

Modèles Statistiques (MS)

ou comment réaliser une étude en incluant de l'analyse de données

Jean-Marc.Vincent@imag.fr
Lucas.Schnorr@inf.ufrgs.br

Laboratoire LIG
Équipe-Projet Inria POLARIS

Grenoble 2017

UE MODÈLES STATISTIQUES

- 1 **ORGANISATION DE L'UE : Modèles Statistiques**
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

ORGANISATION

Équipe pédagogique

Jean-Marc Vincent



Jean-Marc.Vincent@imag.fr
coordination de l'UE
Laboratoire d'informatique de Grenoble
Équipe Inria POLARIS
Évaluation de performances de
systèmes/réseaux/infrastructures à
grande échelle

Lucas Schnorr



schnorr@inf.ufrgs.br
Laboratoire d'informatique de Grenoble
Équipe Inria POLARIS
INF, UFRGS (Porto Alegre, Brésil)
Trace d'applications, analyse de
programmes parallèles, environnement
d'expérimentation

COMMUNICATION AVEC L'ÉQUIPE PÉDAGOGIQUE

Mail et adresses électroniques

Adresse Mail enseignant : Prénom.Nom@imag.fr

SUJET : [MIAGE :MS] sujet explicite

envoyer votre mail avec votre adresse officielle **@etu.univ-grenoble-alpes.fr**
toute adresse de provenance différente risque d'être "grey/black-listée" et d'atterrir dans une poubelle

le mail officiel de la L3-MIAGE est la liste

etu-2016-im2ag-gbl3ie160@univ-grenoble-alpes.fr, toute annonce officielle (quicks, apnées, déplacements de créneaux horaires,...) passera par ce mail (que vous devez lire quotidiennement)

Destinataires

organisation/cours/examens... : Jean-Marc Vincent

les **Travaux Dirigés/Pratiques** : Lucas Schnorr

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE**
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

OBJECTIF PÉDAGOGIQUE DE L'UE MODÈLES STATISTIQUES

Connaissances

Savoir réaliser une étude d'un objet informatique (ou autre) à partir de données observées :

(répondre à une question, formuler une hypothèse et la confirmer)

- ▶ savoir bâtir une expérimentation simple et produire des données d'observation
- ▶ savoir analyser les résultats obtenus (processus d'analyse)
- ▶ savoir restituer les résultats sous forme synthétique (processus de visualisation, commentaires, analyse et synthèse)

En pratique, savoir réaliser une étude argumentée et correctement présentée.

Savoir utiliser un/des environnement(s) adapté(s) :

- ▶ suivi des développements logiciels (historique, versionning, collaboration) : git, github
- ▶ processus d'analyse (analyse statistique, synthèse, visualisation) : R (R-studio, ggplot2)
- ▶ mise en forme et présentation : LaTeX (via un markdown)

ORGANISATION DE LA SEMAINE

Cours /TD : **guidelines** pour une étude rigoureuse et reproductible

Les cours/TD seront organisés à partir d'études de cas :

- ▶ une partie synthétique sur les **concepts**
- ▶ une partie sur des **exemples** illustrant les concepts
- ▶ une partie sur votre étude de cas

Forme du travail

- ▶ travail en binôme/quadrinôme
- ▶ travail public (partageable par toute la promotion (et même plus))
- ▶ synthèse en commun (production de fiches)

Travail personnel :

- ▶ prévoir 1 à 2h de travail en moyenne à la maison pour 1 séance de cours/TD ,
- ▶ exercices à la maison (pour préparer le matériel des séances suivantes)

Évaluation : une note d'UE

- ▶ mini-projet avec une présentation
- ▶ suivi des C/TD

CONTENU INDICATIF

Environnement

- ➊ Introduction / problème
- ➋ Programmation Littérale / RStudio / Rmd
- ➌ Visualisation élémentaire / ggplot2
- ➍ Guideline / Checklist for good graphics

Traitement de données

- ➎ Processus d'analyse /statistiques de base (rappels)
- ➏ Données importation pré et post traitement
- ➐ Manipulation de donnée expérimentales (dplyr)

Mini-projet

- ➑ Mini-projet : spécification/études préliminaires
- ➒ Mini-projet : étude et rapport

Présentation

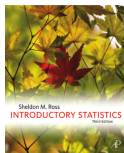
- ➓ Présentation orale (tous les étudiants)

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES**
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

BIBLIOGRAPHIE : OUVRAGES DE RÉFÉRENCE DU COURS

- ▶ **R** Garrett Golemund and Hadley Wickham, R for Data Science, O'Reilly 2016
<http://r4ds.had.co.nz/>
- ▶ **Statistiques** Sheldon Ross Introductory Statistics. Academic Press 2010
Également les polycopiés de Frédérique Leblanc
<http://www-ljk.imag.fr/membres/Frederique.Leblanc/>
- ▶ **Historique** Donald E. Knuth Literate Programming. Academic Press 1983



et évidemment de nombreuses ressources sur le web ...

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES**
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

UN OEIL CRITIQUE

Dernière mise à jour le dimanche 22 janvier 2017 à 00h45



Manuel Valls

31.11%

188603 voix



Sylvia Pinel

1.97%

24657 voix



Vincent Peillon

6.85%

85975 voix



François de Rugy

3.88%

48921 voix



Arnaud Montebourg

17.52%

218885 voix



Benoît Hamon

36.35%

454041 voix



Jean-Luc Bennahmias

1.01%

12609 voix

Votes blancs et nuls

1.3%

16235 voix

UN OEIL CRITIQUE

Dernière mise à jour le dimanche 22 janvier 2017 à 00h45



Manuel Valls

31.11%

108605 voix



Sylvia Pinel

1.97%

24657 voix



Vincent Peillon

6.85%

85975 voix



François de Rugy

3.88%

48921 voix



Arnaud Montebourg

17.52%

218885 voix



Benoît Hamon

36.35%

454041 voix



Jean-Luc Bennahmias

1.01%

12609 voix

Votes blancs et nuls

1.3%

16235 voix



Manuel Valls

31.11%

498114 voix



Sylvia Pinel

1.98%

37703 voix



Vincent Peillon

6.85%

109678 voix



François de Rugy

3.88%

62124 voix



Arnaud Montebourg

17.52%

280519 voix



Benoît Hamon

36.35%

582014 voix



Jean-Luc Bennahmias

1.01%

16172 voix

Votes blancs et nuls

1.3%

20815 voix

et lundi matin

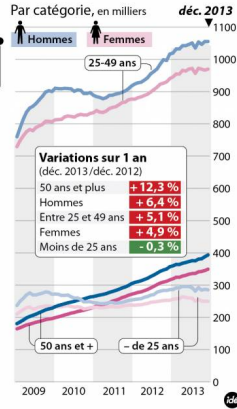
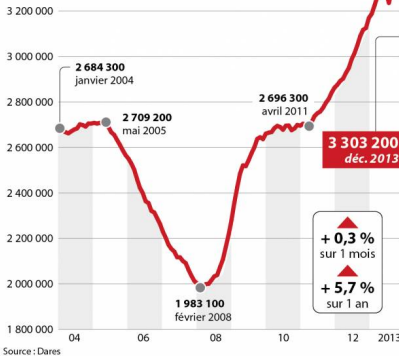
UN OEIL CRITIQUE (2)

- ▶ Performances <http://www.cpubenchmark.net/index.php>
- ▶ <http://www.tylervigen.com/spurious-correlations>
- ▶ <https://www.ined.fr/fr/tout-savoir-population/>
- ▶ taux de réussite au Bac

UN OEIL CRITIQUE (3)

Le chômage

Nombre de demandeurs d'emploi (catégorie A)



Extrait du journal Le Point 2013

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE**
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

FRUSTRATION 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
- ▶ My advisor asked me whether I took care of setting this or this but I can't remember
- ▶ The damned fourth reviewer asked for a major revision and wants me to change figure 3 :(
- ▶ Which code and which data set did I use to generate this figure ?
- ▶ It worked yesterday !
- ▶ 6 months later : why did I do that ?

FRUSTRATION AS A REVIEWER

This may be an interesting contribution but :

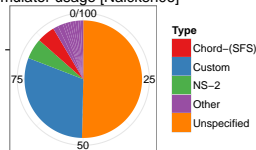
- ▶ This **average value** must hide something
- ▶ As usual, there is no **confidence interval**, I wonder about the variability and whether the difference is **significant** or not
- ▶ That can't be true, I'm sure they **removed some points**
- ▶ Why is this graph in **logscale** ? How would it look 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

A FEW EDIFYING EXAMPLES

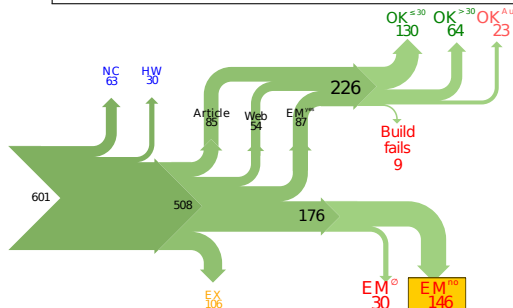
Naicken, Stephen *et Al.*, *Towards Yet Another Peer-to-Peer Simulator*, HET-NETs'06.

From 141 P2P sim.papers, 30% use a custom tool, 50% don't report used tool

Simulator usage [Naicken06]



Collberg, Christian *et Al.*, *Measuring Reproducibility in Computer Systems Research*, <http://reproducibility.cs.arizona.edu/>



- 8 ACM conferences (ASPLOS'12, CCS'12, OOPSLA'12, OSDI'12, PLDI'12, SIGMOD'12, SOSP'11, VLDB'12) and 5 journals
- EM^{no} = the code cannot be provided

THE DOG ATE MY HOMEWORK !!!

► Versioning Problems

*Thanks for your interest in the implementation of our paper. The good news is that I was able to find some code. I am just **hoping** that **it** is a stable working version of the code, and **matches the implementation we finally used for the paper**. Unfortunately, I have **lost some data** when **my laptop was stolen** last year. The bad news is that the code is not commented and/or clean.*

*Attached is the `<system>` source code of our algorithm. I'm **not** very **sure whether it is the final version of the code used in our paper**, but it should be at least 99% close. Hope it will help.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices

*Unfortunately, the server in which my implementation was stored had a **disk crash in April and three disks crashed simultaneously**. While the help desk made significant effort to save the data, my entire implementation for this paper was not found.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon

*Unfortunately the current system is **not mature enough at the moment**, so it's not yet publicly available. We are actively working on a number of extensions and **things are somewhat volatile**. However, once things stabilize we plan to release it to outside users. At that point, we would be happy to send you a copy.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release

*I am afraid that the source code was never released. The code was **never intended to be released so is not in any shape for general use.***

THE DOG ATE MY HOMEWORK!!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left

*⟨STUDENT⟩ was a graduate student in our program but **he left a while back** so I am responding instead. For the paper we used a prototype that included many moving pieces that only ⟨STUDENT⟩ knew how to operate and we did not have the time to integrate them in a ready-to-share implementation before he left. Still, I hope you can build on the ideas/technique of the paper.*

*Unfortunately, the author who has done most of the coding for this paper has **passed away** and the code is no longer maintained.*

THE DOG ATE MY HOMEWORK!!!

- ▶ Versioning Problems
 - ▶ Bad Backup Practices
 - ▶ Code Will be Available Soon
 - ▶ No Intention to Release
 - ▶ Programmer Left
- ▶ Commercial Code

*Since this work has been done at (COMPANY) **we don't open-source code** unless there is a compelling business reason to do so. So unfortunately I don't think we'll be able to share it with you.*

*The code **owned by (COMPANY)**, and AFAIK the code is not open-source. Your best bet is to reimplement :(Sorry.*

THE DOG ATE MY HOMEWORK!!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left
- ▶ Commercial Code
- ▶ Proprietary Academic Code

*Unfortunately, the $\langle \text{SYSTEM} \rangle$ sources are *not meant to be opensource* (the code is partially *property of* $\langle \text{UNIVERSITY 1} \rangle$, $\langle \text{UNIVERSITY 2} \rangle$ and $\langle \text{UNIVERSITY 3} \rangle$.)*

If this will change I will let you know, albeit I do not think there is an intention to make the $\langle \text{SYSTEM} \rangle$ sources opensource in the near future.

*If you're interested in obtaining the code, *we only ask for a description of the research project* that the code will be used in (*which may lead to some joint research*), and we also have a software license agreement that the University would need to sign.*

THE DOG ATE MY HOMEWORK !!!

- ▶ Versioning Problems
- ▶ Bad Backup Practices
- ▶ Code Will be Available Soon
- ▶ No Intention to Release
- ▶ Programmer Left
- ▶ Commercial Code
- ▶ Proprietary Academic Code
- ▶ **Research vs. Sharing**
- ▶ ...
- ▶ ...

In the past when we attempted to share it, we found ourselves spending more time getting outsiders up to speed than on our own research. So I finally had to establish the policy that we will not provide the source code outside the group.

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE**
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 R/KNITR CRASH COURSE Thanks to GitHub SMPE
 - General Introduction
 - Reproducible Documents : knitr

STRUCTURE

Research articles are often structured in this basic order :

Introduction Why was the study undertaken ? What was the research question, the tested hypothesis or the purpose of the research ?

Methods When, where, and how was the study done ? What materials/hardware were used ? How was it configured ?

Results What answer was found to the research question ; what did the study find ? Was the tested hypothesis true ? **Present useful results in a synthetic way with a logical order.**

Discussion What might the answer imply and why does it matter ? How does it fit in with what other researchers have found ? What are the possible bias and points to improve ? What are the perspectives for future research ?

Such structure **facilitates literature review** and is a very effective way to convey information.

If the report is a few pages long then **an abstract is required.**

STEP 0 : TAKING NOTES

Document your :

- ▶ **Hypotheses** : keep track of your ideas/line of thoughts
- ▶ **Experiments** : details on how and why an experiment was run, including failed or ambiguous attempts.
- ▶ **Initial analysis or interpretation** of these experiments : was the outcome conform to the expectation or not ? does it (in)validate the hypothesis ?
- ▶ **Organization** : keep track of things to do/fix/test/improve

Structure :

- 1 General information about the document and organization **conventions** (e.g., directory structure, notebook structure, experimental result storing mechanism, ...)
- 2 Documentation of **commonly used commands** and of how to set up experiments (e.g., git cloning, environment deployment, connection to machines, compiling scripts)
- 3 Experiment results can be either structured **by dates** (↪ add tags) or **by experiment campaigns** (↪ add date/time)

WHICH FORMAT SHOULD I USE ?

- ▶ Wikis are encouraged to favor collaboration but I do not find them really effective
- ▶ Blogging systems are also a way of managing such notebook but they should rather be considered as an effective way to share information with others
- ▶ I recommend to use basic plain-text format and to structure it hierarchically

Here is a [link](#) to an excerpt of the journal of one of my PhD student, managed with git/org-mode. More detailed links are given in slide ??.

Last but not least :

Provide links to Raw Data!!!

WHEN/HOW OFTEN SHOULD I USE IT ?

I have a very intense usage (demo to [general journal](#) and specific [BOINC journal](#)) and I tend to capture a lot of information but you do not have to be as extreme as I am. Here are a few advices :

- ▶ Spending **more than an hour without** at least **writing** what you're working on **is not right** . .
 - **Take a 5 minutes** break and ask yourself what you're doing, what is keeping you busy and where all this is leading you
- ▶ While working on something, you will often notice/think about something you should fix/improve but you just don't want to do it now. Take 20 seconds to write a **TODO** entry.
- ▶ There are moments where you have to **wait for something** (compiling, deployment, . . .). It is generally the perfect time for improving your notes (e.g., detail the steps to accomplish a TODO entry).
- ▶ **By the end of the day** : daily (and weekly) **review** !
 - Update your lists, write what the next steps are
 - **Summarize in a 2-4 lines** (for your advisor) what you did, what was difficult, what you learnt.

STEP 1 : SHARING CODE AND DATA

What kinds of systems are available ?

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

Depends on the level of privacy you expect but you probably already know these tools.

Few handle GB files...

Is this enough ?

- 1 Use a workflow that **documents both data and process**
- 2 Use the machine readable **CSV format**
- 3 Provide **raw** data and **meta** data, not just statistical outputs
- 4 **Never** do data manipulation and statistical tests **by hand**
- 5 **Use R**, Python or another free software to read and process raw data (**ideally to produce complete reports** with code, results and prose)

STEP 2 : 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 even with sessions

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 (e.g., Rstudio)

UE MODÈLES STATISTIQUES

- 1 ORGANISATION DE L'UE : Modèles Statistiques
- 2 OBJECTIF DE L'UE
- 3 RÉFÉRENCES BIBLIOGRAPHIQUES
- 4 PROBLÉMATIQUE : QUELQUES EXEMPLES
- 5 REPRODUCTIBILITÉ : MOTIVATION Thanks to GitHub SMPE
- 6 REPORTING Thanks to GitHub SMPE
 - An IMRAD Report
 - Good Practice for Setting up a Laboratory Notebook
- 7 **R/KNITR CRASH COURSE Thanks to GitHub SMPE**
 - General Introduction
 - Reproducible Documents : knitr

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 SUCH 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 T_EX, L^AT_EX, and related packages and tools. Most of you do not use plain T_EX.
- ▶ 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 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 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 it's 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")  
# or better: install.packages("dplyr")  
install.packages("reshape")  
# or better; install.packages("tidyr")  
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, I will briefly show you **rstudio** (<https://www.rstudio.com/>) and later how to use **org-mode**

RStudio interface showing a Knitr R Markdown document being rendered.

Source Editor (example-r-markdown.md):

```

28
29- ```{r basicconsole}
30 x <- 1:10
31 y <- round(rnorm(10, x, 1), 2)
32 df <- data.frame(x, y)
33 df
34 ```
35
36 ## Plots
37 Images generated by 'knitr' are saved in a figures folder. However,
38 | they also appear to be represented in the HTML output using a [data
39 | URI scheme]( http://en.wikipedia.org/wiki/Data_URI_scheme). This
40 | means that you can paste the HTML into a blog post or discussion
41 | forum and you don't have to worry about finding a place to store the
42 | images; they're embedded in the HTML.
43
44 ### Simple plot
45 Here is a basic plot using base graphics:
46
47 ```{r simpleplot}
48 plot(x)
49 ```
50
51 ```{r simpleplot}
52 plot(x)
53 ```

```

Environment:

Data	df
10 obs. of 2 variables	
Values	
x	integer[10]
y	numeric[10]

Console:

```

~/research/statistics/rmarkdown-meetup-2012/ >
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 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 3.27
5 5 5.04
6 6 6.11
7 7 8.43
8 8 8.98
9 9 8.78

```

Plots:

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 Sweave** 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 preamble :

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

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