Chapter 2 - Research Questions

2.1 What Is a Research Question?

COMING UP WITH A QUESTION IS EASY. Just ask any five-year-old and they can provide you with dozens. Coming up with a good research question is much harder.

What's the difference? The difference, at least in the case of quantitative empirical research, is that a research question is a question *that can be answered*, and for which having that answer will *improve your understanding of how the world works*.

Those are both a little abstract. Let's take them one at a time.

What does it mean to have a question *that can be answered*? It means that it's possible for there to be some set of evidence in the world that, if you found that evidence, your question would have a believable answer. So for example, "what is the best James Bond movie?" can't really be answered.[1] No matter what evidence you find, "best" is ambiguous enough that you can't even imagine the evidence that would settle the question for you. You could get every person on Earth to agree it's *Moonraker* and that still wouldn't necessarily settle the question.

On the other hand, "which era of Bond movies had the highest ticket sales?" can definitely be answered. You look at the ticket sales and see when they were highest. Evidence can tell you the answer to this question.

So we have a question that can be answered. But does it *improve our understanding of how the world works?* What this means is that the research question, once answered, should tell you about something broader than itself. It should inform *theory* in some way. Theory doesn't have to be something as important as the theory of gravity or the theory of evolution. It could even be something as generic as "bread costs more today than last year because bread prices have been generally increasing over time." Theory just means that there's a *why* or a *because* lurking around somewhere. Even hydrogen is a theory - it says that we see material like water behaving a certain way *because* there's a kind of atom that behaves in certain ways and has a certain structure.

Take germ theory, for example. Germ theory says that microorganisms like bacteria and viruses can cause disease. This explains *why* we have diseases, and also *why* disease can spread from one person to another. We don't call it "a theory" because we're uncertain about whether it's true.[2] We call it a theory because it tells us *why*.

A good research question *takes us from theory to hypothesis*, where a hypothesis is a specific statement about what we will observe in the world, like "people who wash their hands will get sick less often." That is, a research question should be something that, if you answer it, helps improve your *why* explanation. Great research questions often come from the theory themselves - the line of thinking being "if this is my explanation of how the world works, then what should I observe in the world? Do I observe it?"

This is easy to miss! Let's keep working with germ theory as an example. We might ponder germ theory for a while and think "Hey, I wonder how small the smallest microorganism is." That's a research question that we could answer with the right evidence, and it's *related* to germ theory, and would be kind of neat to know. However, learning the answer to this question wouldn't really tell us anything new about why we have diseases or why disease can spread from one person to another.[3] Maybe it helps us understand some *other* theory better. That would make our small-microorganism question a better research question for that other theory than for germ theory.[4]

So does asking "which era of Bond movies had the highest ticket sales?" improve our understanding of how the world works? Maybe, for the right theory. Maybe we have a theory that says that action movies were generally at their most popular in the 1980s. Asking about Bond sales over time might tell us a little more information about whether that theory is an accurate explanation of ticket sales.

LET'S WALK THROUGH AN EXAMPLE, starting with our theory. Let's say we have a theory that your curiosity as an adult is harmed by exposure to passive entertainment like TV and movies. Regardless of whether this is actually true or not, it still qualifies as a theory - it explains why we might see certain levels of curiosity in adults.

A natural research question here is "does watching a lot of TV as a child dull your curiosity as an adult?"

Let's check our two conditions for research questions. Could we answer this question? Yes! The data necessary to answer this question might be hard to come by, but we can at least conceive of it existing. If we randomized a bunch of kids to watch different amounts of TV, and then followed them to adulthood and measured their curiosity, that would be some pretty convincing evidence on our research question.[5]

Second, does this research question tell us about how the world works? Yes! If we answered this question, that would pretty clearly inform our theory. If we answered our research question with "no, watching a lot of TV as a child does not dull your curiosity as an adult," then it would be a pretty hard sell to explain adult curiosity by saying it's because of passive entertainment. The research question does help us figure out if the theory is any good.

A good test for whether a research question informs theory is to imagine that you find an unexpected result, and then wonder whether it would make you change your understanding of the world. Let's say that instead of answering the research question "does watching a lot of TV as a child dull your curiosity as an adult?" we use the research question "do kids who watch lots of Sesame Street tend to have lower levels of curiosity later?" We do our research and find that, actually, kids who watch Sesame Street have *higher* levels of curiosity! Uh oh. With this new information, do we have to change our theory? Well, we hem and we haw, and we think about how fond we were of that original theory. And we explain away the Sesame Street result by noting that Sesame Street might be different from most kinds of TV, and also that we just looked at which kids *did* watch Sesame Street, not whether Sesame Street is actually responsible for their curiosity - maybe kids who are more curious in the first place choose Sesame Street.

This ability to see a bad result and still hold on to the original theory tells us that the research question wasn't very good, at least not for this theory.[6] A really good research question, once answered, should be hard to explain away just because it's inconvenient.

So there we have it - "does watching a lot of TV as a child dull your curiosity as an adult?" is a good research question that could be answered with the right data, and would inform our understanding of the world if answered. Granted, the process of actually *answering* that question is another hurdle.[7] But at the very least we know that the question itself is good, even if the answer is elusive.

2.2 Why Start with a Question?

This sounds hard. Why bother? We have a bunch of data at our fingertips. In fact, we're awash in it. There's data everywhere. So why not skip the hard part of deriving a research question from a theory and instead just see what sorts of patterns are in the data?

Well, you could. In fact, a lot of people do. This is called "data mining," and there are people who do that very thing, and manage to do it quite well. They go to the data, look for patterns, and report back. You find a lot of this in the field of data science, but data mining can be done any time you have some data. Just look at the data, see what's in there, and work backwards.

So, sounds good, right? Well, data mining is well and good, but it turns out to be very good at some things, and very bad at others.

The kinds of things that data mining is good at are in *finding patterns* and in *making predictions under stability*.[8] The kinds of things that data mining is less good at are in *improving our understanding*, or in other words *helping improve theory.* It also has a tendency to find *false positives* if you aren't careful.

Finding patterns and making predictions are very valuable. And we probably do want to rely on some sort of data mining for these tasks. After all, there's no way we can really theorize about every possible pattern that could be in the data and think to check it. Doing something that just asks *what* we see rather than *why* is the right angle to take there. Plus, sometimes seeing patterns in data can give us ideas for research questions that we can examine further in other data sources.

It's also probably the best angle to take when we don't care about why! If I don't care why the stock market goes up or down, and I just want to predict if it will or not so I know whether to buy or sell, then data mining may well be the way to go.

But outside of those realms?

Why does data mining have difficulty helping theory? There are a few main reasons.

One of the reasons is that data mining, by definition, focuses on what's in the data, not *why* it's in the data. In other words, it's fantastic at revealing correlations - patterns in the data of how variables we've observed have varied together in the past - but the correlations it uncovers may have little to do with causality, or an understanding of *why* those variables move together.

To introduce an example that will pop up a few times in this book, someone using data mining to try to understand ice cream sales may well notice that the proportion of people who wear shorts is a fantastic predictor of ice cream sales. But shorts-wearing isn't *why* people buy ice cream. They buy ice cream and wear shorts because it's hot. But to a data miner, the shorts/ice cream connection is pretty compelling! After all, shorts can be a great way of *predicting* ice cream eating, even if there's no "why" there.

However, if what we're really interested in isn't predicting ice cream but *explaining why* people eat ice cream, it's pretty tempting at that point to try to invent a story to justify why shorts might actually be the reason people eat ice cream. In the case of ice cream and shorts we can tell that's ridiculous, but it's a lot harder when we don't actually know what's ridiculous and what's not ahead of time.

For example, we'd love to know what causes children to act aggressively. That sounds really important! A data mining exercise here might look through all of the things kids do or are exposed to, and check whether any of them are associated with higher levels of aggression. Maybe kids who play a lot of video games are more likely to be aggressive. So … are the video games responsible? Maybe, maybe not.[9] Data mining is well-equipped to find the relationship but poorly-equipped to tell us *why* that relationship is there. Hopefully someone doesn't get worried and ban all the video games before the researcher can carefully explain the distinction.

Another reason is that, because it's so focused on the data, data mining doesn't really deal in *abstraction*. For example, take a look at a chair. How do you know it's a chair? Well, it's probably got some legs, maybe a back, definitely a flat-ish seating area, and it's clearly designed for sitting in. This is our "chair theory" - we theorize that there are these objects called chairs that have certain chair-like properties, united in the ability to sit on them some distance off the ground. The chair you're looking at now is one example of chair theory.[10]

But what's actually *in the data?* There's no "chair" in the data. There's just a flat bit and some straight-up-and-down bits underneath the flat bit. Data mining would be great at noticing that it sees a lot of flat bits on top of straight-up-and-down bits, but it would not be good at developing "chair theory" for us because it would miss *why* we keep seeing that arrangement - because it allows us to sit on it. A data miner would never guess that the four-legged chair has anything to do with, say, a bean-bag chair, which has no straight-up-and-down bits at all.

False positives are another reason why data mining can be dangerous. Take the video games and aggression example. Okay, sure, maybe the video games aren't *why* the kids are aggressive, but we still found the relationship. Surely there's *something* there.

Well, maybe, and maybe not. Data mining means looking in the data to see what's there. And there's a lot of stuff to look at! If you check, say, a hundred variables and see if they're related to

aggression, *something* is going to pop up as looking related, just by random chance. That random relationship is unlikely to pop up again if you tried another sample. It's only in the sample you have by random chance, which is what makes it a "false positive."

That's one major danger in proceeding in your work without starting with a solid research question. Without a disciplined research question, there's no reason not to just check everything! *Something* is going to pop up as related by random chance if you check enough stuff. It takes a really well-behaved researcher to not pretend that's exactly what they'd been looking for from the start, and to fill in some reason why the 100th relationship they checked makes perfect sense and supports some theory.

There are ways of avoiding false positives while doing data mining - this is something they worry about a lot in data science and have a lot of tools for.[11] But if you're just sort of trawling through a data set to see what you see, you're likely to end up with a whole lot of false positives mixed in with the real positives. You'll have no way of telling one from another.

THAT SAID, DATA MINING ISN'T ALL BAD. There's no way we could possibly *think up* every interesting theory to test. Plenty of theories come from looking at the data in the first place, noticing a pattern, and wondering why the pattern is showing up, or whether the pattern is even real.

The drug Viagra, for example, was initially being tested as a blood pressure medication. The researchers testing it out to see if it worked to lower blood pressure happened to notice, uh, its other effects.

They've done data mining there - instead of coming to the data with a theory, they noticed an interesting pattern in the data.

Of course, the responsible thing to do at that point is to not just take the pattern as given. *That's* where the real problem of data mining is. Instead, they took the interesting pattern they noticed and looked to see if it held up in *other* data - if it replicated - before being sure that the pattern they noticed was real and explained how the drug worked.

Data mining isn't bad. It's just bad as a *final step* if you're trying to explain the world. It can still work as a source of ideas. And heck, maybe it can help you earn a hojillion dollars like Viagra did,

too.

2.3 Where Do Research Questions Come From?

Research questions can come from lots of places. Mostly, curiosity. We want to know how the world works, and that naturally leads to questions!

There are two steps in this process: thinking about theory, and coming up with a research question. Either one can come first.

Perhaps it begins with theory: “I think this is how the world works” or “I wonder if this is how the world works” - that’s your theory. This could be anything from “I think people make the decisions they do because they follow incentives” to “I think plants survive without eating because they collect energy from the sun” to “I think CD sales are down because people stream music now instead.”

With the theory in place, the process continues with our hypothesis: “if this is how the world works, what would I expect to see in the world?” Our above theories might lead to the research questions “will students work harder in school if you pay them for good grades?” or “will plants die if you store them in a dark room?” or “are CDs more popular now in areas with bad Internet connections?” These research questions tell us a *hypothesis to test* such that the result of that test *tells us something about the theory.*

The question might come first. “Will students work harder in school if you pay them for good grades?” we might ask. Then, we might wonder why we came up with such an idea in the first place. Probably because we think students respond to incentives. Or we might wonder, if we answered the question, what sense we might make of it, leading us back to our theory. If you can’t figure out why you would ask the question, it may not be a great research question. Or at least you’d have a hard time getting anyone to care about the answer once you had it.

Let’s be honest, sometimes research questions also come from *opportunity*.

Have a neat data set? Think about what data is available to you and whether any related research questions or theories come to mind.[12]

Or, perhaps you've learned about something unusual or interesting that has happened in the world. Maybe you've learned that a few school districts have decided to try paying their students for good grades. When you hear about something like this, you might ask "what research questions would this allow me to answer?" and from there you have a research question, and from there a theory!

2.4 How Do You Know if You've Got a Good One?

YOU'VE FOLLOWED THE PROCESS. You have a research question in mind. You know it can be answered with data, and you're pretty sure that if you get the answer to it, it will help you learn how the world works.

But is it really a good one? Just a few things to check before you get too far into the process:

- **Consider Potential Results.** A good way to double-check the relationship between your research question and your theory is to *consider the potential answers you might get.* Then, imagine what kind of sense you'd make of that result, or what conclusion you would draw. Let's say you find that students *do* tend to work harder in school when they're paid for good grades. What would this tell us about how students respond to incentives? Or let's say you find that students *don't* work harder when they're paid. What would *that* tell us about how

students respond to incentives? If you can't say something interesting about your potential results, that probably means your research question and your theory aren't as closely linked as you think! Let's say we *do* find that kids who happen to play video games are more aggressive. Can we take that result and claim that video games are a cause of aggression? Not really, for the reasons we've discussed previously. So maybe that research question really isn't linked to that theory very well.

- **Consider Feasibility.** A research question should be a question that can be answered using the right data, if the right data is available. But *is* the right data available? If answering your research question is *possible* but requires following millions of people repeatedly for decades, or trying to measure something that's really hard to measure accurately, like trying to get people to remember what they had for lunch three years ago, or getting access to the private finances of thousands of unwilling people, then that research question might not be feasible. While sometimes you can get around these problems with a clever design, you might want to

consider going back to the drawing board.

- **Consider Scale.** What kind of resources and time can you dedicate to answering the research question? Given a lifetime of effort and considerable resources, you might be able to tackle massive questions like "What causes some countries to become rich and others poor?" Given the confines of, say, a term paper, you could take some wild swings at that question, but you're likely to do a much more thorough job answering questions with a lot less complexity.
- **Consider Design.** A research question can be great on its own, but it can only be so interesting without an answer. So, an important part of evaluating whether you have a workable research question is figuring out if there's a reasonable research design you can use to answer it. Figuring out whether you do have a reasonable research design is the topic of the rest of this book.
- **Keep It Simple!** Answering any research question can be difficult. Don't make it even harder on yourself by biting off more than you can chew! A common mistake is to bundle a bunch of research

questions into one. “What are the determinants of social mobility?” I.e., how someone can move from one social class to another throughout their lifetime. There are *many* determinants of social mobility. You’re unlikely to answer that question well. Instead try “Is birth location a determinant of social mobility?” For another example, how about the question “How was the medium of painting affected by the Italian renaissance?” In a million ways! You’ll get lost and do a poor job on a bunch of minor pieces instead of getting at the whole. Instead maybe “What similar characteristics are there among the countries that adopted the use of perspective in painting most quickly?”

So, consider feasibility, scale, and design. Keep it simple, and think about whether the results you might likely see would tell you anything interesting about the world. After all, learning something interesting and new about the world is our goal!

Previous Next

Page built: 2022-12-09 using R version 4.2.2 (2022-10-31 ucrt)