

# Probabilistic Deterministic Infinite Automata

David Pfau, Nicholas Bartlett, Frank Wood



## Discrete Sequence Modeling

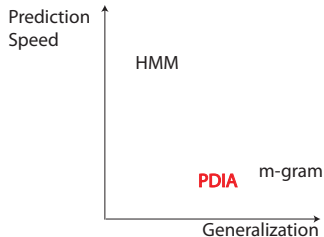
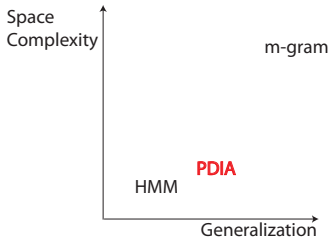
- ▶ 01101010111001011...
- ▶ CGTAACCGATTAC...
- ▶ *Four score and seven...*

## Tasks

- ▶ Conditional Prediction
- ▶ Typicality - Clustering, etc

## Probabilistic Deterministic Infinite Automata (PDIA)

- ▶ Trade off generalization, space complexity, prediction speed
- ▶ Versus Hidden Markov Model (HMM) or  $n^{\text{th}}$  order Markov model (m-gram)

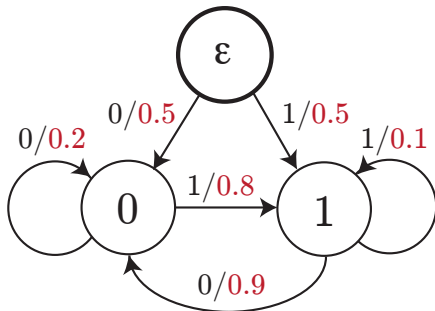


## Probabilistic DFA

- ▶ A PDFA  $\approx$  HMM with one possible path through states given data
- ▶ m-gram  $\subsetneq$  PDFA  $\subsetneq$  HMM

## Bayesian Infinite Automata

- ▶ Define prior over DFA topology
- ▶ Bias towards small DFA that reuse states
- ▶ Let bound on state cardinality  $\rightarrow \infty$



## Results

- ▶ Recovers topology of simple PDFA
- ▶ Generalizes as well as 3rd-order Markov, 1/10th the # of states (natural language)
- ▶ Predicts better with averaging from PDFA of different topology than with single PDFA

## Recap

- ▶ Nonparametric Bayesian learning of simple models for discrete sequences
- ▶ Faster forward prediction than for HMM
- ▶ Favorable tradeoff between model size and generalization