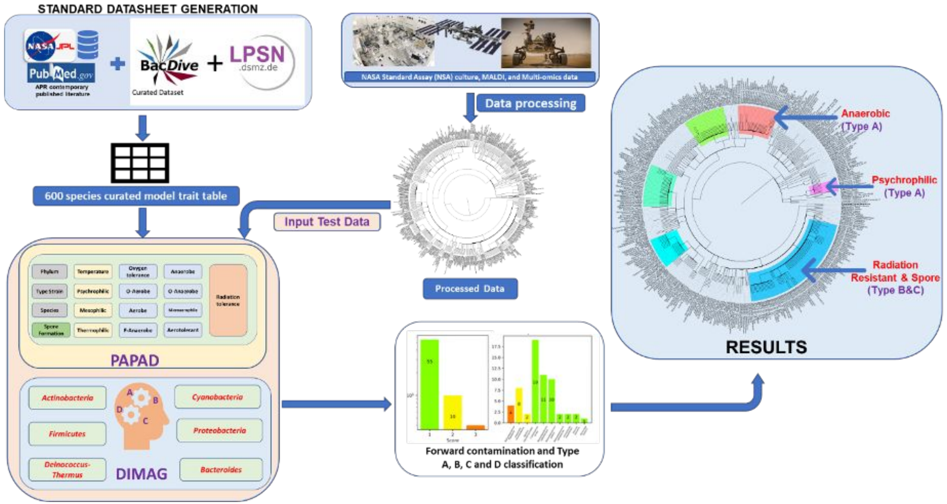
Planetary Protection: Identifying Microbes with potential for Contamination using Data Science



Ashish Mahabal
Center for Data Driven Discovery, Caltech

PSIDA, 21 Jun 2022, ESAC, Spain

With Nishka Arora (Caltech), Moogega Cooper (JPL), Nitin Singh (JPL)

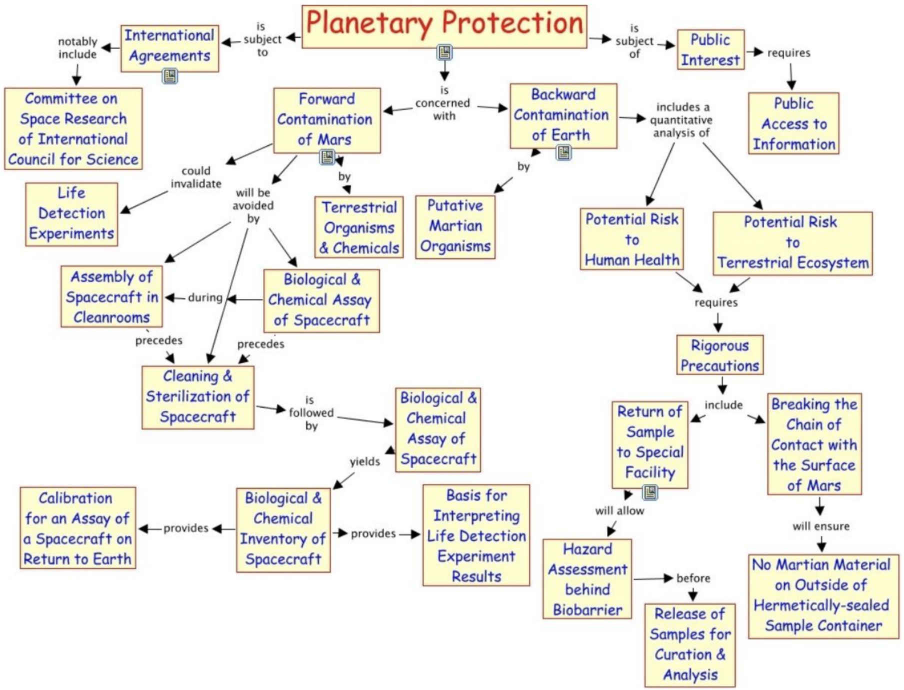
Outline

- Need for Contamination Check
- Types of Contaminators
- Creating curated datasets of known contaminators
- CheckContamination Package
- Extending to more organisms using Data Science
- Next steps

Need for identifying contaminants (and taking action)

Prevent:

- Interplanetary contamination
- Forward contamination (NASA PP Prime Directive)
- Back contamination


Planetary Protection
is subject to
International Agreements
notably include
Committee on Space Research of International Council for Science
is subject of
Public Interest
requires
Public Access to Information
is concerned with
Forward Contamination of Mars
Backward Contamination of Earth
could invalidate
Life Detection Experiments
will be avoided by
by
Terrestrial Organisms & Chemicals
by
Putative Martian Organisms
includes a quantitative analysis of
Potential Risk to Human Health
Potential Risk to Terrestrial Ecosystem
requires
Rigorous Precautions
include
Return of Sample to Special Facility
Breaking the Chain of Contact with the Surface of Mars
will allow
Hazard Assessment behind Biobarrier
before
Release of Samples for Curation & Analysis
will ensure
No Martian Material on Outside of Hermetically-sealed Sample Container
Assembly of Spacecraft in Cleanrooms
during
Biological & Chemical Assay of Spacecraft
precedes
precedes
Cleaning & Sterilization of Spacecraft
is followed by
Biological & Chemical Assay of Spacecraft
yields
Biological & Chemical Inventory of Spacecraft
provides
Calibration for an Assay of a Spacecraft on Return to Earth
provides
Basis for Interpreting Life Detection Experiment Results

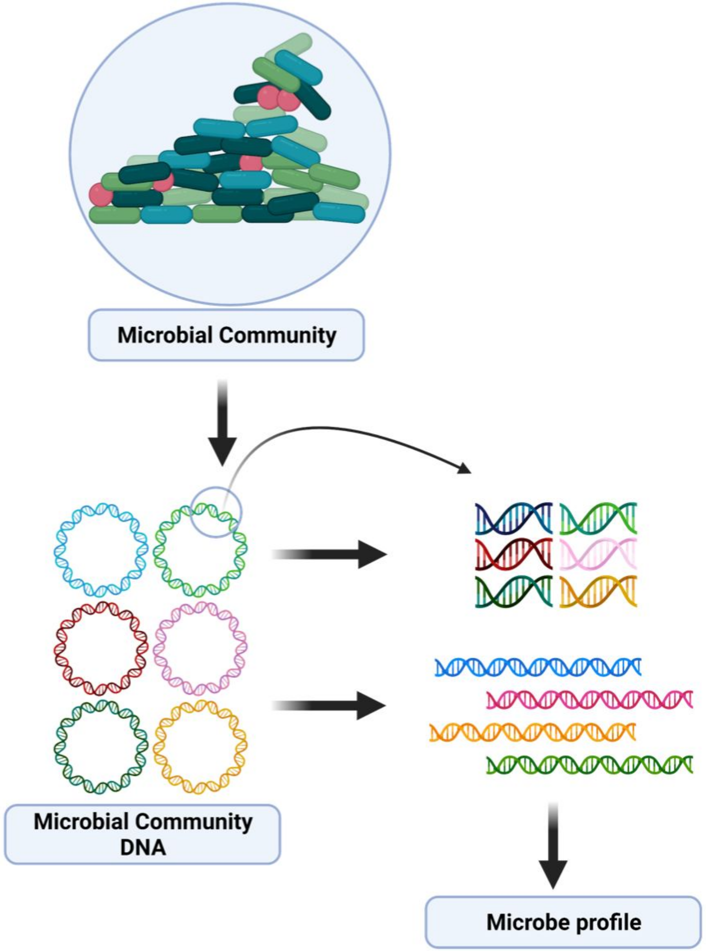
Coustenis et al. 2019 (COSPAR)

Changes needed based on recent studies

Traits of Concern

Species that:

- Survive at extreme temperatures
- Form spores
- Have an anaerobic metabolism
- Are radiation resistant
- Have salt resistance
- Form biofilms
- …



Traits of Concern

Species that:

- **Survive at extreme temperatures**
- **Form spores**
- **Have an anaerobic metabolism**
- **Are radiation resistant**
- Have salt resistance
- Form biofilms
- …

Traits of Concern

Species that:

- **Survive at extreme temperatures**
- **Form spores**
- **Have an anaerobic metabolism**
- **Are radiation resistant**
- Have salt resistance
- Form biofilms

Psychrophilic (<21 C)
Mesophilic (21-39 C)
Extremophilic (>39 C)

Traits of Concern

Species that:

- **Survive at extreme temperatures**
- **Form spores**
- **Have an anaerobic metabolism**
- **Are radiation resistant**
- Have salt resistance
- Form biofilms

Endospores

Traits of Concern

Species that:

- **Survive at extreme temperatures**
- **Form spores**
- **Have an anaerobic metabolism**
- **Are radiation resistant**
- Have salt resistance
- Form biofilms

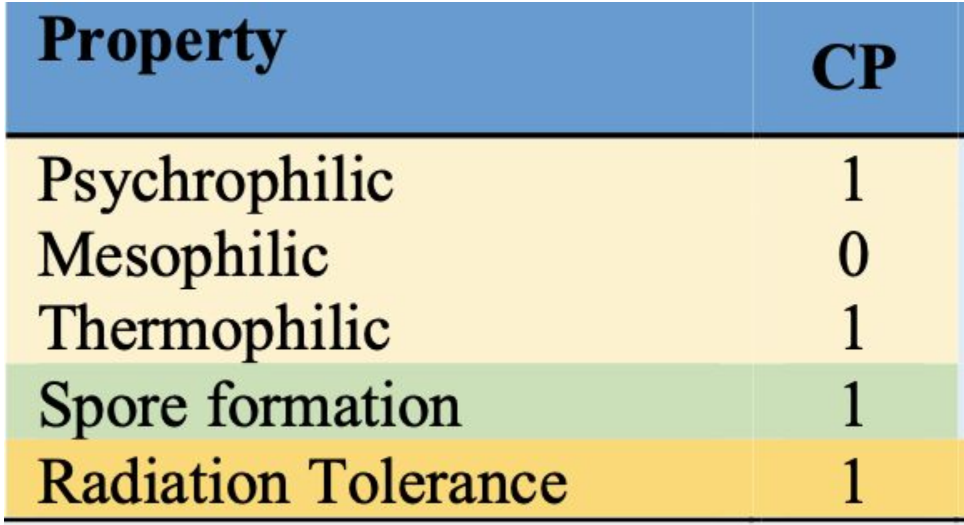
Obligate
Facultative
Microaerophile
Aerotolerant

Curating datasets of known contaminators

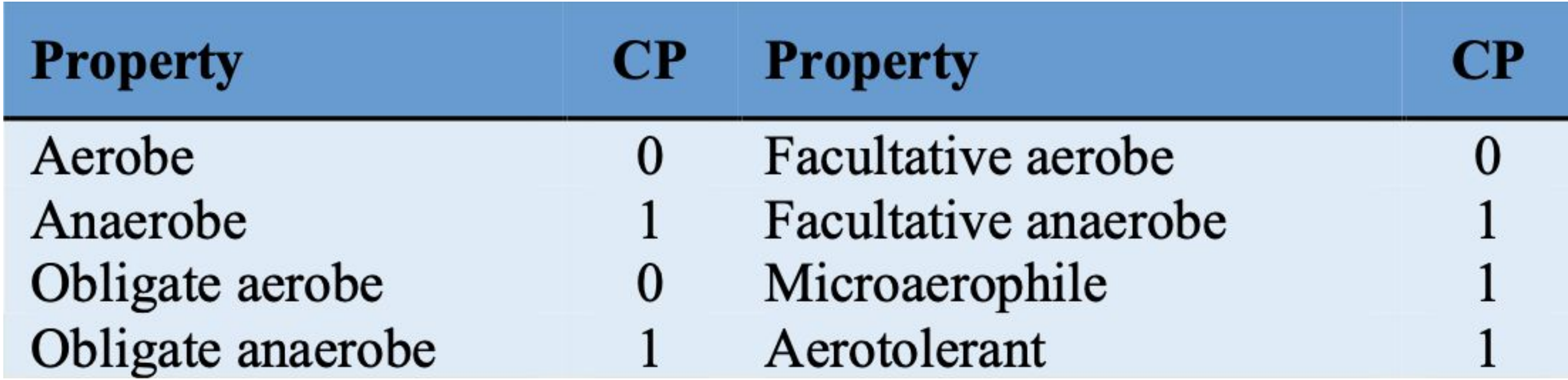
Starting with bacteria phyla:

- Actinobacteria
- Bacteroidetes
- Cyanobacteria
- Deinococcus-Thermus
- Firmicutes
- Proteobacteria

Current set of properties



Property	CP
Psychrophilic	1
Mesophilic	0
Thermophilic	1
Spore formation	1
Radiation Tolerance	1



Property	CP	Property	CP
Aerobe	0	Facultative aerobe	0
Anaerobe	1	Facultative anaerobe	1
Obligate aerobe	0	Microaerophile	1
Obligate anaerobe	1	Aerotolerant	1

Curating datasets of known contaminators

Starting with bacteria phyla:

- Actinobacteria
- Bacteroidetes
- Cyanobacteria
- Deinococcus-Thermus
- Firmicutes
- Proteobacteria

Score 4 Fermicutes:
Bacillus haynesii
Bacillus kiskunsagensis
Bacillus swezeyi
Brevibacillus gelatini
Desulfocucumis palustris
Desulfuribacillus stibiiarsenatis
Kineothrix alysoides
Mobilisporobacter senegalensis
Paenibacillus etheri
Paenibacillus silvae
Scopulibacillus daqui
Sporolactobacillus pectinivorans
Wukongibacter baidiensis

Curating datasets of known contaminators

Starting with bacteria phyla:

- Actinobacteria
- Bacteroidetes
- Cyanobacteria
- Deinococcus-Thermus
- Firmicutes
- Proteobacteria

Score 3 bacteria:
Actinomyces vulturis
Raineyella antarctica
Deinococcus aluminii
Deinococcus saudiensis
Microvirga lupini
Microvirga soli
Hymenobacter deserti

Checking Contamination

```
pip install checkContaminants
```

https://checkcontaminants.github.io/checkSpaceContamination/

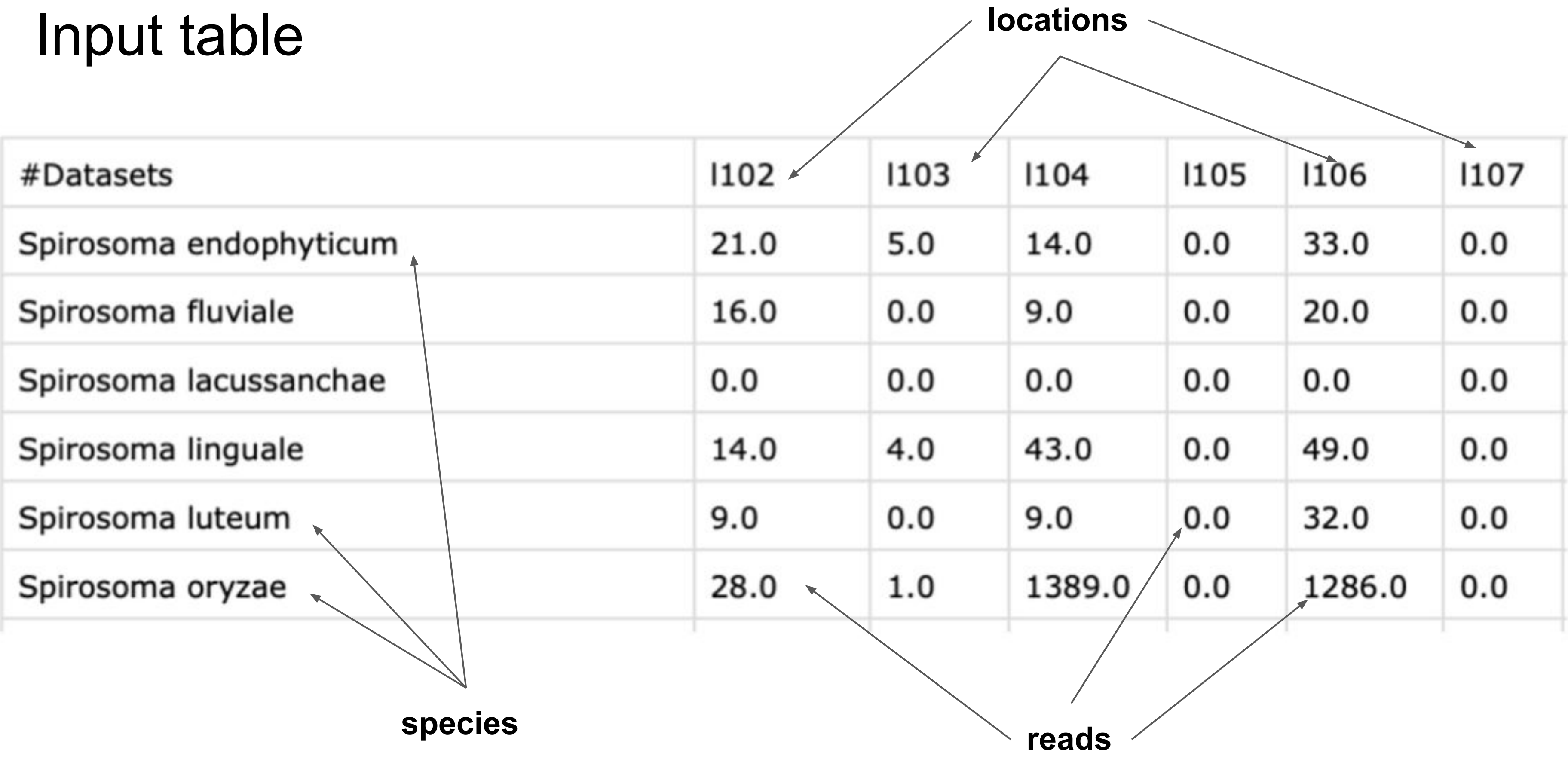
Three “parameters”:

1. Curated Species: List of species with values for important traits
2. Contamination wts for the different parameters (e.g. aerobe = 0, radiation resistant = 1)
3. Threshold of reads (to cater to low biomass needs)

All can be changed by the user

Can also provides weights other than 1/0 to properties

Input table



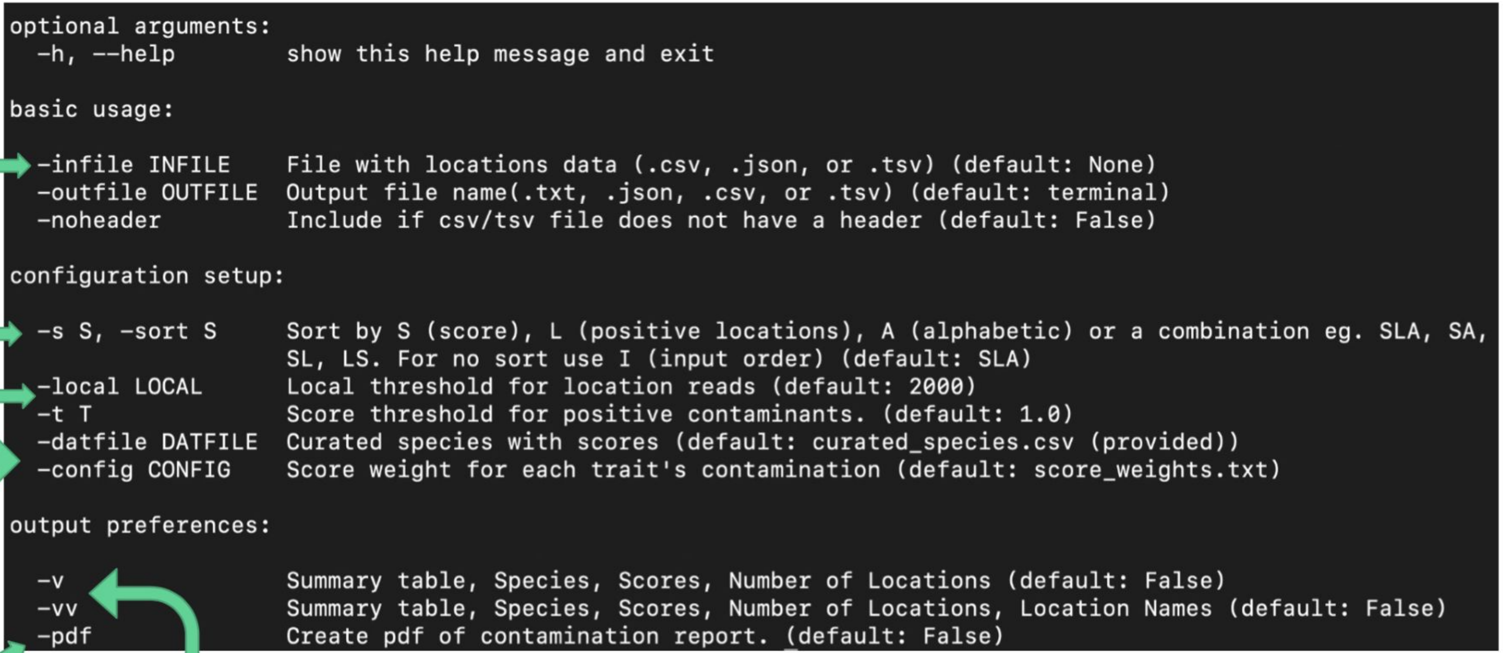
locations

#Datasets	l102	l103	l104	l105	l106	l107
Spirosoma endophyticum	21.0	5.0	14.0	0.0	33.0	0.0
Spirosoma fluviale	16.0	0.0	9.0	0.0	20.0	0.0
Spirosoma lacussanchae	0.0	0.0	0.0	0.0	0.0	0.0
Spirosoma linguale	14.0	4.0	43.0	0.0	49.0	0.0
Spirosoma luteum	9.0	0.0	9.0	0.0	32.0	0.0
Spirosoma oryzae	28.0	1.0	1389.0	0.0	1286.0	0.0

species

reads

Python package: Usage and Options



```
optional arguments:
  -h, --help        show this help message and exit

basic usage:

  -infile INFILE    File with locations data (.csv, .json, or .tsv) (default: None)
  -outfile OUTFILE  Output file name(.txt, .json, .csv, or .tsv) (default: terminal)
  -noheader         Include if csv/tsv file does not have a header (default: False)

configuration setup:

  -s S, -sort S     Sort by S (score), L (positive locations), A (alphabetic) or a combination eg. SLA, SA,
                    SL, LS. For no sort use I (input order) (default: SLA)
  -local LOCAL      Local threshold for location reads (default: 2000)
  -t T              Score threshold for positive contaminants. (default: 1.0)
  -datfile DATFILE  Curated species with scores (default: curated_species.csv (provided))
  -config CONFIG    Score weight for each trait's contamination (default: score_weights.txt)

output preferences:

  -v                Summary table, Species, Scores, Number of Locations (default: False)
  -vv               Summary table, Species, Scores, Number of Locations, Location Names (default: False)
  -pdf              Create pdf of contamination report. (default: False)
```

Input filename is required

Sort results

Set your own thresholds

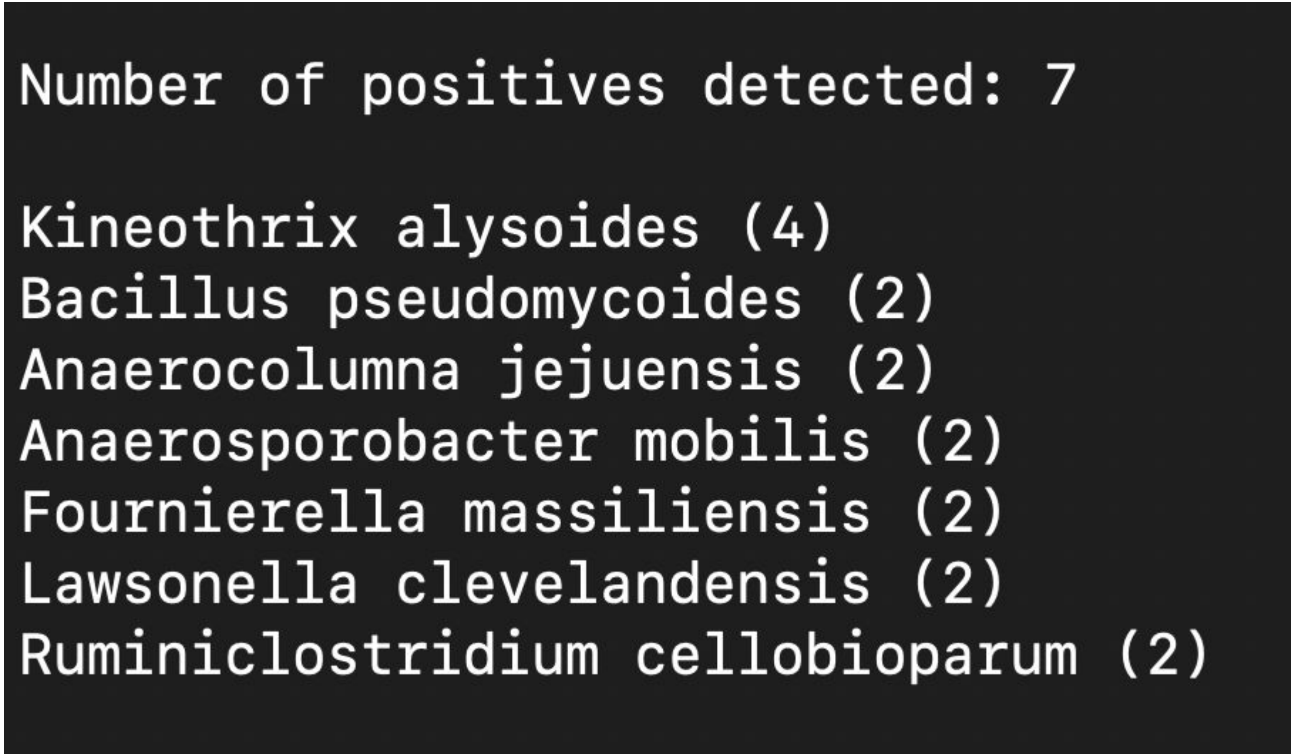
To replace configuration files

Create a pdf with charts and venn diagrams

Change how detailed the output is

Different verbosities of output available

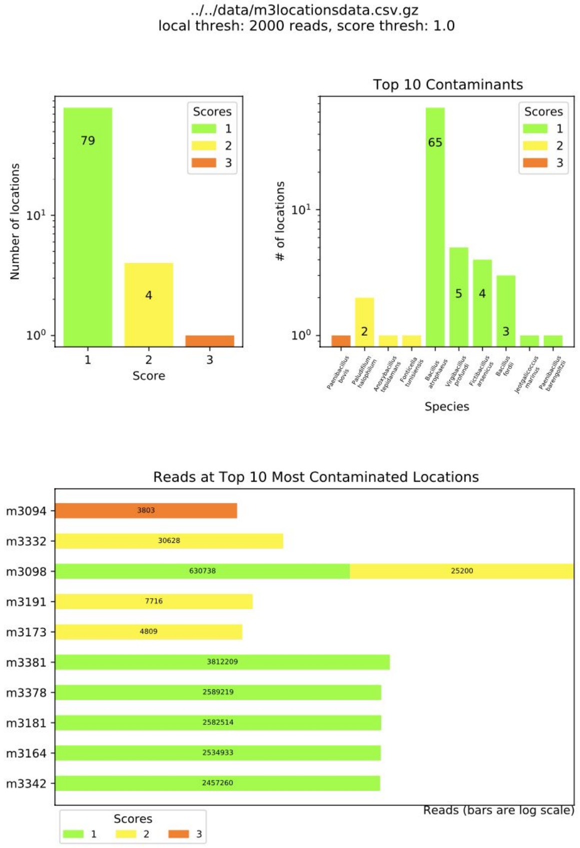
Locations
Scores
Species
Reads

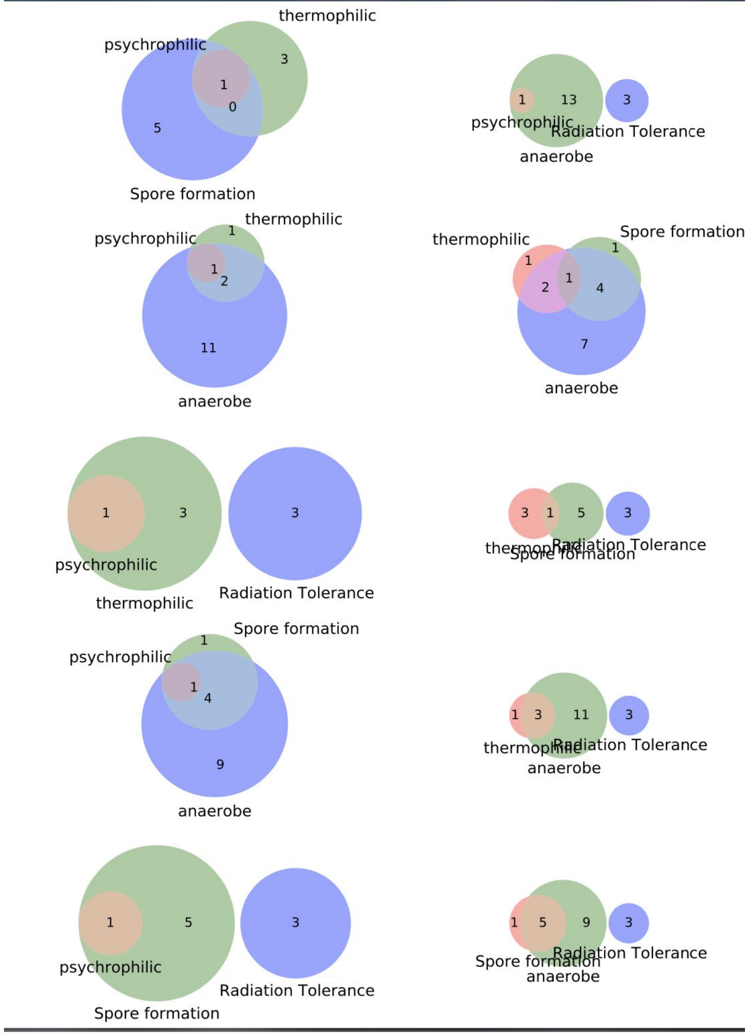


```
Number of positives detected: 7

Kineothrix alysoides (4)
Bacillus pseudomycoides (2)
Anaerocolumna jejuensis (2)
Anaerosporobacter mobilis (2)
Fournierella massiliensis (2)
Lawsonella clevelandensis (2)
Ruminiclostridium cellobioparum (2)
```

Graphical output





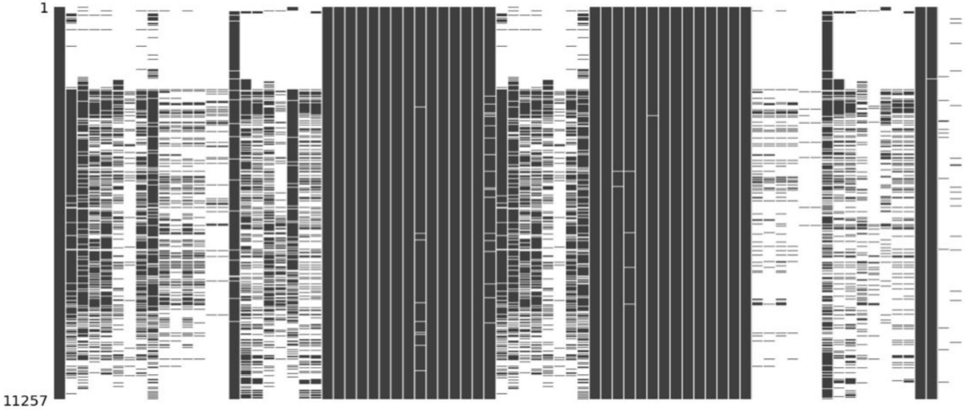
More radiation resistant?

More psychrophilic?

Issues in scaling

Curated sets small

Many holes in properties of bigger samples

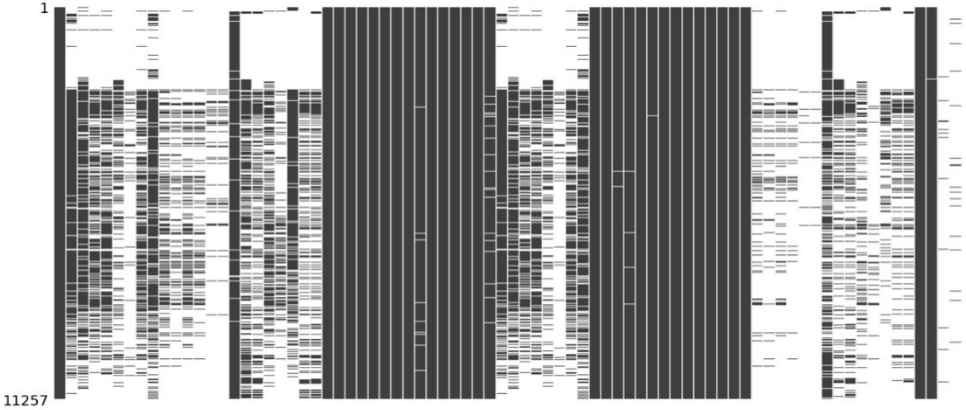


78 columns across 11257 species in GTDB reveal holes

Issues in scaling

Curated sets small

Many holes in properties of bigger samples



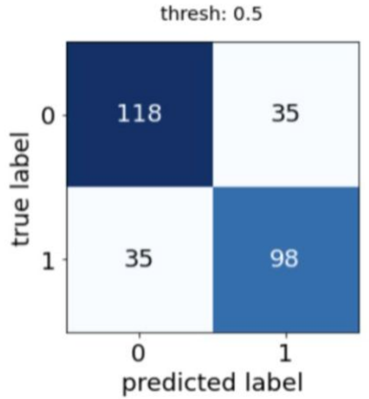
78 columns across 11257 species in GTDB reveal holes

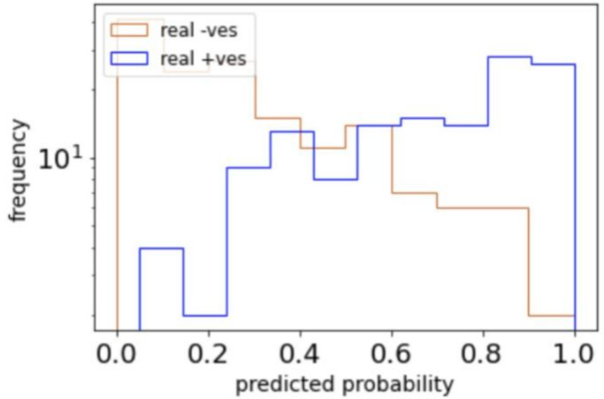
Possible Solution

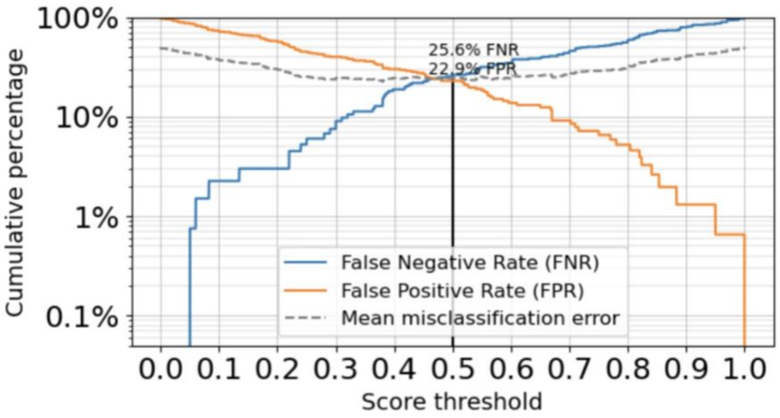
Machine Learning to flag species based on similarity measures

Results from Classifiers: Sporulation

	Random Forest	Naive Bayes	RF with Cross Validation	XGBoost
Accuracy	0.76	0.71	0.76	0.75
Precision	0.74	0.65	0.75	0.73
Recall	0.74	0.78	0.75	0.76
F1 Score	0.74	0.71	0.75	0.74



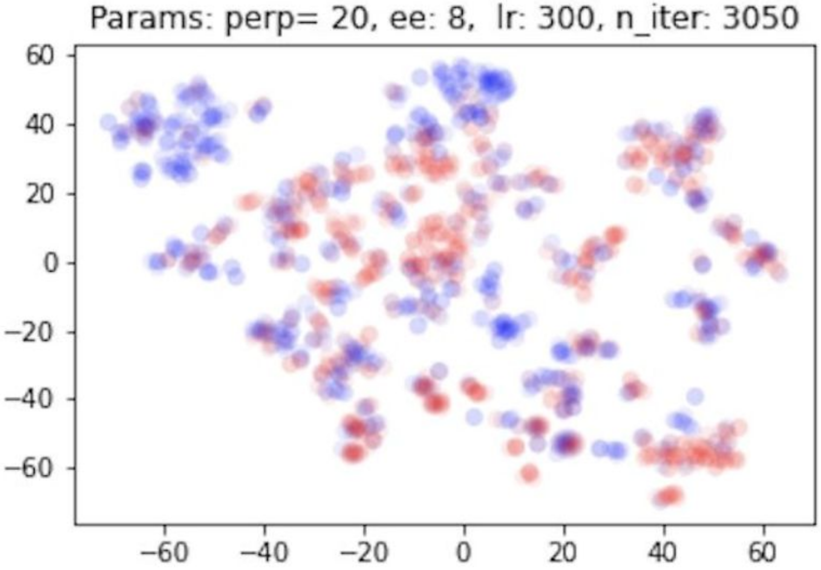




Misclassifications, or outliers?

KL divergence minimised grid search to find optimal hyperparameters revealed
2 blue clusters,
1 red cluster,
otherwise indistinguishable mix of the two types

Initial conclusion: misclassification is likely result of unclear distinction between classes



Blue: Sporulating
Red: Non-sporulating

Towards combining diverse datasets

If we want to generate larger curated datasets, we need to bring together more diverse datasets, including those containing different strains and their properties.

GTDB, NCBI, MALDI-MSI, …

Comparing and combining datasets is non-trivial unless they are standardized.

Towards combining diverse datasets

If we want to generate larger curated datasets, we need to bring together more diverse datasets, including those containing different strains and their properties.

GTDB, NCBI, MALDI-MSI, …

Comparing and combining datasets is non-trivial unless they are standardized.

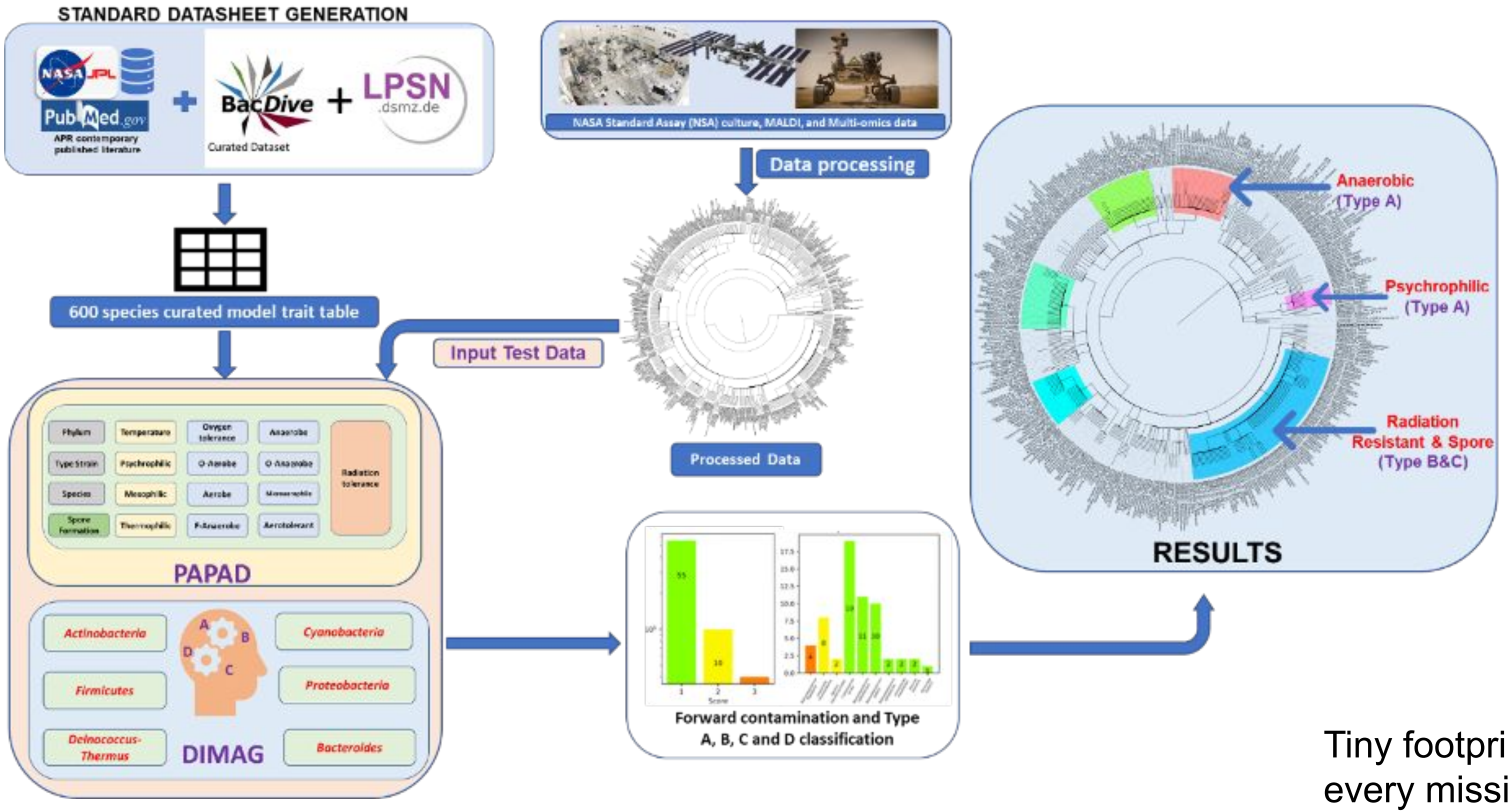
Solution:

Standardize with Data Sheets (https://doi.org/10.1145/3458723), and

Model Cards (https://doi.org/10.1145/3287560.3287596).

JPL DSWG with Subin (Rachael) Kim [SURF]

Planetary Analysis and Protection Assurance Database (PAPAD) and Dual Intelligence Manually Assessed Grouping (DIMAG)



Phylum	Temperature	Oxygen tolerance	Anaerobe	Radiation tolerance
Type Strain	Psychrophilic	O-Aerobe	O-Anaerobe	
Species	Mesophilic	Aerobe	Microaerophile	
Spore Formation	Thermophilic	F-Anaerobe	Aerotolerant	

Tiny footprint tool for every mission

Summary

NASA Jet Propulsion Laboratory
California Institute of Technology

Caltech

CENTER FOR DATA-DRIVEN DISCOVERY

PP requirements evolving

A tool created to check contamination

Need to standardize and merge datasets

Develop larger curated dataset with aid of ML

Extend to Fungi etc.