SOFT COMPUTING FOR TRANSPARENT SYNTHESIS OF GEO BIG DATA

GLORIA BORDOGNA

CNR IREA

CONSIGLIO NAZIONALE DELLE RICERCHE

ISTITUTO PER IL RILEVAMENTO ELETTROMAGNETICO DELL'AMBIENTE

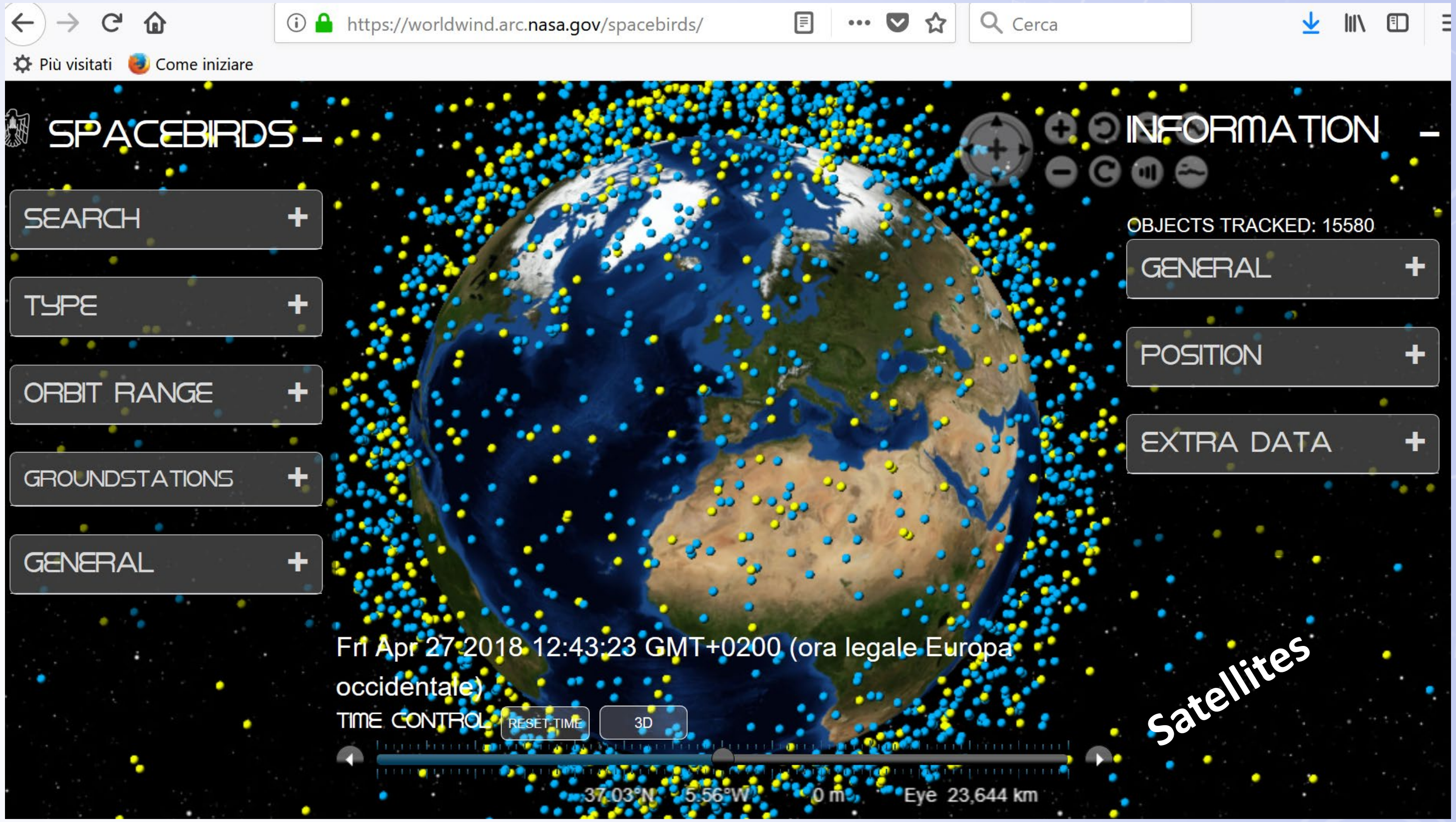
MILANO ITALY

What is Geo Big Data ?

Geo Big Data a BIG Data with a georeference (geofootprint) on Earth

80 % of the 2.5 trillion bytes of data created every day are explicitly or implicitly georeferenced.

[Big Geo Data, A.M. Brovelli, Keynote, OGRS, Perugia, 2016]



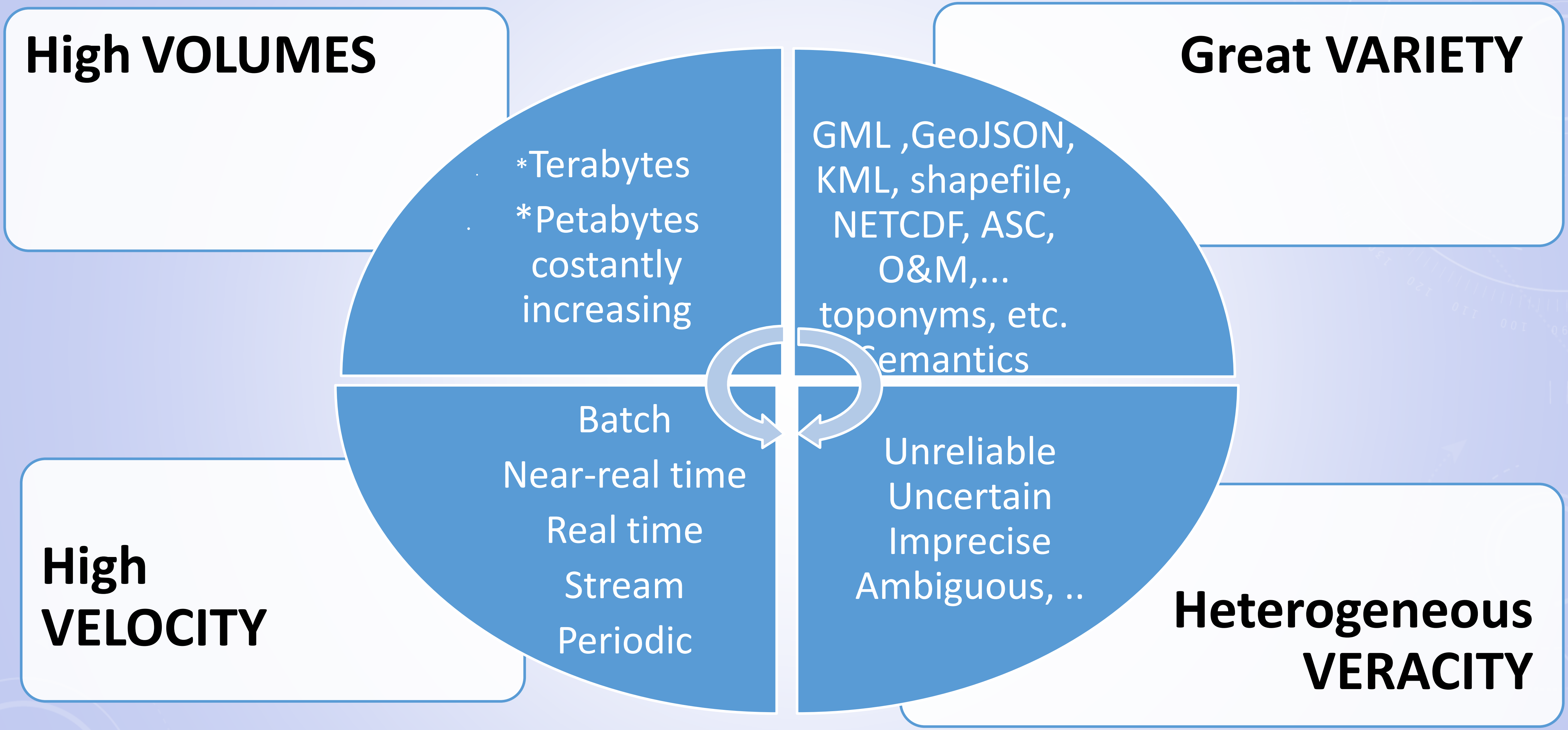
McKinsey&Company

"Data Sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

WIKIPEDIA
The Free Encyclopedia

"Data sets so large and complex that it becomes difficult to process using traditional data processing applications."

Geo Big Data 4Vs



High VOLUMES

*Terabytes

*Petabytes costantly increasing

Great VARIETY

GML ,GeoJSON, KML, shapefile, NETCDF, ASC, O&M,... toponyms, etc. Semantics

High VELOCITY

Batch

Near-real time

Real time

Stream

Periodic

Heterogeneous VERACITY

Unreliable

Uncertain

Imprecise

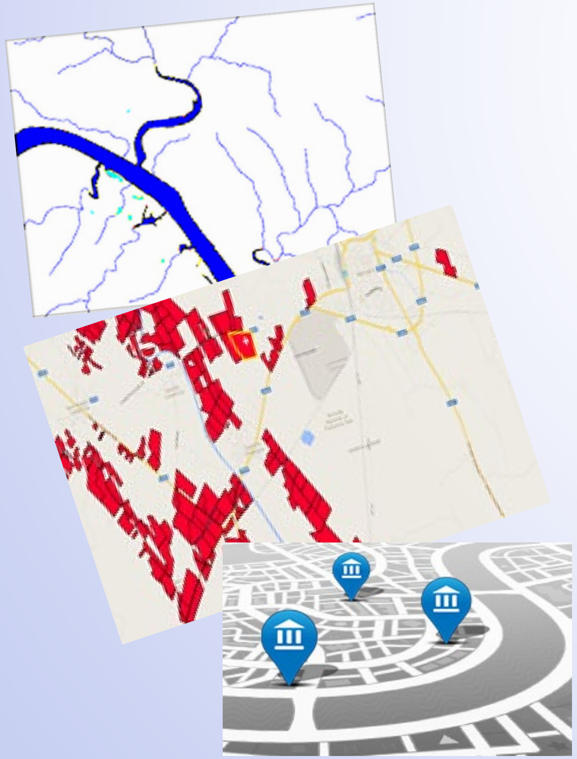
Ambiguous, ..

Geo Big Data are Complex

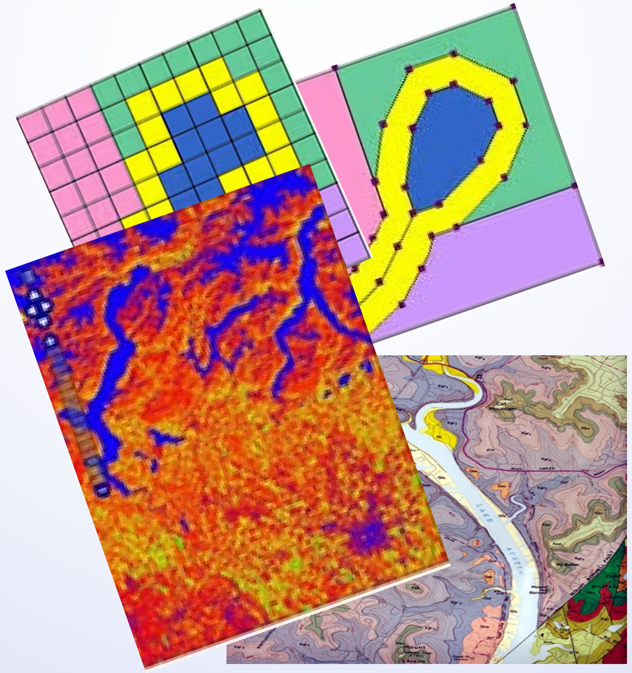
Spatial versus Platial

Tuan, Y.-F. (1977). Space and place: The perspective of experience.

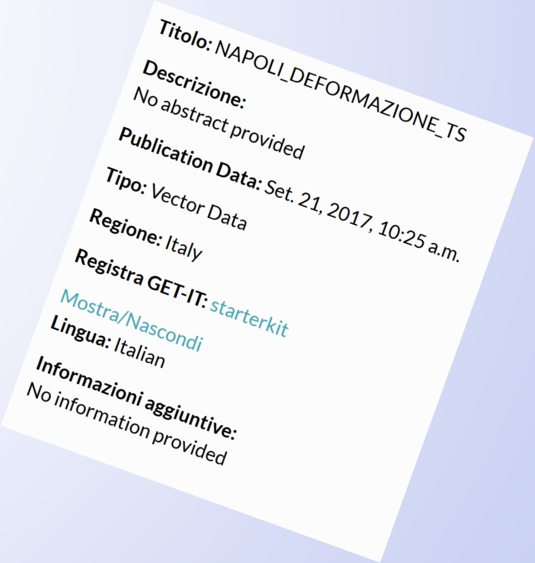
Objects and Features



Field - Coverages



Metadata & text (toponyms)



Titolo: NAPOLI_DEFORMAZIONE_TS

Descrizione:
No abstract provided

Publication Data: Set. 21, 2017, 10:25 a.m.

Tipo: Vector Data

Regione: Italy

Registra GET-IT: starterkit

Mostra/Nascondi

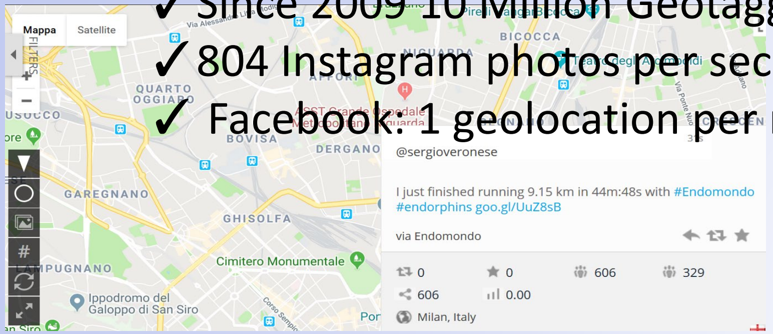
Lingua: Italian

Informazioni aggiuntive:
No information provided

Geo Big Data Sources

Social media

- ✓Since 2009 10 Million Geotagged Tweets per day
- ✓804 Instagram photos per sec 20% georeferenced
- ✓ Facebook: 1 geolocation per min.


@sergioveronese

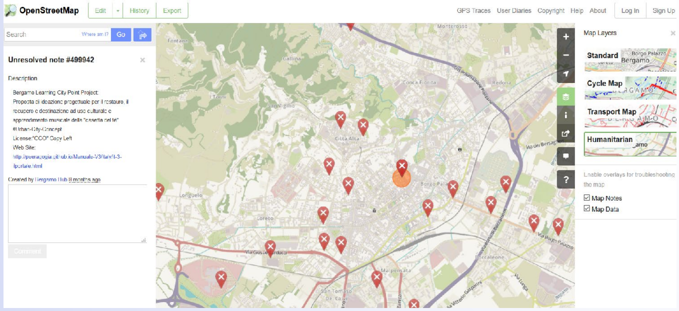
I just finished running 9.15 km in 44m:48s with #Endomondo #endorphins goo.gl/UuZ8sB

via Endomondo

0 0 606 329
606 0.00
Milan, Italy

Volunteered Geographic Information (VGI)

Goodchild , M. F. (2007). Citizens as sensors: GeoJournal


Unresolved note #499942

Description

Bergamo Learning City Point Project:
Proposta di ideazione progettuale per il restauro, il recupero e destinazione ad uso culturale e apprendimento musicale della "casetta del tè"
#Urban-City-Concept
License:"CCO" Copy Left
Web Site:
http://peeragogia.github.io/Manuale-V3/italy/t-3-lportale.html

Created by Bergamo Hub 8 months ago

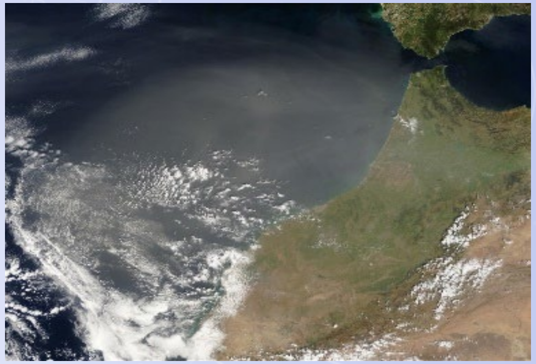
OpenStreetMap (2016 -11- 03 00:00:09 +0000) 6 million users

IoT & cell-phone data

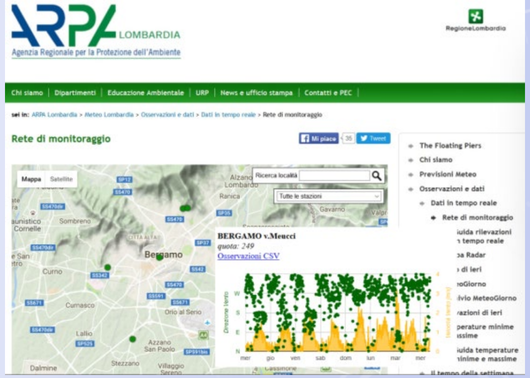
Play store:10 geolocations per min

Remote Sensing Images & products

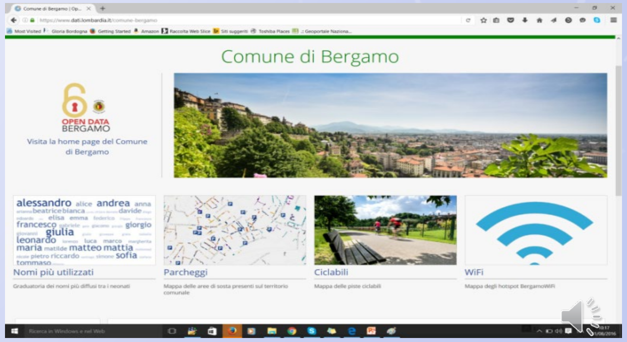
- ✓ Landsat since 1972 . Nasa Earth Observatory hub: (1998 – 2017)
- ✓ EC Copernicus program 972.343.516.862 Tb Sentinel data : (12 Tb per day)



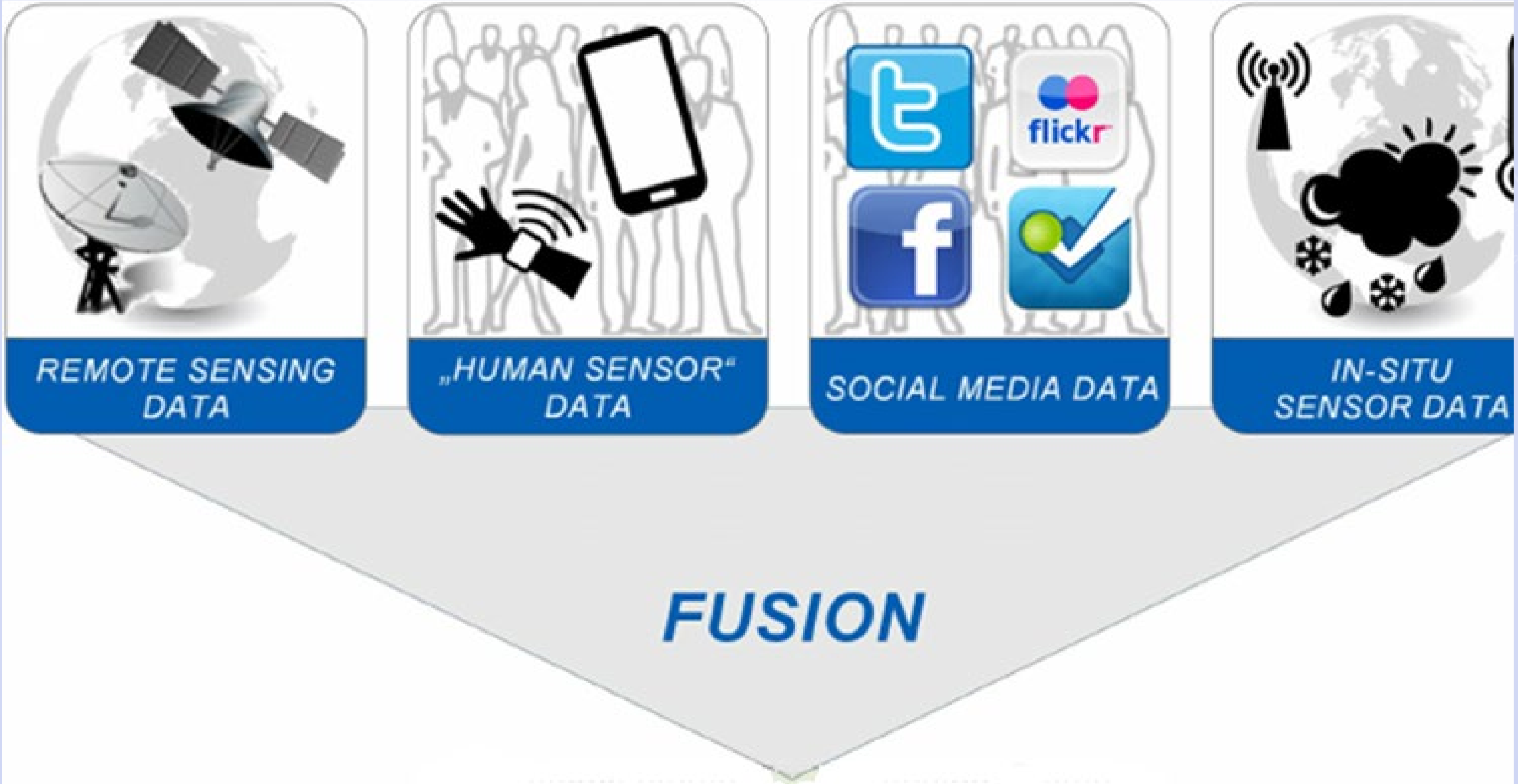
In situ Sensor data



Open data from E-government portals

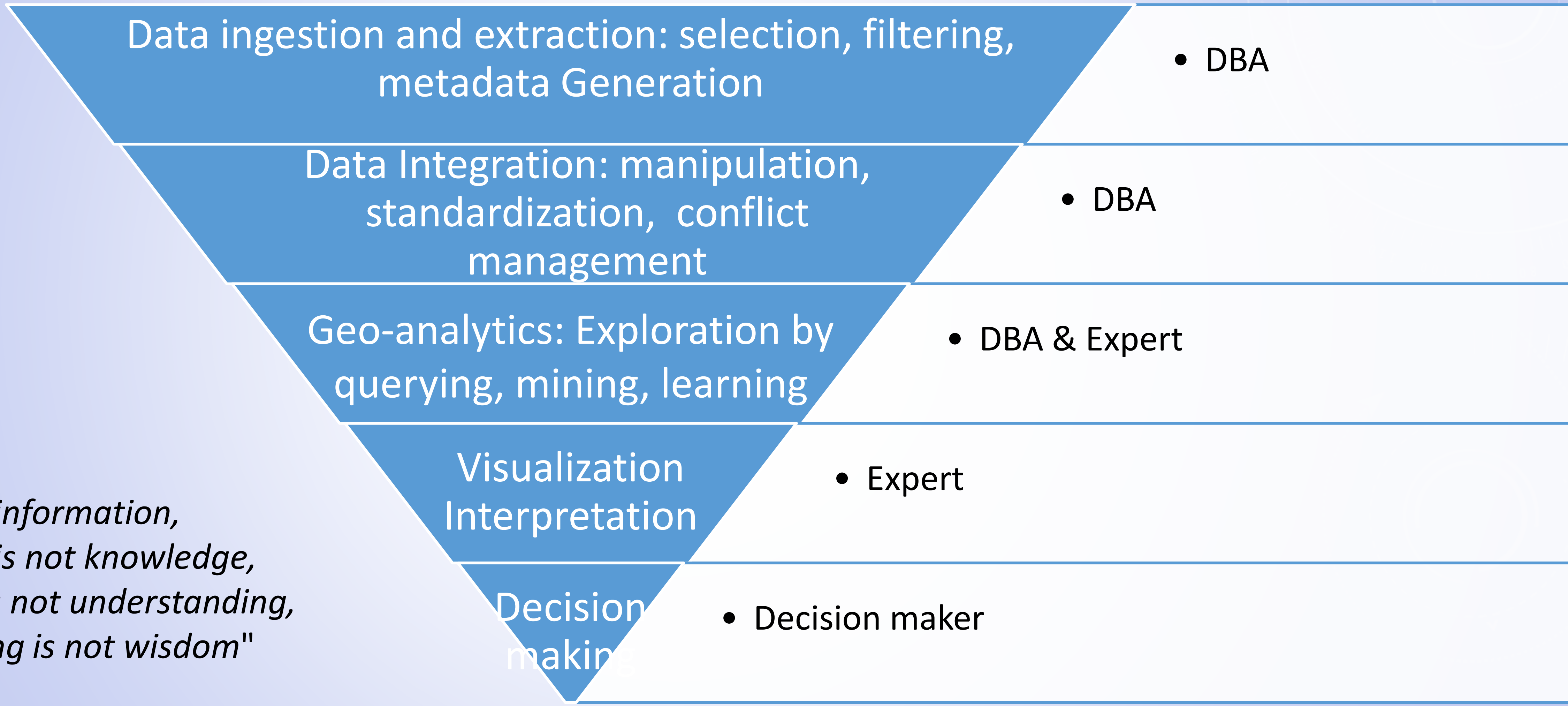


Geo BIG Data Challenge: Multisource Synthesis



To allow better decisions

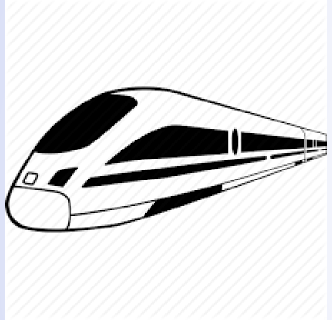
Geo Big Data Value Chain



Data ingestion and extraction: selection, filtering, metadata Generation	DBA
Data Integration: manipulation, standardization, conflict management	DBA
Geo-analytics: Exploration by querying, mining, learning	DBA & Expert
Visualization Interpretation	Expert
Decision making	Decision maker

"*Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom*" (Cliff Stoll)

Solutions



Efficiency: the ability to process High Volumes at High Speed at low cost

Effectiveness: the ability to extract useful information to take decisions:

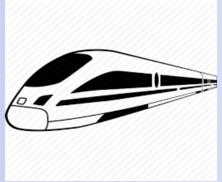
- ✓ Select reliable information considering its Veracity
- ✓ Analyse Geo Big Data by considering its Great Variety

Solutions

High VOLUMES

Distributed data storage

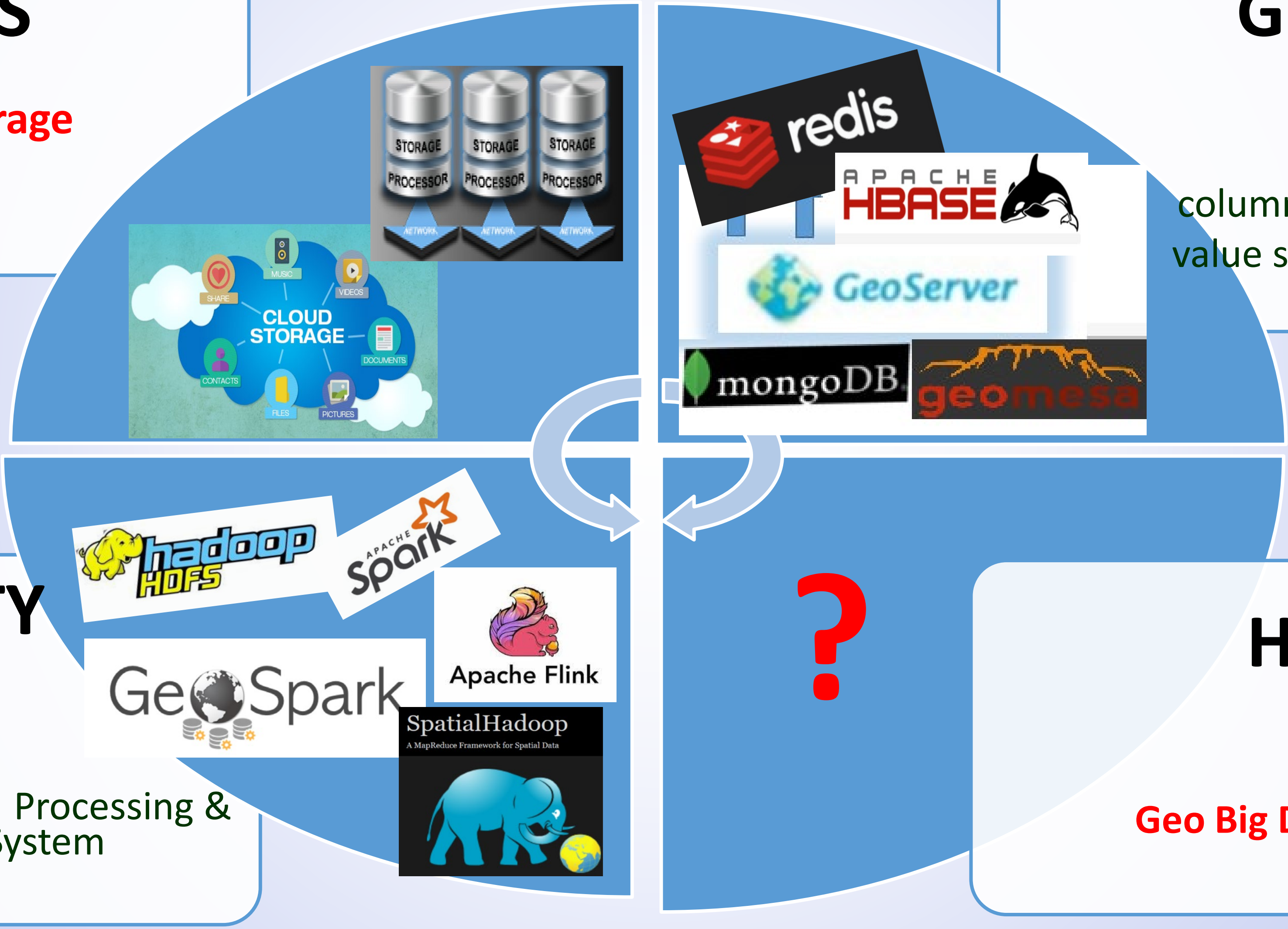
Horizontal scaling



Great VARIETY

NoSQL databases & OGC Web geo services

column & document stores, key-value stores, WMS WFS, CSW, etc



High VELOCITY

Distributed Data Infrastructures

- ✓ Distributed Data Processing & Distributed File System

Heterogeneous VERACITY

Geo Big Data are often assumed as synonymous of facts

Challenges for veracity of Geo Big Data

U. Sivarajah , M. M. Kamal, Z. Irani, V. Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research 70, (2017)

Semantic Interoperability

- ✓ Space versus places
- ✓ Need to represent data and process semantics

Quality assurance & assessment

- ✓ Need to represent and manage imprecision and uncertainty of data;
- ✓ Need to model fitness for use

Flexible & transparent Synthesis

- ✓ Need to cope with distinct needs: redundancy, conflicts, complementarity,..
- ✓ Need of human interpretable results: explainability of the criteria to experts and decision makers

Opportunities offered by Soft Computing

L.A.Zadeh, 1994 Soft computing and fuzzy logic, IEEE Software, 48-56

Soft computing is a branch of AI comprising methodologies that aim to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost. Its principal constituents are fuzzy logic, neuro- computing, and probabilistic reasoning**.**

Semantic Interoperability

- ✓ **Fuzzy sets** allow to represent the semantics of linguistic concepts such as *high, low, big,* etc..

Quality assurance & assessment

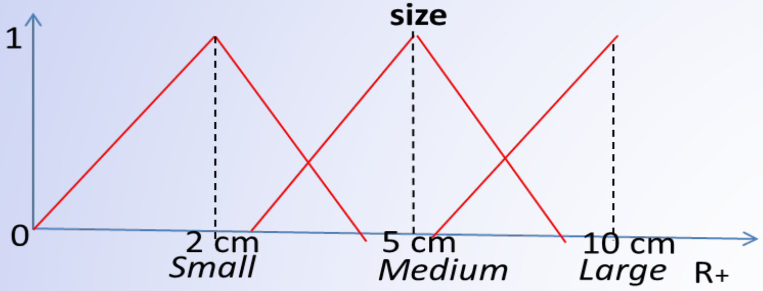
- ✓ **Fuzzy ontologies** allow to represent ill-defined domain knowledge and approximate reasoning to compute fitness for use

Flexible & transparent Synthesis

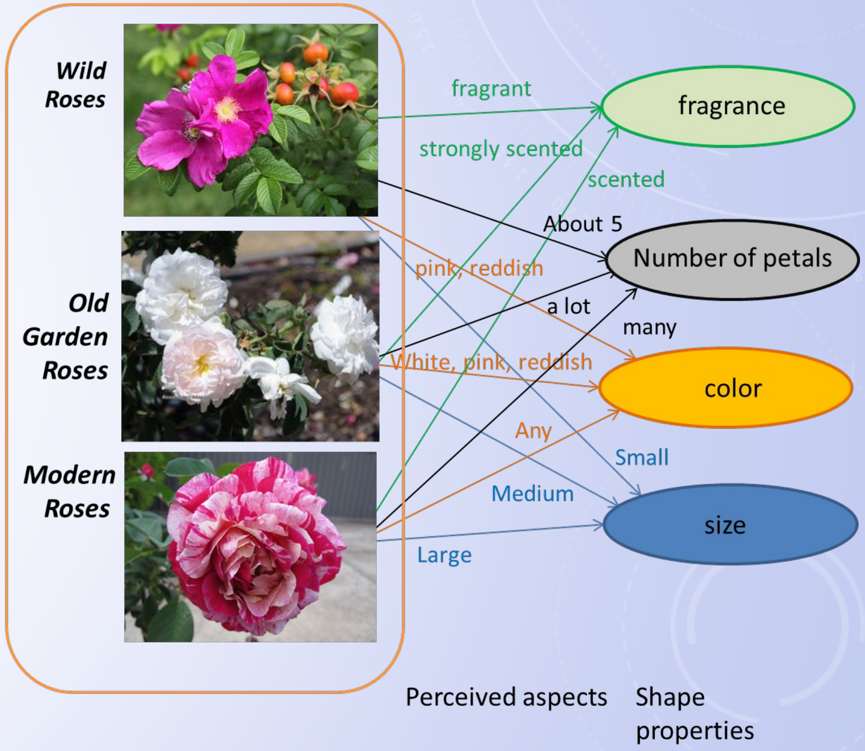
- ✓ **Fuzzy aggregation operators** allow to model decision attitudes and different importance/reliability/trust of data; **Fuzzy clustering** allow to generate groups with faint boundaries

Basic notions of Soft Computing

Membership functions of fuzzy sets define the semantics of linguistic values



Fuzzy ontologies define vague/imprecise knowledge in a doman: ex. Definitions of wild, old-garden and modern roses

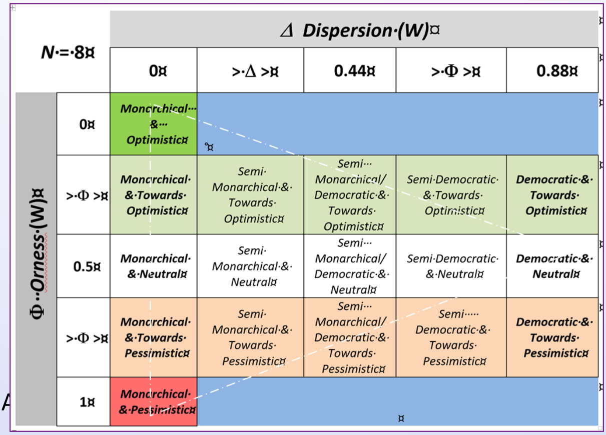


Fuzzy operators allow defining distinct kinds of aggregations of their arguments by satisfying different properties: modeling gradual compensativeness/optimism/democratic behaviours

***Fuzzy operator*: [0,1]^N→ [0,1]**

All <= most <=average <= at least some <=at least 1

T-Norms = AND=Min ([d1, ...,dN])<=
OWA([d1, ...,dN])<=
Max([d1, ...,dN])=OR = T-Conorm



	N = 8	Δ Dispersion (W)				
		0	> Δ >	0.44	> Φ >	0.88
Φ Orness (W)	0	Monarchical & Optimistic				
	> Φ >	Monarchical & Towards Optimistic	Semi Monarchical & Towards Optimistic	Semi Monarchical/ Democratic & Towards Optimistic	Semi Democratic & Towards Optimistic	Democratic & Towards Optimistic
	0.5	Monarchical & Neutral	Semi Monarchical & Neutral	Semi Monarchical/ Democratic & Neutral	Semi Democratic & Neutral	Democratic & Neutral
	> Φ >	Monarchical & Towards Pessimistic	Semi Monarchical & Towards Pessimistic	Semi Monarchical/ Democratic & Towards Pessimistic	Semi Democratic & Towards Pessimistic	Democratic & Towards Pessimistic
	1	Monarchical & Pessimistic				

Case study: Quality assurance of Volunteered Geographic Information

G Bordogna et al. "Contextualized VGI" Creation and Management , ISPRS IJGI, 2016

VGI and in situ georeferenced observations are affected by both imprecision and epistemic uncertainty that degrade the quality of the information

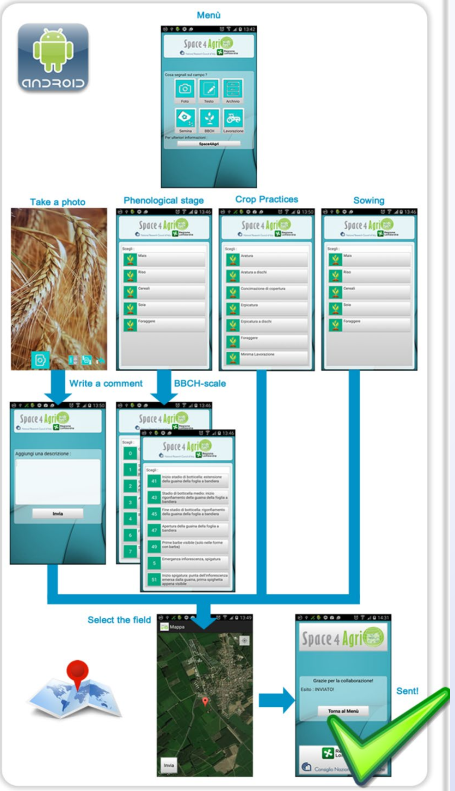
How can be cope with it?

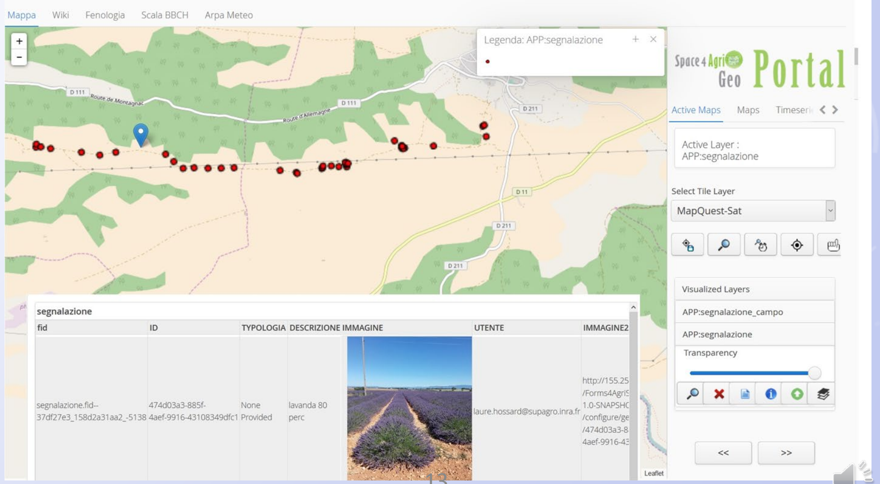
Agronomists and farmers need to geotag crops' types & growth stages :

- ✓ Phenological stages (BBCH ontology)
- ✓ Photos and free text

Problems:

- ✓ vague knowledge:
  Principal growth stage 1: Leaf development
  Principal growth stage 3: Stem elongation
- ✓ variability of phenology
- ✓ uncertainty of landmark



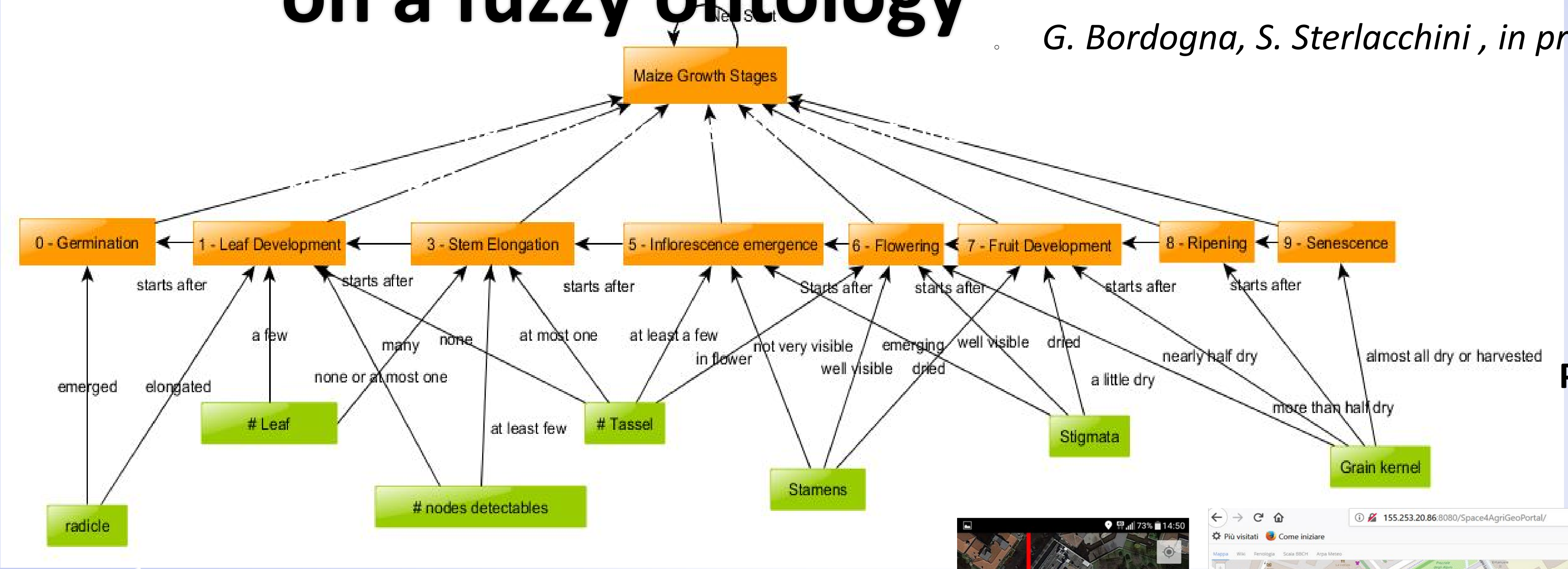


segnalazione

fid	ID	TYPOLOGIA	DESCRIZIONE	IMMAGINE	UTENTE	IMMAGINE2
segnalazione.fid--37df27e3_158d2a31aa2_-5138	474d03a3-885f-4aef-9916-43108349dfc1	None Provided	lavanda 80 perc		laure.hossard@supagro.inra.fr	http://155.25 /Forms4AgriS 1.0-SNAPSHC /configure/ge /474d03a3-8 4aef-9916-43

Case study: Creating in situ observations based on a fuzzy ontology

G. Bordogna, S. Sterlacchini , in proc. of IEA/AIE 2017



Phenological stages

Perceived aspects

Properties

Observation

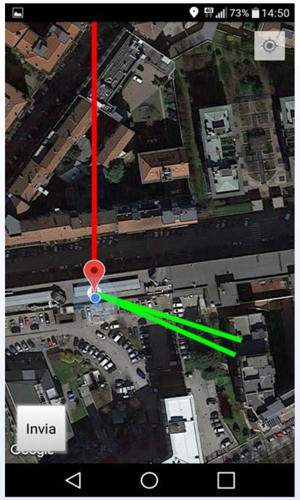
# leafs : *many*

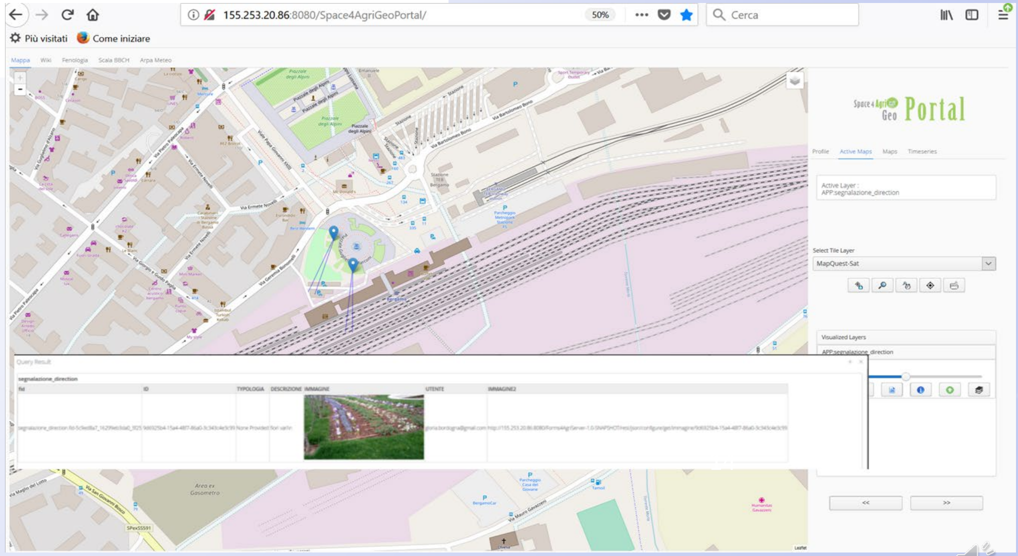
# nodes : *at most 1*

Stage 1: *leaf development 0,5*

Stage 2: *Stem elongation 0,5*

Storage in an a fuzzy database





Flexible & transparent Synthesis of Geo Big Data

Knowledge-based approaches are

- too crisp to generalize when changing study area and observation conditions
- They are transparent & human interpretable (**explainable**)

Machine learning approaches are basically data-driven

- need large sets of Groud Truth Data (GTD) for training often unavailable
- are opaque and do not exploit available knowledge

We do not want «to throw the child with dirty water”

Soft computing :

- It allows representing ill-defined experts' knowledge,
- It allows combining knowledge and data driven approaches with the need of small GTD by explaining learned criteria
- **thus it is compliant with explainable AI**

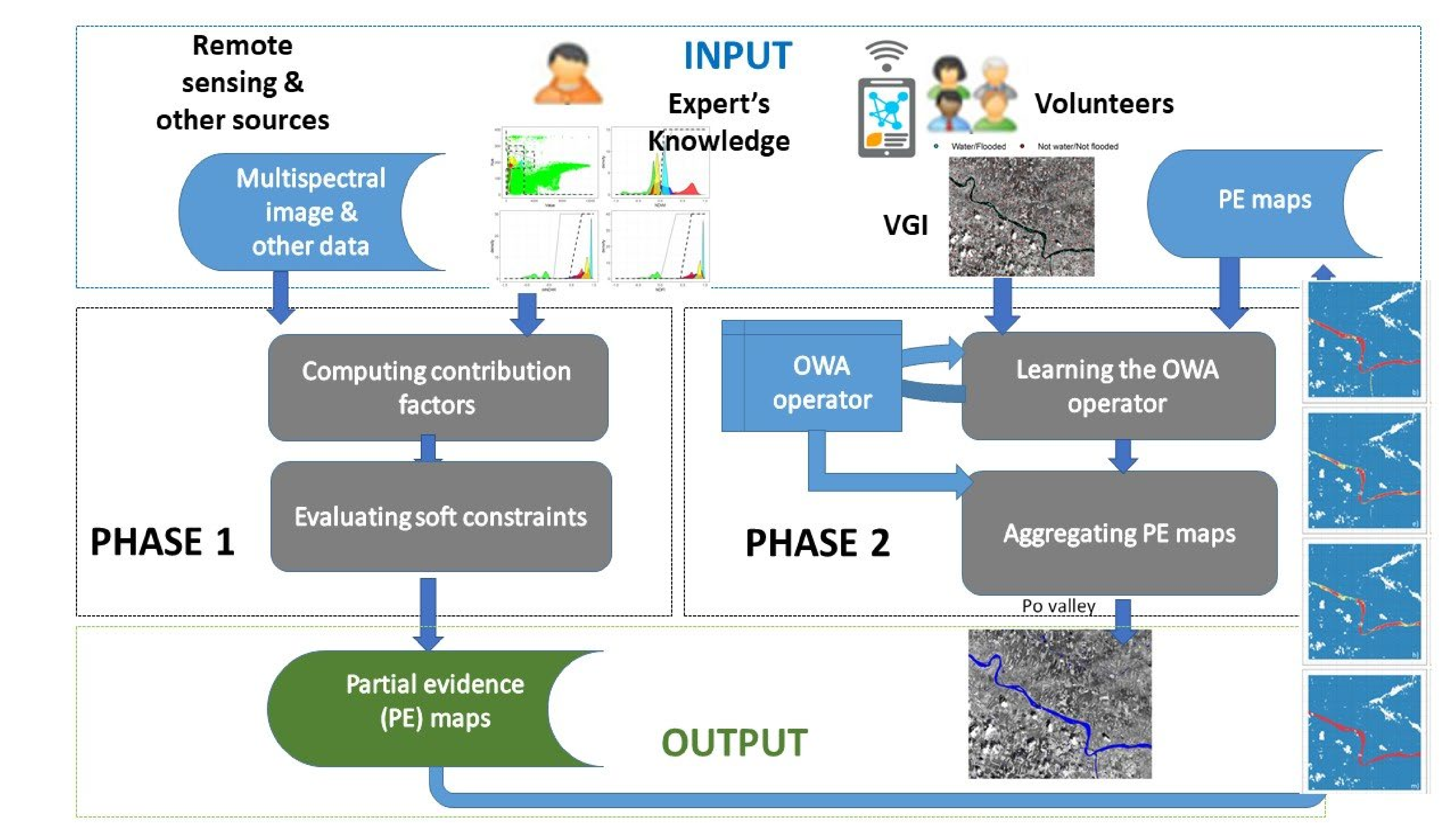
soft approach to flexible synthesis

GIS WORKFLOW:

- **Step 1:** selection of data layers (contributing factors) and application of selection conditions to **segment partial evidence maps**;
- **Step 2: aggregation of Partial evidence maps** by Boolean operators to generate an Environmental Status Indicator (ESI) map

GENERALIZATION

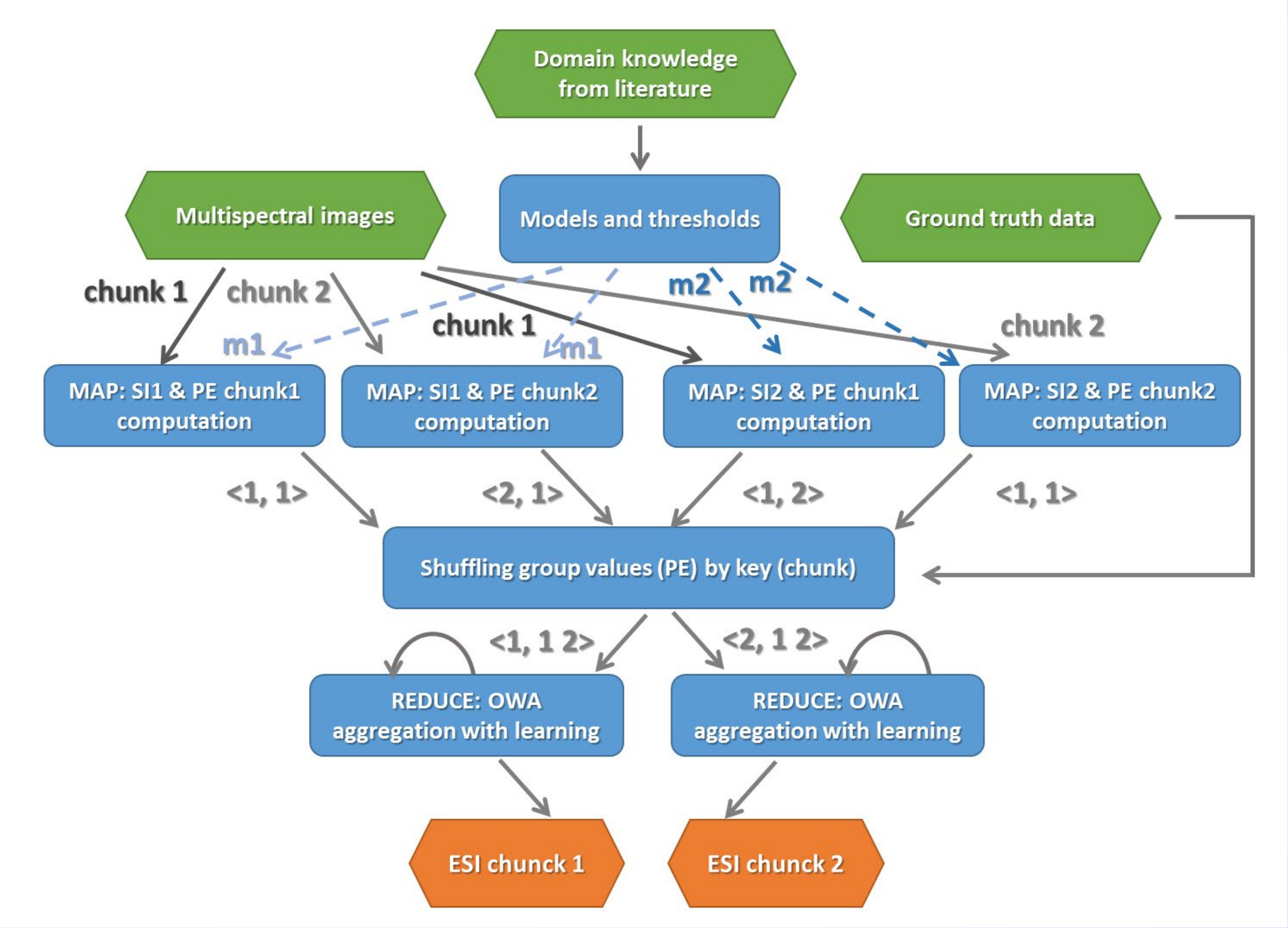
- Step 1: **soft constraints**
- Step 2: **OWA operators,** non-linear mean-like fuzzy operators: either specified by a linguistic quantifier or learned from Ground truth data


INPUT
Remote sensing & other sources
Expert's Knowledge
Volunteers
Water/Flooded
Not water/Not flooded
Multispectral image & other data
VGI
PE maps
Computing contribution factors
Evaluating soft constraints
PHASE 1
OWA operator
Learning the OWA operator
Aggregating PE maps
PHASE 2
Po valley
Partial evidence (PE) maps
OUTPUT

Schema of flexible synthesis Implementation

EFFICIENCY

The 2-step process is applied on each spatial unit (either object or pixel) one independently from others, and thus it can be implemented by exploiting distributed processing



Domain knowledge from literature
Multispectral images
Models and thresholds
Ground truth data
chunk 1
chunk 2
m1
chunk 1
m1
m2
m2
chunk 2
MAP: SI1 & PE chunk1 computation
MAP: SI1 & PE chunk2 computation
MAP: SI2 & PE chunk1 computation
MAP: SI2 & PE chunk2 computation
<1, 1>
<2, 1>
<1, 2>
<1, 1>
Shuffling group values (PE) by key (chunk)
<1, 1 2>
<2, 1 2>
REDUCE: OWA aggregation with learning
REDUCE: OWA aggregation with learning
ESI chunck 1
ESI chunck 2

Case study: Synthesis of multiple spatial data sets

Towards an operational GMES land Monitoring Core Service

geoland

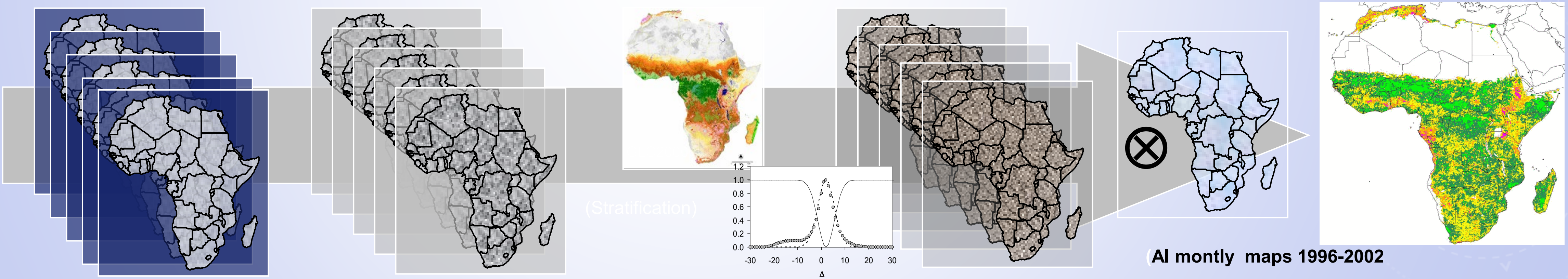
- ✓ *Carrara, G. Bordogna, M. Boschetti, P.A. Brivio, A. Nelson, D. Stroppiana (2008). A flexible multi-source spatial-data fusion system for environmental status assessment at continental scale, International Journal of Geographical Information Science, Vol. 22, 781-799.*
- ✓ *D. Stroppiana, M. Boschetti, P.A. Brivio, P. Carrara, G. Bordogna (2009). A fuzzy anomaly indicator for environmental monitoring at continental scale, Ecological Indicators, Vol. 9, 92-106.*

Synthetic Anomaly Indicator (AI) aggregating contributing factors (partial hints of anomaly) defined as the difference with respect to the reference long term average

Contributing factors' definition

STEP1
Partial Evidence maps computation

STEP 2
Synthesis by OWA aggregation defined by expert thorugh linguistic quantifier most



AI montly maps 1996-2002

(Input decadal maps)

- •10d Phenology
- •10d Rain

(Factors decadal maps)

- •Long term average
- •Difference Δ

(Gaussian interpolation) (Score decadal maps)

- •Stratification by homogeneous areas
- •10d soft constraints definition

- •OWA operator (semantics in between AND –OR)

Case study: Synthesis of remote sensing images & VGI

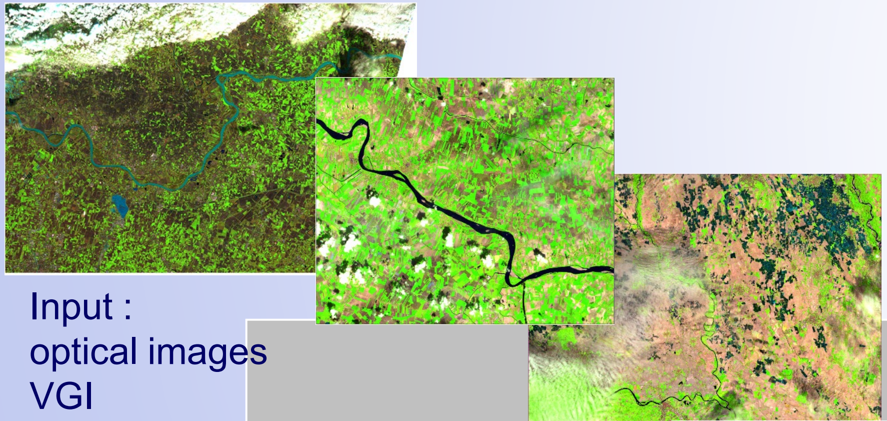
✓ *Goffi et al., Remote Sens. 2020, 12(3), 495; https://doi.org/10.3390/rs12030495*

Mapping standing water areas (flooded areas, water, flooded rice paddies from Sentinel-2 and VGI (in situ observations)

Contributing factors' definition

STEP1
Partial Evidence maps computation

STEP 2
Synthesis by OWA aggregation learned from VGI



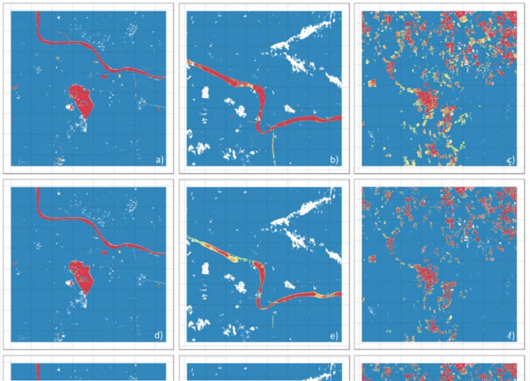
Input :
optical images
VGI

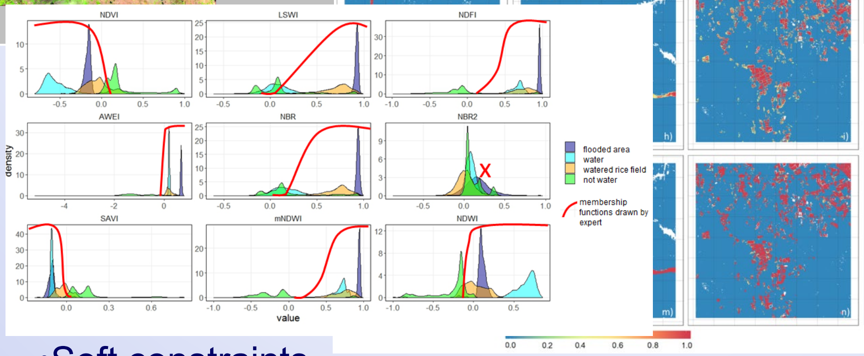
Spectral Index	Formula	Category
AWEI	C1 * (GREEN – SWIR1) – (C2 * NIR + C3 * SWIR2)	**Water**
AWEIsh	BLUE + D1 * GREEN – D2 * (NIR + SWIR1) – D3 * SWIR2	**Water**
mNDWI	(GREEN – SWIR1) / (GREEN + SWIR1)	**Water**
NDWI	(GREEN – NIR) / (GREEN + NIR)	**Water**
NDFI	(RED – SWIR2) / (RED + SWIR2)	**Flooding**
SAVI	(1 + L) * (NIR – RED) / (NIR + RED + L)	**Vegetation**
WRI	(GREEN + RED) / (NIR + SWIR2)	**Water**
HV	F(SWIR2, NIR, RED) where F is defined as in J.F. Pekel et al.1	**Water**

Where C1=4, C2=0.25, C3=2.75, D1=2.5, D2=1.5, D3=0.25, L=0.5
1 High-resolution mapping of global surface water and its long-term changes, Nature volume 540, pages 418-422, 2016

- Water/not water spectral indexes

- Partial evidence maps

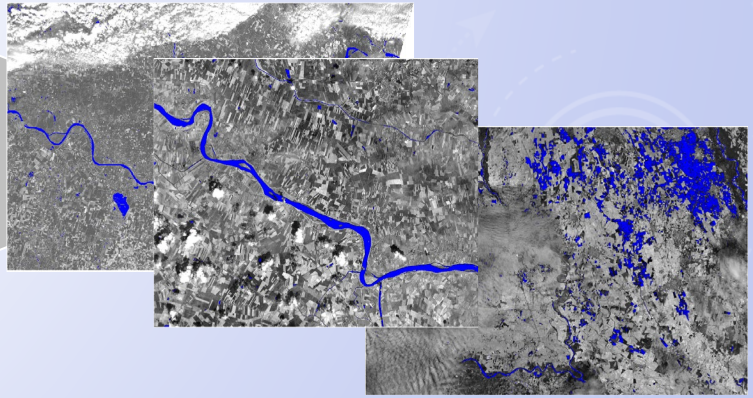




- Soft constraints

⊗

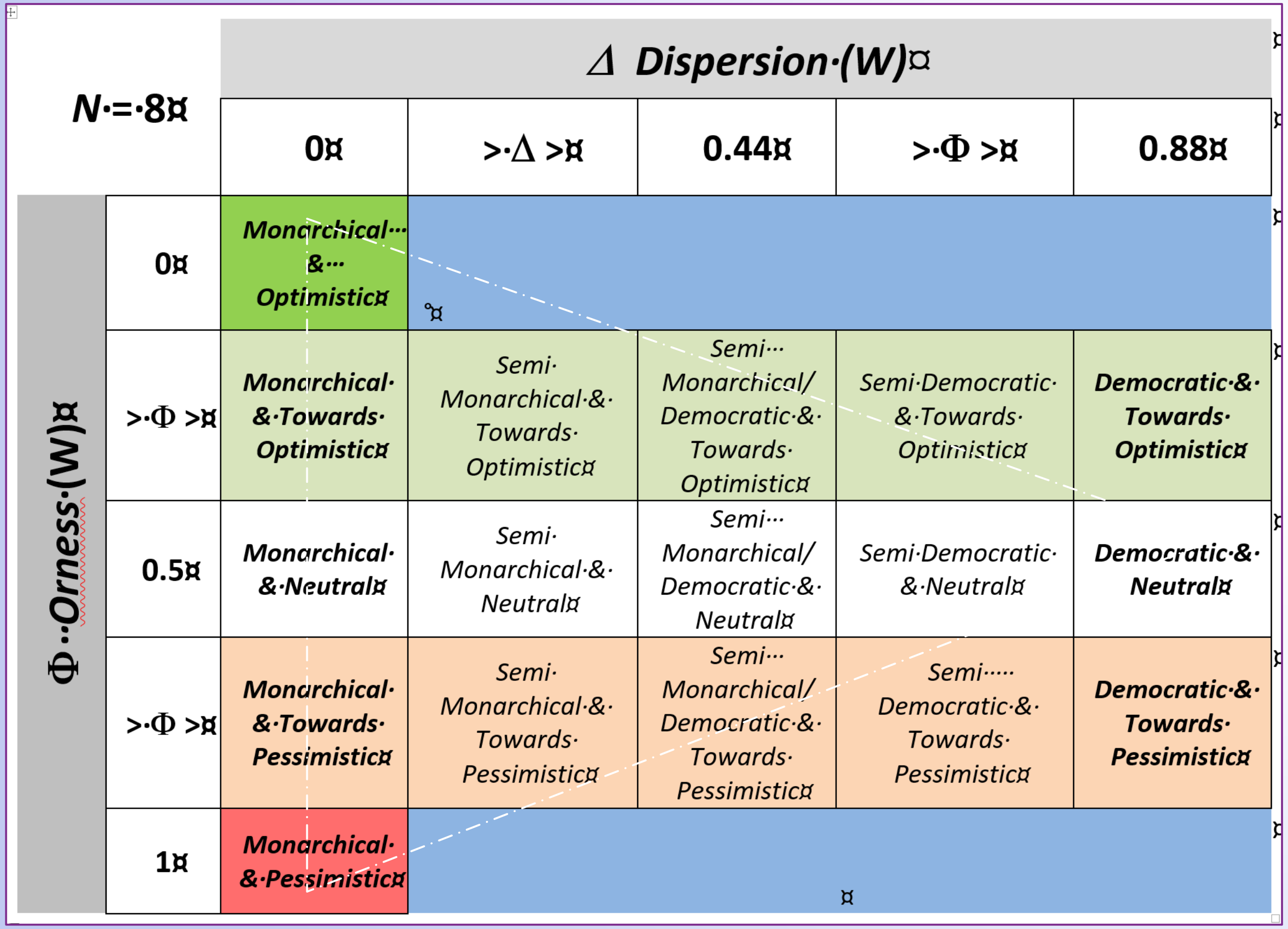
- A distinct OWA operator in each area and their dispersion and Orness
- **A synthetic standing water map**



CASE STUDY: Semantics of the learned OWA operators

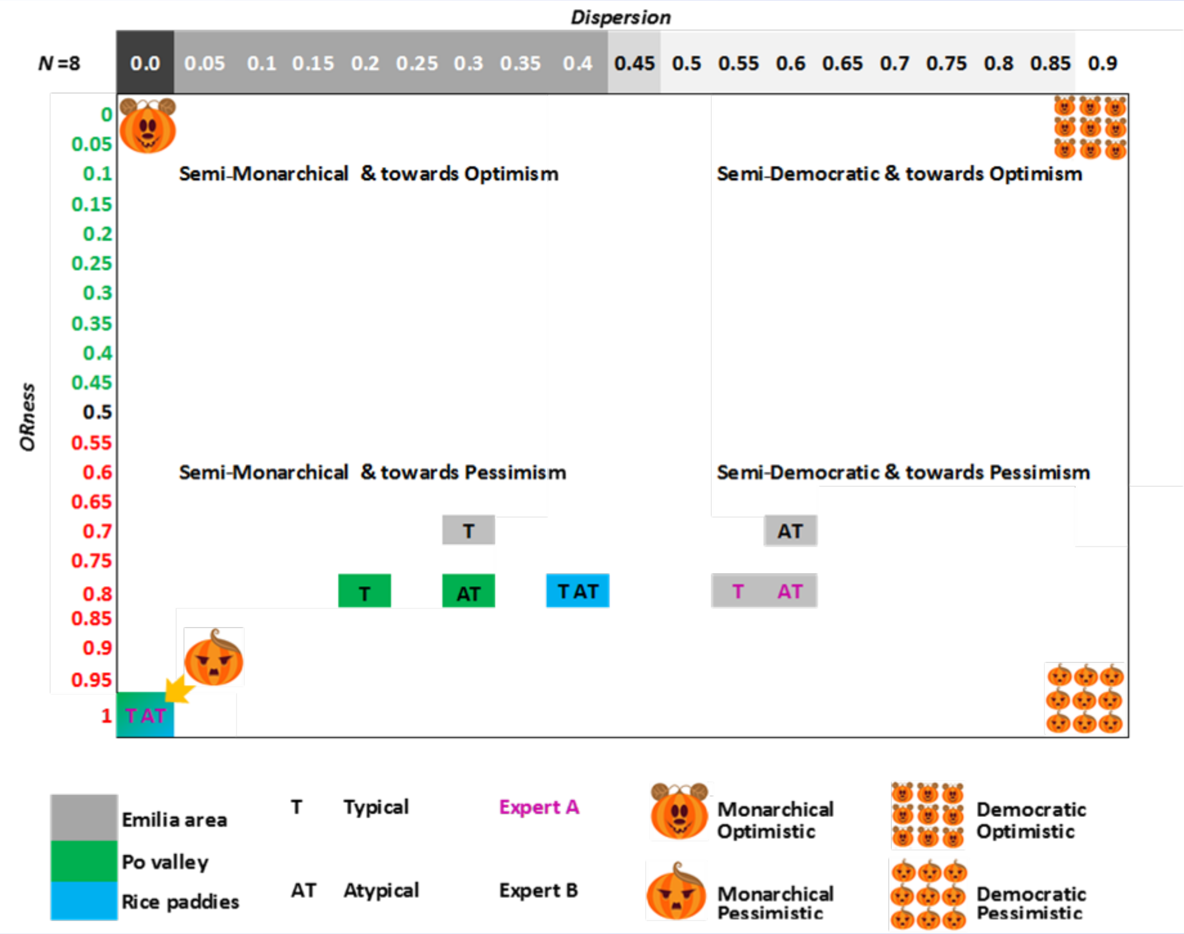
Goffi et al., , Remote Sens. 2020, 12(3), 495; https://doi.org/10.3390/rs12030495

Decision attitudes of OWA operators characterized by Orness and Dispersion



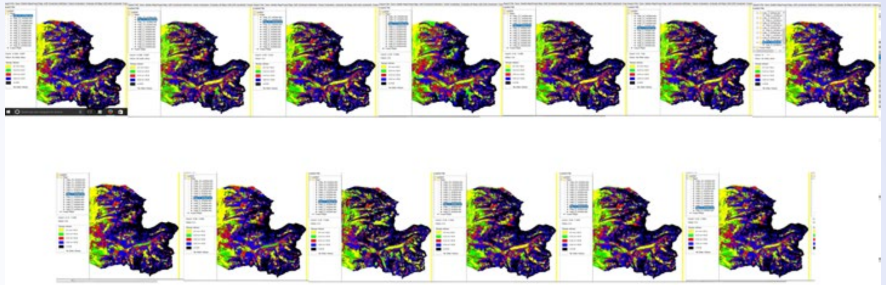
N·=·8¤		*Δ Dispersion·(W)¤*				
		0¤	>·Δ >¤	0.44¤	>·Φ >¤	0.88¤
Φ ·Orness·(W)¤	0¤	***Monarchical··· &··· Optimistic¤***	¤			
	>·Φ >¤	***Monarchical· &·Towards· Optimistic¤***	*Semi· Monarchical·&· Towards· Optimistic¤*	*Semi··· Monarchical/ Democratic·&· Towards· Optimistic¤*	*Semi·Democratic· &·Towards· Optimistic¤*	***Democratic·&· Towards· Optimistic¤***
	0.5¤	***Monarchical· &·Neutral¤***	*Semi· Monarchical·&· Neutral¤*	*Semi··· Monarchical/ Democratic·&· Neutral¤*	*Semi·Democratic· &·Neutral¤*	***Democratic·&· Neutral¤***
	>·Φ >¤	***Monarchical· &·Towards· Pessimistic¤***	*Semi· Monarchical·&· Towards· Pessimistic¤*	*Semi··· Monarchical/ Democratic·&· Towards· Pessimistic¤*	*Semi····· Democratic·&· Towards· Pessimistic¤*	***Democratic·&· Towards· Pessimistic¤***
	1¤	***Monarchical· &·Pessimistic¤***	¤			

In distinct sites different OWAs were learnt



Case study: Synthesis of models

Mapping Landslides Susceptibility with distinct reliability by aggregating results of multiple models as an ensemble approach

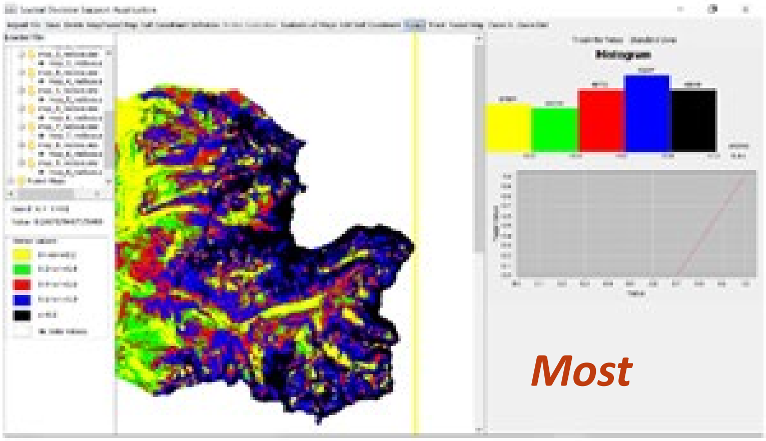


Synthesis agreed by a fuzzy majority modeling a decision maker's attitude to risk

Optimistic attitude: think positive

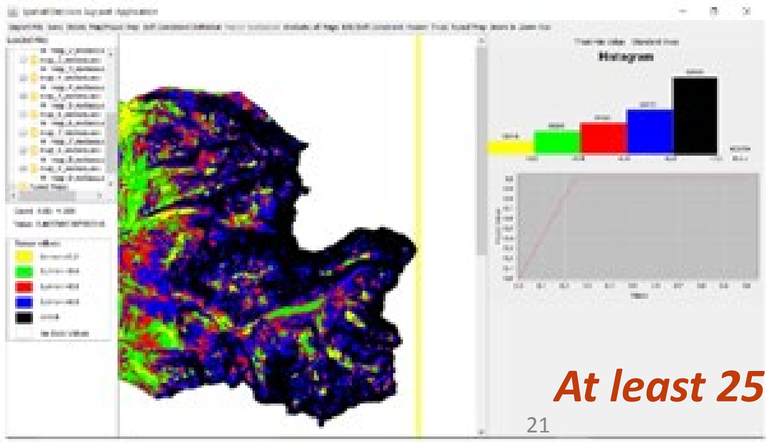
Pessimistic attitude: think negative

Optimistic Fusion



Most

Pessimistic Fusion



At least 25%

Case study: Synthesis of periodic/episodic events reported in Social Networks

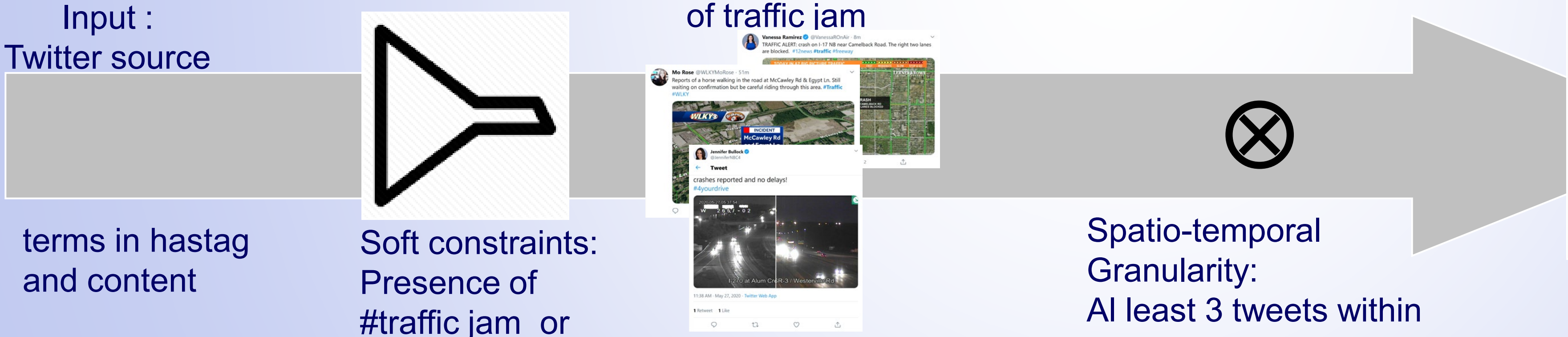
[Paolo Arcaini, Gloria Bordogna, Dino Ienco, Simone Sterlacchini, User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks. Inf. Sci. 340-341: 122-143 (2016)]

- ✓ *Traffic jams*
- ✓ *Sport, festival, music meetings, etc.*
- ✓ *Political elections*
- ✓ *Natural Disasters*
- ✓ …

Contributing factors' definition

STEP1
Partial Evidence computation

STEP 2
Synthesis by spatio-temporal density based aggregation



Input :
Twitter source

•Tweets with Partial evidence of traffic jam

Vanessa Ramirez @VanessaROnAir · 8m
TRAFFIC ALERT: crash on I-17 NB near Camelback Road. The right two lanes are blocked. #12news #traffic #freeway

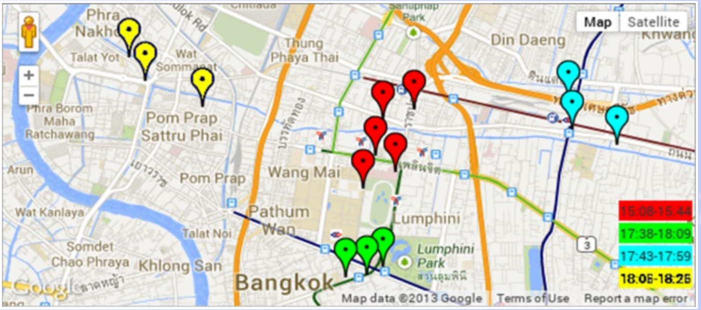
Mo Rose @WLKYMoRose · 51m
Reports of a horse walking in the road at McCawley Rd & Egypt Ln. Still waiting on confirmation but be careful riding through this area. #Traffic #WLKY

Jennifer Bullock @JenniferNBC4
Tweet
crashes reported and no delays!
#4yourdrive
11:38 AM · May 27, 2020 · Twitter Web App
1 Retweet 1 Like

terms in hastag and content

Soft constraints:
Presence of
#traffic jam or
#ingorgo stradale or
# engarrafamento

Spatio-temporal Granularity:
Al least 3 tweets within 0.5km every day



Most frequent Time periods and locations of daily recurring traffic jams in Bangkok during 6-12 2013 from Twitter

Conclusions

Managing **Geo Big Data** call for **flexible and transparent synthesis** capable to model their **veracity** and **decision makers' needs**

Soft computing offers a suitable frameworks to define solutions both **Knowledge & data driven and explainable**

Win-Win solutions since allow several levels of **flexibility**:

- **encode ill-defined knowledge** and **ill-defined decision needs**
- **adapt to local conditions** by exploiting **small ground truth data**
- provide **human interpretability** of the criteria and results

Thanks for your attention!

bordogna.g@irea.cnr.it