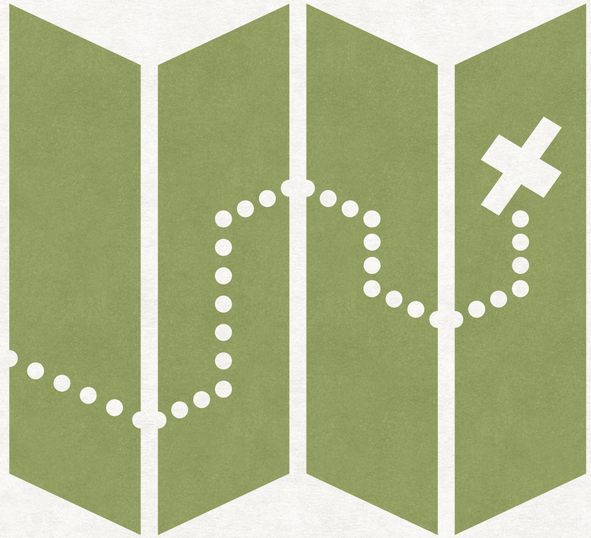
MARK ASCH

# ML/AI FOR SCIENCE
# SCIENCE FOR AI/ML

ML-PREP Training - March-April 2025

# 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

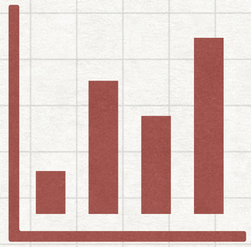
# AI REVOLUTION OF 2020'S

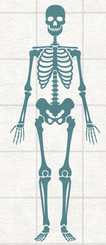
## 3 MAJOR REASONS/CAUSES/FACTORS

# AI REVOLUTION OF 2020'S

## 3 MAJOR REASONS/CAUSES/FACTORS

1. Big Data

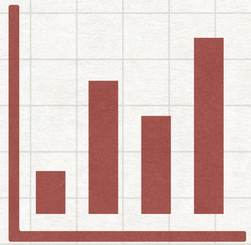


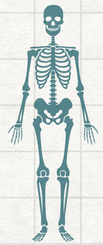




# AI REVOLUTION OF 2020'S

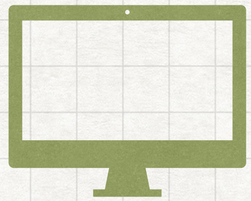
## 3 MAJOR REASONS/CAUSES/FACTORS

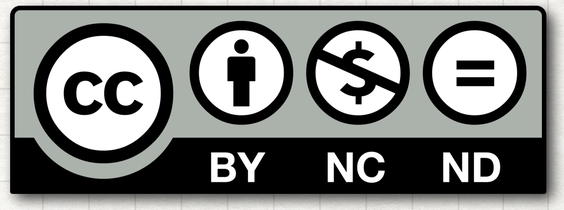
1. Big Data




2. Open source (ML) software librairies

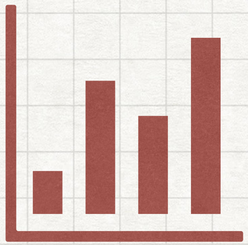


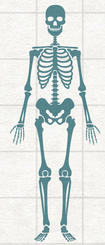


# AI REVOLUTION OF 2020'S

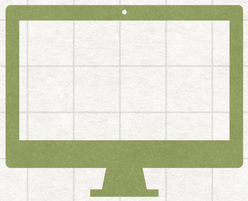
## 3 MAJOR REASONS/CAUSES/FACTORS

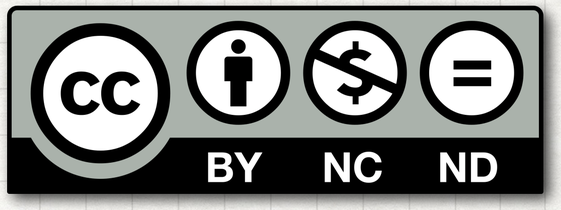
1. Big Data



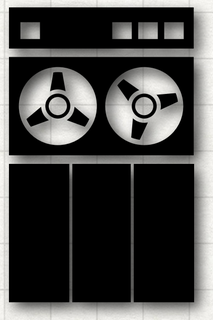
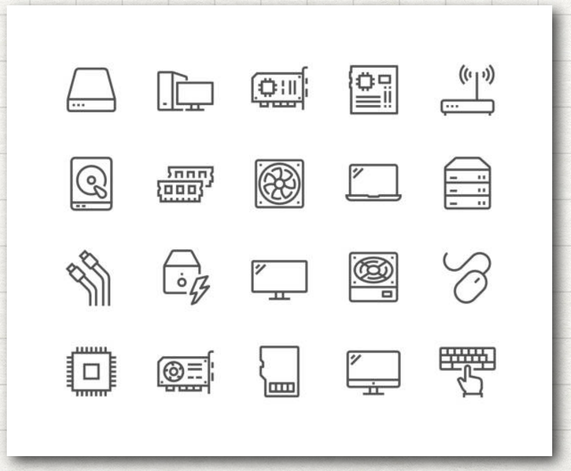
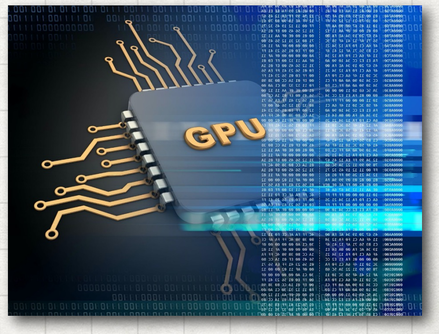




2. Open source (ML) software librairies






3. Access to (cheap) computer hardware and storage

# FACTOR 1: BIG DATA

## THE 4 V'S OF BIG DATA

### Volume
SCALE OF DATA

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

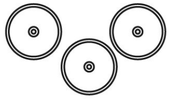

**6 BILLION PEOPLE**
have cell phones
WORLD POPULATION: 7 BILLION

**2.5 QUINTILLION BYTES**
of data are created each day

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

### Variety
DIFFERENT FORMS OF DATA

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

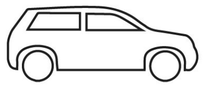
**4 BILLION + HOURS OF VIDEO**
are watched on You Tube each month

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

### Velocity
ANALYSIS OF STREAMING DATA

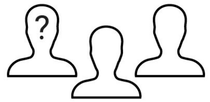
The New York Stock Exchange captures
**1TB OF TRADE INFORMATION**
during each trading session

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



### Veracity
UNCERTAINITY OF DATA

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



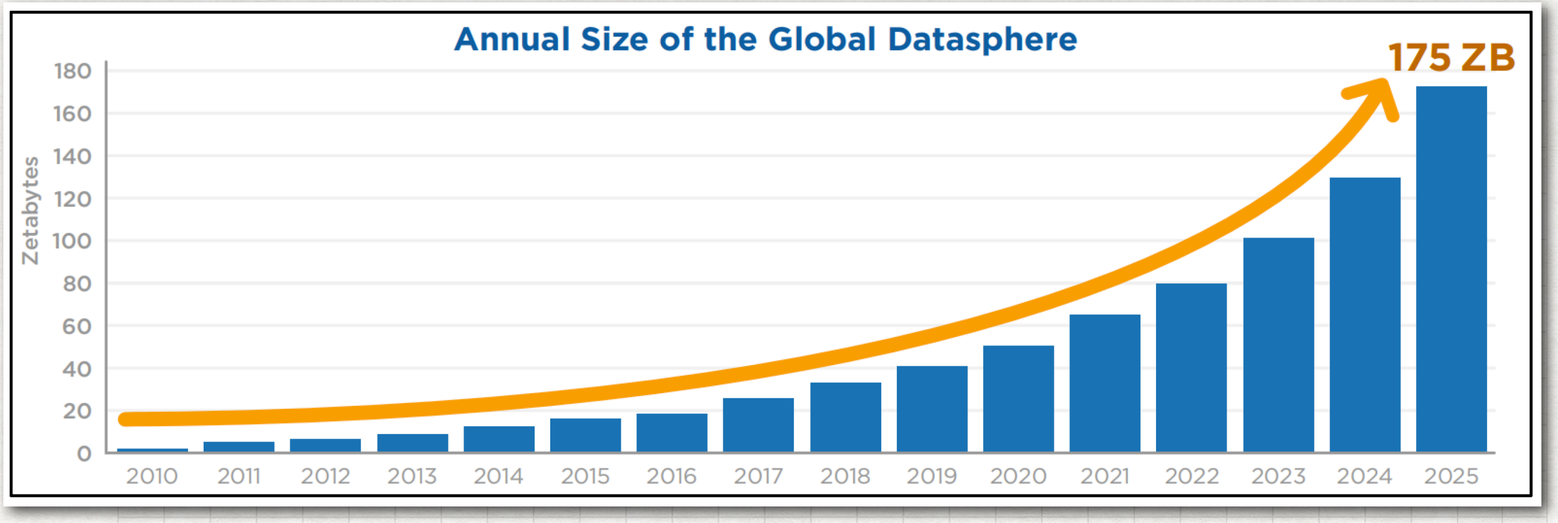

**27% OF RESPONDENTS**
in one survey were unsure of how much of data was inaccurate

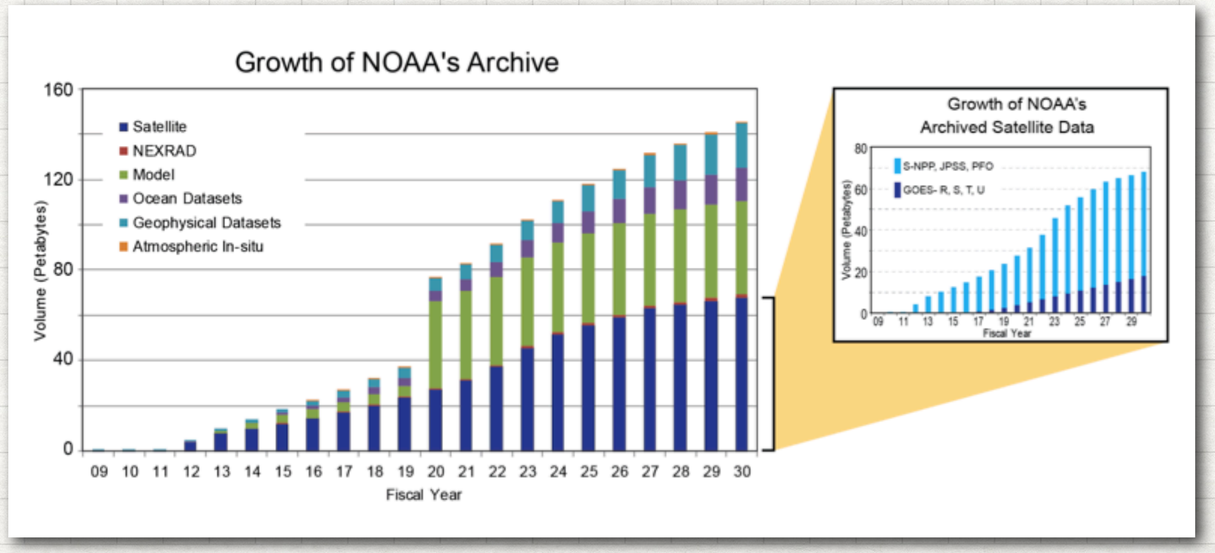

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

# FACTOR 1: BIG DATA

## EXAMPLE: ENVIRONMENTAL DATA IS EVERYWHERE...

1. Satellite Data
2. Sensor Networks
3. Citizen Science
4. Smart Meters and BIMs
5. Mobile Phone Data

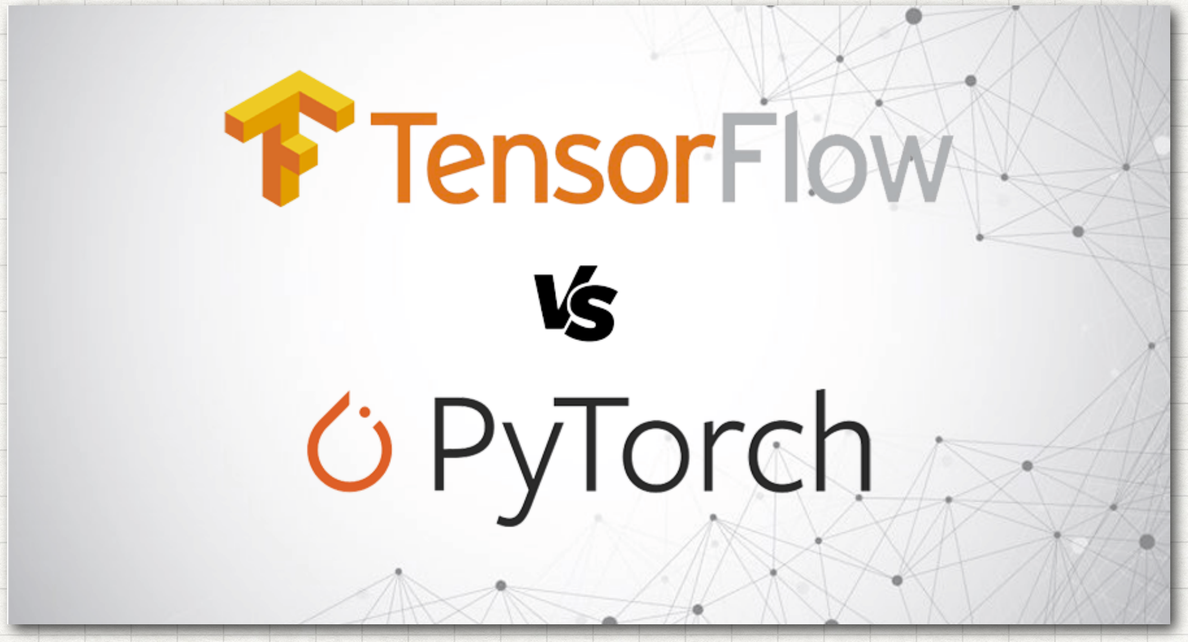


# FACTOR 2: OPEN SOURCE

## 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

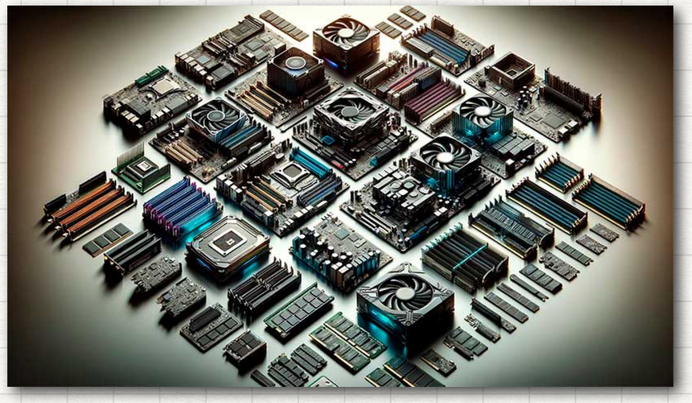


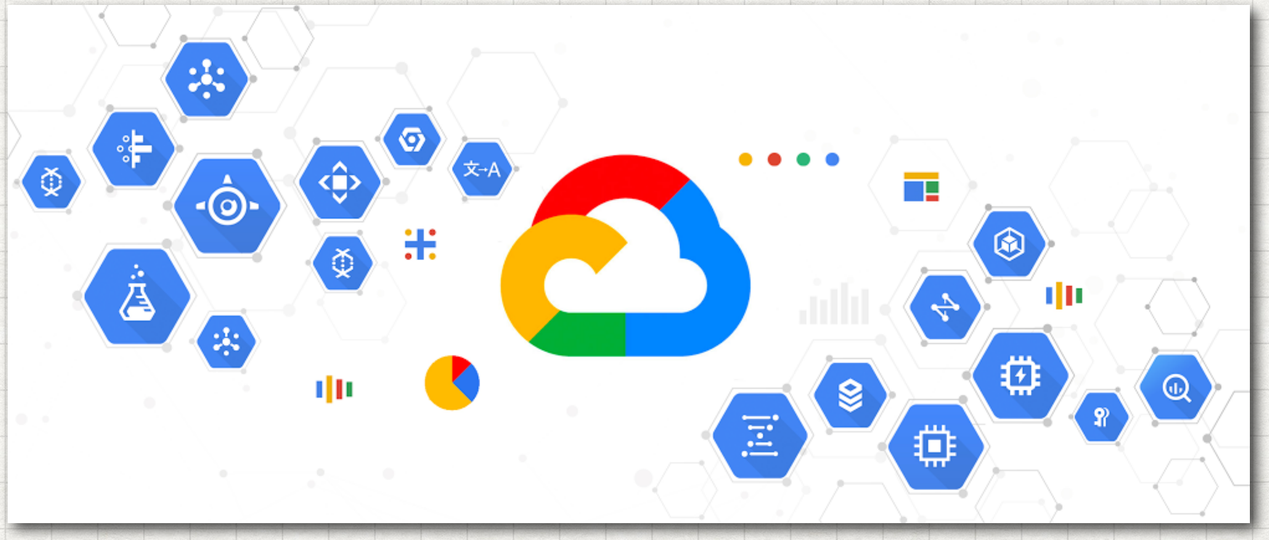


# FACTOR 3: HARDWARE

## SPEED AND STORAGE

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

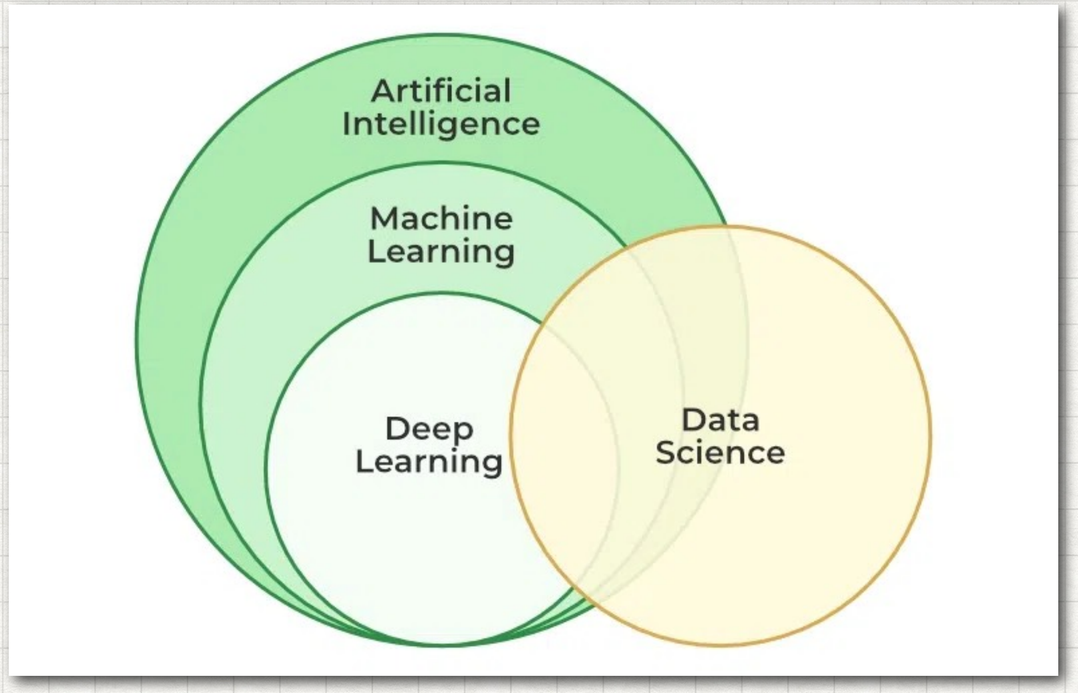




# MACHINE LEARNING

## INTRINSIC PART OF THE AI UNIVERSE

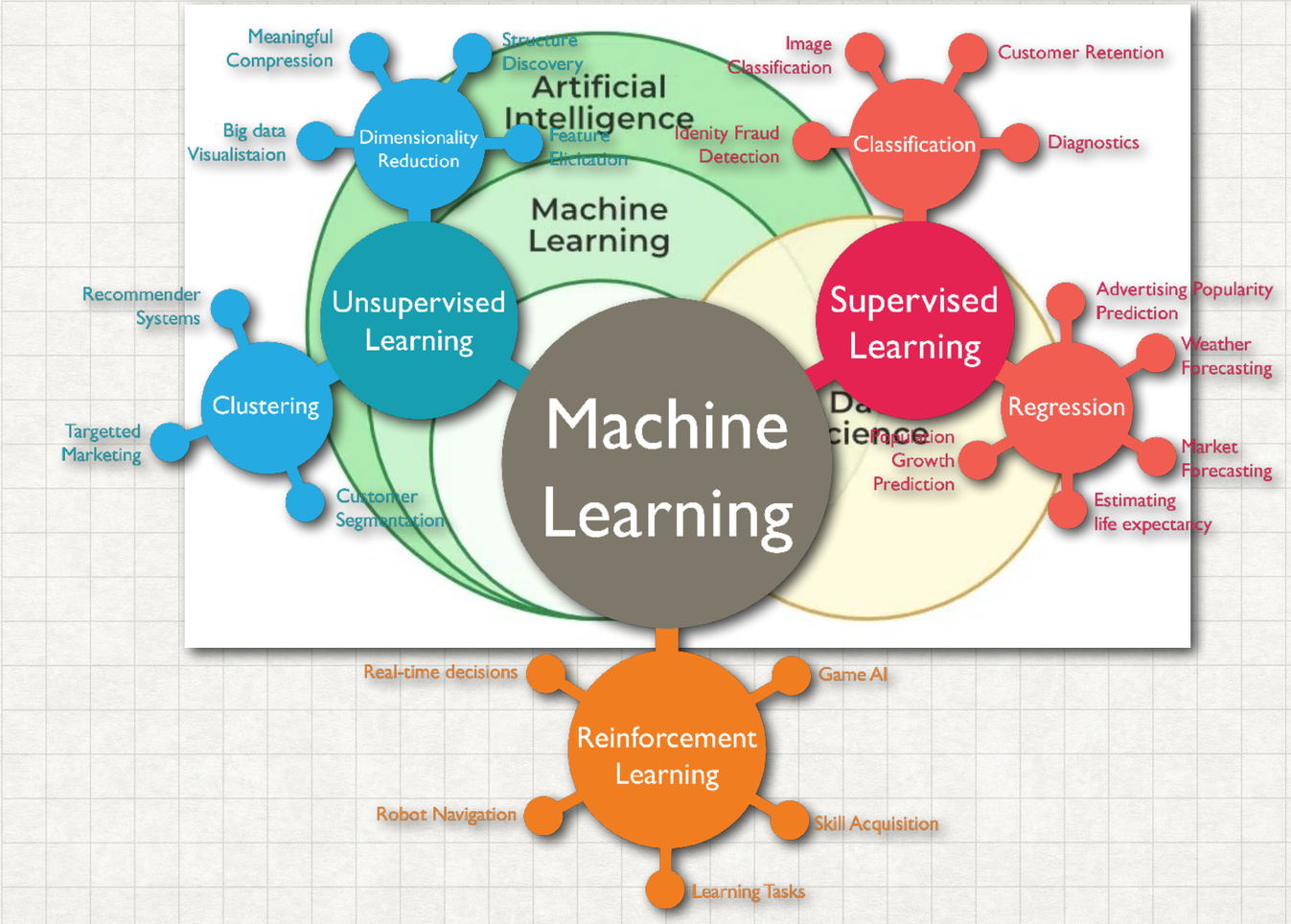
- Supervised
- Unsupervised
- Reinforcement
- Self-supervised = LLM



# MACHINE LEARNING

## INTRINSIC PART OF THE AI UNIVERSE

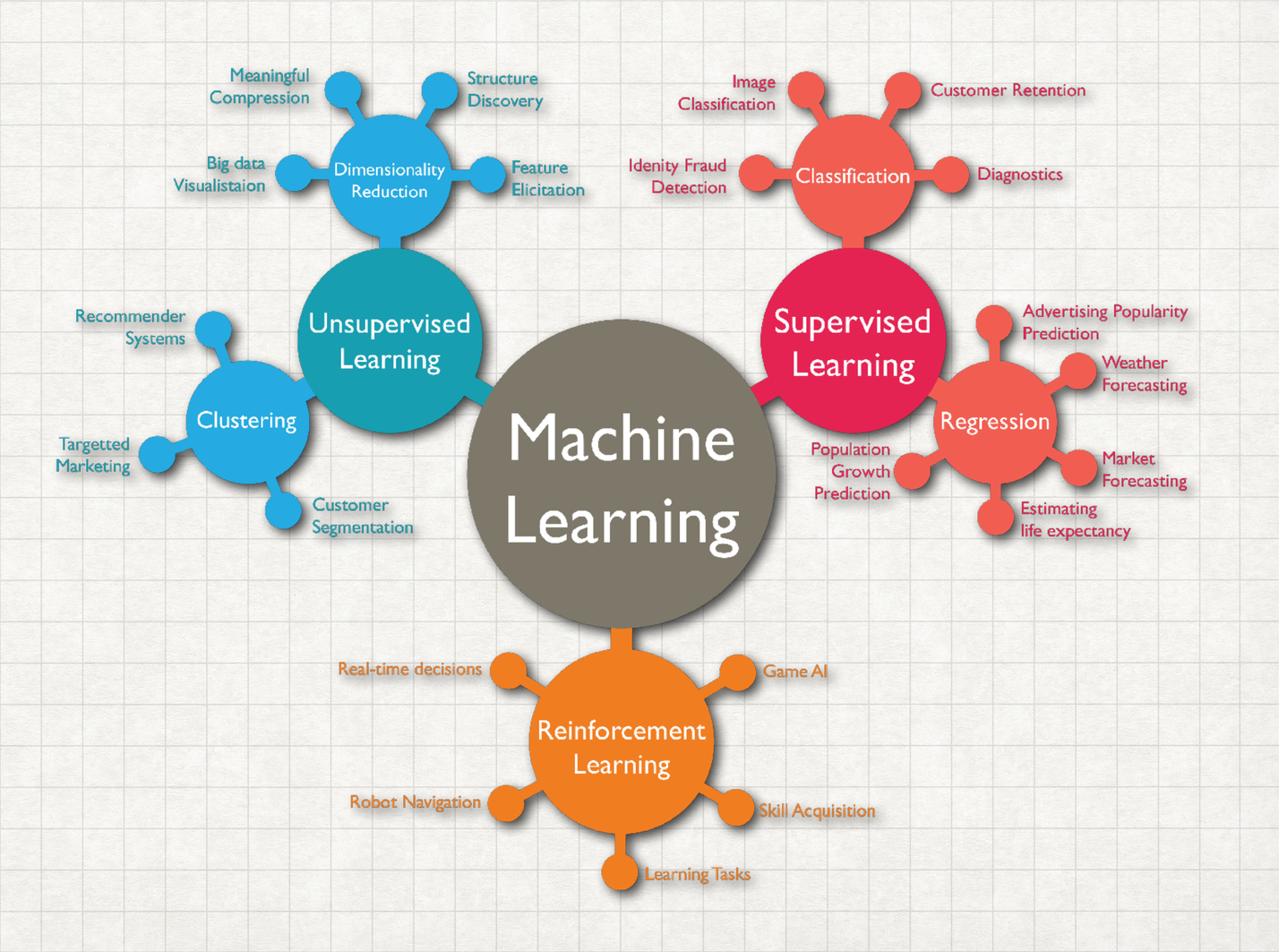
- Supervised
- Unsupervised
- Reinforcement
- Self-supervised = LLM



# MACHINE LEARNING

## INTRINSIC PART OF THE AI UNIVERSE

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



# APPLICATIONS AND USE CASES IN ENVIRONMENT

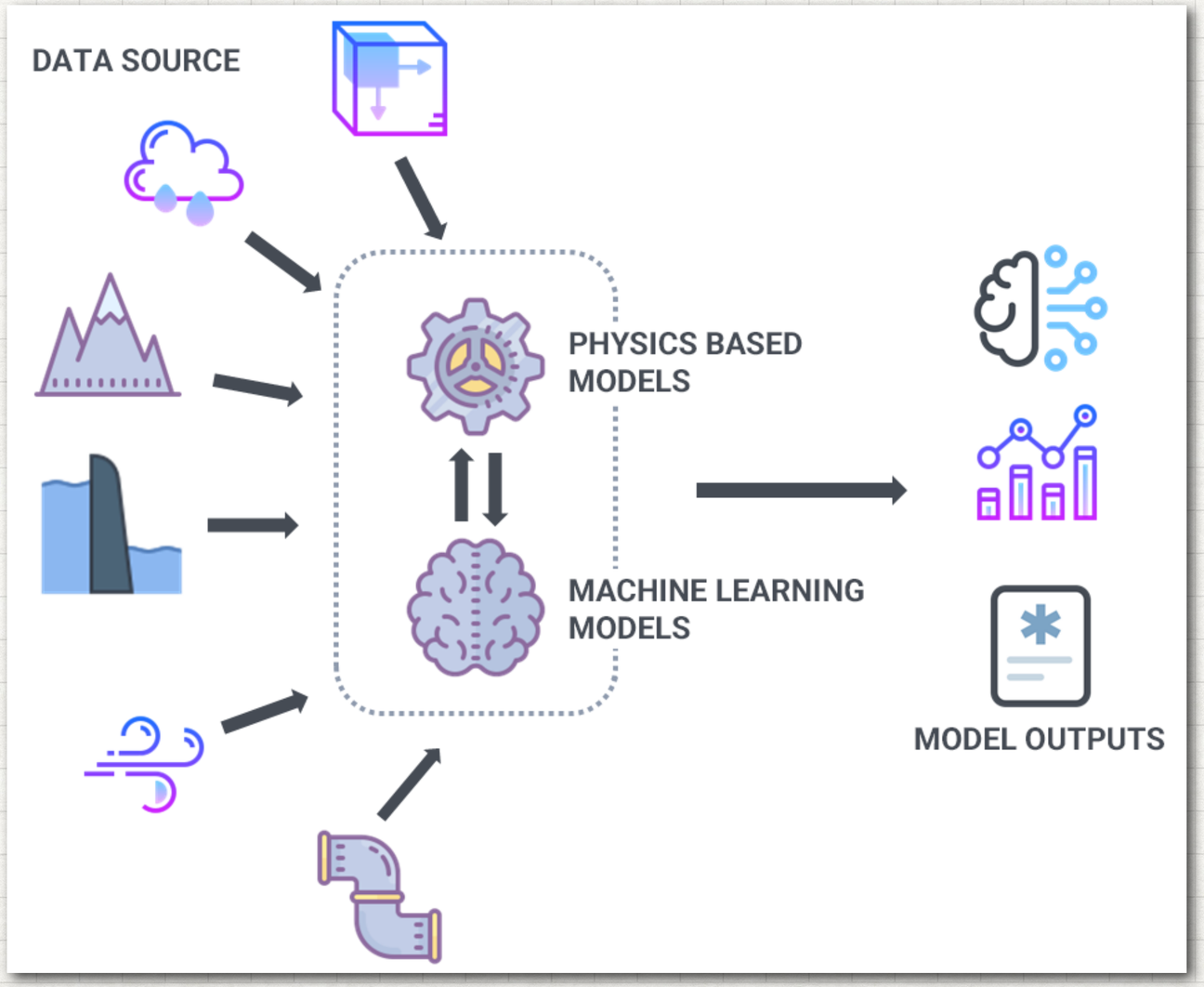
- Climate Science and Weather: Machine learning algorithms can be used for improved weather forecasting and extreme event prediction, climate model downscaling to regional scales, attribution of extreme weather events to climate change, gap-filling in historical climate records, bias correction in climate projections.
- Ecology and Biodiversity: Big data analytics can help automate species identification from images, audio, and video; perform species distribution modeling and habitat suitability analysis, monitor biodiversity from satellite imagery, study animal migration pattern analysis, predict invasive species spread
- Water Resources: Machine learning can accelerate flood prediction and management, Water quality monitoring and contamination detection, Groundwater level forecasting, Drought early warning systems, Optimal reservoir operation
- Agriculture and Food Systems: Big data analytics can help precision agriculture and crop yield prediction, Pest and disease detection, Soil health assessment, Irrigation optimization, Food supply chain efficiency improvements
- Pollution and Air Quality: Machine learning algorithms can perform renewable energy production forecasting, grid optimization and demand response, building energy efficiency optimization, energy consumption pattern analysis, electric vehicle charging infrastructure planning
- Natural Hazards: ML can provide landslide susceptibility mapping, wildfire risk prediction and spread modeling, earthquake aftershock prediction, tsunami early warning, multi-hazard risk assessment
- Circular Economy: Machine learning algorithms can automate waste classification for recycling, material flow analysis in industrial systems, product lifecycle assessment, urban mining and e-waste management, sustainable material design
- Remote Sensing Applications: Big data analytics can provide real-time insights into land use/land cover classification, forest biomass estimation, crop health monitoring, urban heat island mapping, sea level rise impact assessment

# 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.

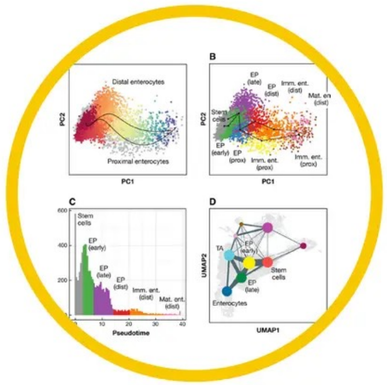


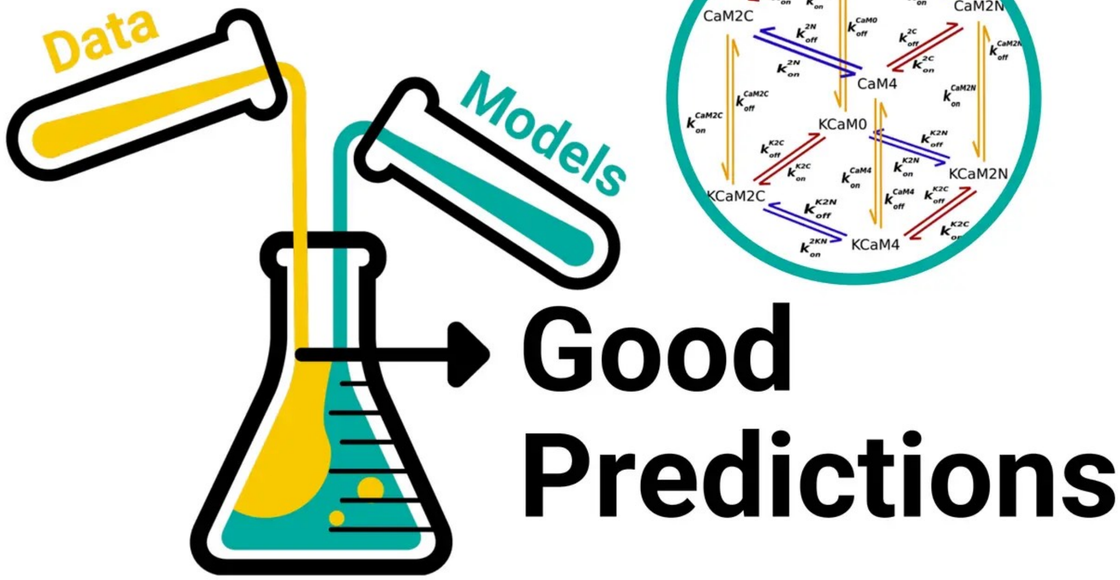
# WHAT IS SCIENTIFIC ML?

## BLENDING

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

How do we simultaneously use both sources of knowledge?







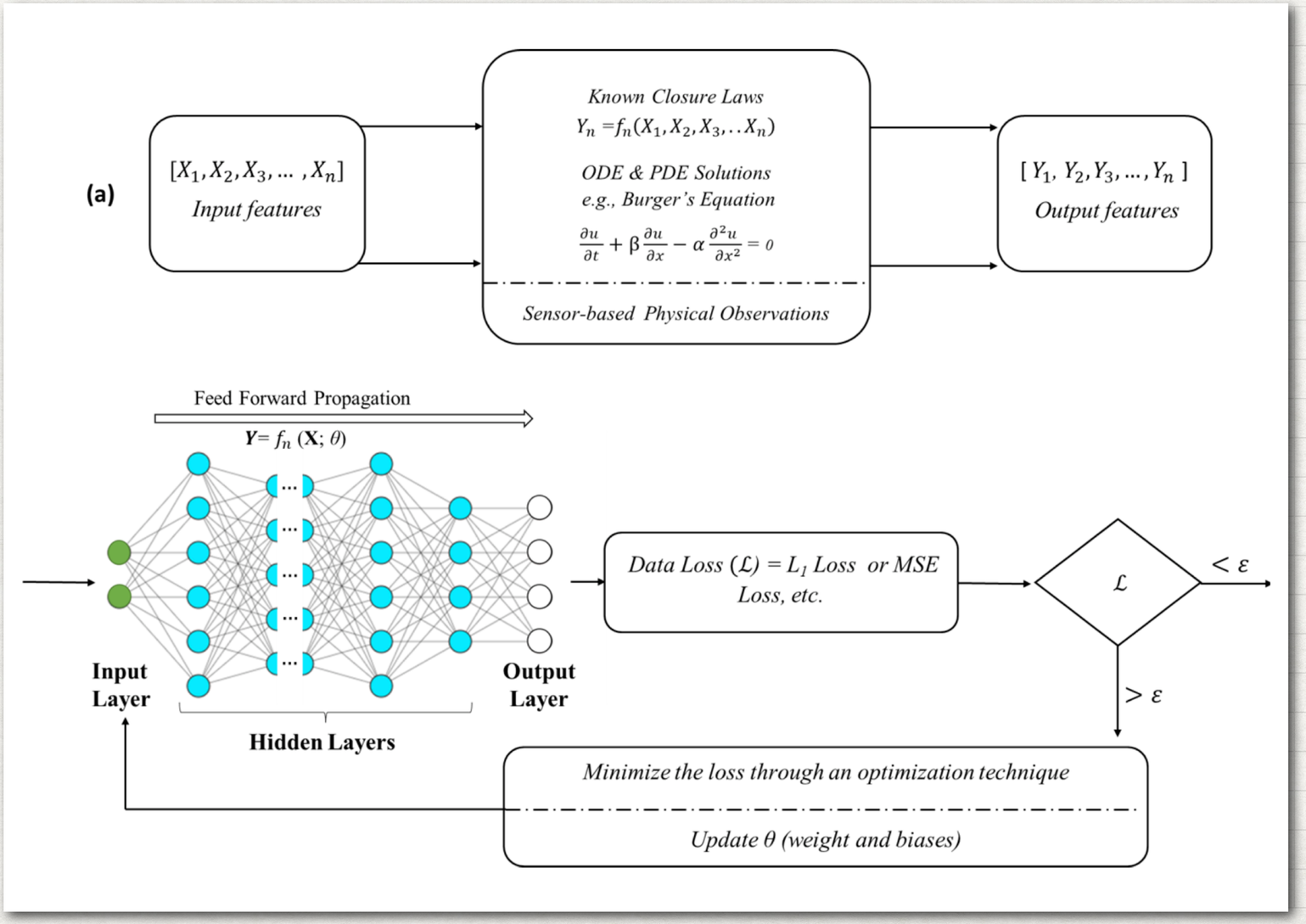
[C. Rackaukas]

# HOW IS SCIENTIFIC ML DONE?

## BLENDING

3 possible paths:

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

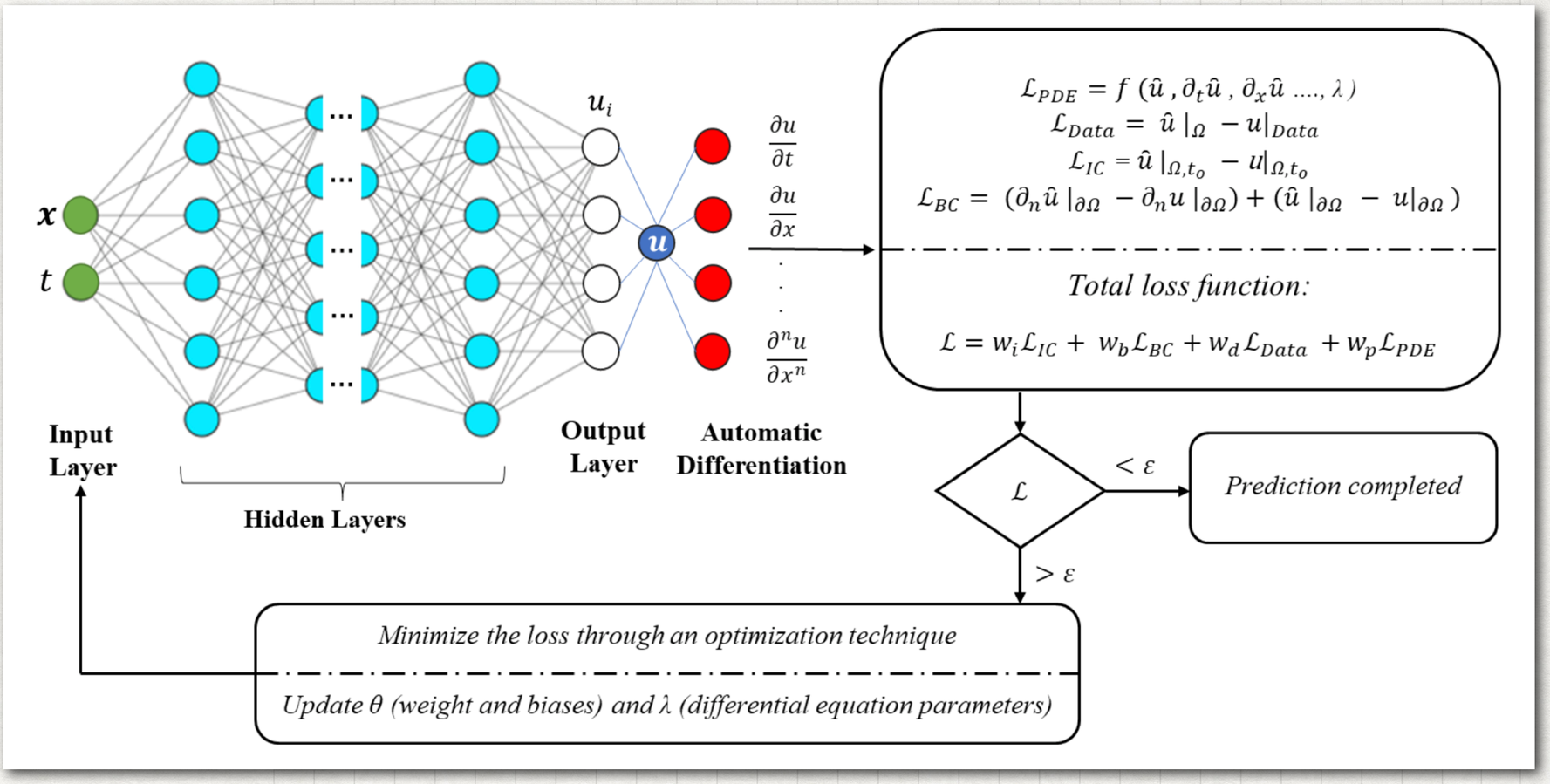


# HOW IS SCIENTIFIC ML DONE?

## BLENDING

3 possible paths:

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

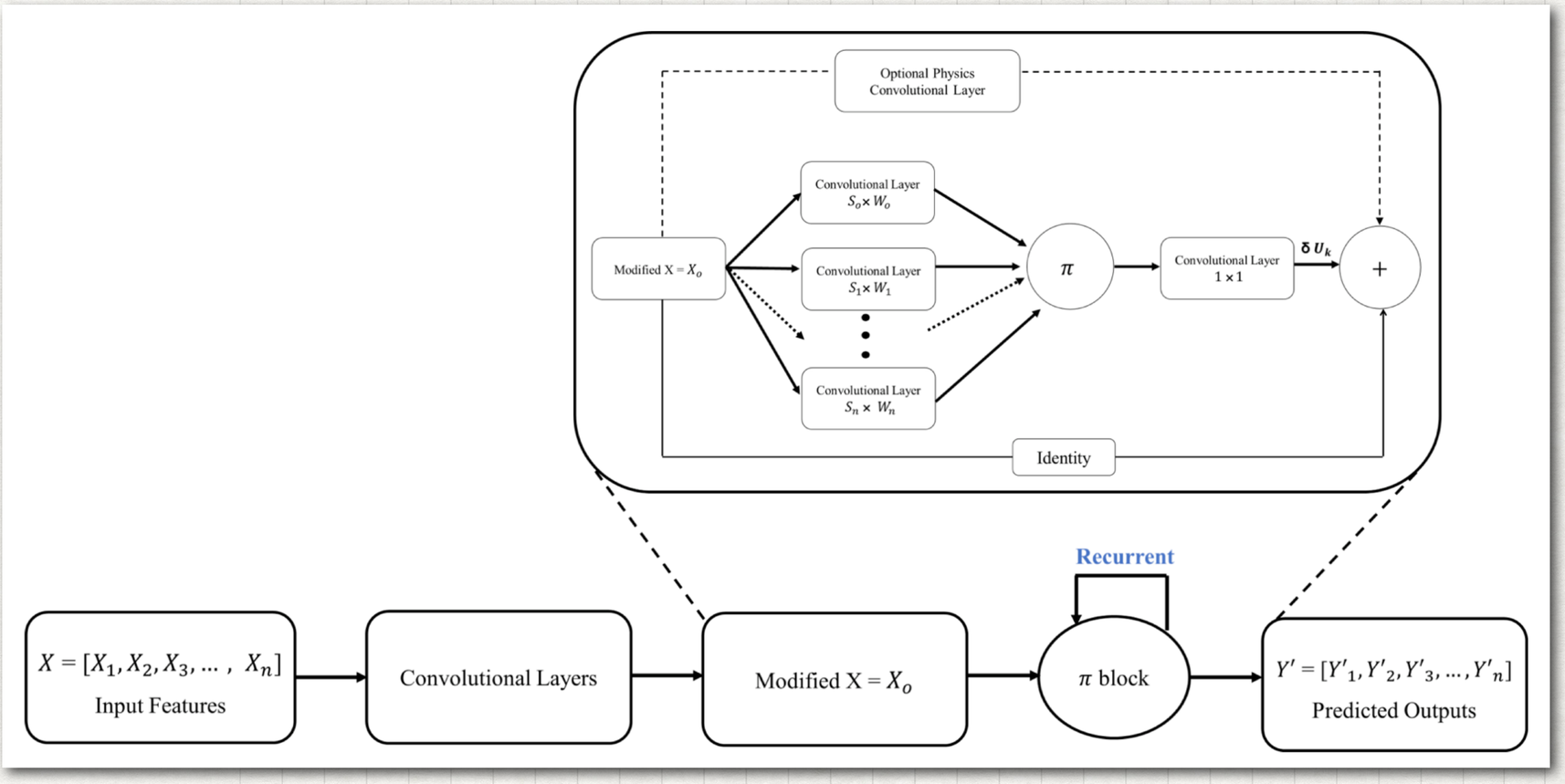


# HOW IS SCIENTIFIC ML DONE?

## BLENDING

3 possible paths:

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



# THE MATH BEHIND SCI-ML

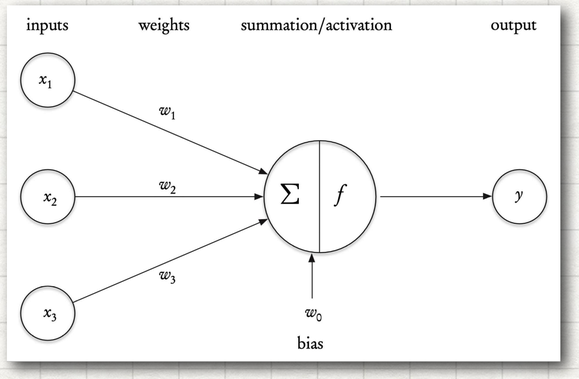
## APPROXIMATION THEORY

- Multi-layer perceptrons - 1950’s - the basis.
- Universal Approximation Property - 1990’s - the theory.
- *Differentiable programming - 2020’s - makes it all possible! (see next slide)*

# THE MATH BEHIND SCI-ML

## APPROXIMATION THEORY

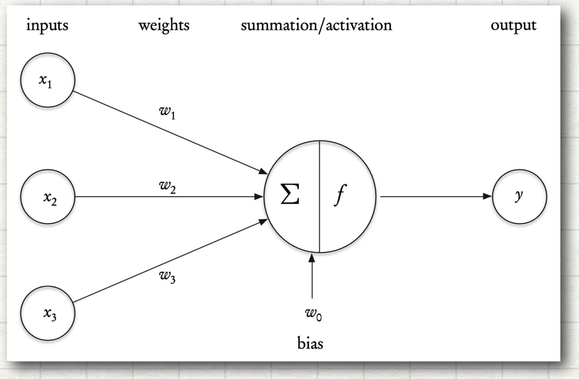
- Multi-layer perceptrons - 1950's - the basis.
- Universal Approximation Property - 1990's - the theory.
- *Differentiable programming - 2020's - makes it all possible! (see next slide)*

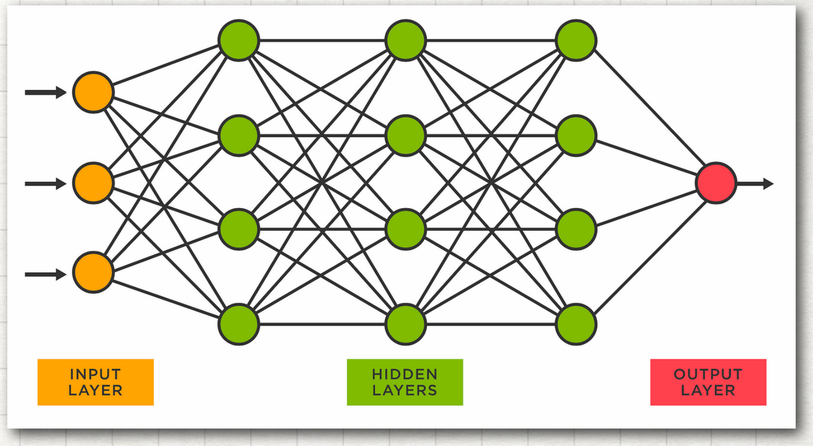


# THE MATH BEHIND SCI-ML

## APPROXIMATION THEORY

- Multi-layer perceptrons - 1950's - the basis.
- Universal Approximation Property - 1990's - the theory.
- *Differentiable programming - 2020's - makes it all possible! (see next slide)*

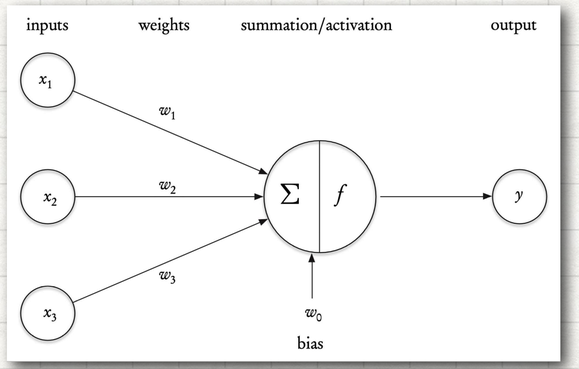


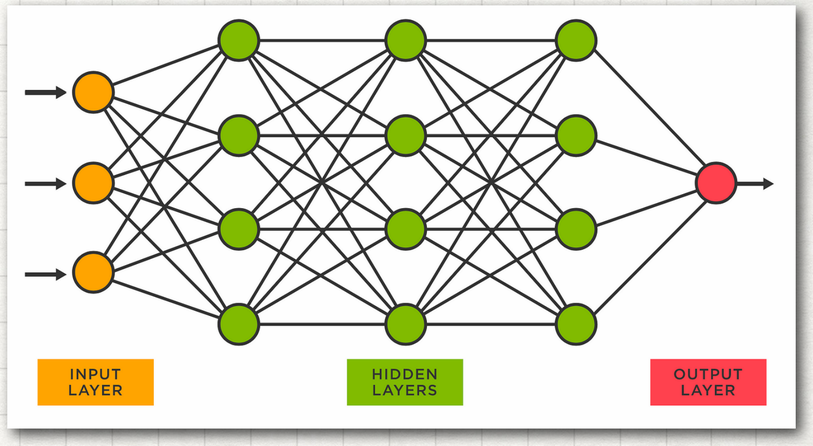
# THE MATH BEHIND SCI-ML

## APPROXIMATION THEORY

- Multi-layer perceptrons - 1950's - the basis.
- Universal Approximation Property - 1990's - the theory.
- *Differentiable programming - 2020's - makes it all possible! (see next slide)*



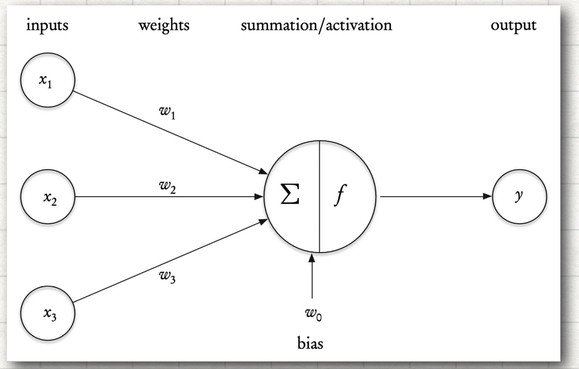

$$y = f\left(w_0 + \sum_{i=1}^{3} w_i x_i\right)$$



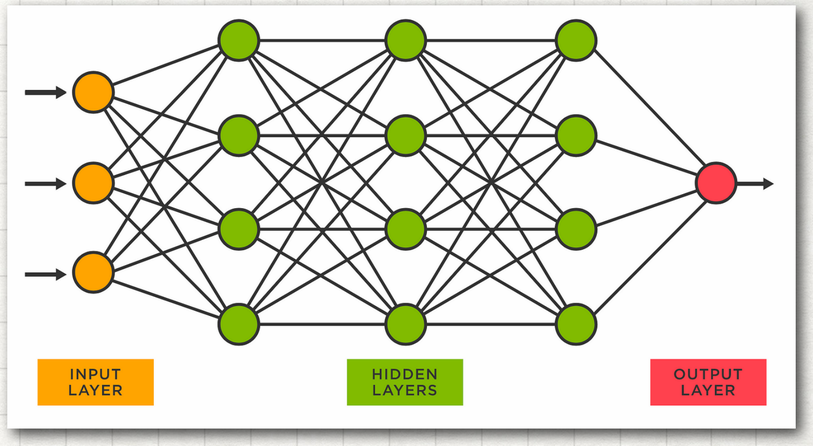
# THE MATH BEHIND SCI-ML

## APPROXIMATION THEORY

- Multi-layer perceptrons - 1950's - the basis.
- Universal Approximation Property - 1990's - the theory.
- *Differentiable programming - 2020's - makes it all possible! (see next slide)*




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




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

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

*are dense in $C(I_d)$.*

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

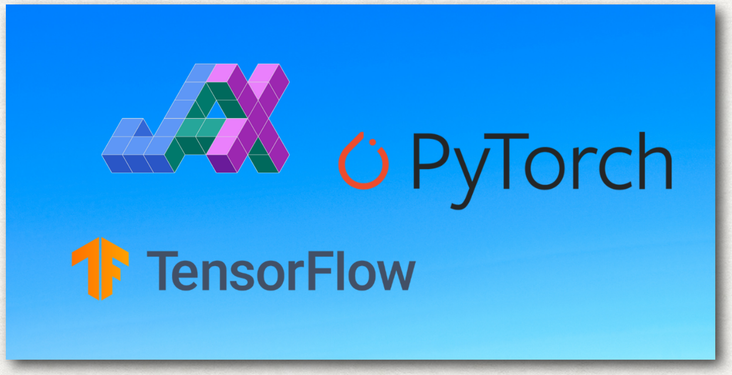
$$\mathcal{M}(\sigma) = \operatorname{span}\left\{\sigma(\mathbf{w} \cdot \mathbf{x} + b) \colon \mathbf{w} \in \mathbb{R}^d,\, b \in \mathbb{R}\right\},$$

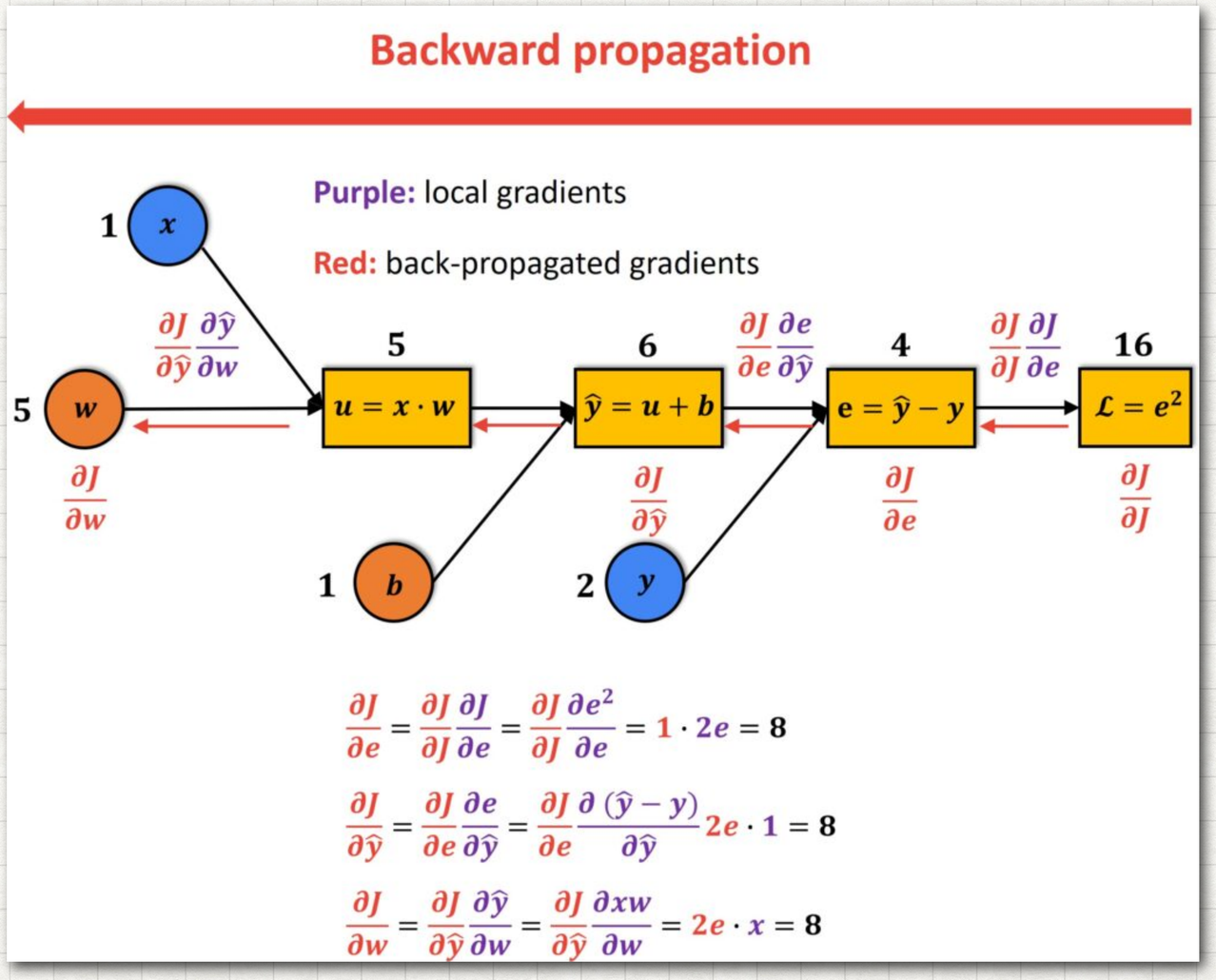
*is dense in $C^{\mathbf{m}^1,\ldots,\mathbf{m}^s}(\mathbb{R}^d) \doteq \cap_{i=1}^{s} C^{\mathbf{m}^i}(\mathbb{R}^d)$.*

# THE CODE BEHIND SCI-ML

## "AUTOGRAD"

- Differentiable programming - 2020's - makes it all possible!
- Compute the gradient of loss J wrt. weights w and minimize J based on:
  - Computational graphs
  - Chain-rule for differentiation



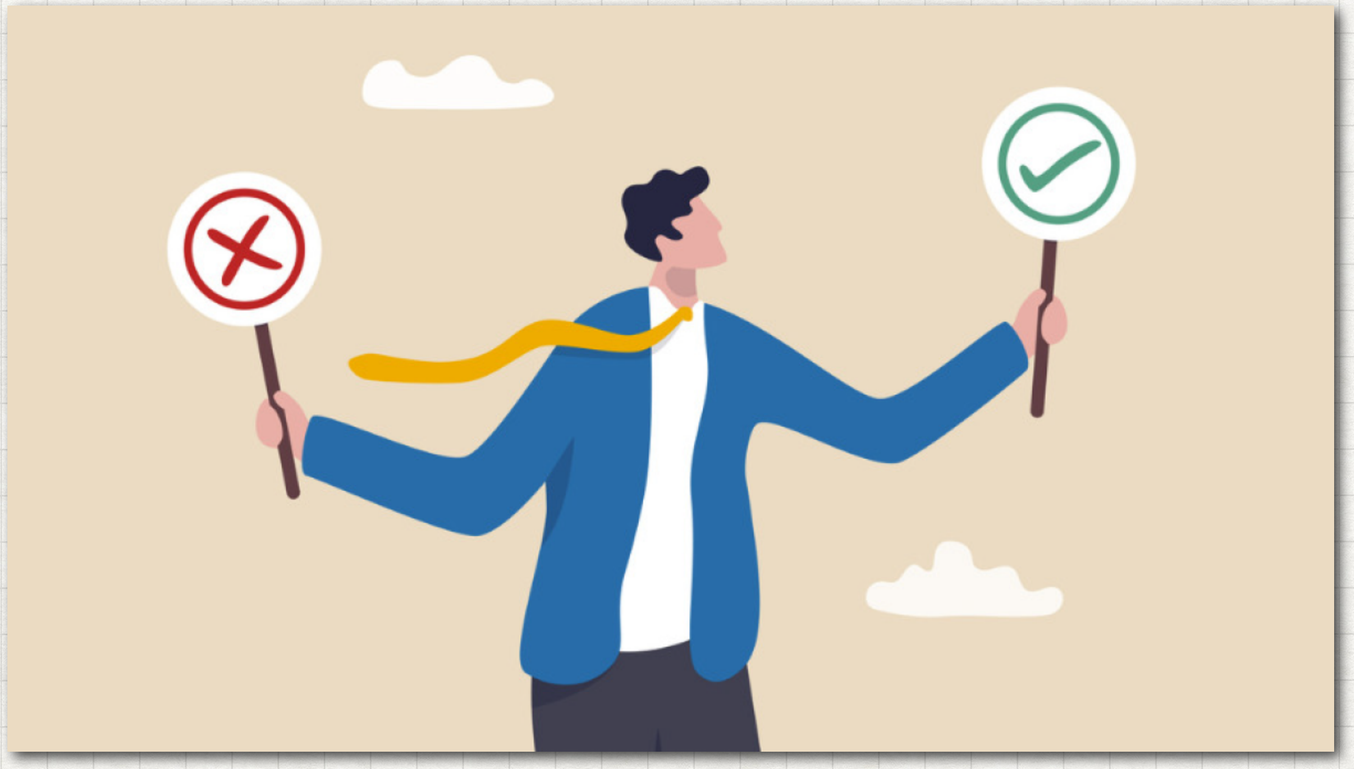



# ETHICAL CONSIDERATIONS

# ETHICS, BIAS, RESPONSIBILITY, TRUST

## DEFINITIONS

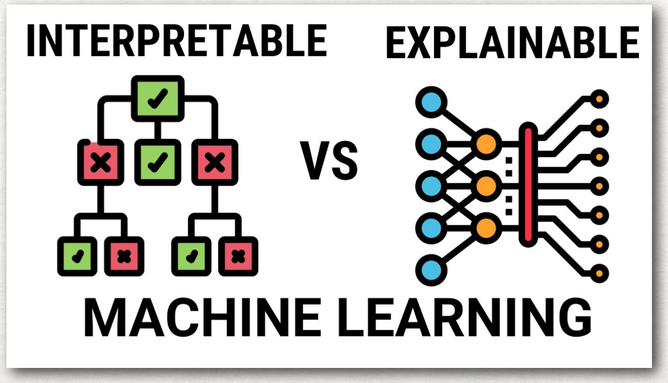
- ETHICS = what is morally good or bad, right or wrong? (norms)
- BIAS = prejudice against a person, object, position.
- TRUST = willingness to assume risk by relying on, or believing in, the actions of another party
- TRUSTWORTHY AI should be lawful, ethical, unbiased.

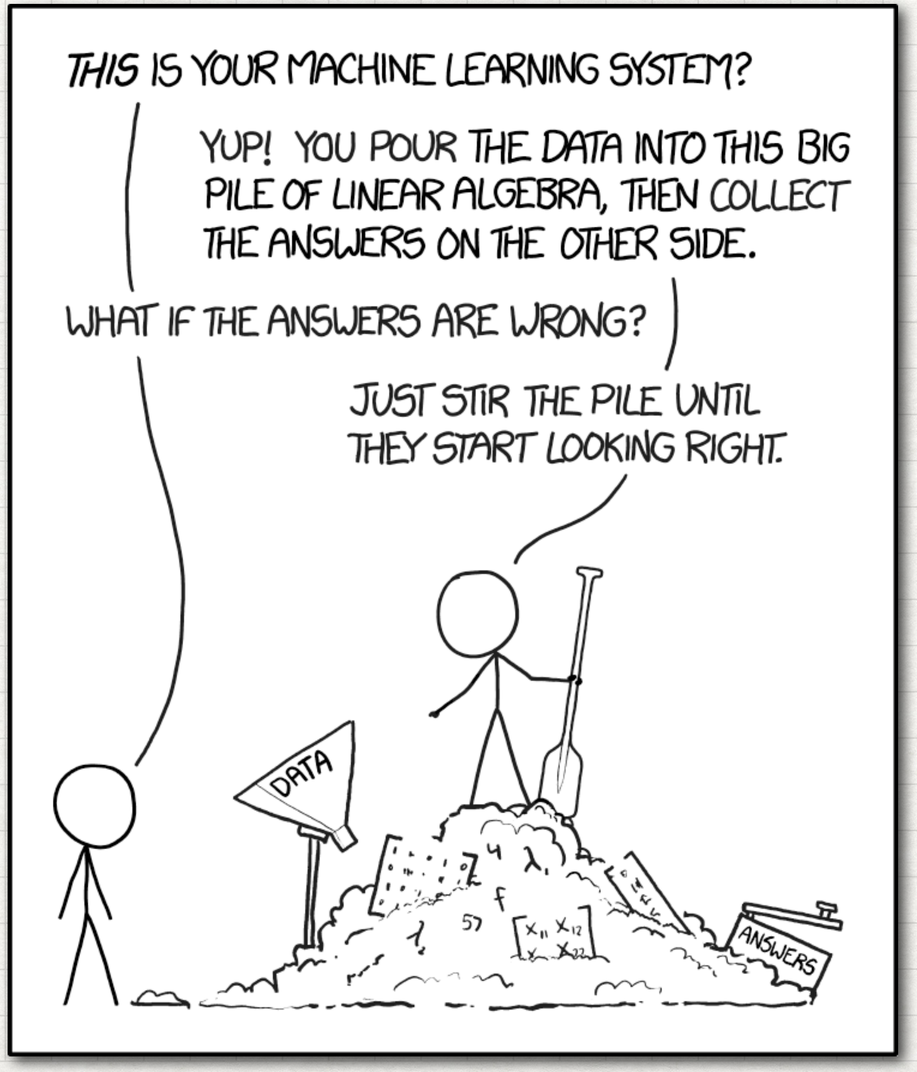


# EXPLAINABILTY & INTERPRETABILITY

## 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

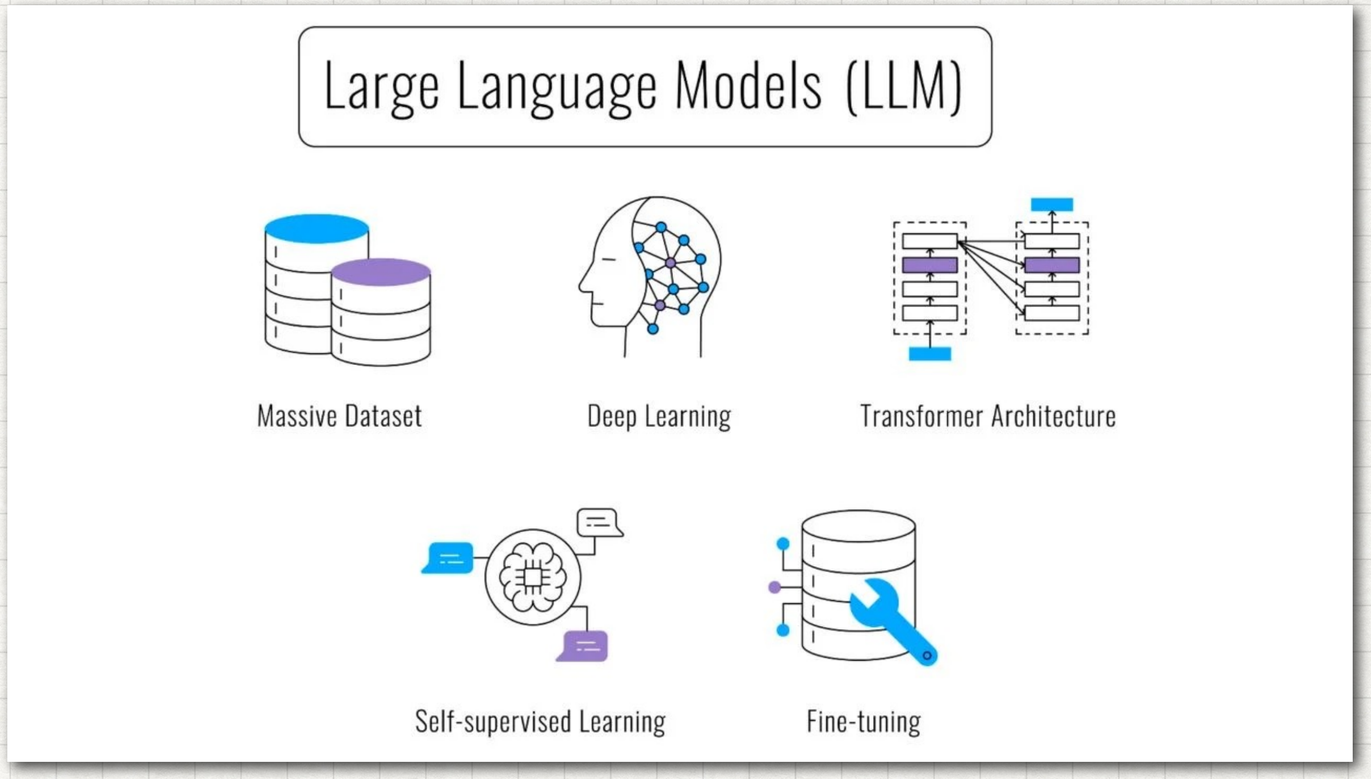
- Training data issues:
  - Non-representative, lack of geodiversity.
  - Faulty, biased training labels.
  - Adversarial effects.
- AI model issues:
  - Learn faulty strategies.
  - Fake something plausible (LLMs!).
  - Non-trustworthy, lack of robustness.
- Societal issues:
  - Lack of consent on data 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

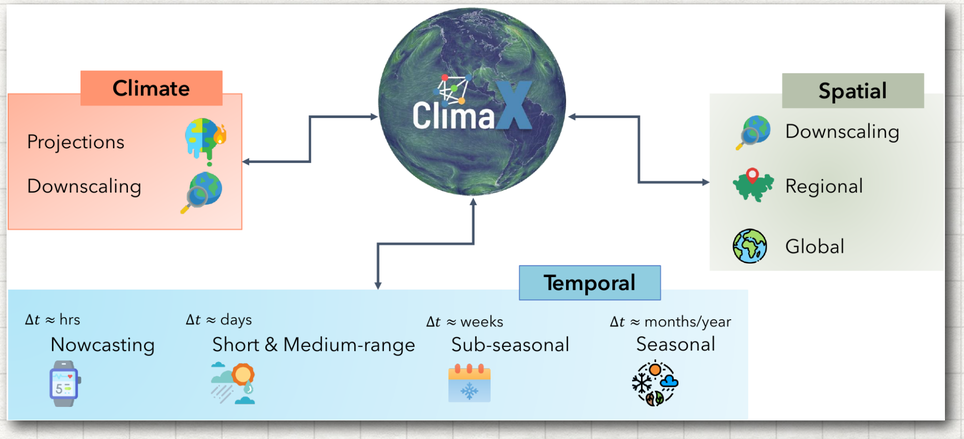
- Economic
  - Job transformation
  - Productivity gains
- Social and Ethical
  - Bias and fairness
  - Privacy concerns
- Educational
  - Personalize learning
  - Critical thinking
- Healthcare
  - Improved diagnostics
  - Patient-health system interactions

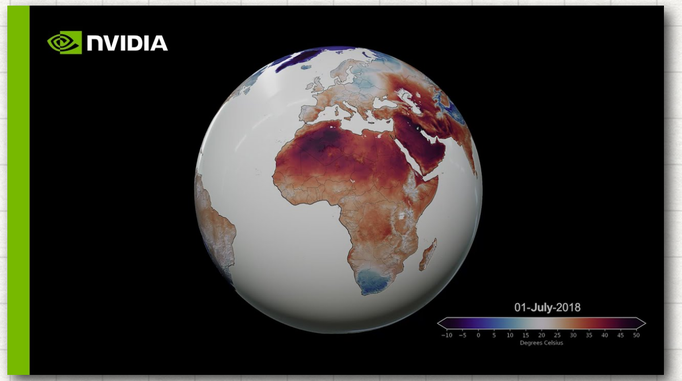
# LLM APPLICATIONS

# NWP & CLIMATOLOGY

## USING LLM'S

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





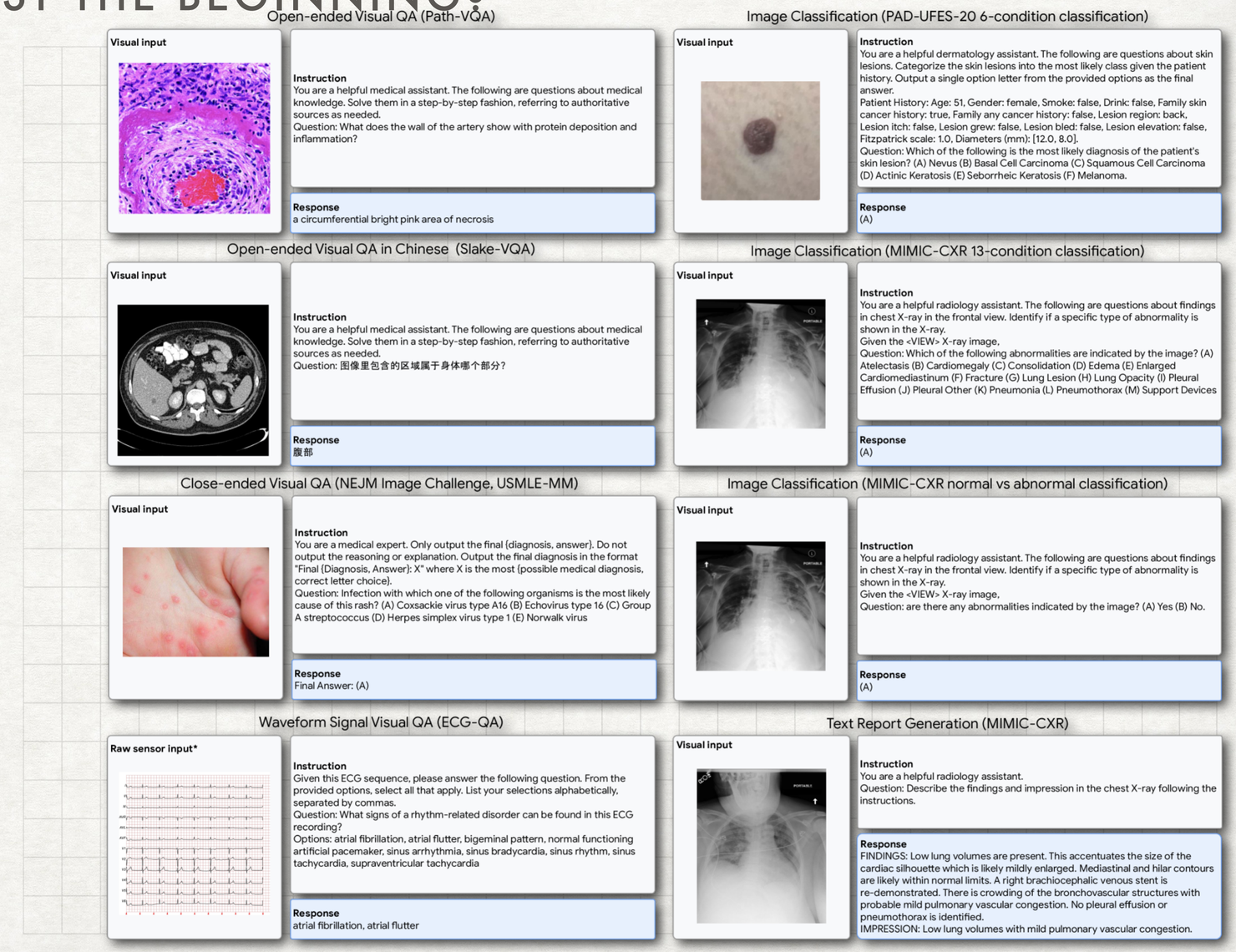
Generative AI to quantify uncertainty in weather forecasting

March 29, 2024 · Lizao (Larry) Li, Software Engineer, and Rob Carver, Research Scientist, Google Research

# LLM FOR HEALTHCARE

## JUST THE BEGINNING?

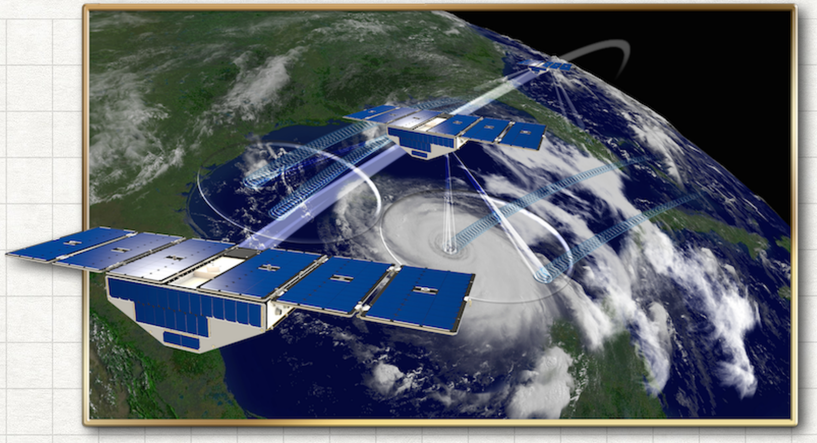
- Google announces Med-Gemini https://arxiv.org/pdf/2404.18416v2 (29 April 2024)
  - Family of multimodal models built upon Gemini specifically designed for the healthcare industry.
  - “Groundbreaking Family of AI Models Revolutionizing Medical Diagnosis and Clinical Reasoning”
  - Models excel in multimodal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

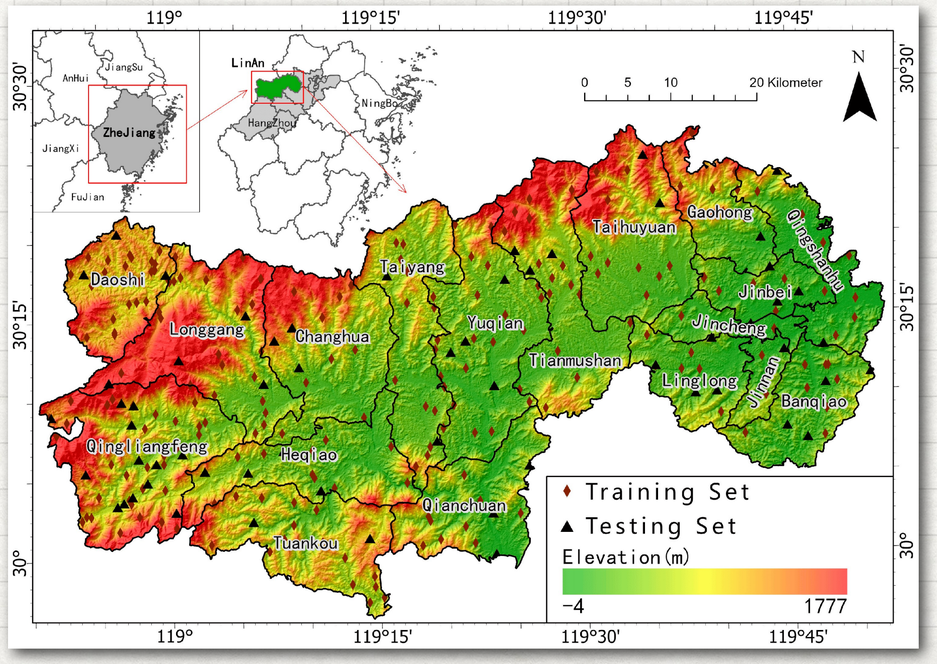


# LLMS FOR ENVIRONMENT

## COMING SOON?

- Climate and weather forecasting—see previous slide.
- Environmental monitoring from satellite imagery databases.
- Policy analysis.
- Biodiversity research.
- Sustainability reporting.
- Landslide susceptibility analysis (will we all be out of a job soon?)

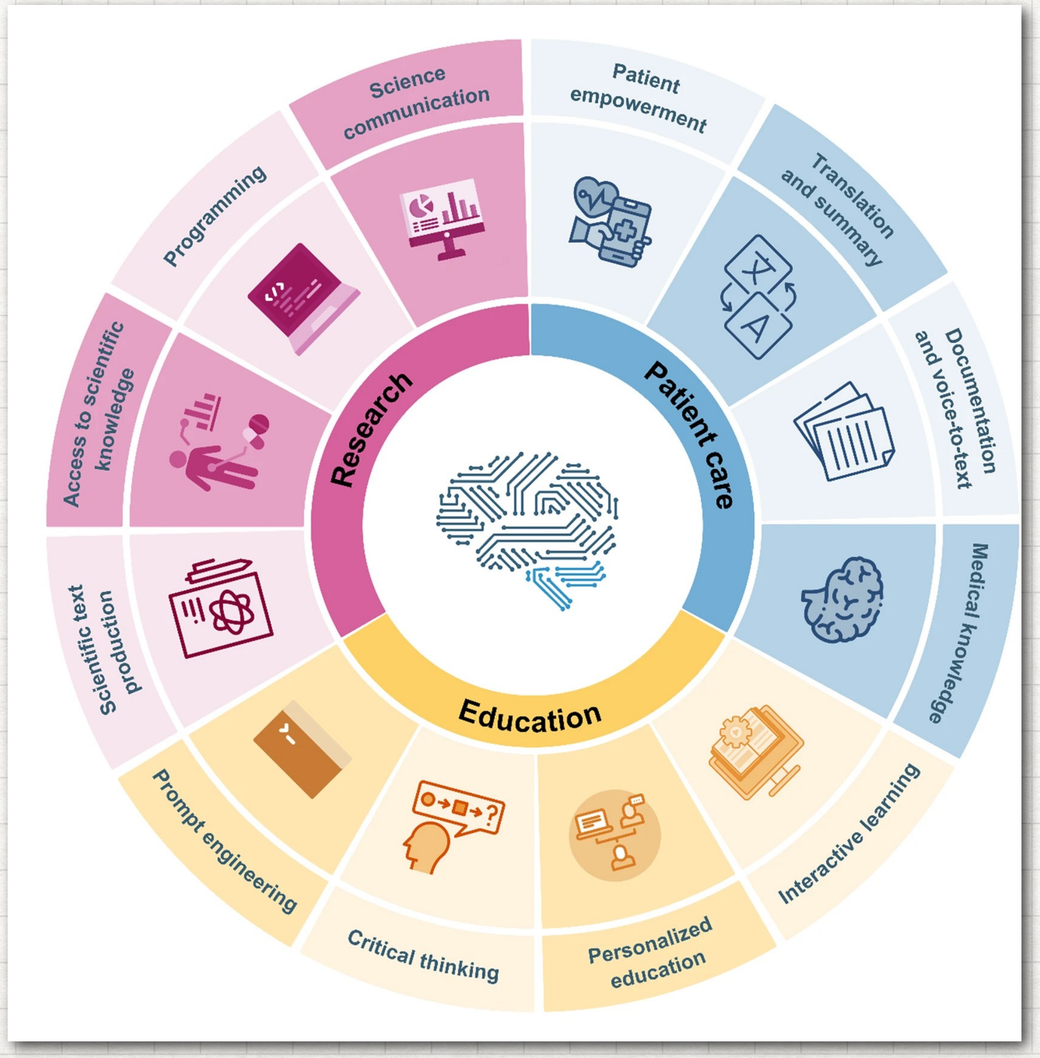




# LARGE LANGUAGE MODELS

## 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?



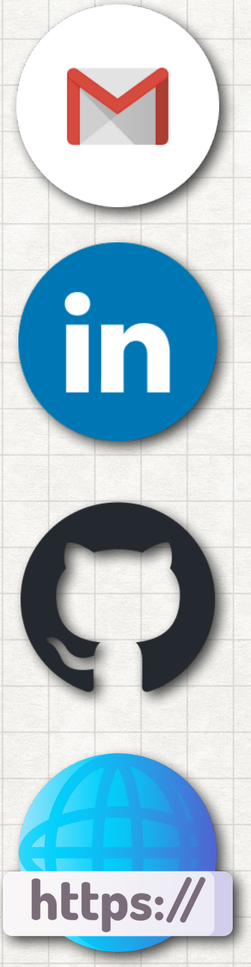

[J. Clusmann, et al. Nature Comm. Medecine, 2023]

# 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 differentitation in machine learning: a survey. Journal of Machine Learning Research, 18 (2017), article 153.