The implication of Big Data for Official Statistics

Diego Zardetto

Istat

THE CONTRACTOR IS ACTING UNDER A FRAMEWORK CONTRACT CONCLUDED WITH THE COMMISSION

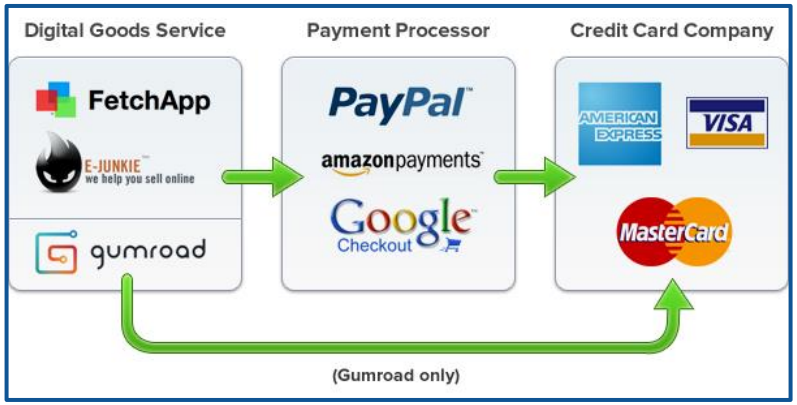
Data Deluge/Big Data (1/3)

UNECE tentative taxonomy

1. Human-sourced information

- Social Networks (Facebook, Twitter, LinkedIn, Pinterest, Tumblr, …)
- Blogs and posted comments
- Pictures (Instagram, Flickr, Picasa, …)
- Videos (Youtube, …)
- Search engine queries
- Mobile data content (text messages, …)
- User-generated maps
- E-Mails
- …

Data Deluge/Big Data (2/3)

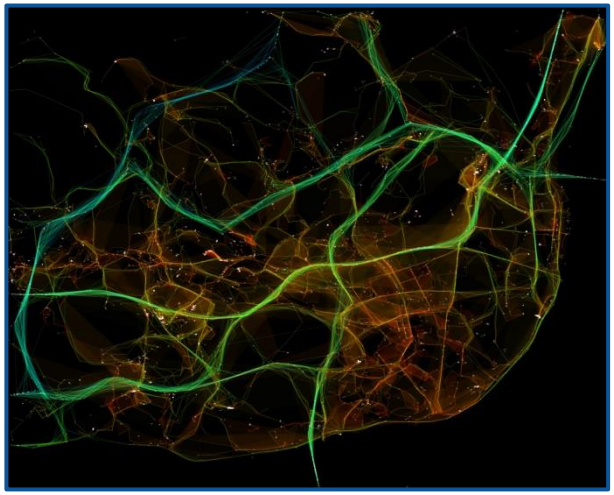


UNECE tentative taxonomy

2. Process-mediated/transaction data

- Commercial transactions
- Banking/stock prices records
- E-commerce
- Telephone Call Detail Records
- Credit cards
- Medical records from Public Health
- …

Data Deluge/Big Data (3/3)



UNECE tentative taxonomy

3. Machine generated data (Internet of things)

- Sensor data
  - ✓ Weather/pollution sensors
  - ✓ Traffic sensors/webcam
  - ✓ Security/surveillance videos/images
  - ✓ …
- Tracking devices
  - ✓ GPS systems
  - ✓ Mobile phone location
  - ✓ Satellite images
  - ✓ …
- Data from computer systems
  - ✓ Logs & Web logs
  - ✓ …

Why do Big Data look so appealing to NSIs?

Possible answer(s): the reactive side

- **Competitive pressure**
  - Private sector may take advantage of Big Data and produce more and more statistics that attempt to beat official statistics on timeliness and relevance
    - ✓ The "Official Statistics" trademark could slowly lose reputation and relevance unless NSIs get on board
- **Funding constraints**
  - Economic crisis (2009-20??) urges organizations to look for ways to increase efficiency and cut costs
    - ✓ Being traditional data collection so cost-intensive, interest in alternative data sources and Big Data is growing

Why do Big Data look so appealing to NSIs?

Possible answer(s): the active side

- **Improving quality of traditional statistics**
  - Providing new auxiliary information that NSIs could exploit to
    - ✓ Build and maintain better sampling frames
    - ✓ Design better samples
    - ✓ Build better Calibration estimators
    - ✓ Soften nonresponse bias further
- **Reducing respondents' burden**
- **Potential for discovering new knowledge**
    - ✓ New well-being indicators
    - ✓ Agriculture and environment statistics
    - ✓ New measures of consumers' confidence
    - ✓ Consumer behavior beyond HBS

Inference in the Official Statistics Realm

- **Outline**
- Official Statistics traditional paradigm
  - Top-down: data are planned
  - Traditional inference approaches
    - ✓ Design based survey sampling theory
    - ✓ Model-assisted approach
    - ✓ Model based inference
- The need of a new paradigm to deal with Big Data
  - Bottom-up: data are already there
  - Exploratory analysis / Knowledge discovery approach
    - ✓ Algorithmic inference: data mining techniques, machine learning, …
    - ✓ Big Data bring plenty of methodological issues and pitfalls

Official Statistics traditional approach (1/2)

- Information needs / Hypotheses
- Design data collection
- Collect data
- Prepare data
- Analyze data
- Obtain information / Confirm or reject hypotheses

Official Statistics traditional approach (2/2)

- **Top-down paradigm**
- Emphasis on
  - **Planning the data** to be later analyzed since the beginning
    - ✓ Target population, Units
    - ✓ Variables, Definitions, Classifications, Questionnaires
    - ✓ Lists and registers to reach units
    - ✓ Methods to select units (randomization), …
  - **Targeting analysis** to specific information needs / hypotheses
    - ✓ Model interpretability perceived as a must
    - ✓ Statisticians always aspire to understand **"how"** something is going on, sometimes even to guess **"why"** it is going on
  - **Using probability theory** as a firm ground to achieve rigorous results in estimation/prediction

The Big Data Paradigm Shift (1/2)

- Data are already here (and everywhere)
- Collect data
- Prepare data
- Explore data (seeking for correlations)
- Tune algorithms
- Discover new knowledge / Validate results

The Big Data Paradigm Shift (2/2)

- **Bottom-up paradigm**
- Emphasis on
  - **Exploring available data**, seeking information value that has not been extracted so far
  - **Trusting the BIG data corpus**
    - ✓ Data tend to be perceived as objective, and discovered correlations as well
    - ✓ Interpretability of mining algorithms is not deemed mandatory
    - ✓ Data scientists seem to be mainly interested in **"what"** is going on, less (or even not at all) on "how" or "why" something is going on
  - **Selecting algorithms based on scalability**
    - ✓ The "Data Exhaust" way: since data could hide valuable insight at all granularities, avoid data aggregation (if feasible)
  - **Using heuristic** techniques for estimation/prediction
    - ✓ Due to the huge data volume often there is no other feasible alternative

Big Data: Methodological Pitfalls

- **Outline**
  - Representativeness (w.r.t. the desired target population)
    - ✓ Selection Bias
    - ✓ Actual target population unknown
    - ✓ Often sample units' identity unclear/fuzzy
    - ✓ Pre processing errors (acting like measurement errors in surveys)
    - ✓ Social media & sentiment analysis: pointless babble & social bots
  - Ubiquitous correlations
    - ✓ Causation fallacy
    - ✓ Spurious correlations
  - Structural break in Nowcasting Algorithms

Big Data vs Primary & Secondary Data Sources

	Primary Sources (e.g. Censuses, Surveys)	Secondary Sources (e.g. Administrative Data)	Tertiary Sources (e.g. Big Data)
Data are designed to be used in statistical production	yes	no	no
Concepts, definitions and classification are stated and known	yes	often	rarely
Target (sub-)population is defined	yes	often	no
Metadata available	yes	often	no
Data are structured	yes	yes	rarely
Data refer to units of the population of interest	yes	usually	no
Data need "heavy" preprocessing to be used in statistical production	no	no	yes
Interest variables are directly available	yes	yes	no
Auxiliary variables are directly available	yes	often	no
Data cover target (sub-)population	yes (census) no (surveys)	often	not yet
Data are representative (or lack of representativeness is intentional and/or can be adjusted for in analyses)	yes	often	no
Data values are "clean"	no	sometimes	rarely

Big Data: why traditional inference methods cannot succeed

- The computational complexity barrier
  - Examples: Matrix inversion (ubiquitous: least squares estimators, GLM maximum-likelihood via Newton-Raphson algorithm) → O(n^3)
  - Most traditional algorithm difficult to parallelize (for achieving Hadoop / MapReduce scalability)
  - …
- Extreme sensitivity to erroneous data / outliers
  - Big data are noisy and unstructured
    - ✓ But due to huge volume cannot apply thorough procedures for Editing & Imputation / Outlier detection

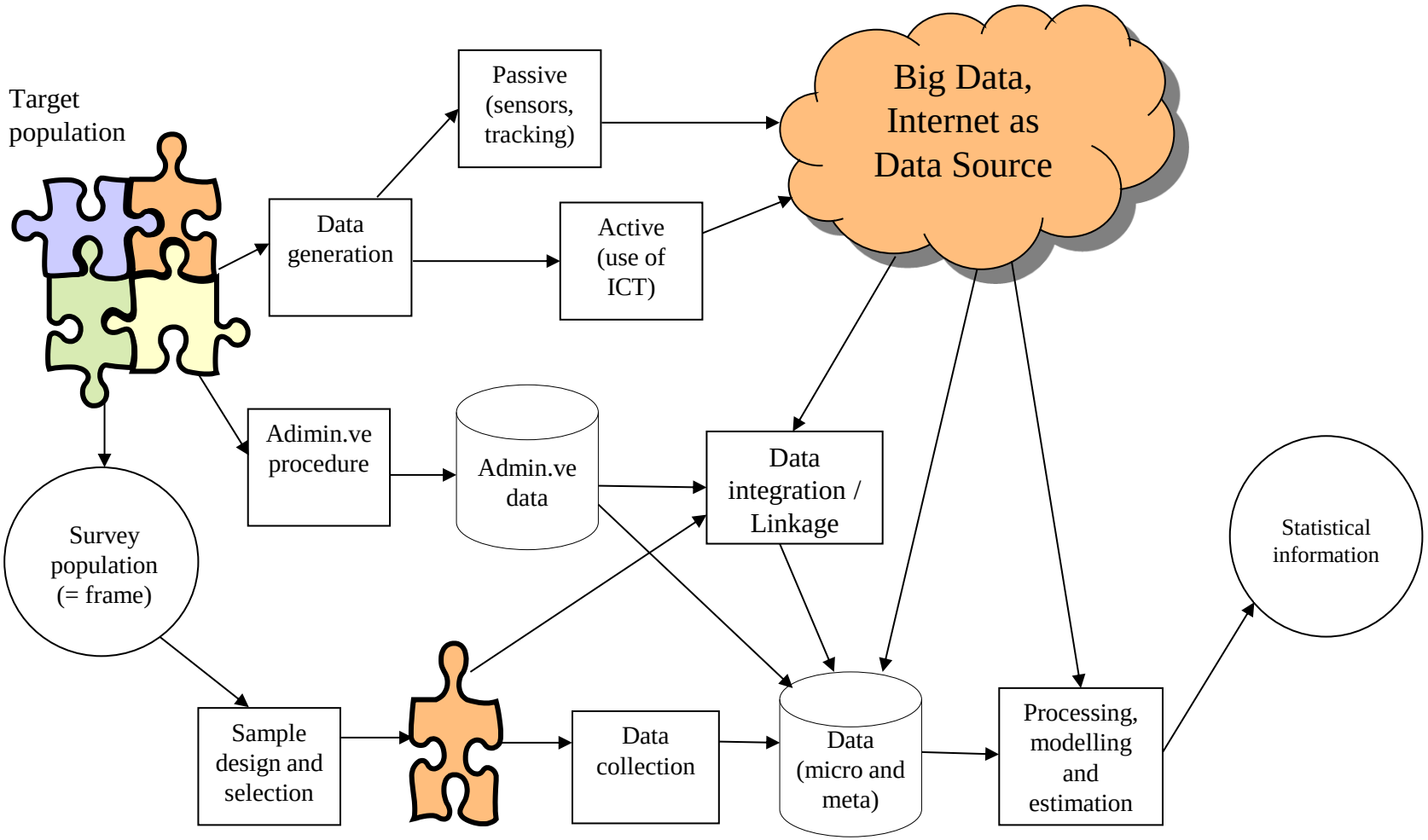
Implications (1/2)

- Current methods in Official Statistics (e.g. design based and model assisted survey sampling theory, regression theory, generalized linear models, small area estimation methods, …) hinge upon specific features of NSI' traditional data, namely
  - **small** amounts of **high quality** data
- These methods:
  - are extremely **sensitive** to **outliers** and **erroneous data** (which explains the tremendous effort put by NSIs in data checking and cleaning activities)
  - typically exhibit **high computational complexity** (power-behavior is the rule, a feature that hinders their scalability on huge amounts of data)
- **Synthesis**: NSIs' statistical methods and Big Data are poles apart, at present
- **Diagnosis**: in order to let Big Data gain ground in Official Statistics, NSIs will have to undertake some radical paradigm shift in statistical methodology

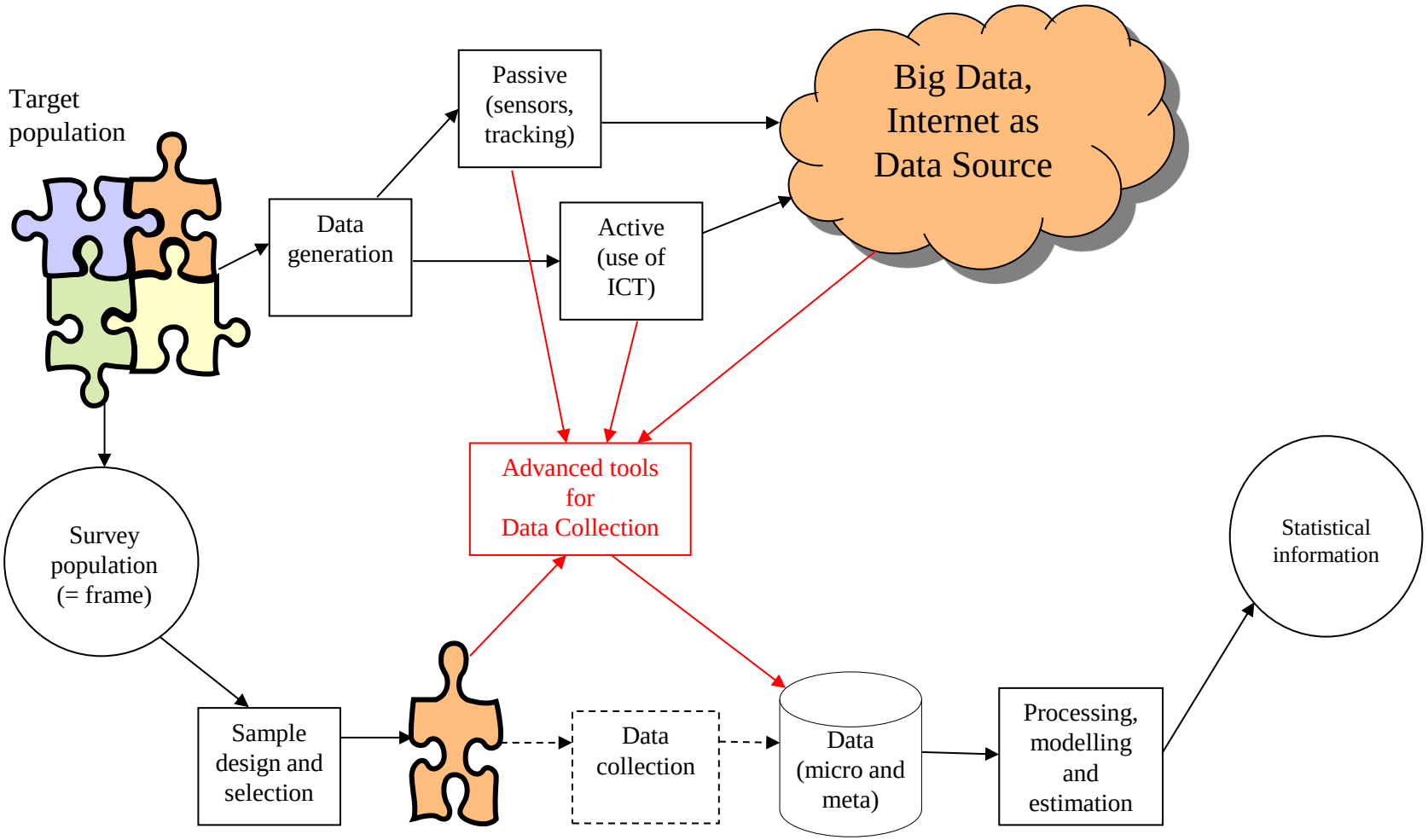
Implications (2/2)

- Despite it is **far from obvious** how to translate such awareness into actual proposals, we deem new candidate methods should be:
  1. more robust (i.e. more tolerant towards both erroneous data and departures from model assumptions), perhaps at the price of some accuracy loss
  2. less demanding in terms of a clear and complete understanding of obtained results in the light of an explicit statistical model (think of Artificial Neural Networks, Support Vector Machines, Classification and Regression Trees, Random Forests, …)
  3. based on approximate (rather than exact) optimization techniques, which:
     - ✓ are able to cope with noisy objective functions (as implied by low quality input data)
     - ✓ typically ensure the mandatory scalability requirement inherent in Big Data processing, thanks to their implicit parallelism (think of stochastic metaheuristics like, e.g., Evolutionary Algorithms, Ant Colonies, Swarm Particles, …)

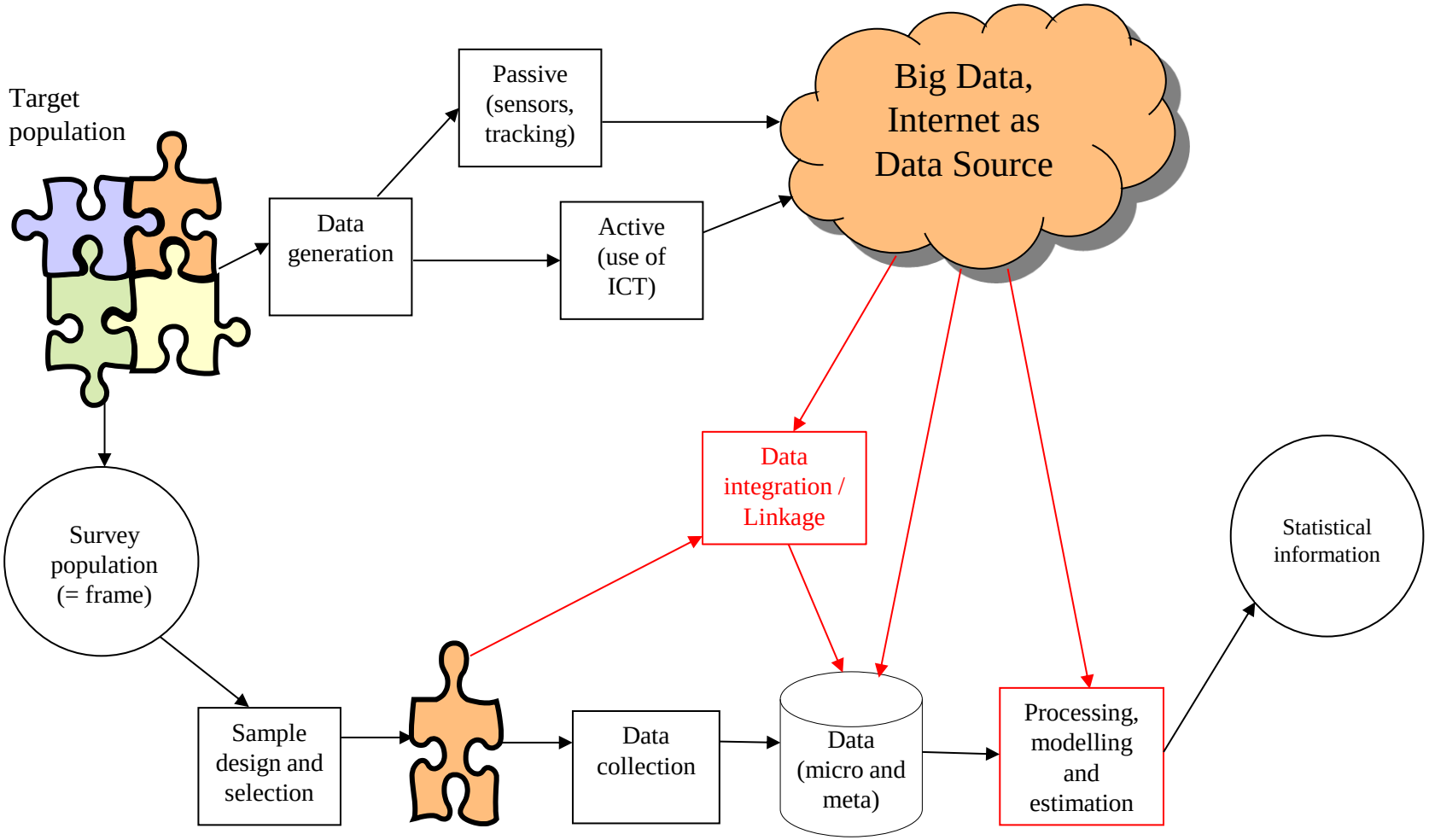
Big Data in Official Statistics: a General Framework



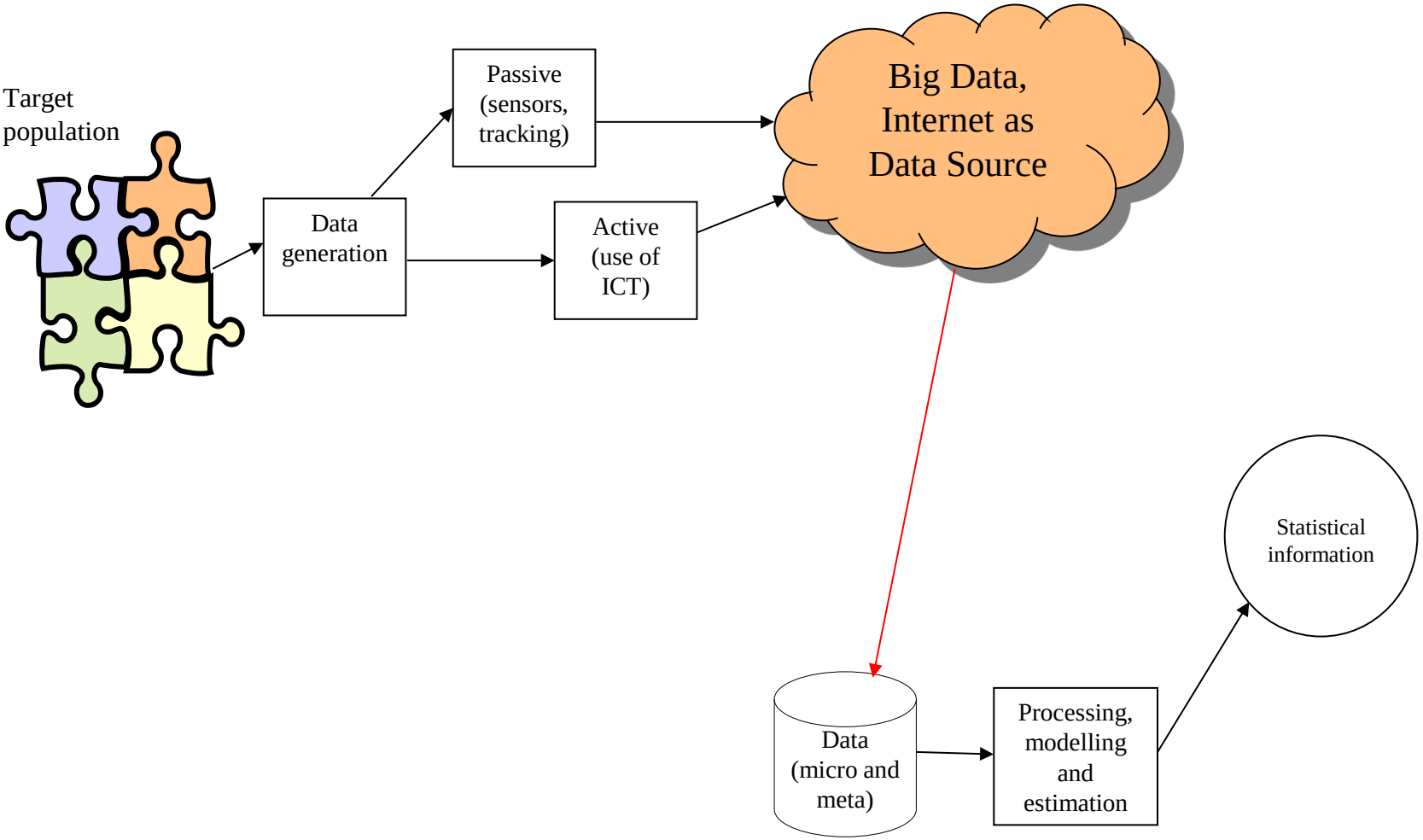
Scenario 1: Alternative Data Collection



Scenario 2: Integrated Use of Big Data and Traditional Data



Scenario 3: Substitution of Traditional Data



Tentative Bibliography (1/2)

- [ECOSOC] "Report of the Global Working Group on Big data for official statistics", United Nations Economic and Social Council, Statistical Commission, session 46, 2015 3-6 Mar
  - http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData.pdf
- [Eurostat] "Big data – an opportunity or a threat to official statistics?", United Nations Economic and Social Council, Statistical Commission, plenary session 62, 2014 9-11 Apr
  - http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/32-Eurostat-Big_Data.pdf
- [Bureau of the Conference of European Statisticians] "In-depth review of big data", Conference of European Statisticians, Paris, 2014 Apr 9-11
  - http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/7-In-depth_review_of_big_data.pdf
- [HLG] "What does Big Data mean for Official Statistics?", United Nations Economic Commission for Europe, 2013 Mar 10
  - http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622
- [ECOSOC] "Big data and modernization of statistical systems", United Nations Economic and Social Council, Statistical Commission, session 45, 2014 4-7 Mar
  - http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf
- [American Association for Public Opinion Research] "AAPOR Report on Big Data", 2015 Feb
  - http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Tentative Bibliography (2/2)

- [P. R. del Castillo] "Reflections on the Use Of Big Data for Statistical Production", CROS portal, 2013 May
  - http://cros-portal.eu/sites/default/files/ReflectionsUseBigDataStatisticalProduction_0.pdf
- [P. Daas et al.] "Big Data as a Source of Statistical Information", The Survey Statistician, 2014 Jan
  - http://isi.cbs.nl/iass/N69.pdf
- [P. Daas et al.] "Big Data and Official Statistics", NTTS conference, Brussels, Belgium, 2013 Mar
  - http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf
- [M. Scannapieco et al.] "Placing Big Data in Official Statistics: A Big Challenge?", NTTS conference, Brussels, Belgium, 2013 Mar
  - http://cros-portal.eu/sites/default/files/NTTS2013fullPaper_214.pdf
- [B. Buelens et al.] "Shifting paradigms in official statistics: from design-based to model-based to algorithmic inference", Discussion paper, Statistics Netherlands, 2012
  - http://www.cbs.nl/NR/rdonlyres/A94F8139-3DEE-45E3-AE38-772F8869DD8C/0/201218x10pub.pdf
- [L. Breiman] "Statistical modeling: The two cultures", Statistical Science, Vol.16, No. 3, 2001
  - http://www.uni-leipzig.de/~strimmer/lab/courses/ss09/current-topics/download/breiman2001.pdf
- [Xiang et al.] "Scalable Matrix Inversion Using MapReduce", HPDC'14 Proceedings of the 23rd international symposium on High-performance parallel and distributed computing, 2014 Jun
  - https://cs.uwaterloo.ca/~ashraf/pubs/hpdc14matrix.pdf