RECON

Building the next generation of statistical tools for outbreak response using R

Thibaut Jombart

5th January 2017

Imperial College London
MRC Centre for Outbreak Analysis and Modelling

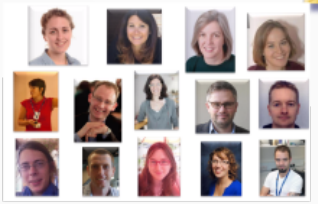
Outline

1. Lessons learnt from the Ebola outbreak response
2. The R Epidemics Consortium
3. RECON projects: from basic needs to active research
4. Methodological dialogue during outbreak response

Ebola response

Lessons learnt from the Ebola response

Lessons learnt from the Ebola response



Imperial College Ebola team

Lessons learnt from the Ebola response

Imperial College Ebola team

Lessons learnt from the Ebola response

Most statistical/modelling tools for situation awareness missing.

What tools do we need?

Some examples:

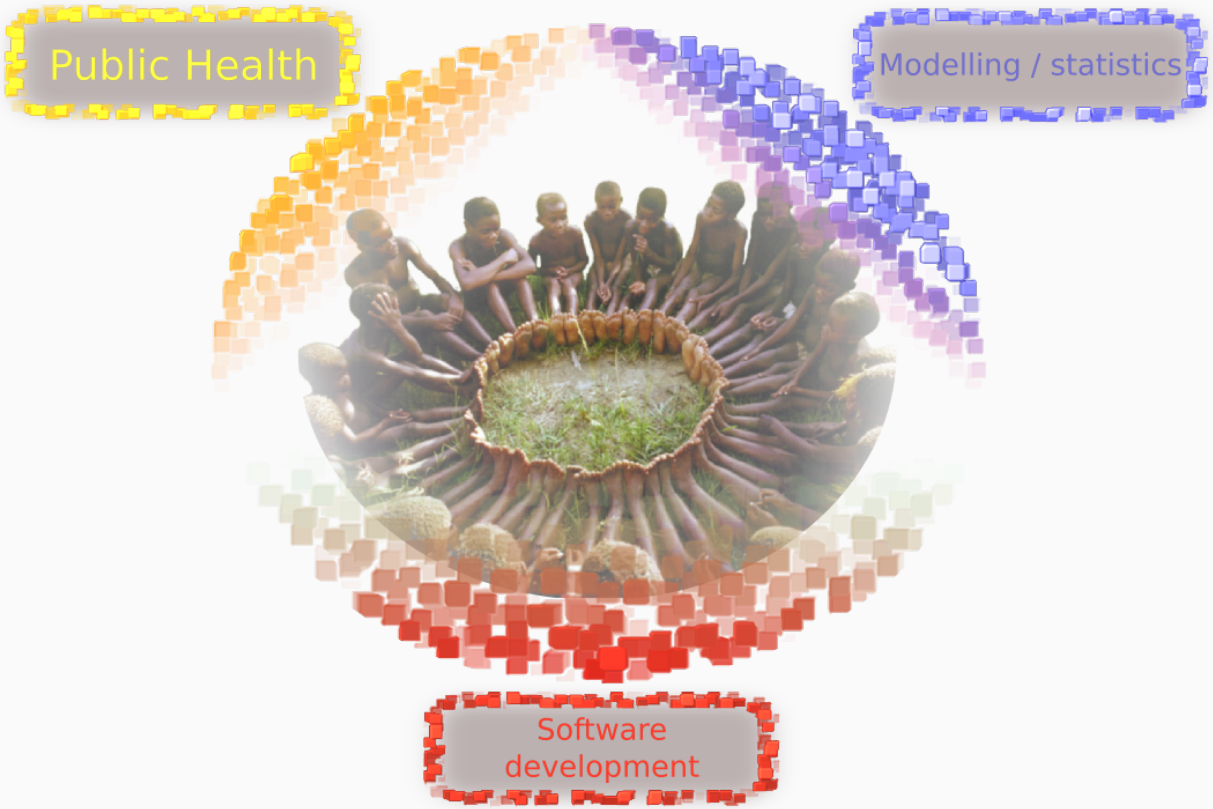
- **data cleaning:** dictionaries, entry matching
- **graphics:** case incidence in space and time, contact tracing
- **parameter estimation:** key delays, transmissibility
- **estimate / test CFR:** gender, health care workers, treaments effects
- **predictions:** case incidence, mortality, evaluate interventions
- **report:** (semi-)automated situation reports

Who do we need to develop these tools?

Who do we need to develop these tools?

Who do we need to develop these tools?

Who do we need to develop these tools?

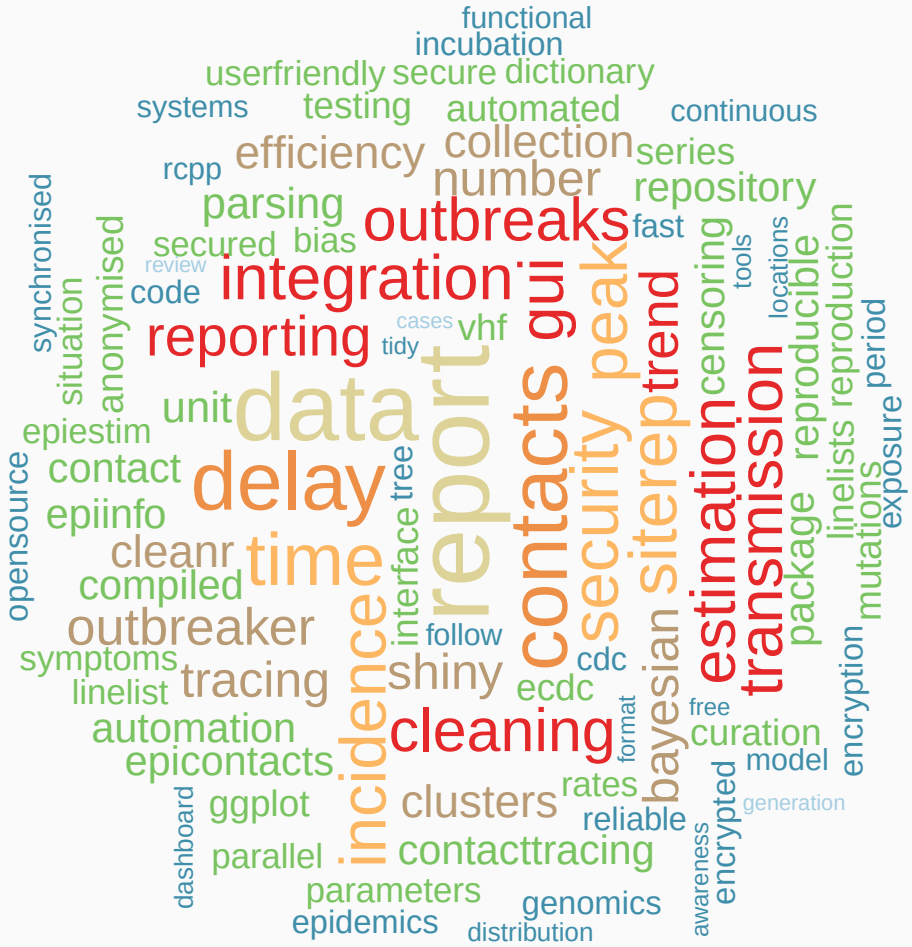


The R Epidemics Consortium

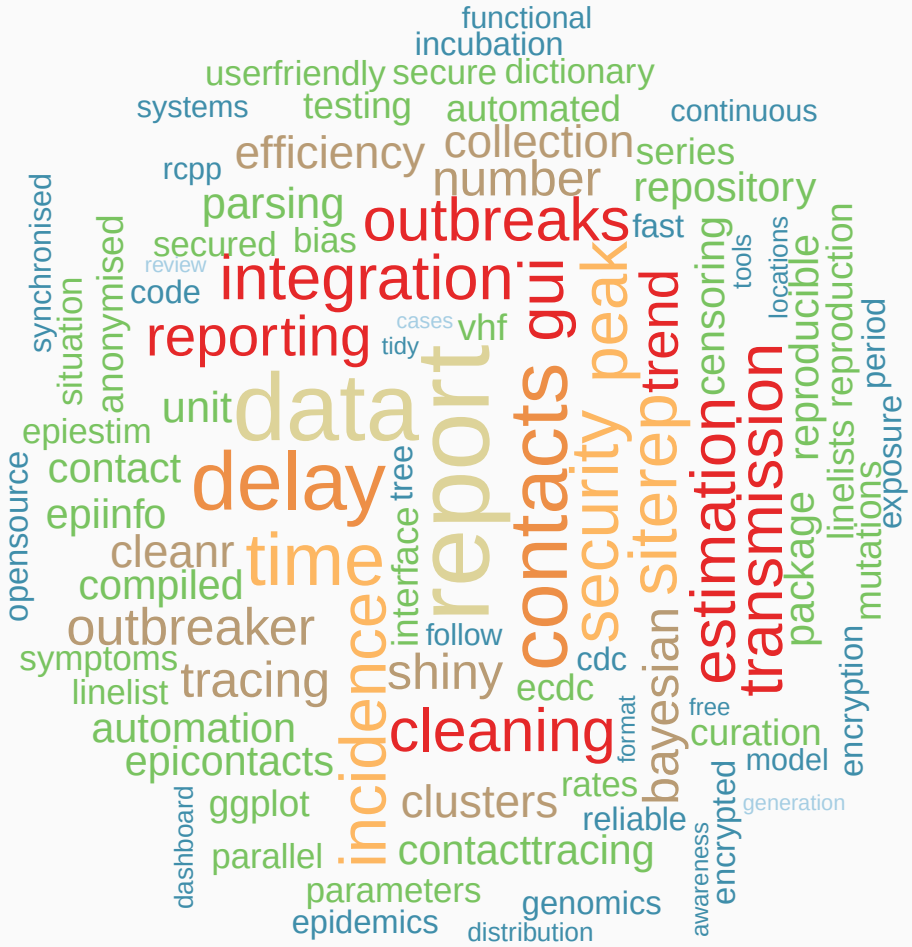
Hackout 3: a R hackathon for emergency outbreak response

Last summer at the *rOpenSci* headquarters (Berkeley)

Hackout 3: from ideas to projects to...



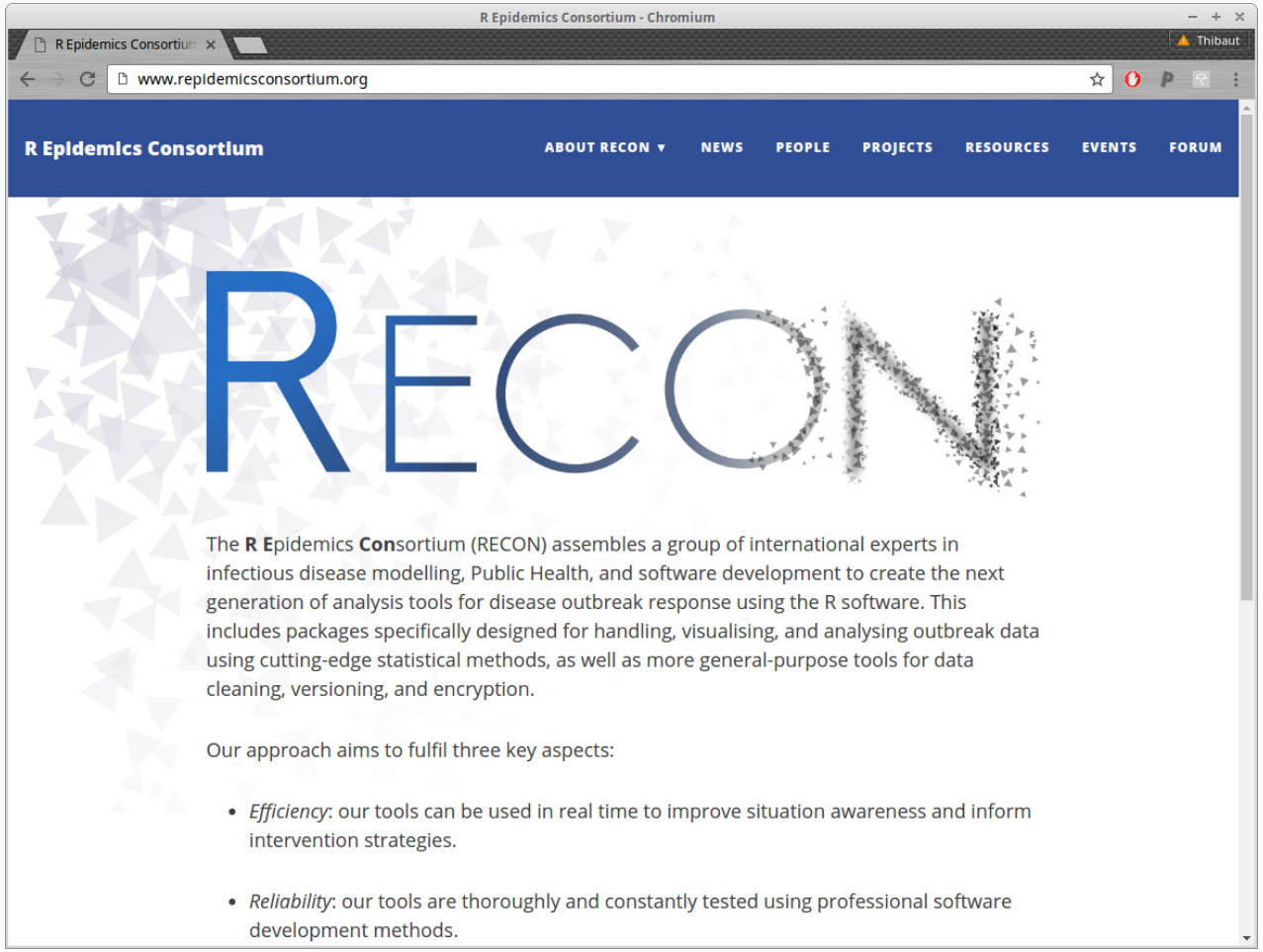
Hackout 3: from ideas to projects to...



How do we keep momentum once the event is over?

RECON: the R Epidemics Consortium

A taskforce to build a new generation of outbreak response tools in .



www.repidemicsconsortium.org

In a nutshell

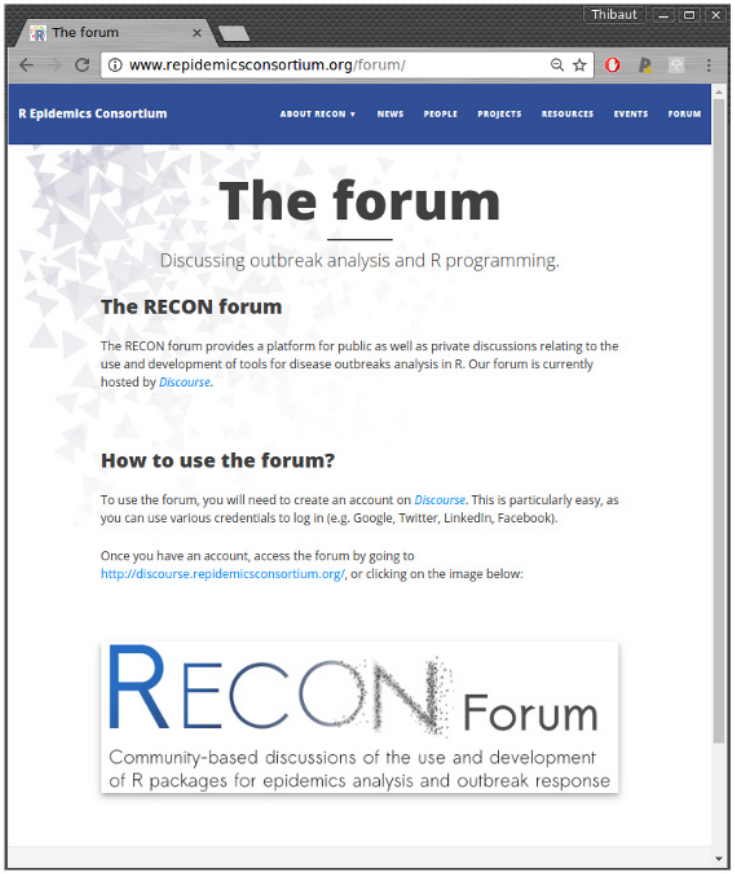
RECON

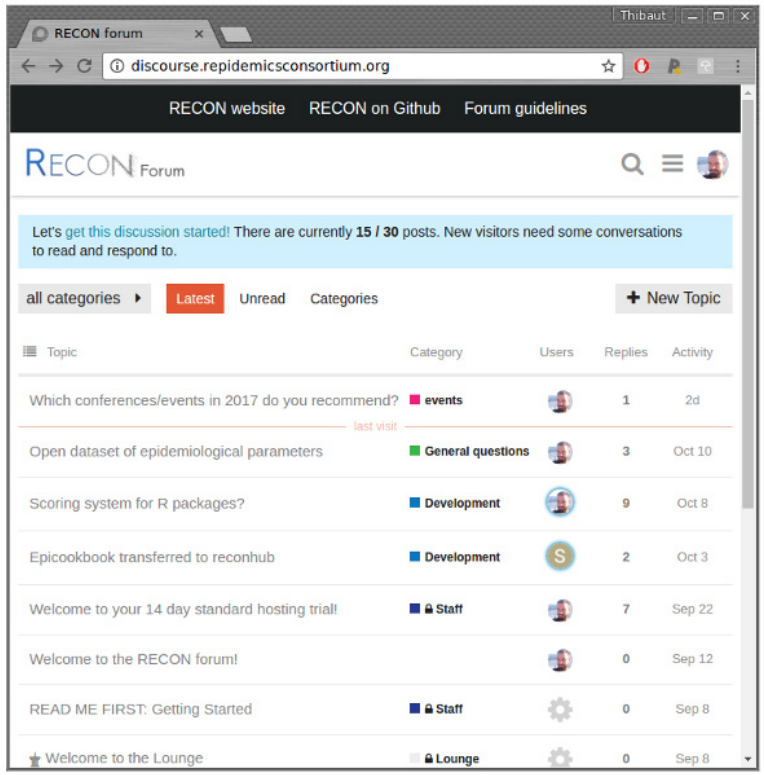
www.repidemicsconsortium.org

- started 6th September 2016
- 48 people (42 members, 6 board)
- 10 countries, $>$ 20 institutions
- $\sim$ 10 new packages coming
- public forum, blog, online resources

The RECON forum

A platform for discussing epidemics analysis in R .

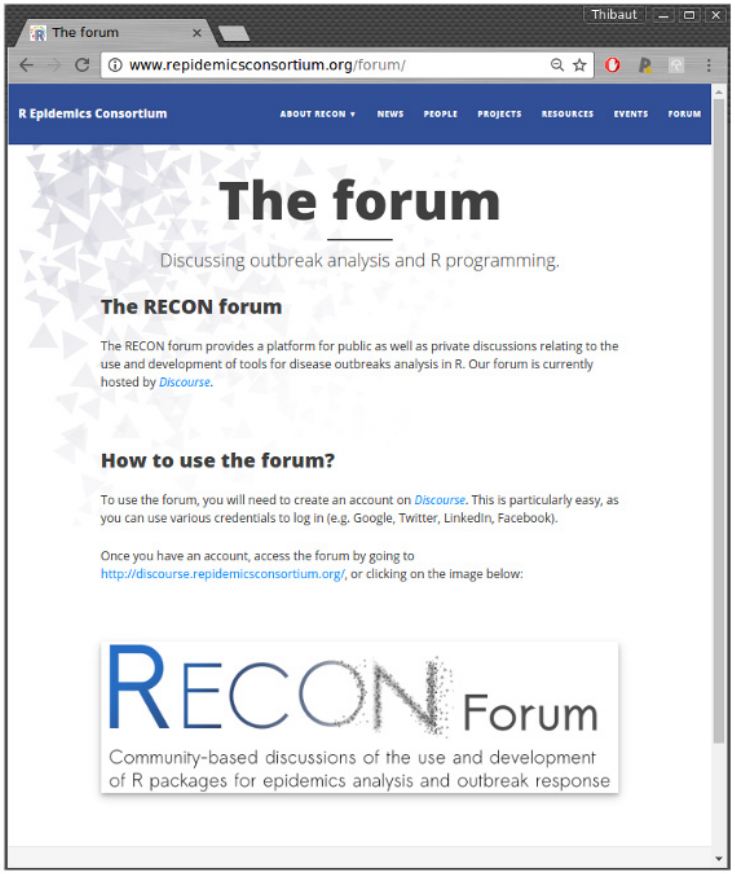


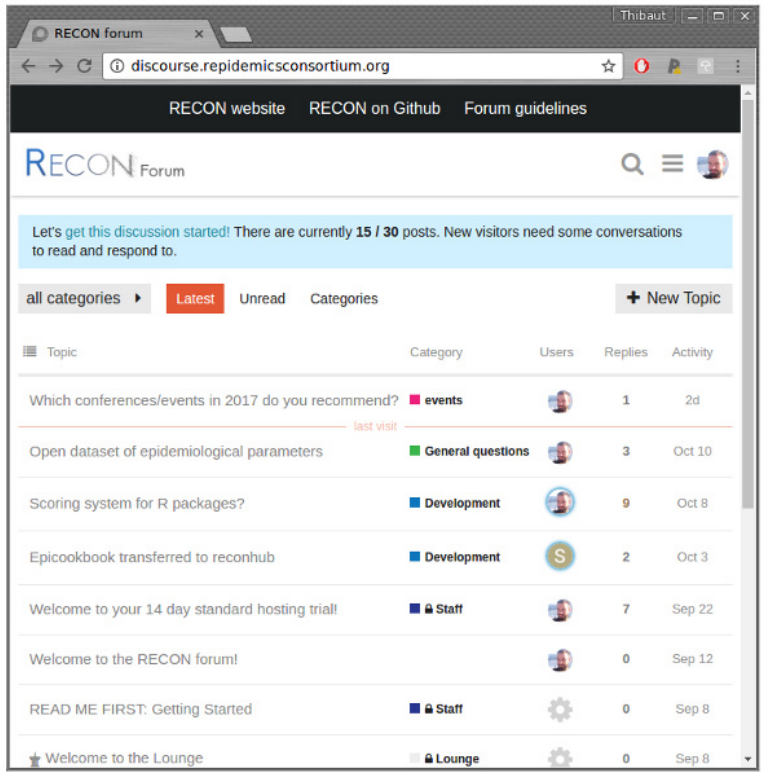


www.repidemicsconsortium.org/forum

The RECON forum

A platform for discussing epidemics analysis in R .





www.repidemicsconsortium.org/forum

Join us!

RECON package: what do we aim for?

- **efficiency**: useful for improving situation awareness in real time; cutting-edge, computer-efficient statistical methods

RECON package: what do we aim for?

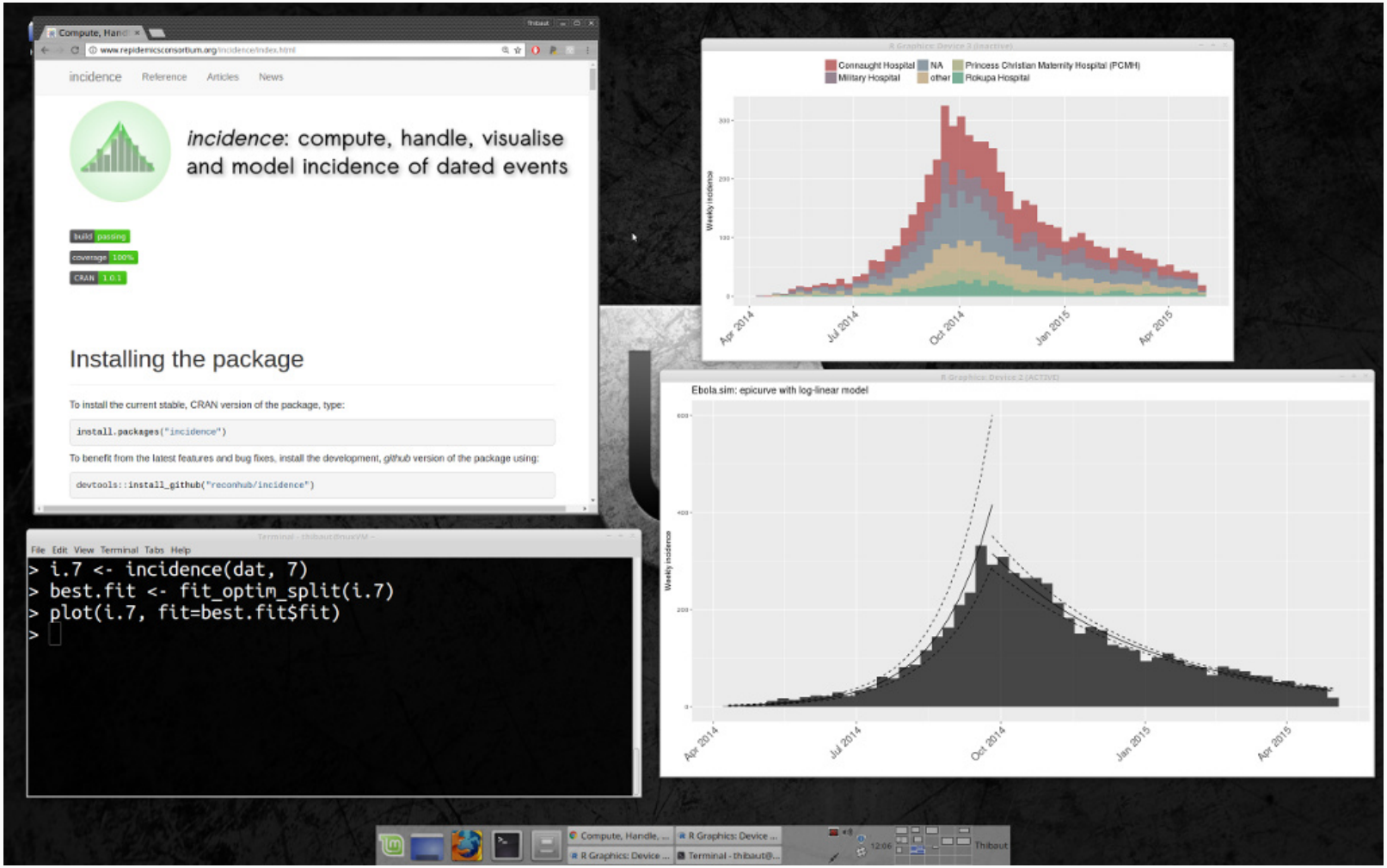
- **efficiency**: useful for improving situation awareness in real time; cutting-edge, computer-efficient statistical methods
- **reliability**: outputs can be trusted; continuous integration, extensive unit testing, code review, good practices

RECON package: what do we aim for?

- **efficiency**: useful for improving situation awareness in real time; cutting-edge, computer-efficient statistical methods
- **reliability**: outputs can be trusted; continuous integration, extensive unit testing, code review, good practices
- **accessibility**: widely available, easy learning curve; extensive documentation, tutorials, websites, forum

RECON projects: from basic needs to active research

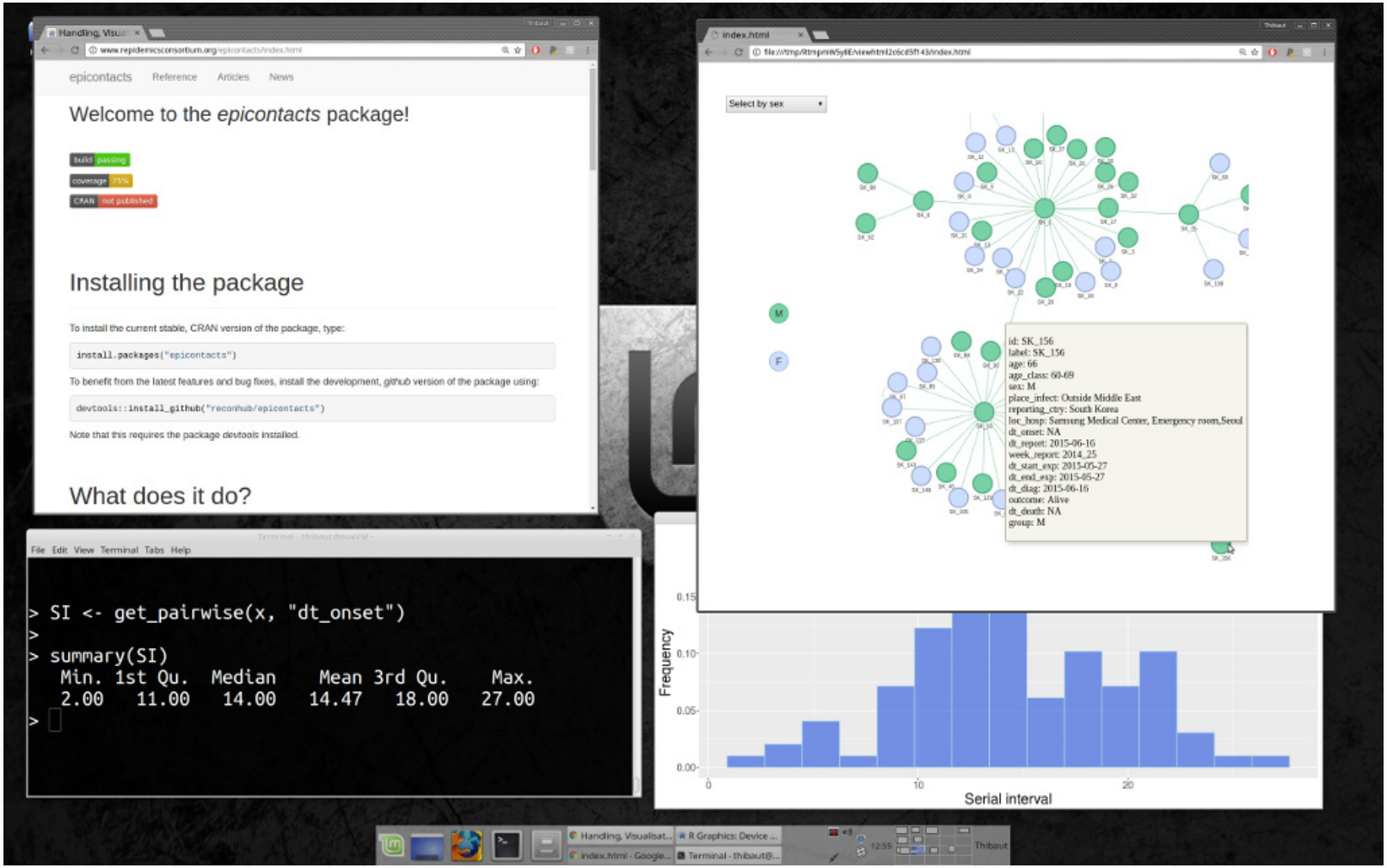
incidence: computation, handling, visualisation and modelling of epicurves



www.repidemicsconsortium.org/incidence

[released]

epicontacts: handling, visualisation and analysis of epidemiological contacts



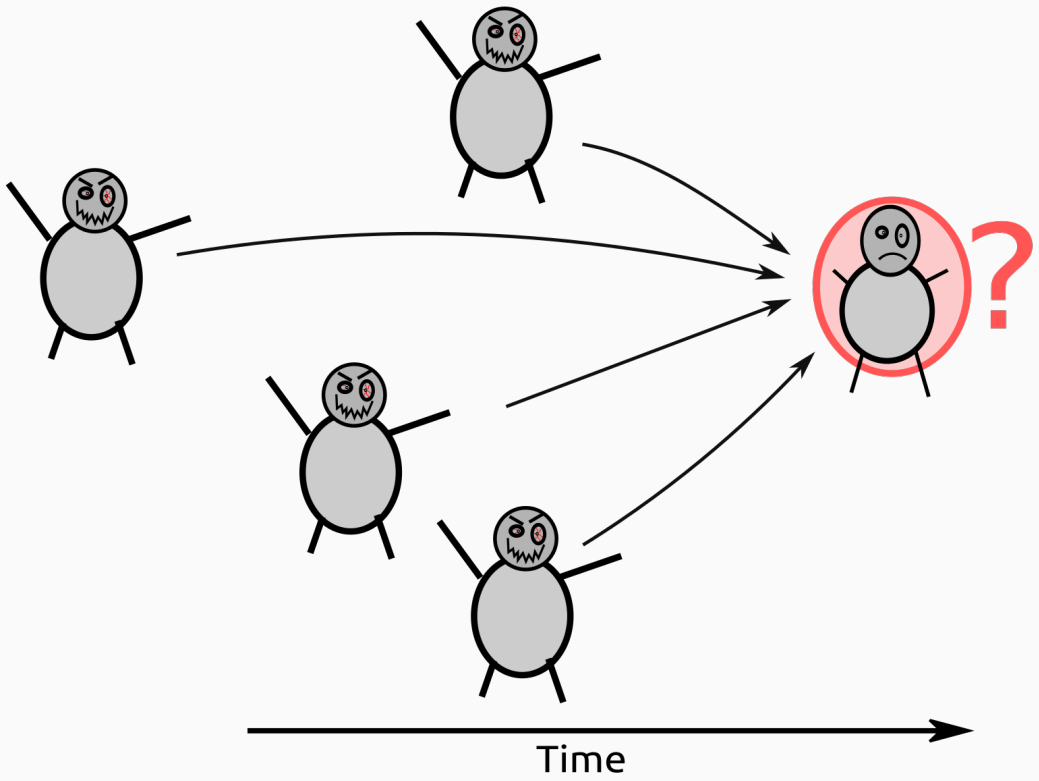
www.repidemicsconsortium.org/epicontacts

[release February 2017]

outbreaker: infering who infects whom from epi/genetic data (1/2)

Use timing of symptoms and pathogen genomes to infer infectors

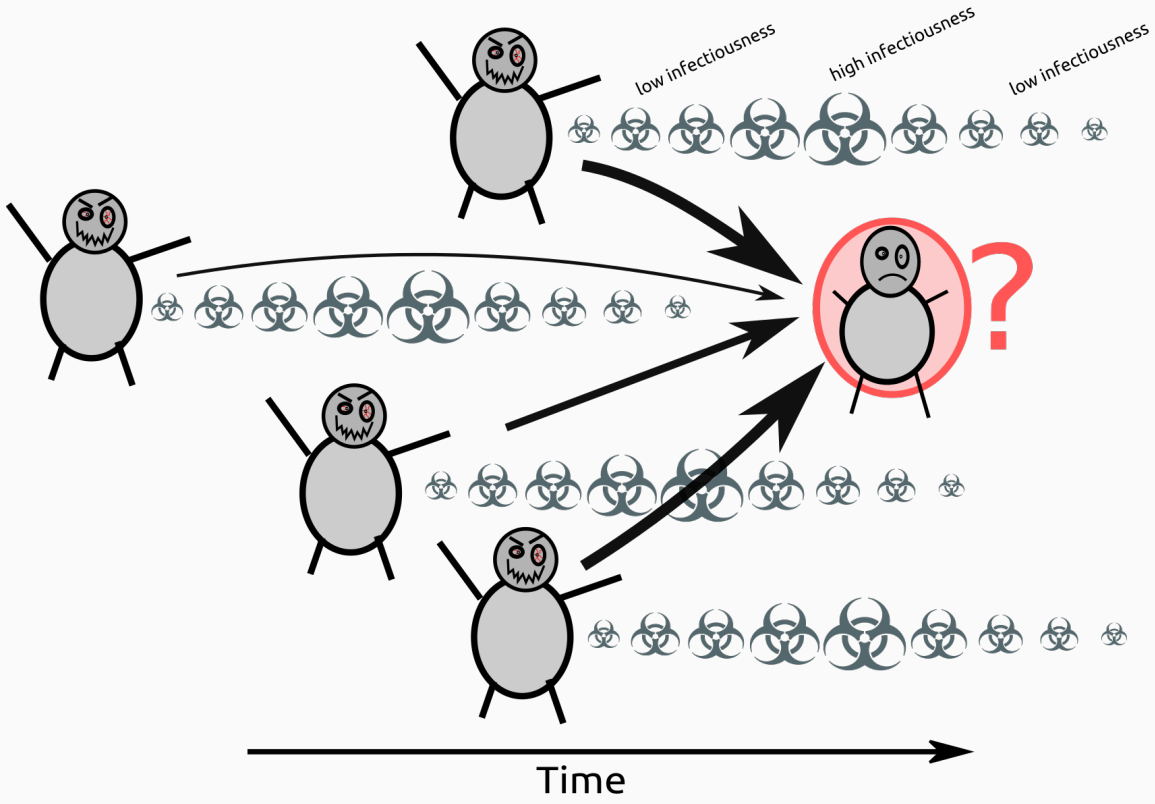
(Jombart *et al*, PLoS Comp Biol, 2014)



outbreaker: infering who infects whom from epi/genetic data (1/2)

Use timing of symptoms and pathogen genomes to infer infectors

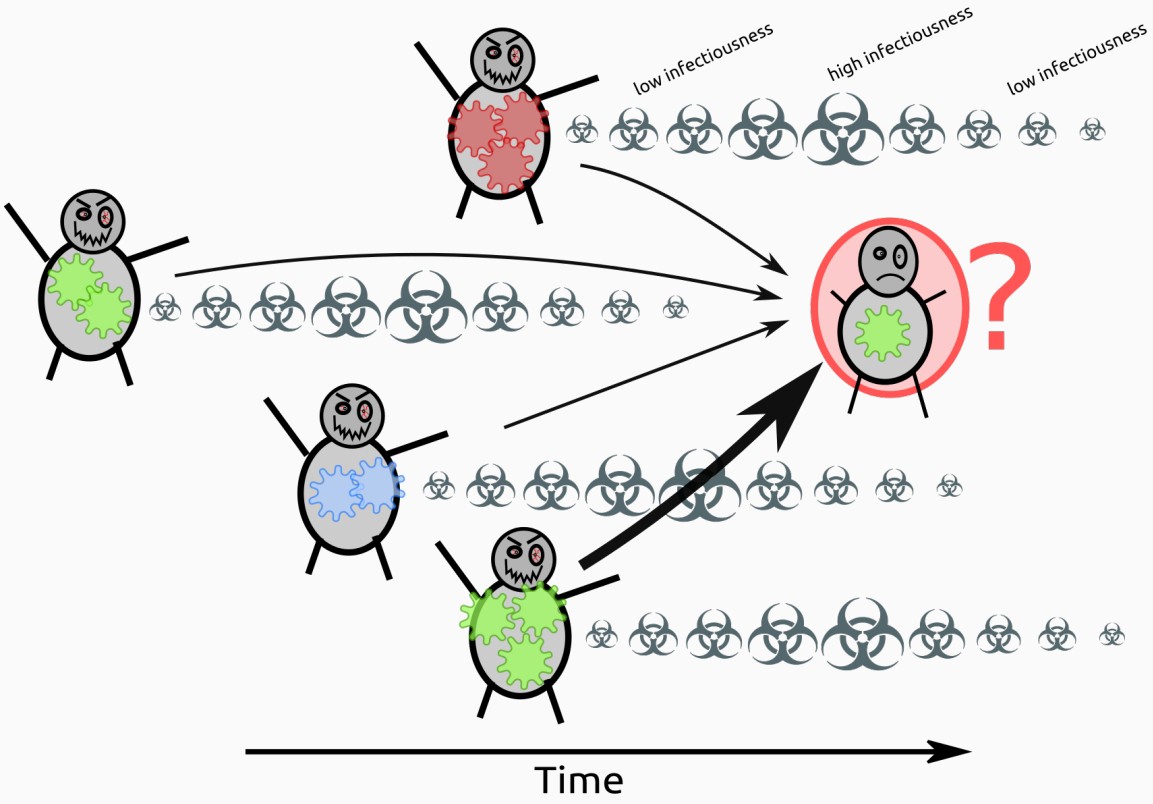
(Jombart *et al*, PLoS Comp Biol, 2014)



outbreaker: infering who infects whom from epi/genetic data (1/2)

Use timing of symptoms and pathogen genomes to infer infectors

(Jombart *et al*, PLoS Comp Biol, 2014)



outbreaker: infering who infects whom from epi/genetic data (2/2)

General approach

- Bayesian framework
- augmented data for ancestries, infection dates, unobserved cases
- MCMC (Metropolis-Hasting) to sample from posterior
- identification of imported cases (outliers)

outbreaker: infering who infects whom from epi/genetic data (2/2)

General approach

- Bayesian framework
- augmented data for ancestries, infection dates, unobserved cases
- MCMC (Metropolis-Hasting) to sample from posterior
- identification of imported cases (outliers)

Likelihood

- $p(\text{transmission tree}) = \prod_i p(\text{branch}_i)$
- $p(\text{branch}) = p(\text{infection date}) \times p(\text{collection date}) \times$
  $p(\text{unobserved cases}) \times p(\text{genetic variation})$

outbreaker: detail of epidemiological likelihood

Likelihood for case i with ancestor α_i:

$$
\begin{aligned}
&p(\text{collection date}) \times p(\text{infection date}) \times p(\text{unobserved cases}) \\
&= p(t_i|T_i^{inf}) \times p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i) \times p(\kappa_i|\pi) \\
&= f(t_i - T_i^{inf}) \times w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf}) \times f_{\mathcal{G}}(1|\kappa_i - 1, \pi)
\end{aligned}
$$

outbreaker: detail of epidemiological likelihood

Likelihood for case i with ancestor α_i:

$$
\begin{aligned}
&p(\text{collection date}) \times p(\text{infection date}) \times p(\text{unobserved cases}) \\
&= p(t_i|T_i^{inf}) \times p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i) \times p(\kappa_i|\pi) \\
&= f(t_i - T_i^{inf}) \times w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf}) \times f_{\mathcal{G}}(1|\kappa_i - 1, \pi)
\end{aligned}
$$

with:

- t_i: collection date of i
- T_i^{inf}: date of infection of i
- κ_i: number of generations between i and α_i
- π: proportion of the outbreak sampled
- f: distribution of colonisation duration (assumed known)
- w: distribution of generation time ($w^{(\kappa_i)}$: κ_i convolutions of w; w assumed known)
- $f_{\mathcal{G}}$: geometric distribution

outbreaker: detail of genetic likelihood

Likelihood for case i with ancestor α_i:

$$\mu^{d(s_i, s_{\alpha_i})}(1-\mu)^{(\kappa_i \times l(s_i, s_{\alpha_i})) - d(s_i, s_{\alpha_i})}$$

outbreaker: detail of genetic likelihood

Likelihood for case i with ancestor α_i:

$$\mu^{d(s_i, s_{\alpha_i})}(1-\mu)^{(\kappa_i \times l(s_i, s_{\alpha_i})) - d(s_i, s_{\alpha_i})}$$

with:

- s_i: sequence of i
- μ: mutation rate
- $d(s_i, s_{\alpha_i})$: number of mutations between case i and α_i
- κ_i: number of generations between i and α_i
- $l(s_i, s_{\alpha_i})$: number of comparable sites between s_i and s_{α_i}

The R package *outbreaker*

Original implementation aimed at computer efficiency:

- C implementation embedded within R package
- multi-platform: linux, MacOS X, Windows, Solaris, ...
- supports parallelization
- post-processing of MCMC, simulations, graphics
- outbreak simulation tool

The R package *outbreaker*

Original implementation aimed at computer efficiency:

- C implementation embedded within R package
- multi-platform: linux, MacOS X, Windows, Solaris, ...
- supports parallelization
- post-processing of MCMC, simulations, graphics
- outbreak simulation tool

Fast, but not very flexible, hard to unit-test, hard to change code.

The R package *outbreaker*

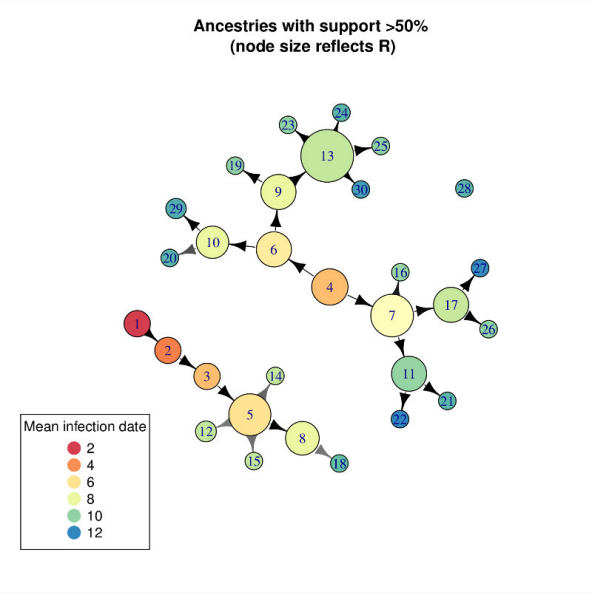
Original implementation aimed at computer efficiency:

- C implementation embedded within R package
- multi-platform: linux, MacOS X, Windows, Solaris, ...
- supports parallelization
- post-processing of MCMC, simulations, graphics
- outbreak simulation tool

Fast, but not very flexible, hard to unit-test, hard to change code.

New methods emerged since.

Methods differ by their treatment of genetic data

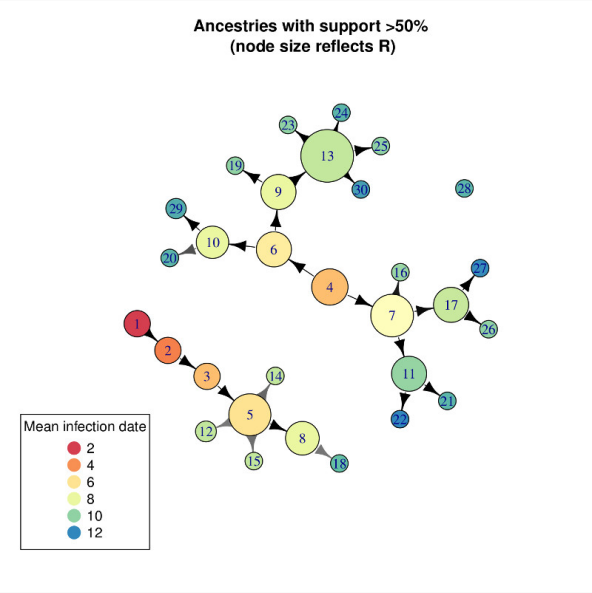


[*outbreaker*]

Quick and dirty

- mutation at transmission
- approximate genetic likelihood
- less accurate but computer efficient
- *e.g.* Ypma *et al.* 2012; Jombart *et al.* 2014

Methods differ by their treatment of genetic data



[*outbreaker*]

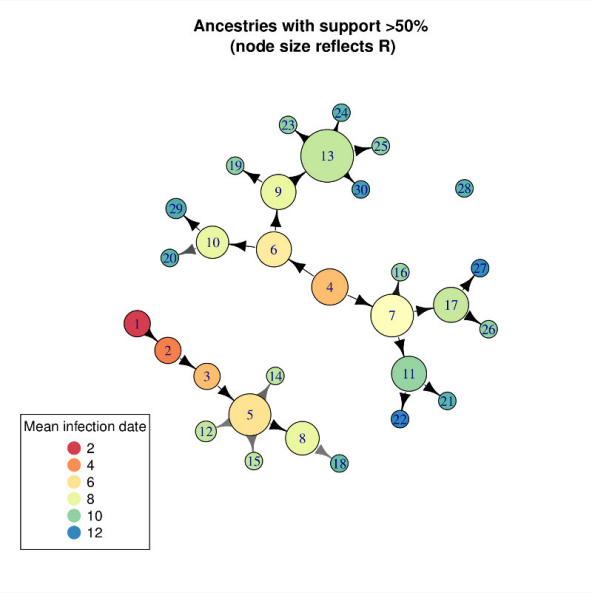
Quick and dirty

- mutation at transmission
- approximate genetic likelihood
- less accurate but computer efficient
- *e.g.* Ypma *et al.* 2012; Jombart *et al.* 2014

Nice and slow

- within-host evolution
- $\sim$ phylogenetic likelihood
- more accurate but costly
- *e.g.* Ypma *et al.* 2013, Didelot *et al.* 2014

Methods differ by their treatment of genetic data



[*outbreaker*]

Quick and dirty

- mutation at transmission
- approximate genetic likelihood
- less accurate but computer efficient
- *e.g.* Ypma *et al.* 2012; Jombart *et al.* 2014

Nice and slow

- within-host evolution
- $\sim$ phylogenetic likelihood
- more accurate but costly
- *e.g.* Ypma *et al.* 2013, Didelot *et al.* 2014

Focus on genome sequences *vs* integrate various types of data.

Are different approaches really that different?

Are different approaches really that different?

Are different approaches really that different?

Different models can lead to very similar implementations.

Are different approaches really that different?

Different models can lead to very similar implementations.

Can we find a general formulation for the different models?

What do these models look like?

- a, b, c: different types of data
- θ: parameters / augmented data

Different types of data often assumed to be *conditionally independent*:

$$p(a, b, c|\theta) = p(a|\theta)p(b|\theta)p(c|\theta)$$

What do these models look like?

- a, b, c: different types of data
- θ: parameters / augmented data

Different types of data often assumed to be *conditionally independent*:

$$p(a, b, c|\theta) = p(a|\theta)p(b|\theta)p(c|\theta)$$

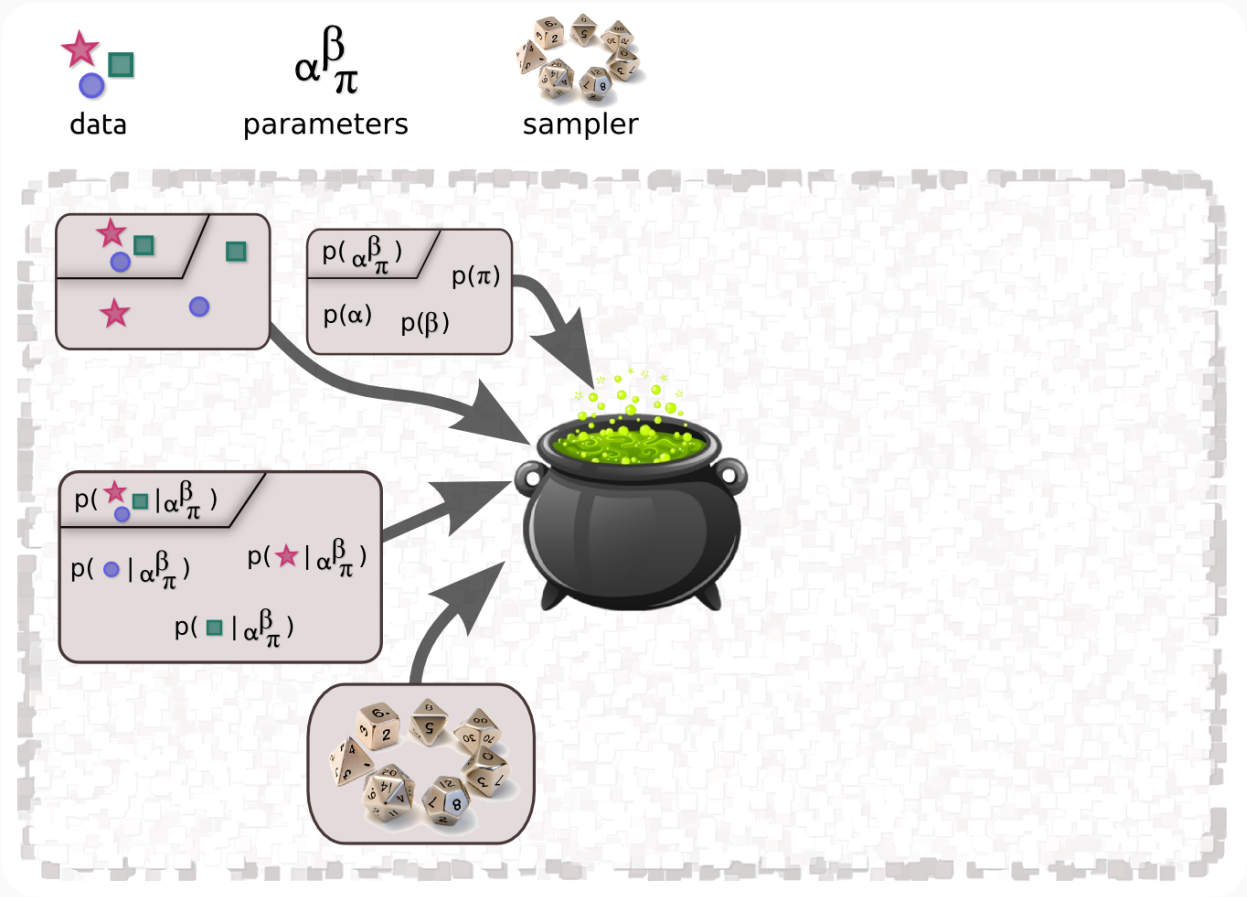
Components can be treated as plugins.

outbreaker2: a general cauldron for cooking methods

Use-your-own: data type, likelihood, prior, MCMC.

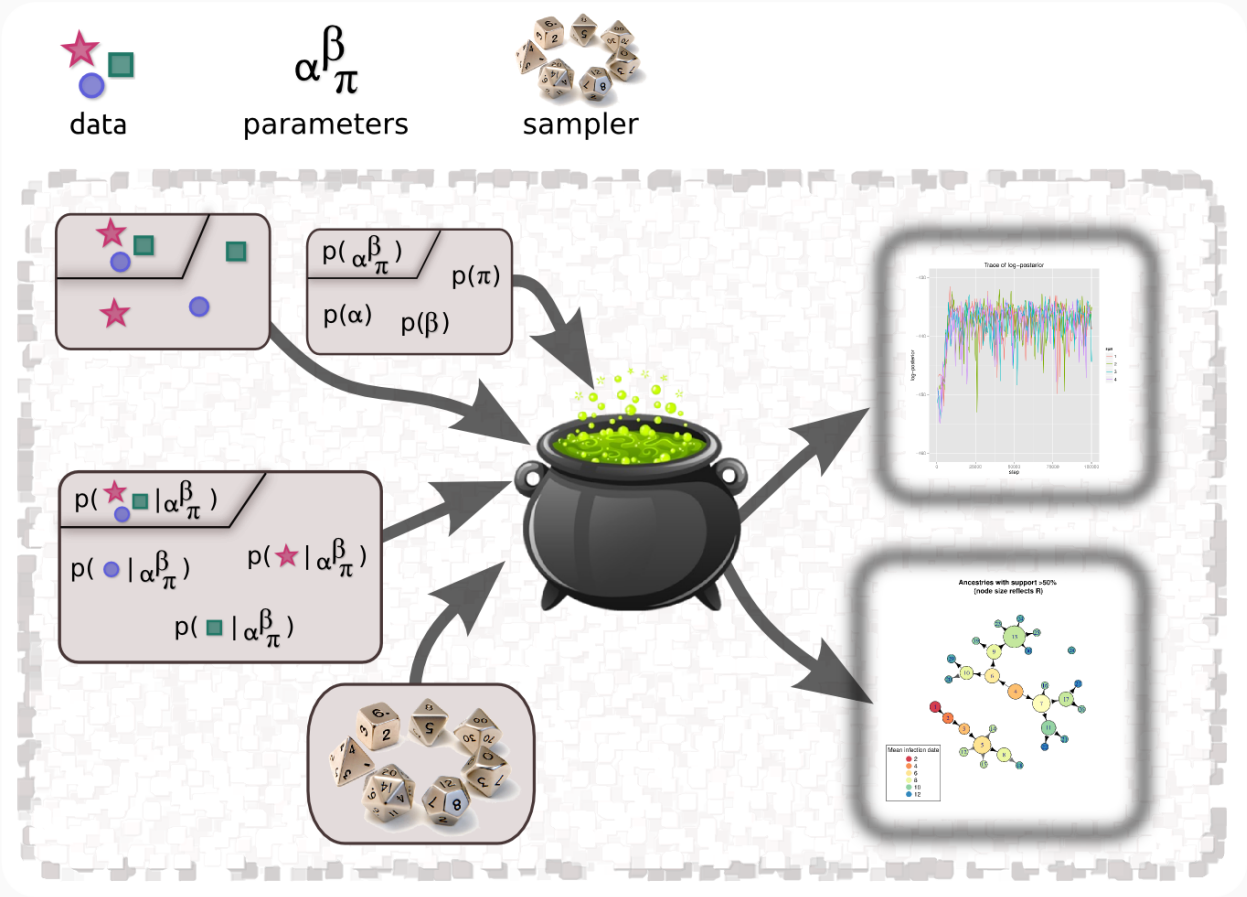
outbreaker2: a general cauldron for cooking methods

Use-your-own: data type, likelihood, prior, MCMC.



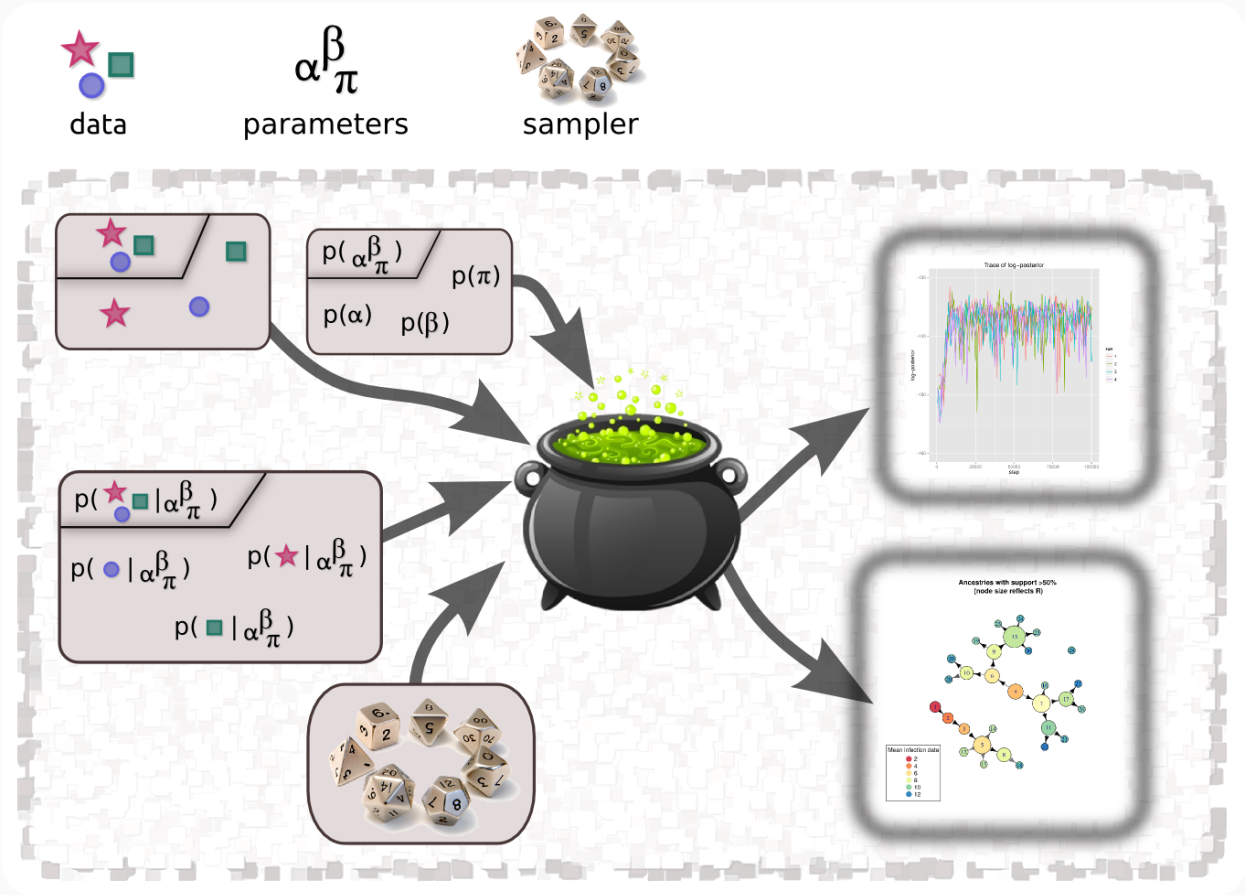
outbreaker2: a general cauldron for cooking methods

Use-your-own: data type, likelihood, prior, MCMC.



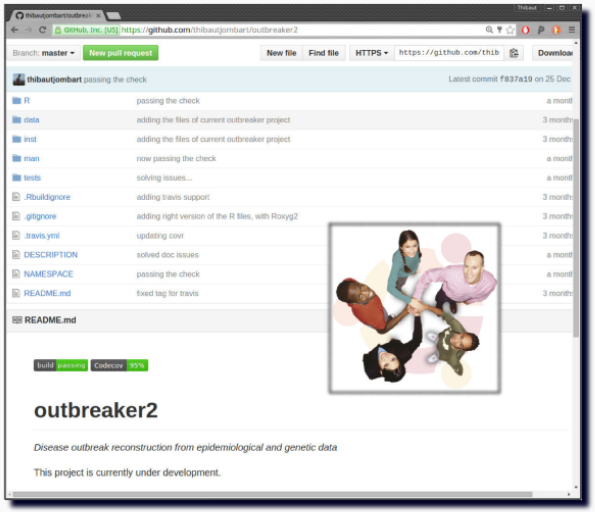
outbreaker2: a general cauldron for cooking methods

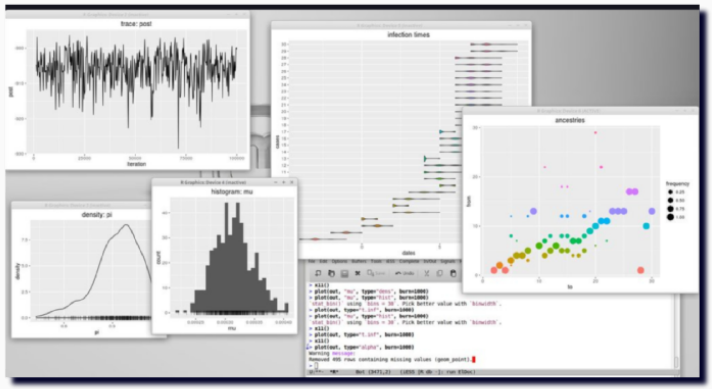
Use-your-own: data type, likelihood, prior, MCMC.



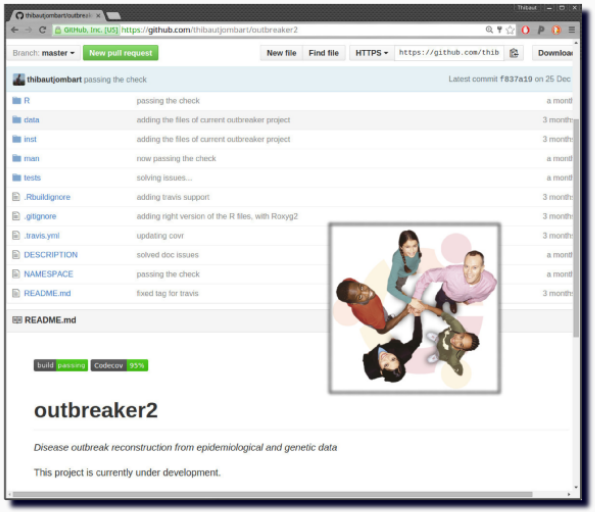
Modularity is key to generalising approaches

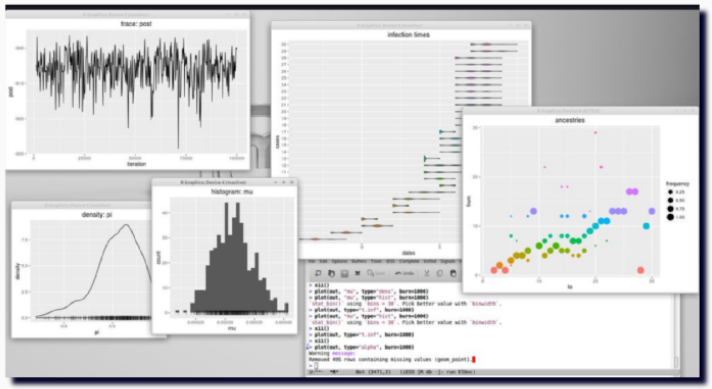
outbreaker2: a general tool for outbreak reconstruction





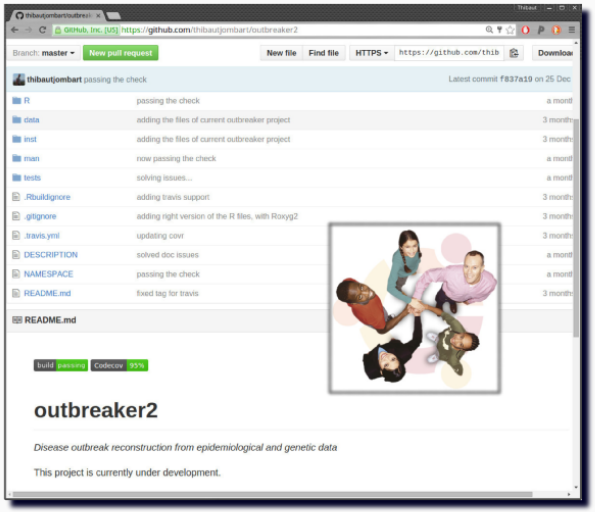
outbreaker2: a general tool for outbreak reconstruction

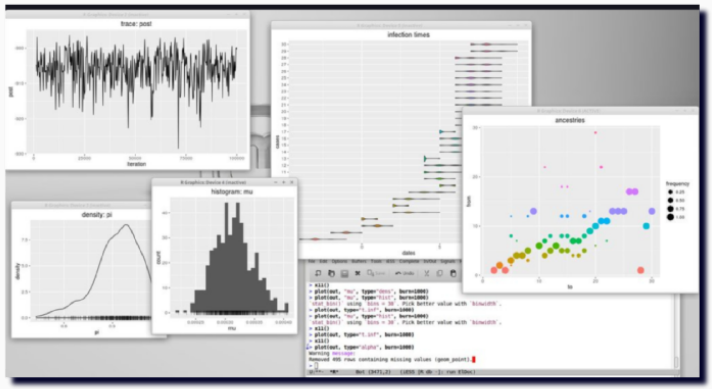




- **modularity**: likelihood, priors, samplers are all modules

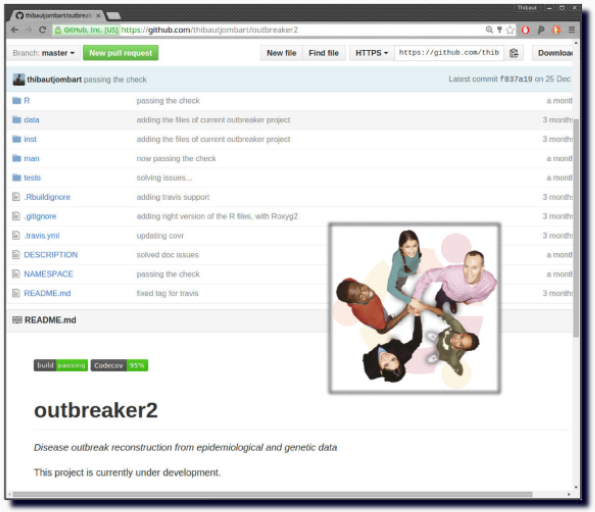
outbreaker2: a general tool for outbreak reconstruction

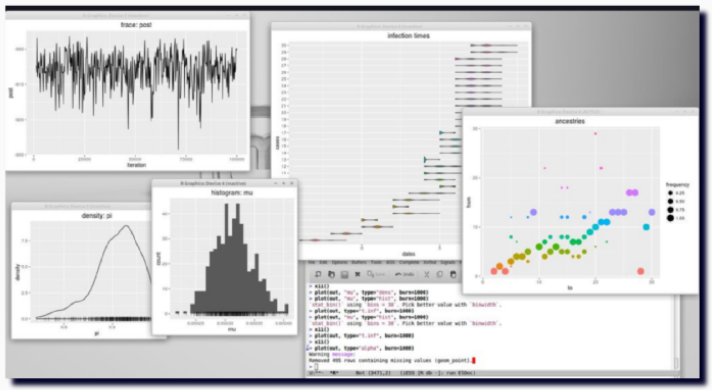




- **modularity**: likelihood, priors, samplers are all modules
- **new 'extensions'**: contact tracing, spatial structure, new MCMC

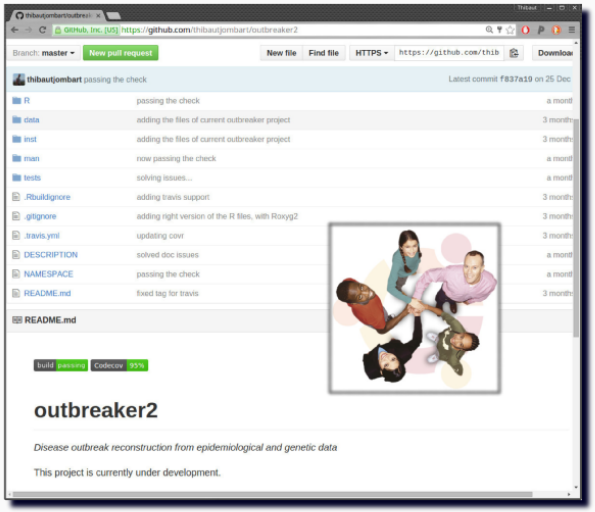
outbreaker2: a general tool for outbreak reconstruction

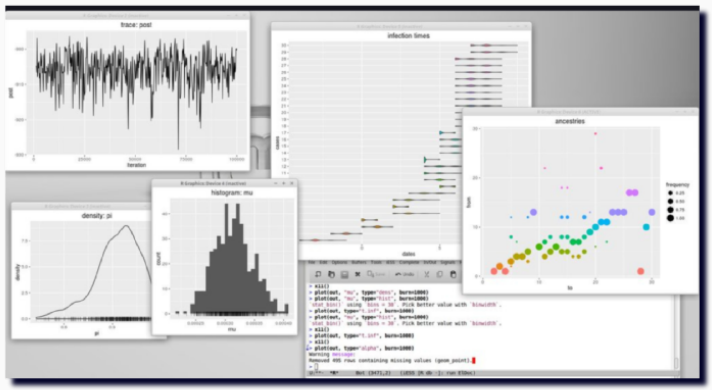




- **modularity**: likelihood, priors, samplers are all modules
- **new 'extensions'**: contact tracing, spatial structure, new MCMC
- **reliability**: continuous integration, extensive unit testing (aiming for 100% coverage)

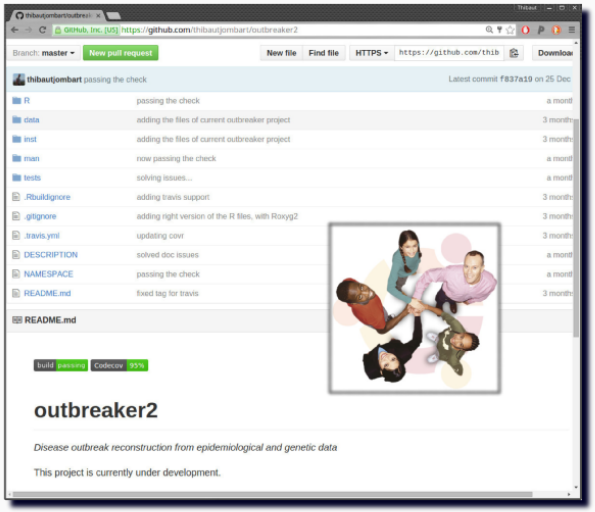
outbreaker2: a general tool for outbreak reconstruction

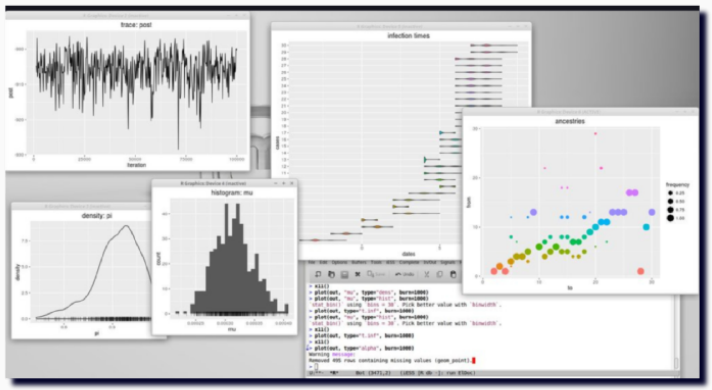




- **modularity**: likelihood, priors, samplers are all modules
- **new 'extensions'**: contact tracing, spatial structure, new MCMC
- **reliability**: continuous integration, extensive unit testing (aiming for 100% coverage)
- **prettier**: plot methods using *ggplot2*, interactive networks visualisation

outbreaker2: a general tool for outbreak reconstruction

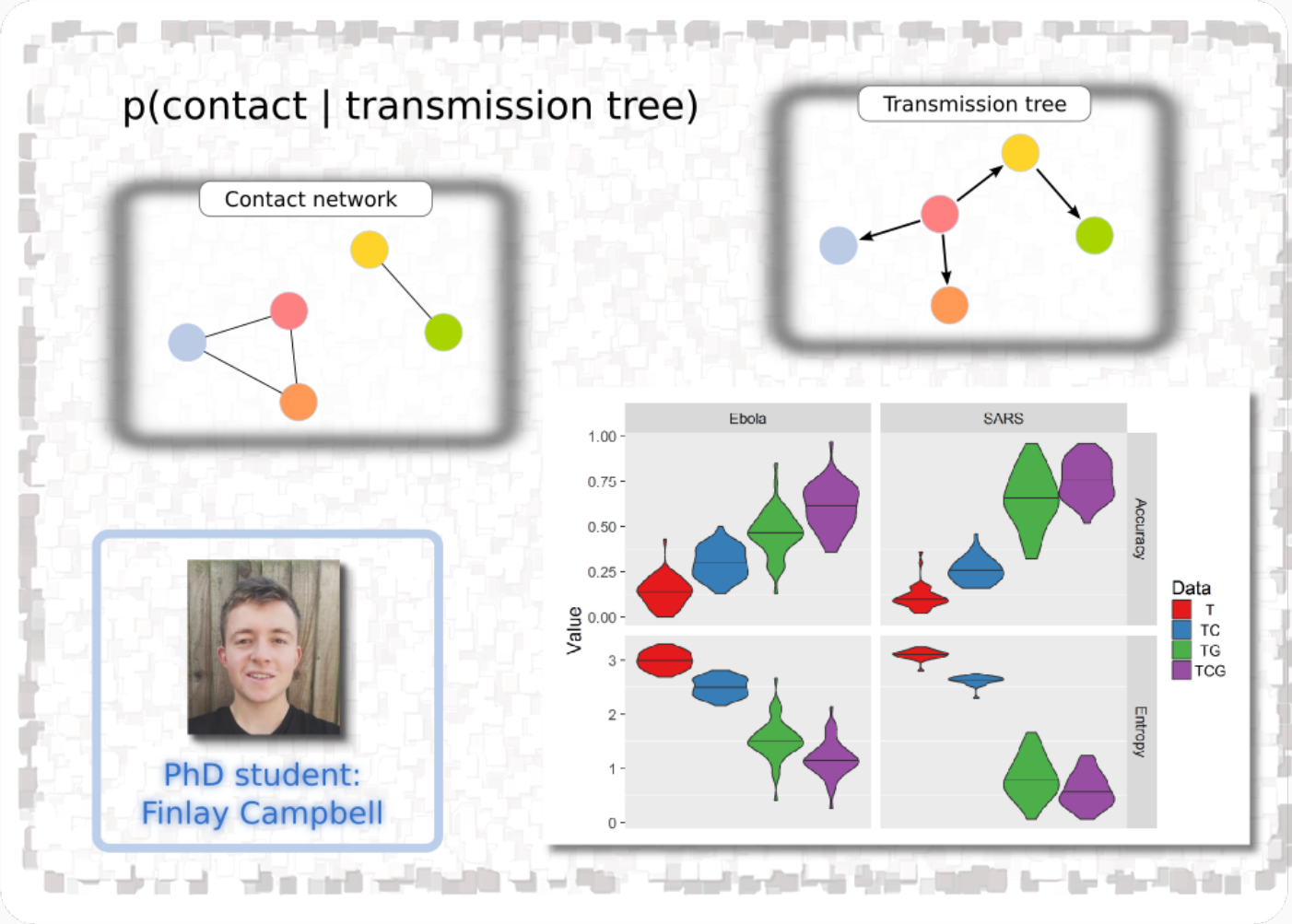




- **modularity**: likelihood, priors, samplers are all modules
- **new 'extensions'**: contact tracing, spatial structure, new MCMC
- **reliability**: continuous integration, extensive unit testing (aiming for 100% coverage)
- **prettier**: plot methods using *ggplot2*, interactive networks visualisation
- should facilitate new contributions

outbreaker2 developments: contact data

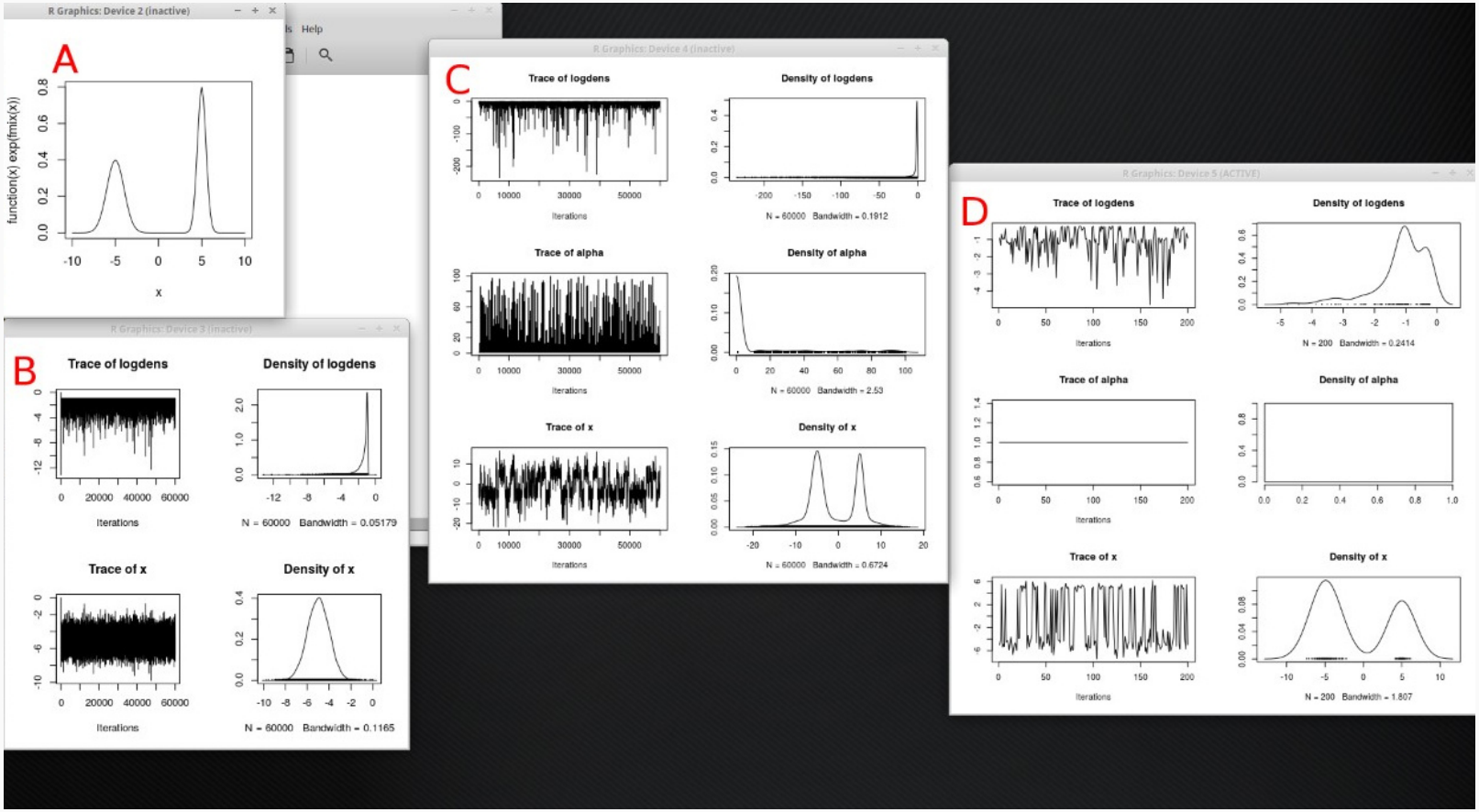
Integration of contact data for improving tree reconstruction.



[collaboration with Neil Ferguson & Anne Cori]

outbreaker2 developments: eMCMC

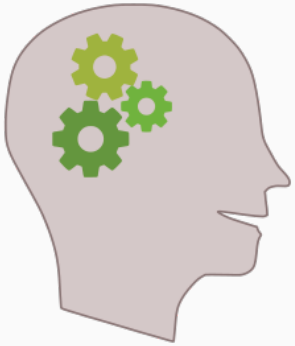
Eruptive MCMC: a new MCMC to sample from multi-modal densities



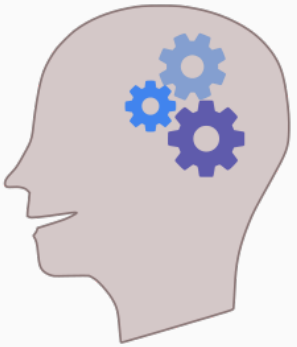
[collaboration with Jukka Corander]

Methodological dialogue

Methodological development relies on an interdisciplinary dialogue

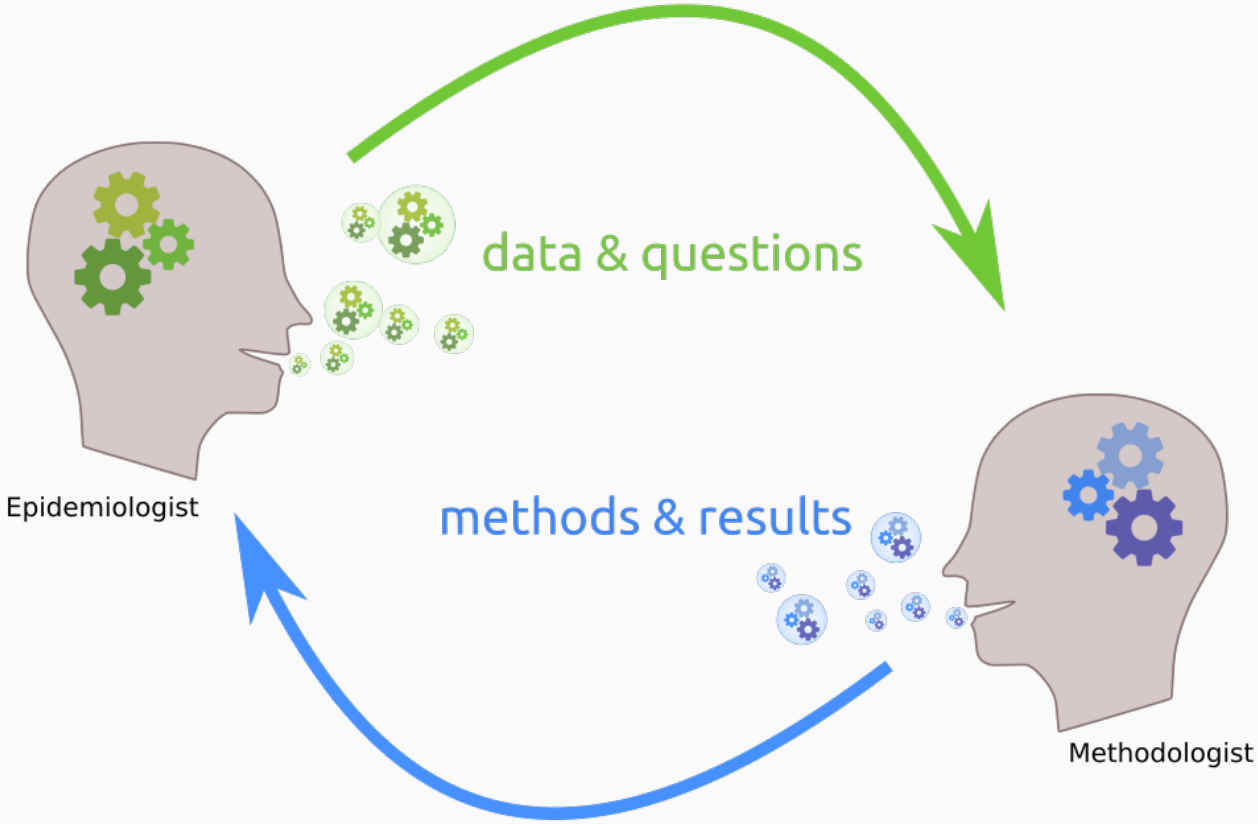


Epidemiologist

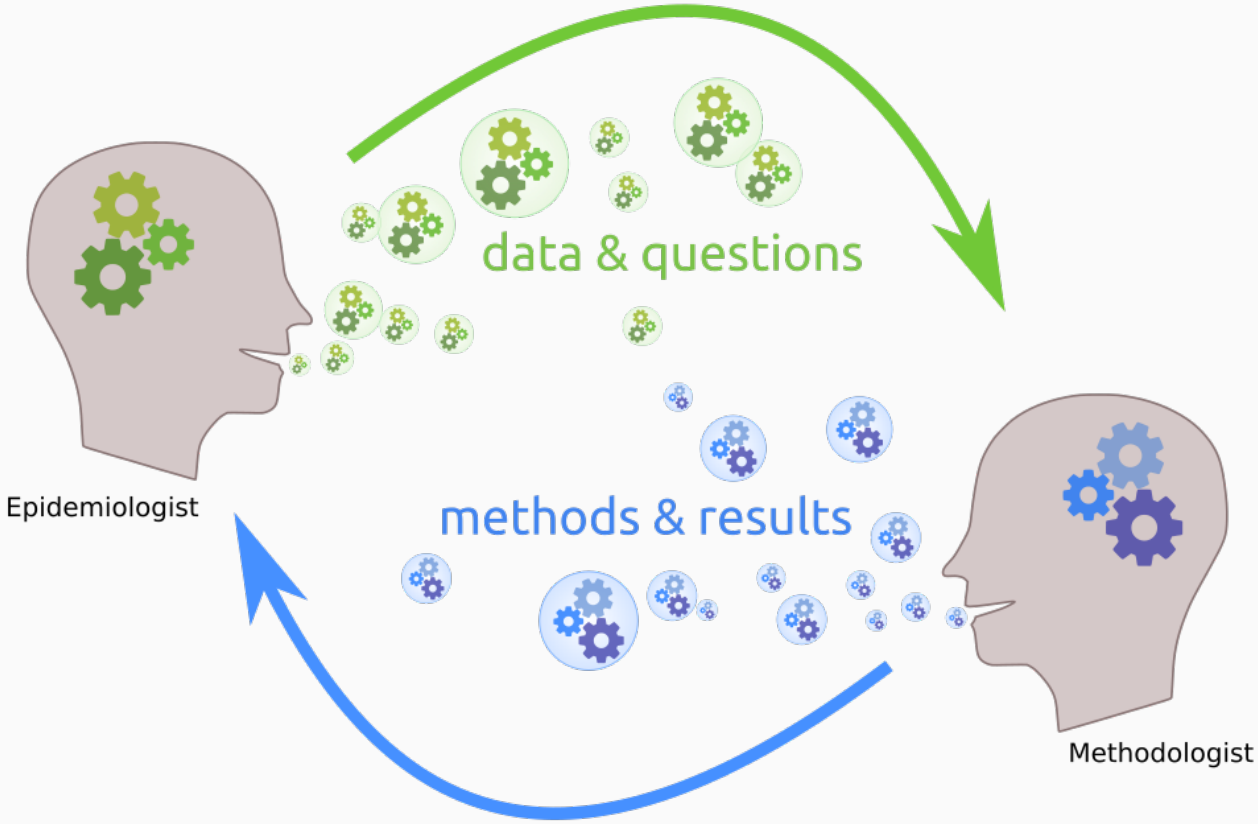


Methodologist

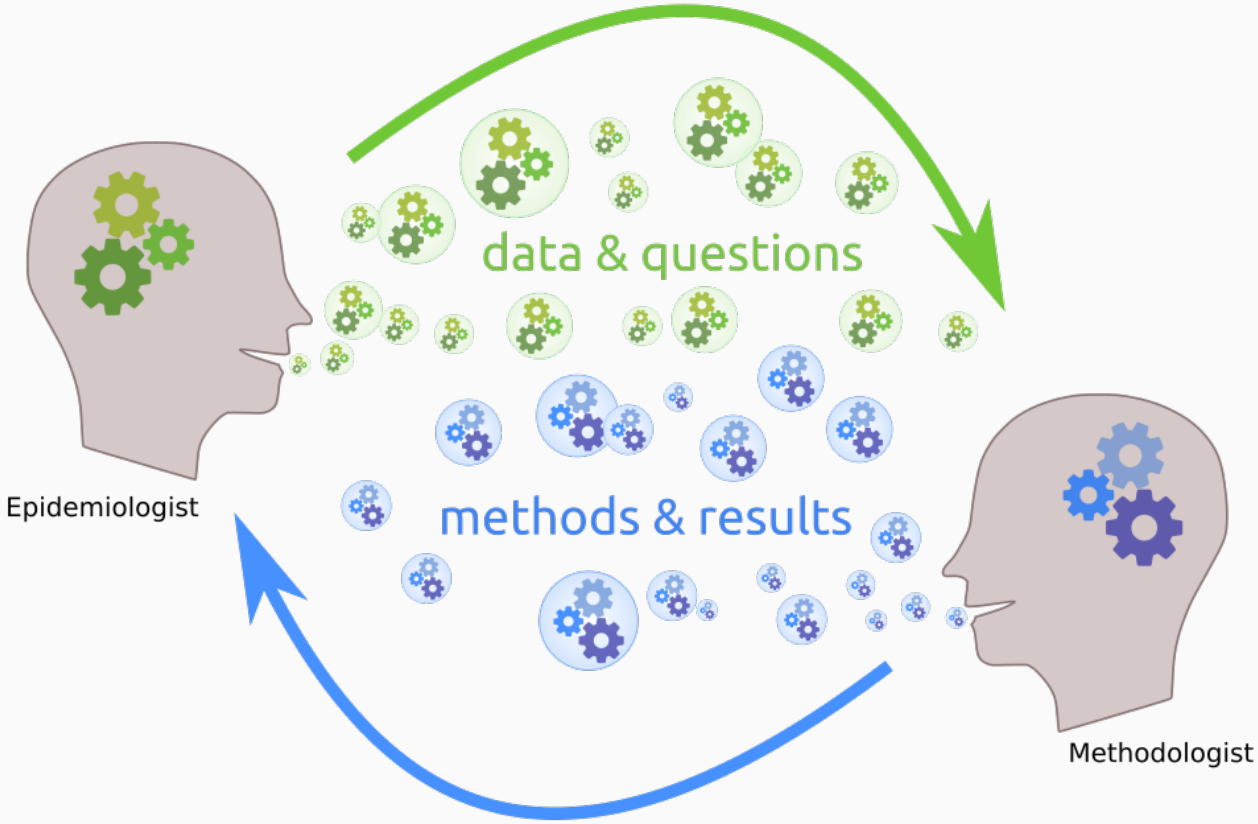
Methodological development relies on an interdisciplinary dialogue



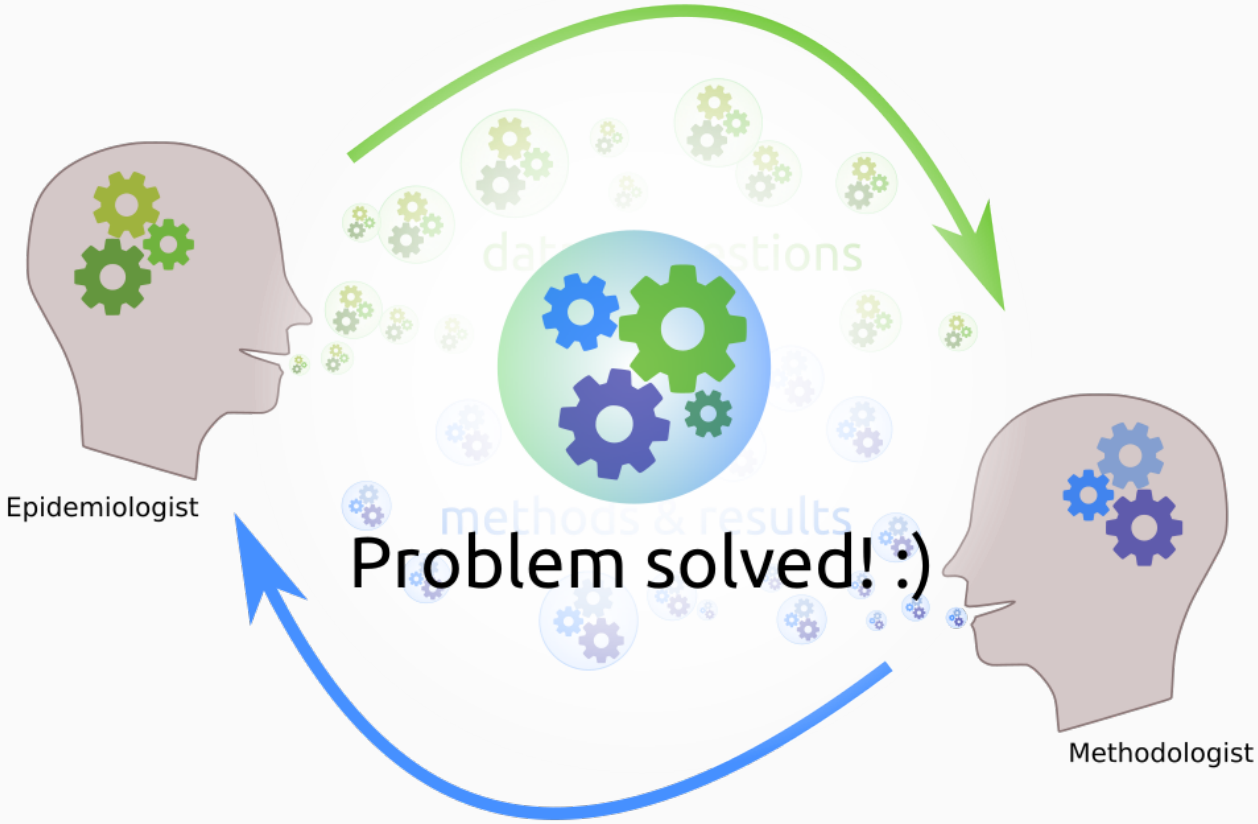
Methodological development relies on an interdisciplinary dialogue



Methodological development relies on an interdisciplinary dialogue



Methodological development relies on an interdisciplinary dialogue



Outbreak response context creates distance and delays

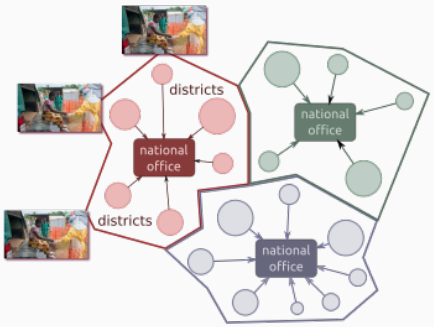
Affected countries

data collection

Outbreak response context creates distance and delays

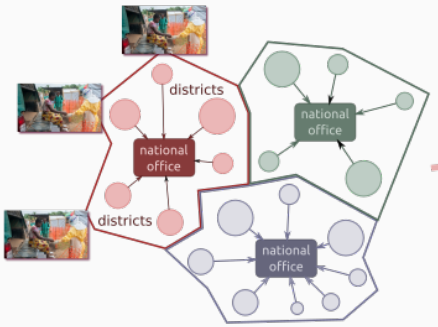
Affected countries

Outbreak response context creates distance and delays



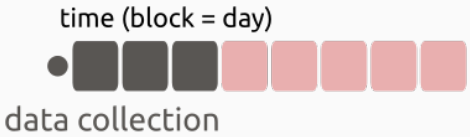
Affected countries

Outbreak response context creates distance and delays

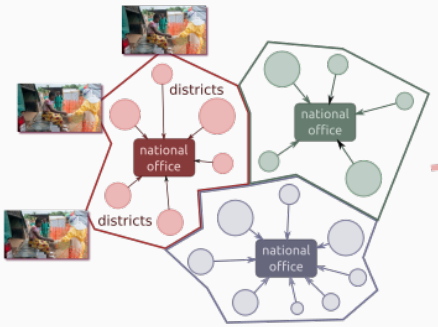


Affected countries

Public Health agencies



Outbreak response context creates distance and delays



Affected countries

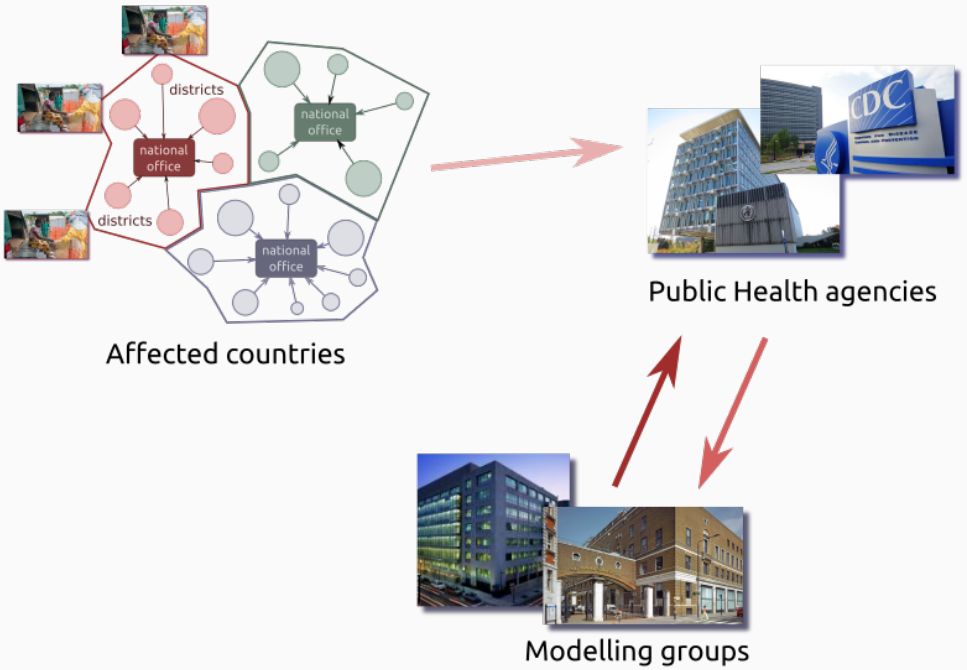
Public Health agencies

Modelling groups

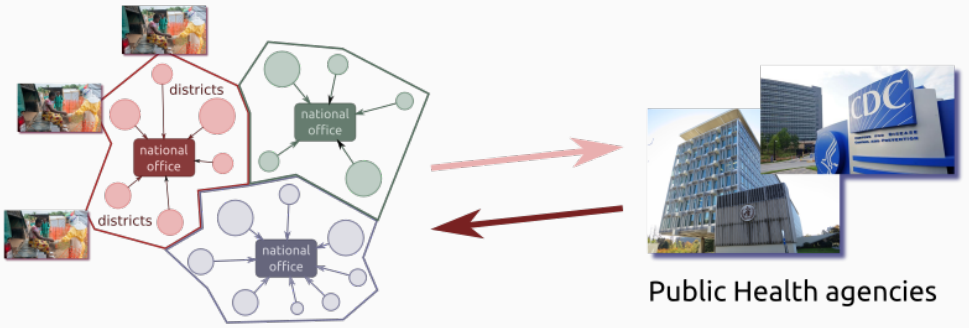
time (block = day)

data collection

Outbreak response context creates distance and delays



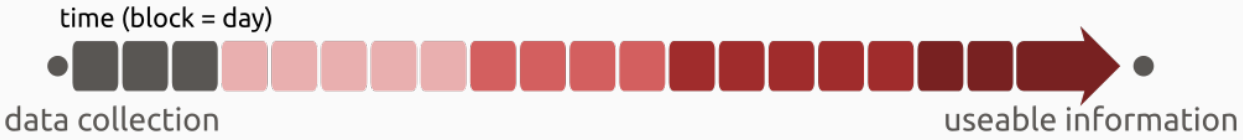
Outbreak response context creates distance and delays



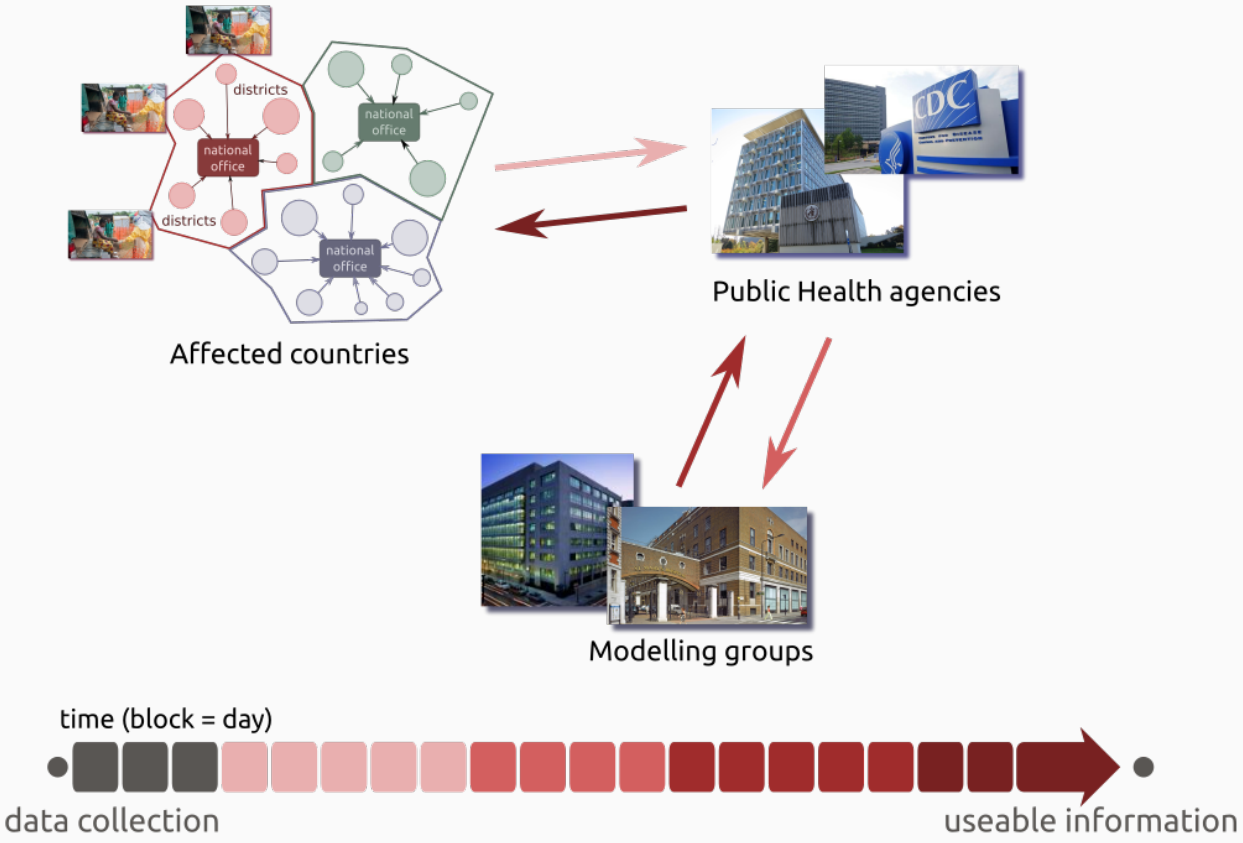
Affected countries

Public Health agencies

Modelling groups

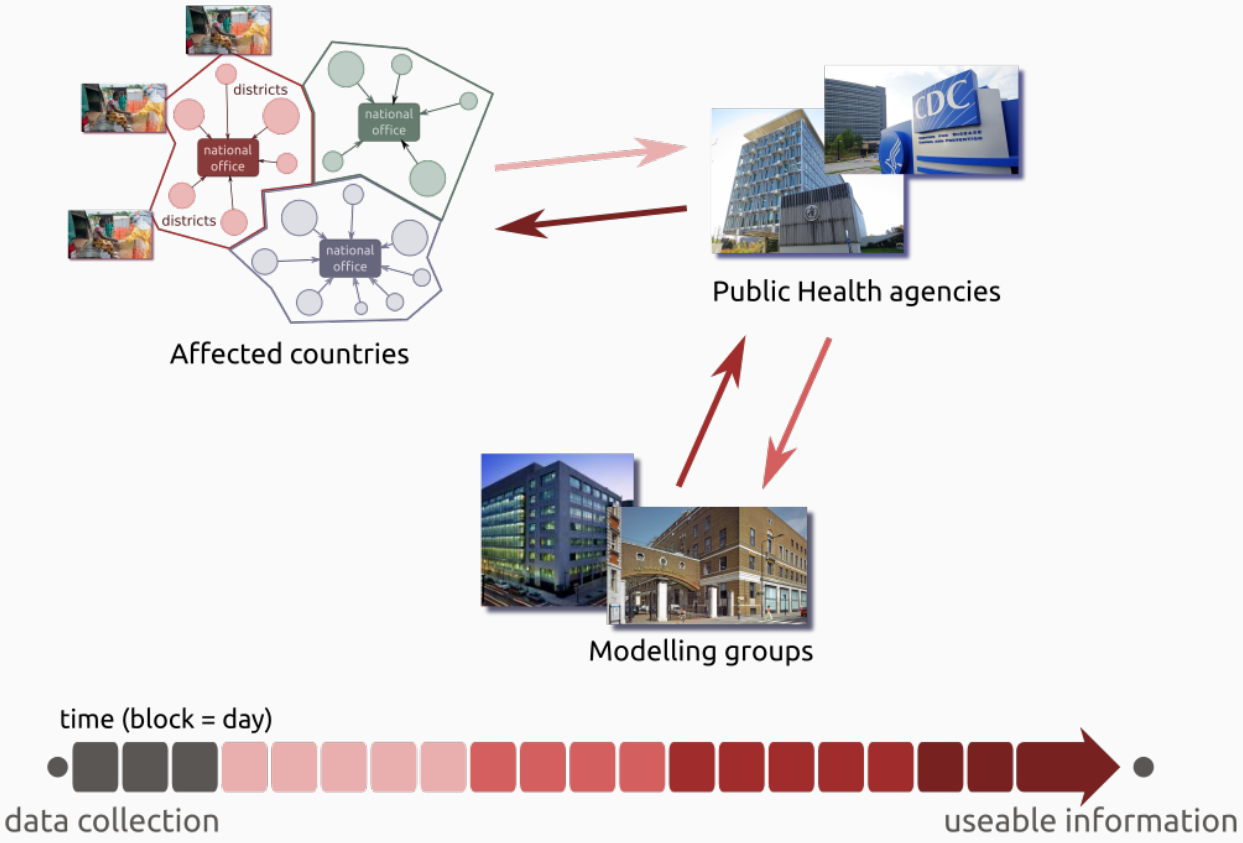


Outbreak response context creates distance and delays



- efficient tools can shorten delays

Outbreak response context creates distance and delays



- efficient tools can shorten delays
- potential of embedding methodologists in outbreak response teams

Summary

- outbreak response creates its own methodological field

Summary

- outbreak response creates its own methodological field
- approaches include: linear modelling, time series analysis, spatial modelling, likelihood-based / Bayesian approaches, graph theory, genetics, ...

Summary

- outbreak response creates its own methodological field
- approaches include: linear modelling, time series analysis, spatial modelling, likelihood-based / Bayesian approaches, graph theory, genetics, ...
- emergency context puts emphasis on reliability and reproducibility

Summary

- outbreak response creates its own methodological field
- approaches include: linear modelling, time series analysis, spatial modelling, likelihood-based / Bayesian approaches, graph theory, genetics, ...
- emergency context puts emphasis on reliability and reproducibility
- RECON is a joint effort for addressing these challenges through free, open-source statistical software development

Summary

- outbreak response creates its own methodological field
- approaches include: linear modelling, time series analysis, spatial modelling, likelihood-based / Bayesian approaches, graph theory, genetics, ...
- emergency context puts emphasis on reliability and reproducibility
- RECON is a joint effort for addressing these challenges through free, open-source statistical software development
- **We are recruiting!**
  Check 'news' on `www.repidemicsconsortium.org`

Thanks to...

- **Vincent Daubin**
- **Imperial College:** Neil Ferguson, Rich Fitzjohn, Anne Cori, Finlay Campbell, Evgenia Markvardt, James Hayward
- **UC Berkeley:** Karthik Ram
- **Groups:** WHO Ebola Response Team, Hackout 1/2/3, RECON members, GOARN
- **funding:** HPRU-NIHR, MRC

More on:

www.repidemicsconsortium.org

Questions?