Chapter 4
Big Data

Felix Chan

Abstract This chapter provides a brief history of *big data* and reflects on its impact on the future of econometric analysis when the volume of data grows exponentially with increasing complex structure. A brief history of data is presented with a focus on its evolution from *data* to *big data* and covers both *structured* and *unstructured* data as well as their convergence and complementarity in terms of data collection, storage, management and methods of analysis. It emphases on the role of technologies in the distribution and analysis of big data, and provides an overview of two data principles and frameworks that facilitate research reproducibility while respecting data generated through first-nation and indigenous population. Finally, the evolution of statistical methods in analysing increasingly complex data and the role of Natural Language Processing (NLP) and Large Language Models (LLM) in the context of economic and econometric research is also discussed and explored.

4.1 Introduction

The importance of big data to economic and econometric analysis has been receiving attention since the early 2010s, see for examples, Letouzé (2012), Varian (2014), and Taylor, Schroeder and Meyer (2014). More generally, these contributions do not just highlight the impacts of big data on econometrics, but also the values in cross-pollination between econometrics, data science and digital technologies.

Inspired by these contributions, this chapter is an attempt to consolidate the history and the current status of three areas namely, the history of economic data, the recent development of data science due to increase accessibility of big data and the development of digital technologies that are relevant to economic and econometric research.

Felix Chan ✉
Curtin University, Perth, Western Australia e-mail: F.Chan@curtin.edu.au

The main objective of the chapter is to reflected on the history, development and the interactions between these areas with the aim to provide some insight on the future development and education of econometrics in the context of (big) data and rapid technological development, such as Generative Artificial Intelligence.

Econometricians have been accustomed to analysing data. However, these are mostly structured and tabulated numerical data, or at least data that can be mapped to a set of numerical values in a straightforward manner. As accessibility to different types of data emerges, such as audio, video and unstructured textual data, and the volume of relevant data becomes large, the term *big data* emerges to describe the situation where the supply of data pushes our ability to understand them. While the term *big data* has only surfaced since 2000s, the challenge of handling data that requires capability beyond existing technologies is not new, and such challenge can be argued as the main driver of technological innovation (see for example, Hollerith, 1889).

The solution to the challenge usually involves both methodological and technological innovations. Perhaps a unique feature of the recent development is that big data is no longer just a challenge to a small number of selected fields, but rather, it has impacts on virtually on all areas, especially in the context of interdisciplinary research. The ability to combine data generated through different fields, allow more insightful analysis to be conducted and it also demonstrates the interconnectedness of our World.

The interconnectedness also creates challenges and their solutions demands standardisation of data between disciplines across many issues, including data governance, data privacy, data collection and management, as well as research best practices. The amount of data available also pushes the boundaries of statistical analysis and disciplines are beginning to learn from each other on the different approaches to extract as much information from these data as possible.

It is also worth noting that the impacts of data science and web-technologies on economic and econometric research are multi-dimensional. The developments of these areas lead to increased avenues of collecting, curating and disseminating (big) data. Perhaps more notably, these developments contribute substantially to Generative Artificial Intelligence (GAI), particularly in the form of Large Language Models (LLM). LLM in particularly, has significant impacts on virtually every part of research workflow, ranging from literature review, data collection, data curation, data creation, data storage, processing and retrieval to programming, data wrangling, statistical analysis, visualisation and presentation of results. This is not just happening in economics and econometrics, but across all disciplines.

Thus, understanding the history and development of these three areas and their interactions would seem to be an important step to understand and perhaps, project the future of economic and econometric research.

The chapter is organised as follows. Section 4.2 discusses the definition of data in the context of academic research. It then provides a brief history of economic data collection and the evolution of their accessibility through various government and non-government agencies. Section 4.3 introduces the concept of 3Vs developed in data science and their applications to describe and measure various characteristics of

big data. Some examples from spatial and network economics is also provided to examine the applications of this framework in the context of econometric analysis. This is followed by Section 4.4 which introduces the FAIR and CARE data principles. These principles facilitate reproducible research and research transparency, as well as providing guidance on respectful data governance on data generated from indigenous and first nation population. The latter would be particular relevant in research such as *random control trails* in developing countries, which is a popular area of economic research. The section also discusses the construction of big data via data linkage as well as the concept of probabilistic data linkage for de-identified data in the context of combining de-identified administrative data that are commonly used in health economics. Section 4.5 provides a brief historical account on the statistical techniques that are used to analyse big data. This section should be viewed as a teaser to other chapters of the book, such as Chapters 10 and 12. The remaining of Section 4.5 focus on Natural Language Process (NLP) and Large Language Models (LLM). The availability of big data makes these techniques and technologies possible and relevant to economic and econometric research. Their impacts are not limited to the generation of results, but they are also fundamentally shifting academic research practice. Section 4.6 provides concluding remarks summarising the lessons learned from each of the sections in the chapter as well as providing some projections to the future of econometrics research and education.

4.2 From Data to Big Data

4.2.1 What is Data?

Given this chapter is about big data, providing a definition of *data* may seem like a sensible starting point. This, however, turns out to be a rather challenging task as defining *data* is not a straightforward assignment. It would appear that the definition of data evolves over time and changes depending on its context. To give an example, let's start with definition from dictionary. The *Cambridge Dictionary* defines *Data* to be

> "*information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer*"[1]

There are some interesting features with this definition. Firstly, the definition seems to suggest that data and information are in some sense equivalent, although this view has been actively debated in the knowledge management literature. For example, Tuomi (1999) points out the importance of distinguishing between *data*, *information* and *knowledge* and this view has been further enforced in Joshi and Krag (2010). These distinctions are consistent with the Data-Information-Knowledge-Wisdom (DIKW)

[1] Cambridge University Press (2021)

pyramid, which origin can be traced back to the work of T.S. Eliot in 1934. In the work "The Rock", T.S. Eliot posed the sequence *"Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"*. Although the DIKW pyramid (or model) remains an actively debated topic in the knowledge management literature as highlighted by Joshi and Krag (2010), it is agreed that data, information, knowledge and wisdom are related but different concepts.

Secondly, the dictionary definition specifically includes digital objects or objects in electronic form that can be stored and used by a computer. While this is an obvious attempt from the dictionary to keep the definition 'up-to-date', it also suggests that the nature of data has evolved over time and its evolution is intimately related to technological progression and therefore, not necessarily future-proof.

Perhaps another useful approach is to compare definitions of data from statistical agencies and research institutions. The Australian Bureau of Statistics defines data as

"Data are measurements or observations that are collected as a source of information. There are a variety of different types of data, and different ways to represent data."[2]

while University of York adopted the definition from Wikipedia[3] and defines data as

"In the pursuit of knowledge, data is a collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted."[4]

These definitions are perhaps more descriptive, but they remain too generic to be useful. In the context of academic research, the term *Research Data* has recently been proposed, which is perhaps more relevant to economic and econometric research. A definition of research data is

"*Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form."*[5]

University libraries often define research data as

"recorded factual material commonly accepted in the scientific community as necessary to validate research findings, excluding drafts, peer reviews, and informal communications"[6]

There is a common theme among these definitions, despite their variations. Specifically, data in this context refer to all materials that are required to reproduce the research outcomes or findings. Therefore, one of the main objectives of research data is to facilitate research *reproducibility*. This, of course, covers the data used during research, but also covers data generated as a result of the research activities, which

[2] Australian Bureau of Statistics (2026).

[3] Wikipedia (2026)

[4] University of York Library (2026)

[5] Openaire (2026)

[6] Oregon State Univesrity Library (2026)

includes programming code or prompts used to generate relevant responses from Large Language Model (LLM) that lead to specific research findings e.g., prompts that instruct LLM to generate numerical data through analysing graphic or audio files.

The last example also highlights the use of *unstructured* data to generate *structured* data. As the name suggests, Structured Data is data organised into a predefined *schema* or *data model*, typically using tables with fixed fields. Economists and econometricians would be familiar with this format, e.g., tabular data that are typically used in econometric analysis where each row represents an observation and each column represents the value of a variable.

Unstructured data is data that has no fixed schema or consistent internal structure, and is stored in its original native format. Examples include transcripts generated through interviews, collections of academic papers through literature reviews or articles and images from other sources such as webpages or newspapers. Collection of audio and video files are also examples of unstructured data. A major challenge in managing unstructured data is the lack of ability to search and identify relevant information.

While econometricians would be familiar with structured data, generating structured data from unstructured data is becoming increasingly common. One example is the construction of a sentiment index, where word count from relevant articles (unstructured data) is often used as part of the construction of a sentiment index. This process requires Natural Language Processing (NLP) to process input texts and attempts to map the content of the article to a measure of, hopefully, coherent, measure of sentiment. In recent times, this process has been facilitated greatly by Large Language Models, even though there are evidence that classical NLP techniques remain relevant in this process, see for example, Kallstenius, Capusan, Andersson and Williamson (2025).

Despite the many definitions of data, this chapter focuses on research data as it appears to be the most relevant definition in the context of academic research. Perhaps more importantly, the focus of research data facilitates and promotes the notion of *research reproducibility*, which, given the increase complexity and availabilities of data, is becoming increasingly important.

Lessons Learned:

1. Definition of data often depends on context and the general definition not be fit-for-purpose in the context of research.
2. Definition of data also changes over time, often related to technological progression and is unlikely to be future-proof.
3. Definition of research data is more appropriate in the context of academic research but can still be too generic to be useful. However, the definition is closely related to research reproducibility, which is becoming increasingly important.

These lessons have implications to the education of future econometricians, where the notion of data has to be broaden. It should not just be about the data that they analysed, but any materials that are used to generate the research findings, including the process of data collection, ingestion and processing, the software and related

programming code that are used to collect, process, clean and analyse the data, including prompts that are used to generate relevant responses from Large Language Models.

4.2.2 Brief History of (Economics) Data

Given the scope of research data, this section focuses on the history of data that are most familiar to economists and econometricians. That is, tabular data that can be used for econometric analysis. The section discusses the availabilities of such data from a historical perspective. The objective is to highlight the increasing complexities and availabilities of these data due to the rapid development in technologies such as databases, the internet, and related web technologies.

Collection of economic data can be dated back to medieval Europe through port records. The primary objective of this data collection was to detail taxes and duties on imports and exports, see for example, Rössner (2020), for a more detailed account on its practice. Given the nature of its purposes, it is perhaps not surprising that the first generation of quantitative documentation for trade flows was driven purely by administrative requirements.

There were also records of the Spanish government systematically collecting detailed trade data for its colonies between 1504 and 1595. Records preserved in the Archive of the Indies included extensive ship-level data on name, type, size, crew, cargo and ports of call. This represented an early system for monitoring colonial commerce. Rogin (1956) provides a detailed account on the relation of this data collection and its influence on the formulation of economic theory.

Statistical materials can also be found in Ain-i Akbari which focus on economic conditions, serving administrative requirements and tax collection during the Mughal Empire era. According to Moosvi (1987), these materials would later enable modern quantitative analysis of the empire's economy.

While it is unclear exactly when economic scholars started using these data to analyse the economy, empirical debate involving both administrators and scholars emerged in the eighteen century. According to Mitra-Kahn (2011), Davenant's civil service and Walpole's Treasury accounts defined the economy through official measurements in Britain but by 1770s, scholars began to displace secular policy advisors in economic analysis.

At around the same time, the British government began collecting weekly price and quantity data for grain traded in market towns starting in 1770. These data were published in the London Gazette through 1914. This is perhaps one of the earliest examples of high-frequency economic data collection serving both policy and market information needs. Readers are referred to Brunt and Cannon (2013) for a more in-depth exposition on the objective and practice of official statistic sharing during that period.

Overall, it can be seen that the earliest purposes for economic data collection were overwhelmingly administrative, focused on tax collection and trade monitoring.

However, by the eighteenth century, data collection increasingly served policy decision-making needs and economic planning as well. The first mentions of academic use appeared in seventeenth century when Sir William Petty and Gregory King began systematic economic measurement. Rice Vaughan's 1675 work and William Fleetwood's 1707 analysis represented the early scholarly applications using economic data for academic analysis rather than purely administrative purposes.

While the practice of data collection has a long history, it was, and perhaps still is, a labour intensive exercise. As a result, economic data collection is usually designed for a very specific purpose, subject to the motivation of the data collection agencies with very little attention to data *interoperability* i.e., lack of attempt to ensure data from different sources could be used together in a meaningful way.

Nevertheless, data accessibility was often a challenge not in terms of institutional silos, but due to physical proximity and methods of copying large amount of data manually. Researchers often extracted data from physical archives and records, or scanned industry and trade journals, which they would also need access. This changes in the mid-1990's when data digitalisation became common practice and data sharing through network and the internet became feasible.

Tables 4.1 and 4.2 provide summaries of selected organisations and statistical agencies on the approximated start dates of their online data portals. There are two interesting observations from these tables. Firstly, organisations that are not statistical agencies had different attitudes towards openly sharing their data before 1990's, despite the consistent efforts in collecting and analysing data within these organisations. These organisations, however, progressively started sharing their data and provided application programming interface (API) that facilitates data extraction from their online portal since mid-1990s. Statistical agencies, however, provide online data portal from 1990's in addition to sharing data through statistical publications.

The turning point for data sharing in 1990's is no coincident as the mid 90's marks the point when the cost of digitalisation of documents and storage of digital object become cheaper than the cost of keeping physical copy of data. The fact that the number of online data portal experienced exponential growth since the 2010 is also consistent with the rapid development of internet infrastructure and web related technologies, particular in the development of web browser technologies and the increase in bandwidth. Goldfarb, Greenstein and Tucker (2015) provides a comprehensive account of these developments and their economic impacts.

It is also interesting to compare the time it took for statistical agencies to start publishing their data online versus organisations that are not statistical agencies, such as the International Monetary Fund and the World Bank. Even though some data might have been available on their website, the availabilities of various API for direct assessing and downloading their available data was not fully developed until 2010's. This is due to various reasons, including the regulatory and resource constraint relative to the core function of the organisations. Data governance, ownership, ethics and cybersecurity are also important factors as highlighted in Rockhold, Nisen and Freeman (2016).

Discussion so far has been focusing on data that are publicly available. Data service providers have also played an important role in providing data for researchers

historically. While financial market data features heavily from these vendors, there are also commercial data service providers that provide economic orientated data. Although some of these data are publicly available, the cost of continually collecting such data, such as stock prices, often exceeds the the cost of subscription, especially in terms of number of researchers that can access such data within an organisation. Moreover, publicly available data only form part of the data required for financial research. These data service providers often provide proprietary and in some cases, commercially sensitive and confidential, data at the firm level, in addition to publicly available data.

It is perhaps useful to point out that the targeted consumers of these data were investors in the early days, and not academics. From that perspective, it is perhaps not surprising to learn that there is a long history of financial data service providers dating back to 1870s. The first technology that allowed these providers to disseminate data is via ticker tape. While such method of dissemination assisted investors in their investment decisions, it was not particularly useful from the perspective of research. Challenges of processing the data into useable format and the computational limitation at the time were two of many obstacles that prevented the use of such data in academia in the early part of 1900s.

As technology progresses, most notably the rapid increase in computation power, allows more rigorous analysis of these market data. Readers are referred to Chapter 3 for more in-depth discussion on the development of financial econometrics and the availability of high frequency data from the financial market.

Since 1990s, with the availabilities of platforms such as Google Finance and Yahoo Finance, some market data becomes accessible to public. These include stock price data up to daily frequency for large number of firms, particularly from the markets in United States. Nevertheless, these financial data providers remained important and relevant to investors and academia. This is reflected by the size of the Market Data industry.

Despite some of these data can be obtained readily through public domain and data market has become more fragmented, the number of financial data providers has increased. Services such as Wealth Hub (2026) and re3Data (2026), become popular for purpose of identifying data service providers and types of products (and data) that they provide. While these are extremely useful and convenient, it is also indicative of the fragmented nature of the (financial) data market.

In addition to data service providers, another main source of accessing data is via data repositories or data exchange platforms such as re3Data (2026), Kaggle (2026), DataLore (2026) and Big Query from Google Cloud Service (2026). These platforms are result of rapid development in web technologies and open source software that facilitate online collaboration of data analysis. These platforms were motivated by the idea that users can collaborate in analysing complexing data, mostly through construction of Jupyter-like notebook Jupyter (2026). Given the motivation was built on collaboration online, data sharing becomes a necessity as part of this process.

These platforms are very useful for purposes of collaboration with sophisticated data sharing protocols that respect different data licenses, see Section 4.4 for further

Table 4.1: Online Data Portal from Selected Organisations

Agency / portal	Approx. start of statistics / data portal	Main types of economic & financial data shared
International Monetary Fund (IMF)	IMF founded 1944; modern online data portals from 1990s–2000s	Macroeconomic indicators (GDP, inflation, employment), balance of payments, government finance, financial soundness indicators, cross-country time series used in World Economic Outlook and other databases
World Bank	World Bank created 1944; World Bank Open Data portal launched 2010	Global development and macro national accounts, poverty, inequality, trade, education, health, infrastructure, environment, financial inclusion and development indicators, mostly at country level
Organisation for Economic Co-operation and Development (OECD)	OECD in current form from 1961; online OECD.Stat portal from early 2000s	Detailed economic, social and environmental statistics for members and partners: national accounts, productivity, prices, taxation, labour markets, trade, innovation, and sectoral indicators
United Nations (UN) Statistics Division and specialised agencies	UN founded 1945; many statistical series from late 1940s onward	Broad economic and social statistics including national accounts, trade, population, industry, tourism (UNWTO), labour (ILO), agriculture (FAO/FAOSTAT), and regional economic indicators via UN regional commissions
Bank for International Settlements (BIS)	BIS established 1930; coordinated international financial statistics from mid-20th century; BIS Data Portal in 2010s	Global banking and financial market international banking statistics, debt securities, credit to private and public sectors, derivatives, FX turnover, property prices, effective exchange rates, and central-bank balance-sheet series
Agency / portal	Approx. start of statistics / data portal	Main types of economic & financial data shared
European Central Bank (ECB)	ECB started 1998; Statistical Data Warehouse launched mid-2000s	Euro area and EU financial and macro monetary aggregates, interest rates, securities, balance of payments, financial stability indicators, bank balance-sheet and market data, plus macroeconomic series for member states
World Trade Organization (WTO)	WTO created 1995 (building on GATT from 1948); WTO data portal active since 2000s	International trade in goods and services, tariffs, trade remedies, and trade-in-value-added indicators across countries and product groups
Food and Agriculture Organization (FAO) – FAOSTAT	FAO founded 1945; FAOSTAT database first released in 1960s, web portal from early 2000s	Agricultural production, prices, trade, food balance sheets, land use, emissions, and related rural and environmental statistics with strong economic content for agriculture and food systems

Table 4.2: Online Data Portal from Selected Statistical Agencies

Country	Name of Agency	Approx. Start of Agency	Approx. Start of Online Accessibility
Australia	Australian Bureau of Statistics (ABS)	1905 (Commonwealth Bureau of Census and Statistics; ABS name from 1975)	Mid-1990s–early 2000s (ABS website and downloadable data; interactive data tools expanded 2000s–2010s)
Canada	Statistics Canada	1918 (Dominion Bureau of Statistics; Statistics Canada name from 1971)	Late 1990s (online CANSIM tables), with modern web data portal in the 2000s
United States	U.S. Census Bureau / Bureau of Labor Statistics (BLS) / BEA	Federal statistical functions since 19th c.; Census Bureau roots 1840s; BLS 1884; BEA 1972	Mid-1990s for major agencies' web sites; extensive downloadable data and APIs from 2000s onward
United Kingdom	Office for National Statistics (ONS) / UK Statistics Authority	Central statistical office lineage from 1941; ONS formed 1996	Late 1990s–early 2000s (National Statistics website); modern ONS data portal and API expanded in 2010s
Euro area / EU	Eurostat (Statistical Office of the European Union)	Eurostat created 1953 (in earlier form), current role formalised with EEC/EC development	Late 1990s–2000s (Eurostat online databases); current data browser and bulk download tools from 2010s
France	INSEE – National Institute of Statistics and Economic Studies	1946 (successor to earlier statistical bodies)	Late 1990s–2000s (INSEE online data); APIs and interactive tools expanded in 2010s
Germany	Federal Statistical Office (Destatis)	1953 (modern office; earlier imperial/state statistics date back to 19th c.)	Late 1990s–2000s (GENESIS-Online databases and web tables)
Italy	Italian National Institute of Statistics (ISTAT)	1926	Late 1990s–2000s (ISTAT data warehouse online), with more advanced web tools in the 2010s
Spain	National Statistics Institute (INE)	1945 (continuing earlier statistical services)	Late 1990s–2000s (INE online data portal)
Netherlands	Statistics Netherlands (CBS)	1899	Late 1990s (StatLine online database) with successive modernisations since
Sweden	Statistics Sweden	As an agency in current form from 1960s; earlier offices from 19th c.	Late 1990s–2000s (online statistical databases)

Table 4.2 Cont: Data Portal from Selected Statistical Agencies

Country	Name of Agency	Approx. Start of Agency	Approx. Start of Online Accessibility
Japan	Statistics Bureau of Japan	Central statistics functions since late 19th–early 20th c.	Late 1990s–2000s (e-Stat and bureau web portals)
China	National Bureau of Statistics of China	1952	2000s (NBS web site and online data series), expanding with English-language tables thereafter
India	Ministry of Statistics and Programme Implementation (MoSPI)	Central statistical functions from 1950s; MoSPI in current form from 1999	2000s (MOSPI and India "Open Government Data" portals)
Brazil	IBGE – Brazilian Institute of Geography and Statistics	1936 (as IBGE; earlier statistical activity predates this)	Late 1990s–2000s (SIDRA and IBGE web databases)
Mexico	INEGI – National Institute of Statistics and Geography	1983 (as INEGI; successor to earlier bodies)	2000s (INEGI web data portal and dynamic tables)
South Africa	Statistics South Africa	1994 (successor to earlier Central Statistical Service)	Late 1990s–2000s (Stats SA online datasets and publications)
Kenya	Kenya National Bureau of Statistics	2006 (building on earlier Central Bureau of Statistics)	Mid-2000s onwards (KNBS website and downloadable data)
Russia	Federal State Statistics Service (Rosstat)	As separate state statistics body from 1987 (continuing Soviet Goskomstat)	Late 1990s–2000s (Rosstat web dissemination)
New Zealand	Stats NZ (Statistics New Zealand)	1956 (as Department of Statistics; statistical roles earlier)	Late 1990s–2000s (Infoshare and later data portals)

details on data licenses. It should be noted however, that the quality of data on these platforms varies and should be used with cautions.

At this point, it may be useful to also provide a brief history of 'big data', or more appropriately the history on the term 'big data'. While there is no consensus on when the term 'big data' was first coined, John Mashey was often credited as one of the first people to have coined the term in the mid 1990 while he was working for Silicon Graphics Industry. Despite the fact that he did not record the term in any

academic paper, the term appears in some of his industry presentations. Diebold (2012) provides an interesting account on the origin of the term.

While there was no formal definition given to what may be considered as big data, John Mashey appeared to use the term to describe data that are too large for the existing technology to process, manage or analysis.

Interestingly, it appears that the usage of the term 'big data' that matches more closely to the modern usage of the term is from an econometrician, Francis X. Diebold. In his discussion to two papers during the Eighth World Econometrics Congress in 2000, he presented his discussions paper as *"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting"* (Diebold, 2003). In that paper, Francis Diebold suggested the (memory) size of the file may be a better measurement for big data rather than the number of observations, which was typically used. Even today, the number of observations remained as an important characteristic in the context of meta-data.

Of course, the framework of measuring big data has been evolved since Diebold (2003), and is the focus of the next section. Reflecting on the history of economics data in the context of technological progress, below are some key lessons learned.
Lessons Learned:

1. (Economic) Data collection has a long history and with specific purposes, which were not typically motivated by academic research.
2. Economic data collection is usually expensive and resources demanding. So they are usually done through well funded organisations (and not by individuals).
3. While these organisations were often willing to share their data, methods of data dissemination were limited until mid-1990s, when digitalisation of data became cheaper and the development of web technologies and internet infrastructure facilitated the dissemination of large amount of data digitally.
4. Such technologies also facilitated increasing avenues of data sharing, for commercial data service providers, government agencies, not-for-profit organisations, collaborative platforms and other organisations.
5. The diverse avenues of data sharing cast questions on data quality, currency and data governance.

4.3 Measuring Big Data

The previous section provided an overview on the development and availabilities of economic (and financial) data and argued that the rapid development of web technologies and improvement of internet infrastructure continue to facilitate dissemination of data. One conclusion of these arguments is that as the availabilities of data increases, so are the volume and complexity of data. This leads to the *big data* phenomenon. This section discusses the different characteristics of big data and the framework that can be used to describe these characteristics wholistically. These characteristics would also shine light on the question of 'how big is big?'.

As mentioned earlier, John Mashey defined *big data* as data that are too large for existing technologies to process, manage or analysis. If one accepts this definition, then the 'big data' problem is not new and it will almost definitely be present into the future. To understand this, let us start with a case study. In 1880, data generated from the US Census was expected to take 10 years to process, given the technology at the time. *Herman Hollerith* saw the opportunity and subsequently invented the first 'computer', using punched card to process the data. His invention reduced 'computation' time from 8 years to approximately 2.5 years, see Cortada (1993) for a comprehensive discussion of the event. This is one of the earlier record of 'big data' problem. Incidentally, Herman Hollerith later found a company called *Computing-Tabulating-Recoding company* based on his invention. This company still exists today but it is better known *International Business Machine* (IBM).

Given the current technologies and the internet infrastructure, as well as the diverse sources of data, it should be clear that John Mashey's definition of big data is not adequate as a measure. In fact, the suggestion put forward in Diebold (2003), while important and closer to the nature of big data today, still does not fully capture all the characteristics of big data.

As argued earlier, technological progress has led to increased variety and complexity of data and the measurement of big data must therefore reflect its characteristics which is in fact multidimensional. In addition to volume, which is typically measured in terms of units of memory in a computer system i.e., bytes, velocity and variety have also been included as characteristics of big data.

Before diving into each of these characteristics, it would be useful to point out that these characteristics of big data implicitly assume the data is a digital object, or collection of digital objects, which can be stored in a digital storage device, such as a computer hard-drive. While some of these characteristics are not limited to digital object in a strict sense, it may not be practical to consider them otherwise. For example, volume can strictly speaking apply to tabular data in physical form, such as the number of statistical year book, and strictly speaking the information contains in a physical book can be measured in bytes. However, such exercise may not be particular useful in the context of measuring big data, especially given these data can actually be digitalised. Therefore, these characteristics are further discussed in the context of big data as digital objects.

4.3.1 The 3Vs: Volume, Velocity and Variety

The concept of the 3Vs, Volume, Velocity and Variety, is often accredited to Doug Laney who was an analyst at Mete Group (later acquired by Gartner). In his research note entitled "3D Data Management: Controlling Data Volume, Velocity and Variety", he discussed the three characteristics of big data and these have slowly become the basic measures of big data. Since then, more characteristics have been added to the 3Vs including *Veracity* (V) and *Value* (another V) in addition *Complexity* (C).

These characteristics are conceptual and general. Not all of them are relevant to data typically used in economic or econometric analysis, at least not at the time of writing this chapter. However, the 3Vs are directly relevant to the economic and econometric discipline, so this section will focus on the 3Vs and provides a brief discussion on the other 2Vs and the C in the context of their potential relevance in economic and econometric research.

Volume refers to the magnitude of data, usually measured in terms of units of memory in a computer system i.e., 'bytes' (8 bits). The use of volume to measure the size of data was discussed in Diebold (2003) and is one of the earlier records of such idea. The idea is less about the size of the digital object, but more about if the digital object can be easily processed, manipulated and analysed in the context of available technologies. While modern computer can easily manage a data with size up to a couple of gigabytes, such data would not have been possible to manage, process or analysis back in 1880. In fact, the problem faced by the US Census office in 1880, could have been around 4 gigabytes, which, while not small, would not have been considered as 'big data' in today's standard.

Velocity refers to the rate at which the data is being generated as well as the speed in which it should be analysed and acted upon. Gandomi and Haider (2015) provides a detailed discussion on velocity in a more general setting. An example in econometric where velocity is important would be ultra-high frequency data typically seen in financial econometrics. Specifically, tick-by-tick market data which records the price of each financial asset when they change due to market operation i.e., selling (supply) and buying (demand). Such data provides technical challenges in terms its demand on computation as well as statistical methods of analysing such data. See Chapters 1 and 3 for further information on time series models for irregular spaced (sampled) data and their applications in financial econometrics.

It is important to note the difference between velocity and sampling frequency. Sampling frequency, to a large extend, is determined by an agent in the data value chain, they can be the consumers of the data, or the creators of the data. Velocity, however, measures the frequency in which the data is being generated and the data generation process may not be controlled by individuals, but rather, the market. Intra-daily data in stock price is determined by the market, which velocity is not constant. However, analysts can choose to sample the intra-daily data in a 5 minutes or 1 minutes frequencies.

Other examples of data with high velocity include data of credit card transactions or data generated through activities on the internet, such as search or social media data. These data tend to arrive at random (as the transaction happens) and can be useful in approximating latent variables, such as sentiment. See Section 4.5 for further discussion.

Gandomi and Haider (2015) defined *Variety* as structural heterogeneity in a dataset. Data can be structured and unstructured and volume alone is not enough to capture the time it takes to process a given dataset. For example, one gigabyte worth of tabular data would be less computationally intensive to process than one gigabyte of text or graphics. Variety helps to identify the potential complexity of the data due to its

structured and/or unstructured characteristics as well as the formats of the data e.g., graphics, text, audio, video, numerical.

4.3.2 Beyond the 3Vs

As the nature of data becomes more complex and their variety increases, characteristics of big data also becomes more complex and it is argued that the 3Vs are no longer sufficient to summarise the characteristics of (big) data. According to Gandomi and Haider (2015), IBM coined *Veracity* as the fourth V, which represents the "*unreliability inherent in some sources of data*". Examples may include construction of latent variables such as sentiment via textual data or data from probabilistic linking i.e., merging multiple de-identified datasets through probabilistic linking algorithms. See Section 4.4 for further information on data linkage. Essentially, this reflects uncertainty in the data and it is akin to *measurement error* problems in econometrics. So in a sense, one can interpret Veracity as a measure of possible measurement errors in the data.

It should also be pointed out that large proportion of econometric analysis are using data provided by external parties and the researchers have no control over their construction. So while not explicitly discussed, veracity has always been an issue in econometric analysis, and the reliability of data often relies on the reputation of the organisations that provided the data, such as official statistics agencies.

SAS also introduces *Variability* and *Complexity* as two additional characteristics of big data. Variability can be viewed as a refined version of Velocity as it does not assume data is being generated at a constant rate. Variability aims to measure the variation it the data flow rates. This is akin to measure 'acceleration' of data. The idea is to understand the dynamic of data generation which may have periodic peaks and droughts. Complexity can be interpreted as a refined version of variety, as it aims to measure the complexity of connecting data from different sources in different formats. For example, one can think of a movie as data set containing millions of picture (frames) and different audio tracks (sounds).

Gandomi and Haider (2015) accredited Oracle for introducing *Value* as the fifth V. The concept of Value is that some data may be large but doesn't necessary contain high value. An example is that data on an individual has low value even though an individual can generate large volume of data. However, if one collects such data from large number of individuals, then the data becomes high value.

The relevance of these concepts in the context of economic and econometric analytics varies and are likely to change over time as technology progresses and data availabilities changes. Volume is fundamental, whether it is being measured simply by the number of observations or the size of the digital object, it is difficult to ignore volume as a measure of big data in the context of econometric analysis.

Velocity is becoming increasingly relevant in recent times, especially in the areas of market sentiment through arrival of news, social media posts or market data from stock exchanges. Note that the arrival of these data are random, and hence their

velocity is not constant, so it seems likely that Variability, the variation in the flow rates, may also become important. Perhaps more importantly, there are implications on the statistical analysis and/or modelling using these data as the duration between the arrival of these data also contains important insight. So techniques such as duration (survival) analysis and even *point processes* may become increasingly important in econometrics.

Variety is also becoming increasingly important, especially in areas such as spatial econometrics, where spatial data from remote sensing or satellite are often combined with data typically seen in macro-economics. Another example is the introduction of Natural Language Processing (NPL), tabulated numerical data is not the only type of data that may be useful in conducting econometric analysis. The measurement of sentiment requires textual analysis and natural language processing. Other possible data source to reflect sentiment could include audio (records of radio or podcasts), social media posts, such as X (twitter) and Facebook.

In both of these cases, combining data from different sources with different formats are required for econometric analysis. These data are often unstructured (free text) or in different digital formats i.e., satellite data is in different format not directly operable with macro-economic data. Thus, skill that facilitate the linkage of data with different varieties is becoming increasingly important.

It seems that Veracity can be framed as a measurement errors problem which has a long history in econometrics. Some of the recent advances in data management, such as probabilistic data linkage, is likely to induce measurement errors problems that have not previously been considered. Section 4.5 provides further discussion on this issue. Given this, it seems reasonable to argue that not only is veracity highly relevant in econometrics, it may even lead to further development in the area of measurement errors.

It is unclear how Value may fits in the context of econometric analysis. To some extent, it highlights the importance of variations within a variable, which is elementary. The main lesson here is that size or volume alone does not guarantee variation and it is often the variation that allows statistical analysis to be powerful enough to produce insight.

4.3.3 Some Examples

It may be constructive to examine some typical economic data through the lens of the 3Vs. This section examines the data analysed in selected papers from two different areas of econometrics and provides a summary of these data from the perspective of Volume, Velocity and Variety. The main idea is to highlight the application of 3Vs and how 3Vs can help identifying the evolution of data in economic and econometric analysis.

Table 4.3: Spatial Econometrics Data and the Big Data 3Vs

Paper	Volume	Velocity	Variety
Anselin (1995)	**Moderate to High**: Scales to large point/pixel/areal datasets	**Moderate to High**: Commonly applied to repeated snapshots; inference becomes costly at high update rates	**Moderate**: Numeric spatial variables / polygons/points/rasters and adaptable to diverse geodata
Henderson, Storeygard and Weil (2012)	**Moderate**: Large spatial grids aggregated to national/regional panels (night-lights)	**Low**: Annual (or similarly low-frequency) panels in typical macro applications	**Moderate**: Raster imagery / national accounts / survey aggregates
Blumenstock, Cadamuro and On (2015)	**High**: Millions of users and potentially billions of call/SMS events; heavy aggregation needed	**High**: High-frequency event logs (near-real-time potential)	**Moderate**: Telecom metadata / geolocation / survey/ground-truth outcomes
Donaldson and Storeygard (2016)	**High**: Potentially massive global raster archives (satellite pixels) and derived products	**Moderate to High**: Periodic revisit cycles; not real-time but can be frequent depending on sensor/product	**Moderate**: Remote sensing imagery / boundaries / economic variables (multi-modal integration)
Jean et al. (2016)	**High**: High-resolution imagery; very large pixel-level inputs feeding ML feature extraction	**Moderate**: Snapshot-based or intermittently updated imagery (depends on source)	**Moderate**: Image data (high-dimensional) / survey poverty labels / geospatial joins
Glaeser, Kominers, Luca and Naik (2018)	**Moderate**: Large granular urban datasets (transactions, platforms, sensors, admin microdata)	**Moderate**: Often high-frequency (daily/weekly) depending on source	**High**: Platform data / administrative data / spatial units / digital traces

Tables 4.3 and 4.4 contain summaries of data analysed in selected papers from *Spatial Econometrics* and *Network Econometrics*. There are some interesting observations from these tables. Firstly, while the volume of economic data grows over time, the increase in volume often due to combining traditional economic data with data from other disciplines, such as data from satellite (spatial economics) or data from social media (network economics). This also leads to increase in variety, since the medium and/or (digital) format of these data are usually different to the tabulated data that are typical in econometric analysis.

Velocity of data remained moderate across spatial and network economics, but has the potential to increase in areas such as financial econometrics, where every single transaction leads to the creation of a new observation.

Table 4.4: Network Economics Data and the Big Data 3Vs

Paper	Volume	Velocity	Variety
Sacerdote (2001)	**Moderate**: Student-level administrative cohort data	**Low**: Static academic outcomes	**Moderate**: Student demographics, GPA, room-mate assignment, survey information
Duflo, Dupas and Kremer (2011)	**Moderate**: Thousands of students across schools	**Low**: Multi-year experimental panel	**Moderate**: Test scores, classroom composition, teacher incentives, treatment indicators
Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012)	**Medium** Sector-level input–output matrices	**Low** Annual macro data	**Moderate**: Industry production flows, intermediate inputs, shock propagation structure
Aral, Muchnik and Taylor (2013)	**Very High**: Millions of platform users	**High**: Real-time interaction logs	**High**: Social ties, exposure timing, behavioural adoption, interaction logs
Elliott, Golub and Jackson (2014)	**Moderate to High**: Bank exposure matrices	**Low**: Regulatory reporting frequency	**Moderate**: Interbank exposures, balance sheets, simulated contagion outcomes
Atkin and Donaldson (2015)	**High**: Firm-level transaction records	**Moderate**: Repeated shipment transactions	**Moderate**: Shipment flows, firm identifiers, prices, geographic distances
Larreguy, Marshall and Snyder Jr. (2020)	**Large**: Nationwide municipality administrative data	**Low**: Electoral cycles	**Moderate**: Audit corruption findings, election results, geographic radio coverage, municipal covariates

Lessons Learned

1. The 3Vs provides a foundational framework to examine the characteristics of big data.
2. The 3Vs also evolves and may not be sufficient as the number of sources and complexity of data increases. More characteristics are expected to be included e.g., Variability, Veracity, Value and Complexity.
3. While some of these characteristics may not be relevant to econometric analysis at present, there is a good chance they will become so.
4. The continual increase in volume and velocity lead to new challenges in econometric analysis at both technical and conceptual levels.

4.4 Data Sources, Storage, Processing, Management and Wrangling

Not only did internet technologies facilitate dissemination of data from different official and non-official sources, it also creates different avenues to collect data beyond geographical borders, as well as facilitating dissemination of data via online platforms or data repositories such as re3Data (2026), Kaggle (2026) or Google Cloud Service (2026). Given the volume and scope of data available, and the abilities to continually generate and collect them, it may be useful to discuss how data from different sources can be stored, process, managed, combined and shared. Again, the main objective is to highlight the how these developments are becoming more and more relevant in economic and econometric research over time.

4.4.1 Data Sources, Data Sharing and Data Licenses

In the context of econometric analysis, data can be broadly divided into three groups namely, publicly available data, proprietary data, usually provided by third party data service providers, and researcher generated data. As argued in Section 4.2, a single research project can consist of one or more data types. For example, publicly available macroeconomic indicators may be used in conjunction with data collected through quasi-experiments from researchers. In addition, any materials, including programming code, used to process and analyse these data are, in themselves, data generated by researchers through the project.

For each of these three groups, the data can be further divided into *structured* and *unstructured* data. Tabular data is an example of structured data that economists and econometricians are most familiar with. Data in this format are typically arranged by rows and columns with each row represents an observation and each column contains the values of a particular variable. Data from survey is another example of structured data. While not in tabular form, each questionnaire has the same structure, which can be easily transformed into tabular data. Transcripts obtained from interviews is an example of unstructured data. While each interview shares a common set of questions, the responses from each interviewee, and the subsequent follow up questions can be very different between each individual. The content of the responses is also heterogeneous. Other examples of unstructured data includes newsfeed, emails, images, video, audio recordings, and text-based documents in different formats e.g., PDF, text files.

A consequence of the rapid development in web technologies and computation is the ability to source both structured and unstructured data. This can be broadly attributed to three interrelated, but independent, developments. First, agencies that have historically shared data making their data even more accessible by leveraging new web-technologies. Some examples includes those listed in Tables 4.1 and 4.2. This category also includes data service providers, such as Bloomberg, S&P Global, Refinitiv, Nielsen, WRDS among others. Up until the early to mid 1990s, a lot of

these providers provided data through physical medium, such as CD-ROM or physical hard-disks, including magnetic tapes. The alternative was via proprietary terminals and private network servers using File Transfer Protocol (FTP) or Secured Shell (ssh). Online portal becomes popular since 2000, as web-technologies became more mature. As web-technologies continued to develop and the quality of various cyber security protocols became more mature, (mass) data sharing via Application Programming Interface (API) became more popular. Users with programming knowledge can leverage the provided APIs to search and download large volume of data in relatively short time.

At the time of writing this chapter, most platforms generally provides their APIs in Representational State Transfer (REST) format which utilises the familiar Hypertext Transfer Protocol (HTTP) technology. In addition, majority of these platforms also provide their APIs in Python, a popular programming language, especially in areas such as data science and machine learning. Other programming languages, such as R, Julia, and Java, often provide additional libraries or modules to access these platforms directly.

Second, organisations, particularly government agencies, beginning to share data with the aim to improve transparency. Platforms such as data.gov (2026) (US), data.europa.eu (2026), (EU), data.gov.uk (2026) (United Kingdom), and data.gov.au (2026) (Australia). These platforms act as single entry point to publicly available data from various government agencies within a single country, e,g., US, UK and Australia, or members from the European Union.

The development of these platforms can be seen as a result of the various open government initiatives since the 1990s. Kambil and Ginsburg (1998) provides a comprehensive analysis on the relation between open government initiatives in various jurisdictions and the development of open data portals from these jurisdictions.

In some sense, the content of data disseminated through these developments are not new. While the catalogues of data available through these organisations may have expanded over time, the nature of these data has not changed significantly. These data usually require large scale operations to collect and it is not something that an individual researcher can easily collect on their own. In other words, these data are not generated or constructed by researchers and researchers have virtually no control over how they are measured or collected.

The third development is the demand of data sharing platforms that allow the sharing of data as well as analysis. This is a result of increasing amount of data generated through research projects along with the recent emergence of data science as well as the rapid advancements in computation and web-technologies.

While economists and econometricians were not usually involved in raw data collection, the recent development in experimental economics and the popularity in conducting *random control trail* for policy evaluation and causal analysis, expedited the need for data storage and sharing platforms. The requirements from research funding bodies on the management and sharing of data generated through funded research also play a role in the development of these platforms. The development of the *American Economic Association's Registry for Randomized Controlled Trails* reflects the demand of data repository for data generated through economic research.

Table 4.5: Platforms and Data Values Chain

Stage of Data Value Chain	Function	Example Platforms
Data creation and modelling	Work with and analyse datasets	Kaggle, Datalore
Data publication	Store and publish datasets with persistent identifiers	Dataverse, Zenodo, Figshare
Repository discovery	Discover repositories hosting datasets	re3data
Standards and governance	Catalogue standards, policies, and repositories	FAIRsharing

However, data sharing platforms are diverse and each serve different purposes. Generally speaking, their purposes can be classified based on the different stages of the data value chain. Table 4.5 provides a summary on the selected set of platforms and their functions relative to the different stages of the data value chain.

As shown in Table 4.5, each of these platforms serves a distinct purpose. Some focus more on data creation and provide opportunities for research communities to collaborate and sharing their data insight, while others focus more on being a central data repositories which allow accessibility of data from different research programs or projects. There are also platforms that provide access and sharing of standards and policies on data and data governance.

Table 4.6: Selected Data Science Platforms

Platform	Primary Function	Typical Users
Kaggle	ML competitions and datasets	Data scientists
DrivenData	Social-impact competitions	Data scientists
Zindi	Regional AI competitions	ML practitioners
CodaLab	Research competitions	Academic AI researchers
Deepnote	Collaborative notebooks	Data science teams
Google Colab	Cloud notebooks	Developers and researchers
Datalore	Notebook collaboration	Data scientists
Hugging Face Hub	ML datasets and models	AI researchers
Papers with Code	Benchmarks and datasets	AI research community

These platforms reflect the merger of two ecosystems in digital data sharing that have independently emerged since 1990s, namely the development of data science as a discipline and the demand on research data infrastructure. The earlier efforts in

Table 4.7: Summary of Data Licensing Options (Open-Access Focus)

License / Model	Commercial	Modification	Attribution	Share-Alike	Fully Open	Typical Governance Objective
CC0 (Creative Commons Zero)	Yes	Yes	No	No	Yes	Maximise diffusion and reduce transaction costs
Public Domain Mark	Yes	Yes	No	No	Yes	Signal no rights reserved
CC BY	Yes	Yes	Yes	No	Yes	Enable reuse while preserving reputational credit
ODC-BY (Open Data Commons Attribution)	Yes	Yes	Yes	No	Yes	Database-specific attribution model
ODbL (Open Database License)	Yes	Yes	Yes	Yes	Yes	Enforce reciprocity in database ecosystems
CC BY-SA	Yes	Yes	Yes	Yes	Yes	Prevent enclosure of derivatives
UK Open Government Licence (OGL)	Yes	Yes	Yes	No	Yes	Public sector open data reuse
Australian Gov Open Access (CC BY 4.0 default)	Yes	Yes	Yes	No	Yes	Publicly funded data dissemination
CC BY-NC	No	Yes	Yes	No	No	Restrict commercial appropriation
CC BY-ND	Yes	No	Yes	No	No	Preserve integrity of original dataset
Data Use Agreement (Controlled Access)	Depends	Often limited	Yes	No	No	Protect confidentiality / sensitive data

digital data sharing between 1990s and 2005 focused on disciplinary data archives. These are often hosted by universities or research institutes. Some of the examples include Inter-university Consortium for Political and Social Research (ICPSR) which expanded digital distribution of social science data. ICPSR is one of the earliest examples of data sharing in social science and played and important role in the development of research data infrastructure. Another example is the *Dataverse* project which serves as the foundation for research data storage and acts as the back-end to discipline specific research data repository, e.g., *the American Economic Association's Registry for Random Controlled Trails*.

From the data science viewpoint, however, there were no major platforms before 2005 for individual research communities to share their data. Often benchmark data were shared through academic websites, FTP servers or datasets bundled with publications. These were informal distributions with no standardised meta-data template and only existed in pockets of research communities.

Between 2006 and 2013, both eco-systems expanded rapidly but in different directions. The research data infrastructure emerged and focus on data governance, discoverability and reproducibility of research. Platforms include re3data.org, DataCite, Zenodo, OpenDOAR and FAIRSharing. These platforms are results of various initiatives that were strongly driven by research funders and libraries.

Data science platforms also emerged during this period. Examples include Kaggle, Datalore and Google Big Query. Table 4.6 provides a selected list of these platforms and a summary of their functions. As shown in Table 4.6, the primary purposes of these platforms are not just about sharing data but also providing a platform for competition and sharing of data analytic techniques, which facilitate additional opportunities for training.

Regardless of their primary functions, these platforms have all contributed to the sharing of data which facilitated the continual creation of (big) data, research transparency and research reproducibility. It is worth pointing out that research data, as discussed in the beginning of this chapter, is not restricted to the actual data analysed in a research project, but also the data generated through the process of research including programming code,[7] supplementary analysis and results and other digital or physical artefacts. These platforms all play a role in sharing these resources.

A critical component in publishing and sharing research data is the ownership of data. In order to protect the creator and the custodian of data and to legitimise data as an intellectual property, various data licenses have been proposed. Naturally, data provided by commercial data service providers would no doubt have their specific terms and conditions in their subscription agreement, but licenses on open source data or data that are being shared from their creator are relatively new development. Table 4.7 provides a summary of these licenses and their intended usage.

Understanding data licenses can be important, especially in the context of data linkages. Big data is often created by linking multiple data but data service providers tend to limit the usage of their data and often restrict their data from being linked to other data. The same may also apply to open data. For further discussion on these licenses, see Carbon et al. (2019) and Yang, Kipp and Chen (2024).

[7] Platforms such as GitHub and Gitlab have also been used often in sharing programming code and other resources, such as Jupyter notebook, for various research projects. As software engineering and development are the primary focus of these platforms, they are not the main focus of this chapter.

4.4.2 Data Linkage

Linking or merging different data to form larger data is not new. This is often used to overcome missing values problem.[8] Another motivation of linking or merging data is to create a larger dataset.

For tabular data, merging can mean including more rows and/or including more columns. Econometric convention suggests that adding more columns usually means including more variables (or covariates) whereas including more rows in tabular data usually means including more observations.

Interoperability becomes important in these kinds of merging. Interoperability measures the extend in which two datasets can be combined. If one wishes to add more columns from one set of data to another, the value on each row must correspond to the same observation unit. Similarly, if one wishes to add rows from one set of data to another, the columns in each row must correspond to the same set of variables. If these are in fact the case, then joining two dataset is a relatively straightforward exercises.

In practice, however, these conditions are seldom satisfied, at least not fully. This often leads to missing values in the resulting dataset. Table 4.8 provides a simple example on a common scenario of missing values as result of merging two datasets. A common practice among empirical researchers is to remove rows with missing values. By doing so, the total number of observations may be less than the number of observations from each of the individual dataset. Table 4.4 provides an example of such case, where removing rows with missing values reduces the merged dataset to only two observations, while both Dataset A and Dataset B have four observations. Though, the merged data would allow an investigation on the relation between score and attendance and in the context of big data with high volume, the reduction in observations may still provide sufficient number of observations for purposes of modelling, notwithstanding other risks such as sample selection bias due to the removal of missing values.

An alternative is to consider more explicit missing values approach, such as the *Expectation-Maximisation* (EM) Algorithm as proposed in Dempster, Laird and Rubin (1977). However, these algorithms are typically computationally intensive and given the size of the data, such approaches may not be feasible in practice.

Data linkage as described above can be labelled as *deterministic linkage*. This is fairly common, in macro-econometrics. For example, data for official interest rates of a set of countries may come from different sources than data for gross domestic products of the same set of countries. Analysing their relations therefore requires the linkage of these data. However, as more data become available, linking economic data with data from other disciplines are also becoming popular. In fact, a number of papers as listed in 4.3 attempts to understand economic activities by linking economic variables with satellite data, e.g., Henderson et al. (2012) and Jean et al. (2016).

[8] An earlier example is Cooley and LeRoy (1981) where the value of credit card transaction was a combination of credit card transactions and credit notes from department stores to overcome the lack of credit card transaction data in the early 1950s.

Table 4.8: Example of Missing Values Created by Dataset Merging

Dataset A: Exam Scores

StudentID	Score
101	85
102	90
103	78
104	92

Dataset B: Attendance

StudentID	Attendance
101	95%
102	88%
105	90%
106	85%

Merged Dataset (Left Join)

StudentID	Score	Attendance
101	85	95%
102	90	88%
103	78	NA
104	92	NA
105	NA	90%
106	NA	85%

Recently, there is also an increase use in administrative data, most notably in Health Economics, where modelling health outcomes may involve analysing administrative data from hospitals, general practices, insurance claims, mortality registries and education registers and even administrative data from relevant authorities such as taxation office and police.

The challenge of this kind of data linkage is that these administrative data are *de-identified*, which means there is no clear common key between the different datasets to ensure the matching of each row. In other words, there is no clear mechanism to identify which row in which dataset correspond to the same individual, or observation.

In the absence of the ability to link the data deterministically, it is still possible to link these data *probabilistically* subject to the availability of sufficient information to identify the characteristics between each observation. Fellegi and Sunter (1969) provides one of the earliest methodologies on how this may be achieved. While variations and extensions to the original Fellegi-Sunter model exist, its fundamental method and idea remain to be one of the most popular to-date, especially in health economics, as pointed out by Harron et al. (2017).

Fundamentally, the Fellegi-Sunter seeks to construct two probabilities, namely the probability that the two rows from the two dataset are a match, and the probability that two given rows from two datasets are not a match based on the difference between a given (common) set of columns in both datasets. These columns contain information that are relevant to the individual such as gender, address, age. Given the probabilities, the method then considers the ratio test statistics between the two probabilities and provides a decision rule based on the ratio. The method can be formulated as

a classical hypothesis testing with the null hypothesis being the event that 'rows match' versus the alternative that the 'rows do not match'. The ratio, essentially, is the familiar ratio test statistics. Fellegi and Sunter (1969) provides two methods on how to calculate these probabilities in practice, while other variations, including the utilisation of discrete choice models, have also been proposed.

The fact that the linkage problem involves the estimation of the two probabilities imply that the resulting dataset contains *measurement errors*. That is, some rows may not belong to the same individual and thus, the values in some of the columns for that individual would be incorrect. To the best of the authors' knowledge, the impact of such measurement errors to estimation, and the subsequently inference, has not been investigated fully. Given the creation of big data often involves probabilistic linking, this would seem to be an important topic for future research.

Administrative data usually contain sensitive and private information. While the data is de-identified, the columns used for purpose of linkage may be sufficient to identify the underlying subject. One approach is to separate the data into two parts, The first part contains only the columns for linkage purposes, while the second part contains the remaining data. Researchers can still implement linkage and then passed on the result to the second group of researchers that would conduct the analysis using remaining data.

Even with such approach, privacy risk may still be higher than it can be tolerated. To alleviate privacy concern, the Bloom filter as proposed in Bloom (1970) is typically used. The filter can be used in conjunction with the traditional probabilistic data linkage techniques such as the Fellegi-Sunter method. However, Bloom filter can be computationally intensive. For further discussion on the applications of the Bloom filter see Randall, Ferrante, Boyd, Bauer and Semmens (2014) and of data privacy issues Chapter 15.

4.4.3 Data Processing and Wrangling

One of the implication from the discussion so far is that researchers can readily create their own 'big data' by linking data from different sources. As data become more available through increasing number of channels, and given the format of these data are often not consistent with each other, there is an increased demand on the researchers to have better skills in data manipulation and wrangling.

Even in the simple case where two sets of data with the common keys of countries can be non-trivial to link. This is due to the fact the presentation of country names are often not standardised. United States of America can be presented as U.S, or U.S.A. A simple join operation between the two datasets could lead to a duplication of two rows, both about the United States and each have missing values.

Other data wrangling challenges may involve understanding the meta-data of the dataset. For example, if a panel data is unbalanced, it would be useful to know how many individuals are there in each time period and whether there are attrition or if new individuals joined at some time point in the sample. These problems are often

Table 4.9: Selected Health Economics Studies Using Linked Administrative Data

Reference	Linked Data Sources	Main Question
Lleras-Muney (2005)	Census data, mortality records	Impact of education on adult mortality
Currie and Walker (2011)	Birth records, hospital admissions, environmental pollution monitoring	Impact of traffic-related pollution on infant health
Finkelstein et al. (2012)	Medicaid lottery, hospital discharge records, credit reports	Impact of Medicaid coverage on healthcare utilisation and financial outcomes
Baicker et al. (2013)	Medicaid lottery data, clinical screening records, insurance data	Effects of Medicaid expansion on health outcomes
Case and Deaton (2015)	National mortality registry, demographic population data	Trends in midlife mortality and "deaths of despair"
Chetty et al. (2016)	IRS tax records, Social Security death records	Relationship between income and life expectancy in the United States
Dobkin, Finkelstein, Kluender and Notowidigdo (2018)	Administrative health records linked to credit reports	Relationships between medical expenses and income
Miller, Johnson and Wherry (2021)	Medicaid enrollment files, hospital discharge data, mortality records	Effects of Medicaid expansion on mortality and health outcomes

manageable in the traditional two-dimensional panel data i.e., individual and time, but may become challenging to track as the dimension of the panel increases. For recent developments in the econometrics of multi-dimensional panels, see Mátyás (2024).

More generally, as data structure becomes more complex, identifying issues such as missing values, outliers and other characteristics of the data becomes more time consuming. A more systematic approach or strategy is becoming increasingly important.

The *split-apply-combine* framework proves to be extremely efficient in dealing with these sorts of problems and the framework has been implemented widely, especially in libraries focus on dataframes for virtually all open-source languages, including Python, R and Julia. The framework has its roots in the 1880s, earlier examples of the framework can be found in Walker (1870) and Hollerith (1889). In fact, the electric tabulate system proposed by Hollerith in his attempt to handle the large census data was also an attempt to automate the split-apply-combine process. Before that, the process were mainly carried out manually.

While the names may differ, the concept of split-apply-combine had been widely adopted by statisticians as a way to manage data. The concept appeared in many well-known textbooks in statistics, e.g., Yule and Kendall (1950) and Fisher (1934),

albeit the term *split-apply-combine* did not appear in those references. The first formalisation of the term is often credited to Hardley Wickham and Wickham (2011) remained an influential reference on the subject.

The core idea of *split-apply-combine* is to divide the data into well-defined groups, then *apply* the required transformation or analysis in each of these groups and finally combine the results from these groups to produce the final insight. The increase in computation power greatly facilitates the split-apply-combine paradigm. However, while the framework has been implemented widely in most of the modern software, a minimal programming effort is required to use them in practice, even with the assistance from Large Language Models (see Section 4.5 for further discussion on Large Language models). Thus, documentation of any wrangling exercises would be important for purpose of reproducibility.

4.4.4 FAIR and CARE

The story so far is that modern technologies are making it easy to create big data by combining data from different sources. The increase in computing power also allows researchers to manipulate, transform and wrangle these big data sets. The process from data collection and ingestion to a tabular form that can be readily analysed can be complex and time consuming. It also creates challenges in terms of research reproducibility, even with detailed documentation. The latter is particular important to ensure academic rigour.

The FAIR data principle is a fundamental principle aiming to facilitate research reproducibility. The technological progress described by this chapter so far have all played an important role to facilitate the FAIR data principle. The FAIR data principle stands for (F)indable, (A)ccessible, (I)nteroperable and (R)esuseable.

As the name would suggest, *finable* allows researchers to locate the source of the data. This could be the data produced by the researchers as well as the *raw* data initially collected by the researchers. However, knowing where the data is does not immediately imply its accessibility. For example, data provided by third party data service providers are often not publicly accessible. In such case, even though researchers provide information on the source of the data, it may not be accessible without additional cost from other stakeholders. This distinguishes data that are findable from being *accessible*. Ideally, research data should also be accessible and accessibility is highly influenced by the various data license as discussed in previous section.

Interoperability facilitates the linking (or combining) data. At the minimum, standardisation of naming convention and measurement of economic variables would ensure interoperability and facilitate data linkage.

Data should also be *reusable*. Note that data in modern era are usually digital objects that may be saved in different formats. Proprietary software, such as Stata, tends to store data in a proprietary format that cannot be readily used by other software. Thus, data saved by Stata may not be reusable unless other researchers also

have access to Stata, or the ability to convert Stata data format into other accessible formats. Proprietary data format may also reduces interoperability.

The FAIR data principle is relatively young and can be traced back to the debates on open data in science in the 1990s. The fact that the open data movement coincides with the development of the internet and web technologies is not a coincidence. As highlighted in Wilkinson et al. (2016), while early open data policies focused mainly on human accessibility, FAIR includes accessibility and readability for machines, which facilitates automation of data workflow and is a main motivation of the FAIR data principle.

Fundamentally, the FAIR data principle seeks to solve various data management issues due to the rise of data-intensive research in 2000s, which is a result of rapid development of the web, and other, related technologies, including improvement in digital data storage and greater computation power to process large volume of data.

The FAIR principle was first developed in the seminal work by Wilkinson et al. (2016), even though various concepts within FAIR had already existed. Since 2016, organisations such as European Commission, National Institutes of Health and Universities have progressively adopted the FAIR data principle in managing their (research) data.

Given the role of the FAIR data principle in facilitation research reproducibility, it is anticipated it will play an increasingly important role across different disciplines, including economics and finance. Indeed, it is now a common practice of journals to require authors to provide access to their data as a condition of paper acceptance. This is indicative of the increasing importance of both findable and accessible.

A related data principle is CARE, which stands for (C)ollective Benefits, (A)uthority to Control, (E)thics and (R)epsonsibility. The CARE data principle focus primarily on data collected through indigenous, or first nation, population. This is particularly important in countries where indigenous, or first nation, population is also part of a disadvantaged, or vulnerable, group. The concept of CARE is to protect such population from being exploit through massive data collection, and is particularly relevant in research that involves random control trails in developing countries.

The principle was first developed by the Global Indigenous Data Alliance (GIDA) at around 2019 in response to open data movements overlooking indigenous right and historical misuse of indigenous data. It has strong link to the Indigenous Data Sovereignty (IDS) movements. For further information about the history of IDS, Global Indigenous Data Alliance (2026b) provides a natural starting point.

The components of CARE can be summarised as follows. *Collective Benefit* stipulates that the data should generate values for the communities and not just the researchers and the institutions. There is an emphasis on shared outcomes and in the context of economics, research that leads to improvement of health or education outcomes of the indigenous population would be two examples on collective benefit.

Indigenous population should also have the rights to govern their own data and how their data should be used. The rights include access, consent, restrict use and defining governance structure. *Authority to Control* makes it obvious the ownership of the data belongs to the indigenous population.

Responsibility refers to the handling of the data from those who collect them. Researchers must act ethically and accountably. Requirements include cultural sensitivity, usage transparency and ongoing engagement with communities. This is related to the fourth component, *Ethics*, where data usage should minimise harm and maximise justice. Specifically, it should protect dignity, privacy and culture values of the indigenous population. Risk assessments should include possible misrepresentation, stigmatization and misuse of sensitive data.

For further information on CARE, see Global Indigenous Data Alliance (2026a).

Lessons Learned

1. Technologies increases the scope and volume of data collection.
2. The ability of linking data from different sources creates greater research opportunities and it allows the testing of hypotheses that were not previously possible.
3. Linking data can lead to higher chance of missing data, potential outliers and measurement errors, especially in the case of probabilistic data linkage. The management of these issues remains an active area of research.
4. FAIR data principle helps to address the challenge of research reproducibility. FAIR principle is not restricted to the raw data analysed by the researchers, but also include data, such as programming code and other meta data, generated through the research process.
5. International standard in data management for economics should be developed, similar to other disciplines, such as astronomy, to ensure interoperability of data between different sources.
6. As data collection becomes more specific and often involved indigenous population, issues around data sovereignty must be addressed. The CARE data principle may be useful in this context, especially for researchers who collect data via surveys, quasi-experiments and experiments.

4.5 Analytic Techniques for Big Data

This section focus on three main topics namely, computation on big data, high-dimensional statistics and computation on textual data. The first focus on the development of computing and the power of modern computers in processing and manipulating large volume of data. The second focus on the development of statistical techniques that identify signals from potentially noisy data. This includes machine learning techniques and lays the foundation of the more in-depth discussion in other chapters. The last topic focus on Large Language Models (LLM) and Natural Language Processing (NLP) and how these would complement the traditional statistical analysis in econometric research.

4.5.1 Computation Requirements

As presented in previous section, big data is often defined as data that are too large for existing computer to handle. One example is the computation of the familiar formula for the Ordinary Least Squares estimator, i.e., $\hat{\beta} = (X'X)^{-1} X'y$. While most researchers would take it for granted these days, there are two technical challenges here in this calculation in the context of big data. The first is the size of the matrix X and y and the second is the calculation of the inverse of the Gram matrix $X'X$ when the dimensions of X is large.

For most data analysis software, the data is usually loaded onto the *Random Access Memory* (RAM) of the computer, before further analysis can take place. So while the computer may have sufficient hard-drive storage to store the data, the computer must also have sufficient RAM in order to load the data for the software to operate. Calculating how much RAM is required to load a data set is a relatively straightforward exercise. For a $m \times n$ matrix, which entries are real number, it would require at least $m \times n \times f$ bytes of RAM where f is the number of bytes that the computer has allocated for a real number, which is typically 8 for most econometric and scientific software. E.g., a data with 1000 observations and 10 variables would take approximately $1000 \times 10 \times 8$ 80 Kilobytes (KB). At the time of this writing, a Personal Computer (PC) will typically has 16 to 32 Gigabytes (GB) of RAM.

While this seems large, it should be noted that RAM is not just about storing the data to be analysed, but they are also needed for the actual computation algorithms. Some of these algorithms would require copies of the data or copies of multiple subsets of the data to be created as part of the algorithms. So loading the data alone is not sufficient to ensure that the computer has sufficient memory to also analyse the data.

Design of a carpark would be a good analogy. Not only is the carpark must be large enough to handle the number of cars, it must also have sufficient space for cars to manoeuvre, so that they can go in and out of the carpark.

The computation of the OLS estimator provides a useful example to illustrate. While the RAM will be hosting the data matrix, the calculation of the OLS estimator also requires the Gram matrix $X'X$ to be created and, depends on the algorithm used to obtain its inverse, additional RAM is required to manipulate $X'X$ to derive its inverse.

Table 4.10 provides a historical account on the amount of memory that can be found in a typical personal computer over past decade and their maximum capabilities in calculating the inverse of a matrix in terms of the size of the matrix. As shown in the table, even though the capabilities of personal computer is increasing rapidly over the past decades, it remains a possible constraint when compare with the size of (big) data, given the calculation of inverse is often just a small part of the data analysis requirement.

In addition to space requirement, the speed of the processor (CPU or GPU) is also instrumental in obtaining objects such as inverse of a matrix. The computational cost of calculating inverse in most algorithms is $O(n^3)$. This means that if one doubles the dimension of a matrix, it would require eight times more work to obtain the

Table 4.10: RAM and Feasible Matrix Inversion (double precision)

Decade	Typical RAM	Max Matrix Size ($n \times n$)
1960s	4 KB – 64 KB	$\sim 20 \times 20$
1970s	16 KB – 256 KB	$\sim 50 \times 50$
1980s	256 KB – 1 MB	$\sim 150 \times 150$
1990s	4 MB – 64 MB	$\sim 700 \times 700$
2000s	256 MB – 2 GB	$\sim 5{,}000 \times 5{,}000$
2010s	4 GB – 32 GB	$\sim 20{,}000 \times 20{,}000$
2020s	16 GB – 128+ GB	$\sim 50{,}000 \times 50{,}000$

inverse and if the dimension is triple, then one expect twenty seven times more work. Therefore, even if there is sufficient RAM to load the data matrix, the computer may still take a long time to implement the required algorithms.

A more efficient approach is to treat the OLS estimator as solution to a system of equations i.e., $A\hat{\beta} = b$ where $A = X'X$ and $b = X'y$. This allows the possibility of obtaining $\hat{\beta}$ using elementary row operations and thus avoid the direct calculation of the inverse.

The main point is that despite the rapid development in computing technologies, computation requirements remain significant constraints for analysing available data. As data becomes more complex and statistical techniques become more sophisticated, efficient storage solution and computation algorithms will continue to be an active area of research.

4.5.2 Statistical Techniques

The presence of big data also creates significant challenges in the analysis of data beyond computational requirements. One example is the possible new source of measurement errors induced by probabilistic data linkage as mentioned earlier. Another example is the case where the number of covariates exceeds the number of observations. In traditional econometrics, asymptotic theory often focus only to the case where the sample grows indefinitely i.e., $n \to \infty$ with the number of parameters, K, remained fixed.

Perhaps more importantly, the classical setting assumes $K << n$, that is, the number of unknown parameters (or covariates) are (much) smaller than the number of observations. This can often be violated in the era of big data where tabular can be *wide* i.e., the number of columns in the data matrix is greater than the number of rows. In such case, the calculation of the OLS estimator is not possible due to the singular nature of the Gram matrix, i.e., $X'X$.

One possible solution is to identify the irrelevant columns. This is equivalent to a variable selection problem, or otherwise known as *subset selection*. In this context, this problem, i.e., variable selection, has been a focus in econometric analysis since its inception, even when $K << n$. The issue of misspecification in the context of over- and under- misspecification can be traced back to Keynes (1939) who pointed out the importance of correct specification in the context of Ordinary Least Squares and its potential bias in the presence of omitted variables.

Table 4.11 provides a brief historical timeline of techniques for variable selections in regression (see more details in Chapter 12). As shown in Table 4.11, there are two main periods in its development. Starting from stepwise regression in the early 70s to the late 70s with the proposal of various information criteria. The first formal shrinkage estimator, *Ridge* had also been proposed by Hoerl and Kennard (1970) during the same period.

For almost two decades, there was limited progress until the seminal work in Tibshirani (1996) which proposed the *Least Absolute Shrinkage and Selection Operator* (LASSO) and the associated development of Least Angular Regression (LARS) algorithms as proposed in Efron et al. (2004), which made the LASSO scalable. Again, one can understand the connection between these techniques and the rapid progress in personal computer as well as web technologies during the same period. Since then, variations of LASSO type estimators continue to emerge, most notable cases include the *Elastic Net* by Zou and Hastie (2005) and SCAD as proposed in Fan and Li (2001). Table 4.12 provides a summary on the various shrinkage estimators.

There have also been active progress in examining the statistical properties of these estimators for purpose of statistical inference. Some of the notable contributions include Knight and Fu (2000), Lockhart, Taylor, Tibshirani and Tibshirani (2014), Lee, Sun, Sun and Taylor (2016), and Fan, Shao and Zhou (2018).

However, classical asymptotic theory typically used in econometrics generally assumed that the number of parameters, K, is fixed. This is not necessary the case in the context of big data. As such, a different asymptotic result has emerged. Specifically, *Oracle Properties* as proposed in Fan and Li (2001), has become standard framework when examining properties of shrinkage estimators. The most notable difference between oracle properties and classical asymptotic result is that the former assumes sparsity in the parameter vector, that is, most elements in the parameter vector is zero. Shrinkage estimators with oracle properties are therefore able to identify which coefficients are zero and for those non-zero elements, the estimators are consistent and asymptotically normal.

There have been some active discussions on the modes of convergences of these results in the context of their practical usefulness in econometric analysis. Specifically, these results are typically proved under pointwise convergence whereas classical asymptotic results can usually be obtained under uniform convergence. For further discussions, see Leeb and Pötscher (2005) and Leeb and Pötscher (2008).

Apart from statistical theory, it is also worth noting that variable selection via shrinkage type estimators are mostly data driven without much economic insight.

Table 4.11: Historical development of variable selection and shrinkage methods in regression

Year	Development	Key contribution	Historical significance
1960s–1970s	Stepwise and subset selection	Forward/backward/stepwise procedures for regression specification	Variable selection treated as a discrete model choice problem; computationally intensive and unstable
1970	Ridge regression	Introduction of L_2 shrinkage (Hoerl & Kennard, 1970)	First formal shrinkage estimator; stabilizes multicollinearity but does not produce sparsity
1974	AIC	Information-theoretic model selection criterion (Akaike, 1974)	Establishes "fit + penalty" framework for model selection
1978	BIC	Bayesian-motivated model selection criterion (Schwarz, 1978)	Stronger penalty for complexity; widely used for parsimonious models
1996	LASSO	L_1-penalized regression with sparse solutions (Tibshirani, 1996)	Converts variable selection into convex optimization; enables sparse estimation
2001	SCAD	Nonconcave penalty reducing shrinkage bias (Fan & Li, 2001)	Introduces nonconvex penalties with oracle properties
2004	LARS	Efficient algorithm for LASSO solution paths (Efron, Hastie, Johnstone & Tibshirani, 2004)	Makes LASSO computationally scalable
2005	Elastic net	Combination of L_1 and L_2 penalties (Zou & Hastie, 2005)	Handles correlated predictors and high-dimensional settings
2006	Adaptive LASSO	Data-dependent penalty weights (Zou, 2006)	Improves variable selection consistency and theoretical guarantees
2010	Coordinate descent / glmnet	Fast algorithms for penalized GLMs (Friedman, Hastie & Tibshirani, 2010)	Enables routine large-scale application of regularization methods
2010s–present	Post-selection inference	Valid inference after model selection (Belloni, Chernozhukov & Hansen, 2014b; Chernozhukov, Hansen & Spindler, 2015)	Integrates LASSO into causal econometrics and high-dimensional inference

Table 4.12: Comparison of variable selection and shrinkage methods

Method	Penalty	Sparsity	Strengths	Weaknesses
Subset selection	L_0 (implicit)	Yes	Direct interpretability; classical approach; aligns with hypothesis testing	Computationally expensive; unstable; ignores model uncertainty
Ridge regression (Hoerl & Kennard, 1970)	L_2	No	Handles multicollinearity; reduces variance; stable estimates	No variable selection; all variables retained
LASSO (Tibshirani, 1996)	L_1	Yes	Sparse solutions; performs selection and estimation simultaneously; convex optimization	Biased for large coefficients; unstable with correlated predictors
Elastic net (Zou & Hastie, 2005)	$L_1 + L_2$	Yes	Handles correlated predictors; grouping effect; suitable for $K > n$	Requires tuning of two parameters; less interpretable than pure LASSO
Adaptive LASSO (Zou, 2006)	Weighted L_1	Yes	Oracle properties; improved variable selection consistency	Requires initial estimator; more complex implementation
SCAD (Fan & Li, 2001)	Non-convex	Yes	Reduces shrinkage bias; retains sparsity; strong theoretical properties	Nonconvex optimization; computationally harder; multiple local minima

The motivation of these techniques are also usually driven by prediction accuracy rather than identifying meaningful economic relations between variables.

From the perspective of building meaningful or interpretable economic model, there was a parallel development within econometrics during the same period i.e., 90s to early 2000s. It is worth noting that model specification has a long history in econometrics and misspecification analysis is a major area of research, see for example White (1996) and the comprehensive list of references within. There was an attempt to develop an automated model building procedures within econometrics combining structural information (economic theory) with a data driven approach. Most notable contributions are Hendry and Krolzig (2001), Krolzig and Hendry (2001) and Doornik (2009). See also Hendry (2024) for a reflection on these developments. Tables 4.13 and 4.14 provide an overview on these contributions and a summary of their limitations as well as their differences to the more data driven approaches, such as shrinkage estimators. Readers are referred to Chapter 12 for an in-depth discussion of these developments.

In general, shrinkage estimators seem to be more useful for the case when the number of variables are close to or exceeds the number of observations. Belloni et al. (2014b) and Chernozhukov, Newey and Singh (2022) also point out that these techniques seem to be particularly useful in addressing the weak instruments problem, as it can be used to generate optimal instrumental variables by combining many, possibly weak, instruments. See also Chan and Mátyás (2022) for further discussion

Table 4.13: Automated model building vs shrinkage: mechanisms and objectives

Approach	Core mechanism	Main objective
General-to-specific (GETS) / LSE approach (Hendry, 2024; Campos, Ericsson & Hendry, 2005)	Sequential reduction from a general unrestricted model using hypothesis tests and diagnostic checking	Empirical congruence, parsimony, and interpretable specification
PcGets (Hendry & Krolzig, 2001; Krolzig & Hendry, 2001)	Computer automation of GETS search over multiple reduction paths	Systematic automatic model selection subject to diagnostic constraints
Autometrics (Doornik, 2009; Castle, Doornik & Hendry, 2011)	Multi-path tree search with block tests, indicator saturation options, and diagnostic control	Robust automatic selection in large dynamic econometric models
Subset selection / stepwise regression (Akaike, 1974; Schwarz, 1978)	Sequential inclusion/exclusion or all-subsets search, often guided by tests or information criteria	Parsimonious specification
Ridge regression (Hoerl & Kennard, 1970)	L_2 penalization	Prediction and stabilization under multicollinearity
LASSO (Tibshirani, 1996; Efron et al., 2004)	L_1 penalization yielding sparse solutions	Variable screening and prediction in potentially high-dimensional settings
Elastic net (Zou & Hastie, 2005; Friedman et al., 2010)	Combined L_1 and L_2 penalization	Sparse prediction with correlated regressors
Post-LASSO / double-selection (Belloni et al., 2014b; Chernozhukov et al., 2015)	Use LASSO for nuisance selection, then estimate target parameters with post-selection inference procedures	Valid inference on low-dimensional causal parameters in high-dimensional settings

and readers may also find Chan, Harris, Singh and Yeo (2022) a useful overview on shrinkage techniques for non-linear models in conjunction with Chapter 10.

The impact of increasing computation power in the late 1990s also awaken the development of Artificial Neural Network (ANN), which was popular in 1970s to early 1980s. However, it was discovered in the early 1980 that for ANN to be truly useful, multi-layers are required which was not feasible at the time, given the limitation of computing power and the complexity in training the network. The introduction of back propagation and the increase in computation power in the 1990s had alleviated these roadblocks. However, since these developments were mostly motivated by computation rather than big data, readers are referred to Chapter 12 for further details.

This discussion above highlighted the impacts of big data from one specific aspect namely, the case when the number of parameters exceeds same size $K >> n$, and as discussed above, this has significant impacts on the development of relevant asymptotic theory. This case highlights two very important aspects when considering the impacts of big data on econometrics analysis. Firstly, as discussed in previous

Table 4.14: Automated model building vs shrinkage: strengths and limitations

Approach	Strengths	Main limitations
General-to-specific (GETS) / LSE approach (Hendry, 2024; Campos et al., 2005)	Strong emphasis on dynamics, parameter constancy, encompassing, and diagnostic adequacy; well suited to time-series econometrics	Path dependence; repeated testing complicates inference; can become cumbersome in very high-dimensional settings
PcGets (Hendry & Krolzig, 2001; Krolzig & Hendry, 2001)	Operationalizes Hendry-style model discovery; transparent reduction logic; tailored to econometric diagnostics	Still fundamentally test-based and discrete; less natural when $K \gg n$
Autometrics (Doornik, 2009; Castle et al., 2011)	More scalable and flexible than early GETS implementations; strong performance in simulations; handles breaks/outliers better than older stepwise procedures	More complex workflow; still not a substitute for substantive identification
Subset selection / stepwise regression (Akaike, 1974; Schwarz, 1978)	Simple and historically influential; intuitive inclusion/exclusion interpretation	Unstable; model uncertainty ignored; computationally expensive in large model spaces
Ridge regression (Hoerl & Kennard, 1970)	Reduces variance; handles correlated regressors well; computationally convenient	Does not produce exact sparsity; less useful when explicit variable selection is desired
LASSO (Tibshirani, 1996; Efron et al., 2004)	Joint estimation and selection; convex optimization; scalable with modern algorithms	Shrinkage bias; may select only one variable from a correlated group; classical diagnostics are not built in
Elastic net (Zou & Hastie, 2005; Friedman et al., 2010)	Better than LASSO when predictors are highly correlated; suitable for $K > n$	Requires tuning of multiple hyperparameters; inference remains nontrivial
Post-LASSO / double-selection (Lockhart et al., 2014; Belloni et al., 2014b; Chernozhukov et al., 2015)	Bridges machine learning and econometrics; useful for many-controls and many-instruments problems	Depends on high-dimensional assumptions; still requires careful research design and identification logic

section, measurement of big data is multi-dimensional, and it is not just about large sample size in the sense of large number of observations, large n.

Secondly, as a consequence of big data being a multi-dimensional object, asymptotic theory becomes more relevant. This is in contrast to the initiation that big data may lead to full information in the sense of obtaining data from the population rather than from a (random) sample.

There are two aspects to this. First, for most econometric applications, time almost always plays an important role and by definition, population data cannot be obtained from the perspective of the time dimension, given future cannot be observed. Second, finite sample econometrics becomes computationally expensive and intractable in

the context of big data, especially when the sample size is large i.e., large n, and asymptotic theory appears to be the only feasible results for inference purposes.

Another important point is that the generation of big data is usually not consistent with any form of sampling design. This means sample selection bias is likely to be prominent with big data. Table 4.15 provides a summary of selected papers that discuss this specific issue.

Table 4.15: Big data and its implications for statistical analysis

Reference	**Main issue**	**Implication for "sample from population"**	**Implication for asymptotic theory**
Fan, Han and Liu (2014)	Noise accumulation, spurious correlation, heterogeneity, incidental endogeneity	Observed data are often by-products of complex systems rather than random draws from a well-defined target population	Requires high-dimensional asymptotics, sparsity assumptions, and theory robust to endogeneity and model-selection error
Varian (2014)	Big data in empirical economics and predictive modelling	Administrative, transactional, and digital-trace data often cover broad behavior but not necessarily the target population of interest	Encourages resampling, cross-validation, machine learning, and computationally feasible approximations alongside classical asymptotics
Donoho (2017)	Shift from classical statistics toward data science	Population-sample logic remains important, but data provenance, cleaning, linkage, and algorithmic generation become part of inference	Asymptotic theory remains useful, but must coexist with computational constraints, prediction goals, and nontraditional data structures
Meng (2018)	Big data paradox; data quality versus quantity	Large datasets are often nonprobability or self-selected samples, so representativeness can deteriorate even when n is huge	Classical LLN/CLT intuition can be misleading: as n grows, bias need not vanish and can dominate shrinking variance
Bradley et al. (2021)	Empirical demonstration of big data paradox in surveys	A massive survey can be less informative than a much smaller probability-based sample if selection bias is strong	Asymptotic precision around the wrong estimand is possible; uncertainty quantification must account for data defect correlation and bias

In addition to extending classical asymptotic theory as discussed above, big data also has significant impacts on inferential framework. Table 4.16 provides a summary of selected papers on how big data impacts on inferential frameworks.

Table 4.16: How big data changes econometric and inferential frameworks

Reference	Framework	What changes under big data	Asymptotic response
van de Geer, Bühlmann, Ritov and Dezeure (2014)	Inference in high-dimensional models	Classical confidence intervals do not survive penalization bias and $K >> n$ settings	Debiasing/desparsifying methods recover asymptotically valid confidence regions and tests for low-dimensional components
Belloni, Chernozhukov and Hansen (2014a)	High-dimensional structural and treatment-effect models	Number of controls can be comparable to or exceed sample size; naive post-selection inference fails	Use approximate sparsity, regularization, and orthogonal estimating equations to recover valid inference on low-dimensional targets
Belloni et al. (2014b)	Treatment effects after selection among many controls	The analyst no longer starts from a small prespecified model; variable selection becomes part of estimation	Post-double-selection yields uniformly valid inference under high-dimensional asymptotics
Belloni, Chernozhukov, Fernández-Val and Hansen (2017)	Program evaluation and causal inference with high-dimensional data	Rich covariate sets, heterogeneous treatment effects, and nuisance functions complicate identification and estimation	Orthogonal or doubly robust moments allow root-n inference for target parameters despite slower nuisance estimation
Chernozhukov et al. (2018)	Double/debiased machine learning	Machine learning estimators of nuisance functions are biased and often nonregular from a classical viewpoint	Cross-fitting and Neyman orthogonality restore asymptotic normality for structural parameters

4.5.3 Natural Language Processing and Large Language Models

The recent development of Large Language Models (LLM) have blurred the line between quantitive and qualitative analysis. The ability to summarise and analyse a large amount of textual data, with the ability to quantify some of the information from the textual inputs have greatly facilitated the generation of big data. The idea of mapping textual data to numerical values with the aim to analyse them in conjunction with other quantitative data using mathematical and statistical techniques is not new, and is a key objective of *Natural Language Processing*.

Jurafsky and Martin (2009) defines Natural Language Processing (NLP) as *" NLP is the field concerned with enabling computers to process, understand, and generate human language"* and LLM can be considered as a specific class of models within NLP with the development of the *Transformer architecture* in 2017 as the pinnacle for its practical use and widespread popularity, including services such as ChatGPT, Copilot and others. The history of NLP and LLM is a subject in its

own right and detailed account of their developments is beyond the scope of this chapter. Nevertheless, Table 4.18 provides a summary of the key milestones in the development of NLP with a focus on its trajectory to the development of LLM.

Even before transformers and LLM, NLP has a long history in financial research, especially in the area of sentiment analysis. Table 4.17 provides a brief chronological summary of selected contributions on sentiment analysis using NLP.

Table 4.17: Early Development of Sentiment Analysis in Natural Language Processing

Period	Key Contribution	Representative Work	Methodology
1960s–1980s	Early content analysis and lexical approaches to evaluating tone	Stone, Dunphy, Smith and Ogilvie (1966)	Dictionary-based (lexicon counting)
1970s	Foundations of term weighting and text representation	Sparck Jones (1972)	Statistical text analysis (IDF)
Late 1990s	Emergence of subjectivity detection (distinguishing opinion from fact)	Riloff and Wiebe (1999)	Rule-based + supervised learning
Early 2000s	First machine learning approaches to sentiment classification	Pang, Lee and Vaithyanathan (2002)	Supervised learning (Naive Bayes, SVM, MaxEnt)
Early 2000s	Unsupervised semantic orientation using word associations	Turney (2002)	Unsupervised learning (PMI-based)
Mid 2000s	Refinement of sentiment lexicons and contextual polarity	Wilson, Wiebe and Hoffmann (2005)	Hybrid (lexicon + machine learning)

As indicated in Table 4.17, the application of NLP in sentiment analysis can be traced back to 1960s with diverse range of techniques. NLP plays a key role in generating numerical data from text to be analysed along with other quantitative data. Lucchetti and Cajueiro (2026) provides a comprehensive introduction to the basic concepts in NLP as well as a survey on its current applications in economics and finance. For more financial specific applications, readers may also find Du, Zhao, Mao, Xing and Cambria (2025) useful.

Perhaps one of the greatest breakthroughs in NLP recently is the development of the *transformers* architecture and its role in developing Large Language Models. The basic idea is to break a piece of text into a set of *tokens*, each token represents a specific letter, word, or phases with an unique identification number. Each of this token is also associated with an unique vector. This process is known as *tokenisation*, which is essentially a process of turning a piece of text into a sequence of vectors. Thus, predicting the next word is the familiar estimation exercise of estimating $\mathbb{E}(x_n|X_n)$ where x_n is the $p \times 1$ vector for the n^{th} token and X_n is the sequence

Table 4.18: Historical Development of Natural Language Processing and the Emergence of Large Language Models

Era	Core Methods	Key Innovation	Limitation	Relevance to LLMs
Rule-based 1950s–1980s	Symbolic rules, grammars	Linguistic structure	Poor scalability	Conceptual foundation
Statistical NLP 1990s–2000s	n-grams, HMMs	Probabilistic modelling	Limited context	Precursor to language modelling
Early neural NLP 2000s–2010s	Feedforward NN, embeddings	Continuous representations	Limited context	Basis for representation learning
Sequence models 2010s	RNN, LSTM, GRU	Sequential modelling	Long-range dependency issues	Transitional stage
Transformers 2017–	Self-attention models	Global context modelling	Data-intensive	Core LLM architecture
Pretrained models 2018–2020	BERT, GPT	Transfer learning paradigm	Task-specific fine-tuning	Direct precursor to LLMs
LLM era 2020–	Large-scale transformers	Scaling laws	High computational cost, opacity	Current dominant paradigm
Alignment era 2022–	RLHF, instruction tuning	Human-aligned outputs	Alignment complexity	Enables practical deployment

of previous vectors (tokens). While the prediction is achieved by training various extension of artificial neural networks, conceptually, LLMs, such as GhatGPT, is akin to a multivariate forecasting exercise common in the econometrics literature. In machine learning and artificial intelligence literature, training is akin to estimating an econometric model. In the context of LLMs, the number of parameters is large, often at the magnitude of billions at the time of writing this chapter.

Table 4.19 provides a brief overview of the major developments of tokenisation, transformers architecture and large language models, while Table 4.20 provides a summary on the size of the data used to train some of the more popular foundational models in LLM. There are two interesting observations from these tables. First, it re-iterates, once again, the importance of the technological progression in 1990's, in terms of the development of web technologies and the reduced cost of digitising data. The latter leads to improved interoperability between data which allows the curation of mass data to be used in training LLMs. Second, the size of these models in terms of the number of parameters is indicative of the amount of data required to train these models. Note that the relation between token size and the amount of data used is not obvious because the tokenisation scheme of each model is likely to be different.

Table 4.20 also reveals that these models cannot be estimated without significant infrastructure investment as well as ongoing supply of electricity. Most of these models are expressed in terms of different types of neural network models with

Table 4.19: Brief Timeline of Major Developments in Tokenisation, Transformers, and Large Language Models

Publication	Area	Major development	Why it matters
Gage (1994)	Tokenisation	Byte Pair Encoding (BPE) introduced as a compression algorithm	Later adapted for subword tokenisation in NLP
Sennrich, Haddow and Birch (2016)	Tokenisation	BPE adapted for neural NLP/subword segmentation	Helped handle rare and out-of-vocabulary words in neural machine translation
Vaswani et al. (2017)	Transformers	Transformer architecture introduced	Replaced recurrence with self-attention and enabled large-scale parallel training
Kudo and Richardson (2018)	Tokenisation	SentencePiece popularised language-independent subword tokenisation from raw text	Important for multilingual and end-to-end pipelines
Radford, Narasimhan, Salimans and Sutskever (2018)	LLMs	GPT showed generative pretraining plus task-specific adaptation	Early modern pretrained transformer LM for transfer learning
Devlin, Chang, Lee and Toutanova (2019)	LLMs	BERT established bidirectional transformer pretraining	Made pretrained encoder models central in NLP
Radford et al. (2019)	LLMs	GPT-2 highlighted zero-shot/multitask behaviour from scale	Helped shift attention toward scaling autoregressive LMs
Kaplan et al. (2020)	LLMs	Scaling laws formalised how performance changes with parameters, data, and compute	Gave practical guidance for building larger LMs
Brown et al. (2020)	LLMs	GPT-3 demonstrated strong few-shot learning at 175B parameters	Marked the mainstream arrival of the large language model era
Ouyang et al. (2022)	LLMs	RLHF/instruction tuning became a central alignment paradigm	Improved helpfulness and instruction-following
Touvron et al. (2023)	LLMs	LLaMA accelerated open-weight foundation-model research	Lowered barriers for academic and open-model experimentation

Table 4.20: Evolution of Large Language Models: Parameters and Training Data Size

Year	Model (Version)	Parameters	Training Data (Tokens)[a]	Approx Size[b]
2018	BERT (base/large)	110M / 340M	3.3B	~13 GB
2019	GPT-2	1.5B	40B	~160 GB
2019	XLNet	340M	32B	~128 GB
2020	GPT-3	175B	300B	~1.2 TB
2021	Switch Transformer	1.6T	1T	~4 TB
2021	Megatron-Turing NLG	530B	270B	~1.08 TB
2022	PaLM	540B	780B	~3.1 TB
2022	OPT	175B	180B	~720 GB
2023	LLaMA	7B–65B	1.4T	~5.6 TB
2023	Falcon	7B / 40B	1T	~4 TB
2023	Mistral	7B	1T	~4 TB
2024	LLaMA 3	8B / 70B	15T	~60 TB
2025	LLaMA 3.1	8B / 70B / 405B	15T+	~60 TB+

[a] B and T denote Billions and Trillions, respectively.
[b] GB and TB denote Gigabytes and Terabytes, respectively.

sophisticated, and usually complex layers and nodes. The computation cost of estimating these models given the number of parameters and their complexity exceeds the capability of most accessible computing equipment of a typical researcher. This includes the curation, management and tokenisation of the training data before training (estimation), which in itself can be a great computation challenge.

Given the complexity of these models and the amount of data required, the quality of training data would therefore be an important consideration. The risk of *Garbage-in-Garbage-out* is particular significant in training these models. This can create subtle but important differences between disciplines on the usefulness of LLMs at different parts of the research workflow. One example is the construction of systematic review and literature review. LLM powered services, such as Elicit, provides extremely efficient way to curate existing papers from the academic literature on a particular research questions. The accuracy and the coverage of the results from these services, however, are restricted to the sources of training data in general. Table 4.21 reproduces the table in Raschka (2025) which revealed the composition of pre-training data for GPT 3, which formed the base model for the first version of ChatGPT.

The main lesson is that publicly available LLMs are being trained on publicly available data, which could be restrictive given the large proportion of academic

Table 4.21: Composition of LLM Training Data

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

research in economics are behind paywall. See Shu and Larivière (2024) and Larivière, Haustein and Mongeon (2015) on discussion on the oligopolistic nature of academic publications as well as the progress of the open access movements. The implication is that using LLMs for research purposes may be more accurate and effective in some disciplines over the others. The accuracy and effectiveness are likely to be correlated with the amount of publications that are publicly available or at the very least, accessible to the specific LLMs. Larivière et al. (2015) revealed that 80% of publications in social science, including economics, are behind paywall, so using LLMs to identify relevant literature in economics may not be as effective as other areas, such as Mathematics, where more papers are available in public domain.

The situation is improving in recent times, given the popularity of services such as SSRN and arXiv, where high quality working papers can be readily accessible publicly. However, the quality of working papers on these platforms vary. Given the rapid development of LLMs and its applications, reproducibility of literature search can be difficult to guaranteed. Therefore, it would be advisable to utilise LLMs, and LLMs powered service, as a starting point and verify the results using the more traditional search paradigms, to ensure reproducibility of results.

In additions to generating numerical data from text and providing efficient method of curating materials from existing literature, LLMs also able to create a new source of data. LLM combined with computer vision allows extraction of data from plots and graphs. To demonstrate, Figure 4.1 contains a bar graph presenting 10 values i.e., $x1$ to $x10$, each drawn from a normal random variate. The graph had been uploaded to ChatGPT with the following prompt:

"Extract the numeric values of each bar based on their height relative to the y-axis and organise the data in a table."

Table 4.22 contains the values extracted from ChatGPT along with the exact values. As shown in the table, while extraction is not likely to be precise, it is often a closed approximation to the original data. Again, data obtained this way can be formulated as a measurement error problem, in which there exists a rich literature in econometrics

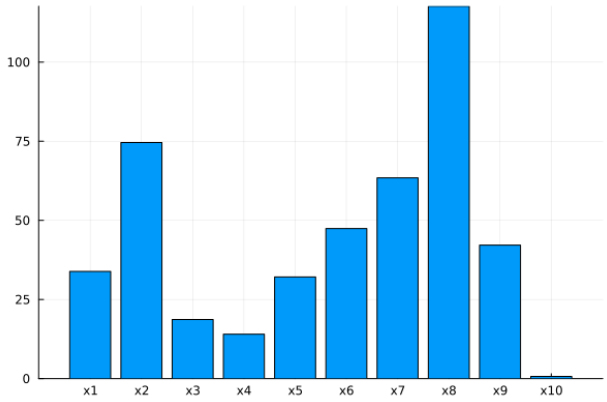

Fig. 4.1: Bar Graph of Simulated Data of Ten Variables

Table 4.22: Data Extracted by ChatGPT with Graphical Input

Variable	Extracted value	Exact value
x1	29	29.640
x2	75	74.275
x3	68	68.851
x4	50	50.535
x5	23	23.466
x6	4	4.235
x7	20	19.996
x8	44	44.429
x9	44	43.970
x10	14	13.643

for problems like this. The main point here, however, is that LLM provides yet another avenue to efficiently curate data from different sources, albeit with some possible errors. However, the degree of such measurement errors can be bootstrapped in principle, since it is possible to simulate graphs and request LLMs to extract their data repeatedly.[9]

Another contribution of LLMs to econometric research is programming code generation. This is a particular strength of LLMs, especially for open-source languages such as Python, R and Julia. Proprietary languages can also benefit provided they have active communities that are willing to share their experience openly, e.g., Stata. Given LLMs has great capability to translate between languages, Stata code can in

[9] This is assuming the service providers have not already provided information on accuracy.

principle be translated to its equivalence in Python with proper support of external libraries and modules. This is a good example where closed-source software does not immediately imply in-accessibility, and community practice and platforms, such as Kaggle, can potentially help to alleviate some of these issues.

It is worthwhile noting two additional approaches that can alleviate discipline specific heterogeneity that cannot be realistically be achieved when training a generic foundational model. Specifically, Retrieval Augmented Generation (RAG) and Retrieval Integrated Generation (RIG). RAG allows user to provide additional data to be used in conjunction with a foundation model. In other words, it provides additional materials for the LLMs to extract information and provides each prompt with a more specific context. In RAG, no additional training is required. On contrary, RIG allows users to change some of the parameters in the foundational model by providing additional data for training purposes. This leads to subject specific LLMs which are more suited in the targeted areas, but it requires intensive computation for training purposes. In practice, both approaches are often implemented in conjunction with each other.

Implicitly or explicitly, RAG or, RAG-like, approaches have been implemented in practice by researchers in economics and econometrics. Table 4.23 provides a summary of selected papers that focus on the applications of RAG or RAG-like approach in economics and financial research as well as their potential advantages. Interestingly, the concept of information retrieval for better economics decision making can be traced back to late 1990s by Varian (1999), before the technical proposal of RAG, which did not seem to appear until the late 2010s and early 2020s, see fore examples, Lewis, Perez, Piktus et al. (2020) and Guu, Lee, Tung et al. (2020).

At the time of writing this chapter, technologies like OpenClaw is becoming popular. OpenClaw is not a LLM but rather, it is a framework that allows creation of different *agents*. These agents can be accessed through LLMs and have the ability to interact with the operating system and other software of the computer. They also have the ability to access web-technologies, such as search engines. In other words, it gives the LLM the ability to complete specific tasks by providing LLM access to use the computer. One common use case is to instruct an agent to search for academic papers that match a specific research question through Google Scholar, while asking another agent to summarise the downloaded papers and turn the summaries into a report. It is also possible to ask agents to access data through the online data portals from the various agencies that mentioned earlier in the chpater, while another agent developing programming code to carry out specific set of econometric analysis.

The impacts of these technologies on economic and econometric research are to be fully realized but it appears to be an emerging area of research, see for examples Korinek (2023b), Korinek (2023a), and Dawid, Harting, Wang, Wang and Yi (2025).

These technologies will fundamentally change the workflow of econometric analysis. When use correctly, it seems to allow researchers to focus much more on the idea and the logic of the research questions rather than being distracted by technical hurdles such as coding error and network authentication issues when accessing data. However, the convenience, and the relatively low (time) cost, of these technologies may also encourage researchers to conduct analysis blindly without

Table 4.23: RAG/RIG and Related Literature in Economics, Econometrics, and Finance

Reference	Area	Relevance to RAG/RIG and Research Workflows
Varian (1999)	Economics	Conceptual foundation for information retrieval and attention in economic decision-making.
Tetlock (2007)	Finance	Early example of using retrieved news data for market prediction.
Loughran and McDonald (2016)	Finance	Survey of textual analysis in finance; provides pre-RAG retrieval-based workflow foundation.
Lewis et al. (2020)	Machine Learning	Foundational RAG framework combining retrieval with generation; improves factual accuracy and knowledge grounding.
Guu et al. (2020)	Machine Learning	Introduces retrieval-augmented pre-training; integrates external knowledge into model reasoning.
Izacard and Grave (2023)	Machine Learning	Demonstrates strong performance of retrieval-augmented models in knowledge-intensive tasks.
Korinek (2023b)	Economics	Discusses LLMs in economic research; highlights literature review automation and need for grounding.
Agrawal, Gans and Goldfarb (2023)	Economics	Explores how generative AI reduces search costs and transforms knowledge production.
Lopez-Lira and Tang (2023)	Finance	Uses textual data for stock prediction; illustrates importance of external information sources.
Li, Wang, Ding and Chen (2023)	Finance	Survey of LLM applications in finance; emphasises retrieval grounding for reliability.
Wu et al. (2023)	Finance/NLP	Evaluates ChatGPT/GPT-4 on financial text analytics tasks.

deeper consideration. This seems to be an issue that perhaps warrant the greatest attention, given all the related ethical and legal issues in terms of intellectual property right, data governance and data privacy.

Lessons Learned:

1. The development of NLP and LLM re-iterated that the challenge arise from data size exceeds the capability of computing technologies is not a new phenomenon.
2. Big data can create new theoretical challenges in statistics and econometrics that have not been considered (deeply) before, including increasing proportion of missing values, different sources of measurement errors, sample selection bias and the needs of new types of asymptotic theory.

3. Development of asymptotic theory in the context of sample selection bias is becoming more relevant in the era of Big Data.
4. NLP and LLM provide another channel in converting textual and graphical data to numerical data and therefore provide new avenues in the curation of big data, albeit with measurement errors in some cases. This also provides a nexus between qualitative and quantitative research.
5. Econometrics is becoming more interdisciplinary in the sense that future econometricians may need more foundation in computer science and information technology, in addition to statistics and mathematics.

4.6 Concluding Remarks

The chapter provided an overview of big data in the context of academic research in economics and econometrics. It began with the definition of data in the context of research and argued that data should not be restricted to the data that were being analysed, but any materials generated through the research life-cycle, especially the materials that are instrumental to ensure research reproducibility, should fall under the umbrella of research data.

The chapter then reviewed the origin and early motivation of economic data collection. The role of technologies was emphasised and it identified that the 1990s was perhaps the most important decade in the context of big data and its applications. During that decade, several foundational technologies emerged, which facilitated the collection and dissemination of data. These include web technologies and the reduced cost of storing digital assets relative to physical storage of data.

In addition to facilitating the collection and discovery of available data, these technologies also facilitate the creation of big data by curating existing data from different sources. The advancement in data linkage allows different data set to be combined in a meaningful way which lead to more insightful and powerful analysis to be conducted. Big data, however, also amplified the issues of measurement errors. Such errors can occur from the initial ingestion to the data linkage process. The recent advances in Large Language Models and computer vision, also allow further data creation and recovery from multi-modal medium. Again, extract of numerical data from these approaches would inevitably contain errors which can be modelled as a measurement error problem.

Given the increased complexity of creation and curation of the (big) data, framework to ensure reproducibility in terms of data construction and research results becomes increasingly important. The FAIR and CARE data principles provide useful starting points to ensure transparency of research and should be embedded as standard research practice.

Framework to describe and to measure big data had also been discussed. The 3Vs framework is useful to describe big data and to widen the understanding on the generation of (big) data to include the speed in which the data being generated as well as the multi-modal nature of data, beyond tabulated numerical data or unstructured

textual data. Graphics, video and audio files are increasingly common in research across different disciplines.

To consolidate multi-modal data, Large Language Models is becoming instrumental. The implications of LLMs in research workflow are yet to be fully understood and appears to be an emerging areas of research. Issues such as ethical applications of LLM,s data privacy and intellectual property rights will undoubtedly be part of this important conversation.

These developments also have impacts on the training of next generation of econometricians. While application of econometrics will become more accessible, the development of econometric techniques will become more technical and will involve a multi-disciplinary approach across computer science, data management, economics and statistics.

In the context of training, perhaps the main question is the identification on the types of knowledge that a future econometrician must know and what types of knowledge one can reasonably rely on technologies, such as Large Language Models or General Artificial Intelligence (GAI). The usage of GAI in assisting the continuous education of future econometricians would also be useful in allowing researchers to up-to-date with the rapid development in research technologies.

Perhaps one can take inspiration from the following quote given the uncertainty around AI.

> " *Honesty compels us to point out that in the 20-year history of* ***Numerical Recipes****, we have never been correct in our predictions about the future of programming languages for scientific programming,* ***not once****! ....*
> *With this edition, we are no longer trying to predict the future of programming languages. Rather, we want a serviceable way of communicating ideas about scientific programming.*"
>
> Press, Teukolsky, Vetterling and Flannery (2007)

With the right planning and implementation, it would appear that LLM and GAI based on the big data can allow researchers to focus on expressing their ideas and produce reproducible results rather than being distracted by technical hurdles, such as finding the right functions from a software documentations or spending hours in configuring LaTeX tables.

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. & Tahbaz-Salehi, A. (2012). The Network Origins of Aggregate Fluctuations. *Econometrica*, *80*(5), 1977–2016.

Agrawal, A., Gans, J. & Goldfarb, A. (2023). The Turing Transformation: Artificial Intelligence, Intelligence Augmentation , and Skill Premiums. *NBER Working Paper*(w31767).

Akaike, H. (1974). A New Look at the Statistical Model Identification. In *Selected Papers of Hirotugu Akaike* (pp. 215–222). New York: Springer. doi: 10.1007/978-1-4612-1694-0_16

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115.

Aral, S., Muchnik, L. & Taylor, S. J. (2013). Engineering Social Contagions: Optimal Network Seeding in the Presence of Homophily. *Network Science*, *1*(2), 125-154.

Atkin, D. & Donaldson, D. (2015). *Who's Getting Globalized? The Size and Implications of Intra-national Trade Costs* (Tech. Rep. No. w21439). Retrieved from https://EconPapers.repec.org/RePEc:nbr:nberwo:21439

Australian Bureau of Statistics. (2026). *Data.* https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/data. (Accessed: 23-01-2026)

Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J., Newhouse, J., ... Finkelstein, A. (2013). The Oregon Experiment—Effects of Medicaid on Clinical Outcomes. *New England Journal of Medicine*, *368*, 1713–1722.

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, *85*(1), 233–298. doi: 10.3982/ECTA12723

Belloni, A., Chernozhukov, V. & Hansen, C. (2014a). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, *28*(2), 29–50. doi: 10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V. & Hansen, C. (2014b). Inference on Treatment Effects After Selection Among High-Dimensional Controls. *Review of Economic Studies*, *81*(2), 608–650.

Bloom, B. H. (1970, July). Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communication of the ACM*, *13*(7), 422–426. Retrieved from https://dl.acm.org/doi/10.1145/362686.362692 doi: 10.1145/362686.362692

Blumenstock, J., Cadamuro, G. & On, R. (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. *Science*, *350*(6264), 1073–1076. doi: 10.1126/science.aac4420

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, *600*(7890), 695–700. doi: 10.1038/s41586-021-04198-4

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in neural information processing systems 33*.

Brunt, L. & Cannon, E. (2013). The truth, the whole truth, and nothing but the truth: the English Corn Returns as a data source in economic history, 1770-1914. *European Review of Economic History*, *17*(3).

Cambridge University Press. (2021). *Cambridge dictionary* (Fifth Edition ed.). Cambridge University Press.

Campos, J., Ericsson, N. R. & Hendry, D. F. (2005). General-to-Specific Modeling: An Overview and Selected Bibliography. *International Finance Discussion*

Papers(838).

Carbon, S., Champieux, R., McMurry, J., Winfree, L., LR, W. & MA, H. (2019). An analysis and metric of reusable data licensing practices for biomedical resources. *PLoS ONE*, *14*(3), e0213090. doi: https://doi.org/10.1371/journal.pone.0213090

Case, A. & Deaton, A. (2015). Rising Morbidity and Mortality in Midlife Among White Non-Hispanic Americans. *Proceedings of the National Academy of Sciences*, *112*, 15078–15083.

Castle, J. L., Doornik, J. A. & Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3*(1), 1–33. doi: 10.2202/1941-1928.1097

Chan, F., Harris, M. N., Singh, R. B. & Yeo, W. B. E. (2022). Nonlinear Econometric Models with Machine Learning. In F. Chan & L. Mátyás (Eds.), *Econometrics with Machine Learning* (pp. 41–78). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-031-15149-1_2 doi: 10.1007/978-3-031-15149-1_2

Chan, F. & Mátyás, L. (2022). Linear Econometric Models with Machine Learning. In F. Chan & L. Mátyás (Eds.), *Econometrics with Machine Learning* (pp. 1–39). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-031-15149-1_1 doi: 10.1007/978-3-031-15149-1_1

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, *21*(1), C1–C68. doi: 10.1111/ectj.12097

Chernozhukov, V., Hansen, C. & Spindler, M. (2015). Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments. *American Economic Review: Papers & Proceedings*, *105*(5), 486–490.

Chernozhukov, V., Newey, W. K. & Singh, R. (2022). Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, *90*(3), 967–1027. Retrieved 2022-09-16, from https://www.econometricsociety.org/doi/10.3982/ECTA18515 doi: 10.3982/ECTA18515

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., ... Cutler, D. (2016). The Association Between Income and Life Expectancy in the United States. *The Journal of the American Medical Association*, *315*, 1750–1766.

Cooley, T. & LeRoy, S. (1981). Identification and Estimation of Money Demand. *American Economic Review*, *71*(5), 825–844.

Cortada, J. W. (1993). *Before the computer: Ibm, ncr, burroughs, and remington rand and the industry they created, 1865–1956*. Princeton, NJ: Princeton University Press.

Currie, J. & Walker, R. (2011). Traffic Congestion and Infant Health: Evidence from E-ZPass. *American Economic Journal: Applied Economics*, *3*, 65–90.

data.europa.eu. (2026). *European Data - European Union.* data.europa.eu. (Accessed: 21-03-2026)

data.gov. (2026). *U.S. Government's Open Data.* https://data.gov/. (Accessed: 21-03-2026)

data.gov.au. (2026). *Australia's Public Data.* data.gov.au. (Accessed: 21-03-2026)

data.gov.uk. (2026). *UK Public Data.* data.gov.uk. (Accessed: 21-03-2026)

DataLore. (2026). *Jetbrains Datalore.* https://datalore.jetbrains.com/notebooks. (Accessed: 20-03-2026)

Dawid, H., Harting, P., Wang, H., Wang, Z. & Yi, J. (2025). *Agentic workflows for economic research: Design and implementation.* Retrieved from https://arxiv.org/abs/2504.09736

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977, September). Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x doi: 10.1111/j.2517-6161.1977.tb01600.x

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Diebold, F. X. (2003). "big data" dynamic factor models for macroeconomic measurement and forecasting: A discussion of the papers by lucrezia reichlin and by mark w. watson. In M. Dewatripont, L. P. Hansen & S. J. Turnovsky (Eds.), *Advances in economics and econometrics: Theory and applications, eighth world congress* (p. 115–122). Cambridge University Press.

Diebold, F. X. (2012). *On the Origin(s) and Dvelopment of the Term "Big Data".* University of Pennsylvania. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152421

Dobkin, C., Finkelstein, A., Kluender, R. & Notowidigdo, M. (2018). The Economic Consequences of Hospital Admissions. *American Economic Review*, *108*, 308-352.

Donaldson, D. & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(4), 171–198. doi: 10.1257/jep.30.4.171

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. doi: 10.1080/10618600.2017.1384734

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics: A festschrift in honour of david f. hendry* (pp. 88–121). Oxford: Oxford University Press.

Du, K., Zhao, Y., Mao, R., Xing, F. & Cambria, E. (2025, March). Natural language processing in finance: A survey. *Information Fusion*, *115*, 102755. Retrieved 2025-11-21, from https://linkinghub.elsevier.com/retrieve/pii/S1566253524005335 doi: 10.1016/j.inffus.2024.102755

Duflo, E., Dupas, P. & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, *101*(5), 1739–1774.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. doi: 10.1214/009053604000000067

Elliott, M., Golub, B. & Jackson, M. O. (2014). Financial Networks and Contagion. *American Economic Review*, *104*(10), 3115–3153.

Fan, J., Han, F. & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314. doi: 10.1093/nsr/nwt032

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273

Fan, J., Shao, Q. & Zhou, W. (2018). Are Discoveries spurious? Distributions of Maximum Spurios Correlations and Their Applications. *Annals of Statistics*, *46*, 989-1017.

Fellegi, I. P. & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., ... Baicker, K. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics*, *127*, 1057–1106.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Edinburgh: Oliver and Boyd.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi: 10.18637/jss.v033.i01

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, *12*(2), 23–38.

Gandomi, A. & Haider, M. (2015, April). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0268401214001066 doi: 10.1016/j.ijinfomgt.2014.10.007

Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, *56*(1), 114–137. doi: 10.1111/ecin.12364

Global Indigenous Data Alliance. (2026a). *Care principles for indigenous data governance.* https://www.gida-global.org/care.

Global Indigenous Data Alliance. (2026b). *History of indigenous data sovereignty.* https://www.gida-global.org/history-of-indigenous-data-sovereignty.

Goldfarb, A., Greenstein, S. & Tucker, C. (Eds.). (2015). *Economic analysis of the digital economy*. University of Chicago Press. doi: 10.7208/chicago/9780226206981.001.0001

Google Cloud Service. (2026). *Big Query.* https://cloud.google.com. (Accessed: 01-04-2026)

Guu, K., Lee, K., Tung, Z. et al. (2020). Realm: Retrieval-augmented language model pre-training. In *International conference on machine learning*.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L. & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, *4*(2), 2053951717745678. Retrieved from https://doi.org/10.1177/2053951717745678 (PMID: 30381794) doi: 10.1177/2053951717745678

Henderson, J. V., Storeygard, A. & Weil, D. N. (2012). Measuring economic

growth from outer space. *American Economic Review*, *102*(2), 994–1028. doi: 10.1257/aer.102.2.994

Hendry, D. F. (2024). A brief history of general-to-specific modelling. *Oxford Bulletin of Economics and Statistics*. doi: 10.1111/obes.12578

Hendry, D. F. & Krolzig, H.-M. (2001). *Automatic Econometric Model Selection Using PcGets 1.0*. London: Timberlake Consultants Press.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Hollerith, H. (1889). An Electric Tabulating System. *The Quarterly: Columbia University School of Mines*, *X*(16), 238–255.

Izacard, G. & Grave, E. (2023). Atlas: Few-shot Learning with Retrieval Augmented Language Models. *The Journal of Machine Learning Research*, *24*(1), 11912–11954.

Jean, N., Burke, M., Xie, M., Davis, W. M. A., Lobell, D. B. & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. doi: 10.1126/science.aaf7894

Joshi, M. & Krag, S. S. (2010, December). Issues in Data Management. *Science and Engineering Ethics*, *16*(4), 743–748. doi: 10.1007/s11948-010-9223-5

Jupyter. (2026). *Jupyter notebook.* https://jupyter.org/about. (Accessed: 23-01-2026)

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing*. Prentice Hall.

Kaggle. (2026). *Kaggle Platform.* https://www.kaggle.com/. (Accessed 20-03-2026)

Kallstenius, T., Capusan, A. J., Andersson, G. & Williamson, A. (2025, July). Comparing traditional natural language processing and large language models for mental health status classification: a multi-model evaluation. *Scientific Reports*, *15*(1), 24102. Retrieved from https://www.nature.com/articles/s41598-025-08031-0 doi: 10.1038/s41598-025-08031-0

Kambil, A. & Ginsburg, M. (1998). Public Access Web Information Systems: Lessons from the Internet EDGAR Project. *Communications of the ACM*, *41*(7), 91–97.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Keynes, J. M. (1939). Professor Tinbergen's Method. *The Economic Journal*, *49*(195), 1–21. Retrieved from papers2://publication/uuid/E6C89FD8-DC5B-420A-96D6-4E6CAD471EE3

Knight, K. & Fu, W. (2000). Asymptotics for LASSO-Type Estimators. *The Annals of Statistics*, *28*(5), 1356–1378.

Korinek, A. (2023a, December). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, *61*(4), 1281–1317. Retrieved from https://pubs.aeaweb.org/doi/10.1257/jel.20231736 doi: 10.1257/jel.20231736

Korinek, A. (2023b). Language models and cognitive automation for economic research. *NBER Working Paper*, *30957*. doi: https://doi.org/10.3386/w30957

Krolzig, H.-M. & Hendry, D. F. (2001). Computer Automation of General-to-Specific Model Selection Procedures. *Journal of Economic Dynamics and Control*, *25*(6–7), 831–866.

Kudo, T. & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Association for Computational Linguistics.

Larivière, V., Haustein, S. & Mongeon, P. (2015, June). The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, *10*(6), e0127502. Retrieved 2025-11-23, from https://dx.plos.org/10.1371/journal.pone.0127502 doi: 10.1371/journal.pone.0127502

Larreguy, H., Marshall, J. & Snyder Jr., J. M. (2020). Publicizing Malfeasance: How Local Media Facilitates Electoral Sanctioning of Mayors in Brazil. *The Economic Journal*, *130*(631), 2291–2327.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact Post-selection Inference, with Application to the LASSO. *Annals of Statistics*, *44*(3), 907–927. doi: 10.1214/15-AOS1371

Leeb, H. & Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, *21*(1), 21–59. doi: 10.1017/S0266466605050036

Leeb, H. & Pötscher, B. M. (2008). Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. *Journal of Econometrics*, *142*(1), 201–211. doi: 10.1016/j.jeconom.2007.05.017

Letouzé, E. (2012, May). Big Data for Development - UN Global Pulse. *UN Global Pulse*, 1–47.

Lewis, P., Perez, E., Piktus, A. et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems.*

Li, Y., Wang, S., Ding, H. & Chen, H. (2023). Large language models in finance: A survey. *arXiv preprint arXiv:2311.10723*. doi: https://doi.org/10.48550/arXiv.2311.10723

Lleras-Muney, A. (2005). The Relationship Between Education and Adult Mortality. *Review of Economic Studies*, *72*, 189–221.

Lockhart, R., Taylor, J., Tibshirani, R. & Tibshirani, R. (2014). A Significance Test for the LASSO. *The Annals of Statistics*, *42*(2), 413–468.

Lopez-Lira, A. & Tang, Y. (2023). Can chatgpt forecast stock price movements? *SSRN Working Paper*.

Loughran, T. & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Lucchetti, A. H. & Cajueiro, D. O. (2026). Language as Data: A Survey of Natural Language Processing for Economics and Finance. *Journal of Economic Surveys*, *40*(1), 340–378. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.70014 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.70014) doi: 10.1111/joes.70014

Mátyás, L. (Ed.). (2024). *The Econometrics of Multi-dimensional Panels*. Springer Cham. doi: doi.org/10.1007/978-3-031-49849-7

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, *12*(2), 685–726. doi: 10.1214/18-AOAS1161SF

Miller, S., Johnson, N. & Wherry, L. (2021). Medicaid and Mortality: New Evidence from Linked Administrative Data. *Quarterly Journal of Economics*, *136*, 1783–1829.

Mitra-Kahn, B. H. (2011). *Redefining the economy: how the 'economy' was invented 1620* (Unpublished doctoral dissertation). City Univesrity London.

Moosvi, S. (1987). *The economy of the mughal empire c. 1595: A statistical study*. Oxford University Press.

Openaire. (2026). *Data.* https://www.openaire.eu/research-data-protected-what-is-research-data. (Accessed: 23-01-2026)

Oregon State Univesrity Library. (2026). *Data.* https://guides.library.oregonstate.edu/research-data-services/data-management-define-data. (Accessed: 23-01-2026)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of emnlp*.

Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. (2007). *Numerical receipes: The art of scientific computing* (Third ed.). Cambridge University Press.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Technical Report*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K. & Semmens, J. B. (2014, August). Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics*, *50*, 205–212. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046413001949 doi: 10.1016/j.jbi.2013.12.003

Raschka, S. (2025). *Build a Large Language Model (From Scratch)*. Manning Publications Co.

re3Data. (2026). *Registry of research data repositories.* https://www.re3data.org/. (Accessed 20-03-2026)

Riloff, E. & Wiebe, J. (1999). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of aaai.*

Rockhold, F., Nisen, P. & Freeman, A. (2016, September). Data Sharing at a Crossroads. *New England Journal of Medicine*, *375*(12), 1115–1117. doi: 10.1056/NEJMp1608086

Rogin, L. (1956). *The meaning and validity of economic theory - a historical approach*. Harper & Brothers:L New York.

Rössner, P. R. (2020). Counting cows and coins. In *History and economic life*. Routledge. doi: 10.4324/9780429506819-7

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics*, *116*(2), 681–704.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Sennrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). Association for Computational Linguistics. doi: 10.18653/v1/P16-1162

Shu, F. & Larivière, V. (2024, January). The oligopoly of open access publishing. *Scientometrics*, *129*(1), 519–536. Retrieved 2025-11-23, from https://link.springer.com/10.1007/s11192-023-04876-2 doi: 10.1007/s11192-023-04876-2

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Stone, P. J., Dunphy, D. C., Smith, M. S. & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Taylor, L., Schroeder, R. & Meyer, E. (2014, July). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, *1*(2), 2053951714536877. Retrieved 2026-03-23, from https://journals.sagepub.com/doi/10.1177/2053951714536877 doi: 10.1177/2053951714536877

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*(3), 1139–1168.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., ... Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tuomi, I. (1999). Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. *Journal of Management Information Systems*, *16*(3), 103–117. doi: https://doi.org/10.1080/07421222.1999.11518258

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of acl.*

University of York Library. (2026). *Data.* https://subjectguides.york.ac.uk/data/lore. (Accessed: 23-01-2026)

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. doi: 10.1214/14-AOS1221

Varian, H. R. (1999). The economics of information technology. *California Management Review*, *43*(4), 8–18.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27. Retrieved from https://www.jstor.org/stable/23723482

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30.*

Walker, F. A. (1870). *Statistical Atlas of the United States: Baed on the Results of the Ninth Census 1870*. United States Census Office. Retrieved 2026-03-16, from https://fraser.stlouisfed.org/title/statistical-atlas-united-states-64/front-matter-574125?page=6

Wealth Hub. (2026). *Historical Data Vendor List.* https://www.wealthhubtrading.com/historical-data-vendor-list. (Accessed:20-03-2026)

White, H. (1996). *Estimation, inference and specification analysis*. Cambridge University Presss.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1). Retrieved 2021-09-26, from http://www.jstatsoft.org/v40/i01/ doi: 10.18637/jss.v040.i01

Wikipedia. (2026). *Data.* https://en.wikipedia.org/wiki/Data. (Accessed: 23-01-2026)

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016, March). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. Retrieved from https://www.nature.com/articles/sdata201618 doi: 10.1038/sdata.2016.18

Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of hlt/emnlp.*

Wu, X. et al. (2023). Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? *arXiv preprint arXiv:2305.05862*. Retrieved from https://arxiv.org/abs/2305.05862

Yang, L., Kipp, M. E. I. & Chen, J. (2024). Understanding research data licensing in the usage categories. *Proceedings of the Association for Information Science and Technology*, *61*(1), 1153-1155. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.1215 doi: https://doi.org/10.1002/pra2.1215

Yule, G. U. & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Charles Griffin and Company.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi: 10.1198/016214506000000735

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x