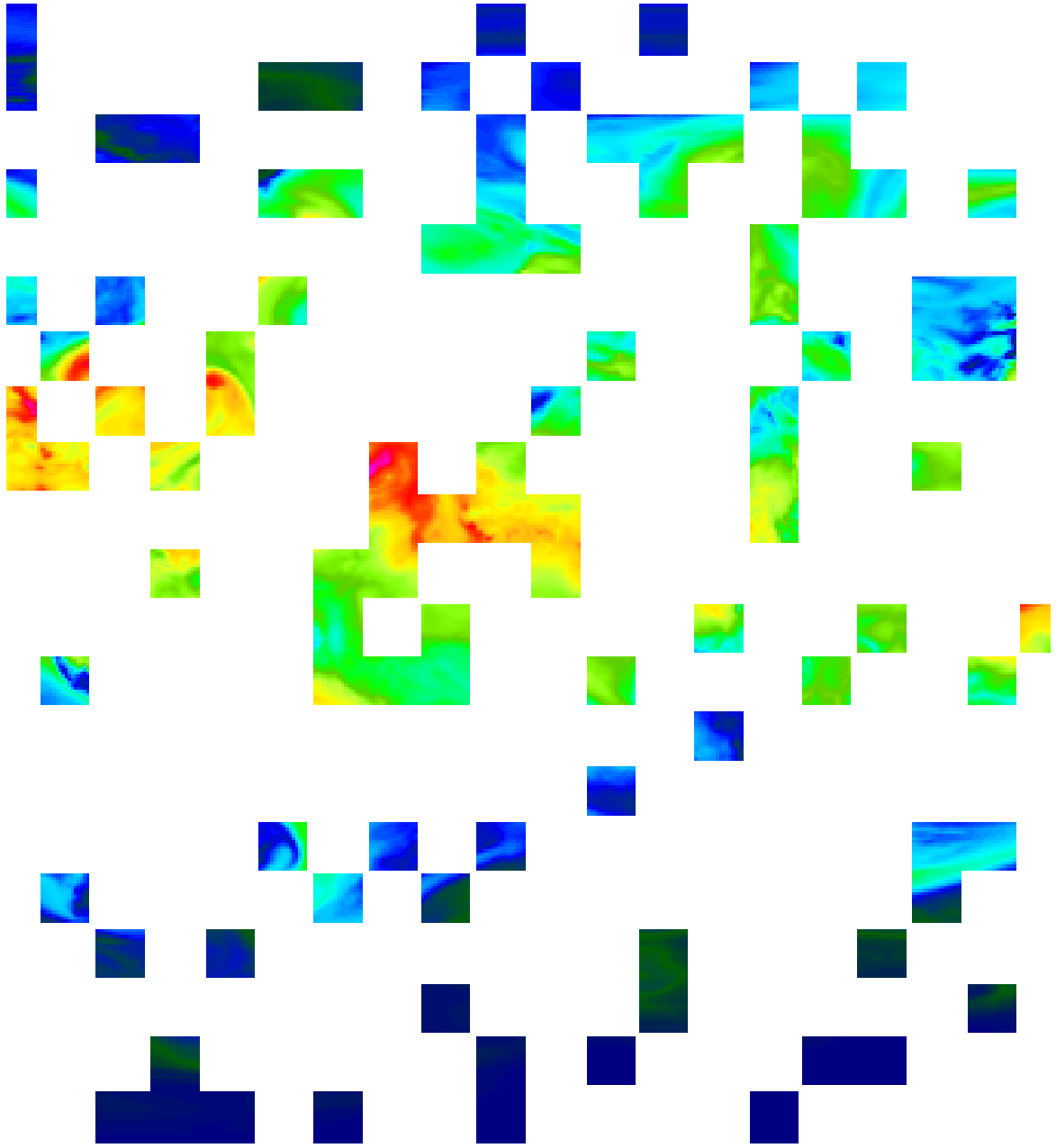
Dr. Johannes Schmude – IBM **Research**

Building Trustworthy AI for Weather and Climate

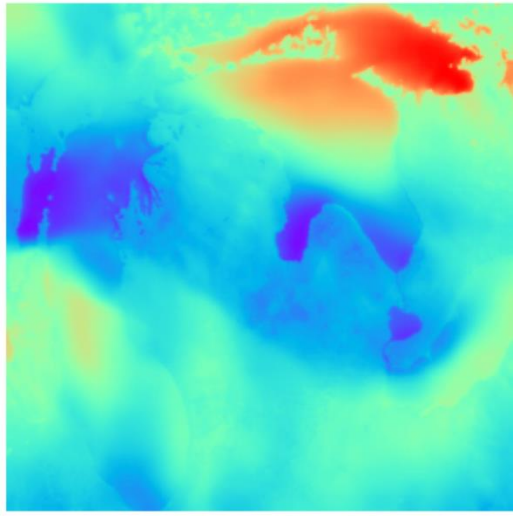
A View from the Private Technology Sector



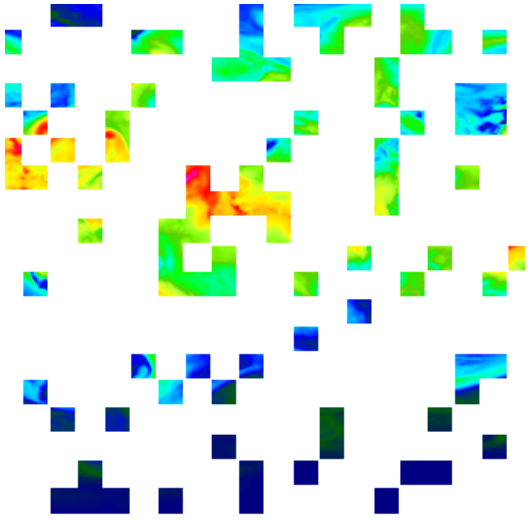
IBM

Public-Private Partnerships as the Driving Force for Large-Scale AI for Science

ECCC IBM
AI Models for Downscaling, Nowcasting and Operational Support

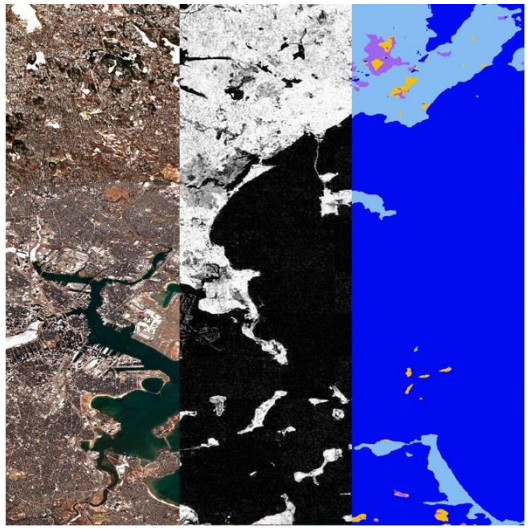


IBM NASA
Prithvi WxC AI Foundation Model for **Weather and Climate**

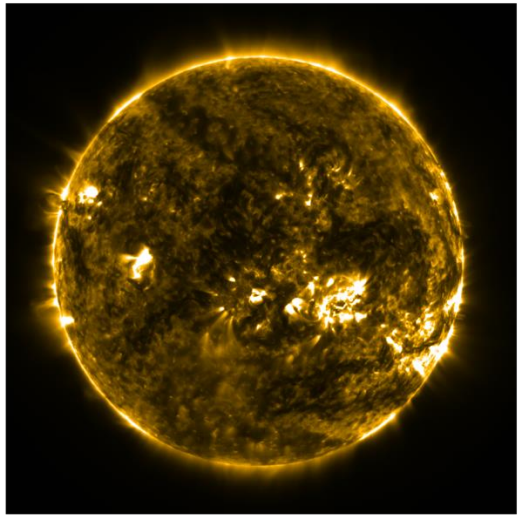


IBM NASA
Prithvi AI Foundation Models for **Earth Observation**

ESA IBM
TerraMind AI Foundation Model for **Earth Observation**



IBM NASA
Surya AI Foundation Model for **Heliophysics**

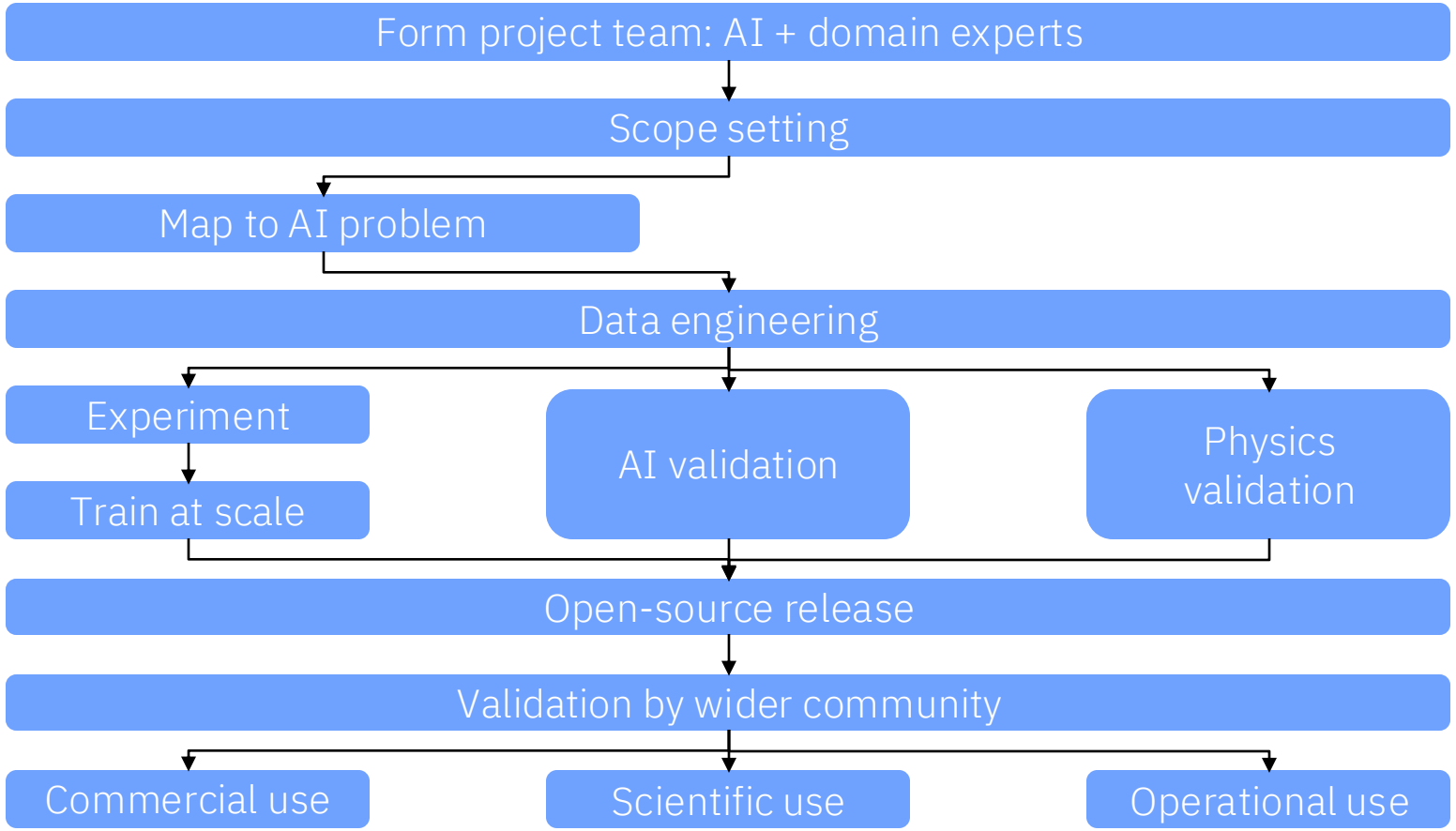


In our work, trust and explainability are not the result of a single test or an evaluation protocol.

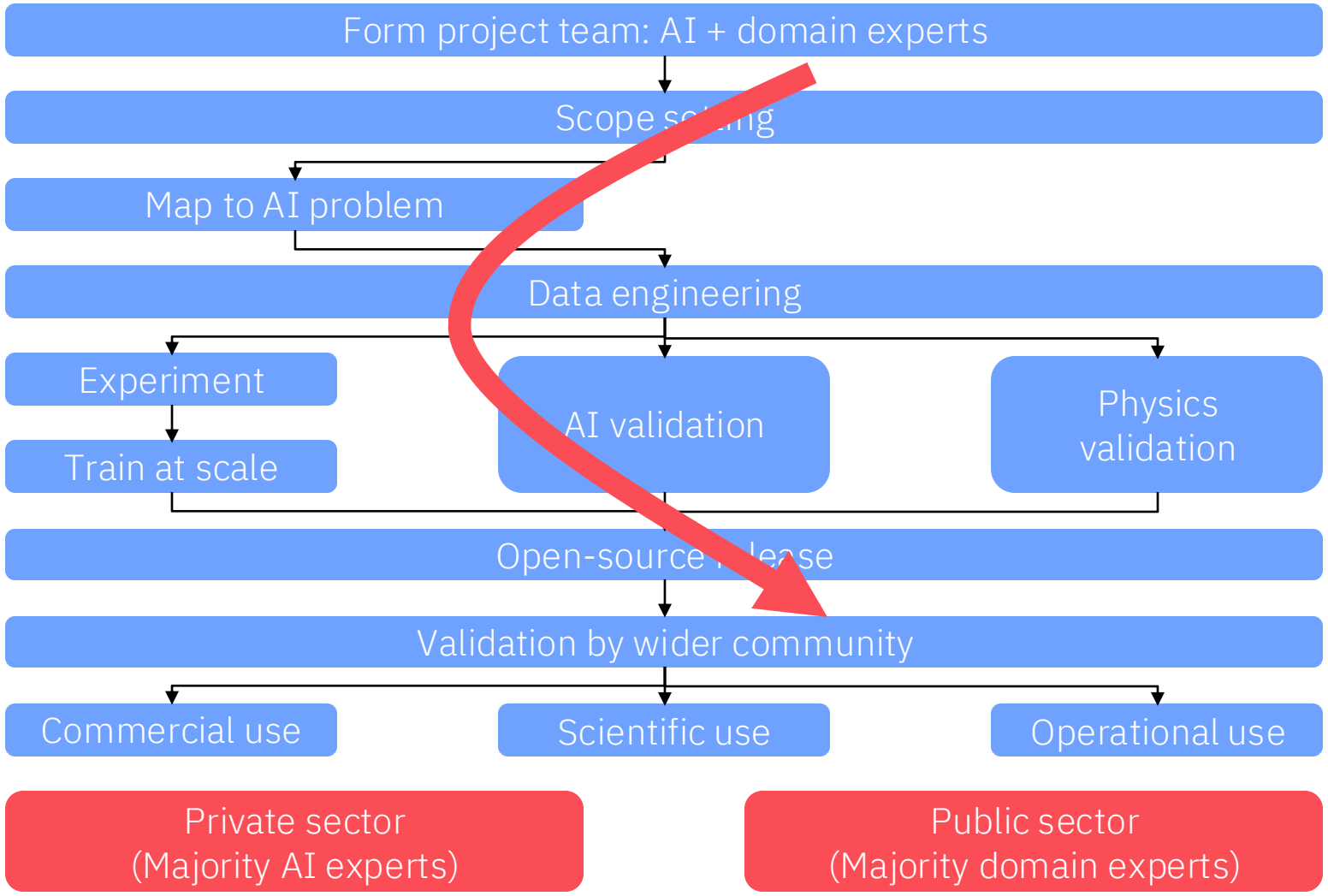
Explainability, validation and trust are the results of continuous collaboration and communication.

They are more processes than thresholds.

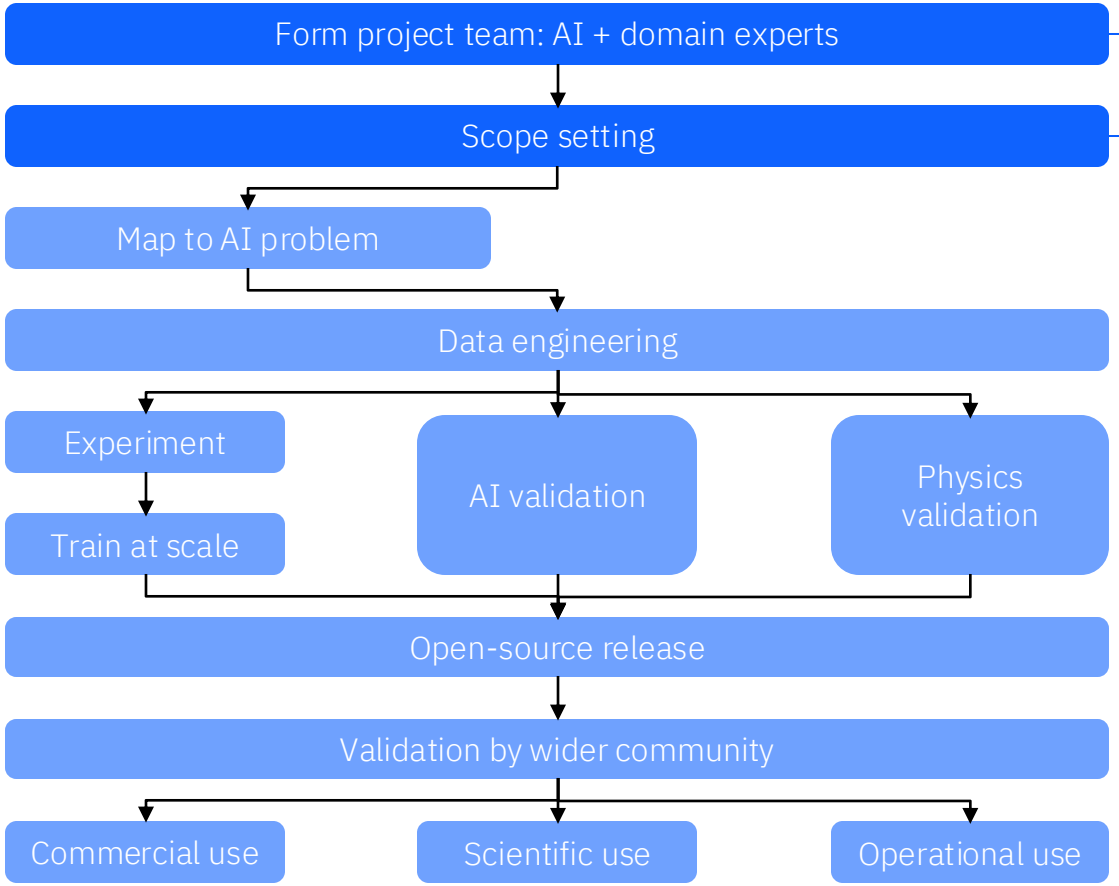
Trust as a Result of Public/Private Sector Workflows



Private/Public Tends to Map to AI/Domain Expertise



Successful Collaborations Require Institutional Alignment



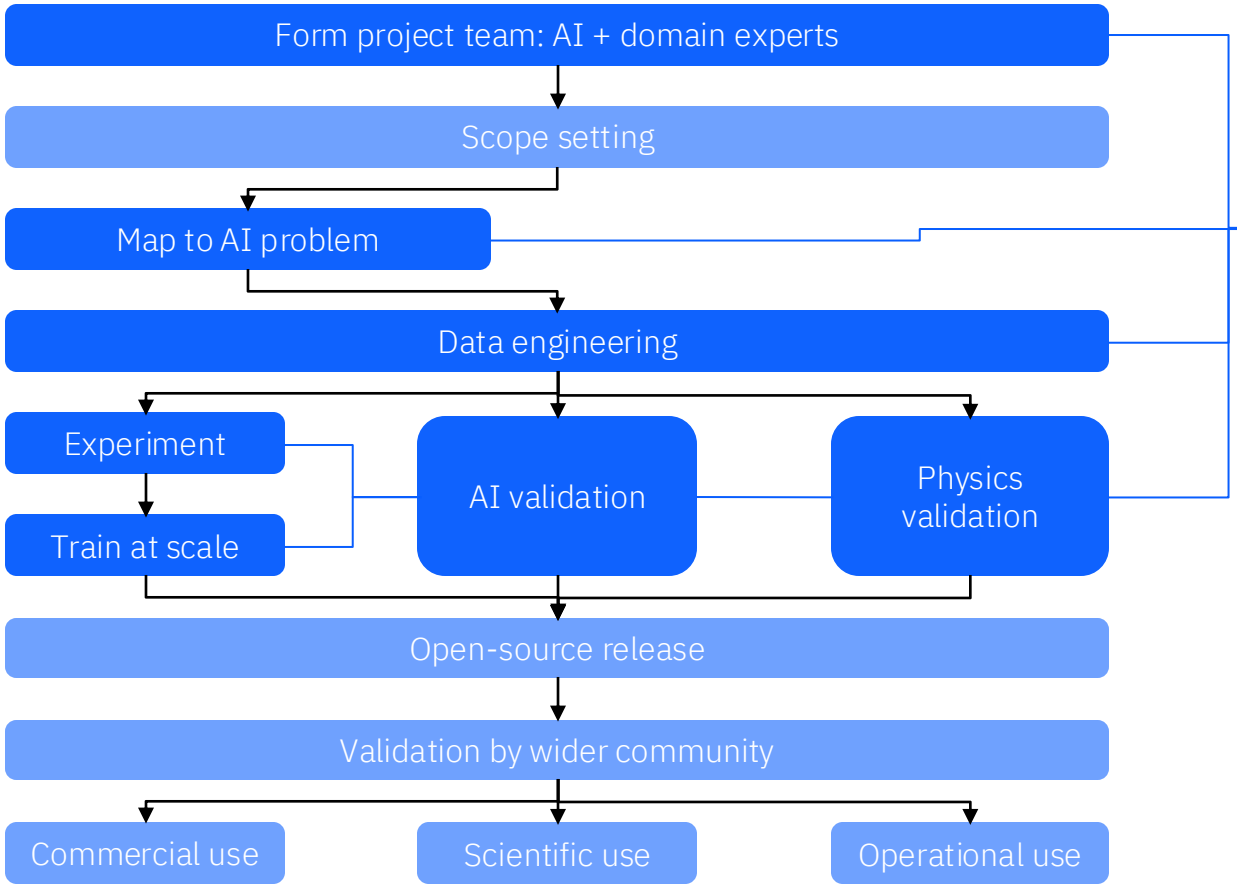
Institutional alignment is key. At executive/strategy level.

Strategy in the private sector can change quickly.

Private sector mandates and incentives include

- As a service
- Commercialization
    - of associated tech and tooling
    - of derived models
    - of model outputs
    - of models
- Corporate social responsibility
- Marketing
- Wider (AI) research

Interdisciplinary Projects Need Interdisciplinary Teams



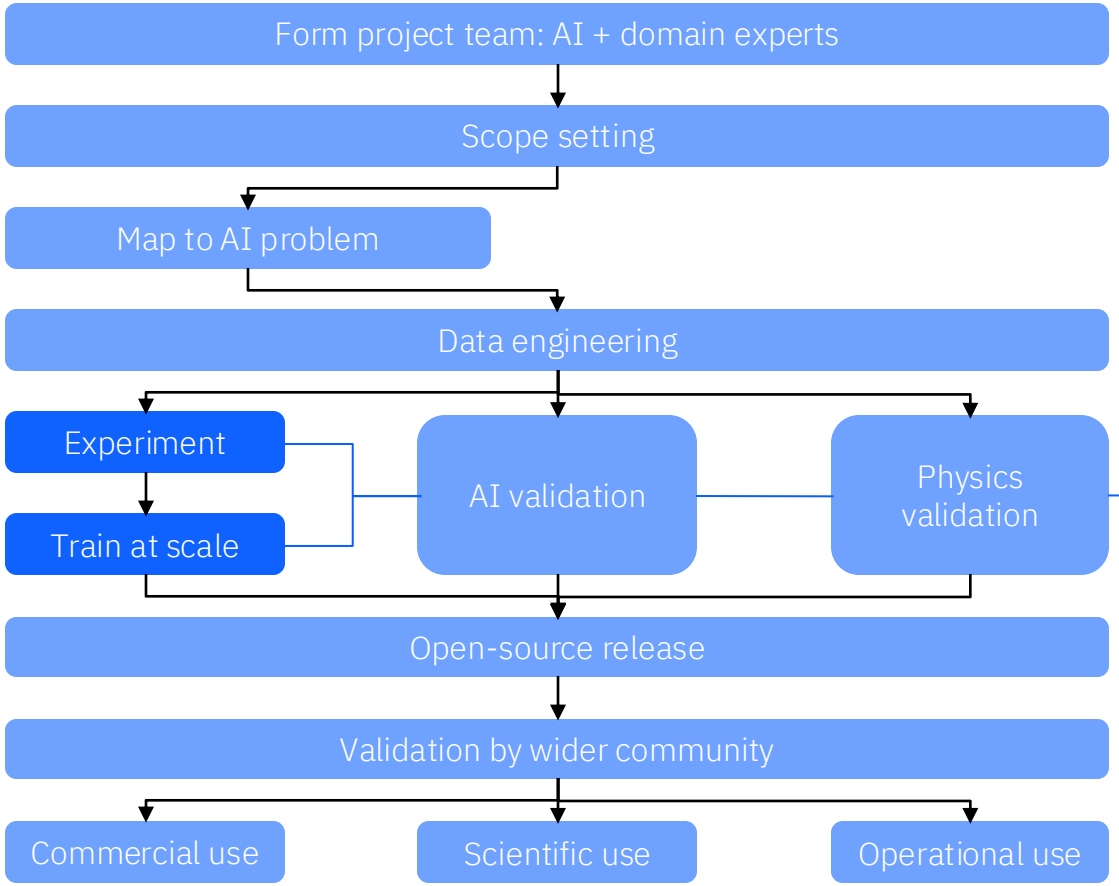
Avoid silos

- Splitting the team into AI experts and domain experts is the seed of mistrust.
- AI development needs clear, fast feedback loops.
- Domain experts need an interest and stake in the putative new AI model.
- Continuous validation leads to bi-directional knowledge transfer.

Organizations involved in Prithvi WxC

- **IBM** Research
- Earth Systems Science Center, **UAH**
- Department of Earth Systems Science, **Stanford**
- Department of Atmospheric Science, **CSU**
- National Center for Computational Sciences, **ORNL**
- **NASA** Goddard Space Flight Center
- **NASA** Marshall Space Flight Center
- **NASA** Headquarters

AI projects are Executed like Laboratory Work

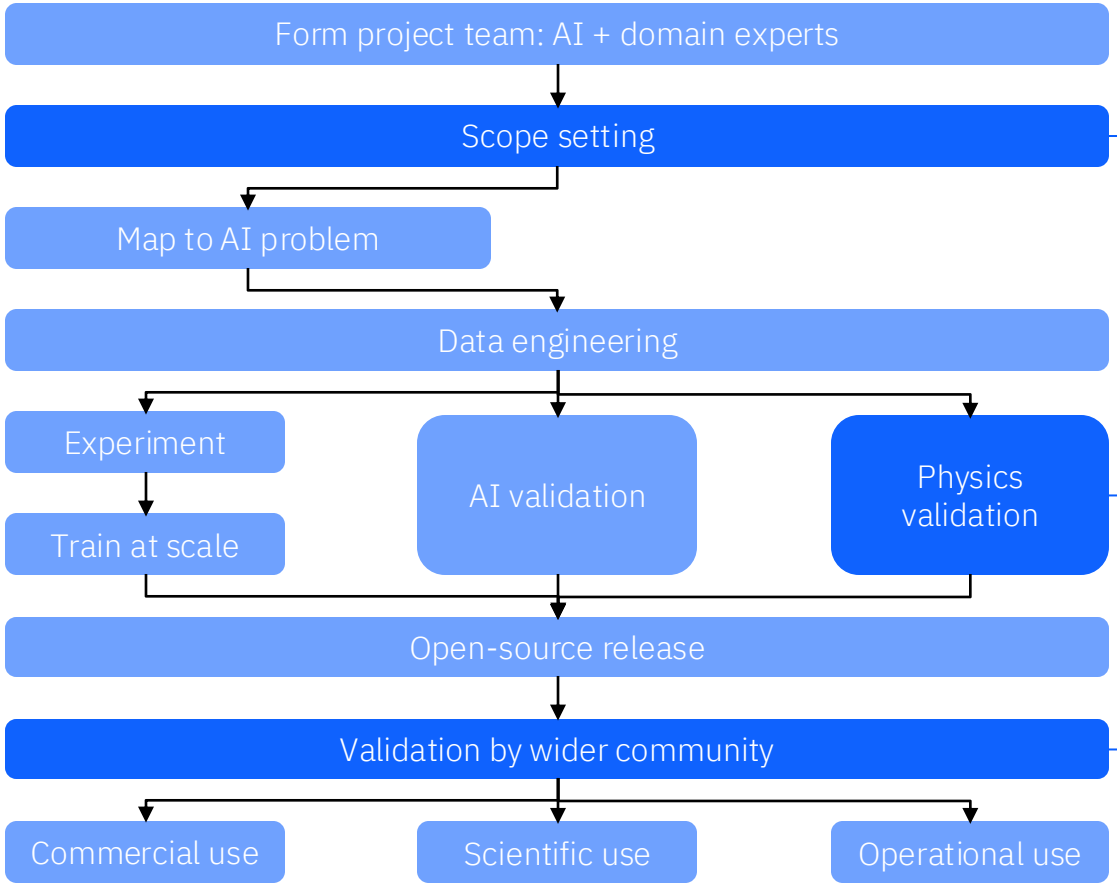


Experimentation phase akin to lab work
- Shared **lab diaries**, code artifacts, results.

Code and data validation
- In multi-organization, multi compute-cluster, multi-researcher collaborations, bugs and errors arise at the interfaces.

Following **best practices from lab work** leads to trust.

Explainability by Process Trumps Explainability by Design



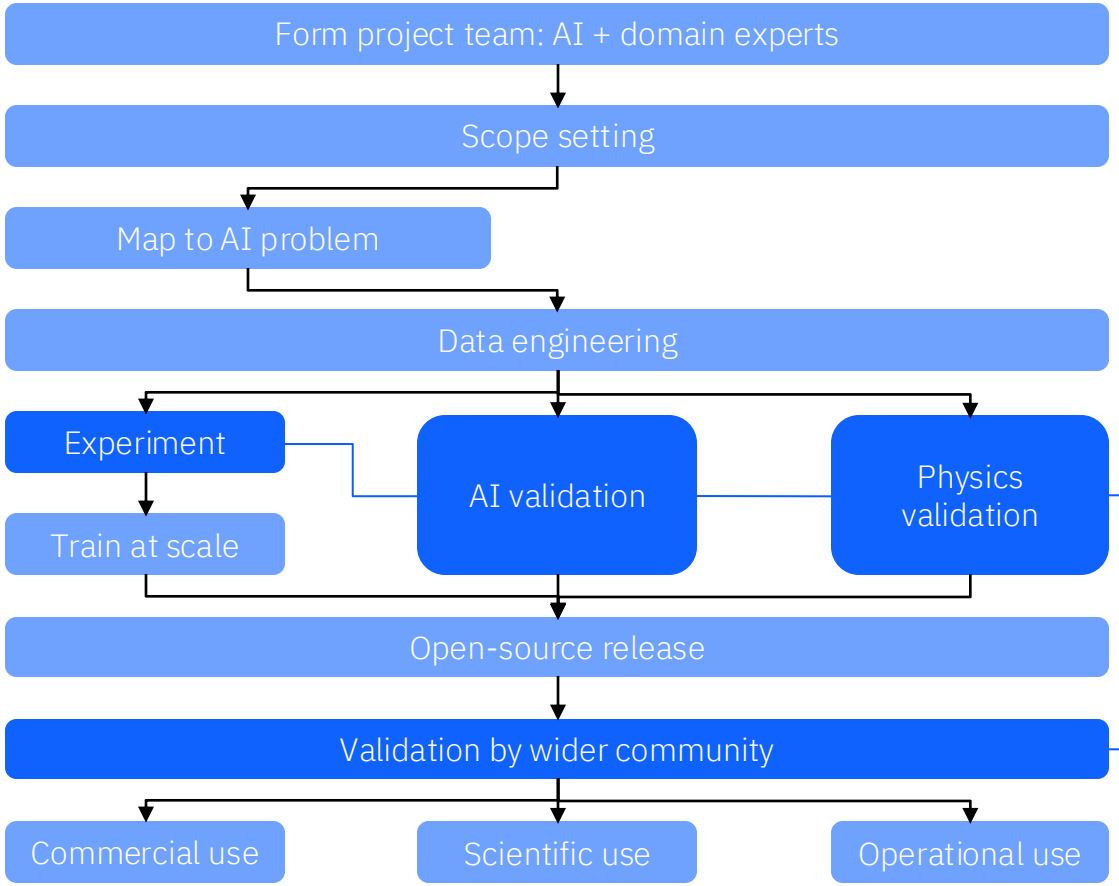
Tempting to bake explainability into scope and architecture.

Physics-inspired architectures are often not optimal.

At validation time, few people ask about physics-inspired training objectives. They ask about accuracy scores and operational risks. (This is only true in first approximation.)

It is better to start a (physics) validation as early as possible.

We need to Define What it Means for AI to be Explainable



The challenge of explainability is that there is **no threshold for something to "be physics"**.

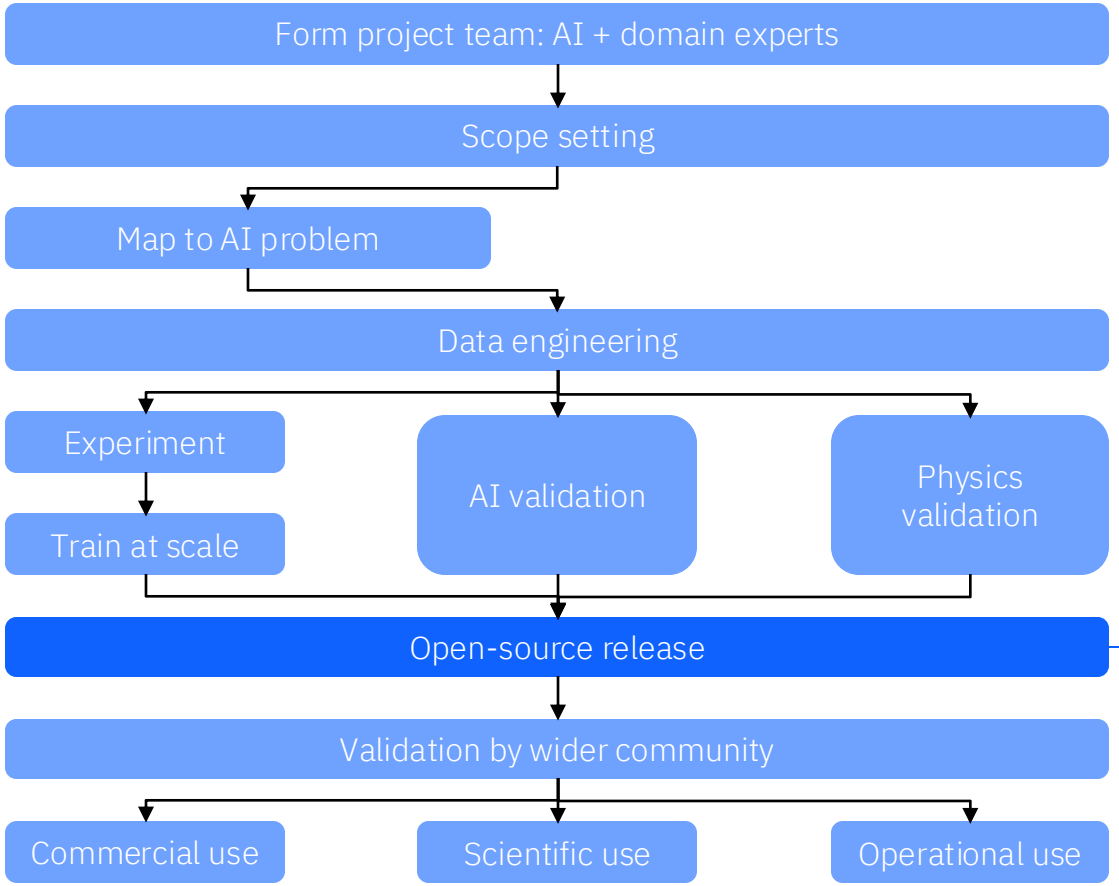
AI models are highly inspectable. You only need a laptop and a basic GPU to inspect the model.

AI models often **lack choke points of information** where code condenses into a mathematical equation or data into a known, measurable quantity.

Developing AI models means constructing such choke points and conducting conclusive experiments.

If we want a firmer notion of explainability, we need to define clear criteria.

The Reality of Open-Source: Open Models, Closed Training



Open-source fosters speed, collaboration, and innovation that creates better software.

A model is not just the model's code
- Model architecture (code)
- Training scripts (code)
- Training data (data)
- Model weights (data)

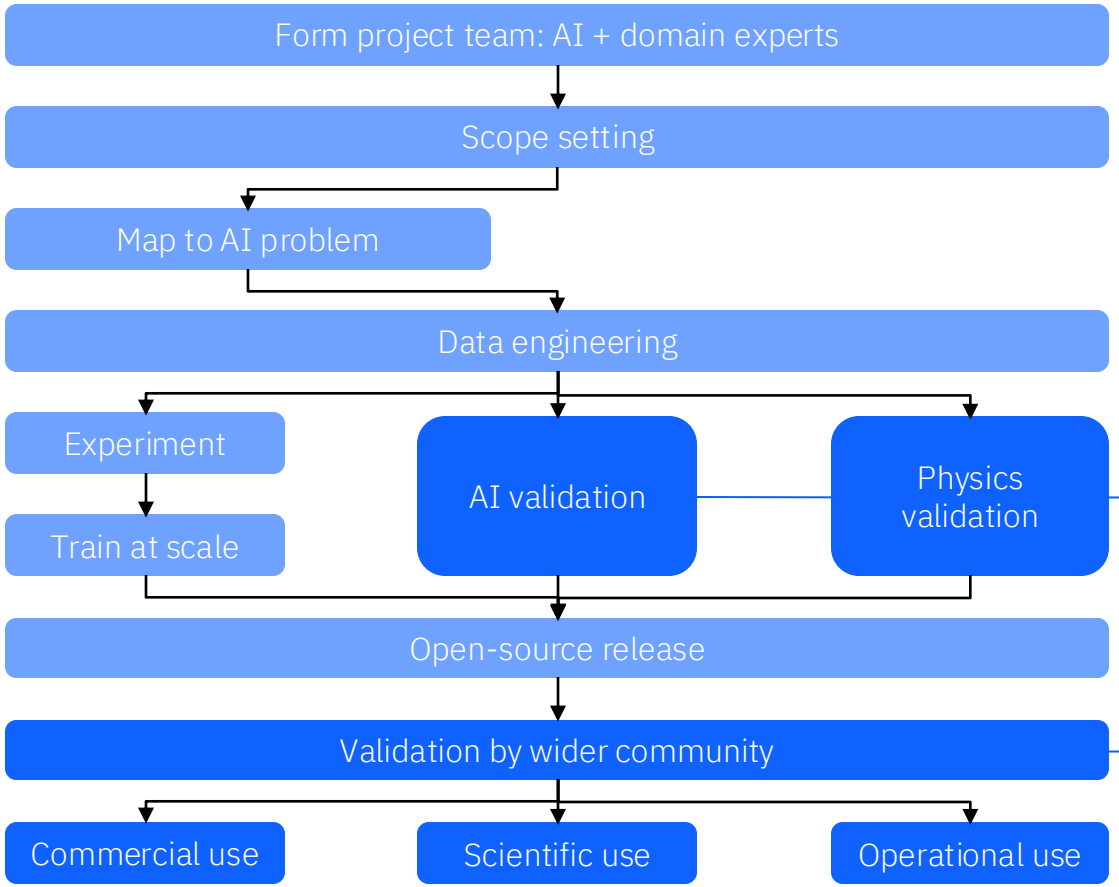
Model architecture: typically open-source
- All major architectures are in open source.
- You probably need 10-20% accuracy improvement to justify closed-source arch. This is unlikely.

Training scripts: often closed-source

Model weights often open source; sometimes not
- This can be due to data licenses.

To keep something closed, it has to be
- A significant improvement
- Hard to reproduce

Accelerating Community Validation via Tooling

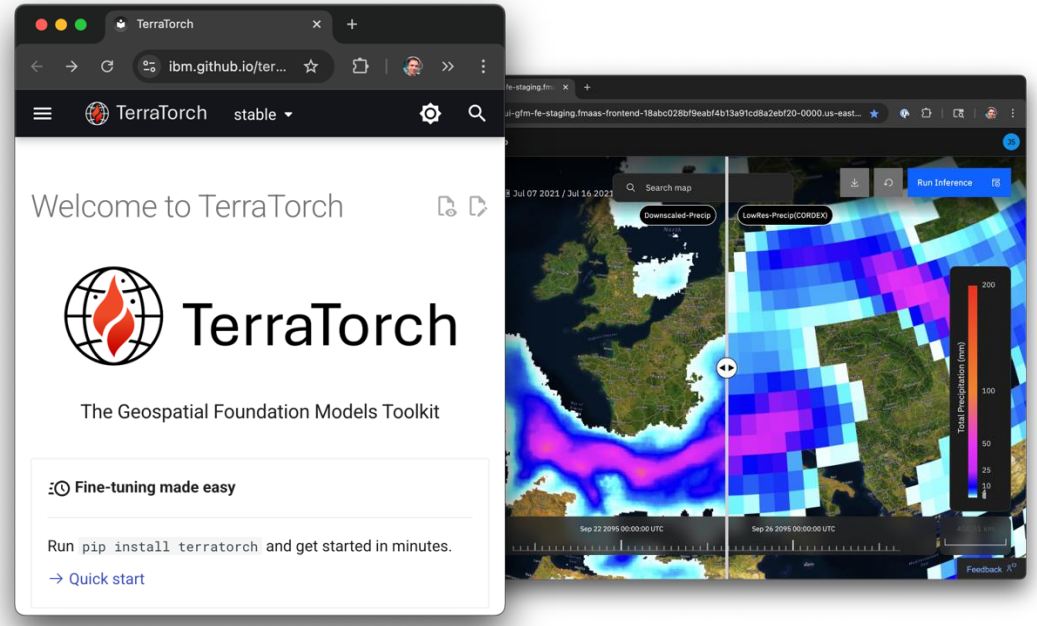


Mainstream AI is clad in benchmarks and state of the art (SOTA) reference scores allowing for (imperfect) comparison.

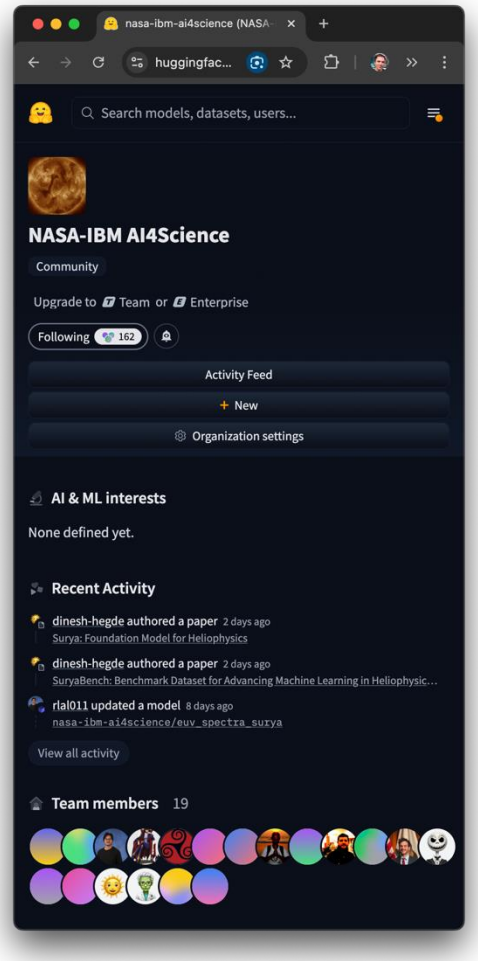
We validate from an AI and physics perspective.

The best validation is community validation.

We accelerate domain expert adoption/validation via tooling.



Partnerships Lead to Open-Source: Our FMs For Science are Available on Hugging Face



ESA IBM
TerraMind
Earth Observation

IBM NASA
Prithvi WxC
Weather and Climate

Prithvi EO (v1/v2)
Earth Observation

IBM NASA
Surya
Heliophysics

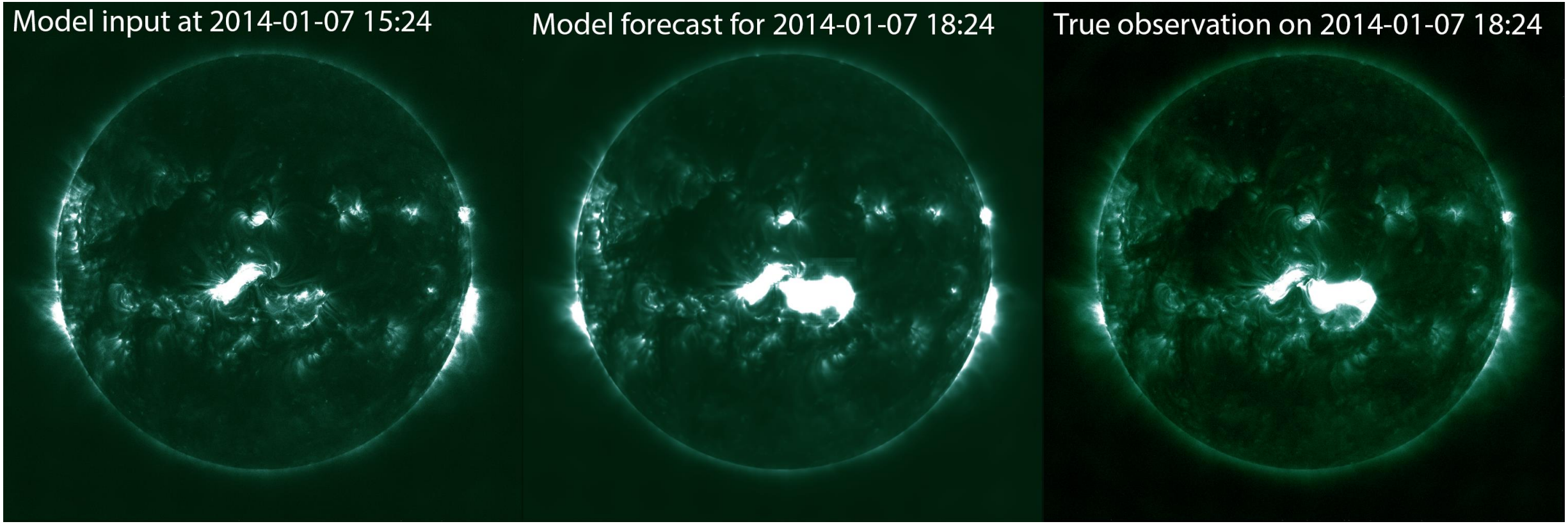
Partnerships Lead to SOTA Performance: TerraMind Outperforms All GeoFMs by at Least 3%

Model	BurnSr*	MADOS*	PASTIS	Sen1Fl11	FBP*	DEN*	CTM-SS	SN7*	AI4Farms*	Avg. mIoU	Avg. Rank
CROMA	82.42	67.55	32.32	<u>90.89</u>	51.83	38.29	49.38	59.28	25.65	55.29	6.61
DOFA	80.63	59.58	30.02	89.37	43.18	<u>39.29</u>	51.33	61.84	27.07	53.59	8.22
GFM-Swin	76.90	64.71	21.24	72.60	67.18	34.09	46.98	60.89	27.19	52.42	10.00
Prithvi 1.0 100M	83.62	49.98	33.93	90.37	46.81	27.86	43.07	56.54	26.86	51.00	11.00
RemoteCLIP	76.59	60.00	18.23	74.26	**69.19**	31.78	52.05	57.76	25.12	51.66	11.22
SatlasNet	79.96	55.86	17.51	90.30	50.97	36.31	46.97	61.88	25.13	51.65	10.67
Scale-MAE	76.68	57.32	24.55	74.13	<u>67.19</u>	35.11	25.42	**62.96**	21.47	49.43	11.44
SpectralGPT	80.47	57.99	35.44	89.07	33.42	37.85	46.95	58.86	26.75	51.87	10.11
S.-S12-MoCo	81.58	51.76	34.49	89.26	53.02	35.44	48.58	57.64	25.38	53.02	10.06
S.-S12-DINO	81.72	49.37	36.18	88.61	51.15	34.81	48.66	56.47	25.62	52.51	10.89
S.-S12-MAE	81.91	49.90	32.03	87.79	51.92	34.08	45.80	57.13	24.69	51.69	12.39
S.-S12-Data2Vec	81.91	44.36	34.32	88.15	48.82	35.90	54.03	58.23	24.23	52.22	10.72
UNet Baseline	**84.51**	54.79	31.60	**91.42**	60.47	**39.46**	47.57	<u>62.09</u>	**46.34**	57.58	4.89
ViT Baseline	81.58	48.19	38.53	87.66	59.32	36.83	44.08	52.57	<u>38.37</u>	54.13	10.28
TerraMindv1-B-single	<u>84.00</u>	65.01	<u>40.80</u>	90.32	–	–	52.66	59.71	27.71	–	–
TerraMindv1-B	82.42	<u>69.52</u>	40.51	90.62	59.72	37.87	**55.80**	60.61	28.12	<u>58.35</u>	<u>3.94</u>
TerraMindv1-L	82.93	**75.57**	**43.13**	90.78	63.38	37.89	<u>55.04</u>	59.98	27.47	**59.57**	**3.44**

Table 6. Performance evaluation of TerraMind across nine benchmark datasets using the PANGAEA evaluation protocol. Higher mIoU values (↑) and lower rank values (↓) are reported. The best model per column is highlighted in bold and the second best is underscored. We indicate unimodal datasets with *, demonstrating that TerraMind outperforms other benchmarks in both unimodal and multimodal settings. All foundation model encoders are frozen, while U-Net and ViT baselines are trained from scratch for each specific task.

Source: https://arxiv.org/pdf/2504.11171

Partnerships Lead to Science:
Surya is the First Model to “Visually” Predict a Solar Flare



Solar flare prediction had been a binary “yes” / “no” task: “Will there be a solar flare in the next 24 hours?”

It was not clear that raw sensor data would contain enough information to model and predict these dynamics.