



Data Mining

Prof. Dr. Eng. habil. Mihai Datcu



Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

The header features a historical map of North America with the text 'THE NORTH' visible. Overlaid on the right side of the map are the mathematical expressions $P(y|x, M)P(x|M)$ in a light, serif font.

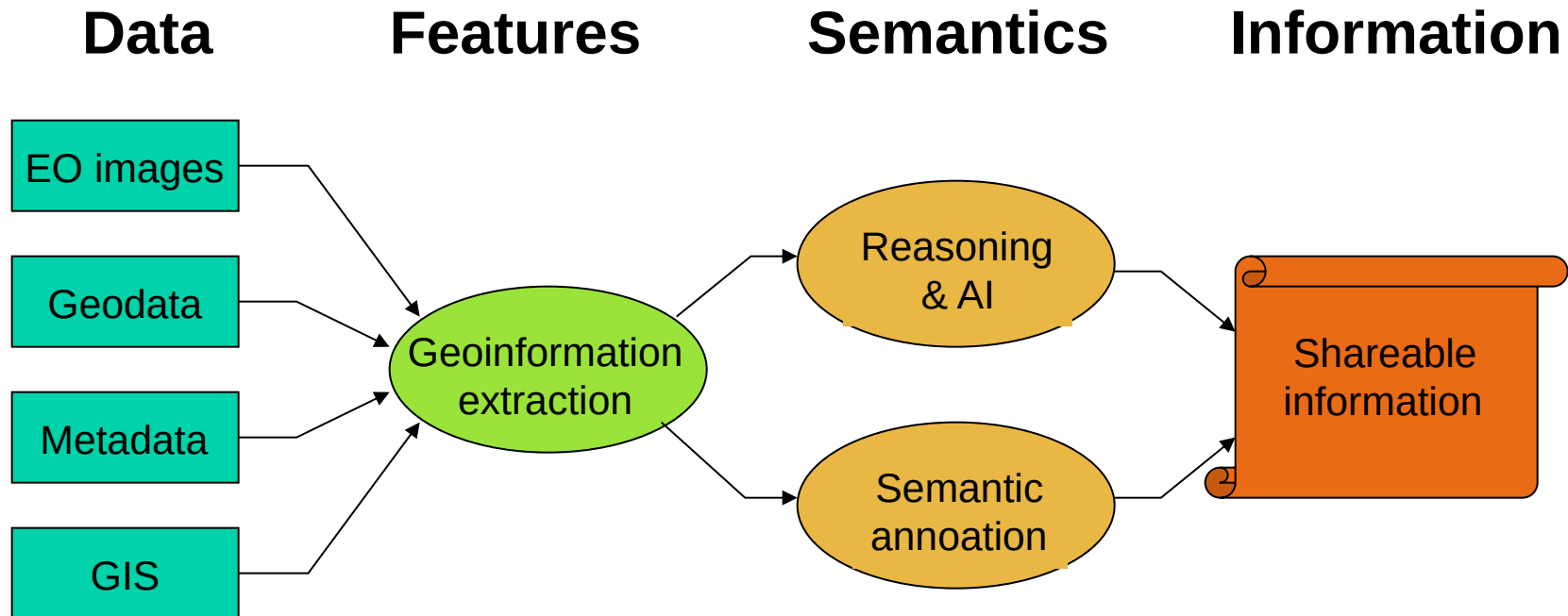
EO Data Mining

EO Data Mining, in analogy to other forms of Data Mining, such as text or semantic web technologies, aims at making the image content accessible.

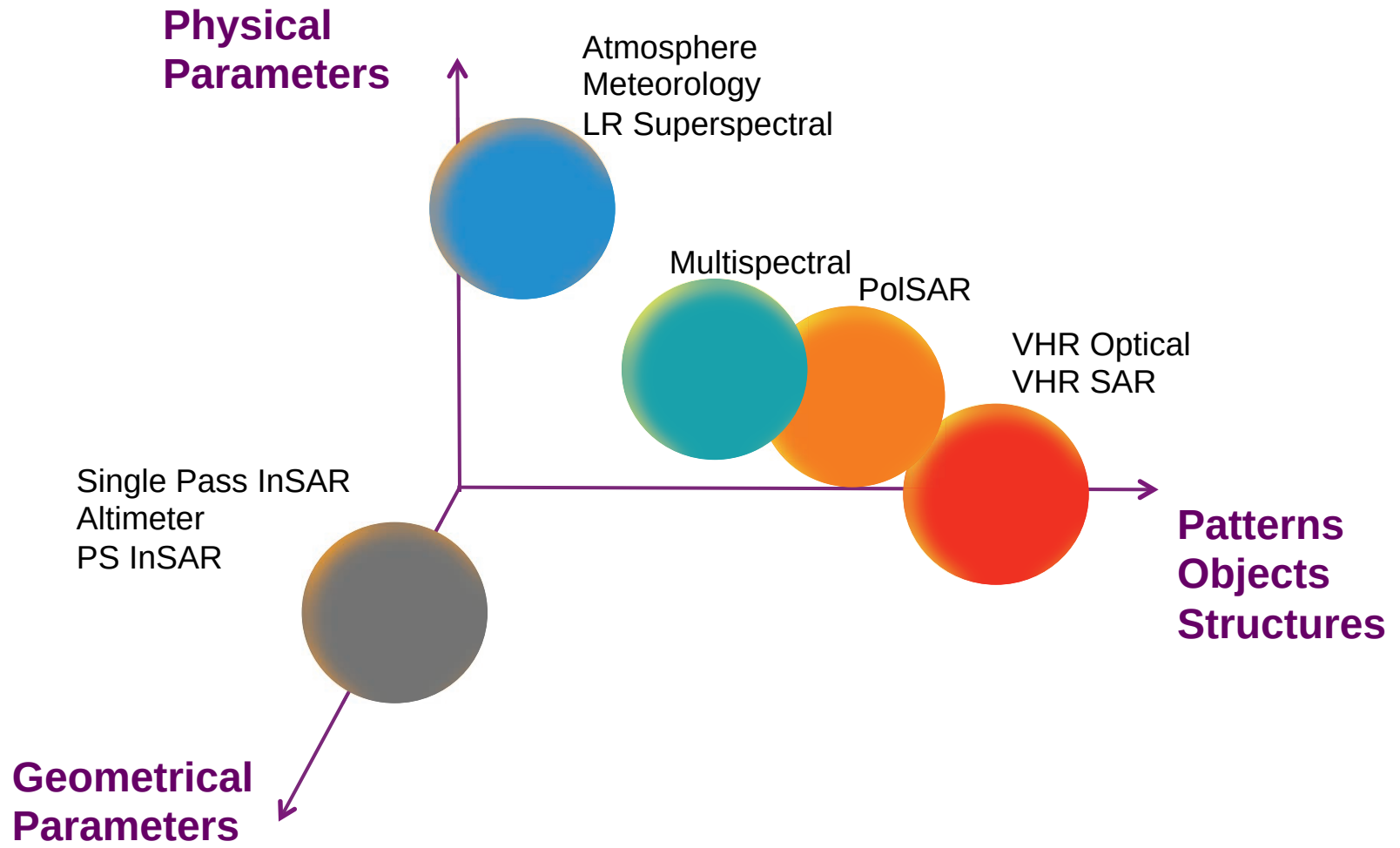
It enables:

- Automatically highlighting of satellite image subsets
- Generates semantic catalogues
- Explores and helps finding specific objects of interest
- Assists image content assessment and understanding
- Supports time-consuming visual interpretation
- Performs analysis on multi-temporal and multi-sensor datasets.

The Concept: from Data to Information



A taxonomy of EO sensors



A [log]History

2014 Sentinel 1 Free data at 5 m resolution

2010 TanDEM-X Free scientific data at 1-3m resolution

2007 CosmoSky Med, RADRSAT 2

TerraSAR-X Free scientific data at 1m resolution

2000 SRTM First time > 80% of continental DEM in 30 m grid

1992 ERS-1 First time a lot of free data !

1974 Sea-SAT First EO spaceborne SAR

1954 Wiley Patent on SAR

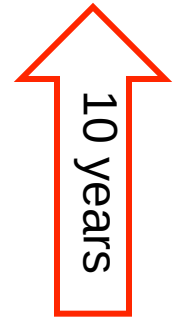
WW I – WW II RADAR

1904 Hulsmeier Patent on Telemobilskope

1897 Marconi First wireless communication

1888 Hertz Experimental validation

1865 Maxwell Theory of electromagnetic waves



KAISERLICHES



PATENTAMT.

History 1900s

PATENTSCHRIFT

— № 165546 —

KLASSE 74 d.

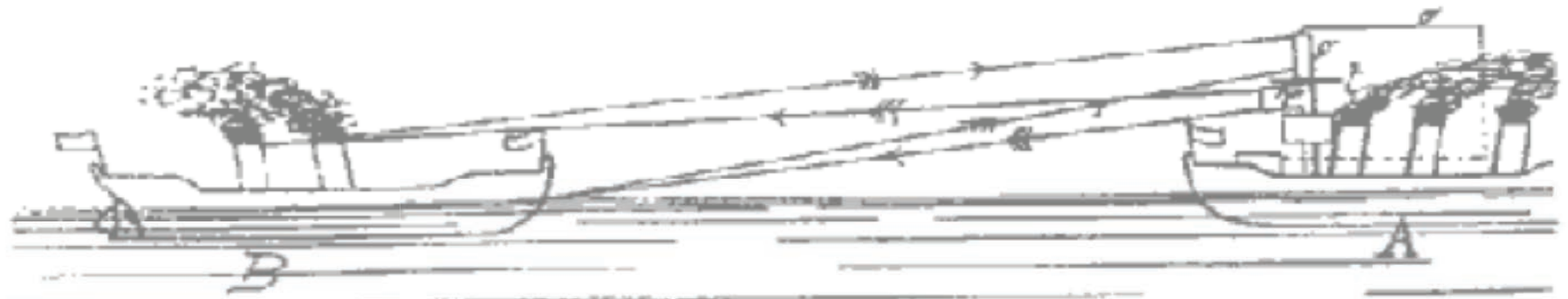
VERGEGEBEN DEN 21. NOVEMBER 1905.

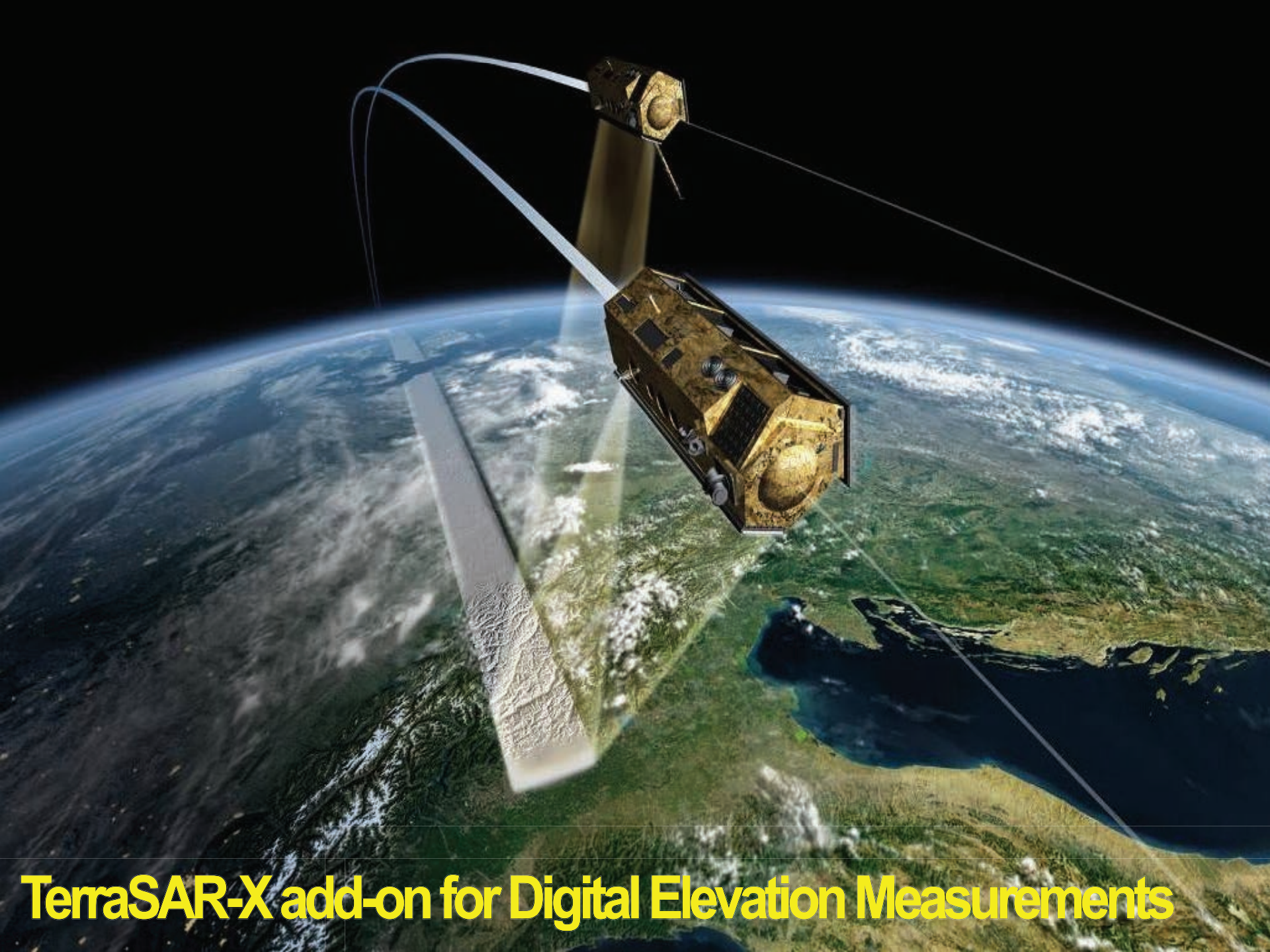
CHR. HÜLSMEYER IN DÜSSELDORF.

Verfahren, um entfernte metallische Gegenstände mittels elektrischer Wellen
einem Beobachter zu melden.

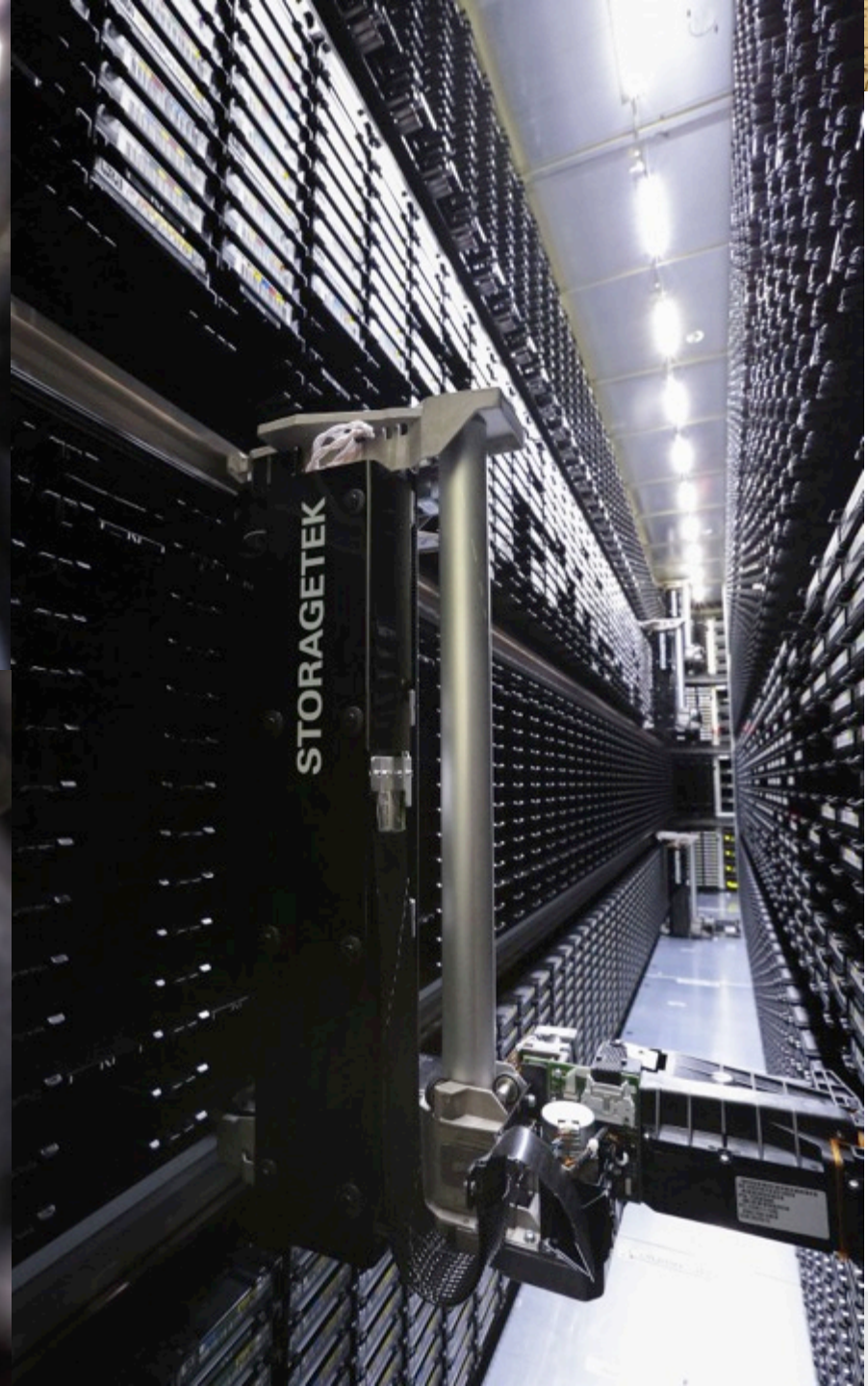
Patentiert im Deutschen Reiche vom 30. April 1904 ab.

Fig. 1.





TerraSAR-X add-on for Digital Elevation Measurements



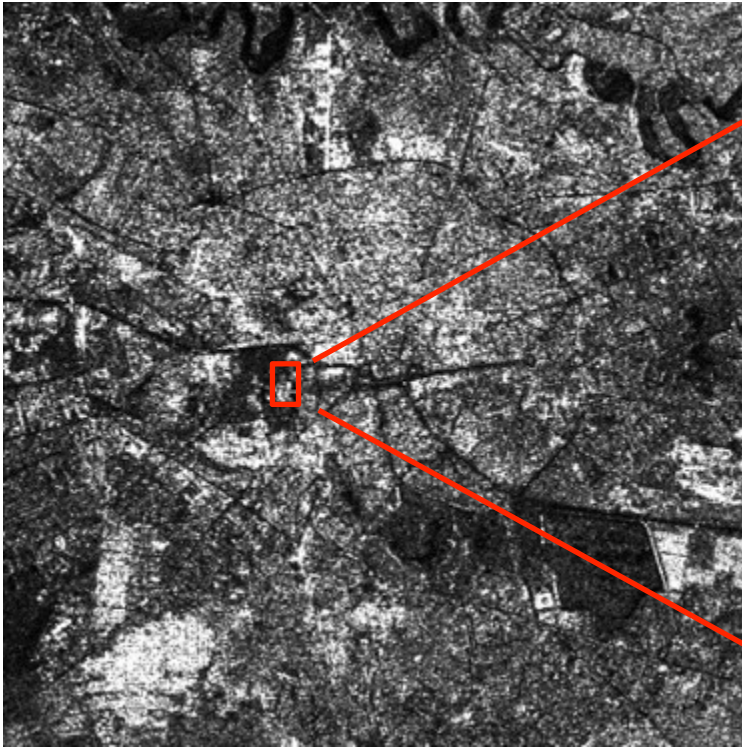


The EO Big Data

- EO Archive = $\mathcal{O}\{10\,000\,000\}$ EO Products
- EO product = “image” + “xml file”
- “image” = 30 000 x 30 000 pixels
- “pixel” = 16/32.../1000s bits, real/complex/multiband
- “xml file” = a lot about the imaging mode, orbit, position, timing
...
- TOTAL = many ZettaBytes



Information vs. data

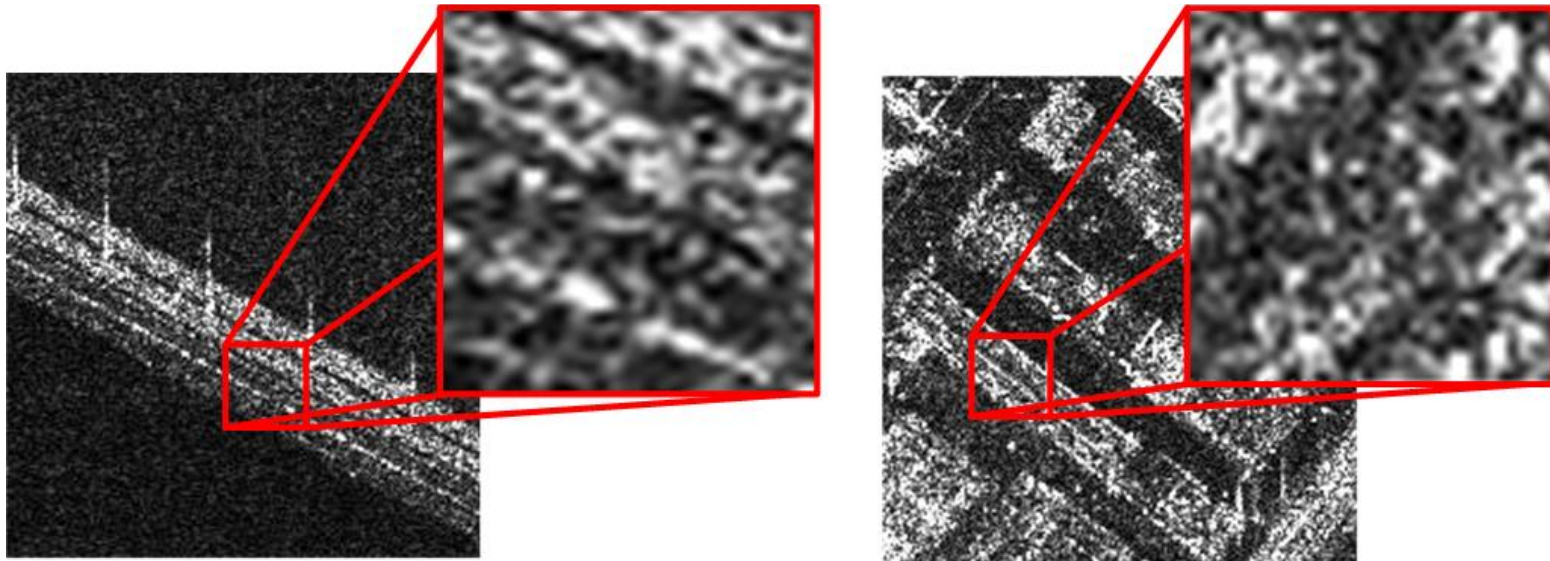


ERS1
24-JUL-1992
512x512 pixels



TerraSAR-X
11-OCT-2008
512x512 pixels

SPATIAL CONTEXT & INFORMATION CONTENT

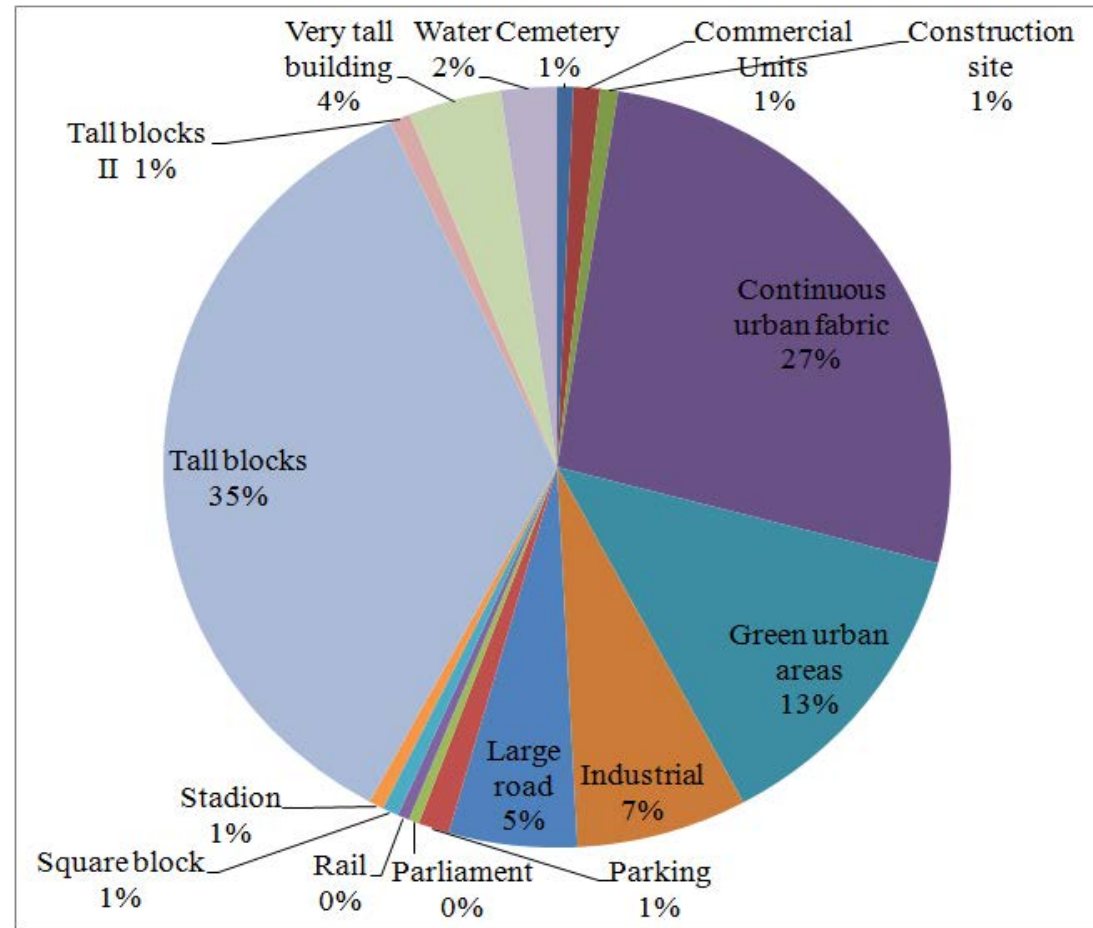


At meter resolution, if context is ignored (the analysis windows being used are relatively small as compared to the object scale), very different scene classes may be confused. The example shows a bridge (left) and buildings (right) which are comprehensible only when the window size is sufficiently large to incorporate relevant context. Otherwise, both scene classes “look” similar and can be confused.



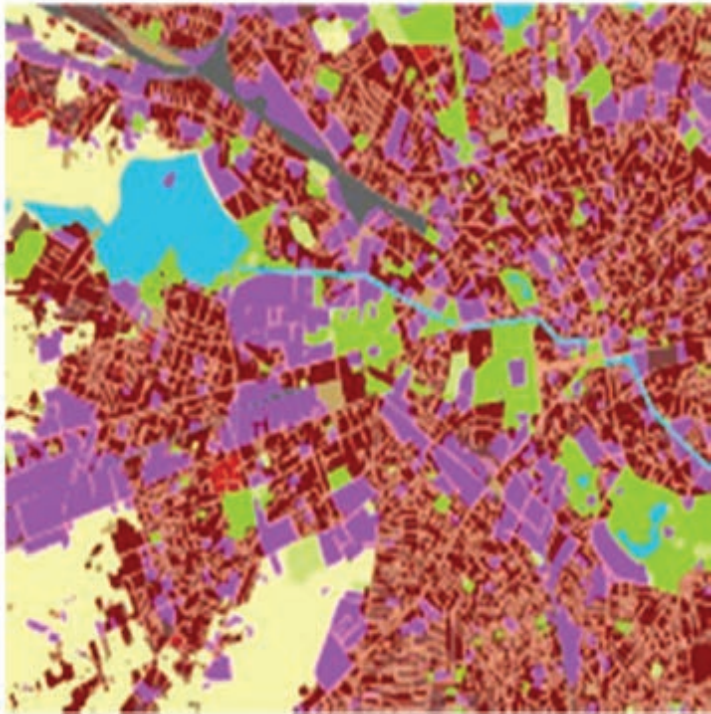
SCENE CATEGORIES & INFORMATION CONTENT

1 HS TerraSAR-X Scene =
up to 10 000 image patches
(100 x 100 m)



EXOGENOUS SOURCES & INFORMATION CONTENT

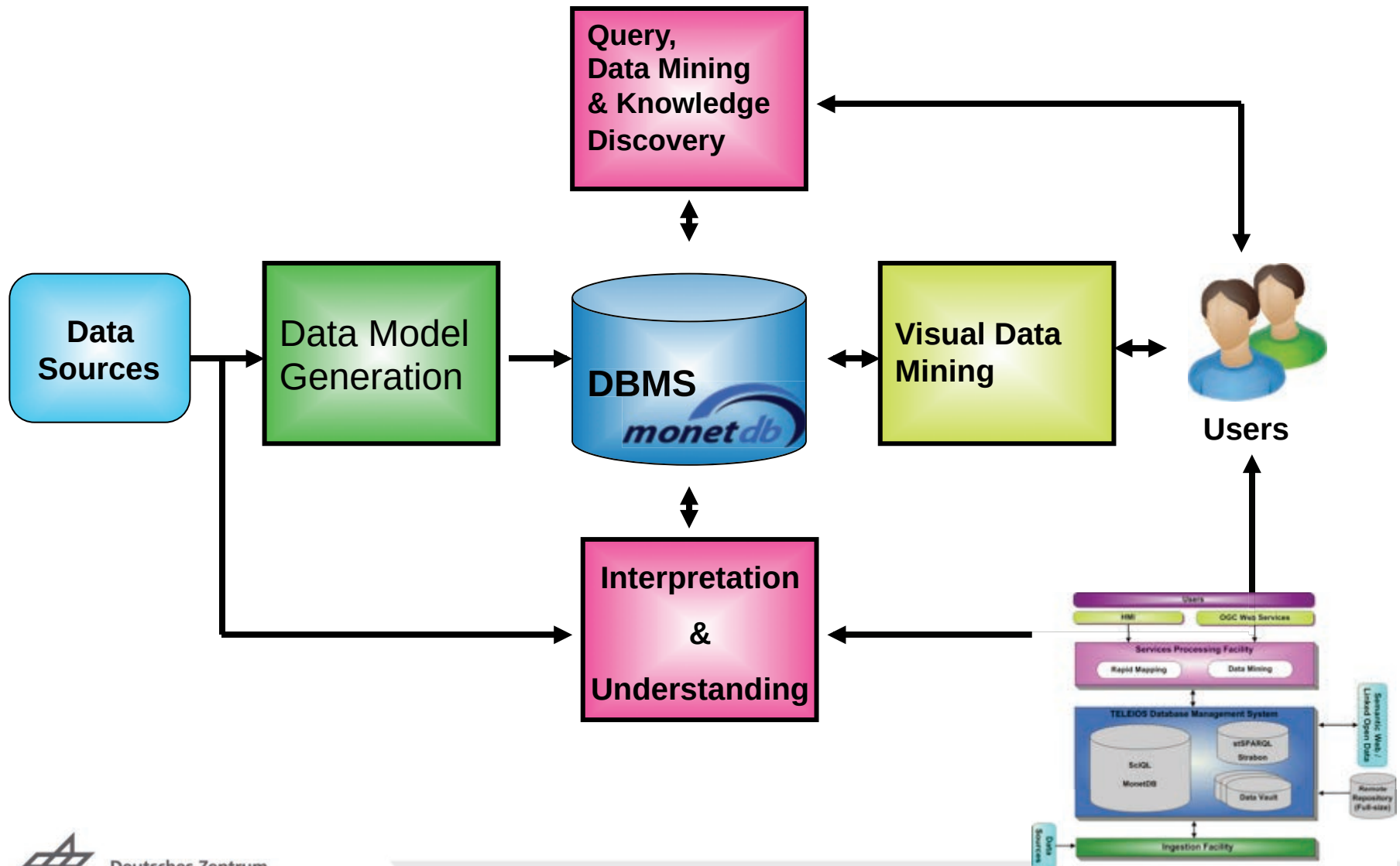
Urban Atlas



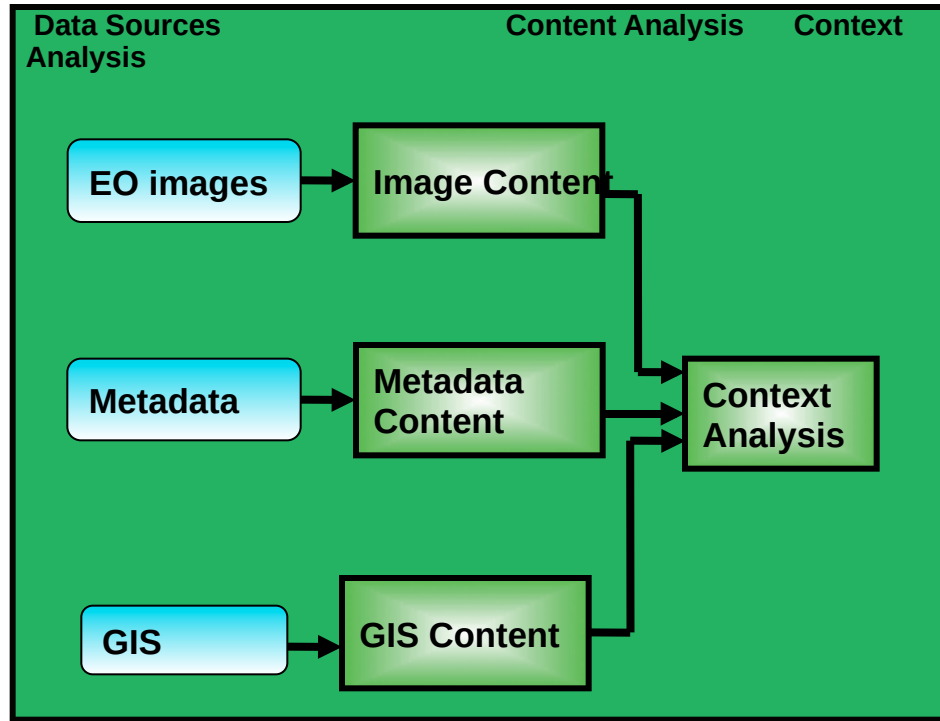
- Continuous urban fabric
- Discontinuous dense urban fabric
- Discontinuous medium density urban fabric
- Discontinuous low density urban fabric
- Discontinuous very low density urban fabric
- Isolated structures
- Industrial, commercial, public, military, private units
- Fast transit roads and associated land
- Other roads and associated land
- Railways and associated land
- Port areas
- Airports
- Mineral extraction and dump sites
- Construction sites
- Land without current use
- Green urban areas
- Sports and leisure facilities
- Agricultural, semi-natural, wetland areas
- Forest
- Water



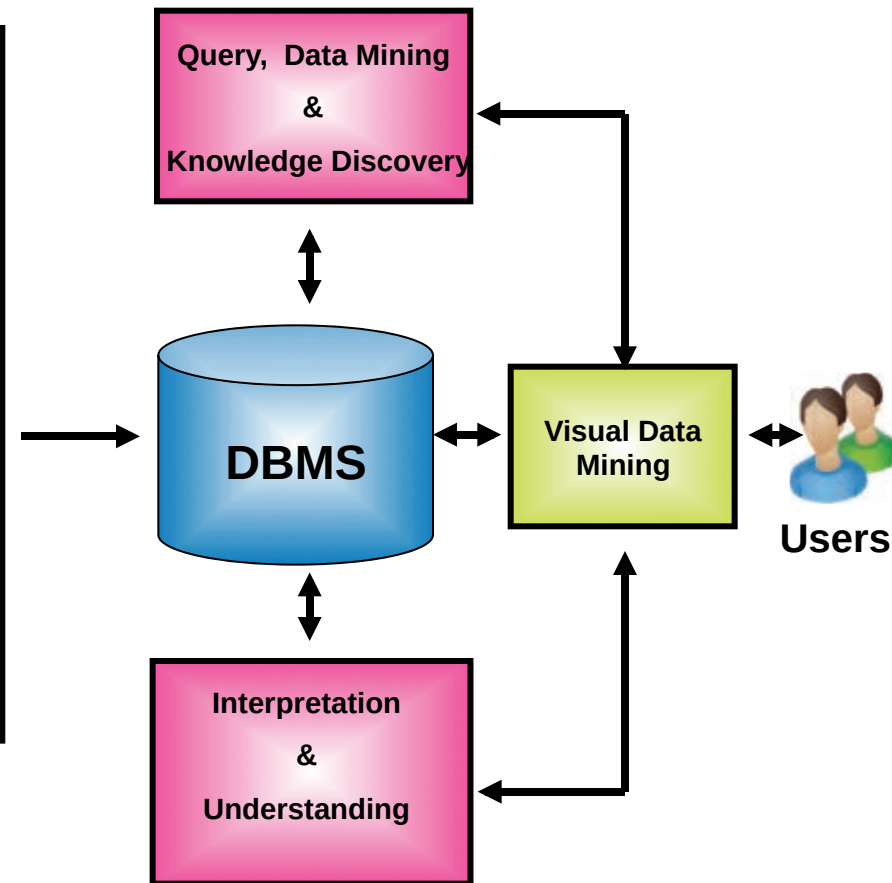
Concept and Components



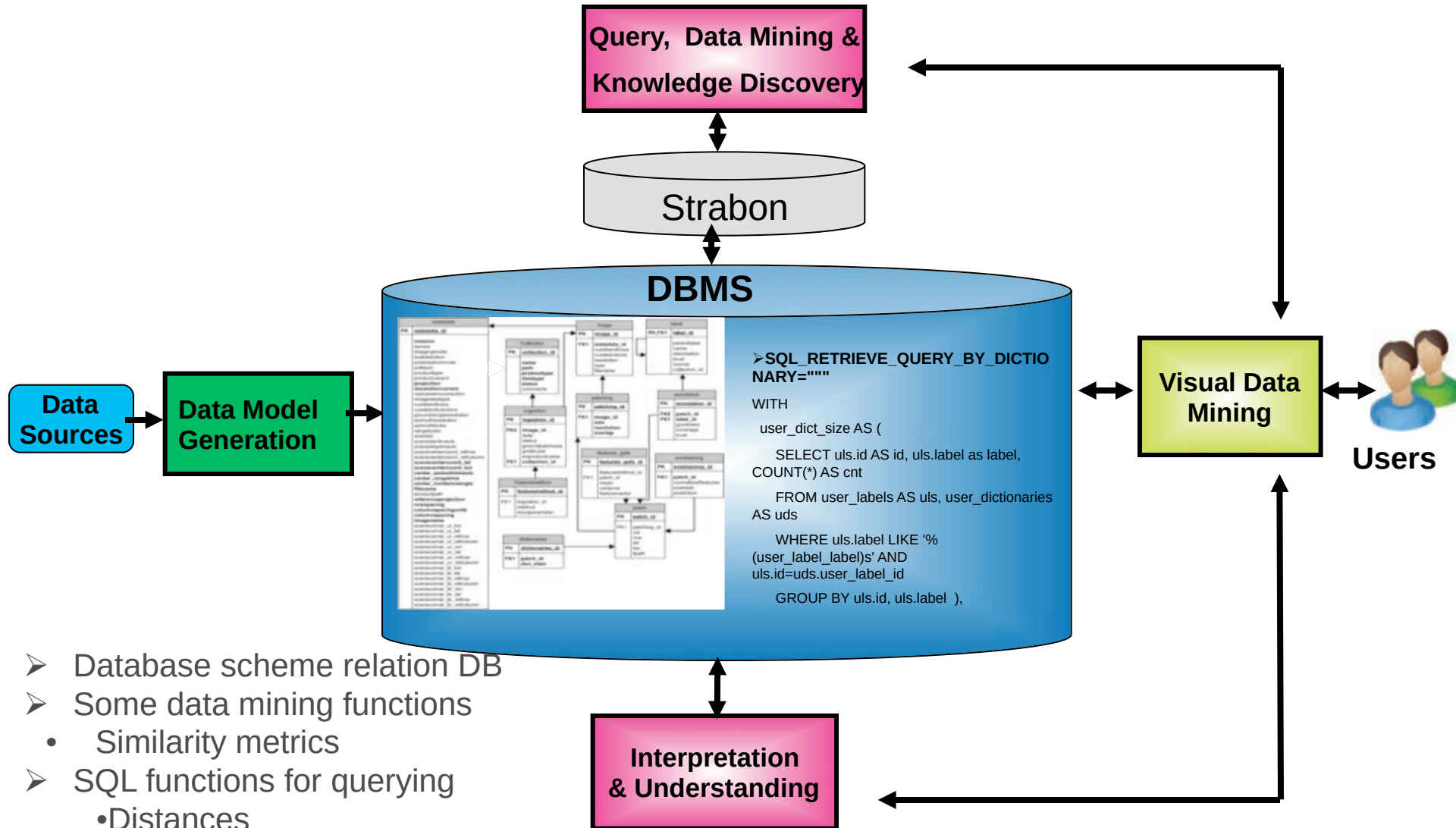
Data Model Generation



- Metadata extraction
- Patch generation
- Feature extraction methods: Gabor filters, Grey-Level Co-occurrence
- Matrix (GLCM), Bag of words, dictionary-based compression features, etc.



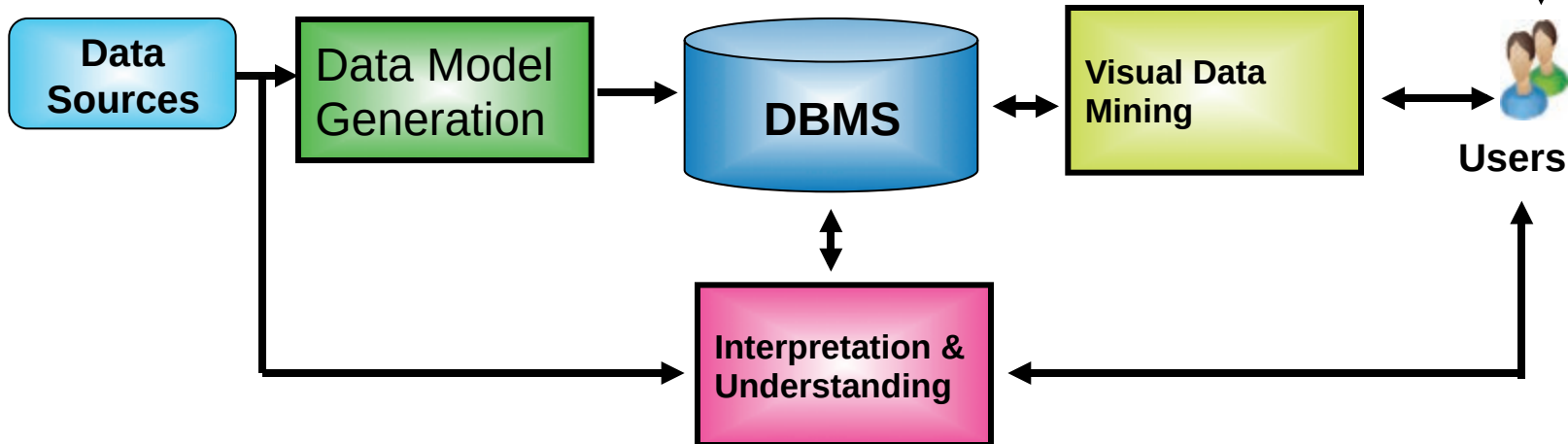
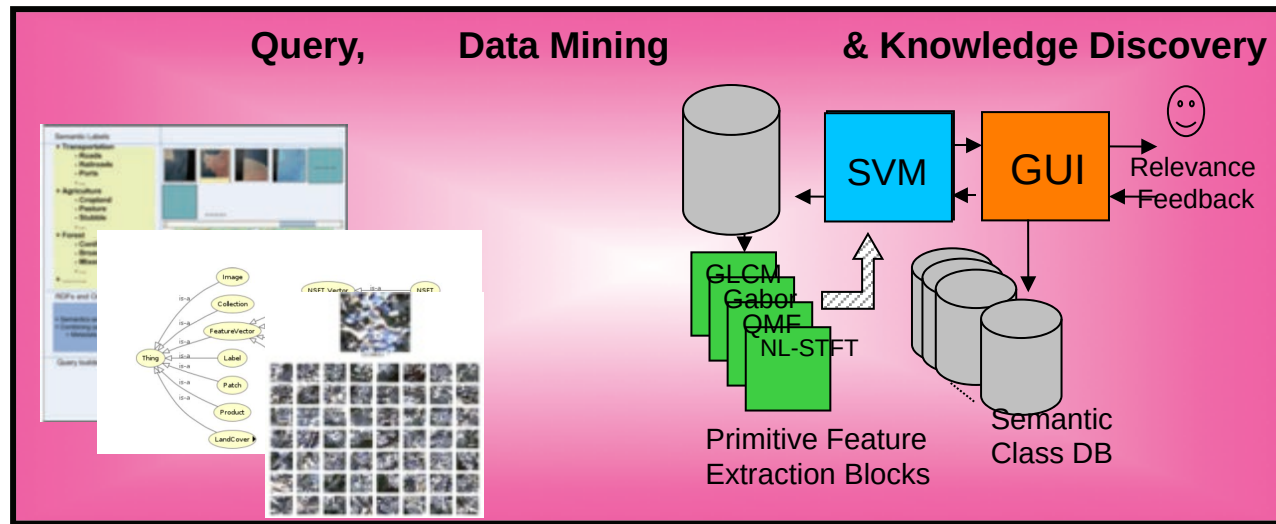
Database Management System



- Database scheme relation DB
- Some data mining functions
 - Similarity metrics
- SQL functions for querying
 - Distances



Query, Data Mining and Knowledge Discovery



Query by Metadata and Ontologies

TerraSAR-X metadata used for RDFs and Queries:

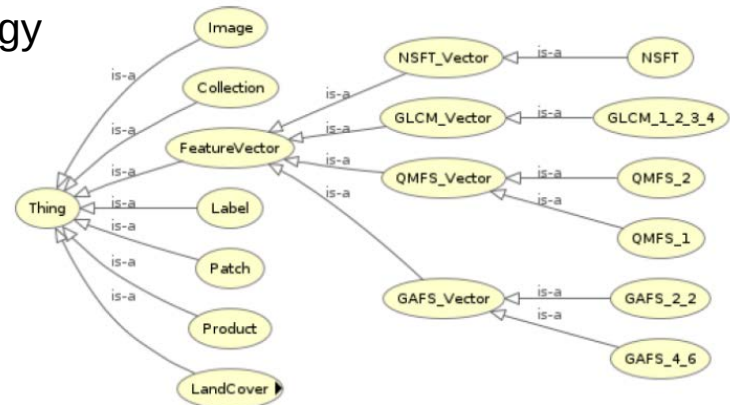
- XML file contains information about productComponent, annotation, imageData, missionInfo, acquisitionInfo, sceneInfo, etc

```
<path>./path>
<filename>TSX1_SAR_MGD_RE_HS_S_SRA_20071124T165907_20071124T165908</filename>
</location>

<path>IMAGEDATA</path>
<filename>IMAGE_HH_SRA_spot_047.tif</filename>
</location>

<imageRaster>
  <numberOfRows>10881</numberOfRows>
  <numberOfColumns>15782</numberOfColumns>
  <rowSpacing units="m">5.0000000000000000E-01</rowSpacing>
  <columnSpacing units="m">5.0000000000000000E-01</columnSpacing>
  <groundRangeResolution>1.00455682210983221E+00</groundRangeResolution>
  <azimuthResolution>1.15718095710302693E+00</azimuthResolution>
  <sceneCornerCoord>
    <refRow>10881</refRow>
    <refColumn>15365</refColumn>
  </sceneCornerCoord>
  <sceneCenterCoord>
    <refRow>5441</refRow>
    <refColumn>7891</refColumn>
    <lat>4.54351575435622905E+01</lat>
    <lon>1.23451472264888142E+01</lon>
  </sceneCenterCoord>
</imageRaster>
```

Ontology



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dlr: <http://www.earthobservatory.eu/ontologies/dlrOntology.owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

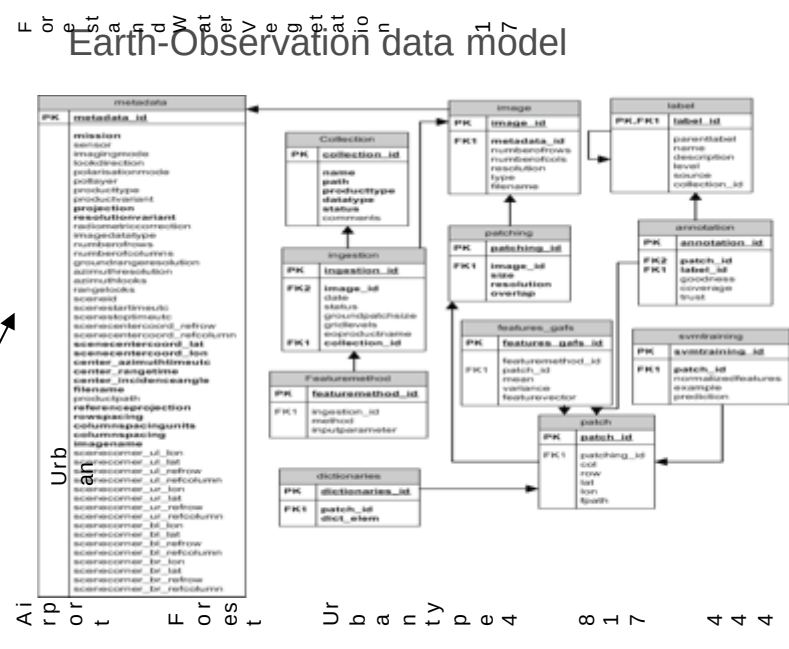
```
SELECT ?xname ?yname ?xml ?p
WHERE {
  ?x rdf:type dlr:Product .
  ?x dlr:hasName ?xname .
  ?y rdf:type dlr:Image .
  ?y dlr:hasName ?yname .
  ?x dlr:hasImage ?y .
  ?x dlr:hasXMLfilename ?xml .
  ?y dlr:consistsOf ?p .
  ?p rdf:type dlr:Patch .
  ?p dlr:hasLabel ?l .
  ?l rdf:type dlr:Label .
  ?l dlr:correspondsTo dlr:Parking .
}
```

queries

Extraction of Metadata from XML annotation file



Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft



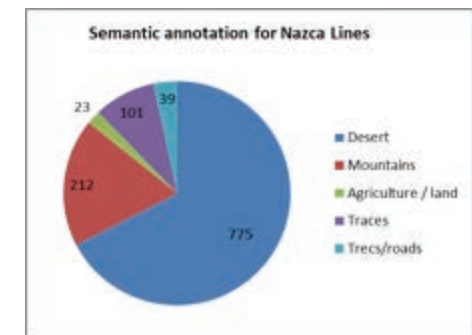
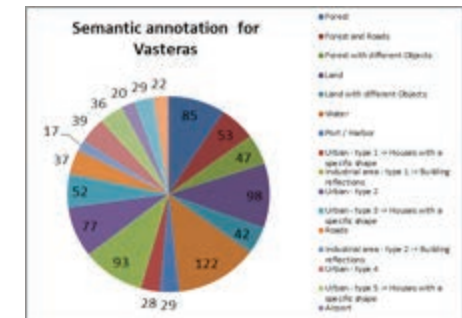
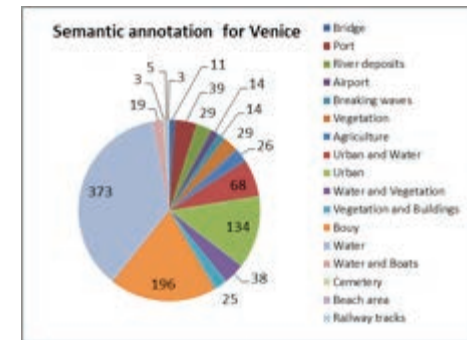
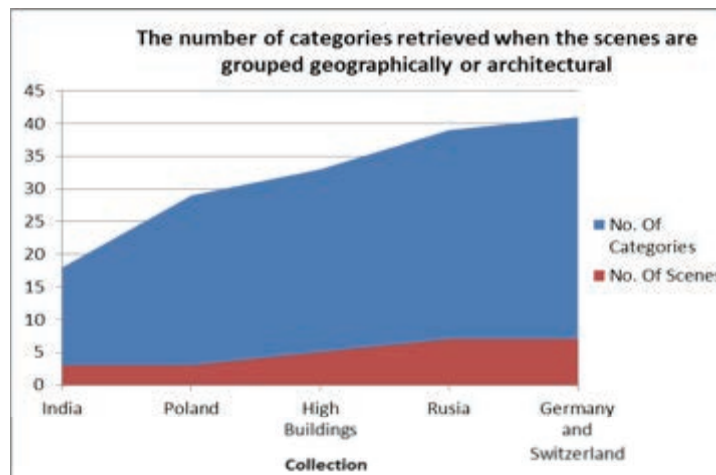
```
SELECT label_id, name,
FROM annotation a
Join label l on
a.label_id=l.label_id
```

 **Deutsches Zentrum
für Luft- und Raumfahrt e.V.**
in der Helmholtz-Gemeinschaft

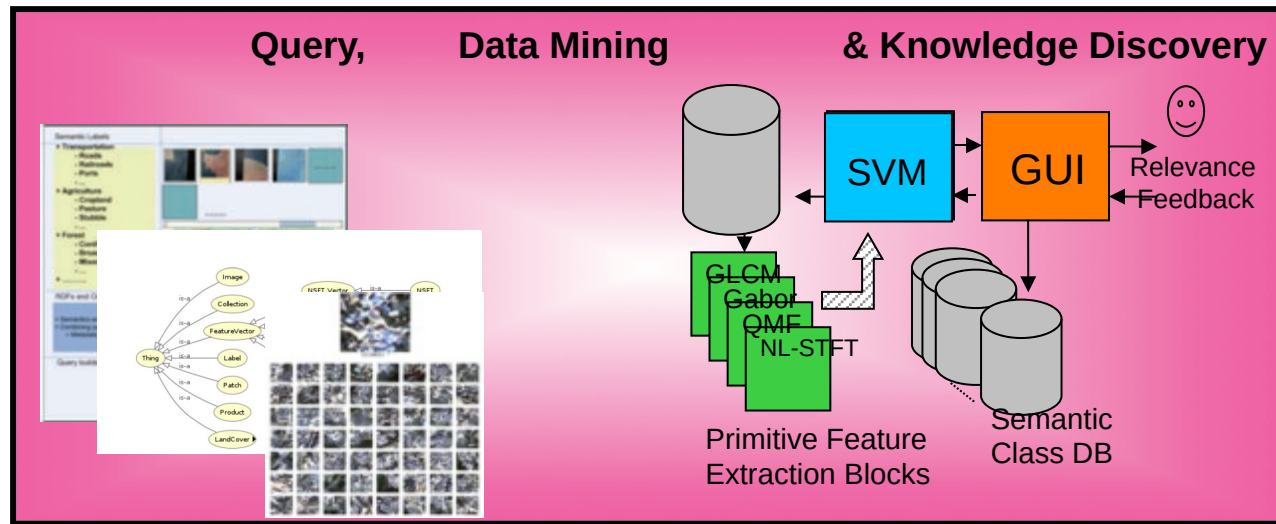
Semantic catalogues



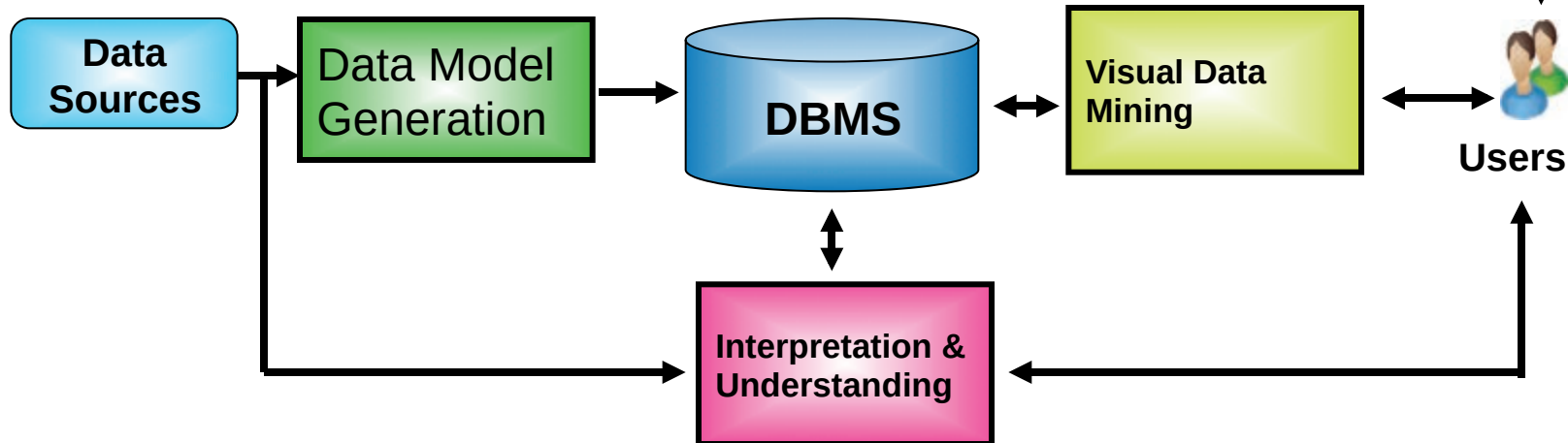
- Bangkok (Thailand);
- Shenyang (China);
- Nazca Lines (Peru);
- Havana (Cuba);
- Venice (Italy);
- Vasteras (Sweden);
- Oran (Algeria);
- Bogota (Columbia).



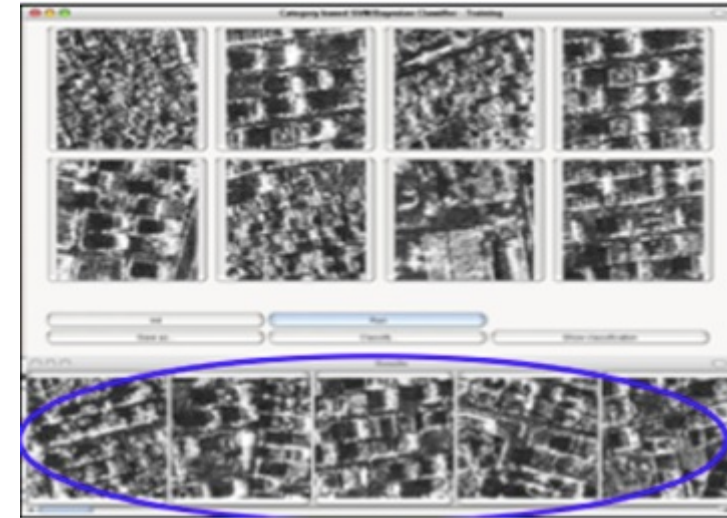
Query, Data Mining and Knowledge Discovery



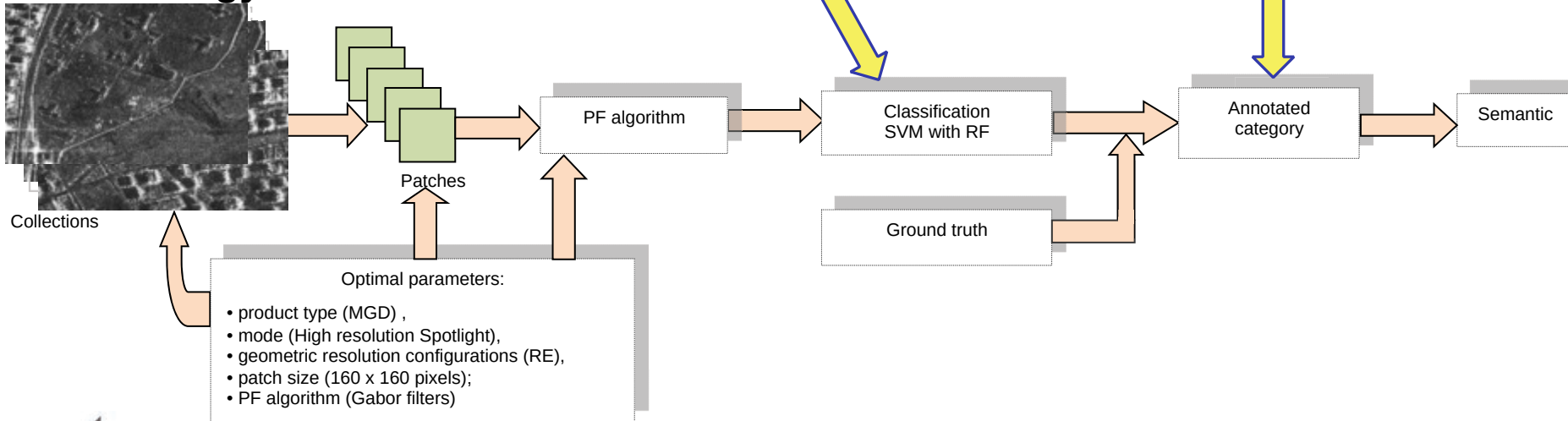
- Queries:
 - Query Builder
 - Ontologies: Strabon
 - SQL
 - Query by Example
- Data Mining
- Semantic Definition



Semantic annotation



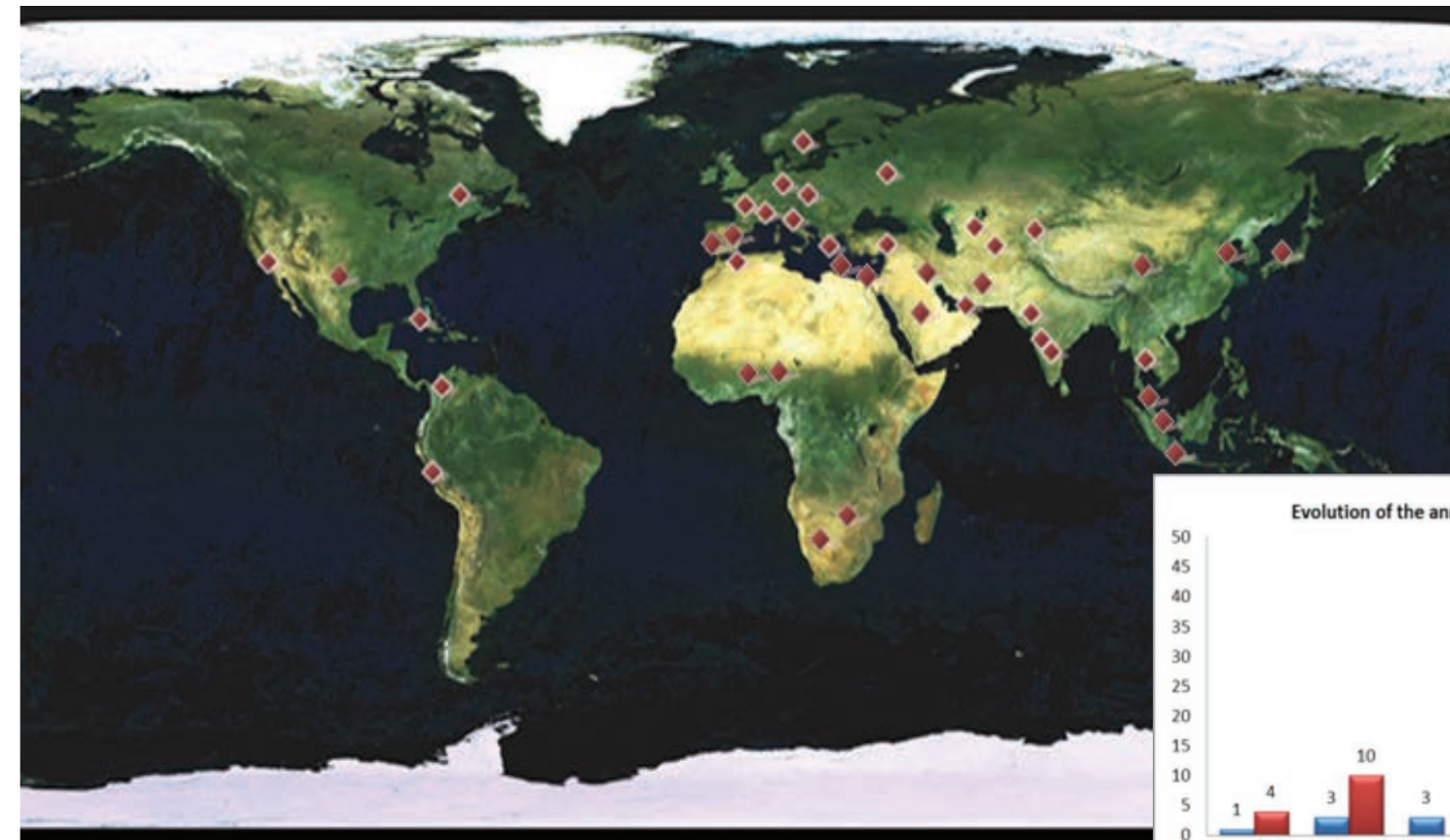
Methodology:



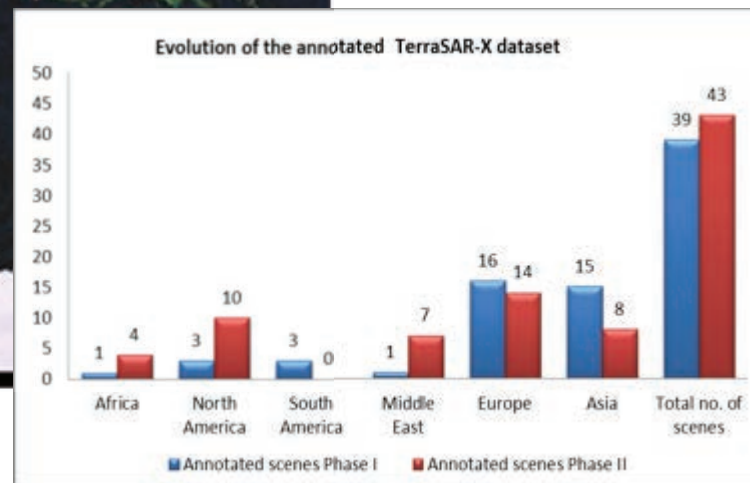


Data Model Generation: Ingested images

Location of the 100 TerraSAR-X scenes and the distribution of the scenes over the World

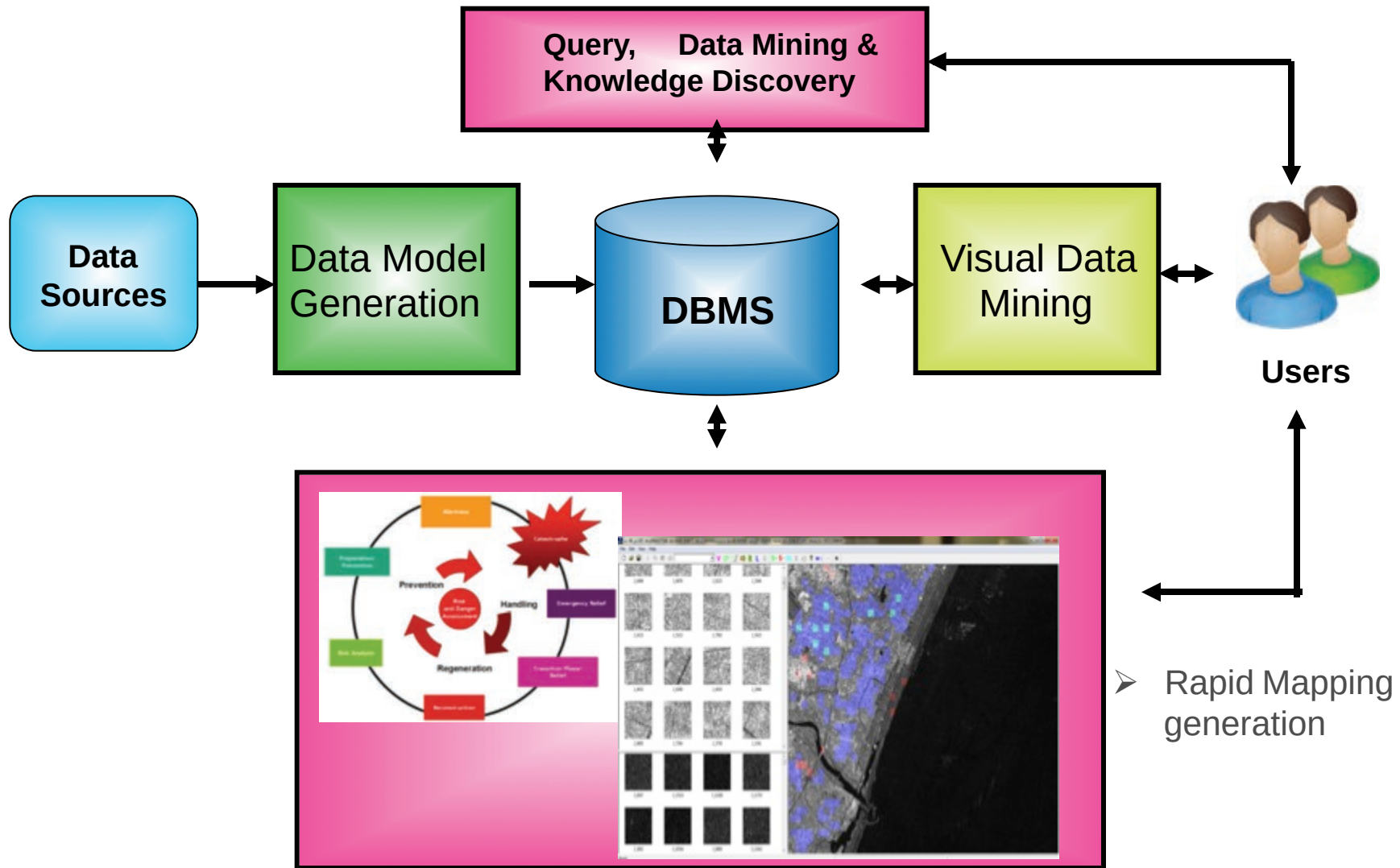


Data Model composed of:
~110000 patches
~320 semantic categories



C.O. Dumitru and M. Datcu, "Information Content of Very High Resolution SAR Images: Study of Feature Extraction and Imaging Parameters", IEEE Trans. Geoscience and Remote Sensing. To be published

Interpretation and Understanding

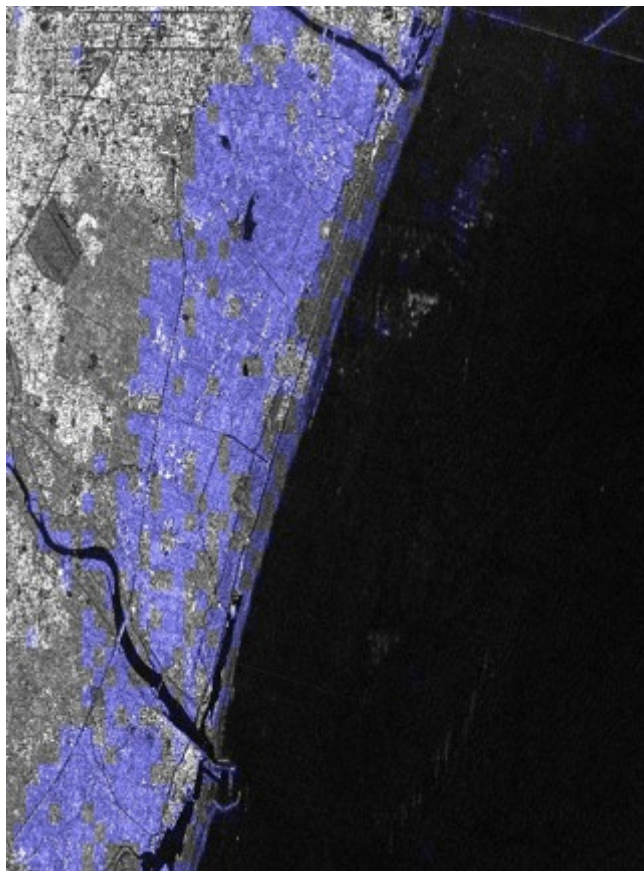


Example of Image understanding

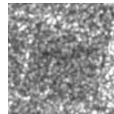
The damages in the agriculture can be clearly seen by comparing the classification in pre disaster image (left figure) with the post disaster image (right figure).

TerraSAR-X scene **before** Tsunami – 20.10.2010
Tsunami – 12.03.2011

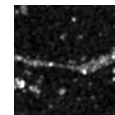
TerraSAR-X scene **after**



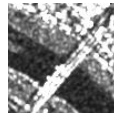
Agriculture



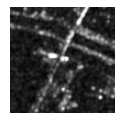
Flooded areas



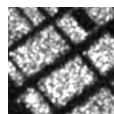
Bridges



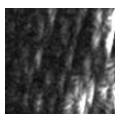
Bridges



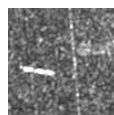
Aquaculture



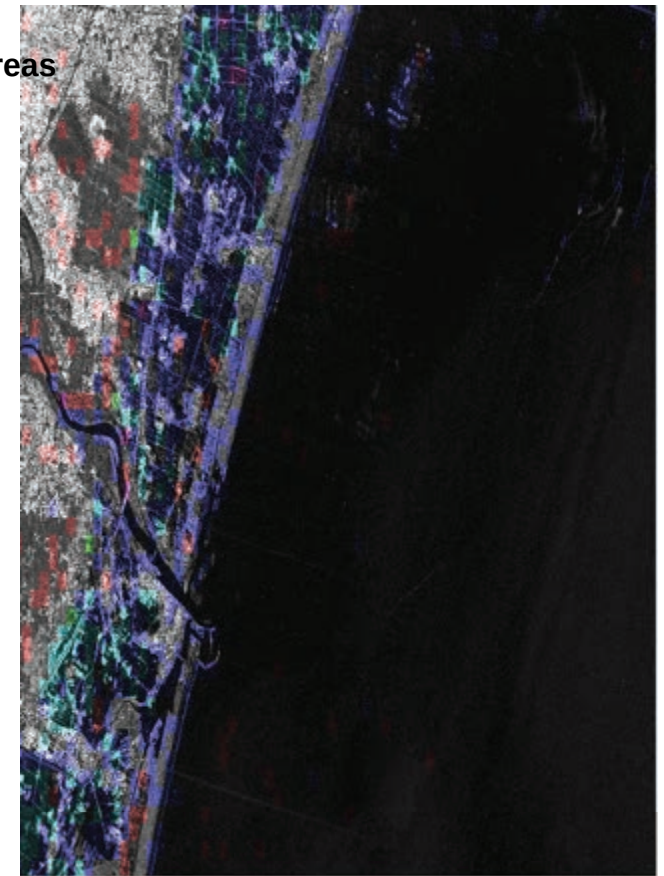
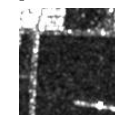
Debris



H. Voltage poles

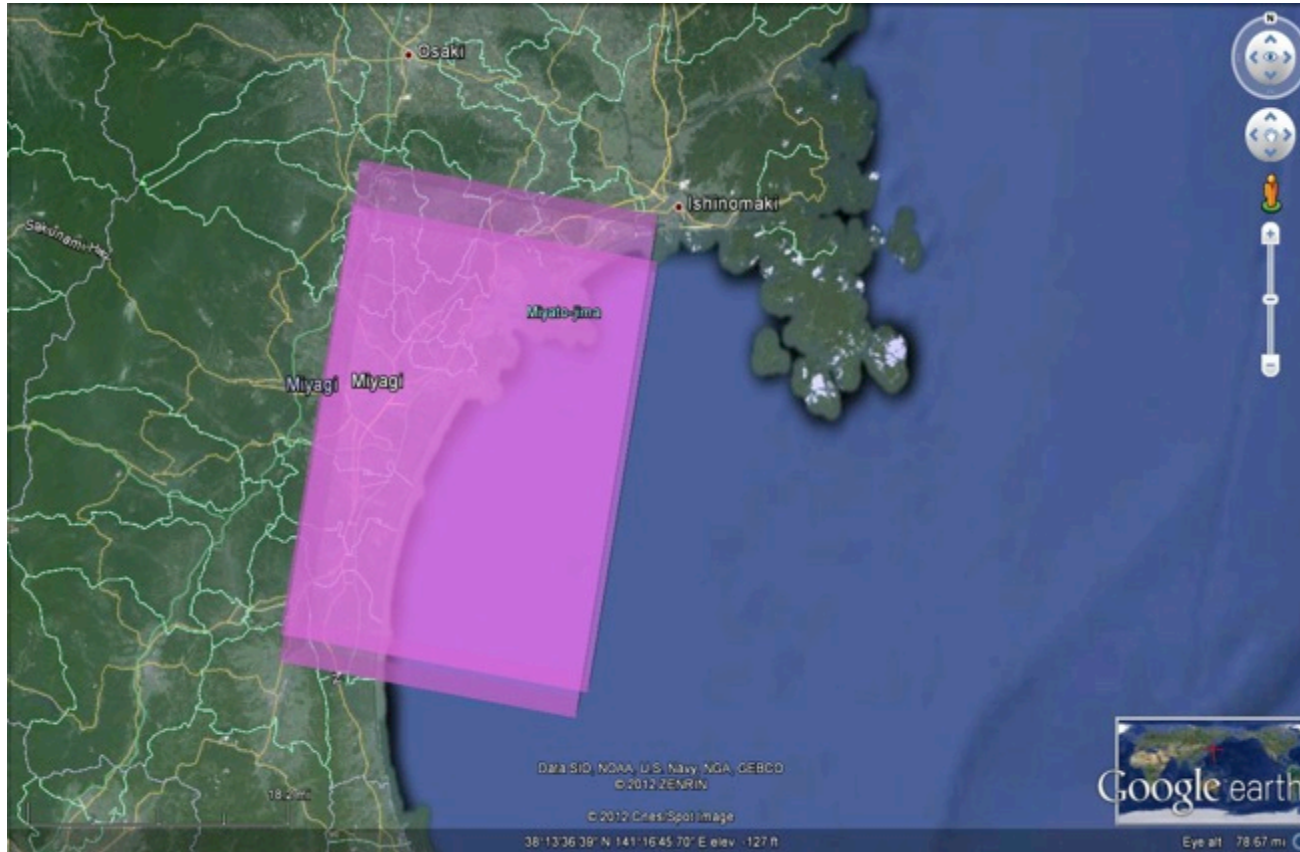


H. Voltage poles

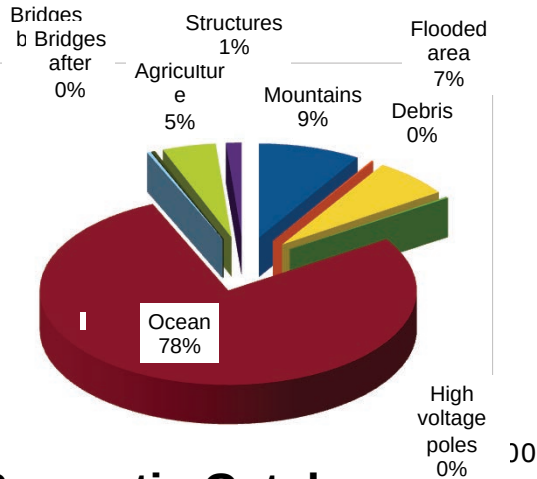


Automated Rapid Mapping scenario: Tsunami in Japan 2011

Two multitemporal TerraSAR-X images used for Japan Tsunami scenario



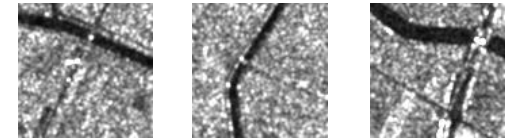
Damage Assessment



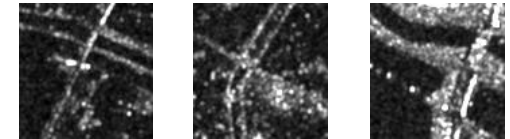
Semantic Catalogue

Semantic class	Total of annotated patches
Mountains	1874
Debris	33
Flooded area	1428
High voltage poles	50
Ocean	16562
Bridges before	66
Bridges after	66
Agriculture	981
Structures	287

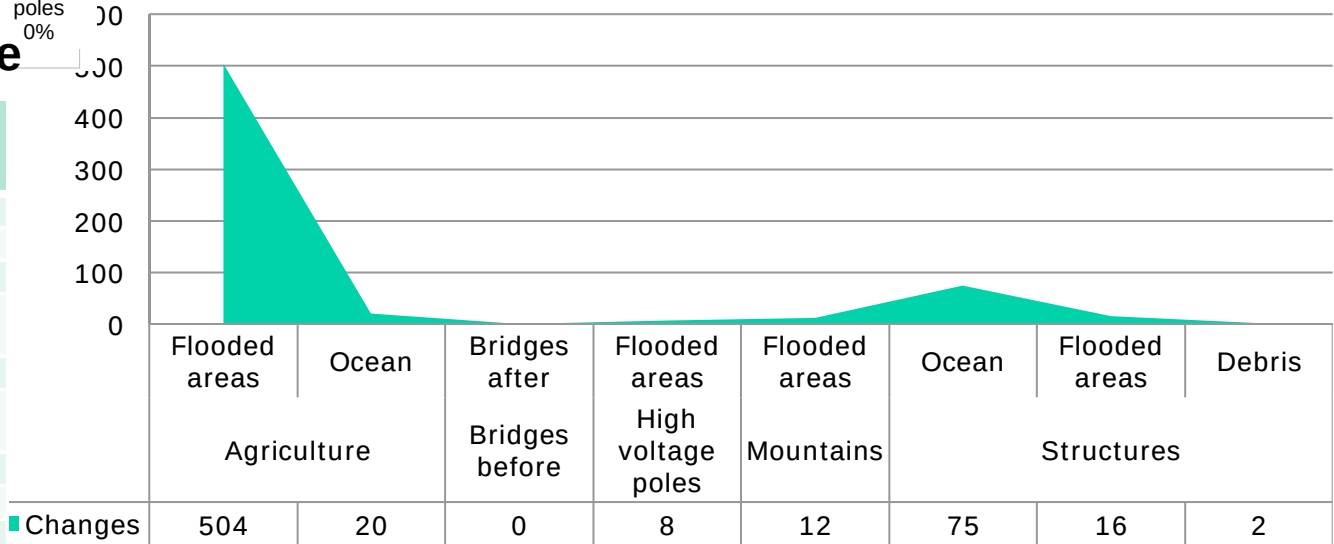
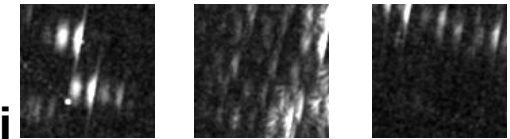
Bridges before Tsunami



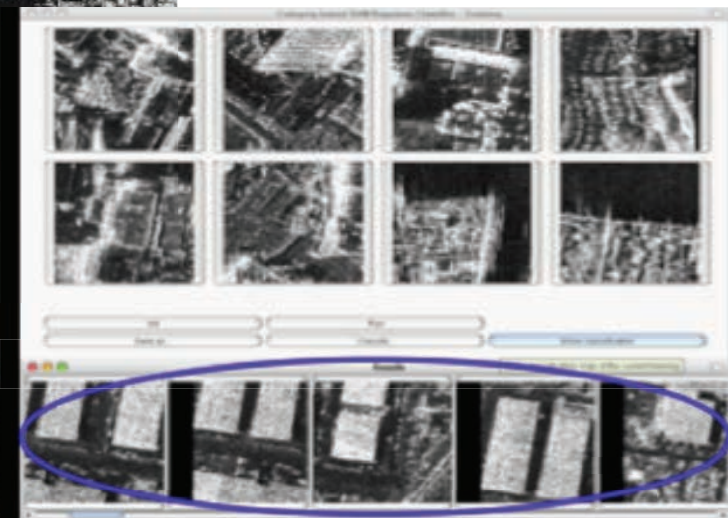
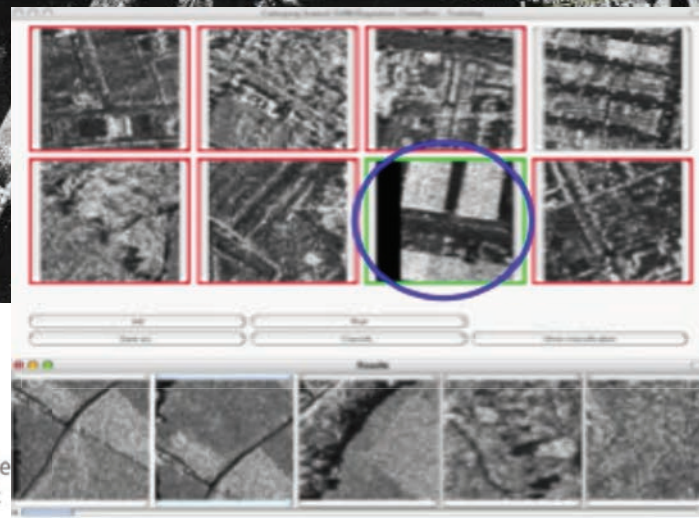
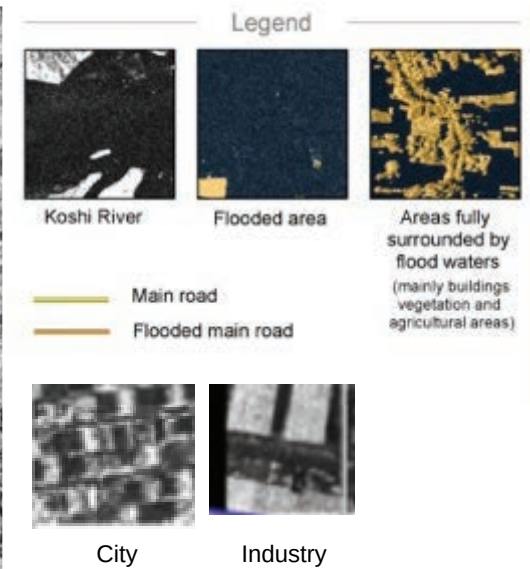
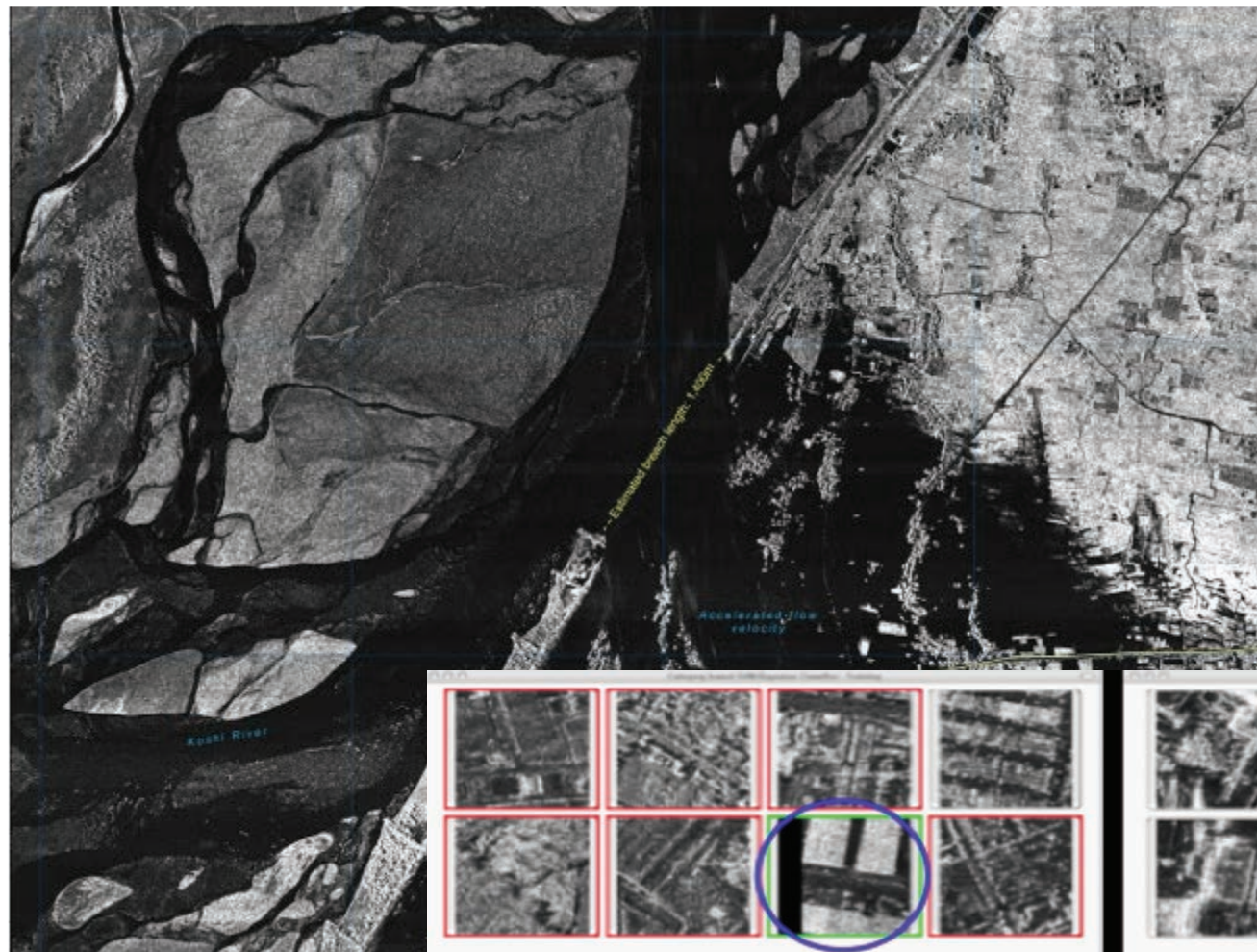
Bridges after Tsunami



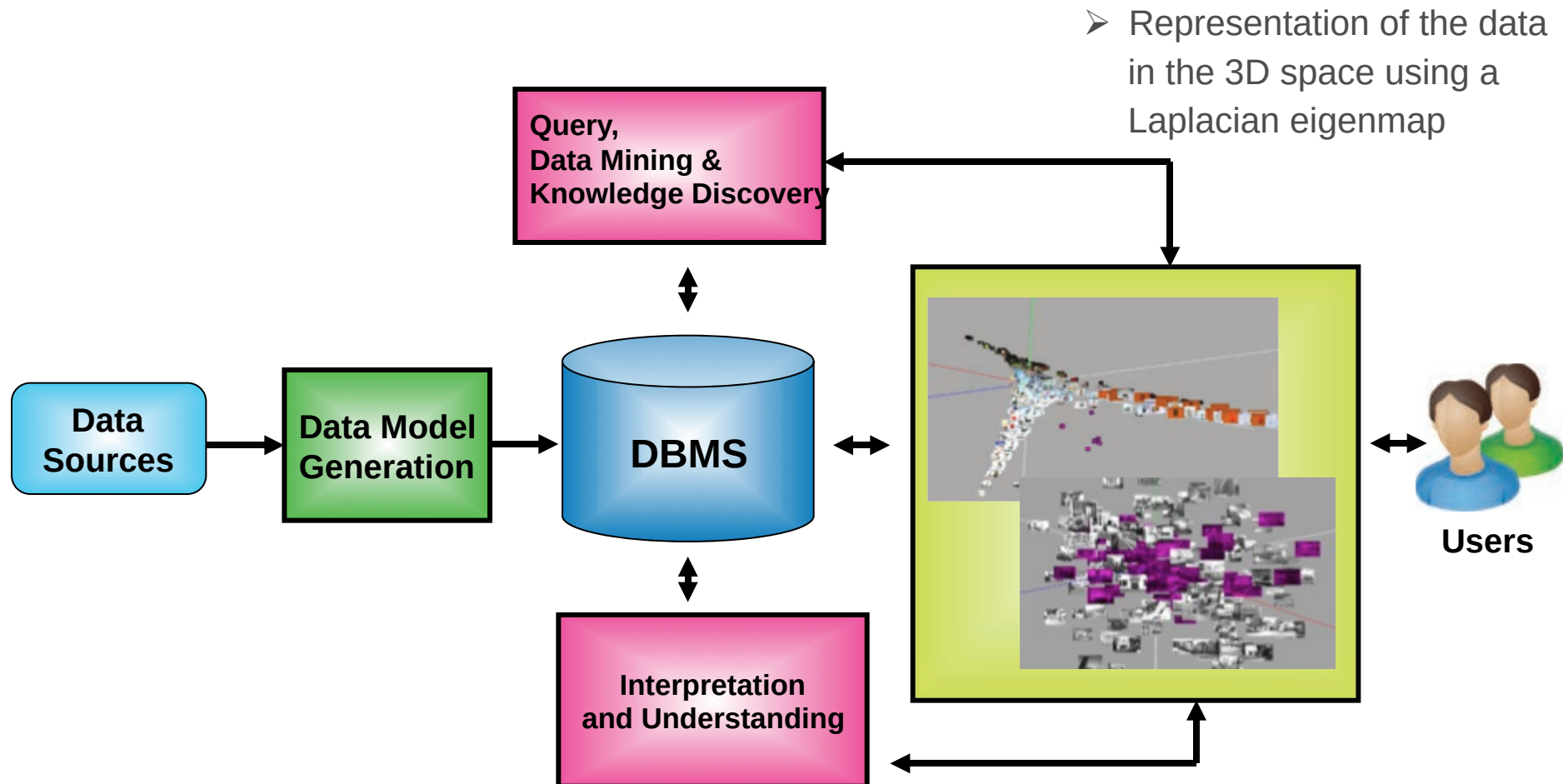
Debris caused by Tsunami



Interactive Learning for Rapid Mapping

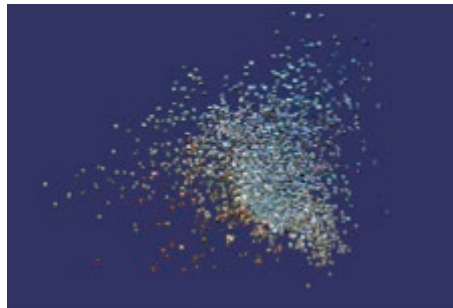
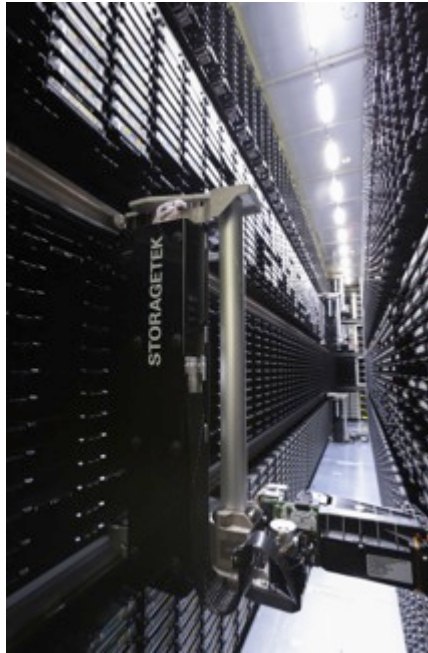


Visual Data Mining



Immersive Visual Information Mining for EO image archives

Navigation inside the EO image collections using the CAVE automatic virtual environment





EO data and Users

Volume and multimodality of data is growing

Data and information is spatio-temporal and unstructured

Users want to have the knowledge

Interactive is the only way of operation

Exploration is the predominant mode of interaction

Context is critical and relevant

Users are interested in information and knowledge independent of conjecture





Big Data trends

Information must be obtained from the data

Databases and search engines were not designed to provide contents

Visualization is very important

HMI are crucial

Cognitive vs. instruments

Reactive systems (may be more then concious ...)

All data must be availabe (no data, no discovery)

Simple, but not naïve

Selective, not only addaptive



The header features a horizontal strip with a historical map on the left, showing labels like 'THE NORTH' and 'London'. On the right, the mathematical expressions $P(y|x, M)P(x|M)$ are displayed in a stylized, overlapping font.

Processing and learning

Pre-processing integrated with post processing

Data cleaning integrated with semantic understanding

Non-visual integrated in visual

Latent integrated in intuitive learning

Learning controlled by unlearning

Information integrated with human centric knowledge

Only realtime



The header features a horizontal strip with a historical map on the left, showing labels like 'THE NORTH' and 'Londen'. On the right, the mathematical expressions $P(y|x, M)P(x|M)$ are displayed in a stylized, overlapping font.

Algorithms and tools

Pattern Recognition and Data Mining Using Data Compression

EO Ontologies

Geospatial Information Retrieval and Indexing Systems
Content Mining, Semantics Modeling, and Complex Queries

Mining Image Time Series

DM supported by Web Services, CLOUD technologies

