Machine Learning Methods for Timing of Biological Events

Tom Dietterich

with Dan Sheldon, Tao Sun, Liping Liu, Evan Goldman, Erin Childs, Olivia Poblacion, Jeffrey C. Miller, Julia A. Jones (OSU)

Steve Kelling, Andrew Farnsworth, Wes Hochachka, Frank Le Sorte, Kevin Webb, Theo Damoulas (CLO)

Rich Caruana (Microsoft Research)

NSF

LTER

The Cornell Lab of Ornithology
Exploring and Conserving Nature

HJ Andrews Experimental Forest
Long Term Ecological Research

OSU
Oregon State UNIVERSITY

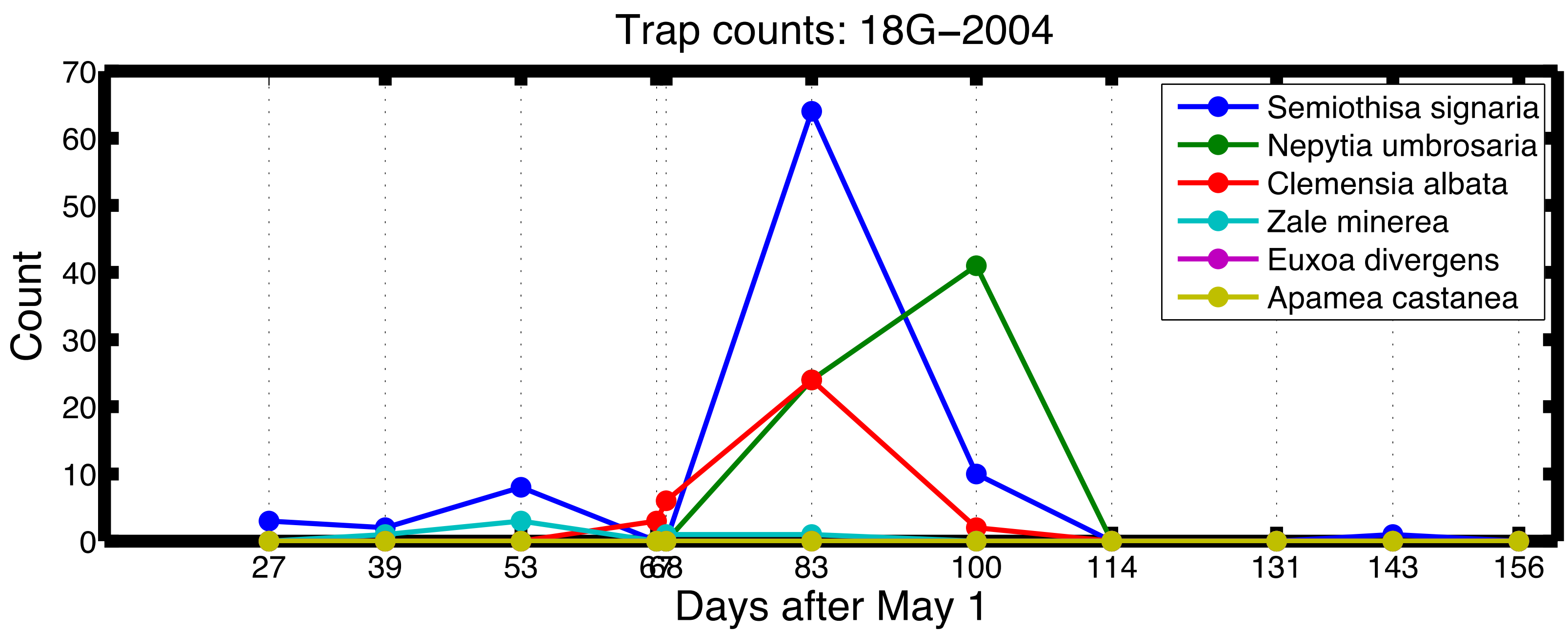
Phenology and Climate

- One potential impact of climate change is to change the timing of life cycle events (“Phenology”)
  - Bird Migration
  - Moth Flight Times
  - Pollinator Flight Times
  - Timing of leaf-out and flowering
- What determines the timing of these events?
  - Day length? (will not change with climate)
  - Temperature, precipitation, wind (will change with climate)
- Phenological asynchrony could lead to major changes in food web structure
  - Local extinctions
  - Rapid evolutionary pressures

Challenges to Data-Driven Modeling of Phenology

- *What we have*: periodic observations of organism "activity"
  - Moth trap counts
  - Bird surveys
- *What we want*: timing of life history events
  - When did adult moths emerge from cocoons?
  - When did migrating birds arrive?
- How to bridge the gap?

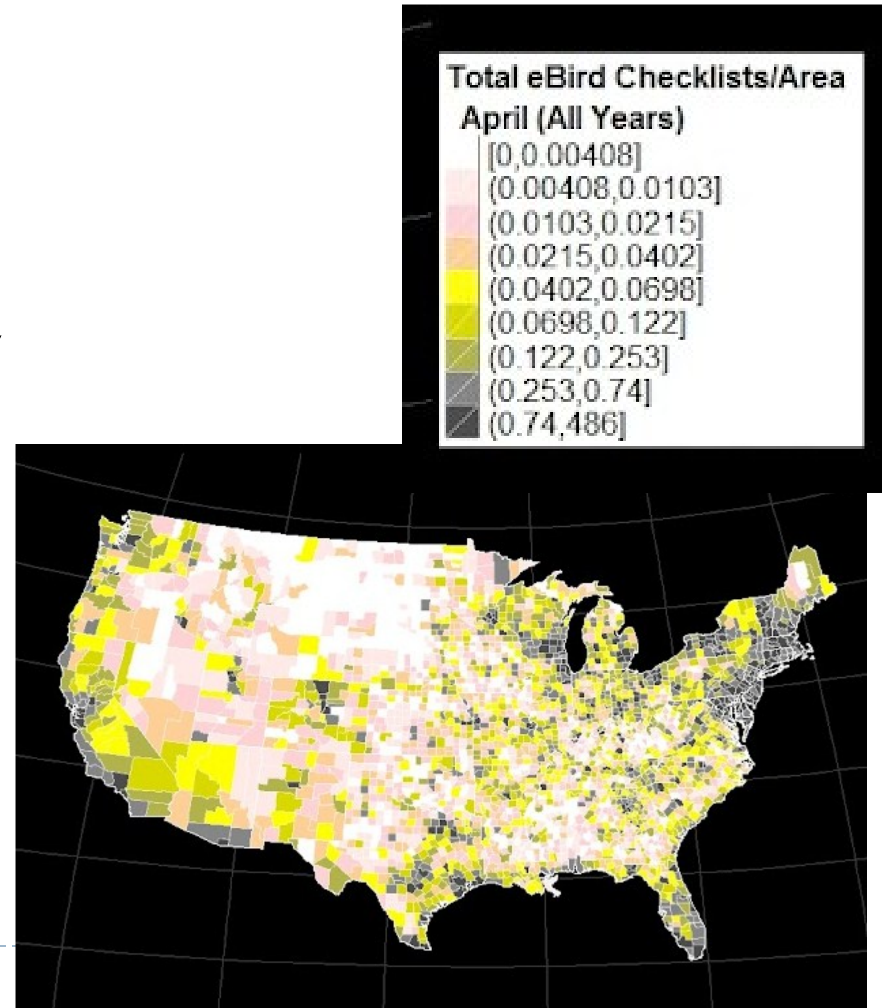
Example: Moth Trap Counts



What was the flight period of Nepytia umbrosaria in 2004?

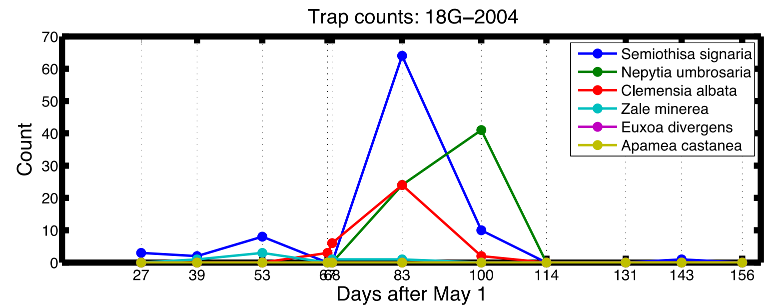
Example: eBird Data

- Bird watchers record their observations in a database through eBird.org.
  - “Citizen Science”
- Features
  - LOTS of data!
    - ~3 million observations reported in May
  - ~3,000 bird species
  - Year-round, Continent-scale



Challenges

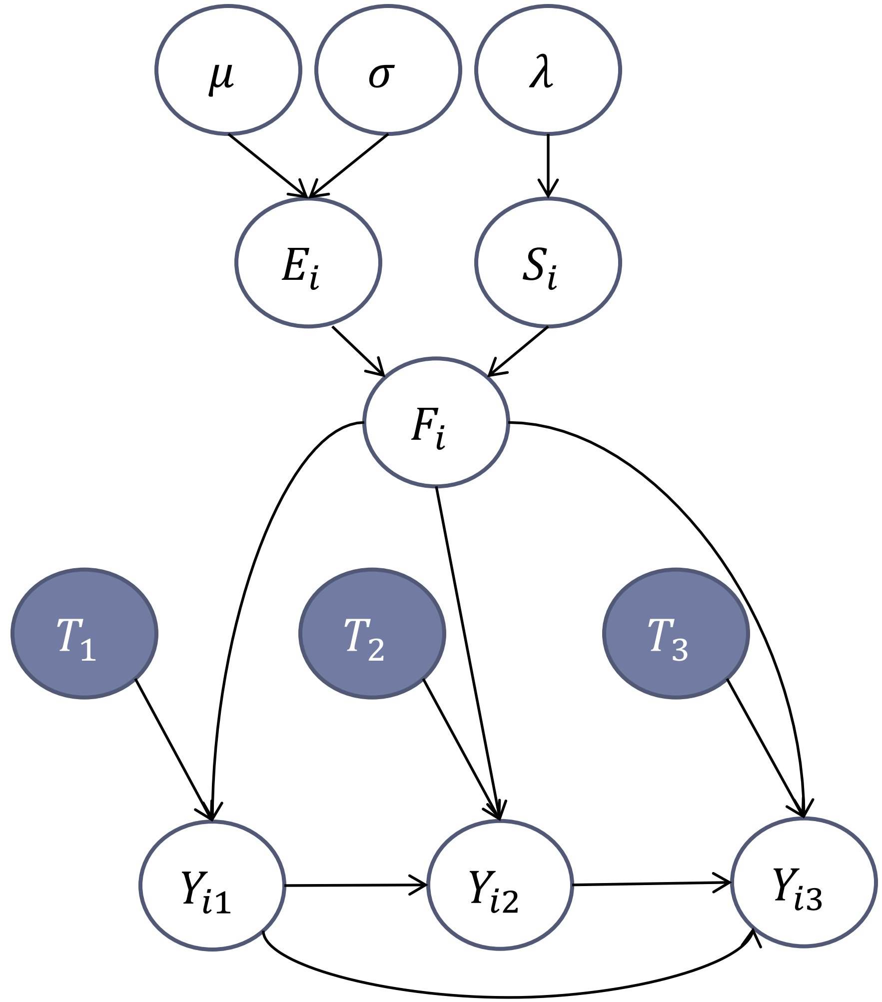
- We do not *directly* observe the events we are interested in
  - Moth emergence
  - Bird arrival
- Surveys are infrequent
  - May miss "peak" activity



- Naïve approaches don't use all of the data
  - Date of first moth, first bird
  - Date of maximum abundance

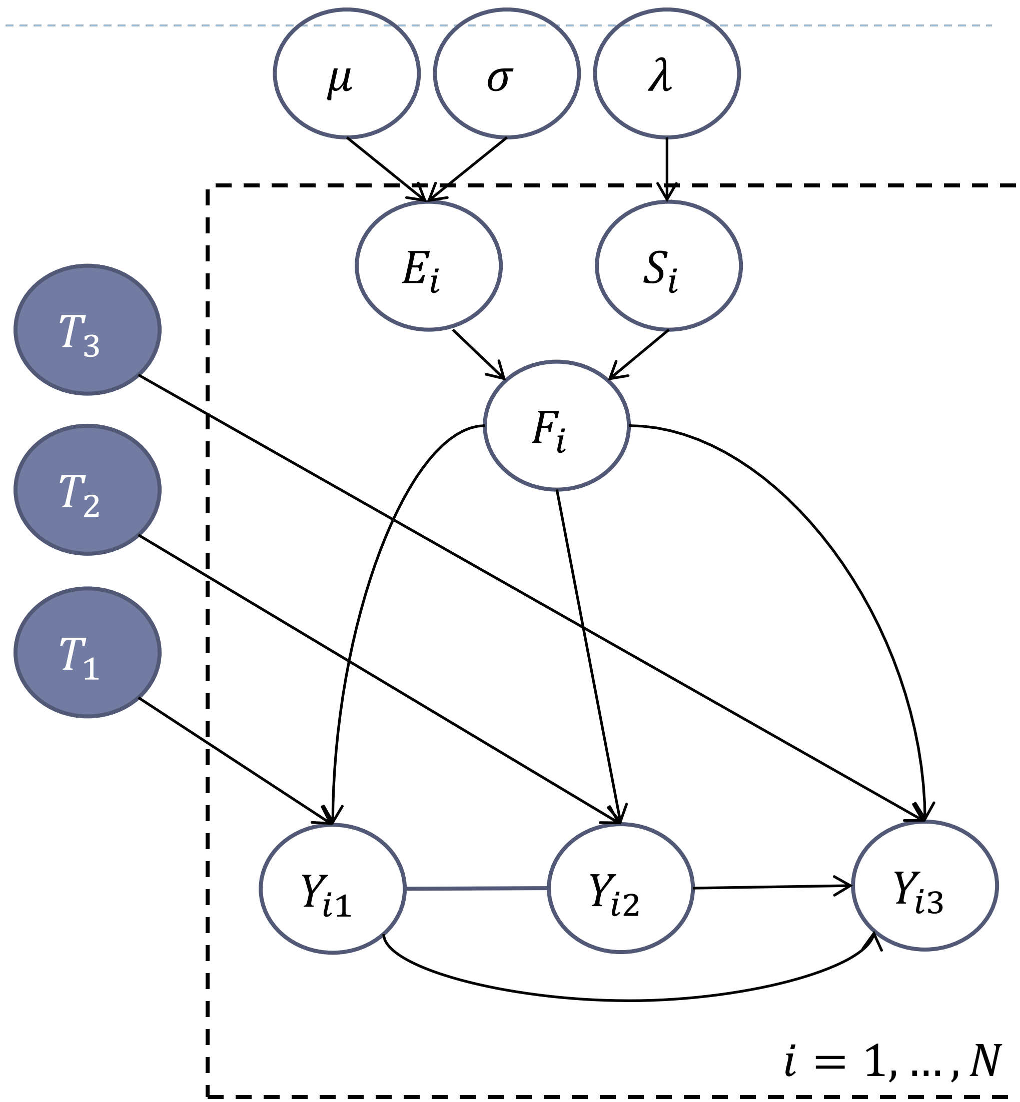
A General Approach: Collective Graphical Models

- Step 1: Define a model of the behavior of individual organism
  - E_i: emergence date for organism i
    - $E_i \sim \mathrm{Norm}(E_i | \mu, \sigma)$
  - S_i: lifespan
    - $S_i \sim \mathrm{Exp}(S_i | \lambda)$
  - F_i: flight period (start, end)
    - start = E_i
    - end = $E_i + S_i$
  - T_t: trapping date
  - Y_{it}: 1 if moth was trapped on date t, 0 otherwise



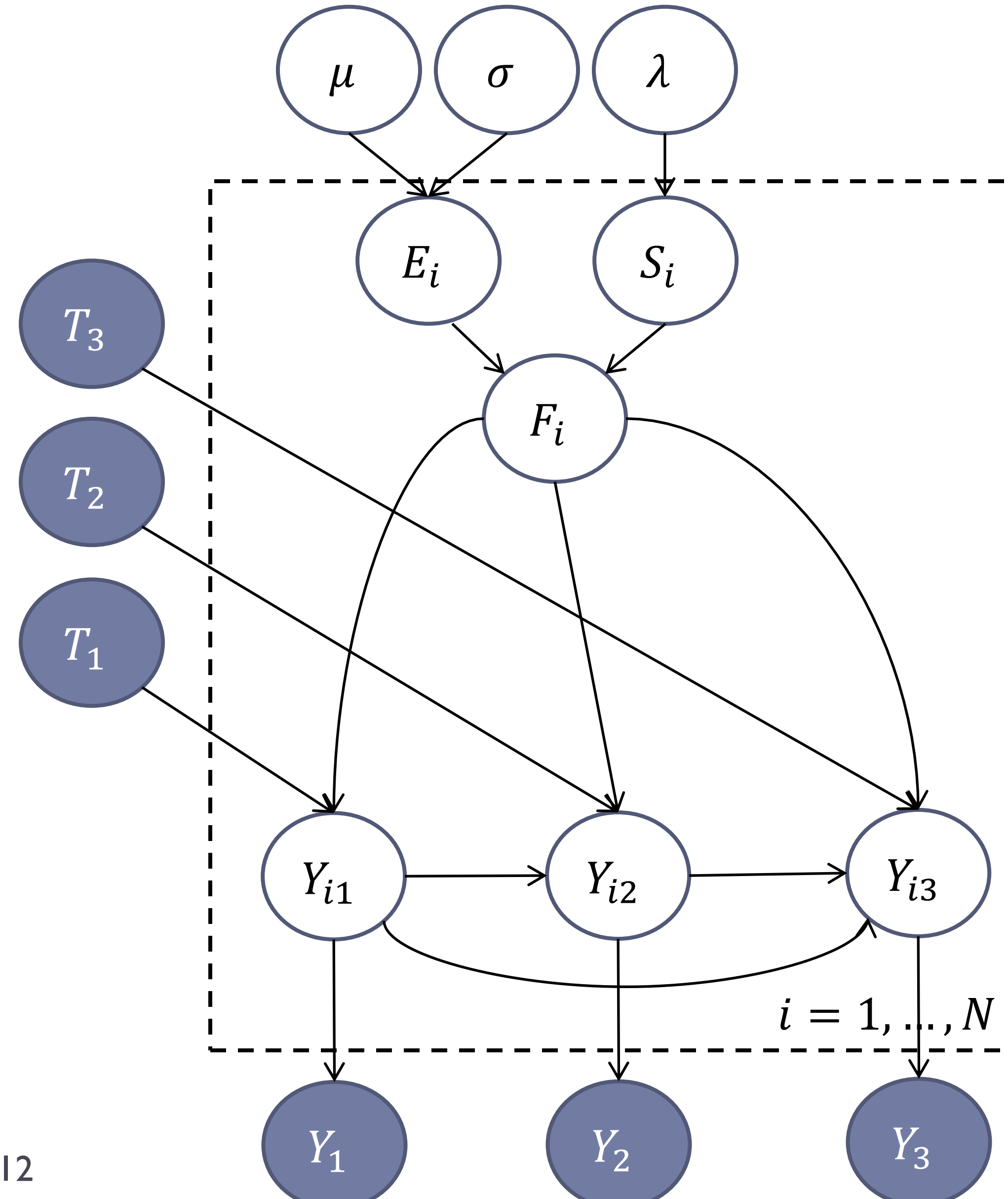
Step 2: Assume a population of iid individuals

- We assume all moths are drawn from the same distribution



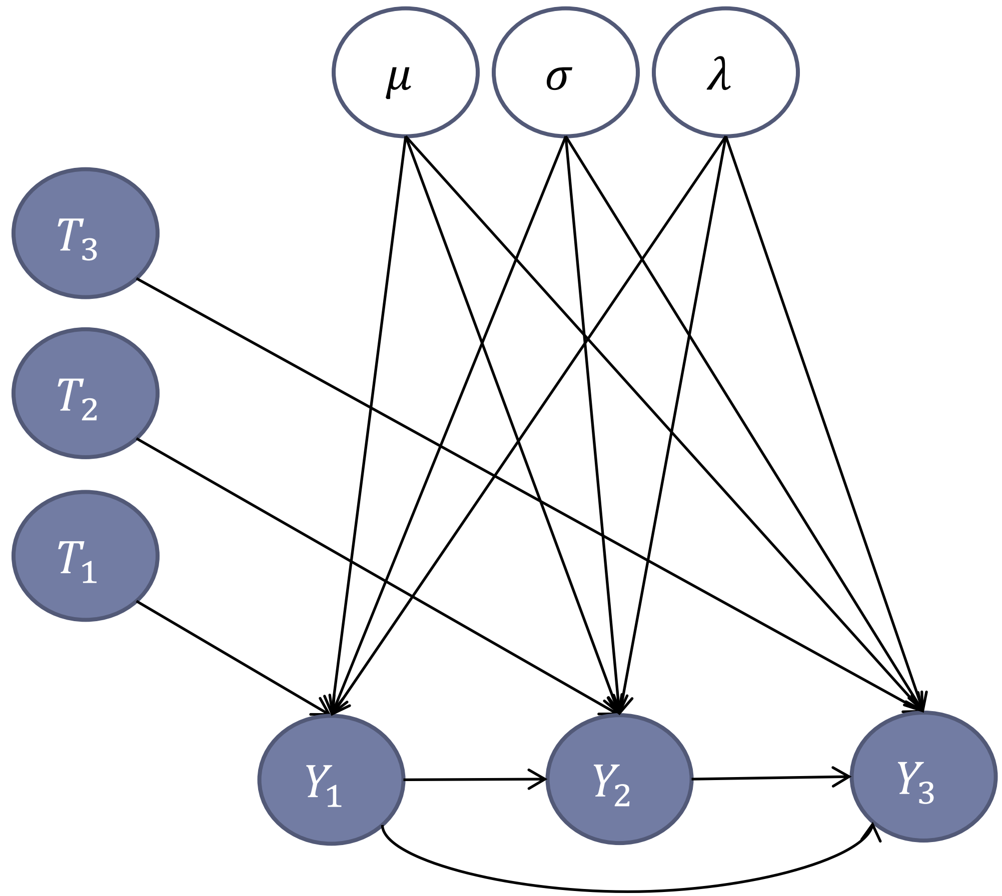
Step 3: Introduce aggregate observation variables

- $Y_1 = \sum_i Y_{i1}$
- $Y_2 = \sum_i Y_{2i}$
- $Y_3 = \sum_i Y_{3i}$



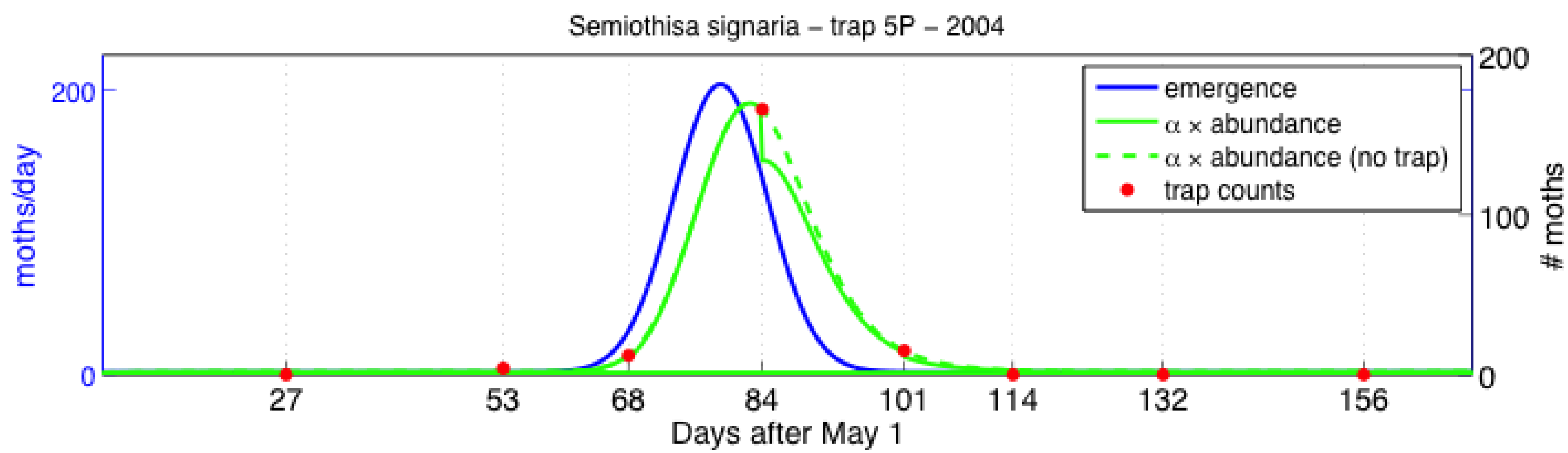
Step 4: Marginalize away the individuals

- Theorem (Dawid & Lauritzen, 1993): Resulting graph has same dependency structure as the individual model
- No combinatorial explosion of dependencies



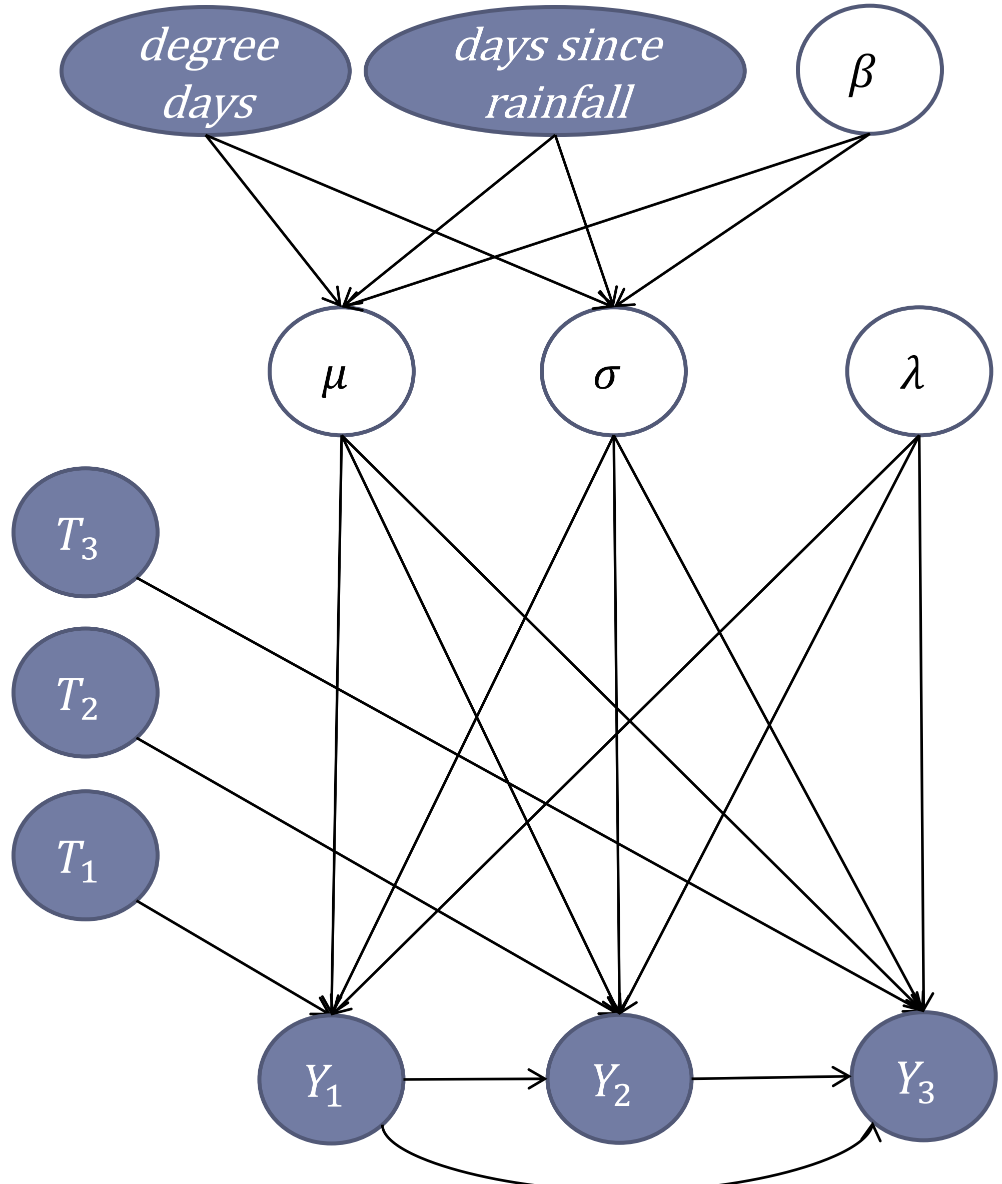
Step 5: Fit via maximum likelihood (etc.)

- Example of fitted model



Modeling Climate Dependence

- Introduce covariates on emergence time
- Linear regression to determine mean and variance



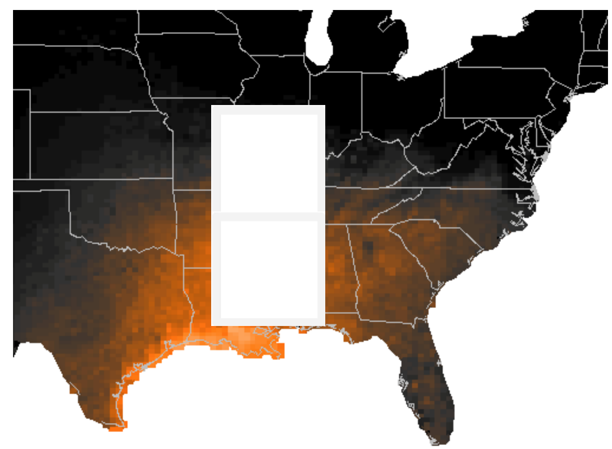
CGM for Bird Migration

- Define grid over US
- Aggregate eBird observations into # birds per cell

Time

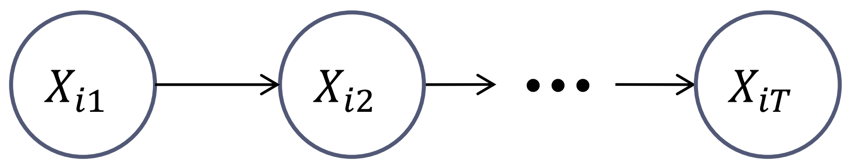
Cell

	1	2	3
A	87	61	22
B	13	39	78

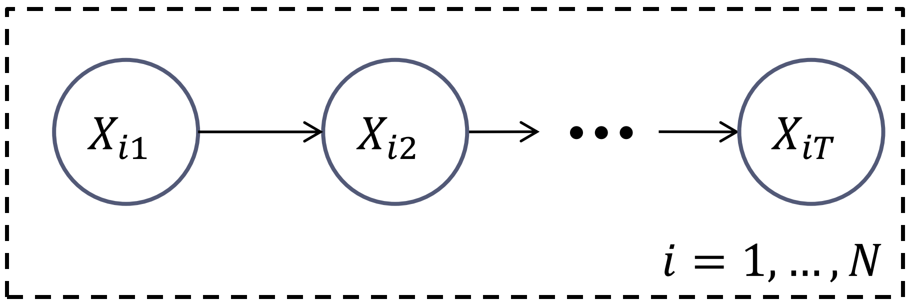


Step 1: Individual Model

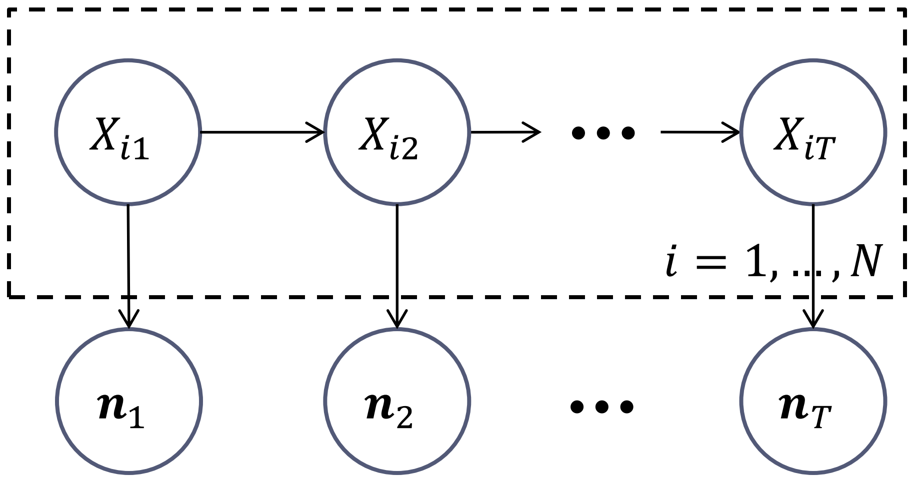
- Each bird is a sample from a Markov Chain
  - X_{it}: Cell of bird i at time t



Step 2: Population of Individuals



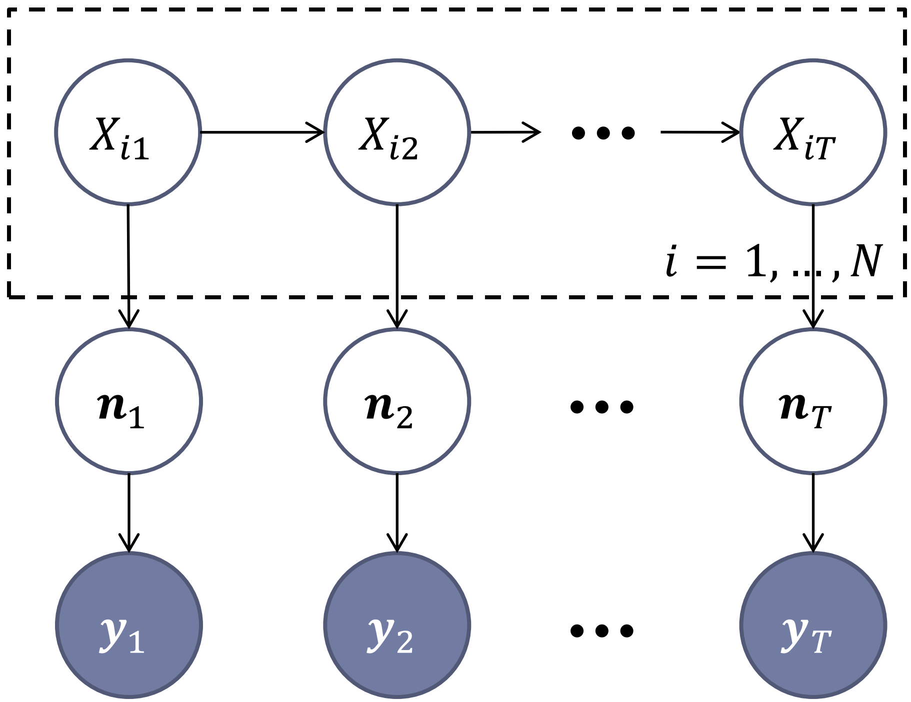
Step 3: Derive Aggregate Counts



true # birds
in each cell at time t

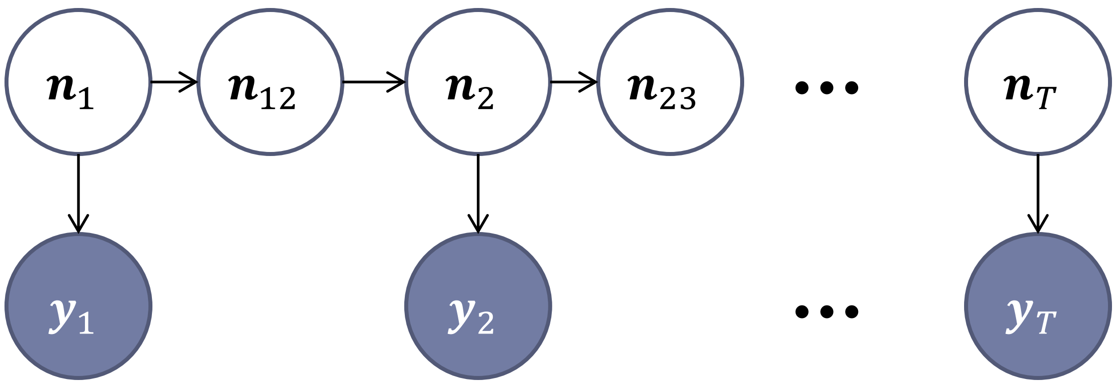
Step 3b: Introduce Stochastic Observation Model

- Each bird is detected with probability d_t by eBirders

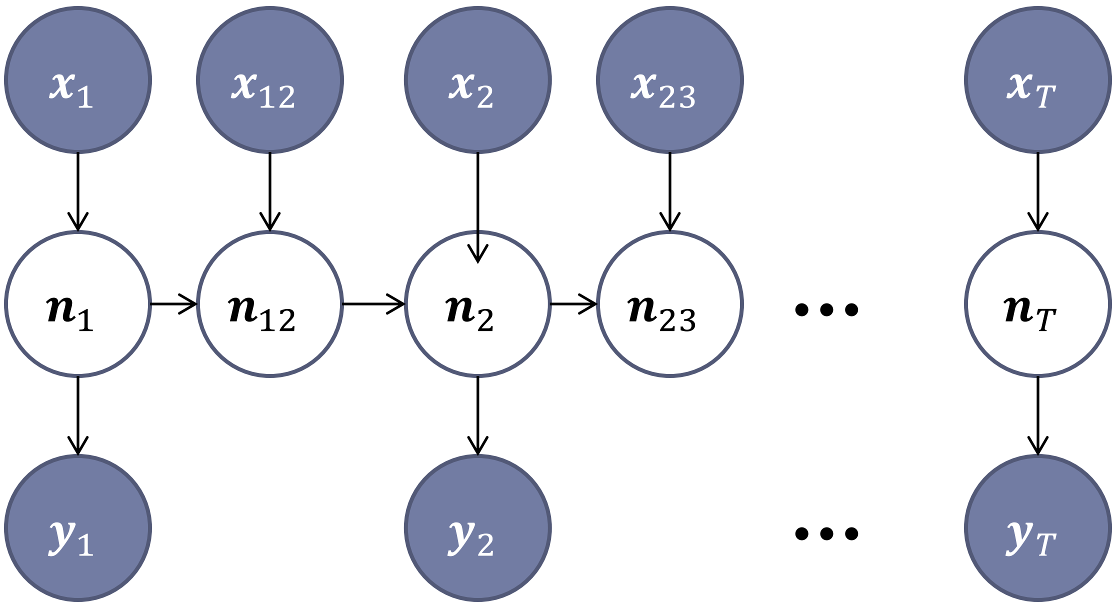


eBird observations in each cell at time t

Step 4: Marginalize away the individuals



Step 5: Add climate covariates



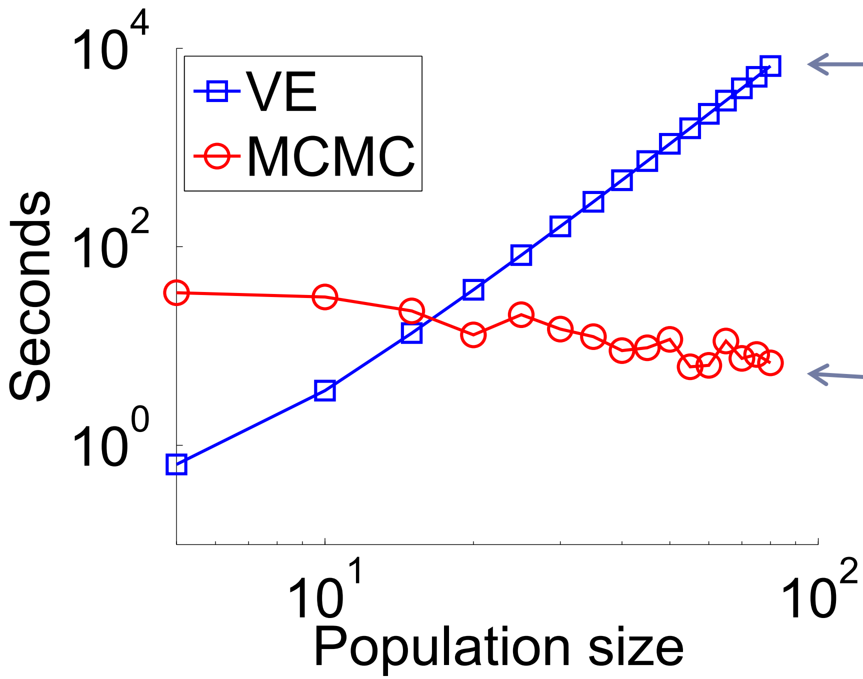
Step 6: Fit via maximum likelihood

- Very challenging inference problem
- State = all ways of partitioning N birds across K sites
- Solution: Gibbs sampling algorithm that takes time independent of N

Gibbs Sampler Experiment

- Running time on simple GCM task

[Sheldon & Dietterich, NIPS 2011]

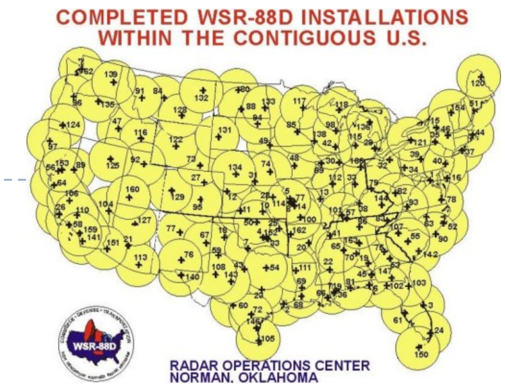


Best exact method (variable elimination; cubic in M)

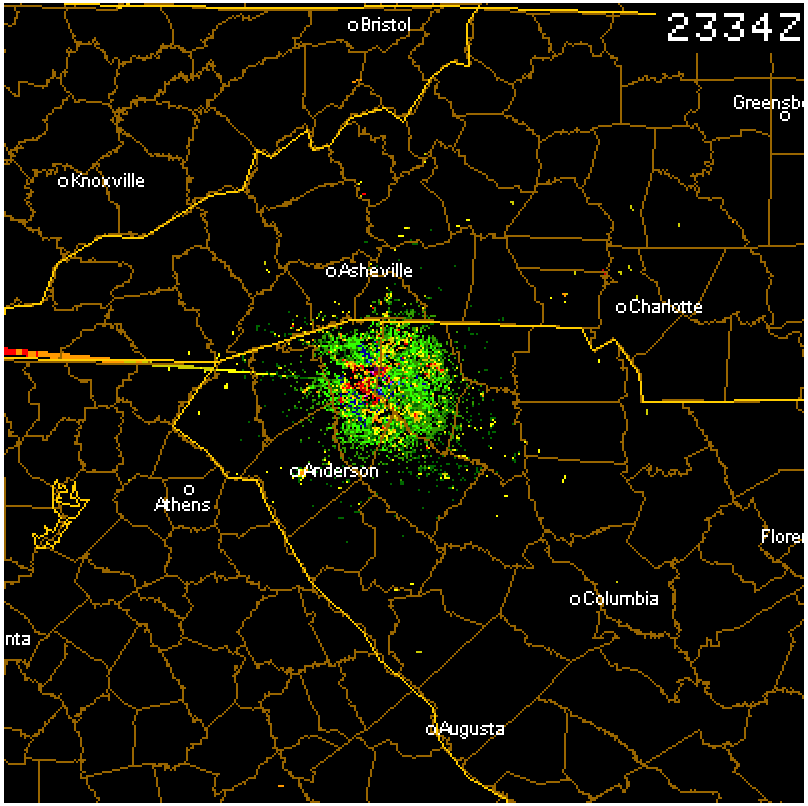
Our method (to 2% relative error)

- Running time *independent of population size*
  - Previous best: exponential

New Project: BirdCast

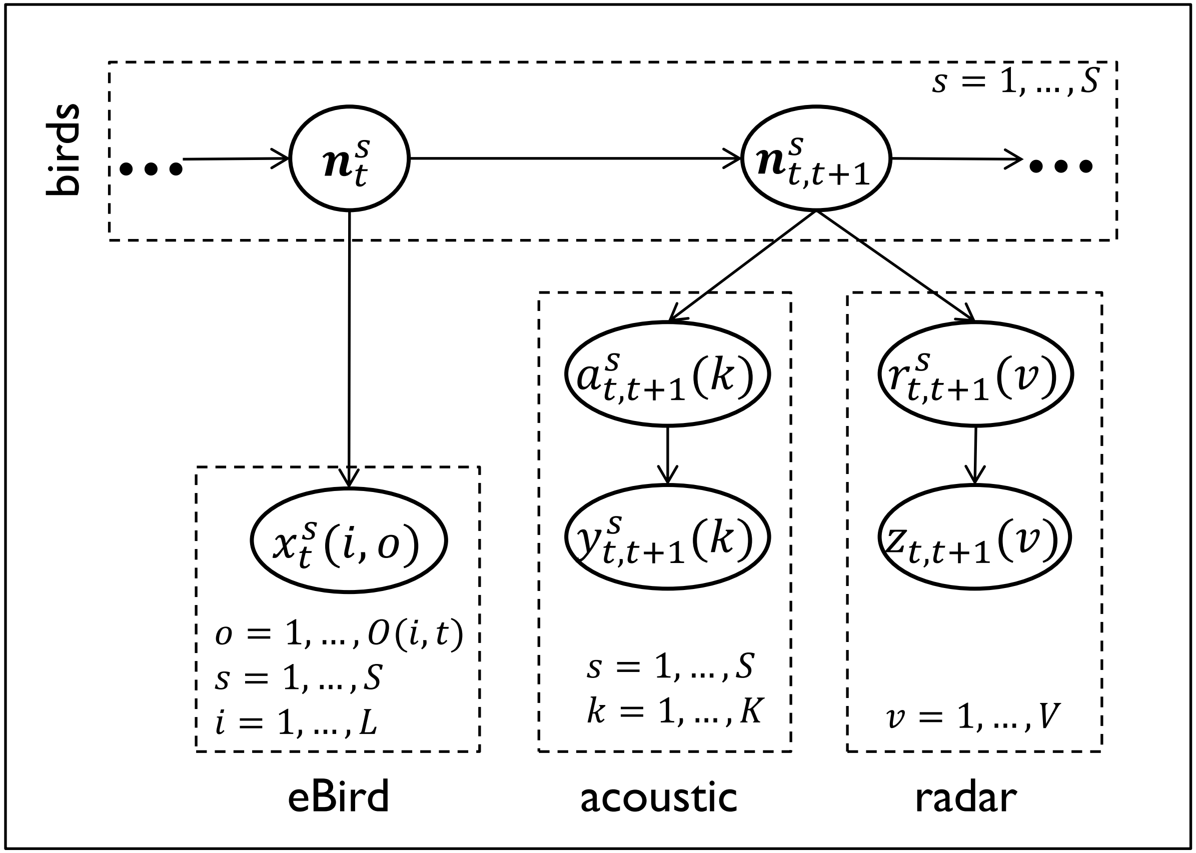


- Goal: Continent-wide bird migration forecasting
- Additional data sources:
  - Doppler weather radar
  - Night flight calls
  - Wind observations (assimilated to wind forecast model)



BirdCast Collective Graphical Model:

- $n_t^s(c)$ = # of birds of species s at cell c and time t.
- $x_t^s(i,o)$ = eBird count for visit o at site i species s and time t
- $y_{t,t+1}^s(k)$ = # of flight calls for species s at site k on the night $(t, t+1)$
- $z_{t,t+1} =$ # of birds (all species) observed at radar v on night $(t, t+1)$
- Occupancy changes each night
- Covariates (not shown): wind, precipitation, land cover, green up, elevation, urbanization



Concluding Remarks

- Collective Graphical Models provide a formalism for modeling phenology from aggregate observations
  - assume a population of iid individuals
  - introduce aggregate observation variables
  - marginalize away individuals
  - fit to data
- CGM Gibbs sampler has running time independent of population size N
  - we do not yet understand dependence on the number of cells K