STATISTICS IN THE BIG DATA ERA[1]

Agostino Di Ciaccio, Giovanni Maria Giorgi

1. Introduction

It is estimated that about 90% of the currently available data have been produced over the last two years. Of these, only 0.5% is effectively analysed and used. However, this data can be a great wealth, the oil of 21st century (Sondergaard, 2011), when analysed with the right approach. In this article, we illustrate some specificities of these data and the great interest that they can represent in many fields.

In section 2 we analyse some common data sources and the big-data characteristics in various application contexts. In section 3 the relevance of the cloud computing is considered and, in section 4, some problematic aspects of big-data are reported. New challenges for the statistical analysis are considered in the section 5 and 6, suggesting some strategies.

2. The data deluge

Doug Laney (2001) defined the concept of big-data as related to three main keywords: *volume*, *velocity* and *variety*.

Volume indicates that data have a large number of units and variables. *Velocity* is needed by real time analysis of streaming data, generated for example by sensors. *Variety* indicates that data comes in all types of formats – from structured, numeric data to unstructured text documents, email, video.

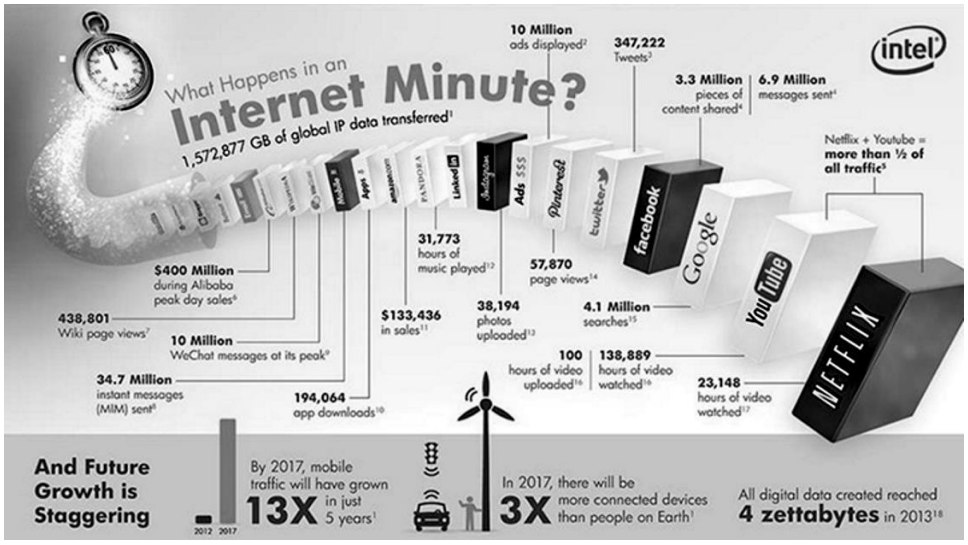
Often, big-data have a different structure than the traditional data-base and are sometimes called *date-lake*. In traditional database, the data set must be carefully designed before you can enter data. Conversely, a *data-lake* indicates a data storage repository that contains a large amount of raw data in native format, with a high level definition of what exists in the lake of data.

[1] Invited paper to the 53rd SIDES Scientific Meeting – Rome 2016.

The United Nations Economic Commission for Europe (UNECE, 2013) classified the big-data sources as:

- *Human-sourced information*: social networks, blogs, pictures, videos, search engine queries, mobile data content, etc.
- *Process mediated/transaction data*: commercial transactions, banking/ stock prices records, e-commerce, credit cards, medical records, etc.
- *Machine-generated data* (Internet of Things): sensor data (weather/ pollution, traffic, security/surveillance,…), tracking devices (GPS systems, mobile phone location, satellite images), data from computer systems (logs & web logs, …).

Figure 1 – *Infographic of big-data generated on Internet (source: intel 2014).*



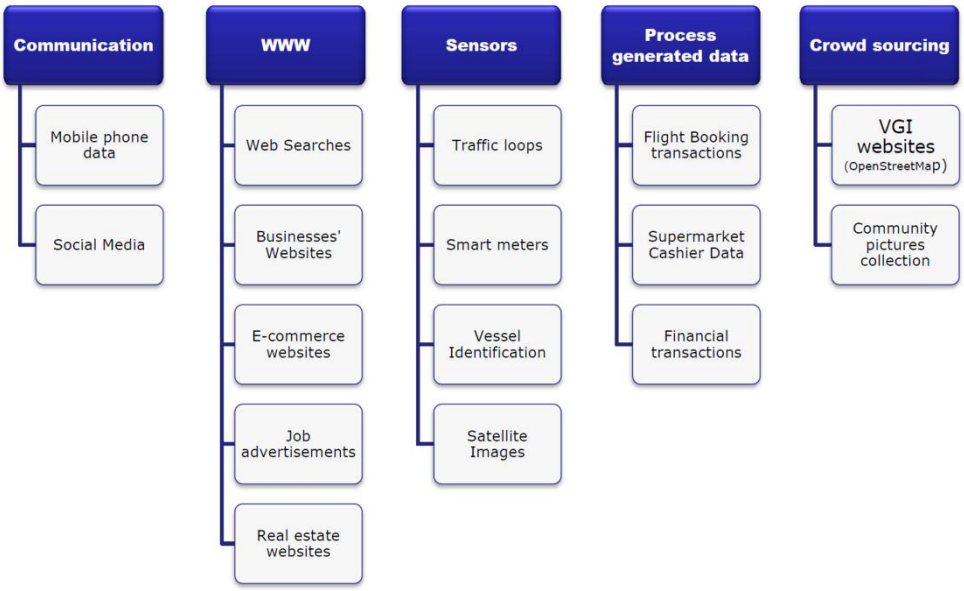
Of course, the Internet is at the moment one of the most prolific sources of data, as is shown in fig. 1. A classification of sources as suggested by EUROSTAT (Skaliotis, 2015) is proposed in fig. 2.

The characteristic of big-data is not just the size. They are often new data types, concerning people's behaviour and beliefs, new types of instruments, and new types of actors.

As pointed out in the introduction, currently we produce a huge amount of data but only a small percentage is actually analysed. In the last few years however, there has been a growing interest in the analysis of these data in various application fields. For example, the predictive power of "*social*" big-data is being used in

many fields like public health or economic development. Global Pulse, an initiative by the United Nations (www.unglobalpulse.org), tries to leverage these data for global development. The group carried out analysis of messages in social networks for several projects, as "*Understanding Immunisation Awareness And Sentiment Through Analysis Of Social Media And News Content* (2015)" or "*Analysing Social Media Conversations To Understand Public Perceptions Of Sanitation* (2014)". The next paragraphs will deepen some applications of big-data analysis.

Figure 2 – *Classification of big-data sources (Eurostat 2015).*



2.1. Textual analysis

Recent years have seen a tremendous growth of text-based data, in particular web pages, news, e-mail and social media. A characteristic of textual data is that they are generated directly by humans, rather than for example by sensors, and are therefore very useful to analyse people's opinions and preferences.

The explosive growth of text data makes very difficult to analyse this data in a timely manner. Therefore, information retrieval systems, able to identify the information quickly and accurately, are necessary. This need has led, for example, to the creation of the well known web search engines. Moreover, it is very useful to

analyse textual data, such as product reviews, forum discussions, and social media to get user's opinions.

Text constitutes unstructured data, which do not conform to well-defined patterns, and are therefore relatively complex to analyse. The optimal processing goal would be to understand the content encoded in the text, but the current technology does not yet allow a computer to accurately understand the natural language text. On the other hand, a wide variety of statistical approaches allow the analysis of textual data. For example, *recursive neural networks* have had significant successes in a number of tasks as in Socher et al. (2013) that used this technique to predict sentence sentiment, with good results.

2.2. Medicine

The large amount of genome sequencing data now make it possible to uncover the genetic markers of rare disorders and find associations between diseases and rare sequence variants.

John Craig Venter was the first to sequence the human genome, with the enormous cost of about 100 million ($). The Venter's target now is to sequence at least one million genomes and use these data, together with information on the health of the DNA donor and the results of other medical tests, to identify new methods of treatment and prevention for many diseases. This approach, named also *precision medicine,* focuses on individual differences in the genes of patients, often in combination with information about their environment, health history and lifestyle. It is a big change in our current treatments, which focuses on generic approaches for as many people as possible, instead of personalized treatments. The potential of this approach has motivated an investment of 215 million ($) from the White House.

2.3. IoT, the Internet of Things

Internet is entering a new phase of growth in which the “things” around us will be connected to the web. Internet of Things (IoT) will radically change the way we interact with our environment.

The large number of sensors, scattered in almost all sectors and connected to Internet, is beginning to result in a massive afflux of data. According to reliable estimates, by 2020, there should be 32 to 50 billion devices connected to Internet, and the volume of data generated is staggering.

Recently, Dutch Telco KPN announced, with great emphasis, that it has completed the national coverage of the Netherlands with a wireless Internet network of things. So far, KPN has inked contracts to connect 1.5 million devices. Similar IoT-networks are going up in France, Germany, South Korea, and elsewhere across the globe.

In its simplest form, the IoT, through networks and sensors integrated with cloud software, allows devices to communicate, analyse and share data.

To understand the relevance of IoT, we can list its application to some classical problems:

- *Predictive equipment maintenance*. Can be used, for example, to manage energy, predict equipment failure or detect other issues. This approach is also being used in the automobile industry, in which the same cars will be able to predict, and prevent, their failures.
- *Moving merchandise more efficiently*. This is one of the classical goals of smart transportation applications in retail, and can be obtained by a very accurate tracking and route optimization.
- *Warehouse automation and optimization*. IoT allows us to monitor sales opportunities in real time and track missed in-store sales. With IoT, we can also understand when the customer needs help or an incentive to purchase, and we can respond proactively.

IoT can create new services and business models. A key role is played by the widespread use of sensors, which can detect events or changes in quantities. Typically, data arising from sensors is in time series format and is often geotagged. Common *smart city* applications include transportation, energy grids, utilities like water, street lighting, parking etc.. For example Barcelona offers smart parking meters that operate on city-wide Wi-Fi, giving residents real-time updates on where to park and allowing them to pay with their phone.

IoT is having a key role also in the *smart home* applications: for example, the rolling shutters or heating can be controlled by a smartphone and some manufacturers already produce Internet-enabled appliances. Also the home security system can allow Internet-based monitoring of the environments and alarms.

3. Big-data on the Cloud

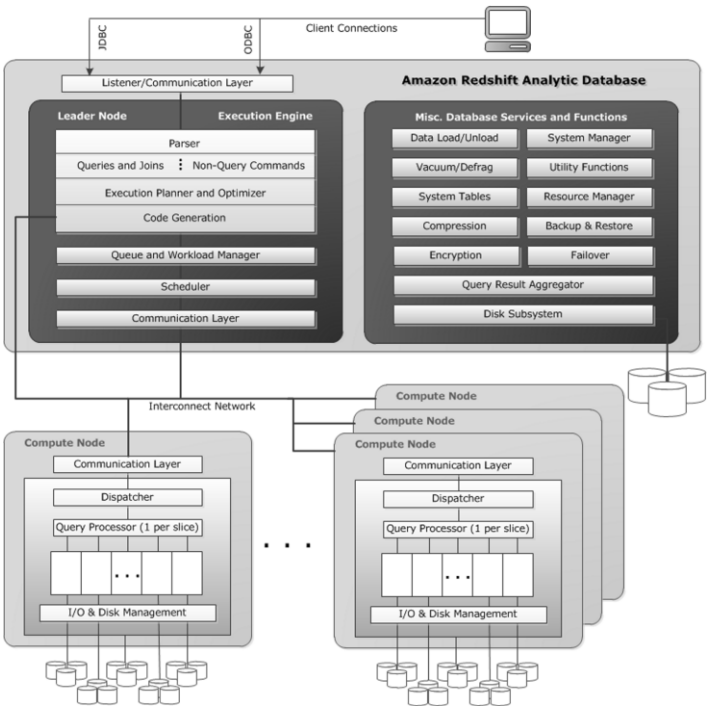
If you need to analyse big-data, it is no longer necessary to have powerful hardware. In fact, you can use many services in the cloud that allows you to manage, structure and analyse large amounts of data. These services also allow the

user not experienced in hardware and software to create procedures that lead to meaningful results.

Currently there are five main web platforms that can be used to analyse big-data on the Cloud. Using one of these platforms, in a few minutes, a cluster can be built for distributed analysis of big-data, obtaining performance and security without the cost of managing a complex hardware infrastructure:

- Amazon Web Services (AWS)
- Google Cloud Platform
- Microsoft Azure
- IBM Analytics
- SAP HANA Cloud Platform

Figure 3 - *High level view of internal components and functionality of the Amazon Redshift data warehouse (aws.amazon.com/it/documentation/redshift/)*



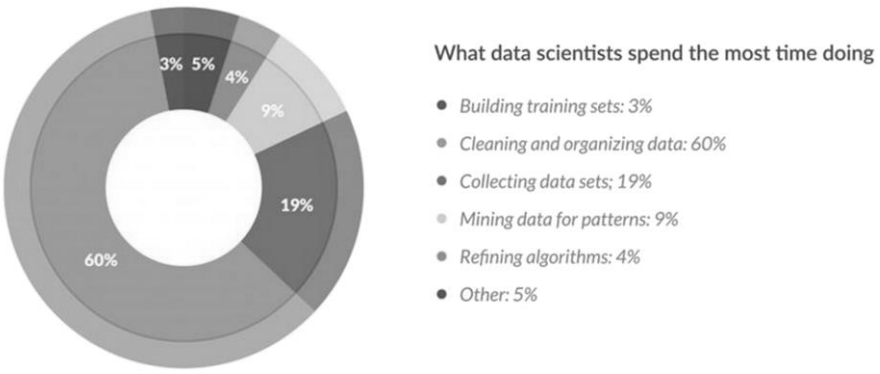
Each platform has its strengths. For example, Google Cloud allows to build complex machine learning models by the powerful TensorFlow framework. This software powers many Google products, with a managed scalable infrastructure,

which is powered by GPUs. Another example is Amazon Redshift, a cloud-based data warehouse (see fig. 3). It is both scalable and, despite its complexity, relatively easy to use.

4. Unpleasant aspects of big-data

The characteristics that big-data possess, involve a number of problems that must be addressed, before and during the statistical analysis. For example, it was observed that '*data preparation*' accounts for about 80% of the work of data scientists (Forbes 2016). In particular, they spend 60% of their time on cleaning and organizing data, while collecting data sets comes second at 19% of their time (fig. 4). Most of them view data preparation as the least enjoyable part of their work.

Figure 4 - *Forbes, march 23, 2016*



Many other problematic aspects have to be considered:

- *Privacy*. Combining someone's personal information with vast external data sets, it is possible to infer new facts about that person, including sensitive information.
- *Coverage/selection bias*. Usually, data are not collected considering a statistical representativeness. The population of interest could have different characteristics than the observed collective.
- *Data security*. The enormous amount of data generated by IoT, often stored on the cloud, is a big problem for data security.

- *Data storage*. Despite the great technical advances, automated tools can create enormous archives making their storage too expensive.
- *Computing power*. Depending on the features, big-data may require fast processors, distributed big-data platforms (such as Hadoop), parallel processing, clustering, MPP, virtualization, large grid environments, high connectivity and high throughputs.

As an example of new privacy and security issues, at the end of 2015, Vtech, a manufacturer of digital toys based in Hong Kong, admitted that cybercriminals had access to the personal data of 6.4 million children by remote control of their toys connected to Internet.

5. Statistical challenges in big-data

The primary value of big-data does not come from the data in their raw form, it comes from the elaboration of data and the products and services that can emerge from the analysis. The radical changes that are emerging in the data management technologies must be accompanied by changes in analytical techniques and the way in which these support decisions. Following the Scheveningen Memorandum on big-data and official statistics (2013), the general directors of the National Statistical Institutes "*Acknowledge that the use of Big-data in the context of official statistics requires new developments in methodology, quality assessment and IT related issues. The European Statistical System should make a special effort to supports these Developments*".

With big-data we have a great opportunity to take advantage of computational and statistical methods to transform raw data into knowledge in several areas such as health and medicine, security, education, and business intelligence.

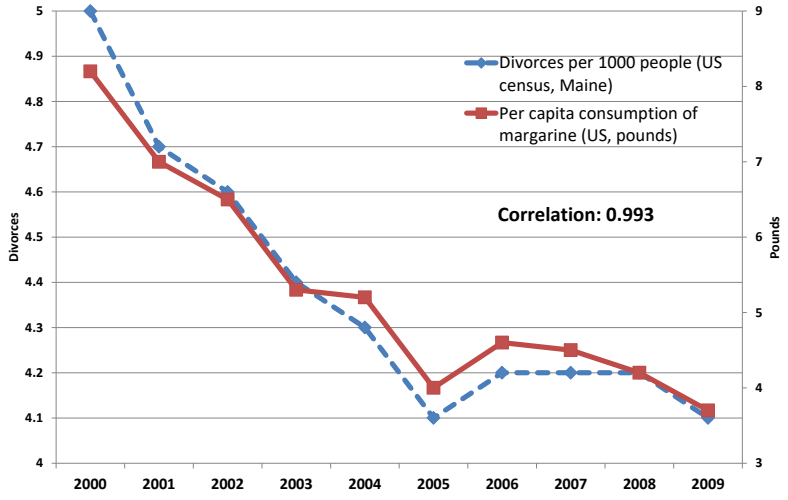
On the other hand, the high dimensionality introduces unique computational and statistical challenges.

- If there is not sampling, then there are not sampling errors. Rather, in this case, we should evaluate the model bias and its reliability or analyse the quality of data.
- In some cases, big-data have much more features than observations. For example, the standard set of microarray data are typically composed of thousands of features (the genes) with only few units.
- The number of fake correlations usually grows with the number of features. Spurious correlation may cause wrong statistical results (see fig. 5).

- When testing sequentially many hypotheses, we must correct for multiple testing. In fact, classical hypothesis procedure defines the test significant 5% (or 1%) of the time, even when the null is true.
- The traditional statistical modelling assumptions are hardly satisfied, which implies biased model parameters and inaccurate statistical tests.
- With big-data, common problems such as sampling bias, missing or incomplete data and sparsity must also be addressed.

Additional problems have been reported in the literature: the population can be unknown or difficult to define, errors can be placed in the pre-processing phase, data coming from social media usually include irrelevant babble or social bots. So, new tools and statistical methods will need to be developed for the pre-processing, classification, summarisation, feature extraction, anonymization and visualization of big-data.

Figure 5 – *A famous spurious correlations(tylervigen.com).*



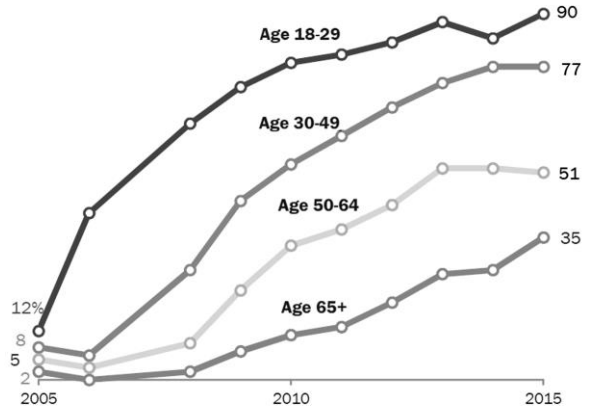
Usually, the major obstacle in the statistical analysis of big-data is the frequent lack of statistical representativity of the observed data. The selection bias can have a different relevance, depending of the analysed phenomena and the aim of analysis.

In recent years, particular attention has been paid to the composition of the collective that uses social networks. Following a research of the Pew Research Center (Perrin, 2015) on USA population, the bias concerns mainly *age* and *socio-economic status*. It is not surprising that young adults are the most likely to use

social media but, today, 35% of all those 65 and older report using social media. Moreover, those in higher-income households were more likely to use social media, while no significant difference was found with respect to gender, educational attainment, racial or ethnic group. A slight difference was found with respect to the urbanization level of the residence place.

Knowing the type of selection bias in the data, allows us to modify the analysis and avoid erroneous results.

Figure 6 - *Percentage of American adults using social media by age.*



Source: Pew Research Center surveys, 2005-2006, 2008-2015. No data are available for 2007.

6. Big-data analytics

In the last years, companies are no longer satisfied to extract detailed information from their archives, but now they require the application of complex predictive models (the analytics).

Among the models suitable for the analysis of big-data, we must consider *neural networks*. One of the advantages of this technique is that it does not require any distributional assumptions on the explanatory variables and the target (if we adopt a supervised approach). On the contrary, neural networks require a great amount of data to assure convergence to the correct parameter estimates. The downside is the difficulty to interpret the model or measure the importance of the variables. Big convolutional or recurrent neural networks have proven to be very effective in classification and cluster analysis (LeCun et al. 2015) and for this purpose they are used, for example, by Google and Facebook.

If the goal is prediction accuracy, averaging many prediction models together, can be a good choice. The idea is that by averaging (or majority voting) several prediction algorithms it is possible to reduce variability without giving up bias. One of the earliest proposals is based on bootstrapping samples and building multiple prediction functions (*Bagging - bootstrap aggregating*, Breiman 1996). *Random forests* extended this idea with classification trees, by a randomization of the features (Breiman 2001), while a further extension that allows to better avoid overfitting are the *extremely randomized trees* (Geurts et al. 2006).

Usually, the prediction algorithms that most frequently win Kaggle statistical competitions, and won the Netflix prize, blend multiple models together. The Netflix movie-rating challenge has become one of the most famous examples for big-data analytics (Bennett and Lanning 2007). Netflix is a movie-rental company that launched a competition in 2006 to try to improve their system for recommending movies to their customers. The Netflix dataset has 17,770 movies (columns) and 480,189 customers (rows). The data matrix is very sparse with "only" 100 million (1%) of the ratings present in the training set. The goal is to predict the ratings for unrated movies, so as to better recommend movies to customers. The winning algorithm used a combination of a very large number of statistical techniques, together with a complex pre-processing (Töscher et al. 2009).

It is known that the use of complex models with thousands of parameters can lead to overfitting, greatly reducing the predictive ability of the model and its reliability.

Overfitting occurs when a non-linear model fits observed random errors instead of the "underlying relationships". Given a sample *D*, once estimated a supervised model $\hat{f}_D$, we want to know its prediction capability on all the possible values of the features and the target. We define Prediction Error (PE):

$$PE = E_{\mathbf{x}} E_Y \left[L\left(Y, \hat{f}_D(\mathbf{x})\right) \mid \hat{f}_D \right]$$

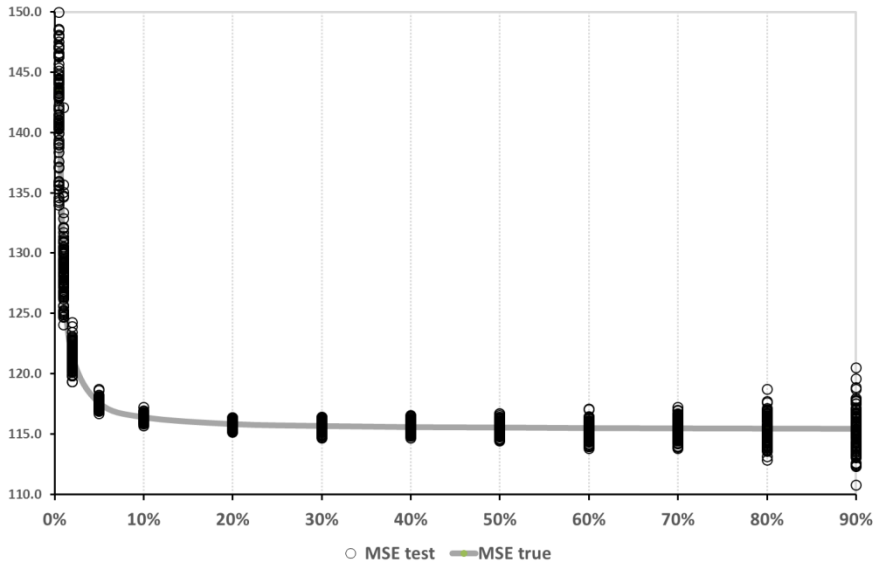
where L(.) is the loss function chosen. With a quadratic loss, the expression is the Mean Squared Error (MSE) computed on the population. In effect, PE is our evaluation of the fit loss for the estimated model on the unobserved population. To assess the model we need to estimate PE, usually by a cross-validation procedure (CV). This consists of splitting the original data-set in two or more parts in order to train the model on a data-set that is different from the data-set used to evaluate the model. In this way, we are able to obtain an evaluation that does not reward the overfitting models.

Several kind of cross-validation procedure can be used to estimate PE: Leave-one-out CV, Hold-out CV, K-fold CV (Di Ciaccio & Borra 2010). If we have a very large number of units, as it is common with big-data, the procedure less

expensive is the Hold-out CV, that splits the sample in only two parts, the training and the test sets.

An excessive number of units can be considered an obstacle for statistical analysis. Actually, even when you have archives with millions of units, it is more effective to have a relatively small training set (for example, 30,000 units) leaving all the other units in the test-set. This is a big advantage from a computational point of view because it allows to estimate the model on a reduced data set, but using the other data for the model evaluation. We can see this property, which may seem counterintuitive, through a simulation.

Figure 7 – *Estimated and true MSE for different hold-out splits (% of training set).*

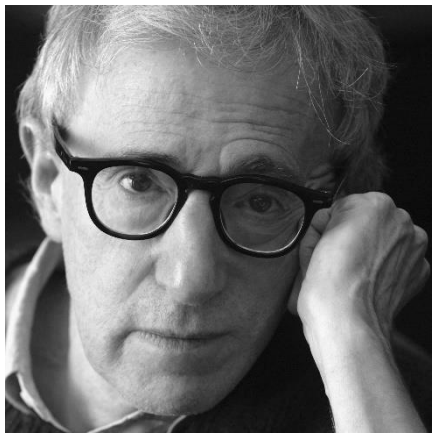


Given a big artificial dataset (100,000 units) with a quantitative target, we applied a small Neural Network to carry out a regression analysis. Hold-out with several split percentages was used to estimate the MSE, replicating the analysis 100 times for each percentage. The 'true' MSE was computed on an independent big-dataset. The results of the simulation, shown in fig. 7, display that with a training-size greater than 20,000 (20% of the sample) we obtain only an increase of the estimator variability of MSE (the circles). We can also deduce that with a large training-size and a large test-size, the hold-out estimator is essentially unbiased with low variability.

7. Dimensionality reduction

Analysing a database of millions of observations is not impossible for the statistical methods (e.g., by sampling the units). With thousands of variables, the number of input features should be reduced before a machine-learning algorithm can be successfully applied. Moreover, automatic statistical methodologies may be required to provide fast, even real-time, predictions, which would require parsimonious models. Indeed, until a few years ago, all calculations were made off-line, and researchers had time to process data with a relatively small sample size. Currently, with many big-data and online procedures, a different approach is needed.

Figure 8 - *Unsupervised dimensionality reduction by PCA (32 components)*



Original image (3800x3800 pixels)

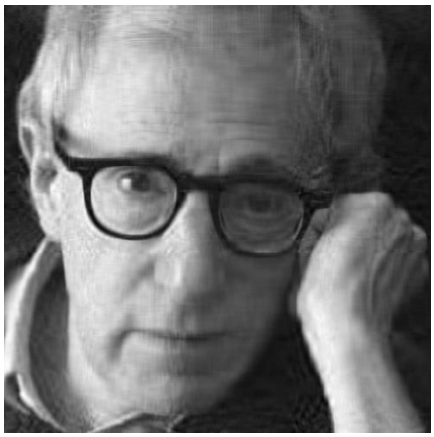


Image rebuilt by PCA

Dimensionality reduction can be performed in two main ways:

- *Unsupervised feature extraction*. Create a smaller set of new features exploiting redundancies in the input data. This can be useful to visualize high-dimension data, seek the intrinsic dimensionality, reduce big-data to manageable dimensions.
- *Supervised feature selection*. Maintain only the most significant features from the original dataset. This allow to discard the irrelevant features, reduce the big-data dimensions, simplify the model interpretation.

In the first approach we can cite the well known *Principal Component Analysis* (PCA), that is a very effective technique if the data structure is almost linear. In fig. 8 it is shown the application of the PCA to a matrix with 3800 rows and 3800 columns corresponding to the image on the left (8-bit grey-scale). The right image is obtained considering 32 columns (the first components of PCA) instead of the original 3800 columns. Effectively, examining fig. 9, it can be observed that 32 components are enough.

If the data structure is non-linear, PCA will overestimate the intrinsic dimensionality of the data. In this case, local approaches are more effective: *local* PCA (Kambhatla & Leen, 1997) or *nearest neighbor algorithm* (Pettis et al. 1979). In the global approaches, in addition to PCA, we can use *manifold methods* or *autoencoder neural networks*.

Figure 9 – *PCA for the image compression: explained variability ratio of each component.*

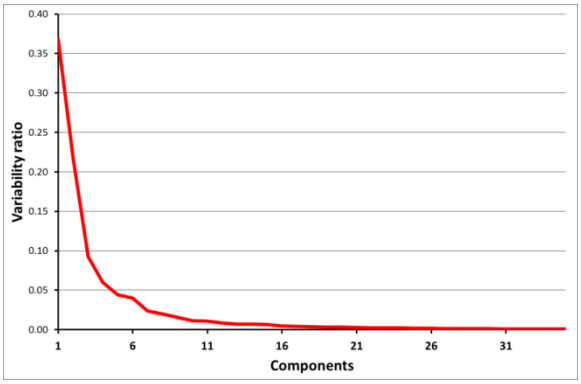
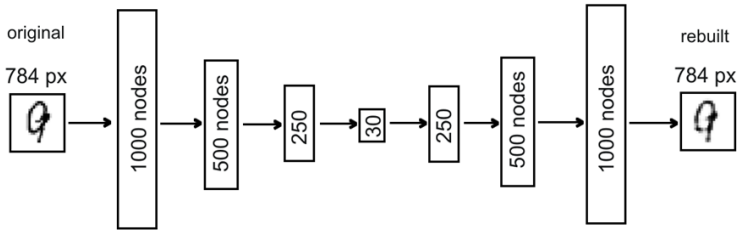
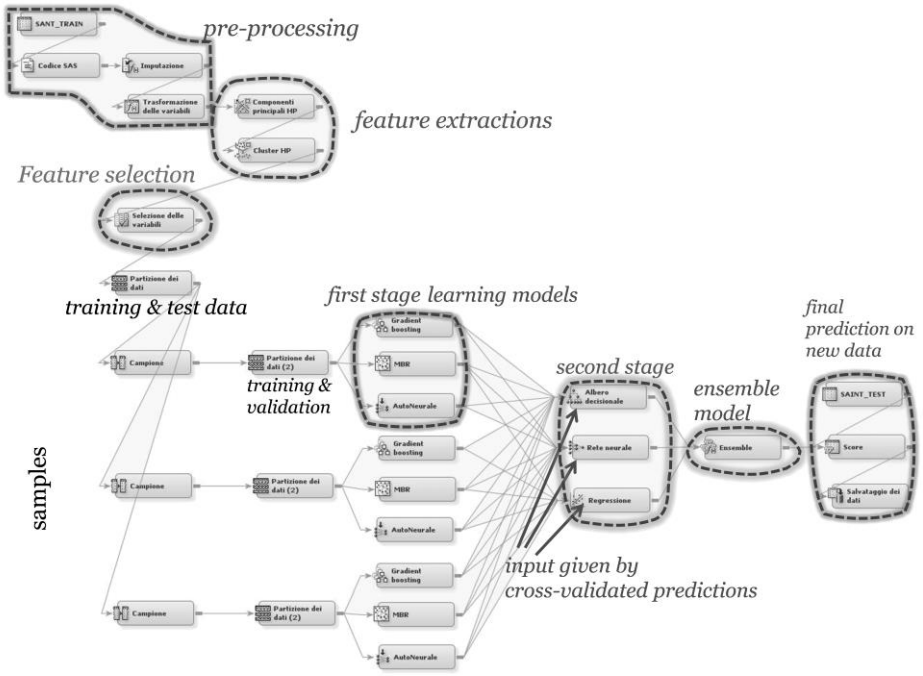


Figure 10 – *The autoencoder used for MNIST data, with 7 hidden layers: 1000-500-250-30-250-500-1000 nodes.*



An autoencoder (Hinton & Zemel 1994) is a multilayer neural network in which the target values are equal to the inputs. This unsupervised learning algorithm has a small central layer to reconstruct high-dimensional input vectors.

Figure 11 – *A complex scheme of analysis with re-sampling and ensemble learning (stacking) using SAS Enterprise Miner.*



In fig. 10 it is shown the structure of an autoencoder applied to the MNIST data set (LeCun et al., 1998). It was shown in literature that high-dimensional data can be converted to low-dimensional codes by training an autoencoder network that works much better than PCA (Hinton et al. 2006).

In general, having a lot of units and variables, we can adopt complex analysis schemes, as shown in fig. 11, where pre-processing, feature extraction and feature selection, sub-sampling, ensemble learning, are used together.

Finally, we may note that also the distinction between supervised and unsupervised approach might not be so strict in the future, if it were possible to obtain a correct classification without having labelled data. Some experiments with very large neural networks on big-data pictures showed that it is possible to train

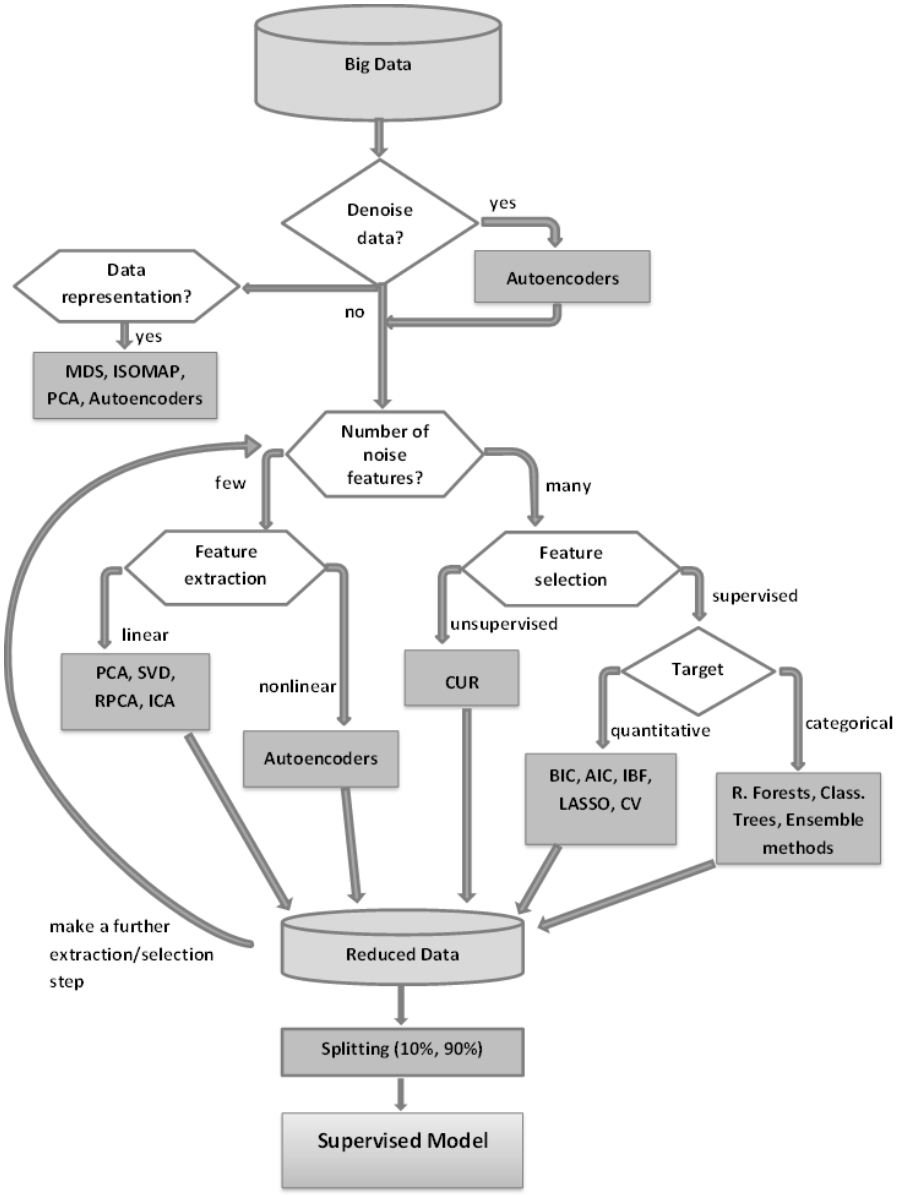
neurons to be selective for high-level concepts using entirely unlabelled data (Le, Ranzato et al. 2012).

8. Conclusion

The analysis of big-data can be approached in several ways, but the underlying problem is again a statistical problem. Learning methods for big-data are statistical methods that generalize and modify the classical techniques for the new data sets. With big-data, the underlying aim remain the same, although with more emphasis on data quality, dimensionality reduction, predictive capability and reliability of the learning models. The analysis of big-data requires the ability to assess the effectiveness of the model with a non-parametric inferential approach, usually referred as generalizability and regularization in the language of machine learning. The model can take advantage of the large amount of units available, but an excessive number of variables requires a dimensionality reduction, to avoid the inclusion a large amount of noise that can make ineffective the estimation of the model. As the flowchart of fig. 12 shows, the dimensionality reduction to be chosen depends on the characteristics of data and the objective of analysis.

Finally, it is useless to look with scepticism these huge data and the techniques that are applied, because this will be the main application field of the statistical analysis in the future, "*There is no need to distinguish big-data analytics from data analytics, as data will continue growing, and it will never be small again."* (Fan & Bifet 2012).

Figure 12 – *A typical flowchart of a big-data analysis.*



References

BENNETT, J., LANNING, S. 2007. The Netflix Prize. Paper presented at the KDD Cup Workshop, San Jose, CA, 12 August.

BREIMAN, L. 2001. Random Forests. *Machine Learning* 45 (1): 5–32. doi:10.1023/A: 1010933404324.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning* 24 (2): 123–140. doi:10.1007/BF00058655

DI CIACCIO, A., BORRA, S. 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods, *Computational Statistics & Data Analysis, 2010, vol. 54, issue 12*, pages 2976-2989.

FAN, W., BIFET, A. 2012. Mining Big-data: Current Status, and Forecast to the Future, *SIGKDD Explorations, vol. 14, issue 2*, pp. 1-5.

GEURTS, P., ERNST, D., WEHENKEL, L. 2006. Extremely randomized trees, *Machine Learning*, issue 1, pp 3-42.

HINTON, G.E., ZEMEL, R.S. 1994. Autoencoders, Minimum Description Length, and Helmholtz Free Energy. *Advances in Neural Information Processing Systems 6*. Cowan, Tesauro & Alspector (Eds.), Morgan Kaufmann: San Mateo, CA.

HINTON G.E., SALAKHUTDINOV R.R. 2006. Reducing the dimensionality of data with Neural Networks, *Science, vol. 313, issue 5786*, pp. 504-507.

KAMBHATLA, N., LEEN T.K. 1997. Dimension reduction by local principal component analysis, *Neural Computation 9 (7)*, 1493–1516.

LANEY, D. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety, *Gartner, Application Delivery Strategies*, 6 February 2001.

LECUN, Y., BENGIO, Y., HINTON, G., 2015. Deep Learning, *Nature 521*, 436–444.

LECUN, Y., BOTTOU, L., BENGIO, Y., HAFFNER, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11):2278-2324.

LE Q.V., RANZATO M., MONGA R, DEVIN M. .et al. 2012. Building High-level Features Using Large Scale Unsupervised Learning, *International Conference in Machine Learning 2012*.

SCHEVENINGEN Memorandum on Big-data and Official Statistics, 2013. DGINS conference.

SKALIOTIS, M., 2015. Big-data in the European Statistical System. Conference by STATEC and EUROSTAT, World Statistics Day 20.10.2015.

SONDERGAARD, P., 2011. Gartner Symposium/ITxpo 2011, October 16-20, Orlando.

TÖSCHER, A., JAHRER, M., 2009. The BigChaos Solution to the Netflix Grand Prize. http://www.stat.osu.edu/~dmsl/GrandPrize2009_BPC_BigChaos.pdf

PEDREGOSA F., VAROQUAUX G., GRAMFORT A. et al., 2011. SCIKIT-LEARN: Machine Learning in Python, JMLR 12, pp. 2825-2830.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C., NG A., POTTS C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP.

PERRIN A. 2015. "Social Networking Usage: 2005-2015." Pew Research Center. October 2015. Available at: http://www.pewInternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

PETTIS K., BAILEY, T., JAIN, A., DUBES, R. 1979. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. Pattern Analysis and Machine Intelligence, 1(1),* pp. 25–37.

UNECE 2013. Classification of Types of Big-data. http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

SUMMARY

It is estimated that about 90% of the currently available data have been produced over the last two years. Of these, only 0.5% is effectively analysed and used. However, this data can be a great wealth, the oil of 21st century, when analysed with the right approach. In this article, we illustrate some specificities of these data and the great interest that they can represent in many fields. Then we consider some challenges to statistical analysis that emerge from their analysis, suggesting some strategies.

Agostino DI CIACCIO, Department of Statistical Science, agostino.diciaccio@uniroma1.it

Giovanni Maria GIORGI, Department of Statistical Science, giovanni.giorgi@uniroma1.it