

RECON

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

Thibaut Jombart

5th January 2017

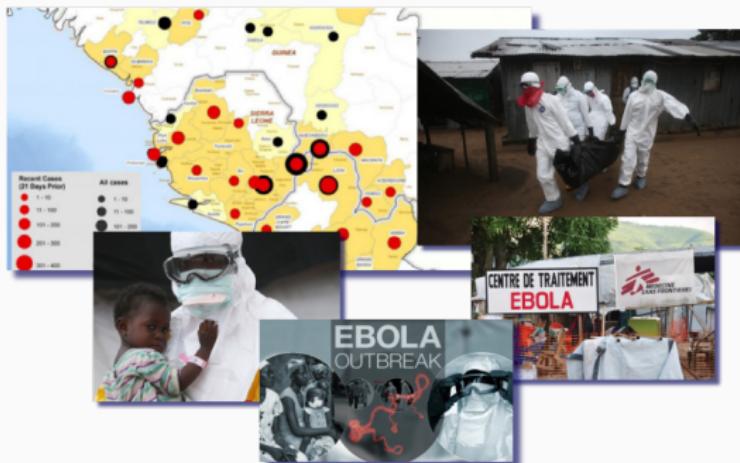
Imperial College London
MRC Centre for Outbreak Analysis and Modelling

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



Lessons learnt from the Ebola response

The image is a collage of various elements related to the Ebola response:

- A map of West Africa showing the locations of Ebola cases.
- A photograph of a WHO Ebola response team in protective suits.
- A photograph of a person in a white protective suit.
- A photograph of a medical facility with a sign that reads "CENTRE DE TRAITEMENT EBOLA".
- A graphic titled "WHO Ebola response team" with a subtitle "Help improving situation awareness".
- A timeline showing the progression of the outbreak: First case (December 2013), WHO notified (March 2014), First data/report (August 2014), and Latest data update (September 2015).
- Three documents or reports displayed vertically.
- A grid of 15 small portraits representing the 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, treatments effects
- **predictions:** case incidence, mortality, evaluate interventions
- **report:** (semi-)automated situation reports

Who do we need to develop these tools?



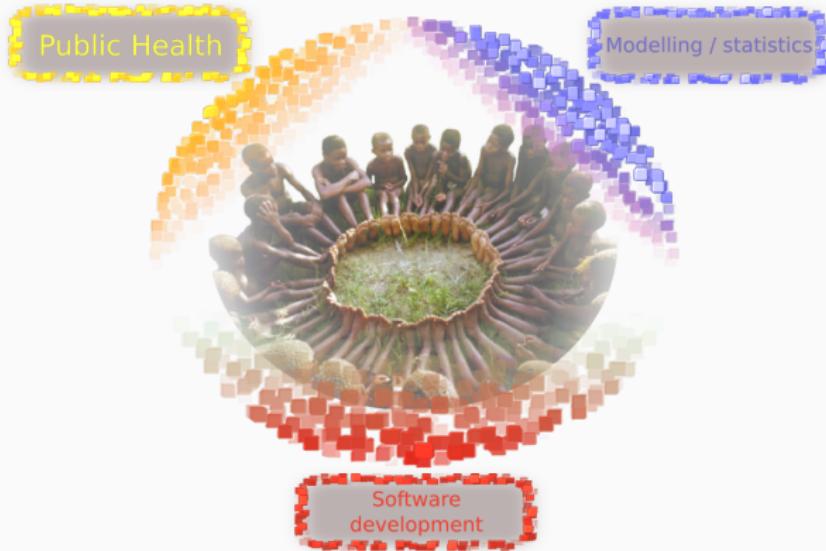
Who do we need to develop these tools?



Who do we need to develop these tools?



Who do we need to develop these tools?



The R Epidemics Consortium

Hackout 3: a hackathon for emergency outbreak response

Last summer at the *rOpenSci* headquarters (Berkeley)



Hackout 3: from ideas to projects to...



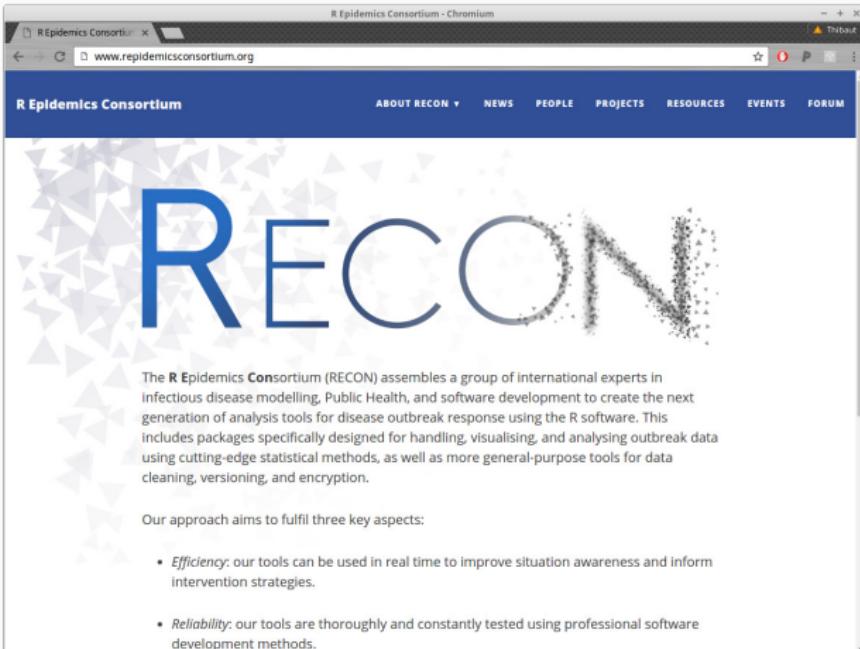
Hackout 3: from ideas to projects to...



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 .



The screenshot shows the homepage of the RECON website. The header features the text "R Epidemics Consortium" and "RECON". The main title "RECON" is prominently displayed in large blue letters, with each letter composed of a cluster of small triangles. Below the title is a descriptive paragraph about the consortium's purpose and approach. A sidebar on the left lists three key aspects: Efficiency, Reliability, and Transparency.

R Epidemics Consortium - Chromium
www.repidemicsconsortium.org

R Epidemics Consortium

ABOUT RECON NEWS PEOPLE PROJECTS RESOURCES EVENTS FORUM

RECON

The R Epidemics Consortium (RECON) assembles a group of international experts in infectious disease modelling, Public Health, and software development to create the next generation of analysis tools for disease outbreak response using the R software. This includes packages specifically designed for handling, visualising, and analysing outbreak data using cutting-edge statistical methods, as well as more general-purpose tools for data cleaning, versioning, and encryption.

Our approach aims to fulfil three key aspects:

- *Efficiency*: our tools can be used in real time to improve situation awareness and inform intervention strategies.
- *Reliability*: our tools are thoroughly and constantly tested using professional software development methods.

www.repidemicsconsortium.org

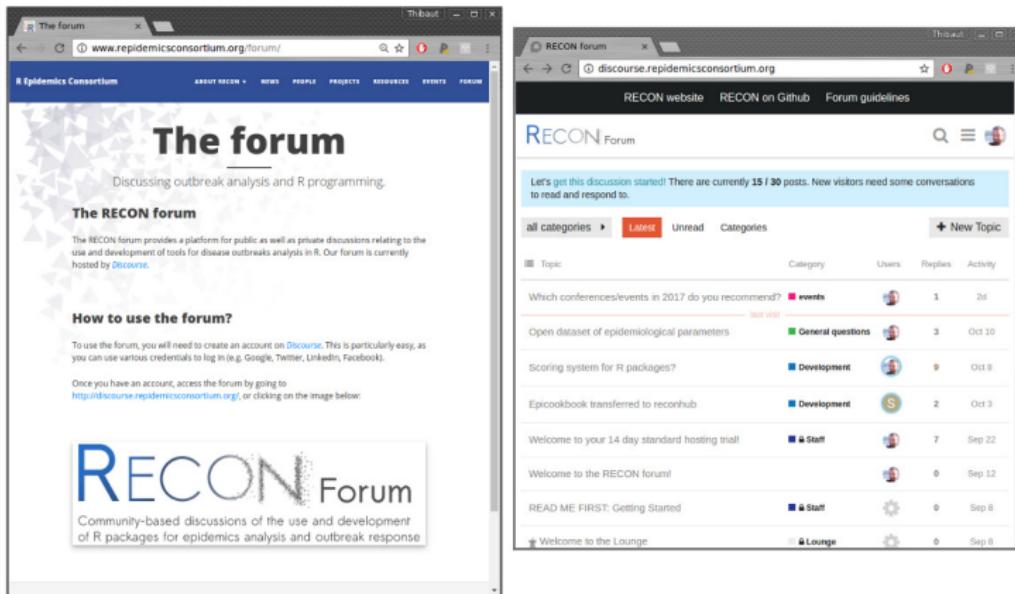
RECON

www.repidemicsconsortium.org

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

The RECON forum

A platform for discussing epidemics analysis in .



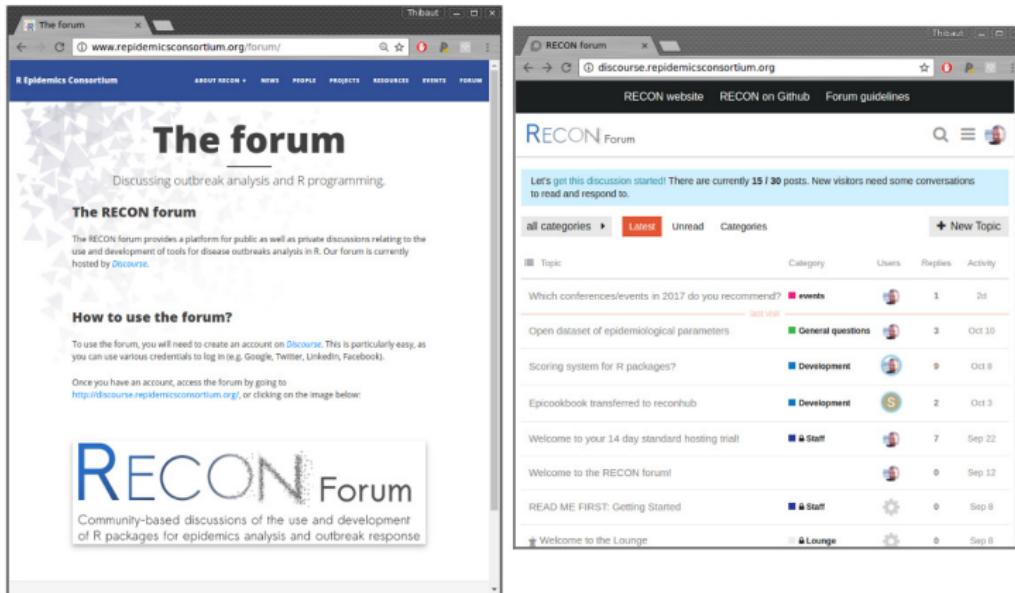
The image shows two side-by-side browser windows. The left window displays the official RECON forum website at www.repidemicsconsortium.org/forum/. It features a dark blue header with the 'Epidemics Consortium' logo and navigation links for About RECON, News, People, Projects, Resources, Events, and Forum. The main content area has a light gray background with a geometric pattern and contains the heading 'The forum' and a sub-section titled 'The RECON forum'. The right window shows a Discourse instance at discourse.repidemicsconsortium.org. The header includes links to the RECON website, GitHub, and Forum guidelines. The main area is titled 'RECON Forum' and shows a list of forum topics. A message at the top encourages users to start conversations. Below it, a table lists various topics with columns for Topic, Category, Users, Replies, and Activity.

Topic	Category	Users	Replies	Activity
Which conferences/events in 2017 do you recommend?	events	2	1	Oct 10
Open dataset of epidemiological parameters	General questions	3	0	Oct 10
Scoring system for R packages?	Development	9	0	Oct 9
Epicookbook transferred to reconhub	Development	2	0	Oct 3
Welcome to your 14 day standard hosting trial!	Staff	7	0	Sep 22
Welcome to the RECON forum!	Staff	0	0	Sep 12
READ ME FIRST: Getting Started	Staff	0	0	Sep 8
Welcome to the Lounge	Lounge	0	0	Sep 8

www.repidemicsconsortium.org/forum

The RECON forum

A platform for discussing epidemics analysis in .



The image shows two side-by-side browser windows. The left window displays the official RECON forum at www.repidemicsconsortium.org/forum/. It features a dark blue header with the 'Epidemics Consortium' logo and navigation links for About RECON, News, People, Projects, Resources, Events, and Forum. The main content area has a light gray background with a geometric pattern and features a large title 'The forum' with a subtitle 'Discussing outbreak analysis and R programming.' Below this, there's a section titled 'The RECON forum' which provides a brief overview of the forum's purpose and its connection to Discourse. Another section, 'How to use the forum?', explains the process of creating an account on Discourse. At the bottom, there's a large 'RECON Forum' logo with the tagline 'Community-based discussions of the use and development of R packages for epidemics analysis and outbreak response'.

The right window shows the Discourse instance at discourse.repidemicsconsortium.org. It has a dark header with links to the RECON website, RECON on GitHub, and Forum guidelines. The main content area is titled 'RECON Forum'. It displays a list of forum topics, such as 'Which conferences/events in 2017 do you recommend?' and 'Open dataset of epidemiological parameters'. Each topic includes details like the category (e.g., events, General questions, Development, Staff, Lounge), the number of replies, and the date it was last updated.

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?

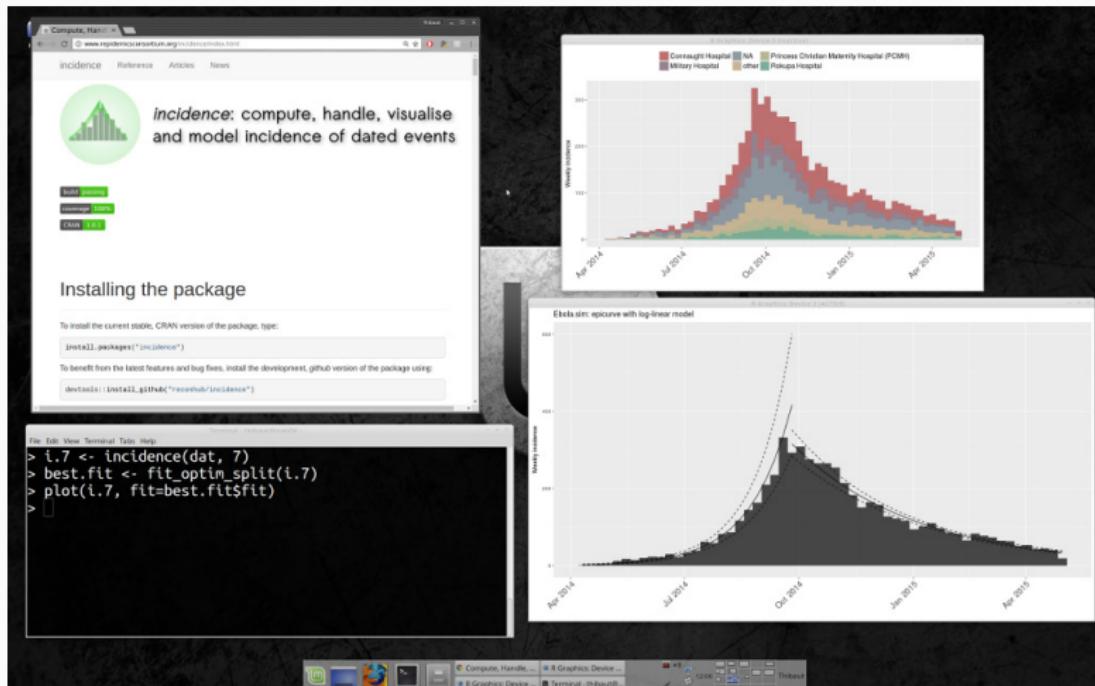
- **efficiency**: useful for improving situation awareness in real time; **cutting-edge, computer-efficient statistical methods**
- **reliability**: outputs can be trusted; **continuous integration, extensive unit testing, code review, good practices**

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

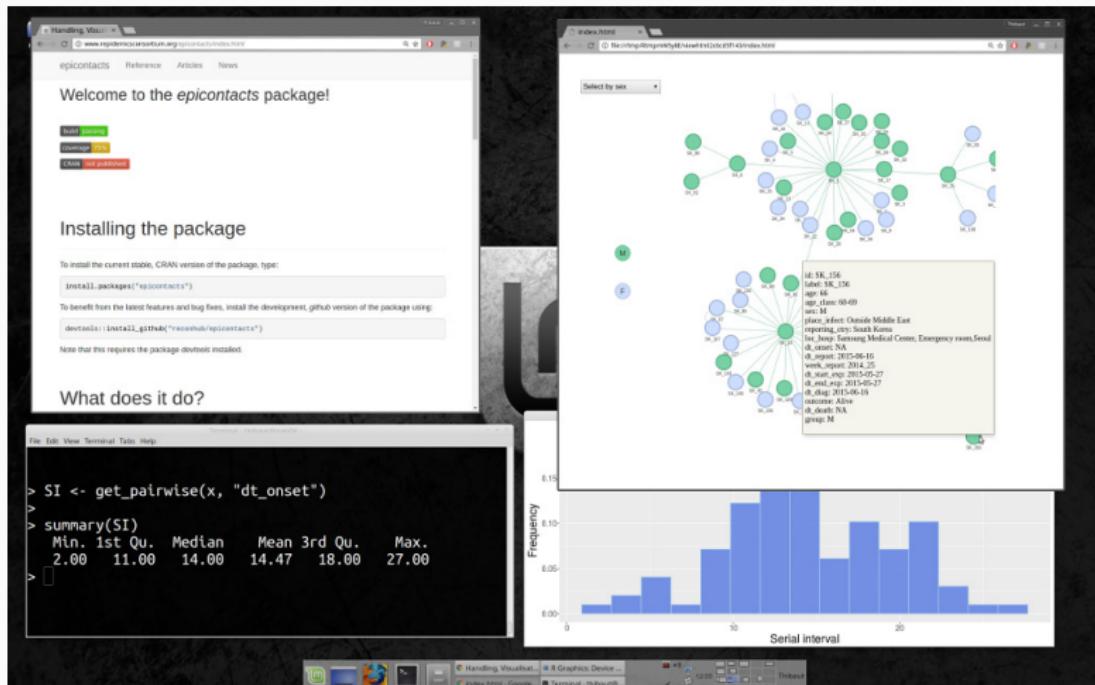
incidence: computation, handling, visualisation and modelling of epicurves



www.repidemicsconsortium.org/incidence

[released]

epicontacts: handling, visualisation and analysis of epidemiological contacts



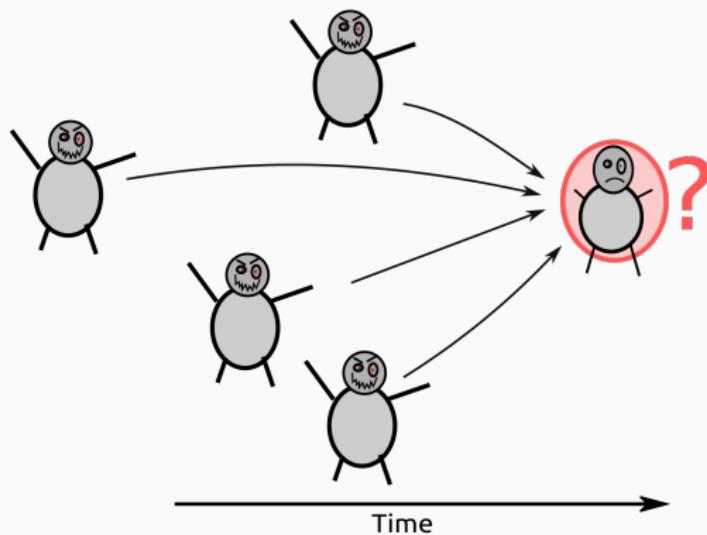
www.repidemicsconsortium.org/epicontacts

[release February 2017]

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

Use timing of symptoms and pathogen genomes to infer infectors

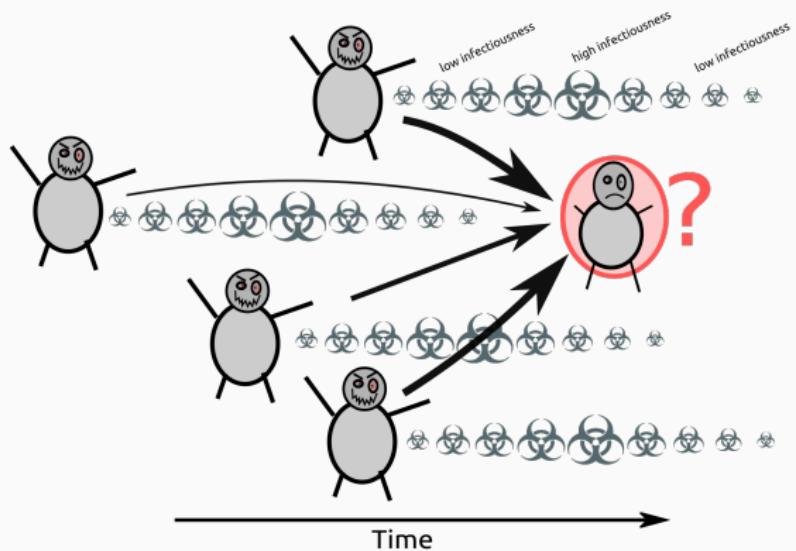
(Jombart et al, PLoS Comp Biol, 2014)



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

Use timing of symptoms and pathogen genomes to infer infectors

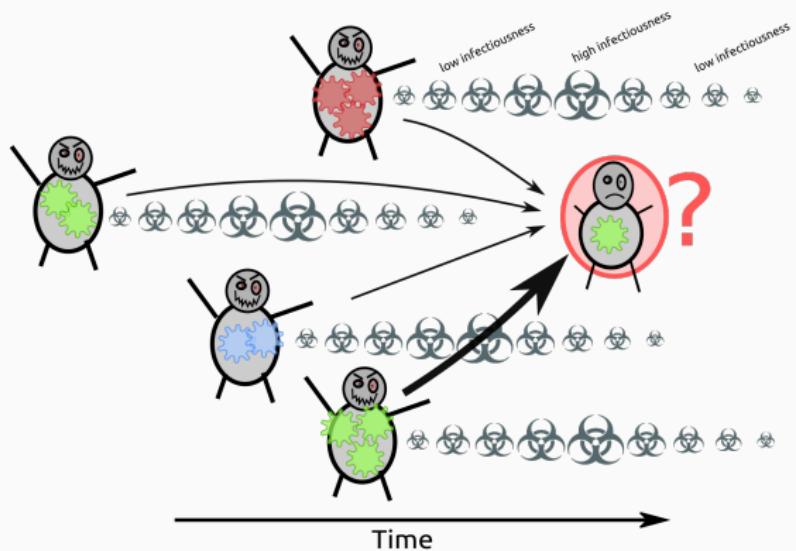
(Jombart et al, PLoS Comp Biol, 2014)



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

General approach

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

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

General approach

- Bayesian framework
- augmented data for ancestries, infection dates, unobserved cases
- MCMC (Metropolis-Hastings) 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}

Original implementation aimed at computer efficiency:

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

Original implementation aimed at computer efficiency:

- C implementation embedded within  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.

Original implementation aimed at computer efficiency:

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

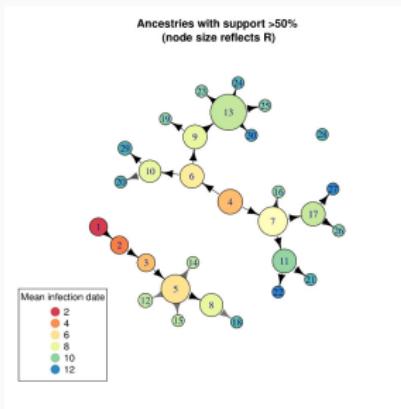
Fast, but **not very flexible**, hard to unit-test, hard to change code.

New methods emerged since.

Methods differ by their treatment of genetic data

Quick and dirty

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

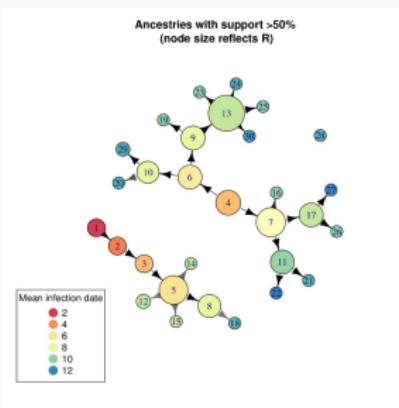


[outbreaker]

Methods differ by their treatment of genetic data

Quick and dirty

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



[outbreaker]

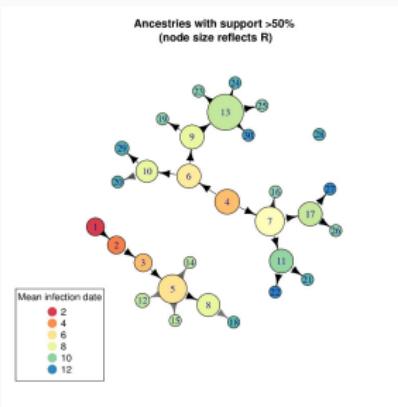
Nice and slow

- 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

Quick and dirty

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



[outbreaker]

Nice and slow

- within-host evolution
- ~ 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**.

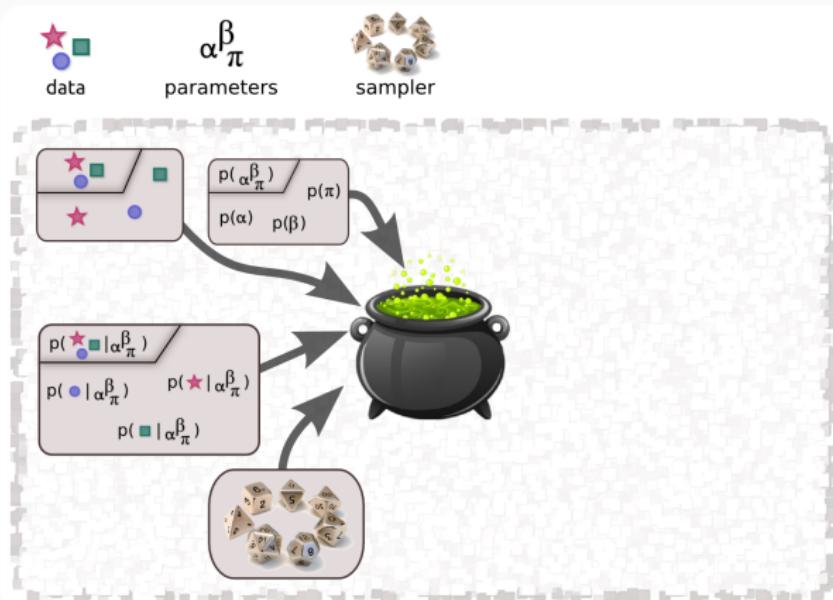
outbreaker2: a general cauldron for cooking methods

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



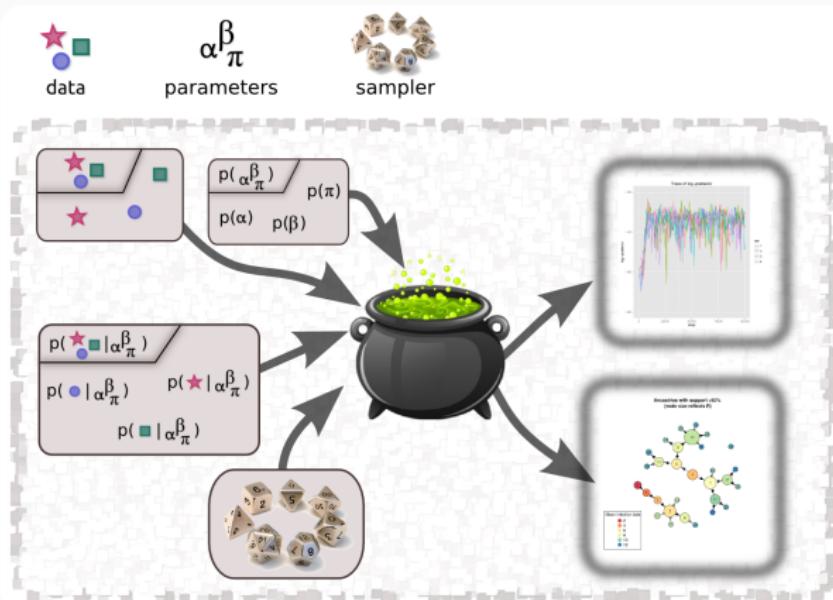
outbreaker2: a general cauldron for cooking methods

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



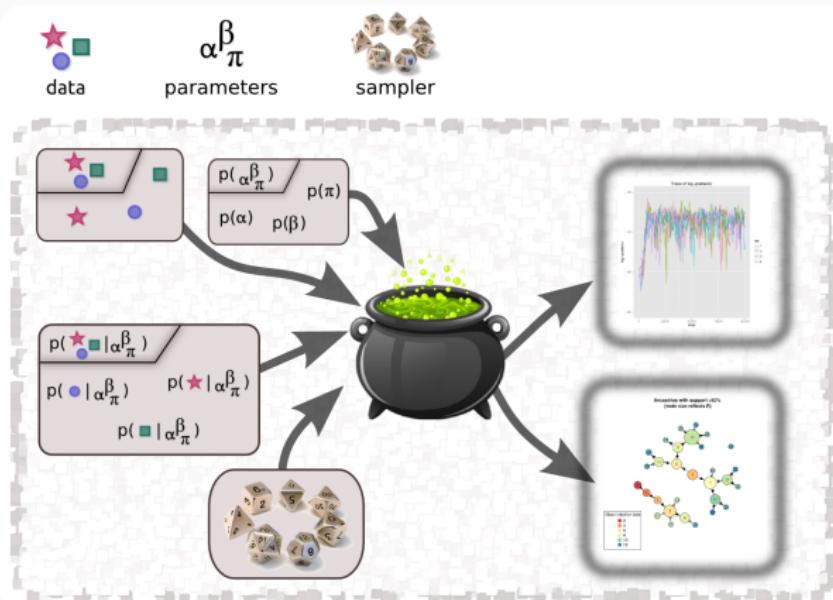
outbreaker2: a general cauldron for cooking methods

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



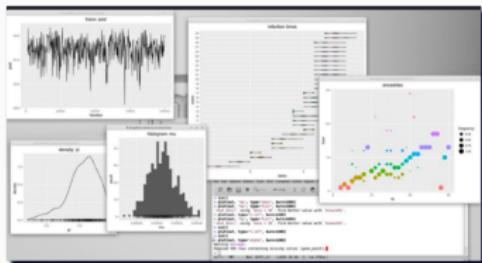
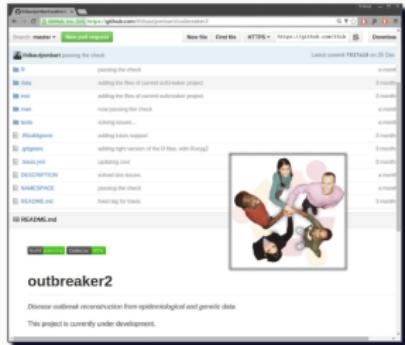
outbreaker2: a general cauldron for cooking methods

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



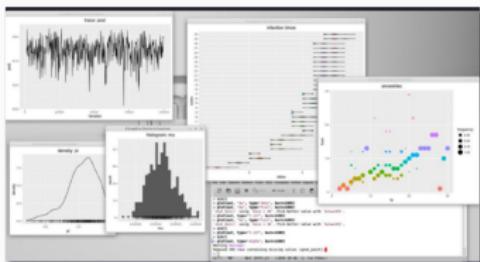
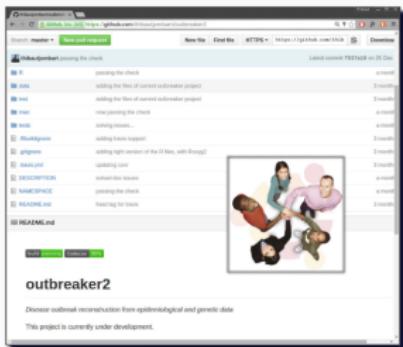
Modularity is key to generalising approaches

outbreaker2: a general tool for outbreak reconstruction

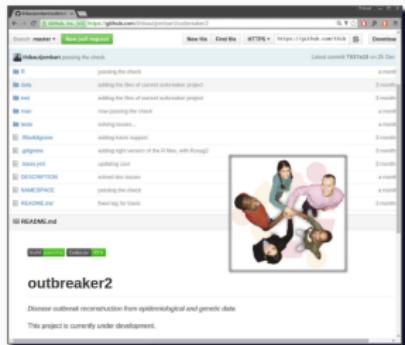


outbreaker2: a general tool for outbreak reconstruction

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



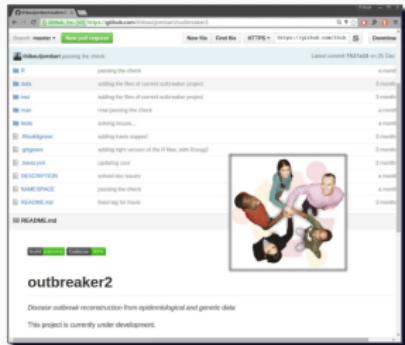
outbreaker2: a general tool for outbreak reconstruction



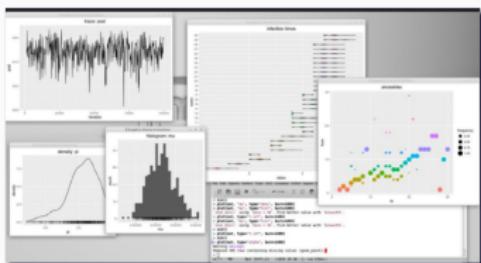
- **modularity:** likelihood, priors, samplers are all modules
- **new 'extensions':** contact tracing, spatial structure, new MCMC



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)

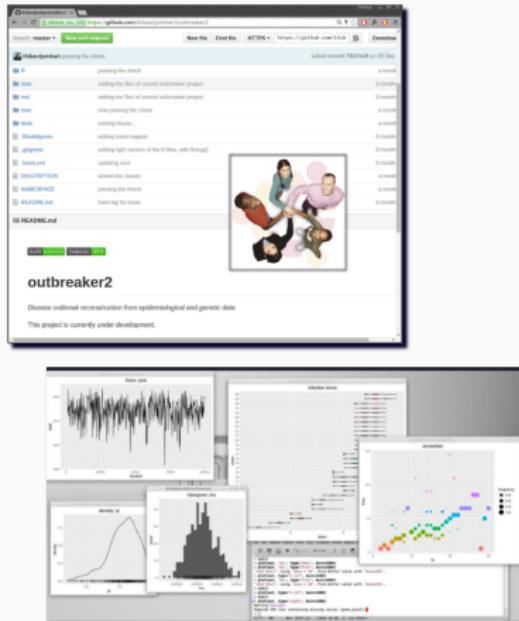


outbreaker2: a general tool for outbreak reconstruction



- **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

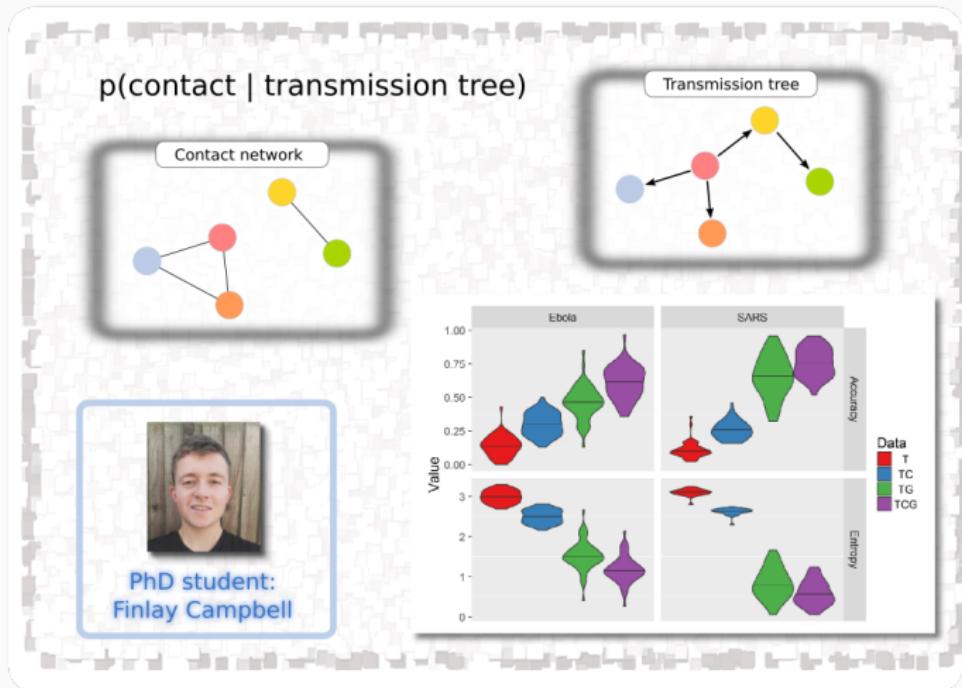
outbreaker2: a general tool for outbreak reconstruction



- **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**

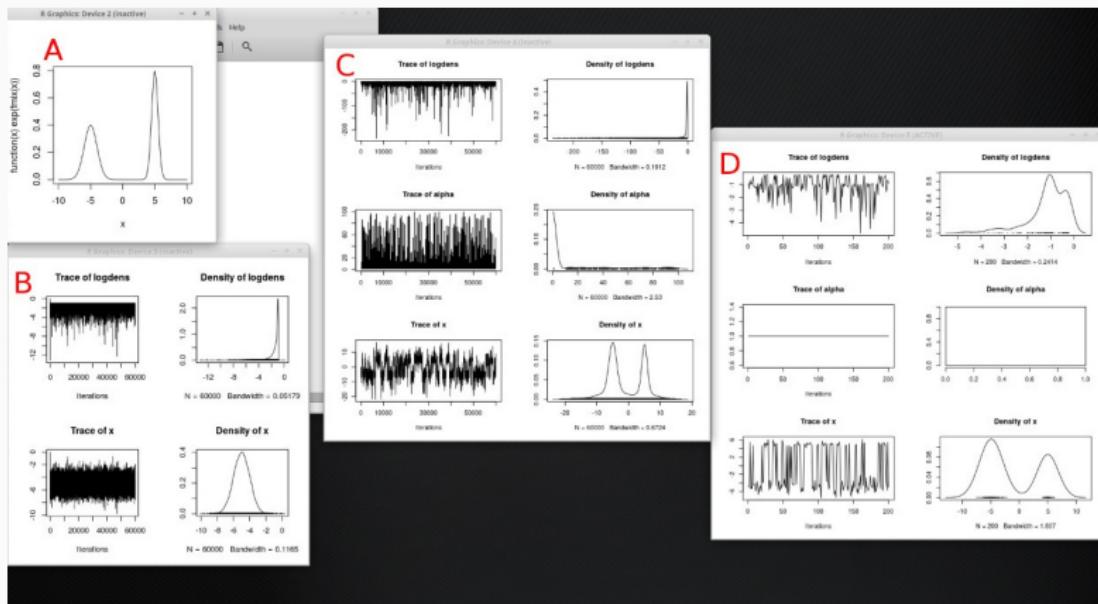
outbreaker2 developments: contact data

Integration of contact data for improving tree reconstruction.



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

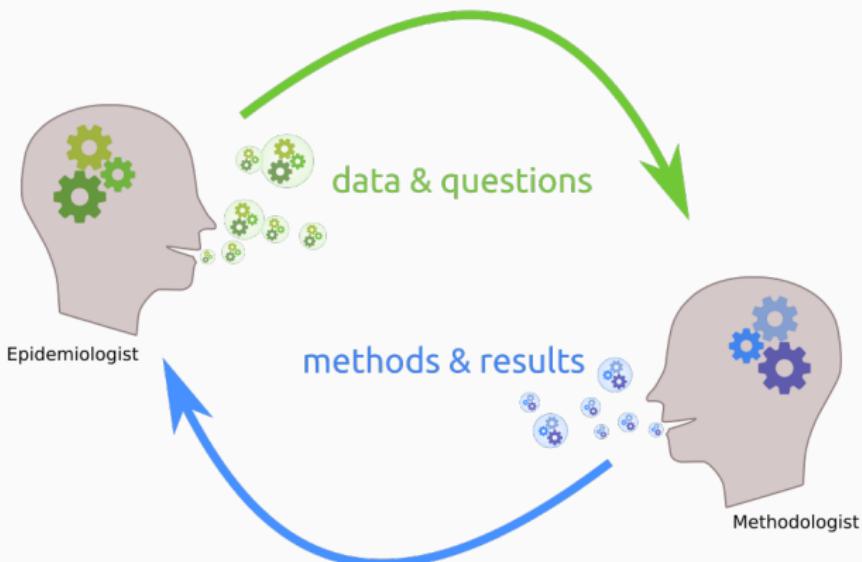


Epidemiologist

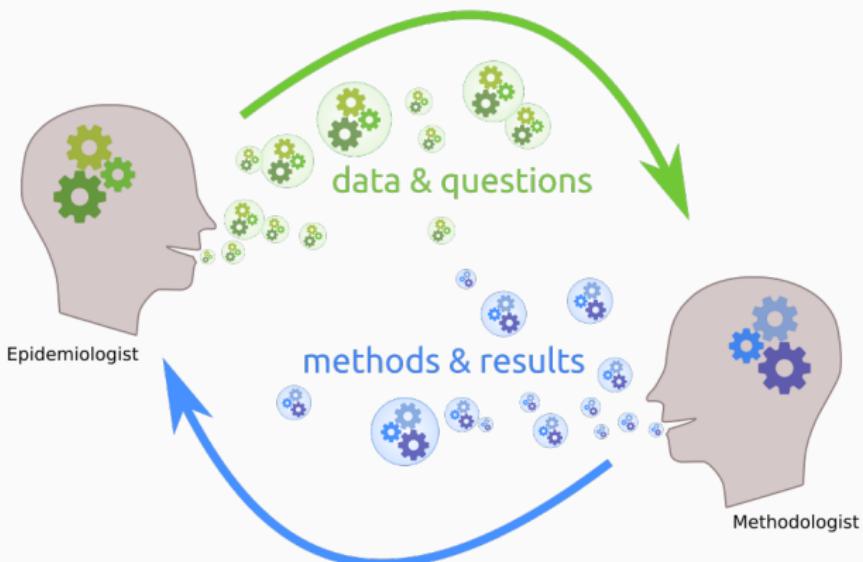


Methodologist

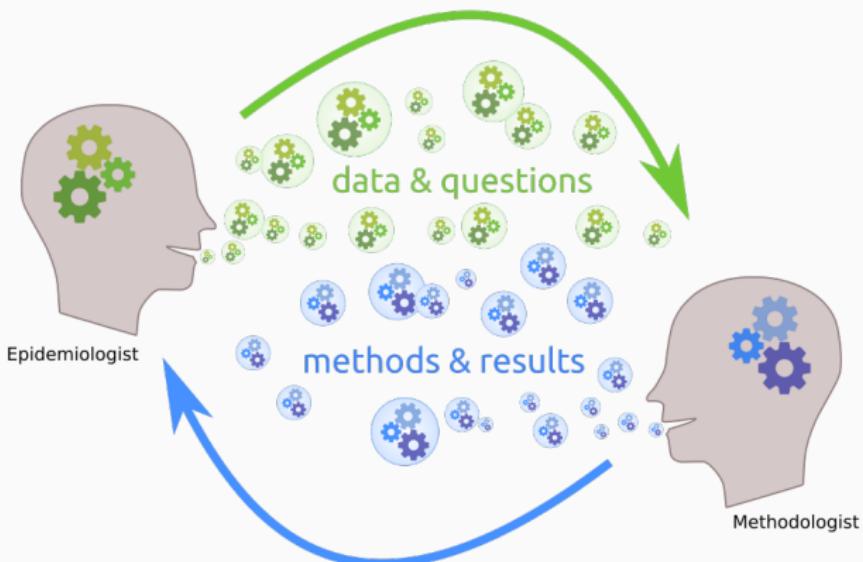
Methodological development relies on an interdisciplinary dialogue



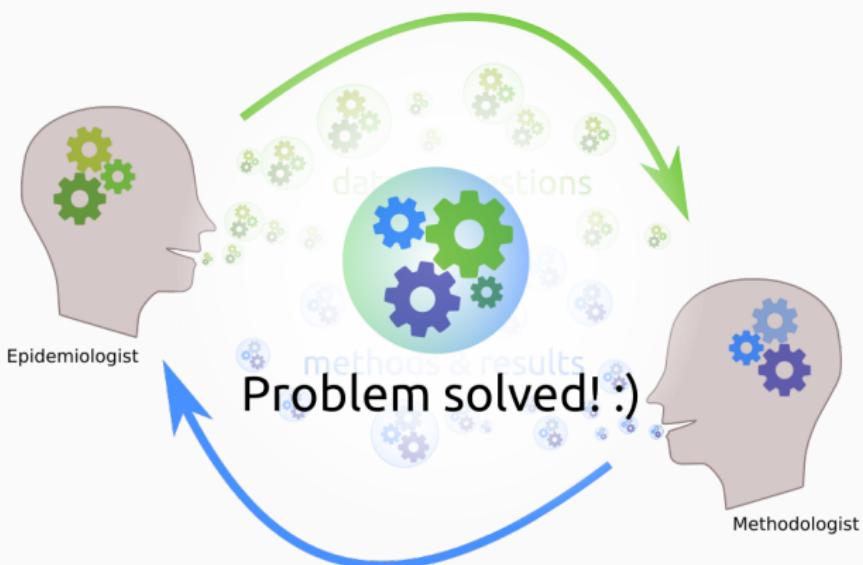
Methodological development relies on an interdisciplinary dialogue



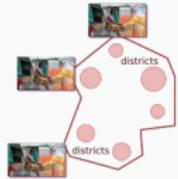
Methodological development relies on an interdisciplinary dialogue



Methodological development relies on an interdisciplinary dialogue



Outbreak response context creates distance and delays

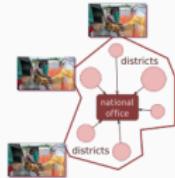


Affected countries



data collection

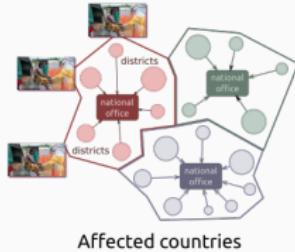
Outbreak response context creates distance and delays



Affected countries

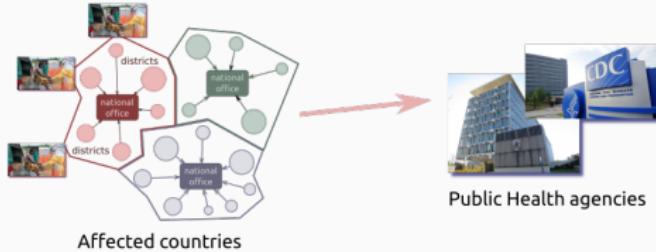
time (block = day)
• 
data collection

Outbreak response context creates distance and delays

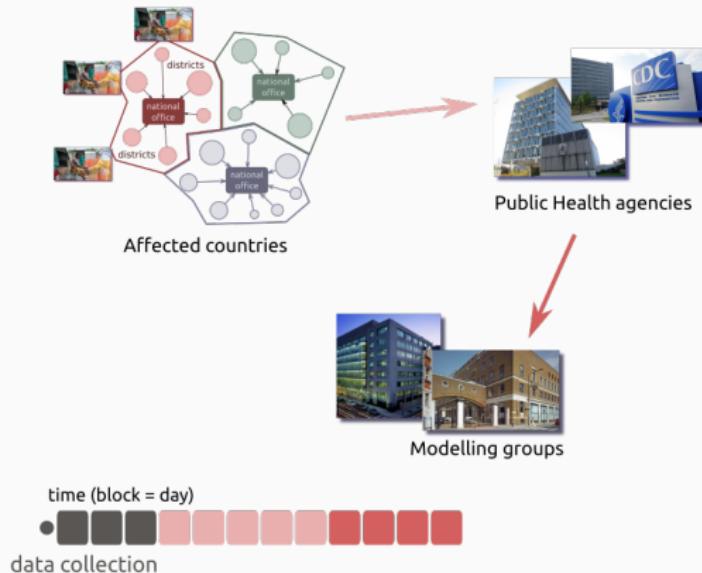


time (block = day)
● data collection

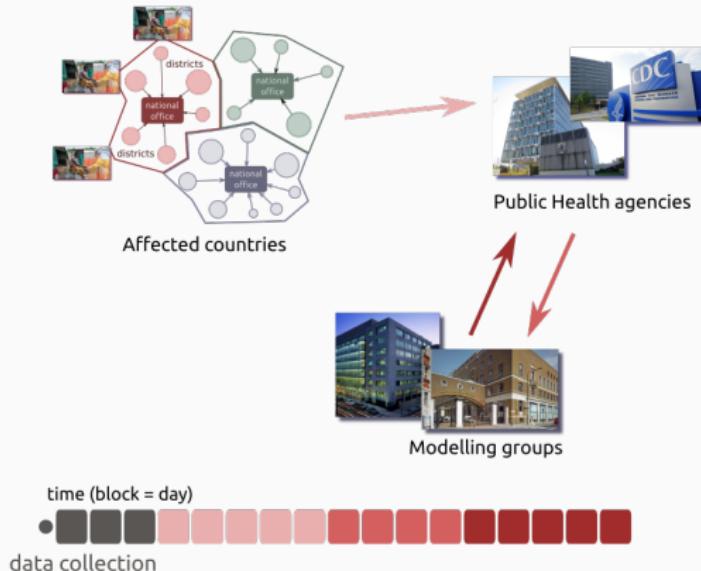
Outbreak response context creates distance and delays



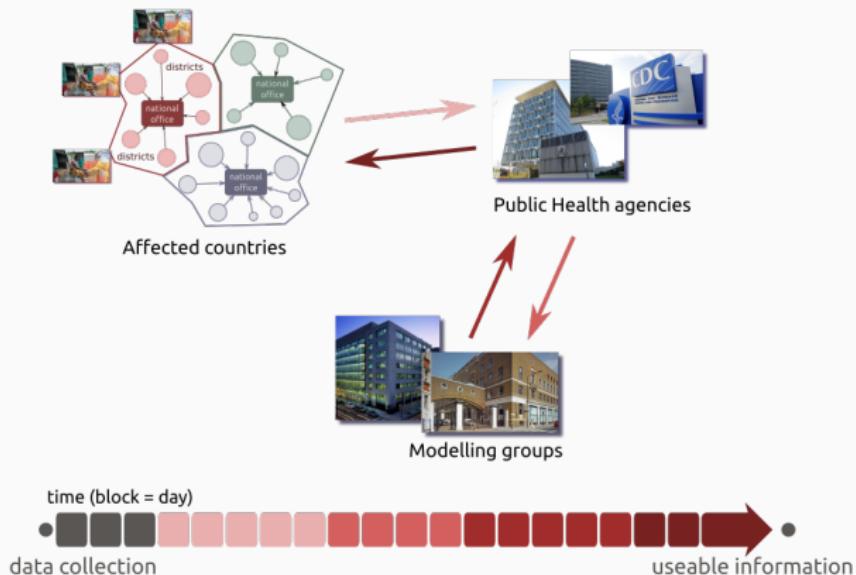
Outbreak response context creates distance and delays



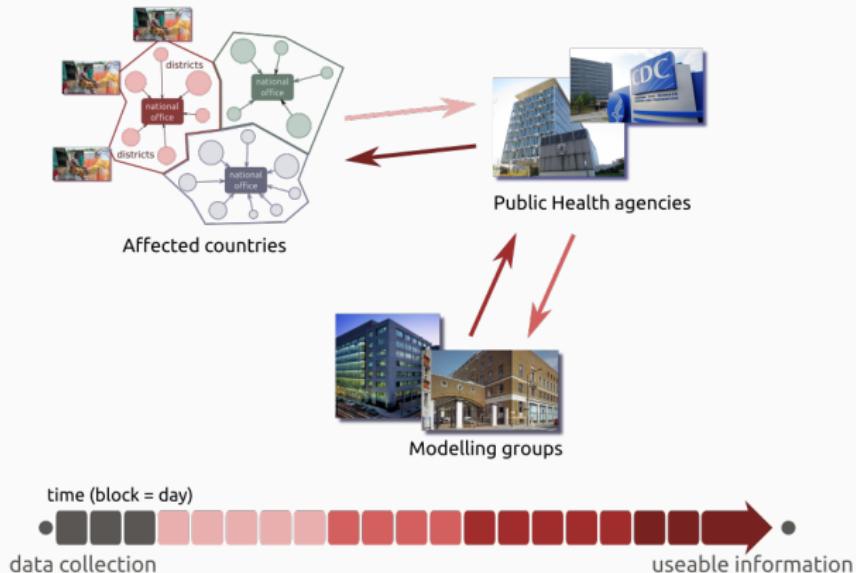
Outbreak response context creates distance and delays



Outbreak response context creates distance and delays

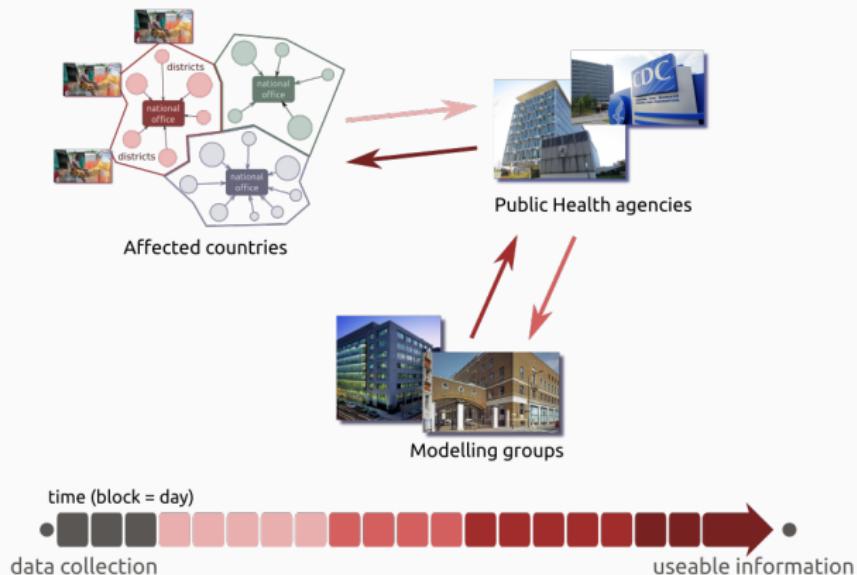


Outbreak response context creates distance and delays



- 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

- 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?