

# Module 1 - Lesson 01

History & Importance of Reproducibility & Transparency

[Melinda K. Higgins, PhD.](#)

2017-08-31



# Outline

- Timeline Reproducible Research & Transparency
- People
- Books
- Literate Programming > Dynamic Documentation > [R]Markdown
- The Big Picture

# Timeline Reproducible Research & Transparency<sup>1</sup>

YEAR	Event
1992	Jon Claerbout coined the term "reproducible research" in his book "EARTH SOUNDINGS ANALYSIS: Processing versus Inversion (PVI)" <sup>2</sup>
1996	CONSORT statement introduced standards for reporting clinical trials <sup>3</sup>
2004	International Committee of Medical Journal Editors (ICMJE) stated they would not publish a clinical trial that had not been registered. <sup>4</sup>
2005	Ioannidis, J. P. A. Why most published research findings are false. PLoS Med. 2, e124 (2005) <sup>5</sup>

1. Timeline partially based on PLOS Blog December 2016<http://blogs.plos.org/absolutely-maybe/2016/12/05/reproducibility-crisis-timeline-milestones-in-tackling-research-reliability/>

2. <http://sepwww.stanford.edu/sep/jon/reproducible.html>

3. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 276:637-639.

4. [http://www.icmje.org/news-and-editorials/update\\_2005.html](http://www.icmje.org/news-and-editorials/update_2005.html)

5. <https://doi.org/10.1371/journal.pmed.0020124>

# Timeline Reproducible Research & Transparency

YEAR	Event
2007	FDA Amendments Act (FDAAA) required more types of clinical trials to be registered (final rules took effect January 2017) <sup>6</sup>
2009	Journal of Biostatistics institutes policy to work with authors to publish articles that meet a standard of reproducibility. <sup>7</sup>
2011	Alsheikh-Ali, et.al. (2011), report the low percentage of researchers satisfying the policies regarding the availability and sharing of their data. <sup>8</sup>

6. <https://clinicaltrials.gov/ct2/manage-recs/fdaaa>

7. <https://academic.oup.com/biostatistics/article/10/3/405/293660/Reproducible-research-and-Biostatistics> & [https://academic.oup.com/biostatistics/pages/General\\_Instructions](https://academic.oup.com/biostatistics/pages/General_Instructions)

8. Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H. & Ioannidis, J. P. Public availability of published research data in high-impact journals. PloS ONE 6, e24357, 2011; <https://doi.org/10.1371/journal.pone.0024357>

# Cancer Testing Falls Apart

The Duke saga began when a prestigious journal, *Nature Medicine*, [published a paper](#) on Nov. 6, 2006, by Dr. Anil Potti, a cancer researcher at Duke University Medical Center; Joseph R. Nevins, a senior scientist there; and their colleagues. They wrote about genomic tests they developed that looked at the molecular traits of a cancerous tumor and figured out which [chemotherapy](#) would work best.

First, though, he asked two statisticians at M. D. Anderson, Keith Baggerly and Kevin Coombes, to check the work. Several other doctors approached them with the same request.

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.

Dr. Baggerly and Dr. Coombes tried to sound an alarm. They got the attention of the National Cancer Institute, whose own investigators wanted to use the Duke system in a clinical trial but were dissuaded by the criticisms. Finally, they [published their analysis](#) in *The Annals of Applied Statistics*, a journal that medical scientists rarely read.

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

2010 Video Presentation by Keith A. Baggerly  
[http://videolectures.net/cancerbioinformatics2010\\_baggerly\\_irrh/](http://videolectures.net/cancerbioinformatics2010_baggerly_irrh/)

# The Excel-Error Heard Around the World



## The Weird and Very Real World of Excel-Error Research

The Rogoff-Reinhart blunder is a prominent example of a very common problem

BY ROBERT LONG | April 18, 2013

They're calling it the "Excel Error Heard Round the World": Kenneth Rogoff and Carmen Reinhart's widely cited paper about the relationship between public debt and economic growth was revealed Monday to have grossly misstated economic growth for high-debt countries, all because of a forehead-smackingly simple error in an Excel spreadsheet. ("It is sobering that such an error slipped into one of our papers despite our best efforts to be consistently careful," the paper's authors said on Wednesday.)

<https://newrepublic.com/article/112951/rogoff-reinhart-and-world-excel-error-research>

# Timeline Reproducible Research & Transparency

YEAR	Event
2012	Begley and Ellis reviewed 53 "landmark" studies and only 6 (11%) had the scientific findings confirmed. <sup>9</sup>
2013	Center for Open Science launches & by 2014 the Open Science Framework has 7000 users with more than 45,000+ and over 15 institutions by 2017 <sup>10</sup>
2014	NIH publishes their guidelines for addressing reproducibility <sup>11</sup>
2015	The Open Science Collaboration reports that they were only able to replicate between 1/3 to 1/2 of the results from 100 studies <sup>12</sup>

9. <http://www.nature.com/nature/journal/v483/n7391/full/483531a.html>

10. <https://cos.io/about/brief-history-cos-2013-2017/> & <https://osf.io/>

11. <https://www.nih.gov/research-training/rigor-reproducibility>

12. Science, 28 Aug 2015; Vol. 349, Issue 6251, aac4716; DOI: 10.1126/science.aac4716; <http://science.sciencemag.org/content/349/6251/aac4716>

# Wide-Spread Gene Name Errors

The screenshot shows a web browser displaying an article from the journal *Genome Biology*. The URL in the address bar is <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7>. The page title is "Gene name errors are widespread in the scientific literature". The authors listed are Mark Ziemann, Yotam Eren and Assam El-Osta. The article is marked as "OPEN ACCESS". The abstract discusses how Microsoft Excel converts gene names to dates and floating-point numbers, with approximately one-fifth of papers containing erroneous conversions.

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta [✉](#)

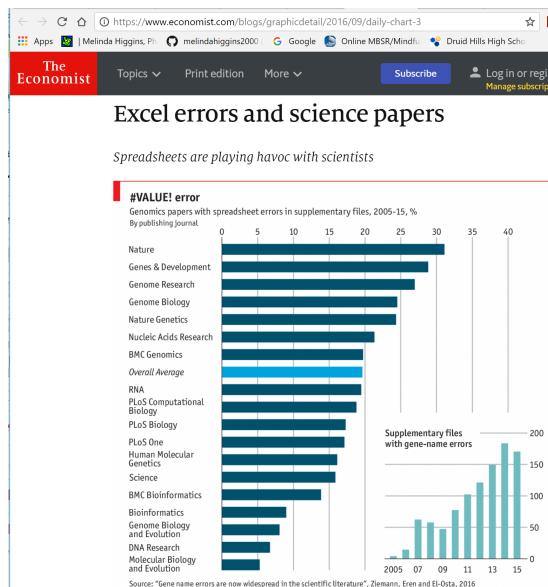
*Genome Biology* 2016, 17:177 | <https://doi.org/10.1186/s13059-016-1044-7> | © The Author(s). 2016  
Published: 23 August 2016

## Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7>

# Wide-Spread Gene Name Errors



<https://www.economist.com/blogs/graphicdetail/2016/09/daily-chart-3>

# People

- Victoria Stodden <https://ischool.illinois.edu/people/faculty/vcs>
  - presentation on History of the Reproducibility Movement  
<https://web.stanford.edu/~vcs/talks/ICERM-Dec102012STODDEN.pdf>
  - co-author "Implementing Reroducible Research" book  
<https://www.crcpress.com/Implementing-Reproducible-Research/Stodden-Leisch-Peng/p/book/9781466561595>
- Roger Peng <http://www.biostat.jhsph.edu/~rpeng/index.html>
  - Associate Editor for Reproducible Research - Biostatistics Journal  
[https://academic.oup.com/biostatistics/pages/Editorial\\_Board](https://academic.oup.com/biostatistics/pages/Editorial_Board)
  - co-author "Implementing Reroducible Research" book  
<https://www.crcpress.com/Implementing-Reproducible-Research/Stodden-Leisch-Peng/p/book/9781466561595>

# People

- John P.A. Ioannidis [https://profiles.stanford.edu/john-ioannidis?  
tab=publications](https://profiles.stanford.edu/john-ioannidis?tab=publications)
  - Professor of Medicine and of Health Research and Policy at Stanford University School of Medicine and a Professor of Statistics at Stanford University School of Humanities and Sciences
- Christopher Gandrud [https://www.iq.harvard.edu/people/christopher-  
gandrud](https://www.iq.harvard.edu/people/christopher-gandrud)
  - research fellow at IQSS (Institute for Quantitative Social Science)
  - Book Author "Reproducible Research with R and RStudio"  
[https://www.crcpress.com/Reproducible-Research-with-R-and-R-  
Studio/Gandrud/p/book/9781466572843](https://www.crcpress.com/Reproducible-Research-with-R-and-R-Studio/Gandrud/p/book/9781466572843)

# People

- Yihui Xie
  - software engineer for RStudio <https://www.rstudio.com/about/>
  - author of "Dynamic Documents with R and knitr" <https://www.crcpress.com/Dynamic-Documents-with-R-and-knitr/Xie/p/book/9781482203530>
  - author of "Bookdown: Authoring Books and Technical Documents with R Markdown" book <https://www.crcpress.com/bookdown-Authoring-Books-and-Technical-Documents-with-R-Markdown/Xie/p/book/9781138700109> and bookdown R package <https://cran.r-project.org/web/packages/bookdown/index.html>
  - author of blogdown R package <https://cran.r-project.org/web/packages/blogdown/index.html>

# People

- Federich Leisch
  - Professor of Applied Statistics at the University of Natural Resources and Life Sciences, Vienna
  - developer of Sweave for creating dynamic reports  
<https://leisch.userweb.mwn.de/Sweave/>
  - co-author "Implementing Rerducible Research" book  
<https://www.crcpress.com/Implementing-Reproducible-Research/Stodden-Leisch-Peng/p/book/9781466561595>

# Books on Reproducibility and Tools of the Trade

## Image



## Book

Implementing Reproducible Research by Victoria Stodden, Friedrich Leisch, Roger D. Peng <https://www.crcpress.com/Implementing-Reproducible-Research/Stodden-Leisch-Peng/p/book/9781466561595>



Dynamic Documents with R and knitr (Chapman & Hall/CRC The R Series) 1st Edition by Yihui Xie <https://www.crcpress.com/Dynamic-Documents-with-R-and-knitr/Xie/p/book/9781482203530>



bookdown: Authoring Books and Technical Documents with R Markdown by Yihui Xie <https://www.crcpress.com/bookdown-Authoring-Books-and-Technical-Documents-with-R-Markdown/Xie/p/book/9781138700109> & read online <https://bookdown.org/yihui/bookdown/>

---

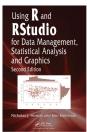
# more books

## Image



## Book

Happy Git and GitHub for the useR by Jenny Bryan; read online  
<http://happygitwithr.com/>



Using R and RStudio for Data Management, Statistical Analysis, and Graphics,  
Second Edition by Nicholas J. Horton & Ken Kleinman; also see [Project MOSAIC](#),  
<http://mosaic-web.org/>



ModernDive: An Introduction to Statistical and Data Sciences via R by Chester Ismay and Albert Y. Kim; read online <https://ismayc.github.io/moderndiver-book/> &  
Getting used to R, RStudio, and R Markdown by Chester Ismay  
<https://ismayc.github.io/rbasics-book/>

... and lots more ... see <https://bookdown.org/>

---

# Literate Programming > Dynamic Documentation

## > [R]Markdown

YEAR      Event

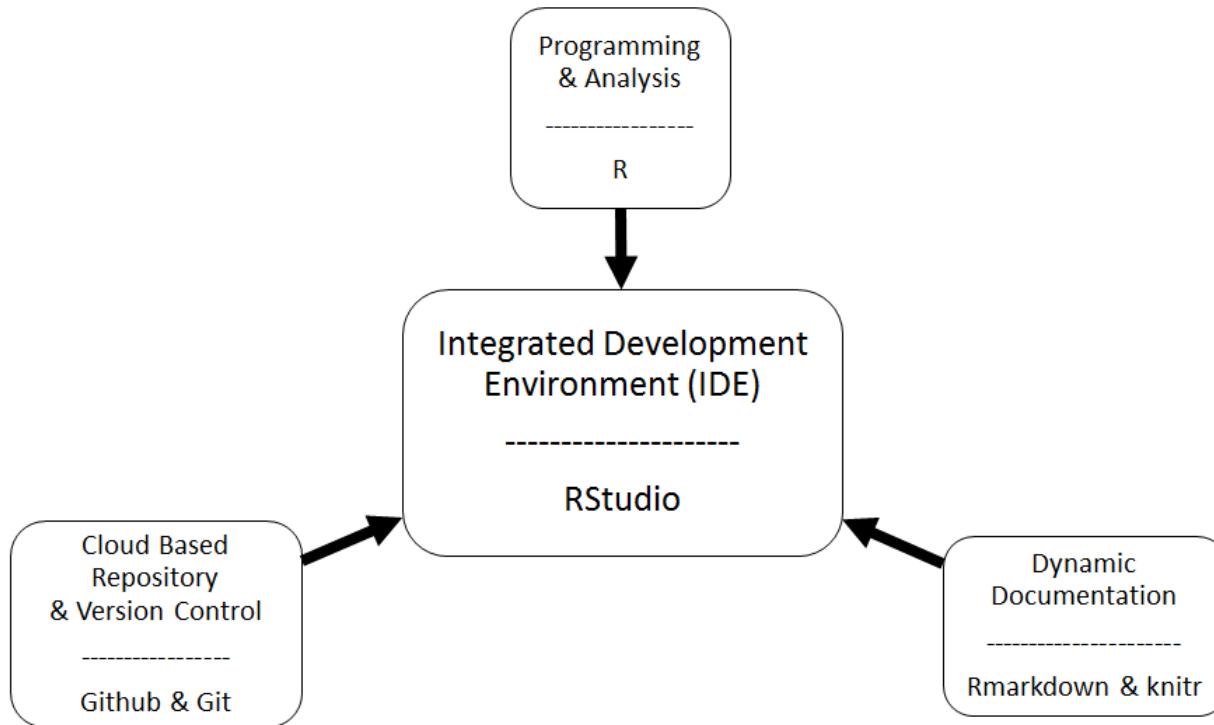
1992	"Literate Programming" is introduced by Donald Knuth as "that (which) combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained, and arguably more fun to write than programs that are written only in a high-level language. <b>The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer.</b> " <a href="http://www-cs-faculty.stanford.edu/~knuth/lp.html">http://www-cs-faculty.stanford.edu/~knuth/lp.html</a>
2002	Friedrich Leisch introduces <b>SWEAVE</b> a program for "Dynamic generation of statistical reports using literate data analysis" <a href="https://leisch.userweb.mwn.de/Sweave/">https://leisch.userweb.mwn.de/Sweave/</a>

# Literate Programming > Dynamic Documentation

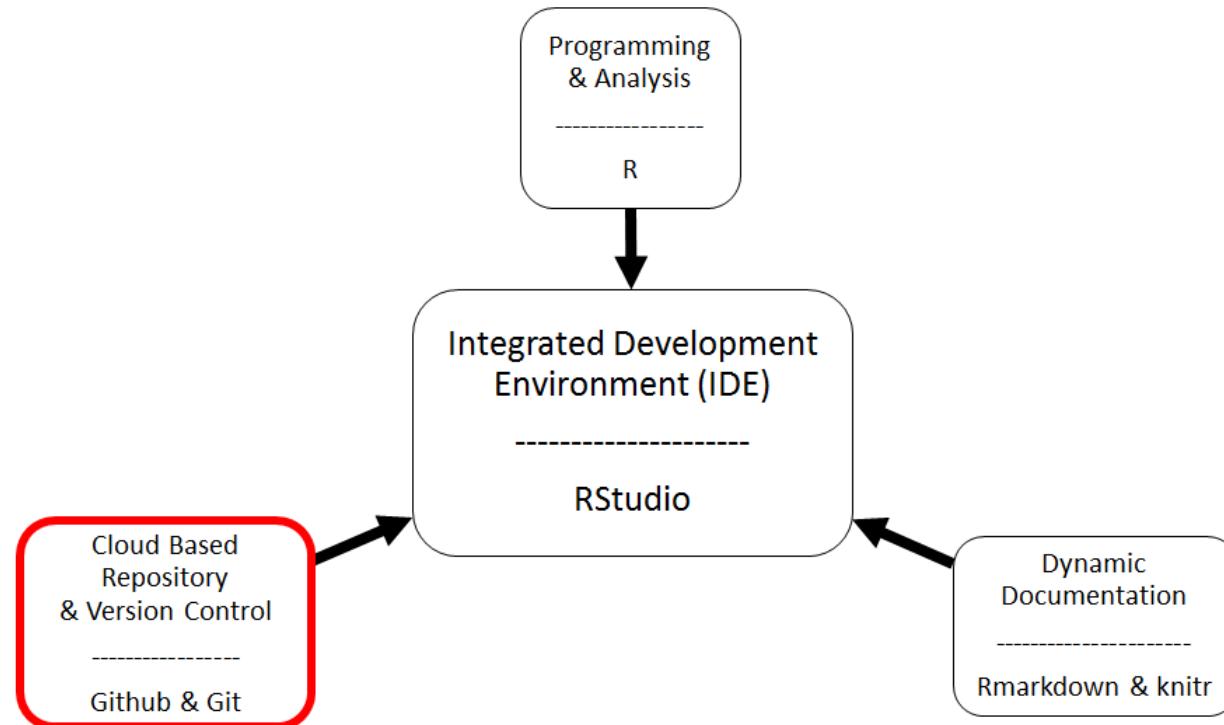
## > [R]Markdown

YEAR	Event
2004	John Gruber created the <b>Markdown</b> language in 2004 in collaboration with Aaron Swartz - their goal was to "write using an easy-to-read, easy-to-write plain text format, and optionally convert it to structurally valid XHTML (or HTML)" <a href="https://daringfireball.net/projects/markdown/">https://daringfireball.net/projects/markdown/</a>
2012	Yihui Xie releases <b>knitr</b> R package released - <b>knitr</b> was inspired by <b>SWEAVE</b>
2014	<b>rmarkdown</b> R package released - extends <b>Markdown</b> to work with R/RStudio environment

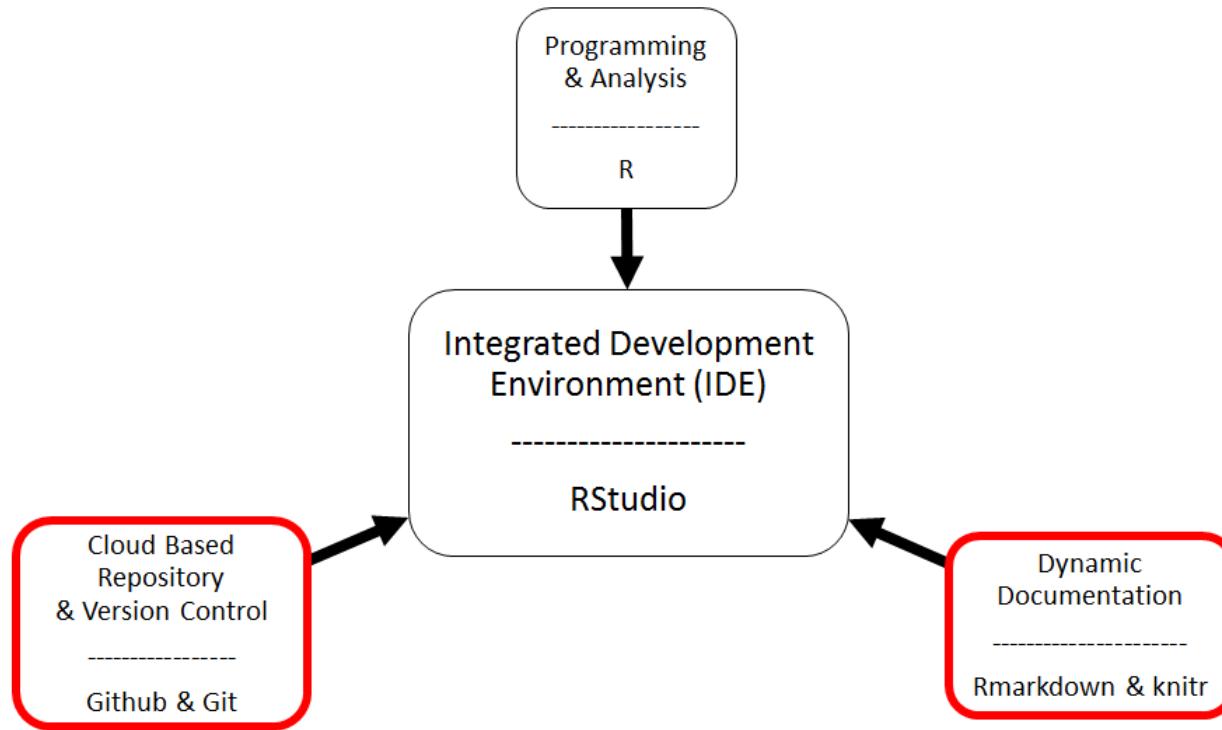
# The Big Picture



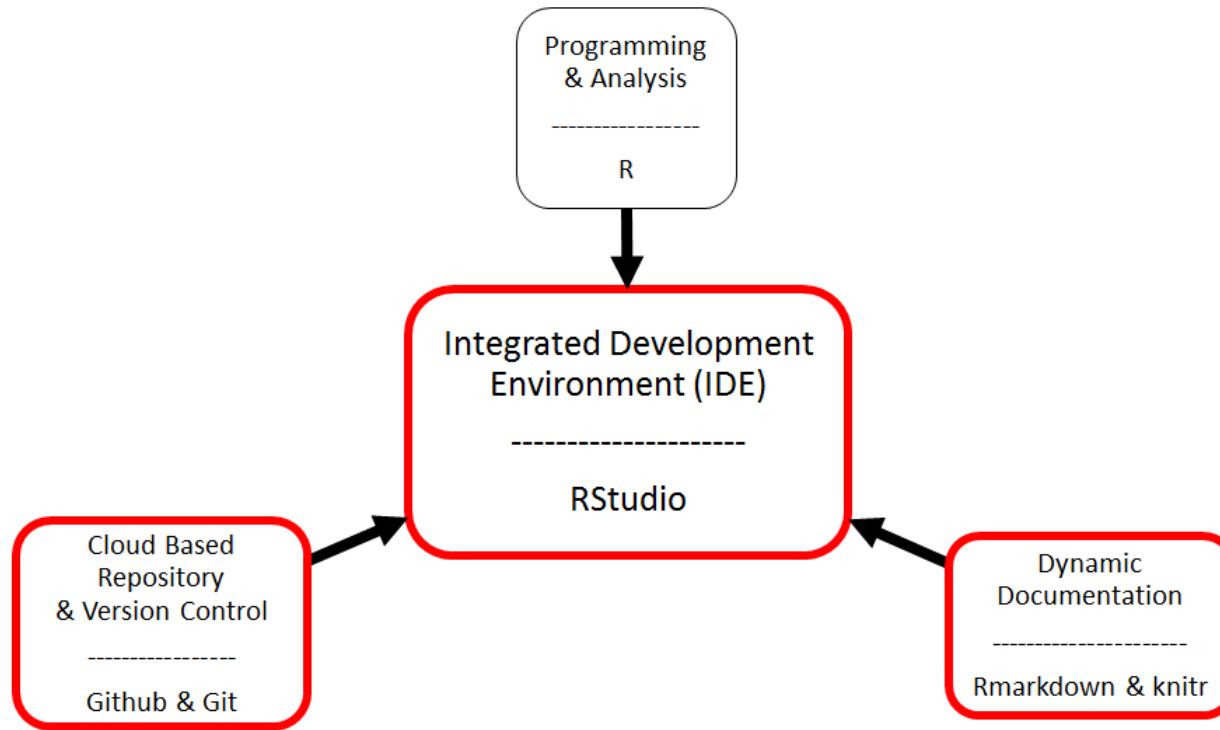
# The Big Picture



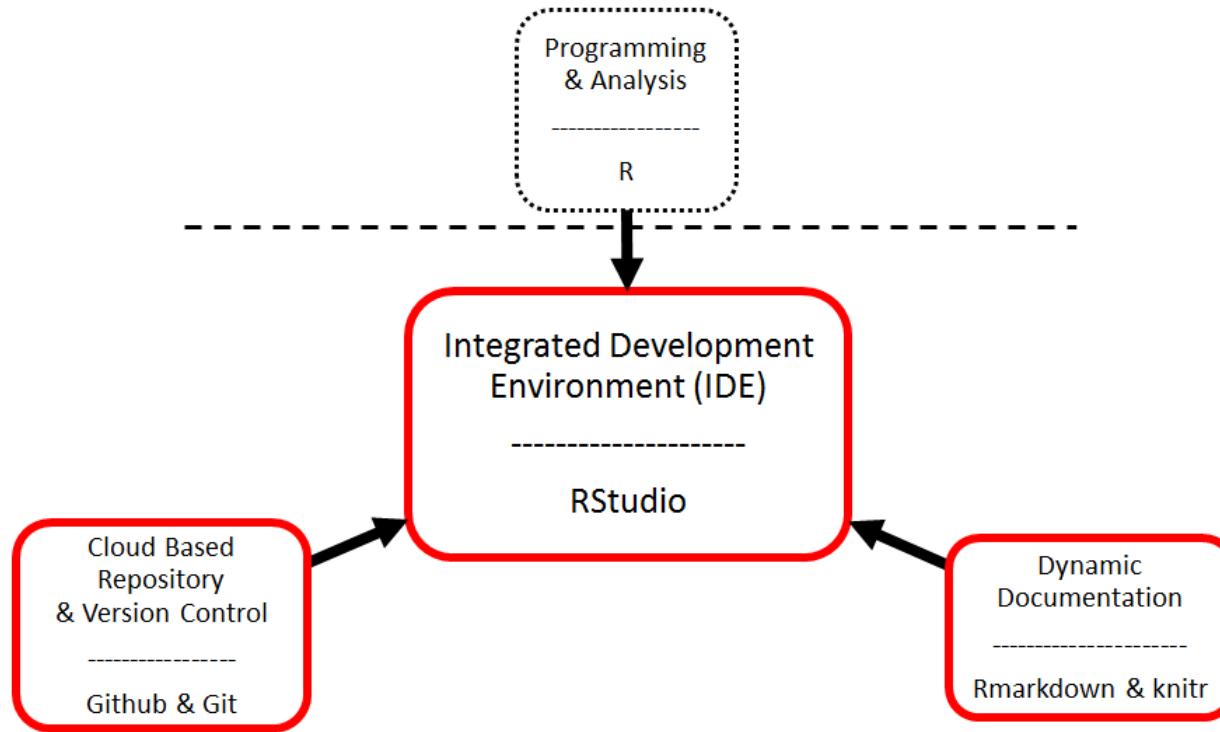
# The Big Picture



# The Big Picture



# The Big Picture



# Next in Lesson 02 ...

Literate Programming

&

Dynamic Documentation