Can Big Data Really Replace Traditional Surveys for the Production of Official Statistics

Danny Pfeffermann

Central Bureau of Statistics, Israel

DataHack- October 2018

What is official statistics? (Encyclopedia.com)

Information collated, processed and disseminated by **national governments** and international bodies **which** link to them.

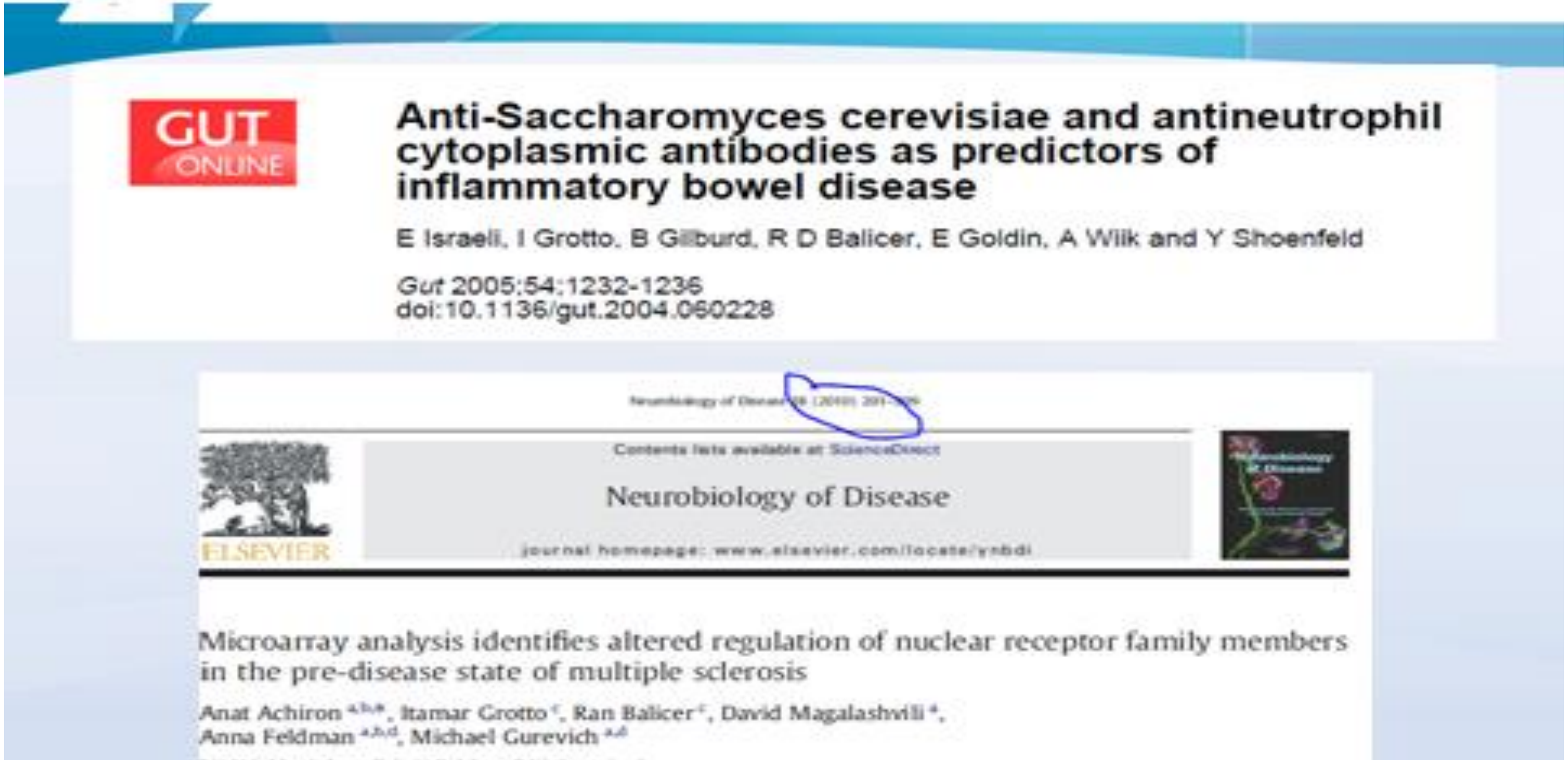
These data are almost invariably **nationally representative**, conforming to **international definitions and classifications** or other well-established conventions.

Official statistics stands in sharp contrast to statistics and data-sets from other sources: academic research, market research, research institutes, commercial organizations, local, regional, and state bodies.

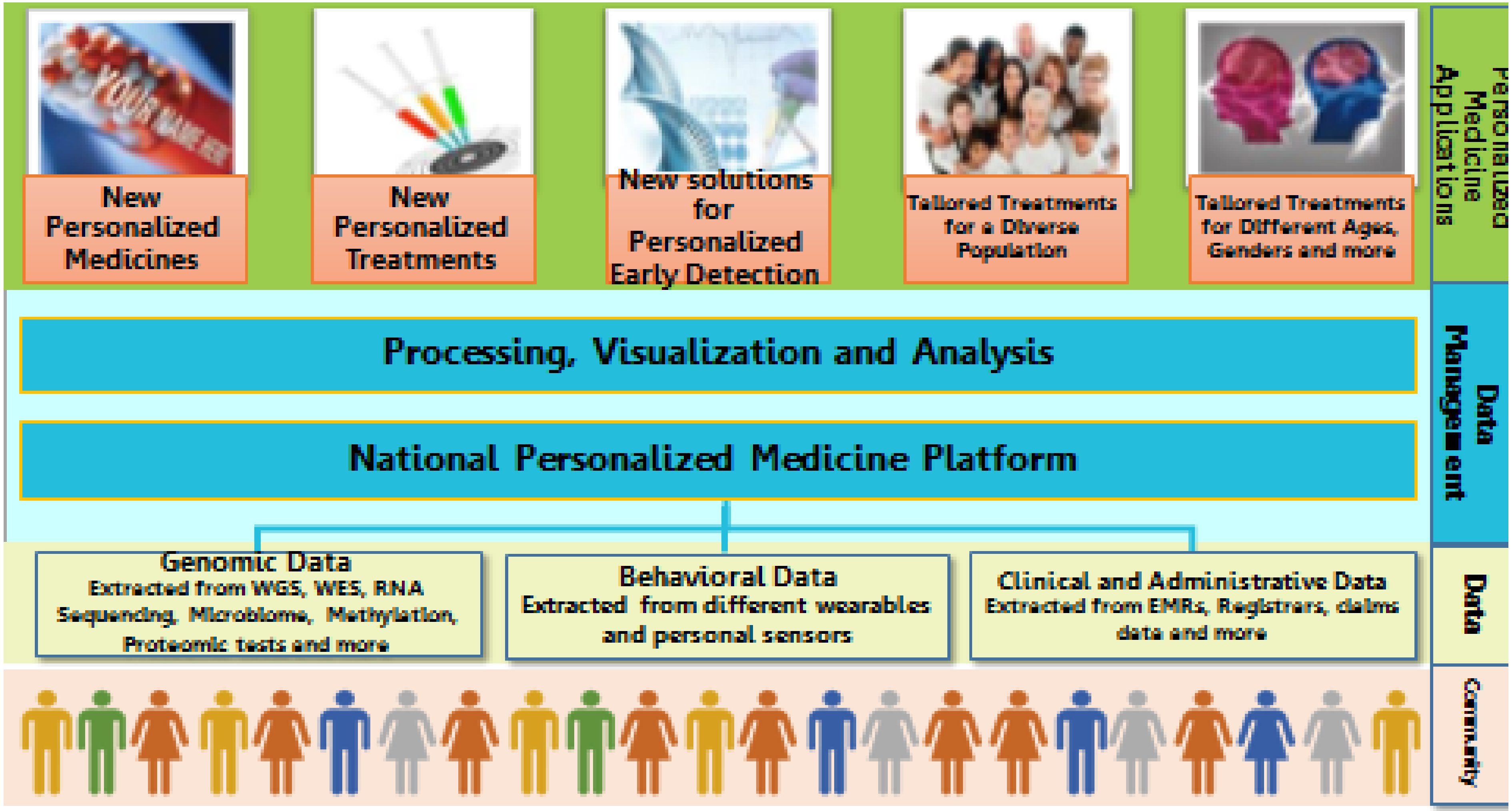
Do big Data fulfil the requirements from official statistics?

Big Data for research (not necessarily relevant for OS)

Applying “Big-Data” to predict diseases



Dream data also for official statistics


New Personalized Medicines
New Personalized Treatments
New solutions for Personalized Early Detection
Tailored Treatments for a Diverse Population
Tailored Treatments for Different Ages, Genders and more
Personalized Medicine Applications
Processing, Visualization and Analysis
National Personalized Medicine Platform
Data Management
Genomic Data
Extracted from WGS, WES, RNA Sequencing, Microbiome, Methylation, Proteomic tests and more
Behavioral Data
Extracted from different wearables and personal sensors
Clinical and Administrative Data
Extracted from EMRs, Registrars, claims data and more
Data
Community

Using Big Data for Official Statistics

Janusz Dygaszewicz, **Central Statistical Office of Poland**

"Currently official statistics are based on data from **state registers** and information obtained from **surveys;**

However, the world is continually changing; there are new phenomena which also require describing with statistical processes. Therefore, it cannot be limited only to the **old data** sources; you need to constantly seek new paths and solutions. The global trend in this field is **Big Data**.

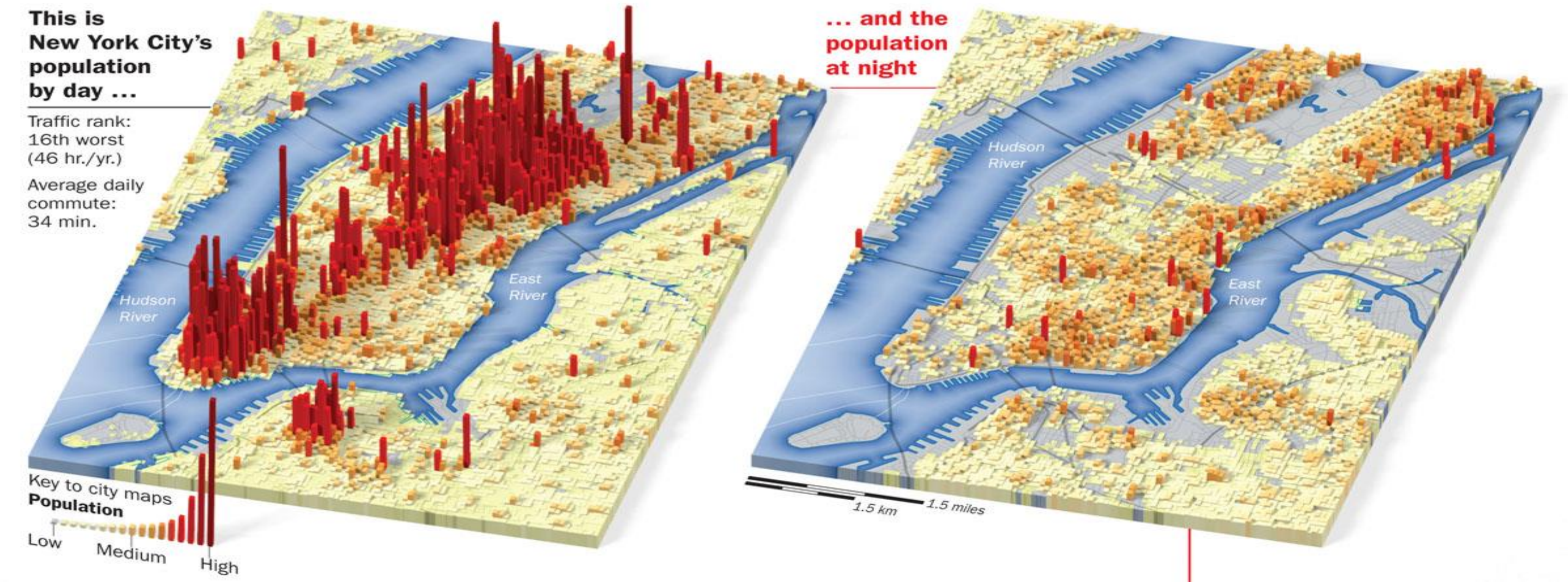
The Potential of Big Data (Janusz Dygaszewicz)

If **"what we know"** represents what we currently produce as **official statistics,** the proportions in the picture **are not right**.

Is the "rest" really needed for official statistics?

Couldn't we estimate it from administrative data & surveys?

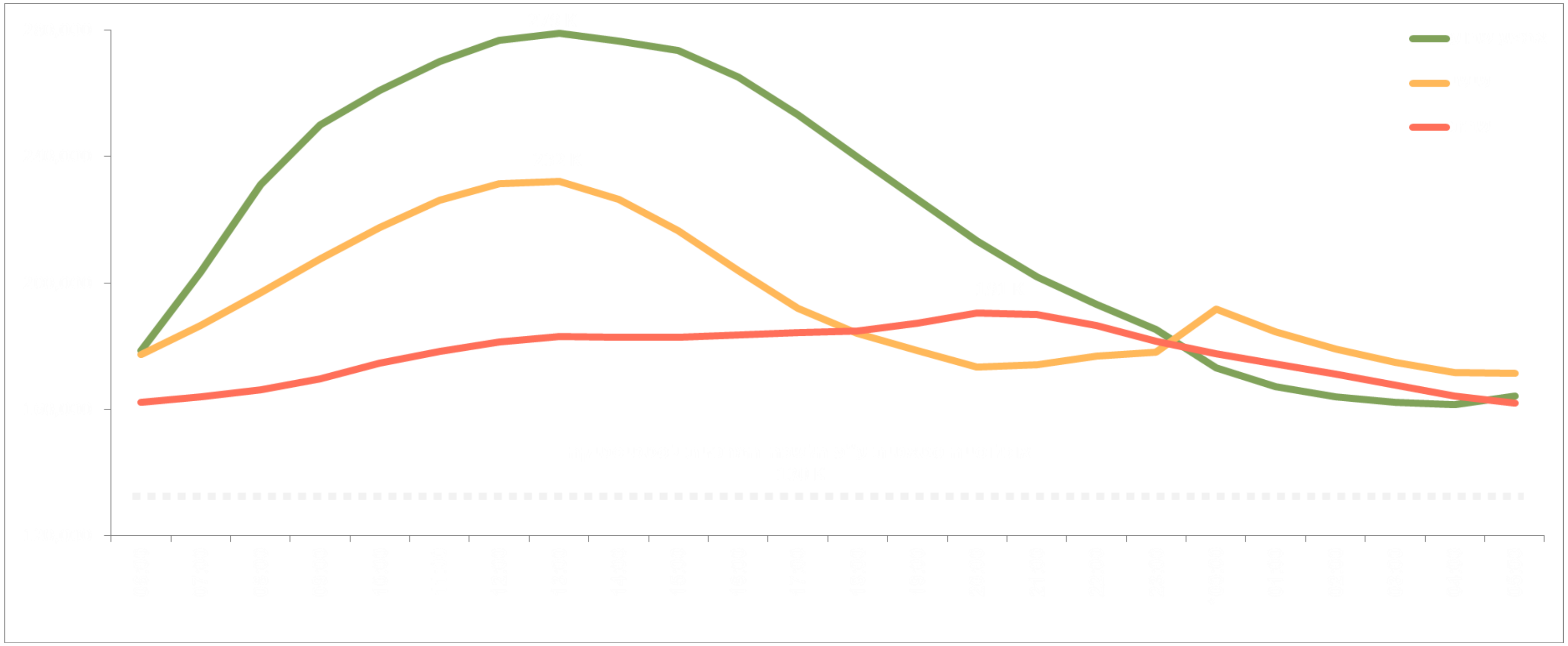
Location data from mobile phones



Practically Impossible to get this information from surveys for every location. Will telephone companies provide us this information? May be. But how will we learn, for example, about the "purpose of the trip" or type of transportation?

לב ת"א – פרופיל יוממות כללי

365 יום בשנה, 24 שעות ביממה, 146 תאי שטח



06:00 13:00 00:00 05:00

--- אמצע שבוע --- שישי --- שבת

Big Data for OS → Big Problems → Big headache

- High dimensionality and extremely large volumes of data.
- Coverage/selection bias (we are talking of **OS**).
- Data accessibility, public permission (**ethics**) new legislation**?**
- Privacy (data protection)**;** disclosure control.
- New sampling algorithms.
- Data storage.
- Heavy computation, new algorithms and analytic tools.
- Integration of files from multiple sources at different times.
- Risks of data manipulation, sudden unavailability, **high costs**.
- Need to train/hire highly skilled experts (**data scientists!!!**)

Do we really get what we need for our official statistics??

Two types of big data

Type 1. Data obtained from sensors, cameras, cell phones**…,** generally structured, accurate, relates to a particular population.

Type 2. Data obtained from social networks, e-commerce**…,** generally diverse, unstructured and appears irregularly.

- Data from different sources may have **different formats**, arrive at **different times** with different degree of reliability, and may be **defined differently**.
- **No such problems with traditional surveys!!**
- **NSOs** need to be prepared that data may **cease to exist**.

Big data is a by-product, not produced for OS purposes.

Other important issues

Coverage (measurement) bias- major concern in use of big data for **OS**.

House sales advertised on the internet do not represent properly all house sales. Opinions expressed in **social networks** may not represent the opinions held by the **general public**.

- **Big not always better!!** Collecting huge amounts of data does not guarantee getting right answers. A small **balanced sample** may provide better insights than **large skewed data**.

Example of bias (measurement errors) in "Big Data"

Population census in Israel

Main purpose: measure the number of residents in each of about 3,000 statistical regions. (~3,000 persons per region).

The Israel **population register** fairly accurate at the **national level**. **Much less accurate** at the small statistical region level, with an average **address error** of about **13%** $\Rightarrow$ relying solely on the register would result in large errors for at least some regions.

Requires special big samples to correct the bias.

The problem will be resolved in the long run by extending and improving the administrative data.

Coverage bias (cont.)

No bias when using big data as **predictors** of other variables.

Examples: Use **BPP** (**5 million commodities**) sold **online** to predict the **CPI**, which **requires two costly surveys;** use **job advertisements** to predict **employment;** use **Satellite images** to predict **crops**.

Requires proper statistical analysis to identify and test (routinely) the prediction models.

Other important issues (cont.)

Sampling: random sampling will continue to play a major role in the era of big data.

✓ Reduces **storage space**, helps protecting **privacy** and **disclosure**, produces **manageable** data sets on which algorithms can run to **fit models** and produce **estimates**.

✓ Sampling from **big, versatile dynamic** data **different** from sampling **finite populations,** requiring **new** sampling algorithms; **e.g.,** sampling from **social networks,** Sampling of **time points…**

✓ If no sampling $\Longrightarrow$ **no sampling errors**.

Which **quality measures** should be computed**? bias? How? Compare to traditional estimates? Measurement errors?**

Other important issues (cont.)

Big Data for sub-populations: NSOs publish estimates for **sub-populations; age**, **gender**, **ethnicity**, **geography,…**

Big data may not contain this information. Requires **massive linkage if** missing information available in other big files.

Data on **sales from supermarkets** contains **no information** on buyers $\Rightarrow$ cannot compare consumption patterns (or types of commodities) of different buyers.

Possible solution: Link sales to buyers by use of **credit card numbers**. **Will credit card companies provide them?**

✓ **Will traditional sample surveys always be needed?**

Other important issues (cont.)

Estimation: at **NSOs** we use **design-based** estimators**, model dependent** estimators**, model-assisted** estimators**,…**

New: ***algorithmic estimators* -** the result of computational algorithms applied to the raw **big data**.

Example: measure of degree of **religiosity (Israel CBS)**. Required merging **12 administrative files with population register** and then apply a complex hierarchical algorithm.

Publication: Big data potentially available for every point in time. **What kind of statistics should be computed and published?** Should official publications from big data be primarily in the form of (online) **graphs and pictures (like currency rates)?**

Computer engineering for OS from big data

No longer **Gigabytes** (~10^9 bytes). **Terabyte** (~10^{12} bytes), **petabyte** (~10^{15} bytes) **&** **Exabyte** (~10^{18} bytes) **New standard**.

- Available computing facilities at **NSOs** cannot store and handle such huge volumes of data.

Possible solution: Use **cloud** storage, management and processing facilities (**Amazon, Microsoft, Israel government?**)

Potential problems with **Data protection**. **Many users**, data distributed over a **large number of processors**.

Other solution: Data centre. Incorporate **all local computers; central management** of storage space **&** processing power of separate servers. **Major challenge**.

Big data for OS- summary remarks

New expensive computing facilities, **new** data processing techniques, **new** linkage methods, **new** visualization methods, **new** sampling methods, **new** analytic methods, **new** measures of error, **new** disclosure control procedures, **new** legislation, **new** types of employees (**data scientists**)**,…**

Big potential advantages: Much more different data sources, timeliness, broader coverage (but possible **coverage bias**), **no** need for sampling frames, **no** questionnaires, **no** interviewers**,…**

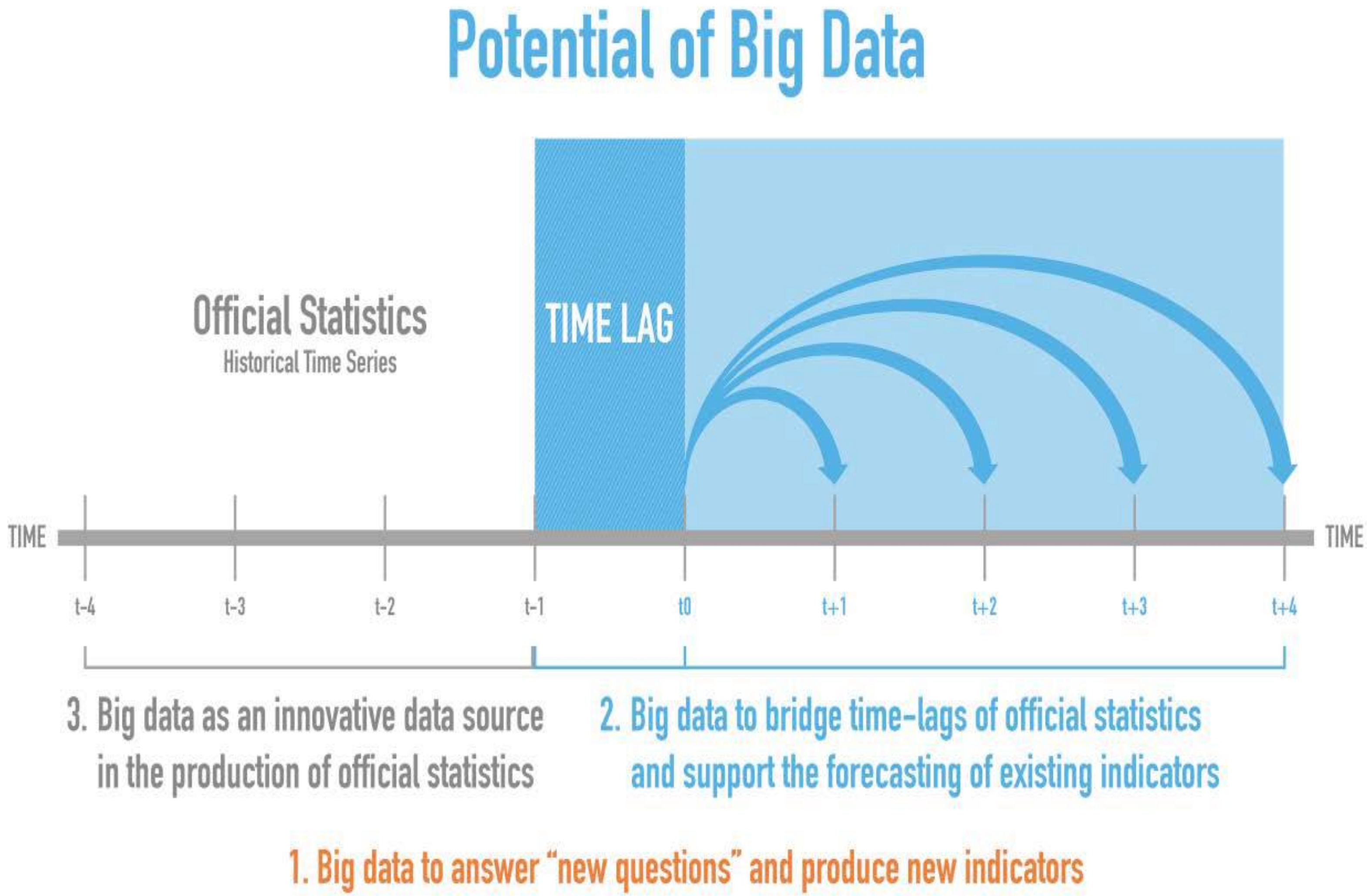
✓ Constant **decline in response rates** in traditional surveys and tightened budgets $\Rightarrow$ **future use of big data inevitable**.

Summary remarks (cont.)

Big data is not a static, but a **dynamic phenomenon**. The systems and networks generating it will continue to evolve and with it the opportunities that big data offers, the challenges that if poses and its statistical implications.

Big data can address new questions and produce new indicators (SDG?, vacant jobs through web scraping?)

However, statistical support will form the key for the evaluation and validation of big data products, their suitability and methodological soundness.


Potential of Big Data
Official Statistics
Historical Time Series
TIME LAG
TIME
TIME
t-4
t-3
t-2
t-1
t0
t+1
t+2
t+3
t+4
3. Big data as an innovative data source in the production of official statistics
2. Big data to bridge time-lags of official statistics and support the forecasting of existing indicators
1. Big data to answer "new questions" and produce new indicators

New challenge for production of traditional OS (Marker 2017)

For 100 years, **NSOs** designed their data collection based on efficient representative samples with good coverage and high response rates $\Rightarrow$ reliable estimates and measures of accuracy.

In the modern era, **big data** are available for which little is known on how they are collected and how **representative** they are.

But the existence of big data has changed the expectation of **timeliness** of data. **NSOs** must figure out how to carry out surveys and censuses **quicker**, or users will rely on **big data** without understanding **what they are losing**.

Accounting for possible coverage bias of big data

Big data may not represent properly the target population!!

An example where the **"sample"** (**big data**) is not representative of the target population.

Other examples: Informative sampling, **NMAR nonresponse..**

Kim (2017) proposes **3** different (possibly combined) procedures to account for **non-representativeness** of big data:

(Stratified balanced) **Reservoir sampling,**
Inverse sampling,
Survey integration.

Survey integration, combine big data with survey data

Basic assumption: Membership of **sample elements** in big data **(B)** set **known**. (Ask the sample members**?**)

Let $\delta_i = 1(0)$ if $i \in B (i \notin B)$. Denote the target variable by ***Y***.

Sample data: $\{(\mathrm{x}_i, z_i, \delta_i); i = 1, ..., n\}$**;**

$\mathbf{x}_i$ = model covariates**,** $\boldsymbol{z_i}$ =variables explaining **B-**membership.

Procedure: Model $\pi_i = \Pr(\delta_i = 1 | \mathrm{x}_i, z_i)$ from sample data$\Rightarrow \hat{\boldsymbol{\pi}}_i$

Use $\boldsymbol{w_i = (1 / \hat{\pi}_i)}$ as weights for inference on **finite population**.

Remarks on proposed procedure

Neat idea but with important limitations:

Assumes existence of a sample with required data

Assumes knowledge of **sample elements'** membership in **B**

Assumes knowledge of variables **z** explaining **B- membership**

Assumes $\Pr(\delta_i = 1 | x_i, z_i, \boldsymbol{y_i}) = \Pr(\delta_i = 1 | x_i, z_i)$.

(**"noninformative sampling"**).

In what follows I outline an alternative procedure based on **Bayes theorem**, which overcomes these limitations (but as everything else in life, **no free lunch**).

Thomas Bayes (1701–1761)

Major impact on probability theory and statistics


ROYAL TUNBRIDGE WELLS
THOMAS BAYES
1702 - 1761
Nonconformist minister
and mathematician
Originator of the statistical
theory of probability, the basis
of most market research and
opinion poll techniques
lived here
1731 - 1761
FOURTH CENTENARY

Bayes Theorem

For Y, C random variables,

$$\Pr(Y = y \mid C = c) = \frac{\Pr(C = c \mid Y = y) \times \Pr(Y = y)}{\sum_j \Pr(C = c \mid Y = y_j) \times \Pr(Y = y_j)},$$

$$f_Y(y \mid C = c) = \frac{\Pr(C = c \mid Y = y)\, f_Y(y)}{\Pr(C = c)} = \frac{\Pr(C = c \mid Y = y)\, f_Y(y)}{\int \Pr(C = c \mid \tilde{y})\, f_Y(\tilde{y}) d\tilde{y}}.$$

C = conditioning variable **indicating sample membership**.

❖ In what follows I assume that the big data are a **"sample"** from the **target population**.

Alternative procedure to account for coverage bias of BD

Population model: $f_p(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for target population outcomes (**census model**),

Big data (B) model: $f_B(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for **B** data.

Denote, as before, $\delta_i = 1(0)$ if $i \in B (i \notin B)$.

$$f_B(y_i \mid \mathrm{x}_i) \overset{def}{=} f(y_i \mid \mathrm{x}_i, \delta_i = 1) \overset{\text{Bayes}}{=} \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

$$\Downarrow$$

$$f_B(y \mid \mathrm{x}_i) = f_p(y \mid \mathrm{x}_i) \text{ **iff** } \Pr(\delta_i = 1 \mid y_i, \mathrm{x}_i) = \Pr(\delta_i = 1 \mid \mathrm{x}_i) \, \forall y_i. \quad \text{(**)}$$

If (**) satisfied, feel free to use B to analyse population data.

Alternative procedure (cont.)

$$f_{B}(y_i \mid \mathrm{x}_i) = \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)\, f_{p}(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}.$$

Target *pdf* is $f_p(y \mid \mathrm{x})$**;** observations only available from $f_B(y \mid \mathrm{x})$. The two distributions connected via **probability link function** $\Pr(\delta \mid y, \mathrm{x})$**;** enables estimating **target population pdf** from observations obtained for **Big data.**

- $f_B(y_i \mid \mathrm{x}_i)$ can be estimated from **B** (or **sample** thereof).

- $\Pr(\delta_i = 1)$ allowed to depend on target variable, **y**. May depend also, or only, on other variable ***z***, but **only need** to model $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$.

Alternative procedure (cont.)

$$f_{B}(y_i \mid \mathrm{x}_i) = \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)\, f_{p}(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

- Inference requires modelling $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ and possibly $f_{p}(y_i \mid \mathrm{x}_i)$, but **no survey data required**.

- Models assumed for $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ and $f_{p}(y_i \mid \mathrm{x}_i)$ **testable** by testing the implied model for $f_{B}(y_i \mid \mathrm{x}_i)$, using **conventional model testing** procedures, since the big data are **known**.

Concluding remarks

- ✓ Use of big data for **OS** is **not straightforward** and requires overcoming many legal, ethical and computational problems **+** development of new methodologies.
- ✓ But use of big data for official statistics is **inevitable** and promises huge possibilities, which cannot be ignored.
- ✓ Under-coverage of big data is a major concern in their use.
- ✓ **Kim (2017)** procedures and the procedure outlined in this presentation are only **first (**but promising) steps to deal with the **undercoverage** problem.
- ✓ Much more theoretical and applied research required.