DATA SCIENCE WORKFLOWS FOR THE CANDELA PROJECT

Mihai Datcu\textsuperscript{1}, Corneliu Octavian Dumitru\textsuperscript{1}, Gottfried Schwarz\textsuperscript{1}, Fabien Castel\textsuperscript{2}, and Jose Lorenzo\textsuperscript{3}

\textsuperscript{1}German Aerospace Center DLR
\textsuperscript{2}ATOS France SA
\textsuperscript{3}ATOS Spain SA

BiDS’19
Munich, 19-21 Febr.2019

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 776193
Machine Learning: CV vs. EO

- **EO**
  - Physical parameters
  - Multi-temporal

- **CV&EO**
  - Labelling

- **EO**
  - Trust me

---

[Image of a scale and a cat with a butterfly]
Preliminaries

- DNN: in 2018 more than 500 papers/month
- Research is often wasted effort
- ML faces a deep reproducibility crisis
- Training data is as important as the learning algorithm
- ML finds any pattern in data, it may be irrelevant
- We need the actual patterns of the Earth processes
- Big EO Data accentuate the crisis

- Solution: In CANDELA we propose a Data Science workflow to insure the quality of the information extraction
CANDELA main objective

CANDELA project main objective is to allow the creation of value from Copernicus data through the provisioning of modelling and analytics tools given that the tasks of data collection, processing, storage and access will be provided by the Copernicus Data and Information Access Service (DIAS).

The goal of the Data Science is to enable the successful integration of heterogeneous datasets, to support the definition and design of the data transformation to information, the use of taxonomies and elements of ontology and semantics, learning, KDD, annotation, data analytics.
Sensory and Semantic Gaps

- Sensory perceptions are **not** 1:1 reproductions of the real world:
  - There are individual representations
- Humans and computers interpret and name objects differently

Fig. 1. Example patches corresponding to the category “urban/residential areas” for the datasets (a) D1, (b) D2, (c) D3, (d) D4, (e) D5, (f) D6, (g) D7, and (h) D8, and corresponding to the category “agricultural fields” for the datasets (i) D1, (j) D2, (k) D3, (l) D4, (m) D5, (n) D7, and (o) D8.
Cross databases semantics

a. Semantic content intersection between datasets.

b. Percentage of exact label matches within the intersected semantic content.


www.candela-h2020.eu
## Training EO 3 bands data sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. images</th>
<th>C</th>
<th>Patch size (pixels)</th>
<th>Type and resolution</th>
<th>Size (zipped)</th>
<th>Applications / target</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCMerced</td>
<td>2100</td>
<td>21</td>
<td>256 x 256</td>
<td>aerial, 30cm</td>
<td>317 Mb</td>
<td>Land use</td>
<td>2010</td>
</tr>
<tr>
<td>WHU-RS19</td>
<td>950</td>
<td>19</td>
<td>600 x 600</td>
<td>Aerial/VHR from, 0.5m</td>
<td></td>
<td></td>
<td>2012</td>
</tr>
<tr>
<td>WHU-RS19</td>
<td>5000</td>
<td>20</td>
<td>600 x 600</td>
<td>screenshots, 26cm - 7.44m</td>
<td></td>
<td>Scene classification in VHR</td>
<td>2015</td>
</tr>
<tr>
<td>RSSCN7</td>
<td>2800</td>
<td>7</td>
<td>400 x 400</td>
<td>GE, 4 scales</td>
<td>348 Mb</td>
<td>Land cover, multiscale</td>
<td>Nov 2015</td>
</tr>
<tr>
<td>AID</td>
<td>1000</td>
<td>30</td>
<td>600 x 600</td>
<td>aerial, 0.5m - 8m</td>
<td></td>
<td>Land cover, multi-resolution</td>
<td>2016</td>
</tr>
<tr>
<td>RSI-CB</td>
<td>24000</td>
<td>35</td>
<td>128 x 128</td>
<td>GE, Bing Maps 0.3–3-m</td>
<td></td>
<td>6 categories, 35 or 45 subclasses</td>
<td>2017</td>
</tr>
<tr>
<td></td>
<td>36000</td>
<td>45</td>
<td>256 x 256</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PatternNet</td>
<td>30400</td>
<td>38</td>
<td>256 x 256</td>
<td>GE, 0.062m – 4.693m</td>
<td></td>
<td>Image retrieval</td>
<td>2017</td>
</tr>
<tr>
<td>DOTA 1.0</td>
<td>188282</td>
<td>15</td>
<td>4000 x 4000</td>
<td>GE mainly ; JL-1 and GF-2</td>
<td>12.5 Gb train val + 6 Gb testing</td>
<td>15 caisses, urban</td>
<td>2018</td>
</tr>
<tr>
<td>SAROptical</td>
<td>10000 pairs</td>
<td>112</td>
<td>112 x 112</td>
<td>TerraSAR-X (1m) spotlight, UltraCAM aerial (20 cm)</td>
<td></td>
<td>SAR and optical joint analysis for dense urban areas</td>
<td>2018</td>
</tr>
<tr>
<td>SEN 1-2 v1</td>
<td>282384 pairs</td>
<td>256</td>
<td>256 x 256</td>
<td>S1 (SAR, VV backscatter, colorized) and S2 (only RGB bands, TOA)</td>
<td>43.7 Gb</td>
<td>SAR to optical image matching</td>
<td>2018</td>
</tr>
</tbody>
</table>
# Training EO multispectral data sets

<table>
<thead>
<tr>
<th>No. images</th>
<th>C</th>
<th>Patch size (pixels)</th>
<th>Type and resolution</th>
<th>Applications / target</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Coffee Scene</td>
<td>2 876</td>
<td>2</td>
<td>64 x 64</td>
<td>SPOT, NIR Red Green false colour JPG</td>
<td>Binary classification (coffee trees or not)</td>
</tr>
<tr>
<td>SAT-4</td>
<td>500 000</td>
<td>4</td>
<td>28 x 28</td>
<td>RGB + NIR, aerial, 1m</td>
<td>Vegetation (e.g. grassland, trees)</td>
</tr>
<tr>
<td>SAT-6</td>
<td>405 000</td>
<td>6</td>
<td>28 x 28</td>
<td>RGB + NIR, aerial, 1m</td>
<td>Land use</td>
</tr>
<tr>
<td>EuroSAT</td>
<td>20 000</td>
<td>10</td>
<td>64 x 64</td>
<td>Sentinel-2, 13 bands or RGB only</td>
<td>Land use and land cover classification</td>
</tr>
</tbody>
</table>
EO SAR training data sets

• MSTAR - an X-band SAR data set used for automatic target recognition (ATR) of military objects
  • In total 17,096 target patches ranging in size from 54×54 pixels to 192×192 pixels with resolution of 1 foot.
    • September 95 Collection contains 20 target types with additional articulation, obscuration, and camouflage views
    • November 96 Collection adds another 27 target types with additional articulation and obscuration cases.

• OpenSARShip – an C-band data set (Sentinel-1) used for ship interpretation
  • In total there are 11,346 ship chips
EO data annotation
CANDELA focus

• In CANDELA, a special attention is given to **re-use and openness**.
  building modules and frameworks on-top of available components
  maximization of benefits from existing assets
  making the solutions available to various user communities

• DLR’s EOLib is an Image Information Mining system for Earth Observation
  processes, extracts, and accesses the content of EO products
  generates higher-level abstractions and semantics
  offers information mining services on the original corpus of EO products
  provides KDD based on the EO content, metadata, semantic annotations,

• EOLib is integrated with the TerraSAR-X Payload Ground Segment (PGS)
The CANDELA analytics modules

EO data

Change detection on time series

Data mining

Data fusion

Change detection indicator

Semantic labels

Semantic search and indexation

Data taxonomy

Front Office Thematic applications

Non EO data
Data Mining and Fusion in CANDELA

Legend:
- CANDELA Component / System
- PGS Component / System
- CANDELA Data Item
- PGS Data Item
- Access to all Data
- Interface
- Data Path
Data Science Workflows

- **Data mining exploration**
  - **Capabilities**
    - Load and ingest EO images together with their metadata
    - Extract and tile the images into patches
  - **CANDELA components**
    - Data Model Generation (DMG) component
  - **Step 1**
  - **Step 2**
    - Automatically ingest all given information into the database
    - DataBase Management System (DBMS) component
  - **Step 3**
    - Visual exploration of the content of the database by giving positive and negative examples
    - Image search component
  - **Output**
    - Visual inspection of the EO image content and types of classes that can be extracted
    - Multi-knowledge and querying component

- **Data mining semantic annotation**
  - **Capabilities**
    - Load and ingest EO images together with their metadata
    - Extract and tile the images into patches
  - **CANDELA components**
    - Data Model Generation (DMG) component
  - **Step 1**
  - **Step 2**
    - Automatically ingest all given information into the database
    - DataBase Management System (DBMS) component
  - **Step 3**
    - Search for the same content for the purpose of grouping and annotation
    - Image search component
  - **Step 4**
    - The content / classified category is semantically annotated and saved into the database
    - Semantic annotation component
  - **Output**
    - Semantic catalogue (dynamic) which is updated during each classification / annotation process
Coarse-to-fine strategy and cascaded learning

- Use of a pyramid of finer image grid levels

- Objective: a finer spatial indexing, and semantic extraction

- Costs: increase of the number of patches to process

- Advantage: at level 100, 70% of the patches are removed, preserving a recall of 90%
Fast learning

Acceleration with two orders of magnitude

Learning with:

- Few
- Controllable
- Trusted
- Trusted

samples


www.candela-h2020.eu
Implementation: Data Model Generation

TerraSAR-X L1b product

TerraSAR-X metadata and image

Tiles with different size

Primitive features: Gabor filters and Weber Local Descriptors

Metadata Extraction ➔ Image Tiling ➔ Quick-looks generation ➔ Primitive Feature extraction ➔ Create the product model
Implementation: Data Mining Data Base

DMDB is a relational database

Main tables are:
- Metadata
- Image
- Tiles
- Features
- Labels

DMDB comprises about
- 8 millions of tiles
- 20 thousand metadata entries.
- 106 semantic labels
Implementation: Data Mining

Metadata parameters are based on XML annotation file of TerraSAR-X L1b products.

- Coordinates (lat/lon)
- Incidence angles
- Acquisition time
- Pixel spacing
- Number of columns/rows
- sensor
- Mission
- orbits

Semantic parameters are based on EO Taxonomy.

- Agriculture
  - Cropland
  - Rice plantation.....
- Bare ground
  - Cliff
  - Desert.....
- Urban area
  - Commercial areas
  - High density residential areas....
- Forest
  - Forest coniferous
  - Forest mixed....
KDD is used to define **semantic annotations of the image content**.

- Goal is to build a model which performs the mapping between low-level image descriptors (primitive features) and high-level image concepts (semantics).
- KDD is based on machine learning methods and relevance feedback mechanisms.
Semantic query

- Storage tanks
- Medium density residential area
- Mixed urban areas

Query Expression:
- Parameter: name
- Operator: =
- Value: Storage...
- Conector: OR
- Value: Medium...
- Conector: AND
Data Fusion: SAR vs. MS EO

TerraSAR-X vs. WordView

The clouds

Complementary features
### Data Fusion: Validation Data Sets

<table>
<thead>
<tr>
<th>SAR instrument</th>
<th>TerraSAR-X</th>
<th>Sentinel-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image location</strong></td>
<td>Bucharest (Romania)</td>
<td>Munich (Germany)</td>
</tr>
<tr>
<td></td>
<td>Washington (USA)</td>
<td>Venice (Italy)</td>
</tr>
<tr>
<td><strong>Acquisition time</strong></td>
<td>Aug. 15, 2009 (Bucharest)</td>
<td>April 24, 2013 (Munich)</td>
</tr>
<tr>
<td></td>
<td>June 22, 2010 (Washington)</td>
<td>Sept. 05, 2012 (Venice)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multispectral instrument</th>
<th>WorldView-2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image location</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bucharest (Romania)</td>
</tr>
<tr>
<td></td>
<td>Venice (Italy)</td>
</tr>
<tr>
<td><strong>Acquisition time</strong></td>
<td>Oct. 29, 2010 (Bucharest)</td>
</tr>
<tr>
<td></td>
<td>Sept. 08, 2012 (Venice)</td>
</tr>
</tbody>
</table>
## Data Fusion: Selected Results

<table>
<thead>
<tr>
<th>No.</th>
<th>Semantic annotation</th>
<th>No. of patches</th>
<th>Multispectral</th>
<th></th>
<th>SAR</th>
<th></th>
<th>Fused images</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>1</td>
<td>Administrative and Monument areas</td>
<td>646</td>
<td>50.29</td>
<td>36.47</td>
<td>44.49</td>
<td>42.30</td>
<td>94.78</td>
<td>73.21</td>
</tr>
<tr>
<td>2</td>
<td>Bridges</td>
<td>24</td>
<td>42.42</td>
<td>58.33</td>
<td>33.45</td>
<td>37.50</td>
<td>80.95</td>
<td>70.83</td>
</tr>
<tr>
<td>3</td>
<td>Broadleaf forest</td>
<td>1061</td>
<td>82.96</td>
<td>41.67</td>
<td>56.57</td>
<td>52.87</td>
<td>95.39</td>
<td>76.06</td>
</tr>
<tr>
<td>4</td>
<td>Cemeteries</td>
<td>72</td>
<td>44.45</td>
<td>36.67</td>
<td>41.10</td>
<td>36.57</td>
<td>91.67</td>
<td>30.56</td>
</tr>
<tr>
<td>5</td>
<td>Grassland</td>
<td>201</td>
<td>41.94</td>
<td>71.14</td>
<td>40.29</td>
<td>77.62</td>
<td>78.00</td>
<td>84.03</td>
</tr>
<tr>
<td>6</td>
<td>High-density residential areas</td>
<td>617</td>
<td>46.45</td>
<td>58.99</td>
<td>43.64</td>
<td>39.66</td>
<td>96.98</td>
<td>57.37</td>
</tr>
<tr>
<td>7</td>
<td>Medium-density residential areas</td>
<td>3120</td>
<td>73.97</td>
<td>57.12</td>
<td>51.51</td>
<td>42.05</td>
<td>94.75</td>
<td>89.58</td>
</tr>
<tr>
<td>8</td>
<td>Mixed urban areas</td>
<td>374</td>
<td>56.00</td>
<td>39.21</td>
<td>53.24</td>
<td>38.72</td>
<td>80.21</td>
<td>40.11</td>
</tr>
<tr>
<td>9</td>
<td>Parking areas</td>
<td>143</td>
<td>60.61</td>
<td>43.97</td>
<td>50.00</td>
<td>37.00</td>
<td>52.76</td>
<td>46.85</td>
</tr>
<tr>
<td>10</td>
<td>Rivers</td>
<td>120</td>
<td>69.37</td>
<td>64.17</td>
<td>59.08</td>
<td>47.50</td>
<td>80.00</td>
<td>80.33</td>
</tr>
<tr>
<td>11</td>
<td>Roads</td>
<td>949</td>
<td>56.37</td>
<td>45.39</td>
<td>47.84</td>
<td>42.33</td>
<td>98.60</td>
<td>22.34</td>
</tr>
<tr>
<td>12</td>
<td>Sports grounds</td>
<td>21</td>
<td>100.00</td>
<td>80.95</td>
<td>52.31</td>
<td>58.10</td>
<td>85.45</td>
<td>79.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.40</td>
<td>52.84</td>
</tr>
<tr>
<td></td>
<td>47.79</td>
<td>46.02</td>
</tr>
<tr>
<td></td>
<td>85.80</td>
<td>62.52</td>
</tr>
</tbody>
</table>
Thank you for your attention