International Journal of
Geo-Information

MDPI

Review

Review of Forty Years of Technological Changes in Geomatics toward the Big Data Paradigm

Robert Jeansoulin

LIGM UMR8049, Univ. Paris-Est, CNRS, 77454 Marne-la-Vallée, France; robert.jeansoulin@u-pem.fr;
Tel.: +33-1-6095-7743

Academic Editor: Wolfgang Kainz
Received: 29 April 2016; Accepted: 19 August 2016; Published: 29 August 2016

Abstract: Looking back at the last four decades, the technologies that have been developed for Earth observation and mapping can shed a light on the technologies that are trending today and on their challenges. Forty years ago, the first digital pictures decided the fate of remote sensing, photogrammetric engineering, GIS, or, for short: of geomatics. This sudden wave of volumes of data triggered the research in fields that Big Data is plowing today: this paper will examine this transition. First, a rapid survey of the technology through the succession of selected terms, will help identify two main periods in the last four decades. Spatial information appears in 1970 with the preparation of Landsat, and Big Data appears in 2010. The method for exploring geomatics' contribution to Big Data, is to examine each of the "Vs" that are used today to characterize the latter: volume, velocity, variety, visualization, value, veracity, validity, and variability. Geomatics has been confronted to each of these facets during the period. The discussion compares the answers offered early by geomatics, with the situation in Big Data today. Over a very large range of issues, from signal processing to the semantics of information, geomatics has made contributions to many data models and algorithms. Big Data now enables geographic information to be disseminated much more widely, and to benefit from new information sources, expanding through the Internet of Things towards a future Digital Earth. Some of the lessons learned during the four decades of geomatics can also be lessons for Big Data today, and for the future of geomatics.

Keywords: geomatics; Big Data; remote sensing; data warehouse; data mining; technology history

1. Introduction

Besides being a buzzword, Big Data reveals the way technologies spring off, evolve, merge together, replace older ones, or bring new life to forgotten ones. In addition, it enlightens the new context in which these modifications are emerging. Let us listen to what professionals are expecting from Big Data: the IBM website (May 2016) states that "*Big Data technology must support search, development, governance and analytics services for all data types—from transaction and application data to machine and sensor data to social, image and geospatial data, and more.*"

However, how was life before Big Data?

This paper reviews the principal issues that the geo-information science (among terms: *geoinformation*, *geospatial data science*, *geomatics*, etc. (all ±1990), we preferably use the latter.) has confronted since its early stages. This retrospective is conducted under the light shed by the Big Data vocabulary: we invoke the popular three Vs (volume, velocity and variety) and additional Vs often suggested (value, validity, veracity, variability, and, occasionally, vulnerability and visualization).

With the 1972 Landsat pictures came the "data management challenges": how to deal with such huge volumes of data; how to bring structure to a two-dimension sampled signal, relating it with terrain information; and how to process such huge data sets in a realistic amount of time? Soon after arrived the data analysis challenge. Geographic information is about the "real world." What is it, really? What is measured by a pixel radiometry, or what is delineated by a list of coordinates? The results provided by image classification algorithms, or by spatial aggregation/disaggregation models, since the late 1970s, brought great added value to the collected data (value), but a value pervaded by an inherent uncertainty, impeding their direct use in decision making (veracity). In addition, when merging different sources of data, which happens more and more, there are new questions. Is merging semantically relevant? Getting at validity, is it syntactically consistent? Finally, the use of so much geo-data, piled up for several decades now all around the globe, requires clarifications that address variability. What are these data the data? What real evolution or differences do they measure? What kind of decision can we built upon them?

In the context of new data, these longstanding questions continue to be raised today, plus some new issues as well (e.g., geo-location and privacy, confidence on data, and legal liability).This paper examines what aspects have been addressed and pioneered by geomatics, which have been ignored, and what lessons, if any, can be learned from these four decades.

2. Four Decades of Geomatics Revisited through the Vs of Big Data

2.1. Volume: Storage and Numerical Processing Requirements

2.1.1. Data, Spatial Data, Storage, Access and Analysis: A Retrospective

Let us look back, 40 years ago, with a linguistic eye, using the Google tool *Ngram Viewer* [1]: it counts occurrences of terms in general literature and shows quite accurately when a term became popular. The science behind the term has generally been published in scientific papers two to five years earlier, but *Ngram* is a marker of the relevance of the technology described by the term. We have benchmarked several groups of terms with this tool (Figure 1):

- Precursor tools for reasoning and image processing: data analysis 1958, pattern recognition 1958, artificial intelligence 1961, principal component analysis 1960, correspondence analysis 1975, image processing 1965, image understanding 1970, and machine learning 1981;
- Data and corporate knowledge: databases 1965–1970, data warehouse 1988–1992, metadata 1991, OLAP 1994, data mining 1995, business intelligence 1996, and analytics 2004;
- Spatial information data, devices, and tools: satellite imagery or remote sensing 1970, Landsat 1972, GPS 1975, and spatial data quality 1990; and
- Internet tools: Internet 1990, email 1992, browser 1992, web and website 1994 (Note: "web" and Internet, are not shown—out of scale versus the other terms—"browser" is a good proxy).

Alas, while still accessible, Ngram stopped including books in 2008 so the terms "cloud computing" and "Big Data", are too "young", and absent in Ngram.

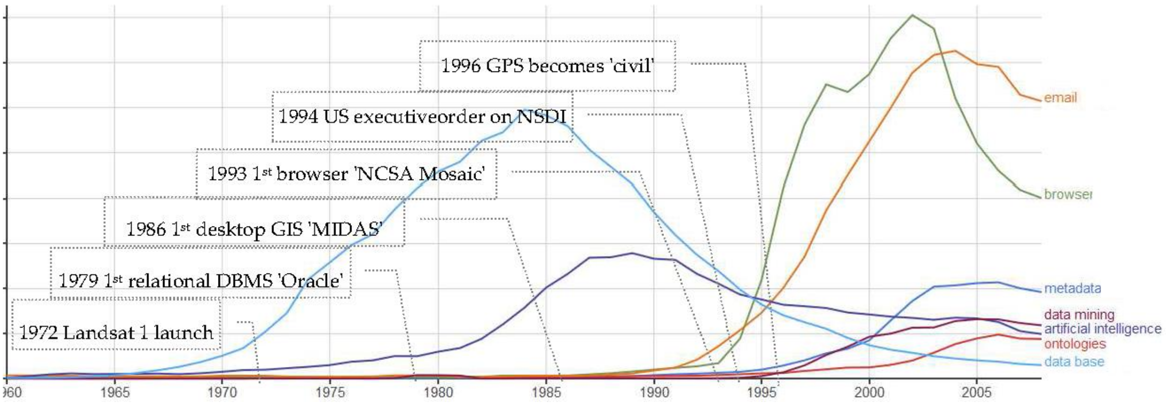


Figure 1. Data engineering: Major events and major buzzwords in the last 50 years.

What lessons can we learn from this historical retrospective?

- Big Data's birth year is around 2010, while spatial data traces back to 1970, before the Landsat data;
- Database and AI tools are at their apogee between 1975 and 1985, extensively used for processing the phenomenal amount of data harvested by satellites: a Petabyte over the decade, at a time when central memory of big computers was limited to a few Megabytes;
- 1995 emerges as a **tipping point**, with the widespread use of the Internet tools, and with data mining, analytics, ontologies taking over databases and AI, whose buzz is somewhat fading. This tipping point marks an important paradigm shift, following Thomas Kuhn's "Structure of scientific revolutions". It does not mark a discovery (e.g., the transistor in 1947) or a technological success (e.g., Landsat launch in 1972) but a collective consciousness (Term used by social theorists like Durkheim, Althusser, and Jung to explicate how patterns of commonality emerge among large groups of autonomous individuals.) that all the technologies related to computation (algorithms, personal devices, network, etc.), and to information (data collections, surveys, literature, news, etc.) should merge together.

2.1.2. Geospatial Data = Huge Data before Big Data

Forty years ago there were no smartphones, no GPS, and no email. Computers were alphanumeric consoles, but satellites started to flood us with Earth imagery. Suddenly, data were there, huge sets of pixels, which were difficult to turn into pictures (no screens, only alphanumeric consoles, and images had to be printed on costly devices): it forced scientists of the time to crunch the data without the chance to view them! The amount of data for a single coverage of the planet by Landsat is phenomenal: one Terabyte. One Petabyte, just of geospatial data, was reached around 1980. In 2010, the amount of data stored for the entire planet reached one Zettabyte [2]. This avalanche of pixels was followed by the digitization of many other data sharing the same fundamental characteristic: they were related to Earth. Geographic information was born. In April 1994, US President Clinton signed the Executive Order: *"Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure" (NSDI)*. This NSDI helped to assemble geographic data nationwide, reducing operating costs, and improving data service and decision-making.

2.2. Variety and Velocity: Data Structure before Unstructured Data

2.2.1. From Ancillary Data, to Metadata

Telecommunication satellites opened space to economic exploitation. Soon after, weather and Earth-observation satellites were the next opportunity. Remote sensing, at first a mere signal processing (sampling and calibration), was confronted with the necessity to link images with terrain coordinates, or with other images, due to the rapid spread of new sensors: therefore inaugurating the use of metadata, even though the first term was "ancillary data". The libraries, confronted with cataloguing issues in the 1960s, developed a markup language *MARC*, to which HTML owes greatly. Then, since the *Dublin Core* in 1995, the prefix "meta" marks a better consideration of what was the first technical way to translate data relationships in computer language. It started with the first Landsat pixel in 1972: a picture-cell (pixel) is the approximation of a ground surface, of which we measure the reflected (or emitted) sampled signal within a wavelength range, integrating diffraction effects, absorption, etc. The purpose was to improve data from a "raw" status, to a status of "corrected data." Pixels are not processed one by one, but as a statistical variable that is assigned to a class, to properties (e.g., a border pixel), which must be described into some additional information. A "processed image" has a lot of such information. Here are two examples:

- Earth observation: The program *"CORINE"* (*"Coordination of Information on the Environment"*: program initiated in 1985 by the European Commission.), which monitors the change in land use in Europe, by updating data from the classes extracted from satellite images, uses metadata extensively [3].

- Automated cartography: How do you make explicit the topology in data vector representation? Topology is implicit, in a correct geometry, but the burden of re-computing it every time is much too heavy. Therefore, in the 1990s, several NGIs were working on what eventually became the *ISO 19101: GI Reference model* (the underlying concept is the polygon-arc-node model, with all topology relationships), as in Table 1, which says that parcels "2" and "3" are adjacent, and their union forms a single hole into parcel "1." Themes and rules can be added: metadata are mandatory for deciphering all that information.

Table 1. Encoding the geometry and the topology of a set of land parcels.

#	Coordinates (or Vertices)	Contains	is in	Touches	Has Hole(s)	Theme, etc.
1	x,y; x,y; x,y;x,y;	**#2;#3**	-	-	1	-
2	x,y; x,y; x,y; x,y	-	**#1**	**#3**	0	-
3	x,y; x,y; x,y	-	**#1**	**#2**	0	-

Therefore, databases and signal processing had to join at some point anyway.

2.2.2. From Data Columns to Data Cubes

The relational model was one major breakthrough in data engineering. Based on a solid mathematical background, it helped to build reliable "transactional" systems, which ignited e-commerce. The archetype of the relational model is the spreadsheet, still very popular in offices.

A new trend emerged, for "analytical" purposes, giving birth in the early 1990s to the OLAP systems used in "Data warehouses" by most major companies, on the promise to discover customers' hidden behavior. The archetype is the "cube" (see Figure 2), based on concepts of: (*a*) "dimensions", along a certain hierarchy (e.g., year, month, day), which can be a partial order (e.g., weeks and months); (*b*) "measures", values recorded for various dimensions (e.g., cancer rates); and (*c*) "facts", any combination of dimensions and measures. Their aggregation forms a data cube [4]. Two dimensions (matrix) or three (cube) can easily be viewed, but more (hypercube) cannot.

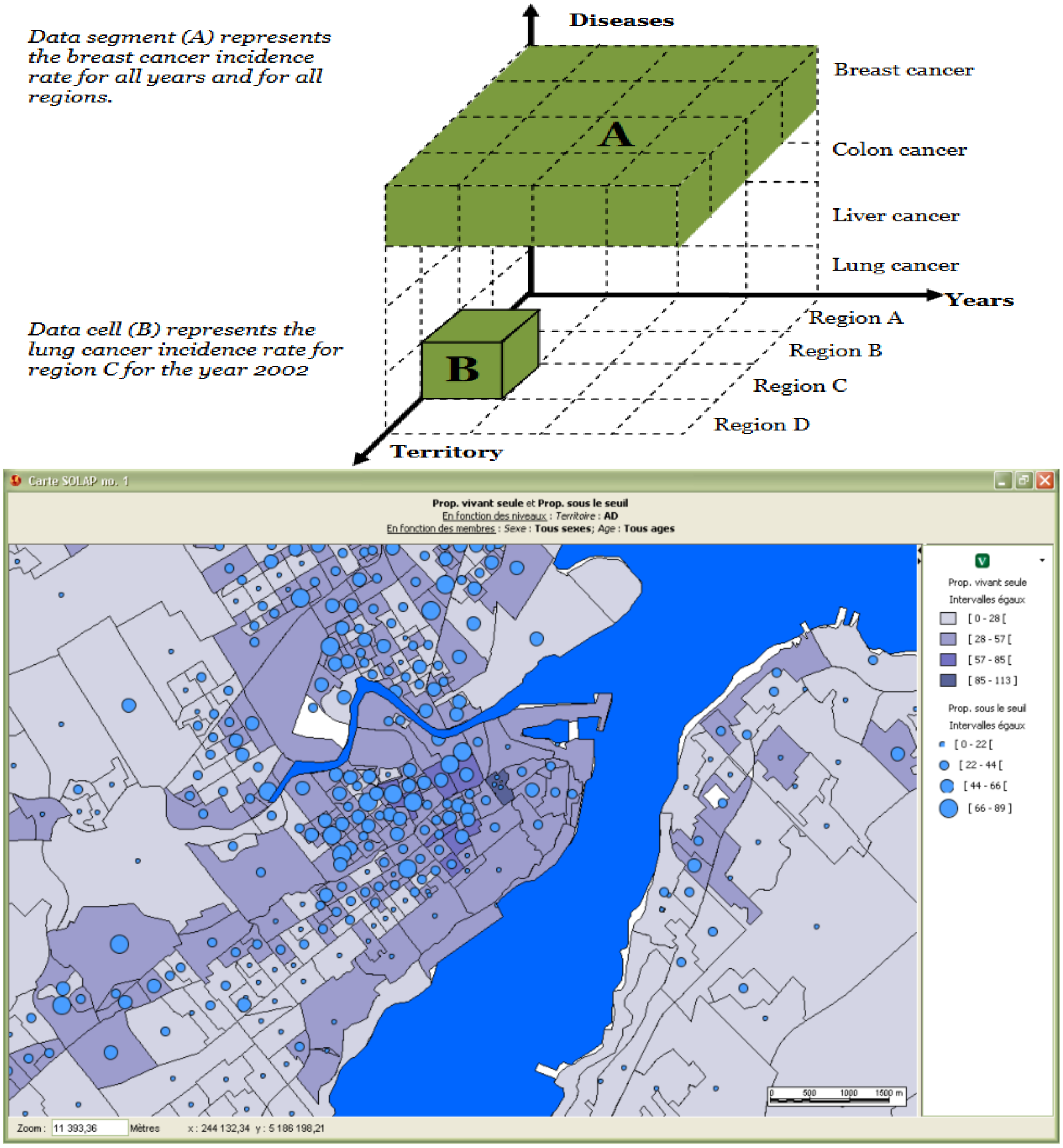
The geographical nature of most data has long been noted: *"[...] approximately 80% of the informational needs of a local government policymaker is related to a geographical location."* (the "80%" has since been largely cited and rarely disputed). This statement is traced back [5] to a 1987 paper by Williams. Data warehouses were facing the issue of managing the spatial data, adding new dimensions to the "thematic" ones: descriptive (e.g., "Quebec"), or geometric spatial dimensions (e.g., vector contour of Quebec), facing again the issue of mixing two or several very different representations of space [6].

Geometry does not match well with tables: if the "info" part of the geographic information can be stored in RDBMS, the "arc" (coordinates) part cannot, nor images (pixels). During two decades, geo-information was confined into the rather closed market of GIS. (a few companies were selling GIS: ESRI, Intergraph, MapInfo, etc. often as hardware-software packages). By the end of the 1990s, several pioneers proposed to combine OLAP and spatial databases: the prototype *GeoMiner* at Simon Fraser [7], the combination GIS–OLAP, leading to SOLAP (the term Spatial OLAP or SOLAP was introduced by Bédard (1997) in reference to the term spatial database) systems, at Université Laval [8].

Experiments were piloted on complex applications, in public transportation or public health. For instance, the integration of geo-referenced indicators**,** for the surveillance of the impacts of climate change [9], provided new means for exploring data at different scales, different regions, and eras, and for visualizing the results in synchronized maps, tables, and charts. Its relevance was confirmed by the project's end-users in the surveillance community. Figure 2 illustrates how to combine the proportion (class) of the population living alone (choropleth map), and the proportion of the population with low income (symbols), translated at the same spatial level.

2.2.3. Streaming, Parallelism, and Pre-Processing in Geomatics

The velocity aspect of this period was multiform. The huge size of remote sensing images confined them to magnetic tapes. It was impossible to load them in main memory: the algorithms developed were based on concepts of splitting, streaming, and parallelizing, as soon as the first so-called vector processors were available (late 1970s). With the OLAP systems, the bottleneck is the time consuming "join" operation, in relational DBMS: the answer is to pre-compute all possible joins—or as many as possible—in order to provide a fast answer for interactive visualization. A drawback is that pre-computation must be performed again after any insert/delete in the database, which required developing efficient methods for the materialization of spatial cubes [10].


Data segment (A) represents the breast cancer incidence rate for all years and for all regions.

Data cell (B) represents the lung cancer incidence rate for region C for the year 2002

Carte SOLAP no. 1

Prop. vivant seule et **Prop. sous le seuil**
En fonction des niveaux : *Territoire* : **AD**
En fonction des membres : *Sexe* : **Tous sexes**; *Age* : **Tous ages**

Zoom : 11 393,36 Mètres x : 244 132,34 y : 5 186 198,21

Figure 2. (**Top**) Multidimensional cube: two examples of incidence rates by territory, time, and diseases. (**bottom**) Combining facts: people living alone (color map), and people with low income (symbols).

However, the velocity aspect has never been addressed in geomatics under the same constraints than those of Big Data today. Velocity in geomatics was an issue: (a) in the processing of high volumes of data; and (b) for interactive visualization, as seen above with data-cubes. At the time, AOL-Mapquest inaugurated Internet mapping, and Google maps were first released, for the United States, in 2004. Then, the largest national mapping agencies launched their own public access portals: Ordnance Survey, IGN, USGS, Geomatics Canada, etc. The network-based environment was not a familiar or friendly one for these agencies, and they had to face problems they were not prepared for, such as time delays and data packet dropout [11]. The resolution of these problems was externalized; as a consequence, the dissemination, and eventually, the market for geographic information fell into the hands of the main players of Big Data.

In addition, the extensive use of video in the Internet nowadays has triggered development of much more efficient streaming algorithms, which could later enable the use of new instruments in geomatics (e.g., drone video and surveillance cameras).

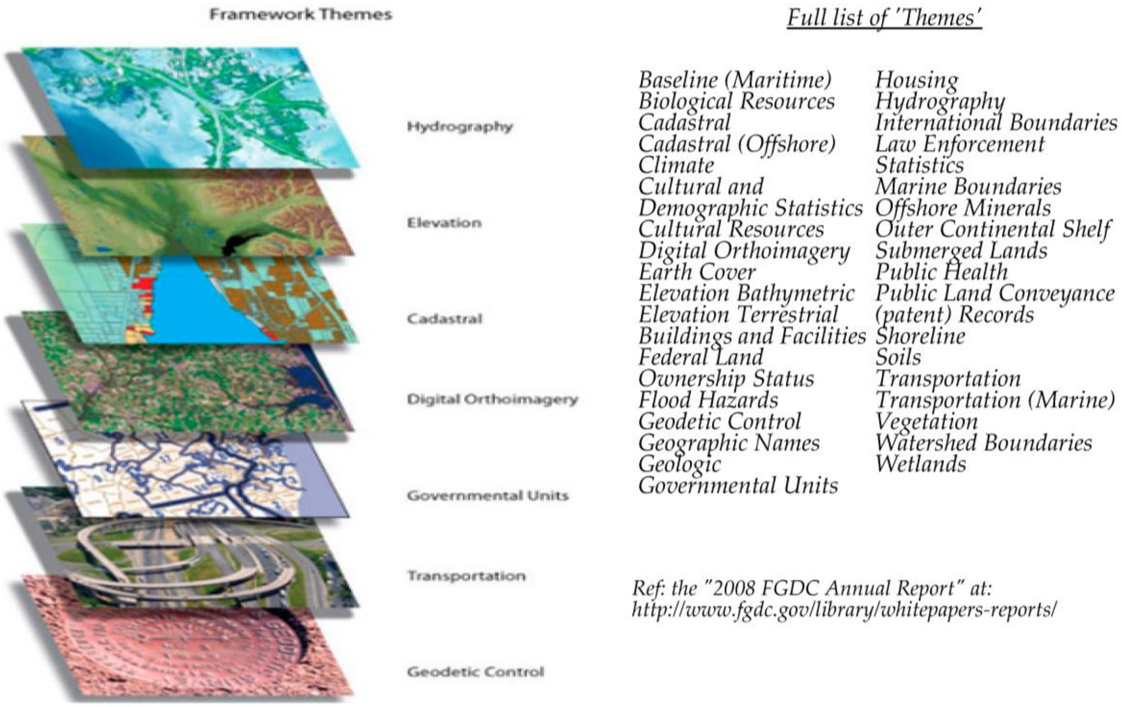
2.3. Value, Validity, Veracity and Variability: Turning Data into Knowledge

2.3.1. Value, Added by Data Processing: From Data Analysis to Data Mining

Since the early 20th century (Nyquist (1928) and Shannon (1949) are among the renowned authors in transmission theory), signal processing has been a field of research in statistical, stochastic, or data transformation approaches, and also for intensive computational methods and efficient algorithms. The Cooley-Tukey algorithm for FFT was published in 1965. Exploratory data analysis (EDA), published in 1977 [12], is linked to Fourier analysis. The term may seem outdated, but EDA contributed to fostering research in high-end computing. For example, let us recall the hierarchical algorithms for pixel classification, used with Landsat images [13], or more sophisticated approaches for unsupervised classification, such as the "dynamic clusters" [14].

Alpilles-ReSeDA (remote sensing data assimilation), a consortium of 10 European partners over the period 1995–1998, focusing on soil and vegetation monitoring at different scales [15], was among the first international efforts to achieve a large scale merging of multi-sensors data (as in Figure 3): satellite images (visible, infra-red, radar), soil samples, weather records, agricultural surveys. Up to six dates and 14 wavelengths, were collected over the whole test site (25 km^2), and made exploitable after pixel registration, albedo and other calibrations. The LAI (leaf area index), a normalized ratio between a green and a near-infrared wavelength, is very discriminant for vegetation, at any date, and between kinds of vegetation at different dates [16]. This discriminative power is demonstrated by the principal component analysis of 6 dates of LAI.

Building classes from raster data (clustering), is computed in the vector space of a few dates of LAI (a dozen of kilo-octets). Assigning a class to the hundreds of thousands of pixels is performed by a simple look-up table. This dissymmetry in cardinality led remote-sensing specialists to use the "divisive", rather than the "agglomerative hierarchical" clustering. It starts with the whole set, then works top-down by successive dichotomies. Top-down clustering is generally more complex than the bottom-up clustering: O(2^n), instead of O(n^2) for most agglomerative methods. However, it has the advantage of being more efficient if we do not generate a complete hierarchy, all the way down [17]. The process builds up the successive dichotomies by a k-means algorithm (k=2), starting with two initial centroids: at each step, the first centroïd is initialized from the peak value of the histograms of the least coherent cluster (the second centroïd can be the second mode of the histogram, or any far different value). The best dissimilarity is obtained by maximizing the "inter-cluster dissimilarity", or maximizing the "intra-cluster similarity". The dendrogram (Figure 4, left) is built top-down, until a relevant number of classes is reached (a few dozen). The height of the dendrogram is proportional to the intra-cluster similarity: it can be used to characterize the "quality" of the clusters obtained for a certain level (a version of this algorithm, *DIANA* –Divisive ANAlysis Clustering-, is now part of the package R).



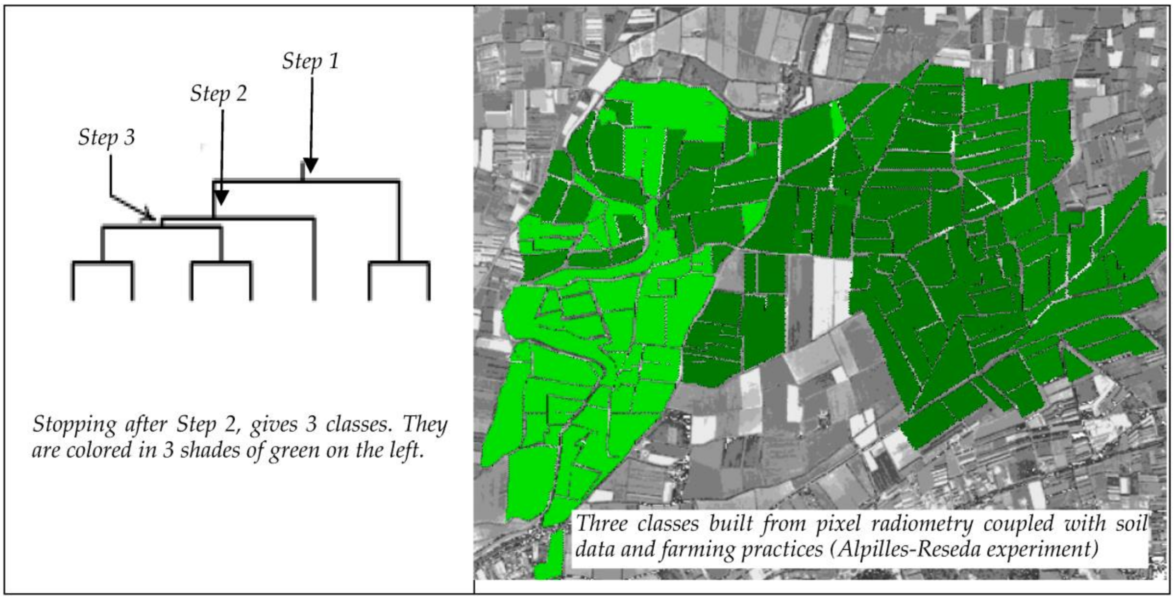
Full list of 'Themes'

Baseline (Maritime)
Biological Resources
Cadastral
Cadastral (Offshore)
Climate
Cultural and
Demographic Statistics
Cultural Resources
Digital Orthoimagery
Earth Cover
Elevation Bathymetric
Elevation Terrestrial
Buildings and Facilities
Federal Land
Ownership Status
Flood Hazards
Geodetic Control
Geographic Names
Geologic
Governmental Units

Housing
Hydrography
International Boundaries
Law Enforcement
Statistics
Marine Boundaries
Offshore Minerals
Outer Continental Shelf
Submerged Lands
Public Health
Public Land Conveyance
(patent) Records
Shoreline
Soils
Transportation
Transportation (Marine)
Vegetation
Watershed Boundaries
Wetlands

Ref: the "2008 FGDC Annual Report" at:
http://www.fgdc.gov/library/whitepapers-reports/

Figure 3. The many themes identified by the Federal Geographic Data Committee, collected by US agencies.



Stopping after Step 2, gives 3 classes. They are colored in 3 shades of green on the left.

Three classes built from pixel radiometry coupled with soil data and farming practices (Alpilles-Reseda experiment)

Figure 4. Hierarchical clustering example (sometimes referred to as phylogenetic tree).

Eventually, all of the pixels of a land parcel are classified according to the class of the average pixel value of that parcel, which allow drawing them on a map (Figure 4, right: three shades of green for the three classes).

It is interesting to mention that, in 1990, Kaufman and Rousseuw were writing "*in the literature, divisive methods have been largely ignored. (In fact, when people talk about hierarchical clustering, they often mean agglomerative clustering.)*" [17]. Furthermore, to note the recent revival (2008): "*There is evidence that divisive algorithms produce more accurate hierarchies than bottom-up algorithms in some circumstances.*

Bottom-up methods make clustering decisions based on local patterns without initially taking into account the global distribution. These early decisions cannot be undone. Top-down clustering benefits from complete information about the global distribution when making top-level partitioning decisions." [18].

We can expect now that, for several situations, the cardinality of the variable space will increase way more than the cardinality of the value space, reproducing the favorable situation of the clustering in remote-sensing: top-down clustering may revive in Big Data.

Another important term, "data mining", traces back to the mid 1990s and is familiar to database scientists as an operational approach of "machine learning." Machine learning, whose foundations are the Turing machine, is more theoretical: based on logics and lambda-calculus. Support vector machines (SVM)—the archetype of machine learning algorithms—were developed as non- probabilistic binary linear classifiers [19], and like similar tools, added unquestionable value to the analysis of many geographical data. Nowadays, these terms are more or less wrapped up in the successive buzzwords of business intelligence, data analytics, and Big Data. Despite other differences, the fact is that the mainstream shifted from signal processing to e-commerce.

2.3.2. Veracity, Data Uncertainty: From Precision to Quality Indicators

Geo-information deals with the "real world." Technically, it states that there is a single world that can be measured by different means, at different scales, from multiple points of view, but eventually everything has to be locatable and consistent (*logically*), because there is a single world.

At first, the quality of measurements was limited to precision: what particular spot on Earth does a particular pixel represent? What is the radiometric contribution of this spot to the pixel value? Image registration and sensor fusion were the most difficult tasks in the 1970s. Soon after, confidence in data classification was the big issue. How much wheat is the USSR really harvesting? The largest US computers of that time were crunching the pixels, day and night, over many weeks, to answer such a question (now, the NSA supercomputers are processing trillions of emails).

When it became easier to merge remote sensing imagery and geographical databases, more complex questions were at hand. Gains in ground resolution opens us new grounds: from the large rectangular crops of the Middle West in the first Landsat images to urban roof gardens. Now geomatics can talk science with sociology, as it did with agronomy in the 1980s.

The challenge about quality is no longer solely about data precision. Several quality indicators have been designed and the quality domain has been structured by international consensus. (the technical Committee ISO TC211 was established to publish standards for geo-information). The standards ISO 19101: "reference model", and ISO 19113: "quality principles" were issued in 2002, reflecting a common ground (see Table 2) between the various National Geographic organizations. National Statistical offices, and international bodies, such as United Nations, OECD (Org. for Economic Co-operation and Development) and Eurostat. One important outcome was precisely to distinguish between causes of uncertainty (accuracy, consistency, and naming), reflecting somehow the differences between veracity, validity and variability.

Table 2. Metadata for representing the quality of spatial data (ISO 2002).

Data Quality Element (Sub-Element)	Description
Completeness (omission, commission, logical consistency)	Presence of features, their attributes and relationships(absence or excess of data, adherence to rules of the data structure)
Conceptual consistency	Adherence to rules of the conceptual schema, to value domains, etc.
Topological consistency	Correctness of explicit topology, closeness to respective position of features
Positional accuracy	Accuracy in absolute point positioning, gridded data positioning
Temporal consistency	Accuracy of temporal attributes and their relationships
Thematic accuracy	Accuracy of quantitative attributes, class correctness

2.3.3. Validity, Data Consistency: Rational Knowledge from Uncertain Knowledge

Understanding the many causes of data uncertainty sheds light on the many approximations made throughout the process of gathering and measuring data, even the simplest datum (e.g., ground temperature).

Considering that data are always somewhat inexact and always depending on a model, which imperfectly represents a partial aspect of the reality, it is important to provide guidelines. Every time we can input a constraint, we can confront the data, and issue a warning for each detected conflict. Figure 5 below is a snapshot to an experiment developed during the European project REV!GIS (Revision in GIS: 5th Framework Program project, involving the universities Keele, Laval, Leicester, Marseilles, Pisa, TUW Vienna, Twente ITC): to revise uncertain flood data using directions of flow as constraints [20]. The top-left and bottom-left diagrams demonstrate how two models, for two adjacent spatial zones, can be independently logically-consistent, but can result into an inconsistent model when merged. By using flow direction and constraints, the min and max estimations of the water height can be improved (shrinking the intervals), or sometimes enlarged for re-establishing global consistency. The process is computationally expensive: such AI algorithms are named NP-hard (for non-polynomial): artificial intelligence may bring solutions and may raise new problems as well.

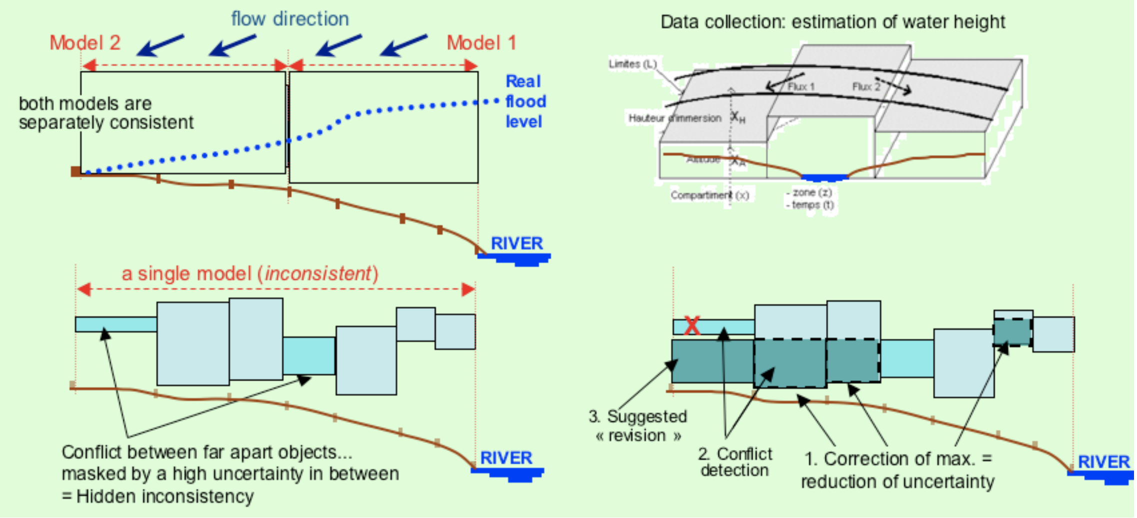

flow direction
Model 2
Model 1
both models are separately consistent
Real flood level
RIVER
Data collection: estimation of water height
Limites (L)
Flux 1
Flux 2
Hauteur d'immersion X_H
Altitude X_A
Compartiment (x)
- zone (z)
- temps (t)
a single model (*inconsistent*)
Conflict between far apart objects...
masked by a high uncertainty in between
= Hidden inconsistency
RIVER
X
3. Suggested « revision »
2. Conflict detection
1. Correction of max. = reduction of uncertainty
RIVER

Figure 5. Estimation of a flooding event, combining two sources of data: flow direction (**top left**) and height (piecewise estimations: **top right**). The bottom line sketches how inconsistencies can be detected between local models (**bottom left**), and then corrected into a unified model (**bottom right**).

This example illustrates circumstances in which there is a necessity to merge (or for "fusion": this term is more popular in the knowledge representation and reasoning community) quantitative information (e.g., direct measurements or pixel analysis) and qualitative information (e.g., domain specific or expert rules, constraints, known impossibilities, or exceptions). This is probably one of the next challenges for Big Data analytics.

In 2008, the publication of "flu trends" by Google looking through Internet queries received special media attention. In 2011, the IBM Watson computer defeated two "Jeopardy" champions: the *DeepQA project* behind Watson is making intense use of geo-information reasoning for answering questions such as "*They're the two states you could be reentering if you're crossing Florida's northern border*" [21]. Developed in the 1980s, Allen's interval algebra, "Mereotopology", or "Region connectedness calculus" RCC algebras [22,23], are precisely used by IBM Watson for constraining queries and making answers more narrow and efficient.

Reasoning under constraints marries well with stochastic reasoning, and Bayes networks algorithms have been applied successfully for developing spatial simulation [24]. For instance, following a three-step approach: (a) an initial graph—a Bayes network—is derived from a set of parameters, possibly correlated by causal relationships, and observed over a territory, at two consecutive years; (b) additional constraints (e.g., geophysical rules) are used to "improve" the computed graph with some a priori expert knowledge; and, finally, (c) this improved graph, fed with values gathered at a new date, will deliver a prediction of what should happen the next year. This approach, which can remind us the "expert systems" of the 1970s, is gaining some popularity with the rise of business analytics. Microsoft has implemented similar approaches in its MS Naive Bayes Algorithm, that the company describes as "useful for quickly generating mining models to discover relationships between input columns and predictable columns."

2.4. Ontologies and Variability: Data Are Acts, Not Facts

Geographers are still classifying terrain features, still zoning land, but much more attention is turned to the meaning of the process, the interpretability of a result, and the suitability for decision-making. Geomatics also raised the question of *what* is in data. Common agreement, data quality and usability are just some of the different aspects of what is often summarized as "what is your ontology?" and the subsequent problem of "ontology alignment" [25,26].

The term "ontology" has been brought to public attention by the introduction of the concept of Semantic web (using markups as an aid for Internet robots to establish better relationships between web fragments). In geomatics, research on ontologies developed because of the extreme variability of naming and representing geographical objects. The question "what is a forest?" is a symptomatic example of the variability that can be introduced, depending on countries and on users [27,28]. It demonstrates that neither "land use" nor "land cover" are neutral terms, but always choices, made in a certain context in time and purpose. A more general question, asked by the philosopher Quine: "what **there** is?", brings closer the "ontologies" (information-science) and the Ontology (philosophy) [29], through the postulate of a single geographical world.

The spatial context, brought by the "there" is the very nature of the geographic information, and if two observations are made at the same location, "*there is*" a relationship between them, whatever they are, even if they belong to two different ontologies. Furthermore, one can ask if an "ontology alignment" is somehow possible. The problem of merging Conceptual Hierarchies using Galois connections, has been addressed in the late 1990s [30].

The already mentioned CORINE-Land cover European CLC project [5], faced this alignment issue, when several countries had to confront their previous land cover surveys, with the necessity to harmonize with Europe. In particular, the British LCMGB (land cover mapping of Great Britain) was built upon a taxonomy divided into 27 classes and 72 variants, versus 44 classes for CLC [31].

Figure 6 shows the (simplified) graphs resulting from two different surveys of the same region. Each land parcel receives a class value from both surveys, and we can derive the two Galois lattices (Figure 6): the partial order (e.g., forest > semi-natural) reflects the fact that every parcel observed and classified as *"forest"*, is also worded as *"semi natural"*. The goal is not to force one taxonomy to fall into the second, rather, to better understand which classes could be the problematic when comparing the ontologies [32]. We accept two postulates: (1) the land parcels belong to the same geographical world, and can be identified directly, providing that geometric errors have been properly corrected; and (2) the classes *"vegetation"* of the two lattices are supposed to be the same common universe. Geography allows making such postulates, what is not necessarily true for other sources of information in Big Data, such as a customer profiles, for instance.

The procedure to obtain the graph of Figure 7 follows the steps:

- Using the same set of parcels as the unique key, a new relation is built, as the union of the two original relations: it is a functional relation.
- A Galois lattice is derived: new nodes emerge, and the new partial order is derived directly from the observations, not from the two original partial orders.

- The new nodes must be interpreted in terms of a mixture of the two original taxonomies: it points out that some original classes are more problematic, e.g., a meadow parcel of the real world could have been classified as herbaceous in Taxonomy 2 (yellow), but either grassland, then semi-natural, or agriculture, in Taxonomy 1 (green).
- Indices of quality can be attached to the original classifications, weighed by the cumulated areas of the related parcels, or by any context data, to quantify the uncertainty attached to every new node.

The overall process gives useful indications on how to take a decision, for instance if the region to monitor is much larger than the test region above, and the goal is to force every parcel classified in one taxonomy to be translated into the other. That was the LCMGB to CLC translation problem.

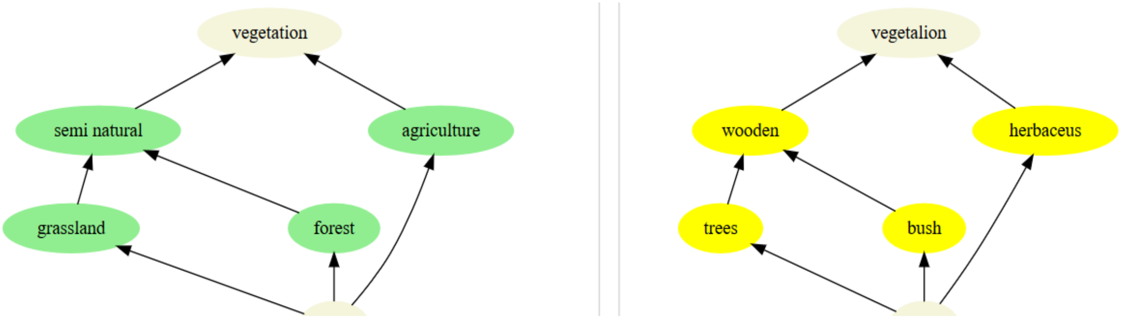


Figure 6. Two ontologies computed as the Galois lattices derived from two observation sets.

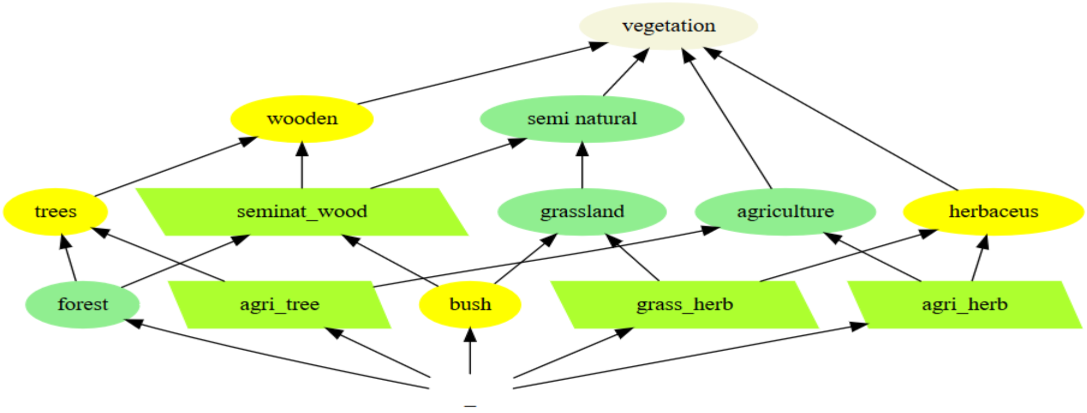


Figure 7. Ontology alignment: combining the two Galois lattices, and naming additional nodes.

This is an a posteriori ontology alignment under spatial constraints. Additional a priori constraints (partial order) may have been introduced, as in the Bayes net example of the previous section.

It is obvious that decision-makers are increasingly relying on data to prepare and make their decisions, but few of them are aware of what is behind these data, and that they result from many undocumented choices and small decision-making processes, at all stages. Hence, data result from a series of decisions, in a certain context that can be represented by ontologies.

Note: the European Environment Agency provides the CORINE-Land-Cover data sets from 1990, 2000, 2006 and 2012. Some modifications in the classification procedure have been introduced through the time. The 1990 data set has been modified, to be "more compatible" with 2000, but new modifications have been introduced later on, and the original 1990 dataset is not available anymore. Hence, a part of the semantics has been lost. This can be a lesson for Big Data: several historical versions of same datasets can be processed on the fly, instead of keeping only the most recent.

In a different scientific field, it is instructive to learn how the archeologists collect data about artifacts, and how they document the collection process: "Data are not Facts, but Acts", is the main thesis of [33], which states that "geographical information should be represented by a set of activities, rather than by a data-set". The issue of the quality of spatial data makes use of ontologies, in particular to differentiating the notion of external quality (quality for the user, or "activated" quality), versus the (internal, or passive) quality declared by the data providers [34].

3. Discussion

Can we say that geomatics was dealing with Big Data before the term was coined?

Let us try to sum up the hurdles that geomatics has had to overcome over four decades, and let us revisit them through the prism of the seven Vs that are invoked to characterize Big Data. Some lessons learned by geomatics may still be useful for Big Data today. Reciprocally, the importance of Big Data as becoming the unique repository for every information collection and dissemination procedure, and for all information processing methods and algorithms, forces the geomatics scientists and engineers to reposition themselves in this new context.

3.1. Volume

Since 1972, the volume of data has been huge, exceeding the regular computing capacities of that time, which necessitated the development of specific tools for reducing computing time to acceptable limits. Now, the main difference is in the much broader use of these data, by many more people. Efficient storage was an issue yesterday for remote sensing, as it is today for Big Data, probably with similarly high stakes. Communication through networks was inexistent then, and the necessity of a rapid access to data is now the challenge. The "*MapReduce*" concept is the major system improvement that enables "Big Data for all", though parallel computing was stammering 40 years ago.

3.2. Velocity

The kind of data as well as the kind of users that we could see in geomatics was not tied to data dissemination issues, until the mid 2000s. Most pictures were static pictures; maps are not updated every day. Some exceptions, such as the processing of weather imagery, were limited to specific professionals. The dissemination of online mapping, including aerial and street views, has profoundly changed the market. The recent and rapidly growing use of drones for close aerial imagery introduces a profound change, as does the massive introduction of environmental sensors (Internet of Things)—without forgetting the surveillance cameras. This is now an opportunity for Big Data players to broaden their scope of activities to almost every niche of Geomatics.

3.3. Variety (and Visualization)

Geomatics data were, at first, mostly measurements, signal, geometry, time series. The link with plain language information was limited to the kind of scarce text as we can find printed on maps. The challenge with data that are, all at once, huge and diverse, was addressed by bringing in some structure: from minimal ancillary data, for images, to first-order logics (the relational model) for data with databases. In the 1990s, the analytical approach introduced additional preprocessors devoted to restructure (i.e., ETL, Extract-Transform-Load) the initial information, along several chosen dimensions.

On the other hand, the realm of Big Data covers the land of "verbum", as in the Gospel of John: "*In principio erat Verbum*". Verbum excerpted from web sites, from social networks includes trillions of sparse, scattered, unrelated docs (unstructured) anywhere on the Internet. So-called "robots" are trying to relate and rank the responses to every request (i.e., indexing the Internet), including complex and multi-morphed requests (data analytics). This "on the fly" preprocessing is a considerable extension of what the "data warehouses" are doing, though much less structured, until recently. XML and ontologies are the key tools for structuring this unstructured realm of verbum.

The concept of unstructured data was not in use 30 years ago: relational databases were just on the rise. The term semi-structured data appeared in 1995, and XML was first established in 1997. However, NoSQL is not erasing SQL: for example, the service Big Query, by Google, uses SQL. In the meantime, the rise of cloud computing greatly facilitates the development of the global "Big Data as a Service" (BDaaS). Teich says [35]: "*the SQL programming language remains the best means for accessing and querying data, whether it's in relational databases, NoSQL systems, or Hadoop clusters*".

3.4. Value, Veracity, Validity, and Variability

"*The three V's—volume, velocity and variety—do a fine job of defining Big Data. [...] Variability, veracity, validity and value aren't intrinsic definitional Big Data properties. They are not absolutes. By contrast, they reflect the uses you intend for your data. They relate to your particular business needs*" [36]. This statement is revealing the importance of this additional set of "Vs" if we consider Big Data not solely as a platform, but as a service (BDaaS).

In the 1970s, the full armory of mathematical tools was (almost) there. Principal component analysis, pattern recognition, complex spatial queries, decision making, exploratory data analysis, etc., are the foundations of machine learning and data analytics. If you look deep into the algorithms of today, you can find the legacy of geo-processing: ask IBM Watson to testify!

Veracity is close to what geomatics names "data quality." This aspect has been extensively studied by the spatial data providers (e.g., NGOs), including for legal reasons. It is not yet the case with Big Data, because relationships and correlations are computed on text information rather than on physical measurements. However, the trend is to give much closer attention to this aspect, as well as the two other facets of uncertainty, which are Validity and Variability. This later term was added as one more "V" for Big Data: "[*by variability*] *I mean variance in meaning, in lexicon. The best example of that would be the variability problem that the [supercomputer] Watson at IBM was trying to take on. [Watson] would get an answer and would have to dissect that answer into its meaning and then to try to figure out what the right question was within that three-second response time*" [37]. This vision is quite similar to the notions explored with the ontologies and their alignment.

3.5. Conclusions

Some 40 years ago, the multiplication of remote sensing imagery, of automated cartography, and a rapidly increasing computing power offered an opportunity to merge and process an enormous amount of data. Geomatics was on that leading edge, as was biology because DNA-sequencing in the 1990s contributed highly to machine learning algorithms, particularly fitted for text processing. However, the specificity of geomatics is the large variety of its sources of information, and therefore, the large variety of challenges to overcome.

This finding legitimates the question of what lessons can be learned from these four decades. The rise of geomatics, as a comprehensive system of technologies, practices and products, then followed by its present relative fade, is symptomatic of greater changes that occurred in that period. The Figure 1 shows that AI and databases have followed a similar fate. Figure 1 also demonstrates that the change has occurred in 1995, when the use of the Internet became the principal vector of almost any other numerical technological development.

Lesson 1: In geomatics, the issue of representing the information has been the highest challenge: (a) to overcome the many formats for raster data, or vector data; (b) to represent complex metadata, up to ontologies; and (c) then to help processes to become seamlessly interoperable, up to information fusion. Big Data can build upon XML or JSON, adding some logics with RDF and OWL, but the challenge of information fusion is still high.

Lesson 2: Uncertainty is everywhere, and real world observations are hampered by variable quality levels of many kinds, which any information fusion process must treat together with the data, what increases the complexity in an exponential way. It took about two decades before quality issues become a major concern in geomatics. In Big Data, which is driven largely by marketing automation, quality awareness is not yet a major concern, but should occur soon, e.g. see [38] for an analysis of a Google failure.

Lesson 3: Kuhn wrote [39]: "The decision to reject one paradigm is always simultaneously the decision to accept another". The release of AOL and Google maps, followed by the widespread of GPS navigators, together with smartphones, decided eventually that geomatics is now part of Big Data. Though not everyone is aware. The impact will be important, for the companies of course, for the universities curricula, and for the kind of jobs that will be offered. The future is probably for a broad generation of "data scientists", equipped with minor degrees in social, or natural sciences, or in law studies.

Big challenges are ahead of us, including for geomatics: crowdsourcing and volunteered geographic information, widespread use of drone imagery, of surveillance cameras, the Internet of Things in urban and natural spaces, to cite the most obvious ones.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BDaaS	Big Data as a Service
DBMS	Data Base Management System (RDBMS: Relational DBMS)
GIS	Geographic Information System
ISO	International Standards Organization
LAI	Leaf Area Index (vegetation monitoring)
NGO	National Geographic Organization
OECD	Organisation for Economic Co-operation and Development
OLAP	On-Line Analytical Process (SOLAP: Spatial OLAP)
PCA	Principal Component Analysis
UNESCO	United Nations Educational, Scientific, Cultural Organization

References

1. Michel, J.B.; Shen, Y.K.; Aiden, A.P.; Veres, A.; Gray, M.K.; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; Orwant, J. Quantitative analysis of culture using millions of digitized books. *Science* **2011**, *331*, 176–182. [CrossRef] [PubMed]
2. Thomson-Reuters. Available online: http://blog.thomsonreuters.com/index.php/Big%20Data-graphic-of-the-day (accessed on 24 August 2016).
3. Kimball, R.; Ross, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2002.
4. Dempsey, C. Where is the Phrase "80% of Data is Geographic". Available online: https://www.gislounge.com/80-percent-data-is-geographic/ (accessed on 24 August 2016).
5. European Environment Agency. CORINE Land Cover—Part1: Methodology. Available online: http://www.eea.europa.eu/publications/COR0-part1 (accessed on 24 August 2016).
6. Bédard, Y.; Lam, S.; Proulx, M.-J.; Caron, P.-Y.; Létourneau, F. Data warehousing for spatial data: Research issues. In Proceedings of the International Symposium: Geomatics in the Era of Radarsat (GER'97), Ottawa, ON, Canada, 25–30 May 1997.
7. Stefanovic, N. Design and Implementation of On-Line Analytical Processing (OLAP) of Spatial Data. Ph.D. Thesis, School of Computing Science, Simon Fraser University, Vancouver, BC, Canada, January 1997.
8. Rivest, S.; Bédard, Y.; Proulx, M.J.; Nadeau, M.; Hubert, F.; Pastor, J. SOLAP: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *ISPRS J. Photogramm. Remote Sens.* **2005**, *60*, 17–33. [CrossRef]
9. Bernier, E.; Gosselin, P.; Badard, T.; Bédard, Y. Easier surveillance of climate-related health vulnerabilities through a Web-based spatial OLAP application. *Int. J. Health Geograph. Apr.* **2009**, *8*, 18. [CrossRef] [PubMed]
10. Han, J.; Stefanovic, N.; Koperski, K. Selective materialization: An efficient method for spatial data cube construction. In *Research and Development in Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 144–158.

11. Qiu, J.; Gao, H.; Ding, S.X. Recent advances on fuzzy-model-based nonlinear networked control systems: A survey. *IEEE Trans. Ind. Electron.* **2016**, *63*, 1207–1217. [CrossRef]
12. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
13. Jeansoulin, R.; Fontaine, Y.; Frei, W. Multitemporal segmentation by means of fuzzy sets. In Proceedings of the 7th LARS Symposium on Machine Processing of Remotely Sensed Data, with Special Emphasis on Range, Forest, and Wetlands Assessment, Purdue University, West Lafayette, IN, USA, 23–26 June 1981; pp. 336–340.
14. Diday, E. The dynamic clusters method in nonhierarchical clustering. *Int. J. Comput. Inf. Sci.* **1973**, *2*, 61–88. [CrossRef]
15. Olioso, A.; Prevot, L.; Baret, F.; Vlevers, J.G.P.W. Spatial aspects in the Alpilles-ReSeDA project. In Proceedings of the Workshop Scaling and Modelling in Forestry: Applications in Remote Sensing and GIS, Montréal, QC, Canada, 19–21 March 1998.
16. Jonckheere, I.; Fleck, S.; Nackaerts, K.; Muys, B.; Coppin, P.; Weiss, M.; Baret, F. Review of methods for in situ leaf area index determination. *Agric. For. Meteorol.* **2004**, *121*, 19–35. [CrossRef]
17. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: New York, NY, USA, 1990.
18. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
19. Cortes, C.; Vapnik, V.N. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
20. Jeansoulin, R.; Wilson, N. Quality of geographic information: Ontological approach and artificial intelligence tools in the Revigis project. In Proceedings of the 8th EC-GI& GIS Workshop, Dublin, Ireland, 3–5 July 2002.
21. Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; Murdock, J.W.; Nyberg, E.; Prager, J. Building Watson: An overview of the DeepQA project. *AI Mag.* **2010**, *31*, 59–79.
22. Randell, D.A.; Cui, Z.; Cohn, A.G. A spatial logic based on regions and connection. In Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, San Mateo, CA, USA, October 1992; pp. 165–176.
23. Euzenat, J.; Bessière, C.; Jeansoulin, R.; Revault, J.; Schwer, S. Dossier Raisonnement spatial et temporel. *Bull. de l'Assoc. Fr. de l'Intell. Artif.* **1997**, *29*, 2–13.
24. Cavarroc, M.-A.; Benferhat, S.; Jeansoulin, R. Modeling land use changes using Bayesian networks. In Proceedings of the 22nd IASTED International Conference on Artificial Intelligence and Applications, Innsbruck, Austria, 16 February 2004.
25. Gruber, T.; Olsen, G. An ontology for engineering mathematics. In Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning, Bonn, Germany, 24–27 May 1994; pp. 258–269.
26. Halevy, A. *Why Your Data Won't Mix?*; ACM Queue: New York, NY, USA, 2005.
27. Comber, A.J.; Fisher, P.; Wadsworth, R. Ignore the ontological aspects of land cover at your peril: A plea for expanded metadata. In Proceedings of the Remote Sensing & Photogrammetry Society Conference, Aberdeen, UK, 6 September 2004.
28. Lund, H.G. *Definitions of Forest, Deforestation, Afforestation, and Reforestation*; Forest Information Services: Gainesville, VA, USA, 2007.
29. Quine, W.V. *On What There is*; Harvard University Press: Cambridge, TN, USA, 1948.
30. Ganter, B.; Wille, R. *Formal Concept Analysis: Mathematical Foundations*; Springer: Berlin, Germany, 1999.
31. Smith, G.M.; Brown, N.J.; Thomson, A.G. *CORINE Land Cover 2000: Semi-Automated Updating of CORINE Land Cover in the UK*; Centre for Ecology and Hydrology, UK Natural Environment Research Council: Monks Wood, UK, 2005.
32. Pham, T.T.; Phan-Luong, V.; Jeansoulin, R. Data quality based fusion: Application to the land cover. In Proceedings of the 7th International Conference on Information Fusion (FUSION'04), Stockholm, Sweden, 28 June–1 July 2004.
33. Jeansoulin, R.; Curé, O.; Ahmed, A.; Gademer, A.; Rudant, J.-P. Geographical information is an act, not a Fact. In Proceedings of the 12th AGILE International Conference on Geographic Information Science, Leibniz University, Hannover, Germany, 2–5 June 2009.

34. Vasseur, B.; Jeansoulin, R.; Devillers, R.; Frank, A. External quality evaluation of geographical applications: An ontological approach. In *Fundamentals of Spatial Data Quality*; Devillers, R., Jeansoulin, R., Eds.; ISTE Publishing: London, UK, 2006; pp. 255–270.
35. Teich, D.A. SQL -vs- NoSQL: Database Design Debate Isn't Even a Real Fight. February 2016. Available online: http://searchdatamanagement.techtarget.com/tip/SQL-vs-NoSQL-database-design-debate-isnt-even-a-real-fight (accessed on 24 August 2016).
36. Grimes, S. Big Data: Avoid 'Wanna V' Confusion. Available online: http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077 (accessed on 24 August 2016).
37. Hopkins, B. Forrester Principal Analyst, in TechTarget Interview, by Mark Brunelli. Available online: http://searchdatamanagement.techtarget.com/news/2240036228/Will-your-organization-benefit-from-big-data-processing-technology (accessed on 24 August 2016).
38. Lazer, D.; Kennedy, R. What We Can Learn From the Epic Failure of Google Flu Trends. Available online: http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/ (accessed on 24 August 2016).
39. Kuhn, T.S. *Structure of Scientific Revolutions*; University of Chicago Press: Chicago, IL, USA, 1962.