Remote Sensing Group

Unsupervised and Self-taught Learning for Remote Sensing Image Analysis

Ribana Roscher

Institute of Geodesy and Geoinformation,

Remote Sensing Group, University of Bonn

igg

universitätbonn

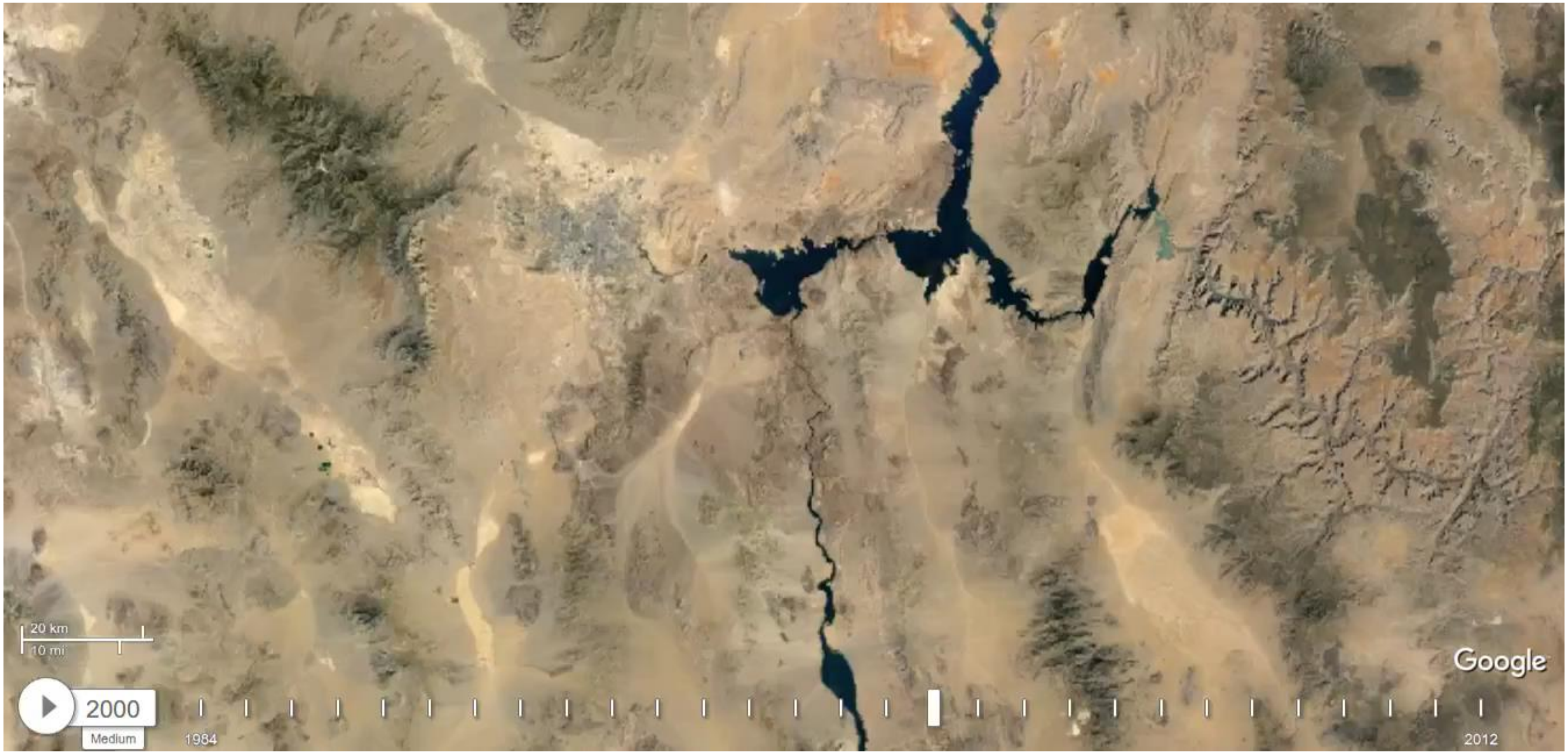
The Changing Earth

https://earthengine.google.com/timelapse/

The Changing Earth

https://earthengine.google.com/timelapse/

The Changing Earth



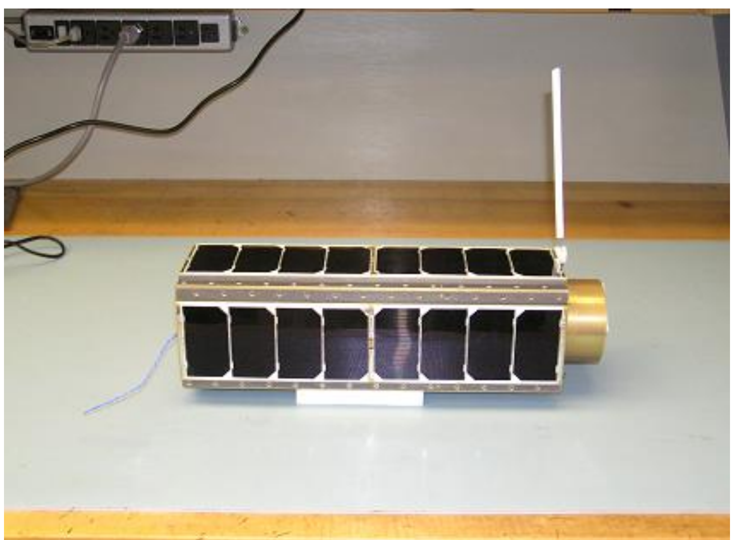
https://earthengine.google.com/timelapse/

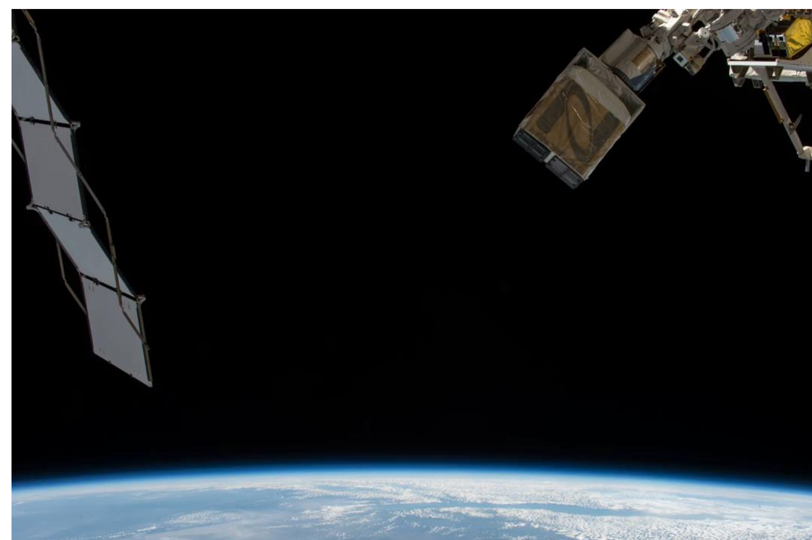
The Changing Earth

https://earthengine.google.com/timelapse/

Monitoring the Change

- High-resolution monitoring possible due to projects such as CubeSats
- International program to bring microsatellite into the orbit
- Statistics mid 2017: over 600 CubeSat missions, 165 active at the moment





Dates: https://sites.google.com/a/slu.edu/swartwout/home/cubesat-database#defs
Images: http://tia.arc.nasa.gov/genesat1/systemsSummary.html

Challenges

- Amount of data (**v**olume)
- Permanent change makes monitoring difficult (**v**elocity)
- Various data sources are not combinable in a trivial way (**v**ariaty)
- Data uncertainty (**v**eracity)

➡ typical Big Data challenges

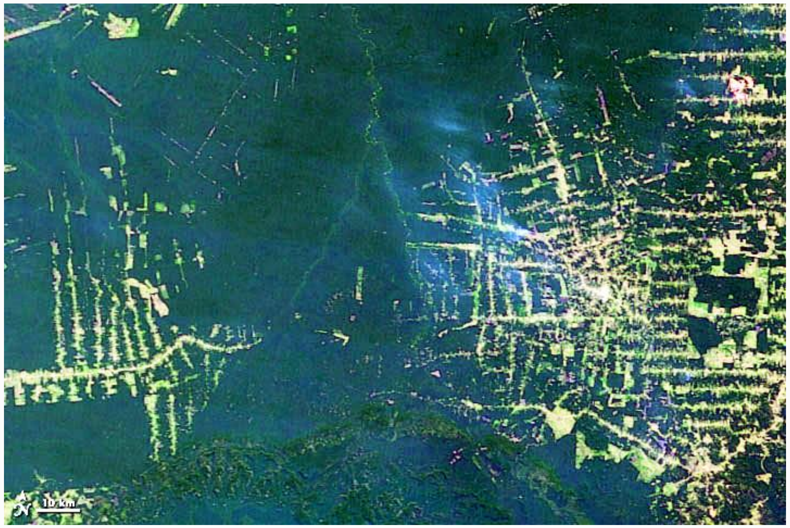
Remote Sensing Tasks

- Self-taught learning for **classification**
- Sparse representation-based spectral clustering for **change detection**
- Archetypal analysis for **unmixing**

Self-taught Learning for Classification

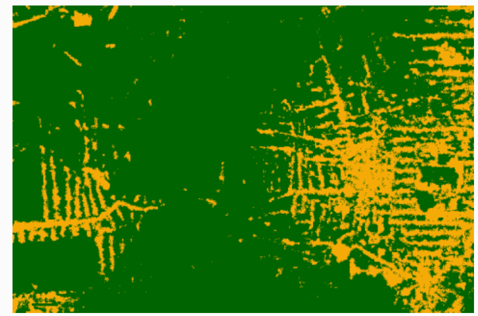
Classification Task

Processed satellite images

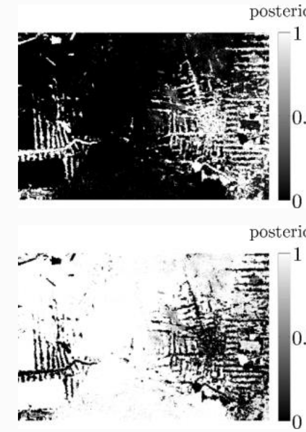


- Pixel with class information (labeled)
- Pixel without class information (unlabeled)

Land use and land cover map



Land use and land cover map



Posterior probabilities

Feature learning

Classification

- Learning step
- Testing step

Evaluation + Post-processing

Paradigms

- Supervised learning
- Semi-supervised learning
- Unsupervised learning
- Self-taught learning

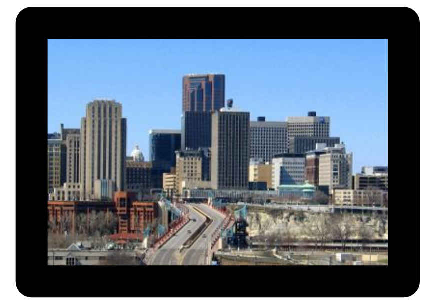
- Other approaches
  - Transfer Learning/Domain adaptation
  - ...

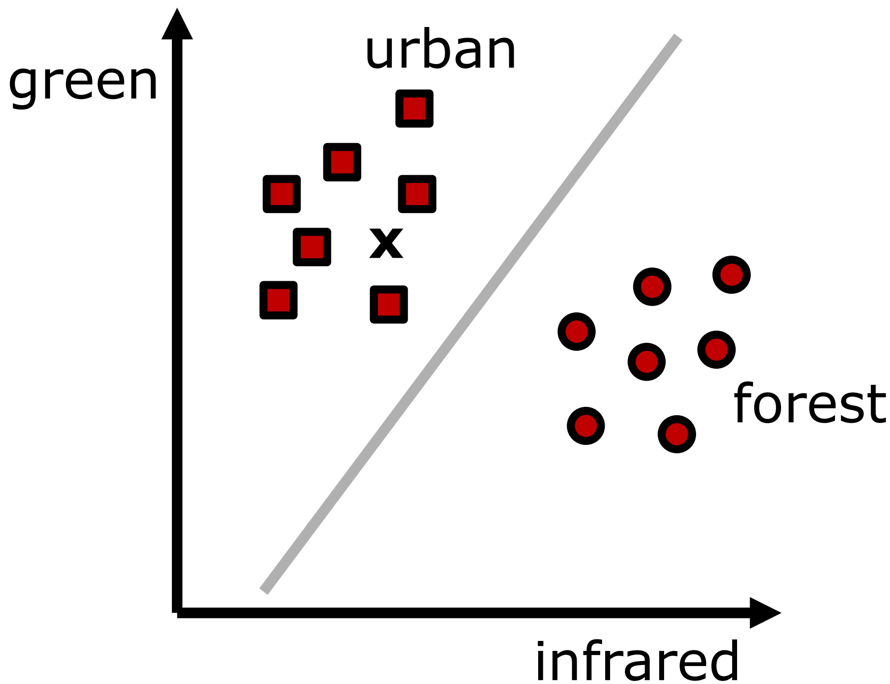
Classification Paradigms

Supervised learning



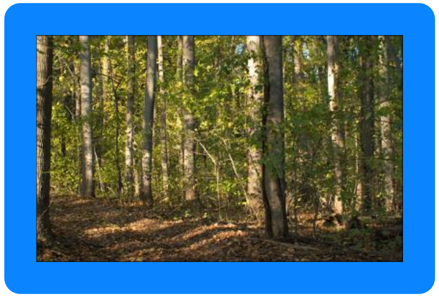


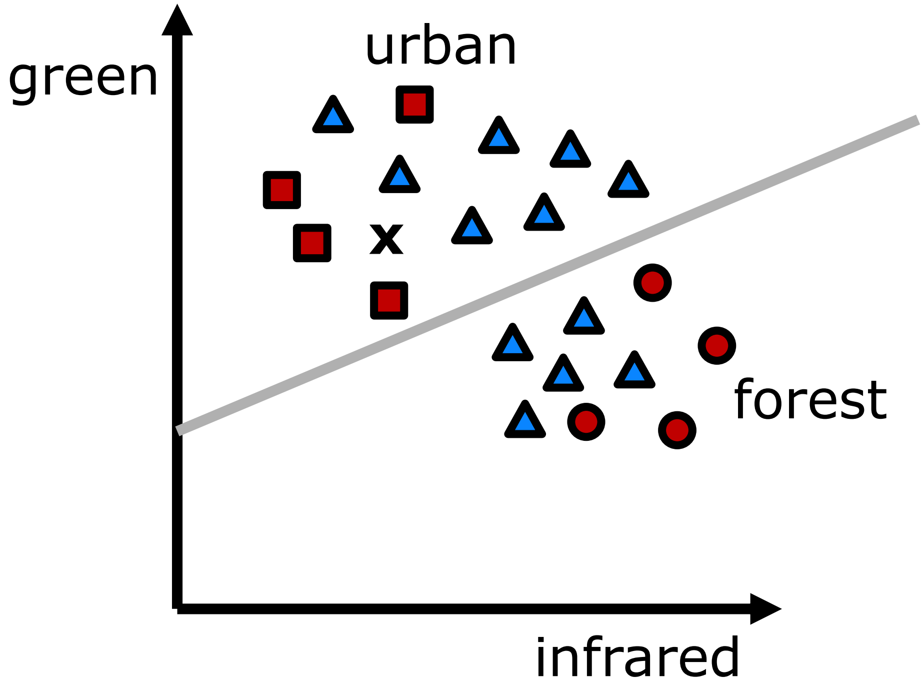




Classification Paradigms

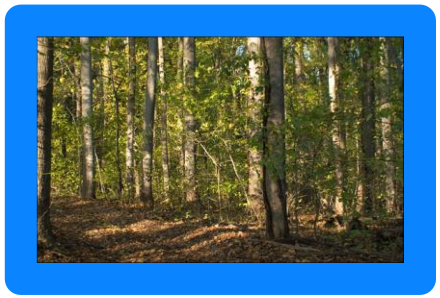
Semi-supervised learning

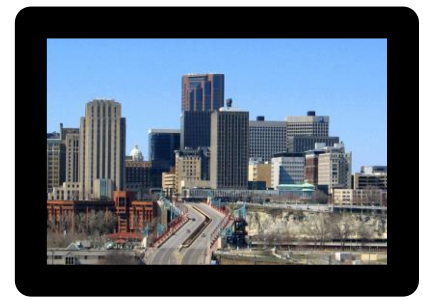


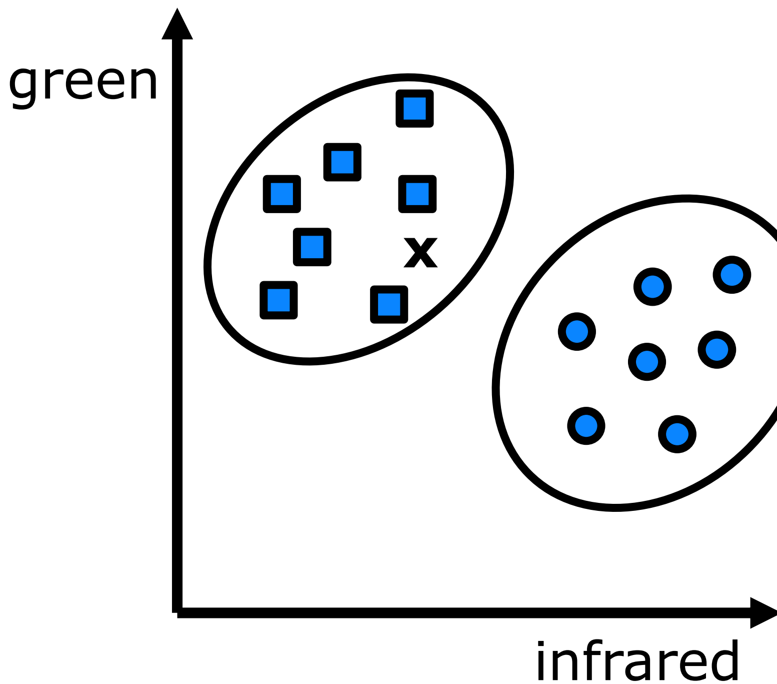


Classification Paradigms

Unsupervised learning

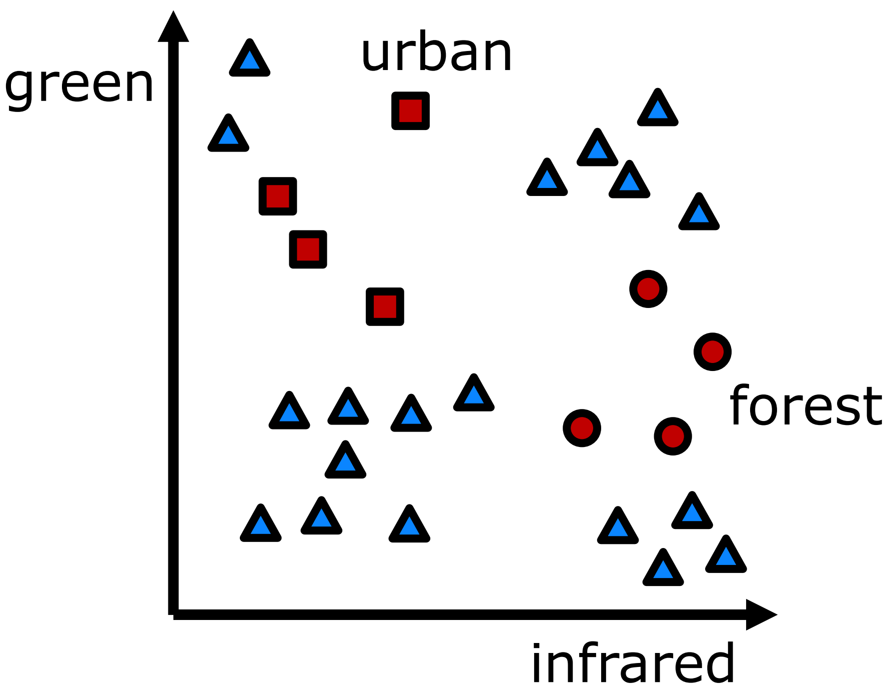





Classification paradigms

Self-taught learning



Feature/Representation Learning

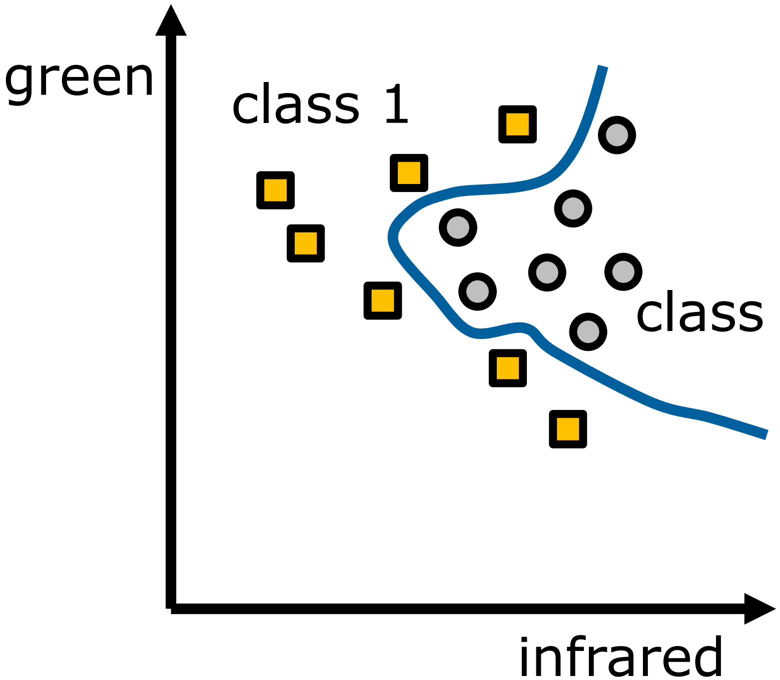
Learning a **new data representation** which is more suitable for classification than the original data representation

Powerful feature representation

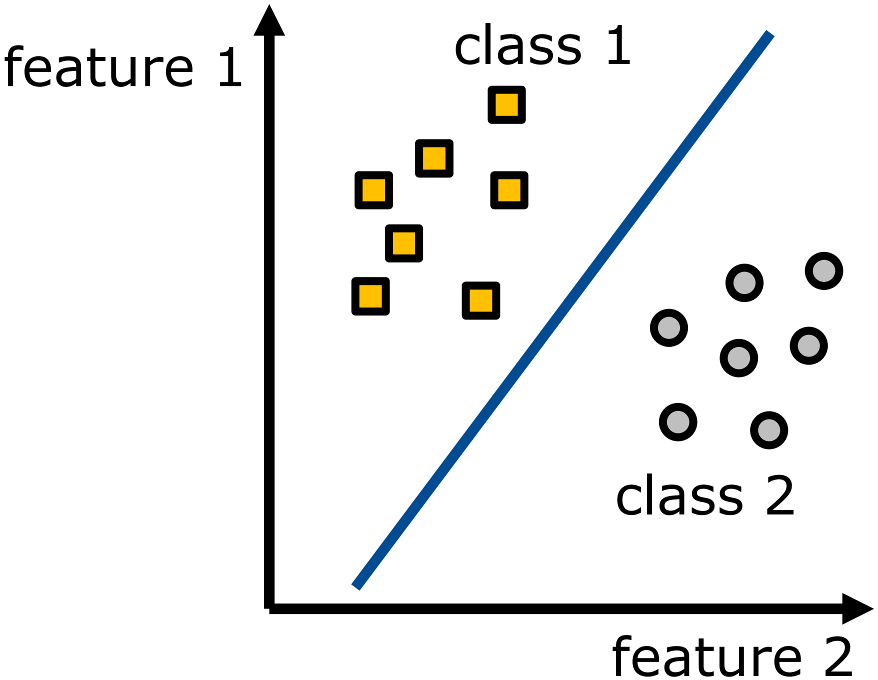
- Discriminative
- Robust
- Lower complexity
- Easier to interpret

Feature Learning

Original representation



Discriminative representation

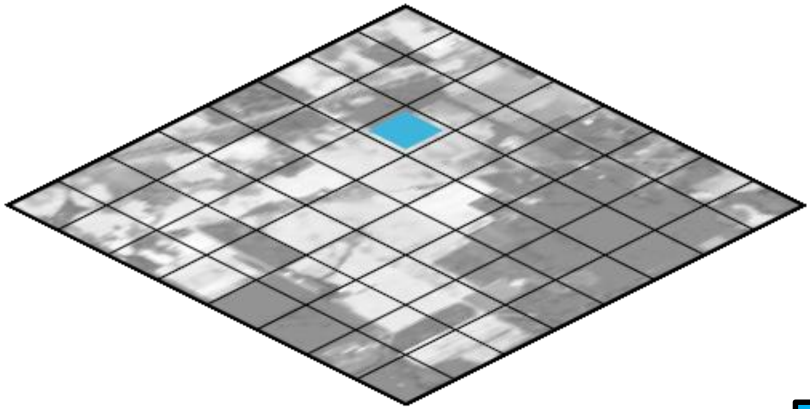


Feature Learning

Learning a new data representation which is more suitable for classification than the original data representation

- Unlabeled data is used in a self-taught learning framework to learn this representation
- Most common approach to self-taught learning is **sparse representation**

Sparse Representation

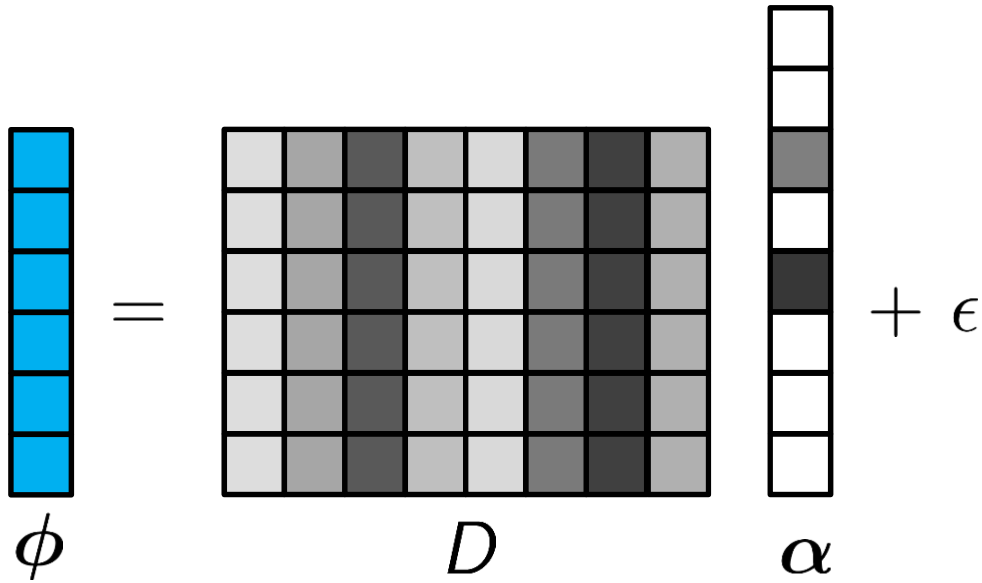


ϕ : pixel/image patch (original representation)

D : dictionary

$\boldsymbol{\alpha}$: sparse activation vector (new representation)

$\|\boldsymbol{\epsilon}\|$: reconstruction error

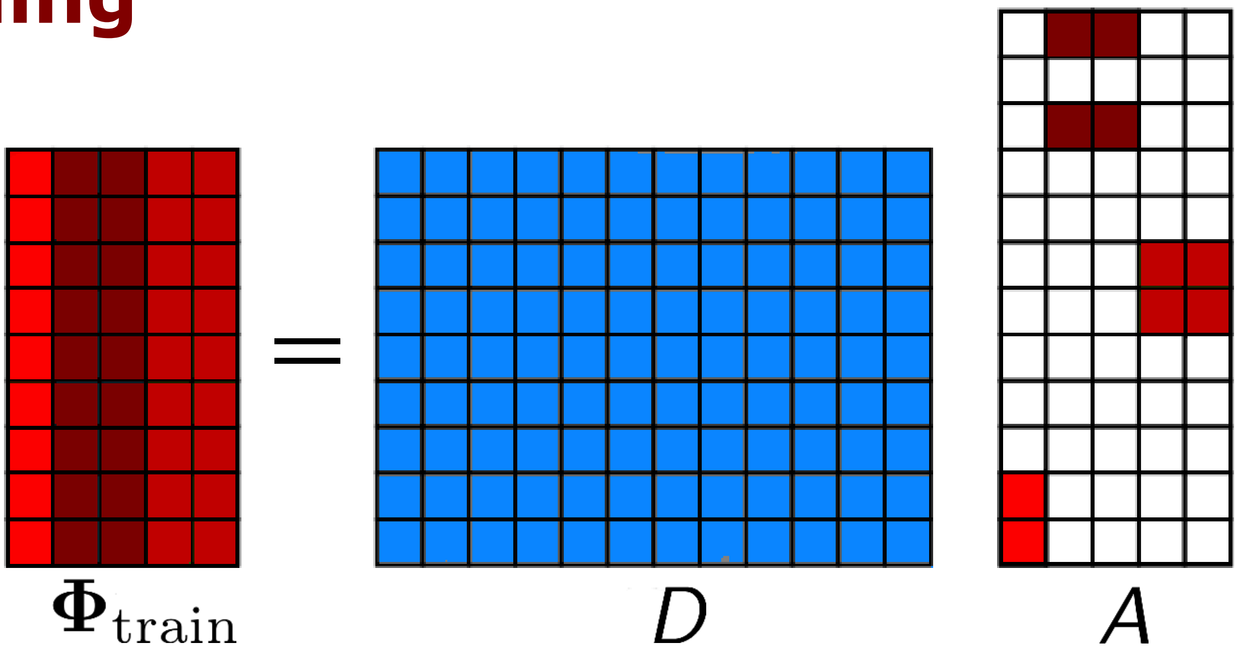


$$\widehat{\boldsymbol{\alpha}} = \mathrm{argmin}\|D\boldsymbol{\alpha} - \phi\| \qquad \text{s.t.} \qquad \|\boldsymbol{\alpha}\|_0 < W$$

$$\widehat{\boldsymbol{\alpha}} = \mathrm{argmin}\|D\boldsymbol{\alpha} - \phi\| \qquad \text{s.t.} \qquad \boldsymbol{\alpha} \succeq \mathbf{0}$$

Self-taught Learning

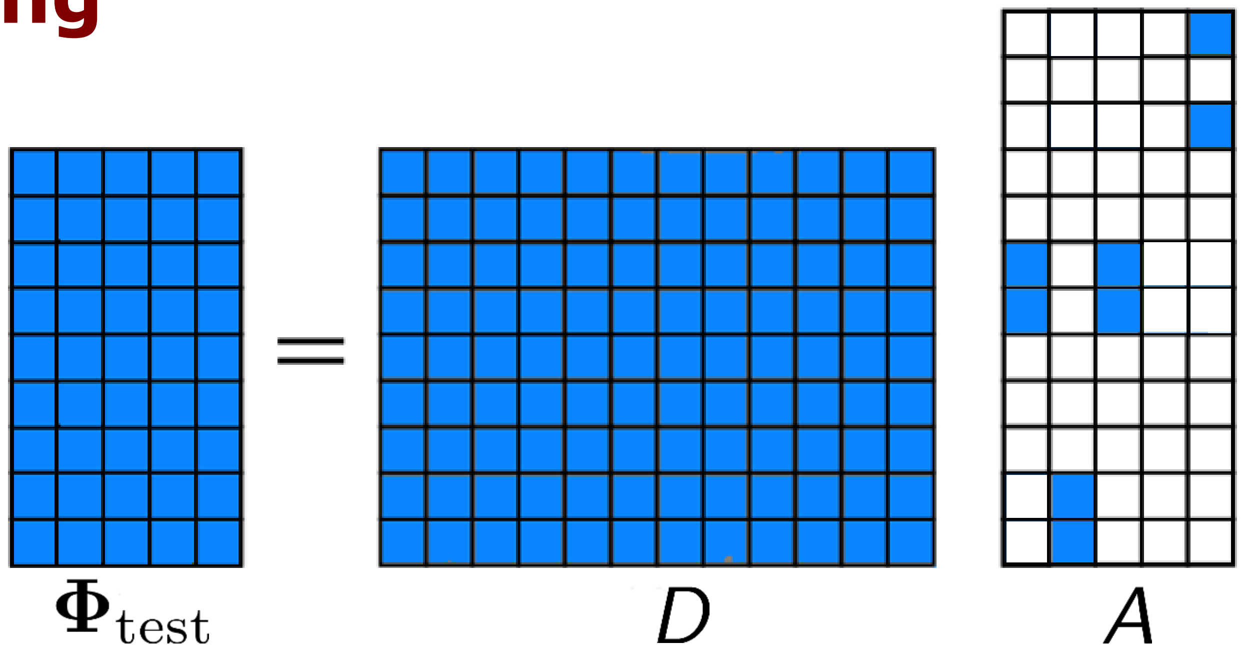
training



- Dictionary contains unlabeled data
- Assumption: samples of the same class are reconstructed with a similar set of dictionary elements and similar weights
- Goal: new representation is highly discriminative

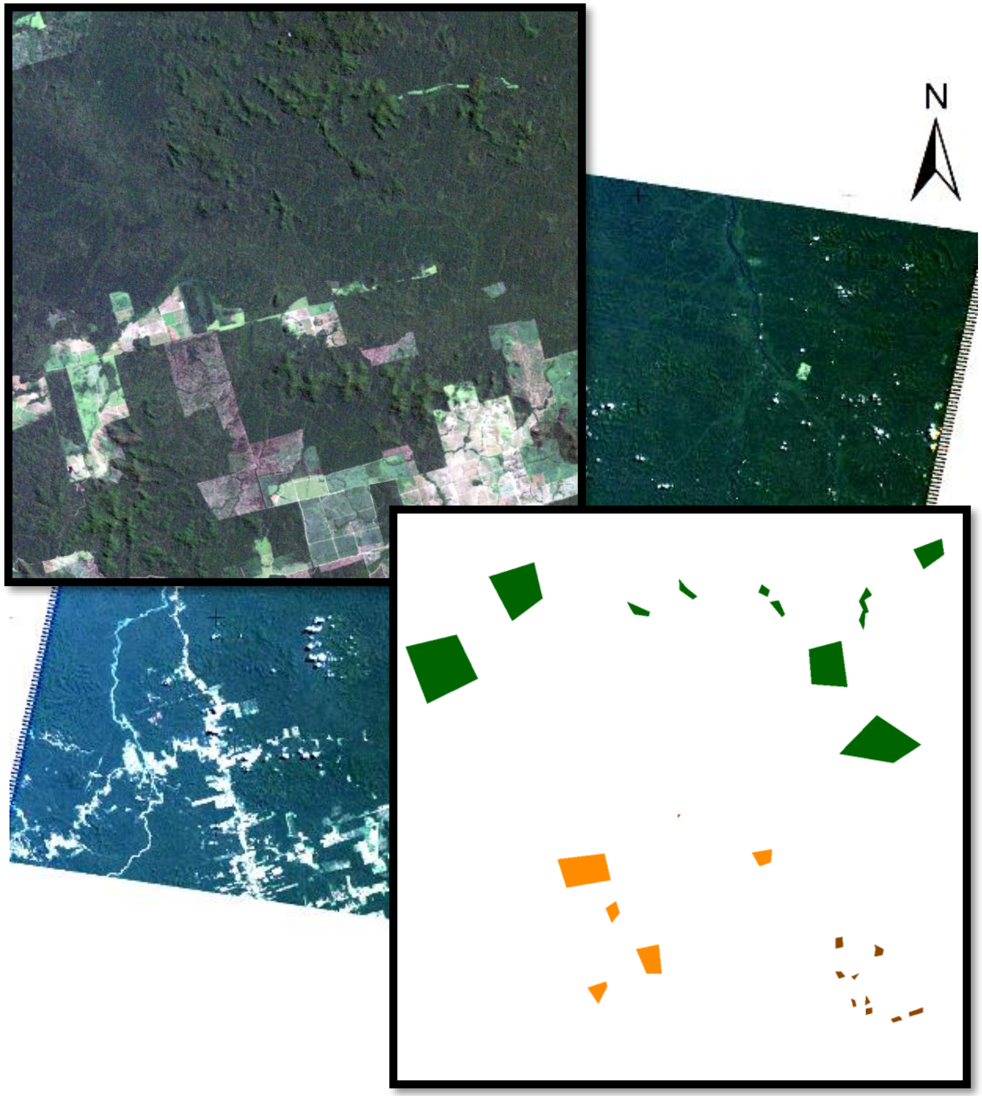
Self-taught Learning

testing



- Dictionary is fixed
- Classify is trained and tested with new representation

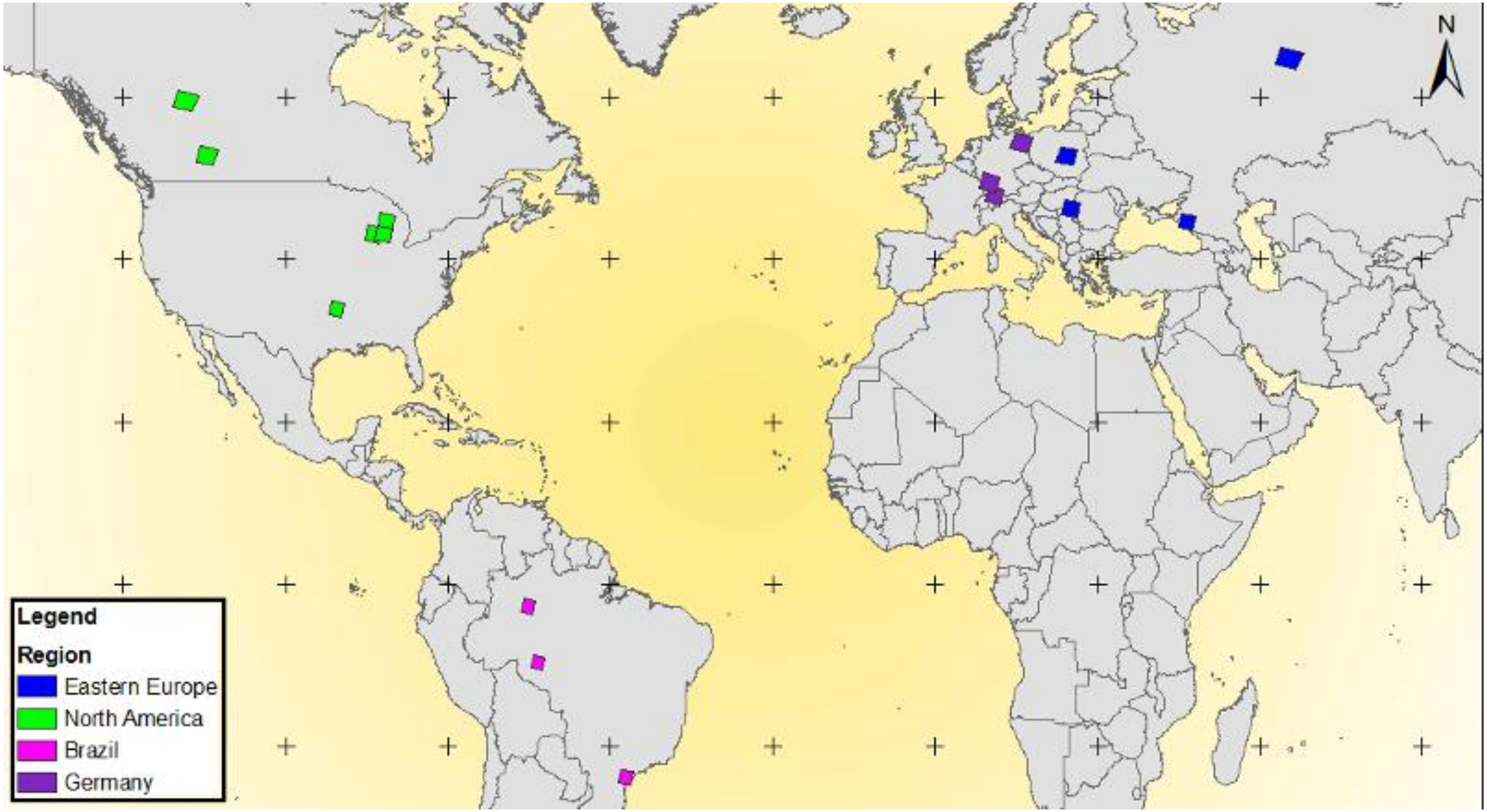
Data Set



- Landsat 5 TM image near Novo Progresso (Brazil)
- Ca. 8000x8000 pixel
- 30x30m spatial resolution
- Area characterized by fire clearing
- Reference information: Forest, deforestation (fire clearing) and arable land
- Subarea: ~900km^2

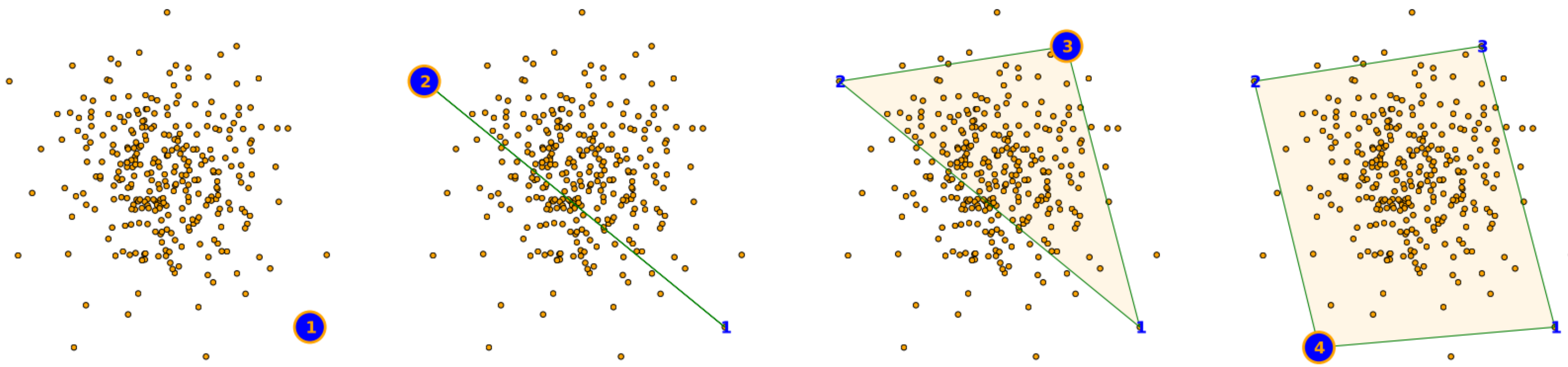
Dictionary Elements

- ~1 Mio. image patches



Archetypal Analysis: SiVM

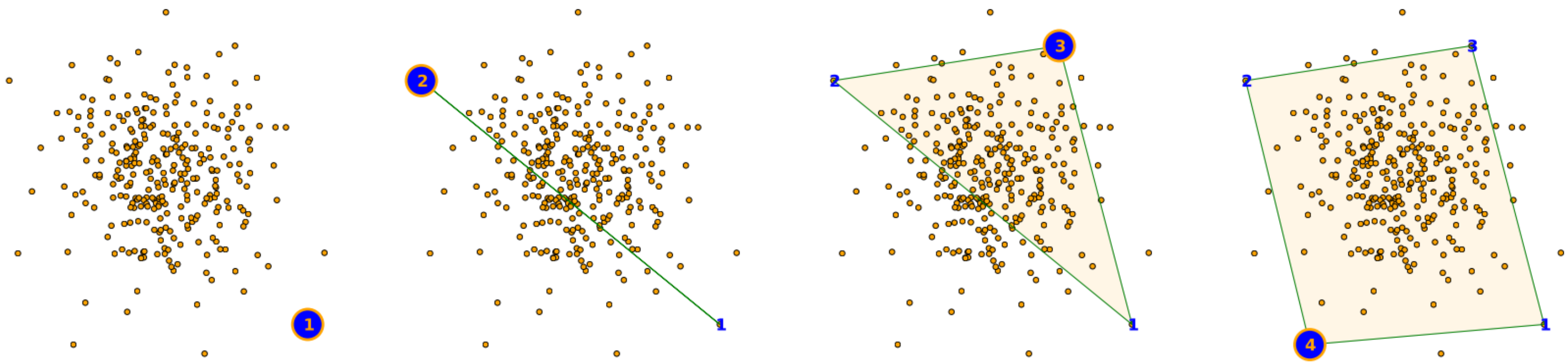
- Archetypal analysis finds the extreme points (archetypes) in feature space
- Efficient determination by **Simplex Volume Maximization (SiVM)**
- Assumption: Convex hull consists of points, which maximize the volume



Christian Thurau, Kristian Kersting, Christian Bauckhage (2010): Yes We Can – Simplex Volume Maximization for Descriptive Web-Scale Matrix Factorization

Archetypal Analysis: SiVM

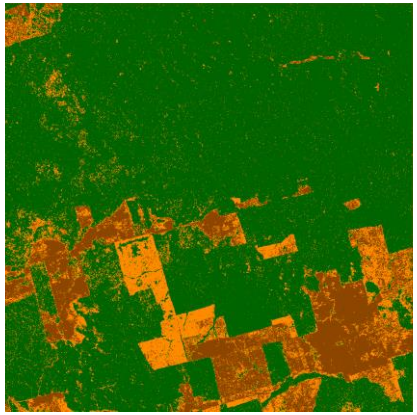
1. Randomly choose (virtual) starting point
2. Choose sample which is farthest away
3. Set this sample as first archetype
4. Choose next sample which is farthest away from all previous archetypes



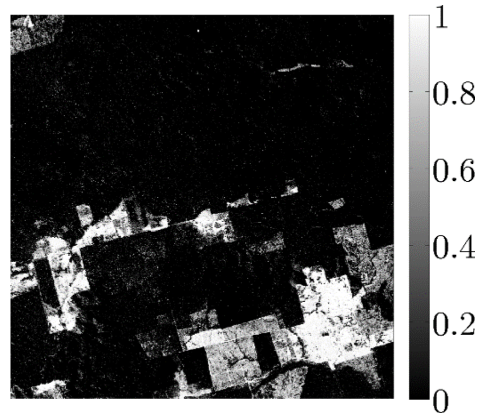
Christian Thurau, Kristian Kersting, Christian Bauckhage (2010): Yes We Can – Simplex Volume Maximization for Descriptive Web-Scale Matrix Factorization

Self-taught Learning Results

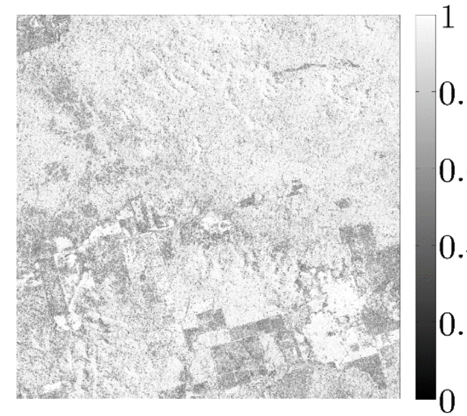
Satellite image



Land cover



Posterior probability: arable



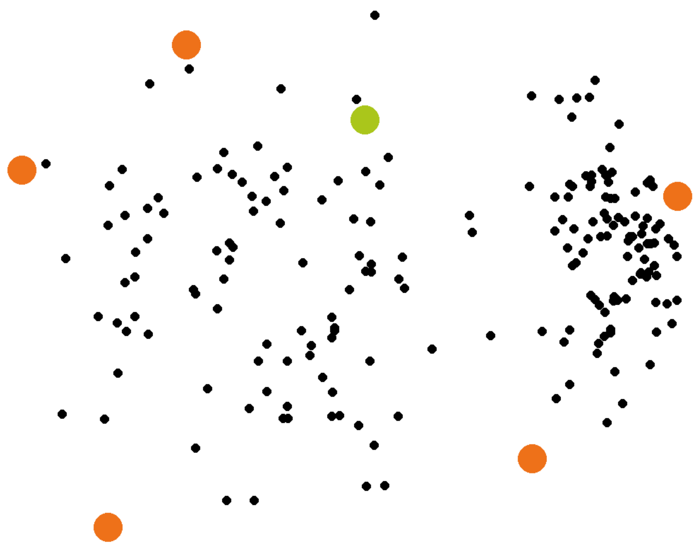
Maximum posterior probability (certainty)

	Original		STL-KSVD		STL-Archetypes	
	K-SVM	LR	K-SVM	LR	K-SVM	LR
Arable	**85.6%**	81.3%	82.7%	83.7%	84.8%	84.0%
Deforestation	80.2%	76.8%	84.0%	82.9%	83.9%	**86.0%**
Forest	98.3%	98.4%	**98.5%**	98.0%	98.2%	98.2%
oa	89.5%	87.4%	90.1%	89.7%	90.5%	**91.0%**
aa	88.1%	85.5%	88.4%	88.2%	89.0%	**89.4%**
Kappa	0.84	0.81	0.80	0.84	0.85	**0.86**

Roscher, R., Römer, C., Waske, B., Plümer, L. (2015). Landcover Classification with Self-taught Learning on Archetypal Dictionaries, IGARSS, Symposium Paper Prize Award

Archetypal Dictionaries

Challenge: Set of archetypes depends on initial point



- Highly variable in high dimensions
- Highly variable if data is normalized (e.g. global contrast normalization)

Archetypal Dictionaries

- Finding the best set of archetypes regarding specific criteria by minimizing

$$U(D) = -\log(e) + \|\gamma\|_2$$

discriminative part (logistic regression CV error)

reconstructive part (reconstruction error)

- Challenge: elements unknown + number of elements unknown

➢ **R**eversible **j**ump **M**arkov **c**hain **M**onte **C**arlo

R. Roscher, S. Wenzel, and B. Waske, "Discriminative Archetypal Self-taught Learning for Multispectral Landcover Classification," in *Proc. of Pattern Recogniton in Remote Sensing 2016 (PRRS), Workshop at ICPR; to appear in IEEE Xplore* , 2016.

Reversible Jump Markov Chain Monte Carlo

Advantages

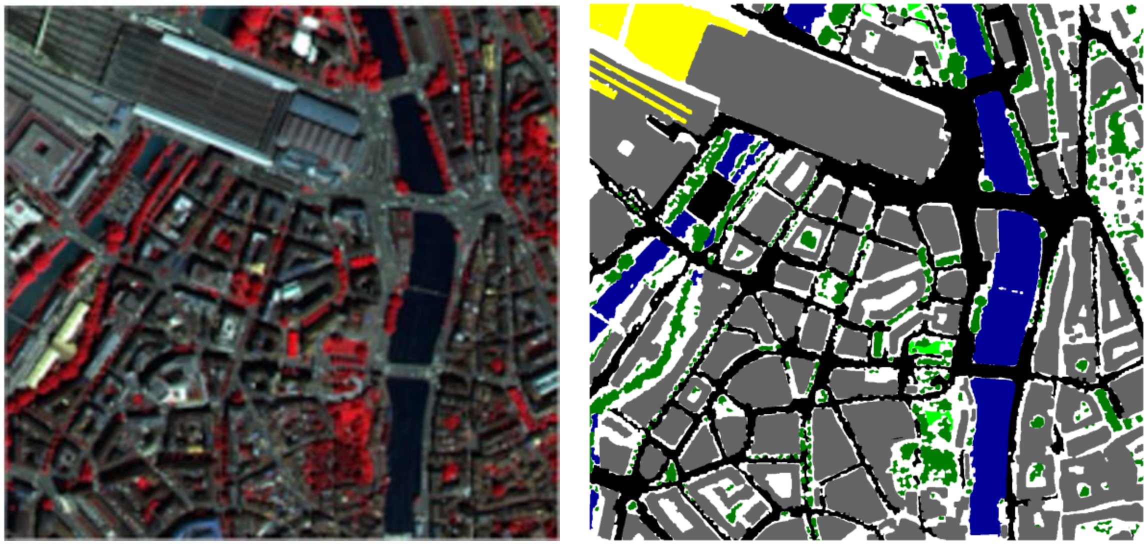
- Finds global optimum

Drawbacks

- Computation of discriminative part and sparse representation is slow

Discriminative STL

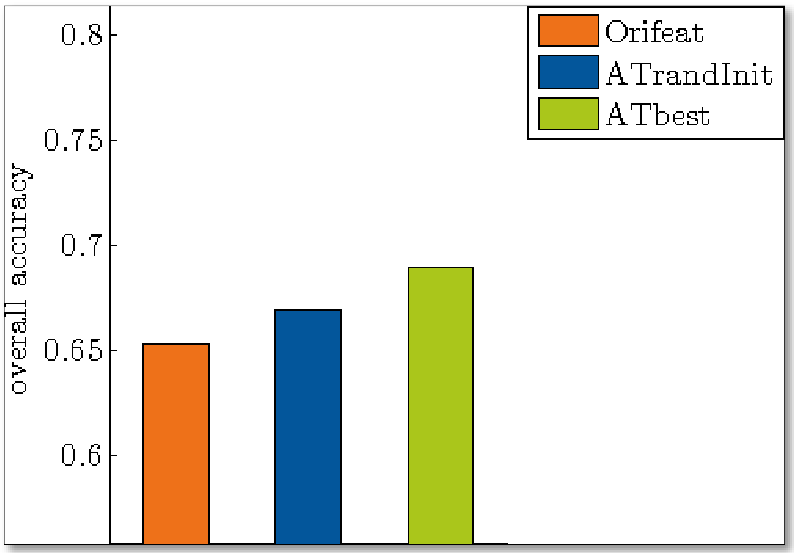
Zurich data set

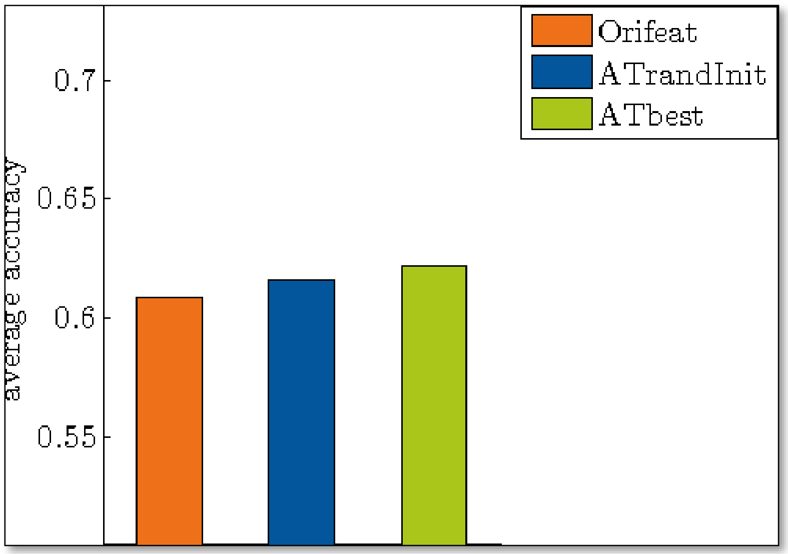


- 20 VHR multispectral images acquired by Quickbird sensor (0.61m/pixel, R-G-B-NIR)
- 8 land cover classes
- Image patches of size 5x5 pixel
- Evaluation by leave-one-out estimation

Discriminative STL

Zurich data set





- Average number of used dictionary elements is 22 with a standard deviation of approximately 6 elements

Roscher, R., Wenzel, S., Waske, B. (2016). Discriminative Archetypal Self-taught Learning for Multispectral Landcover Classification, *Proc. of Pattern Recogniton in Remote Sensing, Workshop at ICPR; to appear in IEEE Xplore*

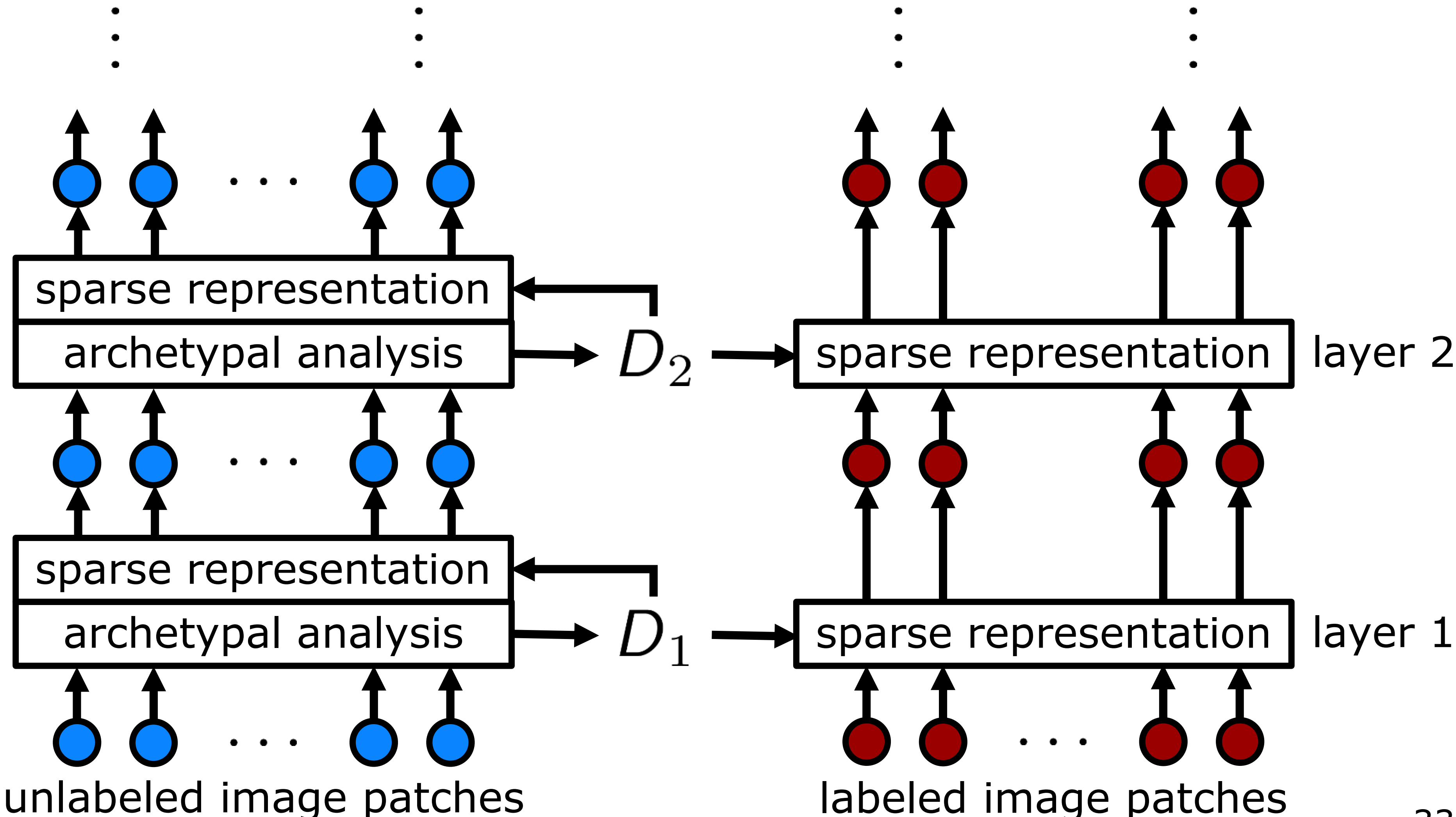
Going Deep (Ongoing Research)

- Deep self-taught learning with sparse representation

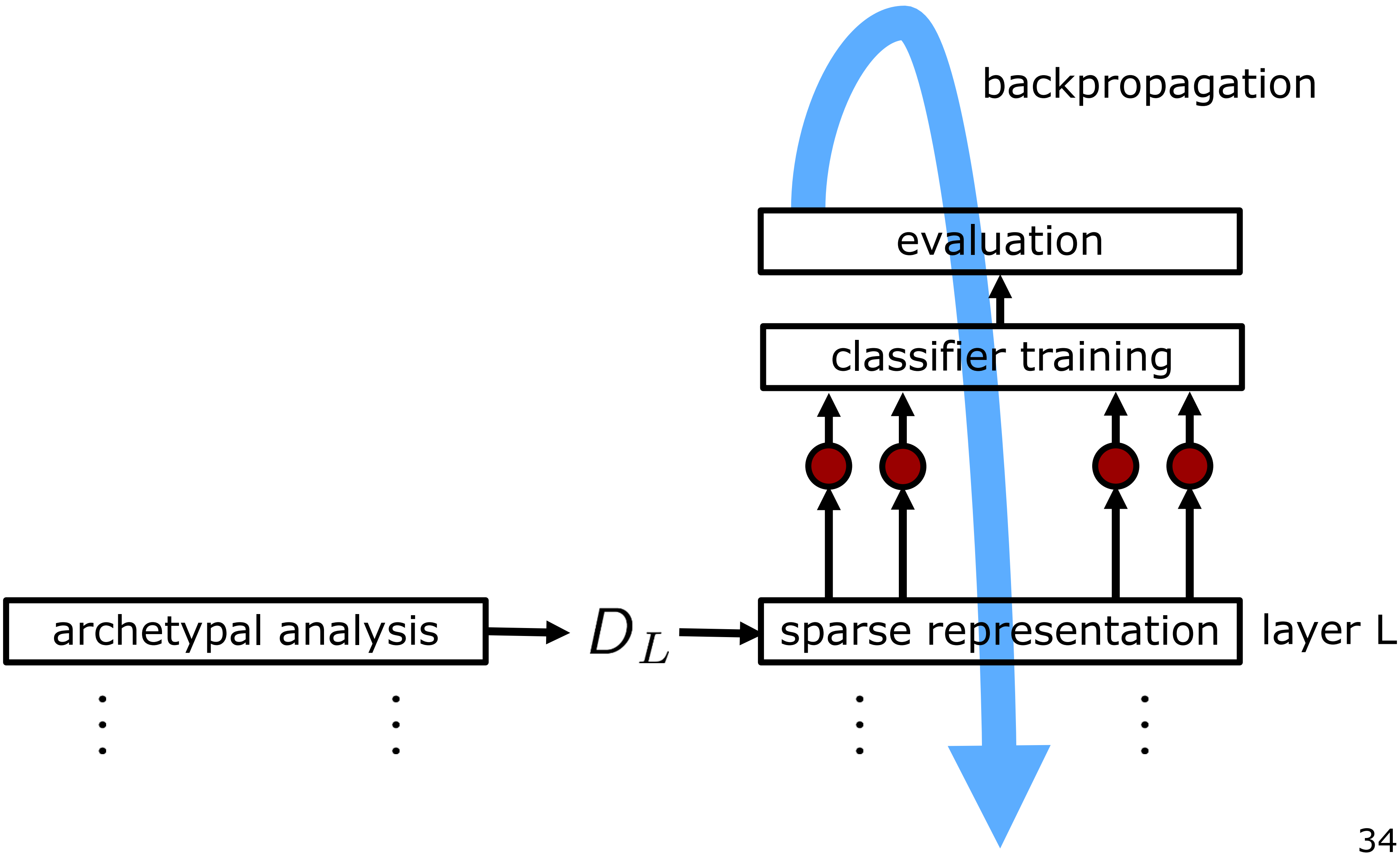
$$\begin{aligned} \phi &= D_1 \boldsymbol{\alpha} \\ \boldsymbol{\alpha}' &= D_2 \boldsymbol{\beta} \\ \boldsymbol{\beta}' &= D_3 \gamma \\ &\vdots \end{aligned}$$

- Output from a previous layer serves as input the next layer
- Feature representation in last layer used for classification

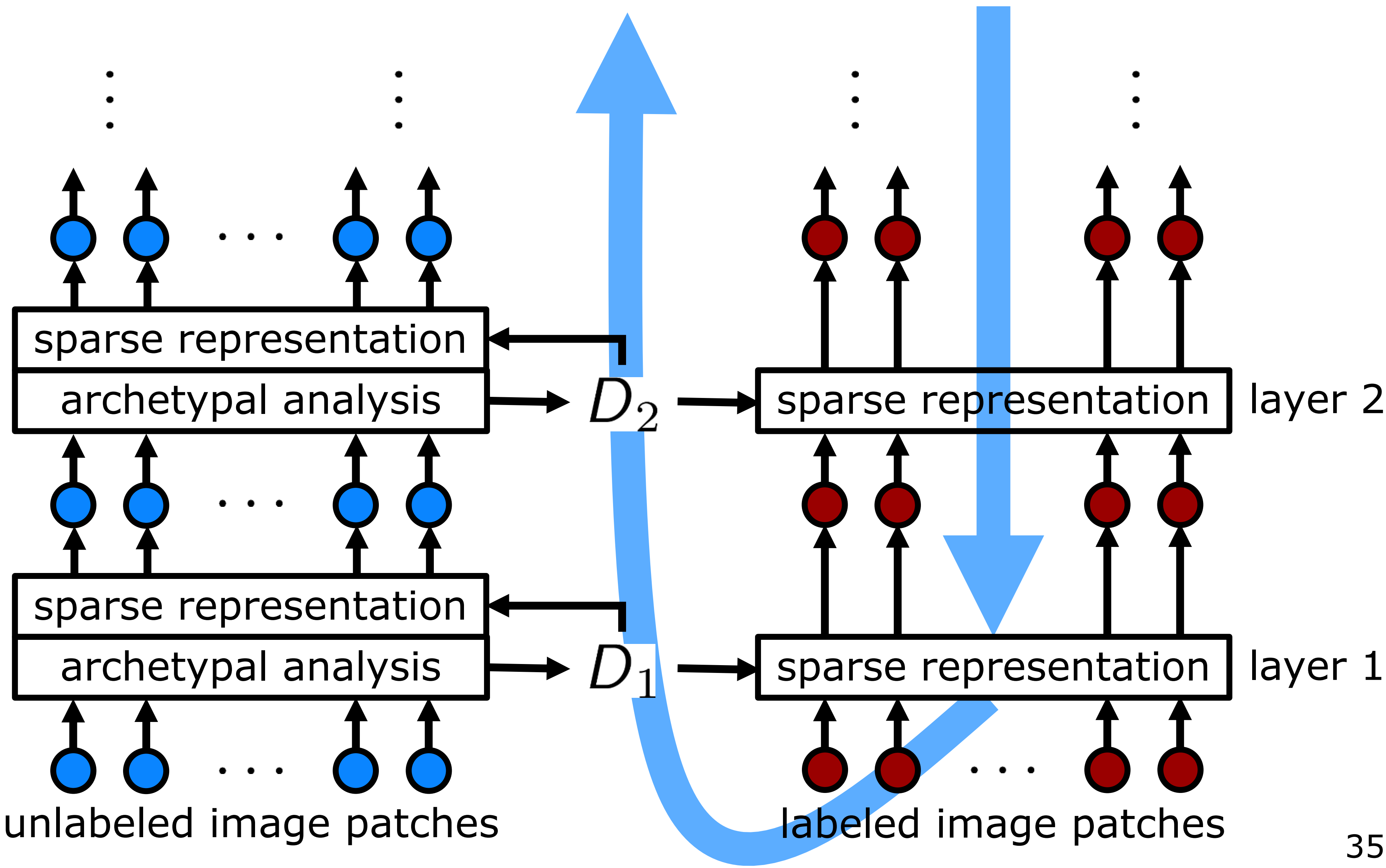
Deep Self-taught Learning



Deep Self-taught Learning



Deep Self-taught Learning



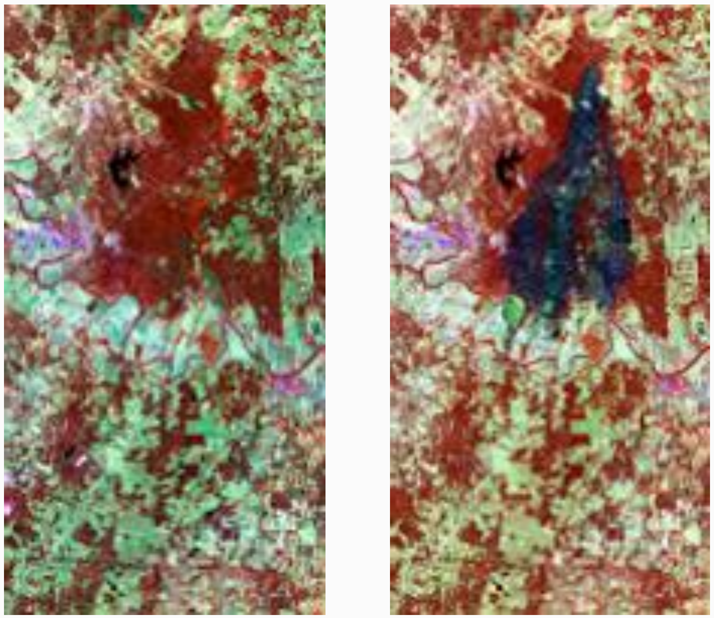
Self-taught Learning - Résumé

- Self-taught learning with sparse representation can find a discriminative feature representation
- Archetypal dictionaries are undercomplete, yet powerful
- Initialization of archetypal analysis influences the classification success
- Extension to Deep STL promising
  - All activations can be interpreted as mixings of archetypes
  - Deeper layers are deeper mixings

Sparse Representation-based Spectral Clustering for Change Detection

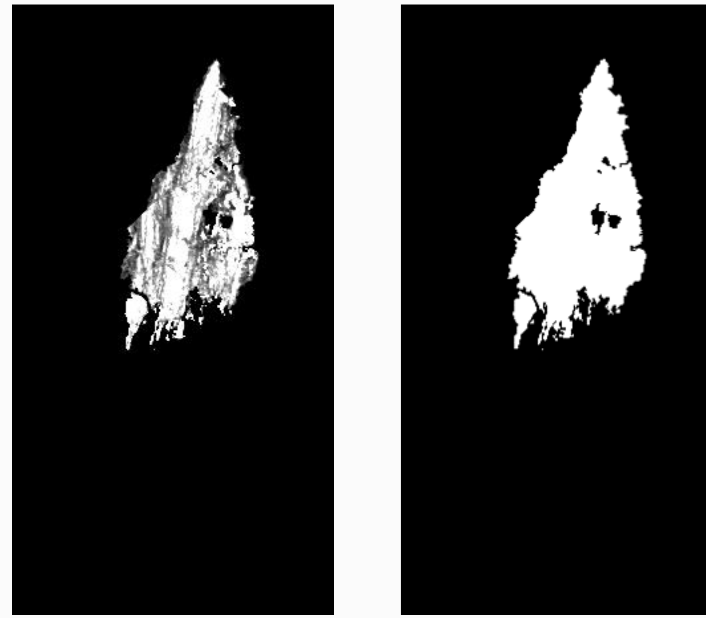
Change Detection Task

Processed satellite images



- 2 images: pre-event and post event

Change maps



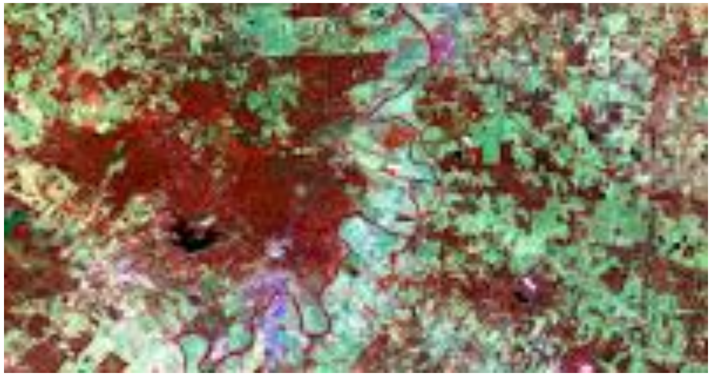
amount of change

binary

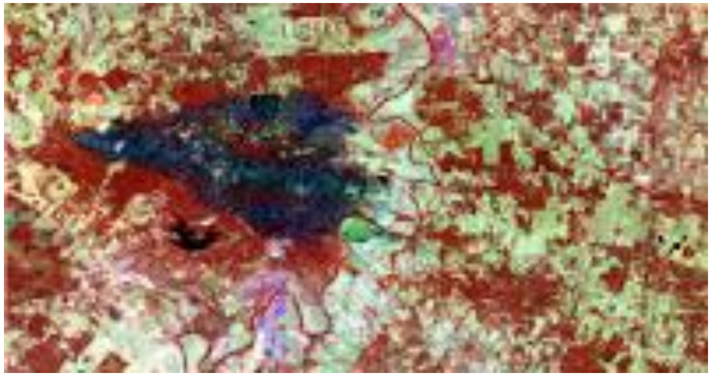
Feature extraction → **Clustering + change assigment** → **Evaluation + Post-processing**

Data set

Bastrop fire dataset (Landsat 5 TM)



Pre-event image



Post-event image

Challenges

- No label information
- Spectral differences due to changing weather conditions, atmospheric conditions, seasonal effects...

Spectral Clustering

Spectral clustering performs clustering on the singular vectors to the smallest singular values derived from a unnormalized Graph Laplacian

$$L = D - W \qquad\qquad D = \mathrm{diag}\left(\sum_m \boldsymbol{w}_m\right)$$

similarity/adjacency matrix

degree matrix

or normalized Graph Laplacian

$$L_{\mathrm{sym}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

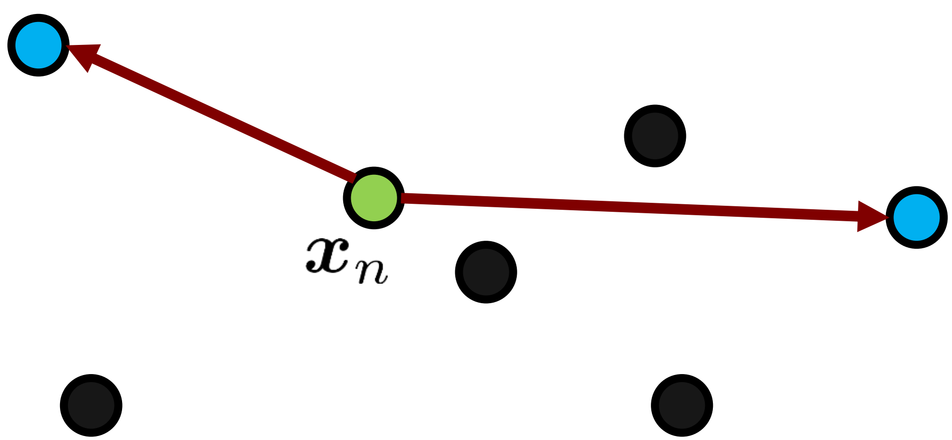
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395-416.

SR for Change Detection

- **Approach**: Clustering on stacked images
- Sparse representation is used to build a sparse adjacency graph W for **spectral clustering**

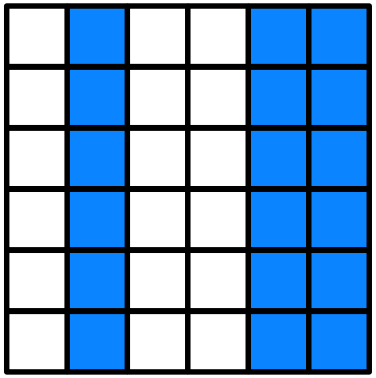
$$\hat{\boldsymbol{\alpha}}_n = \operatorname{argmin}_t \| T\boldsymbol{\alpha}_n - \boldsymbol{\phi}_n \|_2 \quad \text{subject to} \quad \boldsymbol{\alpha}_n \succeq \mathbf{0}$$

$$T = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{n-1}, \boldsymbol{\phi}_{n+1}, \ldots, \boldsymbol{\phi}_N]$$



Sparse Representation for Change Detection

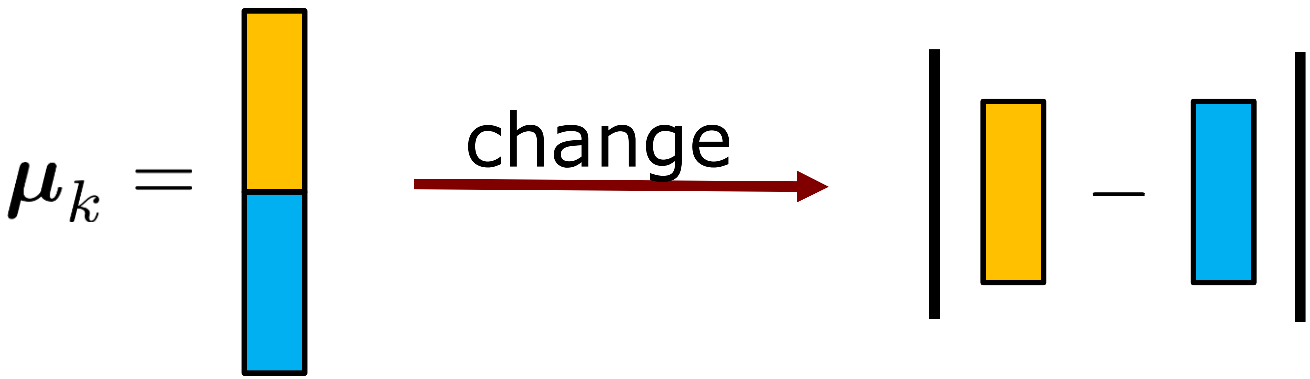
Sparse archetypal adjacency matrix



- Building a sparse representation-based graph is too computational intense
- Using landmarks = archetypes
- Nyström method for large data sets

Change assignment

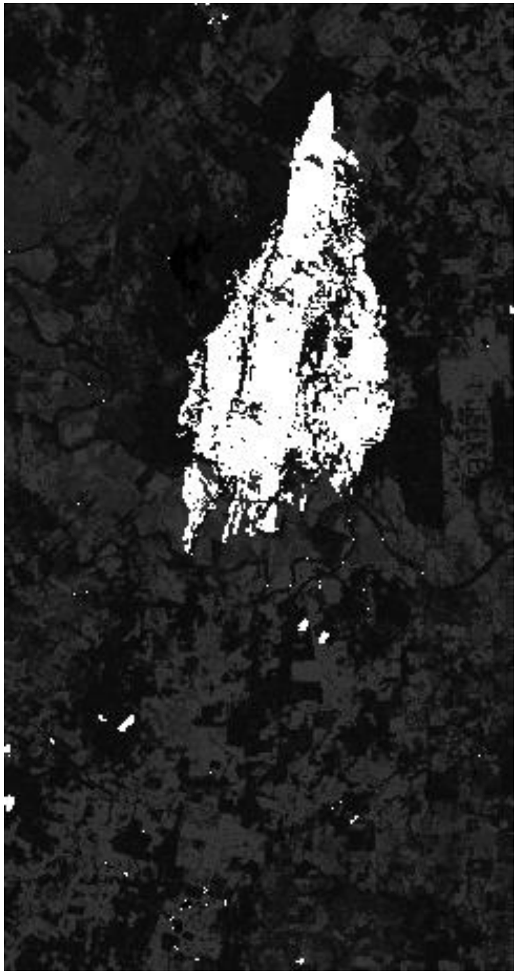
- Change in each cluster is derived from the means obtained from k-means



- Change of cluster mean is assigned to whole cluster

Results

Ground truth

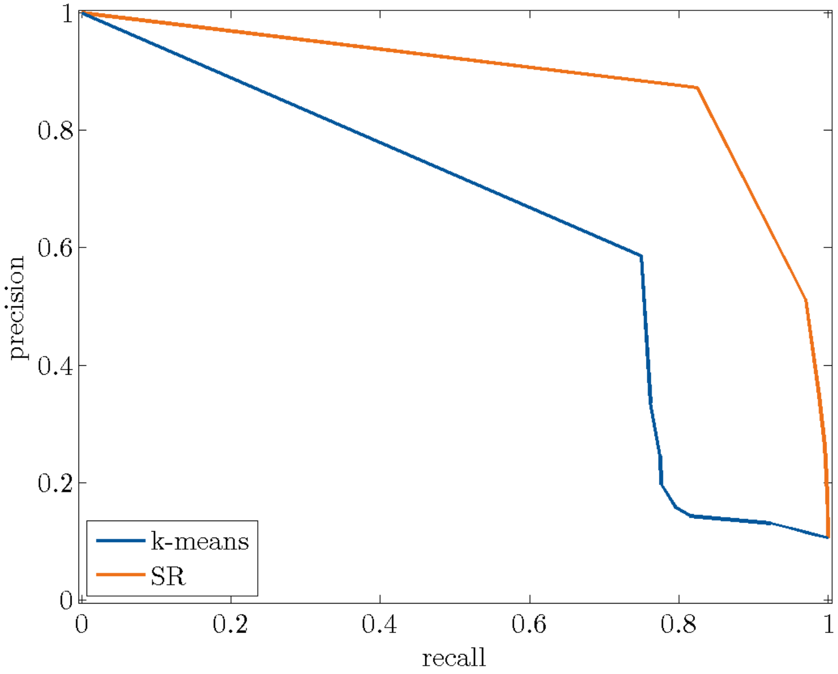


K-means

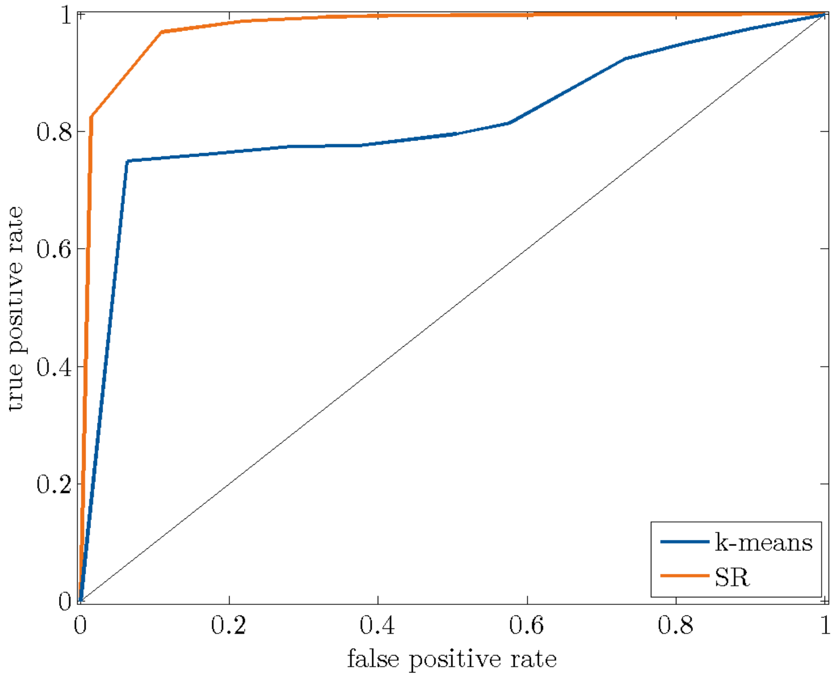
Spectral clustering

Roscher, R., Drees, L., Wenzel, S. (2017). Sparse Representation-based Archetypal Graphs for Spectral Clustering, IGARSS, accepted

Sparse Representation for Change Detection



Precision-recall curve

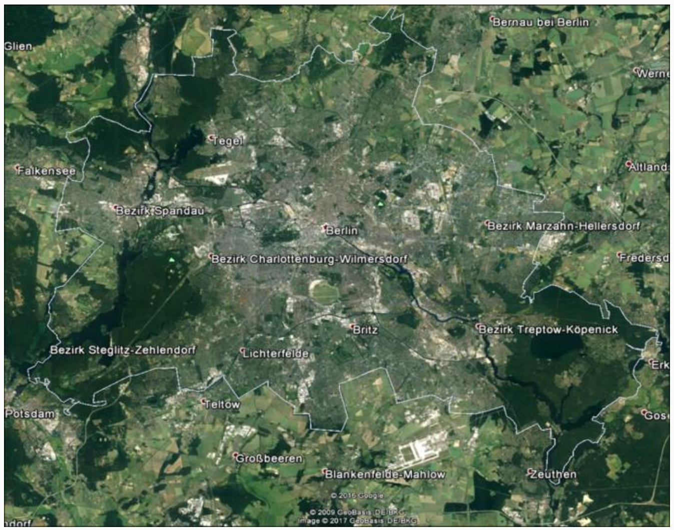


Receiver operating curve

Archetypal Analysis for Unmixing

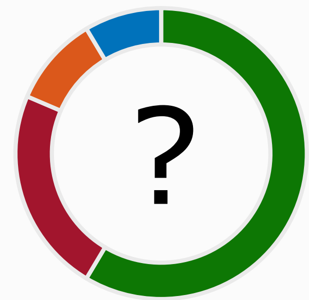
Unmixing Task

Processed satellite image



- Pixel with class information (labeled)
- Pixel without class information (unlabeled)

Sub-pixel quantification



Endmember extraction

- Manually or
- Automatically

Reconstruction by sparse representation

Evaluation

Unmixing Task

1. task: Find suitable endmembers
   - manually derived spectral library
   - archetypal dictionary
2. task: Estimate fractions (activations)
   - Sparse representation

$$\widehat{\boldsymbol{\alpha}} = \operatorname{argmin} \| D\boldsymbol{\alpha} - \phi \| \quad \text{s.t.} \quad \boldsymbol{\alpha} \succeq \mathbf{0}, \ \sum_t \alpha_t = 1$$

Data

Berlin-Urban-Gradient dataset 2009

http://dataservices.gfz-potsdam.de/enmap/showshort.php?id=escidoc:1823890&show_gcmdcitation=false

Data

Study site: Southwest of Berlin

Hyperspectral image

Manually derived spectral library

Reference land cover information

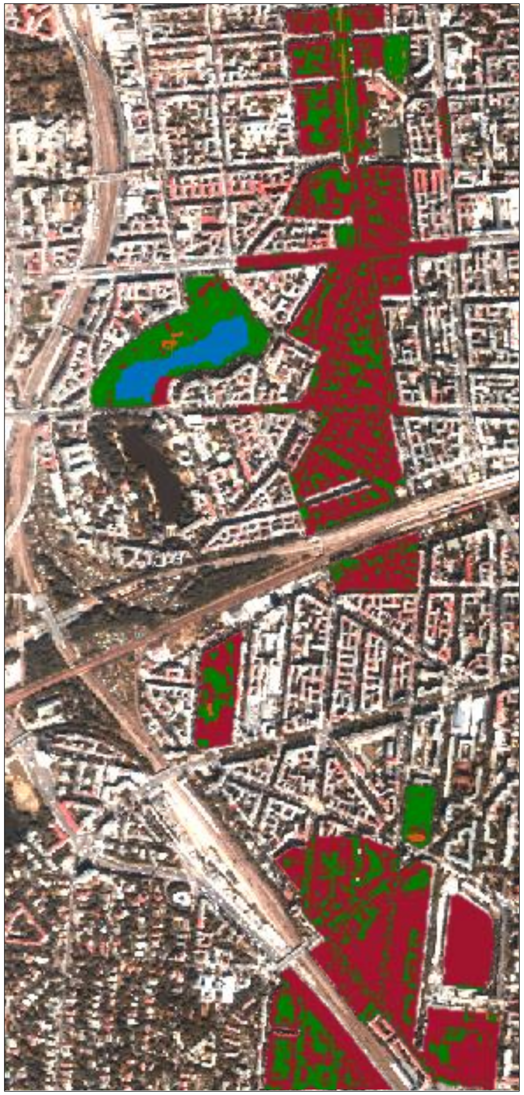
Simulated EnMAP scene

http://dataservices.gfz-potsdam.de/enmap/showshort.php?id=escidoc:1823890&show_gcmdcitation=false

Hyperspectral Data

- Airborne Sensor: HyMap
- 111 spectral bands
- Observed wavelength 450nm – 2500nm
- Spatial resolution of 3.6m
- Visualized as RGB-image with the wavelengths R=640nm, G=540nm and B=450nm

Reference Information



- Reference information was manually obtained
  - digital orthophotos
  - cadastral data
- 4 land cover classes
  - Impervious surface
  - Vegetation
  - Soil & Sand
  - Water

Simulated EnMAP Data

- **Simulated EnMAP** scene of the same area
- Spatial resolution of 30m
- 1495 EnMAP pixels were obtained from the simulation tool, containing the fractions of the land cover classes ranging from 0 to 100%

➢ Task: Reconstruction of fractions of simulated EnMAP data

Archetypal Dictionary vs. Manually Derived Spectral Library

- Archetypal dictionaries were interpreted using reference data

	Archetypal dictionary	**Manually derived library**
Imp. Surface	25	39
Vegetation	12	31
Soil	2	4
Water	1	1
Σ	**40**	**75**

- High total amount of spectra in the manually derived spectral library

Evaluation

		Archetypal dictionary	Manually derived library
$\lVert \epsilon \rVert$		1.1	**0.0**
MAE [%]	**Imp. Surface**	**12.2**	16.0
	Vegetation	11.0	**9.2**
	Soil	2.5	**2.1**
	Water	**1.9**	12.2
	Ø	**6.8**	9.9

- High number of elementary spectra in library results in a small reconstruction error
- All dictionaries achieve similar and satisfactory solutions

Drees, L. and Roscher, R. (2017). Archetypal analysis for sparse representation-based hyperspectral sub-pixel quantification, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-1/W1, 133-139

Summary

- Exploitation of unlabeled samples for learning
  - Self-taught learning
  - Unsupervised learning
- Sparse representation is a versatile tool
- More and more research goes into the direction of unsupervised pre-training in combination with supervised learning

Nyström Method

SVD

$$L_{\mathrm{sym}} = \begin{bmatrix} \boxed{W} & L_{\mathrm{sym},12} \\ L_{\mathrm{sym},21} & L_{\mathrm{sym},22} \end{bmatrix} \qquad C = \begin{bmatrix} W \\ G_{21} \end{bmatrix}$$

- Low rank approximation of Gram matrix

$$G \approx \widetilde{G} = C W_k^{+} C^{\mathsf{T}}$$

Pseudo-inverse of low rank approximation of W

- Singular values and vectores

$$\tilde{S}_k = \frac{N}{K} S_{W,k} \qquad \tilde{U}_k = \sqrt{\frac{K}{N}} C U_{W,k} S_{W,k}^{+}$$