

Reproducible Research with R

Arnaud Legrand

COMPAS tutorials, Neuchâtel, April 2014

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- Reproducible Documents: knitR
- Introduction to R
- Needful Packages by Hadley Wickam

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- Reproducible Documents: knitR
- Introduction to R
- Needful Packages by Hadley Wickam

As a Reviewer

This may be an interesting contribution but:

- This **average value** must hide something.
- As usual, there is no **confidence interval**, I wonder about the variability and whether the difference is **significant** or not.
- That can't be true, I'm sure they **removed some points**.
- Why is this graph in **logscale** ? How would it look like otherwise ?
- The authors decided to show only a **subset of the data**. I wonder what the rest looks like.
- There is no label/legend/... What is the **meaning of this graph** ? If only I could access the generation script.

As an Author

- I thought I used the same parameters but I'm getting different results!
- The new student wants to compare with the method I proposed last year.
- The damn reviewer asked for a major revision and wants me to change this figure. :(
- Which code and which data set did I use to generate this figure?
- It worked yesterday!
- Why did I do that?

My Feeling

Computer scientists have an incredibly poor training in probabilities and statistics

Why should we ? Computer are deterministic machines after all, right? ;)

Eight years ago, I've started realizing how lame the articles I reviewed (as well as those I wrote) were in term of experimental methodology.

- Yeah, I know, your method/algorithm is better than the others as demonstrated by the figures
- Not enough information to discriminate real effects from noise
- Little information about the workload
- Would the “conclusion” still hold with a slightly different workload?
- I’m tired of awful combination of tools (perl, gnuplot, sql, . . .) and bad methodology

Current practice in CS

Computer scientists tend to either:

- vary one factor at a time, use a very fine sampling of the parameter range,
- run millions of experiments for a week varying a lot of parameters and then try to get something of it. Most of the time, they (1) don't know how to analyze the results (2) realize something went wrong...

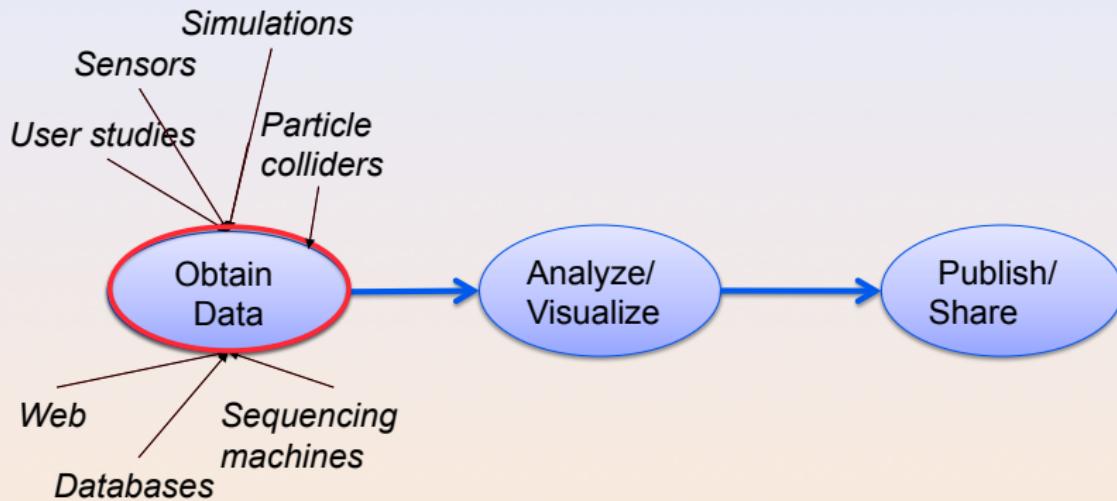
Interestingly, most other scientists do the exact opposite.

These two flaws come from poor training and from the fact that C.S. experiments are almost free and very fast to conduct.

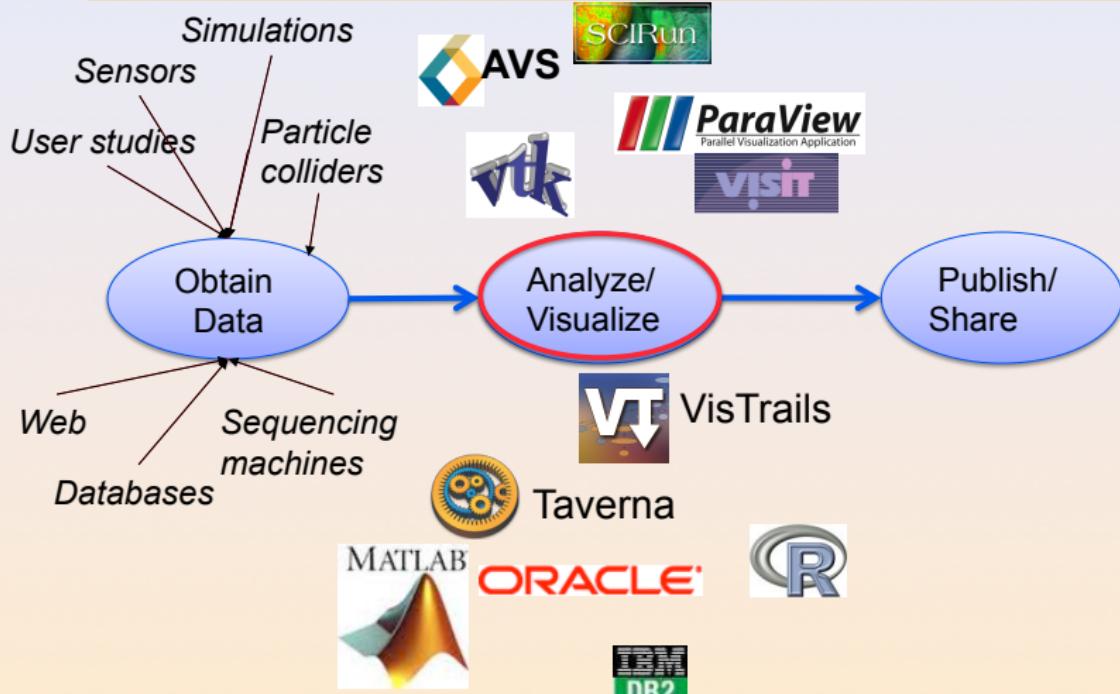
- Most strategies of experimentation have been designed to provide sound answers despite all the randomness and uncontrollable factors;
- Maximize the amount of information provided by a given set of experiments;
- Reduce as much as possible the number of experiments to perform to answer a given question under a given level of confidence.

Takes a few lectures on Design of Experiments to improve but anyone can start by reading Jain's book on The Art of Computer Systems Performance Analysis

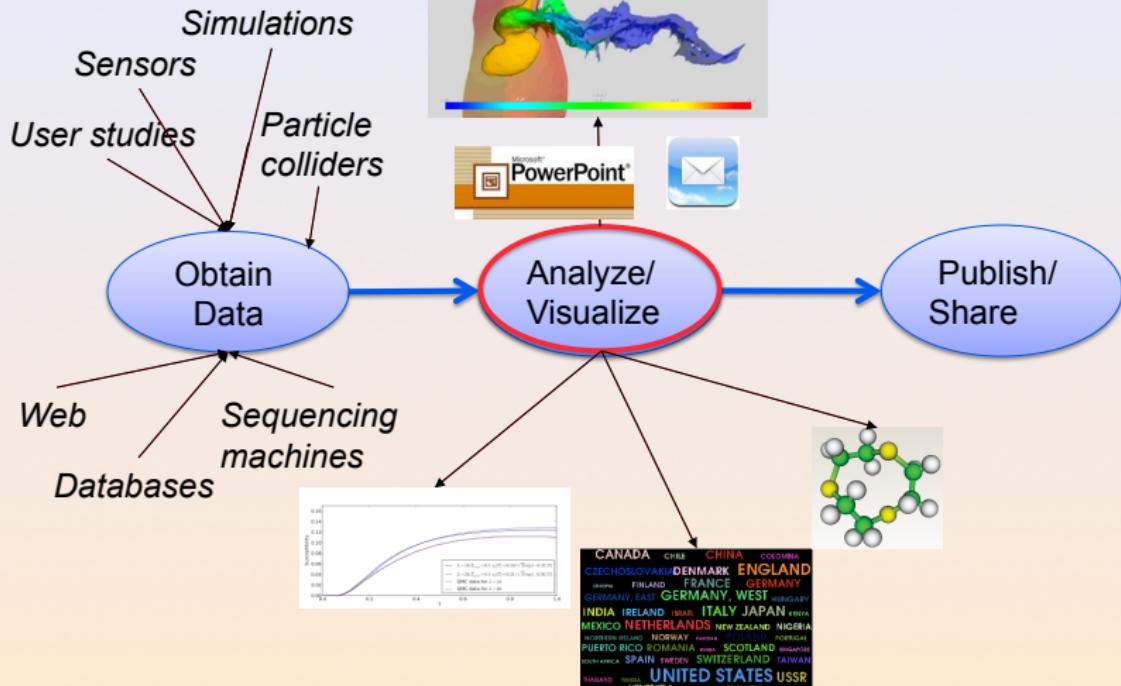
Science Today: Data Intensive



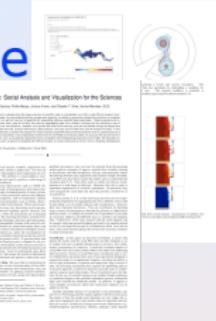
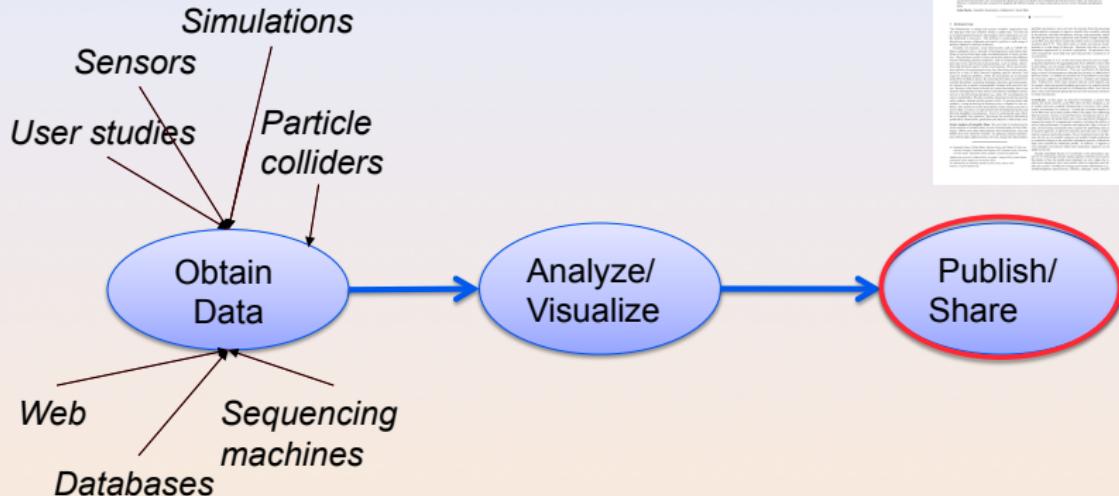
Science Today: Data + Computing Intensive



Science Today: Data + Computing Intensive



Science Today: Data + Computing Inte



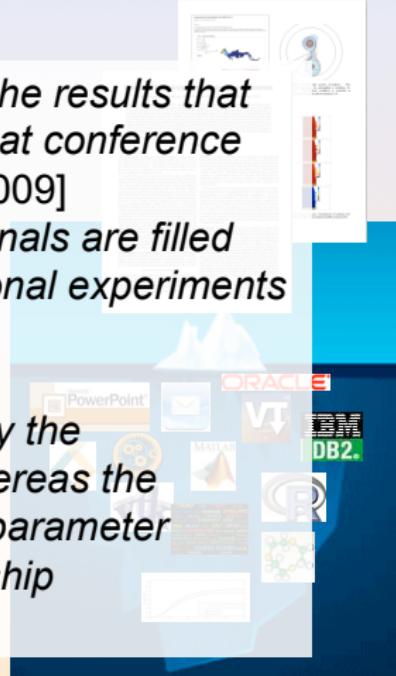
Science Today: Incomplete Publications

- ◆ Publications are just the tip of the iceberg
 - Scientific record is incomplete---to large to fit in a paper
 - Large volumes of data
 - Complex processes
- ◆ Can't (easily) reproduce results



Science Today: Incomplete Publications

- ◆ Publications are just the tip of the iceberg
 - “It’s impossible to verify most of the results that computational scientists present at conference and in papers.” [Donoho et al., 2009]
 - “Scientific and mathematical journals are filled with pretty pictures of computational experiments
- ◆ Can’t really repeat it
 - “Published documents are merely the advertisement of scholarship whereas the computer programs, input data, parameter values, etc. embody the scholarship itself.” [Schwab et al., 2007]



Reproducibility: What Are We Talking About ?

Replicability

Reproducibility

Reproduction of the original results using the same tools

by the original author on the same machine

by someone in the same lab/using a different machine

by someone in a different lab

Reproduction using different software, but with access to the original code

Completely independent reproduction based only on text description, without access to the original code

A Difficult Trade-off

Automatically keeping track of everything

- the code that was run (source code, libraries, compilation procedure)
- processor architecture, OS, machine, date, ...

VM-based solutions

Ensuring others can redo/understand what you did

- Why did I run this?
- Does it still work when I change this piece of code for this one?

Laboratory notebook and recipes

Reproducible Research: the New Buzzword ?

H2020-EINFRA-2014-2015

A key element will be capacity building to link literature and data in order to enable a more transparent evaluation of research and reproducibility of results.

More and more workshops

- Workshop on Duplicating, Deconstructing and Debunking (WDDD) (2014 edition)
- Reproducible Research: Tools and Strategies for Scientific Computing (2011)
- Working towards Sustainable Software for Science: Practice and Experiences (2013)
- REPPAR'14: 1st International Workshop on Reproducibility in Parallel Computing
- Reproducibility@XSEDE: An XSEDE14 Workshop
- Reproduce/HPCA 2014
- TRUST 2014

Should be seen as opportunities to share experience.

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- Reproducible Documents: knitR
- Introduction to R
- Needful Packages by Hadley Wickam

Our Approach: An Infrastructure to Support Provenance-Rich Papers [Koop et al., ICCS 2011]

- ◆ Tools for *authors* to create reproducible papers
 - Specifications that encode the computational processes
 - Package the results
 - Link from publications
- ◆ Tools for testers to repeat and validate results
 - Explore different parameters, data sets, algorithms
- ◆ Interfaces for searching, comparing and analyzing experiments and results
 - Can we discover better approaches to a given problem?
 - Or discover relationships among workflows and the problems?
 - How to describe experiments?

Support different approaches

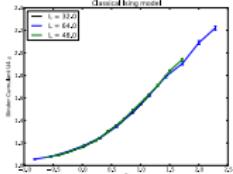
Vistrails: a Workflow Engine for Provenance Tracking

An Provenance-Rich Paper: ALPS2.0

The ALPS project release 2.0:
Open source software for strongly correlated systems

B. Bauer¹ L. D. Carr² H.G. Evertz³ A. Feiguin⁴ J. Freire⁵
S. Fuchs⁶ L. Gamper¹ J. Gukelberger⁶ E. Gulf⁷ S. Guertler⁸
A. Hehn⁹ R. Igashiri¹⁰ S. Isakov¹ D. Koop² P.N. Ma¹¹
P. Mates^{1,2} H. Matsuo¹¹ O. Parcollet¹² G. Pawłowski¹³
J.D. Picon¹⁴ L. Pollet¹⁵ E. Santos¹⁶ V.W. Scarola¹⁶
U. Schollwöck¹⁷ C. Silva¹⁸ B. Surer¹⁹ S. Todo^{11,20} S. Trebst¹⁶
M. Troyer¹¹ M. L. Wall²¹ P. Werner¹ S. Wessel^{1,20}

¹Theoretische Physik, ETH Zürich, 8093 Zürich, Switzerland
²Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
³Institut für Theoretische Physik, Technische Universität Graz, A-8010 Graz, Austria
⁴Department of Physics and Astronomy, University of Wyoming, Laramie, Wyoming 82071, USA
⁵Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah 84112, USA
⁶Institut für Theoretische Physik, Georg-August-Universität Göttingen, Göttingen, Germany
⁷Columbia University, New York, NY 10027, USA
⁸Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany



Classical Ising model

Legend: $L = 32, L = 48, L = 64$

arXiv:1101.2646v4 [cond-mat.str-el] 23 May 2011

1 Correspondence to: juliana.freire@ethz.ch

Workflow

Libraries

ALPS 2.0

matplotlib

Data

Simulation Results

Figure 3. In this example we show a data collapse of the Binder Cumulant in the classical Ising model. The data has been produced by remotely run simulations and the critical exponent has been obtained with the help of the VisTrails parameter exploration functionality.

VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Regular program code

```
figure1 = plot(x)  
save(figure1,'figure1.eps')
```

```
> file /home/figure1.eps saved  
>
```

VCR: A Universal Identifier for Computational Results

Chronicling computations in real-time

VCR computation platform Plugin = Computation recorder

Program code with VCR plugin

```
repository vcr.nature.com  
verifiable figure1 = plot(x)
```

```
> vcr.nature.com approved:
```

```
> access figure1 at https://vcr.nature.com/ffaaffb148d7
```

VCR: A Universal Identifier for Computational Results

Word-processor plugin App

LaTeX source

```
\includegraphics{figure1.eps}
```

LaTeX source with VCR package

```
\includeresult{vcr.thelancet.com/ffaaffb148d7}
```

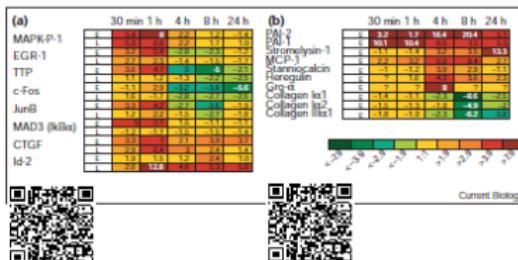
Permanently bind printed graphics to underlying result content

VCR: A Universal Identifier for Computational Results

Research Paper Analysis of replicative senescence Shelton et al. 943

Figure 3

Time course of serum stimulation. (a) Early passage (E; PD30) or late passage (L; PD89) BJ cultures were held in 0.5% serum for 2 days, then stimulated with 10% FBS. RNA levels from cultures at the indicated time points (Cy5 channel) were compared with the uninduced starting culture (Cy3 channel). Positive values indicate higher expression in induced cells; negative values indicate lower expression in induced cells. Question marks indicate that there was insufficient signal for detection. A complete listing of serum-responsive genes from this analysis is provided in Supplementary material. (b) The serum-responsiveness of select senescence-regulated genes in early passage (PD30) BJ fibroblasts.



senescence response appears to overlap substantially with gene expression patterns observed in activated fibroblasts during wound healing [24–26]. MCP-1, Gro- α , IL-1 β and IL-15 are strong effectors of macrophage and neutrophil recruitment and activation [27,28]. The upregulation of Toll (Tlr-4) in senescent fibroblasts confirms the overall immune response behavior at senescence. Tlr-4 is an IL-1 receptor homolog and is implicated in the activation of the gene regulatory protein NF- κ B, a function proposed to be part of the innate immune response [29]. The induction of IL-15 at senescence is also consistent with an innate immune response, as IL-15 can be induced by NF- κ B-dependent transcription [30] and also participates in inflammatory disease processes [28].

Deficiencies in the response of senescent cells to serum stimulation have been reported, and include an inability to induce the expression of *c-fos* mRNA [31] and markers of late G1 and S phase [32]. In response to serum, expression of inflammatory chemokines, matrix-degrading proteases and their modulators is induced in early-passage dermal fibroblasts, and expression of matrix collagens is reduced. This transient burst of activity may represent a natural progression of events in early-passage cells. Transcripts were hyper-induced in serum-stimulated senescent cells, and included markers of the immune system, such as TLR-4, cathepsin-1, and IL-15. These genes are known to be involved in the immune response to infection and tissue damage. The expression of these genes in senescent cells suggests that they may play a role in the immune response to tissue damage or infection.

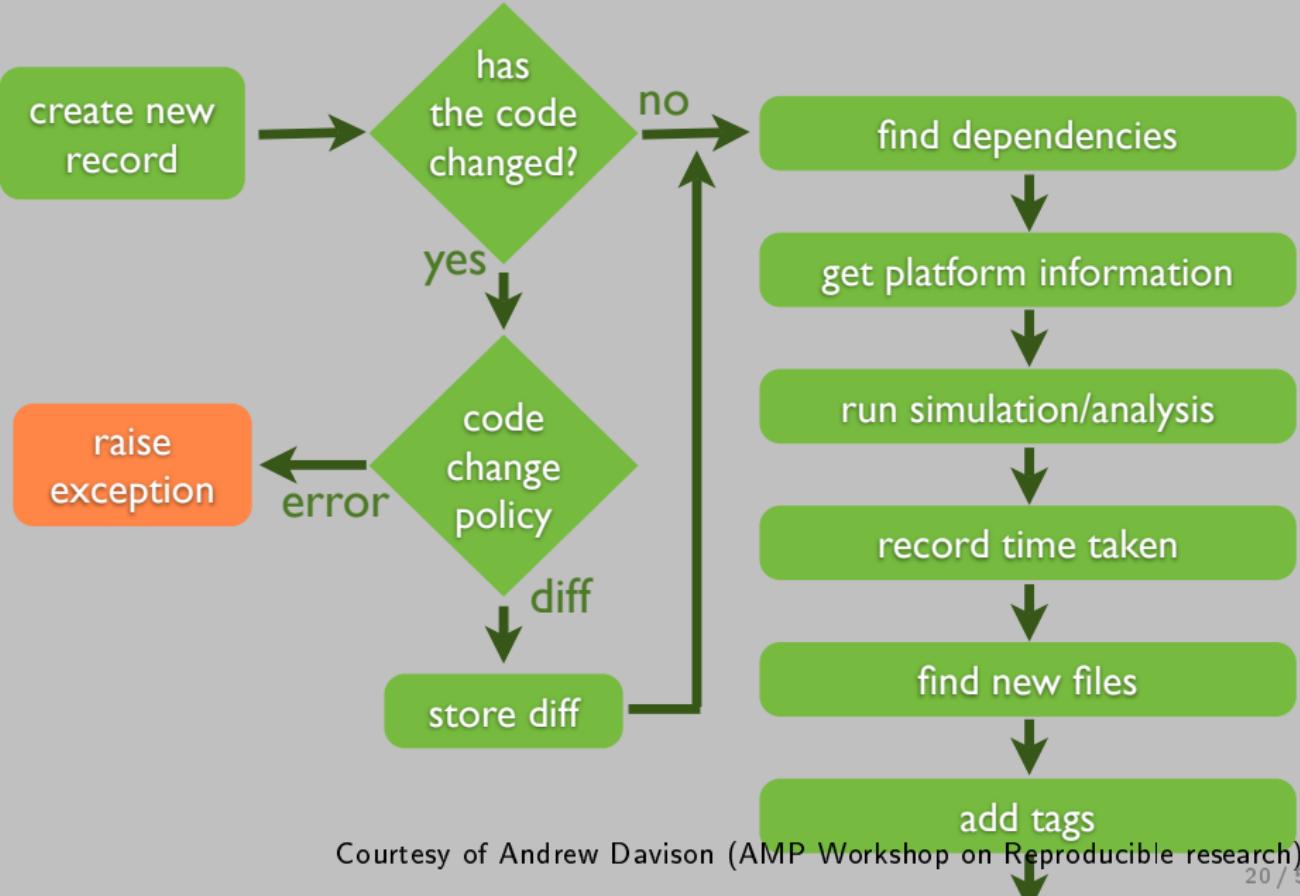
states overlap substantially with those in telomere-induced senescence (W.F., D.N.S., R. Allsopp, S. Lowe, and G. Ferbeyre, unpublished observations) and thus are likely to use many of the same activation processes.

The pattern of gene expression at senescence varies substantially in different cell types. Although the expression of matrix and structural proteins, such as the collagens, keratins and auxiliary factors, is repressed in RPE cells, inflammatory regulators are not induced, in contrast to dermal fibroblasts. Physiologically, this would make sense, as an acute inflammatory response in a tissue critical for normal vision would be likely to have deleterious consequences. However, as the RPE layer has a central role in the deposition and maintenance of extracellular matrix in the retina, decrements in the ability of senescent RPE cells to maintain appropriate expression patterns, as evidenced by decreased expression of collagens, keratins, aggrecan, transglutaminase and so on, would be predicted to have adverse effects on retinal architecture. Dysfunction of the RPE cell layer is considered to be a substantial factor in the development of age-related macular degeneration [36].

Surprisingly, early-passage cells also overexpress many of the markers associated with senescence in dermal fibroblasts, including TLR-4, cathepsin-1, and IL-15. This suggests that early-passage cells may be undergoing a transition from a proliferative to a senescent state, or that they may be responding to external stimuli in a way that is similar to senescent cells.

Courtesy of Marjan Gavish and David Donoho (AMP Workshop on Reproducible research)

Sumatra: a lab notebook



Courtesy of Andrew Davison (AMP Workshop on Reproducible research)

Sumatra: a lab notebook

```
$ smt comment 20110713-174949 "Eureka! Nobel prize  
here we come."
```

Sumatra: a lab notebook

```
$ smt tag "Figure 6"
```

Sumatra: a lab notebook

Sumatra: TestProject: List of records

TestProject: List of records

Delete Include data	Label	Reason	Outcome	Duration	Processes	Simulator		Script			Date	Time	Tags
						Name	Version	Repository	Main file	Version			
<input type="checkbox"/>	20100709-154255		'Eureka! Nobel prize here we come.'	0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:42:55	
<input type="checkbox"/>	20100709-154309			0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:09	
<input type="checkbox"/>	haggling	'determine whether the gourd is worth 3 or 4 shekels'	'apparently, it is worth NaN shekels.'	0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:20	foobar
<input type="checkbox"/>	20100709-154338	'test effect of a smaller time constant'		0.59 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:38	
<input type="checkbox"/>	haggling_repeat	Repeat experiment haggling	The new record exactly matches the original.	0.58 s		Python	2.5.2	/Users/andrew/tmp/SumatraTest	main.py	396c2020ca50	09/07/2010	15:43:47	

So many new tools

New Tools for Computational Reproducibility

- Dissemination Platforms:

[ResearchCompendia.org](#)

[IPOL](#)

[Madagascar](#)

[MLOSS.org](#)

[thedatahub.org](#)

[nanoHUB.org](#)

[Open Science Framework](#)

[The DataVerse Network](#)

[RunMyCode.org](#)

- Workflow Tracking and Research Environments:

[VisTrails](#)

[Kepler](#)

[CDE](#)

[Galaxy](#)

[GenePattern](#)

[Synapse](#)

[Sumatra](#)

[Taverna](#)

[Pegasus](#)

- Embedded Publishing: Courtesy of Victoria Stodden (UC Davis, Feb 13, 2014)

[Verifiable Computational Research](#)

[Sweave](#)

[knitR](#)

[Collage Authoring Environment](#)

[SHARE](#)

And also: Figshare, ActivePapers, Elsevier executable paper, ...

Literate programming

Donald Knuth: explanation of the program logic in a natural language interspersed with snippets of macros and traditional source code.

I'm way too 3133t to program this way but that's exactly what we need for writing a reproducible article/analysis!

Org-mode (requires emacs)

My favorite tool.

- plain text, very smooth, works both for html, pdf, ...
- allows to combine all my favorite languages

Ipython notebook

If you are a python user, go for it! Web app, easy to use/setup...

KnitR (a.k.a. Sweave)

For non-emacs users and as a first step toward reproducible papers:

- Click and play with a modern IDE

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

• General Introduction

- Reproducible Documents: knitR
- Introduction to R
- Needful Packages by Hadley Wickam

Why R?

R is a great language for data analysis and statistics

- Open-source and multi-platform
- Very expressive with high-level constructs
- Excellent graphics
- Widely used in academia and business
- Very active community
 - Documentation, FAQ on <http://stackoverflow.com/questions/tagged/r>
- Great integration with other tools

Why is R a pain for computer scientists ?

- R is **not** really a **programming** language
- Documentation is for statisticians
- Default plots are **cumbersome** (meaningful)
- Summaries are **cryptic** (precise)
- **Steep learning curve** even for us, computer scientists whereas we generally switch seamlessly from a language to another! That's frustrating! ;)

Do's and dont's

~~R is high level, I'll do everything myself~~

- CTAN comprises 4,334 \TeX , \LaTeX , and related packages and tools.
Most of you do not use plain \TeX .
- Currently, the CRAN package repository features 4,030 available packages.
- How do you know which one to use ??? Many of them are highly exotic (not to say useless to you).

I learnt with <http://www.r-bloggers.com/>

- Lots of introductions but not necessarily what you're looking for so I'll give you a short tour.
You should quickly realize though that you need proper training in statistics and data analysis if you do not want tell nonsense.
- Again, you should read Jain's book on **The Art of Computer Systems Performance Analysis**
- You may want to follow online courses:
 - <https://www.coursera.org/course/compdata>
 - <https://www.coursera.org/course/repdata>

Install and run R on debian

```
apt-cache search r
```

Err, that's not very useful :) It's the same when searching on google but once the filter bubble is set up, it gets better...

```
sudo apt-get install r-base
```

```
R
```

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"  
Copyright (C) 2013 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

```
>
```

Install a few cool packages

R has its own package management mechanism so just run R and type the following commands:

- ddply, reshape and ggplot2 by Hadley Wickham (<http://had.co.nz/>)

```
install.packages("plyr")
install.packages("reshape")
install.packages("ggplot2")
```

- knitr by (Yihui Xie) <http://yihui.name/knitr/>

```
install.packages("knitr")
```

IDE

Using R interactively is nice but quickly becomes painful so at some point, you'll want an IDE.

Emacs is great but you'll need *Emacs Speaks Statistics*

```
sudo apt-get install ess
```

In this tutorial, we will use **rstudio** (<https://www.rstudio.com/>).

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- **Reproducible Documents: knitR**
- Introduction to R
- Needful Packages by Hadley Wickam

Rstudio screenshot

RStudio screenshot showing the R Markdown interface, workspace, and console.

File menu: File, Edit, Code, View, Project, Workspace, Plots, Tools, Help.

Code Editor: Displays two files: `markdown-introduction.rmd` and `example-r-markdown.rmd`. The code in `example-r-markdown.rmd` includes R code for generating a scatter plot and its corresponding R Markdown output.

```
28 v ``{r basicconsole}
29 x <- 1:10
30 y <- round(rnorm(10, x, 1), 2)
31 df <- data.frame(x, y)
32 df
33
34
35
36 ## Plots
37 Images generated by 'knitr' are saved in a figures folder. However,
| they also appear to be represented in the HTML output using a [data
| URI scheme]( http://en.wikipedia.org/wiki/Data_URI_scheme). This
| means that you can paste the HTML into a blog post or discussion
| forum and you don't have to worry about finding a place to store the
| images; they're embedded in the HTML.
38
39 ### Simple plot
40 Here is a basic plot using base graphics:
41
42 ``{r simpleplot}
43 plot(x)
44
45
46 ``{r simpleplot}
47 plot(x)
```

Console: Shows the R session history.

```
> set.seed(1234)
> library(ggplot2)
> library(lattice)
> x <- 1:10
> y <- round(rnorm(10, x, 1), 2)
> df <- data.frame(x, y)
> df
   x   y
1 1 1.31
2 2 2.31
3 3 3.36
4 4 4.27
5 5 5.04
6 6 6.11
7 7 7.43
8 8 8.98
9 9 9.38
10 10 9.27
> plot(x)
```

Workspace: Shows the data frame `df` with 10 observations and 2 variables: `x` (integer[10]) and `y` (numeric[10]).

Plots: A scatter plot of `x` vs `y`, showing a positive linear trend.

x	y
1	1.31
2	2.31
3	3.36
4	4.27
5	5.04
6	6.11
7	7.43
8	8.98
9	9.38
10	9.27

Reproducible analysis in Markdown + R

- Create a new **R Markdown** document (Rmd) in rstudio
- R chunks are interspersed with “`{r}`” and “`
- Inline R code: ‘`r sin(2+2)`’
- You can **knit** the document and share it via **rpubs**
- R chunks can be sent to the top-level with Alt-Ctrl-c
- I usually work mostly with the current environment and only knit in the end
- Other engines can be used (use rstudio **completion**)

```
```{r engine='sh'}
ls /tmp/
```
```

- Makes **reproducible analysis as simple as one click**
- Great tool for quick analysis for self and colleagues, homeworks, ...

Reproducible articles with L^AT_EX + R

- Create a new R S_ewave document (Rnw) in rstudio
- R chunks are interspersed with <>>= and @
- You can knit the document to produce a pdf
- You'll probably quickly want to change default behavior (activate the cache, hide code, ...). In the preembule:

```
<<echo=FALSE>>=
opts_chunk$set(cache=TRUE,dpi=300,echo=FALSE,fig.width=7,
              warning=FALSE,message=FALSE)
```

@

- Great for journal articles, theses, books, ...

Outline

1 Reproducible Research

- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- Reproducible Documents: knitR
- **Introduction to R**
- Needful Packages by Hadley Wickam

Data frames

A data frame is a data tables (with columns and rows). `mtcars` is a built-in data frame that we will use in the sequel

```
head(mtcars);
```

| | | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---------|------------|------|-----|------|-----|------|-------|-------|----|----|------|------|
| Mazda | RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda | RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun | 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 4 |
| Hornet | 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 3 |
| Hornet | Sportabout | 18.7 | 8 | 360 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | | 18.1 | 6 | 225 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |

You can also load a data frame from a CSV file:

```
df <- read.csv("http://foo.org/mydata.csv", header=T,  
                strip.white=TRUE);
```

You will **get help** by using ?:

```
?data.frame
```

```
?rbind
```

```
?cbind
```

Exploring Content (1)

```
names(mtcars);  
[1] "mpg"   "cyl"   "disp"  "hp"    "drat"  "wt"    "qsec" "vs"    "am"  
[11] "carb"  
  
str(mtcars);  
  
'data.frame': 32 obs. of 11 variables:  
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
 $ cyl  : num  6 6 4 6 8 6 8 4 4 6 ...  
 $ disp: num  160 160 108 258 360 ...  
 $ hp   : num  110 110 93 110 175 105 245 62 95 123 ...  
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
 $ wt   : num  2.62 2.88 2.32 3.21 3.44 ...  
 $ qsec: num  16.5 17 18.6 19.4 17 ...  
 $ vs   : num  0 0 1 1 0 1 0 1 1 1 ...  
 $ am   : num  1 1 1 0 0 0 0 0 0 0 ...  
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...  
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

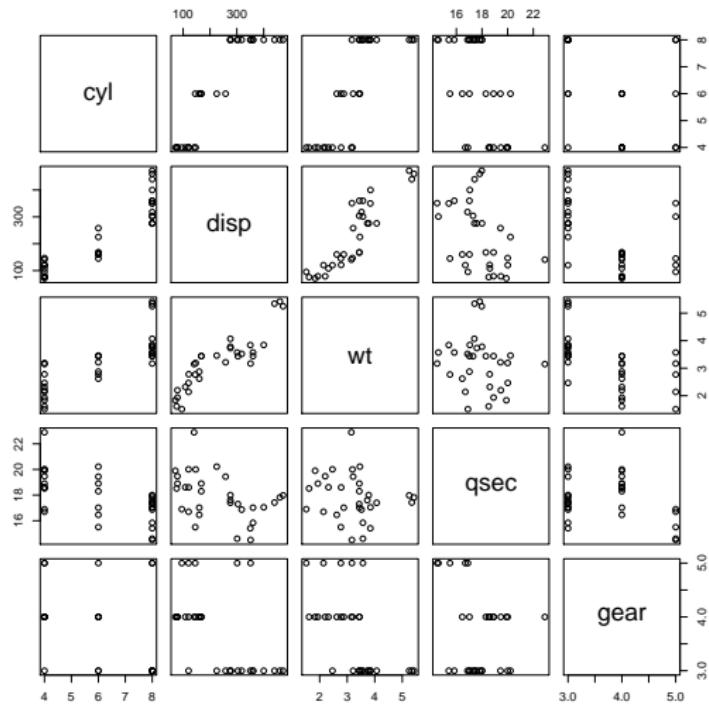
Exploring Content (2)

```
dim(mtcars);  
length(mtcars);  
  
[1] 32 11  
[1] 11  
  
summary(mtcars);
```

| mpg | cyl | disp | hp |
|---------------|---------------|---------------|----------------|
| Min. :10.40 | Min. :4.000 | Min. : 71.1 | Min. : 52.0 |
| 1st Qu.:15.43 | 1st Qu.:4.000 | 1st Qu.:120.8 | 1st Qu.: 96.5 |
| Median :19.20 | Median :6.000 | Median :196.3 | Median :123.0 |
| Mean :20.09 | Mean :6.188 | Mean :230.7 | Mean :146.7 |
| 3rd Qu.:22.80 | 3rd Qu.:8.000 | 3rd Qu.:326.0 | 3rd Qu.:180.0 |
| Max. :33.90 | Max. :8.000 | Max. :472.0 | Max. :335.0 |
| drat | wt | qsec | vs |
| Min. :2.760 | Min. :1.513 | Min. :14.50 | Min. :0.0000 |
| 1st Qu.:3.080 | 1st Qu.:2.581 | 1st Qu.:16.89 | 1st Qu.:0.0000 |
| Median :3.695 | Median :3.325 | Median :17.71 | Median :0.0000 |
| Mean :3.597 | Mean :3.217 | Mean :17.85 | Mean :0.4375 |
| 3rd Qu.:3.920 | 3rd Qu.:3.610 | 3rd Qu.:18.90 | 3rd Qu.:1.0000 |

Exploring Content (3)

```
plot(mtcars[names(mtcars) %in% c("cyl","wt","disp","qsec","gear")])
```



Accessing Content

```
mtcars$mpg
```

```
[1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17  
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30  
[31] 15.0 21.4
```

```
mtcars[2:5,]$mpg
```

```
[1] 21.0 22.8 21.4 18.7
```

```
mtcars[mtcars$mpg == 21.0,]
```

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---------------|-----|-----|------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4 | 21 | 6 | 160 | 110 | 3.9 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21 | 6 | 160 | 110 | 3.9 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |

```
mtcars[mtcars$mpg == 21.0 & mtcars$wt > 2.7,]
```

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---------------|-----|-----|------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4 Wag | 21 | 6 | 160 | 110 | 3.9 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |

Extending Content

```
mtcars$cost = log(mtcars$hp)*atan(mtcars$disp)/  
  sqrt(mtcars$gear**5);
```

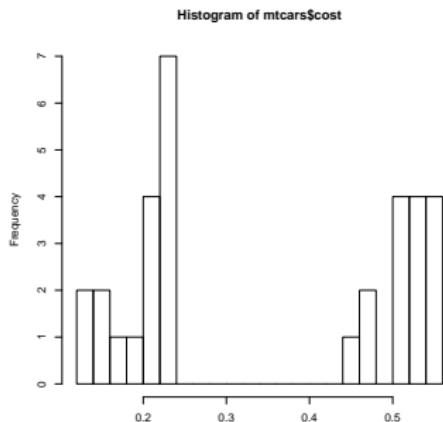
```
mean(mtcars$cost);
```

```
summary(mtcars$cost);
```

```
[1] 0.345994
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|--------|---------|--------|--------|---------|--------|
| 0.1261 | 0.2038 | 0.2353 | 0.3460 | 0.5202 | 0.5534 |

```
hist(mtcars$cost, breaks=20);
```



Outline

1 Reproducible Research

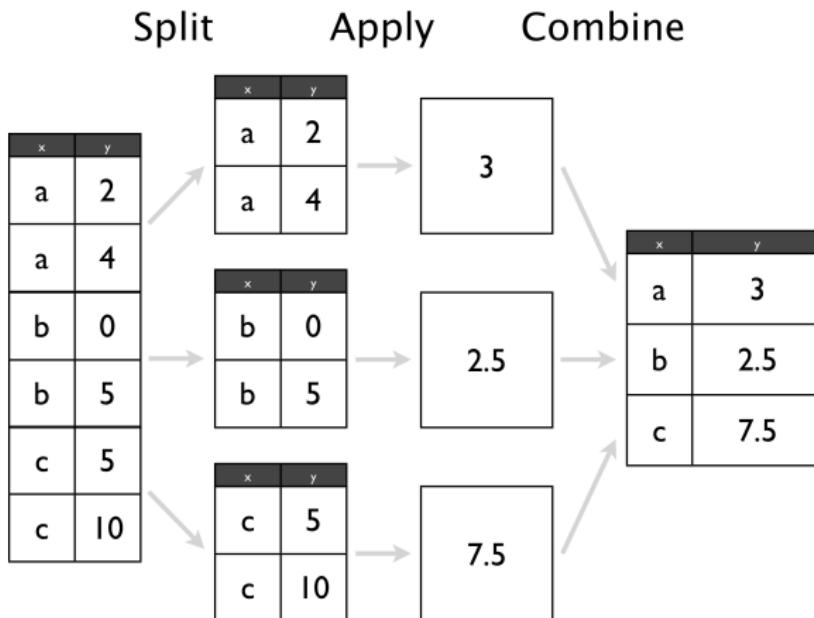
- Looks familiar ?
- Many Different Alternatives

2 R

- General Introduction
- Reproducible Documents: knitR
- Introduction to R
- Needful Packages by Hadley Wickam

plyr: the Split-Apply-Combine Strategy

Have a look at <http://plyr.had.co.nz/09-user/> for a more detailed introduction.



plyr: Powerful One-liners

```
library(plyr)
mtcars_summarized = ddply(mtcars,c("cyl","carb"), summarize,
    num = length(wt), wt_mean = mean(wt), wt_sd = sd(wt),
    qsec_mean = mean(qsec), qsec_sd = sd(qsec));
mtcars_summarized
```

| | cyl | carb | num | wt_mean | wt_sd | qsec_mean | qsec_sd |
|---|-----|------|-----|----------|-----------|-----------|-----------|
| 1 | 4 | 1 | 5 | 2.151000 | 0.2627118 | 19.37800 | 0.6121029 |
| 2 | 4 | 2 | 6 | 2.398000 | 0.7485412 | 18.93667 | 2.2924368 |
| 3 | 6 | 1 | 2 | 3.337500 | 0.1732412 | 19.83000 | 0.5515433 |
| 4 | 6 | 4 | 4 | 3.093750 | 0.4131460 | 17.67000 | 1.1249296 |
| 5 | 6 | 6 | 1 | 2.770000 | NA | 15.50000 | NA |
| 6 | 8 | 2 | 4 | 3.560000 | 0.1939502 | 17.06000 | 0.1783255 |
| 7 | 8 | 3 | 3 | 3.860000 | 0.1835756 | 17.66667 | 0.3055050 |
| 8 | 8 | 4 | 6 | 4.433167 | 1.0171431 | 16.49500 | 1.4424112 |
| 9 | 8 | 8 | 1 | 3.570000 | NA | 14.60000 | NA |

If your data is not in the right form **give a try to reshapeP/melt.**

ggplot2: Modularity in Action

- ggplot2 builds on plyr and on a modular **grammar of graphics**
- obnoxious function with dozens of arguments
- combine small functions using layers and transformations
- aesthetic mapping between **observation characteristics** (data frame column names) and **graphical object variables**
- an incredible **documentation**: <http://docs.ggplot2.org/current/>

The screenshot shows a web browser window with two tabs open. The left tab displays the 'Help topics' section of the ggplot2 documentation, listing various geom types like geom_point, geom_line, and geom_bar, along with their descriptions and icons. The right tab shows a detailed view of the 'geom_point' documentation, featuring a scatter plot of mpg vs wt from the mtcars dataset. The plot uses color to map to cyl (cylinder count), showing a clear negative correlation. Below the plot, the R code for creating it is shown, along with a note about the deprecation of scale_area.

Dependencies

- Depends: stats, methods
- Imports: plyr, digest, grid, gtable, reshape2, scales, gridExtra, MASS
- Suggests: quantreg, Hmisc, maptree, maps, hddim, maptools, multcomp, rpart, testthat
- Extends

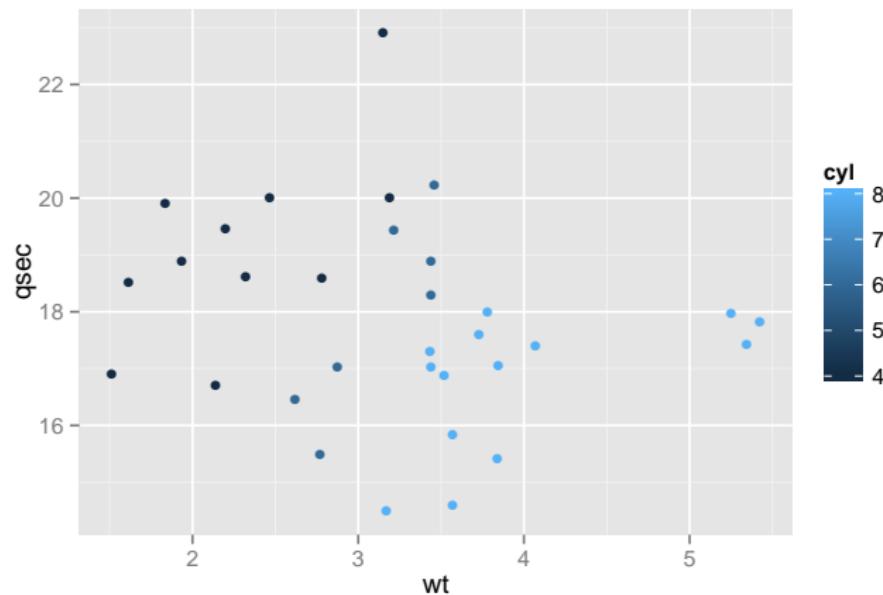
p + geom_point(aes(size = qsec)) + scale_area()

scale_area is deprecated. Use scale_size_area instead.
Note that the behavior of scale_size_area is slightly different by default: it makes the areas proportional to the numeric value. (Deprecated: last used in version 0.9.2)

44 / 52

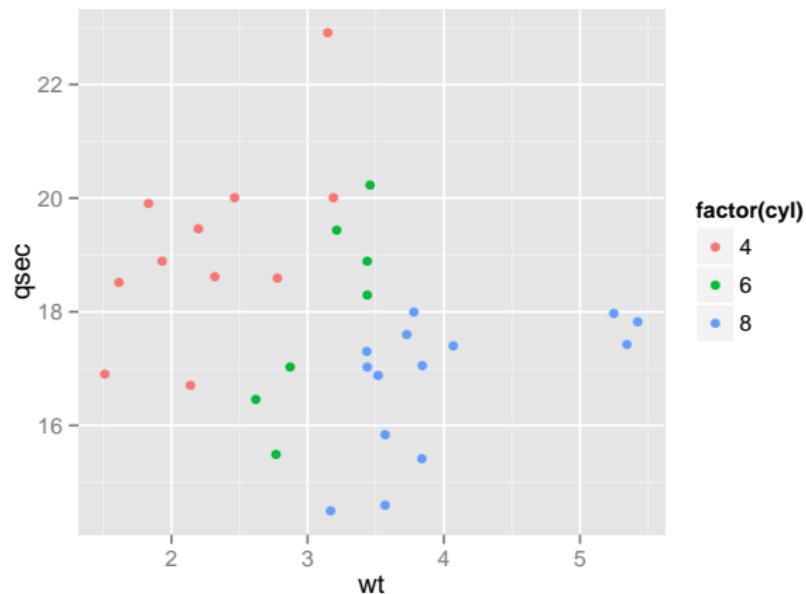
ggplot2: Illustration (1)

```
ggplot(data = mtcars, aes(x=wt, y=qsec, color=cyl)) +  
  geom_point();
```



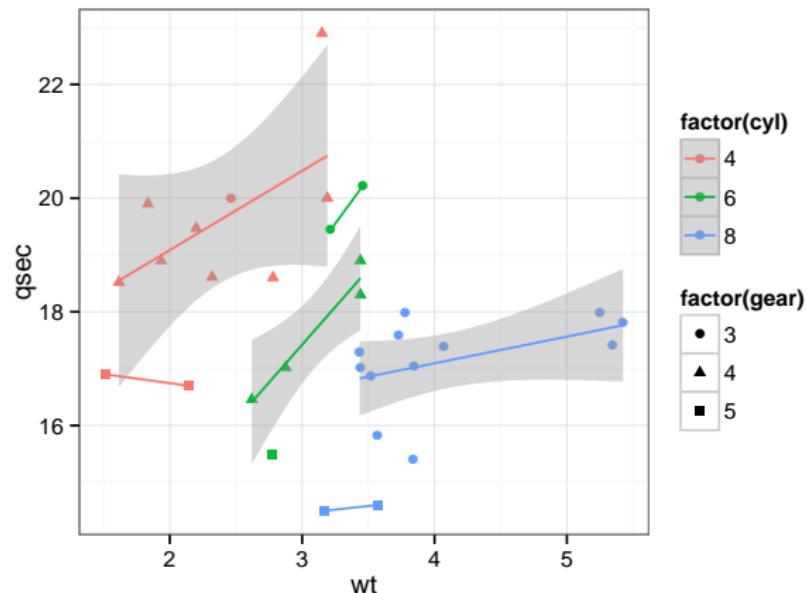
ggplot2: Illustration (2)

```
ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl))) +  
  geom_point();
```



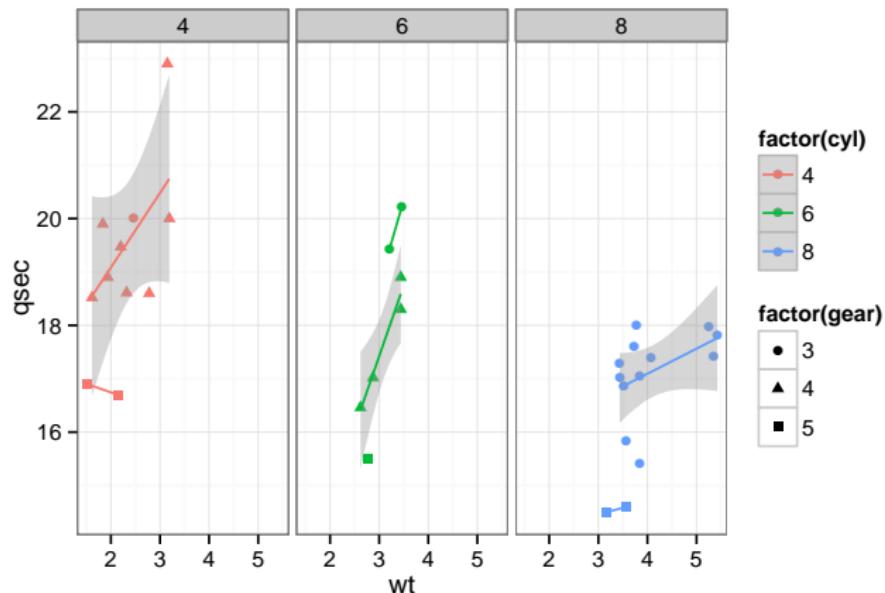
ggplot2: Illustration (3)

```
ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl),  
    shape = factor(gear))) +  geom_point() + theme_bw() +  
    geom_smooth(method="lm");
```



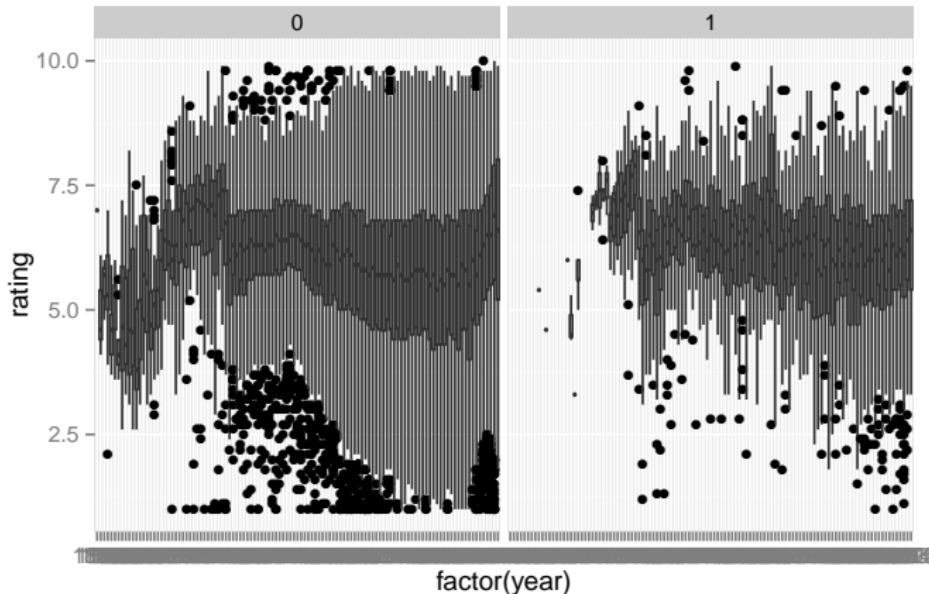
ggplot2: Illustration (4)

```
ggplot(data = mtcars, aes(x=wt, y=qsec, color=factor(cyl),  
    shape = factor(gear))) + geom_point() + theme_bw() +  
    geom_smooth(method="lm") + facet_wrap(~ cyl);
```



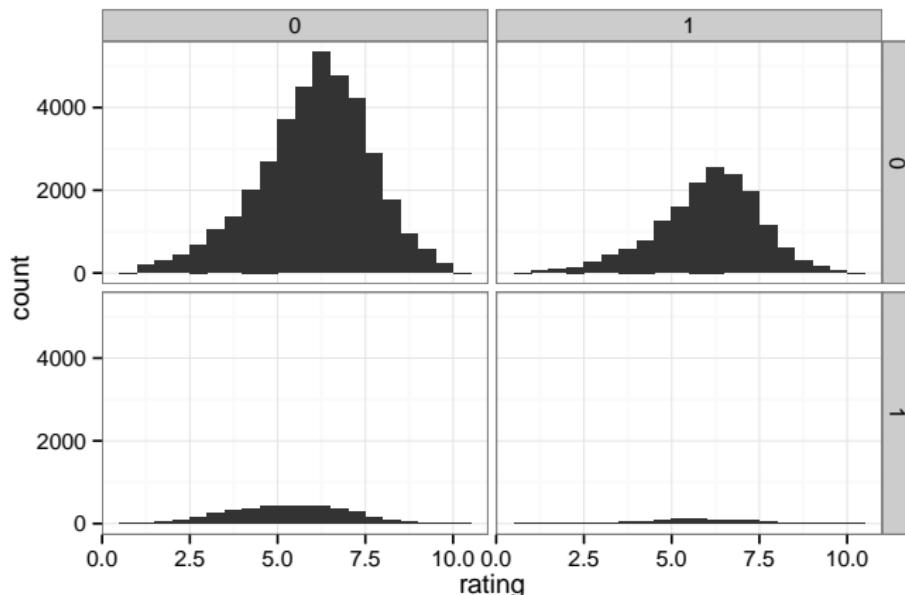
ggplot2: Illustration (5)

```
ggplot(data = movies, aes(x=factor(year),y=rating)) +  
  geom_boxplot() + facet_wrap(~Romance)
```



ggplot2: Illustration (6)

```
ggplot(movies, aes(x = rating)) + geom_histogram(binwidth = 0.5) +  
  facet_grid(Action ~ Comedy) + theme_bw();
```



Take away Message

- R is a great tool but is only a tool. There is no magic. You need to understand what you are doing and get a minimal training in statistics.
- It is one of the building block of **reproducible research** (the *reproducible analysis* block) and **will save you a lot of time**.
- Read at least Jain's book: *The Art of Computer Systems Performance Analysis*.
- Jean-Marc Vincent and myself give a **set of tutorials on performance evaluation** in M2R:

[http://mescal.imag.fr/membres/arnaud.legrand/teaching/2013/
M2R_EP.php](http://mescal.imag.fr/membres/arnaud.legrand/teaching/2013/M2R_EP.php)

- There are interesting **online courses** on coursera
 - <https://www.coursera.org/course/compdata>
 - <https://www.coursera.org/course/repdata>

About these slides

They have been composed in org-mode and generated with emacs, beamer, and pygment/pygments for the pretty printing.