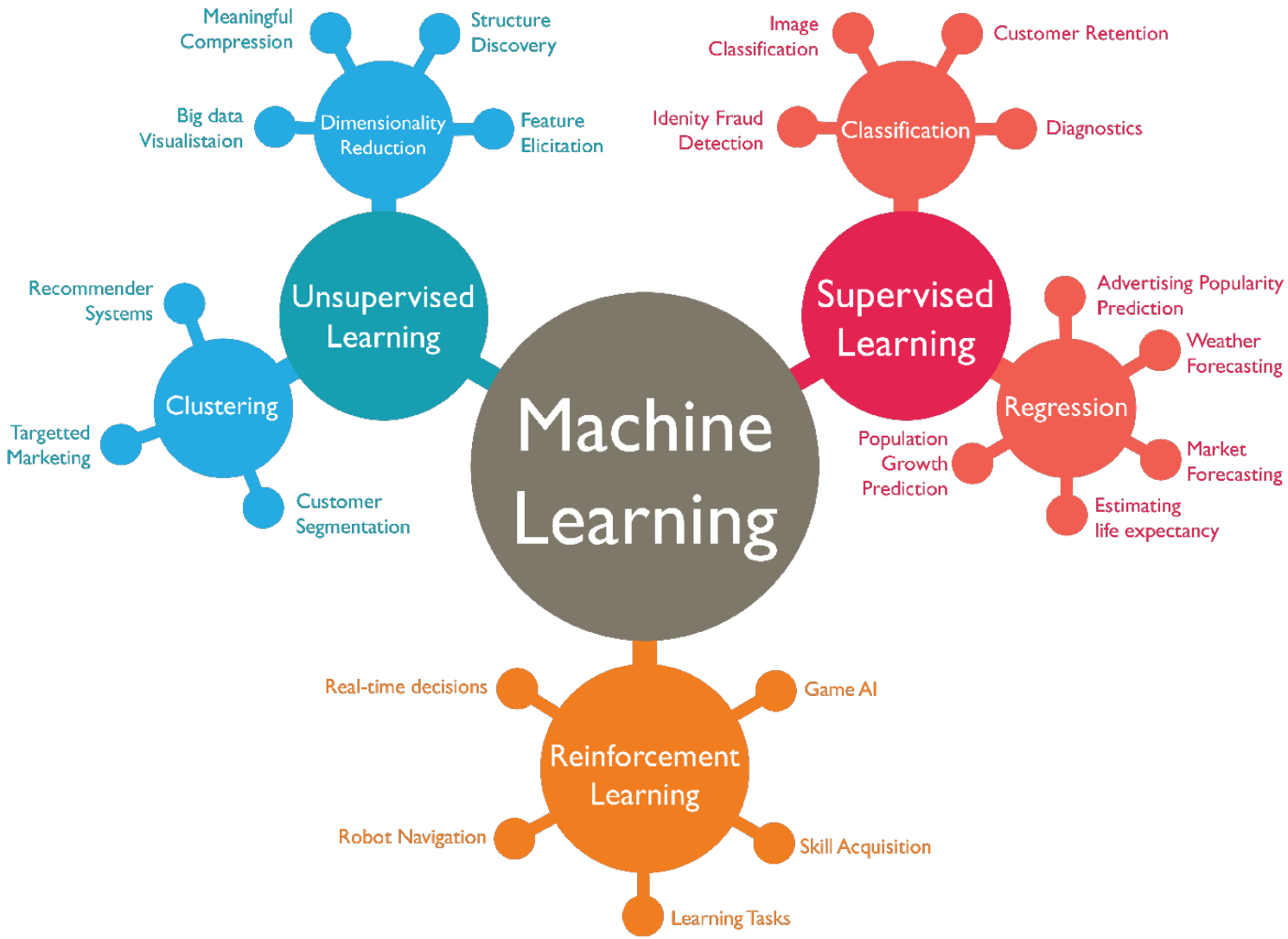
Matthew Fricke

Machine Learning at CARC



Other things I do… autonomous robotics

Melanie Moses BCLab

UNM Newsroom

About Us | News Topics | Schools and Colleges | For News Media

UNM Newsroom / News / UNM's VolCAN team makes history in Canary Islands

UNM's VolCAN team makes history in Canary Islands

Researchers use drones to sample gases from active La Palma volcano

By Kim Delker January 05, 2022

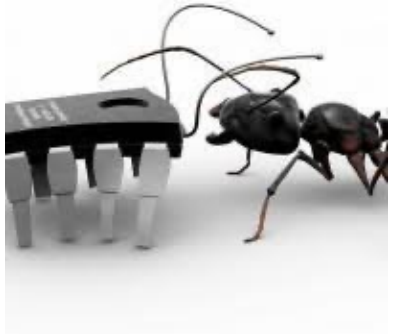
Relate

SCIENCE ADVANCES | RESEARCH ARTICLE

APPLIED SCIENCES AND ENGINEERING

Aerial strategies advance volcanic gas measurements at inaccessible, strongly degassing volcanoes

Other things I do… autonomous robotics

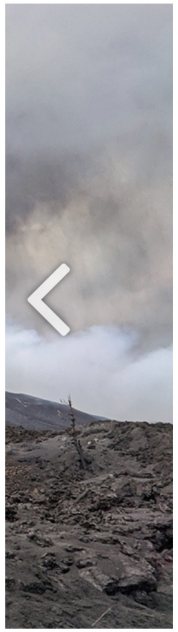


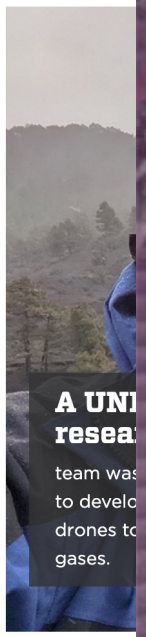
lanie Moses BCLab

UNM Newsroom

About Us News Topics Schools and Colleges For News Media

UNM Newsroom / News / UNM's V





A UNI
resea
team was
to develo
drones to
gases.

NASA MINDS ChiliHouse – Lunar Gardener

UNM's VolCAN
Islands

Researchers use drones to sample gases from active La Palma volcano

By Kim Delker January 05, 2022

Aerial strategies advance volcanic gas measurements at inaccessible, strongly degassing volcanoes

Computational Immunology

Melanie Moses BCLab

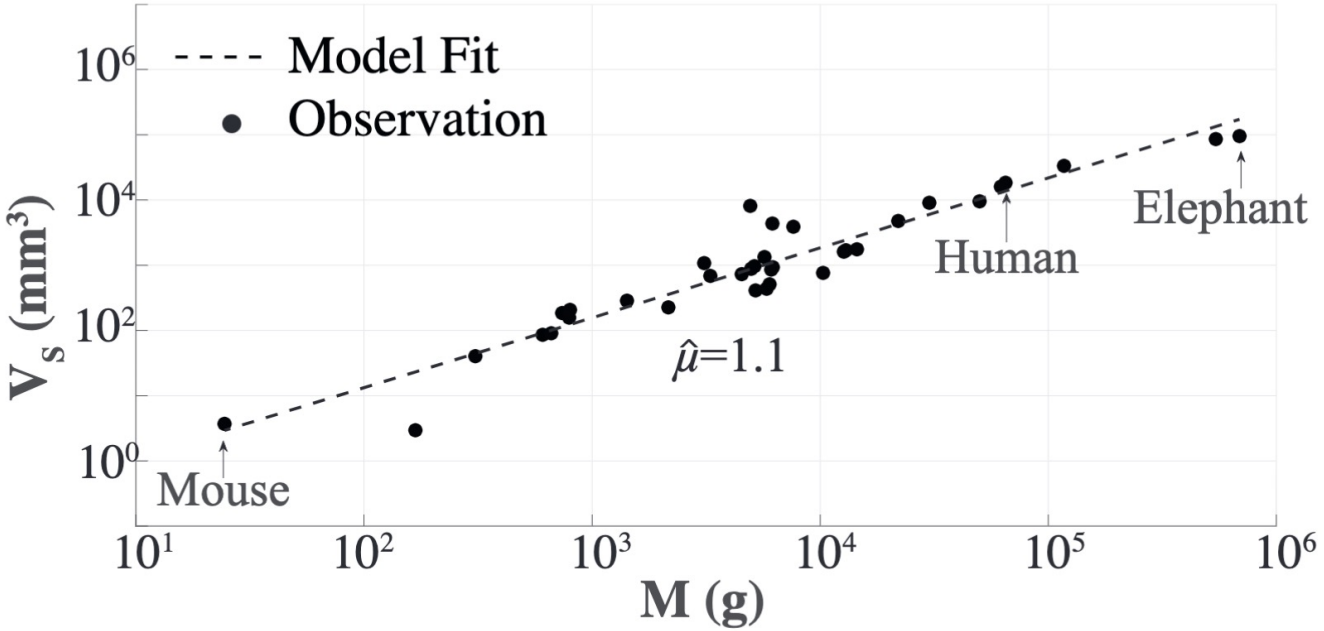


Fig. 1 Spleen Volume Scaling with Animal Mass. Species shown on log-log axes. The dashed line indicates the best model fit. The best-fit exponent for spleens (36 species) is $\hat{\mu} = 1.1$ 95% CI [0.95, 1.2]. $R^2 = 0.91$.

The Sizes and Numbers of Lymph Nodes Support a Scalable Immune Response in Mammals

Jannatul Ferdous[1*], G. Matthew Fricke[1,2†], Judy L. Cannon[3†] and Melanie E. Moses[1,4,5†]

[1]Department of Computer Science.
[2]Center for Advanced Research Computing.
[3]Molecular Genetics and Microbiology.
[4]Biology Department, The University of New Mexico, Albuquerque, USA.
[5]Santa Fe Institute, Santa Fe, USA.

*Corresponding author(s). E-mail(s): jannat@unm.edu;
Contributing authors: mfricke@unm.edu; jucannon@salud.unm.edu; melaniem@unm.edu;
†These authors contributed equally to this work.

Or… How to Build a Scalable Immune System

Machine Learning

NFoLD
NETWORK FOR LIFE DETECTION

Melanie Moses BCLab

INTERFACE

rsif.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:

Agnostic Polymer Detection using Mass Spectrometry using Machine Learning

Victoria Da Poian[*,1,2], Lu Chou[*,1,3], Natalie Grefenstette[*,4,5], G. Matthew Fricke[6], and Christopher P. Kempes[4]

[1]NASA Goddard Space Flight Center, Greenbelt, USA, [2]Microtel LLC, Greenbelt, USA, [3]Georgetown University, Washington DC, USA, [4]Santa Fe Institute, Santa Fe NM, USA, [5]Blue Marble Space Institute of Science, Seattle WA, USA, [6]Department of Computer Science, University of New Mexico, Albuquerque NM, USA.

ELSEVIER

Journal of Computational and Applied Mathematics

Volume 395, 15 October 2021, 113451

Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change

All these projects involved CARC computations.

Topics

- What is machine learning anyway?
- Why should you care?
- Why CARC?

- An Example ML Run at CARC, Monitoring GPU usage
 (and a case for Jupyter Notebooks)

Just from today’s talks…

- Jeremy Hogeveen (Neural Networks)
- Susan Atlas (Neural Networks)
- Sam McKenzie (K-means clustering)
- Tameem Talbash (Neural Networks)
- And more …

Just from today's talks…

- Jeremy Hogeveen (Neural Networks)
- Susan Atlas (Neural Networks)
- Sam McKenzie (K-means clustering)
- Tameem Talbash (Neural Networks)
- And more …

But the GPUs on our Machine Learning Cluster are often idle.

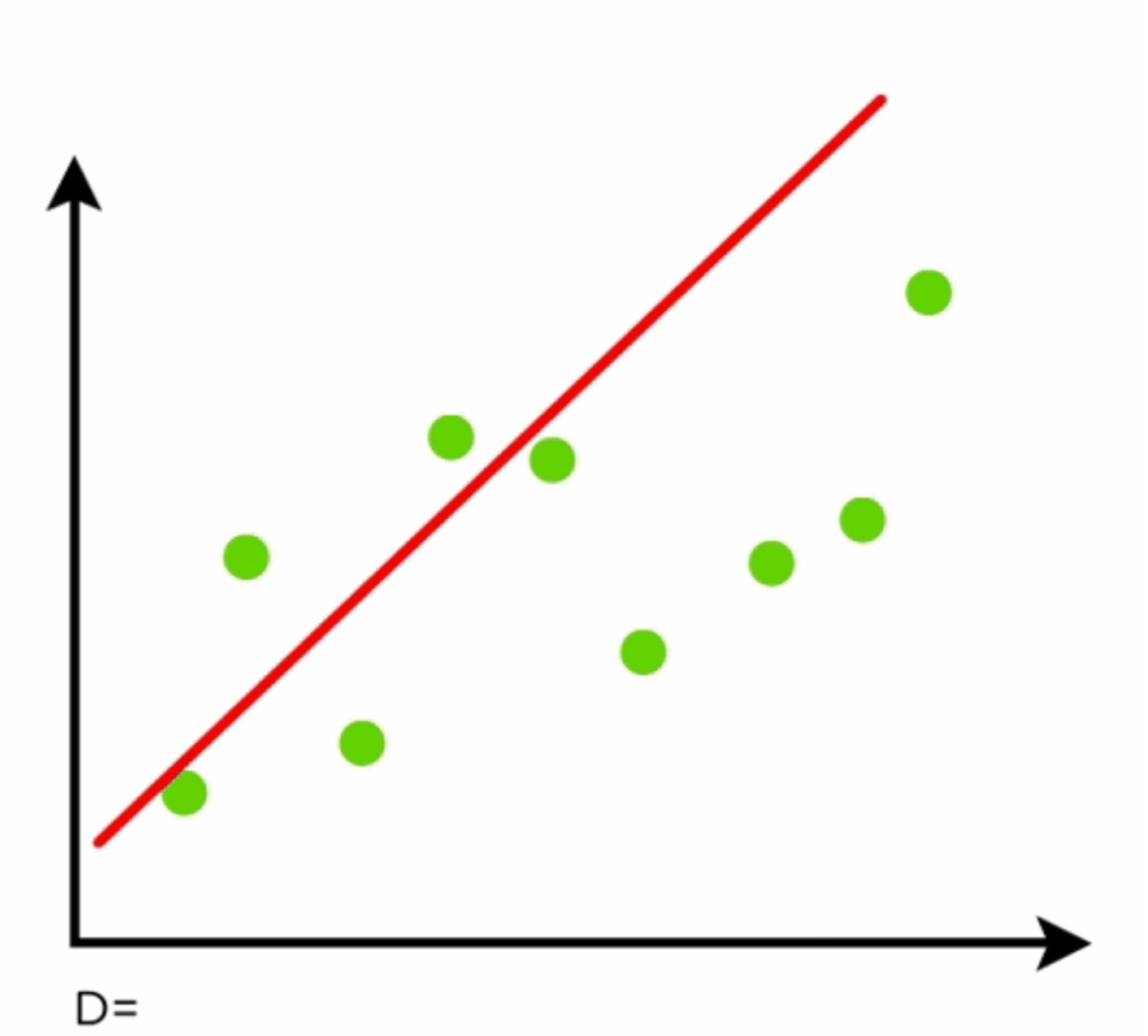
We need CARC ML users to capitalize on those GPUs.

Good for you! Lot's of free GPU resources!

Herbert Simon

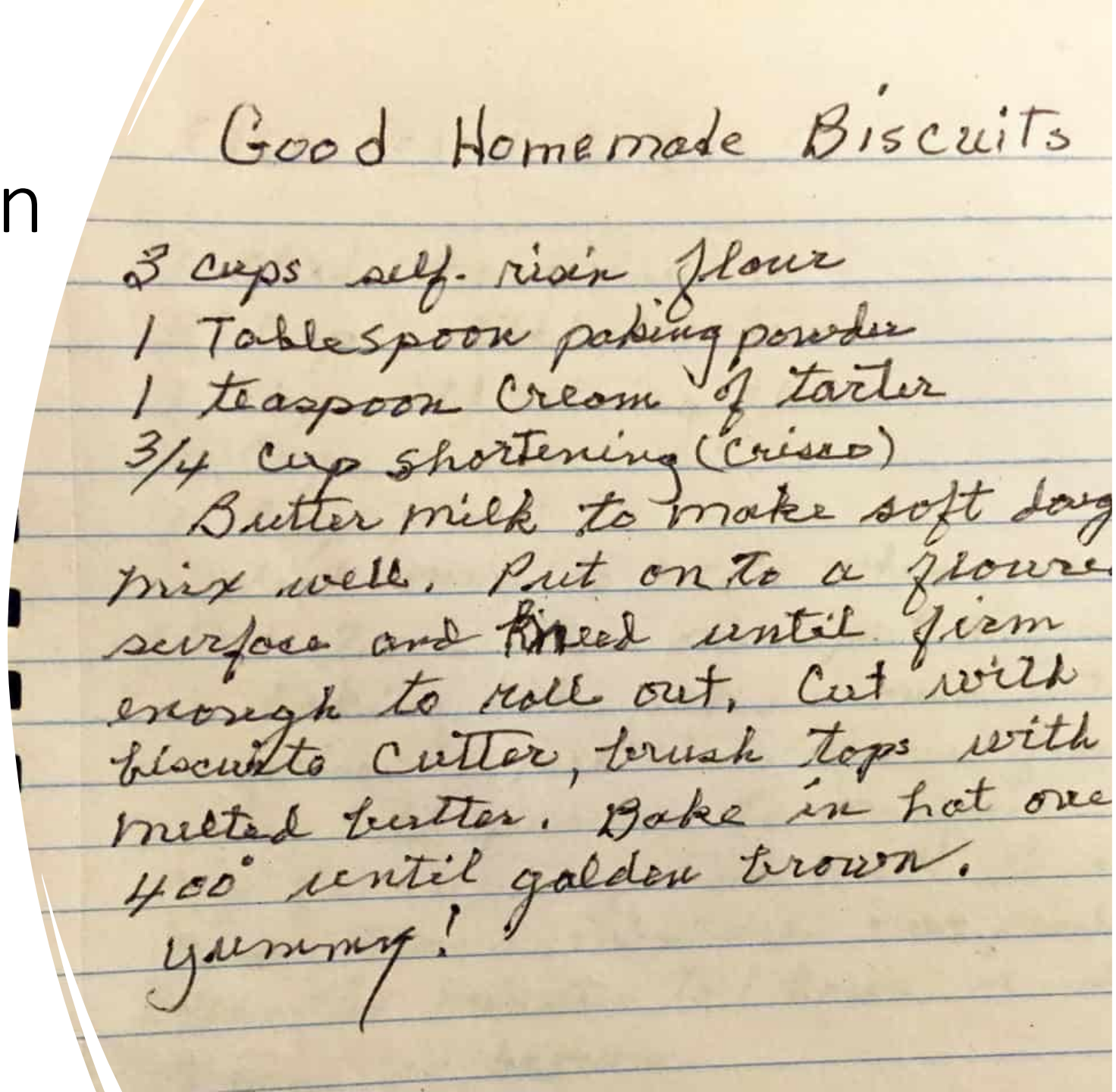
- “Learning is any process by which a system improves performance from experience”
- “Machine learning is concerned with computer programs that automatically improve their performance through experience”
- Professor of Psychology and Computer Science at Carnegie Mellon. Turing award and Nobel winner.

Regression is Machine Learning

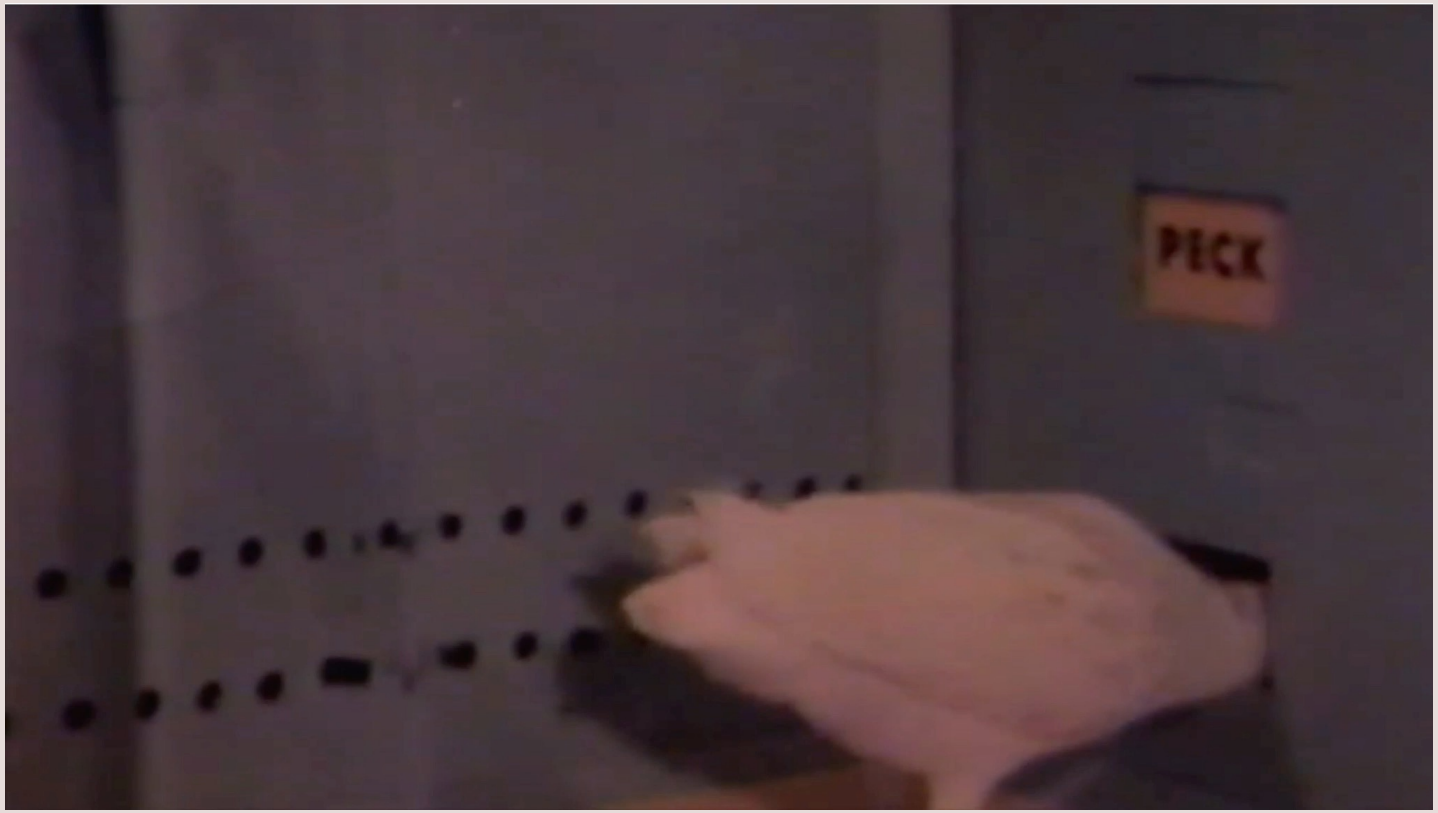


Procedure Driven vs Data Driven

- Procedural programming requires an explicit understanding of the steps that map input to output.



Procedure Driven vs Data Driven

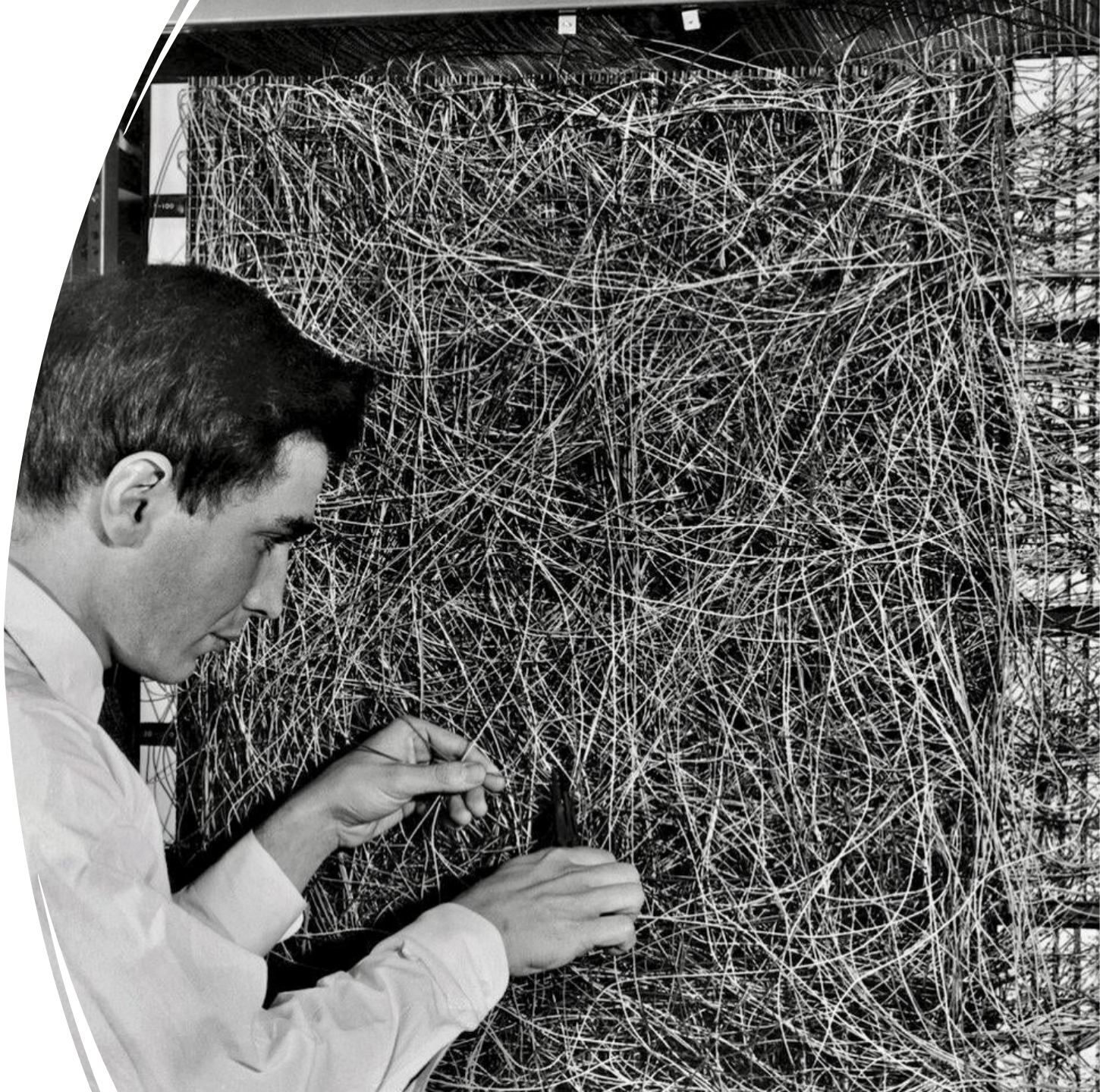


- A data driven program maps input to output based on lots of example inputs and outputs.
- B.F. Skinner 1940s, Psychologist “Operant Conditioning”
- Pigeon trained to classify visual input.
- Food as reward

Neural Networks

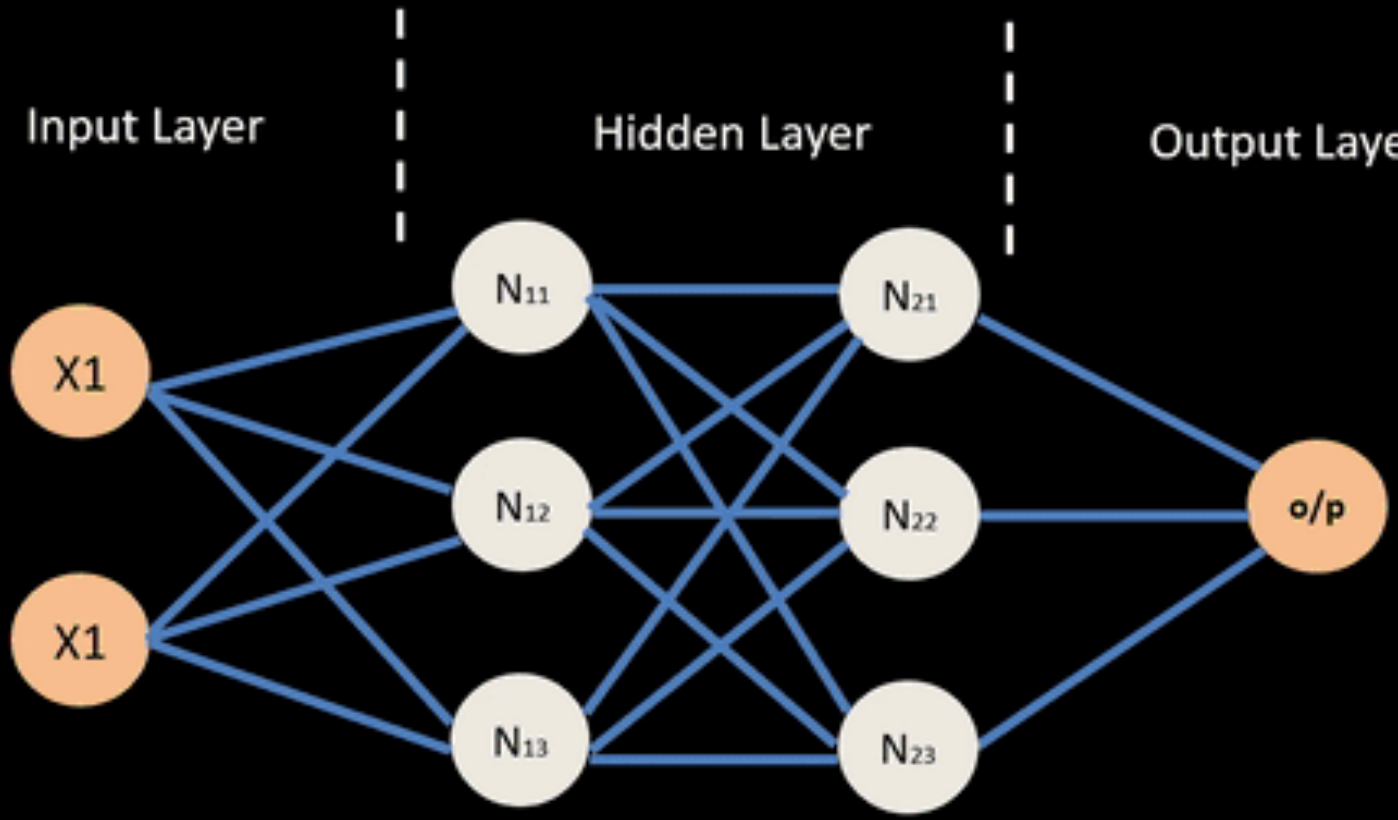
- First Perceptron created in 1958 by psychologist **Frank Rosenblatt**
- Seen here with a Mark I neural network.

- Rosenblatt referenced Skinner's work in several of his papers.



Neural Network – Backpropagation

INPUT

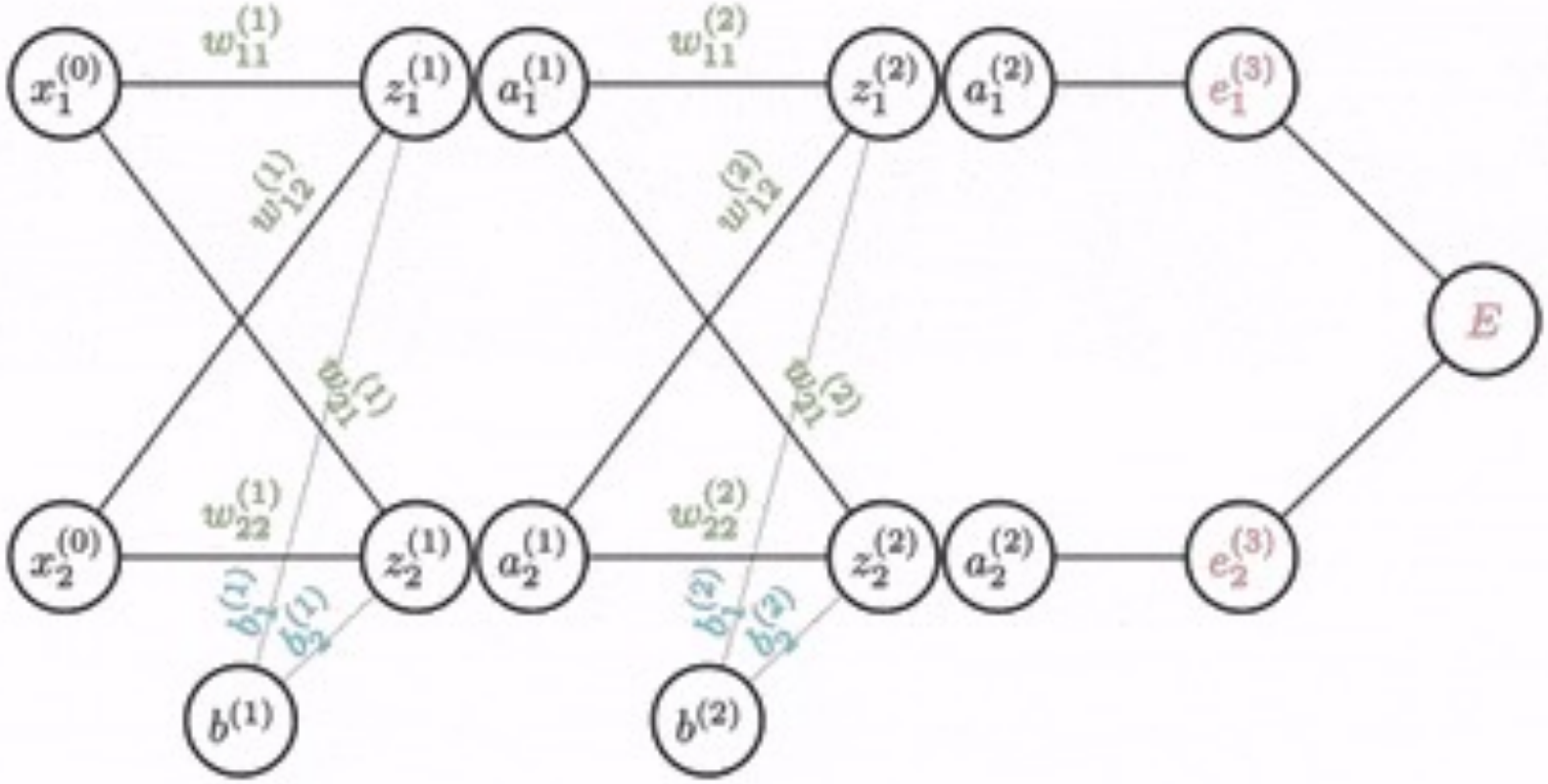


Turn or Peck

Multi-Layer Perceptron

$$\begin{bmatrix} \frac{\partial E}{\partial w_{11}^{(1)}} & \frac{\partial E}{\partial w_{12}^{(1)}} \\ \frac{\partial E}{\partial w_{21}^{(1)}} & \frac{\partial E}{\partial w_{22}^{(1)}} \end{bmatrix} = \begin{bmatrix} \delta_1^L \\ \delta_2^L \end{bmatrix}^T \cdot \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial a_1^{(1)}} & \frac{\partial z_1^{(2)}}{\partial a_2^{(1)}} \\ \frac{\partial z_2^{(2)}}{\partial a_1^{(1)}} & \frac{\partial z_2^{(2)}}{\partial a_2^{(1)}} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial a_1^{(1)}}{\partial z_1^{(1)}} \\ \frac{\partial a_2^{(1)}}{\partial z_2^{(1)}} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}} & \frac{\partial z_1^{(1)}}{\partial w_{12}^{(1)}} \\ \frac{\partial z_2^{(1)}}{\partial w_{21}^{(1)}} & \frac{\partial z_2^{(1)}}{\partial w_{22}^{(1)}} \end{bmatrix}$$

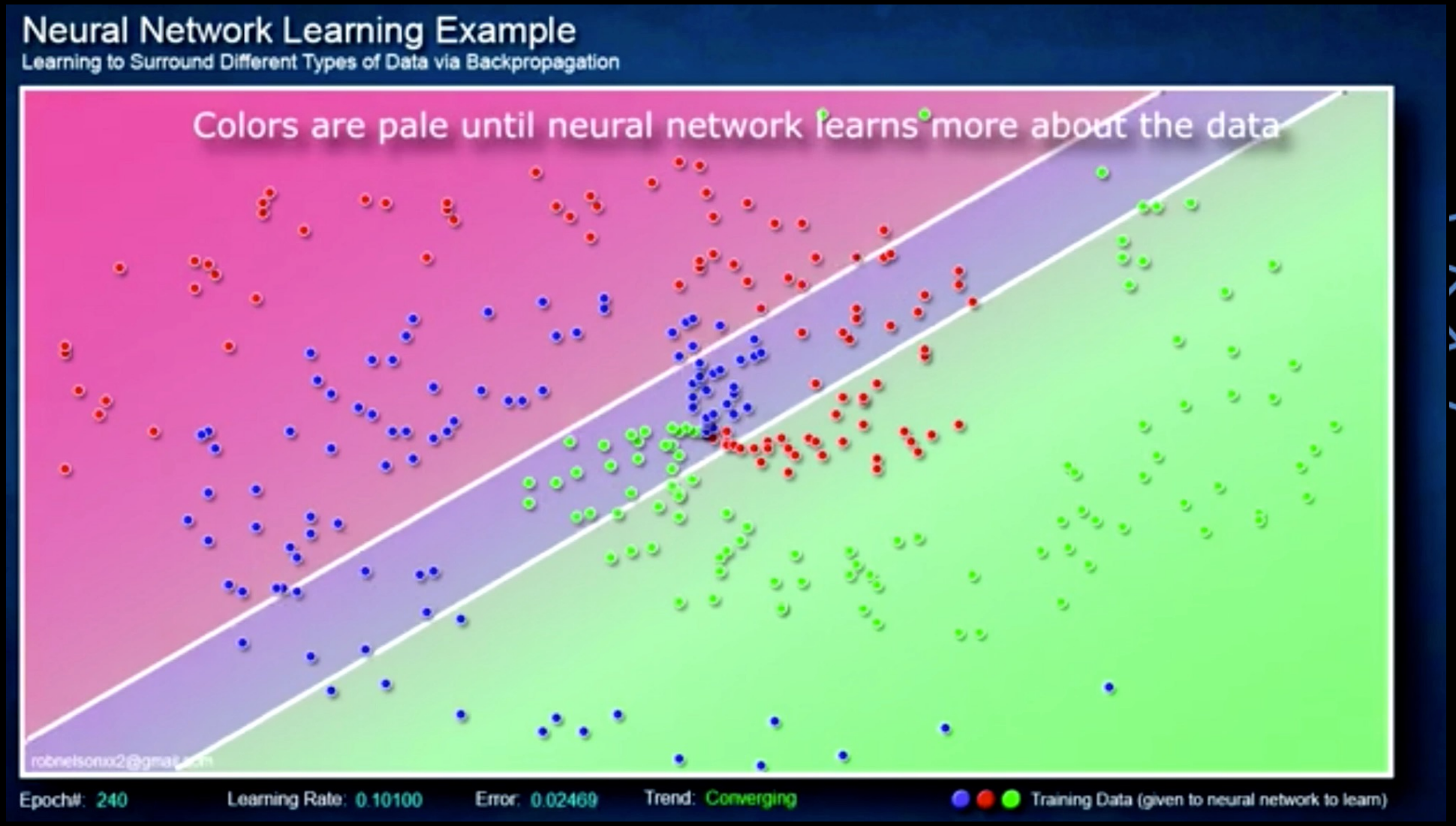
INPUT

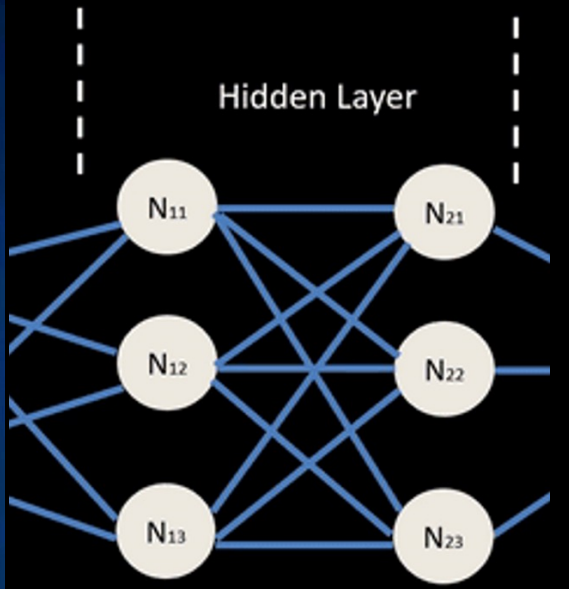


Turn or Peck

Multi-Layer Perceptron

3 Category (red green blue) Classifier of data with 2 features (x,y position)

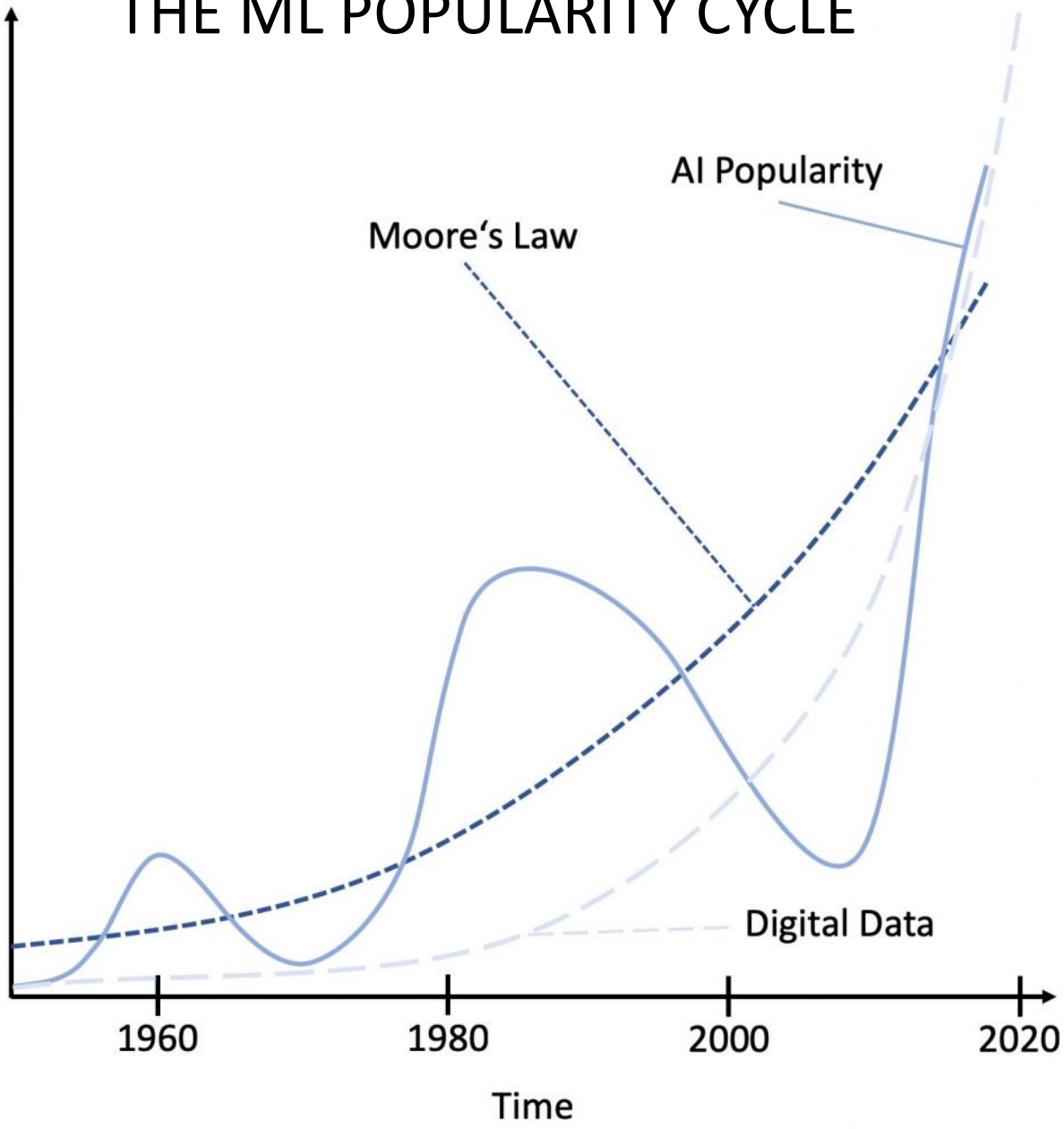




Why Does it Matter

- Machine learning allows us to “solve” (approximate) problems that we cannot solve analytically.

THE ML POPULARITY CYCLE



https://lme.tf.fau.de/pattern-recognition-blog/artificial-intelligence-a-bitter-sweet-symphony-in-modelling/

Current Upswing

TensorFlow

Google

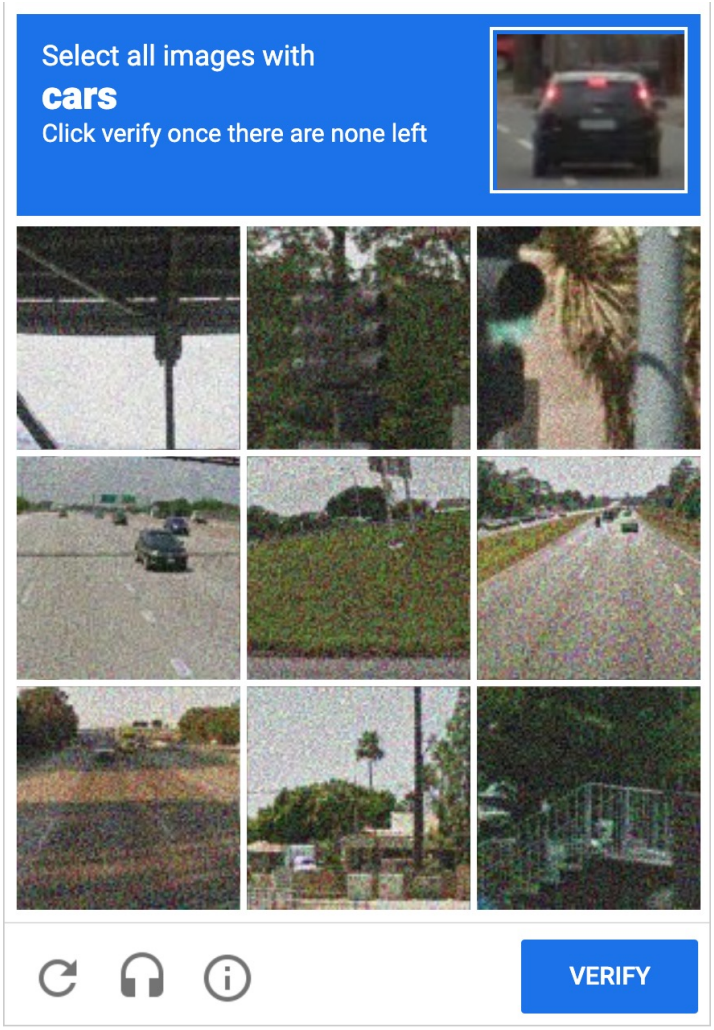
$1.809 Trillion

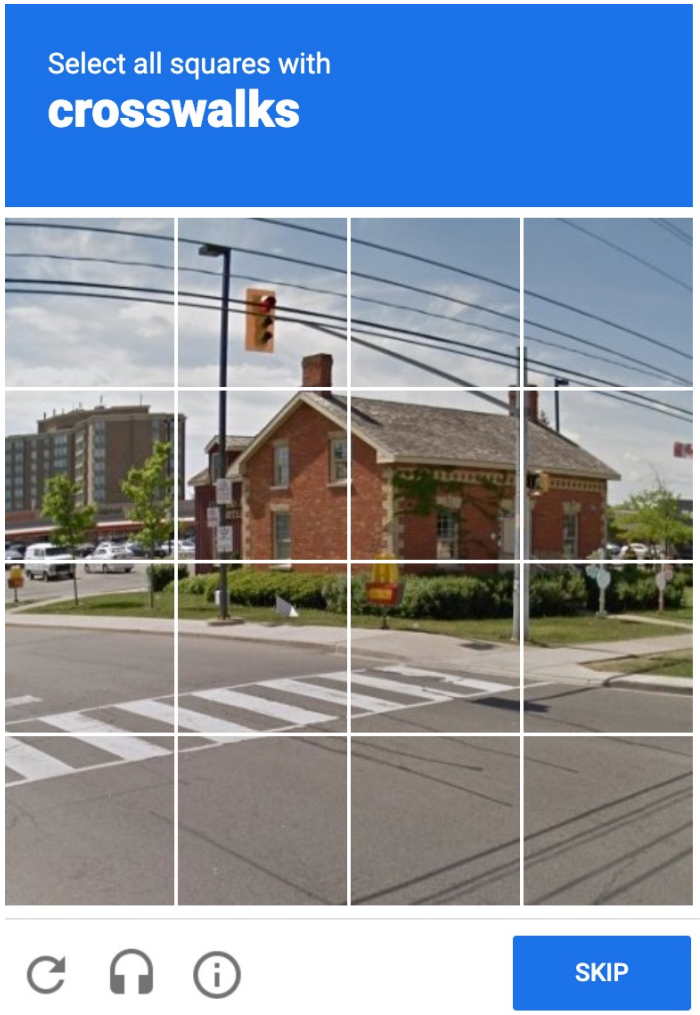
facebook

PYTORCH

$650 Billion

Current Upswing BIG DATA





TensorFlow

Google

$1.809 Trillion

facebook

PYTORCH

$650 Billion

reCAPTCHA is a **free** Google service

Convolutional Neural Networks (CNNs) for Image Classification.

Input image

Convolution Kernel

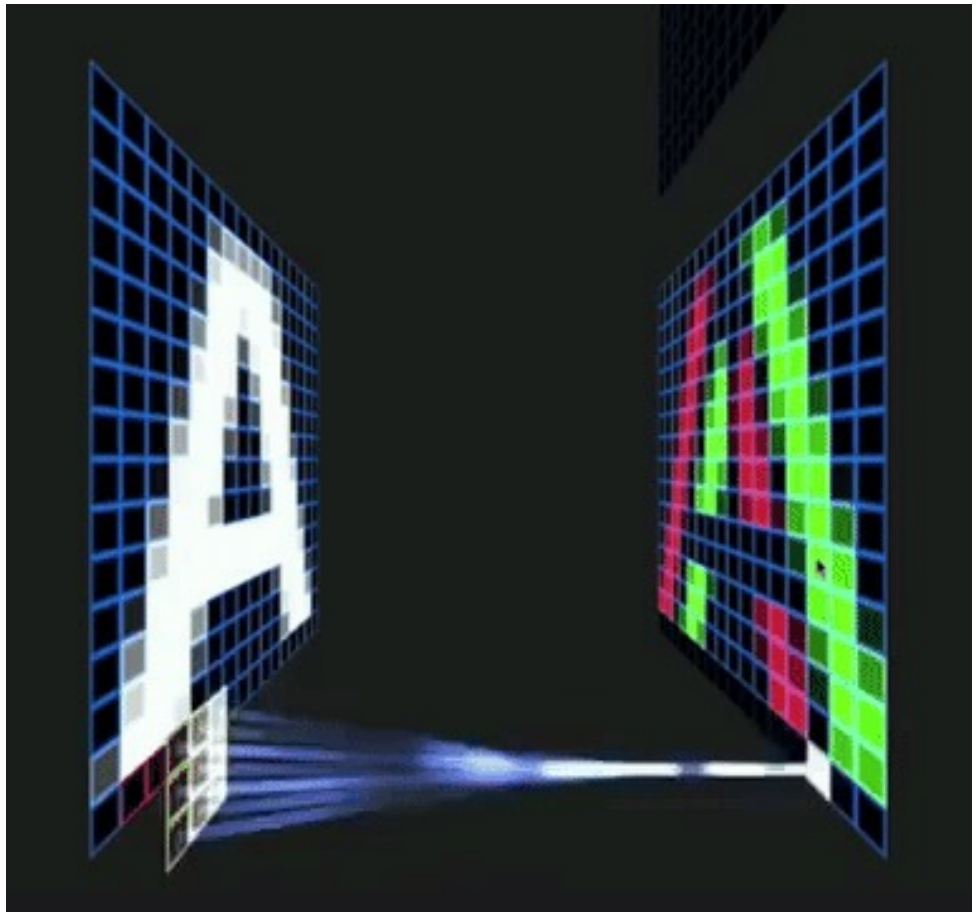
$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

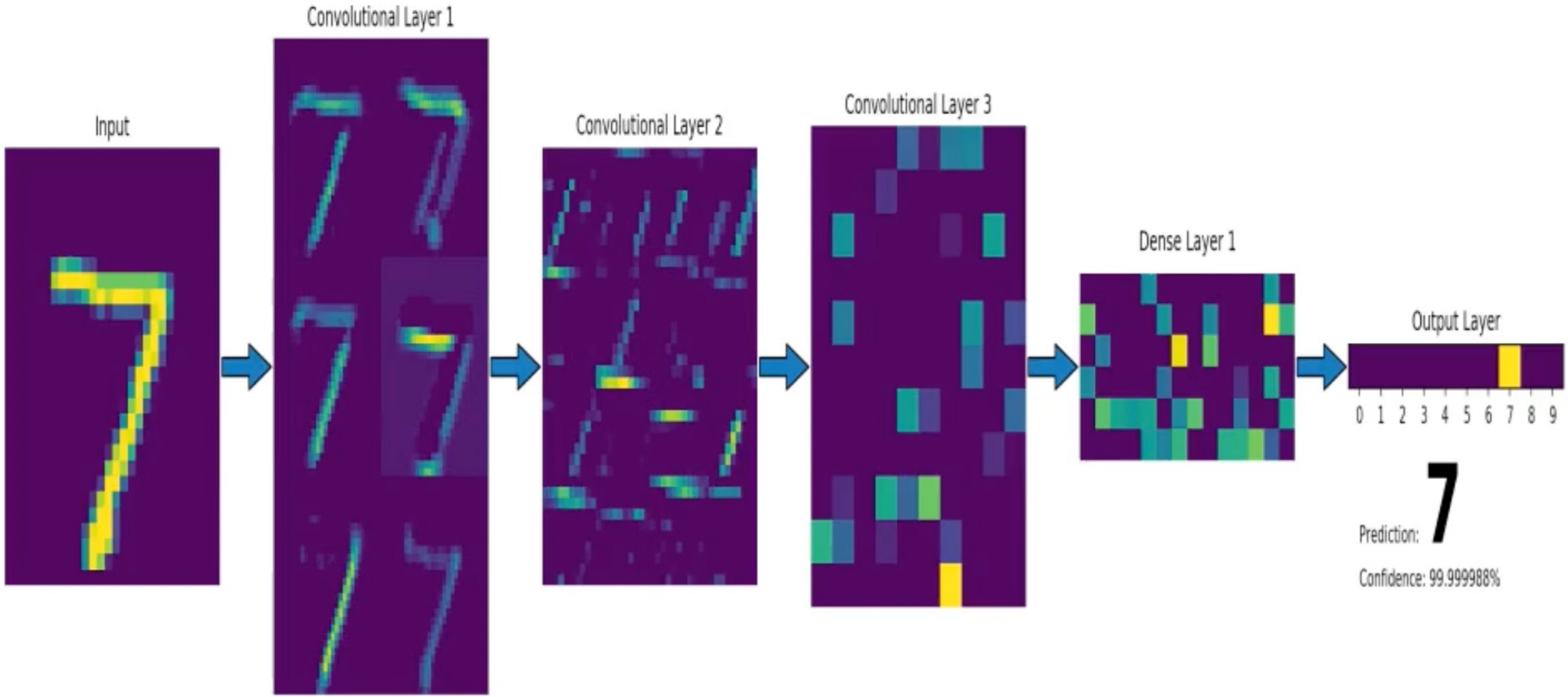
Feature map

Allows the network itself to

- Detect Edges
- Sharpen
- Blur
- Rotate

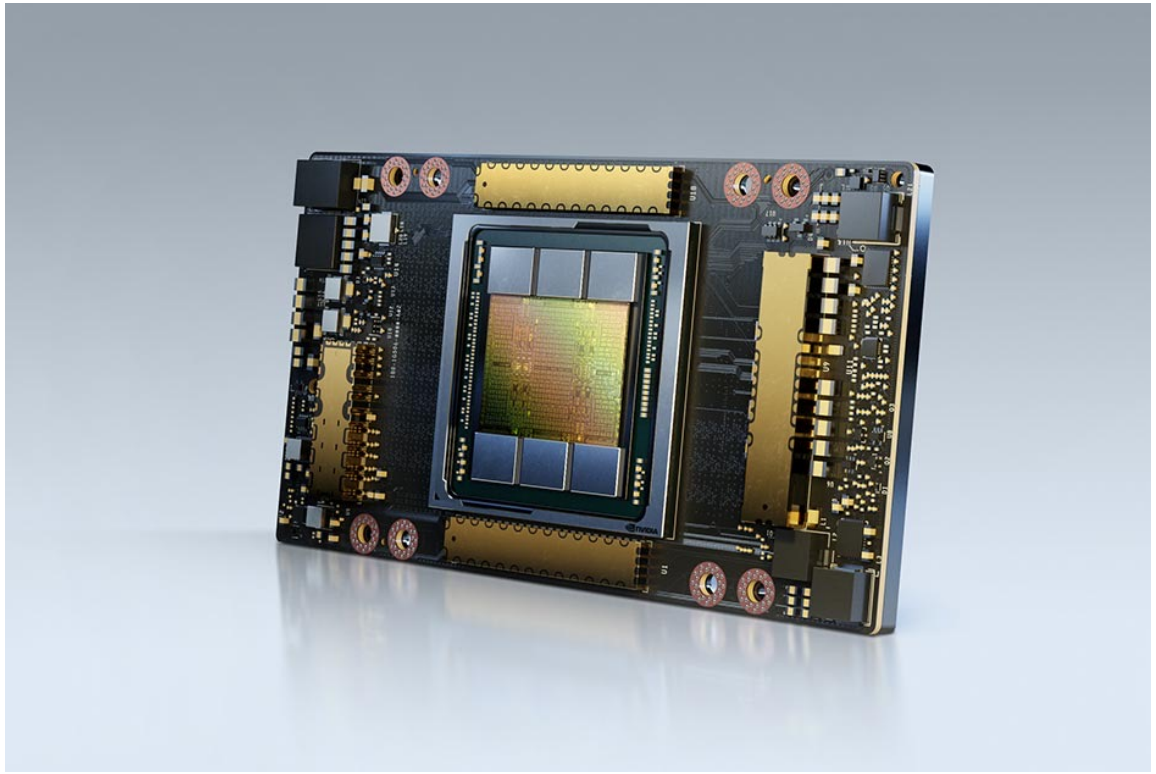
And generally, remove extraneous information and focus on the things that help it classify.




Input
Convolutional Layer 1
Convolutional Layer 2
Convolutional Layer 3
Dense Layer 1
Output Layer
0 1 2 3 4 5 6 7 8 9
Prediction: 7
Confidence: 99.999988%

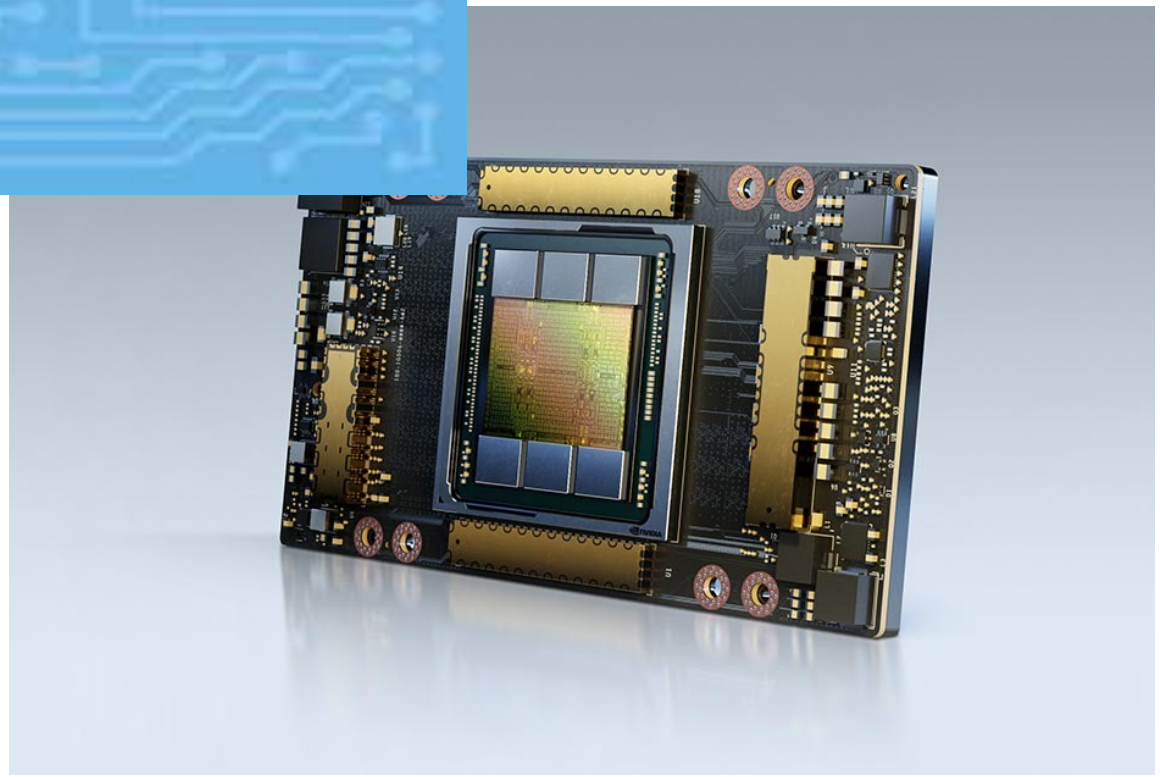
Hardware that’s Good at Matrix Multiplication

NVIDIA CUDA



ML has become so important that it is driving GPU and CPU architectures.

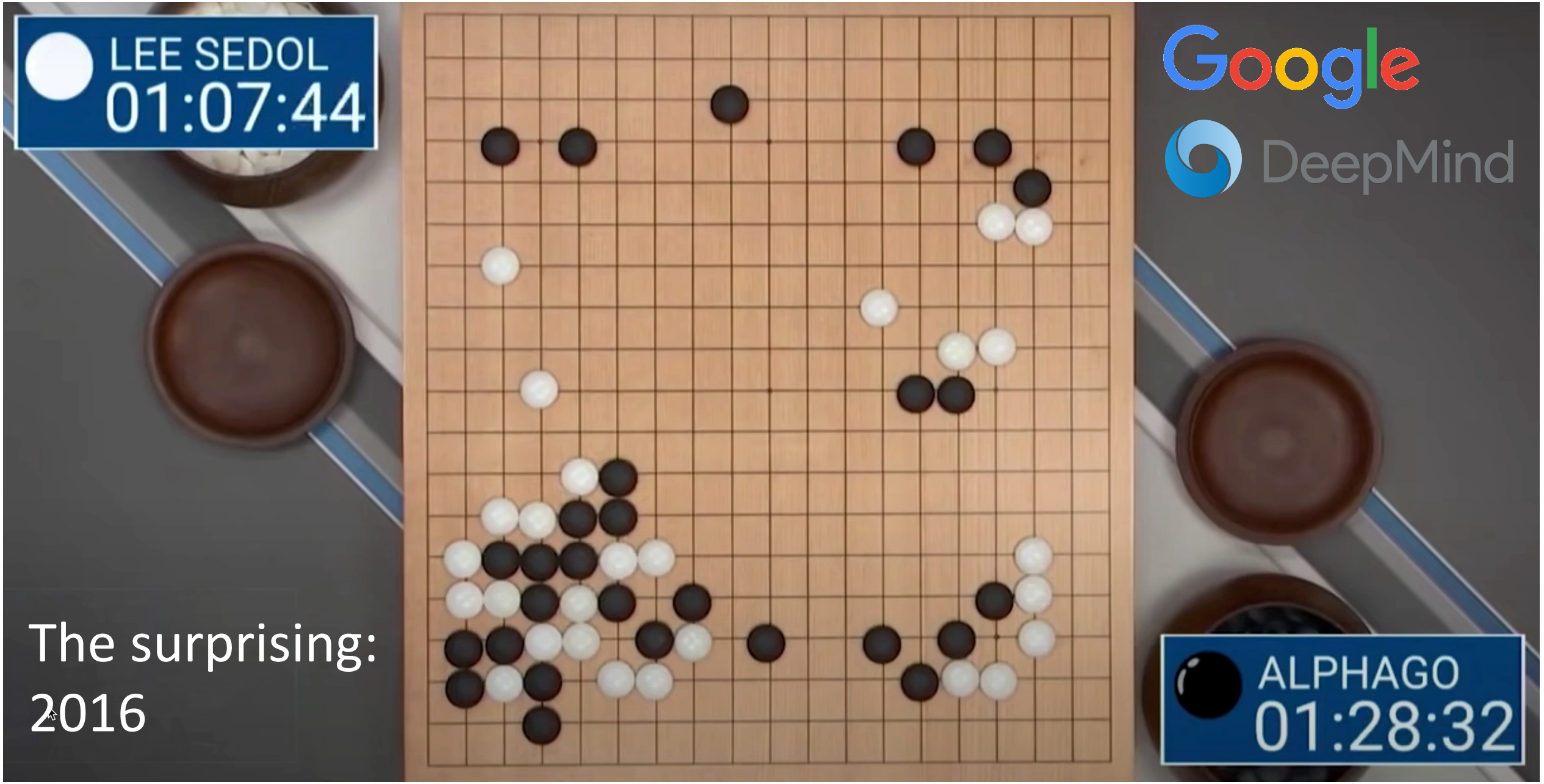
NVIDIA
CUDA



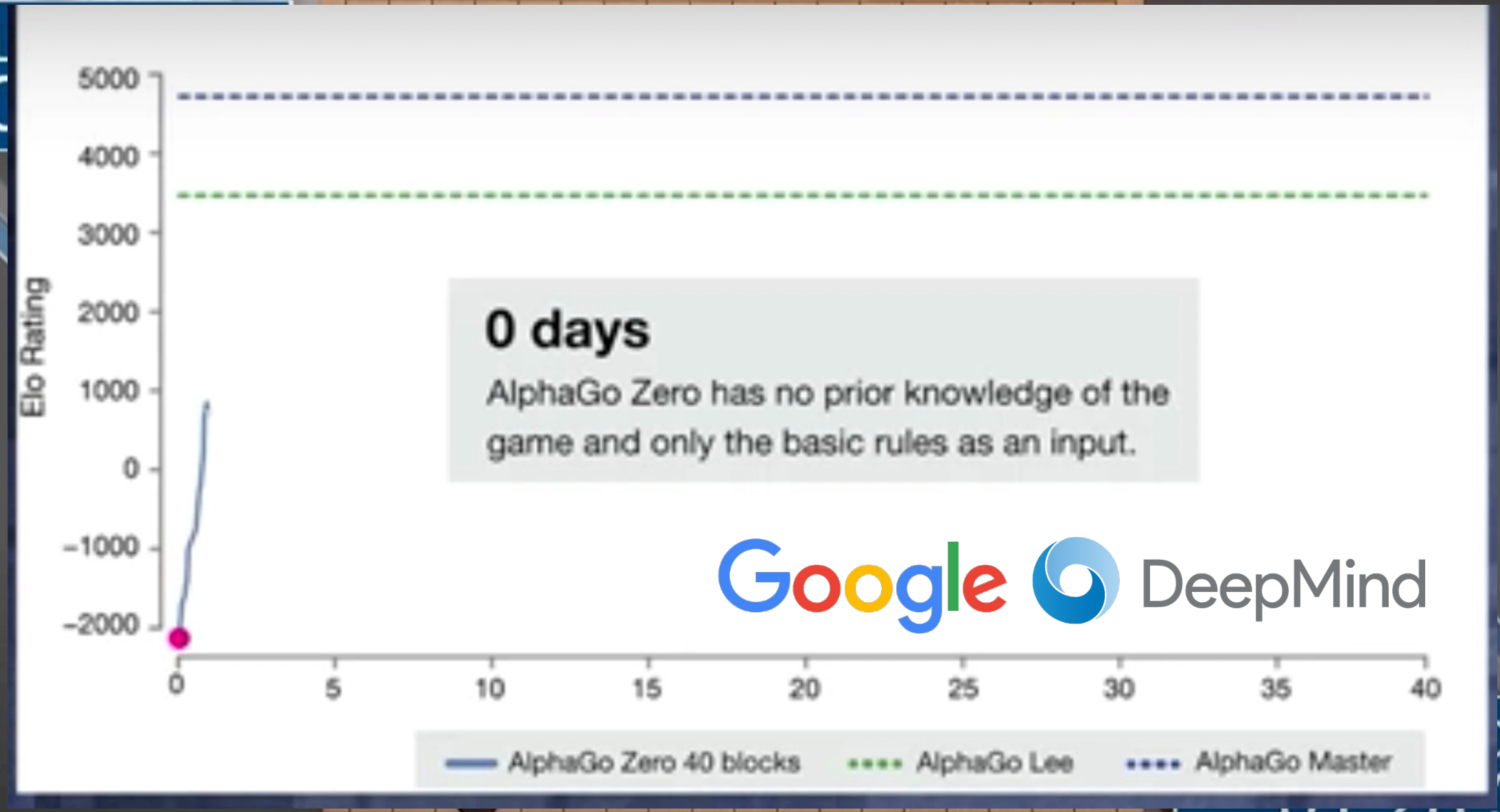
ML has become so important that it is driving GPU and CPU architectures.

The Good and the Bad and the Surprising

- Being able to train machine learning models from vast datasets has some great applications:
- But the model trained depends entirely on the data used to train it.

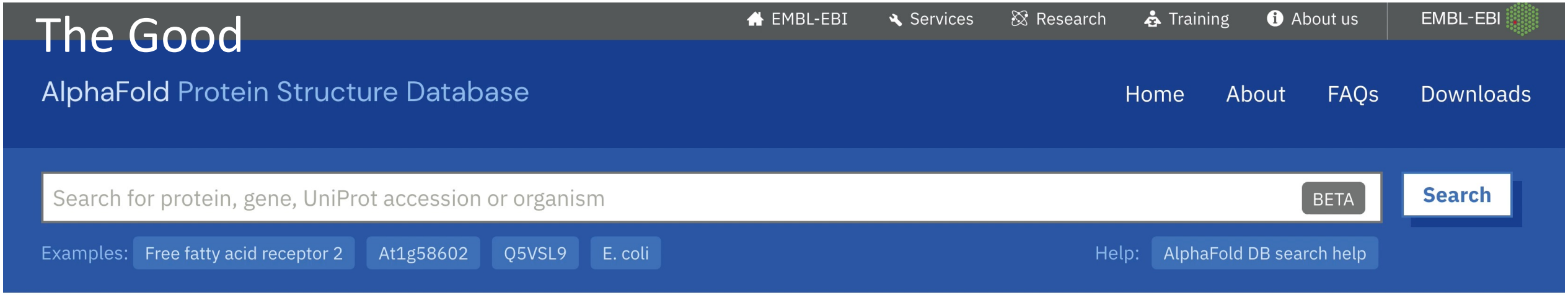


The surprising: 2016


5000
4000
3000
2000
1000
0
−1000
−2000
Elo Rating
0
5
10
15
20
25
30
35
40
0 days
AlphaGo Zero has no prior knowledge of the game and only the basic rules as an input.
Google DeepMind
AlphaGo Zero 40 blocks
AlphaGo Lee
AlphaGo Master

2017

The Good



T-cell immunomodulatory protein homolog

AlphaFold structure prediction

Download **PDB file** **mmCIF file** **Predicted aligned error**

NEW Feedback on structure **Looks great** **Could be improved**

Information

'The game has changed.' AI triumphs at solving protein structures
In milestone, software predictions finally match structures calculated from experimental data – Science Magazine

CARC Alphafold users: - Cristian Bologa - Melanie Moses

Google DeepMind

The Good

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

AlphaFo… Qs Downloads

Search fo… Search

Examples: …

T-ce… …ei…

AlphaFo…

Download…

NEW Fe…

Informat…

'The game… structures

In milesto… …

calculated from experimental data – Science Magazine

…epMind

The Bad

- Machine learning has been applied to
  - Mortgage Approval
  - Parole Hearings in New Mexico
  - Google hiring and promotion

- The input data is the history of parole, mortgage decisions, and hiring.
- See the problem?
-

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Xiaolin, Wu, and X. Zhang. "Automated Inference on Criminality using Face Images." *CoRR.–2016* (2016).

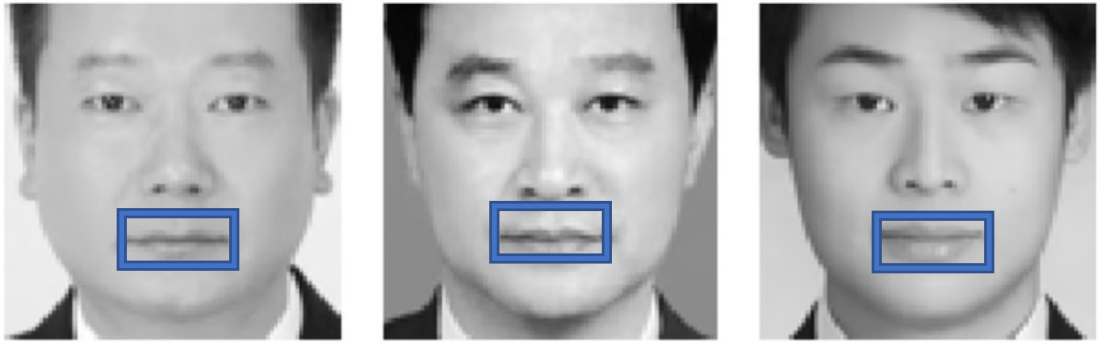
The Bad

- Machine learning has been applied to
  - Mortgage Approval
  - Parole Hearings in New Mexico
  - Google hiring and promotion

- The input data is the history of parole, mortgage decisions, and hiring.
- See the problem?

- These models all learned to associate gender and race with the outcome.

(a) Three samples in criminal ID photo set S_c.



(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Xiaolin, Wu, and X. Zhang. "Automated Inference on Criminality using Face Images." *CoRR.–2016* (2016).

Office of the General Counsel
Rules Docket Clerk
Department of Housing and Urban Development
451 7th Street, SW, Room 10276
Washington, DC 20410-0001.

SANTA FE INSTITUTE

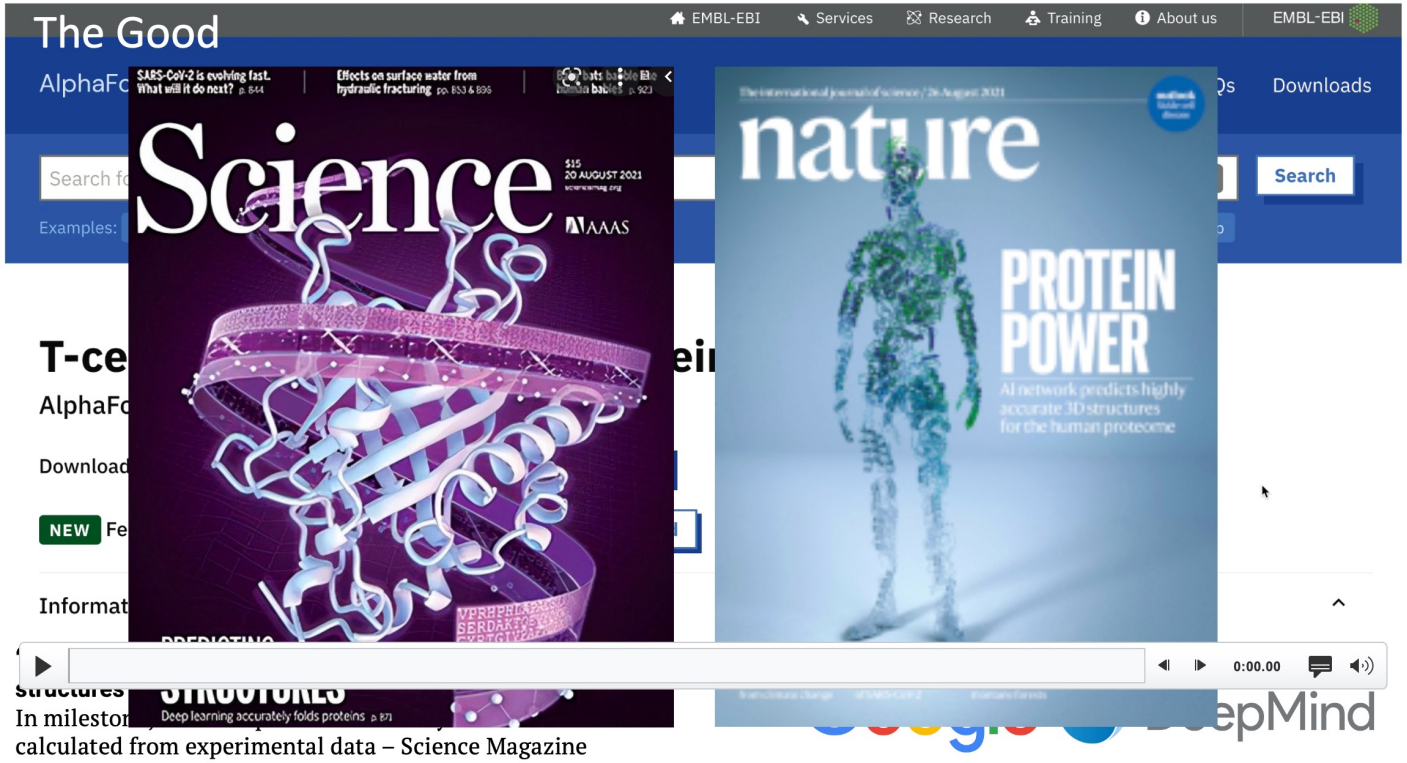
Cris Moore Melanie Moses

Regarding Docket No. FR-6111-P-02, HUD's Implementation of the Fair Housing Act's Disparate Impact Standard

We are a group of computer scientists, social scientists and legal scholars who are writing to express our concern about the proposed amendments to HUD's implementation of the Fair Housing Act, and in particular those amendments related to the use of algorithms.

We believe that the proposed amendments are based on a failure to recognize how modern algorithms can result in disparate impact, even in the absence of discriminatory intent, and how subtle the process of auditing algorithms

ML for CoVID Spike Modelling – GOOD!



In milestor
calculated from experimental data – Science Magazine

Inferring Criminality from Appearance – BAD!

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

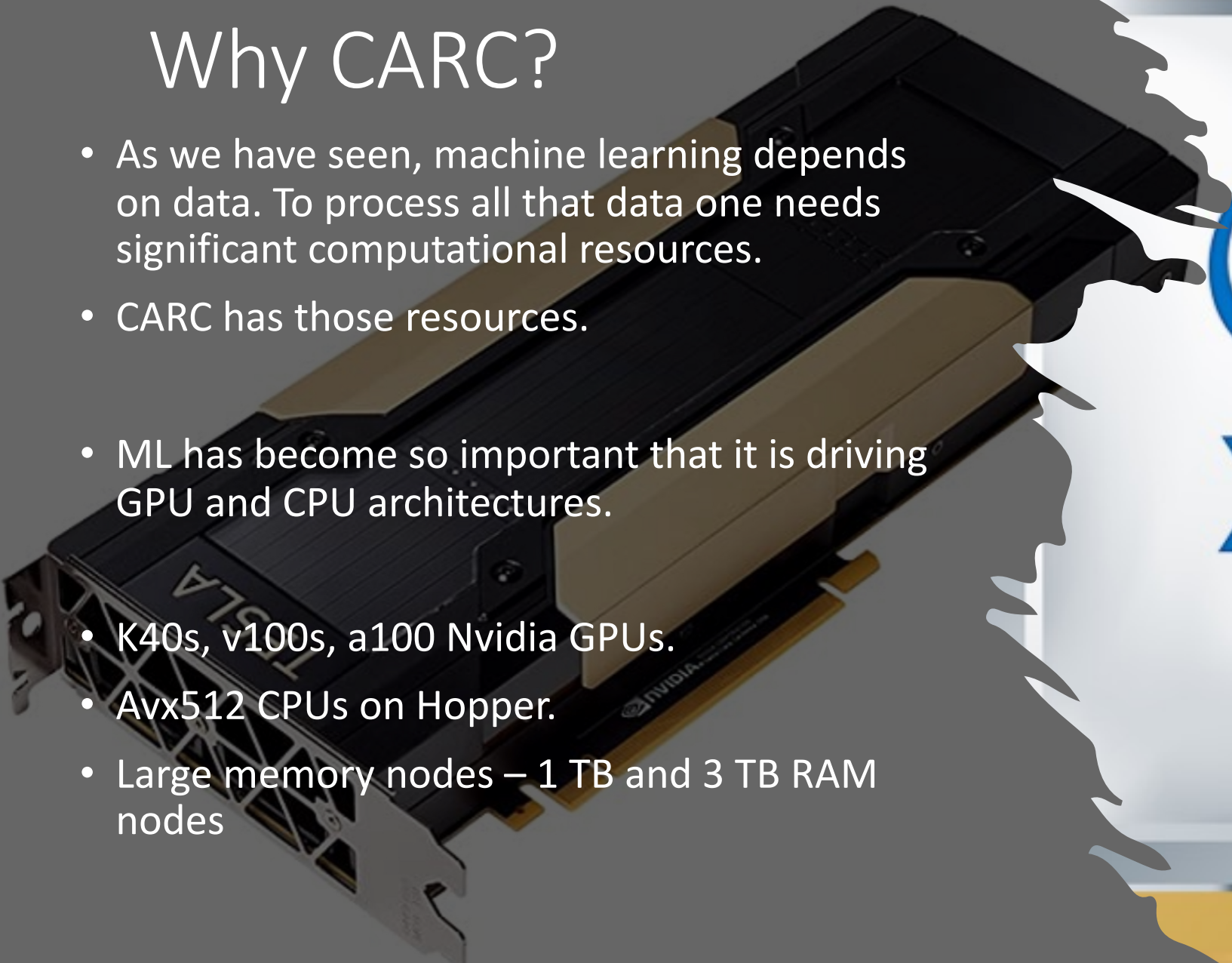
Figure 1. Sample ID photos in our data set.

Why CARC?

- As we have seen, machine learning depends on data. To process all that data one needs significant computational resources.
- CARC has those resources.

- ML has become so important that it is driving GPU and CPU architectures.

- K40s, v100s, a100 Nvidia GPUs.
- Avx512 CPUs on Hopper.
- Large memory nodes – 1 TB and 3 TB RAM nodes



intel®
XEON®
GOLD
inside™

Xena Cluster

*Susan Atlas

x32

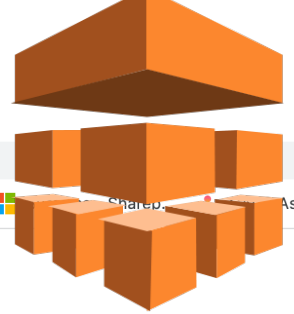
```
mfricke@xena:~ $ qgrok
queues       free   busy   offline  jobs   nodes  CPUs
-----        -----  -----  -----    -----  -----  -----
singleGPU    11     12     0        12     23     368
dualGPU      3      1      0        1      4      64
bigmem-1TB   1      1      0        2      2      128
bigmem-3TB   1      1      0        1      2      128
debug        1      0      0        0      1      16
systems      1      0      0        0      1      16
totals:      18     15     0        16     33     720
```

Machine Learning QuickByte Tutorials

Http://carc.unm.edu

- R
  - R Programming in HPC
  - R at CARC
  - Running R in Parallel with Future
  - Gurobi optimizer with R
- Machine Learning
  - Tensorflow
  - Machine Learning Conda Environments (including PyTorch K40 GPU libraries
  - Tensorflow with multiple GPUs
  - Parallelization with JupyterHub using Dask and SciKit-learn
- Bioinformatics
  - Genomic variant calling with GATK
  - Genome evaluation with QUAST and BUSCO
  - Single genome demographic history with PSMC

https://xena.alliance.unm.edu:8000



Amazon Machine Learning

Google colab

taos.alliance.unm.edu:8000/user/mfricke/tree/workshops/ML

CARC Helpdesk | ParentVUE | CARC AIRS | fricke.co.uk | Fricke Email | Spam Manager | CARC Systems | CARC Business | My UNM | Update | oD Portal

jupyterhub

Files | Running | IPython Clusters

Select items to perform actions on them.

Upload | New

0 / workshops / ML

..

checkpoints_best_cnn

checkpoints_best_mlp

data

GPUvsCPUMatMul.ipynb kB

GPUvsCPUMatMulSaved.ipynb kB

SVHN_classifier.ipynb kB

SVHN_classifierSaved.ipynb kB

TensorBoard.ipynb kB

Notebook:

Julia 1.5.2

Python 3

Python [conda env:.conda-MSML]

Python [conda env:.conda-MSML_TF_GPU]

Python [conda env:.conda-MSML_tensorflow_decision_forest]

Python [conda env:.conda-R]

Python [conda env:.conda-SVM]

Python [conda env:.conda-alphafold-env]

Python [conda env:.conda-chem501]

Python [conda env:.conda-cjs14]

Python [conda env:.conda-comp_immunology]

Python [conda env:.conda-curvefit]

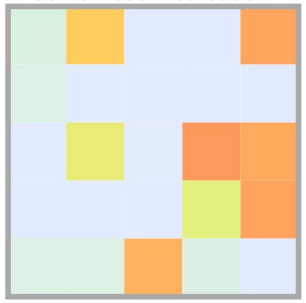
Xena Cluster Grid > Xena Cluster > --Choose a Node

Overview of Xena Cluster @ 2022-04-08 01:49

CPUs Total:	**384**
Hosts up:	**24**
Hosts down:	**0**

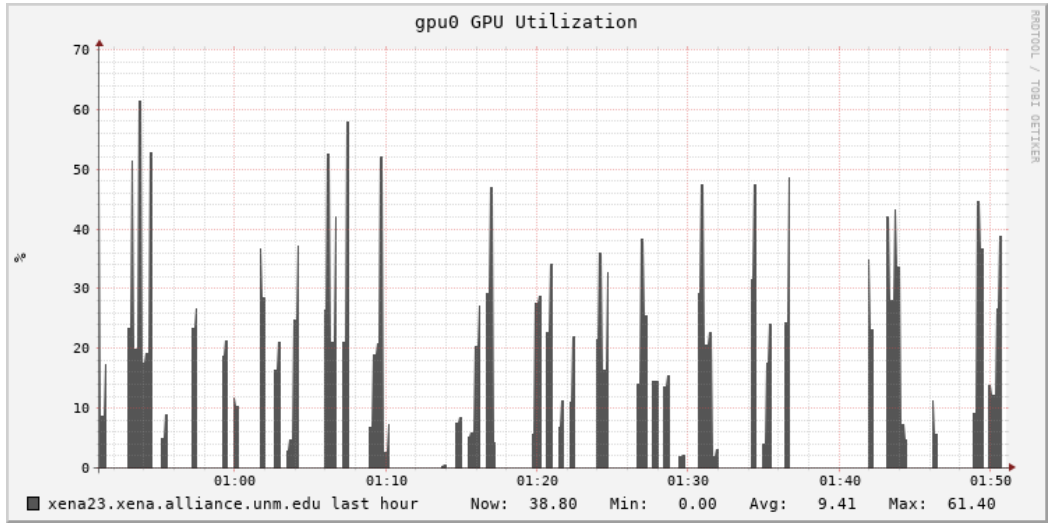
Current Load Avg (15, 5, 1m):
22%, 22%, 22%
Avg Utilization (last hour):
23%

Server Load Distribution

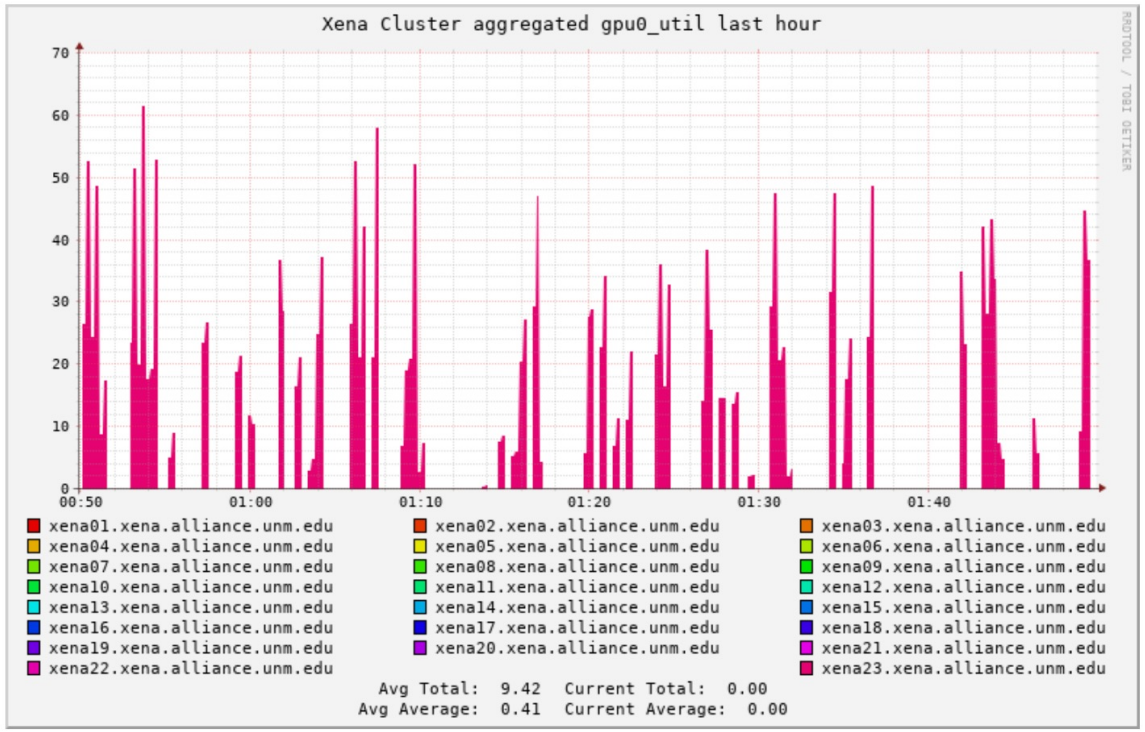


http://xena.alliance.unm.edu/ganglia

Stacked Graph - gpu0_util (%)



Xena23 GPU0 Utilization



Xena Cluster **gpu0_util (%)** last **hour** sorted **by name**

Happy Hour!

Please come to my open office hours!

Send email to help@carc.unm.edu