DANIEL DE MAGALHÃES CHADA

FROM COGNITIVE SCIENCE TO MANAGEMENT SCIENCE: TWO COMPUTATIONAL CONTRIBUTIONS

Escola Brasileira de Administração Pública e de Empresas - EBAPE

Fundação Getulio Vargas

FROM COGNITIVE SCIENCE TO MANAGEMENT SCIENCE: TWO COMPUTATIONAL CONTRIBUTIONS

DANIEL DE MAGALHÃES CHADA

B.Sc. Pontifícia Universidade Católica do Rio de Janeiro, 2007

Dissertação submetida como requisito para a obtenção do grau

Mestre de Administração

Maio 2011

BANCA EXAMINADORA:
Alexandre Linhares (Orientador - EBAPE-FGV)
Horácio Hideki Yanasse (INPE)
Rafael Guilherme Burstein Goldszmidt (EBAPE-FGV)

LOCAL: Rio de Janeiro

DATA: Maio 2011

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Edson de Campos Chada.

1937 – 1994

ABSTRACT

This work is composed of two contributions. One borrows from the work of Charles Kemp and Joshua Tenenbaum, concerning the discovery of structural form: their model is used to study the Business Week Rankings of U.S. Business Schools, and to investigate how other structural forms (structured visualizations) of the same information used to generate the rankings can bring insights into the space of business schools in the U.S., and into rankings in general. The other essay is purely theoretical in nature. It is a study to develop a model of human memory that does not exceed our (human) psychological short-term memory limitations. This study is based on Pentti Kanerva's Sparse Distributed Memory, in which human memories are registered into a vast (but virtual) memory space, and this registration occurs in massively parallel and distributed fashion, in ideal neurons.

RESUMO

Este trabalho é composto de duas contribuições. Uma se usa do trabalho de Charles Kemp e Joshua Tenenbaum sobre a descoberta da forma estrutural: o seu modelo é usado para estudar os *rankings* da revista Business Week sobre escolas de administração, e para investigar como outras formas estruturais (visualizações estruturadas) da mesma informação usada para gerar os *rankings* pode trazer discernimento no espaço de escolas de negócios nos Estados Unidos e em *rankings* em geral. O outro ensaio é de natureza puramente teórica. Ele é um estudo no desenvolvimento de um modelo de memória que não excede os nossos (humanos) limites de memória de curto-prazo. Este estudo se baseia na Sparse Distributed Memory (Memória Esparsa e Distribuida) de Pentti Kanerva, na qual memórias humanas são registradas em um vasto (mas virtual) espaço, e este registro ocorre de forma maciçamente paralela e distribuida, em neurons ideais.

Gratitude bestows reverence,
allowing us to encounter everyday epiphanies,
those transcendent moments of awe that change
forever how we experience life and the world.

— John Milton (1608 - 1674)

ACKNOWLEDGMENTS

Turns out this is the section upon which I put highest importance, but have the most trouble writing. I would like to thank my mother, Lúcia, whose tirelessness, hard work and dedication to her family are forever ingrained in my mind as the meaning of the word nobility. This work would not have been possible without her unwavering support.

I thank my future wife, Mariana, for her endless patience when I was frustrating, kindness when I was frustrated and support when ...always, actually.

I am unaware of how many advisees can honestly say they found great friends in their advisors, but I did in Alexandre Linhares. His capacity to shape an endless flow of raw creativity into highly original scientific work astounds me to this day. His support and optimism were undying when I convinced myself there were no results to be found, or when I found yet *another* bug in the software. I thank him for all this, and for making it all look easy and fun.

My uncle João Bosco and aunt Christina have been champions of my professional and academic development throughout my life, providing support when it was most needed. They are not my godfather and godmother by some twist-of-fate, but nonetheless hold places of honor and respect in my heart larger than any title can bestow.

I thank my colleagues-in-arms: Jarbas Silva, Ariston Diniz and more recently Marcelo Brogliato for countless enlightening conversations in our shared passion of science, computation and cognition.

And to Christian Aranha, my near-advisor through our years of work together: He brought me (almost dragged me, honestly) to the field to which I now hope to dedicate my career, and for that I will always be grateful.

I thank my fellow students at EBAPE/FGV for creating the wonderful atmosphere we shared throughout the course; and the staff at EBAPE/FGV for their professionalism and technical support.

I would also like to thank Eric Nichols for numerous valuable comments during the development of the SDM chapter.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

KT Kemp-Tenenbaum

SDM Sparse Distributed Memory

API Application Programming Interface

Part I

INTRODUCTION

1 COGNITION, DECISION AND ADMINISTRATION

One of the tenets of management science is to develop a comprehensive theory of human decision-making. While the rational-decision actor has been successful in modeling, and in bringing valuable insights into a number of decision scenarios, a number of studies have made clear that humans depart from rationality (for recent summaries, see [Smith e Winterfeldt 2004, Ariely 2008, Ariely 2010]).

Cognitive science, the study of human information-processing, is slowly filling the void between the need for formal models of human behavior and the numerous shortcomings of the rational model. The computational tools generated through the field's explorations into human behavior can provide new paradigms of pattern recognition, exploratory data analysis and information retrieval, all central to the aim of decision science.

For example, in the field of business strategy, [Gavetti et al. 2005, Gavetti e Warglien 2007] postulate that choice, in novel environments at least, is guided by analogy-making. Case studies are a highly popular tool in business education and their value often hinges on the understanding of one situation in terms of another. The literature provides a number of examples of analogical reasoning:

i) The chain "Toys 'R' Us", launched in the 1950s, was tied to the vision and success of supermarkets (the chain was effectively called "Baby Furniture and Toy Supermarket" at a particular point). Afterwards, the launch of the office store "Staples" was based on similar reasoning: "Could we create a Toys R Us for office supplies?" [Gavetti et al. 2005];

ii) In the 1980s the largest European carmaker decided to invest heavily in the U.S. market by introducing a large range of cars that were bestsellers in Europe. Before this decision, the carmaker had 63% of the imported car market in the U.S. But the carmaker was called Volkswagen. The American consumer's experience with Volkswagen consisted of the Beetle, an inexpensive and odd-looking car first sold in 1938. American consumers rejected the idea of a large, well-built, modern-looking, powerful and expensive Volkswagen. To make matters worse, the company decided to withdraw the Beetle from the market, and its share of the imported car market in the U.S. dramatically fell from 63% to less than 4%. The exact same cars were being sold in Europe and in the U.S; the only difference was in the consumers' experiences of what a "Volkswagen" meant. A twenty-thousand-dollar Volkswagen seemed, to Americans, like a practical joke. Similarly, "the new Honda" to an American consumer meant a new car model; to the Japanese, it meant a new motorcycle [Ries e Trout 1993].

iii) When Iranian Ayatollah Ruhollah Khomeini declared a fatwa (a death sentence) to writer Salman Rushdie, the Catholic Church did not stand for the principle of "Thou shalt not kill". It recognized its experience of trying to censor "The last temptation of Christ", a film, and sided with the Iranians. L'Osservatore Romano, a key Vatican publication, condemned Rushdie's book as 'blasphemous'. The Head of the French Congregation, Cardinal Decourtray, called it an 'insult to God'; Cardinal O'Connor from New York made it clear that it was

crucial to "let Moslems know we disapprove of attacks on their religion" [Hofstadter e FARG 1995, Linhares e Freitas 2010].

Decision-makers often obtain strategic insights by understanding one situation in terms of another; however, analogies are but one of the ideas from cognitive science that have crossed the bridge to management science. Neural Networks, or mathematical models of large-scale parallel processing in general, have found use in a number of more traditional management science domains, such as credit-risk evaluation [Piramuthu et al. 1998], the understanding of new product development [Natter et al. 2001], and consumer targeting [Kim et al. 2005], to name a few.

This thesis constitutes two essays on cognitive science. One, in chapter 2, borrows from the work of Charles Kemp and Joshua Tenenbaum, concerning the discovery of structural form: Kemp and Tenenbaum's model is used to study the BusinessWeek Rankings of U.S. Business Schools, and to investigate how other structural forms (structured visualizations) of the same information used to generate the rankings can bring insights into the space of business schools in the U.S., and into rankings in general.

The work of chapter 2 has an exploratory nature and is meant as a relevant incursion into a new concrete usage of cognitive modeling toward administrative practices. Its goal is to illustrate the inherent applicability of the discovery of structural form in exploratory data analysis.

The other essay, in chapter 3, is purely theoretical. It is a study to develop a model of human memory that does not exceed our psychological short-term memory limitations. This study is based on Pentti Kanerva's 'Sparse Distributed Memory' (SDM) [Kanerva 1988], in which human memories are registered into a vast (but virtual) memory space, and this registration occurs in massively parallel and distributed fashion, in idealized neurons [Linhares et al. 2011].

Before we enter into further details, the reader may ask: why these two particular topics? Why the work of Kanerva; or the work of Kemp and Tenenbaum? The response to this question gives a glimpse of the history of this manuscript. In brief, a future model of the human mind, and thus of decision-making, must incorporate features relating these two bodies of work. Human memory has numerous psychological characteristics reflected in Sparse Distributed Memory, and it is possible that this model, or a variation of it, explains how we register information at the most basic level.

Additionally, humans are able to discover different forms in our surroundings, from a linear ordering of aggressive primates, to the self-clustering of prisoners, to the hierarchical trees representing power in the inner circles of the White House. In other words, we are able to perceive not only relationships between entities, but also that these relationships are classifiable into different forms.

Perhaps both models involved in our study, separately and in their own rights, reflect human information-processing abilities. If this is the case, then future mathematical models of decision-making will do well to consider the subject matter which will be explored in the forthcoming chapters.

Part II

FORM, STRUCTURE AND DATA

2

THE MULTIDIMENSIONAL ROAD TO HARVARD OR, WHY RANKING BUSINESS SCHOOLS DOES NOT MAKE SENSE

Science is facts; just as houses are made of stones,
so is science made of facts; but a pile of stones is not a house
and a collection of facts is not necessarily science.

— Henri Poincaré (1854 - 1912)

2.1 INTRODUCTION

Business schools are regularly ranked by Business Week, The Economist, US News & World Report, Fortune, Financial Times, the Wall Street Journal, amongst many other organizations and periodicals. Their publication, beginning in the 1980s, have generated high controversy in the U.S. and abroad. These rankings exert deep influence across the business school landscape [Pfeffer e Fong 2004, Gioia et al. 2000, Gioia e Corley 2002, Corley e Gioia 2000, Alvesson 1990]. They directly affect the perceptions of current students, alumni and prospective students in regards to the quality of the ranked schools. Their influence has repercussions to the extent that schools alter their curricula, fire faculty, and adapt teaching methods with the explicit objective of rising in the ranks. Zell [Zell 2001] elaborates on this change of behavior since the rise of the business school rankings. Pfeffer and Fong [Pfeffer e Fong 2004] explain how i) business schools taylor their curricula in attempts to rise on the ranking; ii) professors "dumb down" their courses in order to receive better reviews from students; iii) the press (rather than academia) has led the way in defining standards of world-class business education and iv) the above points cause a standardization of business schools, which is detrimental to students (who lose options for different types of education) and to schools. Corley and Gioia [Corley e Gioia 2000] explain how “the rankings by these magazines have come to dominate many business schools’ sense-making and action-taking efforts”.

Dichev [Dichev 1999] questions the validity of rankings as a whole, concluding from a cross-rankings correlation that neither the Business Week nor the U.S. News rankings “should be interpreted as a broad measure of school quality and performance”, and that the “absence of positive correlation combined with reversibility in changes implies that one should avoid a broad interpretation of the rankings as measures of the unobservable ’school quality”’. Still others suggest alternate evaluation methods for schools, using different indicators to provide a ’better’ ranking system [Tracy e Waldfogel 1997] or better principles [Cornelissen e Thorpe 2002] in order to better reflect the qualities of each institution. (These, nevertheless, also impose the order structure, which is the critical point of focus here.)

While students and alumni generally regard the rankings as a valid metric of the quality and reputation of the schools, faculty and staff generally share a more adverse view of this system. Among the latter, they are viewed as poor quality indicators of the education provided

by an institution. Furthermore, studies show that there is virtually no correlation between a position in the ranking and academic production [Siemens et al. 2005, Trieschmann et al. 2000]. Other evidence shows that both the rankings themselves [Elsbach e Kramer 1996] and the changes caused by them [Pfeffer e Fong 2004, Zell 2001] elicit responses ranging from mild annoyance to outright rebelliousness amongst faculty. As a testament to the power and influence of rankings, there is empirical evidence showing the correlation between the rankings and the resignations of the deans of schools that score poorly on them [Fee et al. 2005]. From this body of literature one may conclude that rankings hold a huge sway over institutions and their strategies—and over the students' choices concerning which one to attend—despite their clear dissociation from any true measure of the quality of the education at each institution. Similar discussions arise from university–level rankings [Ehrenberg et al. 2001, Liu e Cheng 2005, Florian 2007, Ioannidis et al. 2007].

Yet, there are even deeper problems. Rankings, by their mathematical nature, create serious anomalies—either by placing dissimilar schools in close rank positions, or by placing similar schools in far rank positions. This can be illustrated with a simple example, consisting four schools and two (binary) dimensions, as we will see below.

2.1.1 *Rank anomalies*

A rank is a mathematical structure also known as an *order*: given two distinct entities ϵ_1 and ϵ_2, the statement $\epsilon_1 \prec \epsilon_2$ denotes that ϵ_1 *precedes* ϵ_2, or ϵ_1 *dominates* ϵ_2. The stated meaning in a school ranking is that if school ϵ_1 precedes school ϵ_2, then, generally, ϵ_1 should be preferred to ϵ_2 by prospective students, by faculty in search of job positions, by potential employers of alumni, and by other observers and stakeholders—and the strong phrase, *"best schools"*, is explicitly used in their description. An order, incurred by a ranking, projects schools into a unidimensional, mathematically transitive space, in which there can be no ambiguity, circularity, or niches. Is this unidimensional, transitive, space the best domain to project business schools?

Consider, for the sake of argument, the imaginary land of Simplicia. In Simplicia, there are four business schools. Two business schools, *LE* and *LA,* are found at the island of Laputa—to borrow from Jonathan Swift—and are deeply concerned with theoretical development and (quite literally) blue-sky research. There is no concern with practicalities, hardly any focus at teaching, and case studies and examples are explicitly prohibited. There is one striking difference between schools *LE* and *LA*, though: *LE* is an expensive school, while *LA* is an affordable school. There is absolutely no other difference between the schools: all professors, instalations and every other imaginable characteristic are exactly the same. The other two business schools, *RA* and *RE*, are found in the land of Recordia—a land in which everything must be recorded. These schools sharply focus on example after example, and never attempt to find generalities, similarities, analogies, or models that join characteristics or general ideas from even two individual examples from their vast libraries. In Recordia, philosophy, mathematics, statistics, and metaphors have been banned. At the start of the school year, a lottery selects one thousand examples to be taught that year, with

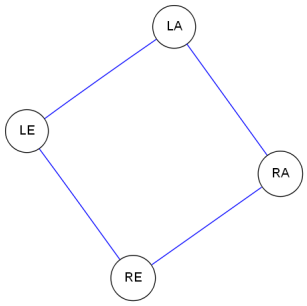


Figure 1: *Each of the four schools of Simplicia is related to two others by one—and only one—of their dimensions.*

no logical sequence between them. As in Laputa, the only difference between the schools is that *RE* is expensive and *RA* is affordable.

What is the structure that relates the schools of Simplicia? There are at least two equally plausible structures: a *grid*, or a *ring*:

- A *grid* structure has two axis x, y in which entities differ—rather like price versus quality, or height versus weight. In this case, the dimensions are (obviously) affordability-exclusivity and a fundamentalist focus on examples-theoretical constructs.

- The *ring* structure also suggests itself: note that, if one starts at any school and moves in either the clockwise or counterclockwise direction, one will rapidly find oneself at the beginning of the journey—rather like fixing a longitude and traveling through different lattitudes will bring one back to the starting point.

The *natural* structure for the schools of Simplicia is either a grid or a ring. One can, of course, *project* these schools into an order, creating a ranking $\{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$. But that ranking will not be its natural structure, as necessarily there will be schools that are ranked next to each other while differing in all dimensions. The rank can respect school similarity (following the ring), or it can prioritize one dimension over another:

- If the rank follows the ring (clockwise or counterclockwise), the first school ϵ_1 will share *one* crucial dimension with ϵ_4 (just as it will with ϵ_2)—but ϵ_1 will share no dimensions with ϵ_3, the third-ranked 'opposite' school. Most importantly, this happens regardless of how the order is construed. In other words, the first school will be significantly more similar to the last school than to the penultimate school which, inconsistently, will be one position closer to the first in the ranking.

- If the rank prioritizes one dimension over another (i.e., following the grid structure), schools ϵ_2 and ϵ_3 will not share dimensions but will be next to each other in the ranking. Students that strongly prefer school ϵ_2 but are accepted only by ϵ_3 and ϵ_4 face a hard prospect, as ϵ_3 will not share any dimension with their preference, while ϵ_4 will share one such dimension. Should students go to ϵ_4 to satisfy one of their preferences? In this case, they would risk the prejudice of the lowest ranked school in the whole of Simplicia. Note that this occurs no matter which dimension is prioritized in the ranking's construction.

Our obvious proposition is this: *Projection into a unidimentional domain loses precious information—and similarity between schools can vanish.* Schools can be close to each other in the ranking, but far from each other in their nature. On the other hand, schools can be far from each other in the ranking, and close to each other in their profile. Let us denote this phenomenon as a *rank anomaly*. In mathematical terms, there can be no topological sort of the schools.

This chapter has two objectives. First, we would like to introduce to the Decision Science community a new research tool that may be widely applied to analyze social, organizational, and economic data. The second objective is to show the power of this method through the analysis of the 2008 data of Business Week's MBA program rankings. The results obtained demonstrate rank anomalies in the published rankings—providing an additional perspective for the critical literature of such rankings.

Before engaging in the study of rankings, let us turn our attention to the more general problem: *the imposition of structure by statistical methods*.

2.1.2 *The imposition of structure*

Structures are imposed by most analytical methods. Clustering methods will always find disjoint sets in data. Ranking (or order-based) methods project entities into a domain that must be isomorphic to either $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{R}$. Decision tree methods will create branchpoints to classify the data, and so forth.

Nature, on the other hand, is indifferent to our methods. Nature presents us with a bewildering array of different forms and structures—as do societies, firms, and other complex systems. One may wonder whether Darwin would have found a ranking of life instead of a *tree of life*, had he applied χ^2 to living creatures; thus never leading to the suggestion of a common predecessor in the past and exploitation of niches in the future. One may also suppose Watson and Crick would find it rather difficult to find the structure of DNA if restricted to decision-tree methods. Humans find structures by studying data and carefully comparing and contrasting this information to previously experienced structures [Linhares e Freitas 2010]. Our analytical methods, however, *impose* structures to data. This imposition can be harmful in a number of ways:

- It may suggest hypotheses which are not warranted. A ranking of living beings, *scala naturae* (or "the great chain of being"), was the unquestioned christian doctrine until Carl Linneaus proposed the tree alternative; this "great chain of being" hypothesis—which goes upwards from rocks, plants, animals, man, spirit, angels and god—suggested a hierarchy of beings that proceeds towards "greatness"; while the *tree of life hypothesis* suggests a common ancestor, speciation, and the exploitation of niches.

- Moreover, the imposition of structures may blind us to important relations hidden in the data. Prisoners generally self-organize into (ethnic) groups. Clustering is able to capture the increased intra-group interaction that dimensionality-reducing methods (such as χ^2 or the use of z-values) cannot. Ranking prisoners in order of "violence propensity", or guards in terms of "abuse of power

propensity", will create so called rank anomalies and will most likely neither reflect nor predict violence between individuals in any meaningful way. One needs to know how individuals interact, not how they rank in a single dimension.

- Finally, the structures may simply be inconsistent with the data, as in the case of rank anomalies in Simplicia. Do these anomalies appear in publicized rankings? If so, can we detect them? (As we will see below, in the "Top-30" Business Schools of America—according to Business Week—the answer to both questions is a resounding *yes*.)

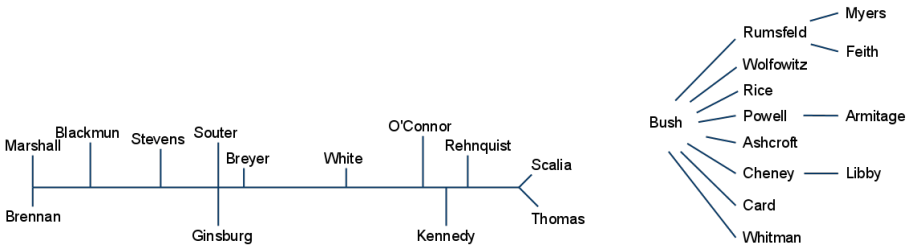
There is, however, no need to pressupose a form when analyzing data. A recent theory from cognitive scientists Charles Kemp and Joshua Tenenbaum [Kemp e Tenenbaum 2008] enables the simulation of the *cognitive discovery of form*. While Kemp and Tenenbaum are mostly interested in their work *as a cognitive theory*, in this study, we present their approach *as a new analytical method*, and we apply it to school rankings. We hereafter refer to the model we will use to compute structures as *KT-Structures* (not to be confused with Hermitian structures, e.g., [Basel 2004]). In the next section we summarize Kemp and Tenenbaum's mathematical model.

2.2 ON KT-STRUCTURES

Kemp and Tenembaum [Kemp e Tenenbaum 2008] have developed a model which, through hierarchical Bayesian inference, can explore and discover the underlying form that best adapts to a given dataset:

> Discovering the underlying structure of a set of entities is a fundamental challenge for scientists and children alike. Scientists may attempt to understand relationships between biological species or chemical elements, and children may attempt to understand relationships between category labels or the individuals in their social landscape, but both must solve problems at two distinct levels. The higher-level problem is to discover the form of the underlying structure. [...] the lower-level problem is to identify the instance of this form that best explains the available data. (p. 10687)

Kemp and Tenenbaum provided, as a psychological theory, a method for the unsupervised learning of form. This method can, moreover, be used as a data analysis method. Statistical methods currently focus on the aplication and optimization of a given structure to data, while presupposing a specific underlying form, such as groupings (e.g., clustering), trees (e.g., hierarchical clustering, minimum spanning tree), or spacial representations (e.g., multidimensional scaling, self-organizing maps, PCA). That is, if one applies a clustering method to a set of data, one stands on the assumption that clusters provide a suitable form to analyse and understand the data. If one applies decision trees, one is projecting that the data can be best understood as having no cycles. Similarly, a school ranking projects schools into a unidimensional, mathematically transitive, lens. The question we pose, therefore, is whether the topological sort brought by the ranking is the best form to analyze the data provided and lead prospective students to optimal decisions concerning school choice.



(a) (b)

Figure 2: *(a) Spectrum extracted from justices' votes; (b) Hierarchy of the Bush cabinet*

There are two important ideas involved in their model: i) the use of a hierarchical Bayesian method to analyse data; and ii) the use of graph grammars and graph redescriptions. Through simple operations based on graph-grammars and Bayesian inference, their algorithm is capable of exploring the space of possible forms (and the particular instances) to find a best-fit representation of the provided dataset. Let us look at each of these ideas in the following subsections.

We take the liberty of reproducing two examples from the dataset used in [Kemp e Tenenbaum 2008]: i) The discovery of a chain that arranges U.S. supreme court justices from liberal (Marshall & Brennan) to conservative (Thomas & Scalia) in fig. 2a. This structure is extracted from a simple record of each justice's votes over approximately 1,500 cases. ii) A hierarchy describing interactions between members of the Bush administration (fig. 2b) based on the Google searches of the form "X told Y", varying X and Y over the different members of the administration.

2.2.1 *Hierarchical Bayesian model: forms, structures, and data*

Bayesian models have been applied with promising results throughout a wide range of problems, stemming from the detection of frauds in financial statements [Zhoua e Kapoor], to real-time telemarketing [Ahna e Ezawa 1997], and leak detection in pipelines [Zhoua et al.]. Examples of *hierarchical* Bayesian models can be found in areas as diverse as marketing [Abe 2009], political science [Lock e Gelman 2010], medicine [Lönnstedt e Britton 2005], language acquisition [Perfors et al. 2011], cognitive science [Kemp 2007, Perfors e Tenenbaum 2009] and artificial intelligence [Damoulas e Girolami 2009].

In Kemp and Tenenbaum's model, starting from dataset D, the algorithm attempts to find a form F and the structure S that best captures the relationships within the dataset. Input data may be expressed either as features and elements or as triangular relational matrices, containing data about the relations between items to be explored. This cognitive aspect of discovery occurs on different levels of abstraction concurrently. The possibilities are generated via graph-grammar splits (see below) and the system seeks then to maximize the posterior probability:

$$P(S, F|D) \propto P(D|S)P(S|F)P(F)$$

That is, what is the probability of a form F and a structure S, given a dataset D? In the hierarchical model, this probability is proportional to the product of i) the probability of dataset D given structure S, ii) the probability of a structure S given the form F, and iii) the probability of form F. Let us look at each of the three probabilities of the right hand side in turn.

As pointed out above, there are many possible forms: trees, hierarchies, rings, clusters, etc, and initially $P(F)$ is given by a uniform distribution over all possible forms of the model.

$P(S|F)$ is given by the number of structures compatible with a given form:

$$P(S|F) \propto \begin{cases} \theta^S & \text{if Sis compatible with F} \\ 0 & \text{otherwise,} \end{cases}$$

i.e., if S is incompatible with F, then $P(S|F) = 0$. Otherwise it can be computed given additional info, such as the Stirling number of the second kind and the number of k-cluster structures for a given form, as described in [Kemp e Tenenbaum 2008]. Graphs with numerous clusters are penalized through parameter θ.

We refer the interested reader to Kemp and Tenenbaum for the mathematical definitions of $P(S|F)$ and for $P(D|S)$, the probability of structure S given prior dataset D.

Their second important idea is the use of *graph grammars* to generate the forms and structures reflecting the data.

2.2.2 *Graphs and graph grammars*

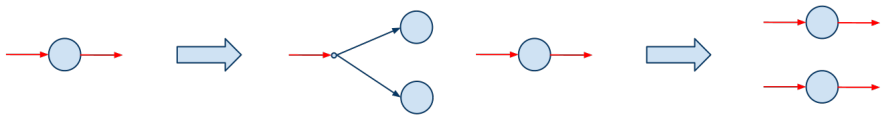
Graph theory provides a mathematical framework to understand objects and their relations. One of the most interesting ideas brought forth in [Kemp e Tenenbaum 2008] was to define the hypothesis space through graph operations and, through Bayesian inference, make use of these simple operations to generate a given structure as a possible fit to the data presented. These generating methods are graph grammars. A particular form (tree, ring, partition, etc.) can be generated by repeatedly applying simple operations in a graph, and by inferring which operation is best suitable to a structure S and dataset D at a given point, one can infer the underlying form F.

Graphs are powerful because they can represent any type of form and provide any kind of structure onto which the data may be projected. Consider, for example, graph grammars for trees and chains:

i) Trees: Suppose all objects are put in a single cluster, C_1. A graph grammar for trees will select a subset of these objects to move to a new cluster C_2, and create a branch point $B_{\{C_1,C_2\}}$ that leads to C_1 and C_2, see Fig. 3a.

ii) Chains: Suppose, once again, that all objects are put in a single initial cluster C_1. A graph grammar for chains will create a cluster C_2, and split C_2 from C_1. No branchpoint is created in this form, of course, see Fig. 3b.

An interesting point concerning these generating processes is that the same operation may also be used on subsequent clusters, i.e., not only on a starting cluster with all objects contained therein. This enables the



(a) (b)

Figure 3: *(a) Tree-generative Graph Operation. (b) Cluster-generative Graph Operation*

'splitting' process to continue until a final structure is reached. Also, items may eventually be moved between clusters, if the model finds that this would create a structure that is more adequate to the data.

We will close our description over the principal points of the original authors' algorithm by focusing on the process of generating the split space via Stirling Numbers.

2.2.3 *Stirling Numbers*

As we have mentioned, the KT algorithm chooses, at each iteration, one best split operation to apply over the graph. This requires generating the possibility space of all potential split locations. In order to generate the hypothesis space of possible splits, the Stirling numbers of the second kind are used. A Stirling number is calculated as follows:

$$S(m,n) = \frac{1}{n!}\sum (-1)^k \binom{n}{n-k} (n-k)^m$$

This enumerates the ways of distributing m distinct objects into n identical containers with no container left empty.

Given this brief summary of KT-Structures, we may now proceed to apply the model to the Business Week Rankings of US-based MBA Programs.

2.3 THE BUSINESS WEEK 2008 RANKING

Our first experiment computes KT-Structures of the Business Week 2008 ranking. The purpose is to compare and contrast the rankings widely used with the KT-Structure. All the 21 possible forms provided in the method were computed, and we concentrated attention to those that suggested rank anomalies. Of these, the tree and hierarchy structures readily presented potential rank anomalies, and we concentrate focus on them here.

2.3.1 *Materials and Methods*

The data used was that provided in the Business Week 2008 MBA program ranking. We ignored the values in the fields "2006 Rank" and "2008 Rank", as we did not want to skew the results towards those generated by Business Week—had we included such dimensions, the correlation between ranking distance and KT-structure distance would become artificially inflated. In the remaining dataset, there are 12 variables, and nothing beyond those values is assumed to have any importance (e.g., to Business Week's methodology, cities "do not exist":

a student living in Bloomington IN has the exact same experience of a student living in New York City—the data is simply oblivious to this information). These dimensions are: graduate poll, corporate poll, intellectual capital, tuition and fees, pre-MBA pay, post-MBA pay, selectivity, job offers, general management, analysis, teaching, and careers. These last four dimensions ranged from A+ to C, and we changed these results to numerical values (A+, A, B, and C were translated to 1, 2, 3, and 4, respectively).

Notice an important aspect here. The model does not know that an A-grade is better than a C-grade, or that a higher post-MBA pay value is better than a lower one. The method does not have any information concerning the meaning of all these variables. But there are strong relations between the data: ranks are provided by orders; tuition, fees and pay are determined by the market, letter grades are obtained through Business Week's polls, etc. The model is able to compute the structures based only on the underlying data, and does not need to understand the *meaning* imbued in each dimension—the problem of meaning is far from solved in cognitive science [Linhares 2000, Linhares e Brum 2007].

While we did not have access to the original polling data utilized in building the original ranking, the relevance of the questions and issues posed here remain. Because of this, we avoid direct criticism of the methodology and steer away from a direct comparison of our results to those of the publication. For a deeper description of how Business Week built their ranking, please see [Gloeckler et al. 2008]

2.3.2 *Numerical experiments*

The most interesting form is a hierarchy; presented in figure 4. At a macro level, this form has some semblance with the original ranking (Figure 5). The KT-Structure distance between two schools (i, j) is measured by counting the number of edges from the origin school's cluster to the destination school. The rank distance, on the other hand, is simply obtained by $|R_i - R_j|$, where R_k is the position of school k in the rank. Note that the domains are quite distinct, as distances in the KT-Structure tend to be smaller, yet, there is positive correlation between the rank and the hierarchy ($r = .65$—and covariance is 8.34). This shows that—at a large scale—there is some agreement between the rank and the KT-Structure.

The striking characteristic of trees and hierarchies—as contrasted to rankings—is the possibility of branchpoints. If the reader will allow a metaphor: given the data, schools are better viewed as cities organized alongside a river rather than as elevator stops on a skyscraper. The glacier melts at the left side of fig. 4, with the cluster comprising Harvard, Stanford and Wharton. As one moves downstream, the differentiating variable (at this point) is post-MBA pay: the first cluster with three schools are the only ones over $120k, the second cluster with values ranging from $105k (Chicago) to $116k (MIT). The cluster comprising Michigan ($105k) and Duke ($100k) is followed by one comprising Cornell ($96k) and NYU ($95k). There are three schools downstream with post-MBA pay of $100k or more (UCLA, Virginia, and CMU), but at this stage many other variables become increasingly relevant, and the tree branches.

A small stream leads to Yale, Maryland, and Olin. A combination of *relatively* undesirable data explains this cluster: the schools share C's

in "general management" and "analysis" (and B's in "careers"), they are low-ranked in the corporate poll (positions 33, 41, and 42), and they are relatively expensive. These traits lead us to interesting distortions between this tree and the rankings.

As a demonstration of the explaining power of the KT-Structure, consider the following example. Suppose a student preferred the University of Washington's Foster School (no. 27), but was rejected there and accepted by two schools: Yale (no. 24) and Georgia Tech (no. 29). The student's choice seems easy, as Yale is no doubt better ranked.

The KT-Structure, however, tells a different story, placing Yale far from the student's preferred Foster. Here is why. If the student chooses Georgia Tech, tuition costs drop slightly from Foster's $64,902 to Georgia Tech's $64,152—while Yale will charge $93,098. Again, if the student chooses Georgia Tech, Foster's "B" in "general management" is also found in Georgia Tech—while Yale holds a "C". Finally, if the student chooses Georgia Tech, Foster's "B" in "analysis" is reflected by an "A" in Georgia Tech's grade—while Yale holds a "C". Georgia Tech, at the 28^{th} position in the corporate poll, is much closer to the preferred Foster's 26^{th} position than Yale (33^{th} position).

Of course, by choosing Yale over Georgia Tech, there are also significant gains in other dimensions, but these are dimensions *which the student did not prioritize by choosing Foster*. The ranking keeps moving further away from the student's preferred school characteristics. The preferred school held the 30^{th} position in the graduate poll; Georgia Tech holds the 31st—but Yale is at the 19^{th} position. In "intelectual capital", the preferred school held the 29^{th} position, while Georgia Tech holds the 26^{th} position—but Yale is number 10. In school selectivity (perhaps a minor concern to our already accepted student), the preferred school accepts 30% of applicants, Georgia Tech accepts 29%—while Yale is much more selective, at 14%.

Of the 12 dimensions considered in building the ranking, Yale differs significantly in 7 dimensions from both the student's preferred school and from Georgia Tech (and also from Brigham Young). This is why the KT-Structure places schools like Maryland (26) close to Washington University's Olin (28), while both are far from the University of Washington's Foster (27) and Georgia Tech (29) (which also resemble each other in many dimensions). Instead of differentiating them, the

School	2008	Grad. poll	Corp. poll	Intel. capital	Tuition & fees	Post-MBA pay	Selectivity (%)	Gen. mgmt. skills	Analysis
Brigham Young	22	27	15	41	$ 37,010.00	$ 90,000	56	A	A
U. of Wash. (Foster)	27	30	26	29	$ 64,902.00	$ 85,000	30	B	B
Georgia Tech.	29	31	28	26	$ 64,152.00	$ 95,000	29	B	A
(group range)		27-31	15-28	26-41	37-64 K	85-95 K	29-56 %	A-B	A-B
Yale	24	19	33	10	$ 93,098.00	$ 97,000	14	C	C
Maryland (Smith)	26	28	42	3	$ 82,435.00	$ 91,000	28	C	C
Wash. U. (Olin)	28	24	41	16	$ 82,672.00	$ 90,000	34	C	C
(group range)		19-28	33-42	3-16	82-93 K	90-97 K	14-34 %	C	C

Table 1: *Rank anomalies. Though schools ranked {22, 24, 26, 27, 28, and 29} seem close in the ranking, they are clearly separable into different clusters.*

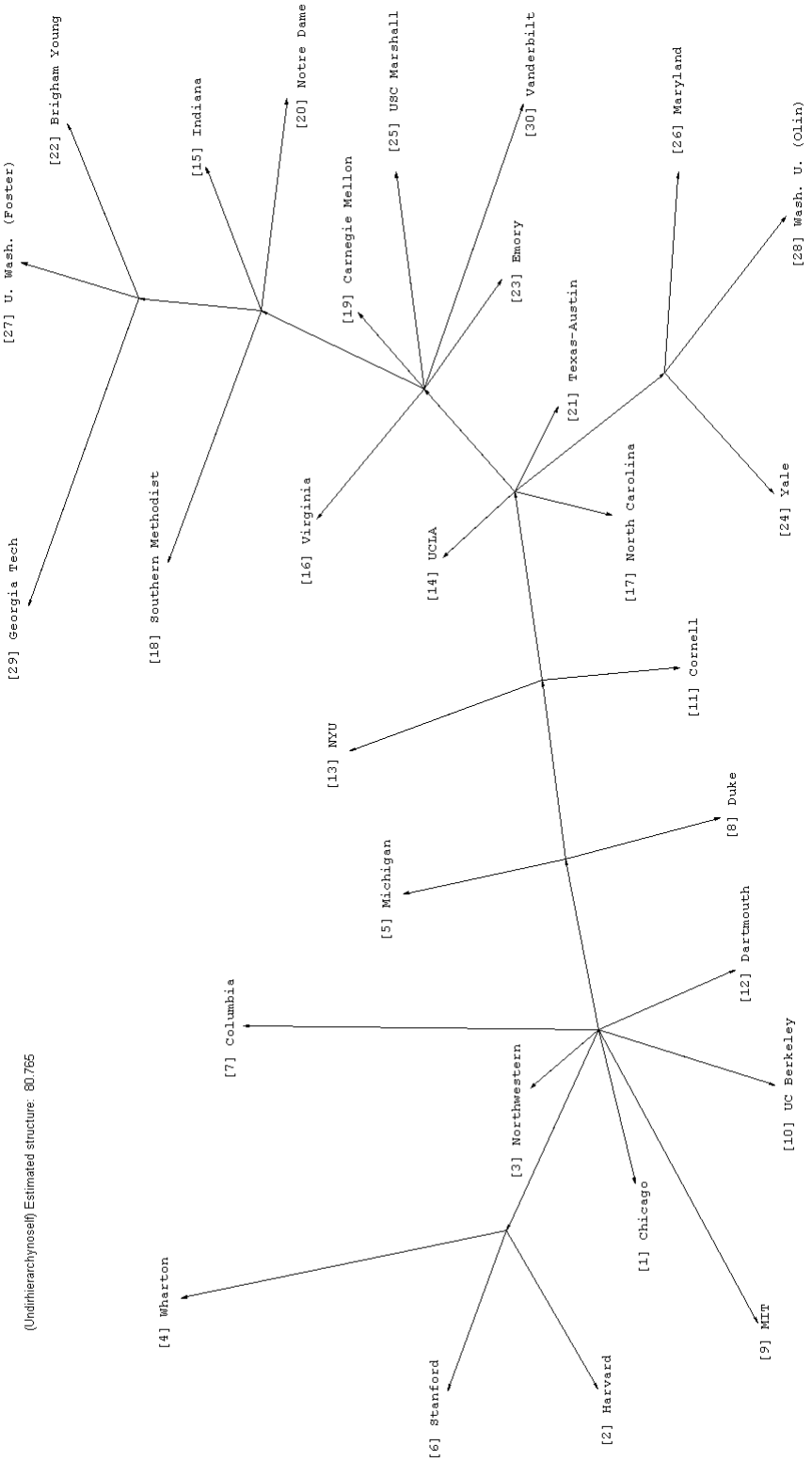


Figure 4: *The generated KT-Structure: an undirected hierarchy with no self-links.*

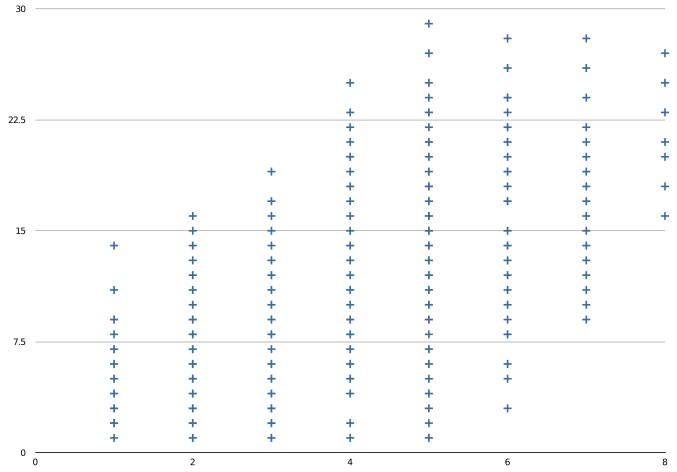


Figure 5: *The KT-Structure distance between schools plotted against their ranked distance.*

rank alternates between these two different groups, obliterating their differences along the way. These are serious rank anomalies.

In sum: if the ontology of the world of MBA programs consisted solely of the 12 dimensions included in the rankings, which is questionable, and if the collected data were an absolutely perfect reflection of reality, which is also questionable, and even if the aforementioned criticisms of rankings brought forth in the literature were all invalid, which also happens to be questionable, this much is true: a student with a strong preference for the no. 27^{th} school would find that school no. 29 seems a better match than school no. 24. If, in an ideal world, popular publications provided other visualizations instead of rankings, there would be no cognitive dissonance in choosing between a school that better reflects one's true preferences versus the "better ranked" one. (We in fact hypothesize that students facing these choices would choose schools according to the KT-Structure more often than according to rank, though we have no way to test this at this point). This is the type of meaningful information which the KT-Structure brings to light and to which a simple rank ordering remains oblivious.

The KT-Structure enables a comparison to other schools in the same cluster and, moreover, highlights the differences between the clusters upstream. A school can move faster if it knows exactly where it is located in this multidimensional space, and sensitivity analysis can be conducted through careful variations of parameters. There are no sudden jumps here; as there are many multidimensional curves on the road to Harvard[1].

2.4 SUMMARY

We introduce to the decision science community, Kemp and Tenenbaum's model for finding structure in data. Instead of presenting it under the perspective of a psychological theory, our goal here is to describe it as a new methodology for research. In our experiments, we

1 Harvard is not used here to express endorsement or any other value judgment. Because of its great wealth, history, faculty, and alumni, it can be argued that Harvard University—and HBS—has become the archetype of the "world-class" University.

have applied the method to the data used to construct school rankings by Business Week (2008). We claim the method provides insights into the multidimensional space in which schools compete, and that the resulting KT-Structures better reflect the multi-faceted reality of business schools and are better representations than the widely disseminated rankings.

Using the very same features used by Business Week, the KT-Structures bring to light anomalies in which schools may be next to each other in the rankings while bearing few resemblances in their numerous dimensions. Conversely, schools can be far in the rankings, but have a large set of similar features. We therefore question the validity of school rankings: A rank is not necessarily the most adequate form to represent (or understand) entities with no dominance relation. Statistical and data mining methods often pressupose a hidden structure, such as a cluster, a tree, or a ranking. The MBA program rankings, however, impose a representational form that is unfit for the type of information they hope to convey. This has sweeping implications to school strategy, positioning, and, because of the wide impact of published rankings, for prospective students and all stakeholders. One can only idealize a world in which the structures that best reflect the data are widely disseminated for public consumption. We hope readers will receive this introduction to KT-Structures with the recognition that this is an innovative, promising approach that deserves to be admitted into the decision scientist's toolbox of research methods.

Part III

SPARSE DISTRIBUTED MEMORY

3

THE EMERGENCE OF MILLER'S MAGIC NUMBER ON A SPARSE DISTRIBUTED MEMORY

In order to more fully understand this reality,
we must take into account other dimensions
of a broader reality.

— John Archibald Wheeler (1911 - 2008)

3.1 INTRODUCTION

Human short-term memory is severely limited. While the existence of such limits is undisputed, there is ample debate concerning their nature. Miller [Miller 1955] described the ability to increase storage capacity by grouping items, or "chunking". He argued that the span of attention could comprehend somewhere around seven information items. Chunk structure is recursive; as chunks may contain other chunks as items: Paragraphs built out of phrases built out of words built out of letters built out of strokes. This mechanism is used to explain the cognitive capacity to store a seemingly endless flux of incoming, pre-registered, information, while remaining unable to absorb and process new (non-registered) information in highly parallel fashion.

Miller's 'magic number seven' has been subject of much debate over the decades. Some cognitive scientists have modeled such limits by simply using (computer-science) "pointers", or "slots" [Gobet e Simon 2000, Gobet et al. 2001]—see [Linhares e Brum 2007, Linhares e Freitas 2010] for debate. However, such approaches do not seem plausible given the massively parallel nature of the brain, and we believe memory limits are an emergent property of the neural architecture of the human brain. As Hofstadter put it a quarter of a century ago [Hofstadter 1985, p. 642]: the "problem with this [slot] approach is that it takes something that clearly is a very complex consequence of underlying mechanisms and simply plugs it in a complex structure, bypassing the question of what those underlying mechanisms might be".

Our objective in this chapter is to study these memory limits as emergent effects of underlying mechanisms. We postulate two mechanisms, discussed in previous literature. The first is a mathematical model of human memory brought forth by Kanerva [Kanerva 1988], called Sparse Distributed Memory (SDM). We also presuppose, following [Kanerva 1993], an underlying mechanism of chunking through averaging. It is not within the scope of this study to argue the validity of SDM as a cognitive model. For incursions on this broader topic, we direct the reader to [Kanerva 1993, Stewart e Eliasmith 2009, Gayler 2003] who discuss the plausibility of this Vector Symbolic Architecture family of models (in which SDM is contained).

This work, while similar in its mathematical foundations, is different from previous capacity analyses: In [Kanerva 1988], the memory capacity analysis relates to long-term memory mechanisms, while we study the short-term memory limits (of this same model). Our work also differs from that of Plate, in that, regardless of the number of items presented, the memory will only store (and subsequently retrieve) a

psychologically plausible number of items. The difference becomes salient in Plate's own description [Plate 2003, p. 139]: "As more items and bindings are stored in a single HRR the noise on extracted items increases. If too many associations are stored, the quality will be so low that the extracted items will be easily confused with similar items or, in extreme cases, completely unrecognizable".

A number of theoretical observations are drawn from our computations: i) a range of plausible numbers for the dimensions of the memory, ii) a minimization of a current controversy between different 'magic number' estimates, and iii) potential empirical tests of the averaging assumption. We should start with a brief description of our postulates: i) the SDM, and ii) chunking through averaging.

3.1.1 *Sparse Distributed Memory*

The Sparse Distributed Memory (SDM), developed in [Kanerva 1988], defines a memory model in which data is stored in distributed fashion, in a vast, sparsely populated binary address space. In this model, (a number of) neurons act as *address decoders*. Consider the space $\{0,1\}^N$: SDM's address space is defined allowing 2^N possible locations, where N defines both the word length and the number of dimensions of the space (e.g., the memory holds binary vectors of length N). In SDM, the data is the same as the medium in which it is stored (i.e. the stored items are N-bit vectors in N-dimensional binary addresses).

SDM uses Hamming distance as a metric between any two N-bit vectors (hereafter memory items, items, elements, or bitstrings—according to context). Neurons, or *hard locations* (see below), in Kanerva's model, hold random bitstrings with equal probability of 0's and 1's—Kanerva [Kanerva 1994, Kanerva 2009] has been exploring a variation of this model with a very large number of dimensions (around 10000). (With the purpose of encoding concepts at many levels, the Binary Spatter Code—or BSC—, also shares many properties with SDM.) With the Hamming distance as a metric, one can readily see that the average distance between any two points in the space is given by the binomial distribution, and approximated by a normal curve with mean at $N/2$ with standard deviation $\sqrt{N}/2$. Given the Hamming distance, and large N, most of the space lies close to the mean. A low Hamming distance between any two items means that these memory items are associated. A distance that is close to the mean $N/2$ means that the memory items are orthogonal to each other. This reflects two facts about the organization of human memory: orthogonality of random concepts, and close paths between random concepts.

Orthogonality of random concepts: the vast majority of concepts is orthogonal to all others. Consider a non-scientific survey during a cognitive science seminar, where students asked to mention ideas unrelated to the course brought up terms like *birthdays, boots, dinosaurs, fever, executive order, x-rays*, and so on. Not only are the items unrelated to cognitive science, the topic of the seminar, but they are also unrelated to each other.

Close paths between concepts: The organization of concepts seems to present a 'small world' topology–for an empirical approach on words, for instance, see [Cancho e Solé 2001]. For any two memory items, one can readily find a stream of thought relating two such items ('Darwin gave *dinosaurs* the *boot*'; 'she ran a *fever* on her *birthday*'; 'isn't it time for

the Supreme Court to *x-ray* that *executive order*?' ... and so forth). Robert French presents an intriguing example in which one suddenly creates a representation linking the otherwise unrelated concepts of 'coffee cups' and 'old elephants' [French 1997]. In sparse distributed memory, any two bitstrings with Hamming distance around $N/4$ would be extremely close, given the aforementioned distribution. And $N/4$ is the expected distance of an average point between two random bitstrings.

Of course, for large N (such as $N \geqslant 100$), it is impossible to store all (or even most) of the space—the universe is estimated to carry a storage capacity of 10^{90} bits (10^{120} bits if one considers quantum gravity) [Lloyd 2002]. It is here that Kanerva's insights concerning sparseness and distributed storage and retrieval come into play: 2^{20}—or a number around one million–physical memory locations, called hard locations, could enable the representation of a large number of different bitstrings. Items of a large space with, say, 2^{1000} locations would be stored in a mere 2^{20} hard locations—the memory is indeed sparse.

In this model, every single item is stored in several hard locations, and can, likewise, be retrieved in distributed fashion. Storage occurs by distributing the item in every hard location within a certain threshold 'radius' given by the Hamming distance between the item's address and the associated hard locations. Different threshold values for different numbers of dimensions are used (in his examples, Kanerva used 100, 1000 and 10000 dimensions). For $N = 1000$, the distance from a random point of the space to its nearest (out of the one million) hard locations will be approximately 424 bits [Kanerva 1988, p.56]. In this scenario, a threshold radius of 451 bits will define an *access sphere* containing around 1000 hard locations. In other words, from any point of the space, approximately 1000 hard locations lie within a 451-bit distance. All of these accessible hard locations will be used in storing and retrieving items from memory. We therefore define the function $A : \{0,1\}^N \times \{1,2,\ldots,N\} \mapsto 2^{\{0,1\}^N}$ and a hard location $\psi_x \in A(x,R)$ iff $\psi_x \in \{0,1\}^N \wedge H(x,\psi_x) \leqslant R$, where A defines an access radius around x of size R (451 if $N = 1000$; H is the Hamming distance).

A brief example of a storage and retrieval procedure in SDM is in order: to store an item x at a given (virtual) location ζ (in sparse memory) one must activate every hard location within the access sphere of x and store the datum in each one. Hard locations carry N adders, one for each dimension. To store a bitstring x at a hard location ψ, one must iterate through the adders of ψ: If the i^{th} bit of x is 1, increment the i^{th} adder of ψ, if it is 0, decrement it. Repeating this for all hard locations in x's access sphere will distribute the information in x throughout the hard locations.

Retrieval of data in SDM is also massively collective and distributed: to peek the contents of each hard location, one computes its related bit vector from its adders, assigning the i^{th} bit of ψ as a 1 or 0 if the i-th adder is positive or negative, respectively (a coin is flipped if it is 0). Notice, however, that this information in itself is meaningless and may not correspond to any one specific datum previously registered. To read from a location x in the $\{0,1\}^N$ address space, one must activate the hard locations in the access sphere of x and gather each related bit vector. The stored datum will be the majority rule decision of all activated hard locations' related bit vectors. If, for the i^{th} bit, the majority of all bit vectors is 1, the final read datum's i^{th} bit is set to 1, otherwise to 0. Thus, "SDM is distributed in that many hard locations participate in storing

and retrieving each datum, and one hard location can be involved in the storage and retrieval of many data" [Anwar e Franklin 2003, p. 342].

All hard locations within an access radius collectively point to an address. Note also that this process is iterative. The address obtained may not have information stored on it, but it provides a new access radius to (possibly) converge to the desired original address. One particularly impressive characteristic of the model is its ability to simulate the ´tip-of-tongue´ phenomenon, in which one is certain about some features of the desired memory item, yet has difficulty in retrieving it (sometimes being unable to do so). If the requested address is far enough from the original item (209 bits if $N = 1000$), iterations of the process will not decrease the distance—and time to convergence goes to infinity.

The model is robust against errors for at least two reasons: i) the contribution of any one hard location, in isolation, is negligible, and ii) the system can readily deal with incomplete information and still converge to a previously registered memory item. The model's sparse nature dictates that any point of the space may be used as a storage address, whether or not it corresponds to a hard location. By using about one million hard locations, the memory's distributed nature can 'virtualize' the large address space. The distributed aspect of the model allows such a virtualization. Kanerva [Kanerva 1988] also discusses the biological plausibility of the model, as the linear threshold function given by the access radius can be readily computed by neurons, and he suggests the interpretation of some particular types of neurons as address decoders. Given these preliminaries concerning the Sparse Distributed Memory, we should now proceed to our second premise: *chunking through averaging*.

3.1.2 *Chunking through averaging*

To chunk items, the majority rule is applied to each bit: given ν bitstrings to be chunked, for each of the N bits, if the majority is 1, the resulting bitstring's chunk bit is set to 1; otherwise it is 0. In case of perfect ties (no majority), a coin is flipped.

We have chosen the term 'chunking' to describe an averaging operation, and 'chunk' to describe the resulting bitstring, because, through this operation, the original components generate a new one to be written to memory. The reader should note, in SDM's family of high-dimensional vector models, called Vector Symbolic Architectures (VSA), the operation that generates composite structures is commonly known as superposition [Stewart e Eliasmith 2009, Gayler 2003, Plate 2003].

Obviously, this new chunked bitstring may be closer, in terms of Hamming distance, to the original elements, than the mean distance $N/2$ between random elements (500 bits if N=1000), given a relatively small ν. The chunk may then be stored in the memory, and it may be used in future chunking operations, allowing, thus, for recursive behavior. Let us denote this averaged bitstring as α. With these preliminaries, we may turn to numerical results.

3.2.1 *Computing the Hamming distance from chunk α to items*

Let $\xi = \{\xi_1, \xi_2, ..., \xi_\nu\}$ be the set of bitstrings to be chunked into a new bitstring, α. The first task is to find out how the Hamming distance is distributed between this averaged α bitstring and the set $\xi = \{\xi_1, \xi_2, ..., \xi_\nu\}$ of bitstrings being chunked. This is, as discussed, accomplished through majority rule at each bit position. Imagine that, for each separate dimension, a supreme court will cast a decision with each judge choosing yes (1) or no (0). If there is an even number of judges, a fair coin will be flipped in the case of a tie. Given that there are $\nu = |\xi|$ votes cast, how many of these votes will fall in the minority side? (Each minority-side vote adds to the Hamming distance between an item ξ_i and the average α.)

Note that the minimum possible number of *minority votes* is one, and that it may occur with either 3 votes cast or two votes and a coin flip. If there are two minority votes, they may stem from either 5 votes or 4 votes and a coin flip, and so forth. We thus have that, for ν votes, the maximum minority number is given by $\lfloor \nu/2 \rfloor$ (and the ambiguities between an odd number of votes versus an even number of votes plus a coin flip are resolved by considering $2\lfloor \nu/2 \rfloor + 1$ total votes). This leads to independent Bernoulli trials, with success factor $p = 1/2$, and the constraint that the minority view differs from the majority bit vote. Let X be a random variable with the number of minority votes. Obviously in this case, $P(1 \leqslant X \leqslant \lfloor \nu/2 \rfloor) = P(X \leqslant \lfloor \nu/2 \rfloor - 1)$, hence we have, for ν items, the following cumulative distribution function of minority votes [Boland 1989]:

$$P(X \leqslant \lfloor \nu/2 \rfloor - 1) = \sum_{i=0}^{\lfloor \nu/2 \rfloor - 1} \binom{2\lfloor \nu/2 \rfloor}{i} p^i (1-p)^{2\lfloor \nu/2 \rfloor - i} =$$

$$= \sum_{i=0}^{\lfloor \nu/2 \rfloor - 1} \frac{(2\lfloor \nu/2 \rfloor)!}{i!(2\lfloor \nu/2 \rfloor - i)!} \frac{1}{2}^{2\lfloor \nu/2 \rfloor} = 4^{-\lfloor \nu/2 \rfloor} \sum_{i=0}^{\lfloor \nu/2 \rfloor - 1} \frac{(2\lfloor \nu/2 \rfloor)!}{i!(2\lfloor \nu/2 \rfloor - i)!}$$

While we can now, given ν votes, compute the distribution of minority votes, the objective is not to understand the behavior of these minority bits *in isolation*, i.e., per dimension on the chunking process. We want to compute the number of dimensions to (in a psychologically and neurologically plausible way) store and retrieve around M items—Miller's number of retrievable elements—through an averaging operation. Hence we need to compute the following:

(I) Given a number of dimensions N and a set ξ of items, the probability density function of the Hamming distance from α to the chunked elements ξ_i,

(II) A threshold T: a number of dimensions in which, if an element ξ_i's Hamming distance to α is farther from that point, then ξ_i cannot be retrieved,

(III) As $|\xi|$ grows, how many elements remain retrievable?

N	$T_{N,2\sigma}$(M = 4 or 5)	$T_{N,2\sigma}$(intermediary value)	$T_{N,2\sigma}$ (M = 6 or 7)
64	27.42	28.51	29.6
128	50.49	52.62	54.75
192	72.85	76.01	79.16
256	94.83	99.02	103.2
320	116.58	121.8	126.99
384	138.17	144.4	150.61
448	159.62	166.88	174.11
512	180.98	189.25	197.49
576	202.25	211.54	220.8
640	223.45	233.76	244.03
704	244.6	255.92	267.2
768	265.69	278.02	290.32
832	286.74	300.09	313.4
896	307.75	322.11	336.43
960	328.72	344.1	359.43
1024	349.66	366.05	382.4
100	40.52	42.2	43.87
1000	341.82	357.82	373.79
10000	3217.7	3375.16	3532.49

Table 2: *Thresholds $T_{N,2\sigma}$ given plausible success factors and dimension combinations.*

Given bitstrings with dimension N, suppose $\nu = |\xi|$ elements have been chunked, generating a new bitstring α. Let $H_N(\alpha, \xi_i)$ be the Hamming distance from the chunked element α to ξ_i, the i^{th} element of ξ. What is the distance from α to elements in ξ? Here we are led to N Bernoulli trials with success factor $p_{|\xi|}$. Since N is large, $H_N(\alpha, \xi_i)$ for $i = \{1, 2, \ldots, \nu\}$ can be approximated by a Normal distribution, we may use $\mu = Np_{|\xi|}$ and $\sigma = \sqrt{Np_{|\xi|}(1 - p_{|\xi|})}$. To model human short term memory's limitations, we want to compute a cutoff threshold T_N which will guarantee retrieval of around M items averaged in α and "forget" items ξ_i if $H_N(\alpha, \xi_i) > T_N$—where M is the Miller's limiting number. Hence to guarantee retrieval of around 95% (2σ) of M items, we have $T_{N,2\sigma} = Np_M + 2\sqrt{Np_M(1 - p_M)}$, where p_M is the success factor corresponding to M. Note that Cowan [Cowan 2000] has argued for a "magic number" estimate of 4±1 items—and the exact cognitive limit is still a matter of debate. The success factor for 4 (or 5) elements is $p_{\{4,5\}}$=.3125; and for 6 (or 7) elements it is $p_{\{6,7\}}$=.34375. By fixing the success factor at plausible values of M ({4,5}, or an intermediary value between {4,5} and {6,7}, or {6,7}), different threshold values $T_{N,2\sigma}$ are obtained for varying N, as shown in Table 2. In the remainder of this study, we use the intermediary success factor $p_M = .328125 = 21/64$ for our computations; again without loss of generality between different estimates of M.

We thus have a number of plausible thresholds and dimensions. We can now proceed to compute the plausibility range: Despite the

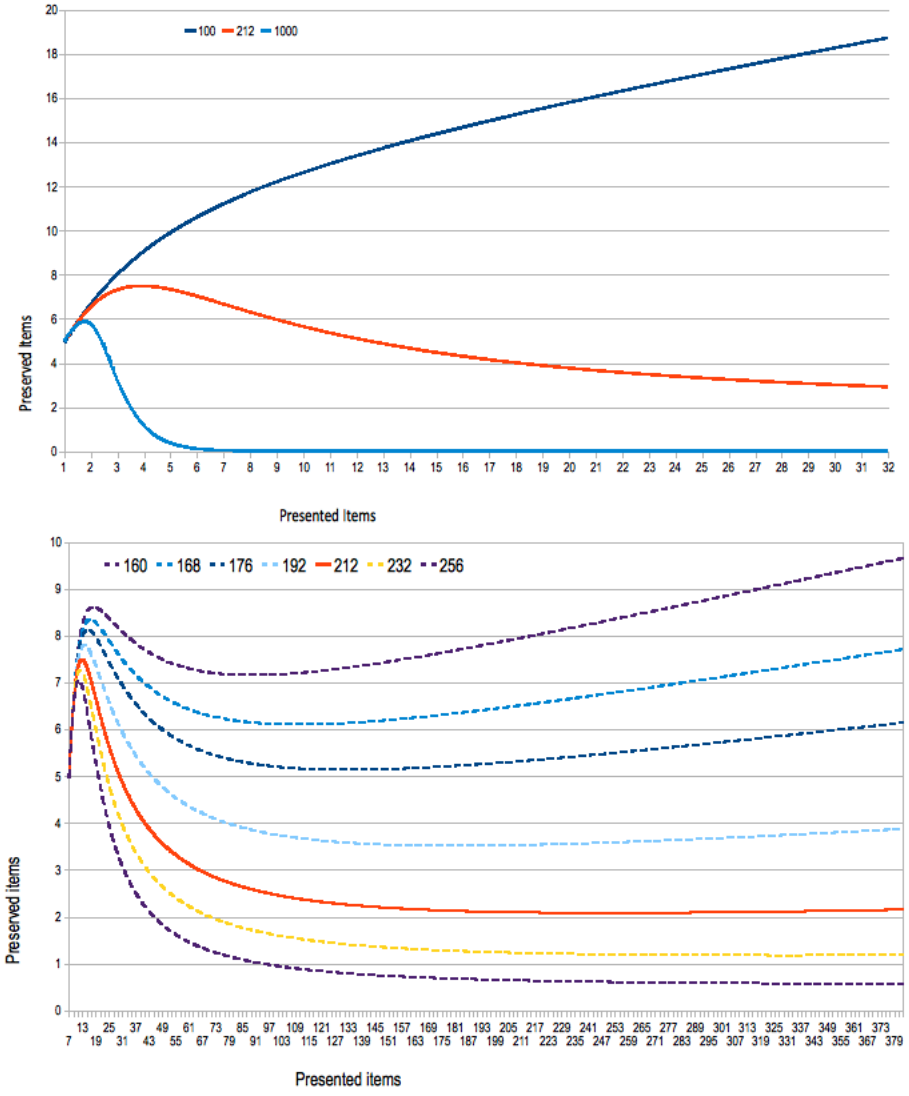


Figure 6: *Behavior at different dimensions and items presented*

implicit suggestion in Table 1 that any number of dimensions might be plausible, how does the behavior of these $(N, T_{N,2\sigma})$ combinations vary as a function of the number of presented elements, $|\xi|$?

3.2.2 *Varying the number of presented items*

Consider the case of information overload, when one is presented with a large set of items. Suppose one were faced with dozens, or hundreds, of distinct items. It is not psychologically plausible that a large number of elements should be retrievable. For an item ξ_i to be impossible to retrieve, the distance between the averaged item α and ξ_i must be higher than the threshold point of the corresponding N. When we have an increasingly larger set of presented items, there will be information loss in the chunking mechanism, but it should still be possible to retrieve some elements within plausible psychological bounds.

Figure 1(a) shows the behavior of three representative sizes of N: 100, 212 and 1000 dimensions. (100 and 1000 were chosen because these

are described in Kanerva's original examples of SDM.) $N = 212$ has shown to be the most plausible number of dimensions, preserving a psychologically plausible number of items after presentations of different set sizes. It is clear that $N = 100$ quickly diverges, retaining a high number of items in a chunk (as the number of presented items grows). Conversely, if $N = 1000$, the number of preserved memory items rapidly drops to zero, and the postulated mechanisms are unable to retrieve any items at all—a psychologically implausible development. Figure 1(b) zooms in to illustrate behavior over a narrower range of N-values and a wider range of presented items. Varying the number of presented items and computing the number of preserved items (for a number of representative dimensions) yields informative results. Based on our premises, experiments show that to appropriately reflect the storage capacity limits exhibited by humans, certain ranges of N must be discarded. With too small a number of dimensions, the model will retrieve too many items in a chunk. With too large a number of dimensions, the model will retrieve at most one or two—perhaps no items at all. This is because of the higher number of standard deviations involved in the dimension sizes: for $N = 100$, the whole space has 20 standard deviations, and $T_{N=100,2\sigma} = 42.2$ is less than 2 standard deviations below the mean—which explains why an ever growing number of items is "retrieved" (e.g., high probability of false positives). For $N = 1000$, the space has over 63 standard deviations, and $T_{N=1000,2\sigma} = 357.82$, is around 8.99 standard deviations below the mean. There is such a minute part of the space below $T_{N=1000,2\sigma}$ that item retrieval is virtually impossible.

With an intermediary success factor p_M between p_4 and p_7 established by the cognitive limits 4 and 7, we have computed the number of dimensions of a SDM as lying in the vicinity of 212 dimensions. Variance is minimized when $N = 212$—and retrieval results hold psychologically plausible ranges even when hundreds of items are presented (i.e., the SDM would be able to retrieve from a chunk no more than nine items and at least one or two, regardless of how many items are presented simultaneously). Finally, given that this work rests upon the chunking through averaging postulate, in the next section we will argue that the postulated mechanism is not only plausible, but also empirically testable.

3.2.3 *The chunking through averaging postulate*

Consider the assumption of chunking through averaging. We propose that it is plausible and worthy of further investigation, for three reasons.

This assumption minimizes the current controversy between Miller's estimations and Cowan's. The disparity between Miller's 7±2 or Cowan's 4±1 observed limits may be a smaller delta than what is argued by Cowan. Our 'chunking-through-averaging' premise may provide a simpler, and perhaps unifying, position to this debate. If chunking 4 items has the same probability as 5 items, and chunking 6 items is equivalent to chunking 7 items, one may find that the 'magic number' constitutes one cumulative probability degree (say, 4-or-5 items) plus or minus one (6-or-7 items).

A mainstream interpretation of the above phenomenon may be that, as with any model, SDM is a simplification; an idealized approximation of a presumed reality. Thus, one may see it as insufficiently complete

to accurately replicate the details of true biological function due to, among other phenomena, inherent noise and spiking neural activity. In this case, one would interpret it as a weakness, or an inaccuracy inherent to the model. An alternative view, however improbable, may be that the model is accurate in this particular aspect, in which case, the assumption minimizes the current controversy between Miller's estimations and Cowan's.

The success factors computed above show that for either 4 or 5 items, we have $p = .3125$, while for 6 or 7 items we have $p = .34375$. If we assume an intermediary value of p—which is reasonable, due to noise or lack of synchronicity in neural processing [Borisyuk et al. 2000]—the controversy vanishes. We chose to base our experiments on the intermediary average value ($p = 21/64 = .328125$), and the results herein may be adapted to other estimates as additional experiments settle the debate.

Moreover, a chunk α tends to be closer to the ξ_i chunked items than these items are between themselves. For example, with $|\xi| = 5$ and $N = 212$, the Hamming distance between a chunk and a random item is drawn from a distribution with $\mu = N21/64$ and $\sigma = (\sqrt{N}\sqrt{903})/64$; in here, from the point of view of the chunked item α, the closest 1% of the space lies at 53 bits, while 99% of the space lies at 84 bits. Contrast this with the distances between any two random, orthogonal, items, which are drawn from $\mu = N/2$ and $\sigma = \sqrt{N}/2$: from the point of view of a random item, the closest 1% of the space lies at 89 bits, while 99% of the space lies at 122. This disparity reflects the principles of *orthogonality between random concepts* and of *close paths between concepts (or small worlds [Cancho e Solé 2001]):* the distance between 2 items from any 5 is large, but the distance to the average of the set is small. Of course, as $|\xi|$ grows, the distance to α also grows (since $\lim_{|\xi|\to\infty} p_{|\xi|} = 1/2$), and items become irretrievable. One thing is clear: with 5 chunked items, the chance of retrieving a false positive is minute.

Finally, the assumption of chunking through averaging is empirically testable. Psychological experiments concerning the difference in ability to retain items could test this postulate. The assumption predicts that $(4, 5)$ items, or more generally that $(2\nu, 2\nu + 1)$ for integer $\nu > 0$ will be registered with equal probability. It also predicts how the probability of $2\nu + 2$ retained items should drop in relation to $2\nu + 1$ if $\nu > 0$. This is counterintuitive and can be measured experimentally. Note, however, two qualifications: first, as chunks are hierarchically organized, these effects may be hard to perceive in experimental settings. One would have to devise an experimental setting with assurances that only chunks from the same level are retrievable–neither combinations of such chunks, nor combinations of their constituting parts. The final qualification is that, as ν grows, the aforementioned probability difference tends to zero. Because of the conjunction of these qualifications, this effect would be hard to perceive on normal human behavior.

3.3 DISCUSSION

Numerous cognitive scientists model the limits of human short-term memory through explicit "pointers" or "slots". In this chapter we have considered the consequences of a short-term memory limit given the mechanisms of i) Kanerva's Sparse Distributed Memory, and ii) chunking through averaging. Given an appropriate choice for the number of

dimensions of the binary space, we are able to model chunks that limit active memory's storage capacity, while allowing the theoretically endless recursive association of pre-registered memory items at different levels of abstraction (i.e., chunks may be chunked with other chunks or items, indiscriminately [Miller 1955, Linhares 2000]). This has been pointed out in [Kanerva 1996], however, in here we use the short-term memory limitations as a bounding factor to compute plausible ranges for N.

Some observations are noteworthy. First, our work provides plausible bounds on the number of dimensions of a SDM—we make no claims concerning Kanerva's recent work (e.g., [Kanerva 2009]). Given our postulates, it seems that 100 dimensions is too low a number, and 1000 dimensions too high. In our computations, assuming $p_M = 21/64$, variance of the number of items retained (as a function of the number of presented items and at least one retrievable item) was minimized at 212 dimensions. This value was chosen as our optimal point of focus for it provided stable, psychologically plausible behavior for a wide range of set sizes. We have concentrated on the SDM and chunking through averaging postulates, yet future research could also look at alternative neural models; for it is certain that the brain does not use explicit slots or pointers when items are chunked. One can reasonably argue: what good can come from replacing one magic number with another? There are two potential benefits: first, by fixing parameter N, we can restrict the design space of SDM simulations and ensure that a psychologically plausible number of items is chunked. Another advantage is theoretical: the number 212 suggests that we should look for neurons that seem to have, or respond majoritarily to, such a number of active inputs in their linear threshold function.

Of course, a single 212 bit vector in SDM does not encode meaningful content at all. The existence of a bitstring can only be meaningful in relation to other bitstrings close to it. Consider, for instance, an A4 sheet of paper, of size 210mm x 297mm (8.3in x 11.7in). A 1200x1200 dots-per-inch printer holds less than 2^{28} potential dots in an entire sheet. While the space of possible black and white printed A4 sheets is a very large set of $2^{139838400}$ possible pages, the vast majority of them, rather like the library of Babel, are composed of utter gibberish. Any single dot needs only 28 bits to be described, and because the dots usually cluster into strokes, chunks can be formed. Moreover, because strokes cluster to form fonts, which cluster to form words, which cluster to form phrases and paragraphs; combinations of large sets of 212 dimensional bitstrings can encode the meaningful content of pages and books—provided those items have been previously chunked in the reader's mind. Without chunks there can be no meaning; this paragraph, translated to Yanomámi (assuming that's possible), would become unreadable to its intended audience and to its authors.

Sparse Distributed Memory holds a number of biological and psychologically plausible characteristics. It is associative, allowing for accurate retrieval given vague or incomplete information (which is relevant given the potential for asynchronous behavior [Borisyuk et al. 2000]); it is readily computable by neurons; it seems suitable for storage and retrieval of low-level sensorimotor information [Assisi et al. 2007], it is a plausible model of the space of human concepts, and it exhibits a phenomenon strikingly similar to the tip-of-the-tongue situation. With

the results presented herein, sparse distributed memory also reflects the natural limits of human short-term memory.

Part IV

CONCLUSIONS AND FUTURE WORK

4 CONCLUSIONS

The exploration of computer models of cognitive behavior has yielded notable advances to our understanding of the human cognitive apparatus, but also a range of tools which may now be explored by other sciences. This work demonstrates the eminent use of some of the tools of cognitive modeling toward management science and administration. The exploration of the boundaries of human rationality, and the application of cognitive models toward decision science proves a vast source of new approaches: analogy-making as a form of knowledge construction and critical analysis; structure acquisition as exploratory analysis and memory models illustrating decision-making capacities.

The independent inquiries put forth in this work remain tied together in their attempt to further the state-of-the-art in decision making and human cognition, a central endeavor and common ground beween the cognitive and management sciences. The following paragraphs hope to emphasize this mutual aim and highlight the results achieved by each individual front.

The initial study investigated the application of the discovery of structural form as a tool for exploratory data analysis. It provided a basis for a new research methodology in decision science and administration. The application of the algorithm on the Business Week rankings gave new insight into the multi-featured space of business schools and the misrepresentations that may arise from a mapping of these features to a ranking structure.

It was shown that schools which share a number of characteristics may be separated when mapped to a ranking and, in likewise fashion, schools that are fundamentally different in several aspects may be placed next to each other. The literature illustrates how the Business Week rankings (and other rankings published periodically) have a deep influence in the decisions of prospective students, of faculty and of the schools' administrations. Specifically, students may base their choice on a published ranking, whereas a different structural representation might better convey the multi-dimensional space of features, and lead to better informed decision-making.

The conclusions drawn by the study were then used to question the validity of rankings, and of imposed structural forms in general.

The second essay, in chapter 3, explored the nature of short-term memory limits from the standpoint of the Sparse Distributed Memory model. The results of the study showed that the model can mimic human chunking behavior, which is limited in its parallel storage capacity. It was further shown that the system retains only a psychologically plausible number of items presented simultaneously while still allowing for theoretically endless recursive associations, i.e., building abstractions upon abstractions, chunks upon chunks.

The application of the discovery of structural form toward data analysis is a work in its infancy and many questions and developments remain to be explored. Furthermore, the exploration of Sparse Distributed Memory as a model of human memory proves fertile ground for explorations into human perception and decision-making. In the

following chapter a (far from complete!) enumeration of avenues for further exploration will be described.

5 AVENUES OF FUTURE EXPLORATION

Theoretical advances of applicable models such as those explored herein can leverage advances in applied social sciences. A framework for the generic discovery of structural form that is reliable, flexible and robust can prove a powerful tool for decision scientists. Furthermore, as described by Kemp [Kemp 2007] the hierarchical Bayesian approach is a useful tool for model-builders to pursue, but as of this publication no generic model building tool has been proposed. The increased automation of hierarchical Bayesian models through the exploration of systemas that can, for instance, infer the number of optimal levels required of them; or explore the space of meta-grammars for hypothesis generation will provide new tools for model-builders in any discipline.

The exploratory work contained in this thesis provides an ample base for further investigation, both into applied decision science and cognitive modeling. We can further illustrate the avenues of future work by enumerating questions that remain, as a brief research agenda:

1. *How may the Kemp-Tenenbaum algorithm be improve in order to better its utility as method of exploratory analysis?*
   A more user-friendly implementation of the structure discovery framework is already under way, and the modularization of the algorithm will permit a number of experiments in improving its precision, robustness and performance. The modularization will also allow the hierarchical Bayesian model to be applied in highly parallel, concurrent fashion (e.g. through the use of modern general programming for graphics processing units - GPGPU), thus allowing the system to scale to much larger data sets and perform deeper explorations of the hypothesis space. Furthermore, this may propitiate the exploration of recursive structures as described in the following question.

2. *What advances can be made to the KT algorithm as a cognitive model?*
   Deeper exploration of the KT algorithm provides a number of points where its process could be made more psychologically plausible. We postulate that a model which incorporates characteristics of FARG models of cognition [Hofstadter e FARG 1995] in the search for best-fit structures can bring the model greater generality of application and psychological plausibility. While the method currently performs a greedy search through the structure space, a *parallel terraced scan* model of hypothesis generation and exploration, which has already been applied in a number of systems [Hofstadter e FARG 1995], should provide enlightening results.
   Moreover, we posit greater plausibiliy *and* flexibility if the KT algorithm could be brought to perform joining operations as well as splitting. As emphasized in chapter 3, humans naturally perform chunking when presented with new data. A chess-player's first view of a game board is not that of one huge group of chess-pieces that are gradually split, but rather the opposite. We join letters into words, words into sentences, sentences into paragraphs, para-

graphs into chapters and chapters into M.Sc. Theses (roughly).
Lastly, the algorithm is currently unable to perform recursive investigations. Imagine a large data set whose natural representation is disjoint clusters, but within each cluster there may be a another internal structure. If prisoners separate themselves into gangs, one would be quick to infer each gang as having its own internal structure of power, be it a chain, hierarchy or something else. We expect that a tool with this capacity will provide interesting results both as a cognitive and as an analytical model; for if there are discrepancies in the internal structures of the highest-level structure, exploring recursive patterns may cause the underlying phenomena that determined the discrepancies to become apparent.

3. *What advances can be made to Sparse Distributed Memory as a cognitive model and decision support system?*
The principal unanswered issues identified in our exploration of SDM are those of 'spreading activation' and 'critical distance'. The first pertains to the activation of similar or near concepts from the initial activation of one. To this author, the idea of 'bear' quickly activates the concepts 'polar', 'wild', 'large' and 'dangerous' whereas to a photographer at National Geographic it may activate 'interesting', 'photo lens' or a host of other concepts. The Sparse Distributed model, as explored herein, is one whose architecture is predisposed towards *convergence,* that is, from an initial point of ambiguity or partial correctness the memory iteratively searches for a *single* correct response to its stimulus. An SDM that allows not only convergence but directed divergence, i. e.a number of valid responses to an initial input will potentially allow: i) the encoding of a slip-net [Hofstadter e FARG 1995] toward high-level analogy encoding and generation, and ii) the storage and retrieval of previously explored hypothesis spaces in hierarchical Bayesian models, through which a hybrid system may be explored.
The second issue, 'critical distance', is already under investigation [Brogliato 2011] and bears on the unexplored nature of the convergence threshold exhibited in Sparse Distributed Memory. There is a threshold distance after which an input stimulus will not converge towards a valid output bitstring. After this threshold the SDM will either fall on a tip-of-the-tongue scenario (never converging or diverging) or diverge altoghether. A better understanding of the underlying mechanisms of this phenomenon, with the critical distance as its main indicator, may further our understanding of SDM and of its plausibility as a cognitive model of human memory and decision-making.

This work is an initial inquiry into the contributions that modern computational cognitive science can provide to the mangement and decision sciences. There is much that may yet be explored in this increasingly crossed frontier, and one hopes this elementary foray may have lent a contribution to this progressively important exchange.

Part V

APPENDIX

APENDIX A: APPLYING KT-STRUCTURES: A TUTORIAL FOR DECISION SCIENTISTS

In this appendix we provide a step-by-step tutorial, so that other researchers may promptly apply their own datasets to this new method.

On the information-processing of the method

The method works by pressuposing initially that all entities (in our case, schools) are contained in a single cluster. The method then, given a specified form F and the dataset D, searches for the best structure that represents the data. In the online supplement we present a video of the method's convergence, from a single all-encompassing cluster to a series of 'splits' and re-adjustments.

Software Requirements

Kemp and Tenenbaum host their code and data sets at http://charleskemp.com/code/formdiscovery1.0.tar.gz. The code is written in the Matlab (Matrix Laboratory) framework, which is proprietary software, though widely available. We are at this stage attempting to execute the code in the open-source alternative, GNU Octave (www.gnu.org/software/octave). We are also starting a translation to Python (www.python.org). The code also has dependencies on the open-source GraphViz package (http://graphviz.org/), an advanced package that enables numerous functions for drawing graphs and trees.

Tutorial

There are many steps that need to be taken in order to execute the method in a new dataset. The following files must be configured:

File `setps.m`: This is one of the parameter configuration files. It has the vectors

```
ps.data = {'demo_chain_feat', 'demo_ring_feat',...
'demo_tree_feat', 'demo_ring_rel_bin',...
'demo_hierarchy_rel_bin', 'demo_order_rel_freq',...
'synthpartition', 'synthchain', 'synthring',...
'synthtree', 'synthgrid', 'animals',...
'judges', 'colors', 'faces', 'cities',...
'mangabeys','bushcabinet', 'kularing',...
'prisoners', 'schools'};
```

and

```
ps.dlocs = {[b, 'demo_chain_feat'], [b, 'demo_ring_feat'],...
[b, 'demo_tree_feat'], [b, 'demo_ring_rel_bin'],...
[b, 'demo_hierarchy_rel_bin'], [b, 'demo_order_rel_freq'],...
[b, 'synthpartition'], [b, 'synthchain'], [b, 'synthring'],...
[b, 'synthtree'], [b, 'synthgrid'], [b, 'animals'],...
[b, 'judges'], [b, 'colors'], [b, 'faces'], [b, 'cities'],...
[b, 'mangabeys'], [b, 'bushcabinet'], [b, 'kularing'],...
[b, 'prisoners'], [b, 'schools']};
```

The user must include the name of the new data file in both these vectors. In our case, the inclusion in `ps.data` and in `ps.dlocs` is of the last entry, `'schools'`, and also `[b, 'schools']`, correspondingly.

File `setrunps.m`: This file needs to be altered according to the nature of the data. Is it feature data? Is it similarity data? Is it relational data?

File `masterrun.m`: this is the main program file. A number of small changes must be made here. First, the directory path in which GraphViz is installed must be set. For example, in a windows machine:

```
[s,w] = system('C:\Program Files\Graphviz\bin\gvedit.exe');
```

The following vectors also need to be changed:

```
thisstruct = [1,3,6];
```

and

```
thisdata = [1:5];
```

Note that the numbers here must reflect the positions given in the vector `ps.structures` and `ps.data` (both found on file `setps.m`). In the above example, the system will load the first five entries of `ps.data`, each at a time, as input to search for structures, and it will search for the types of structures in entries 1, 3, and 6 of `ps.structures` (which are `'partition'`, `'order'`, and `'tree'`).

Obviously, the data files must be in the `\data` subdirectory.

After these steps are complete, typing `masterrun` at the Matlab prompt will start the execution the program.

BIBLIOGRAPHY

[Abe 2009]ABE, M. "Counting your customers" one by one: A hierarchical bayes extension to the pareto/nbd model. *Market Science*, v. 28, p. 541–553, 2009. (Cited on page 12.)

[Ahna e Ezawa 1997]AHNA, J.; EZAWA, K. Decision support for real-time telemarketing operations through bayesian network learning. *Decision Support Systems*, v. 21, p. 17–27, 1997. (Cited on page 12.)

[Alvesson 1990]ALVESSON, M. Organization: From substance to image? *Organization Studies*, v. 11, p. 373–394, 1990. (Cited on page 7.)

[Anwar e Franklin 2003]ANWAR, A.; FRANKLIN, S. Sparse distributed memory for 'conscious' software agents. *Cognitive Systems Research*, v. 4, p. 339–354, 2003. (Cited on page 26.)

[Ariely 2008]ARIELY, D. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: Harper, 2008. (Cited on page 3.)

[Ariely 2010]ARIELY, D. *The Upside of Irrationality: The Unexpected Benefits of Defying Logic at Work and at Home*. New York: Harper, 2010. (Cited on page 3.)

[Assisi et al. 2007]ASSISI, C. et al. Adaptive regulation of sparseness by feedforward inhibition. *Nature Neuroscience*, v. 10, p. 1176–1184, 2007. (Cited on page 32.)

[Basel 2004]BASEL, B. Families of strong kt structures in six dimensions. *Commentarii Mathematici Helvetici*, v. 79, p. 317–340, 2004. (Cited on page 11.)

[Boland 1989]BOLAND, P. J. Majority systems and the Condorcet Jury Theorem. *The Statistitian*, v. 38, p. 181–189, 1989. (Cited on page 27.)

[Borisyuk et al. 2000]BORISYUK, R. et al. An oscillatory neural network model of sparse distributed memory and novelty detection. *Biosystems*, v. 58, p. 265–272, 2000. (Cited on pages 31 and 32.)

[Brogliato 2011]BROGLIATO, M. S. *Understanding the Critical Distance in Sparse Distributed Memory*. Dissertação (Mestrado) — Escola Brasileira de Administração Pública e de Empresas - EBAPE, Fundação Getulio Vargas, 2011. Forthcoming. (Cited on page 40.)

[Cancho e Solé 2001]CANCHO, R.; SOLÉ, R. The small world of human language. *Proceedings of the Royal Society B*, v. 268, p. 2261–2265, 2001. (Cited on pages 24 and 31.)

[Corley e Gioia 2000]CORLEY, K.; GIOIA, D. The rankings game: Managing business school reputation. *Corporate Reputation Review*, v. 3, p. 319–333, 2000. (Cited on page 7.)

[Cornelissen e Thorpe 2002]CORNELISSEN, J.; THORPE, R. Measuring a business school's reputation: perspectives, problems, and prospects. *European Management Journal*, v. 20, p. 172–178, 2002. (Cited on page 7.)

[Cowan 2000]COWAN, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, v. 24, p. 87–185, 2000. (Cited on page 28.)

[Damoulas e Girolami 2009]DAMOULAS, T.; GIROLAMI, M. Combining feature spaces for classification. *Pattern Recognition*, v. 42, p. 2671–2683, 2009. (Cited on page 12.)

[Dichev 1999]DICHEV, I. How good are business school rankings? *Journal of Business*, v. 72, n. 2, p. 201–213, April 1999. (Cited on page 7.)

[Ehrenberg et al. 2001]EHRENBERG, R. et al. Paying our presidents: What do trustees value? *Review of Higher Education*, v. 25, p. 15–37, 2001. (Cited on page 8.)

[Elsbach e Kramer 1996]ELSBACH, K.; KRAMER, R. Members' responses to organizational identity threats: Encountering and countering the business week rankings. *Administrative Science Quarterly*, v. 41, p. 442–476, 1996. (Cited on page 8.)

[Fee et al. 2005]FEE, C. et al. Business school rankings and business school deans: A study of nonprofit governance. *Financial Management*, v. 34, p. 143–166, 2005. (Cited on page 8.)

[Florian 2007]FLORIAN, R. Irreproducibility of the results of the shanghai academic ranking of world universities. *Scientometrics*, v. 72, p. 25–32, 2007. (Cited on page 8.)

[French 1997]FRENCH, R. M. When coffee cups are like old elephants, or why representation modules dont make sense. In: RIEGLER, A.; PESCHL, M. (Ed.). *Proceedings of the 1997 International Conference on New Trends in Cognitive Science*. [S.l.], 1997. p. 158–163. (Cited on page 25.)

[Gavetti et al. 2005]GAVETTI, G. et al. Strategy making in novel and complex worlds: The power of analogy. *Strategic Management Journal*, v. 26, n. 8, p. 691–712, August 2005. (Cited on page 3.)

[Gavetti e Warglien 2007]GAVETTI, G.; WARGLIEN, M. Recognizing the new: A multi-agent model of analogy in strategic decision-making. Revise and resubmit, Administrative Science Quarterly. Oct. 2007. (Cited on page 3.)

[Gayler 2003]GAYLER, R. Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. In: SLEZAK, P. (Ed.). *ICCS ASCS International Conference on Cognitive Science*. Sydney, Australia, 2003. p. 133–138. (Cited on pages 23 and 26.)

[Gioia e Corley 2002]GIOIA, D.; CORLEY, K. Being good versus looking good: Business school rankings and the circean transformation from substance to image. *Academy of Management Learning & Education*, v. 1, n. 1, p. 107–120, september 2002. (Cited on page 7.)

[Gioia et al. 2000]GIOIA, D. et al. Organizational identity, image and adaptive instability. *Academy of Management Review*, v. 25, p. 63–81, 2000. (Cited on page 7.)

[Gloeckler et al. 2008]GLOECKLER, G. et al. *The Best U.S. B-Schools Of 2008*. 2008. Acessed 23-05-2011. Disponível em: <www.businessweek.com/interactive_reports/mba_domestic_2008.html>. (Cited on page 15.)

[Gobet et al. 2001]GOBET, F. et al. Chunking mechanisms in human learning. *Trends in Cognitive Science*, v. 5, p. 236–243, 2001. (Cited on page 23.)

[Gobet e Simon 2000]GOBET, F.; SIMON, H. Five seconds or sixty? presentation time in expert memory. *Cognitive Science*, v. 24, p. 651–682, 2000. (Cited on page 23.)

[Hofstadter 1985]HOFSTADTER, D. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. [S.l.]: Basic Books, 1985. (Cited on page 23.)

[Hofstadter e FARG 1995]HOFSTADTER, D.; FARG. *Fluid Concept and Creative Analogies: Computer Models of The Fundamental Mechanisms of Thought*. New York, NY: Basic Books, 1995. (Cited on pages 4, 39, and 40.)

[Ioannidis et al. 2007]IOANNIDIS, J. et al. International ranking systems for universities and institutions: a critical appraisal. *BMC Medicine*, v. 5, p. 30, 2007. (Cited on page 8.)

[Kanerva 1988]KANERVA, P. *Sparse Distributed Memory*. [S.l.]: MIT Press, 1988. (Cited on pages 4, 23, 24, 25, and 26.)

[Kanerva 1993]KANERVA, P. Large patterns make great symbols: An example of learning from example. In: . Denver Colorado: [s.n.], 1993. (Cited on page 23.)

[Kanerva 1994]KANERVA, P. The spatter code for encoding concepts at many levels. In: MARINARO, M.; MORASSO, P. (Ed.). *ICANN '94, Proceedings of International Conference on Artificial Neural Networks*. London: Springer-Verlag, 1994. v. 1, p. 226–229. (Cited on page 24.)

[Kanerva 1996]KANERVA, P. Binary spatter-coding of ordered k-tuples. In: *In ICANN96, Artificial Neural Networks*. [S.l.]: Springer, 1996. p. 869–873. (Cited on page 32.)

[Kanerva 2009]KANERVA, P. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, v. 1, p. 139–159, 2009. (Cited on pages 24 and 32.)

[Kemp 2007]KEMP, C. *The acquisition of inductive constraints*. Tese (Doutorado) — MIT, 2007. (Cited on pages 12 and 39.)

[Kemp e Tenenbaum 2008]KEMP, C.; TENENBAUM, J. The discovery of structural form. *Proceedings of the National Academy of Sciences*, v. 105, n. 31, p. 10687–10692, 2008. (Cited on pages 11, 12, and 13.)

[Kim et al. 2005]KIM, Y. et al. Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, v. 51, n. 2, p. 264–276, 2005. (Cited on page 4.)

[Linhares 2000]LINHARES, A. A glimpse at the metaphysics of bongard problems. *Artificial Intelligence*, v. 121, p. 251–270, 2000. (Cited on pages 15 and 32.)

[Linhares e Brum 2007]LINHARES, A.; BRUM, P. Understanding our understanding of strategic scenarios: What role do chunks play? *Cognitive Science*, v. 31, p. 989–1007, 2007. (Cited on pages 15 and 23.)

[Linhares et al. 2011]LINHARES, A. et al. The emergence of miller's magic number on a sparse distributed memory. *Public Library of Science (PLOS) One*, v. 6, n. 1, p. e15592, Jan 2011. (Cited on page 4.)

[Linhares e Freitas 2010]LINHARES, A.; FREITAS, A. Questioning chase and simon's (1973) "perception in chess": The "experience recognition" hypothesis. *New Ideas in Psychology*, v. 28, p. 64–78, 2010. (Cited on pages 4, 10, and 23.)

[Liu e Cheng 2005]LIU, N. C.; CHENG, Y. The academic ranking of world universities. *Higher Education in Europe*, v. 30, p. 127–136, 2005. (Cited on page 8.)

[Lloyd 2002]LLOYD, S. Computational capacity of the universe. *Physics Review Letters*, v. 88, p. 237901–1:4, 2002. (Cited on page 25.)

[Lock e Gelman 2010]LOCK, K.; GELMAN, A. Bayesian combination of state polls and election forecasts. *Political Analysis*, v. 18, n. 3, p. 337–348, 2010. (Cited on page 12.)

[Lönnstedt e Britton 2005]LÖNNSTEDT, I.; BRITTON, A. Hierarchical bayes models for cdna microarray gene expression. *Biostatistics*, v. 6, p. 279–291, 2005. (Cited on page 12.)

[Miller 1955]MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, v. 63, p. 81–97, 1955. (Cited on pages 23 and 32.)

[Natter et al. 2001]NATTER, M. et al. The effect of incentive schemes and organizational arrangements on the new product development process. *Management Science*, v. 47, n. 8, p. 1029–1045, 2001. (Cited on page 4.)

[Perfors e Tenenbaum 2009]PERFORS, A.; TENENBAUM, J. Learning to learn categories. In: TAATGEN, N. et al. (Ed.). *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin TX: Cognitive Science Society, 2009. p. 136–141. (Cited on page 12.)

[Perfors et al. 2011]PERFORS, A. et al. The learnability of abstract syntactic principles. *Cognition*, v. 118, n. 3, p. 306–338, 2011. (Cited on page 12.)

[Pfeffer e Fong 2004]PFEFFER, J.; FONG, C. The business school "business": Some lessons from the U.S. experience. *Journal of Management Studies*, v. 41, n. 8, p. 1501–1520, 2004. (Cited on pages 7 and 8.)

[Piramuthu et al. 1998]PIRAMUTHU, S. et al. Using feature construction to improve the performance of neural networks. *Management Science*, v. 44, n. 3, p. 416–430, 1998. (Cited on page 4.)

[Plate 2003]PLATE, T. *Holographic reduced representations: Distributed representations for cognitive structures*. Stanford CA: CSLI Publications, 2003. (Cited on pages 24 and 26.)

[Ries e Trout 1993]RIES, A.; TROUT, J. *The 22 immutable laws of marketing*. New York: HarperCollins, 1993. (Cited on page 3.)

[Siemens et al. 2005]SIEMENS, J. et al. An examination of the relationship between research productivity in prestigious business journals and popular press business school rankings. *Journal of Business Research*, v. 58, n. 4, p. 467–476, April 2005. (Cited on page 8.)

[Smith e Winterfeldt 2004]SMITH, J.; WINTERFELDT, D. von. Decision analysis in management science. *Management Science*, v. 50, p. 561–574, 2004. (Cited on page 3.)

[Stewart e Eliasmith 2009]STEWART, T.; ELIASMITH, M. Oxford handbook of compositionality. In: _____. [S.l.]: Oxford, 2009. cap. Compositionality and Biologically Plausible Models. (Cited on pages 23 and 26.)

[Tracy e Waldfogel 1997]TRACY, J.; WALDFOGEL, J. The best business schools: a market-based approach. *Journal of Business*, v. 70, p. 1–31, 1997. (Cited on page 7.)

[Trieschmann et al. 2000]TRIESCHMANN, J. et al. Serving multiple constituencies in business schools: M.B.A. program versus research performance. *Academy of Management Journal*, v. 43, p. 1130–1141, 2000. (Cited on page 8.)

[Zell 2001]ZELL, D. The market-driven business school: Has the pendulum swung too far? *Journal of Management Inquiry*, v. 10, p. 324–338, 2001. (Cited on pages 7 and 8.)

[Zhoua e Kapoor]ZHOUA, W.; KAPOOR, G. Detecting evolutionary financial statement fraud. Accepted for publication, doi:10.1016/j.dss.2010.08.007. (Cited on page 12.)

[Zhoua et al.]ZHOUA, Z. et al. Bayesian reasoning approach based recursive algorithm for online updating belief rule based expert system of pipeline leak detection. Expert Systems with Applications, accepted for publication. (Cited on page 12.)