Chapter 6. Categorization: Describing Resource Classes and Types

The Discipline of Organizing, **2012**

Robert J. Glushko
Rachelle Annechino
Jess Hemerly
Longhao Wang

6.1 Introduction

For nearly two decades, a TV game show called "Pyramid" aired in North America. The show featured two competing teams, each team consisting of two contestants: an ordinary civilian contestant and a celebrity. In the show's first round, both teams' members viewed a pyramid-shaped sign that displayed six category titles, some straightforward like "Where You Live" and others less conventional like "Things You Need to Feed." Each team then had an opportunity to compete for points in 30-second turns. The goal was for one team member to gain points by identifying a word or phrase related to the category from clues provided by the other team member. For example, a target phrase for the "Where You Live" category might be "zip code," and the clue might be "Mine is 94705." "Things you Need to Feed" might include both "screaming baby" and "parking meter."

The team that won the first round advanced to the "Winner's Circle," where the game was turned around. This time, only the clue giver was shown the category name and had to suggest concepts or instances belonging to that category so that the teammate could guess the category name. Clues like "alto," "soprano," and "tenor" would be given to prompt the teammate to guess "Singing Voices" or "Types of Singers."

As the game progressed, the categories became more challenging. It was interesting and entertaining to hear the clue receiver's initial guess and how subsequent guesses changed with more clues. The person giving clues would often become frustrated, because to them their clues seemed obvious and discriminating but would seem not to help the clue receivers in identifying the category. Viewers enjoyed sharing in these moments of vocabulary and category confusion.

The Pyramid TV game show developers created a textbook example for teaching about categories -- groups or classes of things, people, processes, events or anything else that we treat as equivalent -- and categorization -- the process of assigning instances to categories. The game is a useful analog for us to illustrate many of the issues we discuss in this chapter. The Pyramid game was challenging, and sometimes comical, because people bring their own experiences and biases to understanding what a category means, and because not every instance of a category is equally typical or suggestive. How we organize reflects our thinking processes, which can inadvertently reveal personal characteristics that can be amusing in a social context. Hence, the popularity of the Pyramid franchise, which began on CBS in 1973 and has been produced in 20 countries.

Many texts in library science introduce categorization via cataloguing rules, a set of highly prescriptive methods for assigning resources to categories that some satirize as "mark 'em and park 'em." Similarly, many texts in computer science discuss the process of defining the categories needed to create, process, and store information in terms of programming language constructs: "here's how to define an abstract type, and here's the datatype system."[1] We take a very different approach in this chapter. In the following sections, we discuss how and why we create categories, reviewing some important work in philosophy, linguistics, and cognitive psychology so that we can better understand how categories are created and used in organizing systems. We discuss how the way we organize differs when

we act as individuals or as members of social, cultural, or institutional groups; later we share principles for creating categories, design choices and implementation experience. As usual, we close the chapter with a summary of the key points.

6.2 What Categories Are and Why We Create Them

Categories are **equivalence classes**, sets or groups of things or abstract entities that we treat the same. This does not mean that every instance of a category is identical, only that from some perspective, or for some purpose, we are treating them as equivalent based on what they have in common. When we consider something as a member of a category, we are making choices about which of its properties or roles we are focusing on and which ones we are ignoring. We do this automatically and unconsciously most of the time, but we can also do it in an explicit and self-aware way.[2]

When we encounter objects or situations, recognizing them as members of a category helps us know how to interact with them. For example, when we enter an unfamiliar building we might need to open or pass through an entryway that we recognize as a door. We might never have seen that particular door before, but it has properties and affordances that we know that all doors have; it has a doorknob or a handle; it allows access to a larger space; it opens and closes. By mentally assigning this particular door to the "doors" category we distinguish it from "windows," a category that also contains objects that sometimes have handles and that open and close, but which we do not normally pass through to enter another space. Categorization judgments are therefore not just about what is included in a class, but also about what is excluded from a class. Nevertheless, the category boundaries are not sharp; a "Dutch door" is divided horizontally in half so that the bottom can be closed like a door while the top can stay open like a window.

Categories are **cognitive and linguistic models** for applying prior knowledge; creating and using categories are essential human activities. Categories enable us to relate things to each other in terms of similarity and dissimilarity and are involved whenever we perceive, communicate, analyze, predict, or classify. Without categories, we would perceive the world as an unorganized blur of things with no understandable or memorable relation to each other. Every wall-entry we encounter would be new to us, and we would have to discover its properties and supported interactions as though we had never before encountered a door. Of course, we still often need to identify something as a particular instance, but categories enable us to understand how it is equivalent to other instances. We can interchangeably relate to something as specific as "the wooden door to the main conference room" or more generally as "any door."

All human languages and cultures divide up the world into categories. How and why this takes place has long been debated by philosophers, psychologists and anthropologists. One explanation for this differentiation is that people recognize structure in the world, and then create categories of things that "go together" or are somehow similar. An alternative view says that human minds make sense of the world by imposing structure on it, and that what goes together or seems similar is the outcome rather than a cause of categorization.

Bulmer framed the contrast in a memorable way by asking which came first, the chicken (the objective facts of nature) or the egghead (the role of the human intellect).[3]

A secondary and more specialized debate going on for the last few decades among linguists. cognitive scientists, and computer scientists concerns the extent to which the cognitive mechanisms involved in category formation are specialized for that purpose rather than more general learning processes.[4]

Even before they can talk, children behave in ways that suggest they have formed categories based on shape, color, and other properties they can directly perceive in physical objects. [5] People almost effortlessly learn tens of thousands of categories embodied in the culture and language in which they grow up. People also rely on their own experiences, preferences, and goals to adapt cultural categories or create entirely individual ones that they use to organize resources that they personally arrange. Later on, through situational training and formal education, people learn to apply systematic and logical thinking processes so that they can create and understand categories in engineering, logistics, transport, science, law, business, and other institutional contexts.

These three contexts of **cultural, individual,** and **institutional categorization** share some core ideas but they emphasize different processes and purposes for creating categories, so they are a useful distinction.[6] Cultural categorization can be understood as a natural human cognitive ability that serves as a foundation for both informal and formal organizing systems. Individual categorization tends to grow spontaneously out of our personal activities. Institutional categorization responds to the need for formal coordination and cooperation within and between companies, governments, and other goal-oriented enterprises.

6.2.1 Cultural Categories

Cultural categories are the archetypical form of categories upon which individual and institutional categories are usually based. Cultural categories tend to describe our everyday experiences of the world and our accumulated cultural knowledge. Such categories describe objects, events, settings, internal experiences, physical orientation, relationships between entities, and many other aspects of human experience. Cultural categories are acquired primarily, with little explicit instruction, through normal exposure of children with their caregivers; they are associated with language acquisition and language use within particular cultural contexts.

Two thousand years ago Plato wrote that living species could be identified by "carving nature at its joints," the natural boundaries or discontinuities between types of things where the differences are the largest or most salient. Plato's metaphor is intuitively appealing because we can easily come up with examples of perceptible properties or behaviors of physical things that go together that make some ways of categorizing them seem more natural than others.[7]

Natural languages rely heavily on nouns to talk about categories of things because it is useful to have a shorthand way of referring to a set of properties that co-occur in

predictable ways.[8] For example, in English (borrowed from Portuguese) we have a word for "banana" because a particular curved shape, greenish-yellow or yellow color, and a convenient size tend to co-occur in a familiar edible object, so it became useful to give it a name. The word "banana" brings together this configuration of highly interrelated perceptions into a unified concept so we do not have to refer to bananas by listing their properties.[9]

Languages differ a great deal in the words they contain and also in more fundamental ways that they require speakers or writers to attend to details about the world or aspects of experience that another language allows them to ignore. This idea is often described as **linguistic relativity**. (See the Sidebar).

For example, speakers of the Australian language Guugu Yimithirr don't use concepts of left and right, but rather use compass-point directions. Where in English we might say to a person facing north, "Take a step to your left", they would use their term for west. If the person faced south, we would change our instruction to "right", but they would still use their term for west. Imagine how difficult it would be for a speaker of Guugu Yimithirr and a speaker of English to collaborate in organizing a storage room or a closet.[10]

LINGUISTIC RELATIVITY

Linguistic diversity led Benjamin Whorf, in the mid 20th century, to propose an overly strong statement of the relationships among language, culture, and thought. Whorf argued that the particularities of one's native language determine how we think and what we can think about. Among his extreme ideas was the suggestion that, because some Native American languages lacked words or grammatical forms that refer to what we call "time" in English, they could not understand the concept. More careful language study showed both parts of the claim to be completely false.

Nevertheless, even though academic linguists have discredited strong versions of Whorf's ideas, less deterministic versions of **linguistic relativity** have become influential and help us understand cultural categorization. The more moderate position was crisply characterized by Roman Jakobson, who said that "languages differ essentially in what they *must* convey and not in what they *may* convey." In English one can say "I spent yesterday with a neighbor." In languages with grammatical gender, one must choose a word that identifies the neighbor as male or female. [11]

It is not controversial to notice that different cultures and language communities have different experiences and activities that give them contrasting knowledge about particular domains. No one would doubt that university undergraduates in Chicago would think differently about animals than inhabitants of Guatemalan rain forests, or even that different types of "tree experts" (taxonomists, landscape workers, and tree maintenance personnel) would categorize trees differently.[12]

6.2.2 Individual Categories

Individual categories are created in an organizing system to satisfy the ad hoc requirements that arise from a person's unique experiences, preferences, and resource collections. Unlike cultural categories, which usually develop slowly and last a long time, individual categories are created by intentional activity, in response to a specific situation, or to solve an emerging organizational challenge. As a consequence, the categories in individual organizing systems generally have short lifetimes and rarely outlive the person who created them.[13]

Individual categories draw from cultural categories but differ in two important ways. First, individual categories sometimes have an imaginative or metaphorical basis that is meaningful to the person who created them but which might distort or misinterpret cultural categories. Second, individual categories are often specialized or synthesized versions of cultural categories that capture particular experiences or personal history. For example, a person who has lived in China and Mexico, or lived with people from those places, might have highly individualized categories for foods they like and dislike that incorporate characteristics of both Chinese and Mexican cuisine.

Individual categories in organizing systems also reflect the idiosyncratic set of household goods, music, books, website bookmarks, or other resources that a person might have collected over time. The organizing systems for financial records, personal papers, or email messages often use highly specialized categories that are shaped by specific tasks to be performed, relationships with other people, events of personal history, and other highly individualized considerations.

Traditionally, individual categorization systems were usually not visible to, or shared with, others, whereas, this has become an increasingly common situation for people using web-based organizing system for pictures, music, or other personal resources. On web sites like the popular Flickr site for photos, people typically use existing cultural categories to tag their photos as well as individual ones that they invent. In particular, the typical syntactic constraint that tags are delimited by white space encourages the creation of new categories by combining existing category names using concatenation and camel case conventions; photos that could be categorized as "Berkeley" and "Student" are thus tagged as "BerkeleyStudent." Similar generative processes for creating individual category names are used with Twitter "hashtags" where tweets about events are often categorized with an ad hoc tag that combines an event name and a year identifier like #NBAFinals12. Web-based documents and product pages in web catalogs are commonly categorized with "ReadThis" and "BuyThis" tags that are meaningful for the individuals who created those categories for themselves, but which are not very informative for anyone else.

6.2.3 Institutional Categories

In contrast to cultural categories that are created and used implicitly, and to individual categories that are used by people acting alone, institutional categories are created and used explicitly and rationally, and most often by many people or computational agents in coordination with each other. Institutional categories are most often created in abstract

and information-intensive domains where unambiguous and precise categories are needed to regulate and systematize activity, to enable information sharing and reuse, and to reduce transaction costs. Furthermore, instead of describing the world as it is, institutional categories are usually defined to change or control the world by imposing semantic models that are more formal and arbitrary than those in cultural categories. The rigorous definition of institutional categories enables **classification:** the systematic assignment of resources to categories in an organizing system.

Institutional categories are of two broad types. **Institutional taxonomies** are systems of carefully defined categories designed to make it more likely that people or computational agents will organize and interact with resources in the same way. Laws, regulations, and standards often specify institutional taxonomies that consist of decision rules for assigning resources to new categories and behavior rules that prescribe how people must interact with them.[14] Among the thousands of standards published by the International Organization for Standardization (ISO) are many institutional taxonomies that govern the organization of resources and products in agriculture, aviation, construction, energy, healthcare, information technology, transportation, and almost every other industry sector.[15]

Institutional taxonomies are especially important in libraries and knowledge management. For example, the Dewey Decimal System enables different libraries to arrange books in the same categories, and the Diagnostic and Statistical Manual of Mental Disorders (DSM) in clinical psychology enables different doctors to assign patients to the same diagnostic and insurance categories.[16] Institutional taxonomies are covered in detail in Chapter 7, "Classification."

The second broad type of institutional categories are systems of **institutional semantics**, precisely defined abstractions or information components (Section 3.3.3) needed to ensure that information can be efficiently exchanged and used. The Universal Business Language (UBL) is a library of about 2000 semantic "building blocks" for common concepts like "Address," "Item," "Payment," and "Party" along with nearly 100 document types assembled from the standard components. UBL is widely used to facilitate the automated exchange of transactional documents in procurement, logistics, inventory management, collaborative planning and forecasting, and payment.[17] Institutional semantics are discussed in Chapters 8 and 9.

Institutional categorization stands apart from individual categorization primarily because it invariably requires significant efforts to reconcile or compromise about mismatches between existing individual categories, where those categories embody useful working or contextual knowledge that is lost in the move to a formal institutional system.[18]

Institutional categorization efforts must also overcome the vagueness and inconsistency of cultural categories because the former must often conform to stricter logical standards to support inference and meet legal requirements. Furthermore, institutional categorization is usually a process that must be accounted for in a budget and staffing plans. While some kinds of institutional categories can be devised or discovered by computational processes,

most of them are created through the collaboration of many individuals, typically from various parts of an organization or from different firms.[19] The different business or technical perspectives of the participants are often the essential ingredients in developing robust categories that can meet carefully identified requirements. And as requirements change over time, institutional categories must often change as well, implying version control, compliance testing, and other formal maintenance and governance processes.

Some institutional categories that initially had narrow or focused applicability have found their way into more popular use and are now considered cultural categories. A good example is the periodic table in chemistry, which Mendeleev developed in 1869 as a new system of categories for the chemical elements. The periodic table proved essential to scientists in understanding their properties and in predicting undiscovered ones. Today the periodic table is taught in elementary schools, and many things other than elements are commonly organized using a graphical structure that resembles the periodic table of elements in chemistry, including sci-fi films and movies, desserts, and superheroes. [20]

6.2.4 A "Categorization Continuum"

As we have seen, the concepts of cultural, individual, and institutional categorization usefully distinguish the primary processes and purposes for creating categories. However, these three kinds of categories can fuse, clash, and recombine with each other. Rather than viewing them as having precise boundaries, we might view them as regions on a continuum of categorization activities and methods.

Consider a few different perspectives on categorizing animals as an example. Scientific institutions categorize animals according to explicit, principled classification systems, such as a Linnaean taxonomy that assigns animals to a phylum, class, order, family, genus and species. Cultural categorization practices cannot be adequately described in terms of a master taxonomy, and are more fluid, converging with principled taxonomies sometimes, and diverging at other times. While human beings are classified within the animal kingdom in biological classification systems, people are usually not considered animals in most cultural contexts. Sometimes a scientific designation for human beings, "homo sapiens" is even applied to human beings in cultural contexts, since the genus-species taxonomic designation has influenced cultural conceptions of people and (other) animals over the years.

Animals are also often culturally categorized as pets or non-pets. The category "pets," in the US mainstream, commonly includes dogs, cats, and fish. A pet cat might be categorized at multiple levels that incorporate individual, cultural, and institutional perspectives on categorization – as an "animal" (cultural/institutional), as a "mammal" (institutional), as a "domestic short-hair" (institutional) as a "cat" (cultural), and as a "troublemaker" or a "favorite" (individual), among other possibilities, in addition to being identified individually by one or more pet names. Furthermore, not everyone experiences pets as just dogs, cats and fish. Some people have relatively unusual pets, like pigs. For individuals who have pet pigs or who know people with pet pigs, "pigs" may be included in the "pets" category. If enough people have pet pigs, eventually "pigs" could be included in mainstream culture's pet category.

It is not possible to entirely separate individual, cultural and institutional perspectives on categorization. Individuals form subcultures and contribute to institutions; culture influences individuals and institutions; institutions influence individuals and culture. Categorization skewed toward cultural perspectives incorporate relatively traditional categories, such as those learned implicitly from social interactions, like mainstream understandings of what kinds of animals are "pets", while categorization skewed toward institutional perspectives emphasizes explicit, formal categories, like the categories employed in biological classification systems.

A final example that demonstrates the interplay and conflict between the different contexts of categorization involves the vehicle categories in the US Corporate Average Fuel Economy (CAFE) standards. The CAFE standards sort vehicles into "passenger car" and "light truck" categories and impose higher minimum fuel efficiency requirements for cars because trucks have different typical uses.

When CAFE standards were first introduced in 1975, the vehicles classified as light trucks were generally used for "light duty" farming and manufacturing purposes. "Light trucks" might be thought of as a "sort of" in-between category – a light truck is not really a car, but sufficiently unlike a prototypical truck to qualify the vehicle's categorization as "light." Formalizing this sense of in-between-ness by specifying features that define a "car" and a "light truck" is the only way to implement a consistent, transparent fuel efficiency policy that makes use of informal, graded distinctions between vehicles.

A manufacturer whose average fuel economy for all the vehicles it sells in a year falls below the CAFE standards has to pay penalties. This encourages them to produce "sport utility vehicles" (SUVs) that adhere to the CAFE definitions of light trucks but which most people use as passenger cars. Similarly, the PT Cruiser, a retro-styled hatchback produced by Chrysler from 2000-2010, strikes many people as a car. It looks like a car; we associate it with the transport of passengers rather than with farming; and in fact it is formally classified as a car under emissions standards. But like SUVs, in the CAFE classification system, the PT Cruiser is a light truck.

CAFE standards have evolved over time, becoming a theater for political clashes between holistic cultural categories and formal institutional categories, which plays out in competing pressures from industry, government, and political organizations. Furthermore, CAFE standards and manufacturers' response to them are influencing cultural categories, such that our cultural understanding of what a car looks like is changing over time as manufacturers design vehicles like the PT Cruiser with car functionality in unconventional shapes to take advantage of the CAFE light truck specifications. [21]

Section 6.2 explained what categories are and the contrasting cultural, individual, and institutional contexts and purposes for which categories are created. In doing so, a number of different principles for creating categories were mentioned, mostly in passing.

6.3 Principles for Creating Categories

We now take a systematic look at principles for creating categories, including: enumeration, single properties, multiple properties and hierarchy, family resemblance, similarity, theory and goal-based categorization.

6.3.1 Enumeration

The simplest principle for creating a category is enumeration; any resource in a finite or countable set can be deemed a category member by that fact alone. This principle is also known as **extensional definition**, and the members of the set are called the **extension**. Many institutional categories are defined by enumeration as a set of possible or legal values, like the 50 states in the US or the ISO currency codes (ISO 4217).

Enumerative categories enable membership to unambiguously determined because a value like state name or currency code is either a member of the category or it isn't. But there comes a size when enumerative definition is impractical or inefficient, and the category either must be sub-divided or be given a definition based on principles other than enumeration.

For example, for millennia we earthlings have had a cultural category of "planet" as a "wandering" celestial object, and because we only knew of planets in our own solar system, the planet category was defined by enumeration: Mercury, Venus, Earth, Mars, Jupiter, and Saturn. When the outer planets of Uranus, Neptune, and Pluto were identified as planets in the 18th-20th centuries, they were added to this list of planets without any changes in the cultural category. But in the last couple of decades many heretofore unknown planets outside our solar system have been detected, making the set of planets unbounded, and definition by enumeration no longer works.

The International Astronomical Union thought it solved this category crisis in 2006 by proposing a definition of planet as "a celestial body that is (a) in orbit around a star, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit." Unfortunately, Pluto does not satisfy the third requirement, so it no longer is a member of the planet category, and instead is now called an "inferior planet." [22]

6.3.2 Single Properties

It is intuitive and useful to think in terms of properties when we identify instances and when we are describing instances (as we saw in Section 3.3, "Resource Identity" and in Chapter 4). Therefore, it should also be intuitive and useful to consider properties when we analyze more than one instance to compare and contrast them so we can determine which sets of instances can be treated as a category or equivalence class. Categories whose members are determined by one or more properties or rules follow the principle of **intensional definition**, and the defining properties are called the **intension**.[23]

Any single property of a resource can be used to create categories, and the easiest ones to use are often the intrinsic static properties. As we discussed in Chapter 4, intrinsic static

properties are those inherent in a resource that never change. The size, color, shape, weight, material of composition, and texture of natural or manufactured objects are intrinsic and static properties that can be used to arrange physical resources. For example, an organizing system for a personal collection of music that is based on the intrinsic static property of physical format might use categories for CDs, DVDs, vinyl albums, 8-track cartridges, reel-to-reel tape and tape cassettes.[24] Using a single property is most natural to do when the properties can take on only a small set of discrete values like music formats, and especially when the property is closely related to how the resources are used, as they are with the music collection where each format requires different equipment to listen to the music. Each value then becomes a subcategory of the music category.

The author, date, and location of creation of an intellectual resource cannot be directly perceived but they are also intrinsic static properties. The subject matter or purpose of a resource, its "what it is about" or "what it was originally for," are also intrinsic static properties that are not directly perceivable, especially for information resources.

The name or identifier of a resource is often arbitrary but once assigned normally does not change, making it an extrinsic static property. Any collection of resources with alphabetic or numeric identifiers as an associated property can use sorting order as an organizing principle to arrange spices, books, personnel records, etc., in a completely reliable way. Some might argue whether this organizing principle creates a category system, or whether it simply exploits the ordering inherent in the identifier notation. For example, with alphabetic identifiers, we can think of alphabetic ordering as creating a recursive category system with 26 (A-Z) top-level categories, each containing the same number of second-level categories, and so on until every instance is assigned to its proper place.[25]

Some resource properties are both extrinsic and dynamic because they are based on usage or behaviors that can be highly context-dependent. The current owner or location of a resource, its frequency of access, the joint frequency of access with other resources, or its current rating or preference with respect to alternative resources are typical extrinsic and dynamic properties that can be the basis for arranging resources and defining categories.

These properties can have a large number of values or are continuous measures, but as long as there are explicit rules for using property values to determine category assignment the resulting categories are still easy to understand and use. For example, we naturally categorize people we know on the basis of their current profession, the city where they live, their hobbies, or their age. Properties with a numerical dimension like "frequency of use" are often transformed into a small set of categories like "frequently used, "occasionally used," and "rarely used" based on the numerical property values.[26]

While there are an infinite number of logically expressible properties for any resource, most of them would not lead to informative and useful categories. Therefore, it is important to choose properties that are psychologically or pragmatically relevant for the resource domain being categorized. Whether something weighs more or less than 5000 pounds is a poor property to apply to things in general, because it puts cats and chairs in one category, and buses and elephants in another.[27]

To summarize: The most useful single properties to use for creating categories for an organizing system are those that are formally assigned, objectively measurable and orderable, or tied to well-established cultural categories, because the resulting categories will be easier to understand and describe.

If only a single property is used to distinguish among some set of resources and to create the categories in an organizing system, the choice of property is critical because different properties often lead to different categories. Using the age property, Bill Gates and Mark Zuckerberg are unlikely to end up in the same category of people. Using the wealth property, they most certainly would. Furthermore, if only one property is used to create a system of categories, any category with a large numbers of items in it will lack coherence because differences on other properties will be too apparent, and some category members will not fit as well as the others.

6.3.3 Multiple Properties

Organizing systems often use multiple properties to define categories. There are three different ways in which to do this that differ in the scope of the properties and how essential they are in defining the categories.

6.3.3.1 Multi-Level or Hierarchical Categories

If you have many shirts in your closet (and you are a bit compulsive or a "neat freak'), instead of just separating your shirts from your pants using a single property (the part of body on which the clothes are worn) you might arrange the shirts by style, and then by sleeve length, and finally by color. When all of the resources in an organizing system are arranged using the same sequence of resource properties, this creates a **hierarchy**, a multi-level category system.

If we treat all the shirts as the collection being organized, in the shirt organizing system the broad category of shirts is first divided by style into categories like "dress shirts," "work shirts," "party shirts," and "athletic or sweatshirts." Each of these style categories is further divided until the categories are very narrow ones, like the "white long-sleeve dress shirts" category. A particular shirt ends up in this last category only after passing a series of property tests along the way: it is a dress shirt, it has long sleeves, and it is white. Each test creates more precise categories in the intersections of the categories whose members passed the prior property tests.

Put another way, each subdivision of a category takes place when we identify or choose a property that differentiates the members of the category in a way that is important or useful for some intent or purpose. Shirts differ from pants in the value of the "part of body" property, and all the shirt subcategories share this "top part" value of that property. However, shirts differ on other properties that determine the subcategory to which they belong. Even as we pay attention to these **differentiating** properties, it is important to remember the other properties, the ones that members of a category at any level in the hierarchy have in common with the members of the categories that contain it. These properties are often described as **inherited** or **inferred** from the broader category.[28] For

example, just as every shirt shares the "worn on top part of body" property, every item of clothing shares the "can be worn on the body" property, and every resource in the "shirts" and "pants" category inherits that property.

Each differentiating property creates another level in the category hierarchy, which raises an obvious question: How many properties and levels do we need? In order to answer this question we must reflect upon the shirt categories in our closet. Our organizing system for shirts arranges them with the three properties of style, sleeve length, and color; some of the categories at the lowest level of the resulting hierarchy might have only one member, or no members at all. You might have yellow or red short-sleeved party shirts, but probably don't have yellow or red long-sleeved dress shirts, making them empty categories Obviously, any category with only one member does not need any additional properties to tell the members apart, so a category hierarchy is logically complete if every resource is in a category by itself.

However, even when the lowest level categories of our shirt organizing system have more than one member, we might choose not to use additional properties to subdivide it because the differences that remain among the members do not matter to us for the interactions the organizing system needs to support. Suppose we have two long-sleeve white dress shirts from different shirt makers, but whenever we need to wear one of them, we ignore this property. Instead, we just pick one or the other, treating the shirts as completely equivalent or substitutable. When the remaining differences between members of a category do not make a difference to the users of the category, we can say that the organizing system is pragmatically, or practically complete even if it is not yet logically complete. That is to say, it is complete "for all intents and purposes."

On the other hand, consider the shirt section of a big department store. Shirts there might be organized by style, sleeve length, and color as they are in our home closet, but would certainly be further organized by shirt maker and by size to enable a shopper to find a Marc Jacobs long-sleeve blue dress shirt of size 15/35. The department store organizing system needs more properties and a deeper hierarchy for the shirt domain because it has a much larger number of shirt instances to organize and because it needs to support many shirt shoppers, not just one person whose shirts are all the same size.

6.3.3.2 Different Properties for Subsets of Resources

A different way to use multiple resource properties to create categories in an organizing system is to employ different properties for distinct subsets of the resources being organized. This contrasts with the strict multi-level approach in which every resource is evaluated with respect to every property. Alternatively, we could view this principle as a way of organizing multiple domains that are conceptually or physically adjacent, each of which has a separate set of categories based on properties of the resources in that domain. This principle is used for most folder structures in computer file systems and by many email applications; you can create as many folder categories as you want, but any resource can only be placed in one folder.

The contrasts between intrinsic and extrinsic properties, and between static and dynamic ones, are helpful in explaining this method of creating organizing categories. For example, you might organize all of your clothes using intrinsic static properties if you keep your shirts, socks, and sweaters in different drawers and arrange them by color; extrinsic static properties if you share your front hall closet with a roommate, so you each use only one side of that closet space; intrinsic dynamic properties if you arrange your clothes for ready access according to the season; and, extrinsic dynamic properties if you keep your most frequently used jacket and hat on a hook by the front door. [29]

If we relax the requirement that different subsets of resources use different organizing properties and allow any property to be used to describe any resource, the loose organizing principle we now have is often called **tagging**. Using any property of a resource to create a description is an uncontrolled and often unprincipled principle for creating categories, but it is increasingly popular for organizing photos, web sites, email messages in gmail, or other web-based resources. We discuss tagging in more detail in Section 7.4, "Social/Distributed Classification."

6.3.3.3 Necessary and Sufficient Properties

A large set of resources does not always require many properties and categories to organize it. Some types of categories can be defined precisely with just a few **essential** properties. For example, a prime number is a positive integer that has no divisors other than 1 and itself, and this category definition perfectly distinguishes prime and not-prime numbers no matter how many numbers are being categorized. "Positive integer" and "divisible only by 1 and itself" are **necessary** or **defining** properties for the prime number category; every prime number must satisfy these properties. These properties are also **sufficient** to establish membership in the prime number category; any number that satisfies the necessary properties is a prime number. Categories defined by necessary and sufficient properties are also called **monothetic**. They are also sometimes called **classical categories** because they conform to Aristotle's theory of how categories are used in logical deduction using syllogisms.[30] (See the Sidebar, "The Classical View of Categories").

Theories of categorization have evolved a great deal since Plato and Aristotle proposed them over two thousand years ago, but in many ways we still adhere to classical views of categories when we create organizing systems because they can be easier to implement and maintain that way.

THE CLASSICAL VIEW OF CATEGORIES

The classical view is that categories are defined by necessary and sufficient properties. This theory has been enormously influential in Western thought, and is embodied in many organizing systems, especially those for information resources. This principle of defining categories is conceptually simple and has a straightforward implementation in technologies like database schemas, decision trees, and classes in programming languages.

However, as we will explain, we cannot rely on this principle to create categories in many domains and contexts because there are not necessary and sufficient properties. As a result, many psychologists, cognitive scientists, and computer scientists who think about categorization have criticized the classical theory.

We think this is unfair to Aristotle, who proposed what we now call the classical theory primarily to explain how categories underlie the logic of deductive reasoning: All men are mortal; Socrates is a man; Therefore, Socrates is mortal. People are wrong to turn Aristotle's thinking around and apply it to the problem of inductive reasoning, how categories are created in the first place. But this isn't Aristotle's fault; he was not trying to explain how natural cultural categories arise.

An important implication of necessary and sufficient category definition is that every member of the category is an equally good member or example of the category; every prime number is equally prime. Institutional category systems are often designed to have necessary and sufficient properties because it makes them conceptually simple and gives them a straightforward implementation in technologies like database schemas, decision trees, and classes in programming languages.

Consider the definition of an address as requiring a street, city, governmental region, and postal code. Anything that has all of these information components is therefore considered to be a valid address, and anything that lacks any of them will not be considered to be a valid address. If we refine the properties of an address to require the governmental region to be a state, and specifically one of the U.S. Postal Service's list of official state and territory codes, we create a subcategory for US addresses that uses an enumerated category as part of its definition. Similarly, we could create a subcategory for Canadian addresses by exchanging the name "province" for state, and using an enumerated list of Canadian province and territory codes.

6.3.4 The Limits of Property-Based Categorization

Property-based categorization works tautologically well for categories like "prime number" where the category is defined by necessary and sufficient properties. Property-based categorization also works well when properties are conceptually distinct and the value of a property is easy to perceive and examine, as they are with man-made physical resources like shirts.

Historical experience with organizing systems that need to categorize information resources has shown that basing categories on easily perceived properties is often not effective. There might be indications "on the surface" that suggest the "joints" between types of information resources, but these are often just presentation or packaging choices, That is to say, neither the size of a book nor the color of its cover are reliable cues for what it contains. Information resources have numerous descriptive properties like their title, author, and publisher that can be used more effectively to define categories, and these are certainly useful for some kinds of interactions, like finding all of the books written by a particular author or published by the same publisher. However, for practical purposes, the

most useful property of an information resource is its **aboutness**, which may not be objectively perceivable and which is certainly hard to characterize.[31] Any collection of information resources in a library or document filing system is likely to be about many subjects and topics, and when an individual resource is categorized according to a limited number of its content properties, it is at the same time not being categorized using the others.

When the web first started, there were many attempts to create categories of web sites, most notably by Yahoo! As the web grew, it became obvious that search engines would be vastly more useful because their near real-time text indexes obviate the need for *a priori* assignment of web pages to categories. Rather, web search engines represent each web page or document in a way that treats each word or term they contain as a separate property.

Considering every distinct word in a document as a property stretches our notion of property to make it very different from the kinds of properties we have discussed in the previous two sections of this chapter. We do not need that generality yet, so we will defer further discussion of document representation for search engines until Chapter 8 and stick with our more intuitive and limited concept of property.

6.3.5 Family Resemblance

In general, categorization based on explicit and logical consideration of properties is much less effective, and sometimes not even possible for domains where properties lack one or more of the characteristics of separability, perceptibility, and necessity. Instead, we need to categorize using properties in a statistical rather than a logical way to come up with some measure of resemblance or similarity between the resource to be categorized and the other members of the category.

Consider a familiar category like "bird." All birds have feathers, wings, beaks, and two legs. But there are thousands of types of birds, and they are distinguished by properties that some birds have that other birds lack: most birds can fly, most are active in the daytime, some swim, some swim underwater; some have webbed feet. These properties are correlated, a consequence of natural selection that conveys advantages to particular configurations of characteristics; birds that live in trees have different wings and feet than those that swim, for example. In the end, there is no single set of properties that are both necessary and sufficient to categorize a bird.

There are three related consequences of this complex distribution of properties for birds and for many other categories in cultural or natural (as opposed to man-made) domains. The first is an effect of **typicality** or **centrality** that makes some members of the category better examples than others, even if they share most properties. Most people consider a robin to be a more typical bird than a penguin. [32] Or try to define "friend" and ask yourself if all of the people you consider friends are equally good examples of the category. This effect is also described as **gradience** in category membership and reflects the extent to which the most characteristic properties are shared.

A second consequence is that the sharing of some but not all properties creates what we call **family resemblances** among the category members; just as biological family members do not necessarily all share a single set of physical features but still are recognizable as members of the same family. This idea was first proposed by the 20th century philosopher Ludwig Wittgenstein, who used "games" as an example of a category whose members resemble each other according to shifting property subsets. See the Sidebar "What is a Game?"[33]

WHAT IS A GAME?

Ludwig Wittgenstein (1889-1951) was a philosopher who thought deeply about mathematics, the mind, and language. In 1999 his *Philosophical Investigations* was ranked as the most important book of 20th century philosophy in a poll of philosophers.[34] In that book, Wittgenstein uses "game" to argue that many concepts have no defining properties, and that instead there is a "complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail." He contrasts board games, card games, ball games, games of skill, games of luck, games with competition, solitary games, and games for amusement. Wittgenstein notes that not all games are equally good examples of the category, and jokes about teaching children a gambling game with dice because he knows that this is not the kind of game that the parents were thinking of when they asked him to teach their children a game.

The third consequence, when categories do not have necessary features for membership, is that the boundaries of the category are not fixed; the category can be stretched and new members assigned as long as they resemble incumbent members. Personal video games and multiplayer online games like World of Warcraft did not exist in Wittgenstein's time but we have no trouble recognizing them as games and neither would Wittgenstein, were he alive. Recall that in Chapter 1 we pointed out that the cultural category of "library" has been repeatedly extended by new properties, as when Flickr is described as a web-based photo-sharing library. Categories defined by family resemblance or multiple and shifting property sets are termed **polythetic**.

We conclude that instead of using properties one at a time to assign category membership, we can use them in a composite or integrated way to determine **similarity**. Something is categorized as an A and not a B if it is more similar to A's best or most typical member rather than it is to B's.[35]

6.3.6 Similarity

Similarity is a very flexible notion whose meaning depends on the domain within which we apply it.[36] To make similarity a useful mechanism for categorization we have to specify how the similarity measure is determined. There are four major psychological approaches that propose different functions for frameworks: feature- or property-based, geometry-based, alignment-based, and transformational.[37] Each of these psychological definitions or models of similarity has analogues in or can be applied to organizing systems.

An influential model of feature-based similarity calculation is the contrast model proposed by Tversky. This model matches the features or properties of the two things being compared and computes a similarity measure according to three sets of features: those they share, those the first has that the second lacks, and those that the second has that the first lacks. The similarity that results from the set of shared features is reduced by the two sets of distinctive features. The weights or importance assigned to each of these three sets can be adjusted to explain how items are assigned to a set of categories.

We often use a heuristic version of feature-based similarity calculation when we create multi-level or hierarchical category systems to ensure that the categories at each level are at the same level of abstraction or breadth. For example, if we were organizing a collection of musical instruments, it would not seem correct to have subcategories of "woodwind instruments," "violins," and "cellos" because the feature-based similarity among the categories is not the same for all pairwise comparisons among the categories; violins and cellos are simply too similar to each other to be separate categories given woodwinds as a category.

Geometric models are a second type of similarity framework, in which items are represented as points in a multi-dimensional feature- or property-space and similarity is calculated by measuring the distance between them. How distance is measured depends on the type of properties that characterize a domain. When properties that are psychologically or perceptually combined, a Euclidean distance function best accounts for category judgments; but when properties can be conceptually separated, a "city block" distance function works best to explain psychological data, because that ensures that each property value contributes its full amount. Geometric similarity functions are commonly used by search engines; if a query and document are each represented as a vector of search terms, relevance is determined by the distance between the vectors in the "document space." We will discuss how this works in greater detail in Chapter 9.

Alignment-based similarity models have been proposed for domains in which the items to be categorized are characterized by abstract or complex relationships with their features and with each other. For example, some categories are best understood as metaphors that have become conventionalized, and category judgments are made by aligning and projecting aspects of one item or entity onto another. With this model an entity need not be understood as inherently possessing features shared in common with another entity. Rather, people project features from one thing to another in a search for congruities between things, much as clue receivers in the second round of the Pyramid game search for congruities between examples provided by the clue giver in order to guess the target category. For example, a clue like "screaming baby" can suggest many categories, as can "parking meter." But the likely intersection of the interactions one can have with babies and parking meters is that they are both "Things you need to feed."

Transformational models for calculating similarity assume that the similarity between two things is inversely proportional to the complexity of the transformation required to turn one into the other. For example, one way to perform the **name matching** task of determining when two different strings denote the same person, object, or other named

entity is to calculate the "edit distance" between them, the number of changes required to transform one into the other. Two strings with a short edit distance might be variant spellings or misspellings of the same name.[38]

6.3.7 Theory-based Categories

Another principle for creating categories is organizing things in ways that fit a theory or story that makes a particular categorization sensible. A "theory category" can win out even if "family resemblance" or "similarity" with respect to visible properties would lead to a different category assignment. For example, whales are categorized as mammals by biologists even though they share their most visible properties with fish because the theory of "mammalness" emphasizes the property of nursing with mother's milk.

Theory-based categories based on origin or causation are especially important with highly inventive and computational resources because unlike natural kinds of physical resources, little or none of what they can do or how they behave is visible on the surface (see Section 2.4.1, "Affordance and Capability"). Consider all of the different appearances and form factors of the resources that we categorize as "computers" - their essence is that they all compute, an invisible or theory-like principle that does not depend on their visible properties.[39]

6.3.8 Goal-derived Categories

A final principle for creating categories is to organize resources that go together in order to satisfy a goal. Consider the category "Things to take from a burning house," an example that cognitive scientist Lawrence Barsalou termed an **ad hoc** or **goal-derived** category.[40] What things would you take from your house if your neighborhood were burning? Possibly your cat, your wallet and checkbook, your important papers like birth certificates and passports, and grandma's old photo album, and anything else you think is important, priceless, or irreplaceable – as long as you can carry it. These items have almost no discernible properties in common, except for somehow being your most precious possessions. The category is derived or induced by a particular goal in some specified context.

Similarly, a small towel, a music player with headphones, and a bottle of water have no properties in common but they could be organized together because they are members of "things used at the gym when working out" category. This category would fit very well with the many ad hoc categories that gave contestants so much trouble on the Pyramid game show.

6.4 Category Design Choices and Implications

We have previously discussed the most important principles for creating categories: resource properties, similarity, and goals. When we use one or more of these principles to develop a system of categories, we must make decisions about its depth and breadth. Here, we examine the idea that some levels of abstraction in a system of categories are more basic or natural than others. We also consider how the choices we make affect how we

create the organizing system in the first place, and how they shape our interactions when we need to find some resources that are categorized in it.

6.4.1 Category Abstraction and Granularity

We can identify any resource as a unique instance or as a member of a class of resources. The size of this class - the number of resources that are treated as equivalent - is determined by the properties or characteristics we consider when we examine the resources in some domain. The way we think of a resource domain depends on context and intent, so the same resource can be thought of abstractly in some situations and very concretely in others. As we discussed in Chapter 4, this influences the nature and extent of resource description, and as we've seen in this chapter, it then influences the nature and extent of categories we can create.

Consider the regular chore of putting away clean clothes. We can consider any item of clothing as just that – a member of a broad category whose members are any kind of garment that a person might wear. Using one category for all clothing, that is, failing to distinguish among the various items in any useful or practical way would likely mean that we would keep our clothes in a big unorganized pile.

However, we cannot wear any random combination of clothing items – we need a shirt, a pair of pants, socks, and so on. Clearly, our indiscriminate clothing category is too broad for most purposes. So instead, most people organize their clothes in more fine-grained categories that fit the normal pattern of how they wear clothes. For example, everyone probably separates their shirts, pants, and socks when they put away their clothes after doing their laundry. Some pants and shirts may merit wooden hangers; others may rest in special drawers.

In Section 6.3.2 we described an organizing system for the shirts in our closet, so let's talk about socks instead. When it comes to socks, most people think that the basic unit is a pair because they always wear two socks at a time. If you are going to need to find socks in pairs, it seems sensible to organize them into pairs when you are putting them away. Some people might further separate their dress socks from athletic ones, and then sort these socks by color or material, creating a hierarchy of sock categories analogous to the shirt categories in our previous example. We note, parenthetically, that not everyone works this hard when putting their clothes away; some people toss all the single, unpaired socks in a drawer and then rummage around when they need to find a matching pair of socks. People differ in their preferences or tolerances for the amount of granularity in an organizing system and we need to expect and respect these differences.

Questions of resource abstraction and granularity also emerge whenever the information systems of different firms, or different parts of a firm, need to exchange information or be merged into a single system. All parties must define the identity of each thing in the same way, or in ways that can be related or mapped to each other either manually or electronically.

For example, how should a business system deal with a customer's address? Printed on an envelope, "an address" typically appears as a comprehensive, multi-line text object. Inside an information system, however, an address is stored as separate information components for each printed line, or as a set of distinctly identifiable information components. This fine-grained organization makes it easier to sort customers by city or postal codes, for sales and marketing purposes. Incompatibilities in the abstraction and granularity of these information components, and the ways in which they are presented and reused in documents, will cause interoperability problems when businesses need to share information, some of which may be difficult to detect because of the vocabulary problem.[41]

6.4.2 Basic or Natural Categories

We can describe category abstraction in terms of a hierarchy of **superordinate**, **basic**, and **subordinate** category levels. "Clothing," for example, is a superordinate category, "shirts" and "socks" are basic categories, and "white long-sleeve dress shirts" and "white wool hiking socks" are subordinate categories. Members of basic level categories like "shirts" and "socks" have many perceptual properties in common, and are more strongly associated with motor movements than members of superordinate categories. Members of subordinate categories have many common properties, but these properties are also shared by members of other subordinate categories at the same level of abstraction in the category hierarchy. That is, while we can identify many properties shared by all "white long-sleeve dress shirts," many of them are also properties of "blue long-sleeve dress shirts" and "black long-sleeve pullover shirts."

Psychological research suggests that some levels of abstraction in a system of categories are more basic or natural than others. An implication for organizing system design is that basic level categories are highly efficient in terms of the cognitive effort they take to create and use.[42]

6.4.3 The Recall and Precision Tradeoff

The abstraction level we choose determines how precisely we identify resources. When we want to make a general claim, or communicate that the scope of our interest is broad, we use superordinate categories, as when we ask, "How many animals are in the San Diego Zoo?" But we use precise subordinate categories when we need to be specific: "How many adult emus are in the San Diego Zoo today?'

If we return to our clothing example, finding a pair of white wool hiking socks is very easy if the organizing system for socks creates fine-grained categories. When resources are described or arranged with this level of detail, a similarly detailed specification of the resources you are looking for yields precisely what you want. When you get to the place where you keep white wool hiking socks, you find all of them and nothing else. On the other hand, if all your socks are tossed unsorted into a sock drawer, when you go sock hunting you might not be able to find the socks you want and you will encounter lots of socks you do not want. But you won't have put time into sorting them, which many people don't enjoy doing; you can spend time sorting or searching depending on your preferences.

If we translate this example into the jargon of information retrieval, we say that more fine-grained organization reduces **recall**, the number of resources you find or retrieve in response to a query, but increases the **precision** of the recalled set, the proportion of recalled items that are relevant. Broader or coarse-grained categories increase recall, but lower precision. We are all too familiar with this hard bargain when we use a web search engine; a quick one-word query results in many pages of mostly irrelevant sites, whereas a carefully crafted multi-word query pinpoints sites with the information we seek. We will discuss recall, precision, and evaluation of information retrieval more extensively in Chapter 9.

This mundane example illustrates the fundamental tradeoff between organization and retrieval. A tradeoff between the investment in organization and the investment in retrieval persists in nearly every organizing system. The more effort we put into organizing resources, the more effectively they can be retrieved. The more effort we are willing to put into retrieving resources, the less they need to be organized first. The allocation of costs and benefits between the organizer and retriever differs according to the relationship between them. Are they the same person? Who does the work and who gets the benefit?

6.4.4 Category Audience and Purpose

The ways in which people categorize depend on the goals of categorization, the breadth of the resources in the collection to be categorized, and the users of the organizing system. Suppose that we want to categorize languages. Our first step might be determining what constitutes a language, since there is no widespread agreement on what differentiates a language from a dialect, or even on whether such a distinction exists.

What we mean by "English" and "Chinese" and how we view "English" and "Chinese" as categories can change depending on the audience we are addressing and what our purpose is, however.[43] A language learning school's representation of "English" might depend on practical concerns such as how the school's students are likely to use the language they learn, or on which teachers are available. For the purposes of a school teaching global languages, one of the standard varieties of English (which are associated with more political power), or an amalgamation of standard varieties might be thought of as an instance ("English") of the category "Languages."

Similarly, the category structure in which "Chinese" is situated can vary with context. While some schools might not conceptualize "Chinese" as a category encompassing multiple linguistic varieties, but rather as a single instance within the "Languages" category, another school might teach its students Mandarin, Wu, and Cantonese as dialects within the language category "Chinese," that are unified by a single standard writing system. In addition, a linguist might consider Mandarin, Wu, and Cantonese to be mutually unintelligible, making them separate languages within the broader category "Chinese" for the purpose of creating a principled language classification system.

In fact languages can be categorized in multitude of ways. If we are concerned with linguistic diversity and the survival of minority languages, we might categorize some languages as endangered in order to mobilize language preservation efforts. We could also

categorize languages in terms of shared linguistic ancestors ("Romance languages," for example), in terms of what kinds of sounds they make use of, by how well we speak them, by regions they are commonly spoken in, whether they are signed or unsigned, and so on. We could also expand our definition of the languages category to include artificial computer languages, or body language, or languages shared by people and their pets – or thinking more metaphorically, we might include the language of fashion.

If people could only categorize in a single way, the Pyramid game show, where contestants guess what category is illustrated by the example provided by a clue giver, would pose no challenge. The creative possibilities provided by categorization allow people to order the world and refer to interrelationships among conceptions through a kind of allusive shorthand. When we talk about the language of fashion, we suggest that in the context of our conversation, instances like "English," "Chinese," and "fashion" are alike in ways that distinguish them from other things that we would not categorize as languages.

6.5 Implementing Categories in Technologies for Organizing Systems

We have emphasized the intellectual choices and challenges that arise in the design of a system of categories because, at their essence, categories are conceptual or mental constructs. We use categories in a mostly invisible way when we communicate, solve problems, or organize our kitchens and clothes closets. Sometimes categories are more apparent, as when we see signs and labels in the aisles of department or grocery stores to help us find things, when we put our socks and t-shirts in different dresser drawers, or when we create a system of folders and directories in our file cabinets or on our personal computers.

The most visible implementations of a category system are usually those for institutional categories, especially those that are embodied in the organizing systems for information resources where category membership can be verified by technology and the boundaries between categories are precise. In this final section of the chapter we briefly discuss some of the most important technologies for implementing categories, contrasting those that are appropriate for categories where membership is defined using properties with those that work for categories defined on the basis of similarity.

6.5.1 Implementing the Classical View of Categories

The most conceptually simple and straightforward implementation of categories in technologies for organizing system adopts the classical view of categories based on necessary and sufficient features. This approach results in prescriptive categories with explicit and clear boundaries. Classifying items into the categories is objective and deterministic and supports a well-defined notion of validation to determine unambiguously whether some instance is a member of the category.

The most direct way to implement classical categories is as a **decision tree**. A simple decision tree is an algorithm for determining a decision by making a sequence of logical or property tests. For example, we can classify numbers as prime or not with two tests: is it greater than 1, and does it have any divisors other than itself and 1. More complex

categories need more tests. We can model the CAFE fuel economy standards (Section 6.2.4) that assign vehicles to "car," "truck," and "light truck" categories using a decision tree whose differentiating tests are (1) the maximum number of passengers (cars have 10 or fewer), (2) off-road capability (cars do not have it), and (3) weight (trucks weigh at least 6000 pounds), and (4) function (numerous sub-tests that classify vehicles as trucks even if they weigh less than 6000 pounds).[44]

Precisely because natural language embodies cultural categories, it is not the optimal representational format for formally defined institutional categories. Categories defined using natural language can easily be incomplete, inconsistent, or ambiguous, so they are sometimes defined using "simplified writing" or "business rules" that in the aggregate create a decision tree or network that reliably classifies instances.[45] However, the vast majority of institutional category systems are still specified with natural language, despite its ambiguities. Sometimes this is even intentional to allow institutional categories embodied in laws to evolve in the courts and to accommodate technological advances.[46]

Data schemas that specify data entities, elements, identifiers, attributes, and relationships in databases and XML document types on the transactional end of the Document Type Spectrum (Section 3.2.1) are implementations of the categories needed for the design, development and maintenance of information organization systems. Like the classical model of categorization, data schemas tend to rigidly define resources. "Rigid" might sound negative, but a rigidly defined resource is also precisely defined. Precise definition is essential when creating, capturing, and retrieving data and when information about resources in different organizing systems needs to be combined or compared.[47]

The 100 or so standard document types of the Universal Business Language (mentioned briefly in Section 6.2.3) are XML schemas that define basic level categories like orders, invoices, payments, and receipts that many people are familiar with from their personal experiences of shopping and paying bills. UBL's vast library of information components enables the design of very specific or subordinate level transactional document types like "purchase order for industrial chemicals when buyer and seller are in different countries." At the other end of the abstraction hierarchy are document types like "fill-in-the-blank" legal forms for any kind of contract.

In object-oriented programming languages, **classes** are schemas that serve as templates for the creation of objects. A class in a programming language is analogous to a database schema that specifies the structure of its member instances, in that the class definition specifies how instances of the class are constructed in terms of data types and possible values. Programming classes may also specify whether data in a member object can be accessed, and if so, how.[48]

6.5.2 Implementing Categories That Do Not Conform to the Classical Theory

Unlike transactional document types, which can be prescriptively defined as classical categories because they are often produced and consumed by automated processes, narrative document types are usually descriptive in character. We do not classify something as a novel because it has some specific set of properties and content types.

Instead, we have a notion of typical novels and their characteristic properties, and some things that are considered novels are far from typical in their structure and content.[49]

Nevertheless, categories like narrative document types can sometimes be implemented using document schemas that impose only a few constraints on structure and content. Unlike a schema for a purchase order that uses regular expressions, strongly data typed content and enumerated code lists to validate the value of required elements that must occur in a particular order, a schema for a narrative document type would have much optionality, be flexible about order, and expect only text in its sections, paragraphs and headings. Even very lax document schemas can be useful in making content management, reuse, and formatting more efficient.

Category types that are furthest in character from the classical model are those that are not defined using properties in any explicit way. We do not use technology to help us understand cultural categories like "friend" or "game" that rely on some notion of similarity to determine category membership. However, there are technologies that can create a system of categories that uses similarity as its basis.

In particular, **machine learning** is a subfield of computer science that develops and applies algorithms that accomplish tasks that are not explicitly programmed; creating categories and assigning items to them is an important subset of machine learning. Two subfields of machine learning that are particularly relevant to organizing systems are **supervised** and **unsupervised** learning. In supervised learning, a machine learning program is trained by giving it sample items or documents that are labeled by category, and the program learns to assign new items to the correct categories. In unsupervised learning, the program gets the samples but has to come up with the categories on its own by discovering the underlying correlations between the items; that is why unsupervised learning is sometimes called **statistical pattern recognition**. This generally takes longer, since the program isn't given correct answers to use in improving its performance, as it is in the supervised case. As we pointed out in Section 6.2.1, we learn most of our cultural categories without any explicit instruction about them, so it is not surprising that computational models of categorization developed by cognitive scientists often employ unsupervised statistical learning methods. We will now briefly discuss unsupervised learning and return to supervised learning in Chapter 7, "Classification."

There are far too many unsupervised learning techniques for categorization to even mention them all, let alone describe how they work. The ones that are most relevant for us are called **clustering** techniques and they all share the same goal and a few basic methods. The shared goal is to create meaningful categories from a collection of items whose properties are hard to directly perceive and evaluate; this is especially true with large collections of heterogeneous documents, where goals might be to find categories of documents with the same topics, genre, sentiment, or other characteristic that cannot easily be reduced to specific property tests.

The first shared method is that clustering techniques start with an initially uncategorized set of items or documents from which some measures of inter-item similarity can be calculated.[50]

The second shared method is that categories are created by putting items that are most similar into the same category. Hierarchical clustering approaches start with every item in its own category. Other approaches, notably one called "K-means clustering," start with a fixed number of K categories initialized with a randomly chosen item or document.

The third shared method is refining the system of categories by iterative similarity recalculation each time an item is added to a category. Approaches that start with every item in its own category create a hierarchical system of categories by merging the two most similar categories, recomputing the similarity between the new category and the remaining ones, and repeating this process until all the categories are merged into a single category at the root of a category tree. Techniques that start with a fixed number of categories do not create new ones but instead repeatedly recalculate the "center" of the category by adjusting its property representation to the average of all its members after a new member is added.[51]

The end result of clustering is a statistically optimal set of categories in which the similarity of all the items within a category is larger than the similarity of items that belong to different categories. This is a statistical result produced by a computer, and there is no guarantee that the categories are meaningful ones that can be named and used by people. In the end, clustering relies on the data analyst or information scientist to make sense of the clusters if they are to be used to classify resources. In many cases it is better to start with categories created by people and then teach them to computers that can use supervised learning techniques to assign new resources to the categories.

6.6 Key Points in Chapter Six

- Categories are equivalence classes: sets or groups of things or abstract entities that we treat the same.
- The size of the equivalence class is determined by the properties or characteristics we consider.
- We can describe category abstraction in terms of a hierarchy of superordinate, basic, and subordinate category levels.
- Any particular collection of resources can be organized using a combination of intrinisic, extrinsic, static and dynamic resource properties.
- Broader or coarse-grained categories increase recall, but lower precision.
- Some types of categories can be defined precisely with just a few necessary and sufficient properties.
- An important implication of necessary and sufficient category definition is that every member of the category is an equally good member or example of the category.
- Any collection of resources with sortable identifiers (alphabetic or numeric) as an associated property can benefit from using sorting order as an organizing principle.

- A sequence of organizing decisions based on a fixed ordering of resource properties creates a hierarchy, a multi-level category system.
- Sharing some but not all properties is akin to family resemblances among the category members.
- We use properties one at a time to assign category membership.
- We use properties in a composite or integrated way to determine similarity.
- To make similarity a useful mechanism for categorization we have to specify how similarity is measured.
- For most purposes, the most useful property of information resources for categorizing them is their *aboutness*, which is not directly perceivable and which is hard to characterize.
- Cultural, individual, and institutional categorization share some core ideas but they emphasize different processes and purposes for creating categories.
- Languages differ a great deal in the words they contain and also in more fundamental ways by which they organize words into grammatical categories.
- Individual categories are created by intentional activity that usually takes place in response to a specific situation.
- Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories are needed.
- The rigorous definition of institutional categories enables classification, the systematic assignment of resources to categories in an organizing system.
- The most conceptually simple and straightforward implementation of categories in technologies for organizing system adopts the classical view of categories based on necessary and sufficient features.

---

[1] [CogSci] Cataloguing and programming are important activities that need to be done well, and prescriptive advice is often essential. However, we believe that understanding how people create psychological and linguistic categories can help us appreciate that cataloguing and information systems design are messier and more intellectually challenging activities than we might otherwise think.

[2] [CogSci] Cognitive science mostly focuses on the automatic and unconscious mechanisms for creating and using categories. This disciplinary perspective emphasizes the activation of category knowledge for the purpose of making inferences and "going beyond the information given," to use Bruner's (1957) classic phrase. In contrast, the discipline of organizing focuses on the explicit and self-aware mechanisms for creating and using categories because by definition, organizing systems serve intentional and often highly explicit purposes. Organizing systems facilitate inferences about the resources they contain, but the more constrained purposes for which resources are described and arranged makes inference a secondary goal.

Cognitive science is also highly focused on understanding and creating computational models of the mechanisms for creating and using categories. These models blend data-driven or bottom-up processing with knowledge-driven or top-down processing to simulate the time course and results of categorization at both fine-grained scales (as in word or object recognition) and over developmental time frames (as in how children learn categories). The discipline of organizing can learn from these models about the types of properties and principles that organizing systems use, but these computational models are not a primary concern to us in this book.

[3] [CogSci] However, even the way this debate has been framed is a bit controversial. Bulmer's (1970) chicken, the "categories are in the world" position, has been described as empirical, environment-driven, bottom-up, or objectivist, and these not synonymous. Likewise, the "egghead" position that "categories are in the mind"

has been called rational, constructive, top-down, experiential, and embodied – and they are also not synonyms. See also Lakoff (1987), Malt (1995).

[4] [CogSci] Is there a "universal grammar" or a "language faculty" that imposes strong constraints on human language and cognition? Chomsky (1965) and Jackendoff (1997) think so. Such proposals imply cognitive representations in which categories are explicit structures in memory with associated instances and properties. In contrast, generalized learning theories model category formation as the adjustment of the patterns and weighting of connections in neural processing networks that are not specialized for language in any way. Computational simulations of semantic networks can reproduce the experimental and behavioral results about language acquisition and semantic judgments that have been used as evidence for explicit category representations without needing anything like them. Rogers and McClelland (2004) thoroughly review the explicit category models and then show how relatively simple learning models can do without them.

[5] [CogSci] The debates about human category formation also extend to issues of how children learn categories and categorization methods. Most psychologists argue that category learning starts with general learning mechanisms that are very perceptually based, but they don't agree whether to characterize these changes as "stages" or as phases in a more complex dynamical system. Over time more specific learning techniques evolve that focus on correlations among perceptual properties (things with wings tend to have feathers), correlations among properties and roles (things with eyes tend to eat), and ultimately correlations among roles (things that eat tend to sleep). See (Smith and Thelen 2003).

[6] [CogSci] These three contexts were proposed by Glushko, Maglio, Matlock, and Barsalou 2008), who pointed out that cognitive science has focused on cultural categorization and largely ignored individual and institutional contexts. They argue that taking a broader view of categorization highlights dimensions on which it varies that are not apparent when only cultural categories are considered. For example, institutional categories are usually designed and maintained using prescriptive methods that have no analogues with cultural categories.

[7] [CogSci] This quote comes from Plato's Phaedrus dialogue, written around 370 BCE. Contemporary philosophers and cognitive scientists in discussions about whether "natural kinds" exist commonly invoke it. For example, see Campbell, O'Rourke, and Slater (2011), and Hutchins (2010). Atran (1987) and others have argued that the existence of perceptual discontinuities is not sufficient to account for category formation. Instead, people assume that members of a biological category must have an essence of co-occurring properties and these guide people to focus on the salient differences, thereby creating categories. Property clusters enable inferences about causality, which then builds a framework on which additional categories can be created and refined. For example, if "having wings" and "flying" are co-occurring properties that suggest a "bird" category, wings are then inferred as the causal basis of flying, and wings become more salient.

[8] [CogSci] Pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, particles, and numerals and other"parts of speech" are also grammatical categories, but nouns carry most of the semantic weight.

[9] [CogSci] In contrast, the set of possible interactions with even a simple object like a banana is very large. We can pick, peel, slice, smash, eat, or throw a banana, so instead of capturing this complexity in the meaning of banana it gets parceled into the verbs that can act on the banana noun. Doing so requires languages to use verbs to capture a broader and more abstract type of meaning that is determined by the nouns they are combined with. Familiar verbs like "set", "put", and "get" have dozens of different senses as a result because they go with so many different nouns. We set fires and we set tables, but fires and tables have little in common. The intangible character of verbs and the complexity of multiple meanings make it easier to focus instead on their associated nouns, which are often physical resources, and create organizing systems that emphasize the latter rather than the former. We create organizing systems that focus on verbs when we are categorizing actions, behaviors, or services where the resources that are involved are less visible or less directly involved in the supported interactions.

[10] [CogSci] This analysis comes from (Haviland, 1998). More recently, Lera Boraditsky has done many interesting studies and experiments about linguistic relativity. See Boraditksy (2003) for an academic summary and Boraditsky (2010, 2011) for more popular treatments.

[11] [CogSci] Many languages have a system of grammatical gender in which all nouns must be identified as masculine or feminine using definite articles (el and la in Spanish, le and la in French, and so on) and corresponding pronouns. Languages also contrast in how they describe time, spatial relationships, and in which things are treated as countable objects (one ox, two oxen) as opposed to substances or mass nouns that do not have distinct singular and plural forms (like water or dirt). Deutscher (2011) carefully reviews and discredits the strong Whorfian view and makes the case for a more nuanced perspective on linguistic relativity. He also reviews much of Lera Boraditsky's important work in this area. George Lakoff's book with the title "Women, Fire, and Dangerous Things" (1987) provocatively points out differences in gender rules among languages; in an aboriginal language called Dyirbal many dangerous things, including fire have feminine gender, meanwhile "fire" is masculine in Spanish (el feugo) and French (le feu).

[12] [Citation] Medin et al (1997).

[13] [LIS] The personal archives of people who turn out to be famous or important are the exception that proves this rule. In that case, the individual's organizing system and its categories are preserved along with their contents.

[14] [Law] Consider how the cultural category of "killing a person" is refined by the legal system to distinguish manslaughter and different degrees of murder based on the amount of intentionality and planning involved (e.g., first and second degree murder) and the roles of people involved with the killing (accessory). In general, the purpose of laws is to replace coarse judgments of categorization based on overall similarity of facts with rule-based categorization based on specific dimensions or properties.

[15] [Business] The most "standard" of all standards organization is the International Organization for Standardization (ISO), whose members are themselves national standards organizations, which as a result gives the nearly 20,000 ISO standards the broadest and most global coverage. See ISO.org. In addition, there are scores of other national and industry-specific standards bodies whose work is potentially relevant to organizing systems of the sorts discussed in this book. We encounter these kinds of standards every day in codes for countries, currencies, and airports, in file formats, in product barcodes, and in many other contexts. For example, the familiar MARC record format used in online library catalogs is defined in ISO standard 2709, with its American counterpart ANSI Z39.2.

[16] [LIS] The Dewey Decimal System is the world's most widely used library classification system, but most people do not realize that it is proprietary and it is maintained and licensed for use by the Online Computer Library Center (OCLC). See http://www.oclc.org/dewey/DDC, Similarly, the DSM is maintained and published by the American Psychiatric Association and it makes the APA many millions of dollars a year.

[17] [Business] Work on UBL has gone on for over a decade in a technical committee under the auspices of a standards development consortium called the Organization for the Advancement of Open Information Standards (OASIS), which has developed scores of standards for web services and information-intensive industries. All the finished work of OASIS is freely available at https://www.oasis-open.org; the UBL committee is at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ubl

[18] [Business] And often the particularities or idiosyncrasies of individual categorization systems capture user expertise and knowledge that is not represented in the institutional categories that replace them. Many of the readers of this book are information professionals whose technological competence is central to their work and which helps them to be creative. But for a great many other people, information technology has enabled the routinization of work in offices, assembly lines, and in other jobs where new institutionalized job categories have "downskilled" or "deskilled" the nature of work, destroying competence and engendering a great deal of resistance from the affected workers.

[19] [Business] Similar technical concerns arise in within-company and multi-company standardization efforts, but the competitive and potentially anti-competitive character of the latter imposes greater complexity by introducing considerations of business strategy and politics. Credible standards-making in multi-company contexts depends on an explicit and transparent process for gathering and prioritizing requirements, negotiating specifications that satisfy them, and ensuring conformant implementations – without at any point giving any participating firm an advantage. See the OASIS Technical Committee Process for an example (OASIS, 2012) and Rosenthal et al (2004) for an analysis of best practices.

[20] [Citation] http://www.wired.com/geekmom/2012/03/the-periodic-tables-of-everything-but-elements/

[21] [Business] The Corporate Average Fuel Economy (CAFE) standards have been developed by the US National Highway Traffic Safety Administration (http://www.nhtsa.gov/fuel-economy) since 1975. For a careful and critical assessment of CAFE, including the politics of categorization for vehicles like the PT Cruiser, see the report from the Committee on the Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards, National Research Council (2002).

[22] [Citation] The IAU (iau.org) published its new definition of planet in August 2006. This changed definition of a significant cultural category generated a great deal of controversy and angst among ordinary non-scientific people. A typical headline was "Pluto's demotion has schools spinning," describing the outcry from elementary school students and teachers about the injustice done to Pluto and the disruption on the curriculum. A public television documentary in 2011 called "The Pluto Files" retells the story (Nova 2011).

[23] [Citation] The distinction between intension and extension was introduced by Gottlob Frege, a German philosopher and mathematician (Frege 1892). You might be thinking here that enumeration or extensional definition of a category is also a property test; is not "being a state" a property of California? But statehood is not a property precisely because "state" is defined by extension, which means the only way to test California for statehood is to see if it is in the list of states.

[24] [CogSci] The number of resources in each of these categories depends on the age of the collection and the collector. We could be more precise here and say "single atomic property" or otherwise more carefully define "property" in this context as a characteristic that is basic and not easily or naturally decomposable into other characteristics. It would be possible to analyze the physical format of a music resource as a composition of size, shape, weight, and material substance properties, but that is not how people normally think. Instead, they treat physical format as a single property as we do in this example.

[25] [CogSci] We need to think of alphabetic ordering or any other organizing principle in a logical way that does not imply any particular physical implementation. Therefore, we do not need to consider which of these alphabetic categories exist as folders, files, or other tangible partitions.

[26] [CogSci] Another example: rules for mailing packages might use either size or weight to calculate the shipping cost, and whether these rules are based on specific numerical values or ranges of values, the intent seems to be to create categories of packages.

[27] [CogSci] If you try hard, you can come up with situations in which this property is important, as when the circus is coming to the island on a ferry or when you are loading an elevator with a capacity limit of 5000 pounds, but it just isn't a useful or psychologically salient property in most contexts.

[28] [Computing] Many information systems, applications, and programming languages that work with hierarchical categories take advantage of this logical relationship to infer inherited properties when they are needed rather than storing them redundantly.

[29] [Business] Similarly, clothing stores use intrinsic static properties when they present merchandise arranged according to color and size; extrinsic static properties when they host branded displays of

merchandise; intrinsic dynamic properties when they set aside a display for seasonal merchandise, from bathing suits to winter boots; and extrinsic dynamic properties when a display area is set aside for "Today's Special".

[30] [Citation] Aristotle did not call them classical categories. That label was bestowed about 2300 years later by Smith and Medin (1981).

[31] [LIS] We all use the word "about" with ease in ordinary discourse, but "aboutness" has generated a surprising amount of theoretical commentary about its typically implicit definition, starting with Hutchins (1977) and Maron (1977) and relentlessly continued by Hjorland (1992, 2001).

[32] [CogSci] Typicality and centrality effects were studied by Rosch and others in numerous highly influential experiments in the 1970s and 1980s. Good summaries can be found in Mervis and Rosch (1981), Rosch (1999), and in chapter 1 of Rogers and McClelland (2004).

[33] [Citation] An easy to find source for Wittgenstein's discussion of "game" is Wittgenstein (2002) in a collection of core readings for cognitive psychology (Levitin 2002).

[34] [Citation] The philosopher's poll that ranked Wittgenstein's book #1 is reported by Lackey (1999).

[35] [CogSci] The exact nature of the category representation to which the similarity comparison is made is a subject of ongoing debate in cognitive science. Is it a **prototype**, a central tendency or average of the properties shared by category members, or it one or more **exemplars,** particular members that typify the category. Or is it neither, as argued by connectionist modelers who view categories as patterns of network activation without any explicitly stored category representation? Fortunately, these distinctions do not matter for our discussion here. A recent review is Rips, Smith, and Medin (2012).

[36] [CogSci] Some people consider that the concept of similarity is itself meaningless because there must always be some basis, some unstated set of properties, for determining whether two things are similar. If we could identify those properties and how they are used, there would not be any work for a similarity mechanism to do. Another situation where similarity has been described as a "mostly vacuous" explanation for categorization is with abstract categories or metaphors. Goldstone says "an unrewarding job and a relationship that can't be ended may both be metaphorical prisons… and may seem similar in that both conjure up a feeling of being trapped… but this feature is almost as abstract as the category to be explained." Goldstone (1994), p. 149

[37] [Citation] Goldstone (2004).

[38] [Computing] The "strings" to be matched can themselves be transformations. The "soundex" function is very commonly used to determine if two words could be different spellings of the same name. It "hashes" the names into phonetic encodings that have fewer characters than the text versions. See Christen (2006) and http://www.searchforancestors.com/utility/soundex.html to try it yourself.

[39] [CogSci] The emergence of theory-based categorization is an important event in cognitive development that has been characterized as a shift from "holistic" to "analytic" categories or from "surface properties" to "principles." See (Keil epigenesis of mind), (Rehder and Hastie, 2004).

[40] [Citation] Barsalou (1983).

[41] [Computing] Consider what happens if two businesses model the concept of "address" in a customer database with different granularity. One may have a coarse "Address" field in the database, which stores a street address, city, state, and Zip code all in one block, while the other stores the components "StreetAddress," "City," and "PostalCode" In separate fields. The more granular model can be automatically transformed into the less granular one, but not vice versa (Glushko and McGrath, 2005).

[42] [CogSci] Rosch (1978) calls this the principle of cognitive economy, that "what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible. […] It is to the organism's advantage not to differentiate one stimulus from another when that differentiation is irrelevant to the purposes at hand" (pages 3-4).

[43] [CogSci] For example, some linguists think of "English" as a broad category encompassing multiple languages or dialects, such as "Standard American English," "Appalachian English," and "Standard British English."

[44] [CogSci] Even though they are classical categories, we might also model goal-derived categories as decision trees by ordering the decisions to ensure that any sub-goals are satisfied according to their priority. We could understand the category "Things to take from a burning house" by first asking the question "Are there living things in the house?" because that might be the most important sub-goal. If the answer to that question is "yes," we might proceed along a different path than if the answer is "no." Similarly, we might put a higher priority on things that cannot be replaced (Grandma's photos) than those that can (passport).

[45] [CogSci] Institutional uses of decision trees can also sometimes be thought of as models of goal-derived categories. For example, the US Department of Health and Human Services (HHS) uses several decision trees as part of its efforts to ensure that research programs funded by the department do not harm human subjects. The chart http://www.hhs.gov/ohrp/humansubjects/guidance/decisioncharts.htm#c1for example, is used to determine whether a program can be classified as research involving human subjects, which would mean that the program would have to be reviewed by an Institutional Review Board (IRB).

[46] [Law] When the US Congress revised copyright law in 1976 it codified a "fair use" provision to allow for some limited uses of copyrighted works, but fair use in the digital era is vastly different today; web site caching to improve performance and links that return thumbnail versions of images are fair uses that were not conceivable when the law was written. A law that precisely defined fair uses using contemporary technology would have quickly become obsolete, but one written more qualitatively to enable interpretation by the courts has remained viable. See (Samuelson 2009).

[47] [Computing] For example, in a traditional relational database, each table contains a field, or combination of fields, known as a primary key, which is used to define and restrict membership in the table. A table of email messages in a database might define an email message as a unique combination of sender address, recipient address, and date/time when the message was sent, by enforcing a primary key on a combination of these fields. Similar to category membership based on a single, monothetic set of properties, membership in this email message table is based on a single set of required criteria. An item without a recipient address cannot be admitted to the table. In categorization terms, the item is not a member of the "email message" class because it does not have all the properties necessary for membership.

[48] [Computing] Like data schemas, programming classes specify and enforce rules in the construction and manipulation of data. However, programming classes, like other implementations that are characterized by specificity and rule enforcement, can vary widely in the degree to which rules are specified and enforced. While some class definitions are very rigid, others are more flexible. Some languages have abstract types that have no instances but serve to provide a common ancestor for specific implemented types.

[49] [CogSci] The existence of chapters might suggest that an item is a novel; however, a lack of chapters need not automatically indicate that an item is not a novel. Some novels are hypertexts that encourage readers to take alternative paths. Many of the writings by James Joyce and Samuel Beckett are "stream of consciousness" works that lack a coherent plot yet they are widely regarded as novels.

[50] [Computing] Some approaches represent each item as a vector of property values; documents are usually represented as vectors of frequency-weighted terms; in either case the items can be represented as points in

a multidimensional space of these properties. Similarity can then be calculated by measuring the distance between items in this space. A popular text on machine learning is Witten, Frank, and Hall (2011).

Other approaches start more directly with the similarity measure, obtained either by direct judgments of the similarity of each pair of items or by indirect measures like the accuracy in deciding whether two sounds, colors, or images are the same or different. The assumption is that the confusability of two items reflects how similar they are.

[51] [Computing] Unlike hierarchical clustering methods that have a clear stopping rule when they create the root category, k-means clustering methods run until the centroids of the categorize stabilize. Furthermore, because the k-means algorithm is basically just hill-climbing, and the initial category "seed" items are random, it can easily get stuck in a local optimum. So it is desirable to try many different starting configurations for different choices of K.