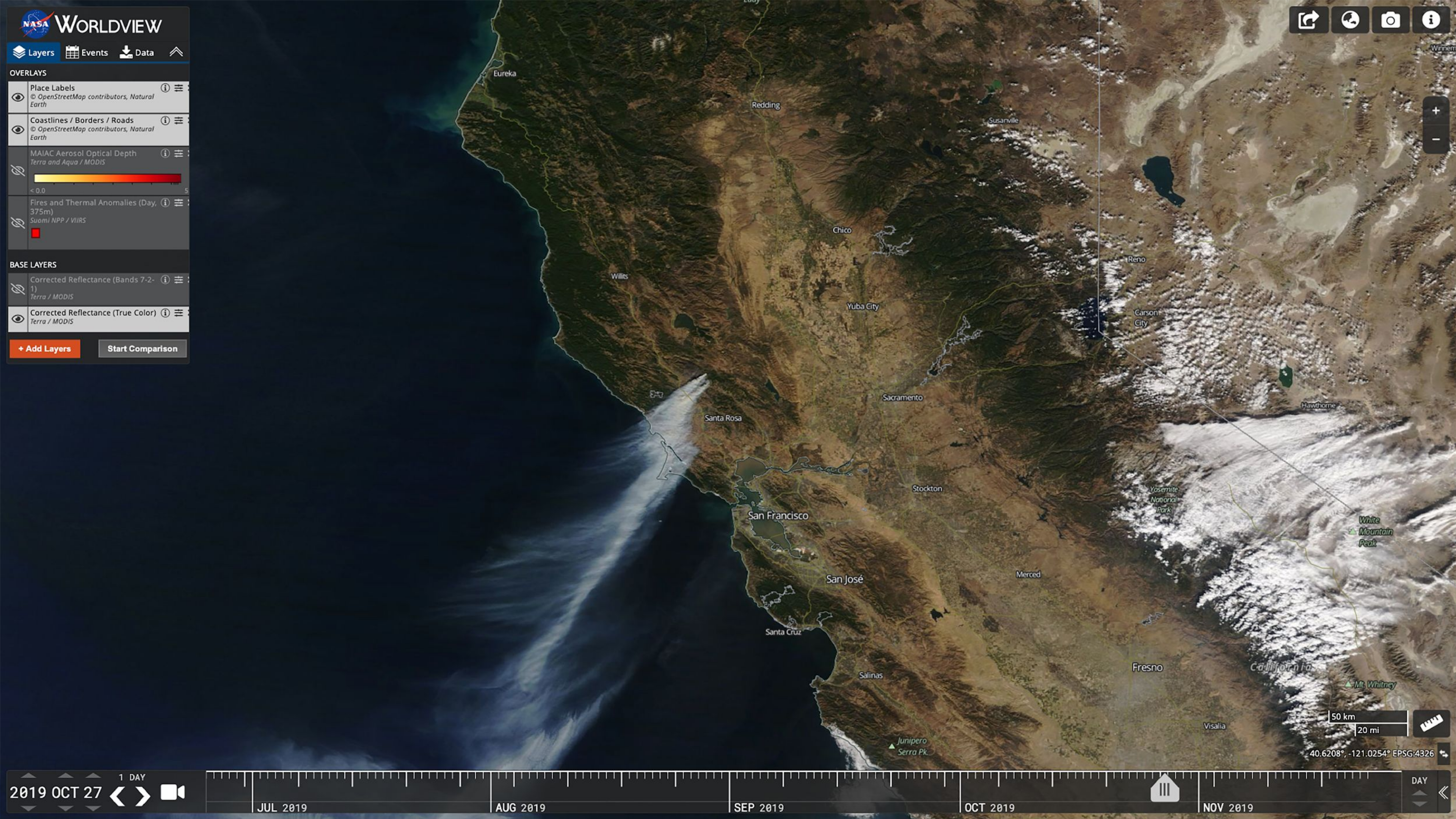
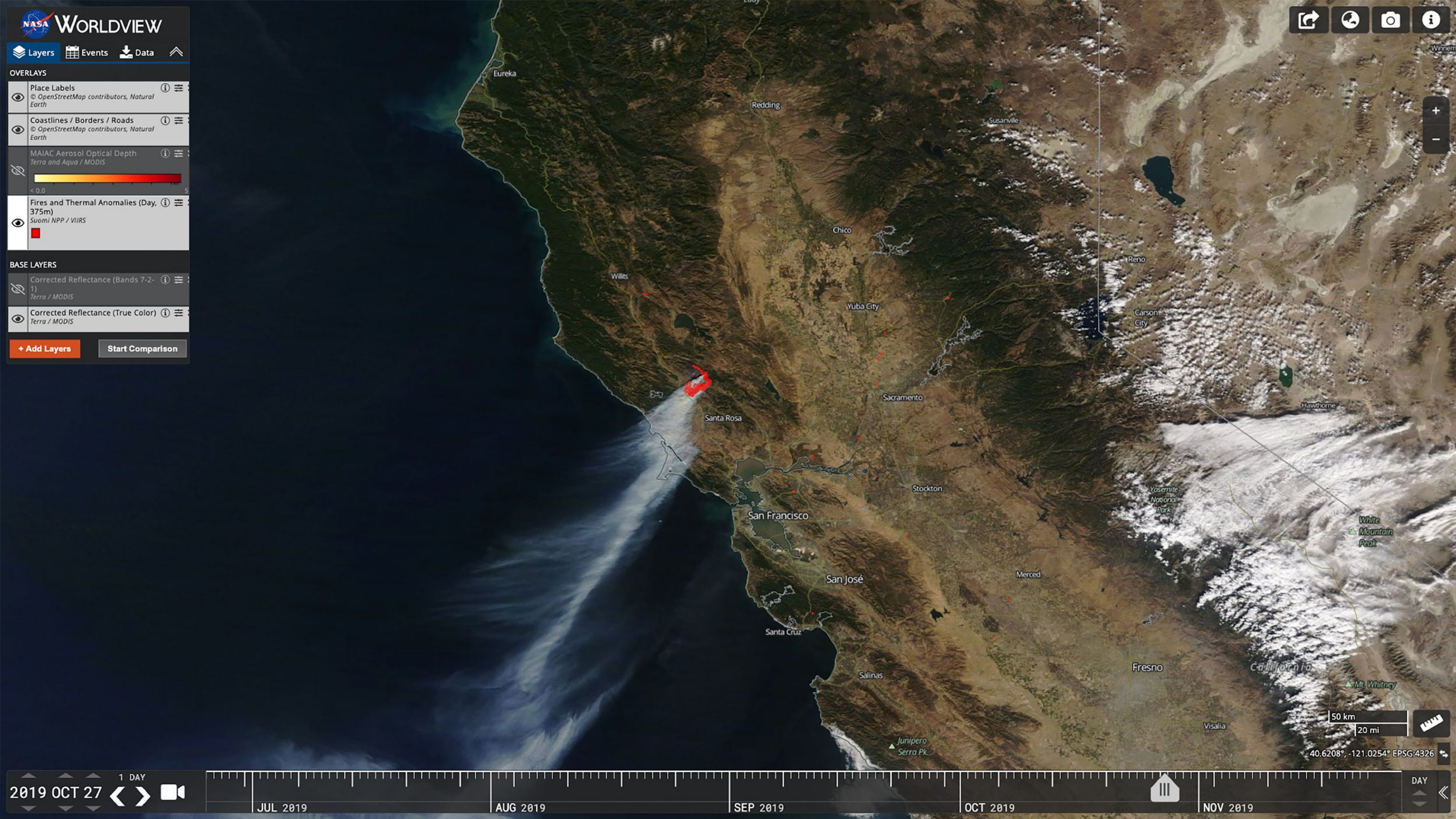
AI & Science: NASA Perspective

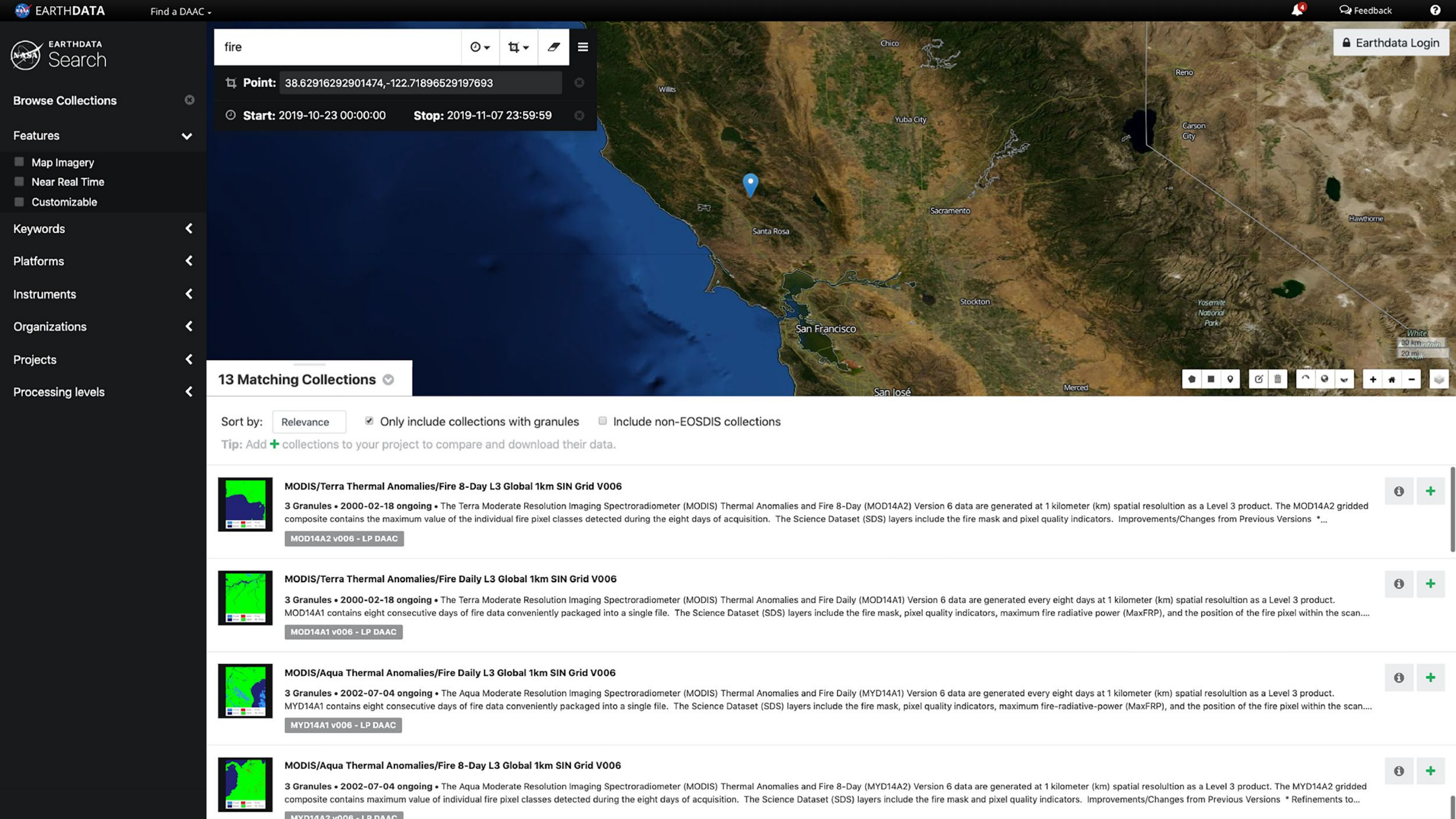
Manil Maskey, PhD

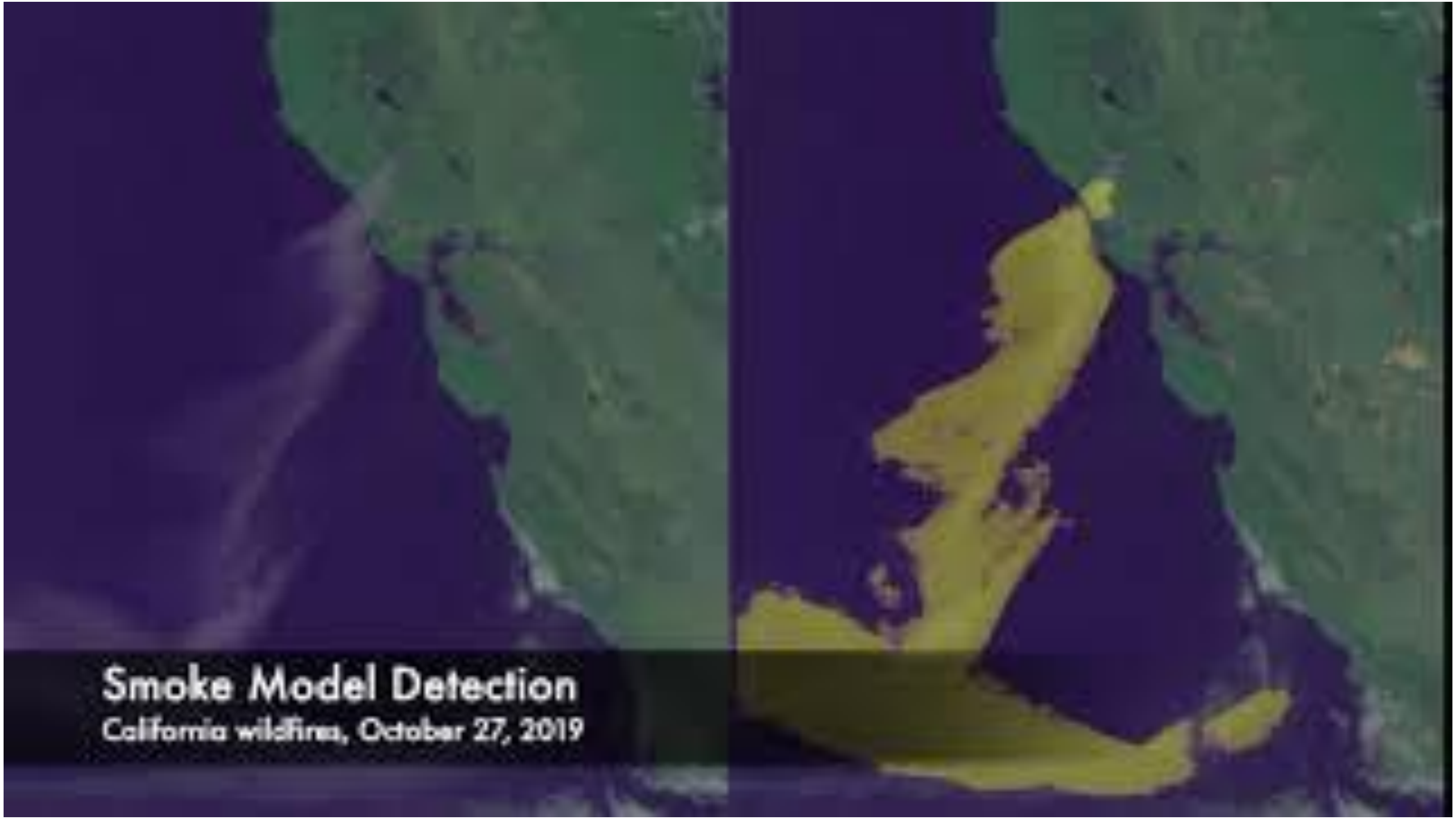
Senior Research Scientist
Data Science & Innovation Lead
NASA MSFC
NASA SMD/HQ

2024 AI Symposium
February 21, 2024

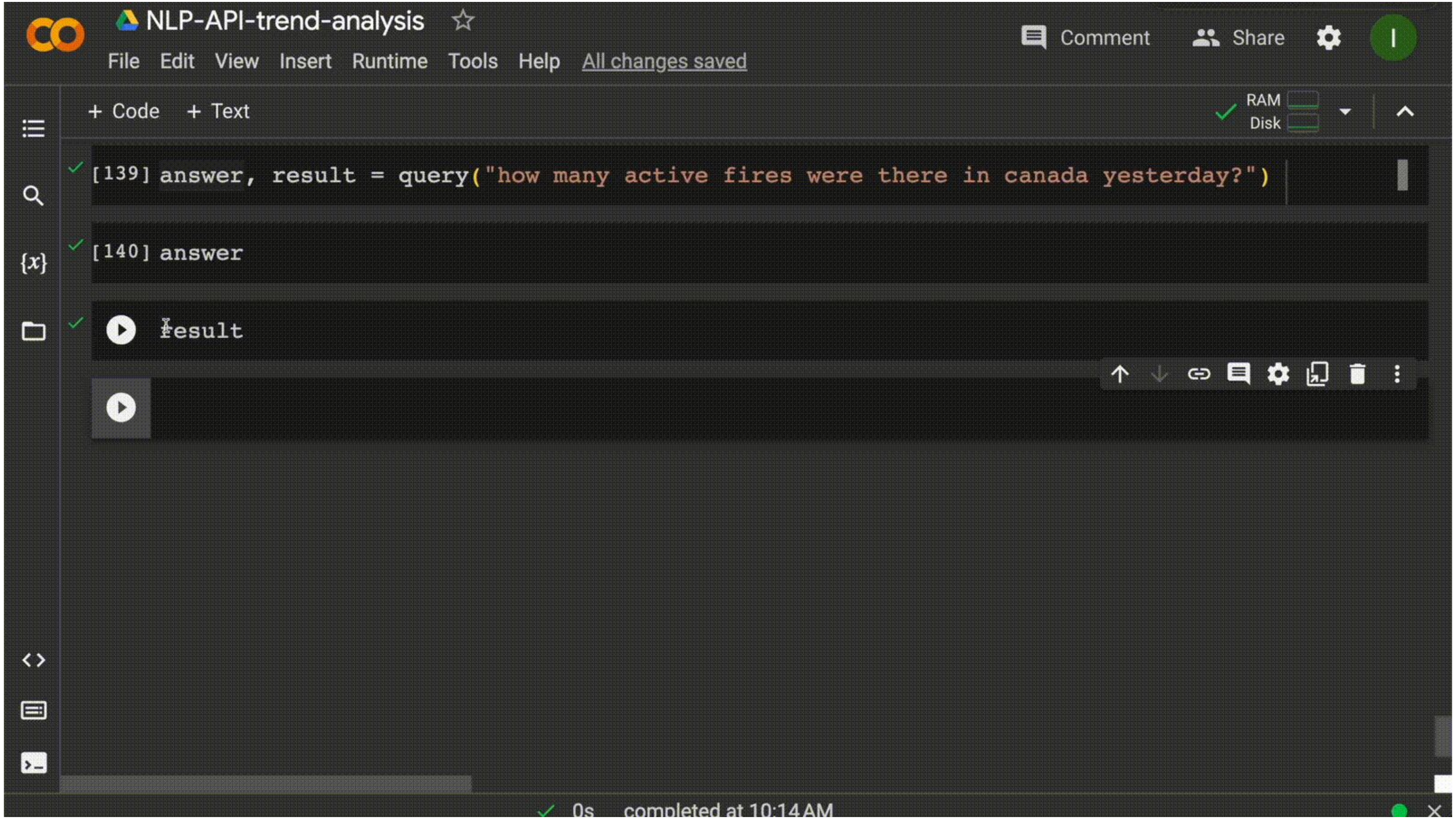

WORLDVIEW
Layers
Events
Data
OVERLAYS
Place Labels
© OpenStreetMap contributors, Natural Earth
Coastlines / Borders / Roads
© OpenStreetMap contributors, Natural Earth
MAIAC Aerosol Optical Depth
Terra and Aqua / MODIS
Fires and Thermal Anomalies (Day, 375m)
Suomi NPP / VIIRS
BASE LAYERS
Corrected Reflectance (Bands 7-2-1)
Terra / MODIS
Corrected Reflectance (True Color)
Terra / MODIS
+ Add Layers
Start Comparison
Eureka
Redding
Susanville
Chico
Reno
Willits
Yuba City
Carson City
Sacramento
Hawthorne
Santa Rosa
Stockton
San Francisco
Yosemite National Park
White Mountain Peak
San José
Merced
Santa Cruz
Salinas
Fresno
California
Mt. Whitney
Visalia
Junipero Serra Pk.
50 km
20 mi
40.6208°, -121.0254° EPSG:4326
1 DAY
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019
DAY


WORLDVIEW
Layers
Events
Data
OVERLAYS
Place Labels
© OpenStreetMap contributors, Natural Earth
Coastlines / Borders / Roads
© OpenStreetMap contributors, Natural Earth
MAIAC Aerosol Optical Depth
Terra and Aqua / MODIS
Fires and Thermal Anomalies (Day, 375m)
Suomi NPP / VIIRS
BASE LAYERS
Corrected Reflectance (Bands 7-2-1)
Terra / MODIS
Corrected Reflectance (True Color)
Terra / MODIS
+ Add Layers
Start Comparison
Eureka
Redding
Susanville
Chico
Reno
Willits
Yuba City
Carson City
Sacramento
Hawthorne
Santa Rosa
Stockton
Yosemite National Park
San Francisco
White Mountain Peak
San José
Merced
Santa Cruz
Salinas
Fresno
California
Mt. Whitney
Visalia
Junipero Serra Pk.
50 km
20 mi
40.6208°, -121.0254° EPSG:4326
1 DAY
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019
DAY


EARTHDATA
Find a DAAC
Feedback
EARTHDATA
Search
fire
Point: 38.62916292901474,-122.71896529197693
Start: 2019-10-23 00:00:00
Stop: 2019-11-07 23:59:59
Earthdata Login
Browse Collections
Features
Map Imagery
Near Real Time
Customizable
Keywords
Platforms
Instruments
Organizations
Projects
Processing levels
Chico
Reno
Willits
Yuba City
Carson City
Sacramento
Santa Rosa
Hawthorne
Stockton
San Francisco
Yosemite National Park
Merced
San José
30 km
20 mi
13 Matching Collections
Sort by: Relevance
Only include collections with granules
Include non-EOSDIS collections
Tip: Add + collections to your project to compare and download their data.
MODIS/Terra Thermal Anomalies/Fire 8-Day L3 Global 1km SIN Grid V006
3 Granules • 2000-02-18 ongoing • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire 8-Day (MOD14A2) Version 6 data are generated at 1 kilometer (km) spatial resolution as a Level 3 product. The MOD14A2 gridded composite contains the maximum value of the individual fire pixel classes detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask and pixel quality indicators. Improvements/Changes from Previous Versions *...
MOD14A2 v006 - LP DAAC
MODIS/Terra Thermal Anomalies/Fire Daily L3 Global 1km SIN Grid V006
3 Granules • 2000-02-18 ongoing • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire Daily (MOD14A1) Version 6 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 3 product. MOD14A1 contains eight consecutive days of fire data conveniently packaged into a single file. The Science Dataset (SDS) layers include the fire mask, pixel quality indicators, maximum fire radiative power (MaxFRP), and the position of the fire pixel within the scan....
MOD14A1 v006 - LP DAAC
MODIS/Aqua Thermal Anomalies/Fire Daily L3 Global 1km SIN Grid V006
3 Granules • 2002-07-04 ongoing • The Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire Daily (MYD14A1) Version 6 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 3 product. MYD14A1 contains eight consecutive days of fire data conveniently packaged into a single file. The Science Dataset (SDS) layers include the fire mask, pixel quality indicators, maximum fire-radiative-power (MaxFRP), and the position of the fire pixel within the scan....
MYD14A1 v006 - LP DAAC
MODIS/Aqua Thermal Anomalies/Fire 8-Day L3 Global 1km SIN Grid V006
3 Granules • 2002-07-04 ongoing • The Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Fire 8-Day (MYD14A2) Version 6 data are generated at 1 kilometer (km) spatial resolution as a Level 3 product. The MYD14A2 gridded composite contains maximum value of individual fire pixel classes detected during the eight days of acquisition. The Science Dataset (SDS) layers include the fire mask and pixel quality indicators. Improvements/Changes from Previous Versions * Refinements to...

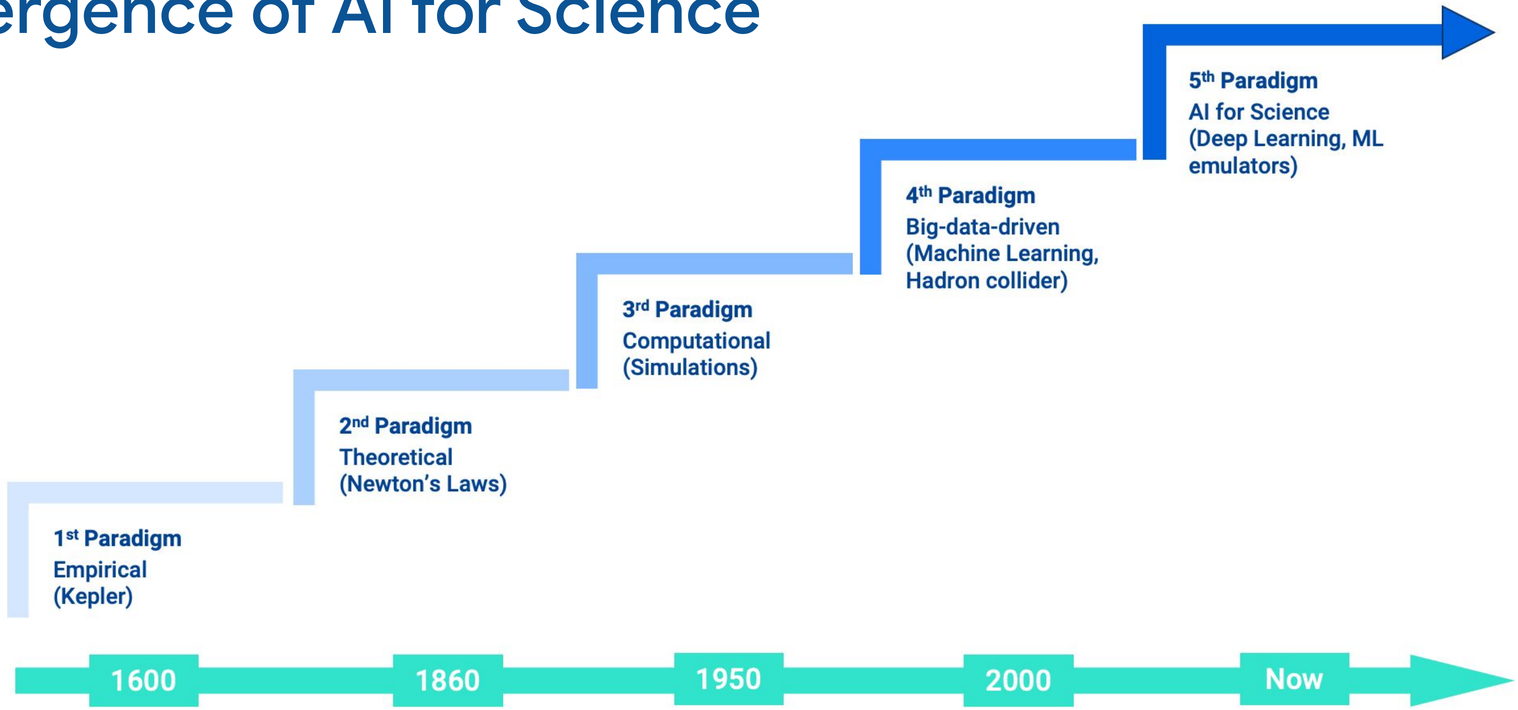

Smoke Model Detection
California wildfires, October 27, 2019

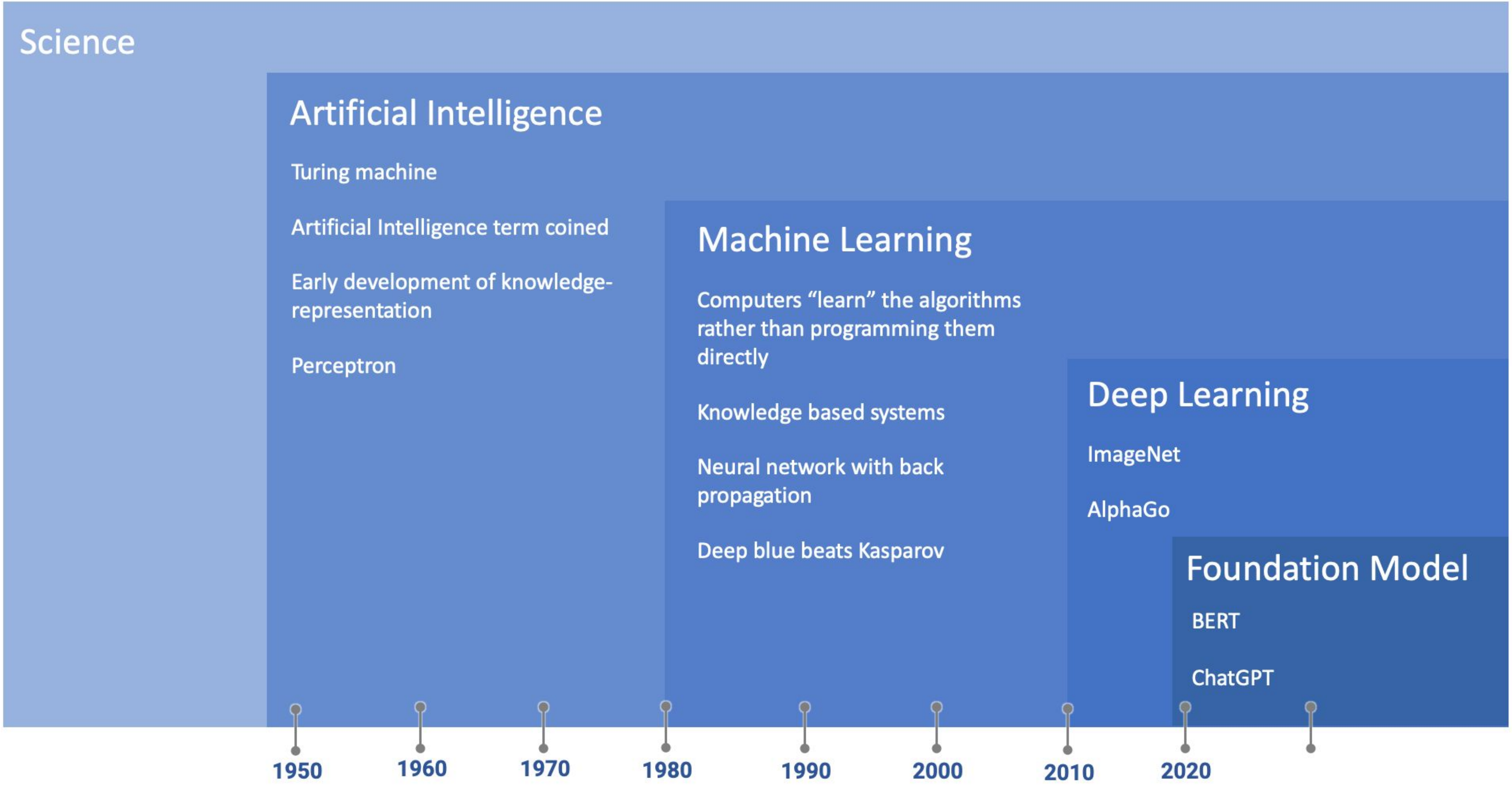
Enhancing Scientific Efficiency through AI



Scientific process is fundamentally changing due to AI

Emergence of AI for Science




Science
Artificial Intelligence
Turing machine
Artificial Intelligence term coined
Early development of knowledge-representation
Perceptron
Machine Learning
Computers "learn" the algorithms rather than programming them directly
Knowledge based systems
Neural network with back propagation
Deep blue beats Kasparov
Deep Learning
ImageNet
AlphaGo
Foundation Model
BERT
ChatGPT
1950
1960
1970
1980
1990
2000
2010
2020

AI in General	AI for Science
Deploy and iterate	Robustness and accuracy; Mistakes may be costly
Benchmark & SOTA	Find shortcomings and understand why
Recognize patterns	Model PDEs, ODEs
Automate human tasks	Augment repetitive tasks
Improvement in efficiency	High-consequence decision making
Less specialized data	Complex and heterogeneous data
Internal evaluation	Scientific community evaluation

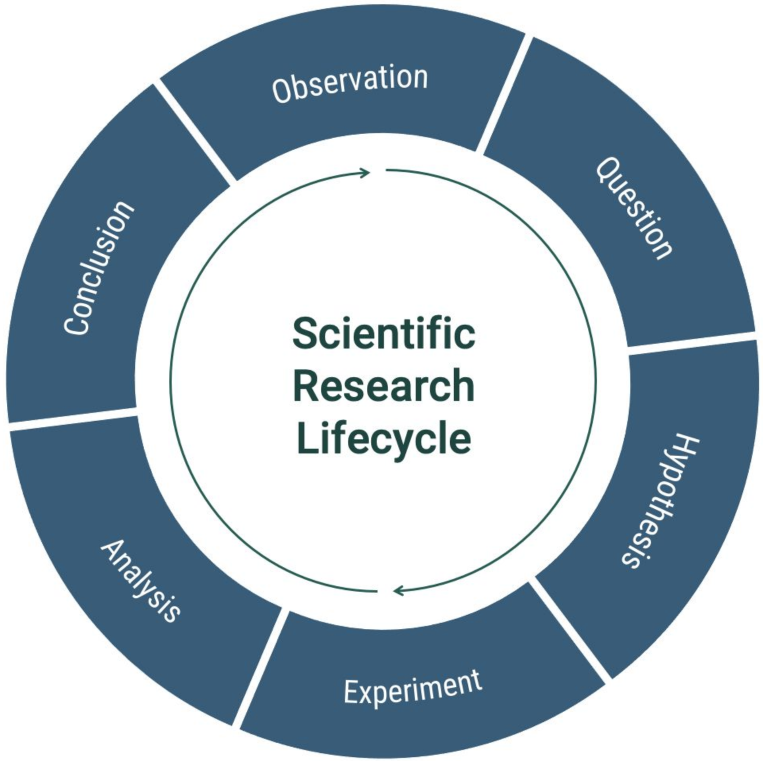
Advancing science: AI integration and necessary advances

Scientists are creative & patient

How can we effectively use AI tools in science?

What advancements in AI are needed to meet our science goals?

Science is a process



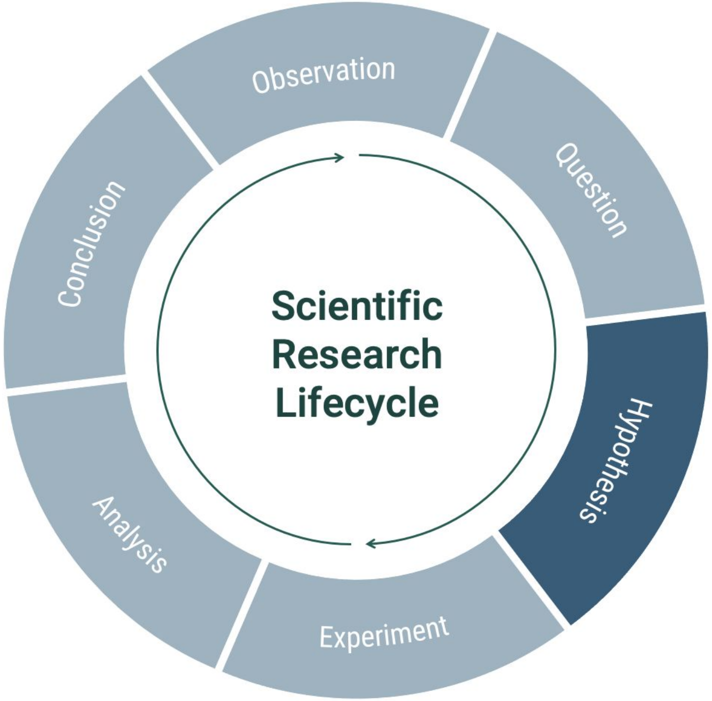
How can we best utilize AI tools in science?

Hypothesis space is too large

Humans can consider a few hypotheses at a time

AI can generate and test more complex hypotheses

Select which hypotheses are worthy of testing

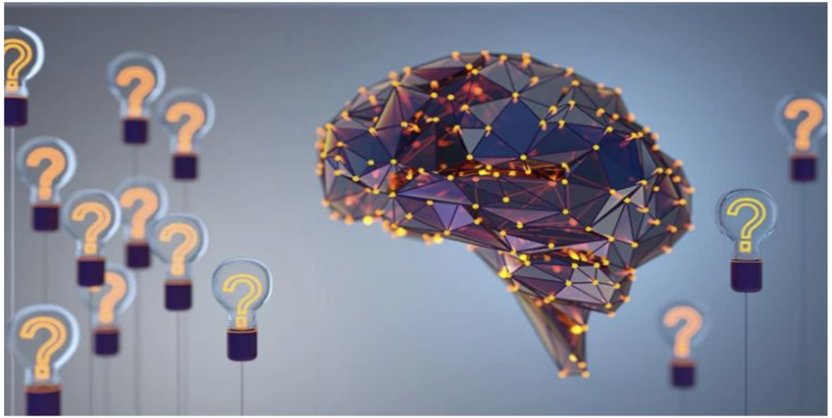


How can we best utilize AI tools in science?

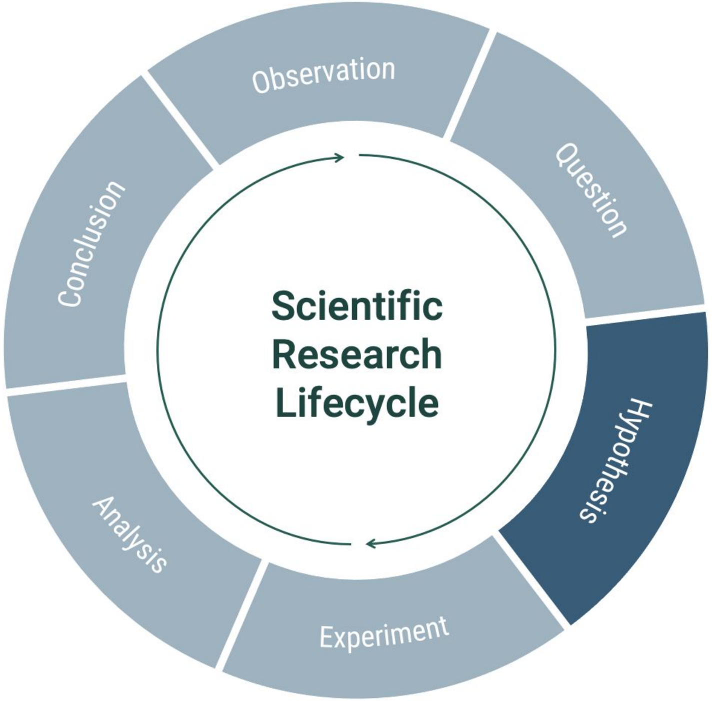
Hypotheses devised by AI could find 'blind spots' in research

Artificial intelligence is asking questions that humans hope to answer.

By Matthew Hutson



Credit: Olemedia/Getty



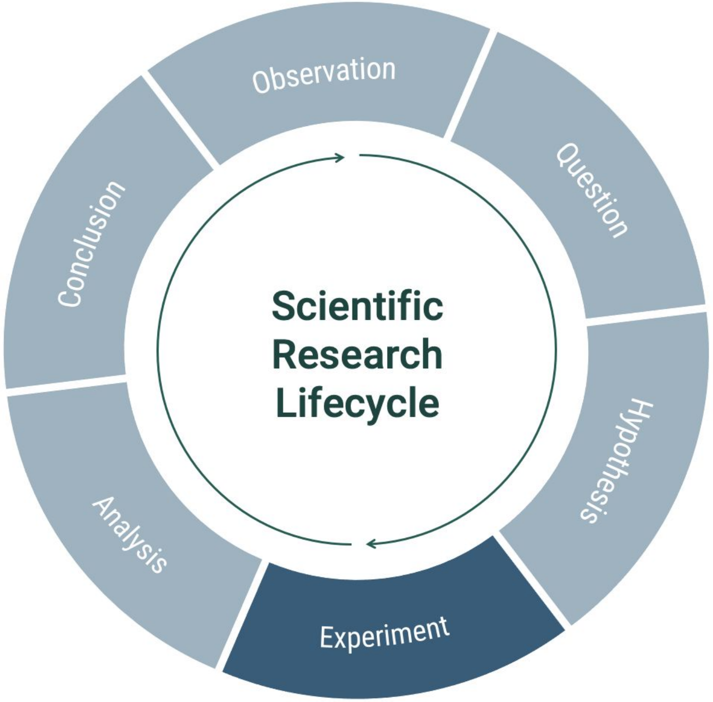
How can we best utilize AI tools in science?

Design of candidates

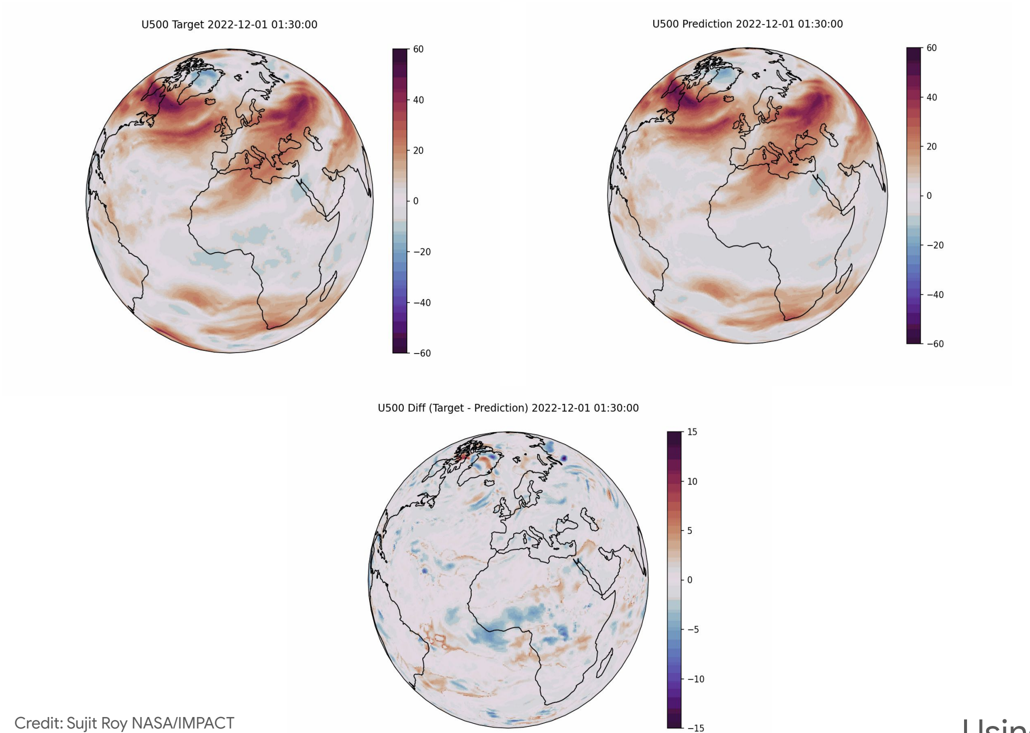
Less expensive experimentations

Which science papers to choose?

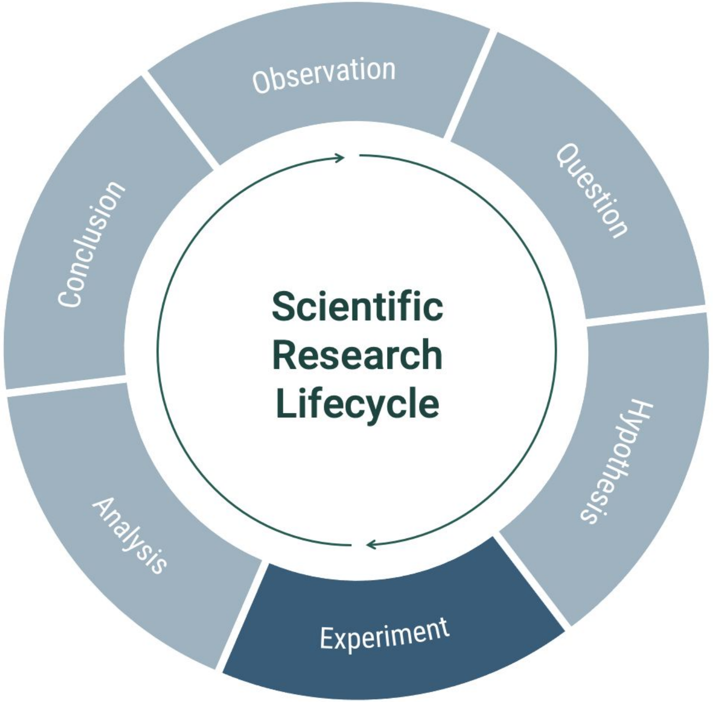
Robotics



How can we best utilize AI tools in science?



Credit: Sujit Roy NASA/IMPACT



Using AI to optimize experiments

How can we best utilize AI tools in science?

Too much data for human cognition

Massive data generated by each experiment

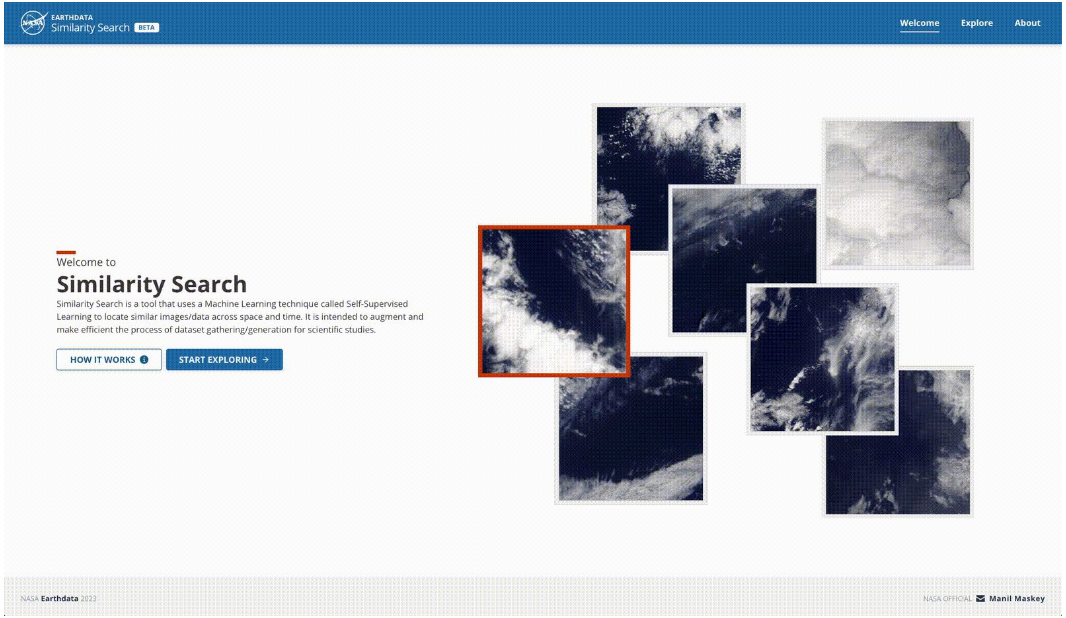
Different researchers publish experiments with different results

Discovering data is searching needle in a haystack

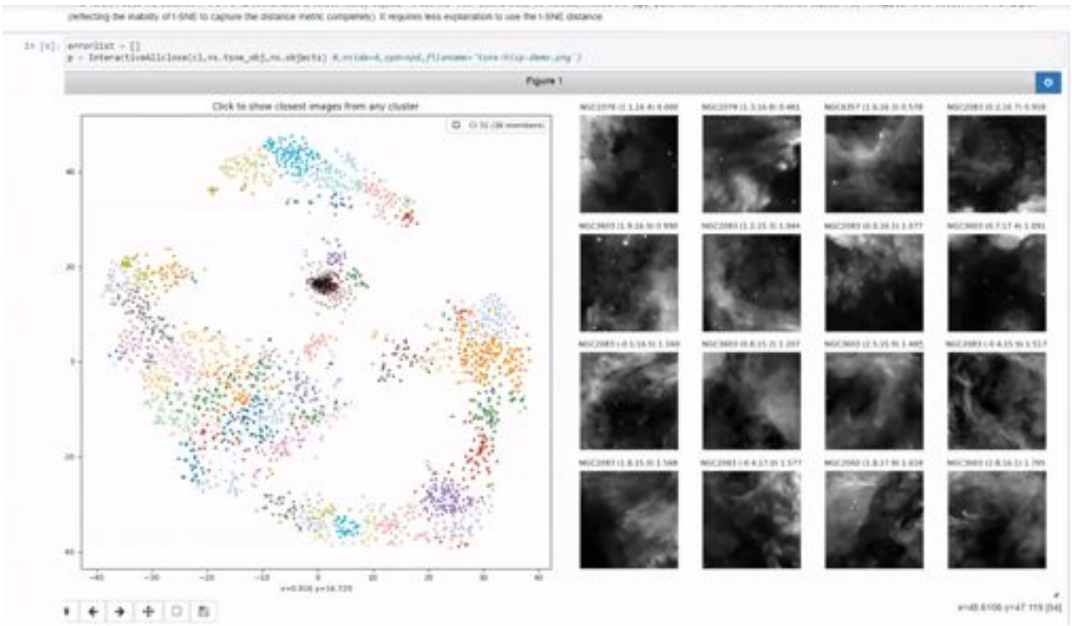
Science phenomena are rare

AI for scaling, speedup

How can we best utilize AI tools in science?



How can we best utilize AI tools in science?

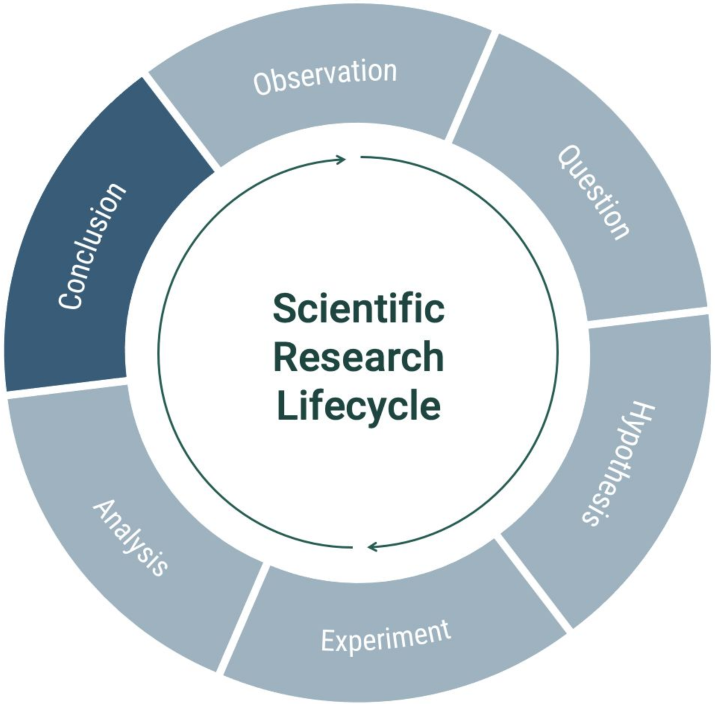


How can we best utilize AI tools in science?

Synthesizing research

Publication

AI writing agents

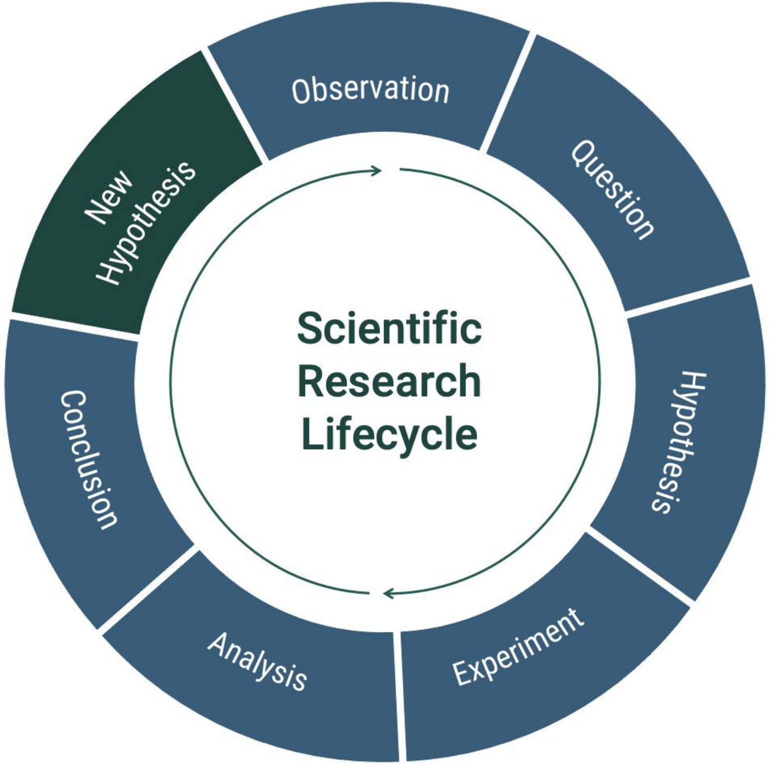


How can we best utilize AI tools in science?

AI has the potential to change every step of the science research lifecycle

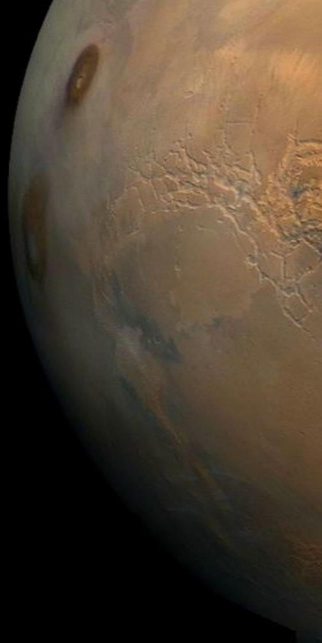
AI can speed up scientific discovery process - no repetitive tasks

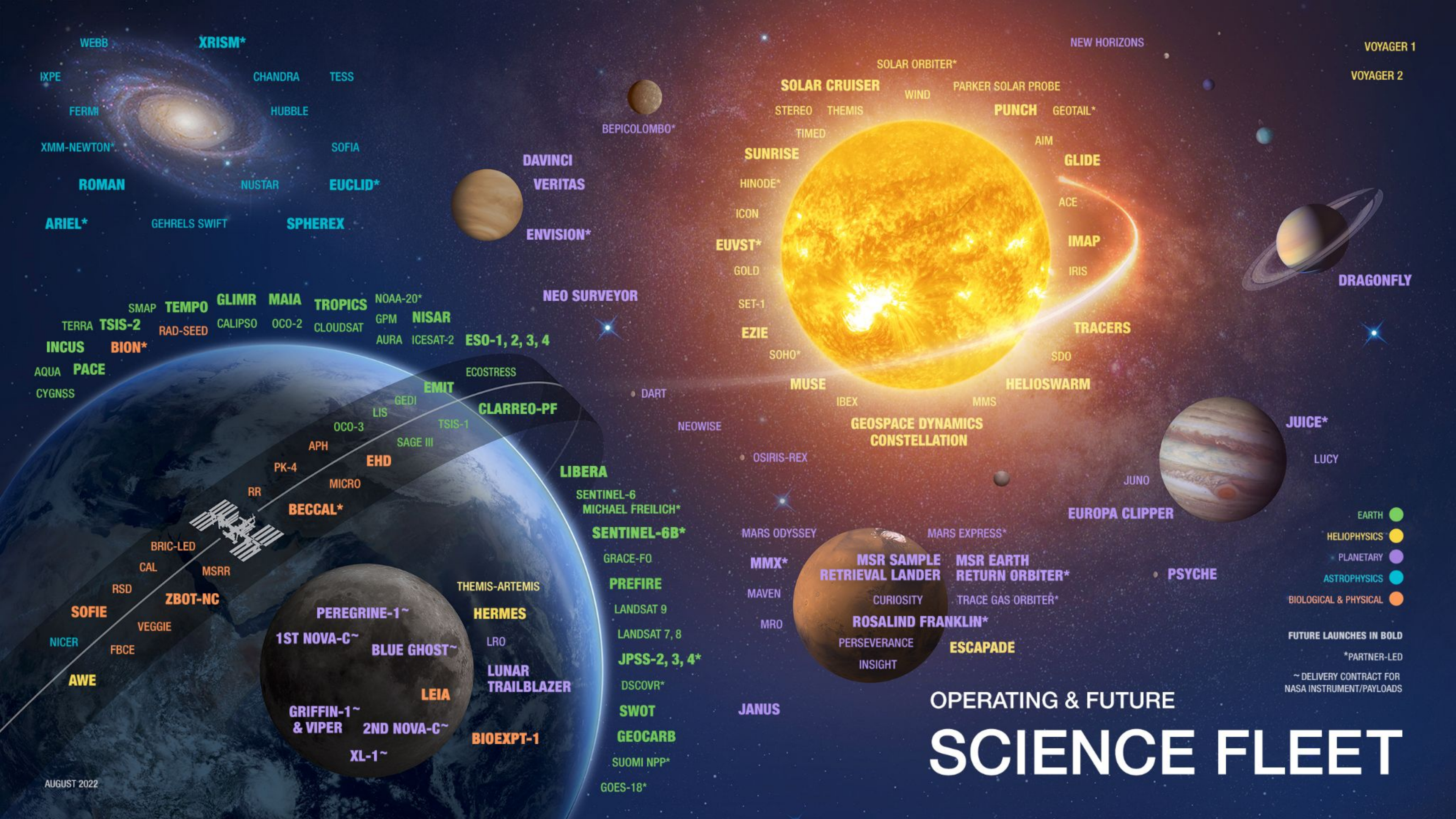
Completely new AI generated hypothesis?

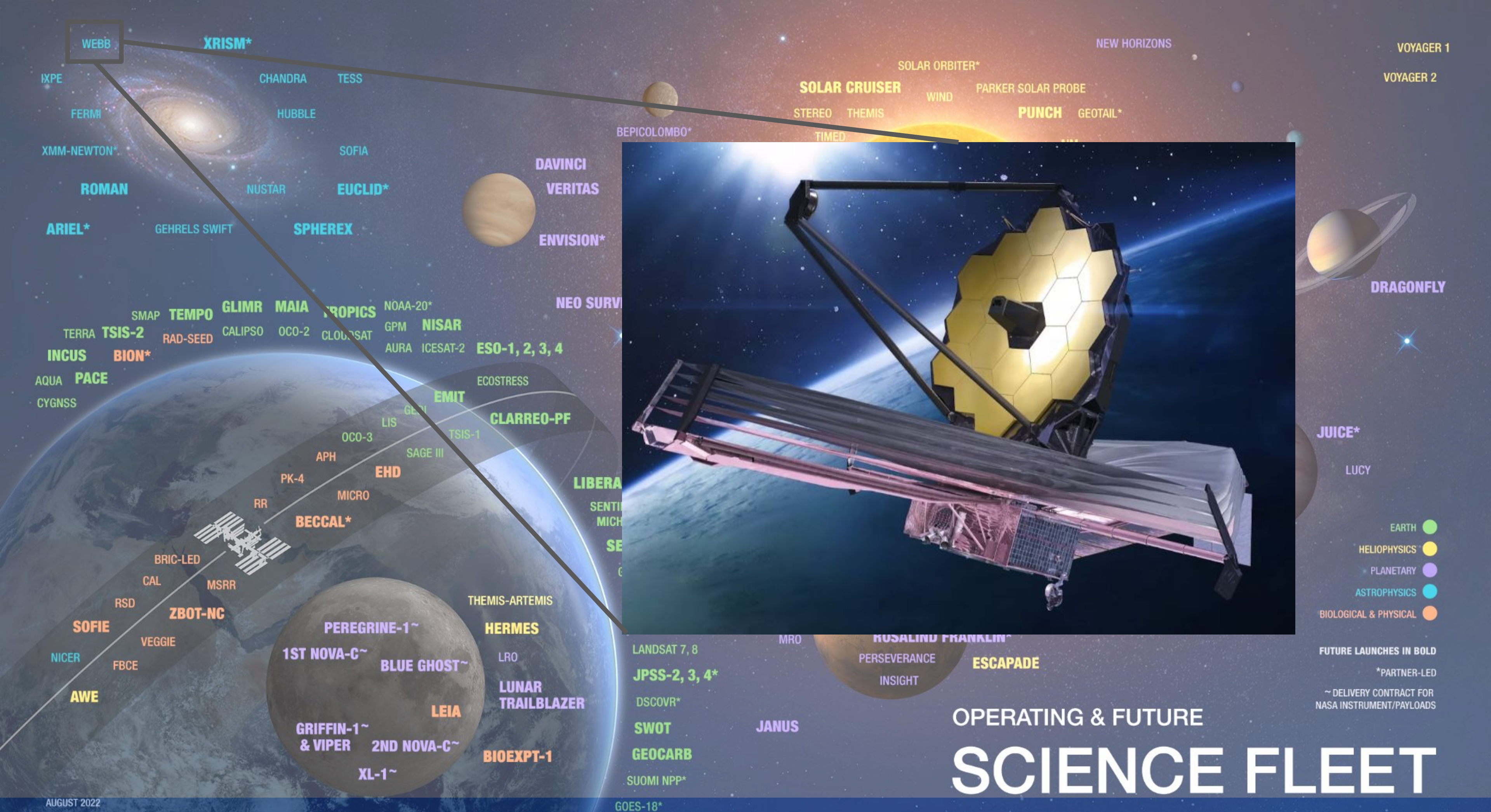


NASA Science

Vision: using the vantage point of space to achieve with the science community and our partners a deep scientific **understanding** of our planet, other planets and solar system bodies, the interplanetary environment, the Sun and its effects on the solar system, and the universe beyond....




WEBB
XRISM*
IXPE
CHANDRA
TESS
FERMI
HUBBLE
XMM-NEWTON*
SOFIA
ROMAN
NUSTAR
EUCLID*
ARIEL*
GEHRELS SWIFT
SPHEREX
BEPICOLOMBO*
DAVINCI
VERITAS
ENVISION*
NEO SURVEYOR
SOLAR ORBITER*
SOLAR CRUISER
WIND
PARKER SOLAR PROBE
STEREO
THEMIS
PUNCH
GEOTAIL*
TIMED
SUNRISE
AIM
GLIDE
HINODE*
ACE
ICON
EUVST*
IMAP
GOLD
IRIS
SET-1
EZIE
TRACERS
SOHO*
SDO
MUSE
HELIOSWARM
IBEX
MMS
GEOSPACE DYNAMICS CONSTELLATION
NEW HORIZONS
VOYAGER 1
VOYAGER 2
DRAGONFLY
JUICE*
LUCY
JUNO
EUROPA CLIPPER
DART
NEOWISE
OSIRIS-REX
SMAP
TEMPO
GLIMR
MAIA
TROPICS
NOAA-20*
TERRA
TSIS-2
RAD-SEED
CALIPSO
OCO-2
CLOUDSAT
GPM
NISAR
INCUS
BION*
AURA
ICESAT-2
ESO-1, 2, 3, 4
AQUA
PACE
CYGNSS
ECOSTRESS
EMIT
GEDI
LIS
CLARREO-PF
OCO-3
TSIS-1
APH
SAGE III
PK-4
EHD
MICRO
RR
BECCAL*
BRIC-LED
CAL
MSRR
RSD
ZBOT-NC
SOFIE
VEGGIE
NICER
FBCE
AWE
LIBERA
SENTINEL-6 MICHAEL FREILICH*
SENTINEL-6B*
GRACE-FO
PREFIRE
LANDSAT 9
LANDSAT 7, 8
JPSS-2, 3, 4*
DSCOVR*
SWOT
GEOCARB
SUOMI NPP*
GOES-18*
THEMIS-ARTEMIS
PEREGRINE-1~
HERMES
1ST NOVA-C~
BLUE GHOST~
LRO
LUNAR TRAILBLAZER
LEIA
GRIFFIN-1~ & VIPER
2ND NOVA-C~
BIOEXPT-1
XL-1~
MARS ODYSSEY
MARS EXPRESS*
MMX*
MSR SAMPLE RETRIEVAL LANDER
MSR EARTH RETURN ORBITER*
MAVEN
CURIOSITY
TRACE GAS ORBITER*
MRO
ROSALIND FRANKLIN*
PERSEVERANCE
ESCAPADE
INSIGHT
JANUS
PSYCHE
EARTH
HELIOPHYSICS
PLANETARY
ASTROPHYSICS
BIOLOGICAL & PHYSICAL
FUTURE LAUNCHES IN BOLD
*PARTNER-LED
~DELIVERY CONTRACT FOR NASA INSTRUMENT/PAYLOADS
OPERATING & FUTURE
SCIENCE FLEET
AUGUST 2022


WEBB
XRISM*
IXPE
CHANDRA
TESS
FERMI
HUBBLE
XMM-NEWTON*
SOFIA
ROMAN
NUSTAR
EUCLID*
ARIEL*
GEHRELS SWIFT
SPHEREX
DAVINCI
VERITAS
ENVISION*
NEW HORIZONS
VOYAGER 1
VOYAGER 2
SOLAR ORBITER*
SOLAR CRUISER
WIND
PARKER SOLAR PROBE
STEREO
THEMIS
PUNCH
GEOTAIL*
BEPICOLOMBO*
TIMED
DRAGONFLY
SMAP
TEMPO
GLIMR
MAIA
NOAA-20*
TERRA
TSIS-2
RAD-SEED
CALIPSO
OCO-2
GPM
NISAR
INCUS
BION*
AURA
ICESAT-2
ESO-1, 2, 3, 4
AQUA
PACE
CYGNSS
ECOSTRESS
EMIT
LIS
CLARREO-PF
OCO-3
TSIS-1
APH
SAGE III
PK-4
EHD
MICRO
RR
BECCAL*
BRIC-LED
CAL
MSRR
RSD
ZBOT-NC
SOFIE
VEGGIE
NICER
FBCE
AWE
THEMIS-ARTEMIS
PEREGRINE-1~
HERMES
1ST NOVA-C~
BLUE GHOST~
LRO
LUNAR TRAILBLAZER
LEIA
GRIFFIN-1~ & VIPER
2ND NOVA-C~
BIOEXPT-1
XL-1~
JUICE*
LUCY
EARTH
HELIOPHYSICS
PLANETARY
ASTROPHYSICS
BIOLOGICAL & PHYSICAL
LANDSAT 7, 8
MRO
PERSEVERANCE
INSIGHT
ESCAPADE
JPSS-2, 3, 4*
DSCOVR*
SWOT
JANUS
GEOCARB
SUOMI NPP*
GOES-18*
FUTURE LAUNCHES IN BOLD
*PARTNER-LED
~ DELIVERY CONTRACT FOR NASA INSTRUMENT/PAYLOADS
AUGUST 2022
OPERATING & FUTURE
SCIENCE FLEET

Science by the NUMBERS

RESEARCH

~10,000 U.S. Scientists Funded
~3,000 Competitively Selected Awards
~$600M Awarded Annually

TECHNOLOGY DEVELOPMENT

~$397M Invested Annually

SMALLSATS/ CUBESATS

57 Science Missions
10 Technology Demos

EARTH OBSERVATIONS

24 Operating Missions
23 Upcoming Missions

SOUNDING ROCKETS

11 Science Missions Launched
43 In Development

MISSIONS

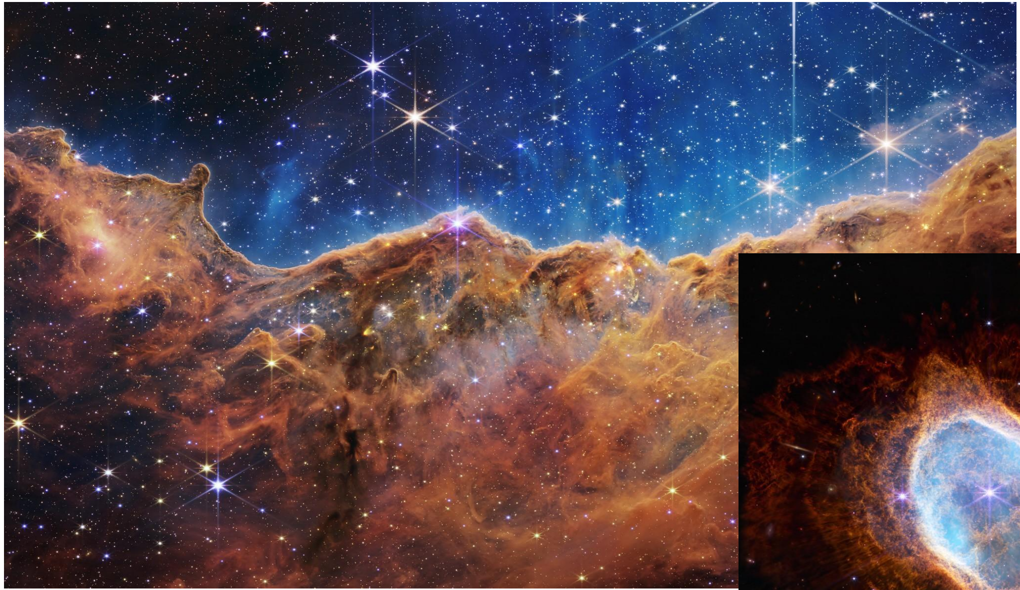
134 Missions from formulation through extended operations

BALLOONS

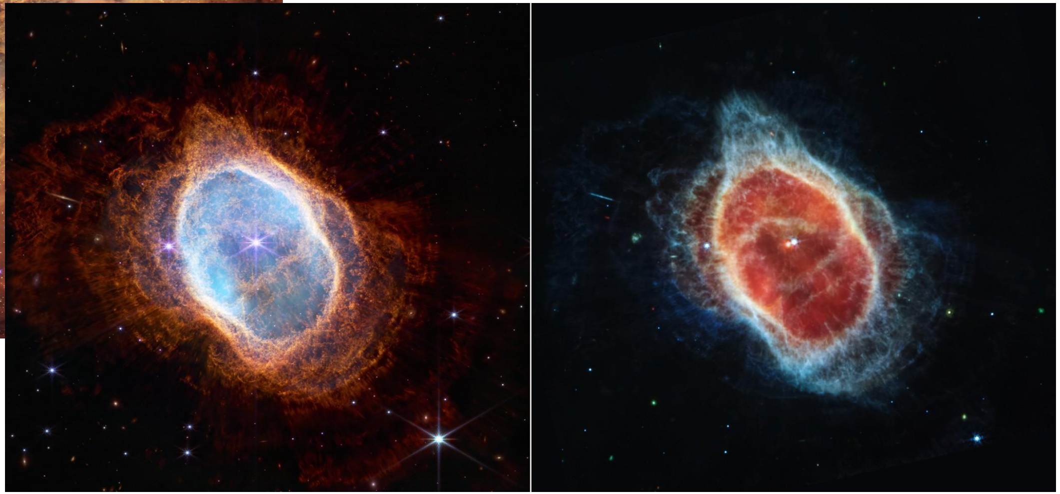
2 Missions Launched
52 Missions in Development

Value of NASA's Science Data

NASA's Strategic Goal 1.1
"Understand the Sun, Earth, Solar System, and Universe."



Carina Nebula – a rare peek of stars in their earliest, rapid stages of formation.



Dying Star's Final 'Performance' – dying star expelling gas and dust

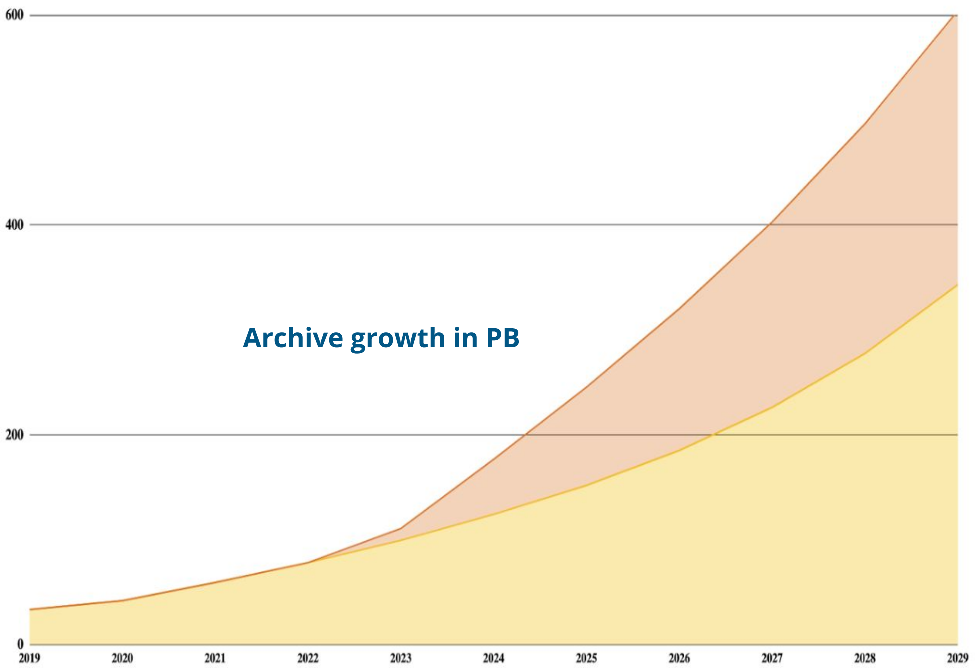

NASA EARTH FLEET
OPERATING & FUTURE THROUGH 2023
SWOT (CNES)
LANDSAT-9 (USGS)
SENTINEL-6 Michael Freilich/B (ESA)
TROPICS (6)
GEOCARB
NISAR (ISRO)
MAIA
TSIS-2
TEMPO
PREFIRE (2)
PACE (NSO)
GLIMR
ICESAT-2
GRACE-FO (2) (DLR)
CYGNSS (8)
NISTAR, EPIC (DSCOVR/NOAA)
CLOUDSAT (CSA)
TERRA (JAXA, CSA)
AQUA (JAXA, AEB)
AURA (NSO, FMI, UKSA)
CALIPSO (CNES)
GPM (JAXA)
LANDSAT 7 (USGS)
LANDSAT 8 (USGS)
OCO-2
SMAP
SUOMI NPP (NOAA) (JAXA)
ISS INSTRUMENTS
EMIT
CLARREO-PF
GEDI
SAGE III
OCO-3
TSIS-1
ECOSTRESS
LIS
JPSS-2, 3 & 4 INSTRUMENTS
OMPS-Limb
LIBERA
03.24.20
INVEST/CUBESATS
RainCube
CSIM-FD
CubeRRT
TEMPEST-D
CIRiS
HARP
CTIM
HyTI
SNoOPI
NACHOS
(PRE) FORMULATION
IMPLEMENTATON
PRIMARY OPS
EXTENDED OPS

Now

~100PB Archive

4B+ Files

Future



NASA Earth science data growth

Enduring Legacy of NASA Science Data

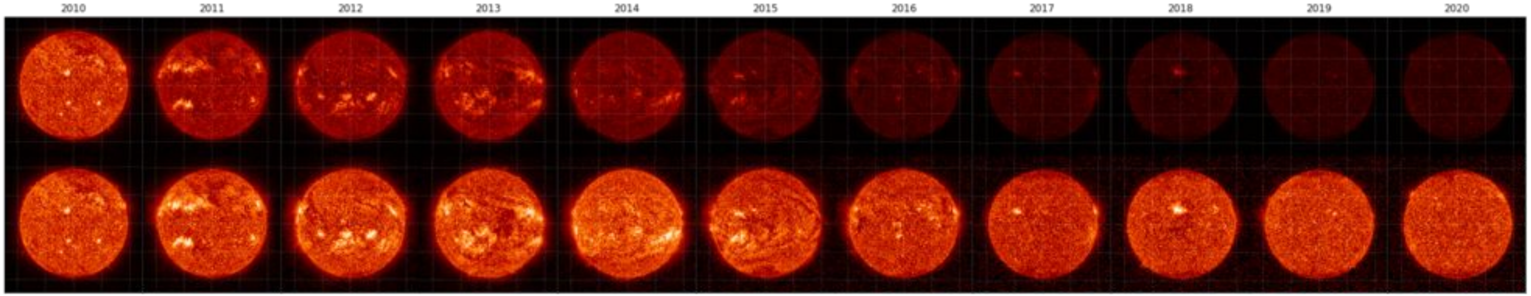
“NASA has become a **knowledge agency**. Long after the Mars Surveyor has gone silent, Hubble has met the same fate as Mir, and the Moderate Resolution Imaging Spectroradiometer has produced its final set of images, what will endure are the volumes of valuable **data** that these instruments and many others have collected over their lifetimes....”

– National Research Council, 2002

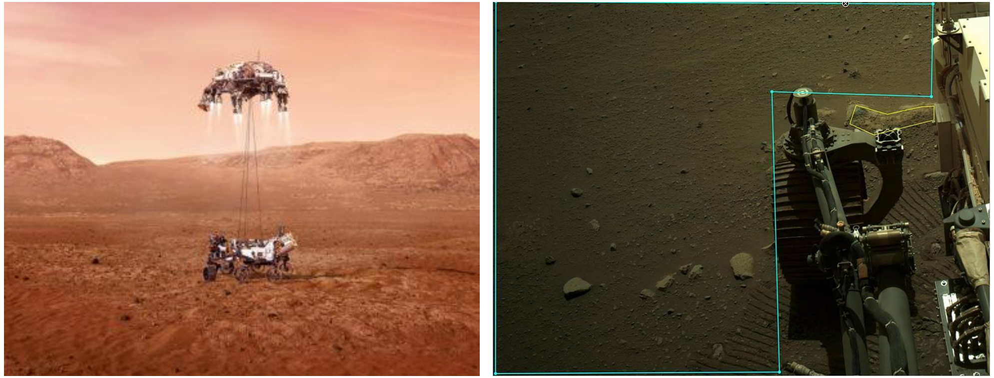
... the challenge of managing overwhelming data and distilling it into meaningful knowledge

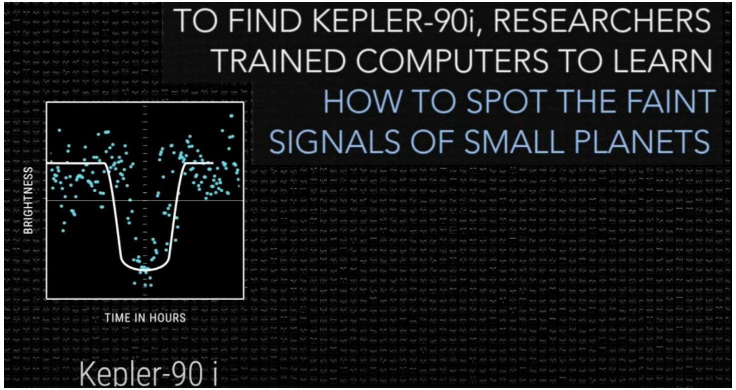
Artist concept of the Mars Global Surveyor
NASA/JPL-Caltech/Corby Waste

Artificial intelligence successes - NASA science



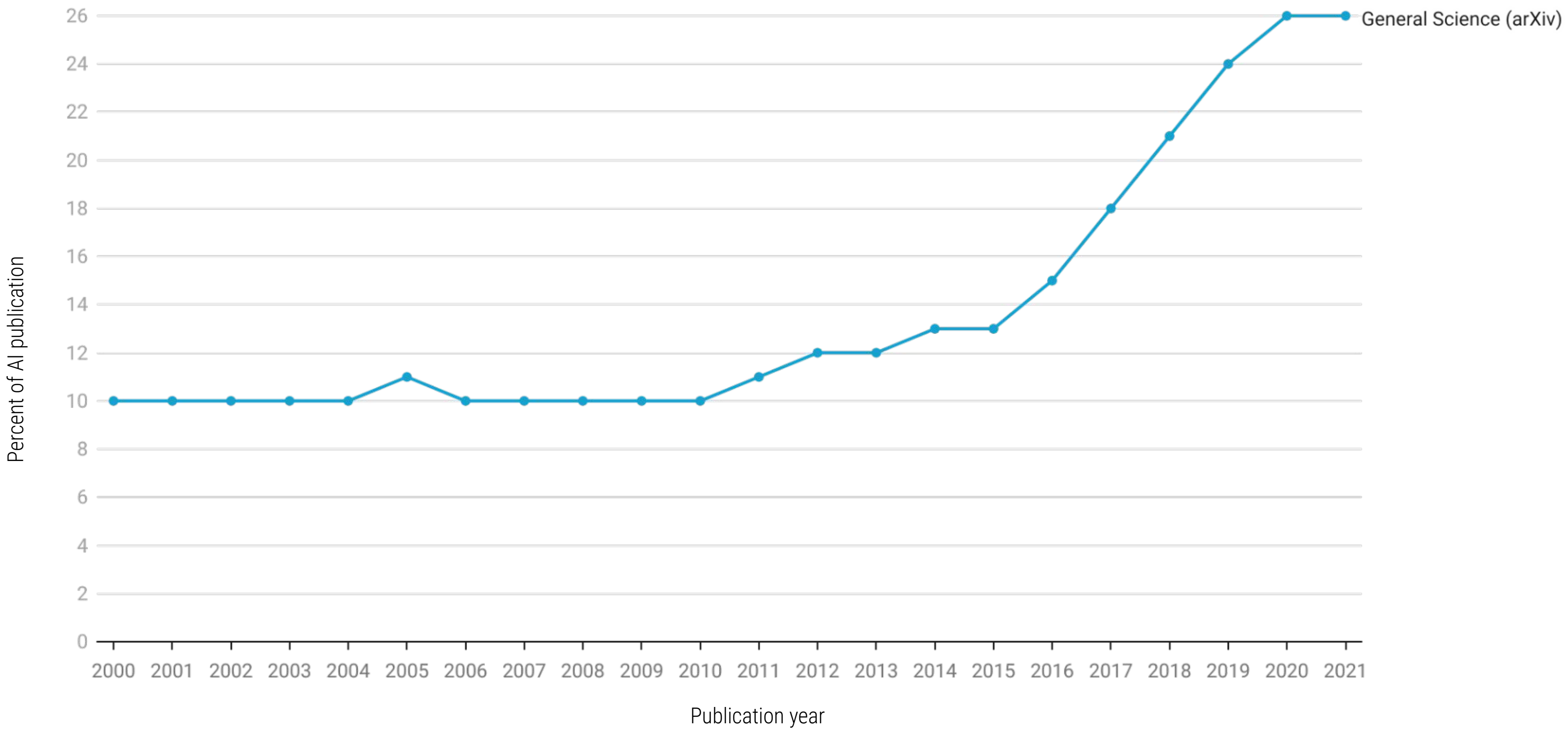
Images correction for degradation using AI



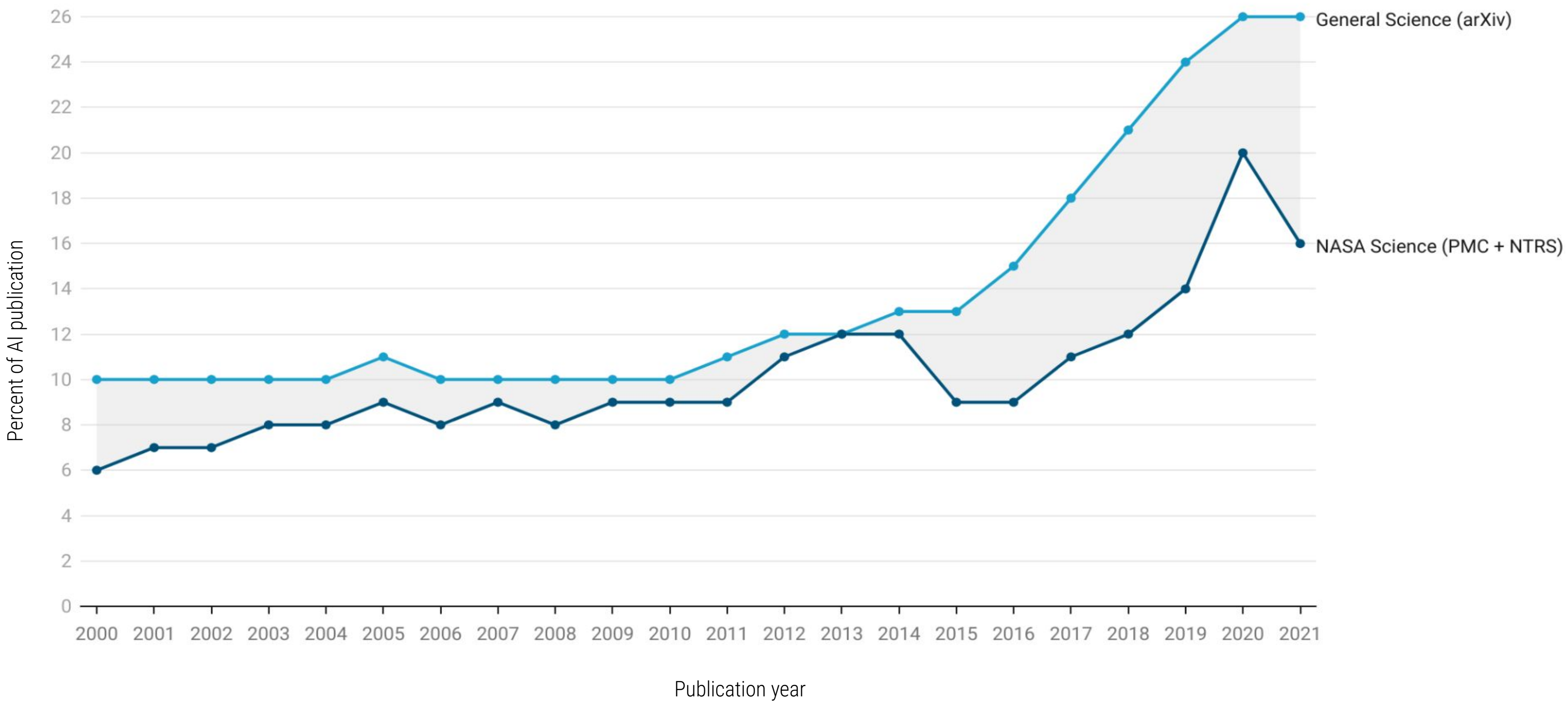


Rover onboard AI systems for guidance

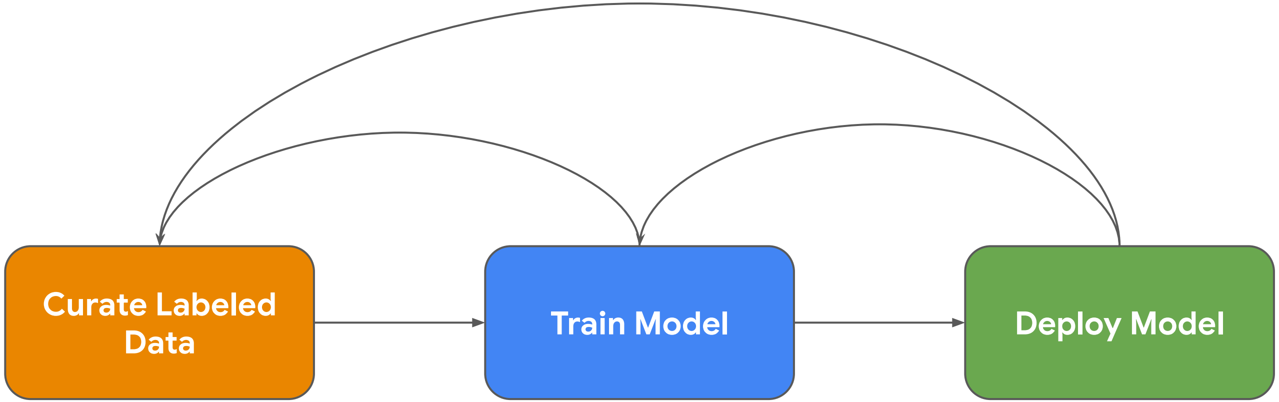
AI Publication Trend for General Science



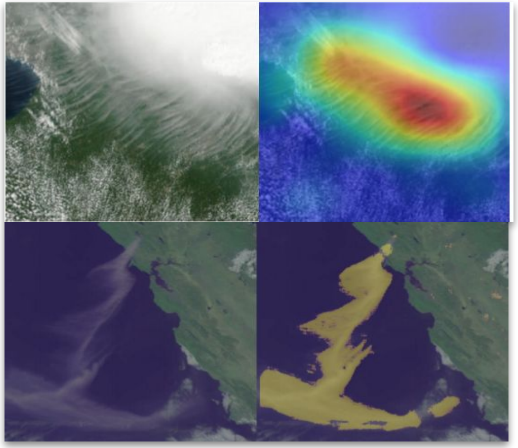
AI Publication Trend for General Science vs. NASA Science



Supervised learning



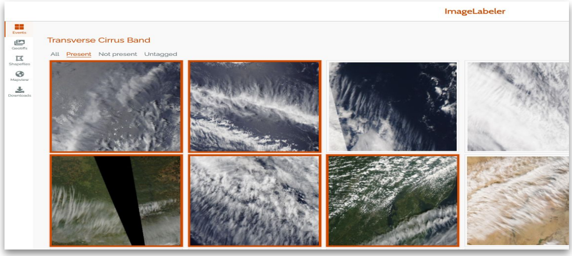
Supervised learning over the years...



Atmospheric phenomena identification

COVID 19 indicators

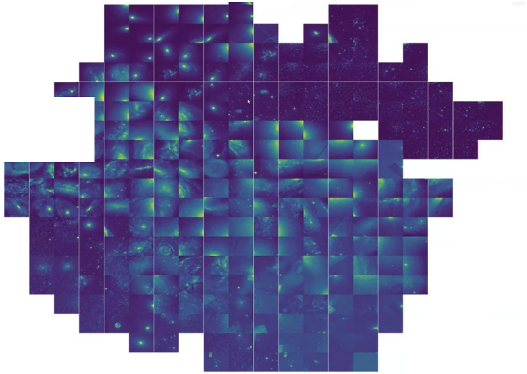
Marine debris segmentation



Labeling tools

Cross-divisional AI

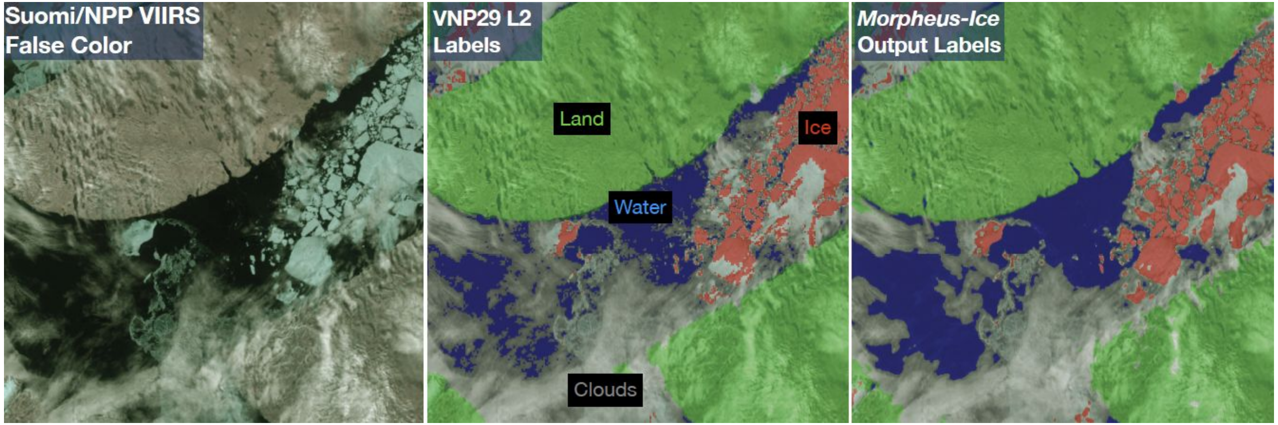
Petabyte scale search on multi-spectral unlabeled data to rapidly curate annotated datasets + Search By Image for NASA Science - *Earth science + Astrophysics*



Cross divisional AI project: Self supervised learning approach developed for GIBS archives applied to Hubble telescope data

Cross-divisional AI

Morpheus-Ice: Pixel-Level Classification of Imaging Data from Astrophysics to Earth Science



AI for science at scale

9.9 billion
trees assessed

326,000
satellite images

89,000
hand-labeled tree crowns

10^7 million
sq. km. mapped

60,000,000
core HPC hours (NASA + NSF)

38
collaborators across academia and government

Tucker and Brandt et al.

An Unexpectedly Large Count of Trees in West African Sahara and Sahel

Semi-arid trees have a crucial role in carbon storage & biodiversity and are not well-documented.

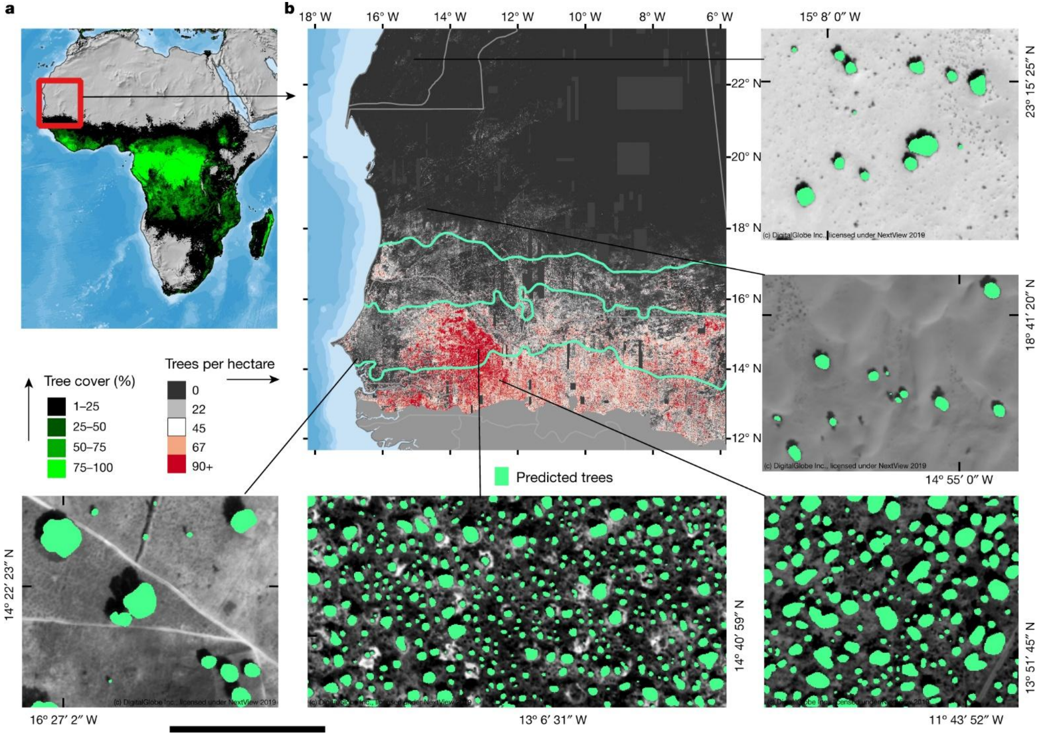
Researchers mapped trees >3 m2 crown area in West Africa, from the Sahara, Sahel, and sub-humid zone using 50 cm satellite imagery, deep learning, and HPC.

1,837,565,501 trees were mapped over 1,300,000 km2

Deep learning and <1 m satellite imagery deliver heretofore unavailable semi-arid carbon storage and biodiversity insights

GEDI and ICESat-2 data can be combined with these data to provide tree heights for larger trees or clusters of trees

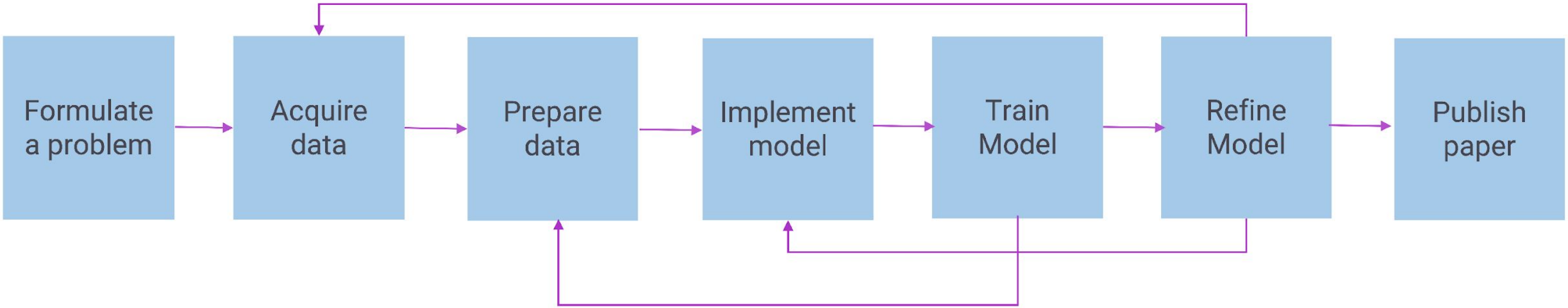
Brandt, M., Tucker, C.J., Kariryaa, A. et al. An unexpectedly large count of trees in the West African Sahara and Sahel. Nature (2020). https://doi.org/10.1038/s41586-020-2824-5



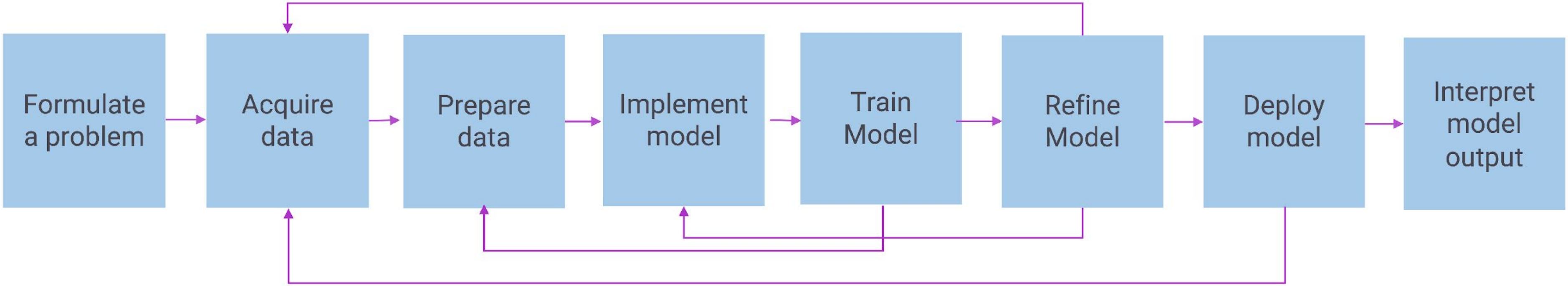
a) All trees >3-m^2 crown area were mapped in the red rectangle using deep learning and <1 m satellite imagery. b) The density of trees per hectare derived from 1,837,565,501 trees at various scales, from local to regional scales.

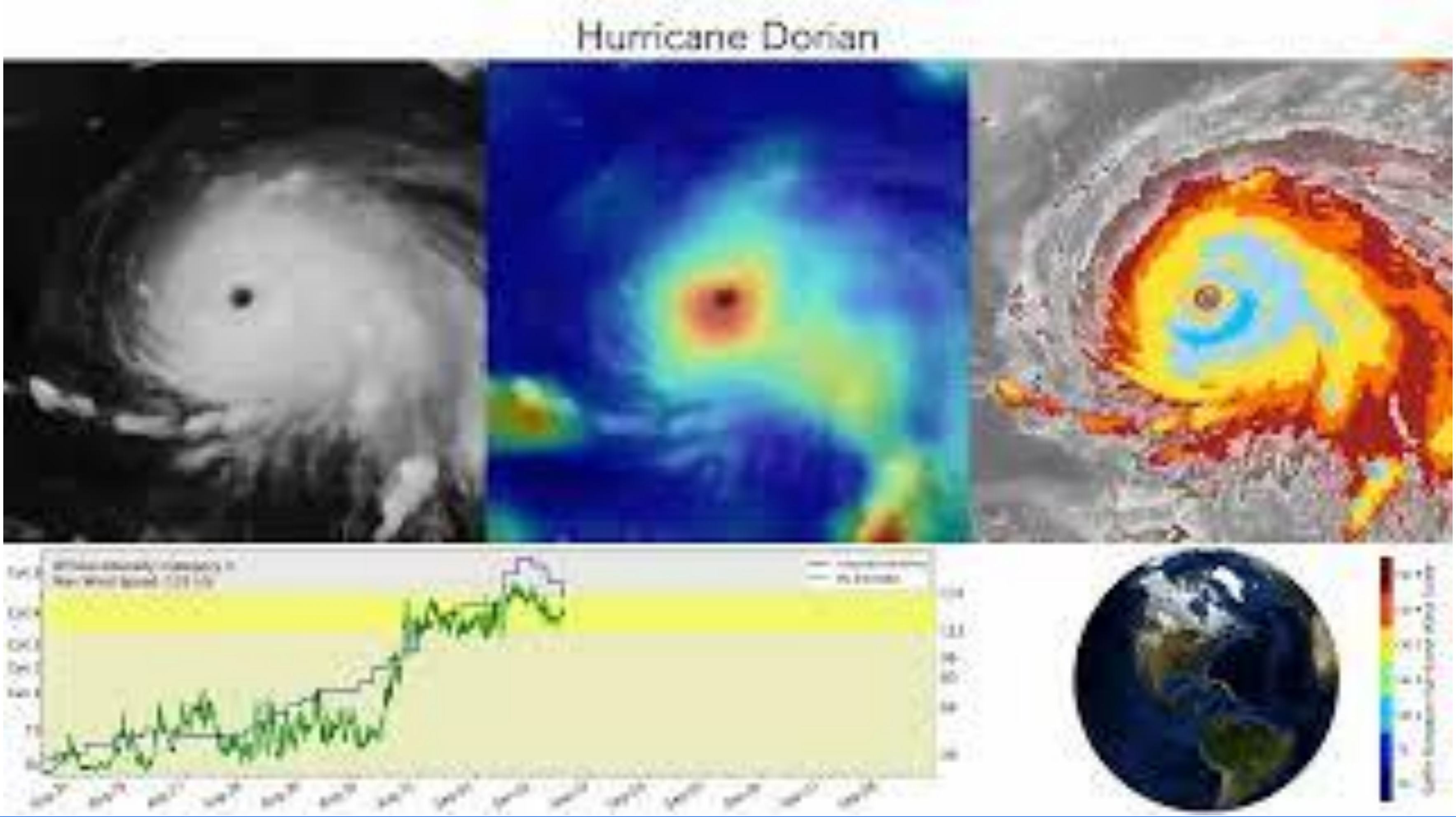
Last mile..

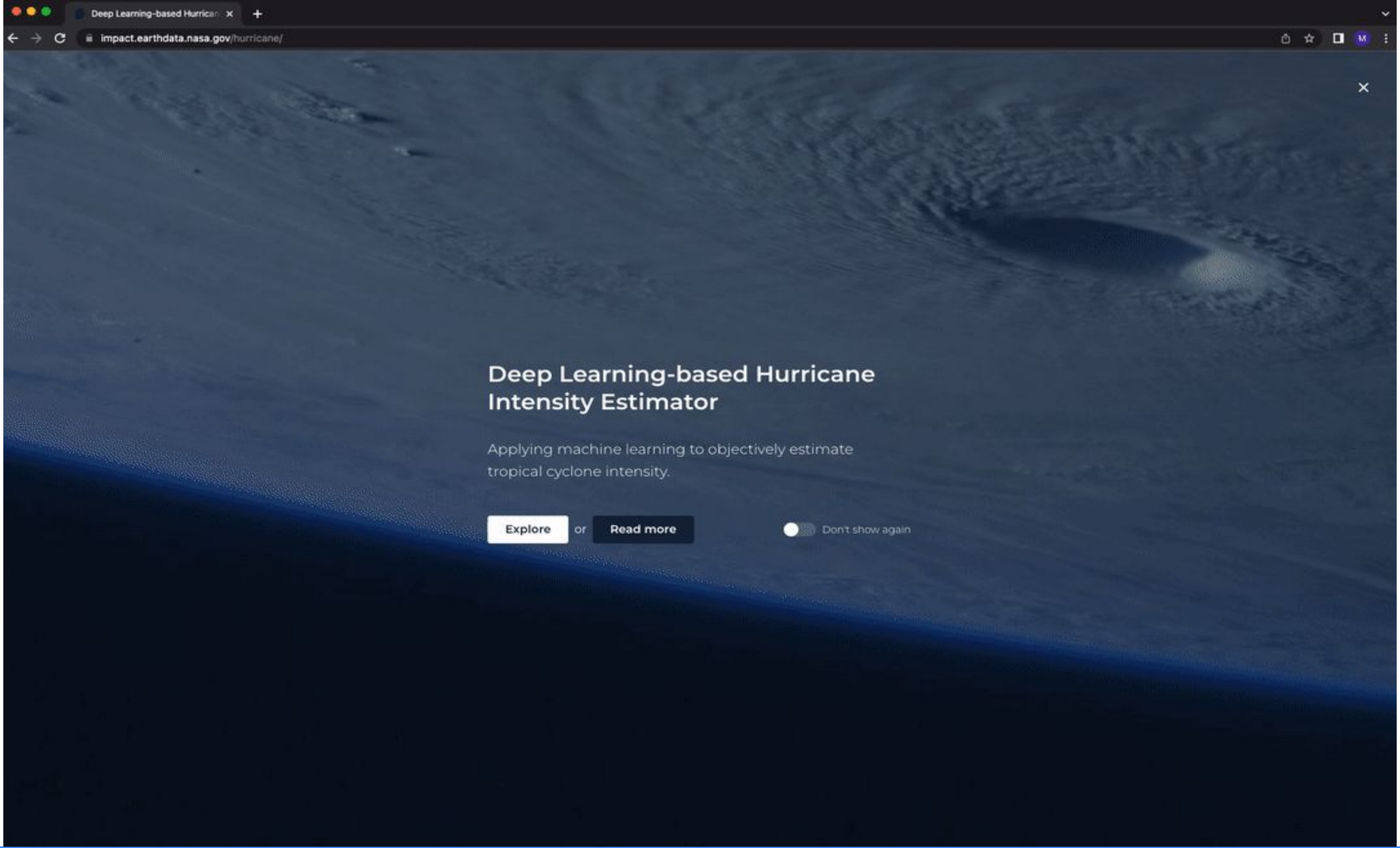
AI in literature



AI lifecycle - iterative




Hurricane Dorian



https://impact.earthdata.nasa.gov/hurricane/

There are challenges with Supervised Learning

Advancing Application of Machine Learning Tools for NASA's Earth Observation Data

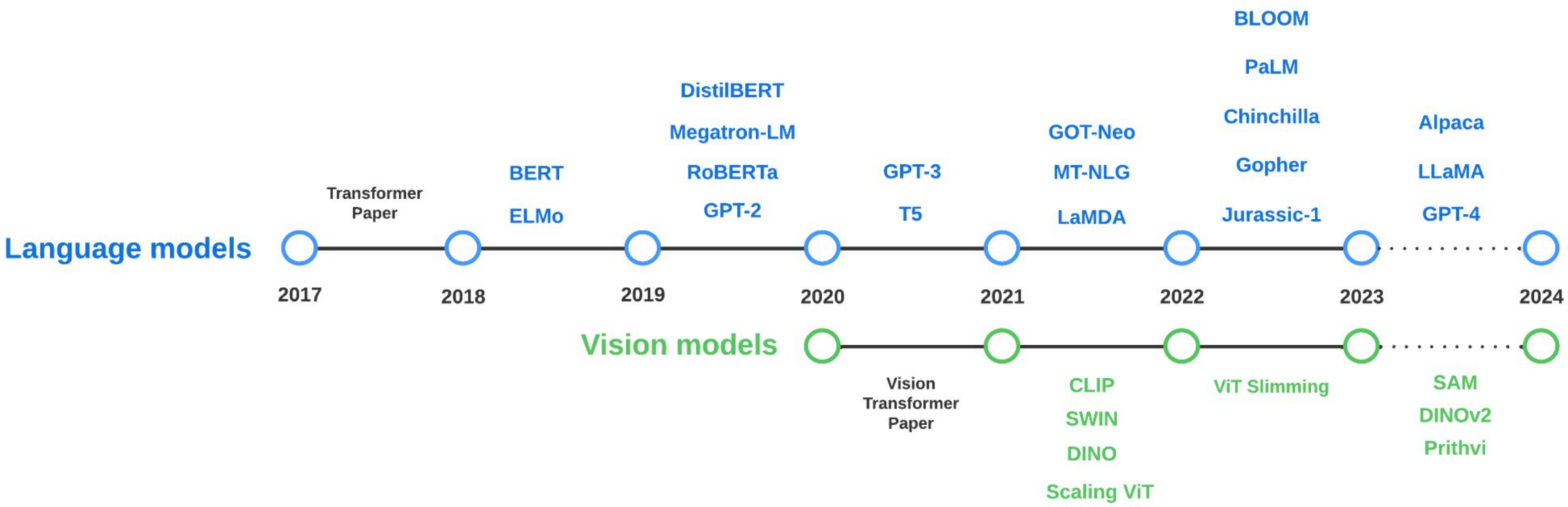
Jan. 21-23, 2020 | Washington, D.C.
Workshop Report

Maskey et al. "Advancing AI for Earth Science: A Data Systems Perspective," AGU Eos 2020

- **Training data** is the main component of supervised machine learning techniques and is increasingly becoming the **main bottleneck to advance applications of machine learning** techniques in Earth science.
- Geoscience models must **generalize across space and time**; however, for supervised learning one needs large training datasets to build generalizable models.

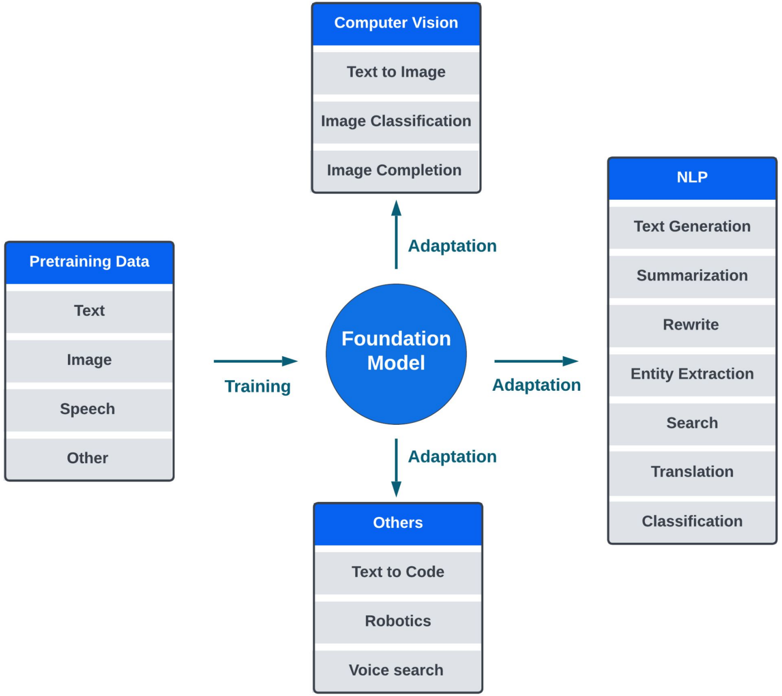
Recent Progress

Self-supervised learning at scale

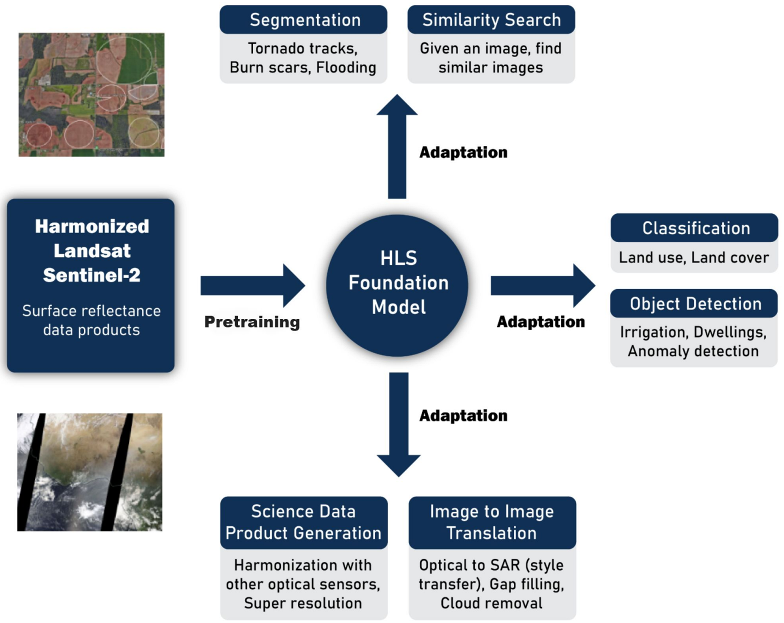


AI foundation models

- Large-scale models pre-trained on vast amounts of data, serving as a starting point for fine-tuning on specific tasks
- Unlike traditional models FMs are pre-trained on general data and then adapted to specialized tasks
- Pre-training captures broad knowledge, allowing for versatility across multiple applications
- Substantially reduce the downstream effort for building AI applications, including the need for large labeled training datasets

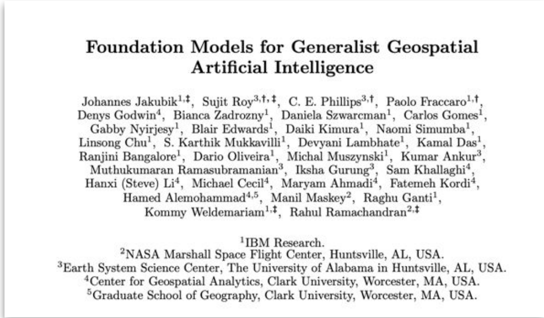


Geospatial foundation model with Harmonized Landsat Sentinel-2: Prithvi



- Build with collaboration with IBM Research
- Initial version released are 100M and 300M parameter models
- Masked Autoencoder where attention mechanism is extended in space and time
- Being evaluated for adaptation for different categories of downstream tasks

Collaborators: IBM, UAH, Clark University, ORNL, Hugging Face



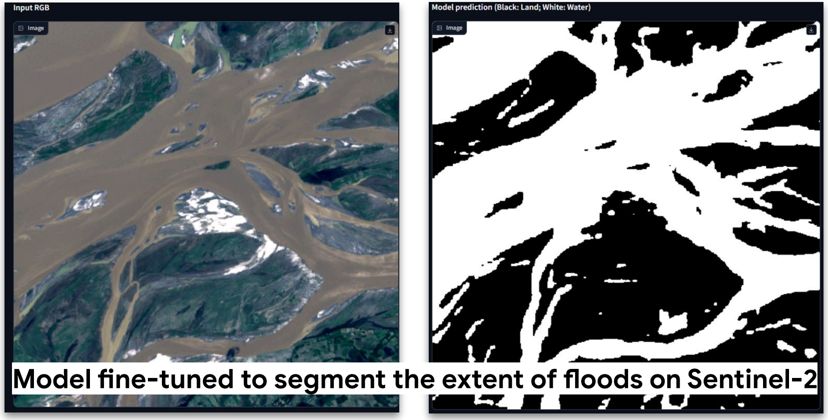
Foundation Models for Generalist Geospatial Artificial Intelligence

Johannes Jakubik[1,‡], Sujit Roy[3,†,‡], C. E. Phillips[3,†], Paolo Fraccaro[1,†], Denys Godwin[4], Bianca Zadrozny[1], Daniela Szwarcman[1], Carlos Gomes[1], Gabby Nyirjesy[1], Blair Edwards[1], Daiki Kimura[1], Naomi Simumba[1], Linsong Chu[1], S. Karthik Mukkavilli[1], Devyani Lambhate[1], Kamal Das[1], Ranjini Bangalore[1], Dario Oliveira[1], Michal Muszynski[1], Kumar Ankur[3], Muthukumaran Ramasubramanian[3], Iksha Gurung[3], Sam Khallaghi[4], Hanxi (Steve) Li[4], Michael Cecil[4], Maryam Ahmadi[4], Fatemeh Kordi[4], Hamed Alemohammad[4,5], Manil Maskey[2], Raghu Ganti[1], Kommy Weldemariam[1,‡], Rahul Ramachandran[2,‡]

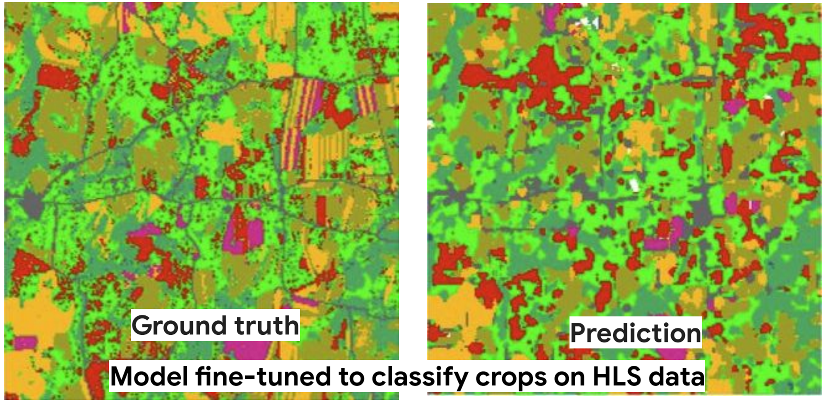
[1]IBM Research.
[2]NASA Marshall Space Flight Center, Huntsville, AL, USA.
[3]Earth System Science Center, The University of Alabama in Huntsville, AL, USA.
[4]Center for Geospatial Analytics, Clark University, Worcester, MA, USA.
[5]Graduate School of Geography, Clark University, Worcester, MA, USA.

https://arxiv.org/pdf/2310.18660.pdf

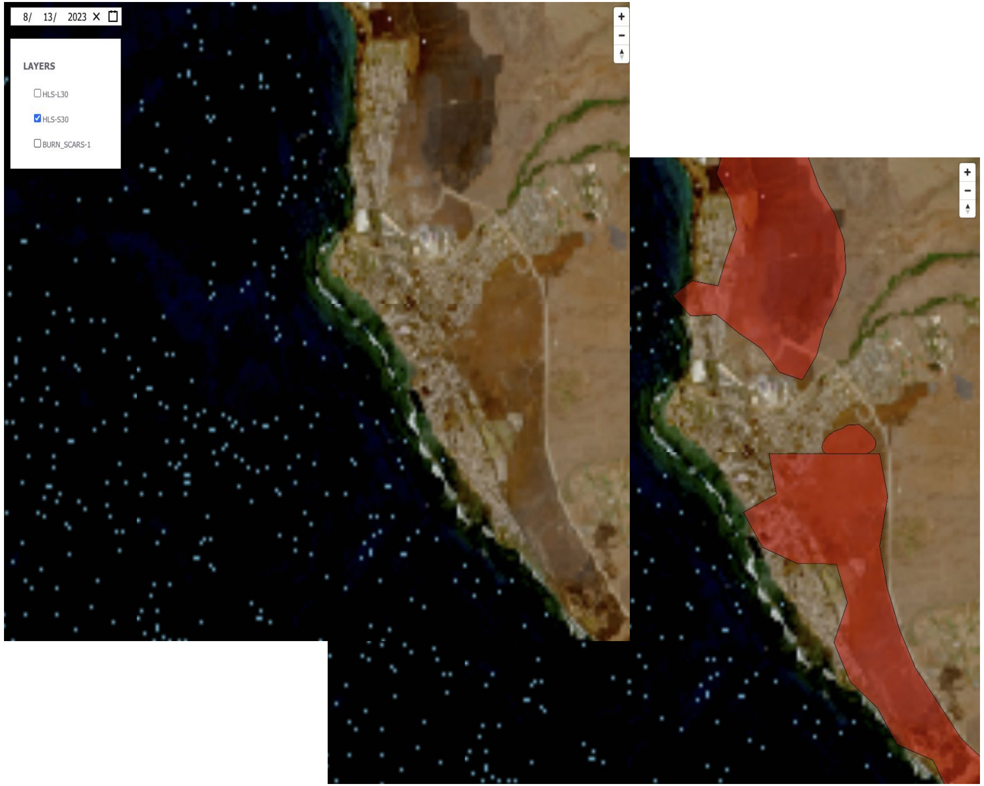
Geospatial foundation model : downstream applications



Model fine-tuned to segment the extent of floods on Sentinel-2



Model fine-tuned to classify crops on HLS data



Burn scar mapping: Maui Fire 8.13.2023

EARTHDATA
AI-powered Earth Insights BETA

Welcome Explore About

Welcome to

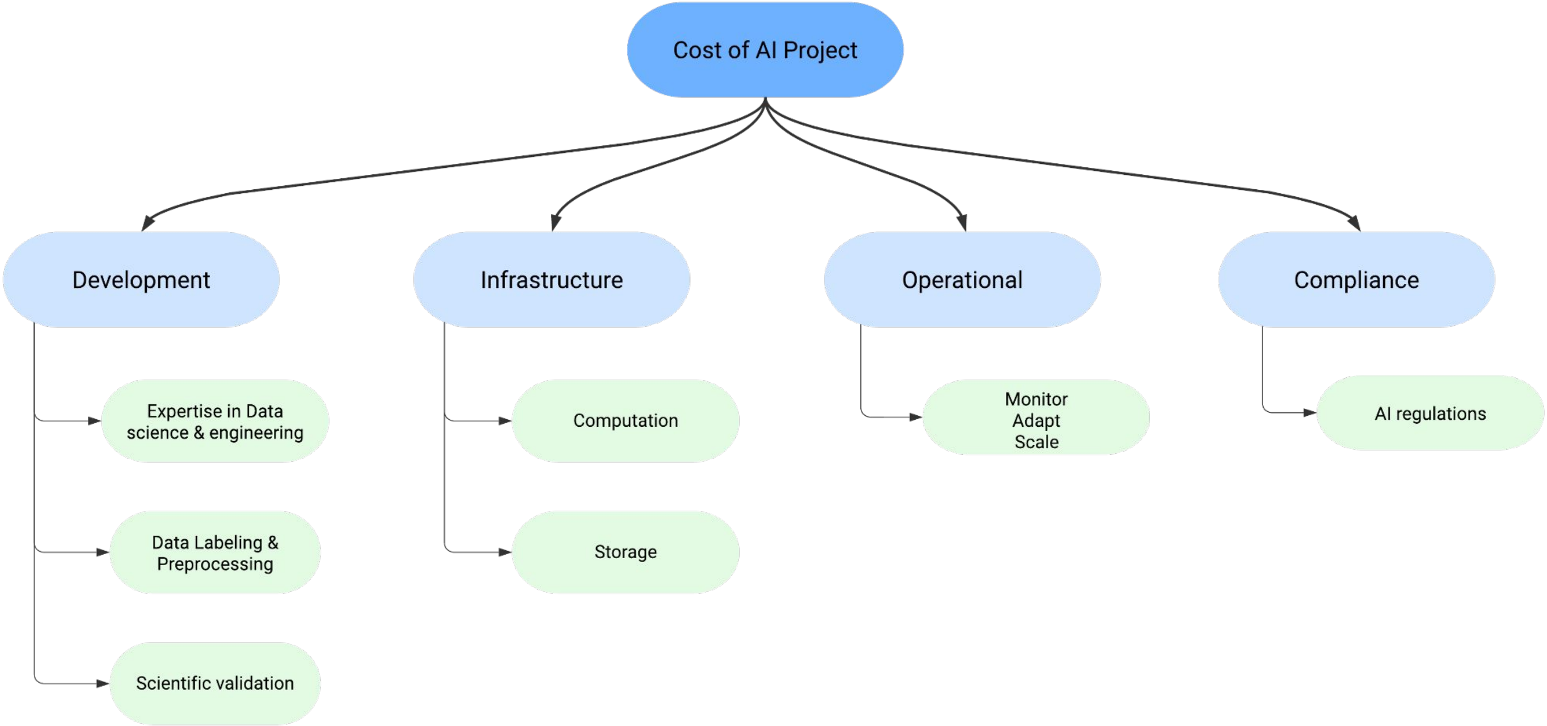
AI-powered Earth Insights

AI-powered Earth Insights is a system that leverages the first of its kind open-source geospatial AI foundation model developed by NASA and IBM Research. It uses the Harmonized Landsat Sentinel-2 Foundation (HLS) data and models that are fine-tuned on Flood mapping and Burn scar segmentation tasks. It allows users to inference on the fine-tuned models and visualizes the results.

ABOUT START EXPLORING →

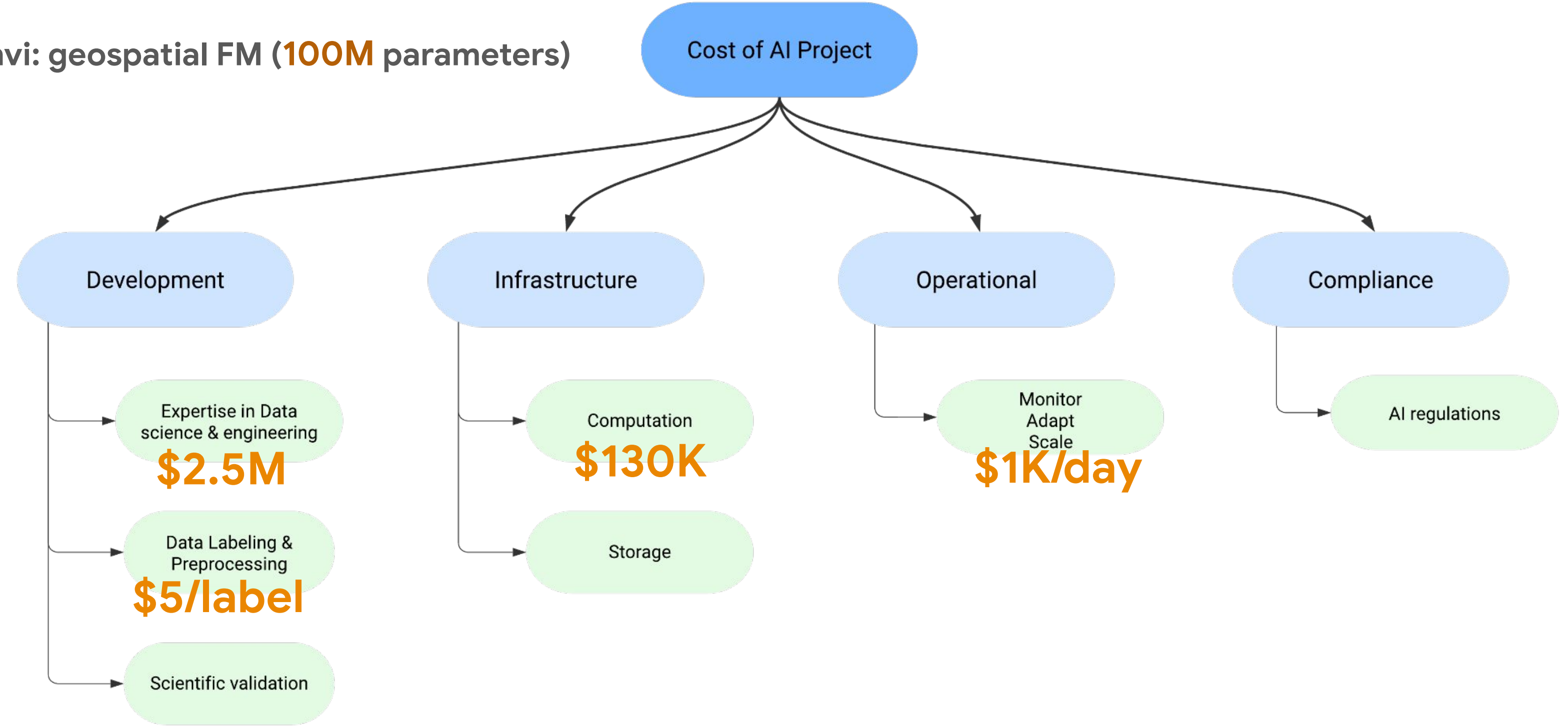
Challenges

AI today is expensive



AI today is expensive

Prithvi: geospatial FM (100M parameters)



Understanding of foundational knowledge

AI foundational knowledge is critical to develop scientific applications using AI

WIRED BACKCHANNEL BUSINESS CULTURE GEAR MORE SIGN IN SUBSCRIBE

Sloppy Use of Machine Learning Is Causing a 'Reproducibility Crisis' in Science

AI hype has researchers in fields from medicine to sociology rushing to use techniques that they don't always understand—causing a wave of spurious results.

Source: WIRED

Ethical issues

Bias and Fairness

Transparency and Explainability

Safety and Security

Human-AI Collaboration

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > comment > article

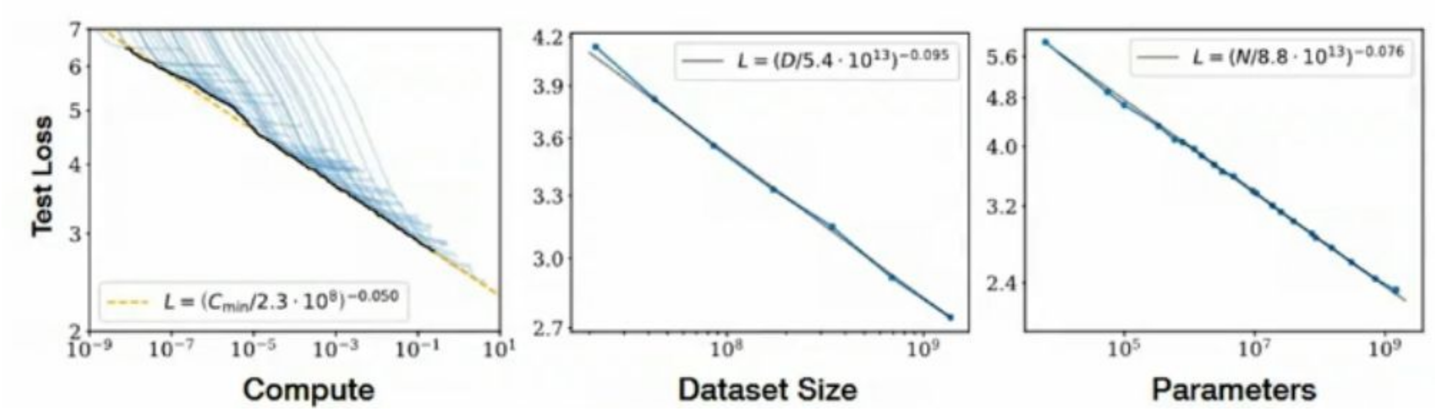
COMMENT | 31 October 2023

Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research

Artificial-intelligence tools are transforming data-driven science – better ethical standards and more robust data curation are needed to fuel the boom and prevent a bust.

Opportunities with Pre-trained Large Models
"value of small data"

Foundation models scale with **data**



Kaplan et al.

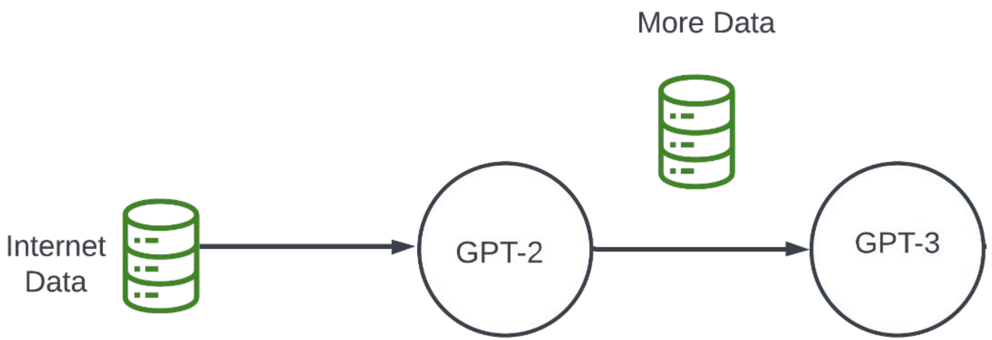
arXiv:2001.08361

GPT evolution

Internet
Data

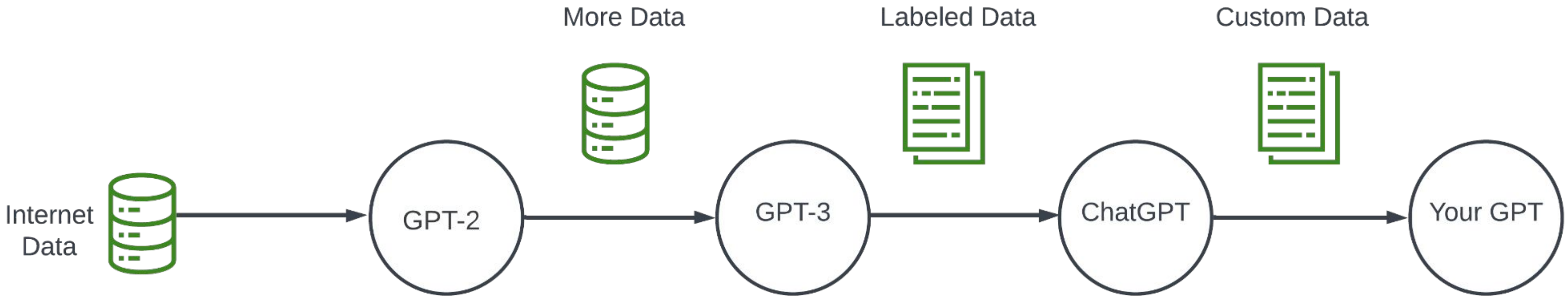
GPT evolution

GPT evolution

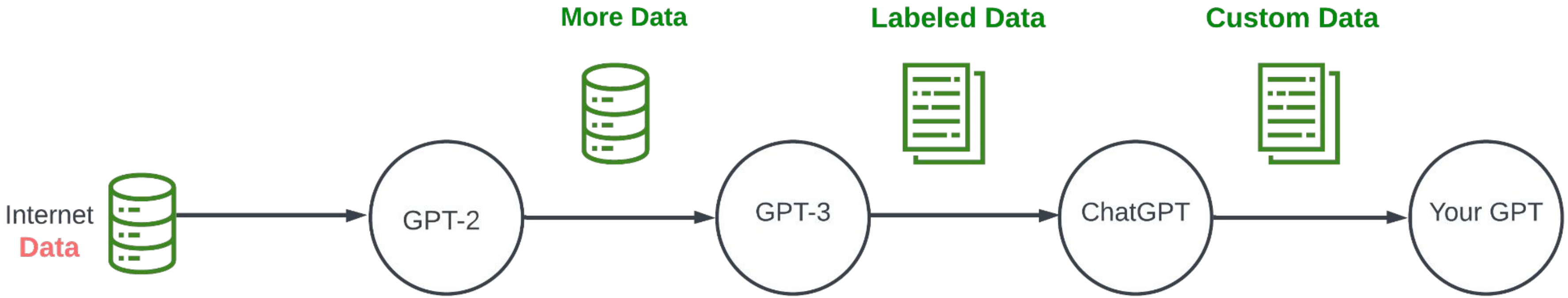


GPT evolution

GPT evolution



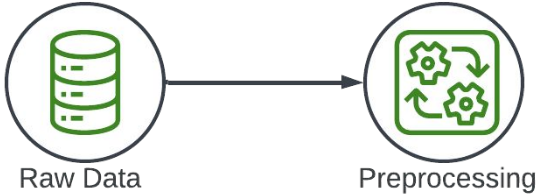
GPT evolution



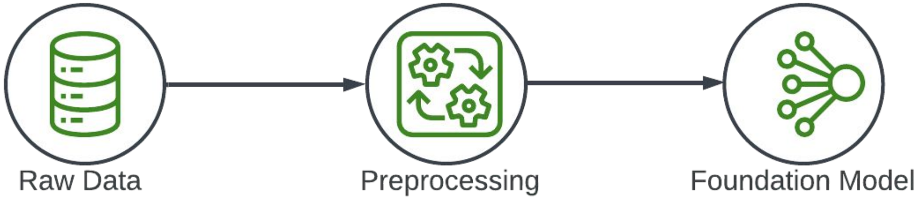
Data is the differentiator, model mostly constant

Foundation model

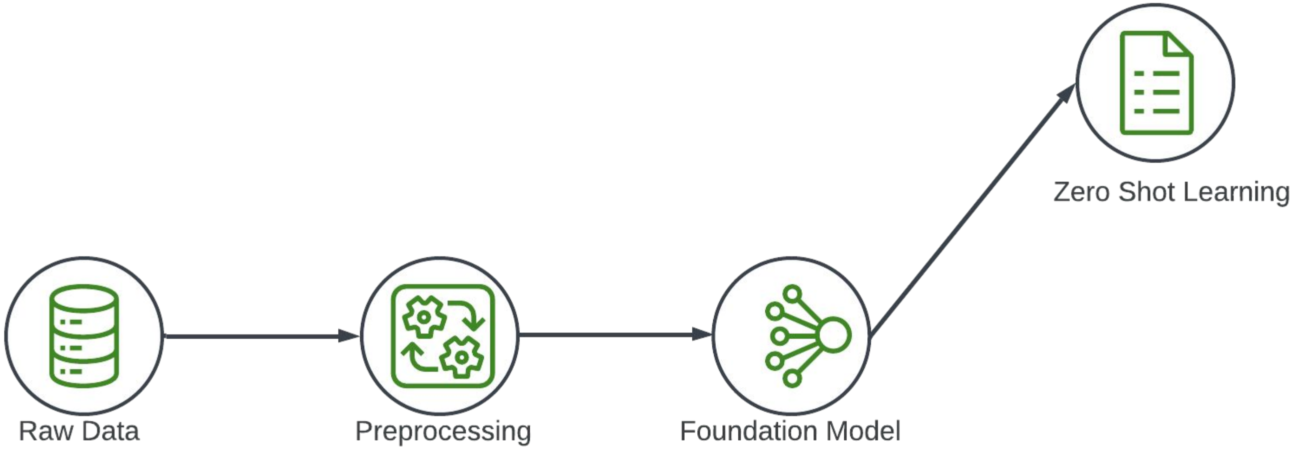
Foundation model



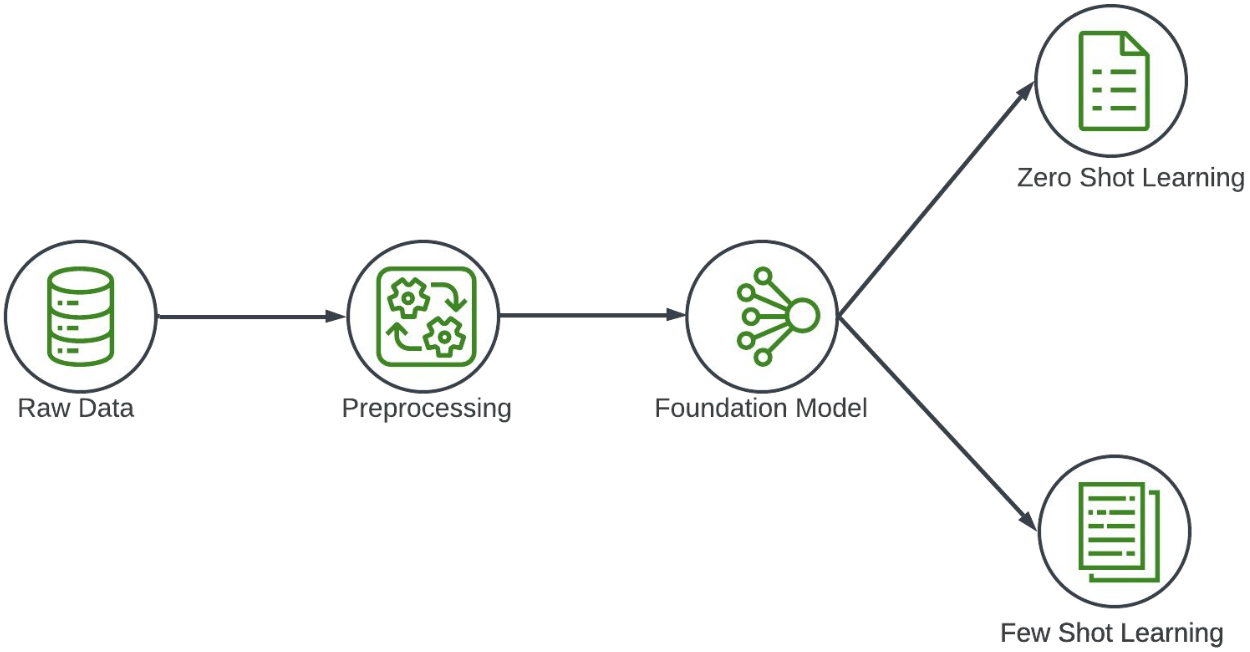
Foundation model



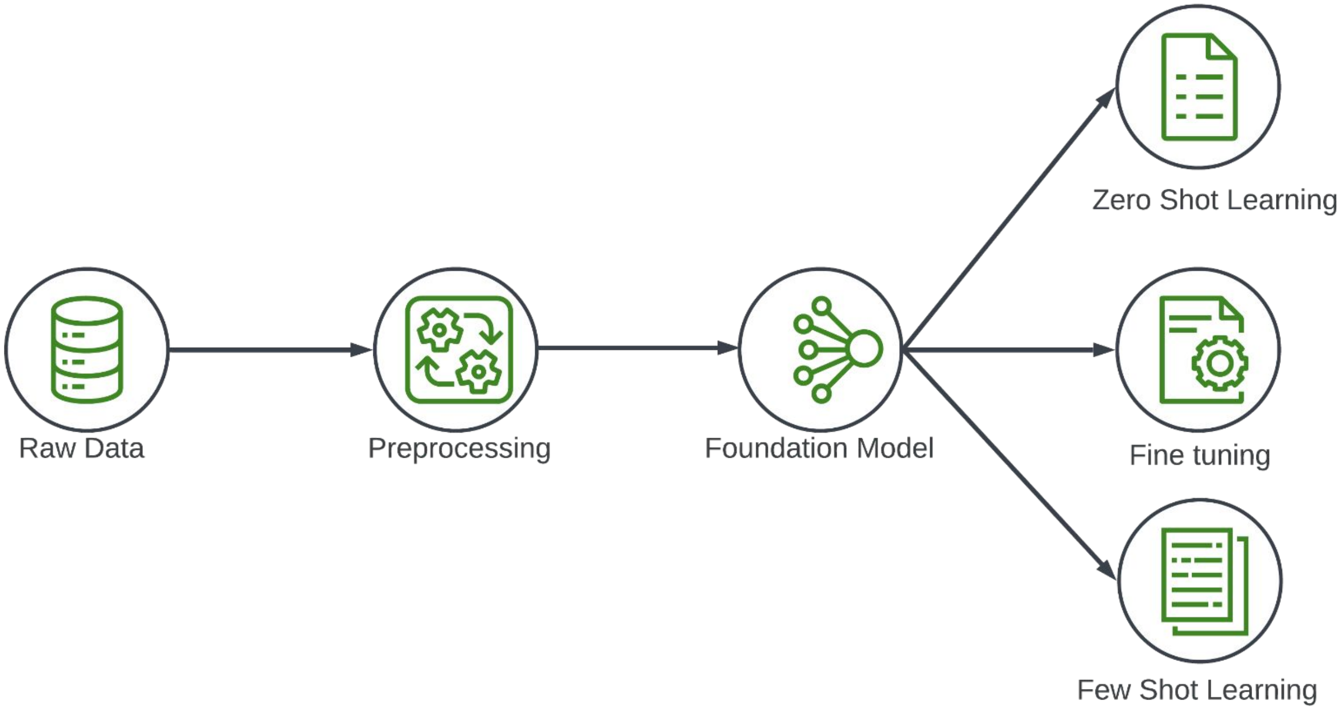
Foundation model



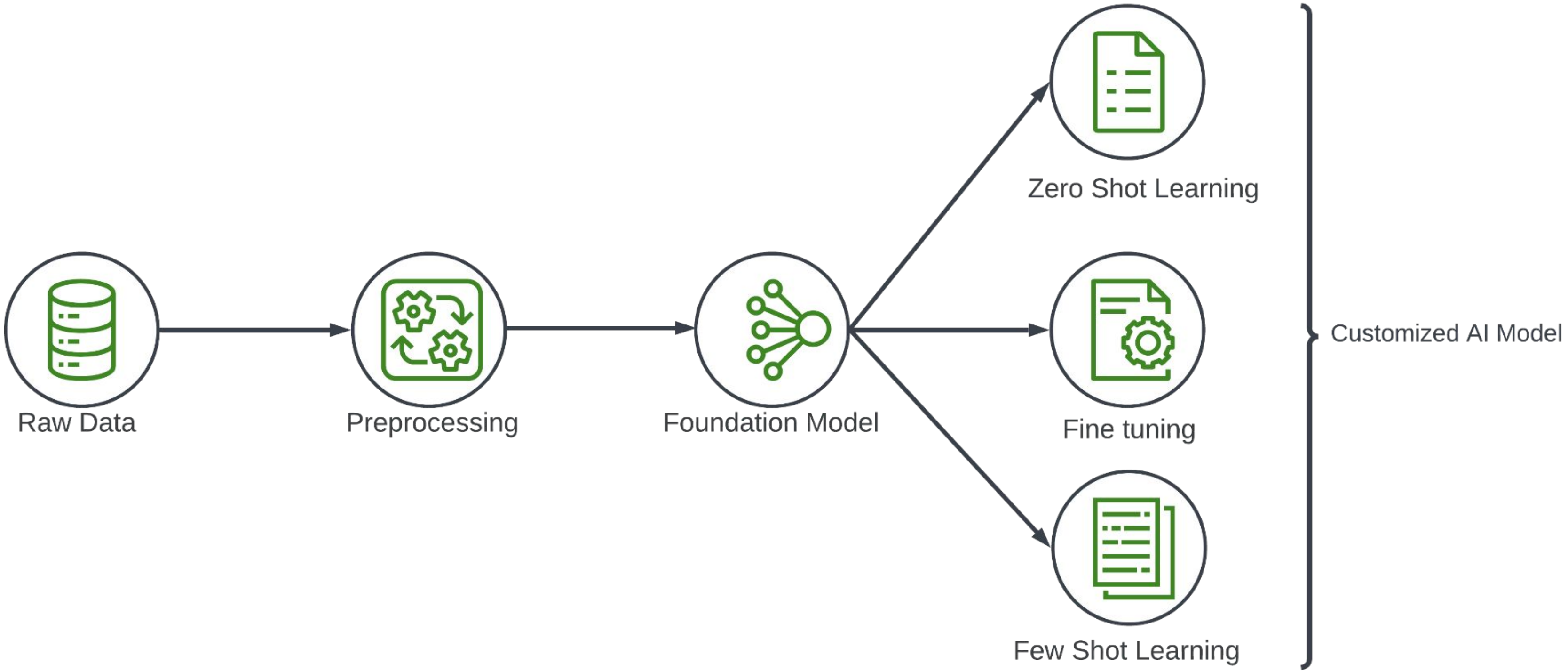
Foundation model



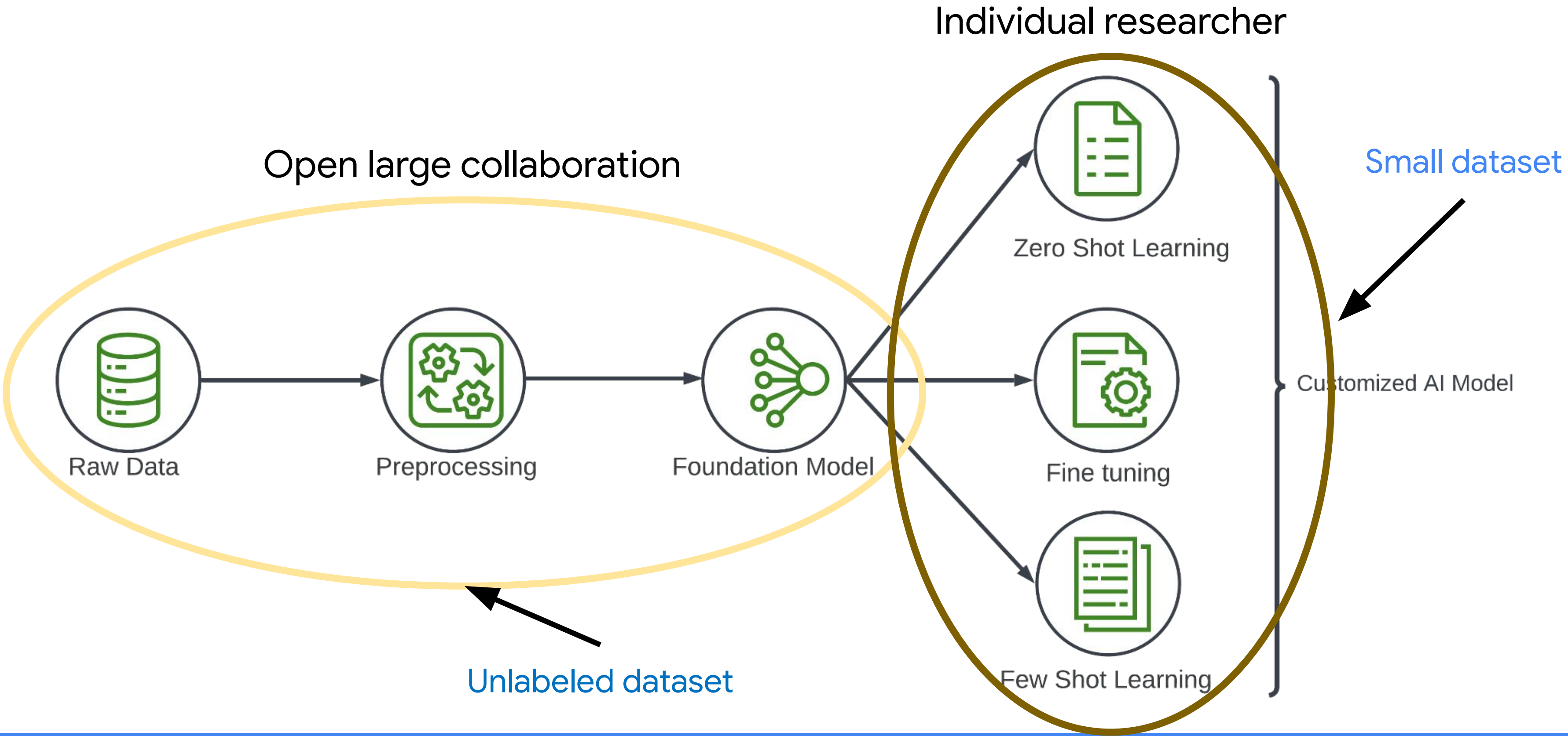
Foundation model



Foundation model

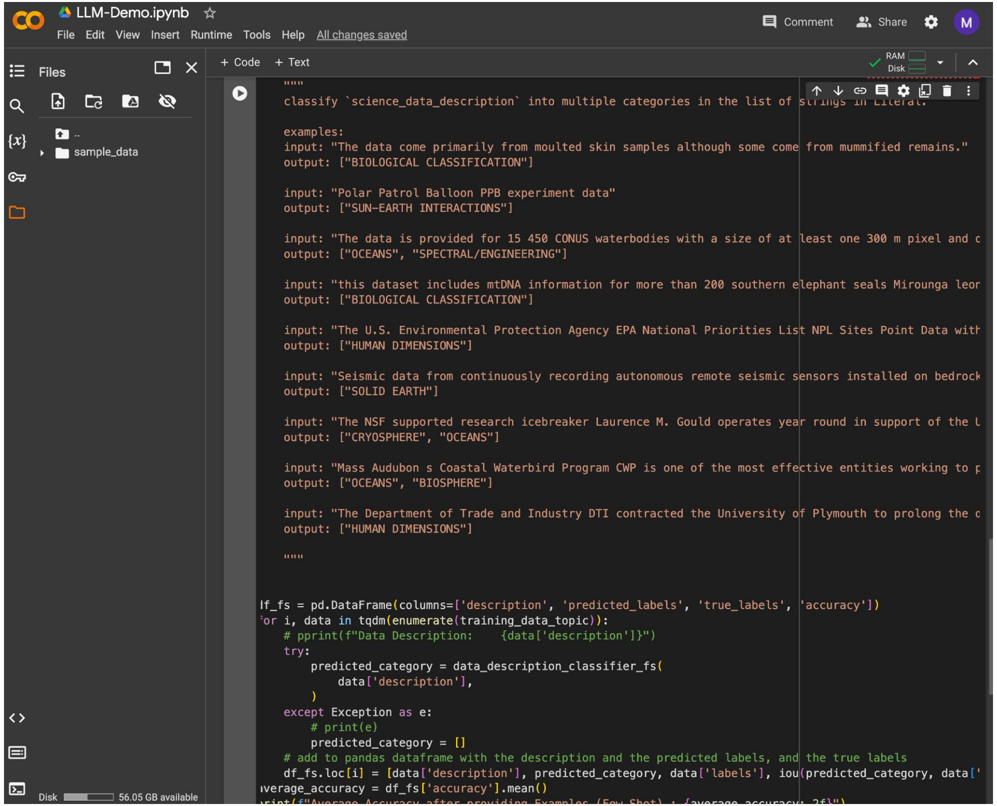


Foundation model

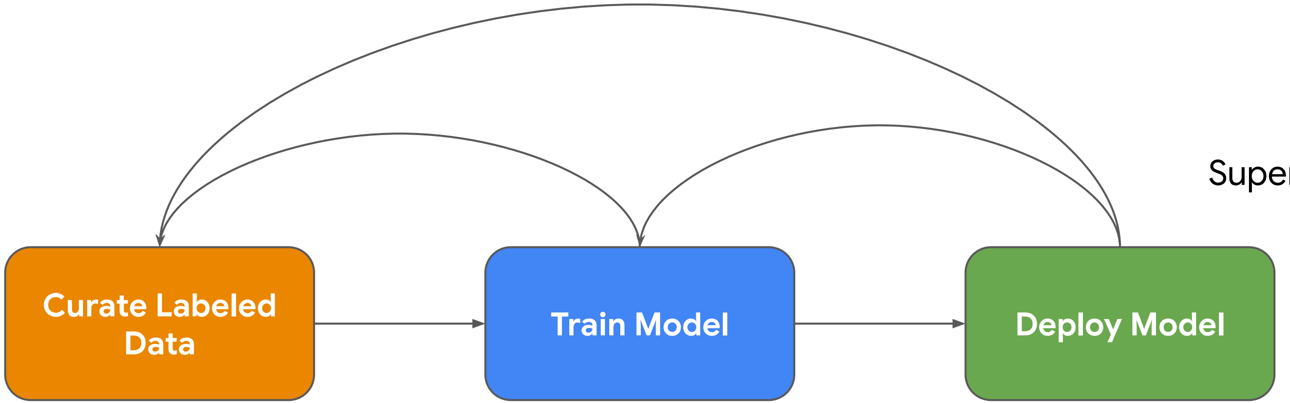


Demo - Live Notebook

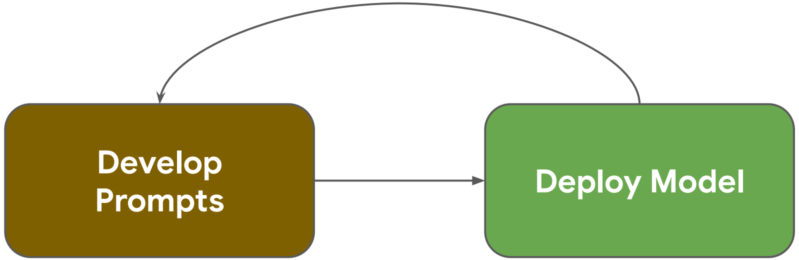
Relevance of small datasets when used with large pre-trained models



Changing paradigm



Supervised learning



Tuning on smaller data

Data & Pretrained Large Models

Foundation models are more data-centric than model-centric

Data requirement for new **custom** model is much smaller using foundation model: **Lower cost**

Foundation model for Science: Open data, Science quality, Transparent, Reuseable

Effectiveness of foundation models is directly tied to the volume of data they are trained on.

Foundation models can perpetuate and amplify biases present in the pretraining data > need for data governance policies (ethics)

The role of human input in data cleaning is critical for the quality of foundation models

Foundation models require new data inputs to stay current and relevant

What AI advances are needed to realize science goals?

AI = Algorithm + Data (lots of data):
Algorithms that can learn from less data?

Novel data management

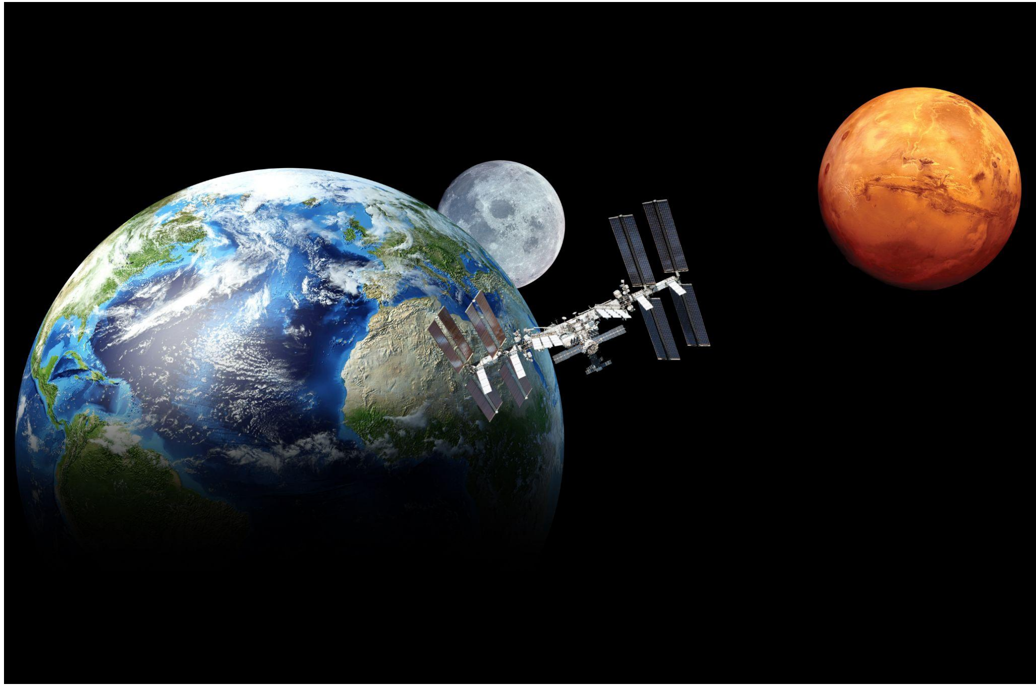
X-disciplinary tools

Optimization techniques

Ethical and XAI

Automated hypothesis generation

Affordable



Transforming Science with AI

AI Integration in every step of the scientific discovery and understanding

Responsible AI outputs

Efficient data utilization and model development

Collaboration and Open Science

Need for foundational knowledge

Infrastructure and ecosystem development

AI+Science @ NASA

Unlimited potential

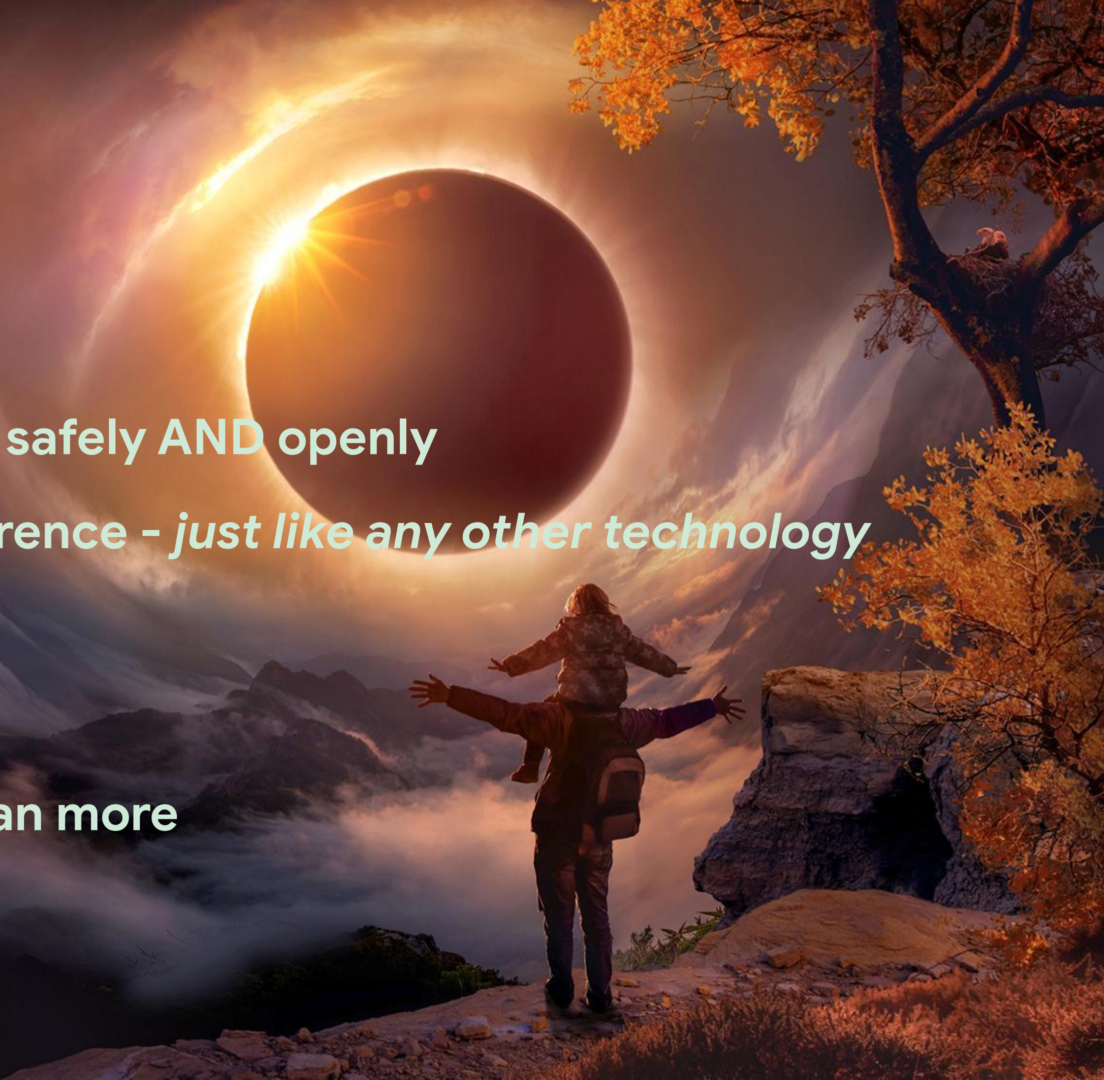
AI must be built responsibly and safely AND openly

How you use it makes a big difference - *just like any other technology*

Rigorous and careful evaluation

Must include diverse input

Making NASA’s science data mean more




Thank you