

50 Years of Data Analysis: From Exploratory Data Analysis to Predictive Modeling and Machine Learning

Gilbert Saporta

► **To cite this version:**

Gilbert Saporta. 50 Years of Data Analysis: From Exploratory Data Analysis to Predictive Modeling and Machine Learning. Christos Skiadas. Data Analysis and Applications 1. Clustering and Regression, Modeling-estimating, Forecasting and Data Mining, ISTE-Wiley, 2019, Data Analysis and Applications, 978-1-78630-382-0. hal-02470740

HAL Id: hal-02470740

<https://hal-cnam.archives-ouvertes.fr/hal-02470740>

Submitted on 11 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

50 years of data analysis: from exploratory data analysis to predictive modelling and machine learning

Gilbert Saporta
CEDRIC-CNAM, France

In 1962, J.W.Tukey wrote his famous paper "The future of data analysis" and promoted Exploratory Data Analysis (EDA), a set of simple techniques conceived to let the data speak, without prespecified generative models. In the same spirit J.P.Benzécri and many others developed multivariate descriptive analysis tools. Since that time, many generalizations occurred, but the basic methods (SVD, k-means, ...) are still incredibly efficient in the Big Data era.

On the other hand, algorithmic modelling or machine learning are successful in predictive modelling, the goal being accuracy and not interpretability. Supervised learning proves in many applications that it is not necessary to understand, when one needs only predictions.


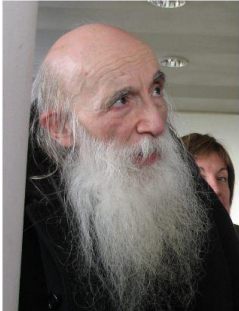
However, considering some failures and flaws, we advocate that a better understanding may improve prediction. Causal inference for Big Data is probably the challenge of the coming years.




Keywords: exploratory data analysis, machine learning, prediction

It is a little presumptuous to want to make a panorama of 50 years of data analysis while David Donoho (2017) has just published a paper entitled "50 Years of Data Science". But 1968 is the year when I began my studies as a statistician and I would very much like to talk about the debates of the time and the digital revolution that profoundly transformed statistics and which I witnessed. The terminology followed this evolution-revolution: from Data Analysis to Data Mining and then to Data Science while we went from a time when the asymptotics began to 30 observations with a few variables, in the era of Big Data and high dimension.

1.The revolt against mathematical statistics

Since the 1960s, the availability of data has led to an international movement back to the sources of statistics "let the data speak" and to sometimes fierce criticisms of an abusive formalization. In addition to John Tukey cited above, here is a portrait gallery of some notorious protagonists in USA, France, Japan, the Netherlands, Italy.

		
John Wilder Tukey (1915-2000)	Jean-Paul Benzécri (1932-)	Chikio Hayashi (1918-2002)

		
Jan de Leeuw (1945-)	J. Douglas Carroll (1939-2011)	Carlo Lauro (1943-)

And an anthology of quotes:

- *“He (Tukey) seems to identify statistics with the grotesque phenomenon generally known as mathematical statistics and find it necessary to replace statistics by data analysis”* (Anscombe, 1967).
- *“ Statistics is not probability, under the name of mathematical statistics was built a pompous discipline based on theoretical assumptions that are rarely met in practice”* (Benzécri, 1972)
- *“The models should follow the data, not vice versa”* (Benzécri, 1972)
- *“Use the computer implies the abandonment of all the techniques designed before of computing”* (Benzécri, 1972)
- *“ Statistics is intimately connected with science and technology, and few mathematicians have experience or understand of methods of either. This I believe is what lies behind the grotesque emphasis on significance tests in statistics courses of all kinds; a mathematical apparatus has been erected with the notions of power, uniformly most powerful tests, uniformly most powerful unbiased tests, etc. etc., and this is taught to people, who, if they come away with no other notion, will remember that statistics is about significant differences (...). The apparatus on which their statistics course has been constructed is often worse than irrelevant---it is misleading about what is important in examining data and making inferences “* (Nelder, 1985)

Data analysis was basically descriptive, non-probabilistic in the sense that no reference was made to the data generating mechanism. Data analysis favors algebraic and geometrical tools of representation and visualization.

This movement has resulted in conferences especially in Europe. E.Diday and L.Lebart initiated in 1977 a series entitled Data Analysis and Informatics, J.Janssen was in 1981 at the origin of biennial ASMDA conferences (Applied Stochastic Models and Data Analysis) which are still continuing.

The principles of data analysis inspired those of Data Mining, which developed in the 1990s on the border between databases, information technology and statistics. Fayaad (1995) is said to have the following definition: *“Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”* . David Hand et al. precised in 2000 *“ I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets”*

The metaphor of Data Mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools. Data Mining is generally concerned with data which were collected for another purpose: it is a secondary analysis of data bases that are collected Not Primarily For Analysis, but for the management of individual cases. Data Mining is not concerned with efficient methods for collecting data such as surveys and experimental designs (Hand, 2000)

2. EDA and unsupervised methods for dimension reduction

Essentially, exploratory methods of data analysis are dimension reduction methods: unsupervised classification or clustering methods operate on the number of statistical units, whereas factorial methods reduce the number of variables by searching for linear combinations associated with new axes of the space of individuals.

2.1 The time of syntheses

It was quickly realized that all the methods looking for eigenvalues and eigenvectors of matrices related to the dispersion of a cloud (total or intra) or of correlation matrices could be expressed as special cases of certain techniques.

Correspondence analyzes (single and multiple), canonical discriminant analysis, are particular principal component analyzes. It suffices to extend the classical PCA by weighting the units and introducing metrics. The duality scheme introduced by Cailliez and Pagès (1976) is an abstract way of representing the relationships between arrays, matrices and associated spaces. The paper by De la Cruz, O. & Holmes, S.P. (2011) brought it back to light.

From an other point of view (Bouroche & Saporta, 1983) the main factorial methods PCA, MCA, as well as multiple regression are particular cases of canonical correlation analysis.

An other synthesis comes from the generalization of canonical correlation analysis to several groups of variables introduced by J.D.Carroll (1968). Given p blocks of variables \mathbf{X}_j one looks for

components \mathbf{z} maximizing the following criterium : $\sum_{j=1}^p R^2(\mathbf{z}, \mathbf{X}_j)$.

The extension of this criterion in the form $Max_Y \sum_{j=1}^p \Phi(Y, X_j)$ where Φ is an adequate measure of association lead to the maximum association principle M.Tenenhaus (1977), J.F. Marcotorchino (1986), G.Saporta (1988) which also includes the case of k-means partitioning

critérium	analysis
$\max \sum_{j=1}^p r^2(c, x_j)$ with x_j numerical	PCA
$\max \sum_{j=1}^p \eta^2(c, x_j)$ with x_j categorical	MCA
$\max \sum_{j=1}^p R^2(c, \mathbf{X}_j)$ with \mathbf{X}_j data set	GCA (Carroll)
$\max \sum_{j=1}^p Rand(Y, x_j)$ with Y and x_j categorical	Central partition
$\max \sum_{j=1}^p \tau(y, x_j)$ with rank orders	Condorcet aggregation rule

Table 1 : Various cases of the maximum association principle

PLS approach to structural equation modelling also provides a global framework for many linear methods as has been shown by Tenenhaus (1999) and Tenenhaus & Tenenhaus (2011).

2.2 The time of clusterwise methods

The search for partitions in k classes of a set of units belonging to a Euclidean space is most often done using the k -means algorithm: this method converges very quickly, even for large sets of data, but not necessarily towards the global optimum. Under the name of dynamic clustering, E. Diday (1971) has proposed multiple extensions, where for example the representatives of classes can be groups of points, varieties etc. The simultaneous search for k classes and local models by alternating k -means and modeling is a geometric and non-probabilistic way of addressing mixture problems. Clusterwise regression is the best known case: in each class a regression model is fitted and the assignment to the classes is done according to the best model. Clusterwise methods allow for non-observable heterogeneity and are particularly useful for large data sets where the relevance of a simple and global model is questionable. In the 1970s, E. Diday and his collaborators developed "typological" approaches for most linear techniques: PCA, regression (Charles, 1977), discrimination. These methods are again the subject of numerous publications, in association with functional data (Preda & Saporta, 2005), symbolic data (de Carvalho et al., 2010), and in multiblock cases (De Roover et al., 2012, Bougeard et al., 2017).

2.3 Extensions to new types of data

2.3.1 Functional data

Jean-Claude Deville (1974) showed that the Karhunen-Loève decomposition was nothing else than the PCA of the trajectories of a process, opening the way to functional data analysis (Ramsay & Silverman, 1997). The number of variables being infinitely not countable, the notion of linear combination to define a principal component, is extended to the integral $\xi = \int_0^T f(t)X_t dt$, $f(t)$ being an eigenfunction of the covariance operator $\int_0^T C(t,s)f(s)ds = \lambda f(t)$.

Deville & Saporta (1980) then extended functional PCA to correspondence analysis of trajectories of a categorical process.

The dimension reduction offered by PCA makes possible to give solutions to the problem of regression on trajectories, a problem that is ill posed since the number of observations is smaller than the infinite number of variables. PLS regression, however, is better adapted in the latter case and makes it possible to deal with supervised classification problems (Costanzo & al., 2006)

2.3.2 Symbolic data analysis

E. Diday is at the origin of many works that have made it possible to extend almost all methods of data analysis to new types of data, called symbolic data. This is the case for example when the cell i, j of a data table is no longer a number, but an interval or a distribution. Here is an example of a table of symbolic data (from Billard & Diday, 2006):

w_u	Court Type	Player Weight	Player Height	Racket Tension
w_1	Hard	[65, 86]	[1.78, 1.93]	[14, 99]
w_2	Grass	[65, 83]	[1.80, 1.91]	[26, 99]
w_3	Indoor	[65, 87]	[1.75, 1.93]	[14, 99]
w_4	Clay	[68, 84]	[1.75, 1.93]	[24, 99]

Table 2 An example of interval data

2.3.3 Textual data

Correspondence analysis and classification methods were very early applied to the analysis of document-term and open-text tables. Refer to Lebart & al., (1998) for a full presentation. Text analysis is now part of the vast field of text mining or text analytics.

2.4 Non linear data analysis

Dauxois and Pousse (1976) extended principal component analysis and canonical analysis to Hilbert spaces. By simplifying their approach, instead of looking for linear combinations of maximum

variance like in PCA $\max V\left(\sum_{j=1}^p a_j x^j\right)$ subject to $\|\mathbf{a}\|=1$, one looks for separate nonlinear

transformations Φ_j of each variable maximizing $V\left(\sum_{j=1}^p \Phi_j(x^j)\right)$. This is equivalent to maximize the

sum of the squares of the correlation coefficients between the principal component c and the transformed variables $\sum_{j=1}^p \rho^2(c, \Phi_j(x^j))$ which is once again an illustration of the maximum association principle.

With a finite number of observations n , this is a ill-posed problem, and one needs to restrict the set of transformations Φ_j to finite dimension spaces. A classical choice is to use spline functions as in Besse (1988).

The search for optimal transformations has been the subject of work by the Dutch school synthesized in the book published under the collective name of Gifi (1999).

Separate transformations are called semi-linear. A different attempt to obtain "truly" nonlinear transformations is the kernelization:

In line with the work of V. Vapnik, Schölkopf et al. (1998) defined a non-linear PCA in the following manner where the entire vector $\mathbf{x}=(x^1, x^2, \dots, x^p)$ is transformed. Each point of the space of the individuals E is transformed into a point in a space $\Phi(E)$ called extended space (or feature space) provided with a dot product. The dimension of $\Phi(E)$ can be very large and the notion of variable is lost. A metric multidimensional scaling is then performed on the transformed points according to the Torgerson method which is equivalent to the PCA in $\Phi(E)$. Everything rests on the choice of the scalar product in $\Phi(E)$: if we take a scalar product which is easily expressed according to the scalar product of E , it is no longer necessary to know the transformation Φ which is then implicit. All calculations are done in dimension n . This is the "kernel trick".

Let $k(\mathbf{x}, \mathbf{y})$ be a dot product in $\Phi(E)$ and $\langle \mathbf{x}, \mathbf{y} \rangle$ the dot product of E . One has just to replace the usual Torgerson's matrix \mathbf{W} by a matrix where each element is $k(\mathbf{x}, \mathbf{y})$, then to doubly center \mathbf{W} in rows and columns : its eigenvectors are the principal components in $\Phi(E)$.

Once defined the kernel-PCA, many works followed "kernelizing" various methods, let us quote: Fisher discriminant analysis by Baudat & Anouar (2000) found independently under the name of LS-SVM by Suykens & Vandewalle (1999), the PLS regression of Rosipal & Trejo (2001), the unsupervised classification with kernels k -means already proposed by Schölkopf et al., canonical analysis (Fyfe & Lai., 2001). It is interesting to note that most of these developments come not from statisticians but from researchers in Artificial Intelligence or Machine Learning.

2.5 The time of sparse methods

When the number of dimensions (or variables) is very large, PCA MCA and other factorial methods lead to results that are difficult to interpret: how to make sense of a linear combination of several hundred or even thousands of variables? The search for so-called "sparse" combinations limited to a small number of variables, that is to say with a large number of zero coefficients, has been the subject of the attention of researchers for some fifteen years. The first attempts requiring that the coefficients be equal to -1, 0 or 1, for example, lead to non-convex algorithms that are difficult to use.

The transposition to PCA of the LASSO regression de Tibshirani (1996) allowed to give exact and elegant solutions. Recall that the LASSO consists in performing a regression with a L^1 penalty on the coefficients, which makes it possible to easily manage the multicollinearity and the high dimension.

$$\hat{\beta}_{lasso} = \arg \min_{\beta} \left(\|y - X\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

Zou et al. (2006) proposed to modify one of the many criteria defining the PCA of a table X : principal components z are such that:

$$\hat{\beta} = \arg \min_{\beta} \|z - X\beta\|^2 + \lambda \|\beta\|^2 + \lambda_1 \|\beta\|_1$$

The first constraint in L^2 norm only implies that the loadings have to be normalized ; the second constraint in L^1 norm tunes the sparsity when the Lagrange multiplier λ_1 varies. Computationally, one gets the solution by alternating a SVD β being fixed fixé, to get the components z and an elastic-net to find β when z is fixed, until convergence.

The positions of the null coefficients are not the same for the different components. The selection of the variables is therefore dimension by dimension. If the interpretability increases, the counterpart is the loss of characteristic properties of PCA such as the orthogonality of the principal components and / or the loadings. Since then, sparse variants of many methods have been developed, such as sparse-PLS by Chun & Keleş (2009), sparse discriminant analysis by Clemmensen et al. (2011), sparse canonical analysis by Witten et al. (2009) and sparse multiple correspondence analysis by Bernard et al. (2012).

3. Predictive modelling

A narrow view would limit data analysis to unsupervised methods to use current terminology. Predictive or supervised modeling has evolved in many ways into a conceptual revolution comparable to that of the unsupervised. We have moved from a model-driven approach to a data driven approach where the models come from the exploration of the data and not from a theory of the mechanism generating observations, thus reaffirming the second principle of Benzécri: *"the models should follow the data, not vice versa"* .

The difference between these two cultures: generative models versus algorithmic models, or models to understand versus models to predict has been theorized by Breiman (2001), Saporta (2008), Shmueli (2010) and taken up by Donoho (2015). The meaning of the word model has evolved: from that of a parsimonious and understandable representation centered on the fit to observations (*predict the past*) one has moved to black box type algorithms whose objective is to forecast the most precisely possible new data (*predict the future*). The success of machine learning and especially

the renewal of neural networks with deep learning have been made possible by the increase in computing power, but also and above all by the availability of huge learning bases.

3.1 Paradigms and paradoxes

When we ask ourselves what is a good model, we quickly arrive at paradoxes.

A generative model that fits well with collective data can provide poor forecasts when trying to predict individual behaviors. The case is common in epidemiology. On the other hand, good predictions can be obtained with uninterpretable models: targetting customers, or approving loans, are not needing a consumer theory. Breiman remarked that simplicity is not always a quality: *“Occam’s Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately in prediction, accuracy and simplicity (interpretability) are in conflict”*.

“Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data” (Breiman, 2001).

Other quotes illustrate these paradoxes:

“Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms” (Vapnik, 2006).

“Statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy” (Shmueli, 2010)

In a Big Data world, estimation and tests become useless, because everything is significant! For instance a correlation coefficient equal to 0,002 when the number of observations is 10^6 is significantly different from 0, but without any interest. Usual distributional models are rejected since small discrepancies between model and data are significant. Confidence intervals have zero length etc. etc. One should keep in mind the famous sentence of George Box: *“All models are wrong, some are useful”*.

3.2 From statistical learning theory to empirical validation

One of the major contributions of the theory of statistical learning developed by Vapnik and Cervonenkis was to give the conditions of generalizability of the predictive algorithms and to establish inequalities on the difference between the empirical error of adjustment of a model to observed data and the theoretical error when applying this model to future data from the same unknown distribution. If the theory is not easy to use, it has given rise to the systematization of the practice of dividing data into 3 subsets: learning, testing, validation (Hastie, Tibshirani, Friedman, 2001).

There had been warnings in the past, like that of Paul Horst 1941: *“the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established”* and the finding of cross validation by Lachenbruch & Mickey, 1968 and Stone, 1974. But it is only recently that the use of validation and test samples has become widespread and has become an essential step for any data scientist. However, there is still room for improvement if we go through the publications of certain areas where prediction is rarely checked on a hold-out sample.

3.3 Challenges

Supervised methods have become a real technology governed by the search for efficiency. There is now a wealth of methods, especially for binary classification: SVM, random forests, gradient boosting, neural networks, to name a few. Ensemble methods are superimposed to combine them (see Noçairi et al., 2016). Feature engineering consists of constructing a large number of new variables functions of those observed and choosing the most relevant ones. While in some cases the gains over conventional methods are spectacular, this is not always the case as noted by Hand (2006).

Software becomes more and more available: in 50 years we have moved from the era of large, expensive commercial systems (SAS, SPSS) to the distribution of free open source packages like R and ScikitLearn. The benefits are immense for the rapid dissemination of new methods but the user must be careful about the choice and often the lack of validation and quality control of many packages: it is not always clear if user written packages are really doing what they claim to be doing. Hornik (2012) was already wondering if there were not too many R packages

10 years ago, in a resounding article C.Anderson had prophesied the end of theory since *“the data deluge makes the scientific method obsolete”*. In a provocative manner he wrote *“Petabytes allow us to say: “Correlation is enough.” We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.”* This was of course misleading and the setbacks of Google's epidemic influenza forecasting algorithm brought a denial (Lazer et al., 2014). Correlation is not causality and drawing causal inference from observational data has always been a tricky problem. As Box et al. put it (1978), *“To find out what happens when you change something, it is necessary to change it”*. The best way to answer causal questions is usually to run an experiment. Drawing causal inference from Big Data is now a hot topic, cf Bottou et al. (2013) et Varian (2016).

Quantity is not quality and massive data can be biased and lead to unfortunate decisions reproducing the a priori that led to their collection. Many examples have been discovered related to discrimination or presuppositions about gender or race. More generally, the treatment of masses of personal data raises ethical and privacy issues when consent has not been gathered, or has not been sufficiently explained. Books for the general public such as Keller and Neufeld (2014) and O'Neil (2016) have echoed this.

5. Conclusion

The past 50 years have been marked by dramatic changes in statistics. The ones that will follow will not be less formidable. The Royal Statistical Society is not afraid to write in its Data Manifesto *“What steam was to the 19th century, and oil has been to the 20th, data is to the 21st.”*

Principles and methods of Data Analysis are still actual and exploratory (unsupervised) and predictive (supervised) analysis are the two sides of the same approach. But as correlation is not enough, causal inference could be the new frontier and could go beyond the paradox of predicting without understanding by going towards understanding to better predict, and act to change.

If the job of statistician or data scientist will be ever more exciting, we believe that it will have to be accompanied by an awareness of its social responsibility.

References

- Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, <http://www.wired.com/2008/06/pb-theory/>
- Baudat, G., Anouar, F. (2000) Generalized discriminant analysis using a kernel approach. *Neural Computation* **12**, 10, 2385–2404
- Bernard, A. , Guinot, C., Saporta, G. (2012) Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis , In: Colubi, A., Fokianos, K., Gonzalez-Rodriguez, G., Kontaghiorghes, E. (eds.) *Proceedings of 20th International Conference on Computational Statistics (COMPSTAT 2012)*, 99–106
- Besse, Ph. (1988) Spline functions and optimal metric in linear principal components analysis. *In Components and Correspondence Analysis* (Van Rijckevorsel et al, Eds.). John Wiley & Sons.
- Billard, L., Diday, E. (2012) *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons.
- Bottou, L. et al. (2013) Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, **14**, 3207–3260
- Bougeard, S. , Abdi, H. , Saporta, G. , Niang Keita, N. (2018). Clusterwise analysis for multiblock component methods , *Advances in Data Analysis and Classification*, **12**, 2, 285-313
- Box, G., Hunter, J.S, Hunter, W.G. (1978) *Statistics for Experimenters* , John Wiley & Sons
- Breiman, L. (2001) Statistical Modeling: The Two Cultures, *Statistical Science*, **16**, 3, 199–231
- Cailliez, F., Pagès, J.P. (1976) : *Introduction à l'analyse des données*, Smash, Paris
- Carroll, J.D. (1968) Generalisation of canonical correlation analysis to three or more sets of variables. *Proceedings, 76 th Annual Convention, American Psychological Association*, **3**, 227-228
- Chun, H. , Keleş, S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B* , **72**, 3-25
- Clemmensen, L., Hastie, T. and Ersboell, K. (2011) Sparse discriminant analysis, *Technometrics*, **53**, 4 , 406-413
- Costanzo, D., Preda, C., Saporta, G. (2006) Anticipated prediction in discriminant analysis on functional data for binary response, *in COMPSTAT'06, A.Rizzi ed. , Physica-Verlag*, 821-828
- De Roover, K. , Ceulemans, E., Timmerman, M.E., Vansteelandt, K., Stouten, J., Onghena, P. (2012) Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol Methods*. **17**, 1, 100-119
- De la Cruz, O., Holmes, S.P. (2011) The Duality Diagram in Data Analysis: Examples of Modern Applications, *Annals of Applied Statistics*, **5**, 4, 2266-2277

- Deville J.C., (1974) Méthodes statistiques et numériques de l'analyse harmonique, *Annales de l'INSEE*, **15**, 3-101
- Deville J.C., Saporta, G. (1980) Analyse harmonique qualitative, in *Data Analysis and Informatics*, E.Diday, North-Holland, 375-389
- Diday, E. (1974) Introduction à l'analyse factorielle typologique, *Revue de Statistique Appliquée*, **22**, 4, 29-38
- Donoho, D. (2017) 50 Years of Data Science, *Journal of Computational and Graphical Statistics*, **26**, 4, 745-766
- Friedman, J.H. (2001), The Role of Statistics in the Data Revolution?, *International Statistical Review*, **69**, 1, 5-10
- Fyfe, C., & Lai, P. L. (2001). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, **10**, 365–374
- Gifi, A. (1990) *Non-linear multivariate analysis*, John Wiley & Sons
- Hand, D., Blunt, G., Kelly, M., Adams, N. (2000) Data mining for fun and profit, *Statistical Science*, **15**, 2, 111-126
- Hand, D. (2006) Classifier Technology and the Illusion of Progress, *Statistical Science*, **21**, 1, 1-14.
- Hastie, T., Tibshirani, R., Friedman, J. (2001) *The Elements of Statistical Learning*, Springer
- Keller, M., Neufeld, J. (2014) *Terms of Service: Understanding Our Role in the World of Big Data*, Al Jazeera America
- Hornik, K. (2012) Are There Too Many R Packages? *Australian Journal of Statistics*, **41**, 1, 59-66
- Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014) The Parable of Google Flu: Traps in Big Data Analysis, *Science*, **343**, 6176, 1203-1205
- Lebart, L., Salem, A. , Berry, L. (1998) *Exploring Textual Data*, Kluwer Academic Publisher, Dordrecht, Boston
- Marcotorchino, F. (1986) Maximal association as a tool for classification, in *Classification as a tool for research*, Gaul & Schader (eds), North Holland, 275-288
- Nelder, J.A. , discussion of Chatfield, C., The initial examination of data , *Journal of the Royal Statistical Society, Series A*, **148**, 214-253, 1985.
- Noçairi ,H., Gomes,C., Thomas, M., Saporta, G.. (2016) Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry, *Electronic Journal of Applied Statistical Analysis*, **9**, 2 340-361
- O'Neil, C. (2016) *Weapons of Maths Destruction*, Crown
- Ramsay, J.O. , Silverman, B. (1997) *Functional data analysis*, Springer

- Rosipal, A., Trejo, L. (2001) Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, *Journal of Machine Learning Research*, **2**, 97-123
- Schölkopf, B., Smola, A., Müller, K.L. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, **10**, 5, 1299-1319
- Suykens, J.A.K.; Vandewalle, J. (1999) Least squares support vector machine classifiers, *Neural Processing Letters*, **9**, 3, 293-300
- Saporta, G. (1988) About maximal association criteria in linear analysis and in cluster analysis, in *Classification and Related Methods of Data Analysis*, H.H.Bock ed., p.541-550, North-Holland
- Saporta, G. (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- Shmueli, G. (2010) To explain or to predict? *Statistical Science*, **25**, 289–310
- Tenenhaus, M. (1977) Analyse en composantes principales d'un ensemble de variables nominales ou numériques, *Revue de Statistique Appliquée*, **25**, 2, 39-56
- Tenenhaus, M. (1999) L'approche PLS, *Revue de Statistique Appliquée*, **17**, 2, 5-40
- Tenenhaus, A., Tenenhaus, M. (2011) Regularized Generalized Canonical Correlation Analysis, *Psychometrika*, **76**, 2, 257-284
- Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288
- Tukey, J.W. (1962) The Future of Data Analysis, *Ann. Math. Statist.* **33**, 1, 1-67
- Vapnik, V. (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- Varian, H. (2016) Causal inference in economics and marketing, *PNAS*, **113**, 7310-7315
- Witten, D., Tibshirani, R., and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 3, 515-534
- Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265-286