# Likelihood-based inference for partially observed processes, with applications to genetic sequence data, panel data and spatiotemporal data

Edward Ionides
University of Michigan, Department of Statistics

Biostatistics Seminar
Ohio State University
Friday 12th April, 2019

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/osu19.pdf`

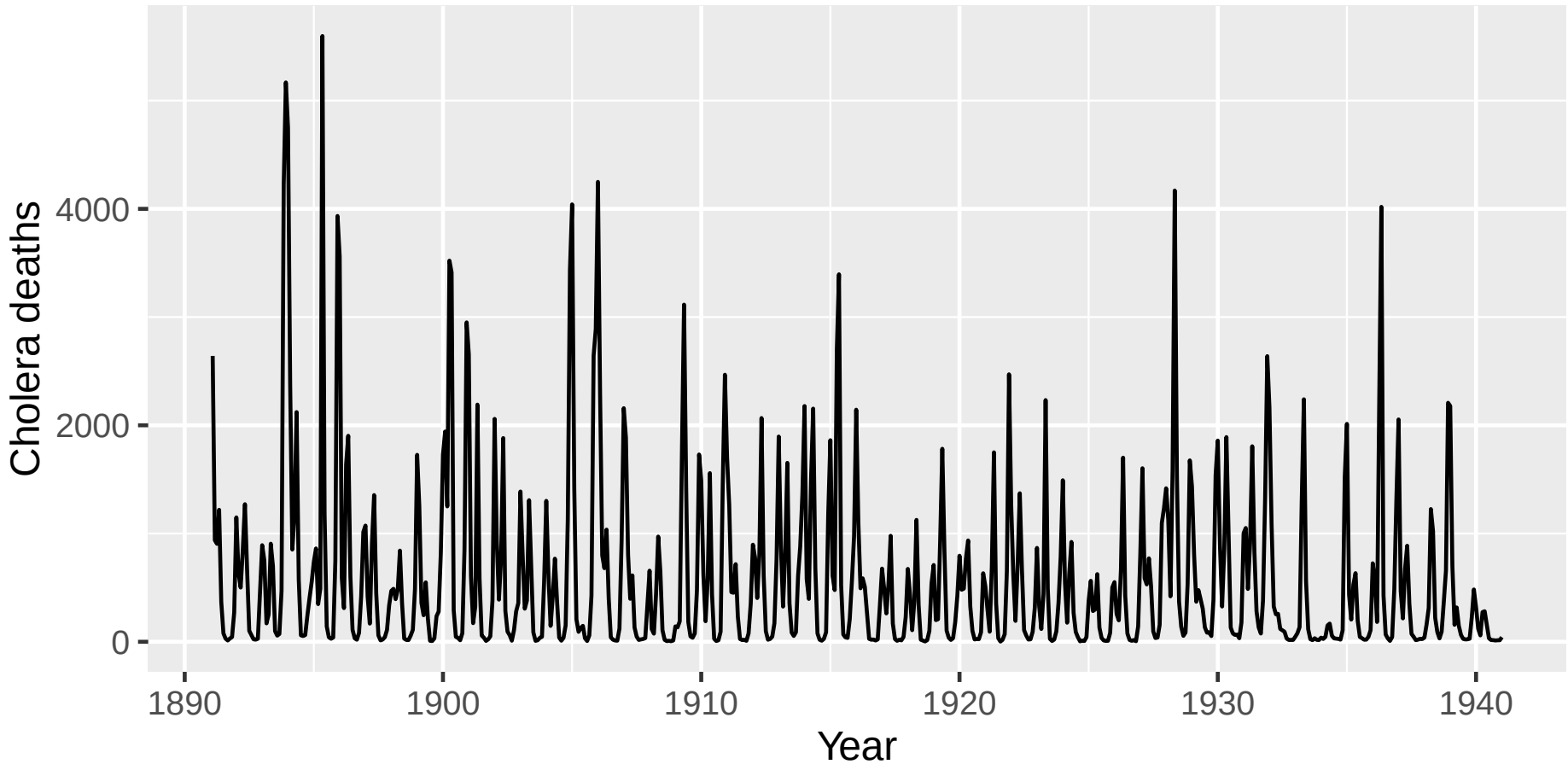
# Four motivating data analysis challenges

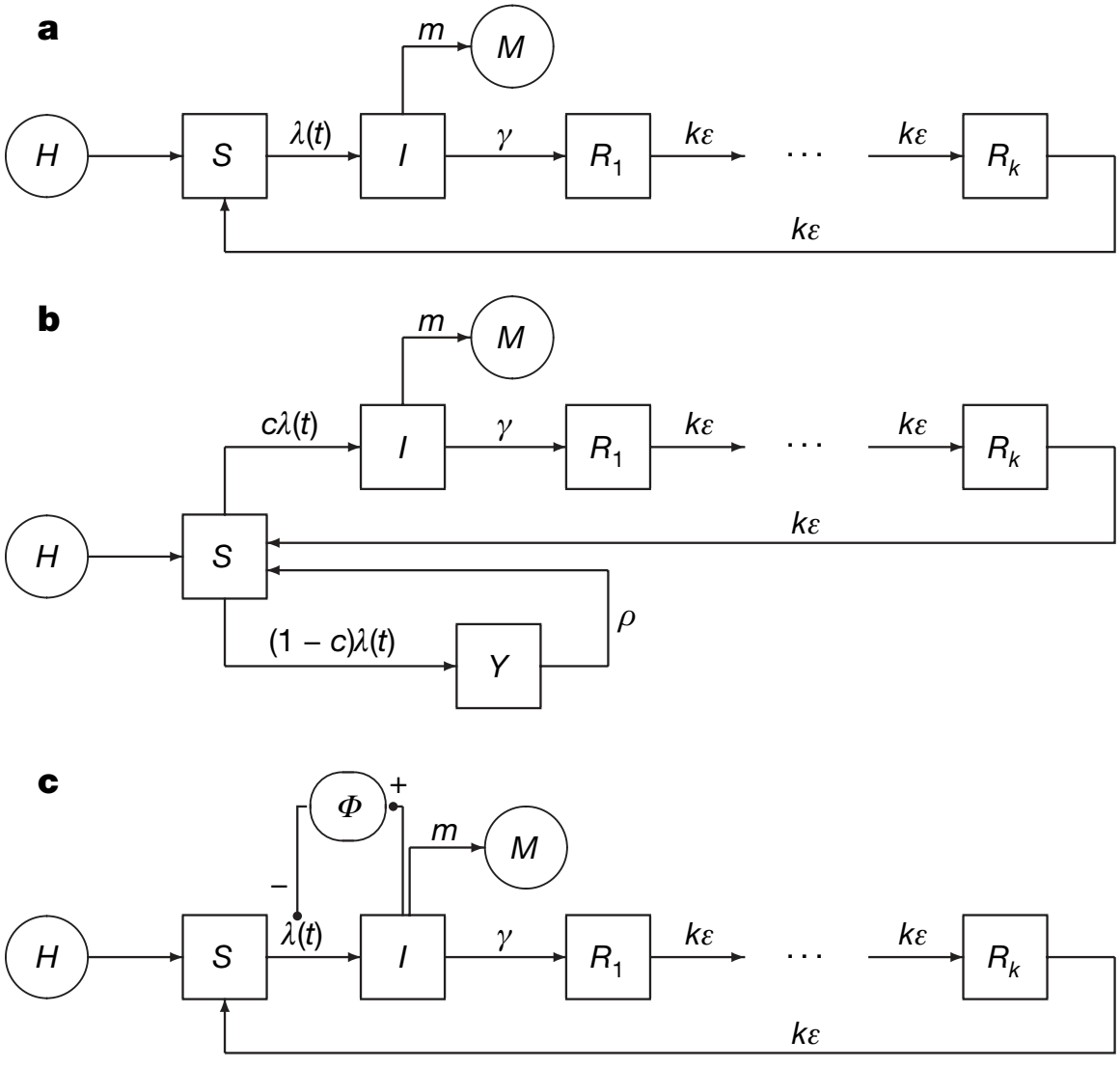
1. Time series analysis: cholera in Bangladesh.
   - Fitting nonlinear dynamic models to a single long time series.
2. Panel data analysis: dynamic variation in sexual contact rates.
   - Observations on a collection of units lead to a panel of time series.
   - Analyzed together, the panel strengthens inferences available from any one time series.
3. Genetic sequence data: HIV transmission within and between demographic groups.
   - Genetic sequences of pathogens can inform transmission relationships between infected hosts.
4. Spatiotemporal analysis: dengue in Rio de Janeiro.
   - Coupling between spatial locations leads to high-dimensional dynamics.

# Commonalities between these four examples

- All examples are partially observed Markov process (POMP) models.
- A POMP model involves a latent dynamic process with the Markov property: the future given the current state does not depend on the past.
- Only noisy and incomplete measurements are available on the latent process.
- Sequential Monte Carlo (SMC) algorithms provide a widely applicable approach for low-dimensional systems.
- Extensions to SMC are required for higher dimensional systems.

# Monthly cholera deaths in Dhaka, Bangladesh, 1891-1940






## Competing POMP models

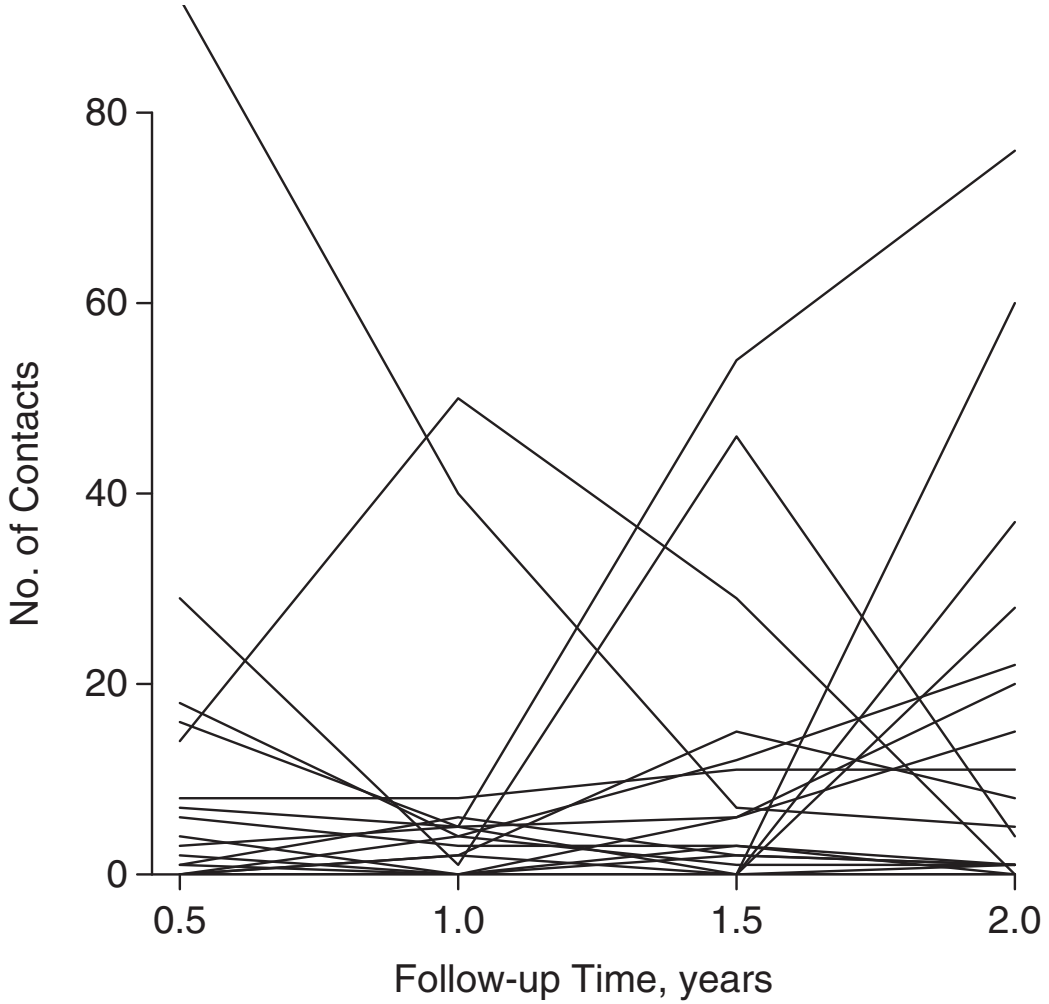
(King et al., 2008)

| | |
|---|---|
| $S$ | Susceptible |
| $I$ | Infected |
| $R_j$ | Recovered |
| $M$ | Mortality |
| $H$ | Population size |
| $Y$ | Asymptomatics in **b** |
| $\Phi$ | Phage in **c** |
| $\lambda$ | force of infection |
| $\gamma$ | recovery rate |
| $\epsilon$ | loss of immunity |
| $m$ | cholera mortality |

# 2. Panel data on sexual contacts

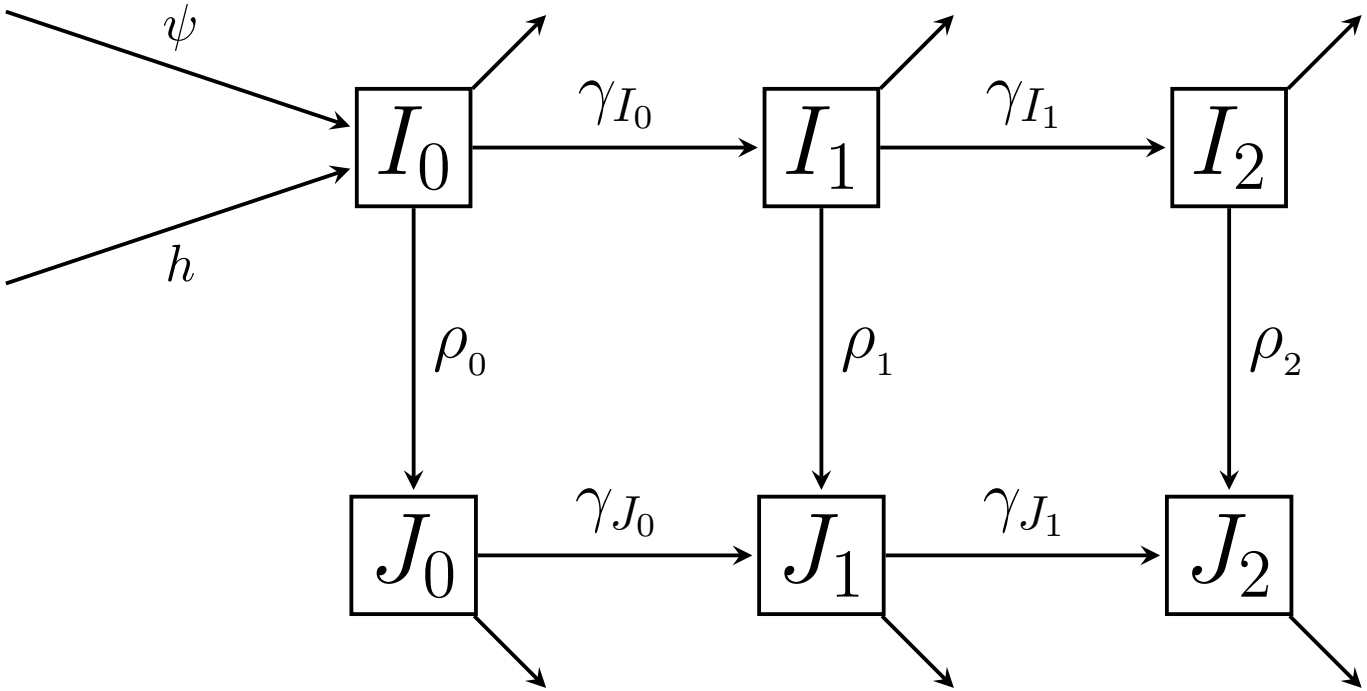
- Mathematical models of HIV transmission struggle to explain observed incidence due to the low measured probability of transmission per sexual contact.
- The anomaly can be resolved by models that include individual-level variability in sexual behavior over time.
- Romero-Severson et al. (2015) constructed behavioral models with various heterogeneities, both between individuals and within individuals over time. These models were fitted to behavioral panel data.
- Collections of POMP models with some shared parameters, but no dynamic interactions, are called **panelPOMP** models.

# Total sexual contacts in 6 month intervals



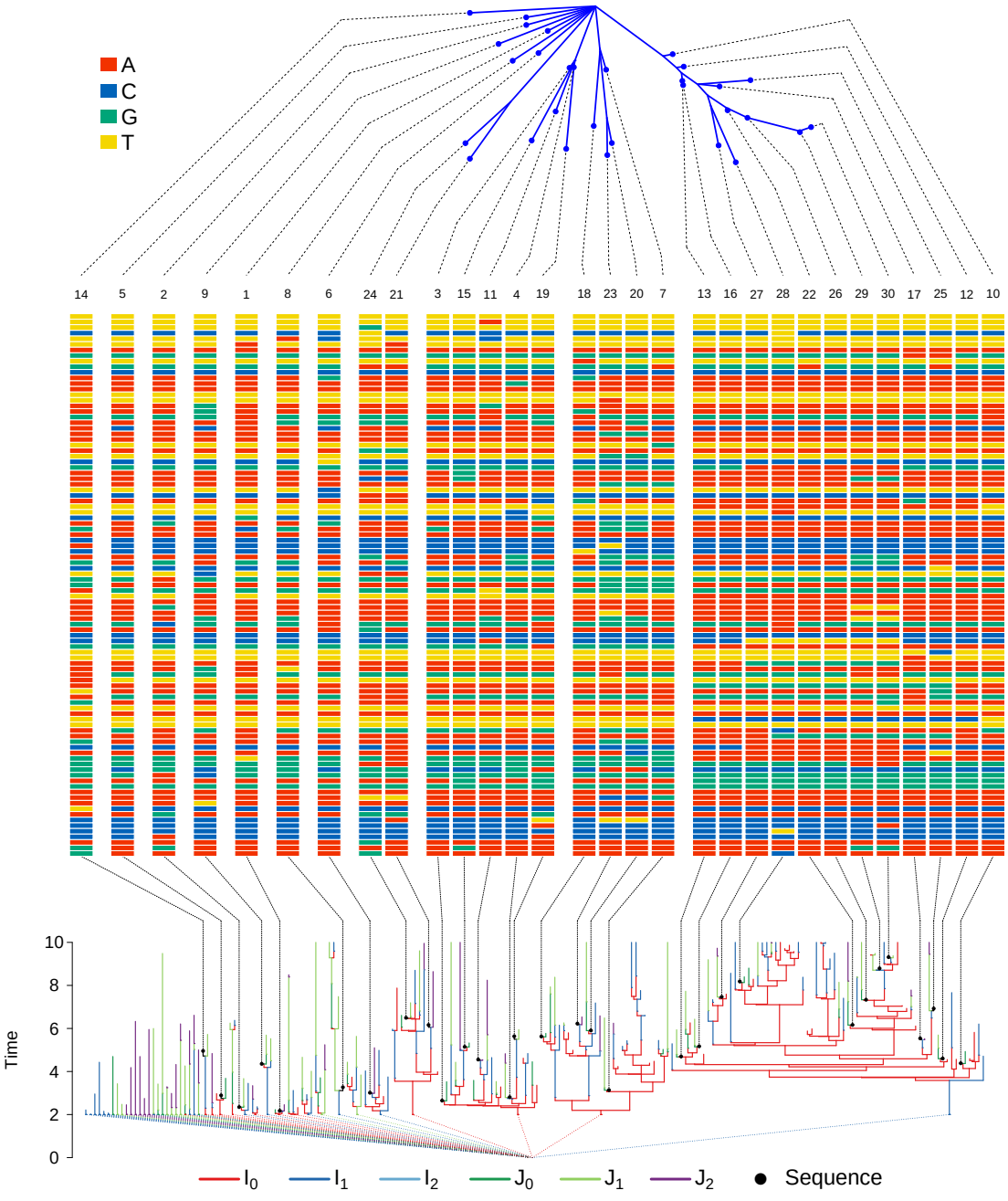

- Time series for 15 units from a panel of 882 gay men who completed a 2 year longitudinal study.
- Sexual contacts were reported in various categories: oral, anal, protected, unprotected, etc. Here, we show total reported contacts.

# 3. Infectious disease dynamics inferred from genetic data




A flow diagram for HIV.

- $I_k$ classes represent undiagnosed infections.
- $J_k$ classes represent diagnosed infections.
- $k = 0, 1, 2$ denotes early, chronic and AIDS stages.
- Infection can come from within, or outside, the study population.
- Genetic data give evidence on infectors as well as infectees.




**A simulated HIV epidemic** (Smith et al., 2017)

Top: phylogeny of observed sequences.

Middle: simulated sequence data from a fitted model.

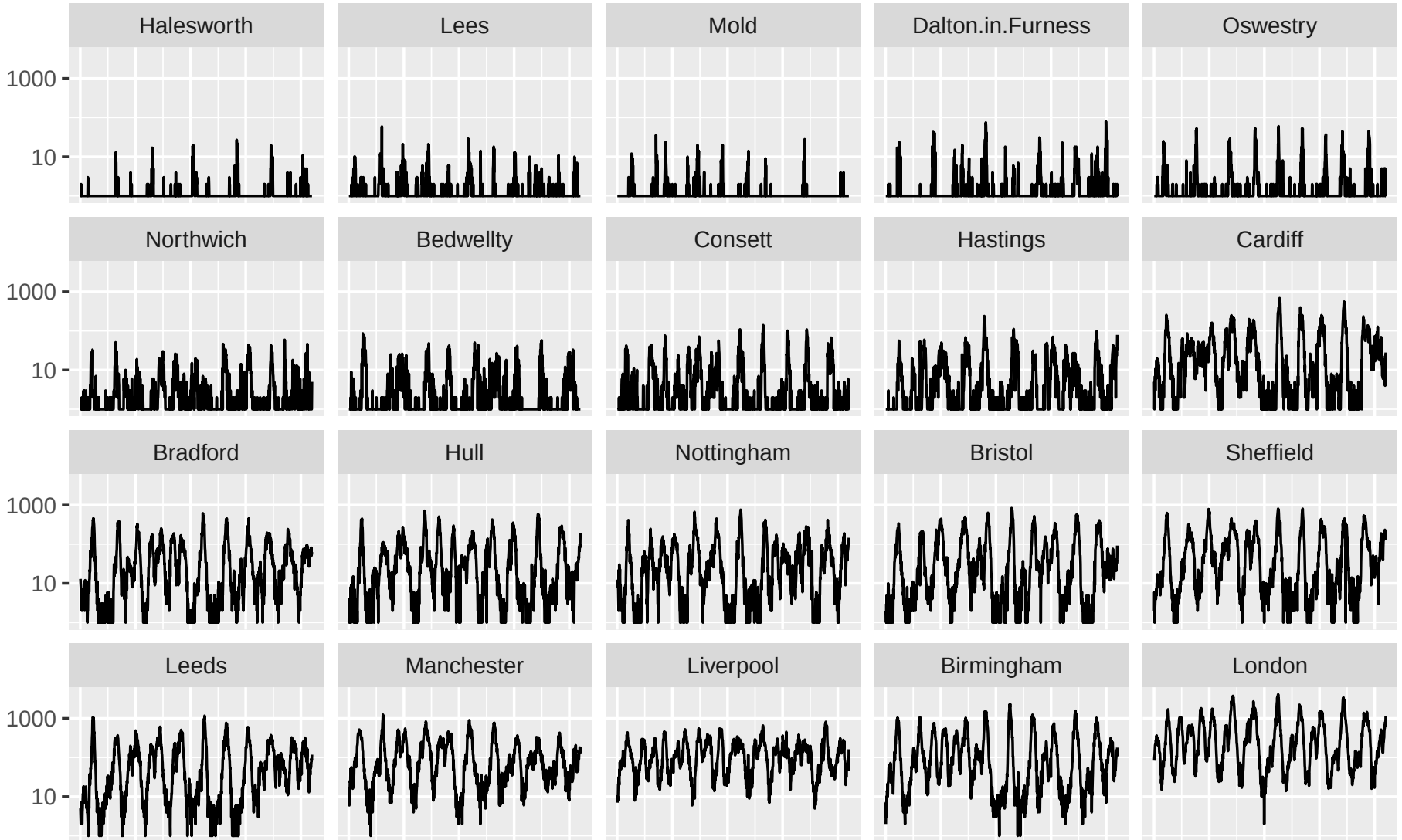
Bottom: Transmission forest for the simulated epidemic.

red: undiagnosed early infection

blue: undiagnosed chronic infection

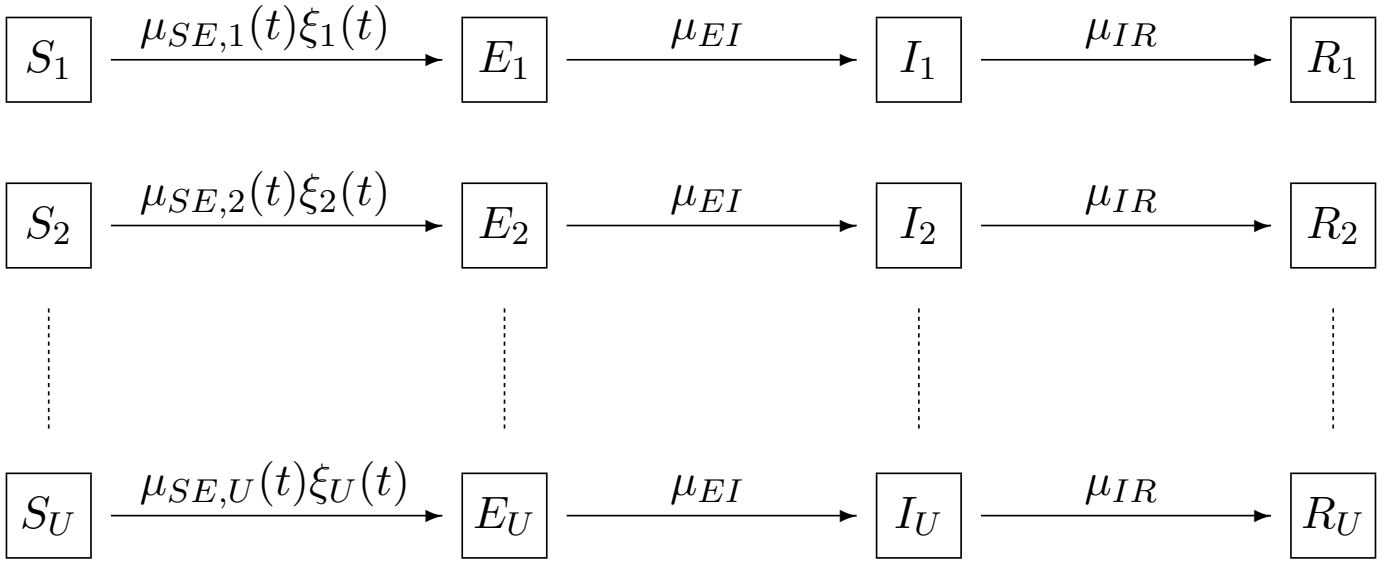
green: diagnosed

# Measles in 20 UK cities, 1944–1965




- A few cities are above a critical community size for sustaining measles.
- What is the effective transmission network between cities?

# A SpatPOMP for measles

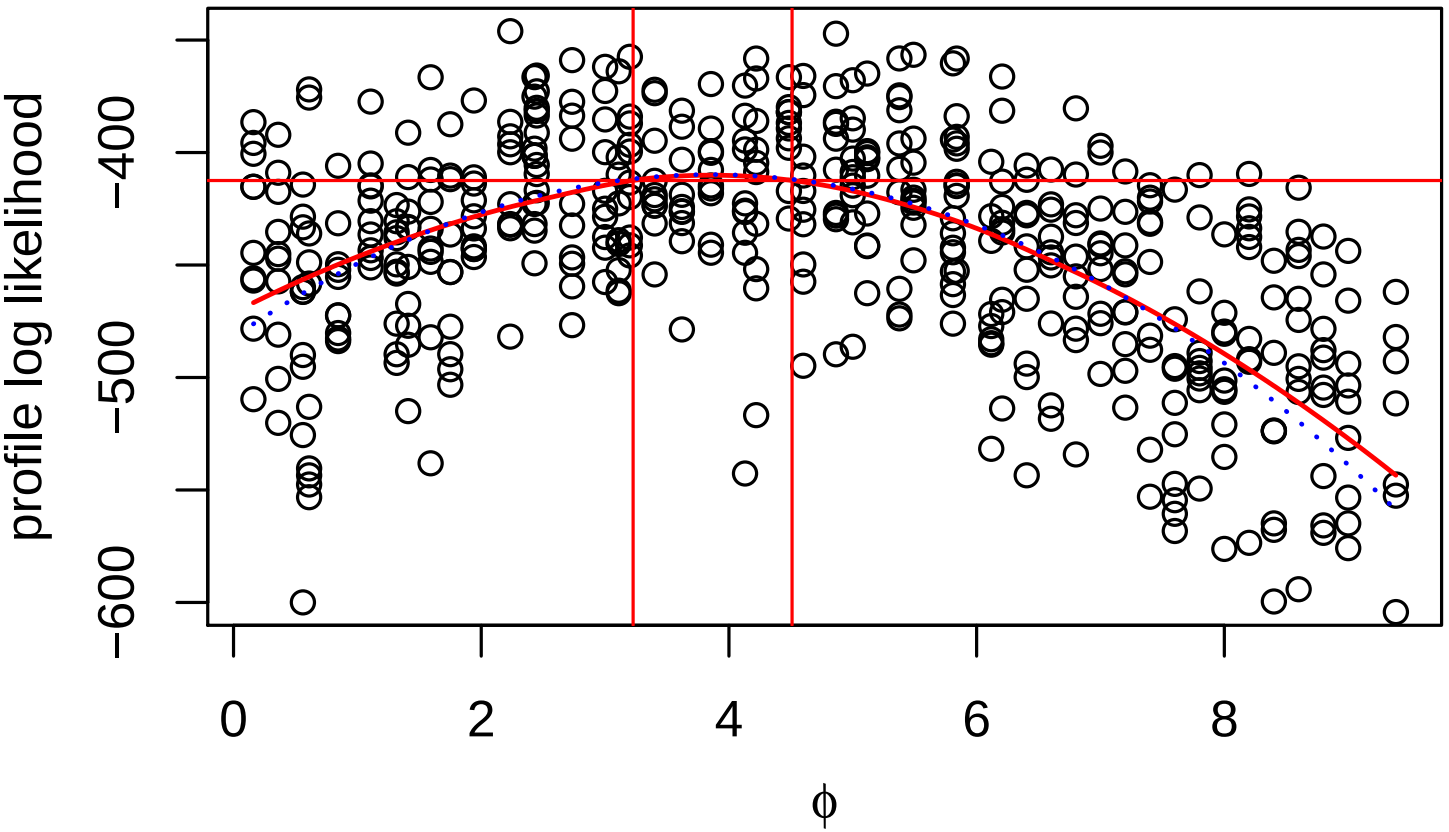



- A **SpatPOMP** is a POMP model with a collection of coupled units.
- Here, units are cities, $u = 1, \ldots, U$
- Coupling arises because $\mu_{SE,u}(t)$ depends on $I_1, \ldots, I_U$
- Data are weekly reported case counts in each city.
- Modeled using coupled over-dispersed Markov chains.

# Innovations for general POMP models

- New Monte Carlo optimization algorithms facilitate likelihood maximization for large partially observed Markov process (POMP) models: **iterated filtering**.
  - Iterated filtering algorithms optimize the likelihood using a sequence of random parameter perturbations, with decreasing magnitude. Sequential Monte Carlo (SMC) provides a tool for numerical solution to this nonlinear filtering problem.
  - Existing variations on expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) do not scale well for these problems.
  - We are doing parametric inference. The main problem using likelihood or Bayesian methods is computational. If existing methods worked computationally, there would be no problem!
- A new perspective on likelihood-based inference via **Monte Carlo profile likelihood**.

# Monte Carlo profile for genetic data on HIV dynamics




- $\phi$ models HIV transmitted by recently infected, diagnosed individuals.
- The profile confidence interval is constructed by a cutoff that is adjusted for the Monte Carlo variability (Ionides et al., 2017).
  - A proper $95\%$ cutoff is 2.35. Without Monte Carlo error, it is 1.92.
  - Each point took approximately 10 core days to compute.
  - Alternative approaches struggle with Monte Carlo likelihood error of order 100 log units.

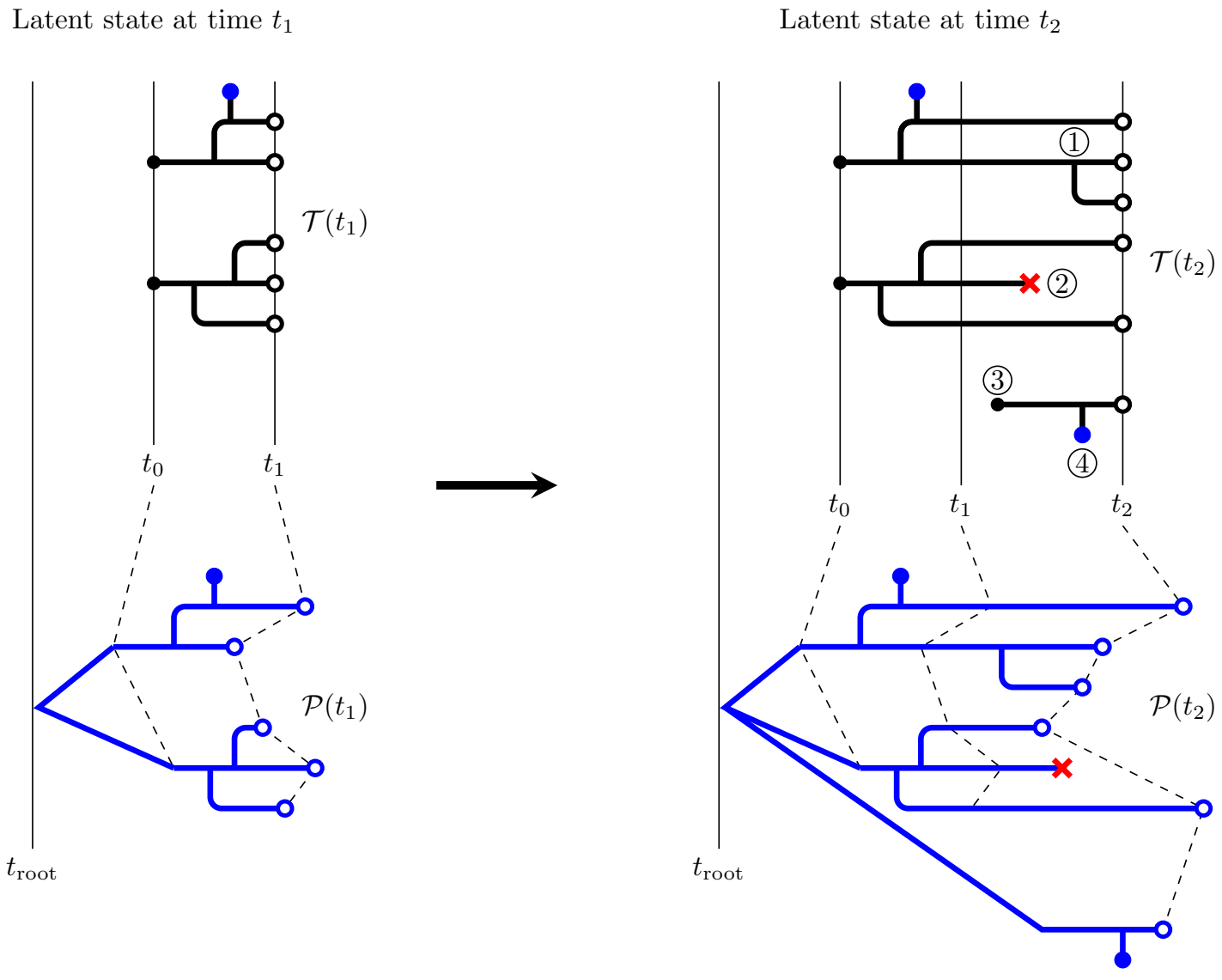
# Previous uses of SMC for phylodynamic inference

- SMC techniques have previously been used for inferring phylogenies (Bouchard-Côté et al., 2012), and for phylodynamic inference conditional on a phylogeny (Rasmussen et al., 2011).
- These approaches avoid the high-dimensional, computationally challenging problem of joint inference.
- Several innovations were necessary to realize computationally feasible SMC on models and datasets of scientific interest.
  - Dimension reduction: constructing the POMP model with genetic sequences only in the measurement model to reduce the dimension of the latent variables.
  - Algorithm parallelization.
  - Hierarchical sampling.
  - Just-in-time construction of state variables.
  - Restriction to a class of physical molecular clocks.
  - Maximization of the likelihood using iterated filtering.

# The latent process for a GenPOMP

- The **latent Markov process**, $\{X(t),\ t \in \mathbb{T}\}$, with $\mathbb{T} = [t_0,\ t_{\text{end}}]$, models the population dynamics and also includes any other processes needed to describe the evolution of the pathogen.
- Suppose we can write $X(t) = \big(\mathcal{T}(t),\ \mathcal{P}(t),\ \mathcal{U}(t)\big)$, where
  - $\mathcal{T}(t)$ is the **transmission forest**,
  - $\mathcal{P}(t)$ is the **pathogen phylogeny** equipped with a relaxed molecular clock,
  - $\mathcal{U}(t)$ represents the **state of the pathogen and host populations**.
- For example, $\mathcal{U}(t)$ may categorize each individual in the host population into classes representing different stages of infection.
- We suppose that $\{\mathcal{U}(t), t \in \mathbb{T}\}$ is itself a Markov process.
- The **plug-and-play property** (Bretó et al., 2009; He et al., 2010) makes our methods applicable to any latent process for which a simulator exists.

# Simulating a GenPOMP from $t_1$ to $t_2$

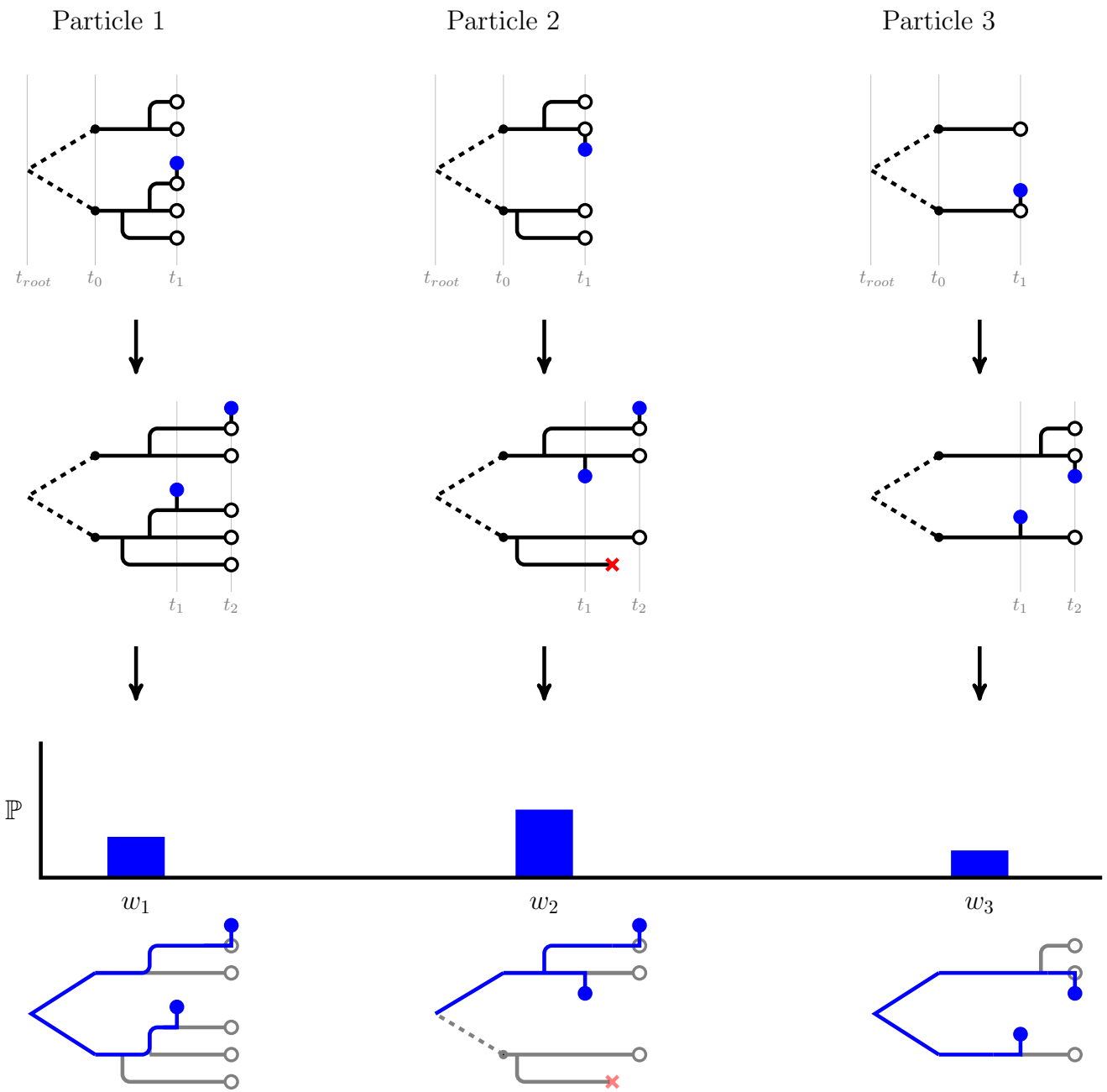



- Black: transmission forest, $\mathcal{T}(t)$. Blue: pathogen phylogeny, $\mathcal{P}(t)$.

# Annotations for the GenPOMP schematic diagram

- The branching pattern of the pathogen phylogeny mirrors that of $\mathcal{T}(t)$ over the interval $[t_0, t_1]$, so pathogen lineages are assumed to branch exactly at transmission events. This simplifying assumption can be changed.
- Randomness in the rate of evolution—a relaxed molecular clock—results in random edge lengths in $\mathcal{P}(t)$.
- At ①, an active node splits in two when a transmission event occurs.
- At ②, an active node becomes a dead node ($\times$) when an infected host emigrates, recovers, or dies.
- At ③, an immigration event gives rise to a new active node with its own root.
- At ④, a sequence node ($\bullet$) is spawned when a sample is taken.

# GenSMC: Sequential Monte Carlo for a GenPOMP




1. **Proposal**. Simulate particles forward from time $t_1$ to time $t_2$. Then select an individual to be sequenced.

2. **Weighting**. Based on the structure of the proposed transmission forest, construct the subtree of the phylogeny that connects the observed sequences. Use this subtree to compute weight of the particle: the conditional probability of the new sequence.

# Dimension reduction: A measurement model integrating the sequence evolution model

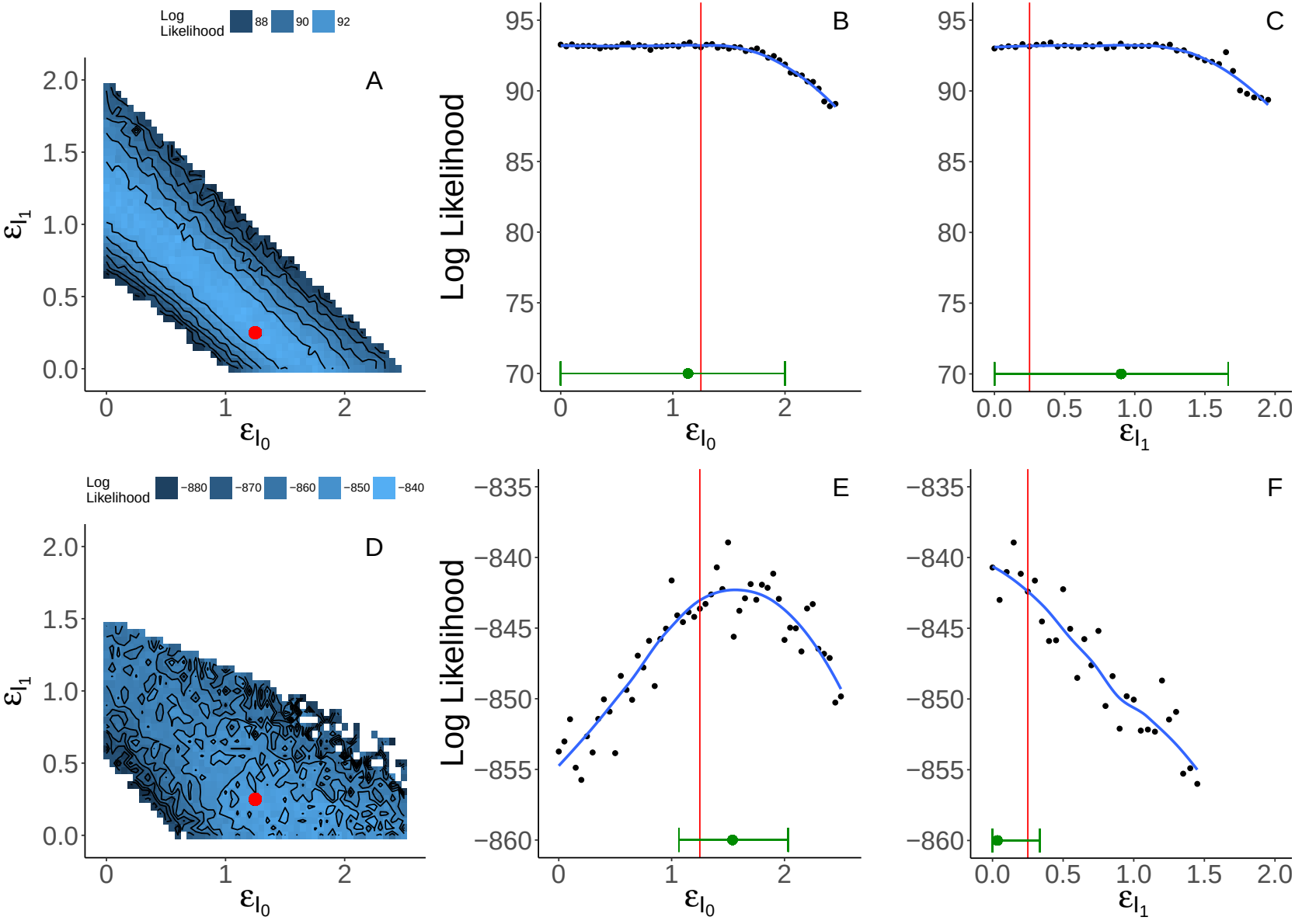
- We put the evolutionary process for the genetic sequences into the measurement model.
- Formally, let a measurement consist of an assignment of a new sequence to an individual in the transmission tree.
- The measurement density involves finding the likelihood of the new sequence given the old sequences and the tree. This likelihood can be computed efficiently by the *peeling* algorithm.
- Particles representing the latent process do not have to include the high-dimensional pathogen genome.

# Restriction to a class of physical molecular clocks

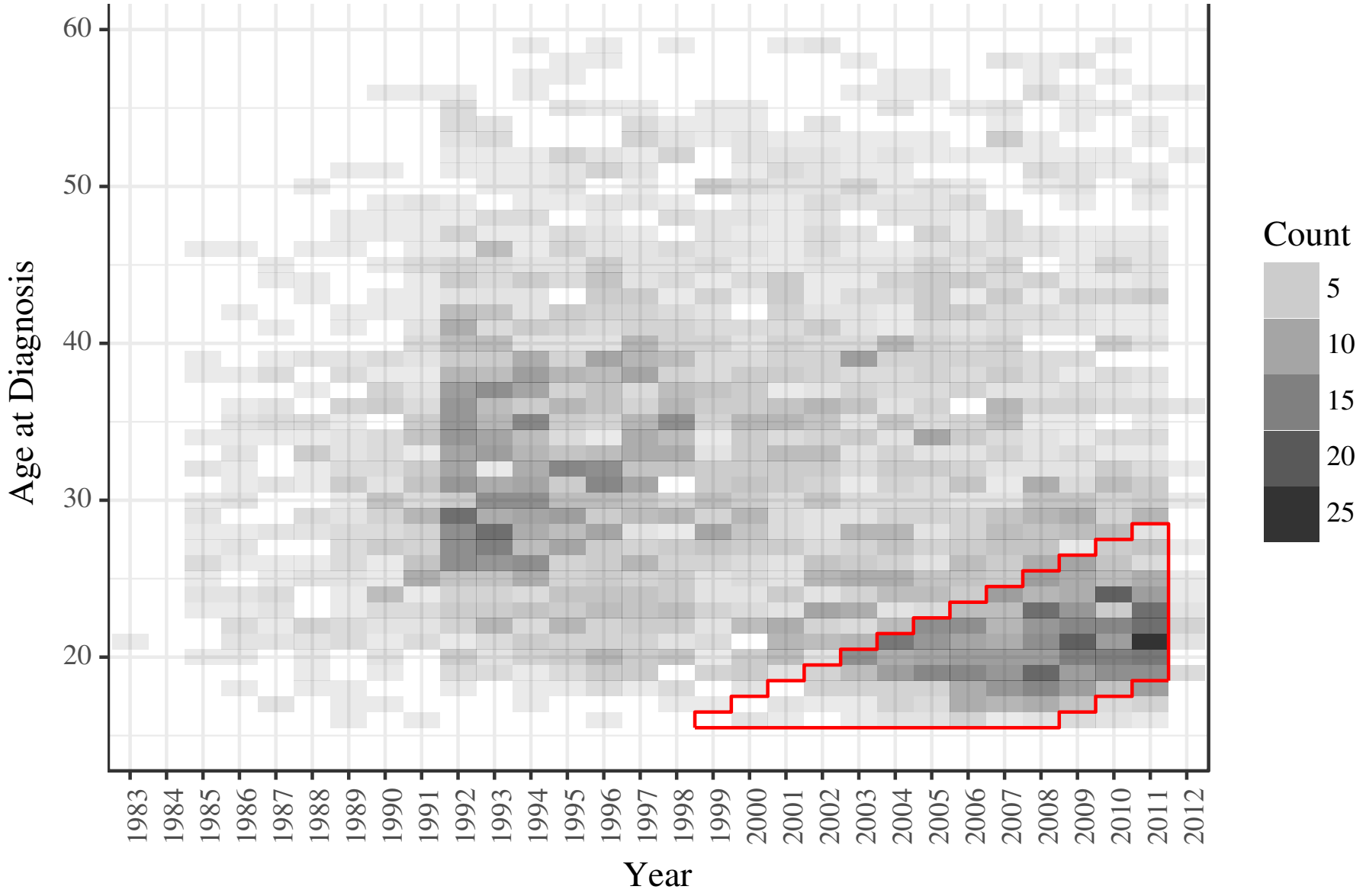
- A strict molecular clock models the rate of evolution as constant through time and across lineages, assuming (i) sequence evolution is Markovian; (ii) no simultaneous mutations.
- These assumptions imply a Poisson-like mean-variance relationship (Bretó and Ionides, 2011).
- Overdispersion (known as a relaxed clock) has been shown to improve the fit of phylogenetic models to observed genetic sequences in many cases (Drummond et al., 2006).
- In our approach, this corresponds to constructing each edge length of $\mathcal{P}(t)$ as a stochastic process on the corresponding edge of $\mathcal{T}(t)$.
- Various forms of such processes have been assumed in the literature, but not all are self-consistent under Markovian assumptions.
- For example, log normal clock perturbations lack an additivity property: adding a node to split a branch must change the evolutionary process along that branch.
- We suppose the relaxed clock is a non-decreasing continuous-valued Lévy process. In practice, we use a Gamma process clock.

# A genPomp simulation study

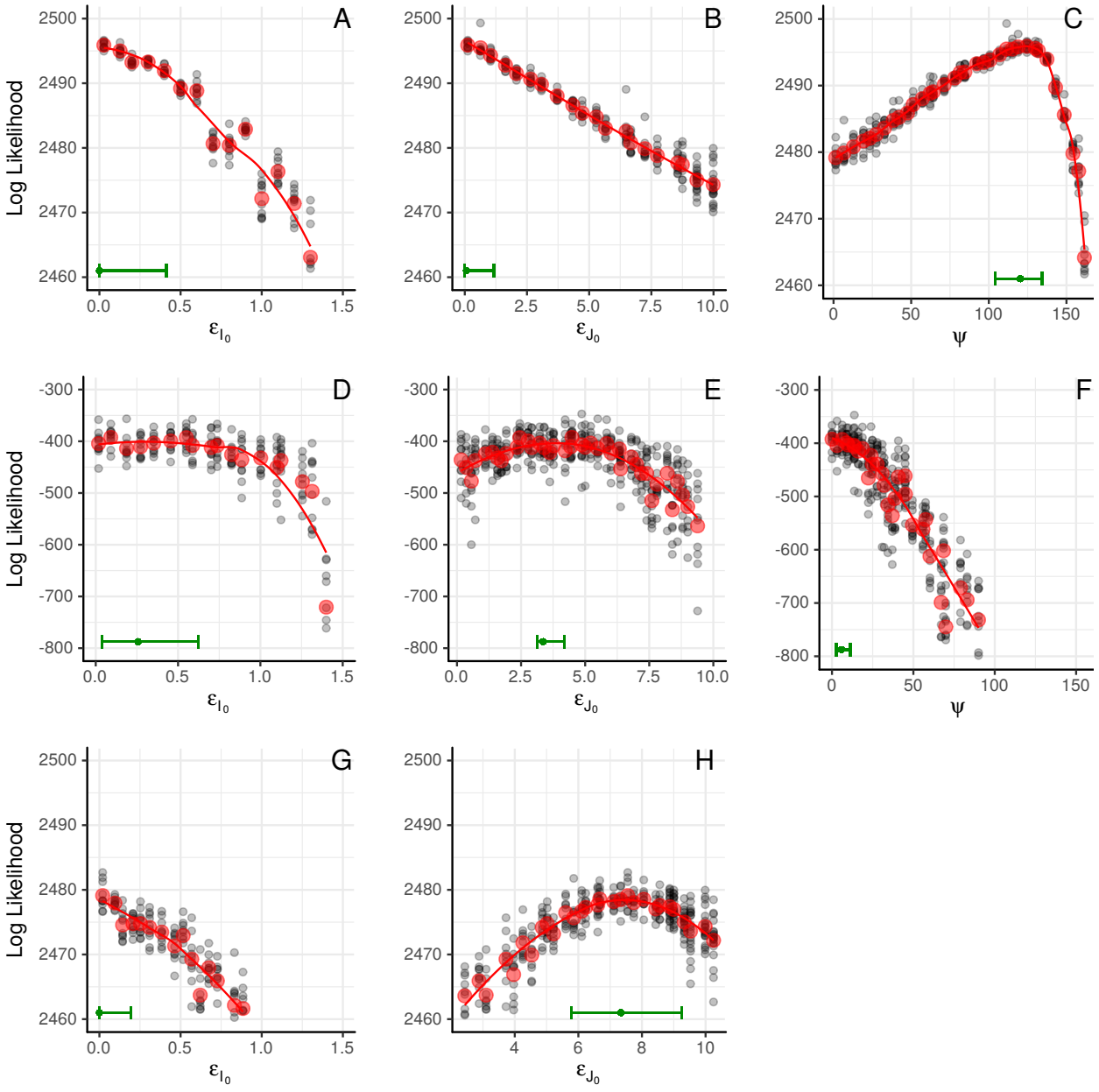
**Top: diagnosis data only. Bottom: including sequence data**



# Detroit data: a young black MSM epidemic



# Detroit data: a young black MSM epidemic




A-C. Diagnosis only.

D-F. Including sequence data.

G-H. Diagnosis only, fixing $\psi = 0$.

# Moving forward from Smith et al (2017, MBE)

- `genPomp` was demonstrated on simulation-based phylodynamic likelihood inference for general dynamic models with order 100 sequences and order 1000 infected individuals.
- Further work is needed to scale to larger systems.
- Having access to the full phylodynamic likelihood facilitates investigations of what (if anything) is lost by 2-step methods and summary statistic methods such as ABC.
- **Preliminary results**: The Volz/Rasmussen likelihood approximation works well if the true phylogeny is known. Phylogenetic uncertainty, especially when the phylogeny is constructed under assumptions different from the latent dynamic system, can lead to substantial bias in estimates and confidence regions.

# Strengths and limitations of the GenPOMP framework

Strengths:

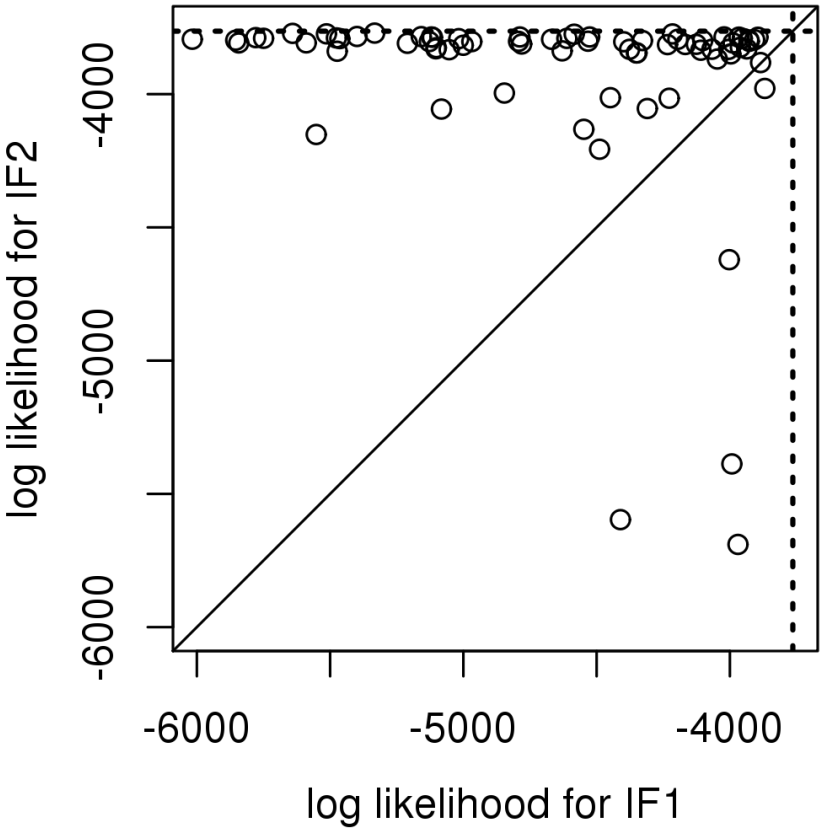
- A large and general model class for population dynamics.
- Statistically efficient inference.
- Can be used to assess loss of information and biases in methods that scale better.

Limitations:

- Computational requirement.
- Some detailed individual-based models may not fit easily into the GenPOMP framework.

# A brief introduction to iterated filtering

- Successful SMC allows likelihood evaluation.
- This likelihood evaluation is both costly and noisy for non-small problems, so requires specialized algorithms to enable effective inference.
- The IF1 iterated filtering algorithm of (Ionides et al., 2006) averaged filtered parameters in a perturbed model, repeating with successively smaller perturbations.
- The IF2 algorithm of (Ionides et al., 2015) simply feeds perturbed particles at the end of one filtering iteration back as starting values for the next iteration, with decreasing perturbations.
- IF1 made possible some previously inaccessible inferences, but IF2 is much better!




**Comparison of IF1 and IF2 on the cholera model.**

**Algorithmic tuning parameters for both IF1 and IF2 were set at the values chosen by King et al (2008) for IF1.**

- Log likelihoods of the parameter vector output by IF1 and IF2, both started at a uniform draw from a large 23-dimensional hyper-rectangle.
- Dotted lines show the maximum log likelihood.

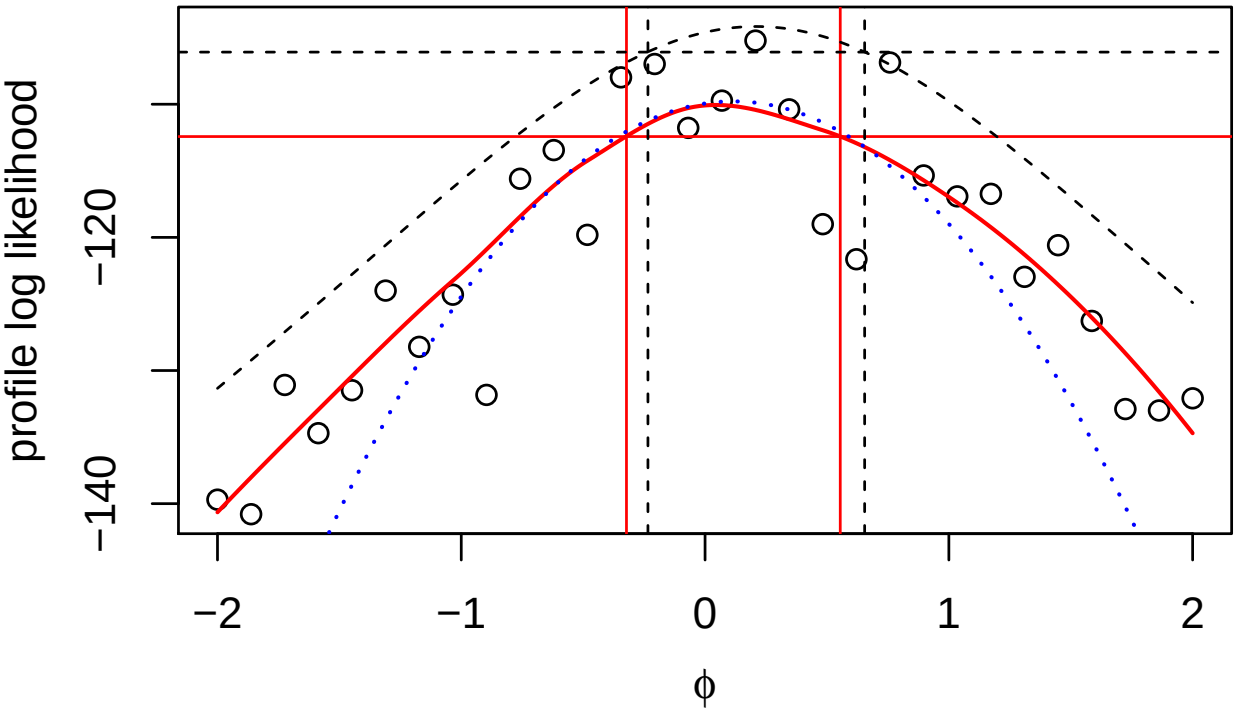
# Monte Carlo adjusted profile (MCAP) confidence intervals

- The usual cutoff $\delta = 1.92$ for a 95% profile confidence interval is based on an asymptotic quadratic log likelihood (Wilks' $\chi^2$ theorem).
- Profile intervals are robust to reparameterization.
- A Wilks limit also applies to give a cutoff for a smoothed Monte Carlo profile based on a quadratic approximation (Ionides et al., 2017),

$$\delta_{MCAP} = z_\alpha^2 \left( a \times \mathrm{SE}_{\mathrm{mc}}^2 + \frac{1}{2} \right),$$

  where $z_\alpha$ is the $1 - \alpha/2$ normal quantile, $a$ is the quadratic coefficient of a quadratic regression near the profile maximum, $\mathrm{SE}_{\mathrm{mc}}$ is the Monte Carlo error of the maximum of this quadratic.
- if $\mathrm{SE}_{\mathrm{mc}} = 0$, the cutoff for $\alpha = 0.05$ reduces to $\delta_{MCAP} = 1.96^2/2 = 1.92$.
- We apply this cutoff after estimating the profile via a locally weighted quadratic smoother.
- We call this procedure a **Monte Carlo adjusted profile (MCAP)**.

# A toy: MCAP for a log normal model




Points show Monte Carlo profile evaluations. Black dashed lines: exact profile and 95% confidence interval. Solid red lines: MCAP confidence interval. Dotted blue line: quadratic approximation.

| | Exact profile | MCAP profile | Bootstrap | Quadratic |
|---|---|---|---|---|
| Coverage % | 94.3 | 93.4 | 93.3 | 93.3 |
| Mean width | 0.78 | 0.88 | 0.94 | 0.92 |

# Software

**pomp**. An R package developed and maintained for 12yr (King et al., 2016). Various tutorials, courses and open-source examples exist.

`https://kingaa.github.io/sbied/`

`https://ionides.github.io/531w18/`

**panelPomp**. An R package extending `pomp` for PanelPOMP models (Bretó et al., 2019)

**genPomp**. A C++ program written for (Smith et al., 2017)

**spatPomp**. An R package extending `pomp` for SpatPOMP models. A preliminary version will be released soon.

# Collaborators

Contributors on the methodological developments:

- Aaron King
- Alex Smith
- Carles Breto
- Joonha Park
- Dao Nguyen
- Kidus Asfaw

Collaborators on the scientific work:

- Ethan Romero-Severson
- Mercedes Pascual
- Jim Koopman
- Erik Volz

# References I


Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174 – 188.

Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Systematic Biology*, 61(4):579–593.

Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3:319–348.

Bretó, C. and Ionides, E. L. (2011). Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, 121:2571–2591.

Bretó, C., Ionides, E. L., and King, A. A. (2019). Panel data analysis via mechanistic models. *To appear in JASA. Available at Arxiv:1801.05695*.

# References II


Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005*, pages 64–69. IEEE.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface*, 7:271–283.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA*, 103:18438–18443.

Ionides, E. L., Breto, C., Park, J., Smith, R. A., and King, A. A. (2017). Monte Carlo profile confidence intervals for dynamic systems. *Journal of the Royal Society Interface*.

# References III


Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA*, 112:719–724.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454:877–880.

King, A. A., Nguyen, D., and Ionides, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, 69:1–43.

Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, 7(8):e1002136.

Romero-Severson, E., Volz, E., Koopman, J., Leitner, T., and Ionides, E. (2015). Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *American Journal of Epidemiology*, 182:255–262.

# References IV

Smith, R. A., Ionides, E. L., and King, A. A. (2017). Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Molecular Biology and Evolution*, 34:2065–2084.