Next Generation Bayesian Methods for Complex Systems: Theory and Implementation

Yousef H. El-Laham

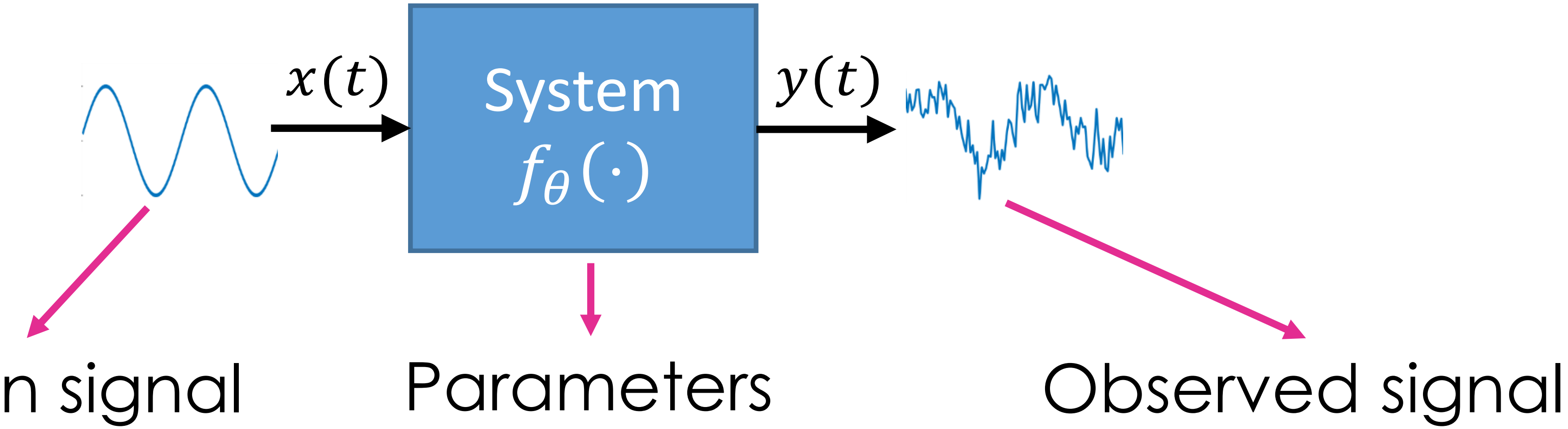
25 July 2019

Outline

1. Motivation
2. Research Overview
3. Computational Aspects
4. Interdisciplinary Work
5. Timeline
6. Conclusions

Complex Systems

- Motivation: Understanding how complex systems work.

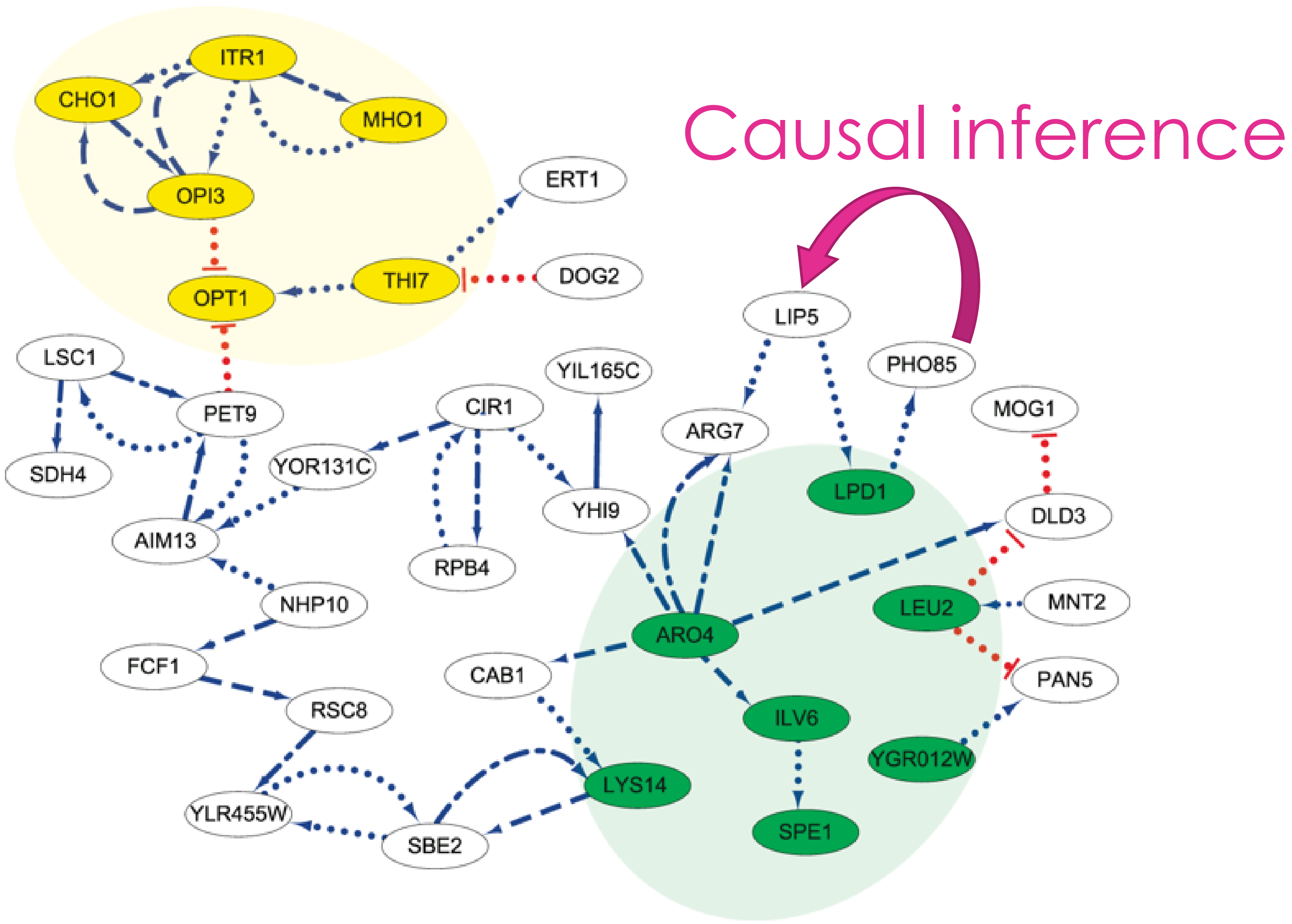


- We call a system "complex" if the number of unknown parameters that represent the system is large.

Financial Market

S&P 500 index from dates 05/2018-05/2019

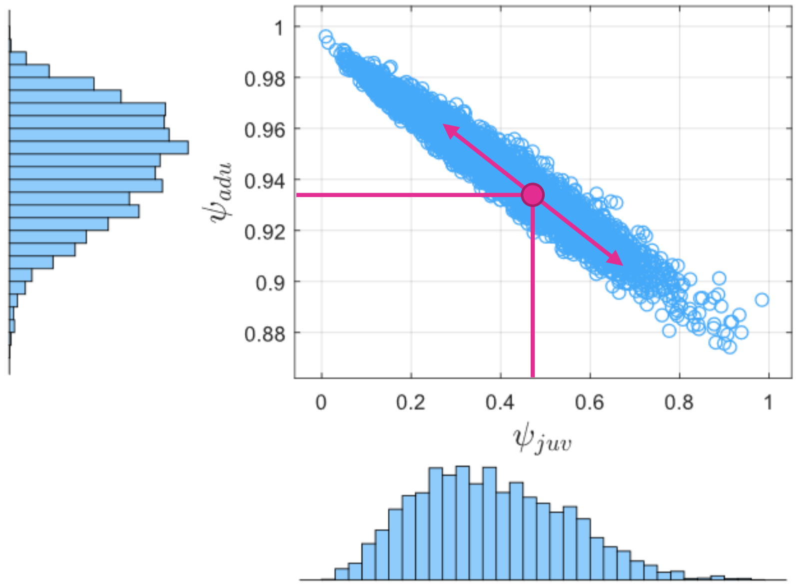
Gene Regulatory Networks



Network describing the interaction of genes.

Population Dynamics

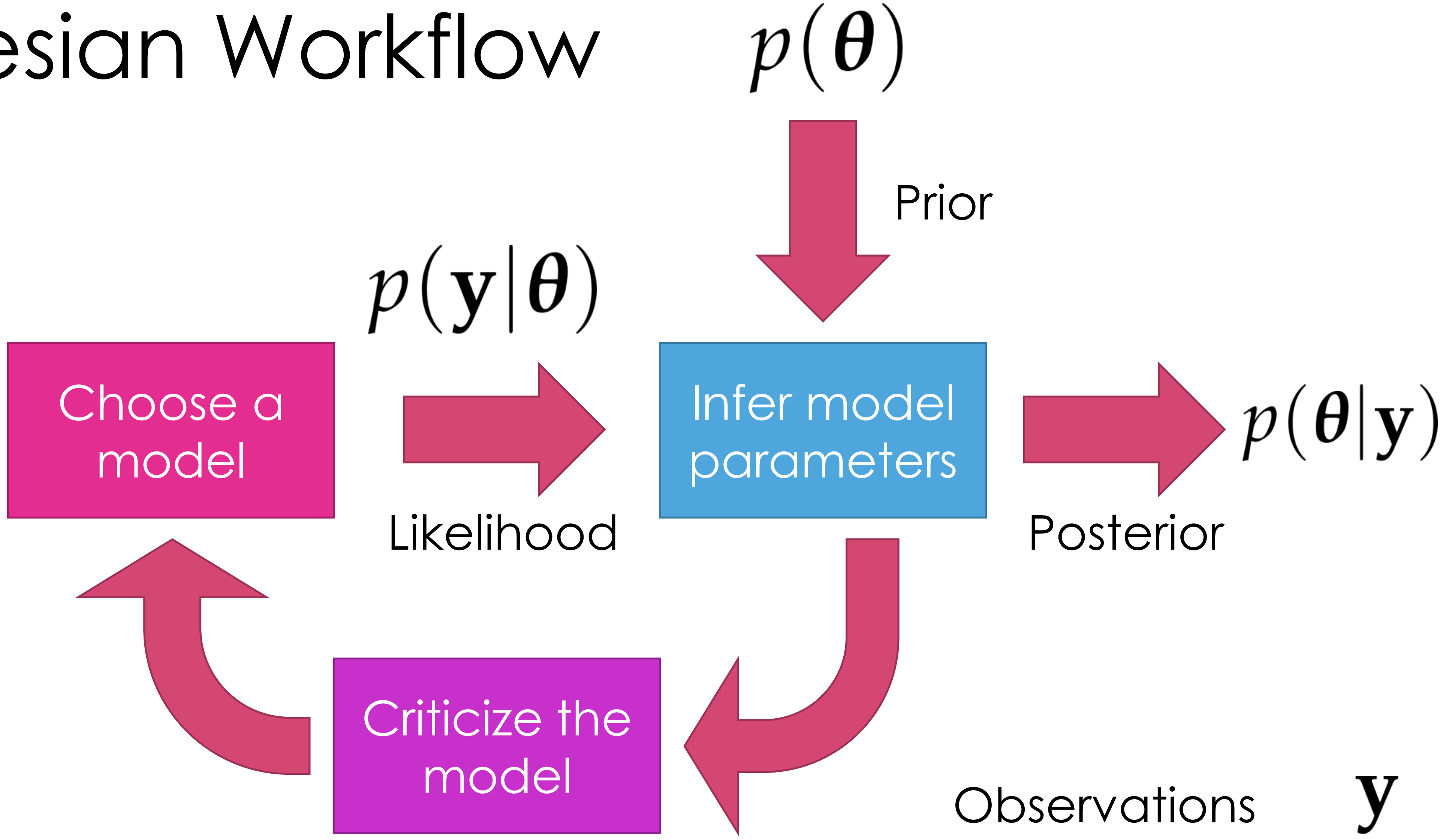
Goal: Parameter estimation!



Cape Bird North per ... abundance counts

$$Count[t] = f(x_t, \psi_{juv}, \psi_{adu}) + error$$

Bayesian Workflow



Observations $\mathbf{y}$

Model parameters $\boldsymbol{\theta}$

Bayes' Theorem

- Our goal is to determine the **posterior distribution** of the unknown model parameters $\boldsymbol{\theta}$.
- We quantify the posterior via Bayes theorem:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto \underbrace{p(\mathbf{y}|\boldsymbol{\theta})}_{\text{likelihood}}\underbrace{p(\boldsymbol{\theta})}_{\text{prior}}$$

The <u>likelihood</u> tells us how likely a given parameter "generated" the data

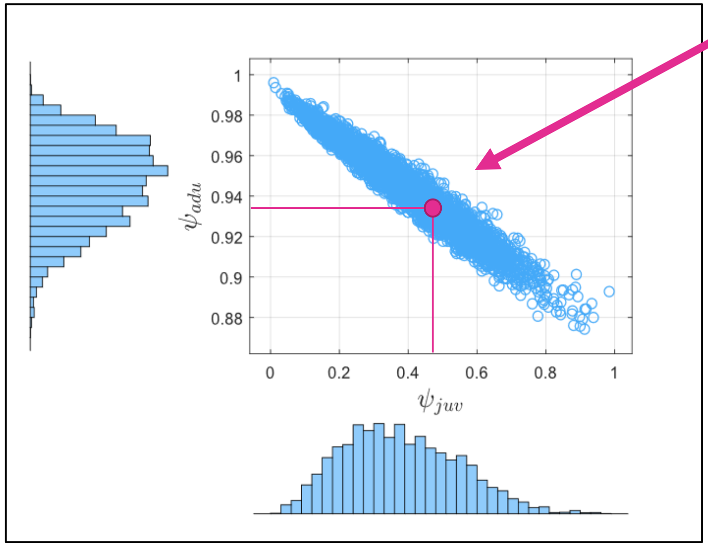
The <u>prior</u> encodes domain expert knowledge about the parameter

What Do We Want?

- We can usually evaluate something proportional to the posterior. But is that enough?
- What kind of quantities are we interested in?

$$\text{Avg}\left[\boldsymbol{\theta}|\mathbf{y}_{1:T}\right] = \int \boldsymbol{\theta} \times p(\boldsymbol{\theta}|\mathbf{y}_{1:t})d\boldsymbol{\theta}$$

In most cases, computing this integral in a closed-form expression is impossible!

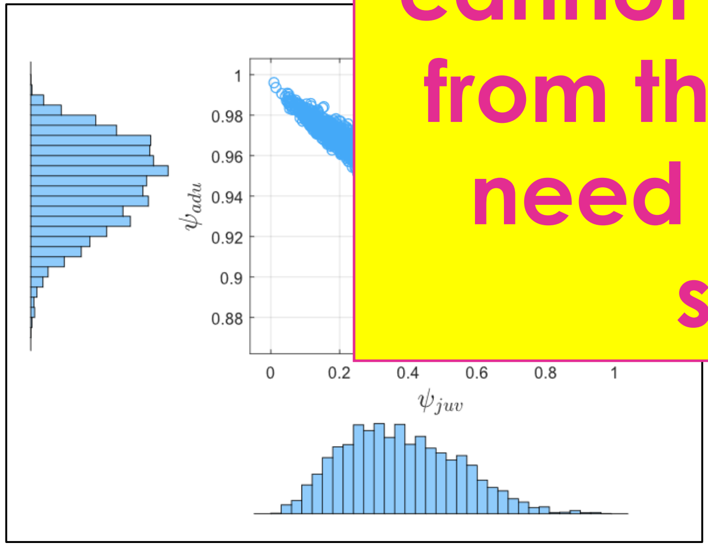


We need to use computers to numerically <u>compute</u> these complex integrals!

Monte Carlo Methods

- The theory of Monte Carlo methods tells us how we can use **random** … se crazy integrals.
- First, we draw … distribution…

Unfortunately, in most useful models, we cannot sample directly from the posterior! We need an alternative solution…



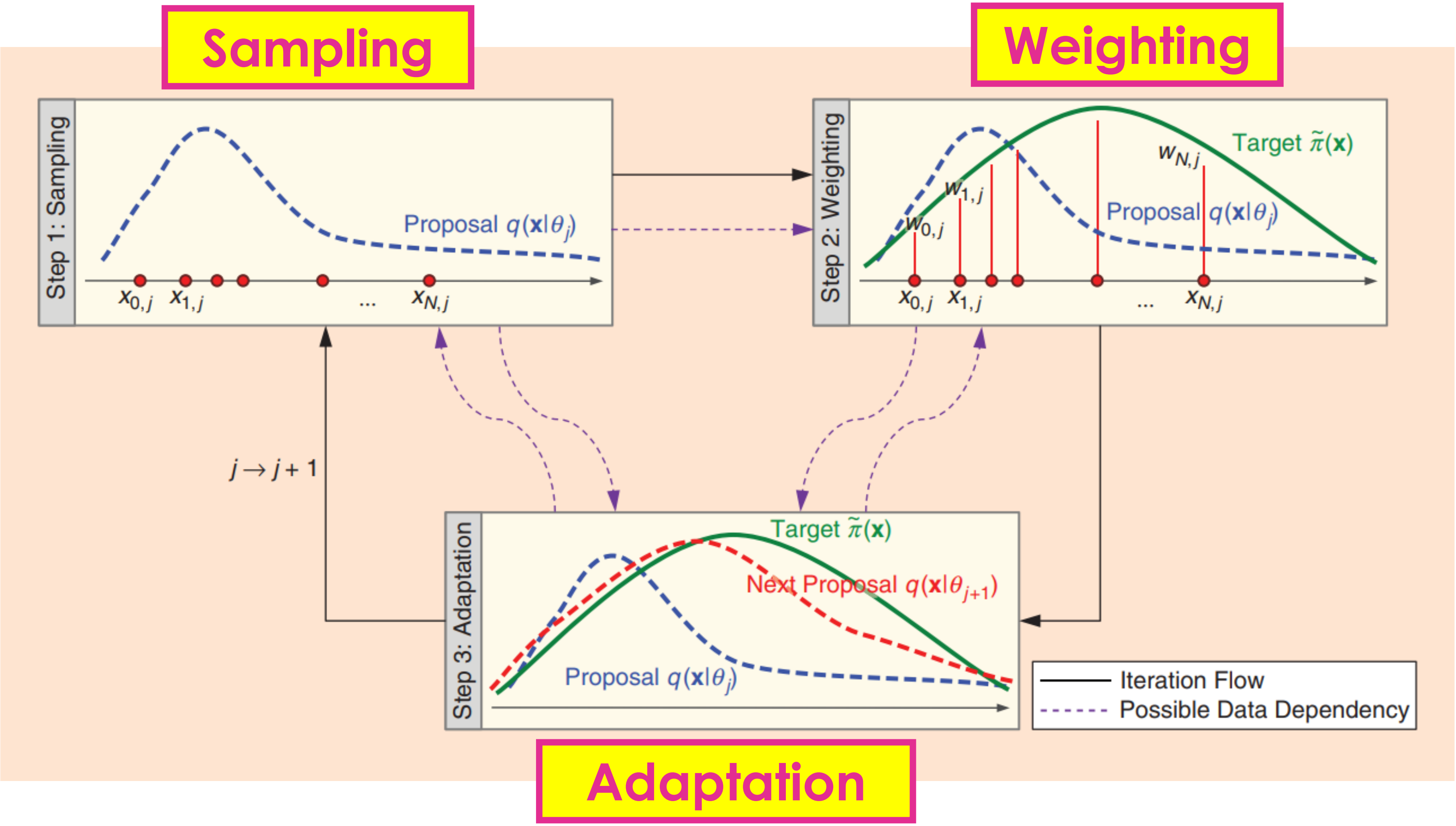
Blue dots = Samples

…, $\boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$ can … roximate the … est:

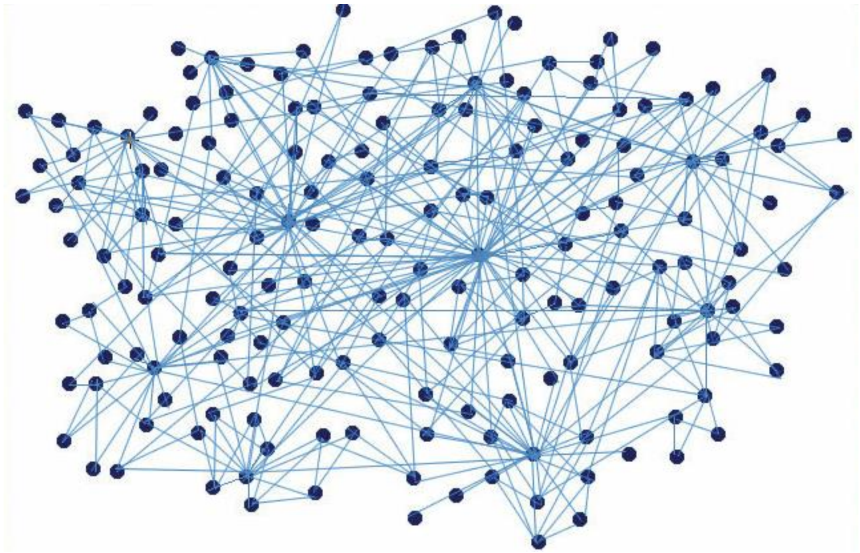
$$\text{Avg}[\boldsymbol{\theta}|\mathbf{y}_{1:T}] = \int \boldsymbol{\theta} \times p(\boldsymbol{\theta}|\mathbf{y}_{1:t}) d\boldsymbol{\theta}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}^{(m)}$$

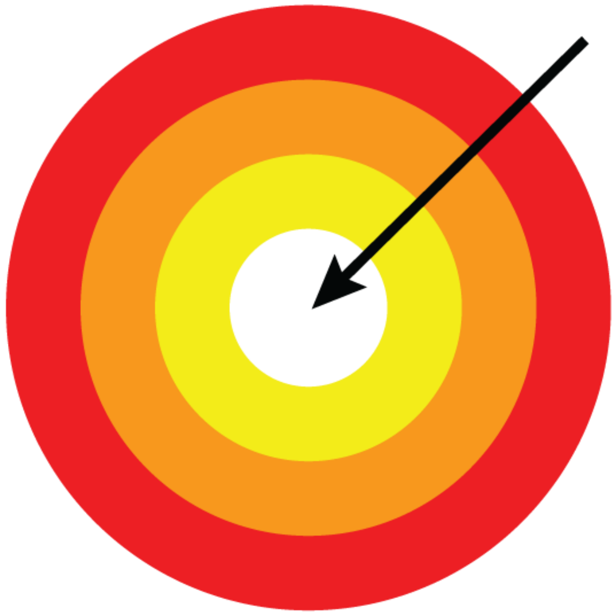
Adaptive Importance Sampling (AIS)



Computational Aspects



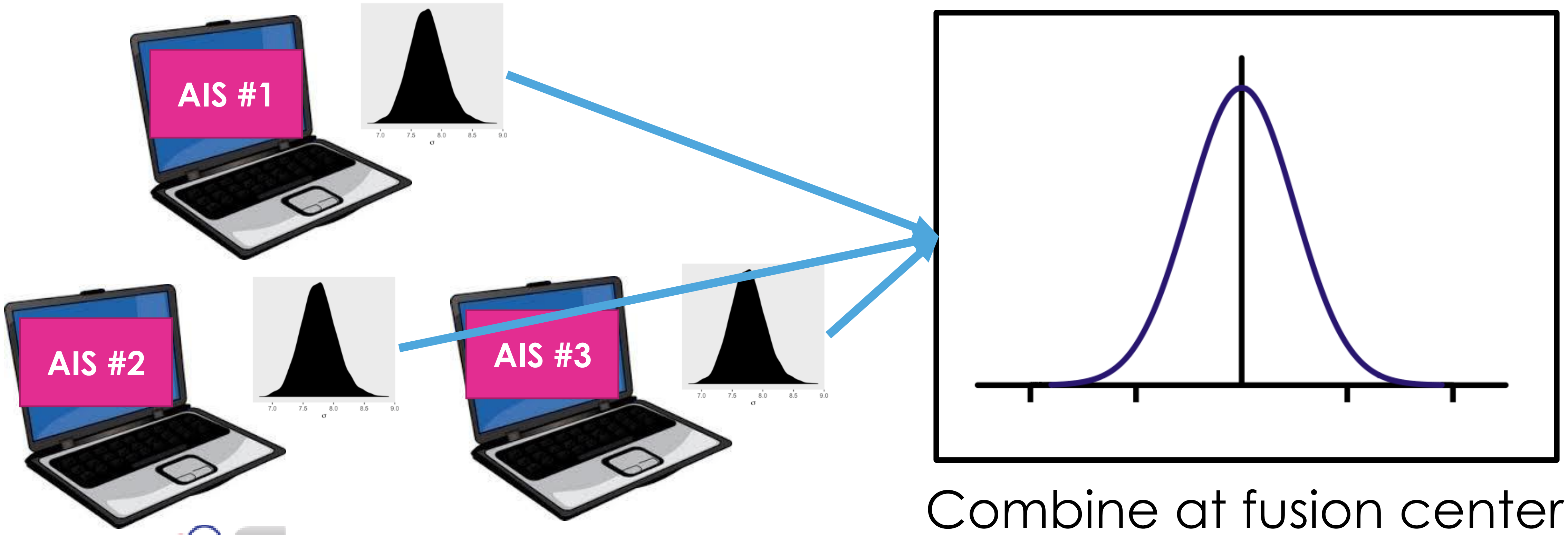
Scalability



Validation

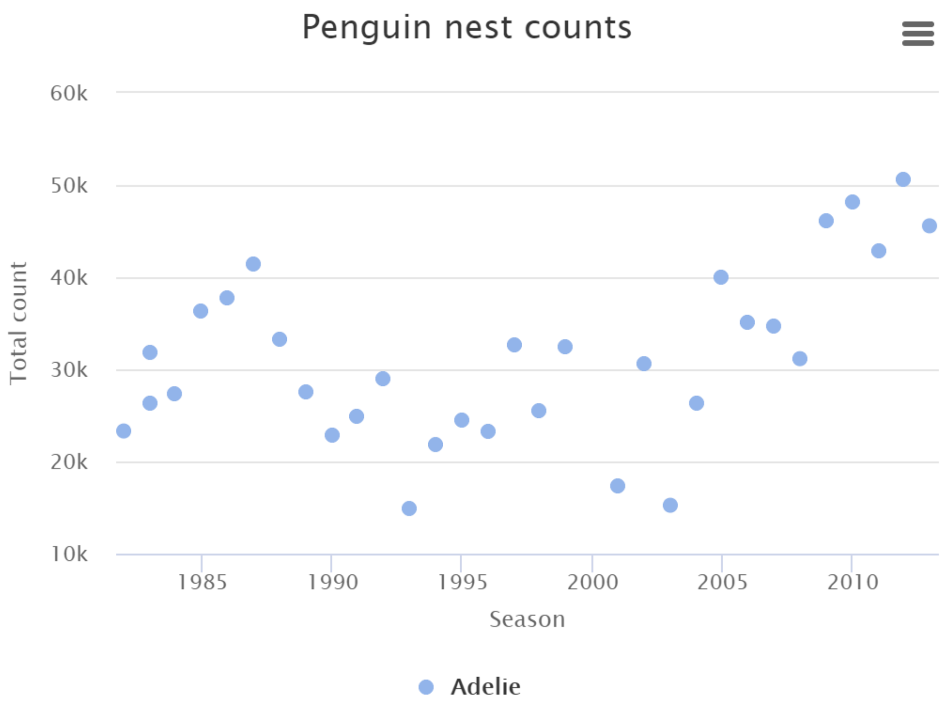
Parallel Implementation

- Luckily for us, AIS is an easily parallelizable algorithm!



Combine at fusion center

Penguin Population Dynamics

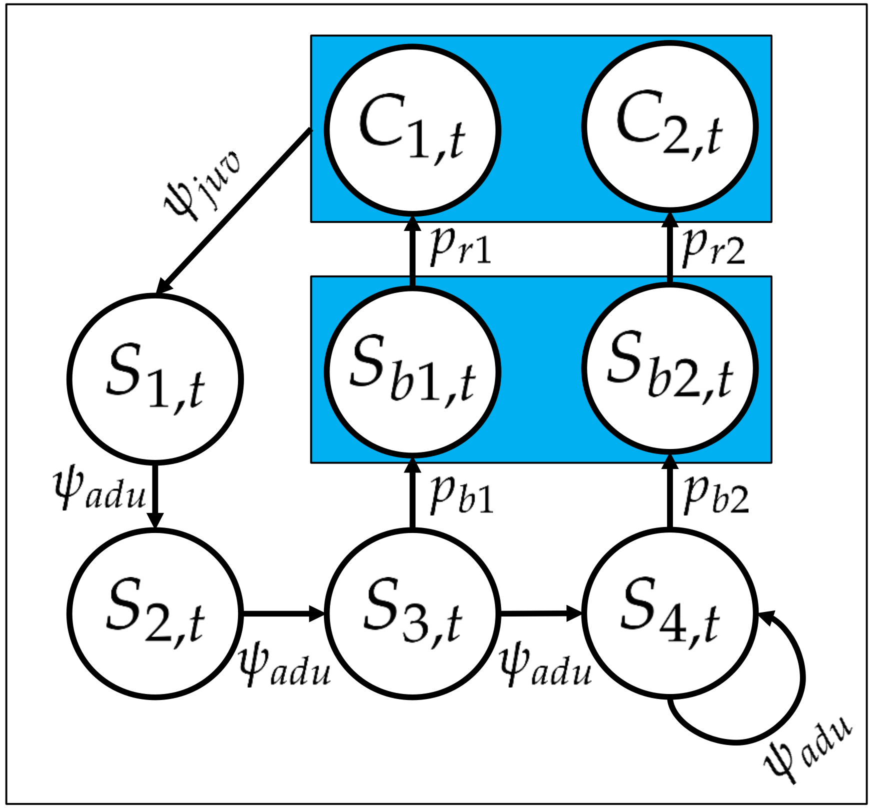


Time-series of abundance counts for a particular penguin colony.

Banded Adélie penguin.

We want to understand this system without using *mark-recapture*!

Stage-Structured Model



Life cycle diagram with J=4.

$$\boldsymbol{\theta} = [\psi_{adu}, \psi_{adu}, p_{b1}, p_{r1}, \ldots]^\mathsf{T}$$

$$\mathbf{x} \left\{ \begin{aligned} S_{1,t}^{(\ell)} &\sim \text{Bin}\left(0.5C_{t-1}^{(\ell)}\psi_{juv}\right) \\ S_{j,t}^{(\ell)} &\sim \text{Bin}\left(S_{j-1,t-1}^{(\ell)}, \psi_{adu}\right), \quad j = 2, \ldots, J-1 \\ S_{J,t}^{(\ell)} &\sim \text{Bin}\left(S_{J-1,t-1}^{(\ell)} + S_{J,t-1}^{(\ell)}, \psi_{adu}\right) \\ S_{bj,t}^{(\ell)} &\sim \text{Bin}\left(S_{j+2,t}^{(\ell)}, p_{bj}^{(\ell)}\right), \quad j = 1, \ldots, J-2 \\ C_{j,t}^{(\ell)} &\sim \text{Bin}\left(2S_{bj,t}^{(\ell)}, p_{rj}^{(\ell)}\right), \quad j = 1, \ldots, J-2 \end{aligned} \right.$$

$$\mathbf{y} \left\{ \begin{aligned} \tilde{S}_{b,t}^{(\ell)} &\sim \mathcal{N}\left(S_{b,t}^{(\ell)}, (\sigma_s S_{b,t}^{(\ell)})^2\right), \\ \tilde{C}_{t}^{(\ell)} &\sim \mathcal{N}\left(C_{t}^{(\ell)}, (\sigma_c C_{t}^{(\ell)})^2\right), \end{aligned} \right.$$

$$S_{b,t}^{(\ell)} = \sum_{j=1}^{J-2} S_{bj,t}^{(\ell)}$$

$$C_{t}^{(\ell)} = \sum_{j=1}^{J-2} C_{j,t}^{(\ell)}$$

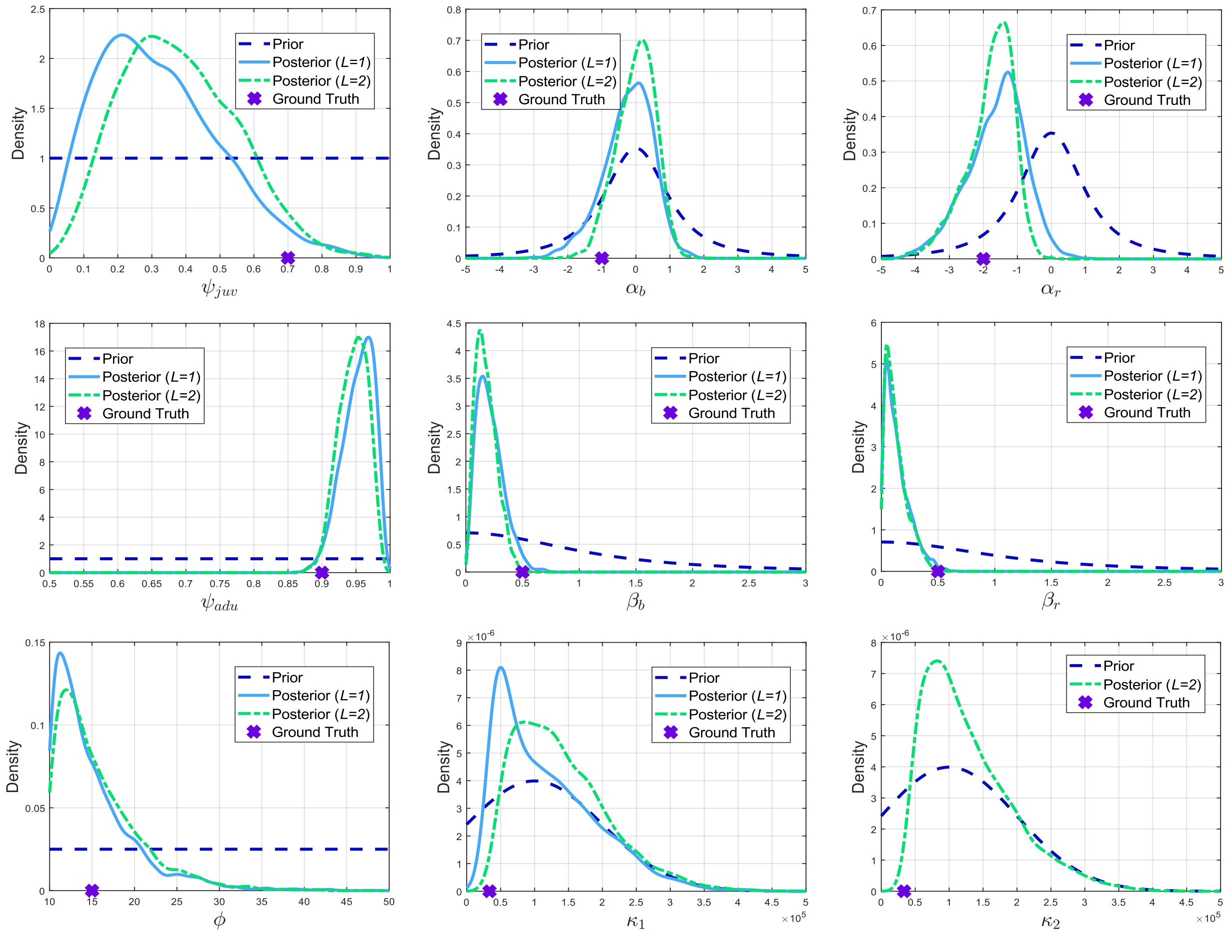
We observe a sum of the states of interest, but not each individual age class!

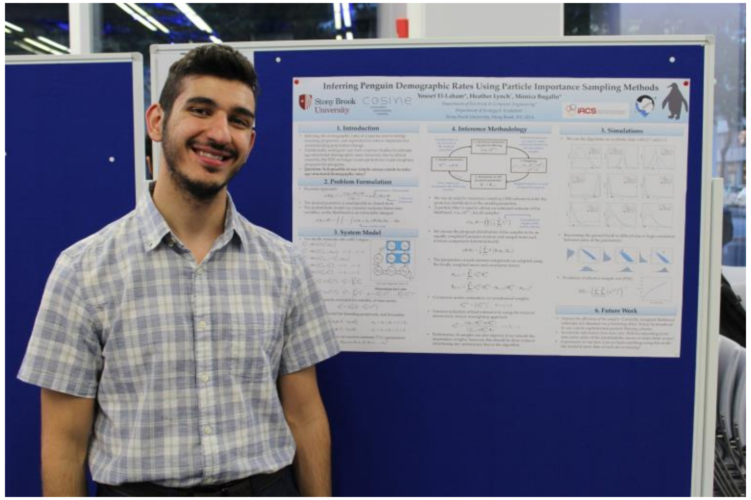
Goals of Interdisciplinary Work

- We address the following research questions:

  1. Using the framework of Bayesian inference, what can we learn about the parameters?
  2. If we use data across multiple sites and assume they share most parameters, does inference become easier or harder?
  3. Is it possible to learn something in the case that there is a lot of missing data?
  4. How can we scale our inference algorithms to hundreds of time-series?

Preliminary Results





Work presented at MLSS 2019 in London!

So far, we have some results on synthetic data!

Timeline

Aug 2019

Sep 2019
Finalize results on synthetic data for penguin problem

Oct 2019
Obtain results for real-penguin data

Nov 2019
Submit a journal article about my work on scaling AIS algorithms

Jan 2020
Try to start a student-run machine learning tutorial series through the IACS

Mar 2020
Begin to address the problems of dealing with the "big-data" scenario

April 2020
Develop an AIS library designed for parallel implementation

Jul 2020
Start writing a draft of my dissertation

Aug 2020

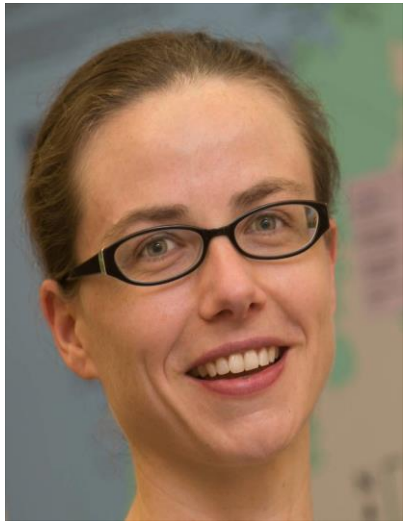
Collaborator (Theory)

Dr. Petar Djurić
Electrical Engineering
Department Chair

Principal Investigator (PI)

Dr. Mónica Bugallo
Electrical Engineering
Director of WISE

Collaborator (Application)



Dr. Heather Lynch
Ecology & Evolution
IACS Faculty Member

My Lab!

cosine
communication
signal processing
networking

MAPPPD
Mapping Application
for Penguin
Populations and
Projected Dynamics

Thank you!