Deep learning for genetic epidemiology

Alexander E. Zarebski
University of Melbourne

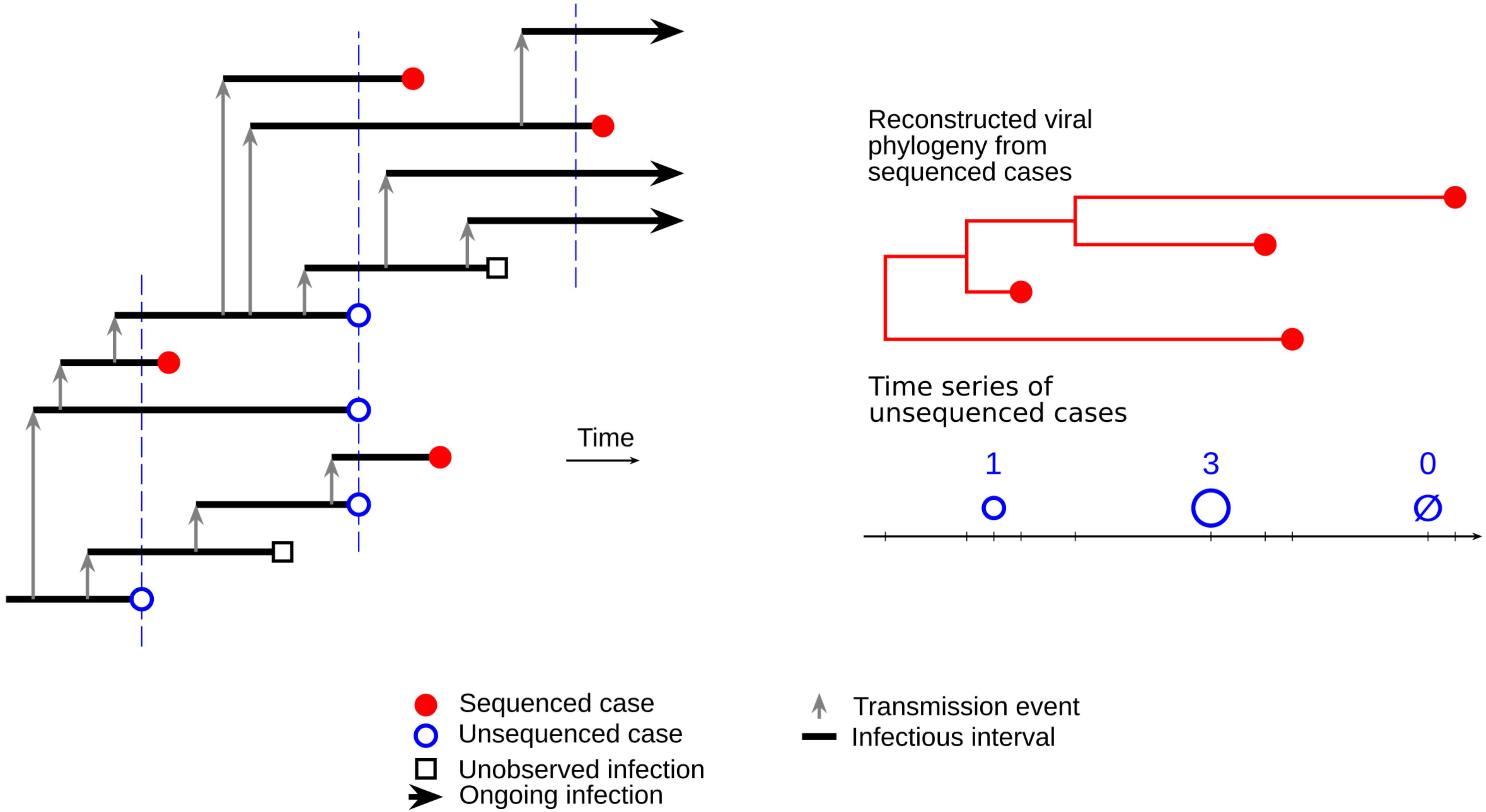
ANZIAM 2024

- Infectious disease epidemiology helps to inform epidemic response: lockdowns, vaccination strategy, etc.
- Knowing the *prevalence of infection* and *reproduction number* is useful for selecting an appropriate response
- Genetic epidemiology uses genomic data and to estimate these quantities

Background: phylodynamics

- *Phylodynamics* uses genomic data to study how a population size has changed over time
- Helps us understand the efficacy of interventions
- Helps us answer questions such as
  - Where has the pathogen come from?
  - What proportion of cases we are observing?
  - Is this variant more transmissible?
  - ...

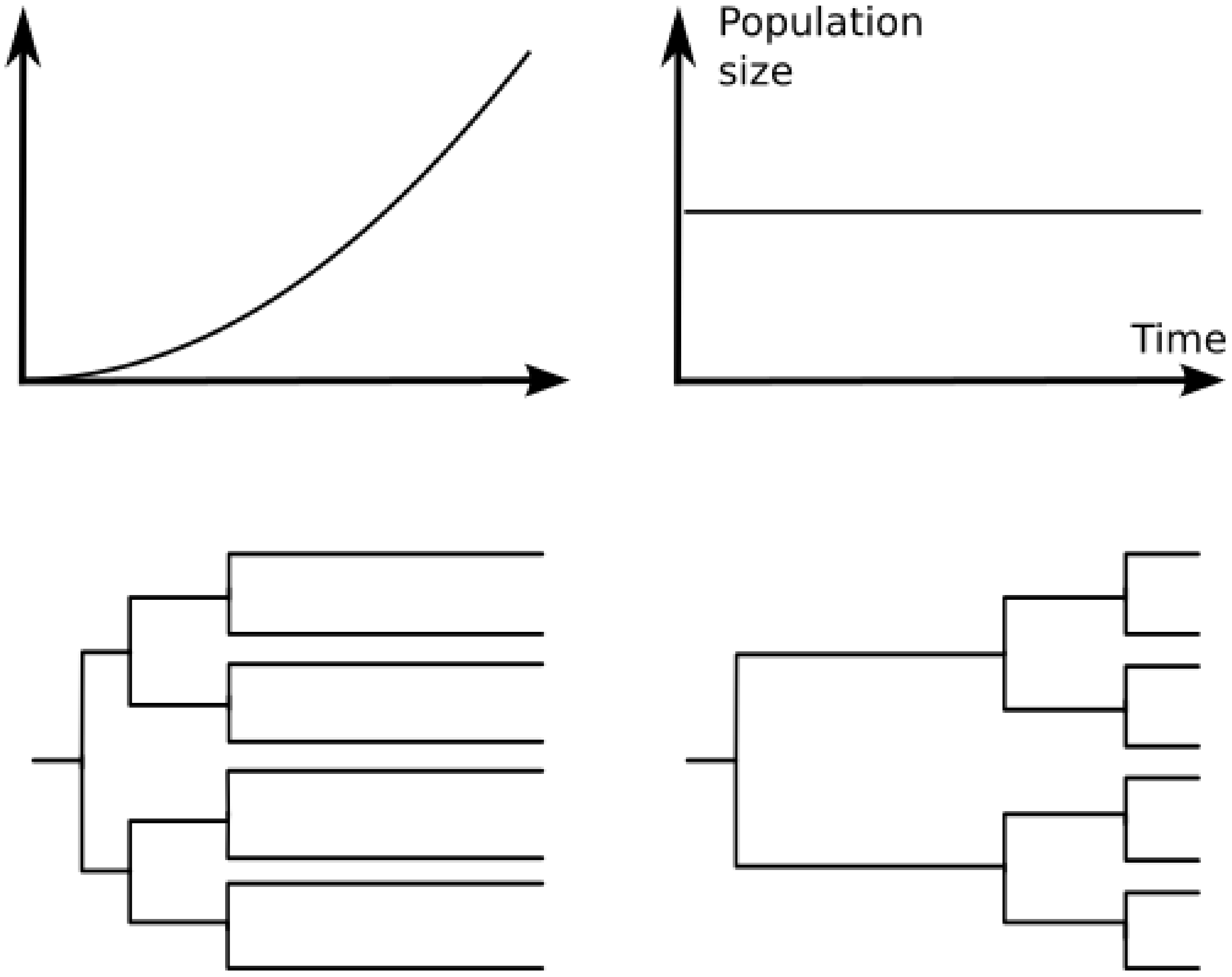
What is a phylogenetic tree?



Reconstructed viral phylogeny from sequenced cases

Time series of unsequenced cases

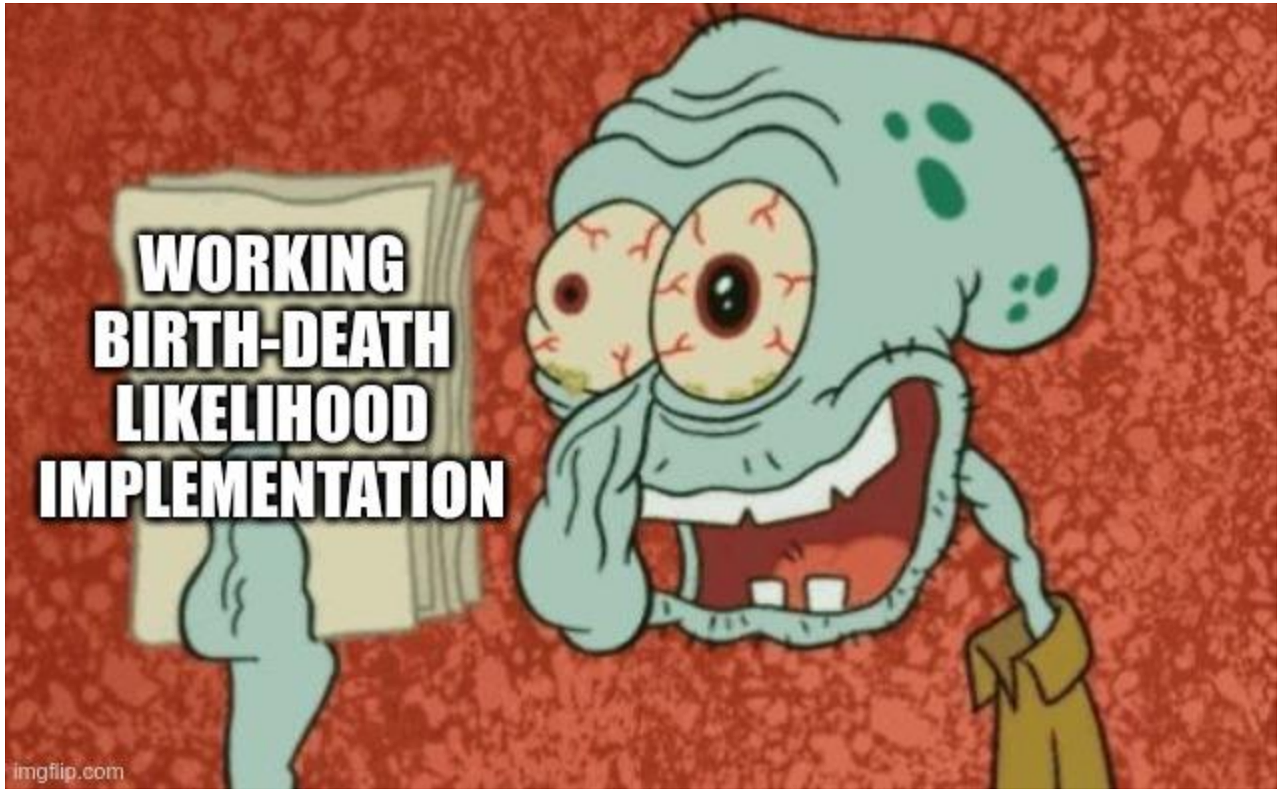
What does this tell me?



Motivation for a new approach

- We want the *prevalence of infection* and the *reproduction number*
- Phylodynamics can be slow; we need answers in a hurry

- *But worst of all…*


WORKING
BIRTH-DEATH
LIKELIHOOD
IMPLEMENTATION
imgflip.com

What about neural networks?

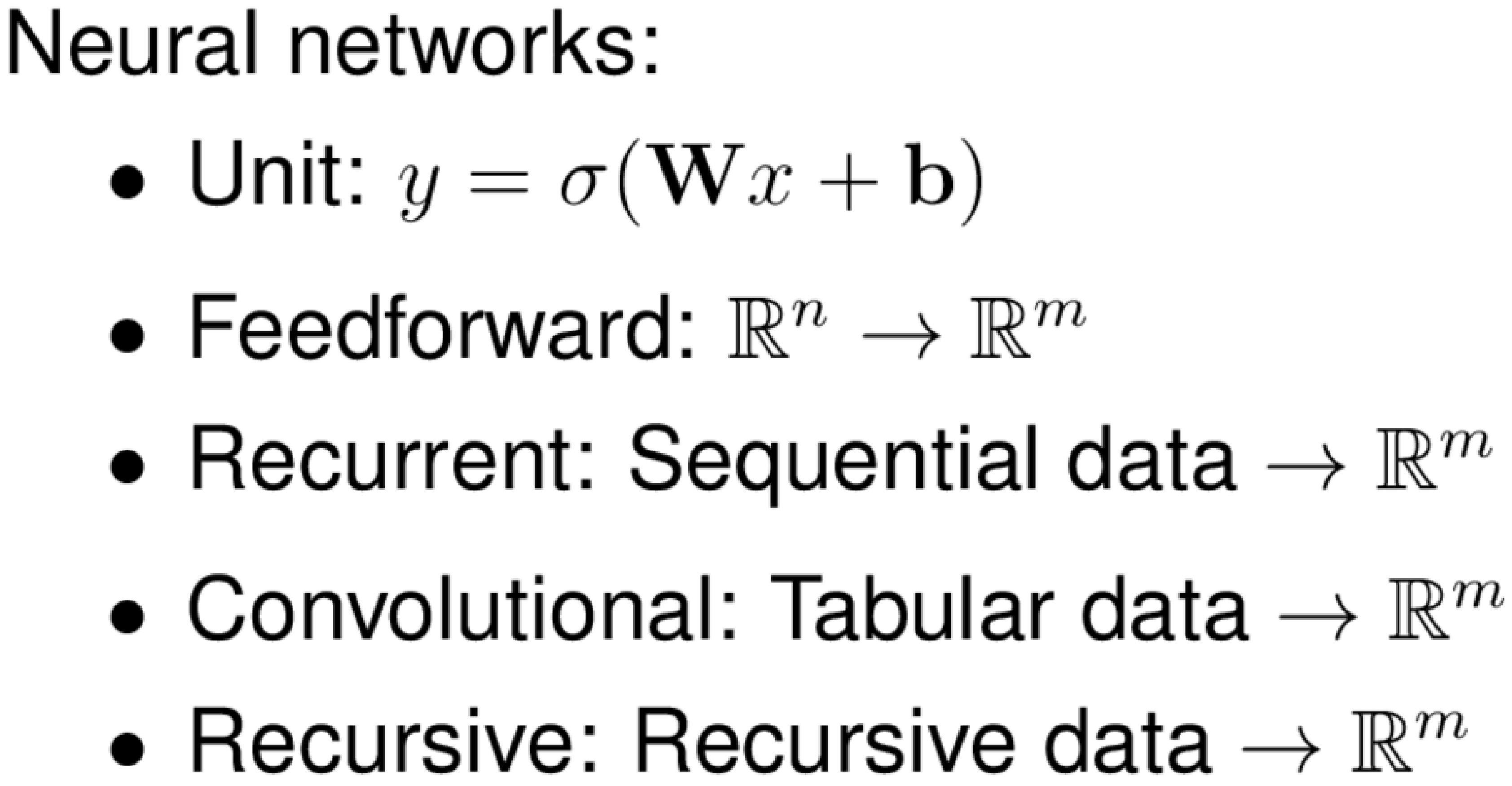
“statistics” : $\mathcal{D} \to \Theta$

“simulation” : $\Theta \to \mathcal{D}$

Simulate $\theta_i \in \Theta$, then $d_i \in \mathcal{D}$ to generate $\{(d_i, \theta_i)\}_i$

Train a neural network to predict θ_i when given d_i

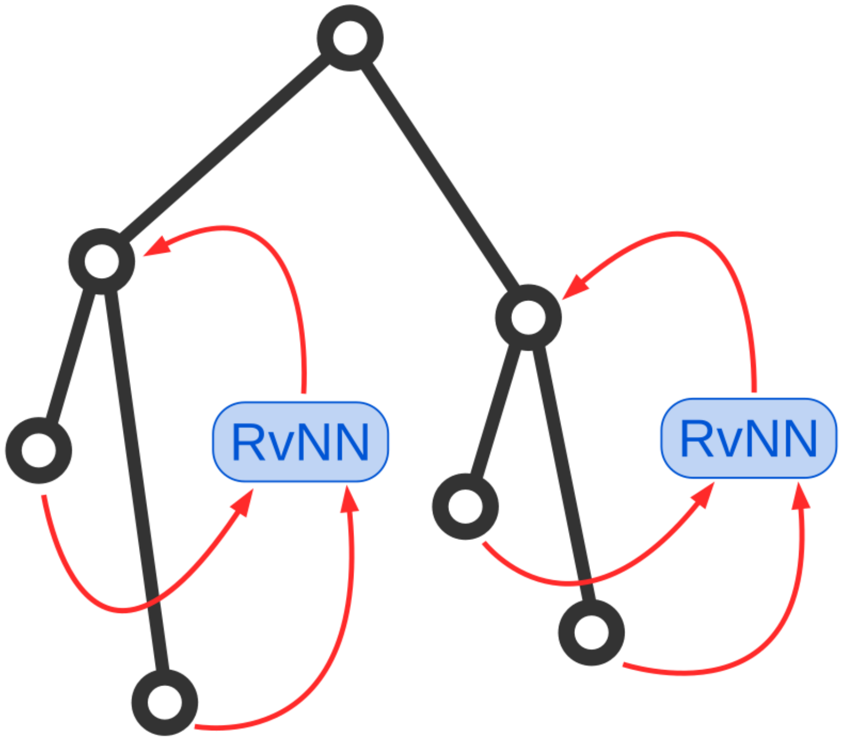
Terminology

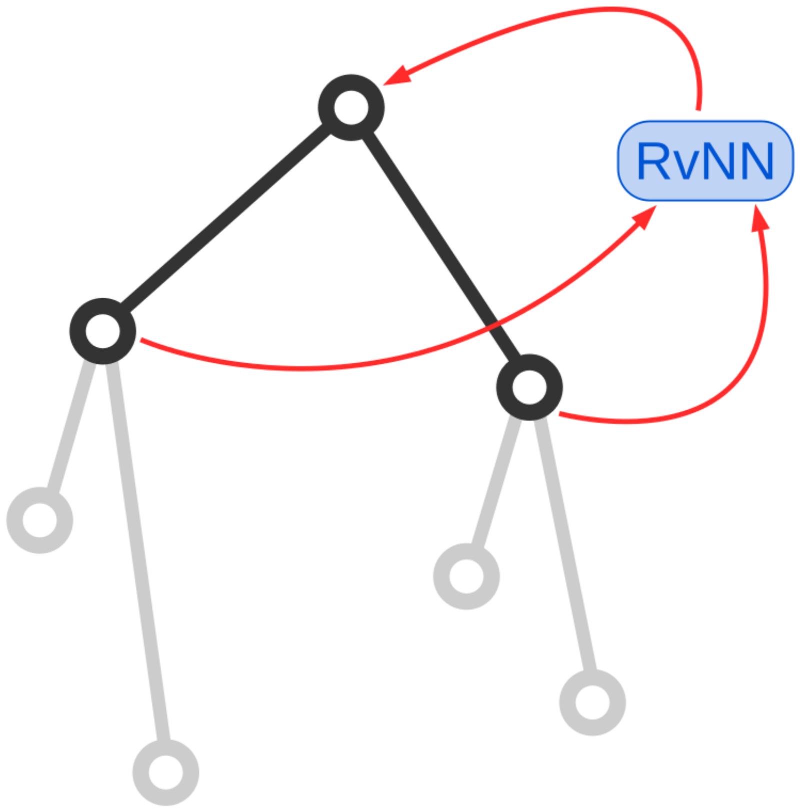


Neural networks:

- Unit: $y = \sigma(\mathbf{W}x + \mathbf{b})$
- Feedforward: $\mathbb{R}^n \to \mathbb{R}^m$
- Recurrent: Sequential data $\to \mathbb{R}^m$
- Convolutional: Tabular data $\to \mathbb{R}^m$
- Recursive: Recursive data $\to \mathbb{R}^m$

What is a *recursive* neural network?





What is a *recursive* neural network?

$\mathcal{T}$ is the set of trees.

$L : \mathbb{R}^{2n+1} \to \mathbb{R}^n$ is a linear map.

σ is an activation function.

$b : \mathcal{T} \to \mathbb{R}_{\geq 0}$ is the branch length

The map $f : \mathcal{T} \to \mathbb{R}^n$ is given by

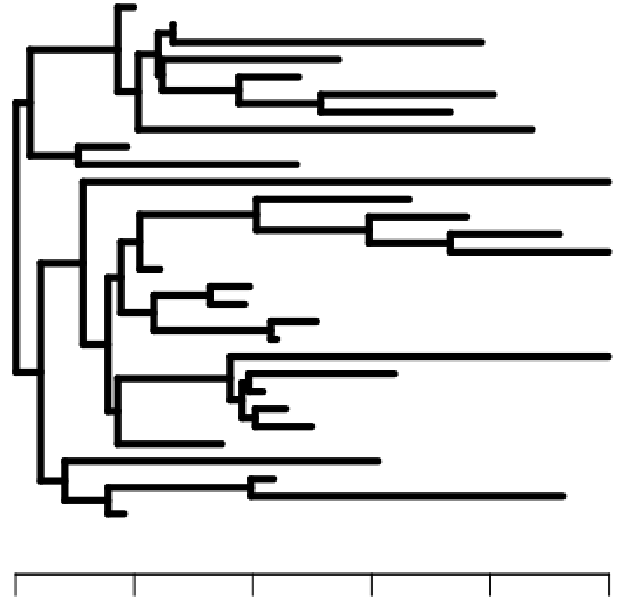
$$f(t) = \begin{cases} \sigma((b(t), \mathbf{0})) & \text{for leafs } t \\ \\ \sigma(L((\sigma(b(t)), f(\text{left}(t)), f(\text{right}(t))))) \end{cases}$$

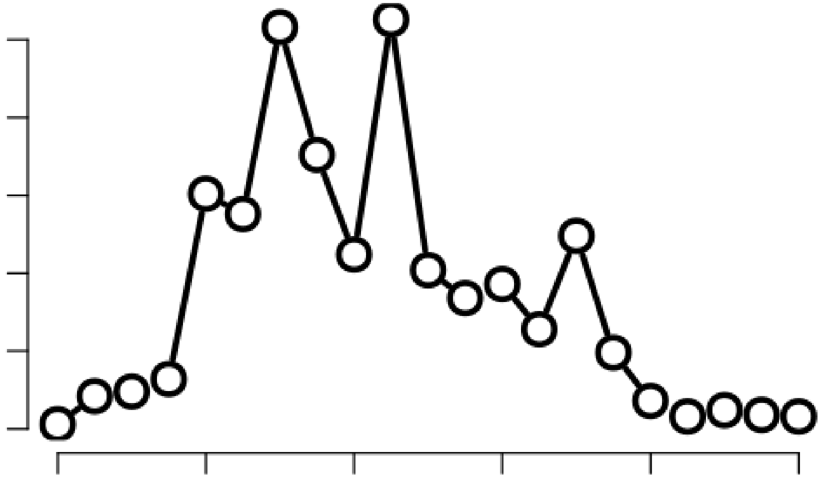
Spoilers!

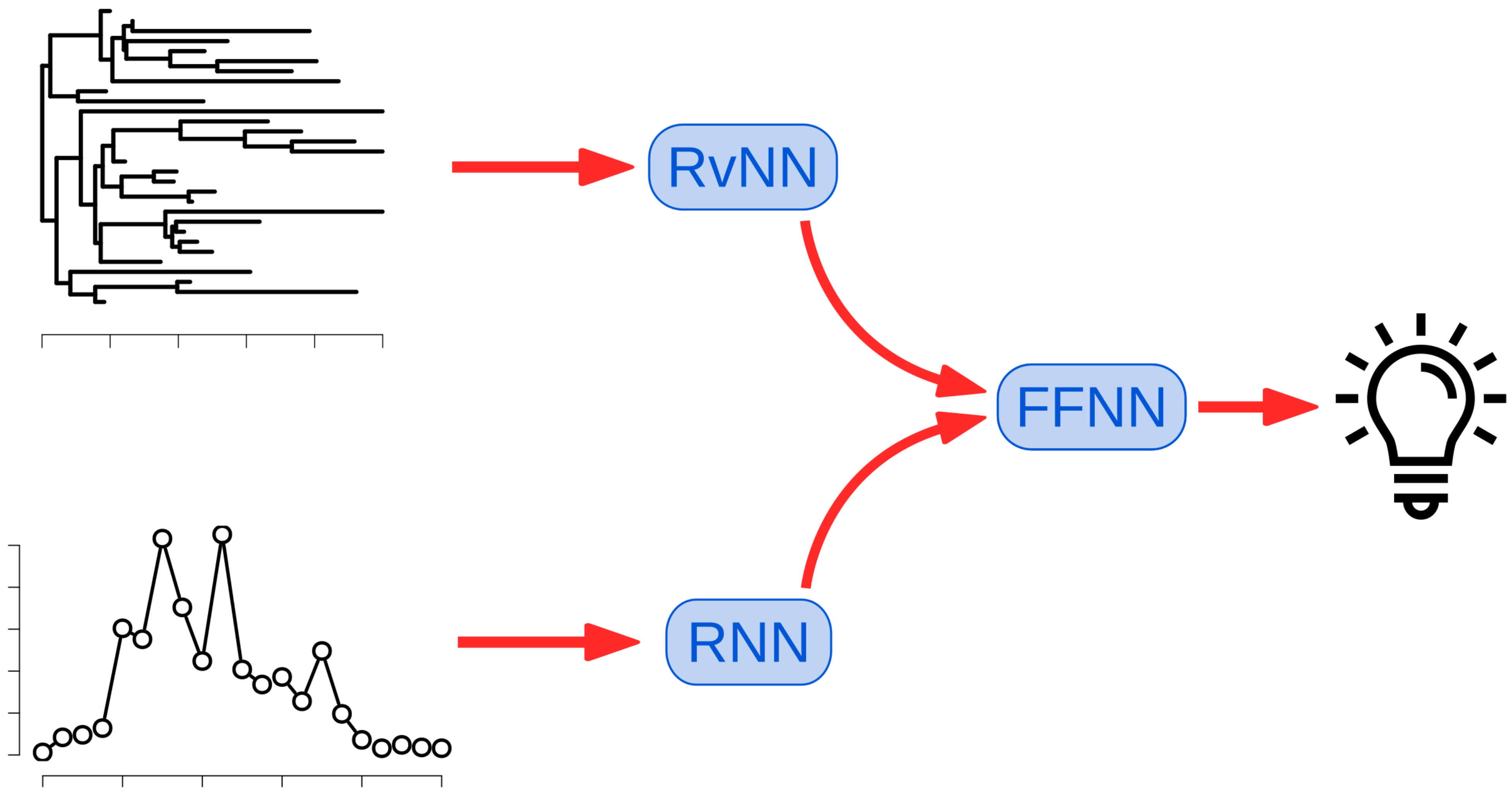
- You can train a neural network to estimate the prevalence of infection and the basic reproduction number
- It will be able to use both genomic data and time series data
- It runs **fast** (once you have trained it)
- You can train it on your laptop

Epidemic model

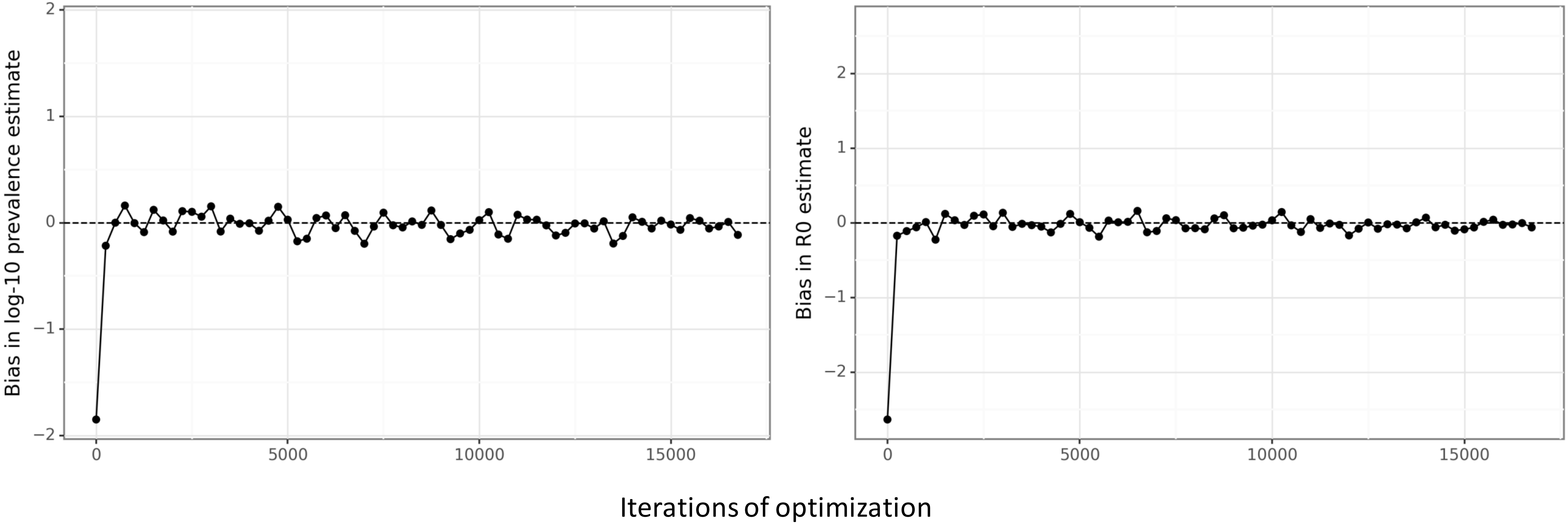
- Birth-death-sampling process
- Infect new cases at rate λ
- Cease to be infectious when:
  - Test cases at rate ω
  - Sequence cases at rate ψ
  - Cases recover at rate μ
- Observe the process
  - Time series from aggregated tests
  - Reconstruct phylogeny from sequences



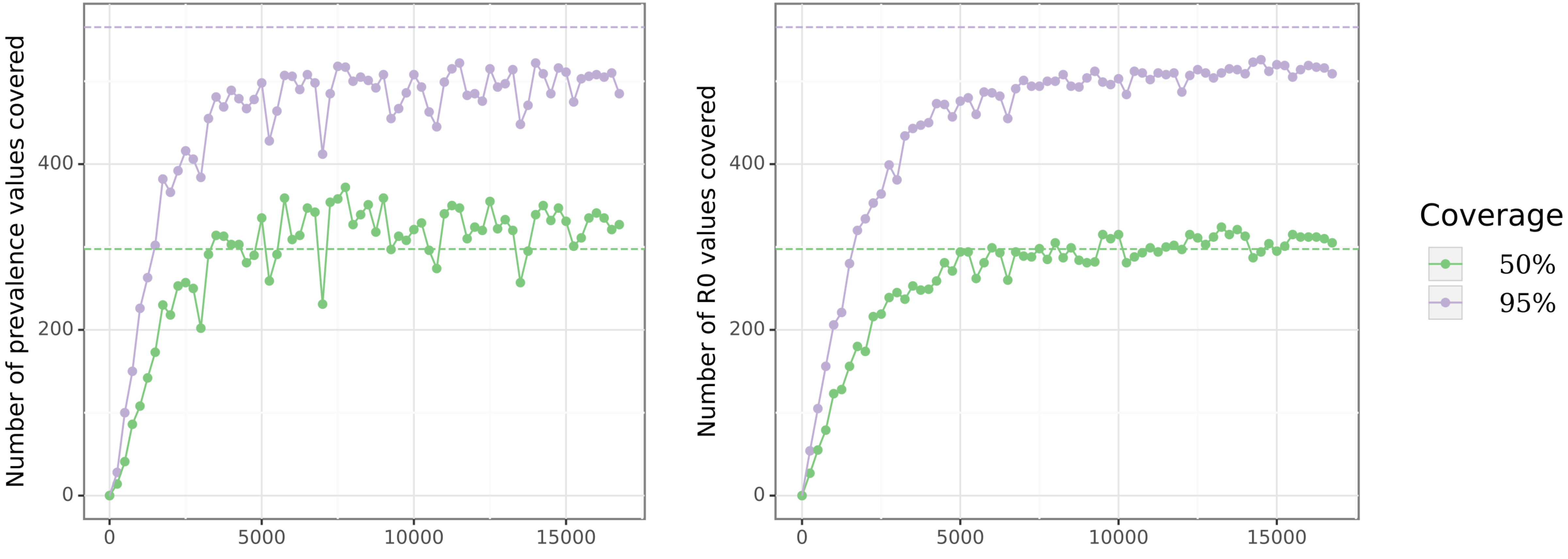



RvNN
FFNN
RNN

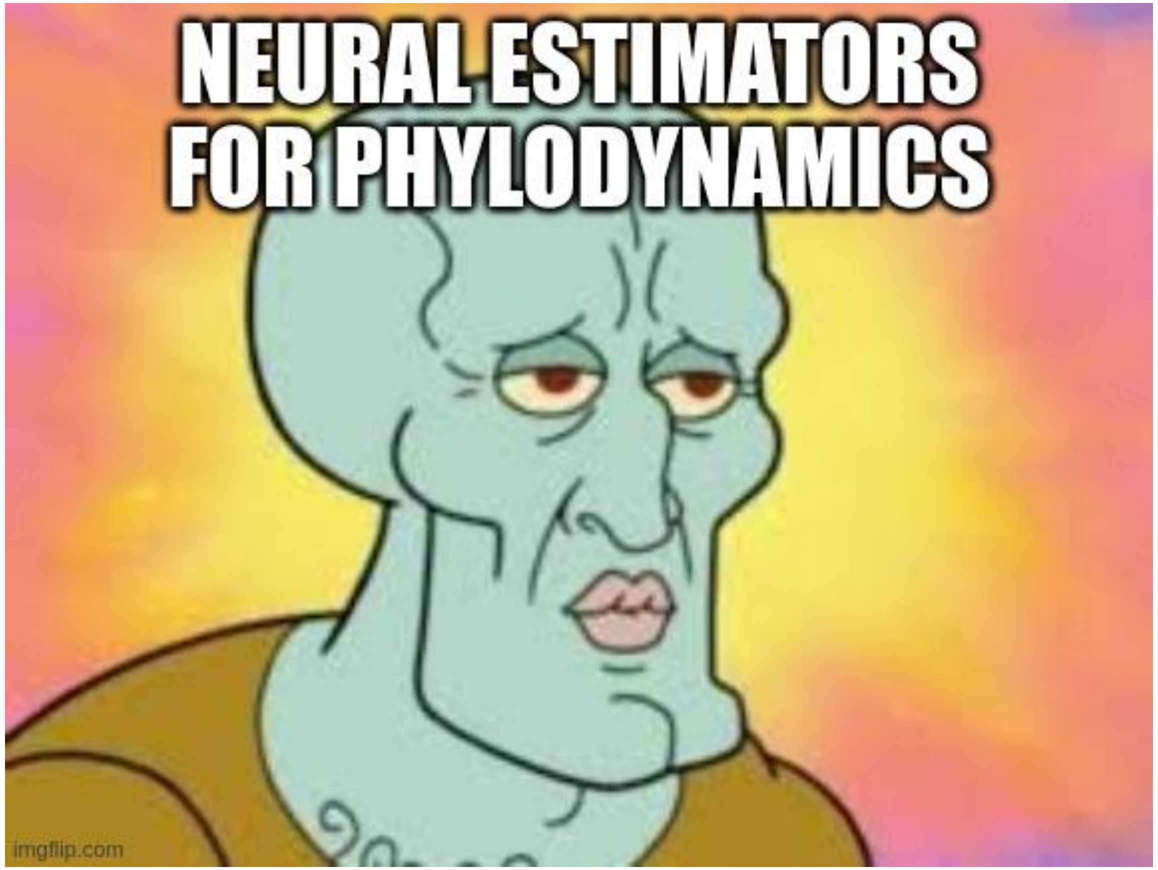
Point estimate bias during training



Uncertainty calibration during training



Iterations of optimization


NEURAL ESTIMATORS
FOR PHYLODYNAMICS
imgflip.com

Take home message(s)

- Finally! Simulation based inference
  - *That sounds familiar; what about the bootstrap, particle filter, and ABC?*
- Phylodynamics can go beyond simple models!
  - *That sounds familiar; what about pMCMC and tree uncertainty?*
- It's an exciting time to be working on phylodynamics

Thank you

- Jennifer Flegg
- Melbourne Mathematical Biology (MMB) Group