



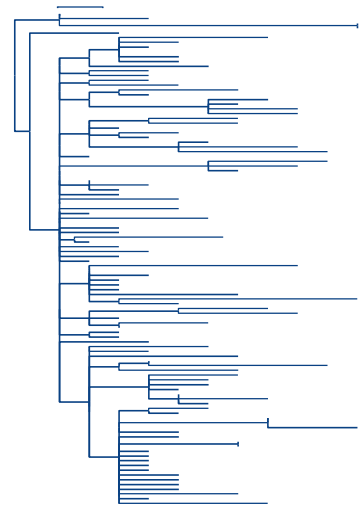
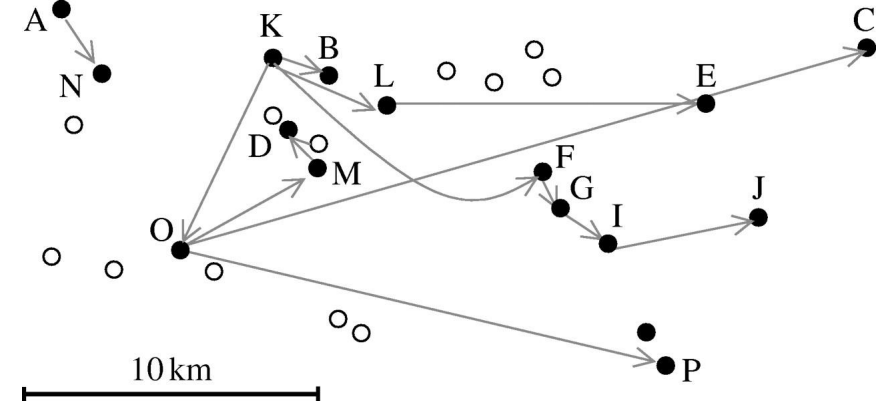
Melbourne Veterinary School
Faculty of Veterinary and Agricultural Sciences

Overview of key tools and approaches for outbreak reconstruction, inference and simulation

Simon Firestone¹, Saritha Kodikara

¹ Melbourne Veterinary School, The University of Melbourne

² Royal Melbourne Institute of Technology University





Outline

Overview of key tools and approaches for outbreak reconstruction, inference and simulation

A Frequentist approach to outbreak reconstruction: Cottam UK FMD 2001, 2007

Bayesian inference and approaches

- Some of my examples: FMD Japan 2010, *M.bovis* in New Zealand
- Gentle introduction to Bayesian inference for analysing epidemics
- Key tools for outbreak reconstruction and parameter inference
- Key tools for outbreak simulation: in peace-time and real-time
- Links to Bayesian approaches and genomic data
- Comparison of the accuracy of outbreak reconstruction algorithms
- Extension of the best model into BORIS

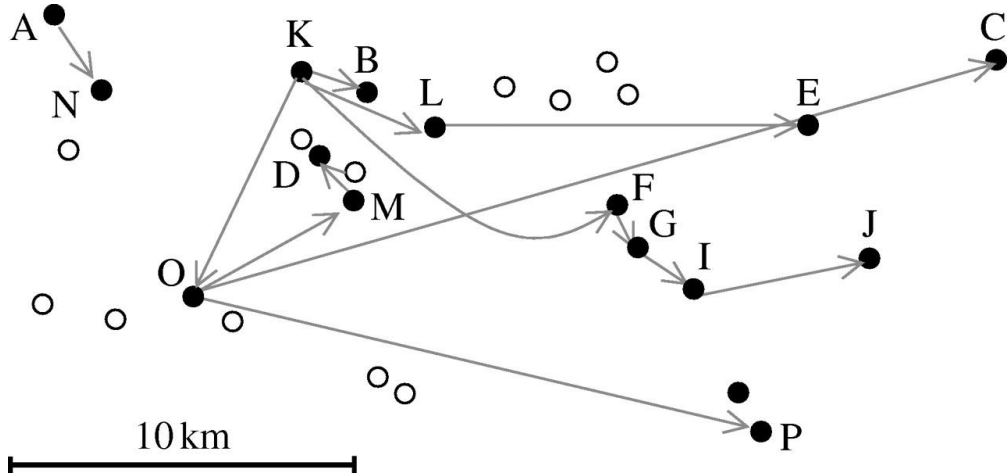
Inferring who infected whom: examples

Integrating Genetic and Epidemiological Data to Determine Transmission Pathways of FMD UK 2001
(Cottam et al, 2008a)

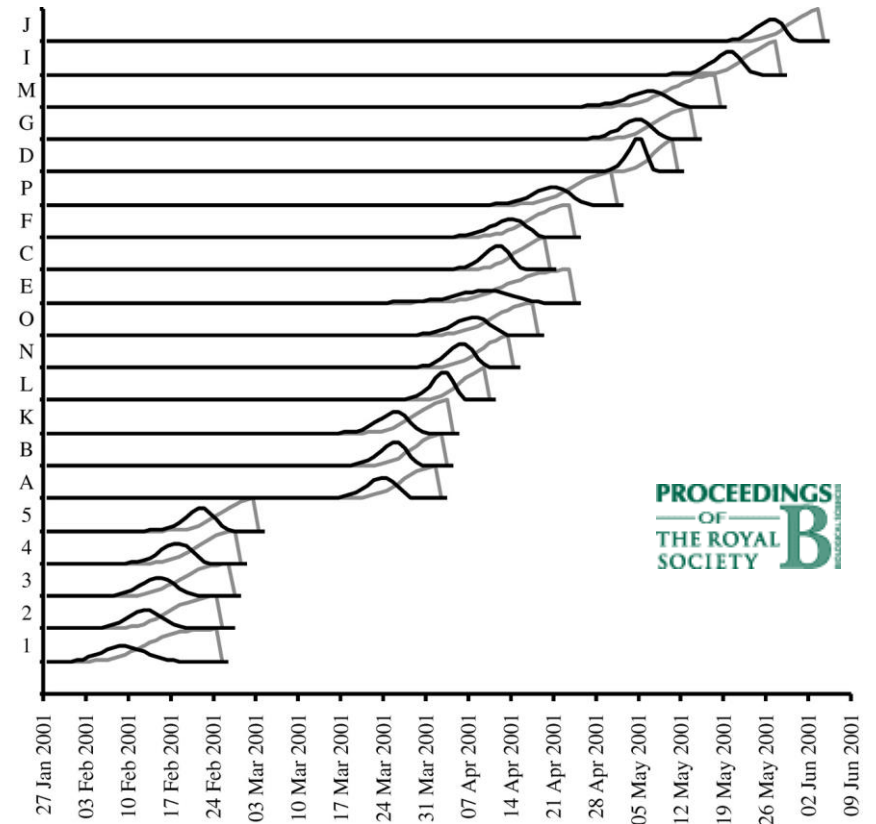
Rank all possible genomic parsimony networks by 'epi' likelihood score.

'epi' likelihood functions for:

- farm i was infectious at time, t
- farm i infected farm j



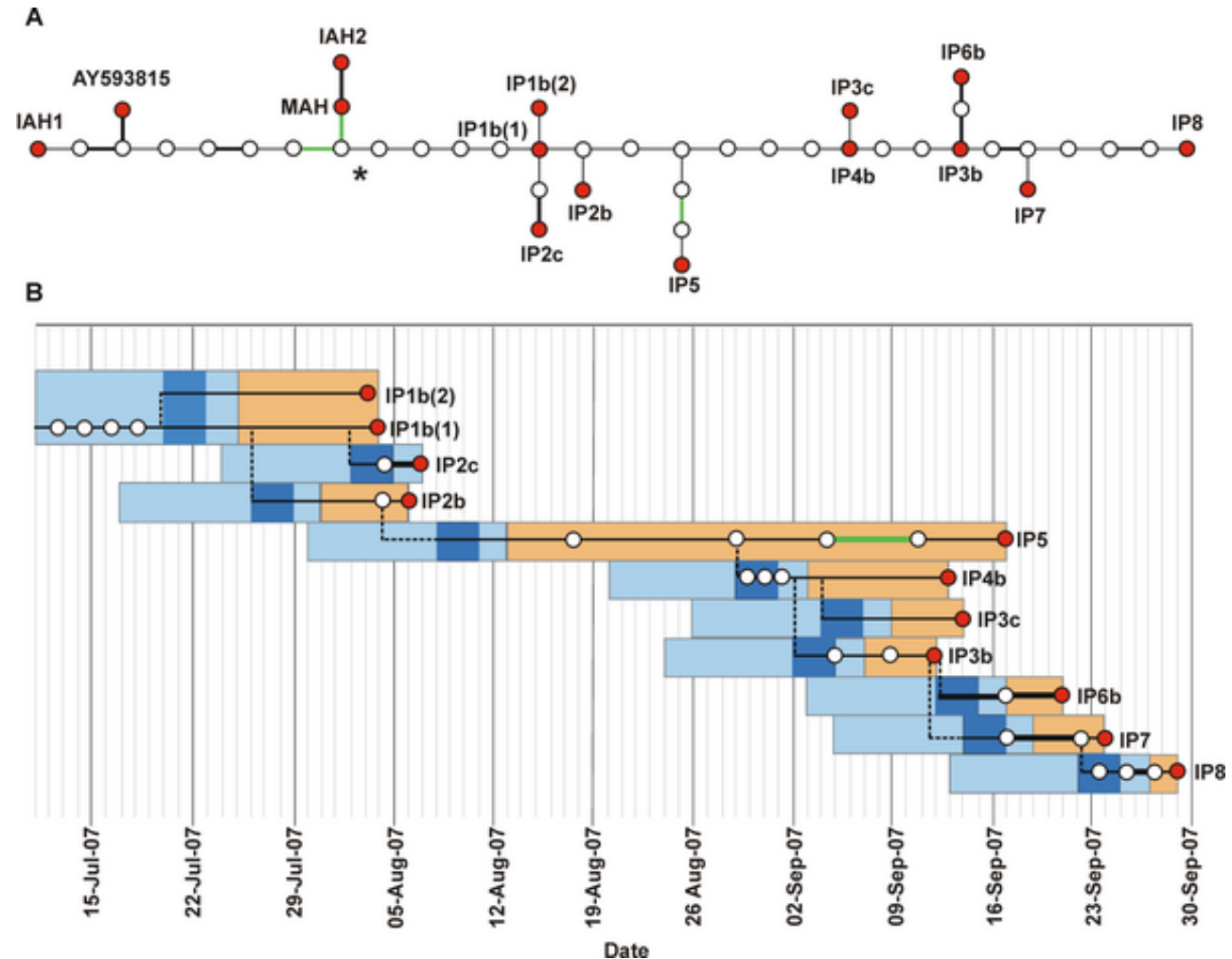
Transmission risk windows



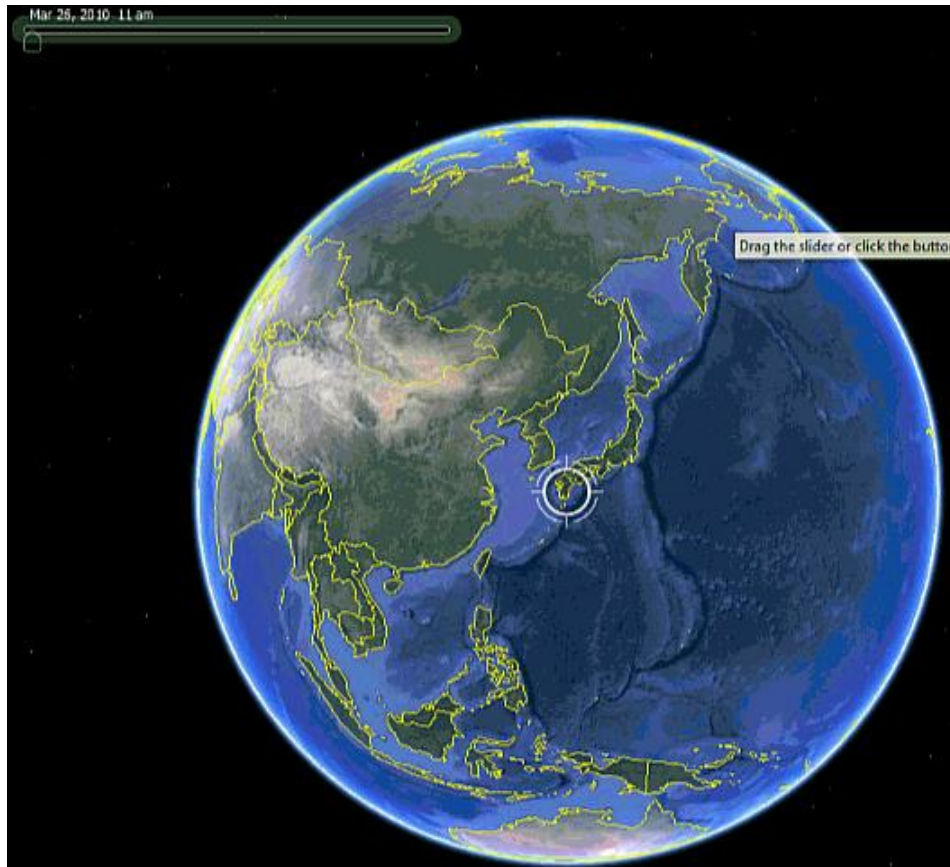
Inferring who infected whom: examples

Transmission Pathways of Foot-and-Mouth Disease Virus in the United Kingdom in 2007

(Cottam et al 2008b)



FMD in Miyazaki Prefecture, Japan 2010



292 premises detected as infected

nearly 290,000 animals were culled

economic impact \approx USD 2.4 billion

suspect 'source' farm #6, detected after index #1

sequences available on 104 clinical samples

FMD Japan 2010:

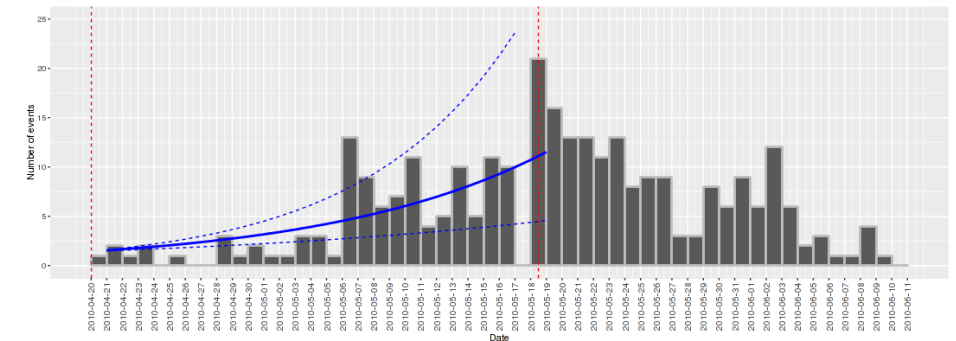
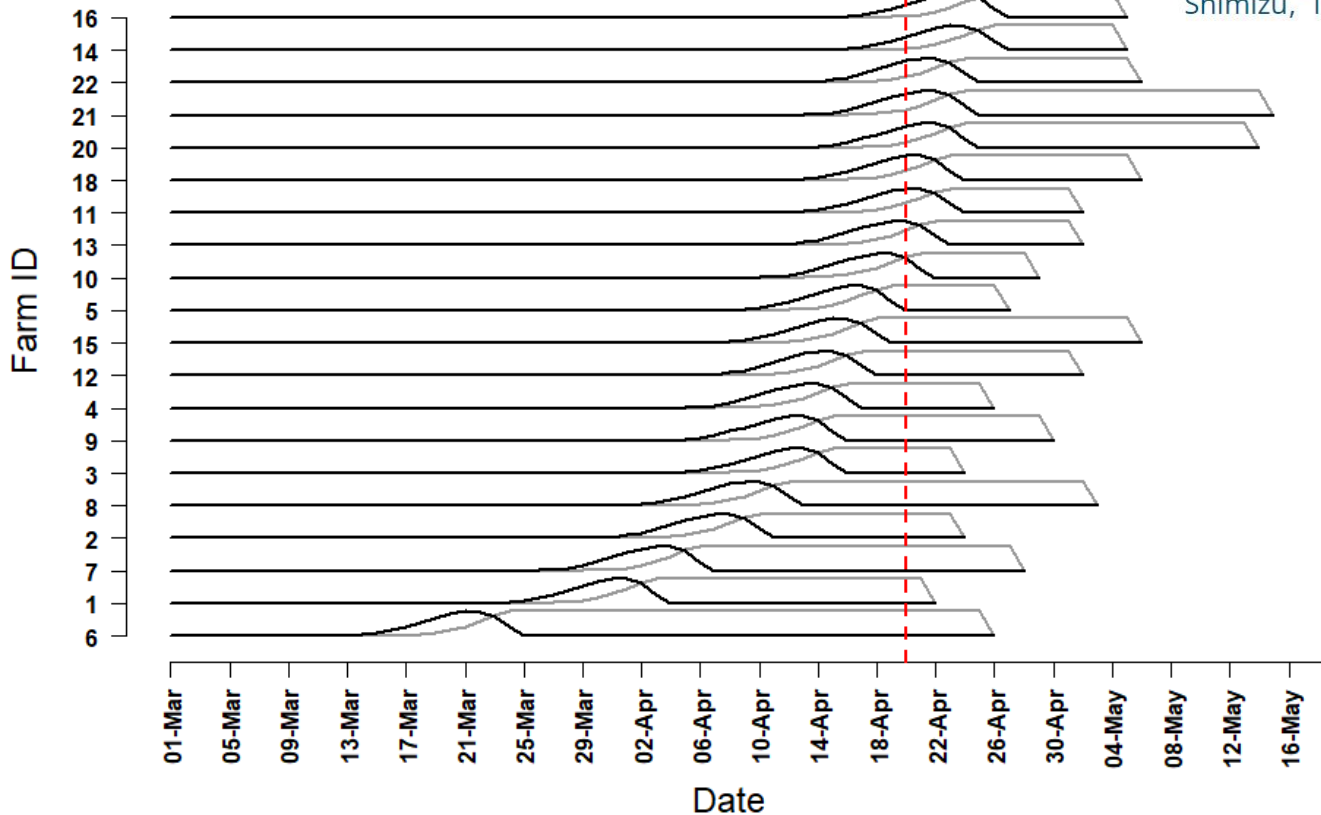


Original Article | [Full Access](#)

Reconstructing a transmission network and identifying risk factors of secondary transmissions in the 2010 foot-and-mouth disease outbreak in Japan

Yoko Hayama , Simon M. Firestone, Mark A. Stevenson, Takehisa Yamamoto, Tatsuya Nishi, Yumiko Shimizu, Toshiyuki Tsutsui

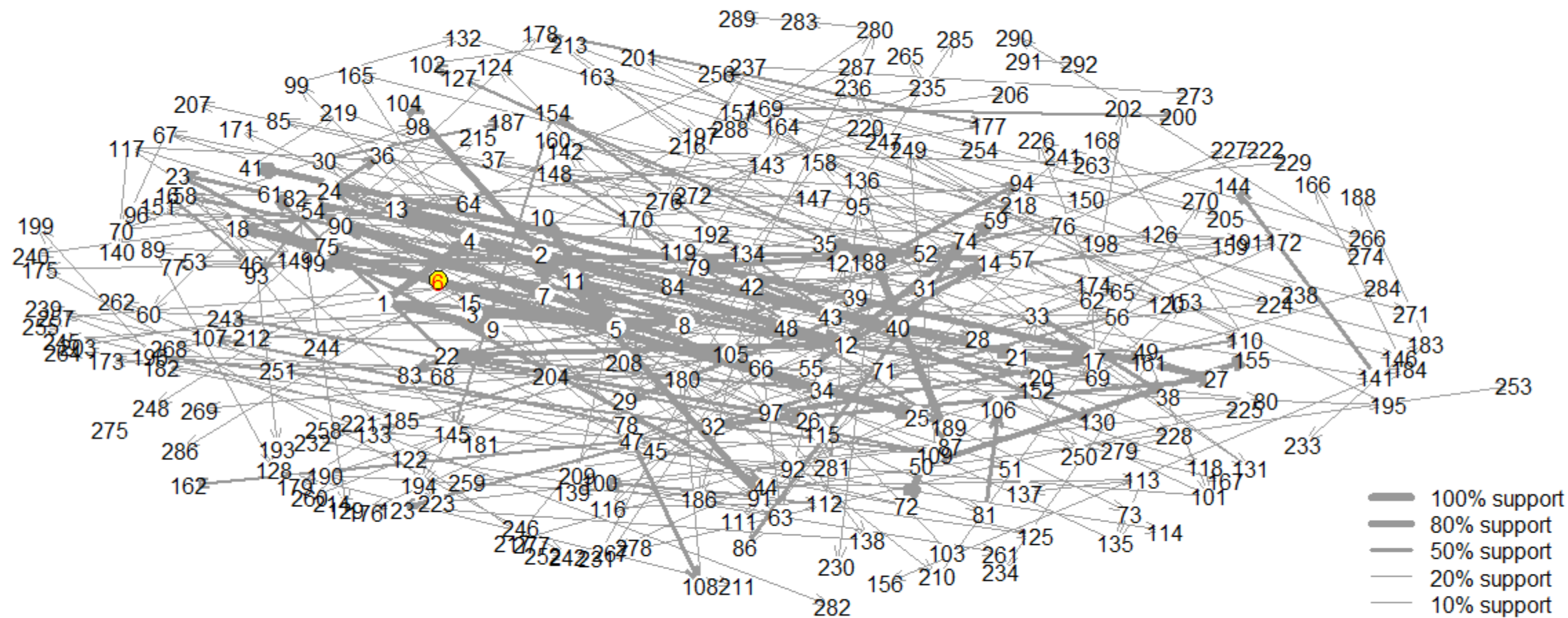
Temporal risk windows



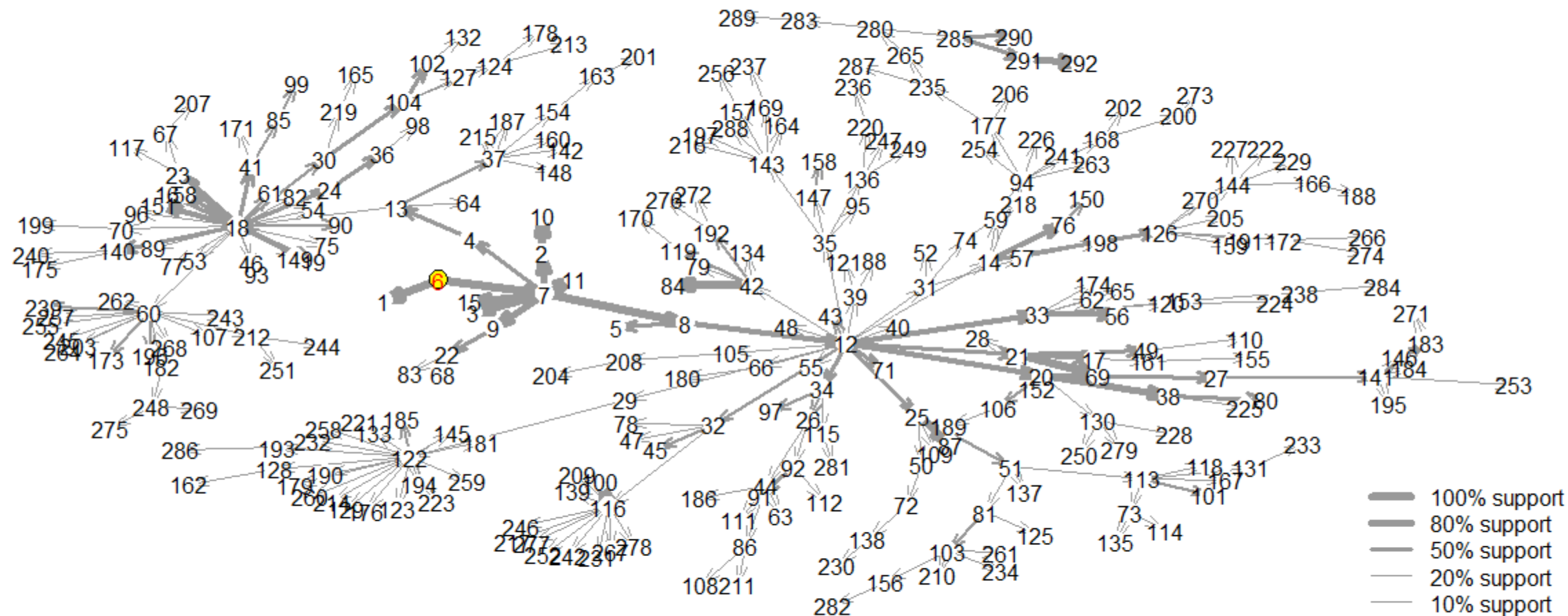
Doubling time: 12 days (95% CI: 11, 15)

R_0 : 1.40 (95% CI: 1.33, 1.46)

Inferred network (i=10,000)

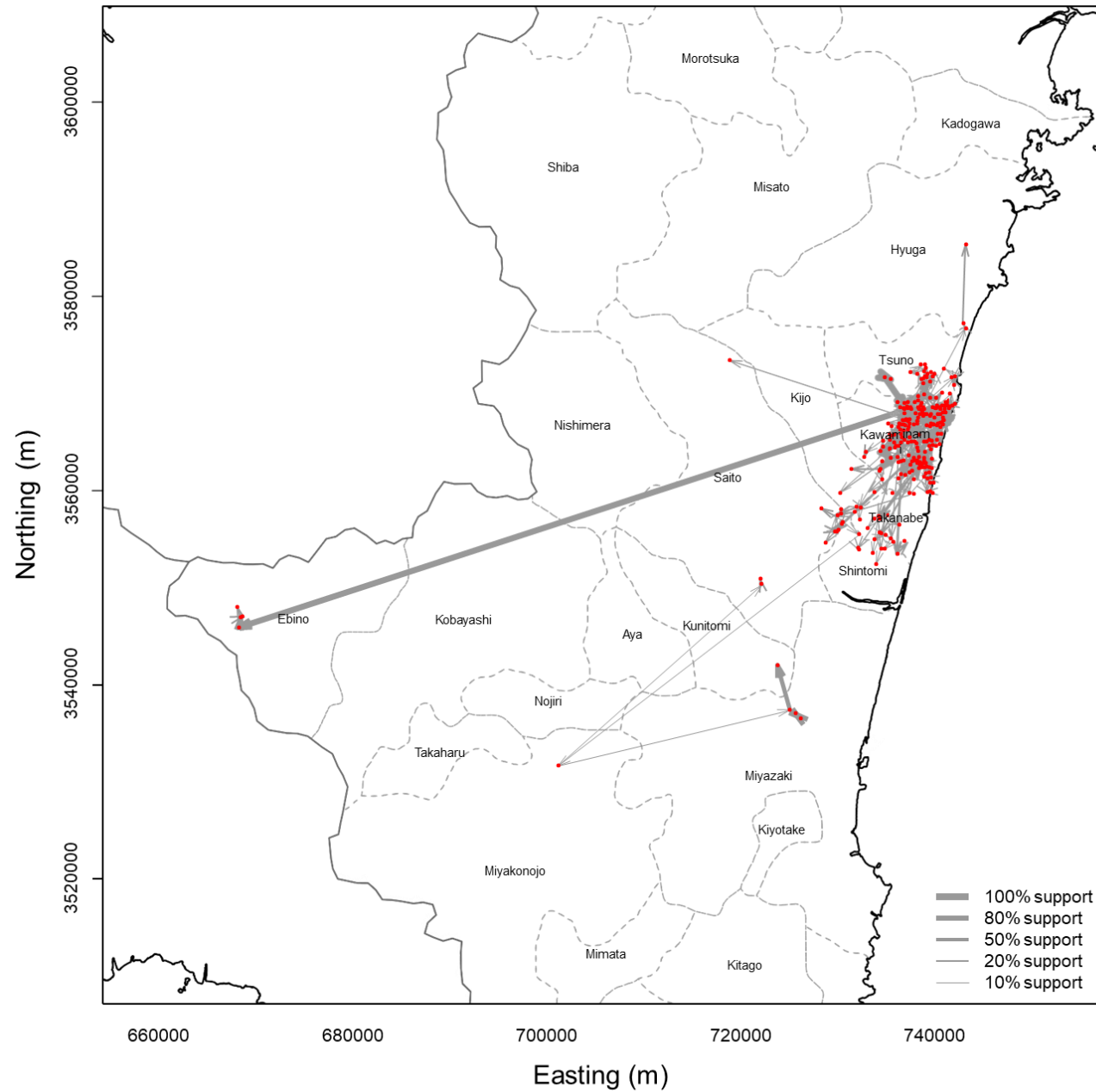


Inferred network in arbitrary space

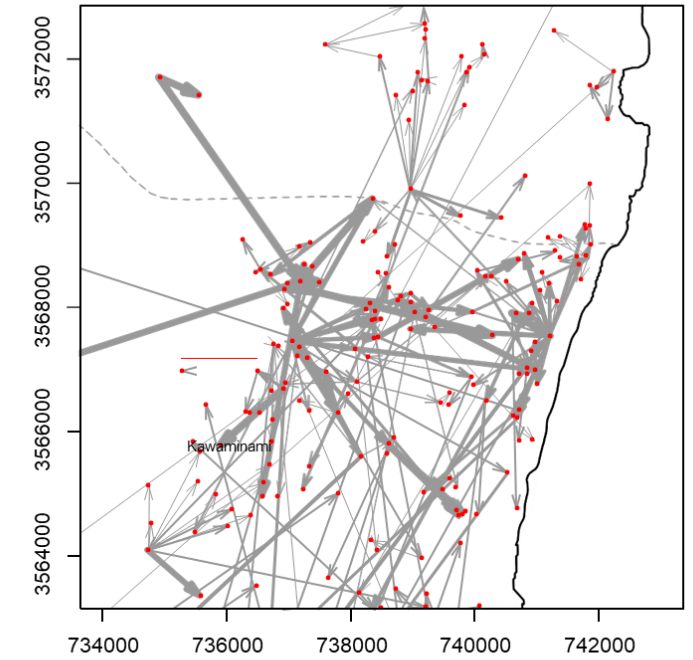


FMD Miyazaki Prefecture, Japan 2010

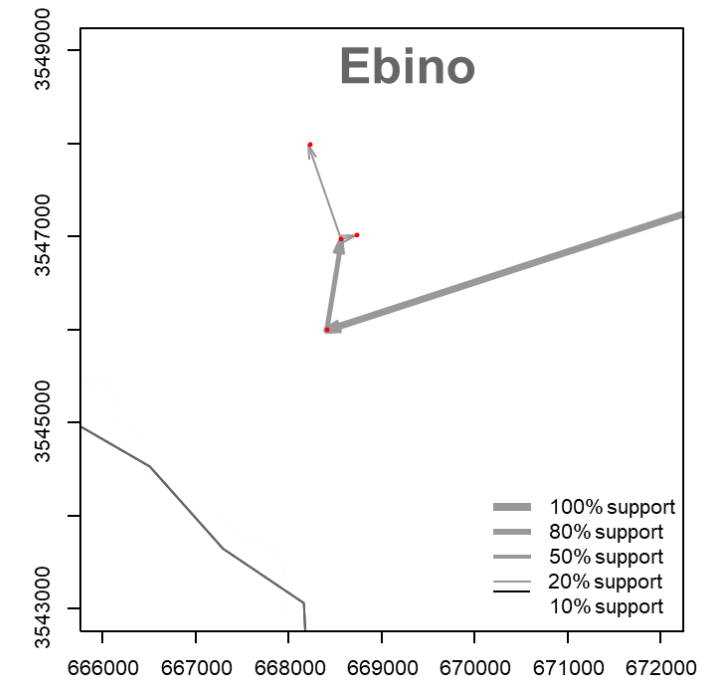
Inferred network



Kawaminami

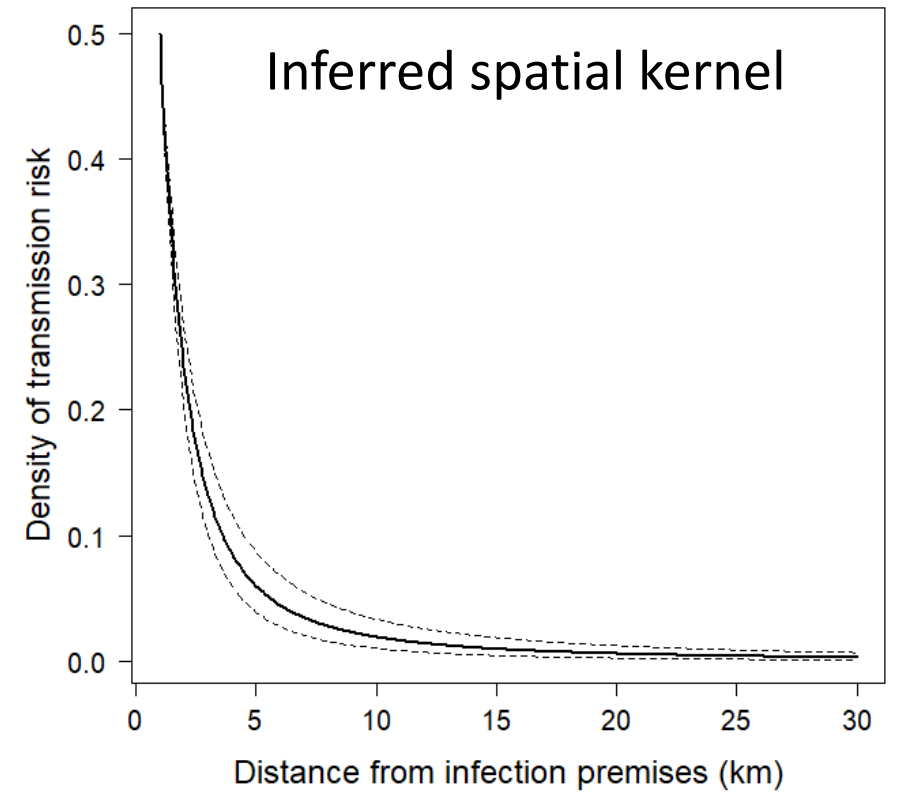


Ebino



FMD Japan 2010: inferred parameters

Model parameter	Median	Units	[95% HPD]
mutation rate	1.83×10^{-5}	substitutions site ⁻¹ day ⁻¹	$[1.63, 2.06] \times 10^{-5}$
delay from origin to outbreak detection	29.4	days	[25.4, 47.3]
number of farms infected at detection	13	IPs	[10, 18]
incubation period	3.0	days	[1.3, 8.9]
latent period	7.1	days	[5.6, 8.8]
infectious period	15.2	days	[13.6, 17.2]



FMD Japan 2010: inferred parameters

Model parameter	Predominant species	Median	[95% HPD]
Infectivity	cattle	1.0	Reference
	pigs	14.7	[7.21, 30.2]
	sheep & other	0.60	[0.03, 3.43]
Effect of farm size	(no. animals) ^x	0.08	[0.00, 0.23]
Susceptibility	cattle	1.0	Reference
	pigs	0.21	[0.01, 1.02]
	sheep & other	0.50	[0.02, 3.53]
Effect of farm size	(no. animals) ^x	0.16	[0.07, 0.24]

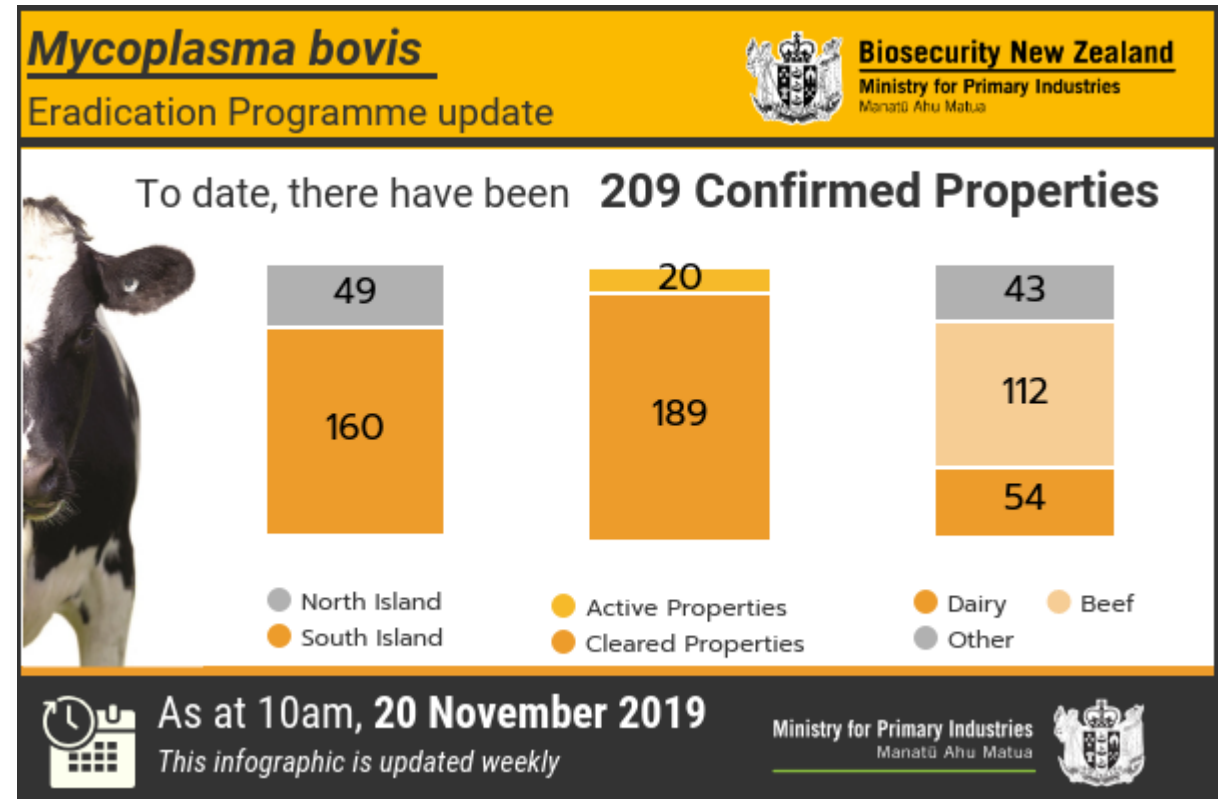
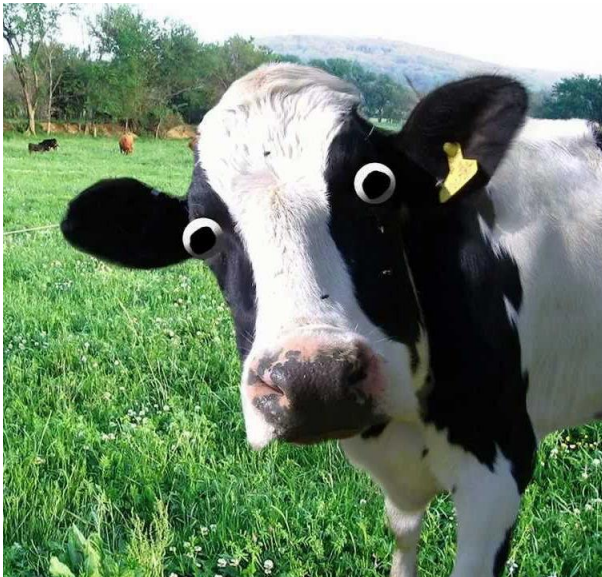
Inferring who infected whom: examples

Mycoplasma bovis in New Zealand (detected July 2017, ongoing)

Mycoplasma bovis



Working together with industry, we aim to eradicate *Mycoplasma bovis* from New Zealand.



<https://www.mpi.govt.nz/protection-and-response/mycoplasma-bovis/>



Outline

A Frequentist approach to outbreak reconstruction: Cottam UK FMD 2001, 2007

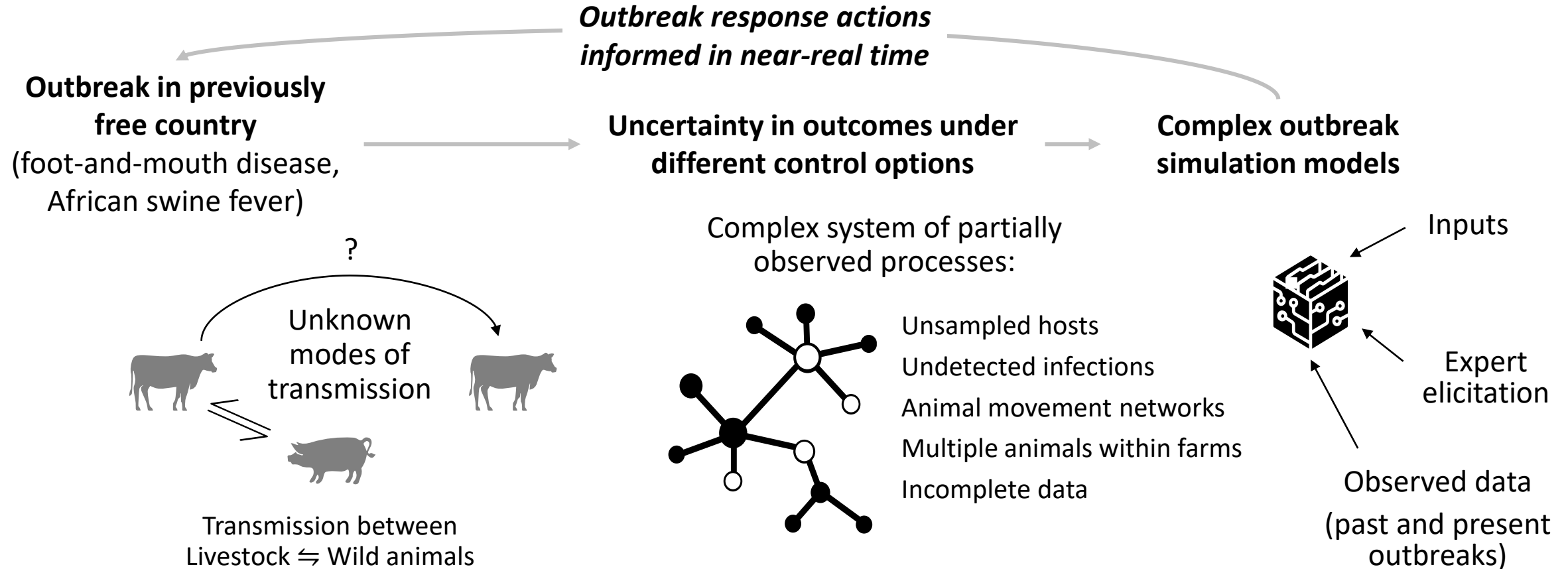
Bayesian inference and approaches

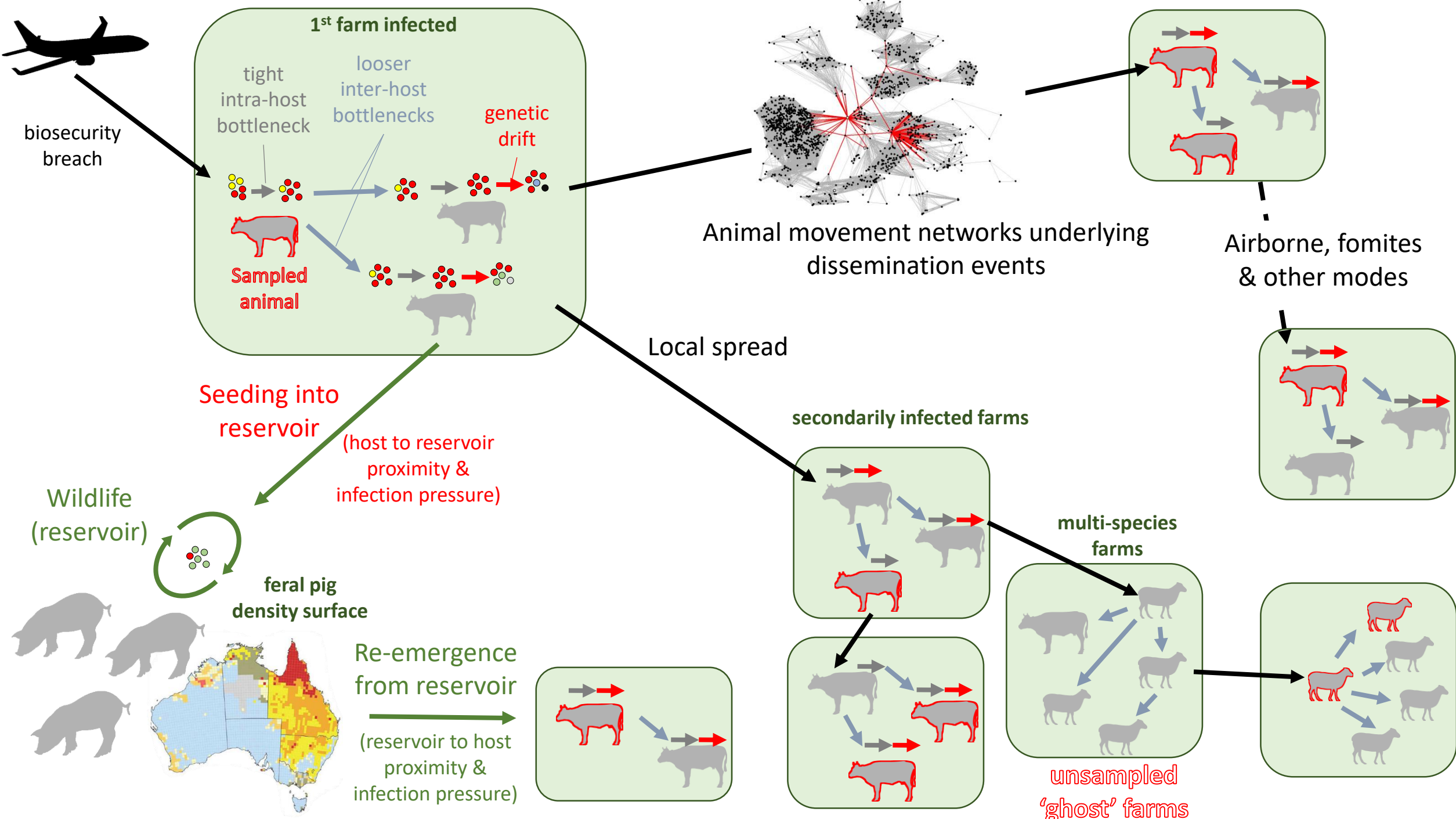
- Some of my examples: FMD Japan 2010, *M.bovis* in New Zealand
- **Gentle introduction to Bayesian inference for analysing epidemics**
- Key tools for outbreak reconstruction and parameter inference
- Key tools for outbreak simulation: in peace-time and real-time
- Links to Bayesian approaches and genomic data
- Comparison of the accuracy of outbreak reconstruction algorithms
- Extension of the best model into BORIS

Bayesian inference for analysing epidemics

Outbreak datasets are only ever partial observations.

Many key processes are unobserved, and many datapoints are missing. Many unknown parameters.

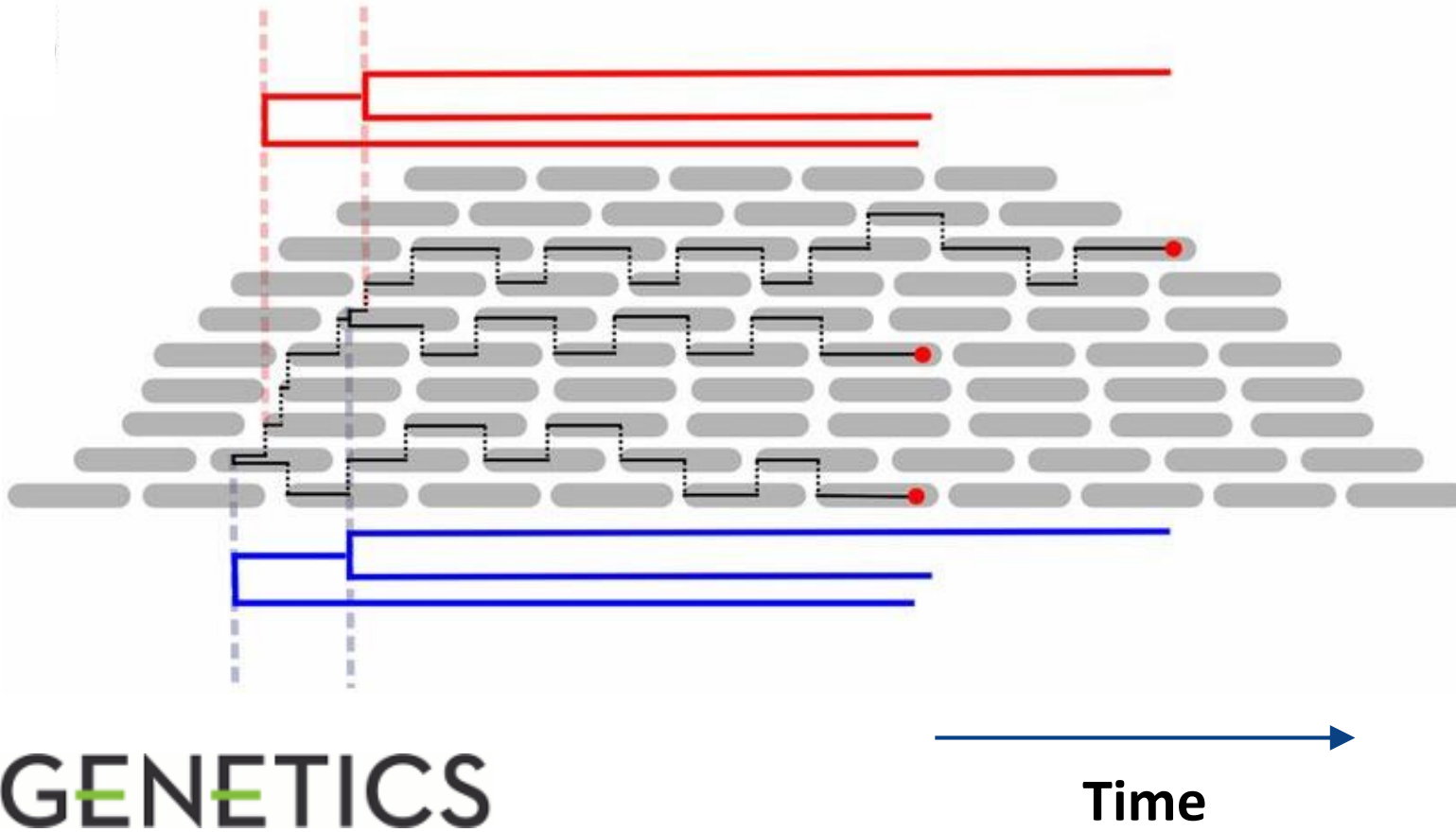




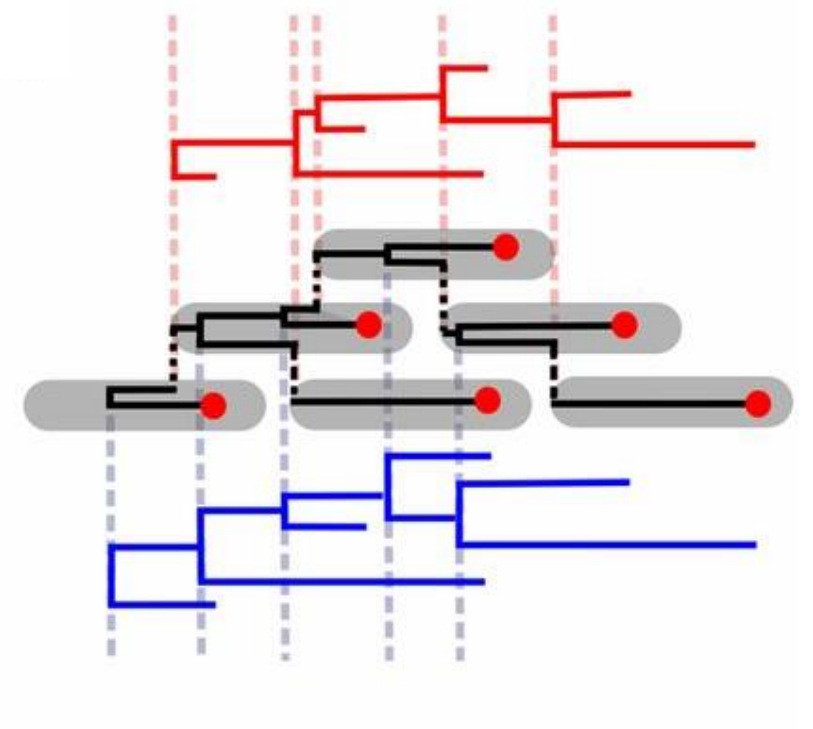
Schematic for viral dynamics

transmission network (red)
phylogenetic tree (blue)

Sparse sampling



Dense sampling

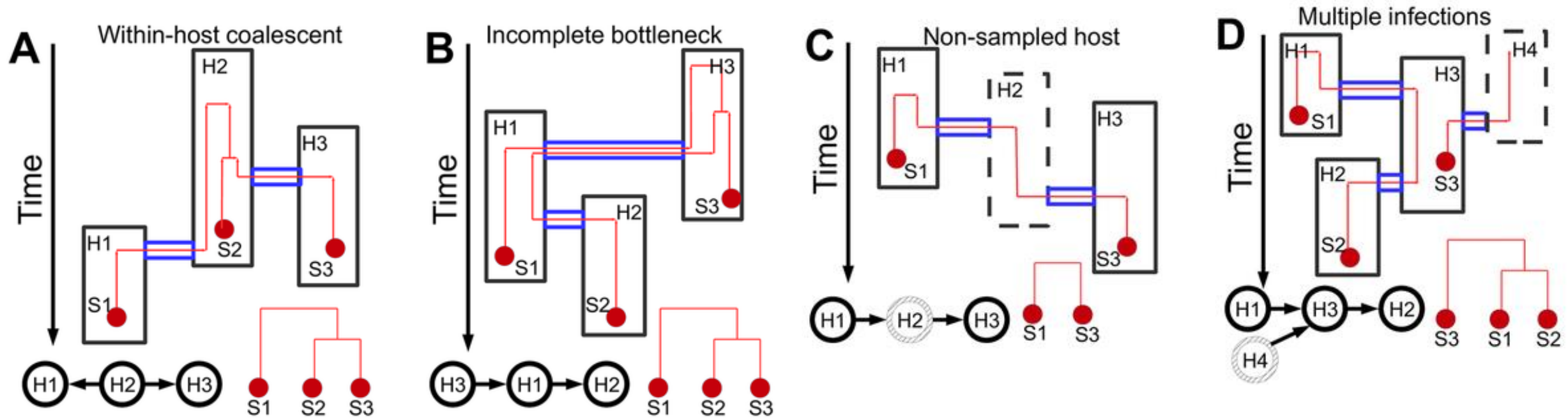


GENETICS

Rolf J. F. Ypma et al. *Genetics* 2013;195:1055-1062

Copyright © 2013 by the Genetics Society of America

Transmission Complexities



De Maio N, Wu CH, Wilson DJ (2016) SCOTTI.
 PLOS Computational Biology 12(9)

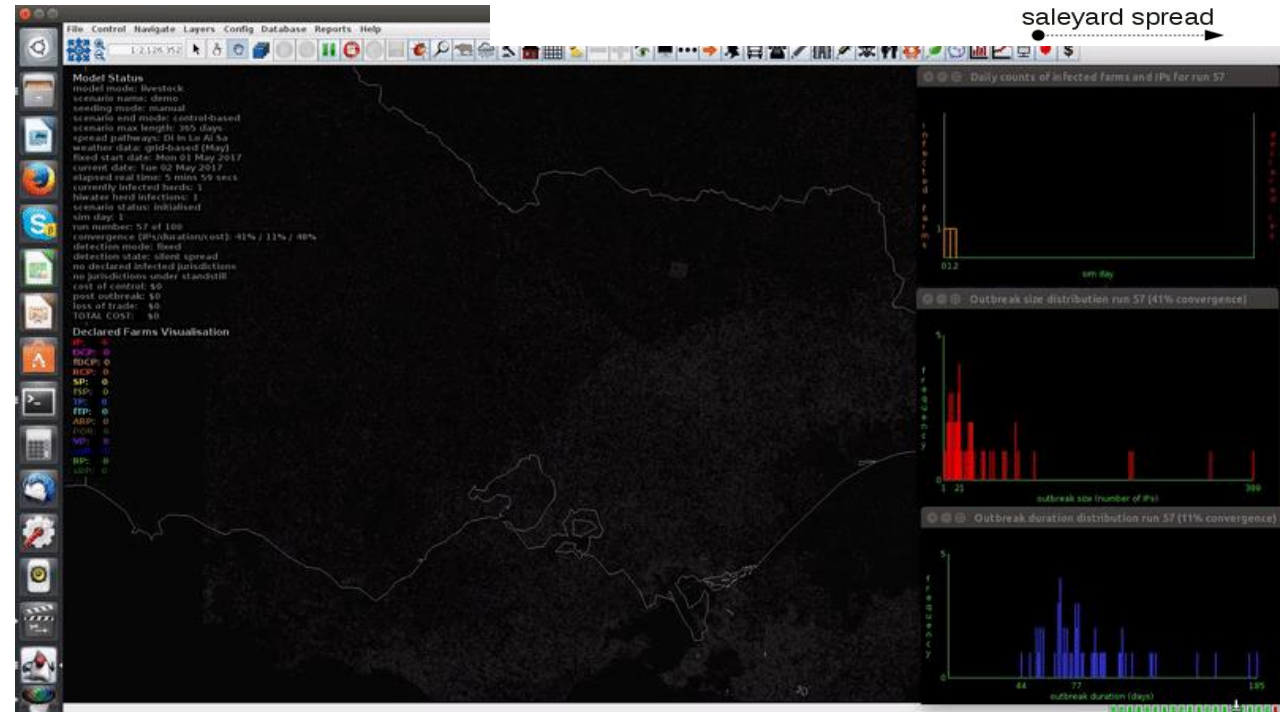
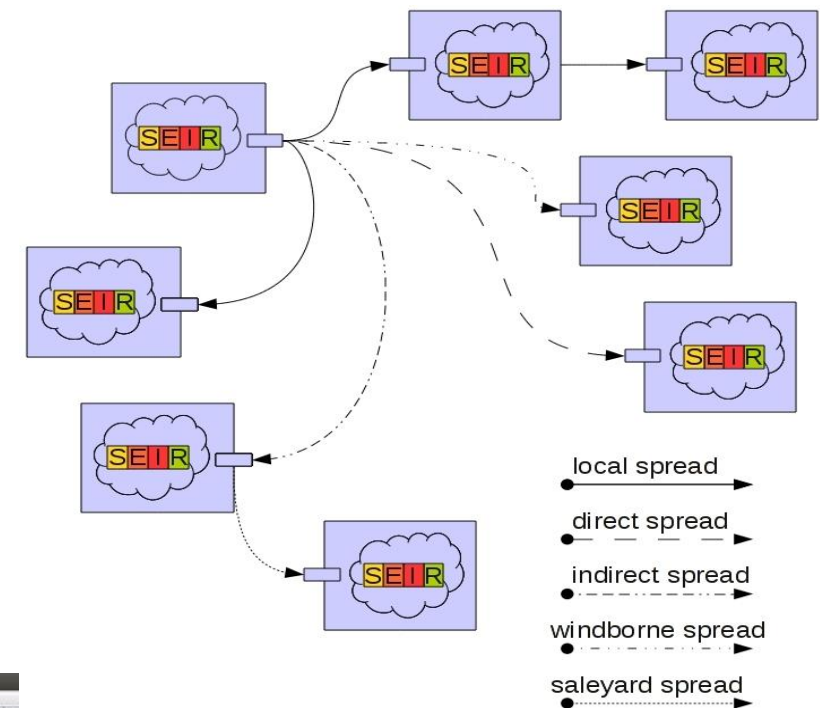


Simulation modelling: FMD AADIS

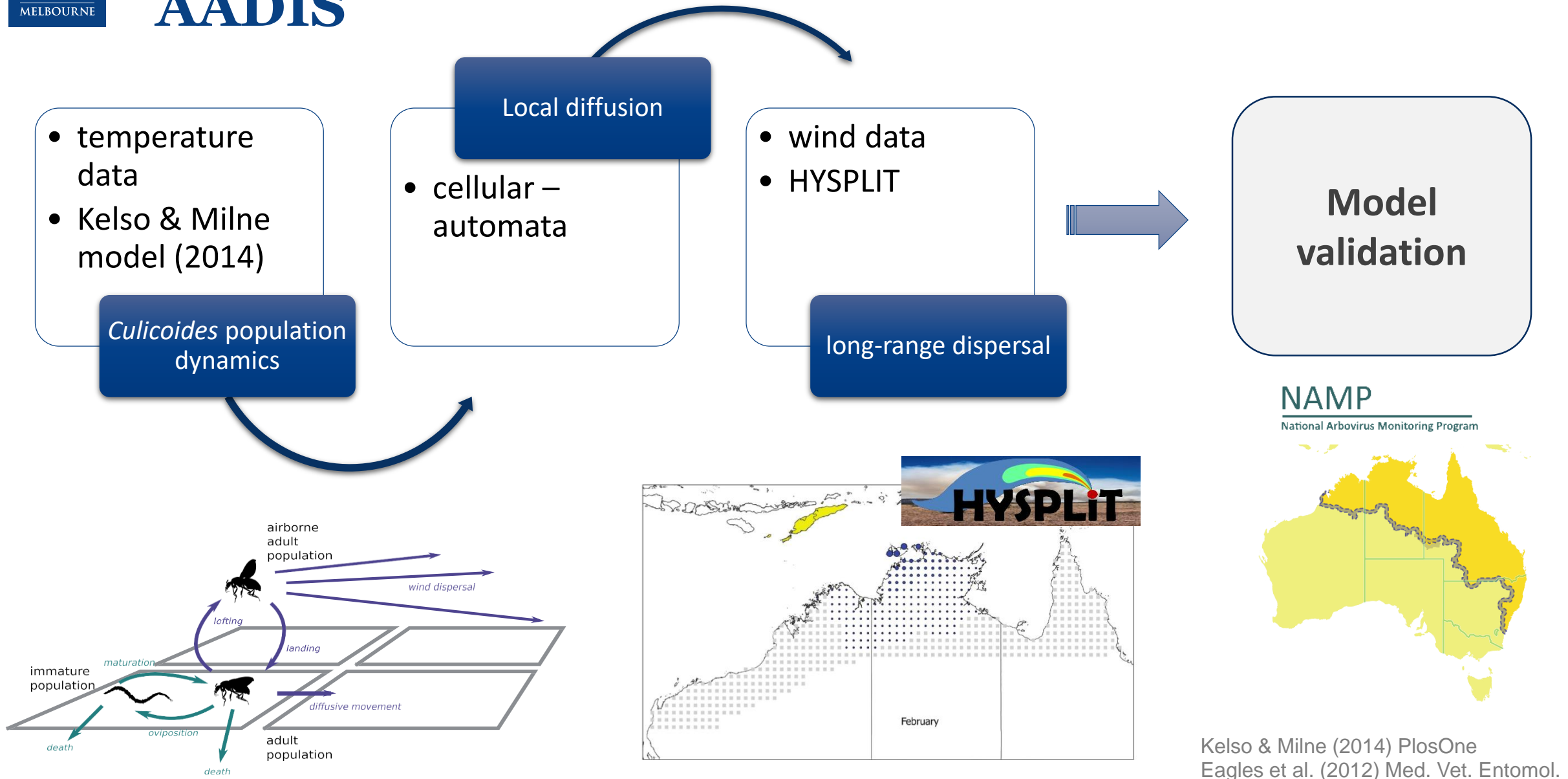
Australian Animal Disease (AADIS) Model

(DAWR/UNE/CEBRA)

- National scale (100M animals across 250K herds).
- Computationally efficient coded in Java (runs on a desktop computer)
- Hybrid
 - population-based SEIR within-herd model
 - individual (agent)-based model to represent between-herd spread, control and eradication

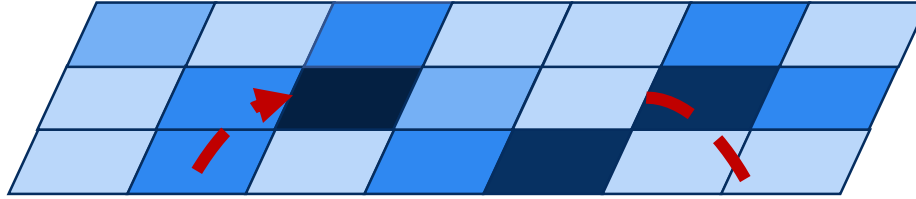


Simulation modelling vector-borne disease: AADIS



Simulation modelling vector-borne disease: AADIS

Vector raster
(grid) layer



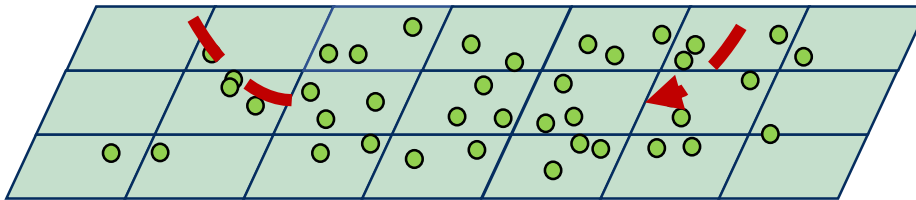
Host to vector
infection pressure

β_{hv}

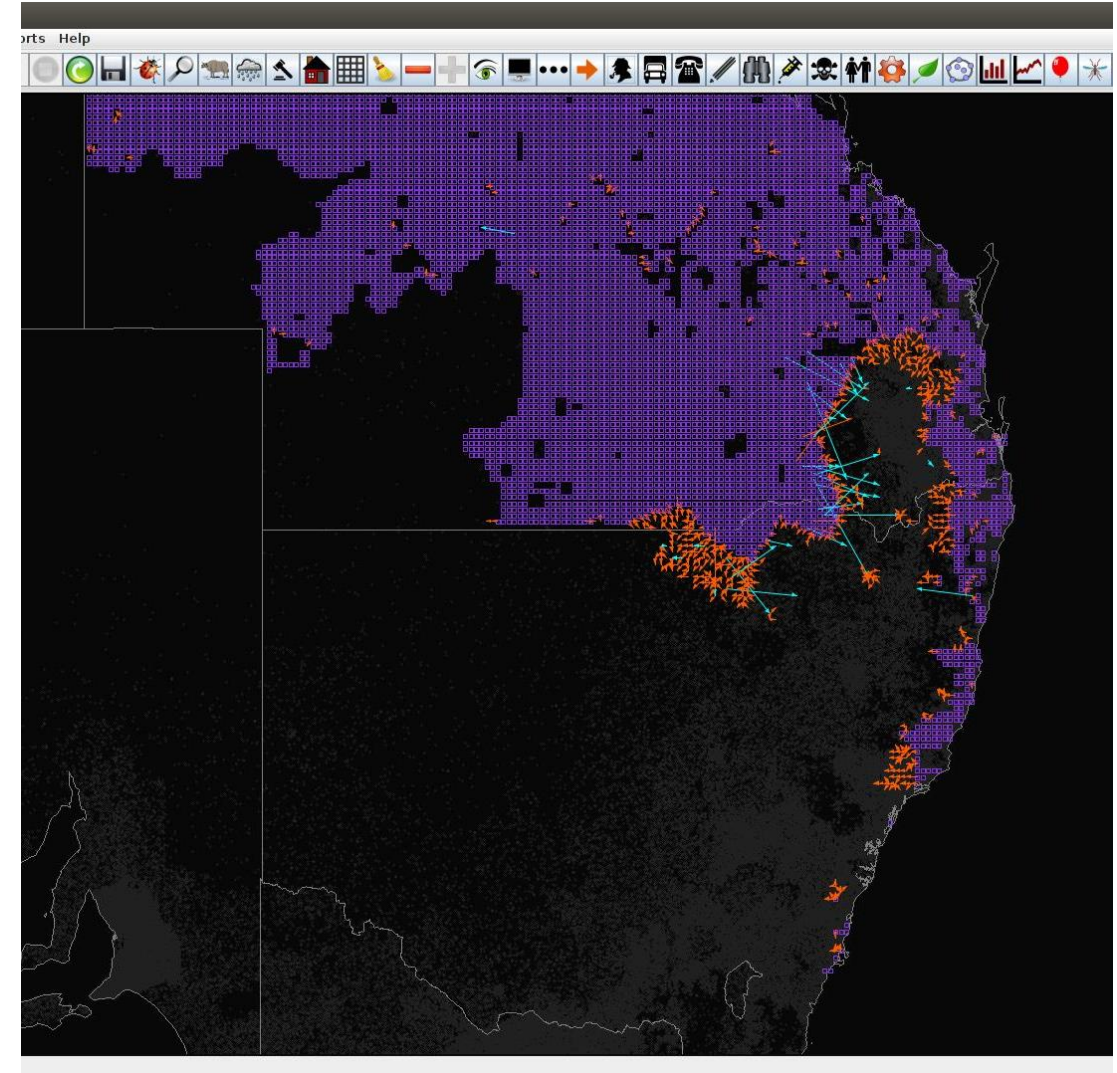
Vector to host
infection pressure

β_{vh}

Host point layer



● S ● E ● I ● R





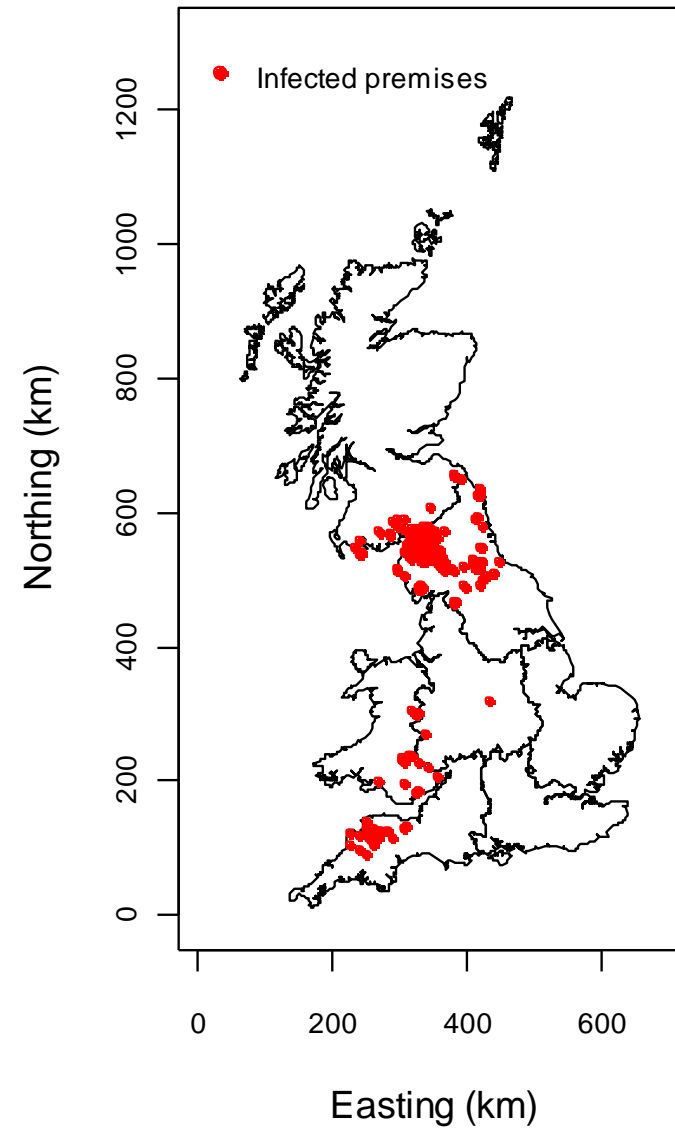
So where do modellers get their parameters?

Complex simulation models are hungry for data on a lot of variables, many with a lot of uncertainty

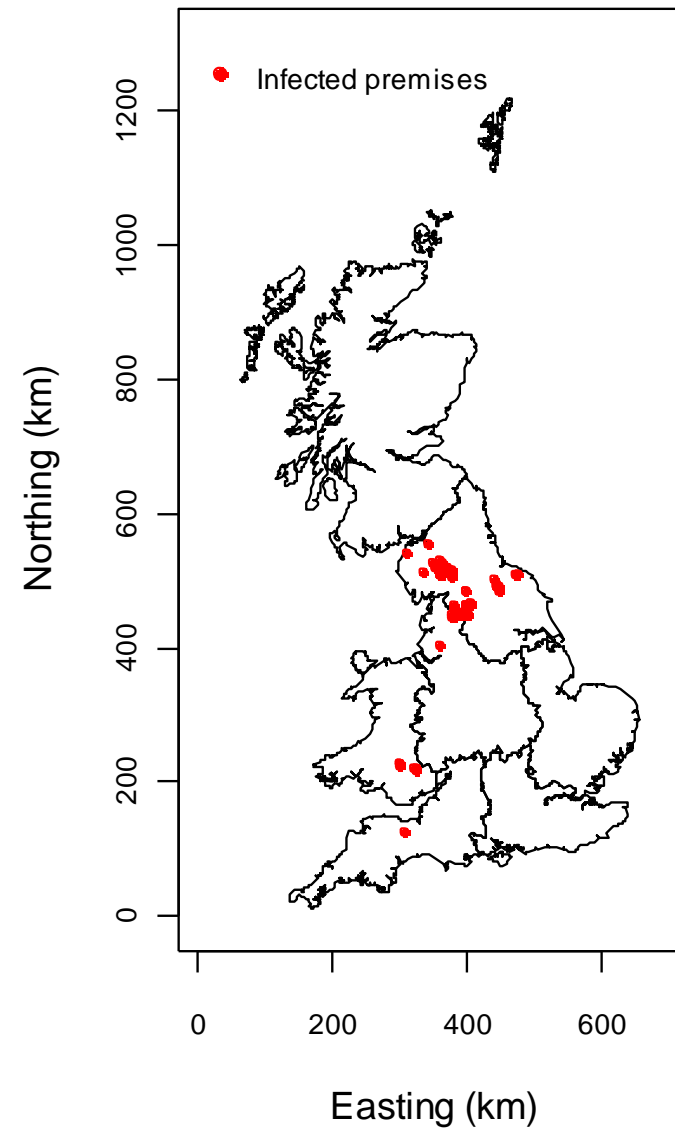
Parameter estimation in a hurry

- fitting / 'tinkering'
- 'brute force'
- ABC (at least it sounds simple!)
 - Approximate Bayesian computation
 - a formal short-cut
- Full Bayesian inference (not in a hurry unless you've done a lot of work already)

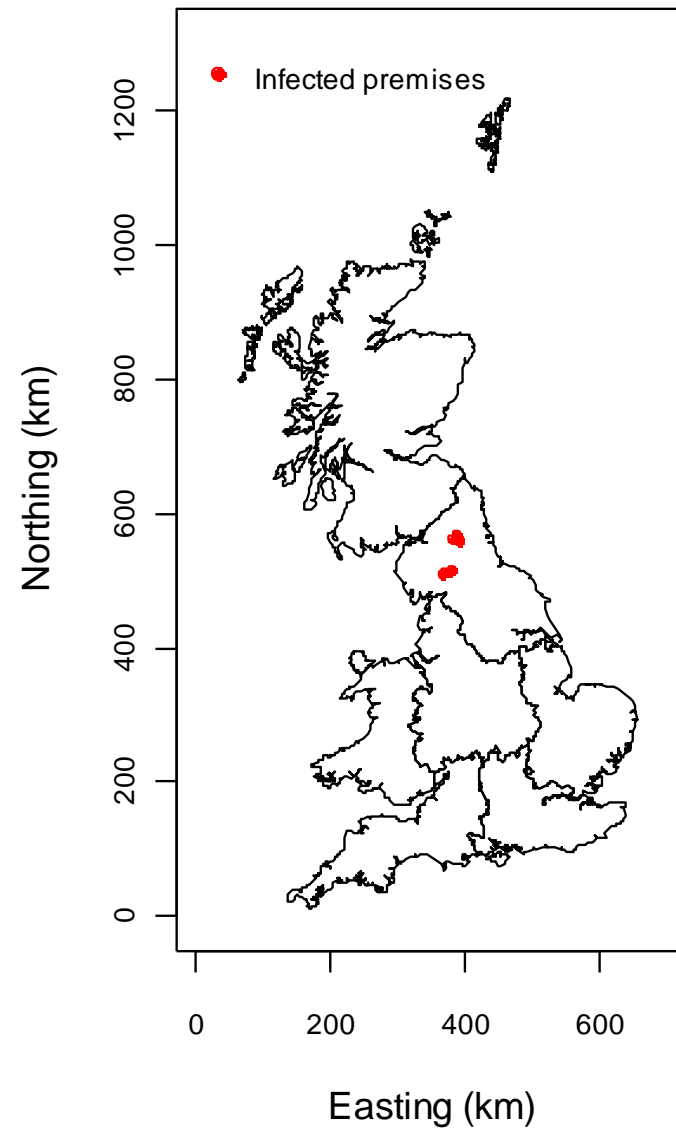
7 Apr – 5 May 2001



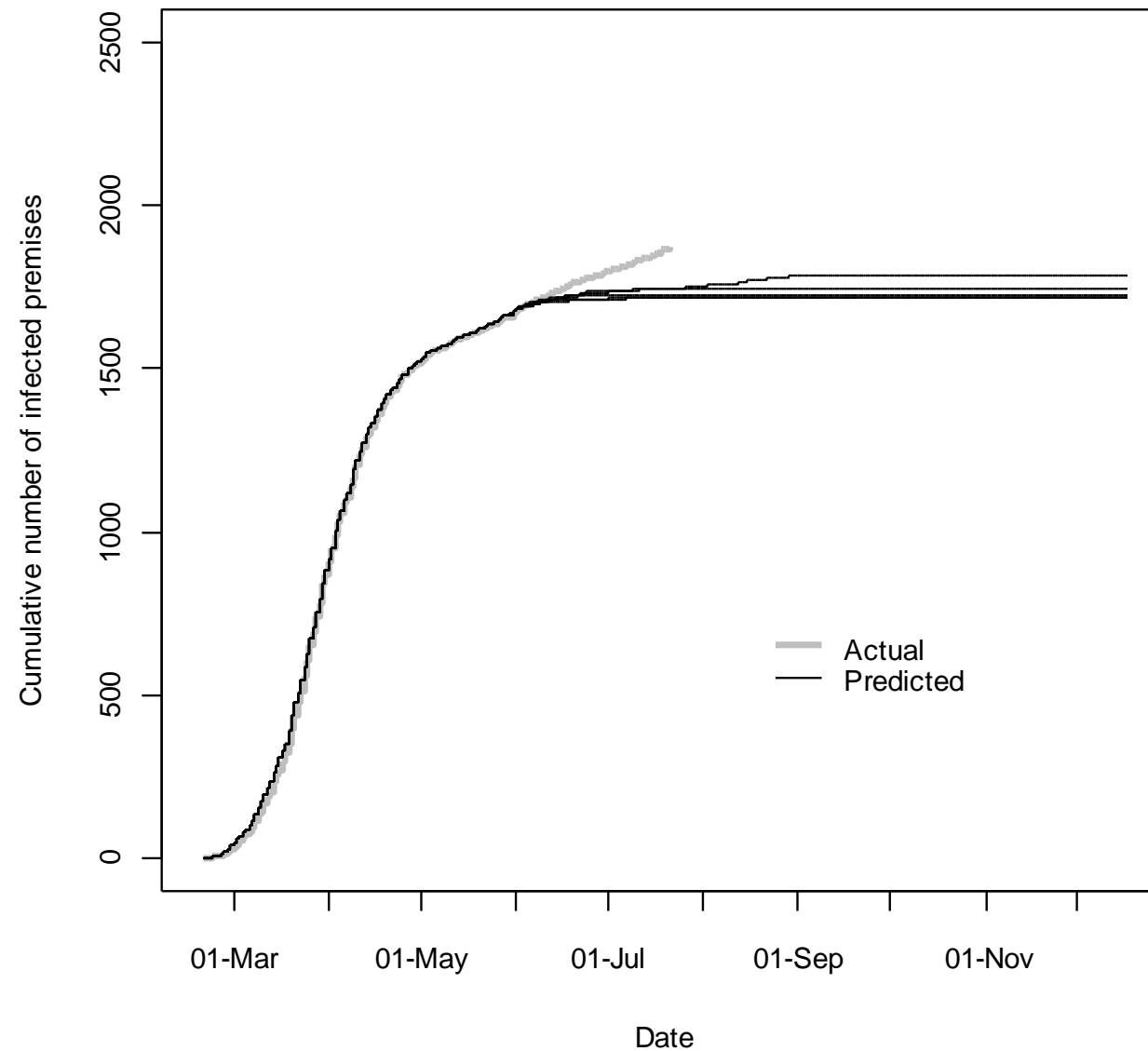
10 Jun – 8 Jul



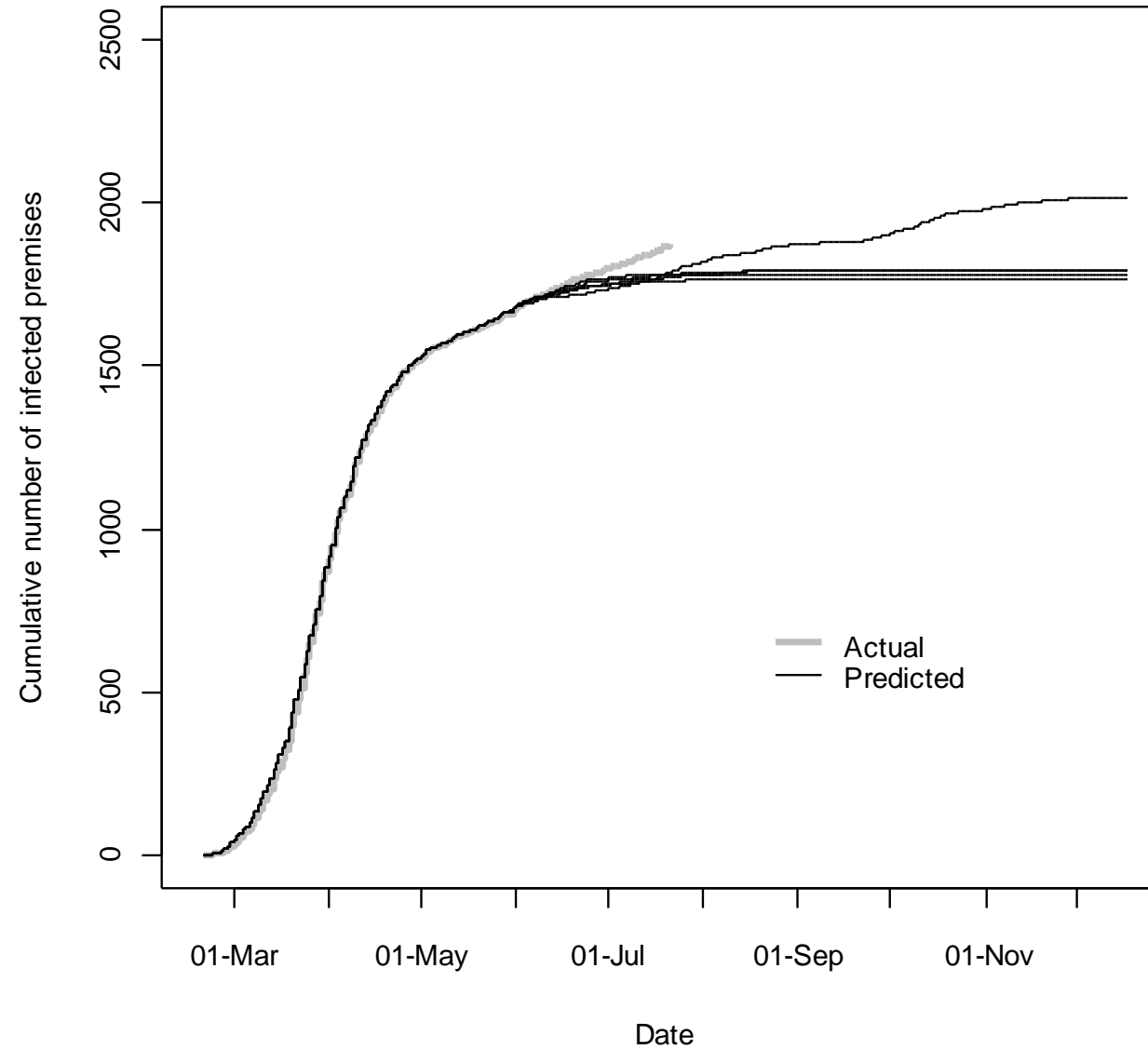
3 Sep – 1 Oct



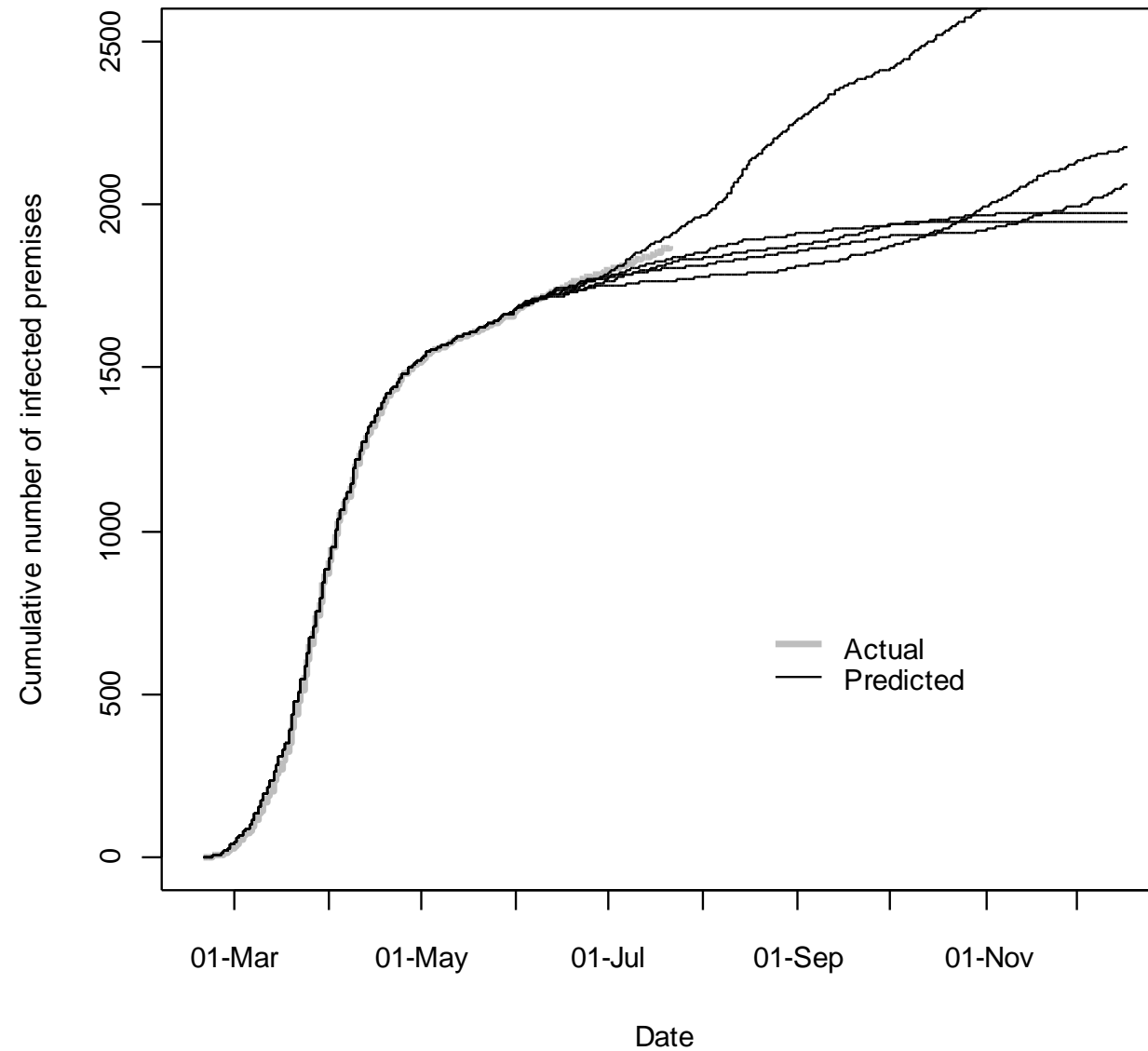
Foot-and-mouth disease in Great Britain February – September 2001. Actual cumulative number of infected premises (grey) and predicted cumulative number of infected premises (black) assuming **10%** high risk movements. Simulation start 1 June 2001.



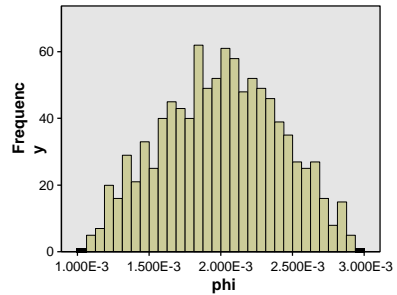
Foot-and-mouth disease in Great Britain February – September 2001. Actual cumulative number of infected premises (grey) and predicted cumulative number of infected premises (black) assuming **20%** high risk movements. Simulation start 1 June 2001.



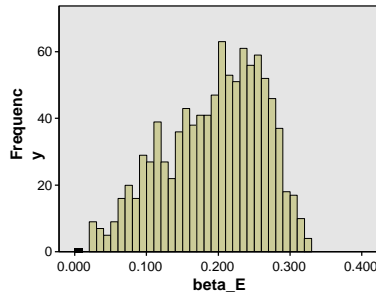
Foot-and-mouth disease in Great Britain February – September 2001. Actual cumulative number of infected premises (grey) and predicted cumulative number of infected premises (black) assuming **30%** high risk movements. Simulation start 1 June 2001.



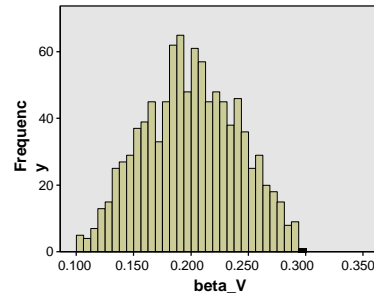
Parameter estimation: 'brute force'



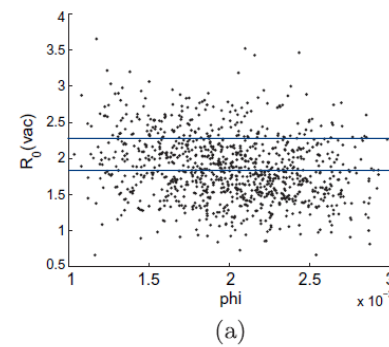
(a) Histogram of the values of ϕ



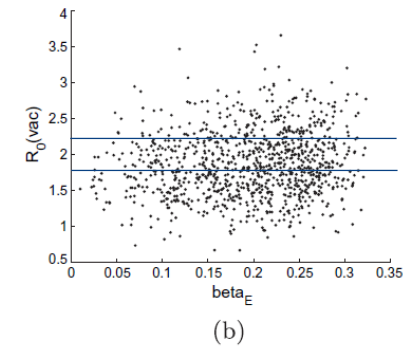
(b) Histogram of the values of β_E



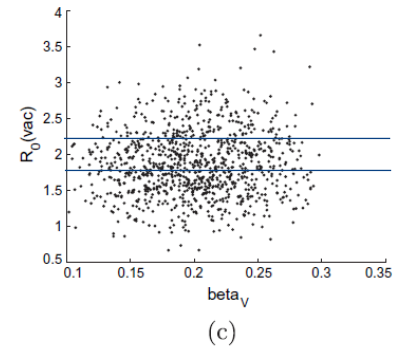
(c) Histogram of the values of β_V



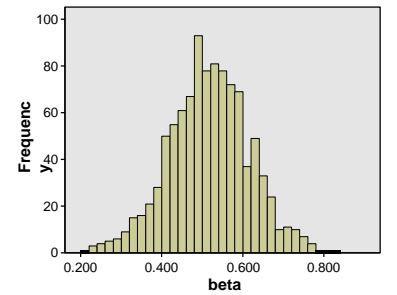
(a)



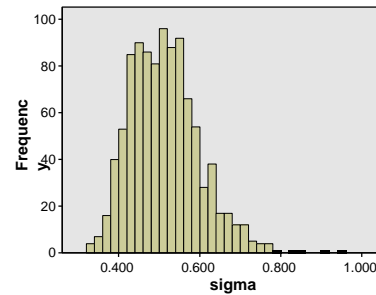
(b)



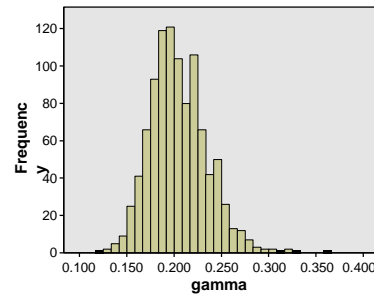
(c)



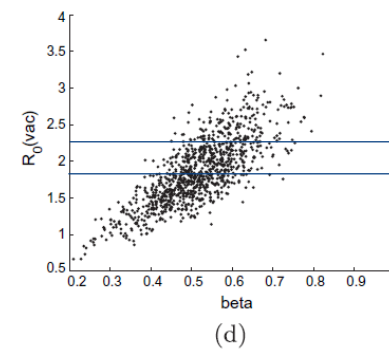
(d) Histogram of the values of β



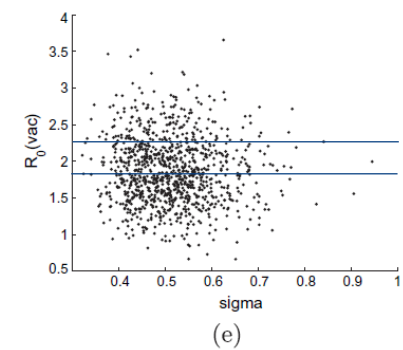
(e) Histogram of the values of σ



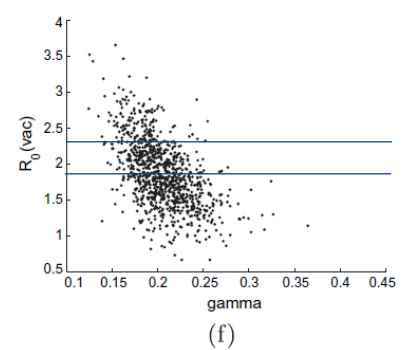
(f) Histogram of the values of γ



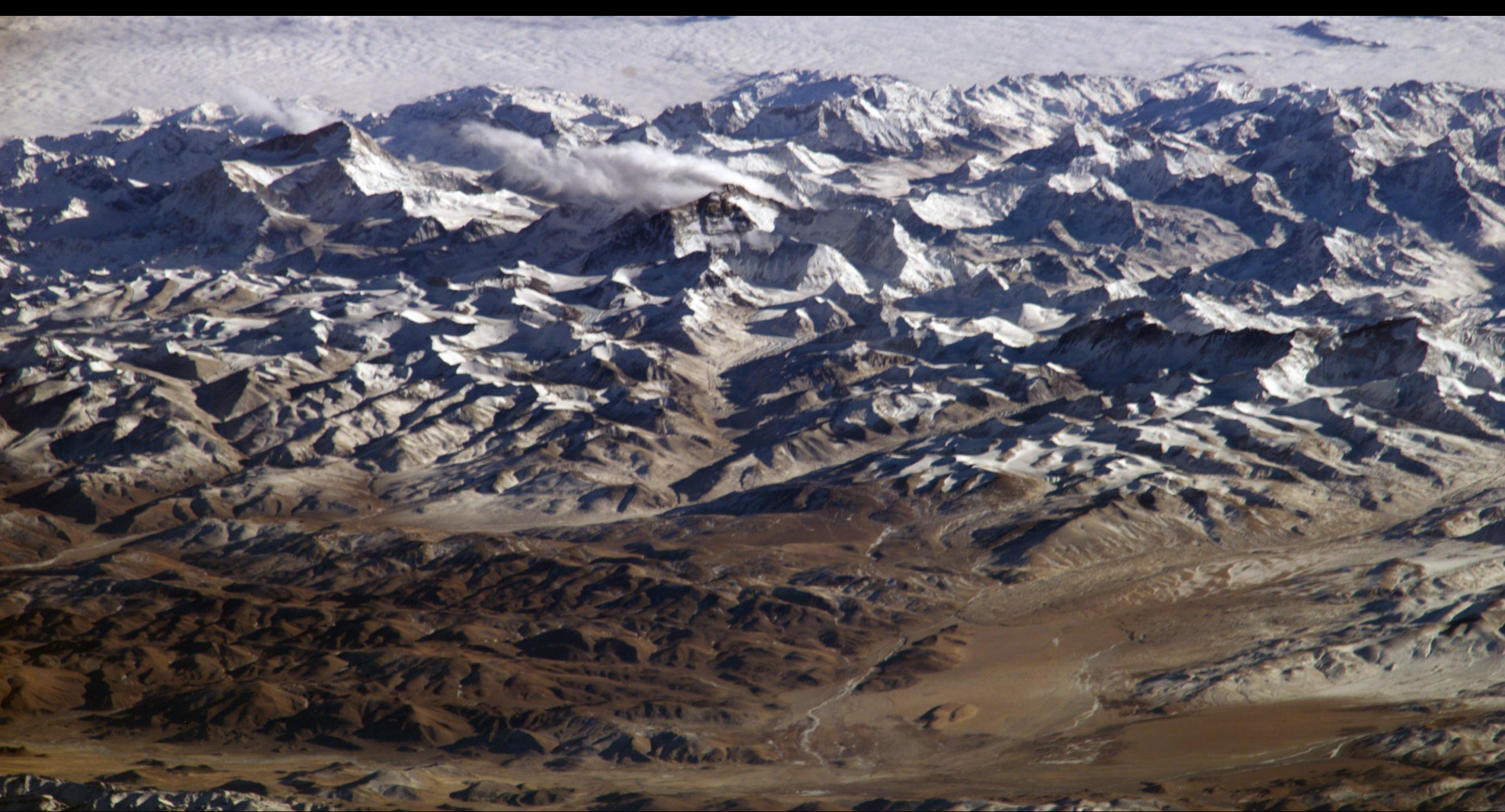
(d)



(e)

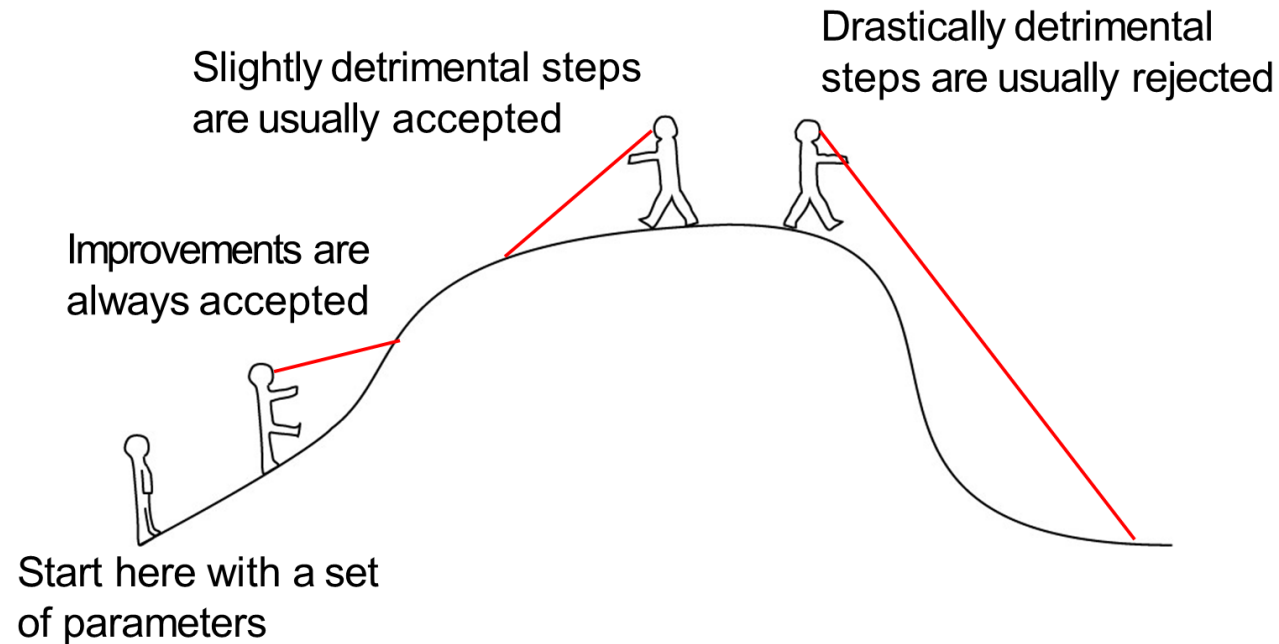


(f)



Bayesian MCMC approaches

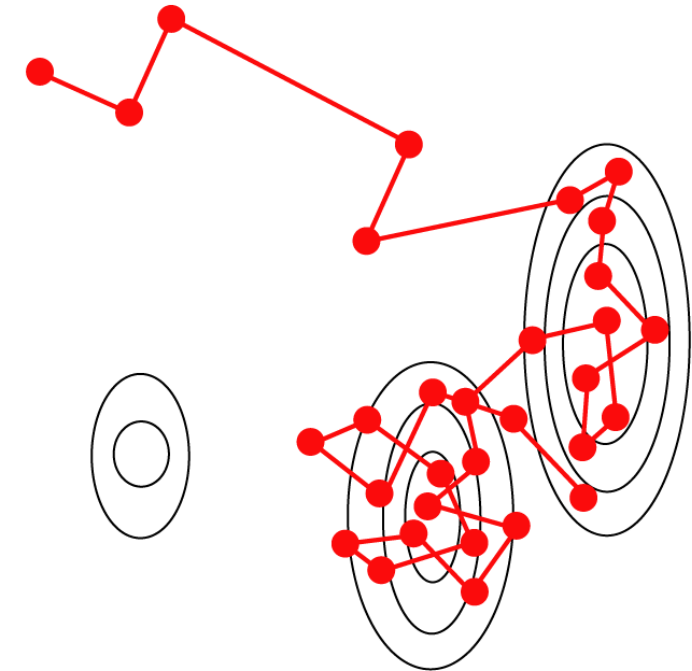
Metropolis-Hastings sampling of Markov Chain Monte Carlo (MCMC) chains



Nicholas Metropolis
Los Alamos, 1953



Wilfred Hastings
Uni of Toronto, 1970

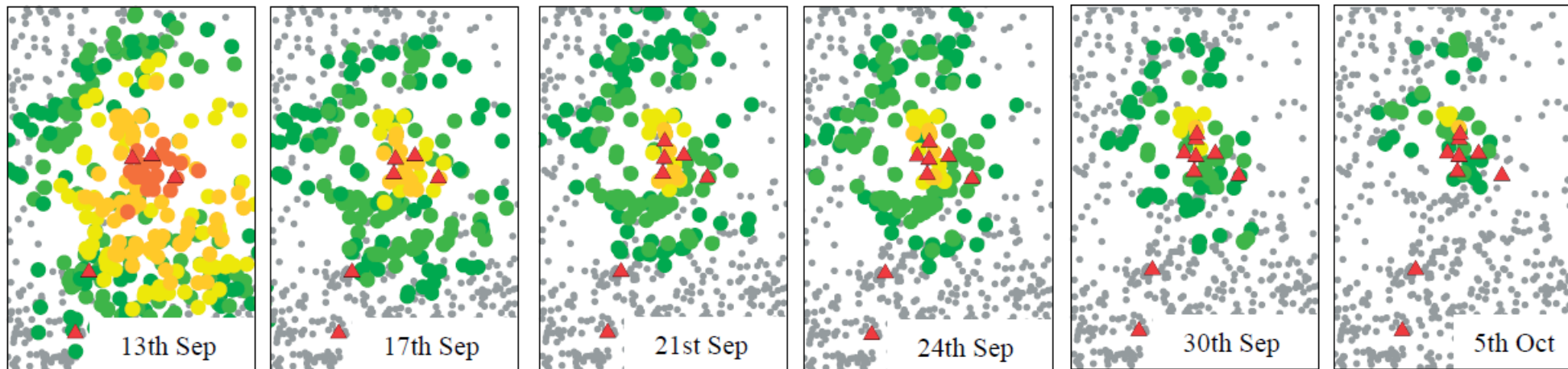


Predicting Undetected Infections

Full Bayesian RJ-MCMC approach

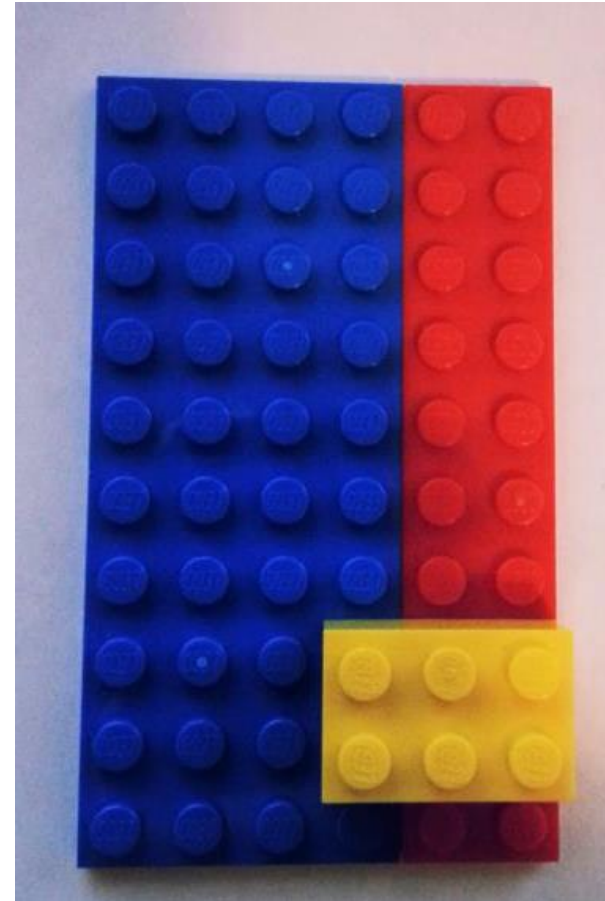
Infers the underlying epidemic process

Fed observed data at a real time point early in an epidemic, infers the likelihood that each premises is an undetected 'ghost' infection yet to be discovered



An intro to Bayesian thinking - using Lego

Rev. Thomas Bayes
(1702-1761)





The Bayesian approach

Bayes' theorem is a way to work out the likelihood of something in the face of some particular pieces of evidence.

Expresses how our belief in a hypothesis (h) should change to account for new evidence (e).

$$P(h|e) = \frac{P(e|h) \times P(h)}{P(e)}$$

$P(e|h)$ **likelihood** of evidence given the hypothesis

$P(h)$ **prior** belief of probability of h being true

$P(e)$ probability of the evidence

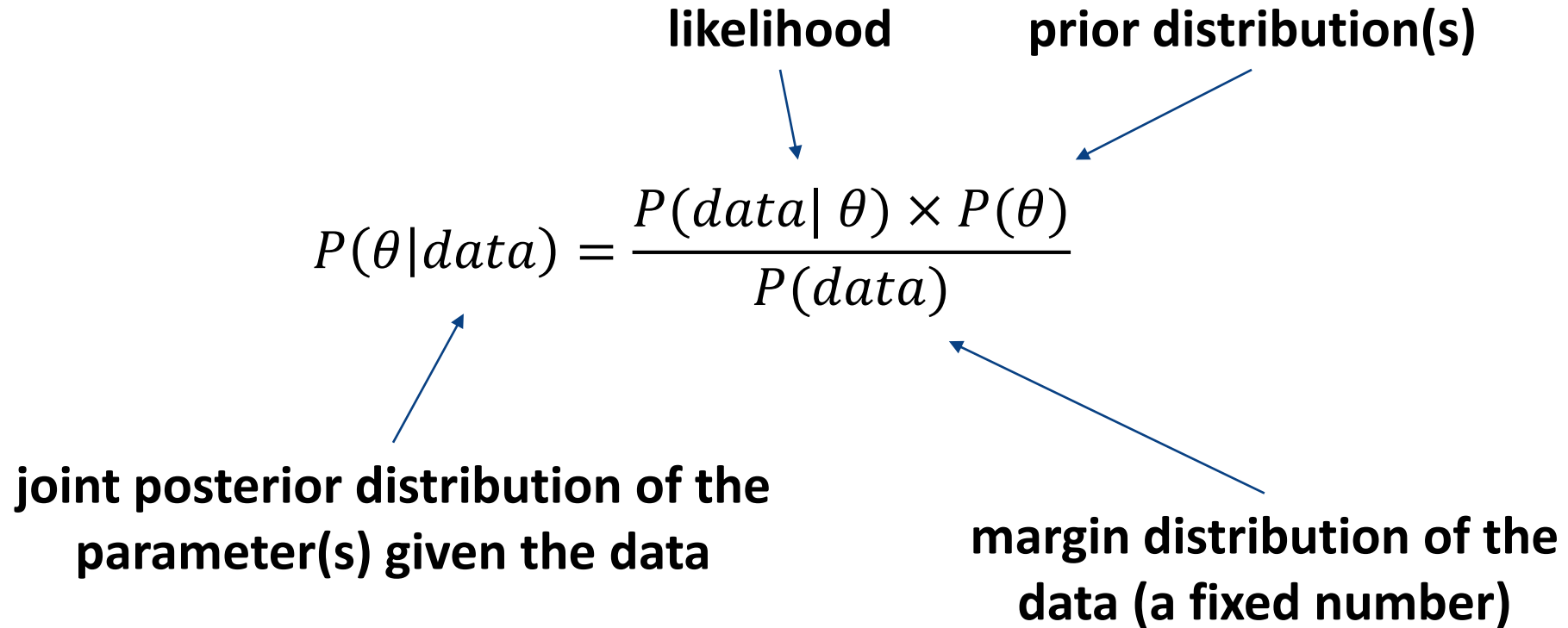
The Bayesian approach

Typically we're interested in the likelihood of some parameters (θ) given some data and priors.

$$P(\theta|data) = \frac{P(data|\theta) \times P(\theta)}{P(data)}$$

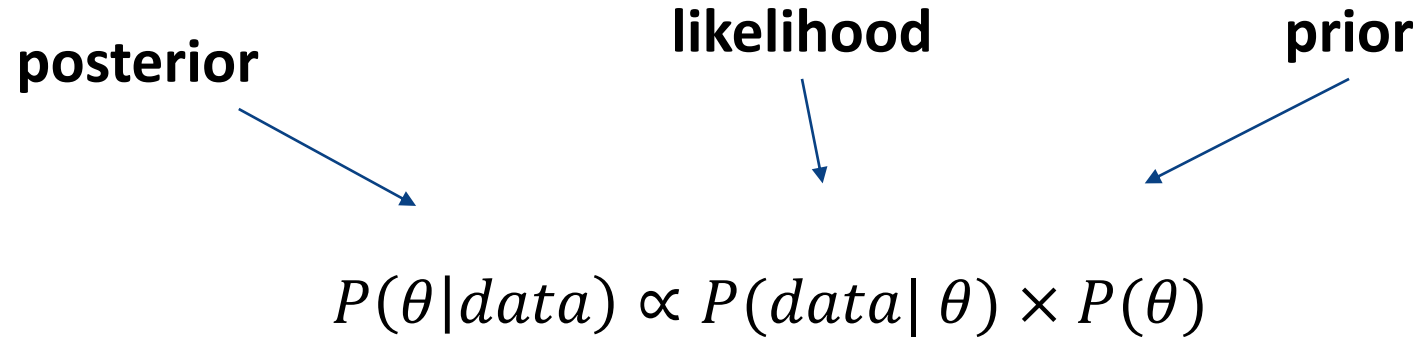
likelihood **prior distribution(s)**

joint posterior distribution of the parameter(s) given the data **margin distribution of the data (a fixed number)**



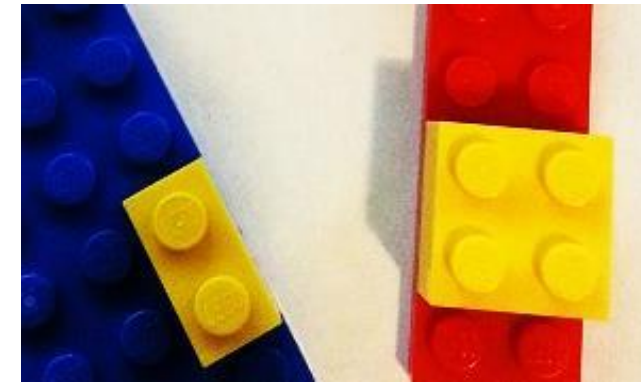
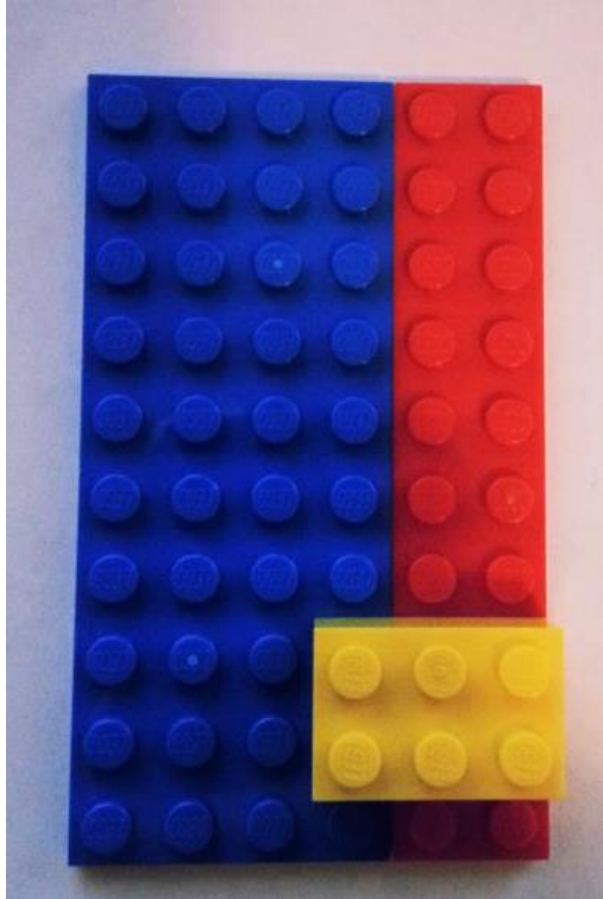
The Bayesian approach

The posterior is proportional to the likelihood and the prior



Bayes theorem with lego

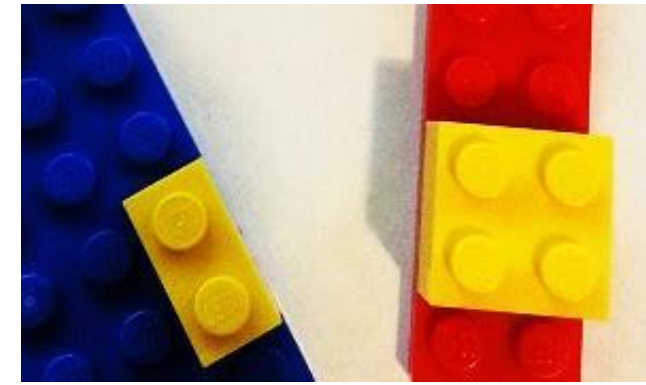
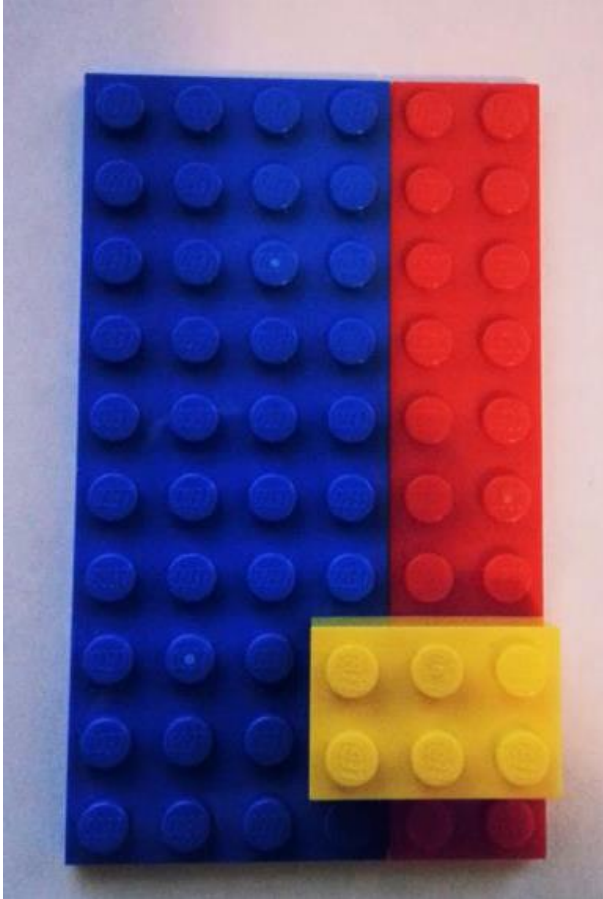
Background



If you find out you're on **red**, then what's the probability of being on **yellow**?

Bayes theorem with lego

Background



Priors:

$P(\text{blue}) =$

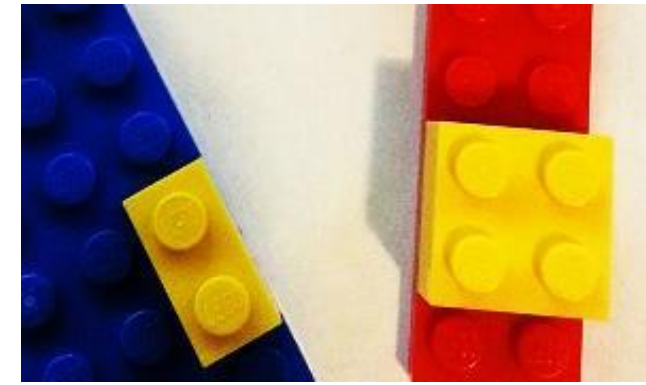
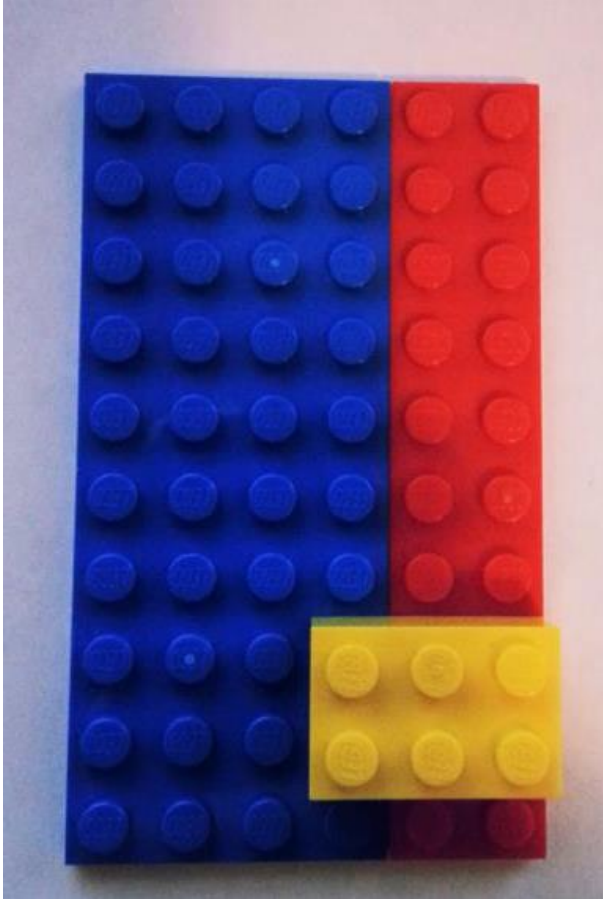
$P(\text{red}) =$

$P(\text{yellow}) =$

If you find out you're on **red**, then what's the probability of being on **yellow**?

Bayes theorem with lego

Background



Priors:

$$P(\text{blue}) = 40/60 = 66\%,$$

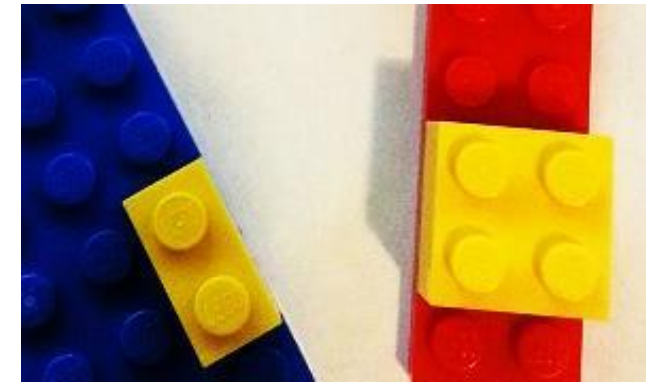
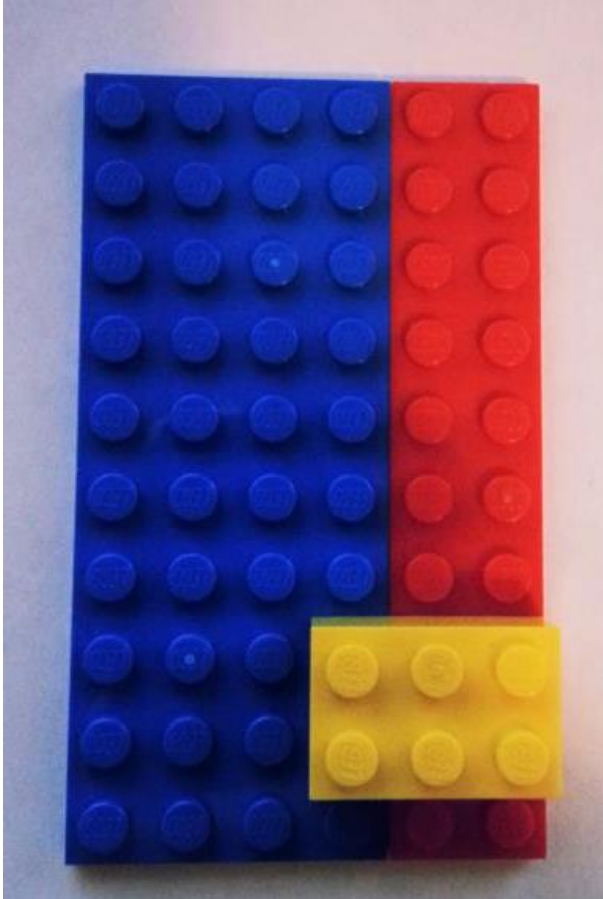
$$P(\text{red}) = 20/60 = 33\%,$$

$$P(\text{yellow}) = 6/60 = 10\%$$

If you find out you're on **red**, then what's the probability of being on **yellow**?

Bayes theorem with lego

Background



Priors:

$$P(\text{blue}) = 40/60 = 66\%,$$

$$P(\text{red}) = 20/60 = 33\%,$$

$$P(\text{yellow}) = 6/60 = 10\%$$

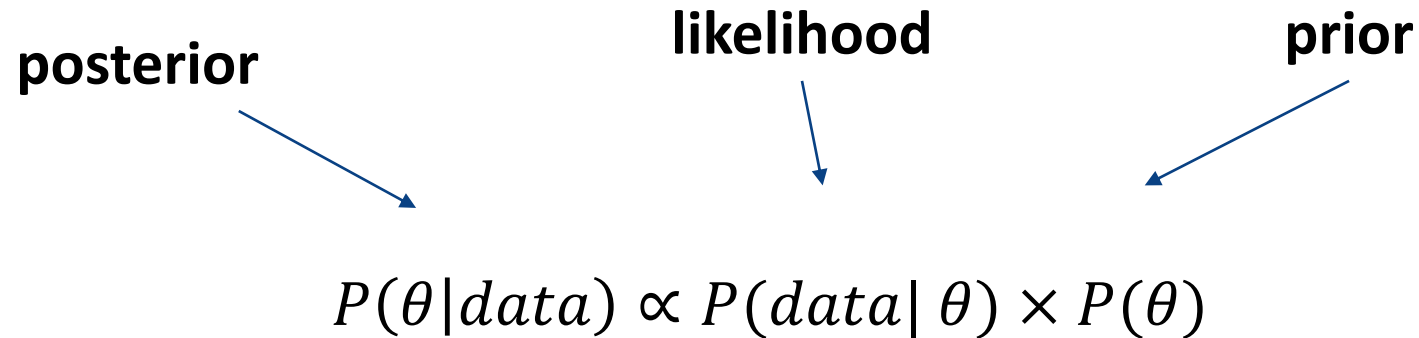
If you find out you're on **red**, then what's the probability of being on **yellow**?

$$P(y|r) = \frac{P(r|y) \times P(y)}{P(r)}$$

$$P(y|r) = \frac{4/6 \times 6/60}{20/60} = \frac{4}{20} = 20\%$$

The Bayesian approach

The posterior is proportional to the likelihood and the prior



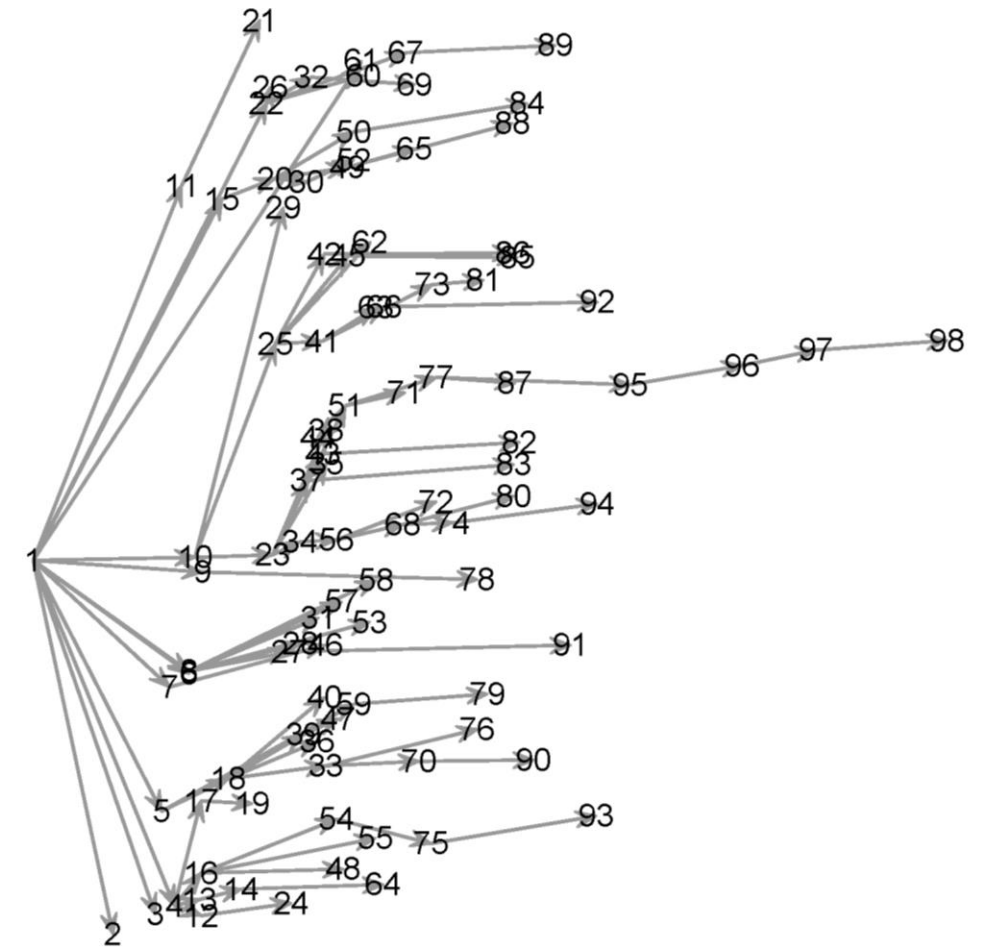
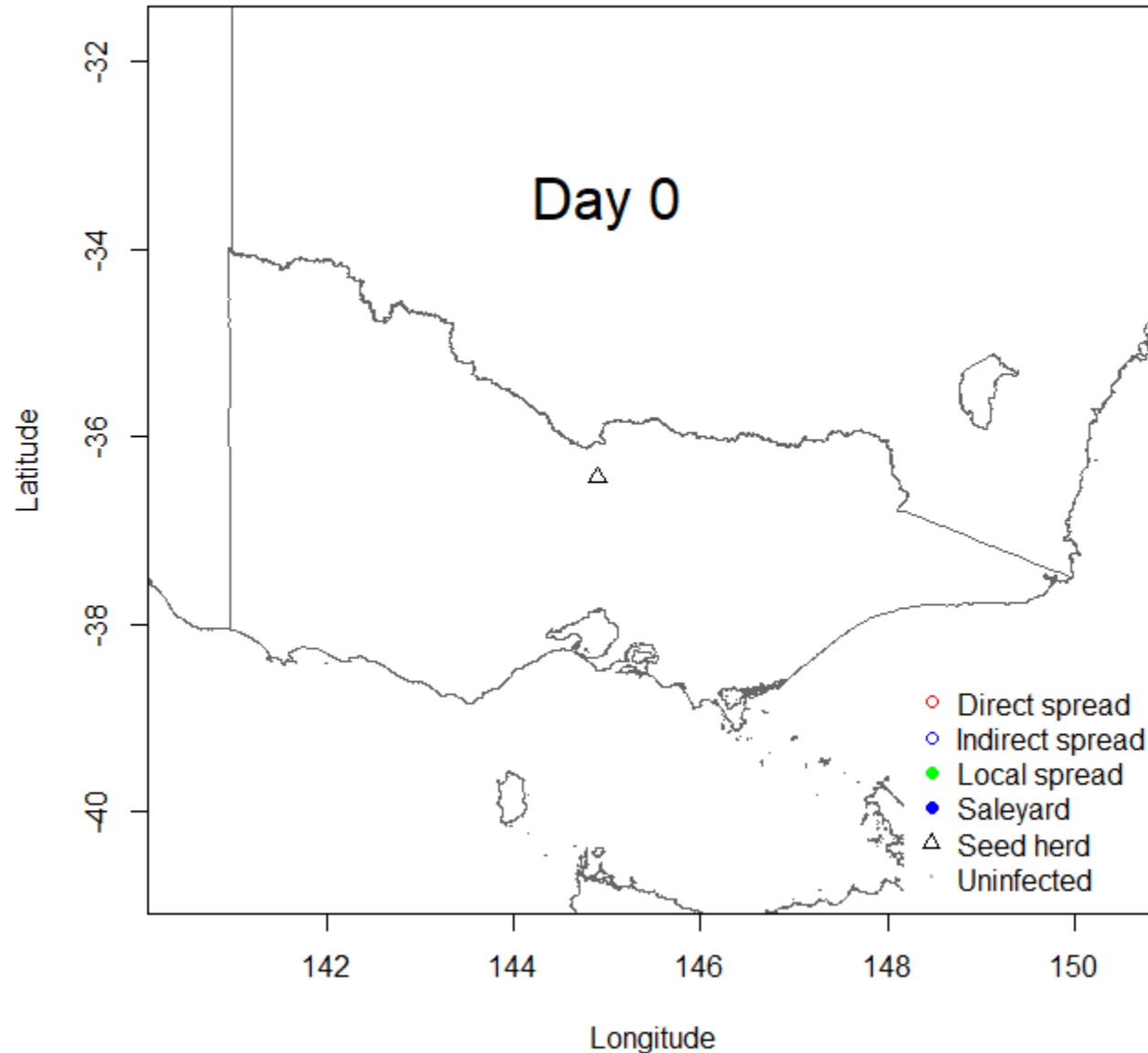
So we need three key things:

data, priors, and a likelihood specification (i.e. a model formulation)

Inferring who infected whom: the tools

- Frequentist
 - Cottam (genomic parsimony trees ranked by epi likelihood)
 - Cottam modified (+ spatial kernel + tracing data)
- Bayesian
 - Lau (C++, infers almost everything!)
 - **Lau + farm-level covariates + contact data + implemented as R package = BORIS**
 - SCoTTi (BEAST2) + within host model = BadTriP
 - Phybreak (R, partial inference)
 - Sampled ancestors (BEAST2, BDFSkyline)
 - Outbreaker 1 (R, partial) + spatial and contact data = Outbreaker 2
 - BeastLier (BEAST1, partitioned model)
 - TransPhyloR (R, RJ-MCMC)

Outbreak simulations: Australian Animal Disease Spread (AADIS) hybrid model (Bradhurst et al, 2015)

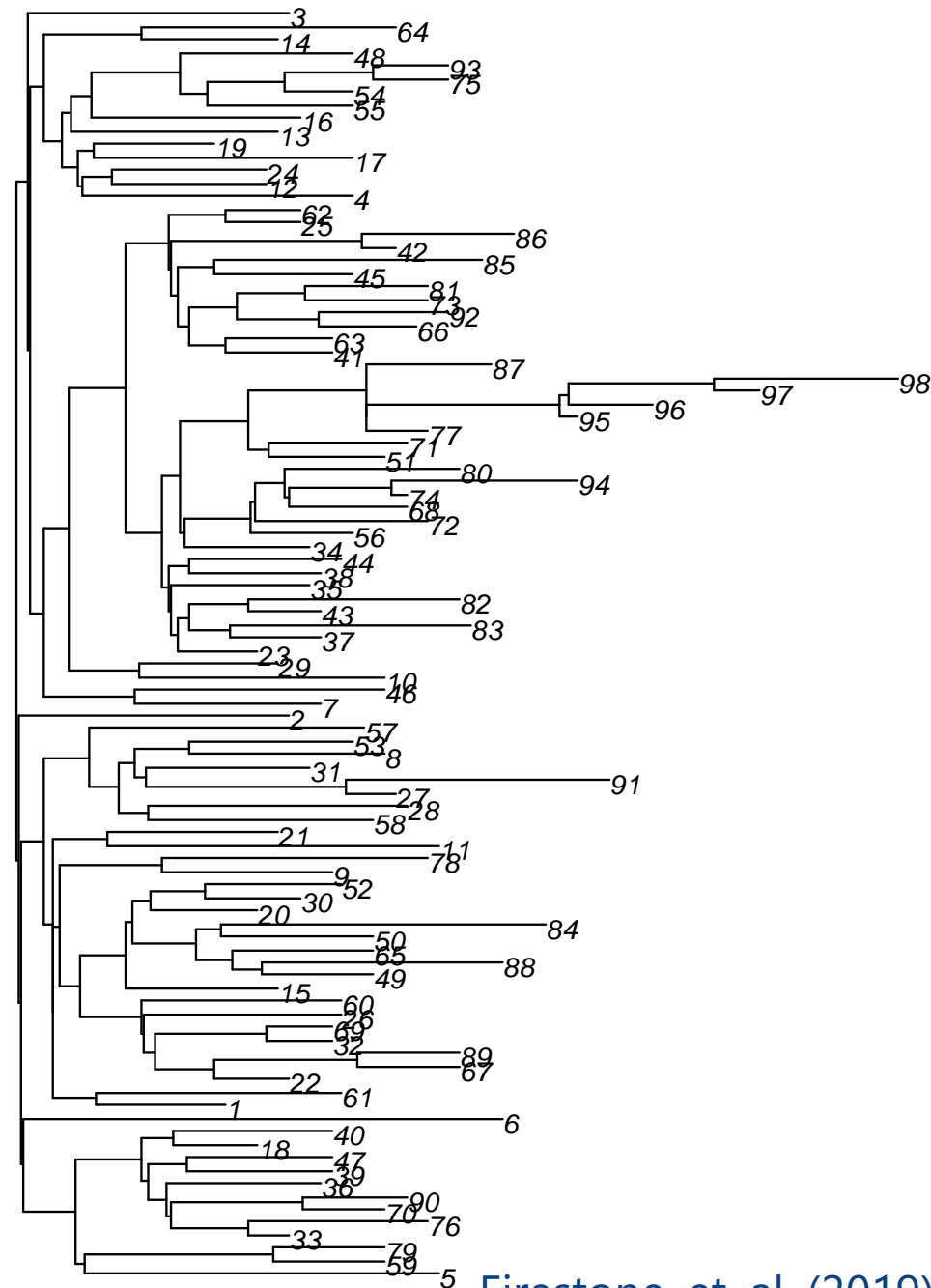




Methods: Genomic data simulations

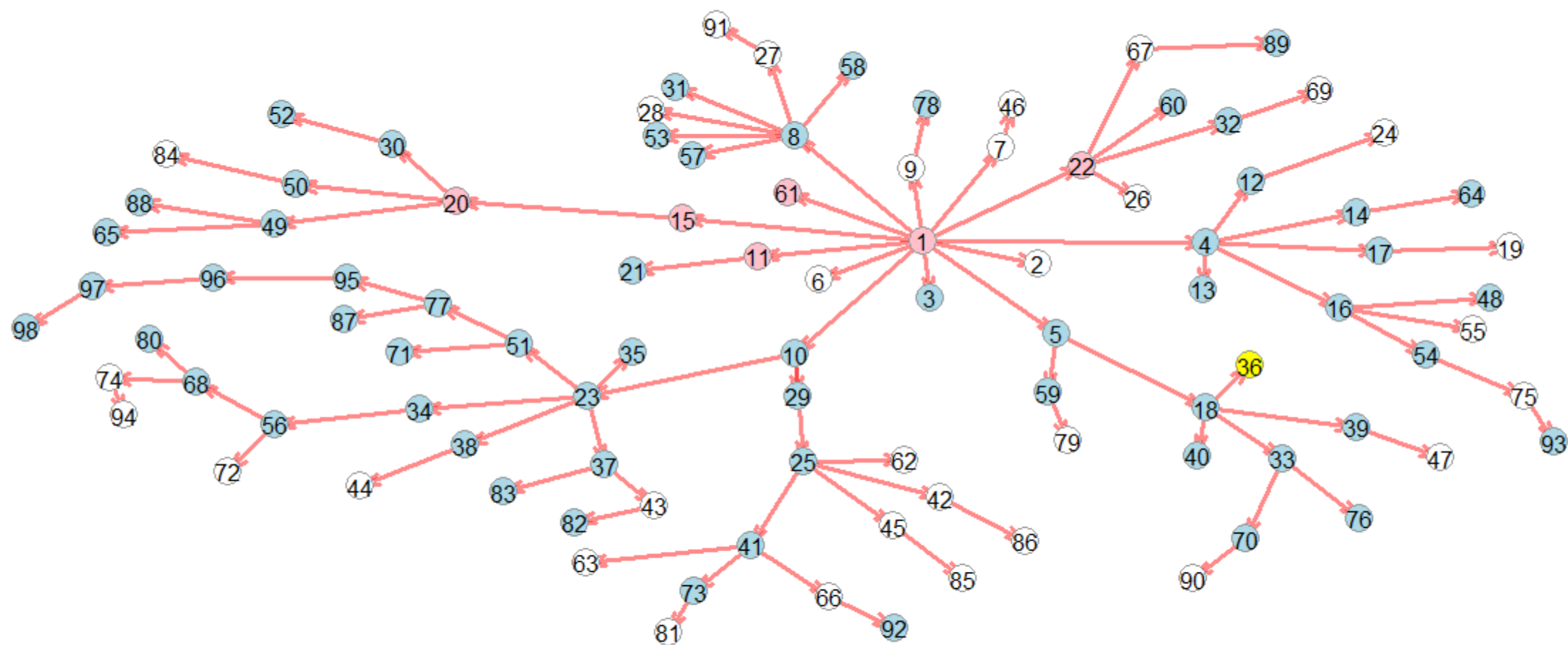
- Phylogenies simulated (nested in transmission tree): VirusTreeSimulator
- Sequences simulated: along phylogenies with Seq-Gen
- HKY model
- 2.168×10^{-5} changes site⁻¹ day⁻¹
- TS/TV = 7.61

(Cottam et al., 2006, 2008; Juleff et al., 2013)

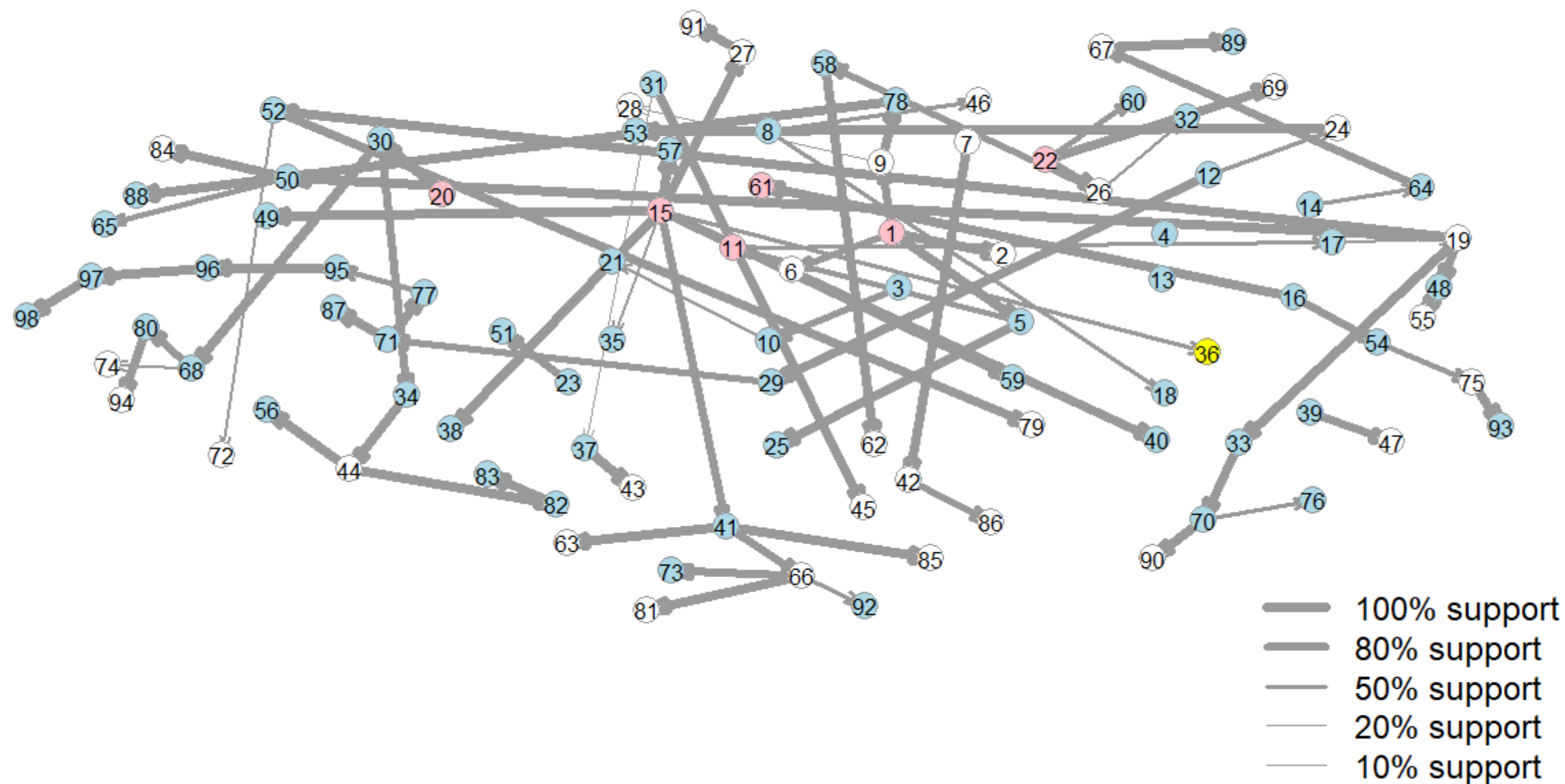


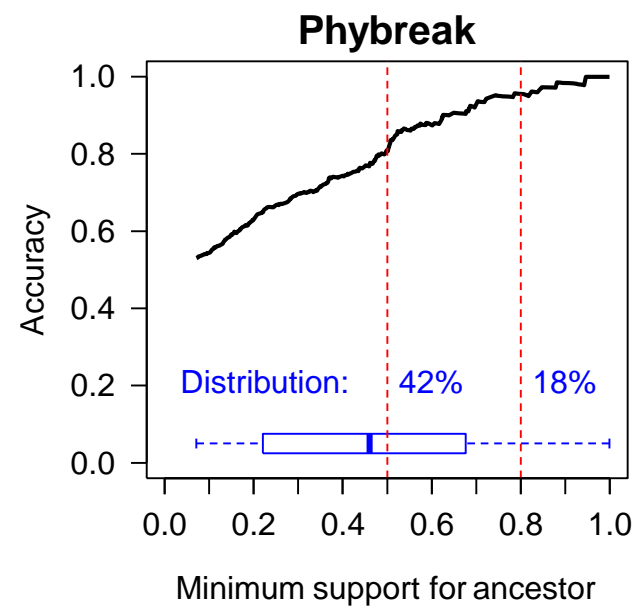
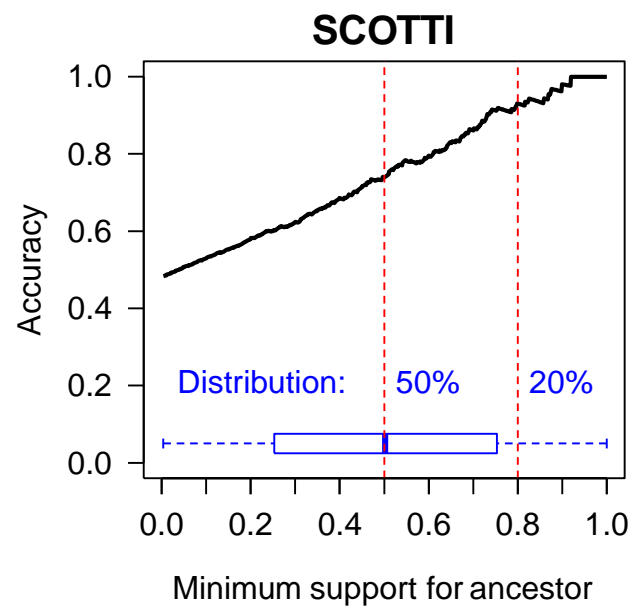
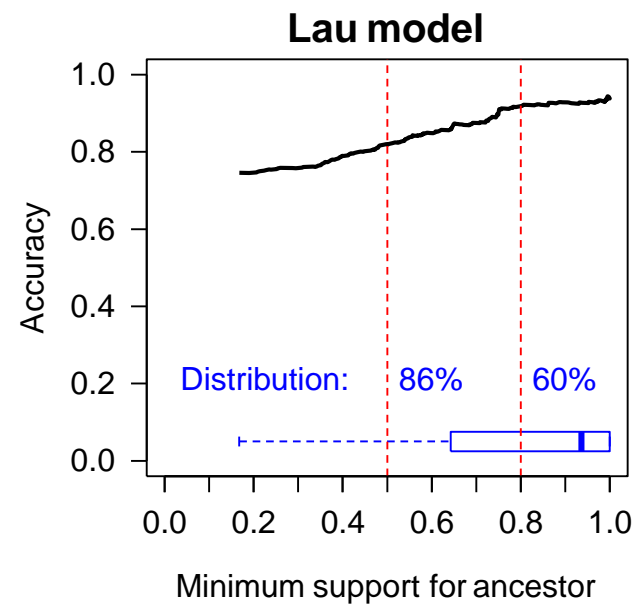
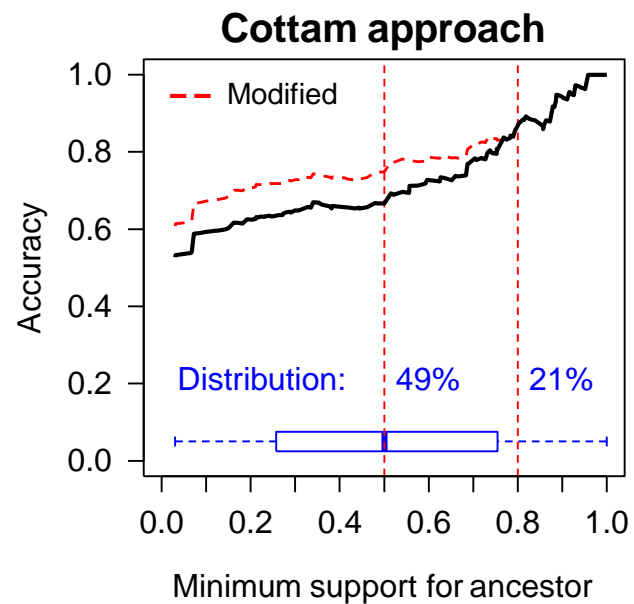
Firestone, et. al. (2019)

True network in arbitrary space



Lau model inferred network (iteration = 100)





Extending the model: BORIS

BORIS: R package for Bayesian Outbreak Reconstruction Inference and Simulation

<https://github.com/sfires/boris>



Modifying the model

Max Lau's original model: the total probability of individual j becoming infected during time period $[t, t + dt]$:

$$P(j, t, dt) = \left\{ \alpha + \sum_{i \in \xi_I(t)} \beta k_{d_{ij}} \right\} dt$$

1° (background)
transmission rate

2° transmission rate
(just based on spatial kernel)

$$P(j, t, dt) = \left\{ \alpha + \sum_{i \in \xi_I(t)} \left(\beta k_{d_{ij}} \times n_i^v Inf_i \cdot n_j^p Susc_j \right) + \sum_{i \in \xi_I(t)} \left(\beta_{mov} \cdot x_{ij} \right) \right\} dt$$



spatial
kernel

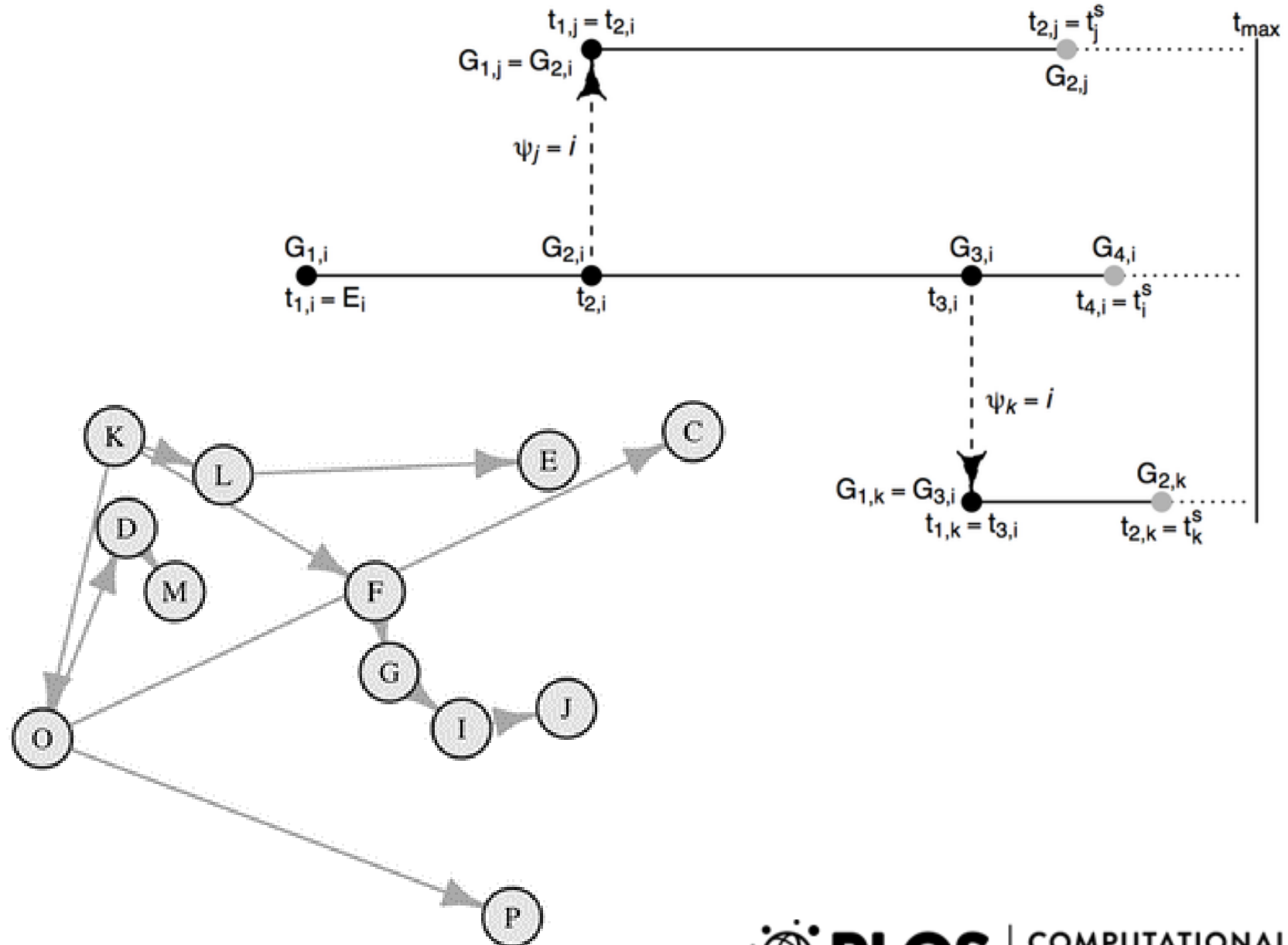


number and type of
susceptible animals



contact-traced
movements

A Systematic Bayesian Integration of Epidemiological and Genetic Data (Lau et al, 2015)



Likelihood components:

In the general multiple-cluster scenario, with complete data $z = (\mathbf{E}, \mathbf{I}, \mathbf{R}, \mathbf{G}, \boldsymbol{\psi})$ and model parameters $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2, p)$, we can express the likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}; z) = & \prod_{j \in \chi_E^{-1}} P(j, \psi_j) \times \exp \{-q_j(E_j)\} \times \prod_{j \in \chi_S} \exp \{-q_j(t_{max})\} \\ & \times \prod_{j \in \chi_I} f_E(I_j - E_j; a, b) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\ & \times \prod_{j \in \chi_{E \setminus I}} \{1 - F_E(t_{max} - E_j; a, b)\} \times \prod_{j \in \chi_{I \setminus R}} \{1 - F_I(t_{max} - I_j; \gamma, \eta)\} \\ & \times \prod_{j \in \chi_E} g(G_{2,j}, \dots, G_{m_j,j} | t_{j,j}, \psi_j, G_{1,j}) \times \prod_{j \in \chi_E} h(G_{1,j} | \psi_j). \end{aligned}$$

Likelihood components:

In the general multiple-cluster scenario, with complete data $z = (\mathbf{E}, \mathbf{I}, \mathbf{R}, \mathbf{G}, \boldsymbol{\psi})$ and model parameters $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2, p)$, we can express the likelihood as

$$\begin{aligned}
 L(\boldsymbol{\theta}; z) = & \prod_{j \in \chi_E^{-1}} P(j, \psi_j) \times \exp \{-q_j(E_j)\} \times \prod_{j \in \chi_S} \exp \{-q_j(t_{max})\} \\
 & \times \prod_{j \in \chi_I} f_E(I_j - E_j; a, b) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\
 & \times \prod_{j \in \chi_{E \setminus I}} \{1 - F_E(t_{max} - E_j; a, b)\} \times \prod_{j \in \chi_{I \setminus R}} \{1 - F_I(t_{max} - I_j; \gamma, \eta)\} \\
 & \times \prod_{j \in \chi_E} g(G_{2,j}, \dots, G_{m_j,j} | t_{j,j}, \psi_j, G_{1,j}) \times \prod_{j \in \chi_E} h(G_{1,j} | \psi_j).
 \end{aligned}$$

Likelihood components:

In the general multiple-cluster scenario, with complete data $z = (\mathbf{E}, \mathbf{I}, \mathbf{R}, \mathbf{G}, \boldsymbol{\psi})$ and model parameters $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2, p)$, we can express the likelihood as

$$L(\boldsymbol{\theta}; z) = \prod_{j \in \chi_E^{-1}} \underbrace{P(j, \psi_j)}_{\textcircled{1}} \times \underbrace{\exp \{-q_j(E_j)\}}_{\textcircled{2}} \times \prod_{\textcircled{3} \in \chi_S} \exp \{-q_j(t_{\max})\}$$

① Likelihood of proposed source at this step in the MCMC, i.e.:

infection of j by ψ_j for all χ_E^{-1} exposed individuals (farms) excluding that with the earliest inferred onset time (which must be a primary infection), which depends on a background transmission rate (α) for primary sources, $l))\}$

or a secondary transmission rate and a spatial kernel for secondary cases:

$$\beta K(d_{\psi_j j}; \kappa)$$

Likelihood components:

In the general multiple-cluster scenario, with complete data $z = (\mathbf{E}, \mathbf{I}, \mathbf{R}, \mathbf{G}, \boldsymbol{\psi})$ and model parameters $\boldsymbol{\theta} = (\alpha, \beta, a, b, \gamma, \eta, \kappa, \mu_1, \mu_2, p)$, we can express the likelihood as

$$L(\boldsymbol{\theta}; z) = \prod_{j \in \chi_E^{-1}} \underbrace{P(j, \psi_j)}_{\text{①}} \times \underbrace{\exp \{-q_j(E_j)\}}_{\text{②}} \times \prod_{\text{③} \in \chi_S} \exp \{-q_j(t_{\max})\}$$

- ② Likelihood that each exposed individual χ_E survived unexposed χ_S until its **inferred exposure time**, assuming the below:
- ③ Likelihood that each unexposed individual survived unexposed χ_S until the end of simulated time (t_{\max}), assuming the below:

$$q_j(s) = \int_0^s \left\{ \alpha + \sum_{i \in \xi_I(t)} \beta K(d_{ij}; \kappa) \right\} dt$$

Likelihood components:

$$\begin{aligned} & \times \prod_{j \in \chi_I} f_E(I_j - E_j; a, b) \times \prod_{j \in \chi_R} f_I(R_j - I_j; \gamma, \eta) \\ & \times \prod_{j \in \chi_{E \setminus I}} \{1 - F_E(t_{\max} - E_j; a, b)\} \times \prod_{j \in \chi_{I \setminus R}} \{1 - F_I(t_{\max} - I_j; \gamma, \eta)\} \end{aligned}$$

④ & ⑥ Likelihood contributions of **inferred latent periods** of each of the individuals that became infectious

⑤ & ⑦ Likelihood contributions of **inferred infectious periods** of each of the individuals that recovered or was removed



Thank you ... questions



Acknowledgements

- This research was supported by an Australian Research Council Discovery Early Career Researcher Award (project number DE160100477)
- Components of this research were undertaken on the Melbourne Bioinformatics Cluster (grant number VR0277)
- National Institute of Animal Health, NARO, Japan.
- Graeme Garner (Australian Govt, DAWR)
- Asia-Pacific Centre for Animal Health (APCAH)
- Chris Jewell, Finlay Campbell, Sebastián Duchêne, Soren Alexandersen, Anthony Chamings, Neil Young



OIE Collaborating Centre
Diagnostic Test Validation Science



Australian Government
Department of Agriculture
and Water Resources



Australian Government
Australian Research Council



simon.firestone@unimelb.edu.au

<http://fvas.unimelb.edu.au/veam>