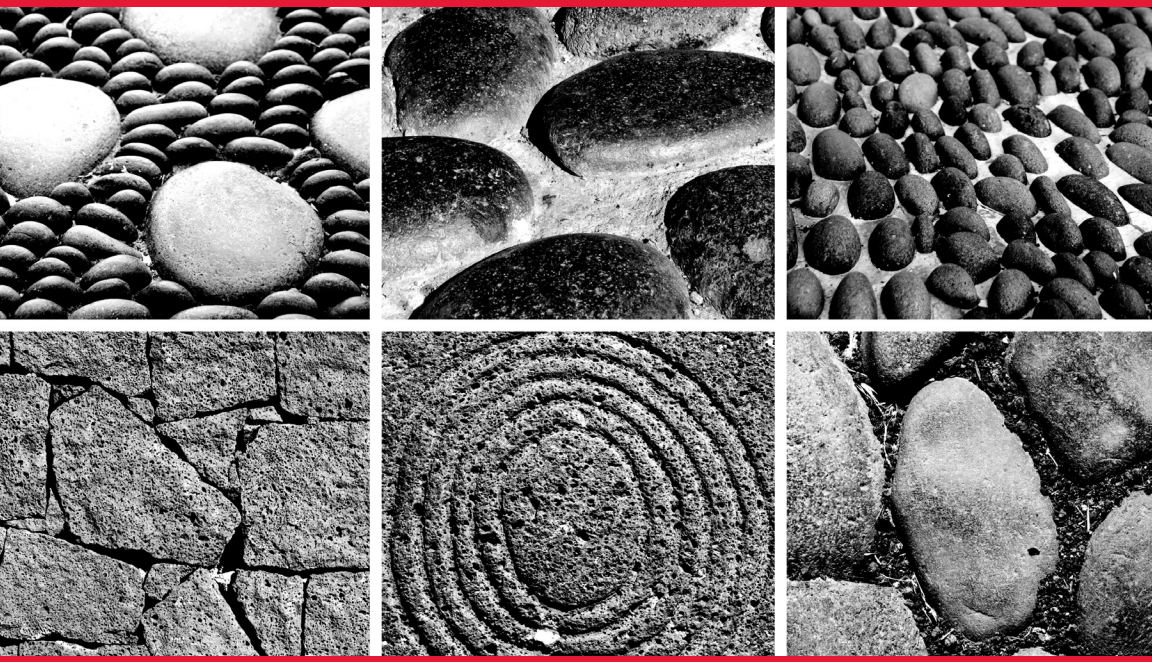
O'REILLY®

Data Preparation in the Big Data Era

Best Practices for Data Integration



Federico Castanedo

SAN JOSE

LONDON

NEW YORK

SINGAPORE

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Data Preparation in the Big Data Era

Best Practices for Data Integration

Federico Castanedo

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data Preparation in the Big Data Era
by Federico Castanedo

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Shannon Cutt
Production Editor: Dan Fauxsmith

Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

August 2015: First Edition

Revision History for the First Edition
2015-08-27: First Release
2015-11-04: Second Release

978-1-491-93895-9

[LSI]

Table of Contents

CHAPTER 1

Data Preparation in the Era of Big Data

Introduction

Preparing and cleaning data for any kind of analysis is notoriously costly, time consuming, and prone to error, with conventional estimates holding that 80% of the total time spent on analysis is spent on data preparation.[1] Accelerating investment in big data analytics—$44 billion total in 2014 alone, according to Gartner—elevates the stakes for successfully preparing data from many sources for use in data analysis.

Substantial ROI gains can be realized by modernizing the techniques and tools enterprises employ in cleaning, combining, and transforming data. This report introduces the business benefits of big data and analyzes the issues that organizations face today with traditional data preparation and integration. It also introduces the need for a new approach that scales, known as data curation, and discusses *how* to deliver the results.

This report will cover the following key topics:

- Starting with the business question
- Understanding your data
- Selecting the data to use
- Analyzing your current data strategy
- Assessing alternative ETL and data curation products

1 T. Dasu and T. Johnson, "Exploratory Data Mining and Cleaning," *Wiley-IEEE* (2003).

- Delivering the results

Starting with the Business Question

What are you aiming and analyzing for, exactly?

We are currently living in the big data era, with huge business opportunities and challenges for every industry. Data is growing at an exponential rate worldwide: in 2016, global Internet traffic will reach 90 exabytes per month, according to a recent Cisco report.[2]

The ability to manage and analyze an unprecedented amount of data will be the key to success for every industry. Data-driven companies like Google, Amazon, Facebook, and LinkedIn have demonstrated a superior position against their competitors. It is well-known that Google based most of their success on their available data, and as they mention in a paper published in 2009[3]: "*We don't have better algorithms. We just have more data*". It has also been reported that data-driven companies can deliver profit gains that are on average 6% higher than their competitors.[4]

To exploit the benefits of a big data strategy, a key question is *how to translate data into useful knowledge*. To meet this challenge, a company needs to have a clear picture of the strategic knowledge assets, such as their area of expertise, core competencies, and intellectual property. Having a clear picture of the business model and the relationships with distributors, suppliers, and customers is extremely useful in order to design a tactical and strategic decision-making process. The true potential value of big data is only gained when placed in a *business* context, where data analysis drives better decisions—otherwise it's just data.

Which Questions to Answer

In any big data strategy, technology should be a facilitator, not the goal, and should help answer business questions such as: Are we making money with these promotions? Can we stop fraud by better

2 "The Zettabyte Era—Trends and Analysis".

3 Alon Havely, Peter Norvig, and Fernando Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* (2009).

4 Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution," *Harvard Business Review* (October 2012).

using this novel approach? Can we recommend similar products to our customers? Can we improve our sales if we wish our customers a happy birthday? What does data mean in terms of business? And so on.

Critical thinking must be used to determine what business problem you want to solve or which questions you wish to answer. As an example, you should have clear and precise answers for the following general questions: Why are we doing this? What are we trying to achieve? How are we going to measure the success or failure of this project?

Articulate Your Goals

There is a false belief that only big companies can obtain benefits from developing a big data strategy. Since data is being generated so fast, and at an exponential rate, any small- or medium-sized enterprise will gain a competitive advantage by basing their business decisions on data-driven products.

However, it is extremely important to articulate clear goals and business objectives from the very beginning. Implementing a well-defined data strategy allows companies to achieve several benefits, such as having a better understanding of their customer base and business dynamics. This investment produces rewarding returns in terms of customer satisfaction, increases in revenues, and cost reduction. Each data strategy should be aligned with tactical and strategic objectives. For example, in the short term, the goal may be to increase the user base and in the mid/long term to increase revenues. In addition to setting goals, it's also important to optimize the appropriate key performance indicator (KPI) at each stage of the strategy. In any big data strategy, starting the implementation by defining the business problem you want to solve is what matters.

Gain Insight

The data you analyze should support business operations and help generate decisions in the company. Any results should be integrated seamlessly with the existing business workflows, and will only be valuable if managers and frontline employees understand and use those results accordingly.

Here are four steps for any company to gain specific insights into their business problems:

1. Start with a business question. For example, if we change the size of our product, will this result in an increase in sales?
2. Come up with a hypothesis. Following our example above, you might hypothesize: a smaller size may increase revenues.
3. Perform an exhaustive analysis of the impact of your decision, before you make it. Gather data using various methods, including controlled and double-blind experiments, and A/B testing.
4. Draw conclusions to your business question, based on the outcome of your experiments and analysis; use these conclusions to aid in your business decisions.

Data silos

One challenge that some companies may face in implementing a big data strategy is the existence of *data silos* among different areas of the company. When data silos are present, your business's data is distributed among the different silos, without communication and interfaces between them.

As part of your big data strategy, you should plan to integrate data projects into a coherent and unified view and, even more importantly, avoid (as much as possible) moving data from one place to another.

In a big data project, input data sources can come from different domains, not only from traditional transactions and social network data, and it is necessary to combine or fuse them. In order to successfully combine your data, it's important to first understand it, and your goals.

Data lakes

Until you have a solid grasp on the business purpose of your data, you can store it in a *data lake*. A data lake is a storage repository that holds raw input data, where it can be kept until your company's goals are clear. An important drawback of data lakes is the generation of duplicate information, and the necessity of dealing with the data variety problem in order to perform correct data integration. Data variety, together with velocity and volume, is one of the "three V's" of big data characteristics. *Data variety* refers to the number of distinct types of data sources. Since the same information can be stored with different unique identifiers in each data source, it becomes extremely difficult to identify similar data.

Understanding Your Data

While "big data" has become a buzzword, the term "data" is actually very broad and general, so it's useful to employ more specific terms, like: raw data, technically-correct data, consistent data, tidy data, aggregated or compressed data, and formatted data—all terms we'll define in this section.

Raw data refers to the data as it comes in. For example, if files are the source of your data, you may find the files have inconsistent elements—they may lack headers, contain wrong data types (e.g., numeric values stored as strings), missed values, wrong category labels, unknown character encoding, etc. Without doing some sort of data preprocessing, it is impossible to use this type of data directly in a data analysis environment or language.

When errors in raw data are fixed, the data is considered to be *technically correct*. Data that is technically correct generally means that each variable is stored using the same *data type,* which adequately represents the real-world domain. But, that does not mean that all of the values are error-free or complete. The next level in the data preparation pipeline is having *consistent data*, where errors are fixed, and unknown values imputed.

When data is consistent and ready for analysis, it is usually called *tidy data. Tidy datasets* are easy to manipulate and understand; they have a specific structure where each variable is saved in its own column, each observation is saved in its own row, and each type of observational unit forms a table.[5]

It is also common to *aggregate* or *compress tidy data* for use in data analysis. This means that the amount of historical data is reduced significantly. Finally, the results obtained from the analysis are provided in *formatted data*.

It is a good practice to store the input data at each different phase: (1) raw, (2) technically correct, (3) consistent/tidy datasets, (4) aggregated, and (5) formatted. That way, it will be easy to modify the data process in each phase, as needed, and minimize the impact on the other phases.

5 Hadley Wickham, "Tidy Data," *Journal of Statistical Software* 59, issue 10 (September 2014).

It is also important to know the *source* of the data at each phase, and which department owns the data or has responsibility for its management.

Selecting the Data to Use

Most machine learning and data analysis techniques assume that data is in an appropriate state for doing the analysis. However this situation is very rare—raw data usually comes in with errors, such as incorrect labels and inconsistent formatting, that make it necessary to *prepare* the data. Data preparation should be considered an *automated phase* that can be executed in a reproducible manner.

If your input data is in file format, it is important to consider *character encoding* issues and ensure that all of the input files have the same encoding, and that it's legible by the processing machine. Character encoding defines how to translate each character of a given alphabet into a sequence of computer bytes. Character encoding is set by default in the operating system and is defined in the locale settings. Common encoding formats, for example, are UTF-8 and latin1.

Data Preparation Methods

Depending on the *type* of your input data, you can use different methods to prepare it for analysis.

For date-time data, it is common to use POSIX formats and store the value as the number of seconds that have passed since January 1st, 1970 00:00:00. This format facilitates computations by directly subtracting or adding the values. Converting input dates into a standard format is not always trivial, because data can be described in many different ways. For instance, July 15 of 2015, 2015/15/07, or 15/07/2015 may refer to the same date.

In the case of categorical variables, the work of classifying dirty input text into categorical variables is known as *coding*. String data are one of the most difficult data types in which to detect errors or inconsistencies in the values. Most of the times, this data comes from human input, which easily introduces inconsistencies. Techniques to deal with string inconsistencies are known as *string normalization* or *approximate string matching*.

On the one hand, *string normalization* techniques transform a variety of strings to a common and smaller set of string values. These techniques involve two phases: (1) finding a pattern in the string, usually by means of regular expressions, and (2) replacing one pattern with another. As an example, consider functions to remove extra white spaces in strings.

On the other hand, *approximate string matching* techniques are based on a distance metric between strings that measures how different two strings are. From a mathematical point of view, string metrics often do not follow the demands required from a distance function. As an example, string metrics with zero distance does not necessarily mean that strings are the same, like in the q-gram distance. One of the most common distances is the generalized Levenshtein distance, which gives the minimal number of insertions, deletions, and substitutions needed to transform one string into another. Other distance functions include Demareu-Levenshtein, the longest common substring, the q-gram distance, the cosine distance, the jaccard distance, and the Jaro-Winkler distance. For more details about approximate string matching, please refer to Boytsov[6] and Navarro[7].

Analyzing Your Current Data Strategy

When data is ready for statistical analysis, it is known as *consistent data.* To achieve consistent data, missing values, special values, errors, and outliers must be removed, corrected, or imputed. Keep in mind that data-cleaning actions, like imputation or outlier handling, most likely affect the results of the data analysis, so these efforts should be handled correctly. Ideally, you can solve errors by using the expertise of *domain experts*, who have real-world knowledge about the data and its context.

Data consistency can be divided into three types:

1. In-record consistency
2. Cross-record consistency

6 L. Boytsov, "Indexing methods for approximate dictionary searching: comparative analyses," *ACM Journal of Experimental Algorithmics* 16, 1-88 (2011).

7 G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys* 33, 31-88 (2001).

3. Cross-data-set consistency

In-record consistency means that no contradictory information is stored in a single record; *cross-record consistency* means that statistical summaries of different variables do not conflict among them, and *cross-data-set consistency* indicates that the dataset being analyzed is consistent with other datasets of the same domain.

Missing Values

Missing values (known as NA) are one of the most basic inconsistencies. Some data analysis methods can deal with NAs, while others may fail when the data has missing input values, or may confuse a missing value with a default category.[8]

NAs are commonly confused with an unknown category; however, these are two different ideas. An NA value states that the information is not available in the dataset, whereas an unknown value indicates that the information is *in* the dataset but it is unknown. If the records may have an unknown category, this should not be confused with the NA values. A simple approach to deal with NAs is to ignore the records that contain them. When the ratio of NAs versus all of the data is high, it is better to use *imputation techniques*.

Imputation Techniques

Imputation techniques correct NAs by estimating values based on other information. The most basic imputation method is to determine the mean of the observed values, or any other measure of centrality. Another method is known as *ratio imputation*, where the estimate X_i is given by an average ratio between x and a covariate y: $X_i = Ry_i$. This is commonly computed as the sum of observed x values, divided by the sum of corresponding y values. It has the property that $x = 0$ when $y = 0$, which is in general not guaranteed in linear regression.

Generalized linear regression models can also be used as an imputation method. In this case, missing values are estimated by using a linear regression from known variables.

8 Maytal Saar-Tsechansky and Foster Provost, "Handling Missing Values when Applying Classification Models," *Journal of Machine Learning Research* 8, 1625–1657 (2007).

Hot deck imputation is a technique that replaces NAs by copying values from similar records in the dataset. It can be applied to numerical or categorical records. A critical decision in using hot deck imputation is how to select similar records.

Here are a few methods for selecting similar records:

- Randomly select one value from the known ones.
- Sorted selection, where the missing value is selected based on the closest value of one or more known auxiliary variables.
- Nearest-neighbor imputation with a specific distance function that computes a measure of similarity between records. A missing value is imputed by finding the nearest or k-nearest records (K-NN). In the case of K-NN, if the missing value is categorical, the level with the higher frequency is chosen, and if it is numerical, the mean would be the value usually taken.

Inconsistencies and outliers

There may be also some obvious inconsistencies in the data, like negative age values. These kind of inconsistencies are easy to detect and fix by using a set of user-defined rules or constraints. However, as the number of variables increases (i.e., high dimensional spaces), the number of rules may increase rapidly, and it may be beneficial to have an automated mechanism to generate them. Furthermore, multivariate rules may be interconnected by common variables, and deciding which variable causes an inconsistency may be difficult.

Sometimes rules can be interconnected, and it is necessary to make a decision about which interconnected inconsistencies should be solved. The principle of Fellegi and Holt[9] minimizes the number of fields being altered—this approach makes sense if the errors occur relatively few times and randomly across variables. Other common inconsistencies for numeric variables are those having special values, such as infinite and Not a Number (NaN). Note: these "not-real" numbers should also be removed before data analysis.

Outliers also require special attention. In general, outliers are very informative because they can indicate a special situation or an error.

9 I.P. Fellegi and D. Holt, "A systematic approach to automatic edit and imputation," *Journal of the American Statistical Association* 71, 17–35 (1976).

For a good overview about outliers, check-out the work of Barnett and Lewis[10] and Hawkins[11].

Whether or not outliers should remain in your data depends on the goal of your analysis. For instance, if you are looking for patterns in the data (like fraud-detection systems), outliers *should* be included and identified accordingly. In other cases, if we are providing some historical analysis, they may be removed to avoid introducing noise. In unimodal and symmetrically distributed data, Tukey's box-and-whisker method is the common technique to detect and visualize outliers. In Tukey's method, outliers are defined as those values larger than each whisker. Each whisker is defined by adding 1.5 times the interquartile range to the third quartile and rounding to the nearest lower observation. This method fails when the distribution of data is skewed, as in exponential or log-normal distributions. One workaround is to transform the data using a logarithm or square root transformation, or use a method that takes the skew into consideration, such as the Hiridoglou and Berthelot method for positive observations.[12]

All of the above methods fix inconsistent observations by modifying invalid values in a record, using information from valid values. Sometimes the cause of errors or inconsistencies in the data can be solved automatically with enough certainty, but there are several cases where it wouldn't be so easy and more advanced methods are required.

Assessing Alternative ETL and Data Curation Products

Extract-Transform-Load (ETL) was the name coined for the first-generation data integration systems. ETL products are used to combine data into a common data warehouse. They have evolved into data curation products by introducing data cleaning phases as well.

10 V. Barnett and T. Lewis, *Outliers in statistical data* (New York: Wiley, 1994).

11 D.M. Hawkins. *Identification of outliers. Monographs on applied probability and statistics* (Chapman and Hall, 1980).

12 M.A. Hiridoglou and J.M. Berthelot, "Statistical editing and imputation for periodic business surveys," *Survey methodology* 12(1), 73–83 (1986).

Data curation involves data cleaning, schema definition/mapping, and entity matching. As mentioned earlier, the process of data cleaning transforms raw data into consistent data that can then be analyzed. *Schema definition/mapping* is the process of making connections among data attributes and features. *Entity matching* is the task of finding different records in the data sources that refer to the same entity. Entity matching is essential when data from multiple sources are integrated, because it allows you to remove duplicate records.

Manual data curation is not an easy or feasible task, since companies usually have hundreds of databases and many thousands of tables. Furthermore, the increasing amount of data being stored introduces scalability problems for doing data curation. Problems also arise when companies acquire other companies, and the same information is stored using different schemas. Therefore, a key problem is often *how* to deal with the data cleaning and curation problem, cost effectively and at large scale.

Crowd-Sourcing Data Curation

Users or domain experts have been involved in the data curation problem in different scenarios.

In early 2008, Facebook launched a tool called Translations—allowing social network users to translate their site into different languages. In doing so, Facebook leveraged their users as a type of human crowdsourcing project to do the hard work of translating the site into several languages, and filed a patent named "*Hybrid, offline/online speech translation system*" describing and protecting their strategy. Twitter also followed a similar approach and relied on volunteers to translate their site.

At LinkedIn, they followed a strategy named "*Data Jujitsu*," to solve data cleaning/curation problems, among others. For instance, to match the employer names of its users—LinkedIn provided users with some features, like type-ahead completion, and asking for the company's ticker symbol or website. This was opposed to leaving a blank text box for users to type in their employer's name, which would generate several varying responses for the same company (e.g., I.B.M or IBM). This was a clever and easy approach to solve a specific data curation problem.

The Crowder research project from Berkeley AMPlab is a hybrid human-machine approach to solve the entity resolution problem.

They developed fast algorithms to detect similar entities and exploit transitivity to reduce the crowd cost required to examine similar candidate pairs.

The SampleClean research project is an extension of the Crowder project, created to deal with large databases. Instead of cleaning the full data sources, it only cleans a sample of the full dataset. Therefore, it provides more accurate results than the original dirty data, without the overhead of cleaning *all* of the data.

For more details on projects using crowdsourced data processing, check out http://www.cs.berkeley.edu/~jnwang/crowd_paper.html.

Data curation will require more efforts in the future because there is a growing interest in integrating structured business data with semi-structured and unstructured data from web pages, time series, etc. Therefore, the *data variety* characteristic of big data will introduce many new challenges.

One Commercial Solution

Tamr, a commercial product focused on the data curation problem at scale, attempts to solve the *variety* issue of big data. Tamr's input data sources can reside in HDFS, CSV files, or relational databases. After reading the input sources, Tamr can generate the schemas of the sources and curate the data. Schema mapping in Tamr takes advantage of metadata features such as attribute name, data type, and constraints, as well as statistical profiles of values or value length. Thus, the product can create a catalog of all data sources spread out across the company and helps users to understand and unify their data. Tamr's idea is to automate the data curation process as much as possible by using machine learning and statistical methods, and only asks the domain expert for input in the cases where it is not clear how to fix the problem. The system also allows the user to define a specific threshold for each inconsistency that requires human intervention.

Entity matching

Entity matching, or *deduplicating* records, is a complex task. The naive approach has a quadratic time complexity, because it needs to check among all possible pairs, which does not scale to large data. Tamr uses a proprietary *blocking technique* called "data binning" to approach the problem. A *blocking technique* divides data into K par-

titions, and the deduplication task takes place independently in each partition. This method cuts the amount of data by a factor of K. However, this approach should be handled carefully—because data is often dirty, it may result in a low recall and many false positives of deduplicated results. The data binning technique generalizes the blocking principle by correlating matching records to all possible ways of partitioning the data. This binning process is linear in the size of the input data (one or single pass algorithm), so it scales.

For those records that are classified as unknown (it may be duplicated or not), Tamr employs an active learning approach by involving domain experts to clarify duplicated candidates and ensure correct classification.

As an example, by using Tamr, a direct business benefit for large companies is to create a tidy catalog of suppliers and get insight into who exactly they are paying for similar items and the terms, so they can compare prices across several providers and optimize costs.

In general, it is a good practice to decouple the transformation or cleaning actions with the dataset where these actions are applied. Unfortunately this will not always be the standard procedure. In general, an analyst or data programmer writes long scripts with several functions to detect and correct inconsistencies. One of the benefits of the Tamr solution is that the inconsistencies are detected without the programming effort, and are decoupled from the input data.

Delivering the Results

Sooner or later the curated data will be consumed and analyzed, often in the form of visualizations. Since the data will be shared with people across different roles, it is necessary to know your audience when considering how to deliver the results. The main idea is that different users require different views of the data.

In order to deliver valuable results, the relevant questions to answer in the analysis phase are: How will the results from the analysis be consumed?, Which insights from this problem can be applied to other problems?, and Are the business users ready and trained to consume the analysis?

Businesses can deal with close approximations before they have an exact result, and most of the time it is enough to have an approxi-

mation when making a decision. So, as a general rule for analyzing data, it is better to be approximately right than precisely wrong.

Different ways in which the results are shared depend on the audience. For example, business users may prefer a memo, while C-level managers are generally interested in visualizations that explain clearly the business insight. For more information on data visualization, check out Wong[13] and Yau[14].

A common business concern at organizations that already have a big data analytics strategy is how to reduce the time between receiving (dirty and messy) data to grasping insights that can translate into action. This report covered some of the best practices for reducing the delay in implementing actionable insights; with these tools and techniques in mind, companies are better positioned to rapidly translate big data into big decisions.

13 Dona M. Wong, *The Wall Street Journal Guide to Information Graphics: The Dos and Dont's of Presenting Data, Facts, and Figures* (W.W. Norton & Company, 2013).

14 Nathan Yau, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics* (Wiley, 2011).

About the Author

Federico Castanedo is the Chief Data Scientist at WiseAthena.com, where he analyzes massive amounts of data using machine learning techniques. For more than a decade, he has been involved in projects related to data analysis in academia and industry. He has published several scientific papers about data fusion techniques, visual sensor networks, and machine learning. He holds a Ph.D. on Artificial Intelligence from the University Carlos III of Madrid and has also been a visiting researcher at Stanford University.