Stickleback project presentation

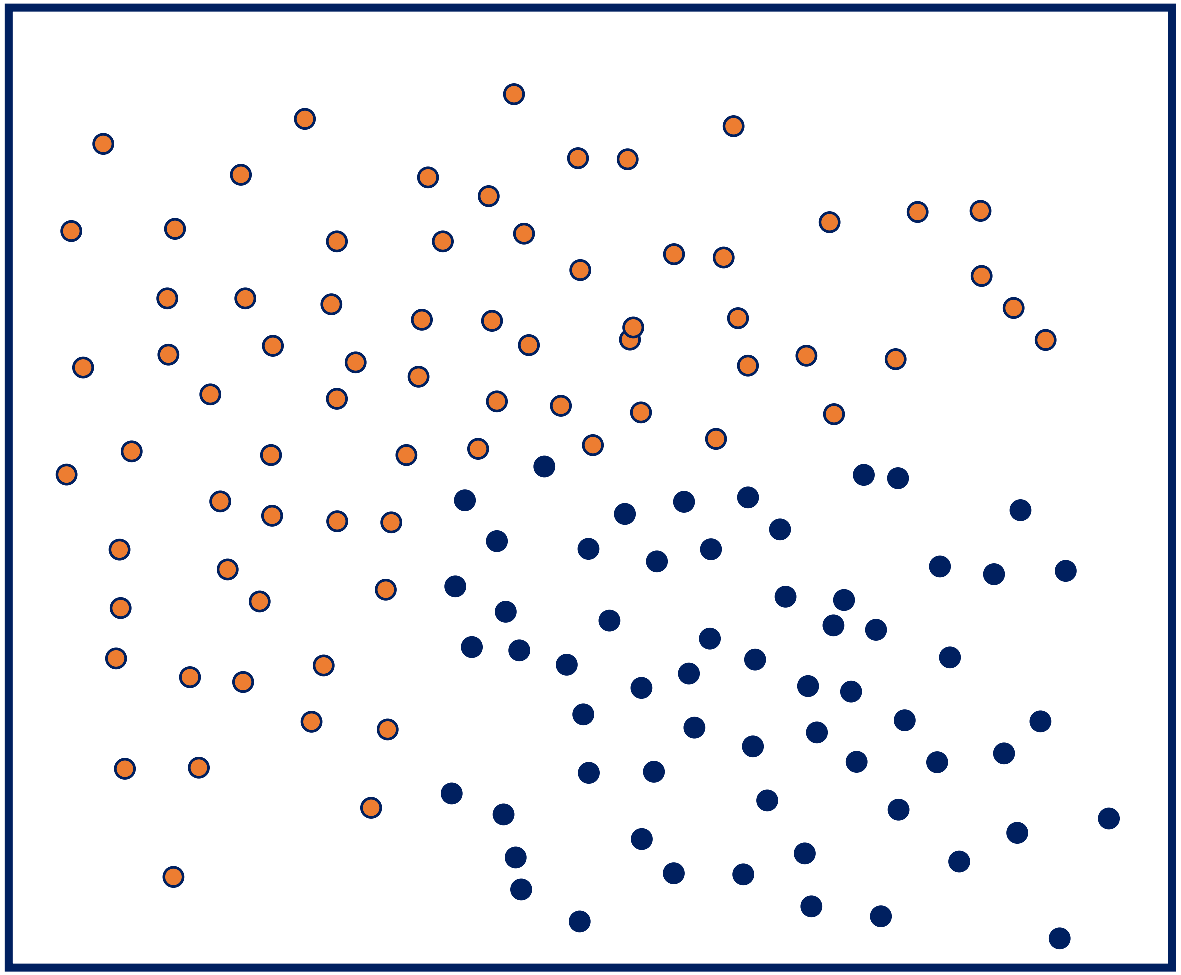
- 5-7 min
- Include the following elements:
  1. General background (see XX et al., 2019, will post others as well)
  2. Background on **your** samples (where collected, what studies used in, etc.)
  3. Sequencing metadata from your sample
     - How many reads? How many mapped? Average depth (assuming 460 Mb genome size)
  4. Syn/Nsyn differences of your samples compared to the reference
     - I'll provide code/instructions later today
  5. Summary/ conclusions

- For 535 students, you'll also write up the presentation into a 1000-3000 word summary (due May 8th, but extensions possible)

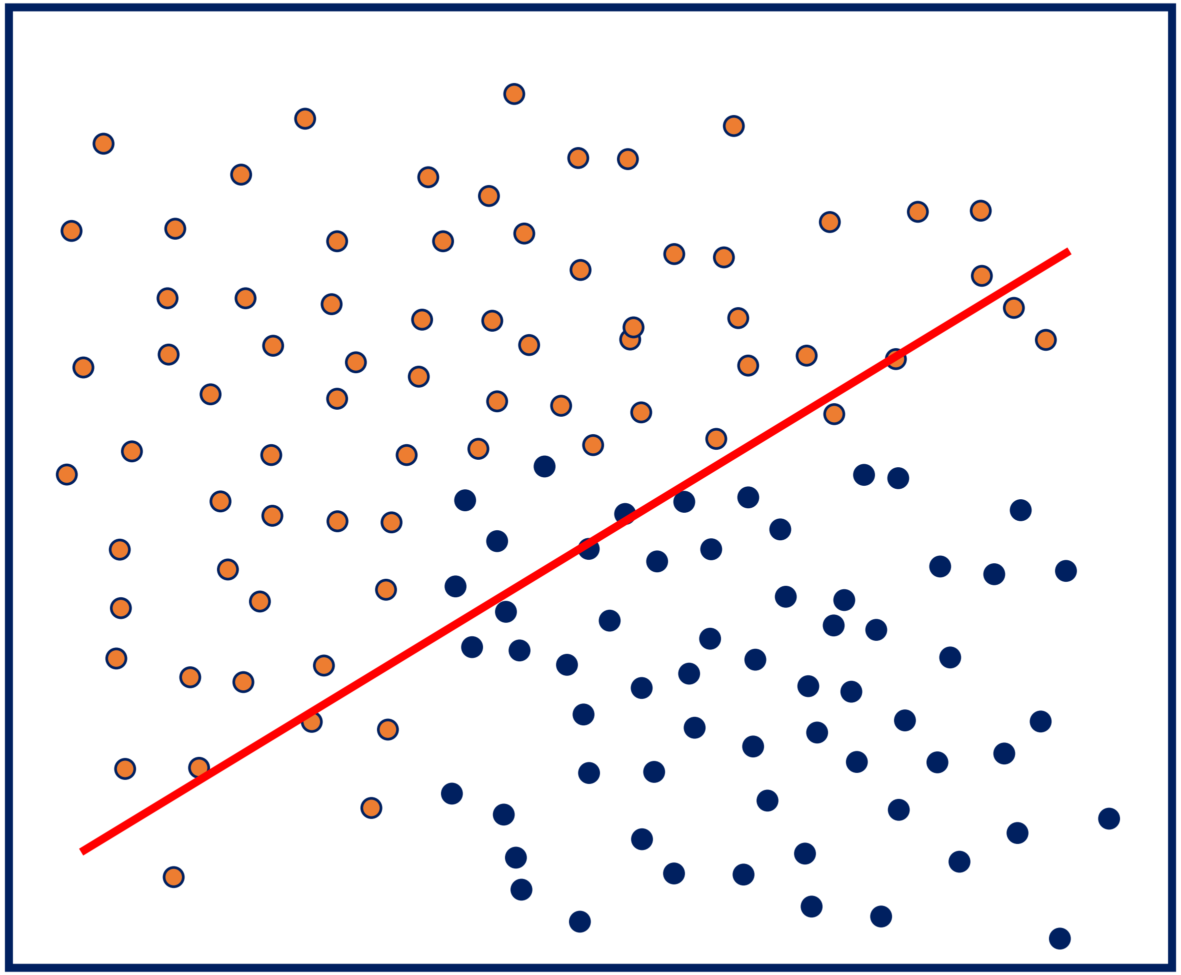
Machine Learning, Simulations, and Modeling

BIOL 435/535: Bioinformatics
April 26, 2022

Machine learning – classifying data



Machine learning – classifying data



Machine learning

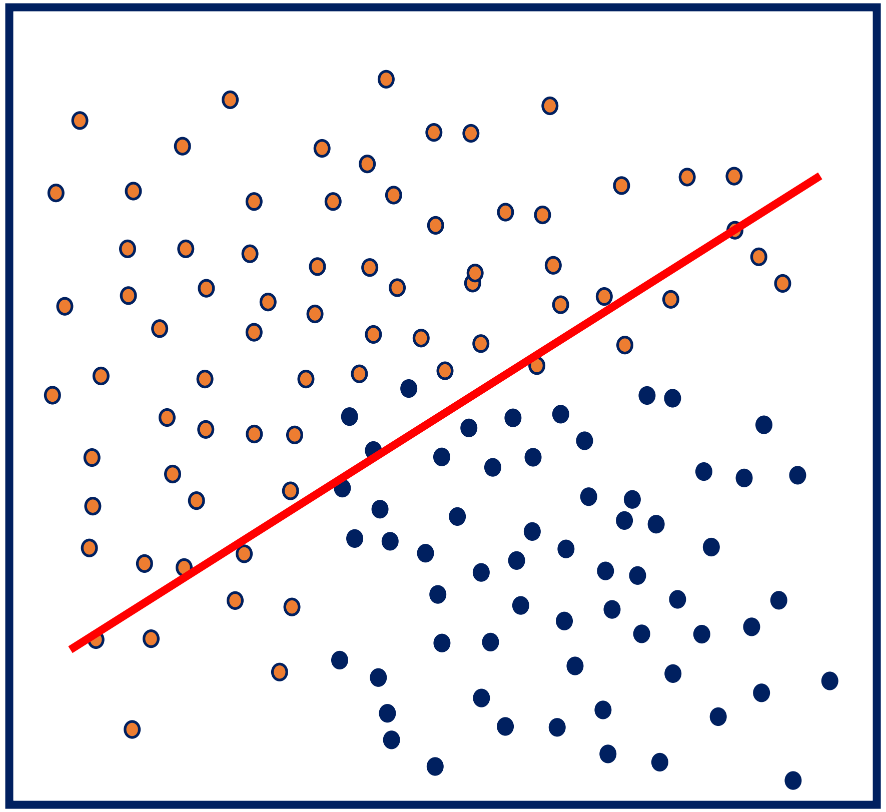
Supervised

- Trains on a **labeled** set of data
- Attempts to classify unlabeled data based on what it learned
- Regression, Support Vector Machines, Decision Trees, Naïve Bayes, etc.

Unsupervised

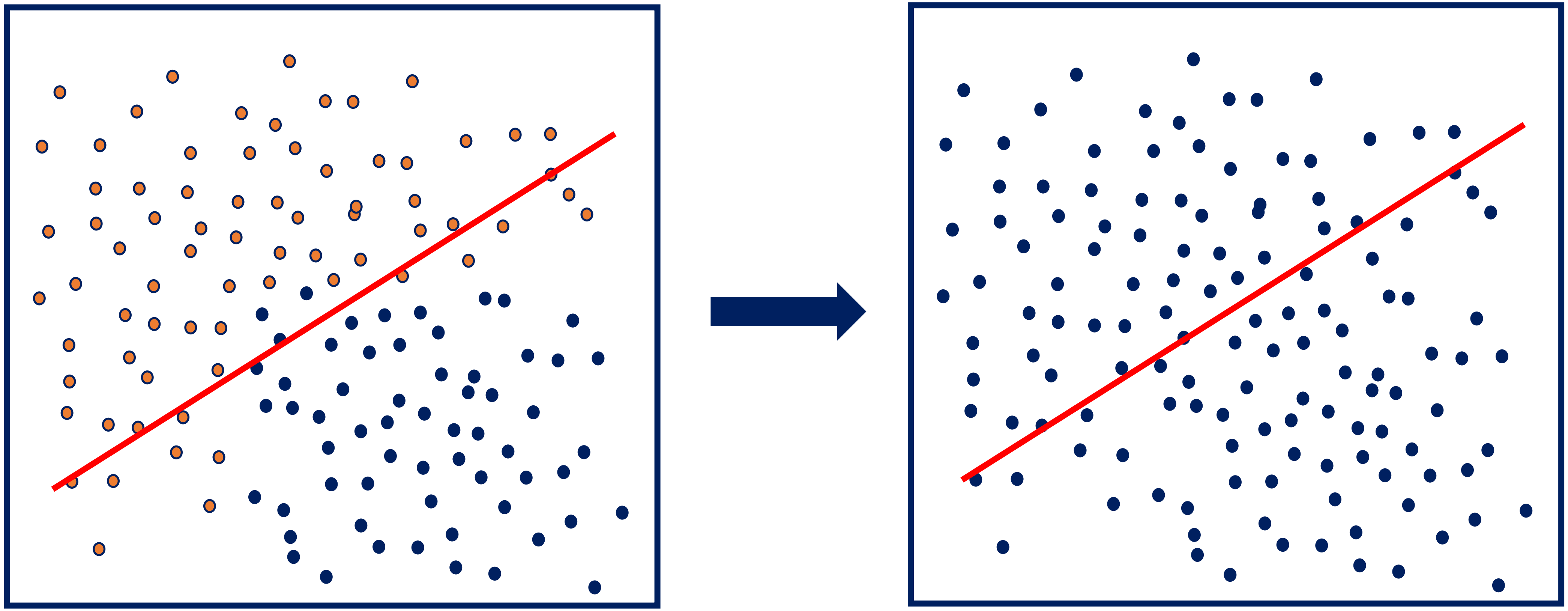
- Trains on an **unlabeled** set of data
- Attempts to mimic the data, then compares to original for errors
- Clustering, principal component analysis, neural networks, etc.

Supervised Machine Learning



Training dataset

Supervised Machine Learning



Training dataset

Test dataset

Markov Chain + Monte Carlo simulation

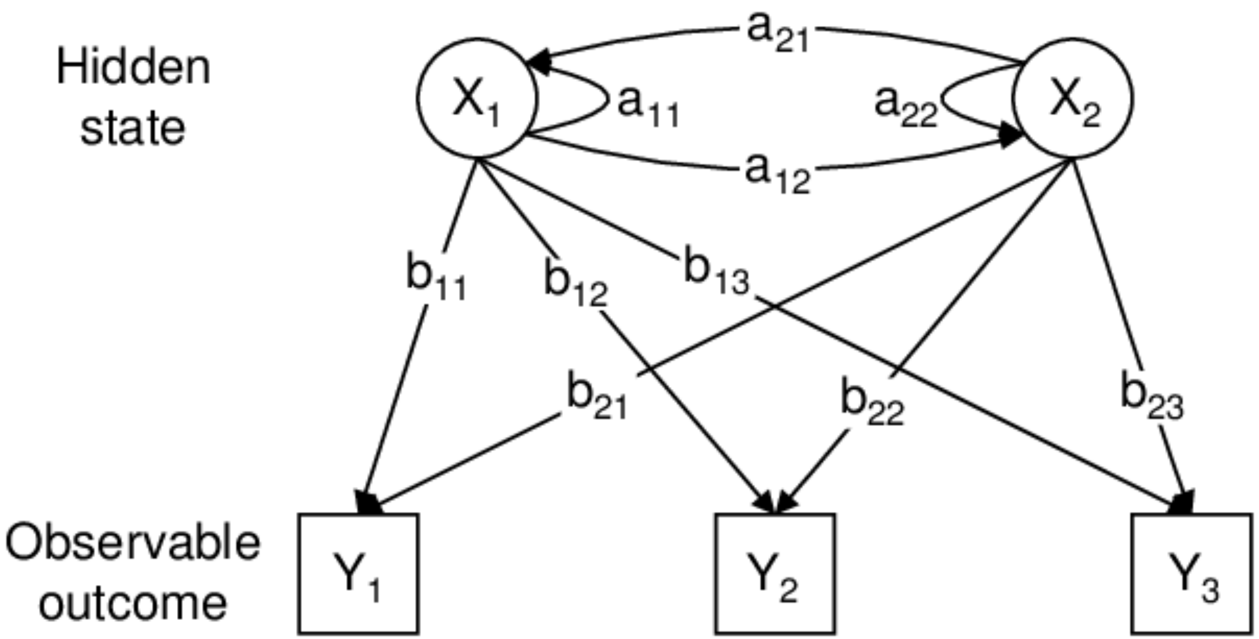
Markovian processes have no memory, but the present state dictates probabilities of future states

Markov chain is a progressive series of “generations” in which the current state is used to estimate the probability of future states

Monte Carlo is a simulation incorporating randomness with weighted probabilities (obtained from the current state)

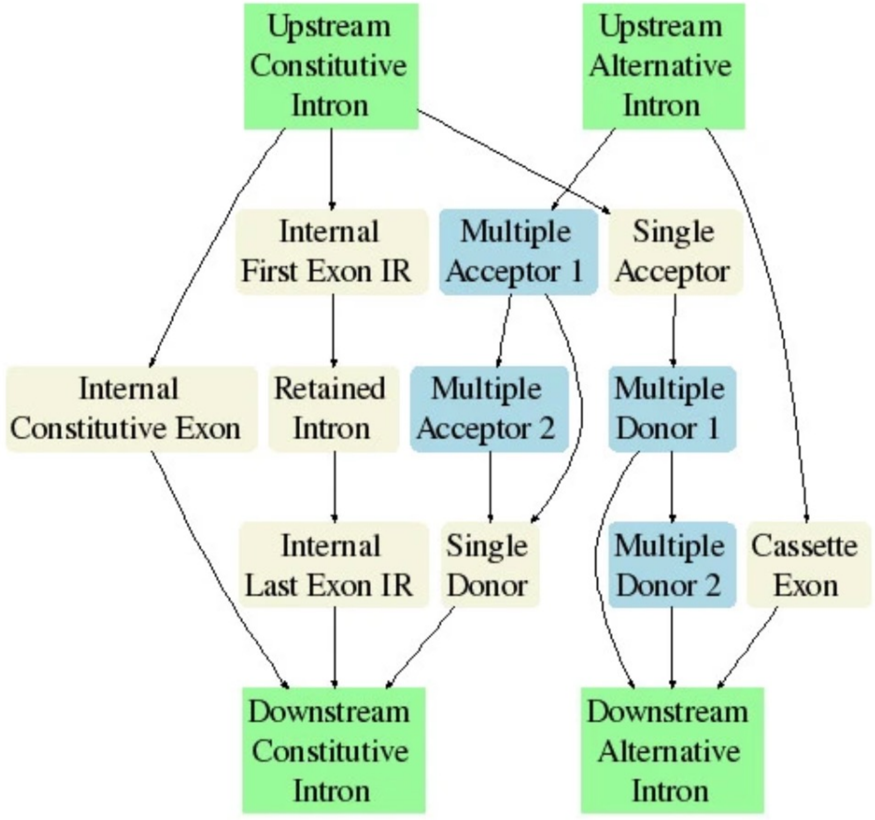
Hidden Markov Models

Markov chain (i.e., no memory) with **hidden states**



Hidden Markov Models

Gene model inference



Upstream Constitutive Intron

Upstream Alternative Intron

Internal First Exon IR

Multiple Acceptor 1

Single Acceptor

Internal Constitutive Exon

Retained Intron

Multiple Acceptor 2

Multiple Donor 1

Internal Last Exon IR

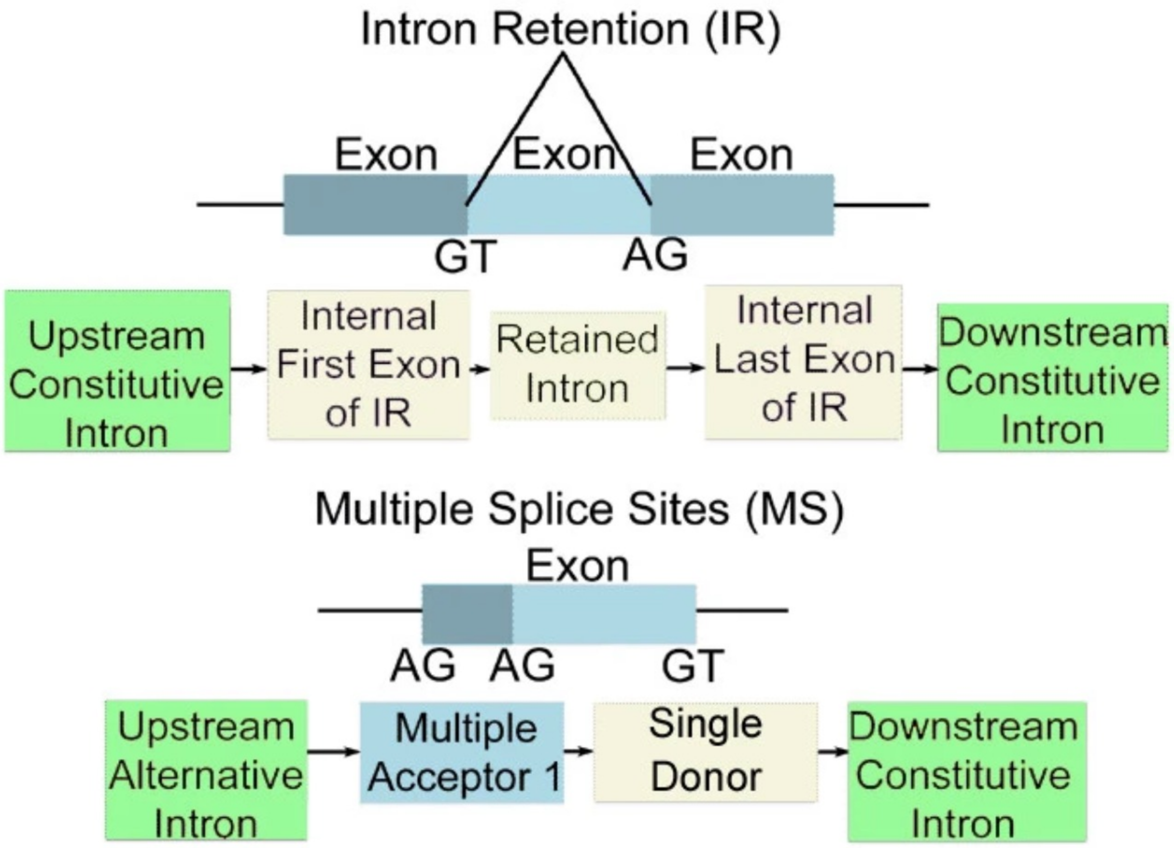
Single Donor

Multiple Donor 2

Cassette Exon

Downstream Constitutive Intron

Downstream Alternative Intron



Intron Retention (IR)

Exon

Exon

Exon

GT

AG

Upstream Constitutive Intron

Internal First Exon of IR

Retained Intron

Internal Last Exon of IR

Downstream Constitutive Intron

Multiple Splice Sites (MS)

Exon

AG

AG

GT

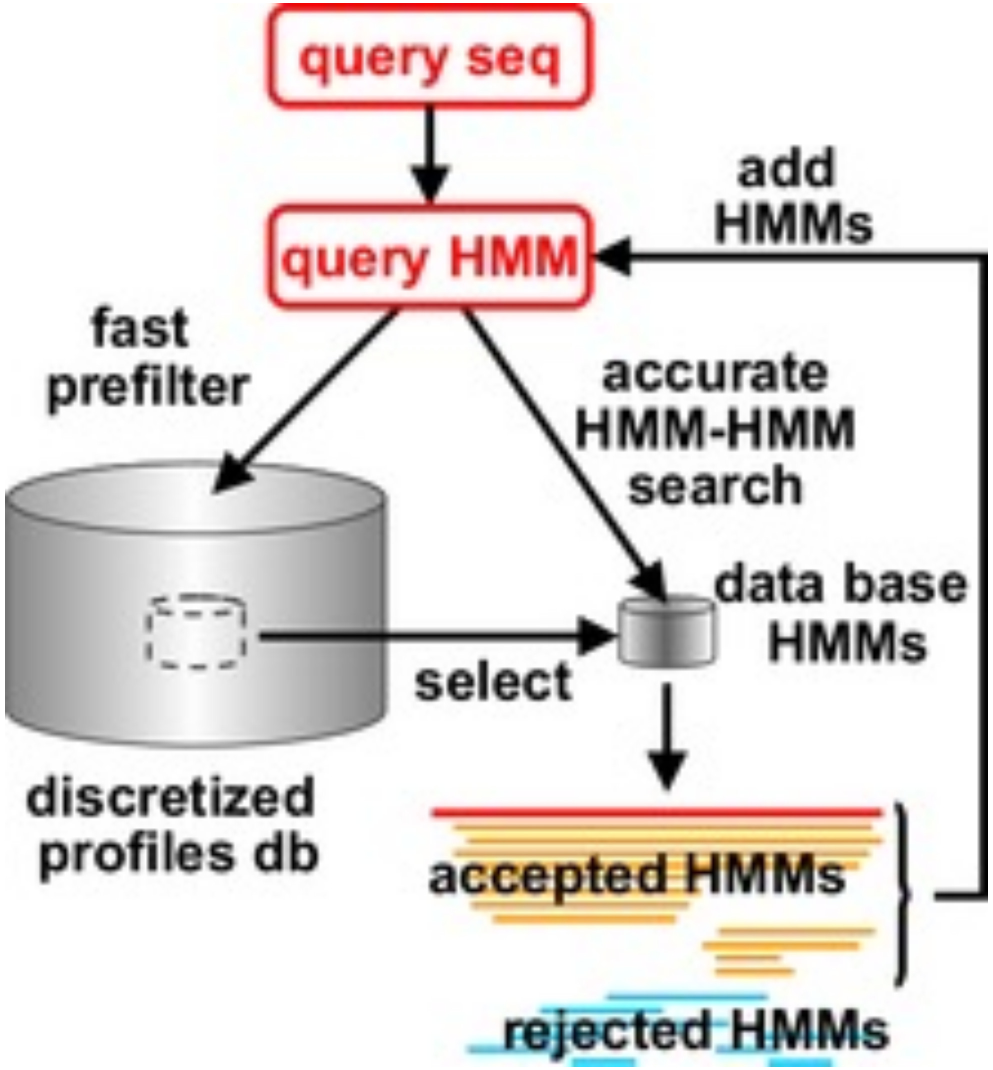
Upstream Alternative Intron

Multiple Acceptor 1

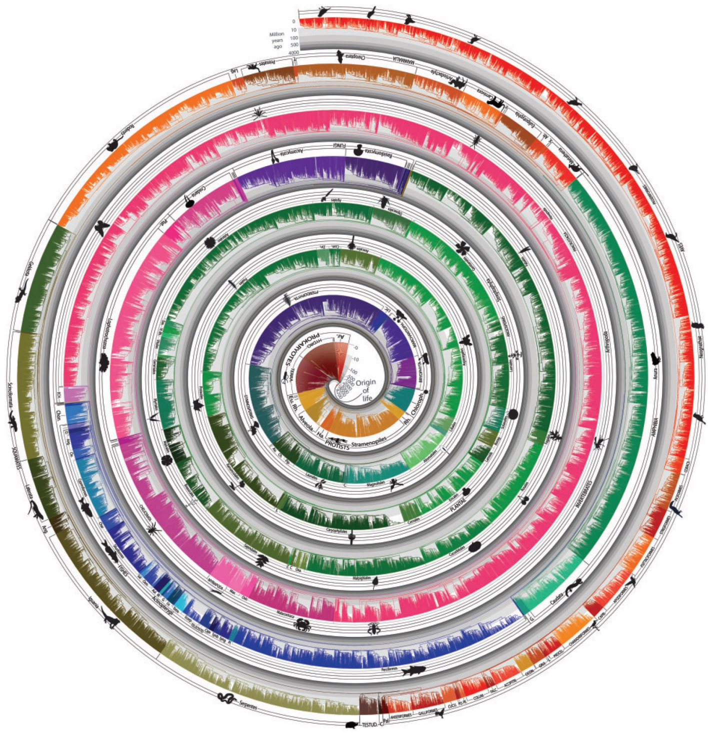
Single Donor

Downstream Constitutive Intron

HHPred

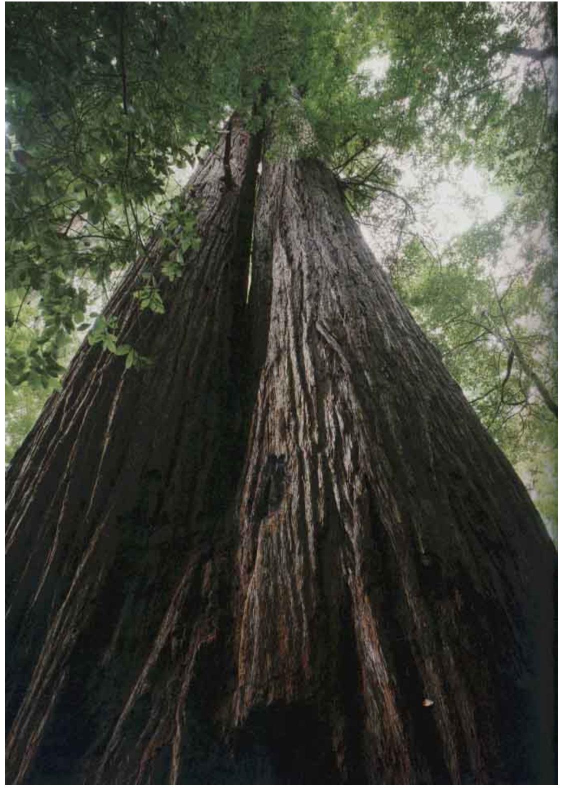

query seq
query HMM
add HMMs
fast prefilter
accurate HMM-HMM search
data base HMMs
select
discretized profiles db
accepted HMMs
rejected HMMs

Supervised Machine Learning – Seek app



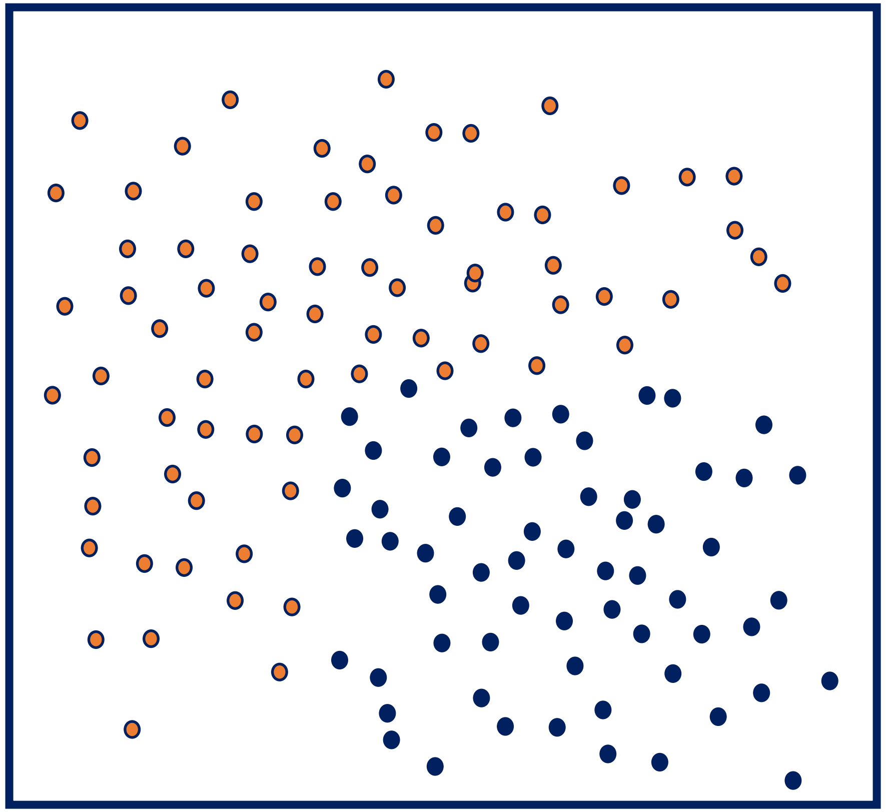
Training dataset (Tree of Life)

On-the-fly
species identification!
Using Decision Tree

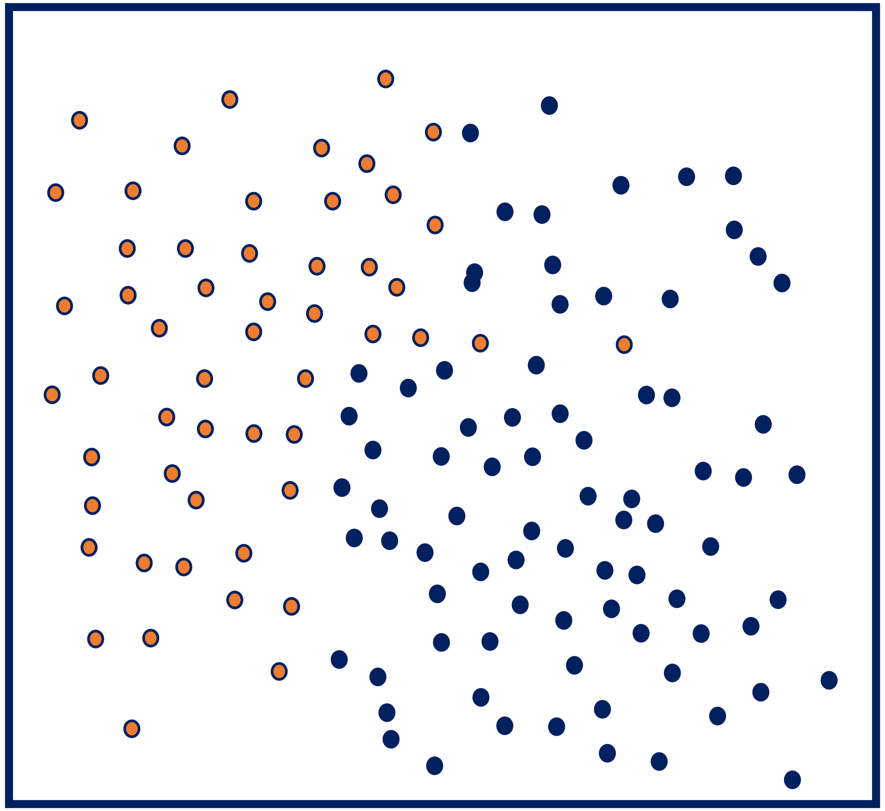


Sequoia sempervirens

Unsupervised Machine Learning

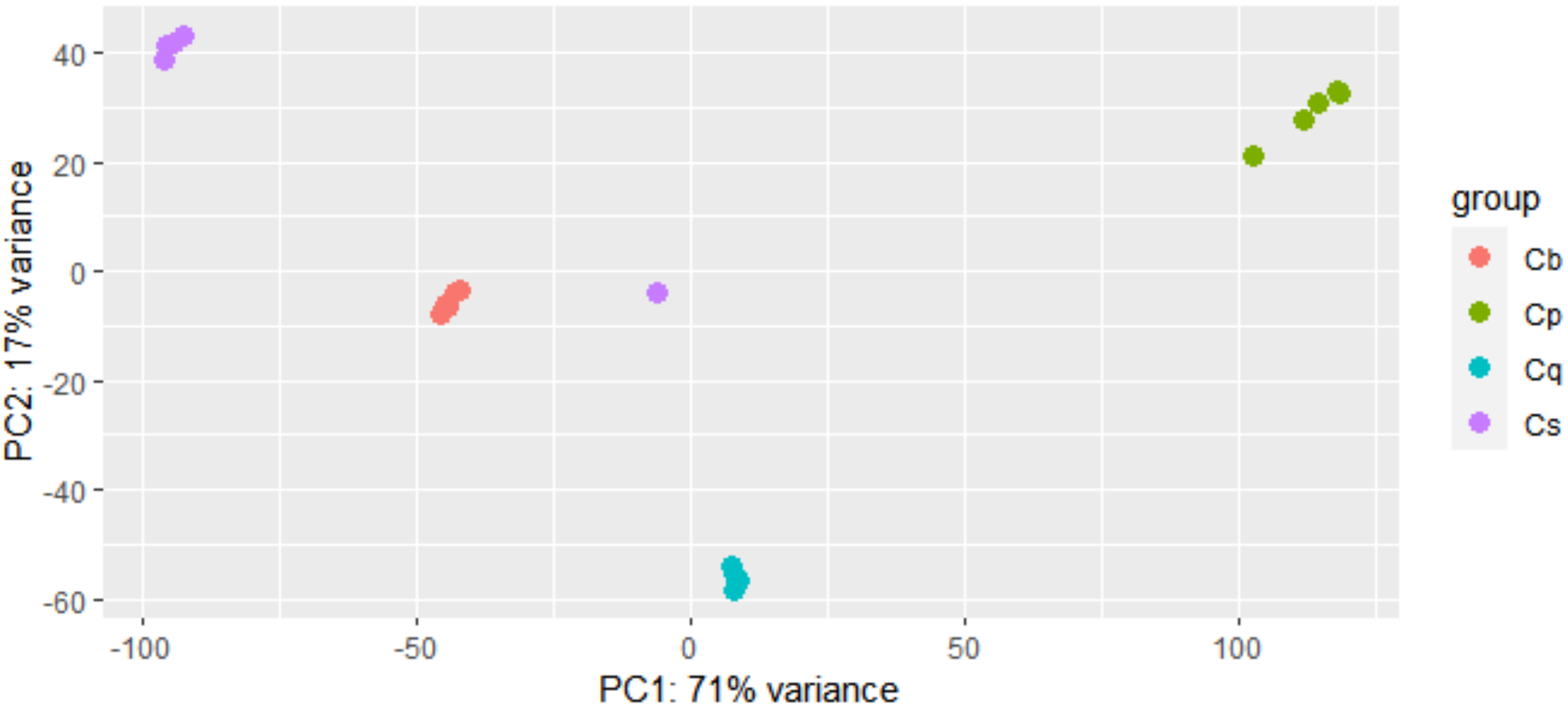


Training dataset



Mimicked dataset

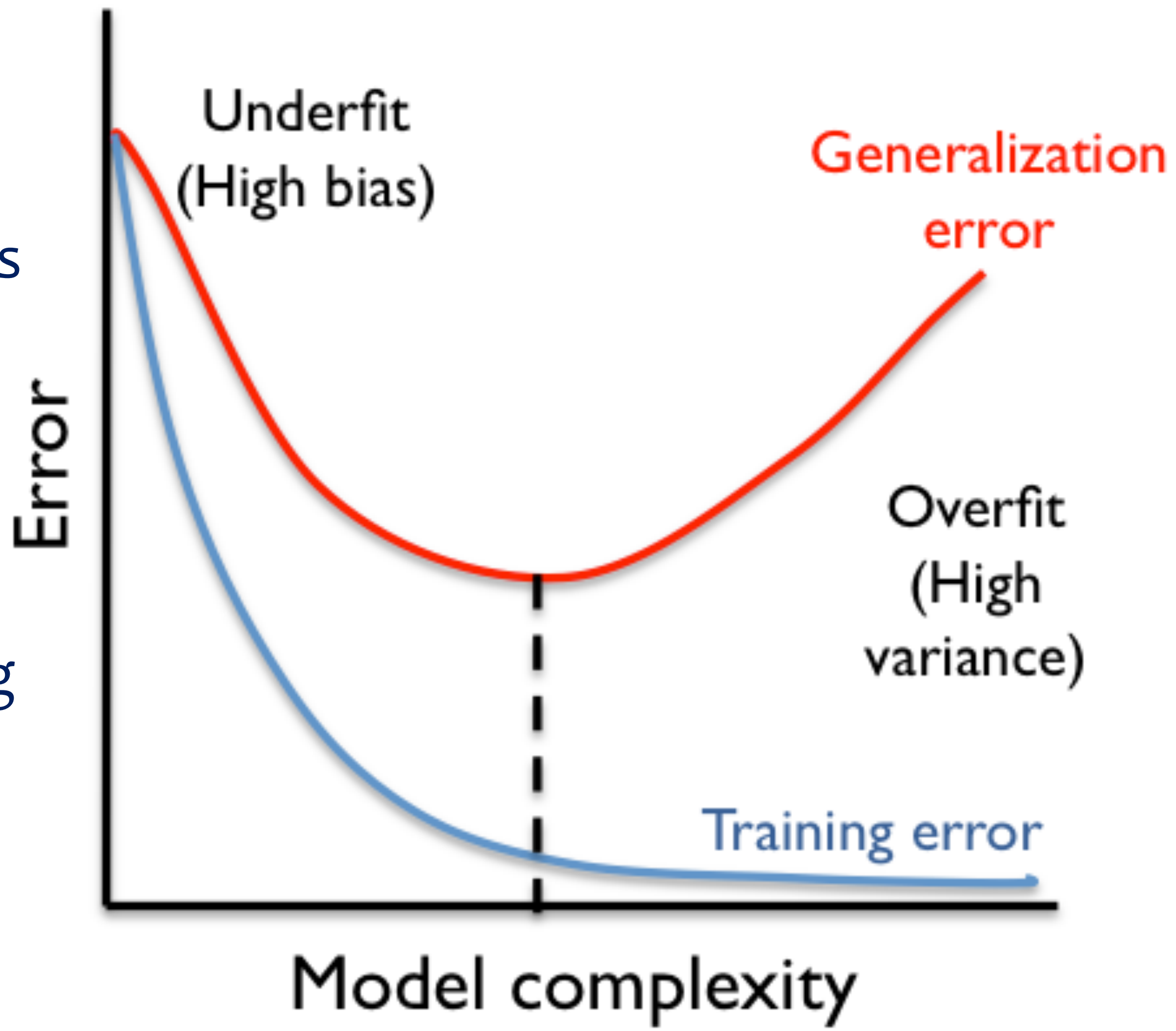
Principle Components Analysis



Trade-offs in training

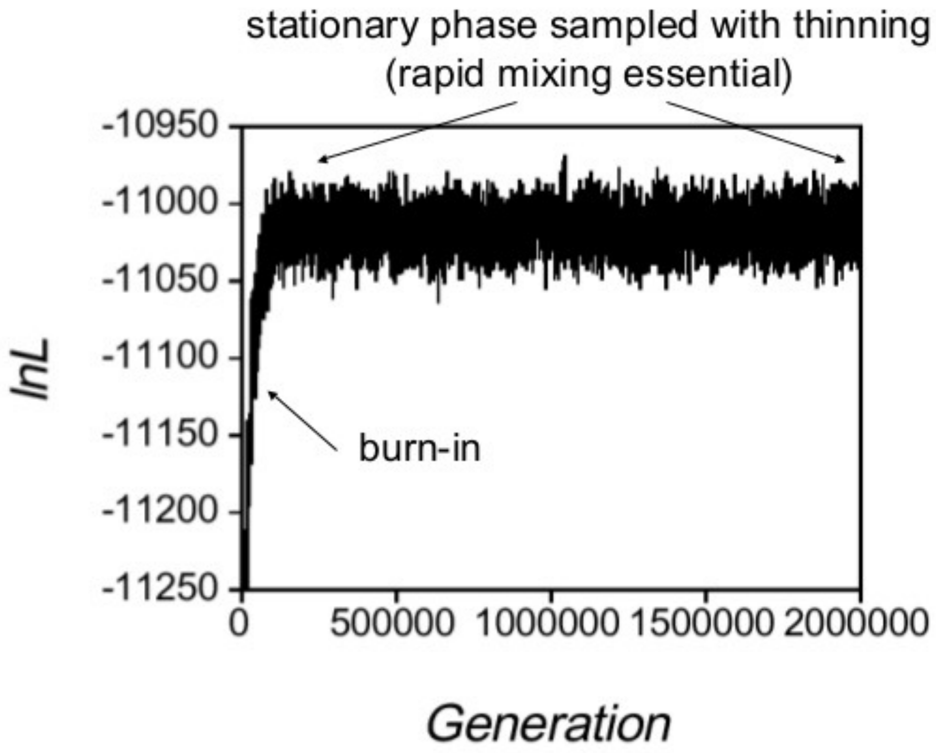
Underfit models (i.e., not enough parameters) have high misclassification rates in training and test datasets

Over-fit model (i.e., too many parameters) will be too keyed in on the training dataset to accurately classify test datasets



Bayesian inference

Uses Markov chain Monte Carlo (MCMC) simulations to estimate posterior probability given a set of prior probabilities and the data



Markov Chain + Monte Carlo simulation

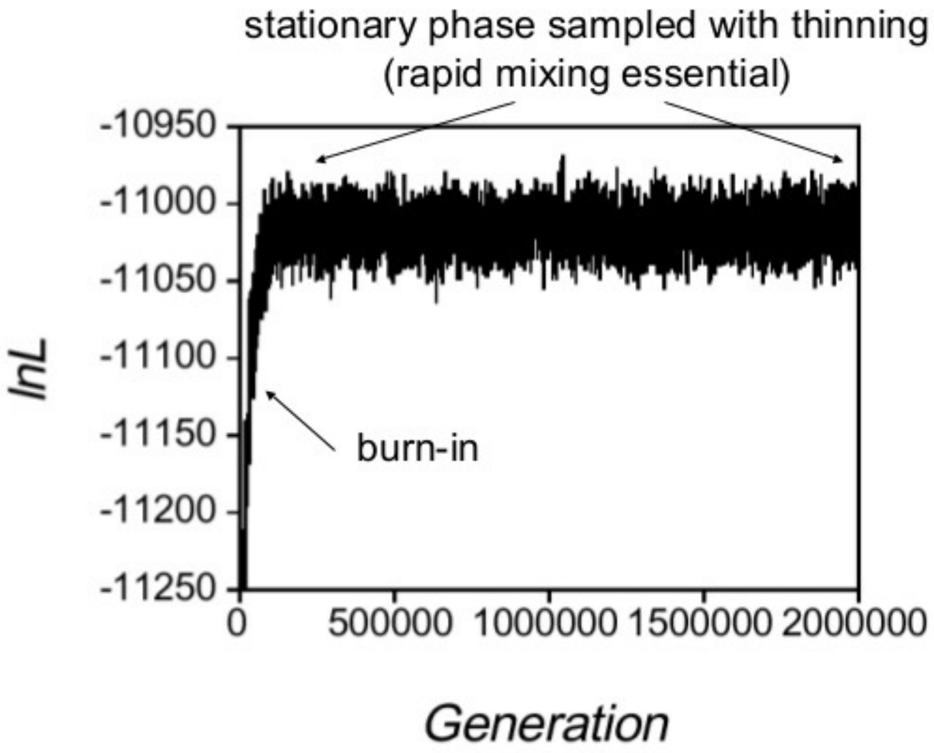
Markovian processes have no memory, but the present state dictates probabilities of future states

Markov chain is a progressive series of “generations” in which the current state is used to estimate the probability of future states

Monte Carlo is a simulation incorporating randomness with weighted probabilities (obtained from the current state)

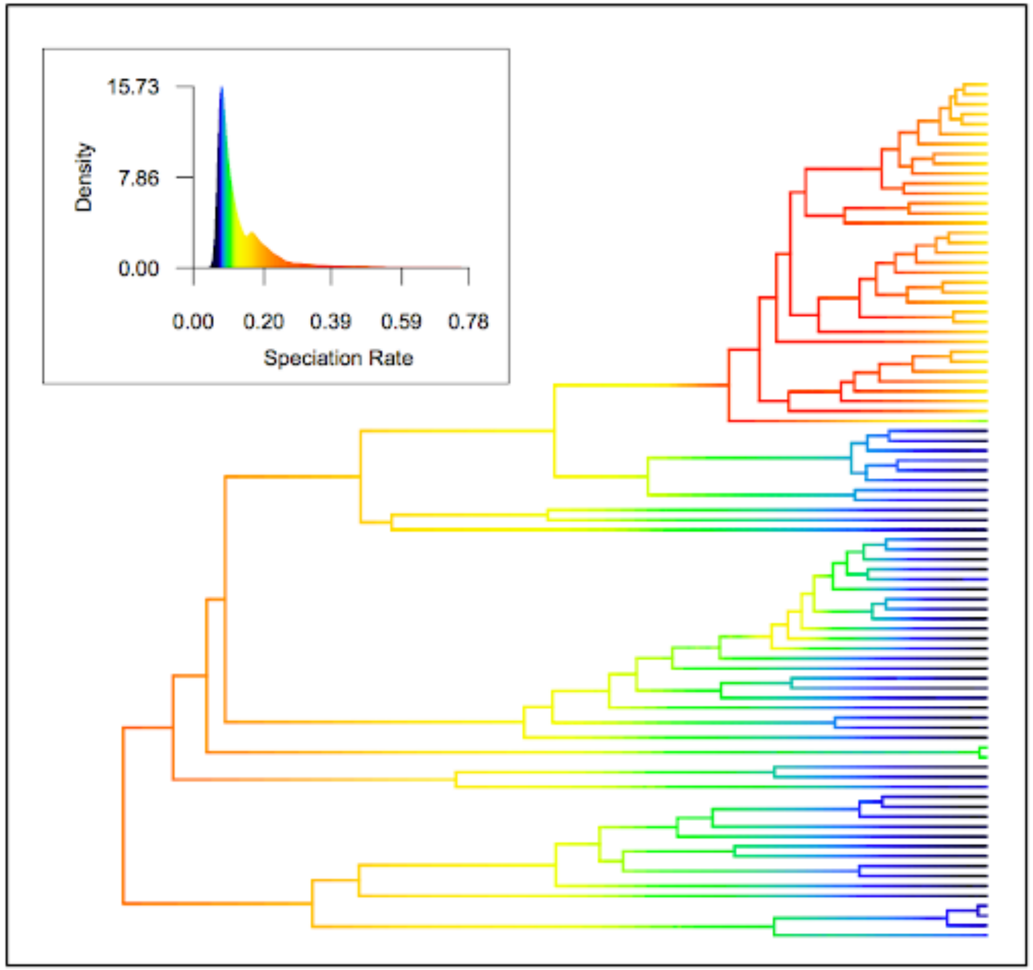
Bayesian inference

Uses Markov chain Monte Carlo (MCMC) simulations to estimate posterior probability given a set of prior probabilities and the data



Bayesian inference – BAMM

Model historical macro-evolutionary trends on trees (e.g., speciation/extinction rate)


Density
Speciation Rate