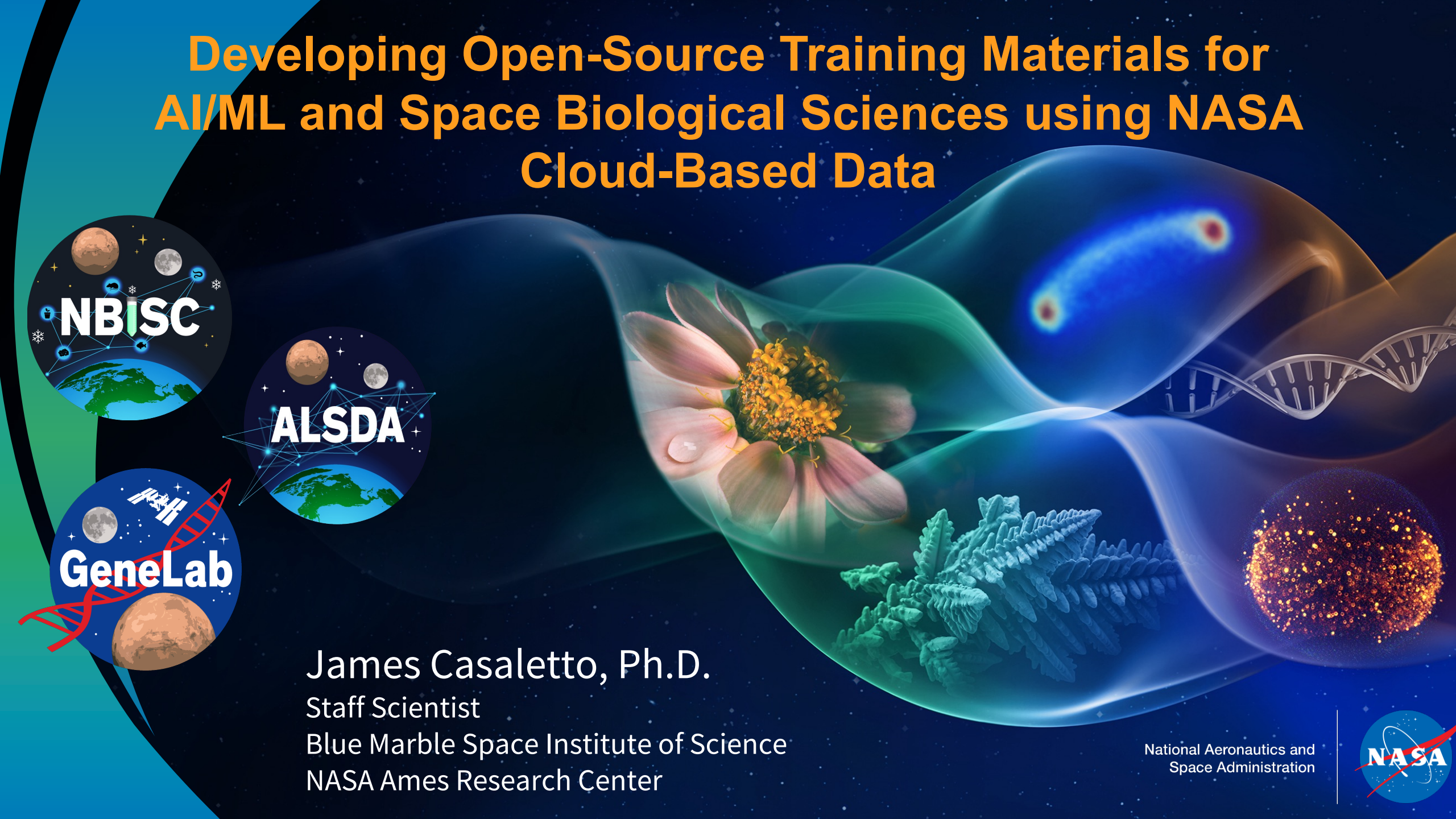
Developing Open-Source Training Materials for AI/ML and Space Biological Sciences using NASA Cloud-Based Data



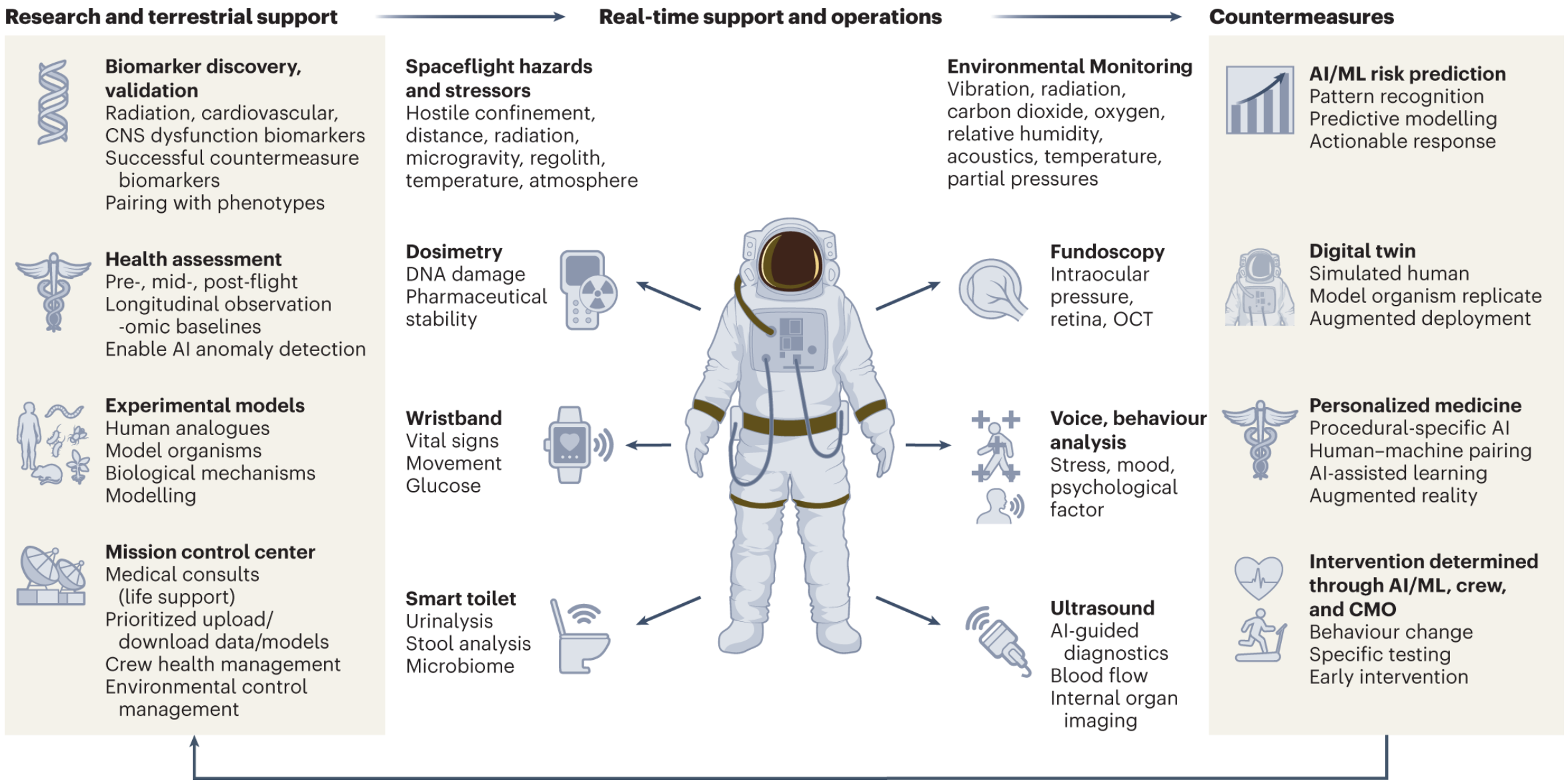
James Casaletto, Ph.D.
Staff Scientist
Blue Marble Space Institute of Science
NASA Ames Research Center

National Aeronautics and Space Administration

NASA

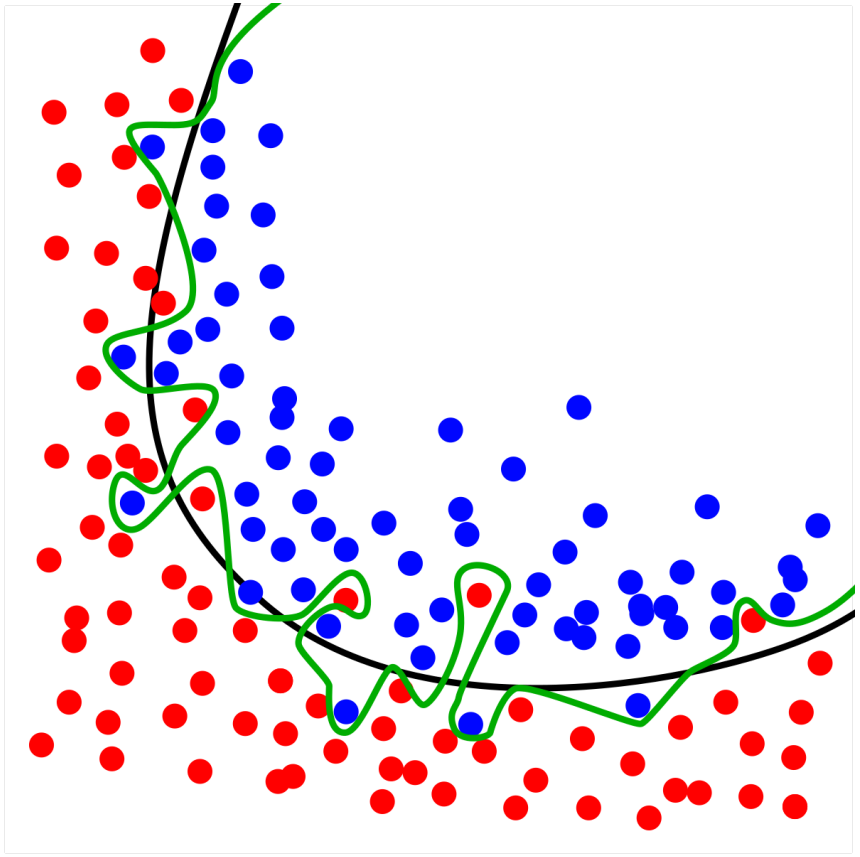
What is the motivation for developing this training?

The moonshot: use AI/ML to help keep astronauts healthy



Space biology is under-understood in part for lack of data

Molecular biology is difficult to study for sparse data

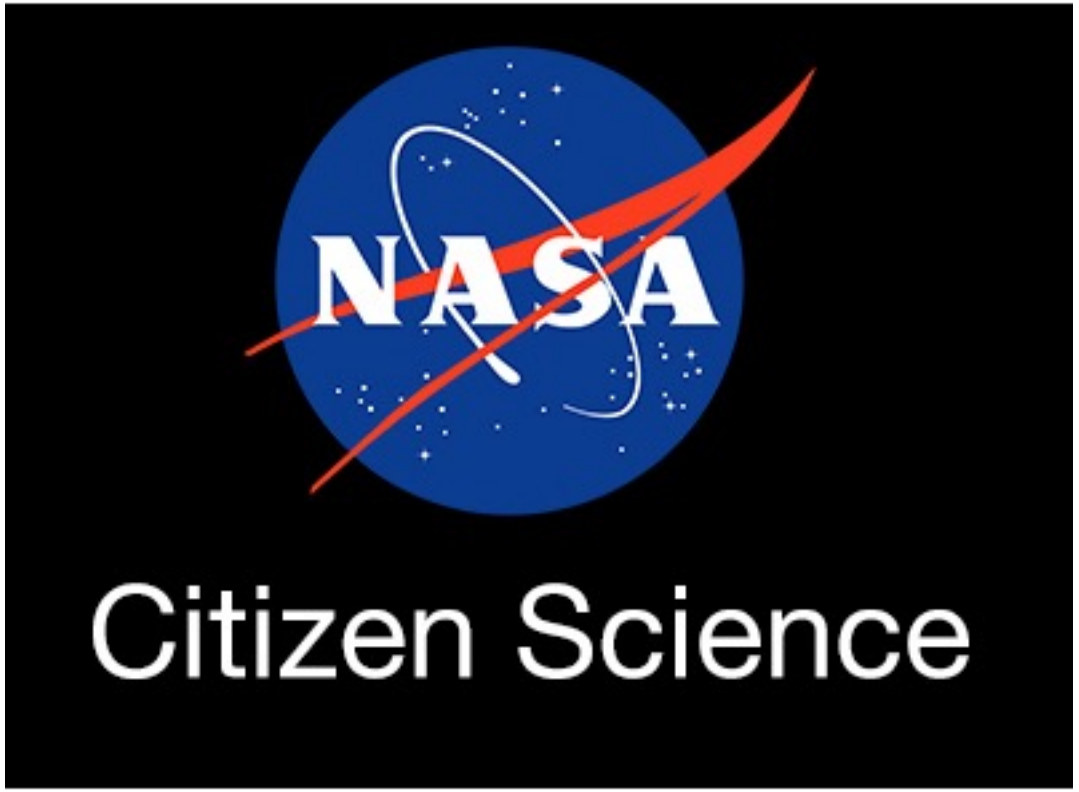


https://www.analyticsvidhya.com/blog/2022/11/explaining-sparse-datasets-with-practical-examples/

AI/ML is a promising approach to study sparse data

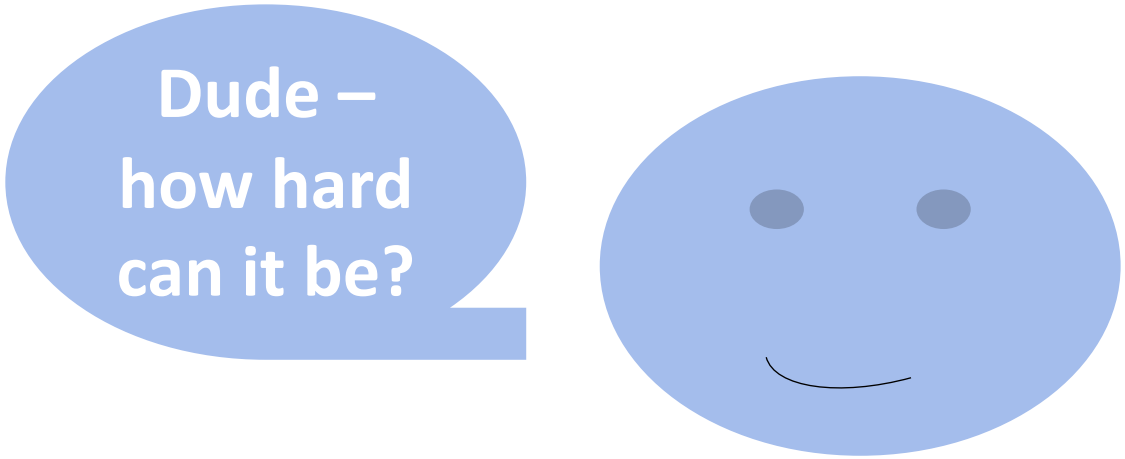
ML Model	Description
Support vector machine	SVM uses support vectors to arbitrarily partition the feature space.
Logistic regression	LR uses regularization to prevent overfitting.
K-nearest neighbors	KNN computes distances between data points.
Multi-layer perceptron	MLP uses gradient descent for optimization.
Decision tree & random forest	DT and RF can work well when coupled with gradient boosting.

Calling all citizen scientists!

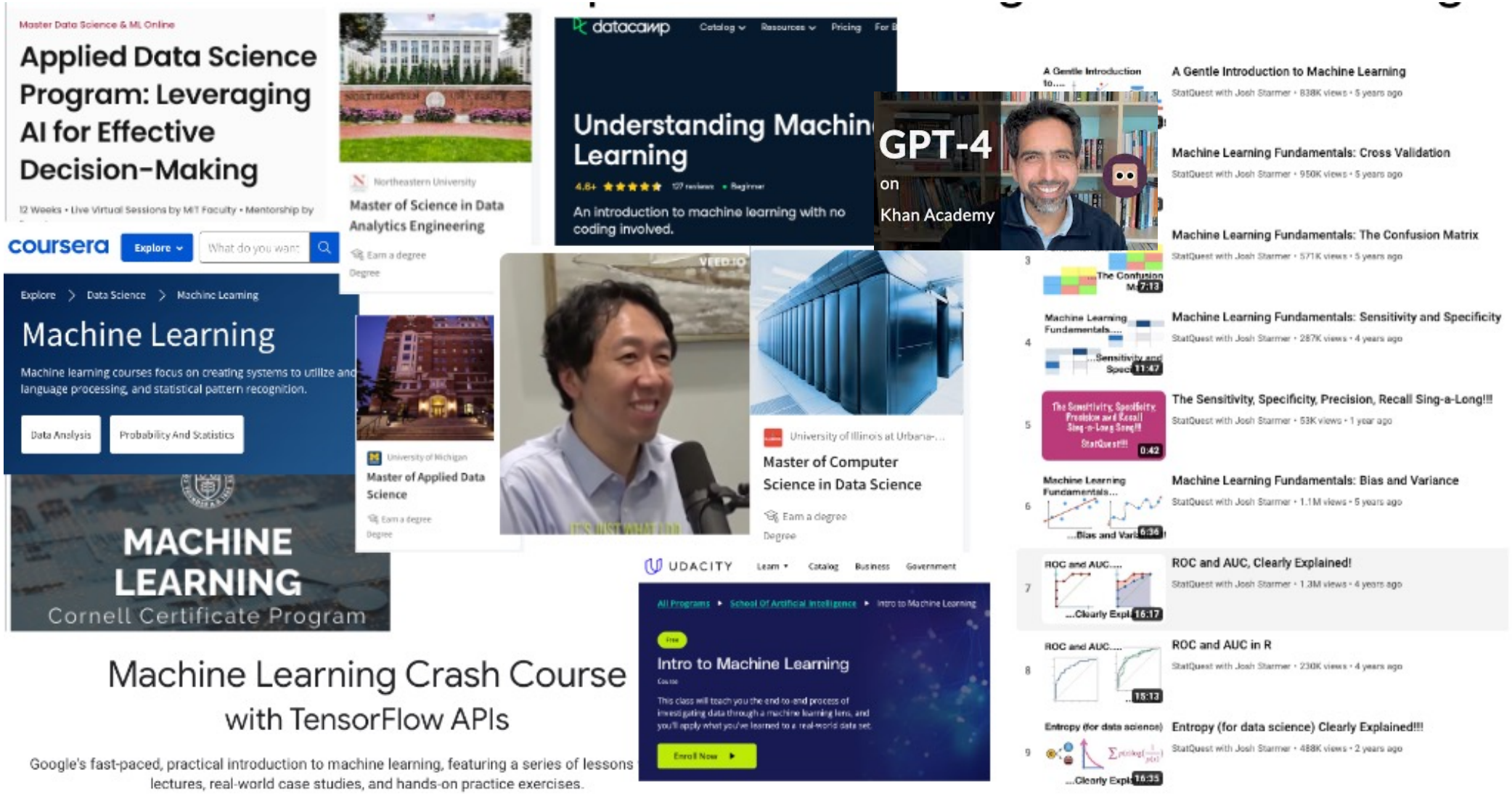


https://www.google.com/url?sa=i&url=https%3A%2F%2Fnesec.strategies.org%2Fstrategic-partnerships%2F&psig=AOvVaw1zxEgOVobVvE71FKZkYIZX&ust=1702452386943000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCOCU18iviYMDFQAAAAAdAAAAABAD

AI/ML, while ubiquitous, is poorly understood by many



There are LOTS of options to learn machine learning



There are NO options to learn AI/ML for space biology

https://github.com/nasa/Transform-to-Open-Science/

What are the choices for developing this training?

Choice 1: How to design the curriculum?

Options	Pros	cons
Design in a vacuum	• simple • control deliverables	• biased • lacks user perspective
Design by committee	• fairly simple • less biased	• lack of expertise • conflict resolution

Choice 2: How to develop the curriculum?

Options	Pros	cons
Build from scratch	• control over product • consistent look/feel	• time and effort • wheel re-invention
Curate collection	• less time/effort • leverage what's good	• inconsistent look/feel • no ML-for-space-bio

Choice 3: How to deploy the curriculum?

Options	Pros	cons
Instructor-led	• high touch/interaction • dynamic	• doesn’t scale • small reach
Self-paced	• scales • global reach	• limited interaction • distractions/motivation

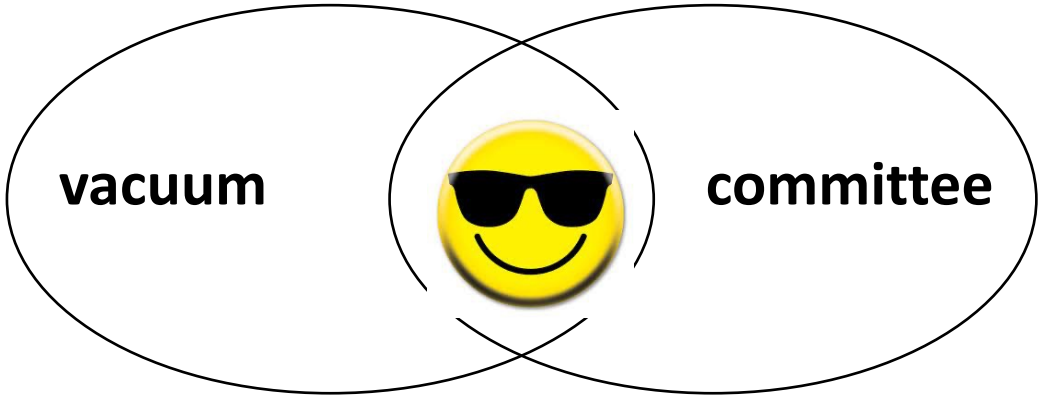
Choice 4: How to deploy the labs?

Options	Pros	cons
Student computer	• zero cost • reproducible results	• may exclude students • inconsistent experience
Cloud IaaS	• scales well • consistent experience	• cost • internet requirement

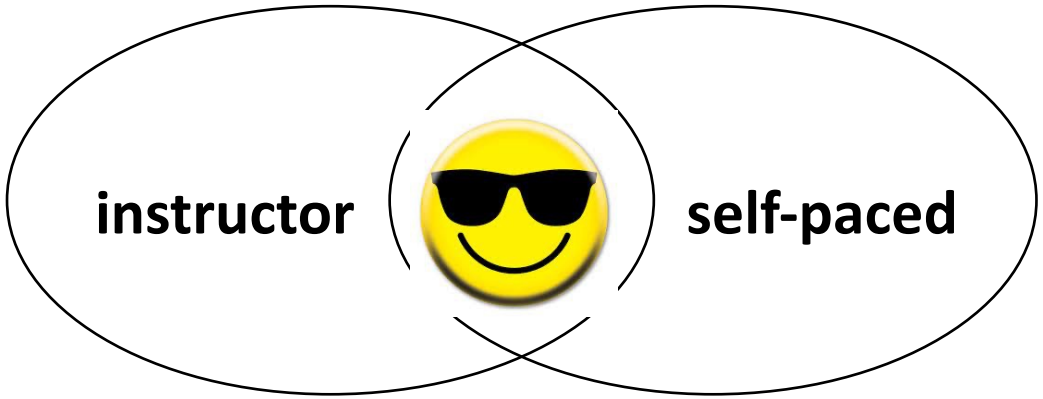
So what’s the plan, at least to start?

Our decision: meet in the middle

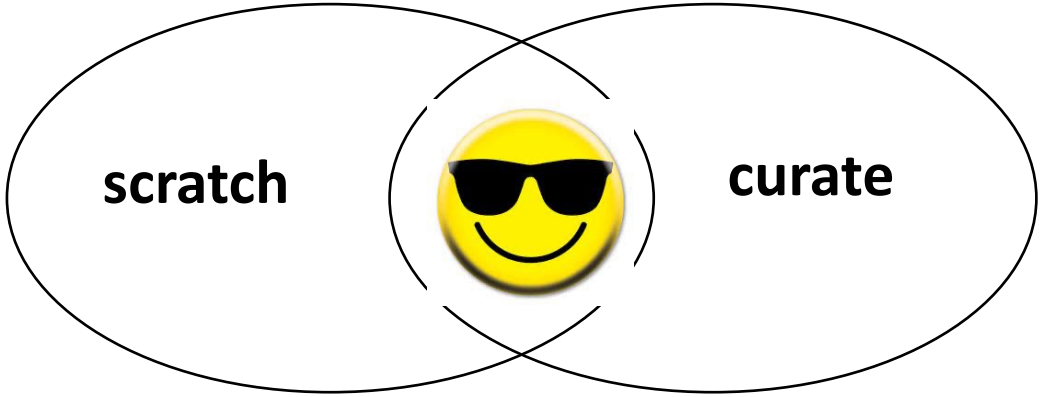
curriculum design



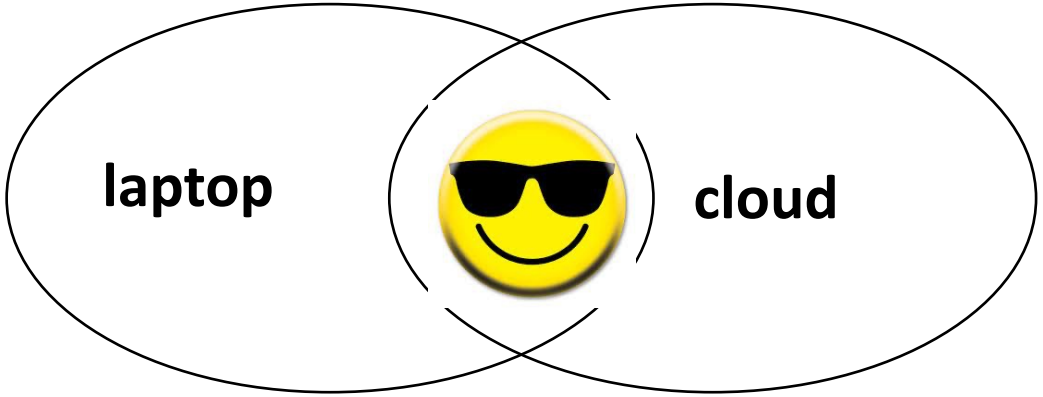
curriculum deployment



curriculum development



lab deployment



Design: start in vacuum, pivot with users

I. intro to course
A. intro to space biology
B. intro to data engineering
C. intro to AI/ML
D. intro to lab environment

II. building basic models
A. using scikit-learn
B. clustering
C. classification
D. regression

III. working with data
A. OSDR & FAIR
B. processing tabular data
C. processing image data
D. visualizing data

IV. interpreting results
A. bioinformatic tools
B. literature search
C. ethical considerations
D. reproducibility

V. advanced topics
A. transfer learning
B. causal inference
C. explainable AI
D. biological interpretations

VI. capstone projects
A. use deep learning to find DNA damage in irradiated immune cells (microscopy)
B. use causal inference machine learning ensemble to find genes correlated to lipid dysfunction (RNA-seq)

Development: build space bio, borrow AI/ML (CC)

coursera

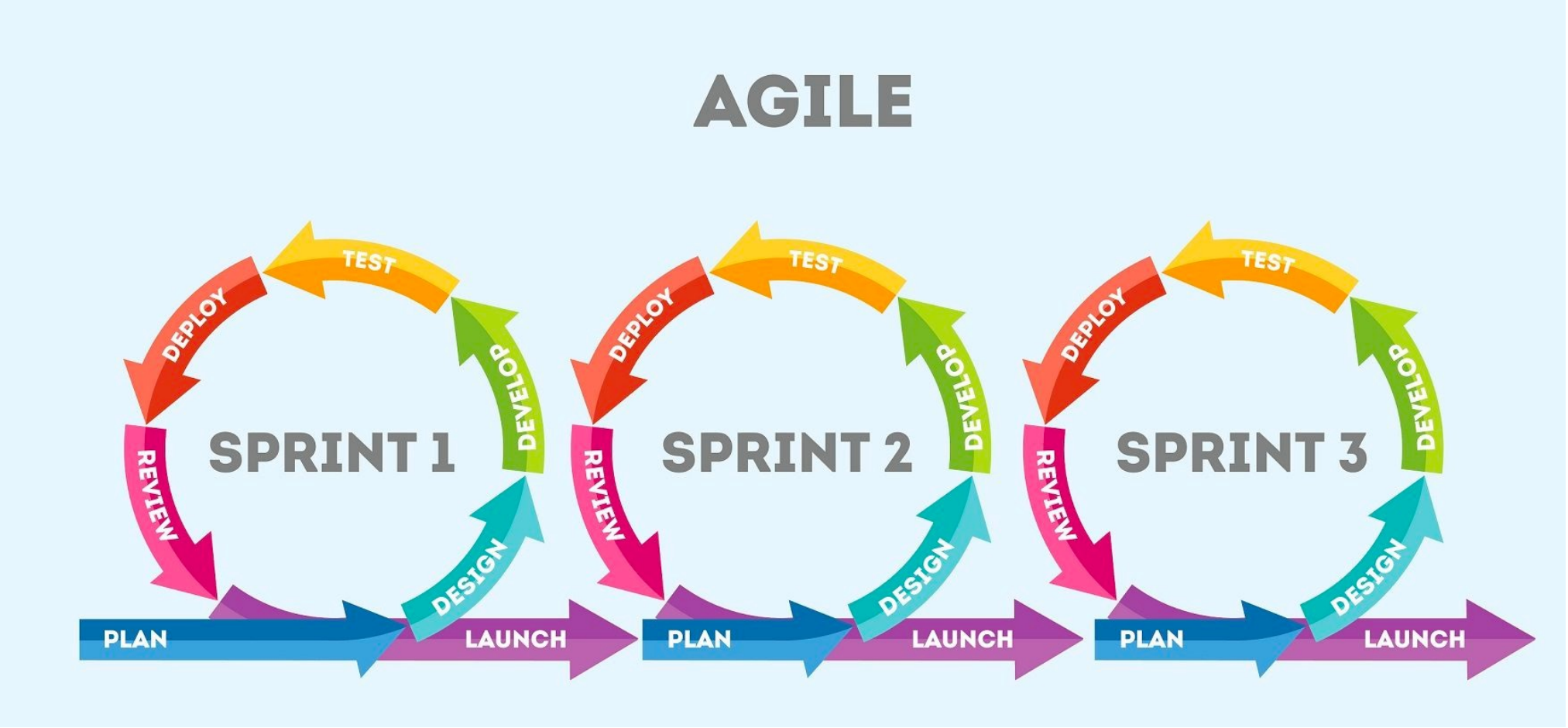
Open Learning

StatQuest!!!

towards data science

Curriculum and lab deployment: Python notebooks

We are borrowing from Agile to manage the project



Which data will we use?

RNA-seq benchmarking data

https://registry.opendata.aws/bps_rnaseq

Microscopy benchmarking data

https://registry.opendata.aws/bps_microscopy

Extra credit: use AI/ML to design, develop, deliver

1. Syllabus design and course structure
2. Course overview, module introductions, and summaries
3. Discussion forums with chatbots
4. Code walkthroughs and commenting
5. Integrate context in case studies
6. Resource recommendations
7. Quiz/assessment generation
8. Design and generate feedback surveys

Join our user community!

https://bit.ly/tops-aiml-users

Acknowledgements

AI for Life in Space
- Lauren Sanders
- Sylvain Costes

AI4LS
Artificial Intelligence for Life in Space

TRANSFORM TO OPEN SCIENCE
NASA

Open Science for Life in Space Teams

GeneLab
ALSDA
NBISC

Compute

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

NASA

Support

- NASA Space Biology Program
- NASA Science Mission Directorate
- NASA Human Research Program
- NASA Biological and Physical Sciences