

Data and Statistics for the Post-2015 Development Agenda  
Implications for Regional Collaboration in Asia and the Pacific, UN ESCAP,  
December 2014



## Using the Data Revolution to provide more effective and efficient statistical production and services

---

### Combining official statistics with new data sources – Australian big data strategy

Gemma Van Halderen  
First Assistant Statistician,  
Population, Labour and Social Statistics  
Australian Bureau of Statistics



## The Way Forward – Making the Business Case

How will Big Data benefit official statistics?

Could it benefit both developed and developing countries?

Will Big Data help in getting timelier and more indicators for the  
Post-2015 development agenda?

Some personal views on the way forward



## How will Big Data benefit the Australian Bureau of Statistics?

Still an open question as we have yet to develop the business case for certain types of Big Data... but promising for

- Satellite Imagery Data
- Mobile phone data
- Longitudinal Linked Employer-Employee Database
- Harnessing own operational data



## Could it benefit developed and developing countries?

- Back to basics
  - Role of the NSO
  - Trust
  - Reputation
  - Principles of official statistics
- Value proposition of Big Data?
  - It is just another data source, after all
  - It is a means to an end

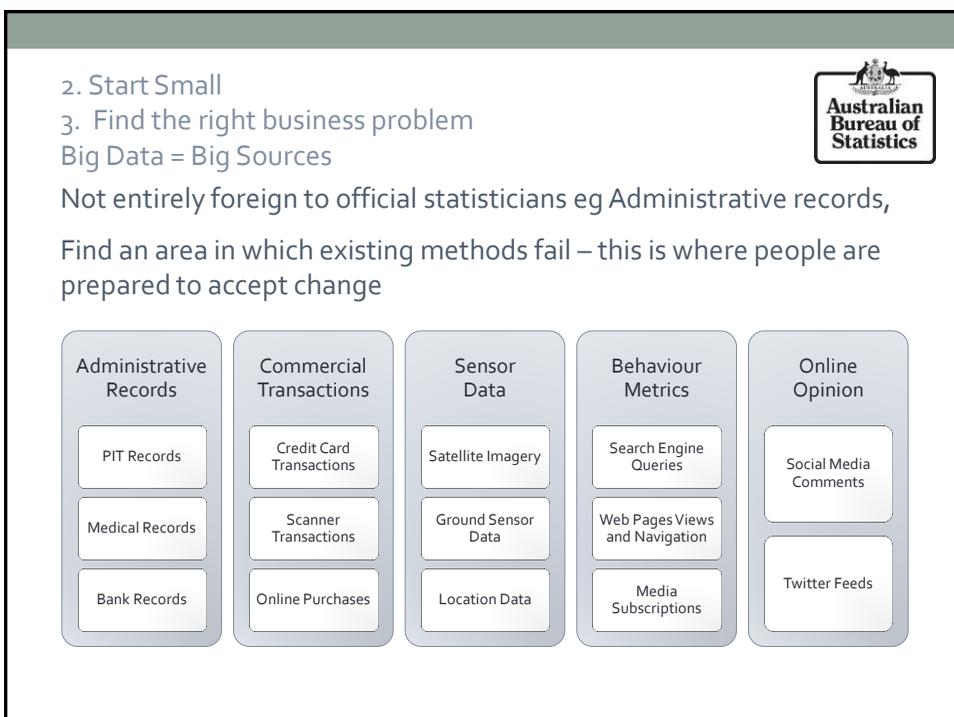
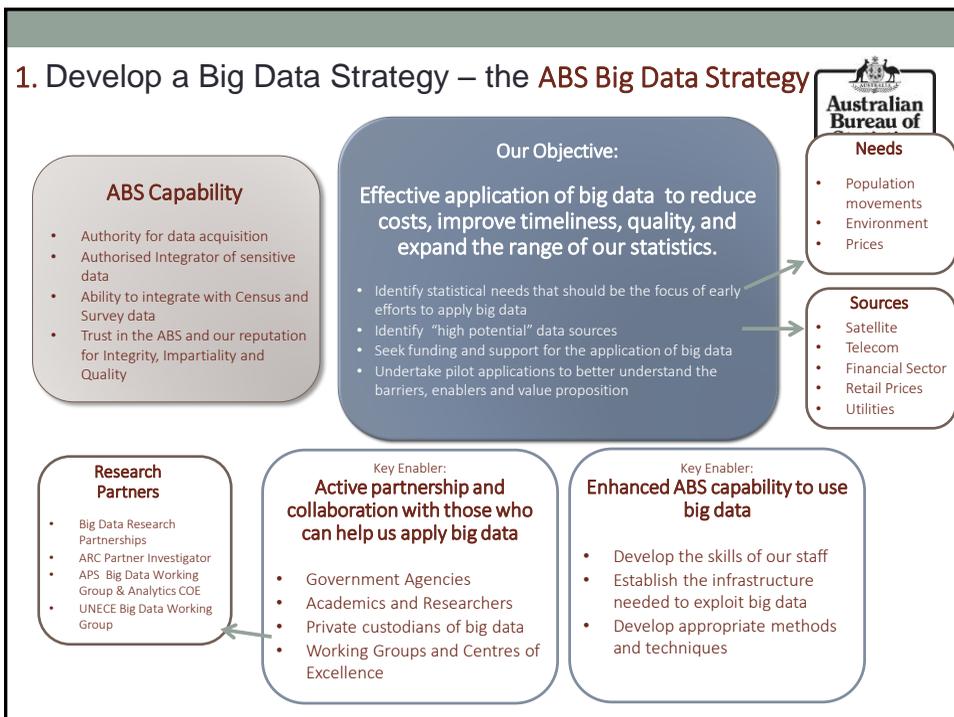


- Countries to make up their own minds about what is best for their country
  - Position the office for the future
    - Other Govt Depts, or private sector providers
- Country specific issues
  - Data acquisition
    - Global guidelines and umbrella agreements
  - Methodology
    - 57 projects grouped into clusters to share research efforts/results
  - IT, legal framework, privacy and confidentiality issues etc.
- Cultural change – management, staff, users on modelling and estimates
  - Trust of official statistics is paramount
    - Trust comes by walking and leaves on a horse back – Dutch saying



### Is this a way forward?

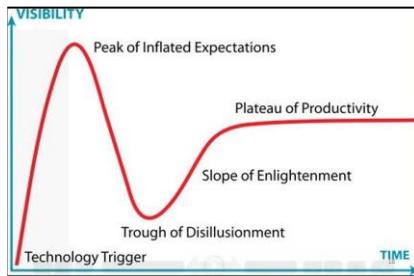
1. Develop a Big Data Strategy
2. Start small
3. Find the right business problem (eg post-2015 indicators)
4. Find business champions
5. Set realistic expectations
6. Develop/Adapt the right methodology
7. Pilot testing
8. Develop a business case – cost benefit analysis



## 5. Big Data = Big Hype? Importance to set right expectations



### Gartner Hype Curve



### Big Data on the Hype Curve



## 6. Avoid Big Traps – Size is not everything



Two broad types of errors in data sets

- Sampling error
  - Dependent on size
- Non sampling error
  - Coverage bias - Big Data population is not the population
  - Self selection bias – squeaky wheels
  - Representation bias – multiple representation
  - Measurement error
  - Increasing the sample size does NOT reduce non-sampling errors

Traps

- Big Data is seen by some as a solution is search of a problem
- Putting the cart before the horse

Develop/adapt the right methodology

- Ground truth and modelling
- Demonstrate results through pilot testing

## 7. Pilot testing



ABS Flagships

What problem we are trying to solve?

What methodology to use for analysis?

## ABS Big Data Research areas - Flagship



Satellite imagery data for agricultural statistics (eg)

Multiply-linked employer-employee data for productivity analysis

Mobile positioning data for measuring population mobility

Predictive modelling of survey non-response behaviour

Data visualisation techniques for exploring large datasets

Predictive modelling of unemployment for small areas

(in decreasing order of progress of development)

## What is our research problem?



*Rather than exclusively through a  
**traditional survey collection**,  
is it possible to use  
**satellite imagery data**  
to estimate the  
**area of land used to grow different crops**  
**and crop yields**  
in Australia?*

## Why?



Potential to reduce costs by

- Reducing the sample size for Agricultural surveys

Provision of more frequent data

Provision of small area data

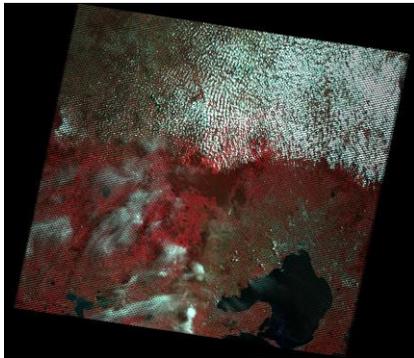
Business case has yet to be established

- Current priority is to test the efficacy of the methodology

# Estimating crop yields from



## Satellite imagery

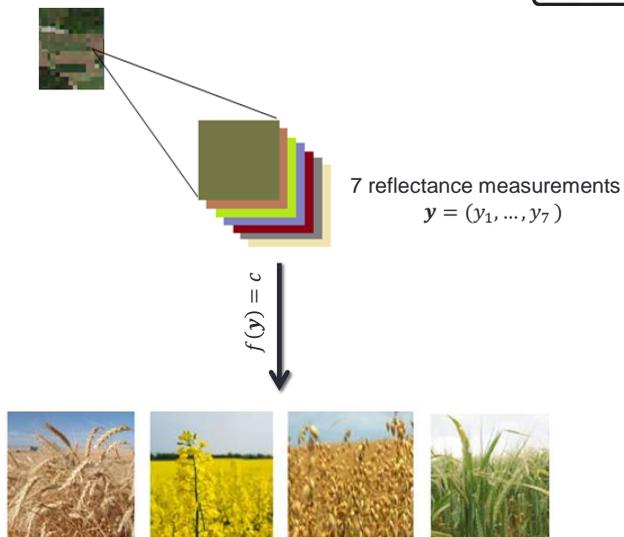


## The data

Landsat 7 imagery from US Geological Survey  
 - reflectance data from 7 freq bands for pixels of 25x25 m<sup>2</sup>

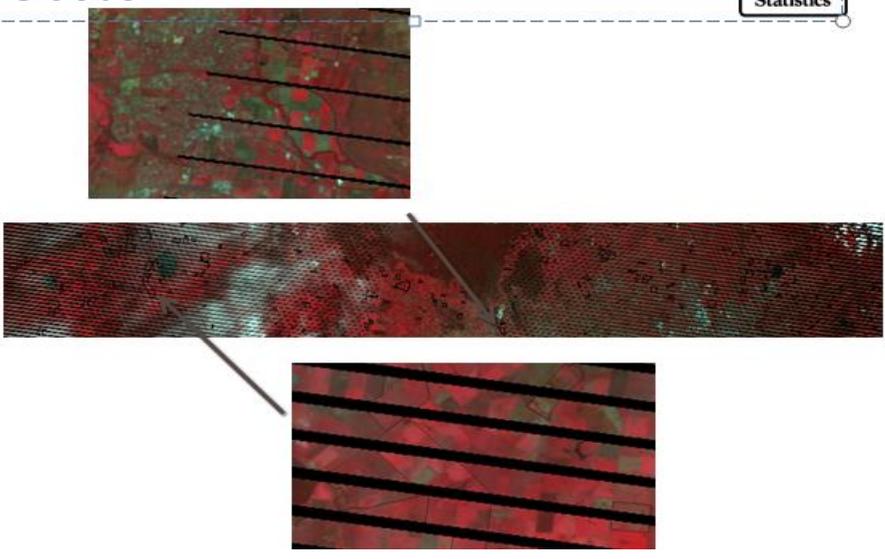
Band1	Band2	Band3	Band4	Band5	Band6	Band7
514	745	888	1908	2112	2233	1356
584	708	953	1763	1940	2233	1378
532	727	985	1872	1961	2233	1290
550	764	985	1981	2197	2233	1489
550	764	969	1981	2069	2233	1356
550	745	985	1945	2048	2233	1312
550	690	921	1799	2197	2182	1512
584	727	888	1727	2175	2182	1489
584	708	888	1763	2154	2130	1512
532	727	904	1763	2133	2130	1489

# Pixel classification and yields



# Missingness from equipment problems

Clouds

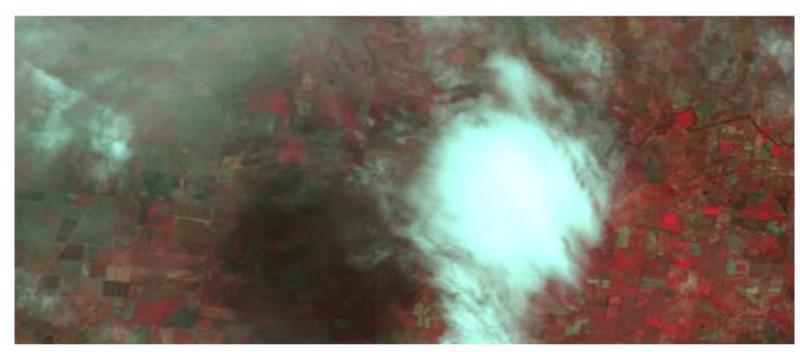


Australian Bureau of Statistics

The slide illustrates missing data in satellite imagery. It features three satellite images of a landscape. The top image shows a small section with several black diagonal lines, indicating missing data. The middle image shows a larger section with a grid pattern and black diagonal lines. The bottom image is a zoomed-in view of the black diagonal lines, showing they are not random but follow a specific pattern.

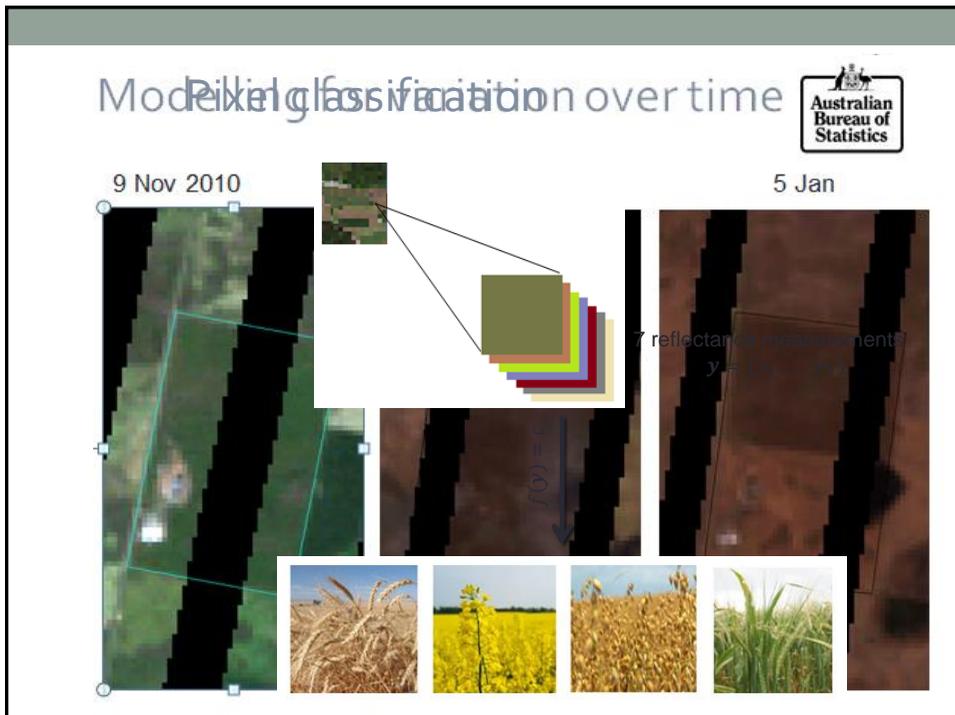
# Perpetual cloud cover

Missing not at random



Australian Bureau of Statistics

The slide illustrates perpetual cloud cover. It features a satellite image of a landscape with a large, bright, circular area of cloud cover. The text 'Missing not at random' is displayed below the title.



## 7. Business case – will it benefit the ABS or official statistics?



Still an open question as ABS have yet to develop the business case for certain types of Big Data... But promising for

- Satellite Imagery Data
- Mobile phone data
- Harness own operational data



In conclusion

## Big Data and Big Challenges



### ABS objective

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

### Challenges

#### Business benefit

Privacy and public trust

Technological feasibility

Data acquisition

Data integrity

#### Methodological soundness

- How to make valid statistical inferences

## Big Data and Big Opportunities



### Possible benefits

- Replace direct data collection
- Complementary direct data collection
- Substitute data items
- New data items
- Supplementary information to improve quality

### Statistical activities

- Sample frames or registers
- Small domain estimation
- Small population group estimation
- Enabling data imputation, editing and confrontation
- Enabling data linking and fusion
- Producing new statistical products
- Improving statistical operations

## Big data and Big Experiments



Will and/or How will Big Data help in getting timelier and more indicators for the Post-2015 development agenda that governments and communities trust?