Copernicus Access Platform Intermediate Layers Small Scale Demonstrator

DATA SCIENCE WORKFLOWS FOR THE CANDELA PROJECT

Mihai Datcu[1], Corneliu Octavian Dumitru[1], Gottfried Schwarz[1], Fabien Castel[2], and Jose Lorenzo[3]

[1]German Aerospace Center DLR
[2]ATOS France SA
[3]ATOS Spain SA

www.candela-h2020.eu

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

Trust me

EO

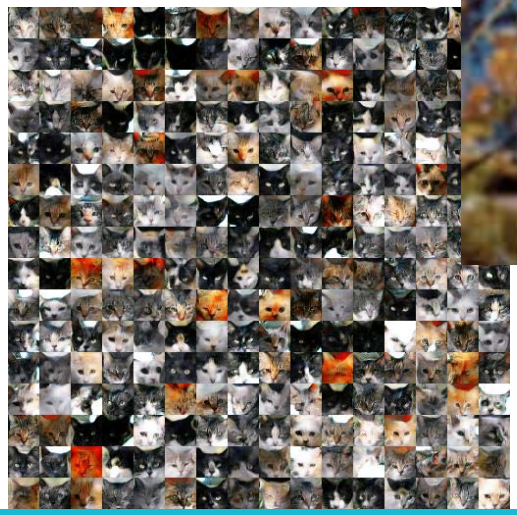
Multi-temporal

EO

Physical parameters

CV&EO

Labelling



Preliminaries

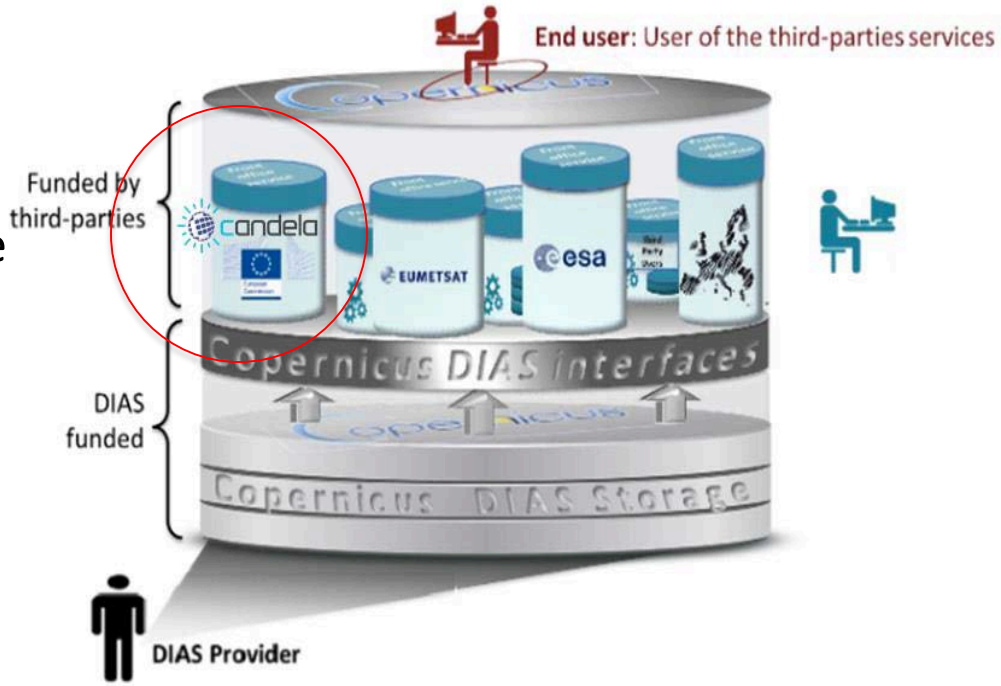
- DNN: in 2018 more than 500 papers/month
- Research is often wasted effort
- ML faces a deep reproducibility crisis
- Training data is as important as the learning algorithm
- ML finds any pattern in data, it may be irrelevant
- We need the actual patterns of the Earth processes
- Big EO Data accentuate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

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

CANDELA main objective

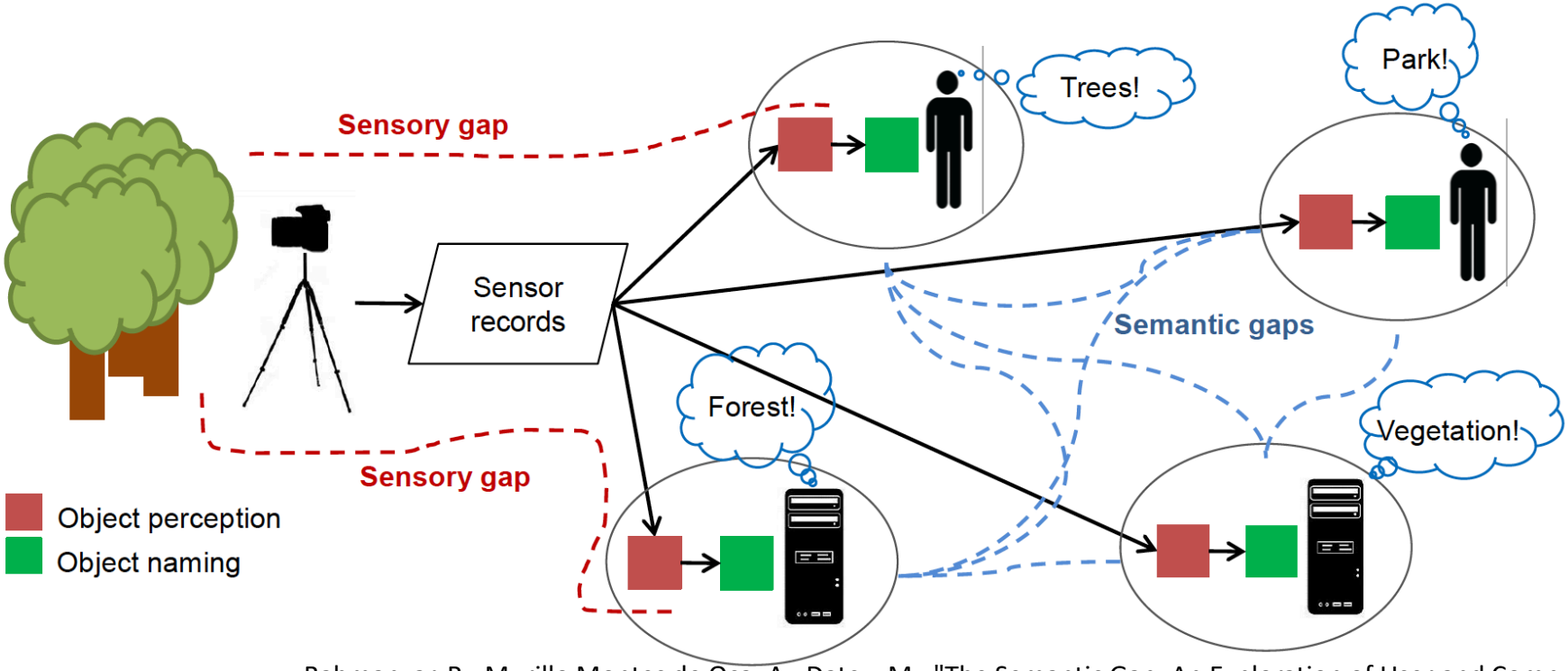
CANDELA project main objective is to allow the **creation of value** from **Copernicus data** through the provisioning of **modelling and analytics tools** given that the tasks of data collection, processing, storage and access will be provided by the **Copernicus Data and Information Access Service (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 <u>not</u> 1:1 reproductions of the real world:
  - There are individual representations
- Humans and computers interpret and name objects differently



Bahmanyar, R.; Murillo Montes de Oca, A.; Datcu, M., "The Semantic Gap: An Exploration of User and Computer Perspectives in Earth Observation Images," in Geoscience and Remote Sensing Letters, IEEE , vol.12, no.10, pp.2046-2050, Oct. 2015

Data Base Biases: Test data sets

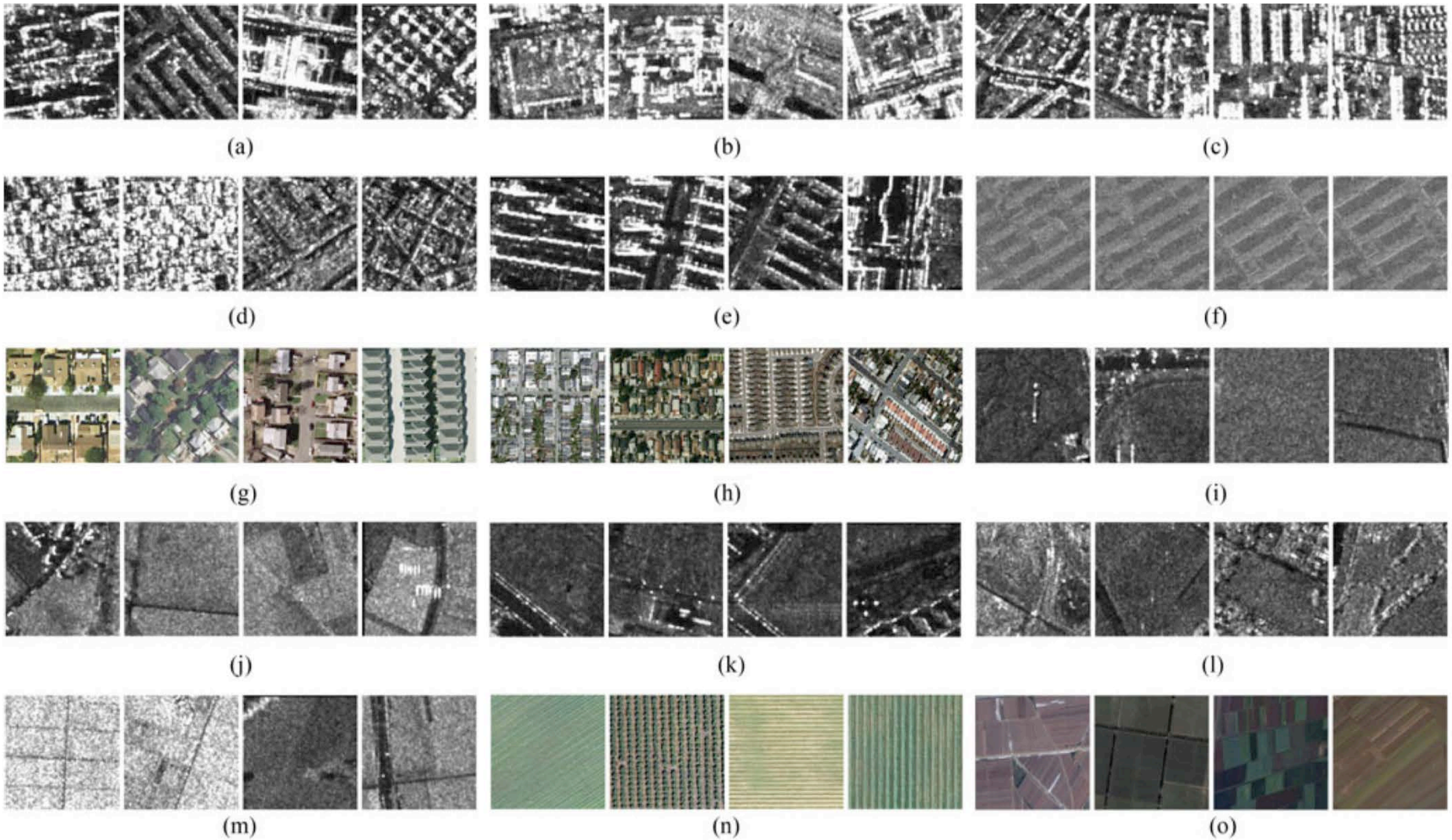


Fig. 1. Example patches corresponding to the category "urban/residential areas" for the datasets (a) D1, (b) D2, (c) D3, (d) D4, (e) D5, (f) D6, (g) D7, ar
(h) D8, and corresponding to the category "agricultural fields" for the datasets (i) D1, (j) D2, (k) D3, (l) D4, (m) D5, (n) D7, and (o) D8.

Cross databases semantics

a. Semantic content intersection between datasets.

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

	D1	D2	D3	D4	D5	D6	D7	D8
D1	100	43	30	25	42	37	36	37
D2	43	100	18	39	37	25	22	38
D3	30	18	100	56	41	61	14	23
D4	25	39	56	100	50	40	20	32
D5	42	37	41	50	100	46	42	65
D6	37	25	61	40	46	100	28	23
D7	36	22	14	20	42	28	100	35
D8	37	38	23	32	65	23	35	100

(a)

	D1	D2	D3	D4	D5	D6	D7	D8
D1	100	91	77	72	36	49	56	38
D2	91	100	100	100	95	44	64	32
D3	77	100	100	100	100	0	100	0
D4	72	100	100	100	100	63	0	44
D5	36	95	100	100	100	76	26	28
D6	49	44	0	63	76	100	21	30
D7	56	64	100	0	26	21	100	6
D8	38	32	0	44	28	30	6	100

(b)

Murillo Montes de Oca, A. ; Bahmanyar, R.; Nistor, N.; Datcu, M., Earth Observation Image Semantic Bias: A Collaborative User Annotation Approach, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 6, pp. 2462 - 2477, 2017

Training EO 3 bands data sets

	No. images	C	Patch size (pixels)	Type and resolution	Size (zipped)	Applications / target	Year
UCMerced	2 100	21	256 x 256	aerial, 30cm	317 Mb	Land use	2010
WHU-RS19	950	19	600 x 600	Aerial/VHR from, 0.5m			2012
WHU-RS19	5 000	20	600 x 600	screenshots, 26cm - 7.44m		Scene classification in VHR	2015
RSSCN7	2800	7	400 x 400	GE, 4 scales	348 Mb	Land cover, multiscale	Nov 2015
AID	10 000	30	600 x 600	aerial, 0.5m - 8m		Land cover, multi-resolution	2016
RSI-CB	24 000 36 000	35 45	128 x 128 256 x 256	GE, Bing Maps 0.3–3-m		6 categories, 35 or 45 subclasses	2017
PatternNet	30 400	38	256 x 256	GE, 0.062m – 4.693m		Image retrieval	2017
DOTA 1.0	188 282	15	4000 x 4000	GE mainly ; JL-1 and GF-2	12.5 Gb train val + 6 Gb testing	15 calsses, urban	2018
SAROptical	10 000 pairs		112 x 112	TerraSAR-X (1m) spotlight , UltraCAM aerial (20 cm)		SAR and optical joint analysis for dense urban areas	2018
SEN 1-2 v1	282 384 pairs		256 x 256	S1 (SAR, VV backscatter, colorized) and S2 (only RGB bands, TOA)	43.7 Gb	SAR to optical image matching	2018

Training EO multispectral data sets

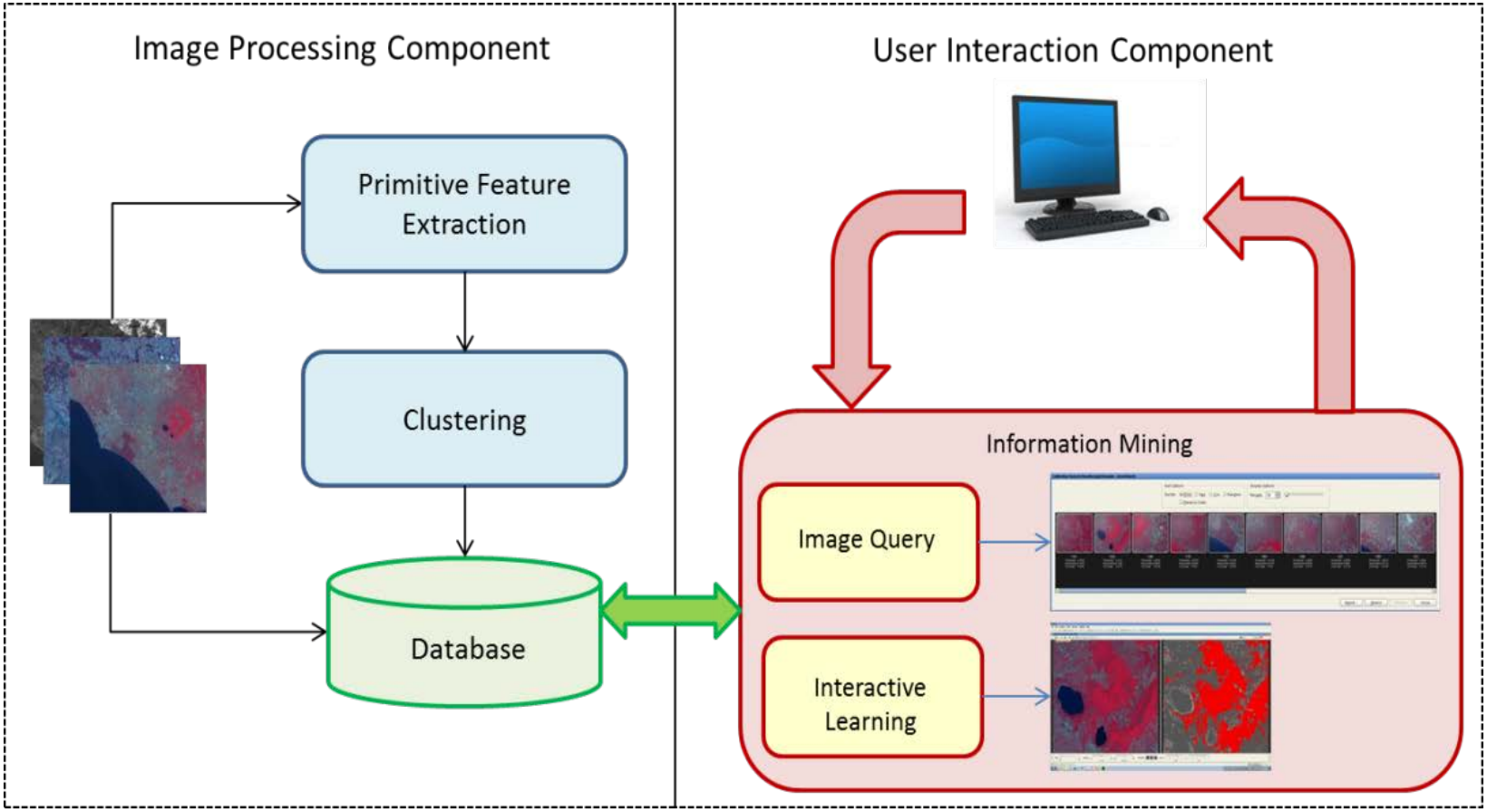
	No. images	C	Patch size (pixels)	Type and resolution	Applications / target	Year
Brazilian Coffee Scene	2 876	2	64 x 64	SPOT, **NIR** Red Green false colour JPG	Binary classification (coffee trees or not)	2015
SAT-4	500 000	4	28 x 28	RGB + **NIR**, aerial, 1m	Vegetation (e.g. grassland, trees)	2015
SAT-6	405 000	6	28 x 28	RGB + **NIR**, aerial, 1m	Land use	2015
EuroSAT	20 000	10	64 x 64	Sentinel-2, **13 bands** or RGB only	Land use and land cover classification	2017

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



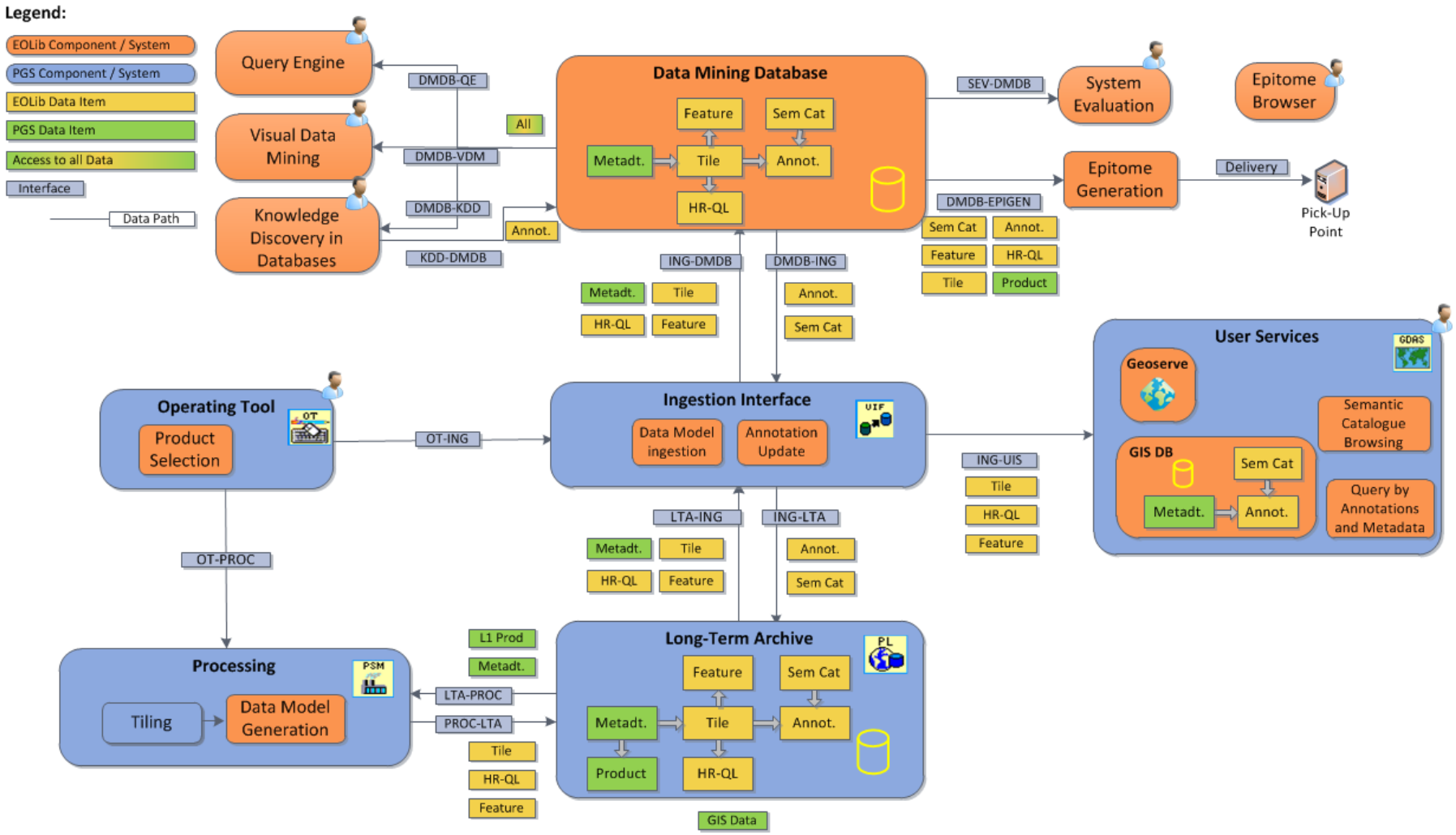
CANDELA focus

- 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

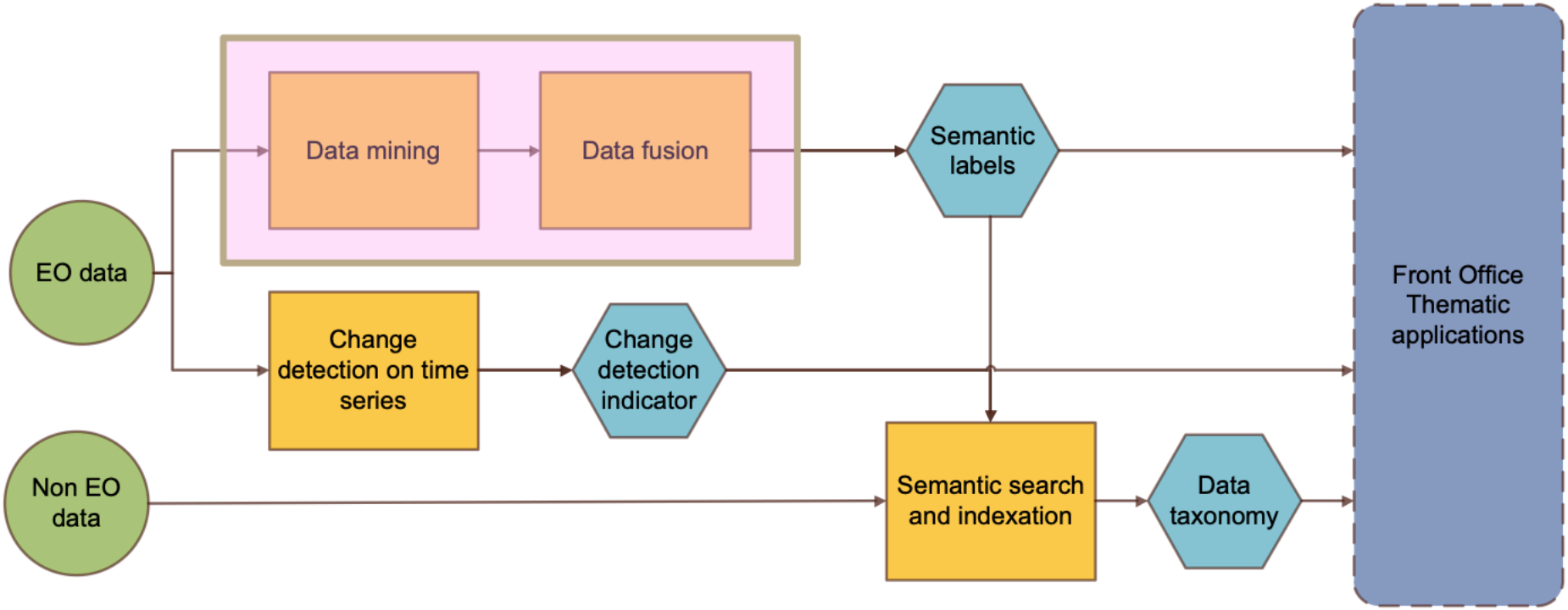
- DLR's EOLib is an Image Information Mining system for Earth Observation
  - processes, extracts, and accesses the content of EO products
  - generates higher-level abstractions and semantics
  - offers information mining services on the original corpus of EO products
  - provides KDD based on the EO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

- EOLib is integrated with the TerraSAR-X Payload Ground Segment (PGS)

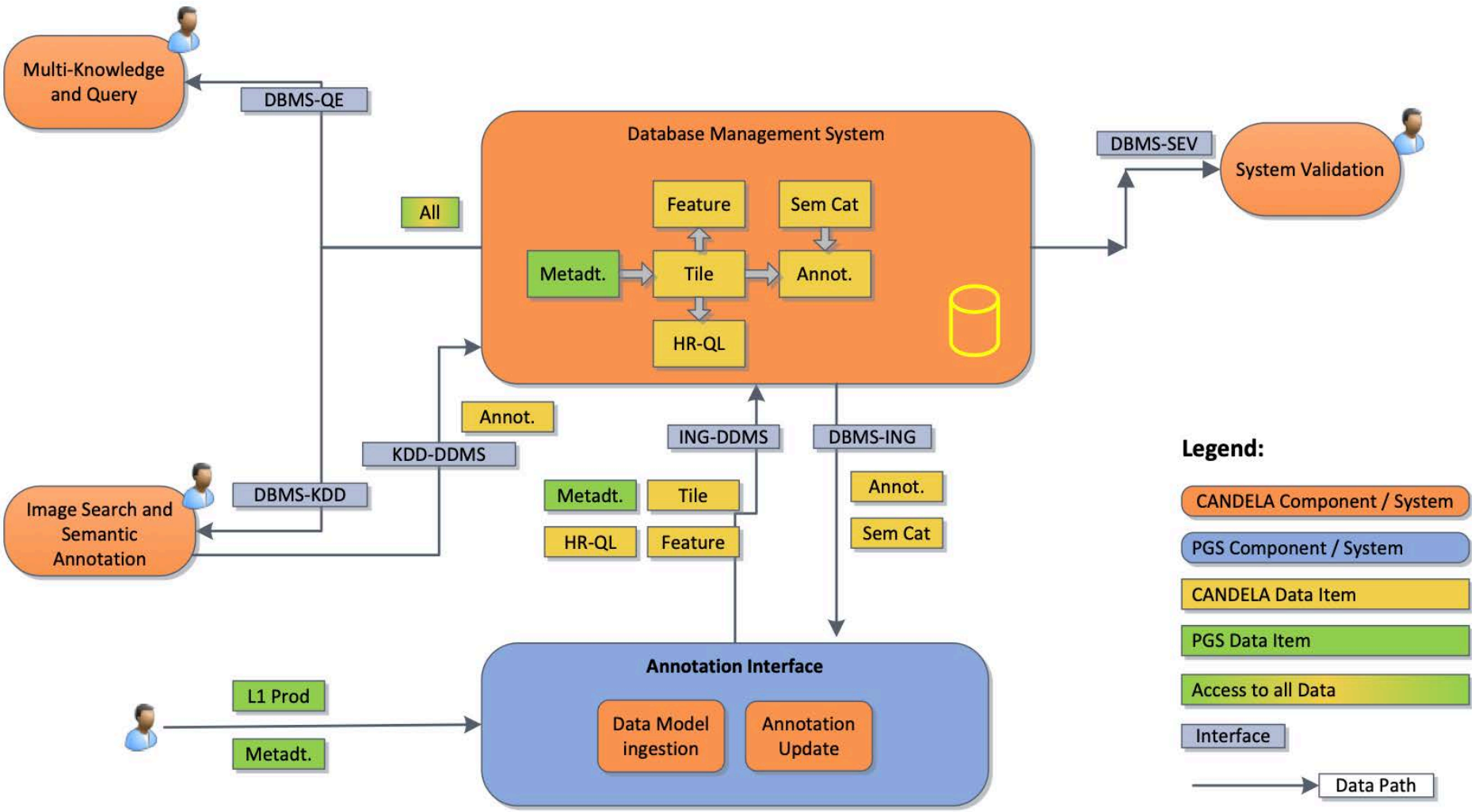
EO Digital Librarian EOLib



The CANDELA analytics modules



Data Mining and Fusion in CANDELA



Data Science Workflows

- Data mining exploration

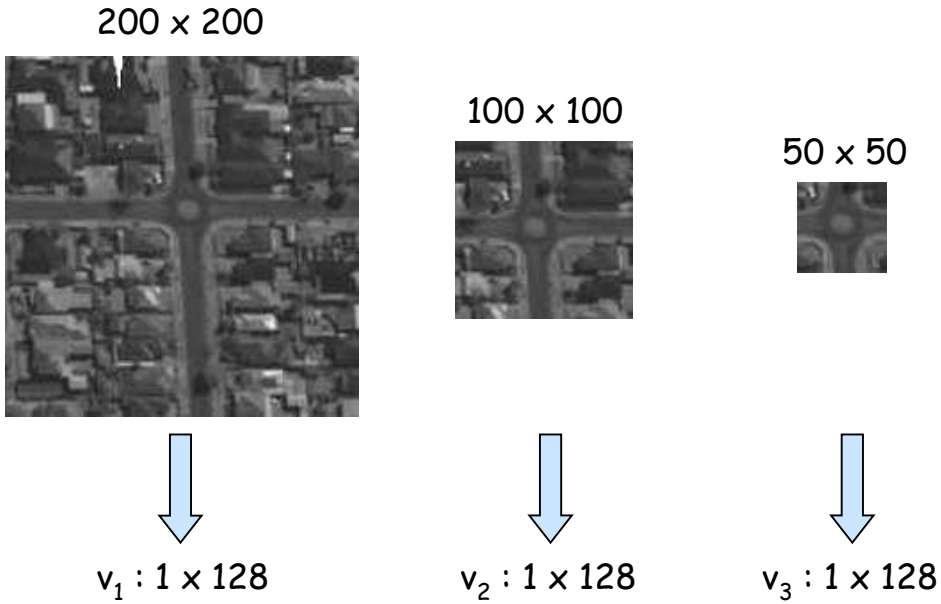
	Capabilities	*CANDELA components*
Step 1	Load and ingest EO images together with their metadata Extract and tile the images into patches	Data Model Generation (DMG) component
Step 2	Automatically ingest all given information into the database	DataBase Management System (DBMS) component
Step 3	Visual 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

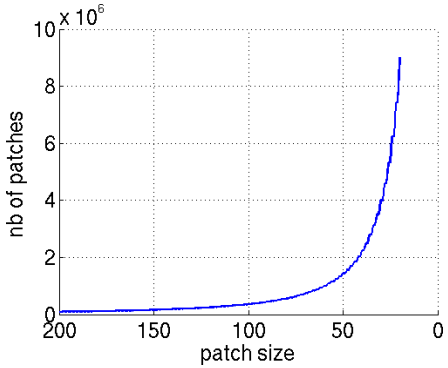
- Data mining semantic annotation

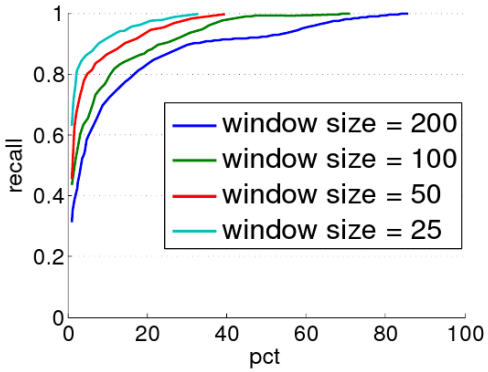
	Capabilities	*CANDELA components*
Step 1	Load and ingest EO images together with their metadata Extract and tile the images into patches	Data Model Generation (DMG) component
Step 2	Automatically ingest all given information into the database	DataBase Management System (DBMS) component
Step 3	Search 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%







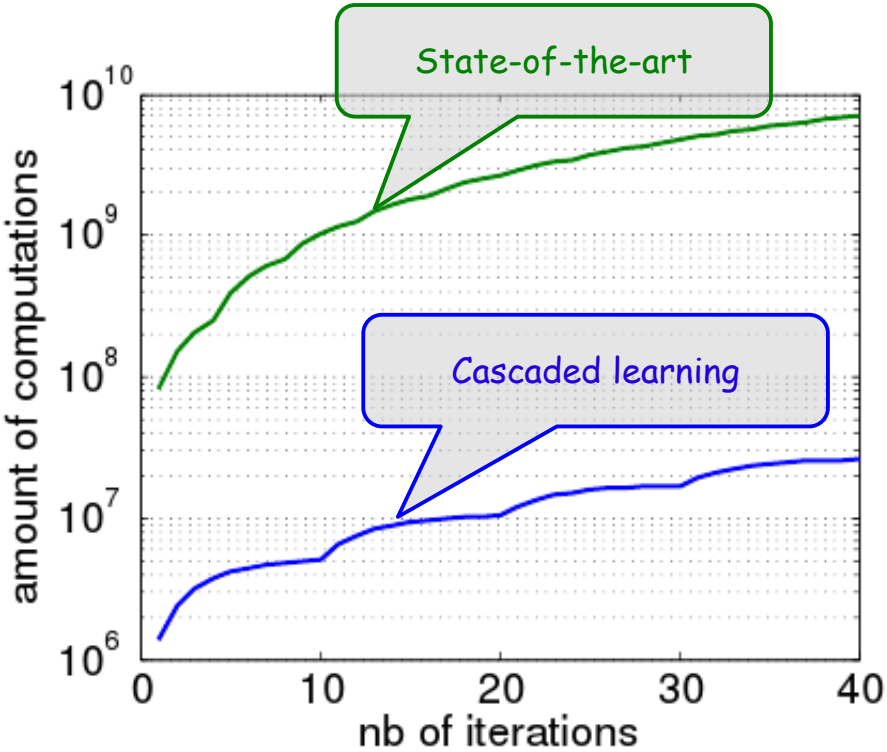
Fast learning

Acceleration with two orders of magnitude

Learning with:

Few
Controllable
Trusted

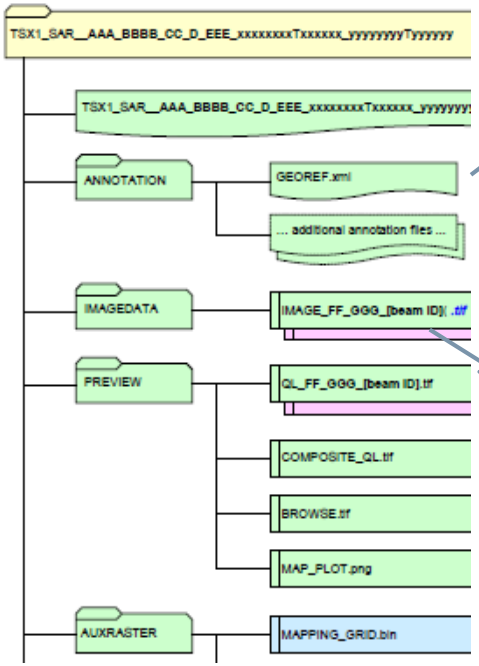
samples



Blanchart, P.; Ferecatu, M.; Shiyong Cui; Datcu, M., "Pattern Retrieval in Large Image Databases Using Multiscale Coarse-to-Fine Cascaded Active Learning," in Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.7, no.4, pp.1127-1141, April 2014

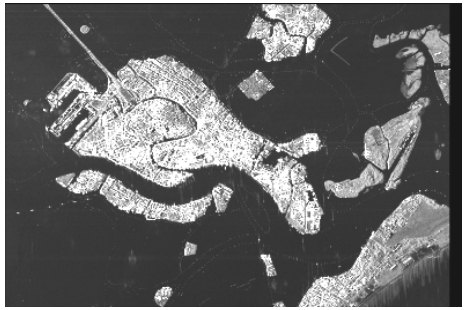
Implementation: Data Model Generation

TerraSAR-X L1b product

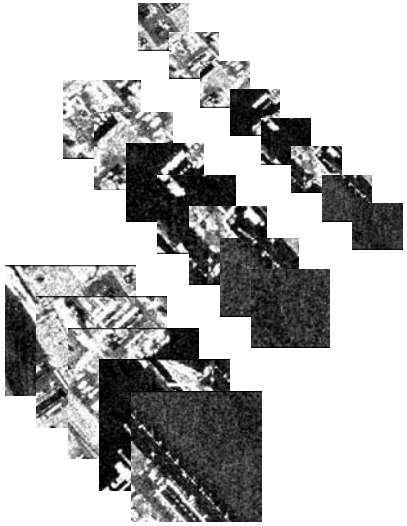


TerraSAR-X metadata and image

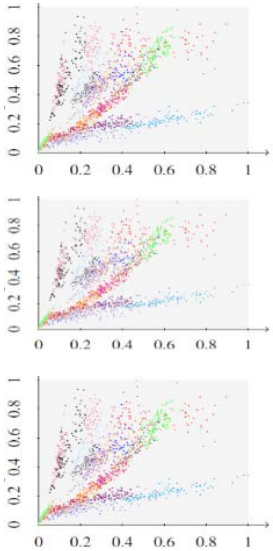
```
<productInfo>
    <missionInfo>
        <mission>TSX-1</mission>
        ...
    </missionInfo>
    <acquisitionInfo>
        ...
    </acquisitionInfo>
    ...
</productInfo>
```



Tiles with different size

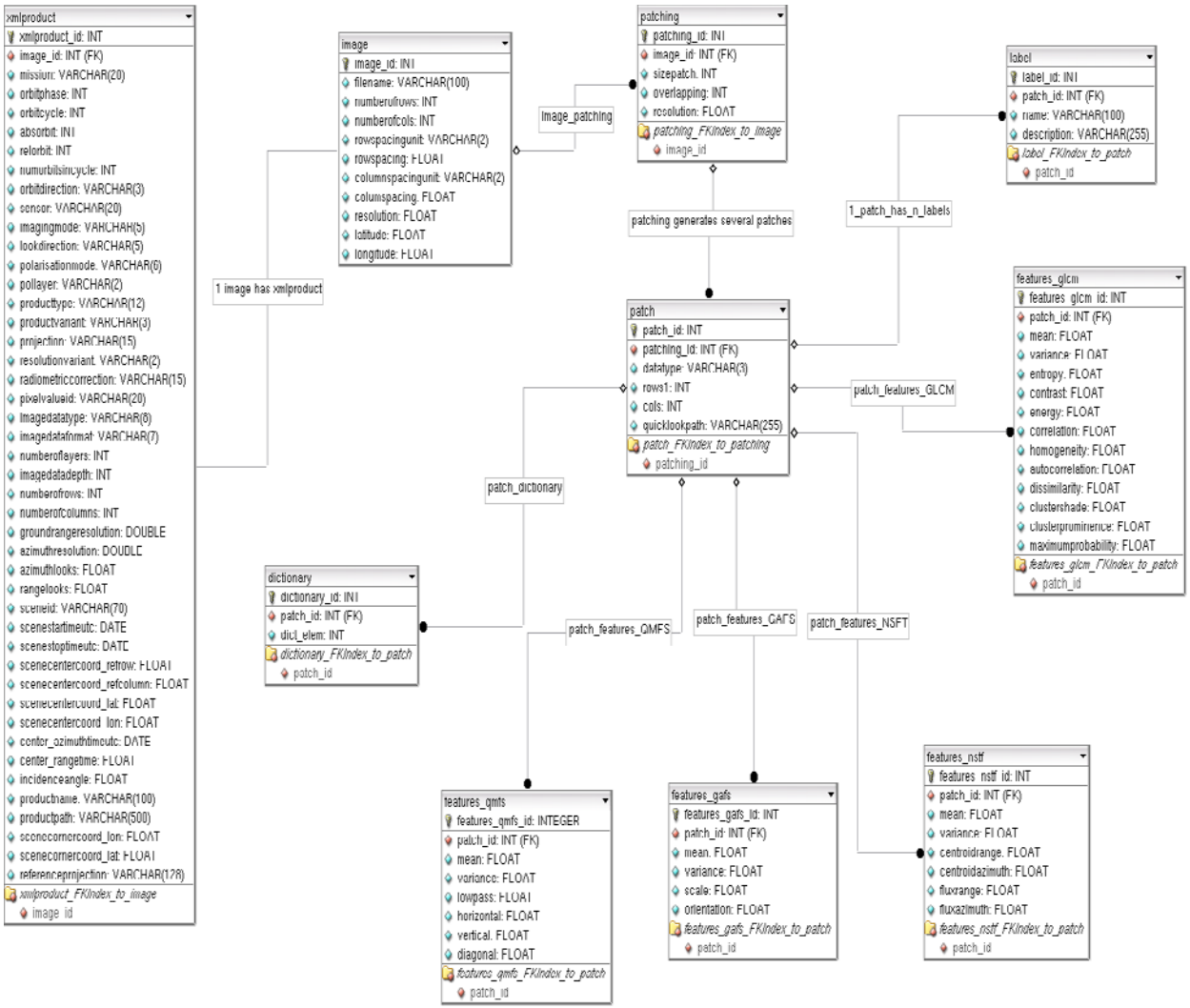


Primitive features: Gabor filters and Weber Local Descriptors



Metadata Extraction → Image Tiling → Quick – looks generation → Primitive Feature extraction → Create the product model

Implementation: Data Mining Data Base



DMDB is a relational database

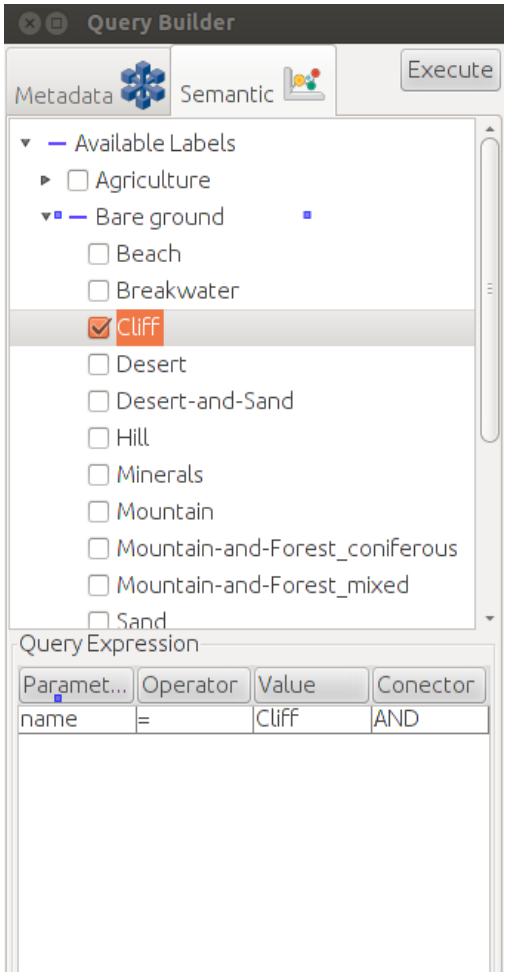
Main tables are:

- Metadata
- Image
- Tiles
- Features
- Labels

DMDB comprises about

- 8 millions of tiles
- 20 thousand metadata entries.
- 106 semantic labels

Implementation: Data Mining



Metadata

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

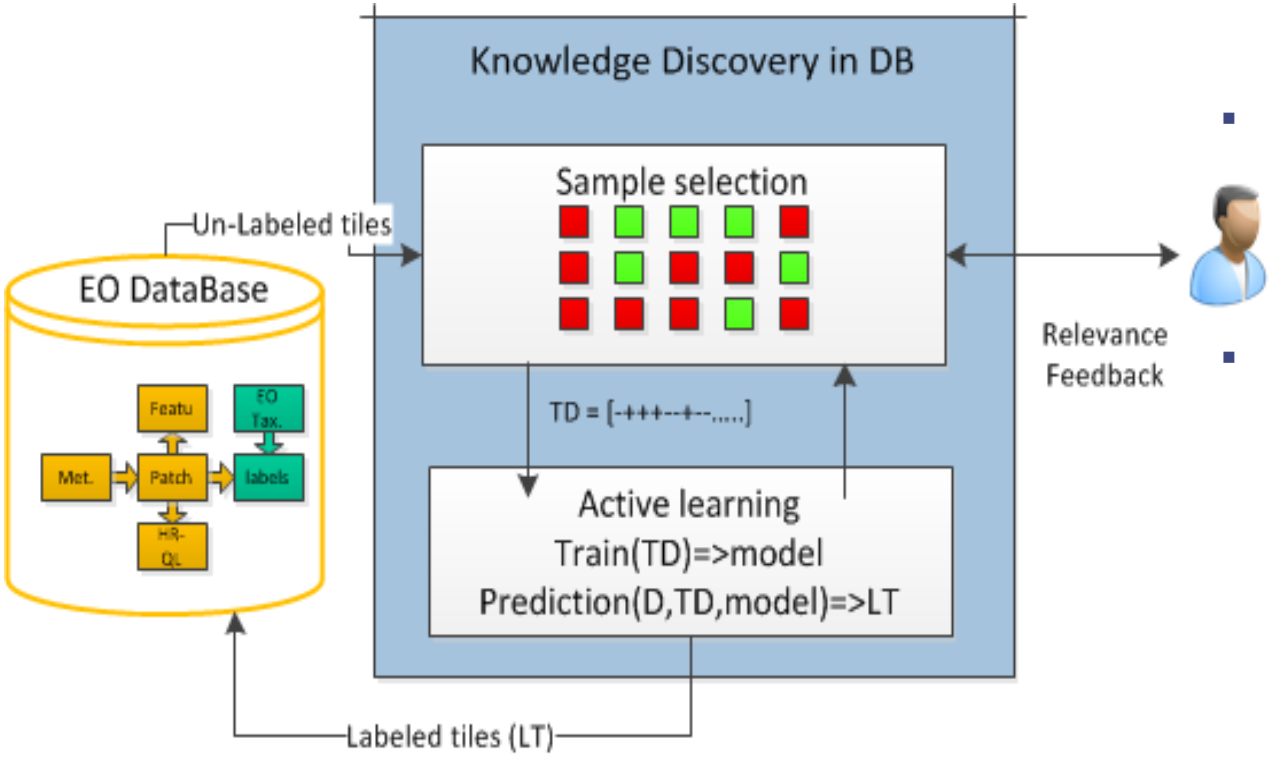
Metadata parameters are based on XML annotation file of TerraSAR-X L1b products

Semantics

- **Agriculture**
  - Cropland
  - Rice plantation…..
- **Bare ground**
  - Cliff
  - Desert…..
- **Urban area**
  - Commercial areas
  - High density residential areas….
- **Forest**
  - Forest coniferous
  - Forest mixed….

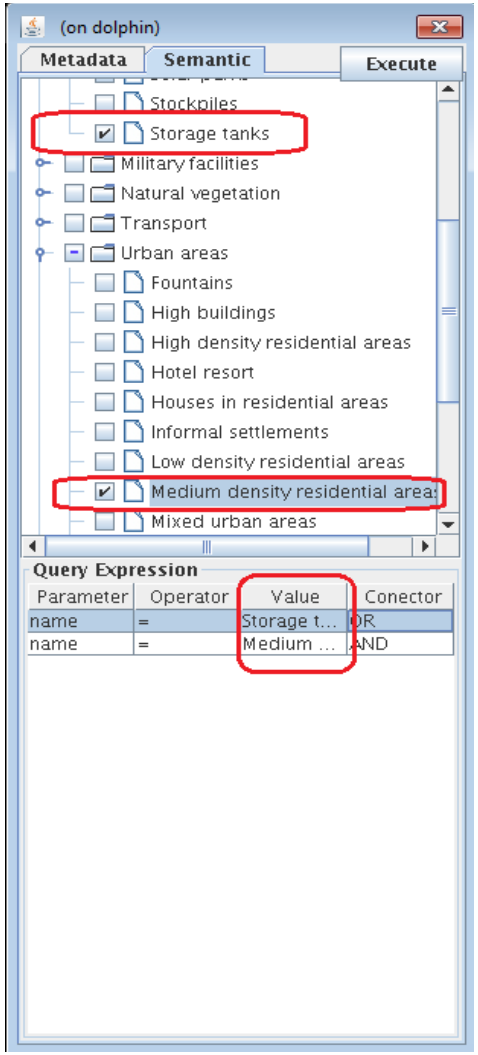
Semantic parameters are based on EO Taxonomy

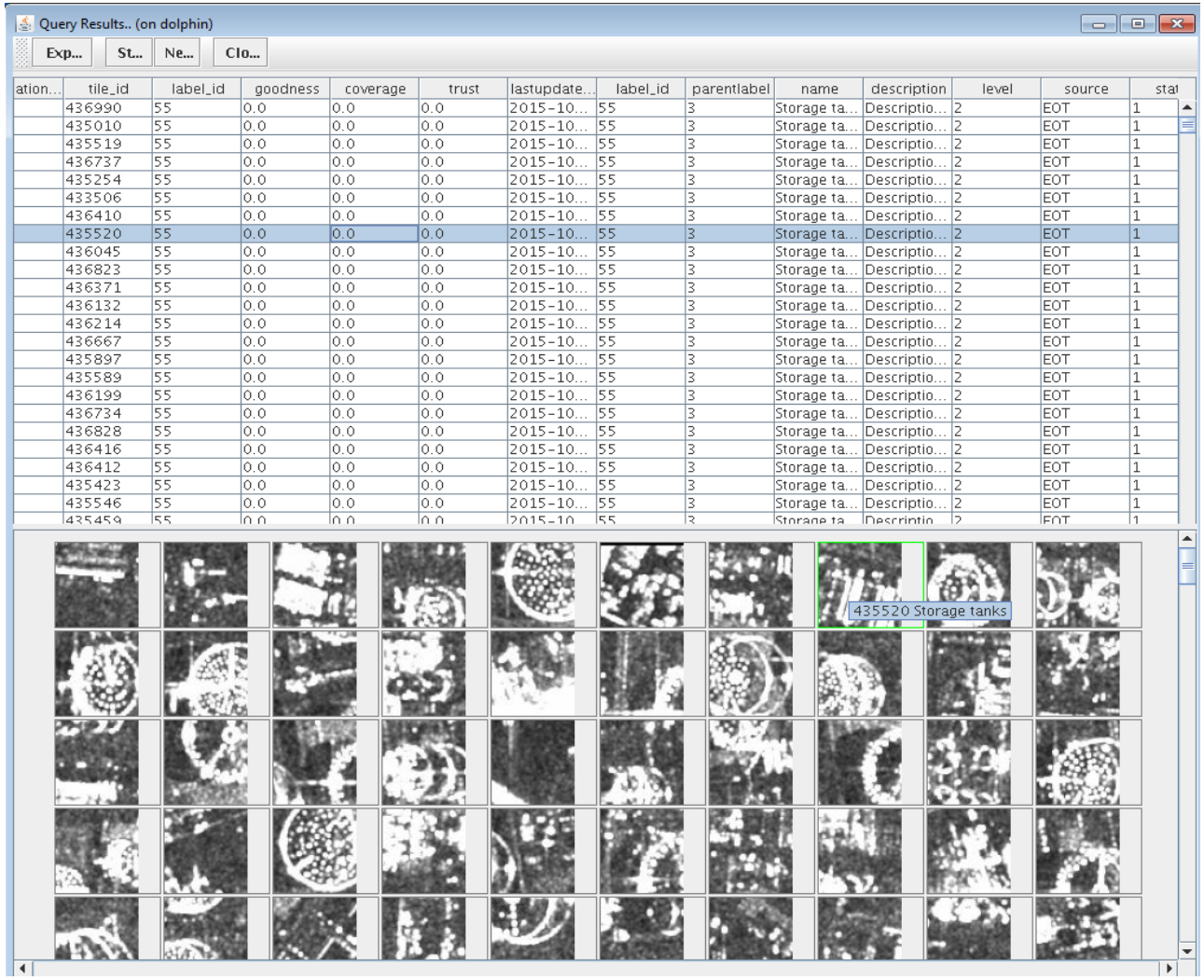
Data Mining



- 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





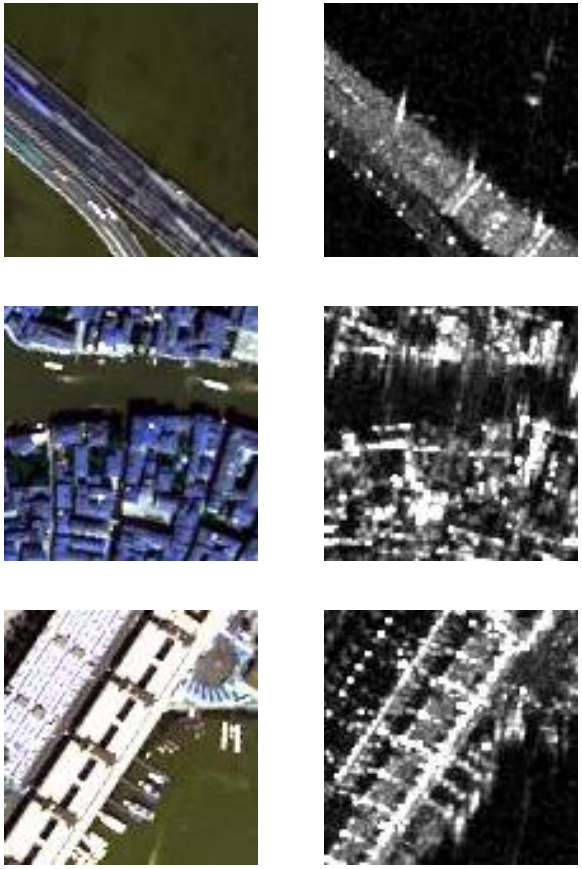
ation...	tile_id	label_id	goodness	coverage	trust	lastupdate...	label_id	parentlabel	name	description	level	source	stat
	436990	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435010	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435519	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436737	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435254	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	433506	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436410	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435520	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436045	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436823	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436371	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436132	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436214	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436667	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435897	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435589	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436199	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436734	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436828	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436416	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	436412	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435423	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435546	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1
	435459	55	0.0	0.0	0.0	2015-10...	55	3	Storage ta...	Descriptio...	2	EOT	1

Data Fusion: SAR vs. MS EO

TerraSAR-X vs. WordView

The clouds

Complementary features



Data Fusion: Validation Data Sets

SAR instrument	TerraSAR-X	Sentinel-1
Image location	Bucharest (Romania) Washington (USA)	Munich (Germany) Venice (Italy)
Acquisition time	Aug. 15, 2009 (Bucharest) June 22, 2010 (Washington)	April 24, 2013 (Munich) Sept. 05, 2012 (Venice)

Multispectral instrument	WorldView-2	
Image location	Bucharest (Romania) Venice (Italy)	Munich (Germany) Washington (USA)
Acquisition time	Oct. 29, 2010 (Bucharest) Sept. 08, 2012 (Venice)	July 12, 2010 (Munich) June 19, 2010 (Washington)

Data Fusion: Selected Results

No.	Seamntic annotation	No. of patches	Multispectral		SAR		Fused images	
			Precision	Recall	Precision	Recall	Precision	Recall
1	Administrative and Monument areas	646	50.29	36.47	44.49	42.30	94.78	73.21
2	Bridges	24	42.42	58.33	33.45	37.50	80.95	70.83
3	Broadleaf forest	1061	82.96	41.67	56.57	52.87	95.39	76.06
4	Cemeteries	72	44.45	36.67	41.10	36.57	91.67	30.56
5	Grassland	201	41.94	71.14	40.29	77.62	78.00	84.03
6	High-density residential areas	617	46.45	58.99	43.64	39.66	96.98	57.37
7	Medium-density residential areas	3120	73.97	57.12	51.51	42.05	94.75	89.58
8	Mixed urban areas	374	56.00	39.21	53.24	38.72	80.21	40.11
9	Parking areas	143	60.61	43.97	50.00	37.00	52.76	46.85
10	Rivers	120	69.37	64.17	59.08	47.50	80.00	80.33
11	Roads	949	56.37	45.39	47.84	42.33	98.60	22.34
12	Sports grounds	21	100.00	80.95	52.31	58.10	85.45	79.00
			60.40	**52.84**	**47.79**	**46.02**	**85.80**	**62.52**

candela

Thank you for your attention

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776193

www.candela-h2020.eu