Science and Technology Facilities Council

AI Benchmarks and Science

Overview, Challenges, Specifics and Opportunities

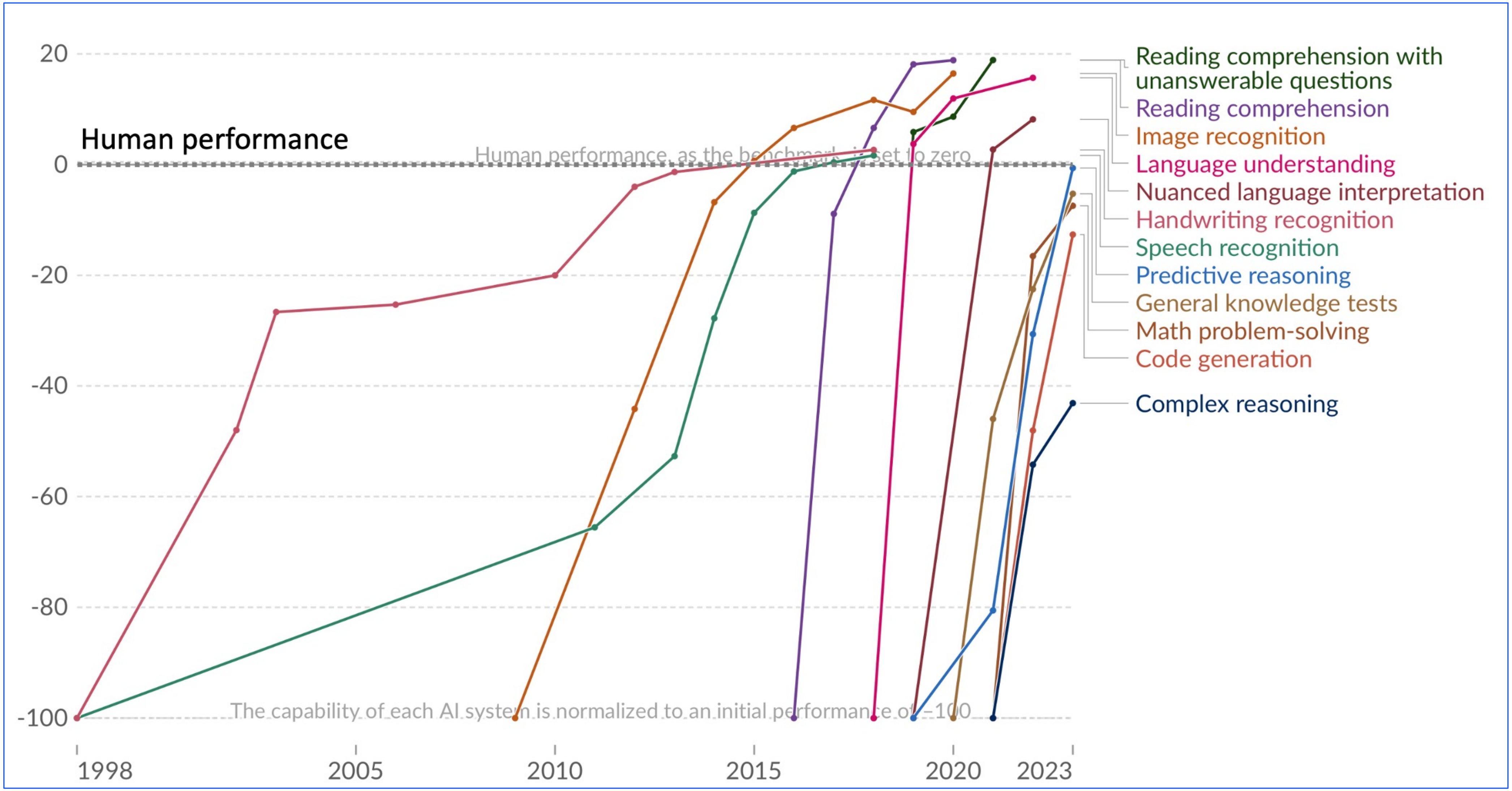
Jeyan Thiyagalingam
Rutherford Appleton Laboratory

t.jeyan@stfc.ac.uk

Big Picture:

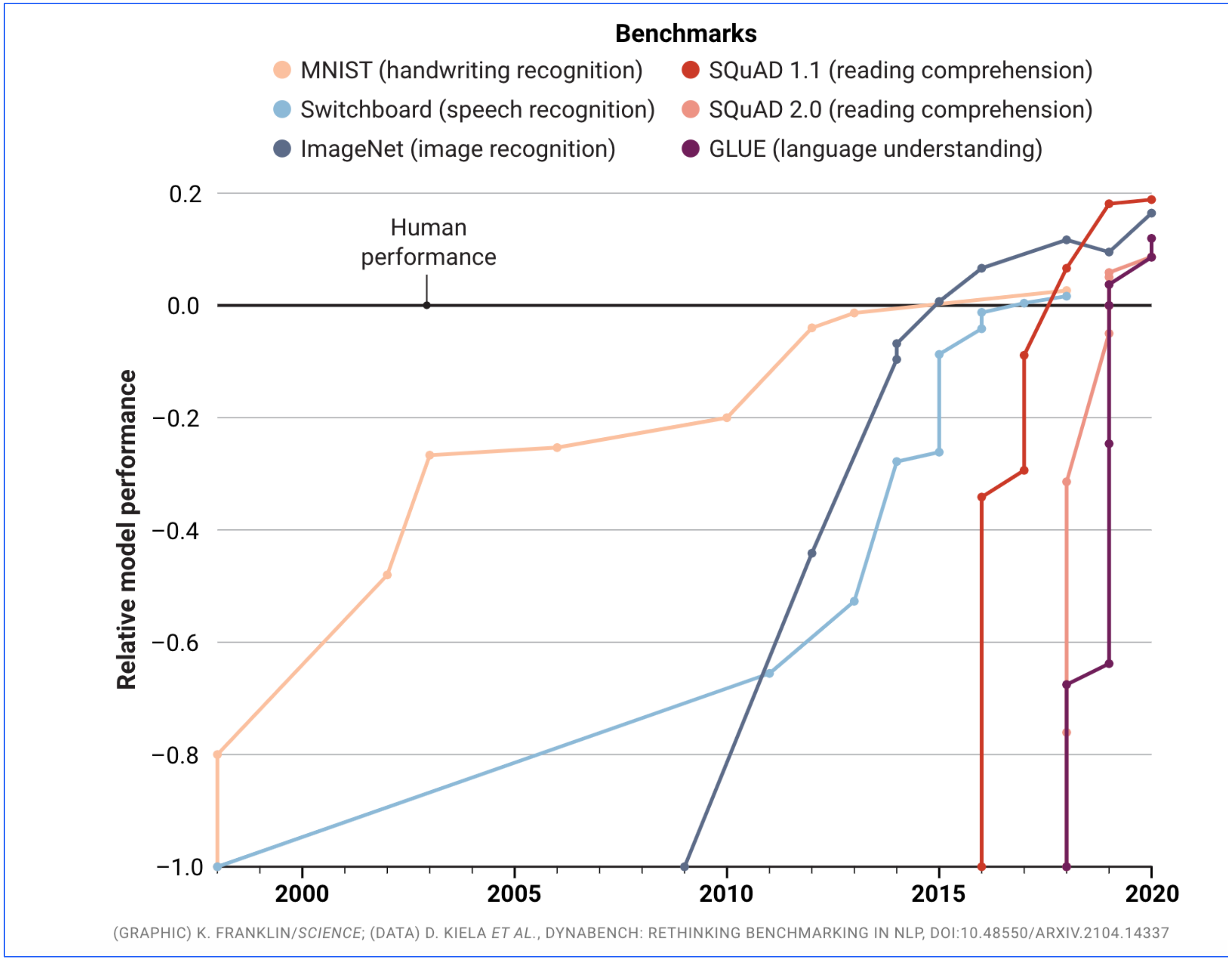
Landscape of AI & AI for Science

AI Systems and Human Level Performance

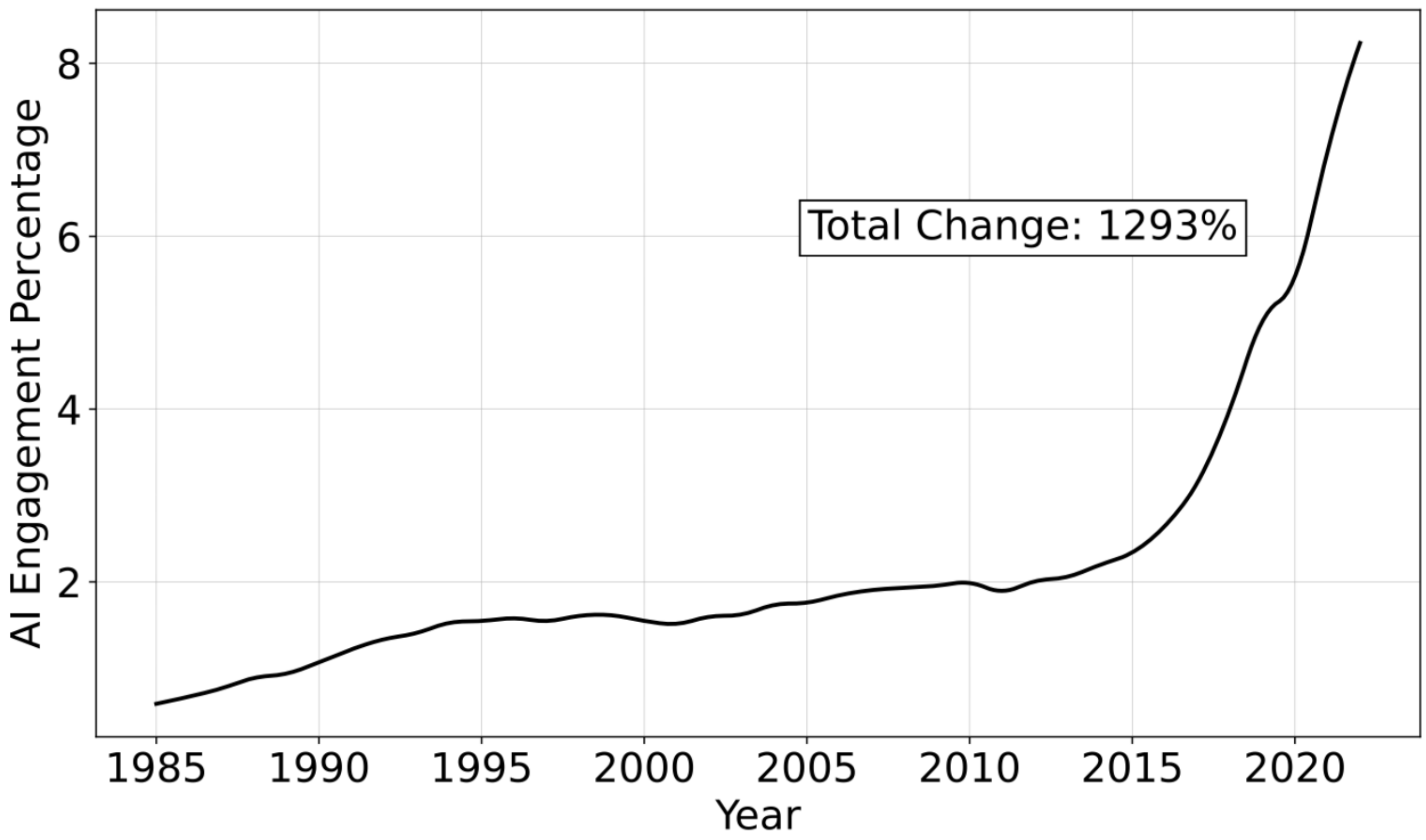


Kiela et al. – Dynabench: Rethinking Benchmarking in NLP (2023)

Ceased Efforts



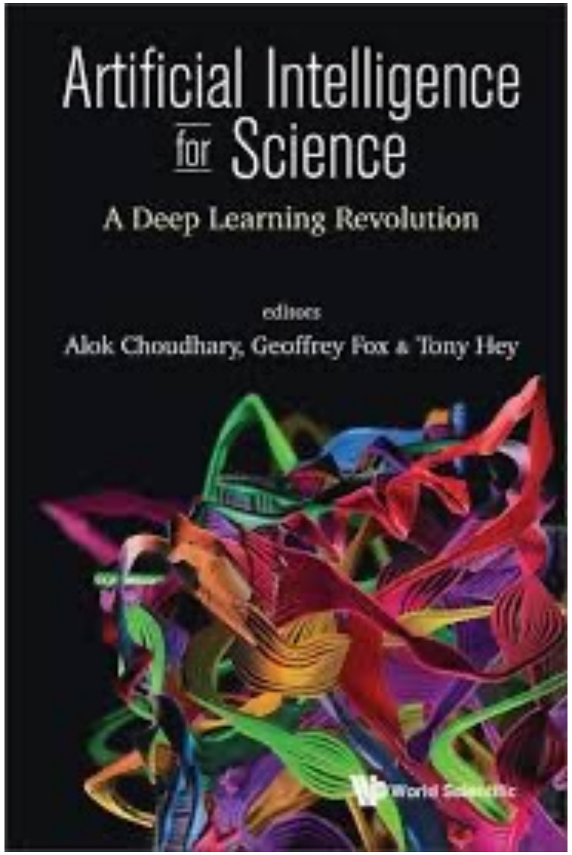
Diffusion of AI within Scientific Fields

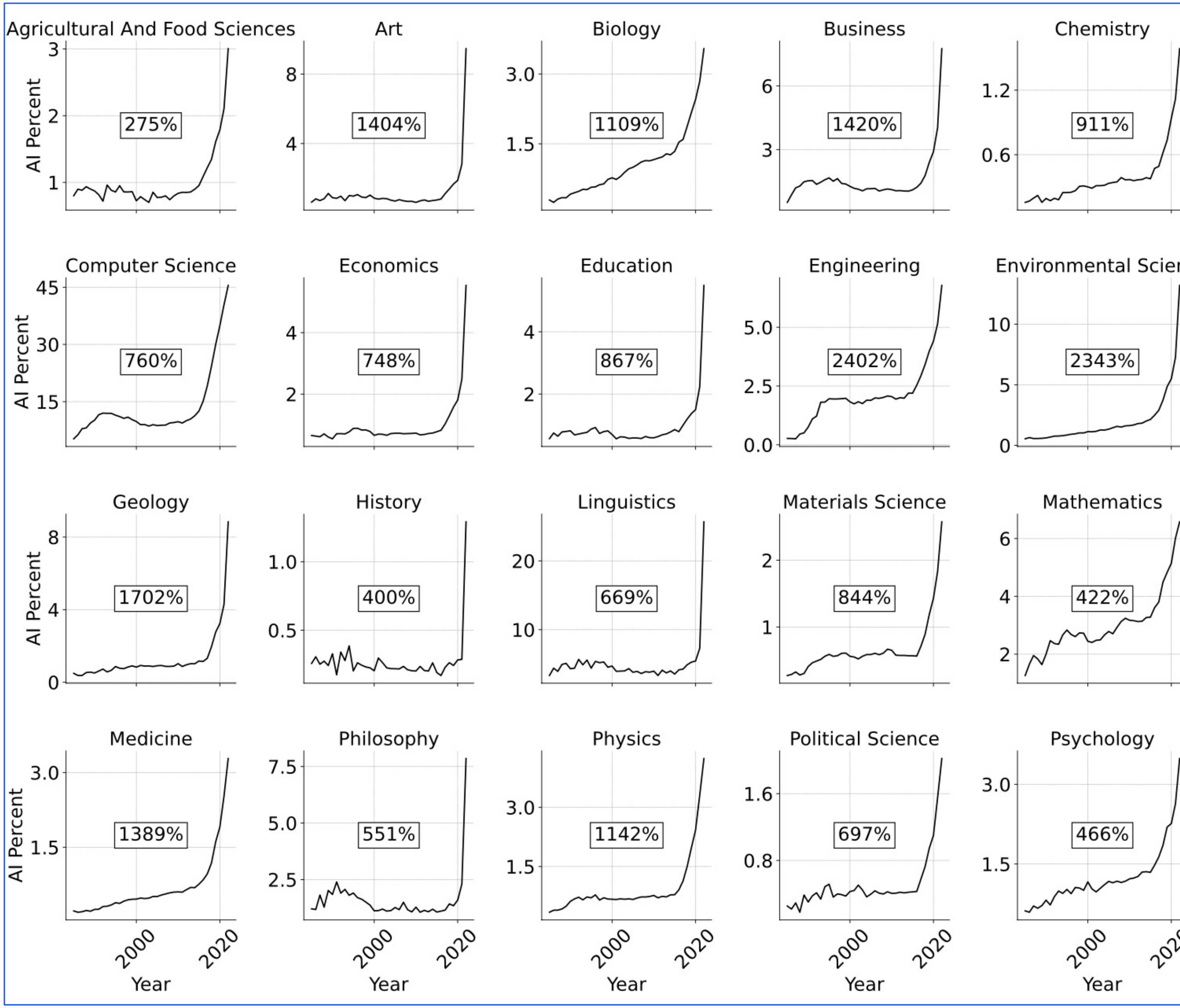


AI Engagement across All Fields is Exponential

Oil & Water: https://arxiv.org/pdf/2405.15828

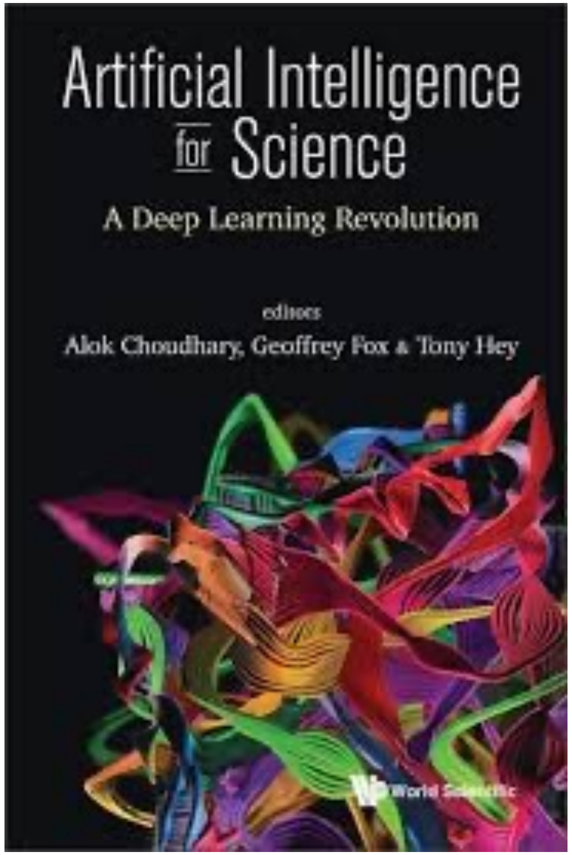
AI Engagement Across 20 Exemplar Fields

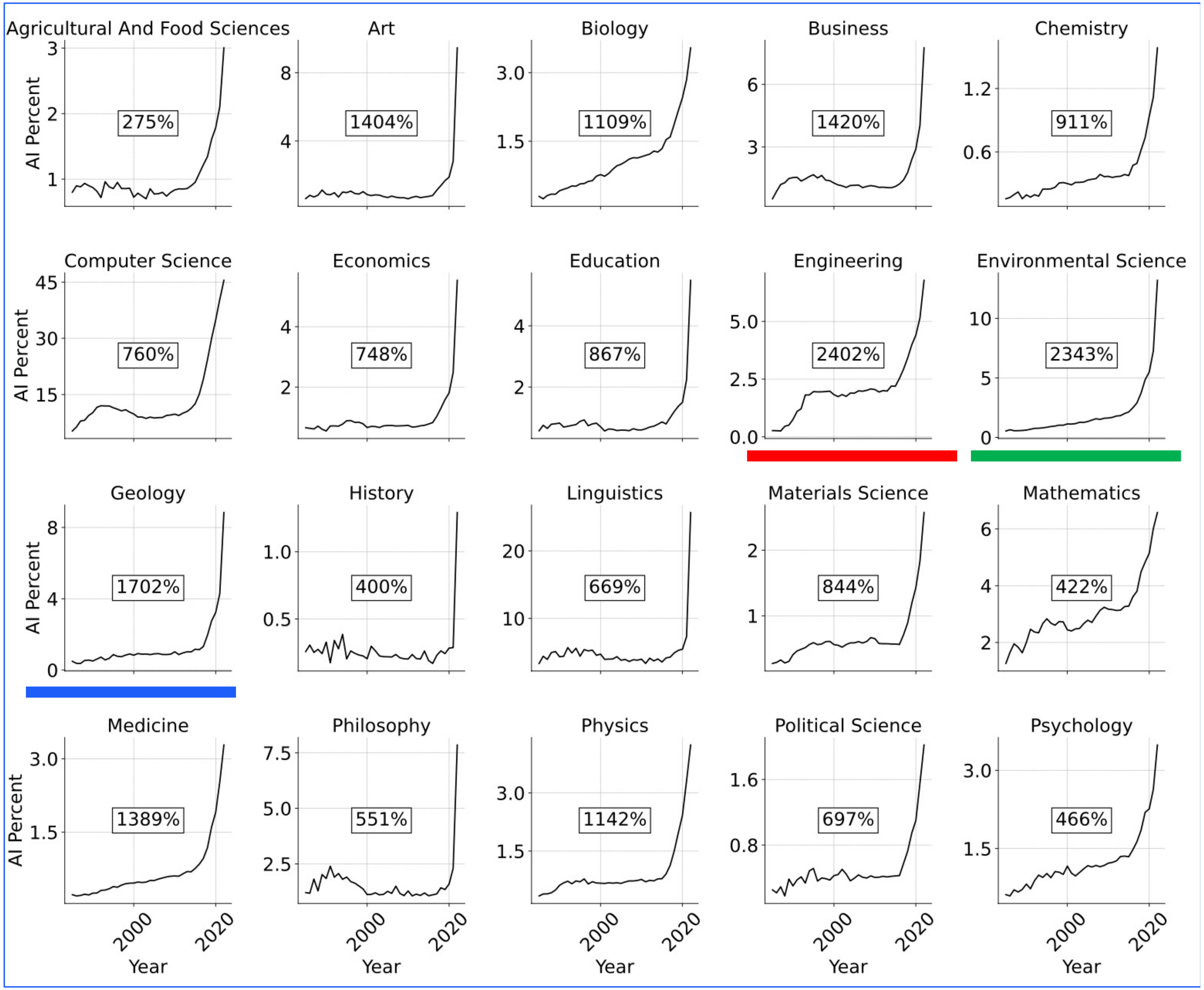




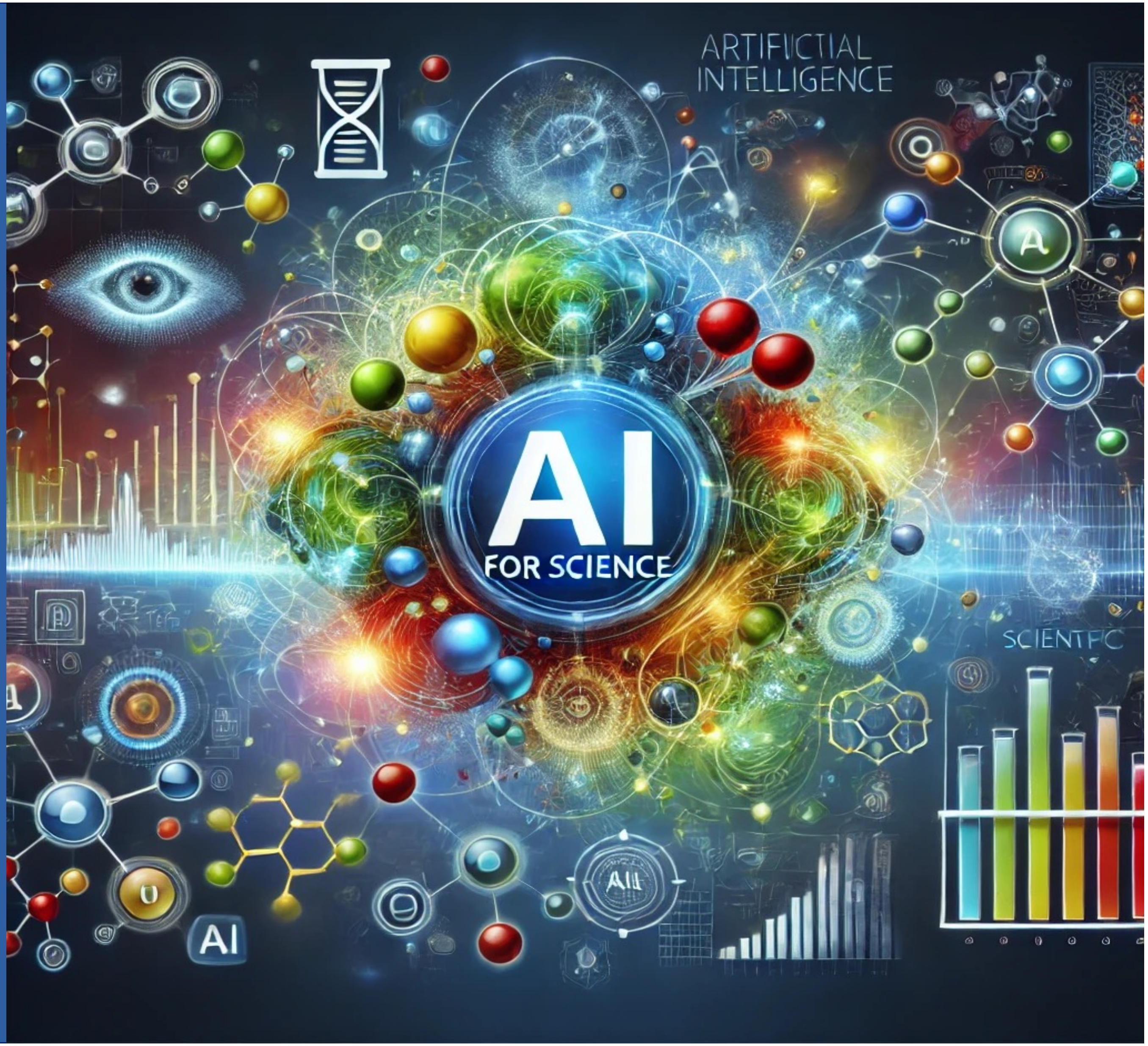
Oil & Water: https://arxiv.org/pdf/2405.15828

AI Engagement Across 20 Exemplar Fields





AI for Science & Benchmarking



Bird's Eye Point of View: Key Areas

Inverse Designs

Energy, Proteins & Polymers

Autonomous Discovery

Materials, Chemistry, Biology
Light & Neutron Sources

Surrogate Modelling for HPC

Climate Ensembles
Exascale apps with surrogates

Software Engineering and Programming

Code generation, Code Translation, Optimisation, Quantum Computing

Prediction and Control of Complex Engineering Systems

Accelerators, Telescopes, Buildings, Cities Reactors, Power Grid, Networks

Foundation Models for Science

Hypothesis Formation, Math Theory and Modeling Synthesis

https://www.anl.gov/ai-for-science-report

Benchmarking & Why

Conventional Notion of Benchmarking?

benchmark

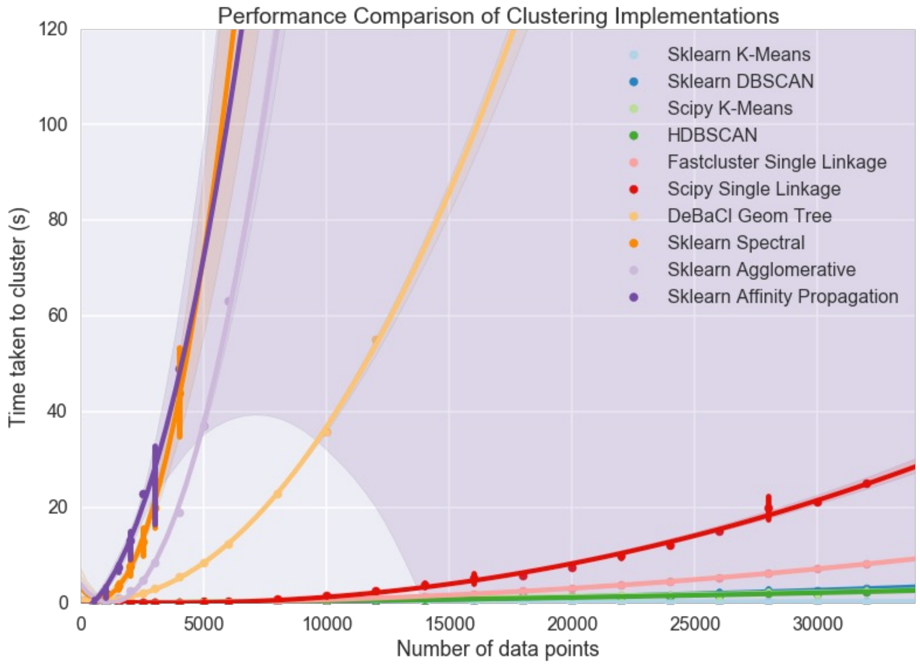
/ˈbɛn(t)ʃmɑːk/

verb

gerund or present participle: **benchmarking**

evaluate (something) by comparison with a standard.
"we are **benchmarking** our performance **against** external criteria"

- give particular results during a benchmark test.
  "the device should benchmark at between 100 and 150 MHz"



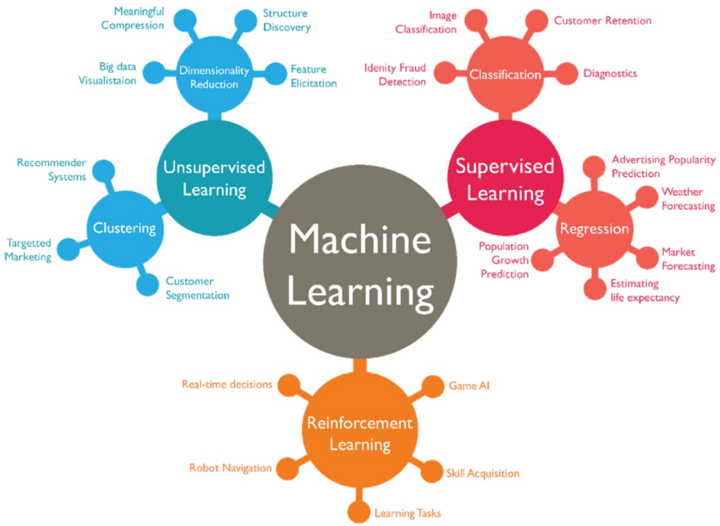
AI Benchmarking

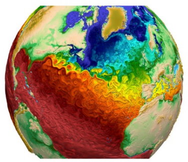
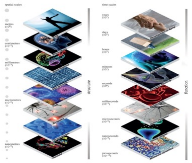
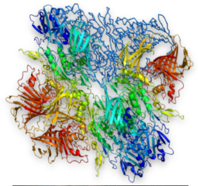
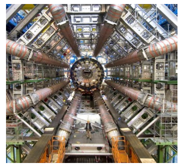
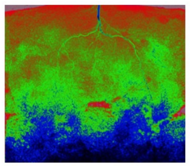
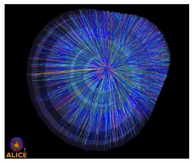
- Benchmarks for assessing
  - AI systems
  - AI models
  - AI frameworks

Challenging Space

- Developing an overall understanding of AI/ML methods is a significant challenge!
  - Too many ML methods!
  - Too many problems!, and
  - Too many systems!

AI Labs.tw	Alibaba	AMD	Andes Technology	Aon Devices	Arm	Baidu
BAAI	Cadence	Calypso AI	Centaur Technology	Cerebras	Ceva	Cirrus
Cisco	Code Reef	Cray	CTuning Foundation	Dell	Dividiti	DDN Storage
Edgify	Enflame Tech	Esperanto	Facebook	FuriosaAI	Google	Groq




Why

- Historically performance (comparison)

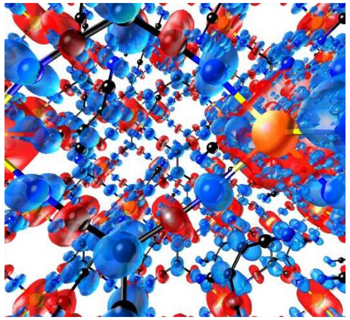
Time-focused metrics (training time, inference time)

- Evaluation different ML techniques
- Evaluation of alternative techniques

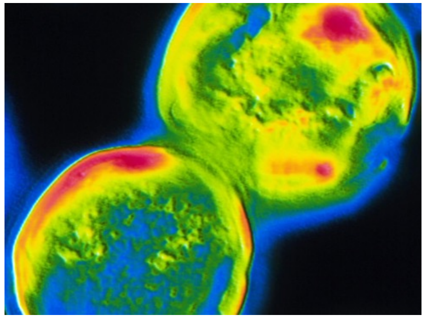
Domain-Specific Metrics (PSNR,IOUs)

Agenda I: Scanning Benchmarks

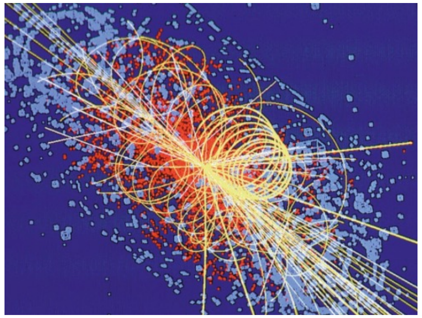
- *Systematically* study / consult multiple domains of sciences



Material Sciences

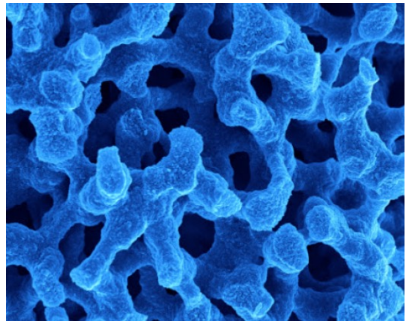


Environmental Sciences



Particle Physics

Astronomy



Life Sciences

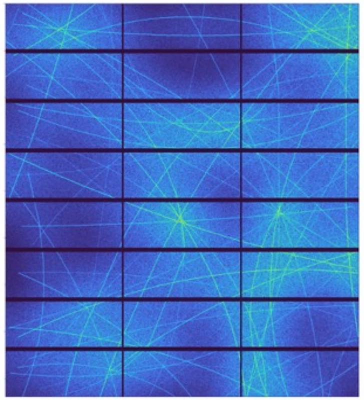
- Identify a set of benchmarks based on:

1
Worthy of solving using machine learning

2
Problem relies on a large and open dataset(s)

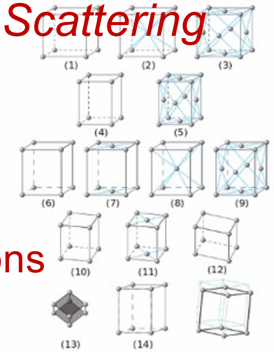
3
(*ideally*)
A reference implementation

- Outcome was a suite of benchmarks – The *SciML Benchmark Suite*

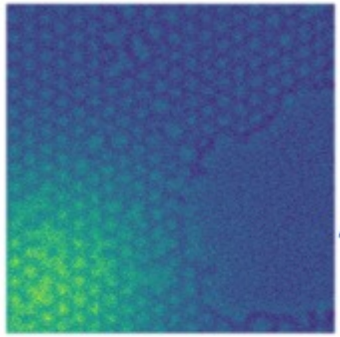


dms_scatter

Diffuse Multiple Scattering

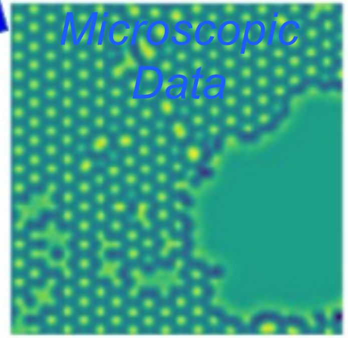


Identifying triple intersections in X-ray images
Dataset of 9GB with 8,060 images

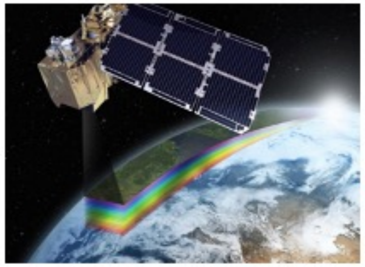


em_denoise

Removing Noise from Electron Microscopic Data

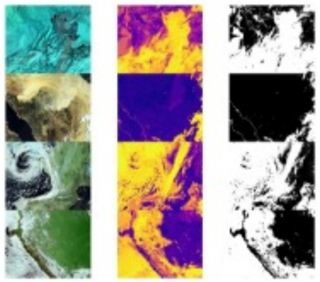


Removing noise of EM images
5GB dataset with 10,000 images of 256x256x1

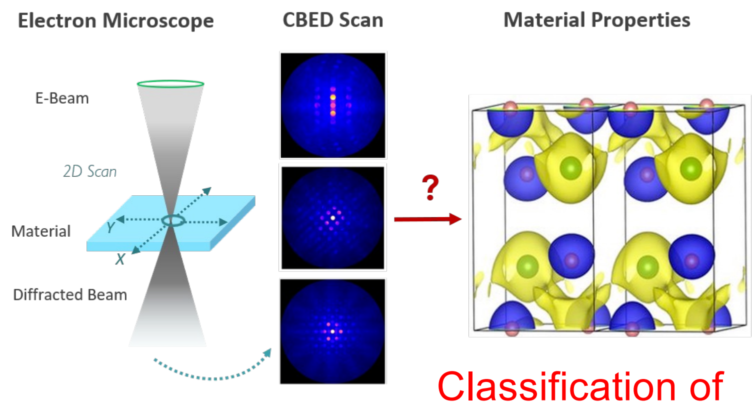


slstr_cloud

Cloud Masking



Identifying pixels that are cloud in satellite images



stemdl:

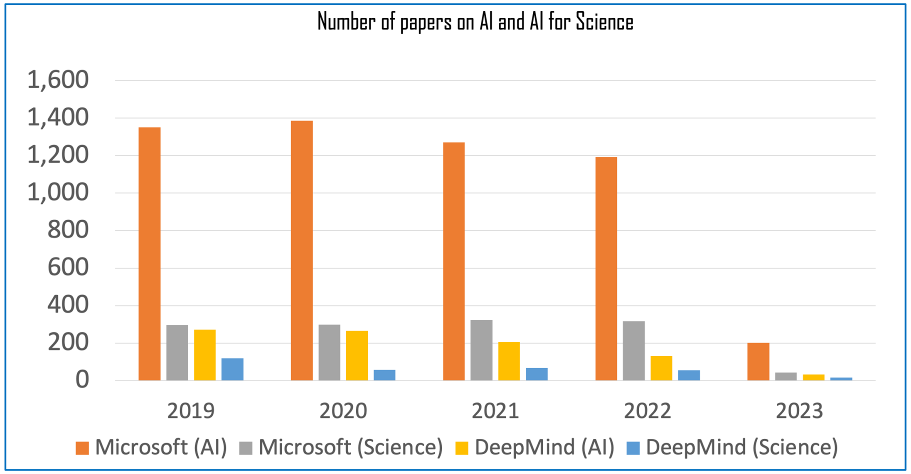
ImageNet for Solid Materials

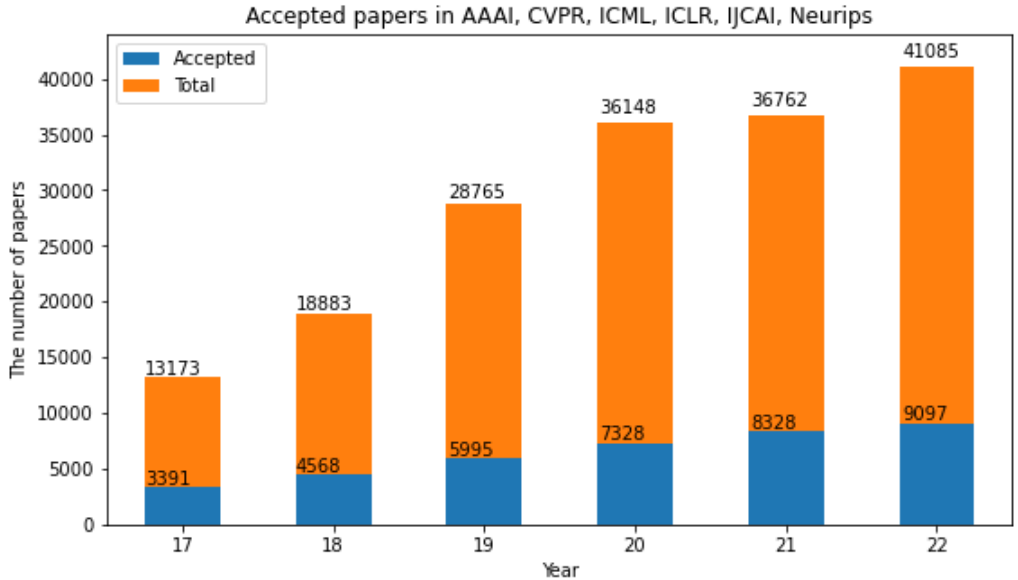
Classification of space groups, 40GB dataset with 230 spacegroups

Why it is / was not practical

- Staying relevant and up-to-date is a serious challenge! (~300 papers/w)

- Surveying the landscape is a difficult job let alone evaluating them





Agenda II: Blueprints

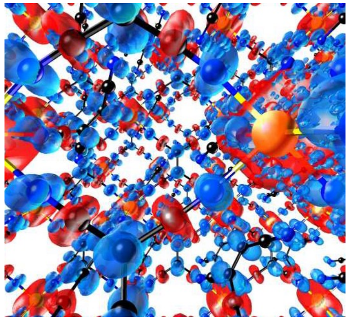
- Instead of individual benchmarks, identifying key family of science cases (across domains), and relevant core ML techniques is more useful

- These, we refer to as blueprints

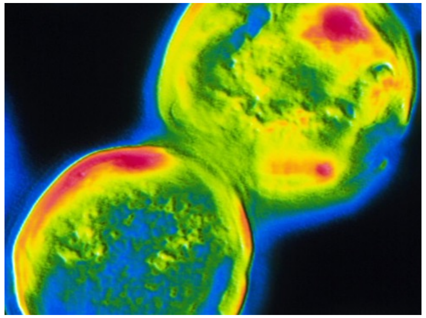
- A single benchmark blueprint will help
  - Developing solutions
  - Understand how ML techniques work
  - Representative of a suite of techniques (and variants therein)

Agenda II: Blueprints

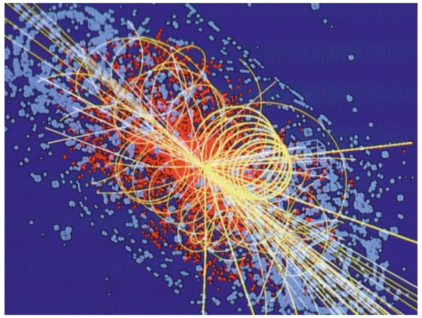
- *Systematically* study / consult multiple domains of sciences



Material Sciences

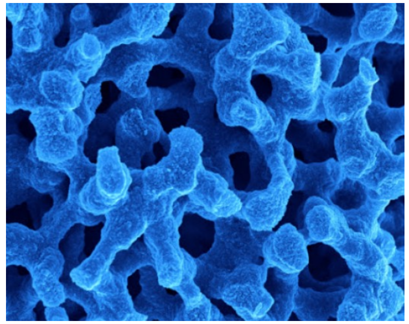


Environmental Sciences



Particle Physics

Astronomy



Life Sciences

- Identify a set of common use-cases or blueprints from each domain:

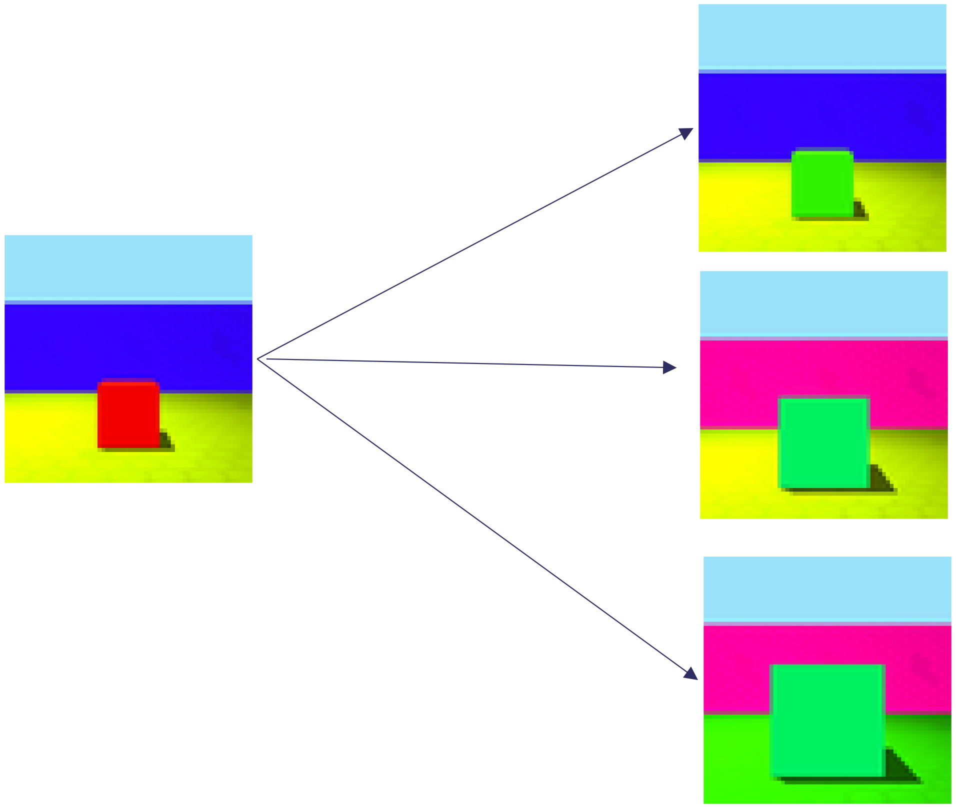
1. Flagship science problems
2. AI/ML Methods that would solve A family of problems
3. Multiple Open datasets
4. (*ideally*) A reference implementation

- Outcome: Benchmark Blueprints

Some Blueprints

Understanding Features

- Methods for disentangling or separating features of the data
- i.e., finding the underlying factors that explains the data



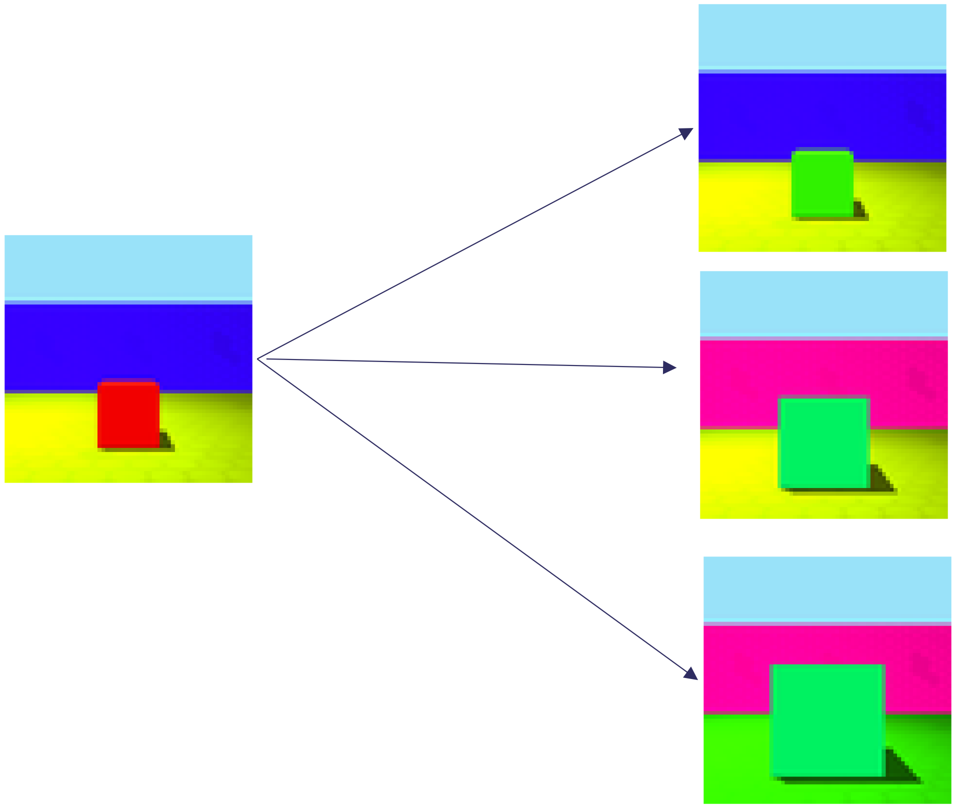
A small size of green cube with blue wall and yellow floor

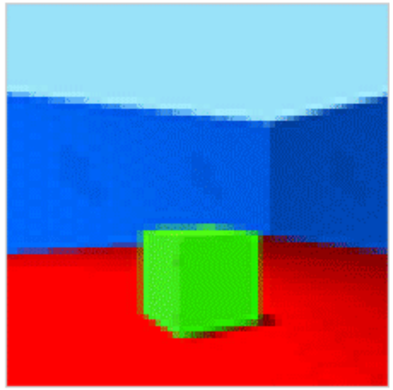
A medium size of green cube with pink wall and yellow floor

A large size of green cube with pink wall and green floor

Understanding Features

- Methods for disentangling or separating features of the data
- i.e., finding the underlying factors that explains the data





Cha & Thiyagalingam, ICLR 23

Understanding Features

- Methods for **disentangling** or separating features of the data
- i.e., finding the underlying factors that explains the data
- If these factors can be separately controlled, i.e., if we can get hold of disentangled representation of the data:
  - Better understanding of the data
  - Better generative models
  - Better inference
  - Provides minimal information for a given task
  - Etc.
- Solves the data-label problem
- That means, we can generate realistically good synthetic data for material science research

Disentanglement: Example Benchmarks

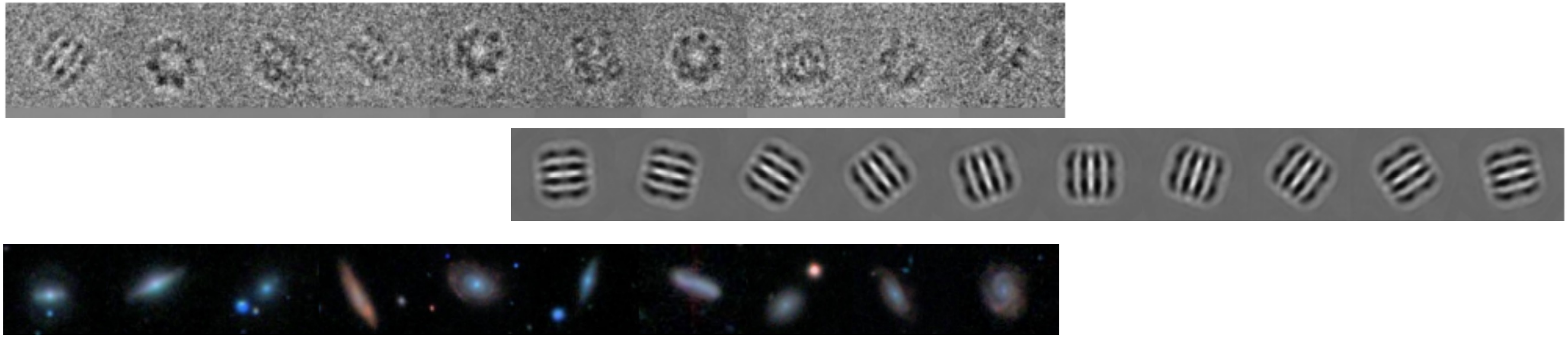
Table 2. Disentanglement scores for the 2D Arrow, 3D Airplane, 3D Teapots, 3D Shape, 3D Face Model and Sprites datasets

Datasets	2D Arrow		3D Airplane		3D Teapots		3D Shape		3D Face Model		Sprites	
Metrics/Models	**DAE**	DIPVAE	**DAE**	DIPVAE	**IB-GAN**	DAE	**FVAE**	DAE	**InfoGAN**	DAE	**DAE**	β-VAE
z-diff ↑	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
z-var ↑	0.85	0.96	1.00	0.96	1.00	1.00	1.00	0.93	0.95	1.00	0.90	0.77
dci-rf ↑	0.88	0.85	0.80	0.54	0.92	0.89	0.99	0.99	0.62	0.65	0.70	0.54
jemmig ↑	0.80	0.75	0.79	0.51	0.60	0.54	0.86	0.87	0.53	0.48	0.59	0.51
dcimig ↑	0.79	0.72	0.75	0.43	0.60	0.53	0.88	0.90	0.54	0.47	0.55	0.43
GF ($\times \frac{1}{100}$) ↓	0.30	2.55	0.19	7.66	0.10	0.002	0.20	0.0009	0.16	0.02	0.005	0.08

Example II:

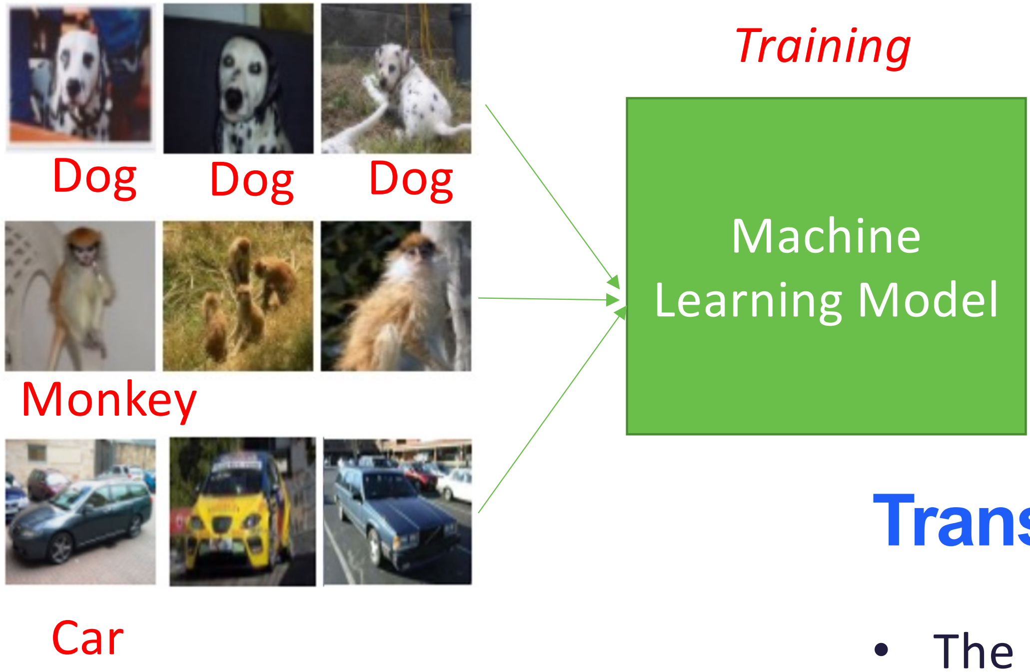
Inference on Rotated Images

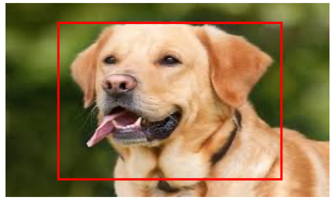
- Classify proteins / galaxies
- ML Models cannot infer rotations

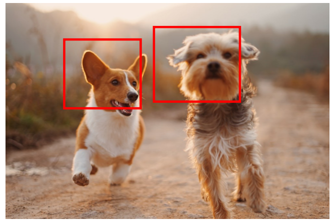


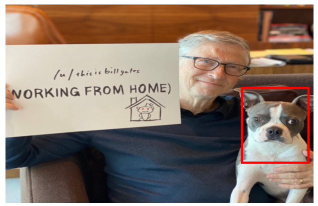
Invariance and Equivariance

- Model learns from supervised examples
- Example: learning to label images of dogs, cars and monkeys









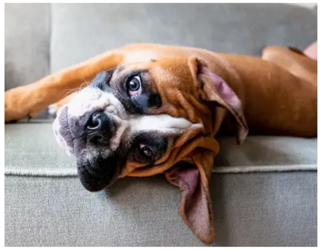
Translational Invariance

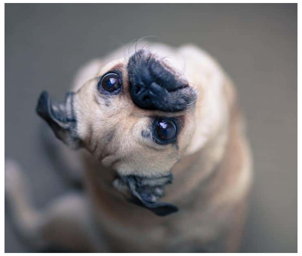
- The location does not matter

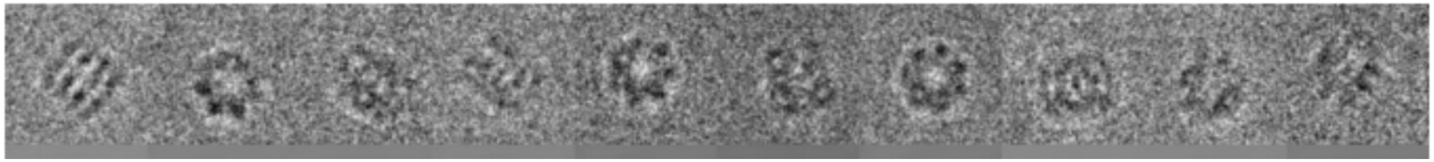
Rotational Invariance

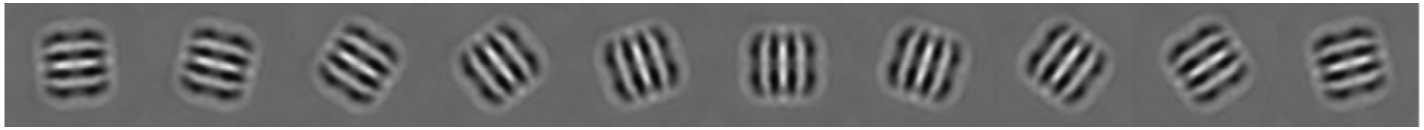
Inference on Rotated Images

- Uh-uh
- ML Models cannot infer rotations





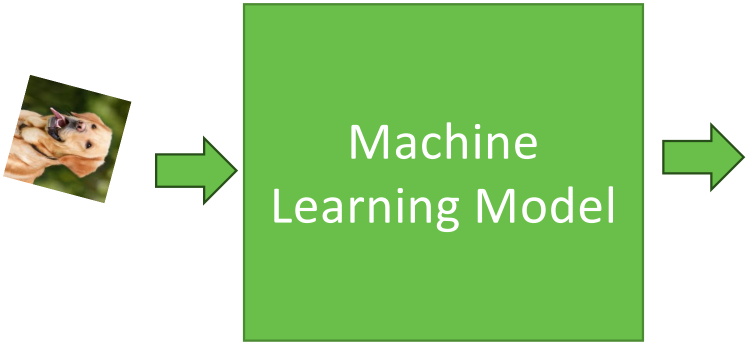


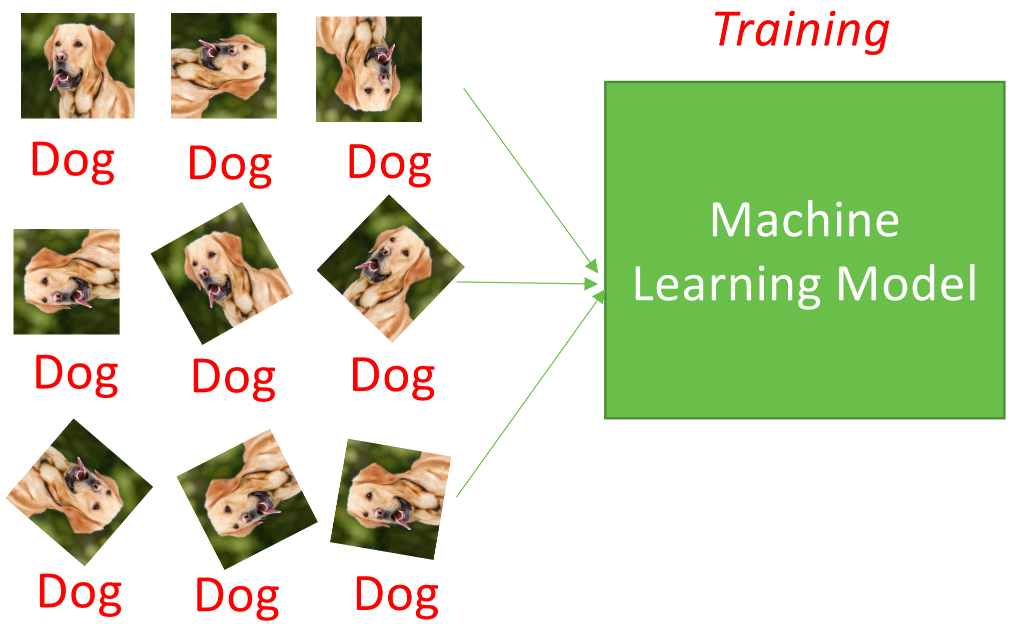


Rotational Invariance and Training

Rotational Invariance

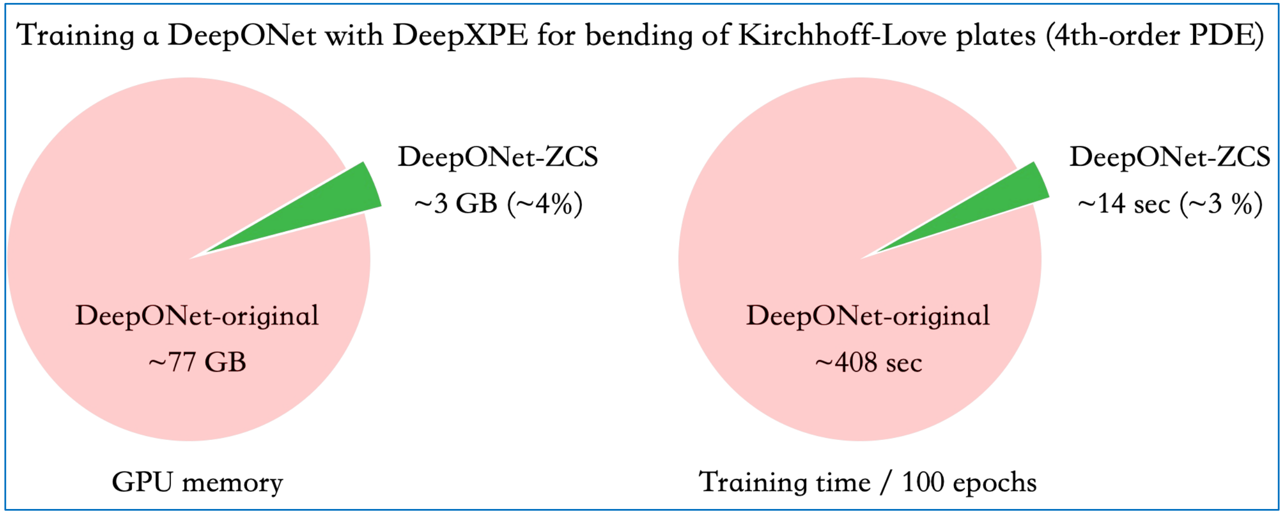
- Can train on rotated images
- But angles are discretised
- Not an elegant solution
  - Volume of data
  - Training times
  - Robustness of inferencing
  - Etc



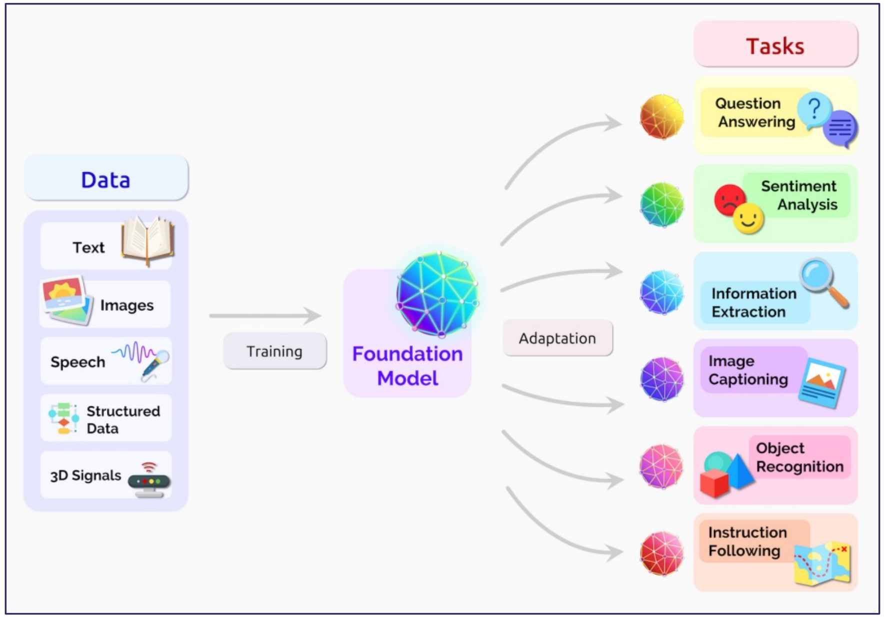


Other Examples

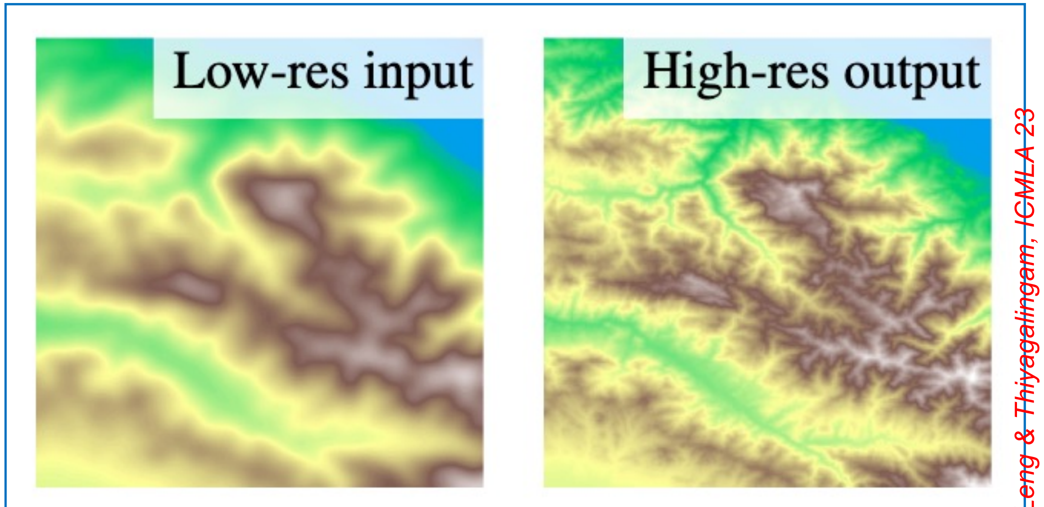
- LLMs for Science
- Super-resolution imaging
- Surrogates
- PINNs
- etc



K Leng et. al., , JCP 24



https://arxiv.org/pdf/2108.07258



(c) A patch for training, 192×192 pixels

Leng & Thiyagalingam, ICMLA 23

Conclusions

- AI has lot to offer for Science
- Benchmarking of various AI methods are crucial
- This not only shapes our understanding, but also the potential solutions
- But this is not an easy task
  - Demands intense international collaborations
  - Volunteers
  - Best practices
  - Science cases, datasets, reference implementations, and
  - Standards

Acknowledgements

- It is not easy to pull a significant effort like this without guidance, steer, and funding ☺
- Lot of people helped, but all these efforts would have been impossible without Tony, Rick, Arjun, Ian, Jamie, Rajeev, et. al.,
- People on the ground who actually executed the visions: More importantly, Juri, Jaehoon, Jason, Samuel, Kuangdai, Susmita, et.al.,
- Collaborators who trust us: APS, ANL, BNL, LBNL, DLS, ISIS, CLF, EPAC, PNNL, TIFR, & NERSC

Funding

UKRI

Department for Science, Innovation, & Technology

UK Atomic Energy Authority

iris

The Alan Turing Institute

ada lovelace centre

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC