Biological Research and Space Health Enabled by Machine Learning to Support Deep Space Missions

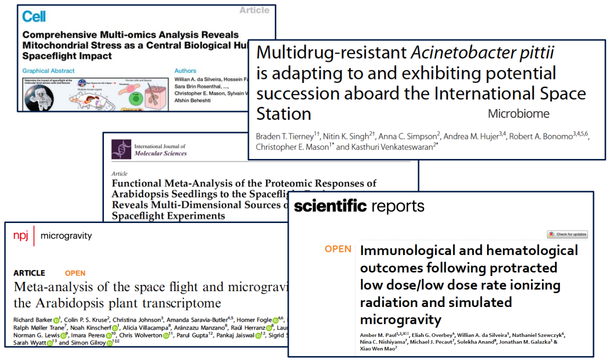
Ryan T. Scott
KBR I Science Lead
NASA Ames Life Sciences Data Archive
NASA Open Science Data Repository
POIS.1 Committee on Space Research 2024
Thursday, July 18, 2024, 10-10:30a
ryan.t.scott@nasa.gov

All authors: Lauren Sanders, Sylvain Costes

BPS
Biological & Physical Sciences

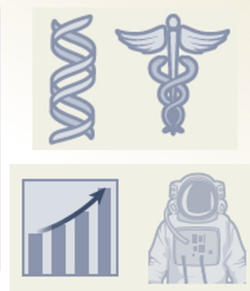
Reusable Data Essential for Basic, Applied Science, Operational Outcomes

Knowledge Discovery

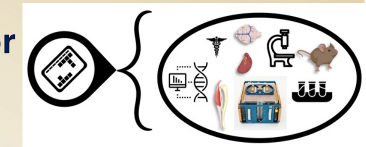


Enable Informatics, AI/ML, KG, DAG, AOP, Omic-Phenotypic Conclusions

Model development
Active and adaptive learning
Causal and network inference
Explainable AI
Federated learning
Few- or one-shot learning
Generative learning (synthetic datasets)
Knowledge graphs
Lifelong learning (living lab)
Out-of-distribution learning
Transfer learning



Telemetry for Biomedical Context

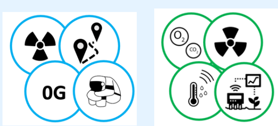


Analysis Across Missions

RR1 vs RR6 vs RR10 vs RR23

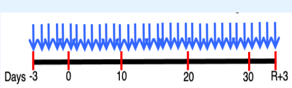
Accumulated vs Total vs GCR vs SAA

CO_2 vs O_2 vs RH vs Temperature



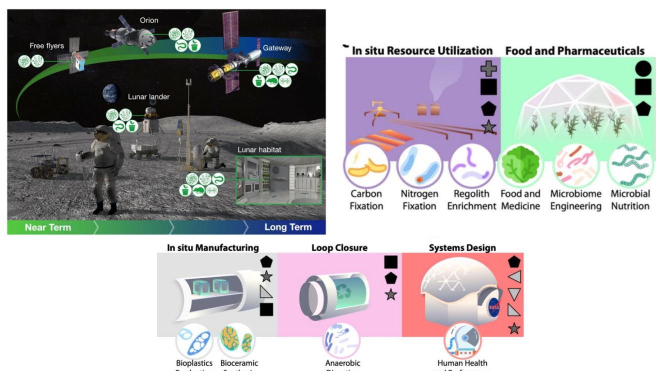
Across Hazards

Across Species, Multimodal Data Types

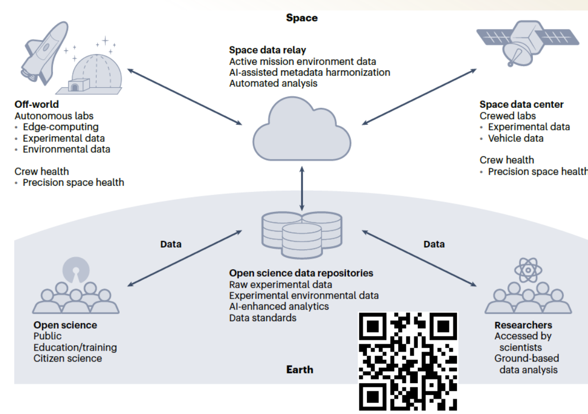


Longitudinal, Biomonitoring, & Experimental

Novel Ecosystem, Platform, Countermeasure Development, Health Risk Quantification



Involve the World in Spaceflight



Scott et al., 2023 https://rdcu.be/c8jSO;
Sanders et al., 2023 https://rdcu.be/c8jSS

July 18, 2024 – #COSPAR2024
ryan.t.scott@nasa.gov

Berliner et al., 2022 Communications Engineering;
Everroad et al., 2021 https://ntrs.nasa.gov/citations/20210023324)

NASA OSDR – By The Numbers

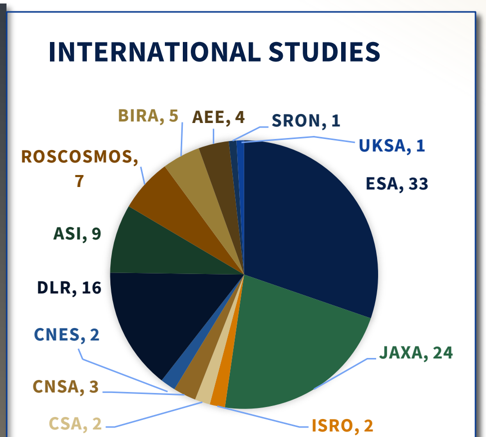
509
Studies

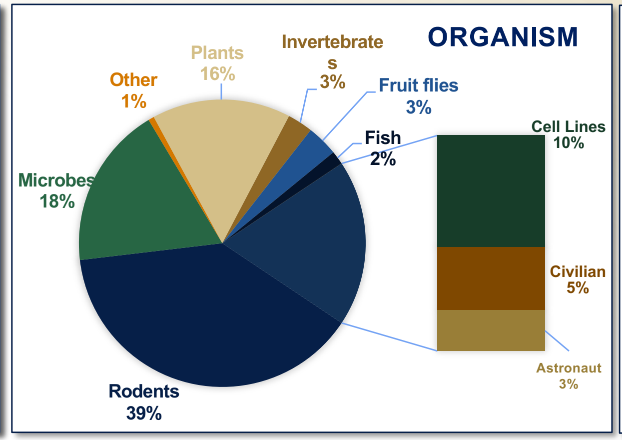
969
Datasets

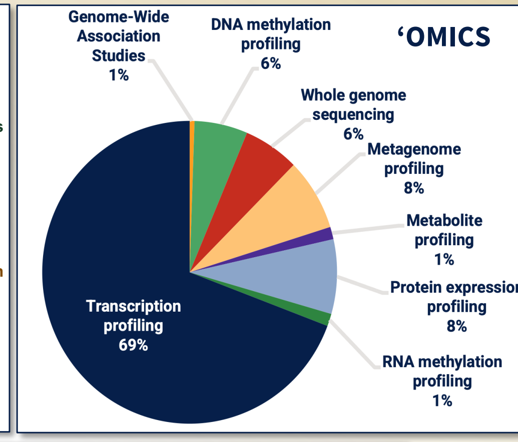
45
Species

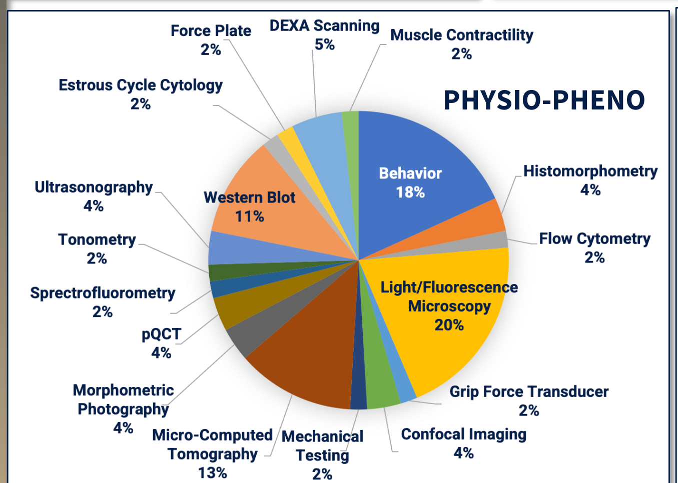
>65
Assays

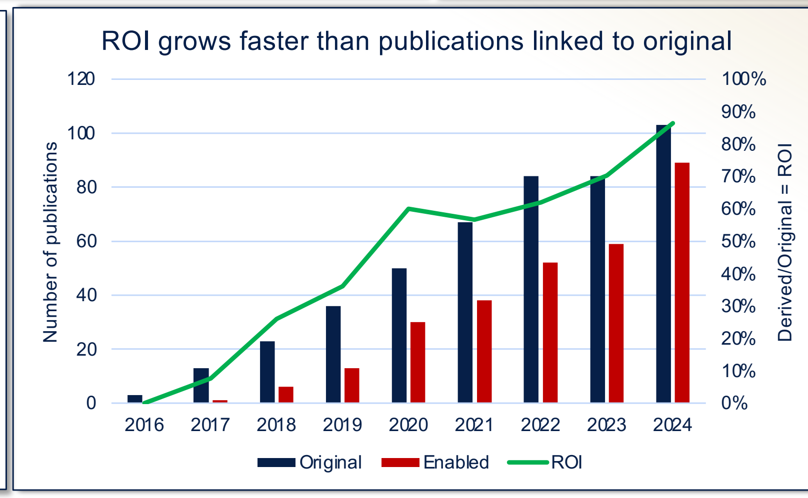
>160TB
Data











89
Enabled Publications linked to OSDR

103
Original Publications linked to OSDR

150+
Datasets used in enabled publications

ryan.t.scott@nasa.gov

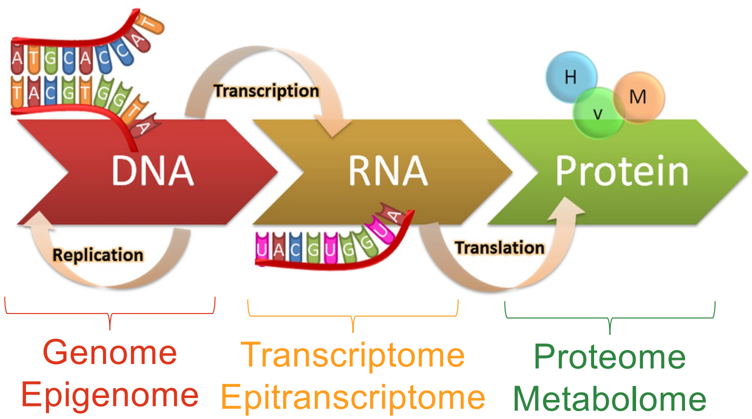
ML and Space Biology/Space Medicine

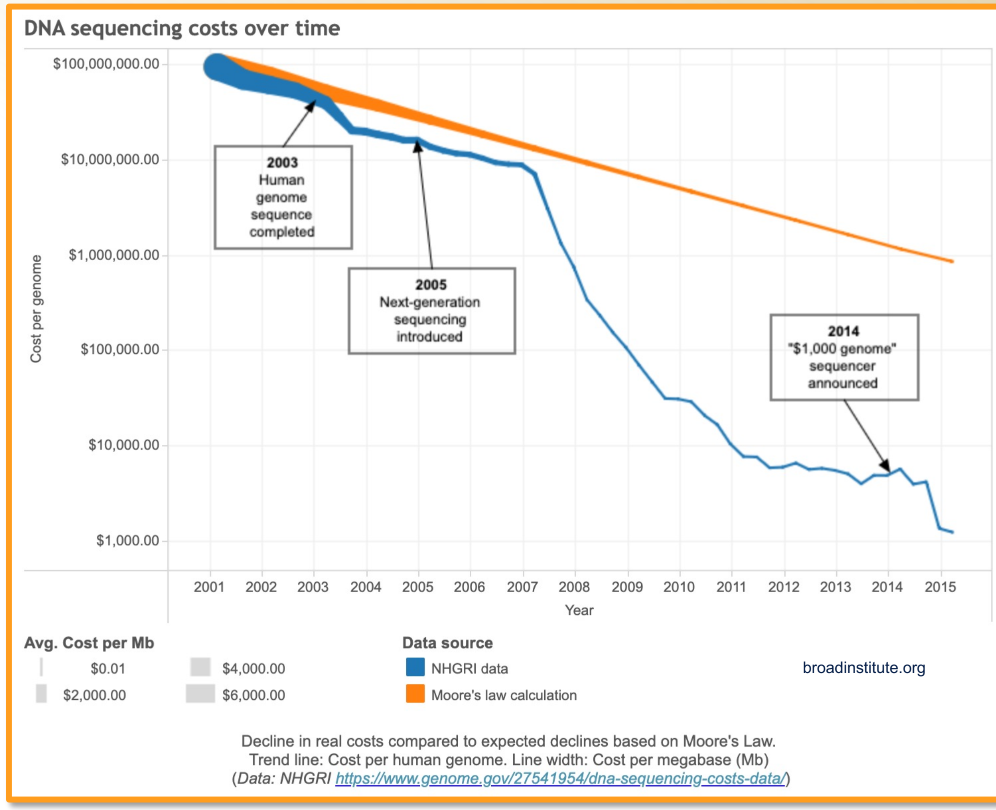
The Challenges of High Complexity, High Dimensionality, Low Sample Size Data

ryan.t.scott@nasa.gov

Paradigm Shift in Biological Data Generation

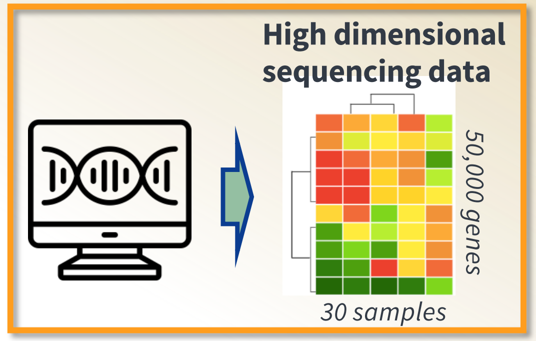
- **Traditional molecular biology** studies a few genes or proteins at a time through
- **High-throughput sequencing ('omics)** gives a readout of the entire genome in a cell or tissue sample





https://genius.com/Biology-genius-the-central-dogma-annotated

Space Biological Data Analysis Challenges

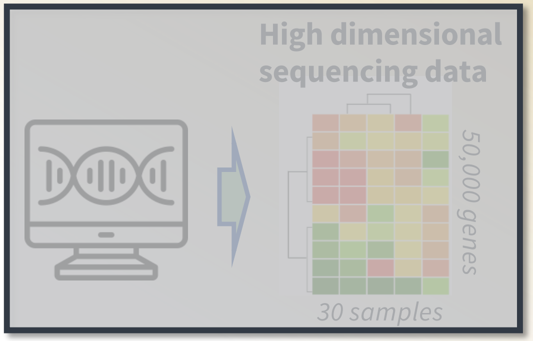


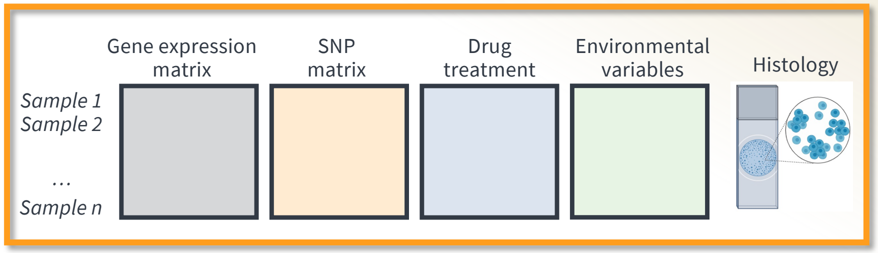
Space Biological Data Challenges

- **Small sample *n***
- **High feature count**
- Heterogeneous data
- Sparse data
- Transfer from model to human

Space Biological Data Analysis Challenges

Space Biological Data Challenges
- Small sample *n*
- High feature count
- **Heterogeneous data**
- **Sparse data**
- Transfer from model to human





ryan.t.scott@nasa.gov

Space Biological Data Analysis Challenges

Space Biological Data Challenges
- Small sample *n*
- High feature count
- Heterogeneous data
- Sparse data
- **Transfer from model to human**

ryan.t.scott@nasa.gov

Complex (Biological) Systems Need Complex Models

STATISTICAL METHODS

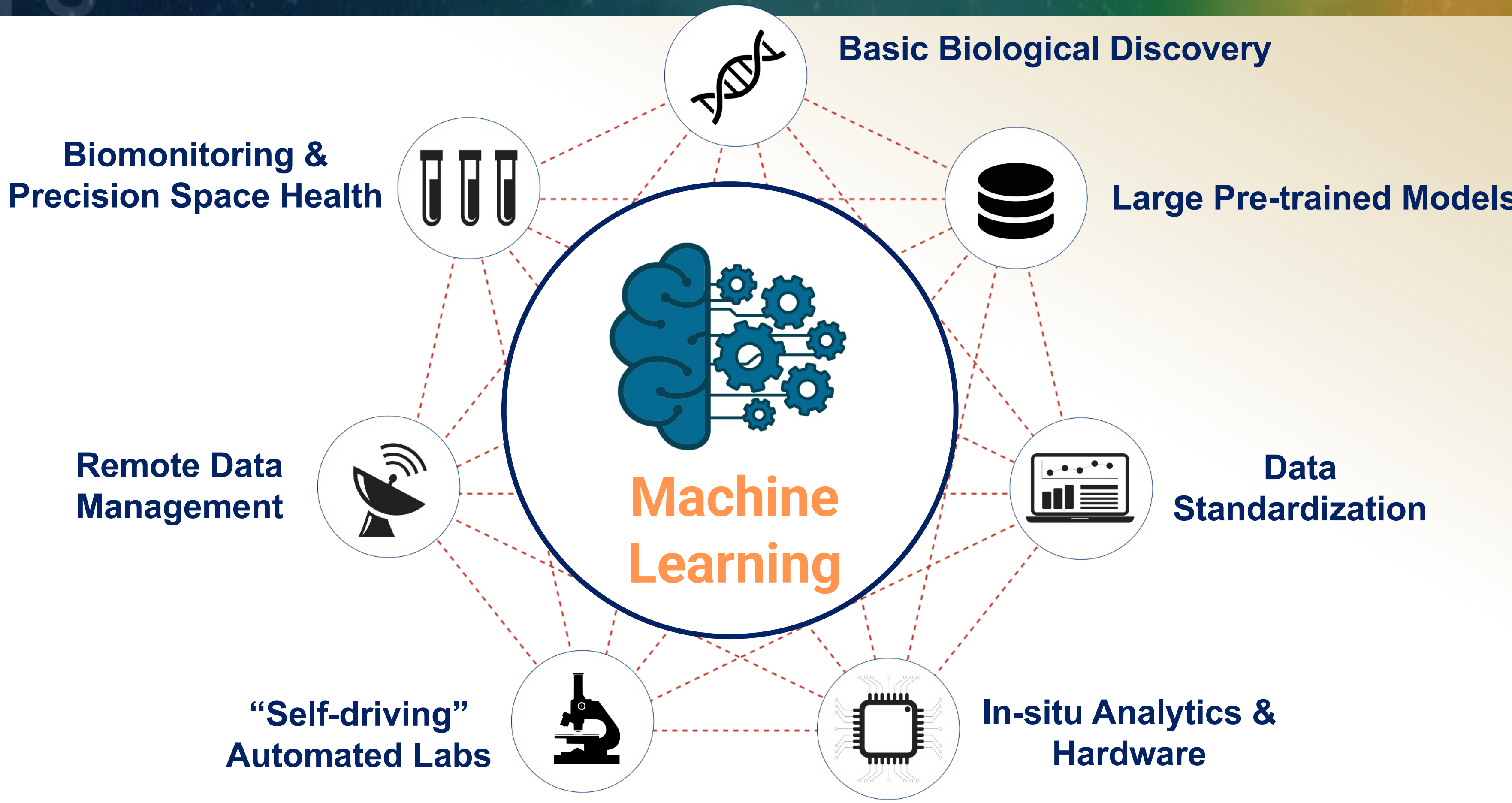
- Draw conclusions from observed data (**inference**)
- Assume specific data distributions
- Examples: hypothesis testing, correlative analysis

MACHINE LEARNING

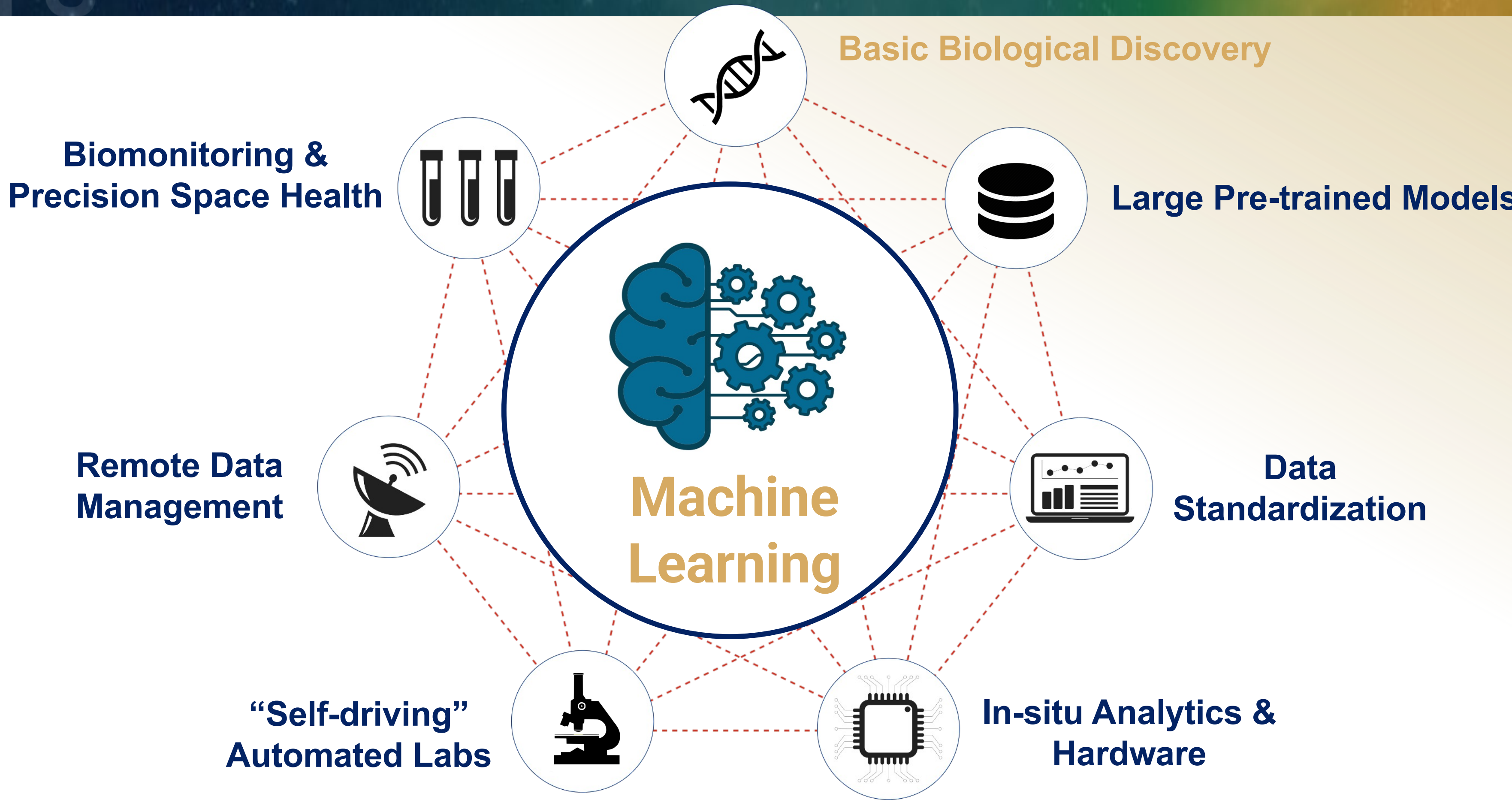
- Learn from data to make predictions on unseen data (**prediction**)
- Able to model nonlinear relationships without assuming a data distribution
- Examples: classification, regression, clustering

Bzdok et al., *Nature Methods* 2018; PMID: 30100822

ryan.t.scott@nasa.gov



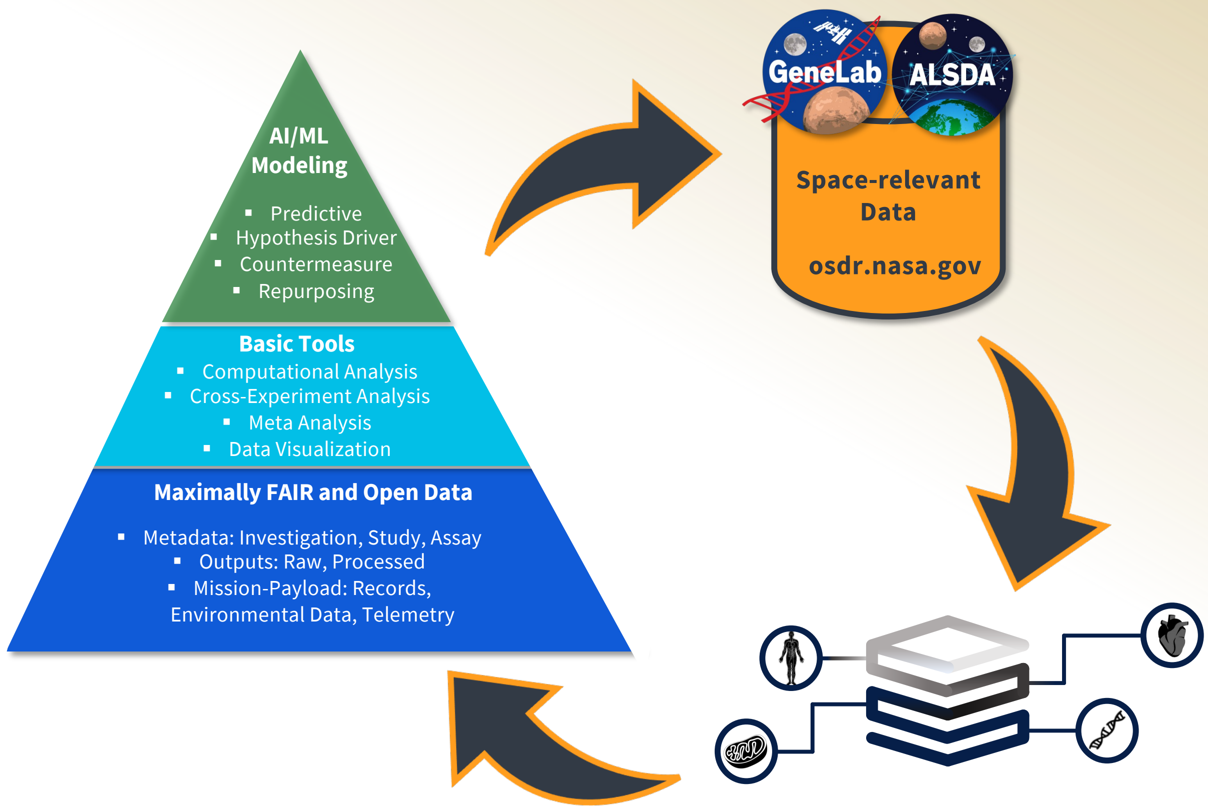
ryan.t.scott@nasa.gov



ryan.t.scott@nasa.gov

“AI for Life in Space” Working Group: AI4LS

Leveraging ML and AI methods to model space biology data from the NASA Open Science Data Repository: **NASA GeneLab** (omics) and **NASA Ames Life Sciences Data Archive** (ALSDA; phen-omics) to better understand the complex effects of spaceflight on living systems across hierarchical biological levels.



Omic to Phenotypic ML Study – Rodent & Muscle

Kevin Li
UC Berkeley
(Now MIT-Harvard PhD)

- RR1 & RR9 Muscle Data
- Qlattice/Abzu
- Li et al., 2023; npj μG

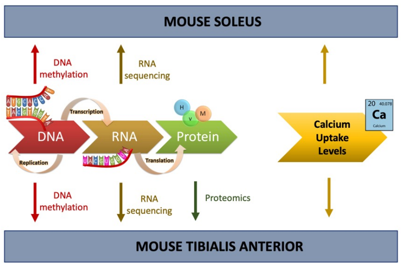
In spaceflown mice,

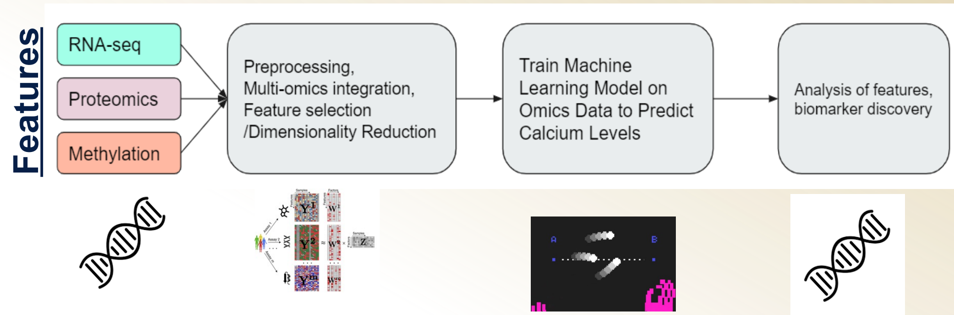
calcium uptake efficiency *decreases* in **soleus** muscle…

but *increases* in **tibialis anterior** muscle

Open Science for Life in Space

OSDR Datasets:
OSD-104
OSD-105
OSD-488



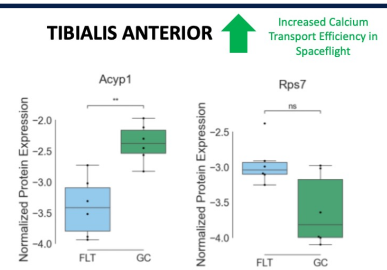


b

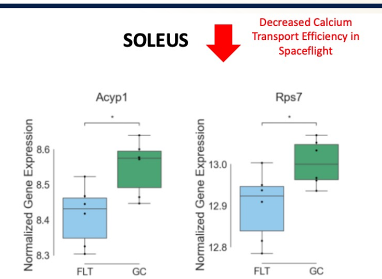
Feature	Models (n)
Acyp1 (proteomics)	89
Rps7 (proteomics)	27
Cct6a (proteomics)	5
Glt28d2 (RNA-seq)	4

c

T1 CV R^2	0.894
T10 CV R^2	0.711
RNA-seq features (n)	12
Proteomic features (n)	38



- Decreased **Acyp1** in flight allows increased calcium transport
- Increased **Rps7** in flight shows low nitrosative stress



- Decreased **Acyp1** in flight means decreased calcium transport
- Decreased **Rps7** in flight shows high nitrosative stress

Li et al., 2023; PMID: 38092777

ryan.t.scott@nasa.gov

Omic to Phenotypic ML Study – Rodent & Liver

James Casaletto
BMSIS/AI for Life in Space

AI4LS Artificial Intelligence for Life in Space

ALSDA

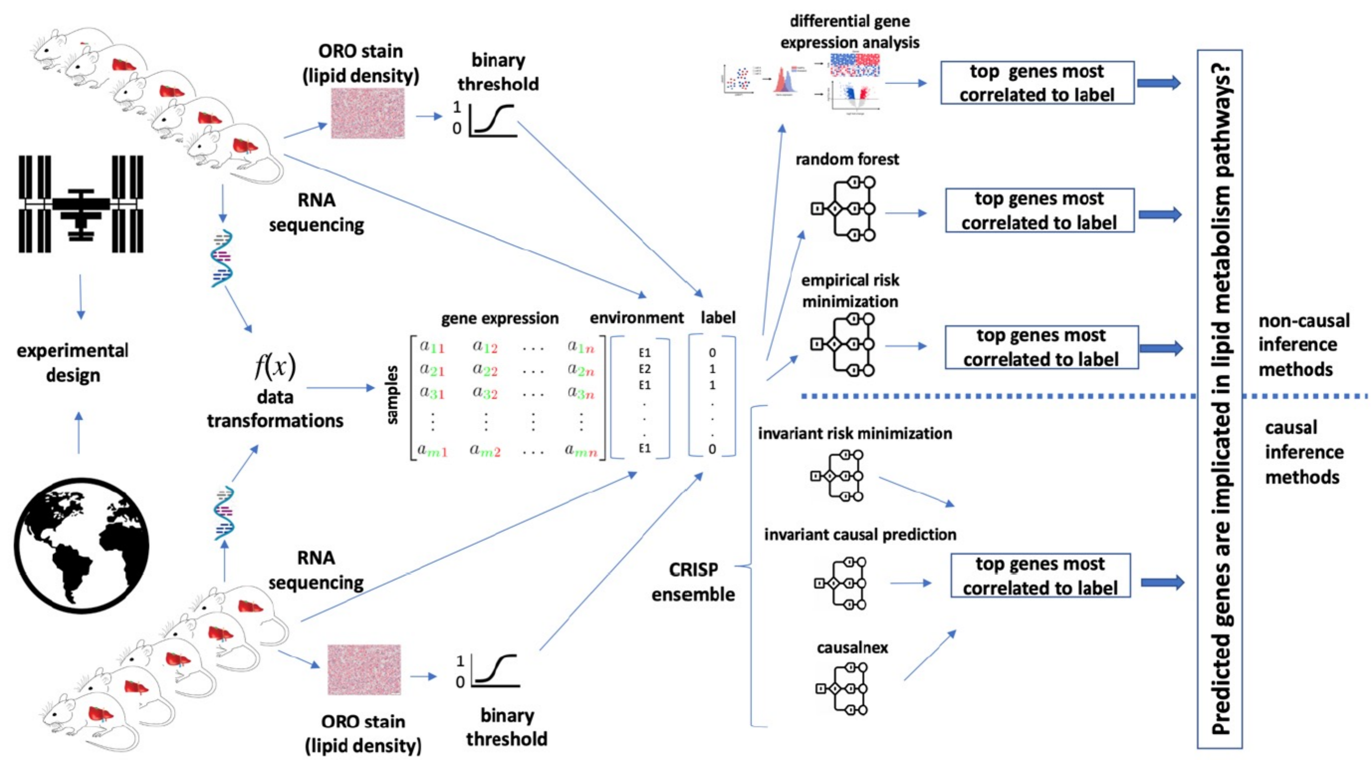
GeneLab

NASA Open Science for Life in Space

OSDR Datasets: OSD-47, OSD-48 OSD-137, OSD-158

- Transcriptomics & Histology
- Part of 2024 Nature Package
- In review: Scientific Reports

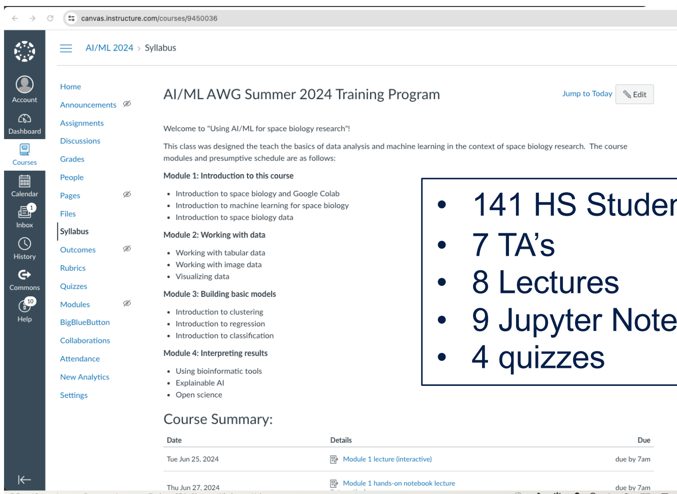
Casaletto et al., 2023 pre-print rsrch sq
https://doi.org/10.21203/rs.3.rs-2332064/v1



Open Access Course to Learn AI/ML for Space Biology

Training in Artificial Intelligence and Machine Learning for Space Biological Sciences Using NASA Cloud-Based Data, 2023-24

Summer of AI synchronous training



- 141 HS Students
- 7 TA's
- 8 Lectures
- 9 Jupyter Notebooks
- 4 quizzes

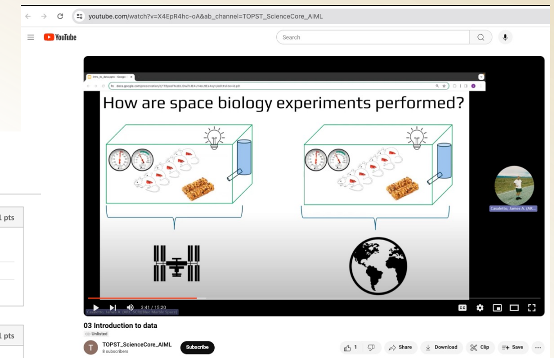
James Casaletto
BMSIS/ AI4LS

Lauren Sanders
Ames/ AI4LS

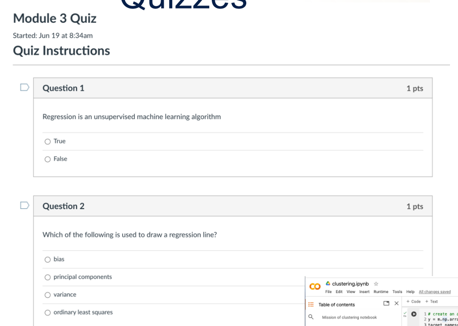
AI4LS
Artificial Intelligence for Life in Space

On-Demand, Public, Free Asynchronous

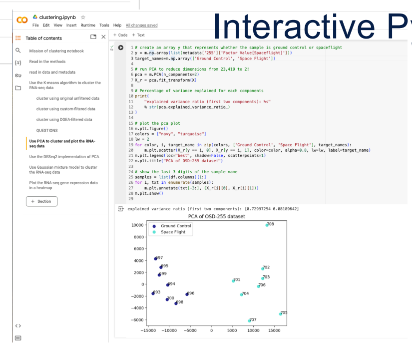
Pre-recorded lectures



Quizzes

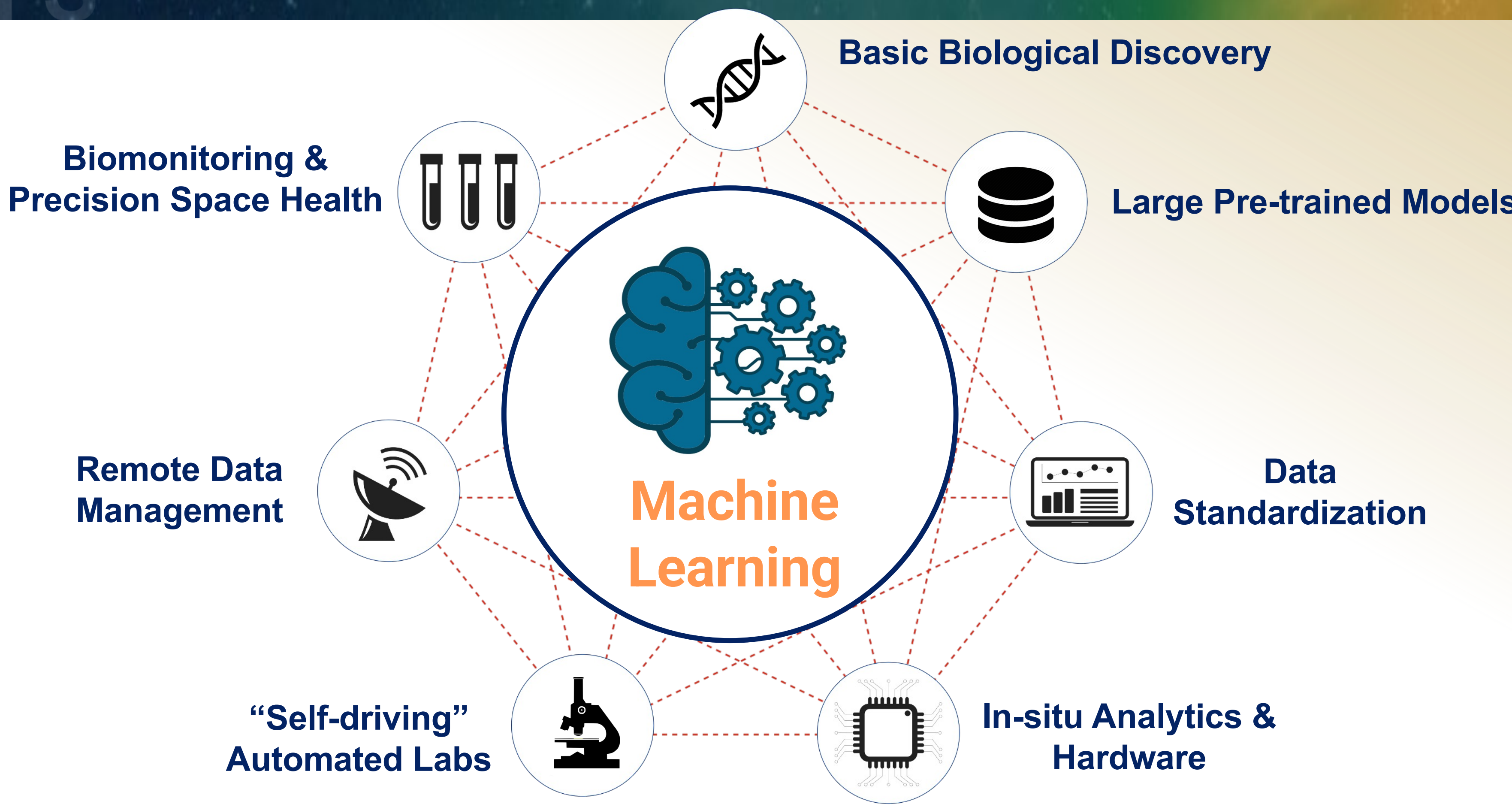


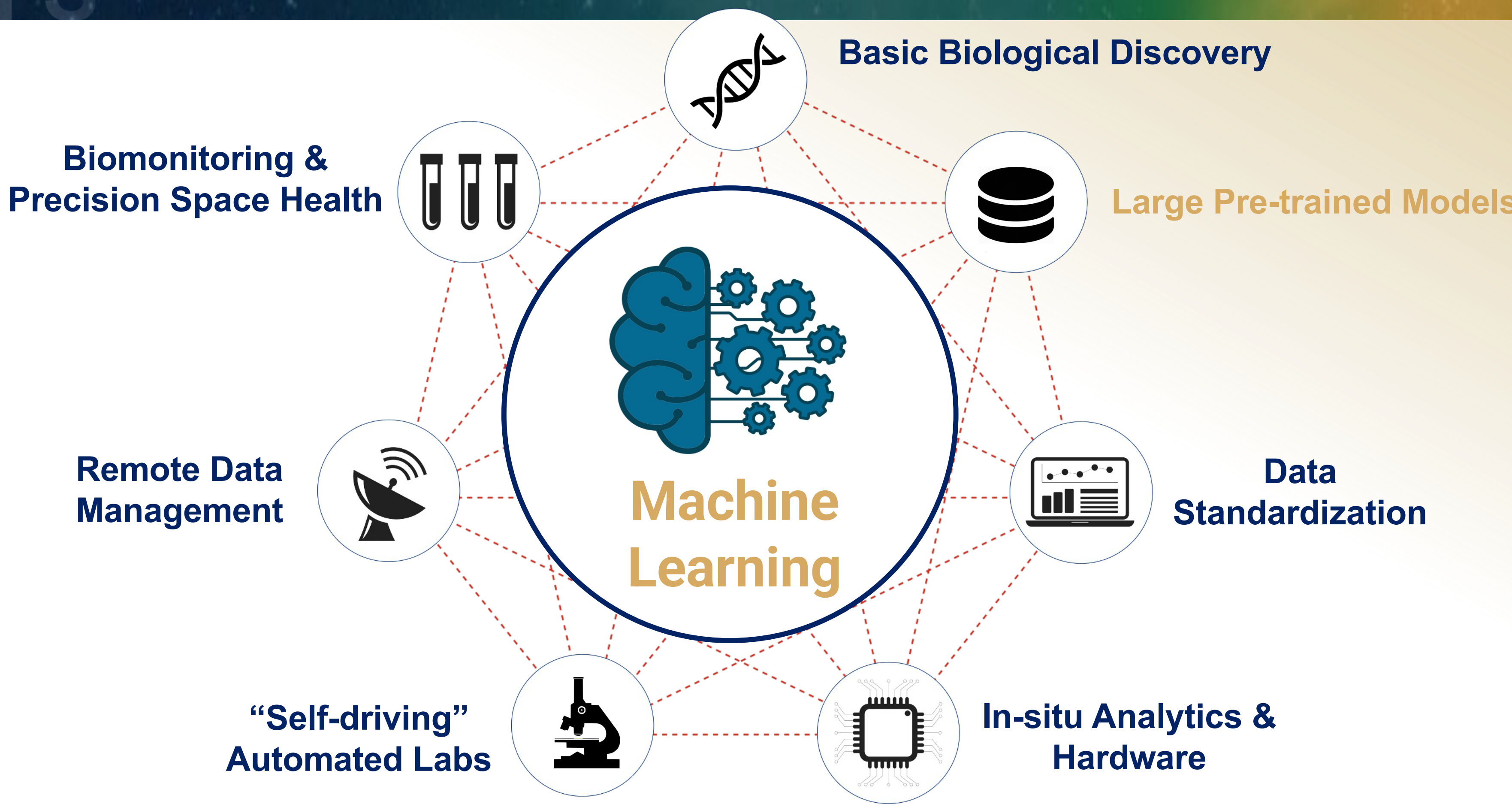
Interactive Python Notebooks



Sign-up!

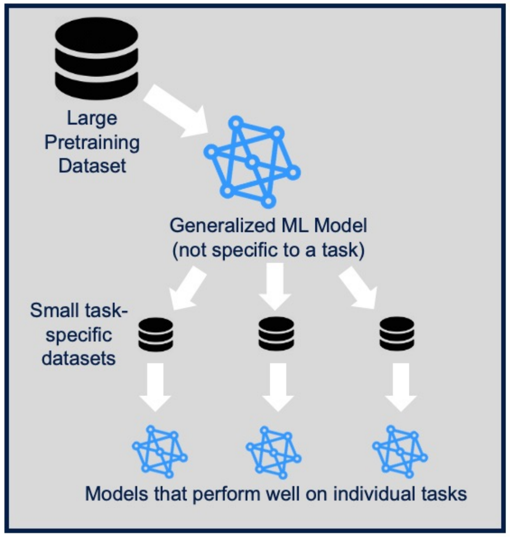
ryan.t.scott@nasa.gov

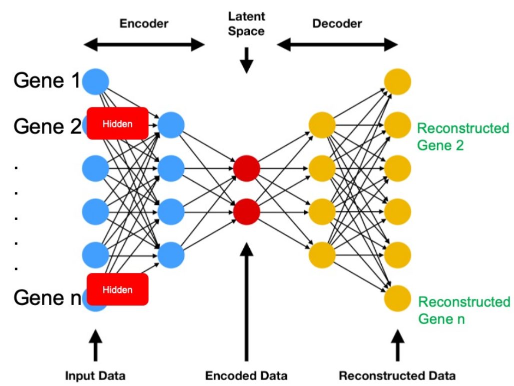

Basic Biological Discovery
Biomonitoring & Precision Space Health
Large Pre-trained Models
Machine Learning
Remote Data Management
Data Standardization
"Self-driving" Automated Labs
In-situ Analytics & Hardware

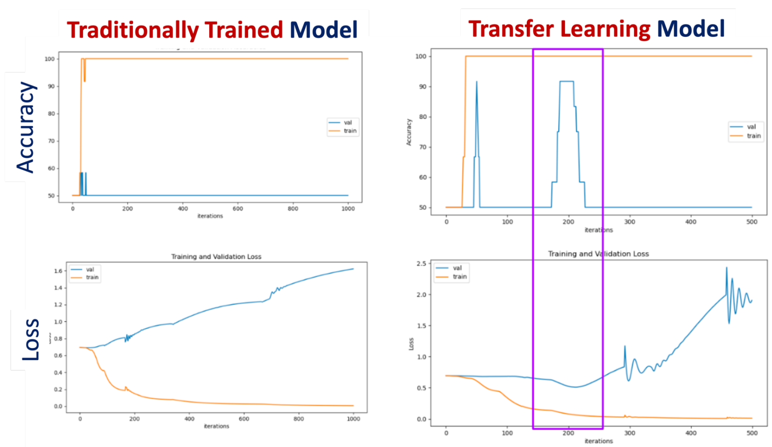

Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
"Self-driving"
Automated Labs
In-situ Analytics &
Hardware

Transfer Learning: Pretrained Models

Large Pre-trained Models *to Connect Biomedical Knowledgebases with* ***Small Spaceflight Datasets***







- **Deep learning model architecture:** scBERT: encoder-decoder
- Self-superviused on unlabled data

recount3

- Pretrained a large encoder to learn gene-gene interaction networks
- Tested the trained model on a downstream, supervised task using a tiny space biology dataset
- Pretrained model outperforms traditionally trained model
- Future vision: "model zoo" of many pretrained models available to the space biology research community

Kevin Li
UC Berkeley
(Now MIT-Harvard PhD)

Lauren Sanders
Ames/ AI4LS

July 18, 2024 – #COSPAR2024
ryan.t.scott@nasa.gov

Using LLMs for OSDR Chatbot, Curation

Aiding Discoverability & Searchability

Walter Alvarado
NASA Ames

Open Science for Life in Space

Display Prompt

What are the effects of space flight on mice liver?

The effects of space flight on mice liver include possible changes in metabolic pathways and alterations in protein levels. However, no significant alterations in protein levels were observed in the livers of spaceflight mice compared to control mice on Earth.

Sources:

https://osdr.nasa.gov/bio/repo/data/studies/OSD-209/

https://osdr.nasa.gov/bio/repo/data/studies/OSD-345/

Conversational Agents: Employs an LLM to enable efficient data retrieval and user assistance

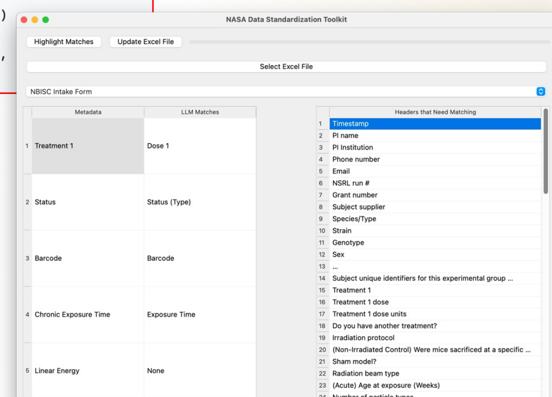
Automated Data Alignment & Verification

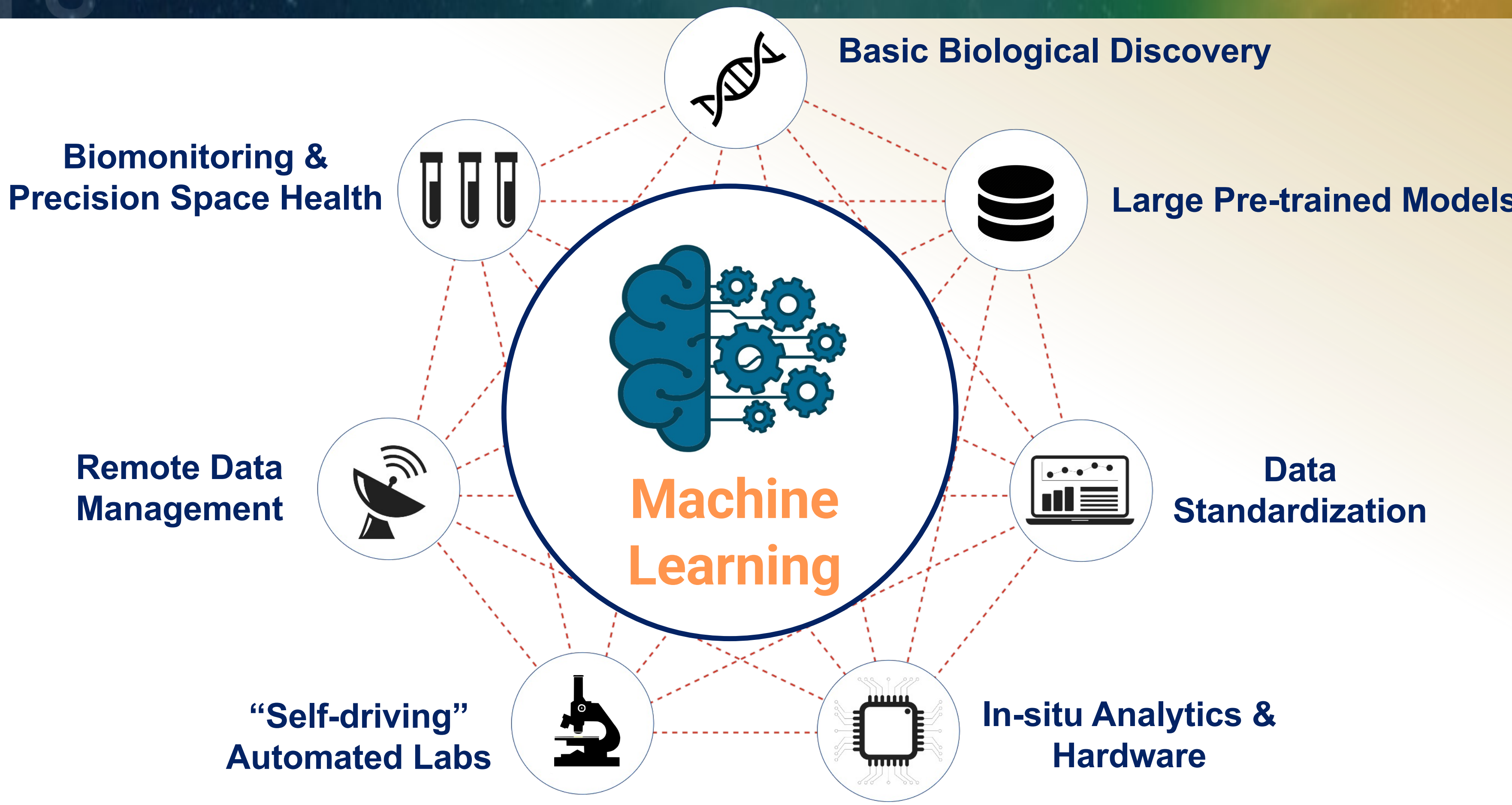
Column in File1	Best Match in File2	Similarity Score	Match Status
Euthanasia method	Euthanasia method	100	Above Threshold
Sample preservation method	Sample preservation method	100	Above Threshold
Biospecimen category	Biospecimen category	100	Above Threshold
Sex	Sex	100	Above Threshold
Genotype	Genotype	100	Above Threshold
Absorbed dose rate	Absorbed dose rate	97	Above Threshold
Number of fractions	Number of fractions	97	Above Threshold
Time point of sacrifice post irradiation	****Time point of sacrifice post-irradiation	93	Above Threshold
Additional notes	Additional Notes	91	Above Threshold
Absorbed dose rate (Unit)	Absorbed dose rate unit	88	Above Threshold
Total absorbed dose (Unit)	Total absorbed dose units	88	Above Threshold
Time between exposures (Unit)	Time between exposures (Hour)	86	Above Threshold
Absorbed dose per fraction	Total absorbed dose per fraction	86	Above Threshold
Time between fractions (Days)	Time between fractions value	84	Above Threshold
Sample storage temperature	Sample storage temperature (Celsius)	84	Above Threshold
Energy (Unit)	Energy unit	83	Above Threshold
Absorbed dose per fraction (Unit)	Total absorbed dose per fraction units	82	Above Threshold
Dose 1 (non-radiation)	No Match Found	0	Below Threshold

```
Matching: Dose 1 (non-radiation)
LLM Match: Treatment 1 dose
Match found!
 Dose 1 (non-radiation) <---> Treatment 1 dose
--------------
Matching: Status (stts_name)
LLM Match: Status of samples (Available, Not-available)
Match found!
 Status (stts_name) <---> Status of samples (Available,
--------------
```

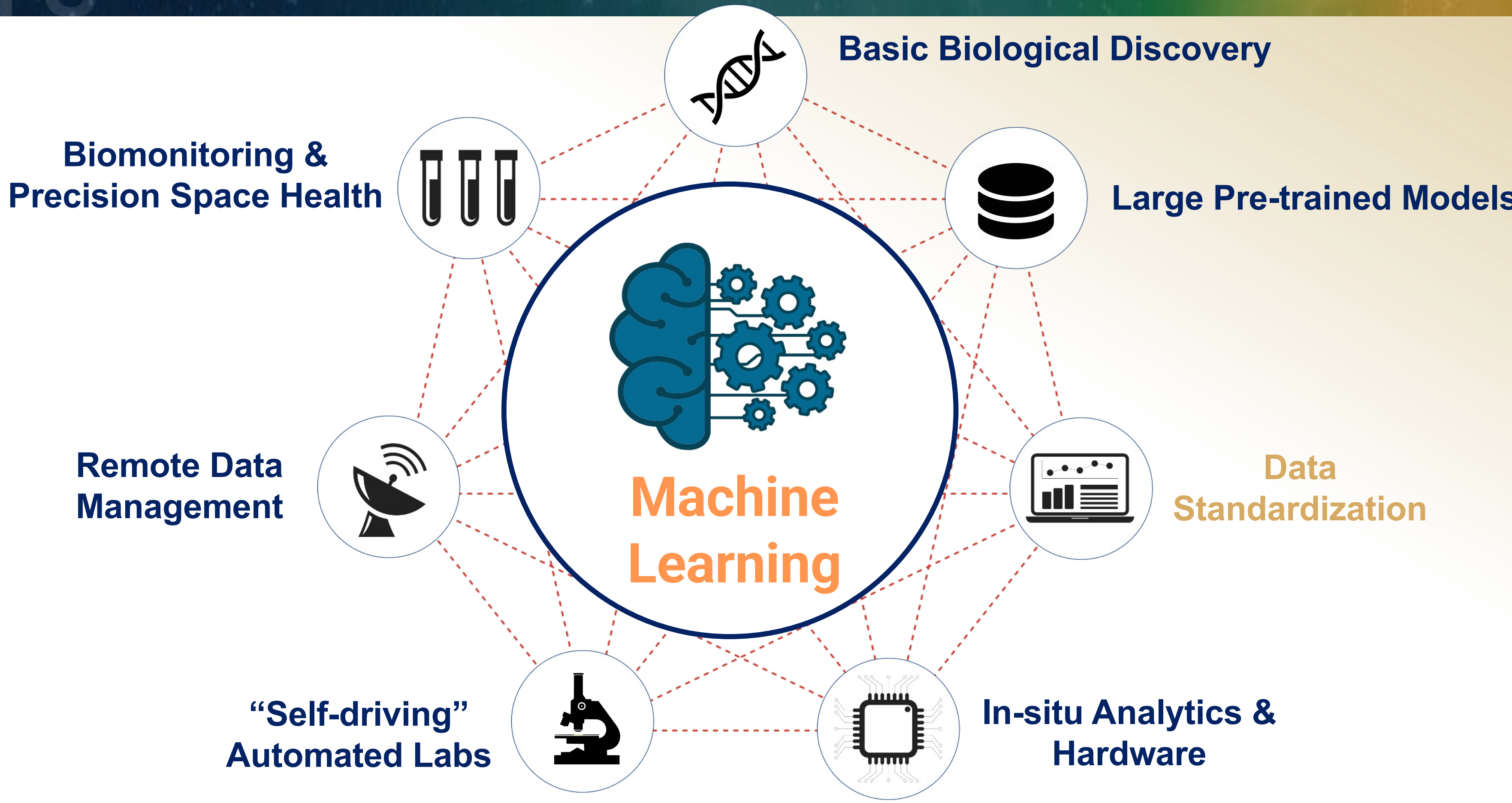
Levenshtein Distance: Measures the difference between two sequences.
LLM Matching: Leverage LLM to find semantic matches.

Customizable Match Threshold: Users can set a threshold for what constitutes a match
Application for OSDR/NBISC:
- Curators can auto-populate relevant metadata from incoming datasets against standardized template





ryan.t.scott@nasa.gov

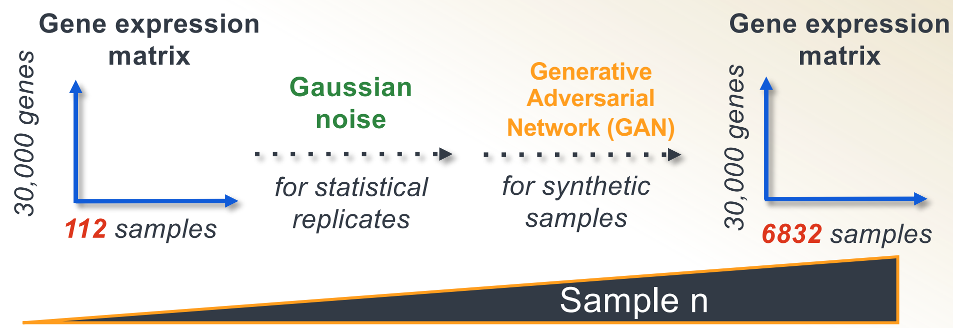


ryan.t.scott@nasa.gov

RNA Sequencing Benchmark Dataset

Scientific motivation: Effects of spaceflight on mouse liver health

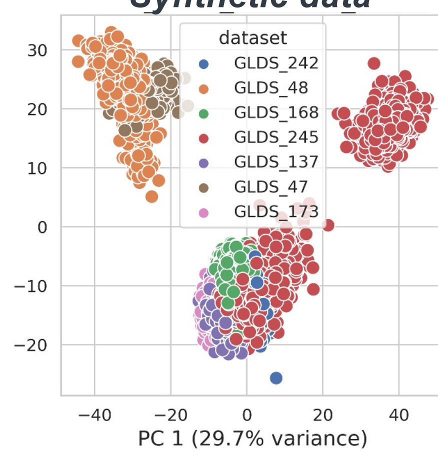
AI readiness pipeline:



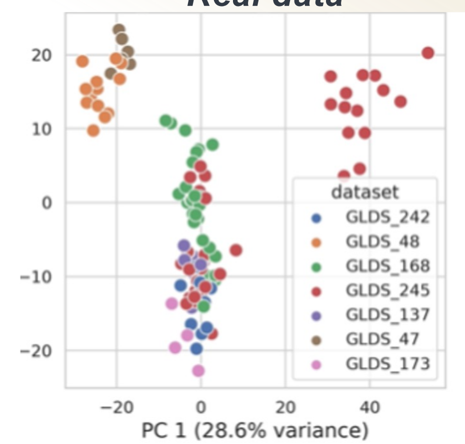
aws

GeneLab Data Set	Tissue	Spaceflight Mission
47	liver	RR1 CASIS
48	liver	RR1 NASA
168	liver	RR1 NASA RR3 CASIS
137	liver	RR3 CASIS
173	liver	STS-135
242	liver	RR9
245	liver	RR6

Synthetic data



Real data

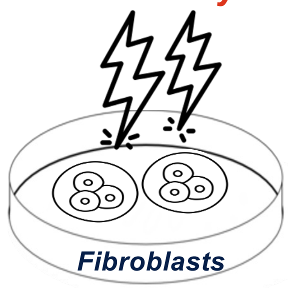


Fluorescence Microscopy Benchmark Dataset

Scientific motivation: simulated space radiation causes cellular DNA damage

15 mouse strains

Simulated galactic cosmic rays

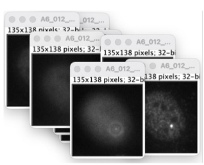


- 53PB1+ immunocytochemistry
- High-throughput imaging
- Automated quantification of 53 PB1+ radiation-induced foci

Penninckx et al. *Radiation Research* 2019

aws

AI readiness pipeline:

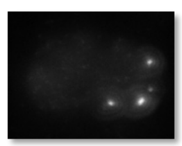


Raw Dataset (n = 94,193):
32-bit Z stacks (9 indices)

DIP lib

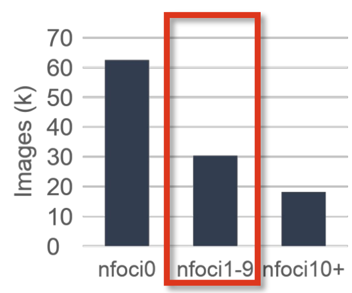
maximum intensity projection

16-bit conversion

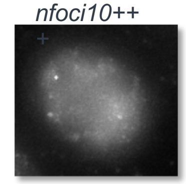


Max Intensity Dataset (n=94,193):
16-bit single-index TIFFs

automatically estimated nfoci

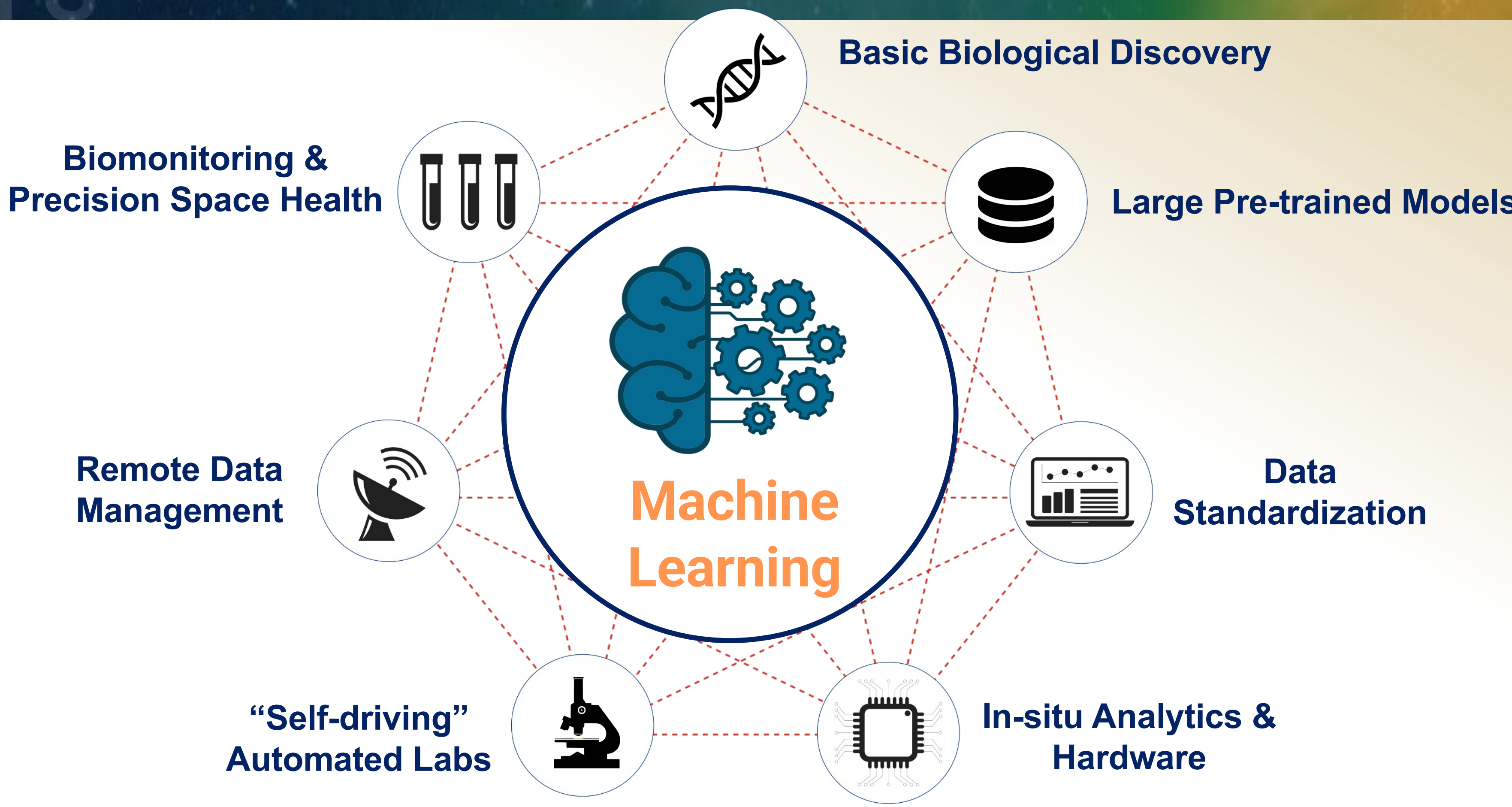


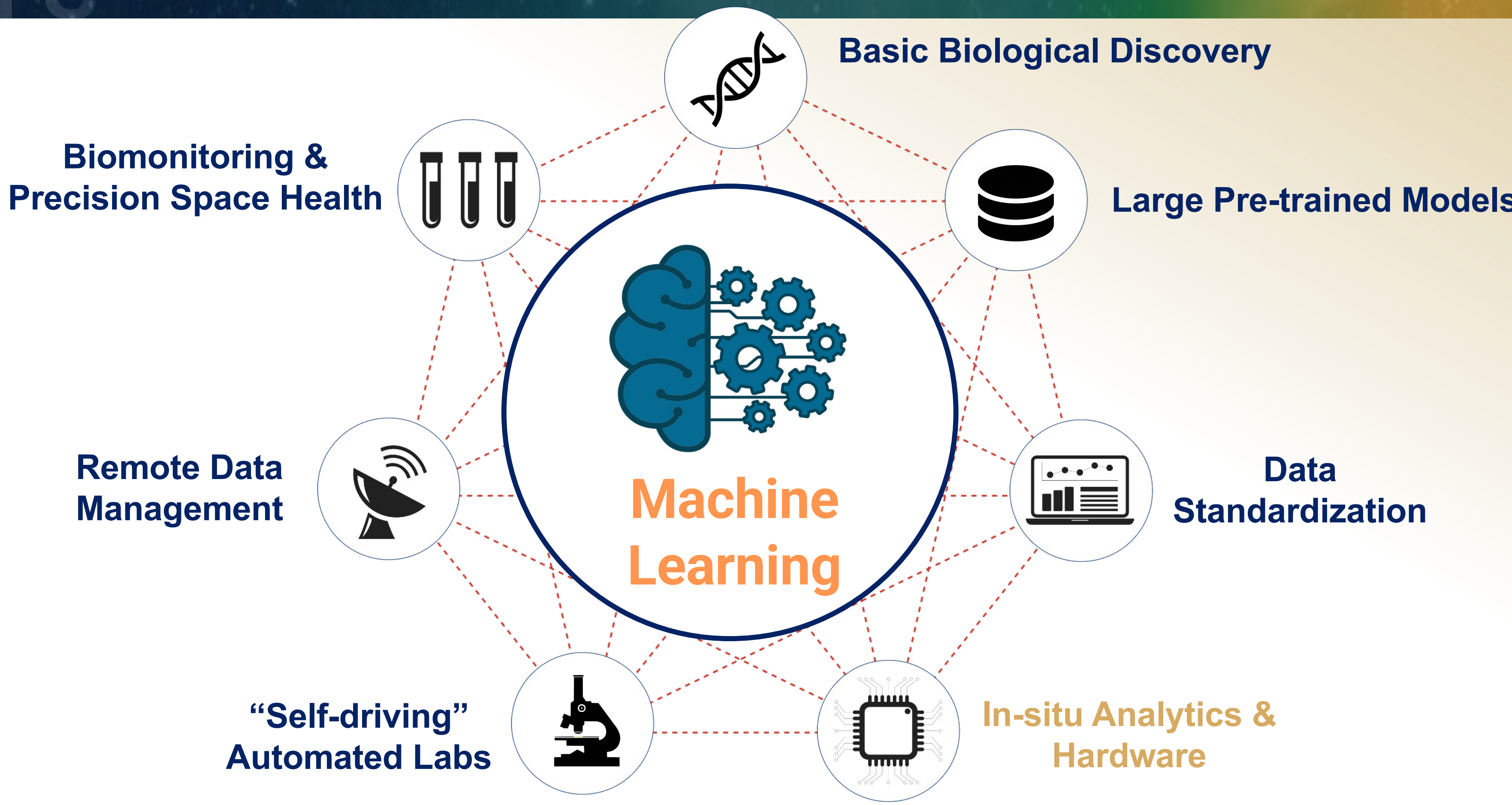
nfoci10++



Labels:

- *nfoci* (number of 53BP1+ DNA damage foci)
- *Radiation* (X-ray, ^{56}Fe…)


Basic Biological Discovery
Biomonitoring &
Precision Space Health
Large Pre-trained Models
Machine
Learning
Remote Data
Management
Data
Standardization
"Self-driving"
Automated Labs
In-situ Analytics &
Hardware



ryan.t.scott@nasa.gov

AI/ML Architecture to Support Deep Space Mission Goals

nature machine intelligence

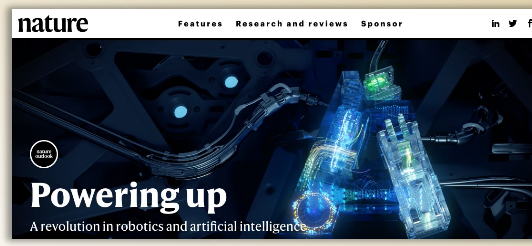
Review article

https://doi.org/10.1038/s42256-023-00617-5

Biomonitoring and precision health in deep space supported by artificial intelligence

Received: 23 December 2021

Ryan T. Scott [1,52], Lauren M. Sanders [2,52], Erik L. Antonsen [3],



nature machine intelligence

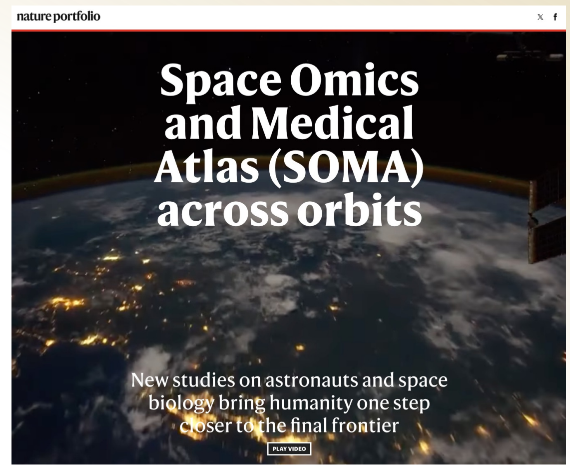
Review article

https://doi.org/10.1038/s42256-023-00618-4

Biological research and self-driving labs in deep space supported by artificial intelligence

Received: 23 December 2021

Lauren M. Sanders [1,52], Ryan T. Scott [2,52], Jason H. Yang [3], Amina Ann Qutub [4],



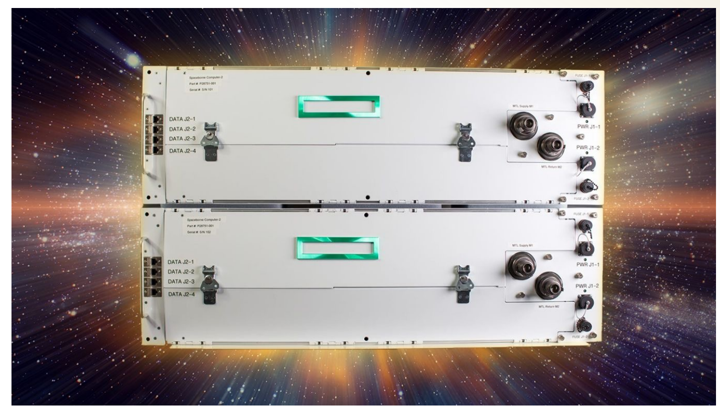
Stanford University · SJSU San José State University · Anschutz · The University of Texas at San Antonio (UTSA) · UCSF · Berkeley Lab Lawrence Berkeley National Laboratory · Mayo Clinic · Baylor College of Medicine · Icahn School of Medicine at Mount Sinai · Salk Institute for Biological Studies · Medical School University of Minnesota Driven to Discover · Johns Hopkins University · Google · Rutgers New Jersey Medical School · Harvard University · KBR · Michigan State University · NASA · Caltech · UAB The University of Alabama at Birmingham · UNF University of North Florida · Weill Cornell Medicine

Scott et al., *Nature Machine Intelligence* 2023
Sanders et al., *Nature Machine Intelligence* 2023

Hardware for ML-enabled in-situ Data Collection and Analysis

NVIDIA Jetson edge AI and robotics platform

HPE's Spaceborne Computer

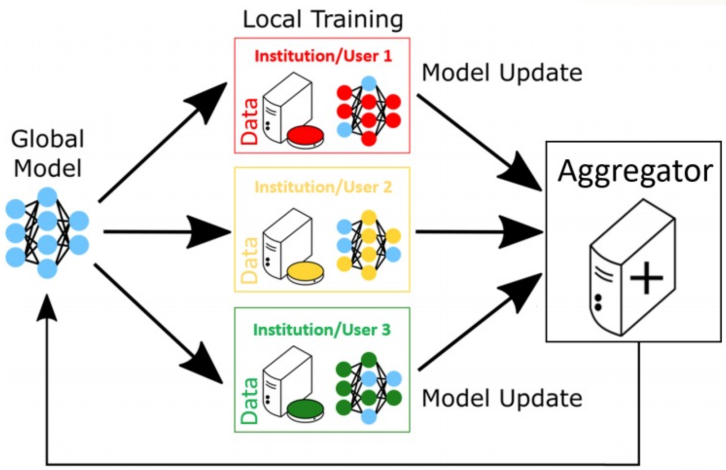


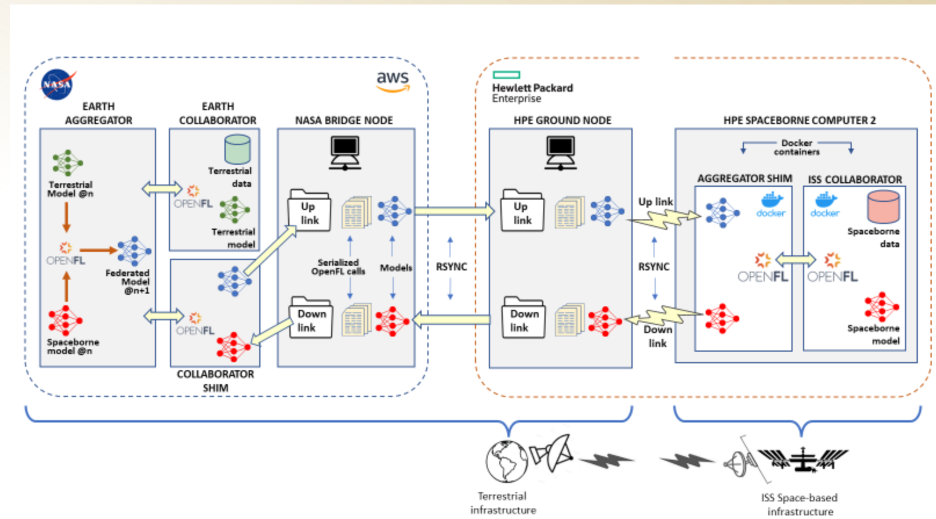
Category	Technology	Relevance to spaceflight
In situ capabilities: small footprint and resilient to environmental factors (radiation, acceleration, vibration)	Neuromorphic processors Edge computing[44]	Space-borne computing with very low power, little or no cooling, high efficacy for AI algorithms and resilience to radiation[146,147] Process and analyse data collected in deep-space missions on board for input to the PSH system

Scott et al., 2023 https://rdcu.be/c8jSO;
Sanders et al., 2023 https://rdcu.be/c8jSS

ryan.t.scott@nasa.gov

Federated Learning Using In-Space Data (FLUID)





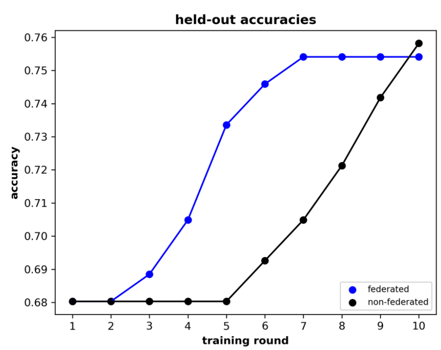
OSDR Dataset: OSD-435

Open Science for Life in Space

Pre-positioned ultrasound echocardiogram data

Used binary classifier to predict dose (HI/LO) from aorta performance metrics (e.g. heart rate, volume, velocity))

Federation survived Loss of Signal!

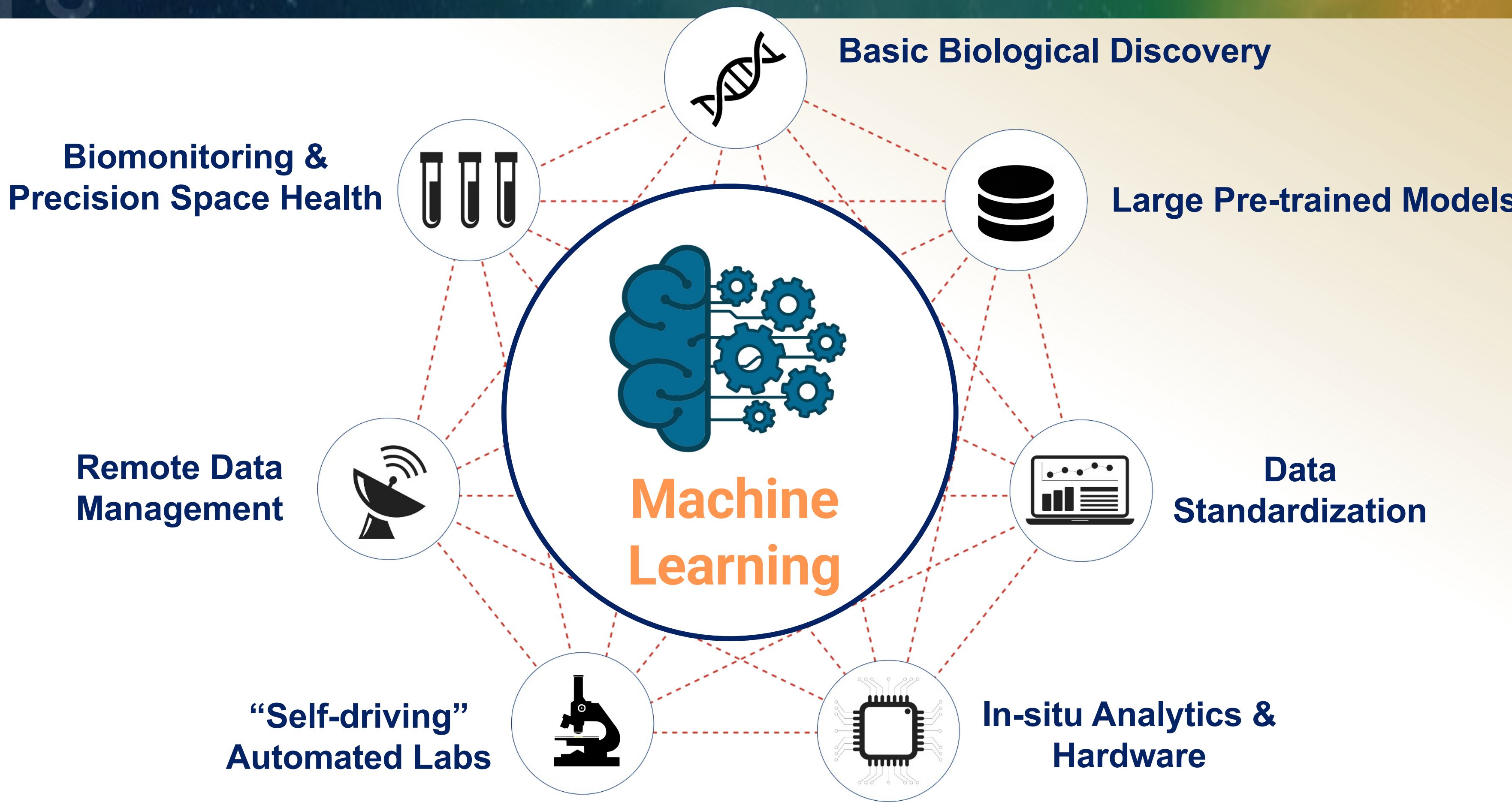


James Casaletto
BMSIS/AI for Life in Space

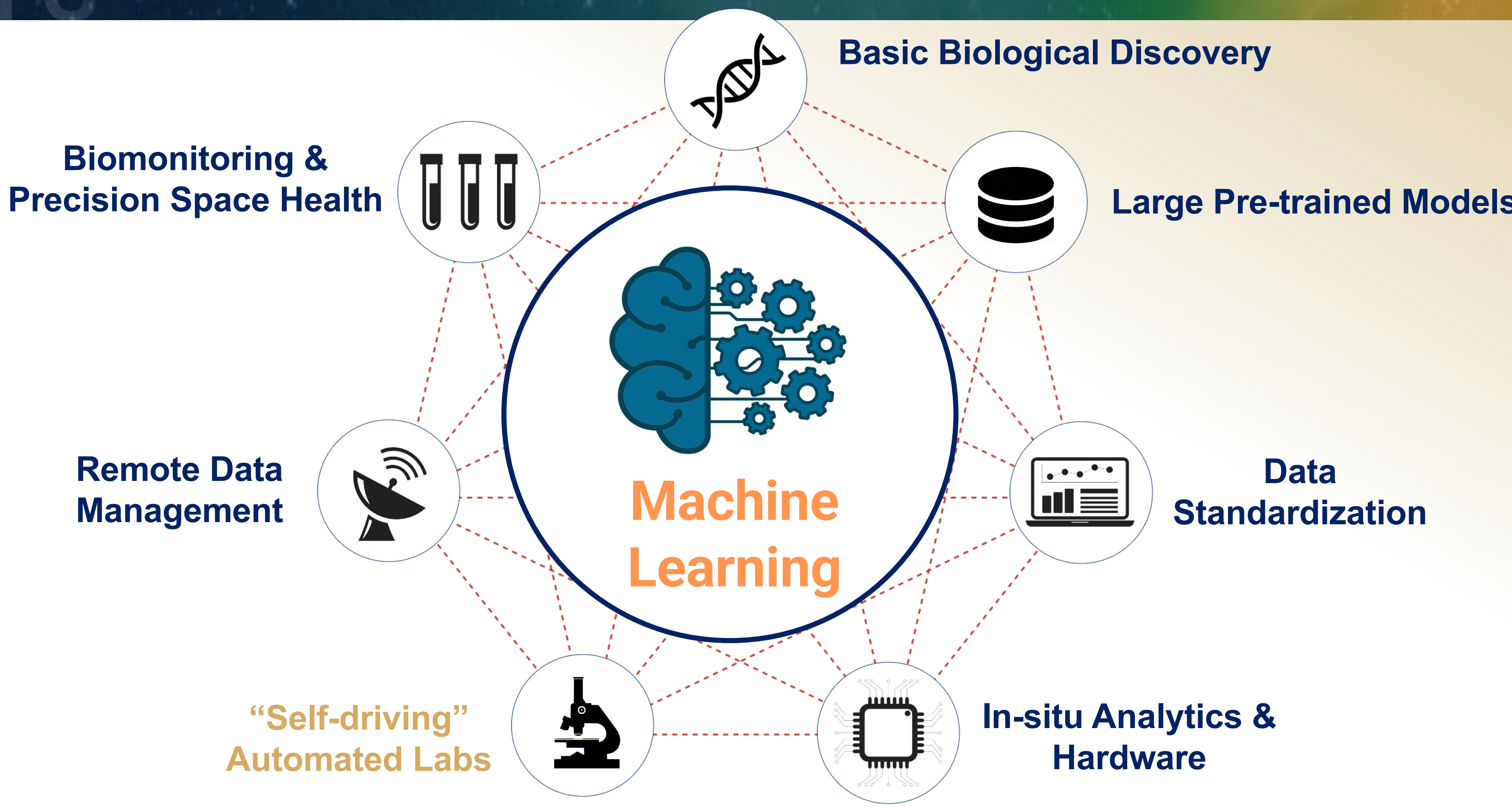
Using federated learning to build machine learning models between Earth and the International Space Station

James Casaletto, Patrick Foley, Mark Fernandez, Lauren Sanders, Ryan Scott, Shubha Ranjan, Shashi Jain, Nate Haynes, Sylvain Costes, Graham Mackintosh,

 ryan.t.scott@nasa.gov



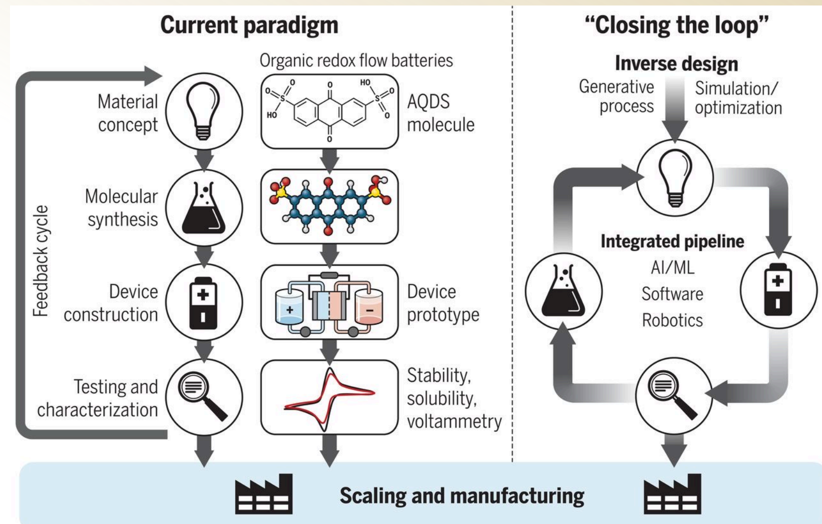
ryan.t.scott@nasa.gov



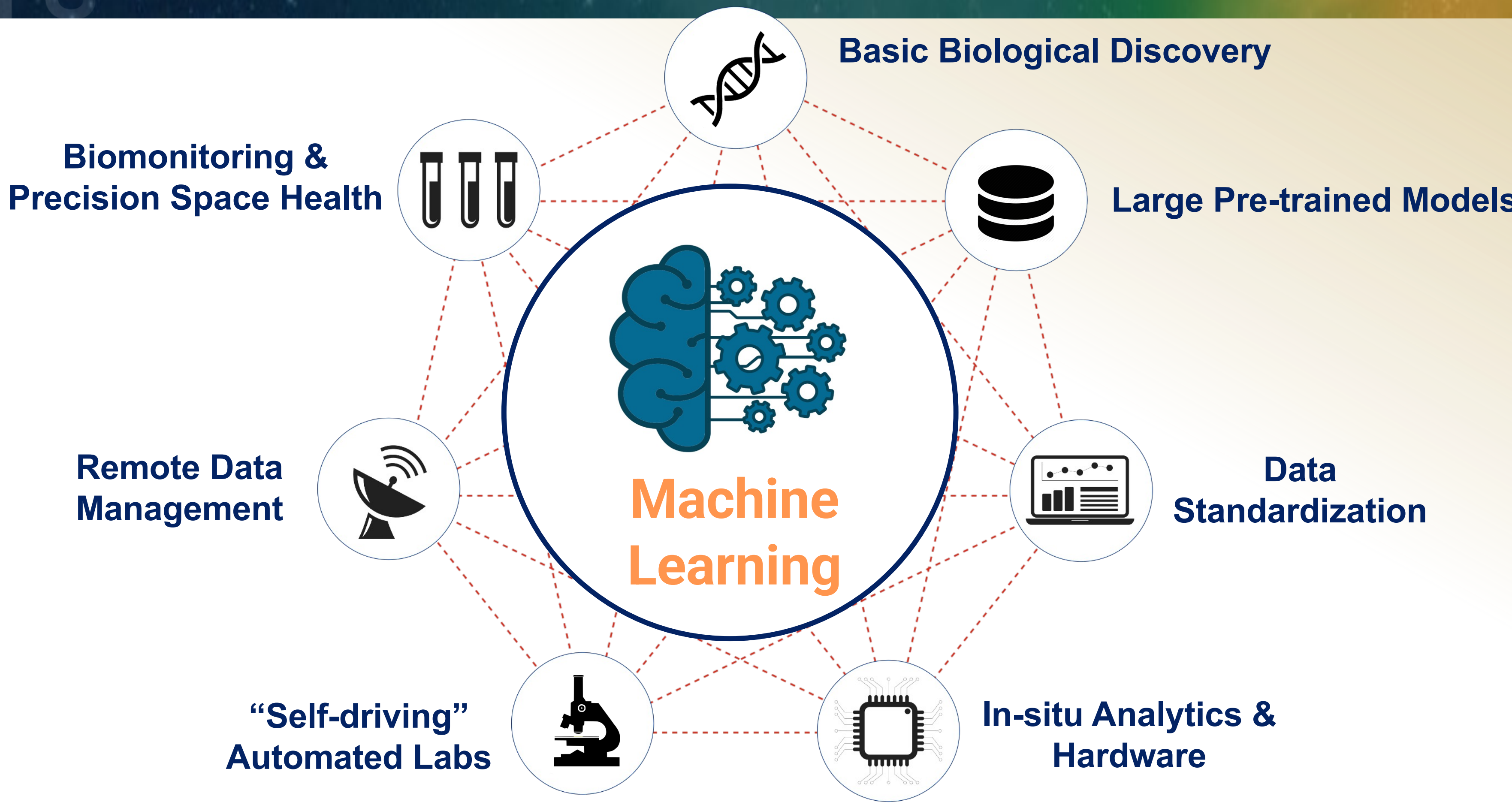
ryan.t.scott@nasa.gov

ML-enabled Automated Science & “Self-Driving” Labs

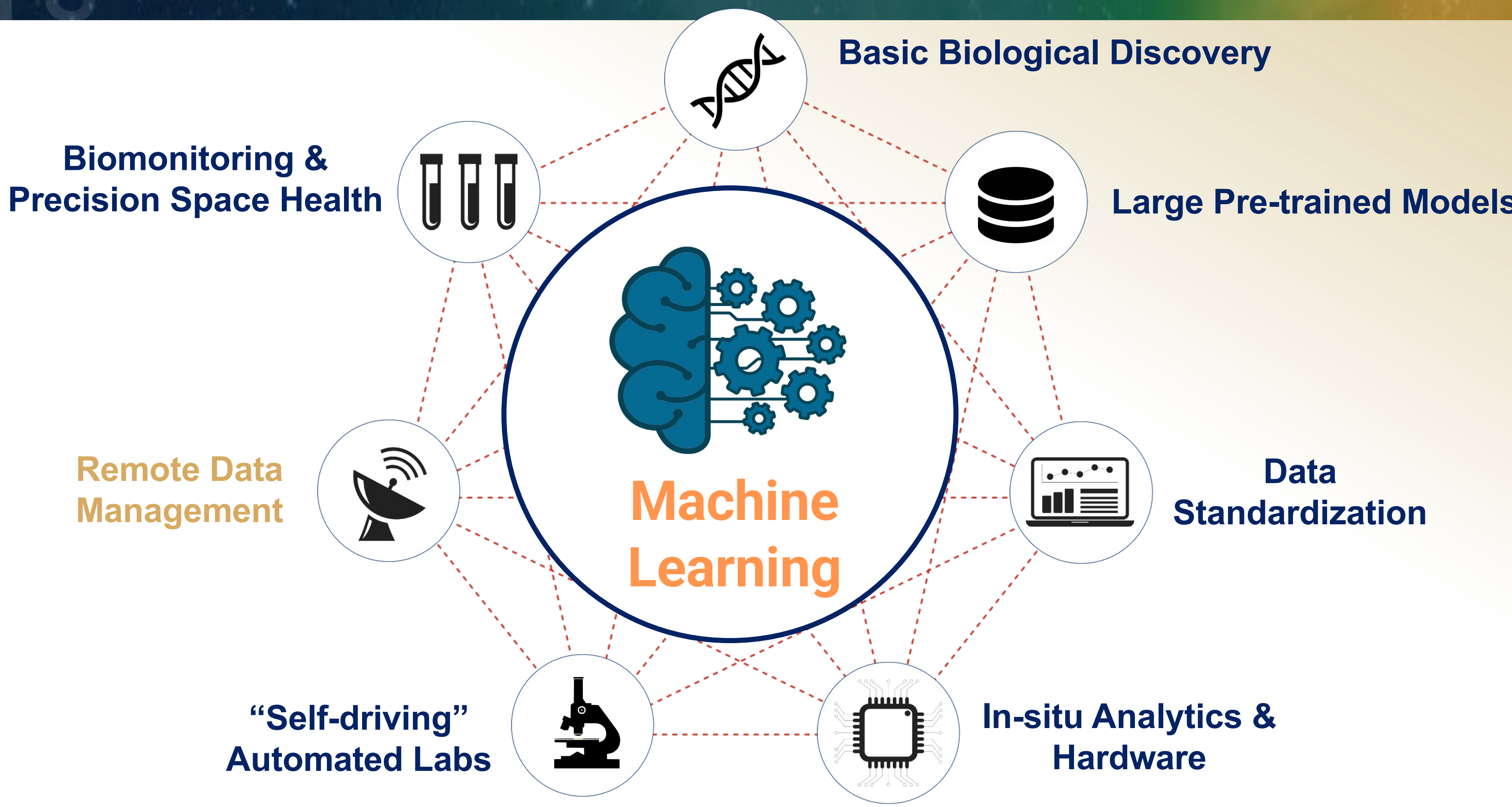
Category	Technology	Relevance to spaceflight
In situ data analysis	Active learning[107]	Train and deploy a model, which continuously monitors and retrains itself with self-assessments and regular human inspection



- “Self-driving” automated laboratory capabilities enable *in-situ* data collection
- Active learning & edge computing would allow *in-situ* data analysis



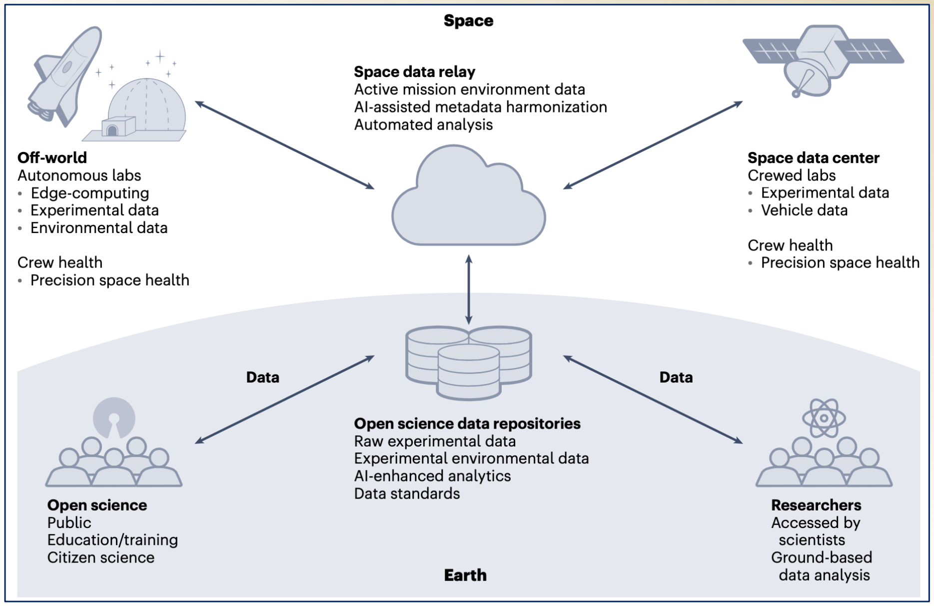
ryan.t.scott@nasa.gov



ryan.t.scott@nasa.gov

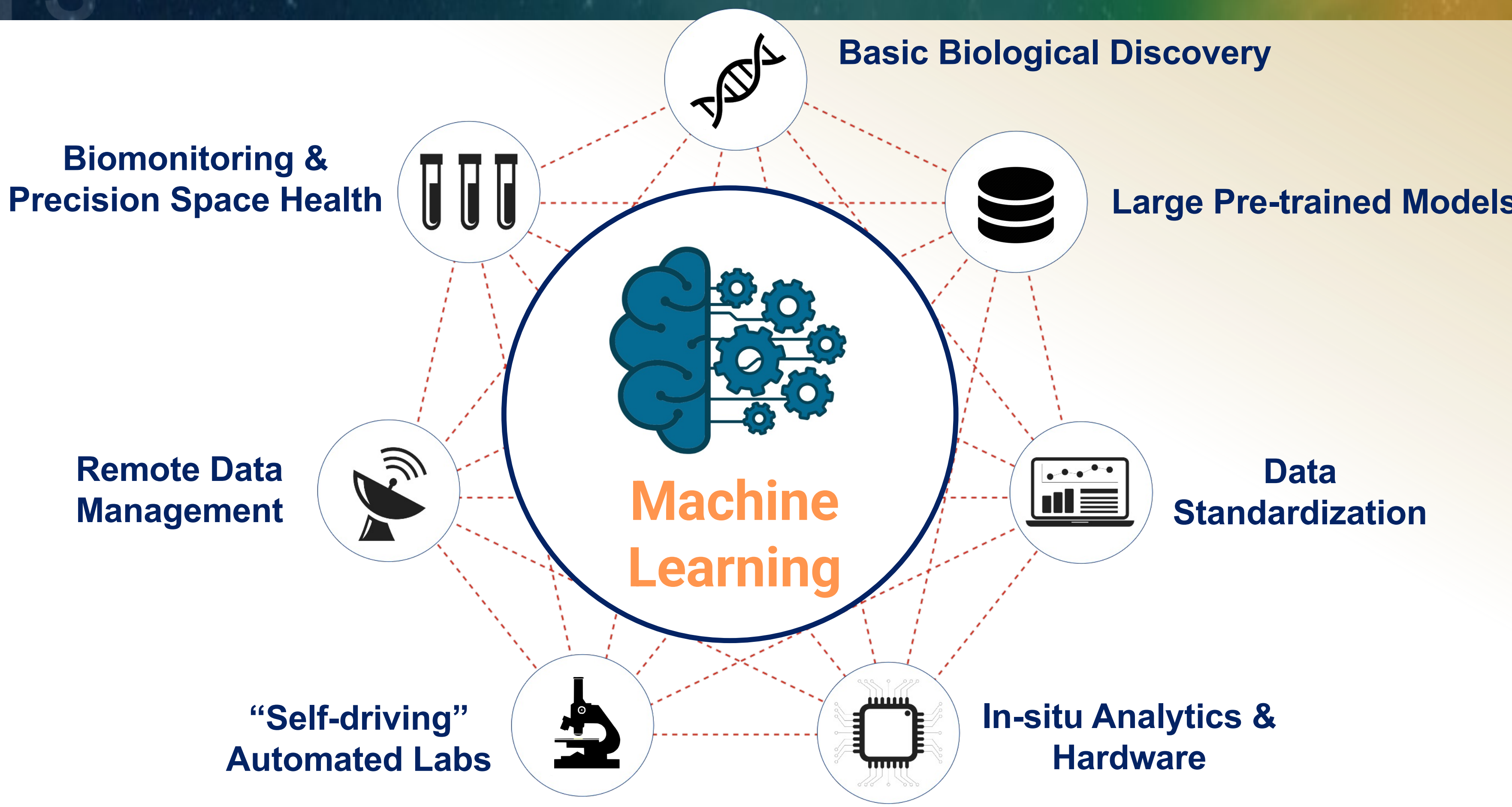
ML Approaches to Support Remote Data Management

Category	Technology	Relevance to spaceflight
Limiting data transfer to Earth	Federated learning[116]	Train a model on data collected in a deep-space mission and on Earth-based data for stronger inference
Distilling and maximizing computing needs in space	Transfer learning Dimensionality reduction[148] TinyML[149] Few-shot learning[150]	Train large models on Earth and deploy on data collected in-flight Identify key features to reduce data size Prune large neural networks to deploy on spacecraft or habitats with operational constraints Learn from few data points by leveraging contextual information

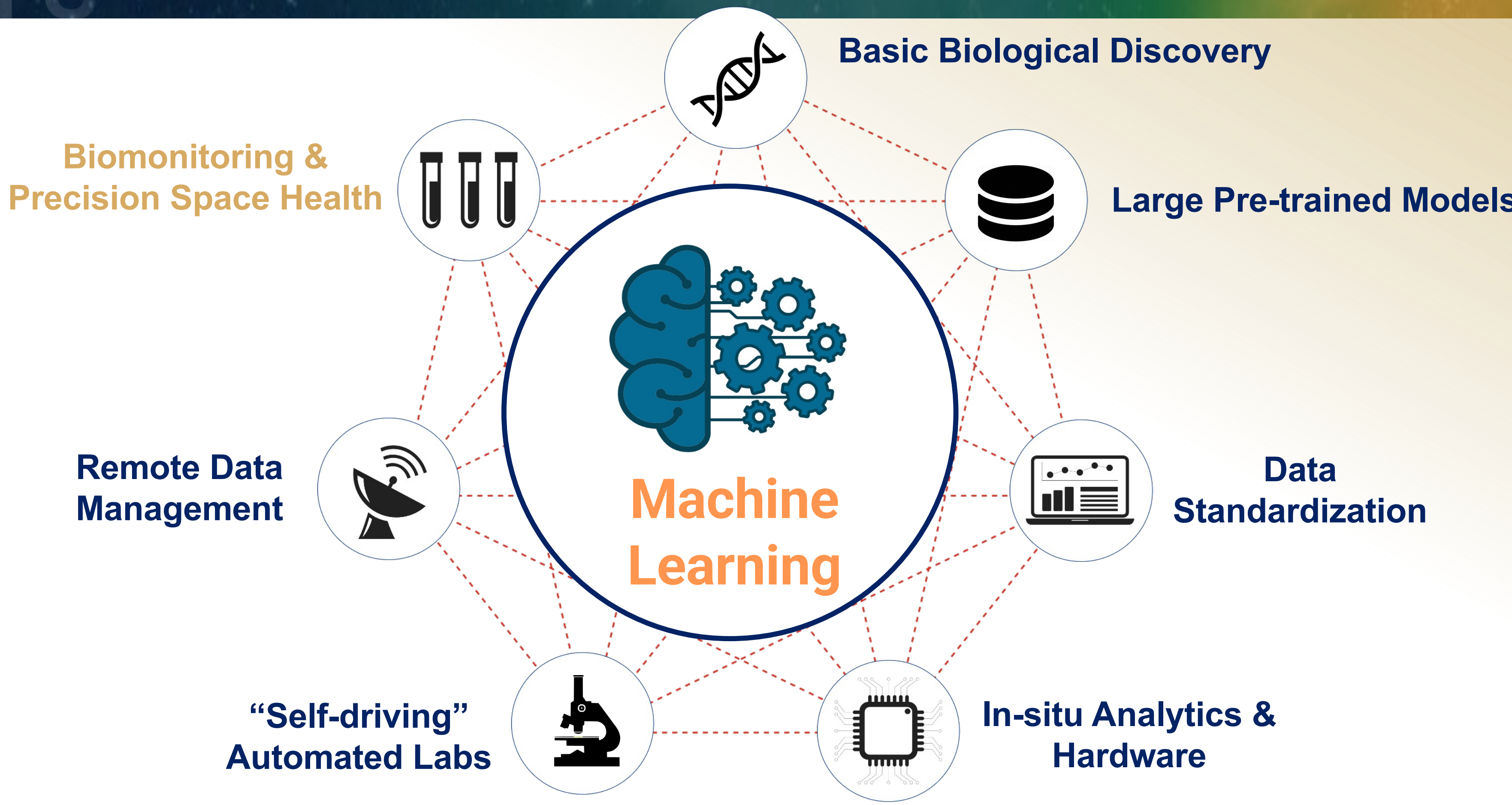


- ML methods such as federated learning, transfer learning, and few-shot learning support deep space data transfer

Scott et al., 2023 https://rdcu.be/c8jSO;
Sanders et al., 2023 https://rdcu.be/c8jSS

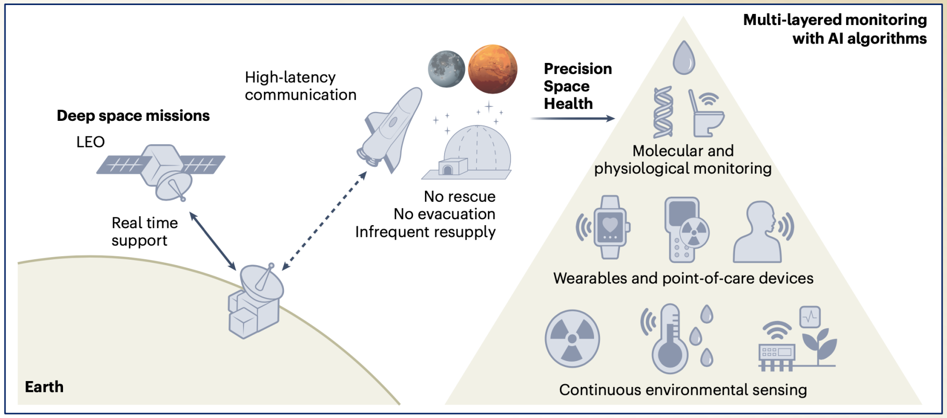


ryan.t.scott@nasa.gov


Basic Biological Discovery
Biomonitoring & Precision Space Health
Large Pre-trained Models
Machine Learning
Remote Data Management
Data Standardization
"Self-driving" Automated Labs
In-situ Analytics & Hardware

ML-Enabled Biomonitoring Approaches

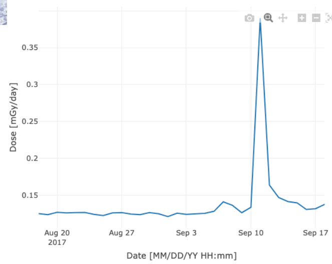
Category	Technology	Relevance to spaceflight
Methods to train on data that differ from inferencing context	Translation[152,153]	For example, train on radiation exposure data in animals and predict radiation risks for human crew members
Methods for when inferencing data are extremely different (for example, outliers) from training data	Generalization: Risk extrapolation[154,155] Domain invariant representation learning[154,155]	Prediction in a situation where an astronaut's biosensor data are outliers compared to the terrestrial clinical data used for model training
Methods for when inferencing data are persistently different from training data	Adaptation[156]	For example, adapting a model trained using terrestrial electrocardiogram data to a 'new normal' of electrocardiogram readings from astronauts whose heart physiology has changed in spaceflight



- Multi-layered monitoring of spacecraft and habitats & *in-situ* computing capabilities for real-time recommendations

Scott et al., 2023 https://rdcu.be/c8jSO;
Sanders et al., 2023 https://rdcu.be/c8jSS

OSDR's AI/ML Analysis Working Group



OSDR AWG
Forum-Space

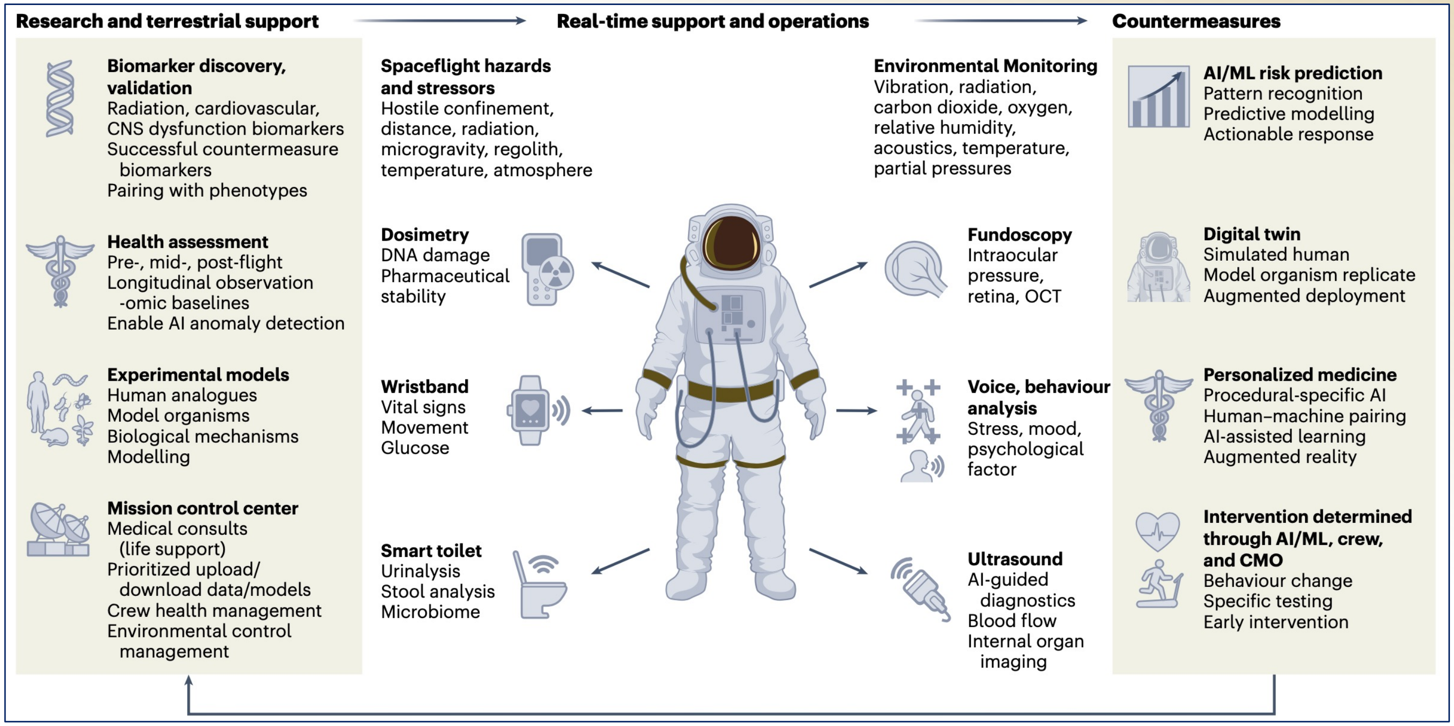
OSD#	Material Type	Omics data	Pheno-Physio
255	Retina	RNAseq (8 GC + 8 FLT)	
568	Retina		Immunostaining Microscopy
557	Eye/Retina		mCT, histology, immunostaining
583	Eye		Intraocular Pressure
397	Retina	RNAseq, methylation (17 simulated mg/rad)	
203	Retina	RNAseq, methylation (60 simulated mg/rad)	
194	Retina	RNAseq (4 BSL, 4 GC, 5 FLT)	
87	Retina	RNAseq (6 FLT)	

Project	Data	Proposed methods
Anomaly detection	Environmental data app	• Dictionary learning with VAE • Vanilla VAE
Data labeler	Microscopy OSD-255, OSD-568, OSD-557, OSD-583, OSD-397, OSD-203, OSD-194, OSD-87	• Lightly + Label Studio • Sklearn semi-supervised learning • FixMatch • MixMatch
Data generator	OSD-255, OSD-568, OSD-557, OSD-583, OSD-397, OSD-203, OSD-194, OSD-87	• VAE • GAN
Digital twin	Data available in papers	• VAE • GAN
Causal inference	benchmark RNA-seq	• RNA-seq augmentation • Invariant Risk Minimization • Augmentation as intervention • ML in biological causal inference • SPOKE

Join the AWG!

July 18, 2024
#COSPAR2024
ryan.t.scott@nasa.gov

ML-Aided Precision Space Health System



- ML to support modeling, prediction and recommendations for a Precision Space Health system for real-time decision-making in deep space

Scott et al., 2023 https://rdcu.be/c8jSO;
Sanders et al., 2023 https://rdcu.be/c8jSS

ryan.t.scott@nasa.gov

ACKNOWLEDGE

AI4LS
Artificial Intelligence for Life in Space

NASA Open Science Data Repository

GeneLab
ALSDA
NBISC
NASA Biological Institutional Scientific Collection
OPEN SCIENCE FOR LIFE IN SPACE

Stakeholders-Management

NASA Space Biology Program
NASA Biological & Physical Sciences Division
NASA Science Mission Directorate
NASA Human Research Program
NASA Postdoctoral Program

Collaborators

All 2021 AI/ML Workshop Participants
Open Science AWG Members

July 16, 2024 – #COSPAR2024
ryan.t.scott@nasa.gov

AI for Life in Space
- Walter Alvarado
- James Casaletto
- Lauren Sanders
- Sylvain Costes

Radiation Biophysics Lab
- Egle Cekanaviciute
- Connie Pasternak
- Sylvain Costes

NASA Center for Climate Simulation: Science Managed Compute Environment
- Aaron Skolnik
- Andre Avelino Paniagua
- Ellen Salmon
- Daniel Duffy

Thursday, 7/18, 5-6:30p
Poster – EC1-Hall1, ID TWT-325

NASA Open Science Data Repository for Space Biology Data Access

Ryan Scott

Friday, 7/19, 10:10-10:30a
F5.2 – Room EC1-211

Elucidating the Spectrum of Human Radiosensitivity: A Five-Year GWAS Analysis of Blood Cell Response to Simulated Space Radiation

Sylvain Costes