

Pattern Discovery in the Crib of Procrustes

Alberto Apostolico*

University of Padova & Purdue University

Abstract. The study of physics purports to concise descriptors or theories, good at predicting a virtually unlimited set of replicas of a phenomenon of a certain nature. The discovery of patterns or structure in discrete objects pursues a similar goal, but it departs from the inference of physical laws in so far as the ensuing generation of unlimited replicas may be a curse rather than a blessing. Decades after the facts, an engineer turned computer scientist and still struggling with his math speculates about the origins of a physicist's fascination with the essence of complexity and structure; and how they can be inferred from examples. Which led to several and still largely unanswered questions, but ultimately helped shaping many a quest for a lifetime.

1 Introduction

We are taught that the *intensional* definition of a set consists of a list of attributes or qualities that uniquely intercept the elements of that set in a broader population or universe. By contrast, and not without some abuse of language, the *extensional* definition of a set is characterized, intensionally, as consisting of the exhaustive list of the members of that set. Likewise, the *extension* of a term is the collection of elements to which it is correctly applied, while its *intension* is the set of features which are shared by every element to which the term applies. Of course, these two descriptors are meant to be fungible: ideally, the intension of a set or term should accurately determine its extension. This bears pragmatic implications of enormous value: it empowers us with the ability to decide, for each newly-encountered item, whether or not it has all the relevant features shared among the objects defined by a term. Intriguingly, as the intension of a term is increased by the more detailed specification of features, the extension of that term tends to decrease, since fewer items now qualify for its application. In Logic, the collection of the attributes in a term is associated with the notion of *connotation*, whereas the collection of objects designated by that term is associated with *denotation*.

* Dipartimento di Ingegneria dell' Informazione, Università di Padova, Padova, Italy and Department of Computer Sciences, Purdue University, Computer Sciences Building, West Lafayette, IN 47907, USA. Work Supported in part by an IBM Faculty Partnership Award, by the Italian Ministry of University and Research under the National Projects FIRB RBNE01KNFP, and PRIN "Combinatorial and Algorithmic Methods for Pattern Discovery in Biosequences", and by the Research Program of the University of Padova. axa@dei.unipd.it

Like the brain, computers appear to be troubled by the denotation of infinite collections or aggregates, a fact perhaps reflected in the school of thought that equates intensionality with intentionality: *intentional* objects (interpreted as the objects of thought) must have intensional properties, the handle for connotation [11]. Thus, in order to serve as effective prostheses of the brain, computers must handle connotations.

2 The powers of abduction

In order for us to ripe the *predictive* powers of intensionality, we must first come to grasps with the way it relates to extensionality. It would appear that the nature of this relationship varies with the case. Still speaking in vague terms, the transition from one mode to the other entails some machinery or tool capable of enabling if not implementing the transition itself. But such a machinery incarnates with varying degrees of power, whereby one could say that some intensional definitions are more extensionally oriented than others. When it comes to mathematical objects, for example, an extensional proclivity seems ingrained in that discipline. A theorem tolerates no exception, whereas one single counterexample will suffice to confute it. By contrast, no example can establish the validity of a theorem, unless all cases were checked.

Extensional orientation in physics is generally credited to Galileo, who paid dearly for it. The physical law encapsulates within some concise descriptors or theory a virtually unlimited set of replicas of a phenomenon of a certain type. Thus, a Law of Physics is not only a mathematical relationship among measurable quantities related to the state and properties of bodies. One might also say that there is an actual machinery that makes sure the intensional law will be implemented. (One could argue whether this is also intentional with “t” replacing “s”, but this goes beyond our scope.)

Disciplines in the natural sciences tend to be more troublesome. The periodic table is full of exceptions, the classification of the species is still being debated, and there is no automated fool-proof rule that takes from the set of a patient’s symptoms to his diagnosis.

In an extensionally oriented world, the two classical pillars of reasoning, deduction and induction –the second one interpreted here in the non-mathematical sense of probable argument– are not of much help in our utilitarian pursuit of predictability. Deductive inference takes from ensemble to specimen, from population to sample, whereas inductive inference follows the opposite direction, from specimen to ensemble, from sample to population¹. Neither process is really conducive to discovery. As we know, C.S. Peirce [13,14] introduced a third way, which he called *abduction* (or retroduction, or hypothesis). This is patterned

¹ The limitations of deduction appear pretty obvious today, less so in a society still impregnated with the remnants of Scholastics such as the one to which Francis Bacon (1620-1690) introduced induction as an outright rebellion against the principle of authority, and “the tendency of the mind to construct knowledge-claims out of itself”, upon which deductive reasoners fed.

after a re-shuffling of the terms in a syllogism, which results in a logical fallacy of the form: All M 's are P 's (rule); All S 's are P 's (result); therefore, All S 's are M 's (case). Rephrased in the jargon of sampling theory, the argument sounds:

All rabbits came from the magician's hat;
all doves in this particular random sample came from the magician's hat,
and thus all doves are rabbits.

As a matter of fact, abduction is not about drawing conclusions as much as it is about building (educated) hypotheses, on the basis of the patterns that can be observed in a phenomenon [14]. Whereas deduction and induction represent two types of symbolic logic, abduction is a form of critical thinking bearing considerable practical yield, as it intertwines with the formation and testing of theories: the unexpected or surprising phenomenon P is observed; Among the hypotheses H_1, H_2, \dots, H_n , Hypothesis H_k is capable of explaining P . Therefore, H_k is pursued.

3 To generate and classify

The thriller "Smilla's Sense of Snow" by Peter Hoeg² reminds us that the likes of Eskimos, Greenlanders and Lappons enjoy over a dozen distinct nuances of snow³. Although, even at our latitudes, it is said that no two snowflakes are identical, we only got one word for snow. One favorite line of E.R. Caianiello was that kids brought up in a highly polluted metropolitan areas thought that black was the color of snow. S. Watanabe⁴ was not sure how did this fit in with his dismal results in the foundations of statistical classification.

² Delta; Reprint edition (1995)

³ Anthony C. Woodbury from University of Texas at Austin set up a compilation (July 1991) from Steven A. Jacobson's Yup'ik Eskimo dictionary (Alaska Native Language Center, University of Alaska, Fairbanks, 1984). This includes: qanuk for 'snowflake', qanir-, qanunge- and qanuggir- for 'to snow'; kaneq for 'frost'; kaner- for 'be frosty/frost'; kanevvluk and the corresponding verb kanevcir- for '(to get) fine snow/rain particles; natquik 'drifting snow/etc', of which the corresponding action is natqu(v)igte-; nevluk and the verb nevlugte- for 'clinging debris'...'lint/snow/dirt...'; aniu, apun, qanikcaq for 'snow on ground'; to which there correspond the actions qanikcir- and aniu-. We also have muruaneq for 'soft deep snow'; qetrar- and qerrettar- 'for snow to crust'; nutaryuk for 'fresh fallen snow on the ground' qanisqineq for 'snow floating on water'. And so on.

⁴ Satoshi Watanabe, Japanese theoretical physicist and one of the founding fathers of Pattern Recognition. My inquiry: "Would you think that the fact that kids in NYC believe that snow is black is an instance of the theorem of the ugly duckling" was tinted with surrealism: our conversation was taking place in a dusty office atop the temporary site of the University of Salerno, an important Soccer Championship game was being broadcast all over town, and thus hardly any other soul inhabited the place at that moment besides the two of us. NYC was much more polluted at the time than it is today

In trying to feed intensional definitions to a computer we realize that these come in two main flavors. In one case, we give some *characteristic* vector of *features*, where each positions of the vector is assigned to tag the presence or absence of a specific feature or property. Objects having vectors that are identical or similar by some measure coalesce in a cluster or class. The other way, which we will recapture later, is to give a *generating process* capable of building all and only the objects in a class. The Theorem of the Ugly Duckling states that if all features in our feature vector are given equal value then the Ugly Duckling is just as similar to a swan as another swan [27]. Thus, our ability to classify rests on our bias, which explicates itself by way of distributing weights unevenly among the various features, on a subjective basis. Much in line with this philosophy, Watanabe was an advocate for Kahrnen-Loewe Transforms [27], due to the ability of that expansion to self-extract the relevant features. One could argue that the Greek knew this (and much more) all along: “Πάντων ξρημάτων μέτρον ο άνθρωπος”⁵, Protagoras of Abdera (485-415 BC) had clearly spelled out in his effort to reconcile existence and change, and then sealed it across for good measure: “of those that exist as well as of those that *do not* exist”. Much later, J.L. Borges would come up with his massively quoted parody of taxonomy⁶ which “deeply shattered” Michel Foucault and others, similarly involved in the quest for “the order of things”.

As said, an alternative notion of class may be based on generative rules, whereby objects endowed with a similar structure must also be assembled in a similar way. Brillouin [7] devoted considerable attention to the problem of defining and measuring the information embodied in structure with the tools of Information Theory. As is well known, Shannon’s theory equates information with surprise, whereby a telegram on a wedding night is not informative [6,24,25]. By this measure, however, a library would become more informative if all books were cut into small fragments to be then tossed up in the air. Still, early classifications of genetic sequences based on measures of self-information claimed some success [17], and at any rate the pursuit of similarly global measures is likely to be revived out of necessity in massive comparative analysis of whole genomes. Brillouin argued that structure is better captured by the redundancy or *negentropy*, which is formally the opposite of Shannon’s information, thereby instituting an interesting duality between the characters of information in stor-

⁵ Man is the measure of everything.”

⁶ In “The Analytical Language of John Wilkins,” from *Otras Inquisiciones* (Other Inquisitions 1937-1952, London: Souvenir Press, 1973), Jorge Luis Borges deals with the predicament of classification: “These ambiguities, redundances, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote 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) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.”

age and in transmission. Kolmogorov [19,22] proposed alternative measures of information, more akin to the structure embodied in objects. For a string x , this measure is expressed as the length of a shortest program that would take a universal computer to synthesize x from scratch. Thus, strings that are hard to synthesize or unpredictable are more random (for this notion, see in particular [26]) than the easier ones. Along these lines, we have a good theory about a series of observations or phenomena whenever we can come up with a generating mechanism or law that is much shorter than the raw series. Unfortunately, the conclusions reached by Kolmogorov seem pendant to the Ugly Duckling: as long as we don't put our own bias, by way of privileging some regularities over others, we won't be able to come up with a theory that is the most compact possible one. Fortunately, morphology is so finely ingrained and functionally implicated in nature that we can adopt morphological structure as the bias meaningfully in many cases of interest. This leads to privileging *syntactic* regularities such as unveiled by grammatical inference techniques (see, e.g., [15,16]), where objects are characterized on the basis of the generative power [12] of minimum description grammars and by the corresponding discriminating ability of acceptor automata. In line with the duality of information mentioned earlier, the very same structural redundancies that are exposed by some such minimal grammars can be exploited in turn to achieve data compression. On these grounds, Lempel and Ziv founded [20] some of the most innovative, effective and elegant data compression methods known to date.

To bridge the statistical and grammatical approaches to classification is an intriguing and so far largely neglected task. Some thought-provoking reflections of the fact that some such bridges might in fact exist arise when we conjugate natural evolution with the formalism of math and computation. This should not come as a surprise. After all, living organisms are the most complex instances of communication, control and computing known to mankind. One of the goals of computational molecular biology is to develop a molecular taxonomy, that is, to derive a classification of species from the structure of their genetic material. In this pursuit, the degree of homology or proximity of a common ancestry may be assessed on the basis of a measure of similarity of genetic sequences [5]. In some relaxed sense, the sequences are treated as feature vectors except that the positions of the individual features is known only approximately. (To compensate for this, the "values" taken by the "attributes" are the same at every positions.) In order to account for the similarity or distance among these vectors, a mechanism of *mutation* is now hypothesized which should lead from one sequence to the other within a minimum number of elementary changes or *edit* operations. At the outset, sequences that are connected pairwise by shorter edit scripts are believed to have evolved more closely. Thus, when cytochrome-c molecules⁷ are grouped on the basis of their pairwise similarity measured in terms of edit distance, we can have a glimpse of the evolutionary process that led

⁷ An ancient protein, essential to the production of cellular energy, which has undergone little changes in millions of years, so that one can look into yeast, plant or human cells and find similar forms of cytochrome-c. Because of its ubiquity and

to their development, as well as of the differentiation of species that went with it. In conclusion, this grouping of macromolecules –based on a generating process that seems just a restricted or specialized version of Kolmogorov’s notion of conditional information– is not far from a taxonomy that regards a bio-sequence as the feature vector representation of self.

Where should we look in an attempt to capture the formation of structure? It must be a spot where some sub-assemblies aggregate to form more complex units. Under the oxymoron: “The sciences of the artificial”, H. Simon [23] tried to pin-down the essence of complexity as part of his attempt at reconciling the analytic or descriptive categories of science with the synthetic or prescriptive categories of engineering. He thought that complexity is constituted by the recursive partition of a system into subsystems, which results from the un-escapable onset of hierarchy, dictated in turn by necessary constraints of stability⁸: hierarchical systems, such as the social, biological, symbolic, etc., are better at survival because their subsystems can settle more easily into intermediate stable states.

4 Procrustes, the sub-semigroup and the Emperor’s new map

The Procrustes⁹ algorithm [10] probably reflected a physicist compulsion to divide forces into “weak” and “strong”. Here the strong forces are those that keep aggregates together while the weak ones are the syntax or glue that holds them together. How can we find the strong forces? It seems reasonable to try and distill off sub-aggregates by looking for unexpectedly high degrees of cohesiveness among the atomic parts. The natural handle for this is statistical analysis. In most natural languages, a strong local cohesiveness among symbols characterizes and actually defines certain sub-structures such as phonemes or syllables, and thus it is precisely these units that may be expected to be exposed in terms of statistical over-representation. Thus, it can happen that the statistical maneuvers that isolate such sub-units in a language also lead to purely syntactic

early availability, cytochrome-c was among the first molecules to be used in studies of molecular taxonomy.

⁸ There might be some kind of interpretive leap in going from one level of the hierarchy to another. The failure of the Logic Theorist program to go much beyond the first two pages of “Principia Mathematica” exposed hierarchy in the engineering of math: Simon liked to spend time explaining his efforts to automate the derivation of “meta-theorems”, the tools that subliminally guide a mathematician and make him aware of what makes of a mere well formed formula also an actually interesting theorem –something a computer won’t grasp as easily.

⁹ Arguably the most interesting of challenges on Theseus’s way to fame, Procrustes (literally, “he who stretches”) was a bizarre host used to adjust his guests to their bed. He lured his victims to his house promising a great meal and a pleasant sleep in his very special bed which he promised to have the very peculiar property of adjusting in length to fit the occupant to the utmost of comfort. Such a unique “one-size-fits-all” feature, however, was achieved by either stretching or chopping off some of the guest’s limbs.

categories of the algebraic theory of formal languages. On the one hand, the syllables or phonemes separate the strong interaction from the weak one. On the other, they form the basis of a peculiar sub-semigroup of the semigroup¹⁰ of all strings on the alphabet of the language: like the symbols in the original alphabet, the syllables can generate by concatenation the set of all words in the dictionary and more; however, the spurious sequences generated by the syllables are not nearly as many as those generated by the original alphabet.

Using words repeatedly and interchangeably as the bed and the guest, Procrustes can now set himself up to the less despeakable task of discovering a *basis* for the primitive patterns of the dictionary. Once found, this basis can generate all past and future phrases ever to be proffered in the language. Problem is, the basis also generates concatenations of units, e.g., syllables, that do not correspond to well formed words and phrases. This is, because we don't seem to have a strict characterization of the dictionary that would be essentially shorter than the dictionary itself. Ideally, the discovery of patterns and association rules thereof would have to concern itself with the task of deriving, from exposure to hidden specimens and partial extensions, adherent intensional characterizations of objects. Since one of the goals is to apply the findings to tasks of classification and prediction, we should not find ourselves in a position to predict more than can happen. However, this kind of duplicity plagues still and rather ubiquitously the increasingly reiterated attacks to the task of pattern discovery, and yet we seem to have no other choice but to imagine Sisyphus happy¹¹.

Because strings constitute such primeval structures and a natural habitat or embedding for most any phenomenon, it is only natural that a formal study of patterns would start with them. Pattern discovery on strings is flourishing by building on a repertoire of sophisticated techniques and tools developed since the early Seventies in pattern matching [3,4]. In the typical problem of pattern matching, e.g., string searching, we are given a text x and a pattern y and the problem is to find all occurrences of y in x . In pattern discovery, we know much less about what we are looking for. Ideally, we would know nothing and yet discover interesting patterns, but we already argued that this rests on shaky grounds methodologically as well as computationally. Some milestones of syntactic string pattern discovery *ante litteram* date back to the beginning of the last century, when Axel Thue showed that over an alphabet of more than two symbols it is possible to write indefinitely long strings not containing any "square", i.e., any pattern of the form ww ¹² (see, e.g., [21]). In a pair of sequences or a sequence family, it is of interest sometime to find a longest common substring,

¹⁰ Elementary algebraic structure consisting of a set and a commutative binary operator, e.g., the set of all finite strings over an alphabet together with the concatenation operator.

¹¹ "Il faut imaginer Sisyphe heureux" is the closing line of *Le Mythe de Sisyphe*, A. Camus' metaphore of the human experience (Gallimard, Paris, 1985). Cf. also J. Monod, *Le Hasard et la Nécessité*, (Seuil, 1973).

¹² On the binary alphabet, a square is an *unavoidable* regularity, while forming a cube www can be avoided. In view of the many tandem repeats that affect the genetic code, it may sound reassuring that the alphabet there contains four letters.

or a longest common subsequence, where the former is a (longest) string w such that any string x in the family may be written as $x = vwz$, and the latter is a (longest) string w that is obtainable from any string in the family by deletion of zero or more consecutive characters.

Whether some prior domain-specific knowledge is given or not, the tenet is that a pattern or association rule that occurs more frequently than one would expect is potentially informative and thus interesting. Accordingly, patterns are sought that are more frequent than one would expect in either one string or a set of strings. To assess the interest of a pattern, measures such as the *quorum* or number of *colors* (i.e., how many sequences contain each one instance of the pattern) or occurrences may be used. Central to these developments is also the notion of an *association rule*, which is an expression of the form $S_1 \rightarrow S_2$ where S_1 and S_2 are sets of data attributes endowed with sufficient *confidence* and *support*. Sufficient support for a rule is achieved if the number of records whose attributes include $S_1 \cup S_2$ is at least equal to some pre-set minimum value. Confidence is measured instead in terms of the ratio of records having $S_1 \cup S_2$ over those having S_1 , and is considered sufficient if this ratio meets or exceeds some pre-set minimum. Since the generation of all candidate associations would be prohibitive in most cases, one uses the observation that the terms of a frequently occurring association must be frequent in their own merit. Resort to abduction is widespread there. For example, having determined that some known enactors of protein translation, the so called promoters, contain significantly over-represented patterns, we find it natural to look for over-represented patterns in trying to identify more promoters. Some Web searching engines [18] sift through daunting masses of documents on premises similar to those that drag us to the most overcrowded restaurant when shopping for a good meal in an unfamiliar neighborhood.

The statistical and syntactic approaches mingle in the effort, in ways that are often subtle and almost always dangerous. Suppose that we wanted to build a table to report, for all substrings in a textstring of n symbols, the number of occurrences of that substring. Since the number of substrings is of the order of the square of n (denoted $O(n^2)$), then the table would contain more entries than the raw data—a far cry from the concise synopsis we had in mind. Even limiting the table to substrings that are over-represented by some measure is not a guarantee that the table would be smaller than the text. In fact, we typically stipulate our stochastic assumptions before bringing in the observable. Likewise, consider the problem of finding, for a given textstring x of n symbols and an integer constant d , and for any pair (y, z) of subwords of x , the number of times that y and z occur in tandem (i.e., with no intermediate occurrence of either one in between) within a distance of d positions of x . In principle there might be n^4 distinct subword pairs in x ! Luckily, an astonishing result of combinatorics on words tells us that the number of states in the finite automaton that recognizes all substrings of a string is linear in the length of the string [8]. A practical consequence of this is that the $O(n^2)$ substrings can be partitioned into a number of equivalence classes that is only linear in n , in such a way that the strings in a class have

precisely the same set of occurrences. Clearly, for two strings to be in the same class one must be a prefix of the other. In the light of this, it suffices in the above to consider a family of only n^2 pairs, with the property that for any neglected pair (w', z') , there is a corresponding pair (y, z) contained in our family and such that: (i) w' is a prefix of w and z' is a prefix of z , and (ii) the tandem index of (w', z') equals that of (w, z) .

The situation looks more dismal when searching for patterns with “don’t care” characters, which are symbols capable of taking up any value from the alphabet. A little reflection establishes that the escalation there is exponential. Assume that on the binary alphabet both *aabaab* and *abbabb* are asserted as interesting patterns. We can give a concise description of both by saying that *a_ba_b* occurs in the string, with “_” denoting the don’t care. By this, however, we have immediately generated the spurious patterns *aababb* and *abbaab*. This problem is reflected in all approaches that resort to “profiles” or weighed matrices in which the i -th column describes the percent composition of the i -th character in a pattern. In these and similar instances, the model seems to introduce more things in our philosophy than are dreamed of in heaven and earth¹³.

In his “Viajes de Varones Prudentes”¹⁴ (Libro Cuarto, Cap. XLV, Lerida, 1658), a J. A. Suárez Miranda narrated by J.L. Borges writes that

... En aquel Imperio, el Arte de la Cartografía logró tal perfección que el mapa de una sola Provincia ocupaba toda una Ciudad, y el mapa del Imperio, toda una Provincia. Con el tiempo, esos mapas desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el tamaño del Imperio y coincidía puntualmente con él. Menos adictas al Estudio de la Cartografía, las generaciones siguientes entendieron que ese dilatado Mapa era Inútil y no sin impiedad lo entregaron a las inclemencias del Sol y de los Inviernos. En los Desiertos del Oeste perduran despedazadas ruinas del Mapa, habitadas por animales y per mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas.

We face widespread and growing risks of building maps bigger than life.

¹³ “There are more things in heaven and earth, Than art dreamt of in your philosophy”- W. Shakespeare, Hamlet, I, v [76].

¹⁴ The piece was written by Jorge Luis Borges and Adolfo Bioy Casares. English translation quoted from J. L. Borges, *A Universal History of Infamy*, Penguin Books, London, 1975: “... In that Empire, the craft of Cartography attained such perfection that the map of a single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of time, these extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same scale as the Empire and that coincided with it point for point. Less attentive to the study of Cartography, succeeding generations came to judge a map of such magnitude cumbersome, and, not without irreverence, they abandoned it to the rigours of Sun and Rain. In the Western Deserts, tattered fragments of the Map are still to be found, sheltering an occasional beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.”

5 Epilogue

There has been a time when apples were just fruit and windows home fixture. Not only AmazonTM, EbayTM and GoogleTM, but even the Personal Computer did not exist at the time of the Procrustes Algorithm. Now, there is at least one department in which the Emperor could use a new map. In the past decade, the framework of human activity has been reshaped forever by technologies ushering into teraflop (i.e., 10^{12} floating point operations per second) machines and data volumes in the terabyte, even petabyte range (1 petabyte is 1 billion times 1 million characters). Enmeshed in a texture of ubiquitous computing, humans are going to be shared among many machines, and face unprecedented problems of knowledge formation, access, management and policy.

The nature and rate of these changes is quantitative on surface but will induce long qualitative leaps, forcing a transition from paradigms of search “by value” and search “by contents” to a new one of search “by meaning”, a paradigm yet to be explored. Automatic or semi-automatic generation of data and relationships thereof will take in an environment that is, like with the famous Heraclitus’ river, never twice the same, and a whole new science and engineering of automated discovery will have to take shape, of which the grounds are just beginning to be laid. The mechanics of discovery, and scientific discovery at that, will undergo changes perhaps comparable to the scientific revolution itself. The implications brought about by such a dramatic change in perspectives have barely begun to be perceived. We shall argue next that even making all this data and information *accessible* leaves still an issue of making it *accessed*. Unless of course we wish to take the view that – to paraphrase the central problem of Episteme – Knowledge might form and exist out there without having to reside even once in a human mind.

A recent in-flight entertainment shows the proud new owner of the latest satellite receiver as he wastes his entire day in pushing the next-channel button in search of what he would like to watch best. The message could not be made more clear: gaining access to information requires an investment of time and resources that dominates over fruition, even precludes it altogether. We thus face a completely new scenario and a wild paradigm shift. The new scenario is that data and information accumulate at a pace that makes it no longer fit for direct human digestion. The paradigm shift is that the bottleneck in communication is no longer represented by the channel or medium but rather by the limited perceptual bandwidth of the final user.

“*Μηδέν εστί*”, goes the opening line by which Gorgias of Leontini (483-376 BC) recited his threefold nihilism: “Nothing exists, if anything existed it would not be knowable, if anything were knowable it could not be communicated”. More than two thousand years worth of episteme have yet to fully sort out the existence of reality outside of us. But if it will become conceivable that the act of knowing is inherently precluded to humans on account of their limited capacity or bandwidth, a modern Gorgias would have to come to terms with an even greater frustration:

Nothing is known;
if anything were known it could not be fetched,
if it could be fetched it could not be digested.

There are very few handles to cope with the information flood. Since there is little hope of implementing any semantics in the common sense of the word, one sure resort is to better understand the syntactic and combinatorial essence of patterns, their structure and organization, and how to go about their discovery [1,2]. There will be increasing need for new and improved techniques for the extraction of prominent features and relationships in data, for the inference of synthetic descriptions and rules, for the generation of succinct visualizations and digests.

At the end of this loosely organized digression, we come back full circle to problems of finding intensions in an extensionally oriented world. This is significant. Researchers know that their work is ultimately not about looking for answers. The unmistakable mark of talent, the reason most scholars strive day in and day out is not getting to know what is the answer, it is to understand what *was* the question. The recent 50-th anniversary issue of the *Journal of the Association for Computing Machinery* opens with an essay by Frederick P. Brooks, jr, entitled “The Great Challenges for Half Century Old Computer Science”[9]. The author gives a list of outstanding problems. Problem Number 1 is as follows:

Shannon and Weaver performed an inestimable service by giving us a definition of information and a metric for information as communicated from place to place. We have no theory however that gives us a metric for the information embodied in structure... ..this is the most fundamental gap in the theoretical underpinning of information and computer science. A young information theory scholar willing to spend years on a deeply fundamental problem need look no further.

We must imagine Procrustes happy.

References

1. Apostolico, A., Pattern discovery and the algorithmics of surprise. In *Artificial Intelligence and Heuristic Methods for Bioinformatics*, P. Frasconi and R. Shamir, Eds., IOS Press, pp. 111-127, 2003.
2. Apostolico, A., and Crochemore, M., “String Pattern Matching for a Deluge Survival Kit”, *Handbook of Massive Data Sets*, J. Abello et al, Eds. Kluwer Acad. Publishers, 151-194, 2001/2.
3. Apostolico, A., and Galil, Z., Eds., *Combinatorial Algorithms on Words*. Springer-Verlag Vol. ASI F 12, 1985.
4. Apostolico, A., and Galil, Z., Eds. *Pattern matching algorithms*. Oxford University Press, 1997.

5. A. Apostolico and Giancarlo, R., Sequence Alignment in Molecular Biology. *Journal of Computational Biology*, **5**, 2:173–196 (1998).
6. Ash., R., *Information Theory*. Tracts in mathematics, Interscience Publishers, J. Wiley & Sons, 1985.
7. Brillouin, L., *Science and Information Theory*. Academic Press, 1971.
8. Blumer, A., Blumer, J., Ehrenfeucht, A., Haussler, D., Chen, M.T. and Seiferas, J.: The Smallest Automaton Recognizing the Subwords of a Text. *Theoretical Computer Science* , 40: 31–55 (1985).
9. Brooks, Frederick P. jr, The Great Challenges for Half Century Old Computer Science *JACM* **50**:1, 25-26 (Special Issue: Problems for the Next 50 Years) (January 2003)
10. Caianiello, E.R. and Capocelli, R.M., On form and language: the Procrustes algorithm for feature extraction”, *Kybernetik.*, **8**:6 223-233, (1971).
11. Chisholm, R., "Intentionality", in Paul Edwards (ed.), *The Encyclopedia of Philosophy* (New York: Macmillan and Free Press), Vol. 3: 201-204, 1967.
12. Chomsky, N., Three models for the description of languages, *IRE Transactions on Information Theory* **2**:3, 113-124, (1956).
13. Crane, T., The Mechanical Mind: A Philosophical Introduction to Minds, *Machines, and Mental Representations* (London: Penguin): 31-37, 1995.
14. Burch, P., *Charles Sanders Peirce*, *The Stanford Encyclopedia of Philosophy*, (Edward N. Zalta, Ed.), <http://plato.stanford.edu/archives/fall2001/entries/peirce/>, Fall 2001.
15. Fu, K. S. and Booth, T. L., Grammatical inference: Introduction and survey – Part I. *IEEE Transactions on Systems, Man and Cybernetics*, 5:95–111, (1975).
16. Fu, K. S. and Booth, T. L., Grammatical inference: Introduction and survey — Part II. *IEEE Transactions on Systems, Man and Cybernetics*, 5:112–127, (1975).
17. Gatlin, L., *Information Theory and the Living Systems*. Columbia University Press, 1972.
18. Kleinberg, J.M., Authoritative sources in a hyperlinked environment, *Journal of the ACM*, **46**:5, 604–632”, (1999).
19. Kolmogorov, A. N., Three approaches to the quantitative definition of information. *Problemi Pederachi Inf.*, 1, (1965).
20. Lempel, A. and Ziv, J., On the complexity of finite sequences. *IEEE Trans. on information Theory*, 22:75–81, (1976).
21. Lothaire, M., *Combinatorics on Words*. Addison-Wesley, Reading, Mass., 1983. Also, second edition: Cambridge University Press, 1997.
22. Martin-Lof, P., The definition of random sequences. *Information and Control*, 9, (1966).
23. Simon, H.A., *The Sciences of the Artificial*, MIT Press, Cambridge, Massachusetts, 1969.
24. Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.
25. Shannon, C.E., Prediction and entropy of printed english. *Bell System Technical J.*, 50–64, (1951).
26. von Mises, R., *Probability, Statistics and Truth*. MacMillan, New York, 1939.
27. Watanabe, S., *Knowing and Guessing*. Wiley, New York, 1969.