International Statistical Institute

Statistical Thinking & Methodology: Pillars For Quality In The Big Data Era

Pedro Luis do Nascimento Silva
President of ISI
Principal Researcher, ENCE, Brazil

(Big) Data Era

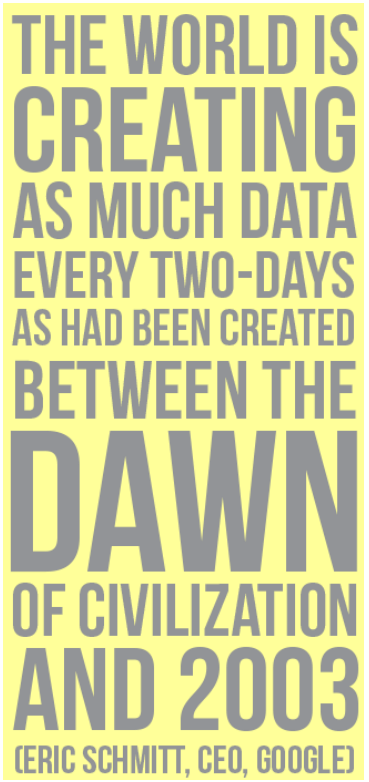
We live in an era of unprecedented volume, availability and access to **data**.

(Big) Data Era

We live in an era of unprecedented volume, availability and access to data.

Global Partnership for Sustainable Development Data (GPSDD)

http://www.data4sdgs.org/#news



(Big) Data Era

We live in an era of unprecedented volume, availability and access to data.

“Data in the world is doubling every 18 months.”

IBM

http://www-01.ibm.com/software/data/demystifying-big-data/

Data Gaps

Despite this *data deluge*, there are glaring **data gaps**.

> "For example, in low-income countries more than 70% of births – almost 20 million children annually – are not registered."

Paris21: http://datarevolution.paris21.org/the-project

Data & Development

"On September 27th 2015, 193 world leaders committed to 17 Global Goals to achieve 3 extraordinary things in the next 15 years.

- End extreme poverty.
- Fight inequality & injustice.
- Fix climate change."

Data & Development

"To reach these Sustainable Development Goals (SDGs), we will need to confront **a crisis** at the heart of solving many of the world's most pressing issues - a crisis of **poor use, accessibility, and production of high quality data** that is stunting the fight to overcome global challenges in every area—from health to gender equality, human rights to economics, and education to agriculture.

The availability and access to **high quality data** is essential to measuring and achieving the SDGs."

http://www.data4sdgs.org/#intro

Central Banks

Play a key role in **shaping policy**.

Understand the role of **relevant**, **accurate and timely data** for:

- Informed debate;
- Policy making;
- Policy evaluation & monitoring.

Operate both as **data producers** and **data consumers**.

Official and Public Statistics

Typical data sources (observational studies):

- **Censuses**
  - Data obtained from **every unit** in the target population.
- **Sample surveys**
  - Data obtained from **samples of units** in the target population.
- **Administrative records**
  - Data obtained for admin purposes, but later used for statistical purposes.

Big Data

New and emerging **data sources**:

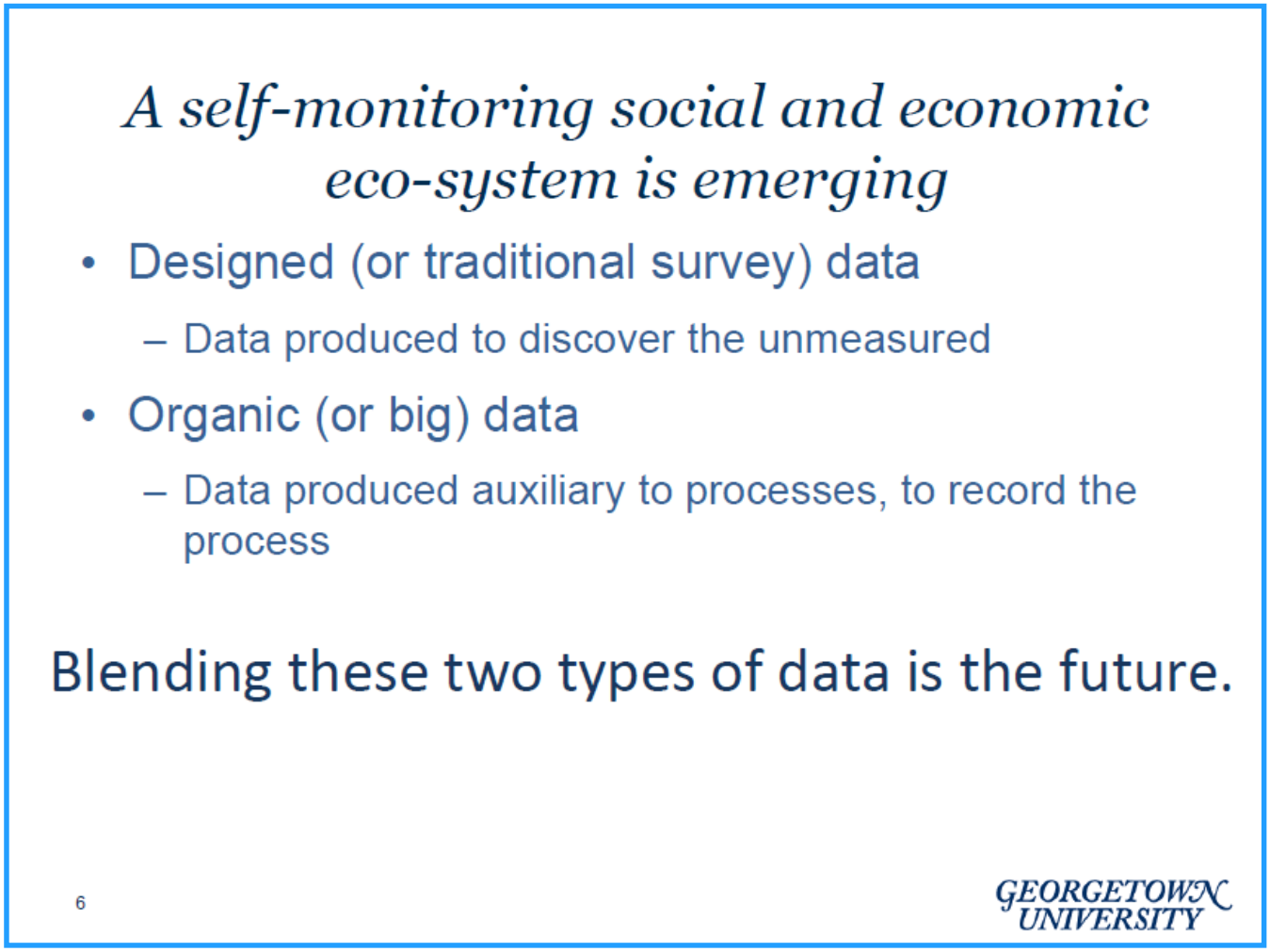
"Big Data are data sources that can be – generally – described as: high **volume**, **velocity** and **variety** of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making."

UNECE Definition 2013

Types of sources:

Social networks (communications; images; searches);
Traditional business data (transactions; records);
'Internet of things' (sensor data).

UNECE Classification:
http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data



Robert Groves

http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/

Big Data Quality Issues

Variability or Volatility

Inconsistence and/or instability of data across time.

Veracity

Ability to trust that data is accurate and/or complete.

Complexity

Need to link multiple data sources.

Accessibility

Need to ensure that data is and will remain available.

Data Quality in the Big Data Era

More data **does not** necessarily **mean** good or better data!

Many of the data available **lack the quality** required for its safe use in many applications.

Challenges are even bigger with Big Data!

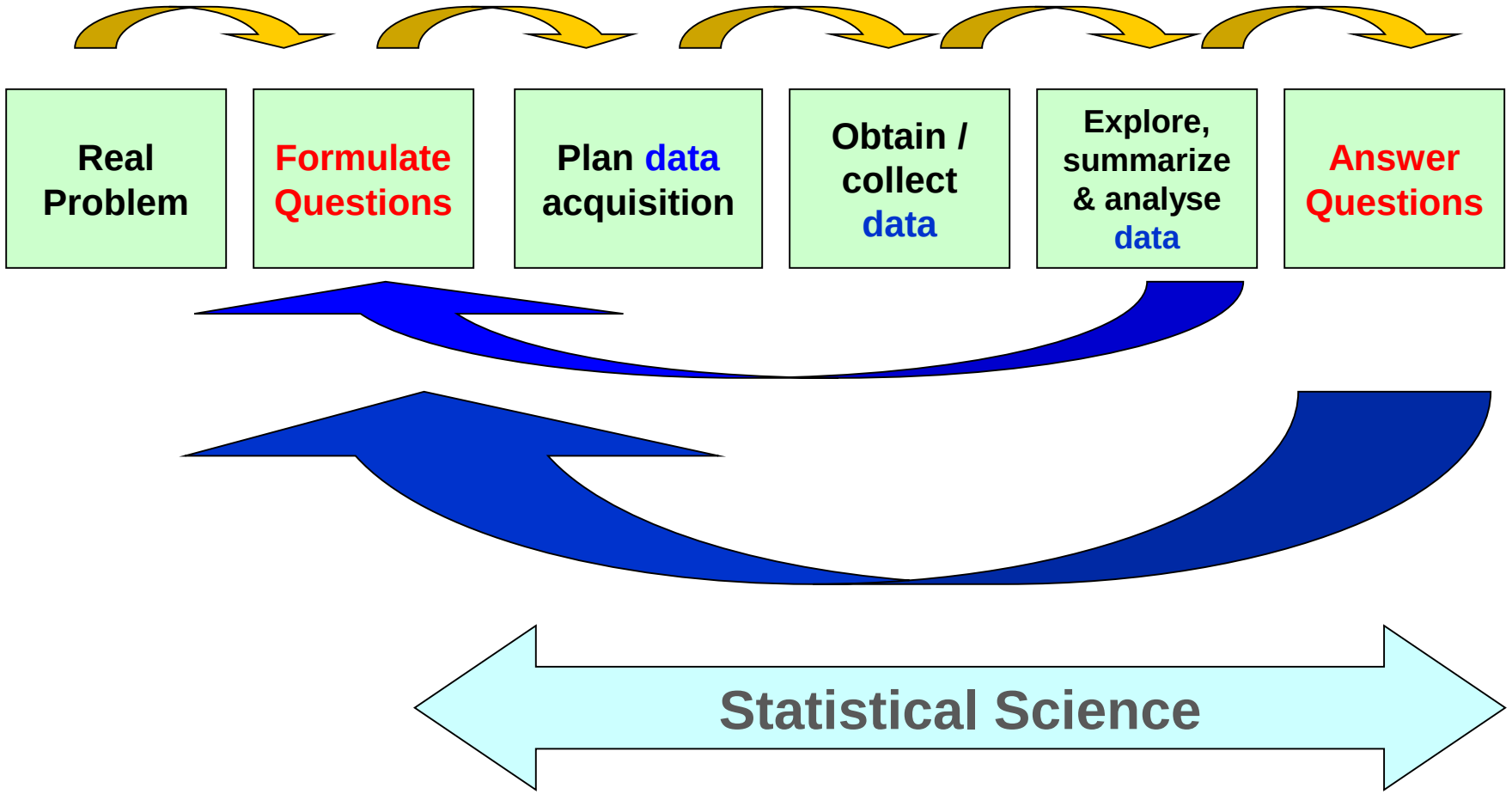
Statistical Science

For all the above reasons, **Statistical Science** has never been in such **evidence** and in such **high demand**.

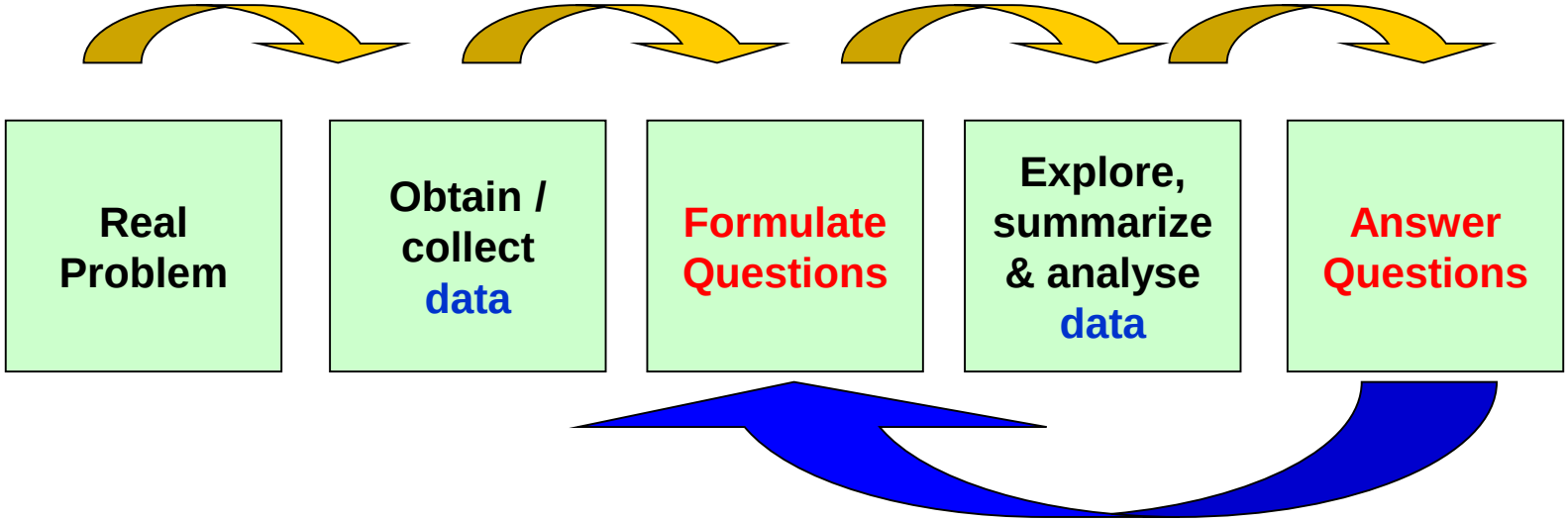
Statistical thinking & methodology offer the essential guidance to obtaining relevant, accurate, current, and cost-effective data.

It also guides the **extraction of useful knowledge from data**, to support decision making.

Conventional Knowledge Generation Process



Knowledge Generation Process in the Big Data Era



Statistical Thinking

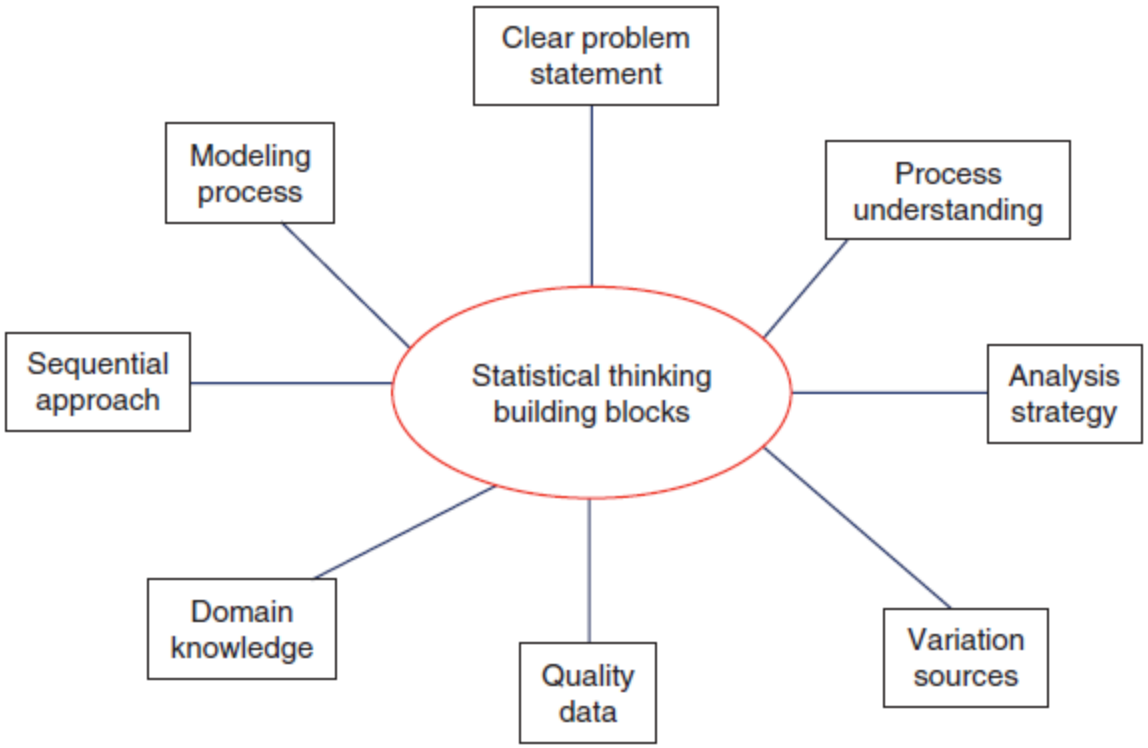


FIGURE 1 | The building blocks of statistical thinking.

Source: Hoerl, Snee & De Veaux (2014)

Statistical Science

Providing solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;

Statistical Science

Providing solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;

Statistical Science

Providing solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;

Statistical Science

Providing solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;
- Using fitted models to answer formulated questions (inference); and

Statistical Science

Providing solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;
- Using fitted models to answer formulated questions (inference); and
- Creating visual displays of data, summaries and key findings revealed from the data.

Obtaining Data

Methods for careful planning and conducting of cost-effective **data gathering** studies:

- Sampling;
- Design of experiments;
- Design for observational studies;
- Measurement protocols (questionnaires, instruments, record keeping protocols, etc.)
- Data checking, cleaning, storage and sharing protocols.

Analysis / discovery

Methods for exploratory and confirmatory data analysis:

- Exploratory data analysis;
- Data mining;
- Hypothesis formulation and testing;
- Model formulation, fitting, selection, diagnostics and interpretation;
- Data summarization, presentation & visualization.

Data Quality

Quality is desirable attribute of all data.

Data quality derives from **quality of the source**(s), **measurement instruments** & **methods**.

Vague concept: **what is data quality**?

Must be defined, so that it can be planned, measured and evaluated.

Data Quality Frameworks

Several important organizations have invested in developing frameworks for data quality:

- ✓ US Office of Management and Budget (2006);
- ✓ Statistics Canada (2009);
- ✓ International Monetary Fund (2012);
- ✓ OECD (2012);
- ✓ UN (2012);
- ✓ IBGE (2013).

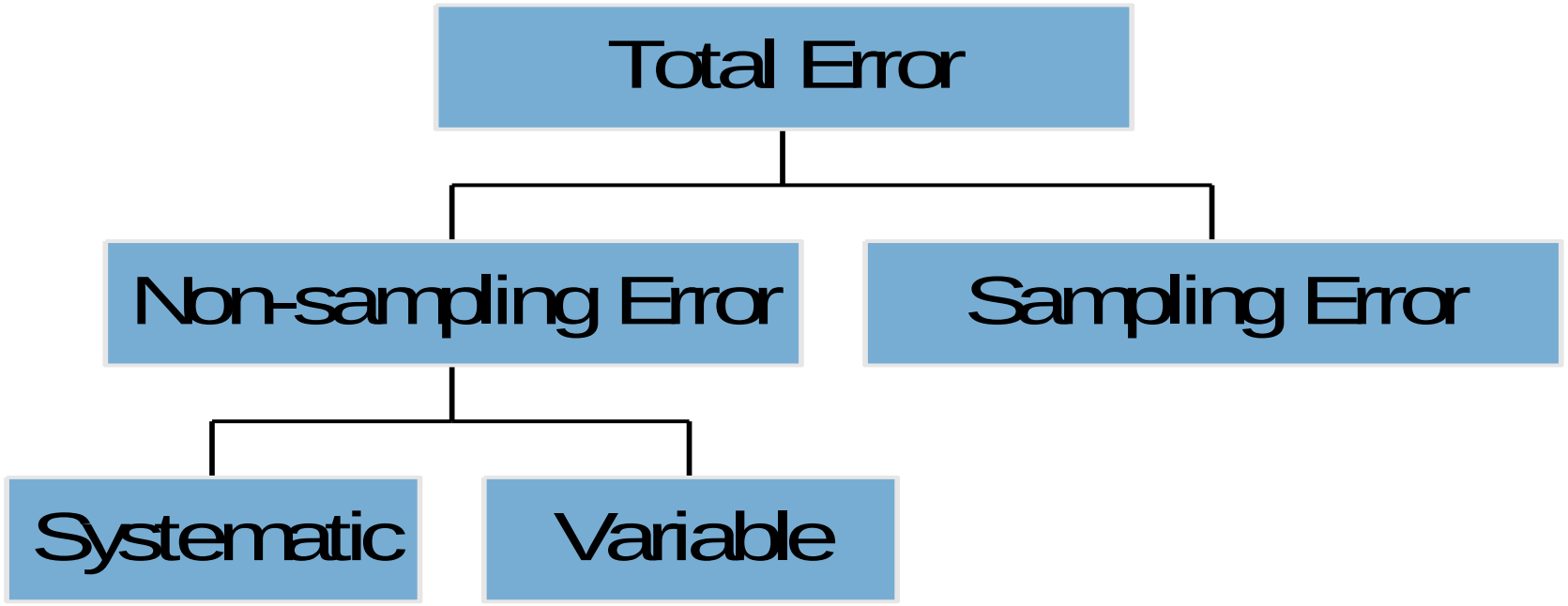
UNECE Framework for the Quality of Big Data

- Institutional / business environment (agency providing the data)
- Privacy and Security
- Complexity
- Relevance
- Time factors

- Accessibility and Clarity
- Usability
- **Accuracy**
- Completeness
- Coherence
- Validity

“Error” in Estimates

Error = Estimate – True Value



Source: United Nations (2005).

Sampling Error

Errors arising due to sampling for observation.

Easier to control.

Bias (systematic error) may be avoided ➔ use **probability sampling.**

Sample design, sample size and **estimator** defined to make **variable sampling error** as small as required.

Sampling Error

Errors arising due to sampling for observation.

Easier to control.

Bias (systematic error) may be avoided ➔ use **probability sampling.**

Sample design, sample size and **estimator** defined to make **variable sampling error** as small as required.

With '**Big Data**', there may no longer be sampling error in many applications!

Non-Sampling Error

Two broad classes of **non-sampling errors**.

Errors due to '**non-observation**':

- Coverage (frames, populations);
- Non-response (collection).

Errors in **observations**:

- Specification;
- Measurement;
- Linking, processing & estimation.

Non-Sampling Error

Two broad classes of **non-sampling errors**.

Errors due to '**non-observation**':

- Coverage (frames, populations);
- Non-response (collection).

Errors in **observations**:

- Specification;
- Measurement;
- Linking, processing & estimation.

With '**Big Data**', non-sampling errors dominate!

Even worse: they may not vanish with large n!

Summarizing

Data quality remains fundamental concern.

Statistical thinking & methodology are essential pillars for promoting:

- **data quality**;
- sound evidence-based decision making.

Big data era will require **more statistical development**, not less.

ISI
Statistical Science for a Better World

References

1. De Veaux, R. & Hand, D.J. (2005). How to lie with bad data. Statistical Science, 20, 3, pp. 231-238.
2. De Veaux, R. Hoerl, R.W. & Snee, R.D. & (2016). Big Data and the missing links. Statistical analysis and data mining, doi: 10.1002/sam.11303.
3. European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.
4. Hoerl, R.W.; Snee, R.D. & De Veaux, R. (2014). Applying statistical thinking to 'Big Data' problems. WIREs Comput Stat, doi: 10.1002/wics.1306.
5. IBGE (2013). Código de Boas Práticas das Estatísticas do IBGE. Rio de Janeiro: IBGE.
6. International Monetary Fund. 2012. *Data Quality Assessment Framework - Generic Framework*.
7. Lyberg, Lars. 2012. "Survey Quality." *Survey Methodology* 38 (2): 107–130.

References

8. Office of Management and Budget. 2006. Standards and Guidelines for Statistical Surveys. Federal Register. Washington, DC.
9. Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Ottawa, Canada: Statistics Canada.
10. Statistics Directorate, OECD. 2012. *Quality Framework and Guidelines for OECD Statistical Activities*.
11. Stigler, Stephen M. (2015). The seven pillars of statistical wisdom. Talk at LSHTM on 29 January 2015.
12. Stigler, Stephen M. (2016). The seven pillars of statistical wisdom. Harvard University Press.
13. United Nations. 2005. *Household Sample Surveys in Developing and Transition Countries*. Ed. Department of Economic and Social Affairs. *Studies in Methods*. Vol. F No. 96. New York: United Nations.

References

14. United Nations. 2005. *Designing Household Survey Samples: Practical Guidelines*. Ed. Statistics Division Department of Economic and Social Affairs. *Studies in Methods*. Vol. F No. 98. New York: United Nations Statistics Division.
15. United Nations. 2012. Guidelines For The Template For A Generic National Quality Assurance Framework (NQAF). http://unstats.un.org/unsd/dnss/qualityNQAF/nqaf.aspx
16. Vale, Steven. 2009. *Generic Statistical Business Process Model*.
17. Weisman, Ethan, Zdravko Balyozov, and Louis Venter. 2010. *IMF's Data Quality Assessment Framework 1*.

IFC Conference 2016 – Most Frequent Words on Paper Titles

IFC Conference 2016 – Most Frequent Words on Paper Titles