The history of the data economy

Part I: The birth of customer insight. By **Timandra Harkness**

The world has changed dramatically over the past 200 years, and the economy has changed with it. Data is now the fuel that drives business – identifying potential markets, shaping new products and targeting would-be consumers.

Practically every major company is investing in statistics, data science and artificial intelligence. Such investments are made so that data can be analysed to identify trends, patterns, and opportunities to deliver a competitive edge. This is true not only for the sectors we typically associate with data, such as retail or finance, but also for categories as diverse as car manufacturers, real estate agents and educational institutions.

Supporting these organisations is an entire industry (multiple industries, in fact) selling, analysing, creating, distributing, or finding other novel ways of profiting from data.

As a share of gross domestic product (GDP), the so-called "data economy" is relatively small. In the European Union, for example, the data market was valued at €75 billion in 2019, with an overall economic impact of about €400 billion – roughly equivalent to 0.5% and 2.5%, respectively, of the EU's total GDP (bit.ly/3sllJmt). But, as a United Nations report explains, the data economy's share in GDP belies its "real market size and economic influence" (bit.ly/3rWehPD). Consider, for example, that Microsoft, Apple, Amazon, Alphabet and Facebook – five of the biggest technology and data firms in the world – had a combined market valuation of more than $5 trillion in 2020. That is more than 10% of the total value of the US stock market (bit.ly/3anTcI6) and almost 6% of global stock market capitalisation (cnb.cx/2ZkMa09).

This is likely just the start of what is to come. As the UN report says: "Data is shaping the future of humanity." So, in order to understand where we may be heading next, *Significance* and *Impact*, the magazine of the Market Research Society, have jointly commissioned a four-part series to explore the past, present and future of the data economy. In part 1, Timandra Harkness tells the story of the birth of customer insight.

Brian Tarran Editor, *Significance*
Katie McQuater Editor, *Impact*

Although today's data-driven economy relies heavily on mathematics, statistics and computer science, its roots owe as much to pragmatic trial and error as to pioneers of social statistics like Adolphe Quetelet. While theoreticians wrangled over how humans varied, and how to quantify this, those who saw the value of data in the nascent mass society were already collecting and using it.

The first people to treat public opinion as a form of data were newspaper publishers in nineteenth-century America, who used "straw polls" of their readers to anticipate election results in print. The *Harrisburg Pennsylvanian*'s 1824 presidential election poll is often cited as the first political poll. It accurately predicted that Andrew Jackson would win the popular vote, though John Quincy Adams was ultimately elected President.

This straw poll approach continued in use till the 1930s. Although they actively went out to survey different groups in the population, newspapers relied more on very large numbers of responses than on any statistical theory to accurately reflect the mood of the nation.

Meanwhile, America's flourishing mail-order market made customer information so valuable that letter-brokers bought and sold customer letters. Those letters, originally solicited by newspaper adverts or leaflets, might include not only names and addresses but also useful details such as medical histories. In 1910, Louen Atkins of Chicago accused his former business partner James Rainey of taking data from his mailing list to poach a customer. The dispute culminated in Rainey shooting Atkins dead.[1]

The research profession

In the early twentieth century, this kind of ad-hoc research began to take more coherent form with the professionalisation of marketing. Market research pioneer Archibald M. Crossley reports applying for a job with a Philadelphia advertising firm in 1918. "My prospective employer asked how I would like to set up a research department. I said: 'I would. What is it?' And his answer was: 'I don't know either.'"

Crossley did some research into research and found that many other advertising agencies already had research departments. Among them, the most influential was probably J. Walter Thompson (JWT), founded in 1878 (and still trading today as Wunderman Thompson, part of the WPP marketing group). Stanley Resor, who took over JWT in 1916, believed that human behaviour, taken *en masse*, could only be understood through statistical and scientific study.

Some magazine publishers also had research departments to help them attract lucrative advertising and make it more effective. Charles Coolidge Parlin, widely regarded as the world's first professional market researcher, was hired by the Curtis Publishing Company in 1911. His extensive research into entire sectors, first agriculture and then automobiles, produced volumes of data and analysis.

Archibald Crossley, duly informed, set up a research department for his new employer. After a stint in the research department of the *Literary Digest* newspaper, he went on to found his own research company in 1926. By Crossley's account, this kind of quantitative research started out partly as a sales technique for advertising firms themselves, to help them compete for clients, but came into its own with the growth of mass media. Advertising spend in the United States increased tenfold between 1900 and 1930, and clients wanted to know whom their radio, and later television, adverts were reaching, and to what effect.

Coming at the same question from another direction, Arthur C. Nielsen set up a business to test the quality of conveyor belts and turbines in 1923, before applying similar methods to market research. As an engineer, Nielsen applied rigorous statistical techniques of probability sampling to new problems like calculating brand market share, and later to measuring broadcast audience habits.

A revolution begins

George Gallup revolutionised quantitative market research by bringing together statistics, journalism and psychology. The method he outlined in his doctoral psychology dissertation, "A New Technique for Objective Methods for Measuring Reader Interest in Newspapers", transferred to human attitudes the method used by inspectors of wheat or water – testing a number of small samples to assess the whole (see "Sampling: Statistical divisions", page 15).

While working as director of research for New York advertising agency Young & Rubicam, Gallup began to widen his focus beyond studying consumer responses to journalism and advertising. This sampling approach could equally be applied, thought Gallup, to public opinion on politics and social issues. In 1932 his research helped his mother-in-law, Ola Babcock Miller, to election as Iowa's Secretary of State. In 1934, Gallup's predictions came within one percentage point of the congressional election results.

The final overthrow of the newspaper straw poll by more statistically robust methods came in 1936. Now running the American Institute of Public Opinion from a small office in Princeton, Gallup used the results of his surveys to produce a regular syndicated column, *America Speaks*.

Gallup challenged the best-performing of the newspapers, the *Literary Digest*, to beat his methods with their straw-poll forecast of the presidential election results. He judged, correctly, that the *Digest*'s straw poll over-represented people with telephones and cars, who were unlikely to vote for Franklin Delano Roosevelt, the Democratic candidate. Gallup's surveys used a quota system to match the electorate demographically, and he correctly predicted a Roosevelt victory. The defeated *Literary Digest* closed down not long afterwards.

The following year, Gallup polling arrived in the UK. Harry Field, a Briton who had worked with Gallup at Young & Rubicam, was despatched to the London School of Economics (LSE) to find a suitable leader for

As Henry Durant's experience grew, he refined aspects of data collection by survey. But he saw problems, too

a British Institute of Public Opinion (BIPO) to mirror its American cousin. Field convinced research student Henry Durant to take on the job.

The role must have appealed both to Durant's political leanings and his lack of private means. The son of a warehouseman, Durant had won a scholarship to Christ's Hospital School and then worked as an insurance clerk before studying sociology at LSE. The £150 per year salary from the BIPO would help support him and his wife while they pursued their academic research careers, and the prospect of giving the public a voice on social and political issues chimed with his left-wing views.[2]

With Durant in post, Field returned to the United States to establish the People's Research Corporation, and then initiate the American Association for Public Opinion Research (AAPOR) and the World Association for Public Opinion Research. Tragically, he was killed in an air crash in Paris before either was established.

The population of inter-war Britain was studied by a number of government bodies, not only as citizens but also as consumers. The Empire Marketing Board, set up to promote the consumption of goods produced within the British Empire, segmented its audience according to social class and sex, placing adverts in the relevant papers, and distributing targeted posters and pamphlets to schools and Women's Institutes. They enlisted advisors from the advertising industry, notably William Crawford, whose 1938 report *The People's Food* ruffled government feathers with its finding that millions of British citizens could not afford to eat properly.

During the Second World War, the distinction between commercial and political polling became almost meaningless. Governments on both sides of the Atlantic took control of information, mindful both of the need to know the level of public support for wartime policies and of the potential power of information delivered to the right audience at the right time.

In the UK, the government brought several research groups together as the Social Survey Unit, run by Louis Moss who had been managing the BIPO. The Unit directly employed researchers and social scientists to supplement official data, as well as farming out survey research to commercial agencies such as JWT's London branch and Britain's largest advertising agency, the London Press Exchange.

Random sampling methods, stratified for occupation, age, sex, and so on, produced tabulated data on vital issues including bicycle use, attitudes to fuel rationing, demand for brooms, and cake consumption in private homes.[3] From 1940, long-running surveys emerged from the Social Survey Unit's work. The National Food Survey, for example, ran for 60 years, until 2000, when it was merged into the Expenditure and Food Survey (bit.ly/3aI3p21).

What people say and do

After the war, both flavours of public opinion – the commercial and the political – continued to be valuable data. The Market Research Society was formed in 1946, with members from public and private organisations. Initially a couple of dozen people meeting over lunch in Soho, within 10 years it had hundreds of members and held its first conference in Brighton in 1957. Henry Durant was its first president.

Durant established his reputation, as Gallup had done, by correctly predicting the outcome of an election. His polling anticipated the Labour victory under Clement Attlee of 1945. Durant's background in both social science and actuarial work proved a good foundation for innovation in polling techniques. He adopted Gallup's "quintamensional design" – five questions designed to find out a respondent's knowledge about an issue, their level of interest in it, their attitude to it, reasons for the attitude, and strength of opinion. As Durant's experience grew, he also refined other aspects of data collection by survey. But he saw problems, too.

In a frank article for *The Incorporated Statistician* in 1954, Durant discusses practical issues in data gathering.[4] How, for example, would you discover drinking habits and consumption though a field survey? You could visit people at home, but unless you weight the responses, people who do not go out much will be over-represented. If you do weight the responses, the "only home one night a week" group will be represented by the smallest sample, giving the least reliable

Timandra Harkness is a presenter, writer and comedian. Her BBC Radio 4 documentaries include *Five Knots* and *Steelmanning*, and she is the author of the book *Big Data: Does Size Matter?*

results. Home interviews could also elicit less accurate answers. "Husbands may not want their wives to know the truth about the amount they drink," says Durant. But if interviewers were instead to be stationed outside pubs, might "there not also be the danger that interviewers will tend to avoid the rough-and-ready types, who in fact do consume more than their due proportion?" asks Durant, reasonably.

Some of these problems were mitigated, to an extent, by the advance of technology. As more households acquired their own telephones, telephone interviews began to take over from face-to-face surveys. Because each telephone number was linked to a specific household, randomised sampling became more practical and, because the interviewer did not need to travel, cheaper to execute. Later still, the internet and smartphones provided easier, lower-cost ways to contact potential interviewees.

By that time, however, survey data directly collected by asking questions had competition. Data generated as a by-product of our everyday activities could potentially reveal far more about us than we would willingly reveal in words. All that was needed were the right techniques to analyse this data. ■

Note

Part II of the *History of the Data Economy* will be published in our August 2021 issue.

Disclosure statement

The author declares no competing interests.

References

1. Robinson, D. J. (2012) Mail-order doctors and market research, 1890–1930. In H. Berghoff, P. Scranton, and U. Spiekermann (eds), *The Rise of Marketing and Market Research* (pp. 73–93). New York: Palgrave Macmillan.

2. Roodhouse, M. (2013) "Fish-and-chip intelligence": Henry Durant and the British Institute of Public Opinion, 1936–63. *Twentieth Century British History*, **24**(2), 224–248.

3. Schwarzkopf, S. (2012) Markets, consumers, and the state: The uses of market research in government and the public sector in Britain, 1925–1955. In H. Berghoff, P. Scranton, and U. Spiekermann (eds), *The Rise of Marketing and Market Research* (pp. 171–192). New York: Palgrave Macmillan.

4. Durant, H. (1954) The Gallup poll and some of its problems. *The Incorporated Statistician*, **5**(2), 101–112.

5. Stigler, S. M. (2003) *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.

6. Lusinchi, D. (2019) Two chapters in the development of human population sampling (1895/1934). In *JSM Proceedings, History of Statistics Interest Group*. Alexandria, VA: American Statistical Association.

7. Lusinchi, D. (2017) The rhetorical use of random sampling: Crafting and communicating the public image of polls as a science (1935–1948). *Journal of the History of the Behavioral Sciences*, **53**(2), 113–132.

8. Crossley, A. M. (1957) Early days of public opinion research. *Public Opinion Quarterly*, 21(1),159–164.

Sampling: Statistical divisions

The prehistory of the sample survey goes back to the eighteenth century, when the mathematician Pierre-Simon Laplace proposed a representative census of France, surveying selected regions instead of the entire country. Adolphe Quetelet's initial enthusiasm for "Laplace's method" waned when he saw the difficulty of choosing a truly representative sample of very heterogeneous regions, and that problem would not be solved until large quantities of tabulated census data became available in the late nineteenth century.[5]

Anders Kiær, director of Norway's Central Bureau of Statistics, put the representative method back on the international agenda in the 1890s, but in practice it failed him, underestimating the size of the disabled population in Norway.[6] By the time Danish statistician Adolph Jensen reported in 1926 that the representative method was recognised as statistically superior,[7] it was already in wide practical use for market research.

Advertising companies were using "reasonably reliable" stratified or quota sampling methods for research before 1920. "In the early days of sampling there was … a tendency to think of cases taken 'at random' as being typical or representative of a universe," recounts polling pioneer Archibald M. Crossley.[8] "When the transition was made from 'at random' to true 'randomization,' the lily was gilded with the phrase 'scientific sampling.' This gilding, I would say, was done at the time of the introduction of the national polls on political issues in the mid-thirties." But even before this improvement, Crossley says, samples were selected to reflect the America portrayed by the Census, and "many probability principles had been used for a long period – e.g. rotation and randomization of blocks, road segments, homes and individuals, and the assignment of specific locations to interviewers."

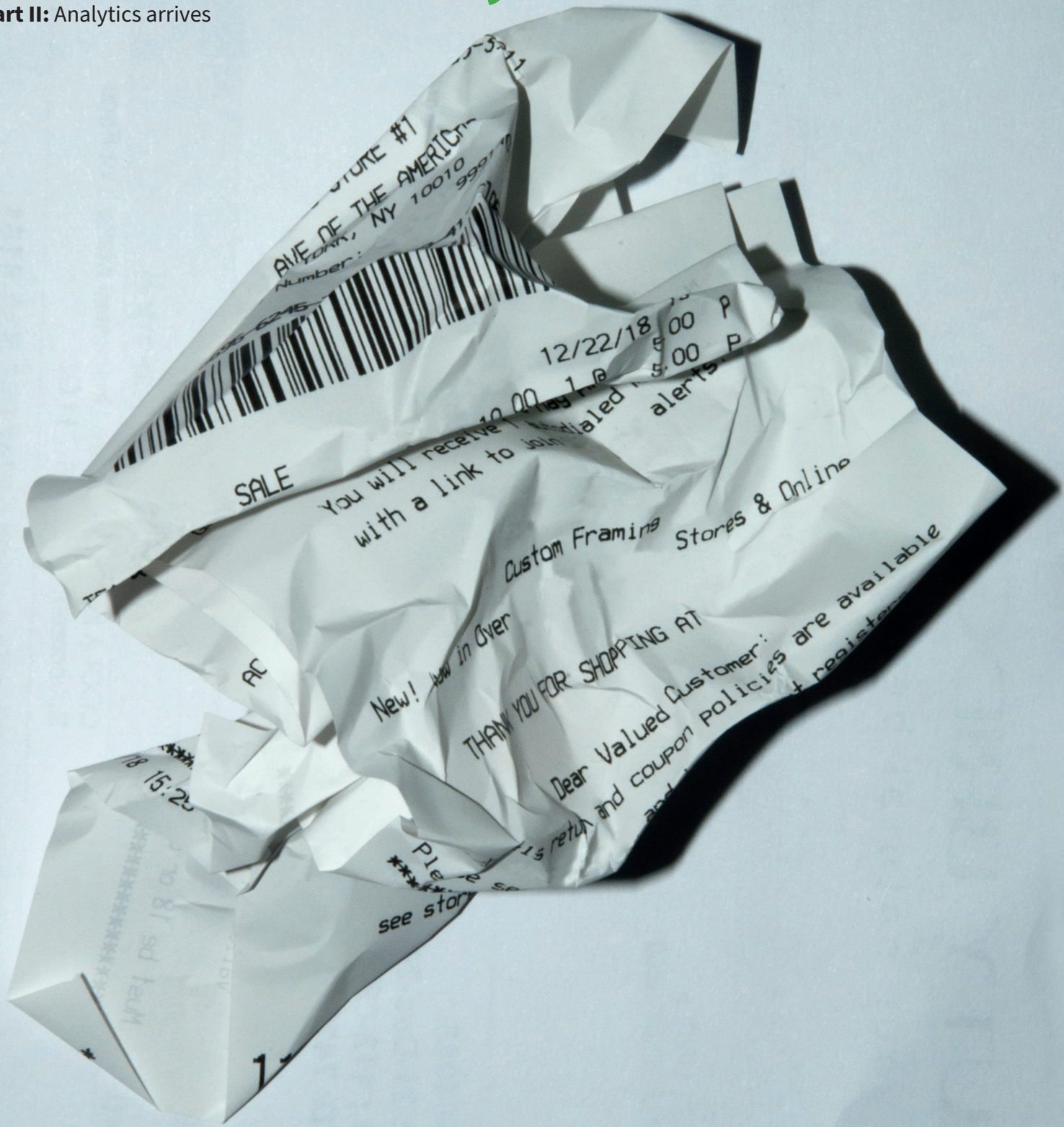
Market researchers were using two sampling methods identified by Jensen: "purposive selection" of representative districts or groups, and random selection using probability theory. For market research, quota sampling, in which respondents are recruited by category – typically sex, age and social class or income – was cheaper and easier to carry out, though it was recognised that properly randomised sampling gave more accurate results. Jerzy Neyman presented a paper to the Royal Statistical Society in 1934, arguing that probability sampling was superior, but in practice, many pollsters treated the two methods as interchangeable or equivalent until the 1940s.[7]

At a 1946 AAPOR meeting, social psychologist Norman Meier defended quota sampling against Morris Hansen, technical advisor to the US Census Bureau, who argued for a better information yield from probability sampling. Because it enabled estimation of sample error, probability sampling was favoured by government surveys, which also had reliable access to population-scale lists and the resources needed to contact a randomly chosen list of targets.

Nevertheless, the quota system continues to be used in market research when interviewers are sent to find human respondents for questionnaires.

The history of the data economy

Part II: Analytics arrives



Timandra Harkness is a presenter, writer and comedian. Her BBC Radio 4 documentaries include *Five Knots* and *Steelmanning*, and she is the author of the book *Big Data: Does Size Matter?*

The world has changed dramatically over the past 200 years. Data is now the fuel that drives business – identifying potential markets, shaping new products and targeting consumers. To understand where we may be heading next, *Significance* has partnered with *Impact*, the magazine of the Market Research Society, to jointly publish a series exploring the past, present and future of the data economy. This second part tells the story of the arrival of analytics and efforts to better understand consumer behaviour using new data sources. By **Timandra Harkness**

"Seeing is the new asking." That's what Stephan Gans, chief insights and analytics officer at PepsiCo, constantly tells his colleagues.

"If you ask a mother, or you ask a dad, 'what do you give your kid in his lunchbox that he takes to school?', you can predict what the answer is going to sound like," says Gans, "because the dad wants to be seen as a responsible dad. So, there's some fruit, there's some this, there's some that…"

But by recording what actually happens, using 24/7 cameras in people's kitchens, researchers found that one conscientious father packed his son's lunchbox every day with carrots, fruit, and so on, and every day the boy came home with his lunchbox completely intact. The child felt eating all those carrots took up too much valuable play time and so bought something quick to eat from the school canteen instead.

"You would have never learned that from the dad," says Gans. And this, in a nutshell, is the problem with asking questions. People might tell you what they believe to be true, or what they want to be true. The actual truth, though, may elude even the most diligent interviewer.

Henry Durant, the first president of the UK Market Research Society, knew of these problems with survey-based data collection, and warned about them, in the 1950s (as we learned in the previous article in this series).[1] So, it is no surprise to observe that, as technology has developed and the data economy has evolved, "seeing" has become "the new asking".

Watch this

Jon Ward, a regional vice-president of sales at eye-tracking company Tobii, offers another example of direct observation being more informative than asking questions.

"A lot of people will say that they're not price sensitive, and then completely demonstrate price-sensitive behaviour when they are not being asked, because it's like saying, 'Are you cheap?' 'Of course I'm not cheap!'" Wearing Tobii's cameras, so compact that they look like Michael Caine spectacles, shoppers can go into physical stores and shop normally. With every detail recorded for analysis, "you see them go in," says Ward, "and yes, they 100% are cheap."

In the age of store cards and barcodes, collecting data on what people buy is easy, but eye-tracking analysis reveals the purchase decision in action: the moment a shopper's eyes scan competitor brands, the back-and-forth comparison between two similar products, and then that final glance at the price label that clinches the deal. Analysing what people look at online is easier still, when volunteers put eye-tracking cameras above their screens and give researchers access to what's on their monitors.

Mike Follett, managing director at eye-tracking company Lumen, thinks all implicit observation techniques share three key advantages over explicitly asking questions.

"The first one is: people don't know what they don't do. It's very, very hard to know that you definitely didn't see something. Given the fact that attention is selective, we have been ignoring the vast majority of human experience to focus on explicitly remembered actions and opinions."

The second advantage is that people are very bad at accurately remembering things that they did do or see, even in situations as important as giving evidence in court.

The third advantage, says Follett, "is that when it comes to marketing, ads might be developed in isolation, but they're always seen in context. So, making sure that you serve up the experience to be as close to reality as possible, and then understand how attention works within that reality, is very important."

Although twenty-first-century technology allows observations to be filmed by cameras smaller than your thumb, and analysed by machine learning algorithms, implicit observation has its roots in twentieth-century methods and before. The first eye-tracking was done by a human looking through a flap in a hoarding and hand-sketching where passers-by were looking.

From the 1950s, behaviourist approaches became more popular as a way to understand human beings. Polymath Herbert Simon brought together his interests in psychology, economics and computing to develop the idea that simply observing what people do is a better guide to what they will do in future than asking them explicitly.[2] We may be rational, but our rationality is bounded by the finite amount of time and attention we have to spend.

Today's data-driven consumer insight combines the kind of individual observation described above with aggregated data, to see overall trends and behaviours.

Media monitoring, the observation of what consumers look at, listen to or read, is as old as media. Early newspaper adverts offered discount coupons to find out how many readers of different newspapers paid enough attention to the advert to cut them out and use them. Companies like Valassis still use coupons today, though online discounts are overtaking paper tokens.

Nielsen, founded in 1923, pioneered automatic monitoring of a household's radio and television habits, supplemented by diaries to track individuals within a household. They also pioneered indexes of retail sales data in the 1930s, combined with panels of consumers who recorded, and later scanned, their actual purchases, enabling advertisers to compare adverts seen with goods bought.

Early point-of-sale systems enabled larger stores and chains to track what was being bought directly. When American chain J. C. Penney installed cash registers linked to mainframe computers at its distribution centre in 1969, as well as speeding up customer transactions, the *Arizona Republic* reported: "the Glendale store manager knew, day by day, just what merchandise was being sold and how much" (bit.ly/3x4mELN). And with the adoption of barcodes from the 1970s onwards, chains like Walmart could track goods in enough detail to predict demand.

In the mid-1990s, UK supermarket chain Tesco became one of the first to offer

Michael Walter/Unsplash.com

The data fusion experiment

The problem the 1989 data fusion experiment was trying to solve was the clash between the desire of clients for a multidimensional picture of a population, and the desire of survey respondents not to spend more than an hour filling in a lengthy questionnaire. If you had two different questionnaires, completed by a similar population, the results of both should, in theory, contain all the information you need – if you combine them in the right way.

Baker, Harris and O'Brien wanted to test a way to fuse two data sets that drew on statistical methods already in use to deal with missing data, which they call "data imputation" or "ascription". They give the example of a weighting system called "hot decking" that was used to infer missing data in the 1981 UK Census.

Respondents were sorted into a relevant order, and "if the sort has been well planned, the person most like the respondent with missing data is the person who came just before him." Missing data are then replaced with the values from the previous respondent.

The core idea of data fusion is that "donor" respondents from one data set can contribute missing variables to "recipient" respondents from another. The resulting data set has the same number of respondents as the "recipient" data set, but more dimensions. Just as in medical donations, the more similar each donor is to their recipient, the better the results.

The first issue, then, is matching donors to recipients. Obviously, donors and recipients need to share some common variables, such as basic demographic categories, but since part of the point is to preserve inter-relationships between variables, not all the shared data will be equally useful. Among the common variables in the experiment were age, sex and class, but also measures like "trying to slim", "foreign car in household" and "have front loading washing machine".

The researchers varied the number of common variables, and of donors, to find out what gave the best results. They borrowed a matching algorithm used by IMS France, based on a chi-squared metric distance between respondents. The French system talks in terms of marriages, matching the "love at first sight" pairs first, in which A is closest to B, and B is closest to A. As the matching algorithm continues its work, it moves on through "childhood sweethearts" to "adultery", since one donor can give their data to more than one recipient.

For the data fusion experiment, the two data sets – donor and recipient – were in fact two halves of the same data set. This meant that, unlike in the real world, the researchers were able to compare the results of the fusion with the true data set.

On the whole, they were impressed with how well the fused data set matched the original, especially when the donor data set was large. "No proponent of fusion would claim that the process can really predict the answers to the missing questions for each individual in the recipient survey," they say. "Rather, fusion should produce acceptable results when the missing data transferred to the recipient file are analysed."

One caveat, however, is that strong associations between variables, like the one between betting and drinking alcohol, were damped down by the fusion process. "Heavy cinema goers are three times as likely to play squash and this is perfectly preserved in the fused data. However, joggers are nearly five times as likely to play squash but the fused data show a substantial regression to the mean."

The other caveat is that this method simulates whole populations well but should not be taken as a genuine result for any individual respondent. "Fusion works by reproducing aggregate data based on groups of respondents. … The important point we wish to make is that the closer one gets to individual respondent data (i.e. small bases) the worse will be the accuracy of the fusion."

Today, this kind of manipulation of data sets would be much easier to do. However, the ability of data aggregators to identify unique individuals and link their own data into one very large data set means it is not always necessary.

shoppers a direct trade of discounts for data: sign up for a "Clubcard" and get targeted offers in return for letting the retailer observe your shopping habits. In Clubcard's first year, 5 million people took the deal. Edwina Dunn and Clive Humby, who helped devise and run the scheme, eventually sold their data company Dunnhumby to Tesco for over £90 million.

Social segmentation

The ultimate aim in monitoring what people see, do, buy and consume is to try to get them to see, do, buy and consume more of the things they like, or that a company thinks they may like; to market to them, in other words. Someone who watches a lot of Netflix may be in the market for a high-end TV, for example. But the fact that a person might like a high-end TV does not necessarily mean they can afford one.

Since the early nineteenth century, merchants have exchanged information on their customers, to predict which ones might be a bad risk for settling their bills. In the 1970s, this stepped up a gear when credit reports and marketing databases began to merge. The Manchester Guardian Society, founded in 1826 "for the Protection of Tradesmen against Swindlers, Sharpers and other Fraudulent Persons", merged with mail order giant GUS in 1996 and eventually became part of international credit bureau and data broker Experian (bit.ly/3vMPY97). A similar process saw American credit bureau Equifax expand into the field of marketing by acquiring both data companies and demographic modelling software. Now it could segment people not only into creditworthy and uncreditworthy, but into marketing categories like "upper crust" and "living off the land'".[3]

Geodemographics, the ability to sort people into demographic and consumer categories linked to where they live, also emerged in the 1970s thanks to newly available computer-readable census information.

The Claritas Corporation, established in 1971, sorted Americans into 40 types, including "money and brains" and "hard scrabble". Its founder once claimed that all he needed was a ZIP code "to predict what you eat, drink, drive – even think".[3] Meanwhile, in the UK, social scientist Richard Webber

used the 1971 Census to study inner-city deprivation in Liverpool. His Classification of Residential Neighbourhoods system formed the basis for the ACORN system later used by data company CACI.[4]

This social segmentation used experimental work being done within companies and independent agencies. And what's important to note is that this work did not solely rely on observing behaviour, or on records of products bought and media consumed. Consumer attitudes were a crucial part of the mix.

From diesel to jet fuel

Phil Barnard, who would later run the global market research group Kantar, describes the UK market research sector of the 1970s as "almost a cottage industry". Research projects would alternate qualitative and quantitative methods to define and explore questions, designing experiments almost like medical trials with Latin squares and control groups.

Even in the early days, when analysis was done with log tables and slide rules, and questionnaires were hand-tabulated on paper, sophisticated statistical models underpinned the work of Barnard and his contemporaries. The Fishbein model, for example, captures consumers' attitudes to a brand or product in mathematical form. If those attitudes change in certain ways, the model helps predict the change in sales or market share.

Cluster analysis could segment people by interests and habits as well as basic demographic categories. "You'd have a battery of questions," says Barnard, "which you knew from your basic research clung together. You'd ask somebody how much they agreed with a particular statement: 'I like going on holiday.' 'I like meeting people'. And from that you would produce a small subset of those questions, maybe three of them. You could use those responses to classify the person on that particular criterion and do that over a number of different dimensions. That would enable you to classify people into different segments."

All these methods and approaches were being used together, often by the same people. But what transformed data from diesel to jet fuel was the ability to combine diverse data sources and create one multidimensional picture.

In 1989 Ken Baker, Paul Harris and John O'Brien presented the results of an experiment in data fusion to the Market Research Society Conference. They concluded that different data sets could be combined to give results comparable to putting a larger questionnaire to one population (see "The data fusion experiment", page 18).

When they wrote up their work several years later,[5] in 1997, they commented that "we have a new buzzword – integrated targeting – the merging/linking/matching of market research databases. Is this the way the industry is moving as the millennium approaches? A lot of researchers would conclude that this development seems to be inevitable."

Writing in the same year, Bill Blyth (chief statistician at Taylor Nelson Sofres) and Tim Bowles said that the problem for market researchers had changed from a lack of data to a proliferation of data from different sources: electronic point of sale data, consumer panel data, pooled retailer records and so on.[6]

Given this proliferation, they wrote, "it is inevitable that market research practitioners will move away from their traditional stance as collectors and purveyors of research data. Since they will have to obtain appropriate market intelligence to establish adjustment factors, they will inevitably become involved in the organisation and analysis of diverse data sources."

The "inevitable" has certainly come to pass. The ESOMAR 2020 industry report[7] describes the impact of data analytics on market research thus: "As new ways to gather data have emerged which do not require a 'real-time', one-to-one personal interaction between researcher and respondent, methodologies have increasingly moved from being an 'active' process or collection to a passive, less intrusive, less conscious (and in some people's view, a more accurate) recording of behaviour and generation of information, for the researcher to use to generate insights.

"The most recent iterations have resulted from a wider application of technology in the industry, which has given rise to a set of methodologies that would have been impossible to apply (or conceive of) otherwise."

Hidden details

ESOMAR estimates the value of analytics to the data, research and insights industry today at around $47 billion, slightly over half the sector's entire worth.

We asked PepsiCo's Gans to put a figure on the value data analytics adds to his work for the soft drinks giant. "I think we spend $2 billion a year just on advertising," he says. "Say that you're 10% more effective in targeting the right consumers and convincing people to buy your brand: you're talking about saving millions and millions."

Lumen's Follett sees the switch to data-led marketing as a revolution comparable to Robert Hooke's *Micrographia*, a seventeenth-century bestseller that introduced readers to images seen through early microscopes. Hooke's sketches of plants and insects revealed hidden details of everyday life; digital ethnography, eye-tracking, media monitoring and loyalty card data have done the same for our understanding of the lives of consumers.

But, of course, there is so much more still to learn – and in the early 2000s, a group of Harvard University students would launch a website to prove it. ■

Note

Part III of "The History of the Data Economy" will be published in our October 2021 issue. The author thanks Adam Phillips and the Archive of Market and Social Research for assistance with researching this article.

Disclosure statement

The author declares no competing interests.

References

1. Harkness, T. (2021) The history of the data economy. *Significance*, **18**(2), 12–15.

2. Simon, H. A. (1992) What is an "explanation" of behavior? *Psychological Science*, **3**(3), 158–159.

3. Laurer, J. (2017) *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. New York: Columbia University Press.

4. Archive of Market and Social Research (n.d.) *Post-War Developments in Market Research*. Wallingford: AMSR. bit.ly/2UeFXDK

5. O'Brien, J., Harris, P. and Baker, K. (1997) Data fusion: An appraisal and experimental evaluation. *Market Research Society. Journal*, **39**(1), 1–52.

6. Bowles, T. and Blyth, B. (1997) How do you like your data: Raw, al dente or stewed? *Market Research Society. Journal*, **39**(1), 163–174.

7. ESOMAR (2020) *Global Market Research 2020: An ESOMAR Industry Repor*t. Amsterdam: ESOMAR.

The history of the data economy

Part III: The new kings and queens of data

Data is now the fuel that drives business – identifying potential markets, shaping new products and targeting consumers. To understand where we may be heading next, *Significance* has partnered with *Impact*, the magazine of the Market Research Society, to jointly publish a series exploring the past, present and future of the data economy. This third part tells the story of the evolution of social media, which created rich and detailed data sources and positioned tech giants as data economies in their own right. By **Timandra Harkness**

Until February 2004, a "face book" was a paper directory that US students received to help them get to know each other, with names, photographs and a few biographical details. That was until Harvard University student Mark Zuckerberg had the idea of creating an online version.

Well over a thousand Harvard students signed up within 24 hours of TheFacebook's launch. Today, Facebook (which dropped the "The" in 2005) claims to have 2.85 billion active users worldwide – over half the world's internet users, and getting on for a third of the planet's human population. Total revenue in 2020 was more than $85 billion.

Timandra Harkness is a presenter, writer and comedian. Her BBC Radio 4 documentaries include *Five Knots* and *Steelmanning*, and she is the author of the book *Big Data: Does Size Matter?*

Any platform that attracts an audience of more than a billion people a day could expect to make money from advertising, but what makes Facebook's advertising space so valuable is the ability to target the right pairs of eyes, and what makes that possible is data.

The ability to gather data from a person's online behaviour, to build a profile of them and target them with online adverts is older than Facebook or its social media predecessors MySpace, Friendster and SixDegrees. In 1996, writer Melanie Warner described the now familiar feeling in an article for *Fortune* magazine (bit.ly/3BToW3y): "You sign on to your favorite website and voila! – up pops an ad for Happy Times Cruise Lines… Sure enough, you work in Connecticut, and you've been thinking about vacationing in the Mediterranean. But how do they know that? Whoever they are."

As Warner goes on to explain, "they" are probably DoubleClick, an advertising broker launched in March of that year. By July, it had profiles for 4 million people, and 25 major websites on its books. "The next time you log on to a DoubleClick site, its software notes your E-mail address, checks out your user profile, and uploads an ad customised for you – within milliseconds of your signing on", she writes.

Underlying some modern iterations of this advertising system is a process known as real-time bidding, which is used today by companies including Google, which bought DoubleClick in 2008. When you log on to a web page linked to an ad network, the network's software will parse available information about you from your log-ins, cookies on your computer, etc., and create a "bid request" at an ad exchange.

"On that exchange, different advertisers will bid for the right to fill that space on your website," says marketing analytics consultant Andrew Willshire. Think of this as a bit like a virtual art auction, where prospective buyers have already told the auction house broadly what sort of painting they are looking to buy and how much they are willing to pay for it. "The ad exchange will weigh up all the potential bidders, and whoever has bid the highest will get the right to serve that impression to the viewer," explains Willshire. "This process happens in thousandths of a second, which is why the adverts look like they are there the whole time to the user."

Much the same thing happens when you log on to Facebook: you will see adverts served to you that companies have bid for, based on your profile. The difference is that social media sites can know their audience like no advertising platform ever before. Not only do they have a comprehensive record of everything you do on their site – the people in your social network, the posts you have "liked" or shared, the locations and activities you have mentioned, and adverts you have previously clicked on – they may also track your activities on other websites, be notified when you open other apps, and add information about you from other companies and data brokers.

Changing the world?

The advent of social media in the early twenty-first century was rapid. By 2011, a UK government report estimated that three in five internet users also used social media, up from under one in five in 2007 (bit.ly/2V1p6Fh). Market researchers, social researchers and others quickly took notice. For example, the government report was commissioned "to explore the ways in which data generated by social media platforms can be used to support social research and analysis at the Department for Work and Pensions". It argued that "when compared to traditional surveys, social media data offer considerable advantages in terms of how quickly results are delivered, the scale at which results can be brought in, and (potentially) how cheaply they can be obtained".

Jake Steadman was one of the many researchers excited about using secondary data from social media and similar sources. "I probably went in just naively assuming it was going to change the world," he says. After years of working for marketing and research agencies, Steadman went to O2 as its first "head of real-time research".

"No one really knew what it was," he says of his new role. "But it meant they [O2] recognised this new and emerging insight source, which was social data, but they didn't really know how to access it or how to use it. And nor did I, I'd never done it before. But we both decided to take a bit of a leap of faith."

Steadman (who has since held roles at Twitter and Deliveroo) is frank about his early enthusiasm for social media data. "I think I arrogantly assumed it was going to replace everything," he says. Nor was he alone in thinking that. Consultant Ray Poynter says: "There was probably a lot of optimism around how much we could do with social media data. One of the mantras that people talked about was, 'Why talk to some of us when you can listen to all of us?' There is a lot of sense in that."

As Poynter explains, social media presented organisations with the means to listen to "real customers talking about their real experiences, on topics we had not thought to ask about or we had not prioritised".

"It's fantastic at answering questions you didn't ask," he says.

Indeed, there are some questions you would never, or could never, put in a survey. For example, it is easy enough to find out what sort of products people like just by asking them. But, if you are looking to create a brand new product, the ultimate goal should be to figure out what people might like in future.

"Take, for example, matcha tea," says Steven King, chief executive of data analytics firm Black Swan. Matcha is the green powder used in the Japanese tea ceremony, prepared by growing green tea in very particular conditions and then grinding it finely. "It really has been around for ages," says King. But then it began to be drunk in different forms and in different places.

King's company specialises in listening, if not to everyone, at least to everyone on social media. Black Swan's data-gathering is designed to pick up on the weak signals associated with big data techniques, like a few visitors to San Francisco trying a new drink in a hip café and then raving about it online.

"People start making [matcha tea] and selling it in a cool café," says King, "then the little cool brands start running quite small production lines – higher costs, but a bit more agile. So, you'll see it in the cool organic shops. From there, it begins to be a slightly larger trend. And then your bigger companies pick it up – and the joy of data is that you saw that whole thing."

Social media data is good for identifying long-term trends, and innovations that straddle contexts. However, says Poynter, there are still questions that are better answered by researchers doing the asking.

Daria Nepriakhina/Unsplash.com

Hypotheticals, adverts or products that do not exist yet, cannot be tested by passive observation of electronic word of mouth.

And, of course, no matter how many social media users rave about a product like matcha tea, it is unlikely that these views are representative of the broader population. Only one in five of the UK population uses Twitter, for example, and it is not a balanced cross-section. Poynter uses the example of care homes: neither residents nor staff are very likely to be using Twitter, so their views will not be accurately reflected on that platform. "So, it's not that you can't use social media," says Poynter. "You've got to think about how and when."

As Steadman ultimately discovered, after the initial giddy wave of social media enthusiasm had passed: "O2, like every brand, did segmentations, and brand measurement, and customer experience measurement and all those kinds of metrics. Social data sits alongside those. It doesn't replace them, it augments and adds cultural context. But you still need to have statistically robust measures in place."

Even Black Swan, whose focus is social media data, combines this source of information with other forms of secondary data, and surveys, to get a multi-dimensional picture. But for King, social media data has advantages over surveys, because you do not influence responses by the way you ask the questions. "By going bottom-up rather than top-down, you've got more granular data that allows you to build models, and then build algorithms and prediction," he says.

This mix of big-picture and granular detail also appeals to statistician Simon Raper, founder of Coppelia Machine Learning and Analytics, but perhaps for different reasons. He uses the example of viewer recommendations for a streaming service. "If I want to understand what the average customer is doing, or something broadly about viewing patterns, I can just take a sample and I don't have to go crazy on the big data stuff." But, he says, "the huge amount of data is going to make it possible for me to answer questions about some niche viewing, like Japanese horror."

Raper likens it to a pixelated image. "If you've got a picture of a crowd, a low-resolution picture doesn't really matter if you just want to make out the shape of the crowd, or groups within it. But if you want to zoom in on someone's face, then you need a really high-definition image." It is almost "paradoxical", he says, that one of the things we can do with enormous amounts

Dimension reduction

However much analysis and research might rely on huge quantities of data and machine learning programs, basic statistical techniques still underpin the work.

"One of the methods that really shines is any form of dimension reduction," says Simon Raper of Coppelia. "The oldest form of this, which is actually Victorian, is principal component analysis (PCA). You can take 100 variables and turn them into three variables that capture as much of the information as possible."

Say, for example, that you want to figure out someone's taste in movies so that you can recommend other films they might like. A streaming platform like Netflix would have information on the films each user has watched through the service, along with a rating of whether each user liked each film. "That would be an enormous data set," says Raper, "with [the] number of rows [equal to] the size of that population, [and where] the number of columns is the complete library of all the films."

While an algorithm can process all that data, the human mind cannot. The goal is to find which variables correlate closely enough to each other to combine them into a single variable, and which need to be kept distinct, to turn all those data points into a model that has predictive power.

It is likely, for example, that liking one Quentin Tarantino film correlates with liking the director's other films, but less likely that a taste for comedy films can predict whether somebody enjoys classic movies (bit.ly/3fthGSb). So, it would make sense to combine all Tarantino films, and others that correlate closely, into a single axis of measurement, but to put comedy and classic movies on different axes.

Using PCA, it is then possible to reduce the mass of data to a few linear measures that explain most of the variation seen in the original data. Plotting each film's position in this mathematical space reveals clusters of similarity, in terms of audience taste.

"Recommendation systems are very much built on this idea," says Raper. "You can think of a movie being in a geometrical space that's massively multi-dimensional. You can't really visualize it, but the recommendation is the other film that's closest to [the film that a user has watched] in that space."

PCA can also be used to group individuals with those similar to them. Cluster analysis takes an agnostic approach to what the groups might be, simply letting mathematical approaches find the most orthogonal axes for that data set, to maximise similarities within groups or clusters, and differences between groups. The emergent patterns can be related back to the data to describe groups by empirical observation rather than theory.

Researchers at Erasmus University in Rotterdam used cluster analysis to find four distinct groups of Facebook users who "liked" a particular football club (bit.ly/3A2anZy). Relating their findings to potential marketing use, they comment: "Clustering of a firm's Facebook fans may improve understanding of strategic segmentation of social media users connected to a firm. Moreover, the cluster results and visualizations can be used to improve targeting of marketing efforts."

The same technique can be used to turn a large and complex data set into a simpler model to make predictions based on fewer data points. Researchers at the UK Office for National Statistics looked for indicators of social capital, a measure of social connections and pro-social attitudes, in the UK Longitudinal Household Survey. They used PCA to turn those indicators into just six questions that could be used to predict an individual's social capital (bit.ly/3ft0c8G). These included strength of feeling about the duty of a citizen to vote, and (perhaps ironically) whether they belong to a social networking website.

of data and processing power is to focus in on the smaller details (see box, "Dimension reduction").

The privacy question

The capacity to zoom in on the individual brings us back to one of the most powerful, and disquieting, aspects of social media – especially when it is being used by the same platform to collect data and to target advertising. Though the predictive strength of data such as Facebook "likes" has been overstated, we undoubtedly share more unsolicited information about ourselves, in digital form, than any humans before us.

Add to that the capacity of analytical algorithms to combine data sets, to find statistical relationships, and to infer interests and demographic details from other data, and market segmentation turns into *micro*segmentation. Facebook can target adverts to niche audiences of a hundred, a dozen, or just a handful of people who fit specific criteria. A female cyclist living in West London who uses a mobile phone, to borrow Facebook's own example (bit.ly/2VcgdJg).

Social media microtargeting techniques were once spoken about enthusiastically and positively (bit.ly/3A2JRPM). Writing about their use in Barack Obama's successful presidential election campaigns, *The Guardian* reported (bit.ly/3jeyDB6) how Obama's 2012 re-election team "harnessed Facebook and other social media to spread the message" and "used cookies to service targeted digital adverts to voters' computers, honing the message according to the individual's age, gender, occupation, interests and voting history".

Yet digital targeting was viewed less favourably after the successful Brexit campaign in the UK and Donald Trump's election as US president in 2016.

One company in particular became the focus for a new awareness of how political campaigning uses social media data: Cambridge Analytica (CA), which claimed to be able to combine data profiling with psychological profiling techniques to help clients win elections. A scandal erupted in March 2018 when it emerged that some of the data available to CA came from users of a third-party Facebook app, and there were suggestions that the data may have been used for political campaigning associated with Brexit. CA ceased operations several weeks after the scandal broke. In October 2018, Facebook was fined £500,000 by the UK Information Commissioner's Office (ICO) for failing to keep its users' data secure (bit.ly/3fsn5cd). Two years later, the ICO concluded that CA was not involved in the Brexit referendum (bit.ly/2WGb1xq).


Chad Madden/Unsplash.com

Still, the damage was done. CA and other similar stories have made many people more wary about what happens to their data. Although the majority of internet users continue to use social media and other apps that track and profile them, half of US adults report having avoided a product or service because of privacy concerns (pewrsr.ch/3jkvAHh). In the USA and UK, tech companies have been called to account by elected politicians. Legislation to regulate data use includes the EU's General Data Protection Regulation (GDPR) and California's Consumer Privacy Act.

One response to increased customer concern and regulation has been for social media companies to monopolise the data they collect. "Facebook in particular has become very, very reluctant to sell information," says Poynter. And to the surprise of some, the tech giants have also been willing to change their data-gathering habits, even to go beyond what the law demands in some cases. Google, for instance, has proposed to block third-party cookies – the type used to track internet activity and target ads – from its Chrome internet browser from 2023 (a number of rival browsers already do so).

"I think everybody was surprised, including the people working at Google, that Google basically did more than it needed to do for GDPR," says researcher Simeon Duckworth. "Everybody imagines, if you take away cookies, the internet will fall apart. Now Google has actually gone and said that it will do that."

Having worked in advertising for many years, Duckworth is sceptical that data-based profiling and targeting is always effective, especially for large advertisers who want to reach a mass audience: the extra expense of tools like microtargeting and A/B testing to refine a marketing message may not be justified by extra effectiveness. Small and niche advertisers are most likely to see results from narrowing down their target audience, but they are also less likely to be able to afford to do it.

For this reason, Duckworth does not think that tougher regulation will destroy the business model of digital advertising. "It's not going to kill the internet, but it is going to redistribute it towards bigger platforms, bigger publishers away from the smaller ones," he says. Large companies like Google and Facebook will have the scale to work within new laws using new techniques. Smaller companies will lose access to data and the benefits of targeting a niche audience.

So, what is the future of the data economy? A fairer, more transparent, less manipulative internet environment? Or simply one in which a handful of big tech companies run the show, and smaller players go under or sell out? The final article in this series will seek to answer these and other questions. ■

Note

Part IV of "The History of the Data Economy" will be published in the December 2021 issue of *Significance*.

Disclosure statement

The author declares no competing interests.

FEATURES

The history of the data economy

Part IV: The future

fabio/Unsplash.com

Data is now the fuel that drives business – identifying potential markets, shaping new products and targeting consumers. This year, *Significance* has partnered with *Impact*, the magazine of the Market Research Society, to jointly publish a series exploring the past, present and future of the data economy. In this fourth and final part, **Timandra Harkness** considers what the coming years have in store for the data-driven industries

Do you want to feel special? Go to coveryourtracks.eff.org and click the "Test Your Browser" button. That's how I found out that my web browser fingerprint is unique among the 220,694 the Electronic Frontier Foundation tested in the previous 45 days.

This was a surprise. It means that even if I refuse tracking cookies – which I do – advertisers can still follow me around different websites, using a combination of innocuous details like my browser version, screen size, graphics set-up and system fonts.

In short, getting rid of third-party cookies, as Google has promised to do from its Chrome browser by late 2023 (bit.ly/3DnjUvM), will not bring the online data economy to a screeching halt. But that does not mean that things will carry on as before. Major changes are afoot in the data-driven industries, spurred by privacy concerns, tightening regulation, and technological advance.

Federated learning

In this series, we have followed the progress of statistical and computing methods for drawing insights from data: from sampling to constructing an "n = all" whole population model, from Victorian techniques of regression and dimension reduction to machine learning models whose detailed workings are mysterious even to those who program them.

The next challenge for those using data to understand people will be to preserve people's privacy and autonomy while drawing conclusions, if not about them personally, about relevant populations. Take Google's cookie announcement. You might, cynically, suggest that it is merely a way for Google to monopolise the ability to target adverts to you. But some of Google's "Privacy Sandbox" proposals have the potential to radically change how researchers and marketers work.

Fledge, developed by Google (bit.ly/3llT3ud), is one potential solution to the problem of, say, being repeatedly targeted by car ads after you have researched a car online. It lets the user's own browser automatically tag a topic of interest for a specified length of time. No central data store will flag the browser or shopper. It is like wearing a lanyard at an event that says "talk to me about cars" but which you can take off at the end of the day, instead of being added to somebody's marketing list for ever.

Another of Google's ideas is FLoC (bit.ly/3oCz8ZW), or federated learning of cohorts. Like Fledge, FLoC moves away from the idea of allowing ad tech companies to amass browsing data on individual web users in a centralised pool. Instead, it creates many "cohorts" of web browsers, grouped by patterns of activity. The web user's own browser then calculates which of these cohorts corresponds most closely to its recent browsing history. That browser-selected cohort is used to target relevant ads to the browser, not the person, who can remain anonymous throughout. The system preserves practical anonymity by letting each user hide in a cohort of thousands of individuals.

The "federated learning" in FLoC refers to a way of training machine learning programs without amassing a large quantity of centralised data. The program is given access to many smaller databases, each of which trains the model locally, with only the results centralised and aggregated.

Such systems have a particular appeal for companies like Google and Apple, who make not only apps but also ecosystems. These companies really are not that interested in collecting data on us, as individuals, argues digital rights researcher Michael Veale. "What they want is the ability to do calculations over your data." He gives the example of voice assistants, like Apple's Siri. This app learns to recognise a user's voice so

Keeping things private

Centralised systems of data collection and analysis enable whoever has been trusted with data to analyse it without revealing private information to anyone else. How can multiple companies (or researchers) learn from decentralised data sources without learning too much about either individuals or their competitors' data?

Secure multi-party computation (SMPC)[1] aims to mathematically emulate the single trusted party.

StJohn Deakins of CitizenMe gives a very simple analogy: "You've got a room with 10 people in, and you want to find out the average shoe size. So, you take a random number, 257, send it to the first person. They add their shoe size [to that number], you go round," and when everybody has added their shoe size you have one total number.

Now you just need to subtract 257 from that total, divide by 10, "and you've got the average shoe size. But you don't know anyone's shoe sizes, because it's gone out around the 10 people." Add in homomorphic encryption, which allows you to do calculations on encrypted data without decrypting it, and you have the potential for securely using private data for research without revealing exactly whose data has contributed what to the result.

First developed in the 1990s,[1] SMPC is one of the "privacy enhancing technologies" discussed in a 2019 Royal Society report.[2] The report gives an example of SMPC use in 2008 to distribute Denmark's EU-fixed sugar beet quotas among producers.

The Danish and Dutch scientists behind that trial, led by Peter Bogetoft, described the reluctance of individual beet farmers to trust the sole beet processor, Danisco, with sensitive commercial information (bit.ly/3lid7NL). "[W]e have therefore become convinced that the ability of multiparty computation to keep secret everything that is not intended to be public, really is useful in practice," say the researchers.[3]

Their method used standard cryptographic techniques including public–private key pairs and modulo arithmetic with large prime numbers.

They did, however, note that some computers took up to a minute to encrypt the 4,000 numbers comprising each bid.

As both mathematical protocols and computer speeds improve, SMPC is becoming fast and efficient enough to be a practical technique. In 2019 Facebook filed a patent (bit.ly/3BnMnRA) for using SMPC to evaluate online marketing campaigns, explaining that "Secure multi-party computations may be used to get attribution results without compromising user privacy" (bit.ly/2Yv7B1i).

as to improve an iPhone user's experience. But by teaching Siri to recognise their voice, the user might have helped train the average Siri model – the one that comes pre-installed on every iPhone – to do a better job of understanding a specific accent.

The improvements you help make as an iPhone user "don't reveal anything about you", says Veale. "They're just improvements to that average model that came to you." And it is the improvements, not the voice data, that are sent to Apple's central algorithm, where "they can be aggregated up and synthesised into a societal improvement, which then gets downloaded again to everybody's device, and vice versa. That's federated learning," says Veale. "And that's private in so far as you're not sending the data, you're just sending the way that the model learned to get a bit better from your data."

You can see how this shift in approach might benefit corporations like Google and Apple, who make, if not always the physical device, certainly the operating system on which a device runs. The user controls their own data, but Google and Apple will be the gatekeepers to it, from which all insights can be drawn.

Personal data stores

This is not the only possible model of the future. How about one in which your data sits, not on a phone or computer made by a Silicon Valley giant, but in a small wooden box on a shelf in your house? That was the vision of a delightfully quirky project called BBC Box (bbc.in/2YrPSYC).

In 2019, a research and development team within the BBC created a hexagonal box – with a whiff of Dr Who's Tardis – containing a Raspberry Pi computer that ran a personal data management system named Databox. The idea was that personal data from a range of different digital services would be stored within the BBC Box, and then it would be up to the user to decide which other apps and programs could access and process that data (see "Keeping things private"). As an example, the BBC developed its own "Profiler" app that would produce an anonymised profile of the Box's owner. That profile – but not the data – could then be exported by the user to a system to produce recommendations of TV shows the user might like.

"Starting from the premise that we're the BBC, and we have a duty of care, not just to our contributors, but to our audience as well, preserving people's privacy is part of that duty of care," says Bill Thompson of BBC R&D. "We are examining models for developing audience insights that don't require us to know anything about you, but that let you tell us enough about yourself … [to build a model giving] a more granular and useful understanding of our audience than we would get by knowing about you particularly."

Obviously, it is not necessary to have a physical container in which to store your data. A virtual container would work just as well. The internet of the future could be a honeycomb of individual data cells, each one containing an individual's personal data.

The creator of the World Wide Web, Tim Berners-Lee, is looking at exactly that model. Concerned that the internet has become a machine for monetised surveillance, rather than an ecosystem of co-operative sharing, he has been working on a new vision of the web called Solid (solidproject.org) – with the name derived from the phrase "social linked data". Solid started at the Massachusetts Institute of Technology and now has its own start-up – Inrupt – to take it closer to fruition. Meanwhile, the same BBC R&D team that built BBC Box is working on an experimental pod-based personal data store (PDS) approach to recommendations, called My PDS. Like the BBC Box, the idea is that

Timandra Harkness is a presenter, writer and comedian. Her BBC Radio 4 documentaries include *Five Knots* and *Steelmanning*, and she is the author of the book *Big Data: Does Size Matter?*

each "pod" pulls together data from different sources – BBC iPlayer, Spotify and Netflix – to create a media profile to which the user can, if they want, grant access to other BBC apps, such as the BBC Sounds app.

All of these projects are experimental prototypes. In Europe, such ideas have been given a leg-up thanks to "the right to data portability" enshrined in the General Data Protection Regulation (GDPR). This right is described by the UK Information Commissioner's Office as allowing "individuals to obtain and reuse their personal data for their own purposes across different services" (bit.ly/3oLG5rP).

Many prototype PDS designs have been built to facilitate this sort of sharing and reuse of data. Some include a dashboard for terms and conditions, so users can be alerted if these change after data has been shared. Others include a token that travels with the data, like a watermark in a digital photograph, specifying what permission has been agreed for its use.

One app, CitizenMe, lets individuals collect data about themselves in a PDS and offer it to places where it could be useful. "The first place is market consumer insights, obviously," says chief executive StJohn Deakins, "because if you've got a large cohort of people with lots and lots of deep multivariate personal data, you can drive a huge amount of insight off the top of that." CitizenMe users might receive offers to share data and answer questions for cash, and they can also donate data for good causes or participate in studies that give them more information about themselves. Deakins says he learned that "people don't really care about the data, but they care about the stories that data tells. Especially about themselves, or people they're close to."

Liz Brandt, chief executive at Ctrl-Shift, sees many opportunities arising from GDPR's right to data portability. For individuals, greater ownership and control of their personal data could allow them to demand a share in the benefits from its use. For businesses and researchers, greater user control might mean that data quality improves, and that they are no longer getting messy, out-of-date or deliberately misleading information from unwilling subjects.

Brandt thinks the UK economy "can gain £27.8 billion in productivity and efficiency through data portability", but just as important, she thinks, are "the new innovative things you can do with it". Realising this potential, though, will require a change to the existing data economy – not just in the UK, but internationally.

This is one thing on which everyone seems to agree: the need for a new system of regulation, of interoperability for apps and programs, and of shared infrastructure to make all the parts work together.

Regulation, consumer inclination and technology are converging towards an expectation of greater privacy and control

Where to now?

It is tempting to believe that in the near future, each of us will have our data tucked away in an individual account, over which we have complete control. Certainly, regulation, consumer inclination and technology are converging towards an expectation of greater privacy and control for the person concerned. If this future does come to pass, the challenge for regulators will be to turn their attention from data to infrastructure. If just a few companies control the systems within which our personal data stores operate, they will arguably have as much power as today's data giants, albeit with less liability for when things go wrong.

It is, however, misleading to think of all data about an individual as "belonging to" that individual. If I use my "smart" railcard to move around a city, every tap in and out of a station generates data for the transport company, and they are not going to stop collecting that data centrally, because they need it to operate their systems, predict demand, and perhaps to understand more about who they are not serving. Nor is my bank going to stop keeping a record of all my transactions.

What is likely to happen is a growing separation between "personal data" and "aggregated data about people". Regulators are already pushing that division by increasing the risks to those who collect, store and use personally identifiable information (PII), with hefty fines for misuse or leaks. Reputational risks, too, mean large tech companies have a vested interest in *not* collecting PII if they can avoid it.

Indeed, Michael Veale believes that, over time, "data is going to become less relevant". For organisations of all stripes, the value of data has always been the information about relationships that it captures: between people, between people and companies, even between people and their devices. Understanding, targeting and influencing people is the end goal, not amassing vast piles of 1s and 0s. When extracting value from data can be done at the point where the data is generated – via on-device processing, for example – why should companies need to accumulate data themselves? They can turn their attention instead to "convincing people to integrate more and more of this stuff" – things like smart watches, digital assistants, and smart refrigerators – "into their homes, lives, bodies," says Veale. "Then these companies are just intermediaries."

"They try to get in the middle of stuff," Veale explains, "whether it's your messages, your payments, your social connections and friends, between you and your music, you and your cooker, because this cumulative power can give them the options to shape your environment, shape your behaviour, shape your interaction to extract money from you."

If this all sounds like the future of the "data economy" is already here, well, maybe it is. Parts of it, at least. But you can bet there's more to come. ■

Disclosure statement

The author declares no competing interests.

References

1. Goldreich, O. (2002) Secure Multi-Party Computation (manuscript). Final revision, 27 October.

2. Royal Society (Great Britain) (2019) *Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis*. London: Royal Society.

3. Bogetoft, P., Christensen, D. L., Damgård, I., *et al.* (2009) Secure multiparty computation goes live. In R. Dingledine and P. Golle (eds), *Financial Cryptography and Data Security* (Lecture Notes in Computer Science, vol. 5628, pp. 325–343). Berlin: Springer.