Sampling and quality control measures for data collection

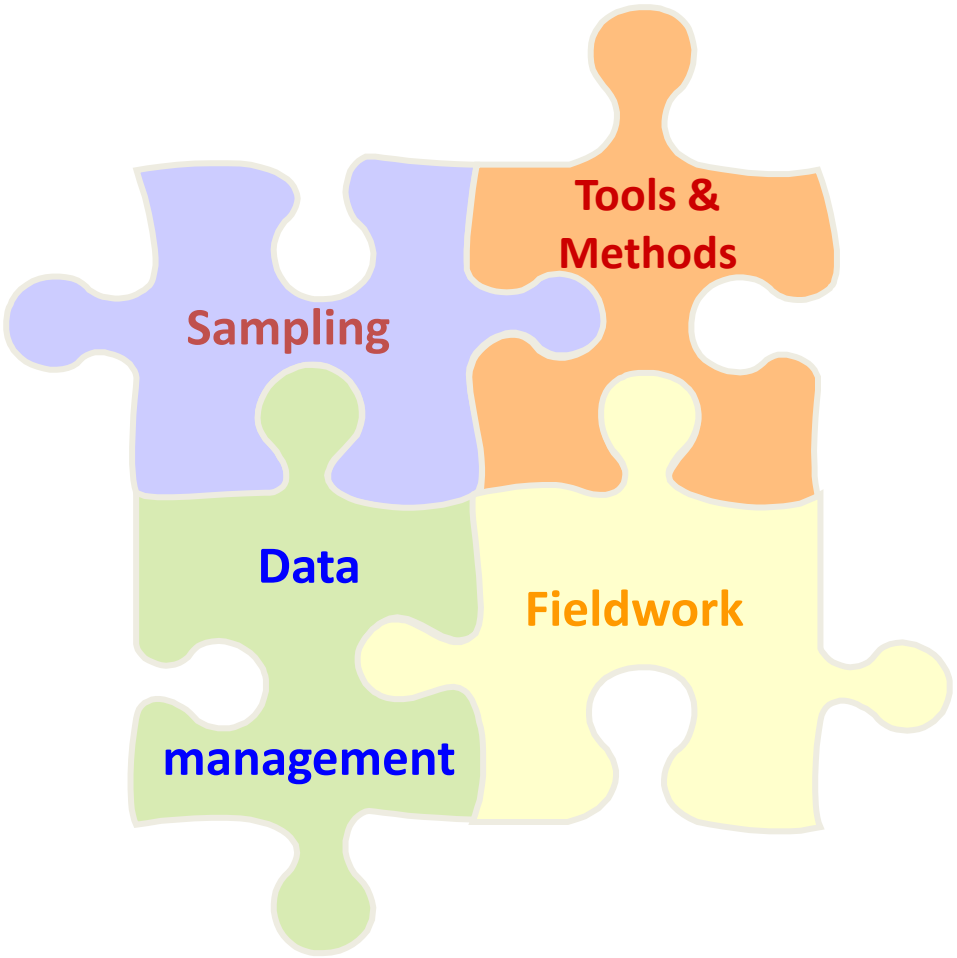
The integration of

sampling design,
development of questionnaires and methods,
organization, training and supervision of field staff,
and computerized quality controls of fieldwork

to provide analysts with reliable data on time

Álvaro Canales and Juan Muñoz
Cape Town, December 2009

The total survey design approach



Scientific Sampling

- Households should be selected through a **documented** process that gives each household in the population of interest a probability of being chosen that is **positive** and **known**
  - This permits making inferences from the sample to the entire population with known margins of error
- Household samples are generally not simple random samples. They are instead
  - Stratified
    - by Region, by Urban/Rural, by Intervention/Control, …
  - Selected in two stages or more
    - Area Units in the first stage/s
    - Households in the last stage

This is called **random sampling** Notice that the selection probabilities do not need to be the same for all households

Only random sampling can do this

In Simple Random Sampling, households are selected with the same probability, and independently of each other

The smallest area units are called sample points. The groups of households selected in each sample point are called clusters

Sampling error

- Sampling error is the result of observing a sample of **n** households (the sample size) rather than all **N** households in the country
- The standard error **e** is a measure of a sample's precision.
  - The chances for the true value of an indicator being farther than **2e** apart from its sampling estimate are about 95 percent.
- The standard error **e** decreases with the square root of the sample size **n**.
  - To reduce the error to one half, the sample size must be quadrupled.
- The size of the population **N** has almost no influence on the size of the sample that is needed to achieve a given precision.
  - To obtain national estimates, big countries and small countries require samples of about the same size.
- Increasing the sample size will generally reduce sampling errors
  - However, it is also likely to increase non-sampling errors

- Why two stages?
  - An updated list of all households in the country is generally unavailable
  - A single-stage sample would be too scattered in the territory
- Why stratification?
  - In order to potentially improve precision, by gaining control of the composition of the sample
  - In order to provide estimates for subgroups that would otherwise be poorly represented (small regions, women-headed households, etc.)

Two-stage sampling solves these problems, but the sample becomes less precise as a result of clustering

These two objectives are generally contradictory in practice

Most stratified samples select households with unequal probabilities. This implies that the survey needs to be analyzed with weights.

The combined result of clustering and stratification is called *design effect*

Design effect (*deff*)

$$deff = \frac{e^2_{Our\ Survey}}{e^2_{A\ Simple\ Random\ Sample\ of\ the\ same\ size}}$$

Our survey will typically have a complex design, with two stages, stratification, etc.

$$deff = \frac{n_{Our\ Survey}}{n_{A\ Simple\ Random\ Sample\ with\ the\ same\ precision}}$$

Deff depends on the indicator being measured
For socio-economic indicators it is typically 3 or more
It can be a little less for demographic indicators
It can be a lot more for infrastructure indicators

Deff depends on the cluster size

Socio-economic surveys try not to exceed 15-20 households per cluster

Demographic surveys may occasionally do more

Sample frames

- An adequate **sample frame** needs to be available before a sample can be selected
  - A sample frame is a list of all units in the population
- The sample frame for the first stage is generally the most recent list of census enumeration areas
  - It needs to be linked to cartography
- The sample frame for the last stage is generally developed specifically for each survey, by way of a household listing operation conducted in all sample points.
  - The time and budget of household listing are
    - Small enough to be considered a marginal part of the overall data collection effort
    - Large enough to be a headache if they are forgotten or underestimated

Household listing issues

- Household listings (and the subsequent selection of the households to be visited) can be prepared
  - by the same fieldworkers who will conduct the interviews, or
  - by independent enumerators
  - The choice is difficult.
- Sample points larger than a few hundred households may require segmentation
  - The sample point is divided into smaller areas of approximately equal size called segments. Then one (or maybe a few) of the segments are randomly selected and listed.
  - Segmentation is a de facto extra sampling stage that is very difficult to supervise. It should only be used as a last resort.
- Beware of imitations and shortcuts
  - Implicit listing (asking the interviewers to select every n-th household on the ground rather than on paper) is not a recommended option.

Excluded strata

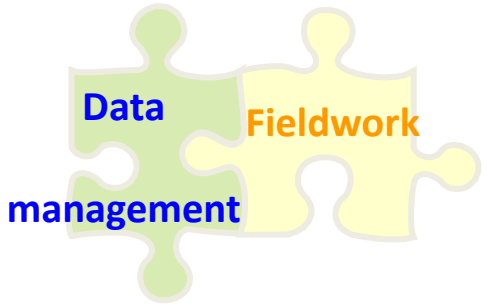
- Parts of the country may need to be excluded because of
  - Outside the program (geographically, organizationally, …)
  - Security reasons
  - Accessibility
  - Nomads
  - Etc.
- That's OK, as long as long as
  - Decisions are properly documented
  - Results are not extrapolated later to the whole country

Nonresponse

- None of the following is a solution for nonresponse
  - Replace nonrespondents with similar households
  - Increase the sample size to compensate for it
  - Use correction formulas
  - Use imputation techniques (hot-deck, cold-deck, warm-deck, etc.) to simulate the answers of nonrespondents
- The best way to deal with nonresponse is to prevent it
  - Some of the above may have a preventive value

To control non-sampling errors

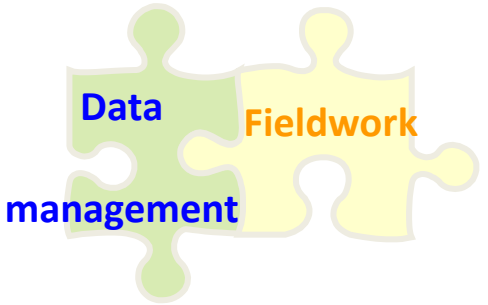
- Manage the survey as an integrated project
- Organize fieldwork on the basis of teams
- Implement computer-assisted field edits – the CAFÉ approach
- Establish strong supervision procedures
- Ensure sufficient training
- Work with a reduced staff over an extended period of data collection



Integrating fieldwork and data management

Computer Assisted Field Edits (the CAFÉ approach)

What happens without integration?



- **A long and frustrating process of "data cleaning" becomes unavoidable**

  The data loose their policy-making relevance

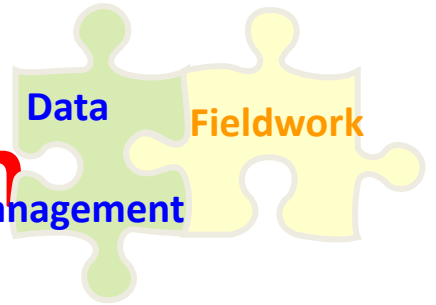
- **Data quality is not guaranteed**

  The process converges (at best) to databases that are *internally consistent*

- **The process entails a myriad of decisions, generally undocumented**

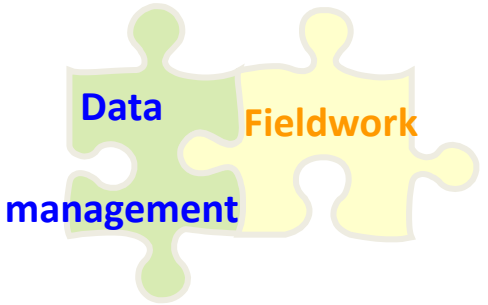
  Users mistrust the data

Benefits of integration

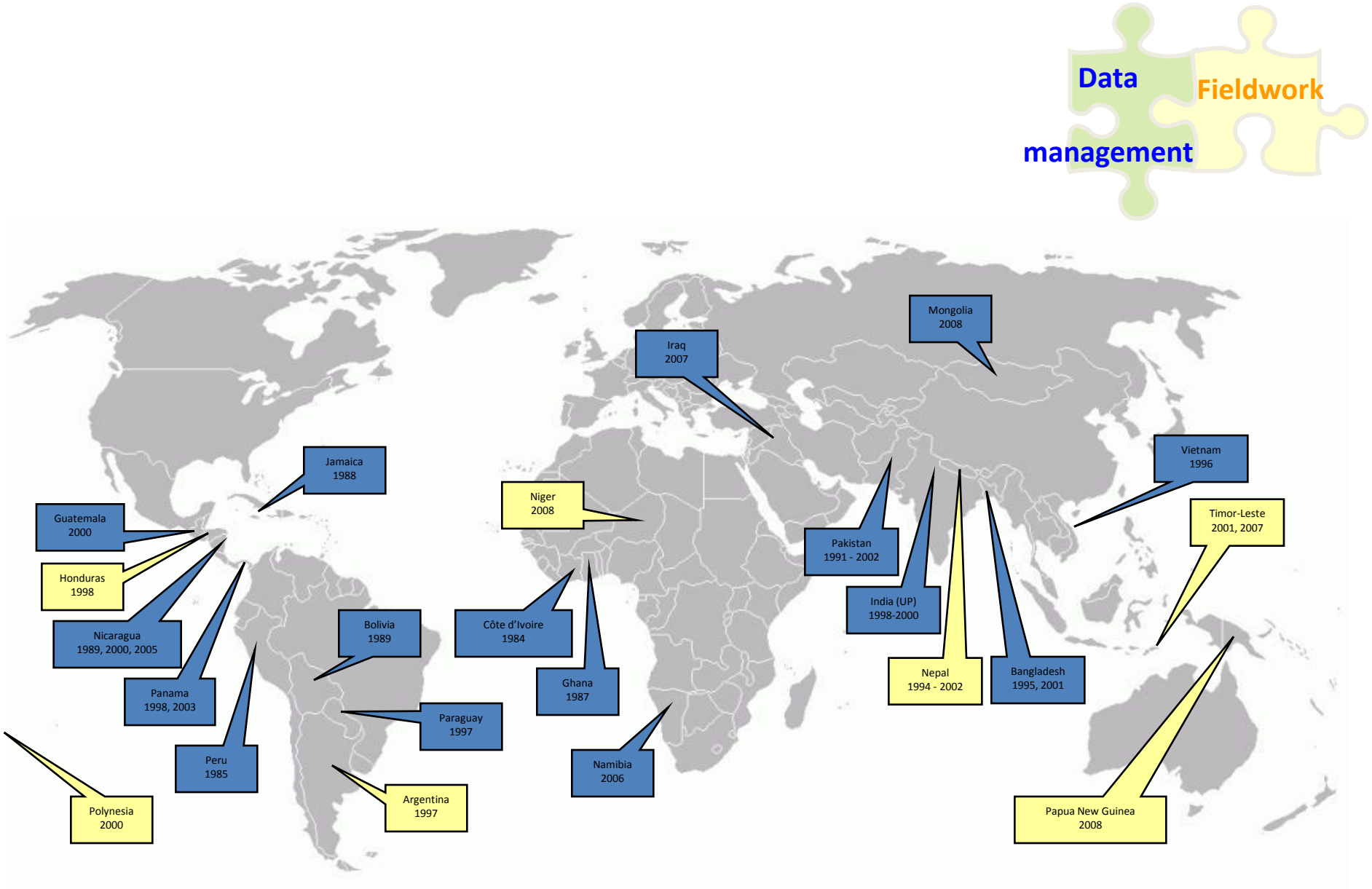


- Provides reliable and timely databases
- Provides immediate feedback on the performance of the field staff, allowing early detection of inadequate behaviors
- Ensures that all field staff applies uniform criteria throughout the full period of data collection
- Solves inconsistencies through direct verification of households reality, rather that through office guesswork
- Is consistent with the total quality culture

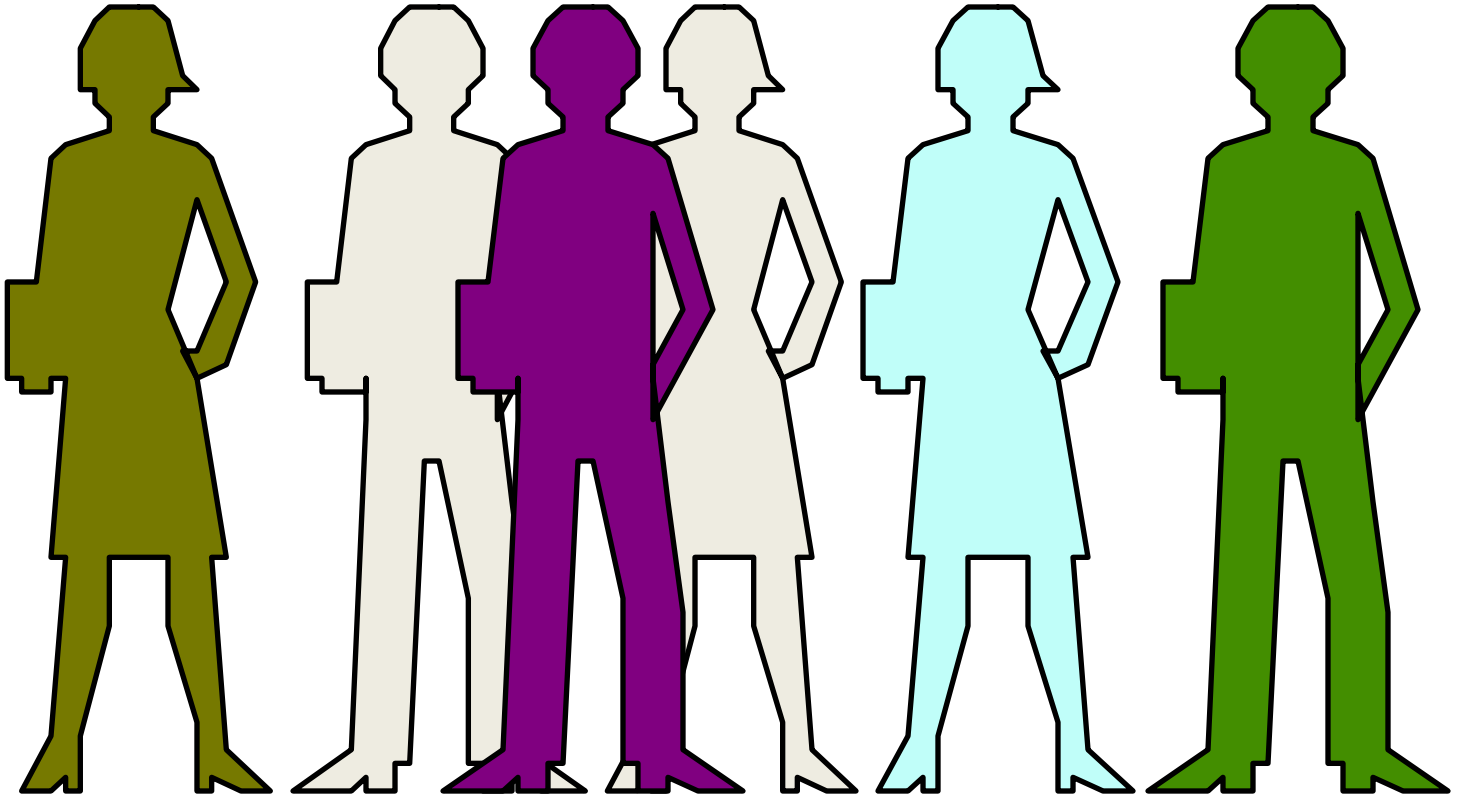
Tactical options



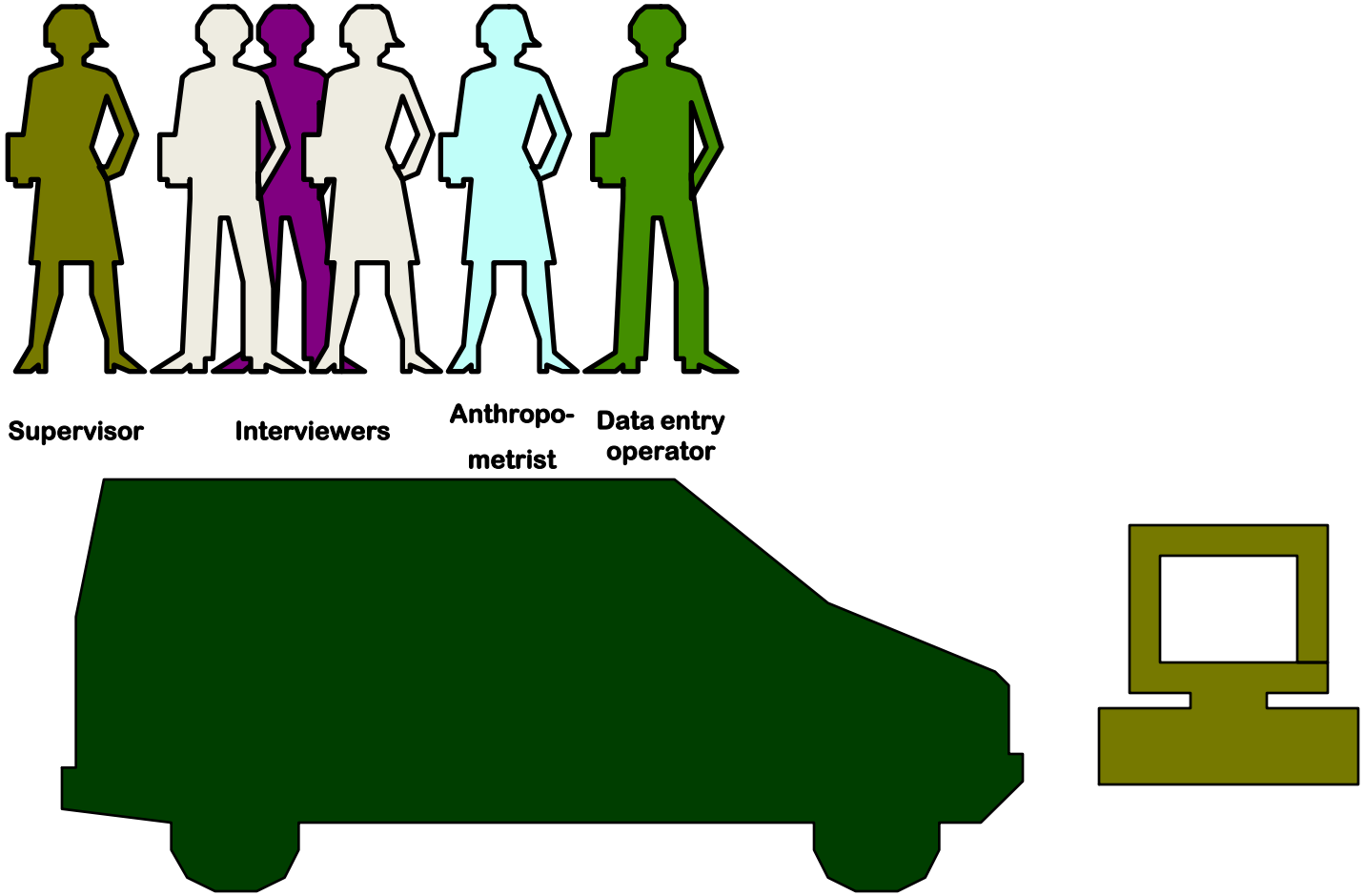
- Two options have proved to be successful
  - Mobile teams with fixed data entry
    - Cote d'Ivoire (1984)
    - Many other countries for almost 25 years
    - Iraq (2006-2007)
  - Mobile teams with mobile data entry
    - Nepal (1992)
    - Many other countries for over 15 years
    - Papua New Guinea (2009)
- Both are based on the team approach
- Neither has tried to get rid of paper questionnaires…
- …yet


Data
Fieldwork
management
Mongolia
2008
Iraq
2007
Jamaica
1988
Vietnam
1996
Niger
2008
Guatemala
2000
Timor-Leste
2001, 2007
Pakistan
1991 - 2002
Honduras
1998
India (UP)
1998-2000
Nicaragua
1989, 2000, 2005
Bolivia
1989
Côte d'Ivoire
1984
Nepal
1994 - 2002
Bangladesh
1995, 2001
Panama
1998, 2003
Ghana
1987
Paraguay
1997
Peru
1985
Namibia
2006
Polynesia
2000
Argentina
1997
Papua New Guinea
2008

Composition of a field team

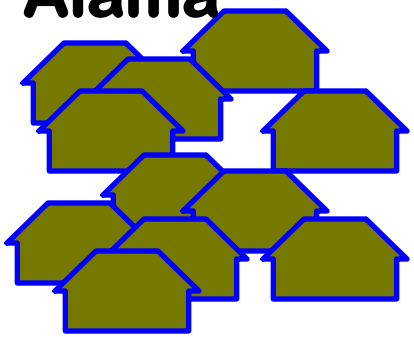


The team and its tools

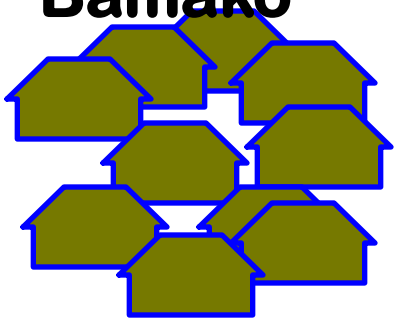


Two sample points visited in a four-week period

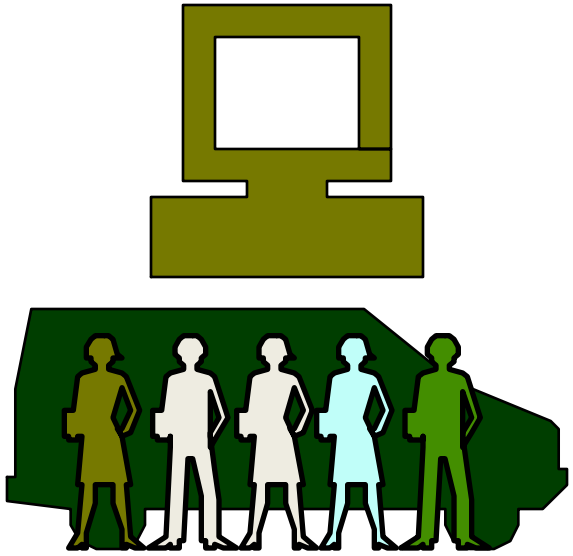
Alama



Bamako

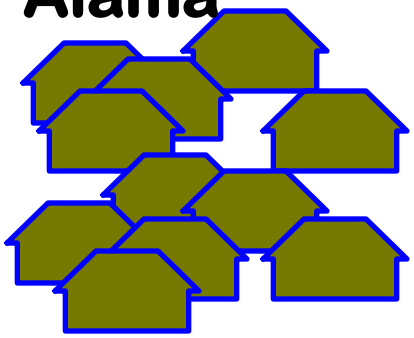


Regional Office

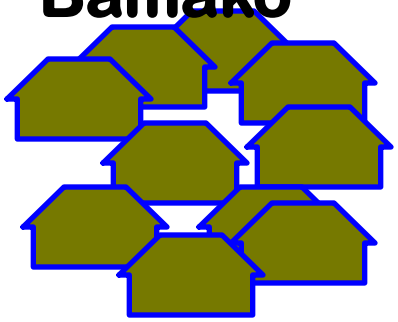


First week

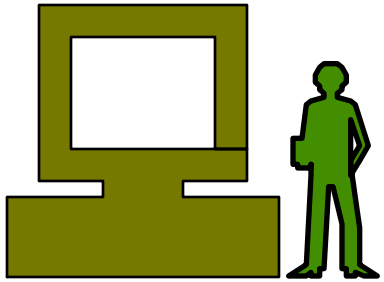
Alama



Bamako



Regional Office

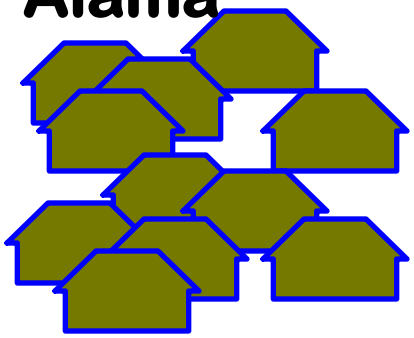


Operator remains in Regional Office

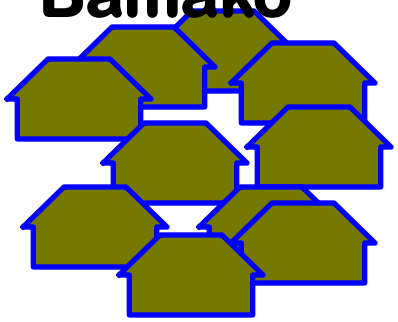
Rest of the team travels to Alama

First week

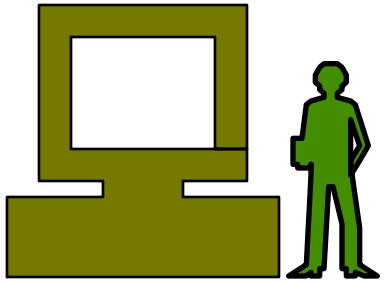
Alama



Bamako



Regional Office

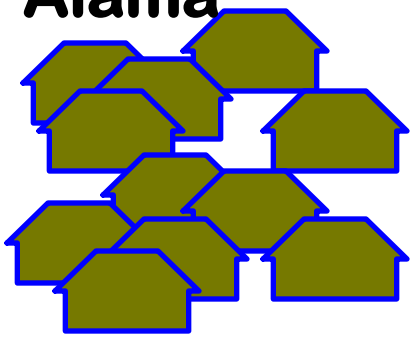


Operator remains in Regional Office

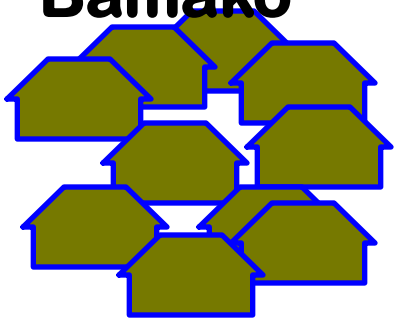
Rest of the team travels to Alama

First week

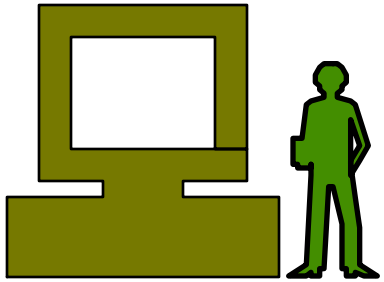
Alama



Bamako



Regional Office

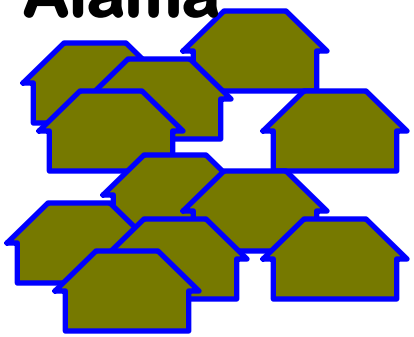


Operator remains in Regional Office

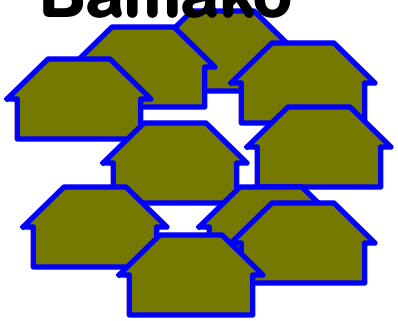
Rest of the team travels to Alama

First week

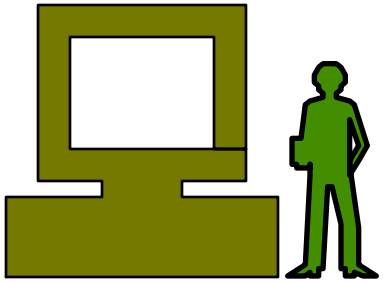
Alama



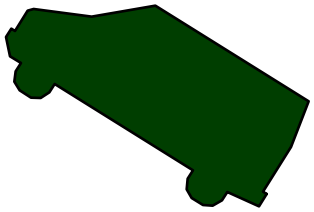
Bamako



Regional Office



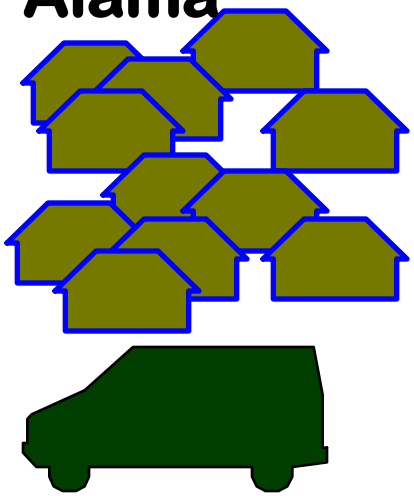
Operator remains in Regional Office



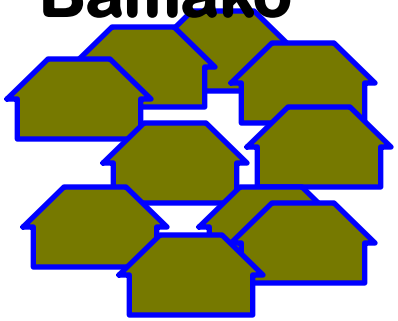
Rest of the team travels to Alama

First week

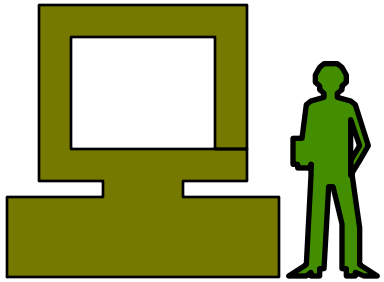
Alama



Bamako



Regional Office

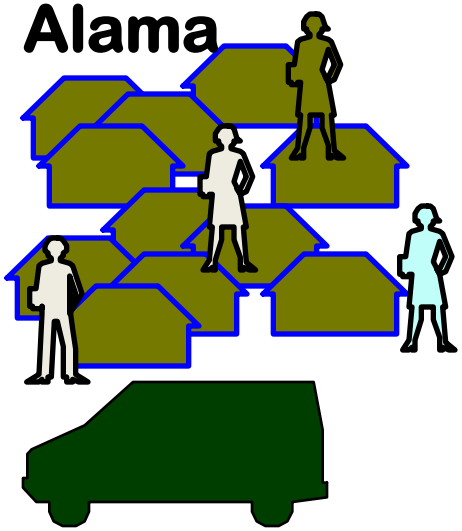


Operator remains in Regional Office

Rest of the team travels to Alama

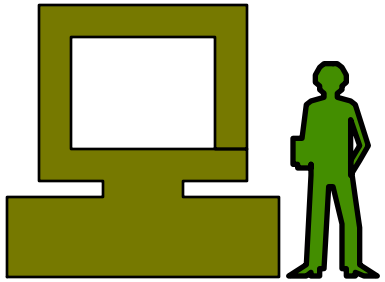
First week

Alama



They complete first half of questionnaires in all selected households

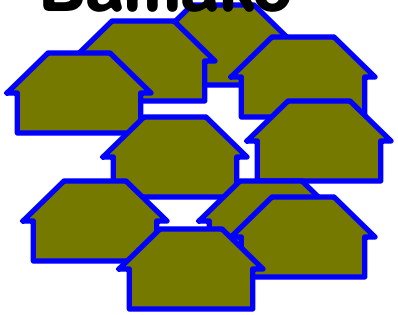
Regional Office



Operator remains in Regional Office

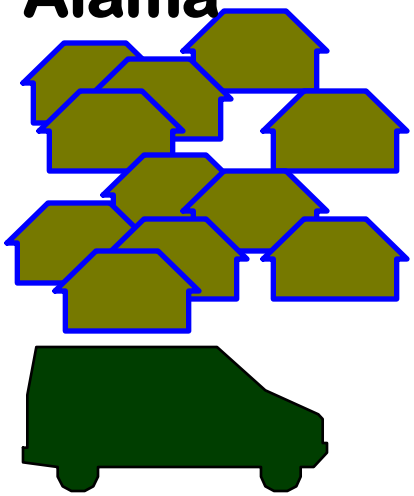
Rest of the team travels to Alama

Bamako

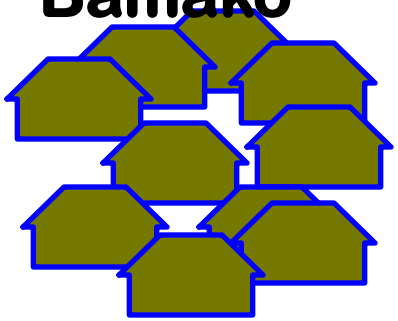


First week

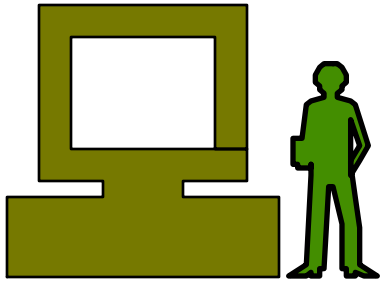
Alama



Bamako



Regional Office

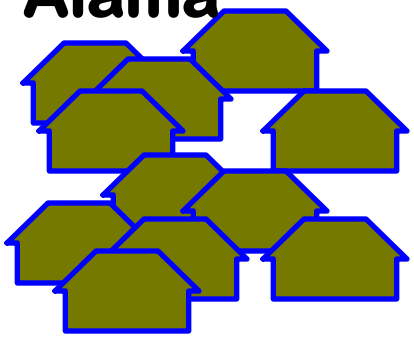


Operator remains in Regional Office

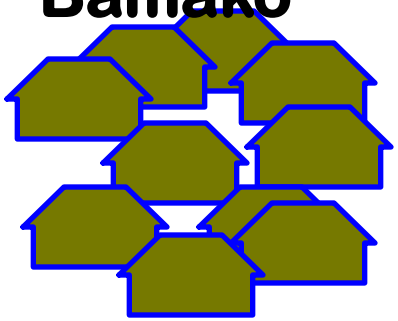
Rest of the team travels to Alama

First week

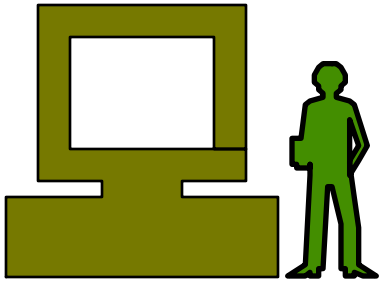
Alama

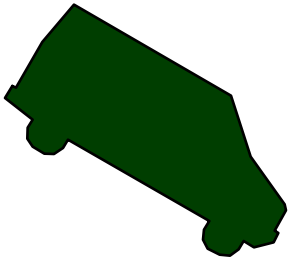


Bamako



Regional Office



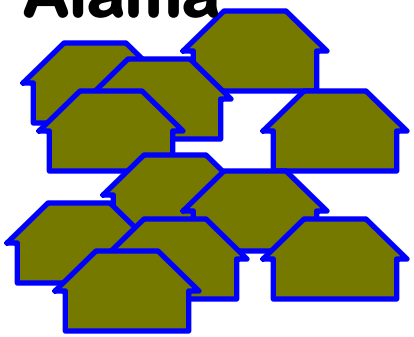


Operator remains in Regional Office

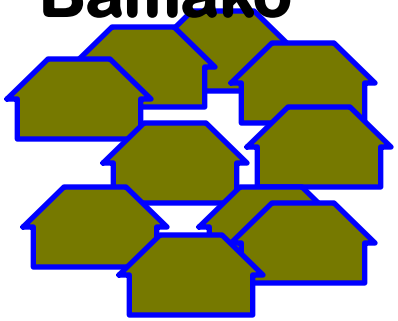
Rest of the team travels to Alama

First week

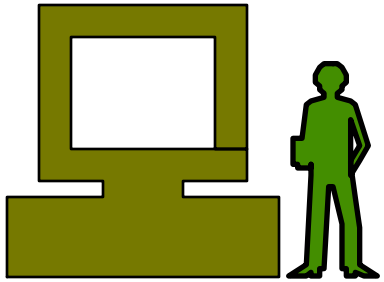
Alama



Bamako



Regional Office

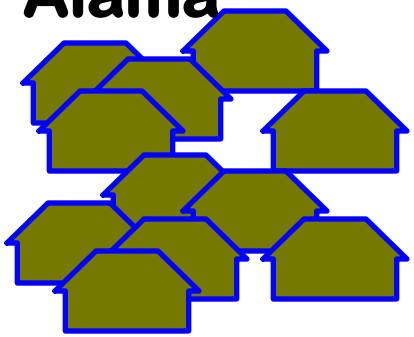


Operator remains in Regional Office

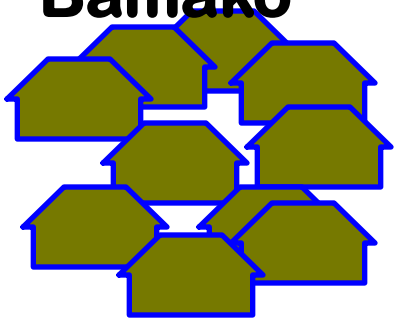
Rest of the team travels to Alama and back

First week

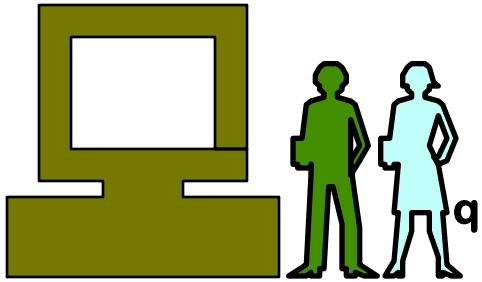
Alama



Bamako



Regional Office

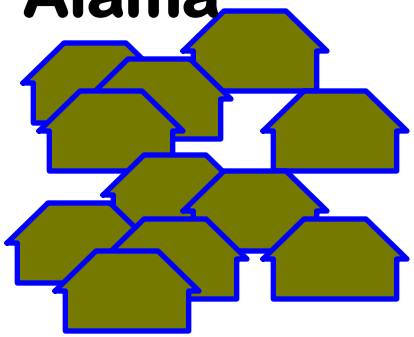


Supervisor gives Alama questionnaires to DEO

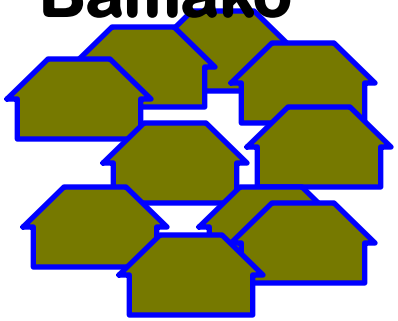
Rest of the team travels to Alama and back

Second week

Alama

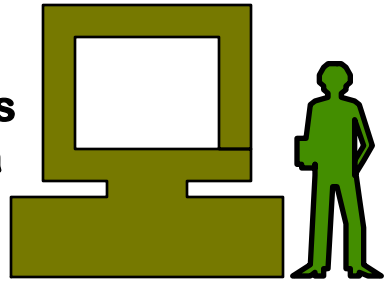


Bamako



Regional Office

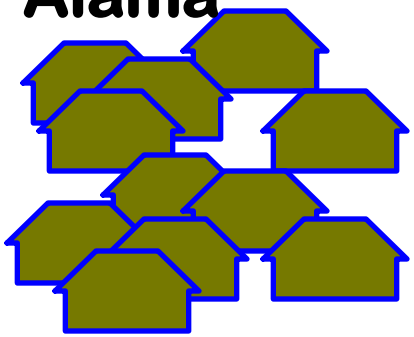
Operator enters first week data from Alama



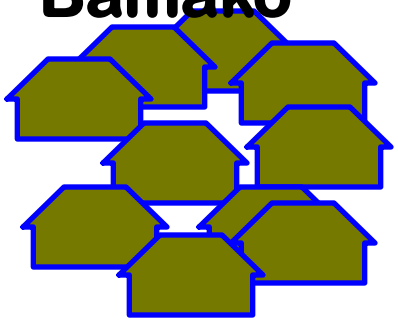
Rest of the team travels to Bamako

Second week

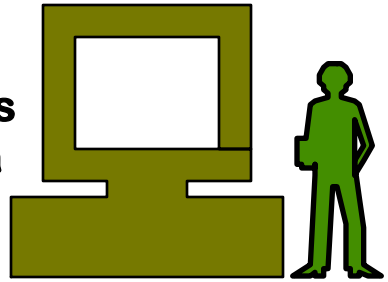
Alama



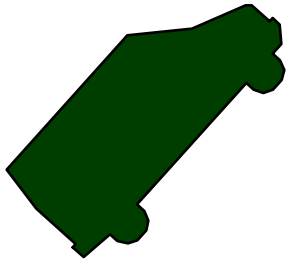
Bamako



Regional Office



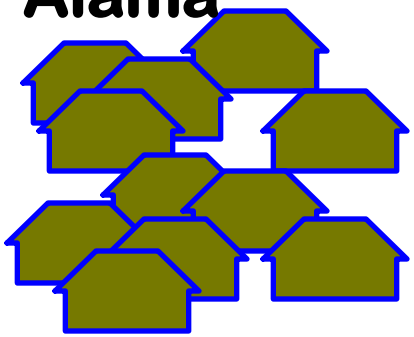
Operator enters first week data from Alama



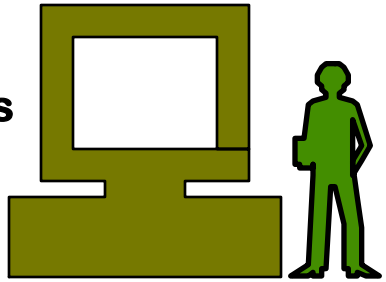
Rest of the team travels to Bamako

Second week

Alama



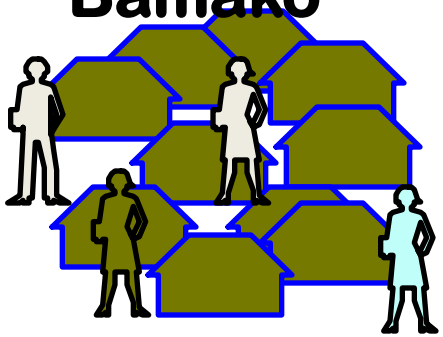
Regional Office

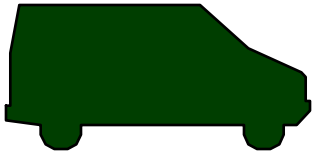


Operator enters first week data from Alama

Rest of the team travels to Bamako

Bamako

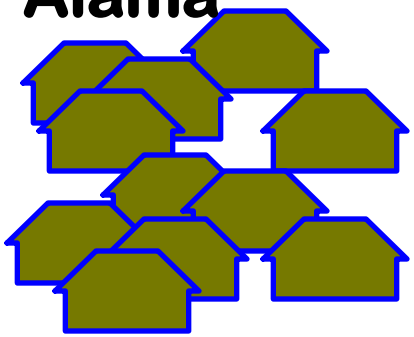




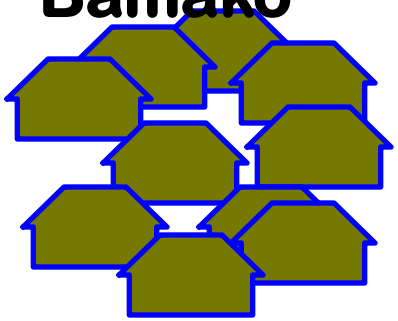
They complete first half of questionnaires in all selected households

Second week

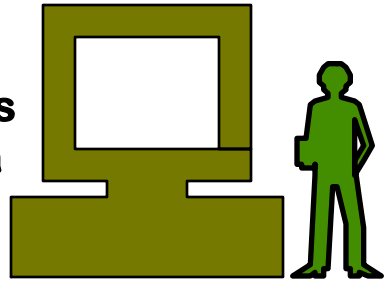
Alama



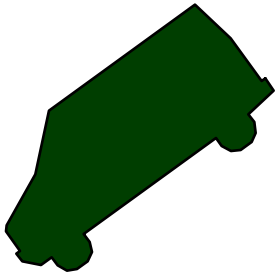
Bamako



Regional Office



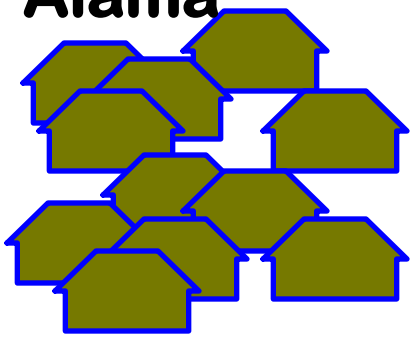
Operator enters first week data from Alama



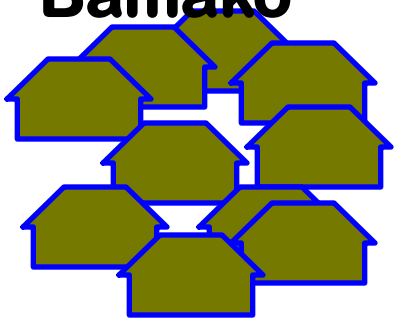
Rest of the team travels to Bamako and back

Second week

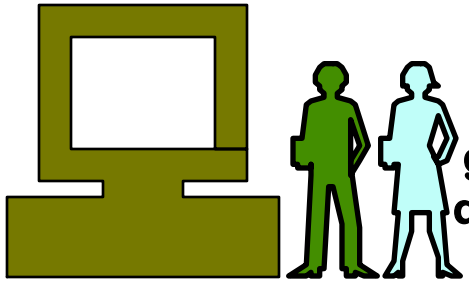
Alama



Bamako



Regional Office

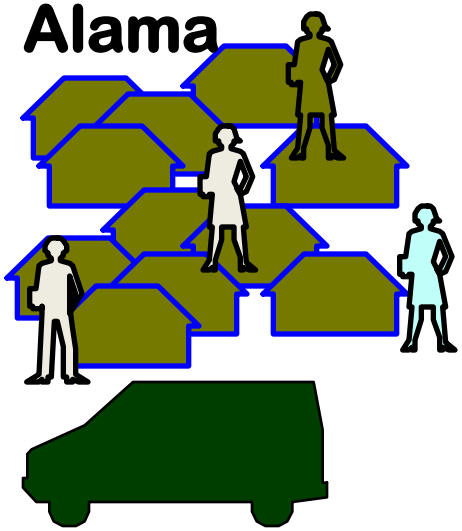


Supervisor gives Bamako questionnaires to DEO. DEO gives back Alama questionnaires with flagged inconsistencies

Rest of the team travels to Bamako and back

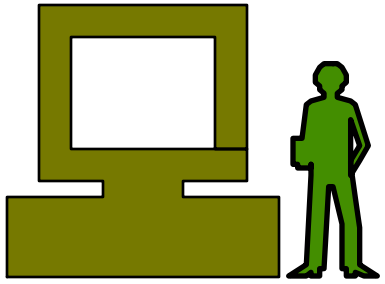
Third week

Alama



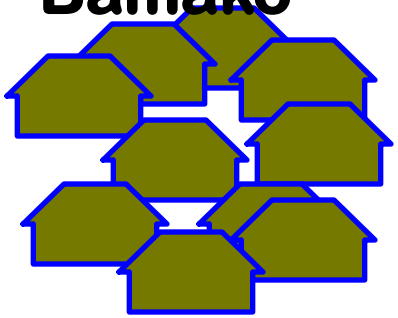
Team completes second half of questionnaires. They correct inconsistencies from first half

Regional Office



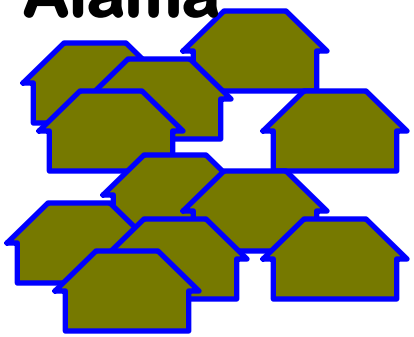
Operator enters first week data from Bamako

Bamako

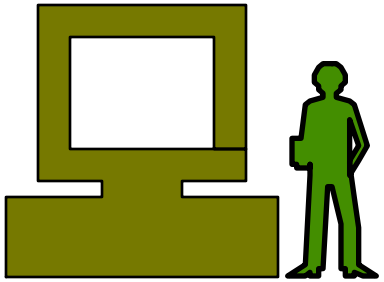


Fourth week

Alama

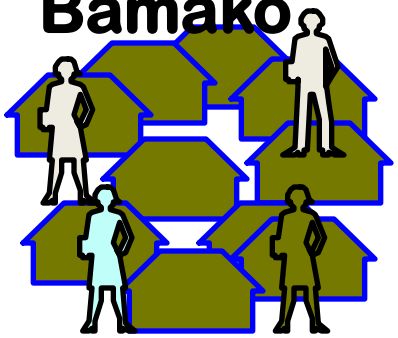


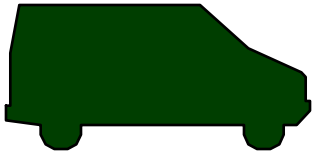
Regional Office



Operator enters second week data from Alama. Corrects inconsistencies from first round

Bamako



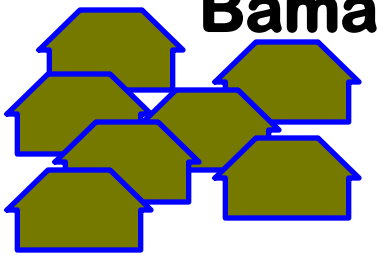


Team completes second half of questionnaires. They correct inconsistencies from first half

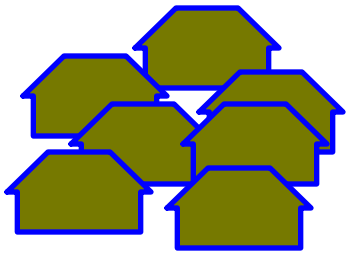

The result is a clean data set, ready for analysis immediately after data collection

Mobile teams with integrated data entry

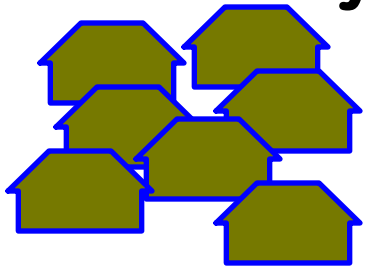
Bamako



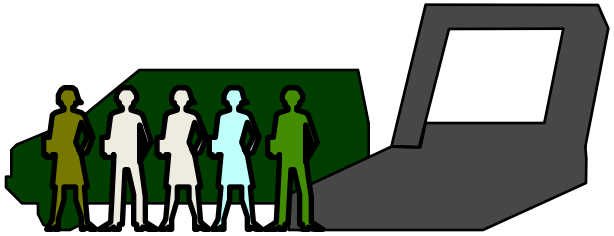
Alama



Cocody

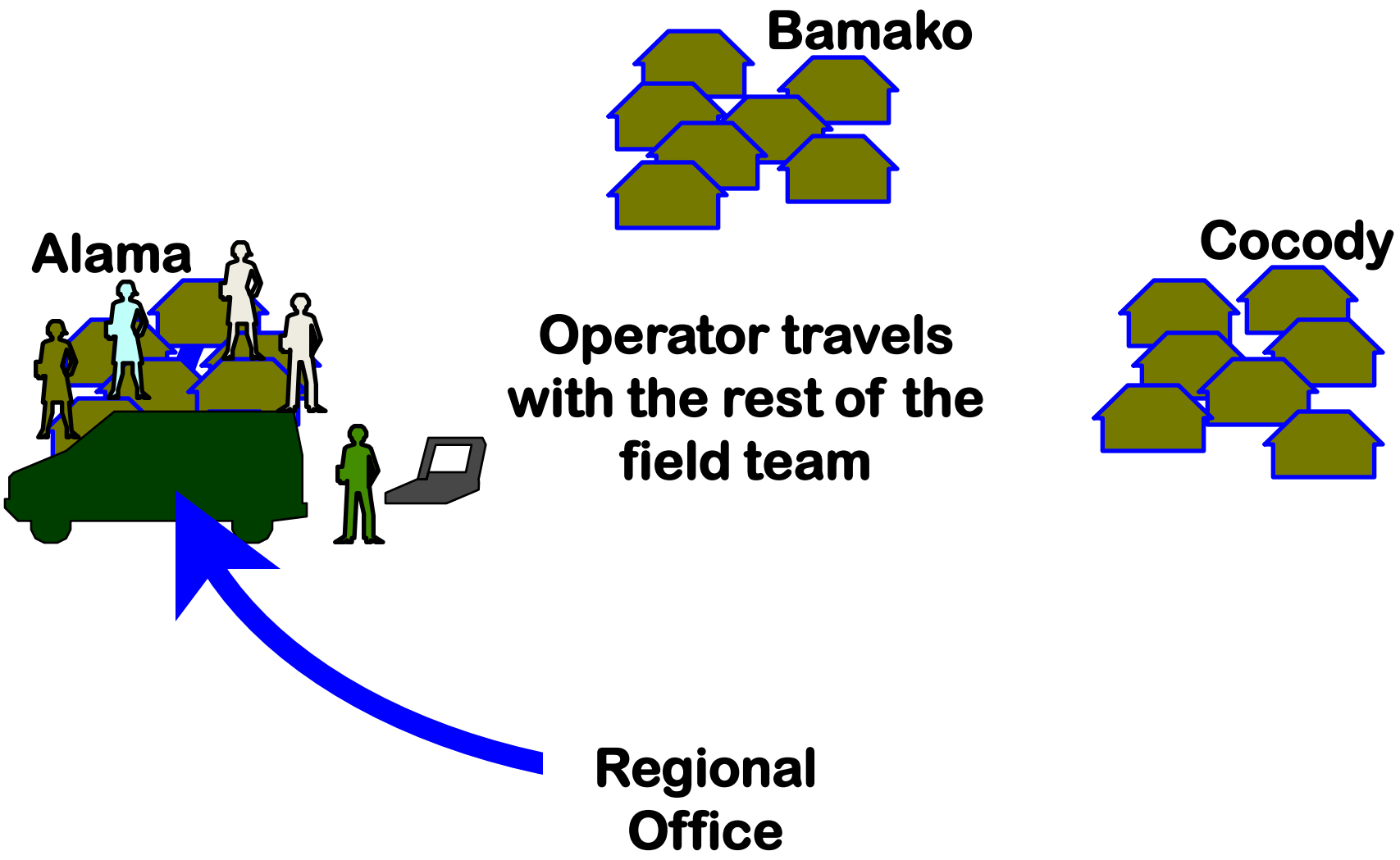


Team works with portable computers and printers

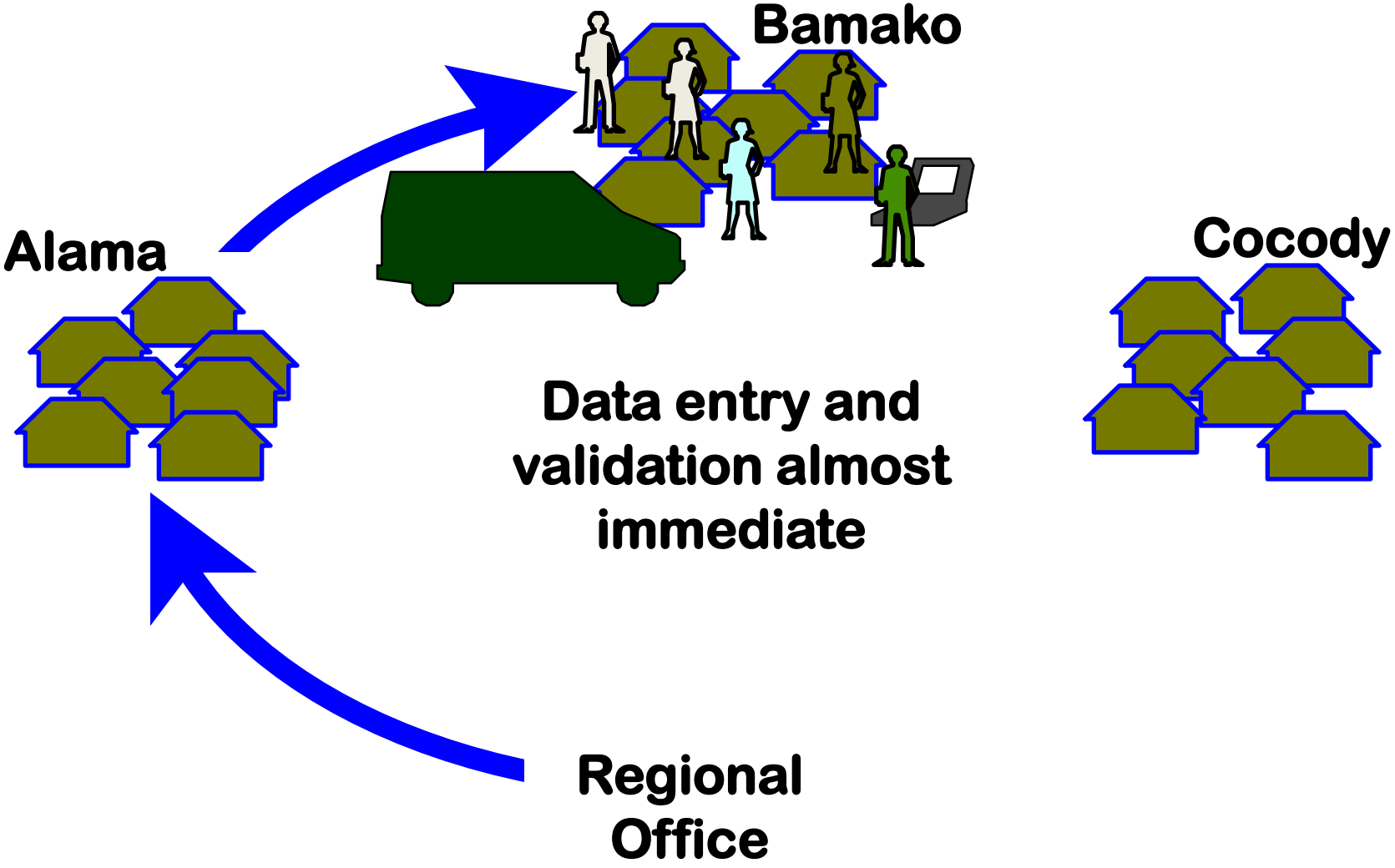


Regional Office

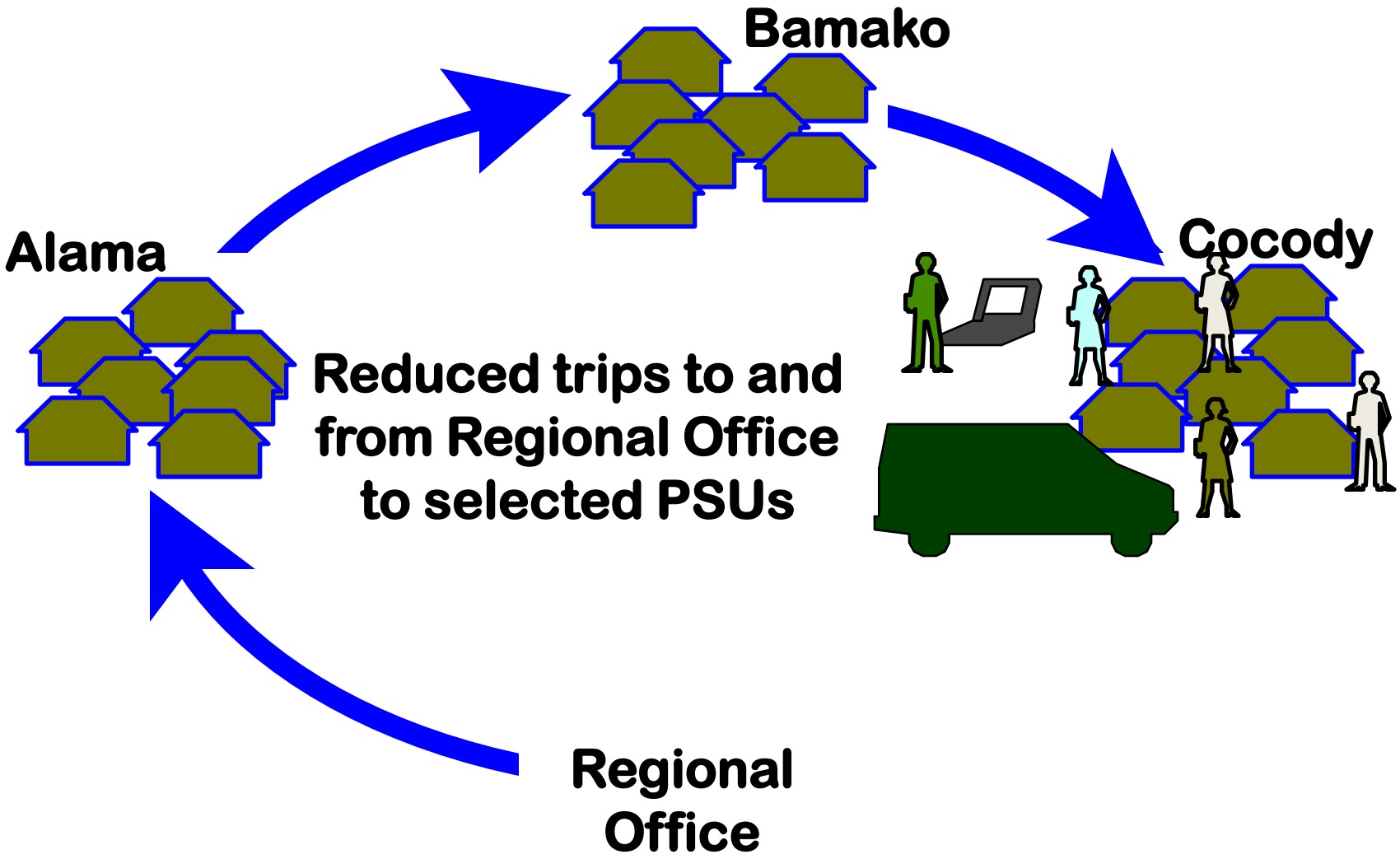
Mobile teams with integrated data entry



Mobile teams with integrated data entry



Mobile teams with integrated data entry



Mobile teams with integrated data entry