

# Foundation Models and Patterns for Science Time Series

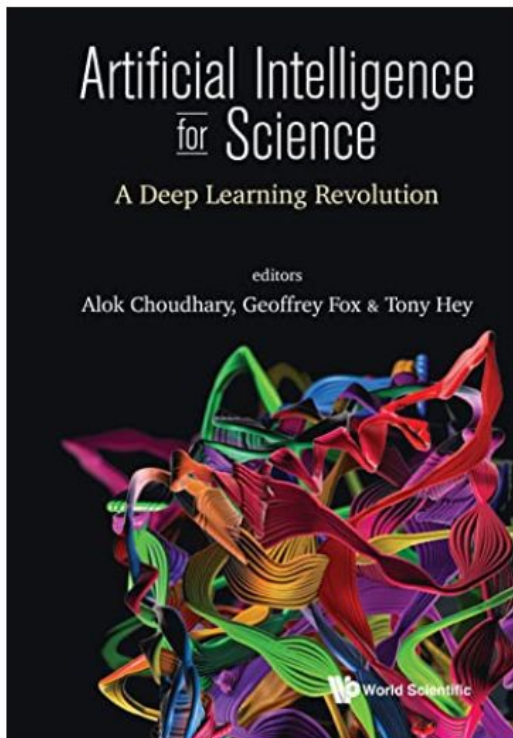
July 15, 2024

IEEE SMC-IT/SCC 2024

IEEE Space Mission Challenges for Information Technology - IEEE Space Computing Conference

Computer History Museum • Mountain View, CA, USA • 15-19 July 2024

**Geoffrey Fox, Biocomplexity Institute and Computer Science Department. University of Virginia**



# Artificial Intelligence for Science: A Deep Learning Revolution



Kindle Edition

by [Alok Choudhary](#) (Editor), [Geoffrey Fox](#) (Editor), [Tony Hey](#) (Editor) | Format: Kindle Edition

5.0 ★★★★★ 1 rating

[See all formats and editions](#)

**Kindle**  
**\$37.49**

Hardcover  
\$148.00

Read with Our **Free App**

2 Used from \$158.12  
8 New from \$144.00

This unique collection introduces AI, Machine Learning (ML), and deep neural network technologies leading to scientific discovery from the datasets generated both by supercomputer simulation and by modern experimental facilities.

Huge quantities of experimental data come from many sources — telescopes, satellites, gene sequencers, accelerators, and electron microscopes, including international facilities such as the Large Hadron Collider (LHC) at CERN in Geneva and the ITER Tokamak in France. These sources generate many petabytes moving to exabytes of data per year. Extracting scientific insights from these data is a major challenge for scientists, for whom the latest AI developments will be essential.

[Read more](#)

[Read sample](#)

## Follow the Author



**Alok N. Choudhary**

[Follow](#)

ISBN-13



978-9811265662

Sticky notes



[On Kindle Scribe](#)

Publisher



**World Scientific Publishing Company**

Publication date



March 21, 2023

Language



English

File size



78220 KB

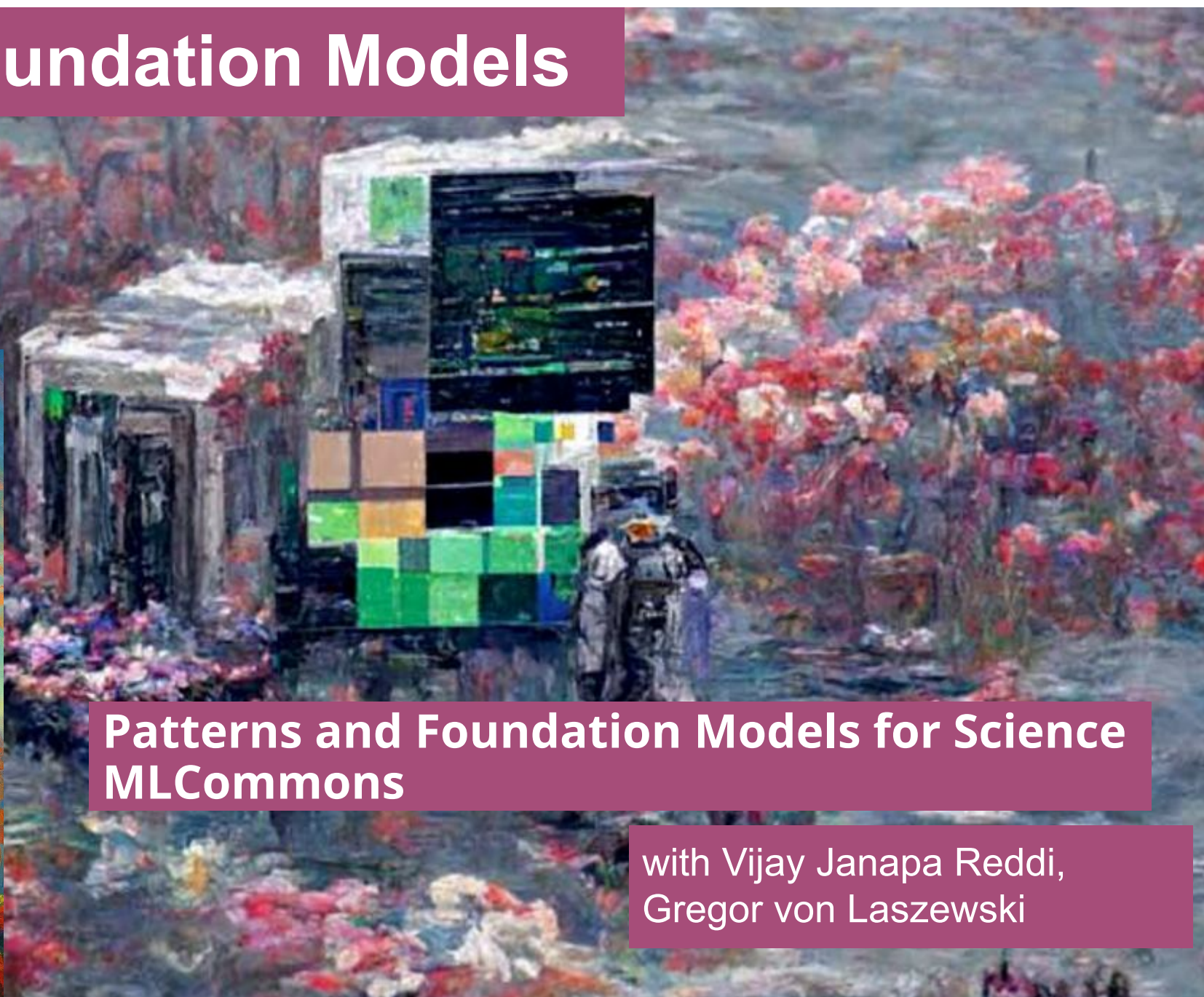


<https://doi.org/10.1142/13123>

**40 Chapters**

**95% of AI Discussed is Deep Learning**

# AI for Science Foundation Models

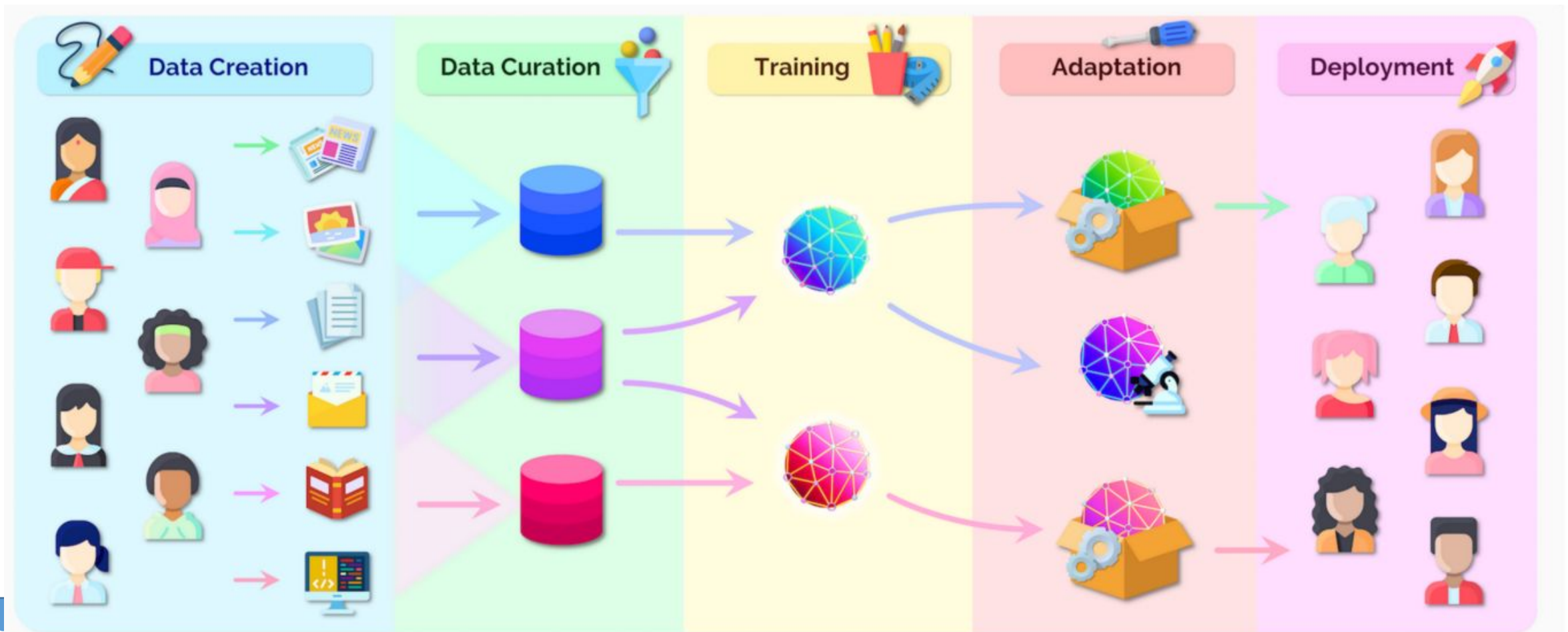


Patterns and Foundation Models for Science  
MLCommons

with Vijay Janapa Reddi,  
Gregor von Laszewski

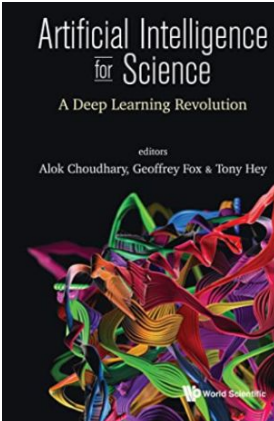
# Big and Foundation Models

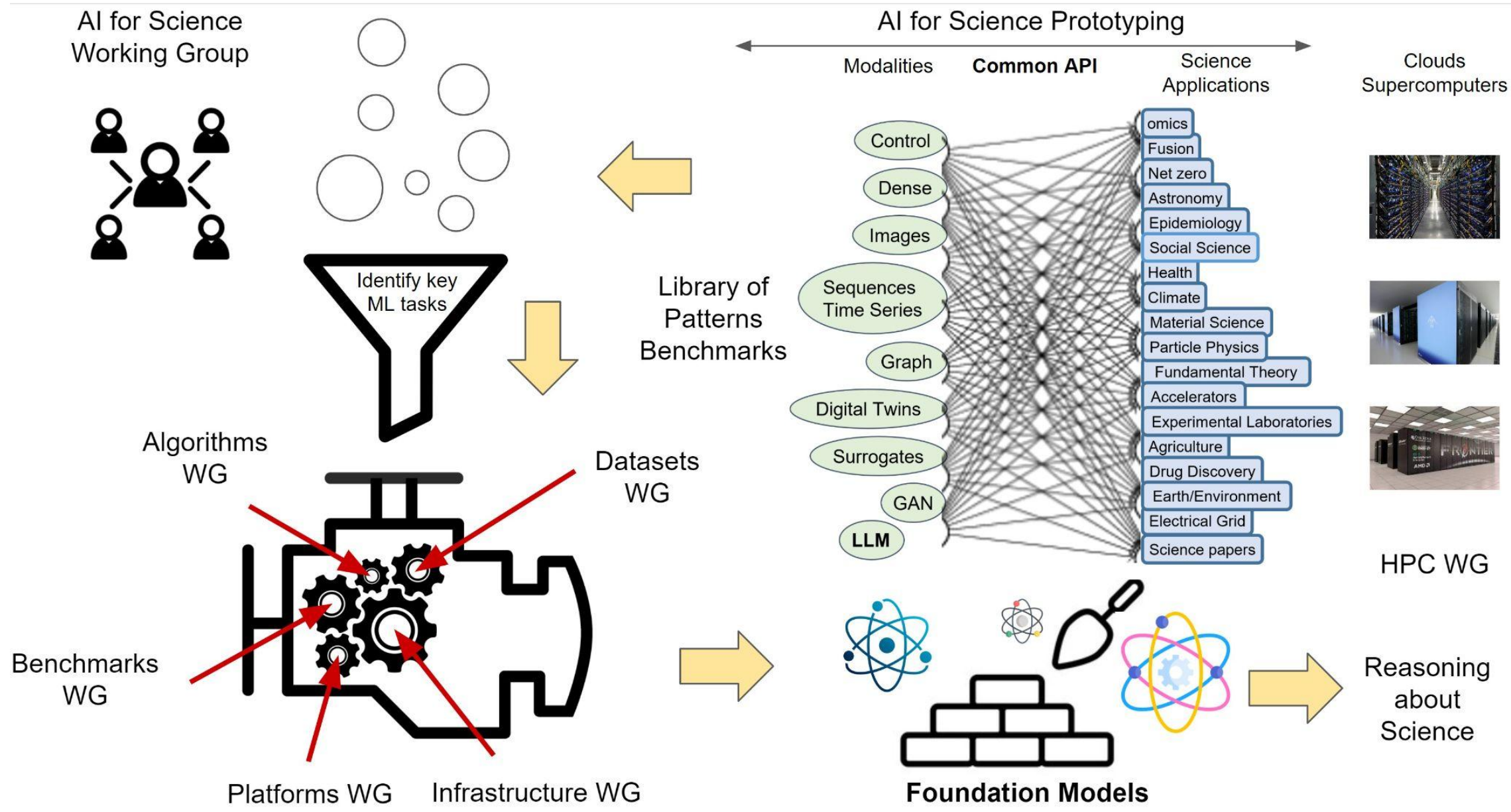
- Transfer Learning builds AI model in one area and then uses it with modest additional training in another area
- Foundation Models follow this to an extreme and train on so much data that model can generalize to cover “everything”



# AI Transforming Science and Engineering

- 95% of research described in book is deep learning based, and future will be more and not less complicated
- **Comparing science with industry AI applications, one finds a much richer set of data and data modalities in science**
- However (by analyzing 40 chapters in book), one can find many similar **patterns** so that approaches can be re-used across many different application domains
- **Strategy 1:** Each domain/research group develops individual AI applications informed by good ideas discovered in other communities
- **Strategy 2:** Applications divide into **patterns** such as time series, imaged-based data, accelerator events, Monte Carlo simulations, and one develops a common AI library for each of ~10 patterns
- **Strategy 3:** Build one or a very few **Science Foundation Models** that are pretrained on core data, such as earthquake simulations or remote sensing data, and then fine-tuned on specialized problems; such as Southern California Earthquakes or identifying icebergs
- **Strategies can co-exist and enhance each other**





# Science Foundation Models in MLCommons

# Foundation Model Summary

This table summarizes Science Foundation models and some of the important studies from commodity applications or general research. <https://sciencefmhub.org/> has a more detailed table with full references of the the 481 references spread over 208 projects.

Histogram of quarters with at least one Publication, Total Articles: 2355 as of 07/14/2024

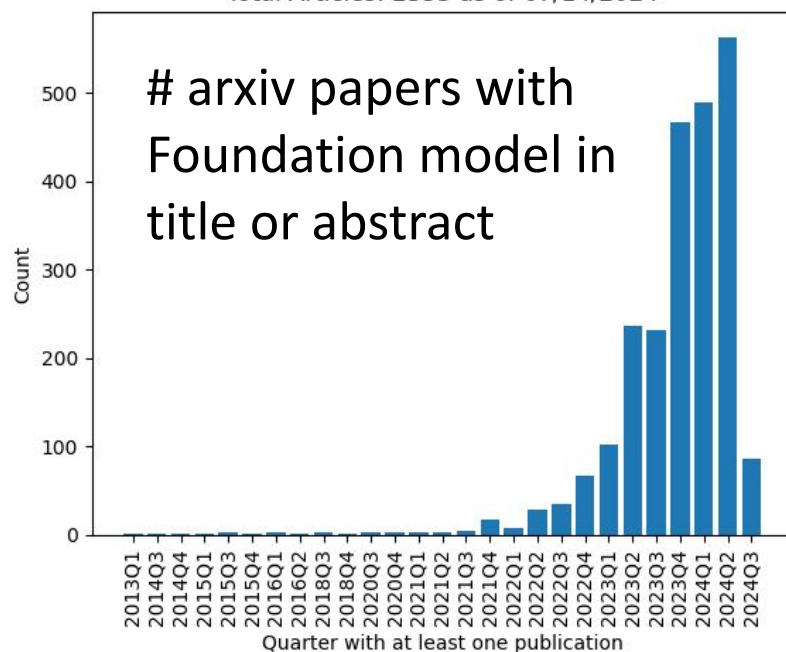


Table 1: Foundation Model Overall Summary 208 Projects 481 Refs	
<b>21 projects, 41 refs Single Modality Image/video-based Models [General]</b>	2020-23: Scaling by Google, MAE: Masked Autoencoders Meta, DINO Meta, Scaling ViT, Ibot, VideoMAE, Context Autoencoder, Segment Anything, Local Mask, Swin Transformer, Swin MAE, DETR, BYOL, ConvNeXt, SimCLR, Mask2Former, ALIGN, SEER. (ViT, Swin, ResNet, EfficientNet, RegNety) (1.2M- 4 billion images from 50 GPU days) I-JEPA, V-JEPA, VitDet
<b>11 projects, 19 refs, Single Modality Image-based models [Remote Sensing ]</b>	2022-23: RingMo, Geograph, SeCo, RVSA, SatMAE, Joint SAR-Optical, Scale-MAE, Billion-scale Remote Sensing, IBM and NASA HLS Prithvi (Swin, ResNet50, ViTAE, ViT) (23K to 2M images, from 1 GPU-day), Mission Critical, SatlasPretrain
<b>15 projects, 28 refs Multi-modal Image/video + Text [General]</b>	2019-23: CLIP, ViLBERT, LXMERT Unified Vision-Language, UniVL, FLAVA, Owl-Vit, Captions, MELTR, Florence-2 (ResNet, BERT, ViT, Transformer)(Upto 400 million image-text pair, 3.6B text annotations, ~40-10,000 GPU days), MM1, Transframer, Yi, Chameleon, DocLLM
<b>2 project 4 refs Multi-modal Image/video + Text [Remote Sensing]</b>	2022: Contrastive Hashing (ResNet18) (13K images); 2024 use CLIP on BgiEarthNet and other data
<b>3 projects 7 refs Multi-modal Image + Text + Audio and others [General]</b>	2022-23: IMAGEBIND, AudioCLIP, mPLUG-2 (ViT, ResNet, Transformer) (7 modalities, upto 14M images)
<b>1 project 2 refs Multi-modal Image + Audio [Remote Sensing]</b>	2023: audiovisual representation learning (ResNet50)(50,000 image-audio pairs)
<b>16 project, 45 refs, Single Modality Image-based models [Climate, Weather]</b>	2022-4: ClimaX, AtmoRep, FengWu-GHR, ClimSim, NowcastNet, Corrformer, ORBIT, CorrDiff, EARTH-2, FourCastNet, FouRKS, GraphCast, MetNet-3, Fuxi-DA, DiffDA, KARINA
<b>3 projects 9 refs Astronomy Single or Multimodal</b>	AstroCLIP Radio Galaxy Zoo
<b>1 project 3 refs Surrogates</b>	Physics Simulations
<b>67 projects 182 refs Time Series</b>	LLM based{Time-LLM, Lag-Llama, AutoTimes, ST-LLM (Traffic), Tempo, LLMTime, Chronos}, InstructTime, GPT-As-Classifer), TARNet, FormerTime, TimeMAE, TS-TCC, TS-GAC, CRT, TS2Vec, FEDFormer, Dlinear, SFA,TFC, LPTM, SCOTT, TimeGPT-1, GHPT, ConvTimeNet, TimesNet, BitCN, PatchTST, N-HITS, Timer, TimeXer, iTransformer, SOFTS, TCDFormer, TiDE (Dense Encoder), TimesFM(Decoder Only), SimMTM, TFT, TSMixer, MLP-Mixer, Automixer, TTM Tiny Time Mixer, SAMFormer, Ms-tct, Dual-Mixer, tsGT, MAMixer, Sparsity for Multivariate Time-Series, TimeMixer, Moirai, TSLANet, Time Evidence Fusion Network, FM4OS, RWKV-TS
<b>17 projects 35 refs Chemistry, Medicine Material Science</b>	2022-23: Uni-Mol, MoLFormer, MUBen, ChemBERTa-2, Chemformer, DeepChem, graph transformer for molecules, GenSLM for genomes, OpenFold based on AlphaFold2, AlphaFold3, Segment Anything for Medical Images, Single Cell RNA ScBERT, ScGPT, Geneformer, ScFoundation, ScSimilarity (Transformer, Graph, LLM, ViT)(100M SMILES, 6.4M genomes, 3D protein structures)
<b>9 projects, 21 refs Pathology</b>	2022-23: UNI, Virchow, DURL, PLM_SSL, SAM-Path, HiPT, Prompt-MIL, SwiFT (ResNet, ViT) (up to 1.5M Whole Slides Images)
<b>6 General Projects/Libraries 13 refs</b>	2022-23: TORCHSCALE, Trillion Parameter Consortium (30B Tokens Text + Science), AI Alliance, MLCommons, DeepSpeed4Science,
<b>13 Projects, 35 refs, Commodity LLMs</b>	2018-23: ChatGPT-4, OpenAI, Llama-2, Meta, Open source, Palm-2, Med-Palm, Google, Bard, Google, Chinchilla, DeepMind, BLOOM, Claude, Anthropic, RedPajama and SlimPajama, BERT, ROBERTa (Transformer, Mixture of Experts)(345M-1.76T parameters)
<b>21 Projects 31 refs LLMs for Science</b>	2020-23: BioBERT, ML-Net, PubMedBERT, BioGPT, BLURB, LinkBERT, BioM-ALBERT, BiomedBERT, Med-Palm2, GeneGPT, CONCH, PLIP, K2, FORGE, OceanGPT, HPC-GPT (Transformer)(68K-200M Science articles, 1.17M Image-text) AuroraGPT. INDUS

# AI for Science Foundation Models



## Earthquake Time Series

with John Rundle, Andrea Donnellan,  
Lisa Grant Ludwig, Alireza Jafari

A beautiful painting of ten small robots under the sun and moon in style of Monet, green color scheme



# Earthquakes and Deep Learning

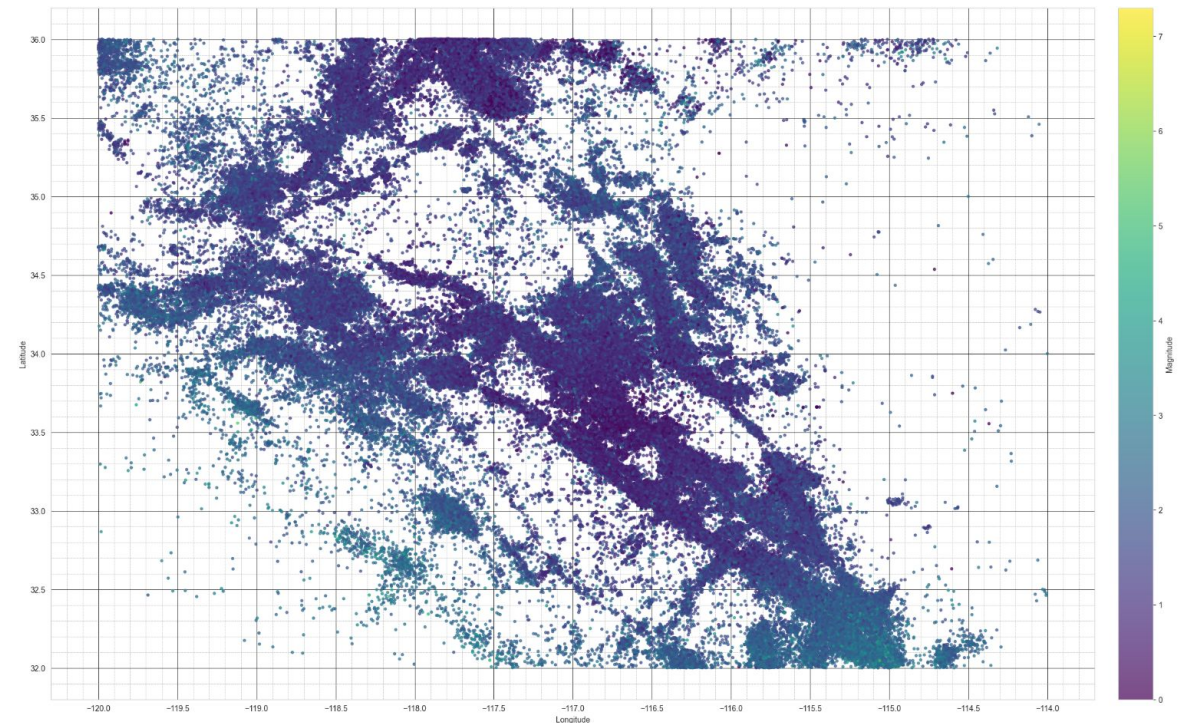
- There are at least **two major computational tasks** related to earthquakes
  - a. Forecast an occurrence of an earthquake; **data-driven?**
  - b. Predict damage once an earthquake happens; **theory-driven?**
- The second b) is complicated but doable as you can get data usable as boundary conditions for prediction of the movement of waves of earthquake energy
- a) is challenging as there is physics (theory) governing the movement of plates that generates an earthquake.
  - a. However you don't know details of plates underground and the friction laws between them
  - b. Further quake is a “phase transition” and not a deterministic motion
- This implies that are “**lots of hidden variables**” and one can hope that deep learning can model these with hidden neurons
- **Looking in a different way, one can observe data about earthquakes (the seismic shocks) but not the data needed for a physics simulation**
- We need to train a neural network to map observed data into future earthquakes
- Dog barking etc also used as possible harbingers (multi-modal data) of an earthquake

# Setting Up Earthquake Time Series

- Studies of 2400 time series for 0.1 by 0.1 degree Latitude 32 to 36 and Longitude -120 to -114 “pixels” (11 km by 11 km)
- Geospatial problems usefully characterised by Nash Sutcliffe efficiency NSE and normalized version NNSE = 1/(2-NSE).
- NSE compares discrepancy between model and data to variation from mean over time

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_m^t - Q_o^t)^2}{\sum_{t=1}^T (Q_o^t - \bar{Q}_o)^2}$$

- Earthquakes are spatially correlated as measured by fault lines
- Further work will look at simulated data where we can switch on and off different physics
- Events combined by adding energies



# Looking at lots of Models, Patterns outperform Foundation Models

Sorted on  
decreasing  
MSE

Hyperparameters  
can change






Radically  
different  
fine-tuning  
methods can  
be used

Model	Architecture	Type	Pre-training	Fine-tuning	MSE	MAE	NNSE
TimeGPT	Transformer	FM	A broad dataset	None	0.01042	0.0593	0.5484
iTransformer-M4	Transformer	FM	M4	Earthquake	0.00702	0.0537	0.5902
TSMixer-M4	MLP	FM	M4	Earthquake	0.00651	0.0535	0.6081
Chronos	Transformer	FM	A broad dataset	None	0.00650	0.0519	0.6087
PatchTST-TrafficL	Transformer	FM	TrafficL	Earthquake	0.00644	0.0501	0.6107
TiDE	MLP	P	None	Earthquake	0.00643	0.0519	0.6110
TSMixer-TrafficL	MLP	FM	TrafficL	Earthquake	0.00643	0.0505	0.6111
TimesNet	CNN	P	None	Earthquake	0.00643	0.0560	0.6112
PatchTST-M4	Transformer	FM	M4	Earthquake	0.00641	0.0504	0.6117
PatchTST-Weather	Transformer	FM	Weather	Earthquake	0.00641	0.0502	0.6119
iTransformer-TrafficL	Transformer	FM	TrafficL	Earthquake	0.00639	0.0513	0.6125
TCN	CNN	P	None	Earthquake	0.00637	0.0535	0.6132
VanillaTransformer	Transformer	P	None	Earthquake	0.00635	0.0498	0.6141
TFT	Transformer+RNN	P	None	Earthquake	0.00635	0.0555	0.6142
GNNCoder 2-layer	GNN	P	None	Earthquake	0.00632	0.0520	0.6153
LSTM	RNN	P	None	Earthquake	0.00631	0.0514	0.6156
DilatedRNN	RNN	P	None	Earthquake	0.00630	0.0510	0.6159
GNNCoder 3-layer	GNN	P	None	Earthquake	0.00629	0.0524	0.6162
GNNCoder 1-layer	GNN	P	None	Earthquake	0.00628	0.0522	0.6166

Table 1. Comparison of the performance of deep learning models for earthquake nowcasting in Southern California, ranked by Mean Squared Error (MSE) in descending order. The table compares various models used in this work, detailing their architectures, types (FM for Foundation Model, P for Pattern), and datasets for pre-training and fine-tuning. In case of Patterns, there is no pre-training, and fine-tuning is a supervised training.

# Does the Catalog of California Earthquakes, with Aftershocks Included, Contain Information about Future Large Earthquakes?

Authors

John B. Rundle   , Andrea Donnellan , Geoffrey Fox, Lisa Grant Ludwig , James P CrutchfieldPublished Online: Wed, 31 Aug 2022 | <https://doi.org/10.1002/essoar.10512008.5> Download PDF Cite Tools  Share 

## Abstract

Yes

## Aftershocks Included, Contain Information about Future Large Earthquakes:

- Our abstract: "Yes."
- **Resolution of seeming Contradiction:**
- The ongoing stream of **reduced #** of medium magnitude **earthquakes magnitude > 3.29** reveals the hidden variables controlling large **earthquakes magnitude >= 6.75**
- Any of the deep learning neural networks can discover these patterns by adding well chosen input streams
- Physics supports this: "seismic quiescence arising from the physics of strain-hardening of the crust prior to major events" <http://doi.org/10.1002/essoar.10510940.4>

## Models for Earth

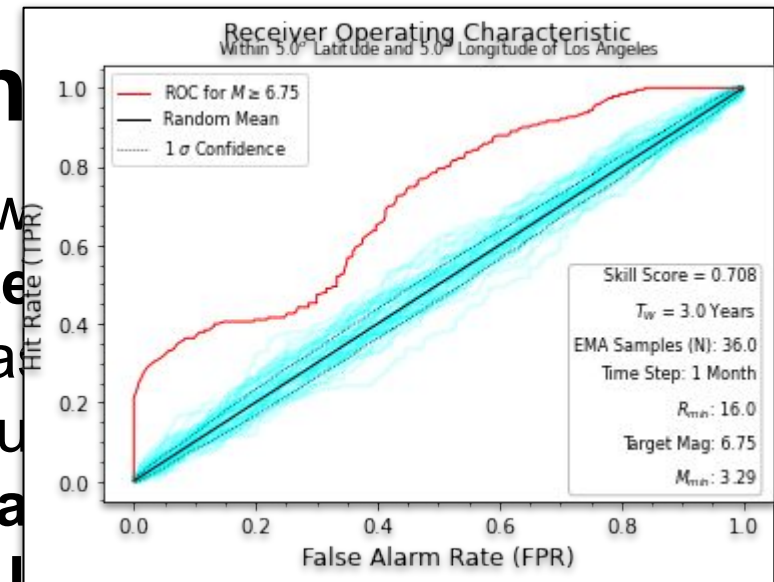
(1974) wrote a paper w

**California, with after**

analysis they did wa

tial probability distribu

**log of California Ea**



# Adding Physics motivated Features to Basic model

Model	Input	MSE	MAE	NNSE
LSTM	Single feature	0.00631	0.0514	0.6156
LSTM	+ Multiplicity	0.00630	0.0506	0.6158
DilatedRNN	Single feature	0.00630	0.0510	0.6159
LSTM	+ Multiplicity + EMA	0.00629	0.0527	0.6162
LSTM	+ EMA	0.00628	0.05177	0.6164
GNNCoder 1-layer	+ Multiplicity	0.00628	0.0520	0.6165
GNNCoder 1-layer	Single feature	0.00628	0.0522	0.6166
GNNCoder 1-layer	+ Multiplicity + EMA	0.00627	0.0517	0.6169
DilatedRNN	+ Multiplicity	0.00627	0.0517	0.6169
GNNCoder 1-layer	+ EMA	0.00627	0.0525	0.6172
DilatedRNN	+ EMA	0.00627	0.0519	0.6174
DilatedRNN	+ Multiplicity + EMA	0.00625	0.0515	0.6174

# AI for Science Foundation Models



## Lessons from Hydrology Time Series

with Eric He

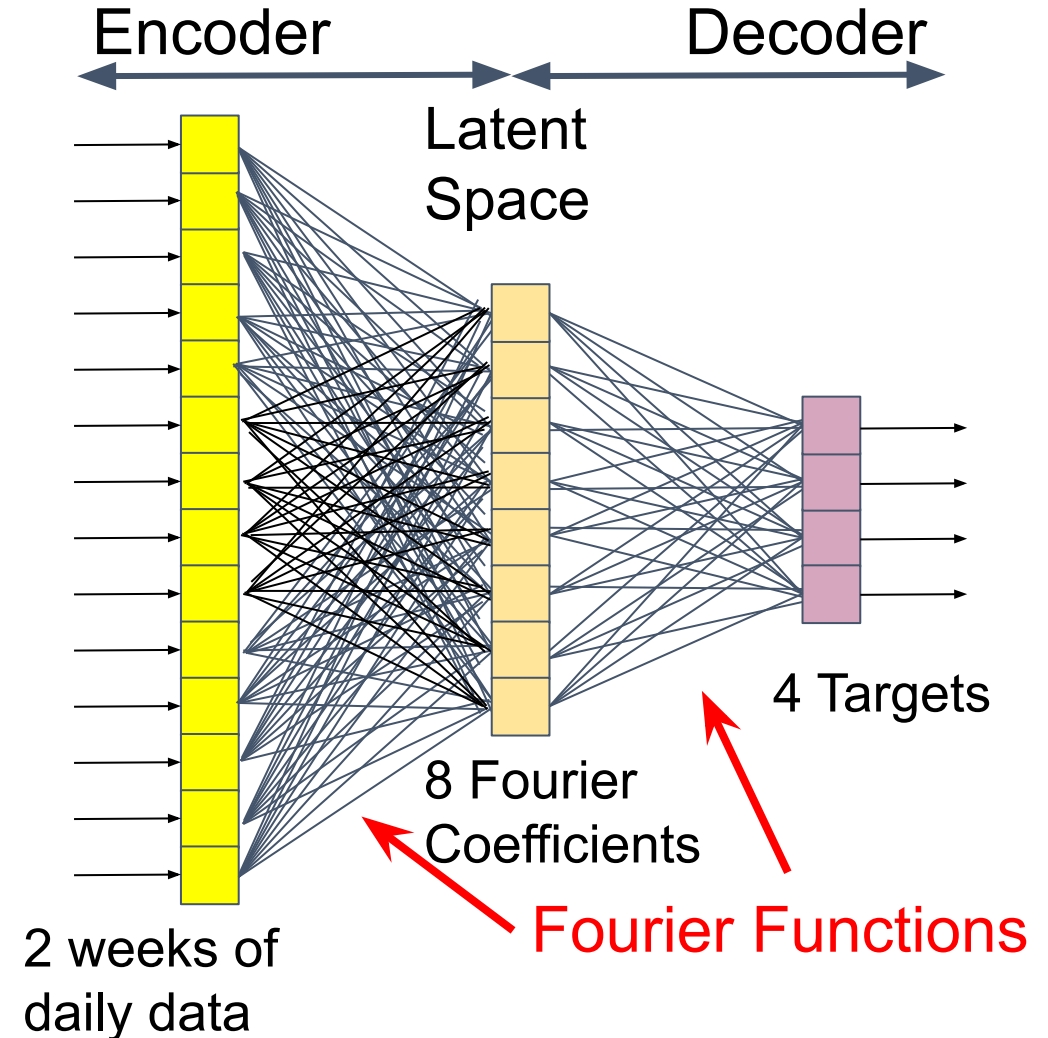
A beautiful painting of ten small robots under the sun and moon in style of Monet, green color scheme

# Orthogonal Functions as Foundation Models I

- Suppose data known over a window of size  $W$ :  $D(t)$  is given as  $d(0)\dots d(W-1)$  for  $t = 0\dots W-1$
- Let  $f(l)(t)$  be an orthogonal polynomial for  $l=0\dots\infty$  with  $L$  as internal layer size (size of Latent space) so we use  $l = 0\dots L-1$ 
  - Fourier series, Legendre polynomials, Laguerre polynomials, wavelets
- Expansion Coefficients are  $a(l=0\dots L-1)$  defined as
- $a(l) = \int f(l)(t) D(t) dt$  integrated over  $t = 0..W-1$
- We can use a numerical integration formula to write  $a(l) = \sum_{t=0}^{t=W-1} d(t)f(l,t)$ 
  - Many different choices; Midpoint, Trapezoidal, Simpson's rule; learn best
- $f(l,t)$  is a weight matrix of size  $L$  by  $W$  mapping input data into expansion coefficients and forms the encoder which is independent of  $D(t)$
- The decoder to predict general time  $T$  is given as
- Predicted( $T(k=0..K)$ ) =  $\sum_{l=0}^{l=L-1} a(l) f(l,T(k))$
- The  $L$  by  $K$  weight matrix depends on  $T(k)$  but not on  $D$

# Orthogonal Functions as Foundation Models II

- This is a simple MLP with three layers of sizes  $W$   $L$  and  $K$
- Different choices of  $L$  and functions  $f(l)(t)$  lead to different Foundation models
- Choice of  $L$  (Latent space size) well understood and studied and could be learnt
- One can also feed in any number of orthogonal functions as “known inputs” or “exogenous variables” e.g.  $\cos(2\pi t/365)$  and  $\sin(2\pi t/365)$  for  $t$  in days - two input known time series



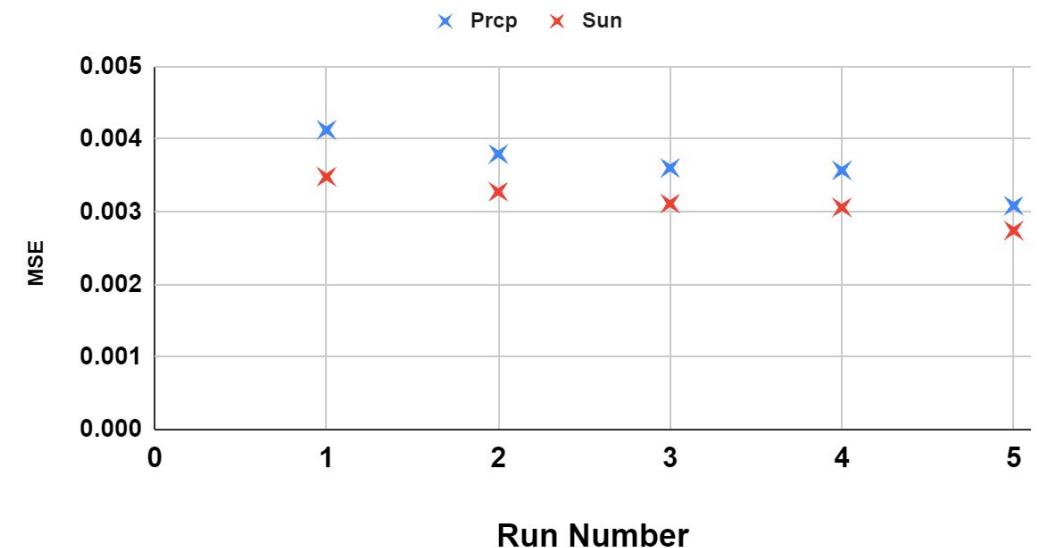


# Sensitivity to “Known Inputs” or “Exogenous Variables” I

- Camels USA -- 7010 days, 671 locations, 27 Static variables (like aridity, mean Temperature), 6 dynamic variables (Precipitation, Streamflow, Solar Radiation, Vapor Pressure, minimum and maximum Temperatures)
- Dynamic variables used in all runs
- 5 studies with different “Known Inputs” which are static and dynamic features known at all times independent of deep learning model
- Model 2-stage LSTM with layers of size 320 (improves over smaller sizes)

Run Number	1	2	3	4	5
six Dynamic	x	x	x	x	x
Linear space-time		x	x	x	x
cos, sin with annual period			x	x	x
27 Static				x	x
11 extra Fourier, Legendre in time					x

MSE for Precipitation and Solar Radiation versus Known Inputs



# Reducing Static Variables with PCA

- Caravan USA data has 482 catchments over 7010 days with 209 static variables
- 3 Dynamic series predicting the next day
- Calculate correlation matrix of static variables over all catchments
- 34 leading PCA eigenvectors capture 90% of the signal
- Solutions with 209 static variables similar to that with smaller number of 34 static eigenvectors.

## Example Static variables

Snow cover extent per month  
 Temperature monthly mean, annual mean/min/max  
 Wetland Class  
 Forest Coverage %  
 Clay fraction in soil  
 Silt fraction in soil  
 Population density  
 Nighttime lights  
 Lithological class  
 Gross domestic product .....

CARAVAN US PCA EXPERIMENT

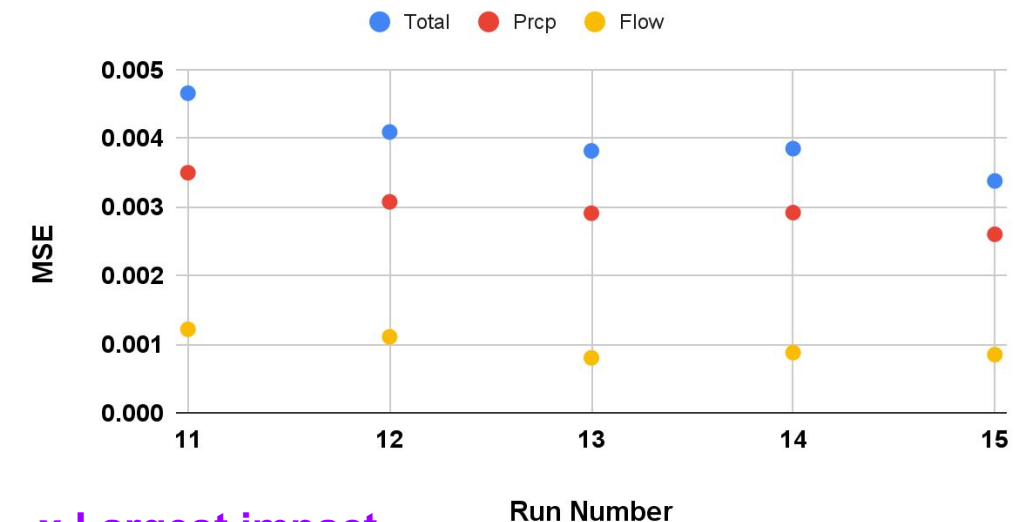
		Original Static		PCA Static	
		MSE	NNSE	MSE	NNSE
<b>Precipitation</b>	Train	0.002920	0.851	0.002994	0.848
	Val	0.004307	0.801	0.004264	0.800
<b>Mean Temperature</b>	Train	0.000468	0.967	0.000468	0.967
	Val	0.000573	0.963	0.000548	0.965
<b>Streamflow</b>	Train	0.000525	0.814	0.000569	0.799
	Val	0.000955	0.703	0.000989	0.703

# Sensitivity to “Known Inputs” or “Exogenous Variables” II

- Caravan Hysets-- 7010 days, 4621 locations, 209 Static variables (like aridity, mean Temperature), 3 dynamic variables (Precipitation, Streamflow, mean Temperature)
- Dynamic variables used in all runs
- 5 studies with different “Known Inputs” which are static and dynamic features known at all times independent of deep learning model
- Model 2-stage LSTM with layers of size 320 (improves over smaller sizes)
- Adding in math functions works well
- Using PCA reduces number of static variables with little impact

Run Number	11	12	13	14	15
Three Dynamic	x	x	x	x	x
Linear space-time	x	x	x	x	x
Cos, Sin with annual period		x	x	x	x
209 Static			x		x
34 PCA of 209 static				x	
add 11 more Fourier, Legendre in time					x

MSE for Total, Precipitation and Flow



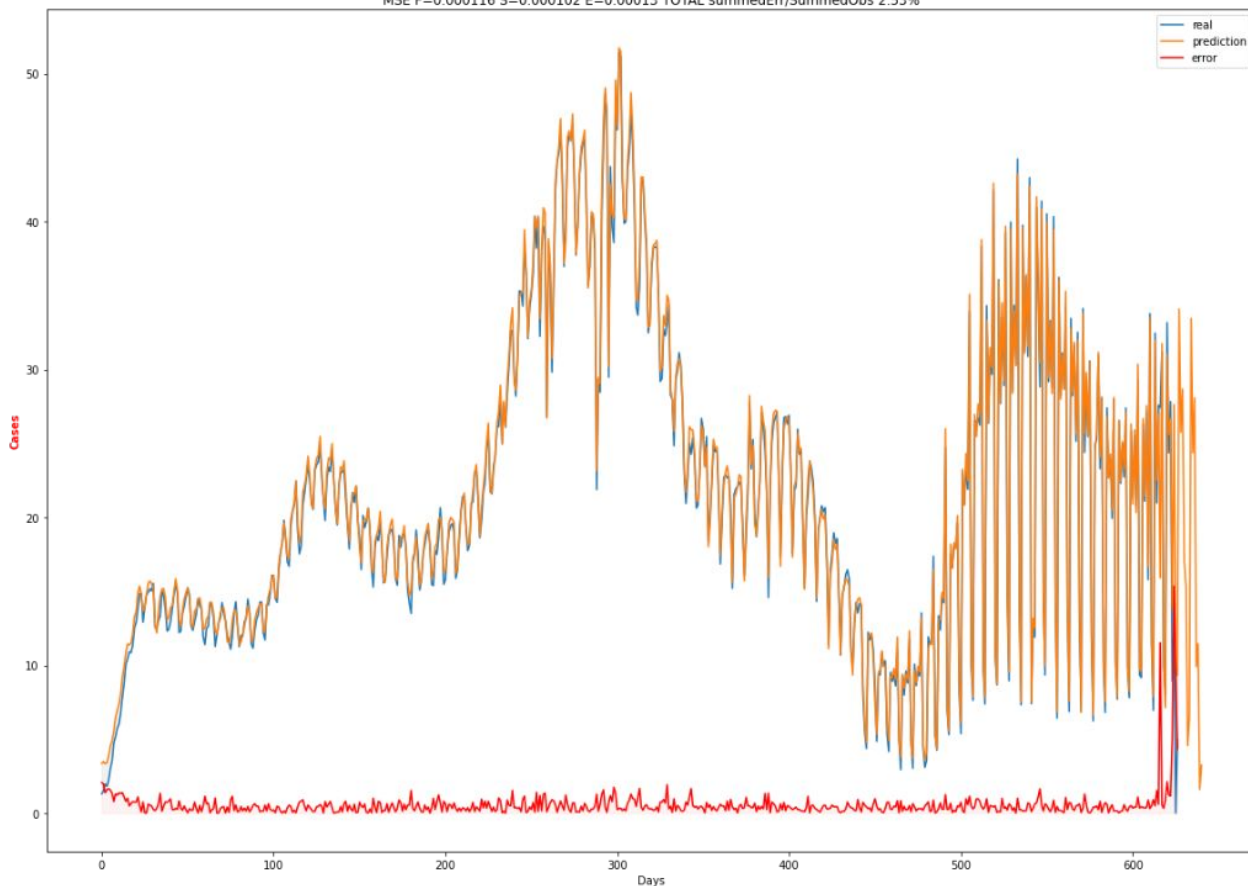
x Largest impact

# LSTM/TFT Description of Covid Data

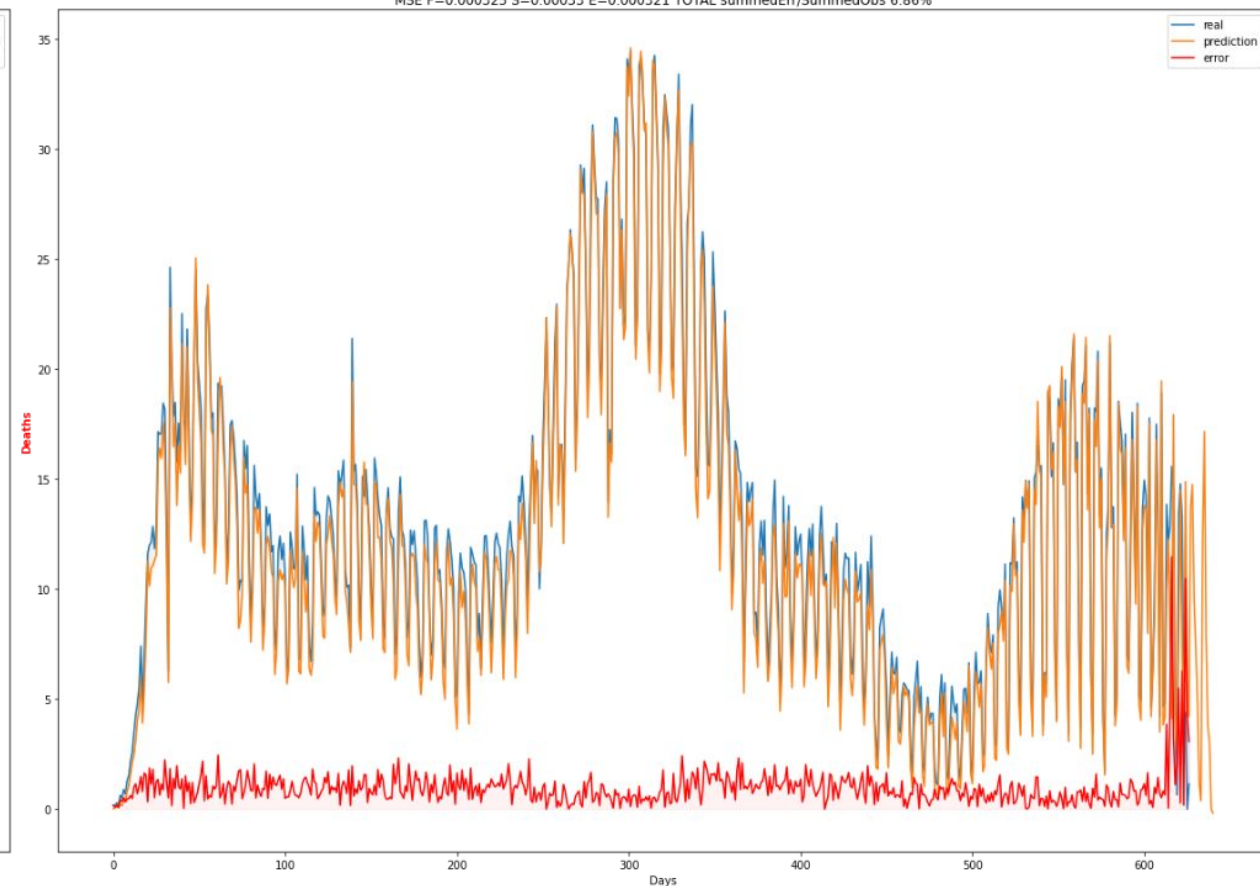
Uses Weekly “known input”

500 most populous counties of 3142 in the USA

Full CovidN21-newTFTv23redo 0.00318 12/05/2021, 19:47:24 UTC  
CovidN21-newTFTv23redo Length=627, Location Summed Results Cases, AVG  
MSE F=0.000116 S=0.000102 E=0.00013 TOTAL summedErr/SummedObs 2.53%

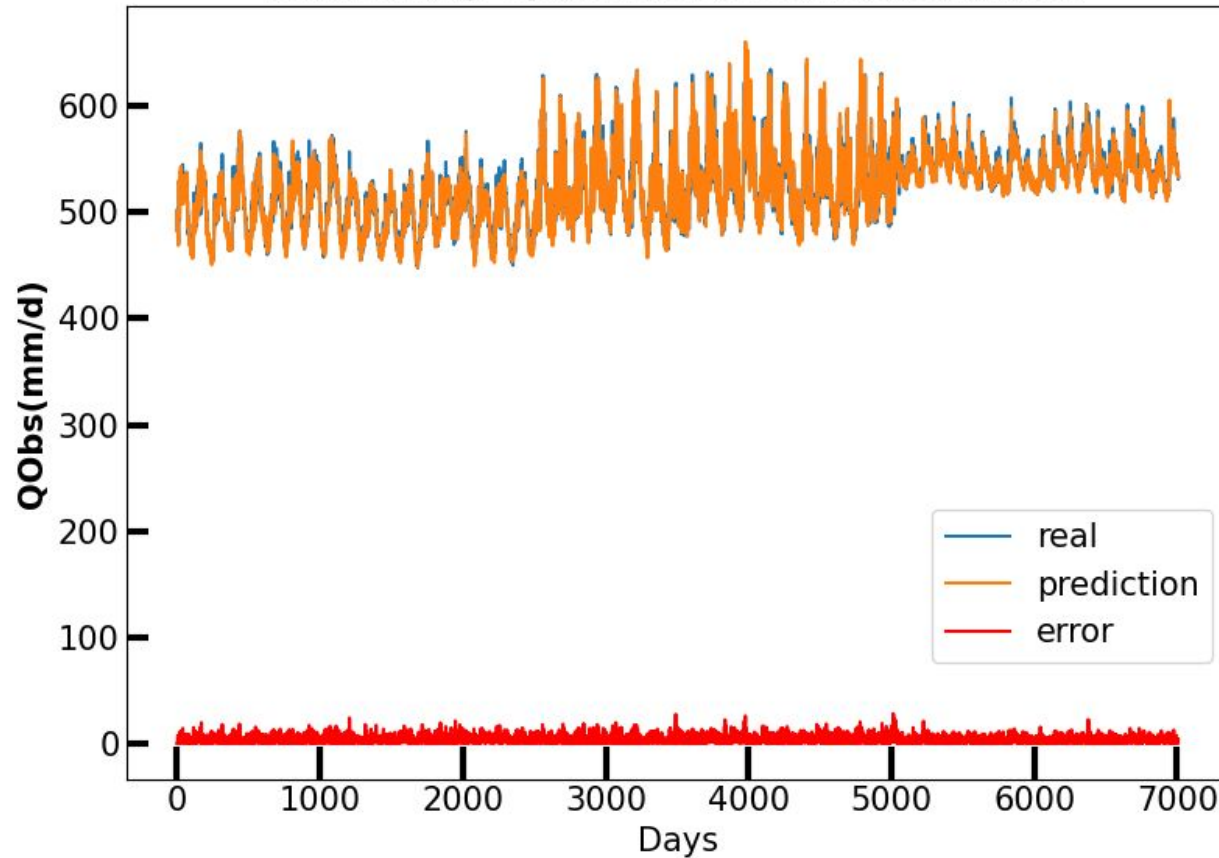


Full CovidN21-newTFTv23redo 0.00318 12/05/2021, 19:47:24 UTC  
CovidN21-newTFTv23redo Length=627, Location Summed Results Deaths, AVG  
MSE F=0.000325 S=0.00033 E=0.000321 TOTAL summedErr/SummedObs 6.86%

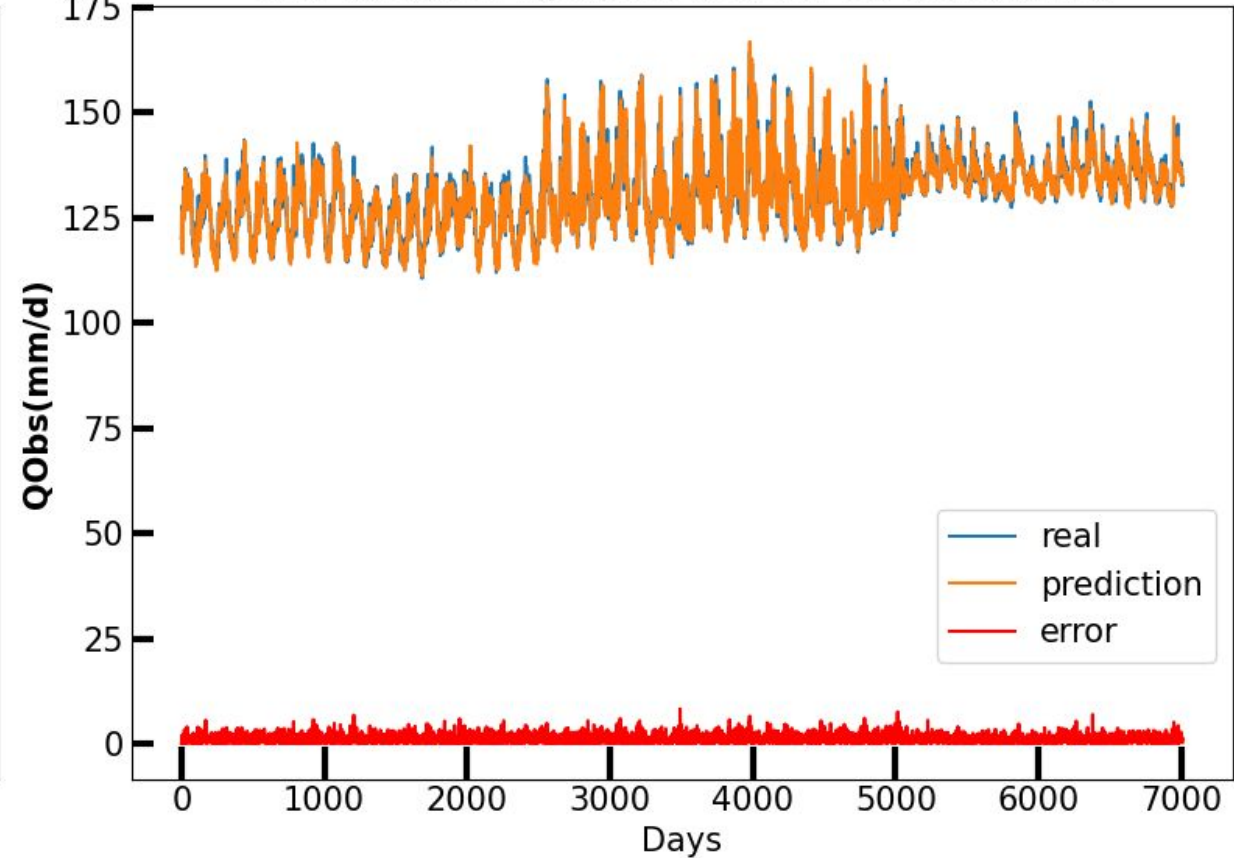


# Annual Behavior seen in Hydrology

Total Training HydroIN3AnewLSTM33 QObs(mm/d)

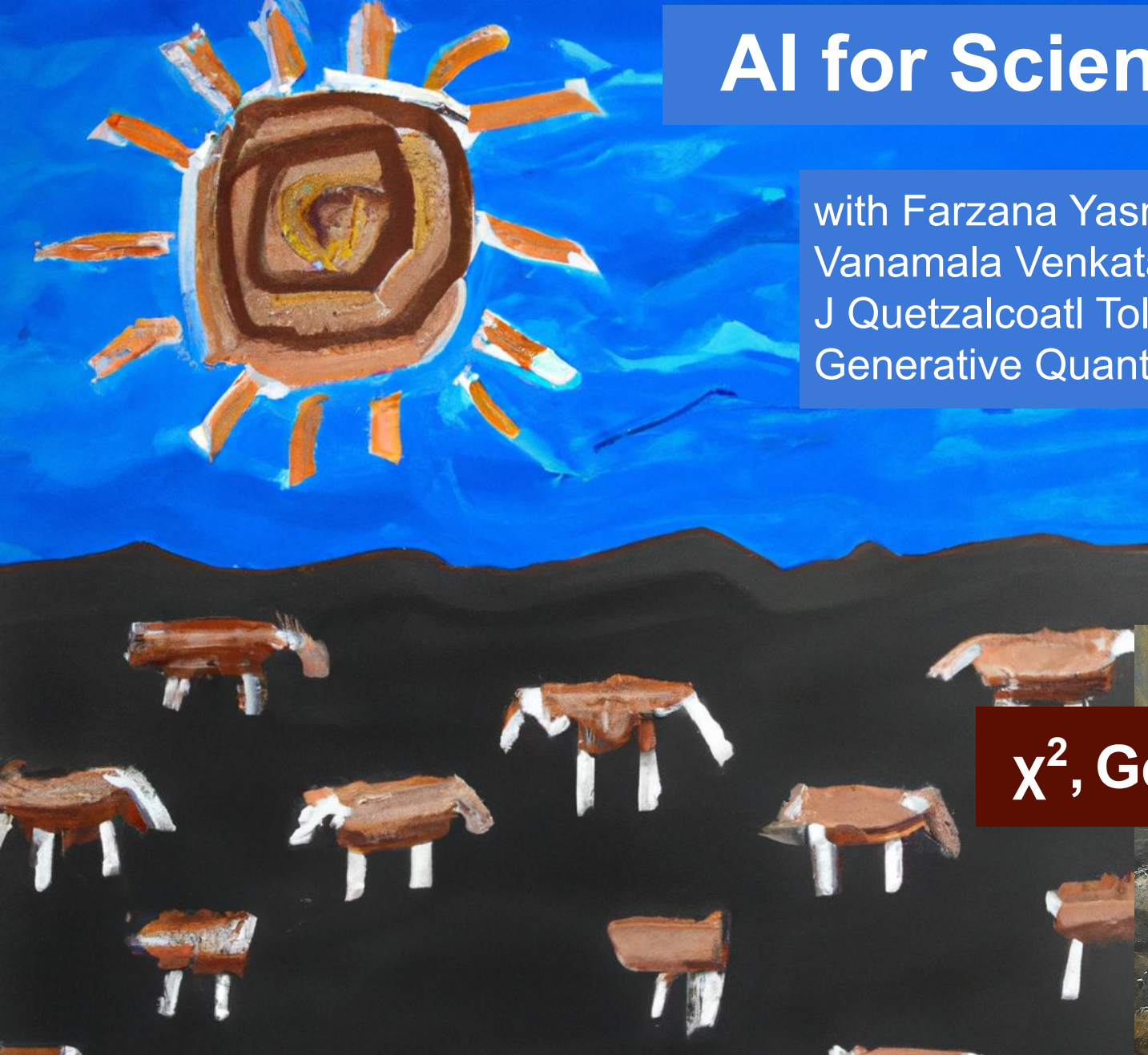


Total Validation HydroIN3AnewLSTM33 QObs(mm/d)

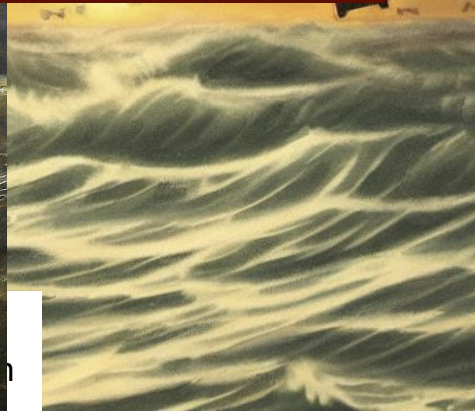


# AI for Science Foundation Models

with Farzana Yasmin Ahmed and  
Vanamala Venkataswamy,  
J Quetzalcoatl Toledo-Marin,  
Generative Quantum group at TRIUMF



$\chi^2$ , Generative AI, and Surrogates

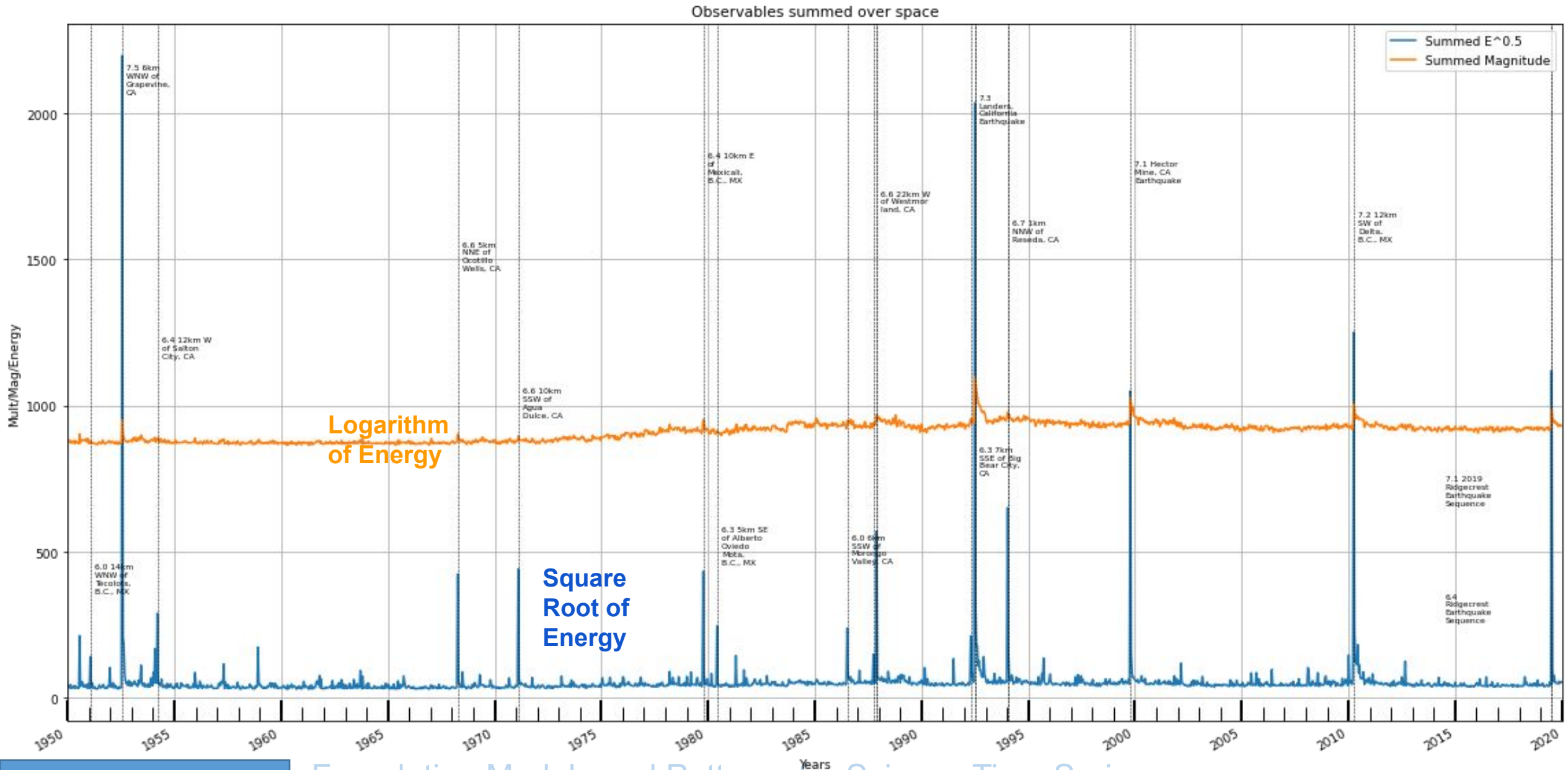


DALL-E Gemini Stable Diffusion: "An ugly painting of the sun shining on a stormy sea with ten small black robots swimming in the sea in style of an amateur"

# Choice of Variables and Errors

- Note science data should and often does have credible estimates of errors in observables
  - Also consider correlations and both systematic and statistical errors
- Note if I observe values  $x_i$ , then in deep learning, I can replace  $x_i$  by  $f(x_i)$  for any function  $f$  with intuitive constraint that  $f(x)$  is a monotonically increasing function of  $x$
- in  $MSE = \sum (\text{Observed} - \text{Predicted})^2$  or  $MAE = \sum \text{Abs}(\text{Observed} - \text{Predicted})$ , one weights large observables more in MSE than MAE
- Often there are huge variations in size and many more small observables than large observables; larger values tend to have larger errors as in  $N \pm \sqrt{N}$  for Poisson
- Further the input data is multiplied by weights and passed through activation layers;
  - Activation will have different impact on large and small observables
- For Earthquakes one typically discusses magnitude =  $\log(\text{Energy } E)$  rather than Energy  $E$  or strain  $\sqrt{E}$ ; a huge difference in loss function
- **Chisq**  $\chi^2 = \sum (\text{Observed} - \text{Predicted})^2 / \text{Error}^2$  is classic or better MSE form is
- **Correlated** is  $\sum_{i,j} (\text{Observed}(i) - \text{Predicted}(i)) (\text{Observed}(j) - \text{Predicted}(j)) C^{-1}(i,j)$

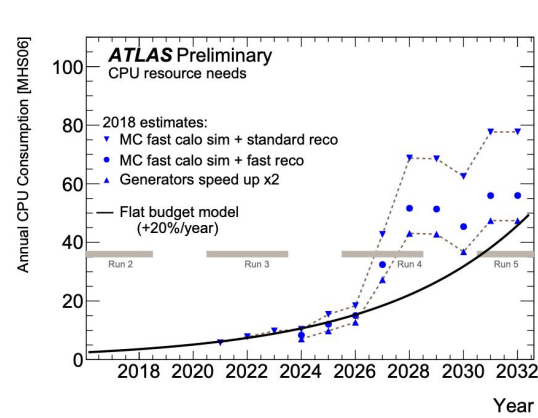
# Comparing Logarithm versus Square Root 1950-2019



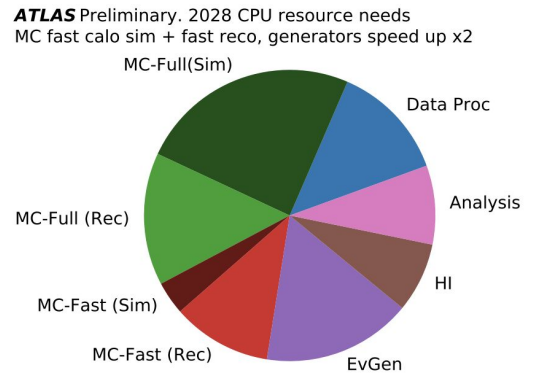


# Generative AI for Kaggle Calorimeter Surrogates

- Generation of simulated events is a significant computing load at LHC
- These are typically generated by GEANT4 from known physics of particle material interactions
- AI surrogates must be generative to mimic a Monte Carlo
- **Errors** are largely proportional to  $\sqrt{\text{Energy}}$  and there are **significant correlations**; often ignored



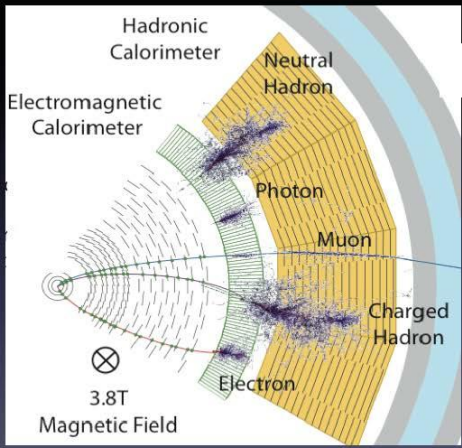
(a)



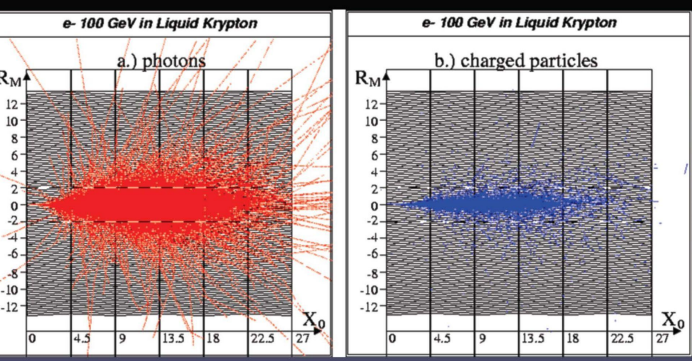
(b)



- In nuclear and particle physics calorimetry refers to the detection of particles through total absorption in a block of matter
  - The measurement process is destructive for almost all particle
  - The exception are muons (and neutrinos) → identify muons easily since they penetrate a substantial amount of matter
- In the absorption, almost all particle's energy is eventually converted to heat → calorimeter
- Calorimeters are essential to measure neutral particles



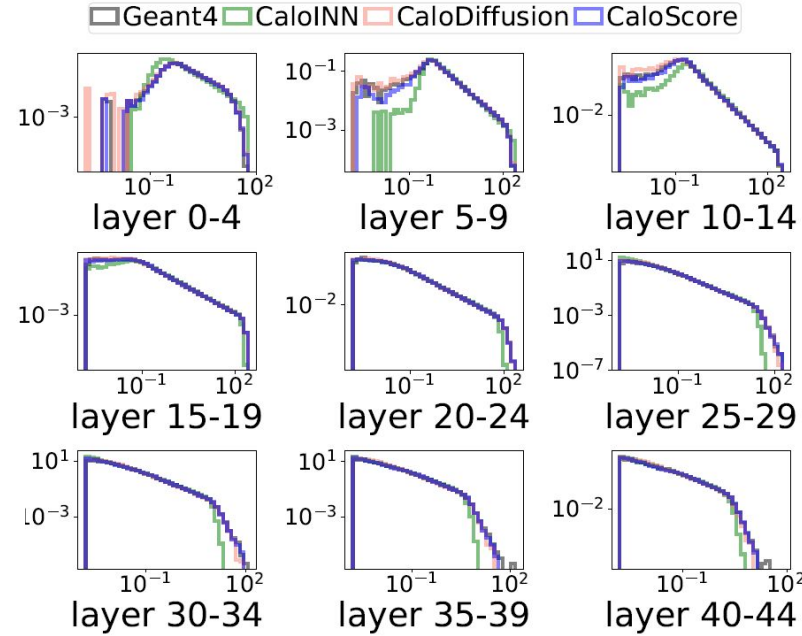
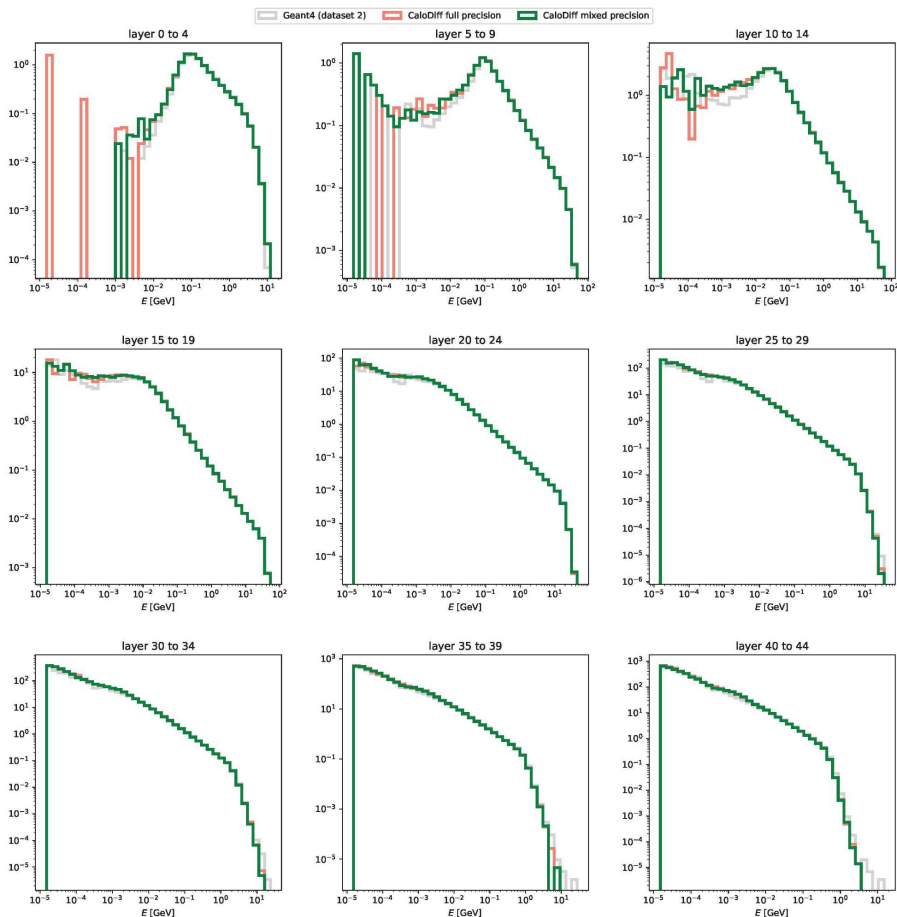
## EM shower development in liquid krypton (Z=36, A=84)



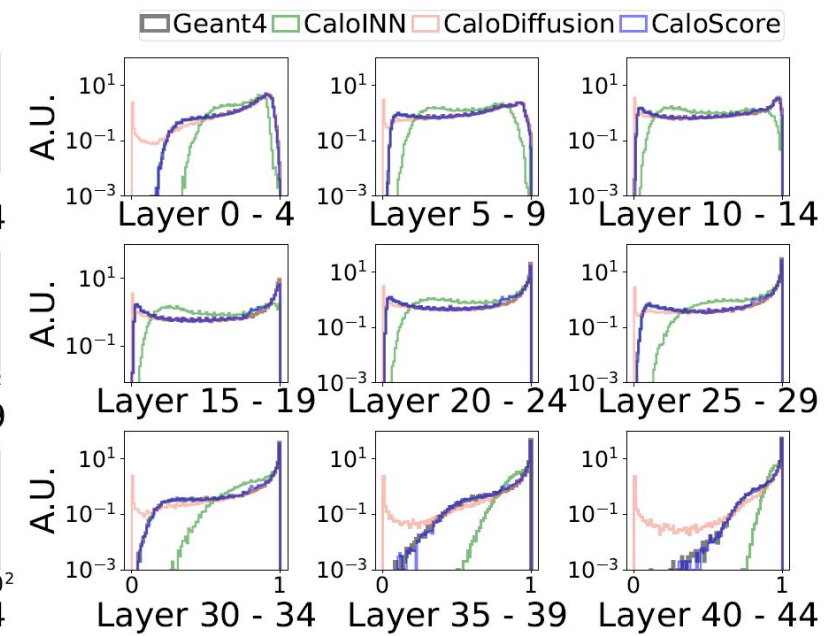
ANT simulation of a 100 GeV electron shower in the NA48 liquid Krypton calorimeter (D.Schinz)

# Comparison of Geant and Generative Calorimeter Simulators

- Correlations are large
- FP16 (.12 sec) versus FP32 (.22 secs) speeds up as in LLM

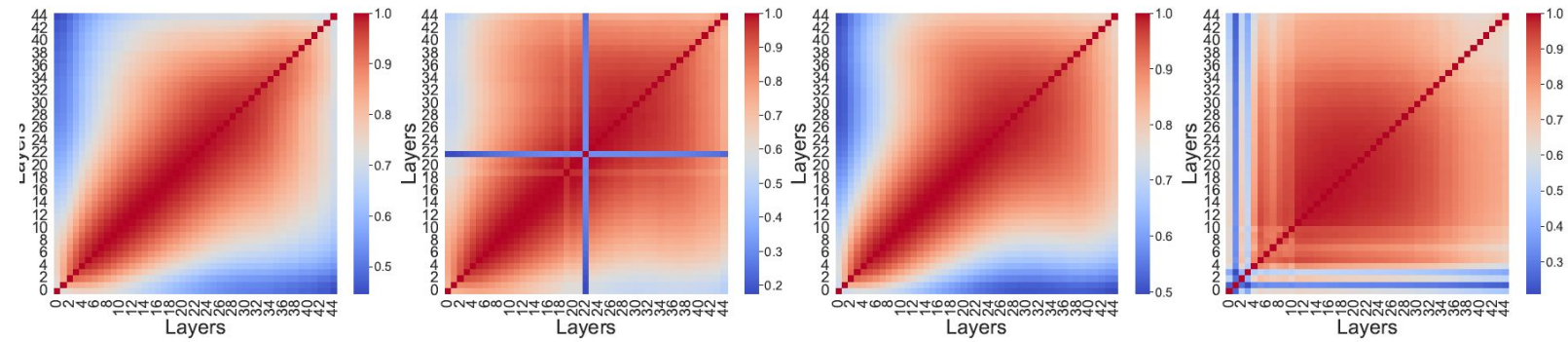


(a) Layer energy distribution (GeV)



(b) Distribution of sparsity

Figure 1: Histogram of two physics observables for dataset 2.



(a) Geant4

(b) CaloScore

(c) CaloDiffusion

(d) CaloINN

# Generative AI Surrogate Methods

- GAN
- VAE Variational Autoencoder
- QVAE VAE with Quantum Spin Generator from Dwave
- Diffusion models
- Normalizing Flow
- There are two terms in QVAE loss
- Classic MSE loss and a KL Divergence that is forcing the distribution to be correct
- It appears one should use real errors and correlations but we don't see this as useful so far

## Timing of Methods

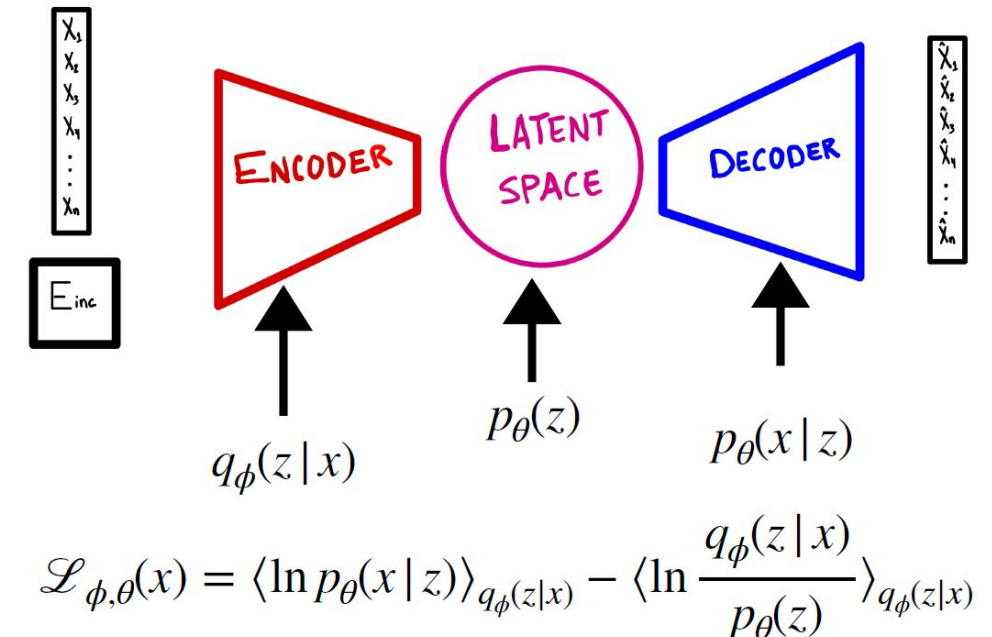
Geant4 ~ 1s

GPU (A100) ~ 2ms

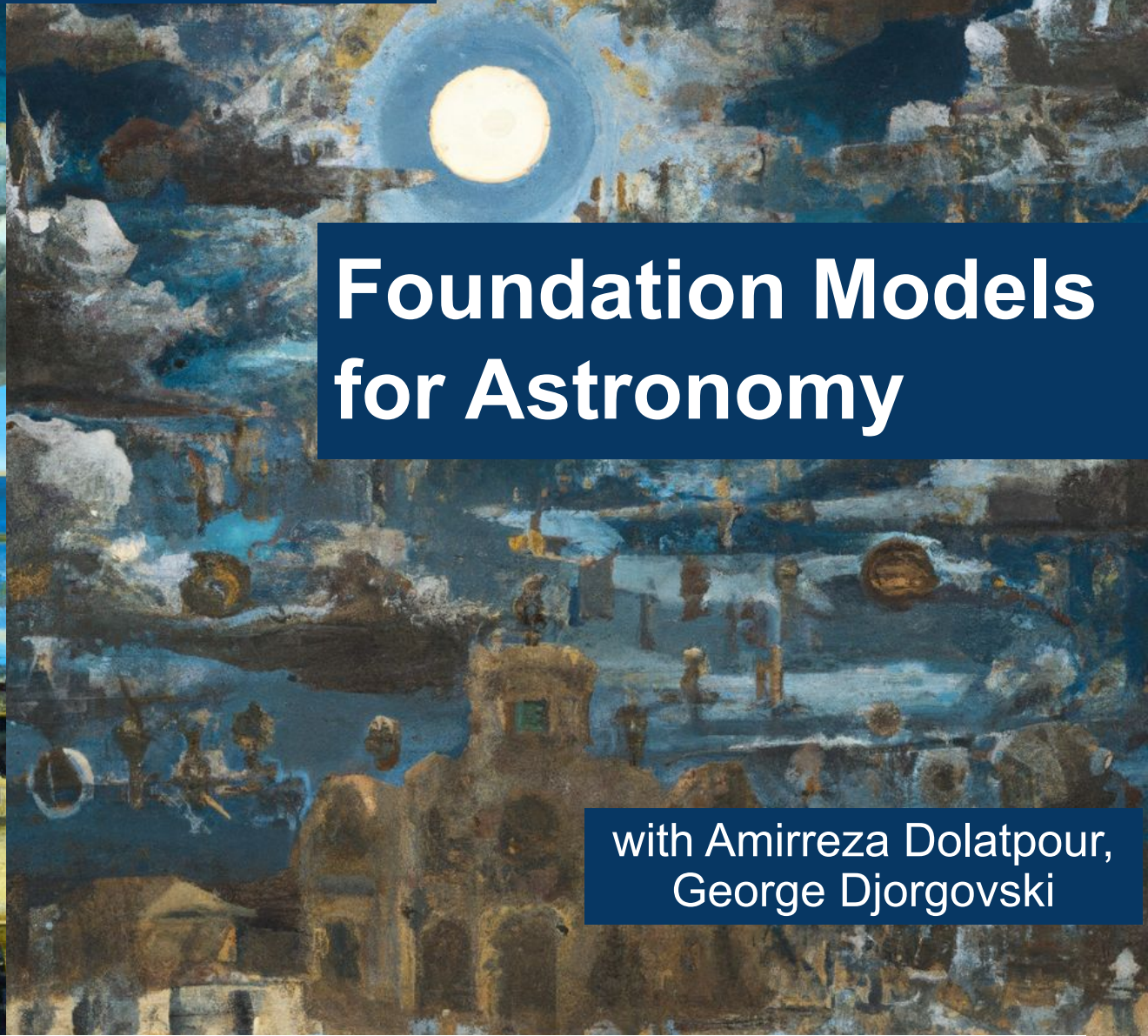
QVAE 0.2ms

QVAE Annealing time ~ 0.02ms

## Variational Autoencoders



# AI for Science Foundation Models



## Foundation Models for Astronomy

with Amirreza Dolatpour,  
George Djorgovski

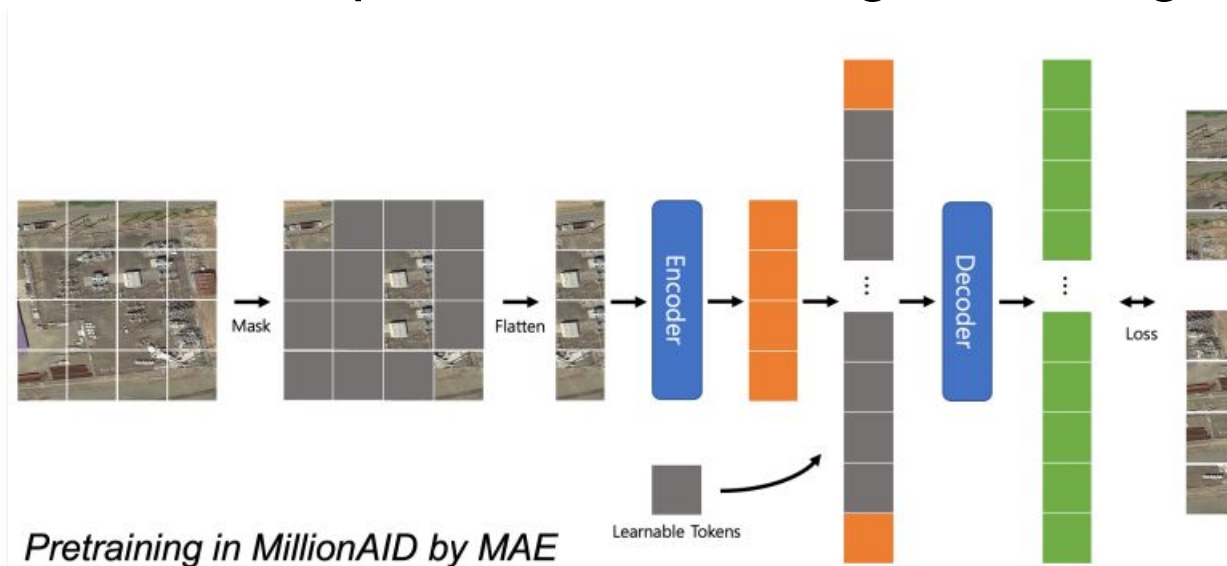
DALL-E

A beautiful painting of ten small black robots swimming in the sea with a supercomputer on an island by Constable

A beautiful painting of sixteen small robots outside a gorgeous palace in stormy cloudy sky in style of Leonardo da Vinci. There is a Sun and a Moon in the sky

# Astronomy/Remote Sensing Foundation Models

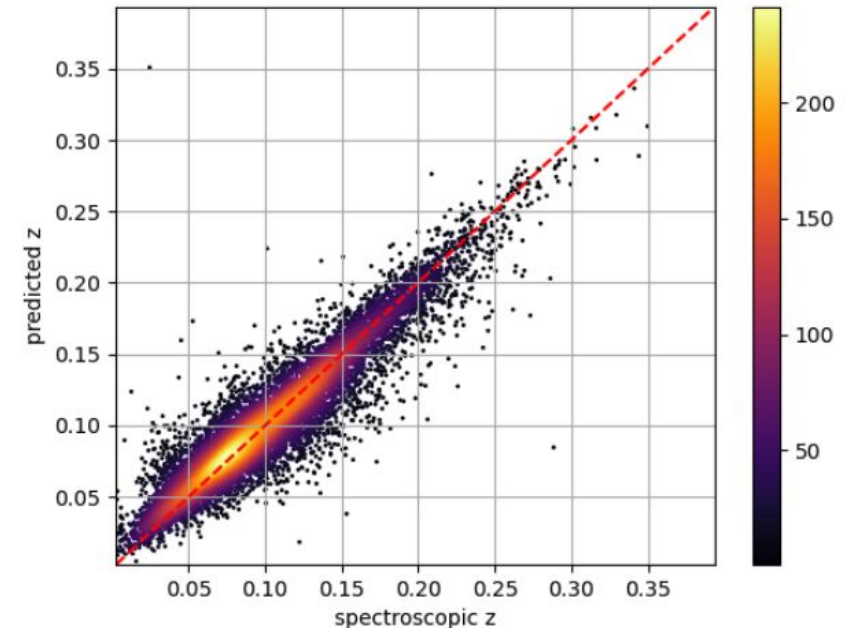
- Train images using vision transformer with masking
- Transformer is alternative to CNN where CNN filters are replaced by transformer attention mechanism
- The Foundation model is pretrained by Masked Autoencoder to recreate masked images
- Images from 96 by 96 to 224 by 224
- Millionaid is a remote sensing dataset
- Batch size 512, form patches and change masking each batch; 75% masked



# Foundation Models for Astronomy



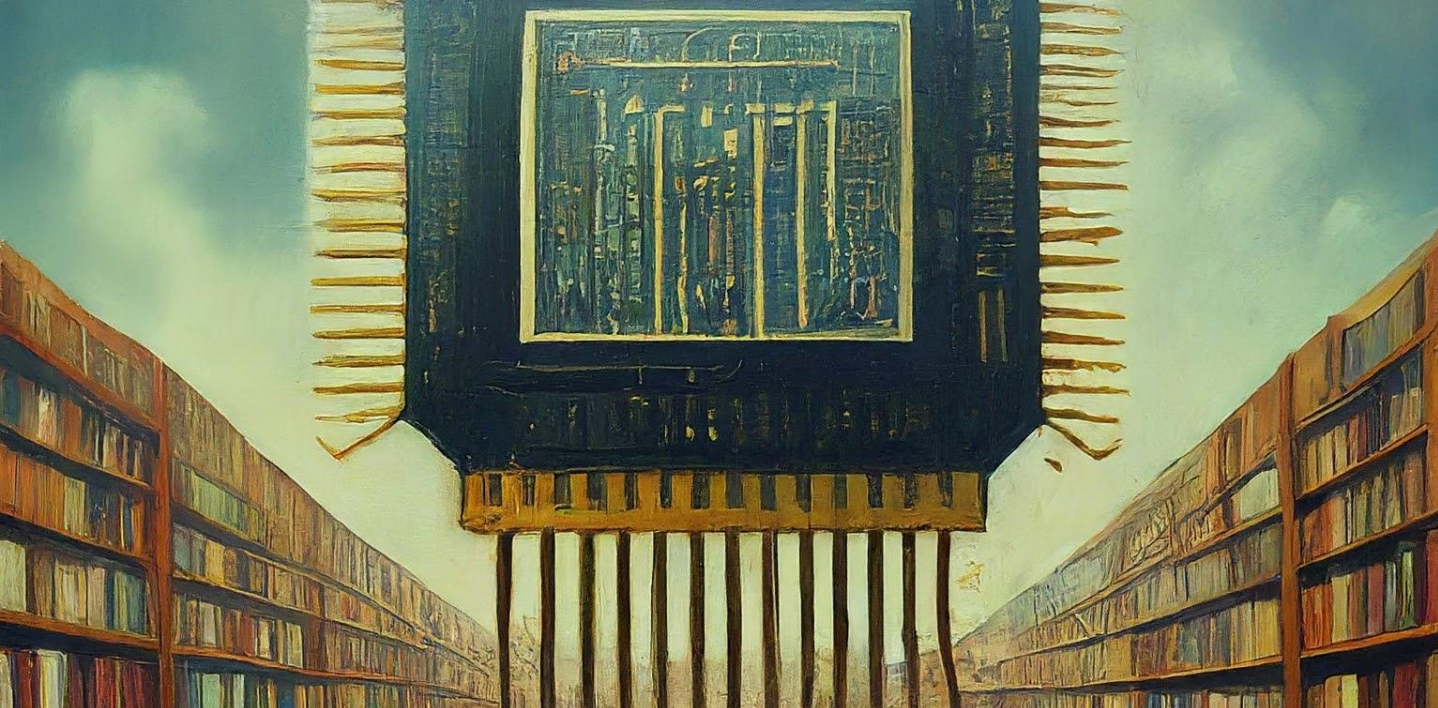
- Similar to many scientific fields that produce images
  - From Universe to electron scales and in many different photon wavelengths
- Illustrate approach with galaxy catalogs
- The distance of a galaxy comes from magnitude and velocity from redshift which is caused by galaxies moving away from the Milky way. This causes the spectrum to shift to the red
- Either observe individual spectral lines or the relative magnitude of different frequency bands
- “All” galaxies have images but not all spectra
- So use “all” images to build a Foundation model to understand structure
- Then fine tune on different data classes: with or without spectra; different frequency band choices to find redshift with measurement accuracy determined by galaxies with spectra



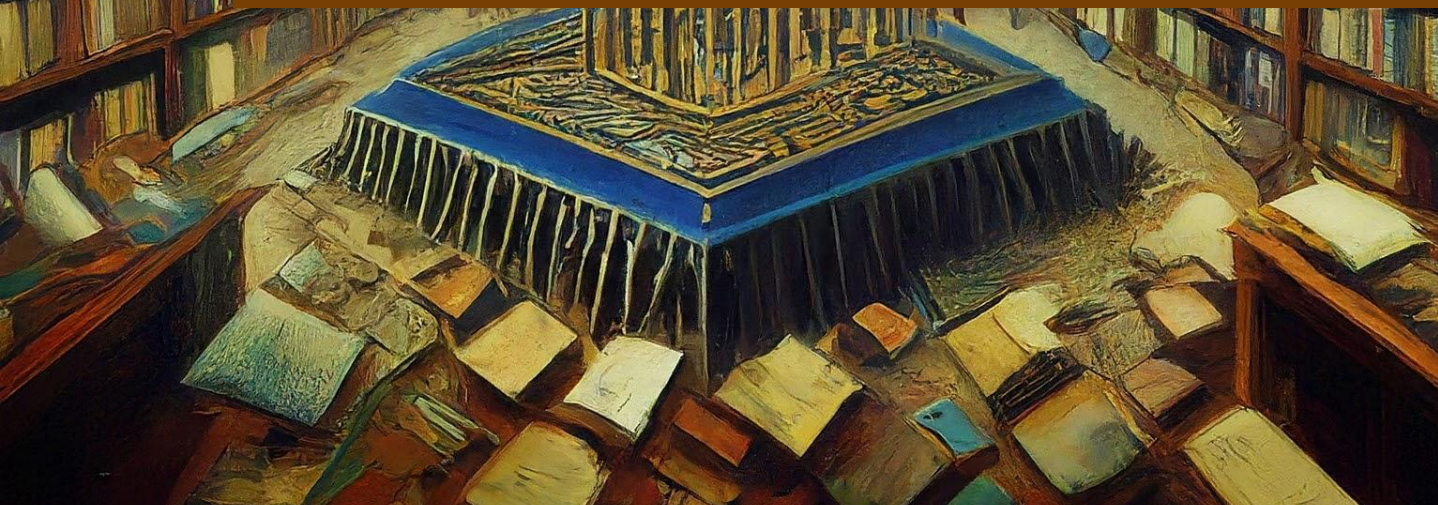
# Foundation Model Compared to Training from Scratch

TABLE IV: Redshift Prediction Using Various Architectures Based on Transformer Layers and CNNs

Architectures		Metrics				
Type	Name	MSE	MAE	Bias	Precision	R <sup>2</sup>
Supervised training (from scratch)	from-scratch plain-ViT-magnitude	0.00077	0.01871	0.00153	0.01736	0.93580
	from-scratch pcm-ViT-magnitude	0.00057	0.01604	-0.00035	0.01458	0.95204
	Henghes et al. [38]	0.00058	0.01568	0.00108	0.01443	0.95176
	from-scratch plain-ViT	0.00097	0.02123	0.00049	0.01957	0.91871
	from-scratch pcm-ViT	0.00063	0.01686	-0.00122	0.01554	0.94764
	Inception-only redshift prediction	0.00064	0.01705	0.00132	0.01593	0.94625
Fine-tuning	plain-ViT-magnitude	0.00068	0.01740	<b>-0.00007</b>	0.01596	0.94334
	pcm-ViT-magnitude	0.00060	0.01655	-0.00095	0.01522	0.94939
	Proposed plain-AstroMAE	0.00056	0.01558	0.00097	0.01429	0.95336
	Proposed pcm-AstroMAE	<b>0.00053</b>	<b>0.01520</b>	-0.00037	<b>0.01391</b>	<b>0.95601</b>
	plain-ViT	0.00086	0.01970	-0.00060	0.01775	0.92790
	pcm-ViT	0.00084	0.01945	-0.00114	0.01737	0.92950
	plain-ViT-inception	0.00059	0.01622	-0.00009	0.01496	0.95029
	pcm-ViT-inception	0.00059	0.01601	0.00042	0.01458	0.95095



# AI for Science Foundation Models and Patterns



# Conclusions

GEMINI A beautiful painting of a library with lots of books surrounding a giant computer chip in style of M.C. Escher, John Constable



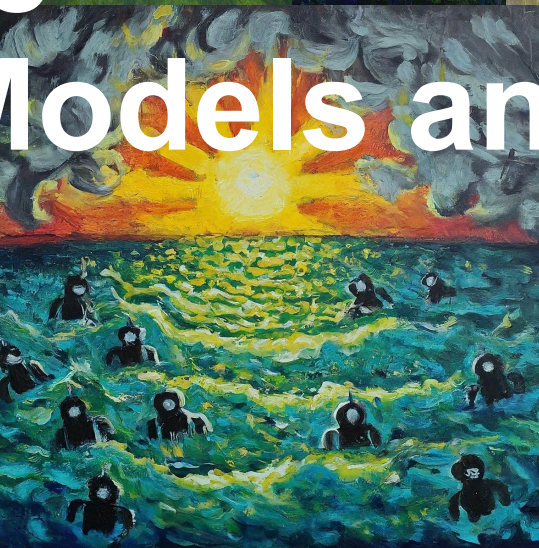
# Lessons for Science Patterns and Foundation Models

- Need to look at both Patterns and Foundation Models
- Patterns successful in Time Series
  - Illustrated for Earthquakes and Hydrology
- Exogenous (Known) Inputs should be included
- Fine-tuning of Time Series Foundation models needs more research
- PCA can reduce input and compute size without losing significant information
- Spatial information can be important and included with graph neural networks for traffic and earthquake nowcasting
- Image Foundation models also sensitive to fine-tuning
- Understanding of errors and correlations in scientific data can be incorporated in statistical distributions for deep learning
- Diffusion models and generative AI broadly important as in recent data assimilation breakthroughs for weather forecasting



# AI for Science

# Foundation Models and Patterns



**The End**