Big data: potential, paradoxes and the renewed importance of statistical thinking

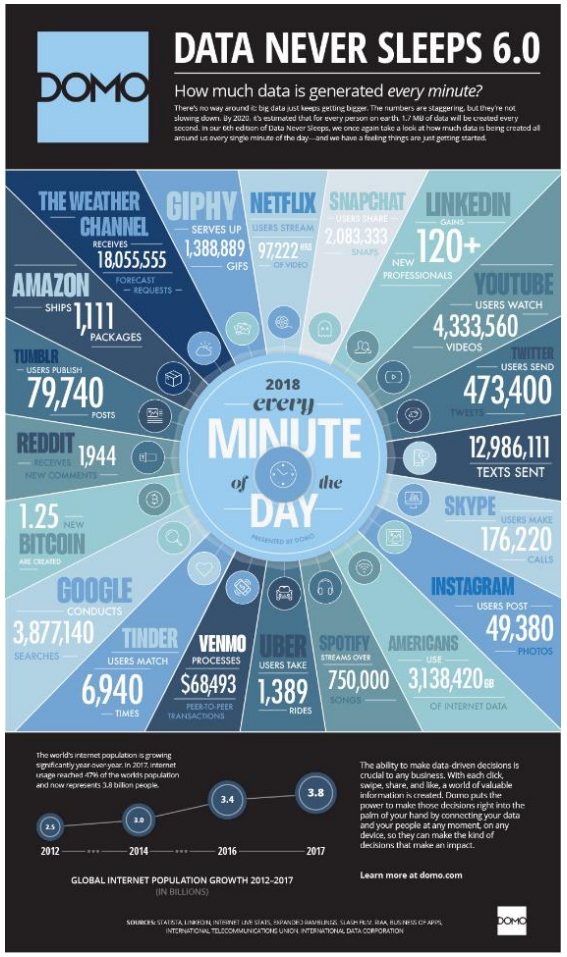
Pedro Silva

2nd Conference on Statistics and Data Science
Salvador - November 2019

The Data Era

We live in an era where **availability and access to data** have no precedent in human history.

According to one recent estimate, humans generate 2.5 quintillion (2.5e18) bytes of data every day on average.



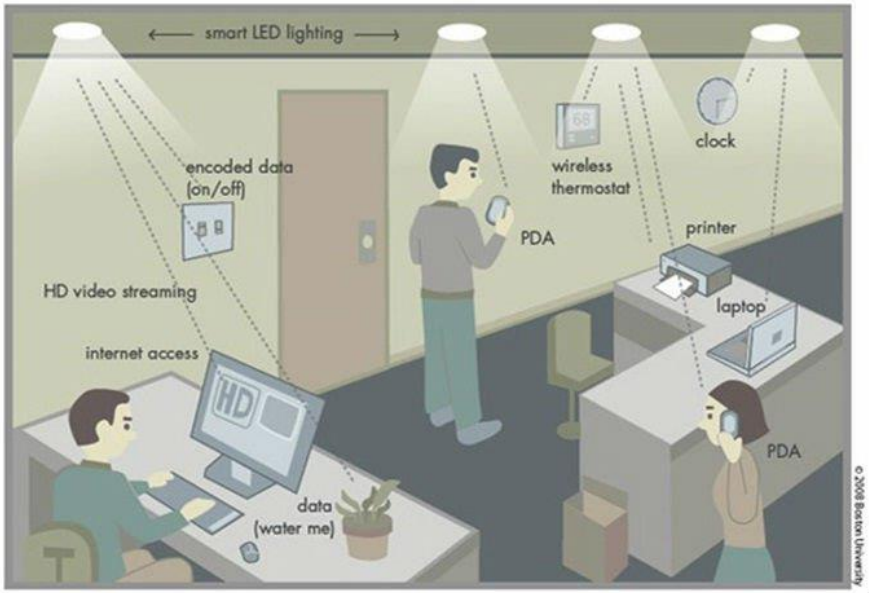
https://www.domo.com/learn/data-never-sleeps-6.

The Data Era

Some call it the '**Big Data Era**'.

I call it the '**Data Era**'.

This is because not all data are 'big', but we surely **produce, access** and **use** a lot of data as part of our daily lives.



The Data Era

"The appearance of **new data sources** due to the expansion of information technologies, and a growing number of **people connected** to information systems are transforming the way data has been traditionally produced, disseminated and used."

Cázarez-Grageda & Zougbede (2019).

"… the public and the private sectors have **embraced big data**, as more and more people recognize that big data can provide insights into the nature of biological processes, precision medicine, climate change, social and economic behaviour, risk assessment, and decision making."

He et al (2019)

Official and Public Statistics

Traditional 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 administrative purposes, but later used for statistical purposes.

Official and Public Statistics

Official and public statistics based on traditional data sources are key part of our **shared knowledge base** about the world.

This knowledge is now being challenged by information coming from new data sources, mostly big data sources.

First, we now can ask **new questions** about the world, but many such questions are only asked after we have collected / obtained / seen the data.

Second, new data may not answer questions we already have, and yet may be used to **challenge** currently available answers.

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

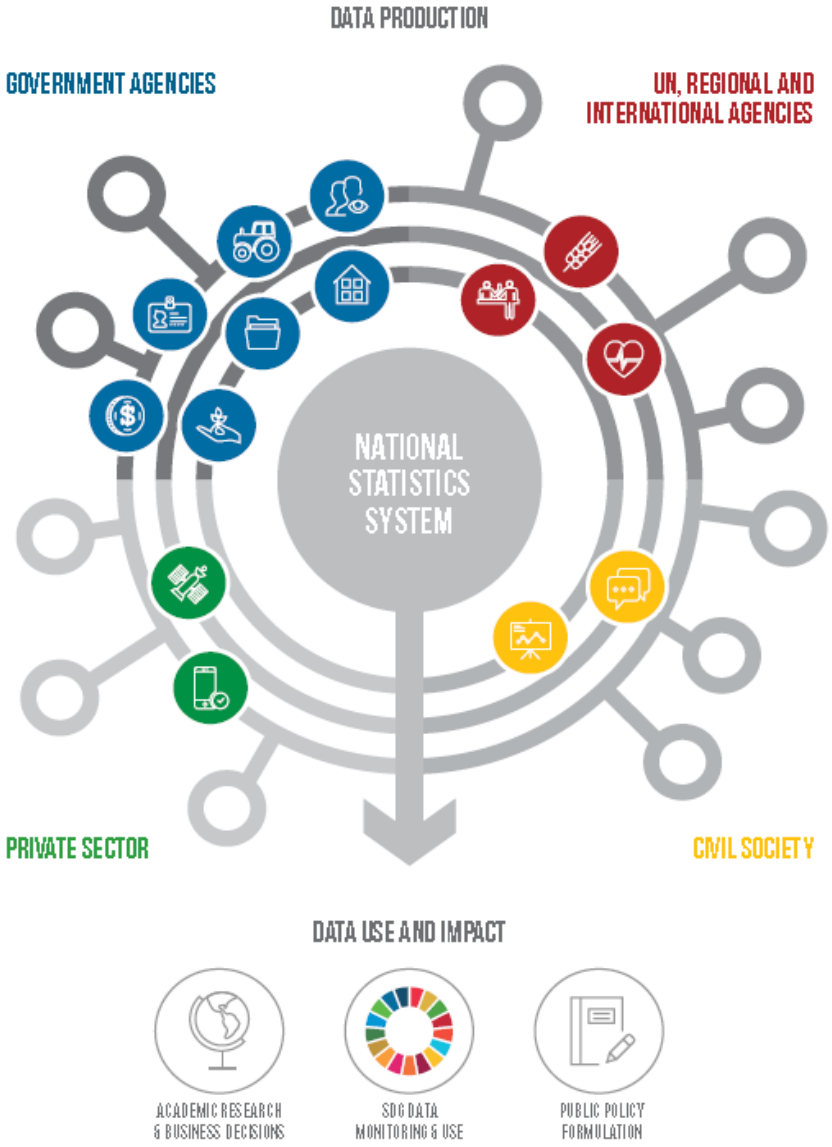
Types of Sources

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



New data ecosystem

Source: Open Data Watch 2016

Data quality

Quality is desirable attribute of all data.

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

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

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

Frameworks for Data Quality

Several important organizations have invested in defining frameworks for data quality.

'Quality frameworks':

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

OECD Quality Framework

Quality Dimension	Description
Relevance	Statistics and data are relevant if they satisfy user's needs.
Accuracy	Refers to the closeness between the values (estimates) provided and the (unknown) true values.
Credibility	Credibility of data products refers to the confidence that users place in those products.
Timeliness	Timeliness of data products reflects the length of time between their availability and the event or phenomenon they describe.
Acessibility	Accessibility of data products reflects how readily the data can be located and accessed.
Interpretability	Interpretability of data products reflects the ease with which the users may understand and properly use and analyse the data.
Coherence	Coherence of data products reflects the degree to which they are logically connected and mutually consistent.
Cost-efficiency	Cost-efficiency with which a product is produced is a measure of the costs and provider burden relative to the outputs.

Source: OECD Statistics Directorate (2012).

Big Data Quality Issues for Official Statistics

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 be available.

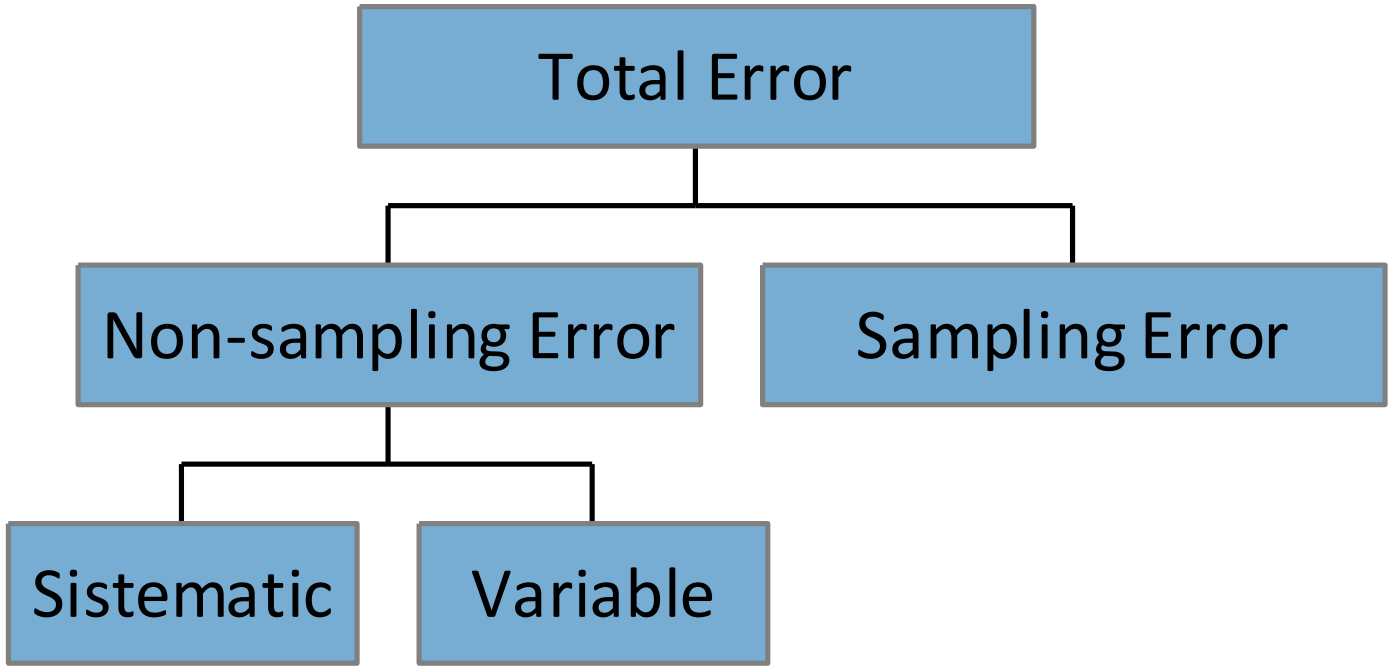
UNECE Framework for the Quality of Big Data

Institutional environment (of data provider / source)	Relevance
Privacy and Security	Time factors
Complexity	Coherence and validity
Completeness	Accuracy (selectivity)
Usability	Accessibility and clarity

Source: UNECE Big Data Quality Task Team (2014)

Error in Estimates

Error = Estimate – True Value



Source: United Nations (2005).

Sampling Error

Easier to control.

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

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

Sampling Error

Easier to control.

Bias (systematic error) can 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 some 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;
  - 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;
  - Processing & estimation.

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

Big Data and Coverage Issues

"If you don't have a website nowadays, you don't exist." – Yuhui Xie, citing talk by Carlos Scheidegger at AT&T in 2012.

My version, as a Statistician from IBGE:

If you don't have access to the internet, you will be invisible!

Internet access in private permanent households	%
Brazil	74.7%
Urban	80.0%
Rural	40.8%

Big Data – Coverage Bias (Meng, 2018)

Register $\mathcal{R}$ covers a fraction c (= m/N) of the population.

Simple Random Sample (SRS) $\mathcal{A}$ of size n (= f × N) from the whole population.

Scenario: c >> f (Register is much larger than the Sample).

Research question

How large must n (or f) be before an estimator based on the Sample is **more accurate** than the estimator based on the Register?

Estimating Population Mean – Big Data x SRS

Sample size needed to estimate population mean with smaller MSE for SRS than for the Register

N	c	m	rho_R,y	
			0.01	0.05
200,000,000	50%	100,000,000	10,000	400
	80%	160,000,000	40,000	1,600
	95%	190,000,000	190,000	7,600

Big Data – Coverage Bias (Meng, 2018)

$U = \{1, 2, ..., N\}$ (Target Population).

$\mathcal{R} \subset U$ is the subset of m units from U covered by the **Register**.

$\mathcal{A} \subset U$ is the subset of n units selected from U by **SRS**.

Goal: estimation of the population mean.

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k$$

Estimators for the Population Mean

Using the **Register**

$$\bar{y}_{\mathcal{R}} = \frac{1}{m}\sum_{k\in\mathcal{R}} y_k = \frac{1}{m}\sum_{k\in U} R_k y_k$$

Using the **SRS**

$$\bar{y} = \frac{1}{n}\sum_{k\in\mathcal{A}} y_k = \frac{1}{n}\sum_{k\in U} A_k y_k$$

where $R_k = I(k\in\mathcal{R})$ and $A_k = I(k\in\mathcal{A})$.

Sources of Error

For both estimators, and under the finite population sampling approach, the only random quantities in the estimators are the **indicators of inclusion** (in the **Register** or in the **Sample**).

For the SRS case, the inclusion mechanism is random and controlled by the analyst.

For the Register case, the inclusion mechanism is typically unknown and out of the control of the analyst.

Accuracy of Estimates

Mean Square Error under SRS:

$$MSE(\bar{y}) = V(\bar{y}) = \frac{1\text{-}f}{n}\frac{N}{N\text{-}1}\sigma_y^2 \cong \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2$$

Mean Square Error under Register:

$$MSE(\bar{y}_{\mathcal{R}}) = E_R(\bar{y}_{\mathcal{R}} - \bar{Y})^2 = E_R\left(\rho_{R,y}^2\right) \times \left(\frac{1}{c} - 1\right) \times \sigma_y^2$$

$$\rho_{R,y} = \frac{Cov_1(R_k; y_k)}{\sqrt{V_1(R_k)V_1(y_k)}}$$

Data defect index

Accuracy of Estimates

To ensure that the SRS gives smaller MSE than the Register, the sample size must satisfy:

$$n \geq \left(\frac{m}{N-m}\right)\frac{1}{E_R\left(\rho_{R,y}^2\right)}$$

Considering N = 200 million, $\rho_{R,y}$ = 0.05, and c = 50%, the MSE of the sample mean from a SRS with n = 400 would be less than or equal to that obtained from a Register covering m = 100 million!

Big Data and Quality

Result is clear: 'blind faith' in the Register can be dangerous and lead to poorer results than those that might be obtained even from small well-designed samples.

Core lesson: new data sources must be assessed with the same rigour as traditional ones (censuses, samples, etc.).

Some Approaches to Address Coverage Bias

Kim & Wang (2018)

Subsampling units from the Register (aiming to compensate for the coverage bias). ➔ Sampling!

Combining the Register with an independent sample using propensity scores for inclusion in the Register.

➔ Data Integration.

Pfeffermann (2017)

Model the conditional probability of Register inclusion given known covariates. ➔ Statistical Modelling.

Subsampling the Register

Four step procedure.

1. **Test for coverage bias** in the Register.
2. If bias is present, **compute importance weights** for units in the Register using information about covariates:

$$w_{1k} = \exp\left(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}}\right) \Big/ \sum_{k \in \mathcal{R}} \exp\left(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}}\right)$$

where $\hat{\boldsymbol{\lambda}}$ satisfies

$$\sum_{k \in \mathcal{R}} \frac{\exp\left(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}}\right)}{\sum_{k \in \mathcal{R}} \exp\left(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}}\right)} \mathbf{x}_k = \sum_{k \in \mathcal{R}} w_{1k} \mathbf{x}_k = \bar{\mathbf{X}}$$

Subsampling the Register

3. **Subsample from the Register** with probabilities proportional to the importance weights:

$\pi_{2k} = n\, w_{1k}$ with $n \leq 1/\max_{k \in \mathcal{R}}\{w_{1k}\}$.

4. **Estimate** the mean from the subsample using:

$$\hat{\theta} = \sum_{k \in s_{\mathcal{R}}} \frac{w_{1k} y_k}{\pi_{2k}} = \frac{1}{n} \sum_{k \in s_{\mathcal{R}}} y_k$$

Crucial assumption: selection to the Register ignorable given covariates: $P\left(R_k=1 \left| \mathbf{x}_k, y_k\right.\right) = P\left(R_k=1 \left| \mathbf{x}_k\right.\right) \; \forall k \in U$.

Combining the Register with a Probability Sample

Here the idea is to use data from a probability sample to obtain weights for units in the Register that can be used for unbiased (consistent) estimation of the population mean.

Data availability set-up:

Source	**Selectivity**	**Covariates x**	**Response** y
Sample $\mathcal{A}$	No	✓	
Register $\mathcal{R}$	Yes	✓	✓

Crucial assumption: selection to the Register ignorable given covariates: $P\left(R_k=1 \left| \mathbf{x}_k , y_k \right.\right)=P\left(R_k =1 \left| \mathbf{x}_k \right.\right) \forall k \in U$.

Combining the Register with a Probability Sample

Assume further that the indicator R_k of whether unit k is included (or not) in the Register $\mathcal{R}$ can be observed for all units in the sample $\mathcal{A}$.

Then use the sample data to fit a model for the register inclusion probabilities:

$$P\left(R_k = 1 \,\middle|\, \mathbf{x}_k\right) = p(\mathbf{x}_k^t \boldsymbol{\lambda}) \quad \forall k \in U$$

For example, the model can be of the logistic type, where:

$$p(\mathbf{x}_k^t \boldsymbol{\lambda}) = \exp(\mathbf{x}_k^t \boldsymbol{\lambda}) / [\mathbf{1} + \exp(\mathbf{x}_k^t \boldsymbol{\lambda})]$$

The model can be fitted using pseudo-maximum likelihood as available, e.g. in the *survey* package from **R** – Lumley (2010).

Combining the Register with a Probability Sample

Given the fitted model, then estimate the population mean from the register using:

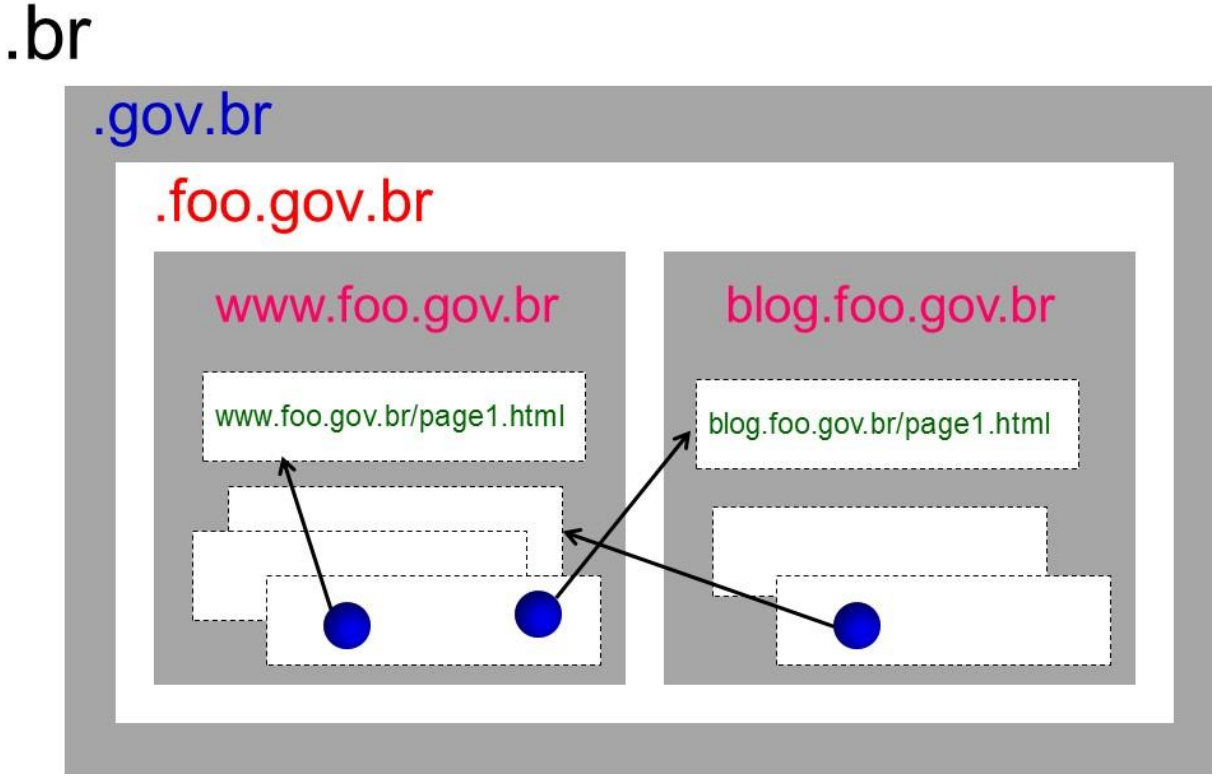
$$\hat{\theta} = \sum_{k \in \mathcal{R}} w_k y_k$$

with $w_k = \left[p(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}})\right]^{-1} / \sum_{k \in \mathcal{R}} \left[p(\mathbf{x}_k^t \hat{\boldsymbol{\lambda}})\right]^{-1}$.

Kim and Wang (2018) also propose a 'doubly robust' estimator by further modelling the response y as a function of the covariates **x**.

Big Data: An Example

Goal: obtain data on Brazilian internet domains.



Big Data: An Example

Pilot project: ".gov.br"

Census of sites and pages within domains registered by branches of the Brazilian government (public sector).

'Robot' visited ALL sites and pages found or connected to domains ending with ".gov.br".

Initial set contained close to 12 K registered domains.

Data collection took around **3 weeks**.

Big Data: An Example

Challenge: ".com.br".

At the time of the exercise, 2.5 million registered domains ending with ".com.br".

Census would be infeasible given technology available ➔ data collection would last ≈ **11 years**.

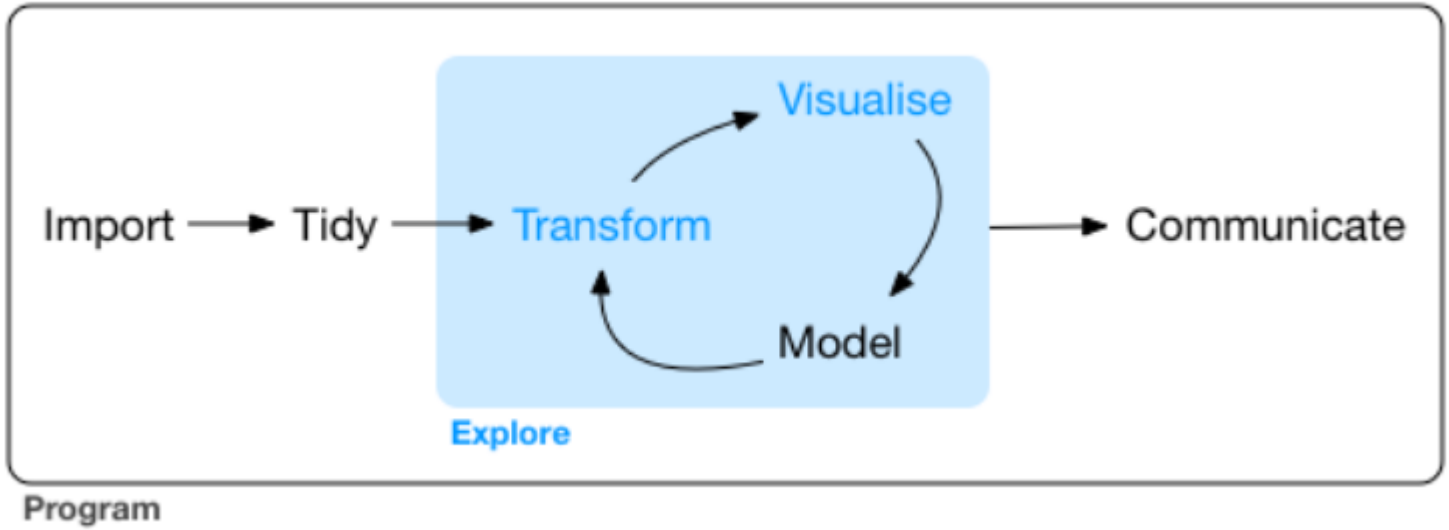
Approach adopted: stratified and clustered sample of n ≈ 4,000 domains ending with ".com.br".

Data collection took around **45 days**.

Source: Silva e Vasconcellos (2011).

Data Science Activities

The activities of DS according to Wickham & Grolemund (2017).



Challenge: no 'concern' for where the data come from or how they were obtained!

Statistical Science: 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, etc.)
- Data checking, cleaning, storage and sharing protocols.

Statistical Science: Analysis / Discovery from Data

Methods for exploratory and confirmatory data analysis:

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

Summarizing

Data quality remains fundamental concern.

Statistical thinking & methodology (**Statistical Science**) remain essential pillars for promoting data quality, as well as good decisions based on data.

Data era will require more statistical development, not less:

- In the past, small n & small p;
- With Big Data, large n or large p or both!

Statistical thinking central to Data Science & Big Data!

References

Cázarez-Grageda, K.; Zougbede, K. (2019). National SDG Review: data challenges and opportunities. Bonn, Germany: PARIS21.

He, X. et al. (2019). Statistics at A Crossroads: Who Is for The Challenge? National Science Foundation, 2019.

IBGE (2013). Código de Boas Práticas das Estatísticas do IBGE. Rio de Janeiro: IBGE.

International Monetary Fund. (2012). Data Quality Assessment Framework - Generic Framework.

Kim, J. K., Wang, Z. (2018). Sampling techniques for big data analysis. Int. Stat. Rev. 1–15.

Lumley, T. (2010). Complex Surveys: A Guide to Analysis Using R. Hoboken: John Wiley & Sons, 276 p.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Annals of Applied Statistics v. 12, n. 2, p. 685–726.

Pfeffermann, D. (2017). Bayes-based Non-Bayesian Inference on Finite Populations from Non-representative Samples: A Unified Approach. Calcutta Stat Assoc Bull, 69(1): 35–63.

Silva, P. L. d. N.; Vasconcellos, M. T. L. D. (2011). Plano amostral para a pesquisa exploratória TICWEB do .com. Rio de Janeiro: MVTL Soluções em Tecnologia Ltda.

Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Ottawa, Canada: Statistics Canada.

Statistics Directorate, OECD. (2012). Quality Framework and Guidelines for OECD Statistical Activities.

Stigler, S. M. (2016). The Seven Pillars of Statistical Wisdom. Harvard University Press.

UNECE Big Data Quality Task Team. (2014). A Suggested Framework for the Quality of Big Data Deliverables of the UNECE Big Data Quality. Geneva.

United Nations (2005). Household sample surveys in developing and transition countries. Vol. F No. 96, Studies in Methods. New York: United Nations. 617 p. http://unstats.un.org/unsd/hhsurveys/sectione_new.htm

United Nations. (2012). Guidelines For The Template For A Generic National Quality Assurance Framework (NQAF). http://unstats.un.org/unsd/dnss/qualityNQAF/nqaf.aspx

US Office of Management and Budget. (2006). Standards and Guidelines for Statistical Surveys. Federal Register. Washington, DC.

Wickham, H.; Grolemund, G. R for Data Science: Import, Tidy, Transform, Visualize and Model Data. Boston: O'Reilly, 2017. 492 p.