
# Towards a methodological framework for the integration of mobile phone data in the production of official statistics

## DGINS 2018

David Salgado (StatSpain - INE)
Bogdan Oancea (INS)

Bucharest, 10-11 Oct, 2018

# Overview

1. Definition revisited $\rightsquigarrow$ Admin Data
2. **Access** to mobile phone data
   - What data?
   - Preprocessing
   - Obstacles
3. Statistical **methodology**
   - From data generation to the whole process
   - Geolocation of network events
   - Data model
   - Inference
4. IT infrastructure
   - For access
   - For computation
5. **Quality** $\rightsquigarrow$ CoP

# Big Data definition revisited

INSTITUTO NACIONAL DE ESTADISTICA

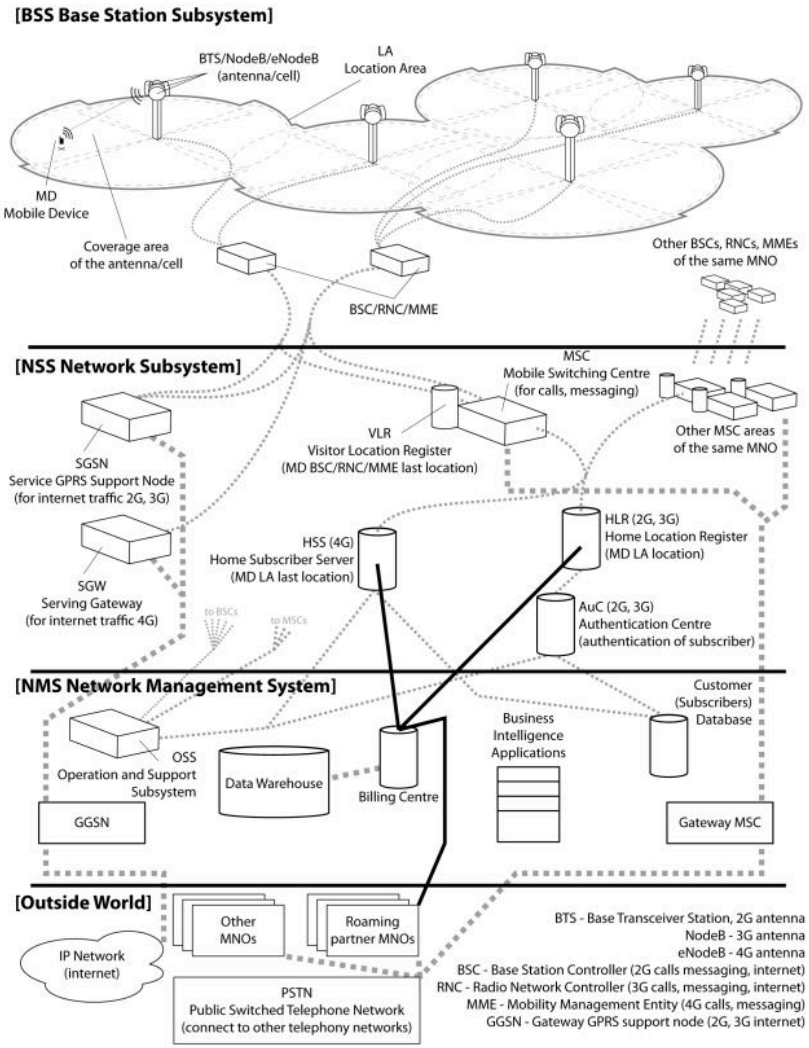
Big Data ***for Official Statistics***

- refer to **third people** and not to data holders;
- are **central in their economic activity**;
- **lack statistical metadata** (since they are generated for very different purposes).

UNECE (wider) Definition for **Admin Data**:

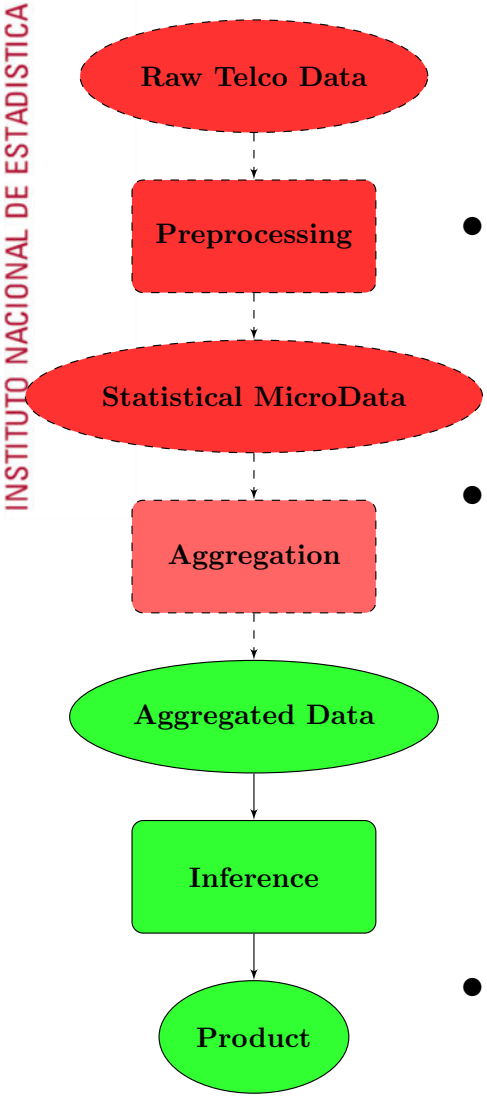
“Data collected by sources **external to statistical offices**.”

# Access to mobile phone data




- Data for statistical purposes do not exist.
- Preprocessing.
- Obstacles:
  - Legal issues?
  - Intellectual property rights and industrial secrecy
  - Costs
  - Public perception on privacy and confidentiality

# Statistical methodology



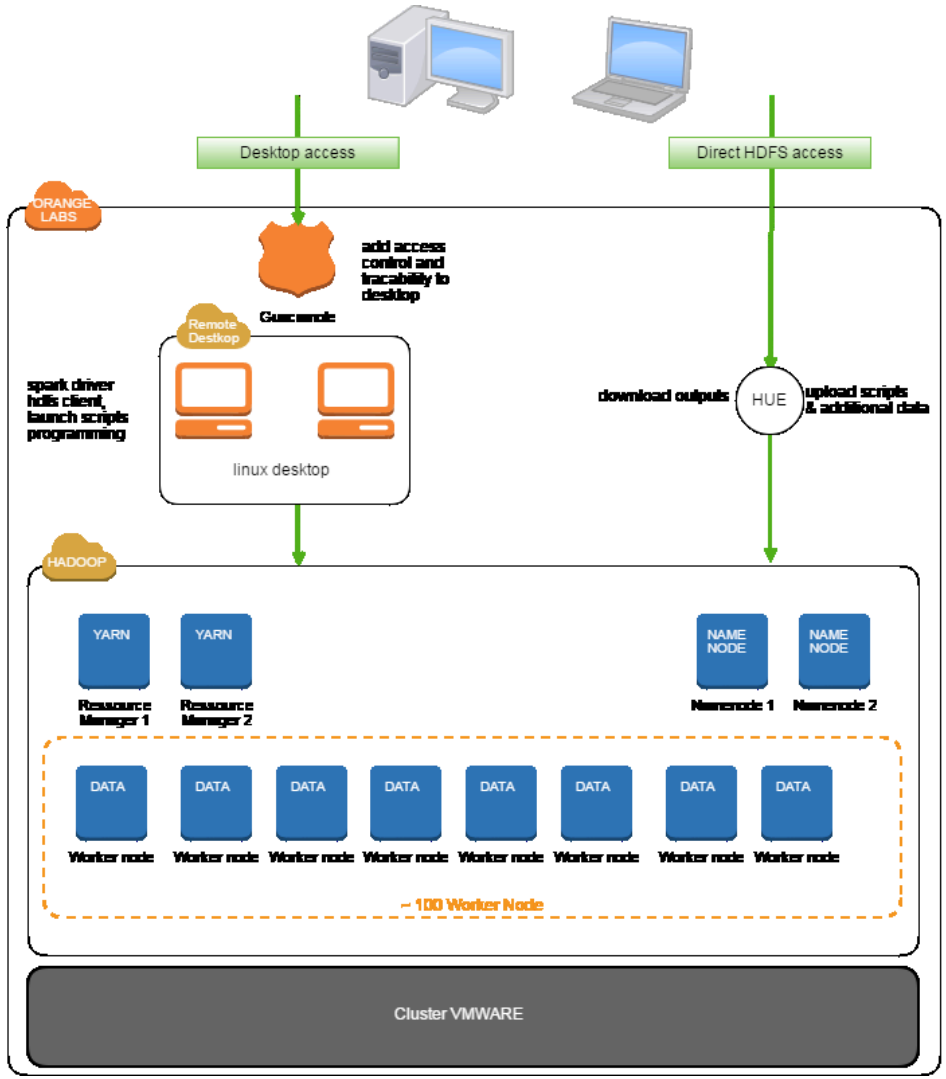

## Mobile Phone Data

- Phase 1: Raw Telco Data Generation
- Phase 2: Statistical MicroData Generation
  Aggregated Data Generation
- Phase 3: Inference

## Issues

- Extended two-phase life-cycle model $\rightsquigarrow$ Admin Data
- Geolocation of network events
- Data model
- Statistical models and data integration

# IT infrastructure: Access



# IT infrastructure: Geolocation

INSTITUTO NACIONAL DE ESTADISTICA

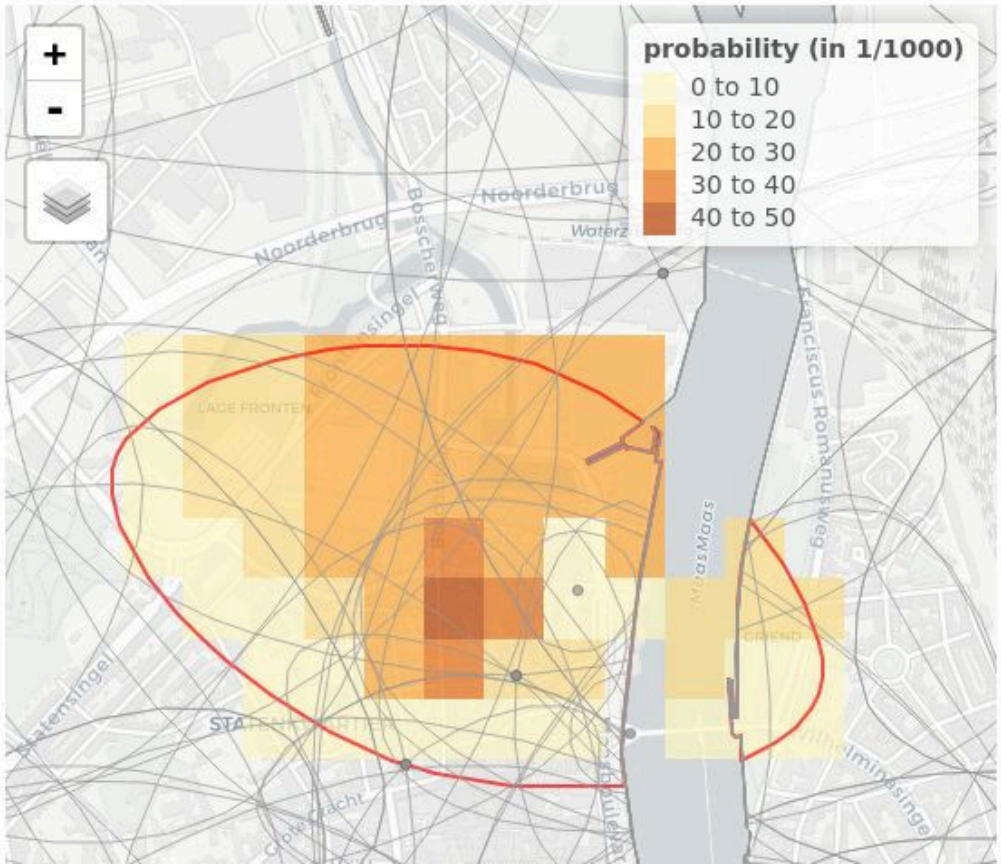
Cell Inspection Tool

**Variable**

- (•) Probability
- ( ) Signal strength (dB)

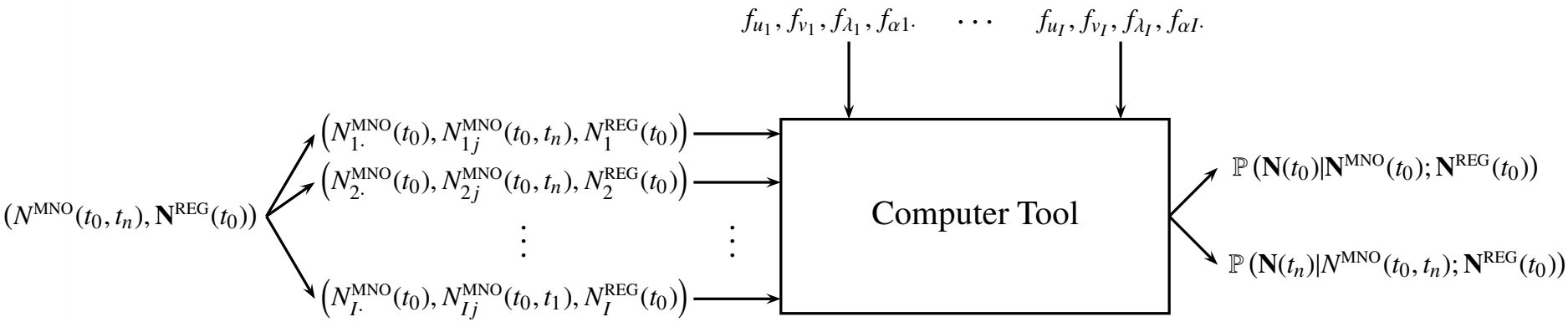
**Selected cells**

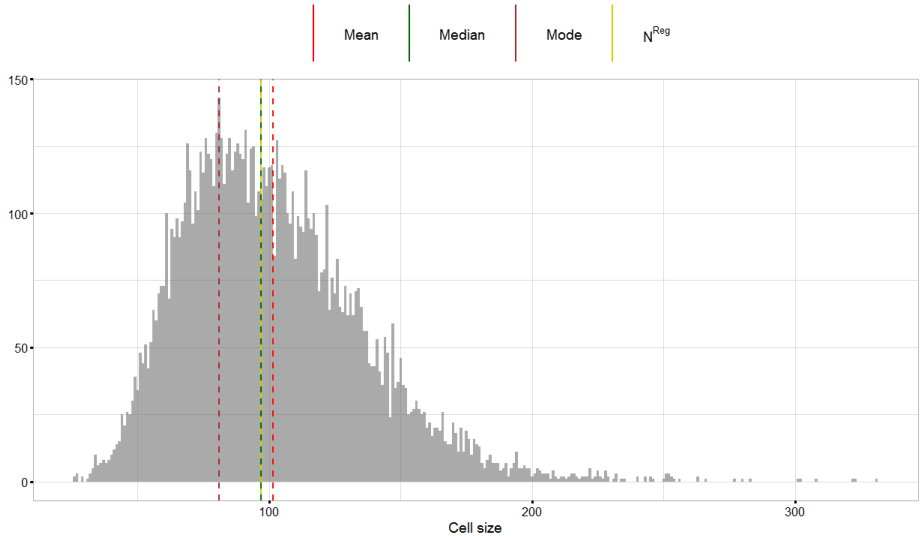
- [x] Cell 1
- [ ] Cell 2
- [ ] Cell 3
- [ ] Cell 4
- [ ] Cell 5
- [ ] Cell 6
- [ ] Cell 7
- [ ] Cell 8
- [ ] Cell 9
- [ ] Cell 10
- [ ] Cell 11
- [ ] Cell 12
- [ ] Cell 13
- [ ] Cell 14



# IT infrastructure: Inference





# Quality issues

- **CoP** affected by two generic facts:
  - MNOs **active part of the production process**.
  - Change of **inferential paradigm**.
  - **Higher** degree of **breakdown**.

- Example: **accuracy** dimension

  From **confidence** intervals to **credible** intervals

  Model **checking** and model **assessment**

# Main conclusions

- **Access blocked**: further work on **perceived risks** and **collaboration**.
- **Total Survey Error** paradigm still **valid**.
- **Geolocation**.
- **No probability sampling**: hierarchical **models**?
- **Computational complexity** for storing, accessing, and computing **in situ**.
- **Quality Assurance Framework** needs revision: **active role** of data holders and change of **inferential paradigm**.

# The future research