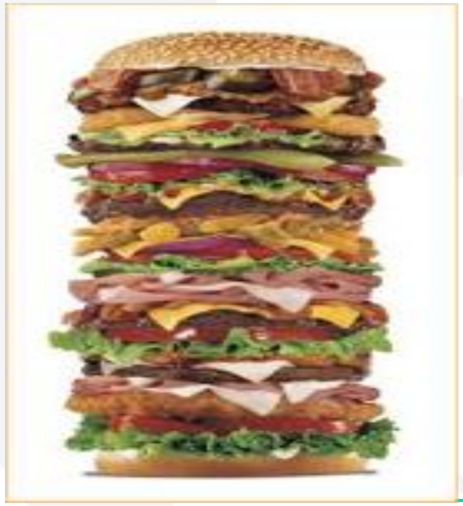
Discrete
input vector

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

**Random** transmission matrix

Discrete
output vector
(speckle intensity)

$$x \qquad H \qquad y = |\, H\, x\, |^2$$



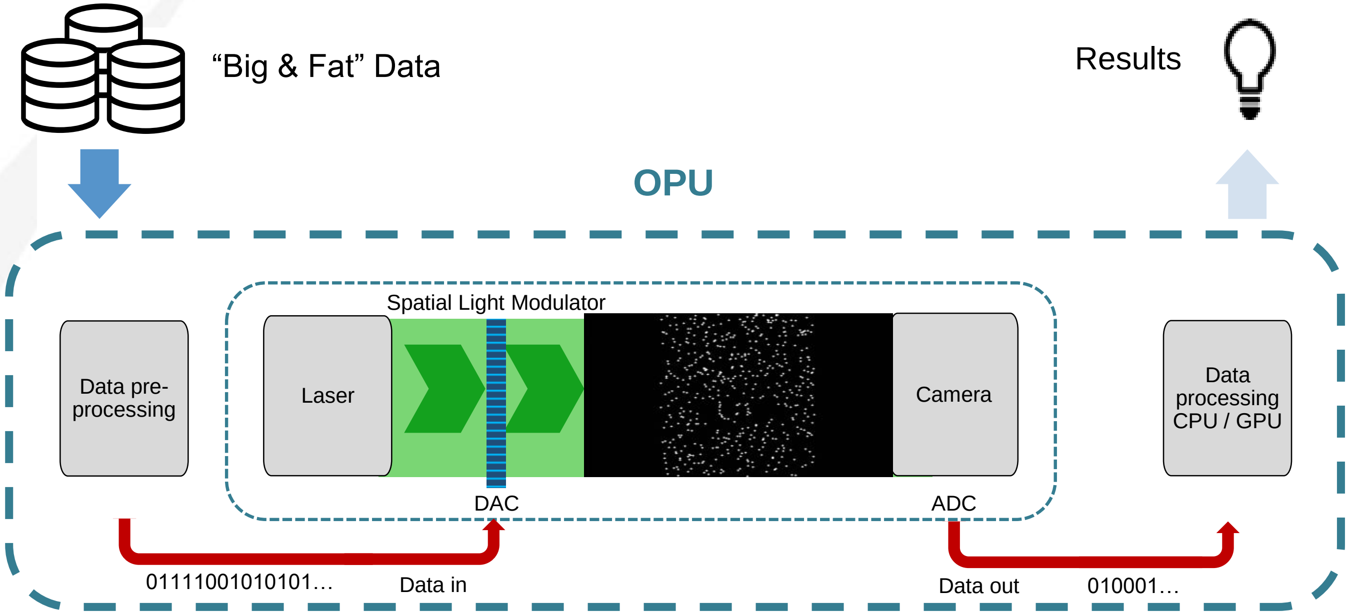
EXTRA-LARGE
H of size higher than
$10^6 \times 10^6$
(TBs of memory)

&

SUPER-FAST

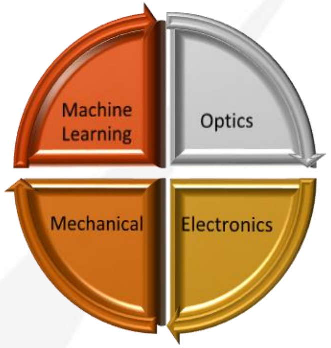
kHz operation
$\rightarrow 10^3$ such
multiplies / s

# How it works

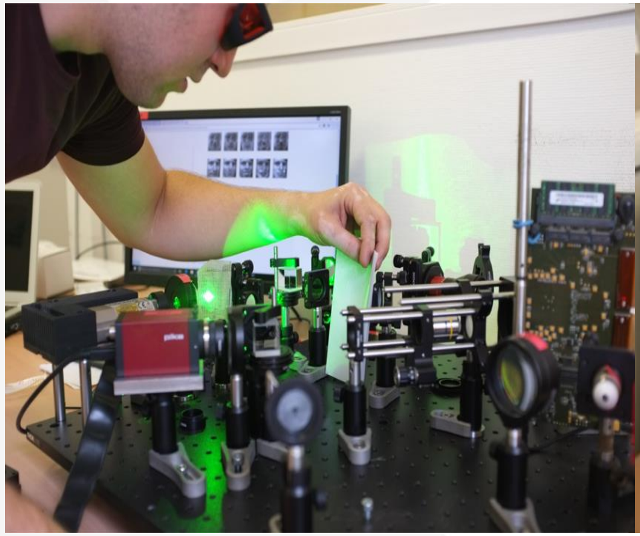


"Big & Fat" Data
Results
OPU
Spatial Light Modulator
Data pre-processing
Laser
Camera
Data processing CPU / GPU
DAC
ADC
01111001010101…
Data in
Data out
010001…

# Climbing up the TRL scale fast




Fast prototyping among a diverse set of expertise




TRL 6-7 (2018-2019)

TRL 3-4 (2016)

TRL 5 (2017)

# Optical Processing Unit

# Why Random Projections?

### Johnson-Lindenstrauss Lemma

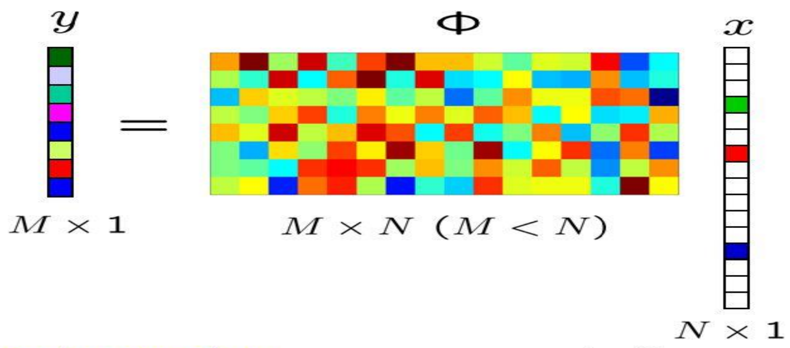
**Lemma** For any $0 < \epsilon < 1$ and any interger n let k be a po sitive interger such that

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

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 $x_i, x_j \in A$

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

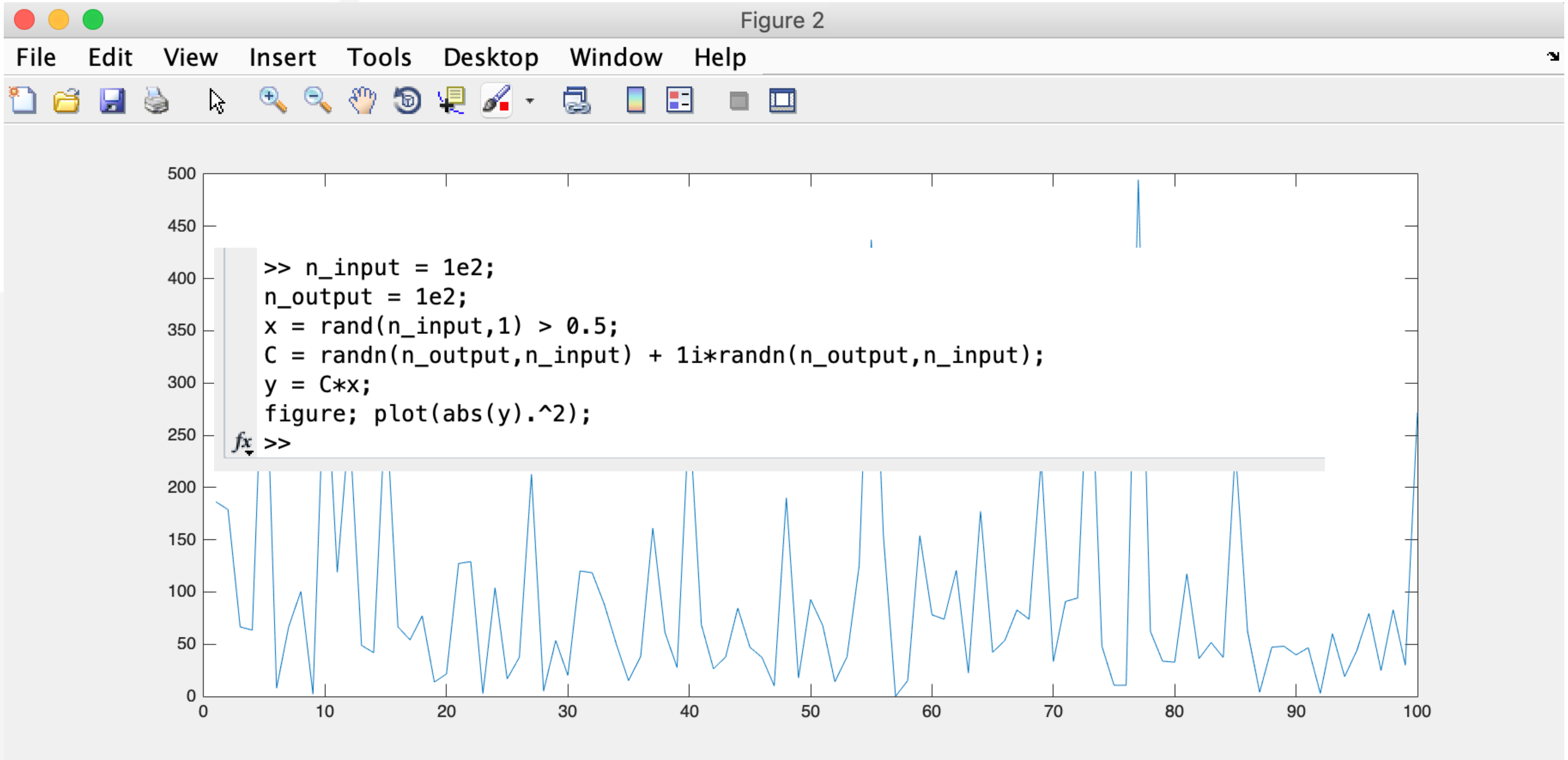
### Dimensionality Reduction




| | |
|---|---|
| **Compressive sensing** | non-adaptive measurements |
| **Sparse Bayesian learning** | dictionary of features |
| **Information theory** | coding frame |
| **Computer science** | sketching matrix / expander |

Source: Volkan Cevher, EPFL

# OPU: No need for large memory



# OPU: No need for large memory

LightOn

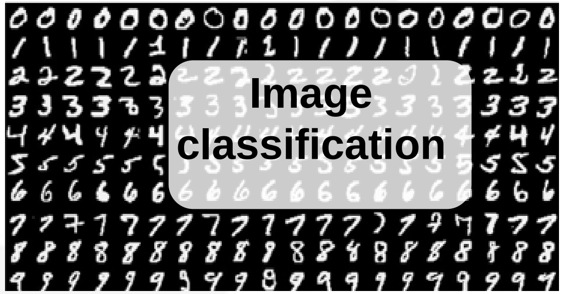
```
n_input = 1e6;
n_output = 1e6;
x = rand(n_input,1) > 0.5;
C = randn(n_output,n_input) + 1i*randn(n_output,n_input);
y = C*x;
figure; plot(abs(y).^2);
Error using randn
Requested 1000000x1000000 (7450.6GB) array exceeds maximum array size preference. Creation of arrays
greater than this limit may take a long time and cause MATLAB to become unresponsive. See array size
limit or preference panel for more information.

fx >>
```

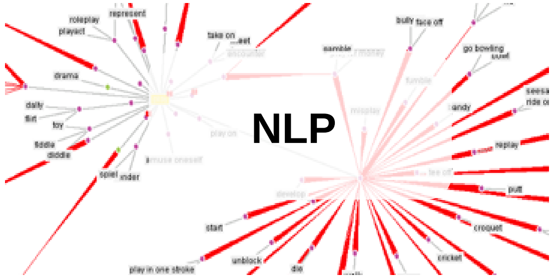
# A Machine Learning example

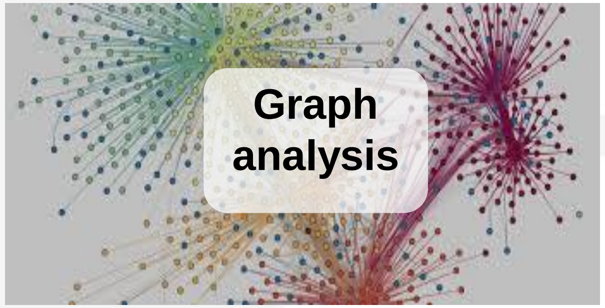
**On a typical Machine Learning training task (transfer learning)**

# Other Uses Already Investigated







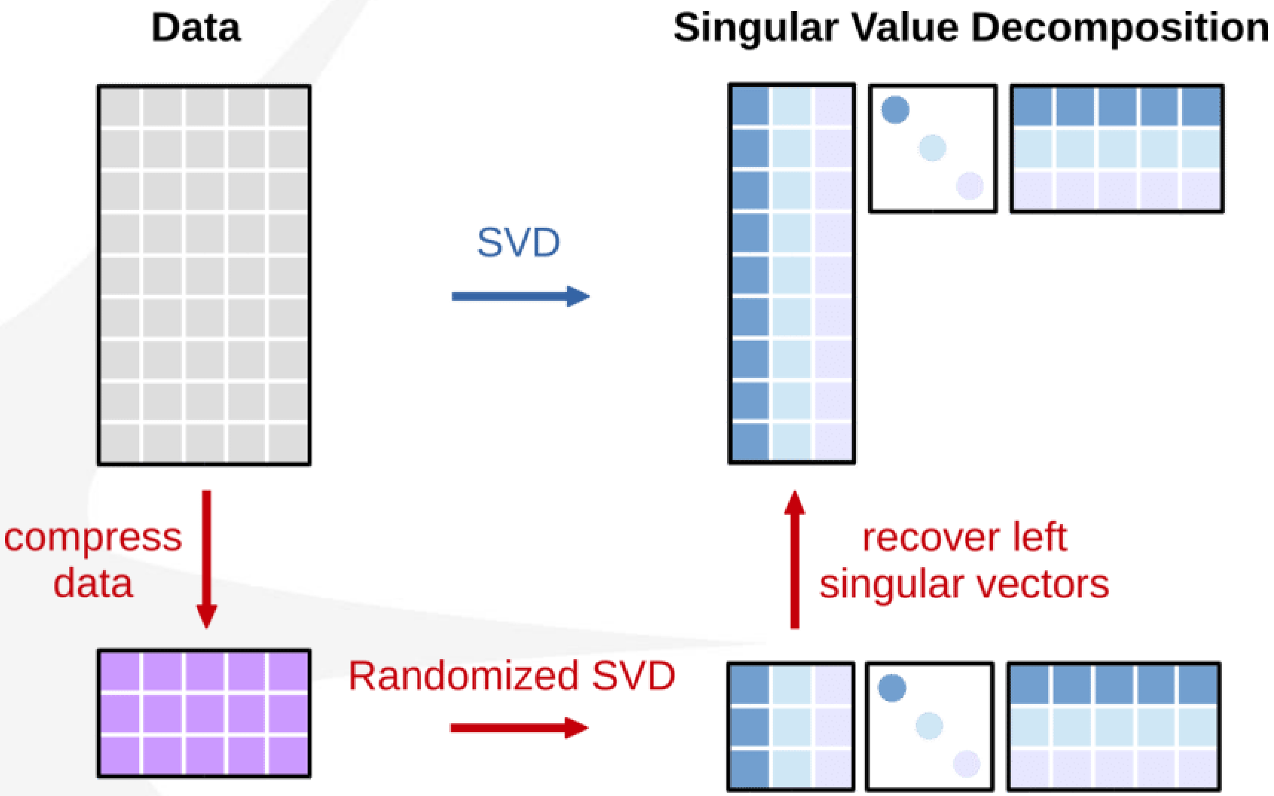



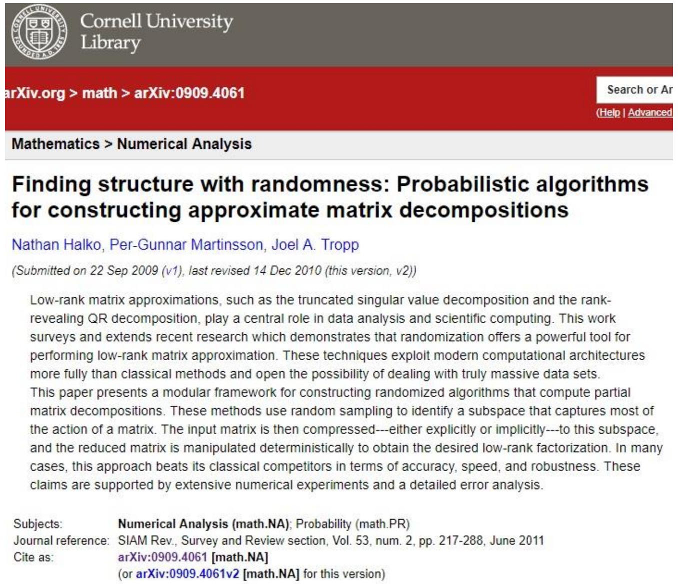


# Recommender Systems

**Recommender Systems shape our lives at scale!**

# Randomized Linear Algebra



Cornell University Library

arXiv.org > math > arXiv:0909.4061

Search or Ar

(Help | Advanced

Mathematics > Numerical Analysis

**Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions**

Nathan Halko, Per-Gunnar Martinsson, Joel A. Tropp

*(Submitted on 22 Sep 2009 (v1), last revised 14 Dec 2010 (this version, v2))*

Low-rank matrix approximations, such as the truncated singular value decomposition and the rank-revealing QR decomposition, play a central role in data analysis and scientific computing. This work surveys and extends recent research which demonstrates that randomization offers a powerful tool for performing low-rank matrix approximation. These techniques exploit modern computational architectures more fully than classical methods and open the possibility of dealing with truly massive data sets.
This paper presents a modular framework for constructing randomized algorithms that compute partial matrix decompositions. These methods use random sampling to identify a subspace that captures most of the action of a matrix. The input matrix is then compressed---either explicitly or implicitly---to this subspace, and the reduced matrix is manipulated deterministically to obtain the desired low-rank factorization. In many cases, this approach beats its classical competitors in terms of accuracy, speed, and robustness. These claims are supported by extensive numerical experiments and a detailed error analysis.

| | |
|---|---|
| Subjects: | **Numerical Analysis (math.NA)**; Probability (math.PR) |
| Journal reference: | SIAM Rev., Survey and Review section, Vol. 53, num. 2, pp. 217-288, June 2011 |
| Cite as: | arXiv:0909.4061 **[math.NA]** |
| | (or **arXiv:0909.4061v2 [math.NA]** for this version) |

**Randomized Matrix Decompositions using R, Aug 2016, N. Benjamin Erichson, Sergey Voronin, Steven L. Brunton, J. Nathan Kutz**


# LightOn.ai