March **2016**

IPSOS VIEWS

Big Data: A Guided Tour

By **Rich Timpone, Ph.D.**
Senior Vice President, Ipsos Science Centre

FOREWORD

Look at its power. Imagine the potential. It often feels like Big Data is omnipresent. Not least in the workplace, where it has become a feature of so many current discussions about strategies and business plans. As the US President's office points out, it can be a force for good: saving lives and making the economy work better. But Big Data also brings with it risks and responsibilities – as witnessed by growing consumer concerns about data protection and information security.

Big Data may now be famous. But it can be hard to pin down – a little mysterious, even. In this Ipsos Views paper, Rich Timpone sets us off on a guided tour of the subject matter – taking us from the all-important definitions through to questions around how it should (and should not) be used.

For those of us involved in the research industry, there are some clear rules we need to follow if we are to make sense of it all. And that's before the role of "traditional" market research techniques come in.

WHAT IS BIG DATA?

"Big Data" is a very broad term, used often - and for a variety of different purposes. But it can be simply defined as **any collection of data or datasets so complex or large that traditional data management approaches become unsuitable** (Cielen and Meysman 2016).

This definition provides a framework for categorising:

- **The types of data that exist**
- **The business problems the data can be applied to**
- **The methods used for capturing data and extracting insight from it**

And, in turn, this framework helps us to understand:

- **Opportunities for leveraging new sources of data**
- **The risks that co-exist with the opportunities**
- **Where there are connections to more traditional market research approaches**

Passive Data

MONITOR

Directly observe and track consumer behaviour, what they say, where they go, what they watch/listen to, and ultimately what they buy. Leverage user generated and other data sources directly.

Active Data

ASK

Measuring what matters, when it matters. Traditional qual and quant data that directly engages individuals to provide their attitudes and behaviours. Includes tracking and deep dive studies.

Interactive Data

INTERACT

Highly involving interactive activities which leverage social sharing, dynamic experience and gamification mechanics to engage consumers in ideation, co-creation and more.

THE CHARACTERISTICS OF BIG DATA

The term "Big Data" is often applied to datasets simply because of their size and/or their complexity. Some definitions are so broad that they lack clarity - for example, in 2014, the Executive Office of the President of the United States issued a report on Big Data that defined it as: "Large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, internet transactions, email, video, click streams, and/or all other digital sources available today and in the future".

The typical industry description of Big Data rests on the "Three Vs", which has its origins in META Group (now Gartner). The three Vs are the elements that fundamentally define Big Data:

- **Volume**
- **Variety**
- **Velocity**

While some commentators say all three elements should be present before data can be considered to be "Big Data", other researchers and clients believe that only one of these elements may be enough.

> *"Some definitions of Big Data are so broad that they lack clarity."*

Volume

As one might imagine from the use of the word "Big", the size of the data source is a core characteristic, though there are few numerical measures of how large a dataset must be before it can be considered "Big". What matters is that the data is so large that traditional processing tools and techniques become ineffective.

Variety

By some estimates, more than 2.5 exabytes (2,500,000,000,000,000,000 bytes) of data are being produced each day. This creates challenges beyond scale and this data comes in many varied forms, including:

- **Simple datasets**
- **Natural text**
- **Online communication (discussions and blogs in social listening as well as web page content itself)**
- **Pictures**
- **Video**
- **Sensor data (from road cameras, satellites and other recording devices)**
- **Digital exhaust (the information leftover from transactions and other activities online)**

Not only are these things complex on their own, the challenges grow further as we combine them. The different forms of data have different units of analysis (such as individual, household, neighborhood, etc.), and traditional data processing techniques often cannot help us to gain insight from them.

Velocity

Modern data is fast, not just because it is being produced quickly, but because it moves quickly and is consumed quickly too.

There are more than 500 million tweets every day on Twitter, and Walmart handles over a million customer transactions every hour. So, there is an opportunity to think about delivering insight in near real-time.

While this speed creates new challenges and opportunities, it also gives rise to questions around the duration of measurement for specific examinations – such as how long is long enough in order to discover what we want to know?

> "*There are more than 500 million tweets every day on Twitter, and Walmart handles over a million customer transactions every hour. So, there is an opportunity to think about delivering insight in near real-time.*"

TYPES OF BIG DATA

Ipsos generally classifies data as:

- **ACTIVE - where respondents are explicitly asked their views in surveys, focus groups etc**
- **PASSIVE - any observational or extracted information where the individual not queried directly**
- **INTERACTIVE - such as from our social communities where dialogue is the source of the insights**

> *"Typically, most Big Data has its origins in the passive domain."*

Common sources of data are:

Behavioural Data

Many companies and clients collect data that track the actual behaviors of customers and individuals from online transactions and company activities, web patterns, mobile activity, software use, to the Internet of Things (IoT) with the modern capturing of sensors from TVs, refrigerators, cars etc.

Purchase Data

Point of sales transactions, as well as credit card and online purchases are being captured and mined for broad insight.

Social Listening

Digital discourse, including tweets, Facebook updates, blogs, vlogs, and online discussions, allow unprecedented ability to 'listen in' on actual exchanges.

Geo-location

A special form of behavioral data - the capturing of GPS data (often as part of the 'digital exhaust' of mobile and other services) allows for geographic analysis and movement patterns.

Unstructured Data

While text data, such as the social listening above, falls in this category, it also includes sources as diverse as photos and videos that are created by individuals. Note that the term unstructured is not meant to be pejorative but emphasises the fact that the sources are not captured in standard data layouts that allow straightforward summary and analysis.

Logistical

The bulk of the data we have been discussing focuses on individuals. Aggregate and logistical data may be analysed on its own or as context for individual behavioural understanding. In this way, we are often examining questions at different levels of analysis or integrating multiple levels. Machine generated as well as explicitly captured sources such as power consumption in national grids or air traffic in metropolitan areas may be valuable to analyse.

THE RISKS OF BIG DATA

The same concerns of traditional research also apply to the study of Big Data. Some of these are highlighted below, not because they are unique to Big Data, but because they are often overlooked due to assumptions about the data or in the excitement that surrounds this new and developing field.

Bad Data

At Ipsos, the quality of the data we use is fundamental to all of our projects, and we do not take it for granted when analysing Big Data. What we collect and how it is classified is crucial – and care must be taken when integrating and fusing data, as this can introduce additional measurement and prediction errors which may result in noise or bias.

We understand that data quality is not simply a question of good or bad but a range from one to the other, and understanding biases and the models to deal with them is a critical element that takes us beyond the simple collection and analysis of data.

> “*What we collect and how it is classified is crucial - and care must be taken when integrating and fusing data.*”

False Positives

Nassim Taleb (Beware the Big Errors of Big Data) argued that the number of spurious correlations grows disproportionately to the number of variables in the data (over 100,000 from analysing a few thousand variables). Big Data studies aren't big just in terms of the number of observations, but also in the number of variables.

Although there are statistical adjustments that can be made to lessen the risk (e.g. the Bonferroni correction, holdout validation, and cross validation), we must take care to identify relationships that are meaningful both statistically and substantively. Just as with traditional research, we must balance the approaches to deal with the increased risk of false positives with approaches to guard against creating false negatives.

Generalisability

Issues around the representativeness of Big Data are sometimes acknowledged, but the full implications are often underplayed. For instance, while it is known that the content of Twitter is unrepresentative, these and other online discourses are often treated as generalisable. While the data they produce is very valuable, we know that those passionate enough to discuss toilet paper or nappies online do not represent a random cross-section of potential purchasers.

This extends to situations where companies look at all the current users of their products and services. We are mindful that this information overlooks previous users and potential purchasers but not current subscribers. This is where blending more traditional active data sources with Big (and other passive) Data is helpful.

In the historical parallel, the Literary Digest poll for the 1936 US Presidential election had an unprecedented sample of 2 million respondents. In spite of this, their prediction was abysmally wrong due to the lack of generalisability of the sample. This lesson that larger does not always mean more representative extends to Big Data today.

Non-representative data is common in market research, so understanding and modelling the biases is crucial if we are to make generalisations. We must be careful not to mistake size for representativeness.

Stability over Time

With so many variables, the possibility for shifts over time should be considered. For example - the Google Flu Trends project initially found that looking at people's search engine patterns could reveal where outbreaks were happening much more quickly than traditional approaches. But when the exercise was repeated, the model led to dramatic over-predictions, because people's behaviour had changed in the meantime.

Underlying Causes

Much of the early value of Big Data is in uncovering surprising relations, like the sale of Strawberry Pop-Tart sales at Wal-Mart in advance of hurricanes. Deciding what to do with these insights can still be difficult and understanding what is behind the patterns is usually important to make the most of the opportunity they present.

Privacy

Even blinded Big Data can be used to identify individuals, as seen in the competition Netflix ran to create an improved recommendation engine. So it is important to apply the same standard of care here as we do when protecting the rights of research participants in 'active data' studies.

TWO MORE Vs

We think two qualities of Big Data are important, its veracity and its value. Unlike the three original Vs (volume, variety, and velocity), they do not describe what Big Data is, but they are important to consider when moving a discussion about Big Data from theory to actionable insight.

Veracity

The accuracy of Big Data is important, and this is true whether examining a single data source or integrating or fusing different sources. Traditional research issues of bias and quality are just as important when studying Big Data.

Value

Any study that we do focuses on providing meaningful and useful insights. So we apply all of our knowledge to avoiding the pitfalls and risks listed earlier.

Some of the promises made about Big Data have been dramatic and ambitious, raising huge expectations. In turn this has created something of a backlash and some disillusionment.

With quotes such as the US Executive Office of the President stating "Big Data is saving lives… making the economy work better… making government work better and saving taxpayer dollars", it is not surprising that many are wondering why they aren't benefiting from the remarkable opportunities.

Our view remains that the value of Big Data is very real, and that there are significant insights to be gained from examining it, many of which are unique.

But the following points are not always clear:

- **What data exists**
- **What questions can be answered with it**
- **What its limitations are**
- **How one can access and mine it**

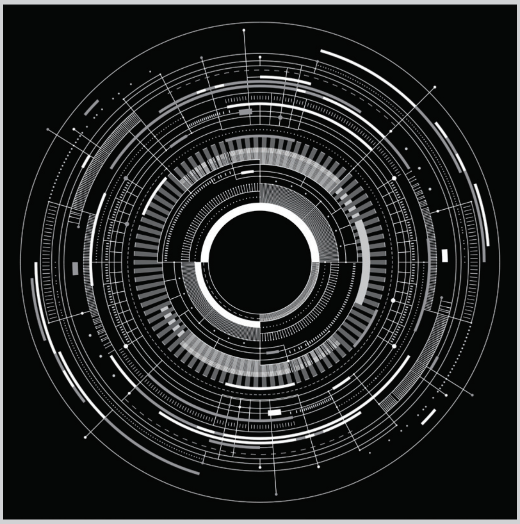
So, while there are technical challenges, we believe the main ones are actually more about research in a more general sense. Big Data is not a black box beyond the understanding of ordinary people or decision makers. The value from Big Data only comes when expectations are realistic, and proper objectives are set.

KNOWING WHAT VERSUS KNOWING WHY

A significant area of debate is whether the discovery of correlations within Big Data is where most future value will be, and that past attempts to understand causal relationships in the data will become much less important. Put more simply, "knowing what, not why, is good enough" (Mayer-Schönberger & Cukier, Big Data).
Of course, there is a spectrum of views about this. At one extreme is the view that "Petabytes allow us to say: Correlation is enough" (Anderson, The End of Theory), but a more moderate position is provided by Mayer-Schönberger & Cukier, namely that "Causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning".

Sometimes, it's about knowing 'what' people do or say alone is enough for action. The classic case of Wal-Mart examining its own sales data in advance of hurricanes and finding spikes in sales of Strawberry Pop-Tarts allowed it to further increase sales. But even this analysis was built by starting with a business question.

Often companies need to understand the 'why' to really understand what they should do in response to what is discovered in the data. Even the Wal-Mart example above would benefit by understanding why, so that it could find out whether other products could be promoted to tap into the same underlying impulses.



BIG DATA NEEDS BIG THEORY

Our view is that there are three components needed to get the maximum value from Big Data:

- **Business Questions and Theory**
- **Statistical Considerations**
- **Appropriate Analytic Techniques**

Having only two of the three creates risks:

- **LACKING THEORY may lead to the 'so what' findings that have driven some companies to question the value of Big Data**
- **LACKING STATISTICAL CONSIDERATIONS may lead to invalid findings and inferences**
- **LACKING ADVANCED ANALYTICS may lead to missing the patterns and insights themselves**

IPSOS VIEWS

With all three components present, we can understand the data, use increasingly sophisticated tools to identify patterns, and determine what is meaningful. Without all three components, results may lack value or, worse, be misleading.

We agree that there is some merit in just knowing 'what' (not 'why'), but believe that even the basic decision about which data source(s) to examine and how to capture the data requires a focus on a business question.

Ipsos has been growing its expertise and capability across every aspect of Big Data – from capture, through analysis, to insight. Our approach, earlier set out in an Ipsos paper from 2014, begins with understanding the underlying business purpose.

We are committed to innovation and staying at the forefront of trends in research, including Big Data and Big Data Analytics.

The tools of Data Science are often discussed interchangeably with Big Data Analytics (Cielen and Meysman 2016) and Ipsos has been growing its capability by linking computational modelling with statistical analysis in traditional research domains as well as Big Data.

In addition to developing new tools and techniques, the Ipsos Science Centre has used tools to help with massive parallel processing of data sets in the hundreds of gigabytes and extending into terabytes.

Although this is changing the balance of research that we do, we will continue to leverage our expertise across all types of data, including traditional forms, and make sure we use the most appropriate tools and resources to best position our clients for success.

So, we do not see Big Data as a simple alternative to traditional methods like surveys. Our view is that the different types of data possess distinct merits for strategic and tactical questions. Depending on the question, these can be employed separately or in coordination.

As discussed earlier, Big Data includes mostly 'passive' data where individuals are not explicitly engaged to answer questions or otherwise interact with researchers.

While this does provide new opportunities to examine what individuals are doing and saying; often pairing this with more traditional active and interactive sources provides the deepest and most actionable insights.

"*We do not see Big Data as a simple alternative to traditional methods like surveys. Our view is that the different types of data possess distinct merits for strategic and tactical questions. Depending on the question, these can be employed separately or in coordination.*"

IPSOS VIEWS

Rich Timpone leads the business team within the Ipsos Science Centre and is a research professional with 20+ years of experience and leadership. His role bridges the R&D efforts of the Ipsos Science Centre, supporting analytic innovation work in areas including Data Fusion, Customer Retention modelling and Bayesian Network analysis, with the delivery of advanced analytic solutions to meet client business needs.

Rich's prior role at Ipsos was leading the Marketing Science team dedicated to P&G which included partnering in the development of their current global brand equity programme.

Rich came to Ipsos with 8 years of client experience at financial service providers VPI and Nationwide Insurance; and prior to moving to the private sector served on the faculties of The Ohio State University and SUNY at Stony Brook for an additional 8 years.

Rich has conducted research, taught, and published extensively in the social sciences and on research methods including advanced statistical analysis and computational modelling. He received his M.A. and Ph.D. in Political Science from SUNY at Stony Brook, and his B.A. from Washington University in St. Louis.

The *Ipsos Views* white papers are produced by the **Ipsos Knowledge Centre.**

www.ipsos.com
@_Ipsos