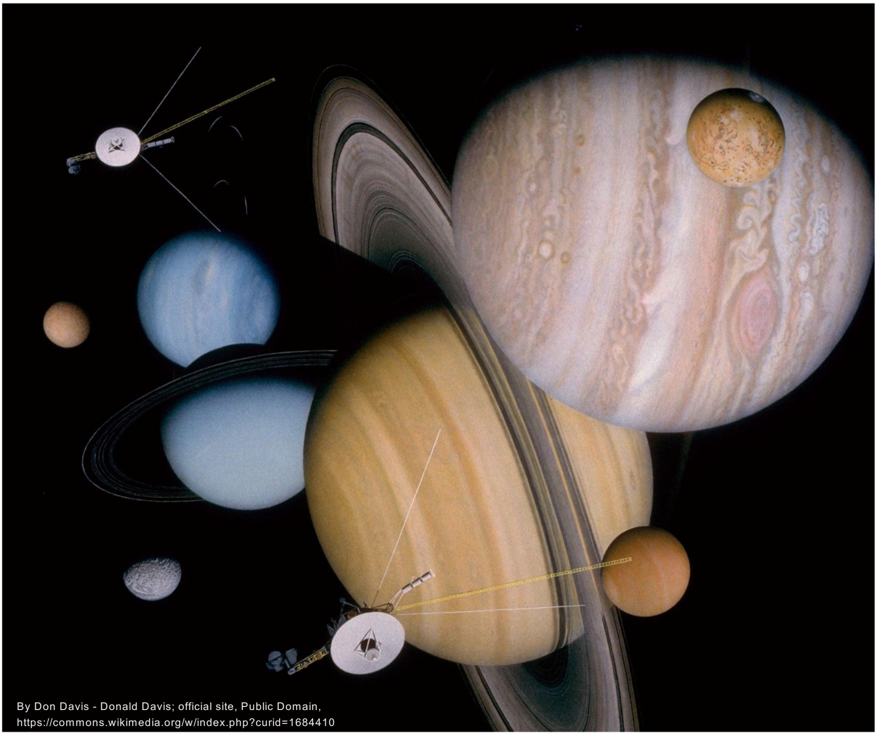
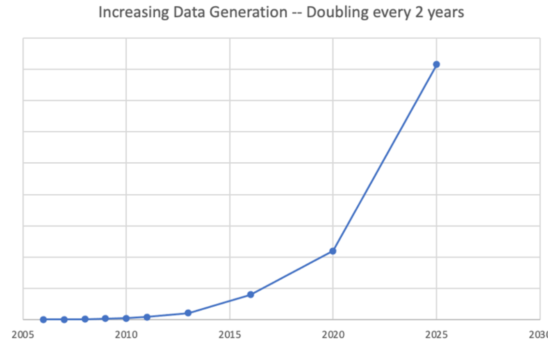
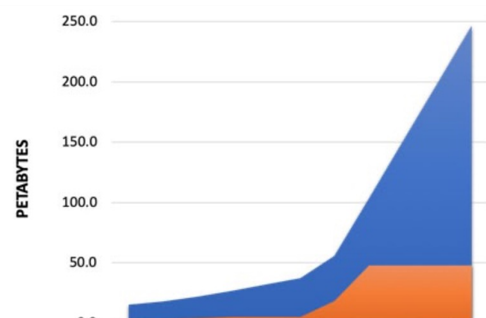
# Are you prepared to leave 99% of your data on another planet?



By Don Davis - Donald Davis; official site, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1684410



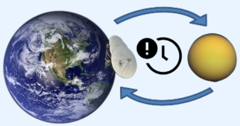




| | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Archive Size (PB) | 15.0 | 17.7 | 21.6 | 26.8 | 32.0 | 37.2 | 55.6 | 103.4 | 151.1 | 198.9 | 246.6 |
| Annual Growth (PB) | 2.6 | 2.8 | 3.9 | 5.2 | 5.2 | 5.2 | 18.4 | 47.7 | 47.7 | 47.7 | 47.7 |

| **Mass Spec Type** | **Mission** | **Launch** | **Samples/Sec** |
|---|---|---|---|
| Quadrupole | MSL | 2012 | 50 |
| Ion Trap | ExoMars | 2028 | 50,000 |
| Orbitrap | Future | TBD | 5,000,000 |

Victoria Da Poian Victoria.dapoian@nasa.gov , Eric Lyness eric.i.lyness@nasa.gov

# Science Autonomy Concept

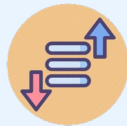


**Communication limitations**

Remote destinations and extreme environments involve longer communication delays and smaller data downlink capacities, while also limiting ground-in-the-loop interactions

**Detection challenges**

Scientists will not be able to guide spacecrafts' instrumentation in detection opportunistic features of interest



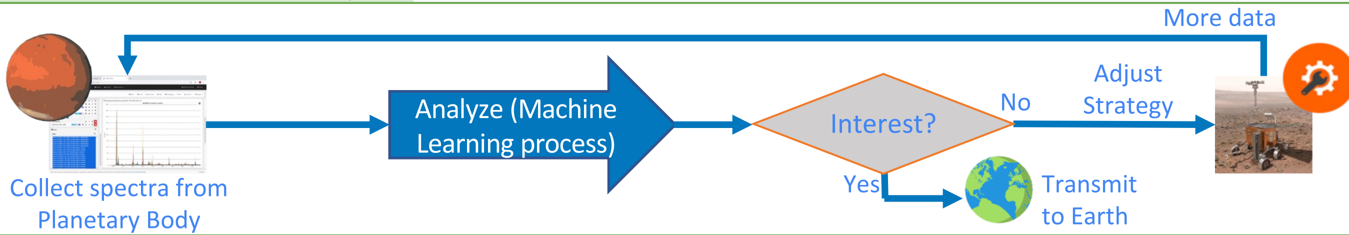
**Data Prioritization**

Future instruments will certainly generate more data: data prioritization is vital to optimize mission science return

The ability of a science instrument to **analyze its own data**:

- to **calibrate** itself
- **optimize ops parameters** based on real-time findings
- **make mission-level decisions** based on scientific observations
- determine which **data** products to **prioritize** and send back first

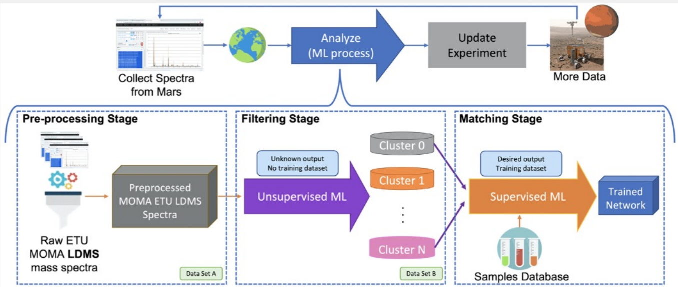
## Long Term Goal: In Situ Analysis



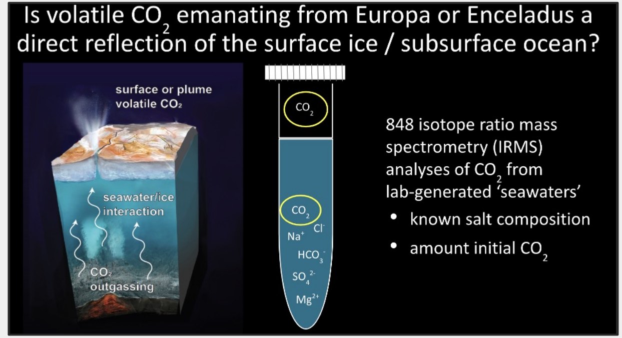

Victoria Da Poian Victoria.dapoian@nasa.gov , Eric Lyness eric.i.lyness@nasa.gov

# Some ML Projects Within NASA Planetary Environments Lab

## MOMA ML for Decision-Making (MOMA Science team, E. Lyness, V. Da Poian)

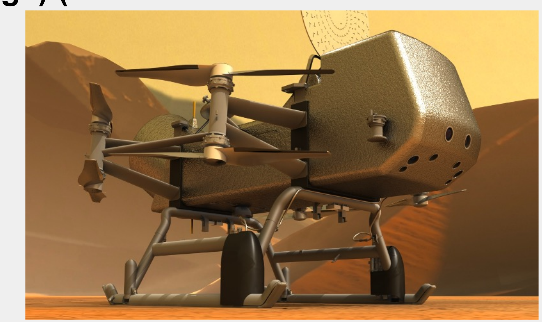



## FLaRe Ocean World Analogs (B. Theiling)




## Innovative Approach (Transfer Learning) on SAM data (V. Da Poian, E. Lyness)


## Dragonfly Automation Ideas (Brainstorm stage) (DraMS Science + Software teams)



Victoria Da Poian Victoria.dapoian@nasa.gov , Eric Lyness eric.i.lyness@nasa.gov

# Machine Learning Introduction

## Types of Machine Learning Algorithms

### Supervised Learning

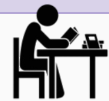
**Data**: every example has features AND labels
→ image labeled “cats” vs “not cats”

**Model**: trained to input features and output labels
→ model makes decision
→ probability view, model learns: $p(Y \mid X)$

**Learning with a teacher**: explicit feedback in the form of labeled examples
→ goal: make predictions
→ + : good performance
→ - : labeled data is difficult to find

**Examples**: Regression, Classification (sort documents by topic), Ranking

### Unsupervised Learning



**Data**: none of the example has labels
→ unlabeled images

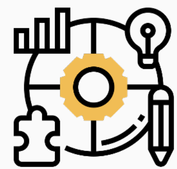
**Model**: trained to input features and reveals its unobserved structure
→ model describes the data
→ probability view, model learns: $p(X)$

**Learning by oneself**: only observed unlabeled examples
→ goal: uncover structure in data
→ + : easy to find a lot of data
→ - : finding patterns of interest

**Examples**: Clustering, Dimensionality reduction (or Manifold learning)

## 3 Primary Components of ML

Data (“experience”)



Method, model, hypothesis

Computational approach combining these 2

Victoria Da Poian Victoria.dapoian@nasa.gov , Eric Lyness eric.i.lyness@nasa.gov

*there is also semi-supervised, reinforcement learning

# ML for Data Constrained Planetary Mission Instruments

Due to the lack of flight-like instruments data, we investigate the use of commercial instruments to train ML algorithms and then tune them on flight-like data. This ML open science challenge (organized with DrivenData) is a proof-of-concept using SAM data onboard Curiosity.

## Idea of Open Science ML Challenge and Setup


**Expectations:** Innovative methods (e.g., transfer learning, data augmentation) to help analyze and interpret the measurements of planetary instruments (limited datasets constraining ML applications)

**Goal:** detect the presence of certain families (rocks, minerals, ionic compounds to help understanding Mars' potential for past habitability) of chemical compounds in geological material samples using evolved gas analysis (EGA) mass spectrometry data collected for Mars exploration missions.

**Features:** each sample is represented by EGA measurements with: *time* (in s), *temp* (in C), *m/z* (of measured ion), *abundance* (rate of ions detected/sec)

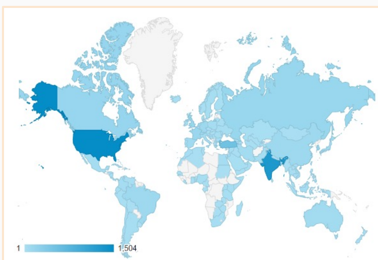
**Labels:** multilabel binary labels for the training set. 10 label classes (Basalt, Carbonate, Chloride, Iron Oxide, Oxychlorine, Silicate, Sulfide, Oxalate, Sulfate, Phyllosilicate) each indicating presence of material in the sample belonging to the respective rock, mineral, ionic compound families

## Implementation w. DrivenData

713 participants from 73 countries

656 submissions (446 submissions performed better than benchmark)

3 winners solutions



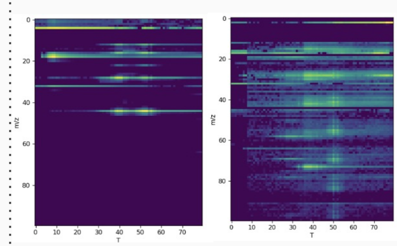

5,870 unique visitors to webpage from 130 countries

DRIVENDATA

## Results: 1st place winner


Software engineer, interest in ML (won several challenges)

- Represented the mass spectrum as a 2D image (temperature vs m/z values) to use as inputs to ML models
- Data augmentation during training




Commercial instrument

SAM testbed

Victoria Da Poian Victoria.dapoian@nasa.gov , Eric Lyness eric.i.lyness@nasa.gov