RIDING THE BIG DATA WAVE: COMPLEMENTING TRADITION WITH TRANSACTIONS

This poster aims to identify the various forms of administrative, online and 'big data' utilised by the Australian Bureau of Statistics for the production and development of Services Producer Price Indexes and Turnover statistics.

BIG DATA IN THE 21st CENTURY:

"90% of the data in the world today has been created in the last two years alone" – IBM.

"... a new kind of profession has emerged, the data scientist, who combines the skills of software programmer, statistician and story teller/artist **to extract the nuggets of gold hidden under the mountains of data**" - The Economist

"The value of big data lies in our ability to extract insights and <u>make better decisions</u>" – Dr. Michael Rappa

This will result in the collection of more relevant, insightful, timely and cost effective official statistics.

ABS OBJECTIVE:

Harness Big Data sources to create a richer, more dynamic and focused statistical picture of Australia for better informed decision making.

AUSTRALIAN GOVERNMENT OBJECTIVE:

The Australian Government aims to be <u>the world leader in the use of big data</u> analytics to drive efficiency, collaboration and innovation within the public sector.
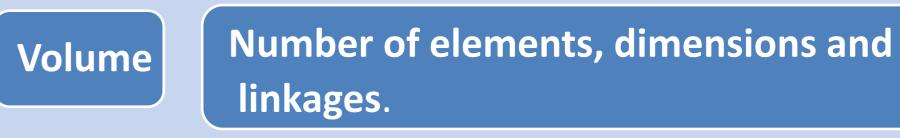
WHAT IS BIG DATA?

Generally large sets of data that is not sourced through traditional survey methodology.

- Web/online data sources
- Administrative data
- Transactions data
- Sensory data
- Satellite imagery

WHY BIG DATA ?

- Higher quality.
- Cost effective.
- Greater coverage.
- Timely.
- Relevance.

THE FOUR DIMENSIONS OF BIG DATA

Volume	Number of elements, dimensions and linkages.
Variety	Diversity of sources, formats, media and content.
Velocity	Speed at which data is being produced, changed, received, and processed .
Veracity	Quality and provenance: quality may be good, poor or undefined.

BIG DATA IN THE SPPI's

The ABS uses online data and administrative data for the following Australian and New Zealand Standard Industrial Classifications (ANZSIC, 2006):

Web/Online price collection:

- *Class 6991 – Professional Photography Services*
- *Class 6611 – Passenger Car Rental/Hiring*
- *Class 4400 – Accommodation Services*

Note: Web/Online pricing data of road tolls, freight levies and publically operated postal services supplements a predominantly surveyed sample for several classes within Division I: Transport, Postal and Warehousing Services.

Administrative data collection:

- Class 6932 – Accounting Services (for small to medium firms only)
- Class 6712 – Non-residential Property Operator Services
- Class 6720 – Real Estate Services

BIG DATA CHALLENGES

Sampling error:
- Quality is dependent on size.

Non-sampling error:

- Coverage bias.
- Self selection bias.
- Measurement bias.

Traps:
- Is big data the solution?

Big data raises <u>new challenges for National Statistical Agencies</u> in respect to the privacy and security of data.

ISSUES FOR CONSIDERATION

- Business benefit.
- Privacy and public trust.
- Technological feasibility.
- Data acquisition - Reliance on external providers.
- Data integrity.
- Methodological soundness.
- Longevity with Big Data sources

SPPID DEVELOPMENT AND BIG DATA

The ABS Consumer Price Index is investing heavily in broadening the scope and use of big data in their collections.

The PPIs are following this initiative in the development of SPPI.

Class 9520 – Funeral, Crematorium and Cemetery Services - web/online data collection with survey data where applicable.

Class 8102 – Higher Education Services - web/online data collection. Non-market (Government) pricing is available through web/online data collection.

Classes within **Division Q – health Care and Social assistance** is based upon utilising transactions data. Household and government expenditure at the service/product level for all states and territories is captured.

Development work on **Class 3109 – Other Heavy and Civil Engineering Construction** - administrative data from representative Engineering Consultancy firms.
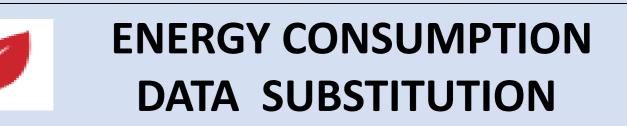
PRICE SAMPLING DEVELOPMENT

The PPI is looking to develop and access new technology that allows for the collection and use of big data for statistical purposes; examples are:
- Web-scraping technology.
- Machine-to-machine technology.
- Purpose built excel programs for the transfer of data into the prices system.

💰 TAX (ATO) DATA

ANNUAL ECONOMIC ACTIVITY SURVEY
Australian Industry cat.no.8155.0

- Business Activity Survey (BAS) data is received from the ATO via the ABS Business Register

- Complements (and supplements) financial survey data.

- Survey uses of BAS data include creation of benchmarks, weights, imputation and data confrontation.

- Replaces direct collection from micro –non-employing businesses; non-employing units contributing less than 2.5% of BAS turnover for each ANZSIC Class.

COMPLEMENTARY ESTIMATES PROGRAM

- Synthetic estimates of Manufacturing industries (Division C) at ANZSIC Class

- Dataset created annually with cat.no. 8155.0 Australian Industry

- Utilises BAS data to model for units not selected in the Economic Activity Survey.

MINING COMMODITY DATA

- Data from state departments complemented by a number of other sources needed to 'fill the gaps', e.g. BREE (Bureau of Resources & Energy Economics) and APPEA (Australian Petroleum and Production Exploration Association).

- Presents National estimates for the mining industry, in respect of employment, major economic and financial aggregates, and quantity of mineral productions across the States and the Northern Territory.

ENERGY CONSUMPTION DATA SUBSTITUTION

National Greenhouse and Energy Reporting Scheme (NGERS)
- Analytical project funded by Department of Industry and Science to measure under-coverage of NGERS dataset using ABS Environment data.

- NGERS energy consumption data to be used for subset of Energy, Water and Environment Survey (EWES) population, where the ABS statistical unit matches to NGERS reporting unit.

- Approximately 15% of total survey sample population in scope of substitution, including large industry players.

INNOVATION IN MICRO DATA LONGITUDINAL ANALYSIS: EXPANDED ANALYTICAL BUSINESS LONGITUDINAL DATABASE (EABLD)

- Project in partnership with Department of Industry and Science, and the ABS

- Firm level analysis of micro-economic drivers of performance, competitiveness and productivity.

- Integrates Tax data with ABS survey data for active businesses between 2001-02 to 2012-13.

- Provides a solid evidence base for productivity analysis, policy development and evaluation.

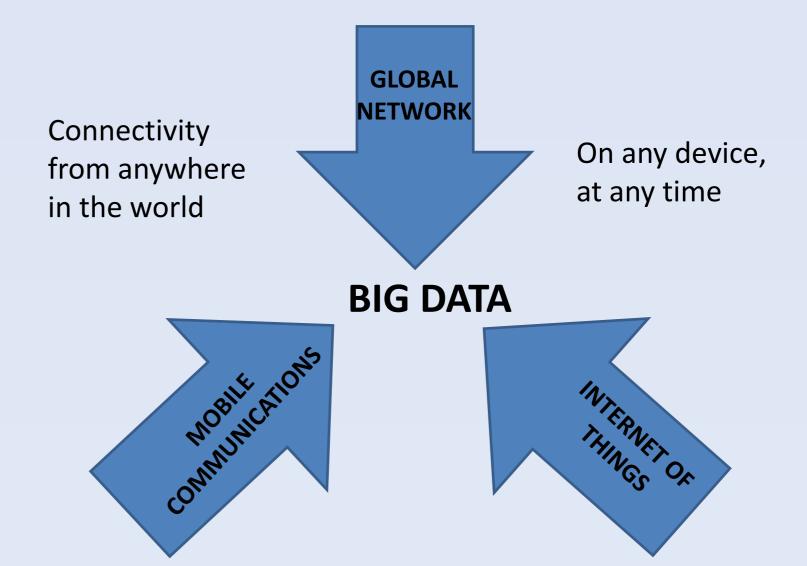- Builds upon the ABS' Business Longitudinal Database (BLD) created in 2005.

DATA SOURCED FROM THE ATO
- Business Activity Survey (BAS) Data
- Business Income Taxation (BIT)
- Pay As You Go (PAYG)

Collision point of global technology sources

GLOBAL NETWORK

Connectivity from anywhere in the world

On any device, at any time

BIG DATA

MOBILE COMMUNICATIONS

INTERNET OF THINGS

FUTURE OF BIG DATA IN TURNOVER STATISTICS
- Standard Business Reporting (SBR)
- Electronic invoicing
- Private sector transactional data