

The Big Opportunity: Audience Research Meets Big Data

A report for the IPA by Richard Marks,
Research The Media Ltd.



Table of contents

About the Author: Richard Marks	3
Foreword by Lynne Robinson	4
The Big Opportunity	5
The Big Picture – A Summary	6
What do we mean by Big Data?	10
The Big Data Revolution	12
What are the strengths of Big Data?	15
What are the strengths of the Industry Currencies?.....	16
Maximising the potential of Big Data	18
1. The Bigger Picture: Context	19
2. What does Big Data represent?.....	20
3. Safety in (big) numbers?	21
4. Opt-in and convenience samples	22
5. Demographics	22
6. The Single Source Opportunity	24
7. Consumer-centric vs. Media-centric Insights	26
8. Verification and Accuracy	26
9. Skillsets	28
10. Privacy	30
11. Big Data + Industry Audience Research = The Big Opportunity	31
The Big Conclusion	33
Footnotes and Bibliography	34
Additional Reading	36



About the Author

Richard Marks

Director, Research The Media Ltd

Richard Marks runs his own consultancy business, Research The Media (www.researchthemedias.com) and is a regular and award-winning speaker at media conferences in the UK and around the world, focusing on the future of media and the implications for the research industry.

Whilst CEO of Kantar Media Audiences, Richard oversaw the introduction and operation of the new BARB panel from 2010 and has been responsible for TV, online and radio audience research in 60 countries, including the TV currencies in Spain, Russia and China.

In the UK media research world, at RSGB and TNS Richard helped create the ground-breaking ROAR and VIPer cross-media studies, set up a pan BBC tracking study, brought the Daily Life study into the 21st century with the use of electronic diaries and consulted on the first IPA Touchpoints. He has also worked on projects for RAJAR, JICREG, Sky, Channel 4, The Guardian, News International and Sky.

More recently he has focused on digital audience measurement, Big Data and innovations such as pioneering the introduction of Return Path Data in the United States, portable meters and online video measurement.



Foreword

Lynne Robinson

Research Director, IPA

Everyone is talking about Big Data. It seems to be permeating into every aspect of our lives. Professionally, it is transforming how we conduct business and how we define and evaluate success. Personally, we are all busy increasing the size of our digital footprints which are the lifeblood of Big Data.

But, what do we really mean by Big Data and, given that Big Data is only really in its infancy, how will it change our lives?

This paper explores what Big Data means in the context of audience measurement, its advantages and disadvantages and its potential future role. Will Big Data totally replace our existing measurement systems or will it augment and enhance them? When and where is it best to use Big Data and when should it be avoided?

We are delighted that Richard Marks has produced this report for the IPA. Richard's career has embraced Audience Research, Big Data and the integration of the two, which means he is eminently qualified to give an unbiased view of this very exciting new world.

The debate surrounding Big Data is only just starting and we hope that the Big Opportunity is a key input into it.



The Big Opportunity

“If you asked me to describe the rising philosophy of the day, I’d say it is data-ism. We now have the ability to gather huge amounts of data. This ability seems to carry with it certain cultural assumptions — that everything that can be measured should be measured; that data is a transparent and reliable lens that allows us to filter out emotionalism and ideology; that data will help us do remarkable things — like foretell the future.”

– David Brooks, *New York Times*¹

Big Data is dominating research discussions at the moment. With server data, social media, rich customer databases and Return Path Data, there is an exciting opportunity to take media research to the next level.



The Big Opportunity – A Summary

Big Data presents a great opportunity to enhance, even potentially to revolutionise, our understanding of audiences, consumers and their interactions with brands and messages.

How can the media industry make the best use of this data? What are the implications for advertisers, agencies, media owners and the industry currencies? How can we make the most of this opportunity?

The need for wider context

We argue that Big Data – from servers, ISPs, customer databases, social media or Return Path Data (RPD) - can greatly enhance the understanding of media audiences. However its growing availability does not remove the need for transparent industry data that gives a bigger contextual picture of the market as a whole, functioning as a jumping off point into proprietary data sets. We believe that there will remain the need to understand first party and proprietary data sets in the context of wider competitor information. Furthermore, survey data can enhance Big Data sets themselves and vice versa.

- The strength of Big Data is that it typically provides highly granular passive measurement of behaviour, at low cost and at speed: volume, velocity, variety and veracity.²
- The strengths of the industry trading currencies are that they typically provide an overview of the whole market, including the competitive context. They offer individual level demographics derived from managed, balanced samples and, critically, their methodologies are transparent and data access is open to all, as opposed to many Big Data sets that are proprietary or bespoke and (understandably) leveraged for commercial advantage.

Long tail measurement

Big Data certainly has the potential to be more accurate than traditional claimed behaviour surveys, particularly when it comes to niche groups. Claimed recall becomes ever more difficult to measure as the myriad of media choices expands, people multi-task their media consumption and also may filter out some media they are exposed to as they did not choose it themselves. Server or Return Path Data data, if correctly filtered and verified, provides a passive measure of exposure and also is able to measure the long tail of media consumption that sample surveys can struggle to detect.

Understanding what Big Data represents

When working with Big Data it is essential to understand the universe that a data set represents and as a result whether or not it is valid to project that data onto a wider audience. Nonetheless users should not fall into the trap of assuming that just because a data set doesn't represent everything it can't represent something: its own discrete universe – just so long as that is clearly understood. Clarity is everything.

There is no safety in big numbers alone

Bigger is indeed typically better when it comes to sample size, but size is not the only key variable when it comes to assessing the reliability of a data set. Equally important is where the data set comes from, whether it is an opt-in sample, if its balance reflects the market it is seeking to represent and what quality controls are in place to remove or correct errors in the data. A balanced, controlled sample of 1,000 individuals with a high response rate will always be more representative than 1,000,000+ customers drawn from an imbalanced, self-selecting or partial sample. Big does not always mean beautiful.

The importance of demographics

Demographic information remains an important part of the trading currencies, not just for advertising purposes but also for understanding audiences and developing and managing the media brands themselves in a competitive context. Typically the size of a data set will be in inverse proportion to the amount of demographics appended, so again a skilled Big Data user needs to understand the implications of this. It is important to bear in mind that the vast majority of Big Data sources measure devices not people.

Single source opportunities

The potential use of Big Data in audience measurement opens up the opportunity to understand the relationship between media exposure and consumer behaviour via data overlays – the holy grail of single source with large sample sizes. This ability to expand audience measurement into measuring outcomes could be a game changer. However we believe that there will remain a need for attitudinal research to augment this, as not all campaigns are simply focused on purchase. The danger of relying solely on Big Data is that this will potentially limit our wider understanding of the impact of media exposure. Similarly the industry increasingly aspires to conduct research that is consumer-centric as opposed to just media-centric, or silo-driven. Surveys like the IPA's TouchPoints can be used to place Big Data learnings within the context of the customer journey. We need to use Big Data to help us achieve that goal, as opposed to replacing one set of industry survey platform silos with a new set of proprietary data silos. Big Data is the means to an end, but it is not the end in itself.

Big Data presents a quantum leap forward for media research

If used in conjunction with relevant industry survey data, it will amplify the strengths of both to create a whole bigger than the sum of its parts. However depending solely on Big Data will result in severely limited insights.

Changing role of analysts

The likely role of data analysts and researchers is evolving in a Big Data world. Greater in-house research and analytical skills will be needed to guide the selection and interpretation of the data sets available - the client analyst will act as personal shopper, ensuring budgets are allocated to the most relevant and reliable data sets. Meanwhile the balance of research agency expertise will adjust to incorporate a higher degree of data curation as opposed to just data creation. In order to maximise the opportunities presented by both Big Data and industry audience services, the skilled data purchaser and user - the data shopper - needs to be asking the right questions.

Ten key questions to ask about Big Data:

- 1 Who owns or controls the data?
- 2 What aspects of a consumers' behaviour are covered and which are not?
- 3 What aspects of competitor performance are covered and which are not?
- 4 Is the methodology transparent? What quality control steps are in place to ensure the accuracy and integrity of the data? Who has appraised the data set?
- 5 What universe does the data represent? If it is being projected to a wider universe, what methodology is being used?
- 6 Is the data a census, derived from a balanced sample or an opt-in sample? If the latter, what was the response rate?
- 7 Is the data longitudinal – can we examine the same people across time?
- 8 What demographics does the data include? How are they obtained? Are they at an individual or household level?
- 9 Is the data compliant with privacy laws? Is its availability likely to be impacted by changes in privacy laws?
- 10 How much work will be needed to turn it from data to insight?

Ten key questions to ask about Audience Research:

- 1 Who commissions and oversees the research?
- 2 How is the sample recruited? What steps are in place to maximise response rates and sample balance?
- 3 Is the data collected passively or using claimed recall? Both can be appropriate but have implications for granularity.
- 4 Is the data collected from an ongoing panel or from separate samples?
- 5 Is the data date-specific or averaged across a time period? How is audience exposure defined? GRPs, TVRs, daily and weekly reach and frequency need to be clearly understood as they have implications for use and interpretation.
- 6 How frequently is the data made available?
- 7 At what level of granularity is the data available?
- 8 What are the analysis limitations imposed by the sample size? At what level of analysis does the data become unreliable? Is this clearly flagged by the analysis package?
- 9 Is the service subject to audit or Technical Committee oversight?
- 10 Are there gaps in the measurement – new technology not yet covered? What plans are in place to fill these gaps?

What do we mean by Big Data?

Big Data is the term that has caught the industry's imagination; it is almost ubiquitous in news headings, conference papers and infographics. Indeed it could be argued that the era of Big Data has truly arrived because:

- Everyone has a view on it.
- Few can actually define it.
- Some are proposing it should be included as a primary economic category in Government statistics.³
- The debate has started on whether it is an asset and should therefore be taxed by the byte!⁴

Dozens of recently published books on Big Data can be found on Amazon, from *Big Data for Dummies* to *Big Data for Chimps*. However almost all these texts focus on either the business applications of Big Data or the IT implications of storing or accessing petabytes, zettabytes or brontobytes⁵ of customer and usage data.

If you are reading this paper hoping for a debate on the relative merits of Hadoop or Oracle then you are going to be disappointed. The focus here is very much on the media research implications, most specifically for audience measurement. To do this, firstly we need to agree what we mean by Big Data in a research and planning context. There has been much comment, not least from the author himself, that Big Data is a misleading term and might better be called Deep Data⁶, but Big Data is the term everyone is using so let's define that.

“The explosion of mobile device ownership and digital connectivity has turned human communications into a rush of information known as Big Data. Simply put, digital devices leave a trail and every time a consumer clicks, hovers, follows a link, makes a comment or even just switches on their phone, that creates a piece of data which, when correctly collected and interpreted, can be used to understand or infer something about them.”

– Cila Warncke, WARC⁷

In a media research context Big Data means:

- Granular media usage data derived directly from Set Top Boxes, web servers, ISPs, mobile networks and user databases. This can either be very large opt-in data sets, managed selected samples or at a census level.
- Customer purchase or usage data derived from online behaviour or offline via EPOS (Electronic Point Of Sale) purchase data integrated with customer loyalty information.
- Social media data from Google, Twitter, Facebook and other services.

Some have argued that the term Big Data really only comes into play when two or more datasets are combined - for example viewing data and purchase data - to create a sum larger than the parts, but for the purposes of this report we will focus on the wider definition of any large dataset that can be used for media planning and consumer insight.

The wider opportunities for using Big Data are exciting, giving media owners, advertisers and agencies a more granular – and often instant – view of media and consumer behaviour.



The Big Data Revolution

“Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves... There’s no reason to cling to our old ways. It’s time to ask: What can science learn from Google?”

– *Chris Anderson, Wired 2008*⁸

“Chris Anderson...wrote in 2008 that the sheer volume of data would obviate the need for theory, and even the scientific method... [T]hese views are badly mistaken. The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Data-driven predictions can succeed-and they can fail. It is when we deny our role in the process that the odds of failure rise.”

– *Nate Silver, 2012*⁹

The views can be extreme – at one end of the spectrum, the view that Big Data can do anything, with the research department of the future consisting of a computer, a man and a dog (the dog’s task is to ensure that the man does not touch the computer). At the other extreme, traditionalists argue that because Big Data often has holes in coverage or imperfections it cannot be used at all. It will be no surprise to learn that we believe the truth lies between those two extremes. We see Big Data as a revolution. However any historian will point out that revolutions are often bloody and messy with a period of anarchy and the eventual winners not always the wide-eyed idealists who manned the barricades at the start. Managed transition, whilst less exciting, is usually more effective.

So where does that leave traditional research practices – doomed inside a castle of established research methods, under siege from Big Data? ¹⁰ Does Big Data wipe away the established order as some are implying?

“Corporate market researchers believe that the leading agency in 2020 is just as likely to be Google, Facebook or a company from outside the industry as it is to be one of the ‘old guard’”

– David Nunan & MariaLuara Di Domenico, *IJMR* ¹¹

Well, within the media industry itself, innovation does not wipe out all that has come before it – television did not kill radio or cinema, the internet has not killed television. Typically innovation takes its place within the ecosystem, improving and complementing but rarely totally replacing. We can see this evolution model if we look back a couple of decades to the advent of EPOS scanning data in consumer research. The availability of this rich and granular new form of purchase data refocused how research was used, but services like Worldpanel and HomeScan adapted to incorporate its use and attitudinal and contextual studies like TGI continue to this day.



“The term ‘big data’ leads logically to there being data that is ‘small’, and many analysts have attempted to differentiate this from the complex and unstructured data sets that comprise big data. The analysis of huge amounts of data begins at choosing smaller amounts to organise and process; this type of data (contained, cleaner and more accessible) is known as ‘small data’. Often it is these smaller, more agile, datasets that can provide the real value in terms of insights drawn.”

– *eConsultancy*¹²

So a key theme for this paper is: how can the media and advertising industry make the most of Big Data? Can Big Data and survey research coexist? Firstly, let’s compare the strengths of each.



What are the strengths of Big Data?

In an audience measurement context, Big Data is certainly a game-changer in many ways:

Passive measurement

Media and consumer behaviour in a digital environment leaves a digital trail - be it tuning on a satellite box, surfing behaviour, click-throughs or online purchase. This data is potentially of a more granular and accurate nature than could ever be detected via a traditional media research survey – both in terms of the amount and detail of information a respondent would be willing to give – or indeed in terms of their ability to accurately recall it.

Granular measurement

What we have is effectively clickstream data - in many cases more data than we actually need - allowing a deeper dive into the data than has been possible with survey and panel-based research, both in terms of the richness of data for each household or individual and in terms of the numbers of records available. So, whilst the advent of digital distribution has facilitated a fragmentation of media choices and extended the long tail of media use, it also provides as a side-product the means to understand that long tail.

Cost

Data collected as a side product of behaviour – digital exhaust – is inevitably far cheaper to gather than bespoke survey research, particularly as costs of data storage plummet.

Speed

The opportunity for instant – or close to it – feedback on behaviour is exciting as we are in a real time age. The opportunity for Real Time Bidding (RTB)¹³ via digital exchanges, perhaps even for TV, excites many.

Data Overlays

The potential to overlay Big Data about media exposure with Big Data about purchasing and consumer behaviour is greater due to the size of the databases.



What are the strengths of the Industry Currencies?

Industry currency research is primarily sample survey based. This has limitations, which is why Big Data has so much to offer, but the currencies do have a number of essential, unique and complementary characteristics that set them apart from Big Data.

A balanced sample

The importance of representativeness will be discussed later on in this report.

Coverage of the whole market

In this sense it could be argued that the industry currencies are already Big Data as they strive to represent the big picture, providing a competitive context for decisions, measuring the totality of the market as opposed to one service. BARB processes each day minute-level data from over 11,000 individuals. NRS interviews 36,000, and RAJAR 110,000 people each year.

Transparency

Typically the surveys are available at a granular level with open methodology and industry oversight via technical subcommittees. Transparency is vital if a currency is to be trusted by its users.

Individual level demographics

As respondents consent for their data to be used, full household and individual level demographic detail are available, essential for profiling, targeting and evaluation.

A managed panel/survey

In the case of BARB which tracks a panel, respondents know they are on a panel and consequently the research agency can check and query anomalies with panel members and ensure a high level of sample continuity from day-to-day, essential for reach and frequency calculation.

Data Availability

Data is typically available to all. Some Big Data sets by their very nature have limited access – such as Facebook, or are in house databases leveraged for commercial advantage. So an attempt to stitch together a total market view from these myriad sets, negotiating politics, PII legislation and political negotiations is unlikely to succeed.

It can be argued that there is a complementary relationship between the two types of data. The following table summarises the relative strengths of each.



Big data & industry surveys compared

	Big Data	Industry Surveys
Coverage of total market	Low Mostly first party or platform based	High
Passivity of measurement	High	Low – BARB excepted
Granularity / long tail measurement	High	Low – as sample centric
Individual level demos	Low	High
Potential to measure addressable / targeted advertising	High	Low – as sample centric
Transparency / impartiality	Low	High – as industry funded
Managed, balanced samples	Low – although less relevant if true ‘census’	High
Potential source for single source / overlays	High	Low – due to limited samples
Speed of delivery	High	Low – BARB excepted
Availability to all	Low	High
Data collection cost	Low	High

Maximising the potential of Big Data

IBM has summarised the potential strengths of Big Data as: volume, velocity, variety and veracity.¹⁴ So how can media research best harness those strengths? How can we ensure that, when it comes to media and advertising, the Big Data revolution is glorious rather than bloody? Let's examine some of the points above to best understand how and where we can safely use Big Data in a media research and specifically audience measurement context.



1. The Bigger Picture: Context

We are used to our industry surveys – BARB, RAJAR, NRS etc. – providing not just detailed feedback on the performance of our piece of content or advertising, but also the wider context. How did we perform relative to the market? What is usual? What constitutes success? A key challenge for Big Data is that it typically can represent a part of the picture but not the whole. Sky’s SkyView and Sky IQ services for example provide extremely granular insights on their customer behaviour, but Sky have never claimed that they represent anything other than Sky subscribers in the UK. For competitive context they – and their advertisers - still rely on BARB to understand how Sky performs in the wider market. Denied that wider context and competitive information media owners and advertisers would find themselves in a solipsistic bubble – with a detailed knowledge of their own customers behaviour, but blind to the rest of the market. However it can be argued that Big Data chimes with the brands growing focus on their own customer base, on customer retention, ARPU¹⁵, lifetime value and share of wallet as opposed to acquisition:

“The prevailing view for how Big Data should operate has often been for brands to analyse their own proprietary data set generated from their own digital assets. This is then used to optimise the consumer interaction with their brand. As such, a rationale has emerged for retaining this analysis work in-house. Yet ‘walled-garden’ approaches only give part of the story, failing to address the activity that led the consumer to touch the brand.”

– Colin Strong, GFK¹⁶

In that sense audience measurement is at a crossroads. Will media owner and advertiser first-party Big Data become a jealously guarded weapon, used to leverage competitive advantage or will the long-established shared currency system still prevail? Data that gives genuine insights is an asset. However, reports from multiple data sets offering conflicting views of success or failure - something that websites have always had to deal with – can be actively dangerous.

We believe that the future is not either/or but rather that industry currencies, providing competitive context and in-house data sets, will coexist with and complement Big Data. Without that wider context, media brands will be left with a myopic view of the market. Building a digital relationship with media consumers rather than just counting them is essential. However data in isolation, trying to use data sets without an understanding of how they slot into the universe as a whole, could suck us into a research black hole; masters of the minutiae of our own domains but blind to the outside world.¹⁷

So when it comes to media research, the focus needs to be on how Big Data can coexist with and benefit the industry surveys, as opposed to a knee-jerk assumption that they will be replaced. Used with appropriate care, Big Data can drive our industry forward. But trying to use datasets without an understanding of how they fit together won't just halt progress - it might set the industry back. So this report examines some of the ways in which the industry currencies could be enhanced by Big Data.

2. What does Big Data represent?

As discussed above, a key factor in the usefulness of any Big Data set is clarity about what it represents and what it does not.

A data set that represents all subscribers to a cable TV network, for example, can give a useful and granular look at how those particular subscribers behave. However on its own it is unlikely to be somehow projectable to the total TV universe, no matter how large the sample. Quite apart from differing package lineups, channel availability and user interfaces there may be other reasons why people have opted for this service over another (cost, availability, telephony services, demographic appeal) that may manifest in differing viewing or behavioural patterns.

Further, it would be possible to overlay, for example, data from one cable company, one ISP and purchasing behaviour from one store card to get a combined view of TV, online and purchasing. However it would be reflecting people who had that particular combination of services – highly useful for qualitative insights and guidance on planning, but an even smaller group within the population as a whole and not usable as a currency measurement of the total market. Unfortunately this aspect can get lost, either in the marketing of a new service which can be reticent to remind potential users or investors about the limitations of a product, or due to a simple misunderstanding of where the data comes from in the first place.

This does NOT mean that the data cannot be used – it means that it needs to be used carefully:

- Because something is representative of something does not mean it can be assumed to represent everything.
- However because something is not representative of everything does not mean it cannot be representative of something.



3. Safety in (big) numbers?

“There is a problematic underlying ethos that bigger is better, that quantity necessarily means quality... Without taking into account the sample of a data set, the size of the data set is meaningless.”

– Danah Boyd & Kate Crawford, *Microsoft*¹⁸

Perhaps the biggest challenge to the correct use and deployment of Big Data in media research is the common assumption that the most important characteristic of a sample is simply its size. Undoubtedly, large samples – or census data – allow a more granular look at behaviour and the data can be mined at a more detailed level to gain insights. However, just because the data set you are analysing has a lot of zeros on the end of it, this does not mean that that it is therefore the only health check that matters. If that were the case then the audience measurement companies would have long ago moved all their data collection to cheaper methods like postal self-completion surveys or online access panels. However industry currencies have largely retained at their heart strictly controlled sample-based recruitment as it is understood that a representative survey is made up of three components:

1. A robust sample size
2. A balanced and representative sample
3. A high response rate – high enough that those who do participate in the survey can be representative of those who did not.

So when it comes to Big Data, big does not always mean beautiful: when appraising a data set, two key questions are:

1. What is it seeking to represent?
2. How representative is it in terms of balance?

There is a reason why the BARB sample size has been set at 5,100 homes: the currency needs to balance sample size against the quality and balance of the sample being recruited. BARB could pick 30,000 homes off a mailing list, or put an advert in *Metro* and recruit a large sample, but it would not be a representative sample. Most likely it would also be heavily weighted towards people who actively want to be on a survey about TV – because they watch a lot of television! Low and non-viewers would be less likely to volunteer and so would need more persuasion.

The mission of the currencies is to represent the whole UK population and that means balance not just raw numbers. However that need for balance and statistical rigour does mean that the currencies can struggle with available sample when it comes to niche and long tail activities, which means that Big Data has a lot to offer when used alongside the currencies to aid insight.

4. Opt-in and convenience samples

As touched on in the above example, another key determinant of the reliability of a Big Data set is how that sample is drawn. If it is a genuine census – i.e. all customers - then this may be less of an issue but often the sample, however large, may be an opt-in or convenience sample.

An opt-in sample may be because respondents needed to agree to something first before their data can be used – clicking an agree box, signing up for a Loyalty Card, responding to an offer. Here the key question is how many people did not agree and so how representative is the sample to the total universe?

A convenience sample is a term used to describe a sample that is easily available as opposed to one that is systematically sampled – this could include a database of people who had entered a competition or attended an event. These can contain impressively large numbers but again will not necessarily represent the total universe. A managed sample is usually preferable – this involves selecting a subset of the sample and filtering for key variables. A balanced, smaller and managed sample will usually be more representative than a larger, convenience sample.

5. Demographics

An important topic related to Big Data is the perceived importance of demographics. In this context Big Data can take many forms:

- A data set in which individual demographics are known – for example a customer database where all customers have to complete a questionnaire on joining / subscribing.

- A data set in which some limited demographics are known – for example subscription packages etc.
- A data set with no demographics appended.
- A data set that has been geo-demographically coded using estimates such as Experian. Typically postal code is used to model demographics for a household as a whole.
- A data set which is at a household as opposed to individual level. For example, data from television Set Top Boxes or PCs will include channel tuning or surfing information but is attributed to a device rather than a person. Again modeling techniques can be used to project individual behavior.

So this boils down to two key questions: is the behavioural data at a household or personal level and what, if any, demographic data is available, again at a household or individual level?

Some have argued that the future of advertising lies in household rather than demographic targeting, so it matters less who is in front of the TV or PC screen at any one point. However given that the holy grail of targeted advertising is the elimination of wastage, it seems unlikely that advertisers will want to exchange one form of wastage for another. Also, to be cynical for a moment, there is a clear correlation between those who argue that demographics are not important and those who do not have access to them in the first place.

Geo-demographic profiling can be an acceptable compromise, but it does need postcode data to be available, which is more readily available in customer and subscriber databases and its use is subject to privacy legislation. Geo-demographics are estimates and probability based as opposed to definitive, so will also include wastage.

This does not mean that the absence or limited availability of demographics means Big Data is therefore not of use, far from it. However it again amplifies the continued importance of the industry surveys, which do provide actual, as opposed to modeled, individual level demographic detail essential to media owners and agencies.

In this context the publishing industry is a useful indicator of the future role of the industry surveys, or JICs. Arguably one form of Big Data has been available to newspapers and magazines for many decades: circulation and distribution data. Highly granular, this information lends itself to data mining and statistical modeling and indeed some of the larger publishing groups have employed statistical gurus to pore over their Big Data and project sales to help minimise wastage in print runs. However, the existence of these huge and rich data sets does not lead to any diminution of the roles of NRS nationally or JICREG locally. There was, and is, still a clear need for survey-based research to provide four important variables:

1. Readership as opposed to purchase
2. Who was reading, how frequently and for how long?
3. What sections did they read?
4. What other publications did they read?



None of these questions could be answered by circulation data. However combining circulation with readership data does allow the calculation of a readers-per-copy figure by demographic that can be applied back on to the circulation or distribution trends to create a more dynamic database than periodic survey data alone can provide. This is the basis of the JICREG database, which is the first industry survey to use Big Data as an input.

It is interesting that one of the earliest examples of Big Data in television media research, SkyView, also follows this model: the tuning data from 30,000 Sky homes has the demographic data from BARB applied to it to create the television equivalent of a Readers-Per-Copy - a PIV (Probability of Individuals Viewing).

The importance of demographics and individuals, as opposed to just device behaviour, is an example of how Big Data can be enriched by survey data, whilst the JIC currencies can in turn be enhanced, allowing more granular measurement of niche and long tail media. This is already in practice with the UKOM hybrid data set produced by comScore.

6. The Single Source Opportunity

Big Data can provide a wealth of behavioural information, tracking consumer purchasing or media consumption - ideally both. The main focus of this report is on Big Data in the context of audience measurement. However one key advantage that Big Data has is the potential - via data matching or overlays - to align exposure data with purchase data, to go one step forward, to provide single source measurement of both exposure and outcomes.

In some cases it can be argued that this data will be inherently more accurate than claimed behaviour, particularly when it comes to niche groups. Claimed activity becomes ever more difficult to measure as the myriad of media choices expands, people multi-task their media consumption and also may filter out some media they are exposed to as they did not choose it themselves. Server or RPD data, if correctly filtered and verified provides a passive measure of exposure. When overlaid with purchasing data, the ability to see how one form of behaviour – media exposure – has influenced another – consumption - is a quantum leap for media planning and Return On Investment (ROI).

However using this data does need two factors to be taken into account – the relative importance of attitudinal behaviour and long-term versus short-term ad effects.

- Some Big Data proponents have argued that the ‘what’ is all that is needed, and that statistically analysing the what will indicate the why – motivations – through correlation and pattern recognition. However this can run the risk of apophenia¹⁹ – seeing correlations and patterns that are not relevant - so we believe there will always a need for attitudinal data to complement audience measurement as opposed to just purchase behaviour data.

- Big Data can measure the final act – the click that orders the product, but for many brands and products that final click is the end of a media journey that has seen them exposed to many different messages across a range of media. Use of Big Data can do a lot to help measure the relevance and effectiveness of advertising in terms of proximity of media exposure to the actual decision, but the wider impact of many campaigns is as much about changing attitudes and consideration.

“Taking the emphasis off of the last click model takes an active effort, and that effort is most effective when deployed programmatically, as part of your data platform strategy. Success means building attribution schemes in your platform that look at multiple sources, including panels, social activity and affinity data, and campaign performance.”

– Econsultancy²⁰

For more on long and short term advertising effects, a recent IPA report, *The Long and the Short of It: Balancing Short and Long-Term Marketing Strategies*²¹ is available.

It may be that social media data can get us some of the way to understanding the attitudinal element and this will be an important area to monitor. However the ability to track such data at an individual level is more limited and it will be skewed towards more active and vociferous consumers. Nonetheless social media data does have long term potential to supplement the what with the why.



7. Consumer-centric vs Media-centric Insights

“I believe in research. Personal, one-on-one research, not the stuff that comes from reports. You’d be surprised what you can learn about tobacco sales from talking to a busboy”

– Don Draper, *Sterling Cooper & Partners* ²²

Earlier we talked about the importance of contextual information. Big Data is by its very nature service-centric. It covers the users of a particular ISP, a particular store, a particular TV service. It can provide a deep dive into specific media behaviour, but it is only partially consumer-centric as it does not provide the totality of a consumer’s media exposure and experience. In exploiting the advantages of Big Data, we need to remind ourselves that people do not exist in media silos, that we need a consumer – as opposed to brand or media – centric view to truly understand behaviour and therefore make the right choices at a high level about our choice and mix of media channels.

Consequently, surveys like the IPA’s own TouchPoints will prove even more important in providing that holistic consumer-centric view that will allow us to deploy big data sets in the most effective and insightful way, allowing a more three-dimensional, consumer-centric insight into behaviour.

8. Verification and Accuracy

“Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.”

– Geoffrey Bowker, *University of California* ²³

Survey-based data is typically held to extremely high standards in terms of quality. The industries JICs have technical sub-committees ensuring industry oversight of these critical currencies that drive the allocation of billions of pounds of advertising spend. For Big Data to truly come of age, and for it to be used as trading currency, then a far greater degree of transparency will be needed to reassure as to the accuracy of the data.²⁴

As we have discussed Big Data is already greatly adding to our understanding of media and advertising but there can be a dangerous misconception that because much Big Data is not sample based, therefore it is inherently the truth and therefore less effort is needed to establish its accuracy and representativeness. However biases can exist in terms of the coverage of the data and even Big Data can have mistakes in it – as any consumer will know, customer databases are not perfect:

“Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big data equation as the numbers themselves... In the near term, data scientists should take a page from social scientists, who have a long history of asking where the data they’re working with comes from, what methods were used to gather and analyze it, and what cognitive biases they might bring to its interpretation”

– Kate Crawford, *Harvard Business Review*²⁵

Return Path Data (RDP) from Set Top Boxes is one of the earliest forms of media research Big Data. However at an early stage it was realized that the data coming back from Set Top Boxes could not be used in its rawest form – it needed extraction, filtration and verification. Using the data in its raw form was, as George Shababb of TNS put it at the time, like trying to “drink from a fire hose”.

So the research agencies producing RPD have put processes in place to account for them, underlining the importance of checking the quality and reliability of a data set however reassuringly big it appears. Otherwise the damage to your business could be the equivalent to that done to a car if it attempted to drive on crude oil. Establishing veracity is vital to the credibility of these new sources of data.



To quote IBM:

“1 in 3 business leaders don’t trust the information they use to make decisions. How can you act upon information if you don’t trust it? Establishing trust in big data presents a huge challenge as the variety and number of sources grows.”

– IBM²⁶

Particularly when it comes to media and advertising research, where the stakes are so high, data used for decision making has to be held to the same high standards and transparency that we have become used to from our industry currencies.

9. Skillsets

“Far too many companies believe that 95 percent of their data and analytics investments should be in data and modeling. But unless they develop the skills and training of frontline managers, many of whom don’t have strong analytics backgrounds, those investments won’t deliver. A good rule of thumb for planning purposes is a 50–50 ratio of data and modeling to training.”

– McKinseys²⁷

The advent of the Big Data era also presents questions about analysts and researchers themselves and in particular the balance between research agency and client side data skills. It may be that the proliferation of Big Data will require an expansion of relevant client side resource. Client analysts may become like personal shoppers for their companies, evaluating and in some cases improving external and in-house datasets, using the criteria outlined in this document.

It is clear that the role and significance of data will only grow. Some claim that the advertising industry is transitioning from the ‘Mad Men’ era to a reliance on ‘Math Men’.²⁸ We believe there will continue to be a role for survey-based research in that mathematics. However for research agencies, there may be a subtle shift in their role to becoming both creators and curators of data. Up to now research agencies have been primarily farmers of data – designing and growing bespoke research data. Arguably the skill set and role is evolving to that of being chefs – selecting a combination of homegrown and bought-in Big Data to cook insightful meals for clients.

“Big Data has tended to come with its share of Big Hype. So long as we’re realistic about its potential, and recognize that our data is only as useful as the human intelligence we bring to it, minus the human biases with which we burden it, Big Data should, indeed, pay significant dividends.”

– Matt Asay, 10gen²⁹

Big Data may shift the balance of what analysts and researchers do - and where they reside in the advertising eco-system – but the need will remain for the human factor in appraising and analysing data sets.



10. Privacy

“Teasing out how to walk the tightrope of privacy and publicity is going to be a critical challenge of our era. It needs you to remember what the data is that you’re chewing on. Never forget that Big Data is Soylent Green.³⁰ Big Data is made of people”.

– Danah Boyd, Microsoft³¹

This report cannot do justice to the whole privacy debate around Big Data, but the two key questions important for our evaluation are:

1. What steps are taken to comply with data privacy?
2. What implications does that have for the representativeness of the data set? Does it mean I am only looking at a subset of people who opted-in? Or does it mean I am looking at just anonymised records without demographics?

Also important is to understand how privacy legislation is likely to evolve, particularly when it comes to overlaying individual information from different sources. What is legal now may not be so in the future, so thinking ahead will be important - baking a data set into your business only to find it is subsequently unavailable, or scaled back, can present headaches.

“There is the risk that market research will be ‘left out of the loop’ as organisations strive for the commercial benefits brought by big data without consideration of the needs for appropriate consideration of personal privacy.”

– David Nunan & Maria Laura Di Domenico, IJMR³²

In the earlier table (page 17) the relatively low costs of Big Data were highlighted. However two factors could drive those costs up:

1. Tighter privacy legislation, particularly around opt-in as opposed to opt-out, could limit the volumes of sample that can be used, even anonymously.
2. As users appreciate the value of their own personal data, then some have speculated that the big winners could be the customers themselves who monetise their own personal data, only surrendering it in return for goods and services.

Either way, it will be important that Big Data is not left out of the loop when it comes to industry oversight, transparency and conforming to industry guidelines on privacy.

11. Big Data + Industry Audience Research = The Big Opportunity

The big theme that emerges from this discussion, the one golden thread is that Big Data presents immense opportunities and is a game changer. However to truly harness its potential and move from Big Data to Smart Data this first party data needs to be used in conjunction with third-party data sets such as the Industry Currencies and relevant survey data.

Data from servers, ISPs and media providers can greatly enhance the granularity of audience measurement. The industry currencies struggle to measure the long tail of content due to limitations in sample size, data collection or methodology. Using Big Data in conjunction with the industry measurement systems presents a way forward that will enhance both first and third party data. Many media Joint Industry Committees around the world have announced initiatives to understand how Big Data could be incorporated into currency measurements. BARB has recently launched its Dovetail initiative³³ examining how server data could be incorporated into the measurement alongside sample-centric measurement, whilst Nielsen in the US recently committed to incorporating Set Top Box data into its local market measurement to replace diary-based measurement.³⁴ MMS in Sweden³⁵ and SKO in The Netherlands³⁶ are publishing video streaming data alongside the core TV currencies. The industry currencies – measuring the whole market at an individual level – are uniquely positioned to provide the context, to being a hub for the incorporation and understanding of first party data sets. This will really allow Big Data to sing. In this context it really is a case of evolution as opposed to revolution.



“Beyond first-party data, third-party intent data can be extremely effective. We’ve seen the best performing third-party data drive returns on par with remarketing”

– Matt Gritzer, *Accordant Media* ³⁷

“Audience measurement and cookie-based targeting supplement each other greatly. Audience measurement creates a standard yardstick that enables marketers to analyze various media avenues, their reach, and effectiveness. Cookie-based targeting on the other hand is a personalized tool that marketers can use to activate their audience data. Audience measurement data helps in decision-making, but it is not actionable like cookie-based data”

– Andy Monfried, *Lotame* ³⁸

To simply assume that Big Data will replace survey based data is to do a disservice to Big Data and actually limit its potential. Bringing together Big Data and survey data really is a case of $1+1 = 3$. Big Data is enhanced and becomes even smarter, more three-dimensional. Big Data is able to reach the parts that survey data alone cannot. Industry survey data is uniquely able to provide the big picture, the context of the market as a whole, with data that is transparent and widely available.



The Big Conclusion

As will have been apparent, we see huge potential in Big Data, but we want to ensure that users make the most informed and therefore optimum use of the new resources available. To do that we need to be informed consumers, knowing what questions to ask as we make our data choices. By focusing on understanding the scope and representativeness of the data and then using it in combination with other data sets, we can move beyond just Big Data to Smart Data:

“Without understanding the consumer mindset there is a danger that Big Data never becomes ‘Smart Data’ and as such the value for marketers is limited. Intelligent integration brings consumer and research knowledge together with Big Data to avoid the risk of chasing the wrong questions and finding false positives.”

- Colin Strong, GFK ³⁹

Potentially, Big Data and survey data are a marriage made in heaven. We are very excited about the groundbreaking opportunities this offers, but we see the education of users as hugely important. Users are advised to consider not only the technical aspects of Big Data - storage and access - but also the research aspects, in terms of the coverage and veracity of the data.

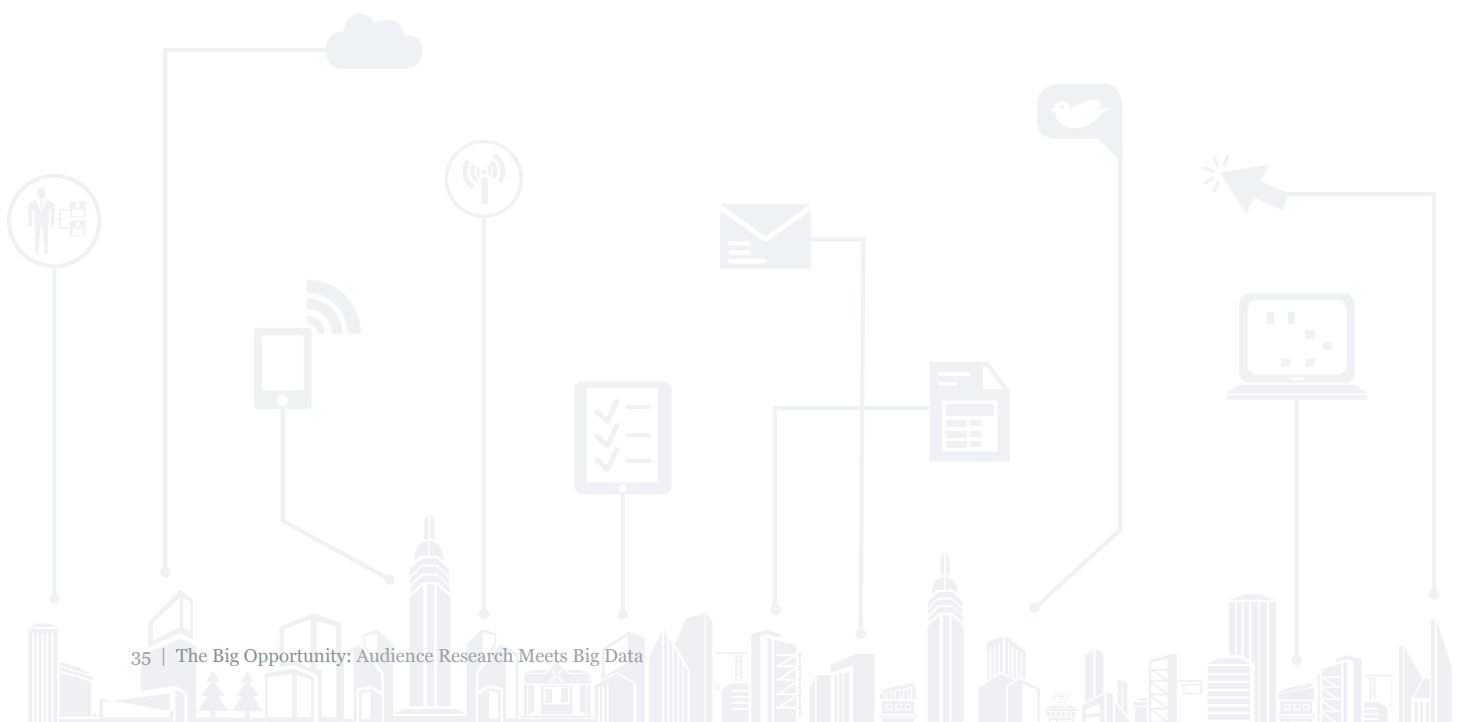
Users need to understand the possibilities and limitations of all the data available to them whatever the source.



Footnotes and Bibliography

- ¹ The Philosophy of Data – David Brooks, New York Times Feb 24 2013
http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html?_r=1&
- ² What is Big Data? - IBM <http://www-01.ibm.com/software/data/bigdata/>
- ³ Beyond Goods and Services: The (Unmeasured) Rise of the Data-Driven Economy - Michael Mandel, Progressive Policy Institute, April 10 2012 <http://www.progressivepolicy.org/2012/10/beyond-goods-and-services-the-unmeasured-rise-of-the-data-driven-economy/>
- ⁴ Elusive Big Data: The Thing And Not The Thing – K.N.C, The Economist, Feb 18 2013
<http://www.economist.com/blogs/graphicdetail/2013/02/elusive-big-data>
- ⁵ <http://www.whatsabyte.com>
- ⁶ Big Data - or Deep Data? – Richard Marks, Mediatel Newline 26 March 2013
<http://mediatel.co.uk/newline/2013/03/26/big-data-or-deep-data>
- ⁷ Big Data Strategies: Key themes from I-Con 2012 – Cila Warncke WARC
- ⁸ The End of Theory: The Data Deluge Makes The Scientific Method Obsolete – Chris Anderson, Wired, June 23 2008
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
- ⁹ The Signal & The Noise – Nate Silver, Penguin Press 2012
- ¹⁰ Why Google Will Crush Nielsen – Frederic Filloux, Guardian 20 May 2012
- ¹¹ Market Research & The Ethics of Big Data – David Nunan & MariaLaura Di Domenico, International Journal of Market Research Feb 2013
- ¹² Big Data Trends Briefing – Digital Cream 2013 - eConsultancy
- ¹³ 5 Ways Real Time Bidding Differs From Direct Buys – Ratko Vidakovic, Market Apr 1 2013 <http://marketingland.com/5-ways-real-time-bidding-differs-from-direct-buys-37384>
- ¹⁴ What is Big Data? - IBM <http://www-01.ibm.com/software/data/bigdata/>
- ¹⁵ <http://glossary.reuters.com/index.php?title=ARPU>
- ¹⁶ Big Data: It May Be Big But Is It Smart? - Colin Strong, GFK
- ¹⁷ Data, Data Everywhere – Richard Marks, RTS Cambridge Sept 2011
- ¹⁸ Six Provocations for Big Data - Danah Boyd & Kate Crawford 21 Sept 2011 http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431
- ¹⁹ *ibid*
- ²⁰ Best Practices in Data Management – Econsultancy Dec 2012
- ²¹ The Long and the Short of It: Balancing Short and Long-Term Marketing Strategies - Les Binet and Peter Field, IPA, June 2013
- ²² <http://blogs.amctv.com/mad-men/2008/07/interview-with-don-draper.php>

- ²³ Memory Practices in the Sciences - Geoffrey Bowker, 2005 MIT Press, Cambridge, Massachusetts.
- ²⁴ Horses For Courses – George Shababb and Richard Marks, ESOMAR Media Symposium Berlin Oct 2010
- ²⁵ The Hidden Biases in Big data – Kate Crawford, Harvard Business Review, April 2013
http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html
- ²⁶ What is Big Data? - IBM <http://www-01.ibm.com/software/data/bigdata/>
- ²⁷ Big Data – What's Your Plan – Stefan Biesdorf, David Court & Paul Wilmott, McKinsey Quarterly, March 2013
http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan
- ²⁸ The Mad Men Years Are Giving Way To The Math Men Era – Chris Moore, AllThingsD, Jan 20 2012
<http://allthingsd.com/20120120/the-mad-men-years-are-giving-way-to-the-math-men-era/>
- ²⁹ Nate Silver Gets Real About Big Data – Matt Asay ReadWrite 29 March 2013
<http://readwrite.com/2013/03/29/nate-silver-gets-real-about-big-data>
- ³⁰ <http://www.urbandictionary.com/define.php?term=soylent%20green>
- ³¹ Privacy & Publicity In The Context of Big Data – Danah Boyd, Microsoft, 29 April 2010 paper to WWW
<http://www.danah.org/papers/talks/2010/WWW2010.html>
- ³² Market Research & The Ethics of Big Data – David Nunan & MariaLaura Di Domenico, International Journal of Market Research Feb 2013
- ³³ www.barb.co.uk/whats-next
- ³⁴ www.mrweb.com/drno/news15835.htm
- ³⁵ http://www.mms.se/wp-content/uploads/_dokument/rapporter/webbrapporter/English/MMS,%20Annual%20Web%20TV%20Report%20-%202012.pdf
- ³⁶ http://www.kijkonderzoek.nl/component/Itemid,47/option,com_kijkcijfers/file,ds-o-p
- ³⁷ Quoted in Best Practices in Data Management – Econsultancy Dec 2012
- ³⁸ ibid
- ³⁹ Big Data: It May Be Big But Is It Smart? - Colin Strong, GFK





Institute of Practitioners in Advertising
44 Belgrave Square
London, SW1X 8QS

www.ipa.co.uk
lynne@ipa.co.uk
+44 (0)20 7235 7020

www.ipa.co.uk/document/The-Big-Opportunity-Full-Report

© IPA June 2013

Designed by APDA-London



[@IPA_Updates](https://twitter.com/IPA_Updates)



[@IPAScotland](https://twitter.com/IPAScotland)



flickr.com/theIPA



youtube.com/theIPA



facebook.com/theIPA



bit.ly/IPALinkedIn

