Module 5 Part 2: Data Analytical Thinking – applying data analytical thinking

Welcome!

This is the continuation of data analytical thinking module. Last time we covered how data can be used to define public problems. Now we will apply a process for using this skill and identify the key risks when using data.

How do we, especially if we are not trained data scientists, think about using data to define the problem better?

How do we define our question? How do we identify what data is available to us? How do we draw inferences from that data that we can rely on?

The answer is, broadly speaking, a three step process for using data to define the problem better: 1) Define the hypothesis, 2) Identify and find the data to validate or disprove the hypothesis, 3) Choose a Method of Analysis.

As we learned in the module on problem definition, we can formulate a hypothesis based on our statement of the problem.

A hypothesis is merely a "proposition which can be put to test to determine its validity." It is a testable suggestion that something is caused by something else.

In this case we want to test that the problem is caused by the stated root causes outlined in our problem definition. We have previously talked to relevant audiences to ascertain if the hypothesis expressed in the problem statement is true, but we can also use data to prove or disprove our definition of the problem.

We can frame the problem definition as a hypothesis by writing down why we think the problem is happening. Our hypothesis may be that children are skipping school because their parents are taking them on off-season holidays to Disney World. Or maybe our hypothesis is that there are not enough cabs on the street when work lets out because there is a financial incentive for cabbies to change shifts at that time.

Perhaps we have a hypothesis that unemployment is increasing because robots are replacing workers. We might hypothesize that the rate of starting new businesses is decreasing because interest rates are increasing, and the cost of capital is going up.

Often the hypothesis includes a theory of change about the best way to solve the problem. Perhaps our hypothesis is that we are going to improve healthcare by spending more on preventive medicine, such as annual checkups, than on after-the-fact procedures because the root cause of poor health is a lack of attention to diet.

This problem definition defines the problem in a way that gets at the underlying behavioral hypothesis about which behaviors lead to better health.

In Chicago, for example, when they developed the city's data-driven project to solve the problem of food borne illness, they defined the goal of the project as changing how they inspected restaurants to increase the speed of finding critical violations in an effort to prevent them from recurring.

Regardless of whether we include the theory of change or not, ensuring that we have defined a specific and actionable problem is essential to take the next step, namely identifying what data to use to test the hypothesis.

It can, of course, be challenging to define a hypothesis without initial background research. When one is not aware of or knowledgeable about the relevant data, it can be difficult to even formulate the question.

For example, coming up with a hypothesis about preventive medicine depends, first, on knowing the statistics about preventable diseases and the relationship between diet and morbidity. The process of determining the hypothesis is iterative and often needs to be revisited after data has been gathered and analyzed.

The next step is to identify the data needed to answer the question and validate the hypothesis. There are mountains of data waiting to be discovered and used for social good.

Generally speaking, there are some sources of reliably available data in the United States. These include spending data at the local, state and federal level. Federal grant and contract data along with census data. Crime, housing, and utility data are prevalent.

You may need to take advantage of the Freedom of Information Act to demand data that should be open and is not.

For example, in 2013, transparency activist Carl Malamud began coordinating an effort to use FOIA to force the IRS to publish nonprofit tax returns.

Malamud used FOIA to request nine nonprofit tax returns from the IRS because the agency would not make the returns available in digital form. Although disclosure of nonprofit returns is required by law and the filers submitted those returns electronically, the IRS wanted to send Malamud image files of the returns. The IRS typically took electronically-filed returns, printed them out, scanned them back in, and sold DVDs with the image files. But because of his successful suit and campaign, the IRS not only turned over Malamud's nine requested returns in a digitally readable format but soon after began to make all electronically filed nonprofit tax returns, which represents about 60% of those filed since 2011, digitally downloadable as open data.

Alternatively, where the data is not available, you may need to gather it through your own survey, a crowdsourcing or a citizen science exercise.
One such example came during the 2011 Fukushima Daiichi Nuclear Disaster in Japan. Distrustful of government-published information, citizens began collecting data of their own using handheld Geiger counters, which was compiled, monitored, and openly shared through a project known as Safecast.

Some data, especially administrative data, may not be publicly available but may still be accessible to a policymaker or accredited researcher through a data lab.

Administrative data is that personally-identifiable information that government collects about us in the course of administering services to us, such as distributing an unemployment or disability benefit, handing out food subsidies, giving us a driver’s license, booking us for a criminal act or releasing us from prison. Third parties such as hospitals and schools also collect data about us that gets reported back to the government, thereby providing an additional source of such information.

To make private data usable while protecting privacy, several governments have turned to the creation of so-called “data labs.”

“Data Labs” or “Policy Labs” are institutions with small groups of data analysts working inside or in tandem with government agencies to make administrative data more usable for evaluation and research. And while organizations vary widely in their implementation, they have all developed models to tap into the skills of highly talented data analysts and to access valuable government datasets responsibly.

For almost 35 years now, Professor Fred Wulczyn and his team at the University of Chicago have worked with states to help them build what he calls “research-valuable data” from the administrative records they maintain for other purposes.

The cornerstone of the Data Center's offerings is the Foster Care Data Archive and associated web tool that supports access to the data needed to generate the evidence needed to support strong foster care programs.

By harmonizing the data across jurisdictions (states and counties), the Archive makes comparative, between- and within-state research possible. The harmonization strategies resolve the challenge of mixing data collected from different state agencies operating under different policy guidelines into a coherent, integrated framework.

There are three key things to check off the list when validating a data set.

First, make sure the data comes from the most authoritative source. If you need census data, get it directly from the Census Bureau. If you need student test scores go directly to the Department of Education or national testing bodies.

Second, take a look at the data and make sure it passes the basic sniff test. Does this data make sense based on what I know about the problem?

Third, try and triangulate the data to verify its accuracy. Triangulation means using more than one method to collect data on the same topic. This is a way of assuring the validity of research through the use of a variety of methods to collect data on the same topic, which involves different types of samples as well as methods of data collection. Thus, try to find another source for the same information and compare.

To answer your research question and investigate your problem definition, you may need to combine data from a number of sources.

For example, in Chicago to develop the city's improved way of inspecting restaurants, city officials self-evidently started with an analysis of the City's historical data on food inspections to predict which establishments were more likely to re-offend.

However, they ultimately looked at a whole host of factors, including three-day average high temperature, nearby garbage and sanitation complaints, nearby burglaries, whether establishments have a tobacco or alcohol license, length of time since last inspection, length of time establishment has been operating and who the inspector was.

Going outside your organizational silo and talking with people across different programs and departments to help identify and find data is key to understanding what data is relevant to the problem and where it can be found.

Once the problem is defined and the data identified and found, then comes the next step of deciding what kind of analysis to do.

One of the most straightforward things to be done with data is simply counting. In the age of big data, we may be counting more things and doing so faster with the aid of a computer but, in the end, a great deal of research involves nothing more than tallying to gain insight.

Princeton Sociologist, Matt Salganik, points to another piece of research involving New York City taxi data.

A 2014 study by his Princeton economics colleague Henry Faber used the taxi data to answer a fundamental question, namely whether, on days on which they could earn more, taxi drivers would drive more consistent with what one would assume from neoclassical economics. Alternatively, would the data reveal, consistent with the assumptions of behavioral economics, that drivers would simply seek to earn a certain amount and, beyond that amount, cease to drive.

In fact, drivers did choose to drive more. Although the conclusion is important, Faber did little more than simply add up taxi driver earnings.

Natural experiments are another source of useful insight that rely on observation without the need to design an experiment or build an algorithm. In a natural experiment, we look for an event that is naturally occurring but that points up societal differences from which meaning can be gleaned.

For example, in 2001, Norwegian tax records became easily accessible online and everyone’s income became transparent, making it possible to draw comparisons between income groups.
UCLA economist Ricardo Perez-Truglia used the income data along with survey data from 1985–2013 to test whether transparency, which allowed people to see a wealth gap that had previously been hidden, impacted people’s happiness and satisfaction. He found that transparency made the rich more satisfied and the poor less satisfied with their lot in life, with implications for the policy debate on tax transparency.

Natural experiments can be very helpful for policy makers as they generally rely on already gathered data and limit the need for original research or constructing designed experiments.

While not always feasible, the gold standard in terms of analysis is the randomized controlled trial (RCT).

An RCT involves taking a population, such as a group of schools or hospitals, and dividing the group into two parts with one half receiving an intervention, program or treatment and the other not receiving it. As with a natural experiment, the researcher studies the resulting differences.

However, the intentional sorting of the population into two or more groups distinguishes the RCT. This is what we commonly do in medical science when we give half the patients the drug and the other half the placebo.

If we observe a statistically significant difference, while holding other factors constant, than we can attribute success to the social program. Randomized controlled trials provide a scientific approach to assessing the efficacy of social programs and taking the ideological guesswork out of the process.

As we discussed in our first module, machine learning is the science of teaching computers to learn.

When used for data analysis, machine learning offers a powerful means for both spotting problems and solving them.

For example, the Rockefeller Foundation has partnered with the Alliance for a Green Revolution in Africa (AGRA) and Atlas AI to fund a predictive analytics project to anticipate damage to with the ultimate goal of improving food security across sub-Saharan Africa. Using satellite imagery, Atlas AI has trained a machine learning algorithm to study crop growth in relation to changing weather, diseases, and pests. By predicting perturbations in agricultural production, it can help government and philanthropy both to anticipate and prevent losses and to target interventions.

At every stage of the data lifecycle from data collection to data processing to data analysis to data use, there are risks. We'll cover some of the key ones here and talk in our next module about how to mitigate them.

Real world problems are often complex and multi-dimensional. Over-reliance on any single method of analysis can be fraught with danger. Using different methods of analysis and multiple data sets related to the problem can help to mitigate this risk.

As we stated at the outset, too, relying only on data analysis of one kind without talking to and learning from humans gives us a one-sided view of the problem. We must do both!

Often the data we're using is incomplete.

If I am trying to measure popular sentiment on an issue using Twitter, I am only measuring the sentiment of those people who use Twitter. The elderly, the poor, the homeless and others who are not big users of social media may go undercounted.

Also, we are often failing to collect the right data. U.S. federal crime data is a good example of how certain data is collected while other relevant data is ignored.

Although the FBI collects, publishes, and makes available for downloading the Uniform Crime Reports estimated monthly aggregates of instances of eight major crimes (murder, rape, assault, robbery, arson, burglary, larceny-theft and motor vehicle theft), we have no similar data store for white collar crime.

The existence or non-existence of data ends up changing our policy priorities. Thus, we must pay careful attention to what's missing.

The problem extends beyond the public sector too - Gartner estimates that 25% of Fortune 1000 companies have information that is inaccurate, incomplete or duplicated.

Data conveys a sense of impartiality and infallibility. Machine learning algorithms, for example, are sometimes introduced to reduce human forms of bias in decision-making. But there is no such thing as unbiased data or unbiased machine learning, making impartial decisions. A machine learning algorithm learns from the historical data that it has been trained on and thus, biased inputs will lead to biased outputs.

As an example, relates Gideon Mann, Chief Data Scientist of Bloomberg LLP, if some of your population is not represented in your training data, the sample the algorithm that you are going to come out with is not going to perform well or be accurate on that part of your training data sample."

It is far too easy to make mistakes when doing data analysis. We make mistakes when looking at data by drawing conclusions that are not supported by the facts at hand.

P-hacking or data dredging is the problem of inferring statistically significant findings in data when none exists. Ideally, one defines the research question and the hypothesis prior to analyzing the data to prove or disprove the claim. If the first hypothesis does not bear out, however, we might try to analyze the same data differently, looking for a new hypothesis. While we are not scientists looking for the perfect experiment, we have to take care not to keep digging and poking simply to find something -- anything -- that causes the data to fit the problem. Establishing data responsibility principles, processes and tools to ensure that data is shared, analyzed and used responsibly and ethically helps to ensure that insights can be gained without harming individuals or groups.

Ultimately, the greatest data risk is the failure to use data to solve public problems in the first place.

Data is playing an increasingly important role in solving big public problems, primarily by allowing citizens and policymakers access to new forms of data-driven assessment of the problems at hand.

It also enables data-driven engagement, producing more targeted interventions and enhanced collaboration. However, IBM estimates that worldwide 80% of the data we collect goes unused. In Europe that number is 85%.

Beth Blauer tells a story of her time leading the StateStat performance management team for Maryland that starkly illustrates the dangers of failing to use data, whether for legal, cultural or technical reasons.

“I had a meeting with our juvenile justice agency and we also had our social services agency and our public safety agency and we were talking about our most violent and dangerous offenders in our state and I asked a very innocent question about how often were our agencies taking our foster care locations--our registered statewide foster care locations--and overlaying it with where our most violent and dangerous offenders lived and matching those addresses. And what I thought was a very innocent question turned into be a very serious problem because the answer was that information is not shared because by law we are prohibited to exchange that information interagency. And my head nearly exploded. I couldn't imagine a scenario where we weren't thinking about this on a regular basis as we're placing children into our own care.”

Combining data with human-centred and participatory approaches is a powerful way to solve public problems. Next, we look more closely at one of the most important sources of data - open data.