A Global Approach to Data Value Maximization

A Global Approach to Data Value Maximization:

Integration, Machine Learning and Multimodal Analysis

By

Paolo Dell'Aversana

Cambridge
Scholars
Publishing

A Global Approach to Data Value Maximization:
Integration, Machine Learning and Multimodal Analysis

By Paolo Dell'Aversana

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-3147-3
ISBN (13): 978-1-5275-3147-5

TABLE OF CONTENTS

PREFACE

Over the past thirty years, I have worked in many projects involving the acquisition, processing and interpretation of geophysical data. Using seismic, electromagnetic and gravity data, I have developed and applied approaches and algorithms for the modeling and inversion of multidisciplinary geophysical measurements. The output of these methods commonly consists of "Earth models" of the spatial distribution of various physical parameters, such as seismic velocity, electrical resistivity, density, fluid saturation, and porosity.

Using large datasets, I have frequently applied Data Science and Machine Learning approaches for supporting and improving my integrated workflow. Sometimes, colleagues, researchers and managers have applied the results of my work to improve their geological models and/or their decisional process. Indeed, a robust model helps in making key decisions, like where to drill a new exploration well.

Unfortunately, the geophysical and geological models are often affected by uncertainties and ambiguities, including the models produced by me, of course. Among the main reasons for such intrinsic indetermination, there is the fact that the exploration target is frequently located at a depth of several kilometers in the terrestrial crust, below complex geological sequences. This often happens, for instance, in hydrocarbon exploration. Consequently, the geophysical response measured at the surface can be characterized by a low signal-to-noise ratio. When geoscientists try to retrieve Earth models from that response, the measurement uncertainties propagate from data to model space, affecting negatively the reliability of the Earth models. These models represent "interpretations" rather than objective information. For that reason, Earth disciplines are a typical example of interpretative science. In other words, geoscientists can produce (very) different Earth models and different interpretations starting from the same experimental observations. The differences depend on many factors, not confined to data quality. These include the personal technical background, the individual experience and sensitivity, the specific ability in using technology for enhancing the signal and for reducing the noise, and so forth.

Under these aspects, geosciences are not very different from many medical disciplines where, for instance, physicians must define a diagnosis based on multidisciplinary observations affected by large uncertainties.

Finally, both geoscientists and physicians must make crucial decisions in uncertain domains.

Over the years, geoscientists, as well as physicians, have learned how to manage experimental errors and model uncertainties. However, there are still many methodological open questions behind their interpretative work.

First, how much do they really understand about the data that they use?

Second, do they properly understand the *meaning* of the models that they retrieve from their data?

Third, how can they extract the maximum informative value from both data and models?

Fourth, how can they optimize the decisional process in uncertain domains using the entire value of information in both data and model space?

Of course, the above questions are not restricted to the domain of the Earth or medical disciplines. The problem of understanding properly the data and the models, exploiting their entire informative value, is generalized to all the main scientific fields. Unfortunately, that problem often remains unsolved: we do not use information in the correct way because we understand it just partially. We waste a great part of its potential value.

The "Gap of Understanding" ("GOU" could be a nice acronym) is related in some way with the number (and with the relevance) of *obscure* steps of the workflow through which we move from data to models, and, finally, from models to decisions.

Assuming that we do our work of data analysis and interpretation in the most honest and scrupulous way, a problem remains in the background. It is data complexity.

The question is that the rapid growth of information and the intrinsic complexity of many modern databases often require extraordinary efforts for exploring the entire volume of information and for maximizing its true value. Besides the *volume* of *Big Data*, there are additional important aspects to take into account. These are *variety*, *veracity*, *velocity*, *validity* and *volatility*. In fact, data complexity increases not only with data volume, but also with the heterogeneity of the information, the non-linearity of the relationships, and the rate by which the data flow changes.

As I said, all that complexity can be a problem. Sometimes we think to solve this problem by just ignoring it. We tend to simplify. Unfortunately, excessive simplification can lead us towards wrong Earth models, wrong medical diagnosis, and wrong financial predictions. Finally, that simplistic approach drives us towards wrong decisions. On the other side, complexity often represents an opportunity rather than a problem. Complexity, if properly managed and correctly understood, can trigger positive changes and innovative ideas. It is an intellectual, scientific, technical challenge.

This book is a systematic discussion about methods and techniques for winning that challenge. The final objective of the following chapters is to introduce algorithms, methods and approaches to extract the maximum informative value from complex information.

As I said, dealing with "Big Data" and with complex integrated workflows is the normal scenario in many Earth disciplines, especially in the case of extensive industrial applications. For this reason, the book starts from the domain of geosciences, where I have developed my professional experience. However, the discussion is not confined to applications in geology and geophysics. It is expanded into other scientific areas, like medical disciplines and various engineering sectors. Similar to geosciences, also in these fields, scientists and professionals are continuously faced with the problem of how to get the maximum value from their datasets. That objective can be obtained using a multitude of approaches.

In the book, algorithms, techniques and methods are discussed in separate chapters, but in the frame of the same unitary view. These methods include data fusion and quantitative approaches of model integration, multimodal data analysis in different physical domains, audio-video displays of data through advanced techniques of "sonification", multimedia machine learning and hybrid methods of data analysis.

Finally, human cognition is also taken into account as a key factor for enhancing the informative value of data and models. Indeed, the basic intuition inspiring me in writing this book is that information is like an empty box if we do not extract any coherent significance from it. This can be a geological, medical, or financial type of significance, depending on the field of application. In other words, the value of information increases if we understand its deep meaning. That intuitive principle is true in science as well as in ordinary life. We can effectively estimate the value of information only after understanding its significance. Consequently, the problem of maximizing the information value is translated into a more general problem: maximizing our capability to extract significance from the information itself.

This methodological approach requires that human sciences are involved in the workflow. In particular, it is important to clarify the concept of "significance of information". This concept is extremely complex and has involved philosophers and scientists for many centuries. For that reason, I have tried to summarize the "question of significance" in different parts of the book, explaining my point of view about it. Especially in the final part, I discuss how modern neurosciences, cognitive disciplines and epistemology

can contribute to the process of maximization of the information value through the analysis of its semantic aspects.[1]

A multitude of examples, tutorials and real case histories are included in each chapter, for supporting the theoretical discussion with experimental evidences. Finally, I have included a set of appendices at the end of the book, in order to provide some insight about the mathematical aspects not explicitly discussed in the chapters.

Due to the multidisciplinary approach that I use in this book, I hope that it can engage the interest of a large audience. This should include geophysicists, geologists, seismologists, volcanologists and data scientists. Moreover, researchers in other areas, such as medical diagnostic disciplines, cognitive sciences and the health industry, can find interesting ideas in the following chapters. No specific background is required for catching my key messages. In fact, this book is aimed mainly at introducing novel ideas and new research directions rather than exhaustively covering specialist topics. Consequently, I have often preferred to discuss the technical details in the appendices, in order to make the discussion more fluid and readable. Furthermore, I have provided the main references and many suggested readings at the end of each chapter for those who are interested in expanding a specific subject.

In summary, the only fundamental requirement for deriving benefit from this book is to read it with an open mind, with the curiosity to investigate the fascinating links between disciplines commonly considered independent.

[1] In the linguistic field, Semantics is the study of meaning. In the semiotic field, it deals with the relations between signs and what they denote. In this book, I use the term “semantic” in a very general sense, for denoting the meaning of words, of sentences, of concepts, and of information in general.

CHAPTER ONE

VALUE OF INFORMATION

Abstract

The rapid growth of information and the intrinsic complexity of many databases, require effective approaches and methods for maximizing the informative content of the data. Beside the *volume* of *Big Data*, there are additional important aspects to take into account. These include *variety*, *veracity*, *velocity*, *validity* and *volatility*. In fact, complexity increases not only with data volume, but also with the intrinsic heterogeneity of sources and types of data, with the non-linearity of the relationships between the data, with the velocity with which the data change, and so on. However, data complexity often represents an opportunity rather than a problem. In order to take profit from it, it is necessary to develop effective workflows for extracting the maximum informative value. In this chapter, after introducing some key aspects of informative complexity, I discuss the concept of Value of Information. I explain how this can be estimated through a Bayesian approach. Then I introduce the roadmap of the process of data value maximization. This combines the benefits of different techniques and methods including Data Fusion, Multimodal data analysis, Audio-Video display and Multimedia Machine Learning, with the additional contribution of Cognitive Sciences and Neurosciences.

Keywords: Big Data, Complex information, Value of Information, Bayes' Theorem.

1.1. Introduction: Big Data and Complex Information

Huge amounts of datasets are continuously created in almost all the modern scientific fields, in many business sectors, through social media and during many other activities of our daily routine. *Big Data* is one of the most inflated expressions used today. It commonly refers to huge *volumes* of data ranging from terabytes to many petabytes.

Beside the memory size occupied by big databases, there are additional important aspects to take into account (Elgendy and Elragal, 2014). V*ariety*

refers to the heterogeneity of sources and types of data, both structured and unstructured. *Velocity* is intended as the rate at which the massive and continuous data flow changes over time and space. Additional important features are *veracity* and *validity*, related with the biases, noise and abnormality in data, and with information accuracy, respectively. *Volatility* is another important aspect concerning how long data are valid and how long they should be stored. Apart from the volume, these features do not concern exclusively Big Data, but also databases of ordinary size.

A general aspect of every type of data is the *Value of Information (VOI)*. This is crucial for Big Data as well as for "standard" databases. It is not a trivial issue to define and estimate the VOI. From a pragmatic point of view, we can think that the VOI is related to how data affect our decisions. In this chapter, I will show how this intuitive idea can be transformed into a useful, quantitative concept.

The full informative value can be properly extracted from the dataset if we are able to handle its intrinsic complexity (Ahmed, S. Ejaz (Ed.), 2017). Thus, *data complexity* represents an additional general aspect of information strictly linked with its value. The reason for that link is not immediately clear. The word "complexity" combines the Latin roots *com* (meaning "together") and *plex* (meaning "woven"). In fact, a complex system is characterized by the inter-dependencies of its components. In particular, in a complex informative system, the different components (measurements, models, uncertainties, methods of acquisition, techniques of analysis, and so forth) are linked and interact in multiple ways. These interactions are often "hidden" and implicit, and must be discovered in some way. Furthermore, the links are often expressed by non-linear relationships that increase exponentially the level of complexity.

This complexity can strongly affect our work, our final decisions, and our performances in using the data, with both negative and positive impacts. For instance, data complexity can increase significantly the computation time, the storage and/or other resources necessary to execute the algorithms. *Computational complexity* is related to the number of an algorithm's steps (time complexity) and/or the number of storage locations that it uses (space complexity) for solving a given computation problem.

On the other side, data complexity can represent an opportunity rather than a problem, but only if we are able to find the relationships hidden in the data. In fact, when we are able to link different pieces of information, we can discover new meaningful structures like data clusters, important correlations, and causal relationships. Finally, that type of *semantic structure* can significantly improve our knowledge. The most intuitive example of this fundamental concept is the discovery of a new scientific law

or, more simply, a new empirical relationship explaining a set of experimental evidences. At first, our observations can appear sparse and disconnected, but if we able to find the proper link(s), then the same data can be re-organized into significant conceptual clusters.

In this book, I will show that complex datasets consisting of multi-physics measurements obtained from multiple and heterogeneous sources, can be properly integrated (Data Fusion, Image Fusion, Model Integration). Such an integration process allows improvement in the solution of difficult problems, like medical diagnosis, geophysical imaging and the decisional process in uncertain domains. Of course, transforming the *disordered complexity* (Weaver, 1948) of a huge and heterogeneous dataset into a coherent model (or a coherent theory) is not a simple process. I will show that it requires specific competences, efficient algorithms, analytical methods, and effective workflows. Despite that intrinsic difficulty, the reward of data fusion and integration can be extremely high in terms of improvement of the VOI.

In summary, the impressive growth of information, and the intrinsic complexity of many databases, require effective approaches and strategies aimed at *maximizing the Value of Information*. In this chapter, I introduce the key factors affecting the VOI and how it can be estimated through a Bayesian approach. Furthermore, I start discussing the roadmap of the process of data value maximization.

1.2. From simple to hyper-complex

Few examples can be useful for clarifying the concepts of data simplicity and information complexity. For the moment I do not provide any formal definition of "simple" and "complex"; however, I illustrate intuitively how complexity grows not only with data volume, but also with other parameters, like data heterogeneity, non-linearity of the relationships, the number of sources of information and spatial dimensionality.

Let us start from a trivial geophysical example of a small and simple dataset, as shown in Fig. 1-1. In this case, just a few well-log measurements[1] are cross-plotted (x-axis: slowness;[2] y-axis: rock density scaled by a factor of 100). The points show some scattering; in fact, the parameter R^2 is significantly less than 1 (it is defined in such a way that it should be equal

[1] Borehole logging is the practice of making a detailed record (a well log) of the geologic formations penetrated by a borehole.

[2] Slowness is a quantity introduced in Seismology. It is the reciprocal of propagation velocity of seismic waves.

to 1 for a perfect fit). However, the measurements follow a clear decreasing linear trend. This dataset can be considered "simple" not only because it consists of a small set of measurements, but also because the variables are correlated in a linear way. In this case, I am assuming that *simplicity* is inversely related to the degree of the mathematical relationship linking (fitting) our data.

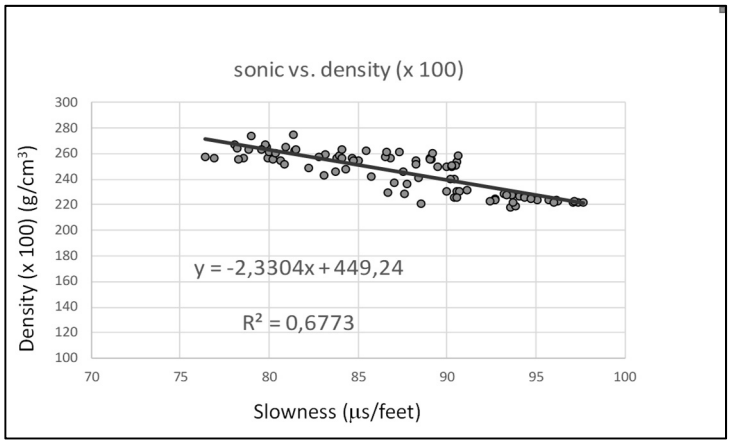

Fig. 1-1. Example of linear relationship fitting a small dataset.

Continuing with this intuitive idea of simplicity, Fig. 1-2 shows an example of a slightly more complex dataset, where the physical measurements are linked through a non-linear relationship. In this case, the water saturation inside an oil reservoir is related to the electrical resistivity through a power-law. The correlation is not perfect, but R^2 is not too far from the unity. This indicates that the empirical relationship fits the data within a good level of approximation. However, there is a certain scattering around the curve fitting the data.

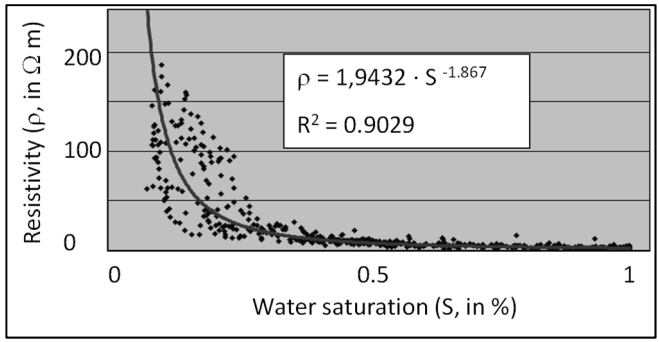

Fig. 1-2. Non-linear relationship example.

At this point of the discussion, an important question arises: what do I mean by the above expression "good level of approximation"? Looking at Fig. 1-2, we can notice that the data with a saturation value below 0.3% are quite scattered and could be split into two separate clusters. These clusters could be fit using two different curves expressed by two different power-laws. This is an example of the well-known trade-off between accuracy and simplicity. If we desire to increase the accuracy of our descriptive models, the counterpart is that the models can become more complex. In this example, "more complex" means that two curves rather than one are necessary for fitting two distinct data clusters.

In many practical cases, scientists prefer to use the simplest model that is able to explain the observations. This approach is known as "Occam's razor". It can be summarized by the following statement: *when there are competing hypothetical solutions to a problem, one should select the one that makes the fewest assumptions*. Unfortunately, in many cases, the decision is not simple; increasing the complexity of the models can be necessary, for instance, for robust physical reasons. In the example of Fig. 1-2, it could be more appropriate to split the data into two clusters rather than fitting all the points with one relationship only. Indeed, the reason for the scattering observed in the figure is that the reservoir consists of two stacked geological layers. These have similar sedimentary properties, but these are not exactly the same.

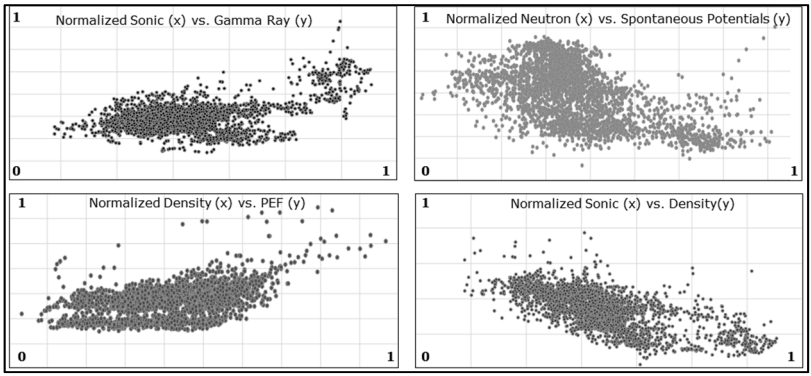

Fig. 1-3. Normalized cross-plots of composite well logs. This dataset is formed by many measurements obtained through different types of acquisition methods. They show some degree of correlation, but can be explained (fit) only partially through simple mathematical relationships.

I will discuss specific methods and algorithms aimed at handling the trade-off between the intrinsic complexity of the data and the simplicity of the models. The following example will aid a better understanding of this point.

Fig. 1-3 shows a dataset with a higher complexity than the previous examples. In this case, there are four cross-plots combining different types of borehole measurements: sonic, gamma ray, neutron, spontaneous potentials, density and PEF (photoelectric absorption) well logs. We can observe "clouds" formed by many measurements. These show some degree of correlation, but this is generally nonlinear and there is a high level of scattering. The reason is that this dataset includes the measurements performed at increasing depths in the well crossing variable geological formations. Consequently, the cross-plots mix trends belonging to different sedimentary strata. In this example, fitting all the data in each panel with a unique mathematical law is geologically inappropriate. In other words, the complexity of our data is too high to be handled with simplistic mathematical relationships, even if we used non-linear laws. In cases like this, it is generally more appropriate to split the data into different clusters, before trying to model them through a simplistic mathematical approach. I will discuss in dedicated sections, how the trends related to different rock formations can be grouped into relatively homogeneous clusters through appropriate techniques of clustering analysis.

Beyond these specific geophysical examples, the key message is that we need to "understand" our datasets as much as possible before trying to "model" them through mathematical and statistical approaches. This statement is particularly true when data complexity increases.

In almost all the natural sciences, as well as in medical disciplines, in social and psychological sciences, we generally deal with systems that are much more complex than a well-log dataset. *Hyper-complexity* is a higher level of complexity typical of datasets characterized by experimental measurements obtained through different methods and sources. Data generally form huge databases (Big Data), belong to different domains and are linked through non-linear mathematical relationships. Typical examples of hyper-complex databases are formed by multidisciplinary geophysical data or by heterogeneous medical data including blood analysis, statistical information, and high-resolution images obtained with different imaging technologies.

Fig. 1-4 shows an example of a multidisciplinary geophysical model. In this case, two sections of two different physical parameters are compared. In the top panel there is a tomography section obtained through inversion of seismic data (travel times); it represents the 2D spatial distribution of seismic propagation velocities. In the bottom panel there is a section of electrical resistivity obtained through inversion of electromagnetic data recorded with the Magnetotelluric method (Cagniard, 1953). The two sections show large-scale similarities in the distribution of the two different parameters, because these are physically linked. In fact, there is a (non-linear) relationship between seismic and electromagnetic measurements and, consequently the models derived from these data are linked to some degree. The physical reason is that both resistivity and velocity change consistently in many types of rocks: frequently, although not in every case, when seismic velocity increases, the same happens for electrical resistivity, as in many carbonate and/or shale rocks.

Fig. 1-4a also shows the *interpretation*, marked by the black curves, of the main trends of the velocity field. These trends represent high-velocity/high-resistivity carbonate and shale thrusts, and low-velocity/low-resistivity basins. The spatial distributions of these geological units are typical of the Southern Apennine belt (Italy), where these data come from. In the bottom panel, Fig. 1-4b, we can see that the resistivity parameter follows approximately the same general geometric trend. There is no rigid spatial correlation between the two sections shown in the upper and lower panels. However, it is clear that some degree of semblance exists, explainable in terms of large-scale geological trends. This is a typical example of hyper-complexity, where two different geophysical domains

(seismic and electromagnetic) are correlated in some way. Thus, the intrinsic complexity of each individual domain is further increased by the fact that these domains are reciprocally linked through a non-linear relationship.

The same type of hyper-complexity, but in the 3D case, is shown in Fig. 1-5. This is an example of the co-rendered imaging of seismic, gravity and electromagnetic information (Colombo et al., 2014). Panel a) shows two seismic cross-sections extracted from a huge 3D seismic cube. Panel b) shows the correspondent velocity models. These highlight the different geological units, including the salt layer. Panel c) shows the density model obtained from gravity data that provides information about the main trend of the basement rocks (the deep layer). Finally, panel d) shows the resistivity models obtained from electromagnetic data inversion. Resistivity adds significant information about the geometry of the salt formation, including interesting shallow details not properly revealed by seismic data. This figure is an example of how different geophysical methods can contribute to define a multi-physics geological model.

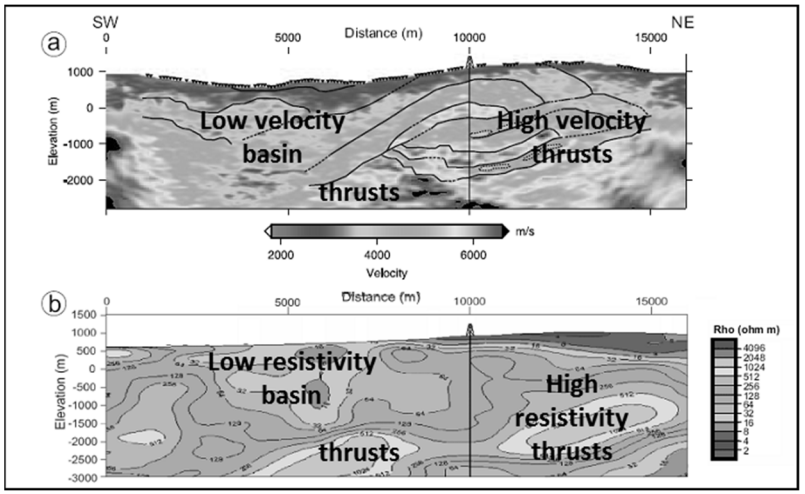

Fig. 1-4. Example of a multidisciplinary geophysical model: 2D section of the tomography seismic velocity model (panel a); 2D section of the resistivity model obtained through inversion of electromagnetic data (panel b) acquired along the same profile. Topography in panel b is strongly smoothed.

Moving from the 2D case of Fig. 1-4 to the 3D case of Fig. 1-5, the hyper-complexity is increased by the huge volume of the dataset and by the higher spatial dimensionality. Thus, we have an example of Big Data and complex information at the same time. Nowadays, this is the ordinary scenario in exploration geophysics, in the hydrocarbon industry as well as in other sectors of the modern geosciences. I will discuss the importance of using the entire complexity of these types of multidisciplinary measurements. Furthermore, I will introduce the most modern approaches and methodologies used for extracting the maximum Value of Information from these hyper-complex datasets.

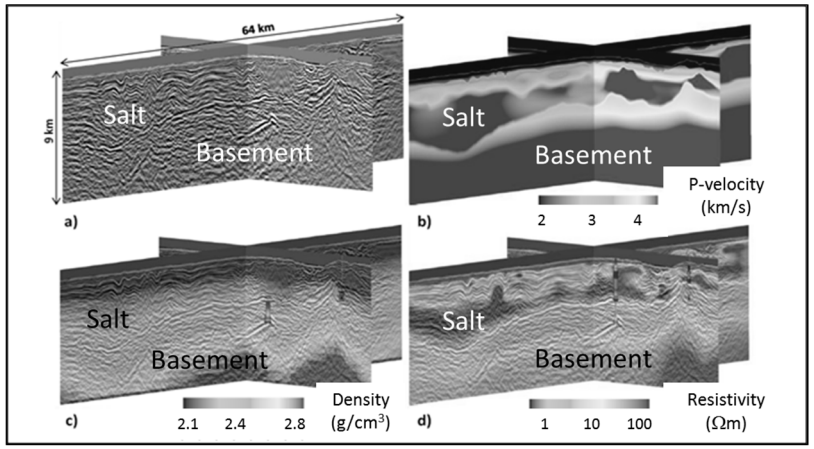

Fig. 1-5. Example of the multidisciplinary geophysical model (Courtesy of Colombo et al., 2014, modified). See text for details.

In many scientific disciplines, including geosciences and medical sciences, as well as in many financial and business sectors, the databases can reach impressive levels of hyper-complexity. The reason is that they are not static, but change continuously over time. Let us consider, for instance, the case of an oil field during production: huge and heterogeneous databases are updated day by day, or even hour after hour, including new information coming from production data, geological and geophysical measurements, and laboratory tests. Extracting the full information value from such *big*, *complex* and *dynamic* databases is the main challenge for many geoscientists, engineers and managers. They must make key decisions that can have an impact quantifiable in terms of many millions of dollars (or

much more). Thus, it is clear that developing effective algorithms, methods and workflows for maximizing the Value of Information of big, complex and dynamic data represents an important objective, in geosciences as well as in other scientific areas.

1.3. Bayes' Theorem and Value of Information

1.3.1. Basic concepts

A crucial question is, why are many databases big, complex and dynamic? An intuitive answer is that in ordinary life, as well as in scientific practice, we continuously acquire and combine new observations with previous data. Consequently, information grows continuously. The second crucial question is whether our knowledge also increases at the same rate as information. This is not a trivial question. In fact, information and knowledge are not the same thing.

For instance, in medical diagnosis, physicians generally combine many different types of information obtained at different times and with different modalities. These can include the patient's anamnesis, symptoms' description, general statistical information, new specific data coming from blood analysis, and results derived from imaging techniques. The same happens in geophysical exploration. For instance, new data acquired through geophysical prospecting techniques are often combined with data coming from previous surveys aimed at estimating the distribution of one or more physical parameters in the subsoil.

The open questions concern if, how and in what measure this massive work of data integration can improve our knowledge. In order to answer these difficult questions, we can try to reformulate them in probabilistic terms.

In medical sciences, as well as in Earth disciplines and in many other scientific sectors, it is important to estimate how the acquisition of new information can update the probability distribution of one or more properties of the system under study. This problem can be formulated in the frame of the Bayesian Statistical Inference Theory (Stone, 2013; Russell and Norvig, 2016). It allows the estimation of "how the *a priori probability distribution* is transformed into the *a posteriori probability distribution*, by incorporating a physical theory (relating the model parameters to some observable parameters) and the actual result of the observations (with their uncertainties)" (Tarantola, 2005, Preface).

Russell and Norvig (2016) consider the *unconditional* or *prior probability* as "referred to the degree of belief in propositions in the absence

of any other information" (p. 485). It represents the best rational assessment of the probability of an outcome based on established knowledge before performing the present experiment.[3] Instead, *posterior probability* is the revised probability distribution of an outcome after taking into consideration new information (such as new experimental observations).

In common scientific practice, estimating the updating of prior probability into posterior probability can be a very difficult task. In fact, when we acquire new information through a new set of observations, "*O*", this is combined with an entire *system of knowledge*. This includes not only previous data and models, but also expectations about the future, constraints, beliefs, and hypotheses. Consequently, the decisions and the behavior of a hypothetical *agent* (a scientist, a manager …) are affected by the new data combined with a *complex state of prior knowledge*.

In many practical cases, we are interested in estimating how the new data are combined with our previous knowledge, and how this integration work affects our decisions. This is the pathway for estimating the Value of Information. In fact, "the value of any particular observation must derive from the potential to affect the agent's eventual physical action" (Russell and Norvig, 2016). The following example helps an understanding of how we can estimate the VOI applying Bayes' Theorem in a simulated case of hydrocarbon exploration (here extremely simplified).[4] Beside the specificity of the example, the approach is general and can be applied in different fields and for many purposes.

1.3.2. An example

Let us imagine that we are going to begin a geological/geophysical survey for exploring the subsoil of a geographical region, with the objective to discover a new oil reservoir. We can assume that our prior information supports, for instance, the positive scenario, *H,* of a possible hydrocarbon discovery in that region, if we drill a well at a given location (*H = oil*). For example, our previous geological studies could suggest that there are favorable conditions for discovering a commercial hydrocarbon accumulation, at a certain depth, in a specific geological formation. Based on that prior knowledge, we can estimate the correspondent prior

[3] In scientific practice, the expression "prior probability" refers to our state of knowledge at a time before a certain experiment.

[4] For a general and exhaustive discussion about this important theorem, I recommend the didactical approach adopted by Russell and Norvig (2016; Chapter 13, section 13.5, pp. 495-499: "Bayes' rule and its use").

probability, let us say *P(oil)*, that our well will drill a commercial oil discovery.[5]

Now, let us suppose that we acquire new geophysical data at the surface, *O*. It is reasonable to expect that this new set of observations will change the prior probability of the scenario *H* into a new *a posteriori* or *conditioned probability*. Qualitatively speaking, the new dataset *O* has value to the extent that it is likely to cause an impact on our state of knowledge and, consequently, on our action plans. Quantitatively speaking, we can apply the Information Value Theory, for estimating the value of the new information *O*. The posterior probability (conditioned by the new data) is commonly indicated with the notation *P(oil|O)*. It indicates the posterior probability of having an *oil discovery scenario*, given the set of new observations *O*.

Our problem is to quantify the value added by the new geophysical information *O*. This problem is commonly formalized using the Bayes formula, well known in statistical applications. It provides the conditioned (posterior) probability $P(H \mid O)$ for a scenario *H* after modification of unconditioned probability (a priori probability) due to a set of observations *O*:

$$P(H|O) = \frac{P(O|H) \cdot P(H)}{P(O)}. \qquad (1\text{-}1)$$

Symbols in formula (1-1) have the following meaning:

a) *P(H)* is the a priori probability (without observations *O*) that the scenario *H* is verified;

b) *P(O|H)* is the probability of the observations *O* when the scenario *H* is verified;

c) *P(O)* is a normalization factor that assures the condition that $0 \leq P(H|O) \leq 1$. Its general form is:

$$P(O) = \sum_{i=1}^{n} P(O|H_i) \cdot P(H_i)\,, \qquad (1\text{-}2)$$

where H_i is the generic scenario i^{th} in a set of *n* scenarios.

The formula 1-1 is often indicated as *Bayes' Theorem*. Its demonstration is immediate (Russell and Norvig, 2016); however, the formula is very intuitive. It says that the product of two probabilistic terms (the denominator

[5] In this example, *P(oil)* denotes a prior probability in the sense that it is unconditioned by any new data.

is just a normalization factor) gives the posterior probability: one term is the prior probability of the scenario *H*. The other one is the conditioned probability of observing the dataset *O* if the scenario *H* is verified.

Returning to our geophysical examples, for instance, I assume that our new geophysical data consist of a set of electromagnetic measurements. Controlled Source Electromagnetic (CSEM) surveying is a geophysical prospecting method that is often used for hydrocarbon exploration, thanks to its sensitivity to the presence of electrically resistive oil (or gas) reservoirs (Constable et al., 2007; Eidesmo et al., 2002). CSEM data frequently add useful information, thus we expect that the new dataset can improve our knowledge, reducing the exploration risk in that area.

As I said, the symbol H_i in formula 1-2 indicates the generic scenario. In this example, it can be the scenario of a commercial oil discovery or the scenario of a dry well. For the sake of simplicity, we can start defining the information *O* as a binary indicator, where *O* is a *positive indicator* if it increases the chance of success (for instance, a commercial discovery), and a *negative indicator* otherwise.[6] Observation *O* is the CSEM dataset, but the same discussion can be done for any other type of information (gravity, seismic, magnetic data).

Let us summarize, schematically, the terms of our example:

a. *P(H) = P(oil)* is the unconditioned (a priori) probability of drilling an oil filled reservoir. This probability can be statistically estimated based on previous theoretical geological studies, for instance.
b. *P(O|H) = P(CSEM/oil)* is the conditioned (a posteriori) probability of having a significant CSEM response (like electromagnetic data with high amplitude) if a commercial oil reservoir is effectively present at that location (where we planned to drill a well). It can be estimated, for instance, through modeling.
c. Applying Bayes' Theorem, the above probabilities allow us to estimate the posterior probability for an oil scenario conditioned by the CSEM response, *P(oil | CSEM)*, as explained in the following.

For a fixed scenario *H* (oil, non-oil …), the set of new observations *O* (that can be a "CSEM response", a "no-CSEM response", and so on) can give a true positive or a false positive. For instance, in the case that we

[6] Following the well-known medical definitions, positive and negative indicators can be further divided into "true positive", "false positive", "true negative" and "false negative". For instance, a significant CSEM response will be a true positive hydrocarbon indicator if it corresponds to a true hydrocarbon discovery. Otherwise, the same CSEM response will be a false positive if it corresponds to a dry well.

observe a significant CSEM anomalous response, the posterior probability for an oil scenario is given by the following expression of the Bayes formula:

$$P(oil|CSEM) = \frac{P(CSEM|oil)\cdot P(oil)}{P(CSEM|oil)\cdot P(oil) + P(CSEM|no\ oil)\cdot P(no\ oil)}. \quad (1\text{-}3)$$

The denominator in formula (1-3) acts as a normalization factor, in order to have $0 \leq P(oil|CSEM) \leq 1$. In this specific example, it represents the sum of the probability to have a true positive and the probability to have a false positive. An analogous formula can be written in the case of no detection of any CSEM anomalous response.

The estimation of the posterior probability through the formula 1-3, allows the calculating of the value of CSEM information (VOI_{CSEM}). It depends on how much the new information affects the exploration risk. In other words, VOI_{CSEM} depends on how the CSEM data affect the value of our hydrocarbon prospect and, finally, our decisional process (if and where to drill).

The decisional impact of the CSEM data can be easily estimated from the difference of the hydrocarbon prospect value based on the exploration risk calculated before and after acquiring the CSEM data.

If the drill cost is indicated with C, the net present prospect value is indicated with V (excluding the drilling cost) and a prior chance of success (without a CSEM response) is indicated with $P(oil)$, then the expected net present value without CSEM information is (Buland et al., 2011):

$$E(V) = P(oil)\cdot V\ . \quad (1\text{-}4)$$

Instead the expected net present value with CSEM information is

$$E(V)' = P(oil|CSEM)\cdot V\ , \quad (1\text{-}5)$$

where $P(oil|CSEM)$ is provided by formula 1-3.

Finally, the VOI of the new CSEM data is given by the estimated value of the prospect with CSEM information (formula 1-5) minus the estimated value of the prospect without CSEM information (formula 1-4), minus the cost C of the additional CSEM information:

$$VOI_{CSEM} = E(V)' - E(V) - C\ . \quad (1\text{-}6)$$

1.3.3. Revisiting the example qualitatively

Summary

The example described above could appear complicated, especially to readers who are not familiar with hydrocarbon exploration. Let us summarize and generalize the entire procedure, without using any mathematical formalism.

a) Our final goal is to calculate the Value of Information (VOI) of new geophysical data that we wish to acquire at the surface to estimate the probability to discover (by drilling an exploration well) a new oil reservoir in a certain prospect area.
b) The value of the new geophysical (electromagnetic) data depends on their impact on our decisions and, finally, on the value of our prospect. Consequently, the VOI of the data depends on how it changes the probability of an oil-discovery scenario if we decide to drill a well at a given location.
c) The Bayes formula allows the estimation of how the prior oil-probability changes into a posterior oil-probability conditioned by the new electromagnetic data.
d) After calculating the posterior oil-probability, we re-estimated the value of our "object of interest" (the drilled prospect). We simply multiplied the prospect value for the new (posterior) oil-probability.
e) Finally, we estimated the VOI of the electromagnetic data by just subtracting the new prospect value (with electromagnetic data) minus the previous prospect value (without electromagnetic data). We also subtracted the cost for acquiring the new data.

Buland et al. (2011) effectively estimated the impact of CSEM data on exploration risk analysis using the Bayesian approach introduced in this chapter. Using a huge industrial database (a real case of Big Data), the authors calculated the exploration risk modification by applying Bayes' theorem when electromagnetic (CSEM) information is taken into account and when it is properly combined with previous data (seismic, well logs, geological knowledge and so forth). The authors confirmed that the CSEM method can provide useful information improving the process of risk evaluation in exploration, increasing significantly the confidence of discovering a commercial hydrocarbon reservoir.

I performed a similar analysis, after expanding the Bayesian method to the integration of seismic, electromagnetic and gravity data (Dell'Aversana, 2016). I applied all the theoretical concepts explained in the previous section

to a real industrial dataset, obtaining positive results about the value added by electromagnetic and gravity data.

Remarks

Beyond the specificity of the geophysical applications here mentioned, the examples done help to clarify some key aspects of the Value of Information related to data integration. New data are effectively useful if they are properly combined with previous information, in order to re-define the entire frame of our knowledge. If that integration process is not properly performed, the new experimental observations risk creating confusion rather than benefits. For instance, it can happen that the new data are in conflict with the actual Earth model, and geologists and geophysicists prefer to ignore the new observations rather than modify their model. Unfortunately, this is a frequent and realistic scenario. The conflict between previous knowledge and new evidences can "complicate" the process of interpretation leading towards inappropriate decisions.

The same considerations are true in other fields that are different from geosciences, such as in medical disciplines. For instance, it is well known that a correct diagnosis can be obtained through the proper combination of all the available prior information with new evidences, including the patient's anamnesis, blood analysis, images of the body interior, and symptoms' analysis. As in the geophysical example done in the previous paragraph, we can use the same Bayesian approach for estimating how much each new medical datum affects the posterior probability of a certain medical diagnosis. The effective value of each individual piece of information depends on how much this is used in the diagnostic process. Unfortunately, for many practical reasons, full data integration often represents just an ideal scenario in medical applications, as well as in geosciences.

Applying an effective integration workflow is a general key requirement for extracting the maximum value from new and old data. For that reason, I will dedicate the next chapter to summarizing the most recent techniques, algorithms, methods and workflow of data integration.

1.4. The roadmap: from significance to value

I would like to conclude this introductive chapter by remarking that the concepts of "Value" and "Significance" are strictly linked. Indeed, information has scarce value if we use it without understanding its significance. This can be considered the basic assumption of this book, and it can be named "*the Semantic Principle of Information Value*". This is a crucial point and it needs to be clarified from the beginning.

In the example discussed in the previous section (1.3.2), I made the implicit assumption that the new set of data, *O*, represents "perfect information". This means that it is a clear (unambiguous) indicator about the presence or the absence of a hydrocarbon reservoir in the subsoil, at a certain location and at a well-defined depth. Furthermore, I assumed that we understood completely its geophysical significance in terms of geological implications on the Earth model. That is never true in the real world, especially in the practice of interpretative disciplines such as geosciences and diagnostic medicine. In general, we make decisions based on our *interpretation* of information that is affected by uncertainties, rather than on perfect data. In other words, our behavior and decisions are driven by the *significance* that we assign to (or that we extract from) information (measurements, processed data, previous knowledge …). For that reason, in the following chapters, I will dedicate part of the discussion to clarifying the fundamental concept of the "significance of information", and how this can be progressively improved through a complex workflow. Fig. 1-6 shows a schematic view of the main steps of this workflow. Besides integration, the Value of Information is increased through the combination of additional methods. These involve *Data Fusion*, *Multimodal data analysis* in different physical domains, the *Audio-Video display* of data through advanced techniques of "sonification", *Multimedia Machine Learning* and *hybrid methods* of data analysis.

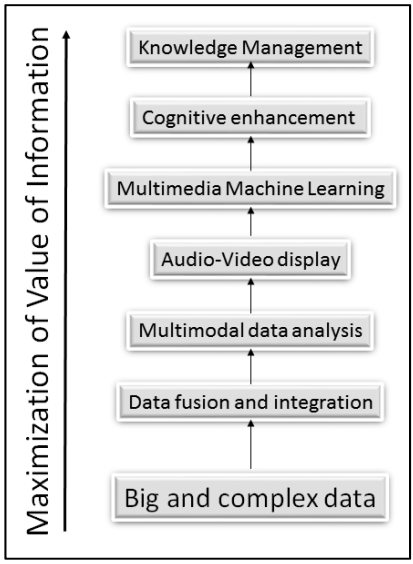

Fig. 1-6. The roadmap for maximization of the Value of Information of big and complex data.

Finally, inferring significance from complex information is a process where human cognition shows its strength, especially in the case of teamwork. Consequently, the workflow of data value maximization must take into account crucial human factors, including individual as well as interpersonal and social aspects. In the final part of the book, I will discuss how modern neurosciences, cognitive disciplines and knowledge management approaches can contribute to the process of information value maximization.

References

1. Ahmed, S. Ejaz (ed.), 2017. *Big and Complex Data Analysis. Methodologies and Applications*. Springer.
2. Buland, A., Løseth, L.O., Becht, A., Roudot, M., and Røsten, T., 2011. The value of CSEM in exploration. *First Break*, v. 29, 69-76.
3. Cagniard, L., 1953. Basic theory of the magneto-telluric method of geophysical prospecting. *Geophysics*, 18, 605-635.
4. Colombo, D., McNeice, G., Raterman, N., Turkoglu, E., and Sandoval-Curiel, E., 2014. Massive integration of 3D EM, gravity and seismic data

for deepwater subsalt imaging in the Red Sea. Exp. Abstracts, SEG, 2014.
5. Constable, S., and Srnka, L., 2007. An introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration. *Geophysics*, 72 (2), WA3-WA12.
6. Dell'Aversana, P., 2016. The value of integration in geophysics. Applications to electromagnetic and gravity data, Expanded abstracts of EAGE Conference and Exhibition.
7. Eidesmo, T., Ellingsrud, S., MacGregor, L.M, Constable, S., Sinha, M.C., Johansen, S.E., Kong, F.N., and Westerdahl, H., 2002. Sea bed logging (SBL), a new method for remote and direct identification of hydrocarbon filled layers in deepwater areas. *First Break*, 20, 144-152.
8. Elgendy, N., and Elragal, A., 2014. Big Data Analytics: A Literature Review Paper. Conference Paper in Lecture Notes in Computer Science. August 2014 DOI: 10.1007/978-3-319-08976-8_16.
9. Russell, S., and Norvig, P., 2016. *Artificial Intelligence: A Modern approach*, Global Edition. Pearson Education, Inc., publishing as Prentice Hall.
10. Stone, J.V., 2013. *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. England: Sebtel Press.
11. Tarantola, A., 2005. *Inverse Problem Theory*. SIAM. ISBN: 978-0898715729.
12. Weaver, W., 1948. Science and complexity, *American Scientist*.

CHAPTER TWO

INTEGRATION OF INFORMATION

Abstract

This chapter is motivated by the assumption that there is a positive correlation between data integration, significance and value of information, as remarked in Chapter 1. The driving principle is that the more integrated are the different pieces of information, deeper is the meaning that we can extract from our data. Indeed, integration of heterogeneous information is crucial in many scientific areas. For instance, in geosciences, an integrated workflow often represents the optimal exploration approach. Especially in complex geological settings, combining complementary methodologies provides results better than using single prospecting techniques. For similar reasons, data, model and image fusion are essential in medical sciences. For example, a robust diagnosis can be obtained through the combination of complementary imaging techniques, blood analysis, accurate study of the patient's anamnesis and so forth. Despite the intrinsic differences, the process by which scientists transform knowledge from data space to model space shows similarities and analogies in many different disciplines. For instance, both medical and geophysical imaging techniques are often based on the same principia of mathematical inversion of multi-source data. Consequently, the processes of image fusion and model integration are also based on similar criteria in Earth and Health sciences. In this chapter, I discuss analogies and similarities between these different disciplines. I assume that a comparative analysis of imaging methods and integration workflows used in medicine and in geophysics can illuminate both fields. I introduce the general aspects of data/image fusion and integration of heterogeneous information. Quantitative integration is strongly based on mathematical inversion. Consequently, I recall the basic concepts of linear and non-linear inversion. Finally, I show how integration represents a fundamental step in the process of data value maximization.

Keywords: Integration, data fusion, image fusion, significance, inversion, joint inversion, geophysical imaging, medical imaging.