Simplicity: a unifying principle in cognitive science?

Nick Chater[a] and Paul Vitányi[b]

[a]Institute for Applied Cognitive Science, Department of Psychology, University of Warwick, Coventry CV4 7AL, UK
[b]Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Much of perception, learning and high-level cognition involves finding patterns in data. But there are always infinitely many patterns compatible with any finite amount of data. How does the cognitive system choose 'sensible' patterns? A long tradition in epistemology, philosophy of science, and mathematical and computational theories of learning argues that patterns 'should' be chosen according to how simply they explain the data. This article reviews research exploring the idea that simplicity drives a wide range of cognitive processes. We outline mathematical theory, computational results and empirical data that underpin this viewpoint.

The cognitive apparatus finds patterns in the data that it receives. Perception involves finding patterns in the external world, from sensory input. Language acquisition involves finding patterns in linguistic input, in order to determine the structure of the language. High-level cognition involves finding patterns in information, to form categories and to infer causal relations.

Simplicity and the problem of induction

A fundamental puzzle is what we term the problem of induction: infinitely many patterns are compatible with any finite set of data (see Fig. 1). So, for example, an infinity of curves pass through any finite set of points (Fig. 1a); an infinity of symbol sequences are compatible with any subsequence of symbols (Fig. 1b); infinitely many grammars are compatible with any finite set of observed sentences (Fig. 1c); and infinitely many perceptual organizations can fit any specific visual input (Fig. 1d). Some patterns are more cognitively natural than others. But why? And are cognitively natural continuations reliable in prediction?

These illustrations are quite abstract; but, importantly, the same issue arises even if the input is arbitrarily rich: although some specific patterns will be eliminated by such enrichment, an infinite number of incompatible patterns will always remain. What principle allows the cognitive system to solve the problem of induction, and choose appropriately from these infinite sets of possibilities?

Any such principle must meet two criteria: (1) it must solve the problem of induction successfully; (2) it must explain empirical data in cognition. We argue that the best approach to (1) is to choose patterns that provide the simplest explanation of the data; and that this approach provides a powerful in-road to (2), in line with a long tradition of psychological research.

The physicist and philosopher Mach [1] proposed the following radical idea: that the cognitive system should (criterion i), and does (criterion ii), prefer patterns that provide simple descriptions of the data. Here, a description must allow the data to be reconstructed, and the simplicity of a description is measured by its length.

Mach's proposal traces its roots back to Ockham's razor; that, in explanation, entities should not be multiplied beyond necessity; and to Newton's statement in the *Principia* that we 'admit no more causes of natural things than are both true and sufficient to explain the appearances'. But to make Mach's proposal precise required a theory of description complexity, which necessitated awaiting further mathematical developments.

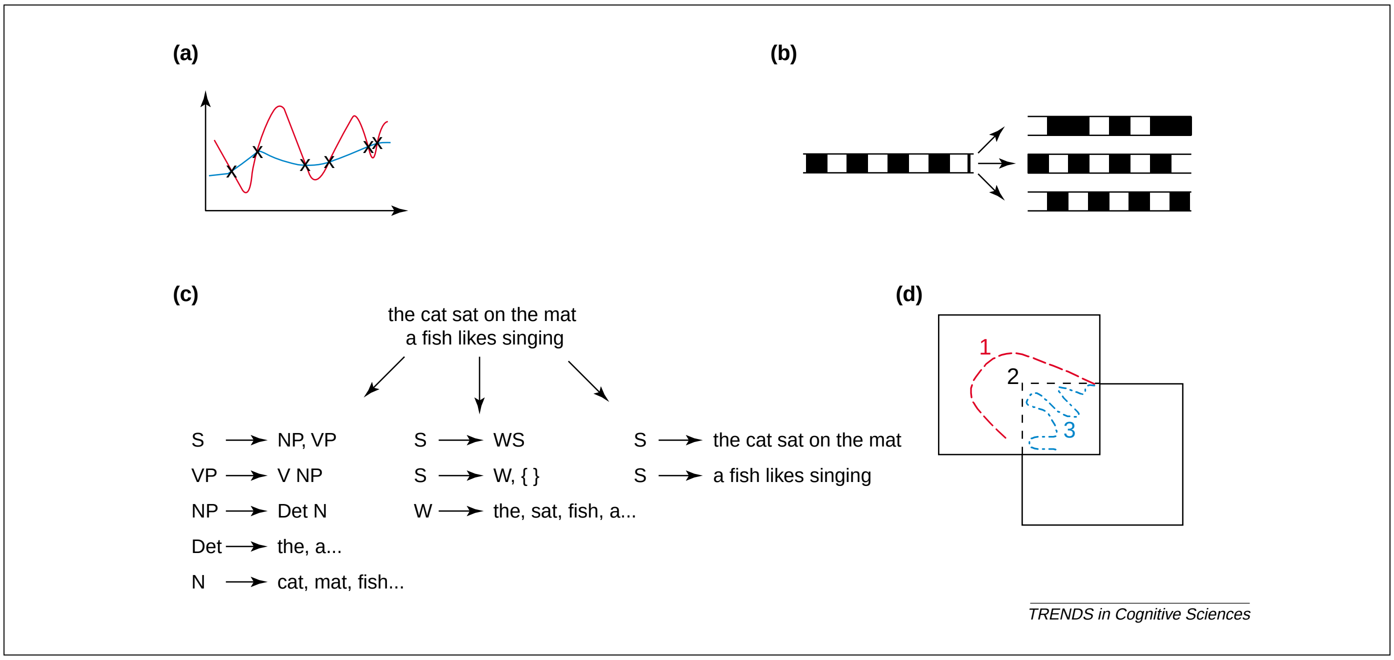
Quantifying simplicity

These mathematical developments came in two steps. First, Shannon's information theory justified $\log_2(1/p)$ as a code length for items with probability p. This is helpful for providing code lengths of highly repetitive data patterns, which can be assigned probabilities, such as low-level perceptual properties, phonemes, words and so on [2]. Second, the critical generalization to algorithmic information theory by Kolmogorov, Solomonoff and Chaitin defined the complexity $K(x)$ of any object, x, by the length of the shortest program for x in any standard (universal) computer programming language [3]. Surprisingly, it turns out that the choice of programming language does not matter, up to a constant additive factor. Moreover, algorithmic information theory turns out to agree closely with standard information theory, where the latter theory applies at all. Crucially, the algorithmic definition of simplicity applies to individual objects, whereas Shannon's definition depends on associating *probabilities* with objects.

Intuitively, then, we can regard the cognitive system's goal as compressing data: coding it in such a form that it can be recovered by some computable process (the mathematics allow that compression may be 'lossy' – i.e. information may be thrown away by the cognitive system, but we do not consider this here). Choices between patterns are determined by the degree of compression they provide – compression thus provides a measure of the

Corresponding author: Nick Chater (n.chater@warwick.ac.uk).

http://tics.trends.com 1364-6613/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S1364-6613(02)00005-0



TRENDS in Cognitive Sciences

Fig. 1. There are always infinitely many patterns compatible with any finite body of data. The general problem is illustrated in (a), where there is an infinite number of continuous functions that can be made to pass through a set of data points. (b) The same issue arises for discrete data: the alternating black and white squares on the left illustrate a sequence of binary data. But, as the right-hand side indicates, the overall pattern of which this data are a part could continue in any way. The 'middle' continuation is more cognitively natural. But why? (c) extends the point to grammar induction from a tiny 'corpus' of language data. Grammar 1 provides a linguistically reasonable analysis; Grammar 2 can produce any word sequence whatever and is clearly wildly overgeneral; Grammar 3 produces just the sentences in the corpus and nothing more. Human learners favour reasonable analyses; but why? (d) The limitless possible hypotheses for elaborating partial perceptual input. Only completion ii. (black dashed line) is seriously entertained, although i. and iii. are also compatible with the data.

strength of evidence for a pattern. This viewpoint forges potential connections between compression and pattern-finding as computational projects. Note that the shortest code for data also provides its least redundant representation; elimination of redundancy has been viewed as central to pattern recognition both in humans [4,5] and machines [6].

More importantly, formalizing simplicity provides a candidate solution to the problem of induction, described above. The infinity of patterns, all compatible with any set of data, are not all equal: the cognitive system should prefer that pattern that gives the shortest code for the data.

Regarding criterion (1) above, there are two beautiful and important mathematical results that justify this choice as a solution to the problem of induction [7]. One is that, under quite general conditions, the shortest code for the data are also the most probable (according to a Bayesian analysis, using the so-called 'universal prior'). A second result is that the shortest code can be used for prediction, with a high probability of 'convergence' on largely correct predictions. Finally, a third powerful line of justification for simplicity as an effective method of induction is its widespread use in machine learning [8,9] and statistics [10].

Simplicity as a cognitive principle

So, simplicity appears to go some way towards meeting criterion (1): justifying why patterns should be chosen according to simplicity. What about criterion (2)? Does simplicity explain empirical data in cognitive science? Table 1 describes a range of models of cognitive phenomena, from low- and high-level visual perception, language processing, similarity judgments, and mental processes in explicit scientific inference. The breadth of domains in which simplicity has proved to be a powerful organizing principle in cognitive modelling is encouraging.

But how does the simplicity principle stand up to direct empirical testing? This question is difficult to answer, for two reasons:

(1) *The representation problem*: Although, in the limit, and assuming the brain has universal Turing-machine power and Kolmogorov complexity is language invariant, many specific, non-asymptotic empirical predictions from simplicity depend on assumptions about mental representation, which will affect what regularities can be detected. And the mental representation of perceptual and linguistic stimuli is highly contentious in cognitive science.

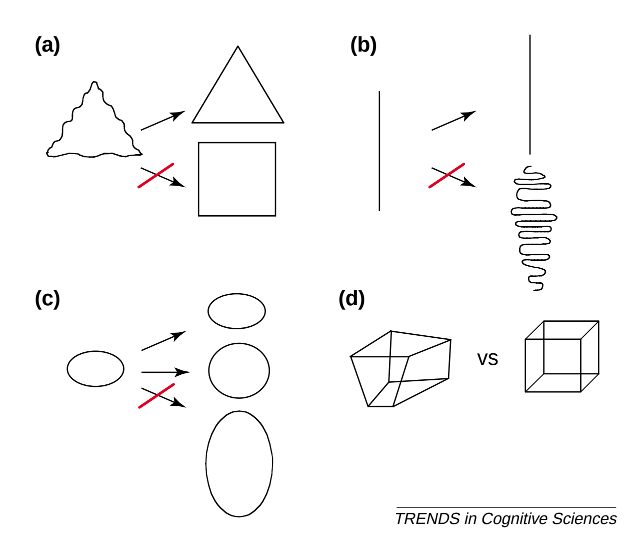
(2) *The search problem*: The cognitive system might prefer the simplest interpretation that it can find, but be unable to find a simple pattern of interest. Thus, without creating a full-scale cognitive model, involving assumptions about representation and perhaps also search, precise predictions from the simplicity viewpoint cannot be obtained [11].

There are, however, several lines of evidence that appear to be consonant with the simplicity viewpoint.

- A vast range of phenomena in perceptual organization, including the Gestalt laws of closure, good continuation and common fate, have been widely interpreted as revealing a preference for simplicity. Box 1 discusses some complex cases. The main theoretical alternative, the Bayesian approach to visual perception [12] is mathematically closely related to the simplicity principle [13].

Box 1. Empirical data

Various qualitative aspects of the resolution of perceptual ambiguity can be understood in terms of simplicity. In each of Fig. Ia–c, the left-hand side schematically represents a visual input, and the right-hand figure represents possible interpretations. Figure Ia illustrates that preferred perceptual organizations typically have a relatively good (although not necessarily perfect) fit with the data – here a somewhat consonant triangle interpretation is favoured over a very unlikely square interpretation. Patterns with good data-fit provide short codes for the data, given the pattern, and are preferred by the simplicity principle.



TRENDS in Cognitive Sciences

Fig. I. Interpretations of perceptual input. (See text for details).

Figure Ib illustrates the complementary preference for simple patterns: the 2-D straight line projected image is thus preferred to a highly irregular curve in the plane, even though, when viewed from one specific angle, this can project a perfect 2-D line. Figure Ic reveals the importance of precision in visual coding. The figure illustrates a preference for interpreting a small ellipse as that ellipse in the plane perpendicular to the viewer, rather than a larger, but geometrically similar, ellipse at a highly skewed angle (another possible interpretation is a circle, at a moderately skewed angle). Thus, data-fit, and, apparently, complexity-of-pattern appear identical here. How can the simplicity principle distinguish the two elliptical interpretations? The answer is that the projection is much more stable for the perpendicular ellipse; for the highly skewed ellipse the angle of orientation must be specified more precisely, costing additional code length, to obtain an equally good fit with the data. Finally, Fig. Id illustrates that simpler interpretations are taken to have causal significance. The right-hand 2-D figure is perceived as a projection of a wire cube; the left-hand figure is perceived as an irregular 2-D figure. Importantly, the joints of the wire cube are perceived as rigid, whereas the joints of the irregular 3-D figure are perceived as potentially flexible. The joints of the cube are perceived as rigid presumably because, otherwise, this 'simple' arrangement would be a remarkable coincidence (analogously, a sequence of 100 heads from a coin would be interpreted as indicating that the coin is biased). Thus, causal structure can be inferred on the basis of simplicity.

Qualitative demonstrations of this kind have also been supplemented by formal theories in psychology that seek to explain the interpretations of perceptual figures as minimizing code length [a,b].

References

[a] Hochberg, J. and McAlister, E. (1953) A quantitative approach to figure 'goodness'. *J. Exp. Psychol.* 46, 361–364

[b] Van der Helm, P.A. and Leeuwenberg, P.A. (1996) Goodness of visual regularities: a non-transformational approach. *Psychol. Rev.* 103, 429–456

- Items with simple descriptions are typically easier to detect in noise [2,11].
- The simplicity of a code for a stimulus quantifies the amount of structure uncovered in that stimulus. The more structure people can find in a stimulus, the easier they find it to process and remember [14] and the less random it appears [15].
- The speed of learning for Boolean concepts (e.g. A or B or C; A and (B or C), etc) is closely predicted by the shortest code length for those concepts [16].
- Similarity can be viewed as a function of the simplicity of the distortion required to turn one representation into the other. This viewpoint makes empirical predictions that are not captured by existing spatial or feature-based theories of similarity, but which have been confirmed [17].
- Shepard's Universal Law of Generalization [18], which implies that items have a probability of confusion that is a negative exponential function of the distance between them in an internal 'space', can be derived from the assumption that the psychological similarity between two objects is a function of the complexity of the simplest

Table 1. Pattern-finding by simplicity: a sample of research[a]

Cognitive process	Data	Codes	Computer science/ mathematical approaches	Cognitive science applications
Low-level perception	Sensoryinput/artificially captured images	Filters in early vision	Image compression [25]	Early vision as compression [23,22]
High-level perception	Sensory input/output of early perceptual Processing	Representations of higher level structure	Pattern theory [26]	Principle of economy [1] Perceptual organization [27,14]
Language acquisition	Linguistic input	Representations of language structure	Text compression [28]	Phonological [29] and morpholgical analysis [30], segmentation [31,24] and grammar induction [32,33]
High-level cognition	High-level representations of knowledge	Similarity, causal relations	Information distance [34] Gencompress [35]	Similarity as representational distortion [18] Categorization by compression [36]
Scientific inference	Scientific data	Theoretical knowledge	Machine induction systems [9] Foundations of statistics [10]	Ockham, Newton Mach's principle of economy [1] Formal measures of simplicity [37,38]

[a] Many pattern-finding problems have been successfully approached by mathematicians and computer scientists using a simplicity principle. In many of these areas, the simplicity principle has also been used as a starting point for modelling cognition.

transformation between them, with minimal additional assumptions [19].

- The physiology of early vision, including receptive-field structures, and phenomena such as lateral inhibition, seems adapted to maximize information compression from visual input [20]. On the other hand, both theoretical and empirical arguments suggest that, the brain also uses highly redundant 'sparse' neural codes for perceptual input [21,22].

Conclusion

Since Mach, a number of theorists have proposed the sweeping idea that much of cognition concerns compression [23], or the elimination of redundancy [24]. This 'simplicity principle' has been developed into a mathematically rigorous method for finding patterns in data [3], has served as the foundation for a broad range of cognitive models, and is consistent with a range of empirical data. We suggest that simplicity is worth pursuing as a potentially important unifying principle across many areas of cognitive science.

Acknowledgements

This work was partially supported by the Human Frontiers Science Foundation, EU Training Network HPRN-CT-1999–00065, the Leverhulme Trust, an ESRC/DTI Link Award, and by Oliver, Wyman and Company. We thank Khalilah Khan and Ursula Richards for invaluable help in preparation of figures and text.

References

1 Mach, E. (1959) *The Analysis of Sensations and the Relation of the Physical to the Psychical*, Dover Publications

2 Hochberg, J. and McAlister, E. (1953) A quantitative approach to figure 'goodness'. *J. Exp. Psychol.* 46, 361–364

3 Li, M. and Vitányi, P. (1997) *An Introduction to Kolmogorov Complexity and its Applications*, 2nd edn, Springer-Verlag

4 Attneave, F. (1954) Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193

5 Barlow, H.B. (1959) Possible principles underlying the transformation of sensory messages. In *Sensory Communication* (Rosenblith, W., ed.), pp. 217–234, MIT Press

6 Watanabe, S. (1960) Information-theoretical aspects of inductive and deductive inference. *IBM J. Res. Dev.* 4, 208–231

7 Vitányi, P. and Li, M. (2000) Minimum description length induction, bayesianism and kolmogorov complexity. *IEEE Trans. Inf. Theory* 46, 446–464

8 Quinlan, J. and Rivest, R. (1989) Inferring decision trees using the minimum description length principle. *Inf. Comput.* 80, 227–248

9 Wallace, C. and Freeman, P. (1987) Estimation and inference by compact coding. *J. Roy. Stat. Soc. Series B* 49, 240–251

10 Rissanen, J. (1989) Stochastic complexity and statistical inquiry. *World Scientific Series in Computer Science* (Vol 15), World Scientific

11 Van der Helm, P.A. and Leeuwenberg, E.L.J. (1996) Goodness of visual regularities: a non-transformational approach. *Psychol. Rev.* 3, 429–456

12 Knill, D. and Richards, W. (1996) *Perception as Bayesian Inference*, Cambridge University Press

13 Chater, N. (1996) Reconciling simplicity and likelihood principles in perceptual organization. *Psychol. Rev.* 103, 566–581

14 Garner, W. (1974) *The Processing of Information and Structure*, Erlbaum

15 Falk, R. and Konold, C. (1997) Making sense of randomness: implicit encoding as a bias for judgment. *Psychol. Rev.* 104, 301–318

16 Feldman, J. (2000) Minimization of Boolean complexity in human concept learning. *Nature* 407, 630–633

17 Hahn, U. *et al.* (2003) Similarity as transformation. *Cognition* in press

18 Shepard, R.N. (1987) Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323

19 Chater, N. and Vitányi, P. (2003) Generalized law of universal generalization. *J. Math. Psychol.* in press

20 Blakemore, C. (1990) *Vision: Coding and Efficiency*, Cambridge University Press

21 Gardner-Medwin, A.R. (2001) The limits of counting accuracy in distributed neural representations. *Neural Comput.* 13, 477–504

22 Olshausen, B.A. and Field, D.J. (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 37, 3311–3325

23 Wolff, J.G. (1982) Language acquisition, data compression and generalization. *Lang. Commun.* 2, 57–89

24 Barlow, H.B. *et al.* (1989) Finding minimum entropy codes. *Neural Comput.* 1, 412–423

25 Fisher, Y. (1995) Fractal Image Compression: Theory and Application. (Fisher, Y., ed.), Springer Verlag

26 Mumford, D. (1996) Pattern theory: a unifying perspective. *Perception as Bayesian Inference* (Knill, D., Richards, W. eds), pp. 25–62, Cambridge University Press

27 Leeuwenberg, E. and Boselie, F. (1988) Against the likelihood principle in visual form perception. *Psychol. Rev.* 95, 485–491

28 Bell, T.C. *et al.* (1990) *Modelling For Text Compression*, Prentice Hall

29 Goldsmith, J. (2002) Probabilistic models of grammar: phonology as information minimization. *Phonological Studies*, 5

30 Goldsmith, J. (2001) Unsupervised learning of the morphology of a natural language. *Comput. Linguist.* 2, 153–198

31 Brent, M.R. and Cartwright, T.A. (1996) Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125

32 Grünwald, P. (1996) Symbolic, connectionist and statistical approaches to learning for natural language processing. In *Lecture Notes in Artificial Intelligence 1040* (Wermter, S., ed.), pp. 203–216, Springer Verlag

33 Clark, R. (2001) Information theory, complexity, and linguistic descriptions. In *Parametric Linguistics and Learnability* (Bertolo, S., ed.), pp. 126–171, Cambridge University Press

34 Bennett, C.H. *et al.* (1998) Information distance. *IEEE Trans. Inf. Theory* 44, 1407–1423

35 Li, M. *et al.* The similarity metric. In *Proc. 14th ACM–SIAM symposium on Discrete Algorithms* (in press)

36 Pothos, E. and Chater, N. (2002) A simplicity principle in unsupervised human categorization. *Cogn. Sci.* 26, 303–343

37 Kemeny, J.G. (1953) The use of simplicity in induction. *Phil. Rev.* 62, 391–408

38 Sober, E. (1975) *Simplicity*, Clarendon Press