An Introduction to Mining Big and Complex Data

Sašo Džeroski
Jozef Stefan Institute, Ljubljana, Slovenia

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Mining Big and Complex Data

- Introduction: Just what is big and complex data?
  - Volume & Velocity (Data Streams)
  - Variety (Structured Inputs and Structured Outputs)
  - Other complexity dimensions (Incompleteness, Context)
- The different tasks of structured output prediction
- Combination with other complexities
  - Semi-supervised
  - SOP on data streams
- Structured output prediction with predictive clustering

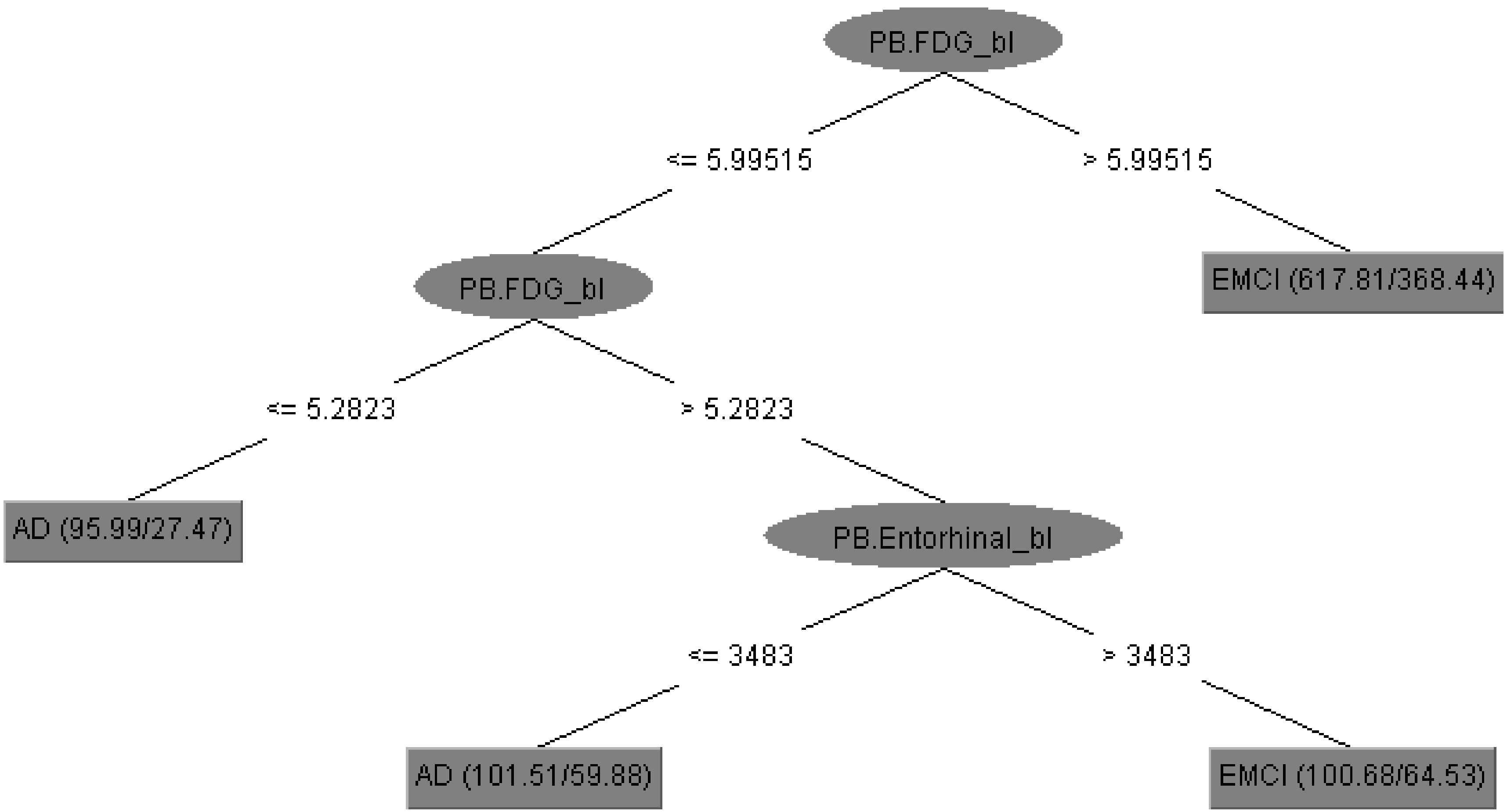
Data mining: Predictive modelling

- 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



Predictive modeling: 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
…	…				…

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**

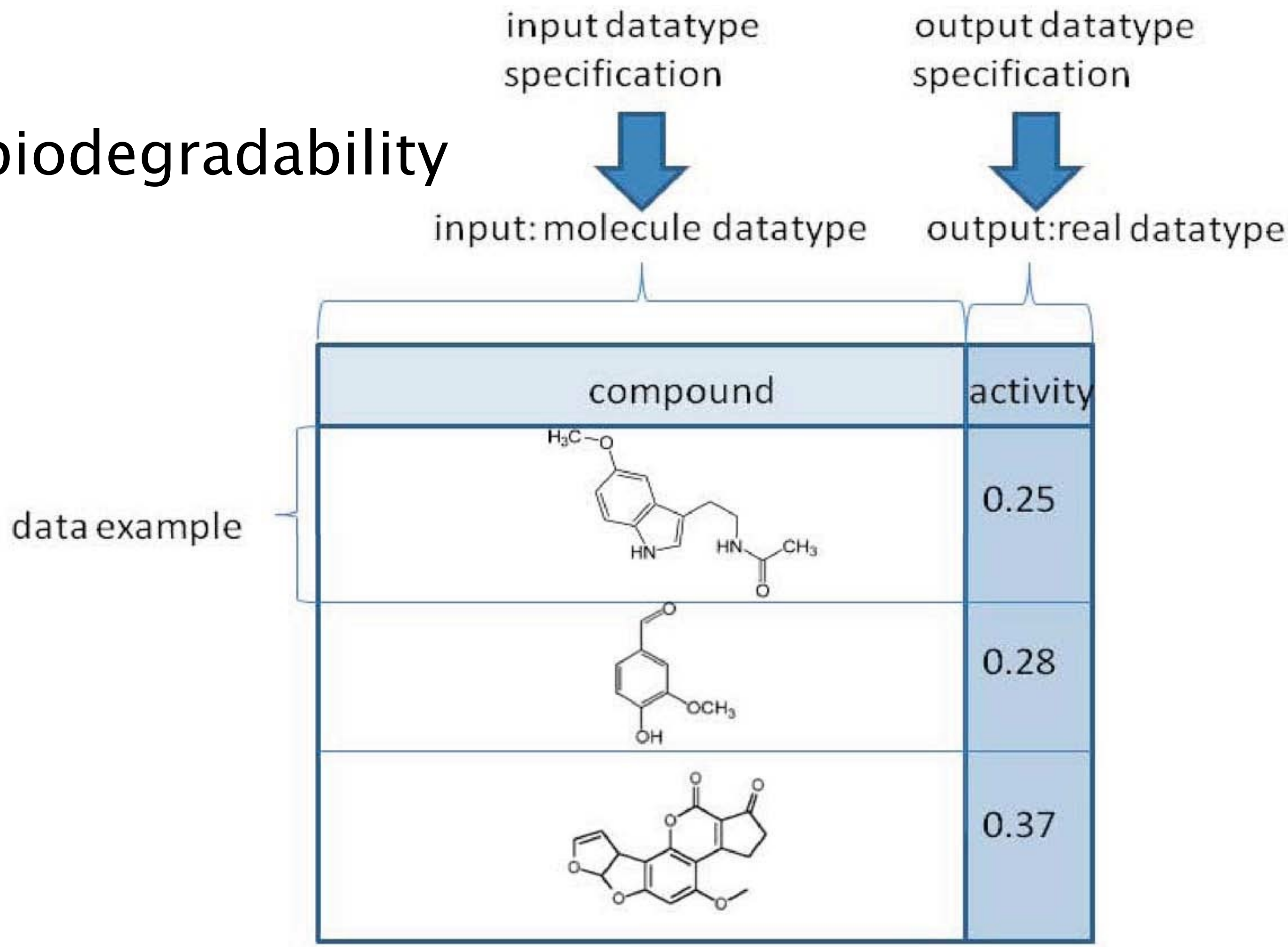
Data streams: Regression

	Descriptive space				Target space
…	…				…
Example n+5	**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

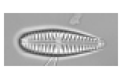
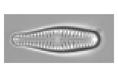
Example:

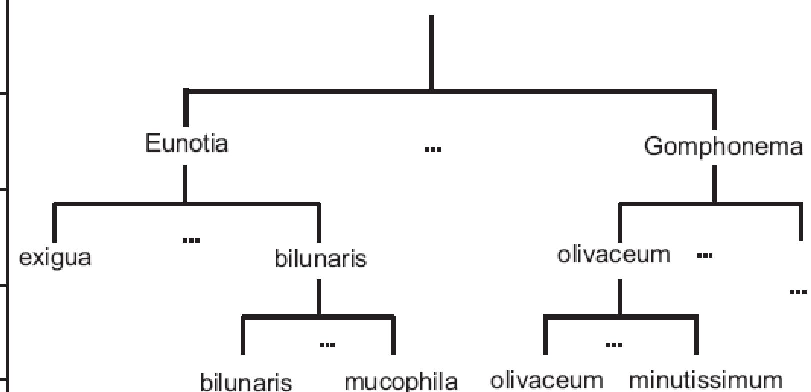
Predicting biodegradability



Big Data: Variety - Structured Output

- Hierarchical classification
- Taxonomic classification of diatoms
- From microscopic images
- Taking into account the taxonomy of diatoms

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
…	…	…	…	…	…	…	…



Structured-output prediction

- Multi-target prediction
  - Classification
  - Regression
  - Mixed
- Multi-label classification
  - Hierarchical multi-label classification
- Predicting (short) time series

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

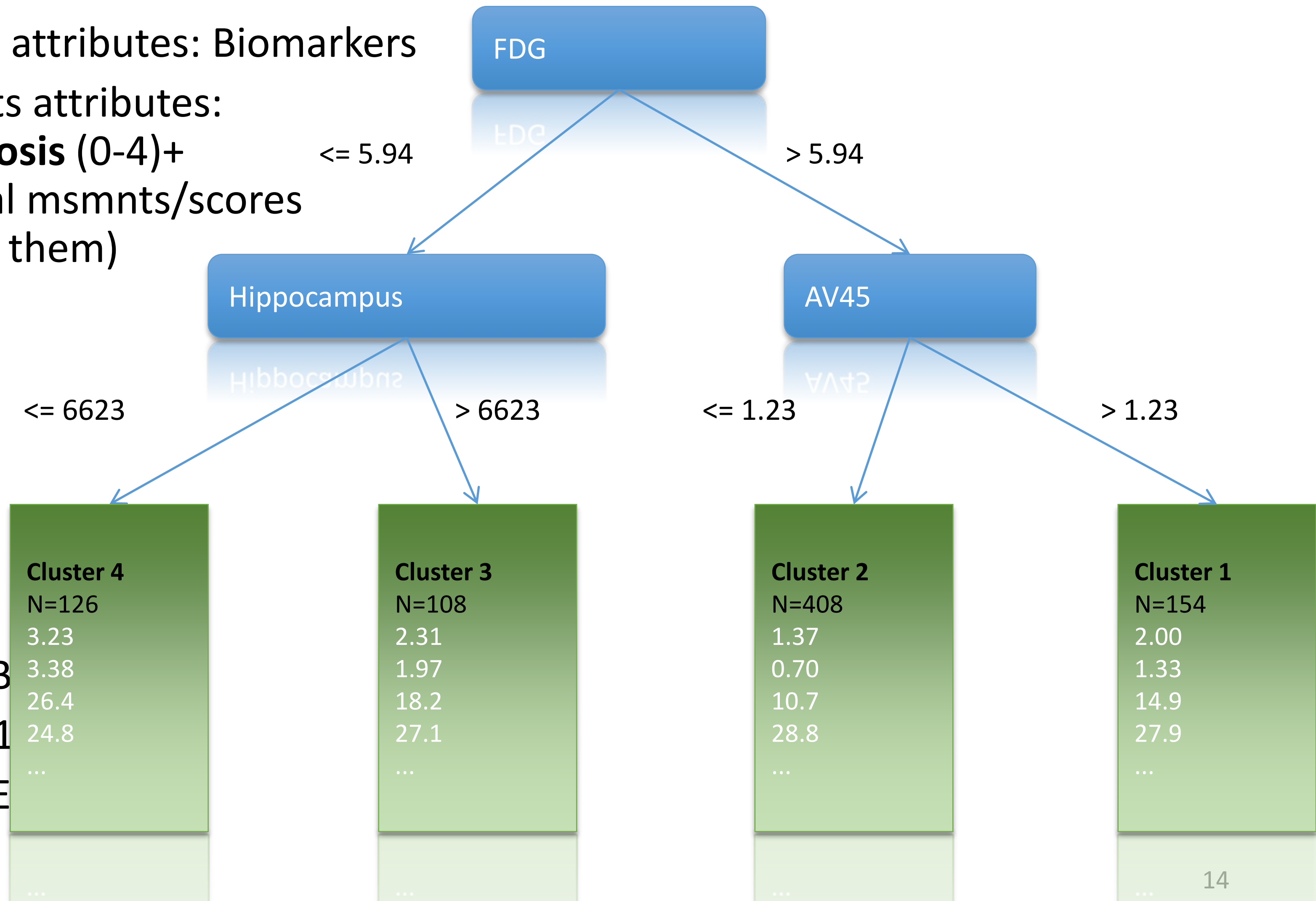
	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…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

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 parter (7 features)

Example MTR model

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



- DX
- CDRSB
- ADAS1
- MMSE
- …

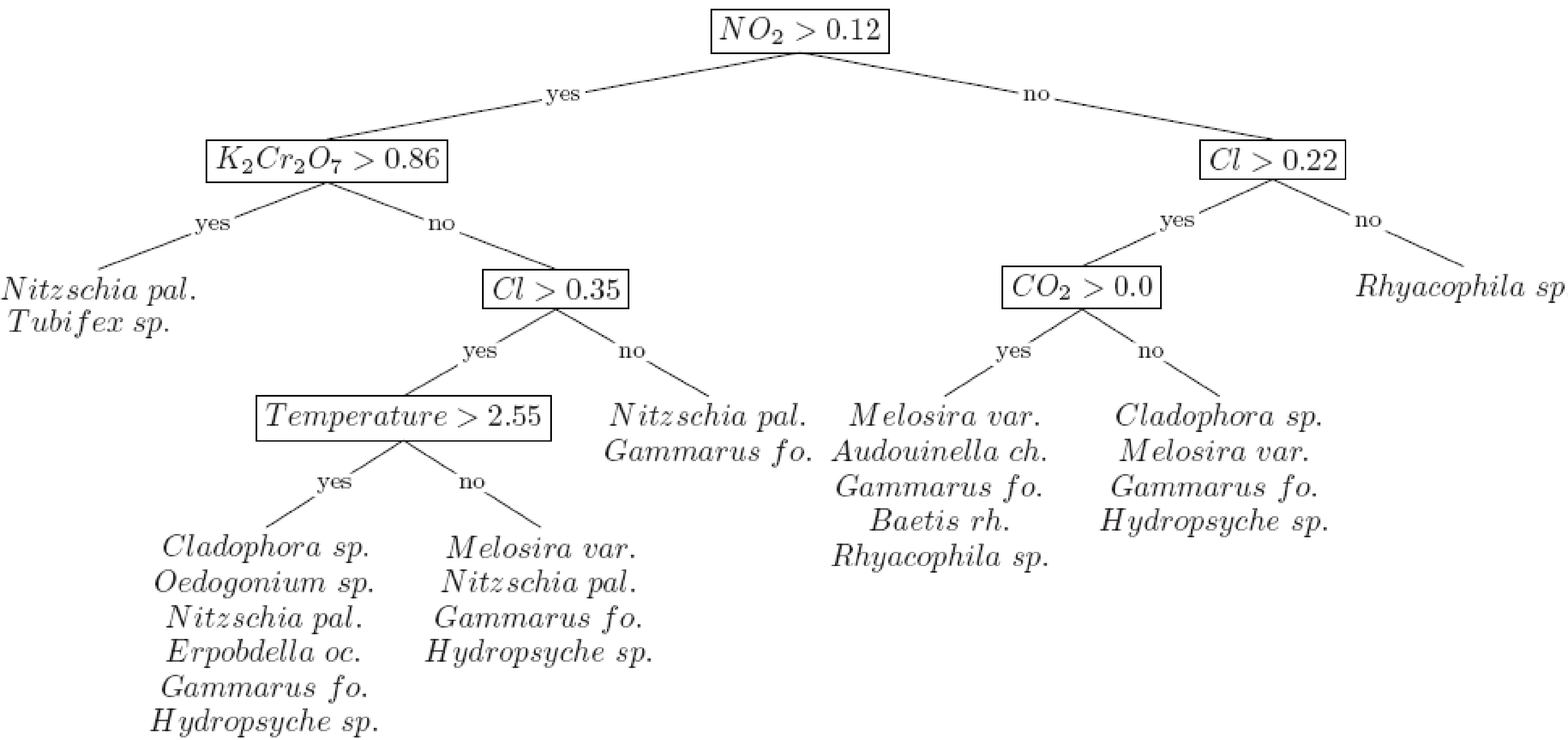
Multi-Target Classification & Multi-Label Classification

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

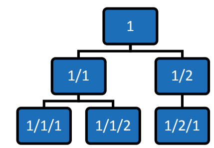
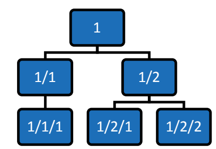
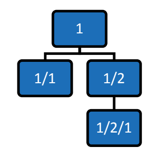
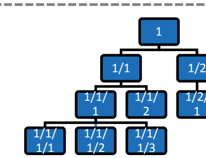
	Descriptive variables						Target variables													
Sample ID	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2	...	*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex sp.*
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
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	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

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 classification: An example

- Gene function prediction
- Input: Tuple of primitives
- Output: A subhierarchy of a hierarchical catalog of gene functions, such as FunCat

Descriptive attributes: $[0.36, -0.49, -0.1, 0.21, -0.34, -0.27, -0.06, \ldots]$

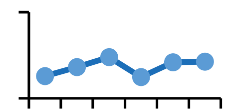
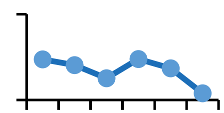
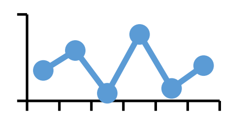
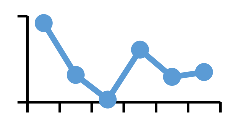
Target hierarchy subset:

metabolism
C-compound and carbohydrate metabolism
regulation of C-compound and carbohydrate metabolism

transcription
RNA synthesis
mRNA synthesis
transcriptional control
transcription activation

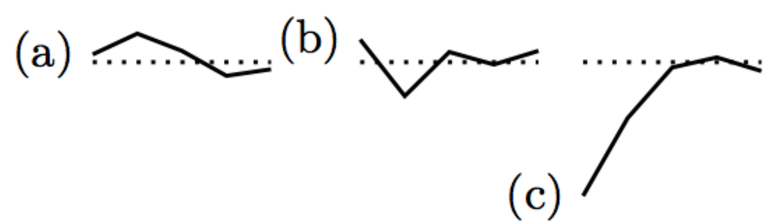
Fig. 1: An example task of HMLC: a single instance from the *cellcycle* dataset (Section 3) is shown, corresponding to one gene. The descriptive attributes are gene properties, the targets are gene functions from the FunCat hierachy.

Time-series prediction

	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	
…	…				…

Predicting short time series

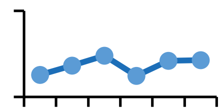
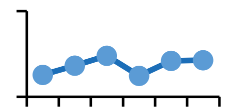
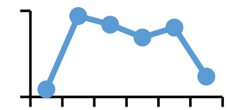
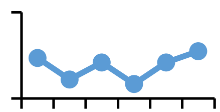
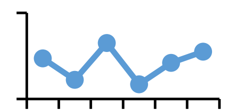
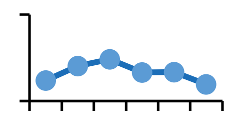
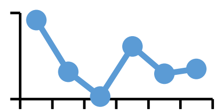
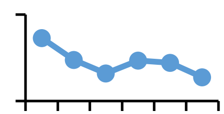
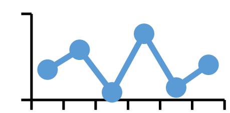
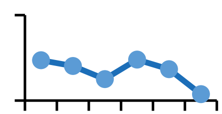
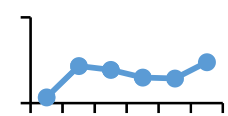
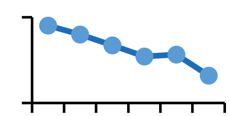
Table 2: An example task of predicting short time series. Three instances (genes) are shown: The descriptive attributes are gene functions, the target is a (short) time series of gene expression values in yeast responding to environmental stress (amino acid starvation in this case).

Descriptive attributes										Target time series
GO 0000282	GO 0000287	GO 0000315	GO 0000322	GO 0000781	GO 0000785	GO 0000790	GO 0000819	GO 0080090		
0	0	0	0	1	1	1	0	1	...	$[0.13, 0.48, 0.19, -0.23, -0.12]$ **(a)**
1	1	1	1	1	0	0	0	0	...	$[0.38, -0.57, 0.17, -0.04, 0.19]$ **(b)**
0	0	0	0	1	1	0	0	0	...	$[-2.25, -0.94, -0.09, 0.08, -0.15]$ **(c)**
...										

Even more complex SOs

- Mixed tuples (diff. arg. of tuple are of diff. types, e.g., a mix of discrete and real-valued targets)
- Besides tuples, sets & sequences of primitive values, also tuples, sets and sequences of structures (e.g., of the previously mentioned types of SOs
  - Tuples of hierarchies (The Gene Ontology has three hierarchies: BF, MP, Cellular Component)
  - Tuples of time series
  - Sets of tuples
  - Sequences of tuples
- …

Predicting tuples of time-series

	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			
...	...				...		

The other complexity aspects

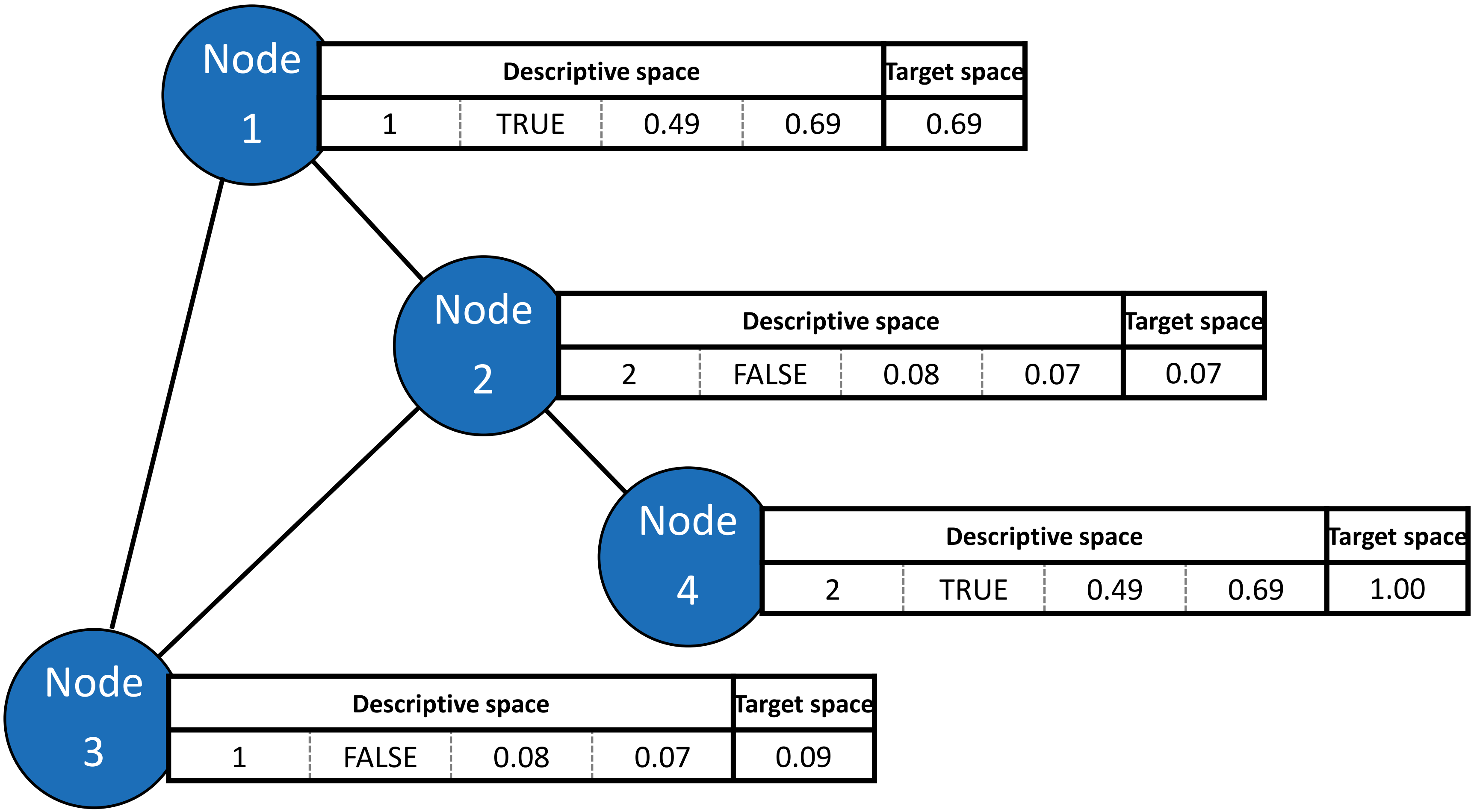
- Incomplete annotations
- Network context

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
…	…				…

Network regression



Motivation for MAESTRA

Each of the individual complexity aspects above **presents a major challenge** to current ML/DM methods

Most approaches to

- Structured output prediction (e.g., multi-label learning)
- Mining data streams (e.g., VFDT)
- Semi-supervised learning (e.g., co-training)
- Learning in a network context (e.g. collective classification)

Consider each of the complexity dimensions individually

Combining complexity dimensions

Simultaneous presence of several complexity aspects is a much harder challenge and is not addressed appropriately by current approaches

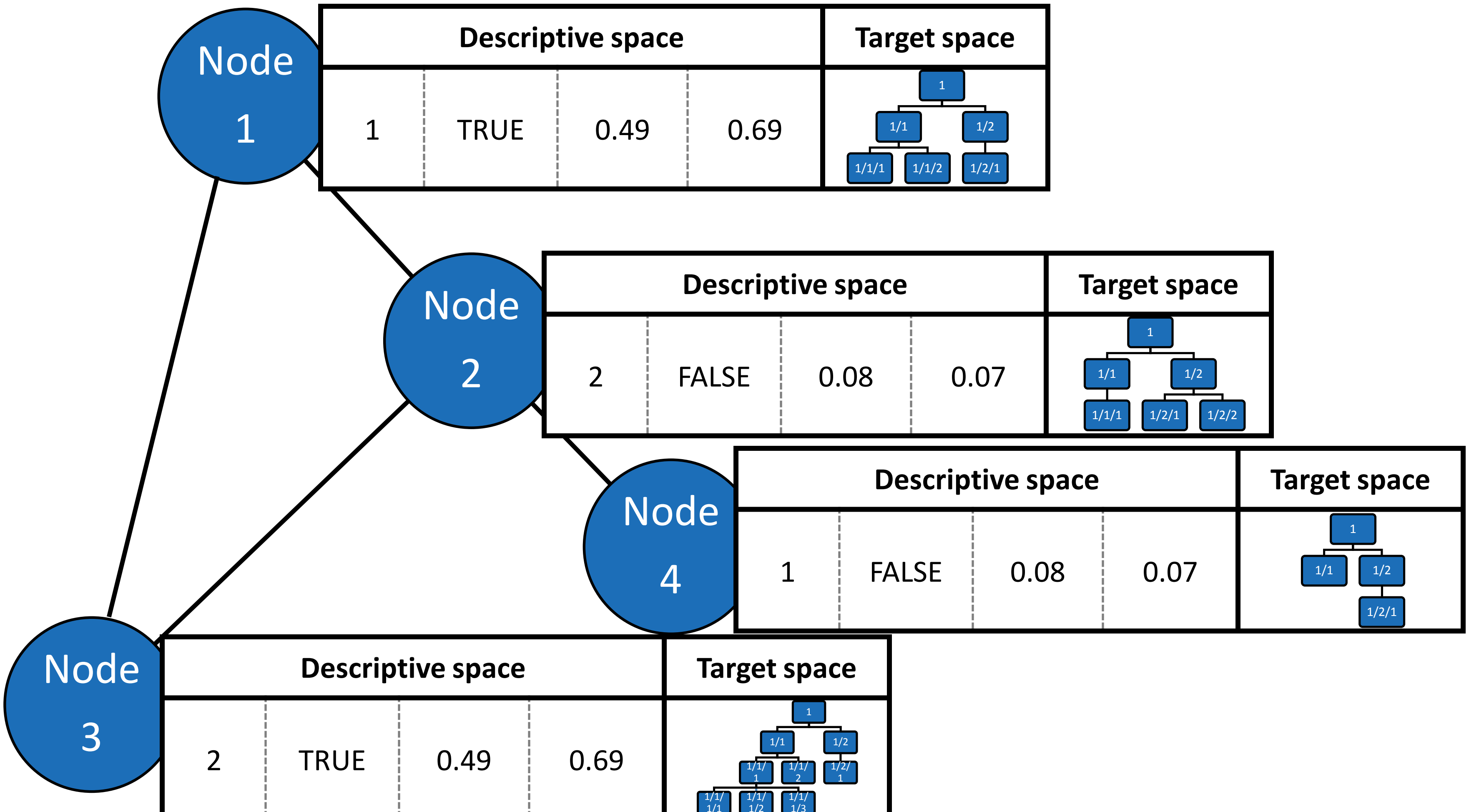
SOP [for different structured outputs] in all cases

- SOP + SSL (Semi-supervised structured-output prediction)
- SOP + Network Data
- SOP + Data Streams
- SOP + SSL + Data Streams
- …
- SOP + SSL + Data Streams + (Dynamic) Network Data

SSL+SOP: 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
…	…				…	…	…

Network +SOP: HMC



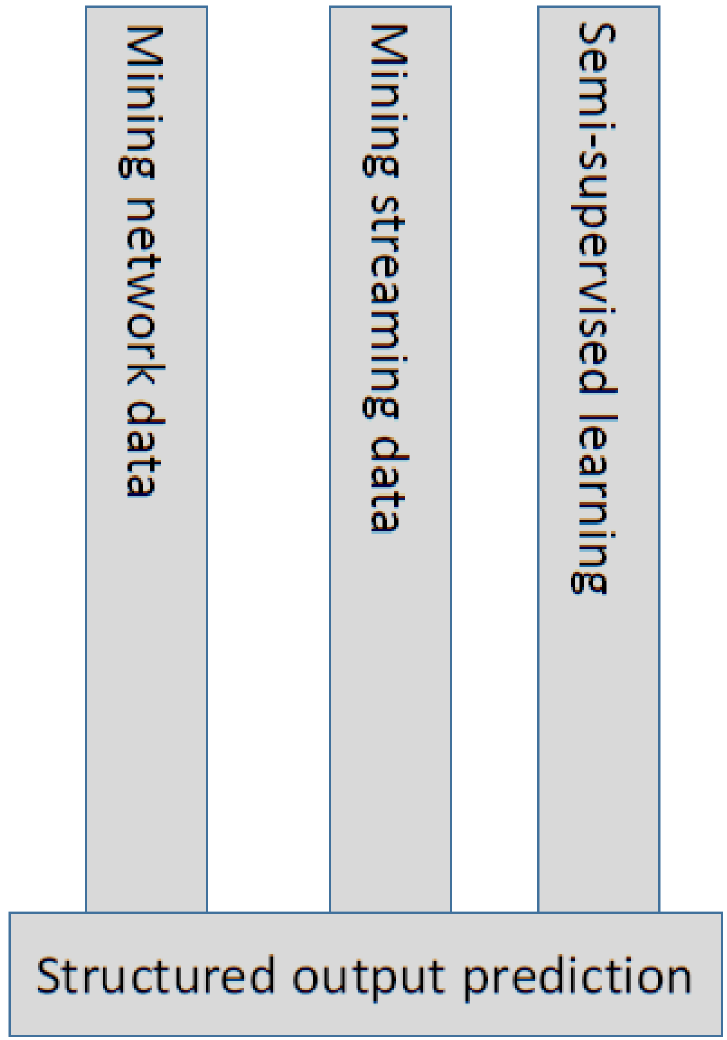
The MAESTRA project: Goals

Develop predictive modelling methods capable of **simultaneously addressing** several (and ultimately **all**) **of the complexity aspects outlined above**: Methods that can handle **massive** sets of **network** data **incompletely annotated** with **structured outputs**

Develop **the foundations** (basic concepts/notions) and **the methodology** (design/implement algorithms) necessary

Demonstrate the potential and utility of the developed approaches on showcase problems

The MAESTRA pillars



MAESTRA Applications

- Life sciences / health
  - Fungal microbiology
  - Predicting gene function

- Sensor networks (smart grid, energy production)
- Social networks (e.g., sentiment analysis/Twitter)

- Multimedia
  - Image annotation
  - Image retrieval

A central approach in MAESTRA (but not the only one :-)

- Learning tree and rule-based models in the context of predictive clustering, which unifies the tasks of **predictive modelling** and **clustering**
- Predictive clustering (PC) allows for
  - **Handling different types of structured outputs**
  - Efficient learning of trees, rules and ensembles thereof
- We are extending PC to consider semi-supervised learning; network context; and to learn from data streams
- We are considering different combinations of complexity aspects (e.g., SOP+data streams+SSL) in this context

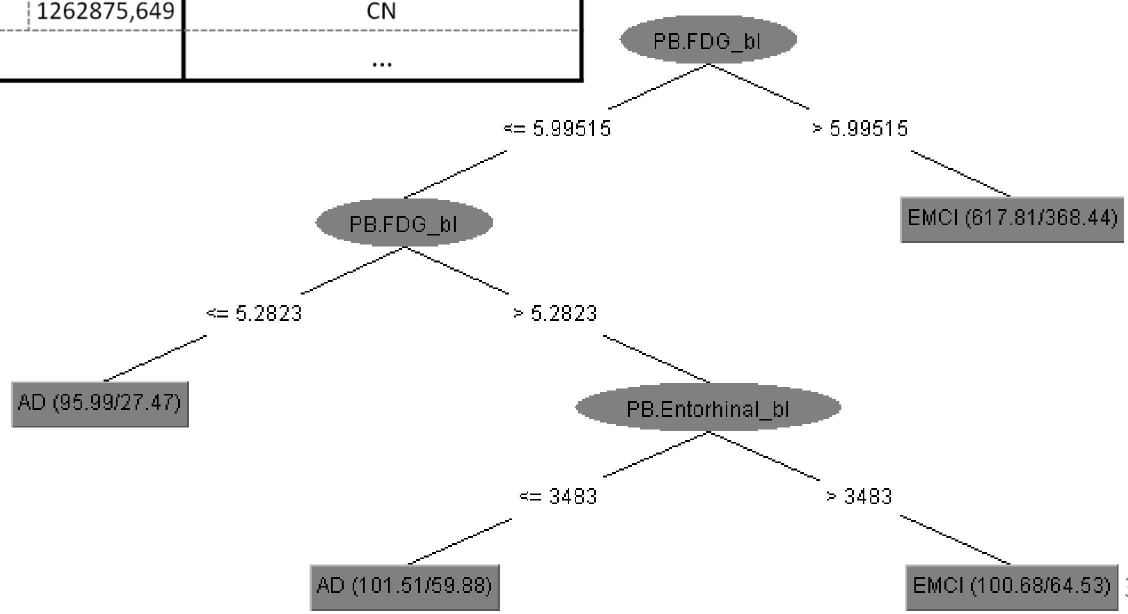
Predictive Clustering for Predicting Structured Os

Predictive modeling

- Input: A table of data, a row is an object, single target

	Descriptive space				Target space
	Gender	Fusiform	Hippocampus	ICV	
Example 1	F	16471	6350	1445040,208	SA, AD
Example 2	M	20680	7440	1610298,246	CN
Example 3	F	18751	6615	1257475,402	CN
Example 4	M	22895	9311	1755672,837	SA, LMCI
Example 5	F	18446	6544	1527253,171	SA, LMCI
Example 6	F	16056	6869	1262875,649	CN
...	...				...

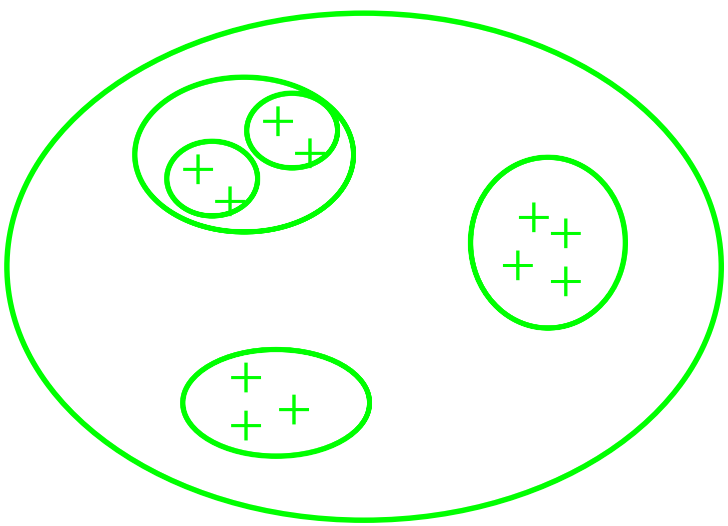
- Output: A predictive model for the target

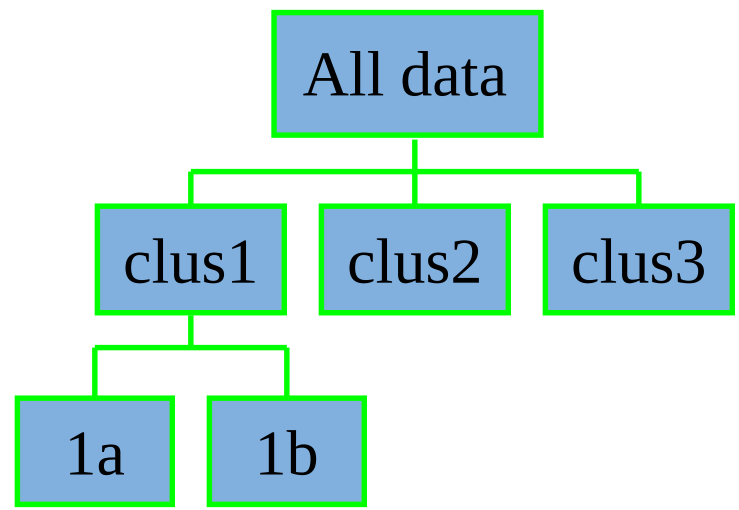


Clustering

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



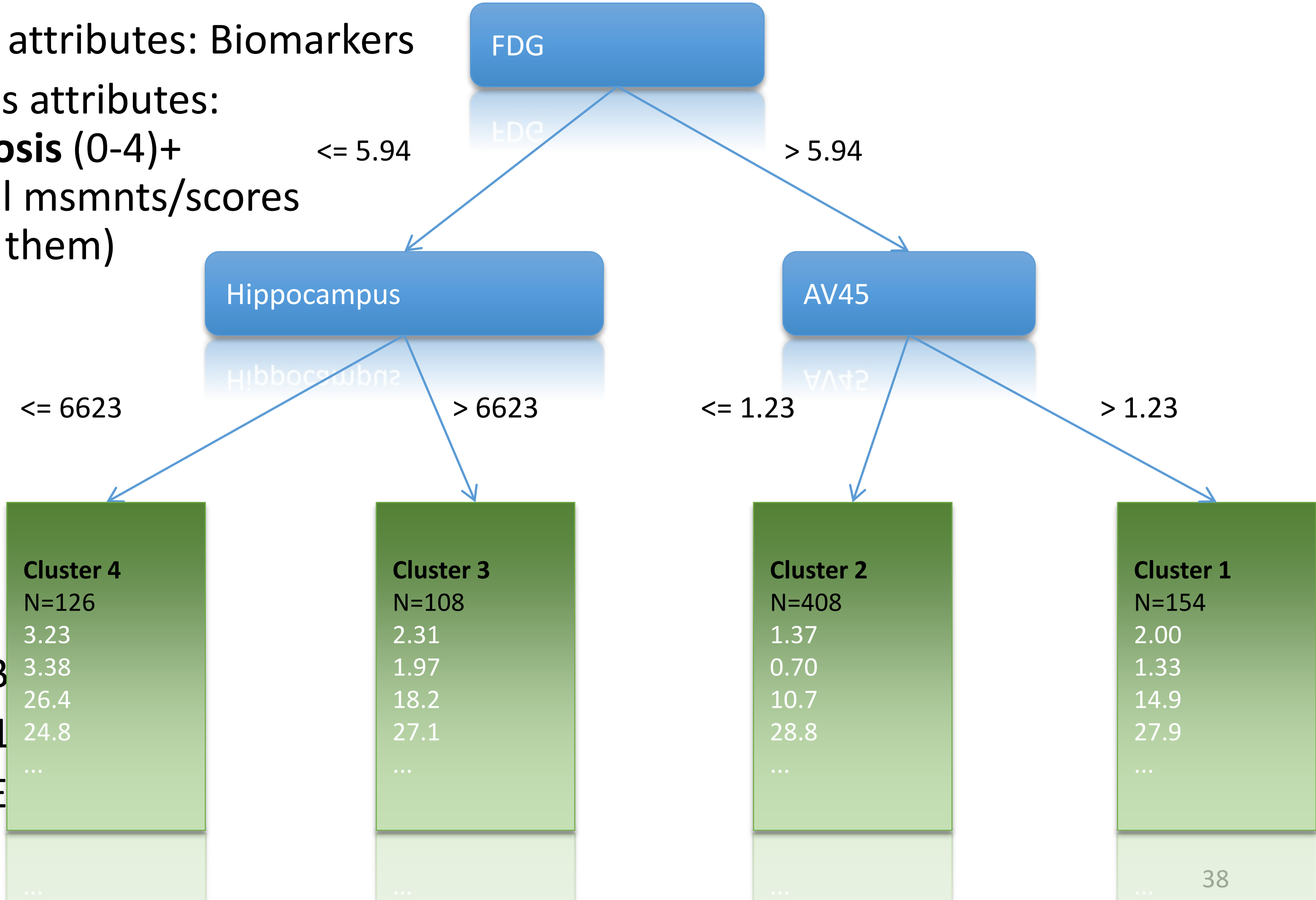


Predictive clustering

- Combines prediction and clustering
- We can have hierarchical clustering (trees) and flat/overlapping clusterings (rules)
- With each cluster, predictive clustering provides
  - A description of the cluster
  - A prediction of the selected targets for that cluster
- The output of PC can be viewed both as a clustering and as a predictive model

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 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 v1, …, vn, which **reduces the most the impurity of the target**
  - Partition S into S1, …, Sn according to A
  - Recursively construct subtrees T1 to Tn for S1 to Sn
  - Result: a tree with root A and subtrees T1, …, Tn

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 v1, …, vn, which **reduces the most the variance** (*measured according to a given distance function d*)
  - Partition S into S1, …, Sn according to A
  - Recursively construct subtrees T1 to Tn for S1 to Sn
  - Result: a tree with root A and subtrees T1, …, Tn

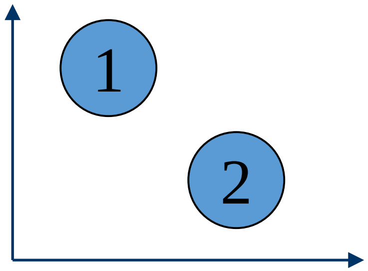
Learning PCTs

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

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

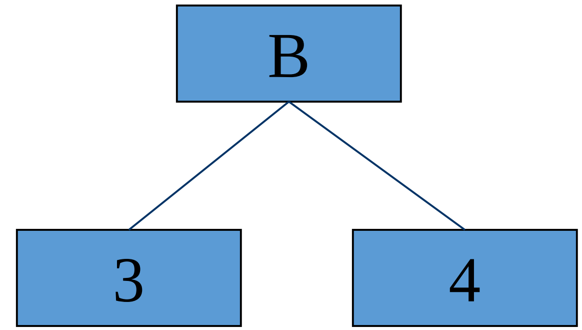
- For the variance, the distance is measured
  - In standard clustering, along all dimensions
  - In prediction, along a single target dimension
  - In predictive clustering, along a structured target, e.g., several target dimensions

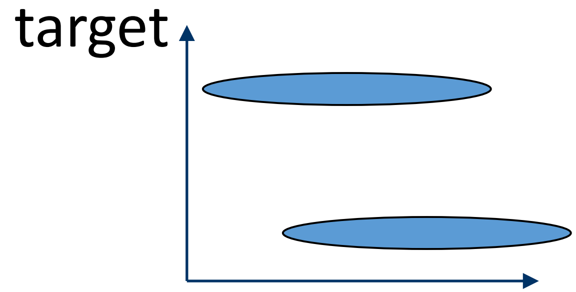
Clustering:



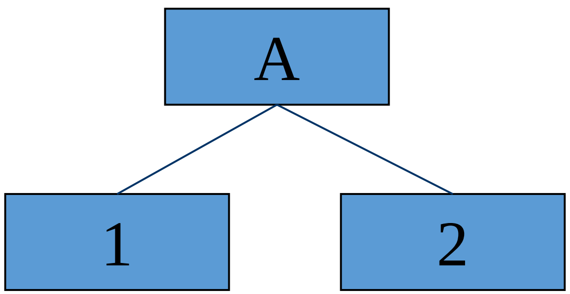
Data divided into clusters 1 and 2 coherent along two dimensions

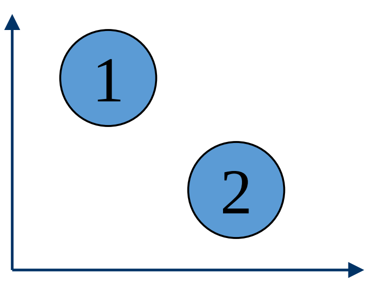
Prediction:





B divides data into clusters coherent along single *target*





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

Distances/variances for SOP tasks

- The algorithm

```
procedure BestTest(E)
  (t*, h*, 𝒫*) = (none, 0, ∅)
  for each possible test t do
      𝒫 = partition induced by t on E
      h = Var(E) − Σ_{E_i ∈ 𝒫} (|E_i| / |E|) Var(E_i)
      if (h > h*) ∧ Acceptable(t, 𝒫) then
          (t*, h*, 𝒫*) = (t, h, 𝒫)
  return (t*, h*, 𝒫*)
```

- Variance for MT regression

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

- Variance for MT classification

$$Var(E) = \textstyle\sum_{i=1}^{T} Entropy(E, Y_i)$$

- Variance for HMLC

$$Var(E) = \frac{1}{|E|} \cdot \sum_{E_i \in E} d(L_i, \overline{L})^2 \qquad d(L_1, L_2) = \sqrt{\sum_{l=1}^{|L|} w(c_l) \cdot (L_{1,l} - L_{2,l})^2}$$

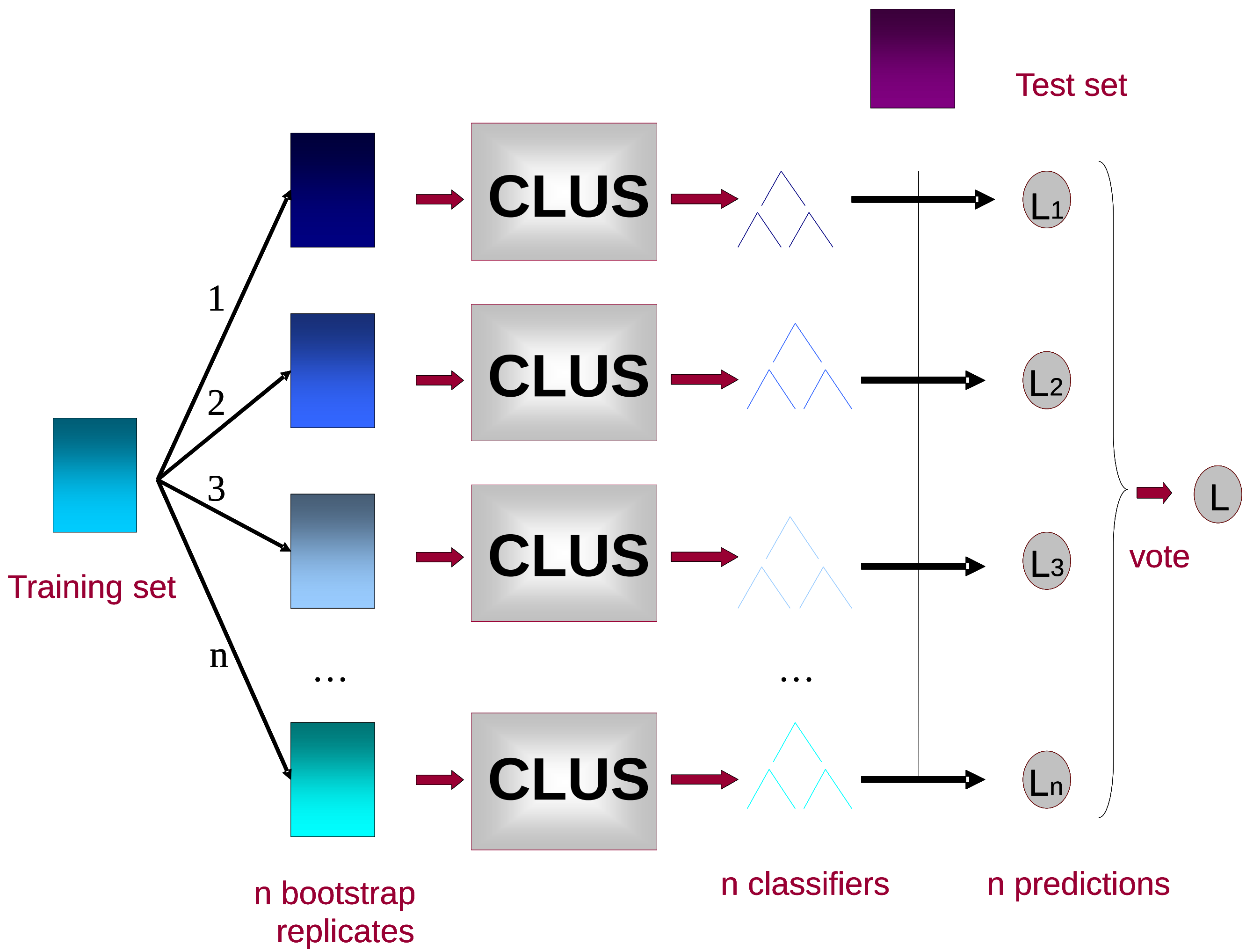
Ensembles of PCTs

- Ensembles of PCTs use several methods for constructing base classifiers
  - Bagging & Random forests
  - Random subspaces & Bagged Random subspaces

- PCTs and Ensembles of PCTs implemented in SW package CLUS, jointly developed by JSI, Ljubljana and KULeuven, Belgium

- Written in Java
- Open source, available for download from http://sourceforge.net/projects/clus

Ensembles of PCTs: Bagging



RandomForests & Feature Ranking

procedure Induce_RF($E, k, f(x)$)
returns Forest, Importances

$F = \emptyset$
$I = \emptyset$
for $i = 1$ **to** k **do**
  $E_i =$ Bootstrap_sample(E)
  $Tree_i = PCT_{rand}(E_i, f(x))$
  $F = F \bigcup Tree_i$
  $E_{OOB} = E \setminus E_i$
  Update_Imp($E_{OOB}, Tree, I$)
$I =$ Average(I, k)
return F, I

RandomForest Ranking

procedure Update_Imp(E_{OOB}, $Tree$, I)

1: Err_{OOB} = Evaluate($Tree$, E_{OOB})
2: **for** $j = 1$ **to** D **do**
3: $\quad E_j$ = Randomize(E_{OOB}, j)
4: $\quad Err_j$ = Evaluate($Tree$, E_j)
5: $\quad I_j = I_j + (Err_j - Err_{OOB})/Err_{OOB}$
6: **return**

procedure Average(I, k)

1: $I^T = \emptyset$
2: **for** $l = 1$ **to** $size(I)$ **do**
3: $\quad I_l^T = I_l / k$
4: **return** I^T

RandomForest Ranking

$$Importance(f_d) = \frac{1}{k} \cdot \sum_{i=1}^{k} \frac{Err_i(f_d) - Err(OOB_k)}{Err(OOB_k)}$$

k is the number of bootstrap replicates and $0 < d \leq D$

Random forest ranking for SOP

- This works for all types of outputs for which we can construct PCTs and ensembles thereof
- Multi-target classification
- Multi-label classification
- Hierarchical multi-label classification
- Multi-target regression

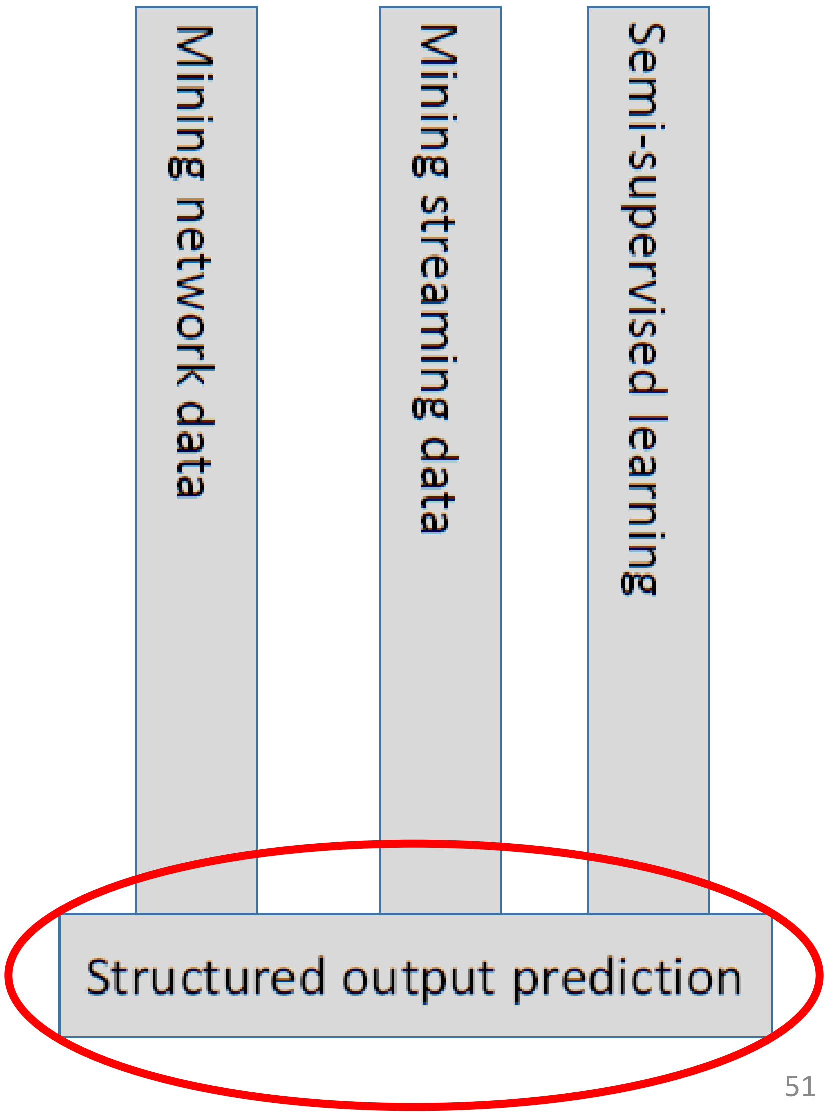
Combination of SOP with other complexity aspects

- RFs of PCTs & feature ranking therewith work with
  - Different types of SOP
  - With different degrees of supervision

- Unsupervised learning/ clustering
- Semi-supervised learning
- Fully supervised learning for different types of SOP
  - Multi-target classification
  - Multi-label classification
  - Hierarchical multi-label classification
  - Multi-target regression

The MAESTRA foundation & pillars

- Incomplete annotations
- Massive/streaming data
- Network context



Coming up next:

Semi-supervised Learning of PCTs

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Followed by:

Learning PCTs to Predict Structured Os from DS

MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Stay tuned for:

Learning in Networks

and

Applications of MBCD

MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Acknowledgements and announcement

We acknowledge European Commission support through the grants

- MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data, grant 612944
- HBP SGA1: The Human Brain Project, grant 720270
- LANDMARK: LAND Management: Assessment, Research, Knowledge base, grant 635201

As well as the Slovenian Research Agency through

- P2-0103 Knowledge technologies
- L2-7509 Structured output prediction …

And announce …

ECML PKDD 2017

SKOPJE, MACEDONIA

18-22 September 2017