Fascinating Tales of a Strange Tomorrow

Julien Simon, Principal Technical Evangelist
Amazon Web Services

julsimon@amazon.fr
@julsimon

1956

Dartmouth Summer Research Project

John McCarthy (1927-2011)
1956 - Coined the term “Artificial Intelligence”
1958 - Invented LISP
1971 - Received the Turing Award

Forbidden Planet

Robbie the Robot

Gazing into the crystal ball

- 1958 Herbert Simon and Allen Newell
  "Within 10 years a digital computer will be the world's chess champion"
- 1965 Herbert Simon
  "Machines will be capable, within 20 years, of doing any work a man can do"
- 1967 Marvin Minsky
  "Within a generation ...
  the problem of creating 'artificial intelligence' will substantially be solved."
- 1970 Marvin Minsky
  "In from 3 to 8 years we will have a machine with the general intelligence of an average human being"

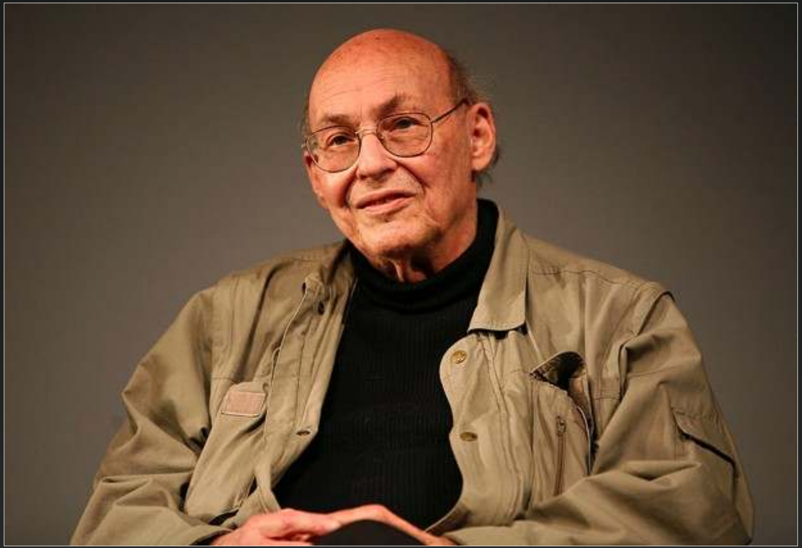
Herbert Simon (1916-2001)
1975 - Received the Turing Award
1978 - Received the Nobel Prize in Economics

Allen Newell (1927-1992)
1975 - Received the Turing Award

https://en.wikipedia.org/wiki/History_of_artificial_intelligence

It’s 2001. Where is HAL?

« No program today can distinguish a dog from a cat, or recognize objects in typical rooms, or answer questions that 4-year-olds can! »



Marvin Minsky (1927-2016)
1959 - Co-founded the MIT AI Lab
1968 - Advised Kubrick on “2001: A Space Odyssey”
1969 - Received the Turing Award

HAL 9000 (1992-2001)

Meanwhile, on the US West Coast…

no – not in Hollywood

Millions of users… Mountains of data… Commodity hardware…
Bright engineers… Need to make money….

Gasoline waiting for a match!

12/2004 - Google publishes Map Reduce paper

04/2006 - Hadoop 0.1

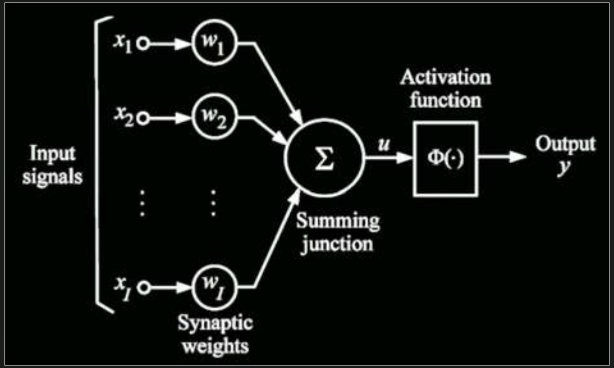
The rest is history

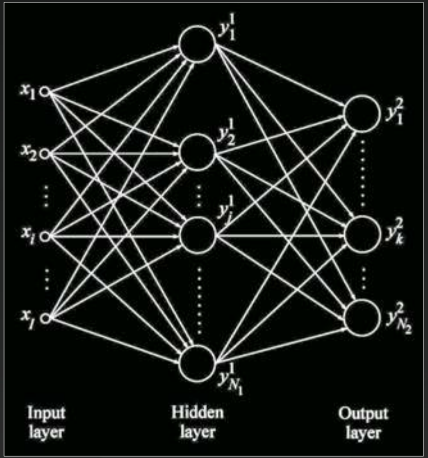
Fast forward a few years

- ML is now a commodity, but still no HAL in sight
- Traditional Machine Learning doesn't work well with problems where features can't be explicitly defined
- So what about solving tasks that are easy for people to perform but hard to describe formally?
- Is there a way to get informal knowledge into a computer?

Neural networks, revisited

- Universal approximation machine
- Through training, a neural network discovers features automatically
- Not new technology!
  - Perceptron - Rosenblatt, 1958
    image recognition, 20x20 pixels
  - Backpropagation - Werbos, 1975
- They failed back then because:
  - Data sets were too small
  - Solving larger problems with fully connected networks required too much memory and computing power, aka the Curse of Dimensionality

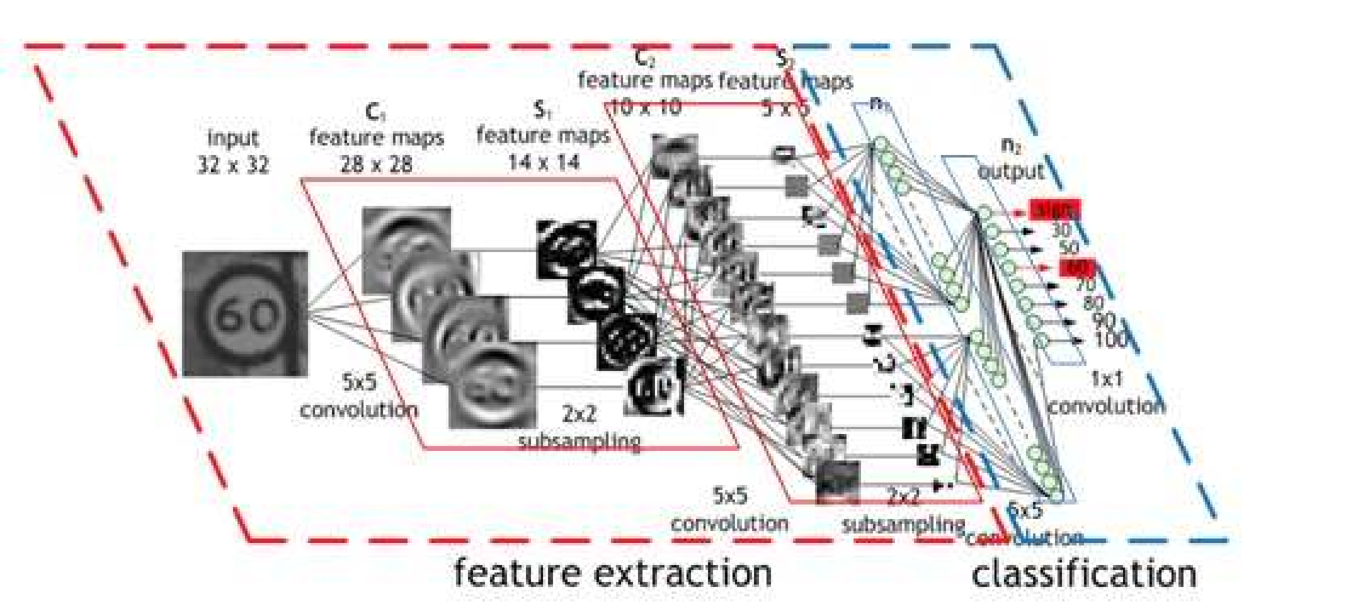




Breakthrough: Convolutional Neural Networks

Le Cun, 1998: handwritten digit recognition, 32x32 pixels

Feature extraction and downsampling allow smaller networks



https://devblogs.nvidia.com/parallelforall/deep-learning-nutshell-core-concepts/

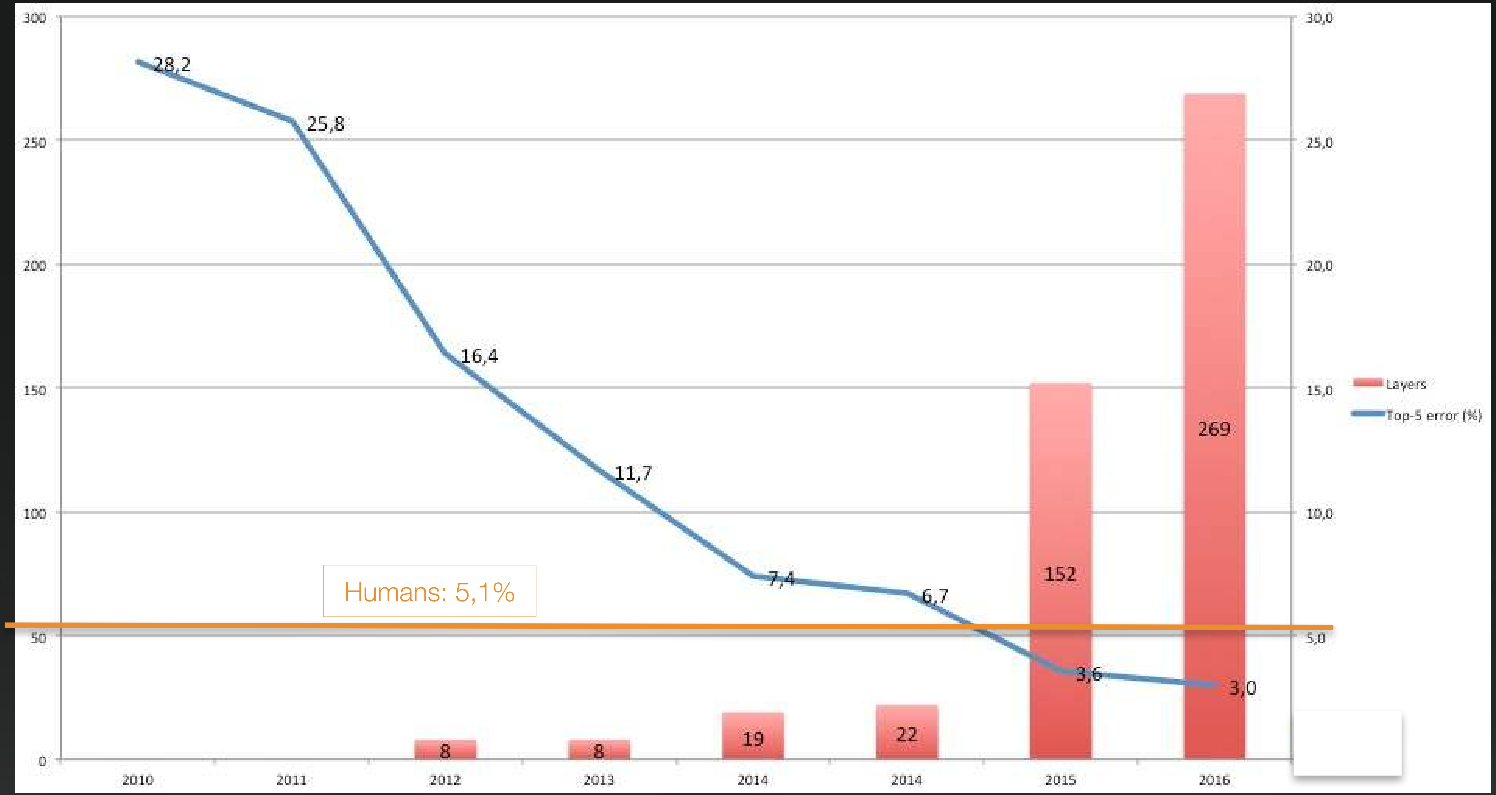
Why it is different this time

- Everything is digital: large data sets are available
  - Imagenet: 14M+ labeled images http://www.image-net.org/
  - YouTube-8M: 7M+ labeled videos https://research.google.com/youtube8m/
  - AWS public data sets: https://aws.amazon.com/public-datasets/
- The parallel computing power of GPUs make training possible
  - Simard et al (2005), Ciresan et al (2011)
  - State of the art networks have hundreds of layers
  - Baidu's Chinese speech recognition: 4TB of training data, +/- 10 Exaflops
- Cloud scalability and elasticity make training affordable
  - Grab a lot of resources for fast training, then release them
  - Using a DL model is lightweight: you can do it on a Raspberry Pi

https://www.nextplatform.com/2016/06/23/hpc-great-ai-supercomputing-stand-gain/

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Same breed?



http://image-net.org/challenges/LSVRC/

Deep Learning at the Edge

- Robots or autonomous cars can't exclusively rely on the Cloud
  - #1 issue: network availability, throughput and latency
  - Other issues: memory footprint, power consumption, form factor
  - Need for local, real-time inference (using a network with new data)

- Field Programmable Gate Array (FPGA)
  - Configurable and updatable to run all sorts of networks
  - Fast enough: DSP cells, Deep Compression (Son Han et al, 2017)
  - Low latency: on-board RAM with very high throughput
  - Better performance/power ratio than GPUs

https://www.xilinx.com/support/documentation/white_papers/wp486-deep-learning-int8.pdf
https://arxiv.org/pdf/1612.00694.pdf

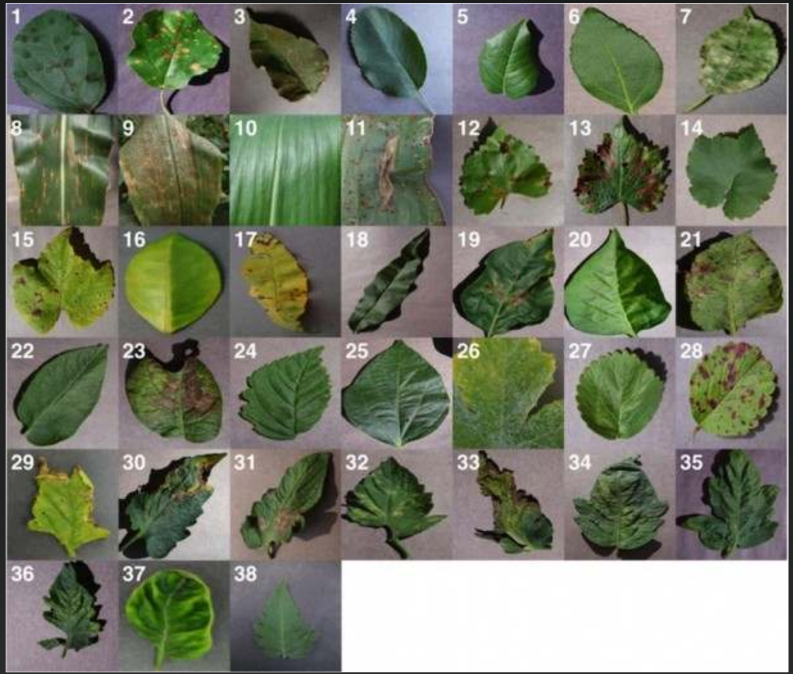
Let’s welcome our new Deep Learning Overlords

Flipping burgers

Flippy

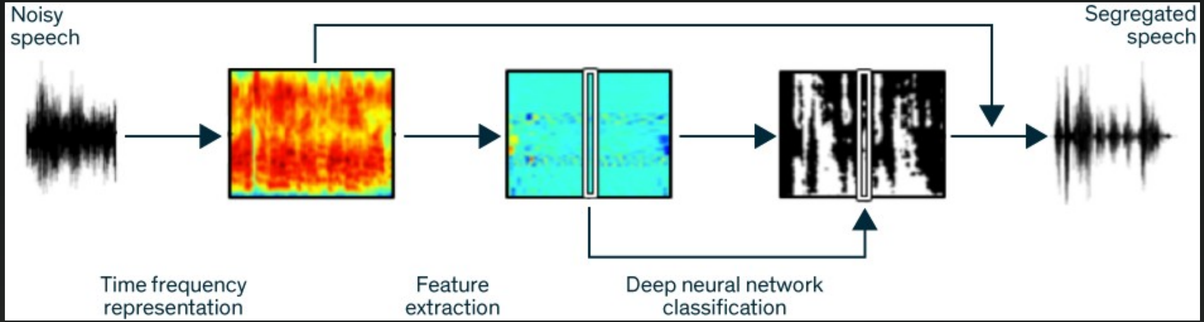
http://misorobotics.com/

Detecting plant diseases



https://blogs.nvidia.com/blog/2016/12/13/ai-fights-plant-disease/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5032846/

Improving hearing aids



http://spectrum.ieee.org/consumer-electronics/audiovideo/deep-learning-reinvents-the-hearing-aid

Amazon Echo

How AWS can help you build

Dive as deep as you need to

Control

High-level services – Amazon Rekognition, Polly, Lex

Platforms – Amazon EMR, Notebooks, Models, FGPA images

Libraries – mxnet, TensorFlow, Caffe, Torch, Theano

Amazon EC2 Instances – CPU, GPU, FPGA

Ease of use

Amazon EC2 Instances

- CPU
  - c5 family (coming soon), based on the Intel Skylake architecture
  - Elastic GPU (preview): on-demand GPU for traditional instances

- GPU
  - g2 and p2 families
  - p2.16xlarge
    - 16 GPUs (Nvidia GK210), 39936 CUDA cores, 23+ Tflops
    - Training a 10 Exaflops network: about 5 days, < $2000

- FPGA
  - f1 family (preview)
  - Up to 8 FPGAs per instance (Xilinx UltraScale Plus)

https://aws.amazon.com/about-aws/whats-new/2016/11/coming-soon-amazon-ec2-c5-instances-the-next-generation-of-compute-optimized-instances/
https://aws.amazon.com/blogs/aws/in-the-work-amazon-ec2-elastic-gpus/
https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/
https://aws.amazon.com/blogs/aws/developer-preview-ec2-instances-f1-with-programmable-hardware/

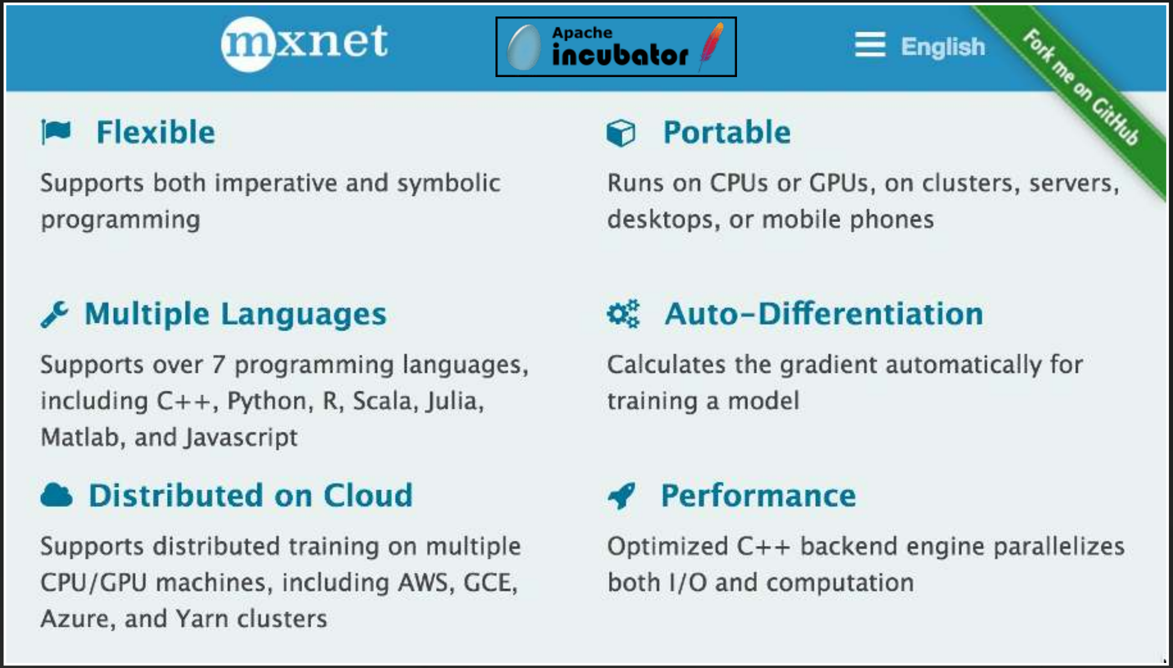
Amazon Machine Images

- Deep Learning AMI (Amazon Linux & Ubuntu)
  - Deep Learning Frameworks
    mxnet, Caffe, Tensorflow, Theano, and Torch, prebuilt and pre-installed
  - Other components
    Nvidia drivers, cuDNN, Anaconda, Python2 and Python3

- FPGA Developer AMI (Centos)
  - Xilinx FPGA simulation & synthesis tools: VHDL, Verilog, OpenCL
  - Software Development Kit: manage Amazon FPGA Images (AFI) on f1 instances
  - Hardware Development Kit: interface your application with AFIs

https://aws.amazon.com/marketplace/pp/B01M0AXXQB
https://aws.amazon.com/marketplace/pp/B06VVYBLZZ

mxnet



mxnet resources

http://mxnet.io/
https://github.com/dmlc/mxnet
https://github.com/dmlc/mxnet-notebooks

http://www.allthingsdistributed.com/2016/11/mxnet-default-framework-deep-learning-aws.ht

https://github.com/awslabs/deeplearning-cfn

Now the hard questions…

- Can my business benefit from Deep Learning?
- Should I design and train my own network?
  - Do I have the expertise?
  - Do I have enough time, data & compute to train it?
- Should I use a pre-trained network ?
  - How well does it fit my use case?
  - On what data was it trained?
- Should I use a high-level service?

- Same questions as Machine Learning years ago ☺

Science catching up with Fiction

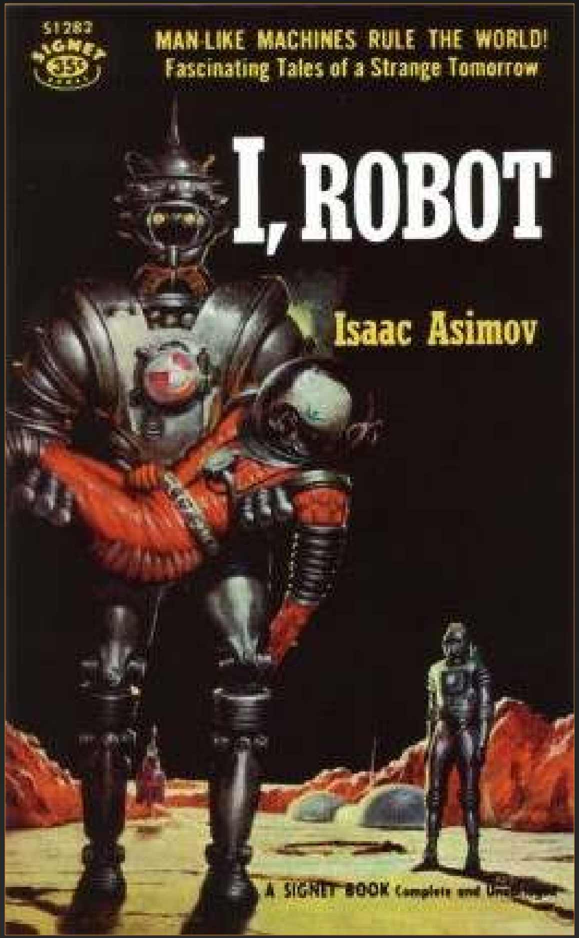
October 2014: Tesla Autopilot

October 2015: 30,000 robots in Amazon Fulfillment Centers

May 2016: AI defeats Lee Sedol, Go world champion

Still: “*The Best AI Still Flunks 8th Grade Science*”

https://www.wired.com/2016/02/the-best-ai-still-flunks-8th-grade-science/



Will machines learn how to understand humans – not the other way around?

Will they help humans understand each other?

Will they end up ruling the world?

Who knows?

Whatever happens, these will be fascinating tales of a strange tomorrow.

Thank you very much for your time!

julsimon@amazon.fr - @julsimon