SCIENCE IN THE AGE OF AI

Harrison B. Prosper
Florida State University

Colloquium, York University, 16 March 2021

Outline

- A Brief History of AI
- Machine Learning
- ML In Science
- Game Changer
- The Future

What Is Artificial Intelligence ?

"…artificial intelligence (AI) refers to any human-like intelligence exhibited by a computer, robot, or other machine."

IBM

https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

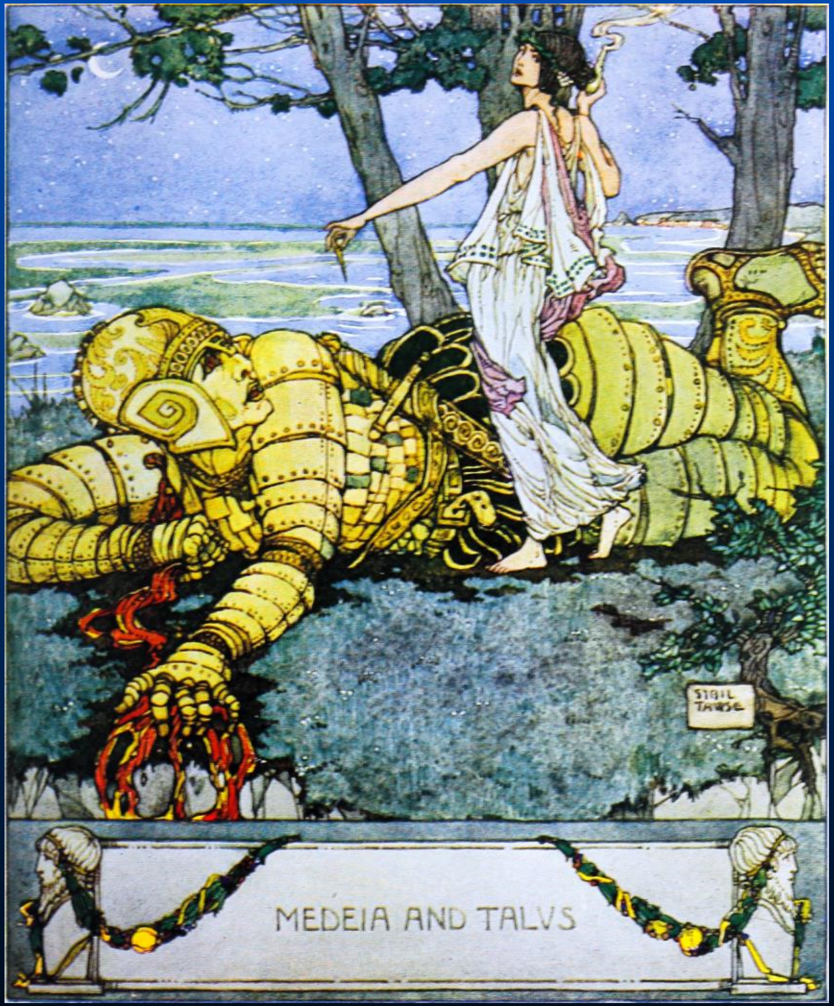
A BRIEF HISTORY OF AI

Talus, Ancient Greece

Stories about artificially intelligent beings abound in ancient civilizations.

India: *spirit movement machines.*

Aka: Robots!



Wikimedia Commons

Moveable Type (Gutenberg Bible, 1456)

By NYC Wanderer (Kevin Eng) - originally posted to Flickr as Gutenberg Bible

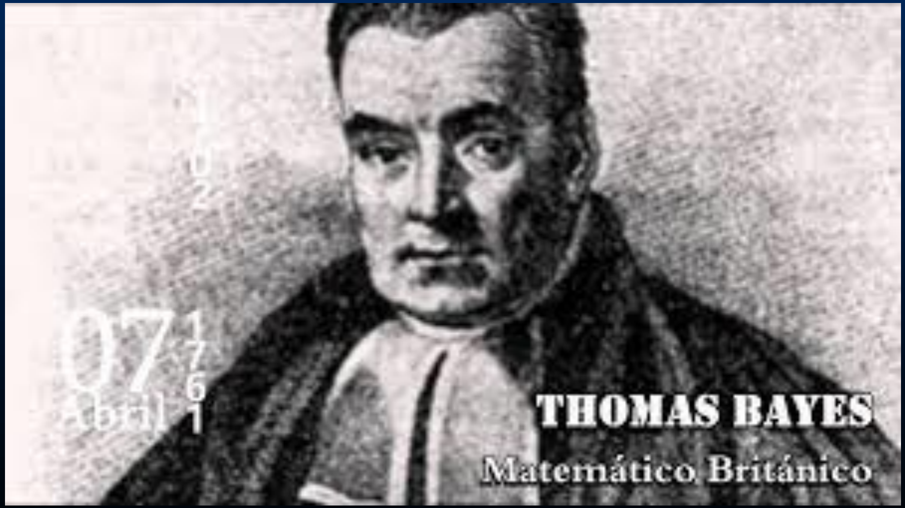
17th century

- Many philosophical ideas about knowledge and reason.

18th century

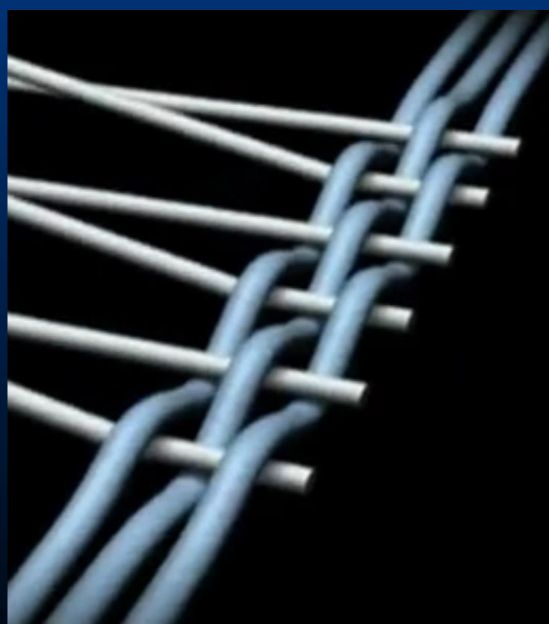
- 1763 – Thomas Bayes publishes important theorem.

$$P(\boldsymbol{H}|D) = \frac{P(D|\boldsymbol{H})P(\boldsymbol{H})}{P(D)}$$

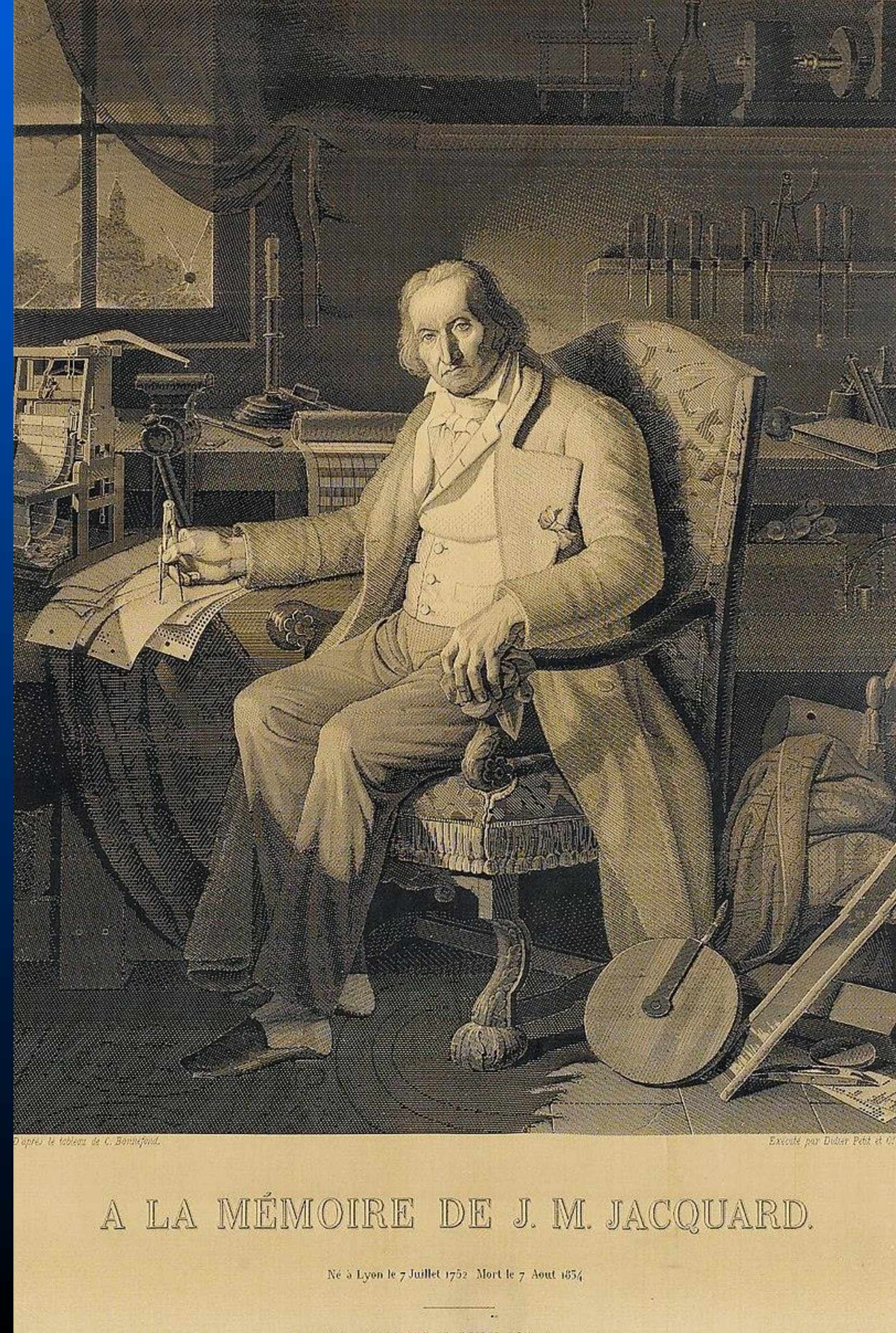


19th century

- 1801 – Joseph-Marie Jacquard invents first programmable machine.

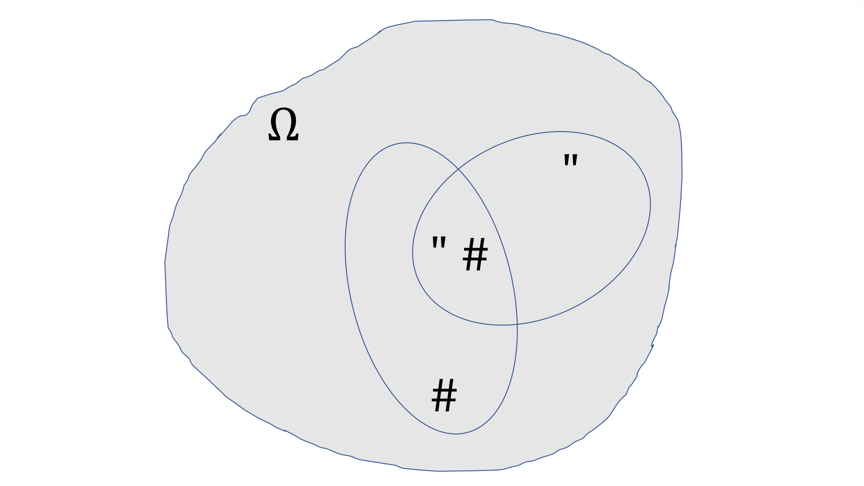


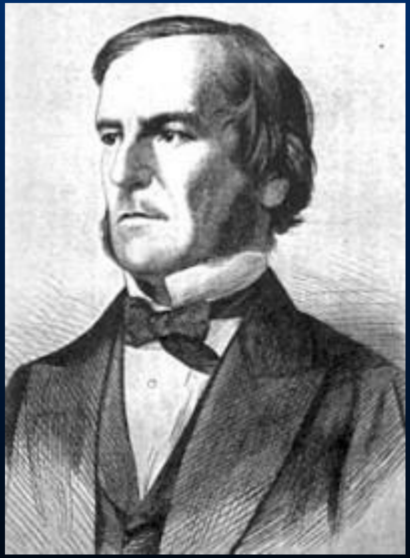
Wikimedia commons



19th century

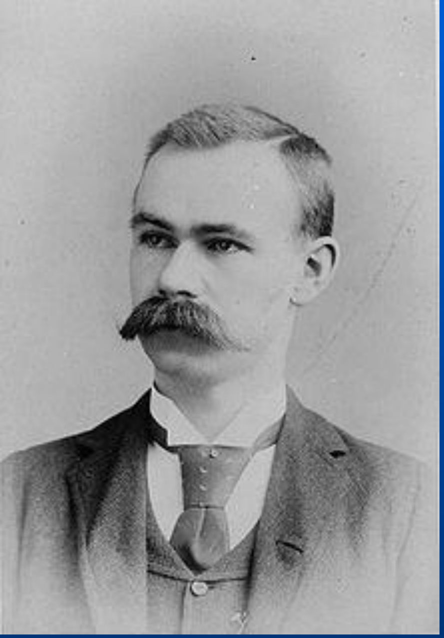
- 1832 – Charles Babbage designs first programmable calculator.
- 1854 – George Boole invents algebra of logic.





1815 -1864

1890 US Census



Herman Hollerith
(1860 – 1929)

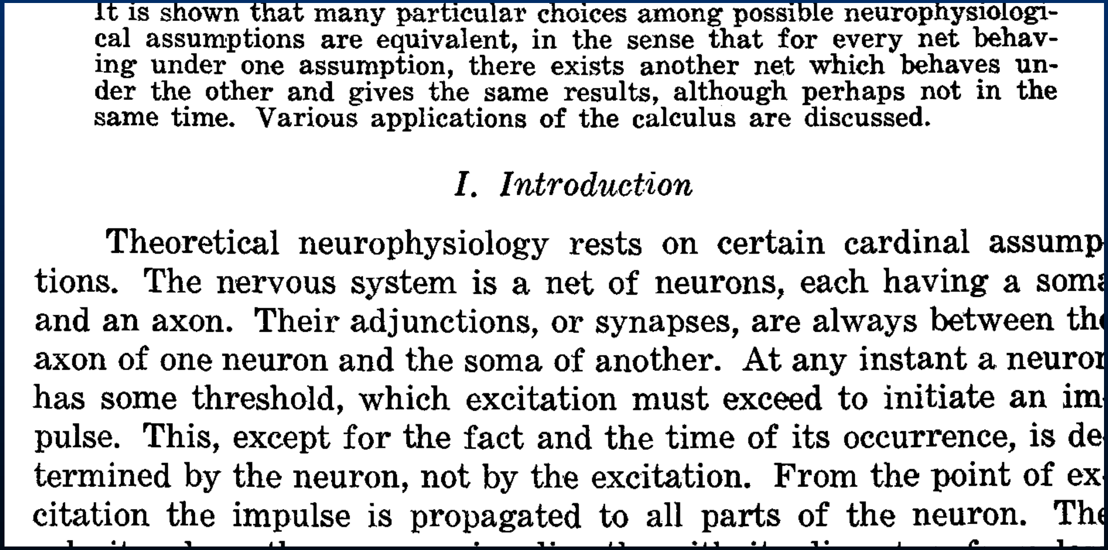
1911: CTR Corp.
1924: IBM

Photo: IBM

Wikimedia commons

20th century (1900 – 1950)

- 1936 – Alan Turing proposes a universal computing machine.
- 1943 – Warren McCulloch and Walter Pitts invent *neural networks* (NN).



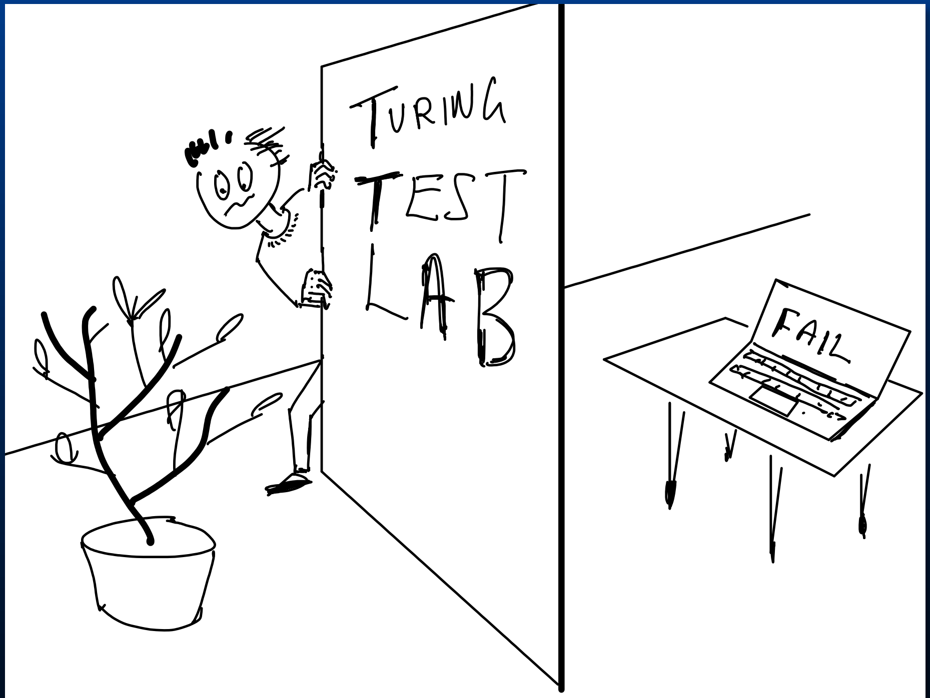
It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

I. Introduction

Theoretical neurophysiology rests on certain cardinal assump tions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between th axon of one neuron and the soma of another. At any instant a neuro has some threshold, which excitation must exceed to initiate an im pulse. This, except for the fact and the time of its occurrence, is de termined by the neuron, not by the excitation. From the point of ex citation the impulse is propagated to all parts of the neuron. The

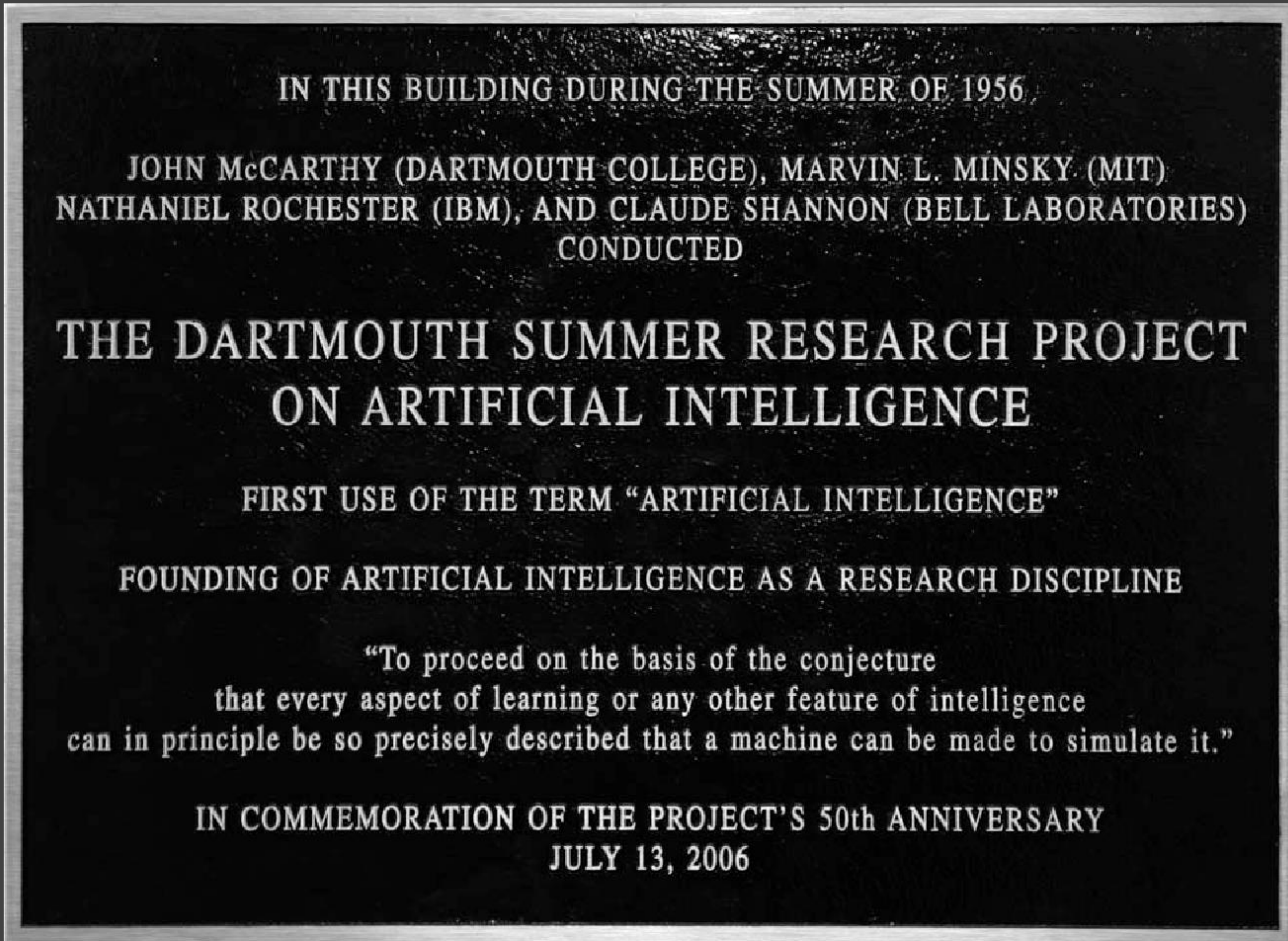
20th century (1900 – 1950)

- 1950 – The Turing Test



"George rethinks his life after failing the Turing Test"

HBP



Courtesy John Moor, DOI:10.1609/aimag.v27i4.1911

1997 World chess champion Gary Kasparov defeated by IBM Deep Blue

Feng-hsiung Hsu
Murray Campbell
IBM Research

Stan Honda/AFP/Getty Images

Source: IBM

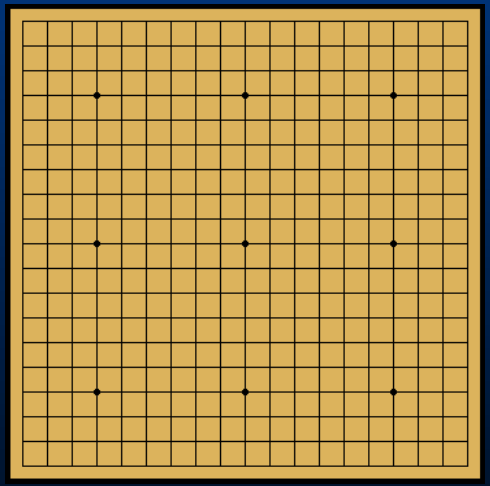
Computer Wins on ‘Jeopardy!’: Trivial, It’s Not
New York Times, Feb. 17, 2011

Carol Kaelson/Jeopardy Productions Inc., via Associated Press

Ken Jennings: “I felt obsolete”
TED Talk

Machine 4, Human 1

2016 – Google's DeepMind AlphaGo program beats Go champion Lee Sodol.



Photograph: Yonhap/Reuters

A Brief History of AI

"York University professors protest their replacement by iPhone 9000s"

Toronto Star, Toronto, Canada, 16 March 2071

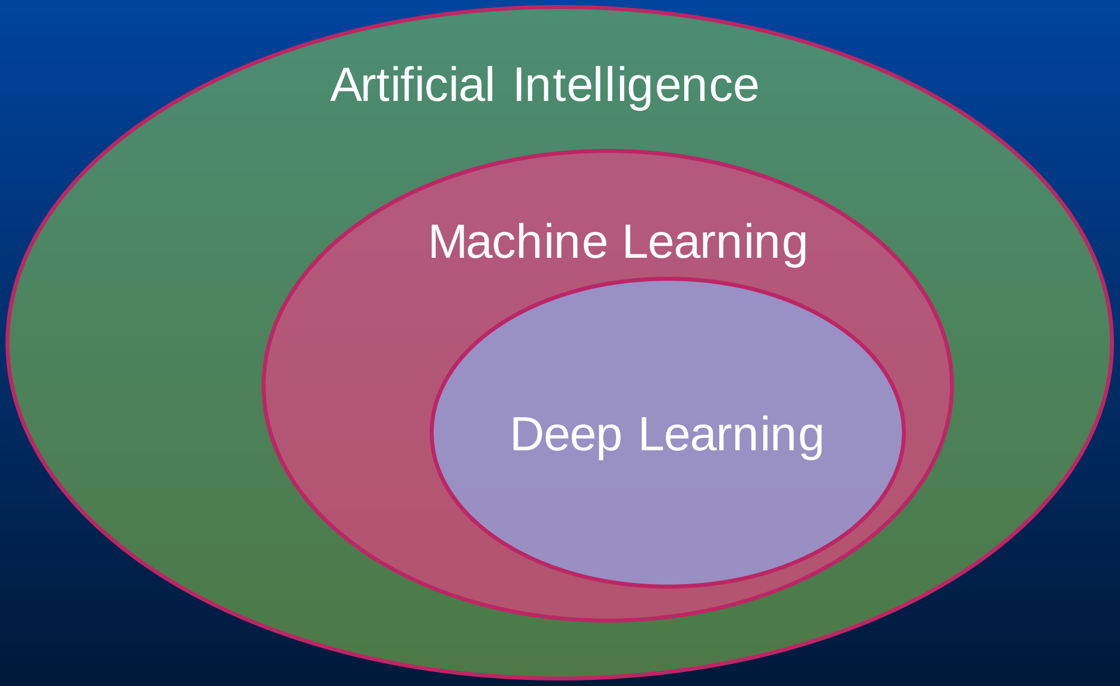
MACHINE LEARNING

“That is positively the dopiest idea I have heard.”

Richard Feynman

Thinking Machines Corporation, summer 1983.

Machine Learning



Machine Learning

The use of computers to fit highly non-linear, recursively constructed, parameterized functions $f(x, \theta)$ to data.

1. Given an objective function (average of loss function, L)

$$F(\theta) = \frac{1}{T}\sum_{i=1}^{T} L(t_i, f_i)$$

2. Solve

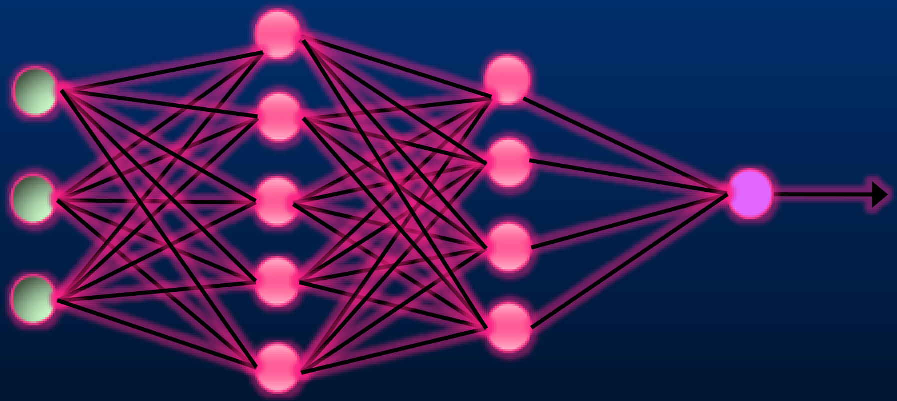
$$\frac{dF}{d\theta} = 0$$

   for θ.

3. Commonly used loss function: $L(t, f) = (t - f)^2$

Deep Learning

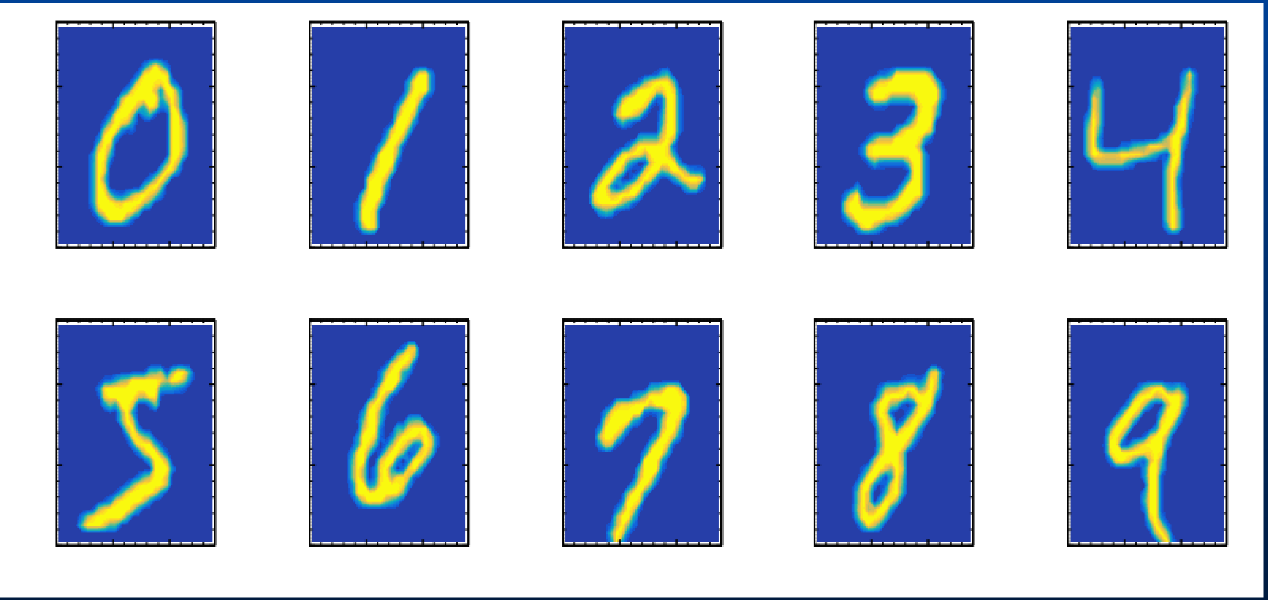
In 2006, University of Toronto researchers Hinton, Osindero, and Teh* developed a sophisticated, workable, method to train deep neural networks.



Geoffrey Hinton

* Hinton, G. E., Osindero, S. and Teh, Y., A fast learning algorithm for deep belief nets, Neural Computation 18, 1527-1554 (2006).

But sophistication is not mandatory! (MNIST)*

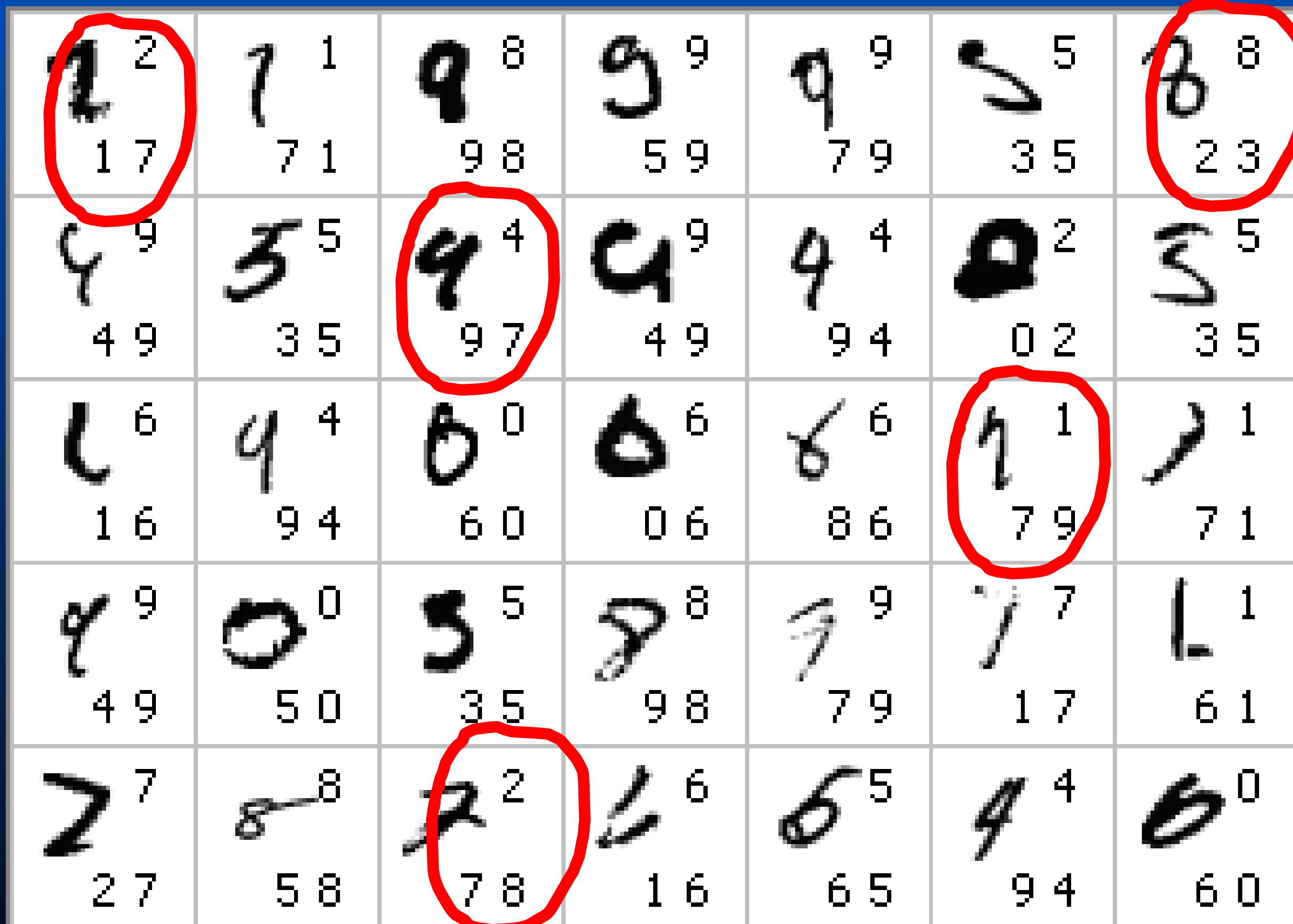


*Cireşan DC, Meier U, Gambardella LM, Schmidhuber J. ,

Deep, big, simple neural nets for handwritten digit recognition.

Neural Comput. 2010 Dec. 22 (12): 3207-20.

(784, 2500, 2000, 1500, 1000, 500, 10)



Upper right: correct answer; lower left answer of highest DNN output; lower right answer of next highest DNN output.

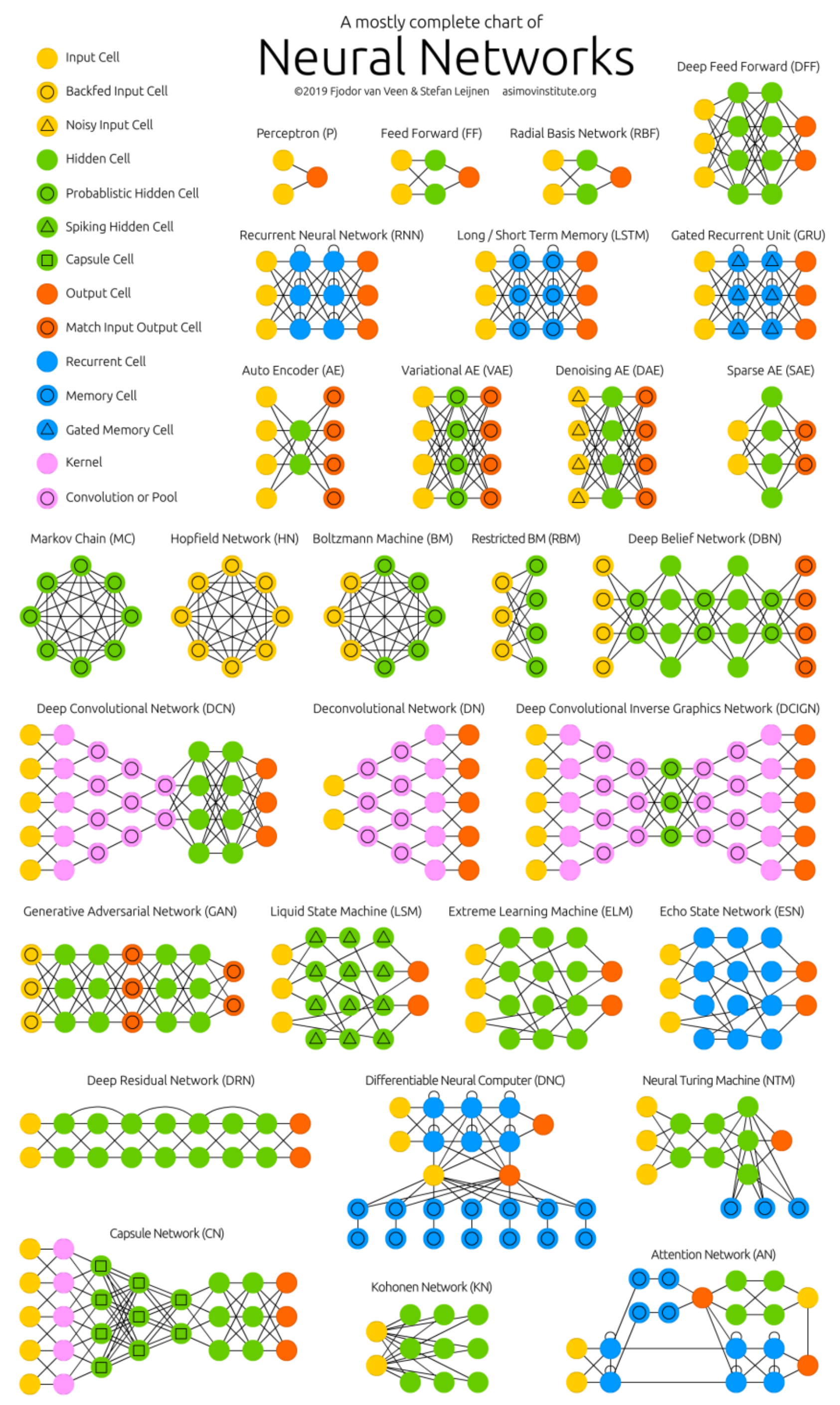
The Deep Learning Revolution

Key advances:

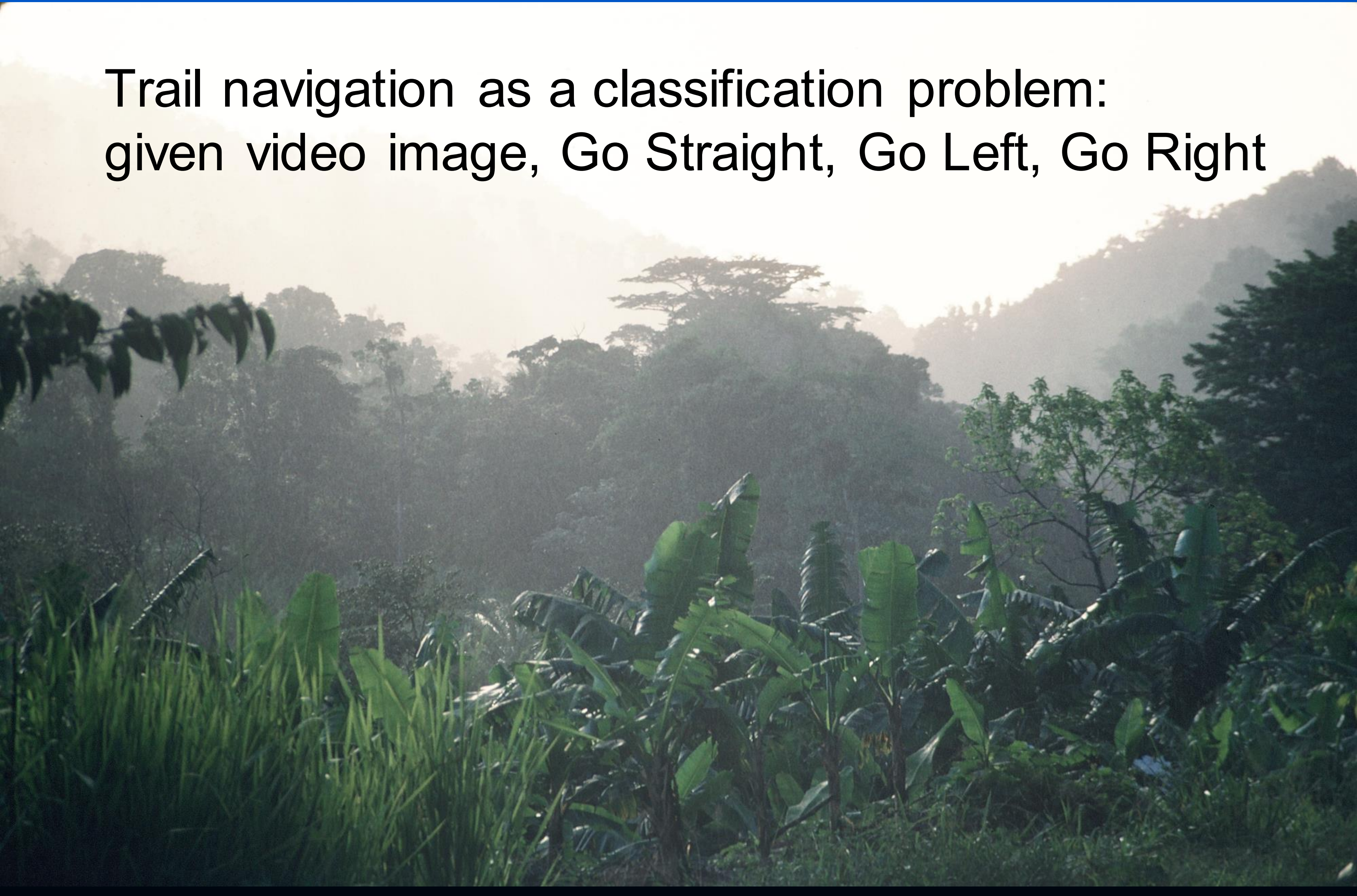
1. Huge highly effective models
2. Huge data sets
3. Parallel computation
4. Effective optimizers
5. Automatic differentiation

Van Veen, F. & Leijnen, S. (2019). The Neural Network Zoo.

https://www.asimovinstitute.org/neural-network-zoo



Trail navigation as a classification problem:
given video image, Go Straight, Go Left, Go Right



IEEE ROBOTICS AND AUTOMATION LETTERS. PREPRINT VERSION. ACCEPTED NOVEMBER, 2015 1

A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots

Alessandro Giusti[1], Jérôme Guzzi[1], Dan C. Cireşan[1], Fang-Lin He[1], Juan P. Rodríguez[1]
Flavio Fontana[2], Matthias Faessler[2], Christian Forster[2]
Jürgen Schmidhuber[1], Gianni Di Caro[1], Davide Scaramuzza[2], Luca M. Gambardella[1]

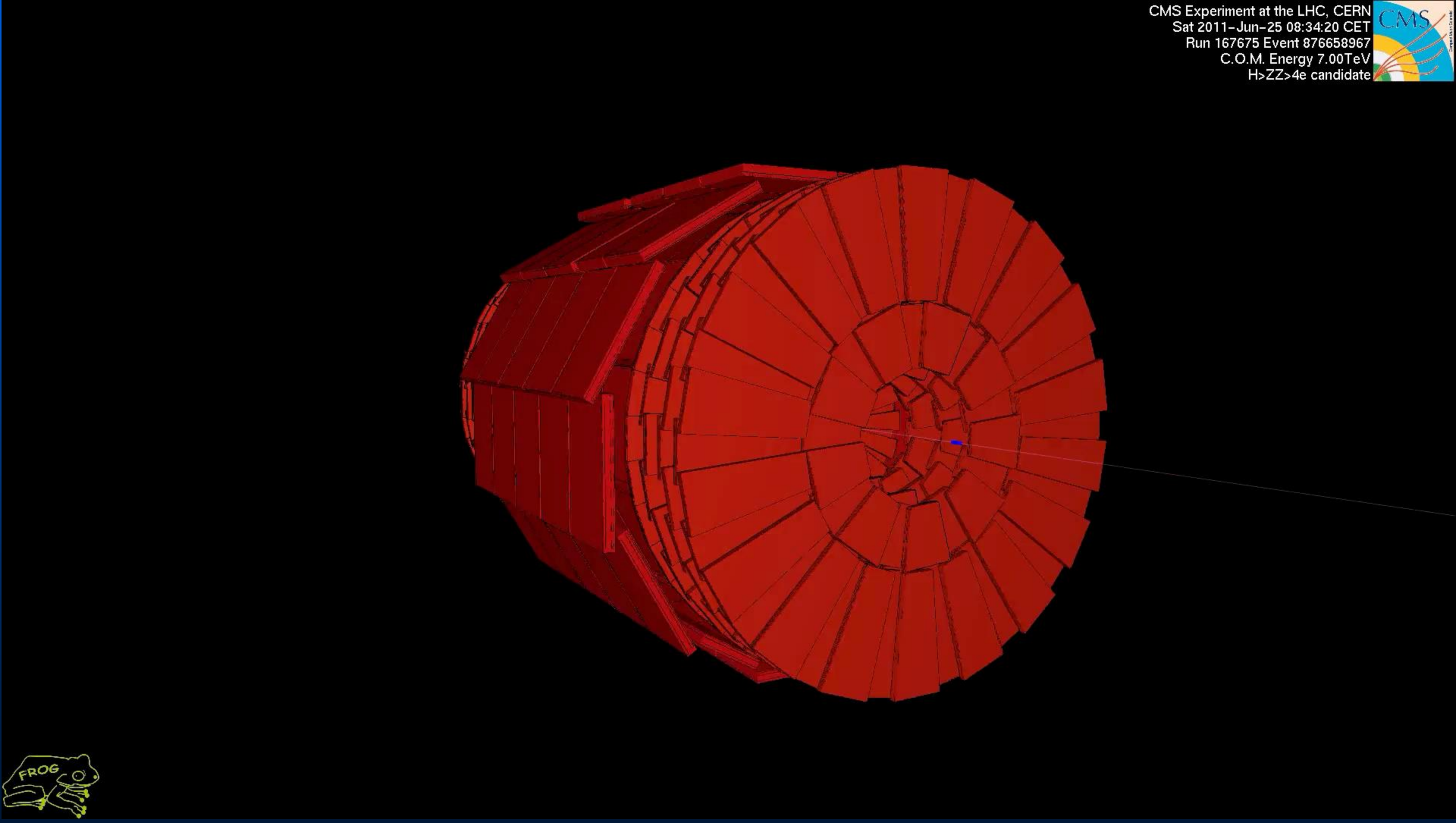
ML IN SCIENCE

AI in Science: Examples

- Particle Physics
- Astrophysics
- Mathematics

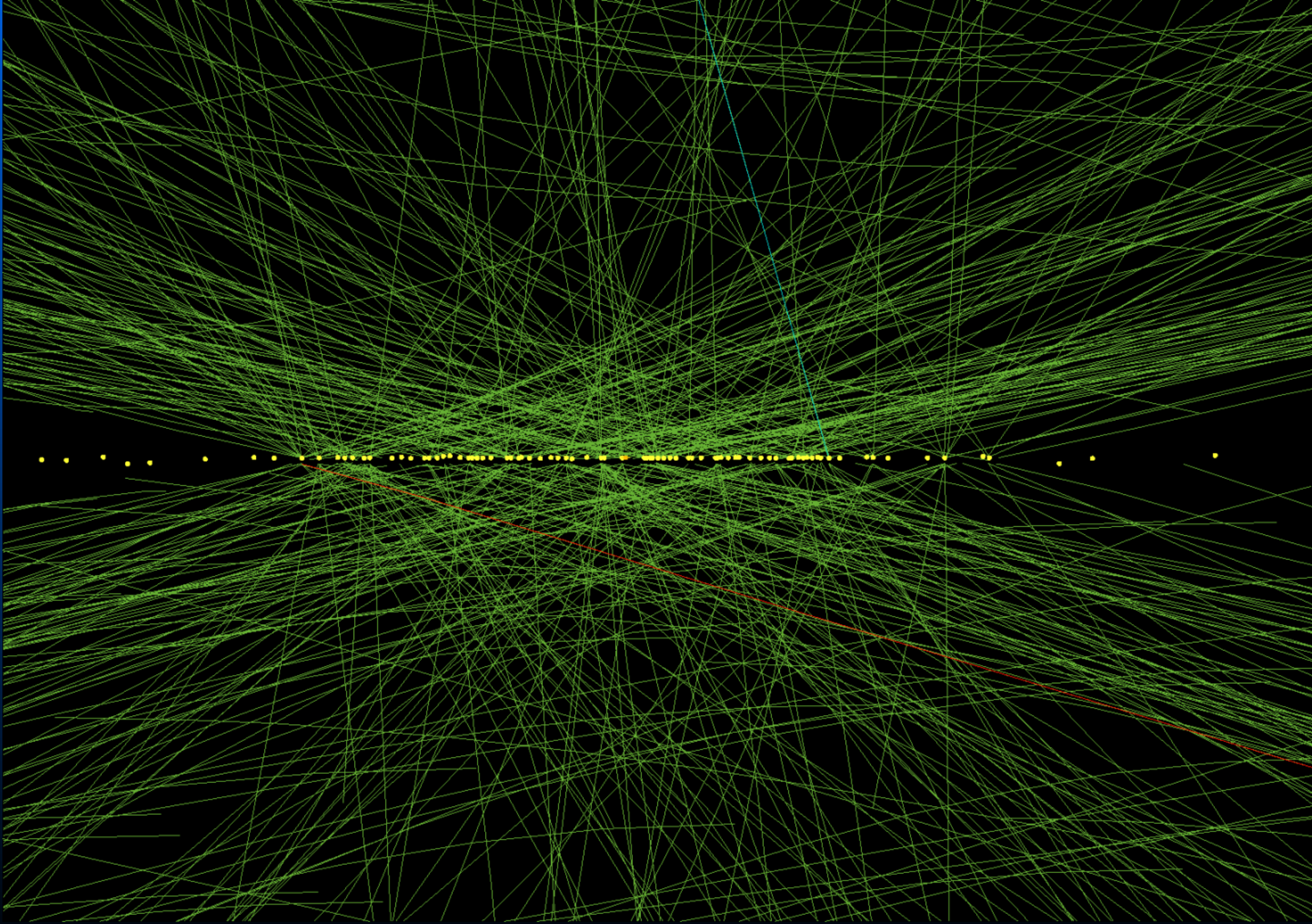
PARTICLE PHYSICS


Collision energy
13 TeV
Total stored energy
720 MJ
Collision rate
1GHz
Length
26.7 km
One Ring to rule them all,
One Ring to find them,
One Ring to bring them all
And in the darkness bind them.
The Large Hadron Collider
Jörg Wenninger



Compact Muon Solenoid

Source: CERN/CMS

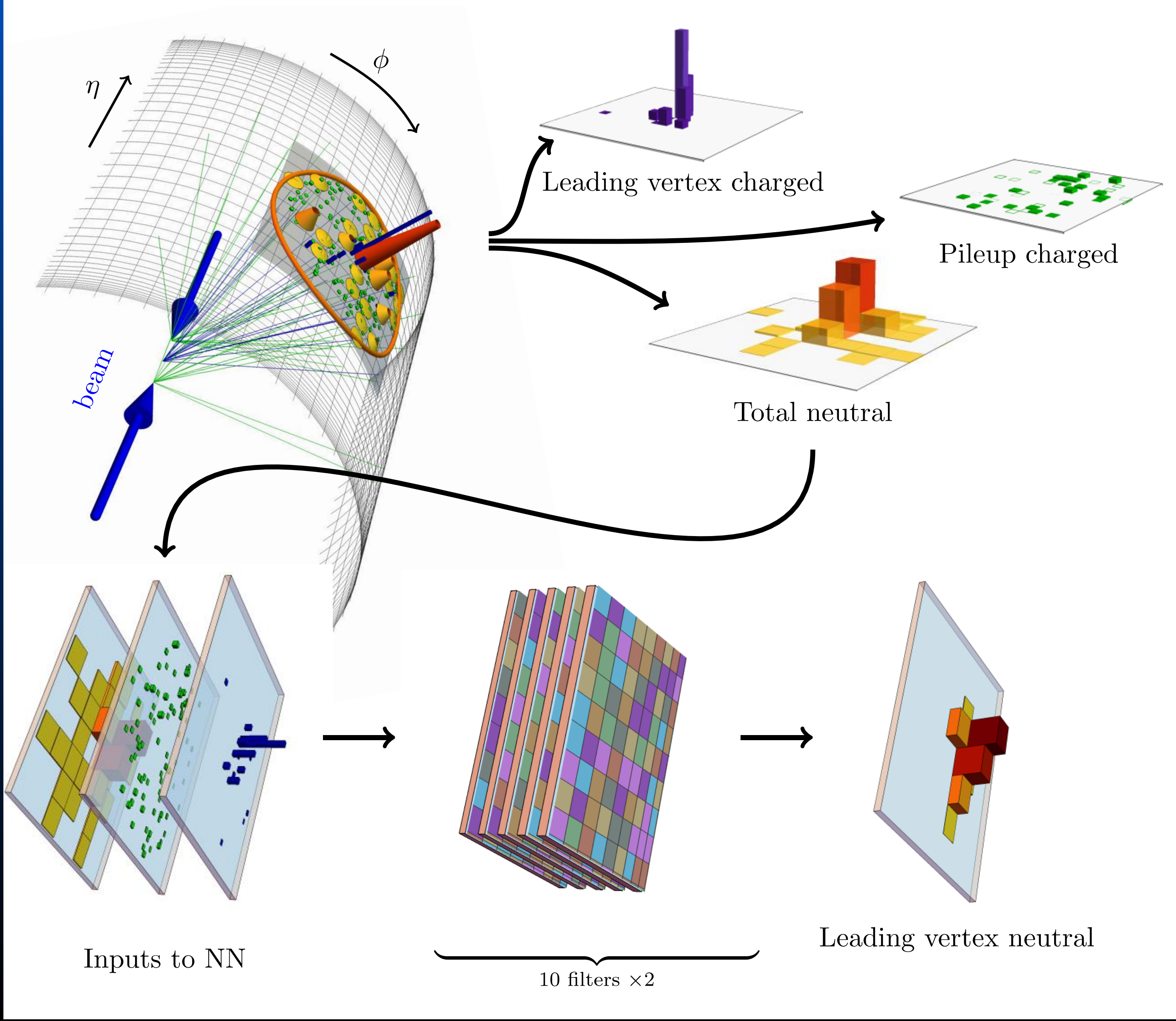


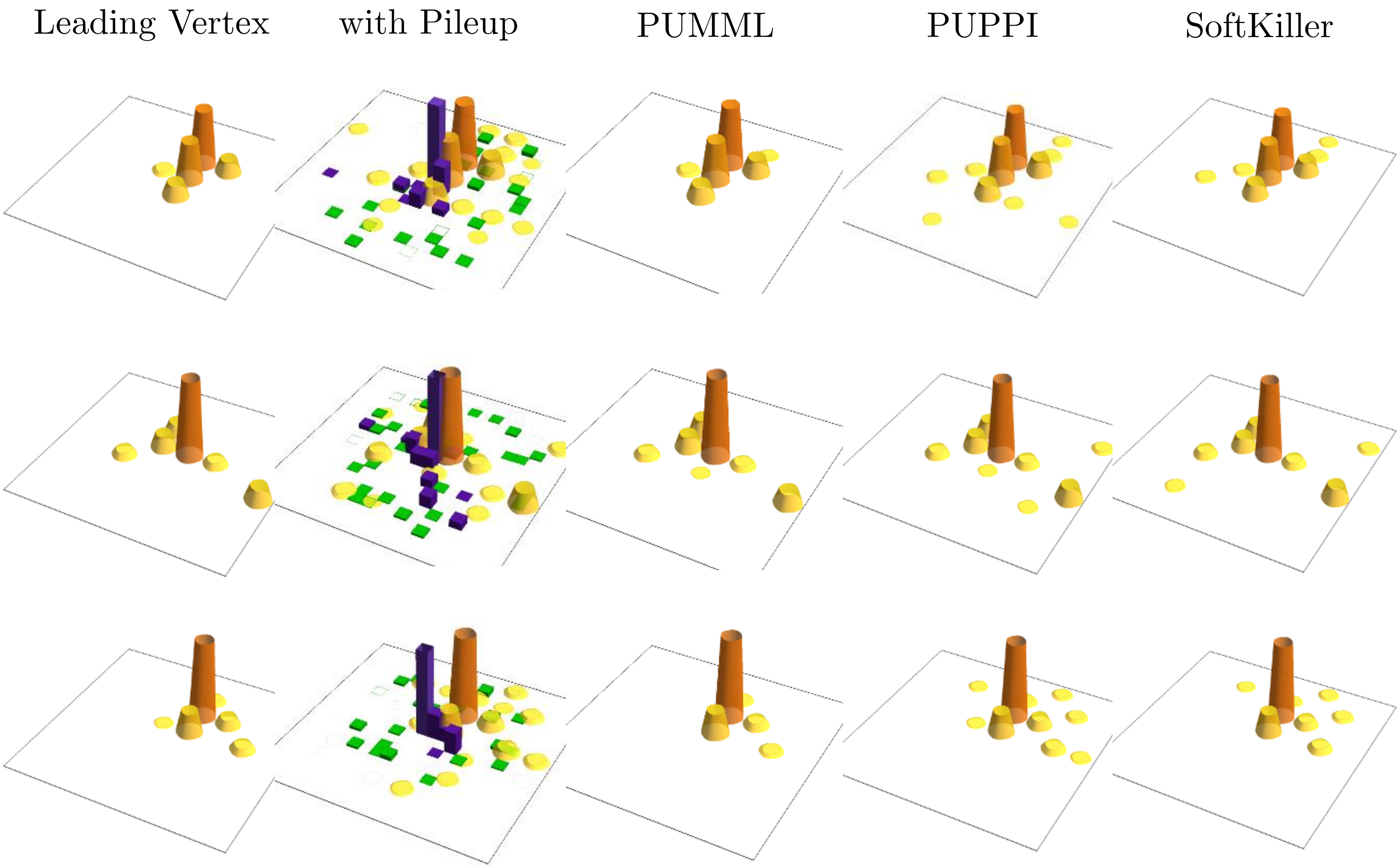
CMS-PHO-EVENTS-2012-006

Pileup Mitigation: PUMML

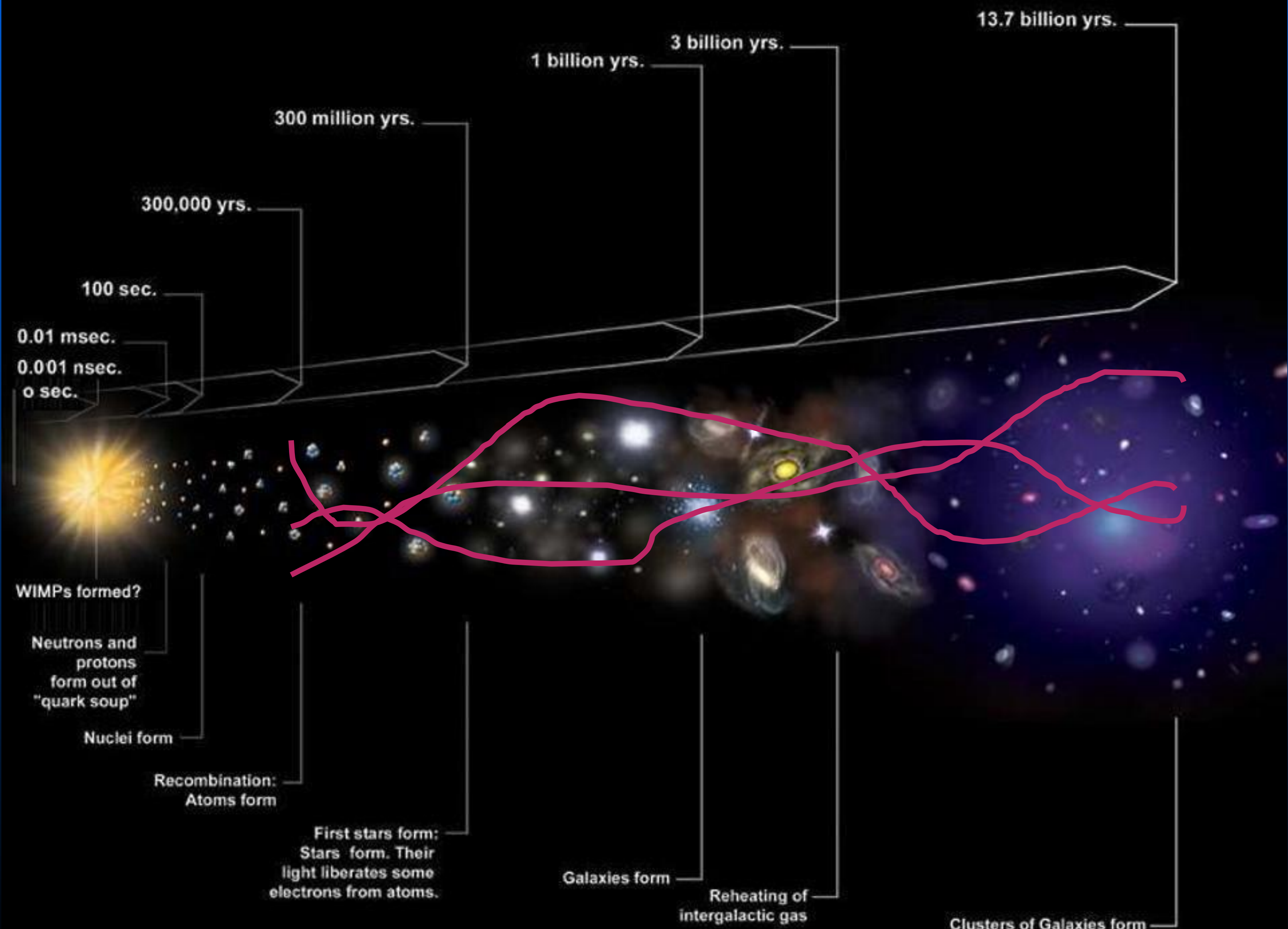
Pileup Mitigation with Machine Learning (PUMML)

Metodiev, Komiske, Nachman, Schwarz, JHEP 12 (2017) 051




Leading Vertex
with Pileup
PUMML
PUPPI
SoftKiller

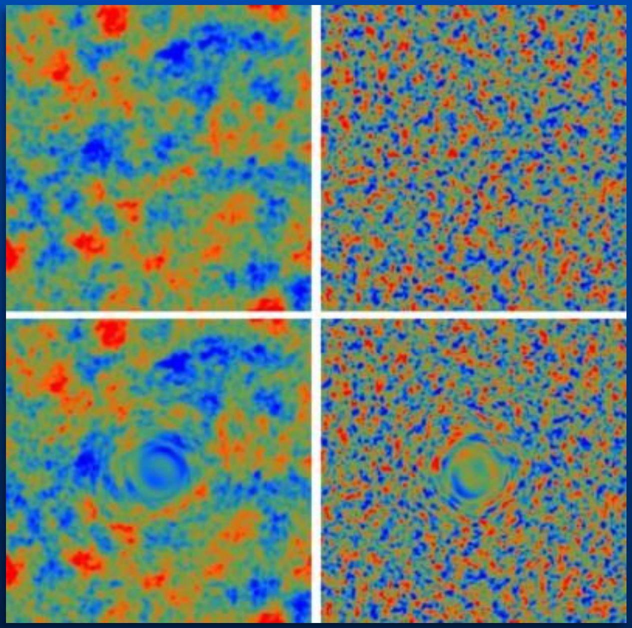
ASTROPHYSICS

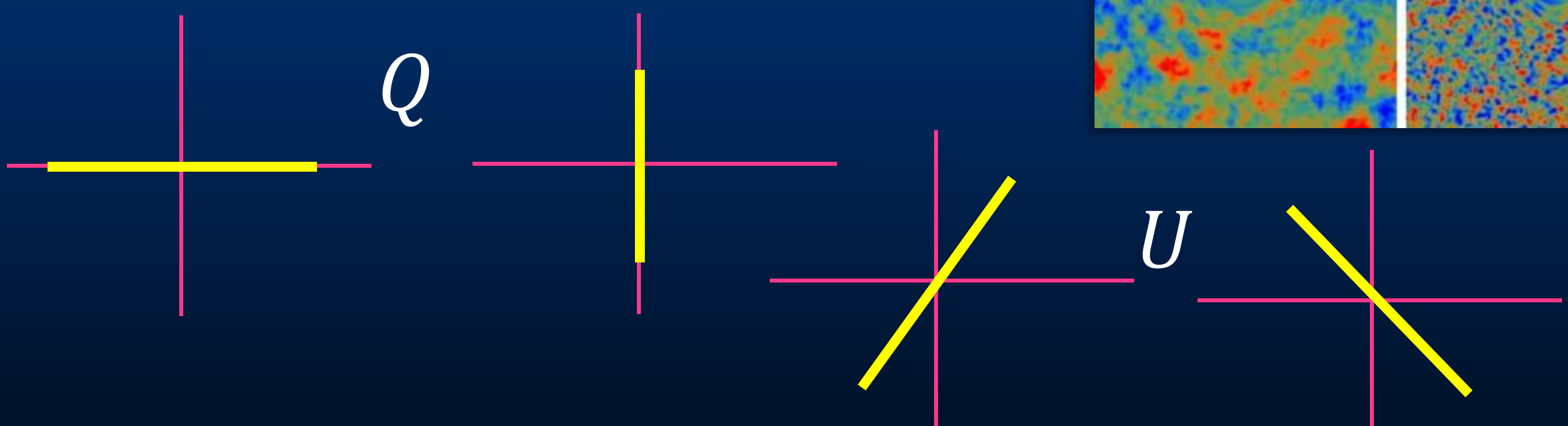


Source: NASA / CXC / M. WEISS

DeepCMB

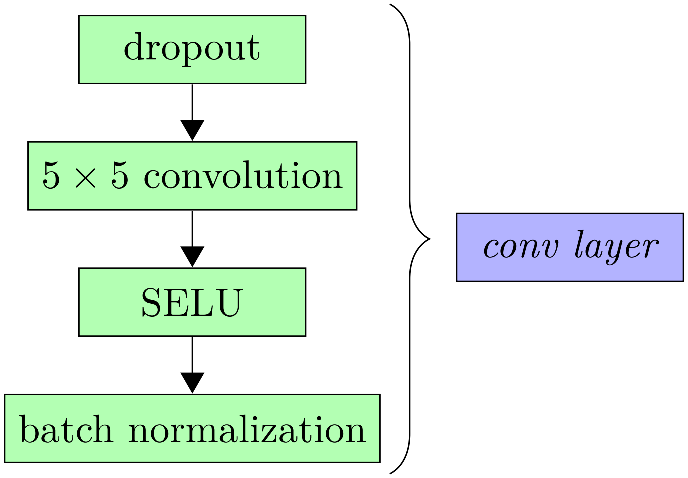
DeepCMB* is a deep neural network that maps 2 gravitationally lensed, 128 x 128 pixel, images of the CMB to 2 un-lensed, 128 x 128 pixel, images.

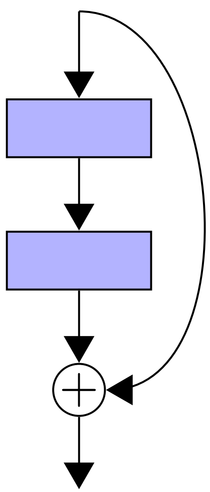


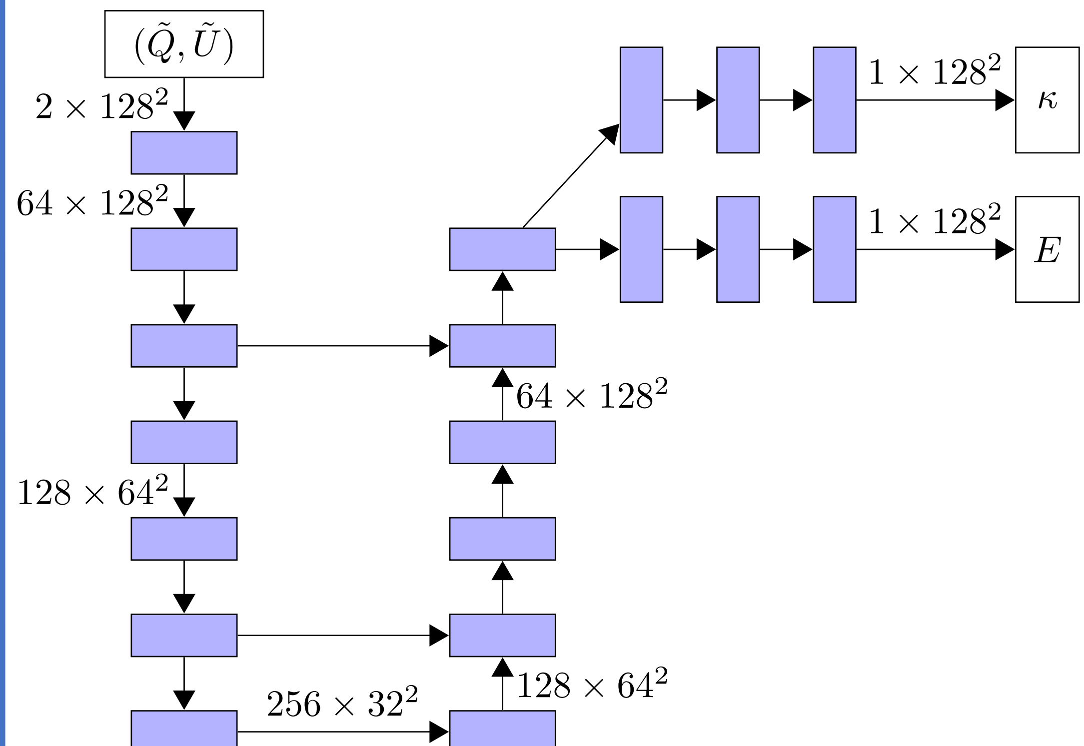


*J. Caldeira, W. L. K. Wu, B. Nord, C. Avestruz, S. Trivedi, K. T. Story, DeepCMB: Lensing Reconstruction of the Cosmic Microwave Background with Deep Neural Networks, **Astronomy and Computing,** 28, July 2019, 100307

Deep CMB: ~ 5 Million Parameters!



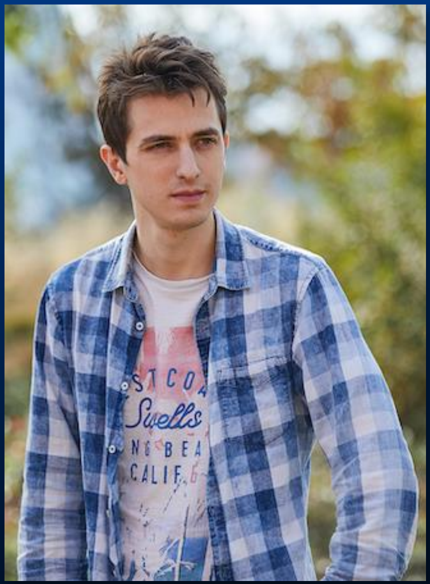




MATHEMATICS

Symbolic Mathematics

In December 2019, Guillaume Lample and François Charton* at Facebook AI Research, Paris, made the startling claim: "*We achieve results that outperform commercial Computer Algebra Systems such as Matlab or Mathematica*."



Lample

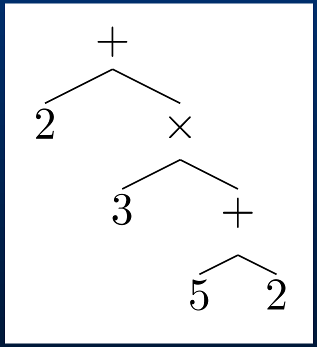
Charton

G. Lample and F. Charton, Deep Learning for Symbolic Mathematics, arXiv: 1912.01412v1.

Symbolic Mathematics

The key idea is to take the idea of mathematics as a *language* seriously. Then, solving a mathematical problem symbolically is analogous to translating from one language to another or rephrasing a sentence.

Consider the expression $2 + 3 \times (5 + 2)$. It is first written as a tree:



Next, the tree is converted to a sequence:

$[+ 2\ \times\ 3 + 5\ 2]$.

Operators, functions, or variables are modeled with specific tokens.

Symbolic Mathematics

The authors' system simplifies, integrates functions, and solves 1st and 2nd order differential equations.

The training data are pairs (x, t) of correctly formed, *randomly generated*, expressions x with associated solutions t.

For example, for integration, at least two approaches are used:

1. Forward: (x, t) where $t = \int x$
2. Backward: (x, t) where $x = Dt$

The Facebook toolkit seq2seq is used to translate one mathematical sequence into another. https:// github.com/facebookresearch/fairseq

Symbolic Mathematics

…and here is a true marvel…

The authors trained their model using the subset of randomly generated functions that sympy can integrate, e.g.,

```python
import sympy as sm
z = sm.Symbol('z')
x = sm.exp(-z)*sm.cos(z)
t = sm.integrate(x, z)
x, t
```

$$\left(e^{-z} \cos(z), \ \frac{e^{-z}\sin(z)}{2} - \frac{e^{-z}\cos(z)}{2} \right)$$

Amazingly, the model was able to integrate functions that sympy could not!

GAME CHANGER

ARTICLE

doi:10.1038/nature24270

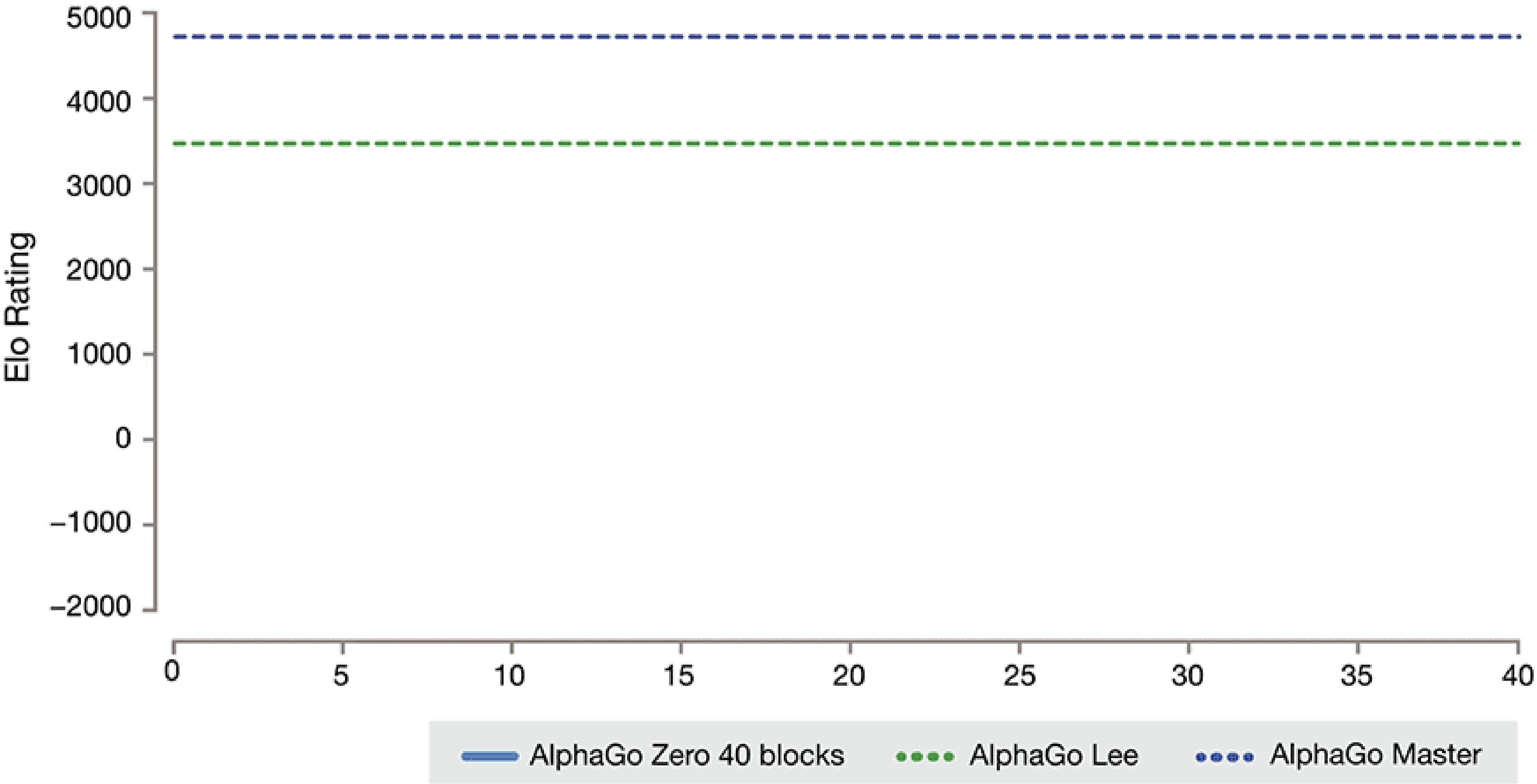
Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.

356 | NATURE | VOL 550 | 19 OCTOBER 2017

https://deepmind.com/blog/alphago-zero-learning-scratch/



THE FUTURE

Science 2071

What might another 50 years of development do to the practice of science?

Potential resources:

- AI-enhanced symbolic mathematics
- High-fidelity, super-fast, simulators
- High-fidelity natural language processors
- Autonomous open data acquisition systems (accelerators, planet-wide biosphere-monitors, planet-wide telescope swarms, planet-wide personal health monitors, …)

If some of this remains, in part, public domain and **If** the Open Science movement takes hold, perhaps science will finally become democratized.

The Good

- Autonomous mathematical assistant
- Autonomous transportation
- Direct brain / machine interfaces
- Personal tutors
- Conflict resolution systems

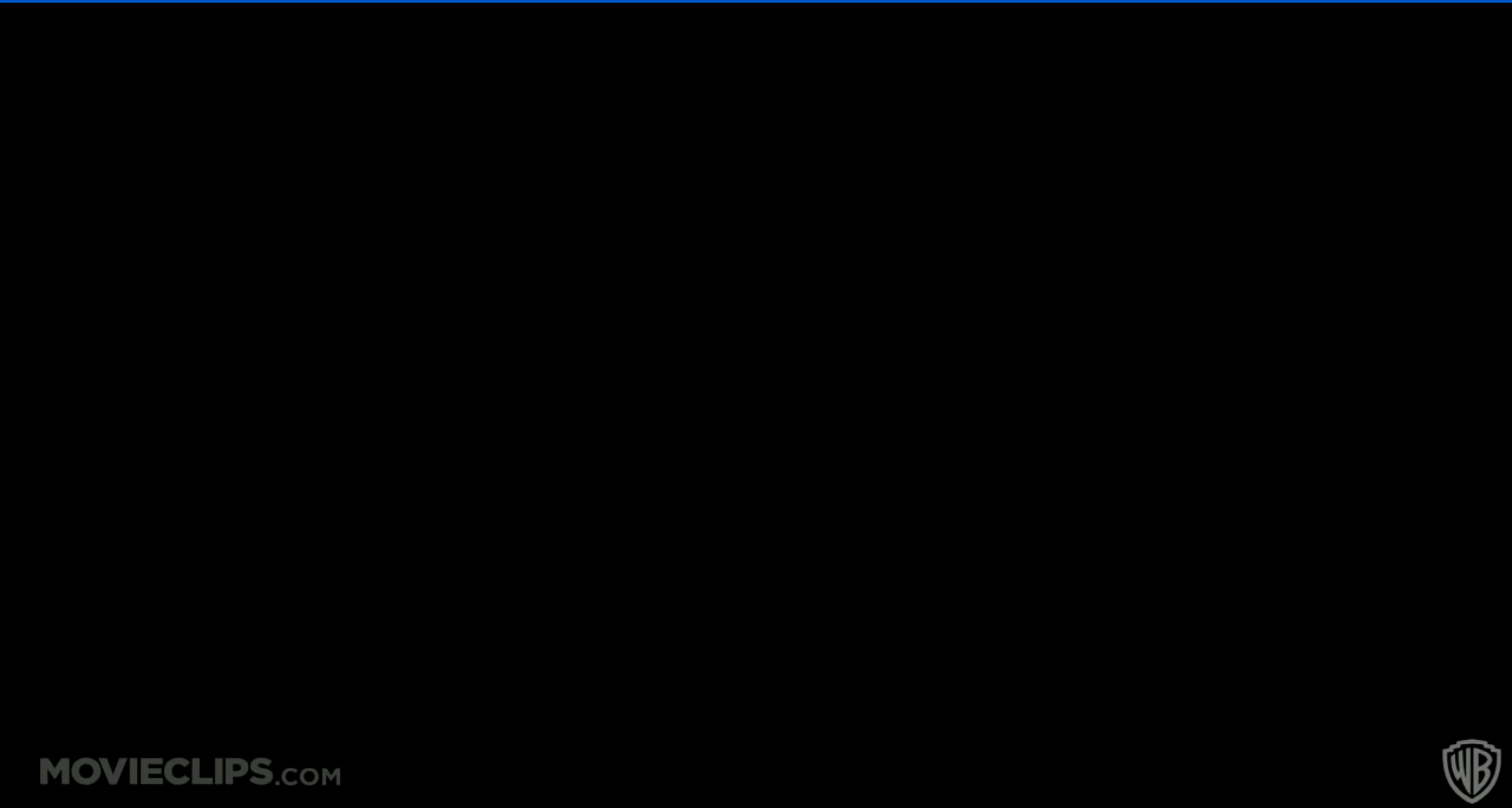
The Bad

- AI-enhanced micro-drones
- AI-enhanced predictive behavioral agents (“profiling”)
- AI-enhanced computer viruses

The Ugly

- The Super-Rich may become the Hyper-Rich.
- The disparity between the haves and have-nots acquires another dimension: Artificial Cognitive Enhancement.
- Social instability arising from acute economic dislocation as AI-enhanced automation displaces more and more traditional jobs.

…and what might happen if we delegate more and more decision making to our tireless assistants: decisions about credit worthiness, loan eligibility, job suitability, likelihood of illness, treatments, life partners, trustworthiness, etc. …?


MOVIECLIPS.COM


The best way to predict the future is to create it.
Peter Drucker
THANK YOU!

SOME TECHNICAL MATERIAL

Machine Learning

Machine learning algorithms fall into five broad categories:

1. Supervised Learning
2. Semi-supervised Learning
3. Unsupervised Learning
4. Reinforcement Learning
5. Generative Learning

Machine Learning

M = Function class

Method

Choose $f\,(x, \theta^*)$ from ***M*** by minimizing the *average loss* (or *empirical risk*)

$$F(\theta) = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{L}(\boldsymbol{y_i}, \boldsymbol{f_i}) + C(\boldsymbol{\theta}),$$

where

$D = \{(y_i\,, x_i)\}$ are training data,

f_i $f(x, \boldsymbol{\theta})$ evaluated at x_i, and

$\boldsymbol{L}(\boldsymbol{y_i}, \boldsymbol{f_i})$, the *loss function*, is a measure of the quality of the choice of function.

$C(\theta)$ is a constraint that guides the choice of $f\,(x, \theta)$.

Minimizing the Average Loss

The average loss function defines a “landscape” in the *space of functions*, or, equivalently, the space of parameters.

The goal is to find the lowest point in that landscape, by moving in the direction of the *negative* gradient:

$$\theta_i \leftarrow \theta_i - \rho \frac{\partial F(\theta)}{\partial \theta_i}$$

Most minimization algorithms are variations on this theme.

Stochastic **G**radient **D**escent (SGD) uses random subsets (*batches*) of the training data to provide *noisy* estimates of the gradient in order to increase the chance of escaping from local minima.

Minimizing the Average Loss

Consider $F(\theta)$ in the limit $T \to \infty$

$$F(\theta) = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{L}(\boldsymbol{y}_i, \boldsymbol{f}_i) + C$$

$$\to \int dx \int dy\, \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{f})\, p(y, x)$$

Since $p(y|x) = p(y, x)/p(x)$ we can write

$$= \int dx\, p(x) \left[\int dy\, \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{f})\, p(y|x)\right]$$

We have assumed the influence of the constraint to be negligible in this limit.

Minimizing the Average Loss

Now, consider the *quadratic* loss $\boldsymbol{L(y, f) = (y - f)^2}$

$$F = \int dx\, p(x) \left[\int dy\, \boldsymbol{L(y, f)}\, p(y|x) \right]$$
$$= \int dx\, p(x) \left[\int dy\, (\boldsymbol{y} - \boldsymbol{f})^{\boldsymbol{2}}\, p(y|x) \right]$$

and its minimization with respect to the choice of function ***f***.

Minimizing the Average Loss

If we change the function $\boldsymbol{f}$ by a small *arbitrary* function $\boldsymbol{\delta f}$ a small change

$$\boldsymbol{\delta F} = 2\int dx\, p(x)\boldsymbol{\delta f}\left[\int dy(y - f)p(y|x)\right]$$

will be induced in *F*. In general, $\boldsymbol{\delta F} \neq 0$. However, if the function $\boldsymbol{f}$ is flexible enough then we shall be able to reach the minimum of *F*, where $\boldsymbol{\delta F} = 0$. But, in order to guarantee that $\boldsymbol{\delta F} = 0$ *for all* $\boldsymbol{\delta f}$ and *for all* ***x*** the quantity in brackets must be *zero*. This yields the important result:

$$\boxed{f(x, \theta^*) = \int y\, p(y \mid x)\, dy}$$

Classification

According to *Bayes' theorem*

$$p(y|x) = \frac{p(x\,|y)\,p(y)}{\int p(x\,|y)\,p(y)\,dy}$$

Let's assign the *target* value ***y* = 1** to objects of class ***S*** and the target value ***y* = 0** to objects of class ***B***.

Then

$$\boxed{f(x,\theta^*) = \int y\,p(y\,|\,x)\,dx = p(1|x) \equiv p(\boldsymbol{S}|x)}$$

That is, the function $f(x,\theta^*)$ equals the *class probability*.

Classification

1. In summary, the result

$$f(x, \theta^*) = p(S|x) = \frac{p(x|S)p(S)}{p(x|S)p(S) + p(x|B)p(B)}$$

depends *only* on the *form* of the loss function, provided that:

   1. the training data are sufficiently numerous,
   2. the function $f(x, \theta)$ is sufficiently flexible, and
   3. the minimum of the average loss, F, can be found.

2. Note, if $p(S) = p(B)$, we arrive at the *discriminant*

$$D(x) = \frac{p(x|S)}{p(x|S) + p(x|B)} \equiv \frac{s(x)}{s(x) + b(x)}$$