

# Why Supercomputing matters to Deep Learning

ESADE – MIBA (FALL 2017)

JORDI TORRES | FRANCESC SASTRE



A photograph of a crowded convention center. The ceiling is dark with a grid of rectangular light fixtures. In the foreground, many people are seen from behind, some wearing backpacks and casual clothing. The background shows more people and what appears to be a poster or display board.

Artificial Intelligence is changing our life



Quantum leaps in the quality of a wide range of everyday technologies thanks to Artificial Intelligence

Speech  
Recognition

We are increasingly interacting with "our" computers by just talking to them



**#1. Alexa**  
*(Amazon Echo)*



**#2. Cortana**  
*(Windows 10 Phone)*



**#3. Siri**  
*(iPhone)*



**#4. Google Now**  
*(Android)*

Natural  
Language  
Processing

*Google Translate* now renders spoken sentences in one language into spoken sentences in another, for **32 pairs** of languages and offers text translation for **100+ languages**.

The screenshot shows the Google Translate mobile application interface. At the top, there's a blue header bar with the text "Google Translate". Above the header, the phone's status bar displays signal strength, battery level at 41%, and the time 19:41. The main screen has "Spanish" on the left and "English" on the right, with a double-headed arrow between them. Below this, under "SPANISH", is the sentence "La inteligencia artificial se usa en muchos sitios". Underneath the sentence are three icons: a camera, a microphone, and a pen. Below the English sentence, there are four small square icons: a "G" logo, a mountain-like icon, a grid icon, and a document icon. The entire interface is set against a light gray background.

Lowi WiFi 19:41 41 %

Google Translate

Spanish ↔ English

SPANISH

La inteligencia artificial se usa en muchos sitios

ENGLISH

Artificial intelligence is used in many places

*Google Translate* now renders spoken sentences in one language into spoken sentences in another, for **32 pairs** of languages and offers text translation for **100+ languages**.

Natural  
Language  
Processing



la inteligencia artificial se usa en  
muchos sitios



Artificial intelligence is used in many  
places.



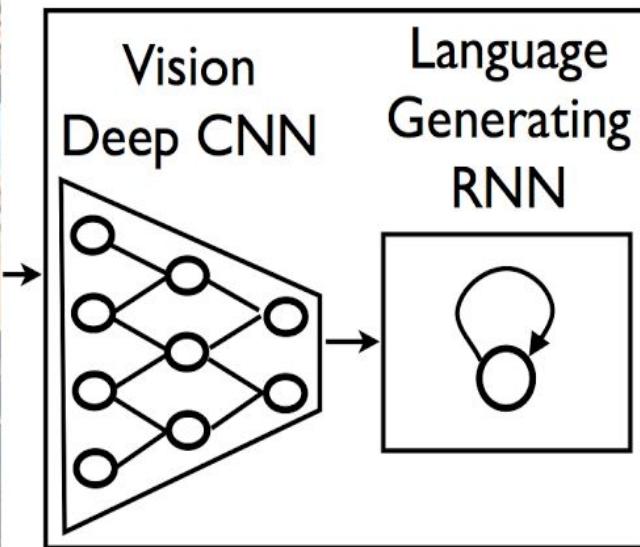
Preparing to speak...

español



English

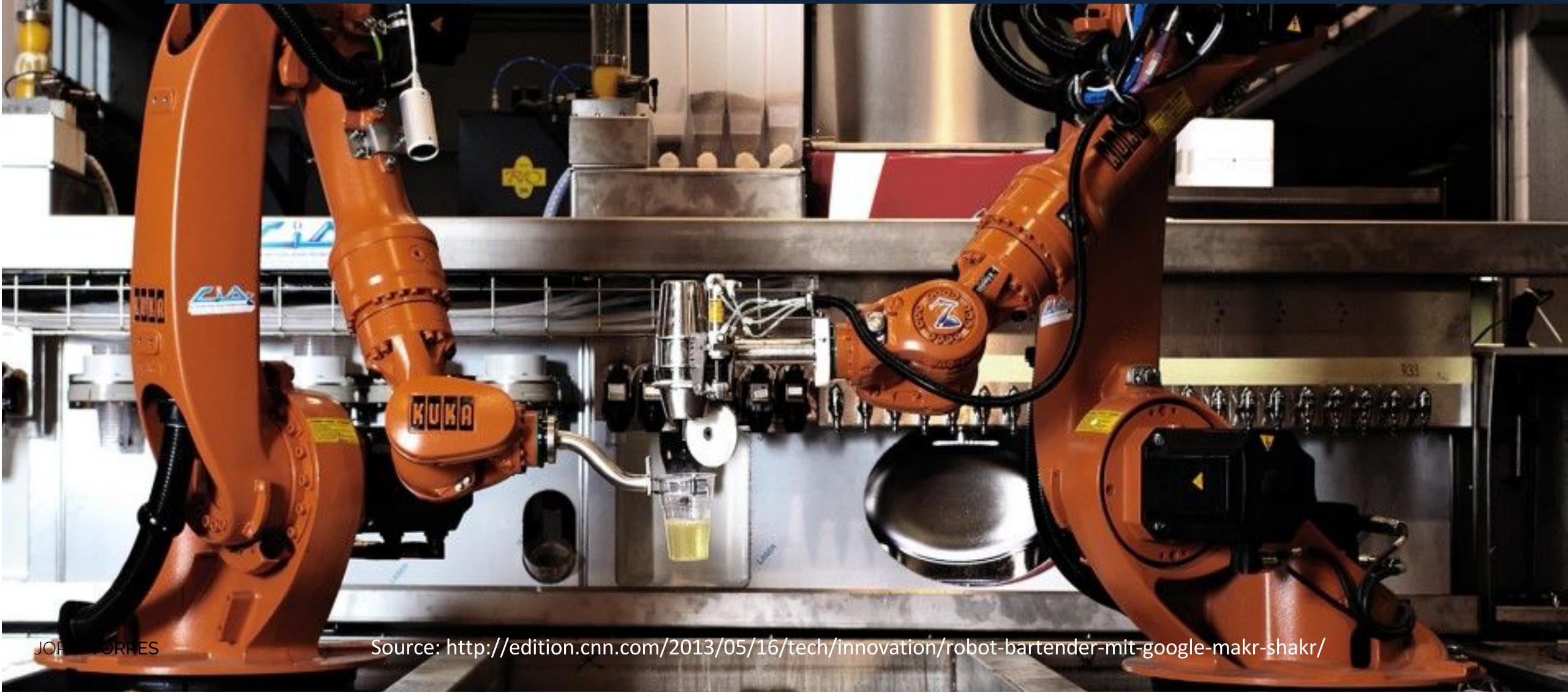
Now our computers can recognize images and generate descriptions for photos in seconds.



**A group of people shopping at an outdoor market.**

**There are many vegetables at the fruit stand.**

All these three areas are crucial to unleashing improvements in **robotics**, drones, self-driving cars, etc.



All these three areas are crucial to unleashing improvements in robotics, **drones**, self-driving cars, etc.



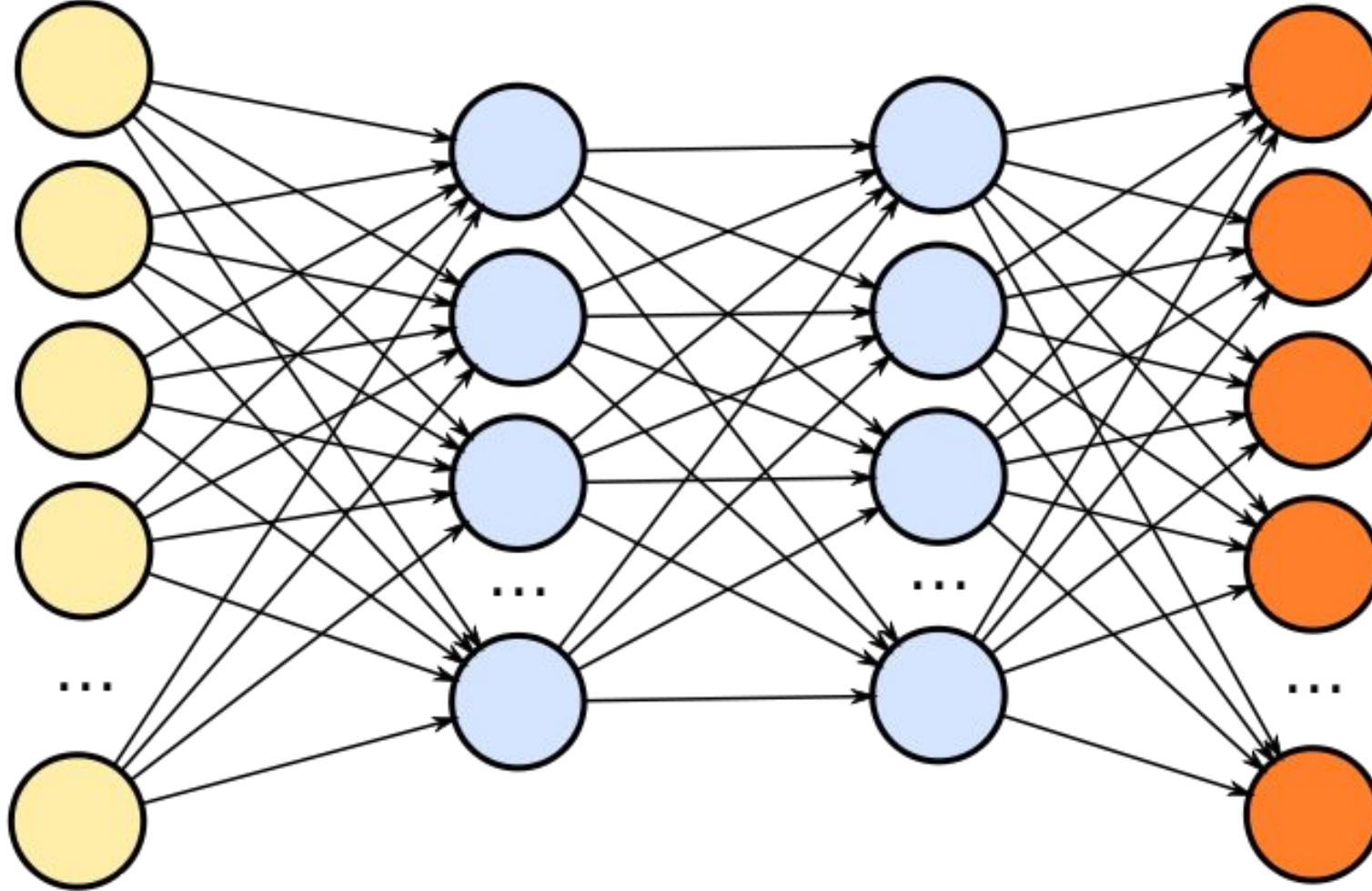
All these three areas are crucial to unleashing improvements in robotics, drones, **self-driving cars**, etc.

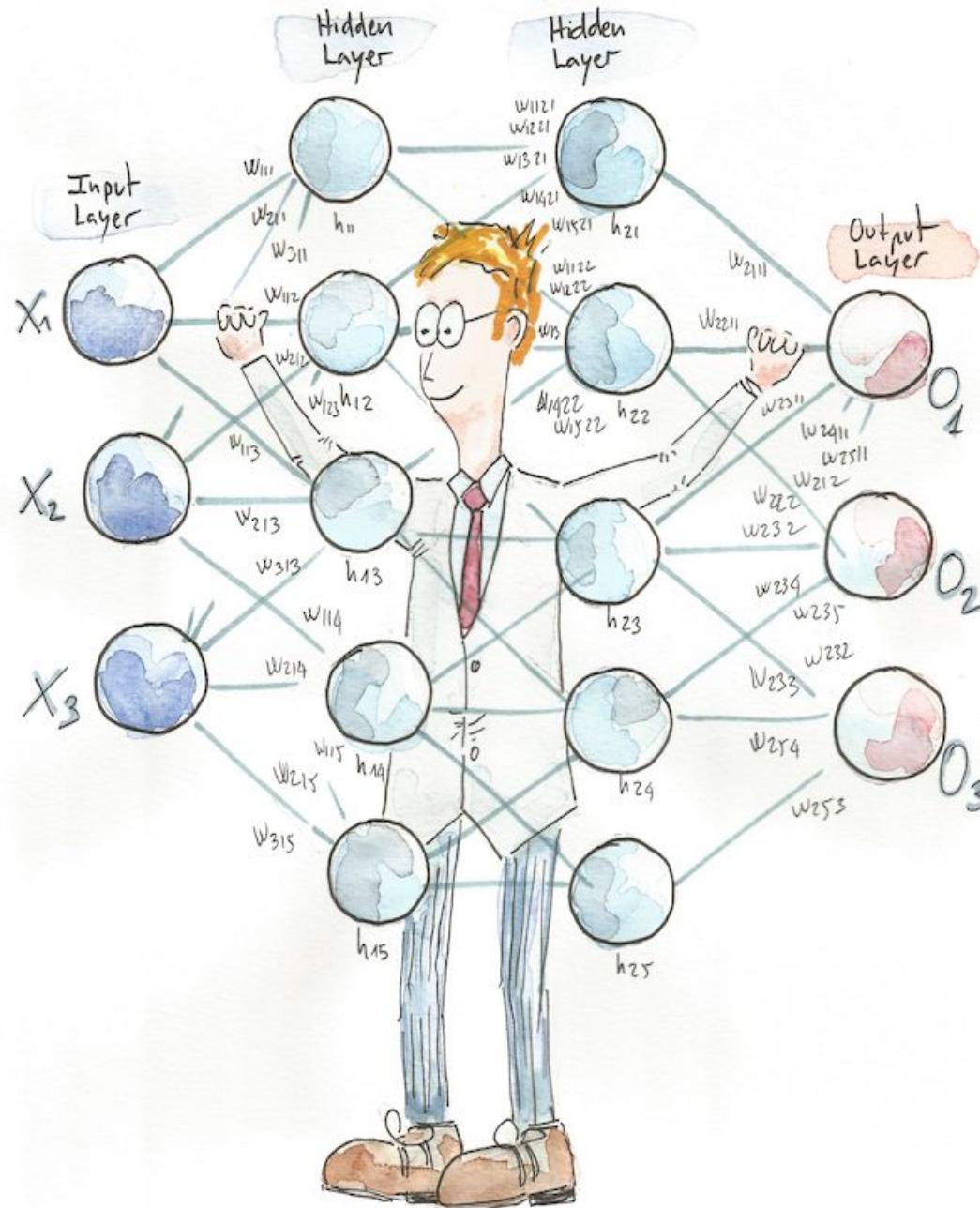


AI is at the heart of today's technological innovation.



Many of these breakthroughs have been made possible by a family of AI known as Neural Networks





Neural networks,  
also known as a  
**Deep Learning**,  
enables a  
computer to  
learn from  
observational  
data

Although the greatest impacts of  
deep learning may be obtained when  
it is integrated into the whole toolbox  
of other AI techniques

Universitat Politècnica de Barcelona

FACULTAT D'INFORMÀTICA  
GUIA DOCENT

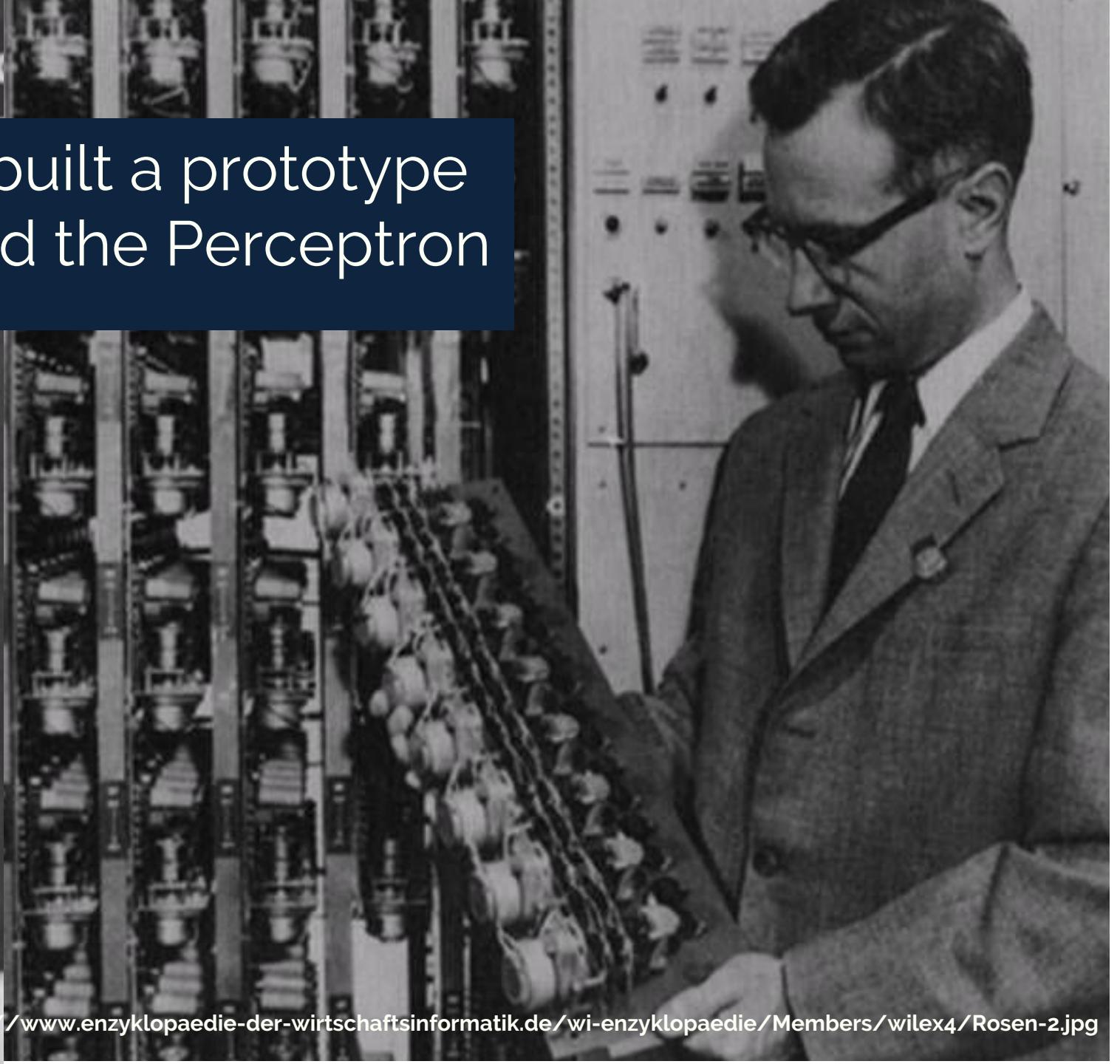
Curs 1982/83

John McCarthy coined the term  
Artificial Intelligence in the 1950s



perceptron

In 1958 Frank Rosenblatt built a prototype neural net, which he called the Perceptron



● "deep learning"

Search term

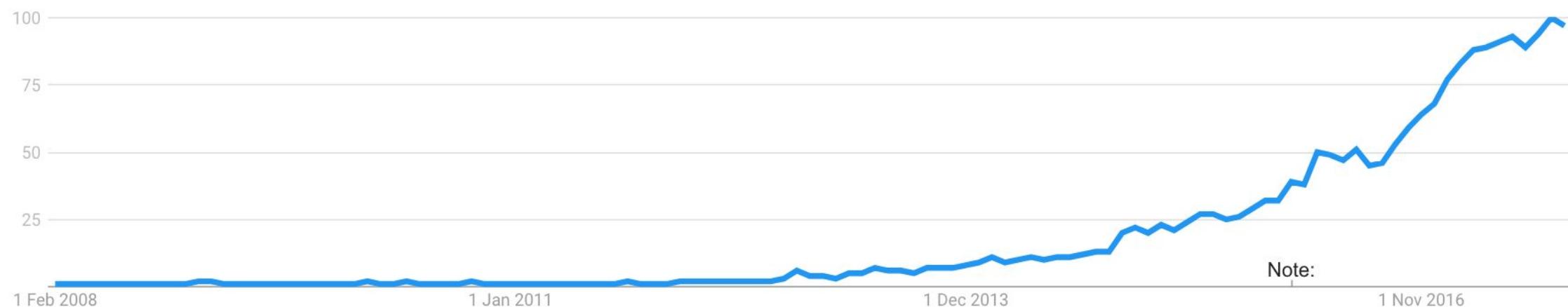
Worldwide ▾

09/01/2008 - 01/10/2017 ▾

All categories

So why did Deep Learning only take off few years ago?

Interest over time



Note:

# One of the key drivers: The data deluge



Source: <http://www.economist.com/node/15579717>

# One of the key drivers: The data deluge

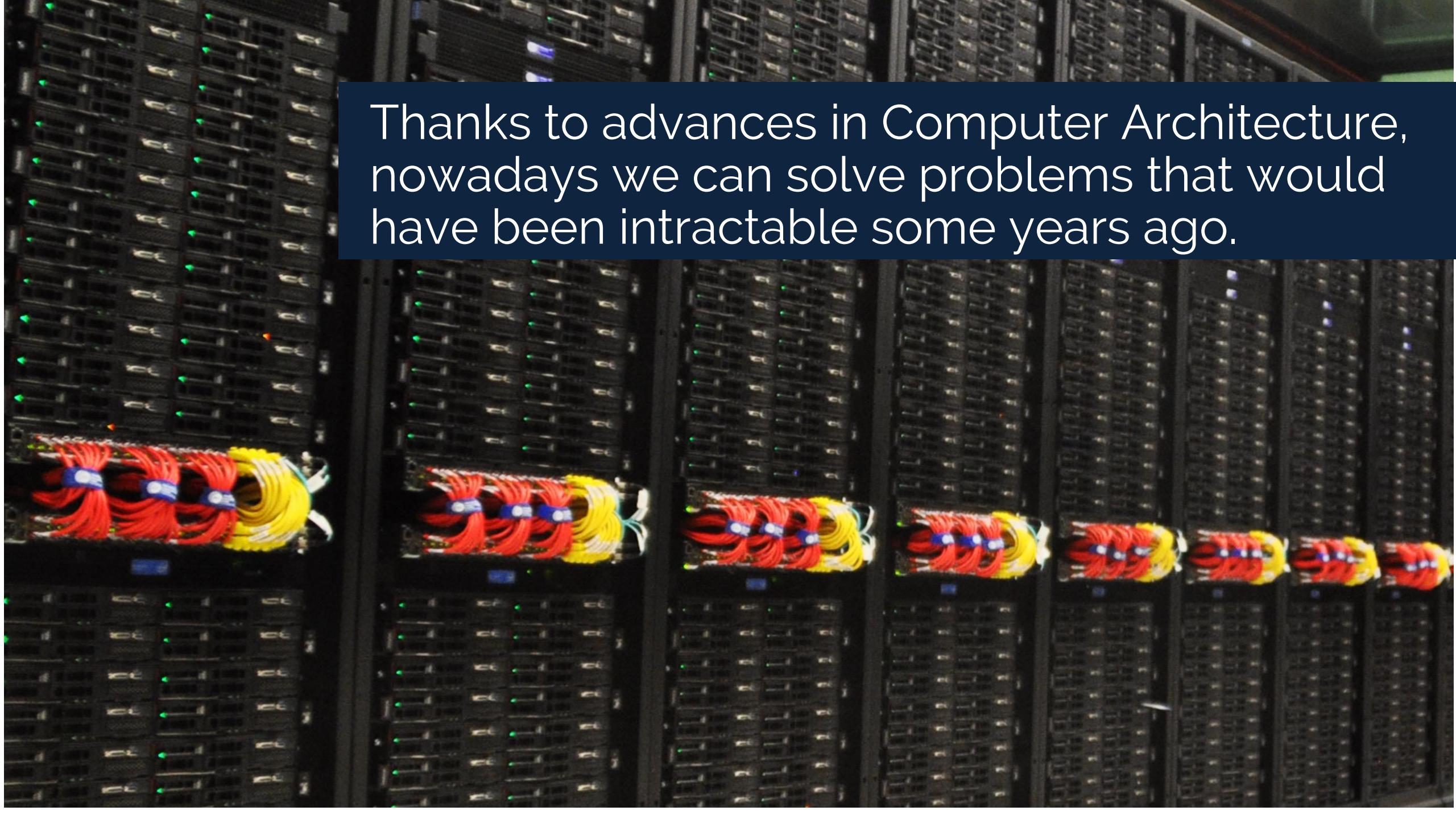
Thanks to the advent of Big Data  
AI models can be “trained” by  
exposing them to large data sets  
**that were previously unavailable.**

Source:<http://www.economist.com/node/15579717>



# Training DL neural nets has an insatiable demand for Computing





Thanks to advances in Computer Architecture, nowadays we can solve problems that would have been intractable some years ago.

# 1982

## FACOM 230 – Fujitsu

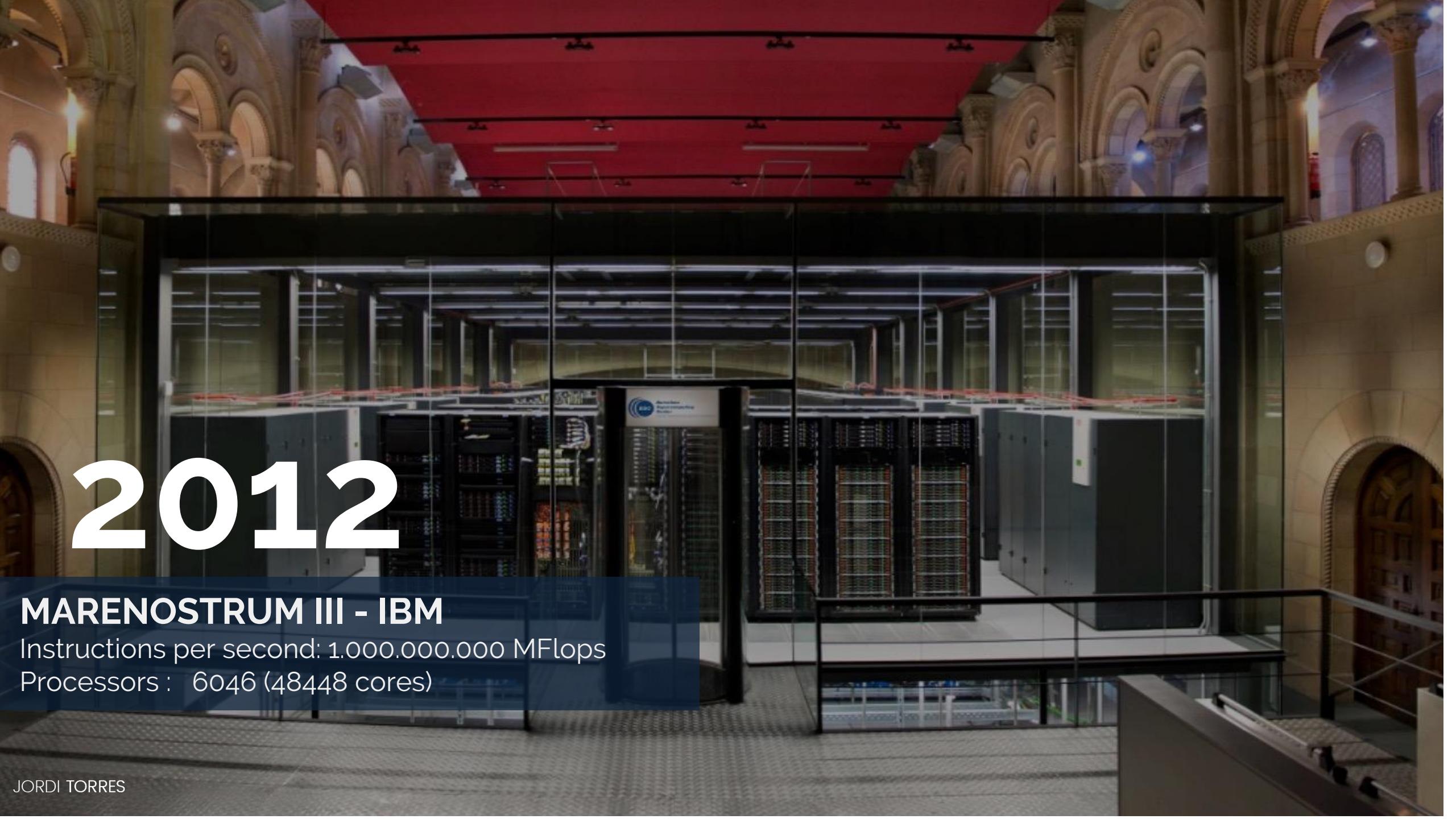
Instructions per second: few Mips \* (M = 1.000.000)

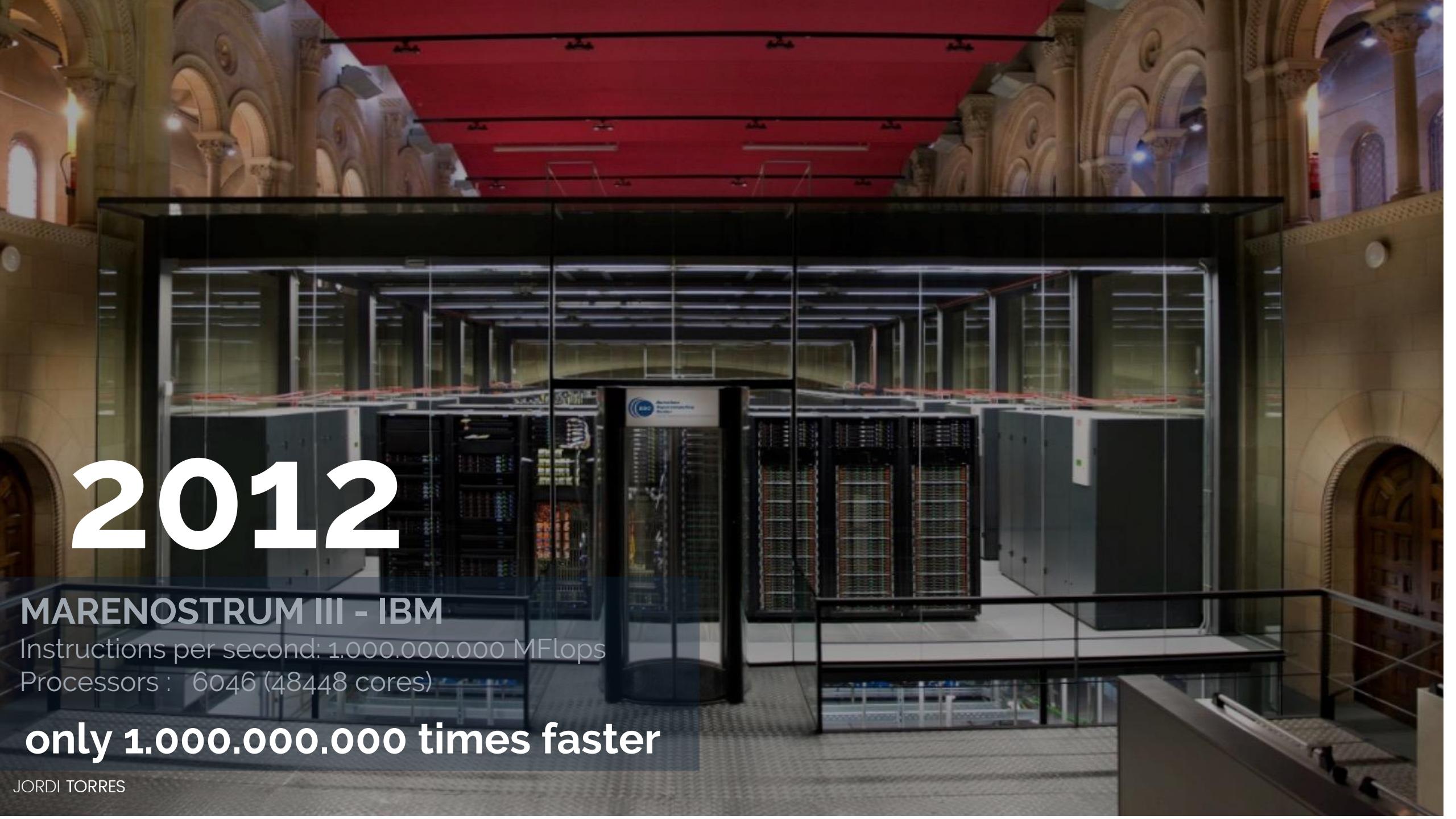
Processors : 1

# 2012

## MARENOSTRUM III - IBM

Instructions per second: 1.000.000.000 MFlops  
Processors : 6046 (48448 cores)





# 2012

## MARENOSTRUM III - IBM

Instructions per second: 1.000.000.000 MFlops

Processors : 6046 (48448 cores)

**only 1.000.000.000 times faster**

# CPU improvements!

Until then, the increase in computational power every decade of "my" computer, was mainly thanks to CPU



# CPU improvements!

Until then, the increase in computational power every decade of “my” computer, was mainly thanks to CPU

Since then, the increase in computational power for Deep Learning has not only been from CPU improvements . . .

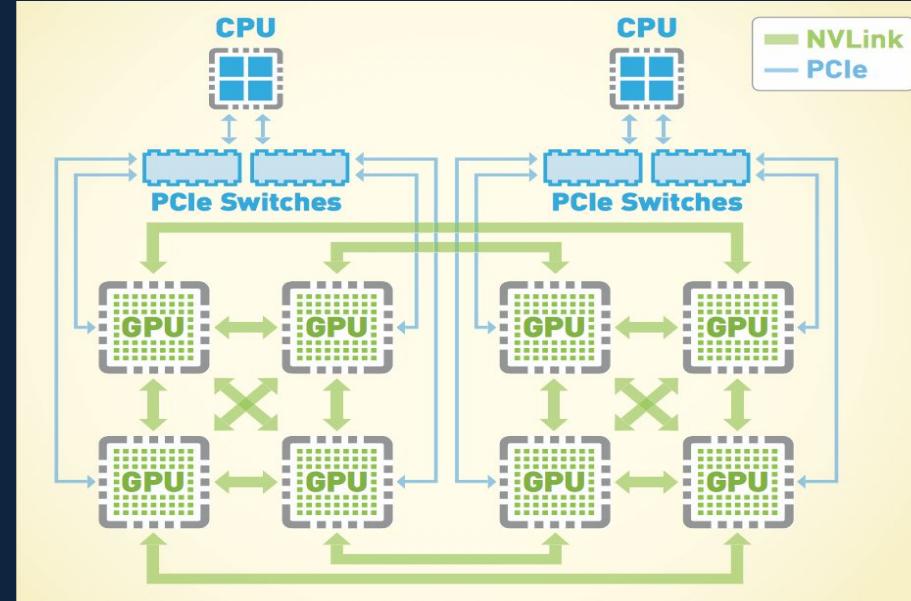




but also from the realization that GPUs (NVIDIA) were 20 to 50 times more efficient than traditional CPUs.

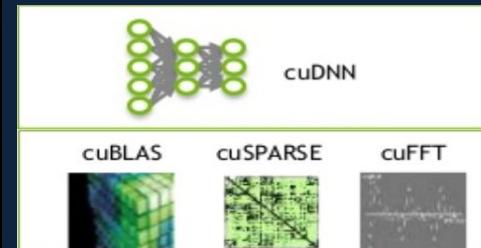
# Deep Learning requires computer architecture advancements

Fast tightly coupled network interfaces



Dense computer hardware

AI specific processors



Optimized libraries and kernels



**COMPUTING POWER**  
is the real enabler!



**Now we are entering into an era  
of computation democratization  
for companies !**

And what is “my/your” computer like now?

And what is “my/your” computer like now?



Source: <http://www.google.com/about/datacenters/gallery/images>

And what is “my/your” computer like now?



Source: <http://www.google.com/about/datacenters/gallery/images>

# Huge data centers!





28.000 m<sup>2</sup>



28.000 m<sup>2</sup>



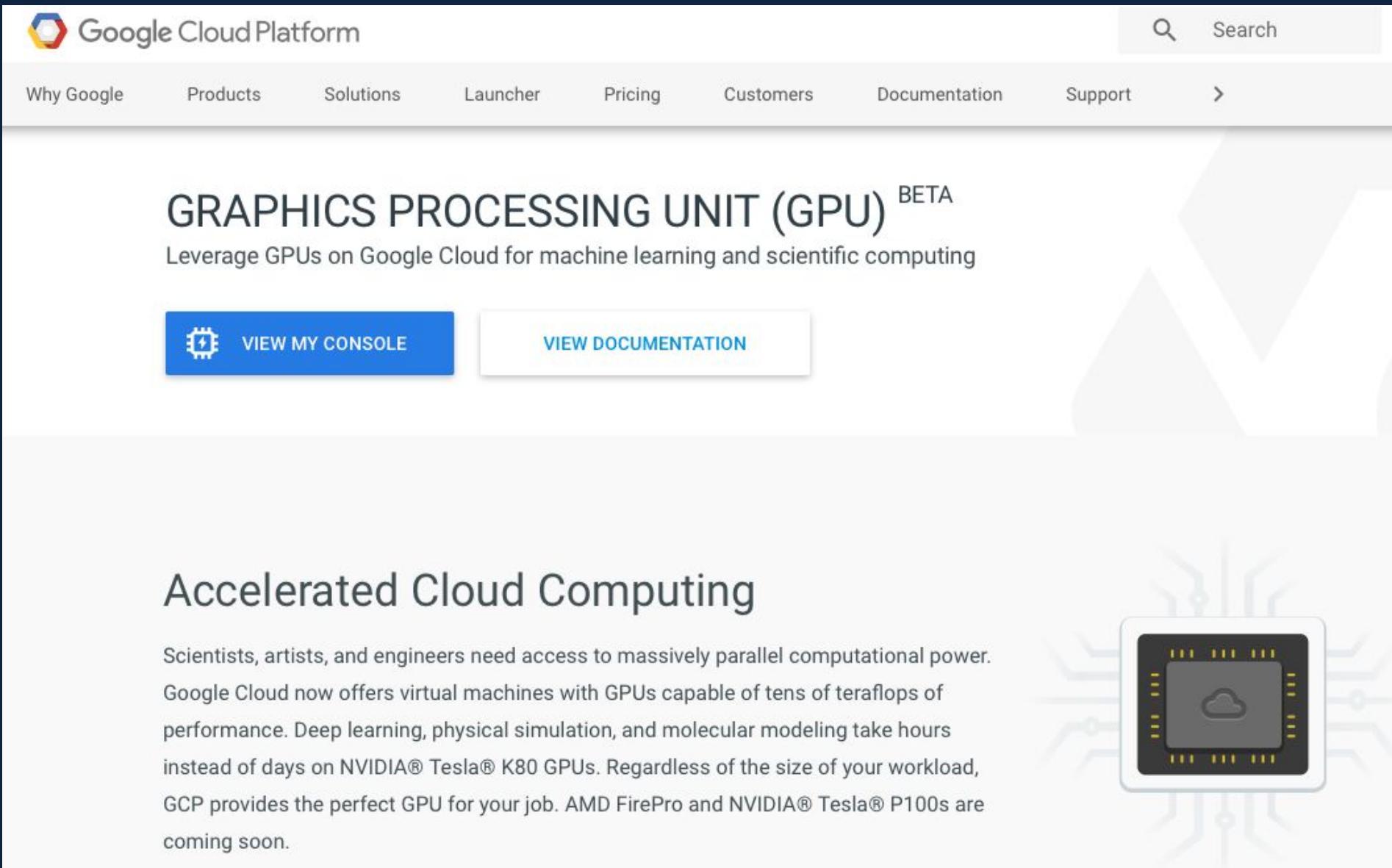
28.000 m<sup>2</sup>

For those (experts) who want to develop their own software, cloud services like Amazon Web Services provide GPU-driven deep-learning computation services

The screenshot shows a web browser displaying a blog post from the AWS Blog. The title of the post is "New P2 Instance Type for Amazon EC2 – Up to 16 GPUs". The post is by Jeff Barr and was published on 29 SEP 2016. It includes a table comparing three instance types: p2.xlarge, p2.8xlarge, and p2.16xlarge, based on various performance metrics.

Instance Name	GPU Count	vCPU Count	Memory	Parallel Processing Cores	GPU Memory	Network Performance
p2.xlarge	1	4	61 GiB	2,496	12 GiB	High
p2.8xlarge	8	32	488 GiB	19,968	96 GiB	10 Gigabit
p2.16xlarge	16	64	732 GiB	39,936	192 GiB	20 Gigabit

# And Google ...



The image shows the Google Cloud Platform GPU landing page. At the top, there's a navigation bar with links for Why Google, Products, Solutions, Launcher, Pricing, Customers, Documentation, Support, and a search bar. The main headline reads "GRAPHICS PROCESSING UNIT (GPU) BETA". Below it, a sub-headline says "Leverage GPUs on Google Cloud for machine learning and scientific computing". There are two buttons: "VIEW MY CONSOLE" (blue background) and "VIEW DOCUMENTATION" (white background). The central text "Accelerated Cloud Computing" is followed by a detailed paragraph about GPU benefits. To the right, there's a graphic of a cloud icon inside a circuit board frame.

Google Cloud Platform

Why Google Products Solutions Launcher Pricing Customers Documentation Support >

Search

GRAPHICS PROCESSING UNIT (GPU) BETA

Leverage GPUs on Google Cloud for machine learning and scientific computing

 [VIEW MY CONSOLE](#) [VIEW DOCUMENTATION](#)

Accelerated Cloud Computing

Scientists, artists, and engineers need access to massively parallel computational power. Google Cloud now offers virtual machines with GPUs capable of tens of teraflops of performance. Deep learning, physical simulation, and molecular modeling take hours instead of days on NVIDIA® Tesla® K80 GPUs. Regardless of the size of your workload, GCP provides the perfect GPU for your job. AMD FirePro and NVIDIA® Tesla® P100s are coming soon.

# And Google ...

 Google Cloud Platform

Why Google Products Solutions Launcher Pricing Customers Documentation Support >

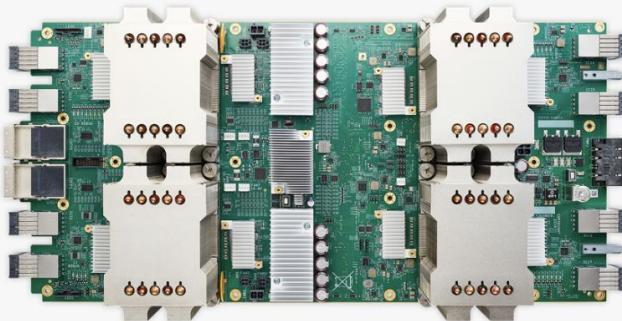
**CLOUD TPU ALPHA**

Train and run machine learning models faster than ever before

 SIGN UP TO LEARN MORE

## Accelerated Machine Learning

Machine learning (ML) has the power to greatly simplify our lives. Improvements in speech recognition and language understanding help all of us interact more naturally with technology. Businesses rely on ML to strengthen network security and reduce fraud. Advances in medical imaging enabled by ML can increase the accuracy of medical diagnoses and expand access to care, ultimately saving lives.



And all major cloud platforms...

Microsoft Azure

IBM Cloud

Aliyun

Cirrascale

NIMBIX

Outscale

...

Cogeco Peer 1

Penguin Computing

RapidSwitch

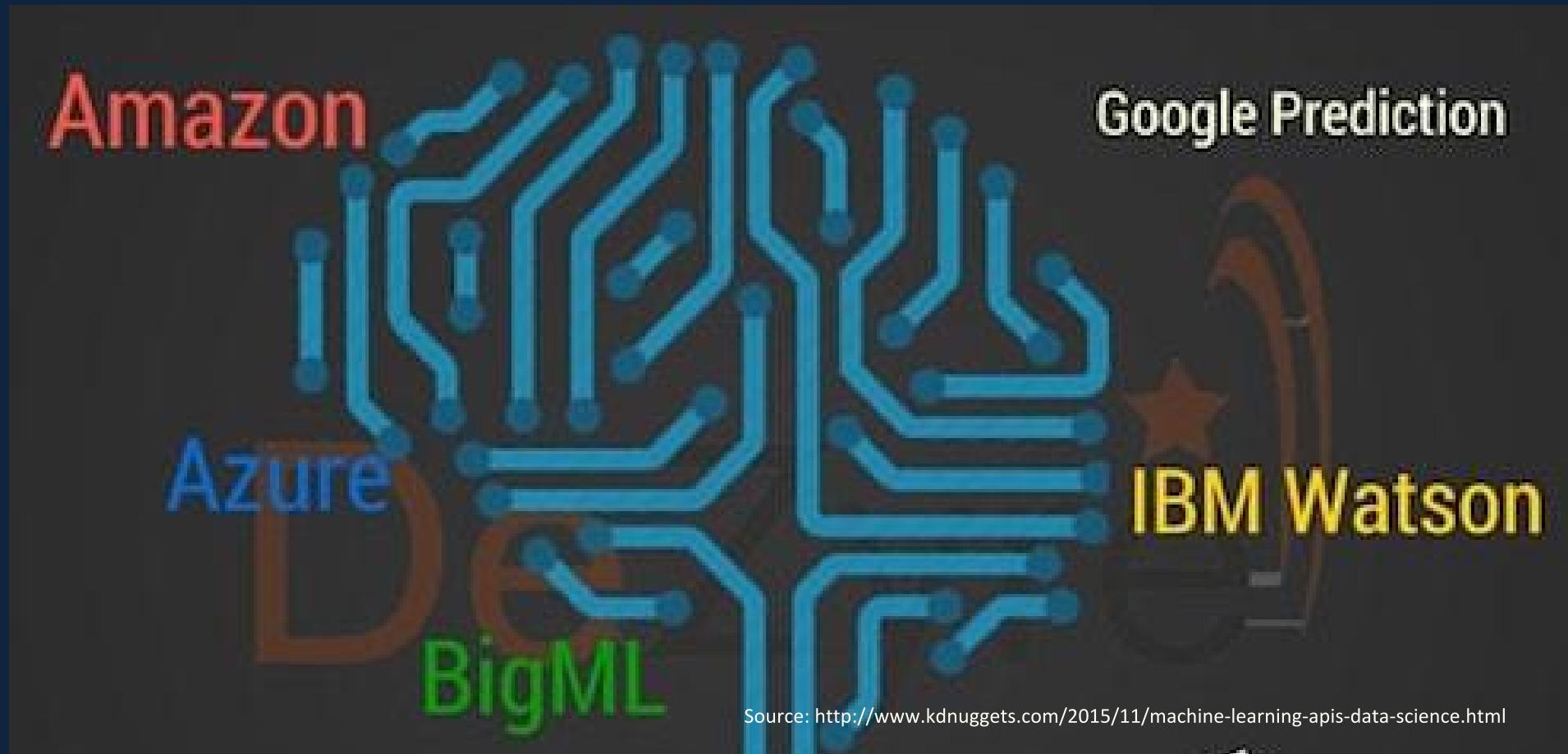
Rescale

SkyScale

SoftLayer

...

And for “less expert” people, various companies are providing a working scalable implementation of ML/AI algorithms as a Service (**AI-as-a-Service**)

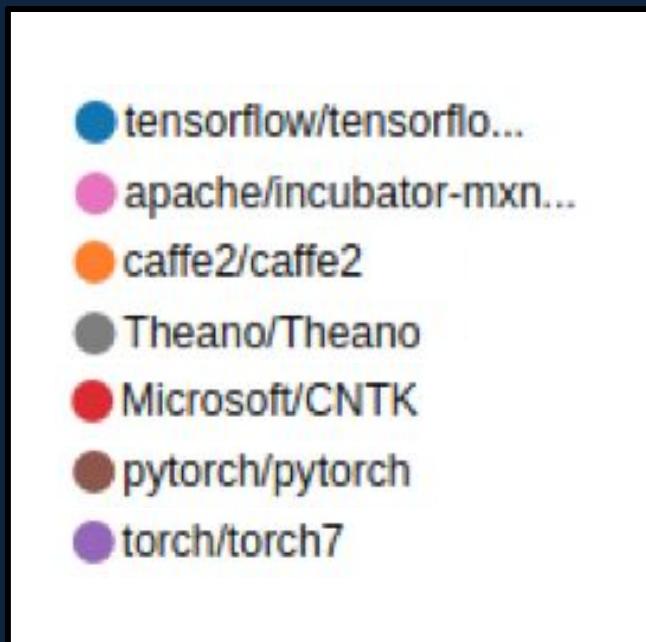


An open-source world for the  
Deep Learning community

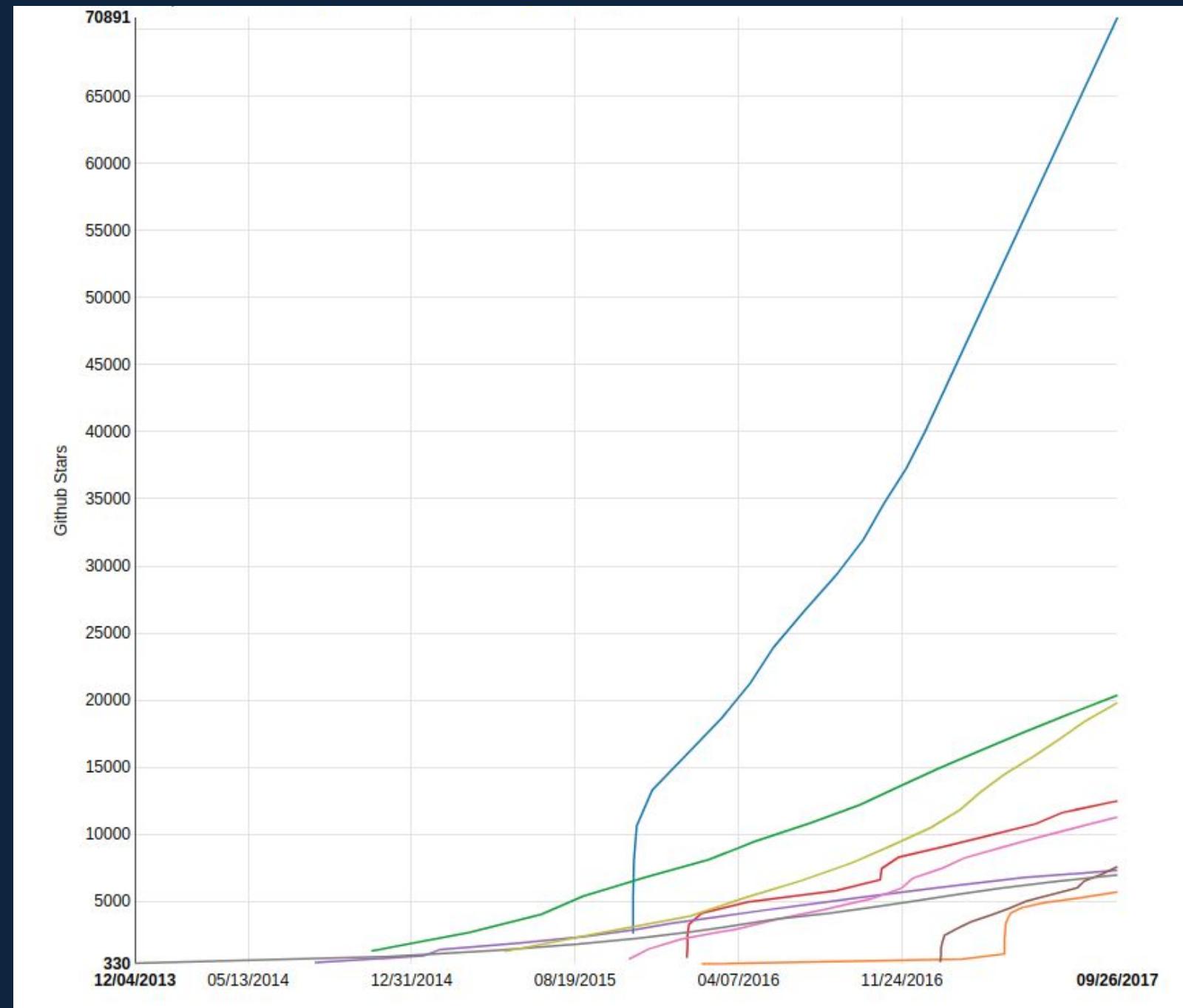
Many **open-source DL software**  
have greased the innovation process



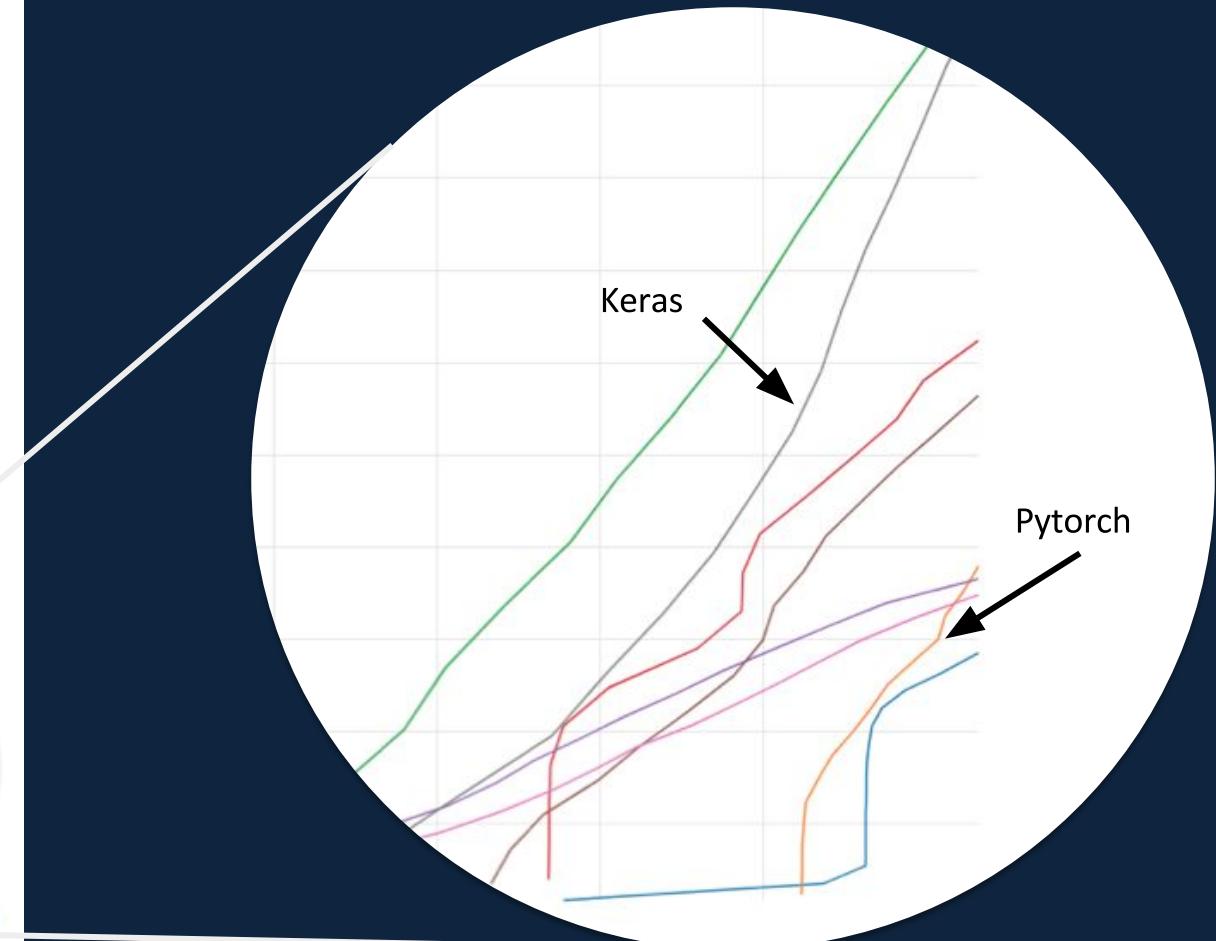
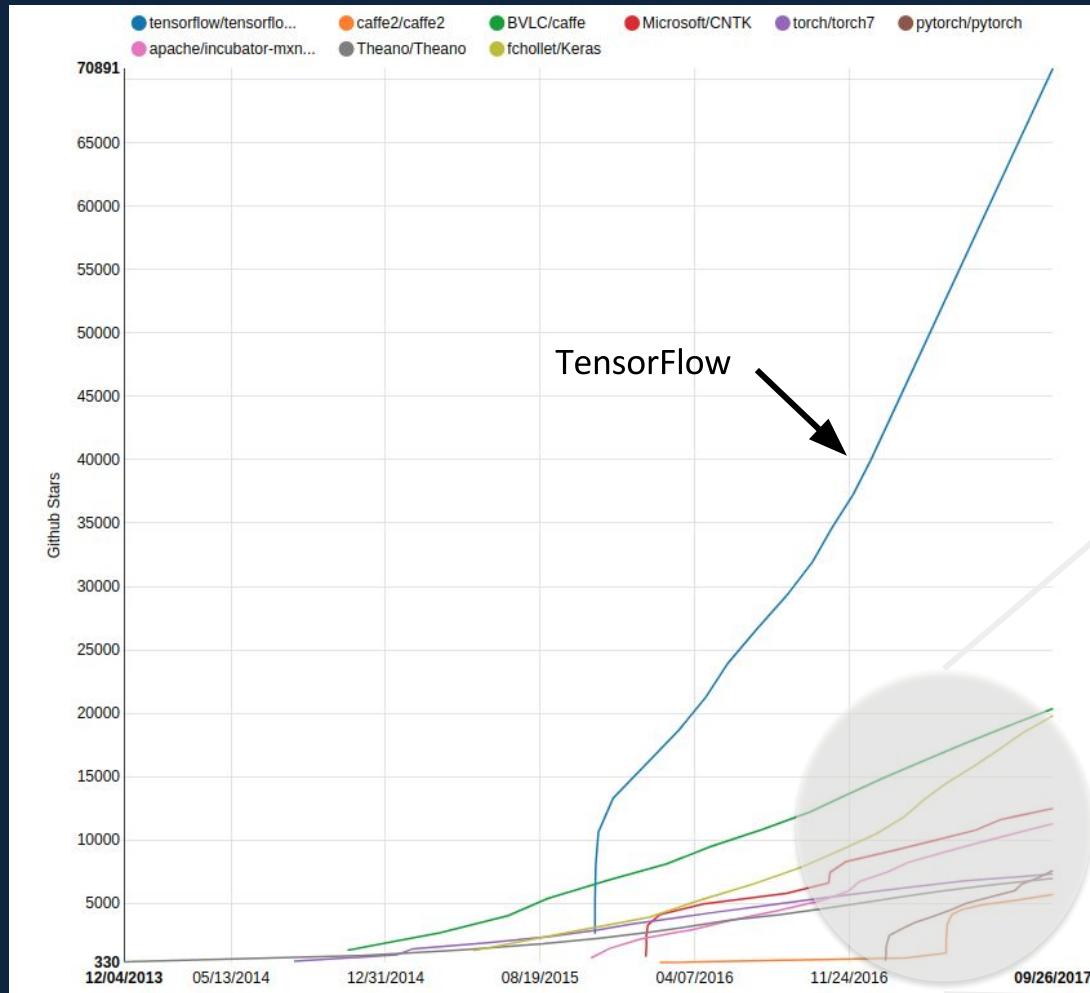
# Github Stars



source: Francesc Sastre



# In this course: we will consider the 3 frameworks with steepest gradient



frameworks with more slope

and no less important, **an open-publication ethic**, whereby many researchers publish their results immediately on a database without awaiting peer-review approval.

The screenshot shows a web browser window with the following details:

- Title Bar:** [1611.10012] Speed/accuracy X
- Address Bar:** https://arxiv.org/abs/1611.10012
- Header:** Cornell University Library
- Breadcrumbs:** arXiv.org > cs > arXiv:1611.10012
- Category:** Computer Science > Computer Vision and Pattern Recognition
- Title:** Speed/accuracy trade-offs for modern convolutional object detectors
- Authors:** Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, Kevin Murphy
- Submission Date:** (Submitted on 30 Nov 2016)
- Abstract:** In this paper, we study the trade-off between accuracy and speed when building an object detection system based on convolutional neural networks. We consider three main families of detectors --- Faster R-CNN, R-FCN and SSD --- which we view as "meta-architectures". Each of these can be combined with different kinds of feature extractors, such as VGG, Inception or ResNet. In addition, we can vary other parameters, such as the image resolution, and the number of box proposals. We develop a unified framework (in Tensorflow) that enables us to perform a fair comparison between all of these variants. We analyze the performance of many different previously published model combinations, as well as some novel ones, and thus identify a set of models which achieve different points on the speed-accuracy tradeoff curve, ranging from fast models, suitable for use on a mobile phone, to a much slower model that achieves a new state of the art on the COCO detection challenge.
- Comments:** A version of this paper is currently under submission to CVPR 2017
- Subjects:** Computer Vision and Pattern Recognition (cs.CV)
- Cite as:** arXiv:1611.10012 [cs.CV]

JORDI TORRES | FRANCESC SASTRE

