RESEARCH · PHILOSOPHY AND PARADIGMS

Theory-Driven versus Data-Driven Approaches to Research and Knowledge Creation

How theory-driven and data-driven research differ, what each does well and badly, why claims that big data makes theory obsolete have not held up, and how the two approaches combine in good practice.

ModQu research article · 1,593 words · 18 references (APA 7th edition)

Research can start from an idea or from the data. In **theory-driven** research, an existing explanation tells the researcher what to look for: hypotheses are derived from theory, data are collected to test them, and the results confirm, qualify or refute the theory. In **data-driven** research, the researcher starts with a body of data and searches it for patterns, letting the data suggest what is worth explaining. The rise of very large datasets and machine learning has sharpened the contrast between the two, with some commentators predicting the end of theory altogether. This article explains the two approaches, the arguments made for each, the lessons of recent experience, and how they can be combined.

Theory-driven research

The theory-driven approach is often described as top-down or deductive. Popper (1959) gave it its classic justification: observation is always selective and guided by expectations, so the best a researcher can do is make those expectations explicit as hypotheses and test them as severely as possible. A theory tells researchers which variables matter, how to measure them and what pattern to expect, and it allows findings to accumulate, because each study tests and refines a shared explanation.

The strengths of the approach are focus and interpretability. A theory-driven study measures what it needs, can be designed to rule out alternative explanations, and produces results whose meaning is clear: the hypothesis was supported or it was not. Its weakness is that it can only find what it is looking for. If the theory leaves out an important factor, a theory-driven study is unlikely to discover it.

Data-driven research

The data-driven approach is bottom-up or inductive. It has a long history. Tukey (1977) championed exploratory data analysis, arguing that researchers should look at data carefully and flexibly before imposing models on it, because the most valuable findings are often unexpected. Glaser and Strauss (1967) made a parallel argument in qualitative research, proposing that theory should be grounded in data rather than imposed from existing frameworks.

Large digital datasets have given the approach new force. Hey et al. (2009) described data-intensive science as a "fourth paradigm" of discovery, alongside experiment, theory and simulation. Leonelli (2016) showed how biology has become increasingly organised around large shared databases, in which data are collected and curated for many future uses rather than to test a particular hypothesis.

The end of theory?

In a widely discussed magazine article, Anderson (2008) argued that with enough data, correlation is sufficient: researchers no longer need models or hypotheses, because patterns in massive datasets can be used directly to predict and act. The article captured a real shift in how some organisations use data, but its central claim has not fared well.

A prominent test case was Google Flu Trends. Ginsberg et al. (2009) reported that search query data could estimate influenza activity in the United States more quickly than official surveillance. The system selected search terms purely by how well they correlated with past flu data, with no theory of why people search as they do. In later seasons, it substantially overestimated flu prevalence. Lazer et al. (2014) attributed the failure partly to "big data hubris", the assumption that large data can substitute for traditional data collection and analysis, and partly to changes in the search engine itself that altered the relationship between searches and illness. Without a theory of how the data were generated, the system could not anticipate when that relationship would break.

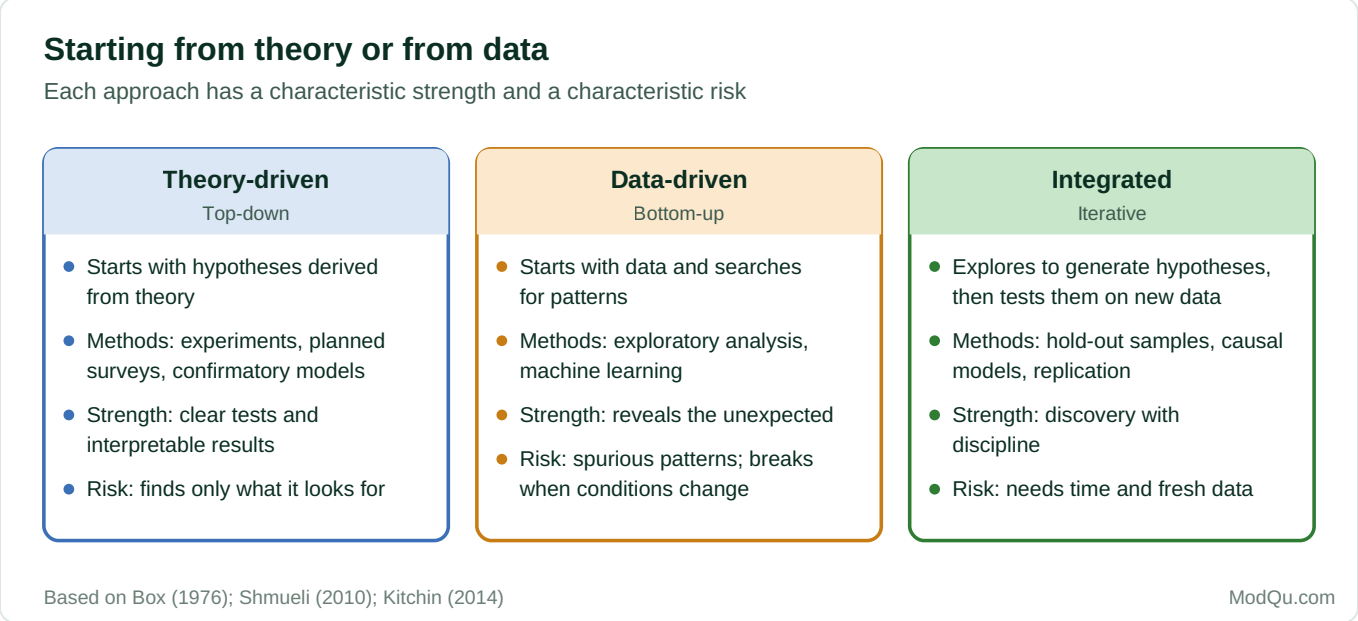


Figure 1. Theory-driven, data-driven and integrated approaches compared. Based on Box (1976), Shmueli (2010) and Kitchin (2014); diagram by ModQu.

Calude and Longo (2017) added a mathematical argument: in sufficiently large datasets, many correlations will arise simply because of the size of the data, regardless of any real relationship. The more variables are searched, the more spurious patterns will be found. Kitchin (2014) and Mazzocchi (2015) argued that data are never theory-free in any case: decisions about what to record, how to categorise it and which patterns to treat as meaningful all reflect assumptions, whether stated or not.

Prediction and explanation

Part of the debate turns on what research is for. Breiman (2001) described two cultures in statistics. One, dominant in academic statistics, assumes a model of how the data were generated and estimates its parameters. The other, growing in computer science, treats the data-generating process as unknown and judges models solely by how well they predict new data. Breiman argued that the second culture had been unjustly neglected and often produced more accurate predictions.

Shmueli (2010) clarified that explanation and prediction are different goals that call for different methods. An explanatory model aims to test causal hypotheses and may sacrifice some predictive accuracy for interpretability. A predictive model aims to forecast new cases as accurately as possible and may include variables with no causal role. Problems arise when one is mistaken for the other, for example when a variable that predicts well is assumed to be a cause that can be manipulated.

Pearl and Mackenzie (2018) pressed this point further, arguing that data alone, however large, cannot answer causal questions such as what would happen if a policy were changed. Answering those questions requires a model of how variables influence one another, which is a form of theory.

Hofman et al. (2017) and Yarkoni and Westfall (2017) proposed a middle position for the social sciences. They argued that explanatory research would benefit from adopting the predictive discipline of machine learning, particularly testing models on data not used to build them, because many explanatory models that fit their own data well predict new data poorly. This is a call for theory-driven research to be judged more severely, not for theory to be abandoned.

Risks on both sides

Each approach has characteristic failure modes.

Theory-driven research can become confirmation-seeking: researchers find what they expect because their designs cannot reveal anything else, or because flexible analysis allows them to make the data fit. Simmons et al. (2011) showed how such flexibility can produce false positive results even when each individual choice seems reasonable.

Data-driven research can mistake noise for signal. Searching many variables for patterns guarantees that some will appear by chance, and patterns that reflect how the data were collected can be mistaken for facts about the world. Without a theory, it is also hard to know whether a pattern will hold in new circumstances.

Combining the approaches

In practice, good research moves between the two. Box (1976) described science as an iterative process in which theory suggests what data to collect and data suggest how the theory should change. Data-driven exploration is valuable for generating hypotheses and spotting anomalies; theory-driven testing is needed to check whether those hypotheses hold up.

Several practices make the combination work.

- **Separate exploration from confirmation.** Patterns found by searching one dataset should be tested on a separate dataset before being treated as findings. Holding back a test sample is standard in machine learning and should be standard in all research.
- **Use theory to judge plausibility.** A pattern with no plausible mechanism deserves more scepticism than one that fits existing understanding, though it should not be dismissed if it replicates.
- **Model how the data were generated.** Knowing who is in a dataset, how they got there and how measurements were made protects against mistaking artefacts for findings, as the Google Flu Trends case showed.
- **Match the method to the goal.** If the aim is prediction, judge models by out-of-sample accuracy. If the aim is to understand what would happen under an intervention, a causal model is needed.

An example: explaining customer churn

Consider a telecommunications company trying to reduce the number of customers who cancel. A data-driven team might train a model on hundreds of variables, from call patterns to billing history, and find that it predicts cancellations accurately. The model might give heavy weight to the number of calls a customer makes to the help line. That is useful for prediction: customers who call often can be offered retention deals. But the model does not say why they call. If calls reflect unresolved technical faults, fixing the faults would reduce cancellations; if they reflect confusion over billing, clearer bills would. A retention offer treats the symptom without addressing either cause.

A theory-driven team might start from established ideas about customer satisfaction and switching costs, measure the relevant constructs in a survey and test which ones explain intentions to leave. Its model would be easier to interpret, but it might miss a factor the theory did not anticipate, such as a competitor's aggressive campaign in particular postcodes that the transaction data would have revealed.

Combining the two is more powerful than either. The data-driven model identifies the help-line pattern; the theory-driven work investigates what lies behind it; a targeted test, such as fixing a known fault for one group of customers and not another, establishes whether the proposed cause is real.

Implications for commercial research

Organisations increasingly hold large amounts of customer and transaction data and are tempted to let it speak for itself. Data-driven analysis can reveal unexpected segments, behaviours and risks. But acting on a pattern usually means intervening, such as changing a price, a message or a product, and the effect of an intervention depends on why the

pattern exists. A business that understands the mechanism behind a pattern can predict when it will hold and when it will not. One that does not is relying on the assumption that tomorrow will resemble yesterday, which is precisely when data-driven predictions tend to fail.

References

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science, 22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1). https://doi.org/10.1177/2053951714528481

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? *EMBO Reports, 16*(10), 1250–1255. https://doi.org/10.15252/embr.201541001

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393