Australian Bureau of Statistics

statistics for informed decision making

Strategic opportunities through applying semantic technologies to modernising official statistics

Dr Siu-Ming Tam

SemStats 2013

22 October 2013

Welcome to SemStats 2013!

The goal of the SemStats workshop is to explore and strengthen the relationship between the Semantic Web and statistical communities…

…to provide better access to the data held by statistical offices.

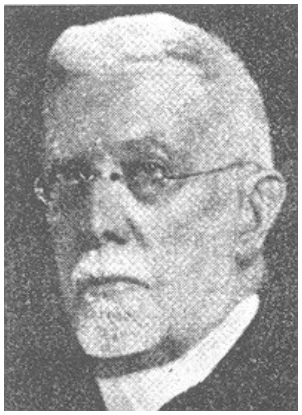
…and to assist in local and global Modernisation of Statistical Production and Services

Statistics

- *Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty* – American Statistical Association
- Origin of the term Statistics
  - Statistics 1.0
  - Statistics 2.0

Change brings uncertainty….

- Statistics help government, business and the community understand and shape change
  - Evidence based decision making
- But what happens when statisticians face change?
  - In technology
  - In meeting increasing and users expectations
  - In scientific methods for producing statistics
  - In budgets!



Kaier

Georg von Mayr in response to Anders Kaier's *On the Method of Representative Enumerations Serving as a Type of the Entire Population of a State* (ISI 1895)

Von Mayr

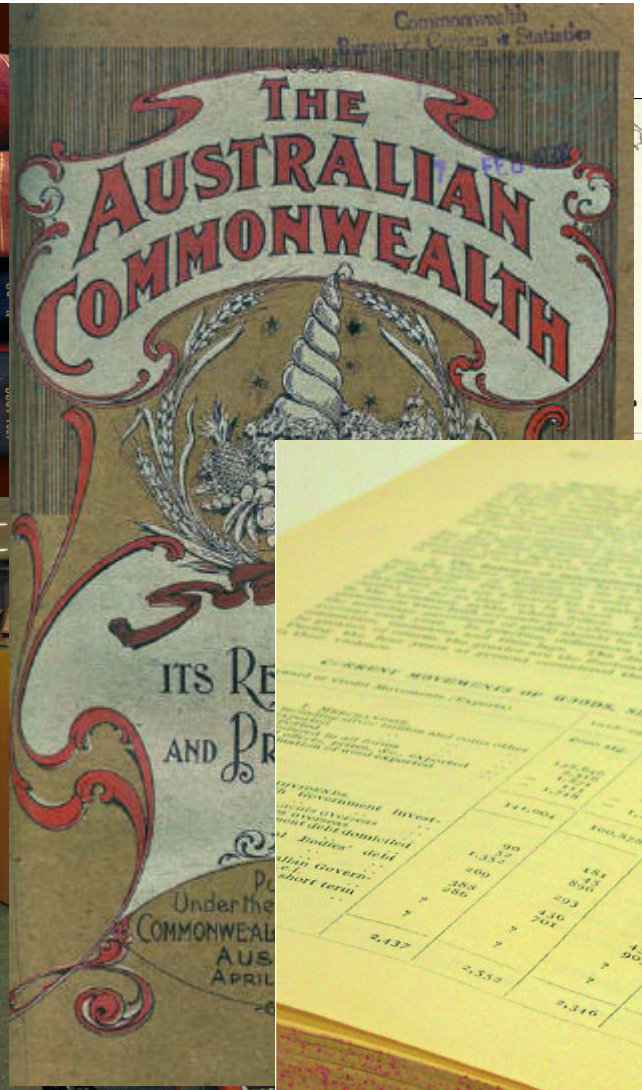
"...I regard as most dangerous the point of view found in his work. I understand that representative samples can have some value, but it is a value restricted to the domain already illuminated by full coverage. **One cannot replace by calculation the real observation of facts.** A sample provides statistics for the units actually observed, but not true statistics for the entire terrain.

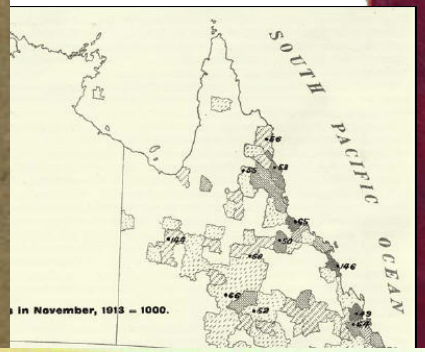
It is **especially dangerous** to propose representative sampling in the midst of an assembly of statisticians. Perhaps for legislative or administrative goals sampling may have uses - but one must never forget that it cannot replace a complete survey. It is necessary to add that there is among us these days a current in the mind of mathematicians that would, in many ways, have us calculate rather than observe. **We must remain firm and say: no calculation when observation can be made**."

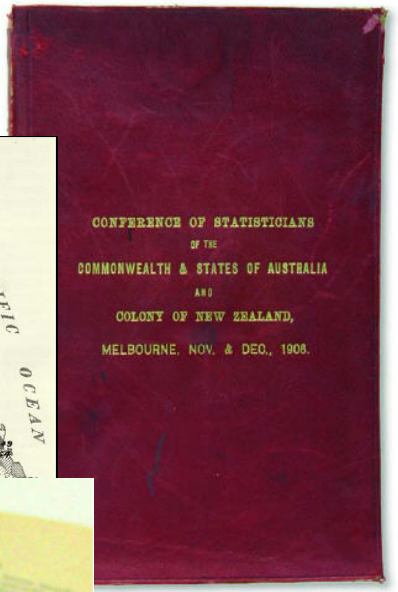
100+ years later…

- Sample surveys for statistical units (persons, businesses) are now “orthodox”
- …. but face to face interviews are expensive (and have other limits, e.g. timeliness, provide load)
  - Can we use Web surveys?
  - Can we use administrative registers instead?
  - Can we sample big data?
  - Can we integrate data from multiple sources to form a well rounded perspective?

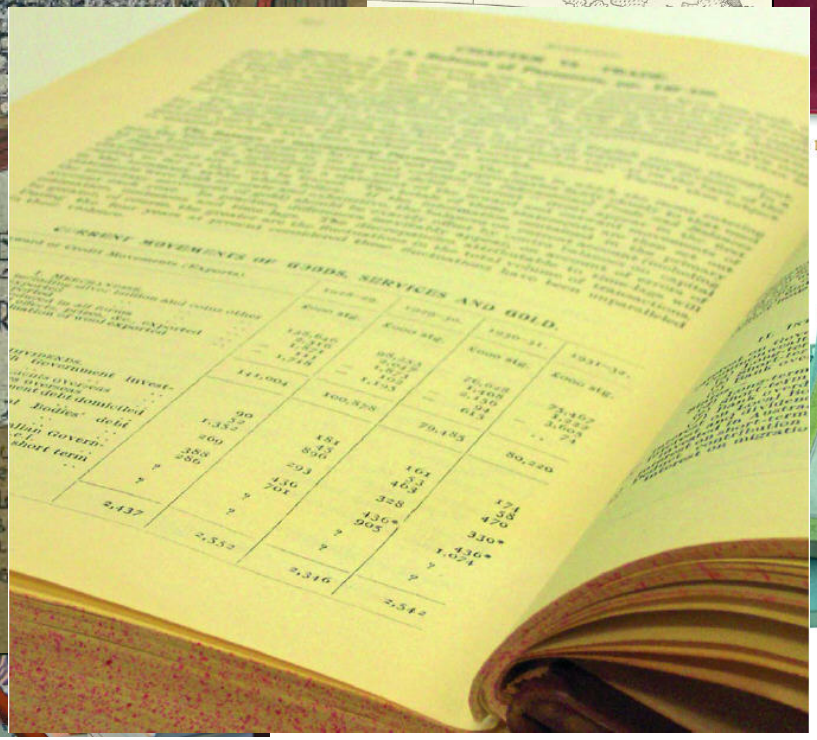
From paper based products



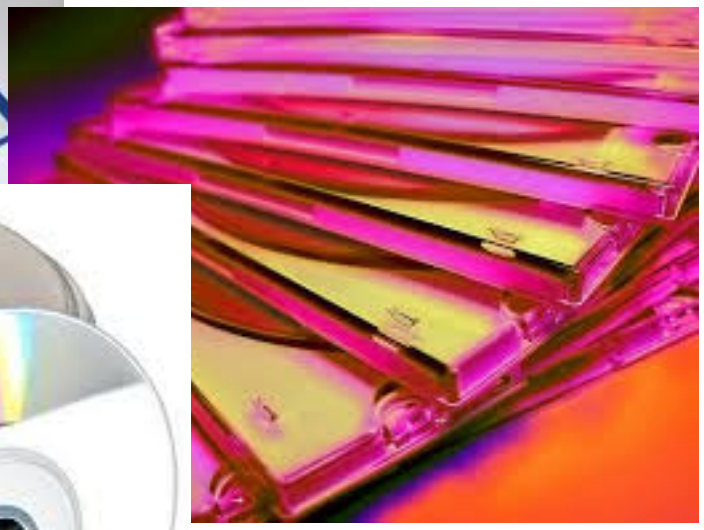




1906 Conference of Statisticians.

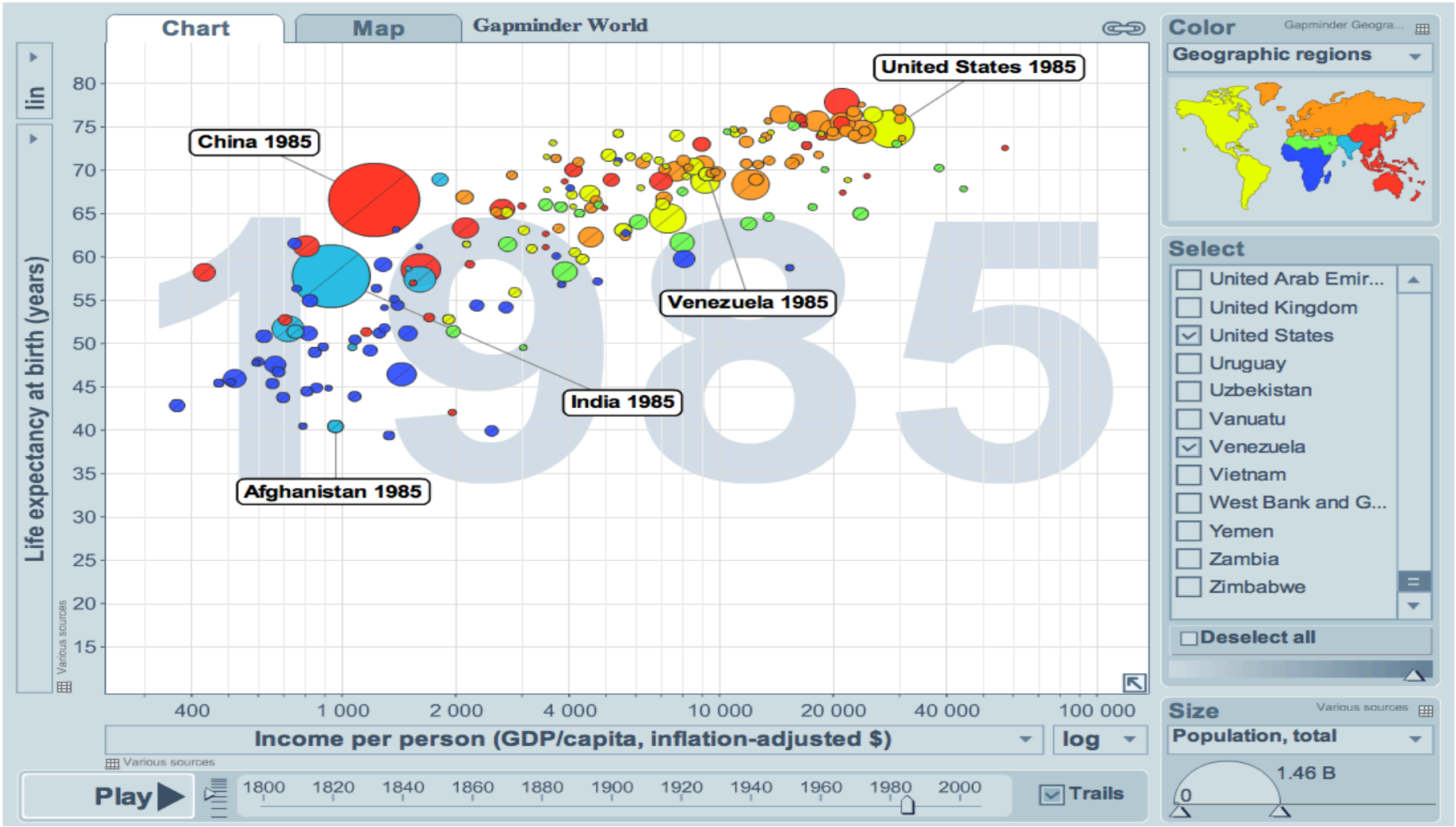


To electronic products



To web-based products

Gapminder World by Hans Rosling



Australian House Price Indexes - Static

6416.0 - House Price Indexes: Eight Capital Cities, Jun 2013 Quality Declaration

LATEST ISSUE Released at 11:30 AM (CANBERRA TIME) 06/08/2013

Summary | Downloads | Explanatory Notes | Related Information | Past & Future Releases

Page tools: Print Page | Print All | Email Notification | RSS | Search this Product

Main Features

Analysis

ABS House Price Index Methodology

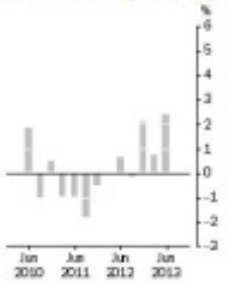
Use of Price Indexes in Contracts

About this Release

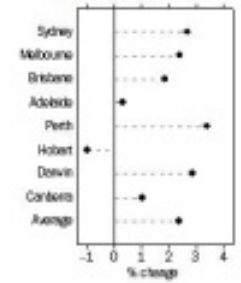
JUNE KEY FIGURES

Established house prices	Mar Qtr 13 to Jun Qtr 13 % change	Jun Qtr 12 to Jun Qtr 13 % change
Weighted average of eight capital cities	**2.4**	**5.1**
Sydney	2.7	6.1
Melbourne	2.4	3.3
Brisbane	1.9	3.7
Adelaide	0.3	0.6
Perth	3.4	11.0
Hobart	-1.0	1.2
Darwin	2.9	7.7
Canberra	1.0	2.6

Established house prices, Weighted average of eight capital cities - Quarterly % change



Established house prices, Quarterly % change - June quarter 2013



Australian House Price Indexes - Interactive

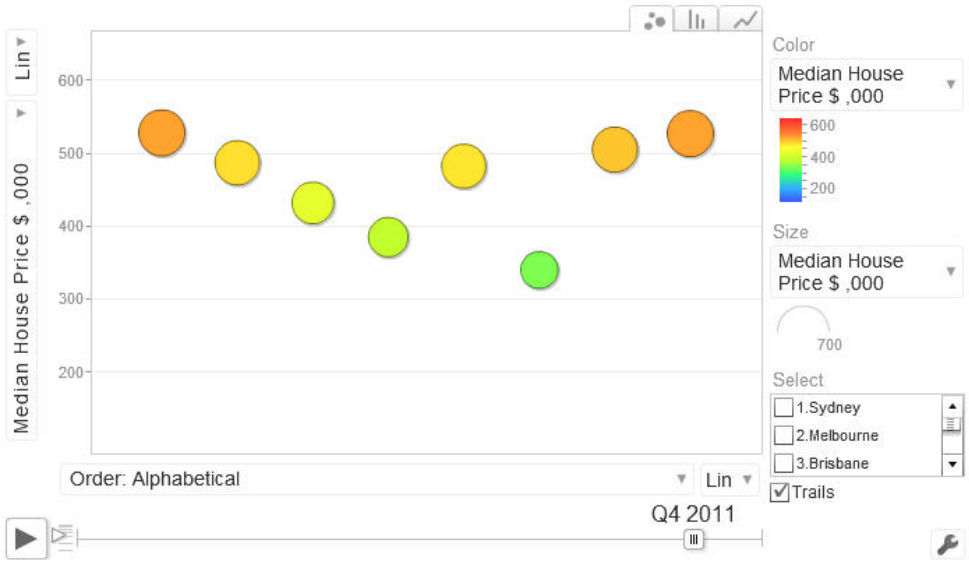
6416.0 - House Price Indexes: Eight Capital Cities

Find out how your state compares using House Price Index data in Google Motion Charts!

Features of Google Motion Charts

- Adjust the speed of the animation by dragging the small white triangle on the right next to the Play button
- Change the colour and size, and select variables (capital city) on the Chart

To view the animation, the free plug-in 'Adobe Flash Player' may need to be installed on your computer.



Each Capital City is represented by a bubble and it's colour and size is related to the Median House Price (unstratified) at that point in time.

The default view is:

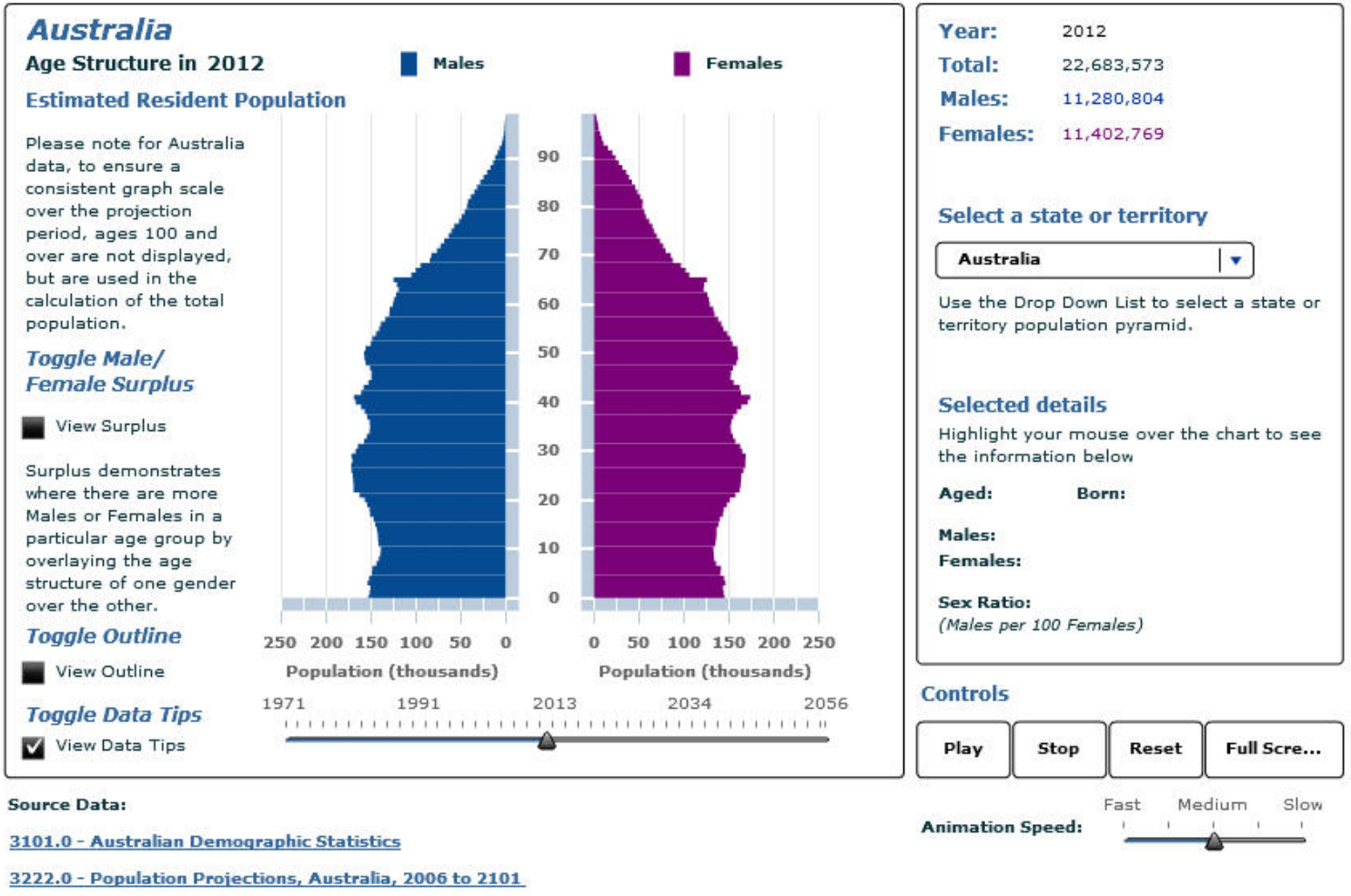
- X-axis: Eight Capital Cities
- Y-axis: Median House Price (unstratified)
- Time: Press Play and the chart is animated from March Quarter 200 onwards

Click Play ▶ to begin the animation.

Source: Cat. No. 6416.0 - House Price Indexes: Eight Capital Cities, June 2013
Table 7. Median House Price (unstratified) from March Qtr 2002 to December Qtr 2012

This page first published 2 July 2010, last updated 9 August 2013

Australian Population Pyramid



Source Data:

3101.0 - Australian Demographic Statistics

3222.0 - Population Projections, Australia, 2006 to 2101

This page first published 27 August 2008, last updated 17 December 2012

Official Statistics

Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation – UN Fundamental Principle

OGD: Oil, Gold, Democracy?
Data is the new gold

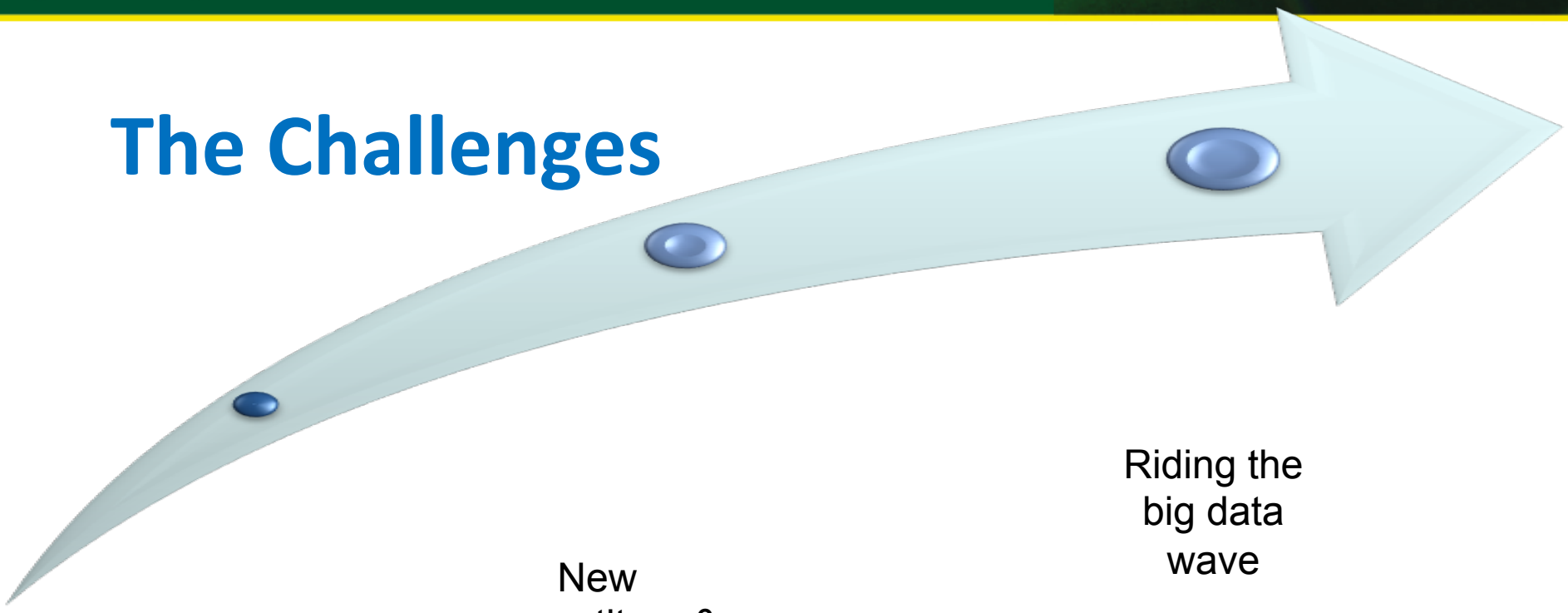
- *Just as oil was likened to black gold, data takes on a new important and value in the digital age…*
  – Neelie Kroes, VP of EC for Digital Agenda
- Public sector information already generates 32 billion € (est) of economic activity each year
- The Open Data package – for data to be able to be re-used released by the EU – will double that

What would be the added return from Linked Open Data?

NSO values include…

- Open access to data
  - Free, equal access, and open licensing system
- Support data with metadata
- Linked open data (?)
  - Tim Berners-Lee's 5 star linked open data format
  - Semantic web is not well understood by statistical leaders
    - Web 3.0 vs Web 2.0
    - SDMX vs DDI

The Challenges



Rapid changes in the environment

New competitors & changing expectations

Diminishing budgets

Increasing cost & difficulty of acquiring survey data

Riding the big data wave

Competition for skilled resources

Changing expectations?

- Users of statistics expect more data, more quickly
- They also want to:
  - "mash up" with data from other sources
  - explore the statistics visually, interactively, with their own analytical techniques
  - discover their own stories

Riding the Big Data wave

- Challenges and opportunities
  - Business comes first
  - Validity of statistical inferences
    - Scientific methods, **including measuring, controlling and communicating uncertainty**, need to be developed & proven
  - Legitimacy, Legacy and Leveraging
    - Replacing collections
    - Blending
    - Gaps
    - Operational efficiency

NSO responses: Modernisation program

- HLG – Europe, Asia and Pacific
  - Products/Process vision
- Products vision
  - Integrated and better access to micro data
  - More small population group/area data
- Process vision
  - Statistics Production Model/ Industralisation
  - Metadata standards
- International collaboration

How can semantic technologies help?

Challenges for Semantic "webbers" and official statisticians

- How do Linked Data open the way to new sources of data?
- How can Semantic Web Technologies help better understand and use new sources of data, including web pages, social media, other sources of big data?
- How can the Semantic Web help us better present our statistics, tell statistical stories, and allow users to interact with data cubes, geospatial, time series and graphical visualisation?
- How can the semantic technologies help us to be more efficient and effective?

Could Semantic Web in any sense, also form "part of the challenge"?

- Is there a risk of fundamentally different ontologies emerging for various topics (eg "Occupation") which compete with relevant statistical classifications that are based on proven conceptual design principles and rigour?
- Is there a risk that poor definition of concepts associated with Linked Data, including incorrect assertions by the provider that their concepts equate semantically with published statistical definitions, lead to misinterpretation and misapplication?

The way forward for Semantic Statistics

- **Articulate the value proposition** of the semantic web for official statistics, particularly to senior leaders of NSOs
- **Raise awareness** of sematic web amongst data managers and those involved in disseminating statistics
- **Continue the dialogue** between the two groups – more Semstats!

Questions?