The intentional arrangement of tools in a working kitchen might look something like Table 8.1:

Table 8.1. A cook's taskonomy

Prep	Oven	Stove
Poultry knife	Oven mitts	Pots and pans
Paring knife	Baking sheets	Wooden spoons
Vegetable knife	Aluminum foil	Wok
Cutting board	Parchment paper	
	Roasting pan	

8.6 Computational Classification

Because of its importance, ubiquity, and ease of processing by computers, it should not be surprising that a great many computational classification problems involve text. Some of these problems are relatively simple, like identifying the language in which a text is written, which is solved by comparing the probability of one, two, and three character-long contiguous strings in the text against their probabilities in different languages. For example, in English the most likely strings are “the”, “and”, “to”, “of”, “a”, “in”, and so on. But if the most likely strings are “der”, “die”, “und”, and “den” the text is German and if they are “de”, “la”, “que”, “el”, and “en” the text is Spanish.

Stop and Think: Office Taskonomy

Think about your personal office space. It may be an interesting hybrid space—it probably contains documents that could be classified in a hierarchical system, but it is also a work space that could lend itself to “taskonomy” organization. Which does it more closely resemble? How have any conflicts between hierarchy and “taskonomy” been resolved?

More challenging text classification problems arise when more features are required to describe each instance being classified and where the features are less predictable. The unknown author of a document can sometimes be identified by analyzing other documents known to be written by him to identify a set of features like word frequency, phrase structure, and sentence length that create a “writeprint” analogous to a fingerprint that uniquely identifies him. This kind of analysis was used in 2013 to determine that *Harry Potter* author J. K. Rowling had written a crime fiction novel entitled *The Cuckoo's Calling* under the pseudonym Robert Galbraith.[514][Com]

Another challenging text classification problem is sentiment analysis, determining whether a text has a positive or negative opinion about some topic. Much academic and commercial research has been conducted to understand the sentiment of Twitter tweets, Facebook posts, email sent to customer support applications, and other similar contexts. Sentiment analysis is hard because messages are often short so there is not much to analyze, and because and because sarcasm, slang, clichés, and cultural norms obscure the content needed to make the classification.

A crucial consideration whenever supervised learning is used to train a classifier is ensuring that the training set is appropriate. If we were training a classifier to detect spam messages using email from the year 2000, the topics of the emails, the words they contain, and perhaps even the language they are written in would be substantially different than messages from this year. Up to date training data is especially important for the classification algorithms used by Twitter, Facebook, YouTube, and similar social sites that classify and recommend content based on popularity trends.

When the relevant training data is constantly changing and there is a great deal of it, there is a risk that by the time a model can learn to classify correctly it is already out of date. This challenge has led to the development of streaming algorithms that operate on data as it comes in, using it as a live data source rather than as a static training set. Streaming algorithms are essential for tackling datasets that are too large to store or for models that must operate under intense time pressure. Streaming approaches complement rather than replace those that work with historical datasets because they make different tradeoffs between accuracy and speed. The streaming system might provide real-time alerting and recommendations, while historical analyses are made on the batch-oriented system that works with the entire data collection.[515][Com]

Stop and Think: Sentiment Analysis

Sometimes, a text message might seem complimentary, but really is not. Is the customer happy if he tweets “Nice job, United. You only lost one of my bags this time.” Think of some other short messages where sarcasm or slang makes sentiment analysis difficult. How would you write a product or service review that is unambiguously positive, negative, or neutral? How would you write a review whose sentiment is difficult to determine?

How a computational classifier “learns” depends on the specific machine learning algorithm. Decision trees, Naive Bayes, support vector machines, and neural net approaches were briefly described in §7.5 Implementing Categories (page 387).