Reproducible Research(1)

문건웅

2018-9-4

강사소개

문건웅

- 가톨릭대학교 의과대학 교수
- 성빈센트병원 순환기내과 재직
- R packages (CRAN)
    - mycor, moonBook, ztable, ggiraphExta, dplyrAssist, editData
    - ggplotAssist, webr, rrtable
- Books
    - 의학논문 작성을 위한 R통계와 그래프(2015, 한나래)
        - 2015년 대한민국 학술원 우수학술도서
    - 웹에서 클릭만으로 하는 R 통계분석(2015, 한나래)
    - Learn ggplot2 Using Shiny App(2017, Springer)
- Web-R.org 운영

내용

- Replication ? Reproducible Research ?
- Reproducible research using RStudio

INTRODUCTION

Again, and Again, and Again ...

REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard. New tools and technologies, massive amounts of data, long-term studies, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts, as are increased pressures on scientists to advance their research. The five Perspectives in this section (and associated News and Careers stories, Readers' Poll, and Editorial) explore some of the issues associated with replicating results across various fields.

Ryan (p. 1229) highlights the excitement and challenges that come with field-based research. In particular, observing processes as they occur in nature allows for discovery but makes replication difficult, because the precise conditions surrounding the observations are unique. Further, although laboratory research allows for the specification of experimental conditions, the conclusions may not apply to the real world. Debate about the merits of lab-based and field-based studies has been a persistent theme over time. Tomasello and Call (p. 1227) further contribute to this debate in their discussion of some obvious barriers to replication in primate cognition and behavior research (small numbers of subjects, expense, and ethics issues) as well as more subtle ones, such as the nontrivial challenge of designing tasks that elicit complex cognitive behaviors.

New technologies continue to produce a deluge of data of different varieties, raising expectations for new knowledge that will translate into meaningful therapeutics and insights into health. Ioannidis and Khoury (p. 1230) outline multiple steps for validating such large-scale data on the path to clinical utility and make suggestions for incentives (and penalties) that could enhance the availability of reliable data and analyses.

Data Replication & Reproducibility

CONTENTS

Perspectives

Replication

- Replication of findings and conducting studies with **independent**
    - Investigators
    - Data
    - Analytic Methods
    - Laboratories
    - Instruments
- Replication is particularly important in studies that can impact broad policy or regulatory decisions

What's Wrong with Replication?

- Some studies cannot be replicated
  - No time, opportunistic : 장기간 연구
  - No money : 대규모 무작위 - 대조군 연구
  - Unique
- 대규모 , 장기간 연구가 아닌 경우
  - replication이 가능할까?

EDITORIAL

nature genetics

Mostly, your results matter to others

High-throughput datasets and analysis protocols are intrinsically difficult to referee. Community standards enforced by journals may be less effective than is widely appreciated. Greater awareness of the needs and value of secondary data users can result in higher-impact papers.

Investigating the compliance of our publications with MIAME standards (minimum information about a microarray; Editorial, *Nat. Genet.* **38**, 1089; 2006), we found that even when authors and referees are aware of community standards and even with editors mandating both data deposition and accession linking as a condition of publication, a proportion of microarray datasets were at that time unavailable or incomplete.

Subsequently, the concept of reporting standards has been extended to proposals asking for minimum information about a proteomics experiment (MIAPE: *Nat. Biotech.* **25**, 887–893; 2007), a molecular interaction (MIMIx: *Nat. Biotech.* **25**, 894–898; 2007), a genome sequence specification (MIGS: *Nat. Biotech.* **26**, 541–547; 2008), *in situ* hybridization or immunocytochemistry (MISFISHIE: *Nat. Biotech.* **26**, 305–312; 2008), a biomedical investigation (MIBBI: *Nat. Biotech.* **26**, 889–896; 2008) and proposed facilities and standards for description and deposition of data generated by genome-wide association studies (dbGAP: *Nat. Genet.* **39**, 1181–1186; 2007 and GAIN: *Nat. Genet.* **39**, 1045–1051; 2007).

On page 149 four teams of analysts treated the findings of a number of microarray papers published in the journal in 2005–2006 as their gold standard and attempted to replicate a sample

issue does indeed explain the limits of the analysts' requirements and critical aims.

Why should we consider the utility of rich datasets to researchers whose aim is reanalysis? Many experiments need to start with reanalysis, for validation or comparison. The journal needs to help our referees to spot-check the results they have been asked to examine. If we can make the papers more accessible to readers, we can make the publication and its associated dataset into a more versatile research tool for the benefit of the whole scientific community. Finally, because the spotlight is on the microarray guidelines as a model for other high-throughput methods, the recommendations of this community can be generalized to other fields.

What of the argument that a research paper is less a tool, more an advertisement published to recruit collaborators? This seems like a good idea, except that collaboration is based on mutual benefit and trust, both of which are engendered when collaborator can verify results for him or herself. In a 2004 editorial, we suggested that it is easier to make your reputation with papers that are useful to other researchers than it is to generate an equivalent number and quality of papers from the same dataset by your own

Repeatability of published microarray gene expression analyses

- Selected articles published in **Nature Genetics** between January 2005 and December 2006 that had used profiling with microarrays
- Of the 56 items retrieved electronically, **20 articles** were considered potentially eligible for the project
- The four teams were from
    - University of Alabama at Birmingham(UAB)
    - Stanford/Dana-Farber(SD)
    - London(L)
    - Ioannina/Trento(IT)
- Each team was comprised of 3-6 scientists who worked together to evaluate each article.

Results

- Result could be reproduced : n=2
- Reproduced with discrepancy : n=6
- Could not be reproduced : n=10
    - No data n=4 (no data n=2, subset n=1, no reporter data n=1)
    - Confusion over matching of data to analysis (n=2)
    - Specialized software required and not available (n=1)
    - Raw data available but could not be processed (n=2)

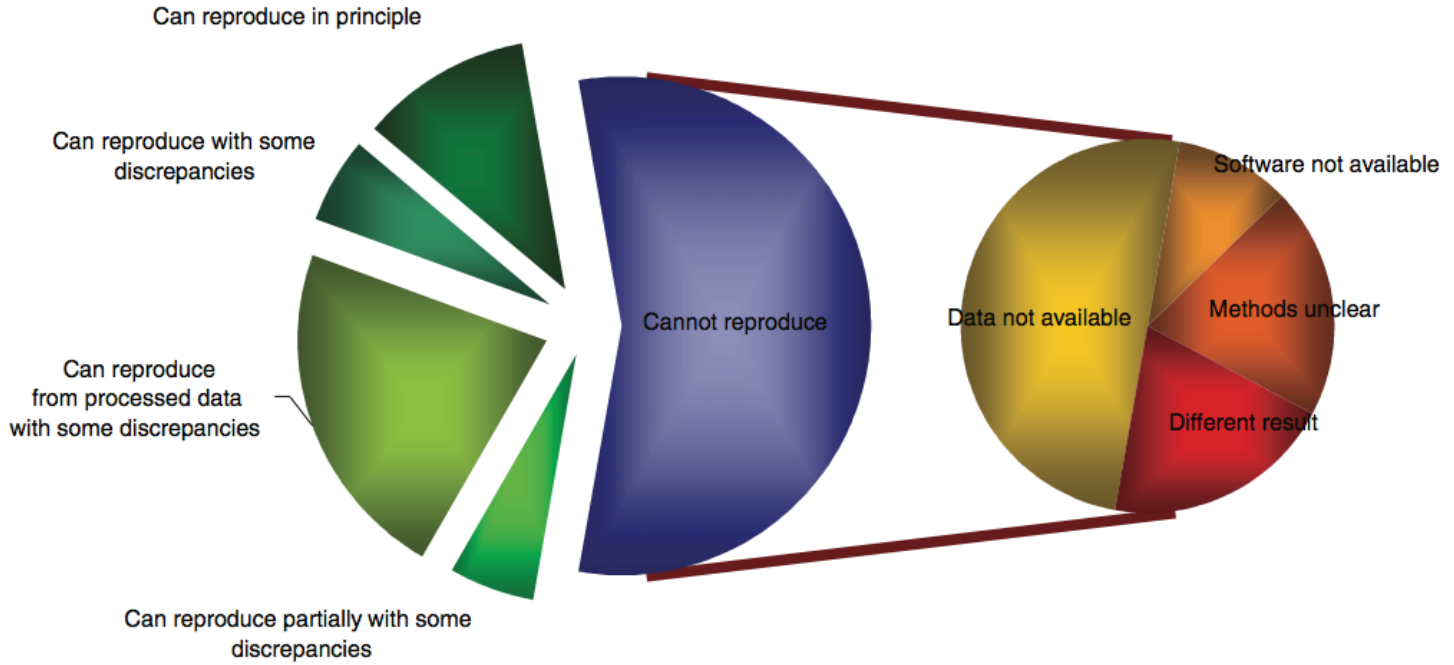
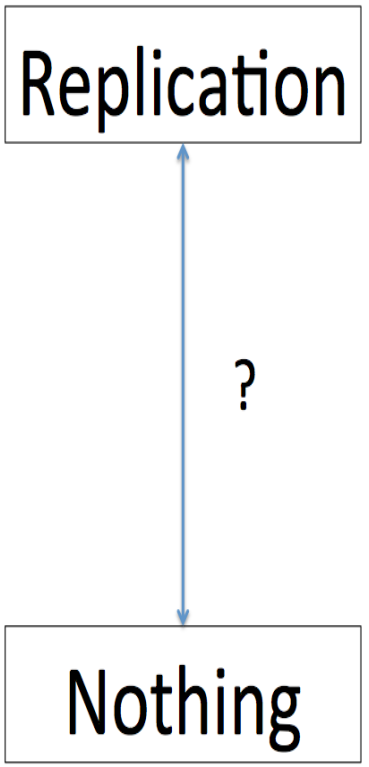
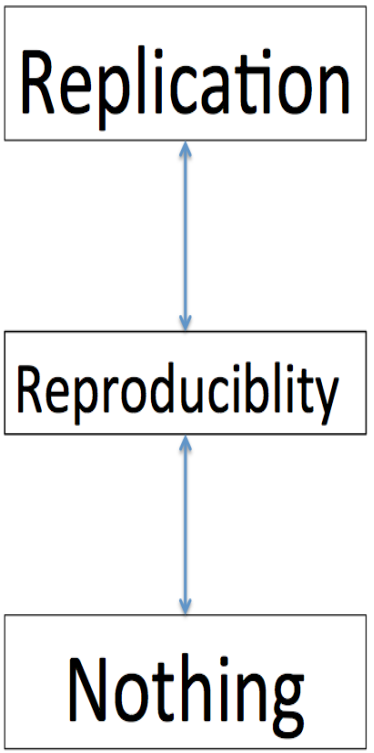


Figure 1 Summary of the efforts to replicate the published analyses.

How Can We Bridge the Gap



How Can We Bridge the Gap



Reproducible Research

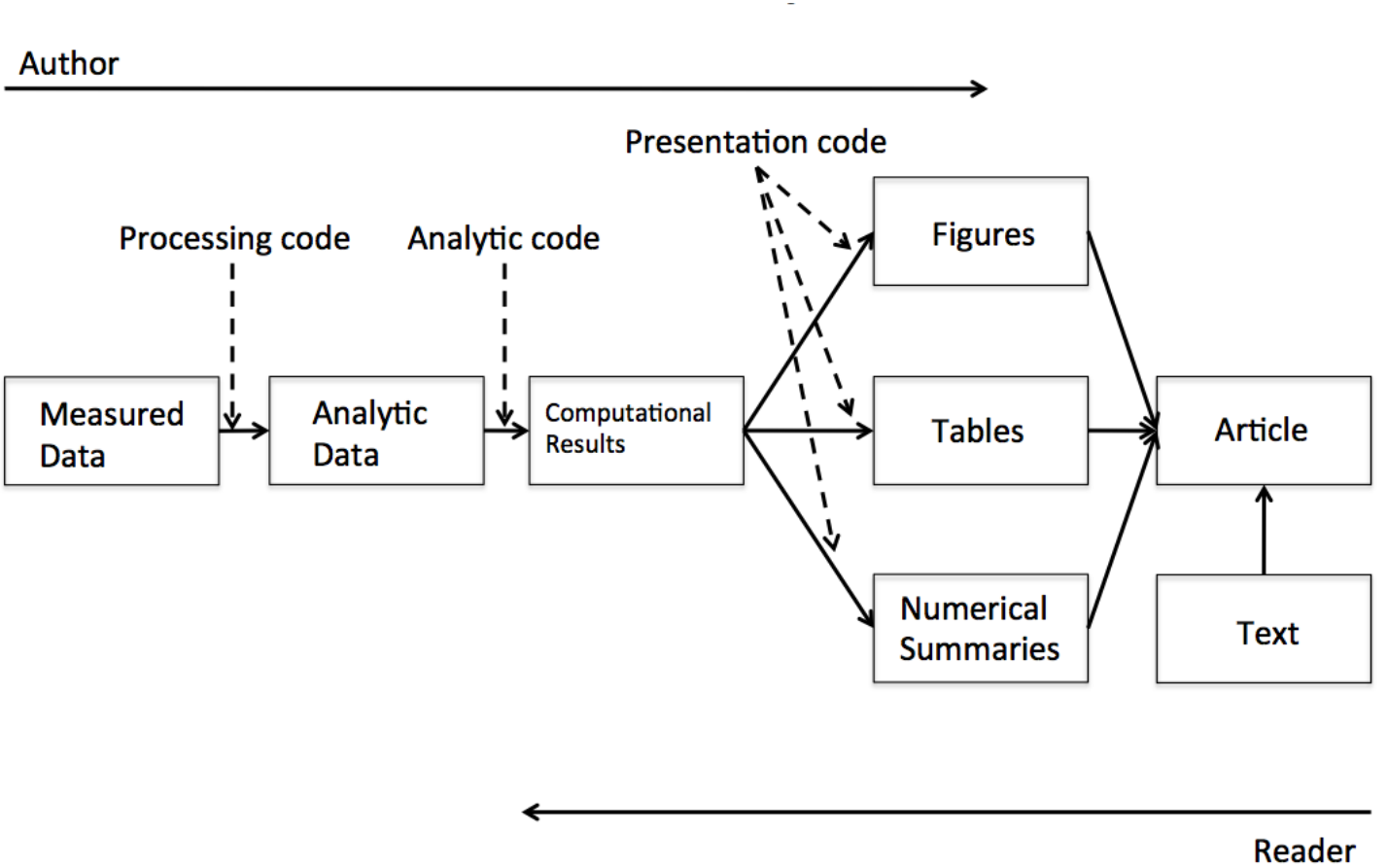
- **data** and the **computer code** used to analyze the data be made **available** to others
- attainable **minimum** reproducibility standard
- fill the gap between full replication of a study and no replication

Why Do We Need Reproducible Research?

- New technologies increasing data collection throughput; data are more complex and extremely high dimensional
- Existing databases can be merged into new “megadatabases”
- Computing power is greatly increased, allowing more sophisticated analyses
- For every field “X” there is a field “Computational X”

Research Pipeline

Research Pipeline



Recent Developments in Reproducible Research

Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Recent Developments in Reproducible Research

Recent Developments in Reproducible Research

REPORT BRIEF MARCH 2012

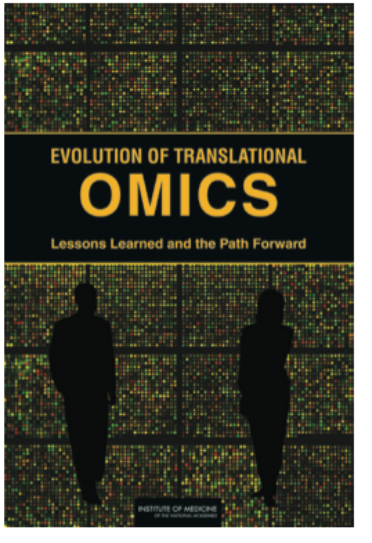
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

Lessons Learned and the Path Forward



Omics ??

- genomics
- transcriptomics
- proteomics
- metabolomics

The IOM Report

In the Discovery/Test Validation stage of omics-based tests:

- **Data/metadata** used to develop test should be made publicly available
- The **computer code** and fully specified computational procedures used for development of the candidate omics-based test should be made sustainably available
- “Ideally, the computer code that is released will **encompass all of the steps of computational analysis**, including all data preprocessing steps, that have been described in this chapter. All aspects of the analysis need to be transparently reported.”

Reproducible research

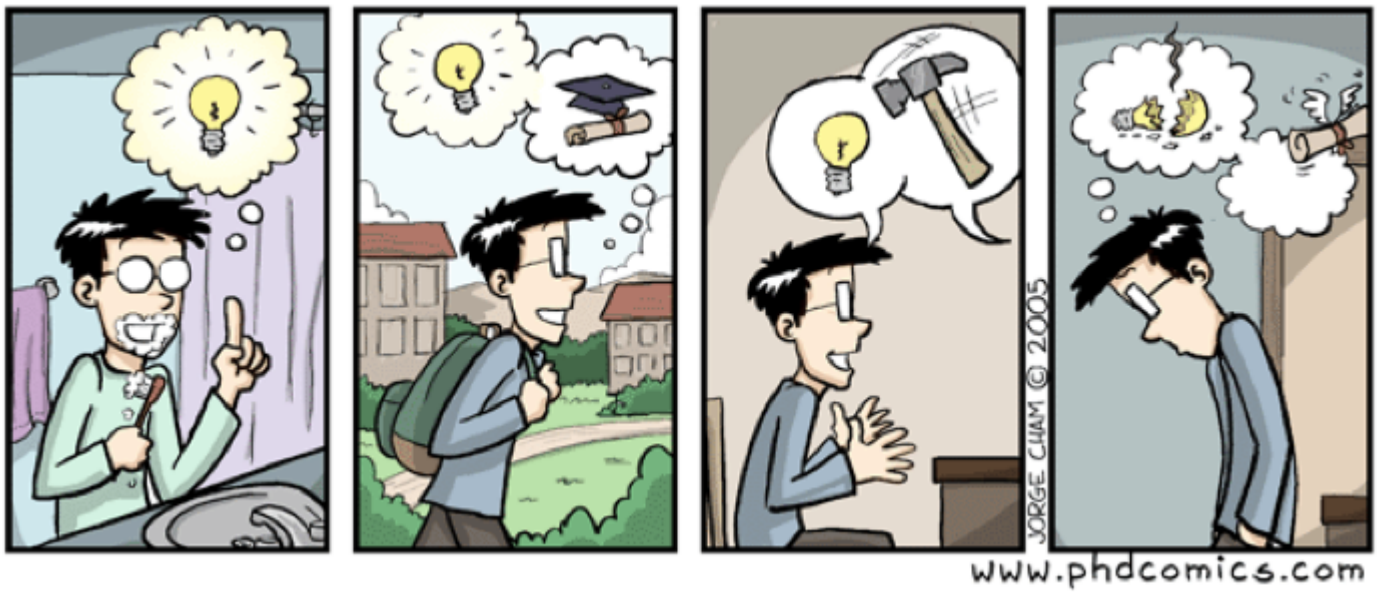
- Literate Programing (문학적 프로그래밍)
- Reproducible Research (재현가능한 연구)

When issues of reproducibility arise

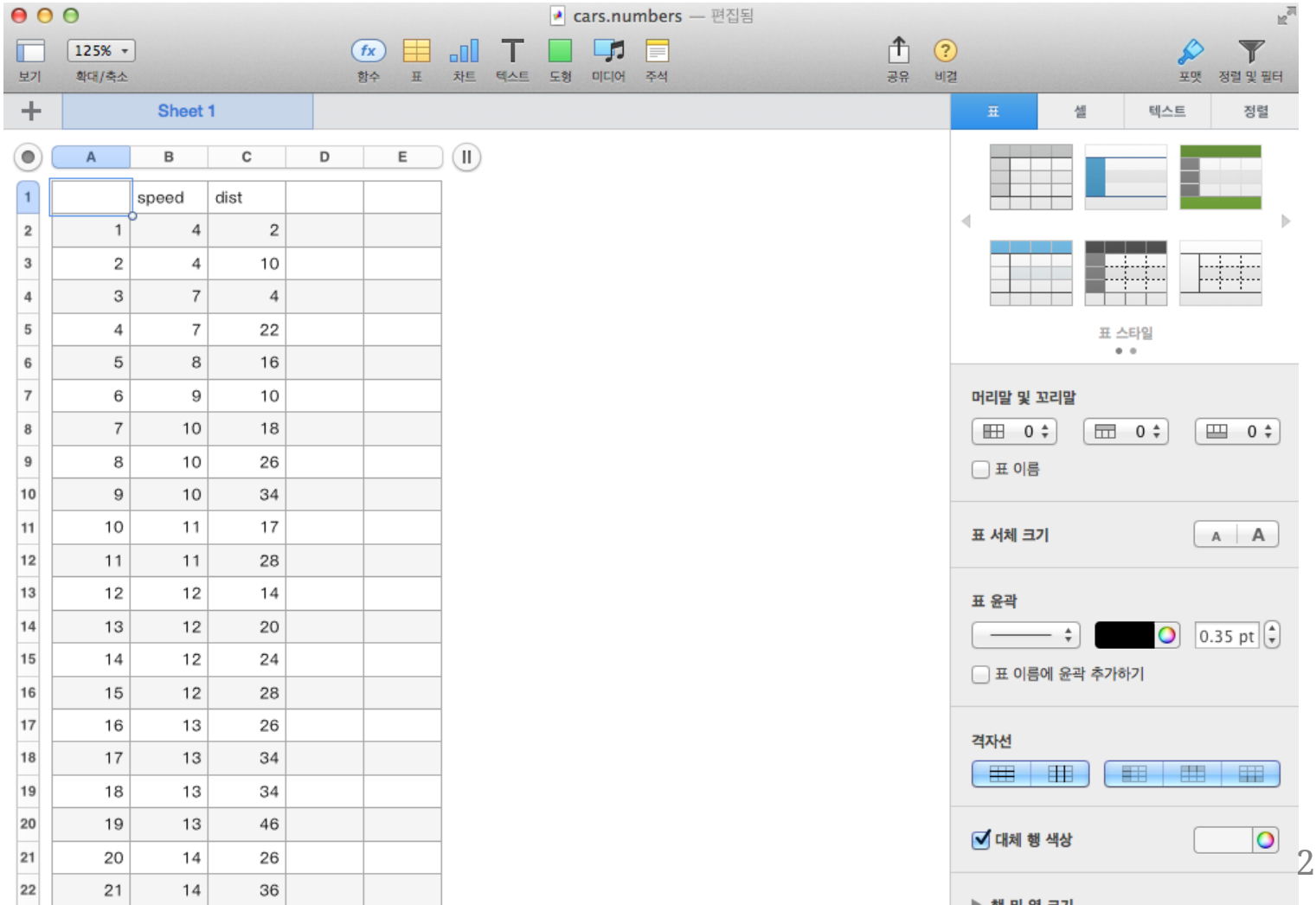
- "Remember that microarray analysis you did **six months ago**?
  We ran a few more arrays.
  Can you add them to the project and repeat the same analysis?"
- "The statistical analyst who looked at the data I generated previously is **no longer available**.
  Can you get someone else to analyze my new data set using the same methods (and thus producing a report I can expect to understand)"
- "Please write/edit the methods sections for the abstract/paper/grant proposal I am submitting based on the **analysis you did several months ago.**"

Typical workflow of many research projects

- First have an idea
- e.g. stopping distance correlate with speed ?

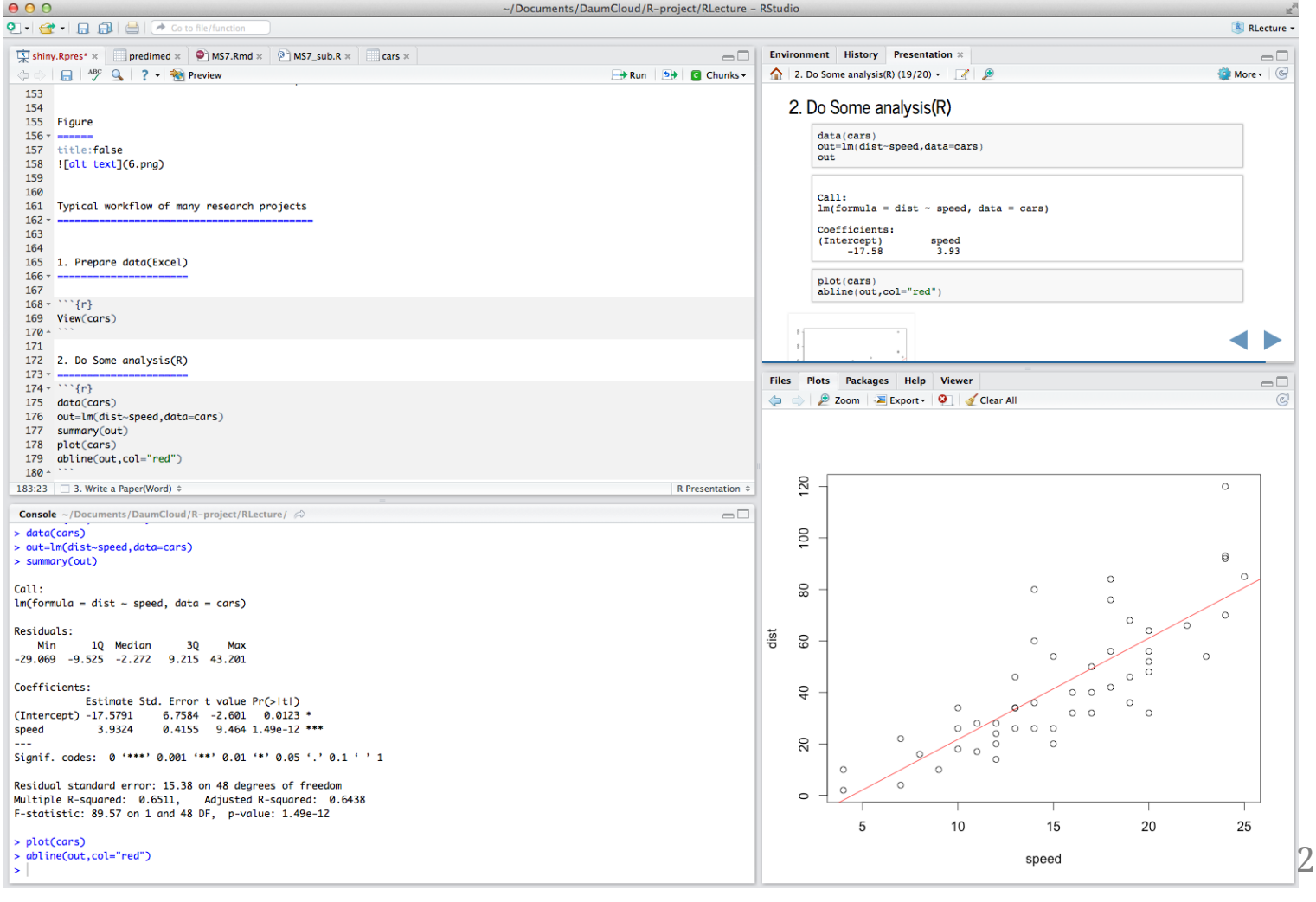


1. Prepare data(Excel/Numbers)

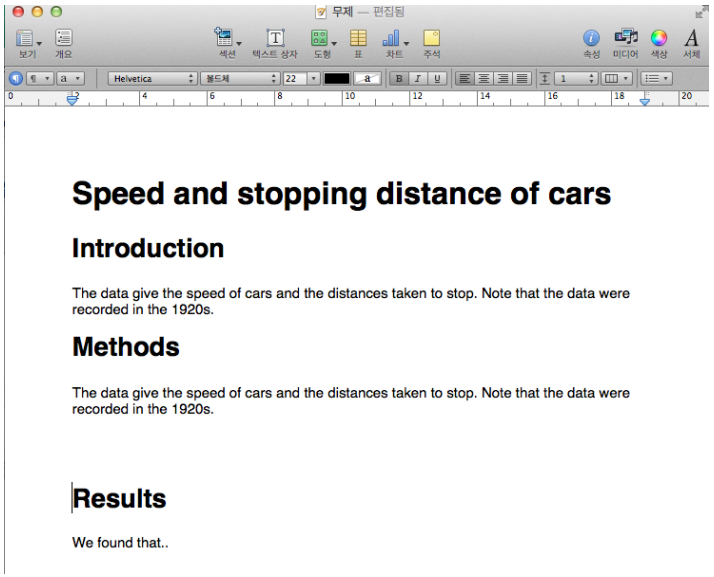


	A	B	C	D	E
1		speed	dist		
2	1	4	2		
3	2	4	10		
4	3	7	4		
5	4	7	22		
6	5	8	16		
7	6	9	10		
8	7	10	18		
9	8	10	26		
10	9	10	34		
11	10	11	17		
12	11	11	28		
13	12	12	14		
14	13	12	20		
15	14	12	24		
16	15	12	28		
17	16	13	26		
18	17	13	34		
19	18	13	34		
20	19	13	46		
21	20	14	26		
22	21	14	36		

2. Do Some analysis(R/SPSS/SAS)



3. Write a report/paper(Word?Pages)

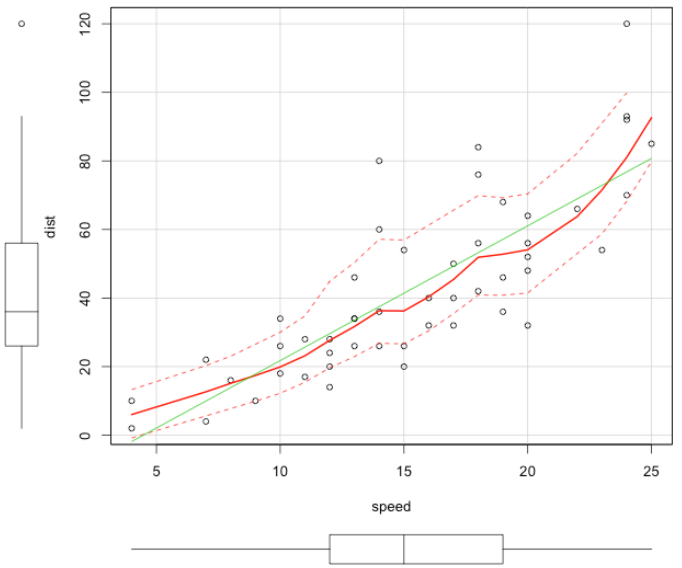


All results(figures, tables) **manually** imported to Word

This workflow is BROKEN

1. Collect and manage data(EXCEL)
2. Analysis (R/SPSS)
3. Writeup(WORD)

Problems brought by the broken workflow



- **What analysis** is behind this figure? Did you account for [ooo] in the analysis?
- **What dataset** was used (e.g. final vs preliminary dataset)?
- Oops, there is **an error** in the data. Can you **repeat** the analysis? And update figures/tables in Word!
- As a coauthor/reader, I'd like to see the **whole research process** (how you arrived to that conclusion), rather than cooked manuscript with inserted tables/figures.

RStudio allows us to fix the disconnect

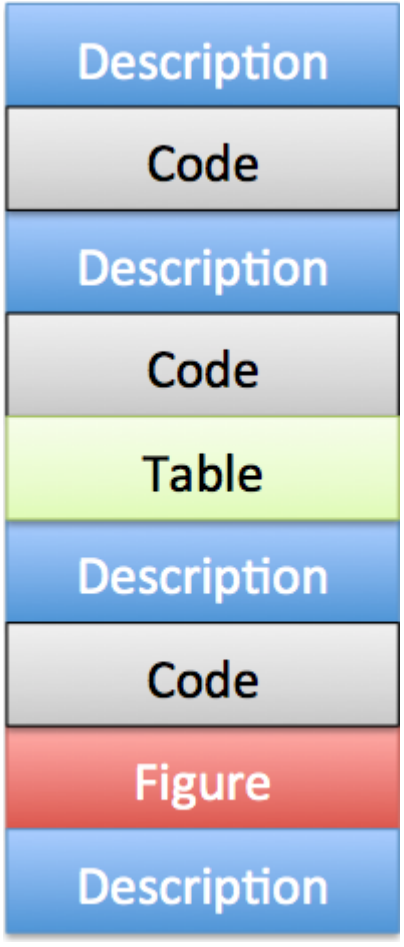
Integrating

- Data management
- Data analysis
- Writing up results

in a single dynamic document

Reproducible research !!

Let's make our project reproducible



- In literate programming, an analytical document is composed of a descriptive narrative “woven” together with software code and computed results.
- Advantages
  - A single document both describes and performs the analysis
  - Enforces reproducibility

If error

- If spotting eror in data, or using different dataset...
- make changes in Rmarkdown and report will update **automatically**

So... Main Advantages

RStudio: Preview HTML

Preview: .../RLecture/sample.html | Log | Save As | Republish | Find

Speed and stopping distance of cars

Introduction

The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.

Methods

```
library(car)
data(cars)
tail(cars)
```

```
##    speed dist
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

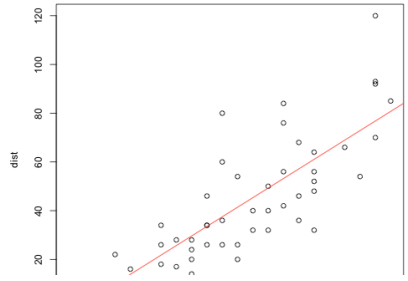
Results

We found that

```
out = lm(dist ~ speed, data = cars)
out
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Coefficients:
## (Intercept)        speed
##      -17.58         3.93
```

```
plot(cars)
abline(out, col = "red")
```



- Data management fully documented (no more manual changes in Excel!)
- Analysis fully documented
- Automated reports
- can publish via Rpubs.com : http://www.rpubs.com/cardiomoon/19541
- can share the project

Summary

- Reproducible research is important as a **minimum standard**, particularly for studies that are difficult to replicate
- Infrastructure is needed for **creating** and **distributing** reproducible documents, beyond what is currently available
- There is a growing number of tools for creating reproducible documents

참고자료

- 의학논문작성을 위한 R통계와 그래프
- R과 knitr를 활용한 데이타 연동형 문서 만들기

Online Resources

- www.rstudio.com
- Web-R.org

온라인교육프로그램

- Coursera : Reproducible Research (Jones Hopkins University)