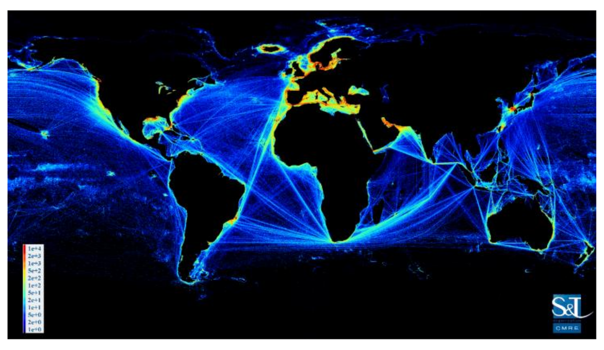
Ontology-based Design of Experiments on Big Data Solutions

Elena Camossi and Maximilian Zocholl
NATO STO CMRE - Science & Technology Organization, Centre for Maritime Research and Experimentation, La Spezia, Italy

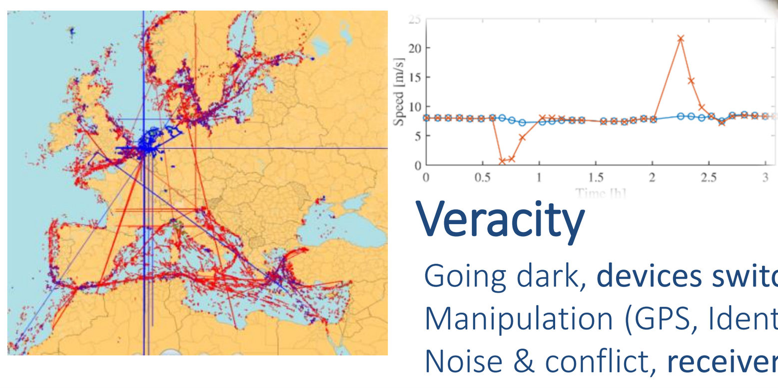
Maritime Situational Awareness

MSA Big Data challenges



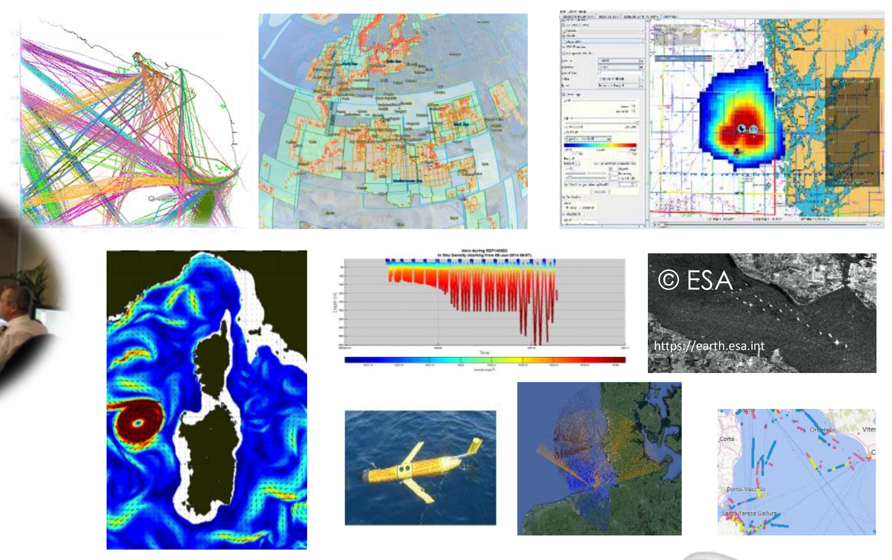
Velocity
AIS-A reporting rate:
2-10 seconds
(3 min. anchor)
256 bits/26.67 ms
(SOTDMA protocol)

Volume
EU: 19 M messages/24h from 80 600 vessels
(18,7 M AIS, 48 K VMS, 39 K LRIT) [EMSA, 2016]
W: 36,146,407 AIS messages /24h [IMISG, 2017]



Veracity
Going dark, devices switched-off
Manipulation (GPS, Identity, Destination…)
Noise & conflict, receivers limited coverage

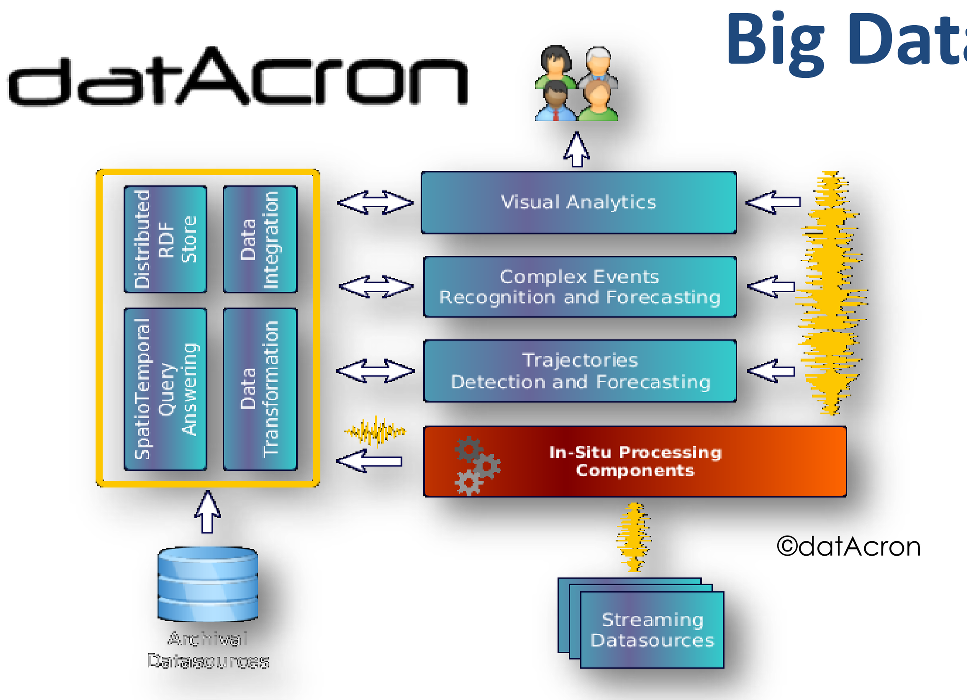
Variety
Heterogeneous Data Integration
Historical & aggregated data, geographical & environmental data, contextual data, meta information
Diverse device types, languages and formats, levels of processing, imperfection types



Semantics
Interoperability and information exchange between nations
Human understanding

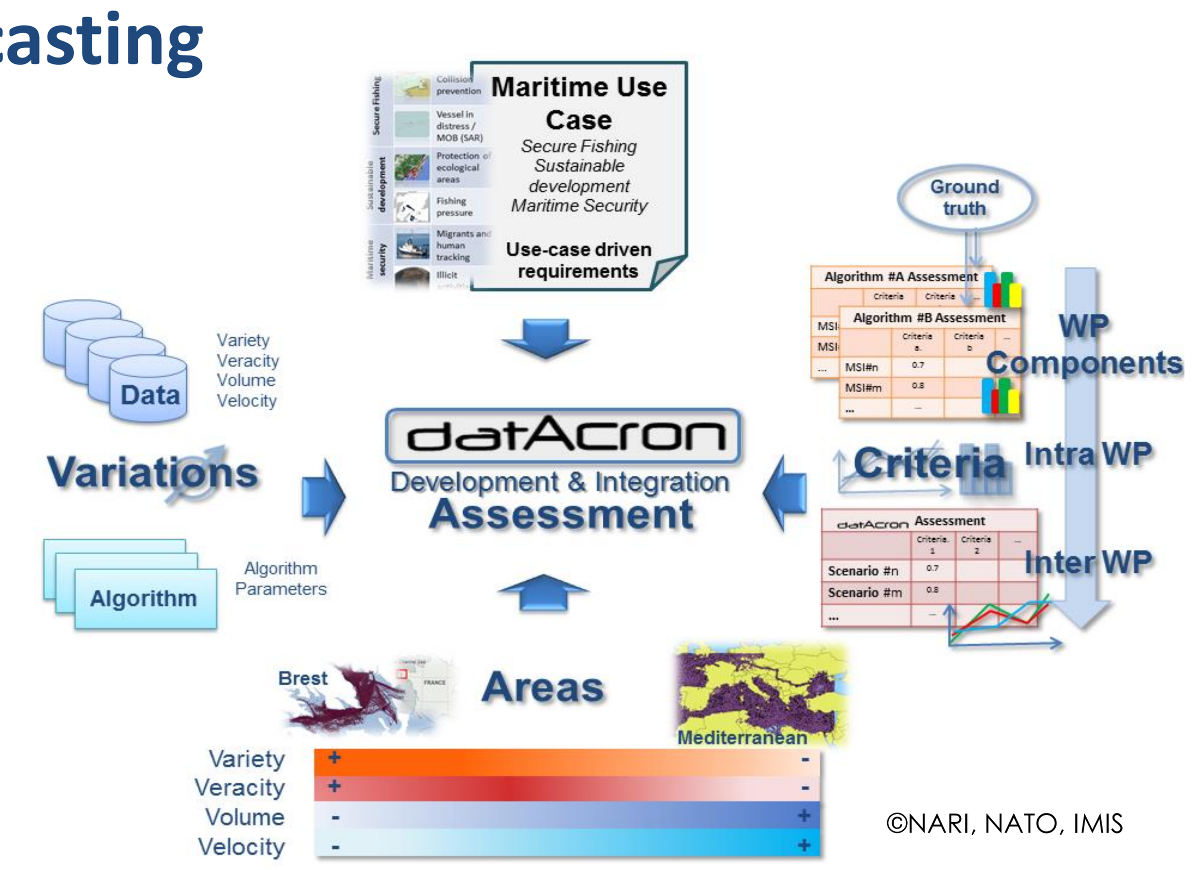
© Budyk Olena

Big Data Analytics for Time Critical Mobility Forecasting



Efficient large-scale mobility data analytics
Detect threats and abnormal activity of large fleets of moving entities in sea and air

Integration of in-situ **streaming data**
Trajectories **detection** and **forecasting**
Recognition and **forecasting** of **complex events**
Maritime Situational Indicators (MSI)
Development of **visual analytics interfaces** for maritime experts and decision-makers



Design of Experiments (DoE)

DoE is a collective of principles, statistical approaches and models for planning and performing experiments as well as analysing their results. An experimental unit is modelled as a system with input and output variables. Factors (controllable input variables) are varied according to an experimental plan that specifies the factor levels (values of the variables) of each experiment. The **goal of DoE** is to choose an experimental plan that is statistically efficient whilst allowing for an aggregation of the experimental results that are consistent with the context of the experiments, as well as to accept or reject the research hypothesis.

DoE for Big Data Solutions: Challenges

- Big data variations translate into a large number of factors with a multiplicity of possible factor levels
- The different components of the big data system implement either deterministic or non-deterministic processes and yield different output types, e.g. continuous or multinomial. The system needs to be unfolded into its components, increasing the number of necessary experiments and introducing the necessity of using different types of DoE. Choosing the wrong DoE results in a reduction of statistical efficiency or the lack of consistency of the results.

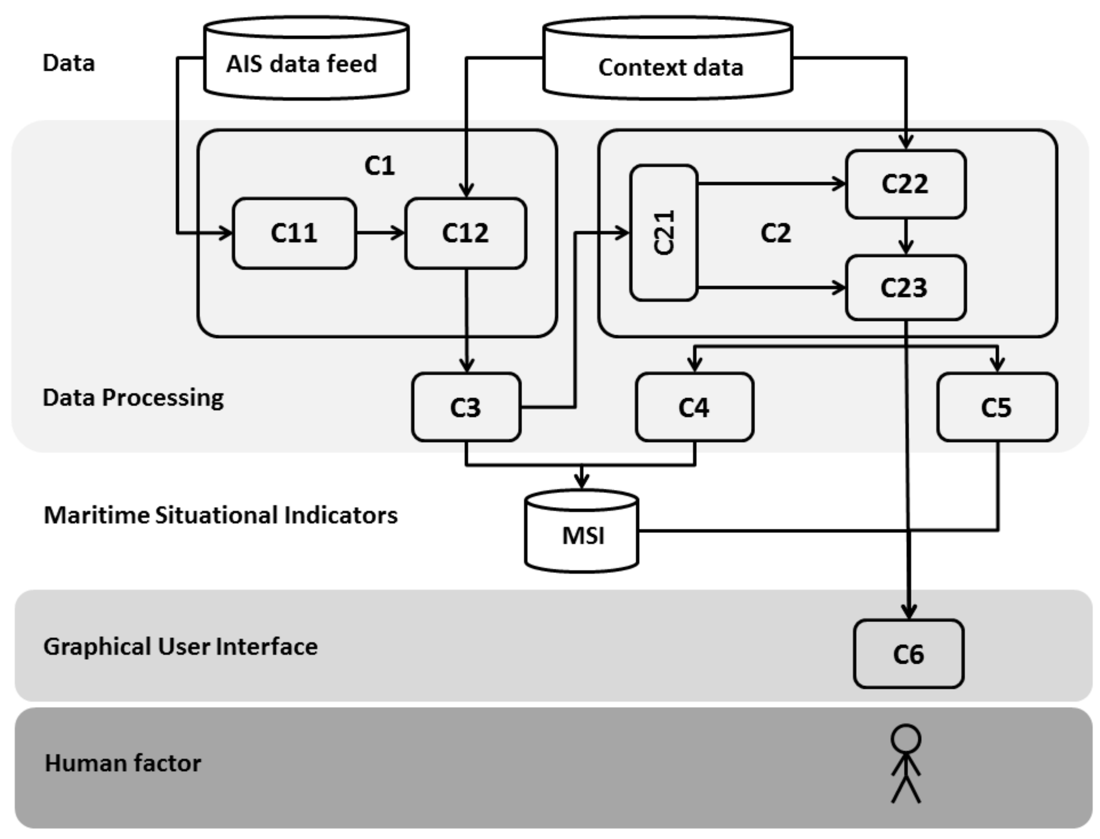
Ontology-based DoE on Big Data Solutions

Proposal and Advantages

- Semantic-wise decomposition of the big data system, supporting the roll-up of the experiment results at component level to obtain the inter-component level evaluation
- Expand the existing formalisations on DoE and leverage the domain knowledge (T-Box) to exploit domain and inter-domain specific restrictions on the factor combinations in order to select from the very large number of possible experiments a representative subset

Deterministic data processing level evaluation
- Research question: Quantify the veracity of components results, applying volume and velocity variations

Non-deterministic scenario-level evaluation
- HF and big data solution. Research question: Quantify true and false detections of the expert user using the system, applying variety and veracity variations
- HF and GUI. Research questions: Can the user assign a meaning to the symbols of the MSI? Can the user distinguish situations in which MSI fit or not the AIS data? Can the user interpret a situation represented by multiple MSI? Can the user distinguish and prioritise situations and assign criticality estimates?



$$Completeness_{EvaluatedConnectedComponents} = \bigcap_{i=1}^{n} C_i$$

$$Throughput_{EvaluatedConnectedComponent} = Min(Throughput_{C_i} \mid C_i \in System)$$

DoE	Automatic assignment	Consistency check
DoEWithReplication	Yes	Yes
DoEWithoutReplication	No	Yes
DoEWithBlocking	Yes	Yes
DoEWithoutBlocking	No	Yes
DoEWithRandomization	Yes	Yes
DoEWithoutRandomization	No	Yes

OWL2 Axioms for modelling the choice of DoE

$$DoEWithoutReplication \sqsubseteq DoE \sqcap \exists hasExpUnit.Deterministic$$

$$\boldsymbol{DoEWithReplication \equiv DoE \sqcap \exists hasExpUnit.NonDeterministic}$$

$$NonDeterministic \equiv \exists hasComponent.NonDeterministic$$

$$\boldsymbol{DoEWithBloking \equiv DoE \sqcap \exists hasNuisanceFactor.Controllable}$$

$$DoEWithRandomization \equiv DoE \sqcap \exists hasNuisanceFactor.Uncontrollable$$

- With increasing number of restrictions (e.g., max number of experiments per user), Optimal Design offers locally optimal Designs
- If only main effects are of interest, full factorial design may be used but the number of experiments are large
- Plackett-Burman reduces the number of experiments but doesn't allow for a separate estimation of main effects and their interactions

References
- M. Zocholl, E. Camossi, A-L Jousselme, and C. Ray: **Ontology-based Design of Experiments on Big Data Solutions** , *Submitted to Semantics 2018*, Vienna, Austria, 10-13 Sept 2018, P&D Session
- E. Camossi and A-L Jousselme: **Information and Source Quality Ontology in Support to Maritime Situational Awareness**, *to be presented at FUSION 2018*, Cambridge, UK, 10-13 July 2018

GA No. 687591
www.datacron-project.eu