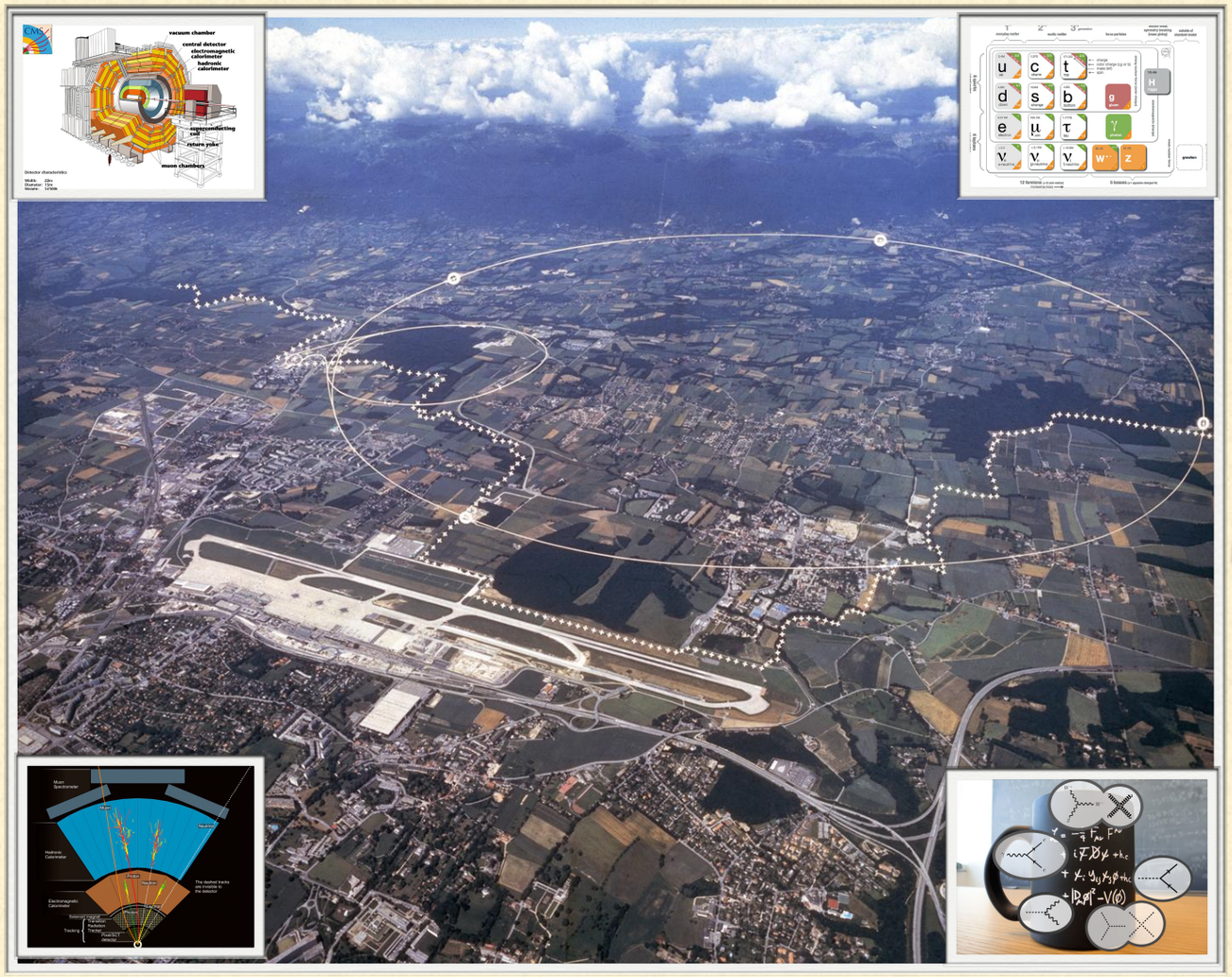
The AI/ML Driven Future of Particle Physics and Cosmology



Partha Konar
THEPH @ PRL

https://www.prl.res.in/~konar/

Partha Konar,
Monalisa Patra,
Sanmay Ganguly

<=========== 75 min ===========>

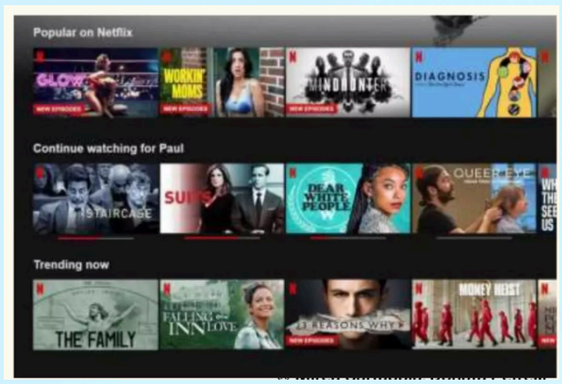
Centre for High Energy Physics

Frontiers of Particle Physics

August 9-11, 2024

While world is mesmerised with different impossibles
done with ML applications in our everyday life,

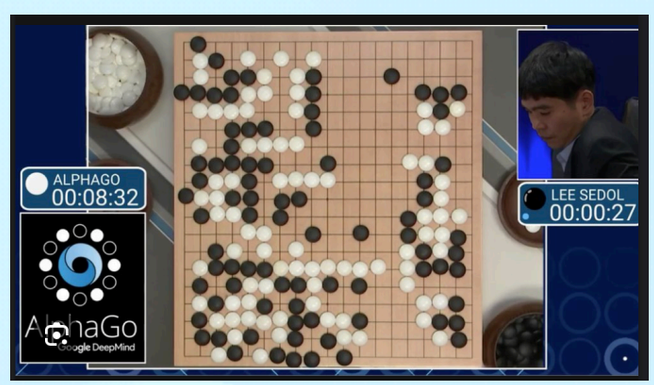
Content Recommendation



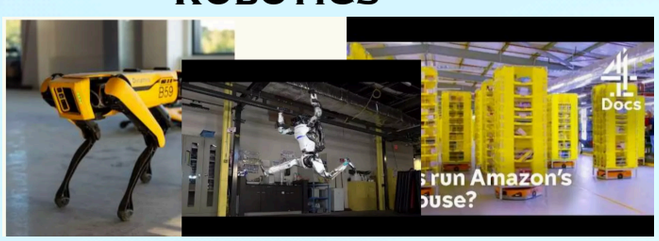
Self-driving Car

Face Recognition

Gaming



Robotics



Creativity Culture music

Virtual Assistance

Artificial Intelligence in Everyday Apps

Predictive Search

Object Detection

News Feed Relevance

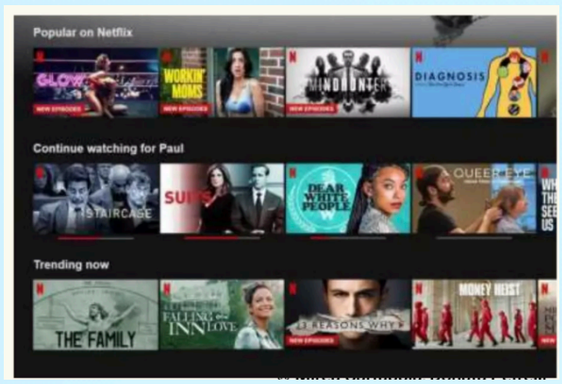
Recommendations

Matching Algorithm

Smart Replies

While world is mesmerised with different impossibles
done with ML applications in our everyday life,

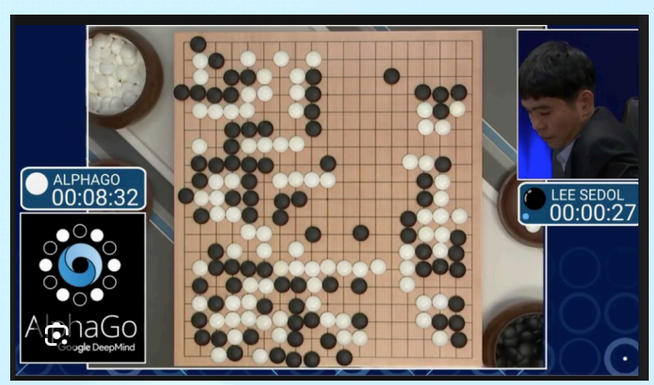
Content Recommendation



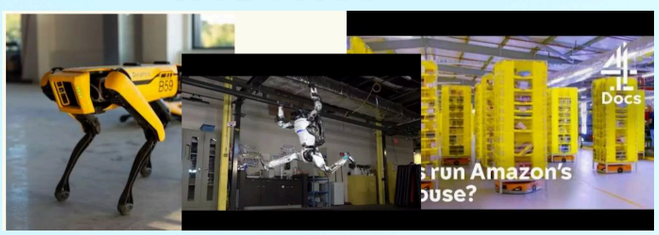
Self-driving Car

Face Recognition

Gaming



Robotics



Creativity Culture music

Virtual Assistance

ChatGPT Gemini

Artificial Intelligence in Everyday Apps

Predictive Search

Object Detection

News Feed Relevance

Recommendations

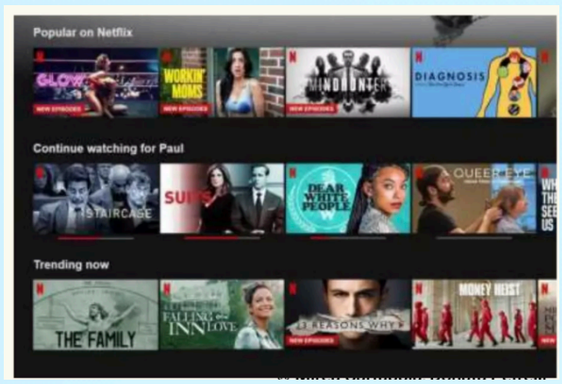
Matching Algorithm

Smart Replies

Dramatic shifts are also happening in almost all research fields —
including Healthcare, Medicine, Finance, Education services etc

While world is mesmerised with different impossibles
done with ML applications in our everyday life,

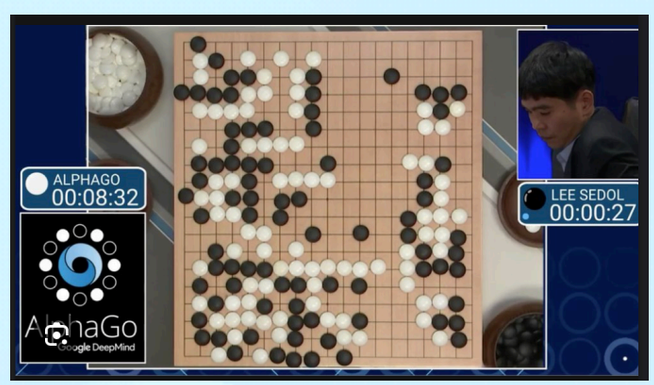
Content Recommendation



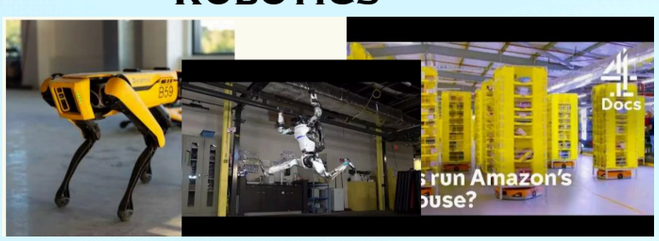
Self-driving Car

Face Recognition

Gaming



Robotics



ChatGPT Gemini

Artificial Intelligence in Everyday Apps

Creativity Culture music

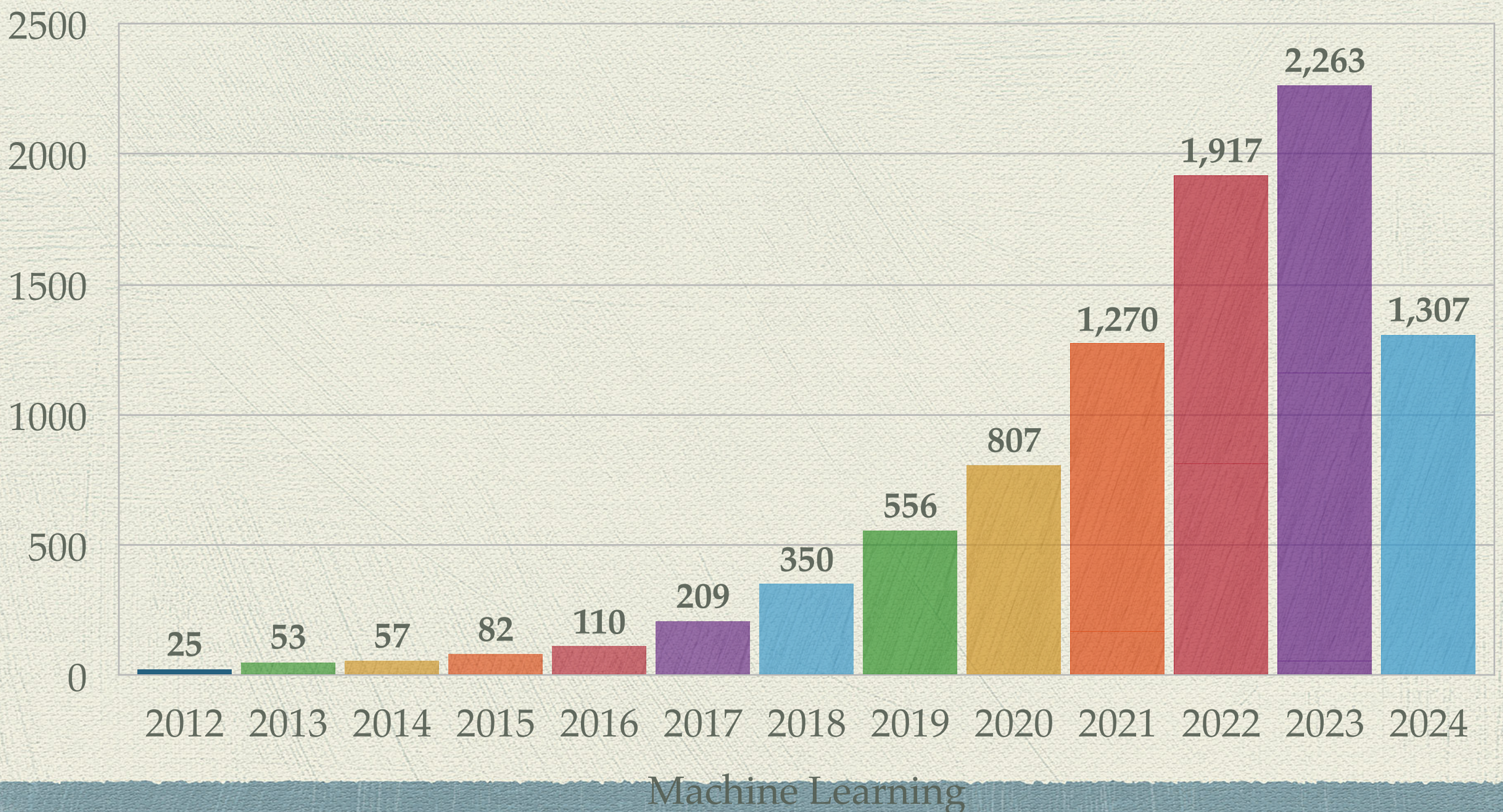
Virtual Assistance

Predictive Search | Object Detection | News Feed Relevance

Recommendations | Matching Algorithm | Smart Replies

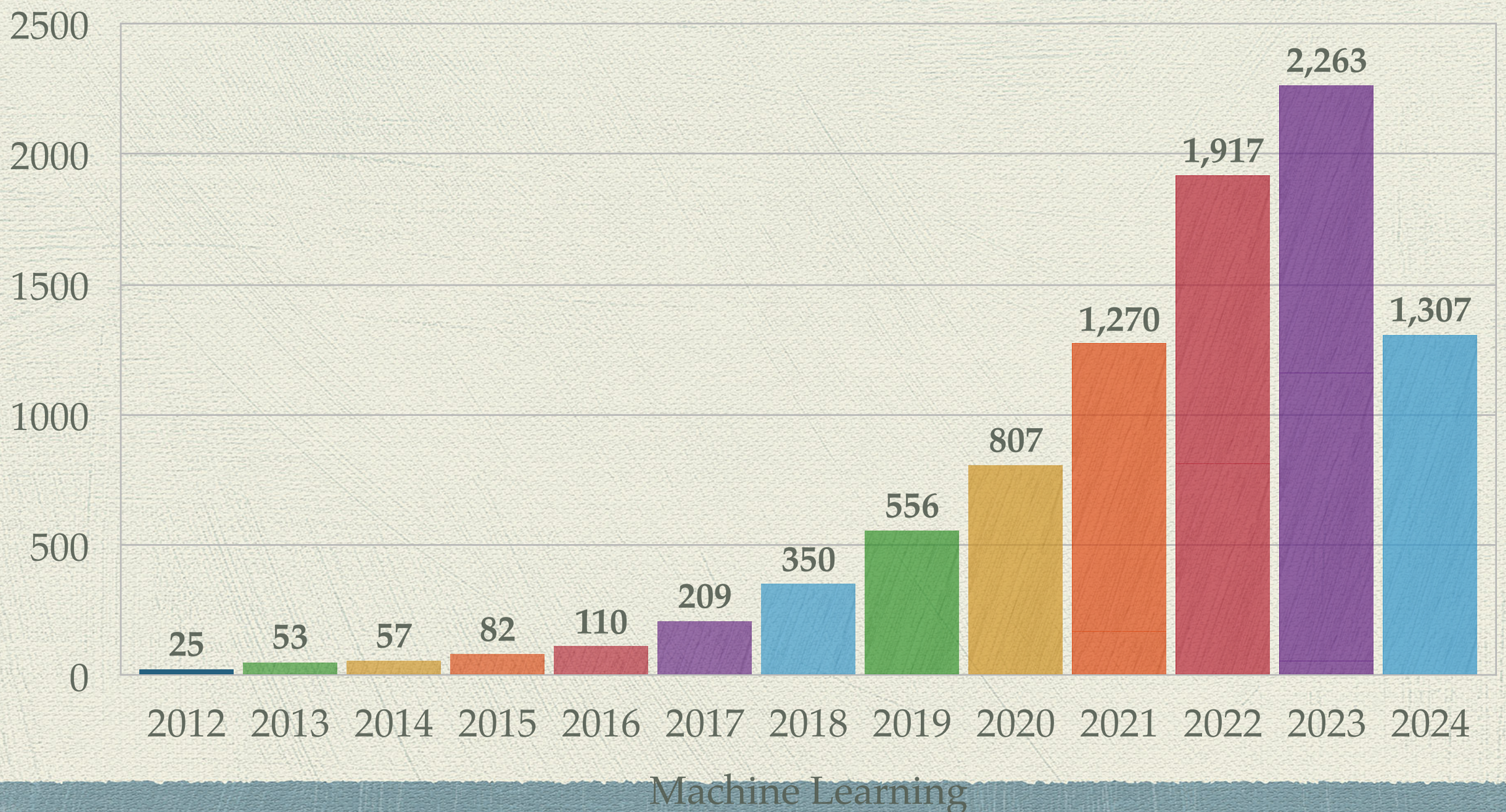
Dramatic shifts are also happening in almost all research fields —
including Healthcare, Medicine, Finance, Education services etc

Several experimental results found their relevance — such scientific discoveries are ML driven


2500
2000
1500
1000
500
0
25
53
57
82
110
209
350
556
807
1,270
1,917
2,263
1,307
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Machine Learning

- Inspire-HEP literatures with `machine learning'

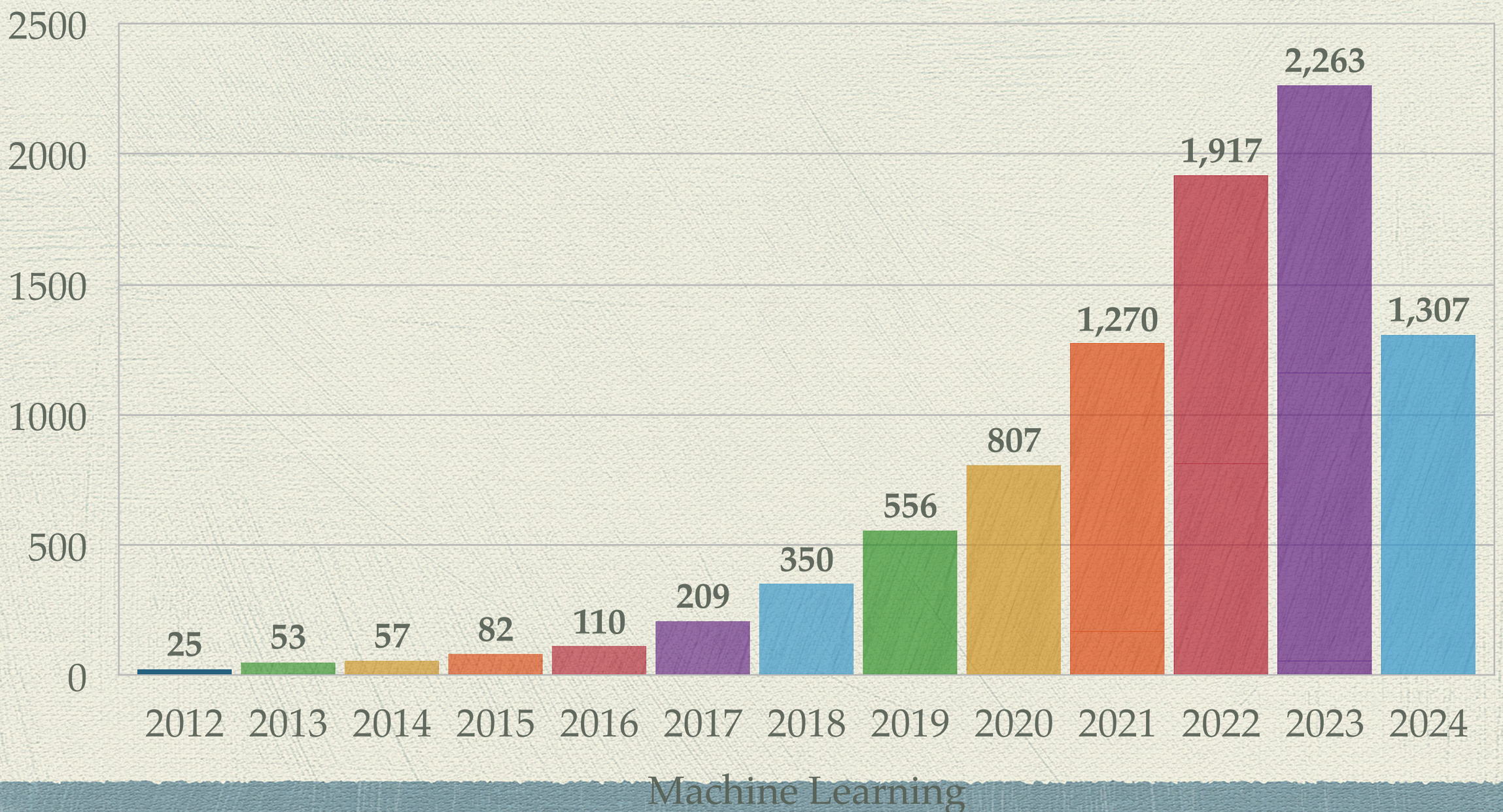
2012 => 2024 : increased by 90 times



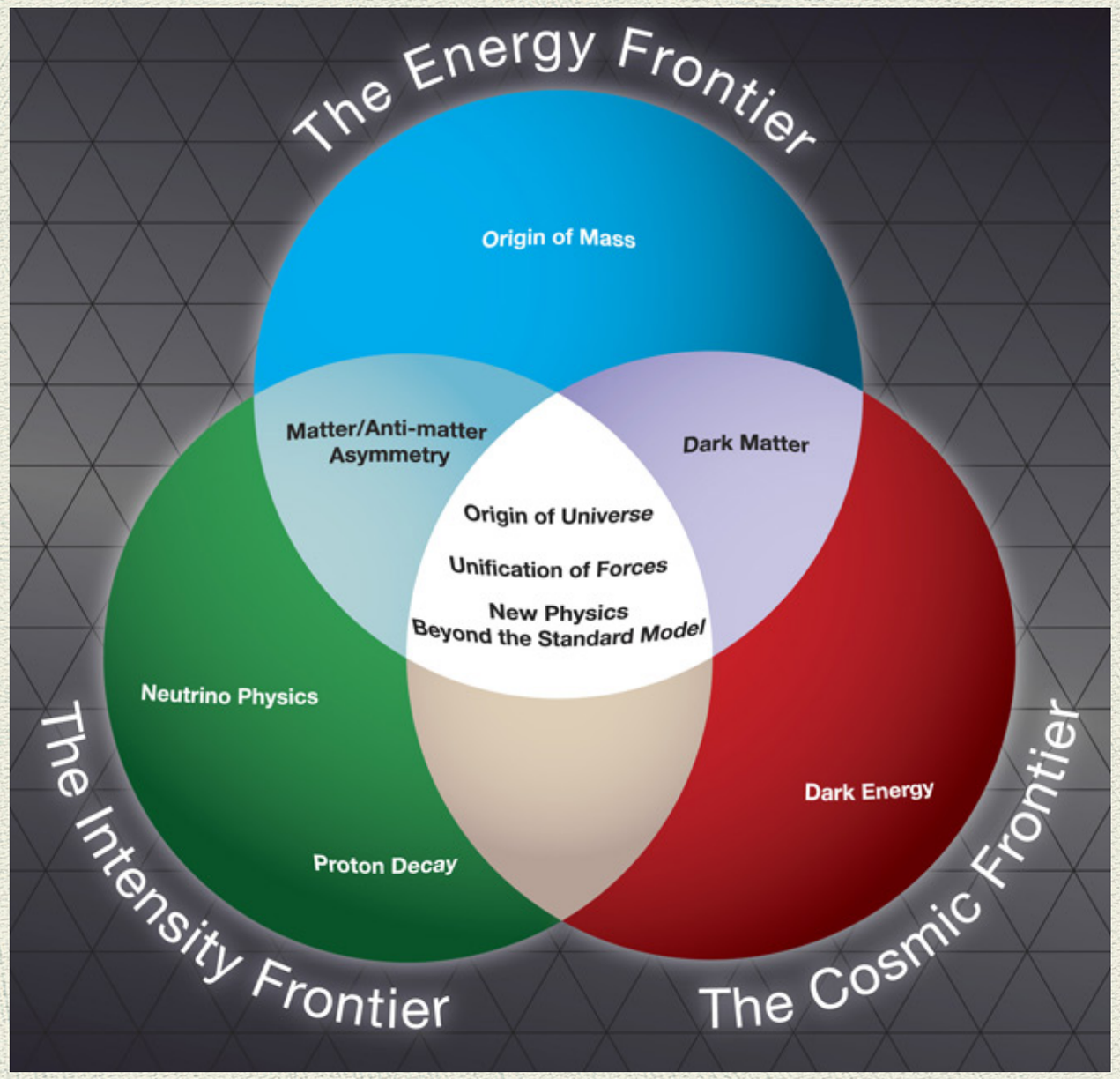
- Inspire-HEP literatures with `machine learning'

2012 => 2024 : increased by 90 times

- **Whereas, SENSEX moved 17k=>74k — only 4.4 times!**



Particle Physics

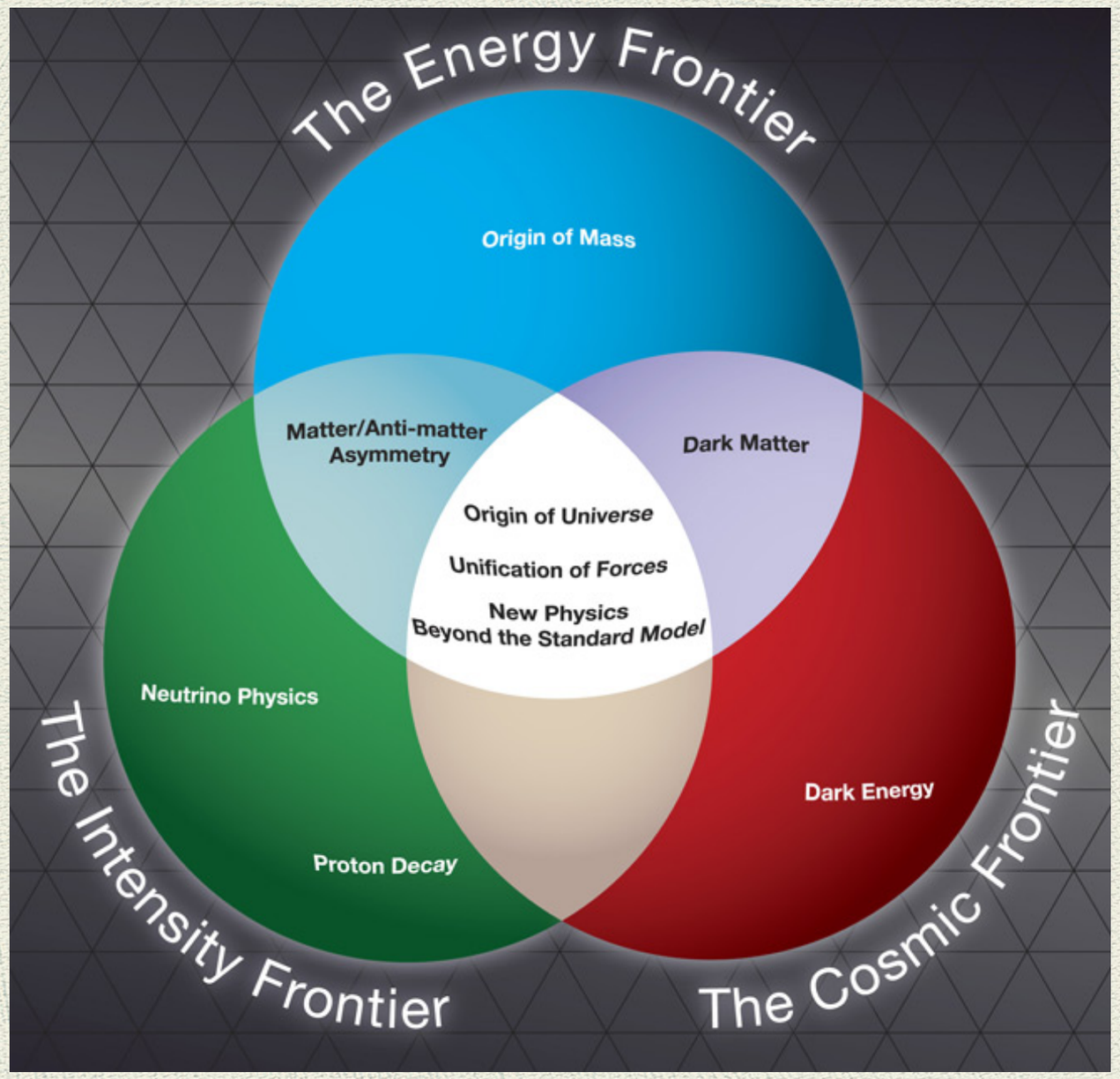


- 2210.05822 : The Future of High Energy Physics Software and Computing

Particle Physics

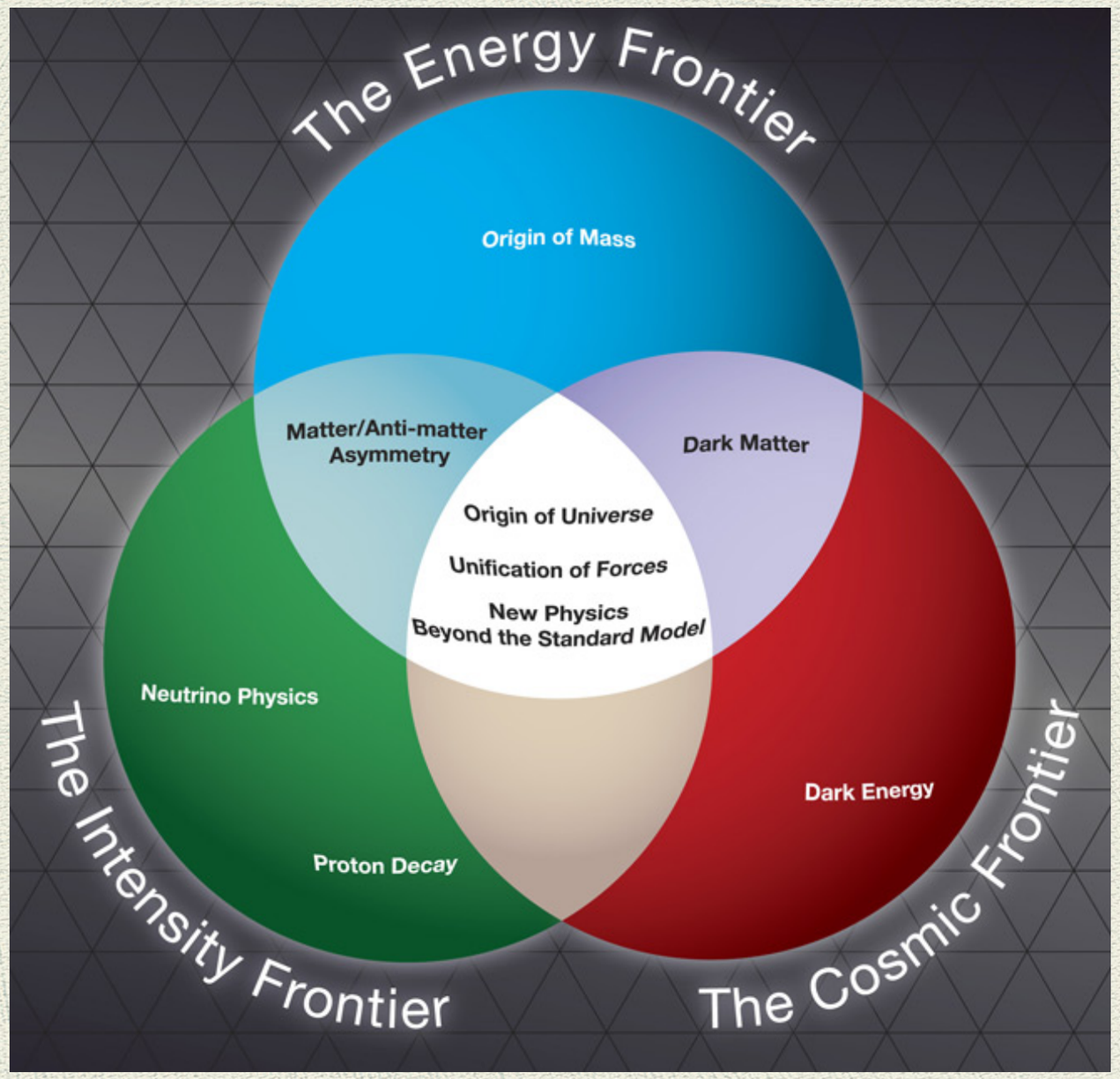
&

Computational Frontier



- 2210.05822 : The Future of High Energy Physics Software and Computing

Particle Physics



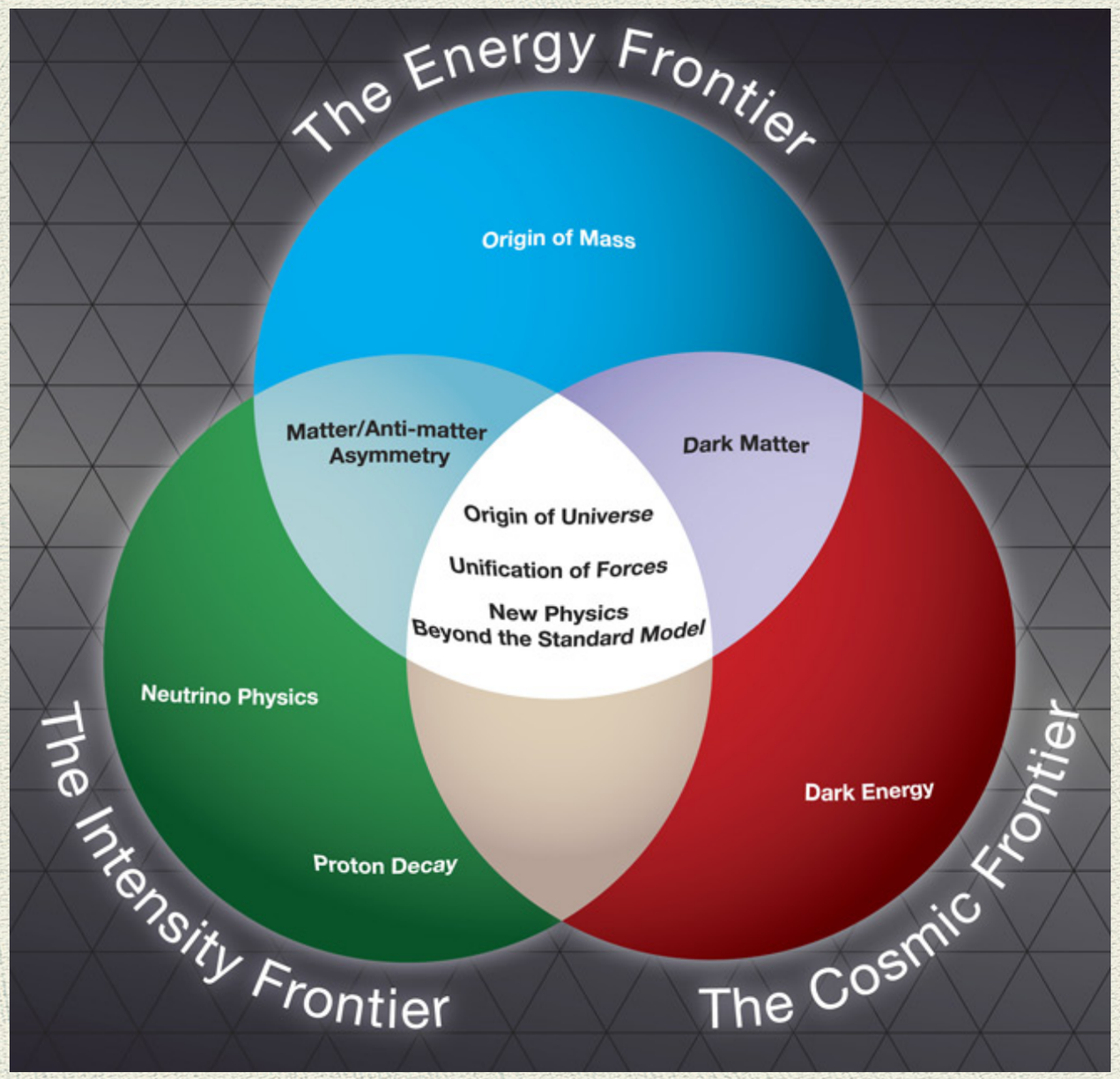
&

Computational Frontier

- **Software and computing essentially present in all fronts**

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics



&

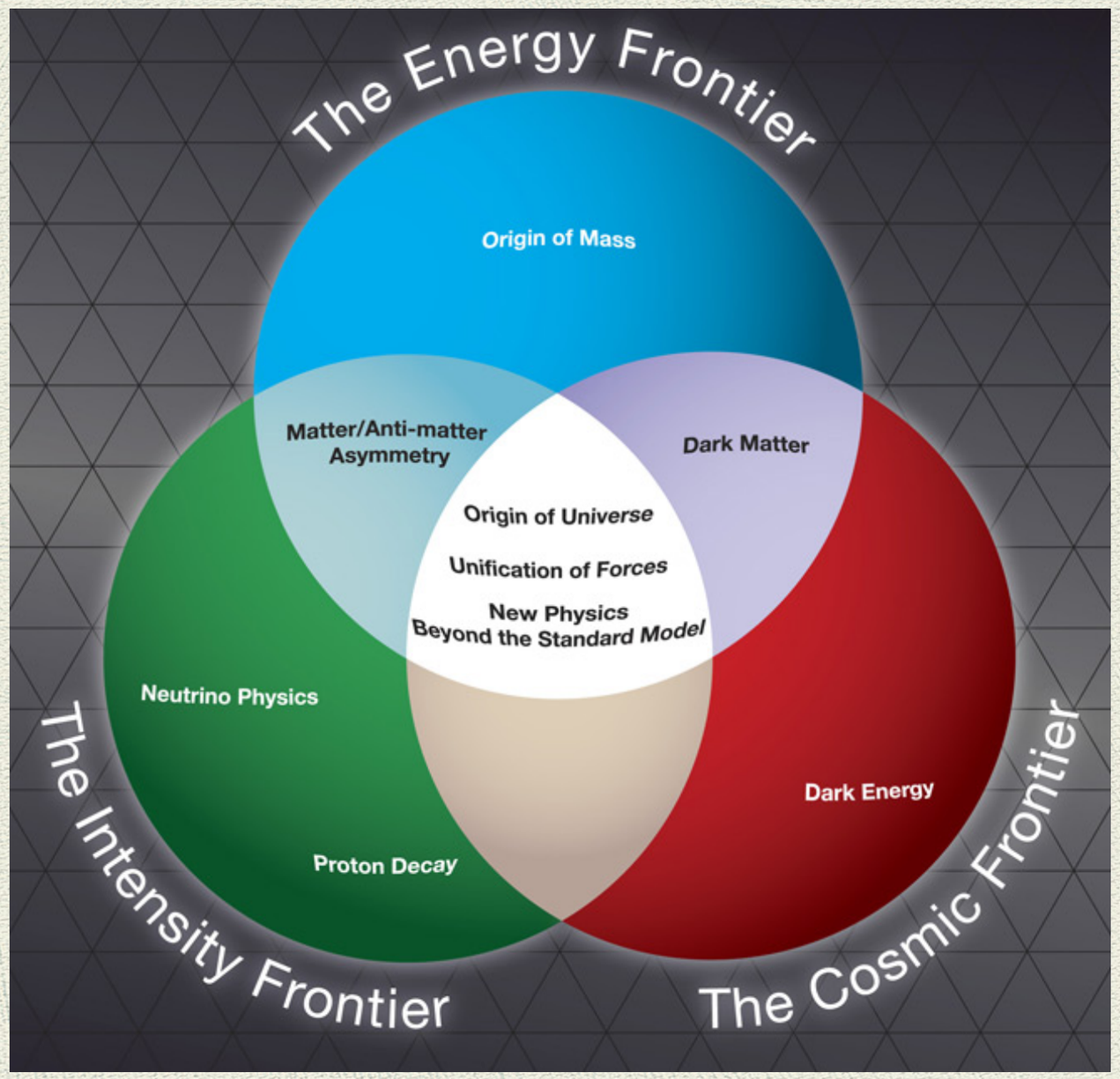
Computational Frontier

- **Software and computing essentially present in all fronts**
- Computing/methodological innovation

  => Full advantage of pristine data by state-of-the-art instruments

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics



&

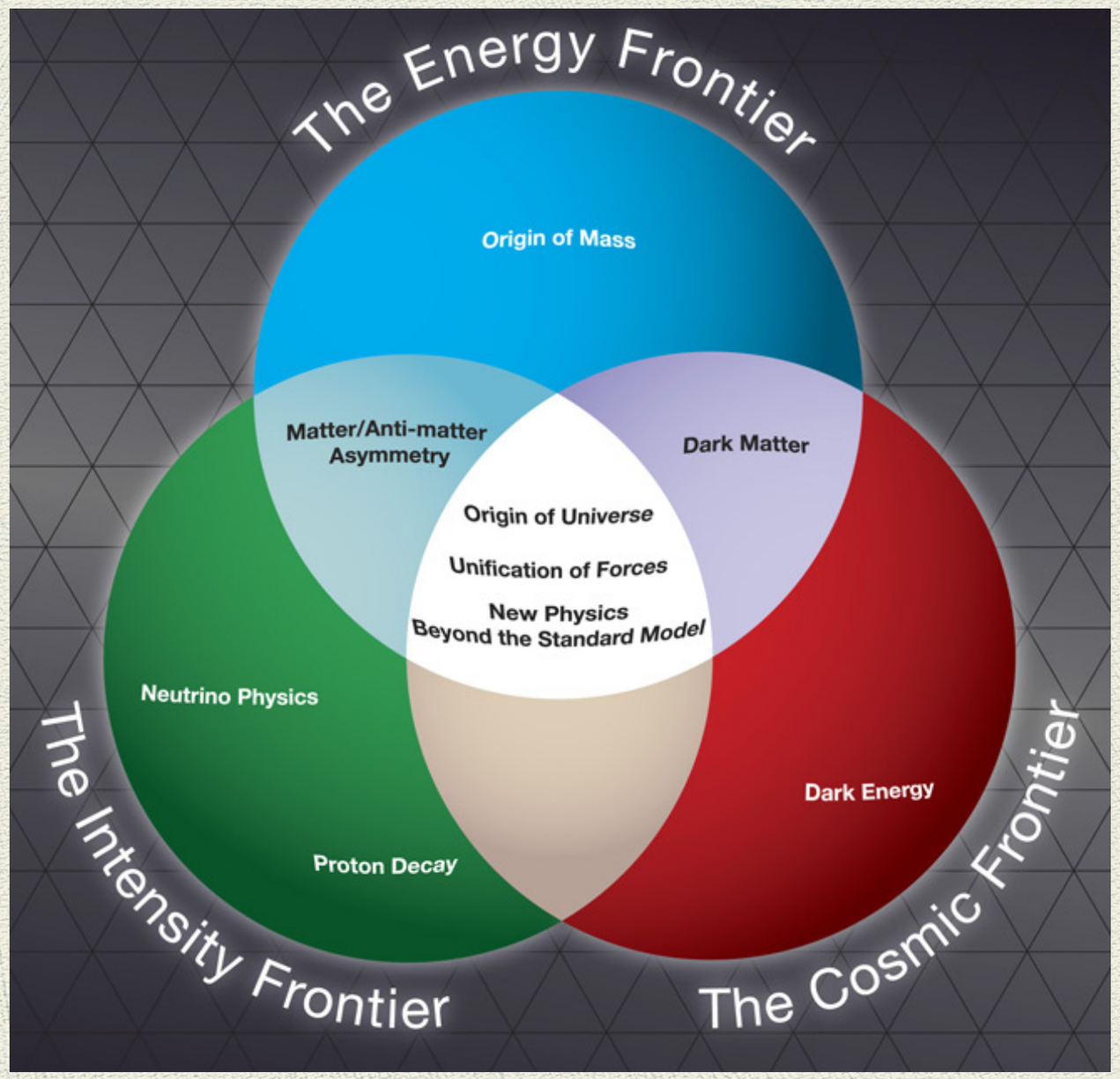
Computational Frontier

- **Software and computing essentially present in all fronts**
- Computing/methodological innovation

  => Full advantage of pristine data by state-of-the-art instruments

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics



&

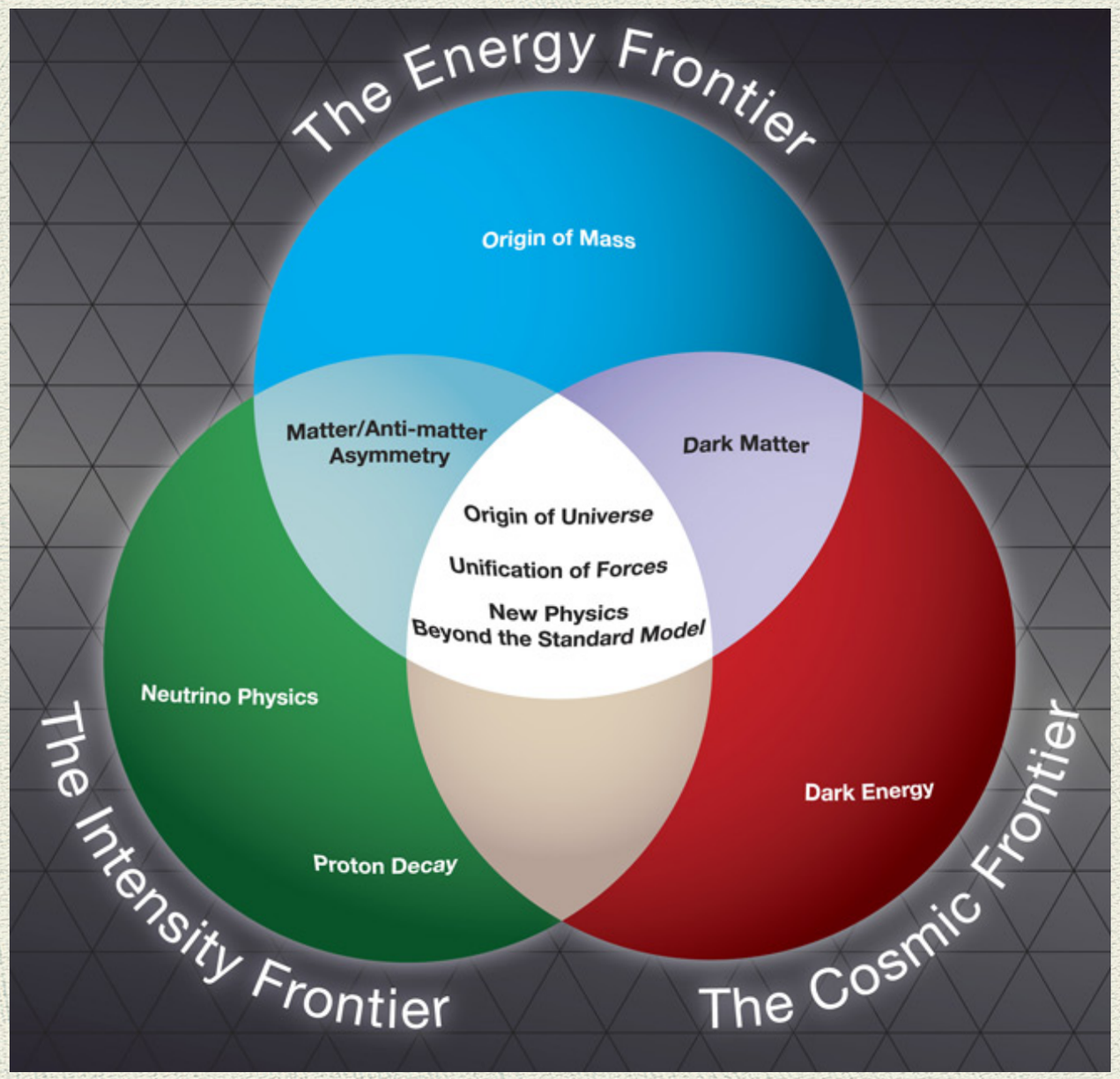
Computational Frontier

- **Software and computing essentially present in all fronts**
- Computing/methodological innovation
  => Full advantage of pristine data by state-of-the-art instruments

- **[2/Top 10] most-cited papers of all time in particle physics — are software programs :**
  GEANT [Detector Simulation Toolkit]
  & PYTHIA [Generation of HEP collision events]

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics



&

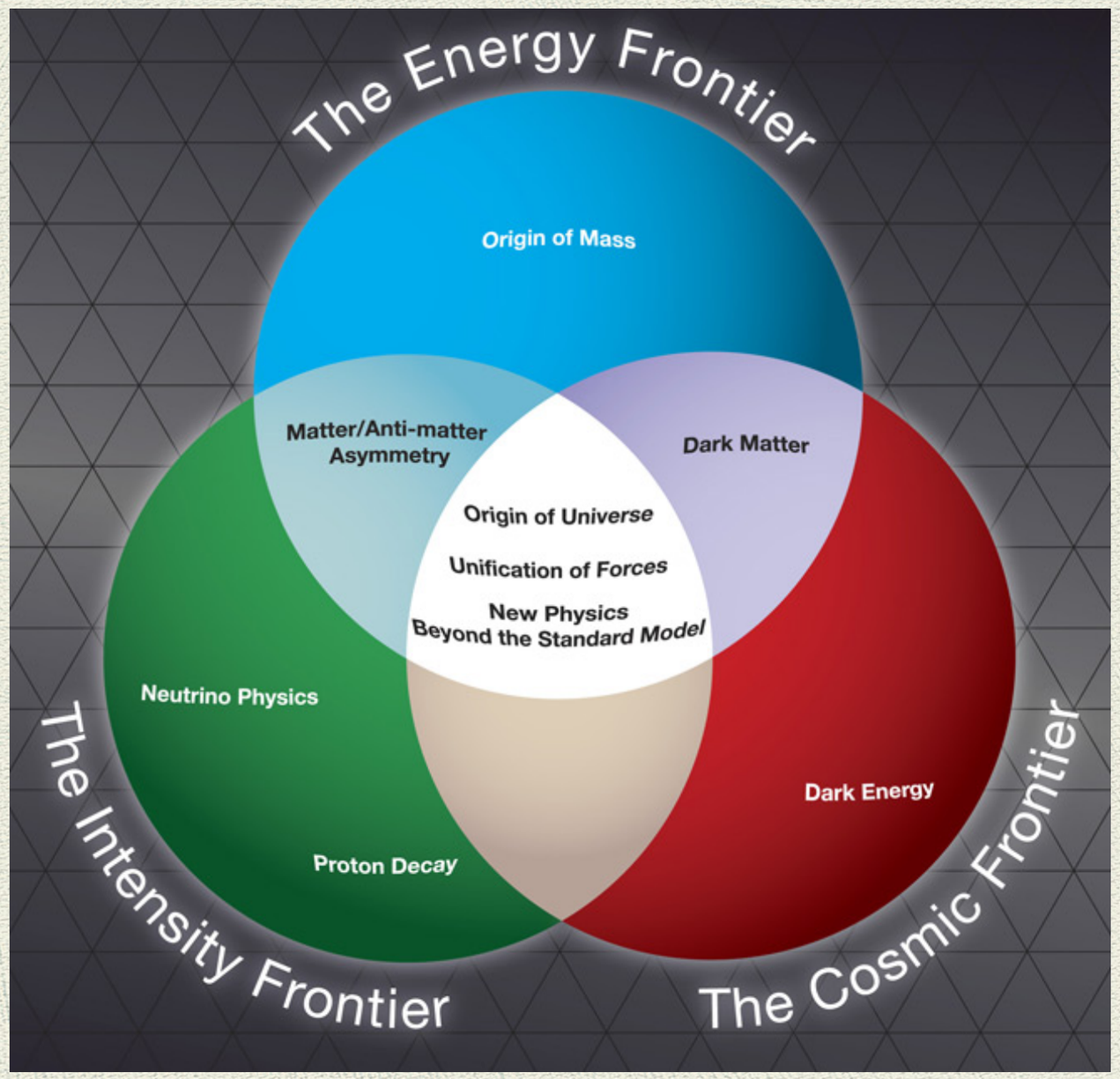
Computational Frontier

- **Software and computing essentially present in all fronts**
- Computing/methodological innovation
  => Full advantage of pristine data by state-of-the-art instruments

- **[2/Top 10] most-cited papers of all time in particle physics — are software programs :**
  GEANT [Detector Simulation Toolkit]
  & PYTHIA [Generation of HEP collision events]

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics



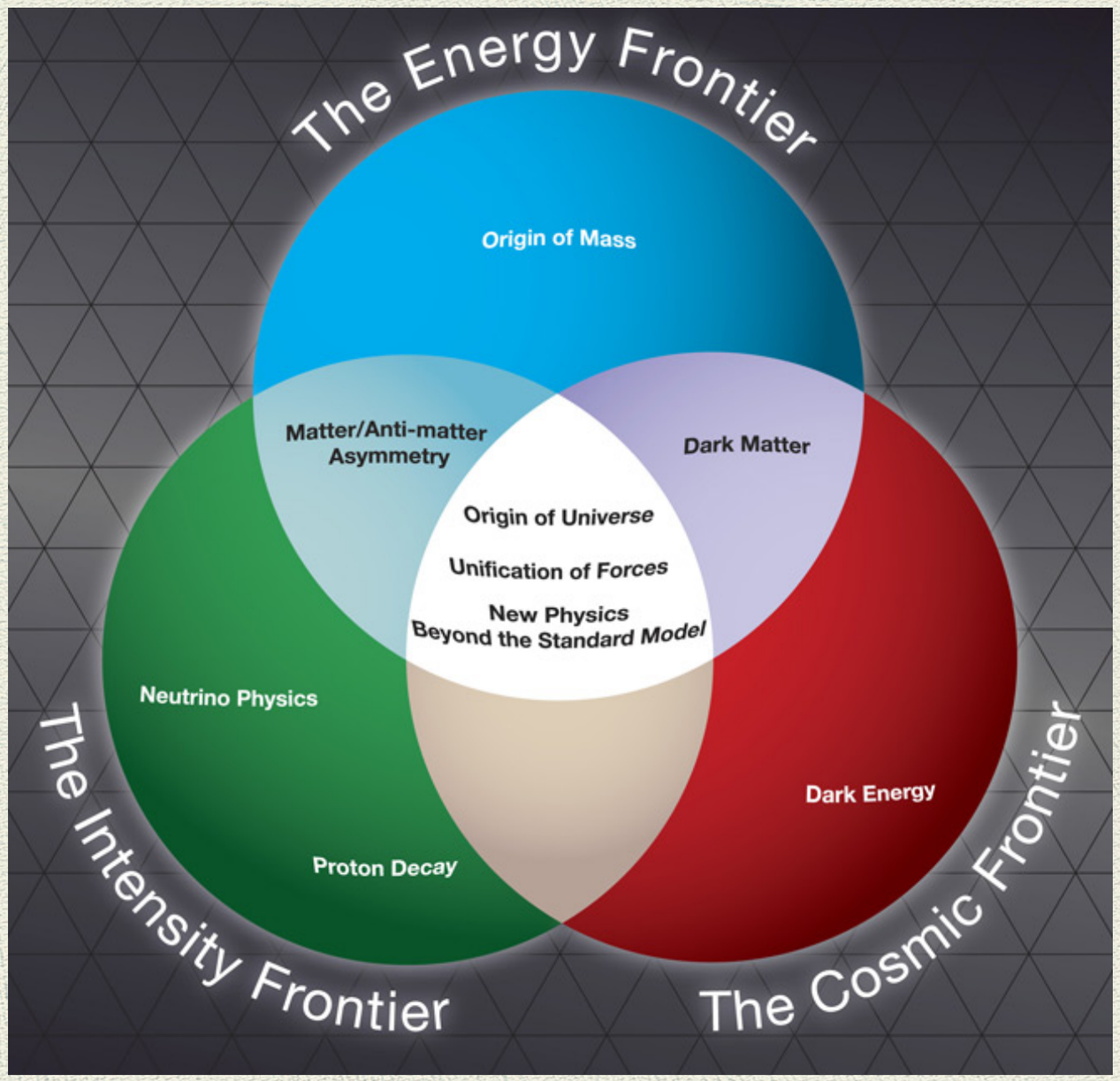
&

Computational Frontier

- **Software and computing essentially present in all fronts**
- Computing/methodological innovation
  => Full advantage of pristine data by state-of-the-art instruments
- **[2/Top 10] most-cited papers of all time in particle physics — are software programs :**
  GEANT [Detector Simulation Toolkit]
  & PYTHIA [Generation of HEP collision events]
- Looking for new physics Beyond the Standard Model (BSM)

◉ **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics

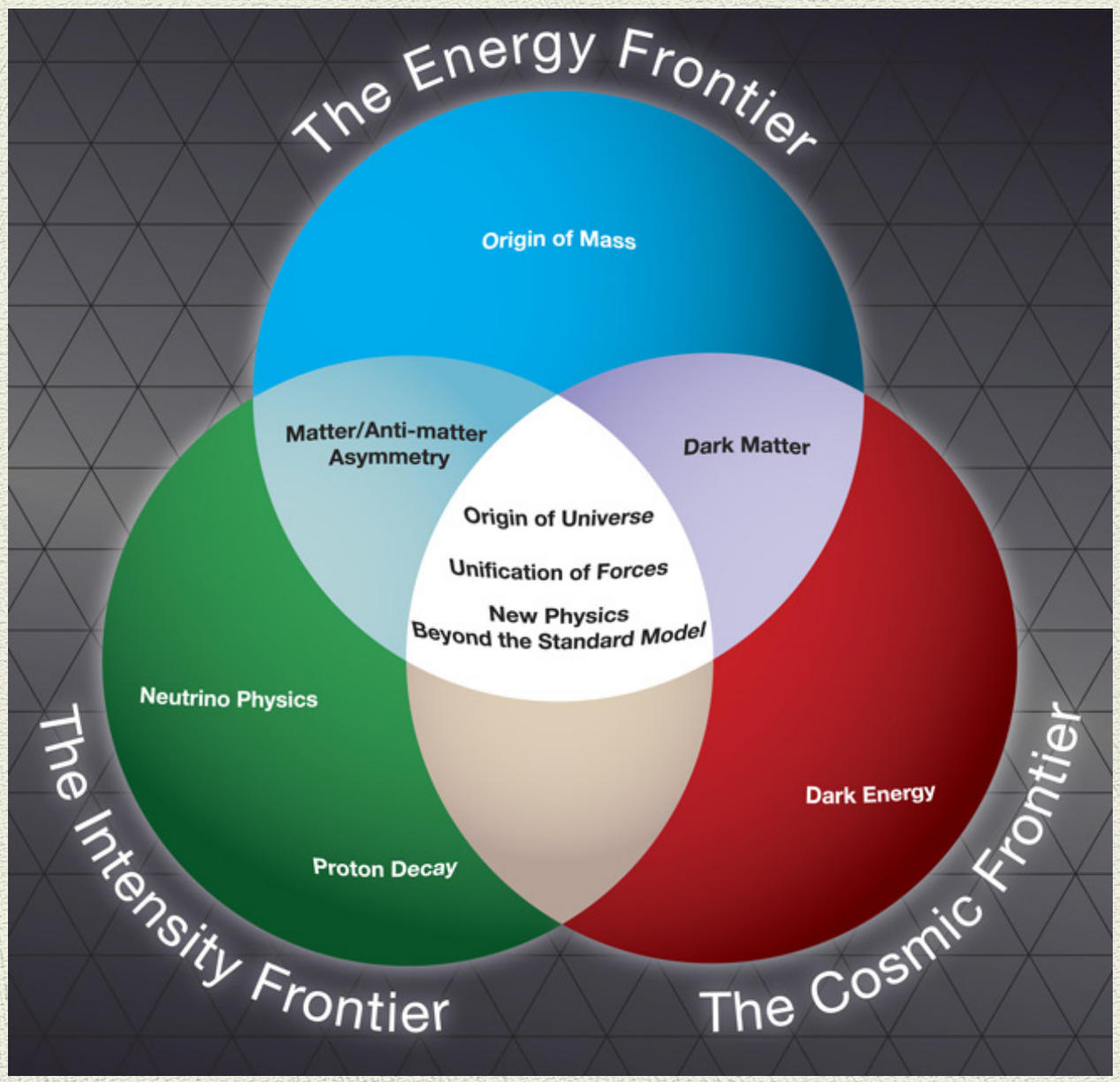


- **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics

&

Computational Frontier

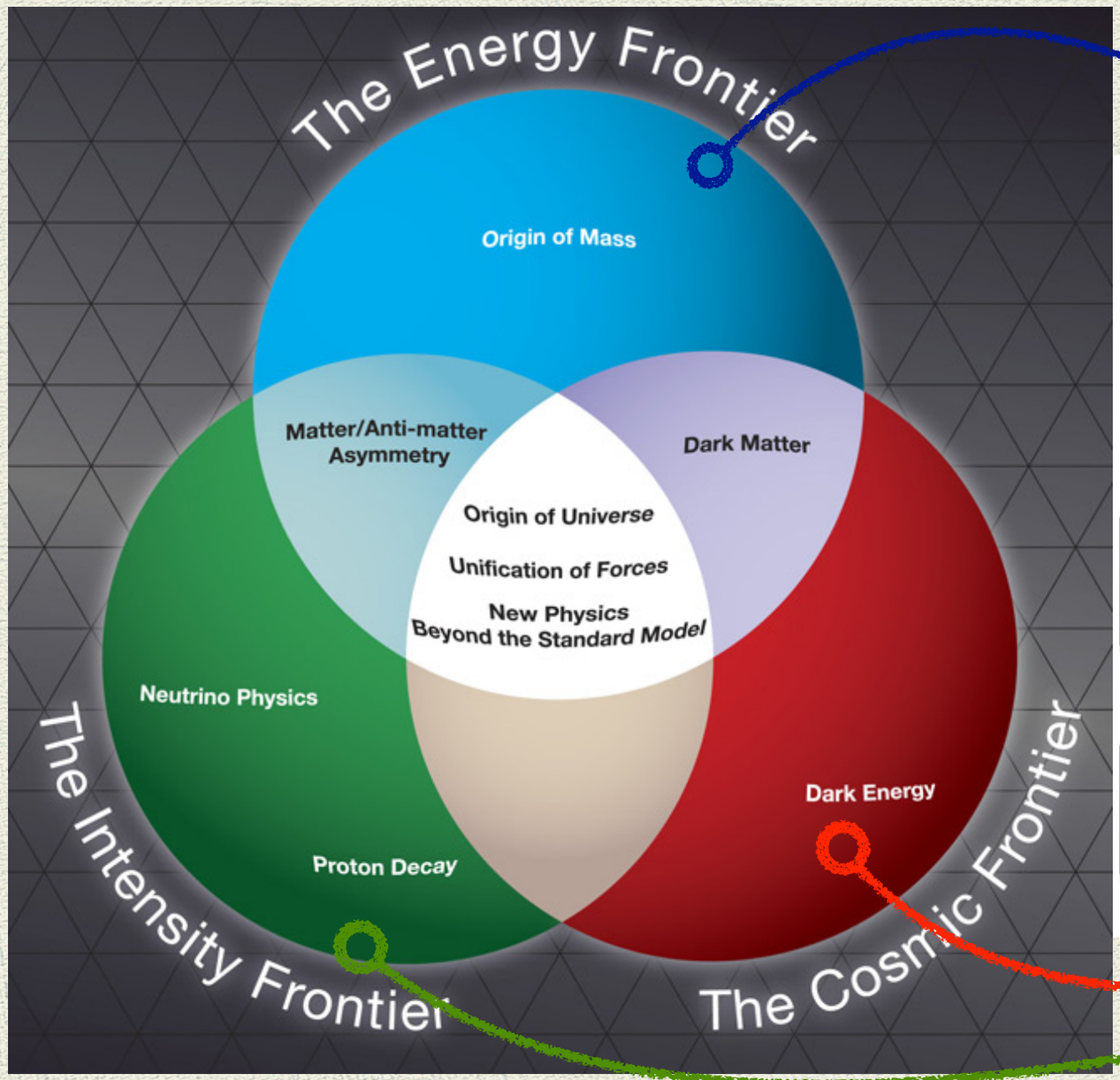


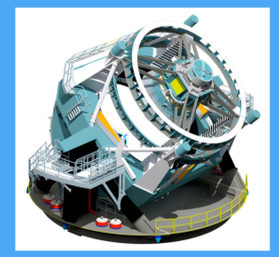
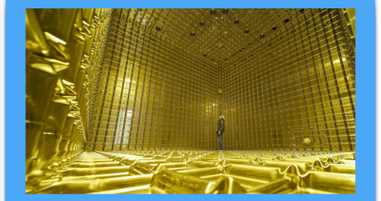
- **2210.05822 : The Future of High Energy Physics Software and Computing**

Particle Physics

&

Computational Frontier



	High-Luminosity Large Hadron Collider	Vera Rubin	DUNE
			
	2020s - 2030s	Vera Rubin	2020s - 2030s
Edge	~MB / event (~200 collisions) 4 x 10^7 event / sec. (40 MHz) ~ 40 TB / sec. ~ 10^9 TB / year = **1 Zettabyte / yr**	1000 x 3.2 Gigapixel images / night 15 TB / night 20k alerts per minute	~ 6 GB / module 5.4 ms readout window ~ 10^8 TB / year = **100 EB / yr** SN trigger ~ 0.5 PB real time processing (1 trigger / month)
Offline	~99.98% of data are removed in real time (~10 kHz/40 MHz) ~10 years of running = **Exabytes of data** (per experiment)	~10 years of running **500 PB image database** 15 PB object database, 40 Billion sources	zero suppression + compression + trigger ~10 years of running **~ 400 PB of data**

- **2210.05822 : The Future of High Energy Physics Software and Computing**

MACHINE LEARNING

FOR HEP COMMUNITY

- Machine learning is not new for HEP community

- Used in low to high level experimental measurements with **track finding**, **calorimeter hit reconstruction**, **particle identification**, **energy/momenta reco**

- Multi Variate Analysis (**MVA**) & Boosted Decision Tree (**BDT**) used extensively on high level variables with primary focus as Classifier
  - **Significant contribution in Higgs discovery**

- **The emergence of modern deep learning era** greatly outperformed the previous state of arts in last one decade or so

- Driving forces -
  - **Advent of graphics processor (GPU) + Increased computing power**
  - **Large available data + Development of advanced ML architectures**

MACHINE LEARNING

AND .. GOING DEEPER

- Classification: Find faint signal against a large background
- Move into higher dimensional space —
  - ☑ Multivariate analysis with High Level Variables
  - ☑ Low Level Variables from detectors (number of dimensions very large)
- Find the Division Boundary in this higher dimensional space
  — Best possible [under-fitting?] but Trustworthy [over-fitting?] way
- Neural Networks based on interconnected nodes in layered structure
  — In analogy with brain neurones
  — Connects different input/ derived data
  — Involve free parameters (weight and bias) [inductive bias?]
  — Optimise “free parameters” using labeled data [Model]

MACHINE LEARNING

AND .. GOING DEEPER

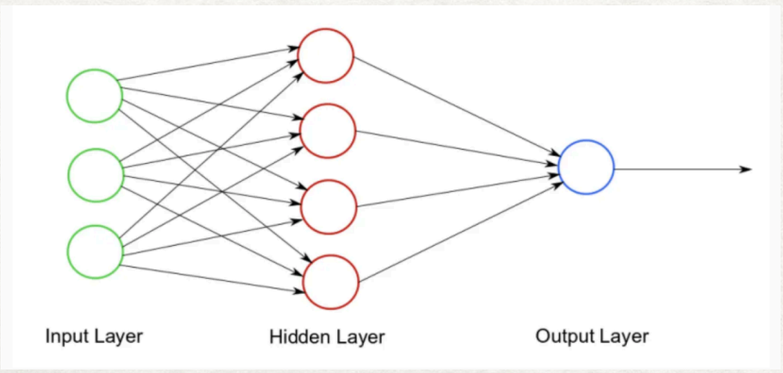
- Universal function approximation: NN with a single hidden layer can approximate any continuous function to any desired precision!

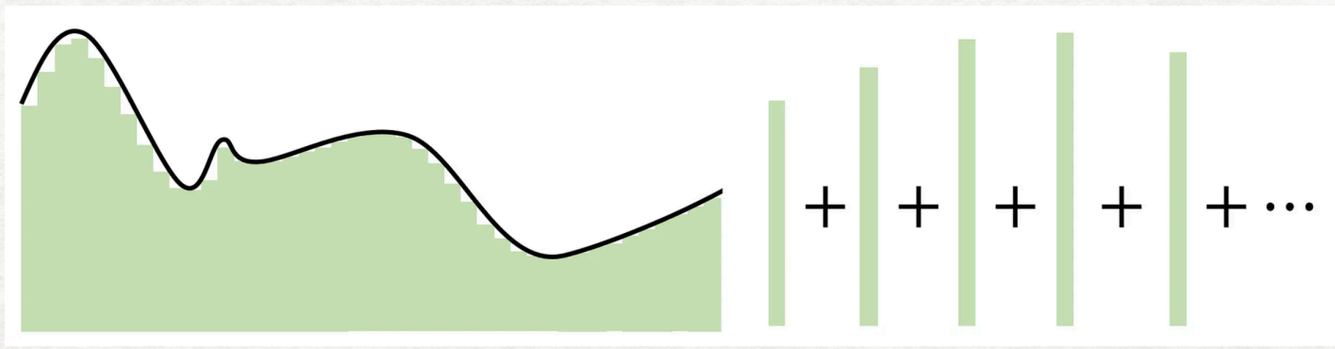
- Deep learning models with **multiple hidden layers** solves the need for infinitely large no of nodes in shallow NN

- Learning **scalable** with data - larger data for better performance

- Deep learning models are now capable of **extracting feature directly from low level data**

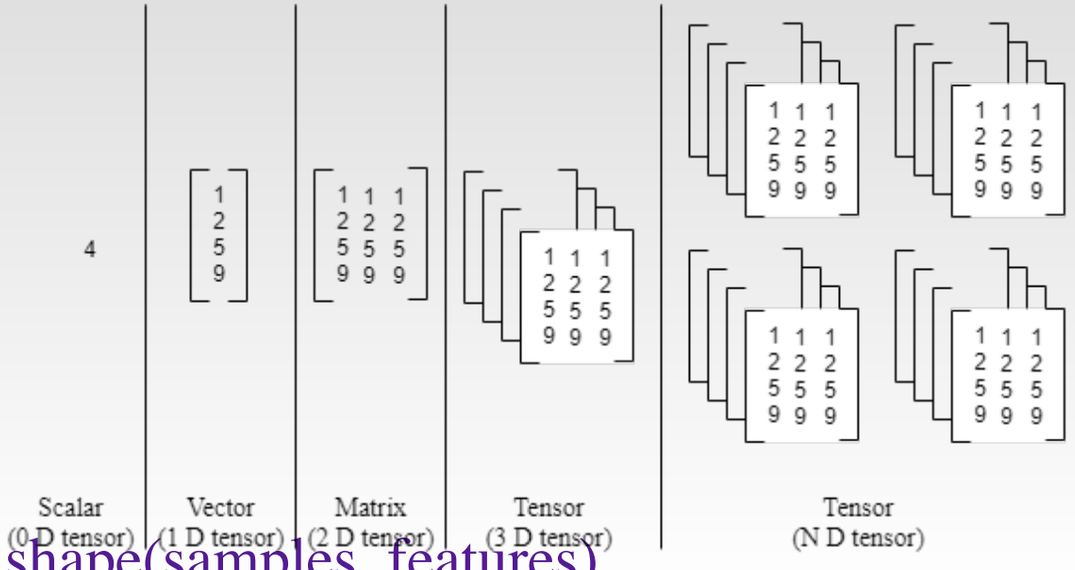
  — End for physics intuitive high level variables from domain experts?





Data Representation

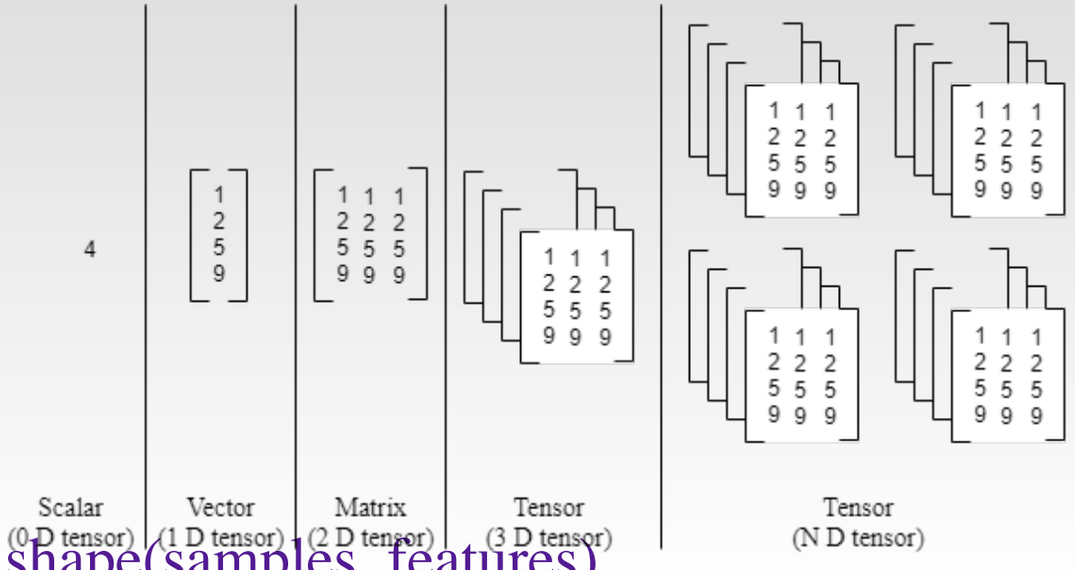
✓ **Fundamental data structure:** Fixed-length Tensors (multi-dimensional arrays)



- **Vector data:** 2D tensor of shape(samples, features)
- **Time series data or sequence data:** 3D tensor of shape(samples, timesteps, features)
- **Images**: 4D tensor of shape(samples, channels, height, width)
  For a grayscale image: channel =1; For RGB image: channels = 3
- **Video**: It is a 5D tensor of shape(samples, frames, channel, height, width)

Data Representation

✓ **Fundamental data structure:** Fixed-length Tensors (multi-dimensional arrays)



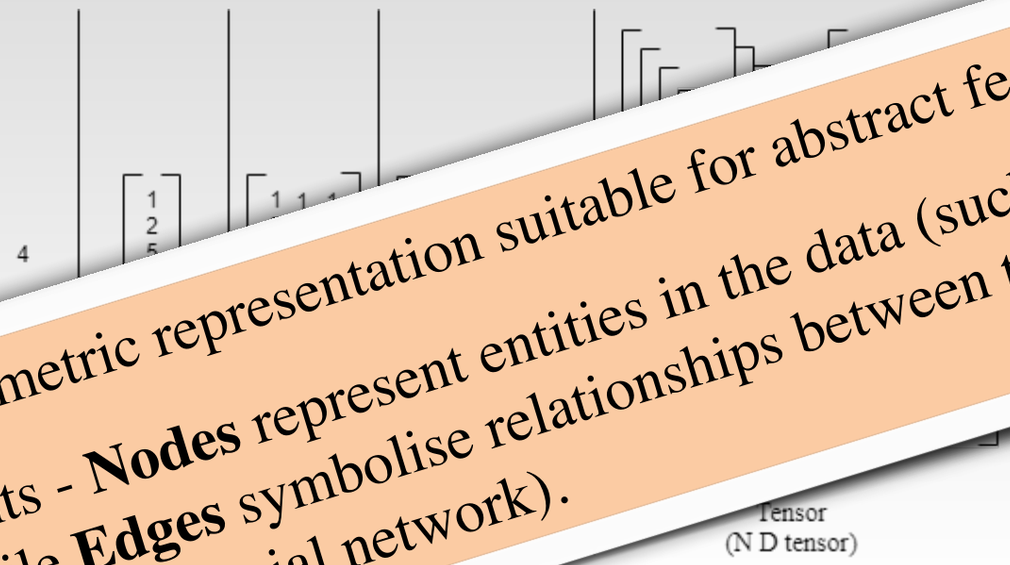
- **Vector data:** 2D tensor of shape(samples, features)
- **Time series data or sequence data:** 3D tensor of shape(samples, timesteps, features)
- **Images**: 4D tensor of shape(samples, channels, height, width)
  For a grayscale image: channel =1; For RGB image: channels = 3
- **Video**: It is a 5D tensor of shape(samples, frames, channel, height, width)

✓ **Note : Euclidean spaces are isomorphic to** $x \in \mathbb{R}^n$; $\vec{x} = \{x_1, x_2, \ldots x_i \ldots x_n\}$
in a n-dimensional linear space

✓ **However some data does not map neatly into** $\mathbb{R}^n$ **=>** Graph Neural Networks seek to adapt existing ML to directly process non-Euclidean structured data as input

Data Representation

✓ **Fundamental data structure:** Fixed-length Tensors (multi-dimensional arrays)



✓ **Point clouds:** A flexible geometric representation suitable for abstract features

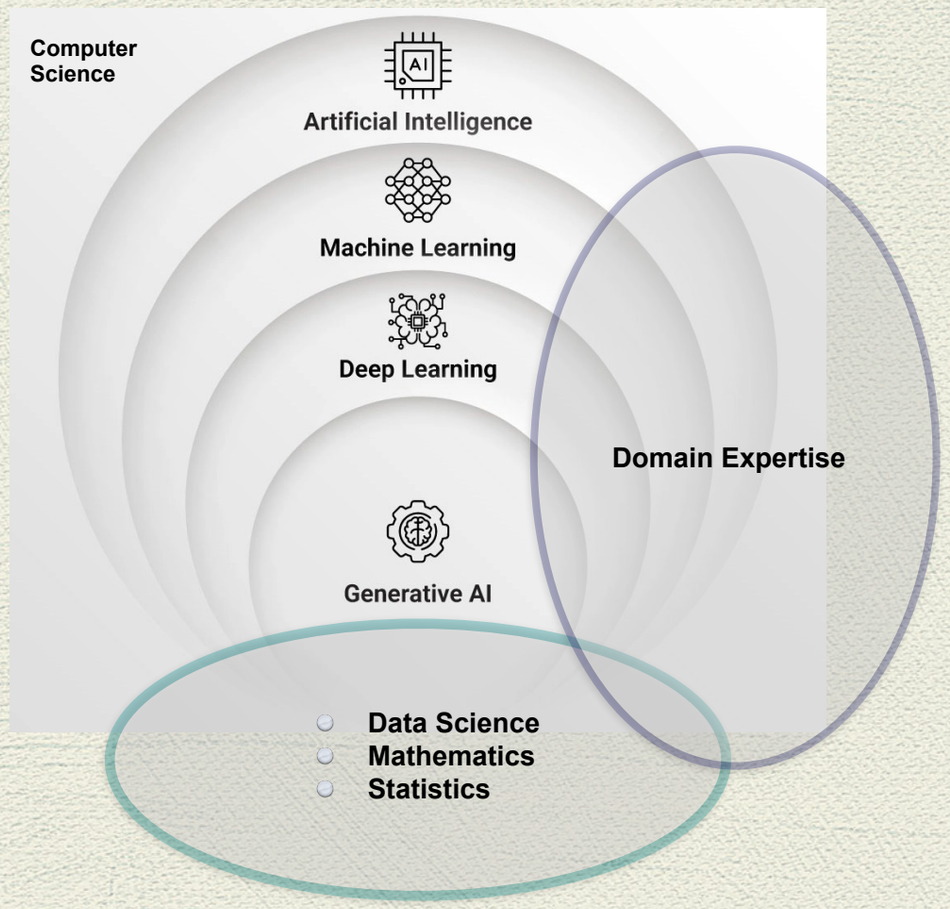
✓ **Graphs:** Two basic elements - **Nodes** represent entities in the data (such as members of an online social network), while **Edges** symbolise relationships between those entities, (such as friendship between members of a social network).

- **Sequences**: 3D tensor of shape(samples, timesteps, features)
- **Images**: 4D tensor of shape(samples, channels, height, width)
  For gray scale image: channel =1; For RGB image: channels = 3
- **Video**: It is a 5D tensor of shape(samples, frames, channel, height, width)

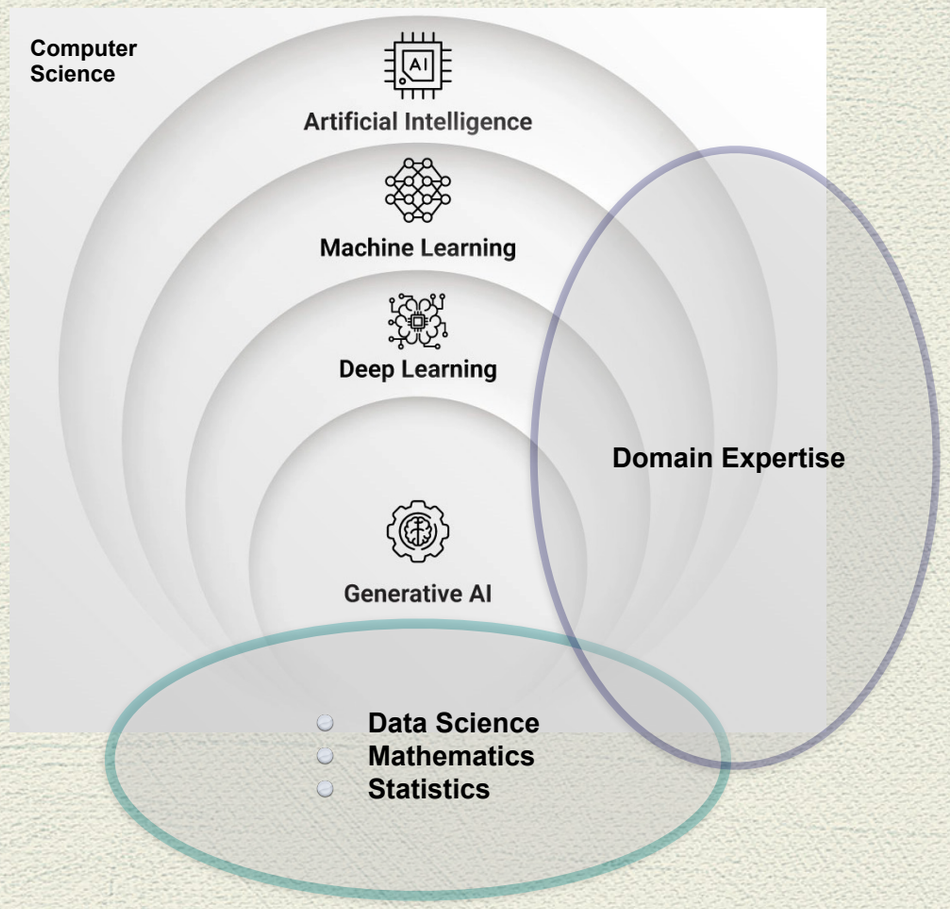
✓ **Note : Euclidean spaces are isomorphic to** $x \in \mathbb{R}^n$; $\vec{x} = \{x_1, x_2, \ldots x_i \ldots x_n\}$ **in a n-dimensional linear space**

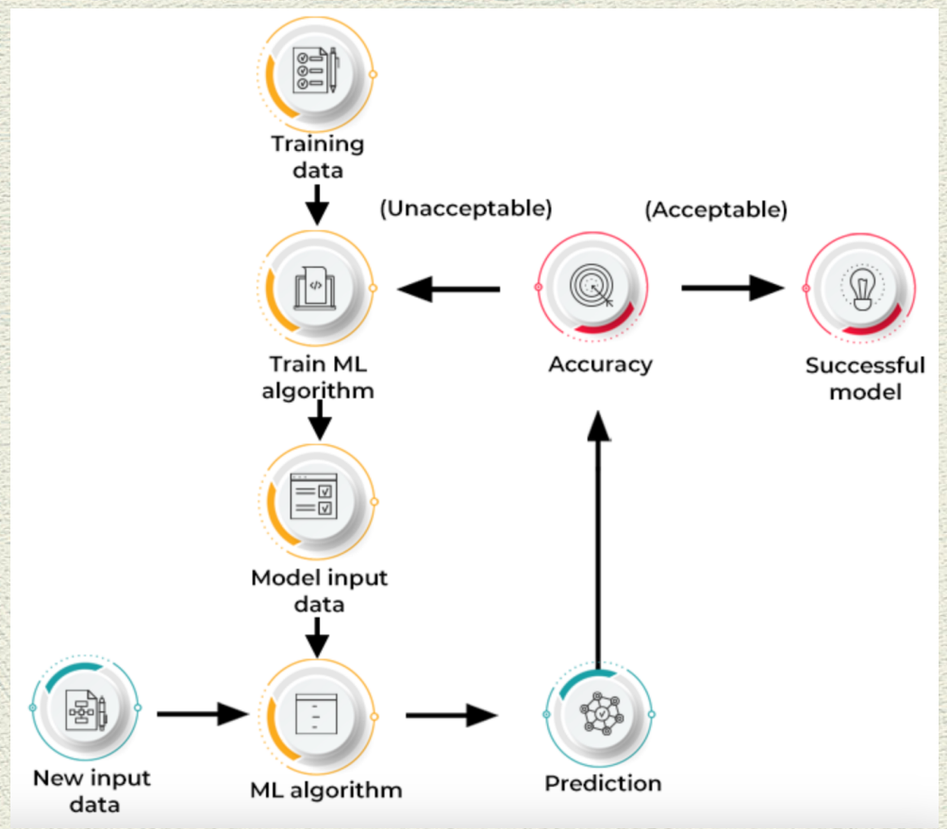
✓ **However some data does not map neatly into** $\mathbb{R}^n$ **=>** Graph Neural Networks seek to adapt existing ML to directly process non-Euclidean structured data as input

How ML works?

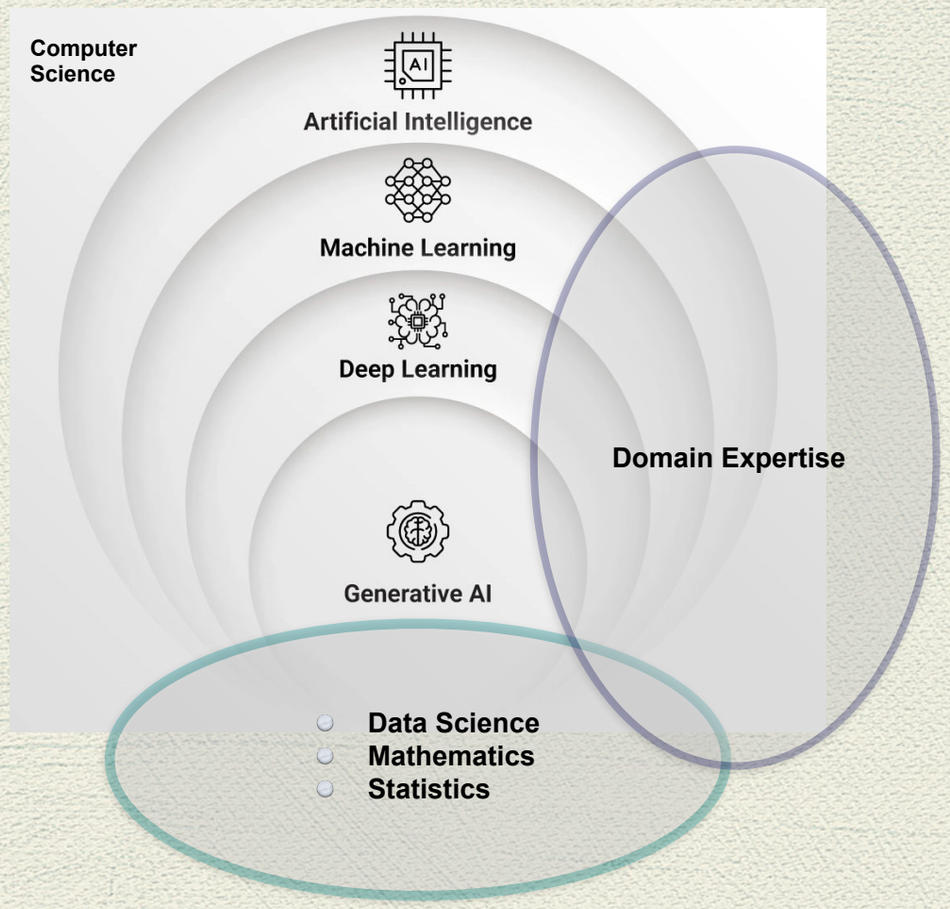

Computer Science
Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Domain Expertise
Data Science
Mathematics
Statistics

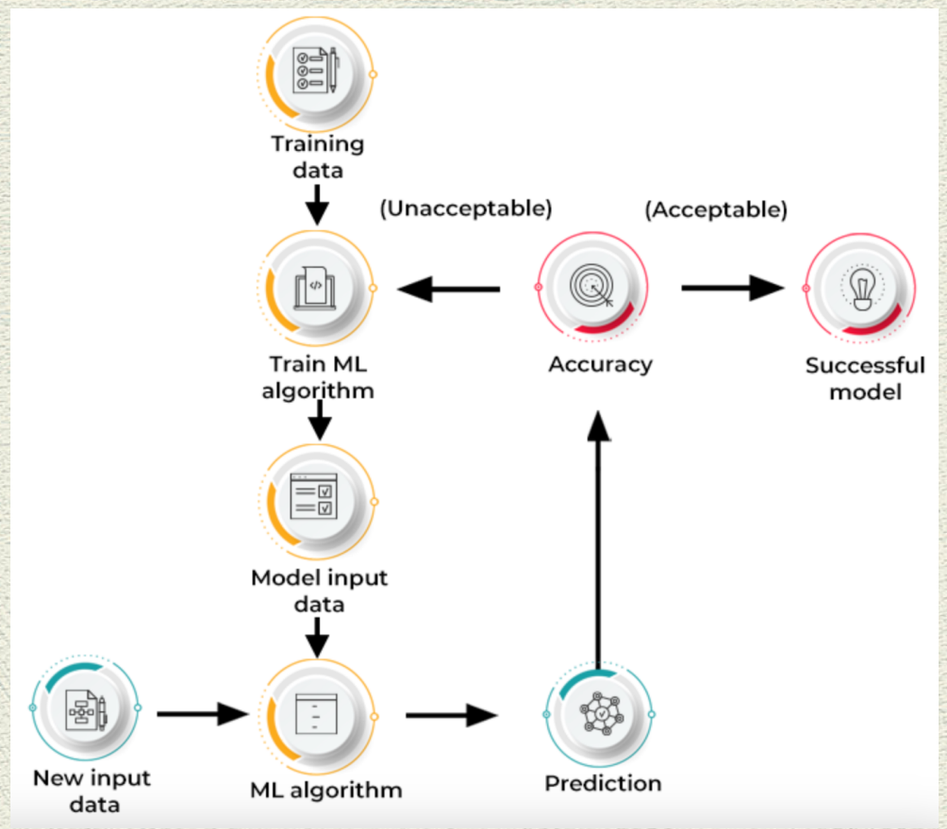
How ML works?


Computer Science
Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Domain Expertise
Data Science
Mathematics
Statistics


Training data
(Unacceptable)
(Acceptable)
Train ML algorithm
Accuracy
Successful model
Model input data
New input data
ML algorithm
Prediction

How ML works?





- **Decision boundary:** surface that divides multi-dim feature space into distinct groups of data points.
- **Training:** ML algorithm discovers the decision boundary
- **Testing**: Then uses to forecast the class of unseen data points.
- Key drivers for its growth

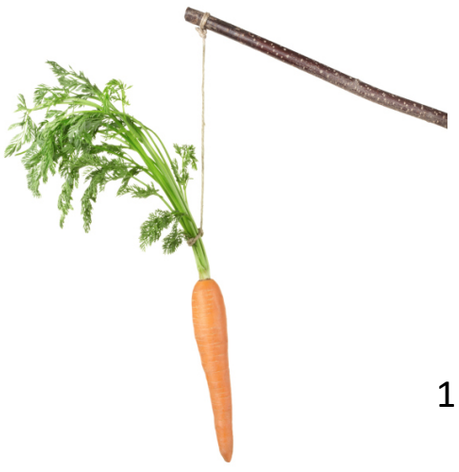
  — **Data, Algorithms & Hardware (graphics processing unit or GPU)**

Three Ways to Learn

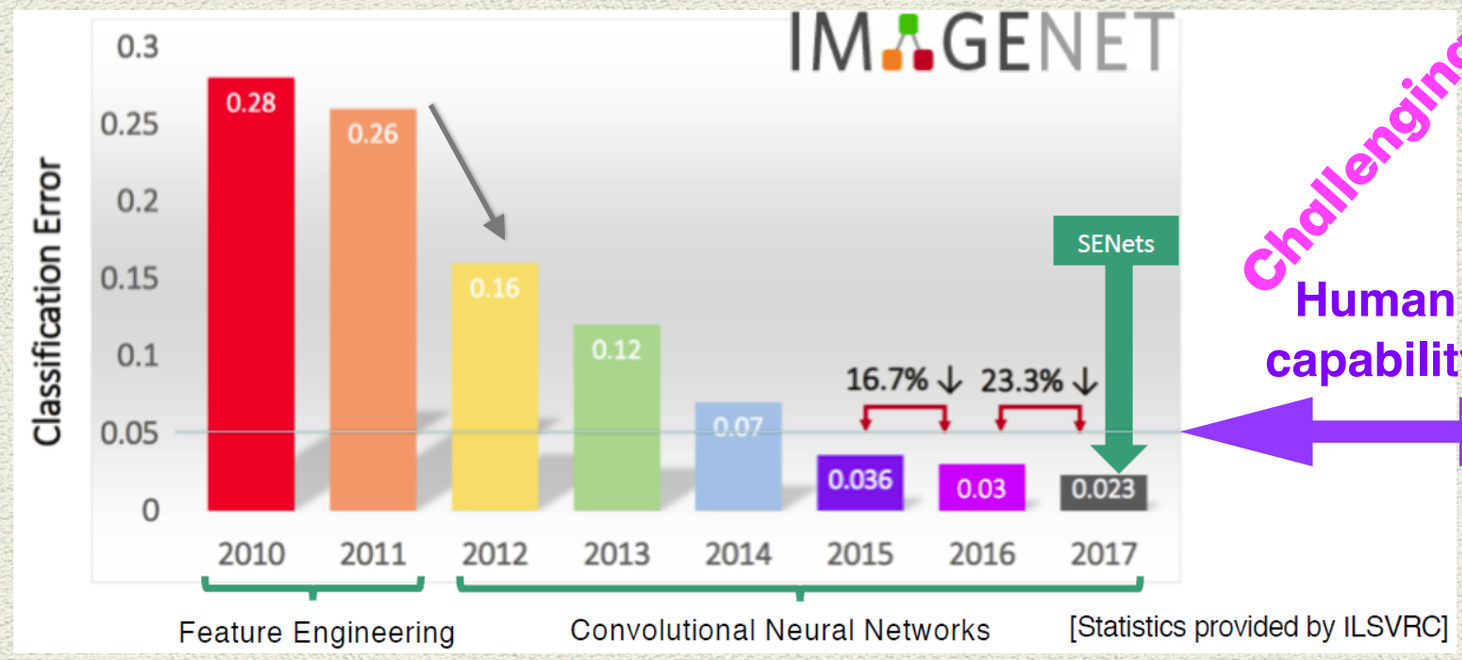
Supervised

Unsupervised

Reinforcement

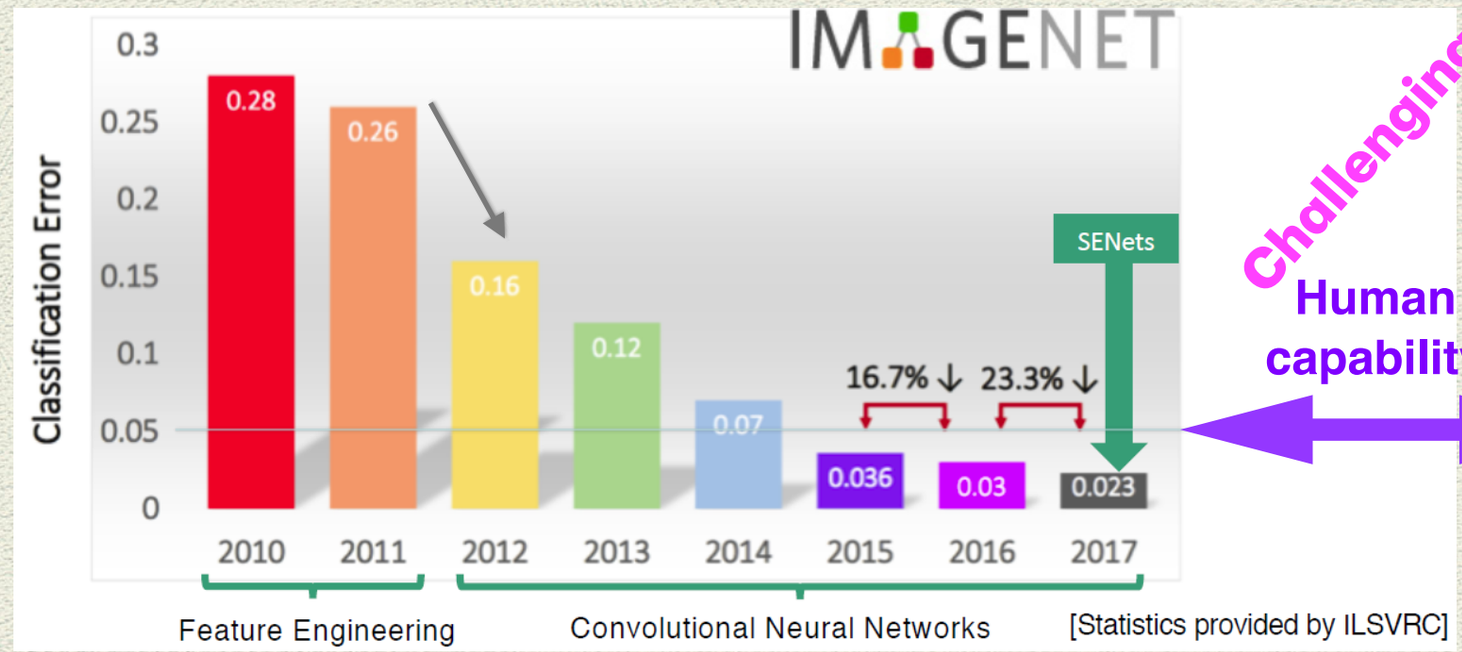


Progress of Deep Learning



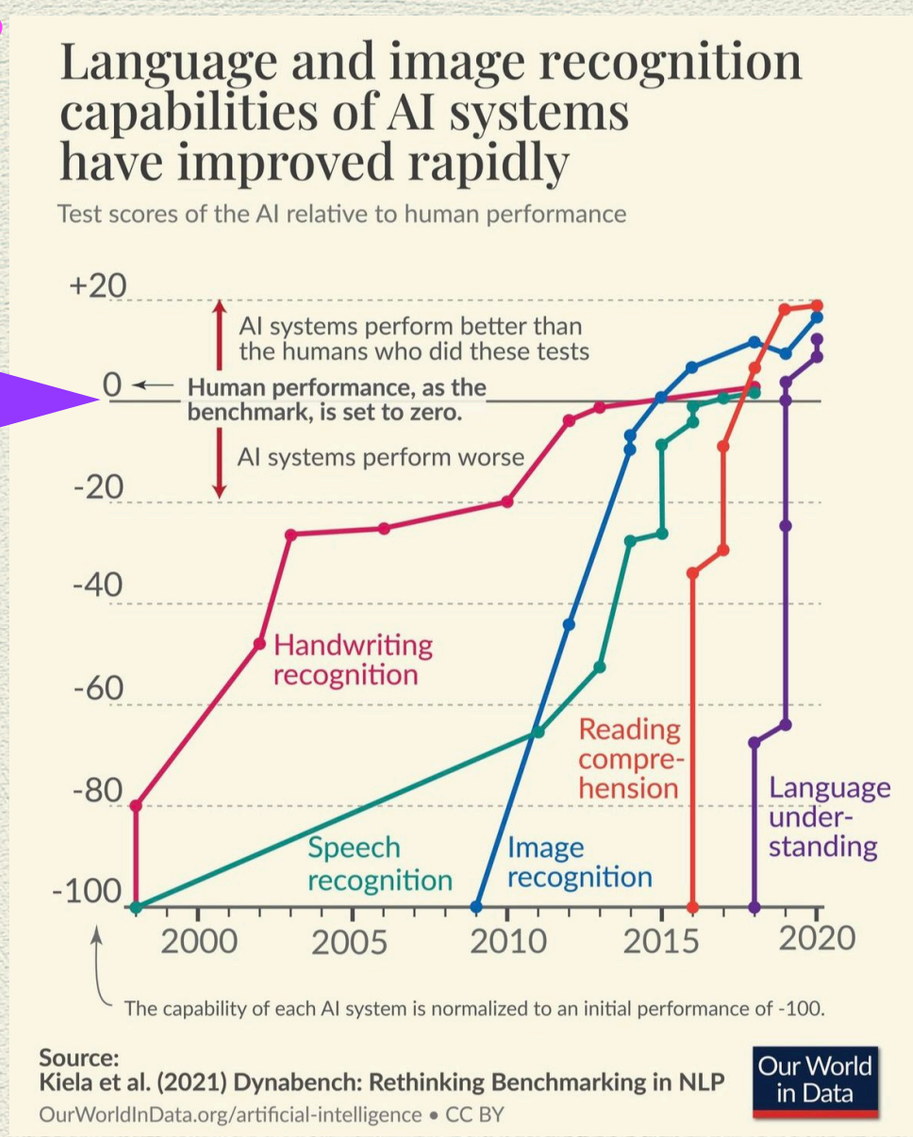
- ImageNet - Large Scale Visual Recognition Challenge (ILSVRC) held each year : largest contest in object recognition
- 2012 - AlexNet [Deep CNN by Alex Krizhevsky etal] ~ 15.4% error (2nd 26.2%!)]
- Since then, these competitions are consistently won by deep convolutional nets

Progress of Deep Learning



Challenging

Human capability



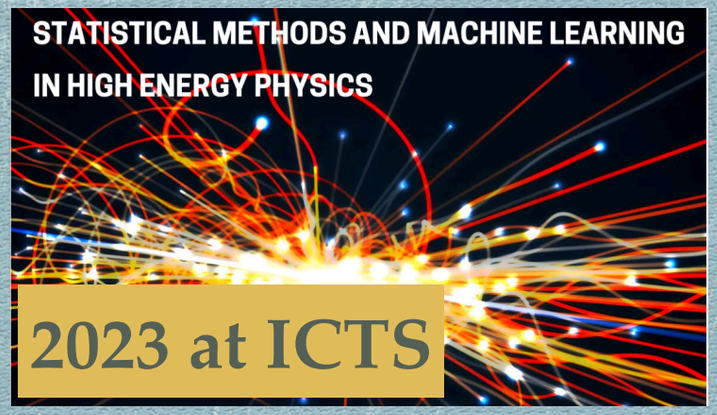
- ImageNet - Large Scale Visual Recognition Challenge (ILSVRC) held each year : largest contest in object recognition
- 2012 - AlexNet [Deep CNN by Alex Krizhevsky etal] ~ 15.4% error (2nd 26.2%!)]
- Since then, these competitions are consistently won by deep convolutional nets

- Imitate (defeat) human intelligence and capability in **visual perception, speech recognition, decision-making, language processing, and so on.**

DEEP MACHINE LEARNING

CATEGORY

Strategy	Representations	Targets / tagging	strategies
Classification	➡ Jet Image ➡ Event Image ➡ Sequence (Recurrent NN) ➡ Graph (Graph NN) ➡ Sets (Point cloud - Graph)	➡ Quarks vs gluons ➡ Boosted H / W / Z / Top tag ➡ New particles and models ➡ Particle tagging at detector ➡ Neutrino flavour	➡ Weak/ Semi/ Un-supervised ➡ Reinforcement Learning ➡ Quantum Machine Learn ➡ Feature Ranking ➡ Optimal Transport
Regression	➡ Parameter estimation ➡ Pileup mitigation ➡ Parton Distribution Func ➡ Symbolic Regression ➡ Function Approximation		
Generative models		➡ GANs ➡ Autoencoders ➡ Phase space generation ➡ Normalizing flows	
Anomaly detection			



Primarily for Students and PDFs working on using deep learning

Preparatory school (Online)
[June 12 - 23 2023]

Lecture + Tutorial
28 Aug - 04 Sep 2023

Workshop
5 - 8 Sep 2023

https://www.icts.res.in/program/ml4hep