Chapter 4
Big Data - Outline

Felix Chan

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

4.1 Introduction

This chapter provides a brief history of *big data* and examines its impact on the future of econometric analysis when the volume of data grows exponentially with increasing complex structure. The chapter begins with a brief history of data focusing on its evolution from *data* to *big data*. It covers both *structured* and *unstructured* data and their convergence in the context of data collection, storage, and management. An overarching theme of this section is the evolution on the definition of *data* and the role of technologies, including analytical techniques, in its evolution.

This is followed by an overview on measuring size and growth of data. This includes concepts such as the *Volume, Velocity, Variety and Veracity* (4Vs) framework on measuring big data and how this framework can be applied in the context of economic research as well as its implication to econometric analysis. The section also covers recent advances in data linkage where de-identified data from different

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

sources can be merged, and thus create a much more comprehensive and holistic data for purpose of analysis and testing economic hypotheses.

Subsequent section discusses data sources, storage and management in the context of data governance and their impacts of modern econometric analysis. This section begins with an overview on possible sources of data relevant to socio-economic research. This includes researcher generated data i.e., surveys or experiments, data obtained from proprietary sources, i.e., third party data service providers, and publicly available data under different open data licensing. This section also summarises methods of sharing and publishing data including data sharing platforms such as Kaggle and the process of publishing data by assigning a *Digital Object Identifier* (DOI) to specific data. The chapter then introduces the FAIR data principle that aims to facilitate research reproducibility, as well as the CARE data principle that aims to respect data generated by first-nation and indigenous population. The section concludes by examining technologies and online platforms that facilitate the analysis of *big data* i.e., the size of data file exceeds available memory on a typical computer (as of 2025).

Finally, the chapter provides a survey on the techniques for analysing *big data*. This section has three parts and will direct readers to Chapters 12, 14, 15 and other relevant chapters for more in-depth expositions. The first part provides an overview on the history of computation. Highlighting the historical cycle on the scale of infrastructure required to perform computation for research. The second part focus on statistical analysis of big data, including various machine learning techniques, such as shrinkage estimators, random forests, artificial neural networks and other ensemble techniques. These techniques are becoming common in the econometric literature and this chapter aims to provide a clear overview on the contributions of these techniques to traditional econometric analysis beyond their ability to produce prediction. This includes their roles in facilitating valid statistical inference, including causal inference, construction of optimal instrumental variables as well as producing interpretable results that would inform policy decisions.

The third part introduces Natural Language Processing (NLP) techniques that are also becoming common to analyse textual and unstructured data, which includes techniques commonly used in qualitative research and in constructing and analysing market sentiment and financial statements. This section discusses fundamental concepts such as tokenisation of words, their representation in a vector space and the estimation of their conditional distribution through algorithms such as word2vec. The chapter concludes by discussing some recent advances in Large Language Models and the techniques to make them more specific for econometric analysis through Retrieval Interleaved Generation (RIG) and Retrieval Augmented Generation (RAG) as well as their implications to future economic and econometric research.

4.2 From Data to Big Data

4.2.1 What is data?

1. Examine the definition of data from a historical perspective.
2. Structured and unstructured data and their roles in economic and econometric analysis.
3. Programming code, search phases and prompt as data.

4.2.2 Brief history of (economic) Data.

4.2.2.1 Economic Data Before 19^{th} Century

4.2.2.2 Economic Data in Early and Mid 20^{th} Century

1. Highlighting this period as the turning due to digital revolution.
2. Cost of digital storage is lower than physical storage.
3. Advances in netoworking technologies and the explosion of internet.

4.2.2.3 Economic Data After Late 20^{th} Century

1. Advances in technologies provide efficient methods of creating, storing and distributing big data.
2. Emerging popularity of data swamp and data lake.
3. Data distribution channels including platforms such as Our World in Data and Kaggle.

Lessons:

1. Advances in data are functions of available technologies.
2. How technologies facilitate the collection and distribution of data.
3. Mid-90s is the turning point when digital storage becomes cheaper than physical storage of data.
4. Late 90's to early 2000s is another turning point. The advances in networking and the internet created platform for sharing files.
5. As a result of these technological advances, the meaning and complexity of data change.
6. Data becomes more accessible but no guarantee of their quality and has implication in research.

4.3 Measuring Big Data

This section discusses the framework to measure the *size* of big data through its historical development.

4.3.1 The History of 4vs

4.3.2 The Origin Story of 3Vs

1. Introduce the first 3Vs namely, Volume, Velocity, Variety.
2. Discuss the impacts of each of these characteristics across different areas in econometrics.
3. These characteristics are required to understand more complex data structure such as high dimensional panel data and network data.

4.3.3 Adding Veracity

1. Examine the addition of the last V, Veracity.
2. Discuss the importance of veracity in ensuring data quality.
3. Discuss veracity in the context of economic data, especially those sourced from unofficial channels.
4. VVVC vs. 4Vs - what happened to complexity (C)? The alternate measurement.

4.3.4 4Vs and Economic Data

Provide examples on describing big data using the 4Vs for each of the following areas of econometrics.

1. Cross Section
2. Time Series
3. Panel and Multi-Dimensional Panel Data
4. Network Data
5. Spatial Data
6. Surveys, Quasi-experimental and Experimental Data
7. Webscrapping and social media data

Lessons:

1. Measurement of data is evolving as data structure getting more complicated.
2. 4Vs is also likely to be evolved (more Vs?).

3. As data becomes more available, veracity becomes more and more important in the context of research.
4. Volume and velocity create new challenge for econometric analysis.

4.4 Data Sources, Storage, Processing, Management and Wrangling

4.4.1 Publicly Available, Researcher Generated and Proprietary Data

1. Evolution of publicly available data through various statistical agencies and non-profitable organisations, such as Internation Monetary Funds.
2. Commerical data service providers. The phenomenon since the digital revolution in the 90s.
3. Researchers generated data:

   a. Surveys
   b. Quasi Experimental Data (development economics)
   c. Experimental Data (behavioural economics)
   d. Webscapping

4.4.2 Data Sharing Platforms and Licensing

1. A brief history of data ownerships and the evolution of data licensing.
2. Data sharing platforms including code sharing platforms e.g., Kaggle, Google's Big Query and GitHub
3. The importance of Application Programming Interface (API) in accessing these platforms for data collections.

4.4.3 Data Linkage

4.4.3.1 Deterministic Data Linking and Merging

1. Reasons of merging data. From not having sufficient observations, e.g., (Cooley & LeRoy, 1981) to consolidating data from different sources.
2. Tools for merging data. From database operations to functions on dataframe objects in software, e.g., Stata, R, Python and Julia.

4.4.3.2 Probabilistic Linkage

1. Sensitive data and de-identification.
2. Probabilistic linking between de-identified data.
3. Case study: Impact from technology such as SeRPCurtin on Health Economics.

4.4.4 Data Processing and Wrangling

4.4.4.1 Data Cleaning

1. Provide an historical account on dealing with outliers.
2. Discuss the rise and pitfalls of *winsorising*.
3. The increasing challenge in identifying missing values and duplicated data as data becomes more complex.
4. The missing value problem arise from merging data from different sources.

4.4.4.2 Split-Apply-Combine

1. The history of split-apply-combine, the basic of data wrangling in practice.
2. Wrangle data into appropriate shape for analysis using split-apply-combine.
3. Application of split-apply-combine to fixed effect estimators in high dimensional panel data.
4. Demonstrate the trade-off between memory requirement and computation time.

4.4.5 FAIR and CARE Data Principles

4.4.6 Findable, Accessible, Interoperable and Reusable

1. Provide a history of the FAIR principle and discuss its role in econometric research.
2. Consolidate that the FAIR data principle is a recognised framework to ensure reproducibility.
3. Understand why it isn't well known among empirical researchers in economics, despite its importance in ensuring reproducibility which is becoming increasingly important, as reflected by recent changes in acceptance policy from prestigious journals.
4. Data becomes more complex and be able to reconstruct the data from its sources is also becoming important.
5. Programming code, search phases and prompt (for LLMs) are also covered under FAIR.

4.4.7 Indigenous Data and the CARE Principle

1. Research on development economics and the economics of well-being often involved data collection from indigenous population, at least in some countries.
2. Understand cultural sensitivity in the context of data collection is important.
3. Discuss the role of CARE data principle in the context of ensuring data sovereignty for its rightful owners.

Lessons:

1. Technologies lead to more detailed data collection.
2. The ability of linking data from different sources create greater research opportunities and allow testing of hypotheses that were not previously possible.
3. This, however, create complexity and new challenges. A rigorous process is required to ensure reproducibility.
4. There is a literature on model complexity. What about data complexity?
5. Linking data lead to higher chance of missing data and potential outliers. What is the best way to handle missing values and outliers. Why winsorizing (blindly) is not the best approach when handling big data.
6. The history on the development of FAIR principle demonstrates the importance of research reproducibility. Social Science research is a little behind in this space but is catching up as reflected by publication policy in various prestigious journals.
7. As data getting more complex, to what extend can process be automated?
8. Encourage the construction of meta-data for purpose of re-reproducibility.
9. The challenge of reproducibility in re-creating the data used for the analysis. This is becoming more important as data construction is becoming more common among economic researchers.
10. The challenge of reproducibility in terms of research results. Programming code should be available and should be treated as part of research data.
11. International standard in data management for economics should be developed, similar to other disciplines, such as physics, to ensure interoperability of data between different sources.
12. As data collection becomes more specific and more involved data collection from 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 or experiments.

4.5 Analytics Techniques for Big Data

4.5.1 A Brief History of Computation

4.5.1.1 From Large to Small, back to Large

1. A brief history of computers.
2. Human as computers, a large team of human to perform calculations.
3. Evolution of 'personal' computer.
4. The era of supercomputer and cloud computing.

4.5.1.2 Computing on 'Big' Data: A Historial Perspective

1. 'Big data' is relative. Data too big for contemporary computers to process is not new.
2. Examine the big data problem from the US census data in 1880.

4.5.2 Statistical Techniques

4.5.2.1 High-Dimensional Statistics

1. Provide a brief history of high dimensional statistics and discuss its relation to big data, laying the foundation for machine learning techniques.
2. High dimensional statistics and big data. How does each characteristic of big data i.e., the 4Vs relate to valid statistical inference.
3. Examine how these relate to the development in different areas in econometrics, e.g., panel data analysis, spatial econometrics, time series, discrete choice and others. Referring to other chapters for more in-depth discussion when appropriate.

4.5.2.2 Machine Learning Techniques

This subsection will only provide an overview. Will refer to Chapter 12 for more in-depth discussion.

4.5.3 Natural Language Processing and Large Language Models

4.5.4 Textual Data and Qualitative Research

1. History of qualitative research and textual analysis.
2. Development of natural language processing.
3. History of sentiment index. An example of using NLP to convert textual data into numerical data.
4. Tokenisation of words and the estimation of their conditional distribution for word predictions.

4.5.5 Large Language Models

1. How tokenisation of words form the foundation of Large Language Models.
2. Vibe coding - using LLM to code econometric routine.
3. Train LLMs locally via Retrieval Interleaved Generation (RIG) and Retrieval Augmented Generation (RAG) to optimised econometric expertise in LLMs.

Lessons:

1. The challenge of data size exceeding the capability of computing technologies has always been there in history.
2. What does data size mean in terms of statistical inference.
3. The definition of ‘big’ is important here. Many columns vs many rows! Not all ‘big data’ are structured and tabular.
4. The ‘bigger’ is the data, more missing values?
5. Does it mean asymptotic theory become even more important?
6. Does it mean small (or finite) sample properties are no longer of interest?
7. The cycle so far is; require large scale infrastructure, as technologies advance, computation can be achieved in smaller devices, e.g., PC and as data becomes more complex, large scale infrastructure is required again.
8. NLP provides another channel converting textual data to numerical data.
9. LLMs can provide assistance to construct econometric routine in any programming language.
10. LLM can be trained locally to optimise its usage for econometrics.
11. Econometrics is becoming more interdisciplinary in the sense that future econometricians will need some foundation in computer science and information technology, in addition to statistics.
12. Implication from the above is the training of econometricians. To what extend should we require some level of programming in the curriculum.

References

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