Let's consider a platform with a set of users and a set of sources that they can choose to consume. On each visit, user i chooses one of the sources to consume; let's suppose that user i chooses source j on their visit with probability $p(i, j)$. We can write this of probabilities as a table as shown in Figure 1, with rows corresponding to users, columns corresponding to sources, and the number in the cell for row i and column j equal to selection probability $p(i, j)$ that user i chooses source j.

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

Figure 1: .

What's happening with these probabilities? It's hard to say that one source has universally more appeal than another, given that user 1 has a high probability of consuming source 1 and never consumes source 4, while the situation is reversed for user 5. But it's hard to say that the sources are completely partitioned either, since people like user 3 seem to consume everything. So is there any more basic structure that we can find in these probabilities?

In fact, we can achieve a lot of descriptive power, and in particular clarify what's happening in Figure 1, if we add one more thing to the model — the idea that sources can belong to multiple *genres*. For purposes of the model we will keep the idea of a genre abstract, but in different application domains genres can encode intuitively natural groupings: different styles of music, different political viewpoints in news sources, different national cuisines in restaurants, and many other categorizations.

Here is how we'll model a user's selection when there are multiple genres. For simplicity, we'll describe things in the case of two genres, and then we'll observe that the same process can work for any number of genres.

- The basic idea is that a user's selection of a source now consists of two steps: first the user chooses which genre they'd like to select from, and then the user selects a source from this genre.

- This means that for each user i, instead of a single activity level, we specify their probability of selecting from each genre: a probability $u_i[1]$ that they select from genre

1, and a probability $u_i[2]$ that they select from genre 2. These numbers add up to at most 1, and with the remaining probability $1 - u_i[1] - u_i[2]$ they leave the platform without selecting anything.

- Next, a given genre specifies a probability for each source. If a user decides to select from genre 1, then they choose source j with probability $s_j[1]$; if instead they decide to select from genre 2, they choose source j with probability $s_j[2]$.

Finally, notice that there's nothing requiring us to have only two genres in this description: we could describe exactly the same process for any number of genres k by saying that user i has probabilities $u_i[1], u_i[2], \ldots, u_i[k]$ of selecting from each of the k genres (where these probabilities add up to at most 1), and then for each source j and genre g, source j has probability $s_j[g]$ of being chosen when a user is selecting from genre g.

Despite the abstract formulation, this selection process is capturing something that we all engage in when we choose things to consume. For example, continuing with the example of restaurant orders, we might decide that we'd like to order Thai food, and then we select from the genre of Thai restaurants; alternately, we might decide that we want a fast-food restaurant where we can pick up our order from a counter and leave, and then we select from the genre of fast-food restaurants. And notice how the genres can overlap: a fast-food Thai restaurant could get chosen for two distinct reasons: one night we might order from it because we're looking for Thai food, and another night we might order from this same restaurant because we're looking for fast food. There are many other stories we can tell like this: for example, a user who watches a college football video might be a fan of football, or they might be a high-school student applying to colleges and trying to learn about life at college from videos. These represent two genres that provide fundamentally different ways of arriving at the same video.

A schematic view of the selection process It's useful to depict the selection process in a diagram as well, as shown in Figure 2. Here we see the two steps a user takes to select a source: from left to right, the first branch represents user i selecting a genre, and then the second branch represents the choice of a source from this genre. And we see in Figure 2 how an item can be chosen for multiple reasons (as in the example of the fast-food Thai restaurant above): there are two paths through the diagram from user i to source j, one via genre 1 and the other via genre 2.

These multiple paths also make clear how to determine the overall probability that user i selects source j: the probability that user i selects source j via genre 1 is the product of the two probabilities on the path through genre 1, $u_i[1]s_j[1]$, and the probability that user i selects source j via genre 2 is similarly the product $u_i[2]s_j[2]$. These two ways of selecting source j are non-overlapping events (i.e. they can't both happen, since user i chooses only

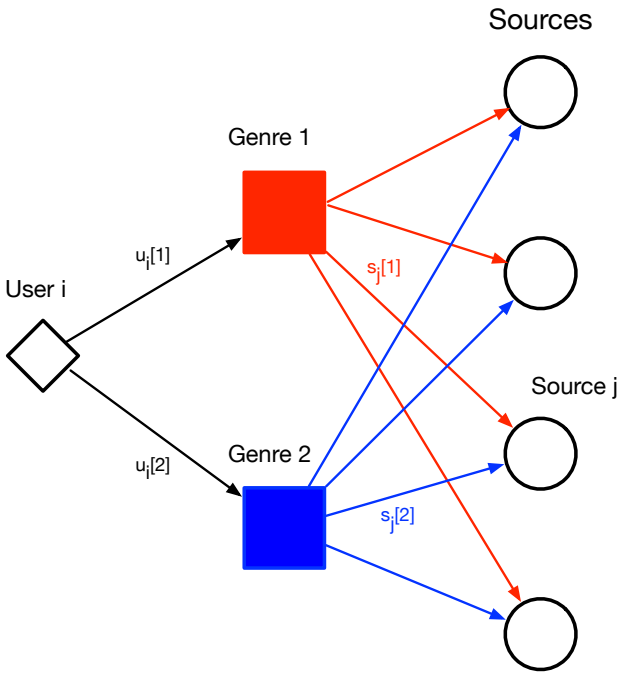

Figure 2: .

one genre for each selection), and so we can directly add their probabilities to get the overall probability $p(i, j)$ that user i selects source j: it is $p(i, j) = u_i[1]s_j[1] + u_i[2]s_j[2]$.

Notice that in this model, the genres are implicit and in a sense invisible to people observing the operation of the platform. We only see users arrive and select sources to consume; the reason why they make these selections, or the fact that there might be two distinct reasons why they could select a particular source, is all taking place below the surface. For this reason, we refer to the overall model as a *latent factor model*: it consists of multiple factors that go into a user's decision (the genres), and they are latent in the sense that we can't directly observe them.

However, even if we can't observe the latent genres directly, we can try to infer genres from the data. For example, suppose we went back to the information in Figure 1 and asked: is there a set of two genres, each with a probability distribution for the sources, and a set of probabilities by which users select genres, that would produce the probabilities $p(i, j)$ that we see in the figure? The answer in this case is "yes", using the following values:

- We let genre 1 have probabilities for the sources given by $s_1[1] = 0.4, s_2[1] = 0.4, s_3[1] = 0.2, s_4[1] = 0$ and let genre 2 have probabilities $s_1[1] = 0, s_2[1] = 0.2, s_3[1] = 0.4, s_4[1] = 0.4$.

- We let user 1 have probabilities over the two genres equal to $u_1[1] = 1$ and $u_1[2] = 0$;

user 2 will have $u_2[1] = 0.5$ and $u_2[2] = 0$; user 3 will have $u_3[1] = 0.5$ and $u_3[2] = 0.5$; user 4 will have $u_4[1] = 0$ and $u_4[2] = 0.5$; and user 5 will have $u_5[1] = 0$ and $u_5[2] = 1$.

- We can now work out that for every user i and every source j, the numbers given above satisfy the equation $p(i, j) = u_i[1]s_j[1] + u_i[2]s_j[2]$. Just to choose one worked example, $p(3, 2) = 0.5 \cdot 0.4 + 0.5 \cdot 0.2 = .2 + .1 = .3$.