Reproducible research and reproducible analyses: literate programming for everyone

Adam J. Richards

Centre national de la recherche scientifique (CNRS)
(French National Center for Scientific Research)
Station d'Ecologie Expérimentale du CNRS à Moulis

Last updated: November 23, 2012

1. Reproducible research
2. Three aspects
3. IpEdit
4. The future

Why are scientific studies so difficult to reproduce?

- Publication/Experimental bias
- Rewards for 'positive results'
- Programming errors or data manipulation mistakes
- Poorly selected statistical tests
- Multiple testing, multiple looks at the data, multiple statistical analyses
- Lack of easy-to-use tools

Ideas adopted from a a presentation by Jim Berger
Image taken from http://wanderingfrance.com

Evidence for a lack of reproducibility

- Bayer Healthcare reviewed 67 in-house attempts at replicating the findings in published research.
  - Less than 1/4 were viewed as essentially replicated.
  - Over 2/3 had major inconsistencies → project termination.
- *Why Most Published Research Findings Are False*
- A presentation by Jim Berger

Bayer Healthcare study: *Nature Reviews Drug Discovery* 10, 712 (2011) [URL]
J.P. Ioannidis. Why Most Published Research Findings Are False **PLoS Med**. 2005 August; 2(8): e124. [URL]

Why bother making our work reproducible?

Nature Medicine - 12, 1294 - 1300 (2006)

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2] Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3,], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

Dr. Baggerly and Dr. Coombes found errors almost immediately. Some seemed careless – moving a row or a column over by one in a giant spreadsheet – while others seemed inexplicable. The Duke team shrugged them off as "clerical errors."

And the Duke researchers continued to publish papers on their genomic signatures in prestigious journals. Meanwhile, they started three trials using the work to decide which drugs to give patients.

source: http://www.nytimes.com/2011/07/08/health/research/08genes.html

Again why bother?

- Retracted: 07 January 2011
- Streamlines the review of manuscripts and grant proposals
- Like writing good code—it ultimately saves time
- Promotes sharing
- Promotes **good practices**

Multiple testing

The tradition in epidemiology is to ignore multiple testing

Over-reliance on the use of *p*-values

The tradition in psychology is to ignore optional stopping; if the *p*-value is close to significant then get one more data point (with no adjustment).

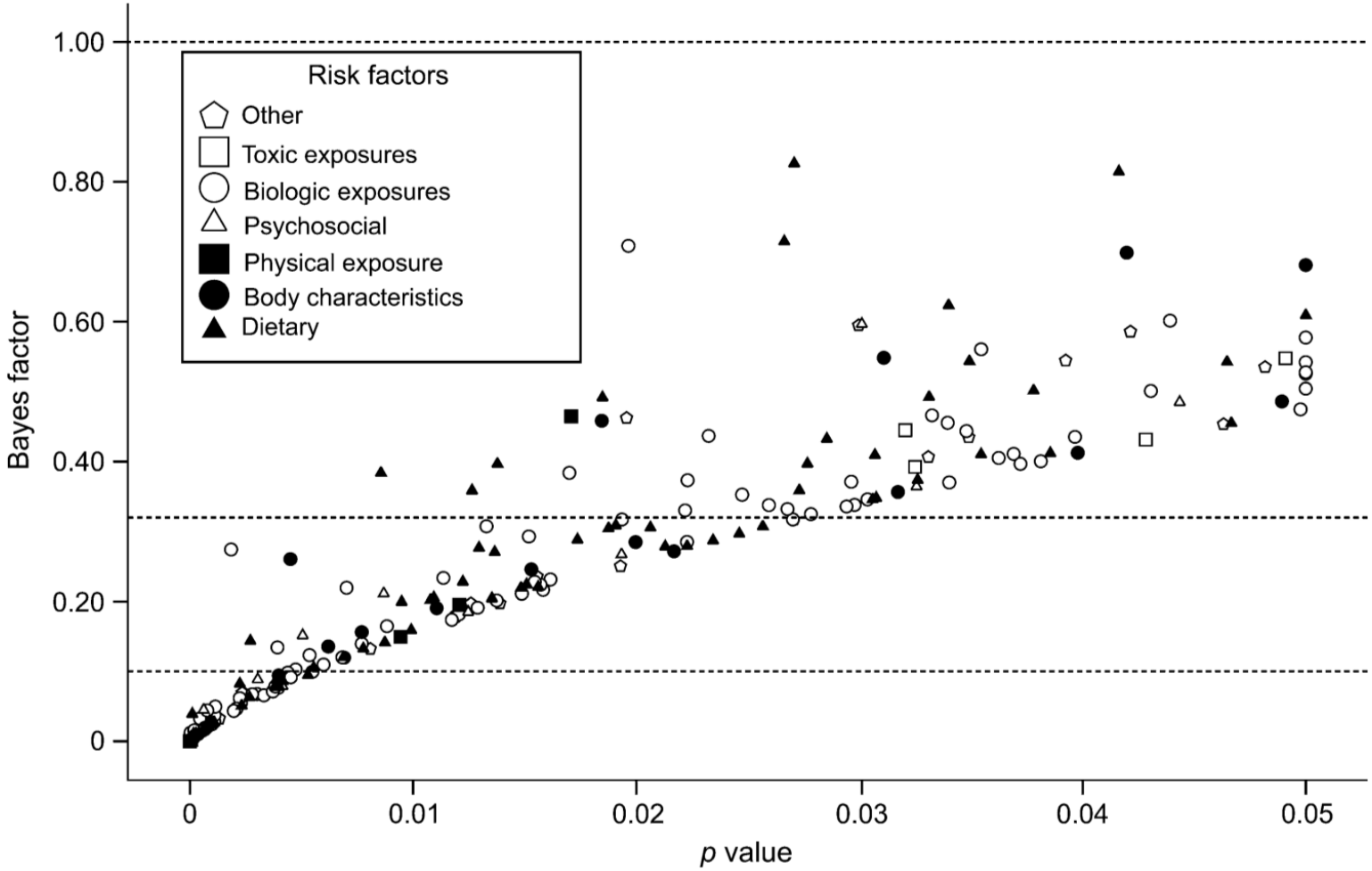


FIGURE 1. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed p value in each study. Shown are calculations assuming θ_A of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

J P A Ioannidis *Am. J. Epidemiol.* (2008) 168 (4): 374-383. [URL]

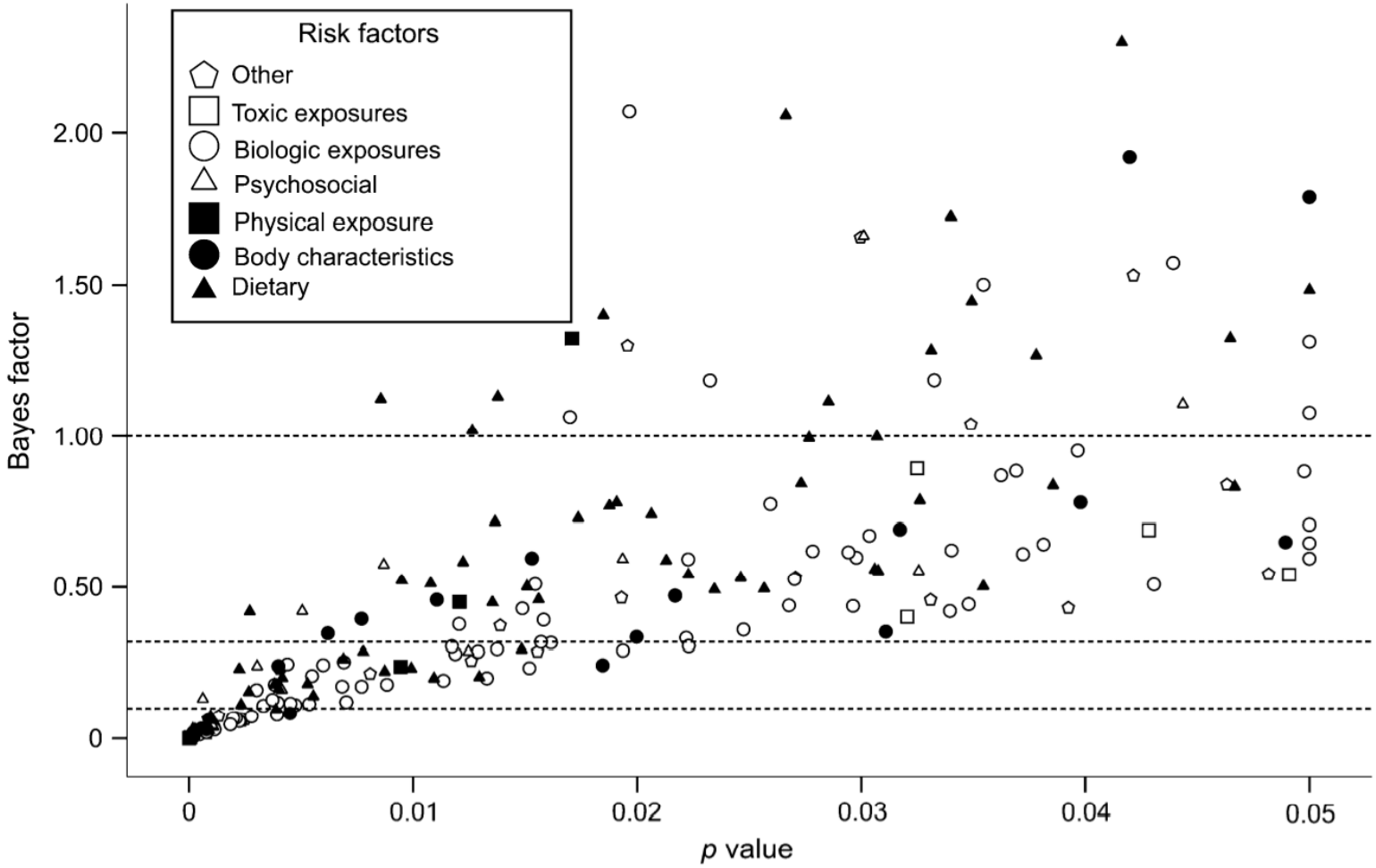


FIGURE 2. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed p value in each study. Shown are calculations assuming θ_A of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

J P A Ioannidis *Am. J. Epidemiol.* (2008) 168 (4): 374-383. [URL]

Reproducible Research

Reproducible research consists two main components:

1. Reproducible experiment
2. Reproducible analysis

Lab/Experiment

Data Sharing	Instruments, samples, buffers	Supplemental Methods
Audit Trail	Version control and backup	Git, Mercurial and SVN
Documentation	All details required to reproduce	Electronic lab notebook

Data Analysis

Data Sharing	raw, standards	public repositories, URL
Audit Trail	Version control and backup	Git, Mercurial and SVN
Documentation	All details required to reproduce	Literate Programming

Data Sharing

1. Must include the raw data and the appropriate meta data
2. Must be free and publicly available
3. Use data standards when available (eg. MIAME or MINSEQE)

Repositories

- Microarray - Gene Expression Omnibus - NCBI
- Transcriptomics data - ArrayExpress - EBI
- Sequencing data - Read Archive - NCBI
- Sequencing data - Trace Archive - NCBI
- Metagenomics - Metagenomics - EBI
- Genome-phenome - European Genome-phenome Archive - EBI

Audit Trail

What kinds of systems are available?

- Good - The cloud (Dropbox, Google Drive)
- Better - Version control systems (SVN, Git and Mercurial*)
- Best - Version control systems on the cloud (GitHub, Bitbucket)

Good practices

1. Use a system that documents both data and process
2. Use the machine readable CSV format
3. Never embed data manipulation and statistical tests
4. Use R, Python or another freely available software to read and process raw data—ideally to produce reports complete with code, results and prose.

* For Windows users look at TortoiseHg.

Create a wiki and track changes to it through a repository

1. Create a repository called 'labwiki' on bitbucket or on another hosted site
2. Download and setup wiki

```
$ hg clone https://<USERNAME>/bitbucket.org/labwiki
$ cd labwiki
$ wget pmwiki.org/pub/pmwikipmwiki-latest.tgz
$ tar zxvf pmwiki-latest.tgz
$ cd pmwiki-*
```

3. Create a config.php file (use the one from the site)
4. add it to your repository

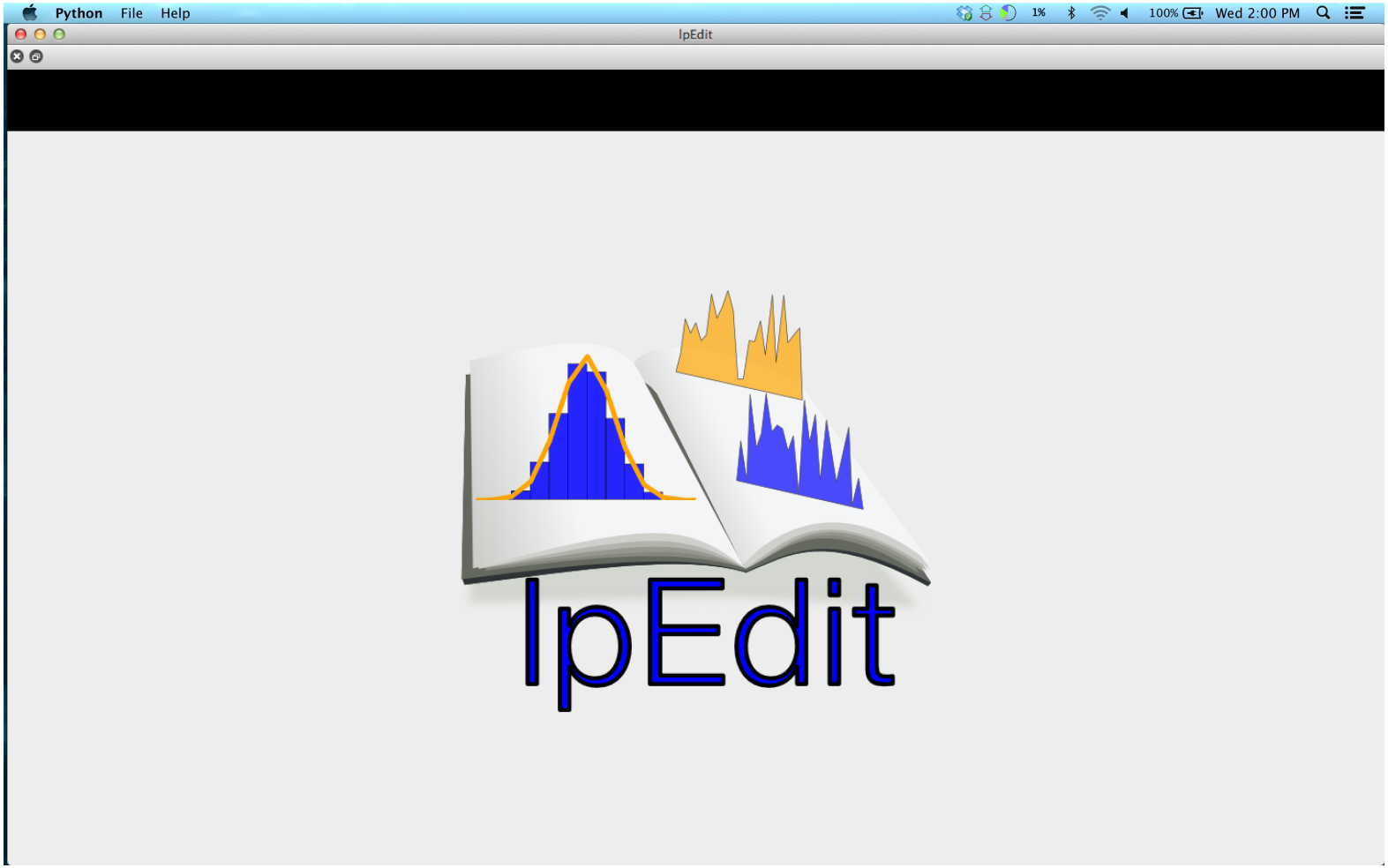
```
$ hg add *
$ hg commit -m "my first commit" --user <ajrichards>
$ hg push
```

5. copy it to a public_html directory (FileZilla if remote)

```
$ cp -r blah /public_html
```

IpEdit

:



About lpEdit

lpEdit has two parts:

1. An editor built specifically for literate programming.
2. A webpage that contains the editor documentation, examples and acts a general resource for reproducible research

lpEdit is

1. cross-platform
2. open source and freely available
3. Easy to use
4. based on QScintilla and written using PyQt4
5. working with both R and Python languages

Literate Programming

The concept was introduced by Donald Knuth in the 1970's. The idea is that we should be able to read and write code as if it were an expression of logic. Thus, the data, code and text used to tell the story must be presented as one single report.

A good literate programming tool should be...

1. Easy to use
2. Produce attractive output formats
3. Be capable of producing PDF, HTML and presentations
4. Should not be editor specific
5. Free and available to everyone
6. Well documented

syntax based on noweb

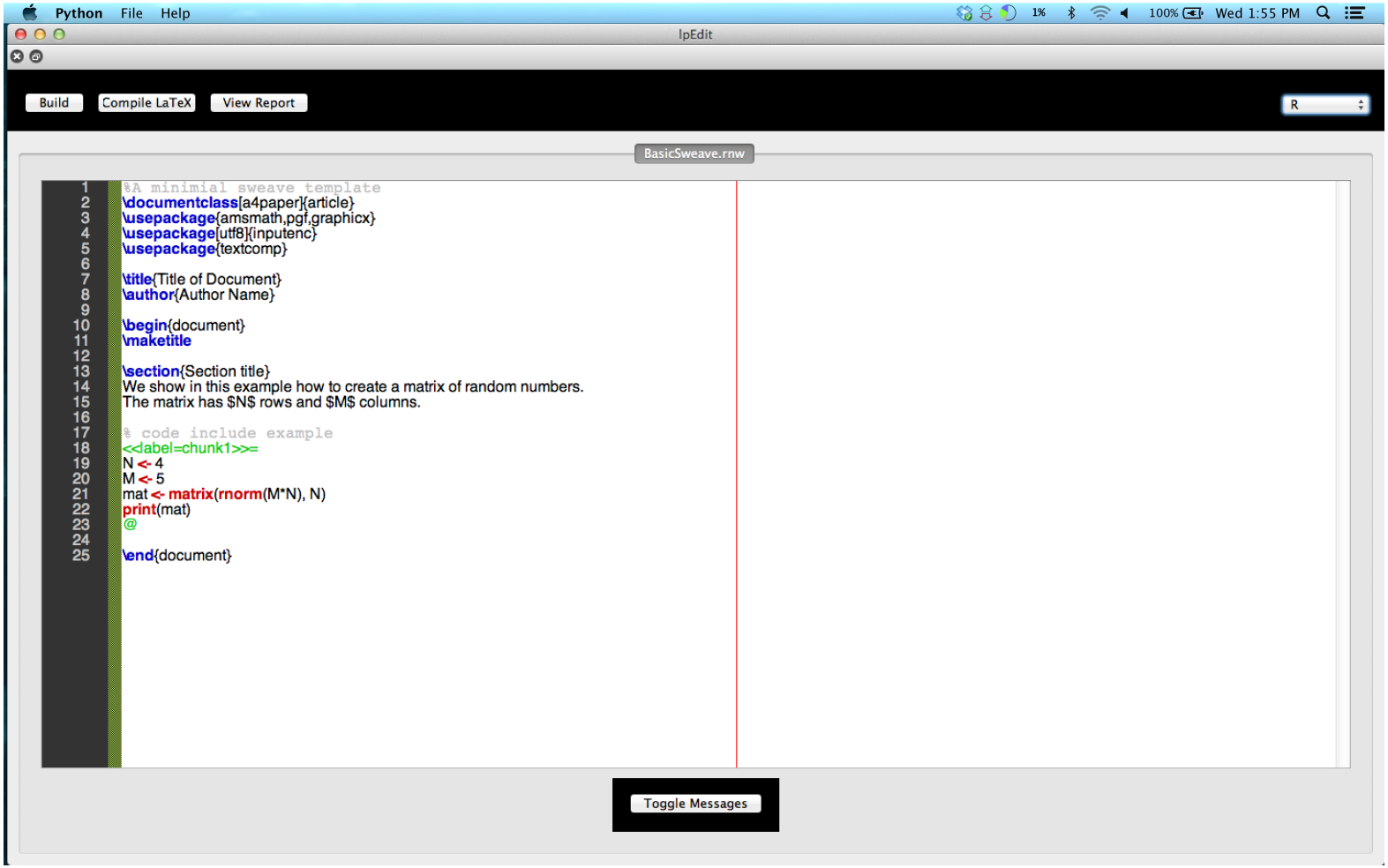
```
This is text that goes around the code

<<label=chunk1>>=
your code
@

This is more text that goes around the code
```

Basic Sweave program

```
\documentclass[a4paper]{article}
\usepackage{amsmath,pgf,graphicx,textcomp}
\usepackage[utf8]{inputenc}
\title{Title of Document}
\author{Author Name}

\begin{document}
\maketitle

\section{Section title}
This matrix has $N$ rows and $M$ columns.

<<label=chunk1>>=
N <- 4
M <- 5
mat <- matrix(rnorm(M*N), N)
print(mat)
@
\end{document}
```

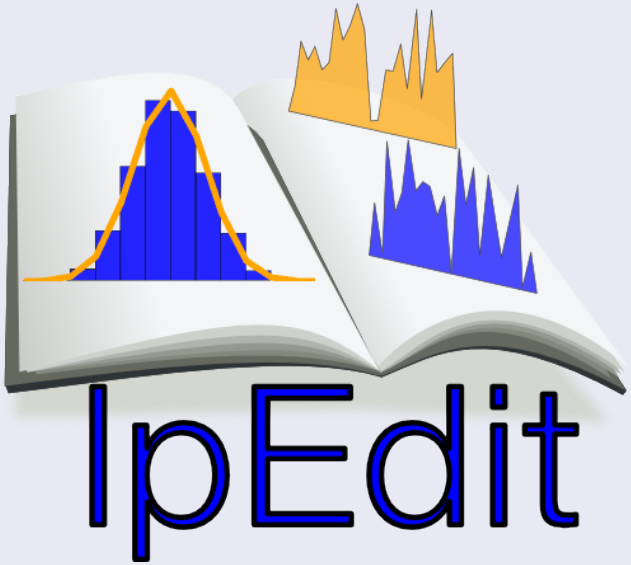
lpEdit - DEME

:


Python File Help
1% 100% Wed 1:55 PM
lpEdit
Build
Compile LaTeX
View Report
R
BasicSweave.rnw
%A minimial sweave template
\documentclass[a4paper]{article}
\usepackage{amsmath,pgf,graphicx}
\usepackage[utf8]{inputenc}
\usepackage{textcomp}
\title{Title of Document}
\author{Author Name}
\begin{document}
\maketitle
\section{Section title}
We show in this example how to create a matrix of random numbers.
The matrix has N rows and M columns.
% code include example
<<label=chunk1>>=
N <- 4
M <- 5
mat <- matrix(rnorm(M*N), N)
print(mat)
@
\end{document}
Toggle Messages

In the future...

- we may include additional languages
- we will have many more examples
- will produce HTML output at the click of a button
- will be working with restructured text markup (reST)
- will have more customization options
- code folding, inline spellcheck



In an ideal world...

Acknowledgments

⋮

Reproducible Analysis

All raw data and associated metadata described in this manuscript were deposited in the public repository *http://myrepository.org* and are available without restriction (for non-commercial purposes) to the general public.

All data manipulations and statistical analyses presented here are documented using literate programming and the complete bundle containing example data, code and the final report is available as supplemental methods.

Importantly, these groups have generously provided funding.

- Duke University - Duke Cancer Institute (DCI)
- Centre national de la recherche scientifique (CNRS)

These people have contributed to the project.

- Kouros Owzar (Duke University)
- Andrzej Kosinski (Duke University)