

# 14<sup>th</sup> International Advanced School of Empirical Software Engineering

# Surveys in Software Engineering

Marco Torchiano



Further coaches

Rafael Maiani de Mello



Daniel Méndez



Guilherme H. Travassos



# Feel free to...



copy, share, and alter,



film and photograph (as long as we look great)



tweet and live blog



today's presentations given that you attribute  
it to its author(s) and respect its right and  
licenses of its parts.

# Who are we?



**Marco Torchiano**

Politecnico di Torino

<http://softeng.polito.it/torchiano/>



**Daniel Méndez**

Technical University of Munich

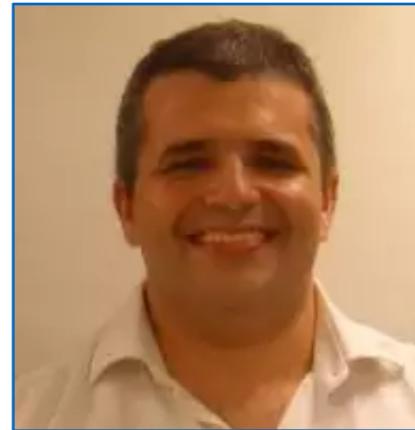
<http://www4.in.tum.de/~mendezfe/>



**Guilherme Horta Travassos**

COPPE, Federal University of Rio de Janeiro

<http://www.cos.ufrj.br/~ght>



**Rafael Maiani de Mello**

Pontifical Catholic University of Rio de Janeiro

<http://inf.puc-rio.br/~rmaiani>

# What do you think?

Why do we need survey research  
in software engineering?



# Session I

Introduction to Survey research

# Big Picture... 1<sup>st</sup> layer

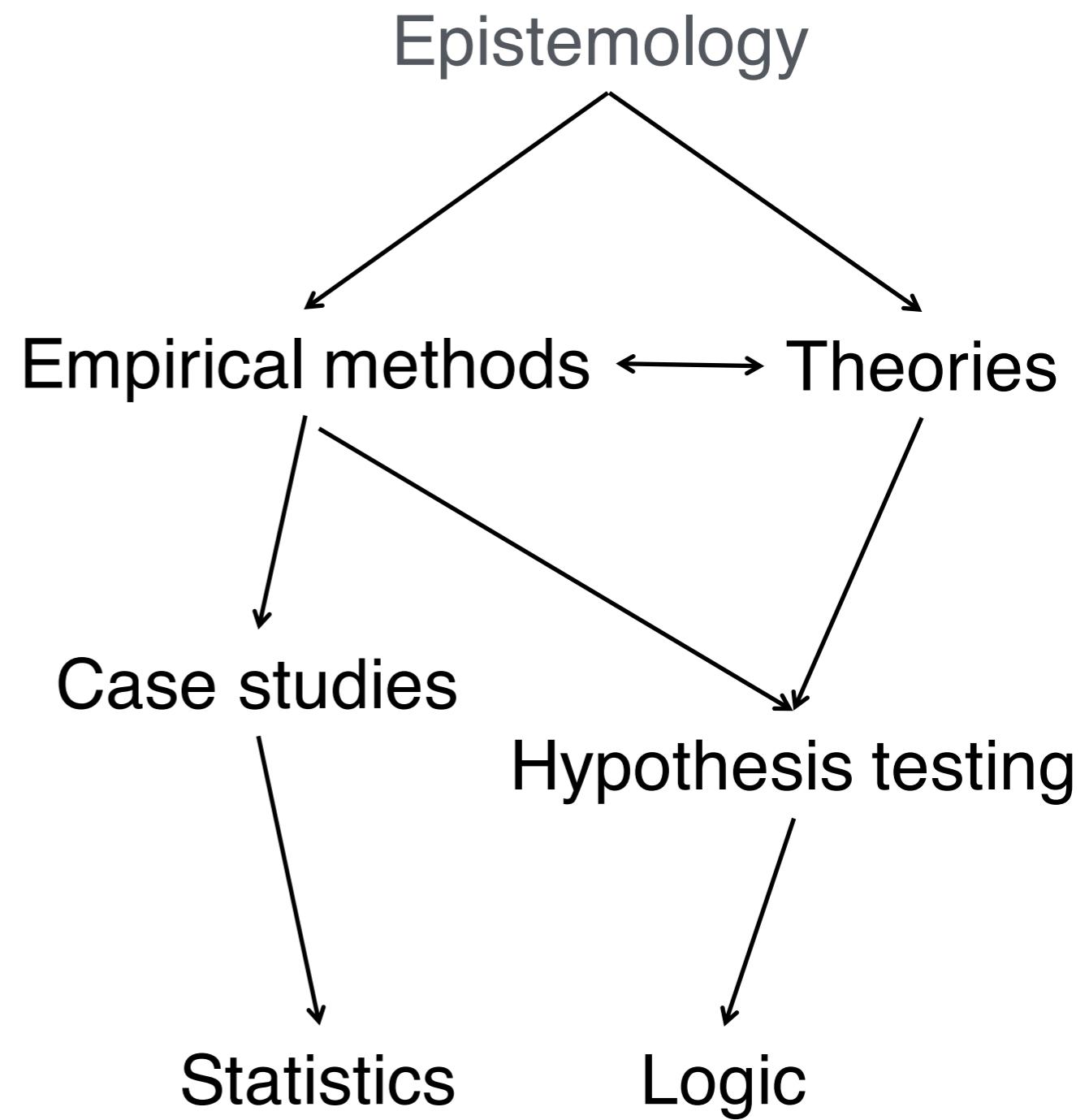
Examples

Philosophy of science

Principle ways of working

Methods and Strategies

Fundamental tools



# ESE relies on every layer!

**Philosophy of science**

**Principle ways of working**

**Methods and Strategies**

**Fundamental tools**

**Setting of Empirical Software Engineering:**

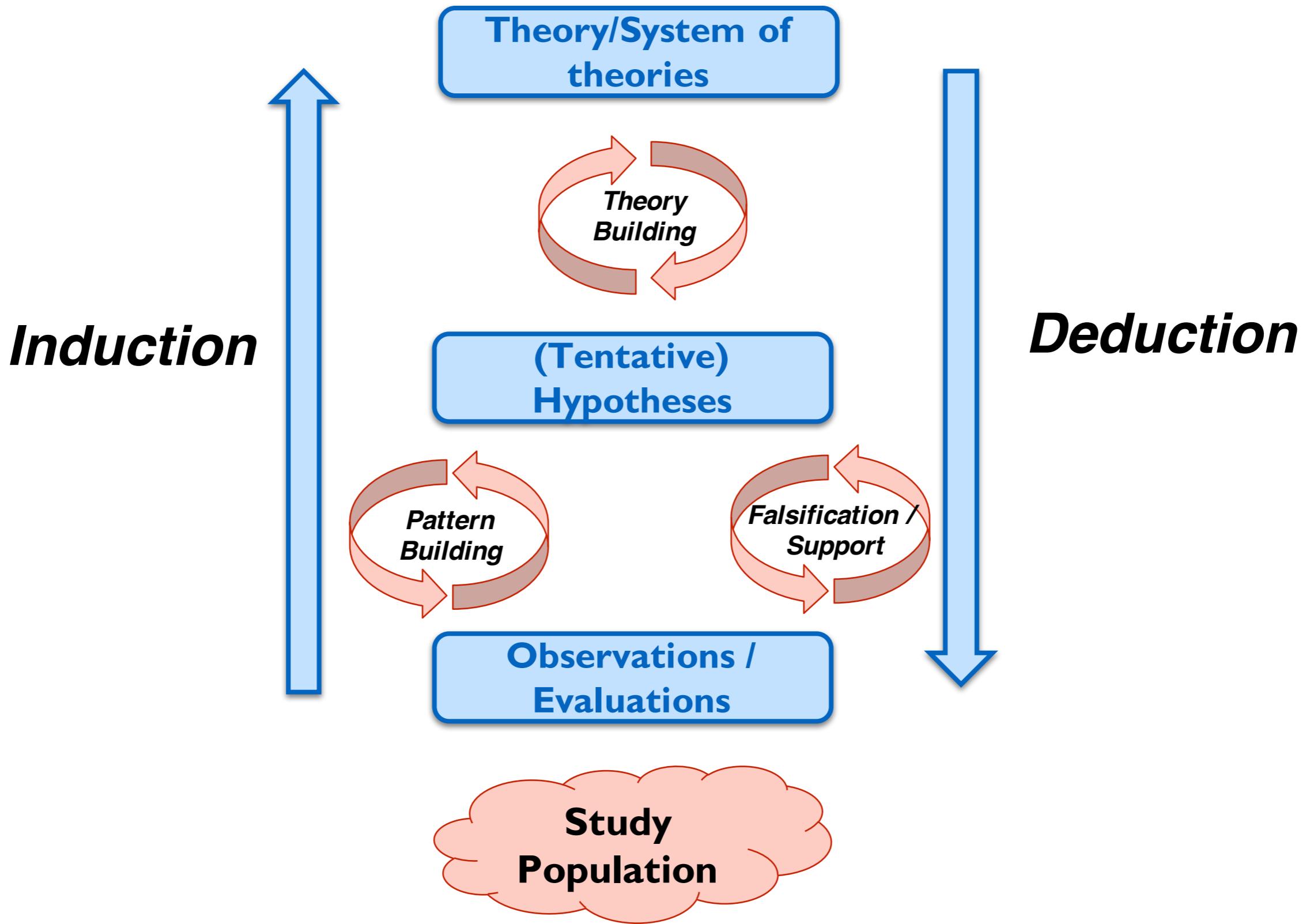
- Theory building and evaluation

are supported by

- Methods and Strategies

Analogy:  
Theoretical and Experimental Physics

# Big Picture... 2<sup>nd</sup> layer



# Big Picture 3<sup>rd</sup> layer: Methods

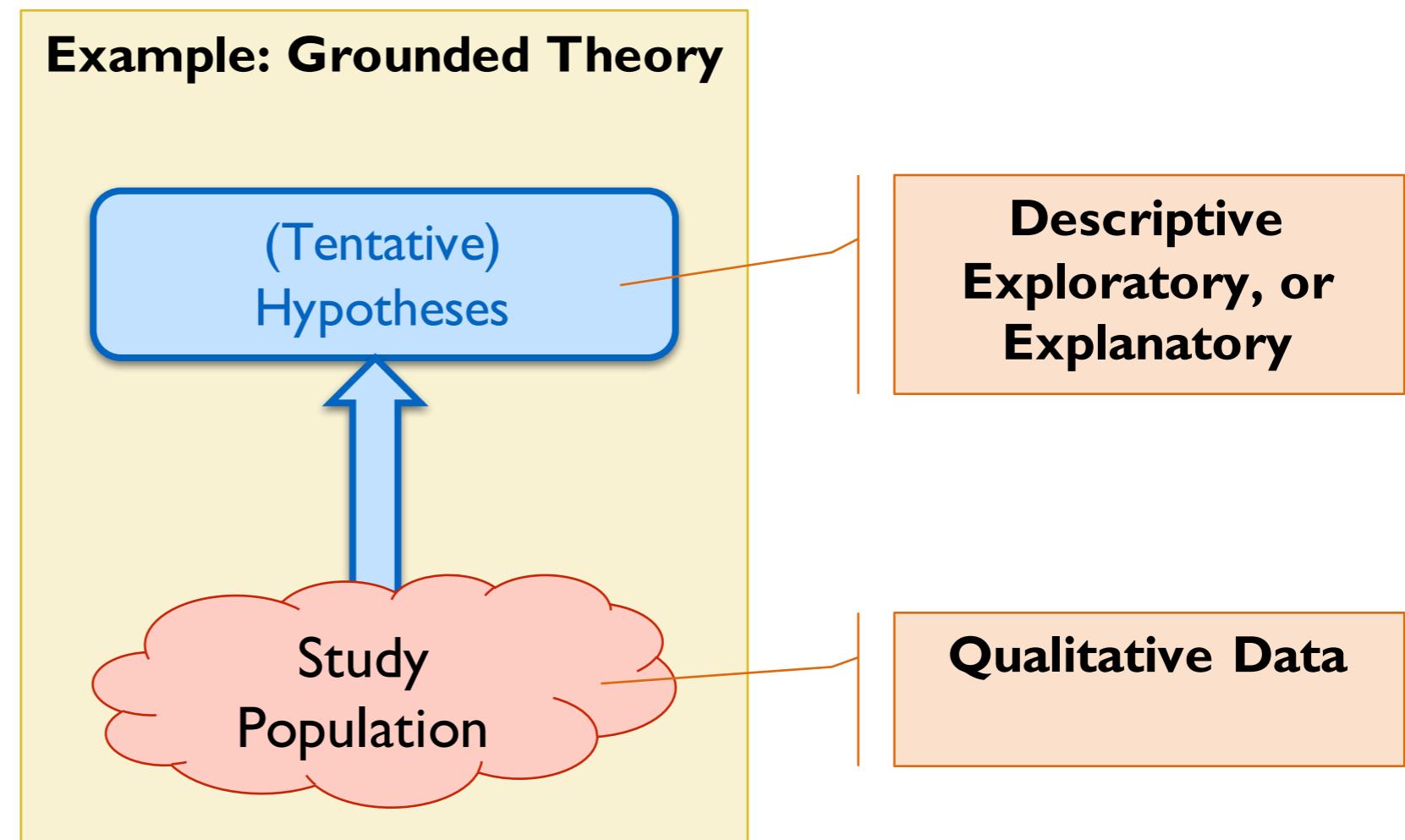
- Each method I can apply...
  - Has a specific **purpose**
  - Relies on a specific **data type**

## Purposes

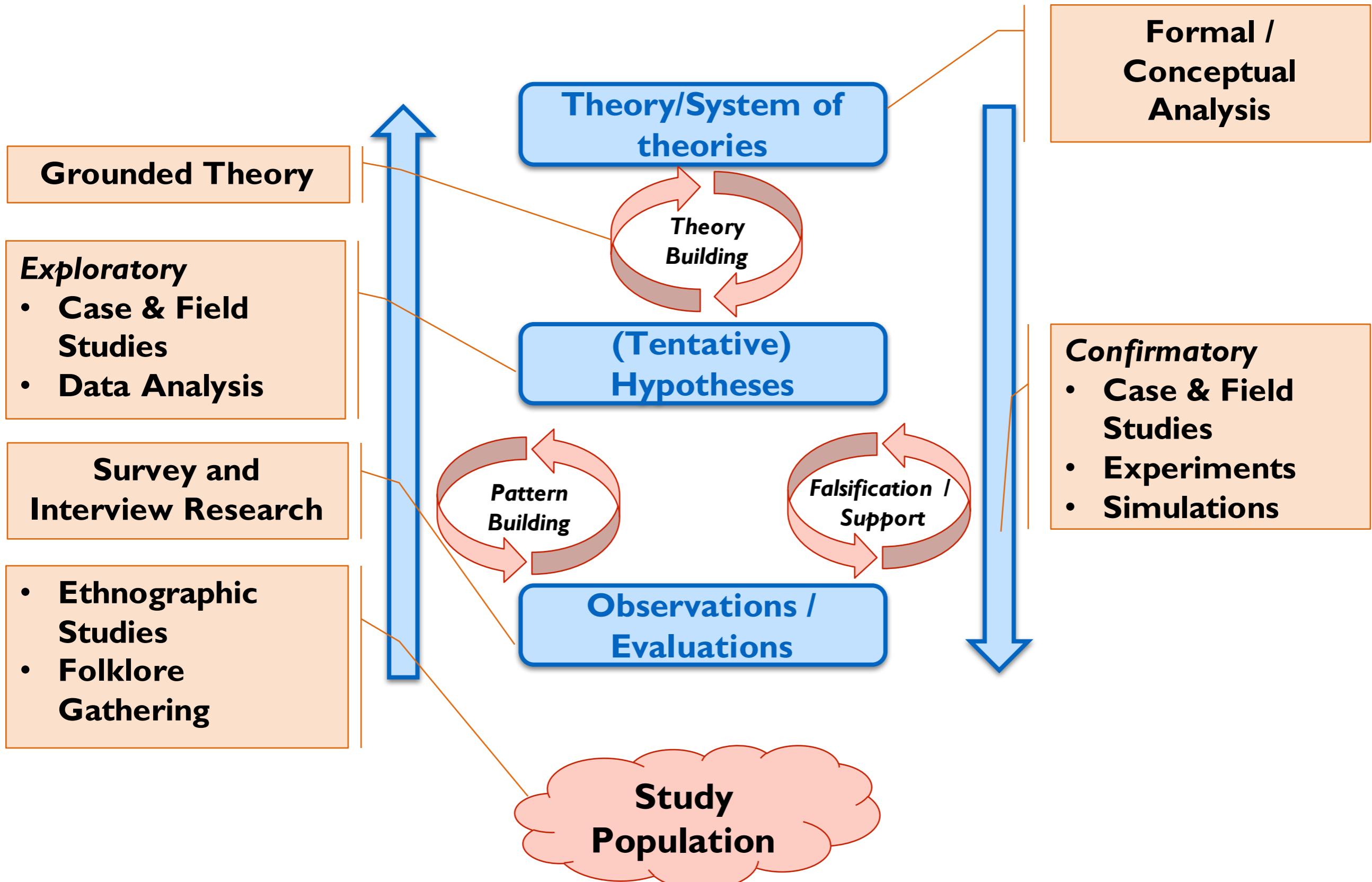
- Exploratory
- Descriptive
- Explanatory
- Improving

## Data Types

- Qualitative
- Quantitative



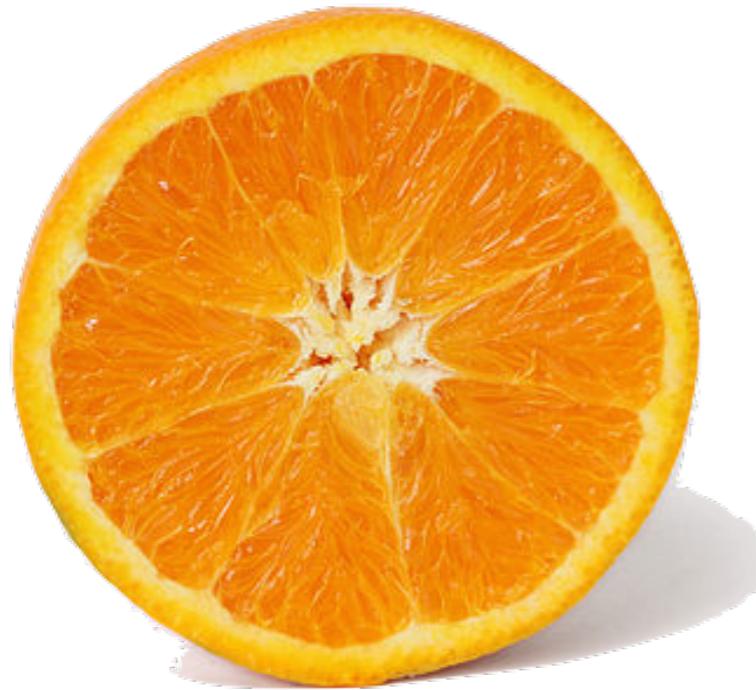
# Big Picture 3<sup>rd</sup> layer: Methods



Further reading: Vessey et al

A unified classification system for research in the computing disciplines

# Observational Studies



Survey  
(Cross-Sectional)



Case study



Case-Control

# Survey

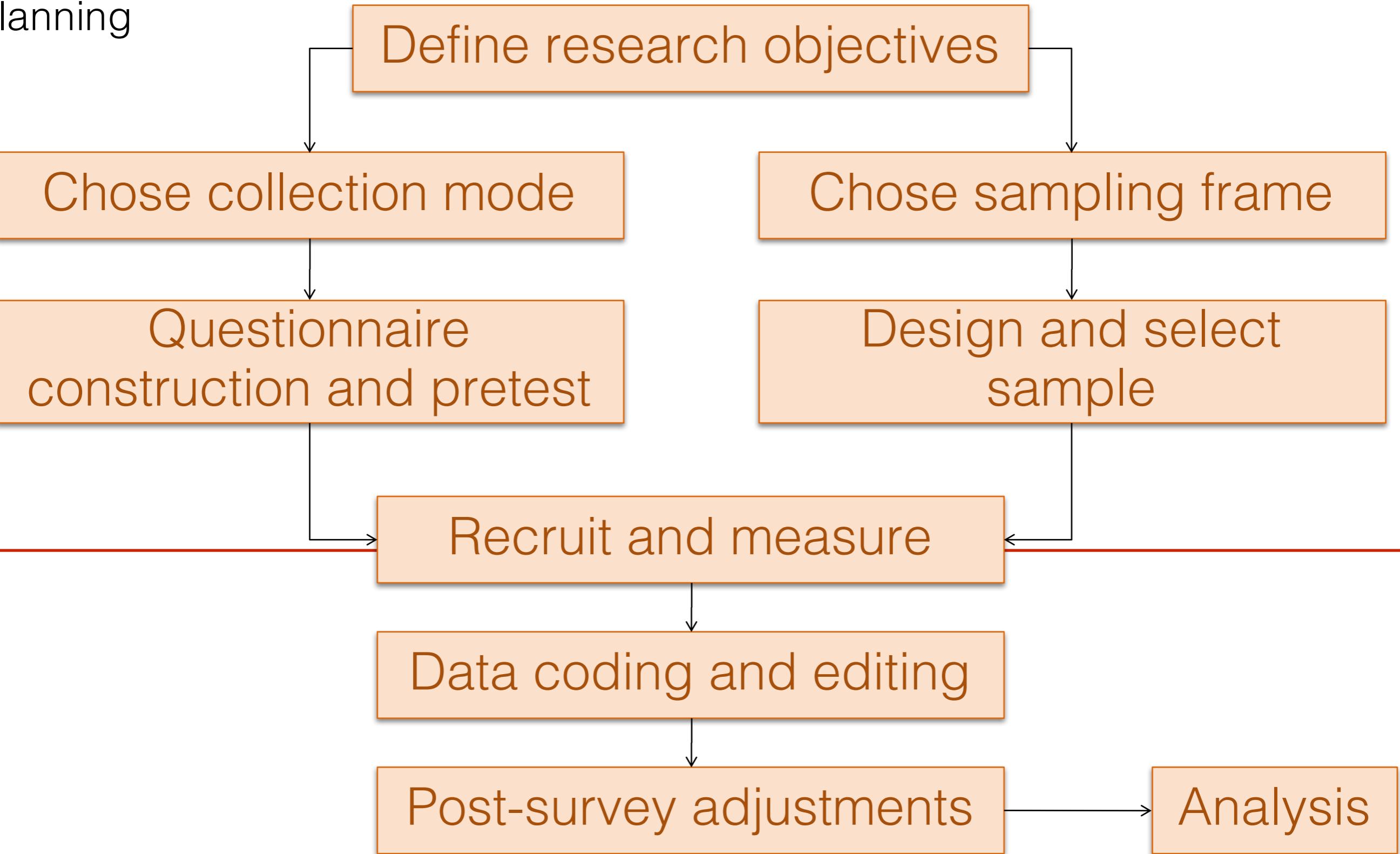
Systematic observational method to gather qualitative and/or quantitative data from (a sample of) entities to characterize information, attitudes and/or behaviors from different groups of subjects regarding an object of study



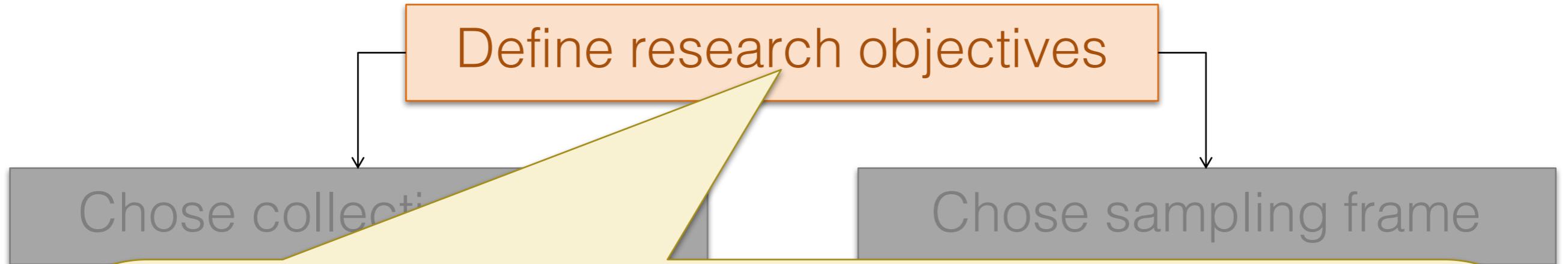
Descriptive statistics +  
Analytic statistics

# Survey process

Planning



# Survey process



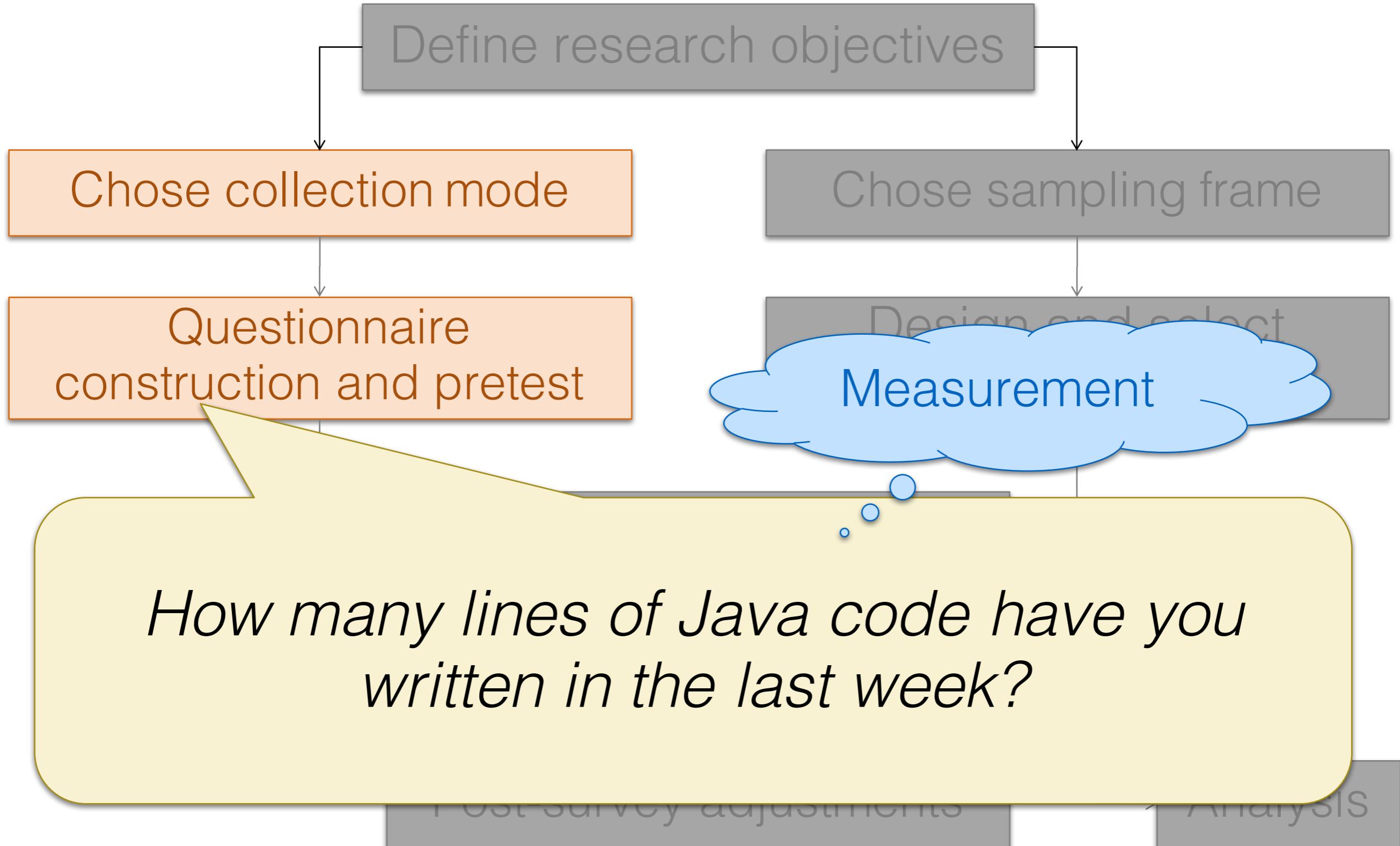
Research question +  
Characterization of  
constructs and population

Parameter/  
Construct

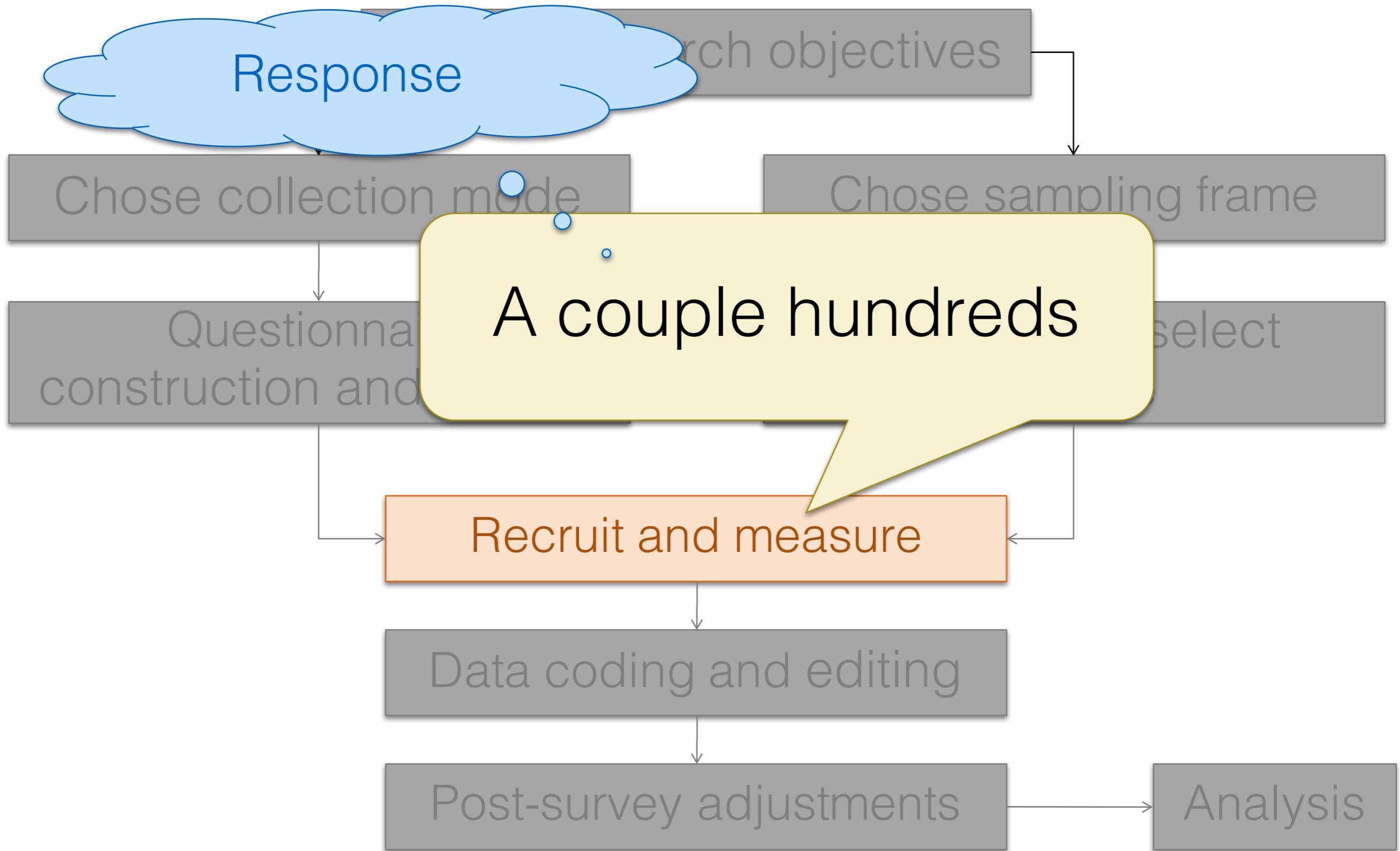
*What is the productivity of  
Java software developers?*

Target population

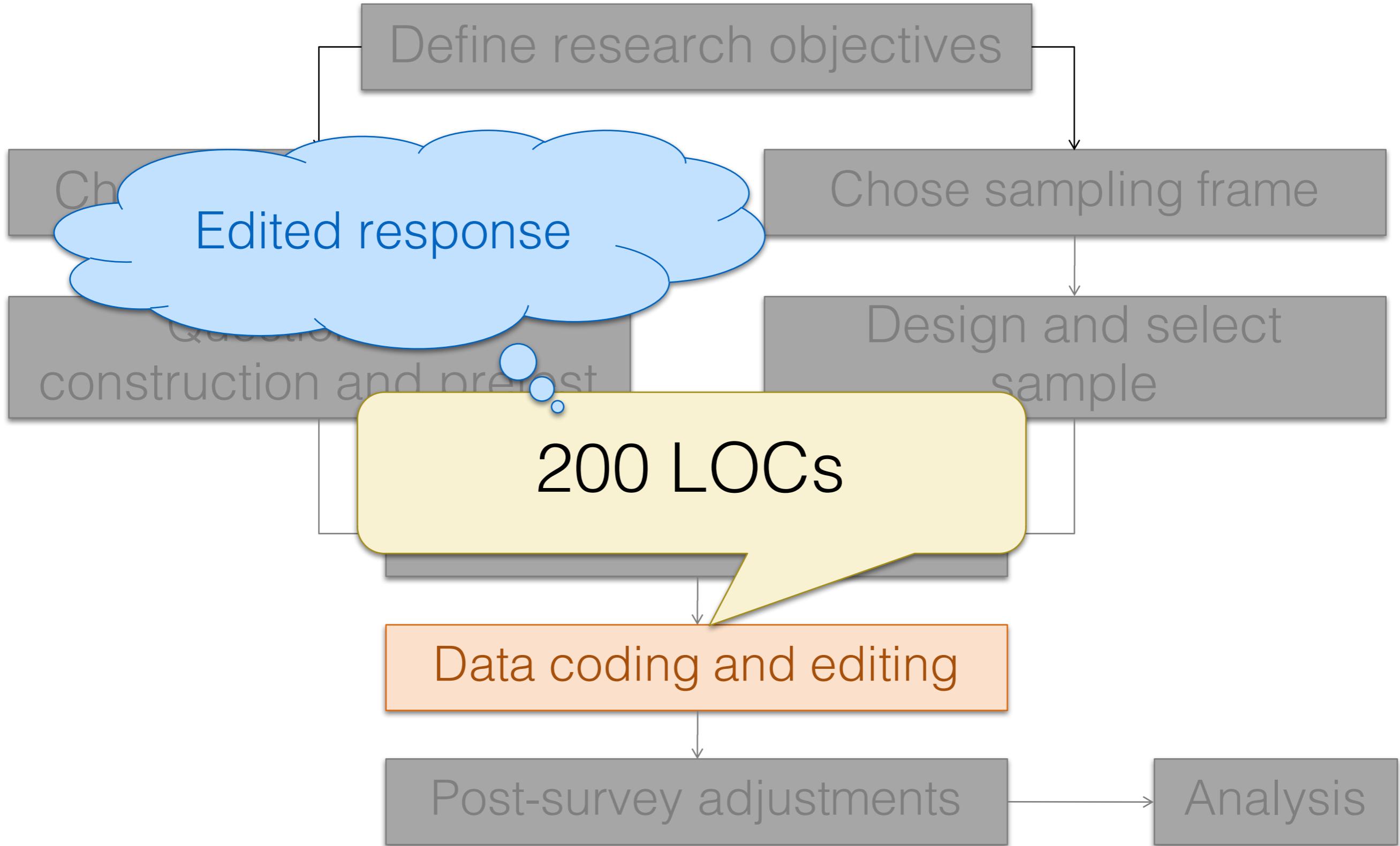
# Survey process



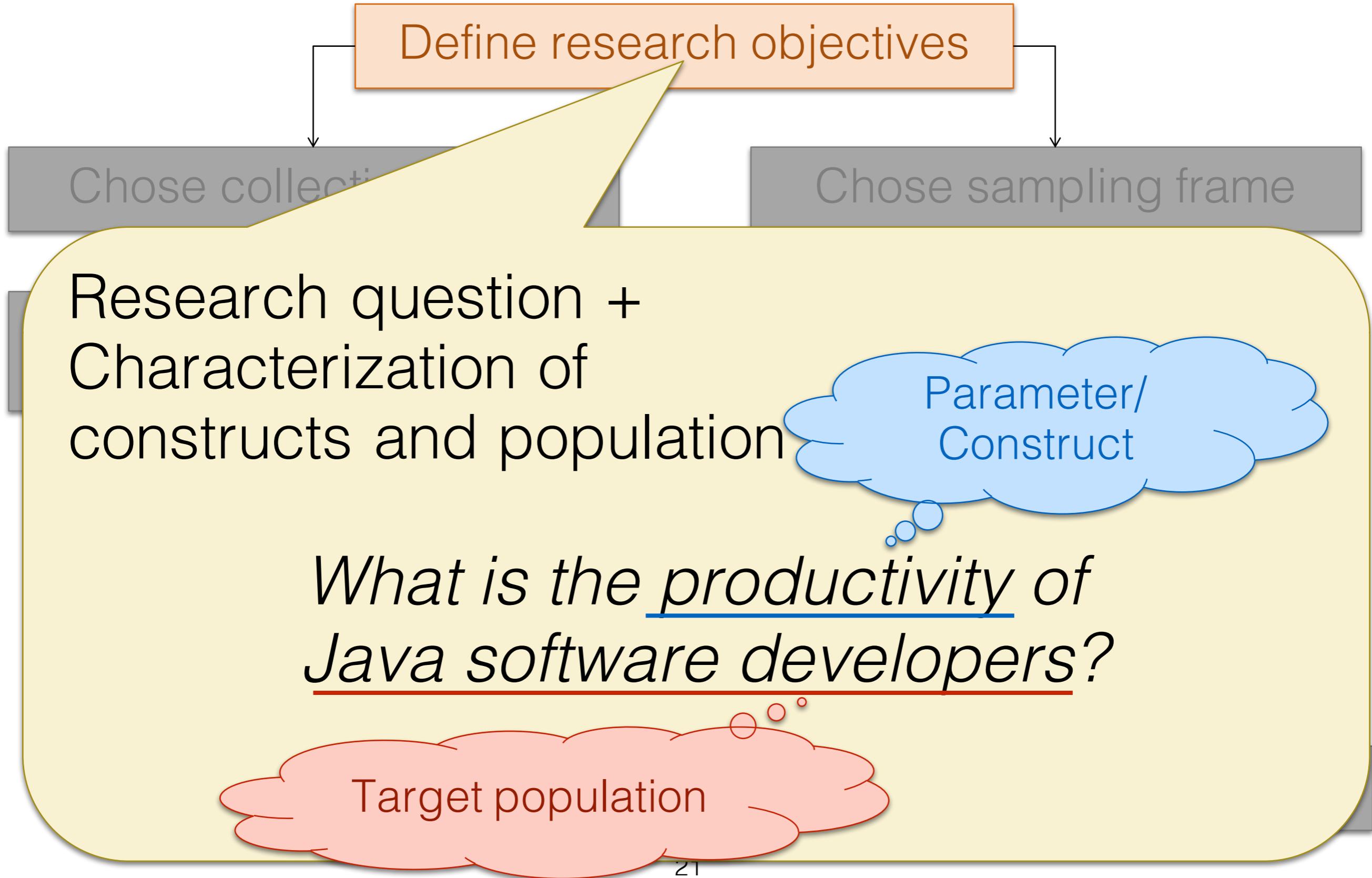
# Survey process



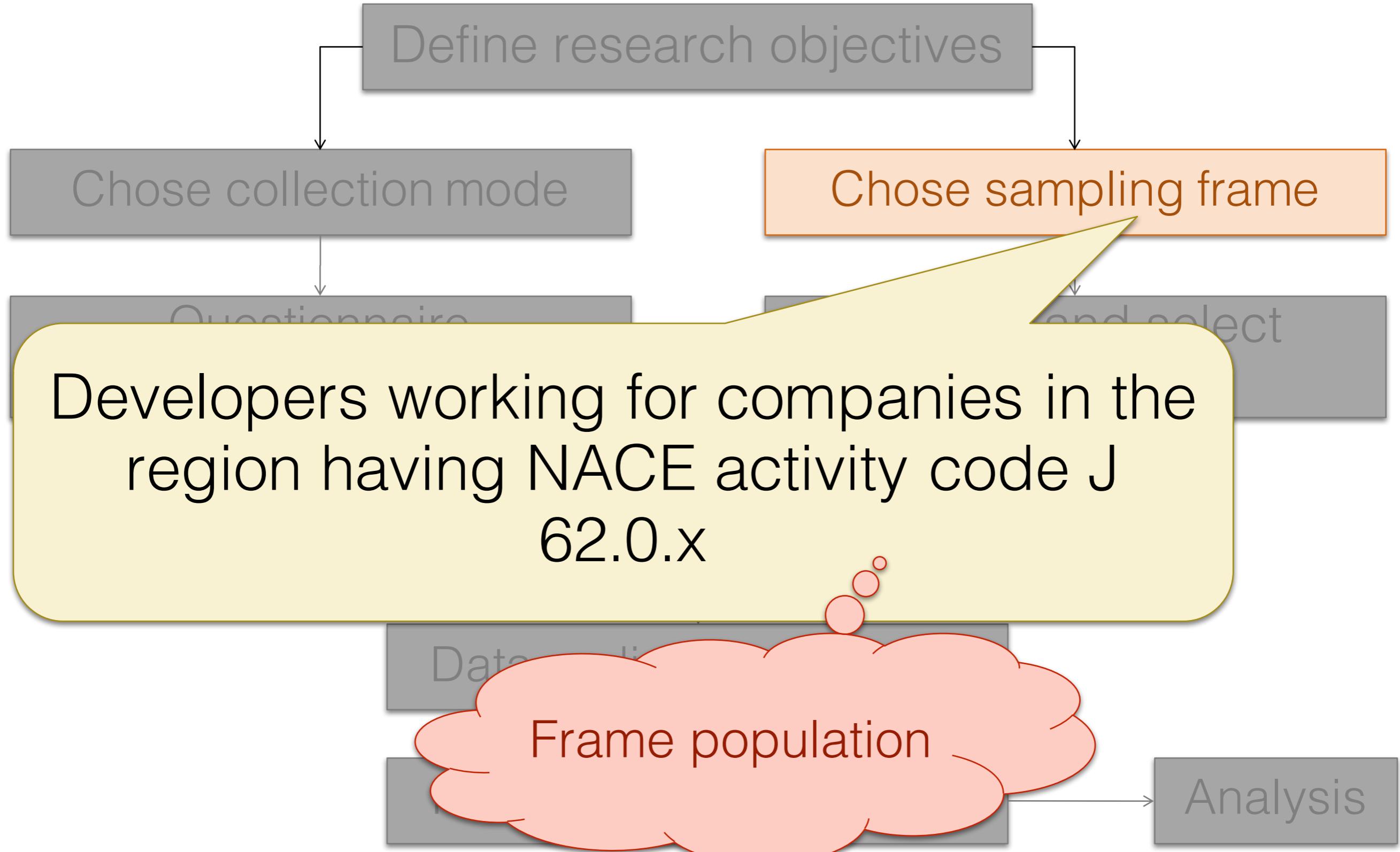
# Survey process



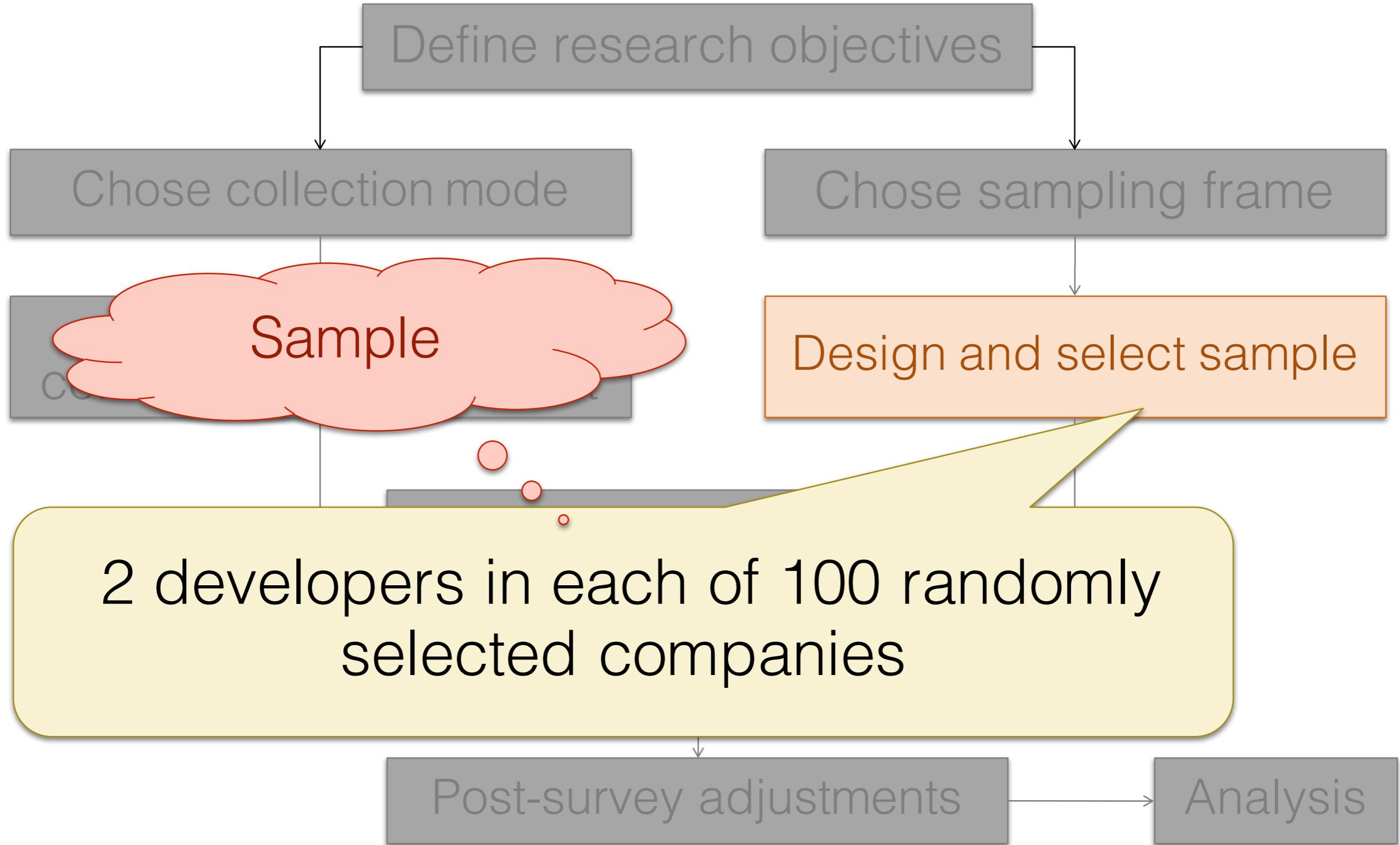
# Survey process



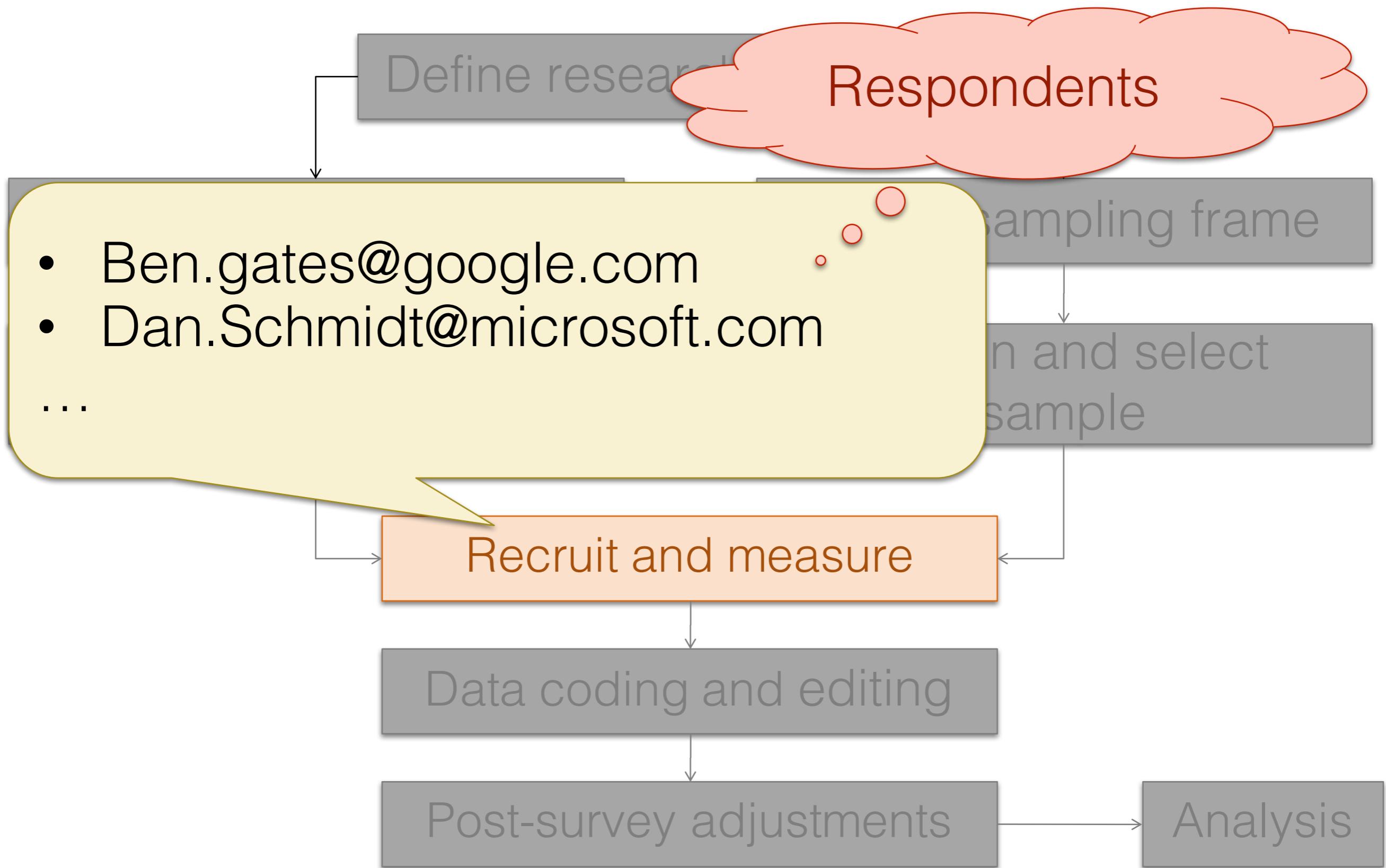
# Survey process



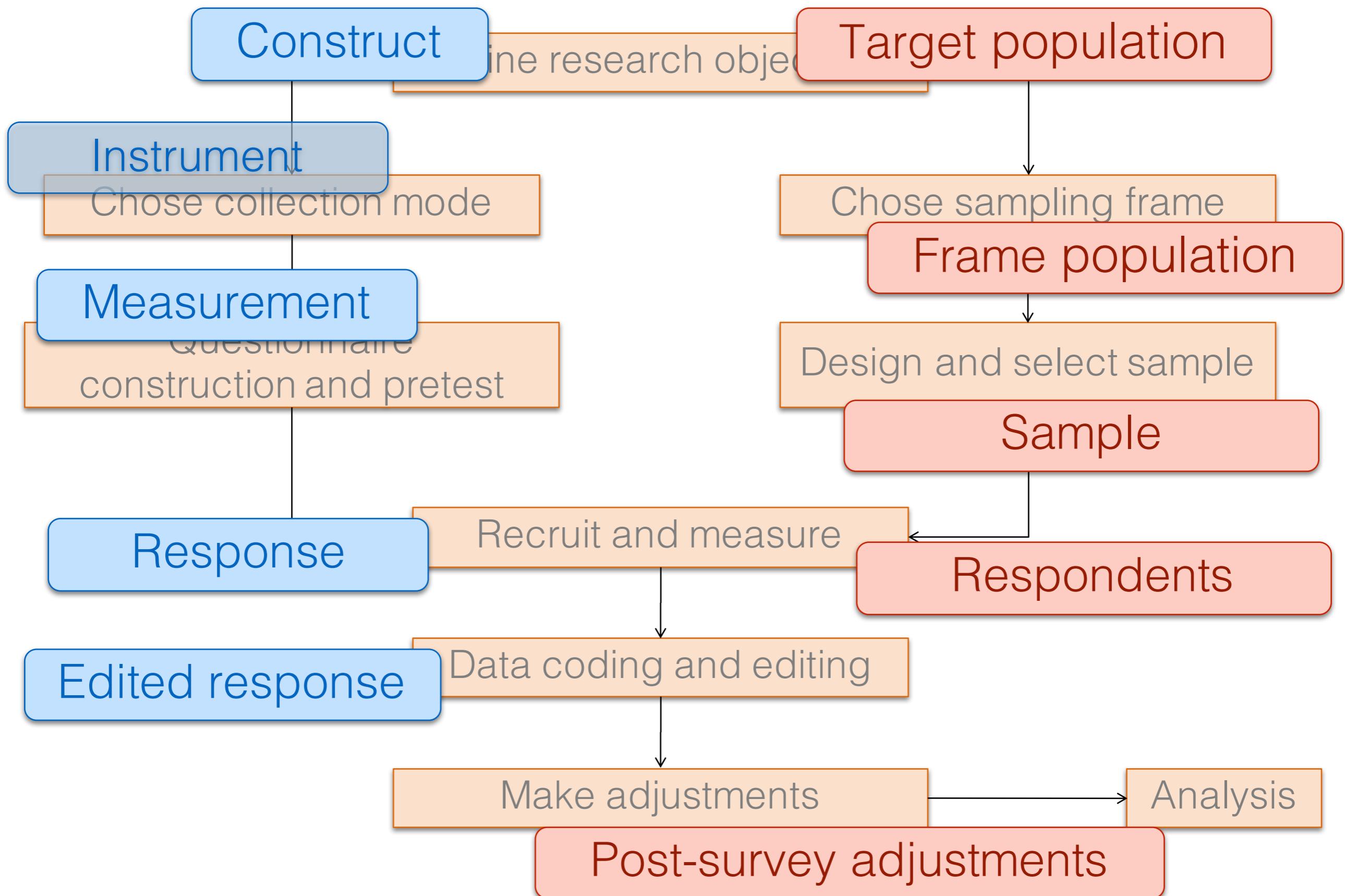
# Survey process



# Survey process

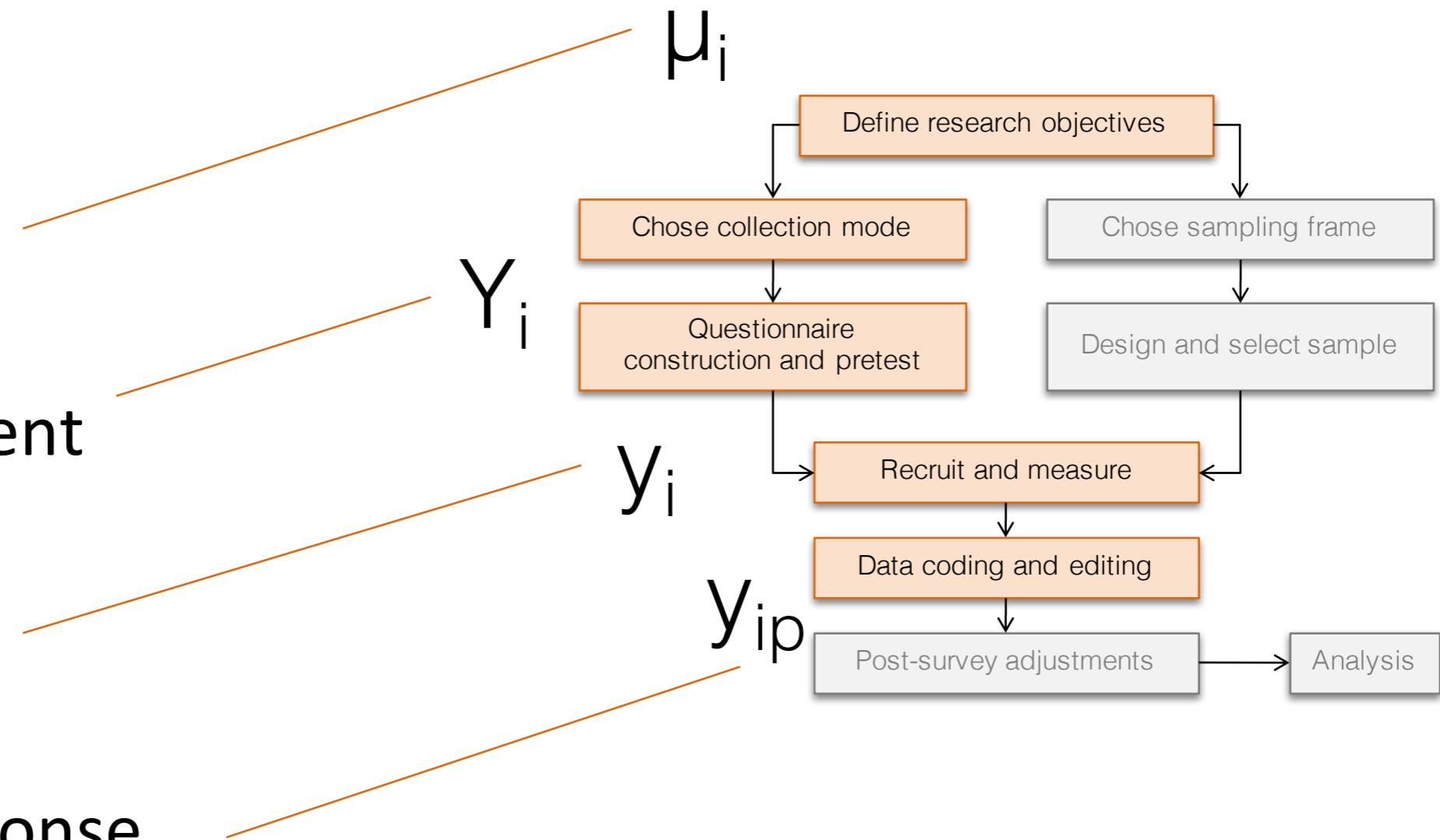


# Measurement vs Representation



# Measurement perspective

- Construct
- Measurement
- Response
- Edited response



# Construct

$\mu_i$

- Element of information sought by researchers
- Examples
  - How many new jobs created
  - How many incidents of crime with victims
  - Which developments tools used
- Formulation
  - Easy to understand
  - Imprecise
  - Abstract

# Construct

- Level of Abstraction / Measurement perspective
  - Directly observable
    - E.g. Staff for a project
    - A few defined ways to measure
  - Non directly observable
    - Intention to adopt a technology
    - No single well-defined measure

# Measurement

Y<sub>i</sub>

- How to gather information about constructs
  - Objective measures
    - Electronic
    - Physical
  - Answers to questions
    - Visual (formal questionnaires)
    - Oral (structured or semi-structured interviews)

# Validity

- Gap between constructs and measurement
  - Ideally the measure is the result of just one among several possible trials
  - In practice the measurement may introduce an error
    - Each trial introduces a different error

$$Y_{it} = \mu_i + \epsilon_{it}$$

- Validity := correlation between  $Y$  and  $\mu$

# Response

$y_i$

- The actual data collected through the survey
  - A question may require
    - Search own memory
    - Access records
    - Ask other persons
  - Closed questions already contain possible answers
  - Sometimes a response is not provided

# Measurement error

- Gap between the ideal measurement outcome and response obtained

$$y_i - Y_i$$

- Response bias
  - Systematic misreporting
- Reliability
  - Variability over several trials

# Edited response

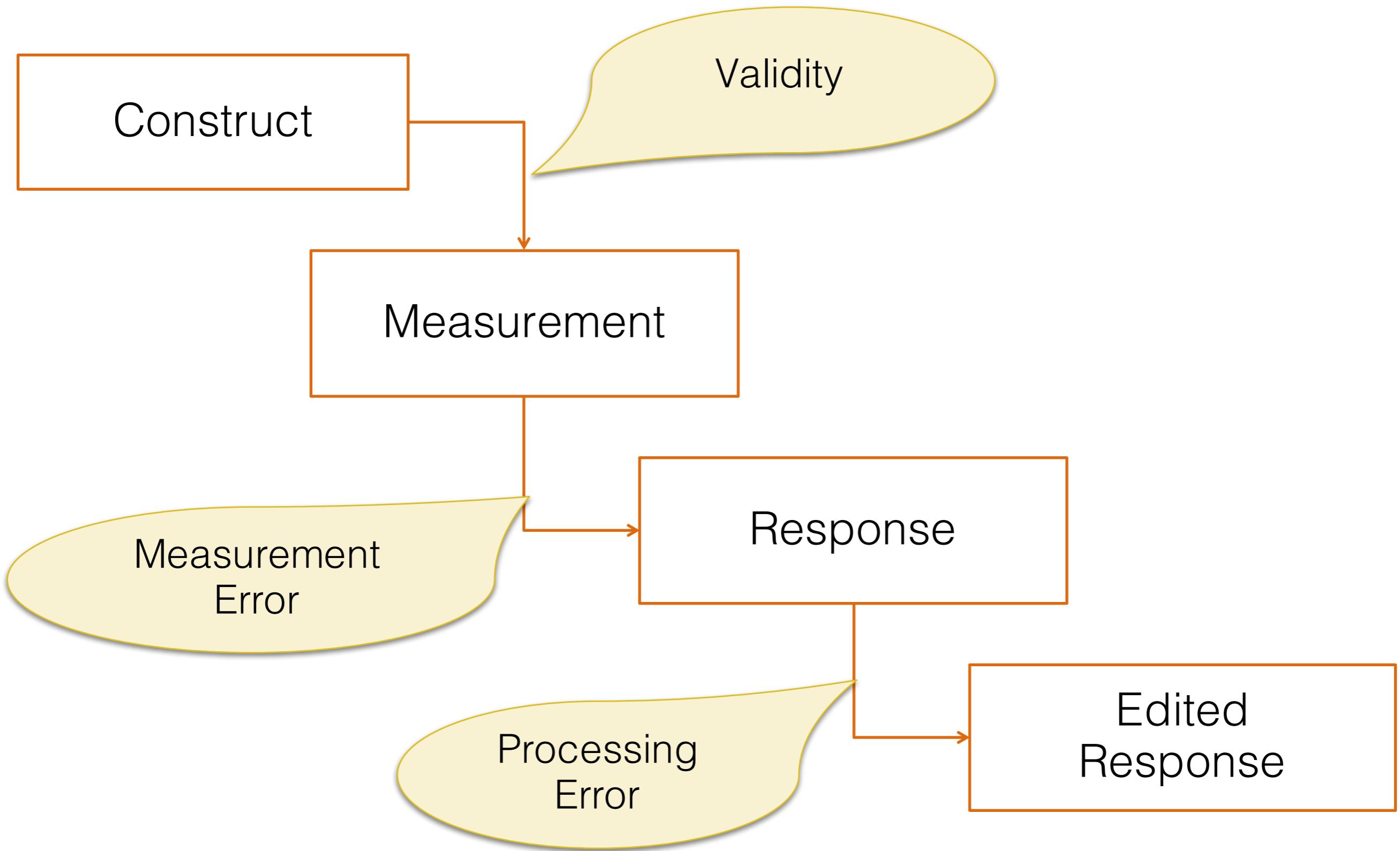
y<sub>ip</sub>

- Review process before using data
  - Range checks
  - Consistency checks
  - Illegible answers detection
  - Skipped questions
  - Outlier detection

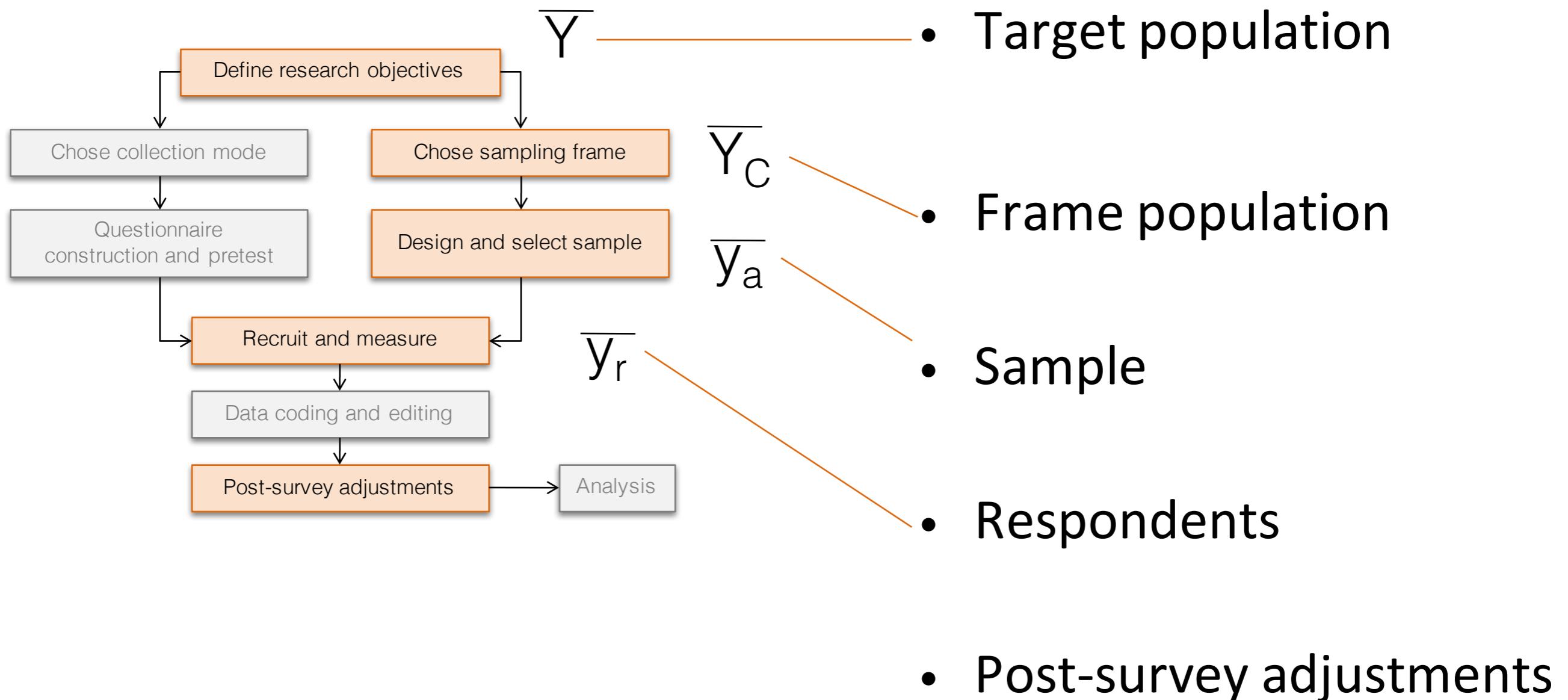
# Processing error

- Gap between variables used in analysis and those provided by the respondent
  - Erroneous outlier identification
  - Coding error

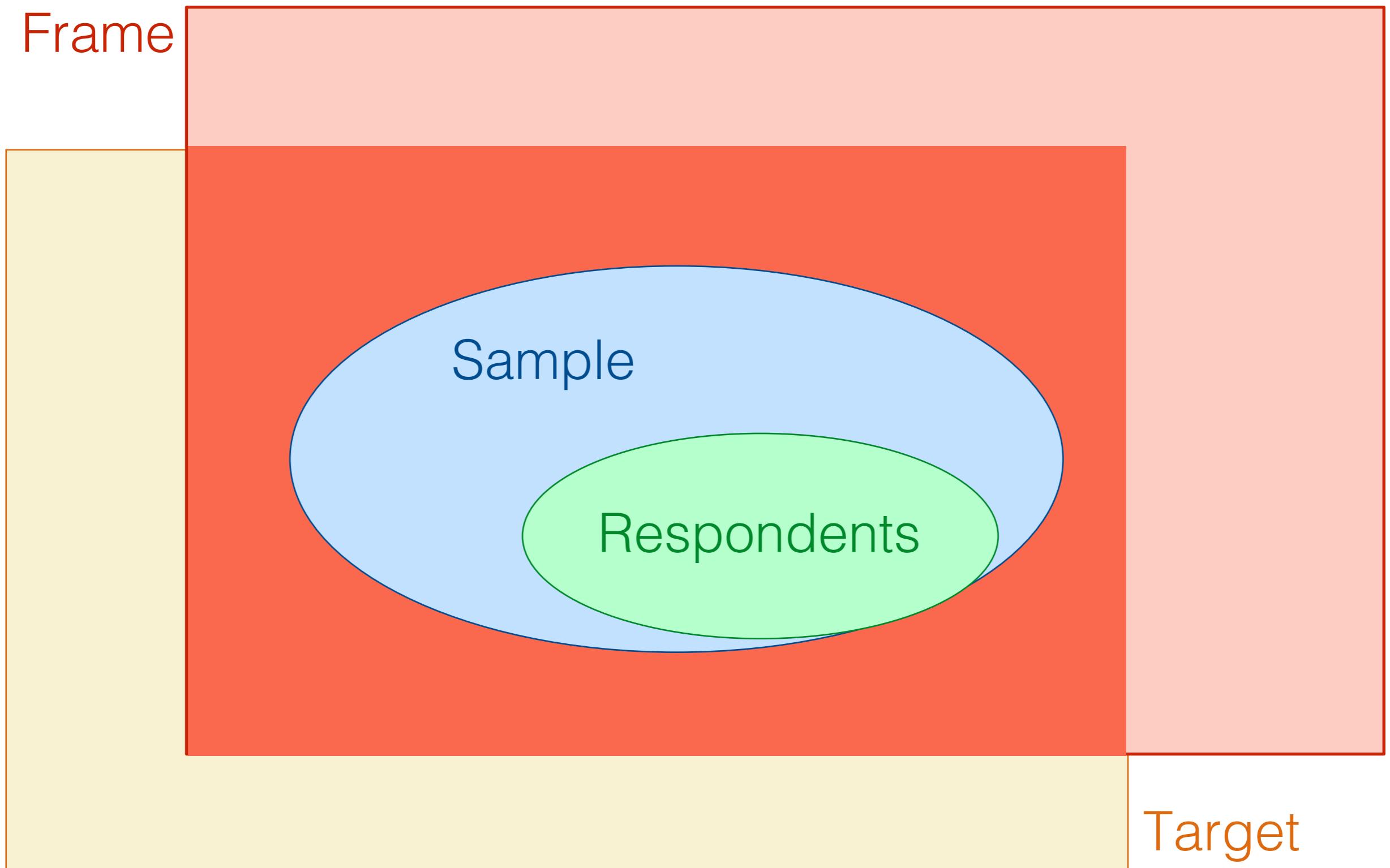
# Errors – measurement pov



# Representation perspective



# Representation perspective



# Target population

- The set of units to be studied
  - Abstract population definition
  - E.g. software projects
    - Time?
    - In software companies only?
    - Italian companies only?
    - Completed or just started projects?

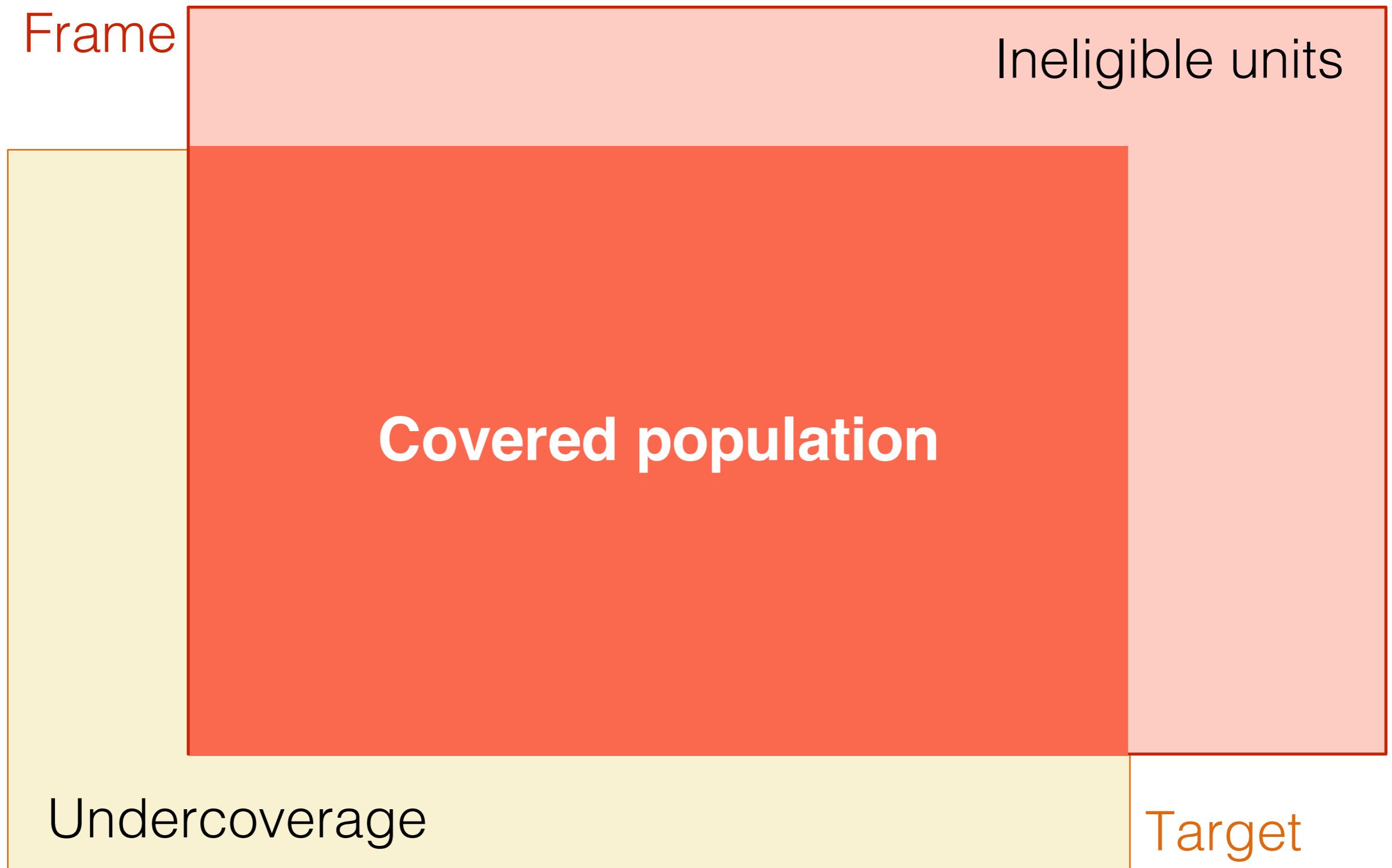
# Frame population

- All units in the sampling frame constitute the frame population
- In theory
  - The subset of target population that has a chance to be selected
- In practice
  - a set of units imperfectly linked to the target population members
    - E.g. telephone numbers

# Framing instrument

- The (conceptual) instrument used to identify the units of study
  - Household phone numbers to get persons
  - Company records to get employees
  - Customer IDs to get customers
  - Social Network IDs to get members
- Warning: often the frame elements are indirectly linked to the units of analysis, through respondents. E.g.
  - UoA: software projects (UoA)
  - Respondents: developers
  - Frame elements: software companies

# Coverage Error



# Reality check

- During 2012 USA Presidential Elections Campaign because of an effect of Federal regulations polling cellphones was more expensive
- As a result, many public polls leave cellphone users out of their samples
- Due to the growing popularity of cellphones as the only point of contact for young voters and minorities, pollsters left key constituencies for Obama out of the polls and skewed the numbers for Romney in some samples

*“That’s why some polls looked so difficult for the president, because they were under-polling the electorate for the president”*

J. Messina (Campaign Manager for Obama)

# Coverage bias

- Two factors
  - 1. Difference between covered and not covered population
    - $\bar{Y}$ : mean of target
    - $\bar{Y}_C$ : mean of covered     $\bar{Y}_U$ : mean of uncovered
  - 2. Proportion of non covered population
    - $C$ : # covered units       $U$ : # uncovered units

$$\bar{Y}_C - \bar{Y} = \frac{U}{C}(\bar{Y}_C - \bar{Y}_U)$$

# Sample

- Units selected from the frame population
  - Time and cost opportunity
- Sampling := Deliberate non-observation
  - May introduce deviation between
    - Sample statistic
    - Full frame statistic

# Sampling Design

- Strategy followed to establish the sample from the frame population
- Non-probabilistic sampling designs include:
  1. Accidental sampling (simply use convenience)
  2. Judgement sampling (apply some technical criteria to sample)
  3. Snowballing (share the sampling decision with part of the subjects)
  4. Quota Sampling (establish fixed quota by groups)

# Sampling Design

- Probabilistic sampling designs include:
  1. Simple random sampling (random select "n" units from the frame population)
  2. Stratified Sampling (simple random sampling from each stratum established in the frame population)
    - Example: To sample java developers by country

# Sample Size Formula

- Recommended when working with probabilistic sampling designs

$$SS = \frac{Z^2 \times p \times (1-p)}{c^2}$$

- **SS:** sample size
- **Z:** Z-value, established through a specific table (Z=2.58 for 99% of confidence level, Z=1.96 for 95% of **confidence level**)
- **p:** percentage selecting a choice, expressed as decimal (0.5 used as default for calculating sample size, since it represents the worst case).
- **c:** desired **confidence Interval**, expressed in decimal points (Ex.: 0.04).

# Sample Size Formula

- Correction formula based on a finite population with a *pop* size

$$SS_f = \frac{SS}{1 + \frac{SS-1}{pop}}$$

Population	Confidence Level	Confidence Interval	Sample Size
10,000	95%	0.01	4,899
10,000	95%	0.05	370
500	95%	0.01	475
500	95%	0.05	217

# Sampling error

- Sampling bias
  - Systematic exclusion of some members
  - Or significantly reduced chance of selection
- Sampling variance
  - Ideal set of samples all drawn from the same frame

$$V_s = \frac{\sum_{s=1}^S (\bar{y}_s - \bar{Y}_C)^2}{S}$$

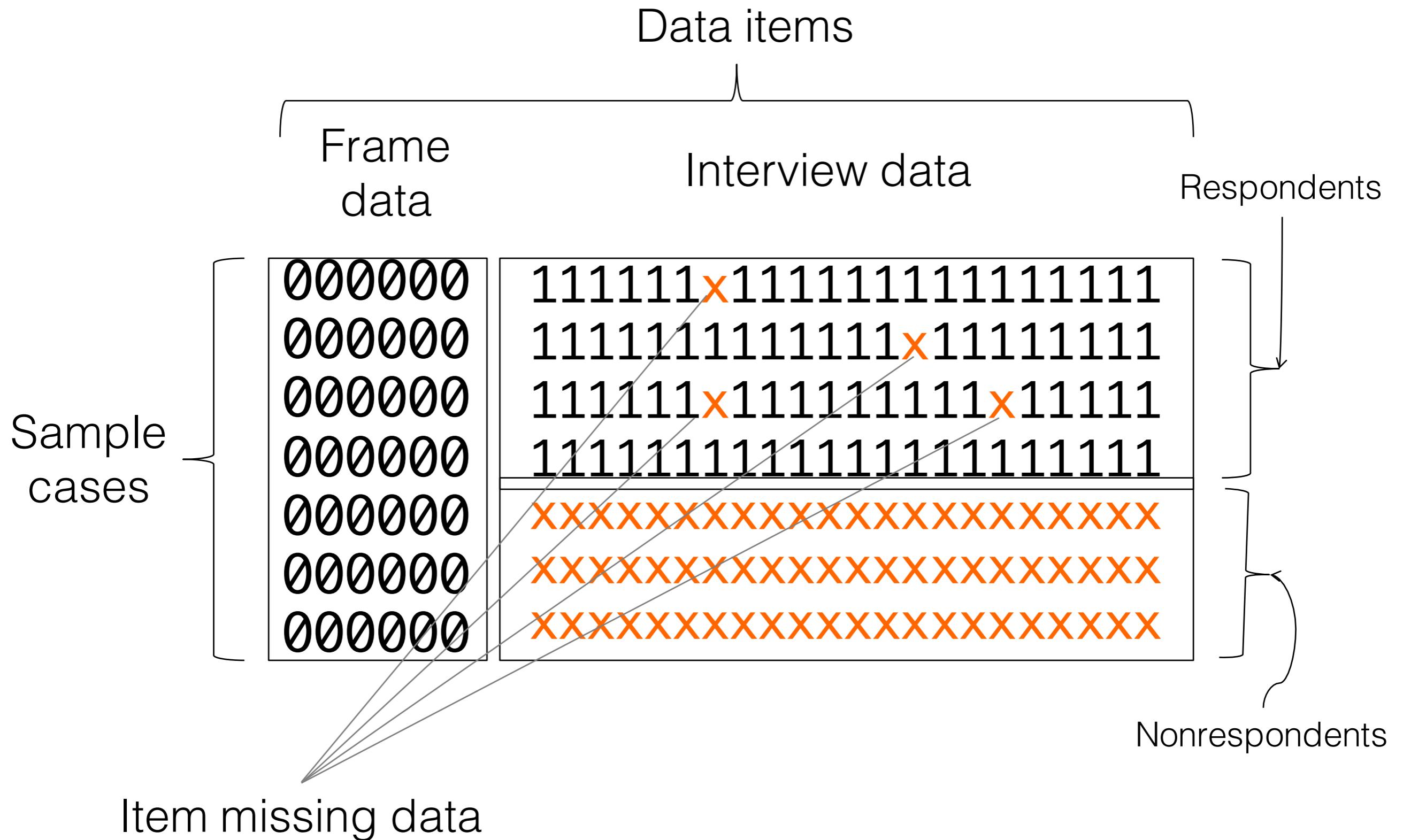
# Sampling error reduction

- Probabilistic sampling
  - ✓ All units have non zero selection probability
- Stratified sampling
  - ✓ Representation of key sub-populations is controlled
- Element samples
  - ✓ As opposed to cluster samples
- Sample size

# Respondents

- The subset of sample for which a measurement could be collected
  - ✓ Item missing data: incomplete measures
- Full participation (i.e. 100% response rate) realistically possible only for inanimate units

# Respondents



# Non-response error

- Non-response bias
  - ✓ Non-response rate:  $m_s / n_s$
  - ✓ Difference between respondents and non-respondents

$$\bar{y}_r - \bar{y}_s = \frac{m_s}{n_s} (\bar{y}_r - \bar{y}_m)$$

# Post-survey adjustment

- Weighting
  - ✓ Compensate under-representation due to
    - ✓ Non response patterns
    - ✓ Mismatch between frame and target population
- Imputation
  - ✓ Item missing data are replaced by estimations

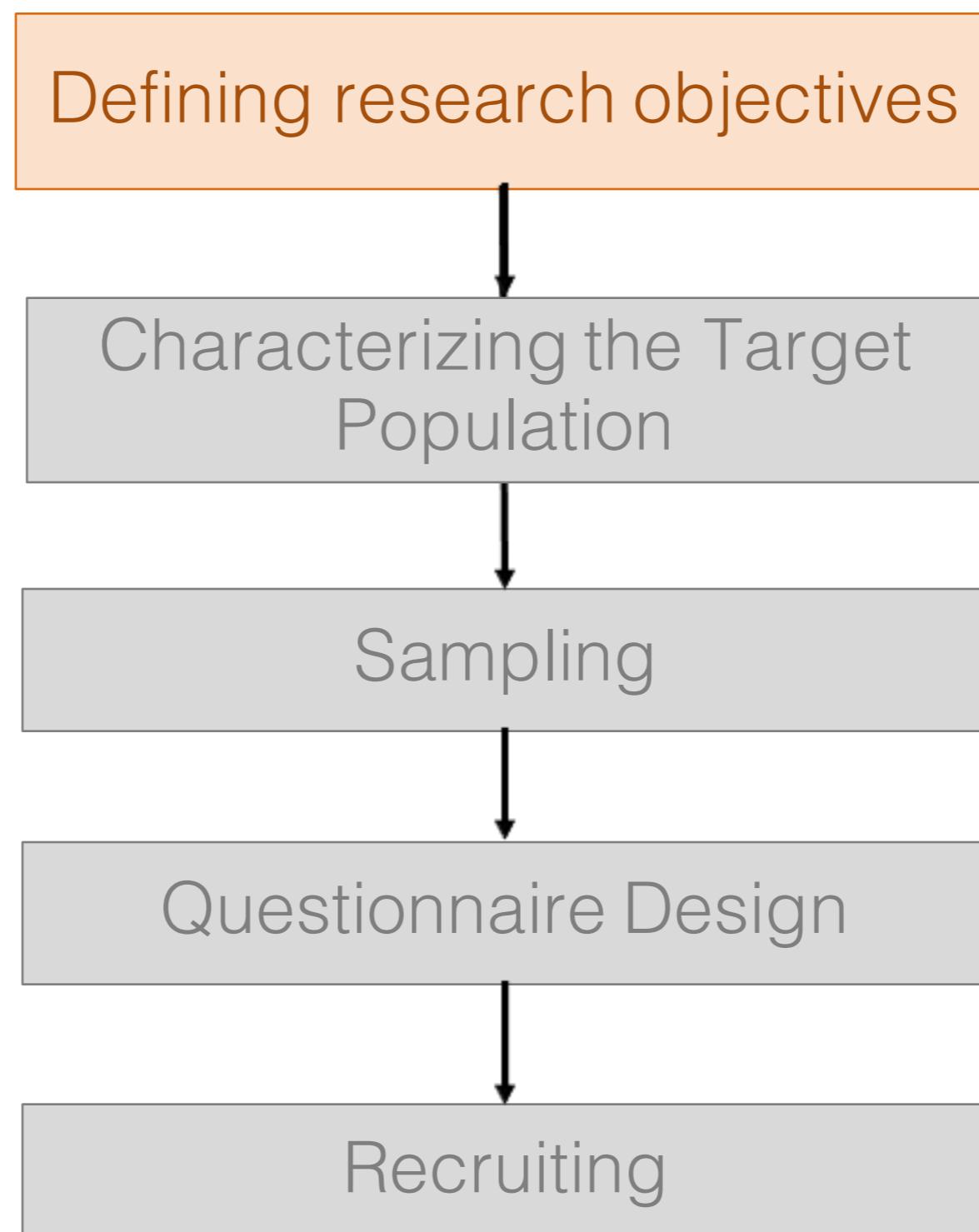
## Session II

# Best Practices on Planning Surveys

**Disclaimer**  
**There is no universal silver bullet!**



# Best Practices



# Design research objectives

**Challenge:**

**Know the limitations of survey research**

- Survey research opts for answers that rely on experiences, opinions, and observations (folklore) of the respondents
- Develop internal questions to help you depicting the research objective
- Opt for descriptive questions (“what is happening?”) or explanatory questions (“why is this happening?”) rather than normative questions (“what should we do?”)

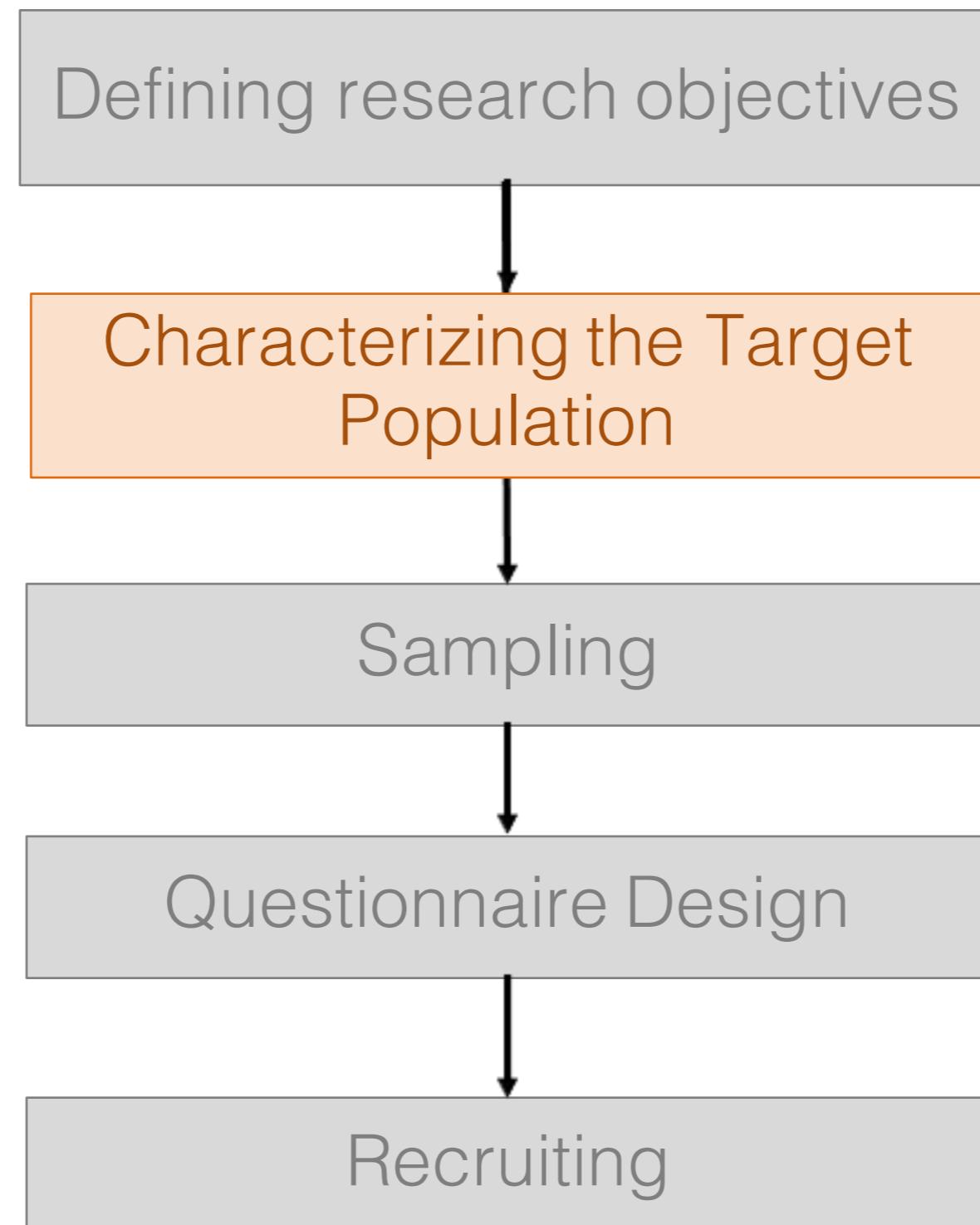
# Design research objectives

**Challenge:**

**Identify the real target population**

- Avoid to restrict the target population based on factors such as its size or its availability.
- Based on the research objectives, answer the following question: “Who can best provide you with the information you need?”, instead of answering “Who are probably available to participate?”

# Best Practices



# Characterizing the Target Population

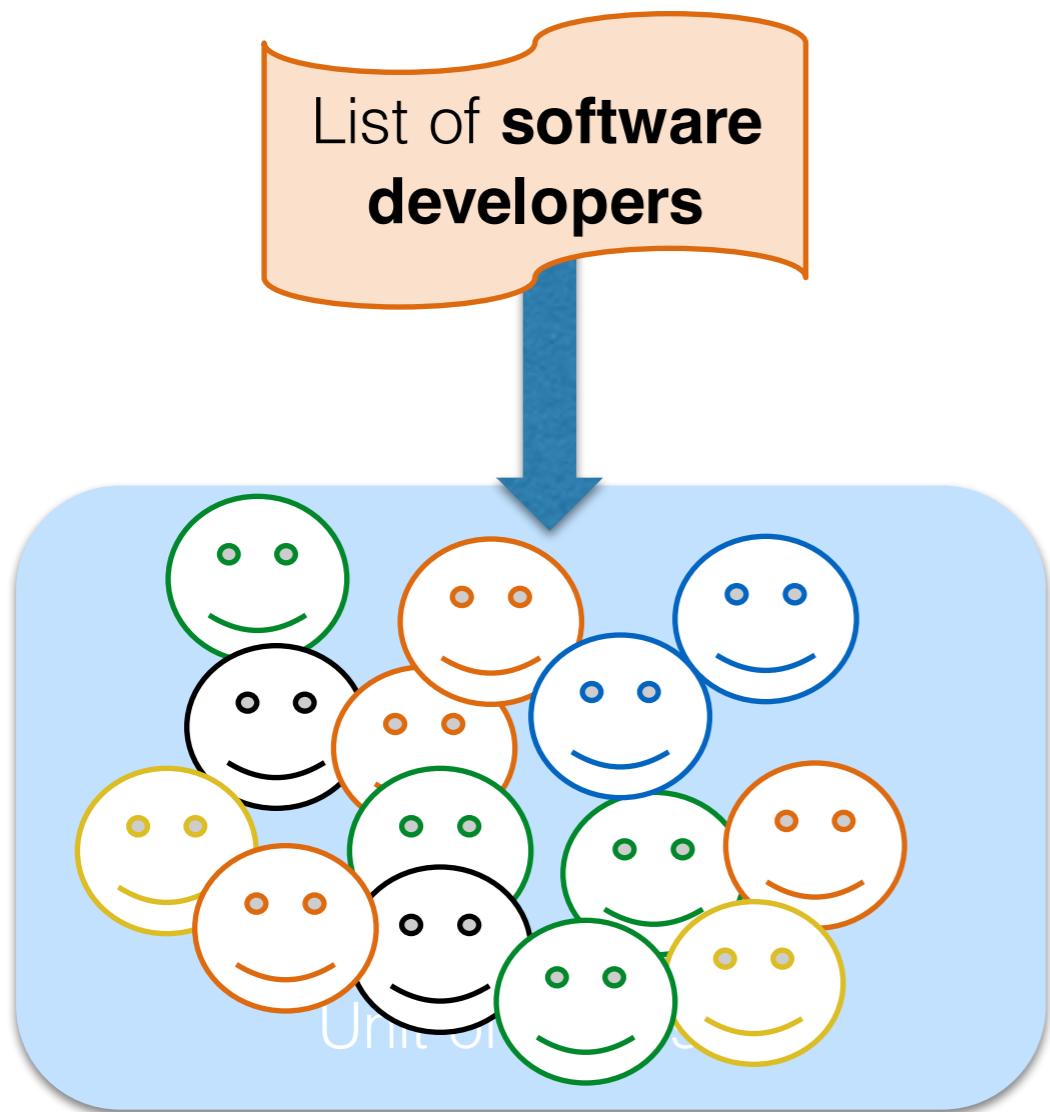
**Challenge:**

## **Identifying the survey unit of analysis**

- The survey subjects (individuals) are typically the survey unit of analysis. However, in some cases it may be a group of individuals, such as households or organizational units/ project teams in SE research
- Based on the research objective, identify which entity should be used to guide sampling and data analysing
  - For instance, investigating **Java developers** programming practice is a research objective different from investigating java programming practice in **software houses**

# Characterizing the Target Population

*How do developers perform code debugging?*



*How code debugging have been performed by developers from software houses?*



# Characterizing the Target Population

## Challenge:

### **Characterizing the subjects and units of analysis in SE surveys**

- Different research objectives may demand different attributes to characterizing individuals/ groups of individuals involved in the surveys
- What attributes are necessary to identify a “representative” population?
- Standards can be especially helpful to provide scales and even nominal values
  - For instance, CMMI-DEV maturity level can be used to characterize organizational units regarding their maturity in software process. RUP roles can be used to characterize subjects’ current position

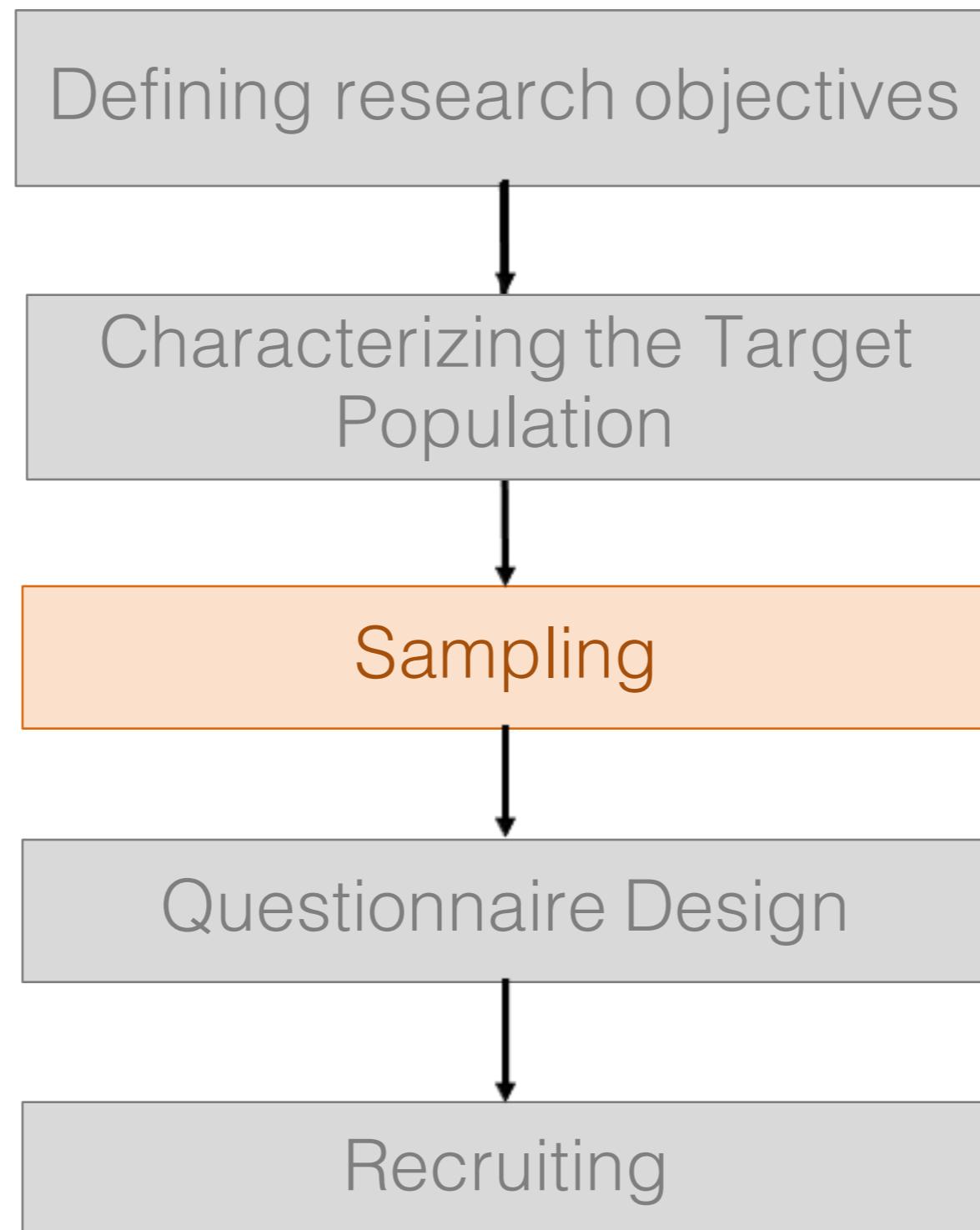
# Characterizing the Target Population

**Challenge:**

## **Characterizing the subjects and units of analysis in SE surveys**

- **Individuals** can be characterized through attributes such as: *experience in the research context, experience in SE, current professional role, location and higher academic degree]*
- **Organizations** can be characterized through attributes such as: *size (scale typically based in the number of employees), industry segment (software factory, avionics, finance, health, telecommunications, etc.), location and organization type (government, private company, university, etc.)*
- **Project teams** can be characterized through attributes such as *project size; team size, client/product domain (avionics, finance, health, telecommunications, etc.) and physical distribution*

# Best Practices



# Sampling

## Challenge:

### Looking for the Frame Population

- Suitable sampling frames are rarely available in SE research. We often need to perform “indirect sampling” (for instance, there is no yellow pages for software projects in a country).
- First of all, you should search for candidates of sources of population. Avoid the convenience on searching candidates, trying to answer: *“Where a representative population from the survey target population or even all target population is available?”*

# Sampling

An universe of imperfect alternatives!



# Sampling

## **A good source of population...**

- ...should not intentionally represent a segregated subset from the target population, i.e., for a target population audience “X”, it is not adequate to search for units from a source intentionally designed to compose a specific subset of “X”
- ...should not present any bias on including on its database preferentially only subsets from the target population. Unequal criteria for including search units mean unequal sampling opportunities
- ... allow identifying all source of population’ units by a distinct logical or numerical id
- ...should allow accessing all its units. If there are hidden elements, it is not possible to contextualize the population

# Sampling

**Challenge:**

## **Looking for the Frame Population**

- In the case of surveys having SE researchers as target population, you can use results from previously conducted systematic literature reviews (SLR) regarding your research theme
  - Social networks addressed to integrate academics such as *ResearchGate* and *Academia.edu* can be also useful in this context.

# Sampling

## Challenge:

### Looking for the Frame Population

- Look for catalogues provided by recognized institutes/ associations/ governments to retrieve relevant set of SE professionals/ organizations.  
Some examples:
  - SEI ([www.sei.cmu.edu](http://www.sei.cmu.edu)) institute provides an open list of organizations and organizational units certified in each CMMI-DEV level.
  - FIPA ([www.tivia.fi/in-english](http://www.tivia.fi/in-english)) provides information regarding Finland IT organizations and its professionals.
  - CAPES ([www.capes.gov.br/](http://www.capes.gov.br/)) provides a tool for accessing information regarding Brazilian research groups.

# Sampling

## Challenge:

### Looking for the Frame Population

- Sources available in the web such as *discussion groups, projects repositories* and *worldwide professional social networks* can be helpful to identify representative populations composed by SE professionals

Such sources can restrict at any moment the access to the content available!

# Sampling

**Challenge:**

## **Looking for the Frame Population**

- How to find the frame population in the source of population?
- Once you have identified a source of population, you need to establish steps/procedures to systematically depicting the survey sampling frame.
  - Such practice is important to assess the samples representativeness, also supporting future re-executions

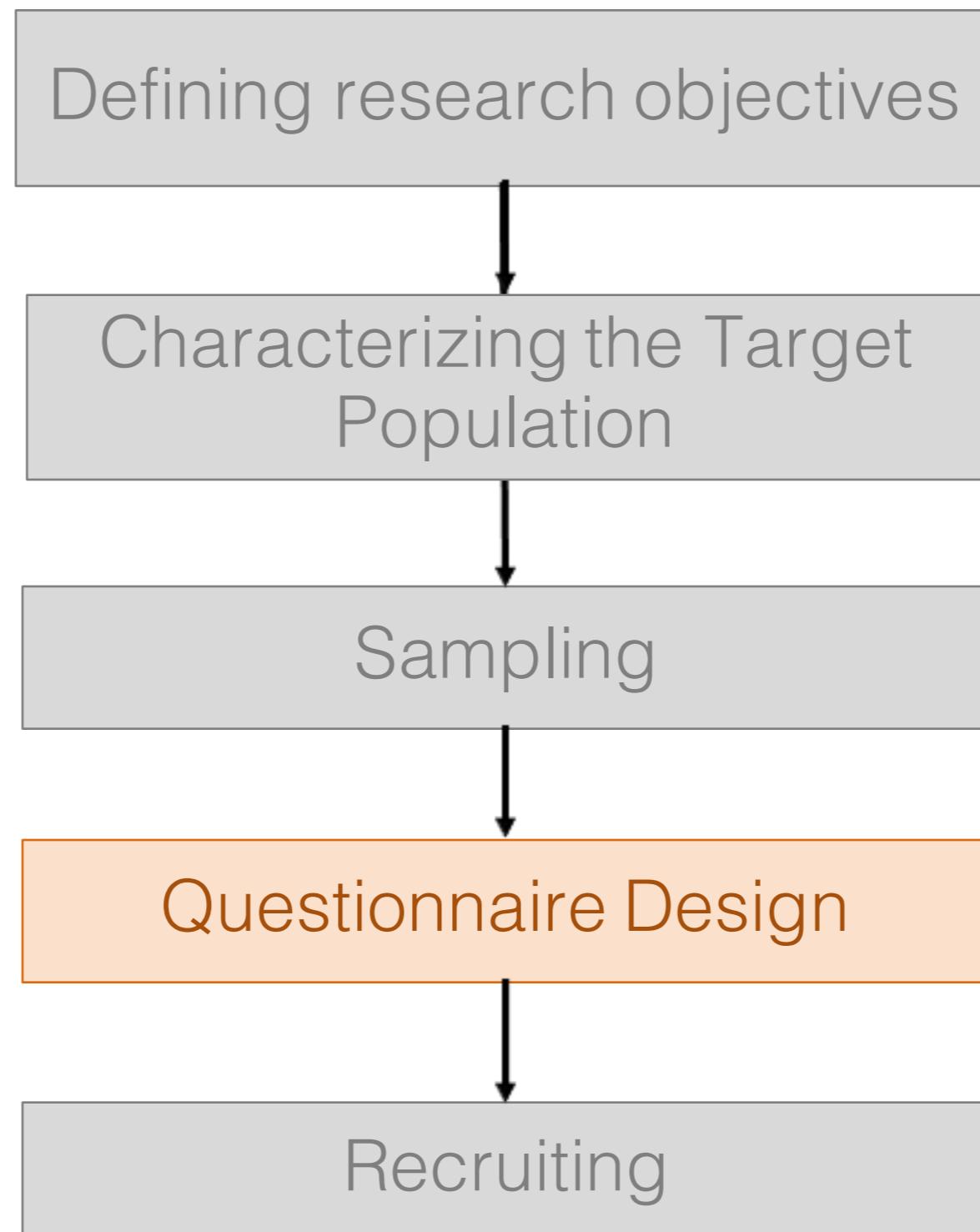
# Sampling

**Challenge:**

## **Establishing the survey sample size**

- Participation rates in voluntary surveys in SE performed over random samples tend to be small (lower than 10%). What are the implications on response rates, but also on representativeness?
- Take preference to probabilistic sampling designs.
  - Independent from the amount of respondents, it will be possible to calculate the results confidence
  - In voluntary surveys with practitioners, establish significantly higher sample sizes, considering the expectation of a very low participation rate

# Best Practices



# Questionnaire Design

**Challenge:**

**To design a clear, simple and consistent survey questionnaire**

**Remember:**

Bad questionnaires can led subjects initially  
willing to participate to give up!

# Questionnaire Design

## Challenge:

**To design a clear, simple and consistent survey questionnaire**

- Use simple and appropriate wording for the survey questions
- Avoid technical terms as much as possible or define them in the questionnaire, according to the survey target population
- Take preference to design short questions **regarding a single concept**
  - Avoid **double barreled** questions
  - Avoid **vague sentences** while writing survey questions

# Questionnaire Design

In your opinion, do you agree or disagree that code refactoring is a need? And what about code smell detection?

- a) I strongly agree
- b) I partially agree
- c) I agree
- d) I disagree

# Questionnaire Design

Code refactoring is an essential practice for improving the understanding of object-oriented code.

- a) Totally agree
- b) Partially agree
- c) Neither agree nor disagree
- d) Partially disagree
- e) Totally disagree

# Questionnaire Design

## Challenge:

To design a clear, simple and consistent survey questionnaire

- Avoid **biased** questions, which can be done by carefully phrasing the questions that do not suggest likely answers or responses
- Avoiding **sensitive** questions
  - In SE context, the sensitive questions can be about respondents income, opinion about organization or management, etc.
- Avoid to ask about far **past events**

# Questionnaire Design

*Do you prefer working in projects following agile methods or those following usual non-agile approaches?*

*Considering the main characteristics of the last 10 software projects you have worked on, please answer the following questions:*

*Asking age, gender, marital status for characterizing requirements engineers*

# Questionnaire Design

## Challenge:

To design a clear, simple and consistent survey questionnaire

- It is important to avoid demanding questions (requiring too much effort from respondents to answer)
- Avoid double negatives

# Questionnaire Design

After reading the attached papers regarding non functional requirements (NFR), please answer the following questions:

1. Which of the following NFR do you disagree are not relevant in the context of real-time systems?

...

# Questionnaire Design

## Challenge:

To design a clear, simple and consistent survey questionnaire

- Be careful on selecting the Response Format!
- Wrong choices of response format may lead you to:
  - Lose precious data
  - Lose the opportunity of applying relevant statistical tests
  - Significantly (and unnecessarily) increase data analysis efforts

# Questionnaire Design

Nominal

- Closed questions
- Statistical analysis based on frequency

Ordinal/  
Likert scale

- Closed questions
- Not necessarily equally distributed intervals
- Significantly restricts statistical analysis

Interval  
Scale

- Closed questions
- Intervals are considered equally distributed
- Statistical analysis is less restrictive than Interval Scale

Free-text

- Open questions
- Allow coding
- Content analysis
- High effort on data analysis

Numeric  
values

- Open questions
- Allow a wide range of statistical analysis

# Questionnaire Design

Do you have experience in Java programming?

( ) Yes

( ) No

How much experience do you have in Java programming?

5 years

How much experience do you have in Java programming?

- a) Very High experience
- ~~b)~~ High Experience
- c) Few Experience
- d) Very Few experience

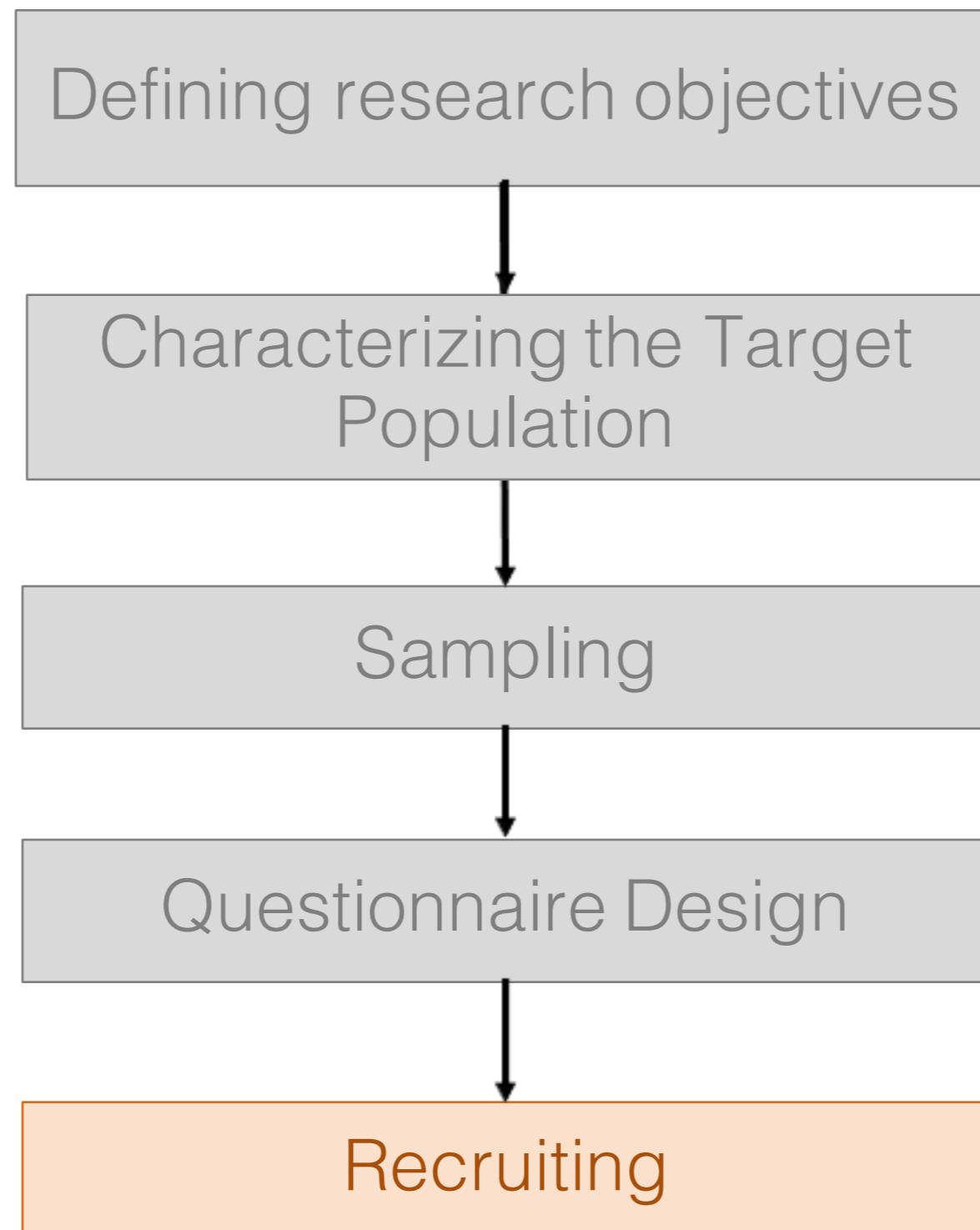
How much experience do you have in Java programming?

*I have been working with Java programming at companies since 2011. Before, I got my first Java certification in 2009, when I started working in personal projects. But I have difficult with object-oriented parts...*

How much experience do you have in Java Programming?

- a) Less than one year
- ~~b)~~ 1 year to 3 years
- ~~c)~~ 3 years to 5 years
- d) More than 5 years

# Best Practices



# Recruiting

**Challenge:**

## **Controlling recruitment and participation**

- Send individual but standard invitation messages
  - It is expected that great most of the individual messages sent will be read
- Avoid "spreading spree": mailing lists, forum invitation messages, crowdsourcing tools (such as Amazon MechanicalTurk)
  - You will have few or no control on who read the invitation. So, who was effectively recruited?
- Never allow forwarding (which is different from snowballing)!
  - It will violate the sample
- Send a questionnaire's individual token to each subject

# Recruiting

**Challenge:**

## **Stimulating participation**

- Reminders should be used with care.
  - Avoid reminding who already had participated
  - Avoid reminding more than once
- The invitation message should clearly characterize the involved researchers, the research context and present the recruitment parameters
- Include in the invitation message a compliment and an observation regarding the relevance of subject participation

# Recruiting

**Challenge:**

**Stimulating participation**

- Establish a finite and not long period to answer the survey
  - One-two weeks
- Offer rewards (raffles, donations, payments, sharing results)
  - Take into account the local policies

# Best Practices



# Piloting the Survey

## Challenge:

**You have only one shot! Once you started the survey, there is usually no way back**

- Pilot the **population and sampling activities**
  - ✓ Use a (smaller) sample of the sampling frame, reproducing all planned steps
  - ✓ Will allow you to check the adequacy of the frame population to your survey.
- Pilot the **questionnaire**
  - ✓ Is it clear, unambiguous, did you maybe miss some questions?
- Pilot the **recruitment**
  - ✓ Does it work effectively?
- Pilot the **data analysis**
  - ✓ Do you have planned for the proper data analysis techniques? What is the necessary data quantity and quality?