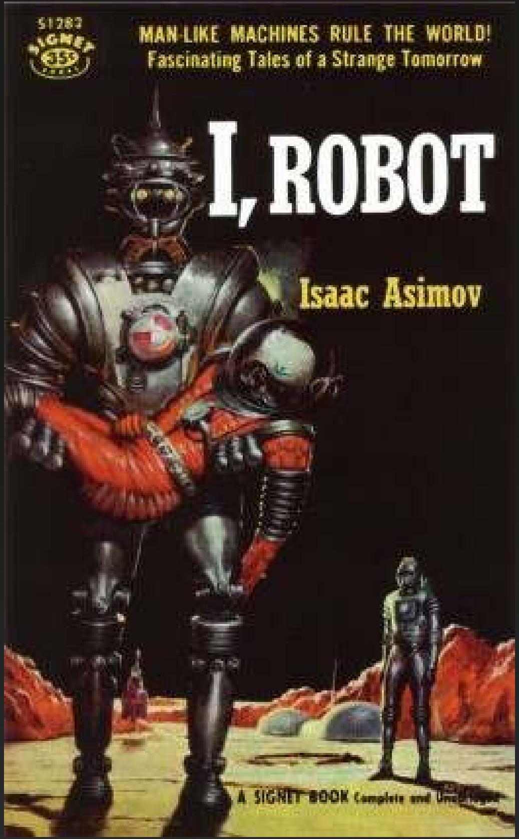
Fascinating Tales of a Strange Tomorrow

Julien Simon, Principal Technical Evangelist, AWS
julsimon@amazon.fr
@julsimon


51282
SIGNET
35¢
MAN-LIKE MACHINES RULE THE WORLD!
Fascinating Tales of a Strange Tomorrow
I, ROBOT
Isaac Asimov
A SIGNET BOOK Complete and Unabridged

1950

1956

Round 1: predictions, predictions…

- 1958, H. A. Simon and Allen Newell: “*within 10 years a digital computer will be the world's chess champion*” and “*within 10 years a digital computer will discover and prove an important new mathematical theorem*”
- 1965, H. A. 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*”

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

It did happen… eventually

May 1997: AI defeats chess world champion

May 2016: AI defeats go world champion

Still, not much came out of AI in the 60s-70s

- Combinatory explosion (exponential time)
- Not enough processing power
- The common sense issue
- “Toy” apps

→ Funding was cut: first AI Winter (1974)

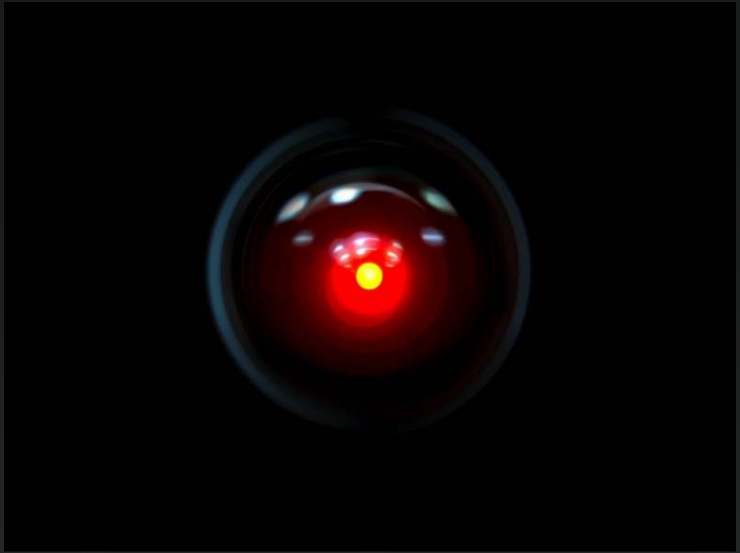
Round 2: LISP Machines (1980s)

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

- Implement LISP instructions with custom hardware
- Very expensive
- Fragmented market
- Wiped out by Moore’s Law and general-purpose workstations (Sun Microsystems etc.)

→ Second AI Winter

“It’s 2001. Where is HAL?” (Minsky)



Meanwhile, on the West Coast…

Google

amazon.com

YAHOO!

Millions of users… Tons of data… Commodity hardware…
Lots of engineers… Need to make money….

Gasoline waiting for a match!

Round 3: the Machine Learning explosion

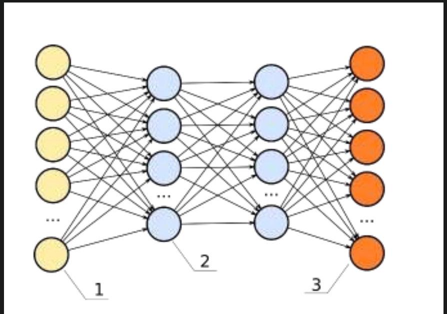
- 12/2004 - Google publishes Map Reduce paper
- 04/2006 - Hadoop 0.1
- 05/2009 – Yahoo sorts a Terabyte in 62 seconds
- Apache projects galore: Hive, Hbase, Spark, etc.

Fast forward 5 years

- ML is now a commodity. Great, but where *is* HAL?
  - Computer vision ?
  - Computer speech ?
  - Natural Language Processing ?
- Traditional Machine Learning doesn’t work well here
  - Training set
  - Features
- A third AI winter, then?

A Blast From The Past

Round 4: neural networks



- "Universal approximation machine" (Andrew Ng)
    - Artificial Intelligence is the New Electricity https://www.youtube.com/watch?v=21EiKfQYZXc
- Through training, a neural network self-organizes
- Patterns and features are discovered automatically
- Simple math, but it requires a lot of computing power
- The more data, the better (unlike traditional ML)

Wait a second, that’s not new at all

- Perceptron for pattern recognition (Rosenblatt, 1958)
- Backpropagation for faster training (Werbos, 1975)
- They failed back then because not enough computing power was available
- This has changed, hasn’t it?

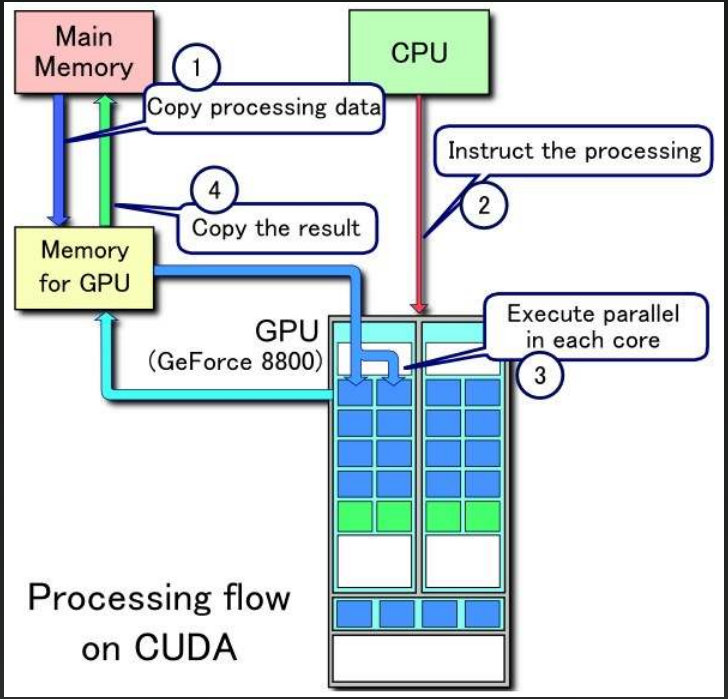
Scaling neural networks

- A neural network performs matrix operations
  - “Easy” to run in parallel, but scale is an issue
  - Product recommendation at Amazon.com: nb of users x nb products

- Deep learning requires many layers
  - Hundreds of layers
  - Training can last for weeks (the more data, the better)
  - That’s a insane amount of math operations

GPUs to the rescue

- General-purpose CPUs are not a good fit
  - We need thousands of cores
  - It would be impractical and expensive to use a huge number of general-purpose servers
- GPUs have been built for math
  - Nvidia K80 GPU: 4,992 cores
  - Multiple GPUs can collaborate inside the same server

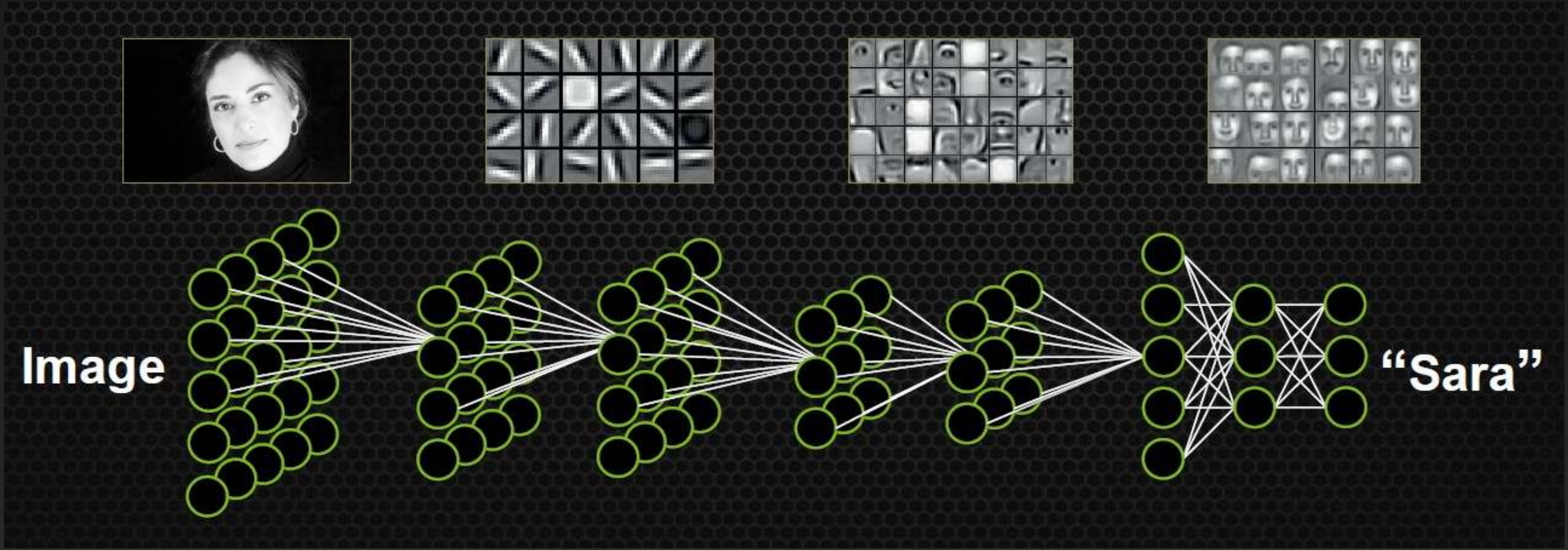


Cloud Computing to the rescue

- Training neural networks requires two things
  - Acquiring and storing lots of data (Petabyte-scale)
  - Running code of lots of GPUs
- Using neural networks requires very little
  - You can run a DL model on a Raspberry Pi!
- Scalability and elasticity are key assets here
  - Use a lot of resources, then release them
  - Pay only for what you need
  - Enjoy the latest GPU technology as it becomes available

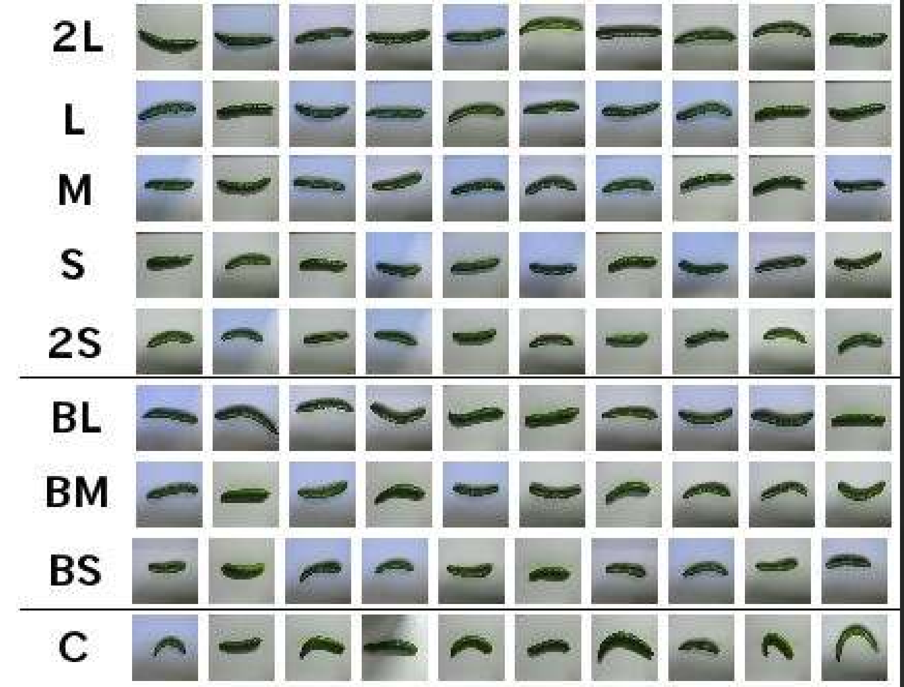
I for one welcome our new DL Overlords

Detecting patterns with neural networks



https://devblogs.nvidia.com/parallelforall/accelerate-machine-learning-cudnn-deep-neural-network-library/

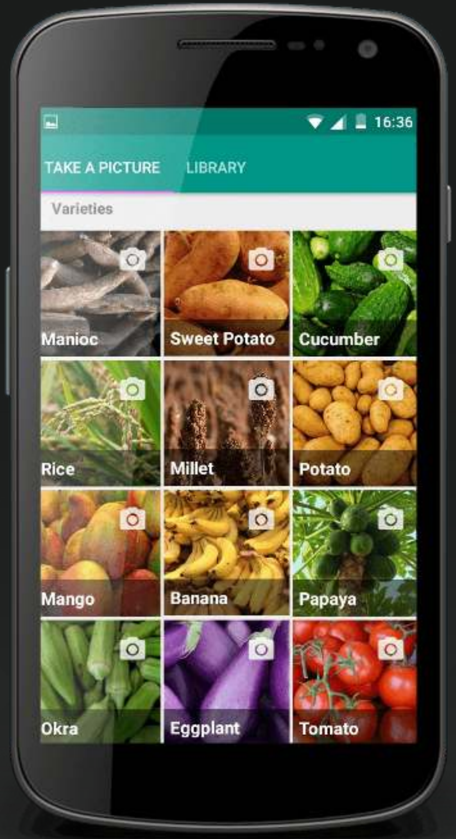
Using Deep Learning to sort cucumbers in Japan



https://cloud.google.com/blog/big-data/2016/08/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow

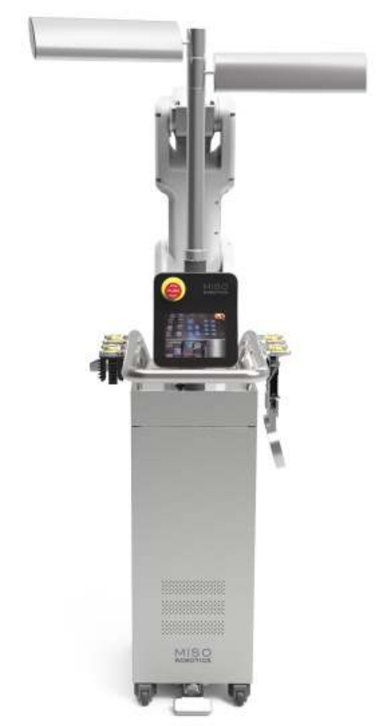
Detecting plant diseases

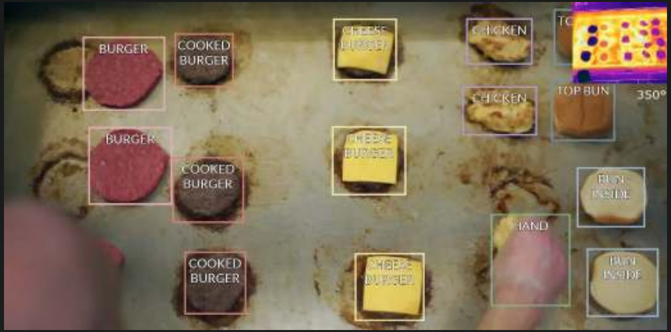
- Mobile application
- Model training on GPUs
- 60 pests identified with 90% accuracy
- Information, advice for treatment, etc.



https://blogs.nvidia.com/blog/2016/12/13/ai-fights-plant-disease/

Flippy, your new burger-flipping buddy





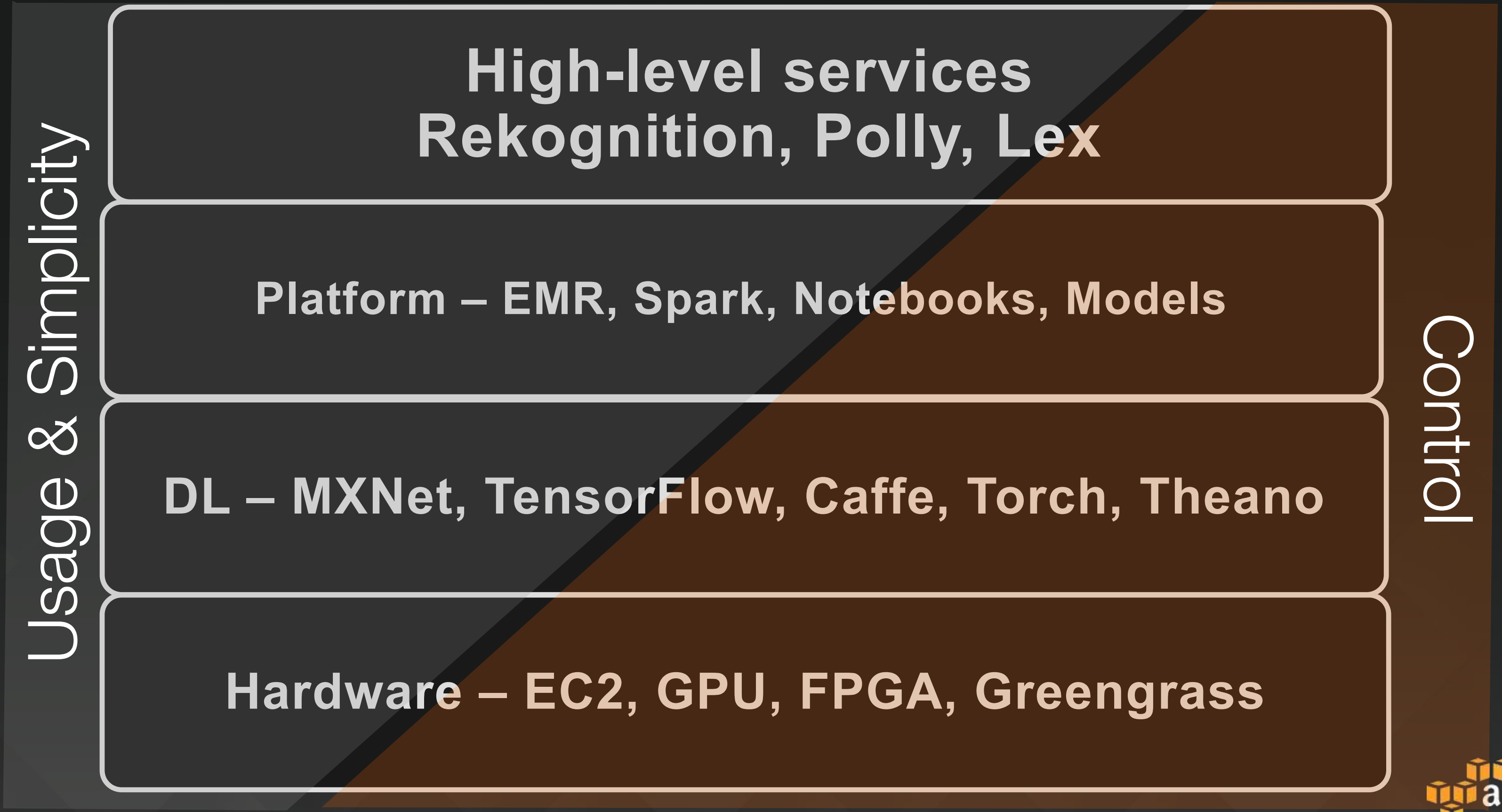
http://misorobotics.com/

Amazon Go

P
3
BBC
NEW
hue
PHILIPS
a
WIKIPE
The Free Encycl
IF
pr
ews
The Economist
wink
we
mo
amazon

Now what?

Dive as deep as you need to (but no more)



Usage & Simplicity

High-level services
Rekognition, Polly, Lex

Platform – EMR, Spark, Notebooks, Models

DL – MXNet, TensorFlow, Caffe, Torch, Theano

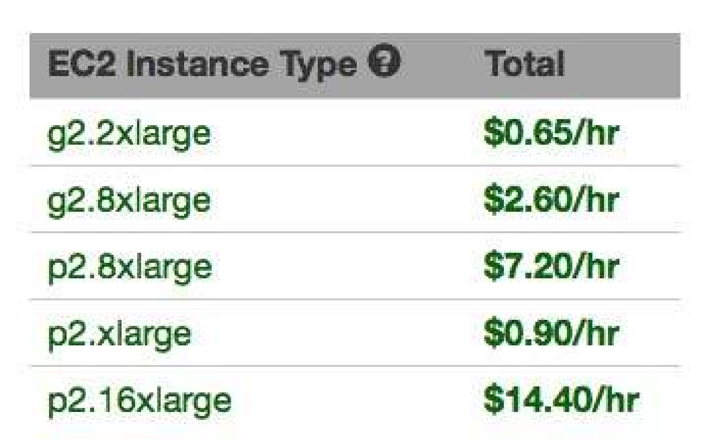
Hardware – EC2, GPU, FPGA, Greengrass

Control

AWS GPU Instances

- g2 (2xlarge, 8xlarge)
  - 32 vCPUs, 60 GB RAM
  - 4 NVIDIA K520 GPUs
  - 16 GB of GPU memory, 6144 CUDA cores

- p2 (xlarge, 8xlarge, 16xlarge)
  - Launched in 09/16
  - 64 vCPUs, 732 GB RAM
  - 16 NVIDIA GK210 GPUs
  - 192 GB of GPU memory, 39936 CUDA cores
  - 20 Gbit/s networking



EC2 Instance Type ?	Total
g2.2xlarge	$0.65/hr
g2.8xlarge	$2.60/hr
p2.8xlarge	$7.20/hr
p2.xlarge	$0.90/hr
p2.16xlarge	$14.40/hr

https://aws.amazon.com/blogs/aws/new-g2-instance-type-with-4x-more-gpu-power/
https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/
https://aws.amazon.com/ec2/Elastic-GPUs/

What about software?

- Nvidia CUDA (drivers & toolkit)
- Many ML/DL libraries support GPUs
  - Tensor Flow, Torch, Theano, Mxnet, etc.
- Setting all of this up is a little tricky

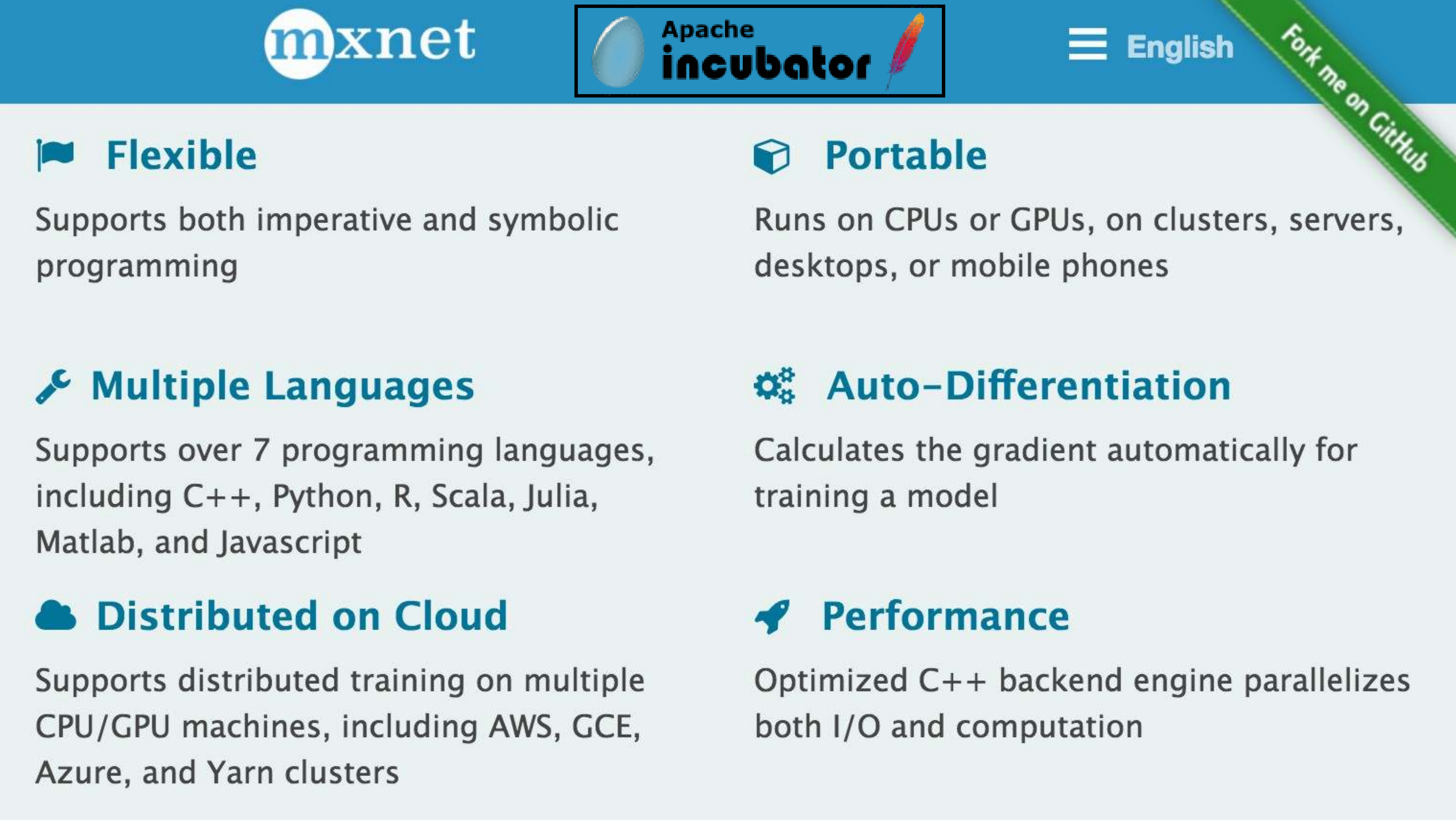
  → Deep Learning AMI

  → Fast.ai (great course) https://github.com/fastai/

AWS Deep Learning AMI

awsmarketplace

- Deep Learning Frameworks – 5 popular Deep Learning Frameworks (mxnet, Caffe, Tensorflow, Theano, and Torch) all prebuilt and pre-installed
- Pre-installed components – Nvidia drivers, cuDNN, Anaconda, Python2 and Python3
- AWS Integration – Packages and configurations that provide tight integration with Amazon Web Services like Amazon EFS (Elastic File System)
- Amazon Linux & Ubuntu



Flexible

Supports both imperative and symbolic programming

Portable

Runs on CPUs or GPUs, on clusters, servers, desktops, or mobile phones

Multiple Languages

Supports over 7 programming languages, including C++, Python, R, Scala, Julia, Matlab, and Javascript

Auto-Differentiation

Calculates the gradient automatically for training a model

Distributed on Cloud

Supports distributed training on multiple CPU/GPU machines, including AWS, GCE, Azure, and Yarn clusters

Performance

Optimized C++ backend engine parallelizes both I/O and computation

Now the hard questions…

- Should I build my own network?
  - Do I have the expertise?
  - Do I have enough time, data & compute to train it?
- Or should I reuse a pre-trained model?
  - How well does it fit my use case?
  - On what data was it trained?
- Or should I use a high-level service?

- Same questions as ML years ago… how did that work out?
- What do *you* think?

“Dogs vs Cats” demo

Keras on P2 instance

Amazon AI demo

Rekognition, Polly & Lex

Thank you!

Julien Simon, Principal Technical Evangelist, AWS
julsimon@amazon.fr
@julsimon