



Multi-target Prediction with Trees & Tree Ensembles

Sašo Džeroski

Jozef Stefan Institute, Ljubljana, Slovenia

Interreg



UNIONE EUROPEA
EVROPSKA UNIJA

ITALIA-SLOVENIJA



TRAIN

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



The basic Machine Learning task: Predictive modeling

- Classification

	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
...

- Regression

	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
...



An example task of Predictive Modelling: Medical diagnosis

- Predictive models focus on a target variable and predict its value from the values of input variables
- Classical problem: Medical diagnosis
- An example: Neurodegenerative diseases
- Target variable: Diagnosis; Possible values:
 - CN - Cognitively Normal (0)
 - SMC - Significant Memory Concern
 - EMCI - Early Mild Cognitive Impairment
 - LMCI - Late Mild Cognitive Impairment
 - AD - Alzheimer's Disease (4)

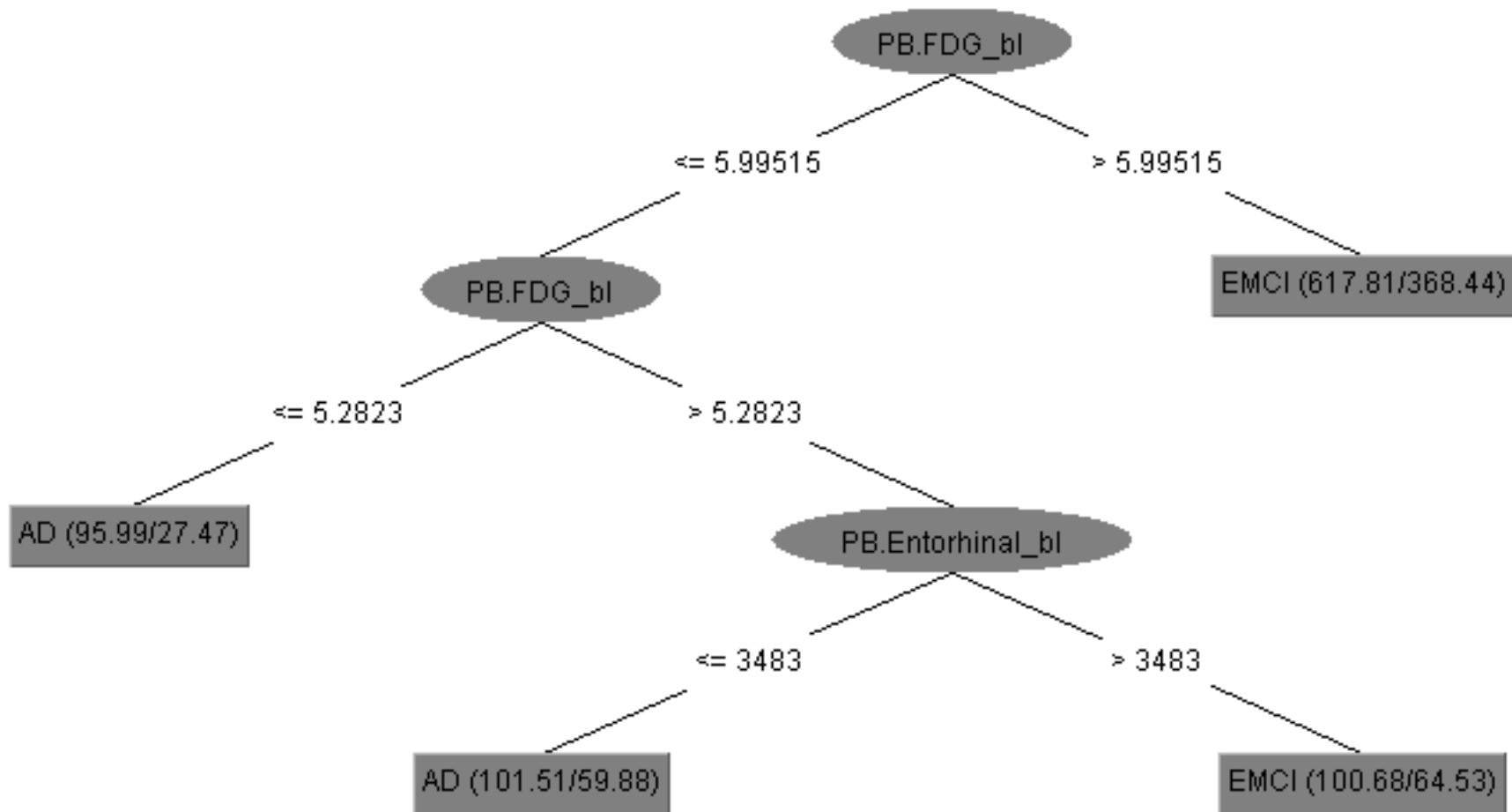


Example task: Descriptive vars.; Biomarkers for Alzheimer's

1. APOE4 – Genetic variations of APOE4 related gene
2. FDG – Positron emission tomography (PET) imaging results with [^{18}F]fluorodeoxyglucose
3. AV45 – Positron emission tomography (PET) imaging results with [^{18}F]-labeled amyloid imaging agent AV45
4. Ventricles
5. Hippocampus
6. WholeBrain
7. Entorhinal
8. Fusiform – Fusiform gyrus
9. MidTemp – Middle Temporal Gyrus
10. ICV – Intracerebral volume [Volumetric data 4-10]



Example: Decision tree for diagnosis





Another example of single-target predictive modeling (classification)

Task: Habitat suitability modeling

Input: Data on locations and habitat suitability

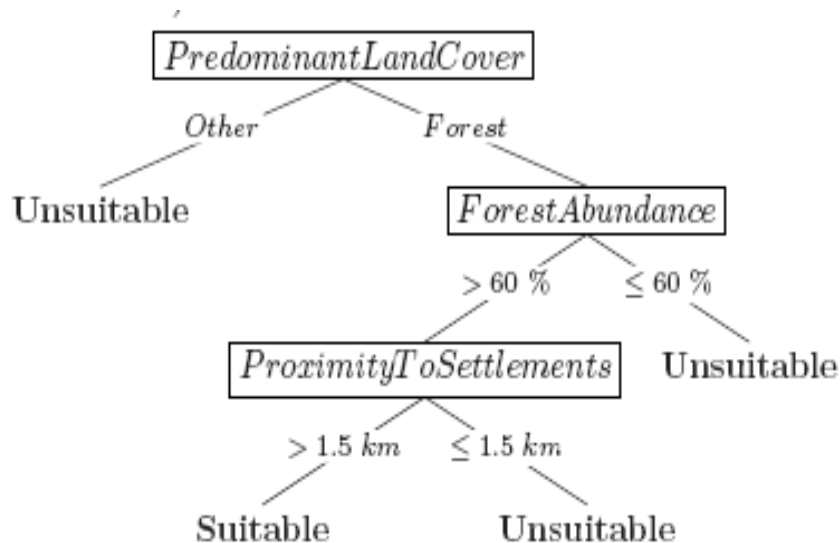
Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
11	Forest	80	21.4	...	Yes
12	Forest	66	13.9	...	Yes
13	Forest	55	50.0	...	No
14	Forest	72	1.2	...	No
15	Grassland	6	19.1	...	No
16	Grassland	0	11.4	...	No
17	Wetland	3	5.8	...	No
18	Water	0	3.9	...	No



Another example of single-target predictive modeling (classification)

Task: Habitat suitability modeling

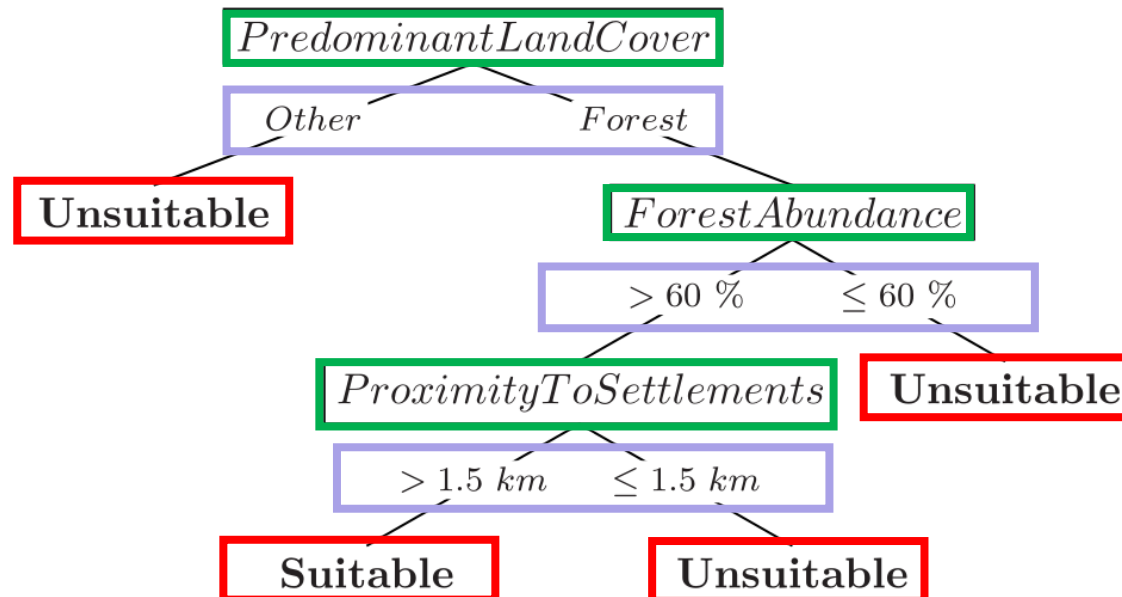
Output: Habitat suitability model

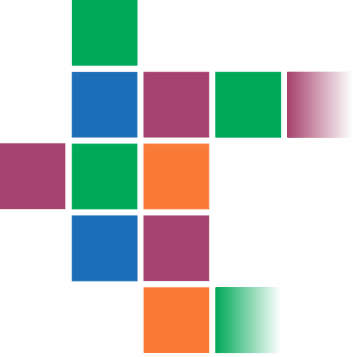


```
IF    PREDOMINANT-LAND-COVER = Forest
AND  FOREST-ABUNDANCE > 60%
AND  PROXIMITY-TO-SETTLEMENTS > 1.5 km
THEN BrownBearHabitat = Suitable
```

What is a decision tree?

- Hierarchically structured predictive model
- **Nodes** – correspond to (environmental) variables
- **Arcs** – possible values of the variables
- **Leafs** – predictions for the target variable





Making a Prediction with a Decision Tree

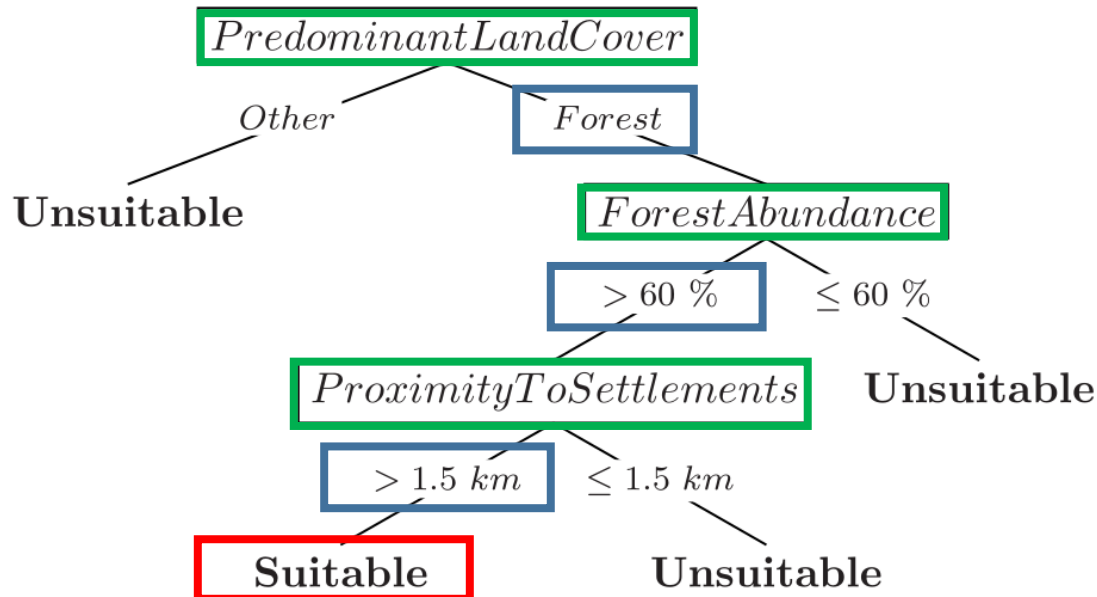
Take as input values of attributes/ independent vars.

Follow branches according to the values of these

Until you reach a leaf

PLC	FOREST-ABUNDANCE	PTS	BBH
Forest	80	21.4	?

Yes





Machine Learning of Decision Trees

Interreg

ITALIA-SLOVENIJA

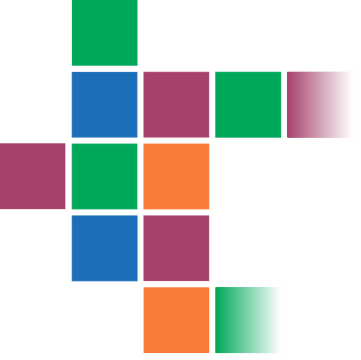


TRAIN



UNIONE EUROPEA
EVROPSKA UNIJA

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



Top-Down Induction of Decision Trees

To construct a tree T from a training set S :

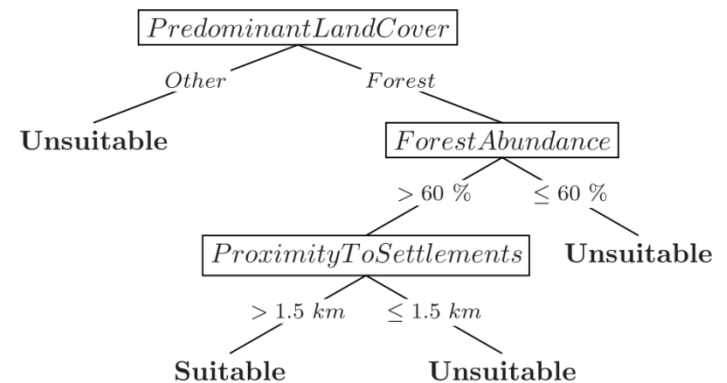
- If **all the examples belong to the same class C** , construct a leaf labeled C
- Otherwise:
 - Select the best attribute A with values v_1, \dots, v_n , which **reduces the most the impurity of the target**
 - Partition S into S_1, \dots, S_n according to A
 - Recursively construct subtrees T_1 to T_n for S_1 to S_n
 - Result: a tree with root A and subtrees T_1, \dots, T_n

TDIDT Illustrated

Input: Set of learning examples S

- 1) Find the best split t (attribute value which results in the biggest reduction of variance considering the target variable)
- 2) Partition the data S into partitions S_v according to t
- 3) For each partition, if stopping criteria met (e.g., all of the examples in partition are of the same class), make a leaf, assign a (prototype) class to leaf
- 4) Otherwise, repeat 1) for each node

Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
l1	Forest	80	21.4	...	Yes
l2	Forest	66	13.9	...	Yes
l3	Forest	78	15.2	...	Yes
l4	Forest	72	1.2	...	No





Mining Big and Complex Data: Dimensions of Complexity

Interreg



UNIONE EUROPEA
EVROPSKA UNIJA

ITALIA-SLOVENIJA



TRAIN

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



Mining Big and Complex Data

- What is big and complex data?
 - Volume & Velocity (Data Streams)
 - Variety (Structured Inputs and Structured Outputs)
- Variety:
 - Different types of data, different tasks of data mining
- MTP is a special case of structured output prediction
 - But you can have more complex outputs than in MTP
- Combination with other dimensions of complexity
 - Semi-supervised ...
 - Data streams
 - Networked data



Big Data: Volume & Velocity

- Large number of columns (high dimensionality)
 - Need feature ranking/selection
- Large number of rows (massive data)
 - Need efficient data mining methods
- Streaming rows (data streams)
 - Need incrementality: Not all data available simultaneously
 - Data instances arrive at **high velocities**, in a **specific order** and their number is **potentially arbitrarily large**
 - The **underlying concept** (distribution) governing the data **can change (concept drift)**
 - We need **fast processing** (due to the high velocity)
 - The large and potentially infinite number of examples demands **economical management of available memory**



Data streams: Regression

	Descriptive space				Target space
...
Example n	1	TRUE	0.49	0.69	0.45
Example n+1	4	FALSE	0.08	0.07	0.12
Example n+2	6	FALSE	0.08	0.07	1.54
Example n+3	8	TRUE	0.00	1.00	3.12
Example n+4	6	TRUE	0.00	0.00	0.05
...



Big Data: Variety - Structured Input

Example:

Predicting biodegradability

input datatype
specification



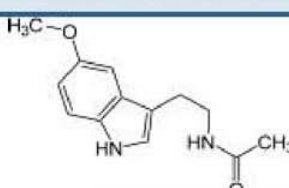
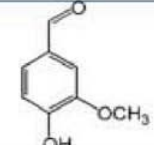
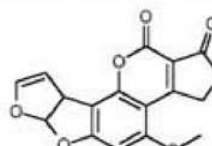
input: molecule datatype

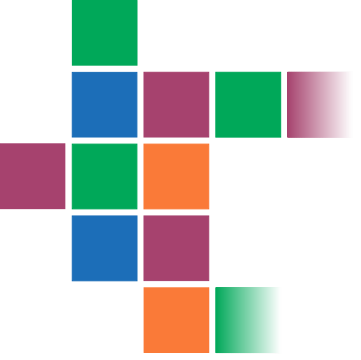
output datatype
specification



output: real datatype




data example

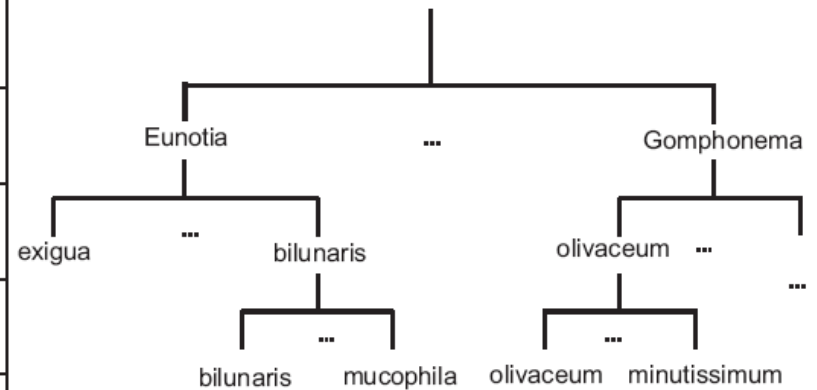
compound	activity
	0.25
	0.28
	0.37

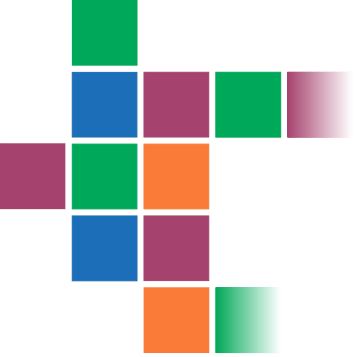


Big Data: Variety - Structured Output

- Hierarchical classification: taxonomy of diatoms
- Classifying microscope images
- The input: A vector of feature values
- The output: Not a scalar value, rather a data structure

image	features/descriptors						taxonomy
	Heuristic shape descriptors						
	48	24	59	66	37	...	olivaceum
	36	25	53	45	15	...	minutissimum
	35	25	56	52	19	...	exigua
...

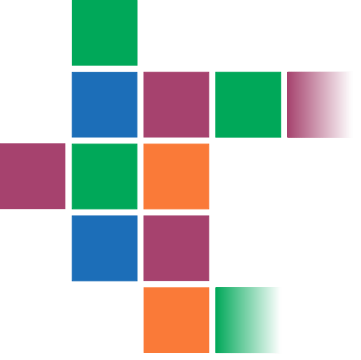




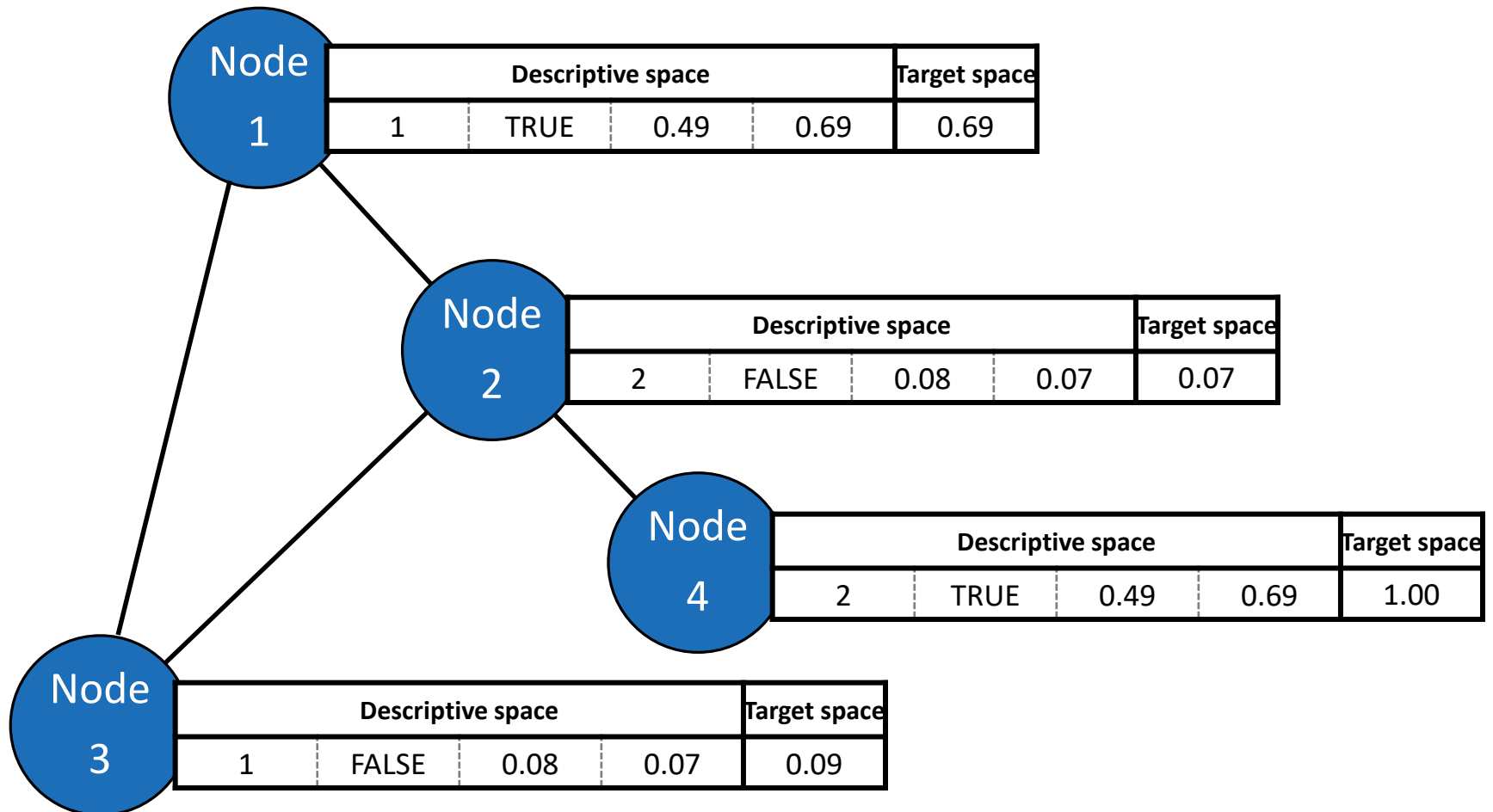
Semi-supervised learning: Classification and regression

	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





The Different Tasks of Multi-Target Prediction

Interreg

ITALIA-SLOVENIJA



TRAIN



UNIONE EUROPEA
EVROPSKA UNIJA



Weather prediction

- STC: Predicting the outlook (sunny, overcast, rain)
- STR: Predicting the temperature (in degrees Celsius)
- MTP: Predicting the weather
 - Outlook
 - Temperature
 - Humidity
 - Quantity of precipitation ...



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
...

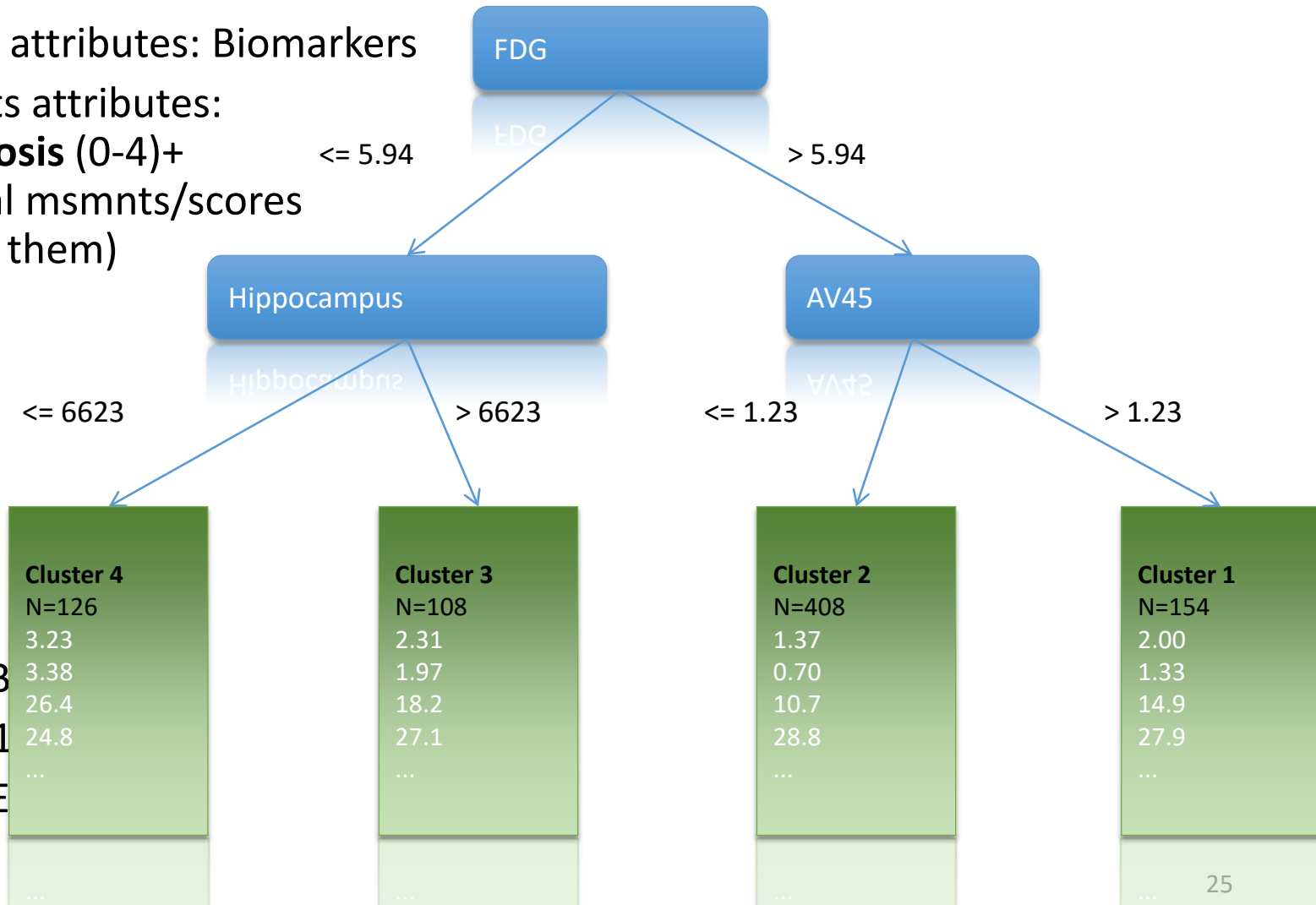


Example MTR task: Target vars.; Clinical scores for Alzheimer's

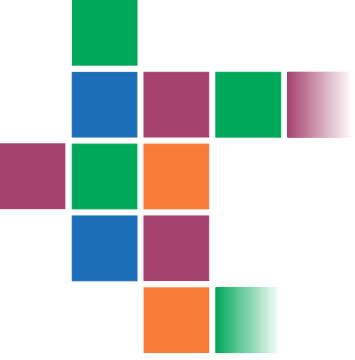
1. CDRSB – Clinical Dementia Rating Sum of Boxes
2. ADAS13 – AD assessment scale
3. MMSE – Mini Mental State Examination
4. RAVLT (immediate, learning, forgetting, perc. forgetting) – Rey Auditory Verbal Learning Test (4 features)
5. FAQ – Functional Assessment Questionnaire
6. MOCA – Montreal Cognitive Assessment
7. Ecog**Pt** (Memory, Language, Visuospatial Abilities, Planning, Organization, Divided Attention, Total score) – Everyday cognition questionnaire – filled in by patient (7 features)
8. Ecog**SP** (Memory, Language, Visuospatial Abilities, Planning, Organization, Divided Attention, Total score) – Everyday cognition questionnaire – filled in by study partner (7 features)

Example MTR model

- Descr. attributes: Biomarkers
- Targets attributes: **diagnosis (0-4)+ clinical msmnts/scores (23 of them)**



- DX
- CDRSB
- ADAS1
- MMSE
- ...



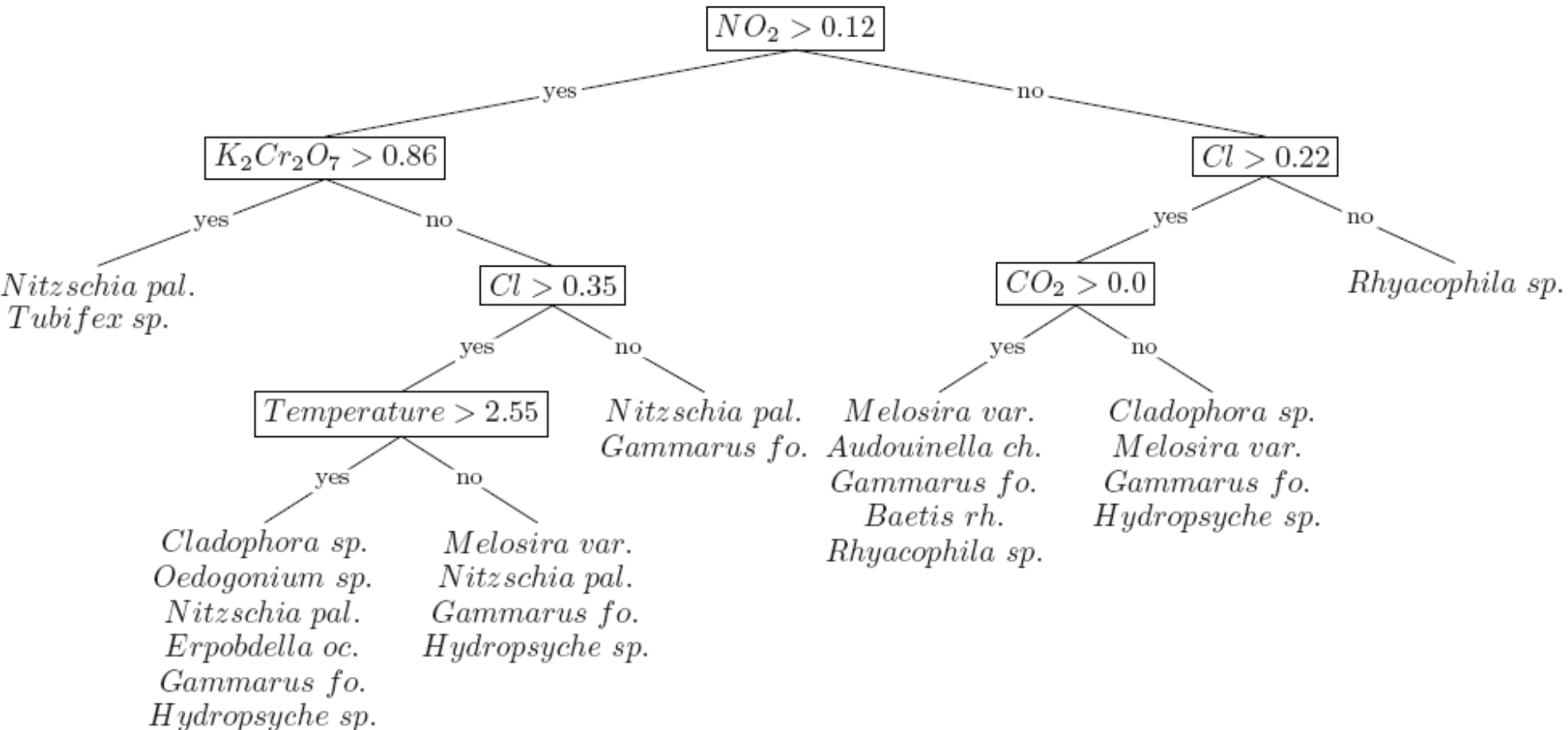
Multi-Target Classification & Multi-Label Classification

- Learning models that simultaneously predict several nominal/**binary** target variables
- Input: A vector of descriptive variables
- Output: A vector of several nominal/**binary** targets

Sample ID	Descriptive variables						Target variables														
	Temperature	K ₂ Cr ₂ O ₇	NO ₂	Cl	CO ₂	...	<i>Cladophora sp.</i>	<i>Gongrosira incrustans</i>	<i>Oedogonium sp.</i>	<i>Stigeoclonium tenue</i>	<i>Melosira varians</i>	<i>Nitzschia palea</i>	<i>Audouinella chalybea</i>	<i>Erpobdella octoculata</i>	<i>Gammarus fossarum</i>	<i>Baetis rhodani</i>	<i>Hydropsyche sp.</i>	<i>Rhyacophila sp.</i>	<i>Simulim sp.</i>	<i>Tubifex sp.</i>	
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1	1

Multi-Label Classification Example

- A decision tree for multi-label classification





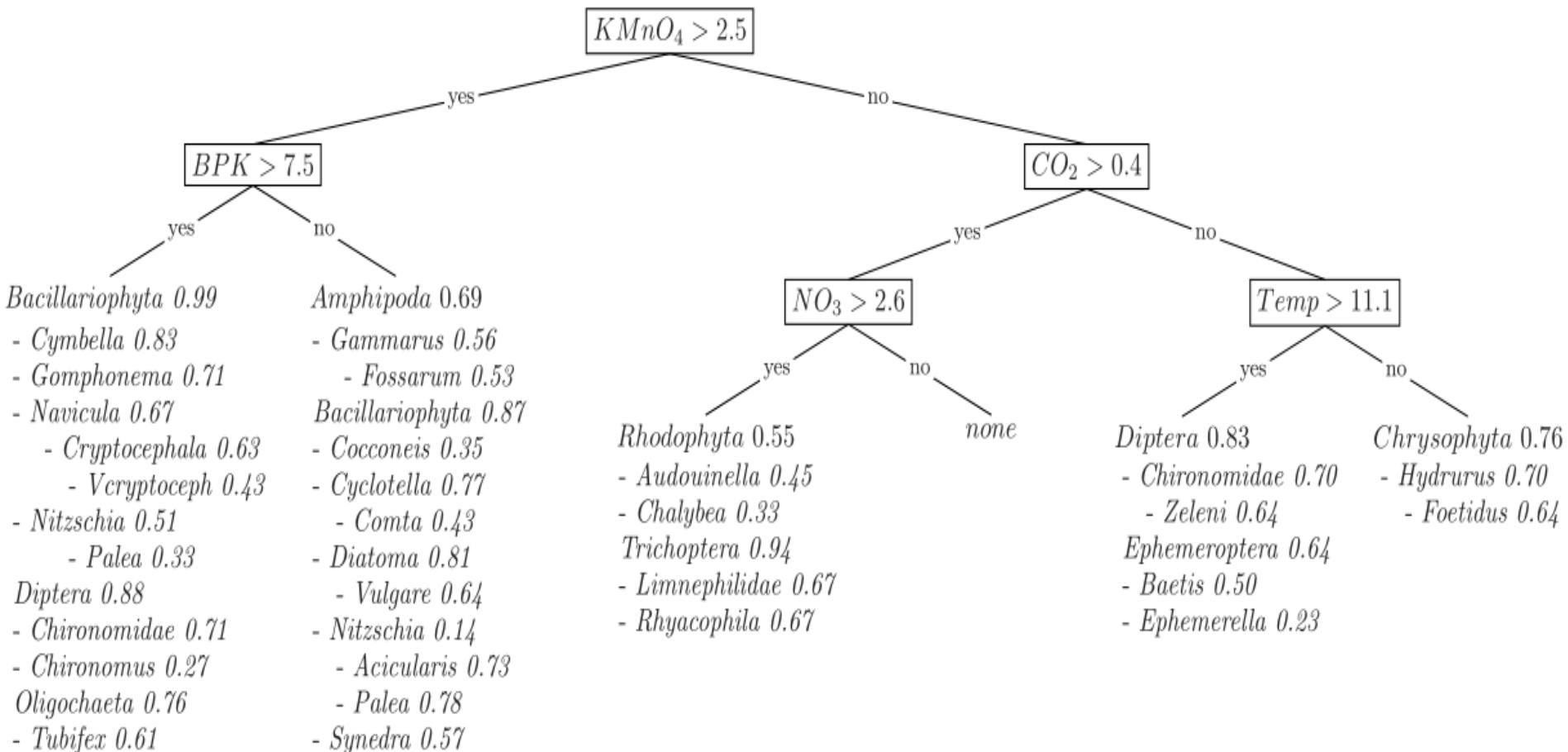
Hierarchical multi-label classification

	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)



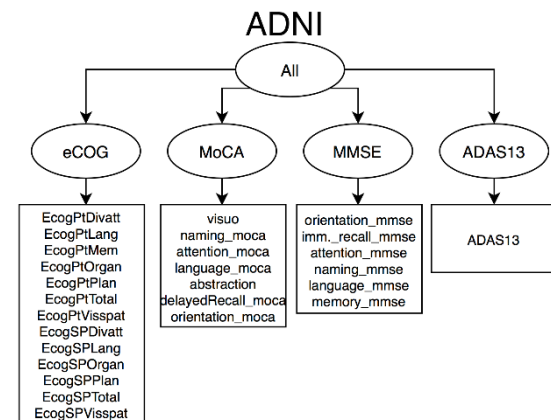


Hierarchical multi-target regression

In HMLC, the binary targets are organized into a hierarchy (tree or DAG)

In HMTR, the continuous targets are organized into a hierarchy

- The target space for the ADNI dataset has a hierarchical structure: The clinical scores are hierarchically organized
- MMSE – Mini Mental State Examination (orientation, immediate recall, attention, naming, language, memory)
- MOCA – Montreal Cognitive Assessment (visuo, naming, attention, language, abstraction, delayed recall, orientation)





Mining Big and Complex Data: Combining Complexities

Interreg

ITALIA-SLOVENIJA



TRAIN



UNIONE EUROPEA
EVROPSKA UNIJA

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



SSL+SOP: Incomplete Annotations

Semi-supervised multi-target regression

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	?	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	?	?	?
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	?	?
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
...



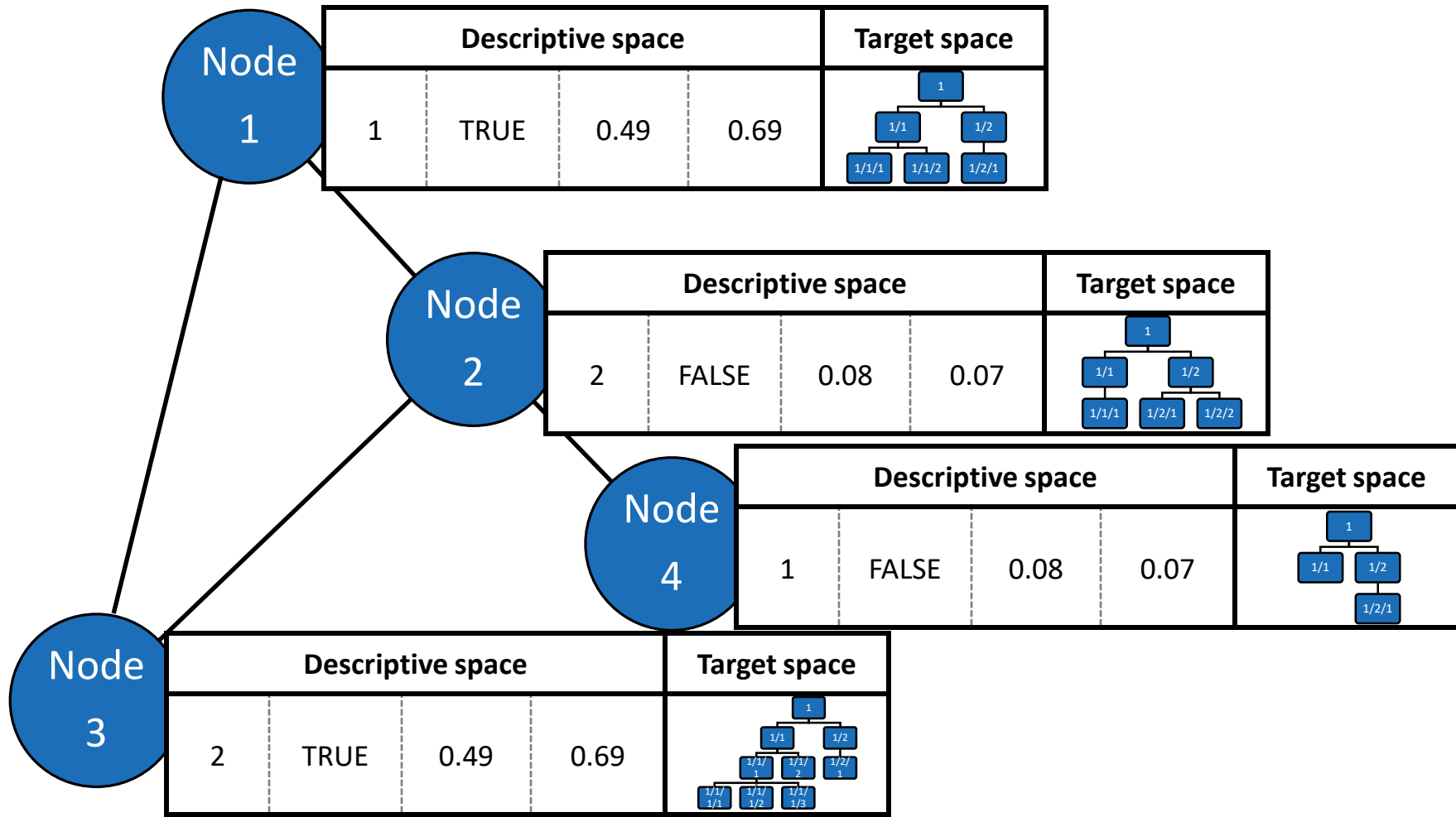
Data streams: (MT) Regression

	Descriptive space				Target space
...
Example n	1	TRUE	0.49	0.69	0.45
Example n+1	4	FALSE	0.08	0.07	0.12
Example n+2	6	FALSE	0.08	0.07	1.54
Example n+3	8	TRUE	0.00	1.00	3.12
Example n+4	6	TRUE	0.00	0.00	0.05
...

	Descriptive space				Target space		
...		
Example n	1	TRUE	0.49	0.69	0.58	0.09	3.99
Example n+1	4	FALSE	0.08	0.07	0.10	1.69	7.57
Example n+2	6	FALSE	0.08	0.07	0.08	0.77	8.86
Example n+3	8	TRUE	0.00	1.00	0.11	3.51	2.50
Example n+4	6	TRUE	0.00	0.00	0.43	2.10	8.09
...



Network +SOP: HMC





Predictive Clustering for Multi-Target Prediction

Interreg

ITALIA-SLOVENIJA



TRAIN



UNIONE EUROPEA
EVROPSKA UNIJA

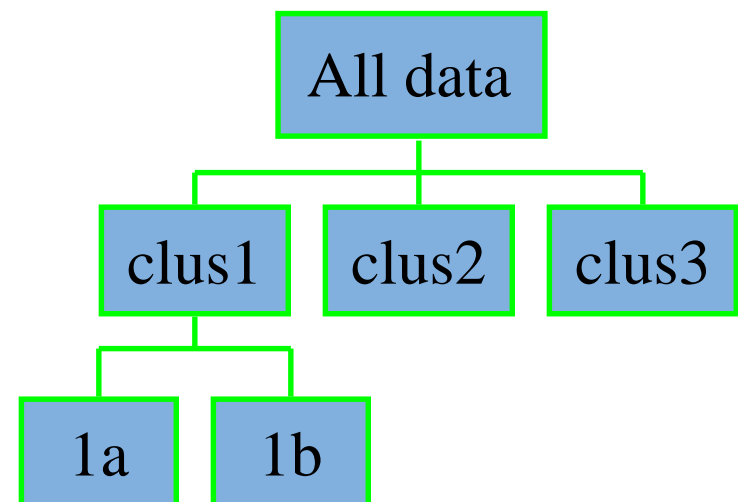
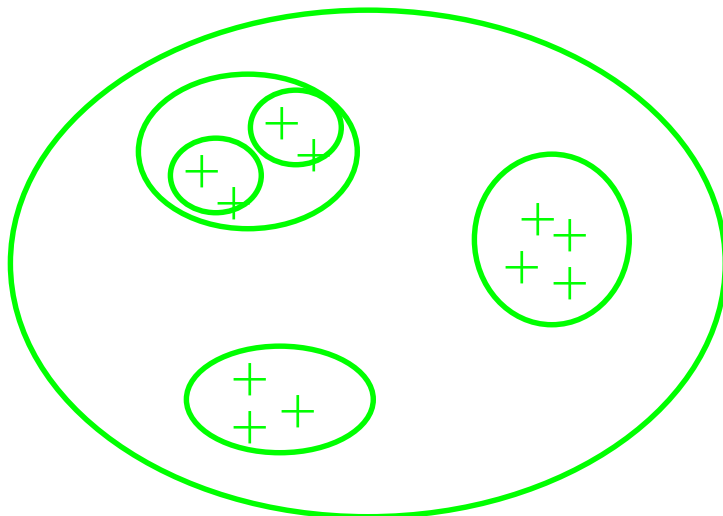
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



Clustering

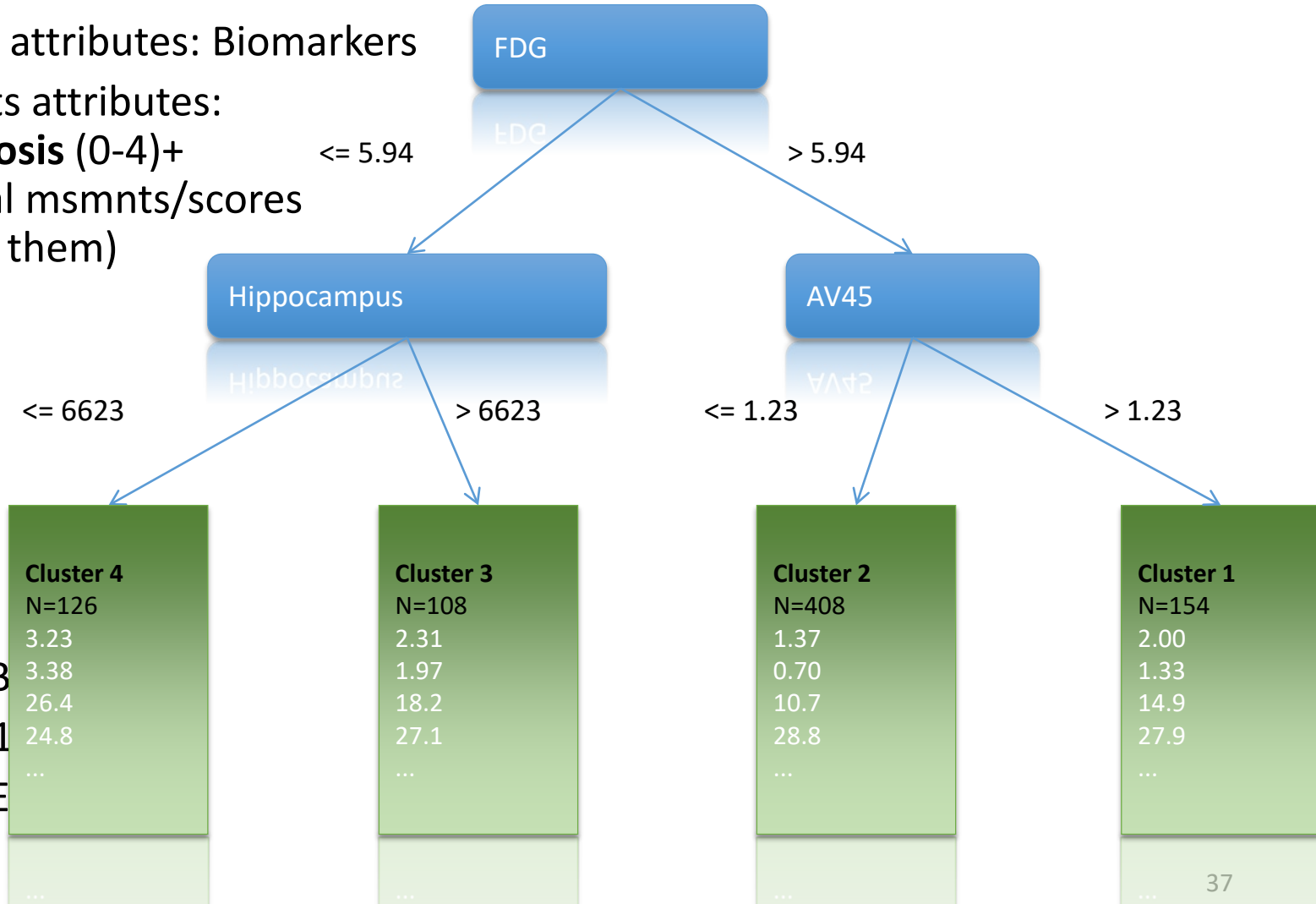
Partition a set of objects into clusters of similar objects

- High similarity of objects within individual clusters, low similarity between objects from different clusters
- Minimize intra-cluster variance (ICV)
- Distance/similarity measure in the example space



Example predictive clustering tree

- Descr. attributes: Biomarkers
- Targets attributes: **diagnosis (0-4)+ clinical msmnts/scores (23 of them)**



- DX
- CDRSB
- ADAS1
- MMSE
- ...



Top-down induction of PCTs

To construct a tree T from a training set S :

- If **the examples in S have low variance**,
construct a leaf labeled $target(prototype(S))$
- Otherwise:
 - Select the best attribute A with values v_1, \dots, v_n ,
which **reduces the most the variance** (*measured according to a given distance function d*)
 - Partition S into S_1, \dots, S_n according to A
 - Recursively construct subtrees T_1 to T_n for S_1 to S_n
 - Result: a tree with root A and subtrees T_1, \dots, T_n



Learning PCTs

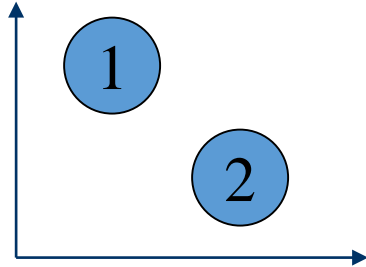
- Recursively partition data set into subsets (clusters) with low intra-cluster variance
 - Variance = avg. squared distance to prototype

$$ICV(S) = \sum_{y_j \in S} d(y_j, p(S))^2$$

- For the variance, the distance is measured
 - In standard clustering, along all dimensions
 - In prediction, along a single target dimension
 - In predictive clustering, along a structured target, e.g., several target dimensions

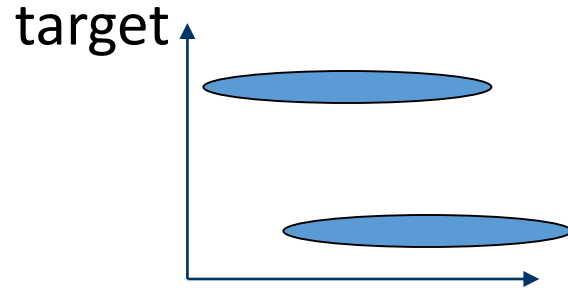
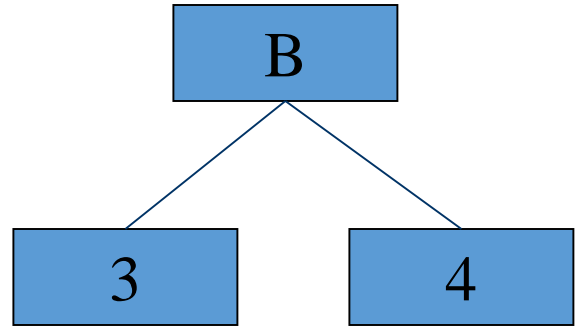


Clustering:

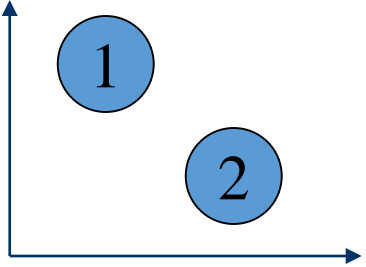
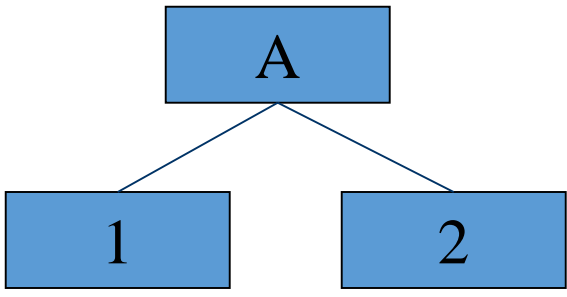


Data divided into clusters 1 and 2 coherent along two dimensions

Prediction:



B divides data into clusters coherent along single *target*



Predictive clustering: A divides data into clusters 1 and 2 coherent along two dimensions



Selecting the best test in a PCT

- Select the test that maximizes variance reduction
- Calculated in line 4

procedure BestTest(E)

1: $(t^*, h^*, \mathcal{P}^*) = (\text{none}, 0, \emptyset)$

2: **for each** possible test t **do**

3: $\mathcal{P} =$ partition induced by t on E

4: $h = \text{Var}(E) - \sum_{E_i \in \mathcal{P}} \frac{|E_i|}{|E|} \text{Var}(E_i)$

5: **if** $(h > h^*) \wedge \text{Acceptable}(t, \mathcal{P})$ **then**

6: $(t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$

7: **return** $(t^*, h^*, \mathcal{P}^*)$



Multi-target regression

- The variance function for MTR
- Is the sum of the variances
- Across all targets

$$\text{Var}(E) = \sum_{i=1}^T \text{Var}(Y_i).$$

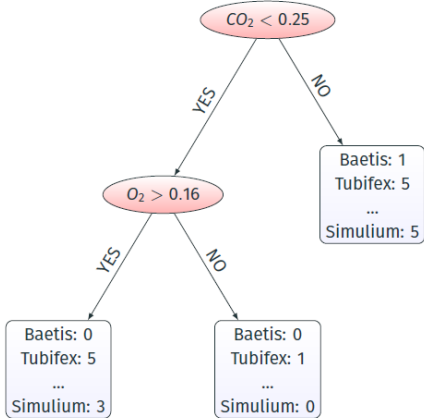
- Normalization is in order
- So that variances are comparable across targets

Ensembles of PCTs

- An ensemble is a set of predictive models, whose predictions are combined [to achieve performance better than that of individual/base predictors]

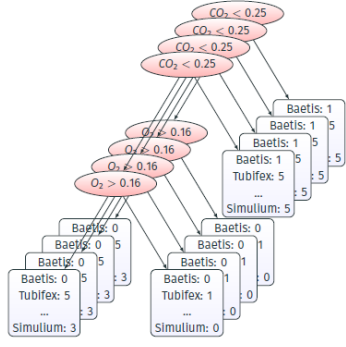
#	Descriptive attributes				Target attributes			
	KMnO ₄	CO ₂	...	K ₂ Cr ₂ O ₇	Baetis	Tubifex	...	Simulium
1	0.66	0.15	...	2.7	3	0	...	3
2	2.05	0.56	...	2.8	0	0	...	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1060	1.3	1.23	...	1.1	5	3	...	1

A single decision tree



#	Descriptive attributes				Target attributes			
	KMnO ₄	CO ₂	...	K ₂ Cr ₂ O ₇	Baetis	Tubifex	...	Simulium
1	0.66	0.15	...	2.7	3	0	...	3
2	2.05	0.56	...	2.8	0	0	...	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1060	1.3	1.23	...	1.1	5	3	...	1

An ensemble of decision trees





Relating the Environment and the Biota: From Habitat models to Community composition

Interreg

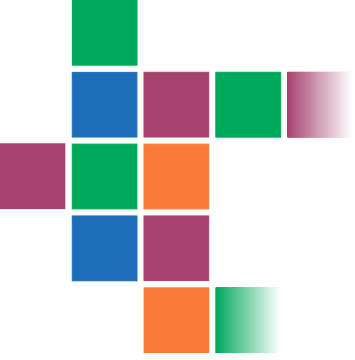
ITALIA-SLOVENIJA



TRAIN



UNIONE EUROPEA
EVROPSKA UNIJA



Environment <-> Biota

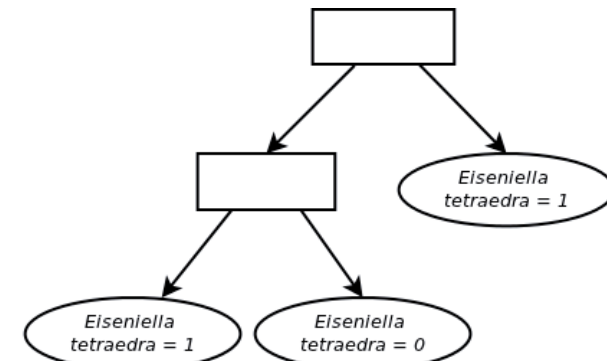
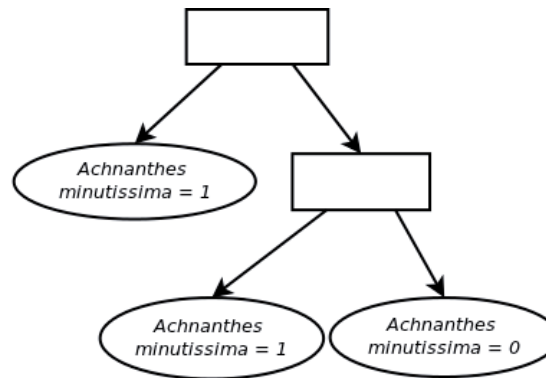
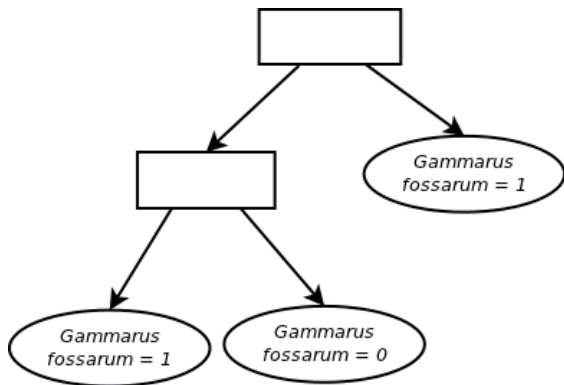
- Predict the biota (or specific components of it)
- At a given site
- From characteristics of the environment at the site
- E.g. predict river water biota from water properties

Sample ID	Descriptive variables						Target variables														
	Temperature	K ₂ Cr ₂ O ₇	NO ₂	Cl	CO ₂	...	<i>Cladophora sp.</i>	<i>Gongrosira incrustans</i>	<i>Oedogonium sp.</i>	<i>Stigeoclonium tenue</i>	<i>Melosira varians</i>	<i>Nitzschia palea</i>	<i>Audouinella chalybea</i>	<i>Erpobdella octoculata</i>	<i>Gammarus fossarum</i>	<i>Baetis rhodani</i>	<i>Hydropsyche sp.</i>	<i>Rhyacophila sp.</i>	<i>Simulim sp.</i>	<i>Tubifex sp.</i>	
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1	1



Habitat modeling

- Model the presence & absence (abundance) of each species separately

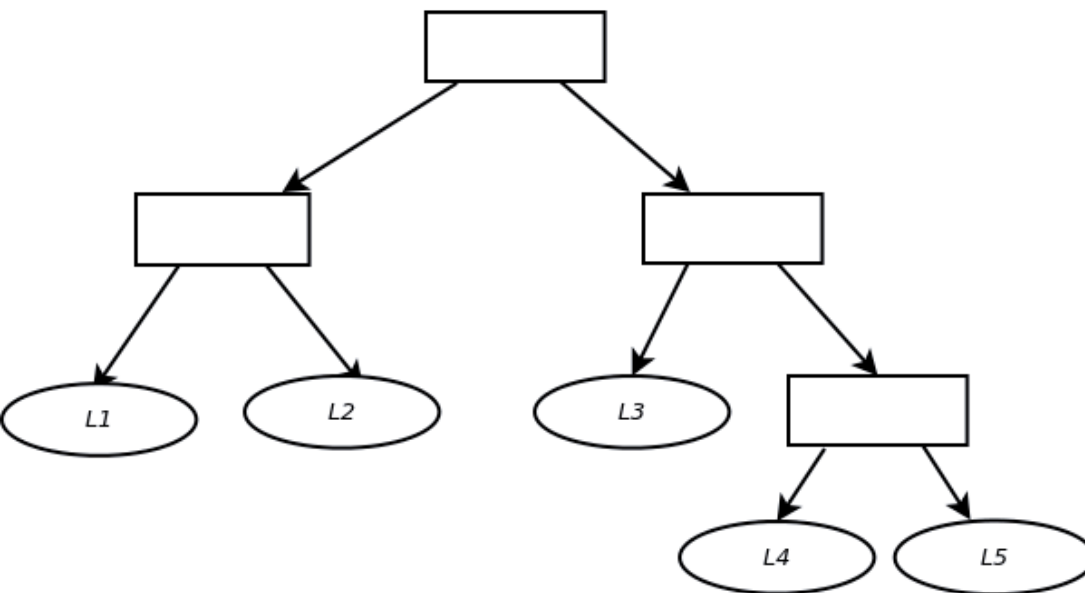


- Binary Classification (Regression)



Predicting species composition

- One model for **all the species at once**



L1:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

L2:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

L3:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

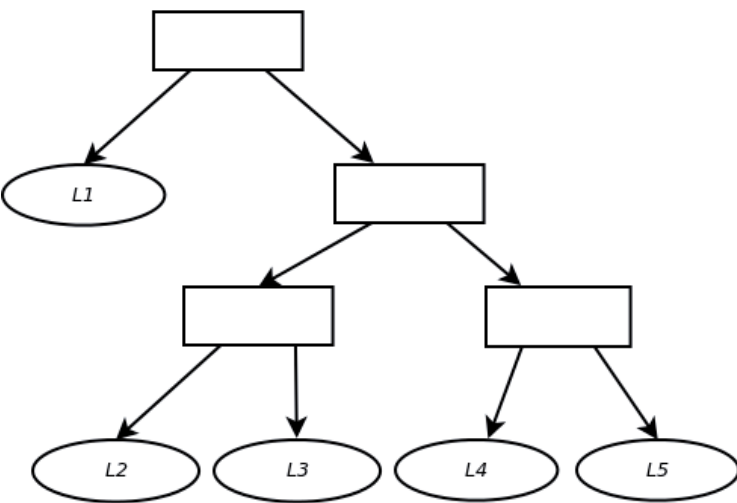
L4:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

L5:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

- **Multi-target classification/regression**

Predicting community structure

- One model for all of the species at once, additionally using the taxonomical hierarchy



L1:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 0

Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima : 1
Eiseniella : 0
Eiseniella tetraedra : 0

L3:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 0
Gammarus lacustris : 1

Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima : 1
Eiseniella : 0
Eiseniella tetraedra : 0

L5:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 1

Bacillariophyta : 1
Achnanthes : 0
Achnanthes minutissima : 0
Eiseniella : 1
Eiseniella tetraedra : 1

L2:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 1

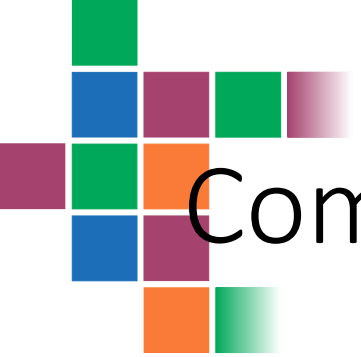
Bacillariophyta : 0
Achnanthes : 0
Achnanthes minutissima : 0
Eiseniella : 0
Eiseniella tetraedra : 0

L4:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 0

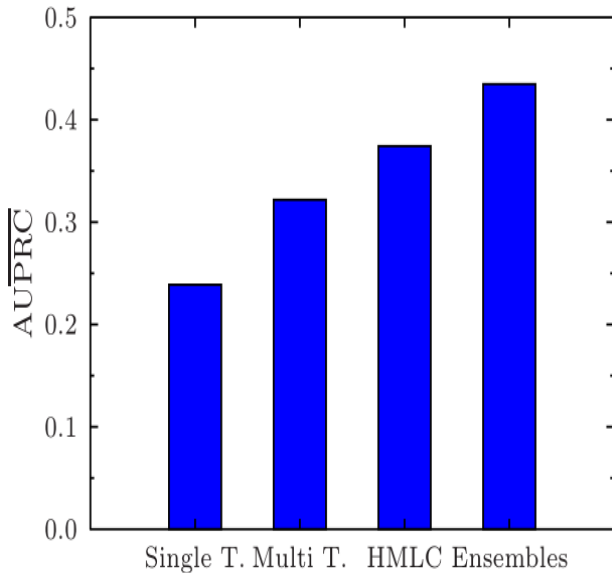
Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima : 1
Eiseniella : 1
Eiseniella tetraedra : 1

- Hierarchical multi-label classification

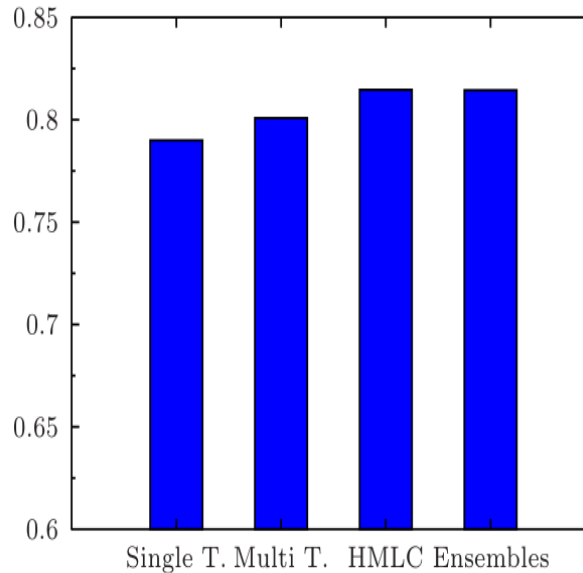


Community structure: Overall results

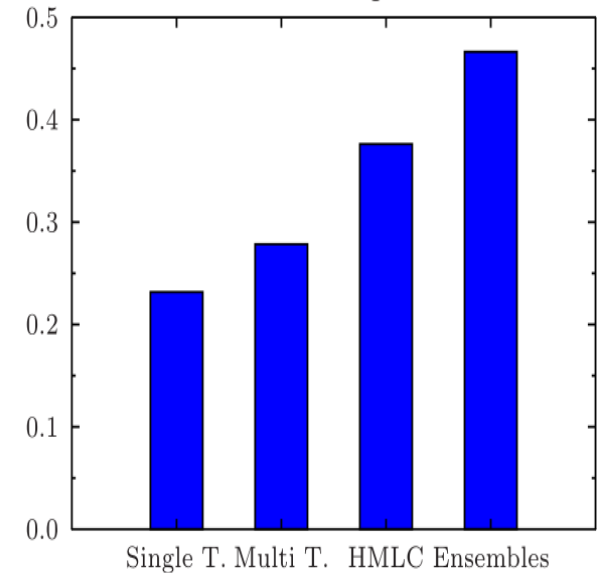
Slovenian rivers



Danish farms

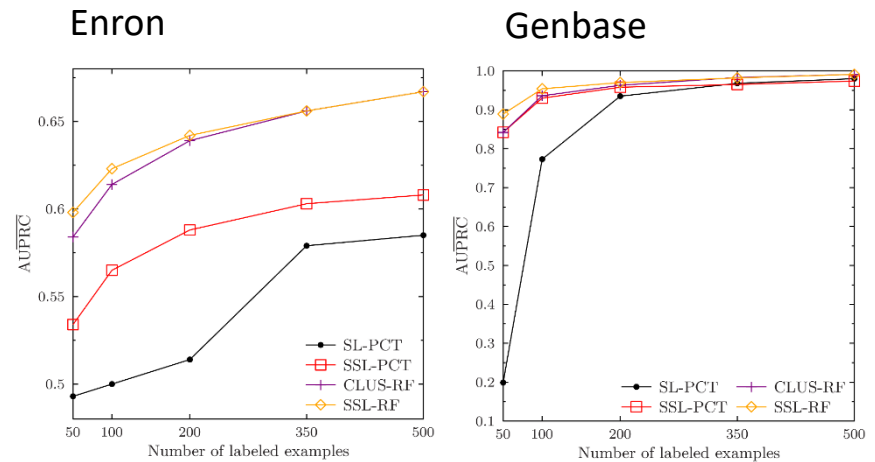


Australian vegetation

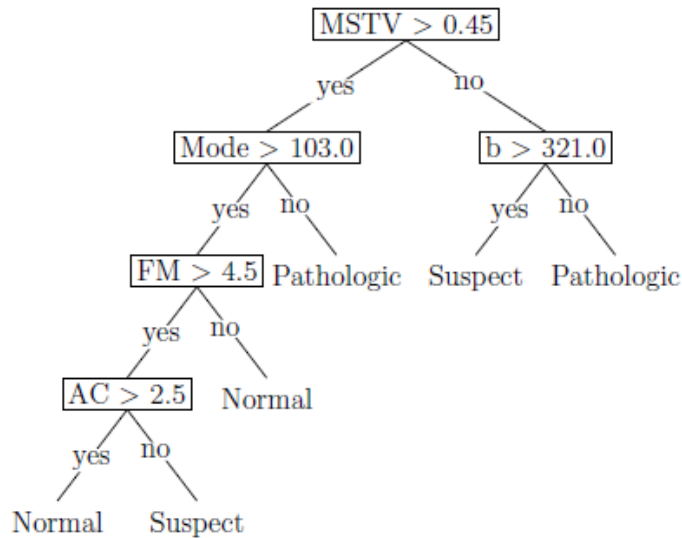


Dataset	Method	AUPRC	O_5	Learning time	Complexity
Slovenian rivers	Single-label	0.239	0.692	23.3	15,336
	HSC	0.309	0.591	10.2	25,035
	Multi-label	0.322	0.007	9.4	1
	HMC	0.374	0.132	0.6	37
Danish farms	Single-label	0.790	0.099	3.7	2605
	HSC	0.808	0.083	1.3	2873
	Multi-label	0.801	0.112	0.7	265
	HMC	0.815	0.065	0.4	259
Australian vegetation	Single-label	0.232	0.715	14,888.2	482,745
	HSC	0.306	0.591	76,023.2	648,970
	Multi-label	0.278	0.684	4639.5	23,699
	HMC	0.376	0.180	313.5	1279

SSL in MTP: Accuracy & interpretability

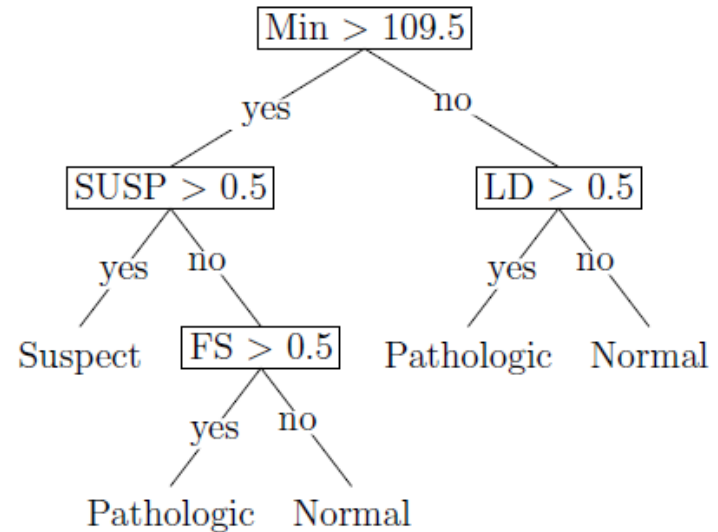


Multi-class classification (Cardiotocography3 Dataset)



Accuracy=81%, 11 nodes

(c) SL-PCT, 50 labeled examples



Accuracy=92%, 9 nodes

(d) SSL-PCT, 50 labeled and 2076 unlabeled examples



Multi-Target Prediction for Virtual Compound Screening

Interreg



UNIONE EUROPEA
EVROPSKA UNIJA

ITALIA-SLOVENIJA



TRAIN

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj



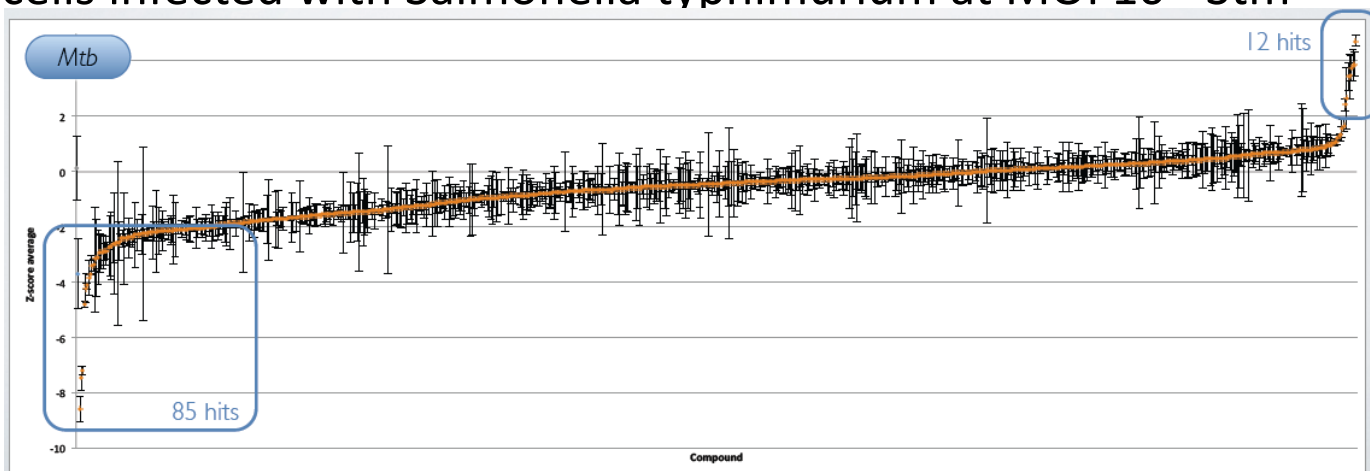
Virtual compound screening

- Descriptive variables refer to compound structure
 - Functional groups
 - Fingerprints
 - Bulk properties
- May also describe the compound in terms of the proteins it targets (e.g. from PubChem)
 - Their functional annotations
 - Pathways they are involved in
 - Proteins that the targets interact with (and/or their functional annotations, pathways they are involved in)
- Target variables describe compound activity and toxicity



Host-targeted Drugs for MTB (Tuberculosis) and STM (Salmonella)

- Library of compounds
 - LOPAC library - Library Of Pharmacologically Active Compounds
 - 1260 compounds
 - Well-characterized compounds, many already applied in clinical practice for a range of conditions
- Flow cytometry (FACS) - measured reduction in bacterial load
 - MeJuSo cells infected with Mycobacterium tuberculosis at MOI 10 – Mtb
 - HeLa cells infected with Salmonella typhimurium at MOI 10 - Stm





MTB&STM: Host-targeted Drugs

- Dataset
 - 964 compounds were found active on human protein targets
 - 711 distinct protein targets were identified
- Each compound is described with (the following features)
 - the respective protein targets
 - functional annotations of the respective protein targets
 - functional annotations of both the respective protein targets and the proteins they interact with
- Targets: bacterial load reduction, host cell viability
- Example antecedents in a rule from a tree:

IF compound targets the protein AAL 06595 THEN ...

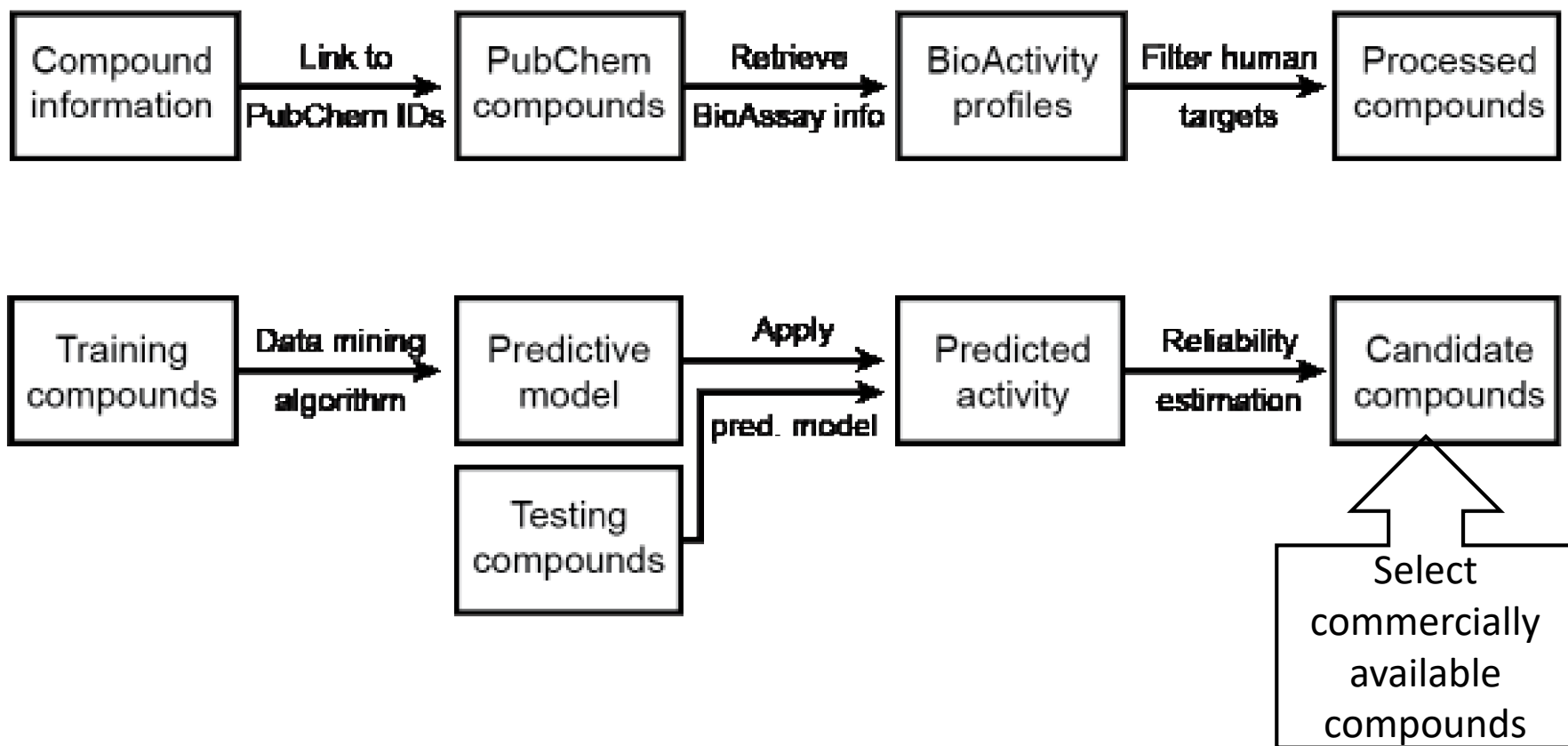
IF a protein with function GO0002637 is targeted THEN ...

THEN bacterial load effect = -5.269 & host cell effect = 0.0475



MTB&STM: Host-targeted Drugs

The Data Analysis Workflow





MTB&STM: Host-targeted Drugs Results

- Greatly increased proportions of hit compounds
 - 5 out of 9 (55.6%) for Mtb and
 - LOPAC primary screen (90 out of 1260 (7.1%) for *Mtb*
- The *in silico* predictive model successfully identified active compounds *de novo*

<i>Abbr.</i>	<i>Compound name</i>	<i>Alternative name(s)</i>	<i>Primary screen z-score</i>	<i>Rescreen z-score</i>	<i>Activity</i>
<i>Mycobacterium tuberculosis</i>					
SU	SU 6656	2,3-Dihydro-N,N-dimethyl-2-oxo-3-[(4,5,6,7-tetrahydro-1H-indol-2-yl)methylene]-1H-indole-5-sulfonamide	-5.79	-10.51	Src family kinase inhibitor
Q	Quinacrine dihydrochloride		-5.25	-9.90	MAO inhibitor
SB	SB 216763	3-(2,4-Dichlorophenyl)-4-(1-methyl-1H-indol-3-yl)-1H-pyrrole-2,5-dione	-6.02	-8.29	GSK-3 kinase inhibitor
G	GW5074	3-(3, 5-Dibromo-4-hydroxybenzylidene-5-iodo-1,3-dihydro-indol-2-one)	-4.86	-6.98	Raf1 kinase inhibitor
T494	Tyrphostin AG 494	N-Phenyl-3,4-dihydroxybenzylideneacyacetamide	-3.83	-6.93	EGFR kinase inhibitor
L	3',4'-Dichlorobenzamil hydrochloride	L-594,881	-3.87	-5.13	Na ⁺ /Ca ²⁺ exchanger inhibitor
H	Haloperidol		-3.77	-2.96	D2/D1 dopamine receptor antagonist



Analyzing data from High-contents Screens

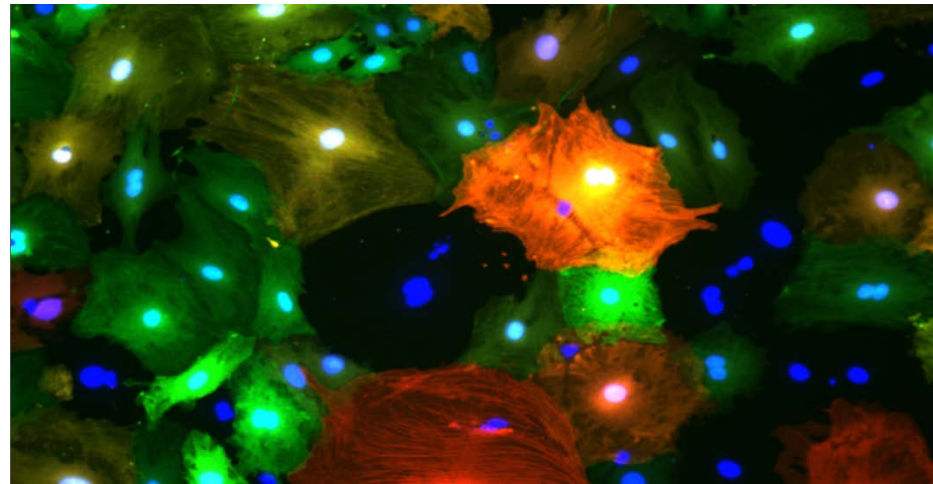
- Compounds described by fingerprints
- Generated by open-source chemoinformatics SW library RDkit
- The FCFP2 fingerprints were used (1024 features)
- Also considered profiles of targeted proteins
- These are the attributes

- Assays photographed under the microscope
- Features extracted from images
- These are then the targets



Reducing fibrosis in myocardial infarction

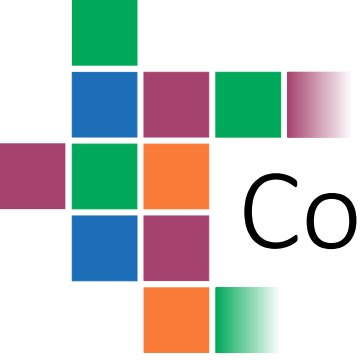
- High content screen using a library of 640 FDA approved drugs (ENZO)
- Identify drugs to reduce fibrosis in myocardial infarction
- Screen used murine cardiac fibroblasts which differentiate into myofibroblasts in culture, expressing increased alpha SMA-RFP and collagen-alpha1-EGFP
- Targets: Intensity of
 - alphaSMA
 - Collagen
- Attributes
 - Fingerprints





Testing the predictions

- Some domain-specific knowledge / constraints applied: Predicted compounds filtered for FDA approved drugs that are not corticosteroids
- SMILE strings used in Chemmine to identify substances with structural similarity to non commercial compounds with high predicted values
- Three related compounds identified which are described in literature to have an anti-fibrotic effect
- Four related compounds identified which were not previously described to have an anti-fibrotic effect
- Tested in the wet-lab and one works really well 😊



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