Quantitative Physical Modeling of Physics Cognition

Steven Hamed Hassani

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Physics
Advisers: Andrea diSessa and William Bialek

June 2012

© Copyright by Steven Hamed Hassani, 2012.

All Rights Reserved

Abstract

Current theories of physics cognition require specification of complex mechanisms for explaining knowledge acquisition. I demonstrate that a quantitative model of physics perception can be constructed by assuming that physics problem perception provides an optimal summary of a set of physics problems. In so doing, I offer the first model of physics cognition determined from a single optimization principle.

I use the model to produce categorizations of physics problems according to surface features. These categorizations suggest, contrary to previous claims, that surface feature perception may in fact be a productive resource for novices: it may provide access to "deep" knowledge originally considered by influential studies as accessible only to experts. The model suggests a potential explanation for why novices often focus on surface features: novices may simply be responding to a set of physics problems in which the surface features of those problems provide relevant information for problem solving.

The model predicts that the initial perception of a physics problem is characterized by the identification of the "surface feature context" of which the problem is a particular example. I use the model to predict potential contexts that individuals may perceive when confronted with a physics problem. Experiments that focus on the initial perception of a physics problem have not been considered previously; I use the model to encourage the construction of such experiments. I speculate that the large differences in novice and expert physics problem classification originally highlighted in influential experiments could reduce substantially when experimental observation is restricted to the initial, perceptual stage of physics problem classification.

Acknowledgements

I could not have explored whether or not quantitative physical modeling provides insight for physics education research without both Andy diSessa and Bill Bialek providing me with both the space and time to carry out this research. For sanctioning this unusual research agenda, along with their advice and support, I cannot fully express my gratitude.

I have enjoyed the hospitality and critical feedback provided by all who have participated in the Patterns group at the Graduate School of Education at UC Berkeley during my time working on this dissertation, including, but not limited to: Jeanne Bamberger, Lauren Barth-Cohen, Janet Casperson, Andy diSessa, Alyssa Dray Sayavedra, Shuly Kapon, Mariana Levin, Colleen Lewis, Angie Little, Hillary Swanson, and Jennifer Wang.

In order to defray the costs of visiting Princeton after spending time at Berkeley, I have called on my brother, Ali Hassani, to accommodate both my wife and myself in his small one bedroom apartment in Manhattan on many occasions, often for multiple months at a time. I'd also like to thank my wife's parents, Karen and Alan Nephew, along with my parents, Manige and Fereidoon Hassani, for providing constant encouragement over my entire tenure in graduate school.

Finally, I would like to reserve my final acknowledgement for my wife, Sara Nephew Hassani. Thank you for your love, no matter the circumstances.

To my wife, Sara.

Contents

Chapter 1

Prospects for Quantitative Physical Models of Physics Cognition

The nature and operation of cognition is one of the crucial frontiers in science. Physical methods have been successfully applied to understanding cognition in many important respects, from the use of neuroimaging in psychology to the development of models for understanding the nature of vision. It is perhaps surprising then, given the amount of time many physicists spend teaching physics, that physical methods have been largely absent in the development of models for understanding physics cognition. In this chapter, I discuss the prospects for quantitative physical modeling in physics cognition. First, I claim that the constructivist orientation to knowledge provides a potentially fruitful starting point for developing quantitative physical models of physics cognition. Second, I claim that quantitative physical methods provide a particular methodological approach to educational research that is both consistent with predominant educational theory and simultaneously allows researchers to probe research questions of relevance to education. Third, I claim that the attempt to find useful overlap between existing theories of physics cognition with quantitative physical methods moves us closer to a general theory of physics practice.

1.1 Physics Education: Significance and Challenges

Teaching physics to students in a manner that is both efficient and effective is of paramount importance to physicists. At the university level, a major responsibility of university physics departments is to teach introductory physics to undergraduates. Introductory physics courses are a staple of university science and engineering programs. At the high school level, physics is a ubiquitous course offering. In fact, 92 percent of all seniors attend a school where physics courses are regularly available (White, 2011). Physics instruction is directly important for students who are interested in STEM careers, and indirectly important for a much broader range of students who can benefit from the rigorous analytical skills physics is in a unique position to teach (Redish, 2000).

Even with the apparent widespread availability of physics instruction at the high school and college level, there are reasons to believe that current physics instructional strategies face some serious challenges. First, it has been alleged that introductory physics courses at the undergraduate level do not introduce a solid understanding of the material (Mazur, 1997). Second, only 37 percent of high school students will take a physics course (White, 2011) despite the apparent availability of such courses in high schools nationwide. Third, the Physics First initiative — started by the American Association of Physics Teachers — aims to improve student exposure to physics classes by teaching physics to students in ninth grade, as opposed to eleventh and twelfth grades (AAPT, 2007). This move creates significant potential challenges: in light of the limited math background of ninth grade students, traditional math-intensive strategies and techniques for physics instruction may require rethinking.

All these reasons — insufficient understandings of physics held by most undergraduates, low rates of physics enrollment in high school, and curricular changes that would involve teaching physics to students with limited math backgrounds — warrants more careful consideration of physics education. It begs a number of questions. To whom is physics accessible? In light of inadequate understandings of physics among undergraduate students, to what extent is physics inherently difficult and to what extent is its difficulty simply a consequence of poor teaching strategies?

How best to teach physics is not a concern that should only be relegated to high school teachers. The academic project of physics is not only to produce basic science — though indeed this is a fundamental and universally recognized objective. It is also to produce science education. Understanding how best to deliver science education to as broad an audience as possible, as effectively and efficiently as possible, should be an important aim for physicists. However, few, if any research studies, to my knowledge, use theoretical, quantitative methods originating from physics to understanding relevant problems in physics education. It is the primary aim of this dissertation to warrant the possibility that physical methods, though not necessarily easily applied in this context, may actually be of service to relevant problems in physics education.

1.2 Physics Cognition as a Participatory Physical Theory

Physicists have often been interested in the role of the observer and of his or her experimental apparatus in determining the character of physical law. The Copenhagen interpretation of quantum mechanics is one important example taught in standard undergraduate physics courses; it states how the act of measurement affects experimental results. Also highlighting the potential role of the observer in constraining physical law are the various forms of the anthropic principle, which state that many aspects of physical law such as the values for various constants of nature take the form or values they do because those values are consistent with intelligent life (Carter, 1973; Barrow and Tipler, 1988). "It from Bit" is a particular anthropic proposal that makes explicit the notion that the character of physical law ultimately derives from our participation in its creation, by stating that our view of the universe ultimately derives from answers to yes or no questions that we ask. This implies that physical law is ultimately information theoretic in origin (Wheeler, 1990). Many critiques have been leveled at anthropic principles, a common one being that it can be invoked easily as explanatory, such that it discourages the search for explanatory theories which do not require invoking the properties of the observer (Penrose, 1989).

While the debate continues among physicists over whether the observer should be invoked when constructing physical law, physics education researchers emphasize that *understanding* of the natural world does in fact depend on the observer. Physics education researchers have documented many instances in which individuals, even after taking courses in introductory physics, do not have the same understandings of the natural world as experts do (Confrey, 1990). Furthermore, physics education researchers are by and large uniform in their view that even individuals who have not had formal physics do interpret the natural world, though not in the same way as experts in physics. Individuals who have not had physics still must participate in the natural world, which means that they must have some knowledge concerning how the natural world operates (diSessa, 1993, 2006). Participation in the natural world breeds intuitive knowledge necessary to operate within it.

Learning physics in the classroom is also a participatory phenomena, which proceeds through interactions among students, the classroom, other peers, and learning materials. Therefore, even the cognitive instantiation of normative knowledge, as taught through learning environments, is constructed in part through a participatory interaction between one's cognition and the environment. Normative physical law may or may not explicitly require the presence of an observer, but its cognitive instantiation is always participatory.

Developing theories of physics cognition for how human beings understand the natural world may differ tremendously from the development of other, more traditional theories of nature: the role of the observer must be explicitly embraced. Since human beings are the ones that practice physics and discover its laws, developing a theory for how humans cognitively understand the natural world may speculatively, in the long run, provide another way to constrain physical law. Of course, this becomes another anthropic argument for constraining physical law; however, physical law, in this case, is constrained by the fundamental processes underlying our own cognition.

Developing theories for physics cognition could also help illuminate nuance in the character of physics knowledge, which is not necessarily codified in physics textbooks. Physicists, as researchers, are charged with the task of finding concise and powerful principles that govern natural phenomena. But physicists also function as educators, charged with teaching the practice of physics to students. Physicists generally recognize that simply stating the laws of physics to students is not sufficient to teach students how to productively practice physics. Physics practice requires that knowledge of these physical laws be supplemented by other skills or knowledge. Unfortunately, physics practice is not well understood by physicists, partially because physicists lack concise principles governing the relevant cognitive processes. Do concise principles of physics cognition exist, which when added to the standard laws of physics transform physics from a theory of the external natural world to a theory of physics practice?

1.3 Macroscopic Physical Theories

Human cognition is a tremendously complex phenomenon, taking place in a system involving $\sim 10^{11}$ neurons and $\sim 10^{14}$ synapses (Williams and Herrup, 1988). From a reductionist orientation to physics, this is not problematic in principle since most physicists believe (or at least hope) that all processes in the natural world are ultimately governed by the fundamental laws of particle physics and general relativity (Weinberg, 1995). In practice however, directly applying reductionist approaches to understanding macroscopic processes in the brain has generally been untenable, not least of which is because the calculations become too difficult to manage — even computationally.

But large systems often display simple regularities in macroscopic behavior that do not require, in order to make macroscopic predictions, an appeal to the fundamental laws of physics. The laws of thermodynamics, which historically appeared before an understanding of atomic structure (Gibbs, 1876), are one important example. Contemporary axiomatic formulations exist for thermodynamics without the need to appeal to quantum mechanics (Callen, 1985).

Physicists have also recognized that simple regularities in macroscopic behavior often appear to be qualitatively different from the evolution of any particular microscopic constituent. This is because particles in a system display collective, emergent behavior, which can only be seen by studying the evolution of the entire system as a group. The system as a collectivity may be governed by different equations than its constituent parts.

For example, consider the time asymmetry in the evolution of macroscopic systems implied by the second law of thermodynamics. The second law states that even though the constituent parts of a system may be governed by fundamental equations that obey time reversal symmetry, macroscopic systems tend to evolve in a particular direction only — that is towards the direction of higher entropy. The second law can be understood from the statistical, reductionist perspective, as emergent from the fact that probabilistic considerations become a dominant governing concern of many large systems (Ben-Naim, 2008).

The second law of thermodynamics is one example of a principle of physics that was known and used productively without a reductionist understanding in place (Gibbs, 1876; Clausius, 1879). Chaos and symmetry breaking in condensed matter physics are two other important examples. Physicists have argued that these examples of emergent principles need not be unique: as one moves towards understanding other systems, like cognitive processes, principles for understanding these phenomena may exist which do not bear any direct resemblance to their lower level constituent principles but are nevertheless predictive of macroscopic phenomena (Anderson, 1972; Bialek, 2002).

1.4 Constructivism and the Importance of Experiential Data

Physical theories of physics cognition must be participatory and therefore anthropic in character. Unfortunately, a quantitative understanding of these processes has been elusive. Yet, a qualitative blueprint does exist for how to think about these processes, which is grounded in research in physics education and is firmly entrenched in the epistemology called *constructivism*.

It is widely believed by physics education researchers that novices are not blank slates; instead, novices bring existing knowledge and skills to bear upon their understandings of the physical world. The observation that novices bring prior knowledge is a perspective that has its roots in constructivism, which is generally attributed to Jean Piaget. Briefly put, constructivism states that learning proceeds through an interaction between experiential data and existing skills and knowledge (Steffe and Gale, 1995; Smith et al., 1993 - 1994).

Because different individuals have different experiences, whether through varying types of instruction in different classrooms, or varying exposure to the natural world, a conclusion to draw from constructivism is that these different experiences should imply differences in knowledge. This means that it is crucial for any legitimate quantitative theory of physics cognition which wishes to explain the origin of knowledge to be able to parameterize the experiences that an individual has.

Surprisingly few, if any, existing quantitative theories of physics cognition explicitly include the role of experience in the development of quantitative modeling. For example, consider the computational programs of problem solving developed by Larkin et al. (1980a,b) — which are possibly the most famous of all quantitative models in physics education. These programs model an individual's knowledge of physics and algebra using a set of "productions". A production tests whether a particular condition exists in either working memory or on paper, and if so, executes an action that adds to that working information. Productions continue to act iteratively on this set of working information until a solution to the problem is reached. The entire problem solving evolution is modeled as a finite sequence of productions.

The goal of writing these programs was to simulate the problem solving processes of novices and experts. The methodology followed by these researchers to construct their models followed from a tradition of specifying cognitive processes using computer algorithms (Newell and Simon, 1976). Unfortunately, a dearth of quantitative models relevant to physics education followed after this early study. It is my claim, which I try to illustrate throughout this chapter, that a paradigm grounded in a marriage of the physical method and the perspective offered by constructivism may renew the search for relevant quantitative models.

From a physicists' perspective, the models of Larkin et al. (1980a,b) may not provide the most sensible starting point for constructing physical models of physics cognition. Even if production rules constitute physics cognition after learning, they are most certainly not fundamental, objective constituents. Constructivism teaches us that knowledge is constructed and subjective, in part, as a result of exposure to experiential data. Much of the normative knowledge of physics included in these models is most definitely learned through instruction; however, the models of Larkin et al. (1980a,b) do not account for experiential data. Quantitative physical modeling in education, if it is to have direct relevance to physics education research that addresses the origin and transformation of knowledge (Vosniadou, 2008), should be able to provide insight into the reason why knowledge has the particular content and structure that it does.

If we accept that the most relevant models in physics education will require a complex knowledge system in order to predict relevant behavior, this does not necessarily imply that models must include this knowledge a-priori. Complexity in knowledge is constructed in part through interaction with a complex world. If the physicist is able to parameterize experiential data adequately, it may be possible that one or more simple cognitive principles, acting on that data, may be able to produce a rich set of cognitive abstractions (i.e. knowledge) that predicts relevant educational phenomena. Since experiential data, like words in textbooks or a teacher's speech, are objective signals, including experiential data as objective model input may provide a direct means of predicting objective behavior from an objective origin. A complex physics knowledge system, rather than having to be stated by fiat in the model, may possibly be derivable quantitatively through the complexity of objective experiential input.

1.5 Functions of Physics Cognition

Even though physics knowledge derives in part from experiential data, constructivism holds that it is also forged in the context of preexisting skills and knowledge. Part of the difficulty in constructing a quantitative physical theory of physics cognition is identifying, and then defining quantitatively, those basic cognitive skills and/or knowledge that are used to organize experiential data. But what form should these basic principles of physics cognition take?

Bialek (2002) offers one potential heuristic for constructing basic macroscopic principles of cognition. Rather than focusing on the complex underlying mechanisms governing cognition, he instead asks theorists to try to identify the most basic functions of the brain. In particular, he notes that living beings may be confronted with a wide array of basic tasks; one important function of our brain is to provide solutions for these basic tasks. Bialek (2002) asks theorists to consider whether a simple notion of optimal performance for these basic tasks can be defined. By constructing optimal solutions to these basic tasks a-priori, they become predictive models against which experimental results can be compared.

Physics cognition serves many complex functions. For example, the task of solving an introductory physics problem is probably a complex task of the brain, in the sense that it probably involves the execution of a number of simple tasks together. The models of Larkin et al. (1980a,b) illustrate this intuition: a large number of production rules were written to simulate the problem solving process. The goal of the research paradigm outlined here is to identify the most basic, fundamental tasks physics cognition confronts, whose simple solutions together may provide solutions to the more complex tasks of physics practice — like physics problem solving.

The complex models of Larkin et al. (1980a,b) meet a relative definition of optimal performance: their expert problem solving model both takes fewer steps and makes fewer errors than the novice model. This overall point of view that novices are seen as suboptimal, relative to experts, is similar to the frame of reference provided by the "misconceptions" movement in science education (Confrey, 1990). In the misconceptions movement, educational difficulties among novices are seen as primarily a result of knowledge that is ill-suited, or suboptimal, for the formation of normative physics knowledge. Relative optimality is also the generally agreed upon defining feature of expertise: expert performance is characterized by more optimal performance, where optimality is defined according to the task, relative to novice performance on the same task (Ericsson and Smith, 1991). Other definitions of relative optimality could be defined, like the ability to solve a wider range of problems in this case. Hypothetically, if the expert model modeled the expert problem solving process through a simulation that took more steps and/or made more errors than the novice model, then the standard view of expert performance would require that the expert model be justified by invoking another definition of relative optimality.

While it is certainly true that there are important aspects of novice physics knowledge that do not match the well-tuned nature of expert cognition, this does not imply that novices may not have access to optimal solutions to some of the most basic tasks confronted in the course of physics practice (diSessa, 2006). It is unresolved exactly what constitute the basic tasks of normative physics practice, which of these tasks students can ultimately solve, and whether any of these practices can actually be characterized as quantitatively optimal solutions.

One way to start the search for basic principles of physics practice is to identify both a relevant but simple task physics cognition confronts, along with the experiential data which could be used to solve that task. The potential benefit to education research is twofold. First, this framework could produce simple principles of physics practice that actually approximate the solutions to the most basic tasks taken by physics cognition. Second, even if individuals were not to use the proposed solution to the identified task, it may be possible that the solution is not being used because it is not being emphasized in instruction. In this case, the proposed principle of physics practice could instead be viewed as a potentially untapped skill, which if marshaled by instruction, could be used for productive purposes. This may provide another possible vehicle for finding productive skills or knowledge upon which to build normative practices in physics.

1.6 Construction and Adaptation

Physics education researchers, following upon the constructivist idea that knowledge is constructed from previous knowledge and experience, have tended to focus on the *process* by which new knowledge is constructed. Many physics education researchers contend that students have difficulty learning physics because existing skills or knowledge either interferes or is not marshaled productively in order to learn formal physics. As a result, research has tended to focus on the necessary process involved in replacing or refining prior conceptions to form more normative conceptions — a process called "conceptual change" (Vosniadou, 2008; diSessa, 1993; Vosniadou, 2002; Vosniadou and Brewer, 1994; Wiser and Carey, 1983; Chi, 2005; diSessa and Sherin, 1998). The quantitative models of problem solving produced by Larkin et al. (1980a,b) were also primarily generated to model the process by which problem solving occurs.

Yet, modeling process quantitatively using a physical method presents potential issues. Perhaps most importantly, a correct description of a process must recognize that the process takes place continuously in time. One of the most fundamental critiques of quantitative approaches to cognition that model cognition algorithmically, such as in Larkin et al. (1980a,b), is that the notion of continuous time does not naturally appear in algorithmic descriptions of those processes. Instead, algorithmic analogies to cognitive processes take place in discrete steps. An algorithmic description that takes place discretely must not be a fundamental description of the cognitive process. This observation has prompted other researchers to look for descriptions of cognitive processes which explicitly include the role of time in those processes — generally using the tools offered by dynamical systems theory (see Port and van Gelder (1998) for a review).

The models constructed here do not primarily aim to model process. Instead of developing complex models which provide temporal predictions by constructing knowledge through a mechanism of continuous interaction between prior knowledge and experience, one might be able to ask whether a notion of an optimal solution to a simple task confronted by cognition exists independent of the mechanism of learning that produced that solution. In this case, one could assume that the entire history of experiences an individual has had up to that time is available as input into the model. By proposing that these solutions to simple tasks confronted by physics cognition may eventually be able to be accessed through learning, a predictive model may be able to be constructed that does not require the specification of complex learning mechanisms.

Predictive models constructed as optimality conditions are commonplace in standard physical theory. For example, the second law for isolated systems is an optimality principle which states that the entropy of isolated systems eventually reaches a maximum value at equilibrium. This principle allows for the prediction of the macroscopic end-state of many thermodynamic processes, without the need for an understanding of the non-equilibrium thermodynamics governing the temporal evolution which ultimately produced that end-state.

A related hypothesis takes another view towards constructed knowledge, which could be very important in a physics learning context like an introductory course in physics. Constructivism asserts that knowledge is both constructed from the history of past experience, but that in addition, it is constructed in order to serve a future function. As such, through the complex learning process, knowledge may also become *adapted* to the potential experiences an individual may have in that context (von Glasersfeld, 1995). For example, do students have knowledge that will help them solve problems on a final exam that they have never before confronted? In this case, physicists could assume that optimal solutions to the basic tasks which constitute problem solving are provided over the entire set of potential experiences. The origin of knowledge would then be explained by not showing how knowledge reflects past experience, but instead by stating that knowledge is optimally adapted to an environment of potential experiences an individual may confront.

1.7 Quantitative Physical Methods Allow Comparative Study

The physical method outlined here provides a vehicle for understanding the role of potential or actual experiences in the origin of knowledge. Though constructivism emphasizes the centrality of the environment of experiences in the origin of knowledge, existing quantitative methods have not generally been able to assess the role that the entire environment plays in the construction of knowledge. For example, the models of Larkin et al. (1980a,b) do not use an environment of experience, past or potential, in stating why problem solving knowledge takes the form it does. The models only use the most immediate experience, which is the physics problem being solved at the moment, as the only "experience" used to predict the realized process an individual takes when solving a problem.

A physical model constructed in the fashion outlined here will be explicitly context dependent. This means that predictions concerning differences in knowledge among two individuals will track differences in past or potential experiences vis-à-vis their respective learning contexts. For instance, differences in the syllabi of two different courses of physics implies differences in learning context between students. If those contexts can be properly parameterized, one may be able to predict differences in physical knowledge that mirrors contextual differences.

Comparative studies are consistently carried out in educational research. For example, the thread in physics education research comparing expert and novice knowledge highlights the cases in which novice knowledge is non-normative relative to expert knowledge (Confrey, 1990; Anzai, 1991). Even though an important goal in physics education research is to understand how educators might teach novices to become more like experts, studies highlighting novice misconceptions generally have not emphasized how educators might overcome these difficulties (diSessa, 2006). Yet, quantitative physical models may be able explain certain deficiencies in novice knowledge as simply due to an adaptation of novices to a deficient environment. Adjusting important features of the novice environment such that it looks closer to the expert environment may provide a step towards mitigating novice difficulties.

1.8 Quantitative Physical Methods May Offer Insight into the Structure of Knowledge

Physics educators generally adhere to the view that theorizing about knowledge is required to make relevant statements about learning processes. It is important to remember that knowledge contained within the brain is not directly observable with present experimental techniques. It is only observable indirectly; for example, through performance on exams or responses to interview questions.

Including theoretical constructs of knowledge that are not directly observable is not problematic from a physicists perspective. Electric fields, virtual particles, and the quantum mechanical wave function are all examples of theoretical constructs in physics which are only observable through their effects on other observables. The theories containing these constructs have been very successful.

In physics education research, a debate exists over the appropriate orientation to the structure of physics knowledge. The debate can be characterized loosely into two competing camps. Those working in the "knowledge-as-theory" perspective have taken the view that the most appropriate abstraction of knowledge is one that is sufficiently coherent and integrated as to be considered "theory-like" (Vosniadou, 2002; McCloskey, 1983). As an example of a particular "knowledge-as-theory" theory, Vosniadou and Brewer (1994) claim that novices in astronomy develop, before formal instruction, mental models of the earth that share the structure and coherence of more normative models but that have incorrect content. Those working in the competing "knowledge-in-pieces" orientation have taken the perspective that novice physics knowledge should instead be viewed as much more fragmented. In this orientation, physics knowledge is viewed as consisting of many elements, with potentially a weak degree of coherence between them. As an example of a "knowledge-in-pieces" theory, diSessa (1993) describes particular physical intuitions that novices have, before they learn formal physics, that they have gained through living and interacting with the world.

Given that many theories of knowledge have been developed to explain data in particular contexts, we cannot legitimately expect that the debate between the "knowledge-as-theory" camp and the "knowledge-in-pieces" camp will be finalized anytime soon. Some researchers have attempted to unify these divergent perspectives by stating the possibility that a knowledge-in-pieces approach may be more applicable in "rich domains" like mechanics, while a knowledge-as-theory approach may be more applicable in contexts in which novices do not necessarily have much personal experience (Özdemir and Clark, 2007; diSessa et al., 2004).

Quantitative physical models may also be able to make strong suggestions concerning the appropriate structure of knowledge. For example, does an optimal solution to a task physics cognition confronts require many knowledge elements, or only a few? It may be the case that the answer to this and other questions concerning the structure of knowledge ultimately depend on both the environment and the task itself, which is consistent with the view that the appropriate theoretical knowledge structure may ultimately be context-dependent.

1.9 Conclusion

In this chapter, I have introduced some of the pieces of a framework for constructing quantitative physical models of physics cognition. In particular, I have argued that physical models of physics cognition are necessarily anthropic. I have suggested grounding these anthropic models in the epistemology of constructivism. I speculate that physicists may be able to find principles governing physics cognition by quantifying the solutions to simple tasks encountered in physics practice.

I have not provided the precise form for the models advocated in this chapter. Even though constructivism may provide the first steps for thinking about physics cognition quantitatively, it does not completely constrain the form of these models. I have suggested that solutions to simple cognitive tasks might be able to be found by considering whether cognition solves these tasks optimally. Yet, the appropriate quantitative notion of optimality, from which a quantitative principle could emerge, has not been motivated in this chapter. The purpose of this research program, broadly speaking, is to attempt to identify generic principles of optimality in cognition at the macroscopic level.

As Bialek (2002) demonstrates, one approach for identifying generic principles is through selecting problems in cognition that may be productively attacked using the physical method. A uniform quantitative framework may emerge for how to think about cognitive problems at the macroscopic level if physicists successfully predict phenomena over a wide enough array of such problems. There are hints that information theoretic notions may indeed be the generic unifying piece (Bialek, 2002; Wheeler, 1990). Later chapters of this dissertation reinforce this suggestion by applying information theory to a particular simple task that physics cognition may confront.

Physics cognition is a particularly rich phenomenon. Though normative physics is learned in classrooms and textbooks, verified in laboratories, and extended in theorists' offices, an understanding of the natural world is not reserved simply for those who practice normative physics. As diSessa (1993) has suggested, individuals who do not practice normative physics also have an understanding of the natural world, simply by experiencing it and having to adapt to it. Physics practice involves more than simply rote memorization of the laws of physics; a large set of other knowledge is required to take what is stated in textbooks and research papers and actually practice physics productively. A large literature in physics education research documents the diversity of physics practice; such literature may provide a starting point for physicists looking for aspects of physics practice that may be quantifiable.

Chapter 2

The Productive Potential of Surface Features

In this chapter, I will identify a potentially quantifiable task confronted by physics cognition. This task, the categorization of physics problems according to "similarity of solution", has generally been thought to be completed differently by novices and experts. I argue that the way novices complete this task may be an indicator of a general productive principle of physics practice, used by both novices and experts. This behavior, the categorization of problems using "surface features," may be indicative of an essential resource for inferring important "deep structure" necessary for solving a problem. Using the outline for physical modeling provided in the previous chapter, I develop a conjecture for how surface features may be used in physics problem solving.

2.1 The Study by Chi et al.

In this chapter, I will critically review a set of experimental tasks examined in the influential study of Chi et al. (1981). In the first two experiments conducted by Chi et al. (1981), both novices and experts in physics were asked to categorize a series of 24 introductory mechanics problems according to "similarity of solution." Novices, who had already taken a course in physics, categorized these problems according to "surface features" — that is, the literal features (like incline plane) used in the problems. Experts in physics categorized the problems according to "deep structure," which are the underlying physical principles (like conservation of energy) that govern the phenomena asked in the problem.

Yet, the reason why novices and experts produced these categorizations was not adequately addressed empirically in this study. Chi et al. (1981) suggests one possible reason for the difference between novice and expert categorizations:

> Only a physicist can detect the similarity underlying the expert's categorization.

Essentially, Chi et al. (1981) argue that only experts are capable of identifying the principles (like conservation of energy) in physics problems. Later, Chi (1993) stated:

> The basic expert-novice result, that experts' knowledge is represented at a "deep" level (however one characterizes "deep"), while novices' knowledge is represented at a more concrete level, has been replicated in many domains, ranging from knowledge possessed by scientists to taxi drivers. This result can also be used to interpret findings in many related cognitive science topics, e.g., analogical reasoning, and concepts and categories.

Chi (2006) restates this argument by stating that one characteristic of expertise is that experts can "perceive the 'deep structure' of a problem or situation." Many others have cited the notion that a canonical feature of expertise is the ability to see the "deep structure" in a situation, and that expert knowledge is organized around schemas of "deep structure" (VanLehn et al., 2005; Anzai, 1991; Hmelo-Silver, 2004; Kalyuga et al., 2003; Aleven and Koedinger, 2002). The paper by Chi et al. (1981) has been identified by the editorial board of the journal *Cognitive Science* as one of the ten classics articles ever to appear in that journal. According to Google Scholar, it is the second most highly cited paper in the history of that journal.

In spite of these accolades, the view that only experts are able to see underlying abstract features has been challenged by others. For example, Smith et al. (1993 - 1994) contends that intuitive physics knowledge of mechanics that all individuals hold, before learning formal physics, contains both "deep structure" as well as "surface structure." Yet, to my knowledge, there does not exist a detailed reassessment of the fundamental claim by Chi et al. (1981) that novices, after having taken a course in physics, are relatively unable to access the "deep structure" of normative physics. One of the goals of this chapter is to explicitly revisit this claim.

2.2 Contextuality in Physics Problem Solving

Chi et al. (1981) was interested in the knowledge novices and experts had available to them for physics problem solving. They interpreted the results of their two categorization experiments using a form of the "perceptual chunking" hypothesis (Chase and Simon, 1973; Gobet and Simon, 2000; Gobet et al., 2001). The perceptual chunking hypothesis contends that the knowledge necessary to perform over the large set of contexts encountered in many domains, such as playing chess or solving physics problems, is managed primarily with "chunks" of knowledge. In the physics problem solving context, rather than the brain storing the solution for all problems an individual may confront, a single chunk of knowledge may be stored that could be useful for a number of problems simultaneously. Chi et al. (1981) hypothesize that many of these perceptual chunks are indexed according to the "types" of problems that an individual perceives: they hypothesize that the particular chunk that is initially cued in the course of problem solving depends on the initial inference of the problem's type. For example, novices may be responsible for many different possible incline plane problems, but rather than memorizing the precise solution path for all of these different problems individually, the brain instead may store relevant knowledge for how to infer an incline plane problem's forces in a perceptual chunk of knowledge indexed by the "incline plane" problem type. This knowledge is then cued if the incline plane type is perceived when confronting a problem, and that knowledge then becomes available in order to infer a solution.

Since knowledge itself is not directly observable, Chi et al. (1981) argue that the categorizations produced by novices and experts in their first two categorization experiments were adequate proxies for the types that those individuals recognize in the course of problem solving. Novices, based on the surface feature categorizations they produced, were assumed to perceive different problems, in the context of problem solving, as belonging to different "surface feature" types (like inclined planes or springs). Experts, because of their "deep structure" categorizations, were assumed to perceive the same problems as belonging to different "deep structure" types (like conservation of energy or Newton's 2nd law).

Many of the deep structure categories produced by the experts spanned multiple surface feature contexts. For example, experts placed a problem involving a spring and a problem involving an incline plane into the same category, because both used the conservation of energy. Chi et al. (1981) took the findings of their experimental categorization task and interpreted them as implying that expert knowledge for problem solving is indexed in a context-independent way, using the normative principles of physics. On the other hand, novice knowledge available for problem solving was interpreted as being indexed, in a context-dependent way, using surface feature types.

In the same study (Chi et al., 1981), a third experiment was conducted that complicates the interpretation of the first two categorization experiments. In this experiment, they asked two novices and two experts to discuss everything they knew about problems of the incline plane type, without providing a particular physics problem to ground the discussion. The experts discussed the applicability of different physics principles for the incline plane type. The applicability of one deep structure principle for an incline plane problem, versus another, was described by the experts as depending upon the particular surface features used in the incline plane problem. For example, Chi et al. (1981) characterizes one expert elaboration of the incline plane type, which they translated into a "production rule" form (Larkin et al., 1980a,b):

1. If problem involves an inclined plane then
   (a) expect something rolling or sliding up or down;
   (b) use F=MA;
   (c) use Newton's 3rd law.
2. If plane is smooth then use Conservation of Mechanical Energy.
3. If plane is not smooth then use work done by friction.
4. If problem involves objects connected by string and one object being pulled by the other then consider string tension.
5. If string is not taut then consider objects as independent.

The third experiment of Chi et al. (1981) complicates the interpretation of the first two experiments by giving credence to the possibility that for experts, the surface features define which particular deep structure principle(s) should be invoked.

Chi et al. (1981) did not notice this important contextuality in the knowledge of the experts who participated in their third experiment. Instead, Chi et al. (1981) highlighted the fact that experts were able to provide a rich, nuanced knowledge system for the incline plane problem type. Indeed, the experts' knowledge of incline planes was well elaborated enough to be able to be translated by the researchers into a formal, production rule form. On the other hand, the researchers were unable to provide such a translation of the novice articulation of the incline plane problem type. They used this as evidence that the novice knowledge system was less elaborated and nuanced then the expert knowledge system.

There is little doubt, given that experts perform better than novices in physics problem solving, that experts have both a larger and more nuanced knowledge system than novices. Yet, Chi et al. (1981) went further with their interpretation of this experiment: they claimed that the difference in the knowledge system between novices and experts was due in large part to the assumed differences in the types that novices or experts recognize in physics problems. Remember that Chi et al. (1981) assumed that the types that novices and experts perceive in problem solving track the categorizations they produced in their experiments. Given this assumption, Chi et al. (1981) was motivated in the third experiment to demonstrate differences between novice and expert knowledge, and then claim that this difference was due to differences in the type that novices and experts recognize — even though they never empirically demonstrated that such a large difference in perceived types even exists.

It is interesting that Chi et al. (1981) did not conduct an experiment in which novices and experts elaborated the knowledge they have about a generic deep structure type (like conservation of energy). They only demonstrated novice and expert elaborations of the incline plane surface feature type. As such, they missed the opportunity to test their claim that an elaborated, generic deep structure type contains productive problem solving knowledge. In fact, their third experiment demonstrates that expert knowledge of deep structure is *context-specific*, raising serious doubts that deep structure perception is independent of surface feature perception.

2.3 Inference from Surface Feature Types

Consistent with the perceptual chunking hypothesis, Chi et al. (1981) assume that an initial perceptual step in problem solving is to recognize a problem's type. They assert that these types are used to pick out chunks of knowledge that may be productive for solving the problem. Chi et al. (1981) step much further, however, by assuming a fundamental difference in recognized types between novices and experts: they assert that novices type problems using surface features while experts type problems using deep structure. Furthermore, they assert that much of the difference in knowledge between novices and experts is due to this assumed difference between novice and expert types. They conclude that because novices do not categorize physics problems according to the principles of physics, novices lack appropriate labels to index productive physics knowledge:

> The knowledge useful for a particular problem is indexed when a given physics problem is categorized as a specific type. Thus, expert-novice differences may be related to poorly formed, qualitatively different, or nonexistent categories in the novice representation.

Essentially, Chi et al. (1981) assume that surface feature typing of physics problems provides an inadequate label to index productive physics knowledge. Given the context-specificity of the deep structure knowledge that experts invoked in the third experiment, this is a peculiar view. One could imagine instead that surface feature types are the appropriate index used for labeling context-specific problem solving knowledge, like the context-specific applications of the principles of physics that the experts in Chi et al. (1981) invoked for the incline plane context. As such, both novices and experts may start their problem solving by using surface feature types to recall relevant chunks of knowledge.

One obvious reason why novices and experts might both use surface features to index deep structure is because the surface features of a physics problem generally constitute one of the few input signals available to make the inferences necessary to solve the problem. All necessary knowledge for solving a physics problem must be able to be inferred from the surface features, including the deep structure of that problem. Chi et al. (1981) recognize this possibility. In their fourth experiment, they presented novices and experts with some physics problems. For each problem, the researchers asked the novices and experts to list the “basic approach” they would use to solve the problem. They were also asked to expound upon the features of the problem statement that led them to infer the basic approach. According to Chi et al. (1981), experts justified their choice of the basic approach in many cases using “second-order” features — that is, using features not explicitly stated in the problem statement. They recognized that these second-order features must originate from surface features:

> [S]ince second-order features must necessarily be derived from more literal surface features that are in the problem statements, it is of interest to see whether the surface features in the problem statement that elicit these second-order features can be identified.

Since experts did not generally articulate the surface features they used in both this experiment and the two categorization experiments, Chi et al. (1981) could not demonstrate how deep structure was inferred from surface features. But this does not mean that experts did not use surface features to make inferences. Surface features are generally the sole input signal used to make inferences in problem solving, for both novices and experts.

Chi et al. (1981) demonstrate that expert knowledge does not simply contain abstract representations of deep structure principles. Experts have available to them context-specific knowledge of the principles of physics, which they can efficiently apply to familiar contexts. For the routine introductory mechanics problems presented by Chi et al. (1981), an expert's ability to apply these principles efficiently hinges partially on the fact that the expert has encountered those problems of that particular surface feature type before, and as such, has stored relevant context-specific knowledge that allows the expert to solve these problems quickly.

Therefore, I claim that the perception of surface feature types, for both novices and experts, provides important information for inferring relevant deep structure.[1] By perceiving a surface feature type for an arbitrary problem, physics cognition may be able to recall potentially relevant knowledge indexed by that surface feature type. Rather than viewing expert physics knowledge as characterized by chunks labeled with abstract deep structure, this view of the structure of physics knowledge hypothesizes that both novices and experts have important physics problem solving knowledge that can be accessed using surface feature types.

2.4 Adaptive Summarization

Since experts may be using surface feature types to make productive inferences, the skill of surface feature categorization by novices demonstrated by Chi et al. (1981) may in fact be a productive skill of normative physics practice. One of this dissertation's primary objectives is to provide a simple quantitative principle that governs this skill. I motivate the qualitative version of this principle in the current section.

While surface feature types index productive knowledge, they also fulfill another function: surface feature types summarize an environment of physics problems. As-

[1] Note that while I hypothesize that routine introductory physics problem solving may start with perception of a surface feature type, it is likely that experts also have more abstract knowledge of physics that allows them to reason productively about less familiar contexts — albeit not as efficiently as when they operate in routine contexts. Efficient problem solving also certainly requires domain independent knowledge that can be used in many contexts (Perkins and Salomon, 1989), such as the ability to manipulate mathematical symbols. Though physics itself may be codified by principles that are context independent, its efficient application should involve knowledge, indexed by perception, which is context dependent.

sume that the initial perceptual step in solving an arbitrary introductory physics problem x is to assign a type $\bar{x}$ to x. I will provide a model that predicts the type $\bar{x}$ for an arbitrary problem x taken from an environment of introductory physics problems X. Taken over an environment X, the set of predicted assignments $x \mapsto \bar{x}$ can be viewed as providing a summary of the environment X into a smaller set of types $\bar{X}$. Note that any arbitrary map $X \to \bar{X}$ provides a summary of the surface features used in that environment, if we let each $\bar{x}$ be denoted with a single label that describes the surface features for all the problems belonging to $\bar{x}$. Therefore, the skill of "typing" can be defined as the ability to assign a particular type $\bar{x}$ to each problem x in an environment of physics problems. The set of predicted responses functions to summarize this environment.

Which particular summary $X \to \bar{X}$ does physics cognition choose? The novices in Chi et al. (1981) who categorized problems according to surface features produced particular surface feature categories. Chi et al. (1981) neglected to provide a discussion for why particular surface feature categories were produced, and not others. They focused their discussion on expert-novice differences.

In this dissertation, I conjecture that knowledge required for typing is adapted to its environment such that it provides an *optimal summary* of the surface features used in that environment. Requiring that the surface feature types provide an optimal summary will constrain the particular surface features types provided by this model. This approach requires that I motivate an appropriate definition of optimal summary, which I discuss in the next chapter.

Note that I am formulating this conjecture using the adaptive view of knowledge that is motivated from constructivism, rather than using the alternative perspective that knowledge is constructed, at least partially, from a history of past experiences. The types that are ultimately produced by physics cognition are partly constructed from actual experiences in the classroom and through practice on particular homework problems; however, it is not necessary to understand how knowledge of surface feature types is constructed to develop a model of surface feature type perception — assuming that an appropriate definition of optimal summary can be motivated.

One benefit of this approach is that a semantics of physics problems, which are the surface feature types produced by the model, can be derived quantitatively in a simple, principled fashion from only the surface features present in the environment. I have in effect conjectured that a semantics derived from the environment can approximate the actual surface feature types produced by individuals.

I will therefore conjecture the following simple principle of physics practice:

> An important part of cognitive perception in physics functions to categorize an environment of physics problems into problem types, where the categorizations produced optimally summarize the surface features used in the environment of physics problems for which an individual is (or has been) responsible.

This is the qualitative version of the quantitative physical principle of physics cognition that I will develop in the next chapter.

2.5 The Difference between Experimental Categorizations and Perceived Types

I have argued that surface feature types provide labels that index productive problem solving knowledge in routine contexts, for both novices and experts. Yet, these types do not match the deep structure categorizations that experts produced in the categorization experiments of Chi et al. (1981). What might explain this difference?

Chi et al. (1981) instructed the individuals to group problems according to their "similarities of solution" — a more or less ambiguous prompt. Instead, imagine that individuals were asked to categorize the problems according to the "law of physics used to solve the problem." In such a modified experiment, the novices, all of whom took and presumably passed an introductory course in physics, may have categorized the problems based on the relevant normative principles like the conservation of energy. The fact that the individuals categorized problems according to an ambiguous prompt may have also contributed to the timing results presented by Chi et al. (1981). In their first categorization experiment, novices took less time to complete the categorization task than experts. Experts may have taken the extra time to select, from the set of context-specific deep structure knowledge elements indexed by the surface feature type, the precise physics principle most applicable to the given problem. Experts may have taken this extra step because they had a different interpretation of the categorization task than novices.

In addition, both experts and novices were not allowed to use pencil and paper to carry out the steps necessary to compare problems based on similarity of solution. It may certainly be the case that experts are able to consider the entire solution of a physics problem more easily than novices without the aid of pencil and paper. Since the solution to a physics problem requires the application of much more knowledge than simply the laws of physics pertaining to a given prompt (VanLehn et al., 2005), the novice might have a difficult time comparing the entire solution of two physics problems without working the two problems out on paper. Yet, if instructed to categorize based on the "law of physics used to solve the problem," the novices may have been completely capable of comparing only the relevant principle of physics of two problems without using pencil and paper. Furthermore, if experts do have the ability to compare the entire solution of two problems without pencil and paper, this does not necessarily mean that experts think in abstract, context-independent ways.

If the hypothesis that novices and experts both use surface feature classification to initially perceive a routine introductory physics problem is correct, it should be measurable in appropriately designed experiments. In the original categorization task (Chi et al., 1981), individuals were given unlimited time to categorize the problems. Suppose instead that both experts and novices were restricted by the researchers to categorize problems within a short time limit. It may be possible that in cases of restricted time that experts would also be observed to categorize with surface features.

2.6 Conclusion

In this chapter, I have argued that surface feature classification may be a general function performed by physics cognition that is potentially useful for inferring relevant knowledge in problem solving. To motivate this hypothesis, I critiqued an article published by Chi et al. (1981) that presented an experimental context in which novices categorized physics problems based on surface features. Rather than viewing surface feature categorization as indicative of a lack of novice knowledge vis-à-vis expert knowledge, I have argued that this skill may in fact be productive.

I have also argued that the difference in categorizations of physics problems exhibited empirically in the paper written by Chi et al. (1981) may not constitute a substantive difference between experts and novices. First, even though novices produce surface feature categorizations, this does not mean they are incapable of inferring deep structure. Surface feature types may be used to infer relevant information about deep structure. Second, the abstract deep structure categorizations that experts produce may not be adequate proxies for the actual types experts initially perceive in physics problem solving contexts. Chi et al. (1981) did not demonstrate that experts invoke deep structure in abstract ways when solving problems. Rather, they demonstrated a rich example of how expert knowledge of a familiar context includes knowledge of how deep structure may be used in that context.

I have implicitly used a heuristic for identifying simple functions of physics cognition: consider how any observed behavior might be productive, including behavior that might look at first glance to be non-normative. This heuristic derives from an optimistic orientation toward learning: if we believe that the average novice, with adequate support, can eventually attain expert status, this means that some aspects of novice cognition will be productive. I take the further step of assuming that certain aspects of novice skill persist as core skills that constitute expert performance, which means that studying novice physics cognition may provide a potential vehicle for identifying general principles of physics practice.

In this chapter, I have used the outline for quantitative physical modeling introduced in the previous chapter to motivate a conjecture concerning physics cognition. As I will show later, the process of translating the qualitative version of this conjecture into a quantitative one is direct, assuming an appropriate notion of an optimal summary can be provided. A single equation will codify the quantitative version of the conjecture.

Chapter 3

Quantitative Physical Model of Surface Feature Perception

In this chapter, I develop a quantitative version of the conjecture for surface feature perception posed in the previous chapter. In so doing, I closely follow the perspective provided by Bialek (2002) for the extraction of semantics. Using information theory, I motivate two competing quantitative notions of an optimal summary, and conjecture that physics cognition produces summaries that reconcile this tension. The model presented here requires that physics cognition be adapted to some particular environment of physics problems, and as such, may be consistent with constructivism.

3.1 Introduction

I assume that one of the first steps that an individual takes when confronting a physics problem is to assign a type to that problem. I assume that he or she does so in order to access relevant problem solving knowledge (see Chapter 2 for further discussion). Suppose further that when an arbitrary problem x is given to individuals who share particular expertise in physics (i.e. novices), different individuals within that class assign different types to this problem. Relative to some particular class of expertise, let $p(\bar{x}|x)$ be the probability that an arbitrarily chosen individual from that class assigns type $\bar{x}$ from a set of types $\bar{X}$ to problem x taken from some environment of physics problems X. The goal of this model is to be able to predict this probabilistic assignment.[1]

In this chapter, I use the information bottleneck method (Tishby et al., 1999; Bialek, 2002) to construct a model that predicts the probabilistic assignment of types by assuming that physics cognition is optimally adapted to a particular hypothetical random process. Let the random process be defined by elevating a particular environment X of physics problems to a random variable, with $p(x)$ being the probability that a particular random problem x is drawn from X.[2] Suppose that for each problem x that is drawn, a type $\bar{x}$ is assigned according to the same rules $p(\bar{x}|x)$ that physics cognition would use if it were to assign a type to the problem x in actuality.

It is important to note that the researcher need not actually create and conduct this random process. The probability measure $p(x)$ need only be a theoretical device used to construct the predictive model. It is helpful however to imagine a realization of this process, in which the researcher creates the rule $p(x)$, knows the rule $p(\bar{x}|x)$, and uses them both in tandem to construct and observe this random process. Visualizing this hypothetical experiment enables one to more easily follow the logic of the argument contained in this chapter and to develop intuition for the character of the model.

In this chapter, I place a probability distribution on an environment of physics problems because it enables the construction of fundamental quantitative measures

[1] I consider the individuals as simply "black boxes" which, when given problems, produce types. I do not attend to aspects of psychology or neuroscience that may provide the mechanism for why particular types are chosen. If this model ends up accurately predicting the probability rules for assigning types, one can later look for internal mechanisms to explain why the model is true (see Chapter 1 for discussion).

[2] The precise form for the probability measure $p(x)$ for the random variable X is left open in this discussion, though it needs to be chosen a-priori in order to make predictions with this model. Two obvious choices to consider, for instance, are whether physics cognition is adapted to a random variable X where $p(x)$ is uniform or to a different random variable X where $p(x)$ is proportional to the number of surface features present in the problem.

for the *information* that the types carry. The information-theoretic quantities that emerge provide a principled approach for deciding whether any particular probabilistic rule set $\{p(\bar{x}|x)\}$ is more or less optimal.

3.2 Detail Provided by the Summary

Information theory provides a principled way for defining parameters which quantify the degree to which the probabilistic assignment $p(\bar{x}|x)$ provided by physics cognition is optimally adapted to the random variable X. I will define these fundamental parameters by examining the amount of information that an arbitrary type $\bar{x}$, from the researcher's frame of reference, carries concerning the random problem drawn from X.

In information theory, the acquisition of information is seen to resolve uncertainty. The entropy $S(X)$ for a random process X provides a quantitative measure of the uncertainty associated with a single draw from the random process (Bialek, 2011):

$$S(X) \equiv -\sum_{x \in X} p(x) \log p(x) \tag{3.1}$$

According to information theory (Shannon, 1948), if a draw from the random variable X occurs such that a particular problem x is instantiated, the identity of the problem x can be understood as providing information $I(x)$, since the problem's identity eliminates the uncertainty that originally existed concerning the problem's identity. This information $I(x)$ carried by the instantiated problem x is defined to be equal to the amount of uncertainty that existed regarding the identity of problem x before x was drawn:

$$I(x) \equiv S(X) \tag{3.2}$$

One can view the assigned type $\bar{x}$, from the researcher's frame of reference, as also providing information about the identity of the problem instantiated from the random variable X. This supposes that the identity of the problem x remains hidden from the researcher's frame of reference, but only the type $\bar{x}$ of that problem is available. The type $\bar{x}$ of the unknown problem will generally reduce the researcher's uncertainty concerning the identity of the problem, but some degree of uncertainty $S(X|\bar{x})$ regarding the identity of the unknown problem will still remain. The entropy formula provided by Eq. (3.1) is used to quantify this uncertainty. The entropy formula in this case uses the conditional probability distribution $p(x|\bar{x})$, which provides the probability that a particular problem x was instantiated from the random variable X, given that the type $\bar{x}$ is known to the researcher:

$$S(X|\bar{x}) = -\sum_{x \in X} p(x|\bar{x}) \log p(x|\bar{x}) \tag{3.3}$$

Note that the conditional probability $p(x|\bar{x})$ can be determined from $p(\bar{x}|x)$ using Bayes' rule:

$$p(x|\bar{x}) = \frac{p(\bar{x}|x)p(x)}{p(\bar{x})} \tag{3.4}$$

where

$$p(\bar{x}) = \sum_{x} p(\bar{x}|x)p(x) \tag{3.5}$$

If $\bar{x}$ is informative, it should decrease the entropy concerning the identity of the unknown, but instantiated, problem. The amount of information $I(X|\bar{x})$ carried by the type $\bar{x}$ about the identity of the unknown problem, from the frame of reference of the researcher, is equal to this decrease in entropy:

$$I(X|\bar{x}) \equiv S(X) - S(X|\bar{x}) \tag{3.6}$$

The average information $I(X;\bar{X})$[3] provided by an arbitrary type in $\bar{X}$ concerning

[3]The semicolon notation is used for the average information $I(X;\bar{X})$ because it turns out that for

a random draw from X is:

$$\begin{aligned} I(X;\bar{X}) &= \sum_{\bar{x}} p(\bar{x})\left[S(X) - S(X|\bar{x})\right] \\ &= S(X) - \sum_{\bar{x}} p(\bar{x}) S(X|\bar{x}) \\ &\equiv S(X) - S(X|\bar{X}) \end{aligned} \tag{3.7}$$

where the notation $S(X|\bar{X})$ denotes a *conditional entropy*, which is the average entropy concerning the identity of the randomly drawn problem from X after conditioning on a type in $\bar{X}$. Bialek (2002) labels the information $I(X;\bar{X})$ as the *detail* provided by the summary.

Note that the detail in the summary $\bar{X}$ is in general less than the complete information provided by the problem itself. Types therefore summarize the random variable X of physics problems through lossy compression. The detail carried by the average type is not sufficient to identify the problem to which the type was assigned.

So far, I have not provided a quantitative conjecture for which particular summaries may be supplied by physics cognition. In the previous chapter I suggested that the types produced by physics cognition may optimally summarize the environment of physics problems, but I did not elaborate on how one might define an optimal summary. In this chapter, I conjecture, following Bialek (2002), that the summaries produced by physics cognition result from a reconciled tension between two different principles of optimality. For the first principle:

> Summaries $\bar{X}$ are more optimal to the extent that they compress the outcome of the random process defined as a draw of a problem from the

any two arbitrary random variables A and B, the average information $I(A;B)$ that an instantiation of the random variable A provides about the random variable B is equal to the average information that an instantiation of the random variable B provides about the random variable A. I will demonstrate this symmetry later in this chapter, and I emphasize this symmetry here by using the semicolon notation.

random variable X. Stated quantitatively, summaries are more optimal to the extent that the detail $I(X; \bar{X})$ is low.

This principle formalizes the function of physics cognition that I discussed in the last chapter: the types $\bar{X}$ act to provide summaries of an environment of physics problems. As summaries, the types should not provide identifying information; the less information they provide about problem identities, the more they summarize. Yet, I have not discussed which of the many possible summaries $\bar{X}$ that compress X may ultimately be produced by physics cognition. One natural hypothesis is that the instantiated summaries simultaneously function to summarize physics problems, while at the same time, function to provide relevant information for problem solving. Yet, as I will demonstrate later in this chapter, the detail $I(X; \bar{X})$ provided by the summaries places an upper bound on the amount of relevant information that $\bar{X}$ may provide about X. This means that acting to compress X with a summary $\bar{X}$, while more optimal with respect to the above principle, may be counterproductive with regard to providing relevant information. In the next section, I introduce a form of information provided by any summary $\bar{X}$ that physics cognition may act to maintain in any instantiated summary.

3.3 Chi et al. (1981) Revisited

In the previous chapter, I introduced the idea that the surface features of physics problems may be used to infer types that cue relevant context-specific problem solving knowledge. To build a quantitative model for the recognition of types using surface features, I first need to define the meaning of the term "surface features." Chi et al. (1981) provide a description for surface features:

> To reiterate, the novices' use of surface features may involve either key-words given in the problem statement or abstracted visual configurations,

> that is, the presence of identical keywords (such as friction) is one criterion by which novices group problems as similar. Yet, novices were also capable of going beyond the word level to classify by types of physical objects. For example, "merry-go-round" and "rotating disk" are classified as the same object....

Chi et al. (1981) therefore state that novices both categorize problems based on the literal words present in problems, but also on the basis of abstractions of the objects present in the problems. By noting that novices list a merry-go-round and a rotating disk as belonging to the same category, Chi et al. (1981) implied that novices perceive the "rotating" aspect of either of these two objects as the relevant feature to maintain in the abstraction.

Yet, it is not clear from the paper written by Chi et al. (1981), when novices are presented with problems that do not involve rotating objects, what particular abstraction of those objects novices may or may not construct. To give another example, consider two problems[4] taken from the "Motion in One Dimension" chapter in Serway and Jewett (2004):

> Problem 44: Emily challenges her friend David to catch a dollar bill as follows. She holds the bill vertically, as in Figure P2.44, with the center of the bill between David's index finger and thumb. David must catch the bill after Emily releases it without moving his hand downward. If his reaction time is 0.2 s, will he succeed? Explain your reasoning.

> Problem 45: In Mostar, Bosnia, the ultimate test of a young man's courage once was to jump off a 400-year-old bridge (now destroyed) into the River Neretva, 23.0m below the bridge. (a) How long did the jump last? (b) How fast was the diver traveling upon impact with the water?

[4] I removed part (c) from problem 45, which involves the speed of sound, to focus on the common context of objects moving in a uniform gravitational field.

Note that a relevant shared "abstracted visual configuration" that individuals may be able to identify in the two problems is that they both can be treated as a *point mass falling*. To focus attention on this particular abstraction presumably requires the use of prior knowledge, whether acquired in the classroom or through experience in the world. This knowledge is needed in order to neglect other features of the problems; for example, that the age of the bridge is irrelevant, or that the location where the context (like Bosnia) occurs is irrelevant, or even that the constitution of the object (like the fact that the object is a dollar bill) under consideration is also irrelevant for the purposes of solving this problem.

It may be difficult for a quantitative model that groups problems based on the literal words present in the problems alone to identify an appropriate abstract context directly from those problems. This is because potentially anomalous surface feature types may be produced from irrelevant correlations in the statistics of the words present in that finite sample. Yet, if the problems are written in an abstract fashion to begin with, this potential issue may be less important. Consider the following problem from the textbook of introductory physics problems written by Snyder and Palmer (1900):

> Energy Chapter, Problem 19: A body weighing 20 g. has a kinetic energy of 1000 ergs. How far would it ascend vertically?

This problem is written in a more terse and abstract fashion than the two problems I provided from Serway and Jewett (2004). In general, the problems provided by Snyder and Palmer (1900) are written more tersely and abstractly than many of the problems written in modern textbooks of introductory physics.[5]

There are at least two other important differences between this problem and the two other problems given above. First, this problem explicitly refers to the "deep

[5] I hope that future work may be able to provide objective criteria, rather than appealing to admittedly vague notions of terseness and abstractness, for identifying whether an arbitrary set of problems is a potentially appropriate environment for modeling physics problem perception.

structure" of energy, while the two other problems do not. Second, the object presented in this problem is ascending rather than descending. Nevertheless, even with these two differences, all three of these problems share a *relevant* abstracted context. They can all be viewed productively as *a point mass in a uniform gravitational field.* The recognition of this relevant abstract context is essential relevant information for solving any one of these three problems. For example, in the problem from Snyder and Palmer (1900), even though the use of energy considerations is suggested, the precise context-specific potential energy form of $U = mgh$ cannot be determined without perception of this relevant abstracted context.

In this dissertation, I will consider the case in which physics cognition is asked to confront the problems of Snyder and Palmer (1900), and whether summarization of the words present in this environment provides relevant information for use in problem solving. For simplicity and objectivity, I will consider every word in every problem as a surface feature in this analysis, but I will ignore all numeric characters.[6]

3.4 Relevant Information Provided by the Summary

Consider the aforementioned process in which a problem is drawn randomly from X. Suppose that an additional form of randomness is added to this hypothetical process after an arbitrary problem x is drawn: suppose a surface feature w is drawn at random from the problem. If each surface feature w in x has frequency $N_x(w)$, let

$$p(w|x) \equiv N_x(w) / \sum_{w} N_x(w) \tag{3.8}$$

[6]I do not weight particular words as more important than others in order to not introduce subjectivity into the model. More details on precisely how I treat the text of Snyder and Palmer (1900) is included in Chapter 4 and the Appendix.

provide the conditional probability of drawing the particular surface feature w from the problem x.

Before the problem is drawn, the probability that a particular surface feature w will be instantiated after a single draw of both a problem and a word is

$$p(w) = \sum_x p(w|x)p(x) \tag{3.9}$$

and the uncertainty over which surface feature will be selected from the environment of surface features W present in X is

$$S(W) = -\sum_w p(w) \log p(w) \tag{3.10}$$

Assuming only the problem x is drawn and its identity known to the researcher, but where the random surface feature of the problem is yet to be drawn, the uncertainty concerning which surface feature will be instantiated is generally smaller than the original entropy $S(W)$. It is provided by:

$$S(W|x) = -\sum_w p(w|x) \log p(w|x) \tag{3.11}$$

I label the difference between these two entropies, following Bialek (2002), as the *relevant information* carried by x. I label this information as relevant because it concerns the surface features: surface features, as discussed extensively in the previous chapter, may index relevant knowledge for problem solving. The relevant information is given by:

$$\begin{aligned} I(W|x) &= S(W) - S(W|x) \\ &= -\sum_w p(w) \log p(w) - \left(-\sum_w p(w|x) \log p(w|x)\right) \end{aligned} \tag{3.12}$$

The average relevant information $I(W;X)$ that an arbitrary problem in X provides is given by:

$$
\begin{aligned}
I(W;X) &= \sum_x p(x)\left[S(W) - S(W|x)\right] \\
&= S(W) - \sum_x p(x) S(W|x) \\
&\equiv S(W) - S(W|X)
\end{aligned}
\tag{3.13}
$$

To quantify the relevant information carried by $\bar{x}$, consider again the scenario in which only the type $\bar{x}$ for a randomly drawn problem is made available to the researcher, but not the identity of the problem. The type $\bar{x}$, in addition to carrying information concerning the identity of the unknown problem, also carries relevant information concerning which surface feature will be drawn from that problem. The uncertainty concerning the surface feature that will be drawn from the unknown problem with known type $\bar{x}$ is:

$$
S(W|\bar{x}) = -\sum_w p(w|\bar{x}) \log p(w|\bar{x}) \tag{3.14}
$$

where the conditional probability $p(w|\bar{x})$ is calculated by summing over all the problems in X:

$$
p(w|\bar{x}) = \sum_x p(w|x)p(x|\bar{x}) = \frac{1}{p(\bar{x})} \sum_x p(w|x)p(\bar{x}|x)p(x) \tag{3.15}
$$

The relevant information carried by $\bar{x}$ concerning which surface features may be drawn from the unknown problem is the decrease in entropy that occurs due to conditioning on $\bar{x}$:

$$
I(W|\bar{x}) = S(W) - S(W|\bar{x}) \tag{3.16}
$$

The average relevant information that an arbitrary type in $\bar{X}$ provides is therefore:

$$I(W;\bar{X}) = S(W) - \sum_{\bar{x}} p(\bar{x}) S(W|\bar{x}) \equiv S(W) - S(W|\bar{X}) \tag{3.17}$$

Note that both the relevant information provided by Eq. (3.17) and the detail provided by Eq. (3.7) take the same form.

3.5 Relevant Information/Detail Inequality

In this section, I demonstrate that the detail $I(X;\bar{X})$ that the summary $\bar{X}$ provides places an upper bound on the amount of relevant information $I(W;\bar{X})$ supplied by that summary. To start, I will first show that the average information $I(A;B)$ that any arbitrary random variable B provides about the random variable A is symmetric in the two random variables. Following Bialek (2011):

$$\begin{aligned}
I(A;B) &= \sum_b P(b)\left[S(A) - S(A|b)\right] && (3.18)\\
&= S(A) - \sum_b P(b) S(A|b) \\
&= -\sum_a P(a)\log P(a) - \sum_b P(b) S(a|b) \\
&= -\sum_{ab} P(a|b)P(b)\log P(a) - \sum_b P(b)\left[\sum_a -P(a|b)\log P(a|b)\right] \\
&= -\sum_{ab} P(a|b)P(b)\log P(a) + \sum_{ab} P(b)P(a|b)\log P(a|b) \\
&= \sum_{ab} P(a|b)P(b)\log\frac{P(a|b)}{P(a)} && (3.19)\\
&= \sum_{ab} P(a,b)\log\frac{P(a,b)}{P(a)P(b)} && (3.20)
\end{aligned}$$

Eq. (3.20) makes apparent that the "mutual information" $I(A;B)$ is symmetric in the random variables A and B. Furthermore, Eq. (3.20) demonstrates that the mutual information is simply the expectation of the random variable $\log\left(\frac{p(A,B)}{p(A)p(B)}\right)$. Note that both the detail $I(X;\bar{X})$ in Eq. (3.7) and the relevant information $I(W;\bar{X})$ in Eq. (3.17) take the mutual information form.

Next, I provide an alternate form for the conditional entropy $S(A|B)$:

$$\begin{aligned} S(A|B) &\equiv -\sum_b p(b) \sum_a p(a|b) \log p(a|b) \\ &= -\sum_{ab} p(a,b) \log p(a|b) \end{aligned} \tag{3.21}$$

Eq. (3.21) demonstrates that the conditional entropy $S(A|B)$ is the expectation of the distribution $\log \frac{1}{p(a|b)}$. Similarly, the conditional entropy $S(A|BC)$ formed by conditioning on a third random variable C is given by:

$$\begin{aligned} S(A|BC) &\equiv -\sum_{bc} p(b,c) \sum_a p(a|b,c) \log p(a|b,c) \\ &= -\sum_{abc} p(a,b,c) \log p(a|b,c) \end{aligned} \tag{3.22}$$

Taking the difference of Eq. (3.21) and Eq. (3.22):

$$\begin{aligned} S(A|B) - S(A|BC) &= -\left(\sum_{ab} p(a,b) \log p(a|b) - \sum_{abc} p(a,b,c) \log p(a|b,c)\right) \\ &= -\left(\sum_{abc} p(a,b,c) \log p(a|b) - \sum_{abc} p(a,b,c) \log p(a|b,c)\right) \\ &= \sum_{abc} p(a,b,c) \log \frac{p(a|bc)}{p(a|b)} \\ &= \sum_{abc} p(a,b,c) \log \frac{p(a,b,c)}{p(a|b)p(b,c)} \end{aligned} \tag{3.23}$$

Note here that $S(A|B) - S(A|BC)$ is the expectation of the random variable $\log\left(\frac{p(A,B,C)}{p(A|B)p(B,C)}\right)$ I will now show that this difference is always non-negative, which will provide a demonstration that conditioning entropy on an additional variable never increases the entropy. To prove this, I will use Jensen's Inequality:

> For a strictly convex function f and random variable X, the expectation of the random variable $f(X)$ is greater than or equal to the function f acting on the expectation of the random variable X — that is, $\mathcal{E}(f(X)) \geq f(\mathcal{E}(X))$, where equality only applies when X is a constant random variable.

A physical reason that explains why this inequality is true is provided by MacKay (2003):

> If a collection of masses p_i are placed on a convex curve $f(x)$ at locations $(x_i, f(x_i))$, then the centre of gravity of those masses, which is at $(\mathcal{E}[x], \mathcal{E}[f(x)])$, lies above the curve.

Since the logarithm is a concave function, I can apply Jensen's inequality in reverse to Eq. (3.23) (Cover and Thomas, 2006):

$$
\begin{aligned}
S(A|BC) - S(A|B) &= \sum_{abc} p(a,b,c) \log \frac{p(a|b)p(b,c)}{p(a,b,c)} \\
&= \mathcal{E}\left(\log \frac{p(a|b)p(b,c)}{p(a,b,c)}\right) \\
&\leq \log \mathcal{E}\left(\frac{p(a|b)p(b,c)}{p(a,b,c)}\right) \\
&= \log \sum_{abc} p(a,b,c) \frac{p(a|b)p(b,c)}{p(a,b,c)} \\
&= \log \sum_{abc} p(a|b)p(b,c) \\
&= \log \sum_{abc} p(a|b)p(b)p(c|b) \\
&= \log \sum_{ab} p(a,b) \sum_{c} p(c|b) \\
&= \log 1 = 0
\end{aligned} \tag{3.24}
$$

which demonstrates that $S(A|BC) \leq S(A|B)$.

These results enable placement of an upper bound on the relevant information $I(W;\bar{X})$ (Roudi and Latham, 2009):

$$\begin{aligned} I(W;\bar{X}) &= S(\bar{X}) - S(\bar{X}|W) && \text{Definition of Mutual Information} \\ &\leq S(\bar{X}) - S(\bar{X}|W,X) && \text{From Eq. (3.24)} \\ &= S(\bar{X}) - S(\bar{X}|X) && \text{Since } p(\bar{x}|w,x) = p(\bar{x}|x) \\ &= I(X;\bar{X}) && \text{Definition of Mutual Information} \end{aligned} \tag{3.25}$$

Note that $p(\bar{x}|x) = p(\bar{x}|w,x)$ holds because the assignment rules are not assumed to depend upon the particular realization of the surface feature in the hypothetical random process. This result holds if W is replaced with any random variable Y derived only from X, assuming the assignment of $\bar{x}$ depends only on the identity of x. Note further that an argument similar to that used in (3.24) applies for the mutual information in Eq. (3.20), resulting in the following inequality:

$$I(A;B) \geq 0 \tag{3.26}$$

The equality applies if and only if A and B are independent random variables — that is, when $p(a,b) = p(a)p(b)$ for all a and b. Inequalities (3.26) and (3.25) together provide upper and lower bounds on the relevant information:

$$0 \leq I(W;\bar{X}) \leq I(X;\bar{X}) \tag{3.27}$$

Inequality (3.27) formalizes the intuition that $\bar{X}$ cannot provide any more information about W than $\bar{X}$ provides about X. This means that $\bar{X}$ should provide at least some information about X, since this is the only way that $\bar{X}$ can provide information about the relevant variable W.

3.6 The Information Bottleneck Principle

I argued in the previous chapter that physics cognition may produce types using surface features. In this chapter, I have demonstrated that when physics cognition interacts with its environment through a hypothetical random process, it can provide information concerning surface features used in that environment. I have also argued that physics cognition may tend toward instantiated summaries $\bar{X}$ in which the detail $I(X;\bar{X})$ that the types provide about the random draw from the environment X is low. Yet, from (3.27), the detail bounds from above the amount of relevant information $I(W;\bar{X})$ the summaries provide. Therefore, I will assert the following competing effect of physics cognition on the random process:

> Perception of types in physics cognition is characterized by summaries $\bar{X}$ which tend toward optimality by providing as much information about the outcome of the random draw of a surface feature from a problem as possible. That is, summaries are more optimal to the extent that the relevant information $I(W;\bar{X})$ is high.

Therefore, the effect of the summaries provided by physics cognition on this hypothetical random process is characterized by two competing factors: first, to provide as little information about the identity of the problem drawn as possible; while second, to provide as much information about the draw of a random surface feature from the problem as possible. Since decreasing the information provided about the identity of the problem will also decrease the information about the surface features of that problem, due to (3.27), these two notions of optimality compete. To reconcile these competing notions, consider the relevant information $I(W;\bar{X})$ as a "benefit" of the summary, treat the $I(X;\bar{X})$ as a "cost" of the summary, and maximize an appropriate weighted difference. This enables a final iteration of this conjecture of physics practice, as a special form of the "information bottleneck" principle described by Bialek (2002):

> Physics cognition produces summaries that are optimal in the sense that they maximize $-\mathcal{F} = I(W;\bar{X}) - TI(X;\bar{X})$, or equivalently minimize the free energy $\mathcal{F}$, where the "temperature" T measures how much weight to associate to the "cost" $I(X;\bar{X})$ relative to the "benefit" $I(W;\bar{X})$.

Different choices for the parameter T imply differences in the particular summaries physics cognition may produce. The choice for the parameter T can be interpreted as a choice of "temperature," given the form of $\mathcal{F}$ as a free energy, and because the necessary conditions that will be demonstrated in Eq. (3.38) for the probabilistic mappings $p(\bar{x}|x)$ take the form of a Boltzmann distribution.

3.7 Necessary Conditions for Solutions to the Information Bottleneck Principle

In this section, I derive a set of necessary conditions for the probabilistic assignment $p(\bar{x}|x)$, following Tishby et al. (1999).

First, remember that the probability functions $p(\bar{x}|x)$ need to be coherent:

$$\forall x : \quad \sum_{\bar{x}} p(\bar{x}|x) = 1 \tag{3.28}$$

I introduce $|X|$ additional Lagrange multipliers $\lambda(x)$ for each of the $|X|$ conditions given in Eq. (3.28), so that the information bottleneck principle becomes equivalent to maximizing the following function:

$$\mathcal{L} \equiv I(W;\bar{X}) - TI(X;\bar{X}) - \sum_{x} \lambda(x) \sum_{\bar{x}} p(\bar{x}|x) \tag{3.29}$$

The two mutual information quantities $I(W;\bar{X})$ and $I(X;\bar{X})$ can be written in terms of $p(\bar{x}|x)$ as well as

$$p(\bar{x}) = \sum_x p(\bar{x}|x)p(x) \tag{3.30}$$

$$p(\bar{x}|w) = \sum_x p(\bar{x}|x)p(x|w) \tag{3.31}$$

which depend only on $p(\bar{x}|x)$ and the nonvarying distributions $p(x)$ and $p(x|w)$. Now expand Eq. (3.29):

$$\mathcal{L} = \sum_{w\bar{x}} P(\bar{x}|w)P(w)\log\frac{P(\bar{x}|w)}{P(\bar{x})} - T\sum_{x\bar{x}} P(\bar{x}|x)P(x)\log\frac{P(\bar{x}|x)}{P(\bar{x})} - \sum_x \lambda(x)\sum_{\bar{x}} p(\bar{x}|x) \tag{3.32}$$

By differentiating with respect to $p(\bar{x}|x)$, using Eq. (3.30) and Eq. (3.31), the first term becomes:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}_1}{\mathrm{d}p(\bar{x}|x)} &= \sum_w \frac{\mathrm{d}p(\bar{x}|w)}{\mathrm{d}p(\bar{x}|x)} P(w)\log P(\bar{x}|w) + \sum_w P(w)\frac{\mathrm{d}p(\bar{x}|w)}{\mathrm{d}p(\bar{x}|x)} \\
&- \sum_w \frac{\mathrm{d}p(\bar{x}|w)}{\mathrm{d}p(\bar{x}|x)} P(w)\log P(\bar{x}) - \sum_w \frac{P(\bar{x}|w)P(w)}{P(\bar{x})}\frac{\mathrm{d}p(\bar{x})}{\mathrm{d}p(\bar{x}|x)} \\
&= \sum_w p(x|w)P(w)\log P(\bar{x}|w) + \sum_w P(w)p(x|w) \\
&- \sum_w P(x|w)P(w)\log P(\bar{x}) - \sum_w \frac{P(\bar{x}|w)P(w)}{P(\bar{x})}P(x) \\
&= \sum_w p(x|w)p(w)\log\frac{p(\bar{x}|w)}{p(\bar{x})} \\
&= \sum_w p(x,w)\log\frac{p(\bar{x},w)}{p(\bar{x})p(w)}
\end{aligned} \tag{3.33}$$

The second term is evaluated in similar fashion. The derivative of $\mathcal{L}$ then becomes:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}p(\bar{x}|x)} &= \sum_w p(x,w)\log\frac{p(\bar{x},w)}{p(\bar{x})p(w)} - Tp(x)\log\frac{p(\bar{x}|x)}{p(\bar{x})} - \lambda(x) \\
&= p(x)\left(\sum_w p(w|x)\log\frac{p(w|\bar{x})}{p(w)} - T\log\frac{p(\bar{x}|x)}{p(\bar{x})} - \frac{\lambda(x)}{p(x)}\right)
\end{aligned} \tag{3.34}$$

In order to write this in a slightly more suggestive form, rewrite the first term:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}p(\bar{x}|x)} &= p(x)\left(\sum_w p(w|x)\log\frac{p(w|\bar{x})p(w|x)}{p(w)p(w|x)} - T\log\frac{p(\bar{x}|x)}{p(\bar{x})} - \frac{\lambda(x)}{p(x)}\right) \\ &= p(x)\left(\sum_w p(w|x)\log\frac{p(w|\bar{x})}{p(w|x)} - T\log\frac{p(\bar{x}|x)}{p(\bar{x})}\right. \\ &\qquad\left. + \sum_w p(w|x)\log\frac{p(w|x)}{p(w)} - \frac{\lambda(x)}{p(x)}\right) \end{aligned} \tag{3.35}$$

Setting this to 0, and rearranging, we obtain:

$$\sum_w p(w|x)\log\frac{p(w|\bar{x})}{p(w|x)} + \sum_w p(w|x)\log\frac{p(w|x)}{p(w)} - \frac{\lambda(x)}{p(x)} = T\log\frac{p(\bar{x}|x)}{p(\bar{x})} \tag{3.36}$$

which implies:

$$p(\bar{x}|x) = p(\bar{x})\exp\left(-\frac{1}{T}\sum_w p(w|x)\log\frac{p(w|x)}{p(w|\bar{x})}\right)\exp\left(\frac{1}{T}\left[\log\frac{p(w|x)}{p(w)} - \frac{\lambda(x)}{p(x)}\right]\right) \tag{3.37}$$

Note that the second exponential factor depends only on x, so the entire term can simply be treated as a normalization factor $Z(x,T)$. Therefore,

$$p(\bar{x}|x) = \frac{p(\bar{x})}{Z(x,T)}\exp\left(-\frac{1}{T}\sum_w p(w|x)\log\frac{p(w|x)}{p(w|\bar{x})}\right) \tag{3.38}$$

where

$$Z(x,T) = \sum_{\bar{x}} p(\bar{x})\exp\left(-\frac{1}{T}\sum_w p(w|x)\log\frac{p(w|x)}{p(w|\bar{x})}\right) \tag{3.39}$$

The formal solutions provided by Eq. (3.38) to the information bottleneck principle are not closed form solutions for the mappings $p(\bar{x}|x)$, since $p(w|\bar{x})$ depends on $p(\bar{x}|x)$ through Eq. (3.15). Instead, they constitute a set of necessary conditions for the mappings $p(\bar{x}|x)$. Furthermore, note that Eq. (3.38) takes the form of a Boltzmann distribution, with the "energy" taking the form of a Kullback-Leibler divergence $D_{KL}(p||q)$ (MacKay, 2003):

$$D_{KL}(p||q) \equiv \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{3.40}$$

so that the "energy" is:

$$E(\bar{x}, x) = D_{KL}(p(w|x)||p(w|\bar{x})) \tag{3.41}$$

The Kullback-Leibler divergence provides a measure of "distance" between the description $p(w|x)$ of the surface features used in x and the summary $p(w|\bar{x})$ of the surface features provided by $\bar{x}$.

Note some important properties of the solutions provided by Eq. (3.38), as argued in a similar context by Bialek (2011). First, if two problems have similar distributions $p(w|x)$, Eq. (3.38) states that the model for physics cognition presented here will tend to map these two problems into the same $\bar{x}$. This is the approach physics cognition takes toward summary, if physics cognition acts to minimize the free energy $\mathcal{F}$: while compressing its environment, physics cognition may try to retain relevant information contained in the environment by grouping problems that have similar surface feature descriptions $p(w|x)$. Second, if physics cognition ultimately produces summaries parameterized by a low temperature T, for those problems x that are mapped significantly onto some particular $\bar{x}$, the "distance" between the summary $p(w|\bar{x})$ and the description $p(w|x)$ must be small. This means that if the temperature T is low, the summaries provided by the types $\bar{X}$ do not highly summarize the surface features in X. The relevant information $I(W;\bar{X})$ and the detail $I(X;\bar{X})$ of the instantiated summary will both tend to be high in this case. If the temperature is high however, the reverse is true: $\bar{x}$ can provide summaries $p(w|\bar{x})$ which deviate significantly from the descriptions $p(w|x)$ provided by the problems which map significantly onto $\bar{x}$.

Bialek (2002) also notes that Eq. (3.38) can also be viewed as providing a set of necessary conditions for the constrained optimization problem to maximize the relevant information $I(W;\bar{X})$, under the constraint of fixed detail $I(X;\bar{X})$: by the method of Lagrange multipliers, the stationary points of the free energy $\mathcal{F}$ provide the set of possible points for which the relevant information $I(W;\bar{X})$ may be maximized, at some fixed detail $I(X;\bar{X})$. From this perspective, T is a Lagrange multiplier which implements the constraint on the detail. If the constraint is relaxed slightly, then the standard interpretation of a Lagrange multiplier implies that the small increase in relevant information $\delta I(W;\bar{X})$ one gains if one allows a small increase in the detail $\delta I(X;\bar{X})$ provided by the summary is given by:

$$\frac{\delta I(W;\bar{X})}{\delta I(X;\bar{X})} = T \tag{3.42}$$

3.8 Conclusion

In the previous chapters, I considered the possibility that physics cognition may be optimally adapted to an environment of potential experiences; however, I did not provide a particular quantitative orientation for constructing such a model. In this chapter, I have motivated a model of type perception of physics problems by considering how physics cognition affects a hypothetical random process, where an environment of potential experiences, in the form of physics problems, provides the bag of objects for the random process.

Following Bialek (2002), I demonstrated that a particular, fundamental effect that physics cognition may have on this hypothetical random process is to provide two forms of information about this process: the "relevant information" and the "detail." I argued that physics cognition weighs maximizing relevant information against minimizing the detail and, as such, chooses summaries that minimize a "free energy." The stationary points of the free energy can also be interpreted as providing possible optimal mappings defined as where the relevant information $I(W;\bar{X})$ is maximized along a particular level set where the detail $I(X;\bar{X})$ is fixed at some value.

It is important to note the limitations of this model. First, the model in this chapter does not predict the language that individuals will use when assigning types to physics problems. It simply predicts the probability of assignment of problems to unlabeled types. It is up to the experimenter to design an appropriate coding scheme for the types that individuals assign. I believe that if the experiment is designed such that the time allotted for typing any given problem is kept short, the experiment may reveal that both novices and experts assign surface feature types (see Chapter 2 for further discussion). Second, an environment of physics problems needs to be chosen a-priori in order for predictions to be made: the model does not provide a rationale for why knowledge would be adapted to any particular environment. Principled choices for the environment exist. For example, to provide predictions for the types individuals assign to problems in introductory mechanics, all of the problems from the mechanics chapters of an introductory textbook may constitute an appropriate environment.

In this chapter, I have emphasized the potential productive capacity of novices by using the surface feature typing behavior attributed to novice cognition in order to propose a general principle of physics practice that may characterize physics cognition at all levels of expertise. In the next chapter, I will demonstrate that the words in a particular textbook of introductory physics problems provides relevant information for problem solving, and I will argue that both novices and experts may have access to this information.

Chapter 4

Surface Feature Types of Introductory Physics

In this chapter, I create deterministic surface feature types of introductory physics using a textbook of introductory physics problems. Following Slonim and Tishby (2000), these types will be constructed using an algorithm motivated from the information bottleneck principle. I examine these types and use them to argue that surface feature perception may provide relevant information for problem solving. Finally, I discuss why the information bottleneck principle, applied to this textbook, describes a fundamental feature of physics as practice.

4.1 The Agglomerative Information Bottleneck Algorithm

The information bottleneck conjecture asserts that the types that physics cognition provides for particular environments X of physics problems can be predicted by minimizing a "free energy" associated with a particular random process constructed from that environment. The temperature T of the free energy parameterizes different ways physics cognition may provide types for a particular environment. Unfortunately, closed form solutions to the information bottleneck principle are not currently known — except for the limit $T \to 0$, where minimizing the free energy becomes simply equivalent to maximizing the relevant information $I(W; \bar{X})$. The relevant information $I(W; \bar{X})$ is trivially maximized when $|\bar{X}| = |X|$, where each $x \in X$ is mapped uniquely and deterministically to a unique $\bar{x} \in \bar{X}$.

To construct non-trivial surface feature types, I utilize an algorithm described by Slonim and Tishby (2000) that is motivated from the $T \to 0$ limit to the information bottleneck principle. Starting from the the trivial $|\bar{X}| = |X|$ solution, choose a pair of problems to merge that minimize the decrease in relevant information. This results in a categorization with $|\bar{X}| = |X| - 1$ members. This process iterates, where after each additional merge a new categorization is produced that has one less element than the categorization that preceded it. The result is a hierarchy of deterministic categorizations, one for each cardinality $|\bar{X}|$ such that $1 \leq |\bar{X}| \leq |X|$. This algorithm is referred to by Slonim and Tishby (2000) as the agglomerative information bottleneck method.

Note that a summary $\bar{X}$ is defined as deterministic if:

$$
\begin{aligned}
p(\bar{x}|x) &= 1 \qquad x \in \bar{x} \\
p(\bar{x}|x) &= 0 \qquad x \notin \bar{x}
\end{aligned}
\tag{4.1}
$$

Even though the typing of the environment X provided by the agglomerative information bottleneck method is deterministic, it provides these types, just as in the standard information bottleneck method, using a random process parameterized by $p(x)$. The probability that a randomly drawn problem belongs to category $\bar{x}$, using Eq. (4.1), reduces from Eq. (3.5) to simply summing over the probabilities of drawing each problem that belongs to $\bar{x}$:

$$p(\bar{x}) = \sum_{x} p(\bar{x}|x)p(x) = \sum_{x\in\bar{x}} p(x) \tag{4.2}$$

The conditional distributions $p(w|\bar{x})$ in Eq. (3.15) that provide summaries of the words present in $\bar{x}$ reduces to:

$$\begin{aligned} p(w|\bar{x}) &= \frac{1}{p(\bar{x})}\sum_{x} p(w|x)p(\bar{x}|x)p(x) \\ &= \sum_{x\in\bar{x}} \frac{p(x)}{p(\bar{x})} p(w|x) \end{aligned} \tag{4.3}$$

That is, the summary $p(w|\bar{x})$ of the surface features used by the problems in $\bar{x}$ is calculated by computing a weighted average of the surface feature descriptions $p(w|x)$ over all the problems which belong to $\bar{x}$.

Slonim and Tishby (2000) recognized that the pair in a given categorization that merges to produce a categorization with one less member is the pair whose summaries $p(w|\bar{x})$ are "closest" together, where the "distance" between the summaries emerges directly from the principle to locally minimize the loss in relevant information. To see why, consider merging two categories $\bar{x}_i$ and $\bar{x}_j$ to form a new category $\bar{x}_k$. The summary of the surface features $p(w|\bar{x}_k)$ provided by $\bar{x}_k$ is from Eq. (4.3):

$$\begin{aligned} p(w|\bar{x}_k) &= \sum_{x\in\bar{x}_k} \frac{p(x)}{p(\bar{x}_k)} p(w|x) \\ &= \frac{1}{p(\bar{x}_k)}\left(\sum_{x\in\bar{x}_i} p(x)p(w|x) + \sum_{x\in\bar{x}_j} p(x)p(w|x)\right) \\ &= \frac{p(\bar{x}_i)}{p(\bar{x}_k)}p(w|\bar{x}_i) + \frac{p(\bar{x}_j)}{p(\bar{x}_k)}p(w|\bar{x}_j) \end{aligned} \tag{4.4}$$

where from Eq. (4.2):

$$p(\bar{x}_k) = \sum_{x \in \bar{x}_k} p(x) = \sum_{x \in \bar{x}_i} p(x) + \sum_{x \in \bar{x}_j} p(x) = p(\bar{x}_i) + p(\bar{x}_j) \tag{4.5}$$

Let $\delta I(\bar{x}_i, \bar{x}_j) = I(W; \bar{X}_b) - I(W; \bar{X}_a)$ denote the reduction in relevant information $I(W; \bar{X})$ that results from the merging of two categories $\bar{x}_i$ and $\bar{x}_j$, where $\bar{X}_b$ gives the categorization before the merge, and $\bar{X}_a$ gives the categorization after the merge. Following Slonim and Tishby (2000), the loss in relevant information that occurs due to a merge of two categories can be calculated:

$$\begin{aligned}
\delta I(\bar{x}_i, \bar{x}_j) &= I(W; \bar{X}_b) - I(W; \bar{X}_a) \\
&= \sum_w p(w|\bar{x}_i) p(\bar{x}_i) \frac{p(w|\bar{x}_i)}{p(w)} + \sum_w p(w|\bar{x}_j) p(\bar{x}_j) \frac{p(w|\bar{x}_j)}{p(w)} \\
&\quad - \sum_w p(w|\bar{x}_k) p(\bar{x}_k) \frac{p(w|\bar{x}_k)}{p(w)} \\
&= \sum_w p(w|\bar{x}_i) p(\bar{x}_i) \frac{p(w|\bar{x}_i)}{p(w)} + \sum_w p(w|\bar{x}_j) p(\bar{x}_j) \frac{p(w|\bar{x}_j)}{p(w)} \\
&\quad - \sum_w p(w|\bar{x}_i) p(\bar{x}_i) \frac{p(w|\bar{x}_k)}{p(w)} - \sum_w p(w|\bar{x}_j) p(\bar{x}_j) \frac{p(w|\bar{x}_k)}{p(w)} \\
&= p(\bar{x}_i) \sum_w p(w|\bar{x}_i) \frac{p(w|\bar{x}_i)}{p(w|\bar{x}_k)} + p(\bar{x}_j) \sum_w p(w|\bar{x}_j) \frac{p(w|\bar{x}_j)}{p(w|\bar{x}_k)} \\
&= p(\bar{x}_i) D_{KL} \left[p(w|\bar{x}_i) || p(w|\bar{x}_k) \right] + p(\bar{x}_j) D_{KL} \left[p(w|\bar{x}_j) || p(w|\bar{x}_k) \right] \\
&\equiv p(\bar{x}_k) D_{JS} \left[p(w|\bar{x}_i) || p(w|\bar{x}_j) \right] \\
&= \left[p(\bar{x}_i) + p(\bar{x}_j) \right] D_{JS} \left[p(w|\bar{x}_i) || p(w|\bar{x}_j) \right]
\end{aligned} \tag{4.6}$$

where the Jensen-Shannon divergence is defined in terms of the Kullback-Leibler divergence in Eq. (3.40) as:

$$D_{JS}\left(p(w|\bar{x_i})||p(w|\bar{x_j})\right) \equiv \frac{p(\bar{x_i})}{p(\bar{x_k})} D_{KL}\left(p(w|\bar{x_i})||p(w|\bar{x_k})\right) + \frac{p(\bar{x_j})}{p(\bar{x_k})} D_{KL}\left(p(w|\bar{x_j})||p(w|\bar{x_k})\right) \quad (4.7)$$

The Jensen-Shannon divergence is another measure for the "distance" between two distributions, defined as a weighted sum of Kullback-Leibler divergences. The loss in information due to the combining of the two categories is, from Eq. (4.6), simply equal to a scaled "distance" between the summaries of the surface features provided by the two categories, where the scale is the probability $p(\bar{x}_k) = p(\bar{x}_i) + p(\bar{x}_j)$. Therefore, at every stage of the merging process, the algorithm provided by Slonim and Tishby (2000) requires merging the two categories which have the smallest scaled distance between their summaries. The result is a hierarchy of categorizations, one at each cardinality $|\bar{X}|$ such that $1 \leq |\bar{X}| \leq |X|$, where the two categories that are combined at each level of the hierarchy have surface feature summaries (parameterized by the conditional distributions $p(w|\bar{x})$) that are the closest, where the distance between two summaries is provided by Eq. (4.6). The weight $p(\bar{x}_k)$ provides a measure for the size of the category; smaller categories tend to be combined first.

4.2 Materials and Methods

In the next two sections, I will provide an example of a deterministic surface feature typing of a particular textbook of physics problems using the agglomerative information bottleneck algorithm described by Slonim and Tishby (2000). I used a textbook of physics problems with expired copyright that was used by secondary school students to prepare for the former Harvard College general entrance examinations (Snyder and Palmer, 1900). I used all the problems from the mechanics chapters in the textbook.

Chapter Headings in Snyder and Palmer (1900)
1. Pressure in Liquids
2. Density and Specific Gravity
3. Tenacity and Elasticity
4. Composition and Resolution of Forces
5. Force and Acceleration
6. Energy
7. Work
8. Coefficient of Friction
9. Gravitation
10. Pendulums
11. Levers, Inclined Plane, Center of Gravity
12. Machines

Table 4.1: This table lists the chapter headings assigned by Snyder and Palmer (1900) in *One Thousand Problems in Physics* for the problems typed in the analysis below.

As discussed in detail in Chapter 3, the problems in this textbook are written in a terse and abstract fashion. The chapter headings assigned by the author for the chapters in the textbook are given in Table 4.1. The most probable words present in the problems in each chapter are given in Table 4.2.

This textbook is available for download from the website `archive.org`; instructions for how to obtain it are included in the appendix. On this website, multiple formats for the textbook exist, including the textbook in plain text format, presumably created using optical character recognition software from the original scanned textbook. Using the structure of this plain text file, I wrote a Unix shell script `otpp_filter.sh` to transform the text file into a particular format expected by the utilities used to create the typings. This filter was applied to the text, and is available in the appendix.

I then applied a utility, `aib`, to implement the algorithm of Slonim and Tishby (2000) on this filtered plain text file. I programmed this utility in C, using a C library implementation of the algorithm by Vedaldi and Fulkerson (2008), along with supporting mathematical routines provided by Galassi et al. (2011). The output of this utility provides another plain text file from which the entire hierarchical set of clusterings for the problems can be extracted. Using the filter `otpp_filter.sh`,

Pressure in Liquids	the of a is on cm pressure box and in with water what filled tube be top side sides i at bottom mercury cubical square long its one if ft
Density and Specific Gravity	of a the in is sp gr g what and water to it cm liquid weighs will its volume how be air on weight when submerged much with block which
Tenacity and Elasticity	a of the wire ibs in mm and diameter is long no as be what by force m will pull how to if much i brass that stretch with stretched
Comp. and Res. of Forces	the of a is to and force ibs at what ft acting an angle in two be on from with other end beam how each it one by forces horizontal
Force and Acceleration	the a of ft in second how body is will from it velocity to per what at seconds far with and be force ball its tower ground long horizontal strike
Energy	a the of energy ft is what second with weighing it its how in body per velocity and to at kinetic ibs will i be seconds mass vertically much from
Work	the a of ft in is how to much work ibs done and weighing be position must plane from gravity its he if by long end body at level an
Coefficient of Friction	the a of ft is plane to coefficient friction body how it will velocity along horizontal in if what with rest weight per second mass ibs inclined moving must an
Gravitation	the of earth a what mass and is surface their miles attraction for how to at would as moon if that are its weigh will i it distance be from
Pendulums	the a of pendulum is what in seconds times at length if to earth vibrations surface long are how vibrates vibration as time lengths minute be relative second that two
Levers, Center of Gravity	the of a ft ibs is and in end from on at to bar weight what center long one lever gravity i other find load point be weights force square
Machines	the of a in is to and what ibs diameter be force ft wheel if must on axle weight by at pulley from which attached applied are pulleys two with

Table 4.2: This table provides a representation of the words used in all of the problems contained in the 12 chapters of *One Thousand Problems in Physics* typed in the next section. This table lists the 30 most probable words in each chapter, sorted with the highest probability words listed first.

I removed all characters that were not alphabetic or not a space, and treated all remaining sequences of alphabetic letters as independent words for use in the analysis. In the appendix, I present the source code for `aib` as well as other supporting utilities I wrote and used in the analysis. There, I demonstrate how to use these utilities may be used to produce a surface feature typing from any set of physics problems.

4.3 Analysis

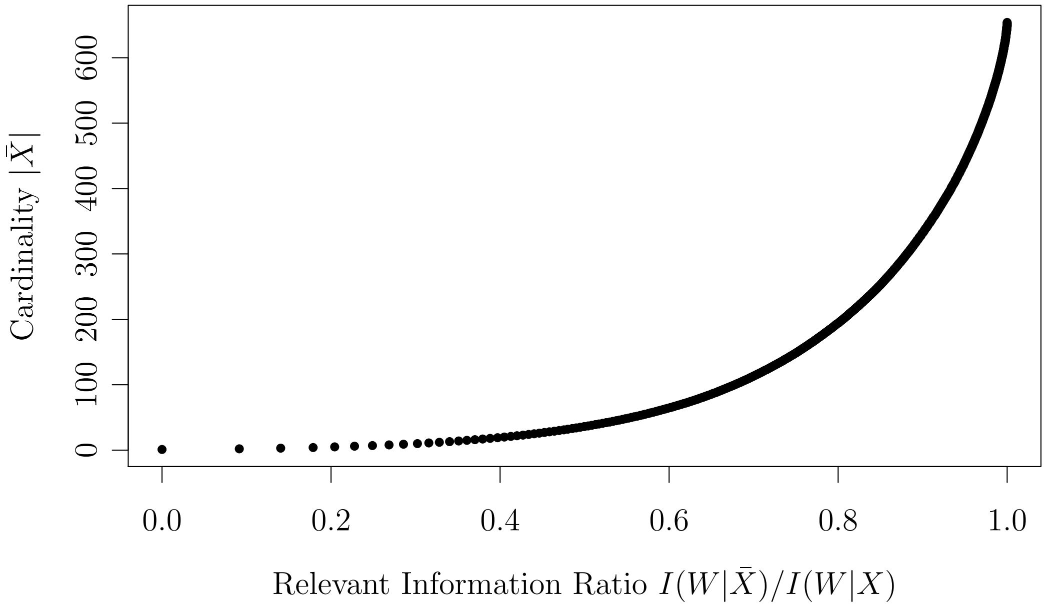

Figure 4.1: This is a plot of the cardinality $|\bar{X}|$ versus the normalized relevant information $I(W;\bar{X})/I(W;X)$ for each categorization produced by the agglomerative information bottleneck algorithm.

In this section, I generate deterministic surface feature types for the mechanics problems in the textbook by Snyder and Palmer (1900) using the algorithm of Slonim and Tishby (2000). In Fig. 4.1, I plot the cardinality $|\bar{X}|$ versus the normalized relevant information $I(W;\bar{X})/I(W;X)$ for each typing $\bar{X}$ produced by the algorithm. Fig. 4.1 demonstrates that as the cardinality of the typing $|\bar{X}|$ is increased, the amount of relevant information $I(W;\bar{X})$ concerning the surface features provided by the typing also increases. This result is expected: if the cardinality of the typing $\bar{X}$ is increased, the average number of problems per type decreases, and therefore the average uncertainty $S(W|\bar{X})$ in the surface features of an arbitrary but unknown problem, assuming knowledge of the problem's type, should decrease.

Though Fig. 4.1 provides a representation of the amount of relevant information concerning the surface features, it does not provide a demonstration of the potential

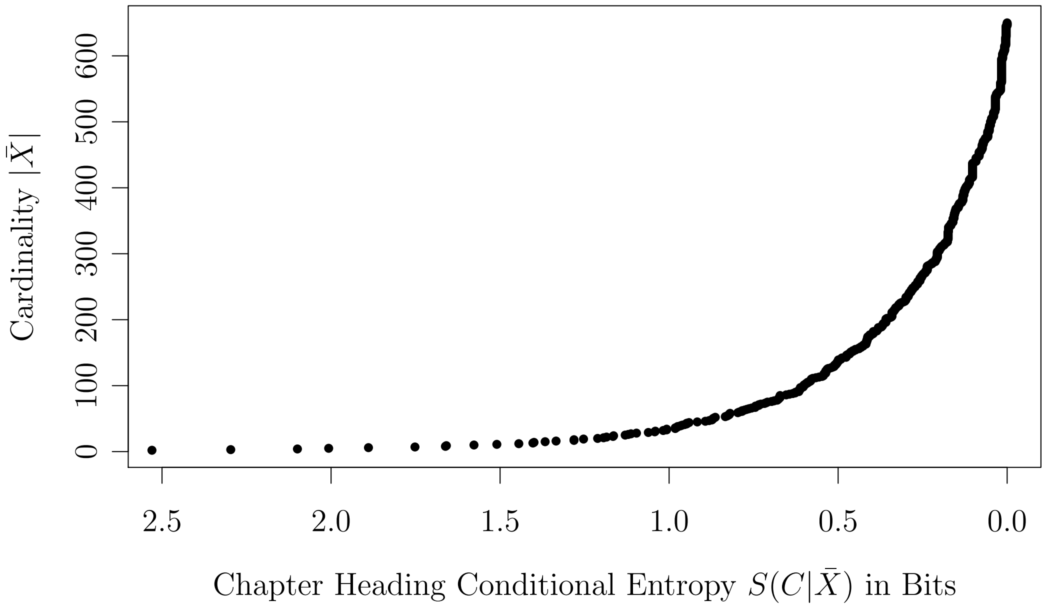

Figure 4.2: This is a plot of the cardinality $|\bar{X}|$ for each categorization produced by the agglomerative information bottleneck algorithm versus the average entropy $S(C|\bar{X})$ that remains in the chapter heading C for an arbitrary problem — after inferring the problem's type from the set $\bar{X}$. The entropy $S(C)$ in the chapter heading of an arbitrary problem, before inference of its type, is $S(C) = 3.286244$. This plot demonstrates that inference of the type for an arbitrary problem, for most of the categorizations produced by the agglomerative information bottleneck algorithm, reduces the uncertainty in the chapter heading for that problem considerably, on average. Furthermore, this plot demonstrates that a typing $\bar{X}$ with only modest cardinality $|\bar{X}|$ is sufficient to imply an average remaining uncertainty $S(C|\bar{X})$ of less than 1.5 bits.

relevance of these typings for problem solving. Consider Fig. 4.2, where I plot the cardinality $|\bar{X}|$ versus the average amount of remaining entropy $S(C|\bar{X})$ concerning the chapter heading for an arbitrary problem — after inferring the problem's type from the set $\bar{X}$. For a particular example, the inference of a surface feature type from the categorization $\bar{X}$ with cardinality $|\bar{X}| = 35$ decreases the average uncertainty concerning the chapter heading of an arbitrary problem from $S(C) = 3.286244$ to about one bit: $S(C|\bar{X}) = 0.981795$.

Note that the author used many of the chapter headings to label "deep structure" for the problems present in that chapter. For example, the "Gravitation" chapter label is a deep structure label provided by the author to denote problems requiring the use

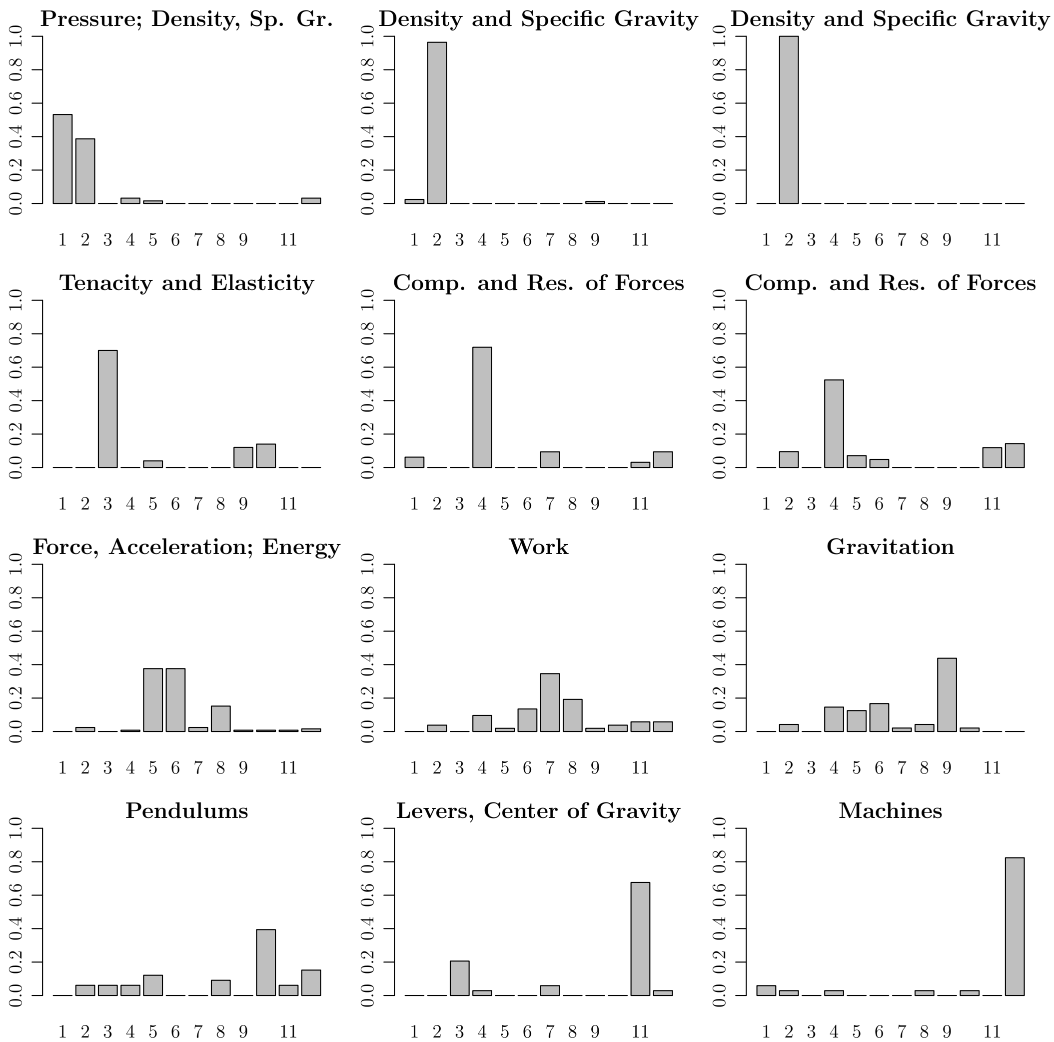

Figure 4.3: This is a plot of the types produced at the 12 cardinality level using the agglomerative information bottleneck method for the 12 chapters of problems included in Table 4.1. Each barplot represents one of the types. The y-axis provides the percentage of problems belonging to that type that originate from the corresponding chapter on the x-axis. The titles for each of the plots provide my ex-post labeling of the types, based on which chapter has the largest percentage of problems belonging to that type.

of Newton's law of gravity. Furthermore, from Fig. 4.2, inference of a type from a surface feature typing $\bar{X}$ with modest cardinality $|\bar{X}|$ reduces the average uncertainty in the chapter heading of an arbitrary problem considerably. This suggests that most of the categorizations $\bar{X}$ produced by the algorithm may provide significant relevant information concerning the deep structure underlying the problems.

Fig. 4.3 provides a plot of the surface feature types generated by the algorithm at the $|\bar{X}| = 12$ cardinality level. Each barplot represents one of the types. On each barplot, the x-axis denotes the chapter, while the y-axis denotes the fraction of problems in that type that belong to that chapter. The titles for each of the graphs indicate my ex-post labeling of the categories, based on which chapter has the largest percentage of problems in that category, since the names for the types are not generated automatically by the principle. The name of the chapters used for the labelings has been taken from Table 4.1. For any category in Fig 4.3, the majority of problems belong to either one or two chapters, implying that this surface feature typing provides, on average, significant information concerning the chapter from which an arbitrary problem originates.

A representation of the content of the categorization at the $|\bar{X}| = 12$ cardinality level is provided by Table 4.3, where I list the most probable words in each category. Note the qualitative similarity between Table 4.2 and Table 4.3, especially for chapters 1-4 and 9-12, which provides another indication that the surface feature typing captures much of the chapter structure assigned by the author to the textbook.

To delve more deeply into the contents of these types, consider two problems taken from the surface feature type in row 3, column 3 on Fig 4.3, that I labeled using the "Gravitation" chapter heading:

> Gravitation Chapter, Problem 7: If moon's diameter is 1/4, and mass 1/80 that of earth, what is the weight of a lb. mass on the moon's surface?

> Gravitation Chapter, Problem 13: If the volume of the sun should decrease

(1,1)	the of a is on in cm and what pressure water box with be side filled tube its will sp top cubical gr sides at mercury block i long if
(1,2)	of a the sp gr in is cm and will liquid what how be it to water submerged its much rod volume on g with must which long wood floats
(1,3)	of the a g in is what sp gr weighs water and it air to volume weight its liquid when b piece body weigh flask metal certain mark mercury which
(2,1)	of a the wire ibs and mm is what diameter in long be no m are force relative as pull if will their by how to two brass with same
(2,2)	the a of is to and ibs ft what at end angle beam an horizontal with weight long rope on force attached from rod wall it by how weighing be
(2,3)	of the a and force ibs to is in two other acting what be each one g end on are which forces at must by cm attached can from cord
(3,1)	a the of ft second is it body in what velocity how with per to will energy its be seconds weighing from plane and far at horizontal ibs must rest
(3,2)	the a of ft to in is how much ibs plane work if and be body an on incline friction inclined done weighing force will from what must coefficient it
(3,3)	the of a is at earth miles mass per to what how and rate ft will surface it i second if hour from moon its far would ball energy that
(4,1)	the a of is in what to at pendulum force weight length and seconds if ibs be ft surface it that edge times distance will on from between i table
(4,2)	the of a ft is in and ibs end from at long bar as center on by that weight one other square lever gravity i how find much weights load
(4,3)	the of in is a to diameter and wheel what ft be ibs if must axle on force from minute by this up at an per pulley attached shaft has

Table 4.3: This table provides a representation of the contents of the typing produced at the 12 cardinality level by listing the most probable words in each category of Fig. 4.3, sorted with the highest probability words in each category listed first. The first column provides the row number and the column number of the category plotted in Fig. 4.3. Note the qualitative similarity between this table and Table 4.2, which provides an indication that the surface feature typing at the 12 cardinality level captures much of the chapter structure present in the textbook.

1/2, its mass remaining the same, what effect would this have upon its attraction for masses upon its surface?

Note that these two problems do not mention gravitation in their prompts, and in fact, only share the words *if*, *mass*, *of*, *surface*, *the*, and *what*. In spite of the lack of shared words, the surface features present in the entire environment provide sufficient statistics necessary to categorize these two problems into the same type. In the entire row 3, column 3 category, 22 out of 30 of the problems, or 73.3 percent of the problems from the gravitation chapter are present in that category. Also, 43.8 percent of the problems in that category are taken from the gravitation chapter. The most probable words in the row 3, column 3 category are provided in Table 4.3.

If a new categorization is generated by the algorithm, this time with 18 types rather than at the 12 cardinality level depicted in Fig 4.3, then a surface feature type is produced which contains 20 of the 22 gravitation problems present in the row 3, column 3 type at the 12 cardinality level. Only 5 problems present in this type at the 18 cardinality level do not originate from the gravitation chapter. Therefore, at the 18 cardinality level, 66.7 percent of the problems from the gravitation chapter are typed the same way, and 80 percent of the problems belonging to that type are from the gravitation chapter. The 12 cardinality level is more coarse than the 18 cardinality level; in the case of the row 3, column 3 type, the algorithm places other problems involving force and energy into the category.

Both the information bottleneck principle and the agglomerative algorithm used here should be viewed fundamentally as methods for summarizing the surface features used in the problems in the environment into a set of common surface feature contexts present in the environment, where every type collects a set of problems that the algorithm has identified as belonging to the same context. For example, in the row 3, column 3 type in Fig 4.3, all 20 gravitation problems present in the surface feature type at the 18 category level use the words: earth, sun, and/or moon. One could, loosely speaking, refer to these problems using the label "earth, sun, or moon," rather than using the gravitation chapter label I appropriated, as a way of tersely describing the surface feature context that the algorithm identified that these problems share. Yet, short labels providing a list of prominent surface features used in the environment should not be treated too literally, since the statistics of the surface features used in the entire environment are used to generate these types. Furthermore, the algorithm does not itself generate a label for these identified surface feature contexts.

The agglomerative information bottleneck algorithm identifies the row 3, column 3 type as a common surface feature context used in this textbook. This identified context commonly requires the use of Newton's law of gravity. If individuals can identify that a problem in this textbook belongs to this particular surface feature context, this identification immediately provides a strong suggestion of the deep structure (i.e. Newton's law of gravity) that may be present in that problem. As I discussed in Chapter 2, Chi et al. (1981) suggest that surface feature types are in general distinct from deep structure types, and that novices may be unable to see the deep structure in a problem because they type based on surface features. Yet, in introductory mechanics, Newton's law of gravity is used commonly when the objects interacting are celestial objects, like the "earth, sun, or moon" that I identified as present in all 20 of the gravitation problems discussed above. At least for gravitation in introductory mechanics, the claim that surface feature perception precludes recognition of important information concerning the deep structure of that context is suggested by the results here to be false.

In Fig 4.3, some of the surface feature types have a significant number of problems from two chapters. Consider the surface feature type in row 3, column 1. This surface feature type has a significant number of problems from both the chapter labeled "Force and Acceleration" and the chapter labeled "Energy" by the author. Recognizing that a problem belongs to this particular surface feature context, though it does provide an indication that the problems in the category either are “Force and Acceleration” problems or “Energy” problems, cannot alone provide an indication of which of the two considerations is most appropriately used to solve the problem. Table 4.3 provides the most probable words used in this category.

This particular surface feature type at the 12 cardinality level was created as a result of a merge of two surface feature types present at the 18 cardinality level, one with 51 members, the other with 73 members. Consider two problems from the category with 51 members:

> Force Chapter, Problem 14: A body is thrown downward from a tower 500 ft. high with a velocity of 50 ft. per second. How long will it take to reach the ground?
>
> Energy Chapter, Problem 19: A body weighing 20 g. has a kinetic energy of 1000 ergs. How far would it ascend vertically?

Even though the first problem is solved most easily using the kinematics equation $y = v_0 t + \frac{1}{2}gt^2$, and the second is solved most easily using the conservation of energy equation $KE_i = mgh$, the two problems were clustered together with this algorithm. Note however that both of these equations belong to a common context that could be labeled as the “point mass in a uniform gravitational field” context. For this particular surface feature type, 45 out of the 51 problems involve a point mass in a uniform gravitational field.

Identifying that either of these problems belong to this surface feature context provides relevant information useful for choosing a context-specific deep structure principle for solving either of these problems. In order to solve the second problem for instance, one cannot simply use the “deep structure” energy considerations hinted in the problem without knowing the context in which energy considerations should be applied. Conservation of energy is an important general principle that exists in physics. In order to solve a problem using the principle of conservation of energy, one must be able to invoke the precise, context-specific version of the principle that is appropriate for the context. In this case, if an individual can identify that this problem belongs to the "point mass in a uniform gravitational field" context, the individual has relevant information necessary to choose the proper form for the potential energy necessary to solve the problem. In the case of the first example, perceiving that the problem involves a point mass moving in a uniform gravitational field provides relevant information in the sense that it allows the individual to restrict the set of physics equations to consider to only those pertinent to the context of a point mass moving in a uniform gravitational field.

Note that both problems above share only the words: *a*, *body*, *how*, *it*, and *of*. Of course, the fact that there is a *body* in the physics problem does not alone imply that it should belong to this common context. The algorithm also recognizes this fact. For example, consider the following problem categorized into a different type at both the 12 and 18 category level:

> Density and Specific Gravity Chapter: Problem 23: A body weighs 540 g. in air and 240 g. in a liquid twice as dense as water. What is (a) the volume of the body? (b) the density?

This problem utilizes the use of the buoyancy force. The algorithm recognized that this problem belongs to a different surface feature context than the two previous problems by using the statistics of the surface features used in the entire environment.

Of course, the algorithm presented here is not perfect in generating surface feature contexts that are relevant for every problem. For example, not all of the problems in the Newton's law of gravity chapter were categorized into the row 3, column 3 type in Fig 4.3, and not all of the kinematic problems taking place in a uniform gravitational field from Chapter 5 were placed into the same category as the problems described above.

Yet, if we examine the surface feature contexts more broadly by turning our attention back to Fig 4.3, each type at the 12 cardinality level provides, for the majority of problems, relevant information for accessing context-specific problem solving knowledge. If each chapter heading is viewed as providing a label for a restricted set of principles necessary for solving problems in that chapter, then an arbitrary surface feature type provides, for the majority of problems, relevant information for selecting a restricted set of potentially useful principles from the set of all possibly useful principles in introductory mechanics.

Further examination of the surface feature contexts generated by the agglomerative information bottleneck method for the textbook by Snyder and Palmer (1900) can be carried out by using the utilities provided in the appendix. The appendix describes how to use these utilities to generate surface feature types directly from the raw plain text file of the book hosted on the `archive.org` website. These utilities can also be used to run the agglomerative information bottleneck algorithm on any other set of problems.

4.4 Conclusion

This chapter generated a set of surface feature types in a principled fashion from a set of introductory mechanics problems in Snyder and Palmer (1900). The generated surface feature types provide a set of common surface feature contexts for the set of introductory mechanics problems used in the textbook.

I have argued that these common surface feature contexts provide relevant information for identifying potentially relevant context-specific principles. Even though the fundamental objective of physics as a field of scientific inquiry is to identify generic principles that govern the natural world, physics practice is *context-specific.* For example, in ordinary problem solving, to apply the generic law of conservation of energy to a particular problem, one must be able to perceive an appropriate context for the problem in order to determine which particular context-specific conservation of energy principle may be useful. In research, identification of the generic principles that govern the natural world often proceeds through the observation of particular contexts in which the principle is operating.

The results presented in this chapter suggest that an appropriately developed principle that asserts the importance of surface features for context recognition may be needed to move physical law toward a theory of physics practice: one that provides the explicit context that individuals recognize and need in order to invoke the context-specific deep structure principle used to solve physics problems. These results demonstrate that summarization of the surface features over a particular set of introductory physics problems written in a terse and abstract fashion is one way of identifying relevant contexts of physics present in that set. The information bottleneck principle, operating again on an appropriately terse and abstract set of introductory physics problems,[1] may be a viable candidate for extending physical law to include physics practice.

As discussed in Chapter 3, algorithms like the one used here, operating on a modern textbook of introductory mechanics, may not do as well as the agglomerative information bottleneck algorithm did with the textbook by Snyder and Palmer (1900) due to potential spurious correlations in the statistics that could be identified as relevant. This of course does not mean necessarily that students presented with an arbitrary problem from these modern textbooks would be unable in general to recognize a relevant context of which the problem is a particular example. It may simply mean that algorithms like the one presented here, operating on a modern set of introductory physics problems, may be inadequate to predict the types that

[1] As mentioned earlier, I hope that objective criteria can be developed for identifying why any given set of physics problems provides a more or less appropriate environment for modeling physics problem perception. Criteria might be able to be productively developed after conducting experiments of the types individuals actually perceive in the initial stage of physics problem classification.

novices perceive for that set of problems. Though Chi et al. (1981) does suggest that individuals are able to perceive "abstracted visual configurations," it is unclear from the data they present whether novices, after having taken a course in physics, would or would not be able to identify relevant physical contexts for the problems similar to the ones demonstrated in the previous section. Further experimentation is needed to assess the ability of novices to identify relevant contexts, either for problems written similarly to those of Snyder and Palmer (1900), or for problems written similarly to those presented in modern textbooks. Given a large set of physics problems written in a way similar to those of Snyder and Palmer (1900), the algorithm presented in this chapter may be able to provide useful insight into what individuals might actually perceive when they first confront a physics problem from that set.

Note that the conjectured principle of physics practice provided in this dissertation does not supply a principle governing how individuals select the particular context-specific equation of physics to apply to a problem, given the set of restricted principles implied by the perceived context. I am also not claiming that this particular deterministic algorithm provides an accurate approximation for the exact types that individuals may perceive; after all, we should not expect that everyone would assign the exact same type to a given problem. Rather, the information bottleneck principle described in Chapter 3 may be a viable candidate for accurately and quantitatively predicting the types individuals perceive, since it allows for the probabilistic assignment of types.

An appropriate experiment to access the types individuals perceive may be easily identified. Suppose that novices and experts were only given 10 seconds from the moment they were shown a problem to describe it, either verbally or on paper. Recall that the subjects in the experiments of Chi et al. (1981) were given unlimited time to categorize the problems, and that experts took longer on average to categorize the problems than did novices. These results of Chi et al. (1981) suggest that the extra time experts used to categorize problems may have contributed to the difference in categorizations between novices and experts. If the time provided to individuals to perceive problems is restricted however, the experimental findings might be markedly different from the original results reported by Chi et al. (1981). Could both experts and those novices who have already taken a course in physics perceive the problems in similar ways if given restricted time to perceive the problems? Could the results, in the aggregate, be described as a summarization of the surface features used in the environment of problems presented to the subjects? These are open and testable questions.

Others have recognized the context-specificity of physics practice. diSessa and Sherin (1998) formulated a theoretical notion of a concept called a coordination class that can be used to qualitatively model normative physics concepts like force. According to coordination class theory, physics concepts like force are not able to be applied by individuals equally well in all contexts. Rather, diSessa and Sherin (1998) claim that knowledge of the concept of force may actually exist as a set of *conceptual projections*. Together, these conceptual projections make up the concept's *coordination class*.

A particular conceptual projection for a particular individual defines the degree to which that individual can see and apply the given concept in a particular context. According to coordination class theory, knowledge of a concept like force is not properly defined as the ability to apply an abstract, context-independent concept, but instead should be defined more precisely as a set of conceptual projections, each of which describes the ability of the individual to apply a concept to a particular context. Yet, coordination class theory does not provide a method to determine, a-priori, what the relevant contexts of introductory mechanics actually are. In the case of routine quantitative physics problems written in a way similar to the problems provided in Snyder and Palmer (1900), the proposed principle of physics practice provided in this dissertation may partially fill that gap.

As coordination class theory posits, and I support here with this conjectured principle of physics practice, expertise in physics is characterized by the ability to operate in diverse contexts. According to these two theories, the relevant grain size for discussing the application of physics knowledge is at the level of the context, rather than at the level of the abstract principle. The model presented in this dissertation provides a principled way for deriving potential contexts for physics problems in which physics cognition may operate.

As I conclude this dissertation, I can offer another perspective concerning the difference between expert and novice categorization discussed by Chi et al. (1981). Recall that both experts and novices were asked to categorize based on a more or less ambiguous prompt of "similarities of solution." As I mentioned earlier, this difference in categorization between experts and novices may have simply been due to a difference in interpretation of this ambiguous prompt.

I speculate that the categorizations the novices produced in the experiment of Chi et al. (1981) may just be reflective of the fundamental reality that the practice of physics is context-dependent. It may be only after enough experience does the student begin to internalize the important idea in physics that physical law can often be described in highly context-independent terms. The expert may simply be viewing physics problems through this lens of physics as abstract law, while the novice may be viewing physics through the lens of physics as practice. Indeed, both the novice and the expert may be right.

Appendix A

Utilities for Computing Surface Feature Types

In this appendix, I discuss the utilities used for the Chapter 4 analysis. The source code for these utilities is included below, and may be used to produce (on Unix systems) the surface feature typings discussed in Chapter 4 directly from the plain text source for Snyder and Palmer (1900) stored on the `archive.org` website. These utilities can also be used to run the agglomerative information bottleneck algorithm on any other set of problems.

The utility `aib` runs the agglomerative information bottleneck algorithm described by Slonim and Tishby (2000), using the `vlfeat` implementation of the algorithm by Vedaldi and Fulkerson (2008) along with supporting matrix routines provided by Galassi et al. (2011). The `aib` utility takes as input a set of problems. A label for each problem is placed on odd lines, while problem texts are placed on even lines. The problem text identified by the label follows the label, with a newline separating the two. If the text file contains N problems, the text file should contain $2N$ lines. For example:

```
1.1
what is the acceleration due to gravity at the surface of the earth
1.2
suppose the earth can be approximated as having uniform density what
would the acceleration be at half the radius of the earth let the
acceleration at the surface be g
...
```

Note that the problem identifier need not take the form $x.y$. Each problem identifier should however uniquely label a single problem. In spite of the visual presence of newlines within the text of problem 1.2 in the example given above, the problem text should not have newlines within it; the newline delimits the end of either a problem identifier or a problem text. The `aib` utility is case sensitive, which is why in this example, as well as for the input into `aib` used in the analysis, I have converted the characters in the input file to lowercase.

Words in the `aib` utility are defined as being separated by space or tab characters. Furthermore, punctuation is not removed by the `aib` utility. Punctuation should be removed prior to processing the set of problems with the `aib` utility if punctuation is to be ignored in the analysis.

The output of the `aib` utility is a representation of the entire hierarchical clustering produced by the agglomerative algorithm. For the purposes of this appendix, it suffices to discuss only the third to last row and second to last row in the output. The third to last row is a space separated list of the N problem identifiers. The second to last row is a space separated list of $2N-1$ integers, for each of the $2N-1$ nodes of a binary tree that represents the entire hierarchical clustering. If the second to last row is considered as a vector v_i, this vector has $2N-1$ components, with the value v_i providing the index of the "parent" category that was created as a result of a merge of category i and some other category. The first N elements in the vector correspond to the N problems in the input, which are the leaves in the tree. The other elements in the array correspond to the created categories. The indices 0 to $N-1$ label the N problems in the input, and the indices from N to $2N-2$ are assigned to the categories generated by the algorithm in the order in which the categories were created by the agglomerative process.

`aib` calculates the required joint probability table $p(w,x)$ in the procedure `probcompute` by assuming that $p(x)$ is proportional to the number of words in the problem. Equivalently, if each word $w \in W$ in x has frequency $N(w,x)$, then $p(w,x) \equiv N(w,x)/\sum_{wx} N(w,x)$. The source code for `aib` is included below. It is intended to be machine readable, but it does require both the open source `vlfeat` library from `http://www.vlfeat.org`, along with the open source gnu scientific library matrix routines, available at `http://www.gnu.org/software/gsl/` in order to compile.

```
#include <stdio.h>
#include <string.h>
#include <stdlib.h>
#include <math.h>
#include <gsl/gsl_vector.h>
#include <gsl/gsl_matrix.h>
#include <gsl/gsl_blas.h>
#include <vl/aib.h>
/* Fixed size of character array that holds a word */
#define MAXWORD 100
#define uint unsigned int


/* A linked list associated with each unique word encountered.
   questno indexes a question in which word encountered.
```

```
   count provides the number of times the word appeared in the
   question indexed by questno.  questno provides a unique
   internal index for each question */

struct cnode {
  uint questno;
  double count;        /* questno provides unique internal */
  struct cnode *next;
};



/* Binary Tree stores words, sorted, as they are encountered
   in the input file. */
typedef struct tnode *Treeptr;

typedef struct tnode{
  char *word;
  Treeptr left;
  Treeptr right;
  struct cnode *first;
  struct cnode *last;
} Treenode;

/* Represents the joint probability p(x,w).  Also stores the
   labels for the rows of the table (the questions), along
   with the labels for the columns (the words) */
typedef struct {
```

```
  char **questionarray;
  char **wordarray;
  gsl_matrix *matrix;
} Probtable;

Probtable *readprob(FILE *fp);
Treenode *addtree(Treenode *p, char *w, uint totalquestions,
                  uint *diffwords);
void *aibanalysis(FILE *fp, Probtable *fulltable,
                  uint **parents);
void printaibanalysis(FILE *fp, Probtable *fulltable,
                      uint *parents);
void assigntoarray(Probtable *fulltable, Treenode *p,
                   uint totalquestions, uint diffwords);
void assigntoarray_internal(Probtable *fulltable,
                            Treenode *p);
int getword(FILE *fp, char *word, int lim);
void probcompute(gsl_matrix *matrix);

int main(int argc, char *argv[]) {

  FILE *inputfile;
  Probtable *fulltable;
  uint *parents;

  if(argc == 1)
    inputfile = stdin;
```

```
  else if((inputfile = fopen(argv[1], "r")) == NULL) {
      printf("Bad filename %s\n", argv[1]);
      return -1;
  }
  fulltable = readprob(inputfile);
  aibanalysis(stdout, fulltable, &parents);
  printaibanalysis(stdout, fulltable, parents);
  return 0;
}

/* Implements the aib algorithm using the vlfeat library.
   The algorithm produces a hierarchical clustering of problems.
   Each node in the binary tree that represents the clustering is
   labeled uniquely.vl_aib_get_parents returns the results of the
   hierarchical clustering, as an array. There are 2*N-1 elements
   in the array, one element for each node in the binary tree
   which represents the clustering.  The first N elements in the
   array correspond to the N problems in the input, these are the
   leaves in the tree.  The other elements in the array
   correspond to the other nodes in the tree.  The value of the
   parents array for a given node in the tree provides the
   "parent" of that node in the tree. */

void *aibanalysis(FILE *fp, Probtable *fulltable,uint **parents) {

  VlAIB *aib;
  uint i;
```

```
  aib = vl_aib_new((double  *) fulltable->matrix->data,
                   fulltable->matrix->size1,
                   fulltable->matrix->size2);
  vl_aib_process(aib);
  *parents = vl_aib_get_parents(aib);
}

/* Prints a list of the question labels, and
   a list of the parents array representing the
   results of the clustering */

void printaibanalysis(FILE *fp, Probtable *fulltable,
                      uint *parents) {

  int i;

  for(i = 0; i < fulltable->matrix->size1; i++)
    fprintf(fp, "%s%s", fulltable->questionarray[i],
           i < fulltable->matrix->size1 -1 ? " ": "\n");
  for(i = 0; i < 2 * fulltable->matrix->size1 - 1; i++)
    fprintf(fp, "%d%s", parents[i],
            i < 2 * fulltable->matrix->size1 - 2 ? " " : "\n");
}

/* Reads the input, and converts the input to the joint
   probability table p(w,x) */
```

```
Probtable *readprob(FILE *fp) {

  Probtable *fulltable = (Probtable *) malloc(sizeof(Probtable));
  Treenode *root = NULL;
  char question[MAXWORD], word[MAXWORD], *tempword;
  int c;
  uint totalquestions = 0, diffwords = 0, inquest = 0;

  fulltable->questionarray = NULL;
  while((c = getword(fp, question, MAXWORD)) != EOF) {
    getc(fp);  /*Throw away return after the question label*/
   while((c = getword(fp, word, MAXWORD)) != EOF) {
      if(isalpha(word[0])) {
        /* Signal for new question*/
        if(!inquest) {
          totalquestions++;
          inquest = 1;
          if(fulltable->questionarray == NULL)
            fulltable->questionarray = (char **)
               malloc(sizeof(char *));
          else
            fulltable->questionarray = (char **)
              realloc(fulltable->questionarray,
                      sizeof(char *) * totalquestions);
          fulltable->questionarray[totalquestions - 1] =
             strdup(question);
```

```
        }
        root = addtree(root, word, totalquestions, &diffwords);
      }
      else if(c == '\n') {
        inquest = 0;
        break;
      }
    }
  }
  assigntoarray(fulltable, root, totalquestions, diffwords);
  return fulltable;
}

/* Adds words to sorted binary tree of words
   as they are encountered in the input */

Treenode *addtree(Treenode *p, char *w, uint totalquestions,
                  uint *diffwords) {

  int cond;

  if (p == NULL) {
    (*diffwords)++;
    p = (Treenode *) malloc(sizeof(Treenode));
    p->word = strdup(w);
    p->last = p->first = (struct cnode *)
                         malloc(sizeof(struct cnode));
```

```
    p->last->questno = totalquestions;
    p->last->count = 1;
    p->last->next = NULL;
    p->left = p->right = NULL;
  }
  else if((cond = strcmp(w, p->word)) == 0) {
    if(p->last->questno == totalquestions) {
      p->last->count++;
    }
    else {
      p->last->next =
         (struct cnode *) malloc(sizeof(struct cnode));
      p->last = p->last->next;
      p->last->questno = totalquestions;
      p->last->count = 1;
      p->last->next = NULL;
    }
  }
  else if(cond < 0)
    p->left = addtree(p->left, w, totalquestions, diffwords);
  else
    p->right = addtree(p->right, w, totalquestions, diffwords);
  return p;
}

/* Takes binary tree of words and converts this tree to
   probability table */
```

```
void assigntoarray(Probtable *fulltable, Treenode *p,
                   uint totalquestions, uint diffwords) {
  fulltable->matrix = gsl_matrix_alloc(totalquestions, diffwords);
  fulltable->wordarray =
    (char **) malloc(sizeof(char *) * diffwords);
  assigntoarray_internal(fulltable, p);
  probcompute(fulltable->matrix);
}

static uint wordno = 0;

void assigntoarray_internal(Probtable *fulltable, Treenode *p) {

  struct cnode *cp;
  uint questno;

  if( p!= NULL) {
    assigntoarray_internal(fulltable, p->left);
    fulltable->wordarray[wordno] = p->word;
    for(questno = 0, cp = p->first; cp != NULL;
        questno++, cp = cp->next) {
      for(; questno < cp->questno - 1 ; questno++)
        gsl_matrix_set(fulltable->matrix, questno, wordno, 0);
      gsl_matrix_set(fulltable->matrix, questno,
                     wordno, cp->count);
    }
    /* Add trailing zeros if reached end of linked list */
```

```
    for(; questno < fulltable->matrix->size1; questno++)
      gsl_matrix_set(fulltable->matrix, questno, wordno, 0);
    wordno++;
    assigntoarray_internal(fulltable, p->right);
  }
}

/* Routine for storing a word from input */
int getword(FILE *fp, char *word, int lim) {

  int c;
  char *w = word;

  while((c = getc(fp)) == ' ' || c == '\t')
    ;
  if(c != EOF)
    *w++ = c;
  if(isspace(c) || c == EOF) {
    *w = '\0';
    return c;
  }
  for(; --lim > 0; w++)
    if(isspace(*w = getc(fp))) {
      ungetc(*w, fp);
      break;
    }
  *w = '\0';
```

```
  return word[0];
}

/* Computes the probability table from the frequency of words
   present in each problem.
   This particular version of probcompute implements the
   assumption that physics cognition is adapted to a random process
   parameterized by p(x) where p(x) scales with the number of words
   present in x */

void probcompute(gsl_matrix *probtable) {

  int total = 0;
  int i, j;

  for(i = 0; i< probtable->size1; i++)
    for(j = 0; j < probtable->size2; j++)
      total += gsl_matrix_get(probtable,i,j);
  gsl_matrix_scale(probtable, 1.0/total);
}
```

The utility `catcontents` takes as input the 3rd to last and 2nd to last rows of the `aib` utility, and produces a set of common contexts at a desired cardinality. It is run using the following syntax:

```
catcontents -c [numcategories]
```

where `[numcategories]` provides the number of types to generate. A list of types is output, one per line, with each line containing a list of the problem identifiers that belong to that category. The source code `catcontents.c` is included below:

```
/* catcontents.c Generates categorization.
   Syntax: catcontents -c [numcategories]
   The command line option provides the number of categories
   to be generated.

   Takes as input two lines. The first line is a space-separated
   list of N unique problem identifiers. The second line is a
   space-separated list of 2N-1 integers, one element for each
   node in the binary tree which represents the clustering. These
   input lines are both generated by the aib utility.

   Outputs a list of categories, one per line.  Each line contains
   a space-separated list of the problem identifiers that belong
   to that category. */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#define MAXWORD 100

/* Structure for a category.  Includes a unique identifier
   for the category, the number of questions in the category
   and an array of questions that belong to the category */

typedef struct  {
  int identifier;
  int numquest;
```

```
  char **questionarray;
} Category;

/* Structure that holds the entire input data.  questionarray
   holds the list of N problem identifiers.  parents represents
   the 2N-1 nodes, where parents[i] provides the parent for the
   ith node in the tree.  totalquestions stores the value for N
*/

typedef struct {
  int totalquestions;
  char **questionarray;
  int *parents;
} Inputdata;

int getword(FILE *fp, char *word, int lim);
Inputdata *readinput(FILE *fp);
Category *categoryprocess(Inputdata *input, int numcategories);
void printcategories(Category *cattable, int numcategories);
int getparents(FILE *fp);

int main(int argc, char *argv[]) {

  int c, numcategories = 0;
  FILE *fp;
  Inputdata *input;
  Category *cattable;
```

```
while(--argc > 0 && (*++argv)[0] == '-') {
  c = *++argv[0];
  switch(c) {
  case 'c':
    if(--argc > 0)
      numcategories = atoi(*++argv);
    break;
  default:
    printf("Illegal option %c\n", c);
    return -1;
  }
}
if(numcategories < 2) {
  printf("Need to request at least two categories.\n",
          numcategories);
  return -1;
}
if(argc == 0)
  fp = stdin;
else if(argc == 1 && (fp = fopen(*argv, "r")) == NULL) {
  printf("Bad filename %s\n", *argv);
  return -1;
}
input = readinput(fp);
cattable = categoryprocess(input, numcategories);
```

```
  printcategories(cattable, numcategories);
}



/* Reads the two rows of input data */

Inputdata *readinput(FILE *fp) {

  int c, i;
  char question[MAXWORD];
  Inputdata *input = (Inputdata *) malloc(sizeof(Inputdata));

  input->totalquestions = 0;
  while((c = getword(fp, question, MAXWORD)) != '\n') {
    input->totalquestions++;
    if(input->questionarray == NULL)
      input->questionarray = (char **) malloc(sizeof(char *));
    else
      input->questionarray = (char **)
        realloc(input->questionarray,
                sizeof(char *) * input->totalquestions);
    input->questionarray[input->totalquestions - 1] =
      strdup(question);
  }
  input->parents =
    (int *) malloc(sizeof(int) * (2 * input->totalquestions - 1));
  for(i = 0; i < 2*input->totalquestions - 1; i++)
```

```
    input->parents[i] = getparents(fp);
  return input;
}

/* Routine for storing a word from the first line of input */

int getword(FILE *fp, char *word, int lim) {

  int c;
  char *w = word;

  while((c = getc(fp)) == ' ' || c == '\t')
    ;
  if(c != EOF)
    *w++ = c;
  if(isspace(c) || c == EOF) {
    *w = '\0';
    return c;
  }
  for(; --lim > 0; w++)
    if(isspace(*w = getc(fp))) {
      ungetc(*w, fp);
      break;
    }
  *w = '\0';
  return word[0];
}
```

```
/* Routine for reading an integer from second line of input */

int getparents(FILE *fp) {

  int c, total = 0;
  while((total = getc(fp)) == ' ' || total == '\t')
    ;
  if(total == EOF || total == '\n')
    return total;
  total = total - '0';
  while((c = getc(fp)) != ' ' && c != '\n')
    total = 10 * total + (c - '0');

  return total;
}

/* Creates an array of categories */

Category *categoryprocess(Inputdata *input, int numcategories) {
  Category *cattable =
    (Category *) malloc(sizeof(Category) * numcategories);
  int i, j, k;

  for(i=0; i < numcategories; i++) {
    cattable[i].identifier = -1;
    cattable[i].numquest = 0;
```

```
    cattable[i].questionarray = NULL;
  }
  for(i = 0; i < input->totalquestions; i++) {
    for(j = i; (2 * input->totalquestions - 1) - input->parents[j] >=
          numcategories; j = input->parents[j])
      ;
    for(k = 0; cattable[k].identifier != j &&
        cattable[k].identifier != -1; k++)
      ;
    if(cattable[k].identifier == j)
      cattable[k].questionarray = (char **)
        realloc(cattable[k].questionarray,
                sizeof(char *) * cattable[k].numquest + 1);
    else {
      cattable[k].identifier = j;
      cattable[k].questionarray = (char **) malloc(sizeof(char *));
    }
    cattable[k].questionarray[cattable[k].numquest] =
       input->questionarray[i];
    cattable[k].numquest++;
  }
  return cattable;
}

/* Outputs the categories */

void printcategories(Category *cattable, int numcategories) {
```

```
  int i,j;

  for(i = 0; i < numcategories; i++) {
    for(j = 0; j < cattable[i].numquest; j++)
      printf("%s%s", cattable[i].questionarray[j],
             j != cattable[i].numquest - 1 ? " " : "\n");
  }
}
```

Using the utilities `aib` and `catcontents`, along with a list of problems structured in the format expected by `aib`, a surface feature categorization for those problems can be produced. For example, if a list of problems is stored in the file `problems`, then the following Unix pipeline can be used to generate a deterministic categorization $\bar{X}$ with $|\bar{X}| = 12$:

```
cat problems | aib | aibcleanup.sh | catcontents -c 12
```

The small shell script `aibcleanup.sh` simply extracts the third to last and second to last rows of output from `aib`, before feeding that output into the input for `catcontents`:

```
#!/bin/sh
#aibcleanup.sh Cleans up aib output
#Deletes first line, last line, and all lines that start with "aib"
#Only two lines remain, first line is a list of every problem id
#Second line is a row which provides a 'breadth-first' representation
#of the knowledge hierarchy

sed '
1d
```

```
/^aib/d
$d'
```

Since the plain text version of the public domain textbook *One Thousand Problems in Physics* by Snyder and Palmer (1900) that can be downloaded from `archive.org` is not in the format `aib` expects, I wrote a shell script, `otpp_filter.sh`[1] that utilizes the structure of the plain text source to transform the text file into the proper format:

```
#!/bin/sh

#otpp_filter.sh Filter for One-Thousand Problems in Physics.
#Input is the raw text source.
#Output is only the mechanics problems
#in an appropriate format for the aib program.

#Raw source for this filter obtained from:
#http://archive.org/details/onethousandprobl00snyduoft
#From here, click All files:HTTP, then download:
#onethousandprobl00snyduoft_djvu.txt
#Raw text file last obtained on April 3, 2012
#at 11:46 am PDT

sed '
#Delete header
1,181d
#Delete line containing only "i"
#since not part of any problem
217d
```

[1]Instructions for how to obtain the plain text source from `archive.org` are included in this file.

```
#Delete table of rectangular
#cross-sections in middle of document.
1381,1790d
#Delete lines indicating the value of
#the g, and the values for force
2198,2202d
2511d
#Delete notes stating that the centrifugal
#force should be ignored
3151d
3288,3289d
#Delete note stating that friction and
#slipping of belts should be ignored
3579d
#Only keep mechanics problems
3881,$d
#Delete Chapter Names
#(Uniquely defined by lines with four adjacent
#capital letters)
/[A-Z][A-Z][A-Z][A-Z]/d
#Delete Empty Lines
/^$/d


#Since problems in the awk filter
#that follow assume that questions
#start with the problem number, a few
#anomalies due to the optical character
```

```
#recognition need to be corrected
#to allow the awk process that follows to
#run smoothly

#Delete anomalous leading period on line 282
282s/^\. //
#Delete anomalous leading comma on line 402
402s/^,//
#Substitute "2^" for "21" on line 2288
2288s/^2\^/21/
#Delete anomalous leading period on line 3692
3692s/^\. //
#Delete leading "5. " on line 3056
#Numbers are eventually ignored anyway.
#(Numbers at the beginning of lines
#are used in the awk script to
#indicate new problems.
3056s/^5\. //' |

awk '
  #Iterate the chapter counter when we hit
  #first problem of chapter
  /^1\. / {chap++}
  #Include newline and chapter number
  #before problem number
  /^[1-9][0-9]*\. / {printf("\n%d.",chap)}
  #Print rest of problem
```

```
  {printf("%s",$0)}
' |

sed '
#Delete empty first line
#introduced in previous awk script
1d
#Place the problem id on separate line
s/^\([1-9][0-9]*\.[1-9][0-9]*\)\. /\1\
/' |

#Join syllables of words that were
#originally hyphenated across lines
sed 's/\([A-Za-z][a-z]\)- \([a-z][a-z]\)/\1\2/g' |

#Only include the mechanics problems
sed '
1311,$d' |

#Delete all characters that are not a space or letters
awk '
NR % 2 == 0 {
  gsub(/[^A-Za-z ]/,"")
}
{ print }' |

#Make all letters lowercase
```

```
tr A-Z a-z
```

With this filter, a deterministic typing of Snyder and Palmer (1900) can be generated directly from the plain text source found at `archive.org` using the following Unix pipeline:

```
cat onethousandprobl00snyduoft_djvu.txt |
otpp_filter.sh | aib | aibcleanup.sh |
catcontents -c [number of categories]
```

If `vlfeat` and the gnu scientific library are properly installed on a Unix system, `aib.c` and `catcontents.c` are compiled as `aib` and `catcontents`, then the following pipeline can be used to generate deterministic typings from the `aib` algorithm with an arbitrary plain text source:

```
cat [text source] | [source filter] | aib |
aibcleanup.sh | catcontents -c [number of categories]
```

This Unix pipeline is packaged below in the shell script `sftyping.sh`:

```
#!/bin/sh
#sftyping.sh [number of categories] [problem source] [filter]
#Produces a typing of [problem source].  Requires the number
#of categories to be specified.  The filter is used to place
#the source in the appropriate format for the aib utility.
#The filter is


#Assumes that aib, aibcleanup.sh, and catcontents are
#either located in the current directory, or somewhere
#in the path.
PATH=$PATH:.
```

```
#Used if no filter is passed as an argument
if [ $# -eq 2 ]
then
  cat $2 | aib | aibcleanup.sh |
  catcontents -c $1
fi

#Used if a filter is passed as an argument
if [ $# -eq 3 ]
then
  cat $2 | $3 | aib | aibcleanup.sh |
  catcontents -c $1
fi
```

Using `sftyping.sh`, I generated Fig 4.3 directly from the plain text version of the public domain textbook *One Thousand Problems in Physics* by Snyder and Palmer (1900) with the following `bargraph.sh` script:

```
#!/bin/sh

#bargraph.sh Produces tex barplot directly
#from raw text source of
#One Thousand Problems in Physics by
#William Snyder (1900) found at
#http://archive.org/details/onethousandprobl00snyduoft

#Adjust this path appropriately
PATH=$PATH:.
```

```
#Pipeline from Original Source to Create
#Fig 4.3 Data
sftyping.sh 12 onethousandprobl00snyduoft_djvu.txt  otpp_filter.sh |
cat2bargraph.sh  > bargraphdata

#R Code for Generating Plot from "bargraphdata"
R --vanilla --slave << RCODE
require(tikzDevice);
tikz("otpp_barplot.tex", width=5.5, height=5.5);

title = c("Pressure; Density, Sp. Gr.", "Machines",
  "Density and Specific Gravity",
  "Comp. and Res. of Forces","Density and Specific Gravity",
  "Force, Acceleration; Energy","Gravitation","Work","Pendulums",
  "Comp. and Res. of Forces","Levers, Center of Gravity",
  "Tenacity and Elasticity");

par(mfcol=c(4,3), mar=c(3,2,1.6,1), ps=9);
h <- read.csv("./bargraphdata", header=FALSE);
order <- c(1,12,6,9,3,4,8,11,5,10,7,2);

for(i in order)
barplot(as.numeric(h[i,]), main=title[i],names.arg=1:12,
  ylim = c(0,1));

dev.off();
```

RCODE

This script calls the statistical software `R` to create the `tex` source `otpp_barplot.tex` for Fig 4.3. This script also generates the file `bargraphdata` which is a list of data used to generate the plot, with each row a comma separated list for the height of each bar within that type. This plot relies on the shell script `cat2bargraph.sh`, which takes as input the category contents output by `catcontents` and outputs the barplot data used in Fig 4.3:

```
#!/bin/sh


#cat2bargraph.sh Produces bargraph data in Fig 4.3.
#Takes as input a particular categorization.
#Each line contains one category in the categorization.
#On each line, a space separated list of questions
#belonging to that category is expected.
#Each question should take the format x.y, with x
#indicating the chapter number, and y the question
#number.


awk '
{
  for(i=0; i<=NF; i++) {


#Initialize the variables minchapter and
#maxchapter
      if(NR == 1 && i == 1) {
        minchapter = int($i);
        maxchapter = int($i);
```

```
      }

#chapcount(i,j) stores number of problems
#from chapter j in category i
      chapcount[NR, int($i)]++;

#numquest(i) stores number of problems
#in category i
      numquest[NR]++;

      if(int($i) < minchapter)
          minchapter = int($i);
      if(int($i) > maxchapter)
          maxchapter = int($i);
  }
}
END {
    for(i=1; i <= NR; i++) {
        for(j=minchapter; j<=maxchapter; j++)
                printf("%.3f%s", chapcount[i, j]/numquest[i],
                        j < maxchapter ? ", " : "");
        printf("\n");
    }
}'
```

With `sftyping.sh`, one can produce a deterministic surface feature categorization for an arbitrary plain text source using the following syntax, just as was done above in the `bargraph.sh` script:

```
sftyping.sh [number of categories] [problem source] [filter]
```

The last argument to `sftyping.sh` allows a filter to be provided to convert the problem source into the format described at the beginning of the appendix, just as was done above for the plain text source of Snyder and Palmer (1900). If the problem source is already in that format, this argument is optional.

Bibliography

AAPT. Physics first: An informational guide for teachers, school administrators, parents, scientists, and the public, 2007.

V. A. Aleven and K. R. Koedinger. An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26 (2):147 – 179, 2002.

P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

Y. Anzai. *Toward a General Theory of Expertise*, chapter Learning and Use of Representations for Physics Expertise, pages 64–92. Cambridge, 1991.

J. D. Barrow and F. J. Tipler. *The Anthropic Cosmological Principle*. Oxford, 1988.

A. Ben-Naim. *Entropy Demystified: The Second Law Reduced to Plain Common Sense*. World Scientific, 2008.

W. Bialek. *Physics of Bio-Molecules and Cells*, chapter Thinking About the Brain, pages 485–577. Les Houches Summer School, Vol. 75. Springer, 2002.

W. Bialek. Biophysics: Searching for principles, 2011. URL `http://www.princeton.edu/~wbialek/PHY562.html`.

H. B. Callen. *Thermodynamics and an Introduction to Thermostatics*. Wiley, 1985.

B. Carter. Large number coincidences and the anthropic principle in cosmology. *IAU Symposium 63: Confrontation of Cosmological Theories with Observational Data*, pages 291–298, 1973.

W. G. Chase and H. A. Simon. Perception in chess. *Cognitive Psychology*, 4(1):55 – 81, 1973.

M. T. Chi. This week's citation classic. *Current Contents*, 42:12, 1993.

M. T. Chi. *Cambridge Handbook of Expertise and Expert Performance*, chapter Two Approaches to the Study of Experts' Characteristics, pages 21–30. Cambridge, 2006.

M. T. Chi, P. J. Feltovich, and R. Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152, 1981.

M. T. H. Chi. Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, 14(2):pp. 161–199, 2005.

R. Clausius. *The Mechanical Theory of Heat*. Macmillan, 1879.

J. Confrey. A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education*, 16:3–56, 1990.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

A. A. diSessa. Toward an epistemology of physics. *Cognition and Instruction*, 10 (2/3):105–225, 1993.

A. A. diSessa. *Cambridge Handbook of the Learning Sciences*, chapter A History of Conceptual Change Research: Threads and Fault Lines. Cambridge University Press, 2006.

A. A. diSessa and B. L. Sherin. What changes in conceptual change. *International Journal of Science Education*, 20(10):1155–1191, 1998.

A. A. diSessa, N. M. Gillespie, and J. B. Esterly. Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28(6):843 – 900, 2004.

K. A. Ericsson and J. Smith, editors. *Toward a General Theory of Expertise*. Cambridge, 1991.

M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. *Gnu Scientific Library: Reference Manual*. Network Theory Ltd., 2011.

W. J. Gibbs. On the equlibrium of heterogenous substances. *Transactions of the Connecticut Academy*, 1876.

F. Gobet and H. A. Simon. Five seconds or sixty? presentation time in expert memory. *Cognitive Science*, 24(4):651–682, 2000.

F. Gobet, P. C. Lane, S. Croker, P. C.-H. Cheng, G. Jones, I. Oliver, and J. M. Pine. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6):236 – 243, 2001.

C. E. Hmelo-Silver. Problem-based learning: What and how do students learn? *Educational Psychology Review*, 16(3):235–266, 2004.

S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller. The expertise reversal effect. *Educational Psychologist*, 38(1):23–31, 2003.

J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon. Models of competence in solving physics problems. *Cognitive Science*, 4:317–345, 1980a.

J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon. Expert and novice performance in solving physics problems. *Science*, 208:1335–1342, 1980b.

D. J. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

E. Mazur. *Peer Instruction*. Prentice Hall, 1997.

M. McCloskey. Intuitive physics. *Scientific American*, 248(4), 1983.

A. Newell and H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 1976.

G. Özdemir and D. B. Clark. An overview of conceptual change theories. *Eurasia Journal of Mathematics, Science & Technology Education*, 3(4):351–361, 2007.

R. Penrose. *The Emperor's New Mind*. Oxford, 1989.

D. Perkins and G. Salomon. Are cognitive skills context-bound? *Educational Researcher*, 18(1):16–25, 1989.

R. F. Port and T. van Gelder, editors. *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT, 1998.

E. Redish. Who needs to learn physics in the 21st century and why? In *Plenary Lecture, Physics Teacher Education Beyond 2000, GIREP International Conference, Barcelona*, 2000.

P. E. Roudi and Y. Latham. Mutual information. *Scholarpedia*, 4(1):1658, 2009.

R. A. Serway and J. W. Jewett. *Physics for Scientists and Engineers*. Thomson Brooks/Cole, 2004.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. *ACM SIGIR*, pages 208–215, 2000.

J. P. Smith, A. A. diSessa, and J. Roschelle. Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *The Journal of the Learning Sciences*, 3(2):115–163, 1993 - 1994.

W. H. Snyder and I. O. Palmer. *One Thousand Problems in Physics*. Ginn and Company, 1900. URL `http://archive.org/details/onethousandprobl00snyduoft`.

L. P. Steffe and J. Gale, editors. *Constructivism in Education*. Erlbaum, 1995.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *The 37th Annual Conference on Communiciation, Control, and Computing*, pages 368–379, 1999.

K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence*, 15:147–204, August 2005.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

E. von Glasersfeld. *Constructivism in Education*, chapter A Constructivist Approach to Teaching, pages 3–16. Erlbaum, 1995.

S. Vosniadou. *Reconsidering the Processes of Conceptual Change*, chapter On the Nature of Naive Physics, pages 61–76. Kluwer, 2002.

S. Vosniadou, editor. *International Handbook of Research on Conceptual Change*. Taylor and Francis, 2008.

S. Vosniadou and W. F. Brewer. Mental models of the day/night cycle. *Cognitive Science*, 18:123–183, 1994.

S. Weinberg. Reductionism redux. *New York Review of Books*, 1995.

J. A. Wheeler. *Complexity, Entropy, and the Physics of Information*, chapter Information, Physics, Quantum: The Search for Links. Addison Wesley, 1990.

S. White. Physics in us high schools: Dispelling the myths (at least some of them). Presentation, 2011 American Association of Physics Teachers Winter Meeting, 2011.

R. W. Williams and K. Herrup. The control of neuron number. *Annual Review of Neuroscience*, 11(1):423–453, 1988.

M. Wiser and S. Carey. *Mental Models*, chapter When Heat and Temperature Were One, pages 267–298. Erlbaum, Mahwah, NJ, 1983.