Diving in: a replicable analysis

Konrad Hinsen

CBM, CNRS Orléans and Synchrotron SOLEIL

konrad.hinsen@cnrs.fr

August 30, 2018

Outline

M3-S1: What is a replicable analysis?

M3-S2: Case study: incidence of influenza-like illness

M3-S3A: Data import (Jupyter)

M3-S3B: Data import (RStudio)

M3-S3C: Data import (OrgMode)

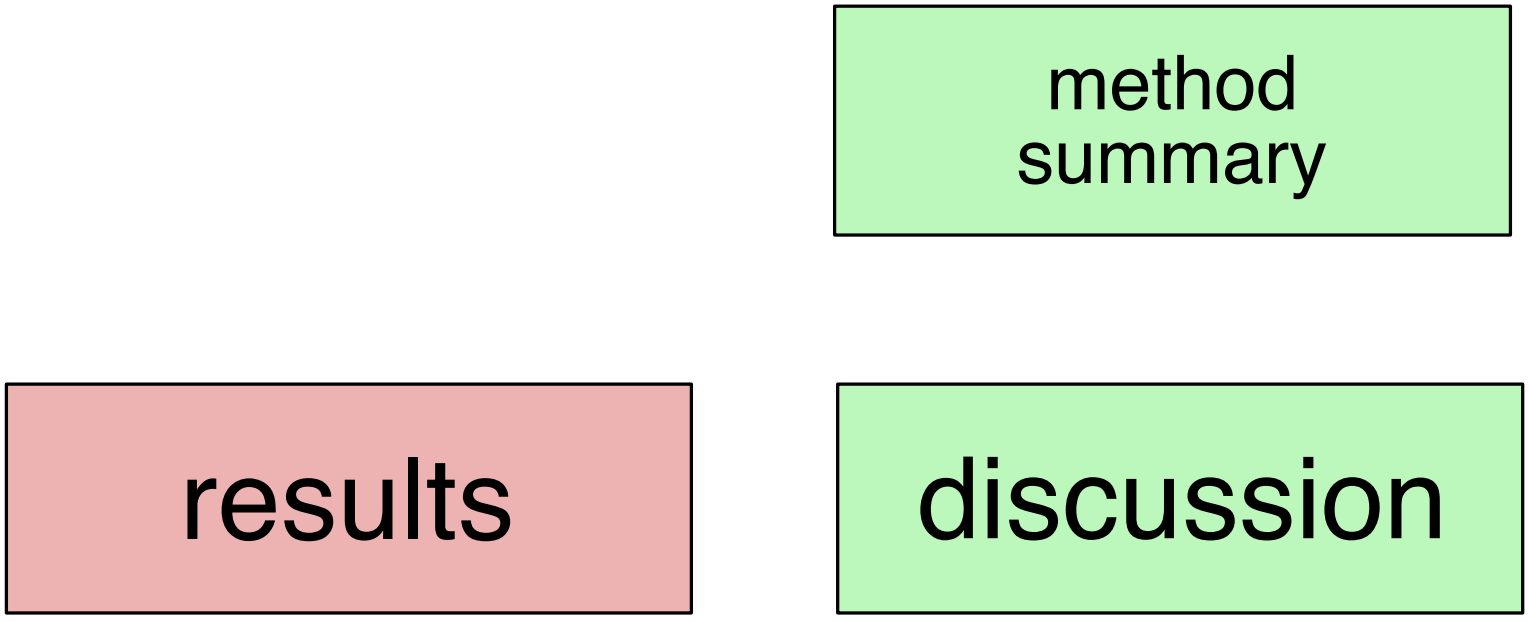
M3-S4A/B/C: Verification and inspection

M3-S5A/B/C: Obtaining answers to a few questions

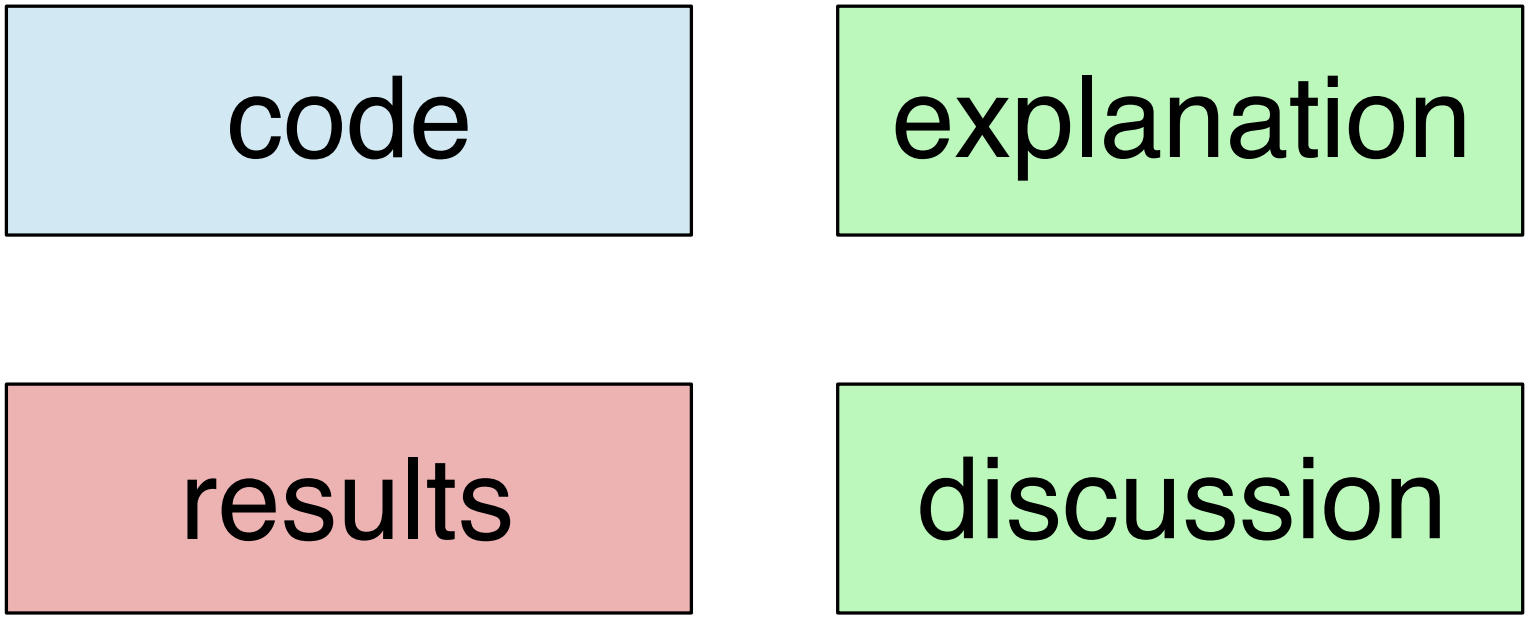
3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Traditional data analysis



Replicable data analysis



Why do it replicably?

- Easy to re-do if the data change
- Easy to modify
- Easy to inspect and verify

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Take-home message

- No manual editing of data
- Everything is done in code!

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Technical choices

- Jupyter notebook
- Python 3 language
- Libraries:
  - pandas
  - matplotlib
  - isoweek

Take-home message

- The data are read directly from the source
- Missing data must be handled

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Technical choices

- RStudio development environment
- R language
- Library: parsedate

Take-home message

- The data are read directly from the source
- Missing data must be handled

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Technical choices

- Emacs editor + Org mode
- Languages:
  - Python 3 for pre-processing
  - R for analysis

Take-home message

- The data are read directly from the source
- Missing data must be handled

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Take-home message

- Preprocessing
  - Adapt the data to the software's conventions
  - Facilitates the analysis
- Verify as much as possible
  - Visual inspection
  - Validation code

3. Diving in: a replicable analysis

1. What is a replicable analysis?
2. Case study: incidence of influenza-like illness
3. Data import
   - Jupyter
   - RStudio
   - OrgMode
4. Verification and inspection
   - Jupyter
   - RStudio
   - OrgMode
5. Obtaining answers to a few questions
   - Jupyter
   - RStudio
   - OrgMode

Questions

1. Which years have seen the strongest epidemics?
2. What is the frequency of weak, average, and strong epidemics?

Take-home message

- A replicable analysis must contain all data processing steps in an executable form.
- It is important to explain all choices that have an impact on the results.
- This requires making many technical details explicit, because that is where most mistakes happen!