Data Science for Space

Sašo Džeroski

Jozef Stefan Institute
Jozef Stefan International Postgraduate School
Bias Variance Labs

Ljubljana, Slovenia

Just what is artificial intelligence?

Artificial Intelligence: Concerned with performing tasks that are deemed to require intelligence when done by humans

Artificial Intelligence **is not just** Machine Learning

- Knowledge representation
- Reasoning
- Planning
- (Machine) Learning
- Natural Language Processing
- Perception (e.g., Computer Vision)

Machine Learning

Just like learning is crucial for human intelligence, Machine Learning is the key to Artificial Intelligence

Learning = Improving performance at a task (e.g., land cover classification) with experience, where experience may include observed or actively collected data

Machine Learning **is not just** learning from massive data

E.g., in reinforcement learning, agents learn to act in an environment by receiving feedback/ reinforcement

Data Science

Most of Machine Learning is about learning from data

Also known under the names of Data Mining and

Data Science

Leaning models from data

- Supervised learning = Predictive modeling [Classification, Regression]
- Unsupervised learning = Clustering [Descriptive modeling]

Machine Learning **is not just** Deep Learning

Different types of models

- Black-box models (incl. Deep Learning)
- Understandable models (Explainable artificial intelligence)

Recent Trends in ML / Data Science

Learning understandable models

Mining big and complex data

- Multi-target prediction
- Semi-supervised learning
- Mining data streams
- Mining data in context [Spatio-temporal, Network]

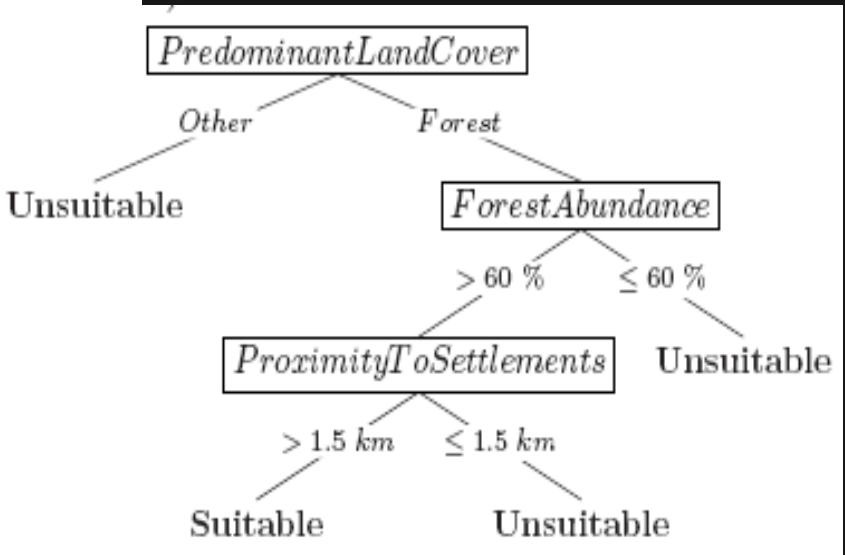
Learning Understandable Models

Input: Data

Output: Predictive model

Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
l1	Forest	80	21.4	...	Yes
l2	Forest	66	13.9	...	Yes
l3	Forest	55	50.0	...	No
l4	Forest	72	1.2	...	No
l5	Grassland	6	19.1	...	
l6	Grassland	0	11.4	...	
l7	Wetland	3	5.8	...	
l8	Water	0	3.9	...	

```
IF   PREDOMINANT-LAND-COVER = Forest
AND  FOREST-ABUNDANCE > 60%
AND  PROXIMITY-TO-SETTLEMENTS > 1.5 km
THEN BrownBearHabitat = Suitable
```



Predictive modeling classics: Classification and regression

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	Yes
Example 3	1	FALSE	0.08	0.07	No
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	Yes
…	…				…

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	0.84
Example 2	2	FALSE	0.08	0.07	0.75
Example 3	1	FALSE	0.08	0.07	0.11
Example 4	2	TRUE	0.49	0.69	0.52
Example 5	3	TRUE	0.49	0.69	0.35
Example 6	4	FALSE	0.08	0.07	0.78
…	…				…

Multi-target prediction

- Classification

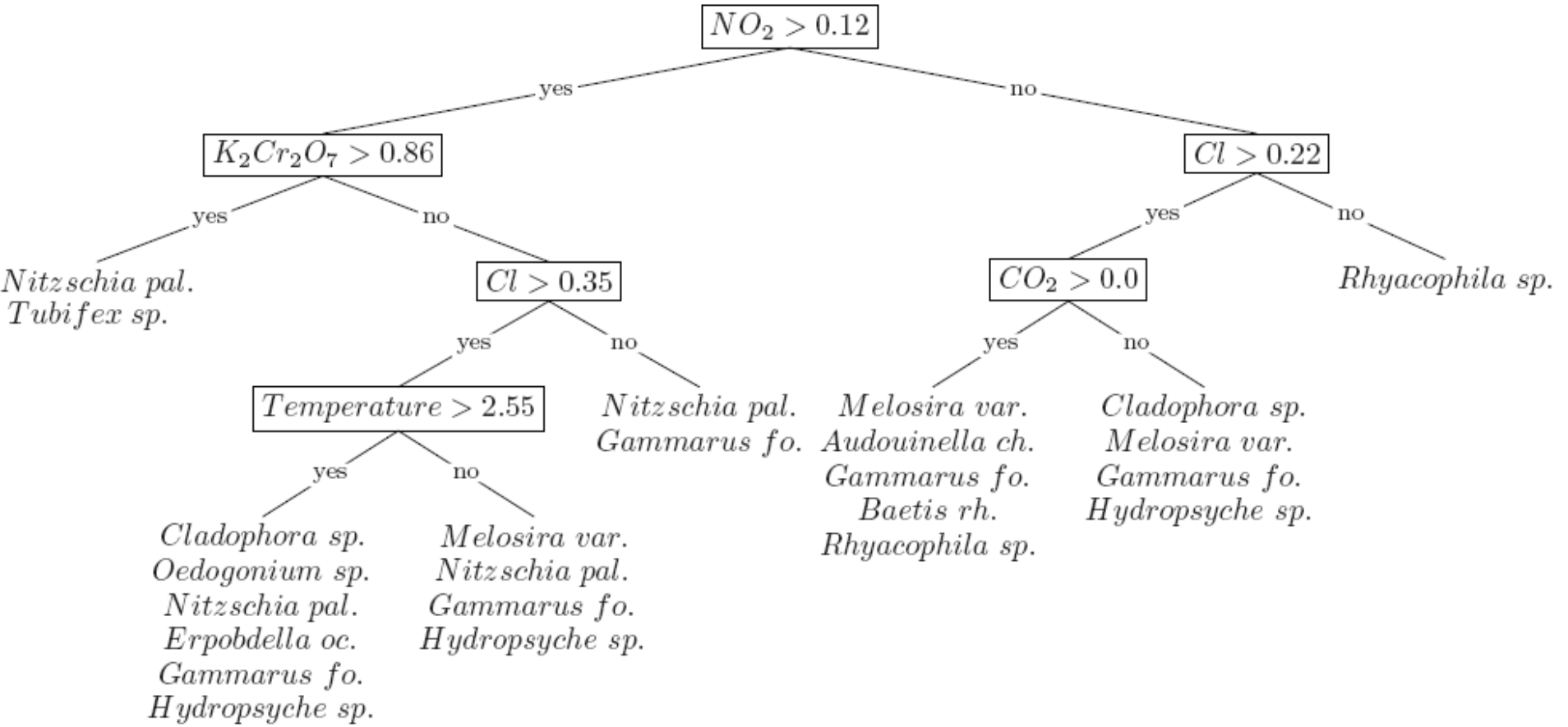
	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	Yes	Blue	Rain
Example 2	2	FALSE	0.08	0.07	Yes	Green	Sun
Example 3	1	FALSE	0.08	0.07	Yes	Blue	Cloudy
Example 4	2	TRUE	0.49	0.69	Yes	Green	Sun
Example 5	3	TRUE	0.49	0.69	No	Blue	Sun
Example 6	4	FALSE	0.08	0.07	Yes	Red	Cloudy
…	…				…	…	…

- Regression

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

Multi-Target/Label Classification

- A single tree predicting presence of many species

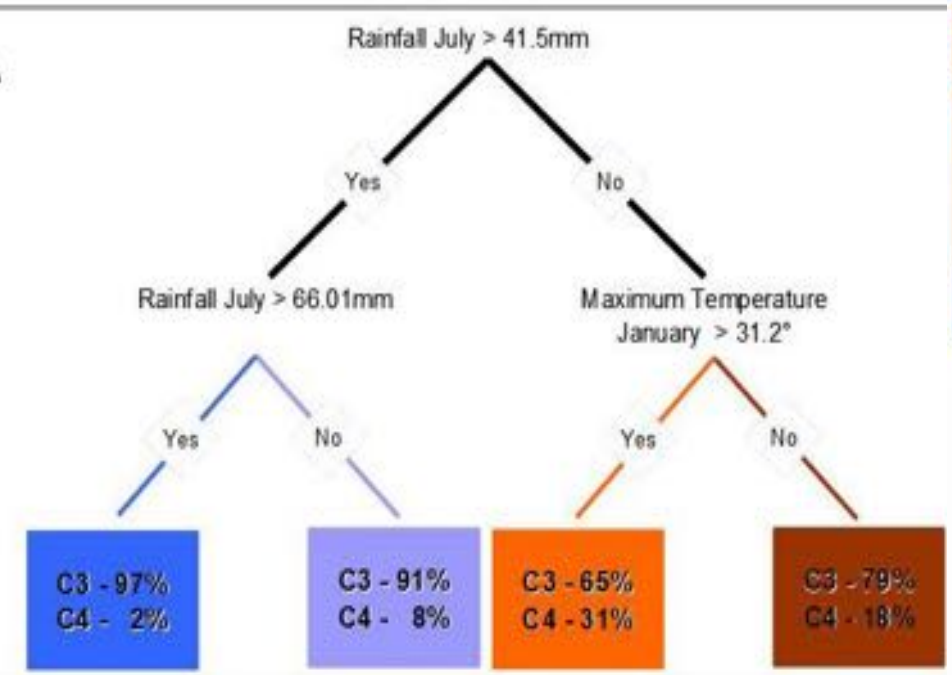


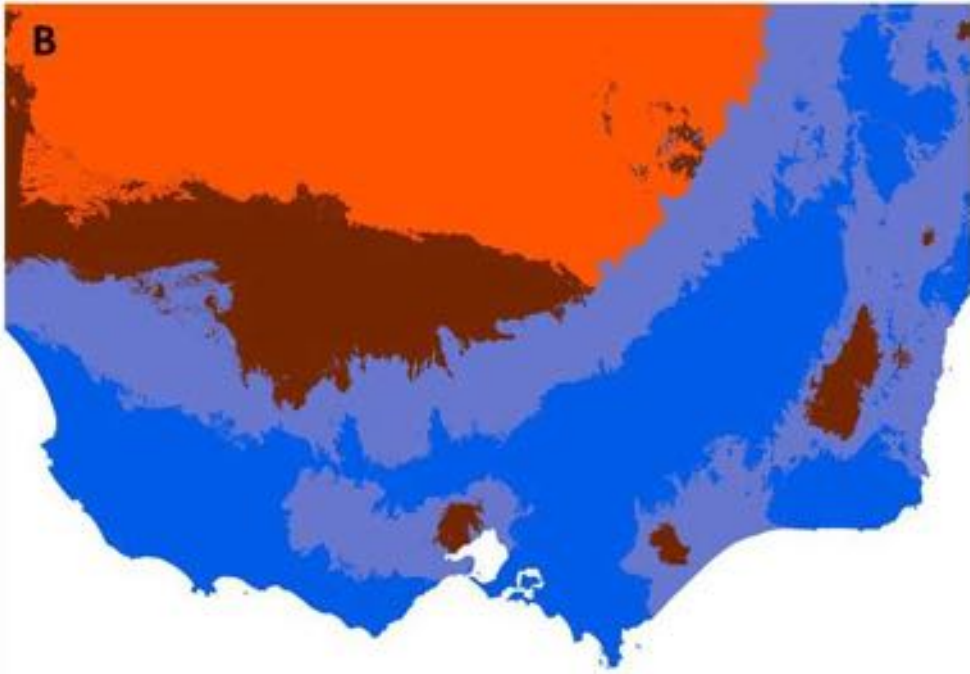
Multi-Target Regression

Predicting plant traits in Victoria, Australia:

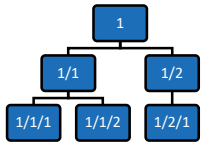
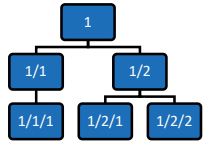
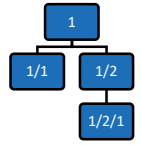
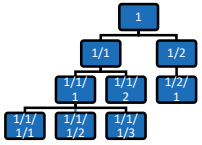
Plant photosynthetic type (carbon fixation pathways)

C3: cool-season-active; C4: warm-season-active



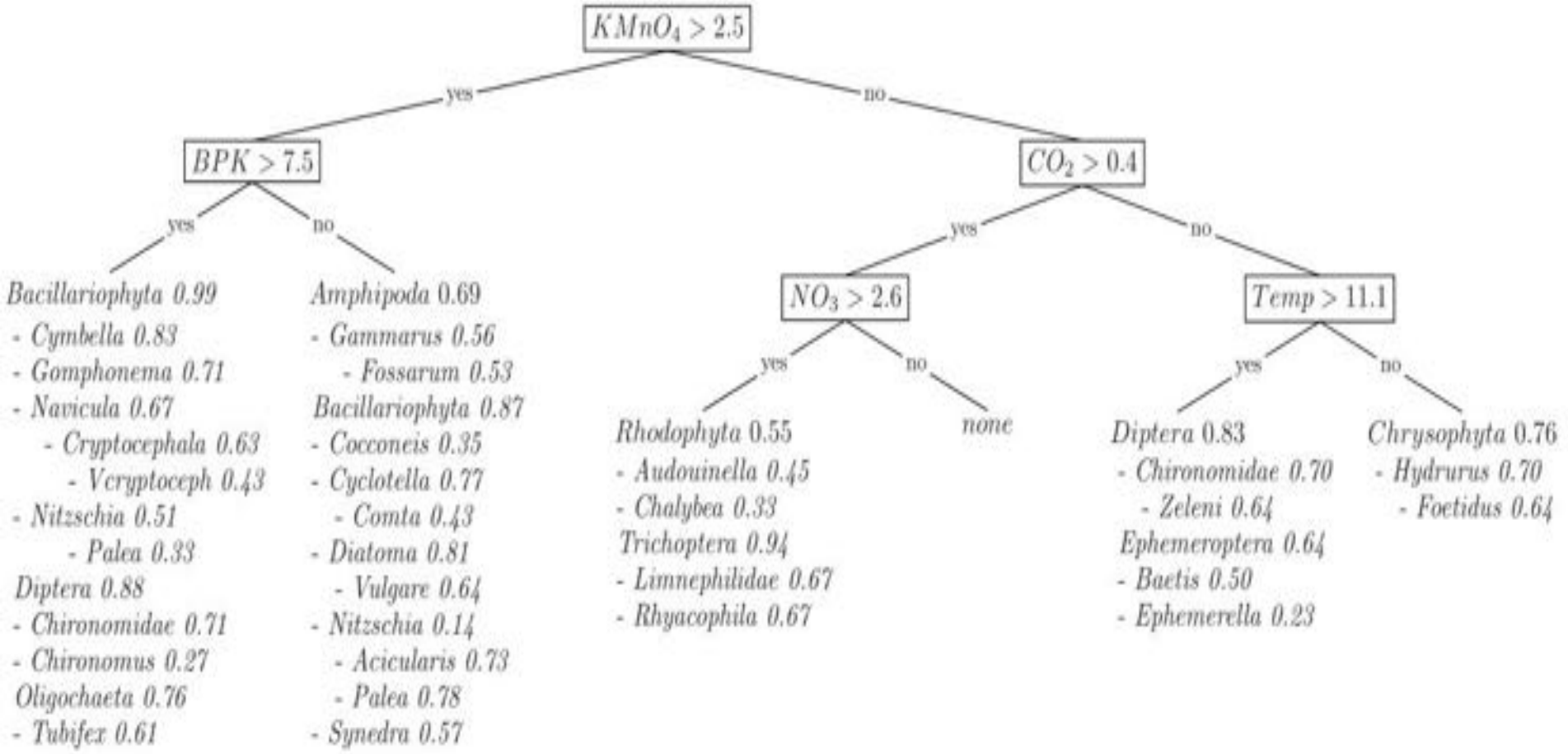


Hierarchical multi-label classification

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	
Example 2	2	FALSE	0.08	0.07	
Example 3	1	FALSE	0.08	0.07	
Example 4	2	TRUE	0.49	0.69	
…	…				…

Hierarchical multi-label classif.

- Predicting community structure (consider taxonomy)



Data streams: Classification

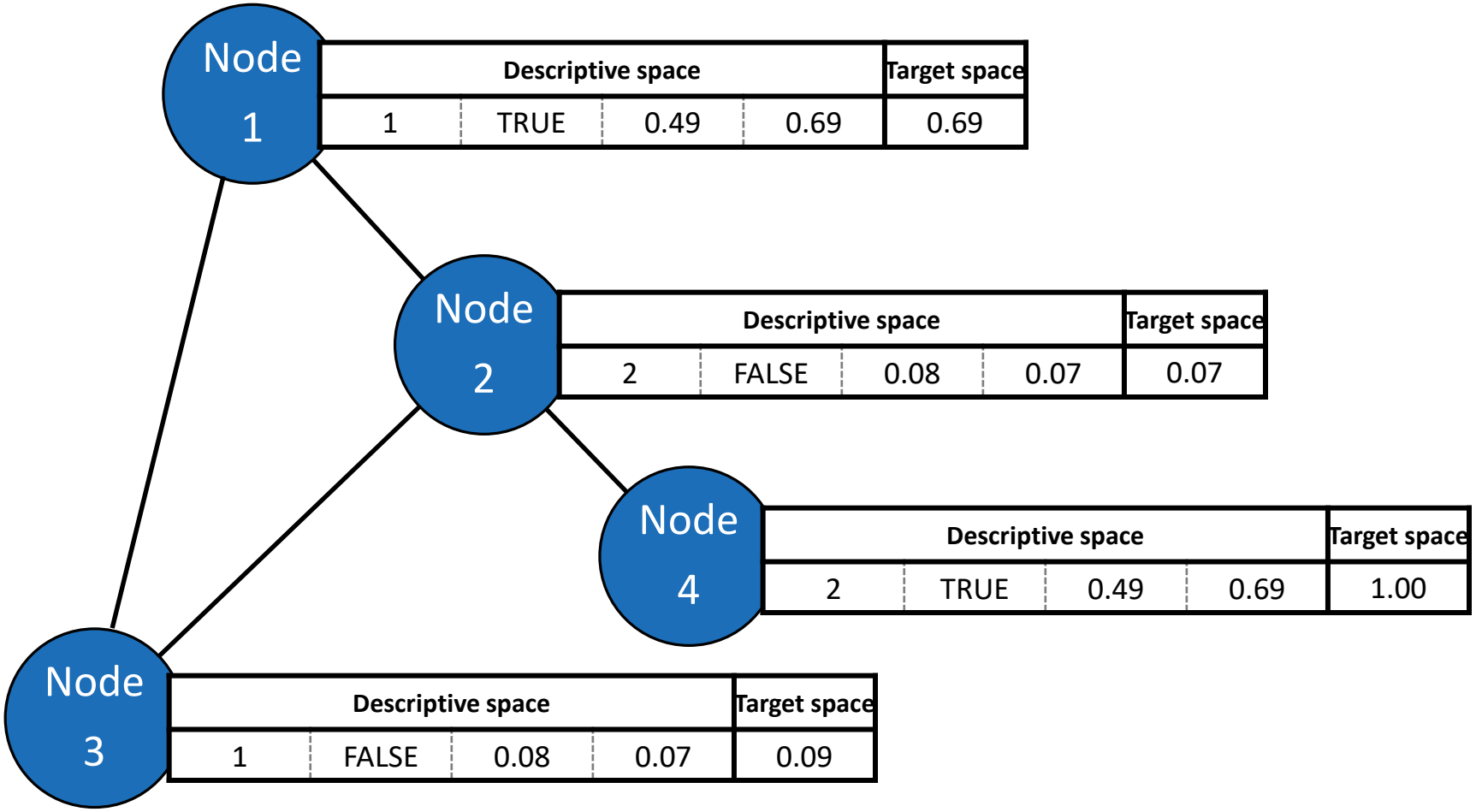
	Descriptive space				Target space
…	…				…
Example n+5	1	TRUE	0.49	0.69	Yes
Example n+1	4	FALSE	0.08	0.07	Yes
Example n+2	6	FALSE	0.08	0.07	Yes
Example n+3	8	TRUE	0.00	1.00	No
Example n+4	6	TRUE	0.00	0.00	Yes
…	…				…

Semi-supervised learning: Classification and regression

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	**?**
Example 3	1	FALSE	0.08	0.07	**?**
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	**?**
…	…				…

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	0.84
Example 2	2	FALSE	0.08	0.07	**?**
Example 3	1	FALSE	0.08	0.07	0.11
Example 4	2	TRUE	0.49	0.69	**?**
Example 5	3	TRUE	0.49	0.69	**?**
Example 6	4	FALSE	0.08	0.07	0.78
…	…				…

Data in context: Spatio-temporal, network



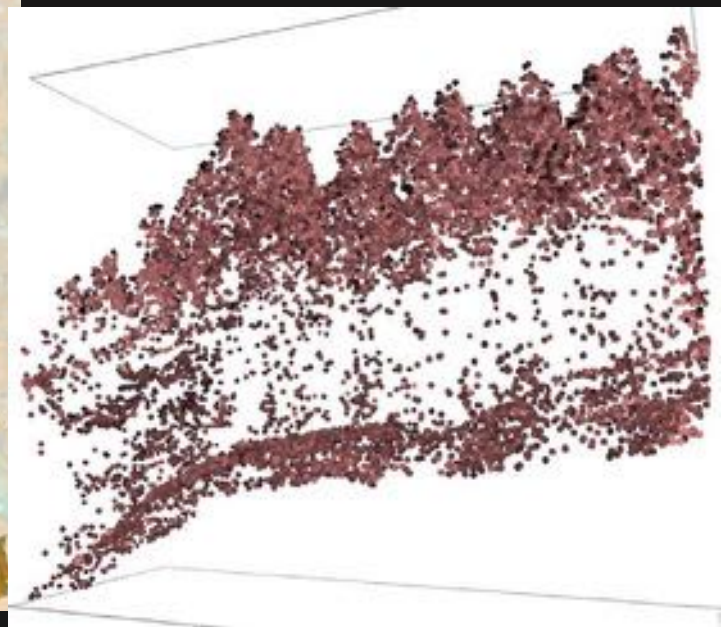
Space-related data

- Scientific data & Operations data
- Data coming from Earth observation
  - Different types of satellite images
  - Different resolutions
  - Different modalities/spectral bands
- Operations Data: spacecraft position, spacecraft on-board commands, instrument operations, science activities, power consumption, …

From satellite images + LIDAR to forest height and density

Input: Landsat images, Multi-temporal, Multi-spectral

+ LIDAR for a small patch

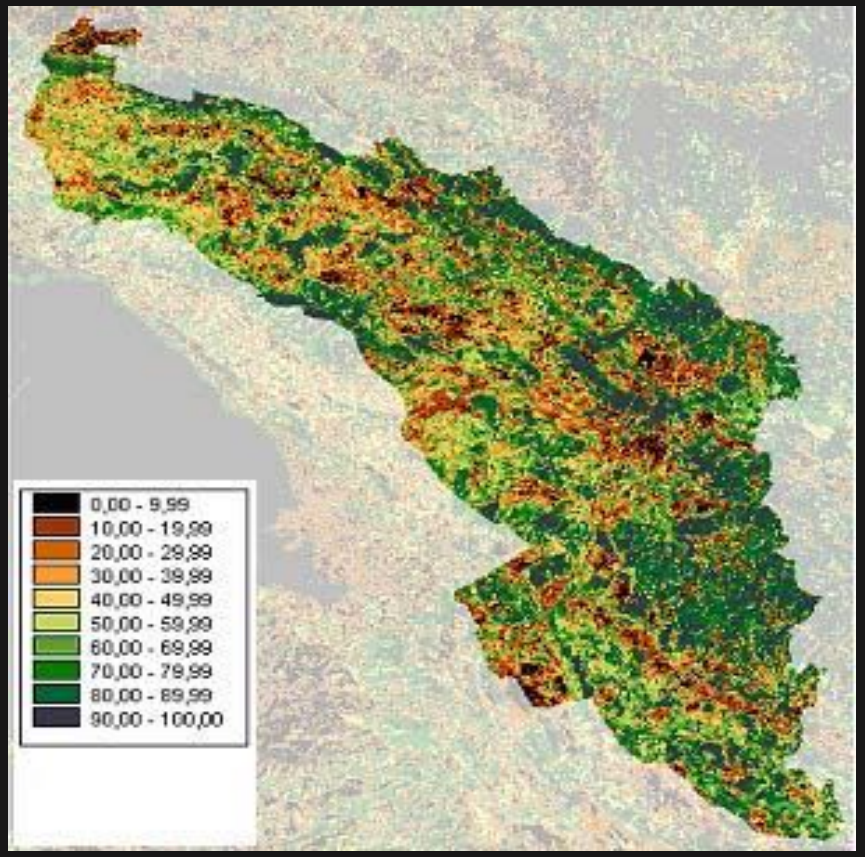


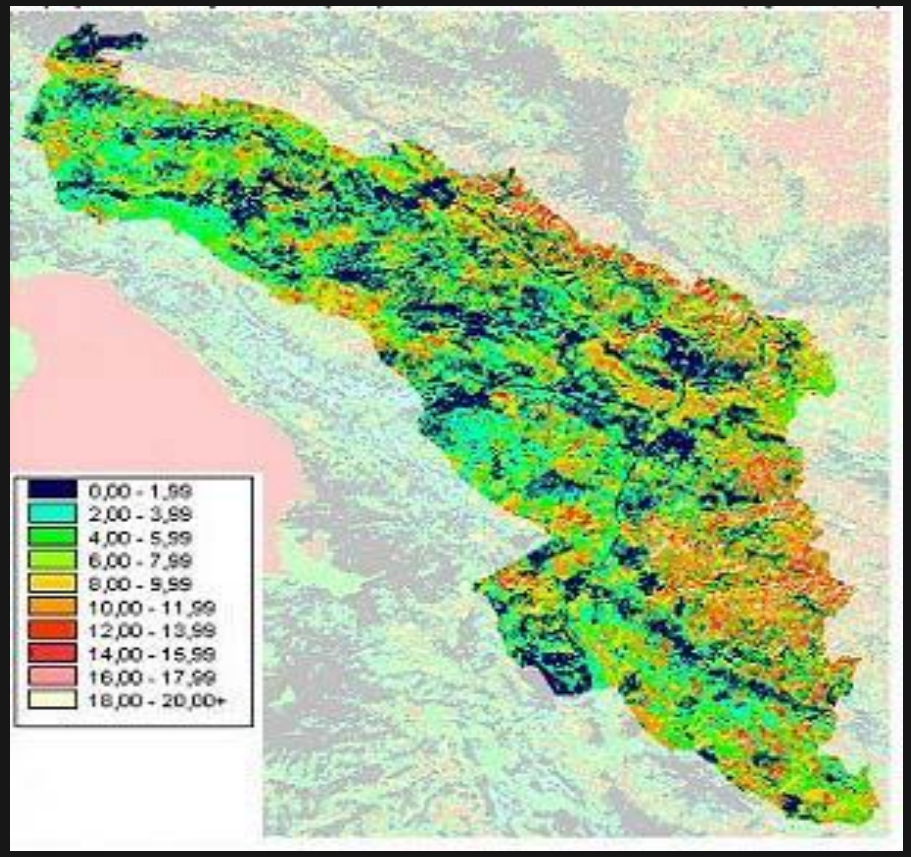
Inputs: Satellite images+LIDAR

- For a small part of the Karst region
- Calculate height & density from LIDAR
- Also vertical vegetation profiles
- Use features from satellite images to learn predictive model (multi-target)

Output: Maps of height&density

- For the Karst region in Slovenia





Hierarchical classification scheme

Level 1	Level 2	Level 3
1 Artificial surfaces	11 Urban fabric	111 Continuous urban fabric
		112 Discontinuous urban fabric
	12 Industrial, commercial and transport units	121 Industrial or commercial units
		122 Road and rail networks and associated land
		123 Port areas
		124 Airports
	13 Mine, dump and construction sites	131 Mineral extraction sites
		132 Dump sites
		133 Construction sites
	14 Artificial, non-agricultural vegetated areas	141 Green urban areas
		142 Sport and leisure facilities
2 Agricultural areas	21 Arable land	211 Non-irrigated arable land
		212 Permanently irrigated land
		213 Rice fields
	22 Permanent crops	221 Vineyards
		222 Fruit trees and berry plantations
		223 Olive groves
	23 Pastures	231 Pastures
	24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
		242 Complex cultivation patterns
		243 Land principally occupied by agriculture, with significant areas of natural vegetation
		244 Agro-forestry areas
3 Forest and semi natural areas	31 Forests	311 Broad-leaved forest
		312 Coniferous forest
		313 Mixed forest

The MEX Challenge

The Mars Express Orbiter is an aging aircraft

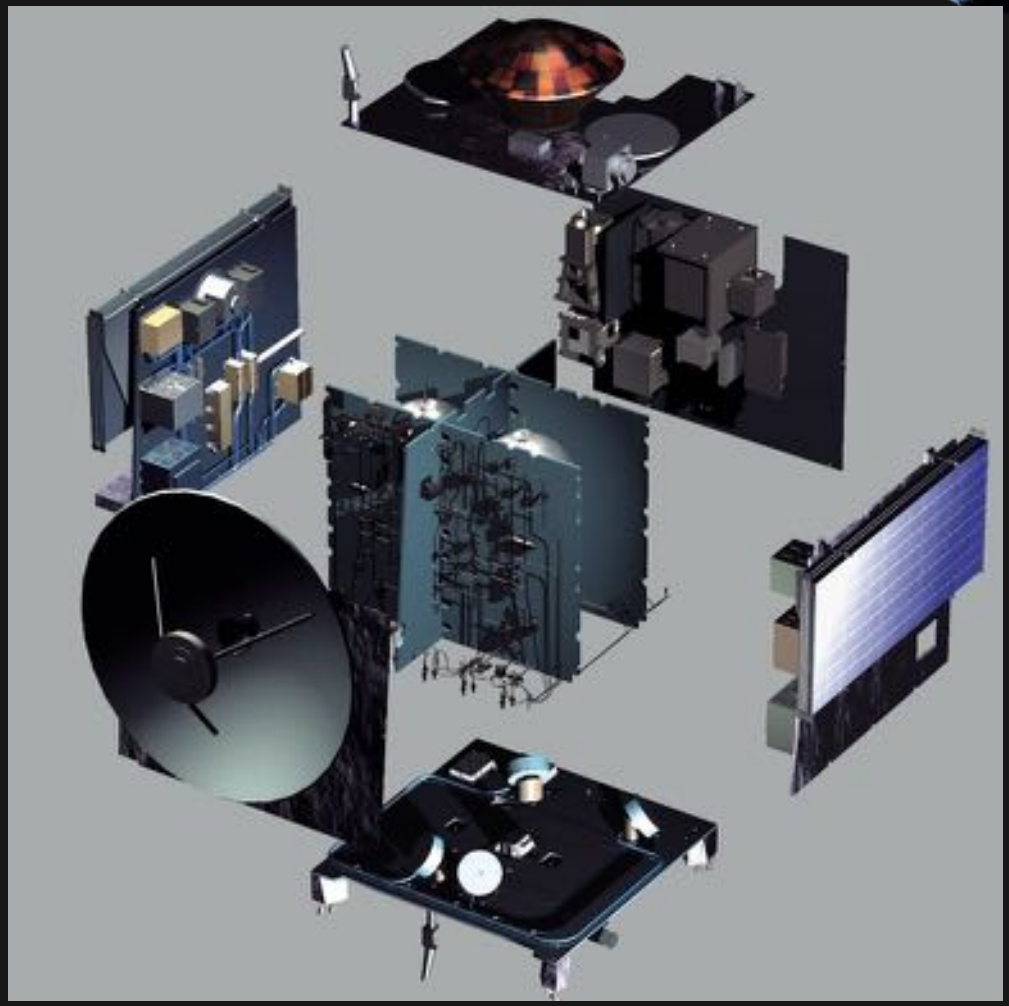
It has aging batteries

Highly variable power demand

The challenge is to

- Maximise the return of scientific data and, at the same time,
- Ensure safety and long-term health of the aircraft

The MEX Challenge



Different systems need different temperature:

- Electronics – room temp.
- Imaging sensors – low temp.

<u>Given</u>: data for three Martian years

<u>Predict:</u> required power for 33 thermal lines for the fourth Martian year (at 1 h resolution)

The MEX Challenge Data

Descriptive features:

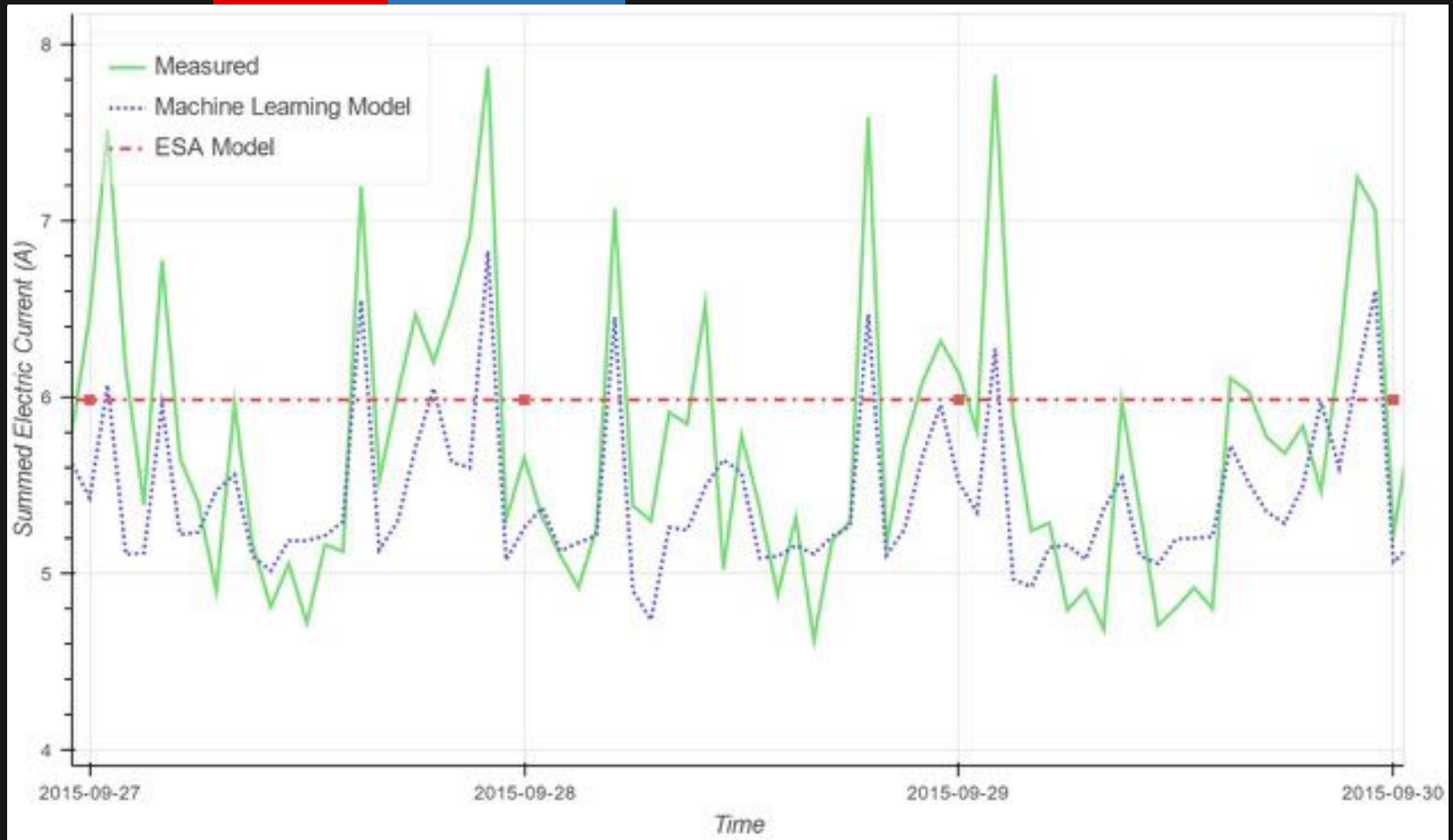
- Solar Aspect Angles (SAAF)
- Detailed Mission Operations Plan / Commands (DMOP)
- Flight Dynamics with pointing events (FTL)
- Long term data(LTDATA)
- Miscellaneous events (EVTF)

33 features for thermal power lines (targets):

- Measured electric current

Final Results

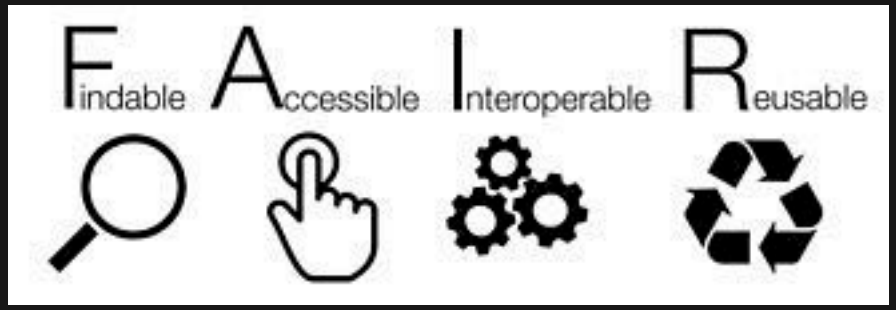
Team	ESA	MMMe8	Redrock	Formax	Alex	Luis
RMSE	0.4903	**0.0792**	0.0803	0.0819	0.0838	0.0884



The Future of Data Science in EO

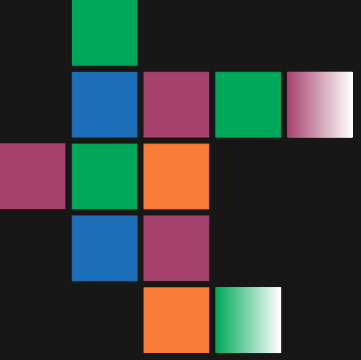
- More and more data becoming available
- More and more SW for data analysis is also becoming available
- So what's keeping us back?

EO (incl. ESA) Data is BIG! But is it FAIR?



- Findable
- Accessible
- Interoperable
- Reusable

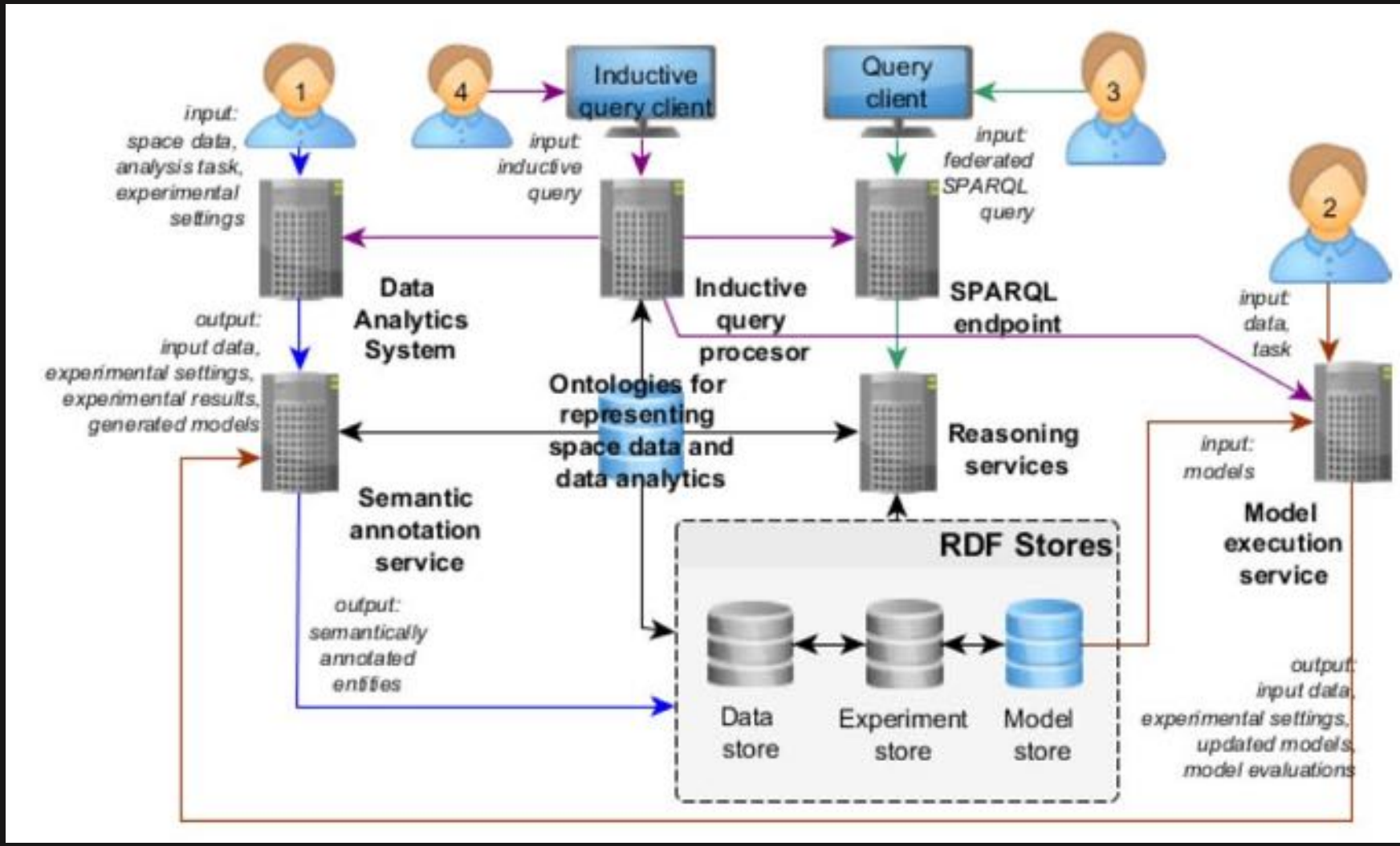
For data to be FAIR, it needs to be accompanied with some knowledge (annotations)

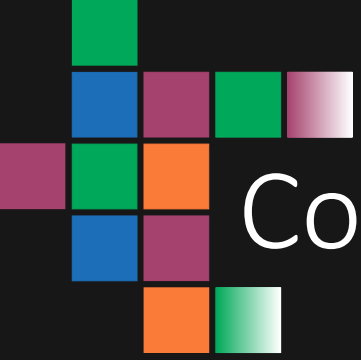


Ontologies, Open Science, Reproducible Research

- Both data and SW need to follow FAIR principles (Findable, Accessible, Interoperable and Reusable)
- Ontologies for describing the elements and processes of data science (data mining)
  - Data
  - Data Mining Tasks
  - Data Mining Algorithms
  - Data Analytics Processes/Workflows
- Together with domain ontologies, this allows
  - Precise description of the data analyses performed
  - Matching algorithms and data
  - Automated construction of data analysis workflows

Data Science Infrastructure: Accessing both Data and Models



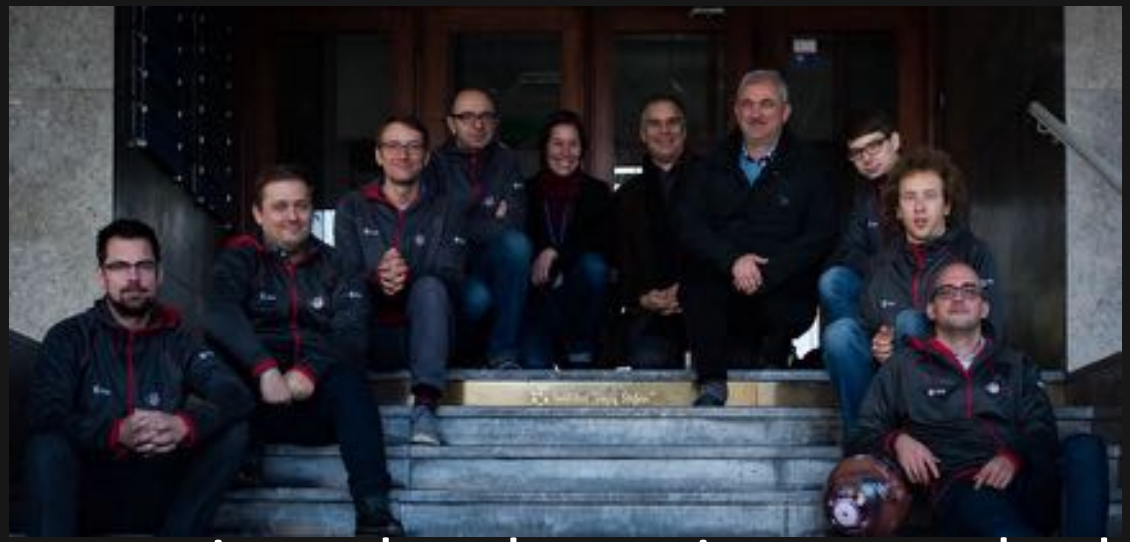


Conclusions

- Exciting new technology for mining big and complex data
- Can handle different aspects of complexity
  - Different types of structured outputs
  - Big data and data streams
  - Partially annotated data, network data
- Efficient, works fast!

  [What's the environmental footprint of deep learning?]
- Can produce accurate models
- Can produce understandable models

Thanks & Invitation to collaborate



- We develop cutting edge data science methods
- We have experience in a wide range of domains
- We have top talent, HPC infrastructure, R&D ecosystem

Jožef Stefan Institute,
Ljubljana, Slovenia

Bias
Variance
Labs