Reproducible data science to inform outbreak response

Report from the North-Kivu Ebola outbreak

Dr Thibaut Jombart (@TeebzR)
London School of Hygiene and Tropical Medicine
Imperial College London
R Epidemics Consortium (RECON)

UserR!2019
Toulouse
11 July 2019

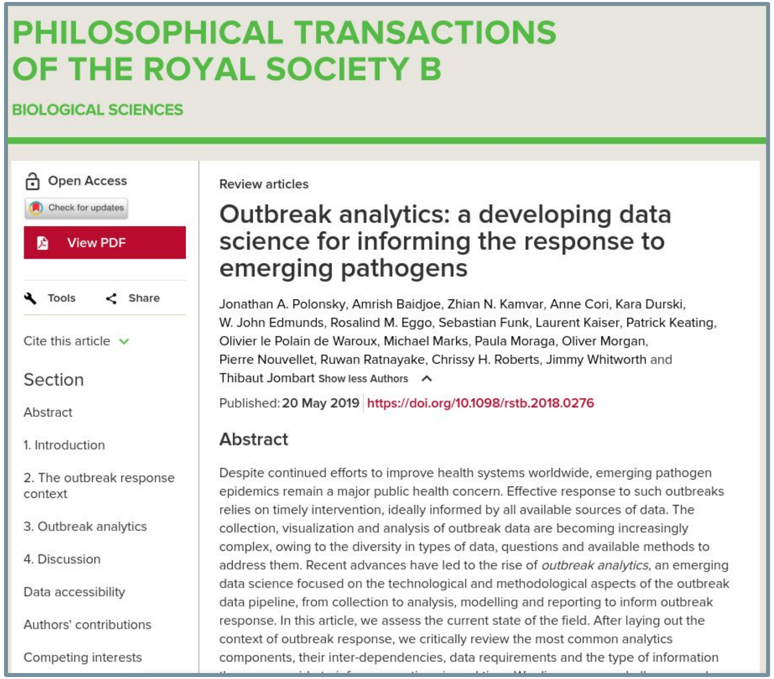
Context

Outbreak analytics

RECON

Ebola in North Kivu

On the emergence of "outbreak analytics"



PHILOSOPHICAL TRANSACTIONS
OF THE ROYAL SOCIETY B

BIOLOGICAL SCIENCES

Open Access

Section

Abstract

1. Introduction

2. The outbreak response context

3. Outbreak analytics

4. Discussion

Data accessibility

Authors' contributions

Competing interests

Review articles

Outbreak analytics: a developing data science for informing the response to emerging pathogens

Jonathan A. Polonsky, Amrish Baidjoe, Zhian N. Kamvar, Anne Cori, Kara Durski, W. John Edmunds, Rosalind M. Eggo, Sebastian Funk, Laurent Kaiser, Patrick Keating, Olivier le Polain de Waroux, Michael Marks, Paula Moraga, Oliver Morgan, Pierre Nouvellet, Ruwan Ratnayake, Chrissy H. Roberts, Jimmy Whitworth and Thibaut Jombart Show less Authors

Published: 20 May 2019 | https://doi.org/10.1098/rstb.2018.0276

Abstract

Despite continued efforts to improve health systems worldwide, emerging pathogen epidemics remain a major public health concern. Effective response to such outbreaks relies on timely intervention, ideally informed by all available sources of data. The collection, visualization and analysis of outbreak data are becoming increasingly complex, owing to the diversity in types of data, questions and available methods to address them. Recent advances have led to the rise of *outbreak analytics*, an emerging data science focused on the technological and methodological aspects of the outbreak data pipeline, from collection to analysis, modelling and reporting to inform outbreak response. In this article, we assess the current state of the field. After laying out the context of outbreak response, we critically review the most common analytics components, their inter-dependencies, data requirements and the type of information

https://doi.org/10.1098/rstb.2018.0276

- DoB: Polonsky et al. (2019) Phil. Trans. R. Soc. B 374
- Data science mixing statistics, mathematical modeling, computer simulations, database infrastructure, GIS, genetics, software engineering
- At the crossroad of public health institutions, private sector, and academia
- Aims to inform response to emergencies in real-time
- **Lack of available tools**

RECON: bringing data science into health emergencies

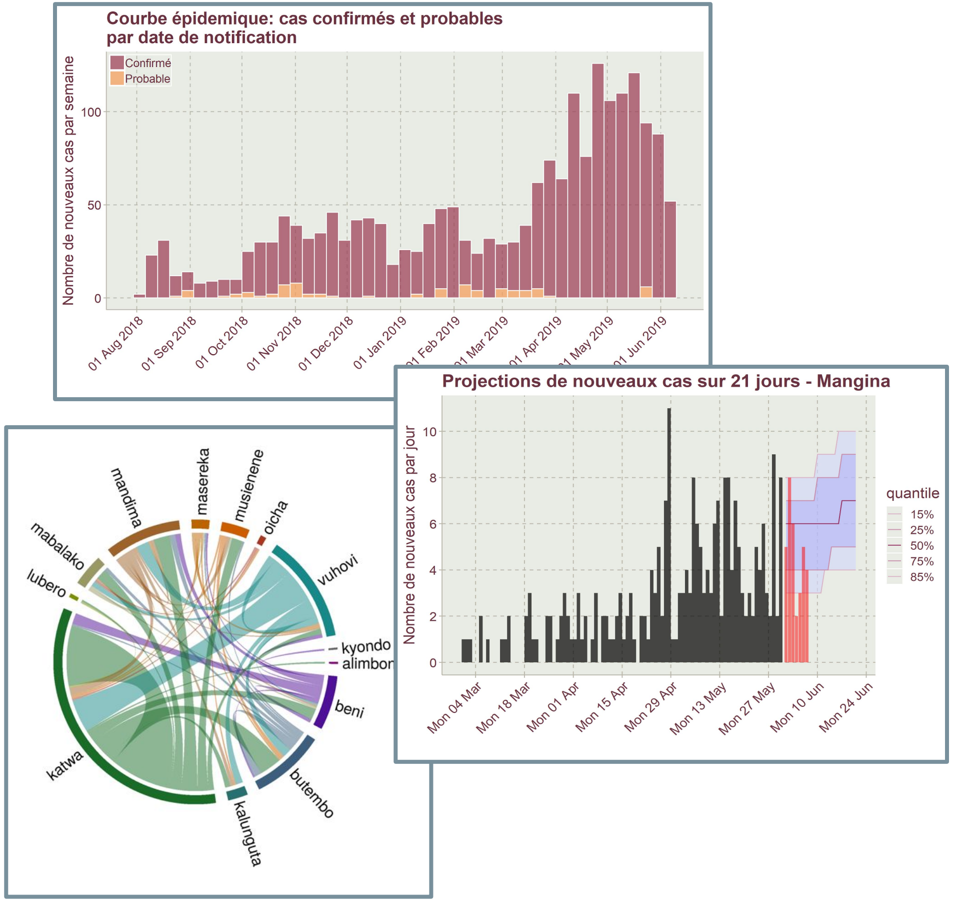
https://www.repidemicsconsortium.org/
https://www.reconlearn.org/

- Origin: Hackout 3 (rOpenSci / Imperial College London), Berkeley, 2016
- NGO for open analytics resources for health emergencies and humanitarian crises
- ~35 members, 200-300 subscribers
- Packages: 10 on CRAN, 15-20 in development
- Events: short courses, workshops, hackathons
- **Deployments** to support response to emergencies

Ebola in North-Kivu & Ituri, DRC

- Largest Ebola epidemic in DRC, 2nd largest in the world
- August 2018 - today:
  - >2400 cases (confirmed / probable)
  - 67% deaths
- Difficulties due to military conflicts
  - Threats to local population
  - Threats to response staff and facilities
- **First deployment of an analytical cell** as part of the Emergency Operations Centre

Outbreak analytics cell: aims and challenges



- **Multiple (messy) data sources**, no global database
- **Independent updates** of different databases
- **Needs**: data cleaning, visualisation, in-depth analyses, forecasting
- **Routine** versus ***ad-hoc*** analyses
- Need for **regular results updates** and **traceability**
- Bad internet, different platforms, low R literacy

An analysis infrastructure for the response to Ebola

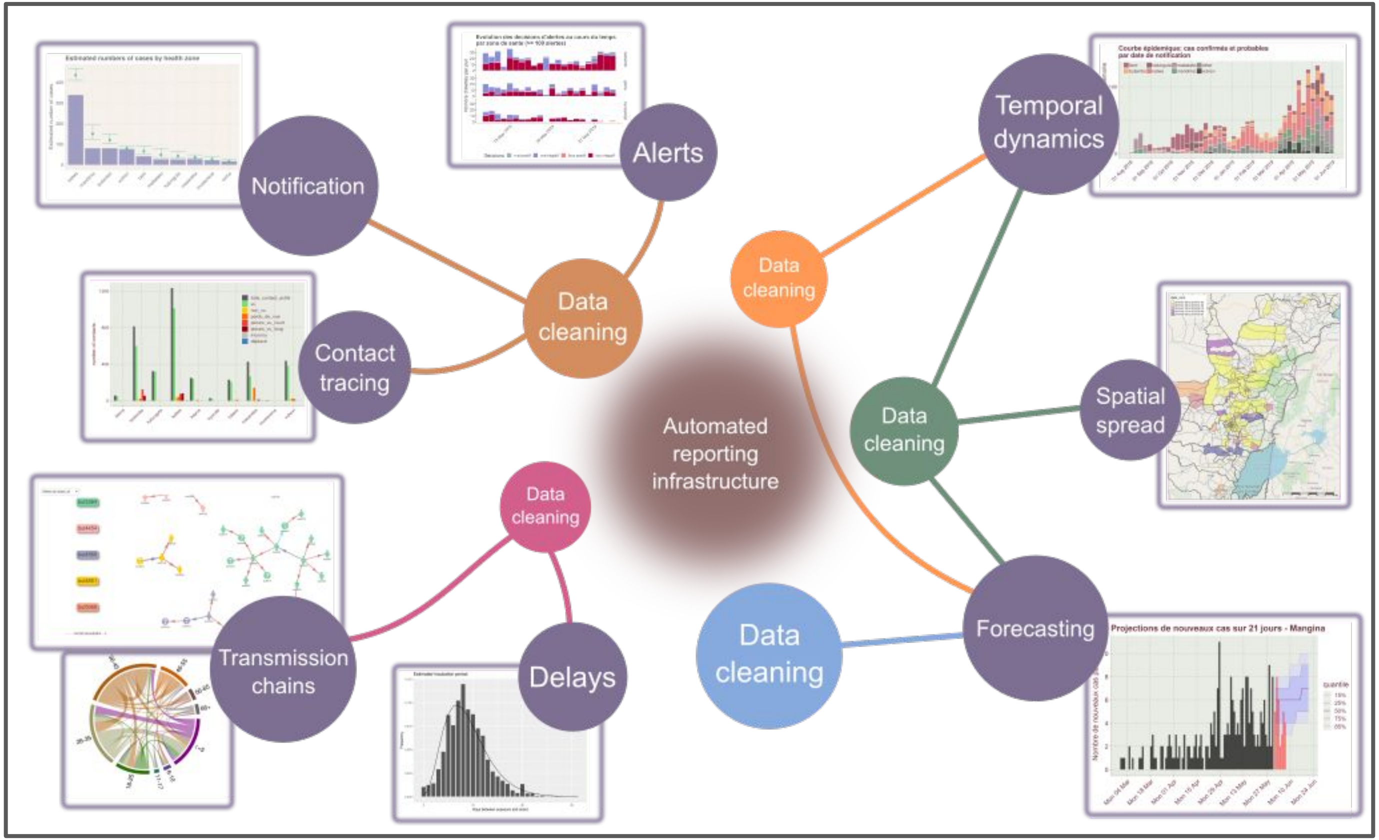
Data cleaning using *linelist*

Tidier markdown workflows with *reportfactory*

Taking R offline: RECON deployer

Illustrations

Overview of the analysis infrastructure

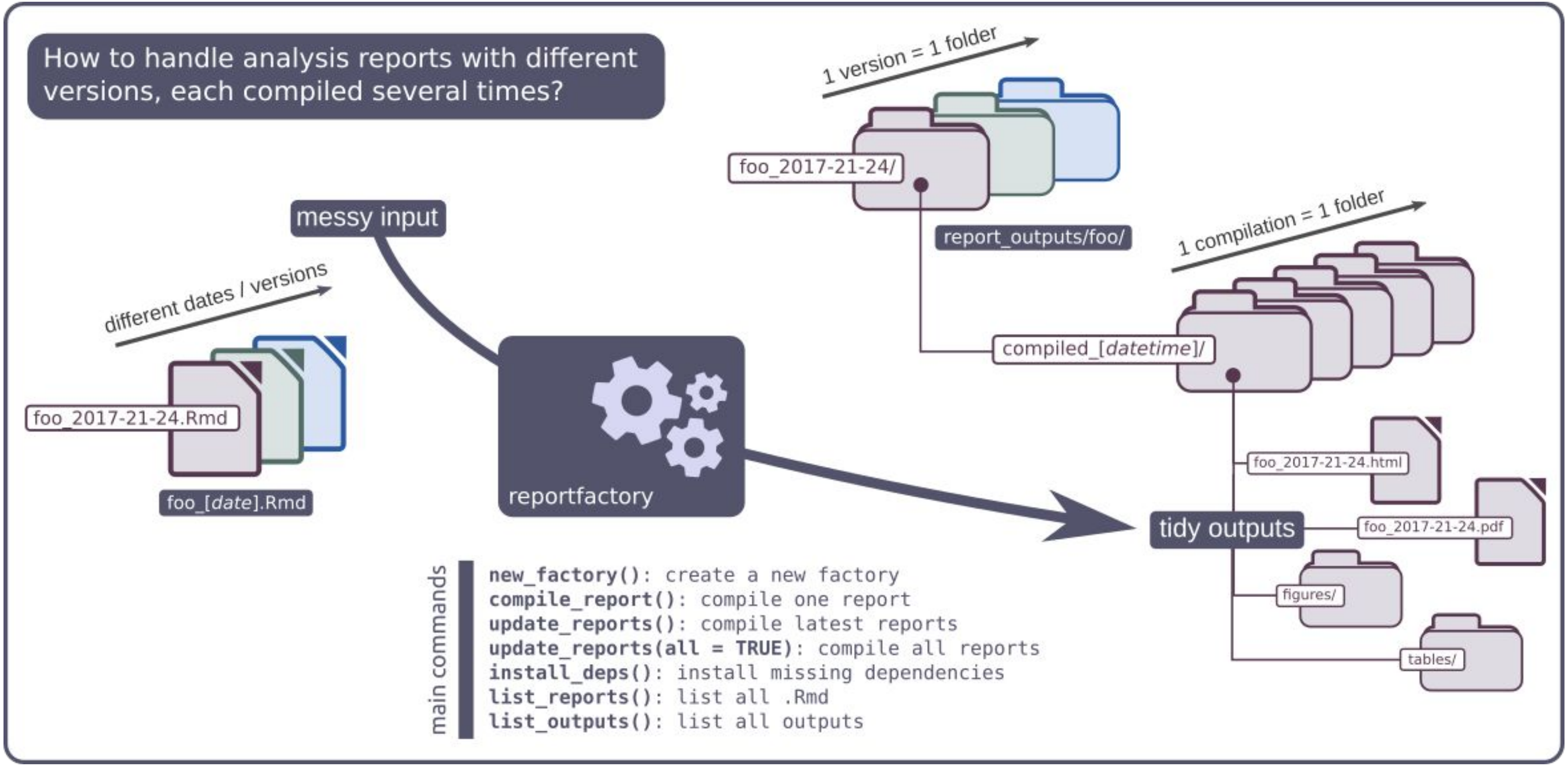

Estimated numbers of cases by health zone
Evolution des décisions d'alertes au cours du temps par zone de santé (N= 109 alertes)
Notification
Alerts
Data cleaning
Contact tracing
Data cleaning
Temporal dynamics
Courbe épidémique: cas confirmés et probables par date de notification
Automated reporting infrastructure
Data cleaning
Spatial spread
Data cleaning
Transmission chains
Delays
Data cleaning
Forecasting
Projections de nouveaux cas sur 21 jours - Mangina
Nombre de nouveaux cas
Quantile

Tidier rmarkdown workflows with *reportfactory* : use case

Original requirements

- Handle multiple `.Rmd` reports
- Handle multiple (dated) versions of the same report
- Separate data, scripts, `.Rmd` sources, outputs
- Generates time-stamped outputs
- Update all reports in one go
- Handle dependencies on packages
- Non-invasive: use of standard `.Rmd`, no config file
- Easy to use: accessible by people new to R
- Offline: does not require internet
- Portable: work on any platform

What does the *reportfactory* do?

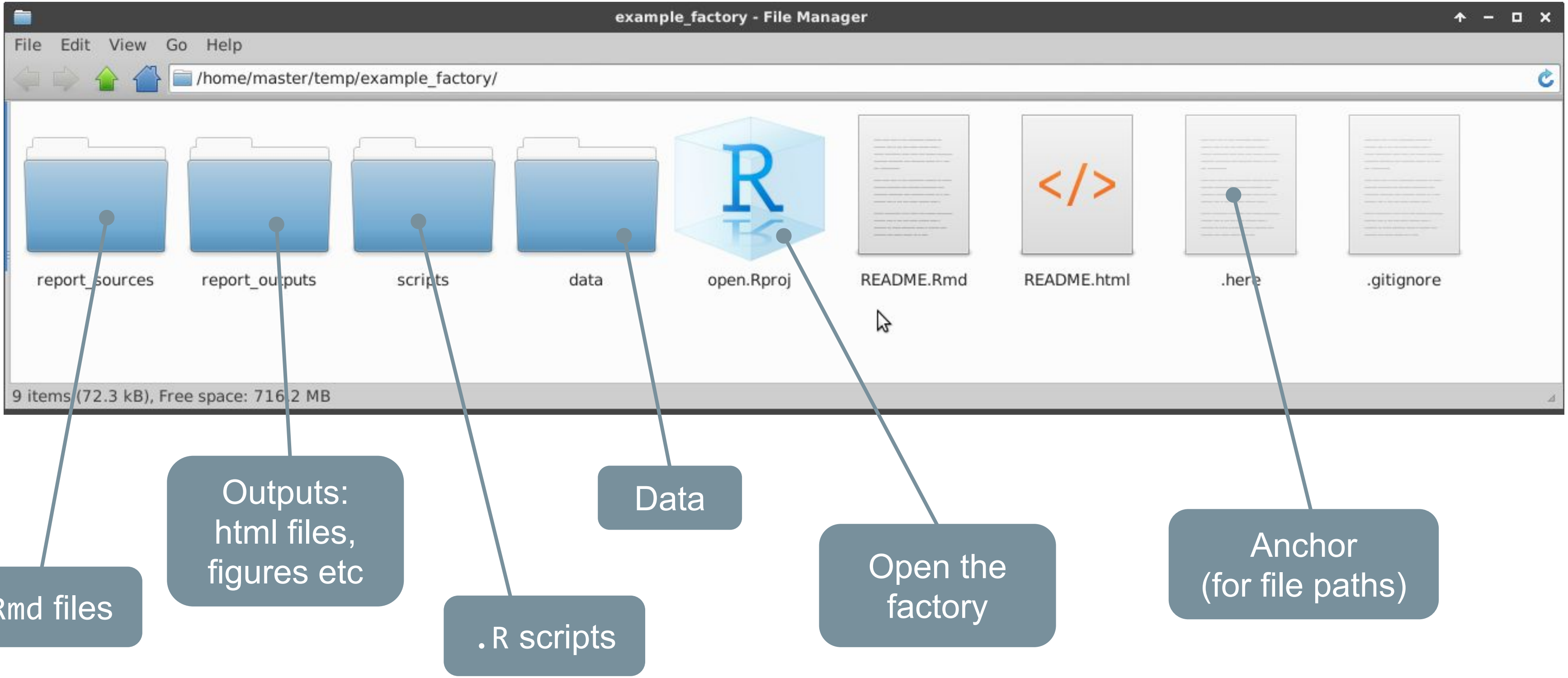


How to handle analysis reports with different versions, each compiled several times?

```
new_factory(): create a new factory
compile_report(): compile one report
update_reports(): compile latest reports
update_reports(all = TRUE): compile all reports
install_deps(): install missing dependencies
list_reports(): list all .Rmd
list_outputs(): list all outputs
```

reportfactory : basic structure

Creating a new factory: `new_factory()`


example_factory - File Manager
File
Edit
View
Go
Help
/home/master/temp/example_factory/
report_sources
report_outputs
scripts
data
open.Rproj
README.Rmd
README.html
.here
.gitignore
9 items (72.3 kB), Free space: 716.2 MB

.Rmd files

Outputs: html files, figures etc

.R scripts

Data

Open the factory

Anchor (for file paths)

reportfactory : other functionalities

Other functionalities

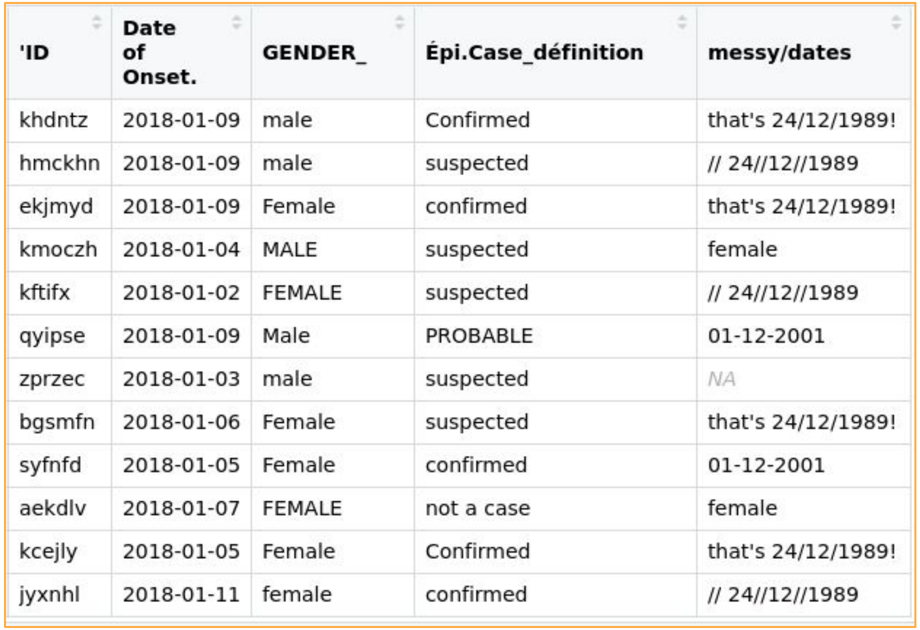
- List / install dependencies: `list_deps() / install_deps()`
- List reports: `list_reports()`
- Compile all recent reports: `update_reports()`
- Compile specific report: `compile_report()`
- Archive old reports: `archive_reports()`
- … : **contributions welcome!**

(join us, we have cookies)

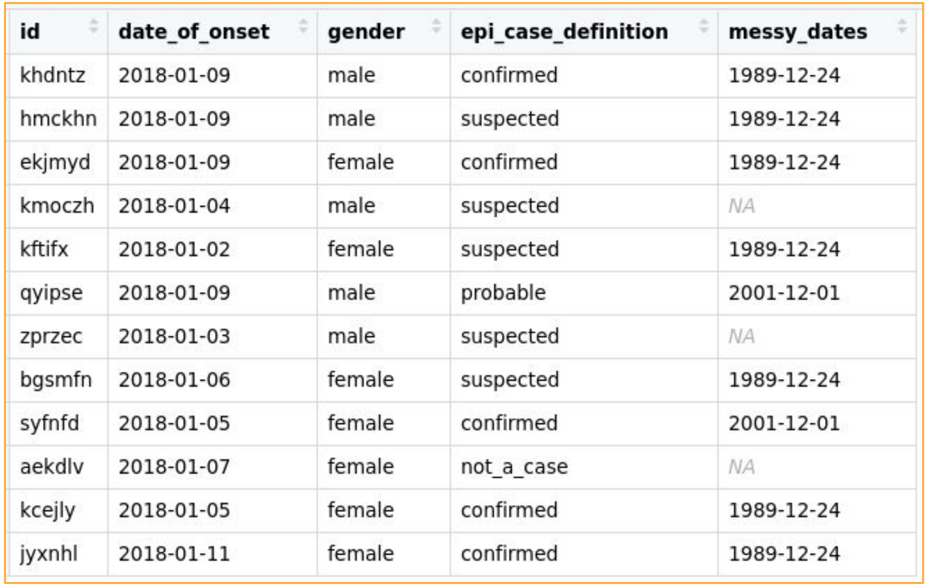
Data standardisation using *linelist*

```
x %>% clean_data()
```

Capitalisation
Accents
Separators
Dates



'ID	Date of Onset.	GENDER_	Épi.Case_définition	messy/dates
khdntz	2018-01-09	male	Confirmed	that's 24/12/1989!
hmckhn	2018-01-09	male	suspected	// 24//12//1989
ekjmyd	2018-01-09	Female	confirmed	that's 24/12/1989!
kmoczh	2018-01-04	MALE	suspected	female
kftifx	2018-01-02	FEMALE	suspected	// 24//12//1989
qyipse	2018-01-09	Male	PROBABLE	01-12-2001
zprzec	2018-01-03	male	suspected	NA
bgsmfn	2018-01-06	Female	suspected	that's 24/12/1989!
syfnfd	2018-01-05	Female	confirmed	01-12-2001
aekdlv	2018-01-07	FEMALE	not a case	female
kcejly	2018-01-05	Female	Confirmed	that's 24/12/1989!
jyxnhl	2018-01-11	female	confirmed	// 24//12//1989

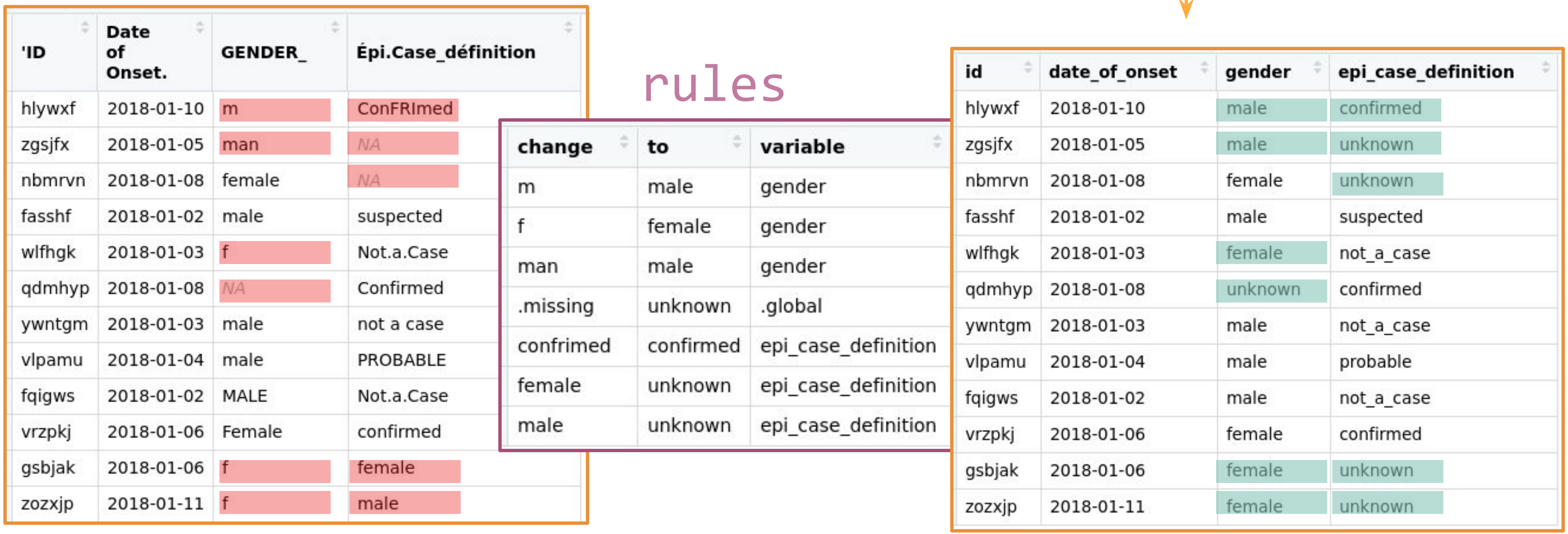


id	date_of_onset	gender	epi_case_definition	messy_dates
khdntz	2018-01-09	male	confirmed	1989-12-24
hmckhn	2018-01-09	male	suspected	1989-12-24
ekjmyd	2018-01-09	female	confirmed	1989-12-24
kmoczh	2018-01-04	male	suspected	NA
kftifx	2018-01-02	female	suspected	1989-12-24
qyipse	2018-01-09	male	probable	2001-12-01
zprzec	2018-01-03	male	suspected	NA
bgsmfn	2018-01-06	female	suspected	1989-12-24
syfnfd	2018-01-05	female	confirmed	2001-12-01
aekdlv	2018-01-07	female	not_a_case	NA
kcejly	2018-01-05	female	confirmed	1989-12-24
jyxnhl	2018-01-11	female	confirmed	1989-12-24

Dictionary-based cleaning using *linelist*

```
x %>% clean_data(wordlists = rules)
```

Typos
Re-levelling
Variable-specific rules



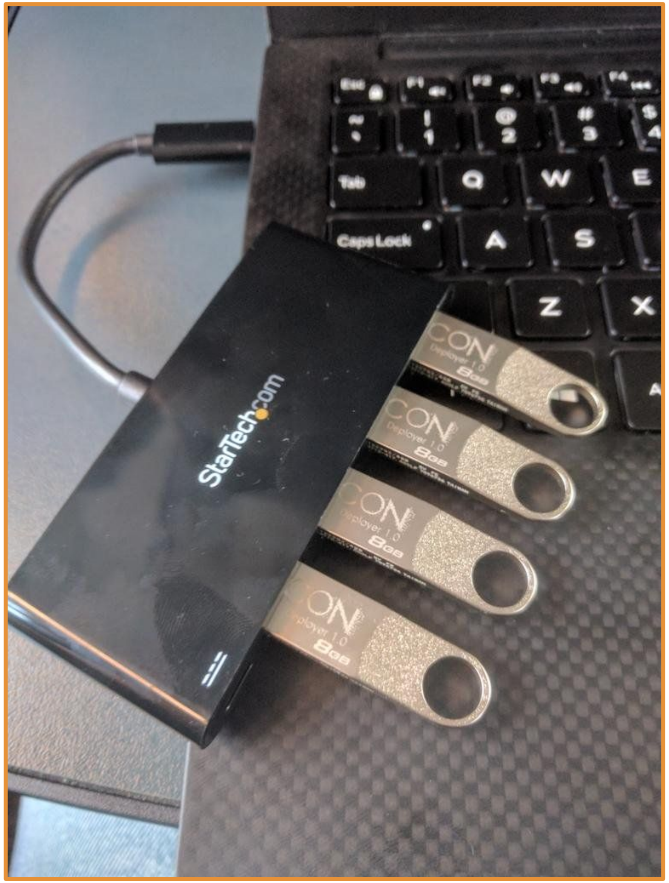
'ID	Date of Onset.	GENDER_	Épi.Case_définition
hlywxf	2018-01-10	m	ConFRImed
zgsjfx	2018-01-05	man	*NA*
nbmrvn	2018-01-08	female	*NA*
fasshf	2018-01-02	male	suspected
wlfhgk	2018-01-03	f	Not.a.Case
qdmhyp	2018-01-08	*NA*	Confirmed
ywntgm	2018-01-03	male	not a case
vlpamu	2018-01-04	male	PROBABLE
fqigws	2018-01-02	MALE	Not.a.Case
vrzpkj	2018-01-06	Female	confirmed
gsbjak	2018-01-06	f	female
zozxjp	2018-01-11	f	male

rules

change	to	variable
m	male	gender
f	female	gender
man	male	gender
.missing	unknown	.global
confrimed	confirmed	epi_case_definition
female	unknown	epi_case_definition
male	unknown	epi_case_definition

id	date_of_onset	gender	epi_case_definition
hlywxf	2018-01-10	male	confirmed
zgsjfx	2018-01-05	male	unknown
nbmrvn	2018-01-08	female	unknown
fasshf	2018-01-02	male	suspected
wlfhgk	2018-01-03	female	not_a_case
qdmhyp	2018-01-08	unknown	confirmed
ywntgm	2018-01-03	male	not_a_case
vlpamu	2018-01-04	male	probable
fqigws	2018-01-02	male	not_a_case
vrzpkj	2018-01-06	female	confirmed
gsbjak	2018-01-06	female	unknown
zozxjp	2018-01-11	female	unknown

Taking R offline using the *deployer*



The RECON deployer

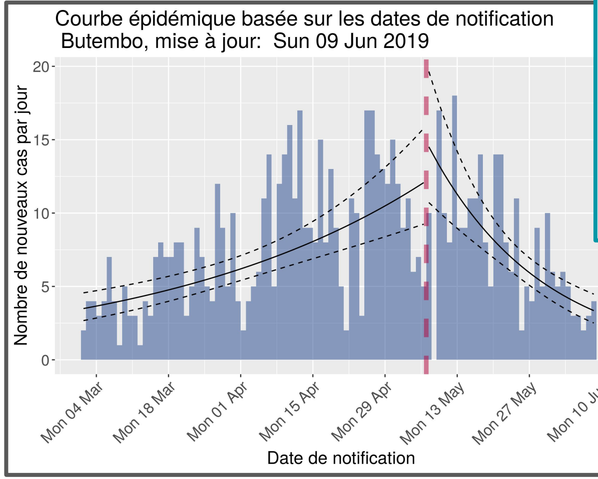
- USB stick with latest R, Rtools, Rstudio for Windows, MacOSX, Linux
- Local package repository - instance of *nomad*: https://github.com/reconhub/nomad
- ~2000-3000 CRAN packages
- ~10-20 github packages
- Cheatsheets
- Website: https://github.com/reconhub/deployer

Making a difference

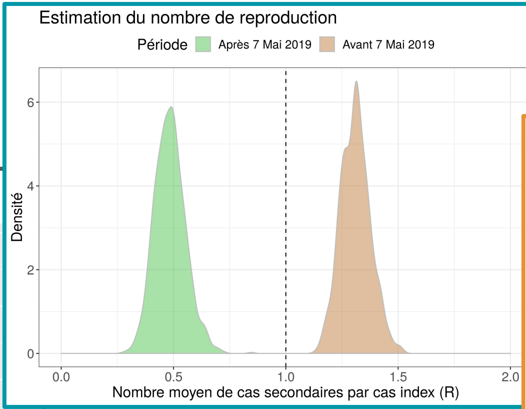
Showing what works
Join the movement

Making a difference: showing what works

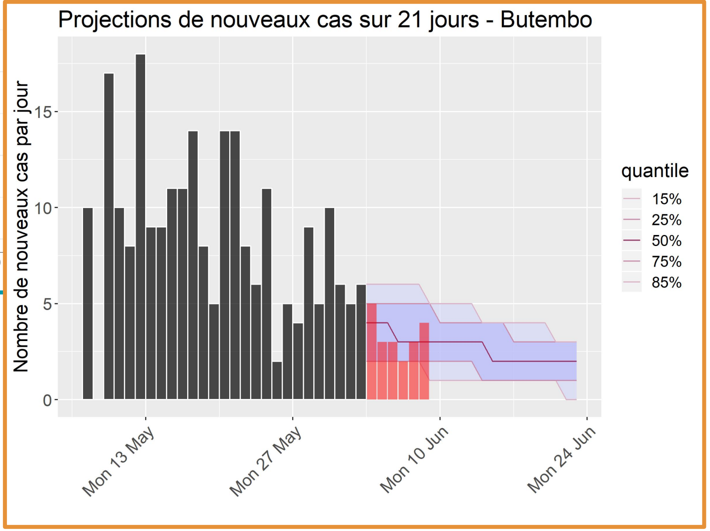
Trend shift with new control strategy



Transmission became < 1



Predicted control within weeks



Confirmed effectiveness of new control strategy

Join the movement!

Outbreak analytics

- Still an emerging field
- Funding and training gaps
- Data scientists needed!

The good stuff

- Help respond to health emergencies and humanitarian crises
- Work with visible impact
- Exciting data challenges
- Lots of potential for capacity building: the **next generation of data scientists needs to be in-country**

Thanks to

Ebola response (institutions): MoH DRC, WHO, GOARN, UK Public Health Rapid Support Team, MSF, UNICEF, WFP

Ebola Emergency Operation Centre, Goma: Xavier de Radiguès, Mathias Mossoko, Christopher Jarvis, Flavio Fingers, Madeleine Crowe, Pascale Lissouba, Simone Carter, Emilie Buttarelli, Nabil Tabal, Richy NGombo, Patrice Ndji, Yannick Tutu, Germain Mweha-Nzuko, Dorothée Bulemfu, Annie Mutombo, Emilia Sana, Franck Mboussou, Luigi Minikulu, Aaron Aruna, Michel Yao, Ibrahima Socé-Fall

LSHTM: John Edmunds, Olivier le Polain, Daniel Bausch, Bayard Roberts, Sebastian Funk, Patrick Keating

Imperial College London: Zhian Kamvar, Rich FitzJohn, Finlay Campbell

Funding:
GCRF project RECAP (ES/P010873/1), UK PH RST, HPRU-NIHR, MRC

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Imperial College London

RECON

Get these slides