

There's nothing wrong with not knowing what a word means, but there is something wrong with not asking! You will likely find that asking clarifying questions about vocabulary gets you even more insight into the underlying data problem.

3. Simulation is a useful technique in data science. It can be useful practice to simulate fake datasets from a model to understand the generative process better, for example, and also to debug code.

Case Study: RealDirect

Doug Perlson, the CEO of RealDirect, has a background in real estate law, startups, and online advertising. His goal with RealDirect is to use all the data he can access about real estate to improve the way people sell and buy houses.

Normally, people sell their homes about once every seven years, and they do so with the help of professional brokers and current data. But there's a problem both with the broker system and the data quality. RealDirect addresses both of them.

First, the brokers. They are typically “free agents” operating on their own—think of them as home sales consultants. This means that they guard their data aggressively, and the really good ones have lots of experience. But in the grand scheme of things, that really means they have only slightly more data than the inexperienced brokers.

RealDirect is addressing this problem by hiring a team of licensed real-estate agents who work together and pool their knowledge. To accomplish this, it built an interface for sellers, giving them useful data-driven tips on how to sell their house. It also uses interaction data to give real-time recommendations on what to do next.

The team of brokers also become data experts, learning to use information-collecting tools to keep tabs on new and relevant data or to access publicly available information. For example, you can now get data on co-op (a certain kind of apartment in NYC) sales, but that's a relatively recent change.

One problem with publicly available data is that it's old news—there's a three-month lag between a sale and when the data about that sale is available. RealDirect is working on real-time feeds on things like when people start searching for a home, what the initial offer is, the time between offer and close, and how people search for a home online.

Ultimately, good information helps both the buyer and the seller. At least if they're honest.

How Does RealDirect Make Money?

First, it offers a subscription to sellers—about \$395 a month—to access the selling tools. Second, it allows sellers to use RealDirect's agents at a reduced commission, typically 2% of the sale instead of the usual 2.5% or 3%. This is where the magic of data pooling comes in: it allows RealDirect to take a smaller commission because it's more optimized, and therefore gets more volume.

The site itself is best thought of as a platform for buyers and sellers to manage their sale or purchase process. There are statuses for each person on site: active, offer made, offer rejected, showing, in contract, etc. Based on your status, different actions are suggested by the software.

There are some challenges they have to deal with as well, of course. First off, there's a law in New York that says you can't show all the current housing listings unless those listings reside behind a registration wall, so RealDirect requires registration. On the one hand, this is an obstacle for buyers, but serious buyers are likely willing to do it. Moreover, places that don't require registration, like Zillow, aren't true competitors to RealDirect because they are merely showing listings without providing any additional service. Doug pointed out that you also need to register to use Pinterest, and it has tons of users in spite of this.

RealDirect comprises licensed brokers in various established realtor associations, but even so it has had its share of hate mail from realtors who don't appreciate its approach to cutting commission costs. In this sense, RealDirect is breaking directly into a guild. On the other hand, if a realtor refused to show houses because they are being sold on RealDirect, the potential buyers would see those listings elsewhere and complain. So the traditional brokers have little choice but to deal with RealDirect even if they don't like it. In other words, the listings themselves are sufficiently transparent so that the traditional brokers can't get away with keeping their buyers away from these houses.

Doug talked about key issues that a buyer might care about—nearby parks, subway, and schools, as well as the comparison of prices per square foot of apartments sold in the same building or block. This is

the kind of data they want to increasingly cover as part of the service of RealDirect.

Exercise: RealDirect Data Strategy

You have been hired as chief data scientist at *realdirect.com*, and report directly to the CEO. The company (hypothetically) does not yet have its data plan in place. It's looking to you to come up with a data strategy. Here are a couple ways you could begin to approach this problem:

1. Explore its existing website, thinking about how buyers and sellers would navigate through it, and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of RealDirect user-behavior data could be used to inform decision-making and product development. Come up with a list of research questions you think could be answered by data:
 - What data would you advise the engineers log and what would your ideal datasets look like?
 - How would data be used for reporting and monitoring product usage?
 - How would data be built back into the product/website?
2. Because there is no data yet for you to analyze (typical in a start-up when it's still building its product), you should get some auxiliary data to help gain intuition about this market. For example, go to https://github.com/oreillymedia/doing_data_science. Click on Rolling Sales Update (after the fifth paragraph).

You can use any or all of the datasets here—start with Manhattan August, 2012–August 2013.

- First challenge: load in and clean up the data. Next, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.
- Once the data is in good shape, conduct exploratory data analysis to visualize and make comparisons (i) across neighborhoods, and (ii) across time. If you have time, start looking for meaningful patterns in this dataset.

3. Summarize your findings in a brief report aimed at the CEO.
4. Being the “data scientist” often involves speaking to people who aren’t also data scientists, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?
5. Most of you are not “domain experts” in real estate or online businesses.
 - Does stepping out of your comfort zone and figuring out how you would go about “collecting data” in a different setting give you insight into how you do it in your own field?
 - Sometimes “domain experts” have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn’t understand (“comps,” “open houses,” “CPC”)? Sometimes if you don’t understand vocabulary that an expert is using, it can prevent you from understanding the problem. It’s good to get in the habit of asking questions because eventually you will get to something you do understand. This involves persistence and is a habit to cultivate.
6. Doug mentioned the company didn’t necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain.