Clustering: Science or Art?[*]

Isabelle Guyon
ClopiNet
955 Creston Road, Berkeley, CA 94708, USA
isabelle@clopinet.com

Ulrike von Luxburg
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
ulrike.luxburg@tuebingen.mpg.de

Robert C. Williamson
Australian National University and NICTA
Canberra, Australia
Bob.Williamson@anu.edu.au

November 10, 2009

Abstract

This paper deals with the question whether the quality of different clustering algorithms can be compared by a general, scientifically sound procedure which is independent of particular clustering algorithms. In our opinion, the major obstacle is the difficulty to evaluate a clustering algorithm without taking into account the context: why does the user cluster his data in the first place, and what does he want to do with the clustering afterwards? We suggest that clustering should not be treated as an application-independent mathematical problem, but should always be studied in the context of its end-use. Different techniques to evaluate clustering algorithms have to be developed for different uses of clustering. To simplify this procedure it will be useful to build a "taxonomy of clustering problems" to identify clustering applications which can be treated in a unified way.

Preamble

Every year, dozens of papers on clustering algorithms get published. Researchers continuously invent new clustering algorithms and work on improving existing ones.

> *People who work on end-use problems ("applications") remain rather untouched by most of these papers. They continue to use their favorite algorithms, usually k-means and linkage algorithms.*

Researchers who publish papers about new clustering algorithms always struggle with the same question: How can they convince a reader that their algorithm is "good"?

> *Applied people don't really care. They don't believe that there is an algorithm which can always discover what they are looking for, and they don't think that there exists "the true clustering" of a data set anyway. They continue to use k-means.*

[*] Authors in alphabetical order

Researchers treat clustering as if it were a scientific discipline. They try to come up with various scores to assess the quality of clustering algorithms.

> *Applied people consider clustering rather as an art or a craft. If used with skill, it can be a useful tool. No more, no less.*

1 Introduction

In his famous Turing award lecture, Knuth (1974) says of Computer Programming that "It is clearly an art, but many feel that a science is possible and desirable." Whether clustering is art or science is, in fact, an old question. *"Is taxonomy art, or science, or both?"* so asked Anderson (1974) in reviewing the state of systematic biological taxonomy. He justifiably went on to claim

> Discussions of taxonomic theory or practice that refer to the concepts "science" and "art" without finer delineation will be less clear and less productive than discussions that first attempt definitions of specific concepts in the panoply of science such as precision or objectivity or repeatability or confidence and then apply these explicitly in evaluating alternatives and specific steps within the taxonomic process. (Anderson, 1974, p. 59)

Whilst it might seem self-evident that making clustering (or taxonomy if you prefer) more "scientific" is desirable, it is not obvious how this can be achieved.

Our perspective is that clustering is, in machine learning terminology, "unsupervised classification." The aim in clustering is to assign instances into classes, which are not defined a priori, and which are supposed to somehow reflect the structure of the entities that the data represents. Note it is common in the broader, non machine learning literature, to use "classification" when talking of clustering (Bowker and Star, 1999, Farris, 1981).
Clustering is a very basic human activity. It has been argued at length how fundamental it is in understanding the world (Bowker and Star, 1999)[1]. It is often presumed that for any situation where clustering may used there is a single "right" clustering. The presumption seems to be based on the notion that categories exist independent of human experience and intent, which has been increasingly discredited — for a cognitive science perspective see (Lakoff, 1987).
Thus one may be tempted to take refuge in mathematics. But this too is dangerous:

> The strive for objectivity, repeatability, testability etc. is a perfectly right attitude as long as their proper place in the "hierarchy of aims" is maintained, but becomes very harmful if — as Polish proverb says — the nose is expected

[1] The reason why Borges' famous strange classification below strikes as so bizarre is precisely because it makes us wonder what on earth would it be like to understand the world in that extraordinary manner — "the impossibility of thinking that" (Foucault, 1970).

> These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia called the Heavenly Emporium of Benevolent Knowledge. In its distant pages it is written that animals are divided into (a) those that belong to the emperor; (b) embalmed ones; (c) those that are trained; (d) suckling pigs; (e) mermaids; (f) fabulous ones; (g) stray dogs; (h) those that are included in this classification; (i) those that tremble as if they were mad; (j) innumerable ones; (k) those drawn with a very fine camel's-hair brush; (l) etcetera; (m) those that have just broken the flower vase; (n) those that at a distance resemble flies. (Borges, 1942)

> to serve the snuff-box: if these *tools* dominate over the *purpose* of scientific research... (Hołyński, 2005, p. 487).

Clustering seems to hold a fascination for mathematicians and engineers and as a consequence there is a large literature on domain independent clustering techniques. However for several decades there has been criticism that the technologies the mathematicians and engineers develop do not appeal to people who use classification structures as a central part of their work (Farris, 1981). Users of classification method disagree in principle with the notion that clustering is a domain-independent subject:

> I suspect that one of the reasons for the persistence of the view that classification is subject-independent is that classificatory theorists have been largely insulated from sources that would inform them otherwise (Farris, 1981, p.213).

This lack of interest from users of fancy machine learning clustering methods is seemingly paradoxical. Supervised classification techniques *are* widely used and appreciated by many people solving real problems. And clustering seems to simply be "unsupervised classification." Supervised clustering can be easily made into a well defined problem with loss functions which precisely formalize what one is trying to do (and furthermore can be grounded in a rational way in the real underlying problem). The difficulty with unsupervised clustering is that there is a huge number of possibilities regarding what will be done with it. And depending on the use to which a clustering is to be put, it may be wonderful or terrible.

Interestingly there is an analogy we can draw with supervised classification. If one is not prepared to commit to a particular loss function (as a formal codification of the use to which your classifier will be put) one can still proceed — one can just try and estimate the underlying probability distribution; see the discussion of loss functions and ROC curves in (Reid and Williamson, 2009). This suggests that in so far as clustering is domain independent, it is about estimating probability distributions or functionals of them. If that is accepted, then it can be easily formalized using loss functions. Furthermore, the analogy suggests that in so far as clustering is not about estimating such functionals, it is *essentially* domain dependent and any evaluation measures must thus necessarily be also domain dependent.

Many of the arguments about the right way to cluster, or how to compare clustering methods are a side-effect of the fact that there are a very wide diversity of clustering problems. Even within a particular domain of application (say the classification of biological organisms), there can be very diverse (and opposing) views as to what a valuable classification is; for example compare *phenetics* which attempts to classify on the basis observable characteristics ignoring phylogeny with *cladistics* which is avowedly phylogenetic. (See the historical account of the battles in Hull, 1988.) The arguments between adherents of either of these camps are not resolvable (even in principle) by coming up with some domain-independent means by which clustering algorithms are judged — their conflict stems from a disagreement as to what is the real question that needs to be answered.

2 Different uses of clustering and their evaluations

Clustering is a pervasive activity with data and knowledge. It affects knowledge representation and discovery (Kwasnik, 1999). It defines infrastructures that have real political significance (Bowker and Star, 1999). It forms the basis for systematic biology, and the need for classification remains ever-present (Ruepp et al., 2004). It is done pervasively. Let us just sketch two very distinct purposes for clustering:

1. **Data preprocessing.** Clustering is used as an automated pre-processing step in a whole chain of processing steps. For example, we cluster customers and products to compress the contents of a huge sales data base before building a recommender system. Or we cluster the search results of a search engine query to discover whether the search term was ambiguous, and then use the clustering results to improve the ranking of the answers. In such situations, the whole purpose of clustering is to improve the overall performance of the system. This can be quantified, but it is intrinsically problem dependent how to interpret such numbers.

2. **Exploratory data analysis.** Here, clustering is used to discover aspects of the data which are either completely new, or which are already suspected to exist, or which are hoped not to exist. For example, one can use clustering to define certain sub-categories of diseases in medicine, or as a means for quality control to detect undesirable groupings that suggest experimental artifacts or confounding factors in the data.

Evaluating the usefulness of clustering algorithms

It seems natural that certain clustering algorithms might be "more useful" for data preprocessing than for exploratory data analysis, and that it is unreasonable to expect a general evaluation procedure for clustering algorithms which is application independent. This suggests that the methods for evaluating clustering algorithms have to be different as well. In the end, all methods of evaluation should try to assess whether the clustering algorithm is useful, not in an abstract sense, but for serving the purpose we have in mind. We do not really care how the clustering algorithm works, as long as it achieves the goal we have set. In particular, it is pointless to discuss whether clustering "is" density-level set estimation, information-theory, or graph theory. Any algorithm, no matter how it works internally, has to be useful for the particular application we have in mind. Let us just give a brief outline how evaluation procedures in the above mentioned applications might look like.

Clustering for data preprocessing. This is about the only setting where it is straightforward to evaluate a clustering algorithm, at least from a methodological point of view. One can interpret the clustering as just one link in a whole chain of processing steps. Put more extremely, the clustering (algorithm) is just one more "parameter" which has to be tuned, and this tuning can be achieved similarly as for all other parameters, for example by cross validation over the final outcome of our system as a whole. We do not directly evaluate the "quality" of the clustering, and we are not interested in whether the clustering algorithm discovers "meaningful groups". All we care about is the "usefulness" of the clustering for achieving our final goal. For example, to build a music recommender system it might be a useful preprocessing step to cluster songs or users into groups to decrease the size of the underlying data set. In this application we do not care whether the clustering algorithm yields a meaningful clustering of songs or users, as long as the final recommender system works well. If it performs superior when a particular clustering algorithm is used, this is all we need to know about the clustering step. Thus, in the data preprocessing setting it is clear how the clustering can be evaluated, at least in principle. How this can be implemented in practice will vary from application to application, and it still might be challenging. But at least, from a methodological point of view, it is clear what has to be done.

Clustering for exploratory data analysis.
Exploratory data analysis should "present the data to the analyst such that he can see

patterns in the data and formulate interesting hypotheses about the data" (Good, 1983). As far as we know, it has never been attempted to assess whether clustering in general (or a particular clustering algorithm) has the potential of being useful for exploratory data analysis. Additionally to the technical side how to perform clustering, this question also has a psychological aspect. After all, it is the human user who is going to explore the data and hopefully detect a pattern. It is a very interesting problem to try to evaluate clustering (algorithms) for exploratory data analysis. One idea could be to ask humans to use a particular clustering algorithm to generate hypotheses, and later go and evaluate these hypotheses on independent data. Major obstacles in this endeavor are how to evaluate whether a hypothesis is "interesting," and how to perform a "placebo clustering" as a null model to compare to.

In this setting, it is very likely that the particular choice of the clustering algorithm is not very relevant, compared to the design of the human computer interaction interface. Exploring data is mainly performed visually, and the visualization and data manipulation capabilities of the system will likely be responsible for success or failure of the attempt to discover structure in the data. Thus we believe it unlikely that a mathematical loss function can be found that is an accurate proxy for the overall perceived quality of such a system.

Evaluation procedures that are not particularly helpful

In most papers about clustering algorithms, one or several of the following methods is used to show the success of a clustering algorithm:

Evaluation on artificial data sets. Clustering algorithms are applied to artificial data sets, for example points drawn from a mixture of Gaussians. Then the clustering results are compared against the "ground truth". Such a procedure can make sense to evaluate the statistical performance of a clustering algorithm under particular assumptions on the data generating process. It cannot be used to evaluate the usefulness of the clustering.

Evaluation on classification benchmark data sets. Clustering algorithms are applied to classification data sets, that is data sets where points come with class labels. Then the class labels are treated as the ground truth against which the clustering results of different algorithms are compared. High agreement with the ground truth is interpreted as good clustering performance.

In our opinion, this approach is dangerous. The underlying assumption is that points with the same class labels form clusters. This might be the case for some data sets but not for others. There might even exist a more natural clustering of the data points which is not reflected in the current class labels. Or, as it is often the case in high-dimensional data, different subspaces of features support completely different clusterings. It might very well happen that a clustering algorithm discovers a very reasonable clustering, but achieves a very bad classification error. So classification error by itself cannot be used as a valid score to compare clusterings.

If, on the other hand, clustering is performed as a pre-processing step for a classifier to be used later on, then it does not make sense to evaluate the classification performance of the clustering algorithm. As described above one needs to evaluate the whole processing chain.

Evaluation on real world data sets. Sometimes people run their algorithm on a real data set, and then try to convince the reader that the clusters "make sense" in the application[2]. For example, proteins are grouped according to some known structure. This is more or less a qualitative version of the approach using benchmark data sets. It can make sense if the clustering algorithm is intended for use in exploratory data analysis in this particular application, but does not carry any further meaning otherwise.

Clustering quality scores. There exist many different scores intended to evaluate the results of clustering algorithms. Typically, these scores take into account within-cluster similarity, between-cluster similarity, and the sizes of the different clusters. We argue that all these scores are unsuitable to evaluate the quality of clustering algorithms in an objective, domain-independent way. Scores are useful on the level of algorithms where they can be used as an objective function in an optimization problem, and it is a valid research question how different scores can be optimized efficiently. However, across different algorithms these score tell only little about the usefulness of the clustering, and for every score preferring one algorithm over the other one can invent another score which does the opposite. A unique, global, objective score for all clustering problems does not exist.

Statistical significance

So far we have been talking about the usefulness of a clustering. Where do considerations about statistical significance come into play? In a statistical setup, the basic assumption is that the given data points are samples from some underlying distribution. There are many data sets where such an assumption makes sense (customers are samples from the "set of humans"; a particular set of hand-written digits just contains a few instances out of a much larger set of "all possible hand written digits"). In such a setting, statisticians are concerned with the statistical significance of the results obtained on finite data sets. In the case of clustering: if we would draw another sample of data points, would we construct similar clusters as on the data set at hand? Or do our clusters just "fit noise" and are completely meaningless? Can we assign a confidence score to a particular clustering of a particular data set which tells us how confident we are that the clustering shows true structure of the underlying space? In many branches of science, it is a strong requirement to report such confidence scores. For example, in biology one hardly encounters papers where clustering plays a fundamental roles and no "p-value" is assigned to the clustering.

But how does this statistical aspect fit into our discussion above? Is the "significance" of a clustering independent of its "usefulness"?

In the preprocessing setting, statistical considerations do not play any role, as long as the system works. If the clustering algorithm gives different results on different samples, but the system works on either of these results, then we are fine. For example, many people use k-means to cluster a huge set of texts, say, in a set of manageable size. If we want to cluster a data set of 10^6 items into 10^3 clusters, it does not really matter where these clusters are, as long as they serve for the compression. We can have a completely different clustering each time we get a new data set, and the overall system might still work fine.

Interestingly, it is in the exploratory setting that statistical significance is important. After all, a user does not have infinite time to inspect all sorts of meaningless clusterings, and

[2] Compare Farris: "A clustering method is selected in each application for its ability to manufacture a grouping most in accord with the subjective feelings of a "professional taxonomist." (That taxonomist, of course, will then claim vindication of his views; they have been verified by an "objective" method!) One must wonder what value might be attributed to a method chosen primarily for its failure to contradict preconceptions (Farris, 1981, p. 208)"

we cannot hope to generate a meaningful hypothesis from nothing. This insight is quite striking: it is the “soft” exploratory data analysis setting where we have the “hardest” statistical requirements for our clustering algorithms.

In this sense, there are some applications of clustering where statistical significance is a necessary requirement for an exhilarative clustering algorithm to be useful. We want to stress that statistical significance alone is never sufficient, in the end we need the usefulness in the sense discussed above.

3 A systematic catalog of clustering problems

We have seen that clustering is used in a variety of contexts with very different goals and that clustering results cannot be evaluated without taking into account this context. Johnson explains at length the inadvisability of hoping for general classifications (suitable for all possible uses)[3]:

> It has been repeatedly stressed . . . that the values of classifications should be assessed according to the range of their purposes (Johnson, 1968).

> Difficulties in identifying problems have delayed statistics far more than difficulties in solving problems. This seems likely to be the case in the future too. Thus it is appropriate to be as systematic as we can about unsolved problems. . . . Different ends require different means and different logical structures. . . . While techniques are important in experimental statistics, knowing when to use them and why to use them are more important. (Tukey, 1954)

So how can we now proceed? A straightforward way would be to compile a table of different clustering applications and corresponding evaluation procedures. A more effective approach might be to come up with a way to treat several clustering applications with similar methods, so one does not have to start from scratch for every new application. One way forward is to go to the meta level and systematically build a taxonomy or catalog of clustering problems (Hartigan, 1977). We believe that this should be done in a solution agnostic way: define the problem in a purely declarative manner without trying to say how it should be solved. We conjecture that such a taxonomy will be of considerable help. This is tantamount the the research program “left to the reader as an exercise” in I.J. Good’s self-referentially titled *The Botrylogy of Botryology* (Good, 1977).

We do not yet attempt to suggest how this catalog of clustering problems look like, but want to point out several “dimensions” of clustering applications which might turn out to be important in such an endeavor. These dimensions should be independent of a particular application domain or a particular clustering algorithm.

Qualitative — quantitative. A fundamental distinction is whether the clustering results are used in a quantitative or qualitative context. A qualitative context is often the one of exploratory data analysis. Here we are interested in certain properties of the clusterings, but not in any final scores. Examples for qualitative uses of clustering are:

[3] The fact that classification depends upon its use and there is no universal scheme possible is a remarkably old idea. Gilmour and Walters (1964, p.5) quotes Mercier (1912, p.152): “The nature of the classification that we make . . . must have direct regard to the purpose for which the classification is required. In as far as it serves the purpose, the classification is a good classification, however ‘artificial’ it may be. In as far as it does not serve this purpose, it is a bad classification, however ‘natural’ it may be.”

Detecting latent structure, defining categories in data for later use (e.g. different subcategories of a disease), verifying that there is structure in the data , verifying that no unexpected clusters show up (quality control), verifying that expected clusters are there. Examples for quantitative uses of clustering are data preprocessing, data compression, clustering for semi-supervised learning.

Exploratory — confirmatory. There exist two distinct branches of data analysis: confirmatory data analysis and exploratory data analysis (Tukey, 1977). While exploratory data analysis is used to discover patterns in data and to formulate concrete hypotheses about the data, confirmatory data analysis deals with the question of how to validate a given hypothesis based on empirical data. Clustering is employed in both contexts.

Unsupervised — supervised. Even though it sounds slightly paradox to machine learners, there exist different levels of supervision in clustering problems. Situations where one does not have a clue what one is looking for are rather seldom, and in many cases additional information can be used. For example in exploratory data analysis, the analyst usually has a good idea what he is looking for.

Inductive bias. In most applications of clustering, one has a "bias" what one is looking for. This bias affects the type of clusters one tries to construct. For example, in some applications the focus might be to join similar data points; for example, when detecting a chemical compound with similar properties to a given one. In other applications, it might be more important to separate different points, for example to identify emerging topics in a stream of news. Such preferences are called the inductive bias. Examples are (see Figure 1):

- compact clusters — chain-like clusters
- peak-based — gap-based
- flat clustering — hierarchical clustering

Meaningful categories — useful categories. Do we need to find clusters that represent some understanding of how the data were generated? Not necessarily. Be it for exploratory analysis or for preprocessing in a prediction task, groupings, which emerge from the data representation may be useful. For instance, in speech recognition, it is common to use vector quantization as preprocessing, a method making no assumption about how data were generated. Similarly, in image processing, connected component methods do not attempt to uncover the mechanism by which data were generated. However, in some applications, clustering can be understood as a method for uncovering latent data structure, and prior knowledge may guide use to elect the most suitable approach. For instance, phylogenies are usually modeled as a diffusion process and clustering approaches to such problems usually attempt to uncover the underlying tree/hierarchy, hence are tackled hierarchical clustering methods. On the other hand, if we can presume that the data were generated by a shallow process, there is no reason to use a hierarchical model. For instance, handwriting used to be taught with a few methods in the US, hence, handwriting styles could be clustered by assuming just a 2 level random process: drawing the method, then drawing the writer; Gaussian mixtures or k-means type of algorithms would lend themselves well to such problems. A third type of model is constraint-based and assumes symmetric interactions between samples (like in a Markov random field of an Ising model of magnetism). For instance, dress-code can be assumed to emerge from peer-pressure and result in clusters of people dressing in a similar way. Graph partitioning methods may lend themselves better to such cases. Evidently in either of these three cases (diffusion model, component mixture model, and "magnetic" model), methods of clustering not

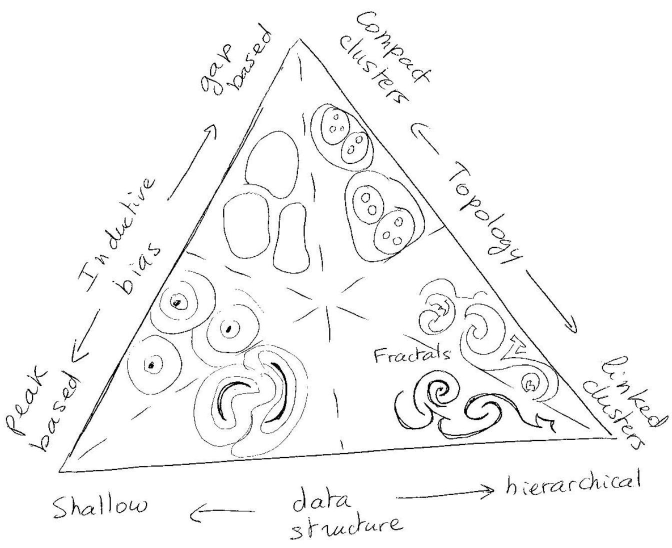

Figure 1: Some aspects of the inductive bias in clustering.

modeling the data generative process might "work well", if all we are interested in is making groups efficient for compressional of prediction. But the understanding that we gain is very different and it may make a lot of difference if we want to *predict the consequences of actions* or *devise policies to attained desired goal.* For instance, consider the problem of epidemiology. Patients with the same symptoms may be clustered assuming one of the three models described above. In the first case (hierarchical model), a genetic mutation may be responsible for a disease and tracing a population to a common ancestor may be useful to diagnose and treat patients. In the second case (mixture model), patients may be incurring disease because of one categorical factor of variability (environment, diet, etc.) like in the case of scurvy and lack of vitamin C. Note that this is a very real problem. People used to think before the "theory of germs" became well accepted after Pasteur's experiments that infectious diseases came from an inner predisposition, not from outside germs. Maybe an appropriate cluster analysis would have proved the theory of predisposition (presumably an inherited trait) wrong. In the third case (magnetic model), disease transmission by contagion can be well modeled in this way and a different type of disease prevention policy devised.

All the different properties mentioned above are inherent in the nature of the *problem* we want to solve with clustering, and our list is by no means exhaustive. We would like to stress that we are not looking for properties of clustering *algorithms.* For example, distinctions we do not believe to be helpful in our quest are: parametric vs. non-parametric, frequentist vs. Bayesian, model-based vs. model free, information-theoretic vs. probabilistic, etc.

What is unlikely to work

Universal "Benchmark data sets". The UCI approach to supervised classification[4] is bad enough. But given the diversity of end uses to which clustering is put, such an approach seems hopeless for clustering. One need only look at how problematic it is to study taxonomic repeatability within a particular domain to be daunted (Moss, 1971).

[4] Whereby there is a small fixed set of data sets, divorced from their real end use, that are used as a simple one-dimensional means of evaluating machine learning solutions.

Attempting to make things more "scientific" merely by throwing mathematics at it without thinking through the more general issues discussed above. Why won't this work? Well it did not in the past:

> "The theoreticians of numerical taxonomy have enjoyed themselves immensely over the past decade (though not without developing several schools with scant respect for each other!). The mushrooming literature is quite fascinating and new developments tumble after each other. Anyone who is prepared to learn quite a deal of matrix algebra, some classical mathematical statistics, some advanced geometry, a little set theory, perhaps a little information theory and graph theory, and some computer technique, and who has access to a good computer and enjoys mathematics (as he must if he gets this far!) will probably find the development of new taximetric methods much more rewarding, more up-to-date, more 'general', and hence more prestigious than merely classifying plants or animals or working out their phylogenies." (Johnson, 1968, p. 224).

> "The classificatory component of taxonomy cannot itself be made into a science by ill-founded philosophy of essentially arbitrary numerical procedures." (Johnson, 1968, p. 235)

We believe that if the clustering researchers want to have an impact in real applications (as it is, for example, the case for support vector machines for supervised classification), then it is time to start a dialog with practitioners and devise meaningful evaluation procedures for the different applications of clustering.

References

Sydney Anderson. Some Suggested Concepts for Improving Taxonomic Dialogue. *Systematic Zoology*, 23(1):58–70, March 1974. URL http://www.jstor.org/stable/2412240.

Jorge Luis Borges. El idioma analítico de John Wilkins. In *La Nación*. 8 February 1942. Translated and Republished as "John Wilkins' Analytical Language," pages 229–232 in *The Total Library: Non-fiction, 1922–1986*, Penguin, London, 1999.

Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, Mass., 1999.

James S. Farris. Classification Among the Mathematicians (Review of "Classification and Clustering," by J. Van Ryzin). *Systematic Zoology*, 30(2):208–214, June 1981. URL http://www.jstor.org/stable/2992422.

Michel Foucault. *The Order of Things: An Archaeology of the Human Sciences*. Random House, 1970.

J.S.L. Gilmour and S.M. Walters. Philosophy and classification. In W.B. Turrill, editor, *Vistas in Botany, Volume IV: Recent Researches in Plant Taxonomy*, pages 1–22. Pergamon Press, Oxford, 1964.

I.J. Good. The botryology of botryology. In J. van Ryzin, editor, *Classification and Clustering: Proceedings of an Advanced Seminar conducted by the Mathematics Research Center, The University of Wisconsin-Madison*, pages 73–94. Academic Press, 1977.

I.J. Good. The philosophy of exploratory data analysis. *Philosophy of Science*, 50(2), 1983.

J. A. Hartigan. Distribution problems in clustering. In J. Van Ryzin, editor, *Classification and Clustering: Proceedings of an Advanced Seminar conducted by the Mathematics Research Center, The University of Wisconsin-Madison*. Academic Press, 1977.

Roman B. Hołyński. Philosophy of science from a taxonomist's perspective. *Genus*, 16(4): 469–502, December 2005.

David L. Hull. *Science as a Process*. University of Chicago Press, 1988.

L.A.S. Johnson. Rainbow's End: The Quest for an Optimal Taxonomy. *Proceedings of the Linnean Society of New South Wales*, 93(1):1–45, 1968. Reprinted in *Systematic Zoology*, 19(3), 203–239 (September 1970).

Donald E. Knuth. Computer Programming as an Art. *Communications of the ACM*, 17 (12):667–673, December 1974.

B.H. Kwasnik. The role of classification in knowledge representation and discovery. *Library Trends*, 48(1):22–47, 1999.

George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, 1987.

Charles Mercier. *A New Logic*. William Heineman, 1912.

W. Wayne Moss. Taxonomic repeatability: An experimental approach. *Systematic Zoology*, 20(3):309–330, September 1971.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. URL `http://arxiv.org/abs/0901.0356`. arXiv preprint arXiv:0901.0356v1, January 2009.

A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539, 2004.

J. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1), 1977.

John W. Tukey. Unsolved problems of experimental statistics. *Journal of the American Statistical Association*, 49(268):706–731, December 1954.