atively 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.

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.

7.2.5 Computational Categories

Computational categories are created by computer programs when the number of resources, or when the number of descriptions or observations associated with each resource, are so large that people cannot think about them effectively. Computational categories are created for information retrieval, predictive analytics, and other applications where information scale or speed requirements are critical. The resulting categories are similar to those created by people in some ways but differ substantially in other ways.

The simplest kind of computational categories can be created using descriptive statistics (see §3.3.4). Descriptive statistics do not identify the categories they create by giving them familiar cultural or institutional labels. Instead, they create implicit categories of items according to how much they differ from the most typical or frequent ones. For example, in any dataset where the values follow the normal distribution, statistics of central tendency and dispersion serve as standard reference measures for any observation. These statistics identify categories of items that are very different or statistically unlikely outliers, which could be signals of measurement errors, poorly calibrated equipment, employees who are inadequately trained or committing fraud, or other problems. The “Six Sigma” methodology for process improvement and quality control rests on this idea that careful and consistent collection of statistics can make any measurable operation better.

Many text processing methods and applications use simple statistics to categorize words by their frequency in a language, in a collection of documents, or in individual documents, and these categories are exploited in many information retrieval applications (see §10.4.1 and §10.4.2).

Categories that people create and label also can be used more explicitly in computational algorithms and applications. In particular, a program that can assign an item or instance to one or more existing categories is called a classifier. The subfield of computer science known as *machine learning* is home to numerous techniques for creating classifiers by training them with already correctly categorized examples. This training is called *supervised learning*; it is supervised

CAFE Standards: Blurring the Lines Between Categorization Perspectives

The *Corporate Average Fuel Economy (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 introduced, 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.[407][Bus]

because it starts with instances labeled by category, and it involves learning because over time the classifier improves its performance by adjusting the weights for features that distinguish the categories. But strictly speaking, supervised learning techniques do not learn the categories; they implement and apply categories that they inherit or are given to them. We will further discuss the computational implementation of categories created by people in §7.5.

Supervised and Unsupervised Learning

Two subfields of *machine learning* that are relevant to organizing systems are *supervised* and *unsupervised* learning. In *supervised learning*, a machine learning program is trained with 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 same items 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*.

In contrast, many computational techniques in machine learning can analyze a collection of resources to discover statistical regularities or correlations among the items, creating a set of categories without any labeled training data. This is called *unsupervised learning* or *statistical pattern recognition*. As we pointed out in §7.2.1 Cultural Categories (page 353), 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.

Many computational categories are like individual categories because they are tied to specific collections of resources or data and are designed to satisfy narrow goals. The individual categories you use to organize your email inbox or the files on your computer reflect your specific interests, activities, and personal network and are surely different than those of anyone else. Similarly, your credit card company analyzes your specific transactions to create computational categories of "likely good" and "likely fraudulent" that are different for every cardholder.

This focused scope is obvious when we consider how we might describe a computational category. "Fraudulent transaction for cardholder 4264123456780123" is not lexicalized with a one-word label as familiar cultural categories are. "Door" and "window" have broad scopes that are not tied to a single purpose. Put another way, the "door" and "window" cultural categories are highly reusable, as are institutional categories like those used to collect economic or health data that can be analyzed for many different purposes. The definitions of "door" and "window" might be a little fuzzy, but institutional categories are more precisely defined, often by law or regulation. Examples are the *North American Industry Classification System (NAICS)* from the US Census Bureau and the *United Nations Standard Products and Services Code (UNSPC)*.

A final contrast between categories created by people and those created computationally is that the former can almost always be inspected and reasoned about by other people, but only some of the latter can. A computational model that categorizes loan applicants as good or poor credit risks probably uses proper-

ties like age, income, home address, and marital status, so that a banker can understand and explain a credit decision. However, many other computational categories, especially those that created by clustering and deep learning techniques, are inseparable from the mathematical model that learned to use them, and as a result are uninterpretable by people.

A machine learning algorithm for classifying objects in images creates a complex multi-layer neural network whose features have no clear relationship to the categories, and this network has no other use. Put another way, machine learning programs are very general because they can be employed in any domain with high dimensional data, but what they learn cannot be applied in any other domain.

7.3 Principles for Creating Categories

§7.2 The What and Why of Categories (page 351) 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.

We now take a systematic look at principles for creating categories, including: enumeration, single properties, multiple properties and hierarchy, probabilistic, similarity, and theory- and goal-based categorization. These ways of creating categories differ in the information and mechanisms they use to determine category membership.

7.3.1 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 United States or the ISO currency codes (ISO 4217).

Enumerative categories enable membership to be unambiguously determined because a value like state name or currency code is either a member of the category or it is not. However, this clarity has a downside; it makes it hard to argue that something not explicitly mentioned in an enumeration should be considered a member of the category, which can make laws or regulations inflexible. Moreover, 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.[408][Law]

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