MINI-PAPER

Big Data Terminology—Key to Predictive Analytics Success

by Mark E. Johnson
Dept. of Statistics, Univ. Cent. Florida

Abstract

With all of the hype surrounding Big Data, Business Intelligence, and Predictive Analytics (with the Statistics stepchild lurking in the background), quality managers and engineers who wish to get involved in the area may be quickly dismayed by the terminology in use by the various participants. Singular concepts may have multiple names depending on the discipline or problem origin (business analytics, machine learning, neural networks, nonlinear regression, artificial intelligence, and so forth). Hence, there is a pressing need to develop a coherent and comprehensive standardized vocabulary. Subcommittee One of ISO TC69 is currently developing such a terminology standard to reside in the ISO 3534 series. In addition to the technical statistical-type terms, it could also include a discussion of some of the software facilities in use in dealing with massive data sets (HADOOP, Tableau, etc.). A benefit of this future standard is to shorten the learning curve for a Big Data hopeful. This paper describes the initial steps in addressing the terminology challenges with Big Data and offers some descriptions of forthcoming products to assist practitioners eager to plunge into this area.

Introduction

Big Data and predictive analytics are at the forefront of discussions involving the state of the statistics profession and its future prospects. Many statisticians who formerly made a living off the Six Sigma juggernaut are gravitating to Big Data applications for the usual reason—that is where the money is! Of course, to take advantage of industry thirst for consultants who are well-versed in the area, considerable training may be entailed, as some of the old weapon standby's (think the seven basic tools) are not exactly ready for Big Data prime time. One of the first barriers to entry into this area is the terminology associated with the tools, hardware and software associated with extra actionable information from massive data sets. It is not just the statistical terms—many of which have different aliases depending on the application area, but also many other computer and information technology lingos and proper names that are emerging in the Big Data world. The following is a potpourri list of terms, packages or buzzwords that could be encountered in short order when looking into the Big Data world:

> Apache Accumulo, Apache Hadoop, Big Memory, Cask, Cloudera, Data Analytics, data lifecycle, Distributed computing, federated databases, Hortonworks, HPCC, IaaS, Internet of Things, Map Reduce, NoSQL, R, Rattle, schema on read, Spark, SQL, Sqrrl, Tableaux, Talend, Tranreality gaming, Tuple space, Unstructured data, plus many others—not an exhaustive list by any means.

With a bit more exploring, an A-to-Z list could probably be concocted. Is it even worth sorting through these names and determining those that are worth pursuing? Does one need to go back to school to participate in the Big Data enterprise? If one asks about the job market associated with Big Data, then for the avariciously inclined, Bill Snyder of InfoWorld posted the following salaries for analysts having expertise in the following Big Data packages or software:

MapReduce: $127,315
Cloudera: $126,816
HBase: $126,369
Pig: $124,563
Flume: $123,186
Hadoop: $121,313
Hive: $120,873
Zookeeper: $118,567
Data Architect: $118,104

These salaries certainly get the attention of our graduate and undergraduate students. Amazingly, there are reports of upwards of 200,000 positions in the Big Data industry! Such information has not gone unnoticed by university administrators. Our university is no exception having added new faculty positions in big data for the past few years. Also, the sponsored research group offered a Big Data Grants Day last fall to foster the search for external funding and to encourage collaborations with local industries and the government sector who had large data set yearning to be analyzed. No doubt other universities are also on such quests.

Knowing that packages flourish and wane and programming for others may not be a long-term objective, a broader viewpoint could be sought. Here terminology plays a role, although admittedly it is perceived as rather dry.

As noted, terminology entails not only the lexicon of statistics but also the areas of information technology include the hardware possibilities involved in storage and retrieval ("the cloud") and software tools (e.g., commercial packages Hadoop and Tableaux) and techniques (distributed computing and methods for handling unstructured data). Hence, getting started in Big Data poses formidable challenges for the traditional Six Sigma practitioner or the recent/current statistics student whose faculty may also be struggling to keep up with the changing times.

The bottom line is that there is no magic elixir available to conquer the wall of Big Data expertise. A first step is confront terminology with a healthy respect for the ecosystem in which the data resides. In this paper some efforts underway will be described. These include work within ISO TC69 SC1 (International Standards Organization Technical Committee 69 on Application of Statistical Methods, Sub-Committee 1 on Terminology and Nomenclature), the Ad Hoc Working Group 7 within TC69 on Big Data, and finally the joint efforts of TC69 with the IEC/ISO JTC1 WG9 led by Wo Chang of NIST. Nancy Grady of Scientific Applications, Inc. (SAI) has also been extremely helpful in these efforts.

Section 2 of this paper addresses the case for consideration of terminology and attempts to describe Big Data, as a consensus definition is illusive. The next three sections of this paper deal with three groups actively involved in this effort. Section 2 describes the terminology efforts within ISO TC69, SC1 in particular with the somewhat narrow focus on predictive analysis terms. A broader group, namely the ISO TC69 AdHoc Group 7 is covered in Section 3. This group at the recent June 2016 London meeting tackled a gap analysis of TC69 standards versus Big Data needs. Finally, the status of the joint efforts of TC69 with IEC/ISO JTC1 WG9 will be highlighted in Section 4. Under the sponsorship of NIST, this group has produced some excellent documents on a tentative Big Data computational ecosystem which has propelled TC69 greatly in their own progress.

The Case for Terminology and What is Big Data?

The need for terminology standards was illustrated at the Fall TCC talk upon which this paper is based by noting the following list of terms:

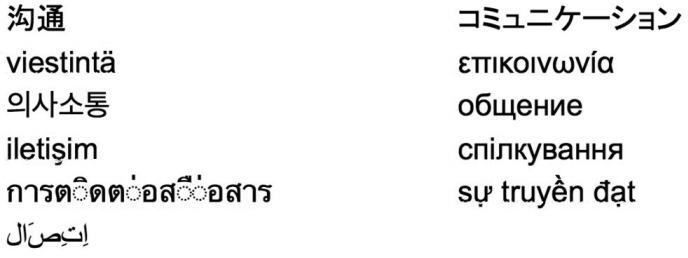

Figure 1: Terminology

In the absence of a terminology standard containing these words, one basically has a collection of seemingly unrelated gibberish. In fact, each of the above 11 items are synonyms for the same term (see the end of the paper for the answer). Terminology standards attempt to define terms in a coherent system so that each term so defined relates to a single, unique concept. "One term-one concept" is a mantra of ISO TC37 on Terminology and Other Language and Content Resources. This notion is critical in the realm of international standards in which the English writing style of a standard should be amenable to translation to other languages. Without terminology standards, the standards community ends up wasting time haggling over the wording of documents owing frequently to an inherent disagreement on terms.

Big Data poses a particular challenge for a definition as it embodies a number of concepts which have a perspective that depends on the individual user. NIST has made a valiant effort at characterizing big data with the following abstract in *NIST Big Data Interoperability Framework: Volume 1 Definitions.*

> *"Big Data is a term used to describe the new deluge of data in our networked, digitized, sensor-laden, information-driven world. While great opportunities exist with Big Data, it can overwhelm traditional technical approaches and its growth is outpacing scientific and technological advances in data analytics." (page iii)*

The NIST document goes on to say:

> *"The term Big Data has been used to describe a number of concepts, in part because several distinct aspects are consistently interacting with each other. To understand this revolution, the interplay of the following four aspects must be considered: the characteristics of the datasets, the analysis of the datasets, the performance of the systems that handle the data, and the business considerations of cost effectiveness." (page 4.)*

In light of the one term-one concept, a definition of Big Data would appear hopeless. However, a working definition, if not following the requirements of ISO TC37 is provided by this group:

> ***Big Data** consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis. (page 5)*

Listed explicitly in this definition are the so-called four "V-s": volume, variety, velocity and variability. Volume refers to the absolute size of the data set for which "big" is in the size of the beholder's storage capacity. Variety refers to diverse data types (nominal, continuous, text, etc.) from various domains and residing possibly on multiple repositories. Velocity naturally concerns the rate of data generation. The NIST document refers to Variability as the change in other characteristics. What this means is elaborated in a later section in which the variability is described as the change in data over time, including the flow rate, the format, or the composition. This is not variability in the statistical sense. A change in flow rate could necessitate an increase in devoted resources or nodes to handle a surge in volume while a format change could require a separate node for special processing. The statistical folks naturally object to variability to be used to describe non-constancy of the other V's. Our current preference is "volatility" to reflect non-constancy in the on-going generation of the data. Other "V's" not at the stature of the aforementioned ones but ought to be considered in the Big Data context include:

Veracity: accuracy of the data

Value: value of the analytics to the organization

Validity: appropriateness of the data for its intended use

The verdict is not out with respect to the ultimate set of V's, although the NIST gang of four offer a starting point. Other candidates mentioned at the Fall TCC conference for tongue-in-cheek consideration were variations on the words—verisimilitude, vapulatory and veridicous.

MINI-PAPER

Terminology for Predictive Analytics, ISO TC69 SC1

Within ISO TC69, the home for terminology is Sub-Committee One (officially ISO TC69/SC1 Terminology and symbols) for which the author is the current Chair (with term limits, this chairmanship will end in 2019). The other sub-committees in TC69 are, as follows:

SC4 Applications of statistical methods in process management
SC5 Acceptance sampling
SC6 Measurement methods and results
SC7 Applications of statistical and related techniques for the implementation of Six Sigma
SC8 Application of statistical and related methodology for new technology

Even if there were an SC on Big Data, terminology related to Big Data would be the responsibility of SC1. SC1 is responsible for the ISO 3534 series on terminology consisting of the following four parts:

ISO 3534-1 *Part 1: General statistical terms and terms used in probability*
ISO 3534-2 *Part 2: Applied statistics*
ISO 3534-3 *Part 3: Design of Experiments,* Geneva: ISO.
ISO 3534-4 *Part 4: Survey Sampling,* Geneva: ISO.

A new *Part 5: Predictive Analytics for Big Data* is under development. The structure of this document is standardized by ISO to be, as follows:

Foreword
Introduction
1 Scope
2 Normative References
3 Terms and Definitions
Annex A (informative) Concept diagrams
Annex B (informative) Methodology used to develop the vocabulary

The normative references include the other 4 parts of ISO 3534 while the challenge is to develop Section 3. The following is a rough outline as of the London meeting in 2016 of Section 3:

Initial Outline of Terms and Definitions for Predictive Analytics in Big Data

3.1 Supervised Problems
  3.1.1 types of models (response variables)
    y continuous (classical regression)
    y binary (logistic regression)
    y categorical (discriminant analysis / classification rules)
  3.1.2 types of model fitting
    least squares (including linear regression)
    decision trees (including boosting, bagging)
    neural networks (from a non-linear regression perspective)
    principal components regression

3.1.3 process of fitting
model selection
quality of fit
overfitting
3.1.4 methods for big data applications
lasso to achieve zero coefficients
kernel smoothing methods
ridge regression to control multi-collinearity
3.1.5 other considerations
partial least squares
scalability of methods
distributed computing aspects
3.2 unsupervised models
3.2.1 outlier detection (in particular, multi-dimensional)
3.2.2 cluster analysis
3.2.3 market basket analysis (graph/networks/linkages)
3.2.4 Miscellaneous

This outline is sufficient at this point to launch a new work item proposal within ISO TC69 which will lead to a call for member country experts who will ultimately write the standard and respond to comments from participating countries. This effort will proceed in parallel with the ongoing activities within the ISO TC69 AdHoc group on Big Data, described in the next section.

ISO Ad Hoc Committee 7

The present structure of ISO TC69 consists of the following six sub-committees:

SC1 Terminology and symbols
SC4 Applications of statistical methods in process management
SC5 Acceptance sampling
SC6 Measurement methods and results
SC7 Applications of statistical and related techniques for the implementation of Six Sigma
SC8 Application of statistical and related methodology for new technology

From these sub-committee names, the positioning of Big Data is obviously not apparent. Consequently, in 2015 TC69 established in an ad hoc Committee (AHC7) to investigate the future role of TC69 with respect to Big Data standards. This group has performed a gap analysis of existing statistical standards and has recognized that many of these standards apply to Big Data even if they were not originally conceived with Big Data in mind. In particular, this group concluded that all the statistical process control standards apply as control charts fundamentally handle streaming data. Likewise, the acceptance sampling standards could also serve since their standards can handle arbitrarily large populations.

Some isolated efforts for specific standards on Big Data have been suggested but have not advanced to the new work item status. Sub-committee 1 as noted previously is working on a terminology document with emphasis on predictive analytics.

MINI-PAPER

Joint effort with ISO/JTC1

Supported by NIST, the ISO/Joint Technical Committee 1 has drafted seven documents of interest to the Big Data community by providing an Interoperability Framework consisting of the following seven volumes:

Volume 1, Definitions
Volume 2, Taxonomies
Volume 3, Use Cases and General Requirements
Volume 4, Security and Privacy
Volume 5, Architectures White Paper Survey
Volume 6, Reference Architecture
Volume 7, Standards Roadmap

These tomes do a great service to the statistical community since they provide a computing ecosystem in which arguably any big data situation can apply. Also, very neatly this group has designated a placeholder for statistical contributions. A much simplified representation of their framework is given in Figure 2. A much more detailed version of this figure appears in Volume 2, page 13.

A massive amount of detail is provided in the NIST Seven Volume set and is great reading to get a feel for the manner in which IT folks address Big Data (and elaborating on the contents of the four box other than the statistical provider box). The center box in Figure 2 provides the placeholder as envisioned by our IT partners, with the statistical role to be further elucidated. The five specific statistical items in a nutshell are elaborated below:

Collection: Moving data from its repository to an accessible location (large volumes with attention to confidentiality) possibly including sampling.

Preparation/Curation: Validating the data as pulled from the repository, cleansing of obvious mistakes and duplicate records, partitioning the data for distributed computing (as required).

Analytics: Discovering value in the large data sets (e.g., correlations and trends) and establishing a structure useful for further summaries and analyses (possibly parallel, summary tables or relational data bases), addressing complexity (execution time of methods), handling real time or streaming data, human-in-the-loop discovery lifecycle.

Visualization: exploratory data presentations for understanding, explicatory views of analytical results (real-time presentation of analytics), telling the story.

Access: data export for querying, consumer analytics hosting, analytics as a service hosting.

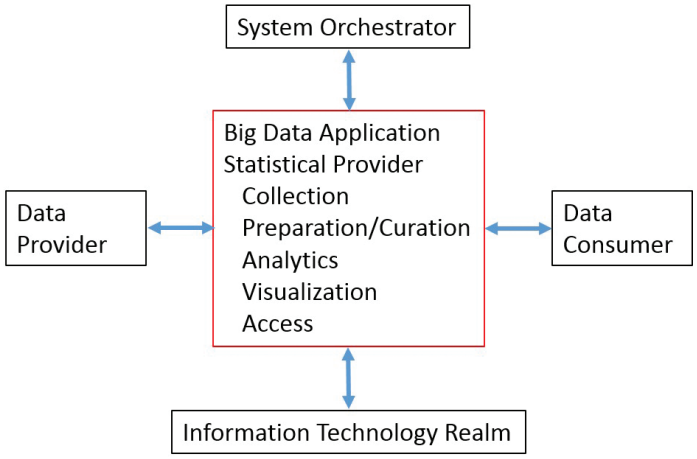

Figure 2: (Over) Simplified representation of the computing ecosystem

These descriptions are not necessarily likely to match the typical statisticians understanding of the five steps provided. *Collection* may involve some sampling from the full data set but we detect a great interest by the IT professionals to use all of the data (after all, they have gone to a great deal of trouble to generate it in their computing frameworks!). *Preparation/curation* is a bit like what most of us think of as data preparation, although we would emphasize preliminary calculations to facilitate some of our predictive analytics tools (e.g., discretizing some non-linear responses, grouping some variables such as regions rather than states, etc.). With regard to *Analytics*, we could offer considerably more than correlations and trends (not that these two are irrelevant). The five activities are placeholders rather than final prescriptions. *Visualization* is very reminiscent of the Tukey approach of exploratory and confirmatory analysis followed by implementation. Some might think that visualization would take place early in an analytical investigation but mega-millions of data points are not necessarily amenable to the tools used on small data sets. *Access* is a bit more cryptic but seems to entail implementation of the discoveries made in the full-blown process (e.g., real time scoring of future items based on the analysis with analyses to be updated efficiently). Volume 2, page 21 notes, "The access activity of the Big Data Application Provider should mirror all actions of the Data Provider, since the Data Consumer may view this system as the Data Provider for their follow-on tasks."

The good news here is that the NIST experts (Wo Chang in particular) involved with ISO/JTC1 recognize the role of statistics in this enterprise and we in turn recognize the great value in having a computation ecosystem in place to guide our own work. M. Boulanger, T. Kubiak and I are collaborating with Wo Chang using an actual large data set on health care fraud to exercise the ecosystem. Already this exercise has helped us to understand their IT ecosystem and allowed us to refine what we bring to the table for the IT professionals. We hope to complete a white paper on this topic by the end of 2016.

Conclusions

Successful Big Data applications are due to collaborative efforts of information technologists, the computer experts and statisticians. Terminology awareness is an initial barrier for the statistics community to participate effectively. As noted, much of the heavy lifting in terms of describing the computational infrastructure has been laid out by the ISO/JTC1 group with thoughtful consideration of the role of statisticians in the endeavor. Obviously, the development of international standards designed directly for Big Data problems is in its infancy. It has been argued here that terminology for predictive analytics is a necessary first start with further standards driven by continued collaborations among the interested parties.

In closing, perhaps one wonders why we simply rely on normative dictionaries to handle the terminology "problem." The Oxford English Dictionary is a recognized authority on definitions and offers as its definition of Big Data:

> Computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.

Such a definition devoid of any sense of uncertainty is unpalatable to statisticians. Another partial definition encountered in preparing the paper is attributable to

Cathy O'Neill: "'Big data' is more than one thing, but an important aspect is its use as a rhetorical device, something that can be used to deceive or mislead or overhype." We are clearly far from a consensus definition of Big Data so meanwhile we shall move ahead in terminology tackling those areas of Big Data that support standards and are in harmony with successful Big Data applications.

Acknowledgements

This paper came to be owing to a presentation that I gave at the Joint Technical Communities Conference held in Orlando, FL, October 22-23, 2015, entitled, "Big Data Terminology—Key to Predictive Analytics Success." Matthew Barsalou, editor of the ASQ Statistics Division Statistics Digest kindly invited me to write a mini-paper based on this presentation.

This was the second year of the Joint Technical Committees Conference (partially subsidized by the ASQ Statistics Division). Gordon Clark is the ASQ Statistics Division Representative to this group and has worked with this conference since its inception. Mindy Hotchkiss was instrumental in having this presentation be a part of the program.

[Answer to Section 2 riddle. Each of these expressions were Google translations of "Communication."]

References

ISO 3534-1 (2006). *Statistics—Vocabulary and symbols—Part 1: General statistical terms and terms used in probability*, Geneva: ISO.

ISO 3534-2 (2006). *Statistics—Vocabulary and symbols—Part 2: Applied statistics.*, Geneva: ISO.

ISO 3534-3 (2013). *Statistics—Vocabulary and symbols—Part 3: Design of Experiments,* Geneva: ISO.

ISO 3534-4 (2014). *Statistics—Vocabulary and symbols—Part 4: Survey Sampling,* Geneva: ISO.

NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, Draft NIST Big Data Interoperability Framework: Volume 1, Definitions. April 6, 2015.

NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, Draft NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies. April 6, 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1

William G. Hunter Award 2016: Dr. A. Blanton Godfrey

The Statistics Division of the American Society for Quality (ASQ) is pleased to announce that Dr. A. Blanton Godfrey is the recipient of the 2016 William G. Hunter Award. The William G. Hunter Award was established by the Statistics Division in 1987 to recognize the many contributions of its founding chair at promoting the use of applied statistics and statistical thinking. The attributes that characterize Bill Hunter's career - consultant, educator for practitioners, communicator, and integrator of statistical thinking into other disciplines - are used to help decide the recipient.

Blan Godfrey is a well-known to many in industry and academia as a visionary leader in applied statistics and quality. He has made impactful contributions in a wide array of application areas that include new technology development, manufacturing, product reliability and quality, and healthcare quality. Blan is currently the Joseph D. Moore Distinguished University Professor in North Carolina State University's College of Textiles where he also served from 2000 to 2014 as the dean. He was Chairman and CEO of Juran Institute, Inc. from 1987 to July 2000. From 1973 to 1987 he was a Member of the Technical Staff at AT&T Bell Laboratories, the last five years as Head of the Quality Theory & Technology Department. Blan's interests cover many areas of mathematical and applied statistics and quality management. He has had a long involvement in health care quality management. He currently serves as a Member and past Chair of the Board of Directors for the Institute for Healthcare Improvement. Blan contributed to the creation of the Malcolm Baldrige National Quality Award and served as a judge for the first three years of the award. Blan has given seminars, consulted, or taught courses in over sixty countries and his written materials have been translated, collectively, in over fifteen languages. He has personally worked with many of the top executives of leading companies throughout the world.