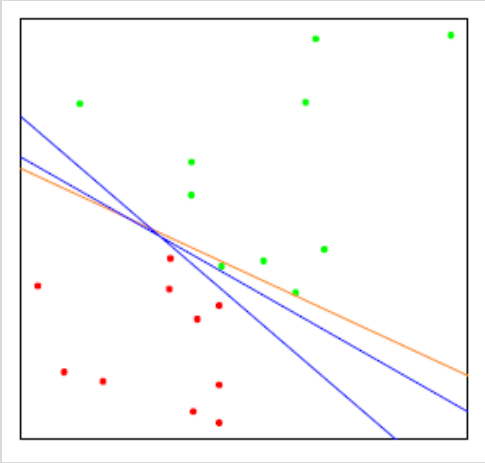


Parallel & Scalable Machine Learning

Introduction to Machine Learning Algorithms

Dr. – Ing. Morris Riedel
Adjunct Associated Professor, University of Iceland
Research Group Leader, Juelich Supercomputing Centre, Germany
Dr. Gabriele Cavallaro
Post Doctoral Researcher, Juelich Supercomputing Centre, Germany

LECTURE 6

Classification Applications, Challenges, and Solutions

January 16th, 2018
JSC, Germany

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH FORSCHUNGSZENTRUM

PRACE PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE
TRAINING PORTAL

Review of Lecture 5

- Supervised Classification
  - Finding a target distribution for realistic learning situations
  - Assume unknown probability distribution over the input space
  - Hypothesis search with M models and we pick one
- Statistical Learning Theory

Unknown Target Distribution

target function $f : X \to Y$ plus noise $P(y|\mathbf{x})$

(ideal function)

$\mathbf{x} = (x_1, \ldots, x_d)$ ← Probability Distribution $P \ on \ X$

Hypothesis Set $\mathcal{H} = \{h\}; \ g \in \mathcal{H}$

(set of candidate formulas)

Final Hypothesis $g \approx f$

'Approximately' 'Probably' 'Number of samples N important for learning'

$$\Pr \ [\ | \ E_{in}(g) - E_{out}(g) \ | > \epsilon \ ] \ <= \ 2Me^{-2\epsilon^2 N}$$

Learning Algorithm ('train a system') $\mathcal{A}$ → 'learn: get error smaller' $E_{in}(g) \approx 0$ 'generalize well for unseen data' $E_{out}(g) \approx E_{in}(g)$

(set of known algorithms)

$$\Pr \ [\ | \ E_{in}(g) - E_{out}(g) \ | > \epsilon \ ] \ <= \ 4m_{\mathcal{H}}(2N)e^{-1/8\epsilon^2 N}$$

Shift the view to the data and we exchange M with growth function that is indeed depending on N

Outline

Outline of the Course

1. Introduction to Machine Learning Fundamentals
2. PRACE and Parallel Computing Basics
3. Unsupervised Clustering and Applications
4. Unsupervised Clustering Challenges & Solutions
5. Supervised Classification and Learning Theory Basics
6. Classification Applications, Challenges, and Solutions
7. Support Vector Machines and Kernel Methods
8. Practicals with SVMs
9. Validation and Regularization Techniques
10. Practicals with Validation and Regularization
11. Parallelization Benefits
12. Cross-Validation Practicals

Day One – beginner

Day Two – moderate

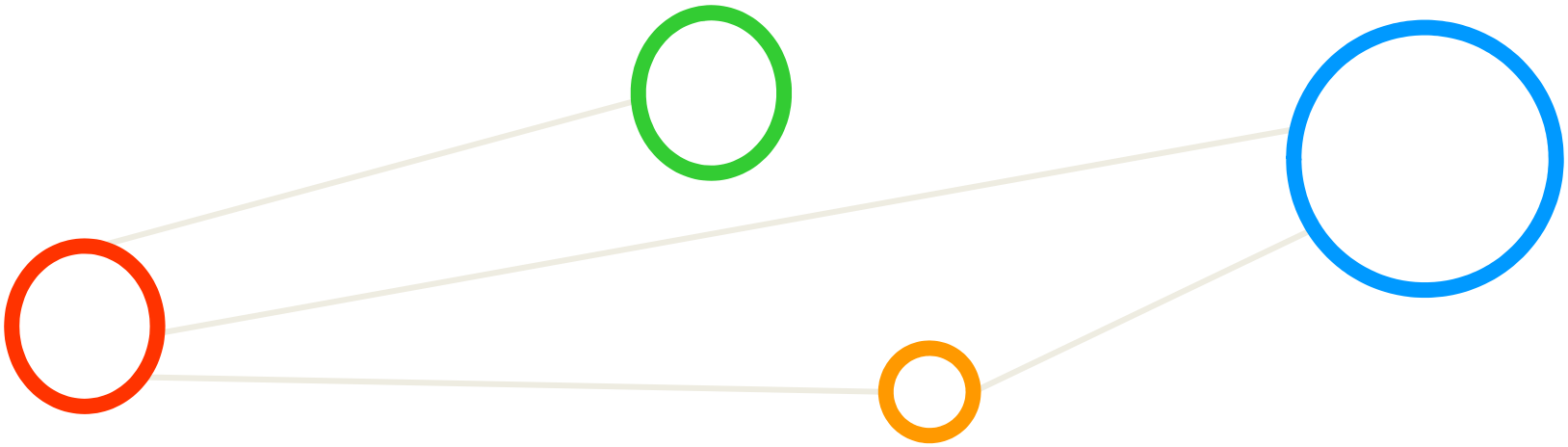
Day Three – expert

Outline

- Remote Sensing Applications
  - Introduction to Application Domain
  - Hyperspectral Datasets
  - Rome Dataset
  - Indian Pines Dataset
  - Explore need for Parallelization

- Classification Challenges & Solutions
  - Selected Key Challenges in Learning from Data
  - Review Practice Experience & Applications
  - Challenge One: Non-Linear Seperable Datasets
  - Pocket Algorithm as Extension to PLA
  - Challenge Two: Problem of Overfitting

Remote Sensing Applications



Tsunami December 2004

A devastating tsunami hit many coastal regions on the Indian Ocean

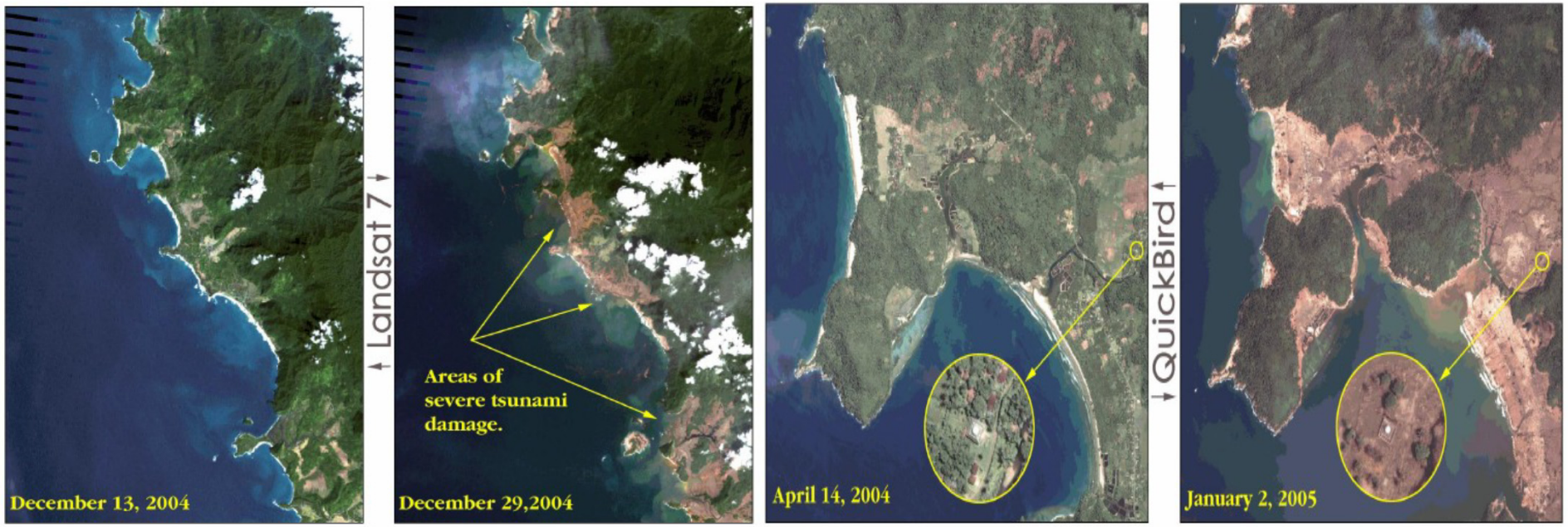
- On of the most devastating natural disasters in recorded history
- 14 countries were hit

[1] The 2004 Indian Ocean Tsunami

[2] Five years after Indian Ocean tsunami

- <u>How to assess the damage?</u>
- <u>Which are the most hit points?</u>
- <u>How to plan the Humanitarian aid?</u>

Tsunami December 2004



[3] Tsunami Damage, Gleebruk, Indonesia January 7, 2005

- Landsat 7
  - 185 Km swath width
  - Panchromatic: 15m
  - Multispectral (7 bands): 30m
  - Thermal band: 60m
  - Revisit interval: 16 days

[4] Landsat 7

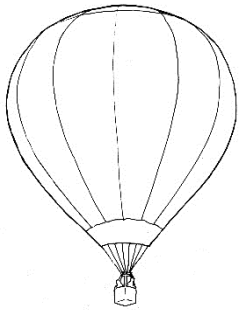
- Quickbird
  - 16.8 Km swath width
  - Panchromatic: 65cm
  - Multispectral (4 bands): 2.62m
  - Revisit time: 1-3.5 days

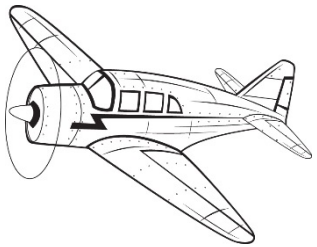
[5] DigitalGlobe's QuickBird satellite sensor

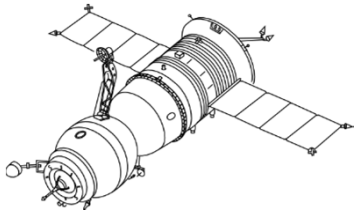
Definition and Milestones

Remote: without physical contact

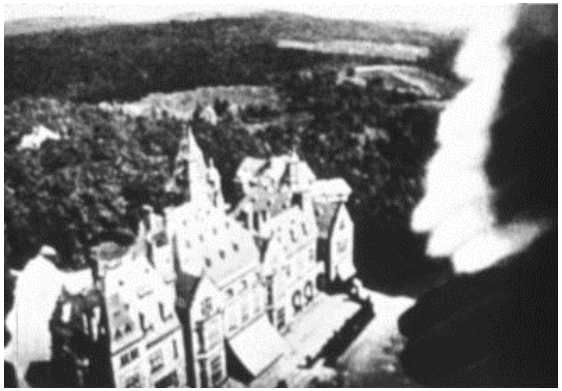
Sensing: measurement of information







[6] Balloons (1868)



[7] Pigeons (1903)

[8] Airplane (1909)

[9] Satellite

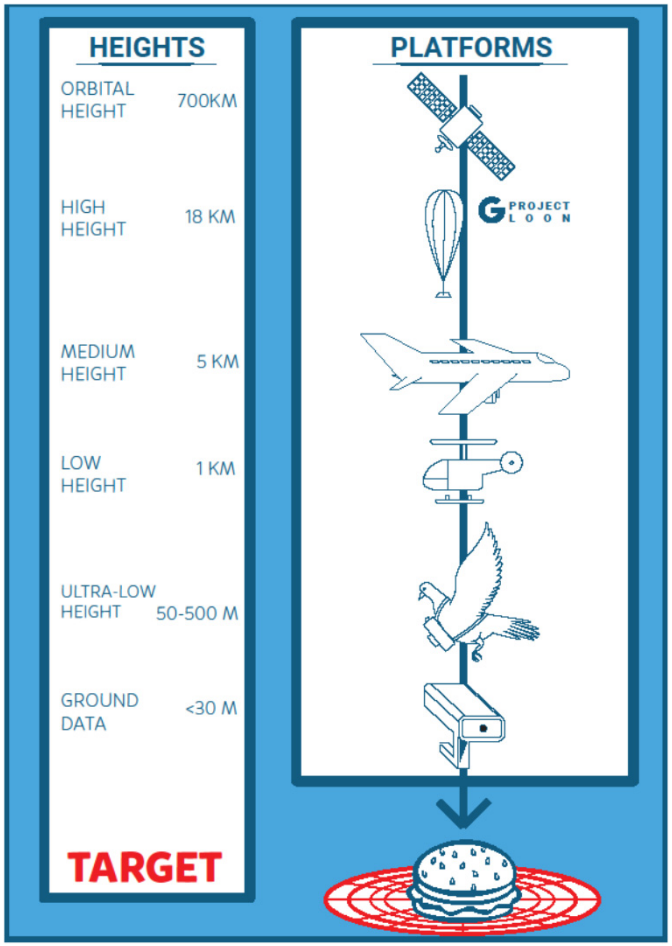
Platforms and Sensors

Active Sensors: own source of illumination

- Radar
- Lidar
- Sonar
- Laser Altimeter

Passive Sensors: natural light available

- Radiometer
- Imaging Radiometer
- Spectrometer
- Spectroradiometer



Sentinel 2

- Launch dates:
  - Sentinel-2A - 23/06/2015
  - Sentinel-2B – 07/03/2017
- Repeat cycle: 5 days with both satellites
- Multispectral sensor: 13 bands

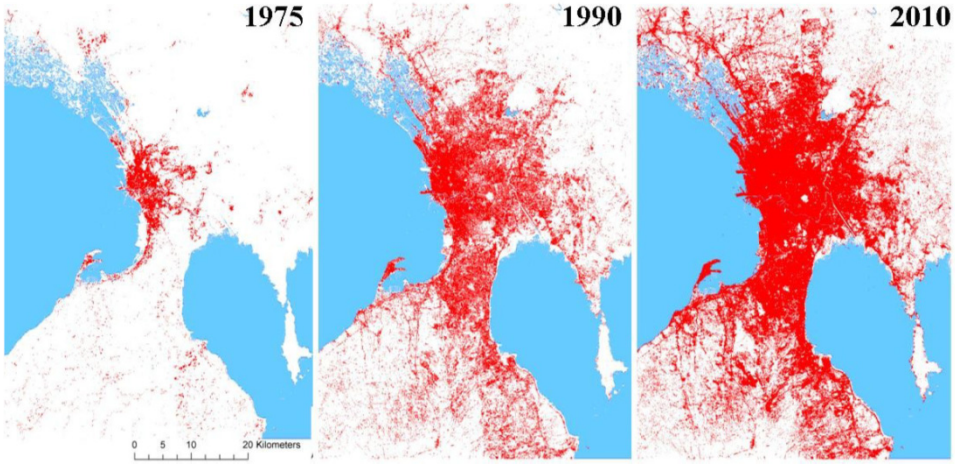
[10] Sentinel-2 global coverage

Application Domains

- Suitable for many applications
- Non invasive method
- Satellite platforms
  - Invaluable view
  - Repetitive and consistent

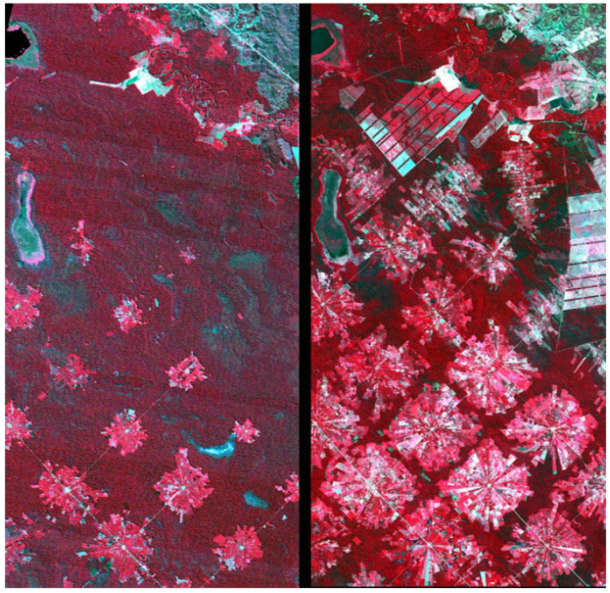
Application Examples

- Environmental assessment and monitoring
  - Urban growth
  - Hazardous waste



[11] Population growth from 1975 - 2010 of Manila

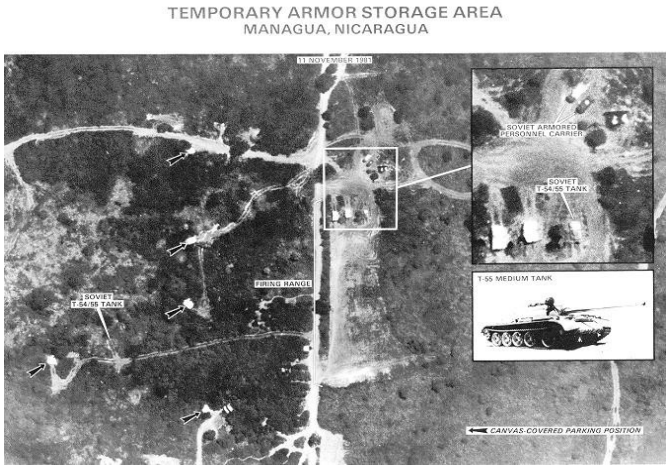
- Global change detection and monitoring
  - Atmospheric ozone depletion
  - Deforestation
  - Global warming



[12] Deforestation in Bolivia from 1986 to 2001

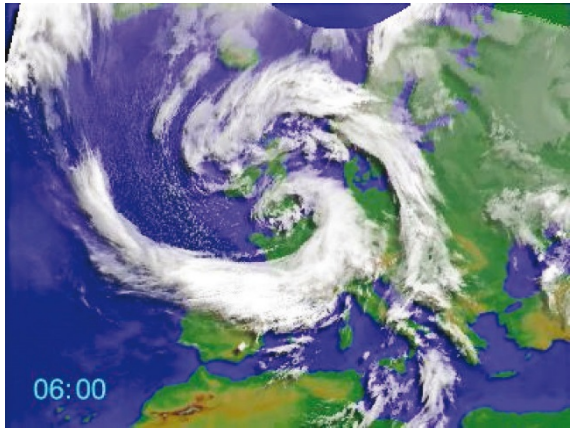
Application Examples

- Military surveillance and reconnaissance
  - Strategic policy
  - Tactical assessment



[13] Temporary Armor Storage Area

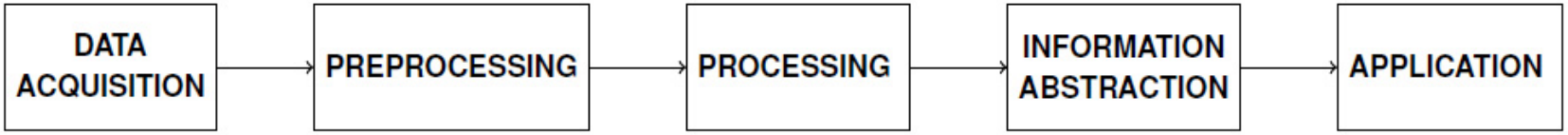
- Meteorology
  - Atmosphere dynamics
  - Weather prediction



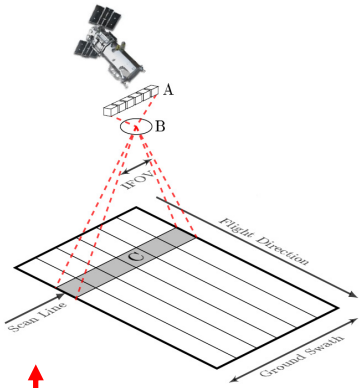
[14] NOAA image

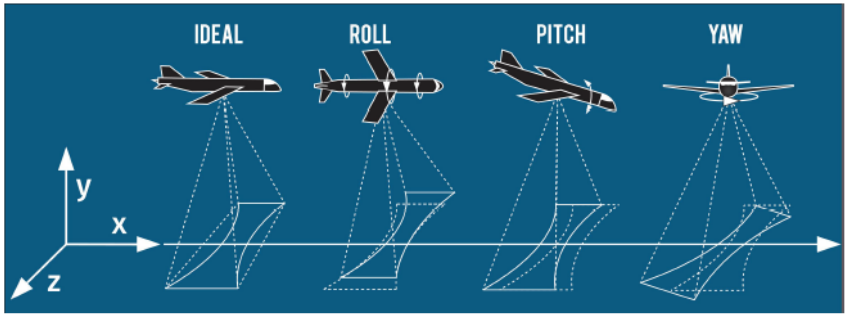
Big Data Analytics

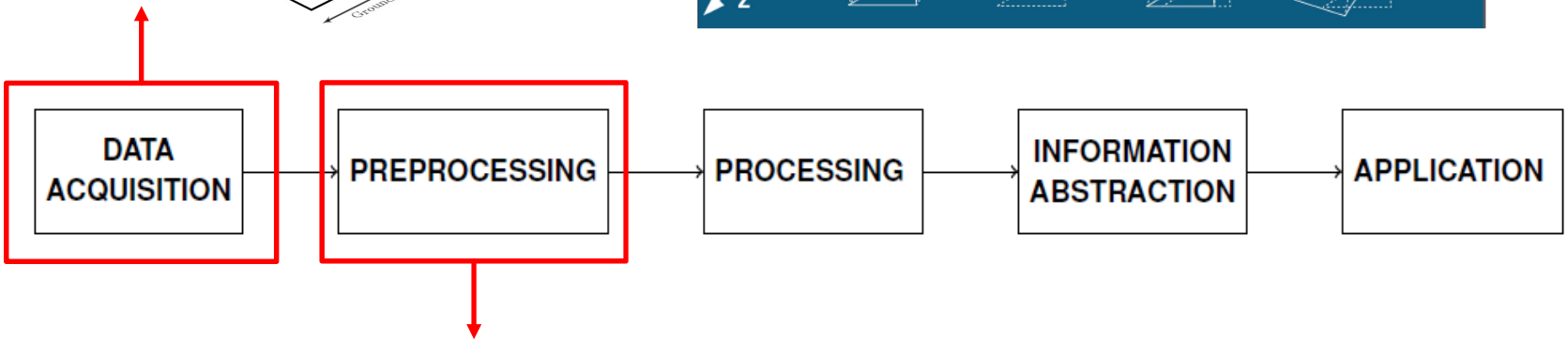
- New analysis challenges
  - 5Vs: *Volume*, *Variety*, *Velocity*, *Veracity* and *Value*
- Scalable methods
- Underlying infrastructures



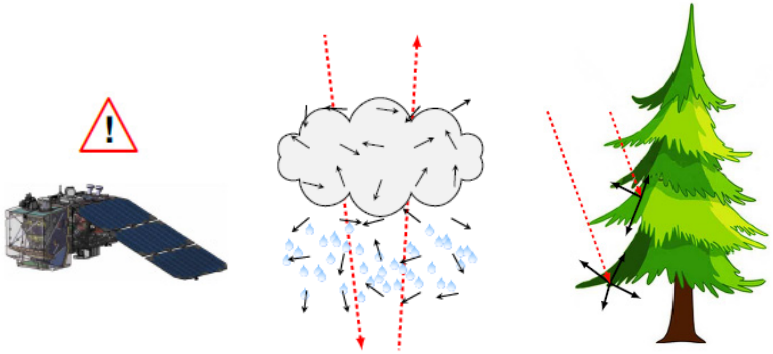
Data Acquisition and Preprocessing





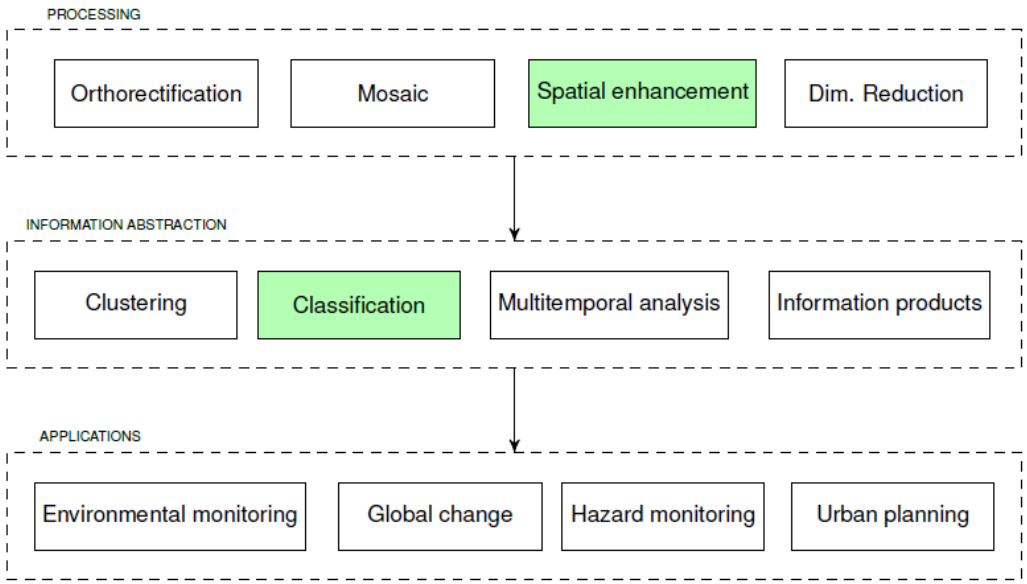


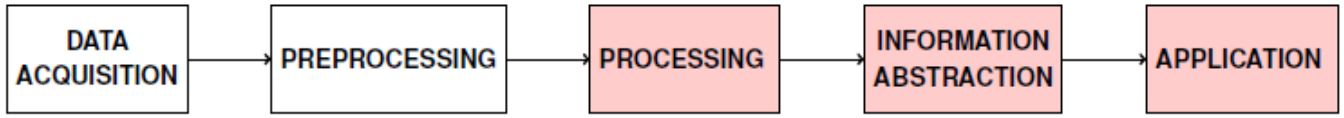
- Data are complex, noisy and may contain errors:
  - Sensors malfunction
  - Atmospheric distortion
  - Multiple scattering



Data Interpretation

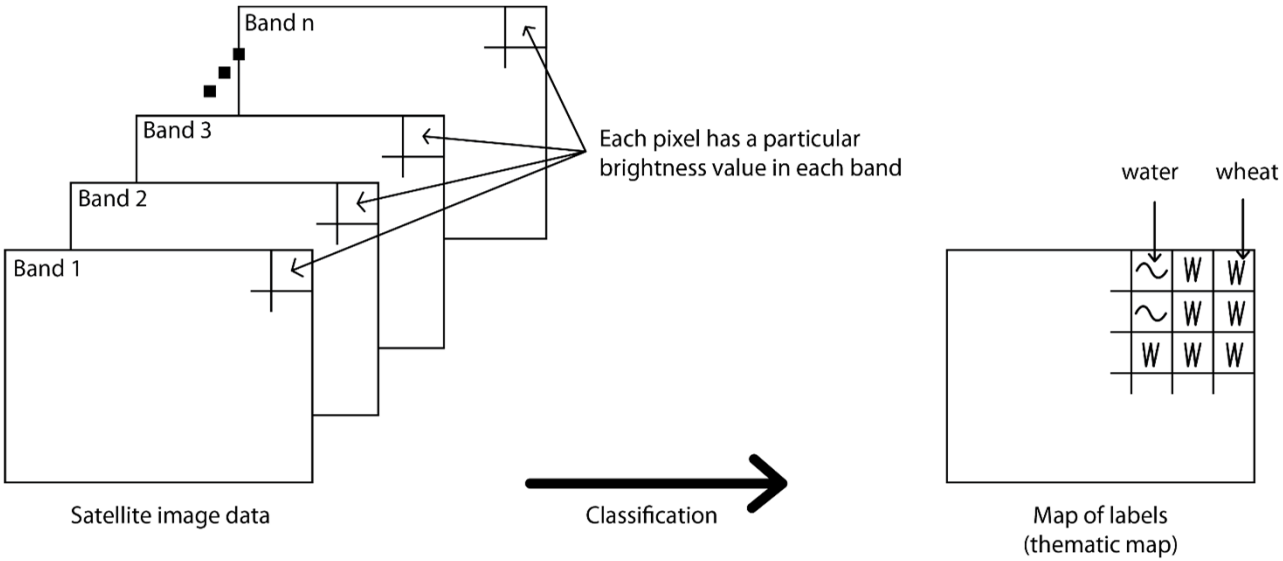
- Once the data is acquired and corrected
- How to interpret the data content and extract information for a specific task?
  - Manual vs Automatic photo-intepretation





Classification of Remote Sensing Images

- Perhaps the most common form of image interpretation



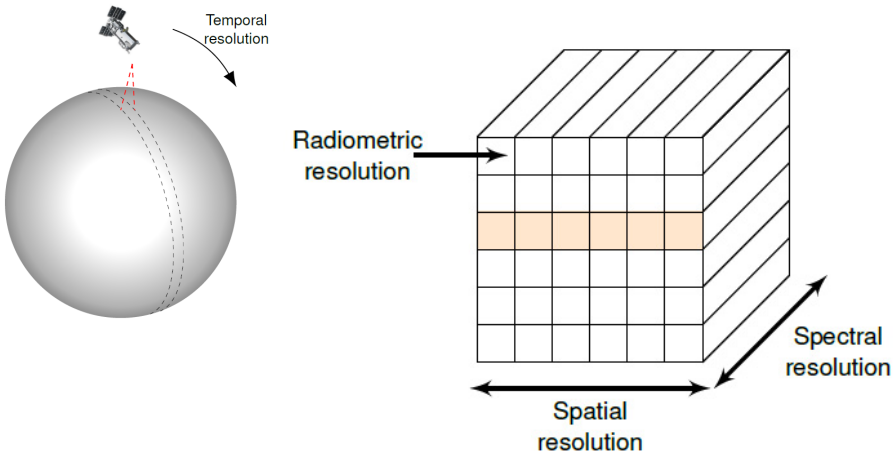
1. Feature transformation
   Spectral or spatial transformation of the input data

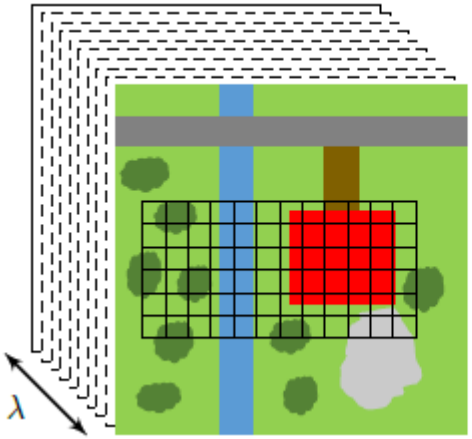
2. Training of the classifier
   The classifier is trained on the data (the problem is learned)

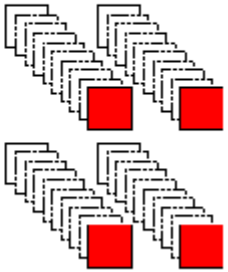
3. Labeling of the data
   The classifier performs a decision on all the data

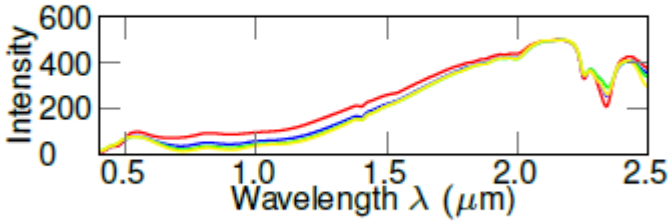
The Value of the Data

- The information content of remote sensing images depends upon various factors
  - The sensor resolution
  - The equipment unreliability
  - The type and amount of noise, etc.

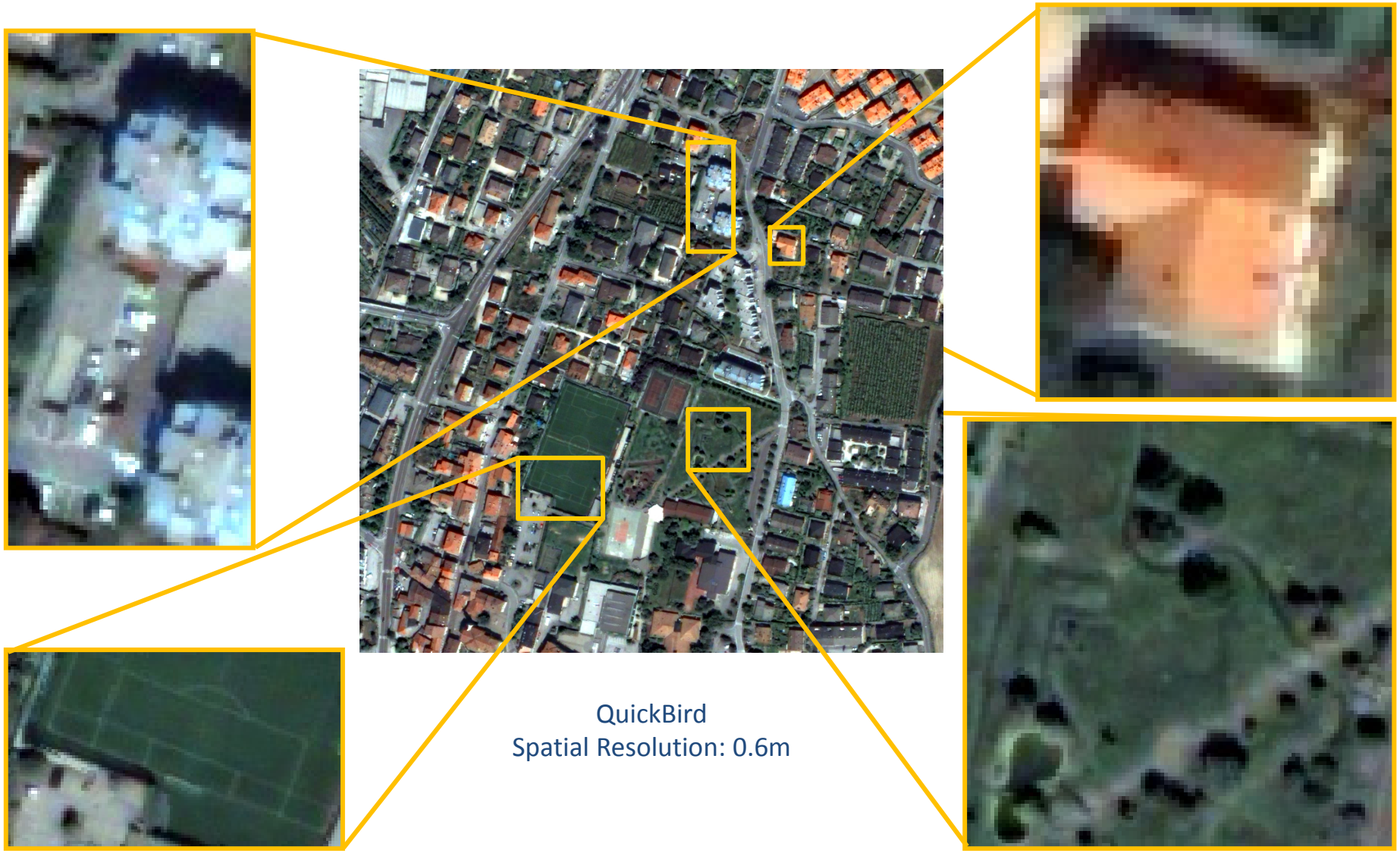






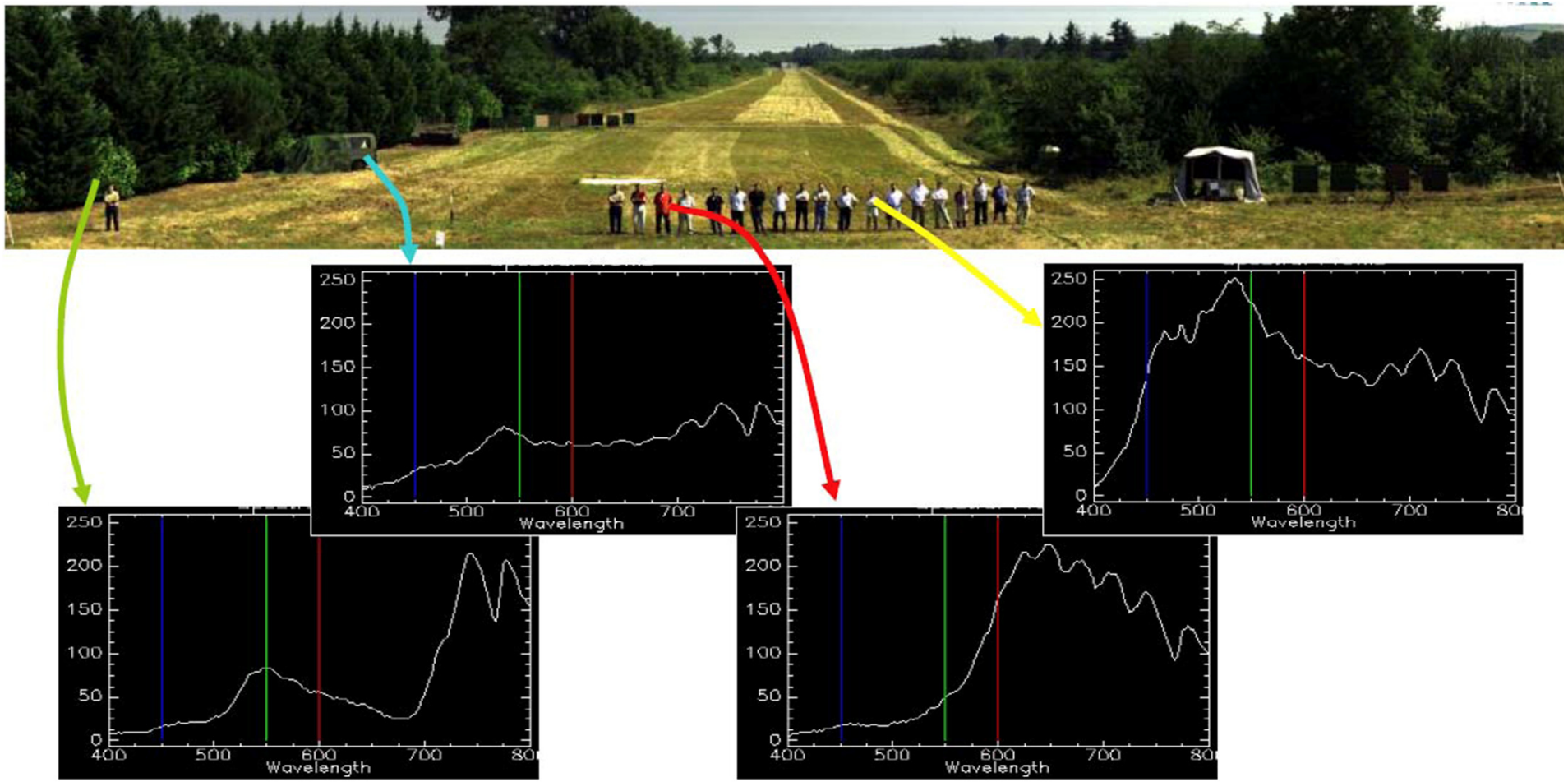


The Spatial Information



QuickBird
Spatial Resolution: 0.6m

The Spectral Information

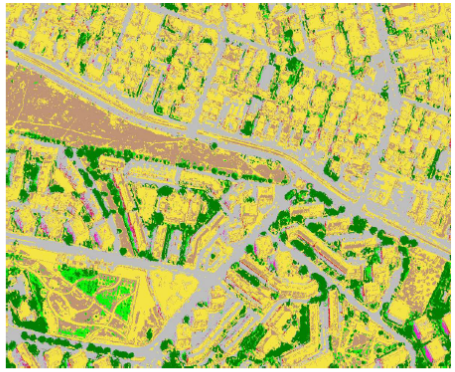


Improved spectral diversity : hyperspectral imagery

Pixel-Wise Spectral-Spatial Classifiers

Examples of maps obtained by classifying different features

Buildings
Blocks
Roads
Light Train
Vegetation
Trees
Bare Soil
Soil
Tower



spectral features

spectral + **spatial features**

The spatial features increases the discrimination of the thematic classes

How to Extract Spatial Features?

- Very High Spatial Resolution images: huge amount of details

WorldView-2 Panchromatic - Resolution 0.46 [m]

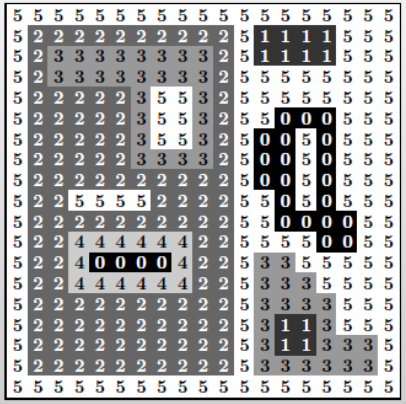
- Sub-metric resolution
- Allows for accurate analysis
- Objects with different scales and shapes

Spatial Simplication of the image

Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

gray-scale image



5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	1	1	1	1	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	1	1	1	1	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	5	0	0	0	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	0	0	5	0	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	0	0	5	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	0	0	5	0	5	5	5
5	2	2	5	5	5	5	2	2	2	2	5	5	0	5	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	0	0	0	0	5	5
5	2	2	4	4	4	4	4	4	2	2	5	5	5	5	0	0	5	5
5	2	2	4	0	0	0	0	4	2	2	5	3	3	5	5	5	5	5
5	2	2	4	4	4	4	4	4	2	2	5	3	3	3	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	1	1	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	1	1	3	3	3	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	3	3	5
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

tree of shapes

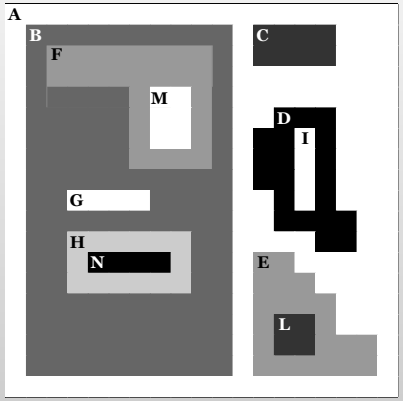
flat zones

Spatial Simplication of the image

Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



tree of shapes

flat zones

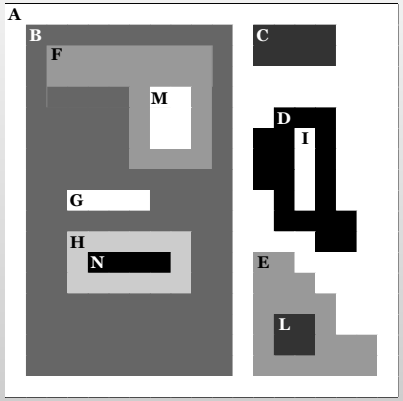
Flat zone: *set of connected iso-intensity pixels (connected component)*

Spatial Simplication of the image

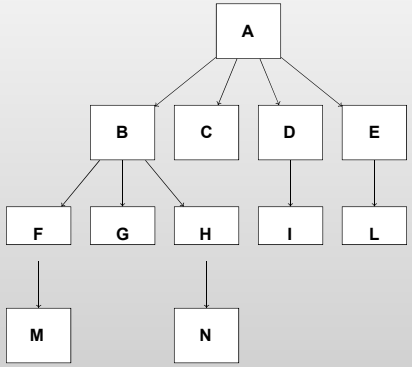
Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



tree of shapes



flat zones

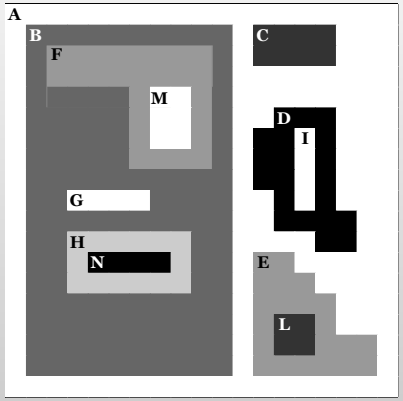
Tree representation: *each node corresponds to a flat zone*

Spatial Simplication of the image

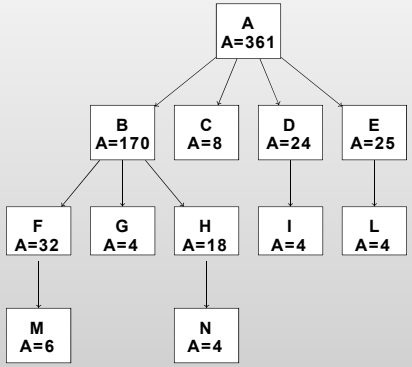
Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



area attribute



flat zones

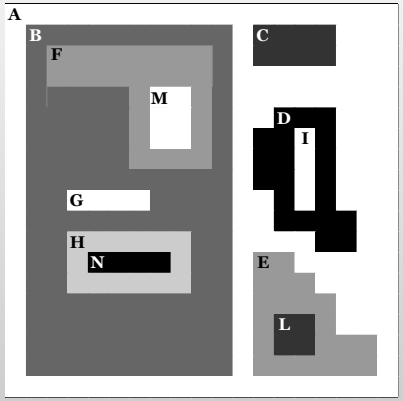
Attribute: *computed on each node*

Spatial Simplication of the image

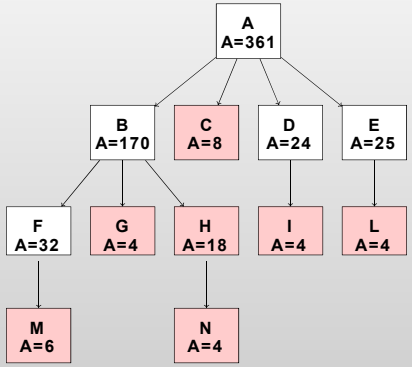
Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



$T = A \leq 20$



flat zones

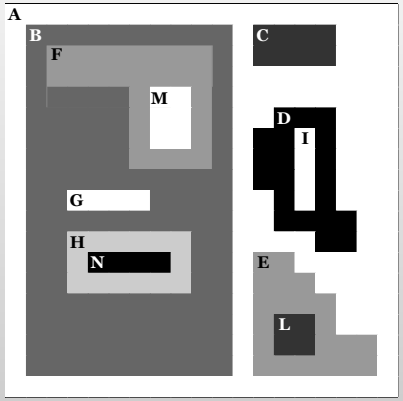
Criterion: *binary predicate evaluated on all the nodes*

Spatial Simplication of the image

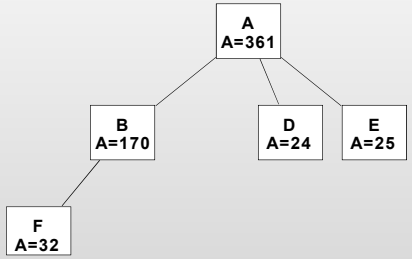
Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



$T = A \leq 20$



flat zones

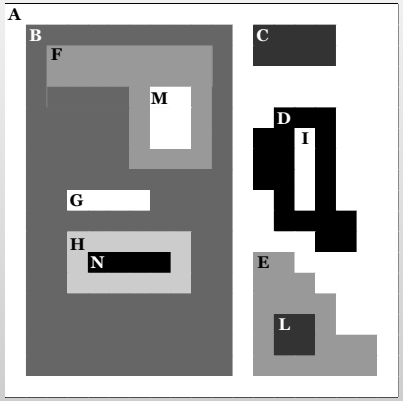
Pruning: *removing the nodes that do not fulfill the criterion*

Spatial Simplication of the image

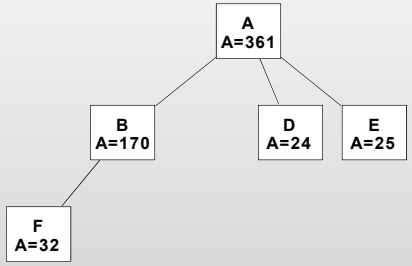
Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

flat zones



$T = A \leq 20$



flat zones

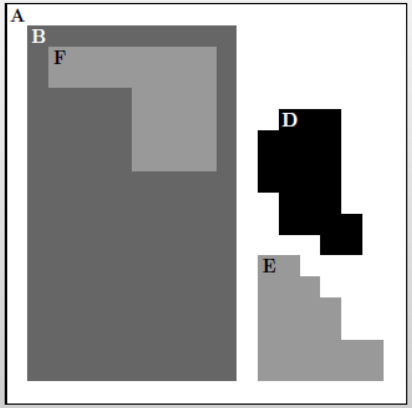


Image simplified: *transform the pruned tree back to an image*

Spatial Simplication of the image

Mathematical Morphology connected operators

- Simplification can be performed by **attribute filters**
- They operate only on flat zones according to a criterion
- Efficiently implemented on tree representations

gray-scale image

5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	1	1	1	1	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	1	1	1	1	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	5	0	0	0	5	5	5
5	2	2	2	2	2	3	5	5	3	2	5	0	0	5	0	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	0	0	5	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	0	0	5	0	5	5	5
5	2	2	5	5	5	5	2	2	2	2	5	5	0	5	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	0	0	0	0	5	5
5	2	2	4	4	4	4	4	4	2	2	5	5	5	5	0	0	5	5
5	2	2	4	0	0	0	0	4	2	2	5	3	3	5	5	5	5	5
5	2	2	4	4	4	4	4	4	2	2	5	3	3	3	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	1	1	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	1	1	3	3	3	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	3	3	5
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

Which details should be removed?

It depends on the application

image simplified

5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	5	5	5	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	5	5	5	5	5	5	5
5	2	3	3	3	3	3	3	3	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	5	5	5	5	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	5	0	0	0	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	0	0	0	0	5	5	5
5	2	2	2	2	2	3	3	3	3	2	5	0	0	0	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	0	0	0	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	0	0	0	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	0	0	0	0	5	5
5	2	2	2	2	2	2	2	2	2	2	5	5	5	5	0	0	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	5	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	5	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	5	5	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	3	3	5
5	2	2	2	2	2	2	2	2	2	2	5	3	3	3	3	3	3	5
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

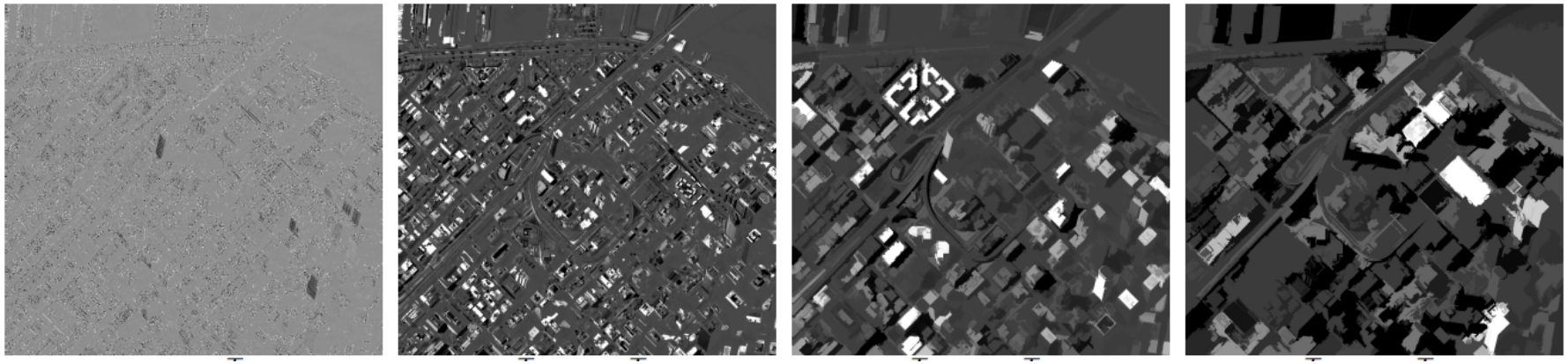
Simplification leads to: *removing or attenuating undesired details*

Self-Dual Attribute Profiles (SDAPs)

Representative: they contain salient structures of the input image

Non-redundant: same objects are present only in one or few levels of the SDAP

WorldView-2 Panchromatic
Resolution 0.46 [m]



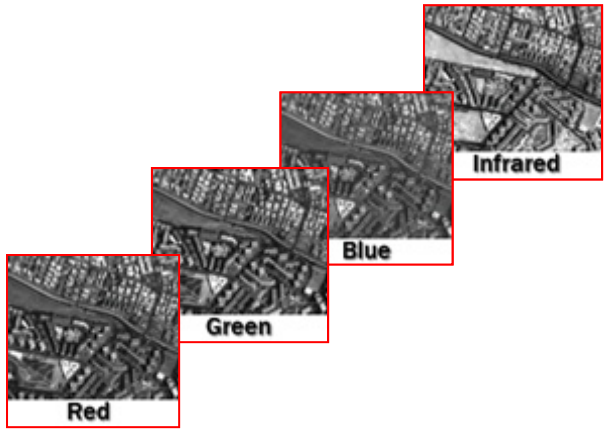
Example Rome Dataset

- Geographical location: Image of Rome, Italy
  - Multispectral data obtained by Quickbird satellite sensor

- High-resolution (0.6m) panchromatic image

(Reasoning for picking SVM: Good classification accuracies on high dimensional datasets, even with a small ‚rare' number of training samples)

Low-resolution (2.4m) multispectral images



Classes

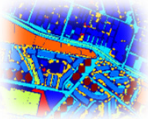
Class	Training	Test
Buildings	18126	163129
Blocks	10982	98834
Roads	16353	147176
Light Train	1606	14454
Vegetation	6962	62655
Trees	9088	81792
Bare Soil	8127	73144
Soil	1506	13551
Tower	4792	43124
Total	77542	697859

[15] Rome Image dataset

B2SHARE
Store and Share Research Data

Inspect and Understanding the Rome Dataset

Class + *Input* + *Features* + *Input* + *Feature* + *Input* + *Features* + *Input* + *Features* + *Input* + *Features*



Labels

10 filtered images

Pan

10 filtered images

Green

10 filtered images

Blue

10 filtered images

Red

10 filtered images

Infrared

- Each pixel vector is stored as a line with the libSVM format
- E.g.,

2 1:0.364706 2:0.360784 3:0.356863 4:0.356863 5:0.349206 6:0.306878 7:0.419355 8:0.453608 9:0.368421 10:1 11:1 12:0.423529 13:0.403922 14:0.403922 15:0.369919 16:0.320833 17:0.302564 18:0.481481 19:0.483516 20:0.32 21:0.625 22:0.833333 23:0.376471 24:0.376471 25:0.372549 26:0.358566 27:0.318367 28:0.243386 29:0.455446 30:0.4 31:0.319149 32:0.368421 33:0.4 34:0.556863 35:0.54902 36:0.436 37:0.322176 38:0.215962 39:0.151079 40:0.257576 41:0.267857 42:0.266667 43:0.277778 44:0.4375 45:0.360784 46:0.360784 47:0.368627 48:0.368627 49:0.363636 50:0.353846 51:0.347826 52:0.335294 53:0.333333 54:0.978723 55:1

[16] G. Cavallaro & M. Riedel et al., 2014

Inspect and Understanding the Rome Dataset

- Data is publicly available in EUDAT B2SHARE tool

Rome data set OK

by [Unknown]

Dec 22, 2016

Last updated at Jan 11, 2018

Abstract: Attribute area

PID: 11304/4615928c-e1a5-11e3-8cd7-14feb57d12b9 Copy

EUDAT

B2SHARE
Store and Share Research Data

[15] Rome Image dataset

Files

Name	Size
sdap_area_all_test.el	419.97MB
sdap_area_all_training.el	46.65MB
sdap_area_panch_test.el	114.76MB
sdap_area_panch_training.el	12.75MB

Basic metadata

Open Access	True ✔	
License		
Contact Email		
Publication Date	2014-05-22	
Contributors		
Resource Type	Category	Other
Alternate identifiers	86	
	Type	B2SHARE_V1_ID
	http://hdl.handle.net/11304/a2892f1c-7e13-42e9-be24-88ae9da847b1	
	Type	ePIC_PID

(persistent handle link for publication into papers)

```
[train001@jrl12 rome]$ pwd
/homea/hpclab/train001/data/rome
[train001@jrl12 rome]$ ls -al
total 580256
drwxr-xr-x 2 train001 hpclab       512 Jan 14 21:52 .
drwxr-xr-x 6 train001 hpclab       512 Jan 14 21:47 ..
-rw-r--r-- 1 train001 hpclab 419974873 Dec 22  2016 sdap_area_all_test.el
-rw-r--r-- 1 train001 hpclab  46652874 Dec 22  2016 sdap_area_all_training.el
-rw-r--r-- 1 train001 hpclab 114763982 Dec 22  2016 sdap_area_panch_test.el
-rw-r--r-- 1 train001 hpclab  12745692 Dec 22  2016 sdap_area_panch_training.el
```

Exercises – Explore the Rome Dataset

Example Indian Pines Dataset

NASA – Jet Propulsion Laboratory (JPL)

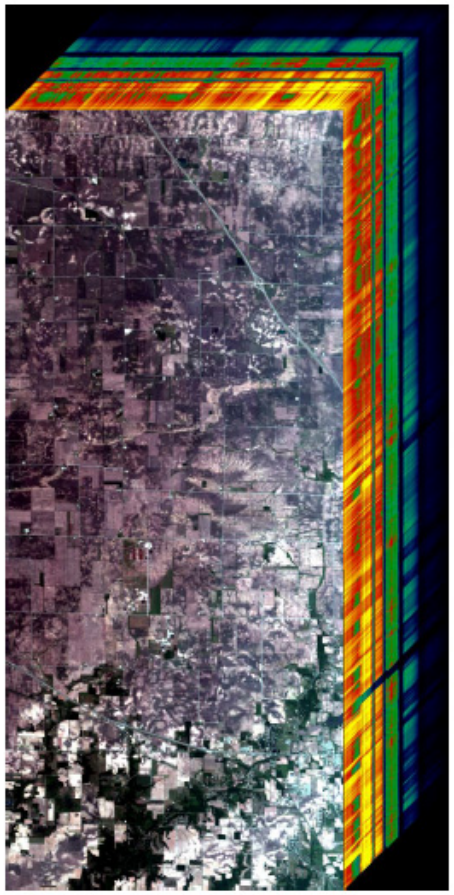
- Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)
- 224 contiguous spectral channels
- Spectral range of 0.4 µm to 2.5 µm (10 nm spectral resolution)
- Spatial resolution of 20 m

[17] AVIRIS

Indian Pine Dataset

June 1992

- Agricultural fields with a variety of crops
- Challenging classification problem
- Similar spectral classes and mixed pixels



[18] Indian Pines dataset

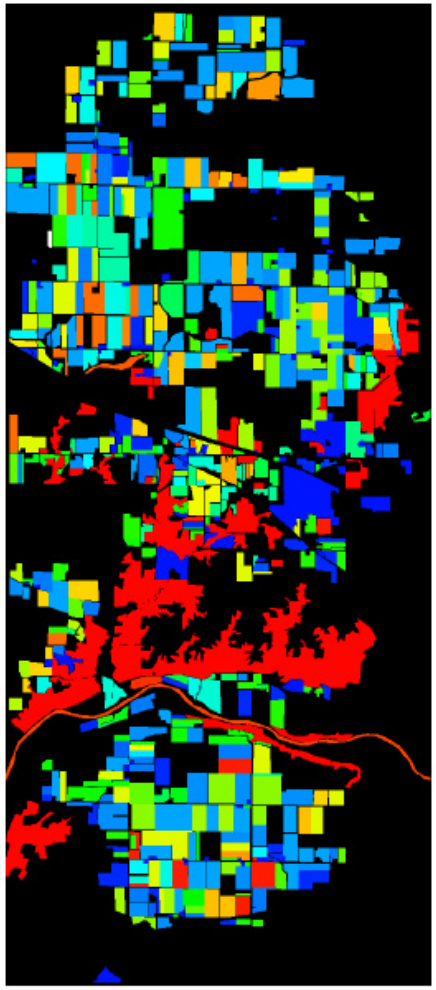
Preprocessing

Corrected by JPL

- 1417×617 pixels (∼600 MB)
- 200 bands (20 discarded, with low SNR)
- 58 classes (6 discarded, with $\leq$ 100 samples)

Class		Number of samples		Class		Number of samples	
number	name	training	test	number	name	training	test
1	Buildings	1720	15475	27	Pasture	1039	9347
2	Corn	1778	16005	28	pond	10	92
3	Corn?	16	142	29	Soybeans	939	8452
4	Corn-EW	51	463	30	Soybeans?	89	805
5	Corn-NS	236	2120	31	Soybeans-NS	111	999
6	Corn-CleanTill	1240	11164	32	Soybeans-CleanTill	507	4567
7	Corn-CleanTill-EW	2649	23837	33	Soybeans-CleanTill?	273	2453
8	Corn-CleanTill-NS	3968	35710	34	Soybeans-CleanTill-EW	1180	10622
9	Corn-CleanTill-NS-Irrigated	80	720	35	Soybeans-CleanTill-NS	1039	9348
10	Corn-CleanTilled-NS?	173	1555	36	Soybeans-CleanTill-Drilled	224	2018
11	Corn-MinTill	105	944	37	Soybeans-CleanTill-Weedy	54	489
12	Corn-MinTill-EW	563	5066	38	Soybeans-Drilled	1512	13606
13	Corn-MinTill-NS	886	7976	39	Soybeans-MinTill	267	2400
14	Corn-NoTill	438	3943	40	Soybeans-MinTill-EW	183	1649
15	Corn-NoTill-EW	121	1085	41	Soybeans-MinTill-Drilled	810	7288
16	Corn-NoTill-NS	569	5116	42	Soybeans-MinTill-NS	495	4458
17	Fescue	11	103	43	Soybeans-NoTill	216	1941
18	Grass	115	1032	44	Soybeans-NoTill-EW	253	2280
19	Grass/Trees	233	2098	45	Soybeans-NoTill-NS	93	836
20	Hay	113	1015	46	Soybeans-NoTill-Drilled	873	7858
21	Hay?	219	1966	47	Swampy Area	58	525
22	Hay-Alfalfa	226	2032	48	River	311	2799
23	Lake	22	202	49	Trees?	58	522
24	NotCropped	194	1746	50	Wheat	498	4481
25	Oats	174	1568	51	Woods	6356	57206
26	Oats?	34	301	52	Woods?	14	130

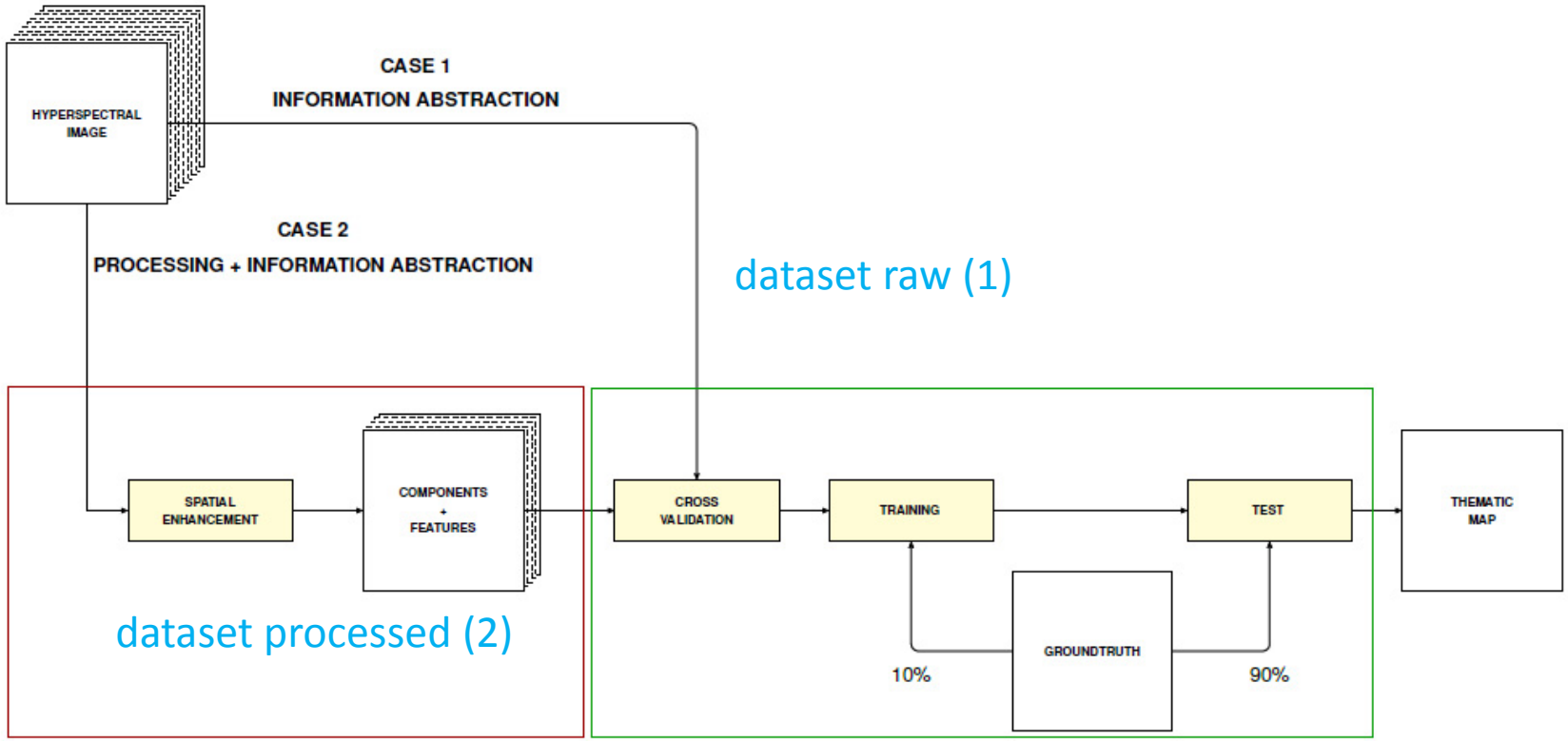
[19] G. Cavallaro and M. Riedel, et al. , 2015



(non-linearly separable) dataset

Experimental Setup

Two Cases



Feature Enhancement & Selection

- Kernel Principle Component Analysis (KPCA)
- Extended Self-Dual Attribute Profile (ESDAP)
- Nonparametric weighted feature extraction (NWFE)

[19] G. Cavallaro and M. Riedel, et al. , 2015

Inspect and Understanding the Indian Pines Dataset

- Dataset raw (1)

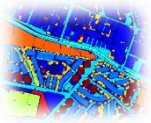
Class + *Original Spectral Bands*

200 spectral bands

48 1:0.365 2:0.361 3:0.356 ..... 209:0.333 220:0.978 libSVM

- dataset processed (2)

Class + *Transformed Features*



30 image features

48 1:0.245 2:0.360 3:0.326 ..... 29:0.241 30:0.878 libSVM

Available Datasets – Training Data Example

- *Indian Pines Dataset Processed – Training*
  - Indian_processed_training.el
  - LibSVM data format: class feature1:value1 feature2:value2

[20] Indian Pines Raw and Processed

```
[morris@login-0-1 pisvmexamples]$ head /home/morris/BIGDATA/172-IndianPines/indian_processed_training.el
48 1:0.69021 2:0.599122 3:0.864571 4:0.265246 5:0.566572 6:0.561181 7:0.665698 8:0.430511 9:0.717402 10:0.514414 11:0.335391 12:0.486979 13:
0.58382 14:0.503849 15:0.420948 16:0.624377 17:0.43594 18:0.324333 19:0.614007 20:0.431993 21:0.600198 22:0.577696 23:0.528974 24:0.458141 2
5:0.543731 26:0.382421 27:0.33752 28:0.543062 29:0.5737 30:0.467784
48 1:0.675361 2:0.585579 3:0.85587 4:0.249317 5:0.579717 6:0.5633 7:0.634459 8:0.434777 9:0.719665 10:0.510279 11:0.330765 12:0.478097 13:0.
559455 14:0.505097 15:0.396293 16:0.627475 17:0.425701 18:0.32044 19:0.637625 20:0.437285 21:0.620149 22:0.535474 23:0.525151 24:0.511439 25
:0.504738 26:0.385925 27:0.303322 28:0.572963 29:0.570833 30:0.491508
51 1:0.554975 2:0.499039 3:0.578284 4:0.648481 5:0.867561 6:0.17041 7:0.841141 8:0.54466 9:0.746417 10:0.459481 11:0.356237 12:0.420116 13:0
.248721 14:0.541081 15:0.444207 16:0.286447 17:0.364712 18:0.522284 19:0.647555 20:0.538008 21:0.51632 22:0.437087 23:0.402312 24:0.574872 2
5:0.633603 26:0.426417 27:0.397177 28:0.400566 29:0.470729 30:0.51207
51 1:0.561296 2:0.495371 3:0.330829 4:0.513766 5:0.556074 6:0.387238 7:0.932418 8:0.643317 9:0.735262 10:0.216349 11:0.284832 12:0.499805 13
:0.272644 14:0.289776 15:0.699631 16:0.142032 17:0.161307 18:0.657442 19:0.72633 20:0.446231 21:0.542728 22:0.526951 23:0.433363 24:0.647875
 25:0.679122 26:0.760654 27:0.302971 28:0.515992 29:0.602717 30:0.34428
46 1:0.870891 2:0.768472 3:0.783699 4:0.28838 5:0.632466 6:0.418282 7:0.508729 8:0.437482 9:0.610721 10:0.480019 11:0.291032 12:0.427472 13:
0.428984 14:0.447632 15:0.452797 16:0.527423 17:0.583162 18:0.42837 19:0.557041 20:0.290409 21:0.547233 22:0.642817 23:0.535524 24:0.4478 25
:0.601117 26:0.405341 27:0.315288 28:0.488306 29:0.659466 30:0.469213
46 1:0.87119 2:0.779909 3:0.769342 4:0.292747 5:0.643116 6:0.416394 7:0.530679 8:0.4283 9:0.61224 10:0.474311 11:0.325682 12:0.468692 13:0.4
2888 14:0.461495 15:0.471023 16:0.524454 17:0.527418 18:0.398571 19:0.555335 20:0.35517 21:0.557823 22:0.588179 23:0.564881 24:0.444141 25:0
.576903 26:0.495859 27:0.344496 28:0.541081 29:0.539253 30:0.494238
15 1:0.869845 2:0.781471 3:0.412471 4:0.318165 5:0.414446 6:0.320591 7:0.605434 8:0.726664 9:0.590541 10:0.457315 11:0.355967 12:0.429357 13
:0.509808 14:0.421272 15:0.5406 16:0.532659 17:0.568507 18:0.404705 19:0.523209 20:0.52817 21:0.575417 22:0.570611 23:0.578825 24:0.452228 2
5:0.484941 26:0.542836 27:0.335545 28:0.491274 29:0.656853 30:0.546783
15 1:0.892436 2:0.790123 3:0.485313 4:0.318438 5:0.478488 6:0.329743 7:0.602466 8:0.581215 9:0.597853 10:0.442717 11:0.409894 12:0.479275 13
:0.51758 14:0.445422 15:0.483192 16:0.549861 17:0.49607 18:0.469129 19:0.475024 20:0.638295 21:0.578851 22:0.586814 23:0.579959 24:0.463105
25:0.550809 26:0.490199 27:0.376347 28:0.468157 29:0.55987 30:0.42689
15 1:0.809988 2:0.508887 3:0.782026 4:0.284838 5:0.682891 6:0.361665 7:0.243857 8:0.330213 9:0.456187 10:0.385119 11:0.230994 12:0.403049 13
:0.333648 14:0.473923 15:0.570014 16:0.45283 17:0.314596 18:0.574783 19:0.708582 20:0.292825 21:0.531681 22:0.616414 23:0.506682 24:0.493297
 25:0.396451 26:0.599563 27:0.329084 28:0.541152 29:0.56897 30:0.466168
8 1:0.843723 2:0.734137 3:0.686274 4:0.335633 5:0.6219 6:0.250049 7:0.575193 8:0.50452 9:0.623444 10:0.496143 11:0.263468 12:0.482361 13:0.5
46293 14:0.402207 15:0.730217 16:0.378631 17:0.433787 18:0.654169 19:0.513311 20:0.202298 21:0.530649 22:0.696446 23:0.631684 24:0.447328 25
:0.425982 26:0.40432 27:0.568027 28:0.461795 29:0.411903 30:0.428683
```

Publicly Available Datasets – Location

- *Indian Pines Dataset Raw and Processed*

B2SHARE
Store and Share Research Data

[20] Indian Pines Raw and Processed

Indian pines: raw and processed

by [Unknown]

Dec 22, 2016

Last updated at Jan 11, 2018

Abstract: 1) Indian raw: 1417x614x200 (training 10% and test) 2) Indian processed:1417x614x30 (training 10% and test)

PID: 11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f Copy

EUDAT

Files

Name	Size
indian_processed_test.el	105.59MB
indian_processed_training.el	11.73MB
indian_raw_test.el	747.13MB
indian_raw_training.el	83.01MB

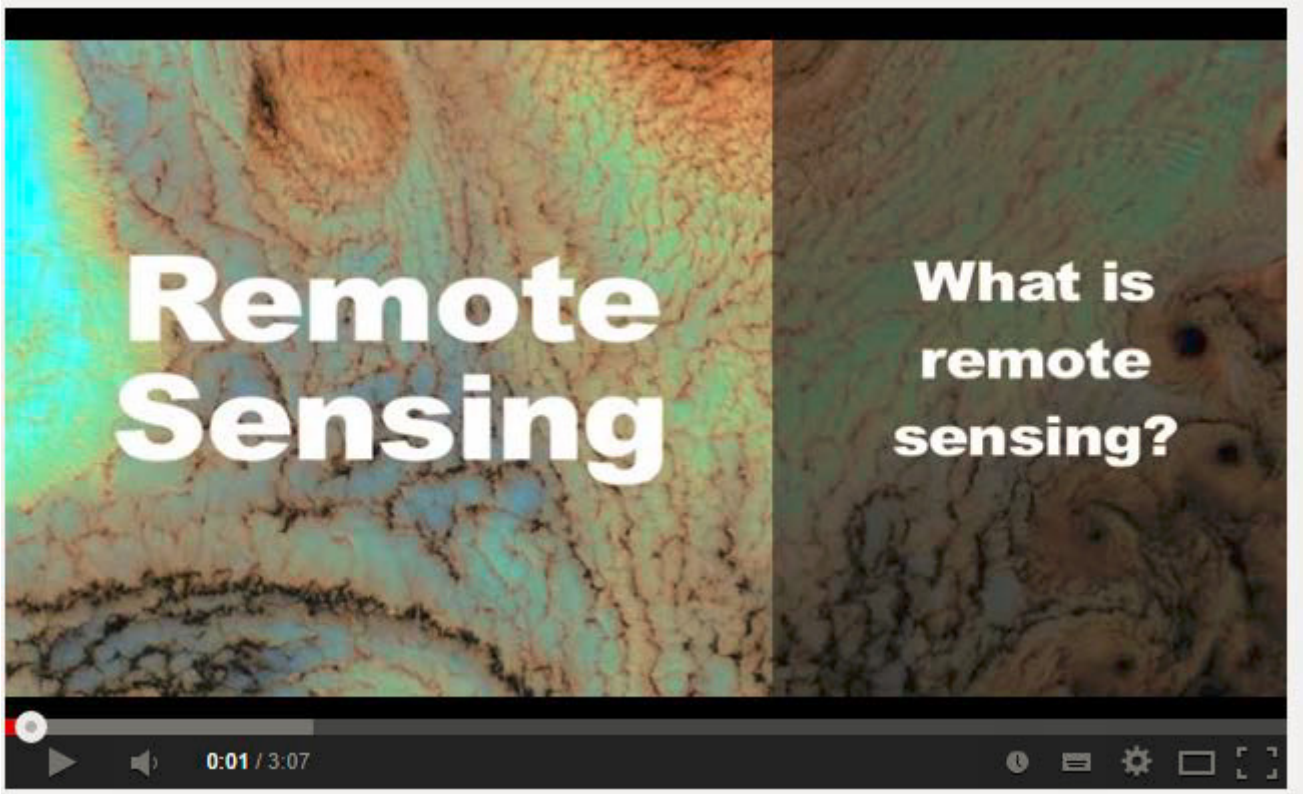
Basic metadata

Open Access	True ✔	
License		
Contact Email		
Publication Date	2015-02-04	
Contributors		
Resource Type	Category	Other
Alternate identifiers	172	
	Type	B2SHARE_V1_ID
	http://hdl.handle.net/11304/9ec5eac8-61b4-4617-ae1c-1f8c8cd3cd74	
	Type	ePIC_PID
Publisher	https://b2share.eudat.eu	
Language	en	

```
[train001@jrl07 indianpines]$ pwd
/homea/hpclab/train001/data/indianpines
[train001@jrl07 indianpines]$ ls -al
total 925344
drwxr-xr-x 2 train001 hpclab       512 Jul  7  2016 .
drwxr-xr-x 5 train001 hpclab       512 Jan 14 13:10 ..
-rw-r--r-- 1 train001 hpclab        36 Jul  7  2016 b2share.txt
-rw-r--r-- 1 train001 hpclab 105594346 Feb  5  2015 indian_processed_test.el
-rw-r--r-- 1 train001 hpclab  11732509 Feb  5  2015 indian_processed_training.el
-rw-r--r-- 1 train001 hpclab 747125597 Feb  5  2015 indian_raw_test.el
-rw-r--r-- 1 train001 hpclab  83014311 Feb  5  2015 indian_raw_training.el
```

Exercises – Explore Indian Pines Dataset

[Video] Remote Sensing



[10] YouTube Video, 'Remote Sensing'

Classification Challenges & Solutions

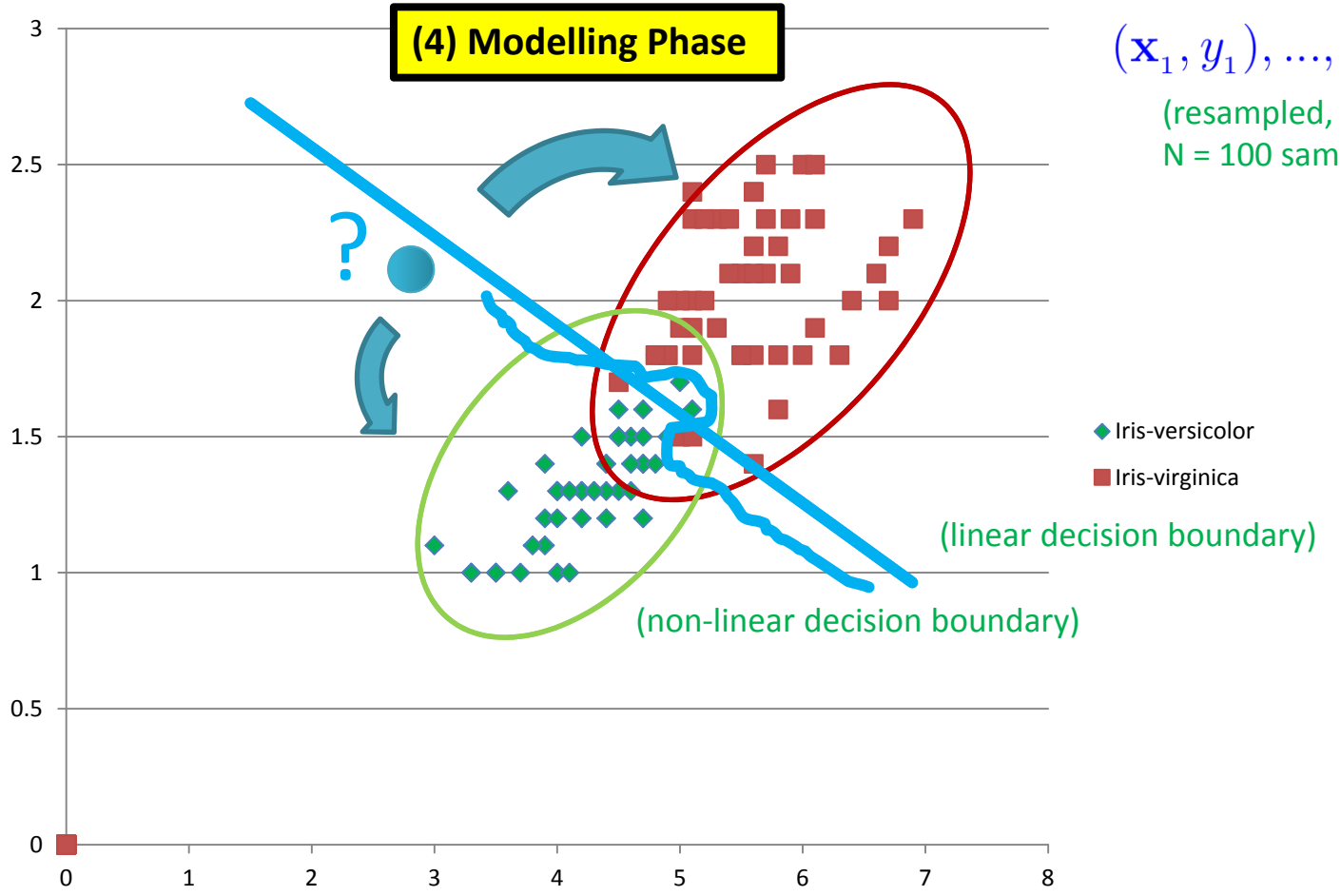
Key Challenges: Why is it not so easy in practice?

- Scalability
  - Gigabytes, Terabytes, and Petabytes datasets that fit not into memory
  - E.g. algorithms become necessary with out-of-core/CPU strategies
- High Dimensionality
  - Datasets with hundreds or thousand attributes become available
  - E.g. bioinformatics with gene expression data with thousand of features
- Heterogenous and Complex Data
  - More complex data objects emerge and unstructured data sets
  - E.g. Earth observation time-series data across the globe
- Data Ownership and Distribution
  - Distributed datasets are common (e.g. security and transfer challenges)

- **Key challenges faced when doing traditional data analysis and machine learning are scalability, high dimensionality of datasets, heterogenous and complex data, data ownership & distribution**

[21] Introduction to Data Mining

Challenge One - Non-linearly Seperable Data in Practice



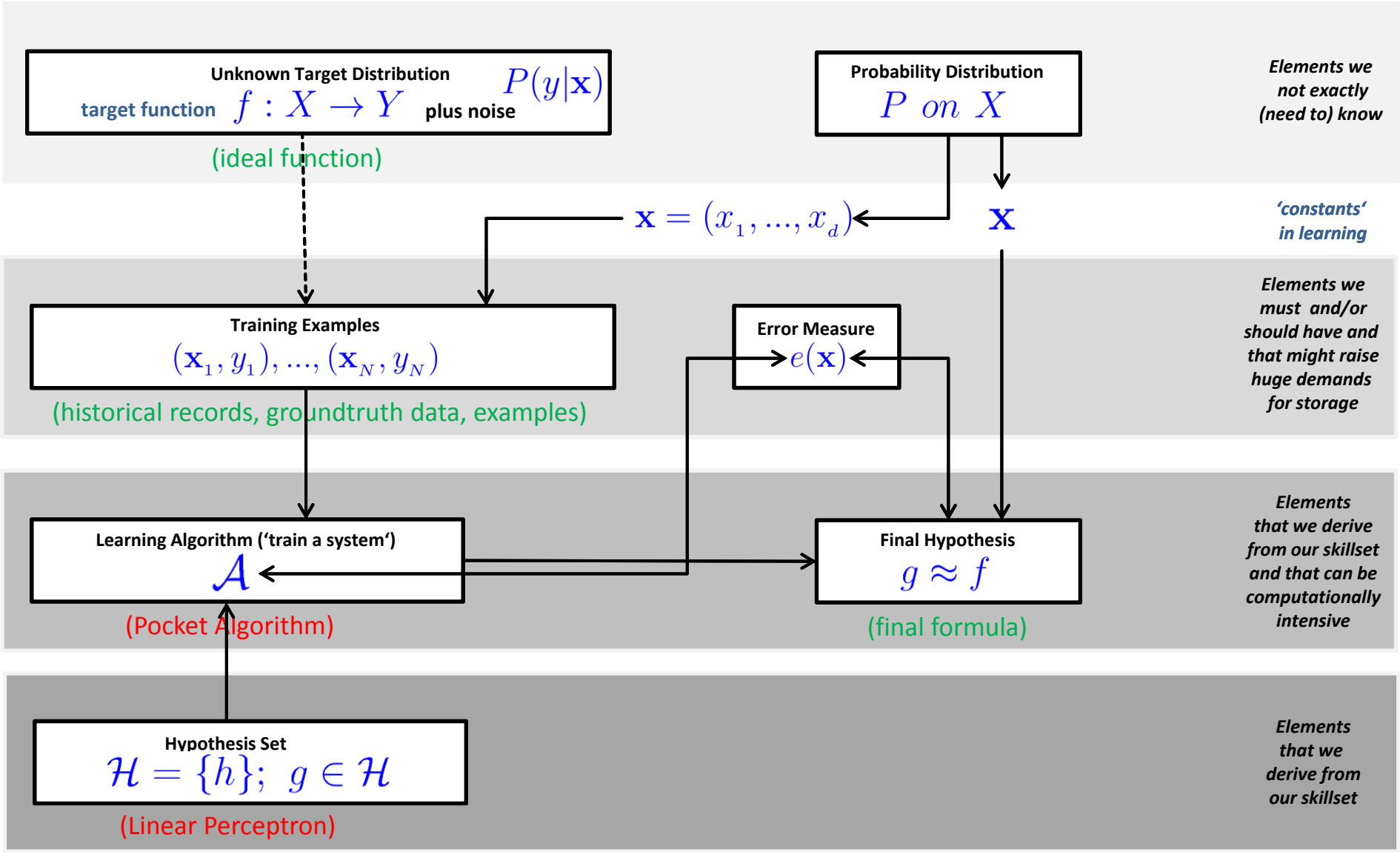
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

(resampled, again
N = 100 samples)

(lessons learned from practice: requires soft-thresholds to allow for some errors being overall better for new data)

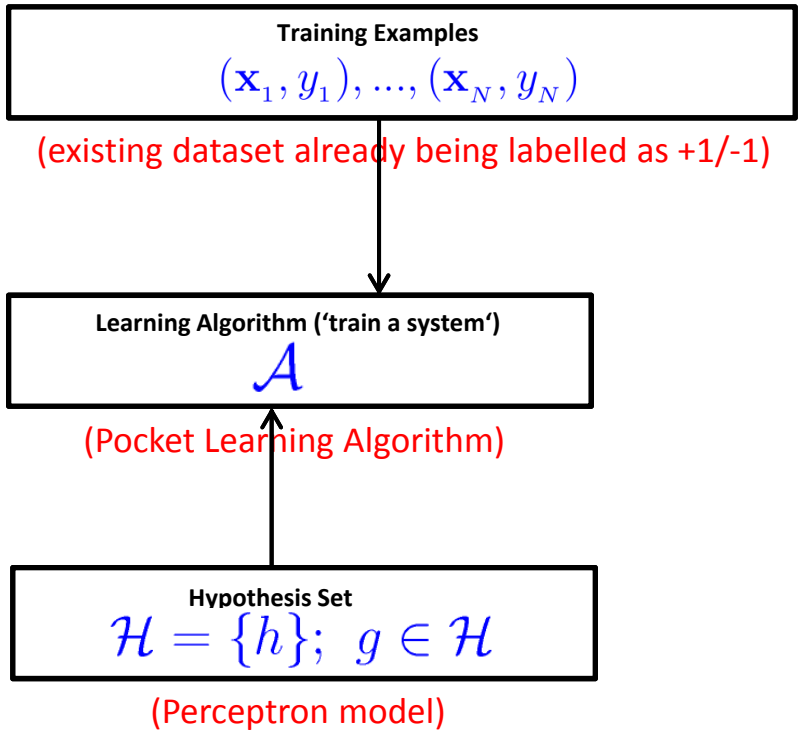
(lessons learned from practice: requires non-linear decision boundaries)

Solution Tools: Linear Perceptron Hypothesis Set & Pocket



Smart Adhancement of PLA – Pocket Algorithm

- When: If we believe there is a linear pattern to be detected
  - No assumption: can be non-linearly seperable data

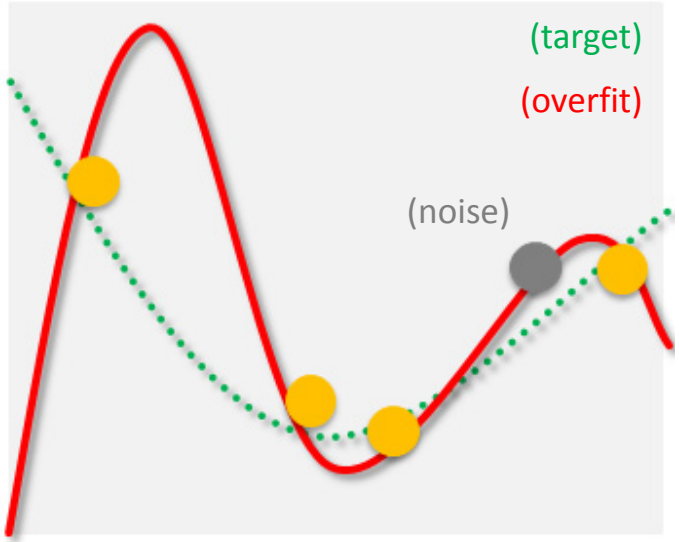


- Basis is still the PLA
- Idea: Put the best solution so far 'in a pocket'
- Best means: Error measure checks per iterations
- Works with non-linearly seperable data
- Needs fixed iterations number (otherwise no convergence of algorithm)

Challenge Two – Problem of Overfitting

- **Overfitting refers to fit the data too well – more than is warranted – thus may misguide the learning**
- **Overfitting is not just 'bad generalization' - e.g. the VC dimension covers noiseless & noise targets**
- **Theory of Regularization are approaches against overfitting and prevent it using different methods**

- Key problem: noise in the target function leads to overfitting
  - Effect: 'noisy target function' and its noise misguides the fit in learning
  - There is always 'some noise' in the data
  - Consequence: poor target function ('distribution') approximation
- Example: Target functions is second order polynomial (i.e. parabola)
  - Using a higher-order polynomial fit
  - Perfect fit: low $E_{in}(g)$, but large $E_{out}(g)$



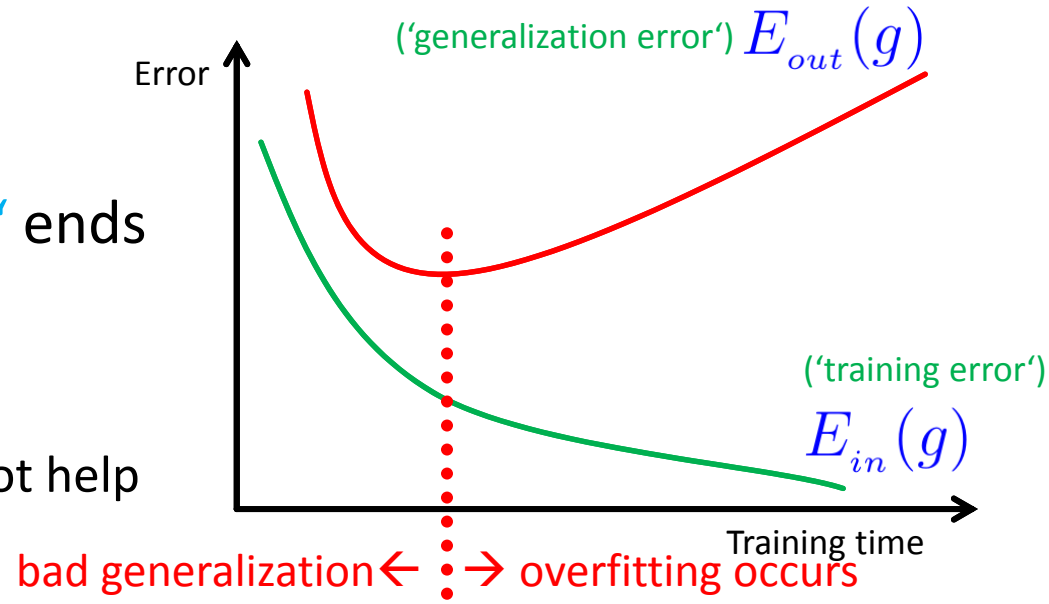
(but simple polynomial works good enough)
('over': here meant as 4th order, a 3rd order would be better, 2nd best)

Problem of Overfitting – Clarifying Terms

- **A good model must have low training error (E_{in}) and low generalization error (E_{out})**
- **Model overfitting is if a model fits the data too well (E_{in}) with a poorer generalization error (E_{out}) than another model with a higher training error (E_{in})**

[21] Introduction to Data Mining

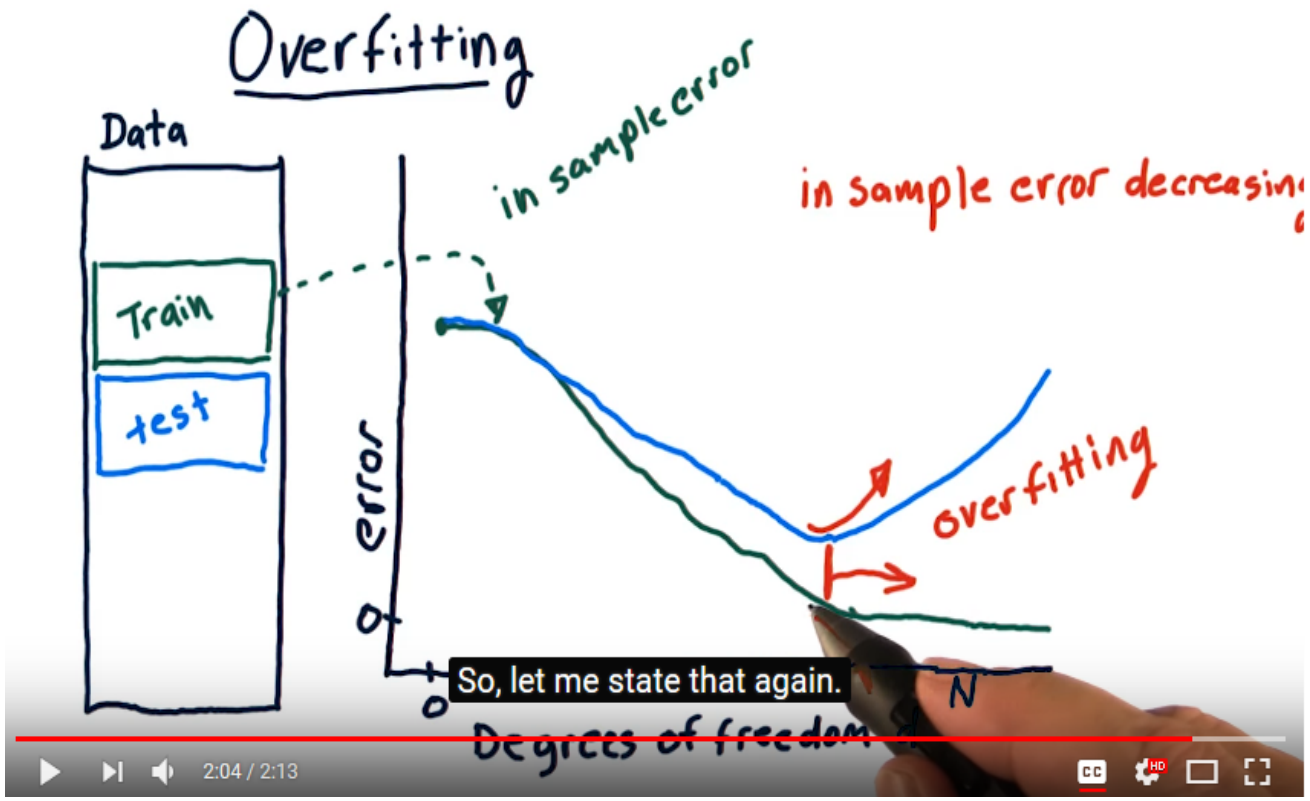
- Overfitting & Errors
  - $E_{in}(g)$ goes down
  - $E_{out}(g)$ goes up
- 'Bad generalization area' ends
  - Good to reduce $E_{in}(g)$
- 'Overfitting area' starts
  - Reducing $E_{in}(g)$ does not help
  - Reason 'fitting the noise'



- **The two general approaches to prevent overfitting are (1) regularization and (2) validation**

➢ **Lecture 9 provides details on validation to be considered as another method against overfitting**

[Video] Overfitting Challenge



[22] YouTube Video, 'Overfitting'

Lecture Bibliography

Lecture Bibliography (1)

- [1] The 2004 Indian Ocean Tsunami
  Online: https://www.thoughtco.com/the-2004-indian-ocean-tsunami-195145
- [2] Five years after Indian Ocean tsunami, affected nations rebuilding better – UN
  Online: http://www.un.org/apps/news/story.asp?NewsID=33365#.WltdyK6nGpp
- [3] Tsunami Damage, Gleebruk, Indonesia January 7, 2005
  Online: https://earthobservatory.nasa.gov/IOTD/view.php?id=5145
- [4] Landsat 7: the seventh satellite of the Landsat program
  Online: https://landsat.gsfc.nasa.gov/landsat-7/
- [5] DigitalGlobe's QuickBird satellite sensor
  Online: https://www.satimagingcorp.com/satellite-sensors/quickbird/
- [6] Arc de Triomphe and the Grand Boulevards from a balloon, Paris, France
  Online: https://www.pinterest.com/pin/291608144593636608/
- [7] Aerial photographs of Schlosshotel Kronberg, near Frankfurt, Germany
  Online: https://en.wikipedia.org/wiki/Pigeon_photography
- [8] First aerial photography taken from an airplane, 1909,near Rome, Italy
  Online: http://professionalaerialphotographers.com/content.aspx?page_id=22&club_id=808138&module_id=158950
- [9] CORONA: American's First Satellite Program: first photograph.
  Online: https://www.oneonta.edu/faculty/baumanpr/geosat2/RS%20History%20II/RS-History-Part-2.html
- [10] Sentinel-2 global coverage
  Online: https://www.youtube.com/watch?v=upXdz6QvI3c

Lecture Bibliography (2)

- [11] Population growth from 1975 - 2010 of Manila
  Online: https://manilabydaniellaandisabel.weebly.com/location-and-characteristics.html
- [12] Deforestation in Bolivia from 1986 to 2001
  Online: https://www.satimagingcorp.com/gallery/more-imagery/aster/aster-deforestation-bolivia/
- [13] Temporary Armor Storage Area, Managua, Nicaragua, 11 November 1981
  Online: https://commons.wikimedia.org/wiki/File:Temporary_Armor_Storage_Area,_Managua,_Nicaragua,_11_November_1981_%E2%80%93_CIA_IMINT.jpg
- [14] National Oceanic and Atmospheric Administration (NOAA): image of Europe, October 11, 2000
  Online: http://www.noaa.gov
- [15] Rome Image dataset
  Online: https://b2share.eudat.eu/records/daf6c389e54340b4b1416cf874251e77
- [16] G. Cavallaro, M. Riedel, et al., "Smart data analytics methods for remote sensing applications," *in the Proceedings of the IEEE Geoscience and Remote Sensing Symposium*, Quebec City, QC, 2014, pp. 1405-1408.
- [17] Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)
  Online: https://www.jpl.nasa.gov/missions/airborne-visible-infrared-imaging-spectrometer-aviris/
- [18] Indian Pines dataset: 220 Band AVIRIS Hyperspectral Image
  Online: https://purr.purdue.edu/publications/1947/1
- [19] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson and A. Plaza, "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," *in the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4634-4646, Oct. 2015.
- [20] Indian Pines Raw and Processed
  Online: http://hdl.handle.net/11304/9ec5eac8-61b4-4617-ae1c-1f8c8cd3cd74

Lecture Bibliography (3)

- [21] Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN 0321321367, English, ~769 pages, 2005
- [22] YouTube Video, ‘Overfitting‘,
  Online: https://www.youtube.com/watch?v=mfzHchd5La8

Slides Available at http://www.morrisriedel.de/talks