Big data and Earth observation
New challenges in remote sensing images interpretation

Pierre Gançarski

ICube
CNRS - Université de Strasbourg

2014

Remote sensing image

What's it ?

- Image captured by an aerial or satellite system :
  - optic sensors : spectral responses of various surface covers associated with sunshine
  - radar sensors
  - Lidar
  - ...

Thanks to the LIVE lab (A. Puissant) and the IPGS lab (J.-P. Malet) for providing images.

Remote sensing image

What's it ?

- Image captured by an aerial or satellite system :
  - optic sensors : spectral responses of various surface covers associated with sunshine
  - radar sensors
  - Lidar
  - ...

Thanks to the LIVE lab (A. Puissant) and the IPGS lab (J.-P. Malet) for providing images.

Remote sensing image

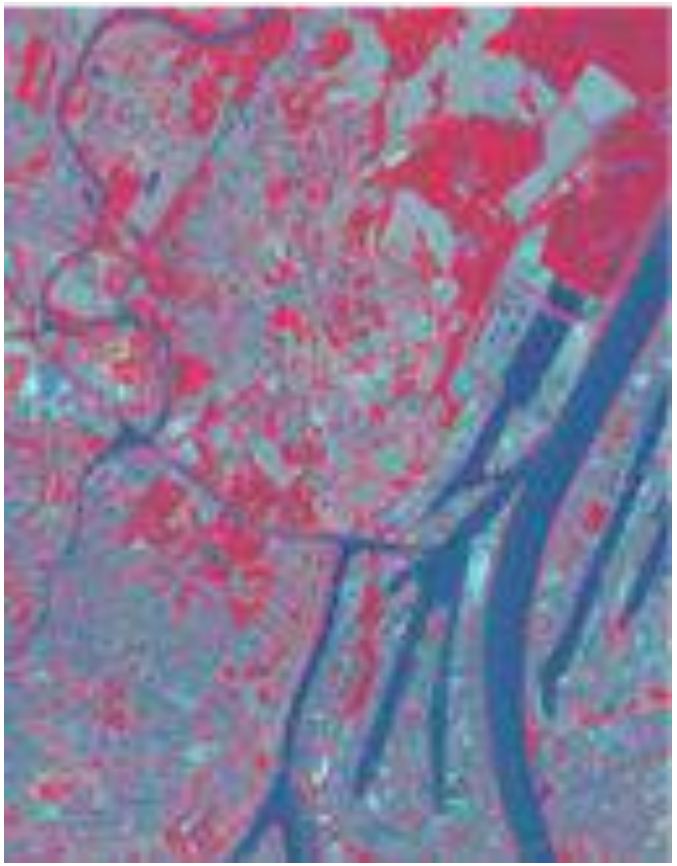
Three dimensions

- spatial resolution : surface covered by a pixel (from 300m to few tens of centimetres)
- spectral resolution : number of spectral information (from blue to infrared) corresponding to the number of sensors
- radiometric resolution : linked to the ability to recognize small brightness variations (from 256 to 64000 level)

Remote sensing image

Spatial resolutions

- High spatial resolution (HSR) : 20 or 10m



Green

Red

Near infrared (NIR)

Remote sensing image

Spatial resolutions

- Very high spatial resolution (VHSR) : from 5m to 0.5m

Remote sensing image

Three dimensions

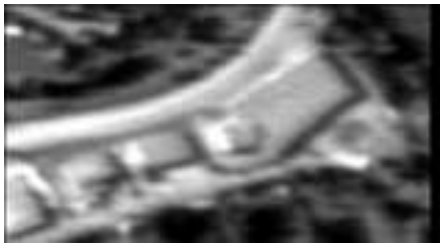
- spatial resolution : surface covered by a pixel (from 300m to few tens of centimetres)
- spectral resolution : number of spectral information (from blue to infrared) corresponding to the number of sensors
- radiometric resolution : linked to the ability to recognize small brightness variations (from 256 to 64000 level)

Remote sensing image

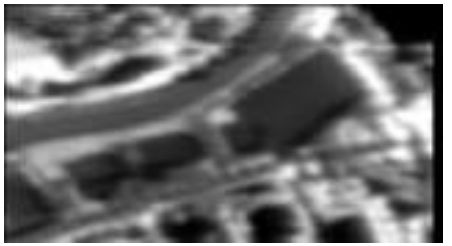
Spectral resolutions

- Hight spectral resolution : One hundred of radiometric bands (or more)

Aerial view



band #1



band #22

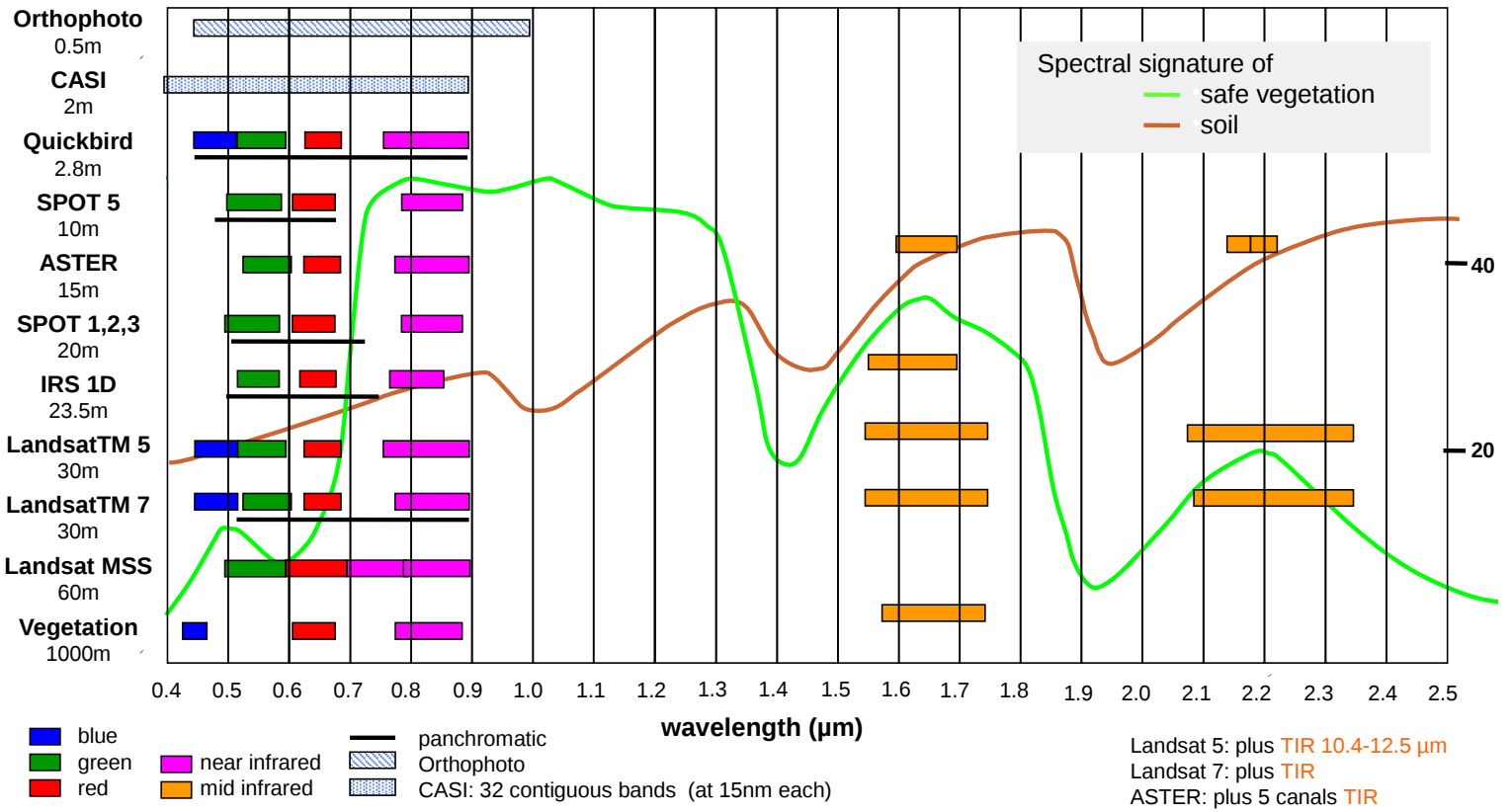
band #29

band #38

Remote sensing image

Image interpretation

- Discrimination between (kinds of) objects can depend on the spectral resolution



Remote sensing image

Three dimensions

- spatial resolution : surface covered by a pixel (from 300m to few tens of centimetres)
- spectral resolution : number of spectral information (from blue to infrared) corresponding to the number of sensors
- radiometric resolution : linked to the ability to recognize small brightness variations (from 256 to 64000 level)

Image interpretation

Semantic gap

- There are differences between the *visual* interpretation of the spectral information and the semantic interpretation of the pixels
- The semantic is not always explicitly contained in the image and depends on domain knowledge and on the context.

$\Longrightarrow$ This problem is known as the *semantic gap* and is defined as the lack of concordance between low-level information (*i.e.* automatically extracted from the images) and high-level information (*i.e.* analyzed by geographers)

Big - Data Science

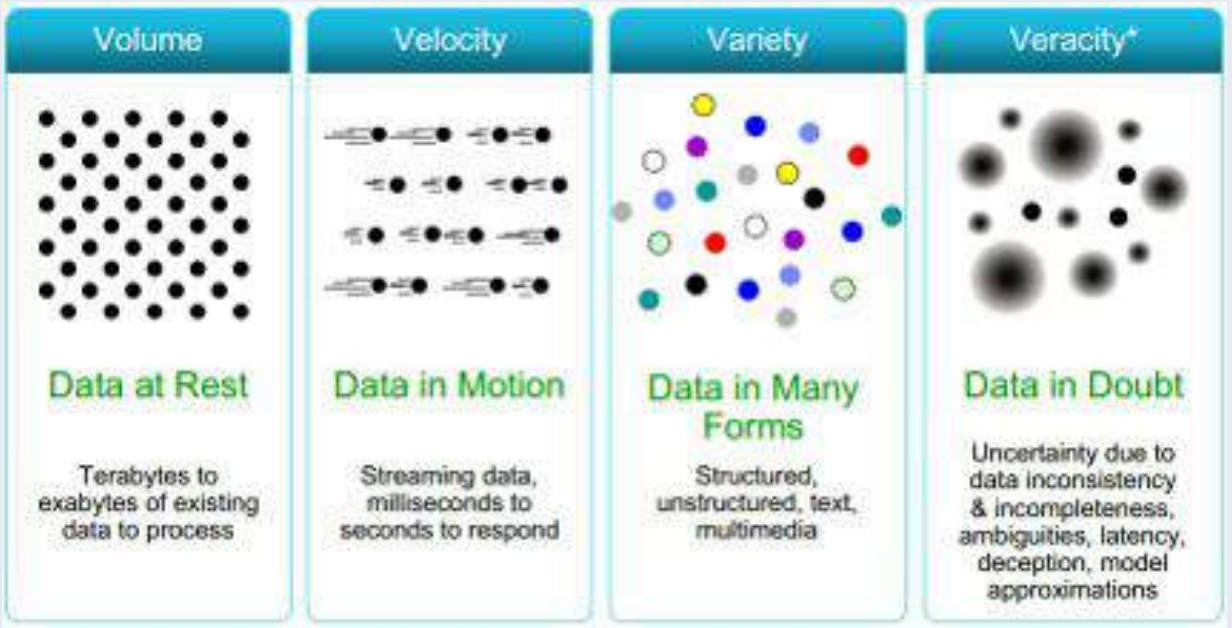
Big Data

- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications

Big - Data Science

Big Data

- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
- Data scientists break it into 4 dimensions : 4 V's of Big Data



Michael Walker 2012

Big - Data Science

Big Data : Volume

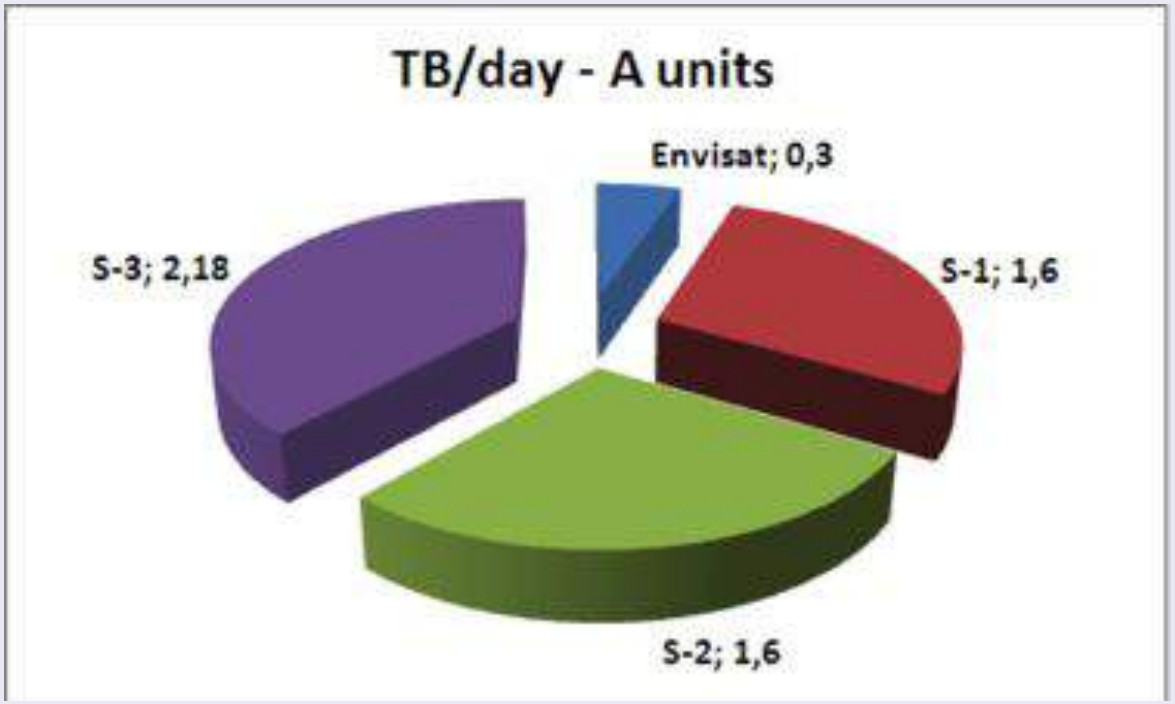
- Systems produce more data than ever before and at a pace unmatched in human history.
  - Internet : in 2020, around 10 zettaoctects, (10 000 billions Go) by month.
  - Large Hadron Collider (LHC) : 5 pétaoctets of scientific data by year

$\rightsquigarrow$ Earth observation : Copernicus - European system for monitoring the Earth

Big - Data Science

Sentinel : data volume

- Copernicus collects data from multiple sources : earth observation satellites (called Sentinel) and in situ sensors such as ground stations, airborne and sea-borne sensors.
- Sentinel : #1 - 1,6 To/day, #2 - 1,6 To/day, #3 - 2,2 To/day



Big - Data Science

Big Data : Variety

- Data comes in many forms
  - Purchase history
  - Social graphs, tweets, blog posts
  - Scientific data
  - Images and videos

$\rightsquigarrow$ Earth observation : Radar, optical images ; texts ..

Big - Data Science

Sentinel : data variety

- Sentinel #1 : radar imagery for land and ocean services. (Sentinel-1A, 3 April 2014. - Sentinel-1B, 2016).
- Sentinel #2 : high-resolution optical imagery (2016).
- Sentinel #3 : high-accuracy optical, radar and altimetry data (2016)
- Sentinel #4 : atmospheric composition (around 2020)
- Sentinel #5 : atmospheric composition (around 2020)

\+ in situ sensors such as ground stations, airborne and sea-borne sensors

\+ existing systems : Pleiades system, HYPXIM program (Hyperspectral imagery), ...

Big - Data Science

Velocity

- Systems produce more data than ever before and at a pace unmatched in human history.
  - Twenty years ago, commercial life revolved around monthly measurements. Today, measurements occur as frequently as every second.

$\rightsquigarrow$ Earth observation : Sentinel Program - 5 To a day

Big - Data Science

Sentinel : data velocity

- A global revisit of the French territory
  - every 10 days (early 2015) : 20 Go to 35 Go a day
  - every 5 days (2016) : 40 Go to 70 Go a day
  - "Ultimate" objective : every day (2020 ?) 80Go à 140 Go/jour

Big - Data Science

Big Data : Veracity

- Data can come from everybody and from everywhere. It's hard to know which information is accurate and which is out of date.

⇝ Earth observation :

  - sensor error, cloud, atmospheric distorsion ...
  - confidence in documents from the WEB

Challenge #1 : Multilevel analysis

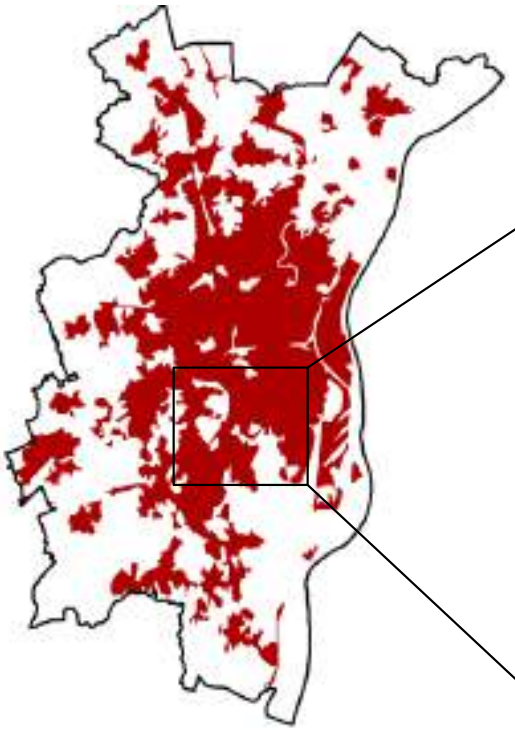
Context

- Multi levels analysis : there are a need to locate, identify and analyze geographic objects at different scales
- For instance :
  - Urban planners are interested in up-to-date land cover and land use information on urban objects at several spatial (1 :100,000 to 1 :5,000) and temporal scales
  - Geologists and geophysicists are interested by accurate mappings of landslides, providing relevant informations about potential environmental and/or human risks and by a better comprehension of the landslides structure by analyzing their different sub-parts (scarp, track and deposit areas)

Challenge #1 : Multilevel analysis

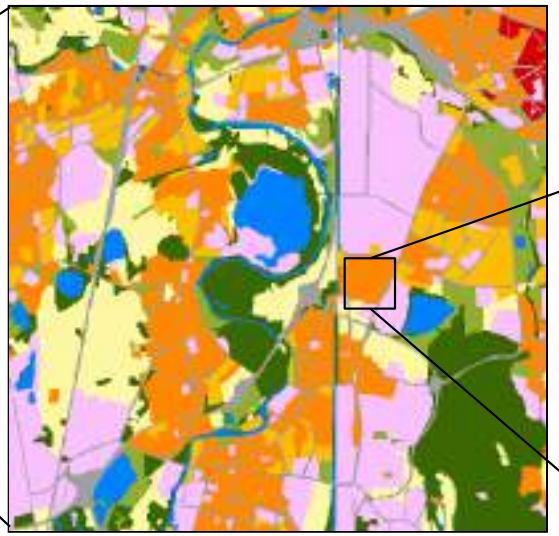
Urban multilevel analysis

Urban area
1:100,000 – 1:50,000



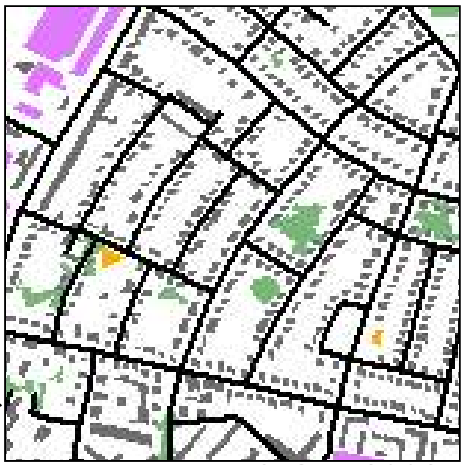
CLC ©IFEN, 2000

District
1:50,000 – 1:10,000



BDOCS ©CIGAL, 2000

Urban object
1:5,000 – 1:2,000



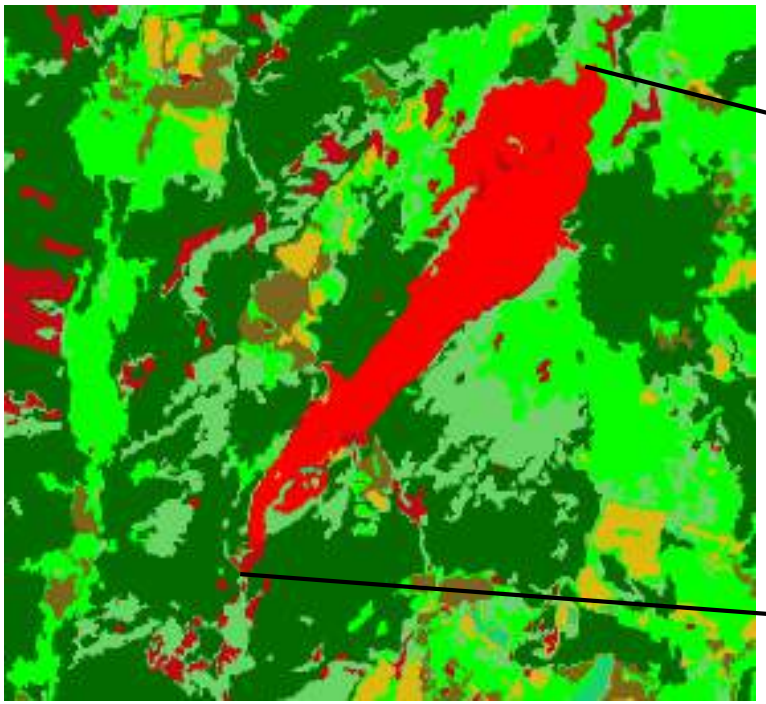
BDTopo ©IGN, 2002

Strasbourg

Challenge #1 : Multilevel analysis

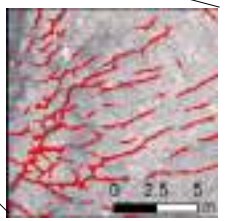
Landslide multilevel analysis

Landslide structures



Landslide sub-parts

Fissures



Challenge #1 : Multilevel analysis

Context

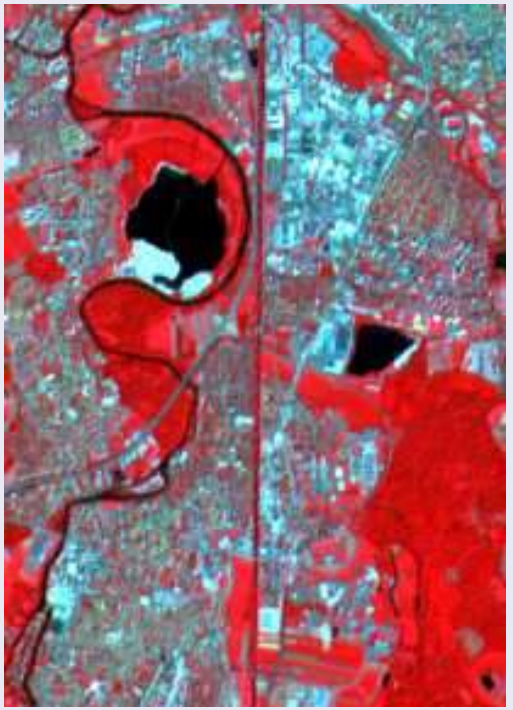
- Multi levels analysis
- Acquiring this information by ground survey techniques is complex, difficult and time-consuming

→ The increasing availability of remote sensing images is an opportunity to extract, characterize and identify the objects of interest.

Challenge #1 : Multilevel analysis

Urban analysis

20 m ©SPOT 4



2.8m © Quickbird

- MSR - 20m : Usable to extract and characterize urban areas
- VSR - 2.8m : Usable to extract and characterize urban objects

Challenge #1 : Multilevel analysis

Lansdlide

20 m ©Landsat

5m © RapidEye

0.5m © IGN

- MSR - 20m : Usable to extract and characterize natural areas
- HSR - 5m : Usable to extract and characterize landslide object
- VHRS - 2.5m : Usable to extract and characterize object sub-parts (fissures...)

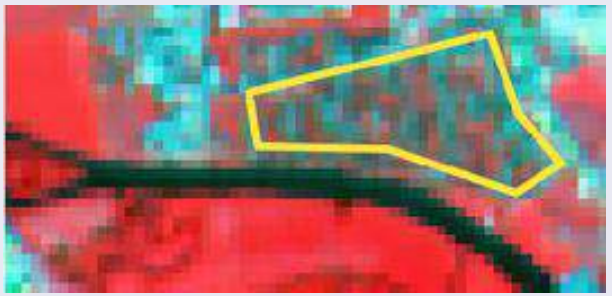
Challenge #1 : Multilevel analysis

Major problem

- How analyse areas at a semantic level which does not directly correspond to a image resolution ?

Urban blocks ?

- Homogeneous patterns of urban elements defined by the minimal cycles closed by communication ways



Urban blocks on MSR image

Urban blocks on HSR image

$\rightarrow$ Neither resolutions (HSR or MSR) is adapted

Challenges

My point of view

- Multilevel analysis : use of all the data in the same time

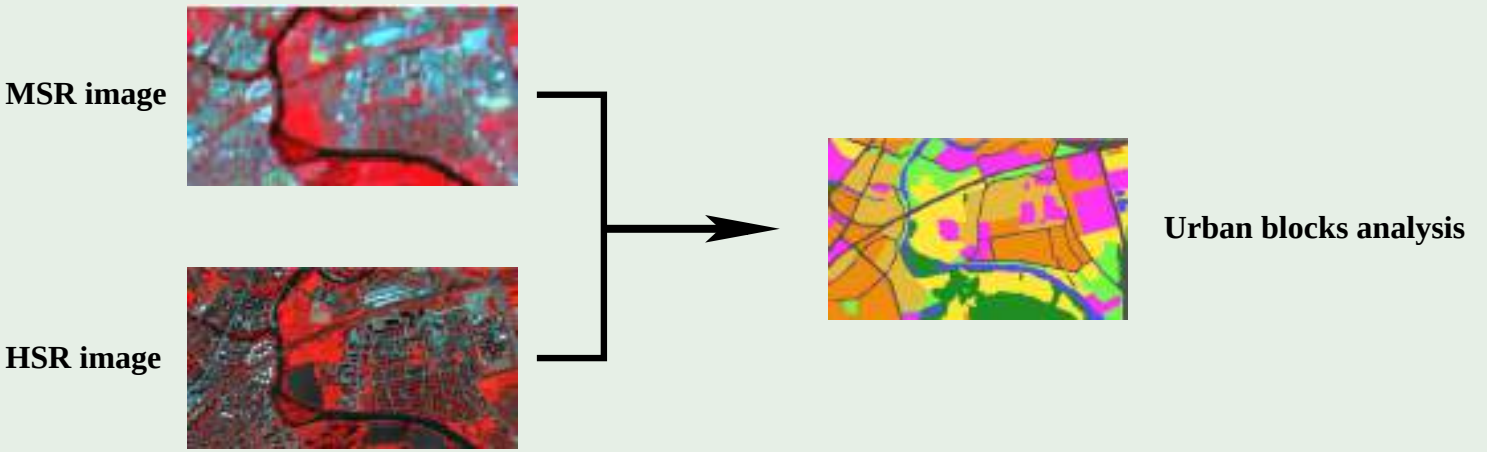
Challenge #1 : Multilevel analysis

Major problem

- How analyse areas at a semantic level which does not directly correspond to a image resolution ?
- How use all the image avaiable on the studied area ?

Idea

- Using the complementarity of the different resolutions
- For instance :



Challenge #2 : Analysis methods

Context

- A (too ?) large panel of methods exists and each of them is often effective but ...
- None is superior to all others in all cases. Its depends
  - on objectives of the mining
  - on time available
  - on kinds of data
  - ...

$\Rightarrow$ In fact, success of data mining, depends in almost all cases, on the ability of the expert to select and configure the algorithm to be used

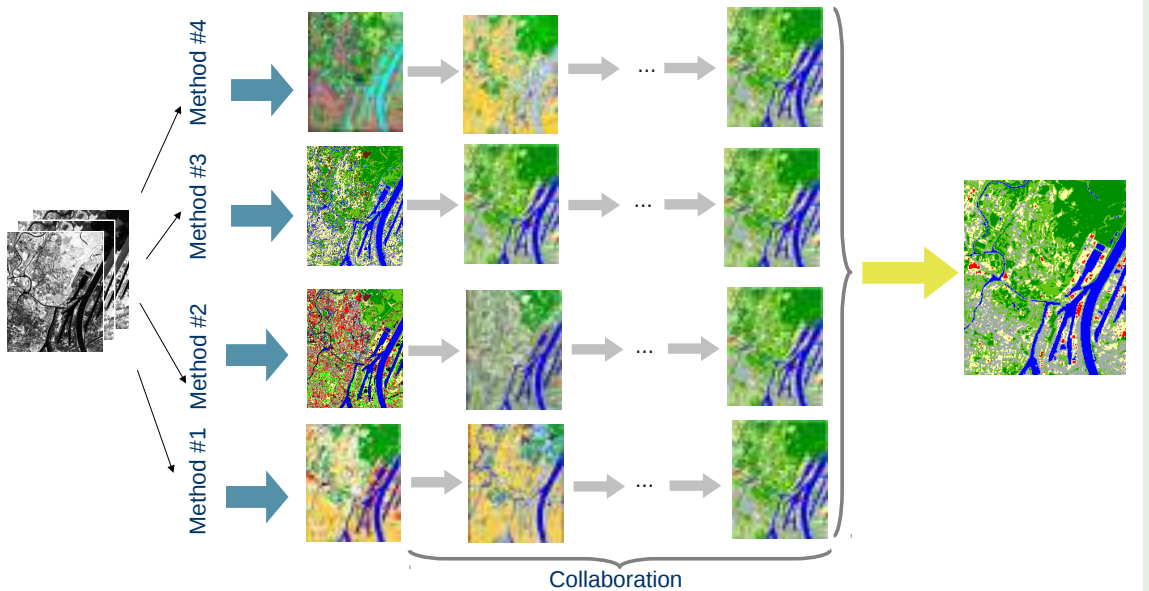
Challenge #2 : Analysis methods

Major problem

- How use the efficacy and complementary of this methods without need of deep knowledge about them ?

Idea

- Using them in a collaborative way



→ Collaborative Multistrategy collaborative classification

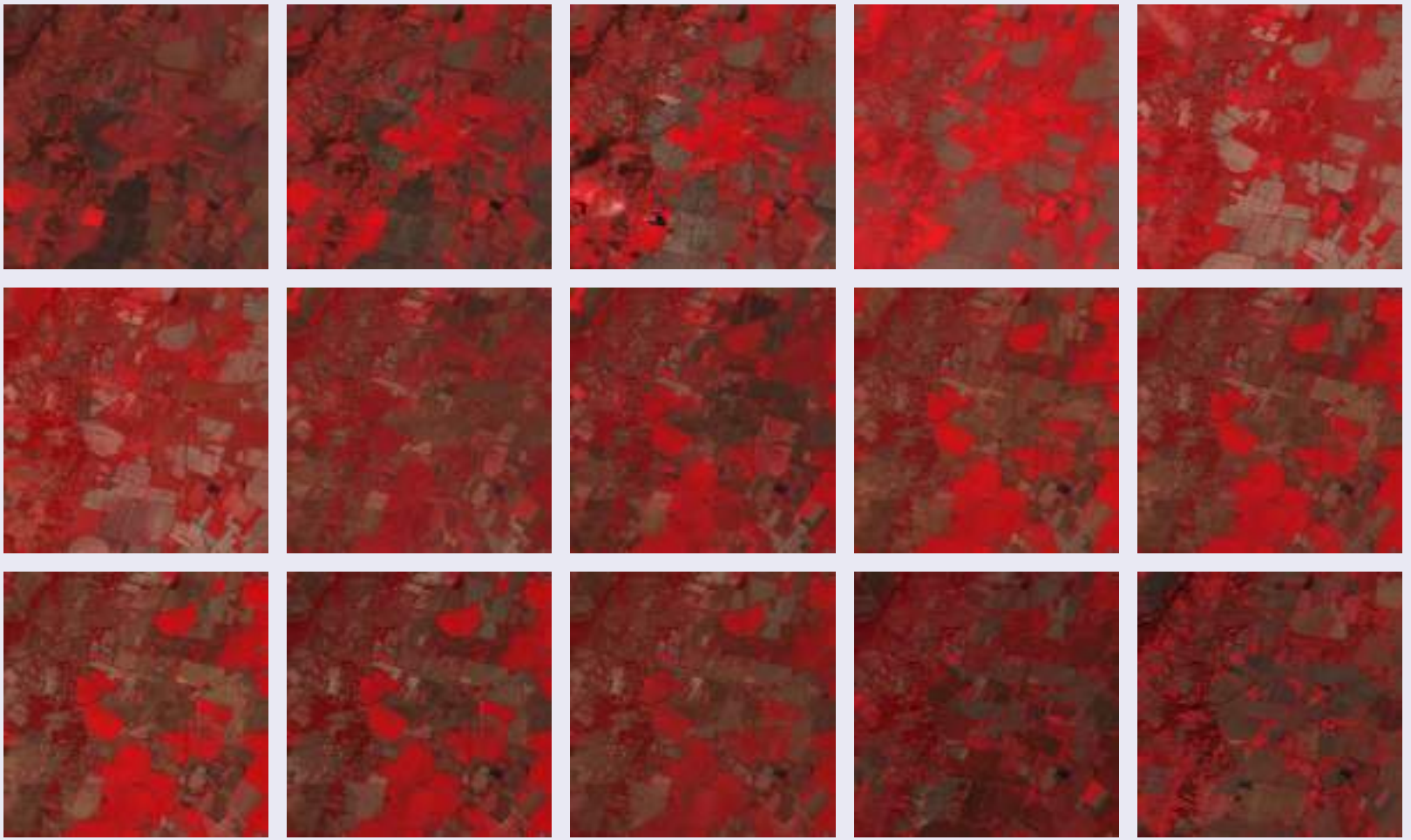
Challenges

My point of view

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods

Challenge #3 : Multitemporal methods

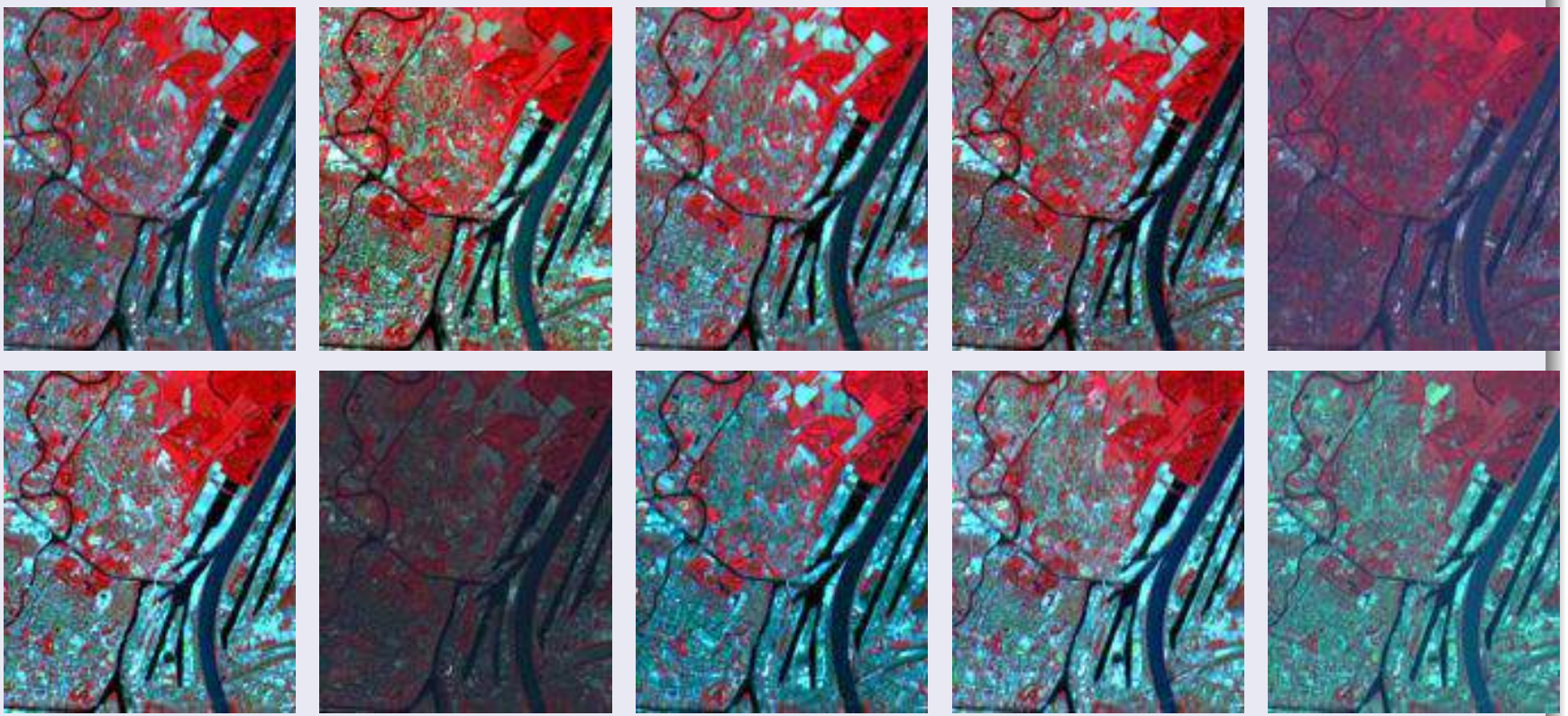
Context : Satellite Image Time Series



© NSPO

Challenge #3 : Multitemporal methods

Context : Satellite Image Time Series



Challenge #3 : Multitemporal methods

Two types of change

- long-term changes (ex : urbanization)
- cyclic changes (ex : agriculture)

Phenomenons are temporally distorted

- Some phenomenons are temporally distorted.
  - Each urbanization state (bare soil, new house, old house) can be temporally distorted.
  - Agronomy
    - Shifted harvest/collect from one parcel to another.
    - Ripening/Maturation speed difference.
- but belong to the same thematically class.

Challenge #3 : Multitemporal methods

Two types of change

- long-term changes (ex : urbanization)
- cyclic changes (ex : agriculture)

Observation of these phenomenons is irregular

- the observed phenomenon may be past $\rightarrow$ we have to deal with available images ;
- clouds can hide the phenomenon on certain dates ;
- images remain expensive $\rightarrow$ we may want to choose the temporal sampling of the image series depending on the period ;
- the satellite may be unavailable.

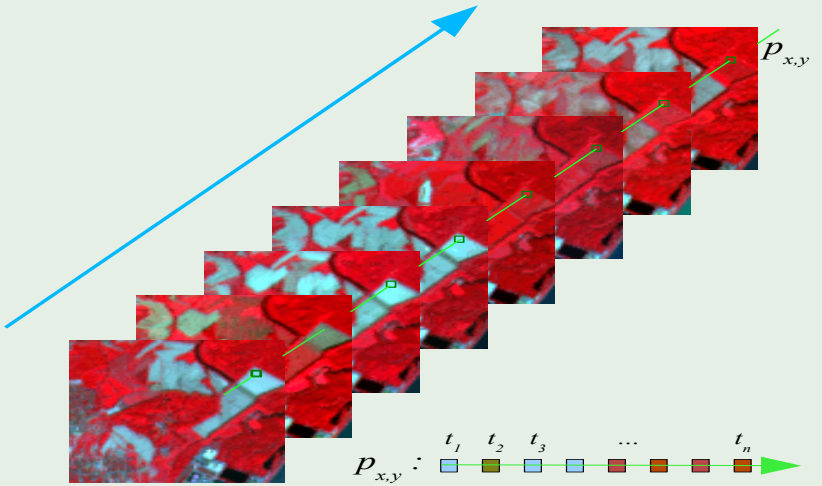
Challenge #3 : Multitemporal methods

Major problem

- How extra information from such SITS ?

Idea

- Define each pixels radiometric value serie as a sequence :



- Use of classical algorithms on the sequences to find global significative behaviors $\rightsquigarrow$ urbanization,agricultural cycle ...

Challenges

My point of view

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kind of analysis

Challenge #4 : High frequency of observations

Context

- High frequency of observations impacts methodologies :
  - to update databases or background knowledge
  - to extract geographic object evolutions
- High frequency associated to big volume makes that is unrealistic to be able to process all the data for each new observation

Challenge #4 : High frequency of observations

Major problem

- How to extract information from image time series with high frequency ?

Idea

- Take advantage of this high frequency to define new types of applications :
  - agricultural monitoring and forecasting in the short term for instance
- Assuming that in such series, the changes are minimal between two observations, define incremental methods able :
  - of adapting the classifications or indices extracted from images
  - of updating the thematic concepts (drift concepts)

Challenges

My point of view

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kinds of analysis
- High fréquency in image acquisition : new incremental methods

Challenge #5 : Background knowledge

Context

- Semantic gap is defined as the lack of concordance between low-level information (i.e. automatically extracted from the images) and high-level information (i.e. analyzed by urban experts)
  - in order to reduce this gap, image analysis methods using region-based (or object-based) approaches are developed these last ten years
- Some initiatives have focused on the use of domain knowledge for classifying urban objects
  - A major issue in these approaches is formalization and exploitation of such knowledge : building a knowledge-base is a difficult task because the knowledge is most of the time implicit and held by the domain experts.

Challenge #5 : Background knowledge

Major problem

- How to exploit domain knowledge to facilitate extraction of relevant information from images (time series) ?

Idea

- Build a knowledge-base as an ontology :
  1. Identification of the geographic concepts
  2. Formalization of these concepts
- Definition of relations between thematic objects (from the ontology) and image-based objets $\rightarrow$ need of an ontology associated to the images
- Integration of mechanisms able to take in account such knowledge into classification algorithms

Challenges

My point of view

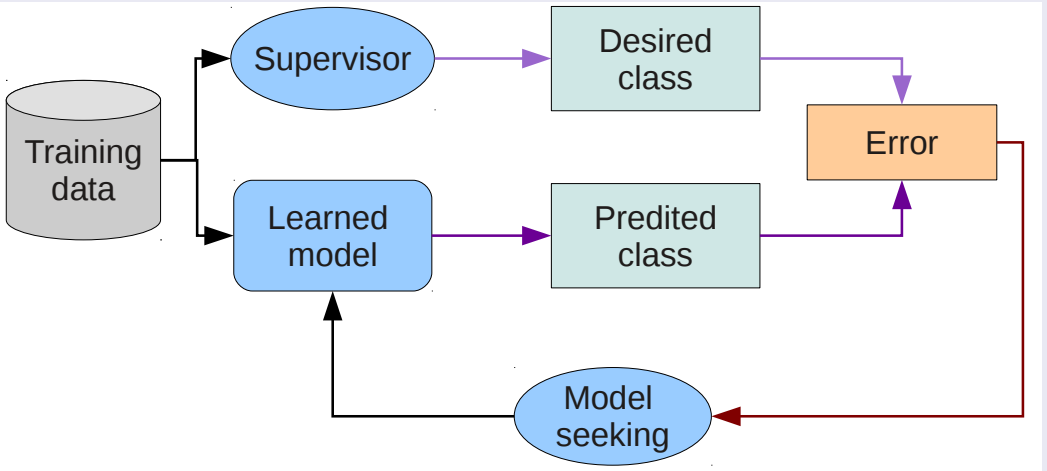
- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kinds of analysis
- High fréquency in image acquisition : new incremental methods
- Background knowledge : need to strengthen the links between geographer and computer scientists.

Challenge #6 : Lack of expertise

Context

- Supervised classification is the task of inferring a model from labelled training data : each example is a pair consisting of an input object and a desired class.

$\Rightarrow$ The classes to be learned are known a priori



- Classical methods : Support vector machine, Decision tree, Artificial neural network. . .

Challenge #6 : Lack of expertise

Supervised classification : case of VHSR

How many classes ?

- 10 ?
  - Road
  - Building ...
- 50 ?
  - Road
  - Street
  - Red car
  - Blue car
  - Lighted (half) roof
  - Shadowed roof ...

Challenge #6 : Lack of expertise

Supervised classification : case of VHSR

How to give enough examples by class with high number of classes ?

Challenge #6 : Lack of expertise

Supervised classification : case of VHSR

- Lack of sample in VHRS images
- It does not exist (for the moment) exploitable ground truth about geographic object evolutions. Build such one :
  - is labor-intensive, time-consuming, expensive
  - requires that the experts have defined the types of sought changes

Challenge #6 : Lack of expertise

Major problem

- How to exploit images or image time series with no or insufficient training dataset (lack of ground truth) ?

Idea

- Unsupervised methods : clustering
- Supervised clustering : use of (very) small training dataset to guide clustering algorithm
- Integration of human-expert in the process (Active learning) :
  1. Starting from (relatively few) training examples, the system learns a model
  2. It evaluates the model using the unlabeled objects
  3. From this evaluation, the system ask to the expert, the label of specific objects (e.g., at the border of two classes)
  4. These objects are added to training examples
  5. The process is iterated until a given quality criterion is satisfied

Challenges

My point of view

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kinds of analysis
- High frequency in image acquisition : new incremental methods
- Background knowledge : need to strengthen the links between geographer and computer scientists.
- Lack of expertise : use of unsupervised (or guided) approaches

Challenge #7 and so on : scale-up, data and knowledge quality

Major problem

- #7 Scalability :
  - How to exploit huge (distributed) dataset of images ?
  - How to deal with the huge picture (20000 x 20000 pixels, for example) that can not fit into memory ?
- #8 Quality of data/knowledge :
  - How to find, evaluate and correct errors in raw data (image) or segmentation data ?
  - How to find, evaluate and correct errors or incoherencies in knowledge base (how to trust the expert ?)
- #9 Robustness of algorithms :
  - How to take into account errors in raw data (image) or segmentation data ?
  - How to take into account errors or incoherency in knowledge base

Challenge #7 and so on : scale-up, data and knowledge quality

Major problem

- #7 Scalability
- #8 Quality
- #9 Robustness of algorithms :
- #10 Multisource :
  - How to use other kinds of data : texts, domain ontologies ?
  - What kind of knowledge to be extracted ?
  - How to combine methods from others domains (geography, linguistic, ...)

Challenges

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kinds of analysis
- High fréquency in image acquisition : new incremental methods
- Background knowledge : need to strengthen the links between geographer and computer scientists.
- Lack of expertise : use of unsupervised (or guided) approaches
- Scalability : need to rethink the algorithms.
- Quality : define method to evaluate and correct error or imprecision in data as well as in knowledge
- Quality : define algorithm able to take into account error/imprecision in data as well as in knowledge
- Multisource : combine all the information available on the studied areas regardless their media

Challenges

- Multilevel analysis : use of all the data in the same time
- Multistrategy collaborative classification : take advantage of complementarity of heterogenous methods
- Multitemporal analysis : opportunity of new kinds of analysis
- High fréquency in image acquisition : new incremental methods
- Background knowledge : need to strengthen the links between geographer and computer scientists.
- Lack of expertise : use of unsupervised (or guided) approaches
- Scalability : need to rethink the algorithms.
- Quality : define method to evaluate and correct error or imprecision in data as well as in knowledge
- Robustness : define algorithm able to take into account error/imprecision in data as well as in knowledge
- Multisource : combine all the information available on the studied areas regardless their media

but above all

The most important challenge : what kind of new services can be offered to end-users ?

(Out of scope of my speech...)

Summary

A challenging domain ...

- Multi levels analysis : need to use all the available data source and methods
  - Multi strategy classification
  - Multi resolution classification
- Temporal analysis
  - Multi temporal classification
  - Time series incremental analysis
- Knowledge-based analysis
- Interdisciplinary to integrate/merge/ ... data and knowledge from different domain (STIC, SHS, ...) : images analysis, data mining, text analysis, environment sciences, geosciences, ...