Data collection and selective data editing in a systemized and integrated way: an experience in progress at Statistics Spain

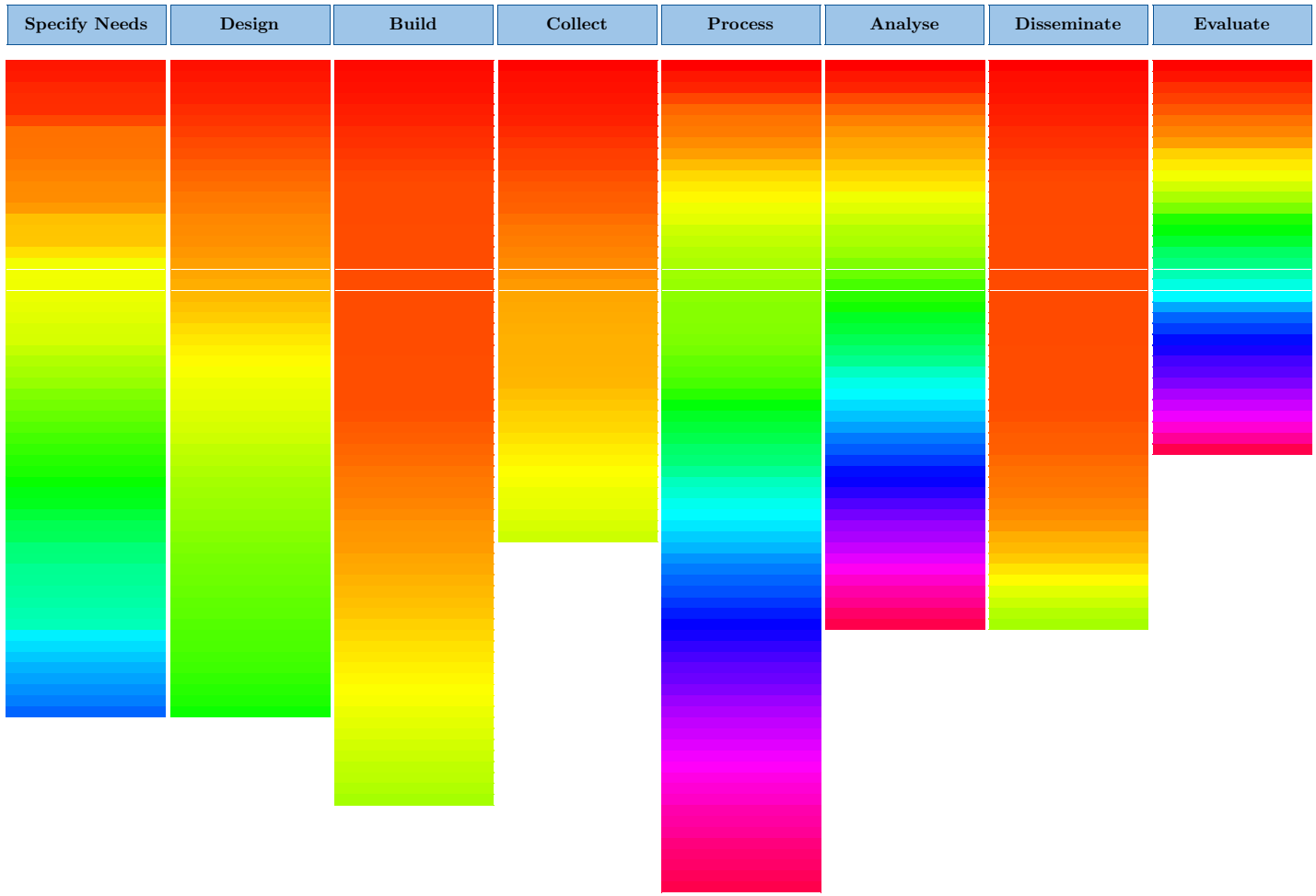
Workshop on the modernisation of statistical production

J.M. Bercebal, J.L. Maldonado, M.A. Martínez-Vidal, and **David Salgado**

Statistics Spain

Geneva, 15-17 April, 2015

Starting point: level of standardization



A priori subjective perception (no algorithm involved)

Streamlining data collection. I

INSTITUTO NACIONAL DE ESTADISTICA

integración de la recogida de información y su administración | iria

Statistics Spain has designed, developed and is currently deploying **IRIA**: a parameterised computer system to design, build, edit, exploit and manage data collection in business and household surveys with heterogeneous characteristics.

- **Modularity** and **configurability**.
- **Extensibility**.
- **Ease-of-use** (no need of deep computer skills).
- **Multi-mode** data collection.
- Component **reusability**.

Streamlining data collection. II

INSTITUTO NACIONAL DE ESTADISTICA

IRIA **Manager**

IRIA **Designer**

IRIA **Engine**

IRIA **Data Collection**

Streamlining data editing. I

Streamlined **E&I strategies** focused on more efficient **error detection**

Efficient error detection based on two principles:

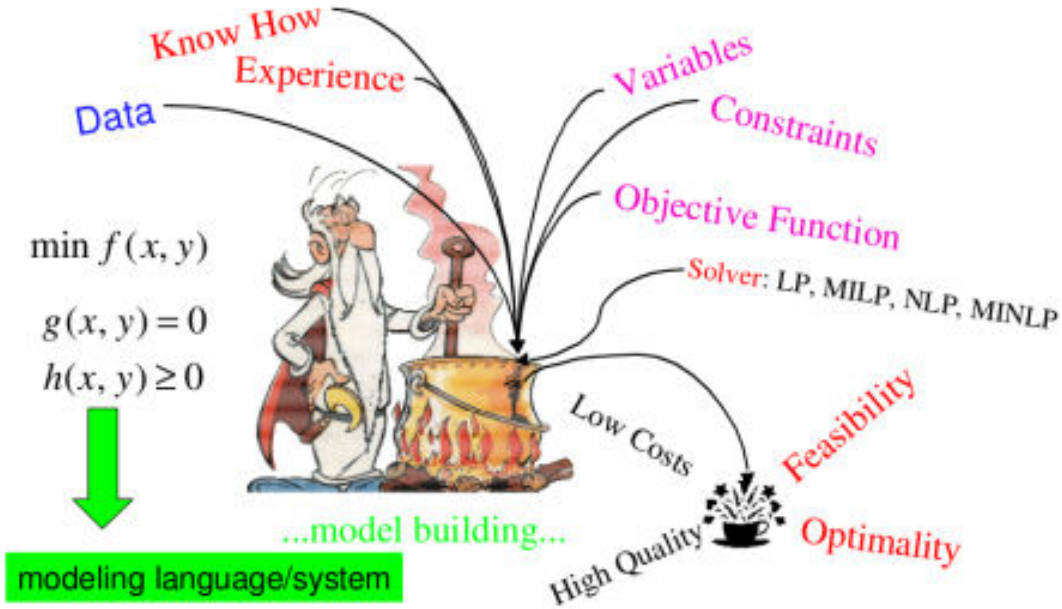
(i) **Minimization** of resources for **interactive** tasks.

(ii) Data **quality assurance**.

Different optimization problems according to **aux info**.

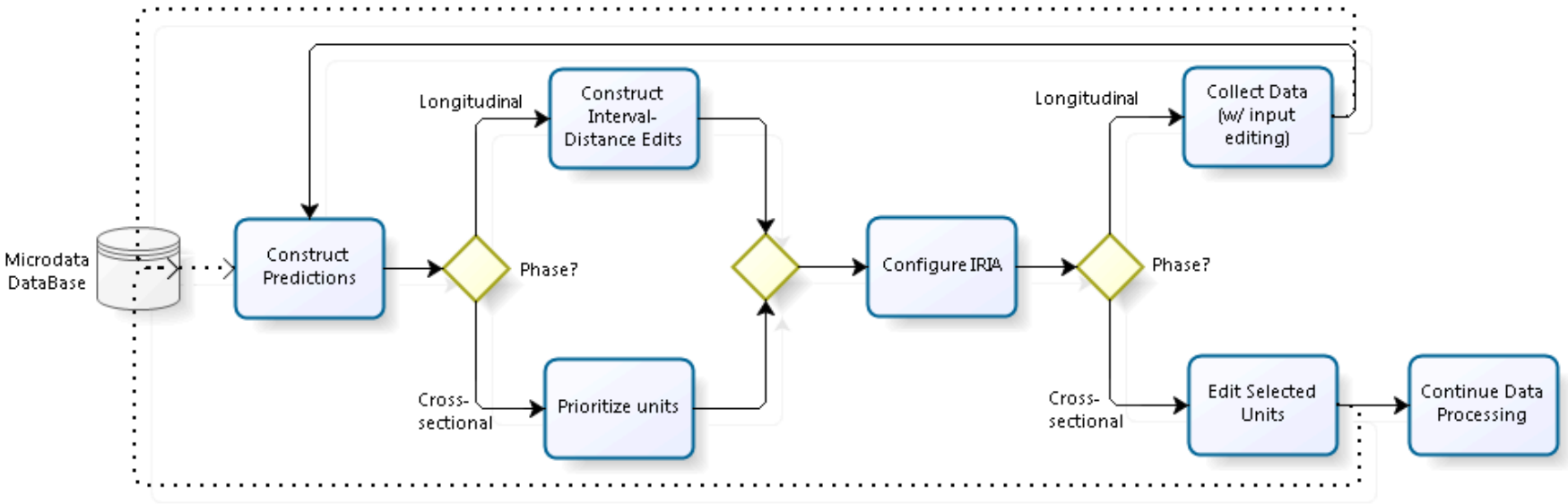
Streamlining data editing. II

Optimization



- provides **statistical foundations** for the heuristic traditional approach by using **statistical models**;
- naturally **automatises** the unit **selection** process even customising edits for **each variable** and **each unit** on **each time period**;
- opens the possibility to perform **selective editing** upon **categorical** variables (*still under research*).

Implementation in production



Some details

- IRIA **designed**, **developed** and **deployed without stopping** production.
- Computation of **edit values** (intervals) not yet integrated in IRIA.
- Heavy computation of edit values (intervals) undertaken with **R packages** developed on purpose:
  - **Microdata database** provisionally prototyped with a **key-value pair structure** as a plain filesystem.
  - Heavy use of **OOP principles** (S4 classes) in the statistical programming.
  - **Modularity** achieved through **packages** (more than 15 packages developed).

Some lessons

In designing and developing statistical production . . .

- **computer system design principles** (data abstraction, modularity, . . .) as leverage intimately linked with **statistical theory**:
  - efficient **software development principles** (OO, . . .) fully considered **in designing statistical routines**;
- efficient professional profile as a **fusion of statistician and computer scientist** (data scientist?);
- **optimal resource allocation** as a principle;

Some obstacles

From our experience the **organization** must face ...

- **cultural reluctance** to new organization of the statistical production;
- **legacy code** consuming resources for its maintenance;
- **legacy human capital** hard to train in new computer skills.

How to **change** the **statistical production system** without stopping at all the production in a **severely resource-restricted environment**?