



MARK ASCH



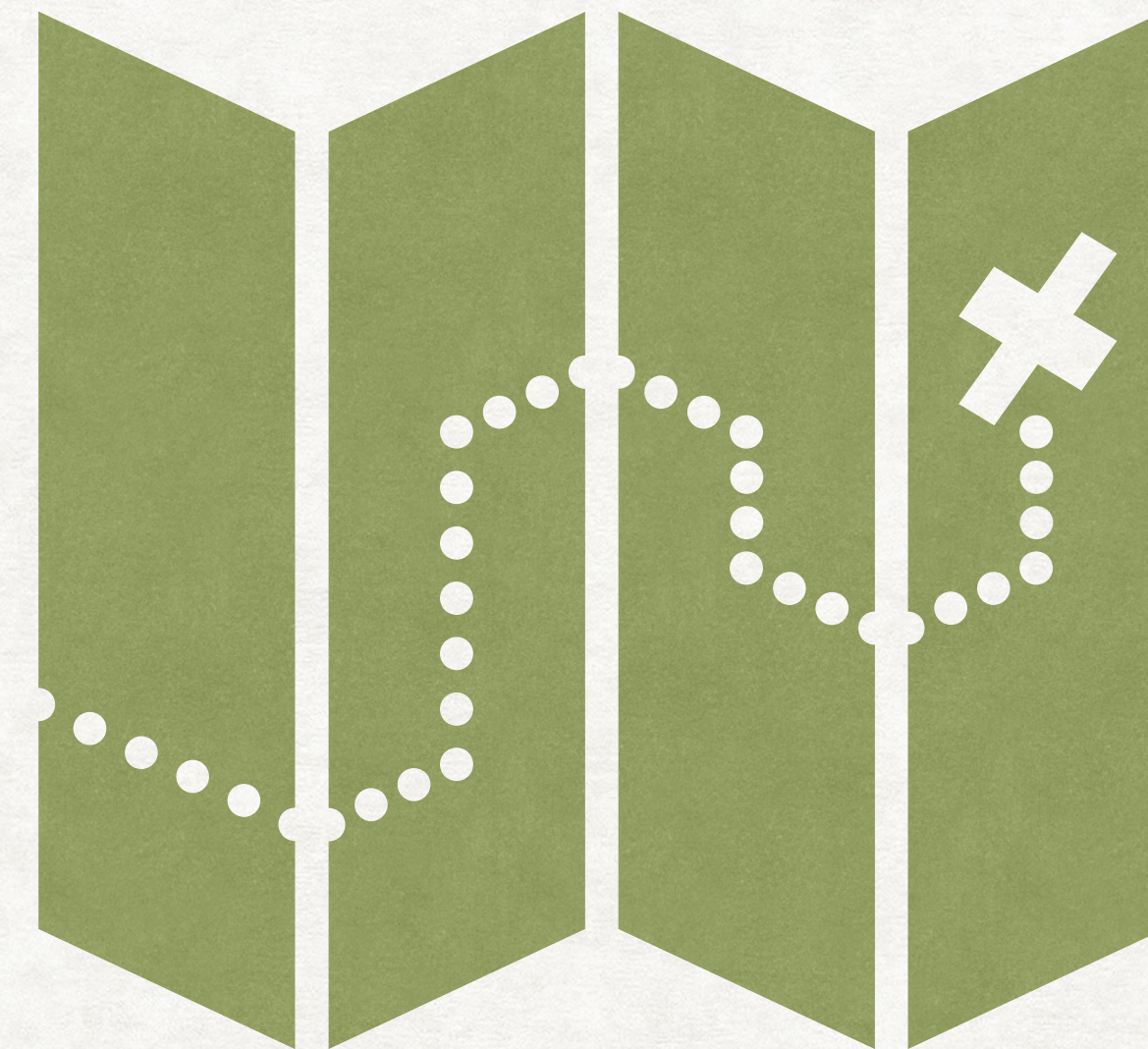
AI FOR SCIENCE SCIENCE FOR AI



2024 MSP Annual Convention, Tacloban City, Philippines - May 30 to June 2, 2024

PLAN

1. ML/AI revolution (ML for Science)
2. Scientific ML (Science for ML)
3. Ethics, bias, responsibility (certifiability)
4. LLMs for science - "quo vadimus"?



“

AI IS ONE OF THE MOST TRANSFORMATIVE AND VALUABLE
SCIENTIFIC TOOLS EVER DEVELOPED.

BY HARNESSING VAST AMOUNTS OF DATA AND
COMPUTATIONAL POWER, AI SYSTEMS CAN UNCOVER
PATTERNS, GENERATE INSIGHTS, AND MAKE PREDICTIONS
THAT WERE PREVIOUSLY UNATTAINABLE*.

— Rick Stevens, ISC 2024

”

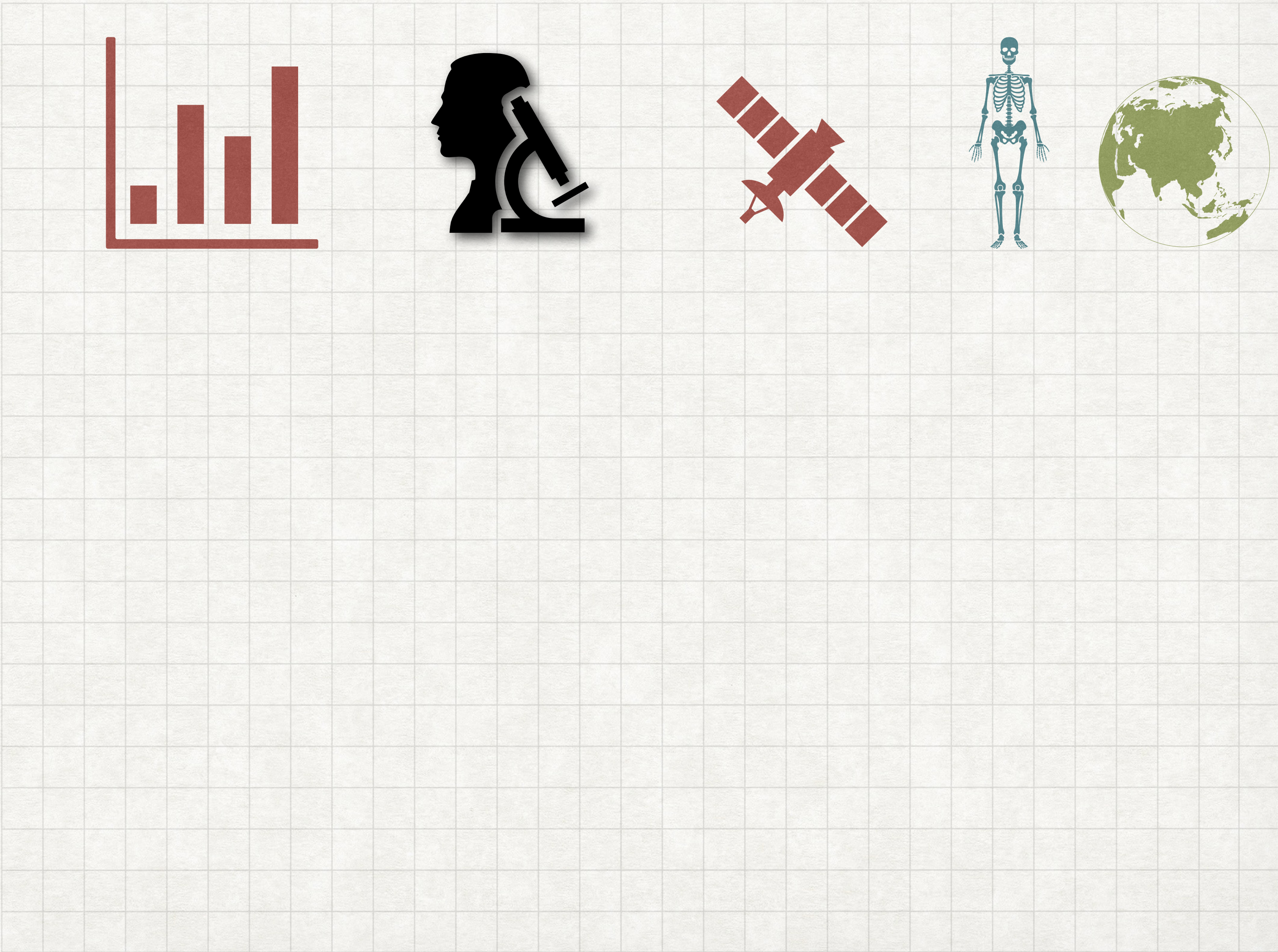
*However, these tools are absolutely **useless** if we cannot **trust** the information we receive from them.

ML/AI FOR SCIENCE

AI REVOLUTION OF 2020'S

3 MAJOR REASONS/CAUSES/FACTORS

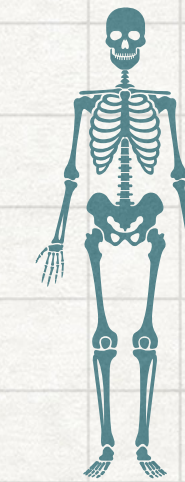
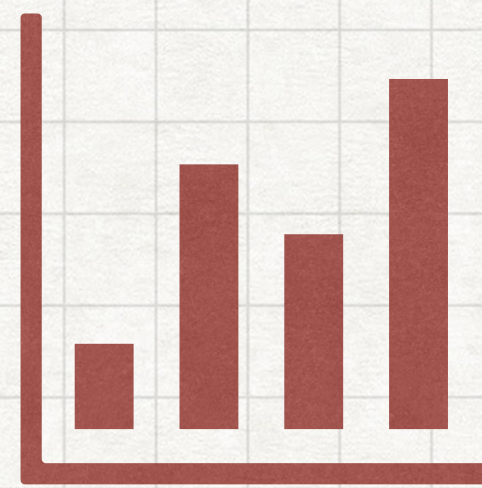
1. Big Data



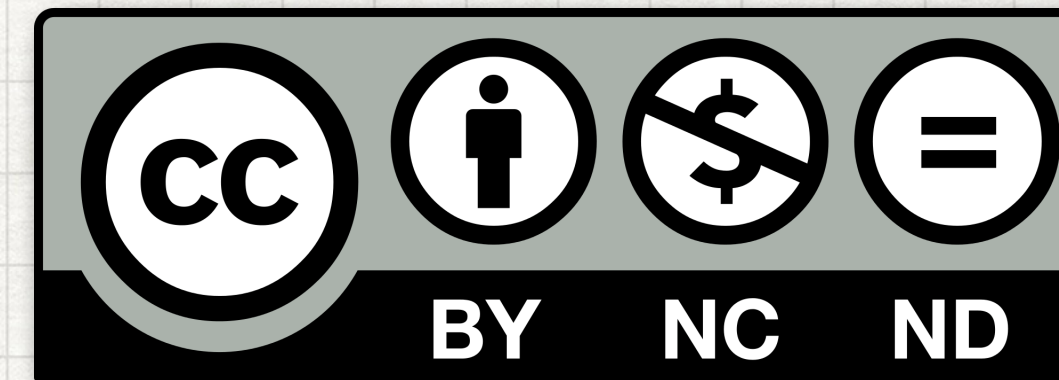
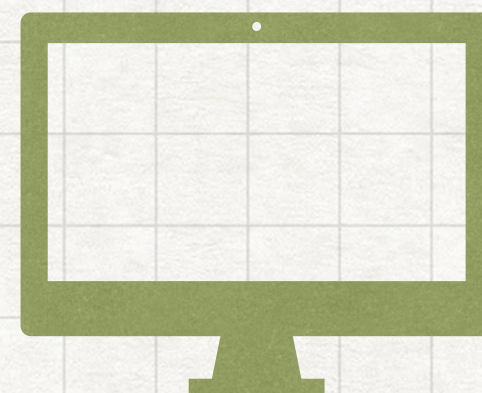
AI REVOLUTION OF 2020'S

3 MAJOR REASONS/CAUSES/FACTORS

1. Big Data



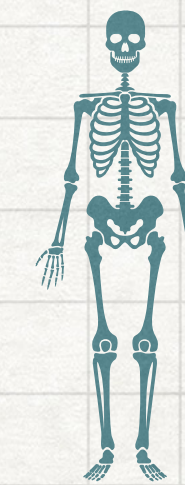
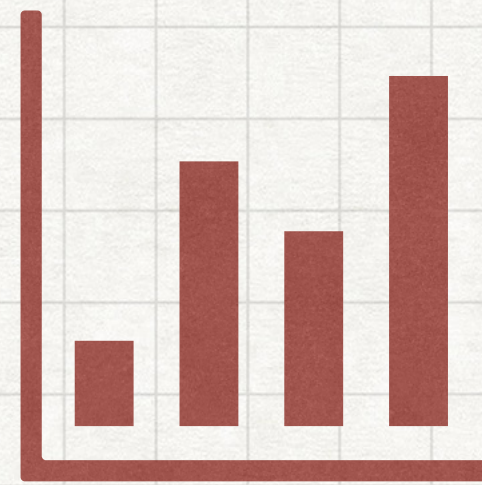
2. Open source (ML) software libraries



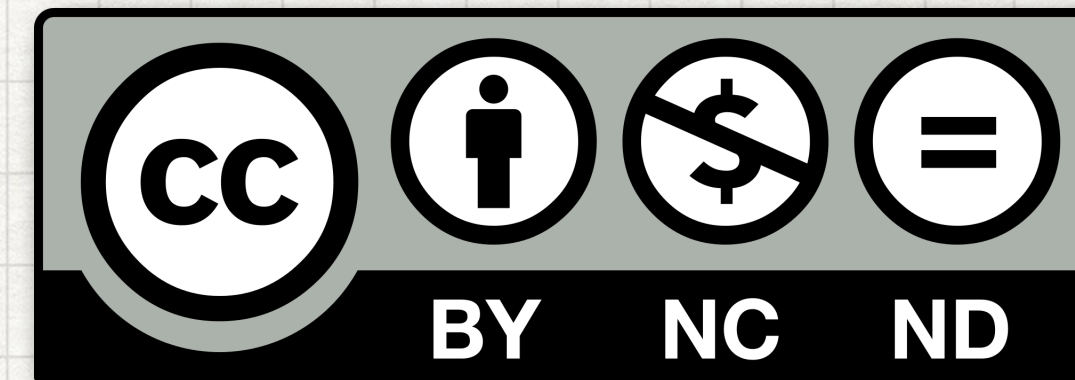
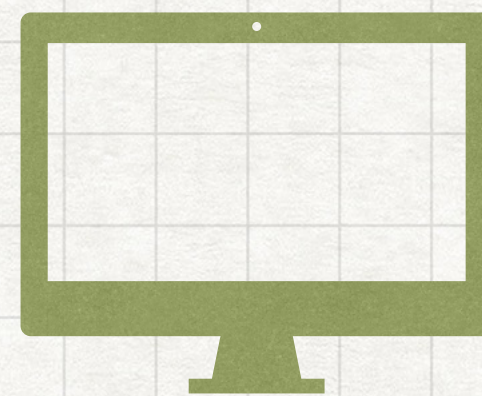
AI REVOLUTION OF 2020'S

3 MAJOR REASONS/CAUSES/FACTORS

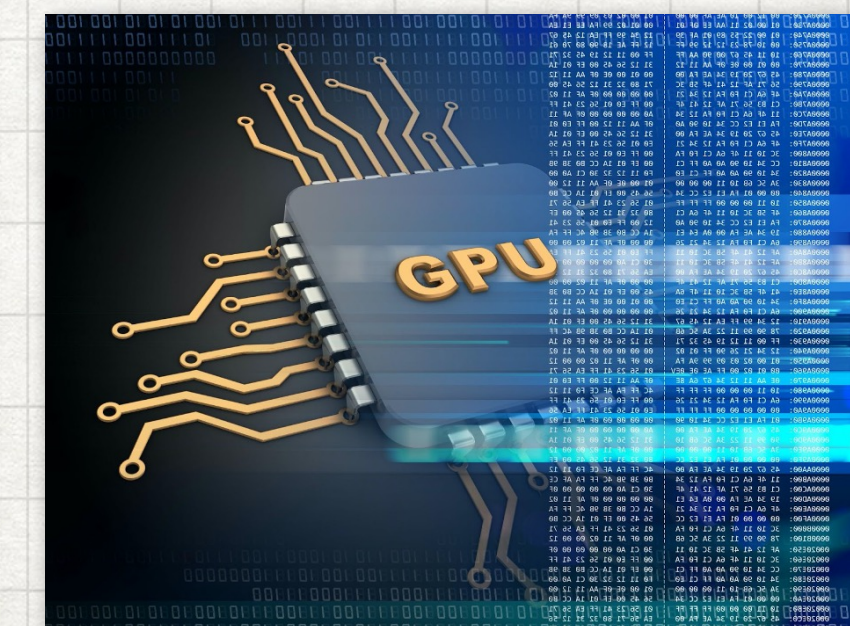
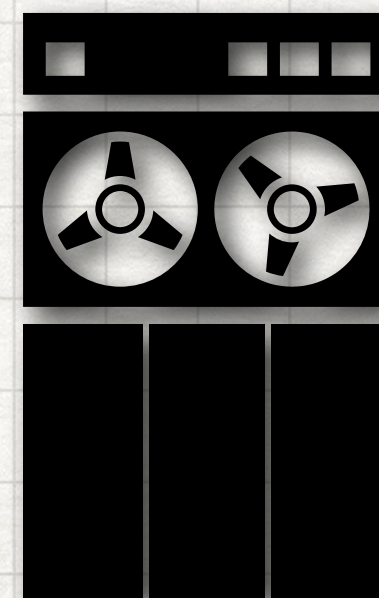
1. Big Data



2. Open source (ML) software libraries



3. Access to (cheap) computer hardware and storage

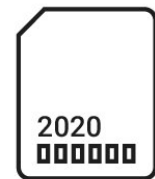


FACTOR 1: BIG DATA



THE 4 V'S OF BIG DATA

40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005



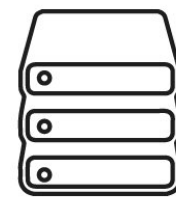
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION



Volume

SCALE OF DATA

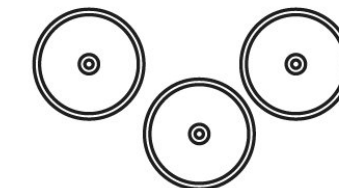
2.5 QUINTILLION BYTES
of data are created each day



Most companies in the U.S. have at least **100 TERABYTES** of data stored



As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES**



30 BILLION PIECES OF CONTENT are shared on facebook every month



Variety

DIFFERENT FORMS OF DATA

4 BILLION + HOURS OF VIDEO are watched on You Tube each month



4 MILLION TWEETS are sent per day by about 200 million monthly active users



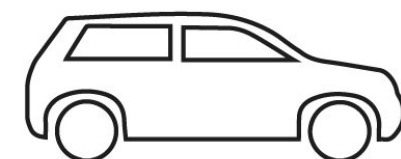
The New York Stock Exchange captures **1TB OF TRADE INFORMATION** during each trading session



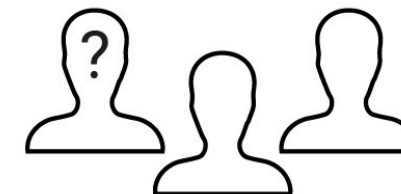
Velocity

ANALYSIS OF STREAMING DATA

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Veracity

UNCERTAINTY OF DATA

27% OF RESPONDENTS in one survey were unsure of how much of data was inaccurate

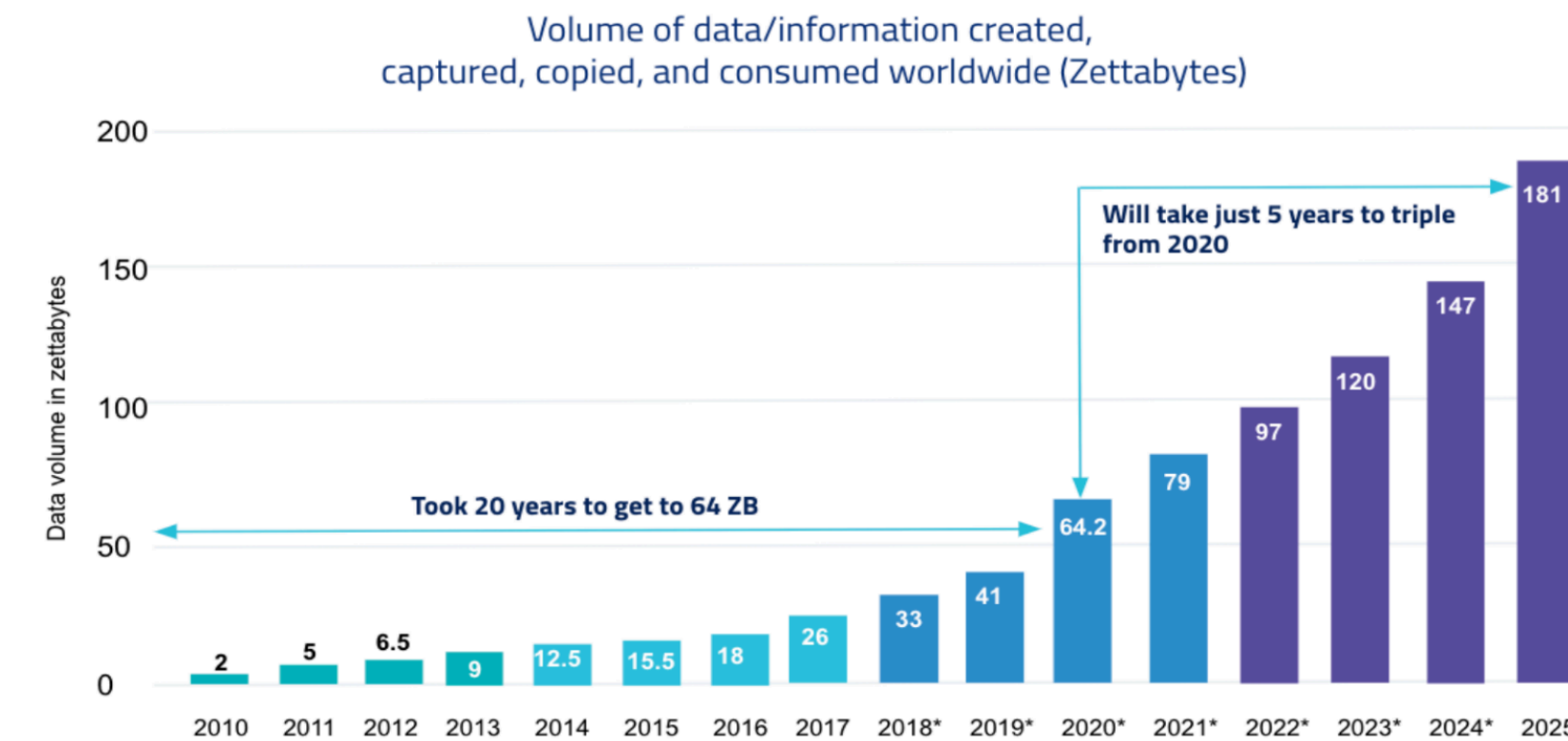


FACTOR 1: BIG DATA

EXAMPLE: HEALTH DATA IS EVERYWHERE...

1. Healthcare Data
2. Diagnostics Data
3. Omics Data
4. IoT and Wearables Data
5. Consumer Data

Big Data Grows Ever Bigger



35%
of all data will
be life sciences
+ healthcare by
2025

Increasingly digital lives are leading to an explosion in healthcare data

>2,000

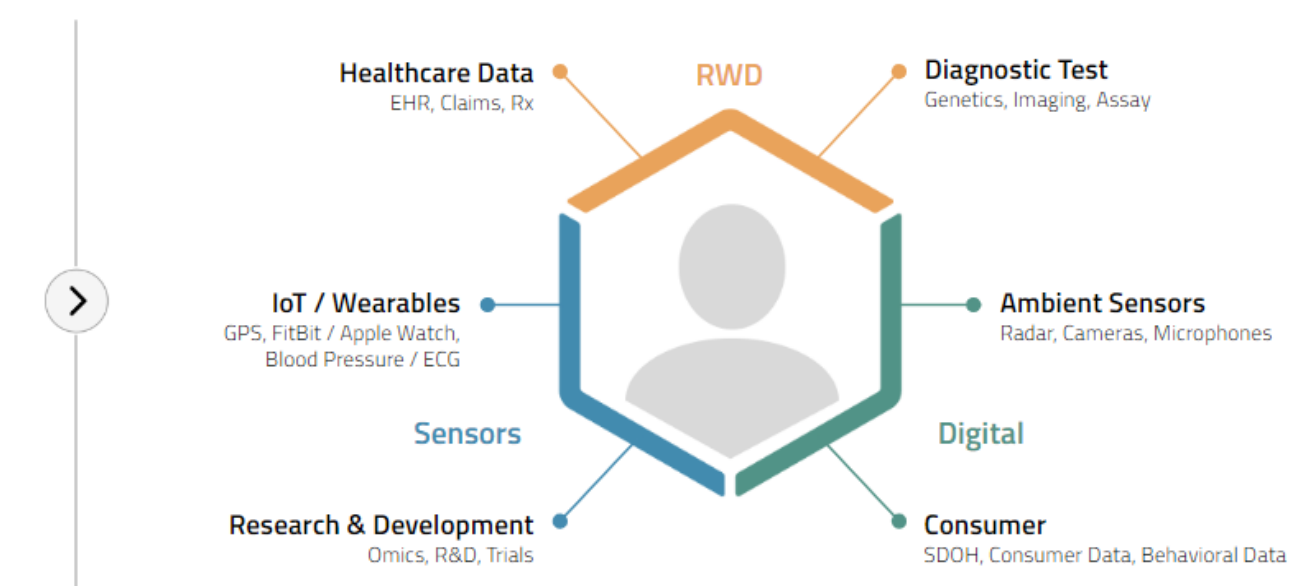
Exabytes of healthcare data generated annually

36%

CAGR growth rate

5,000+

Daily data interactions per person by 2025



FACTOR 2: OPEN SOURCE

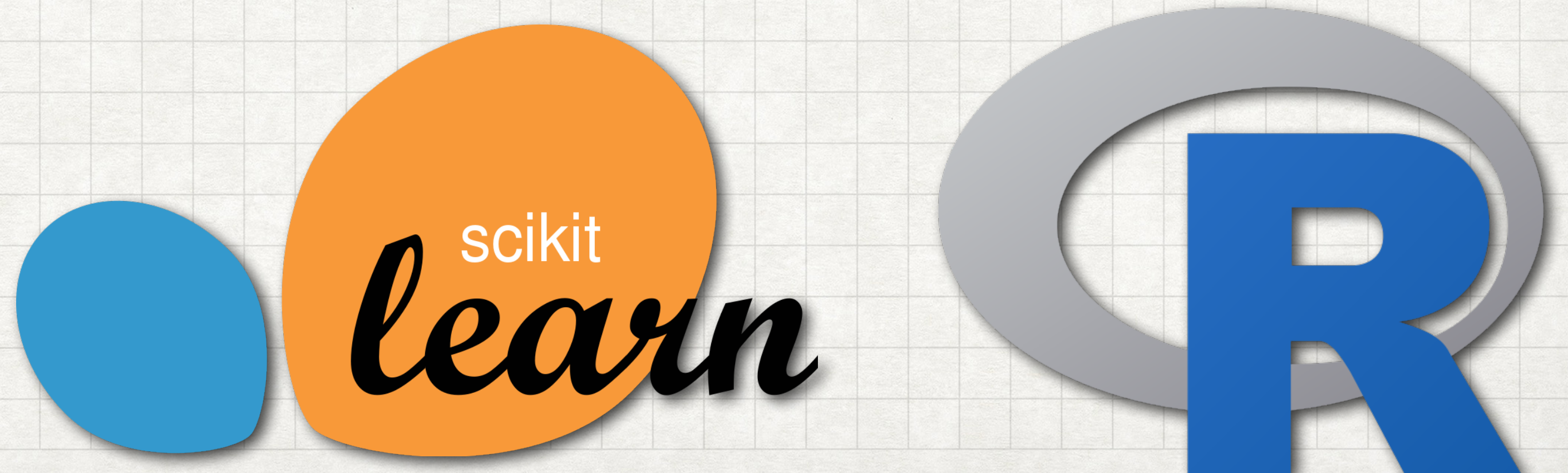
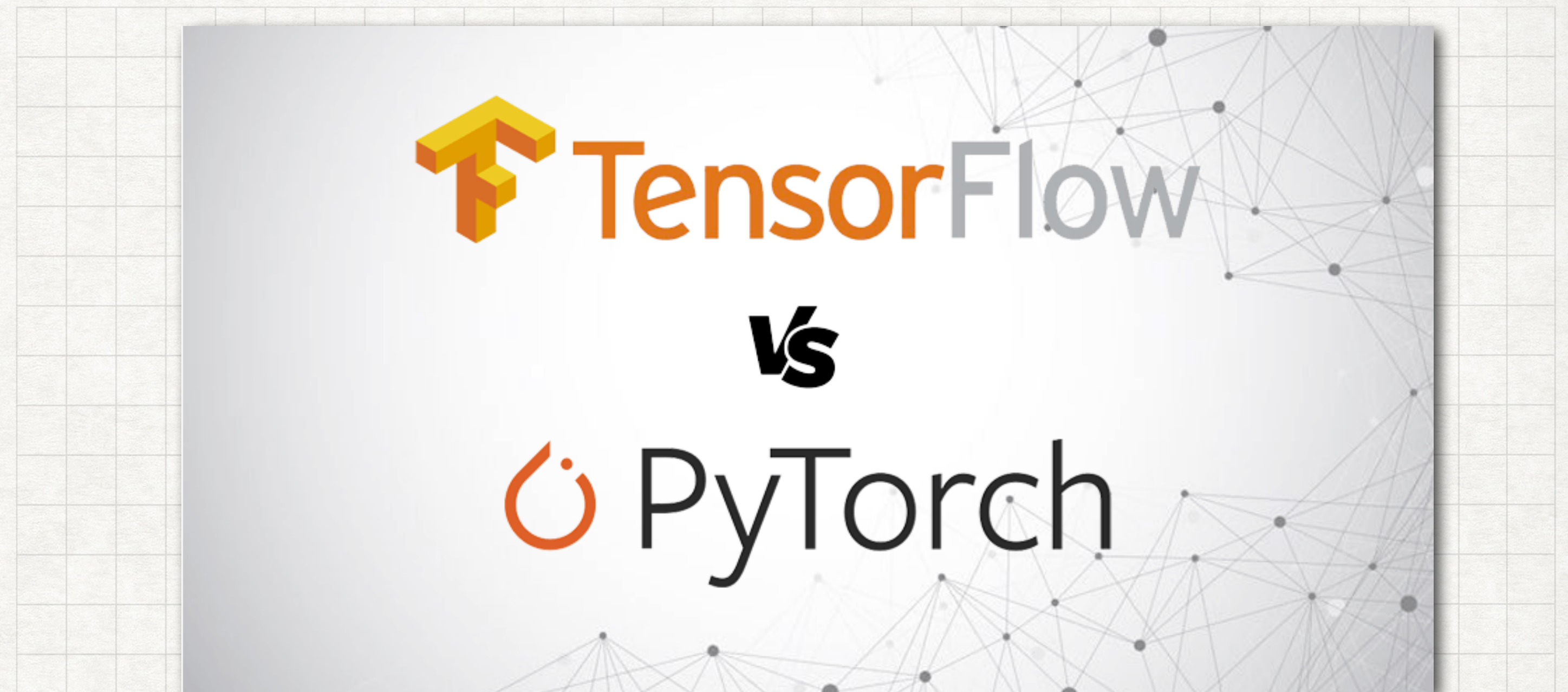
SOFTWARE IS THE KEY

1. Driven by GAFAM...

1. TensorFlow, Keras
2. PyTorch

2. Open is the default

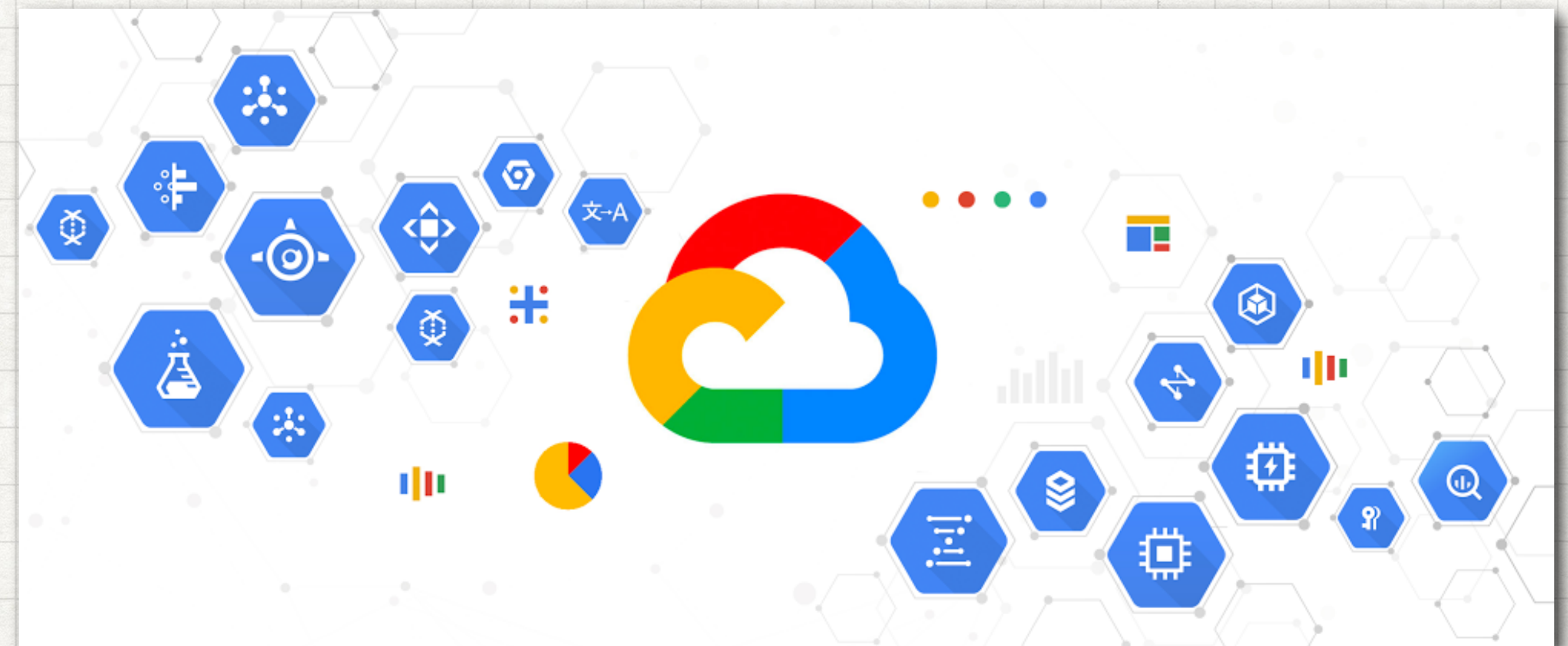
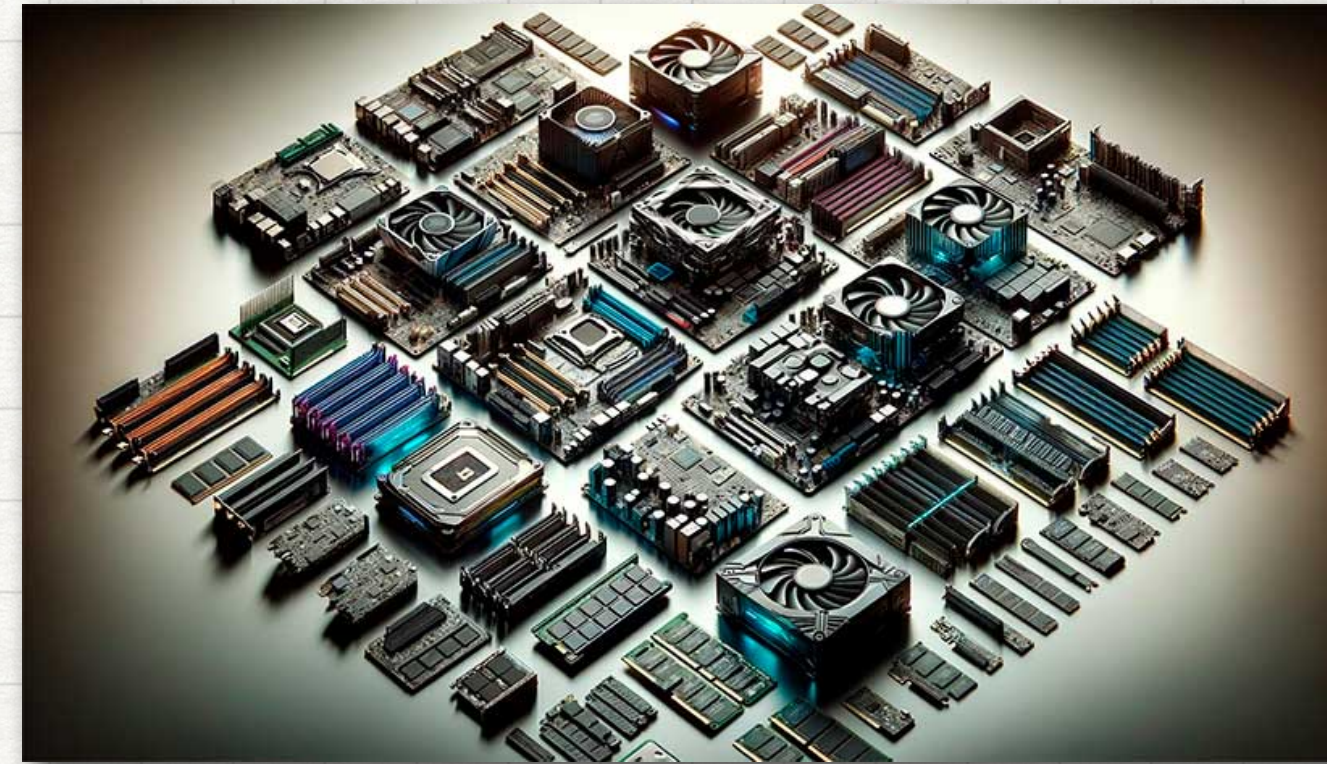
1. Scikit-Learn
2. R
3. JaX/Autograd



FACTOR 3: HARDWARE

SPEED AND STORAGE

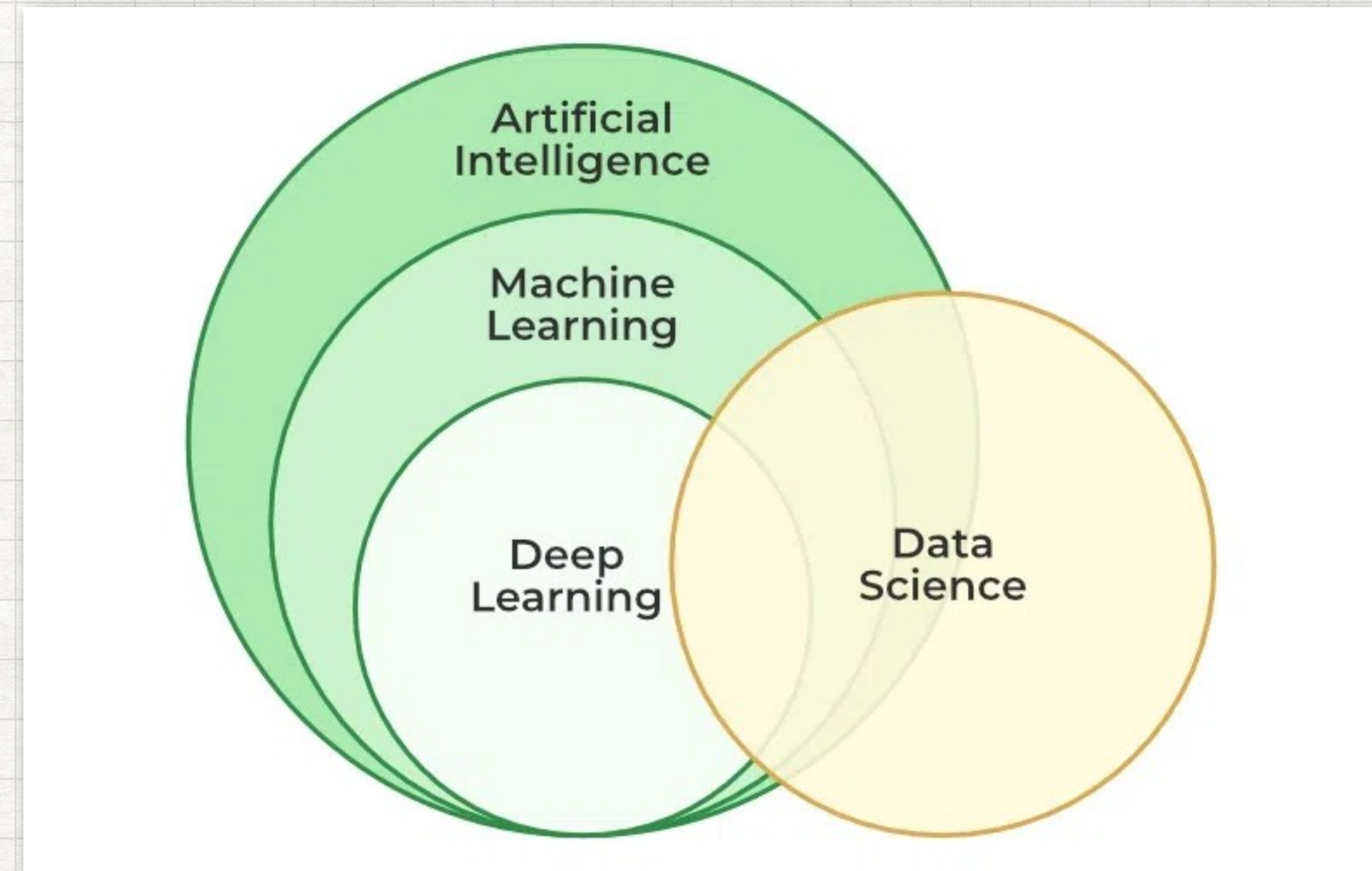
1. CPUs
2. GPUs
3. Storage
4. Cloud: AWS, Azure, Google CoLab, ...



MACHINE LEARNING

INTRINSIC PART OF THE AI UNIVERSE

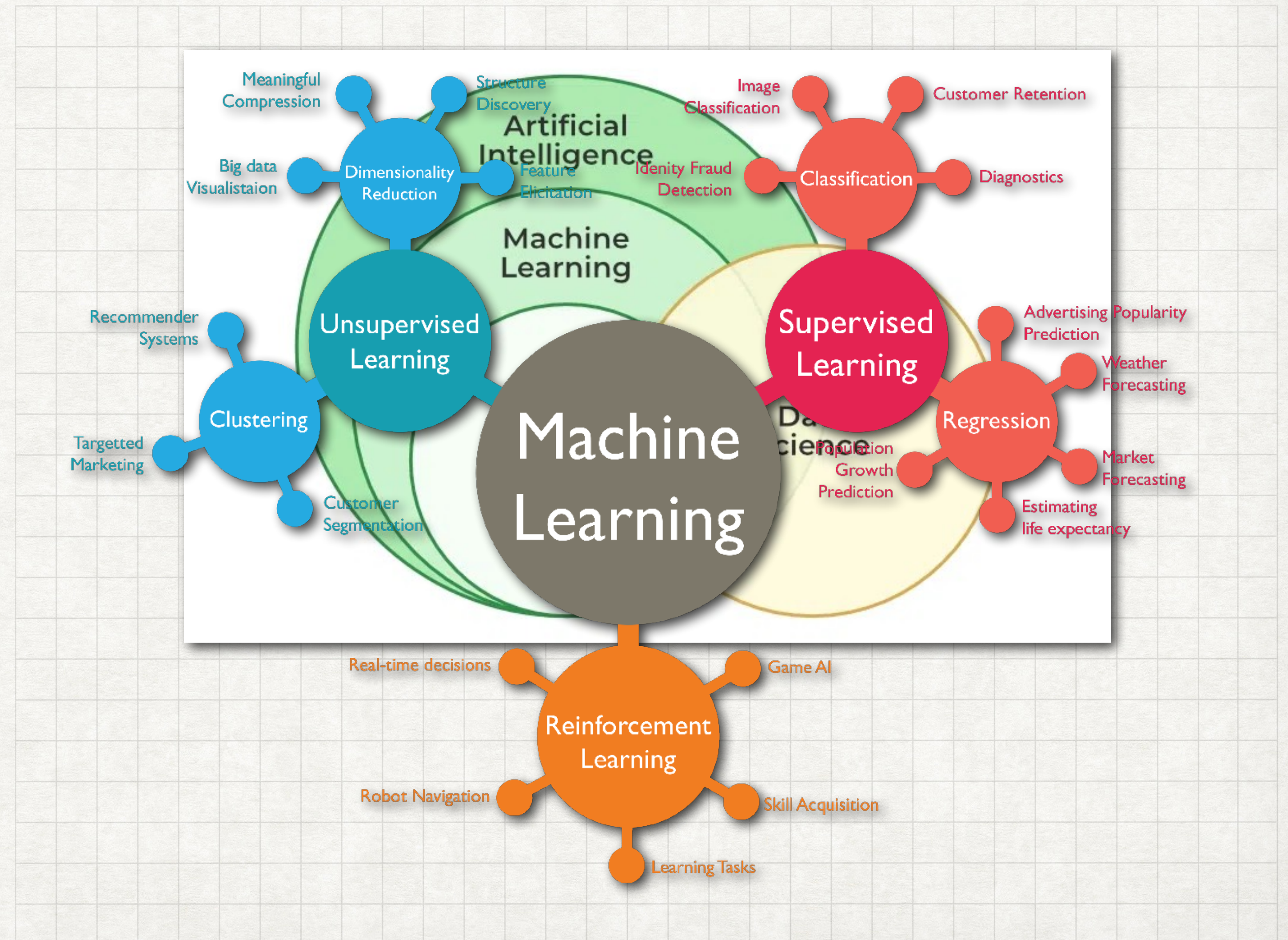
- Supervised
- Unsupervised
- Reinforcement
- Self-supervised = LLM



MACHINE LEARNING

INTRINSIC PART OF THE AI UNIVERSE

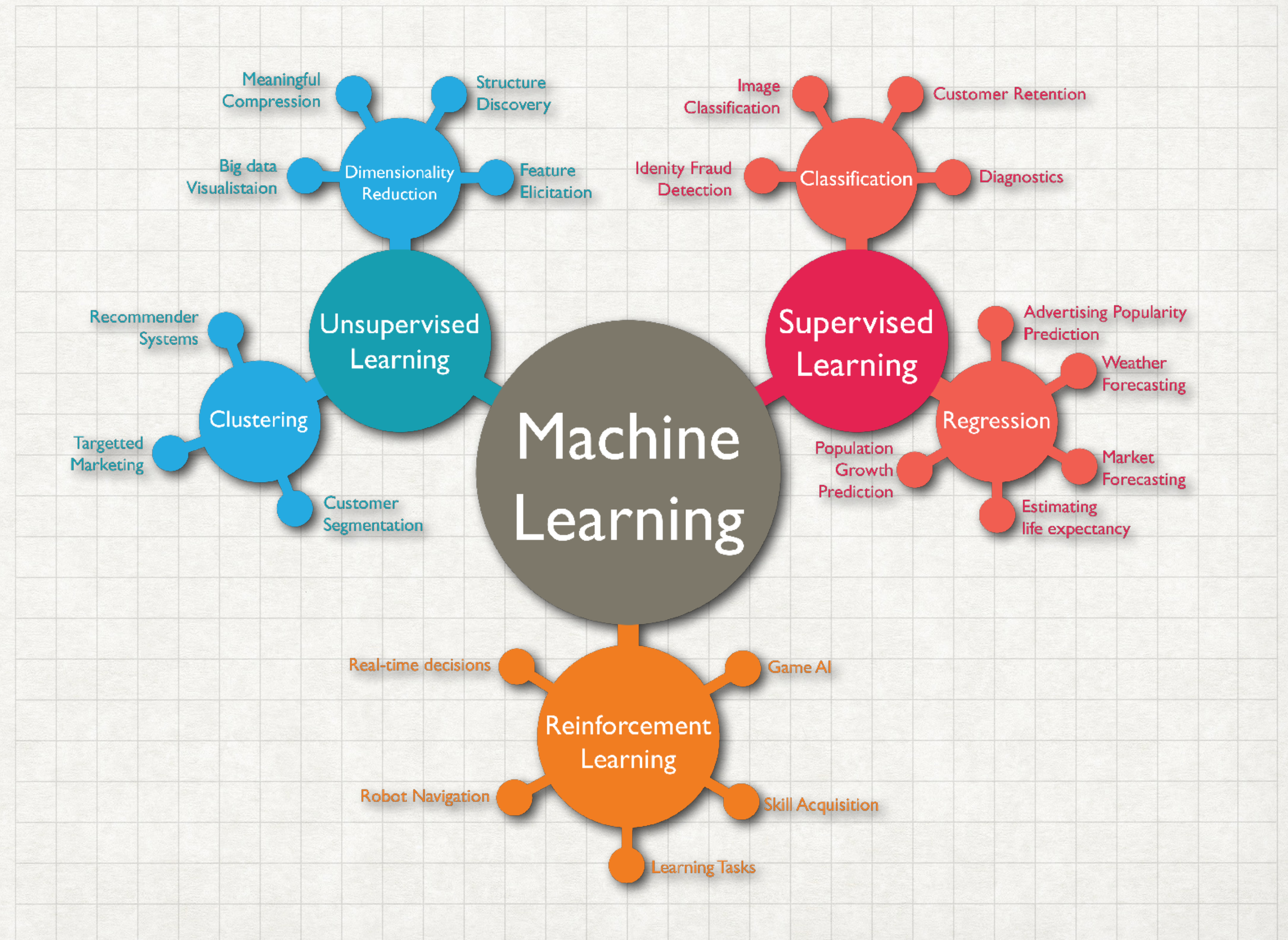
- Supervised
- Unsupervised
- Reinforcement
- Self-supervised = LLM



MACHINE LEARNING

INTRINSIC PART OF THE AI UNIVERSE

- Supervised
- Unsupervised
- Reinforcement
- Self-supervised = LLM



APPLICATIONS AND USE CASES IN HEALTHCARE

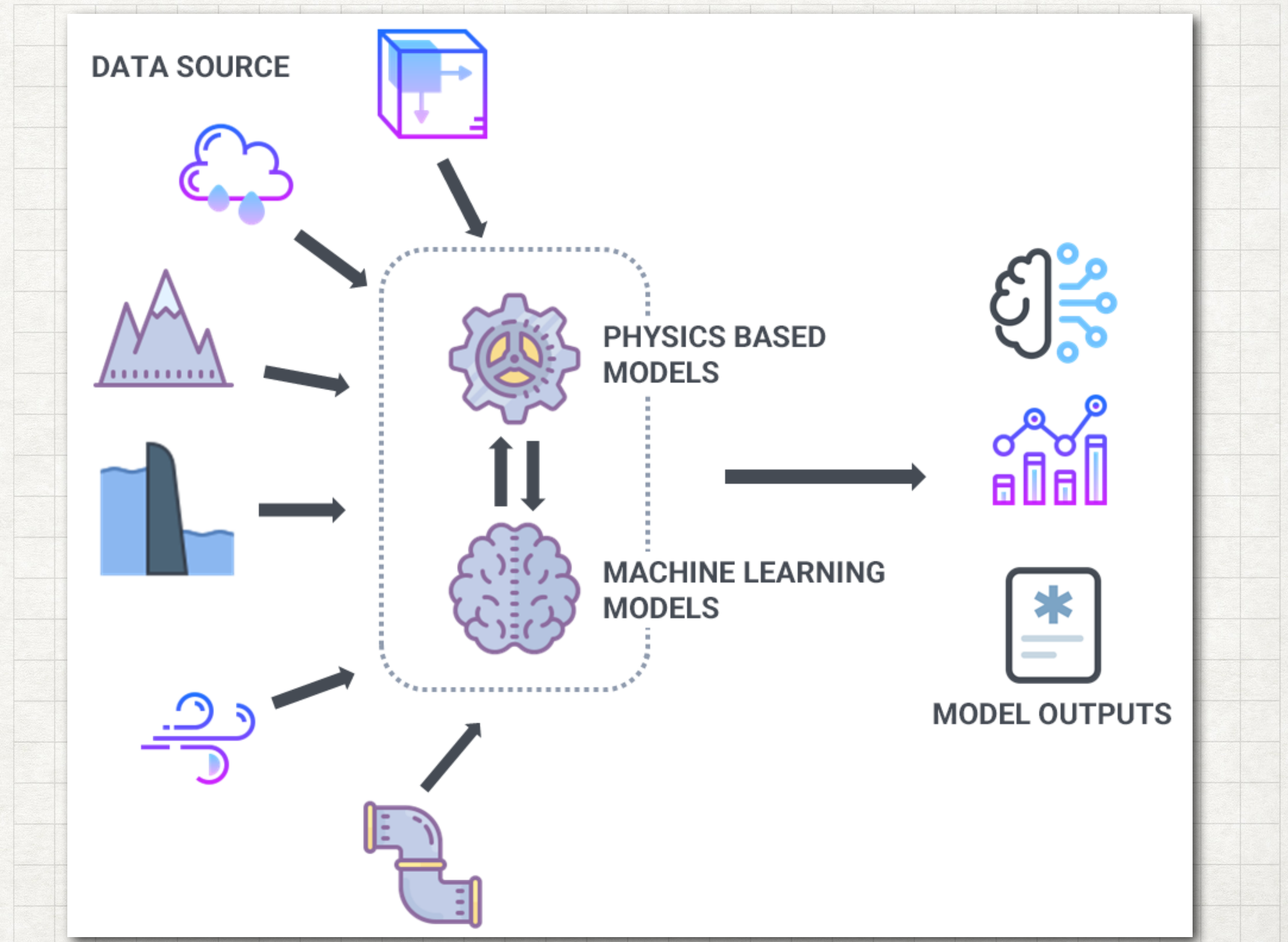
- **Disease diagnosis:** Machine learning algorithms can analyze medical images, electronic health records, and genomic data to assist healthcare professionals in diagnosing diseases such as cancer, Alzheimer's, and rare genetic disorders.
- **Treatment personalization:** Big data analytics can help healthcare providers tailor treatments to individual patients based on their unique genetic makeup, lifestyle factors, and medical history.
- **Drug discovery:** Machine learning can accelerate the drug discovery process by predicting the effectiveness and safety of new compounds, identifying potential drug targets, and optimizing clinical trial designs.
- **Population health management:** Big data analytics can help healthcare organizations identify populations at risk of chronic diseases, monitor disease outbreaks, and evaluate the effectiveness of public health interventions.
- **Mental health assessment:** Machine learning algorithms can analyze speech patterns, facial expressions, and social media data to detect early signs of mental health disorders such as depression and anxiety.
- **Wearable technology:** Sensor data from wearable devices such as fitness trackers and smartwatches can be used to monitor patients' vital signs, activity levels, and sleep patterns, enabling real-time monitoring and personalized care.
- **Medical imaging:** Machine learning algorithms can automate the analysis of medical images such as MRIs and CT scans, improving accuracy and reducing the need for manual interpretation by radiologists.
- **Clinical decision support:** Big data analytics can provide healthcare providers with real-time insights into patient health status, treatment options, and potential risks, enabling evidence-based decision making and improving patient outcomes.

SCIENCE FOR ML/AI

WHAT IS SCIENTIFIC ML?

TWO WORLDS UNITED

Scientific Machine Learning (SciML) is a field of research that combines traditional *scientific modeling* with *machine learning* techniques. It aims to develop new methods and tools for solving scientific problems that are more accurate, efficient, and generalizable than traditional methods.

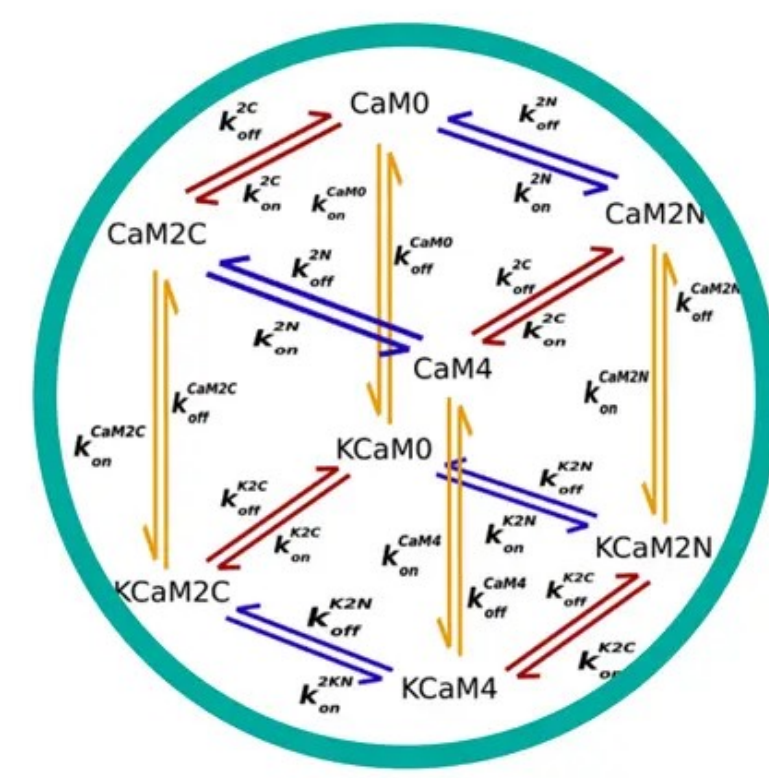
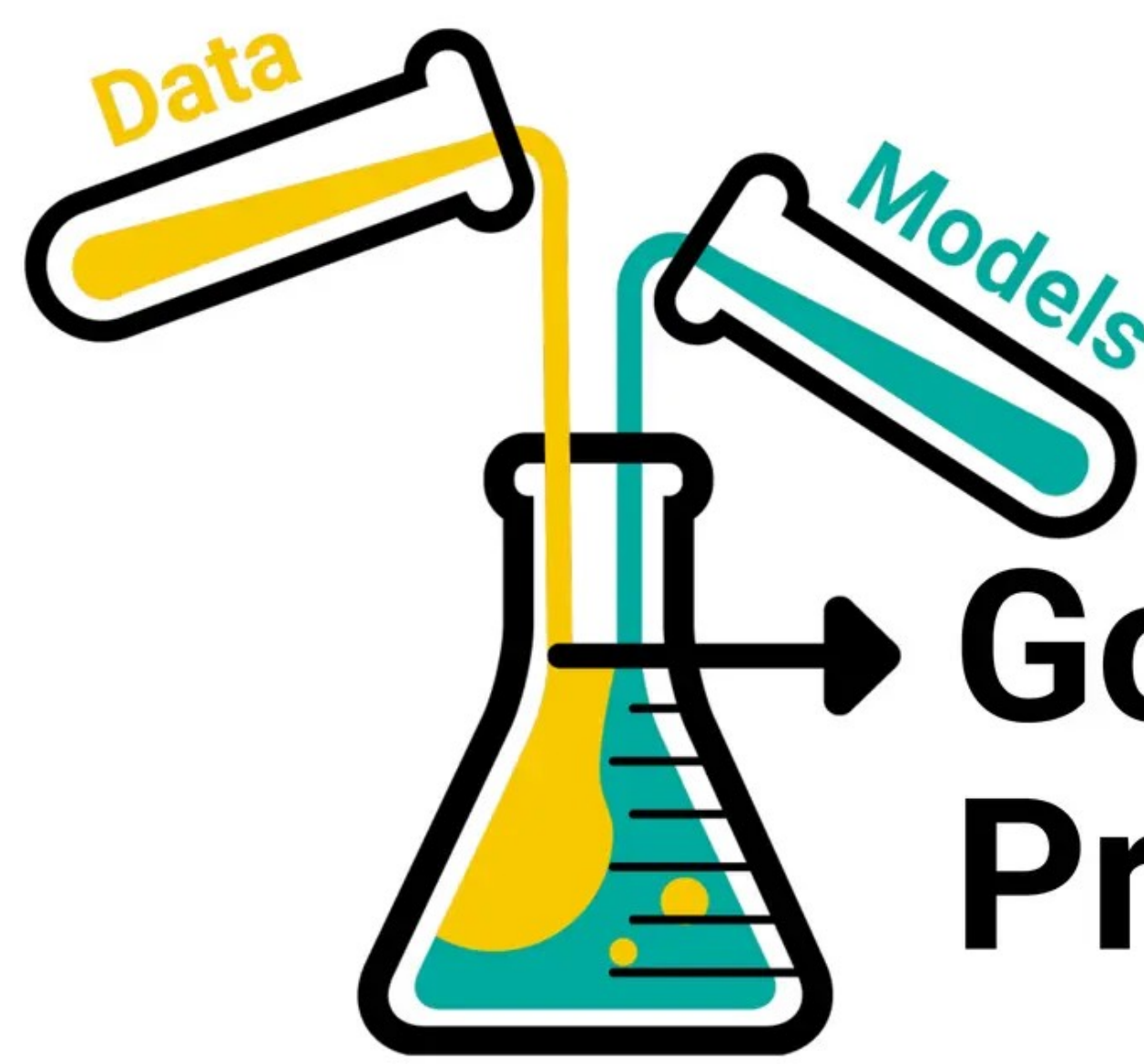
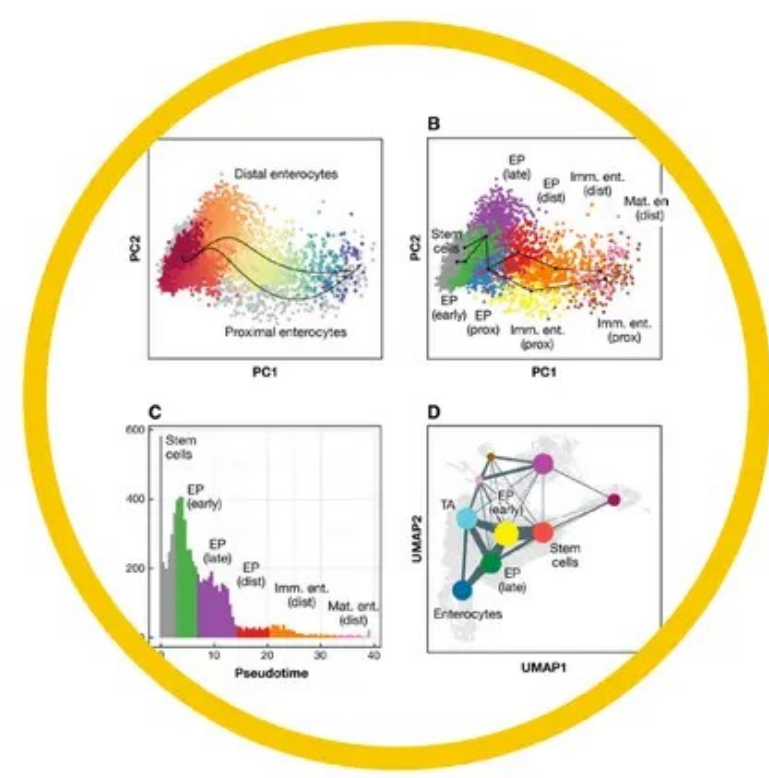


WHAT IS SCIENTIFIC ML?

BLENDING

Scientific Machine Learning is model-based data-efficient machine learning

How do we simultaneously use both sources of knowledge?



Good Predictions

HOW IS SCIENTIFIC ML DONE?

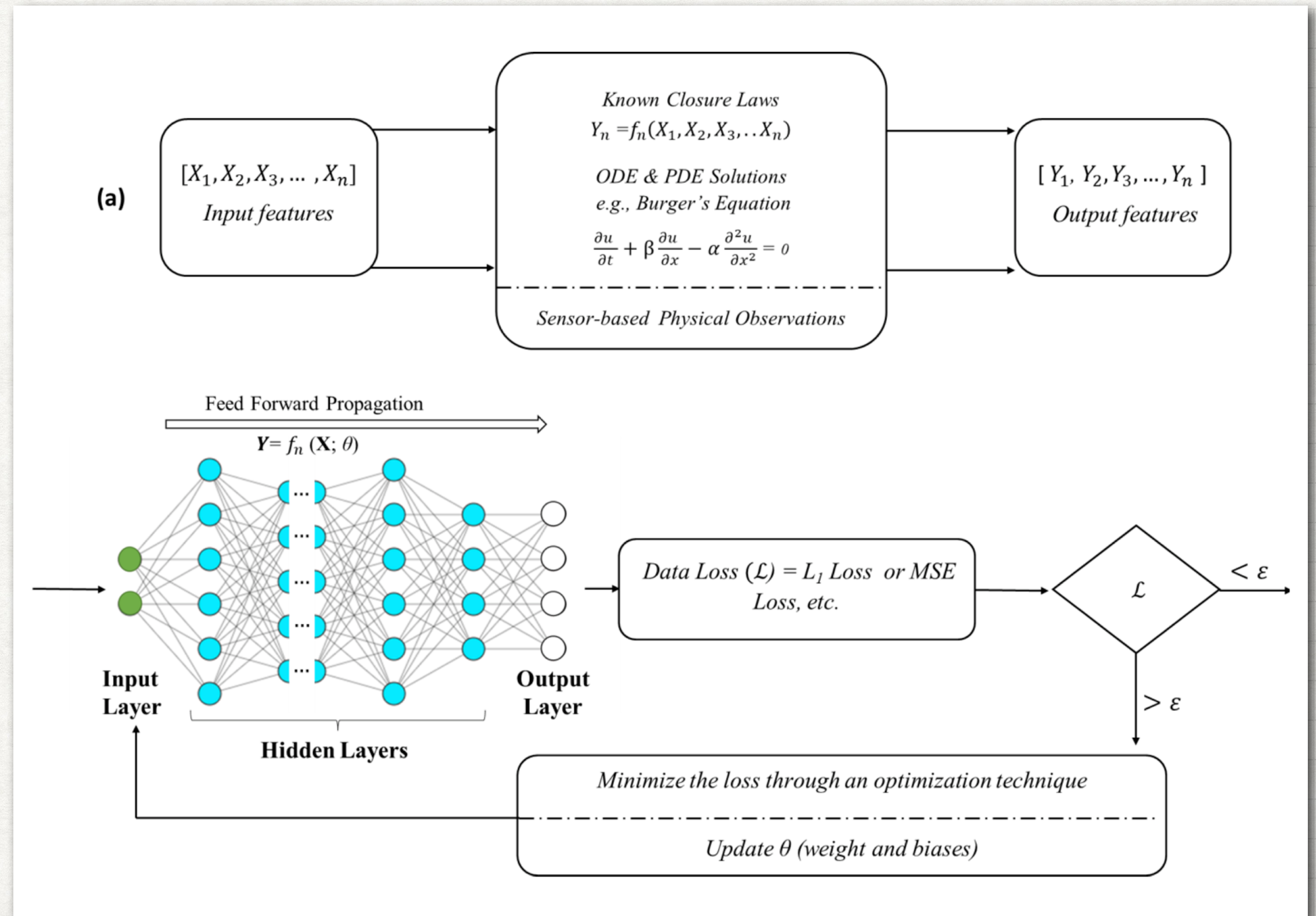
BLENDING

3 possible paths:

1. Physics-guided NNs

2. Physics-informed NNs

3. Physics-encoded NNs

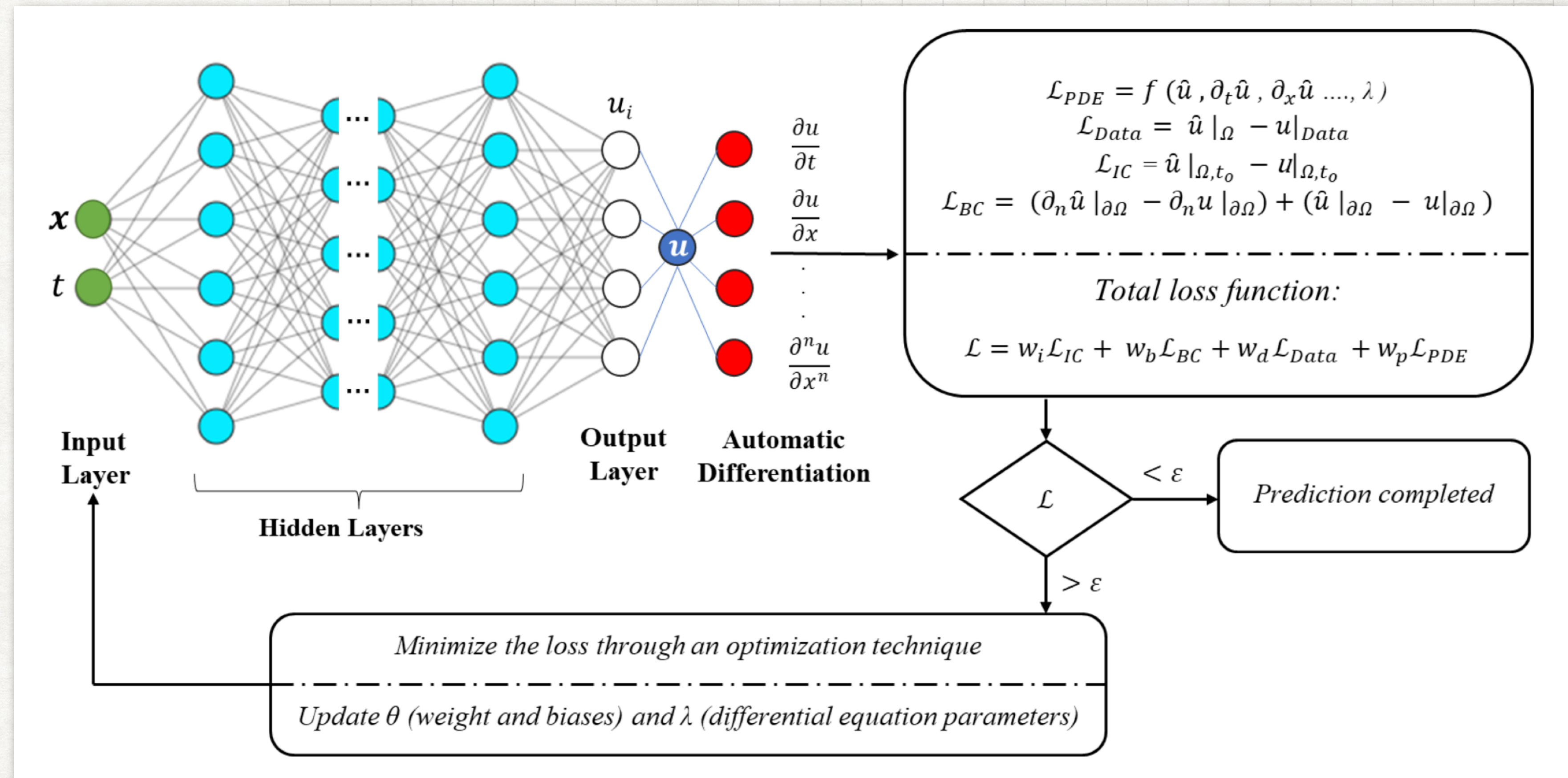


HOW IS SCIENTIFIC ML DONE?

BLENDING

3 possible paths:

1. Physics-guided NNs
2. Physics-informed NNs
3. Physics-encoded NNs

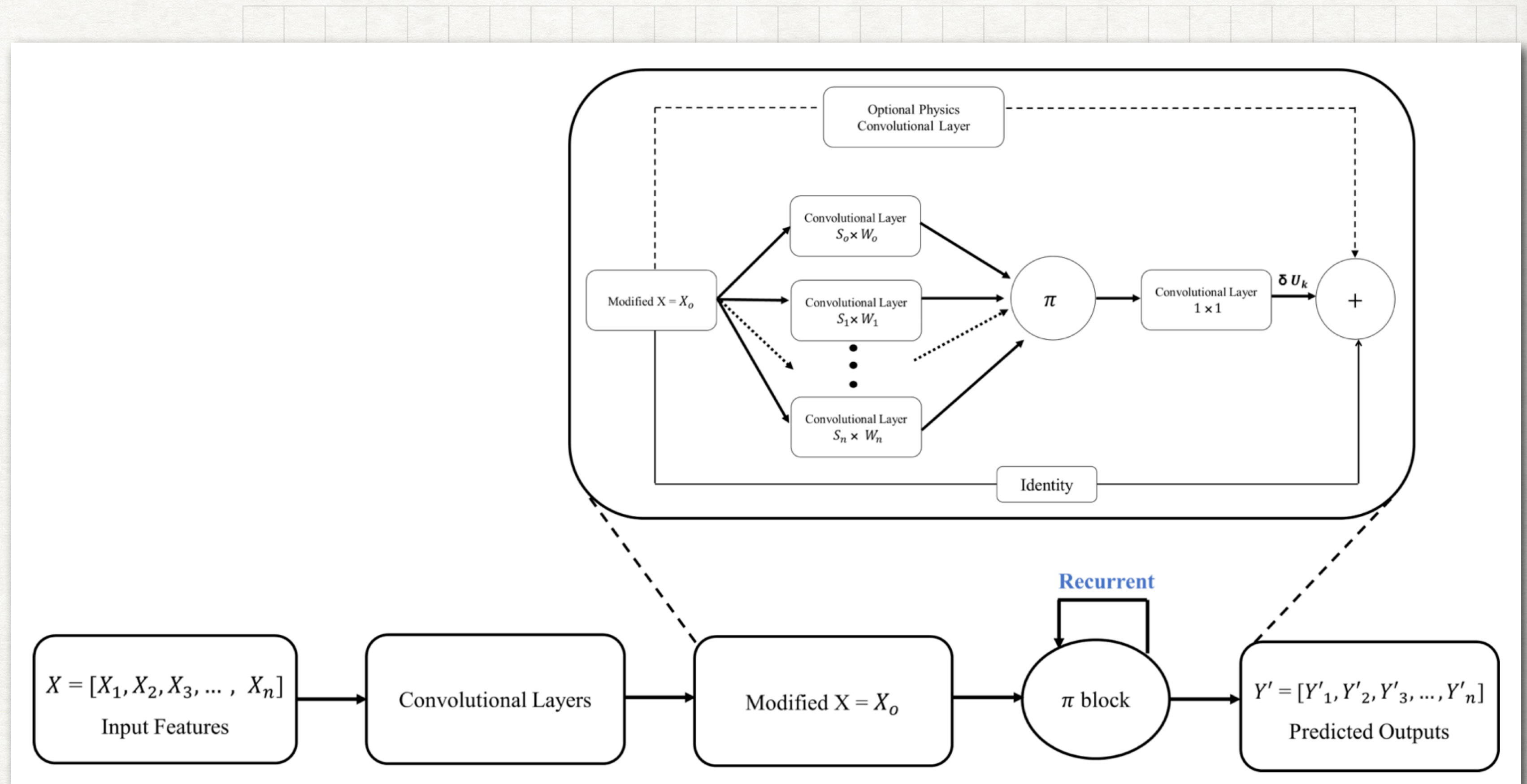


HOW IS SCIENTIFIC ML DONE?

BLENDING

3 possible paths:

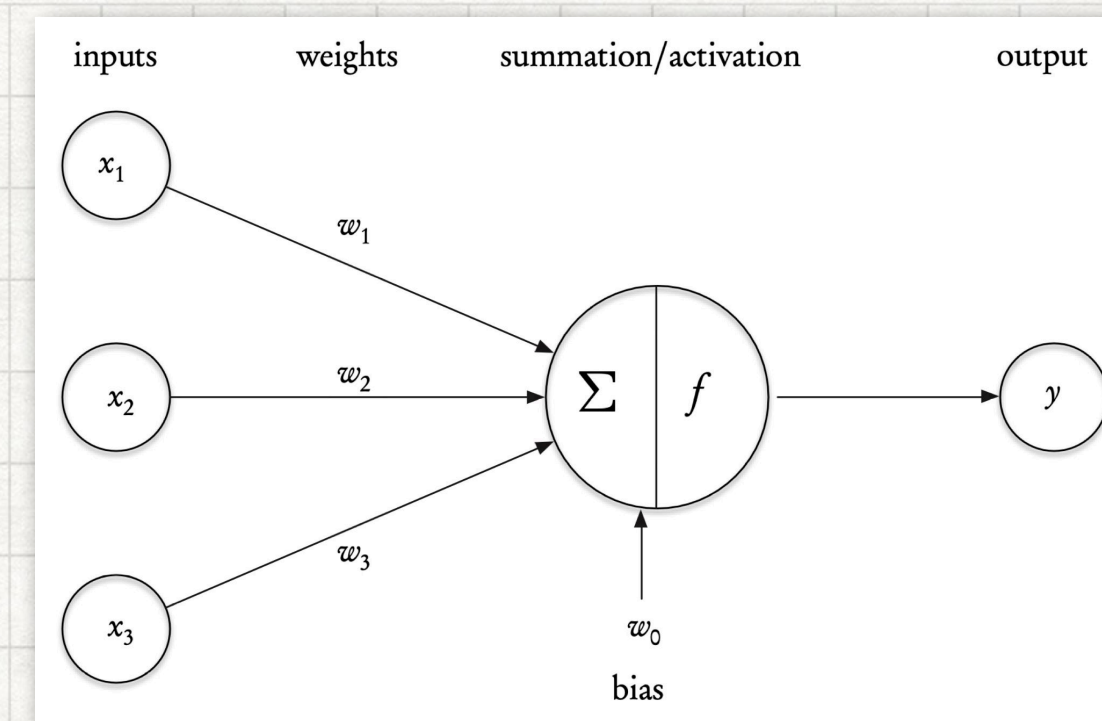
1. Physics-guided NNs
2. Physics-informed NNs
3. Physics-encoded NNs



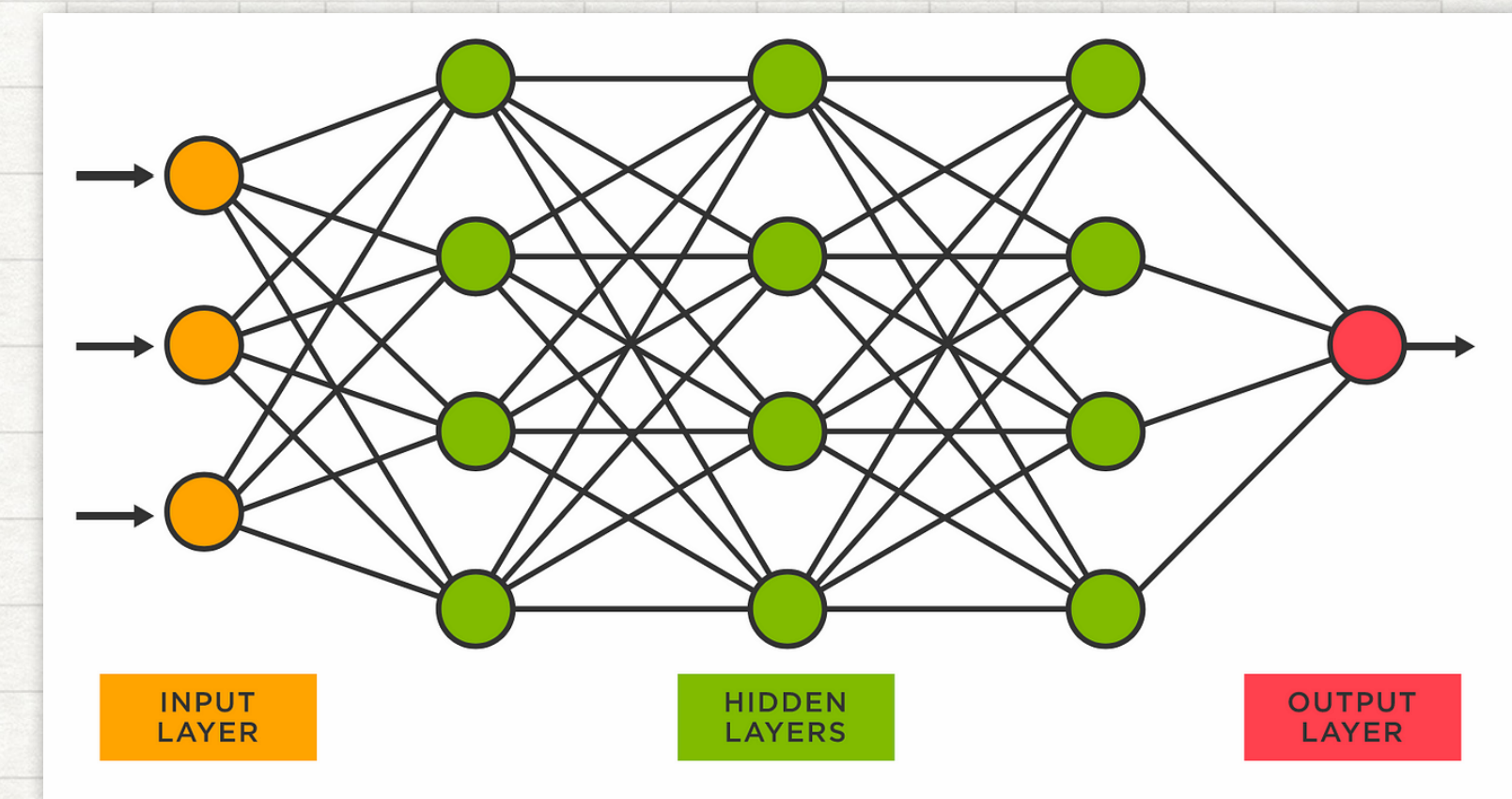
THE MATH BEHIND SCI-ML

APPROXIMATION THEORY

- Multi-layer perceptrons - 1950's - the basis.
- Universal Approximation Property - 1990's - the theory.



$$y = f \left(w_0 + \sum_{i=1}^3 w_i x_i \right)$$



Theorem 1 (Cybenko 1989). *If σ is any continuous sigmoidal function, then finite sums*

$$G(x) = \sum_{j=1}^N \alpha_j \sigma(w_j \cdot x + b_j)$$

are dense in $C(I_d)$.

Theorem 2 (Pinkus 1999). *Let $\mathbf{m}_i \in \mathbb{Z}^d$, $i = 1, \dots, s$, and set $m = \max_i |\mathbf{m}^i|$. Suppose that $\sigma \in C^m(\mathbb{R})$, not polynomial. Then the space of *single hidden layer* neural nets,*

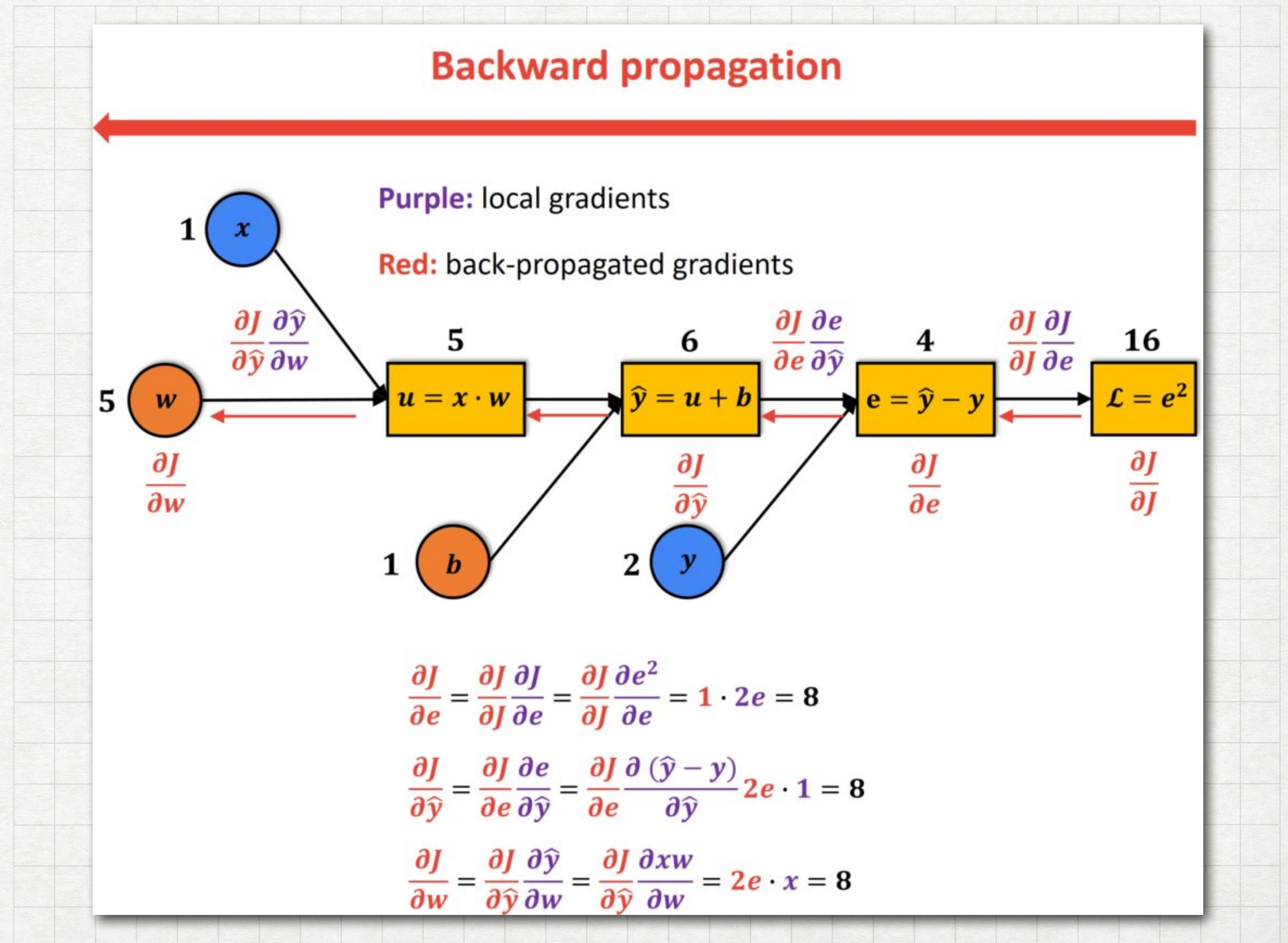
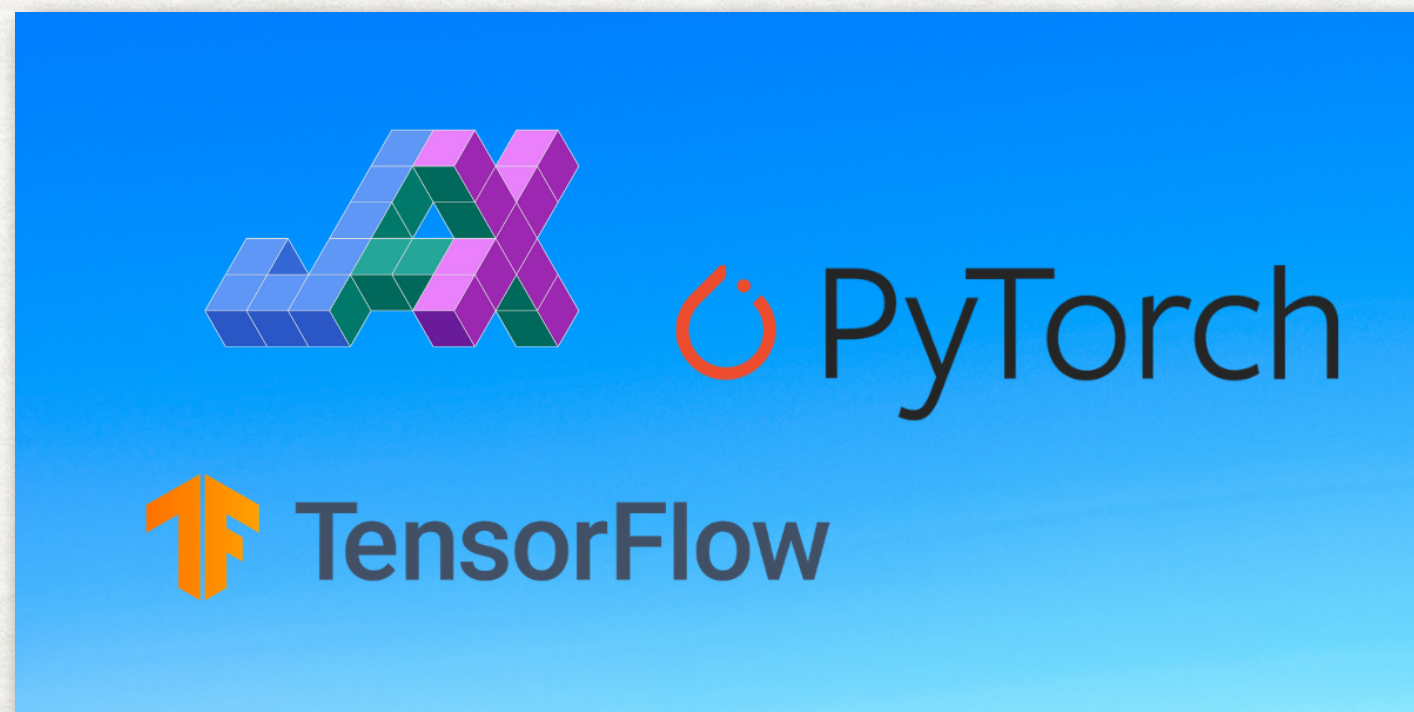
$$\mathcal{M}(\sigma) = \text{span} \{ \sigma(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R} \},$$

is dense in $C^{\mathbf{m}^1, \dots, \mathbf{m}^s}(\mathbb{R}^d) \doteq \cap_{i=1}^s C^{\mathbf{m}^i}(\mathbb{R}^d)$.

THE CODE BEHIND SCI-ML

"AUTOGRAD"

- Differentiable programming - 2020's
- makes it all possible!
- Based on:
 - Computational graphs
 - Chain-rule for differentiation



ETHICAL CONSIDERATIONS

ETHICS, BIAS, RESPONSIBILITY, TRUST

DEFINITIONS

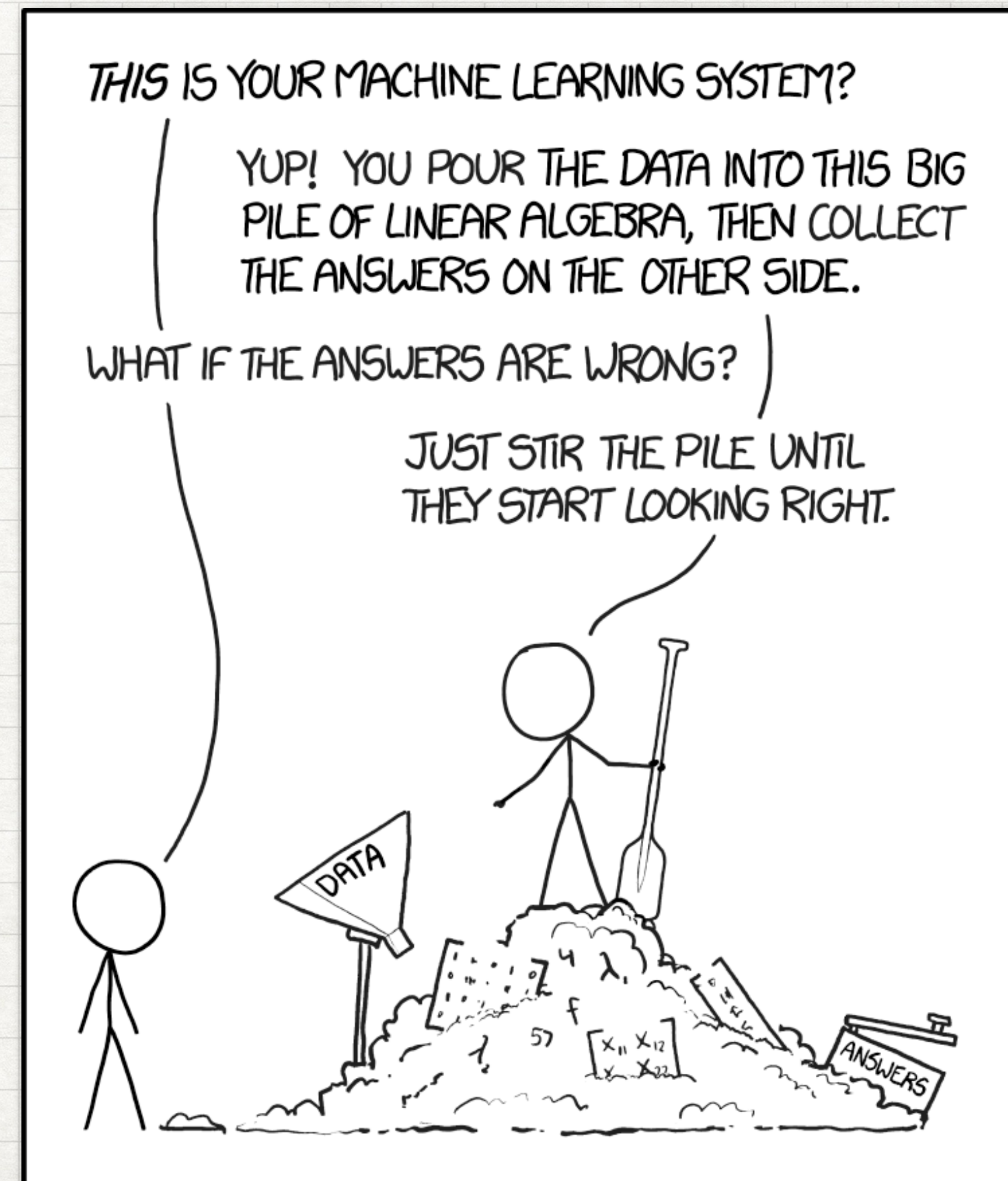
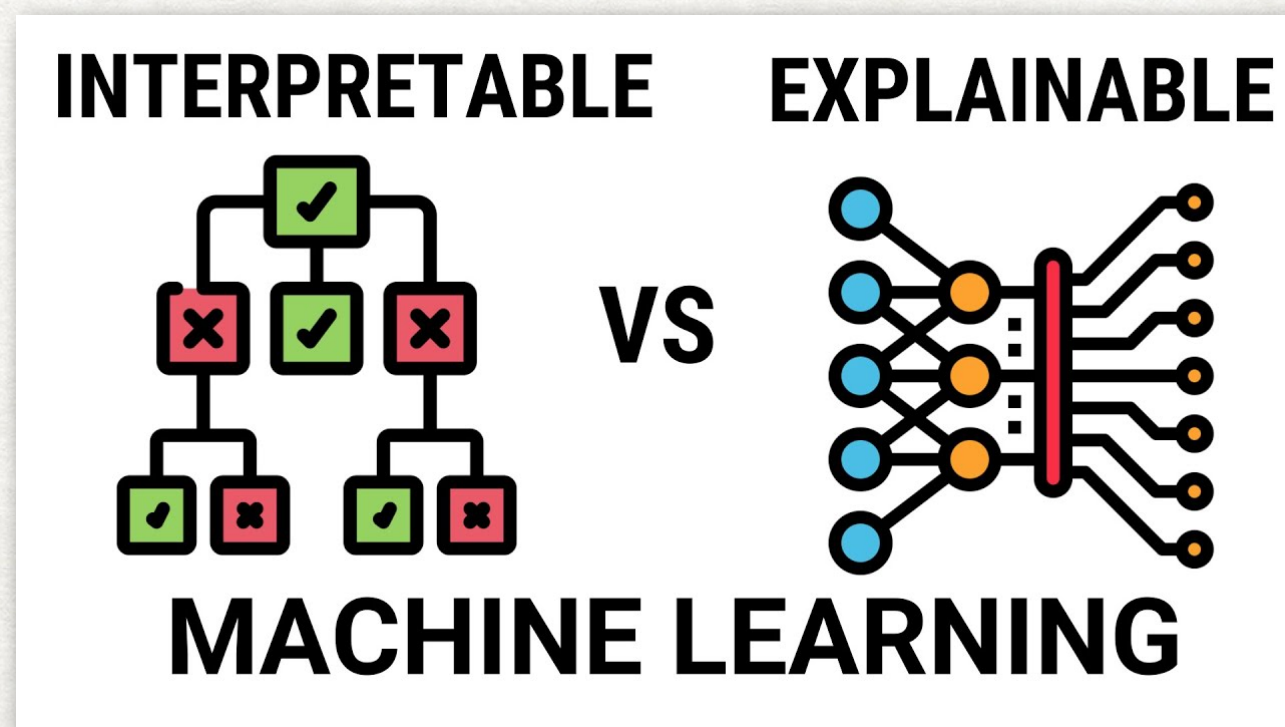
- **ETHICS** = what is morally good or bad, right or wrong? (norms)
- **BIAS** = prejudice against a person, object, position.
- **TRUST** = willingness to assume risk by relying on, or believing in, the actions of another party
- **TRUSTWORTHY AI** should be lawful, ethical, unbiased.



EXPLAINABILITY & INTERPRETABILITY

DEFINITIONS

- **INTERPRETABLE AI** = can be understood by humans without additional explanation = permits a decision of trust = not a black box
- **EXPLAINABLE AI** = can be explained post hoc, after training, in a way that makes it understandable = transparency in black boxes = feature importance, effects, interactions



HOW CAN AI GO WRONG?

TRAINING, MODELS, SOCIETY

- **Training data issues:**

- Non-representative, lack of geodiversity.
- Faulty, biased training labels.
- Adversarial effects.

- **AI model issues:**

- Learn faulty strategies.
- Fake something plausible.
- Non-trustworthy, lack of robustness.

- **Societal issues:**

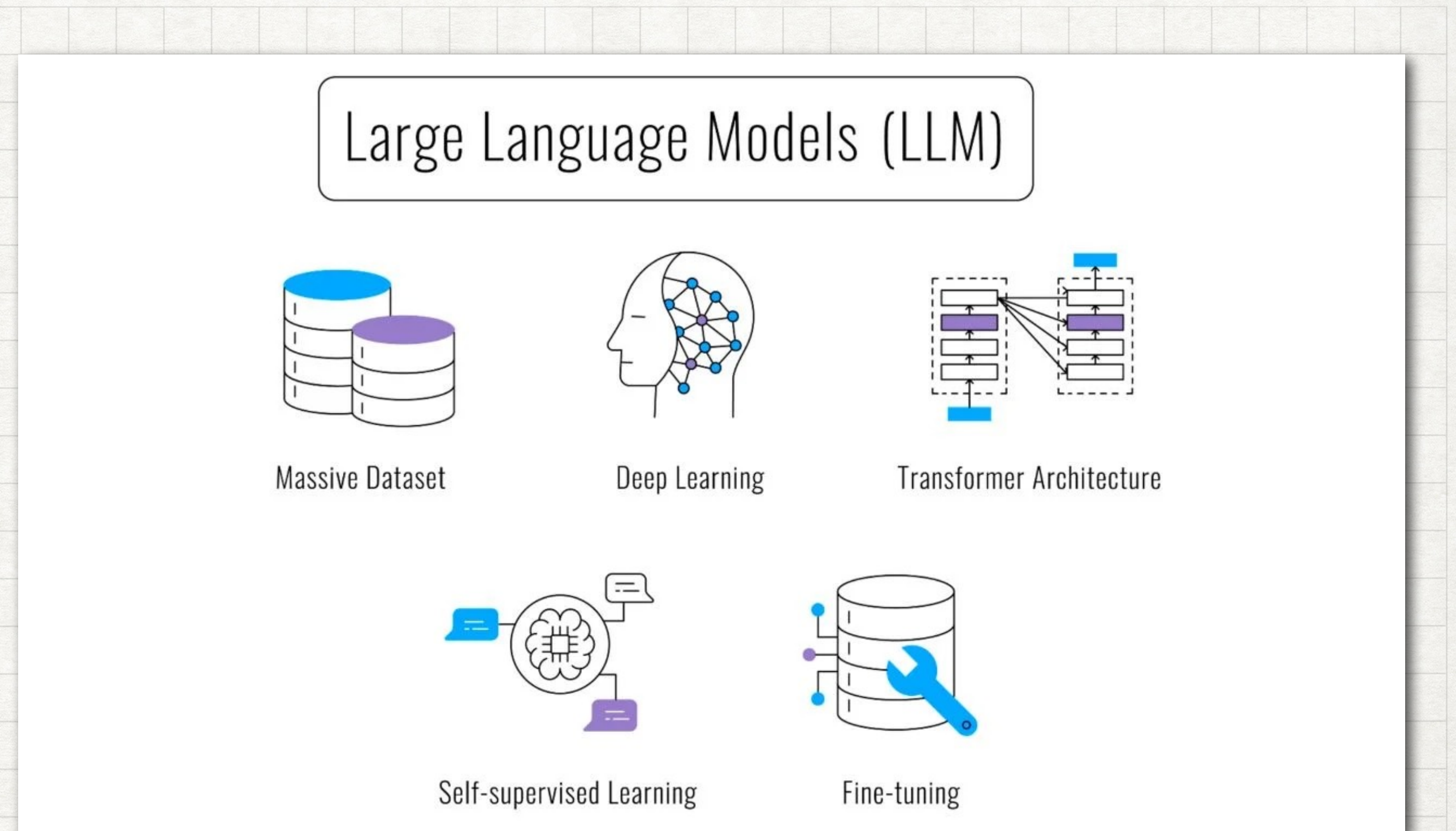
- Lack of consent on data collection.
- Disenfranchise scientists.
- CO2 emissions.
- Globally applicable AI approaches may stymie burgeoning efforts in developing countries.

**LLM'S - THE FUTURE OF
SCIENCE?**

LARGE LANGUAGE MODELS

DEFINITION AND FUNCTIONS

- Deep **learning** models, trained on massive **data** sets, that perform **NLP** tasks (translate, predict, **generate**).
- A glorified **chatbot/parrot**???
- YES, but
 - Can be trained to “speak” physics, biology, epidemiology, etc., etc.
- Tools such as fine-tuning, RAG (prompt engineering).



LARGE LANGUAGE MODELS

APPLICATIONS

- Current:

- **NLP** - text generation, chatbots
- **Education** - tutoring, languages
- **Healthcare** - medical record analysis, symptom checker
- **Business** - customer support, content creation, market analysis
- **Software** - code generation, documentation

- Future

- **R&D** - NWP, drug discovery, research assistant
- **Human-Machine interaction**
- **Autonomous Systems** - assistive technologies, robotics

LARGE LANGUAGE MODELS

IMPACT

- Economic

- Job transformation
- Productivity gains

- Social and Ethical

- Bias and fairness
- Privacy concerns

- Educational

- Personalize learning
- Critical thinking

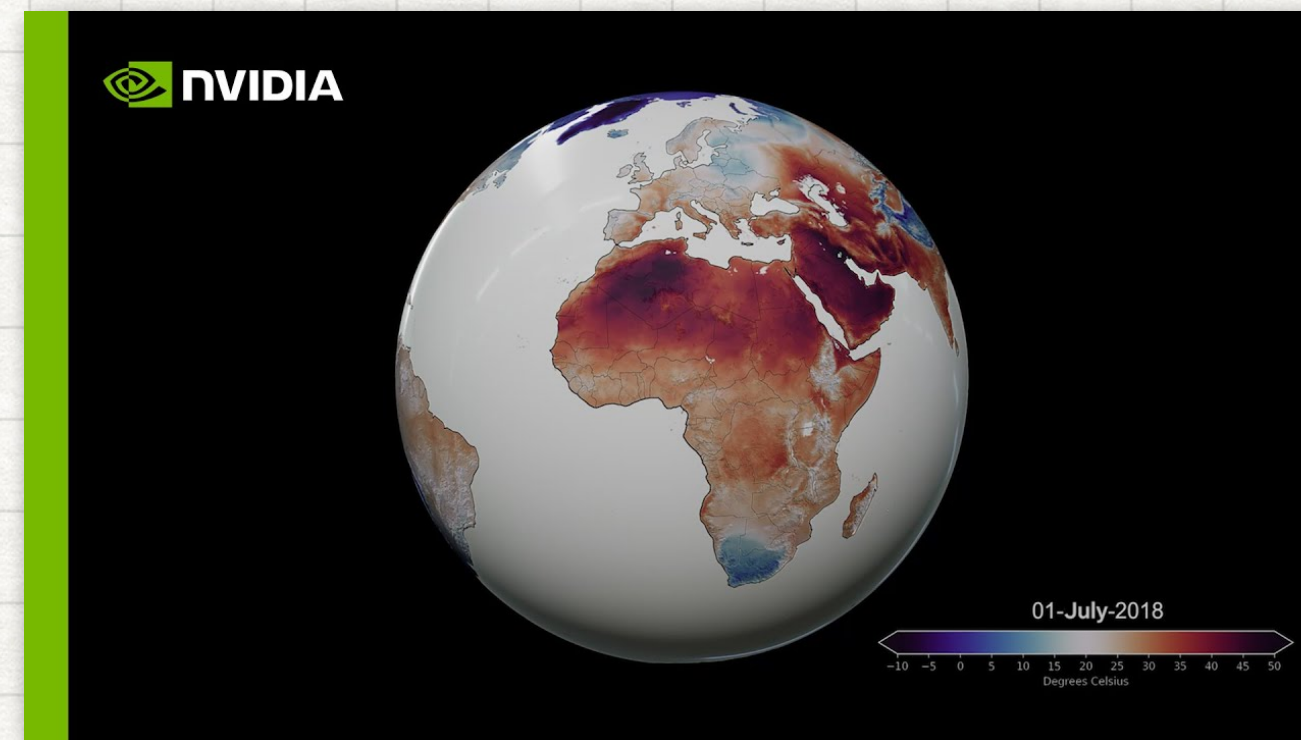
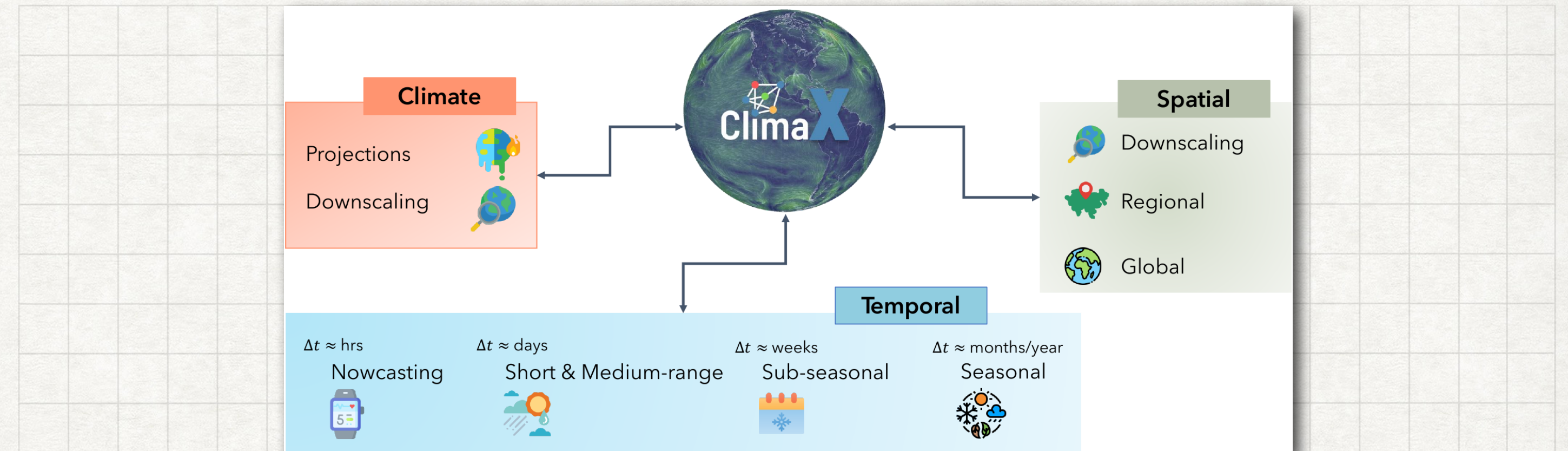
- Healthcare

- Improved diagnostics
- Patient-health system interactions

LLM APPLICATIONS

NWP & CLIMATOLOGY USING LLM'S

- NVIDIA, Huawei, Microsoft, Google
- Trained on (lots of) historical reanalysis data
- Claims:
 - 10 000X speedup!
 - Accuracy for nowcasts
- Is health data next?



Generative AI to quantify uncertainty in weather forecasting

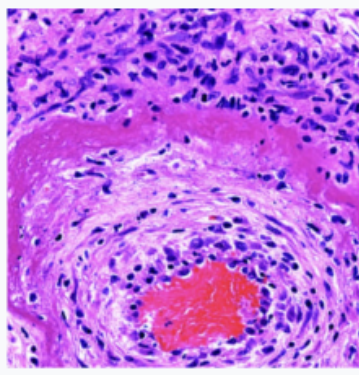
LLM FOR HEALTHCARE

JUST THE BEGINNING?

- Google announces Med-Gemini <https://arxiv.org/pdf/2404.18416v2> (29 April 2024)
- Family of **multimodal** models built upon Gemini specifically designed for the healthcare industry.
- “Groundbreaking Family of AI Models **Revolutionizing** Medical Diagnosis and Clinical Reasoning”
- Models excel in **multimodal** tasks, with substantial improvements in analyzing medical images and videos and accurately retrieving information from long health records

Open-ended Visual QA (Path-VQA)

Visual input




Instruction
You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.
Question: What does the wall of the artery show with protein deposition and inflammation?

Response
a circumferential bright pink area of necrosis

Image Classification (PAD-UFES-20 6-condition classification)

Visual input



Instruction
You are a helpful dermatology assistant. The following are questions about skin lesions. Categorize the skin lesions into the most likely class given the patient history. Output a single option letter from the provided options as the final answer.
Patient History: Age: 51, Gender: female, Smoke: false, Drink: false, Family skin cancer history: true, Family any cancer history: false, Lesion region: back, Lesion itch: false, Lesion grew: false, Lesion bled: false, Lesion elevation: false, Fitzpatrick scale: 1.0, Diameters (mm): [12.0, 8.0].
Question: Which of the following is the most likely diagnosis of the patient's skin lesion? (A) Nevus (B) Basal Cell Carcinoma (C) Squamous Cell Carcinoma (D) Actinic Keratosis (E) Seborrheic Keratosis (F) Melanoma.

Response
(A)

Open-ended Visual QA in Chinese (Slake-VQA)

Visual input




Instruction
You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.
Question: 图像里包含的区域属于身体哪个部分?

Response
腹部

Image Classification (MIMIC-CXR 13-condition classification)

Visual input




Instruction
You are a helpful radiology assistant. The following are questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray.
Given the <VIEW> X-ray image,
Question: Which of the following abnormalities are indicated by the image? (A) Atelectasis (B) Cardiomegaly (C) Consolidation (D) Edema (E) Enlarged Cardiomeastinum (F) Fracture (G) Lung Lesion (H) Lung Opacity (I) Pleural Effusion (J) Pleural Other (K) Pneumonia (L) Pneumothorax (M) Support Devices

Response
(A)

Close-ended Visual QA (NEJM Image Challenge, USMLE-MM)

Visual input




Instruction
You are a medical expert. Only output the final (diagnosis, answer). Do not output the reasoning or explanation. Output the final diagnosis in the format "Final [Diagnosis, Answer]: X" where X is the most (possible medical diagnosis, correct letter choice).
Question: Infection with which one of the following organisms is the most likely cause of this rash? (A) Coxsackie virus type A16 (B) Echovirus type 16 (C) Group A streptococcus (D) Herpes simplex virus type 1 (E) Norwalk virus

Response
Final Answer: (A)

Image Classification (MIMIC-CXR normal vs abnormal classification)

Visual input




Instruction
You are a helpful radiology assistant. The following are questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray.
Given the <VIEW> X-ray image,
Question: are there any abnormalities indicated by the image? (A) Yes (B) No.

Response
(A)

Waveform Signal Visual QA (ECG-QA)

Raw sensor input*




Instruction
Given this ECG sequence, please answer the following question. From the provided options, select all that apply. List your selections alphabetically, separated by commas.
Question: What signs of a rhythm-related disorder can be found in this ECG recording?
Options: atrial fibrillation, atrial flutter, bigeminal pattern, normal functioning artificial pacemaker, sinus arrhythmia, sinus bradycardia, sinus rhythm, sinus tachycardia, supraventricular tachycardia

Response
atrial fibrillation, atrial flutter

Text Report Generation (MIMIC-CXR)

Visual input



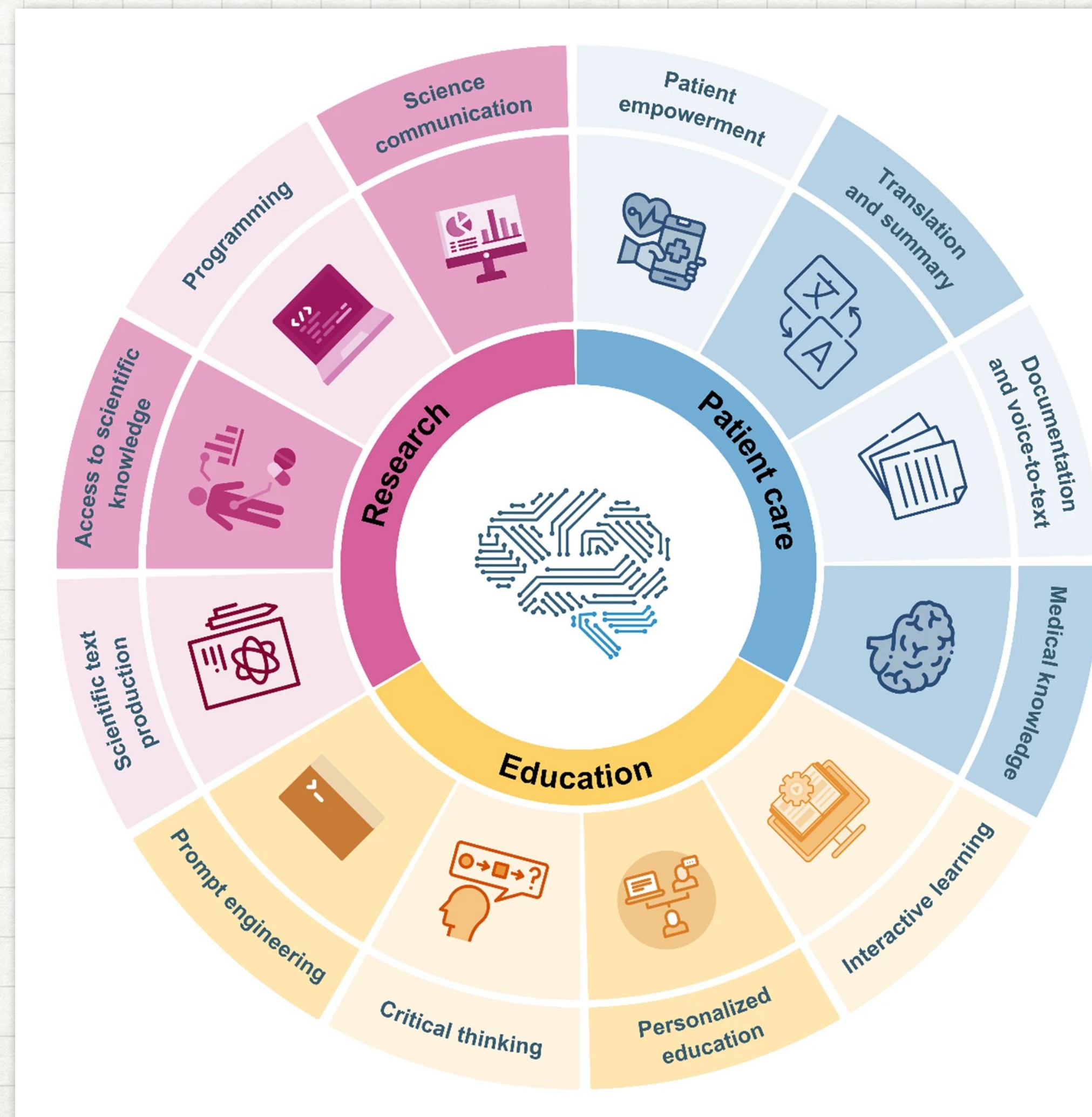
Instruction
You are a helpful radiology assistant.
Question: Describe the findings and impression in the chest X-ray following the instructions.

Response
FINDINGS: Low lung volumes are present. This accentuates the size of the cardiac silhouette which is likely mildly enlarged. Mediastinal and hilar contours are likely within normal limits. A right brachiocephalic venous stent is re-demonstrated. There is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. No pleural effusion or pneumothorax is identified.
IMPRESSION: Low lung volumes with mild pulmonary vascular congestion.

LARGE LANGUAGE MODELS

TOMORROW?

- Small number of **very large**, intelligent LLMs
- Many smaller, **specialized** LLMs
 - Healthcare, legal, finance, etc.
- **Personal** LLMs (your "story")
- Uses in **education**
 - How to educate?
 - What to teach?



THANK YOU!

CONTACT


DETAILS

- mark.asch@u-picardie.fr
- <https://www.linkedin.com/in/mark-asch-8a257130/>
- <https://github.com/markasch>
- <https://markasch.github.io/DT-tbx-v1/>



REFERENCES

BOOKS & PAPERS

- M. Asch 
- S A Faroughi, N Pawar, C Fernandes, M. Raissi, S. Das, N K Kalantari, S K Mahjour. Physics-Guided, Physics-Informed, and Physics-Encoded Neural Networks in Scientific Computing. arXiv, 2023. <https://arxiv.org/pdf/2211.07377>
- S. Cuomo, V. Schiano Di Cola, F. Giampaolo, G. Rozza, M. Raissi, F. Piccialli. Scientific Machine Learning Through Physics Informed Neural Networks: Where we are and What's Next. Journal of Scientific Computing (2022) 92:88.
- L Lu, P Jin, G Pang, Z Zhang, GE Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nature Machine Intelligence 3 (3), 218-229 (2021).
- L. Lu, X. Meng, Z. Mao, G. Karniadakis. DeepXDE: A Deep Learning Library for Solving Differential Equations. SIAM Review, 63, 1 (2021).
- E. Darve, K. Xu. Physics constrained learning for data-driven inverse modeling from sparse observations. J. of Computational Physics, 453. (2022).
- M. Raissi, P. Perdikaris, G. E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. of Computational Physics, 378, pp. 686-707, 2021.
- N. Kovachki, Z. Li, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkulmar. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. J. of Machine Learning Research 24 (2023).
- A. Baydin, B. Pearlmutter, A. Radul, J. Siskind. Automatic differentiation in machine learning: a survey. Journal of Machine Learning Research, 18 (2017), article 153.

