

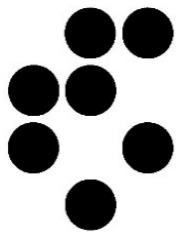


Predictive Clustering for Image Annotation & Retrieval

Sašo Džeroski

Jozef Stefan Institute, Ljubljana, Slovenia

(joint work with Ivica Dimitrovski, Dragi Kocev and Suzana Loškovska)



MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY
ANNOTATED, AND STRUCTURED DATA



Talk Outline

- Predictive clustering
 - From predictive modeling and clustering to predictive clustering
 - Predictive clustering for predicting structured outputs
 - Learning predictive clustering trees
 - Ensembles of predictive clustering trees
- Image annotation with PCTs and ensembles
 - Taxonomical classification of diatom images
 - Hierarchical annotation of medical images
- Visual codebook construction with PCTs and ensembles
 - Supervised for image annotation
 - Unsupervised for image retrieval



Predictive Modelling/ Supervised L.

- 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)
- Descriptive vars.: genetic and image markers

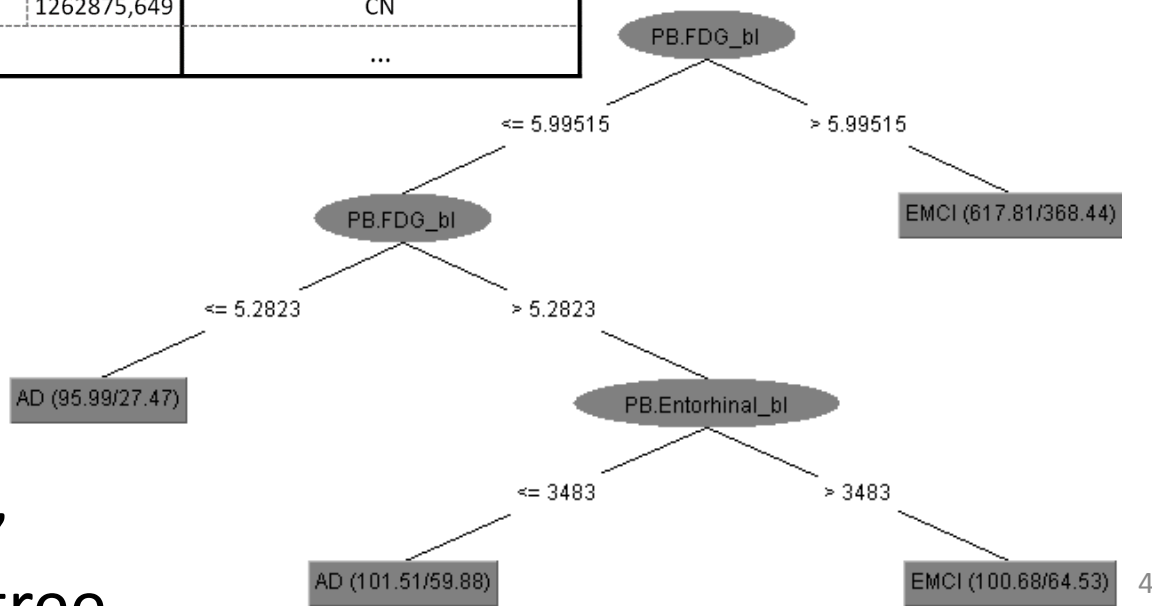


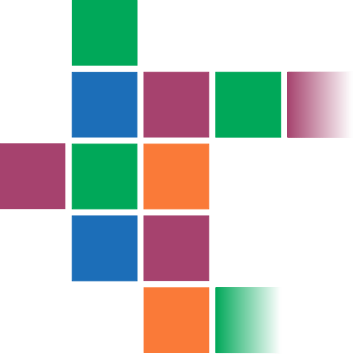
Predictive Modelling

- 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, e.g. decision tree





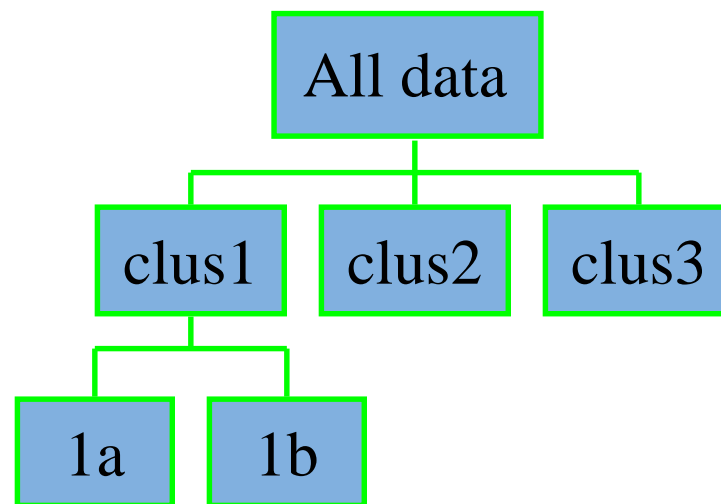
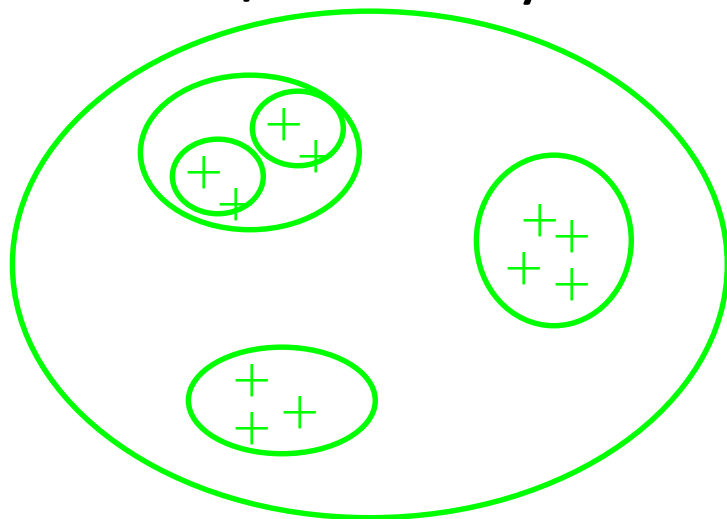
Top-Down Induction of Classification (Regr.) Trees

To construct a tree T from a training set S :

- If **all the examples belong to the same class C (the values of the target have low variance)**, construct a leaf labeled with the class value C (the target average)
- 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

Clustering/ Unsupervised L.

- 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





Basic Clustering Approaches

- K-Means clustering
 - Randomly assign instances to k clusters, then repeat:
 - Calculate centroids of clusters, reassign instances to clusters
 - Until convergence (i.e., cluster assignment doesn't change)
- Hierarchical agglomerative clustering
 - Start with each instance as a cluster, then repeat
 - Merge the two closest clusters
 - Until all instances are in one single cluster



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 (cf. next example)

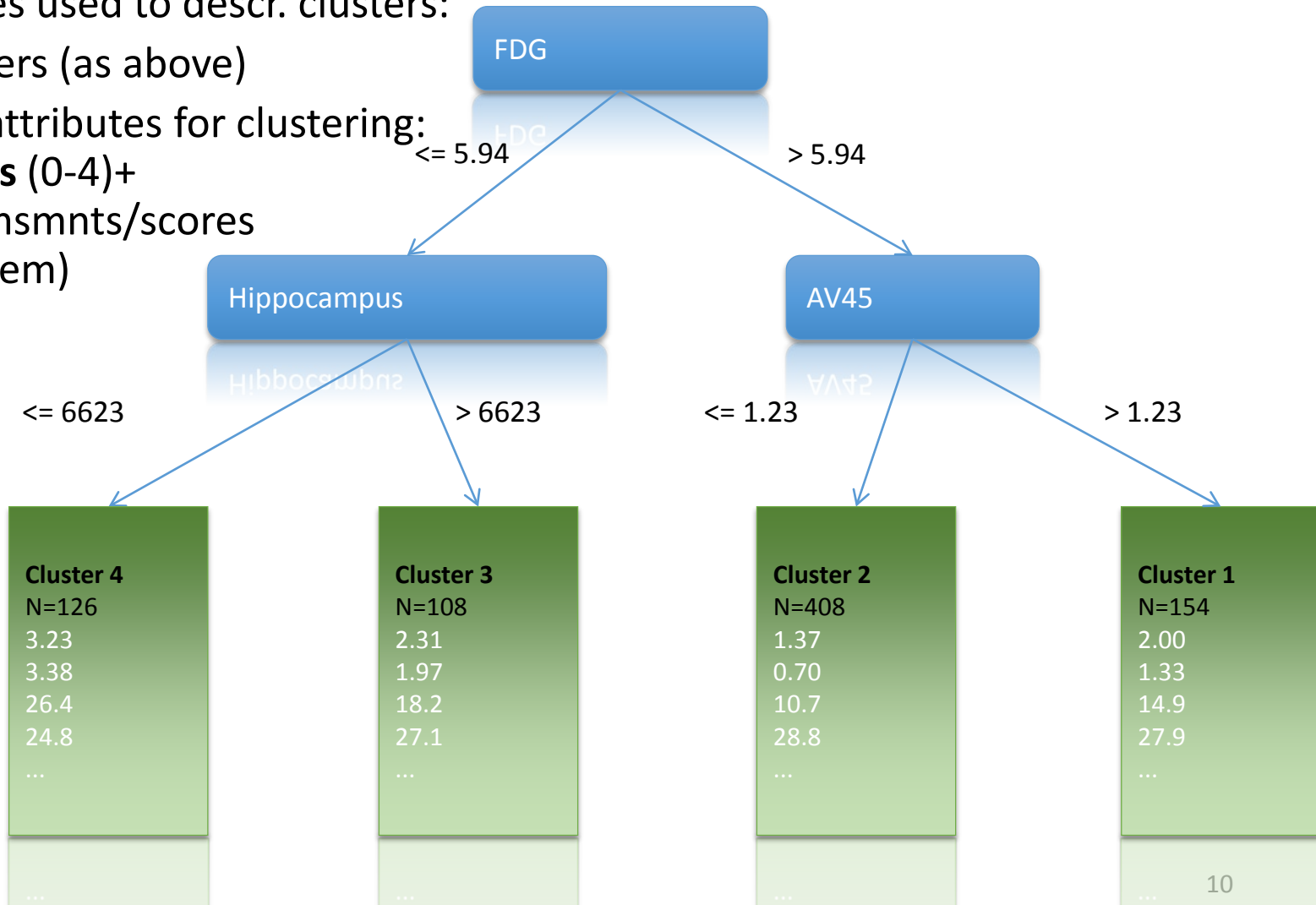


Example Task: Cluster Alzheimer's Patients wrt. Clinical Scores

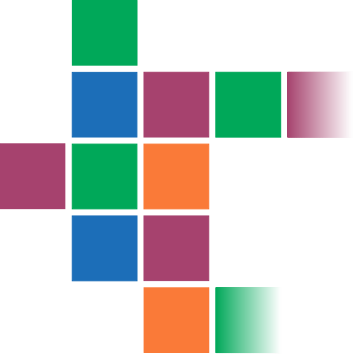
1. CDRSB – Clinical Dementia Rating Sum of Boxes
2. ADAS13 – AD assessment scale
3. MMSE – Mini Mental State Examination
4. RAVLT (immediate, learning, forgetting, perc. forgetting) – Rey Auditory Verbal Learning Test (4 features)
5. FAQ – Functional Assessment Questionnaire
6. MOCA – Montreal Cognitive 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 Predictive Clustering Tree for Multi-Target Regression

- Attributes used to descr. clusters:
Biomarkers (as above)
- Targets attributes for clustering:
diagnosis (0-4)+
clinical msmnts/scores
(23 of them)



- DX
- CDRSB
- ADAS13
- MMSE
- ...



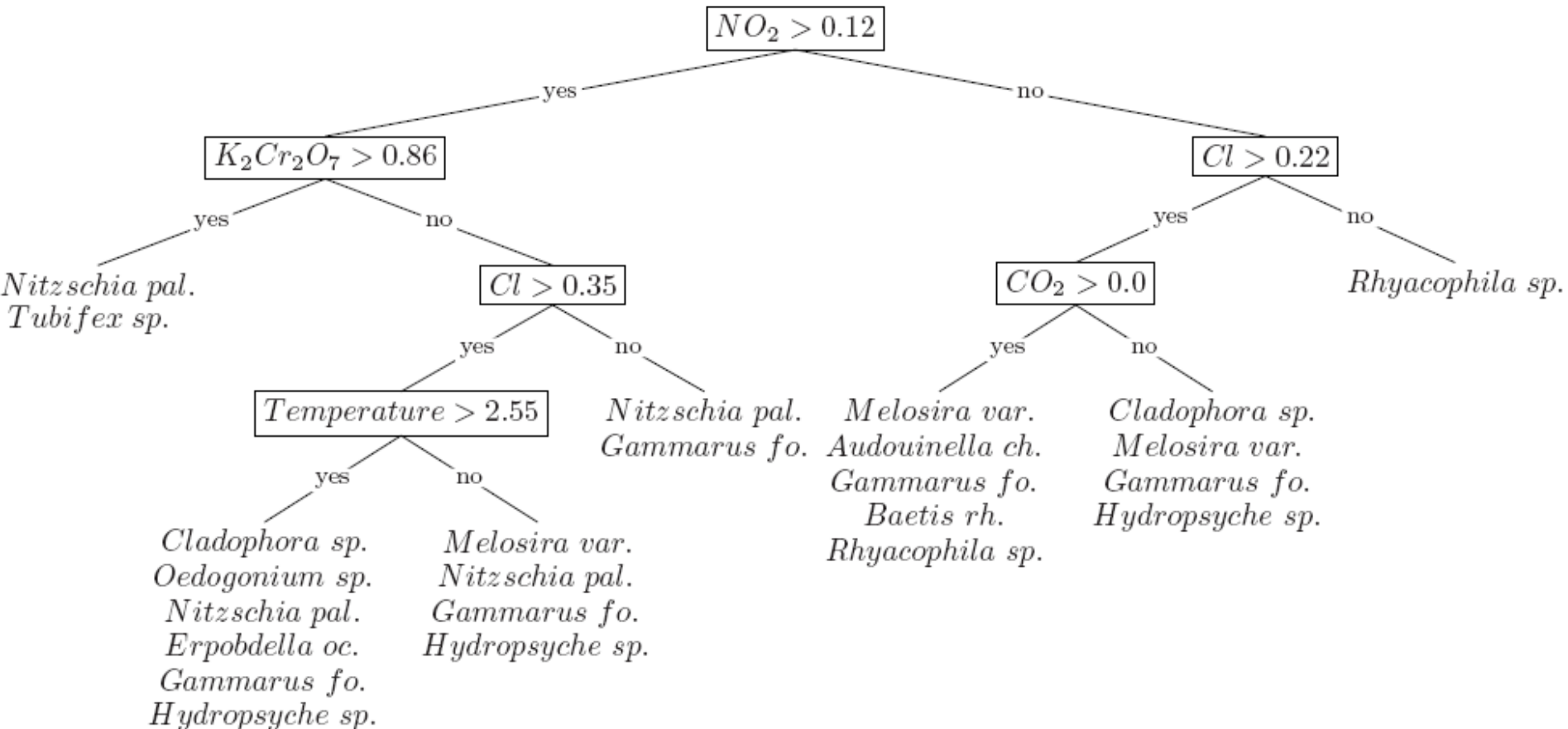
Multi-Label Classification

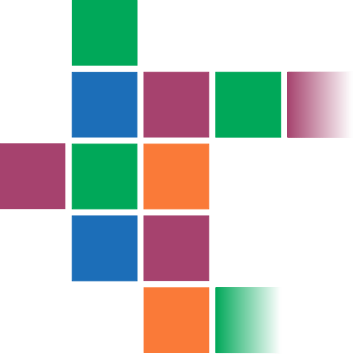
- Special case of multi-target prediction (incl. MTR & MTC)
- Learning models that simultaneously predict several binary target variables (a set of labels)
- Input: A vector of descriptive variables (as for STC/STR)

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

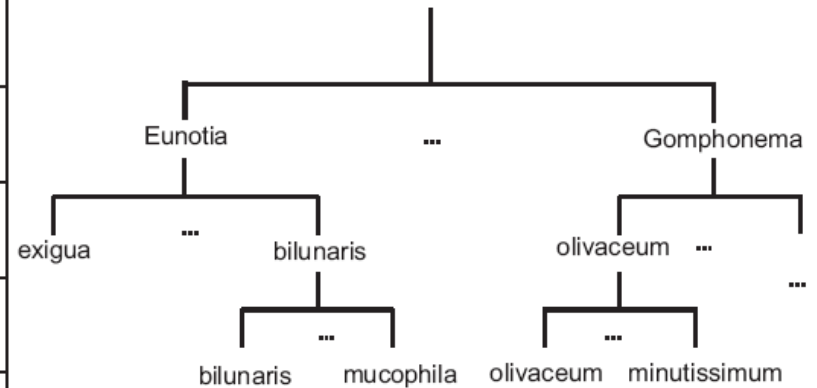




Hierarchical Multi-Label Classification (HMC)

- Labels organized in a hierarchy
- Taxonomic classification of diatoms
- From microscopic images
- Taking into account the existing 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
...





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

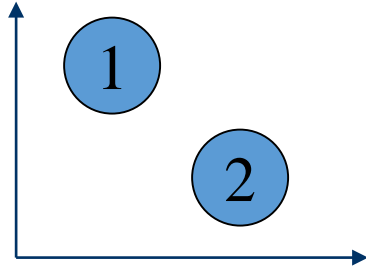
- Recursively partition data set into subsets (clusters) with low intra-cluster variance
 - Variance = avg. squared distance to prototype

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

- 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

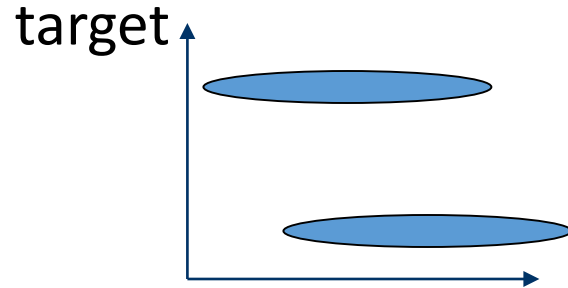
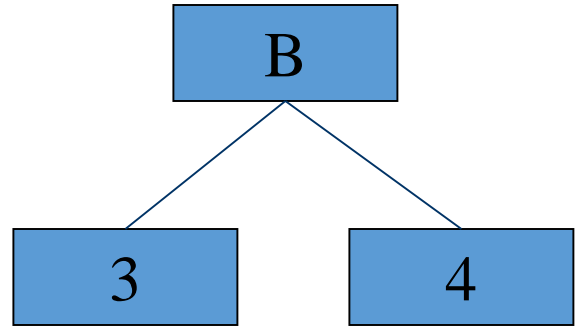


Clustering:

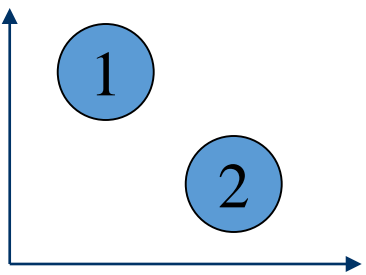
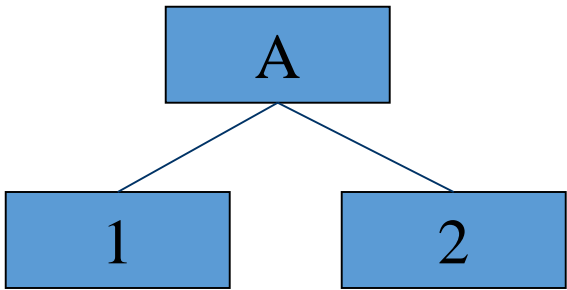


Data divided into clusters 1 and 2 coherent along two dimensions

Prediction:



B divides data into clusters coherent along single *target*



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



Distances/variances for SOP tasks

- The algorithm
- Variance for MT regression

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

- Variance for MT classification

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

- Variance for HMLC

$$\text{Var}(E) = \frac{1}{|E|} \cdot \sum_{E_i \in E} d(L_i, \bar{L})^2$$

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

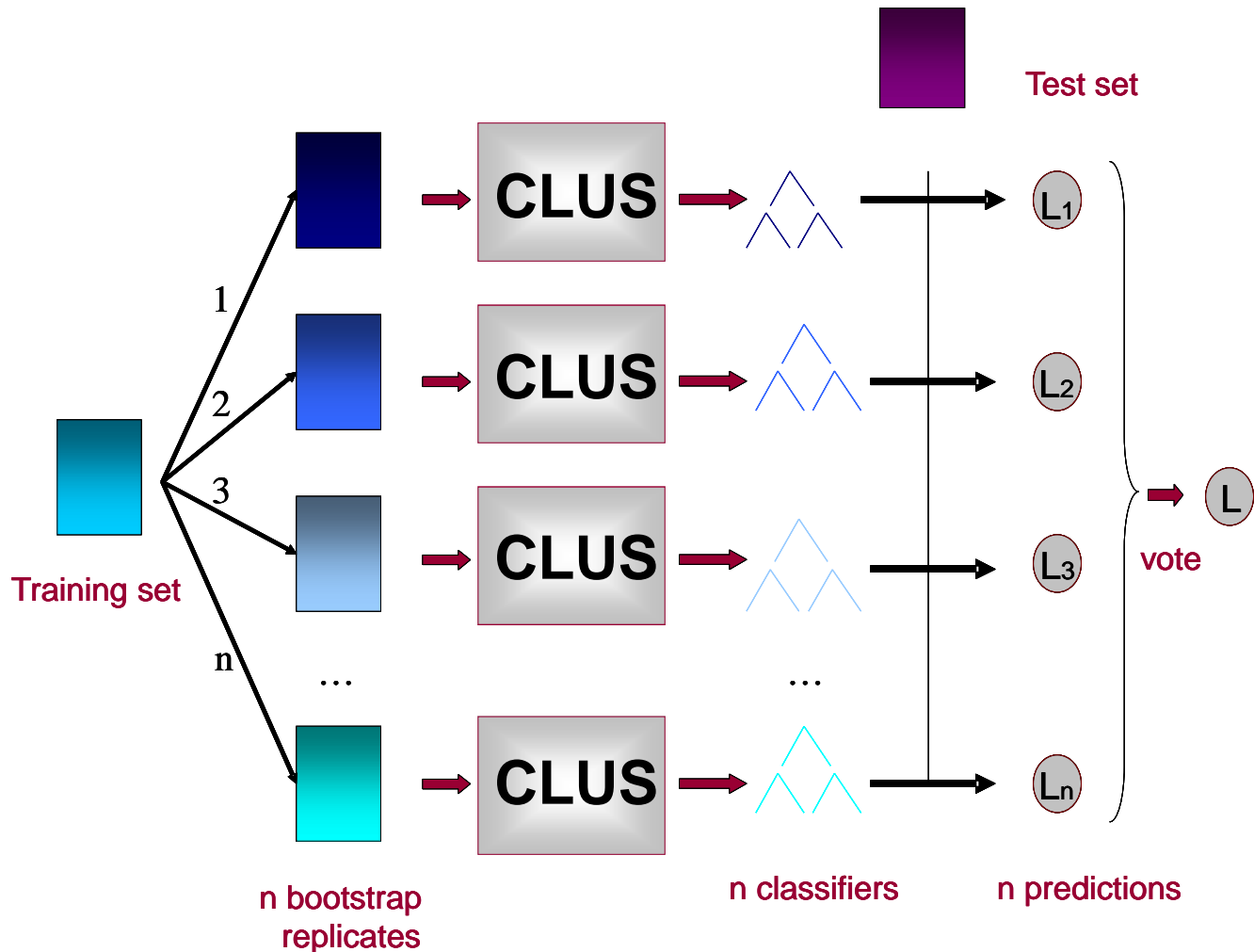
$$d(L_1, L_2) = \sqrt{\sum_{l=1}^{|L|} w(c_l) \cdot (L_{1,l} - L_{2,l})^2}$$



Ensembles of PCTs

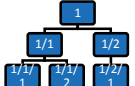
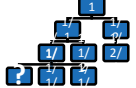
- Ensembles of PCTs use several methods for constructing base classifiers
 - Bagging & Random forests
 - Random subspaces & Bagged Random subspaces
- PCTs and Ensembles of PCTs implemented in SW package CLUS, jointly developed by JSI, Ljubljana and KULEuven, Belgium
- Written in Java
- Open source, available for download from <http://sourceforge.net/projects/clus>

Ensembles of PCTs: Bagging



SSL: Incomplete Annotations

- Some examples have labels, some don't, some incompl.

	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	

	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

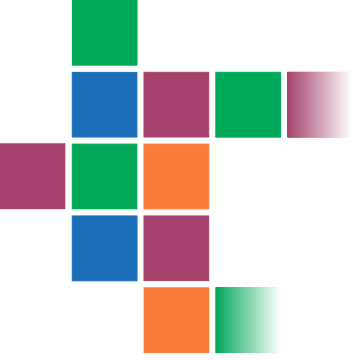


Semi-Supervised Learning w. PCTs

- New definition of variance that includes both targets and attributes, e.g., for MTR

$$Var(E) = \frac{1}{T + D} \cdot \left(w \cdot \sum_{i=1}^T Var(Y_i) + (1 - w) \cdot \sum_{j=1}^D Var(X_j) \right)$$

- T = #target attributes, D = #descriptive attributes
- w = weight parameter, trades-off focus on
 - Prediction ($w=1$)
 - Clustering ($w=0$)
- w tuned by internal cross-validation on labeled part



SSL: Calculating Variance for Attributes with Missing Values

Variances of individual target (Y_i) and descriptive (X_i) attributes:

$$\text{Var}(Y_i) = \frac{\frac{N-1}{K_i-1} \cdot \sum_{j=1}^N (y_{ij})^2 - N \cdot \left(\frac{1}{K_i} \cdot \sum_{j=1}^N y_{ij} \right)^2}{N}$$

N = number of examples,

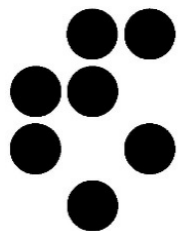
K_i = number of examples with ***non missing values***

In extreme cases ($K=0$), est. var. with var. of parent node:

- (1) leafs of the decision tree may contain only unlabeled examples
- (2) in a leaf, some descr. attributes may have only missing values.



Image Annotation and Retrieval with PCTs

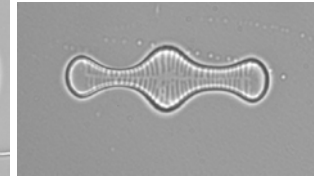
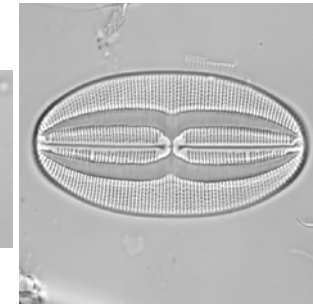
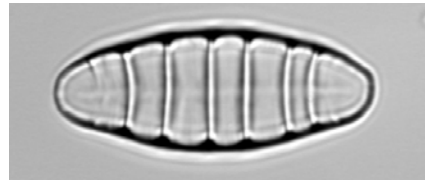


MAESTRA

LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA



Taxonomic Identification of Diatoms from Microscope Images



- Automated diatom classification
 - image processing (feature extraction from images)
 - image classification (labels and groups the images)
- Labels organized in a hierarchy
- Predict all different levels in the hierarchy of taxonomic ranks: genus, species, variety, and form
- Goal of the complete system: assist a taxonomist in identifying a wide range of different diatoms



Feature Extraction from Images

- Contour extraction, then
- Simple geometric properties
 - length, width, size and the length-width ratio
 - Simple shape descriptors: rectangularity, triangularity, compactness, ellipticity, and circularity
- Fourier descriptors (30 coefficients)
- SIFT histograms (key-point detection+)
 - Invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint
 - Cluster key-points, assign KPs to clusters, hist.



Diatom Classification Results

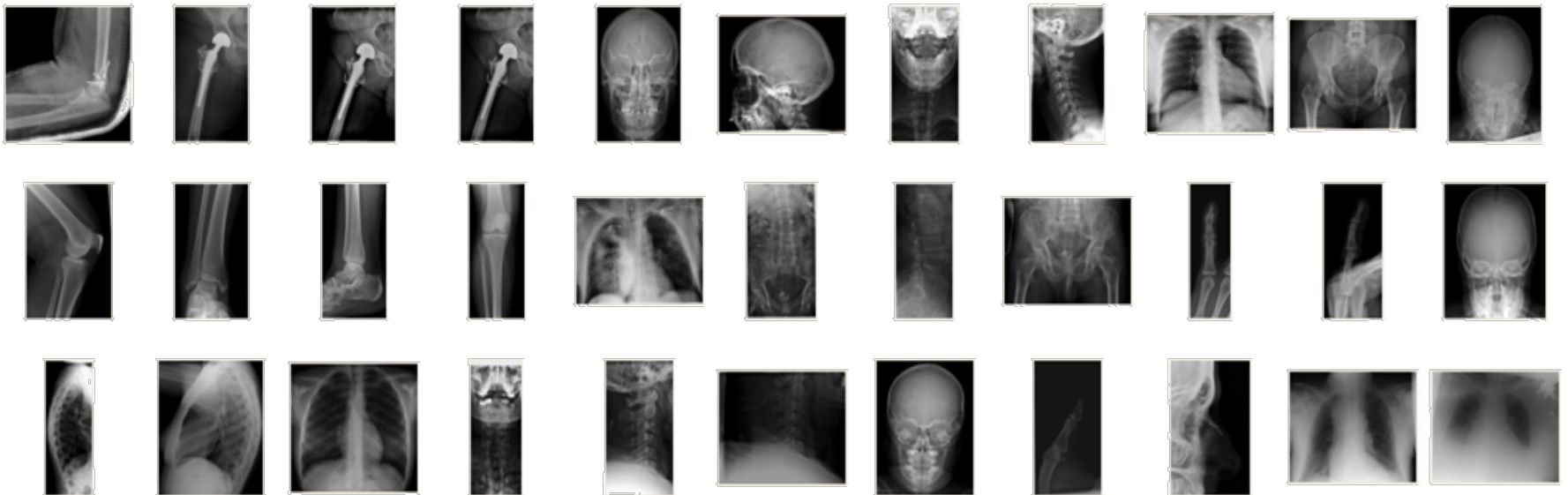
- Predictive performance of the different feature sets and their combinations

Classifier	Descriptors	# features	Overall recognition rate [%]		
			55 diatom taxa	48 diatom taxa	37 diatom taxa
Bagging	Geometric and shape descriptors	9	76.3	76.7	77.2
	Fourier descriptors	30	86.7	88.1	88.6
	SIFT histograms	200	88.4	89.2	91.3
	Geometric and shape desc.+Fourier desc.+SIFT hist.	239	96.2	98.1	98.8
Random Forests	Geometric and shape descriptors	9	76.3	76.7	77.2
	Fourier descriptors	30	86.6	88.1	88.7
	SIFT histograms	200	88.2	87.9	91.1
	Geometric and shape desc.+Fourier desc.+SIFT hist.	239	96.2	98.1	98.7



Medical Image Annotation

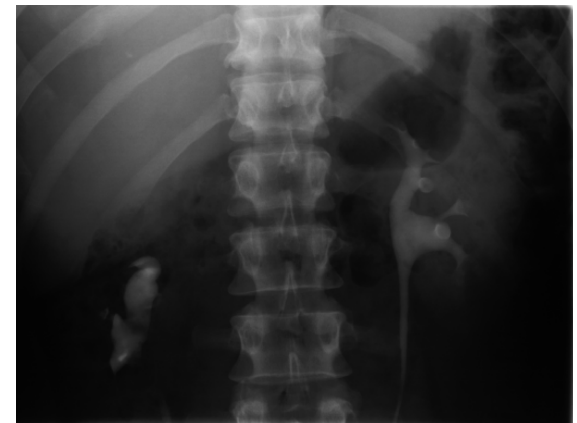
- ImageCLEF2009 Challenge
 - 12677 annotated x-ray images; 1733 non-annotated images
- Hierarchical classification according to two labeling sets:
 - ImageCLEF2007: 116 IRMA codes
 - ImageCLEF2008: 196 IRMA codes








IRMA Coding System

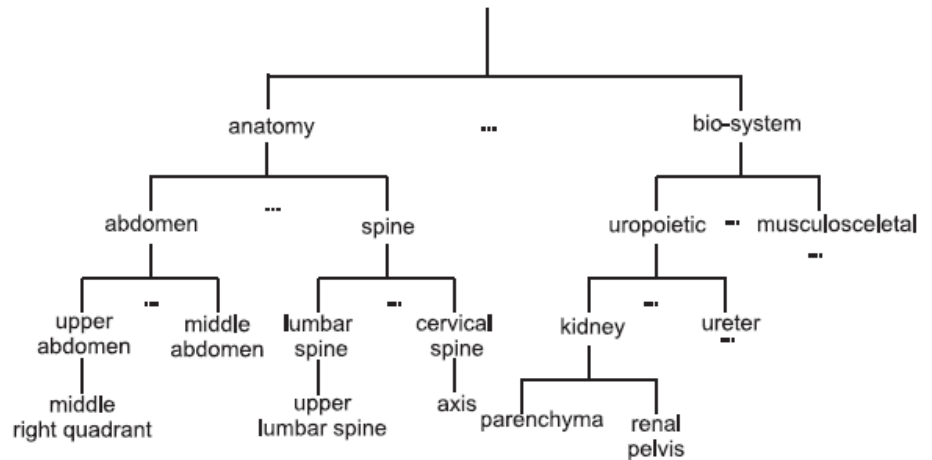
- Four axes marked with {0, ..., 9, a, ..., z}
 - T (Technical): image modality
 - D (Directional): body orientation
 - A (Anatomical): body region
 - B (Biological): biological system
- IRMA code: TTTT – DDD – AAA – BBB
- The code is strictly hierarchical
 - 5 uropoietic system
 - 51 uropoietic system, kidney
 - 512 uropoietic system, kidney, renal pelvis



Medical Image Annotation

- Set of images with their visual descriptors and annotations
- Annotations with IRMA codes, hierarchical

image	features/descriptors					annotations/labels	
		—	/	\	ξ		
	48	24	59	66	37	...	cervical spine@ musculoskeletal system
	36	25	53	45	15	...	middle abdomen@renal pelvis
	35	25	56	52	19	...	lumbar spine@ musculoskeletal system
...

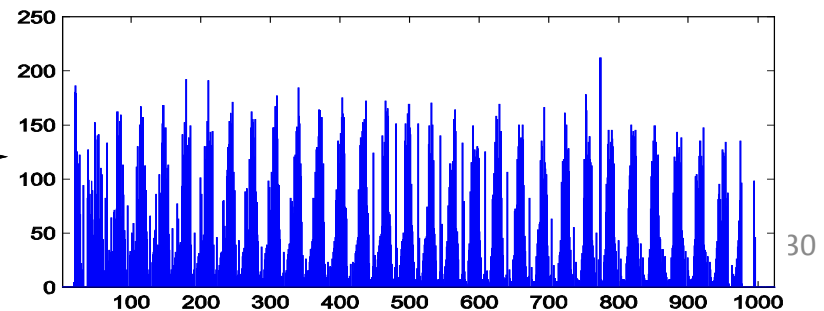
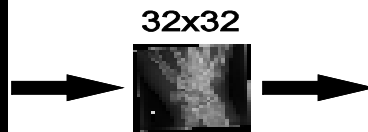




Feature Extraction

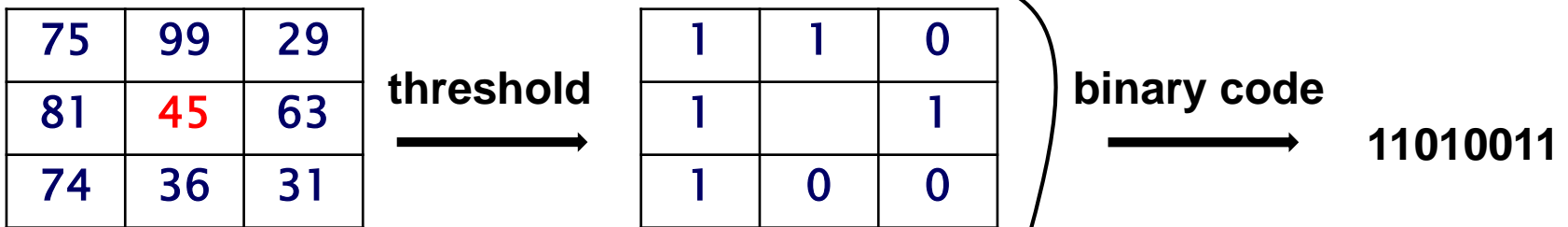
- Local Binary Pattern (LBP) histograms
- Edge Histogram Descriptor (EHD)
- Scale Invariant Feature Transform (SIFT) histograms
- Raw pixel representation (RPR)
 - Scale the image to a common size (32x32 pixels)
 - Represent the image by a feature vector that contains image pixel values

512x446

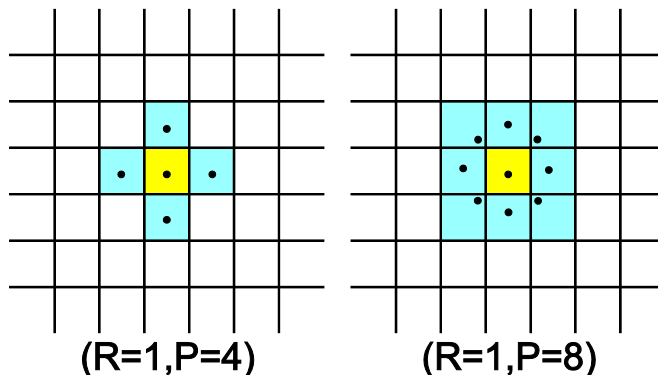


Local Binary Patterns

- Binary code to describe the local texture pattern in a circular region thresholding each neighborhood on the circle by the gray value of its center

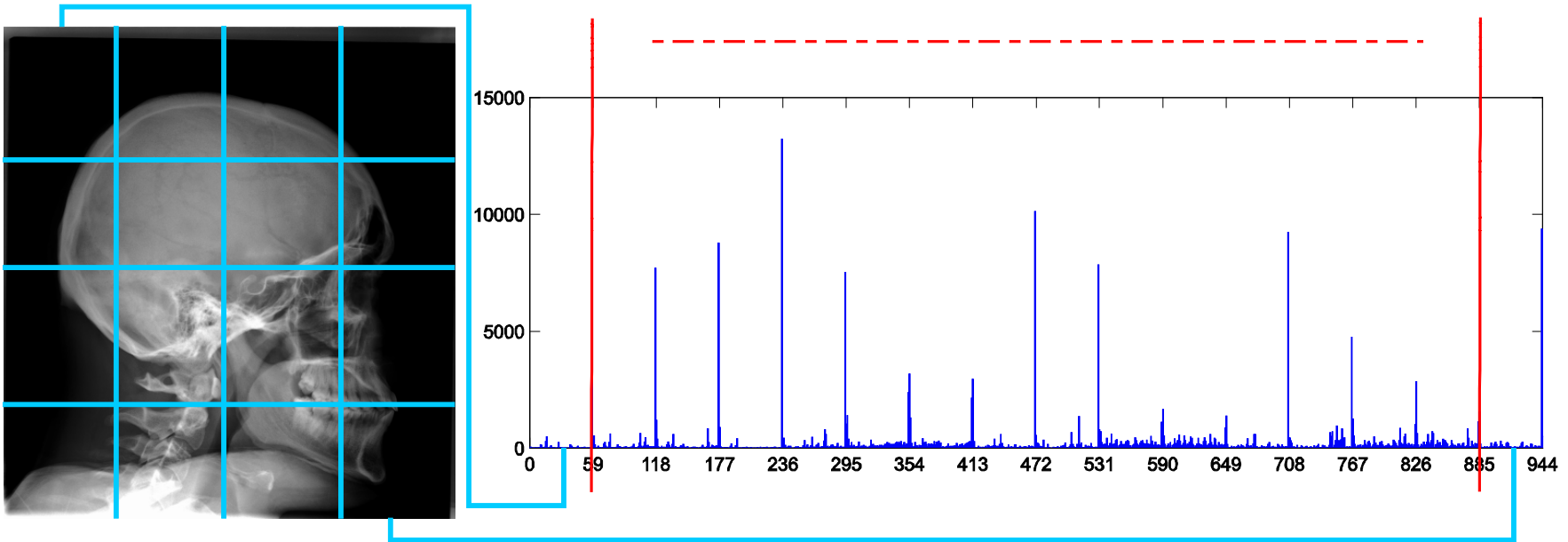


- Circular symmetric neighborhood with different radius R and number of points P



LBP Histograms

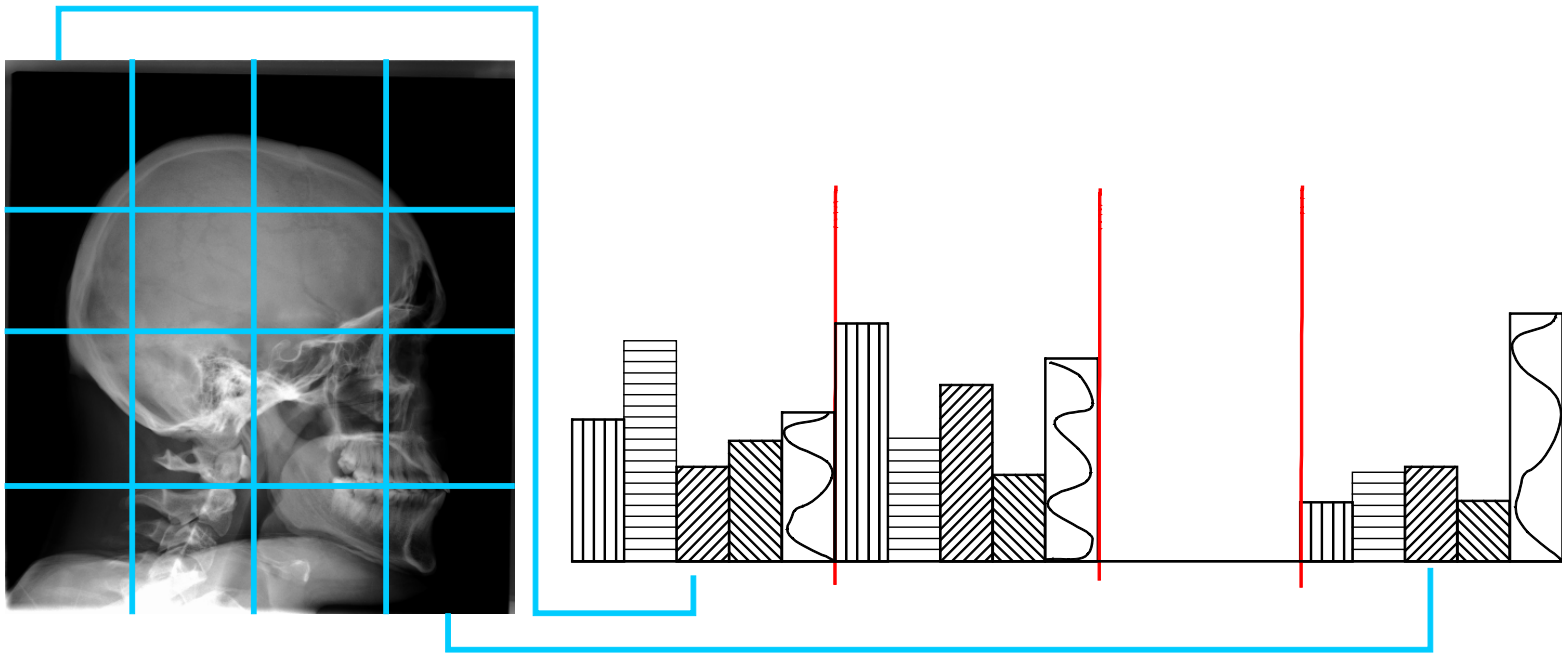
- Image divided in 4x4 parts
- From each sub-image extract ULBP(1,8)





Edge Histogram Descriptor

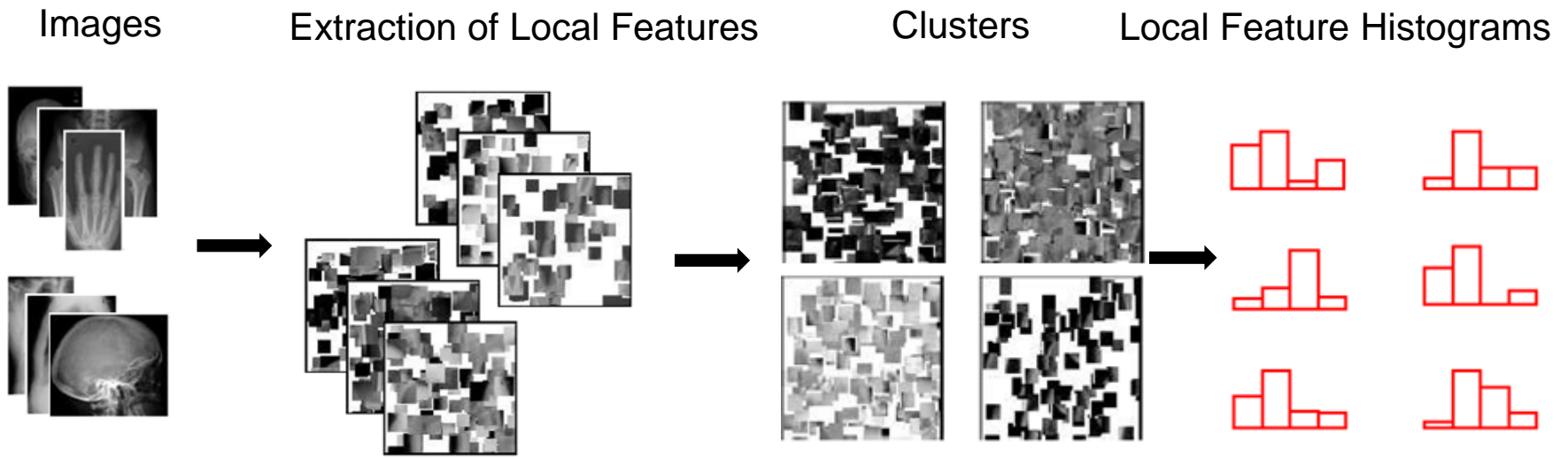
- Sharp change of luminous intensity
- Information about the shapes of the objects
- Frequency and the directionality of the brightness changes in the image





SIFT: Bag of Visual Words

- Extract local SIFT features
- Construct visual word dictionary
- Using K-means clustering
- Vocabulary size – number of visual words
- Local feature histogram





Medical Image Annotation

Comparative study of ensembles of PCTs for HMC and collections of SVMs, one per label

Summary of results

- Ensembles (RFs) of PCTs for HMC perform better
 - Lower hierarchical error measure
 - Higher overall recognition rate
 - Best results on these datasets so far
- RFs of PCTs for HMC are also much more efficient/faster



Constructing BOW Codebooks w PCTs

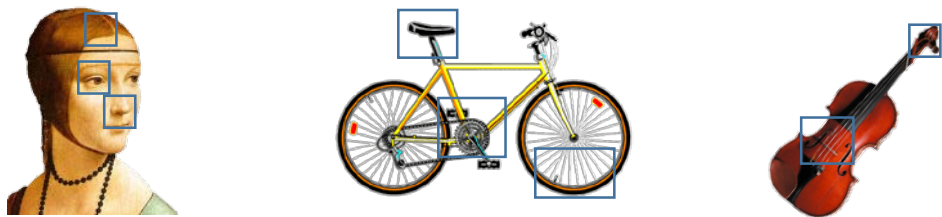
- Visual codebook construction
 - Unsupervised for image retrieval
 - Supervised for image annotation
- Image annotation with hierarchically structured labels (medical X-ray images) and general images
- We used (small) ensembles of PCTs for constructing BOW codebooks
- We learned to annotate using collections of SVMs



Bag-of-Visual-Words (BoVW)

1. Extract features

- Select key points/patches/regions



- Calculate descriptors/features of the selected patches

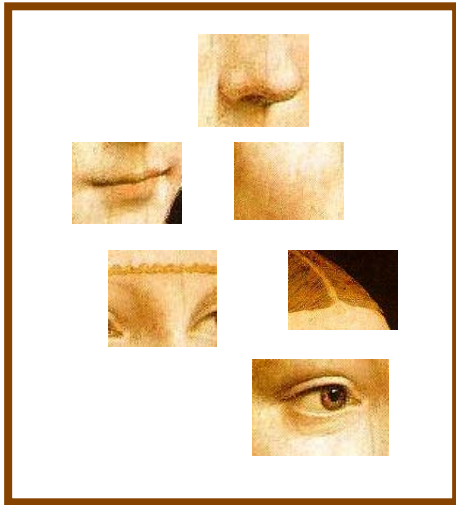
	12 45 78 ...
	34 56 124 ...
	1 6 84 ...



BoVW: Learning a Visual Codebook

2. Learn a visual codebook

- Input: Set of descriptors
- Output: Clusters (Visual Words)



Visual Word 1



Visual Word 2

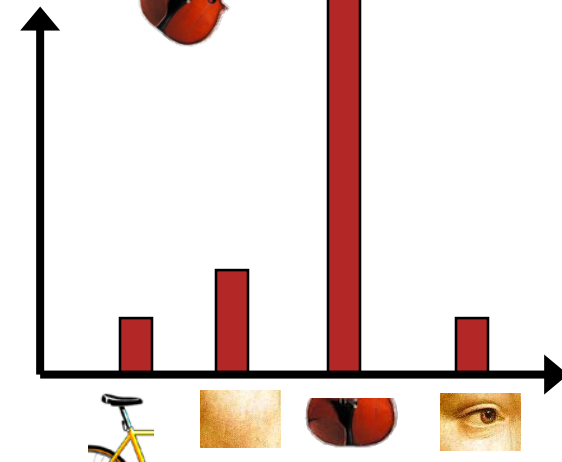
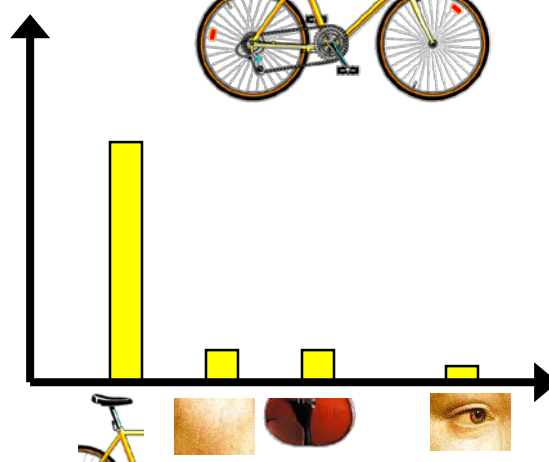
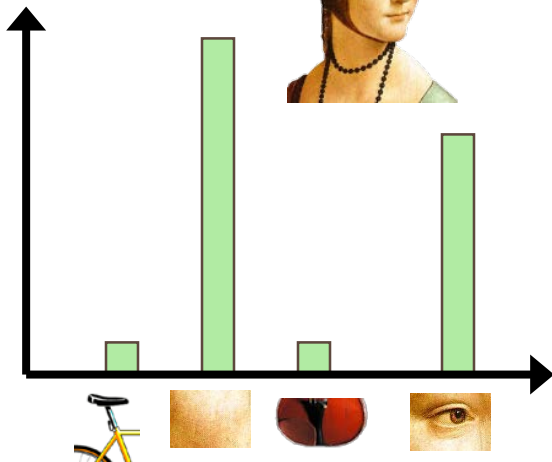


Visual Word 3



BoVW: Representing Images

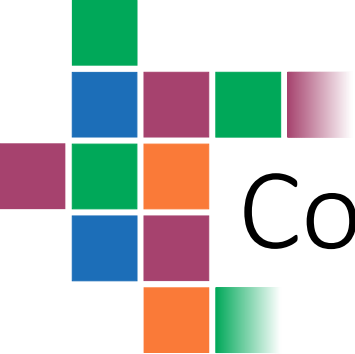
1. Extract features
2. Learn a visual codebook
3. Represent the images by histograms (distribution of the patches over the visual words)





Related Work

- Construction of a visual codebook is a bottleneck in the bag-of-visual-words approach
- k-means to cluster local image regions into visual words
 - Serious limitations for large scale object retrieval
- Hierarchical k-means, approximate k-means and extremely randomized tree ensembles
 - Improve the efficiency at the cost of decrease of the discriminative power of the obtained codebook
- **Our method:** Visual codebook construction using predictive clustering trees to alleviate the efficiency issues and increase the predictive power

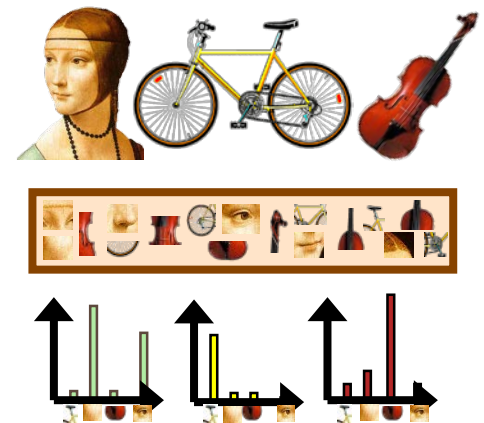
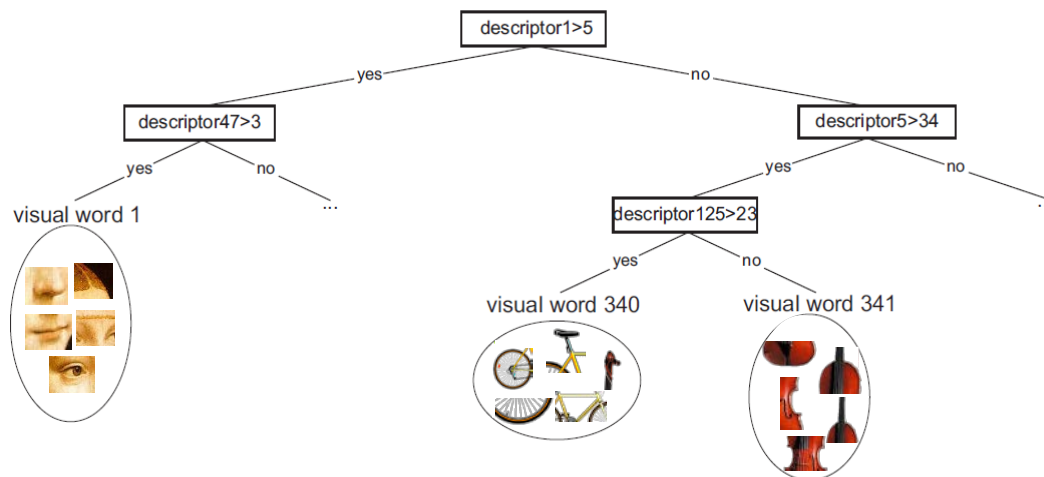


Codebook: Random forest of PCTs

- Here we use a small number of trees in the forest
- Large scale object retrieval
 - Random forest of PCTs for multi-target regression
 - Descriptive and target space are the same
- Multi-label image annotation
 - Random forest of PCTs for multi-label classification
 - Use the annotations of the images to guide the construction of the visual codebooks

Visual Codebook

- Each tree leaf is a visual word
- Each image is described with a histogram of the number of regions per visual word



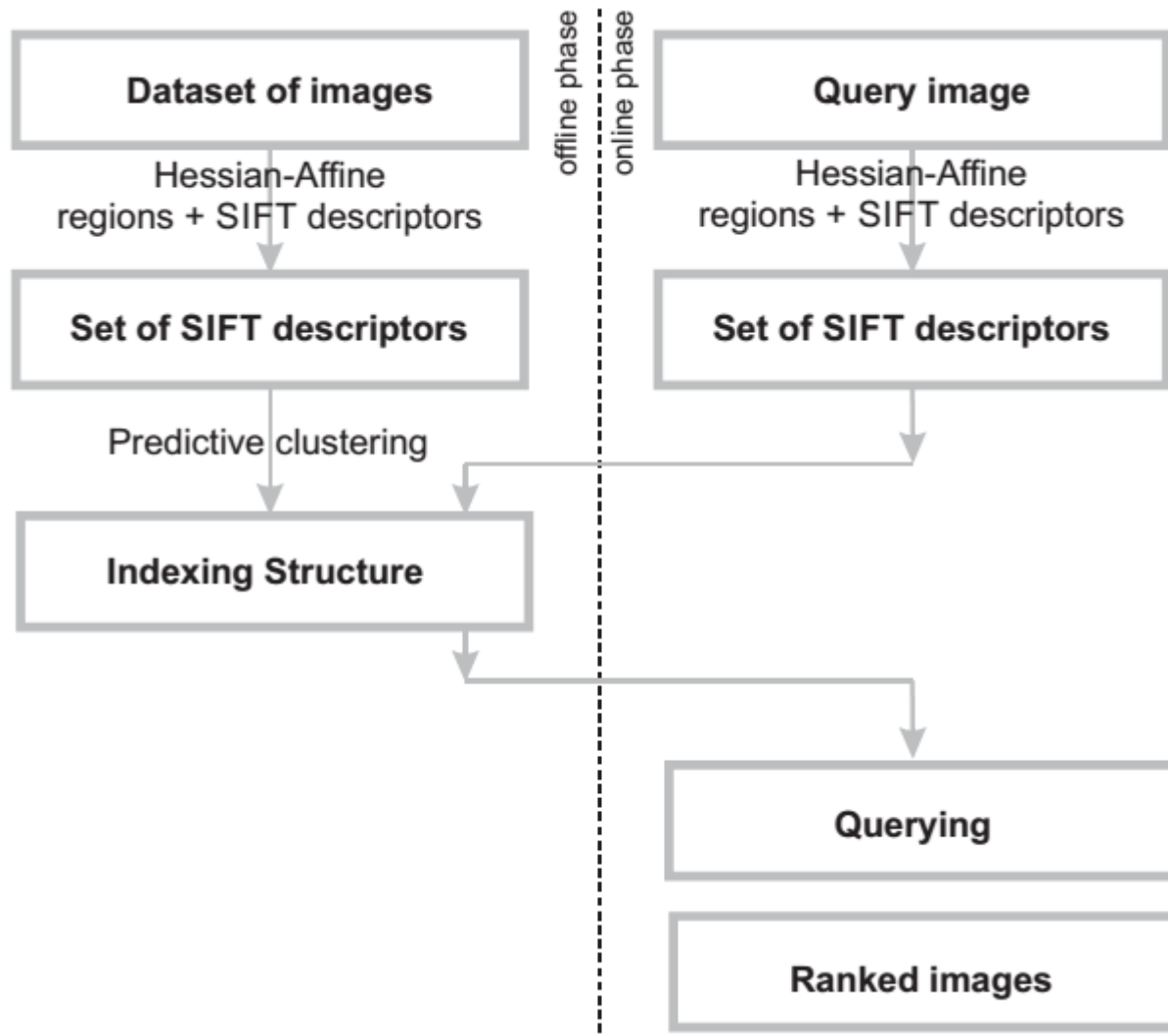
- PCTs are computationally efficient in both construction and prediction, but rather unstable: small random forest of PCTs to obtain the overall codebook
- Concatenation of the codebooks of each PCT



Data Description

- Oxford5k dataset: 5062 high-resolution images of Oxford landmarks
- Paris dataset: 6412 high-resolution images
- Pythia: 5555 high-resolution images
- PASCAL VOC 2007: 9963 images, 20 labels, 1.46 labels per image
- ImageCLEF@ICPR: 8000 images, 53 labels, 8.68 labels per image
- ImageCLEF 2010: 8000 images, 93 labels, 12.06 labels per image
- Oxford100K: 100K images from Flickr by searching the 145 most popular tags
- Oxford1M: 1M images from Flickr by searching the 450 most popular tags
- Challenges: substantial variations in scale, viewpoint and lighting conditions of the images and the objects

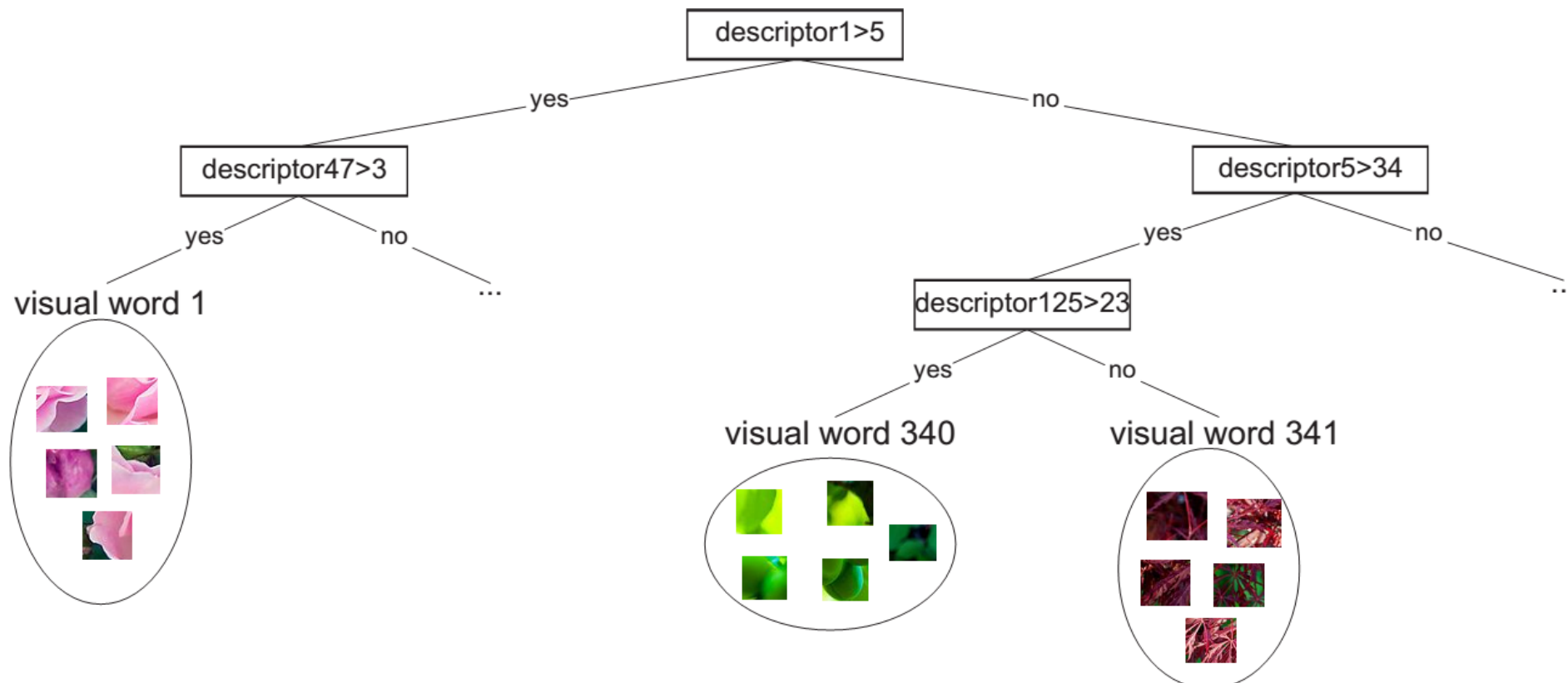
Unsupervised Codebook Constr.






Unsupervised PCTs

- The descriptive space is simultaneously used as a target space





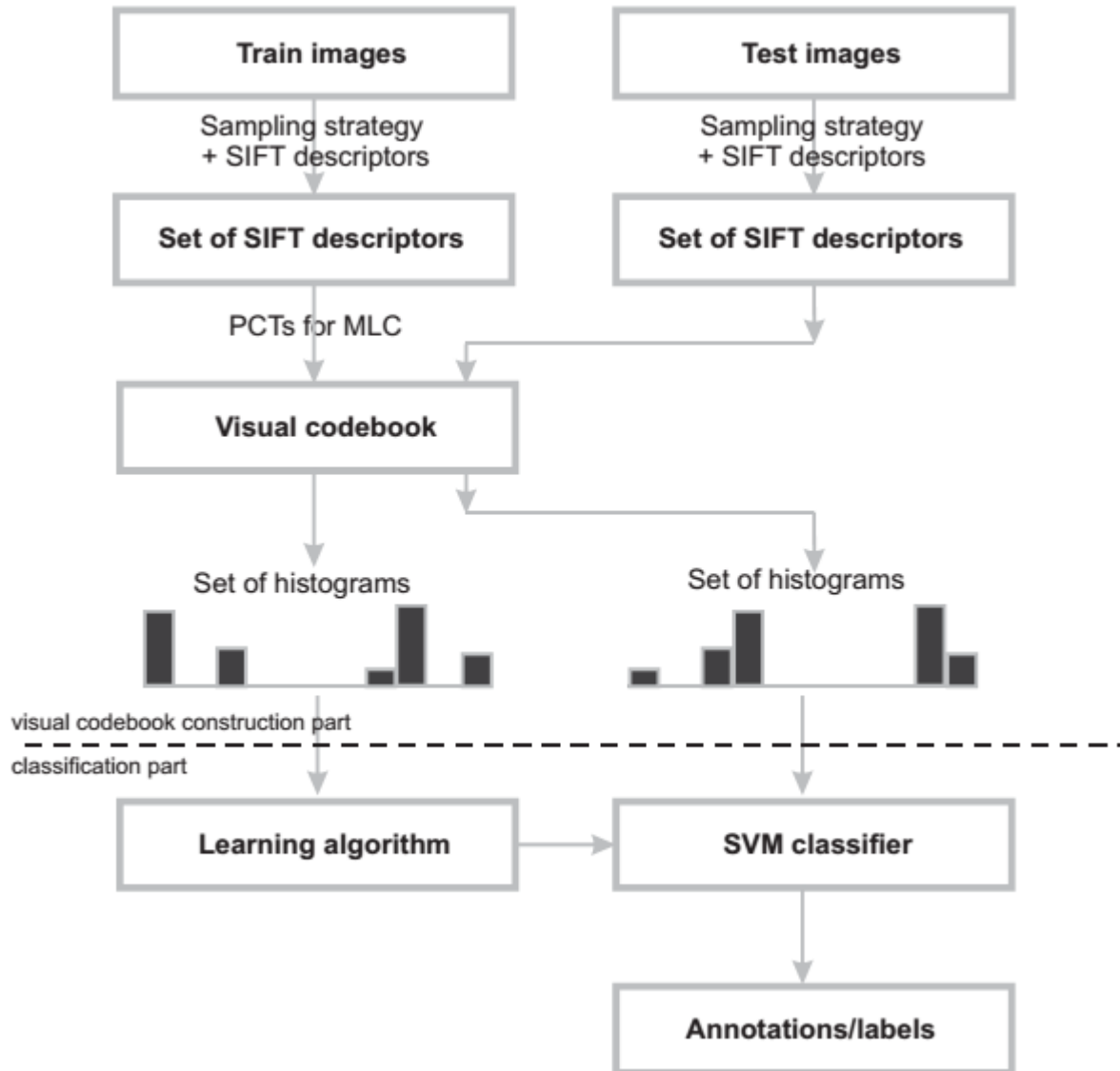
Large Scale Object Retrieval: Performance & Scalability

Comparison of the retrieval performance (given as mean average precision)

Image dataset	Without Spatial re-ranking			With Spatial re-ranking		
	AKM	ExtraTrees	RF of PCTs	AKM	ExtraTrees	RF of PCTs
<i>Oxford5K</i>	0.680	0.675	0.712	0.720	0.710	0.761
<i>Paris</i>	0.687	0.661	0.701	0.688	0.673	0.710
<i>Pythia</i>	0.164	0.172	0.213	0.170	0.189	0.234

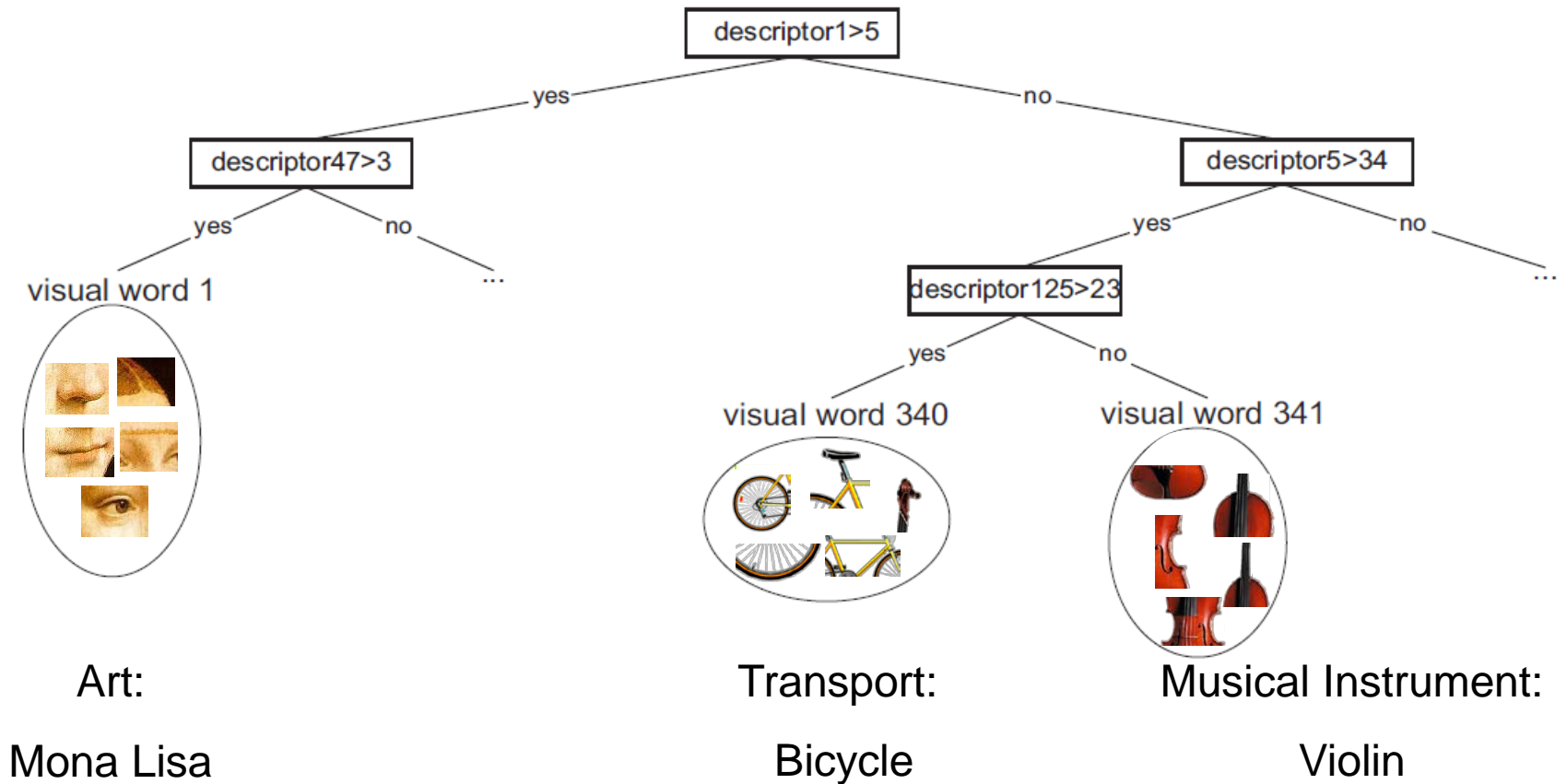
- Spatial re-ranking of a short-list of top ranked results to further boost the retrieval performance
- Better results with larger codebooks and when considering more descriptors
- The retrieval performance of our method is better than the one of both approximate k-means and ensembles of extremely randomized trees
- We are also more efficient than the competition
 - 24.5 times faster than k-means
 - 1.6 times faster than AKM

Supervised Codebook Construction





Supervised Codebook Construction





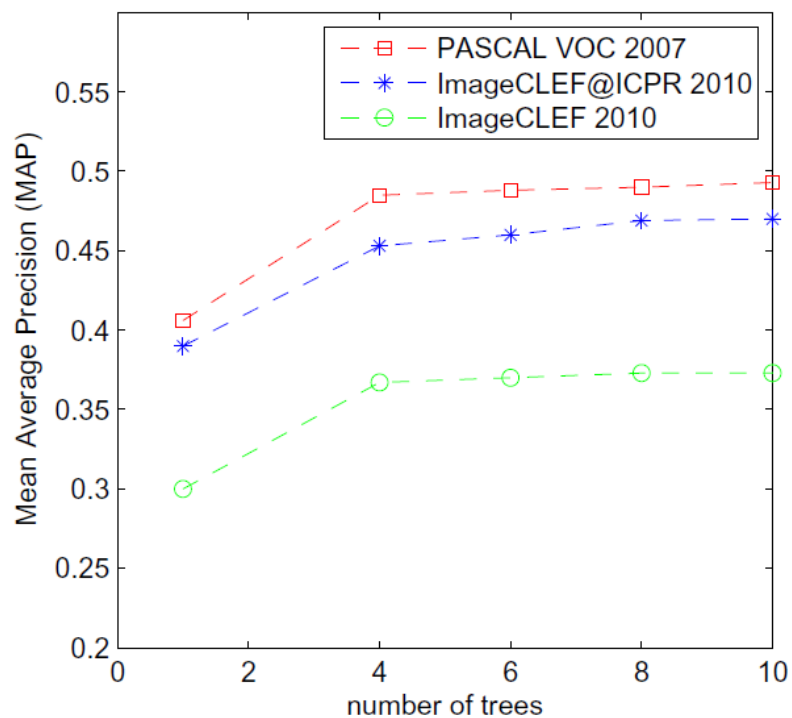
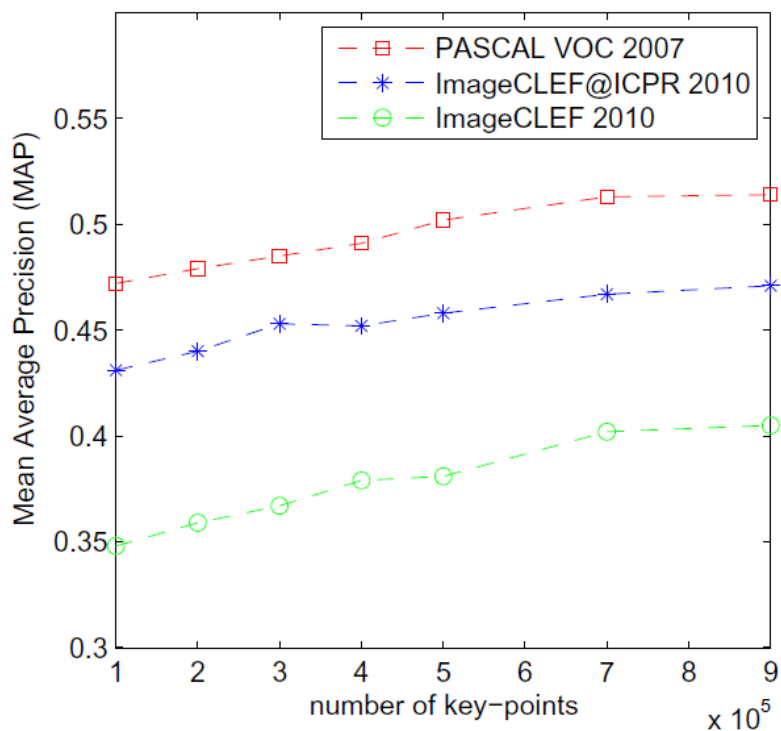
Multi-Label Image Annotation

Image database	Efficiency [s]		Performance [MAP]	
	<i>k</i> -Means	PCTs for MLC	<i>k</i> -Means	PCTs for MLC
PASCAL VOC 2007	12334.820	456.114	0.477	0.485
ImageCLEF@ICPR 2010	11977.230	466.829	0.425	0.453
ImageCLEF 2010	11209.750	544.740	0.329	0.367

- The visual codebook constructed with random forests of PCTs for MLC outperforms the one constructed with *k*-means on all three databases: It is more discriminative
- The improvement is larger for the databases with a larger average number of labels per image
- Dimitrovski et al., Pattern Recognition Letters 2013

Codebooks Learnt from more KPs have Better Performance

- PCTs for MLC are ~40 times more efficient than k-means
- Codebooks using larger number of key-points can be constructed
- Codebooks of 4000 words, diff. no. of KPs, diff. no. of trees in forest





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





Thank you ...

- For your attention.
- Questions welcome!



© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

TARNOWSKI