The slides are for educational purpose only.

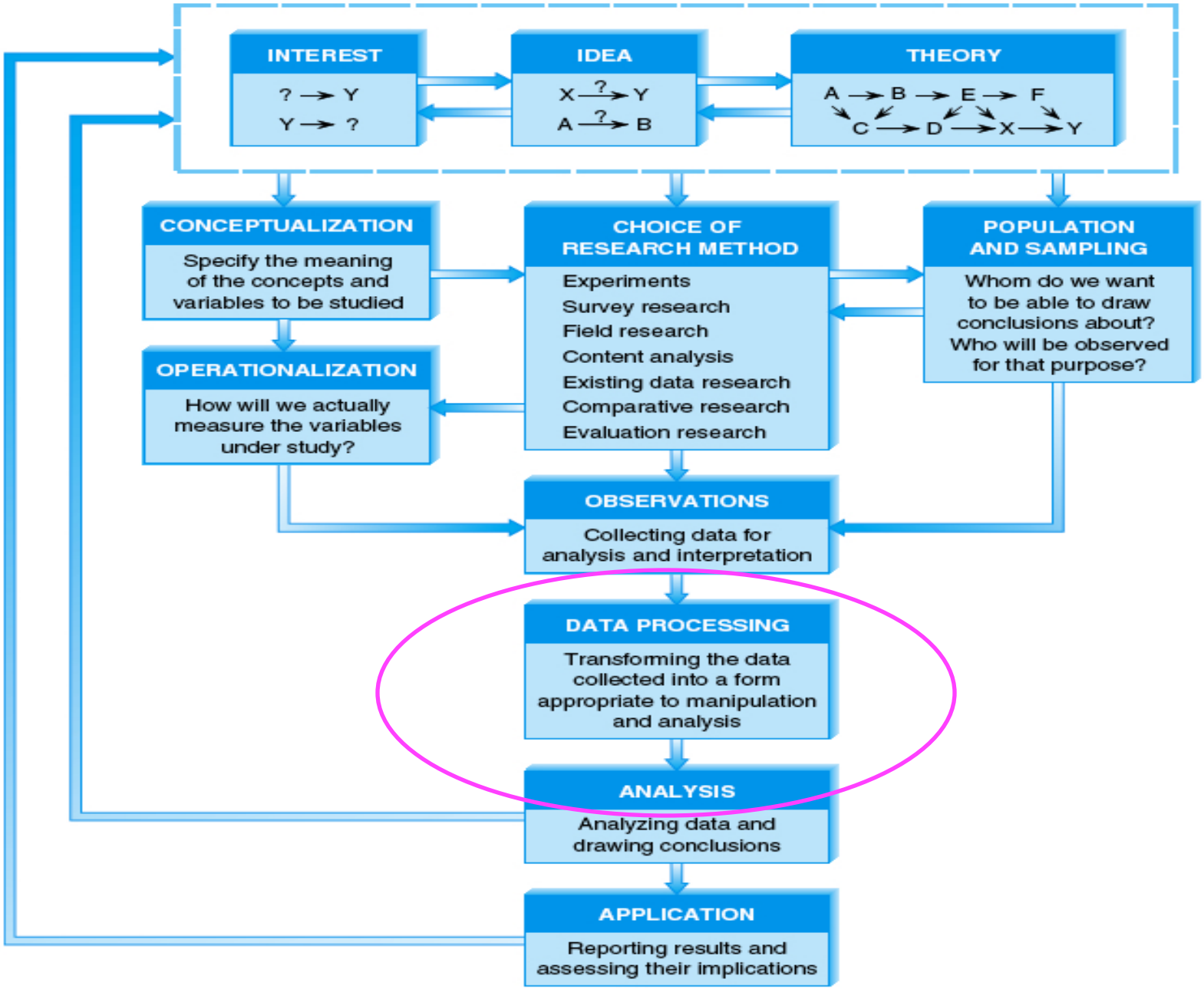
COM5507 Social Media Data Acquisition and Processing
Week 9. Data processing: Overview & issues and cases in data cleaning

Lectured by: Dr. Xinzhi ZHANG
Research Assistant Professor, Department of Journalism
Hong Kong Baptist University
7 Nov 2018 @ CityU M3090

Agenda

- An overview of data processing
- Five steps of data processing and data analysis
- Data processing and data exploration

- Data cleaning (normalization) in Python
- Numeric data processing in Python
- Text data processing in Python


INTEREST
? → Y
Y → ?
IDEA
X →? Y
A →? B
THEORY
A → B → E → F
C → D → X → Y
CONCEPTUALIZATION
Specify the meaning of the concepts and variables to be studied
CHOICE OF RESEARCH METHOD
Experiments
Survey research
Field research
Content analysis
Existing data research
Comparative research
Evaluation research
POPULATION AND SAMPLING
Whom do we want to be able to draw conclusions about? Who will be observed for that purpose?
OPERATIONALIZATION
How will we actually measure the variables under study?
OBSERVATIONS
Collecting data for analysis and interpretation
DATA PROCESSING
Transforming the data collected into a form appropriate to manipulation and analysis
ANALYSIS
Analyzing data and drawing conclusions
APPLICATION
Reporting results and assessing their implications

Data Processing

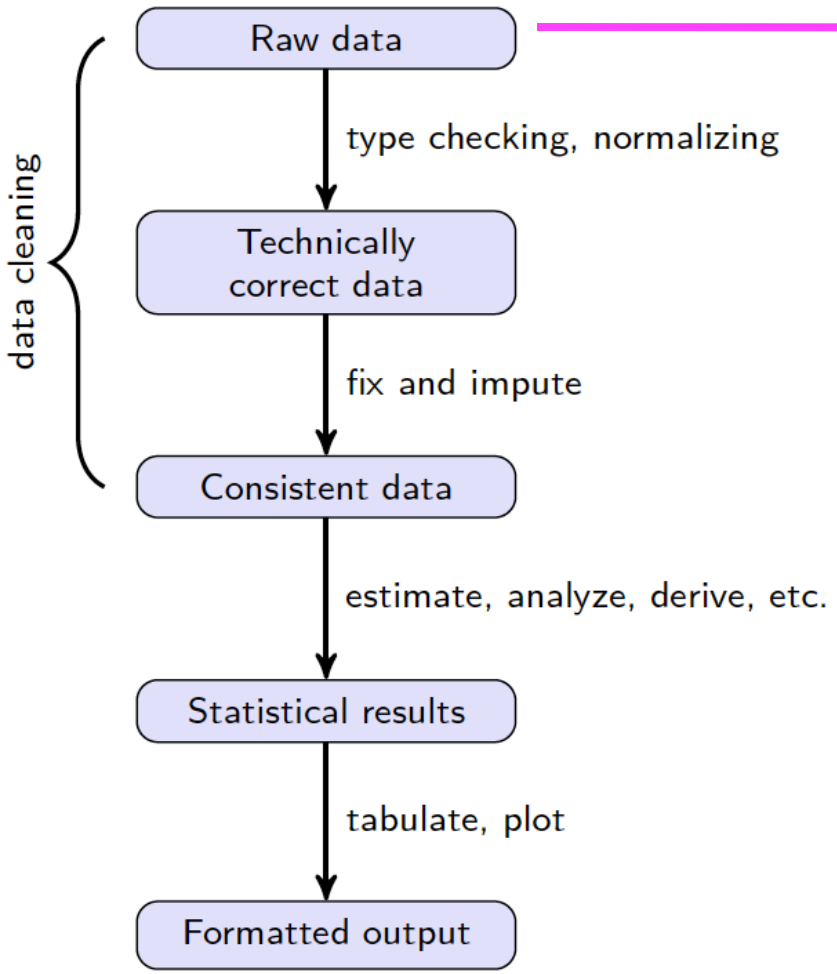
- The purpose of data processing is to transfer collected data (data from web scraping, downloading, or fieldworks) into machine-readable form and ready for data analysis.
- After that, one can use mathematics and statistics to analyze the data with statistical packages.

Data Processing

- Data processing is also called “data preparation:” cleaning the data, normalizing it, and putting it a form that it can be useful for data analysis work.
- It is also called pre-processing (预处理) or coding (coding here in Chinese should be translated as 编码, not 编程, whereas the latter is expressed as programming).

Data Processing

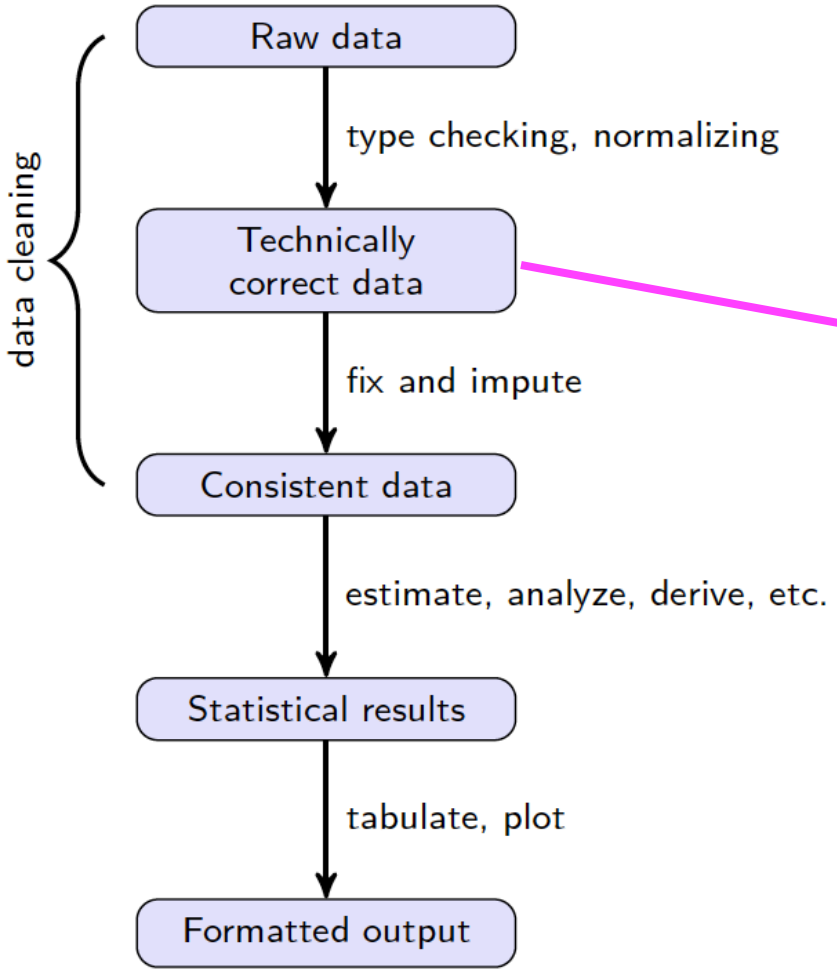
- Why data processing is important?
- https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0



Raw data files may lack headers, contain wrong data types (e.g. numbers stored as strings), wrong category labels, unknown or unexpected character encoding and so on.

Figure 1: Statistical analysis value chain

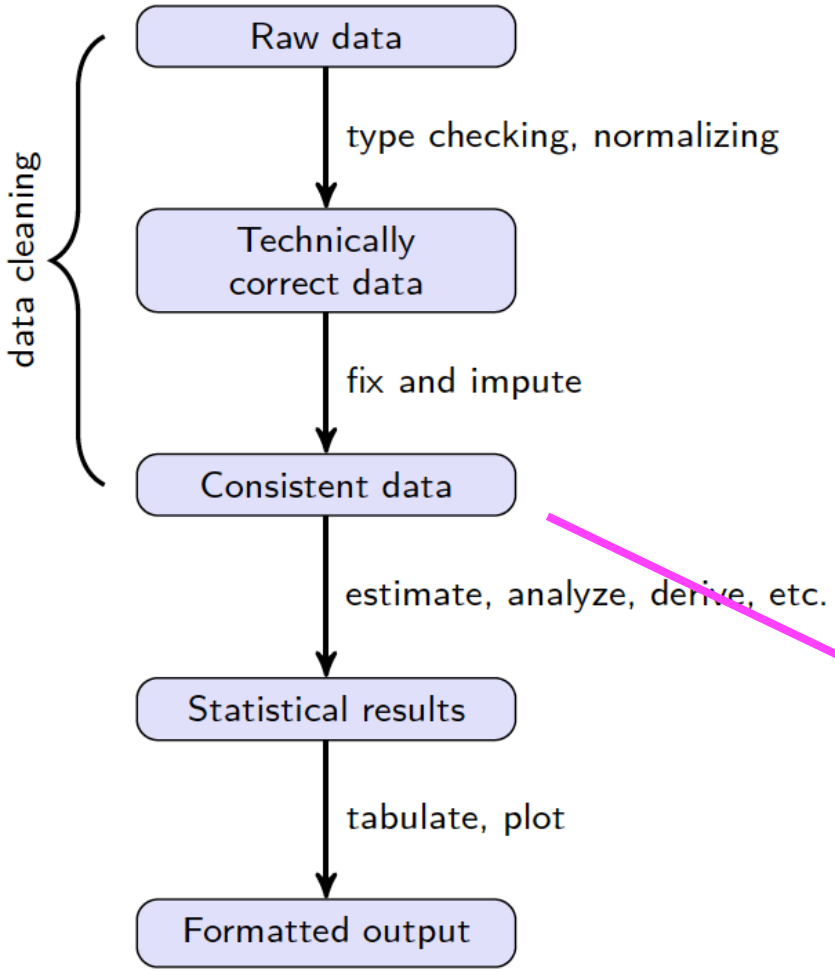
Source of figure 1: Edwin de Jonge and Mark van der Loo



Technically correct data is that, in this state data can be read into a data frame, with correct names, types, and labels, without further trouble.

Figure 1: Statistical analysis value chain

Source of figure 1: Edwin de Jonge and Mark van der Loo



Consistent data is the stage where data is ready for statistical inference. It is the data that most statistical theories (or data analytical methods) use as a starting point.

Figure 1: Statistical analysis value chain

Source of figure 1: Edwin de Jonge and Mark van der Loo

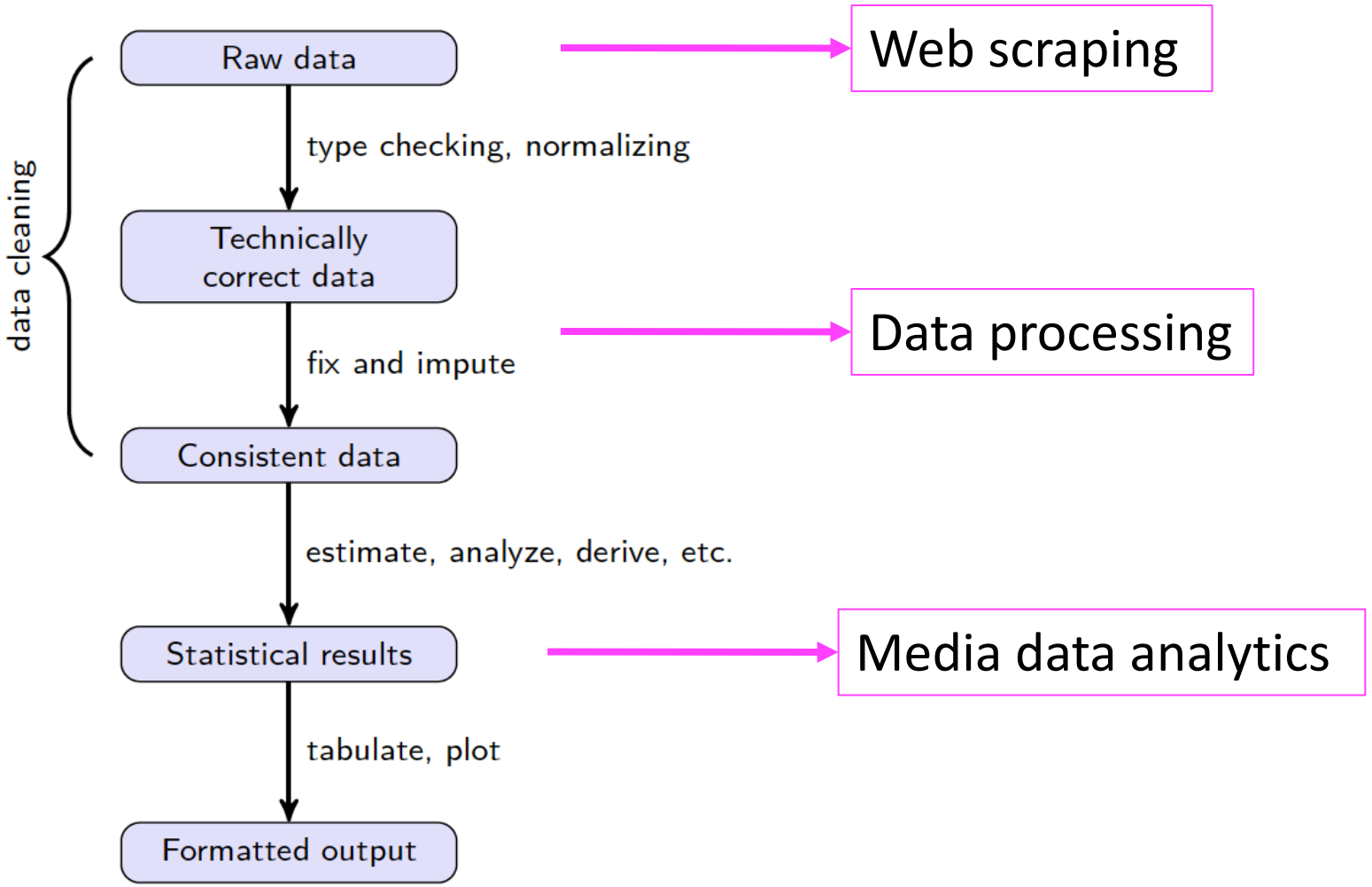


Figure 1: Statistical analysis value chain

Source of figure 1: Edwin de Jonge and Mark van der Loo

Data Processing

- Our goal is to produce **consistent data**.

Machine-readable Data File

- In the field of media and communication, we work on "Spreadsheet" or "Data Frame" (Python Pandas or R)
- Case: a collection of values that belong to a unique subject (unit) in the data file.
  - Example: a person, a news article, a country…
- Variable: a logical grouping of attributes, which describe characteristics or qualities of an object.
  - Example: Age, race, weight, name, scores on a test, and time measured….
- Value: represents the observed attribute of a specific variable of a case
  - Example: 25 years old, Asian, 120 pounds, A..
  - Scale: The possible values the variable can assume form the scale for measuring the variable.

Key issues in data processing

- At “dataset level”
  - Compiling a codebook
  - Examine the shape (dimension)
  - Keeping or dropping variables
  - Merge
  - Join
  - Concatenate
- At “case” level
  - Indexing (“numbering”)
- At “variable” level
  - Data type (string? integer? date?)
  - Renaming, creating, recoding
  - Missing values
  - Normalization (“USA” vs “US” vs “U.S.A.”…)

Coding: Codebook

- In the data processing, one needs to create a codebook first.
- A codebook is a document that describes the locations of variables and lists the assignments of codes to the attributes composing those variables.
- A codebook is the primary guide used in the coding process.
- A codebook is the guide for locating variables and interpreting codes in the data file during analysis.

Codebook

- A codebook at least includes the followings:
  - variable name
  - variable label: the description of the variable, usually the question on the questionnaire
  - value definition: you assign a number to each value of the variable: exclusive and exhaustive
  - Define missing values

Coding

- When we define values, we assign numbers to each possible value.
- Each value and each assigned number has a correspondence.
- These numbers are just the “names” for peculiar answers. They don’t have numerical meanings.
  - Assign peculiar numbers
  - Define those values

18 years old ➔ 18 19 years old ➔ 19 20 years old ➔ 20 Refused to answer ➔ ???	Male ➔ 1 Female ➔ 2 Blank ➔ ??	Disagree ➔ 1 Neither disagree nor agree ➔ 2 Agree ➔ 3 Don’t know ➔ ???

Coding

- Missing values should be defined.
- Examples:
  - Don’t know ➔ -100
  - Refusal ➔ -101 (-100)
  - Blank ➔ -102 (-100)
- Be consistent!

Dataset-level processing

- Merge
- Join
- Concatenate
- https://pandas.pydata.org/pandas-docs/stable/merging.html

Case and variable processing

- A demonstration of Python Pandas data cleaning for numerical and string data