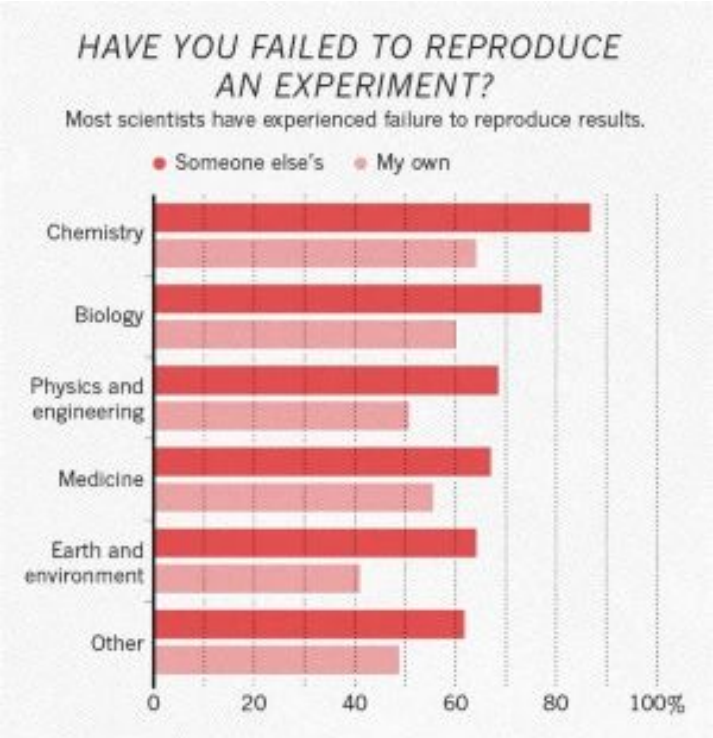
Writing a Reproducible Manuscript in R

Qian Zhao
Postdoctoral Scholar, Biomedical Data Science

Stanford | LIBRARIES

What is reproducible research?

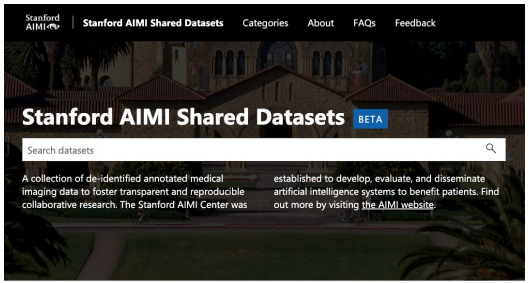


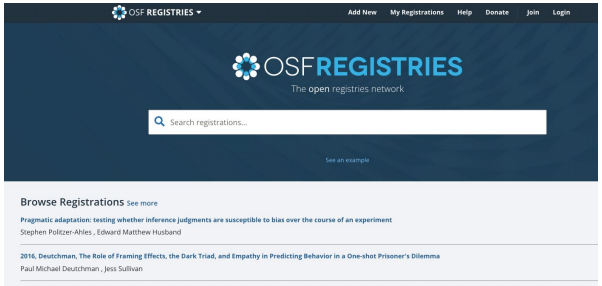
- **Reproducible** means results can be duplicated using the **same materials and procedures**

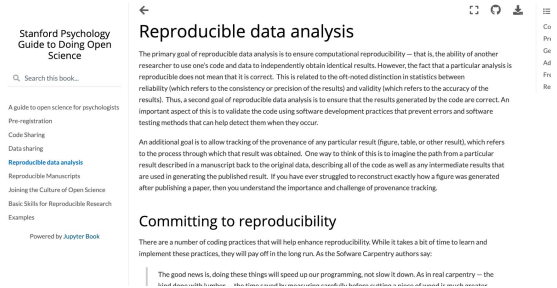
- **Replicable** means results can be confirmed using **new data** collected with **same procedures**

Baker, M. 1,500 scientists lift the lid on reproducibility. (2016) *Nature* 533(7604):452-4.

We can make our research more reproducible by improving practices!







Open Data

Pre-registration

Transparent Data Analysis

Other initiatives: Registered reports, publishing replication studies etc.

Creating a Reproducible Report

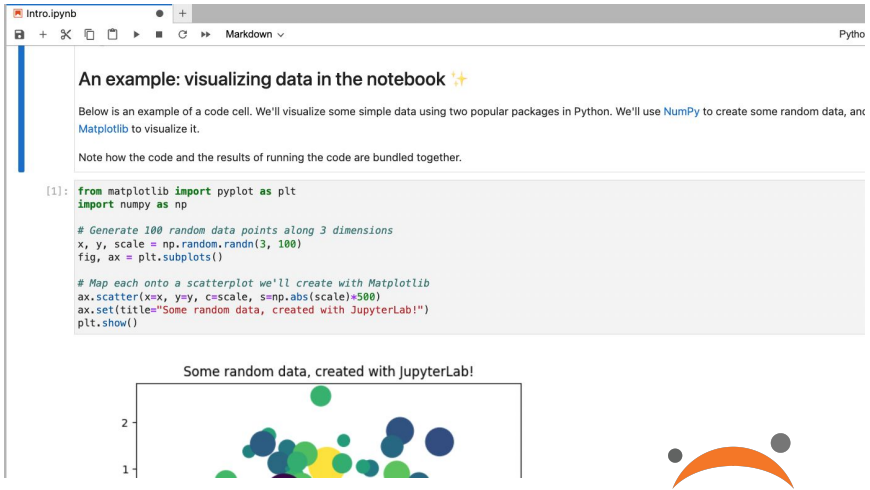
```
\begin{abstract}
Due to its probabilistic nature, Null Hypothesis Significance Testing (NHST) is subject to decisi
\end{abstract}

\keywords{NHST, reproducibility project, nonsignificant, power, underpowered, effect size, Fisher

\section*{Author note}
All research files, data, and analyses scripts are preserved and made available for download at \
\newpage

<<Dependencies, echo=FALSE, print=FALSE>>=
# Running in Sweave or running the code in Rstudio project
# If you want to run in sweave, comment out the line with run <- ''
run <- '../'
# run <- ''

# Set the cran mirror
```



Executable report with **documentation, code, figures** etc. in the same document

Creating a Reproducible Manuscript

- R packages providing templates to format manuscripts into different guidelines
  - **papaya** – American Psychological Association
  - **rticles** – American Chemical Society, American Geophysical Union, American Economic Association, Institute of Mathematical Statistics, IEEE, etc.
- Other platforms
  - Stencila
  - Curvenote

papaja: Prepare APA Journal Articles with R Markdown

CRAN v0.1.1 | repo status Active | Last commit march | R-CMD-check failing | codecov 82% | Bugs 5 open | Questions invalid

papaja is an award-winning R package that facilitates creating computationally reproducible, submission-ready manuscripts which conform to the American Psychological Association (APA) manuscript guidelines (6th Edition). **papaja** provides

- an R Markdown template that can be used with (or without) RStudio to create PDF documents (using the apa6 LaTeX class) or Word documents (using a .docx-reference file).

rticles

Reproducible manuscript preparation with RMarkdown application to JMSACL and other Elsevier Journals

Daniel T. Holmes [a, c], Mahdi Mobini [a, b], Christopher R. McCudden [e, d] [f]

Creating a Reproducible Manuscript in R

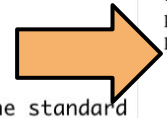
```
---
title: Fisher's Analysis of Iris Data
thanks:
  - ref: T1
    text: Based on the article "The Use of Multiple Measurements in Ta
Problems" by R. A. Fisher (1936)
runtitle: Iris Data
```

```
# Fisher linear discriminant analysis

In a 1936 article, @fisher1936 considered the question: what linear function of
the four measurements
\begin{equation}
X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4
\end{equation}
maximizes the \textit{ratio} of the difference between the means to the standard
deviation within species?

The observed means and their differences are shown in Table \ref{tab:iris_mean}.
We can also compute the sum of squares and products of deviation from specific
means of each species (Table \ref{tab:compute_var}).
```

```
```{r compute_mean, echo = F}
iris_means <- dplyr::tibble(
  Variable = c("Sepal length", "Sepal width",
               "Petal length", "Petal Width"),
  Versicolor = colMeans(subset(iris, Species == "versicolor")[,-5]),
  Setosa = colMeans(subset(iris, Species == "setosa")[,-5]),
```



Submitted to the Annals of Applied Statistics

FISHER'S ANALYSIS OF IRIS DATA*

BY QIAN ZHAO [1]

[1]*Department of Biomedical Data Science, Stanford University, qzhao1@stanford.edu*

I use an analysis of the Iris data set to illustrate how to use the "rticles" package to create a reproducible manuscript.

1. The iris data. The Iris data, collected by Dr. E. Anderson, contains measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I.versicolor*. Figure 1 shows pictures of the two species. The data includes four measurements: sepal length, sepal width, petal length, and petal width. A few rows of the data are shown in Table

(a) *Iris setosa* (b) *I.versicolor*

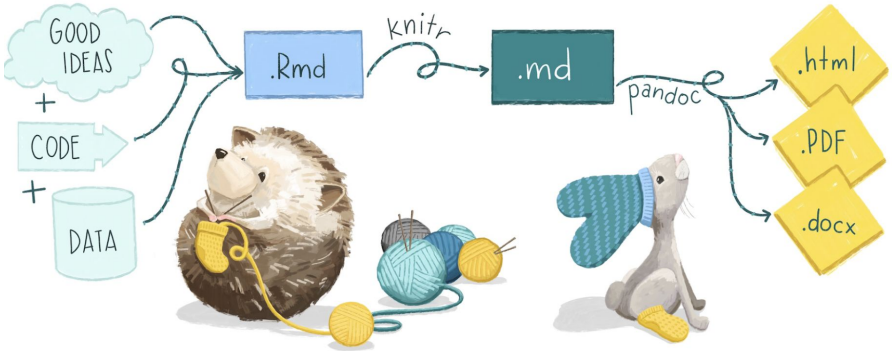
Figure 1: Two iris species

TABLE 1
First few rows in the iris data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
```

Writing a reproducible manuscript has many benefits!

- Efficient, Transparent, Reproducible, Collaborative
- Resources
  - Stanford psychology guide to doing open science
  - Awesome reproducible research



Artwork by Allison Horst