

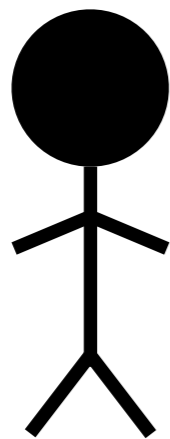
Iterated Learning of Multiple Languages from Multiple Teachers



David Burkett and Tom Griffiths
University of California, Berkeley



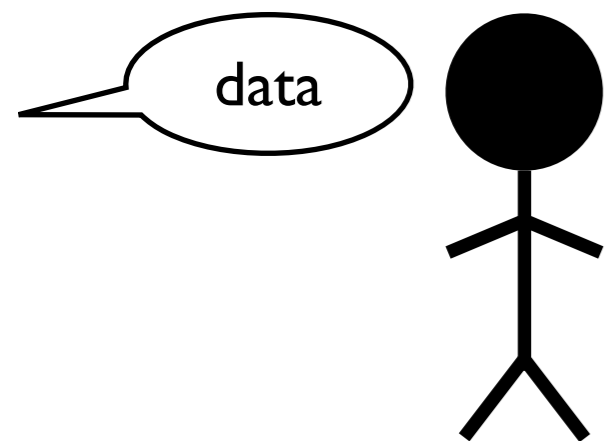
Iterated Learning



(Kirby, 2001)

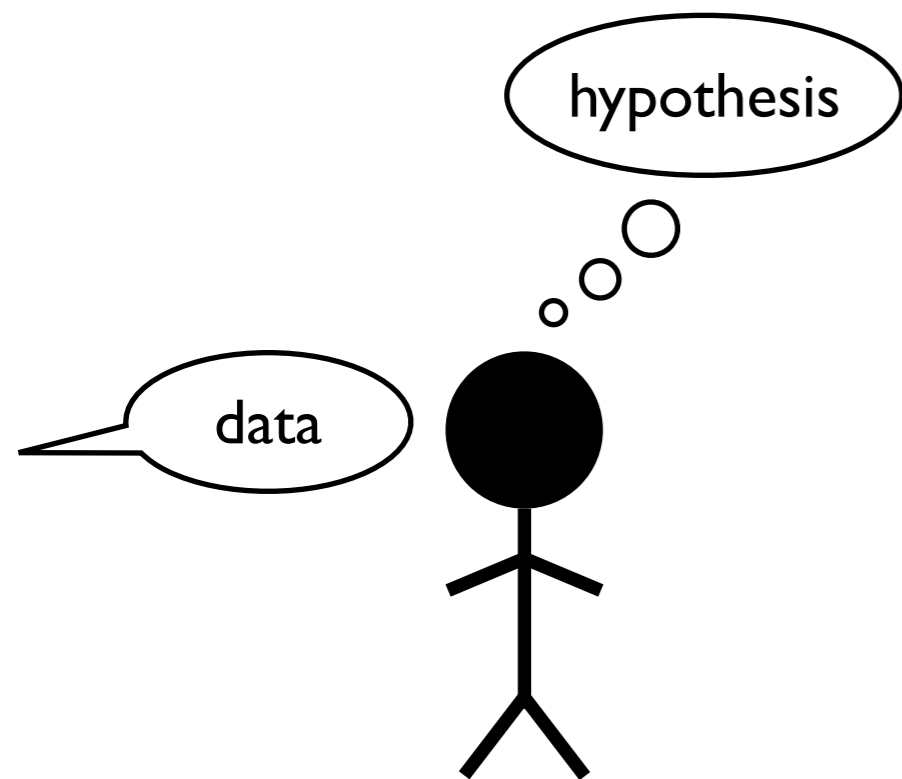


Iterated Learning



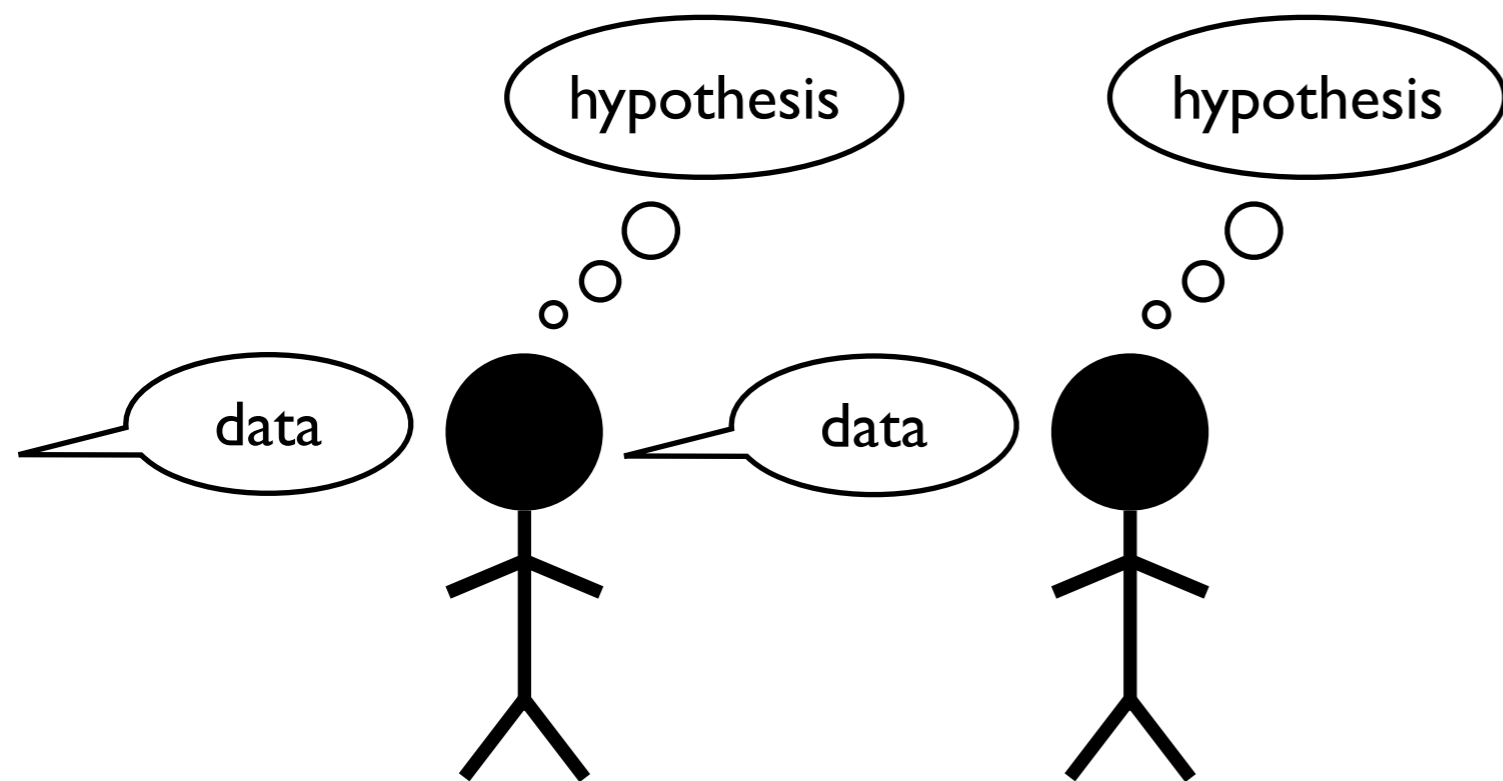
(Kirby, 2001)

Iterated Learning



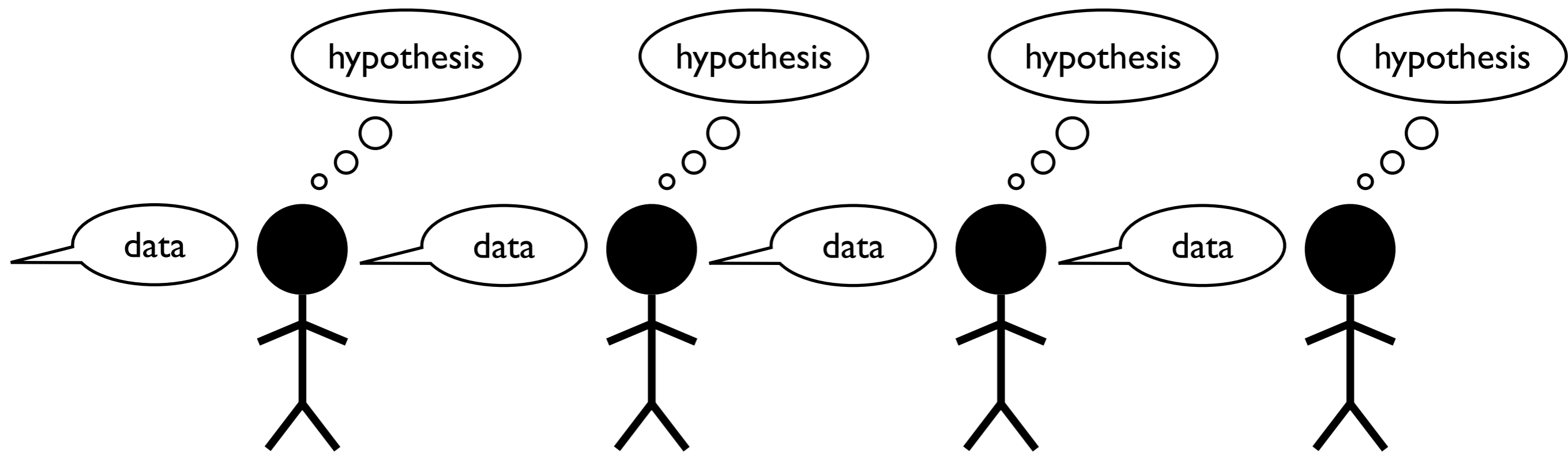
(Kirby, 2001)

Iterated Learning



(Kirby, 2001)

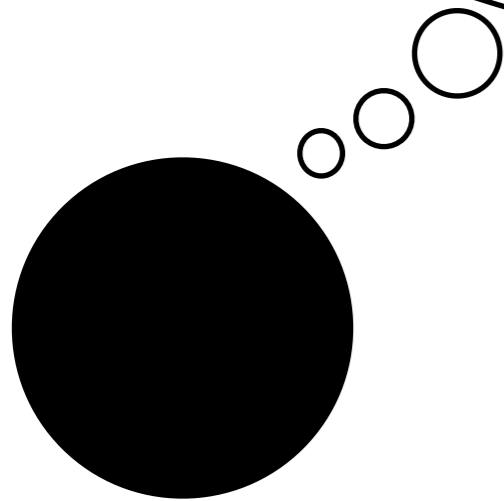
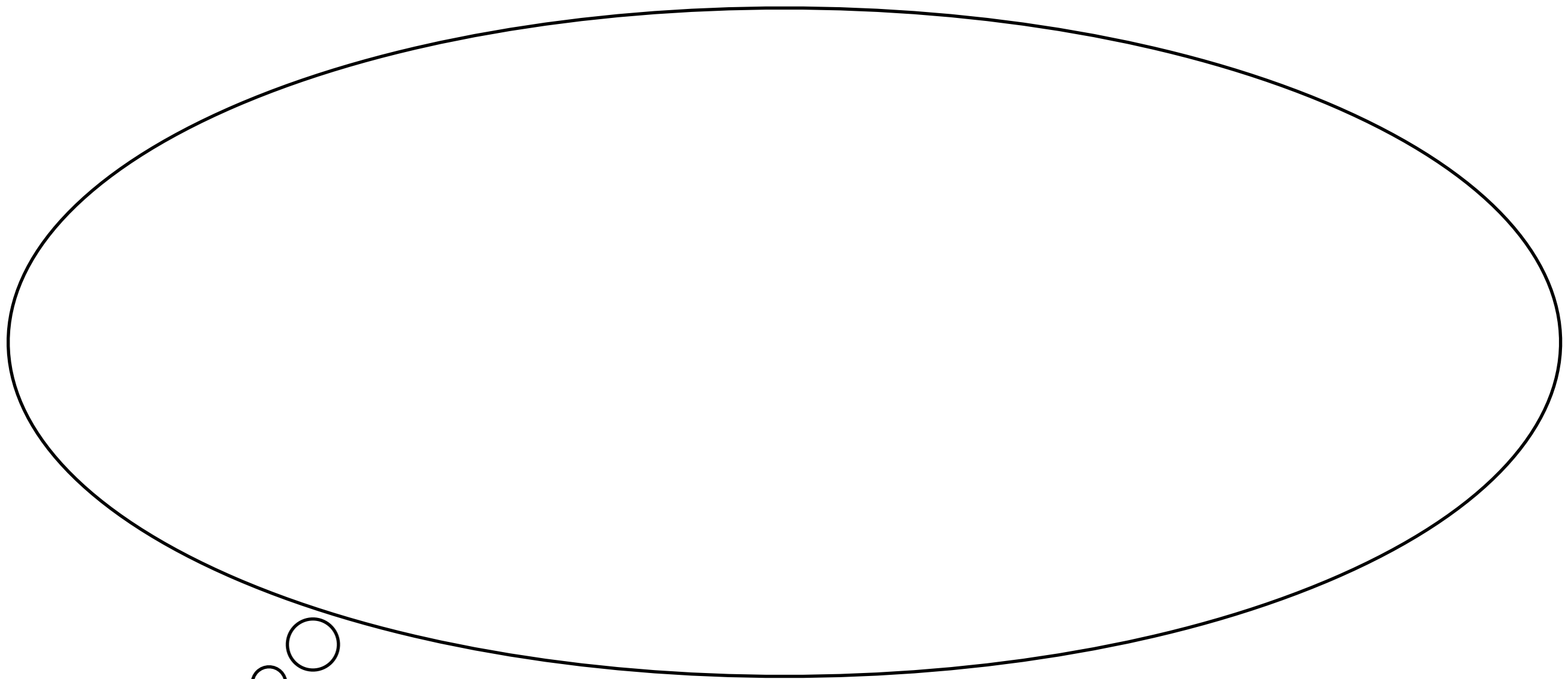
Iterated Learning



(Kirby, 2001)



Language Learning



(Kirby, 2001)

Language Learning

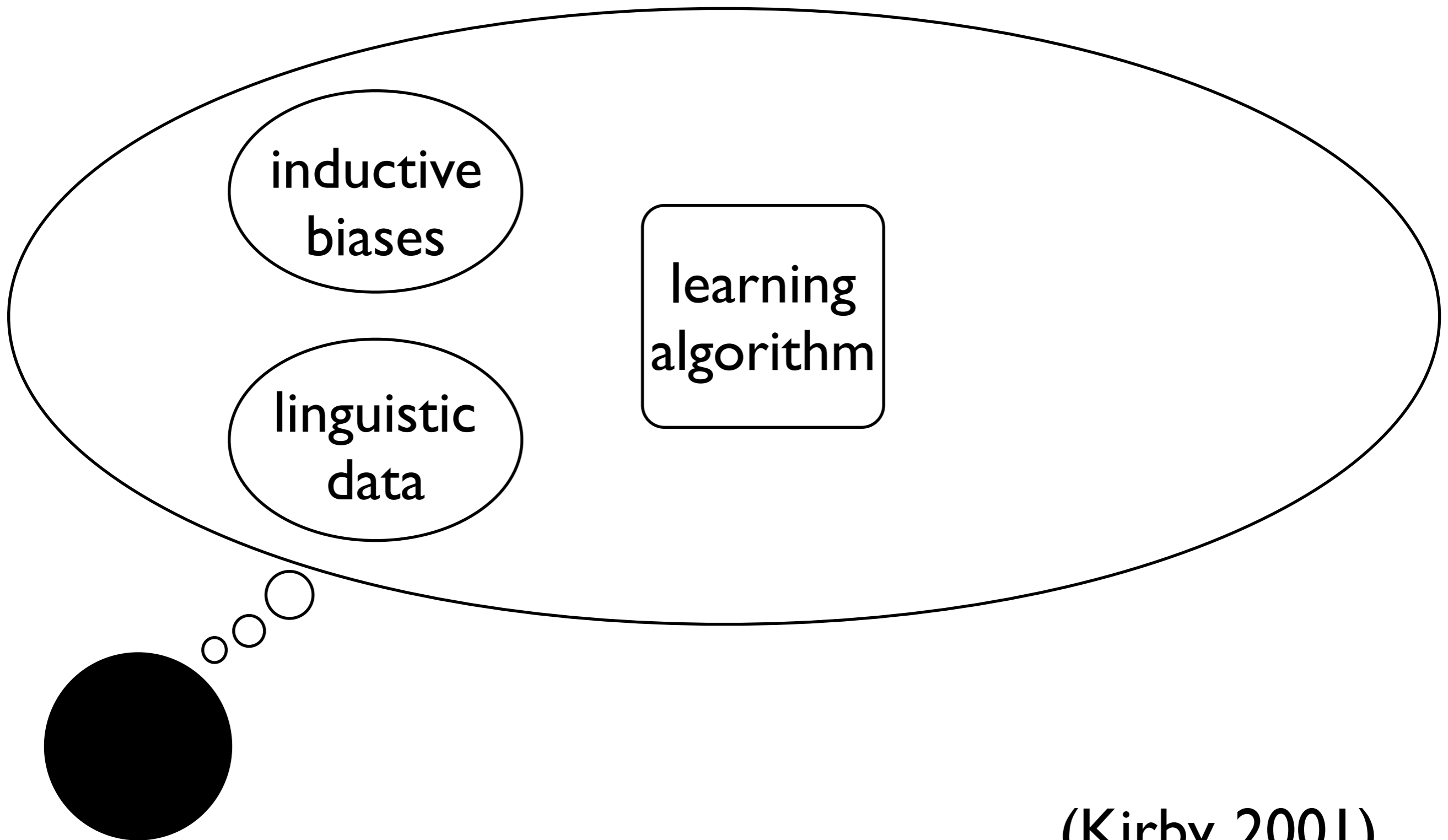


inductive
biases

linguistic
data

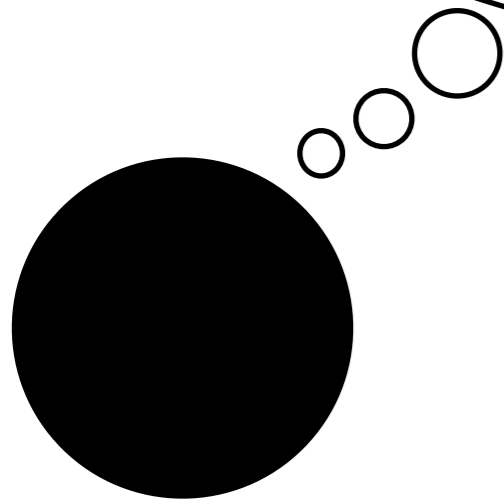
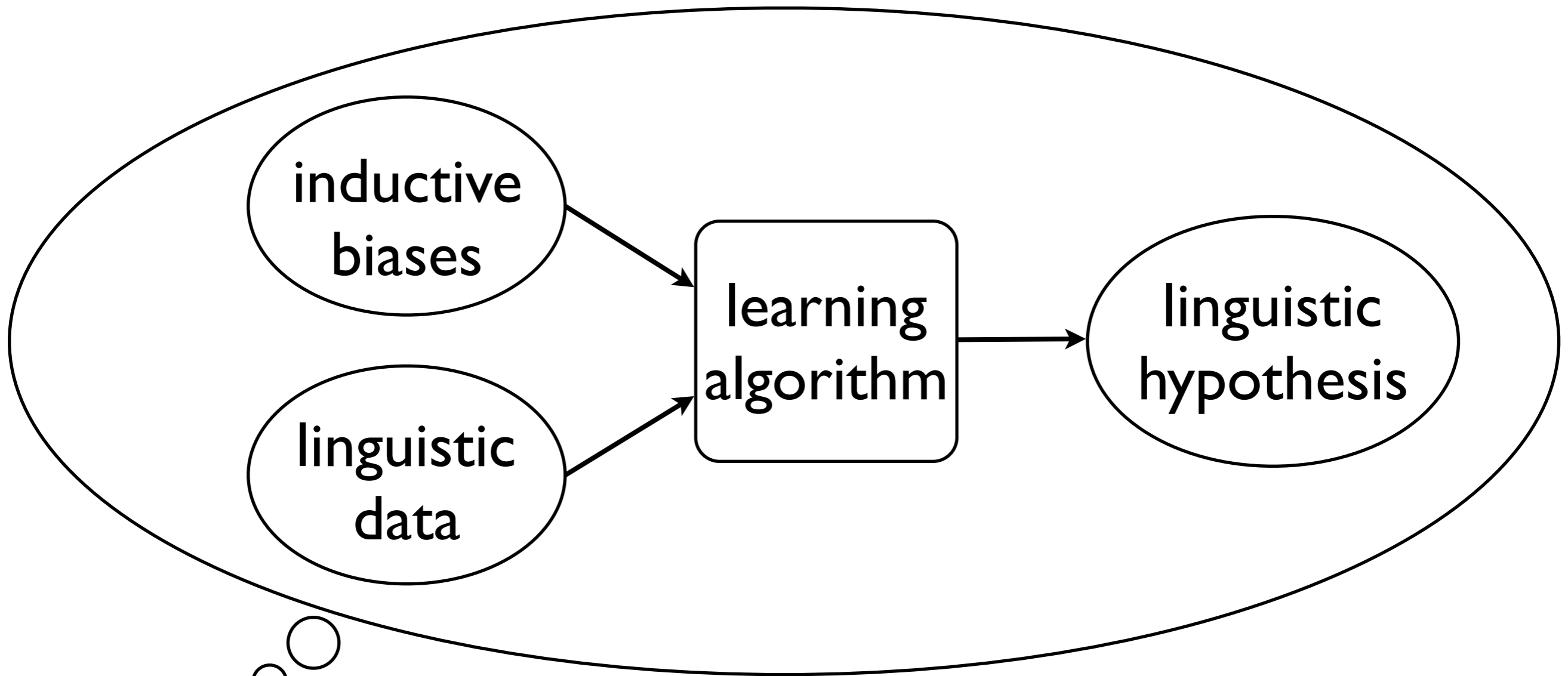
(Kirby, 2001)

Language Learning



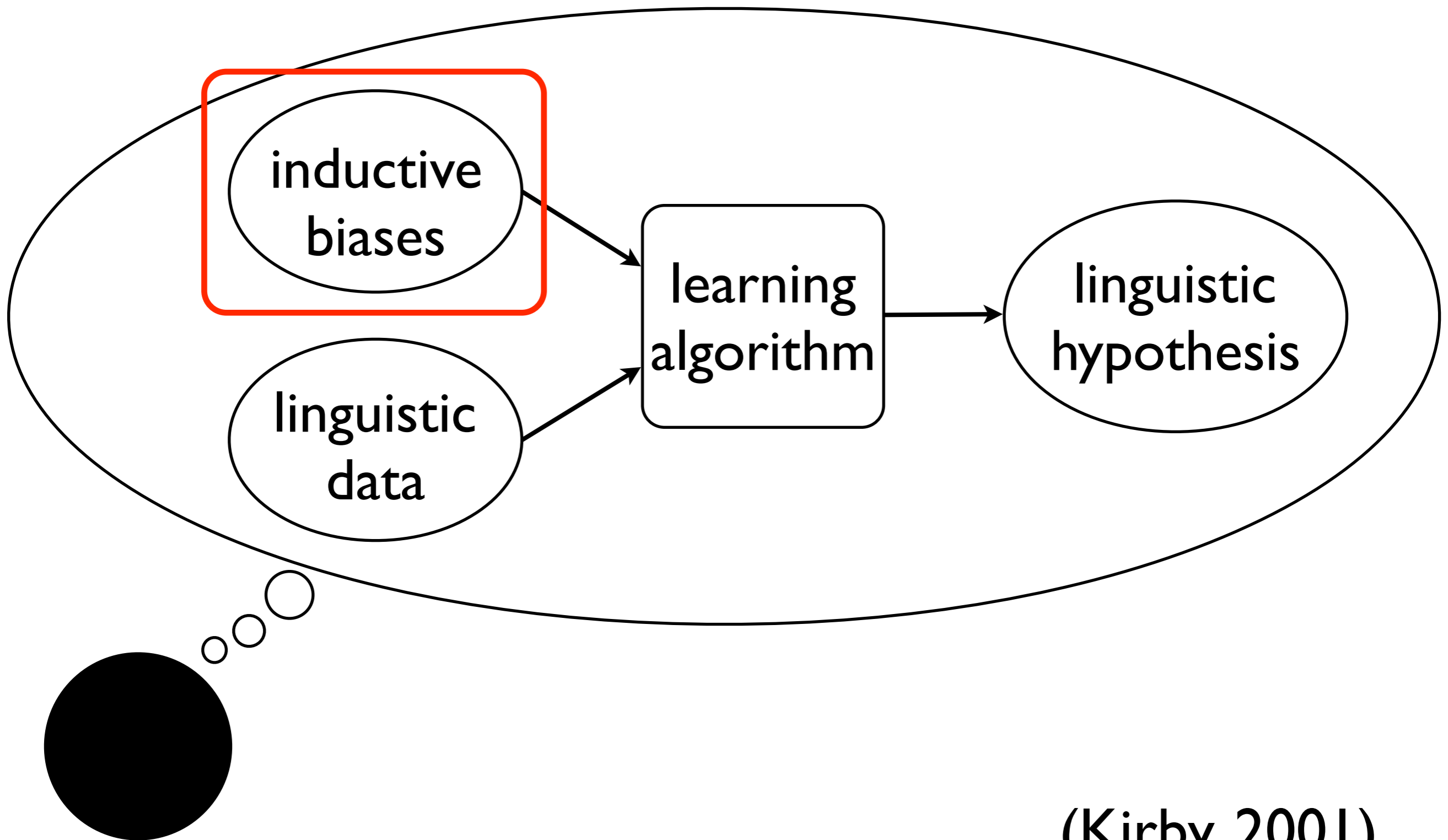
(Kirby, 2001)

Language Learning



(Kirby, 2001)

Language Learning



(Kirby, 2001)



Inductive Biases



Inductive Biases

- Iterated learning and language evolution:

How do learners' inductive biases affect the languages used by a population?



Inductive Biases

- Iterated learning and language evolution:

How do learners' inductive biases affect the languages used by a population?

- Current work:

How does social structure influence the effect of inductive biases?



Outline



Outline

- Learning from a single teacher



Outline

- Learning from a single teacher
- Learning from multiple teachers



Outline

- Learning from a single teacher
- Learning from multiple teachers
- Learning multiple languages



Outline

- Learning from a single teacher
- Learning from multiple teachers
- Learning multiple languages



Bayesian Learners



Bayesian Learners

linguistic
hypothesis

h

Bayesian Learners

linguistic
hypothesis

h

inductive
biases

$p(h)$

Bayesian Learners

linguistic
hypothesis

h

inductive
biases

$p(h)$

linguistic
data

d

Bayesian Learners

linguistic
hypothesis

h

inductive
biases

$p(h)$

linguistic
data

d

data
likelihood

$p(d|h)$

Bayesian Learners

linguistic
hypothesis

h

inductive
biases

$p(h)$

linguistic
data

d

data
likelihood

$p(d|h)$

learning
algorithm

$$p(h|d) \propto p(h)p(d|h)$$



Bayesian Learners

Example



Bayesian Learners

Example

h	$p(h)$
rhotic	3/4
non-rhotic	1/4

Bayesian Learners

Example

h	$p(h)$
rhotic	3/4
non-rhotic	1/4

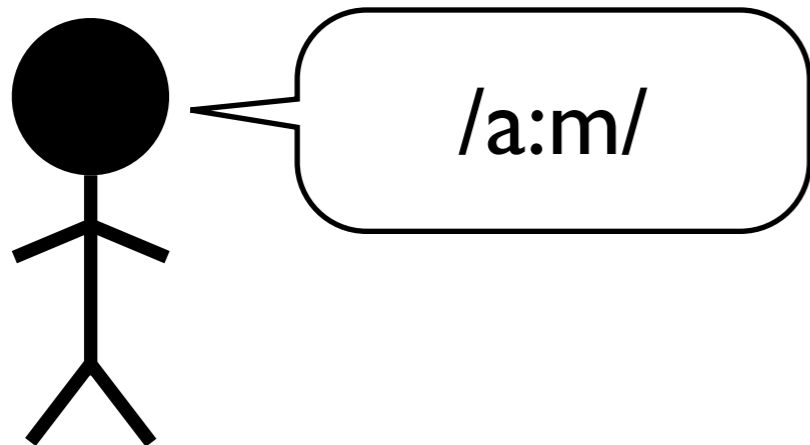
h	d	$p(d h)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5

Bayesian Learners

Example

h	$p(h)$
rhotic	3/4
non-rhotic	1/4

h	d	$p(d h)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5

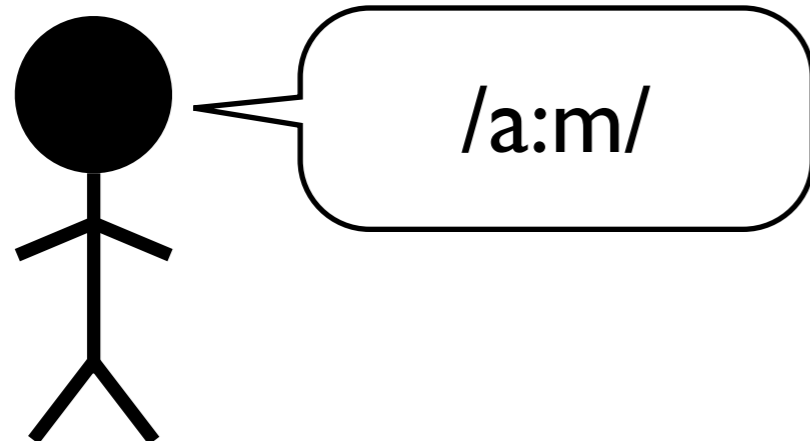


Bayesian Learners

Example

h	$p(h)$
rhotic	3/4
non-rhotic	1/4

h	d	$p(d h)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5



h	$p(h d)$
rhotic	5/13
non-rhotic	8/13



Bayesian Learners

Example

h	$p(h d)$
rhotic	5/13
non-rhotic	8/13

Bayesian Learners

Example

h	$p(h d)$
rhotic	5/13
non-rhotic	8/13



Bayesian Learners

Example

h	$p(h d)$
rhotic	5/13
non-rhotic	8/13



Card

Choice

Bayesian Learners

Example

h	$p(h d)$
rhotic	5/13
non-rhotic	8/13



Card

Choice

Ace through 5

rhotic

Bayesian Learners

Example

h	$p(h d)$
rhotic	5/13
non-rhotic	8/13



Card

Choice

Ace through 5

rhotic

6 through King

non-rhotic



Chains of Learners

(Griffiths & Kalish, 2007)



Chains of Learners

d_0

(Griffiths & Kalish, 2007)



Chains of Learners

$$d_0 \xrightarrow{p(h|d)} h_1$$

(Griffiths & Kalish, 2007)



Chains of Learners

$$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1$$

(Griffiths & Kalish, 2007)



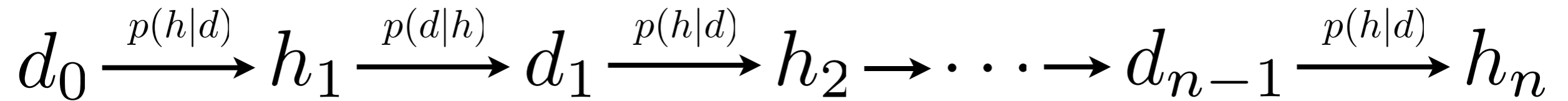
Chains of Learners

$$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1 \xrightarrow{p(h|d)} h_2$$

(Griffiths & Kalish, 2007)

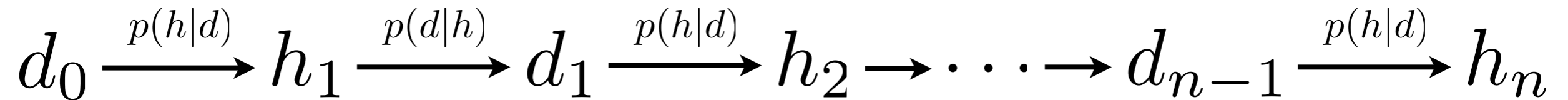


Chains of Learners



(Griffiths & Kalish, 2007)

Chains of Learners

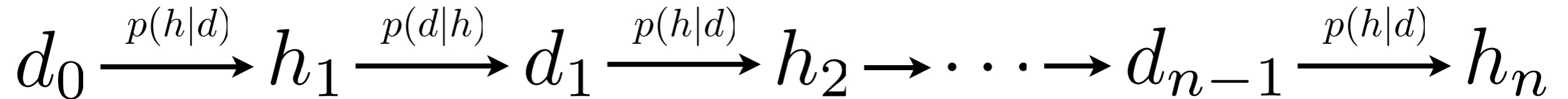


- You can compute $p_n(h)$, the probability that $h_n = h$

(Griffiths & Kalish, 2007)



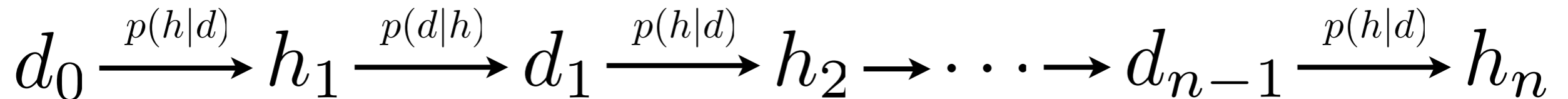
Chains of Learners



- You can compute $p_n(h)$, the probability that $h_n = h$
- For a single chain of learners, the probability distribution over languages converges to the prior.

(Griffiths & Kalish, 2007)

Chains of Learners



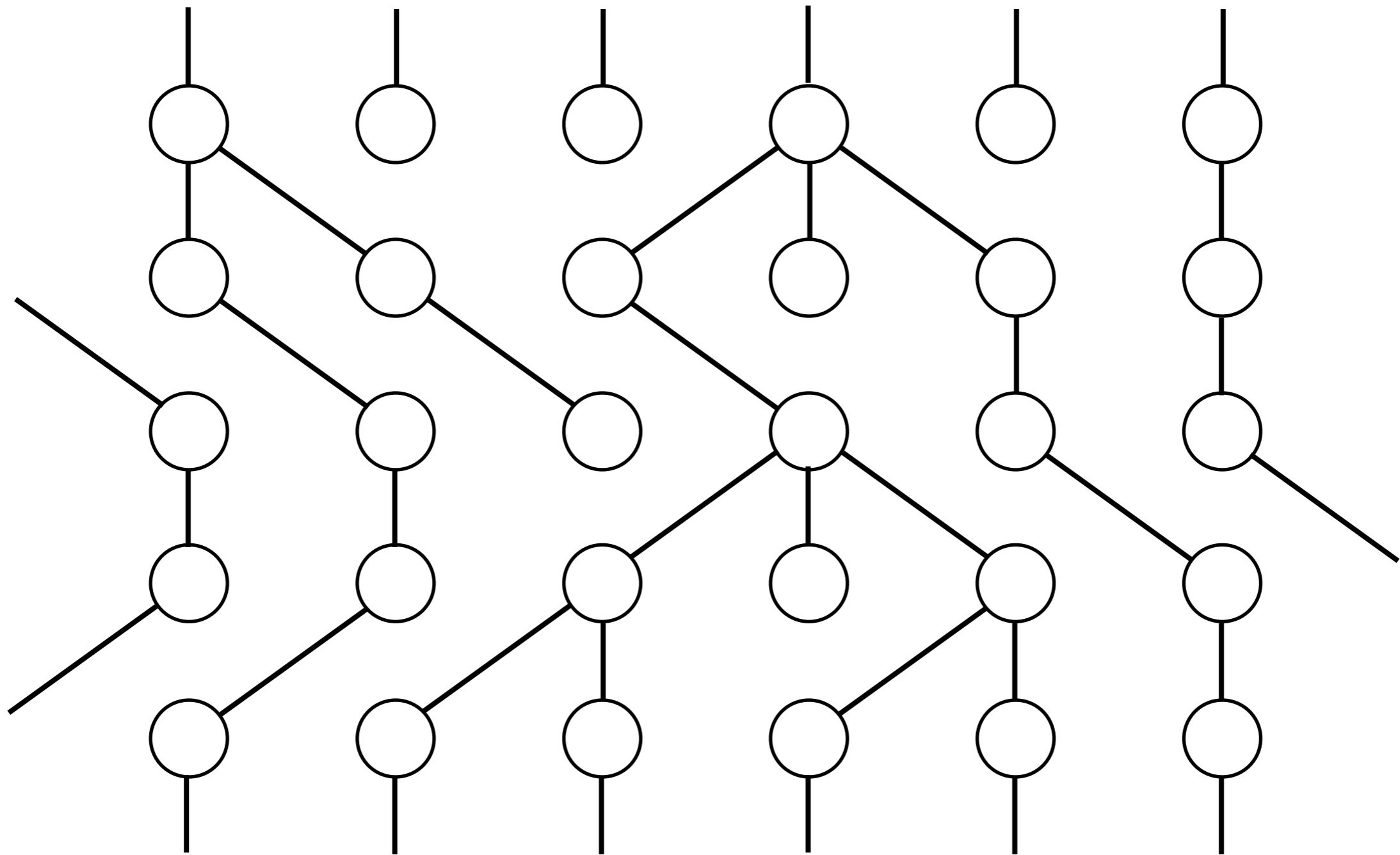
- You can compute $p_n(h)$, the probability that $h_n = h$
- For a single chain of learners, the probability distribution over languages converges to the prior.

$$\lim_{n \rightarrow \infty} p_n(h) = p(h)$$

(Griffiths & Kalish, 2007)

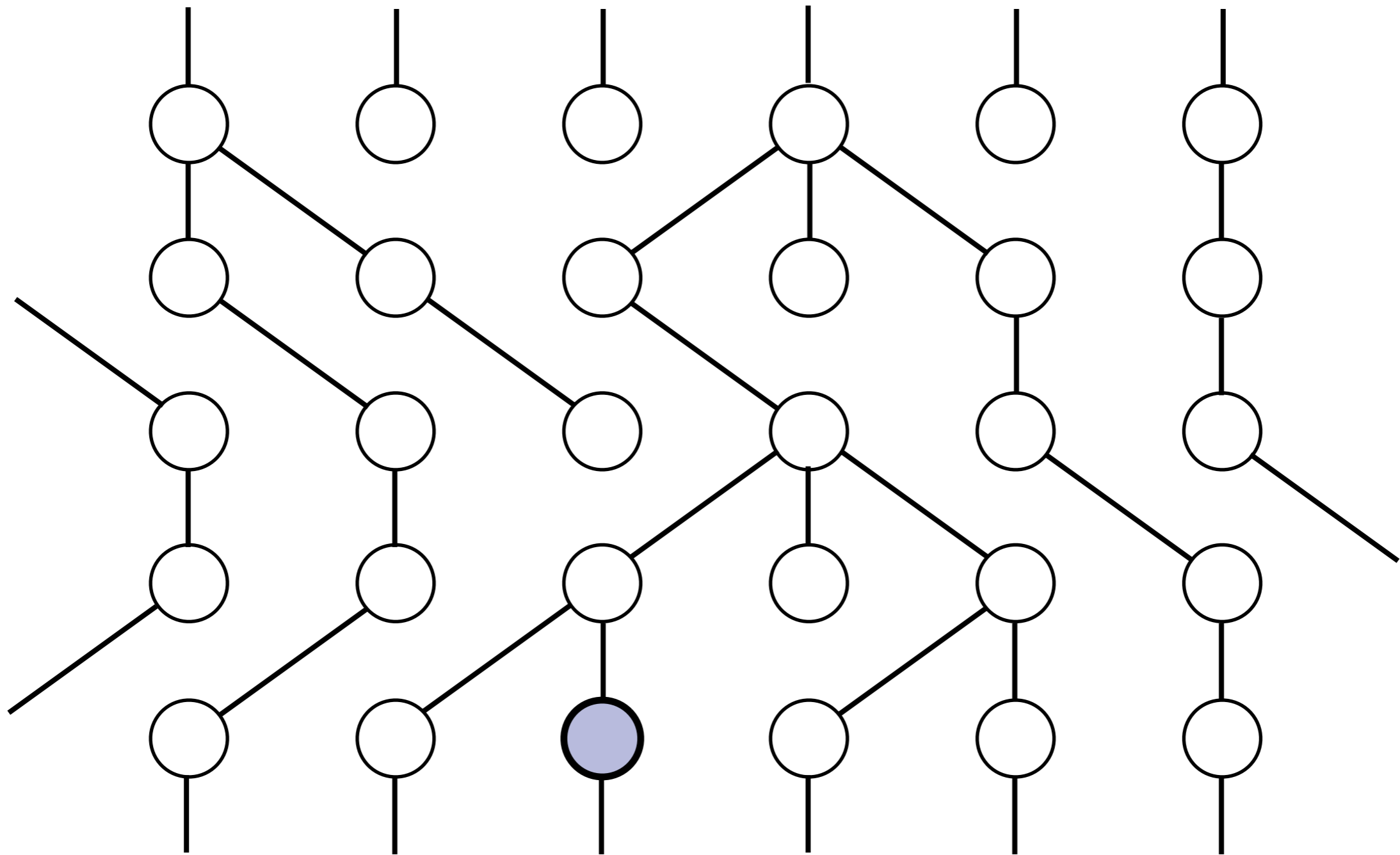


Populations of Learners



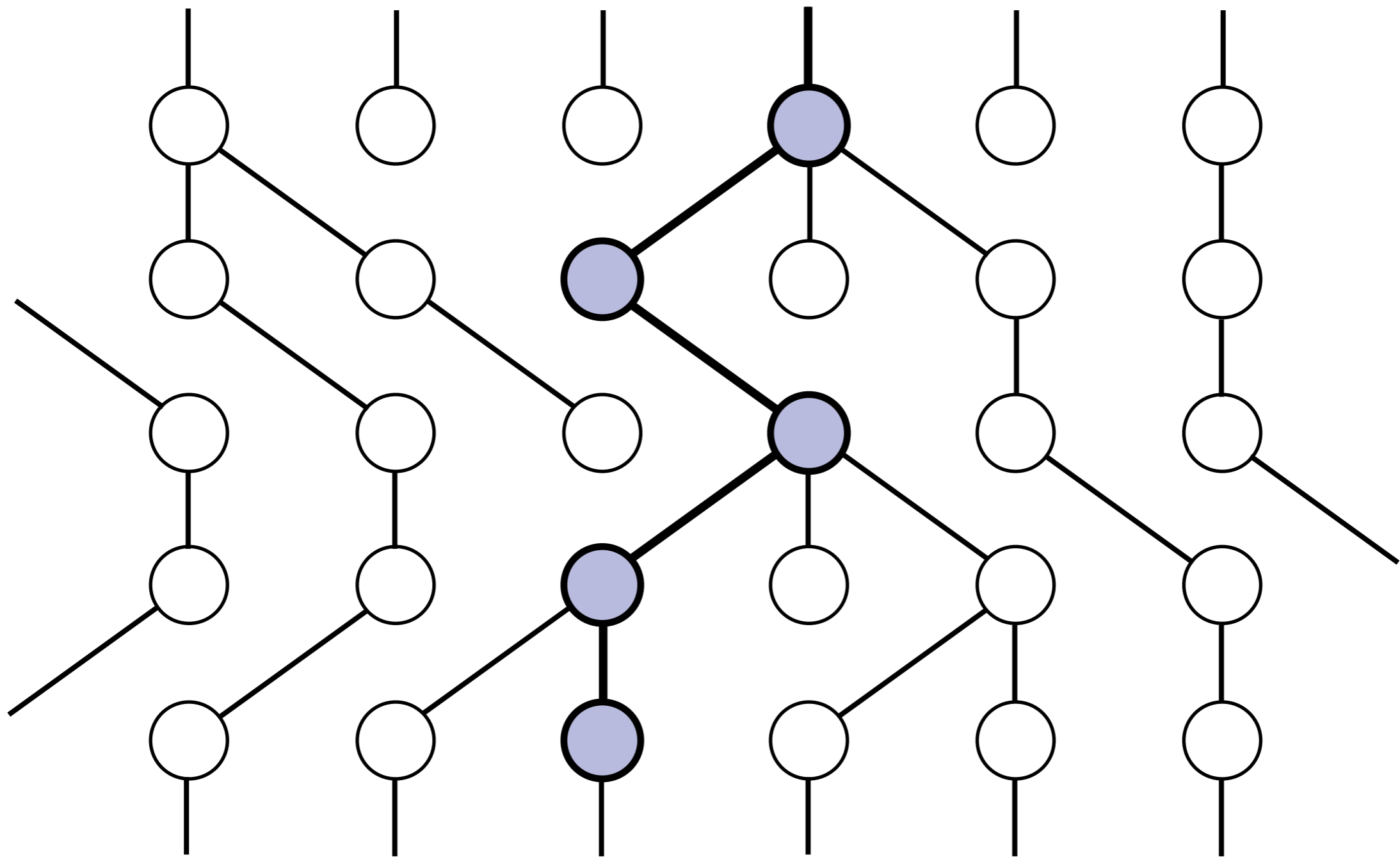


Populations of Learners





Populations of Learners





Outline

- Learning from a single teacher
- Learning from multiple teachers
- Learning multiple languages



Single Teacher



Single Teacher

- The single teacher assumption has been met with some amount of criticism.



Single Teacher

- The single teacher assumption has been met with some amount of criticism.
- *“Models...assuming...a single source of learning input and then iterating that single learner over multiple generations do not embrace the full Darwinian variational picture.”*
(Niyogi & Berwick, 2009)



Single Teacher

- The single teacher assumption has been met with some amount of criticism.
- *“Models...assuming...a single source of learning input and then iterating that single learner over multiple generations do not embrace the full Darwinian variational picture.”*
(Niyogi & Berwick, 2009)
- However, iterated learning is a general framework.



Single Teacher

- The single teacher assumption has been met with some amount of criticism.
- *“Models...assuming...a single source of learning input and then iterating that single learner over multiple generations do not embrace the full Darwinian variational picture.”*
(Niyogi & Berwick, 2009)
- However, iterated learning is a general framework.
- Next step is to analyze the effects of multiple teachers with Bayesian learners.



Multiple Teachers

(Smith, 2009)



Multiple Teachers

- Learners get data from multiple teachers, who may speak different languages.

(Smith, 2009)



Multiple Teachers

- Learners get data from multiple teachers, who may speak different languages.
- Formally:

(Smith, 2009)



Multiple Teachers

- Learners get data from multiple teachers, who may speak different languages.
- Formally:
 - Data are sets of words $d = \{w_0, w_1, \dots, w_b\}$

(Smith, 2009)



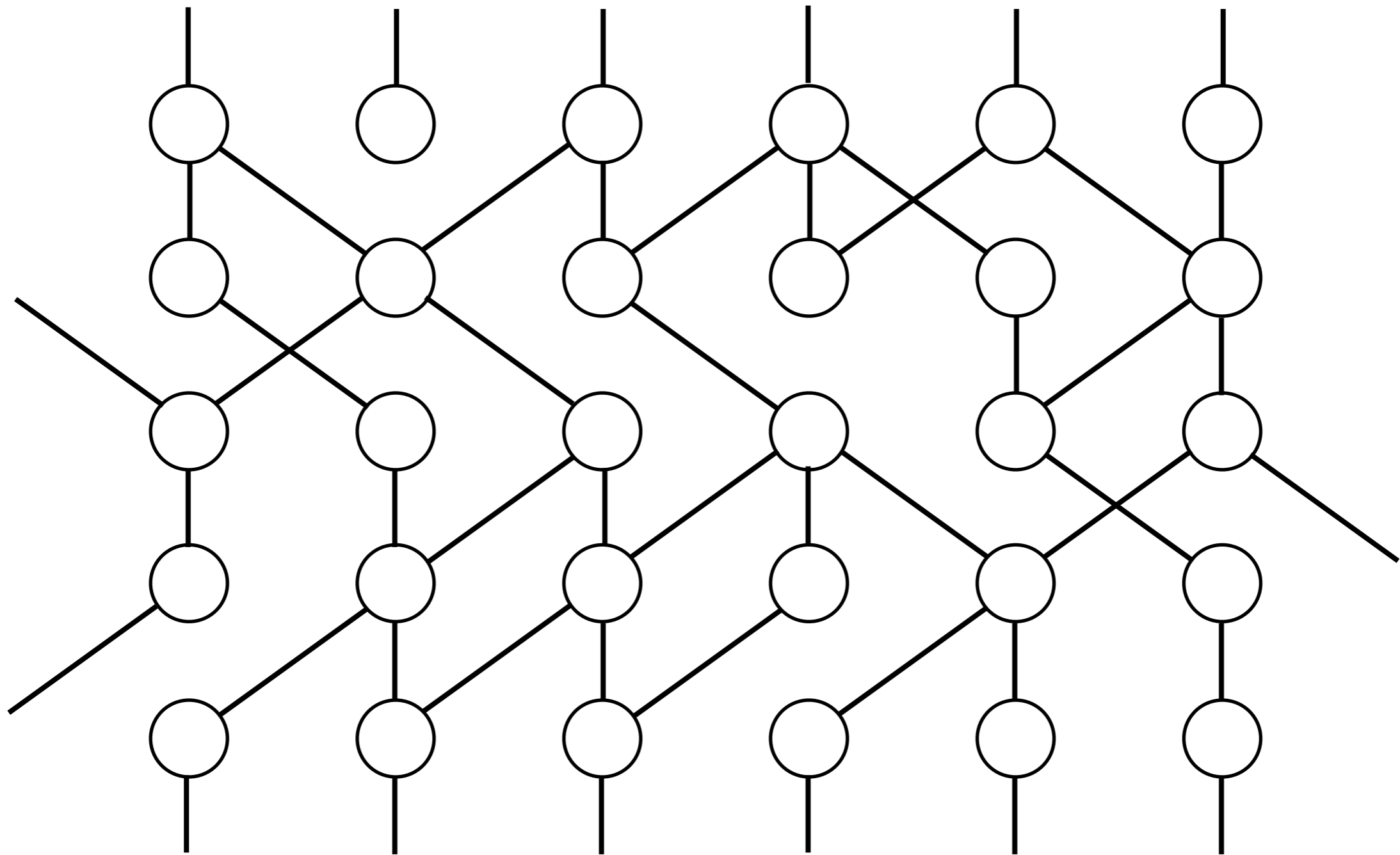
Multiple Teachers

- Learners get data from multiple teachers, who may speak different languages.
- Formally:
 - Data are sets of words $d = \{w_0, w_1, \dots, w_b\}$
 - Words are independent $p(d|h) = \prod_{w \in d} p(w|h)$

(Smith, 2009)

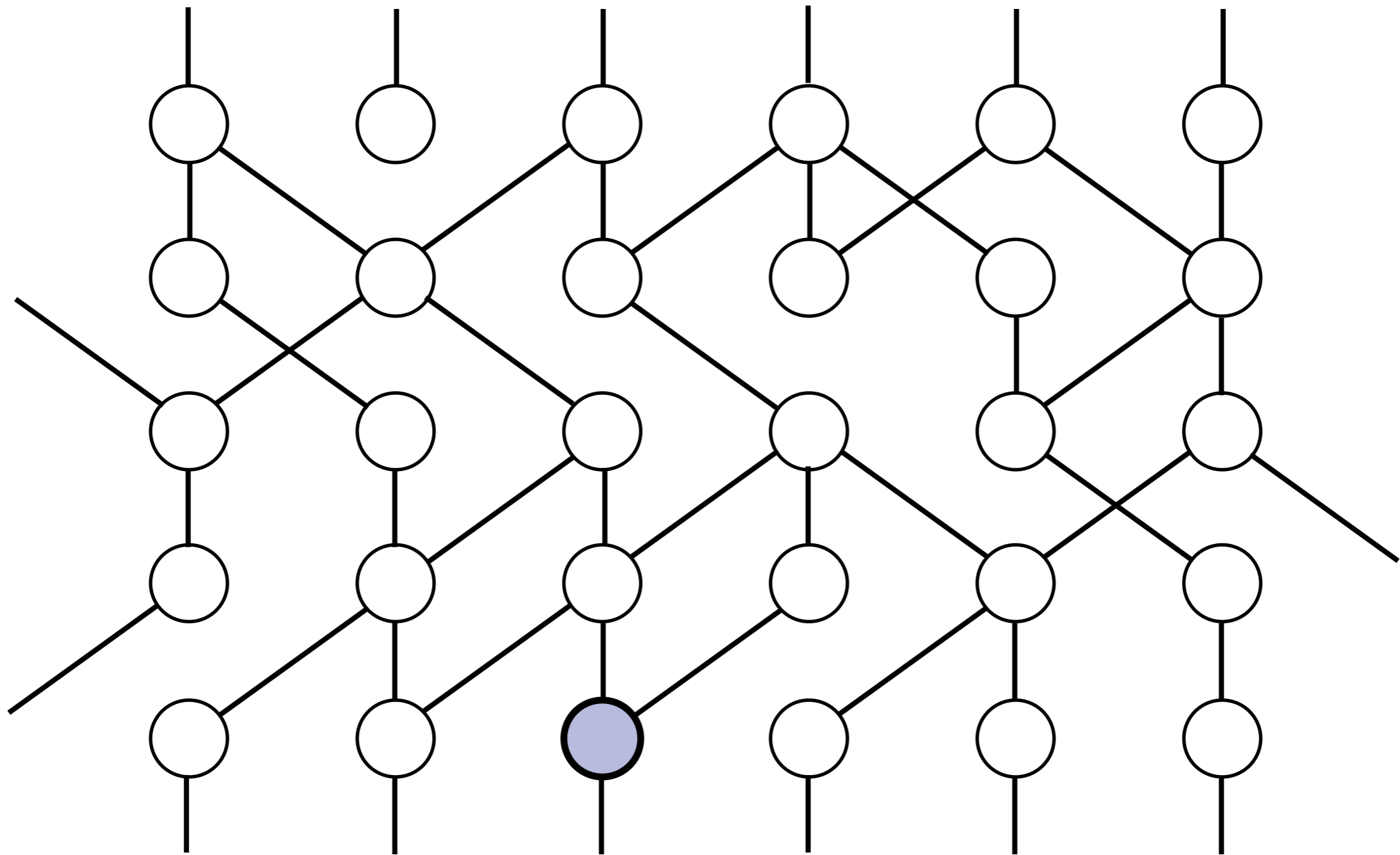


Multiple Teachers



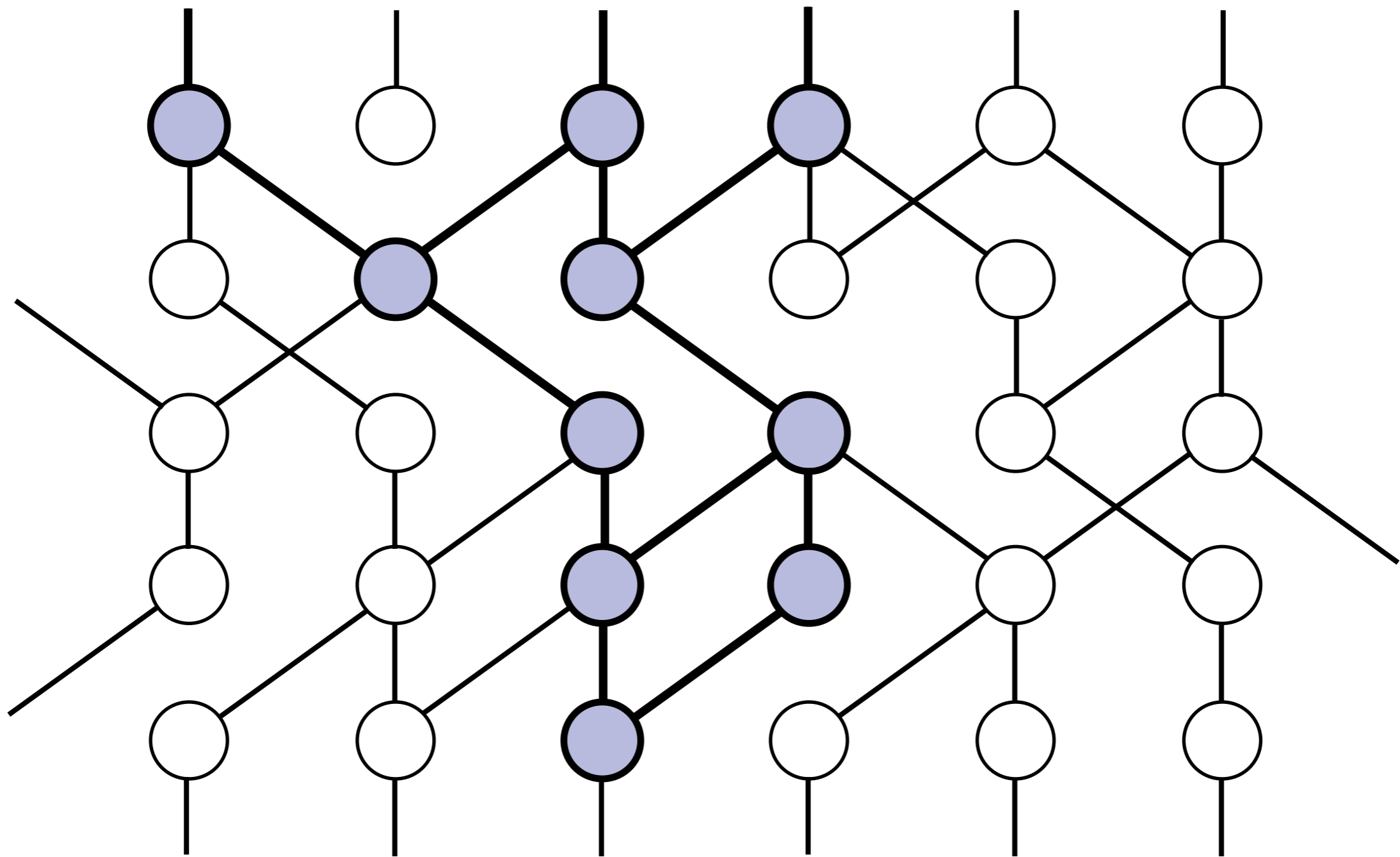


Multiple Teachers





Multiple Teachers





Simulations: Two-Language Setting



Simulations: Two-Language Setting

- Two words: $\{w_0, w_1\}$



Simulations: Two-Language Setting

- Two words: $\{w_0, w_1\}$
- Two languages: $\{h_0, h_1\}$



Simulations: Two-Language Setting

- Two words: $\{w_0, w_1\}$
- Two languages: $\{h_0, h_1\}$
- Likelihood: $p(w|h_i) = \begin{cases} 1 - \epsilon & w = w_i \\ \epsilon & w \neq w_i \end{cases}$



Simulations: Two-Language Setting

- Two words: $\{w_0, w_1\}$
- Two languages: $\{h_0, h_1\}$
- Likelihood: $p(w|h_i) = \begin{cases} 1 - \epsilon & w = w_i \\ \epsilon & w \neq w_i \end{cases}$
- Prior: $p(h_0) = 0.6$

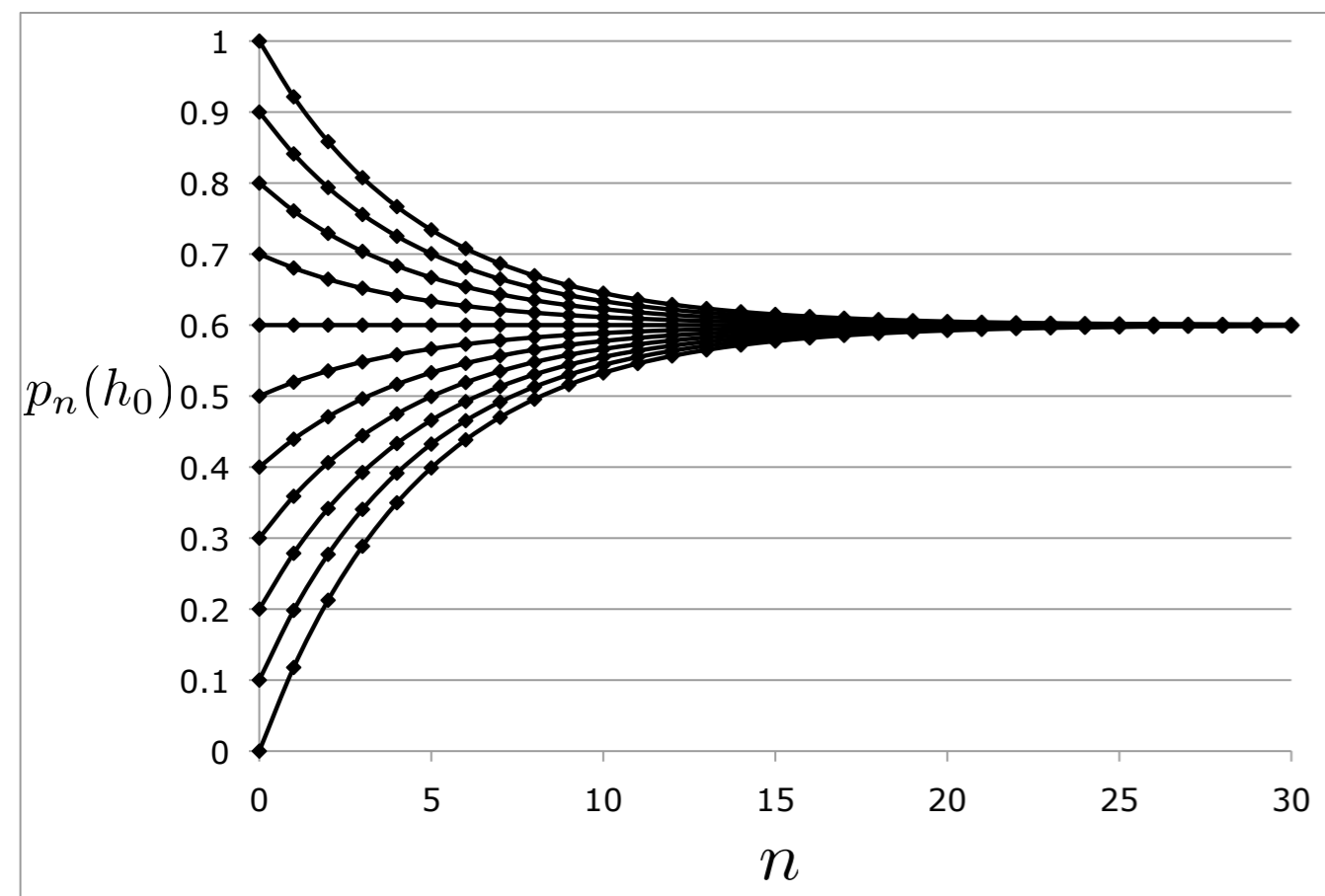


Simulations: Results

(Smith, 2009)

Simulations: Results

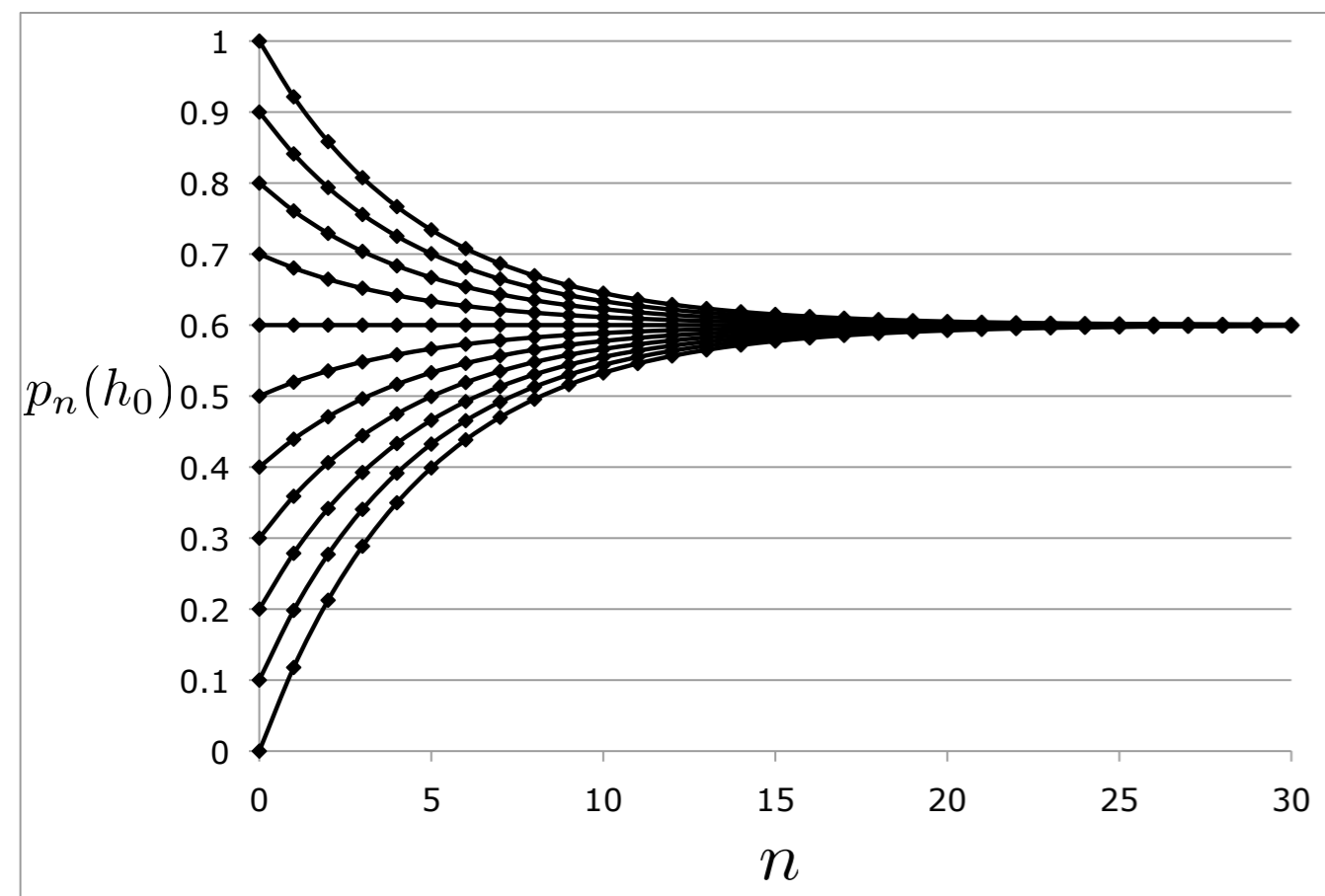
One Teacher



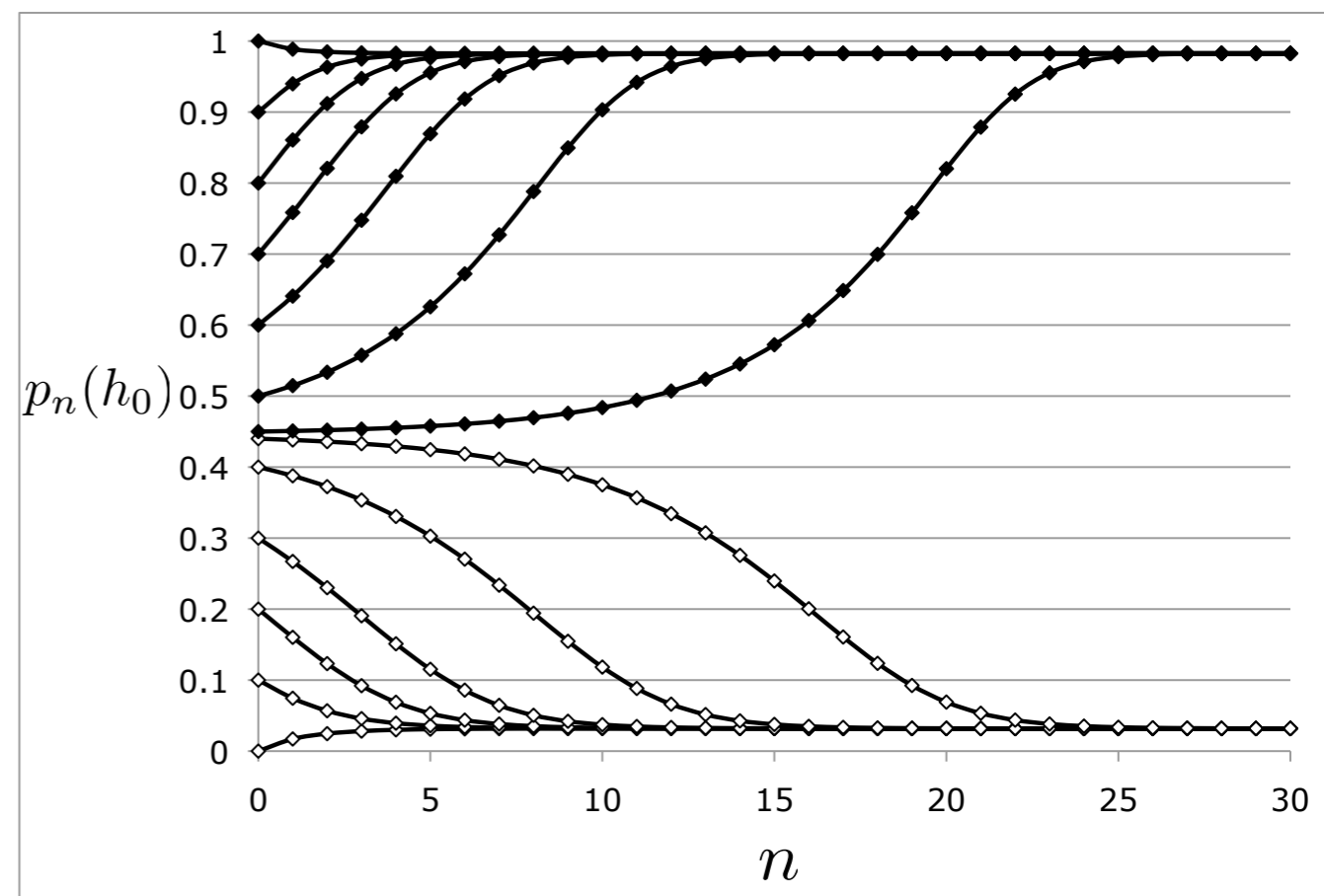
(Smith, 2009)

Simulations: Results

One Teacher



Multiple Teachers



(Smith, 2009)



Outline

- Learning from a single teacher
- Learning from multiple teachers
- Learning multiple languages



Multiple Languages



Multiple Languages

- Problem: prior does not match reality!



Multiple Languages

- Problem: prior does not match reality!
- Assumes all teachers speak the same language.



Multiple Languages

- Problem: prior does not match reality!
 - Assumes all teachers speak the same language.
- Solution: Allow agents to learn multiple languages.



Multiple Languages

- Problem: prior does not match reality!
 - Assumes all teachers speak the same language.
- Solution: Allow agents to learn multiple languages.
 - Simulation 1: Distributions over words
 - Simulation 2: Compositional vs holistic setting



Multiple Languages

- Problem: prior does not match reality!
 - Assumes all teachers speak the same language.
- Solution: Allow agents to learn multiple languages.
 - Simulation 1: Distributions over words
 - Simulation 2: Compositional vs holistic setting
 - Simulation 3: Two-language setting



Distributions over Languages



Distributions over Languages

- Distributions over words are called languages l

Distributions over Languages

- Distributions over words are called languages l

l	w	$p(w l)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5

Distributions over Languages

- Distributions over words are called languages l
- Hypotheses h are distributions over languages $p(l|h)$

l	w	$p(w l)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5

Distributions over Languages

- Distributions over words are called languages l
- Hypotheses h are distributions over languages $p(l|h)$

l	w	$p(w l)$
rhotic	/a:rm/	5/6
rhotic	/a:m/	1/6
non-rhotic	/a:rm/	1/5
non-rhotic	/a:m/	4/5

h	l	$p(l h)$
rho-1/2	rhotic	1/2
rho-1/2	non-rhotic	1/2
rho-9/10	rhotic	9/10
rho-9/10	non-rhotic	1/10



Distributions over Languages: Priors



Distributions over Languages: Priors

- Imagine you've heard 20 words.



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .
- What is your prior belief about the language for w ?



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .
- What is your prior belief about the language for w ?
- Desiderata:



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .
- What is your prior belief about the language for w ?
- Desiderata:
 - l_0 is three times as likely as l_1 .



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .
- What is your prior belief about the language for w ?
- Desiderata:
 - l_0 is three times as likely as l_1 .
 - There is some chance of a new language l_2 .



Distributions over Languages: Priors

- Imagine you've heard 20 words.
- You believe 15 are from l_0 , 5 are from l_1 .
- Now, you hear a new word w .
- What is your prior belief about the language for w ?
- Desiderata:
 - l_0 is three times as likely as l_1 .
 - There is some chance of a new language l_2 .
 - The more words you've already heard, the less likely this word is from a new language.



Distributions over Languages: Priors



Distributions over Languages: Priors

- Solution: Dirichlet Process (DP) prior



Distributions over Languages: Priors

- Solution: Dirichlet Process (DP) prior
- Track n_l , the count of words believed to be from l



Distributions over Languages: Priors

- Solution: Dirichlet Process (DP) prior
- Track n_l , the count of words believed to be from l
- Concentration parameter α controls how often the learner expects to hear new languages



Distributions over Languages: Priors

- Solution: Dirichlet Process (DP) prior
- Track n_l , the count of words believed to be from l
- Concentration parameter α controls how often the learner expects to hear new languages
- When you hear a new word, the prior chooses a language according to:

$$p(l) \propto \begin{cases} n_l & \text{you've heard } l \text{ before} \\ \alpha & l \text{ is a new language} \end{cases}$$



Simulations





Simulations

- Exact inference with a DP prior is intractable.



Simulations

- Exact inference with a DP prior is intractable.
- Use Monte Carlo simulation:



Simulations

- Exact inference with a DP prior is intractable.
- Use Monte Carlo simulation:
 - Each generation has a fixed set of agents A_n .



Simulations

- Exact inference with a DP prior is intractable.
- Use Monte Carlo simulation:
 - Each generation has a fixed set of agents A_n .
 - Word production probability:

$$p_n(w) = \frac{1}{|A_n|} \sum_{a \in A_n} \sum_l p(l|h_a)p(w|l)$$



Simulations

- Exact inference with a DP prior is intractable.
- Use Monte Carlo simulation:
 - Each generation has a fixed set of agents A_n .

- Word production probability:

$$p_n(w) = \frac{1}{|A_n|} \sum_{a \in A_n} \sum_l p(l|h_a)p(w|l)$$

- Use Gibbs sampling to approximate $p(h|d)$.



Simulations

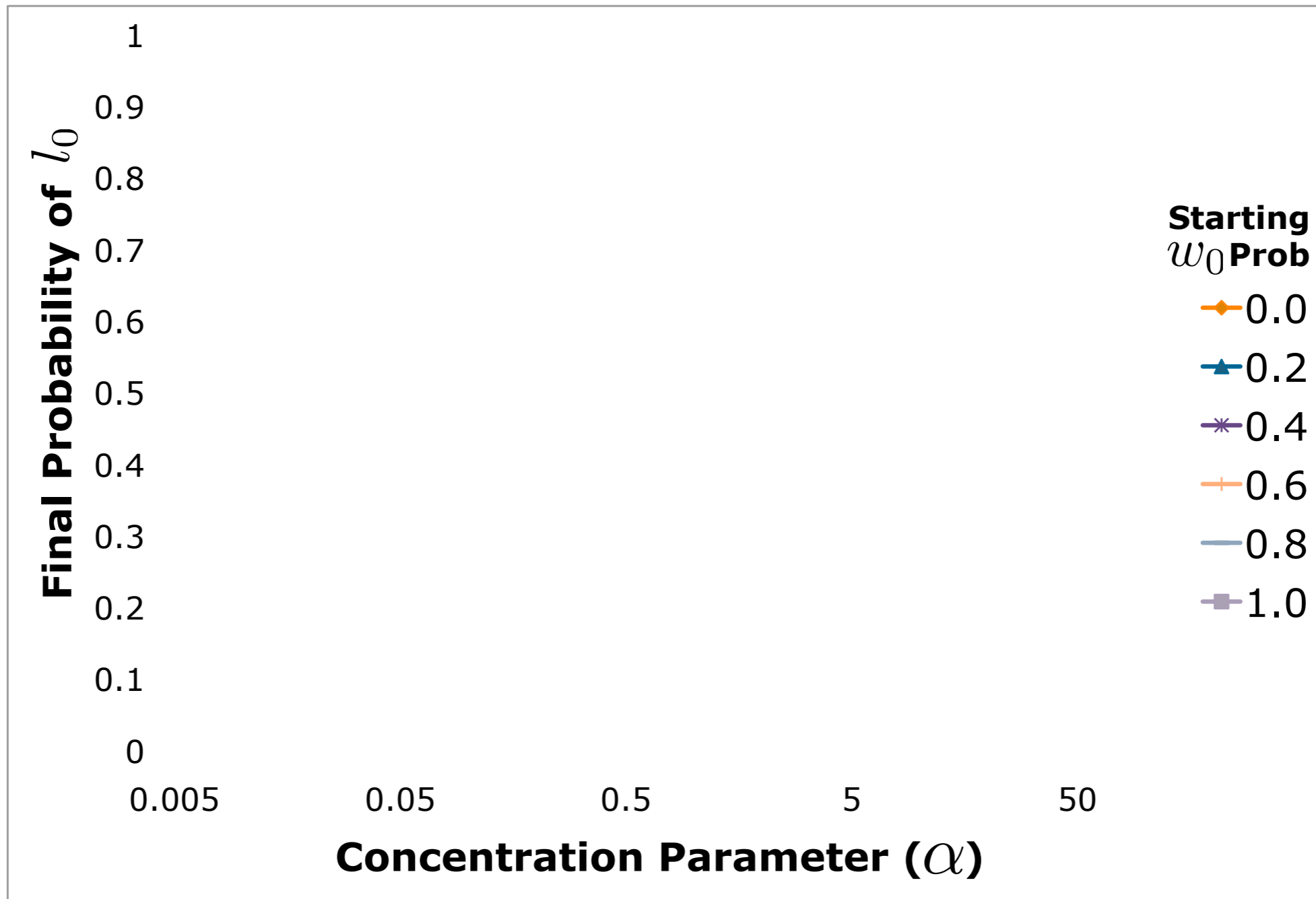
- Exact inference with a DP prior is intractable.
- Use Monte Carlo simulation:
 - Each generation has a fixed set of agents A_n .

- Word production probability:

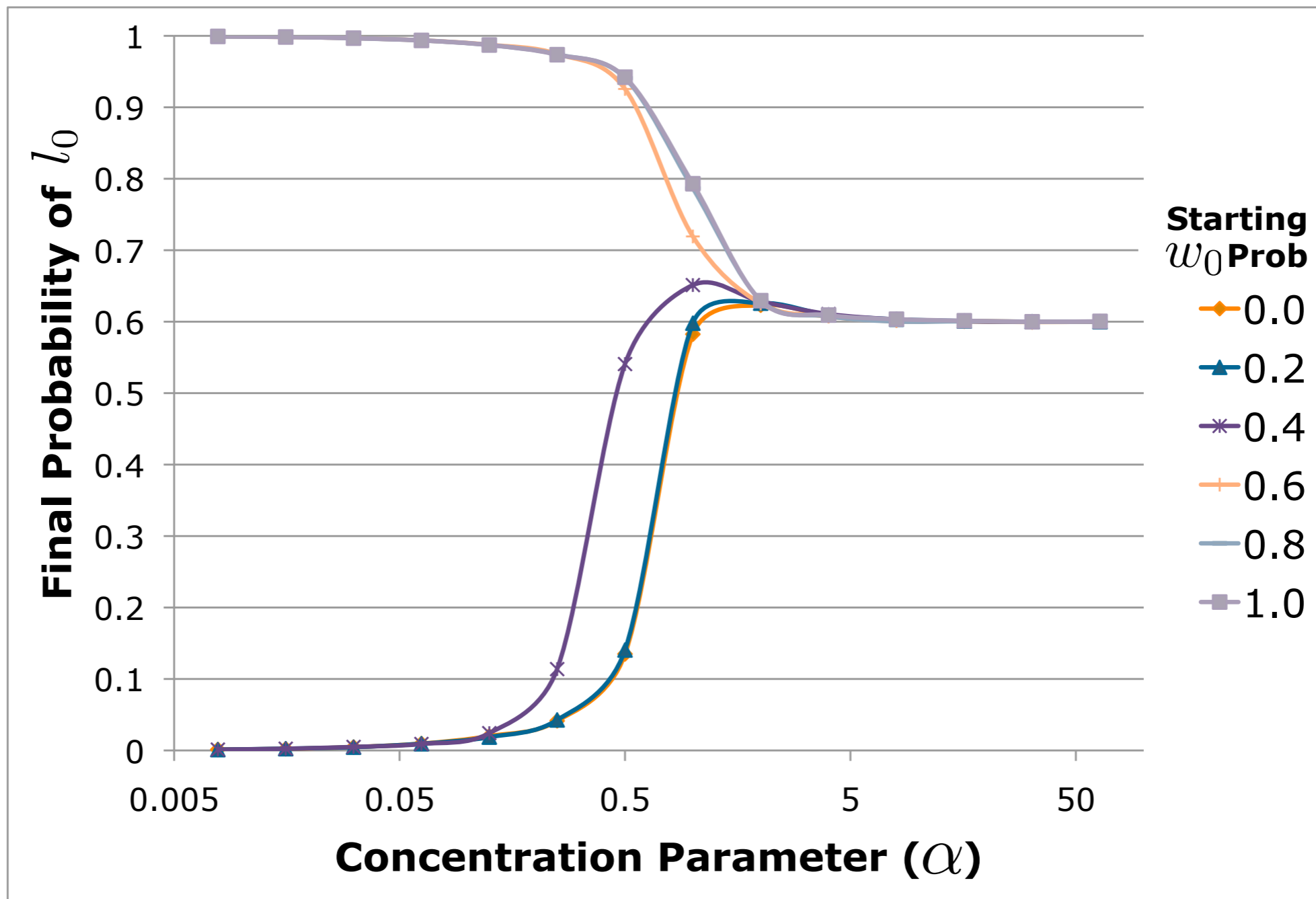
$$p_n(w) = \frac{1}{|A_n|} \sum_{a \in A_n} \sum_l p(l|h_a)p(w|l)$$

- Use Gibbs sampling to approximate $p(h|d)$.
- Run to convergence several times and then average the results.

Simulations: Results

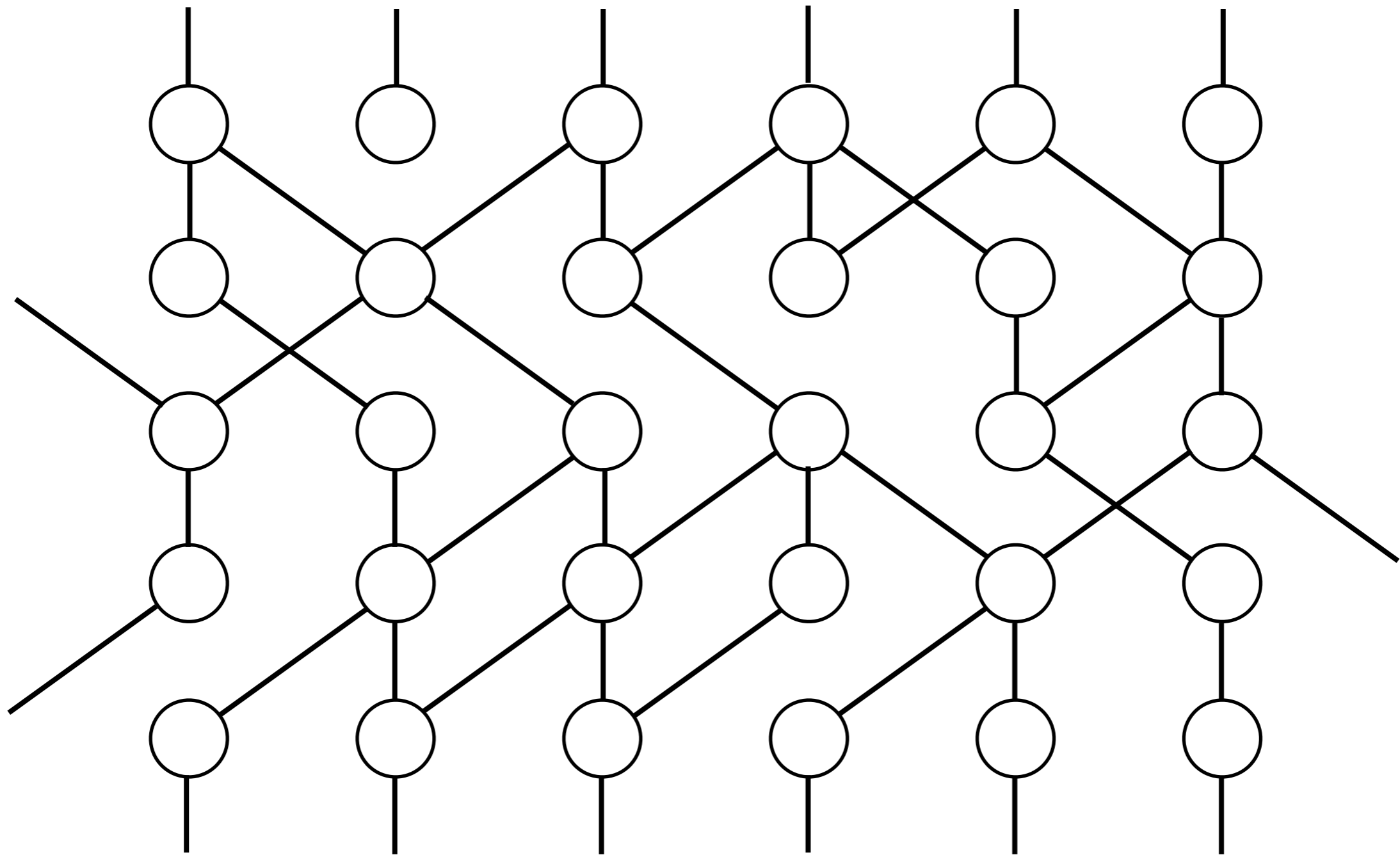


Simulations: Results



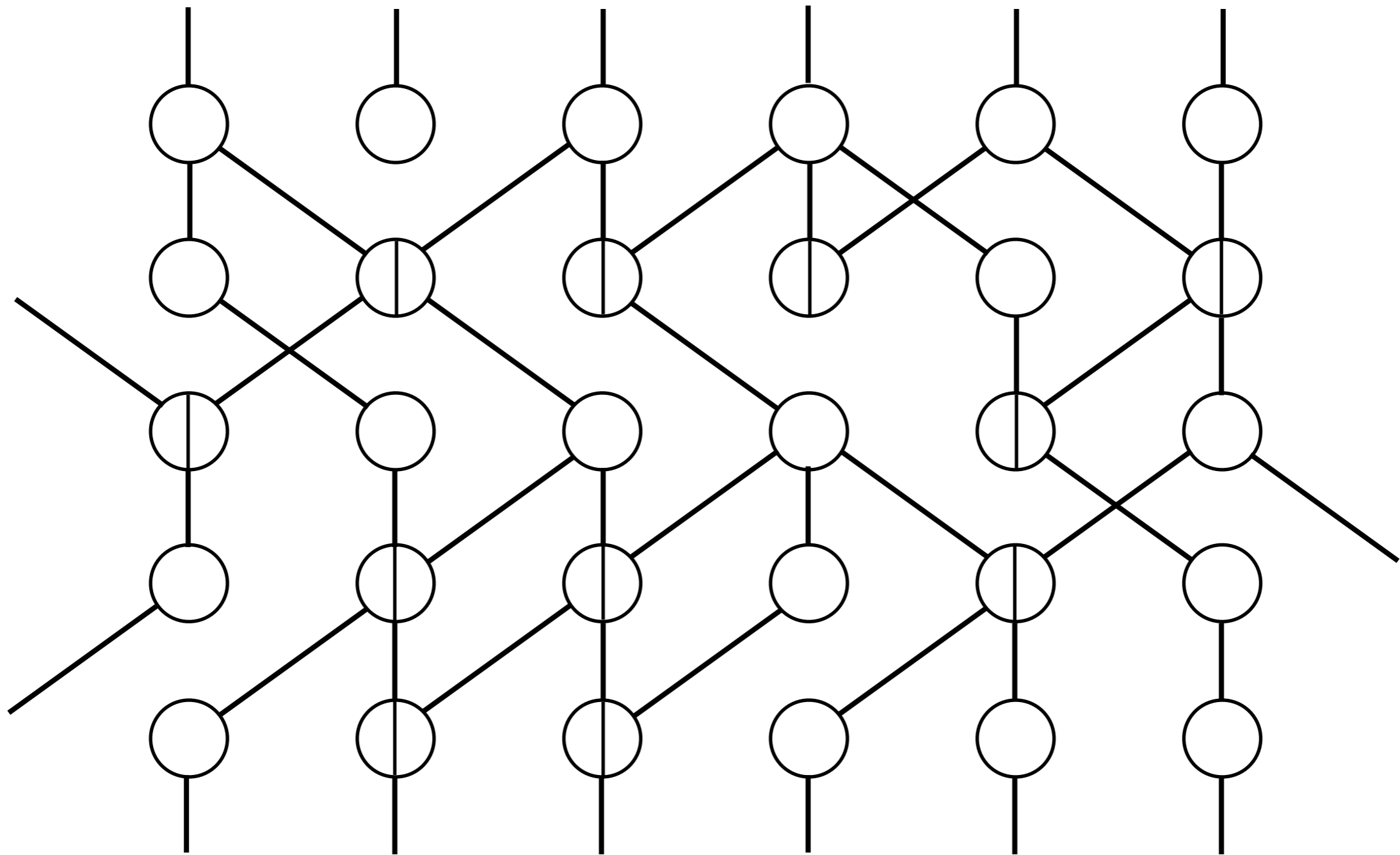


Multiple Languages



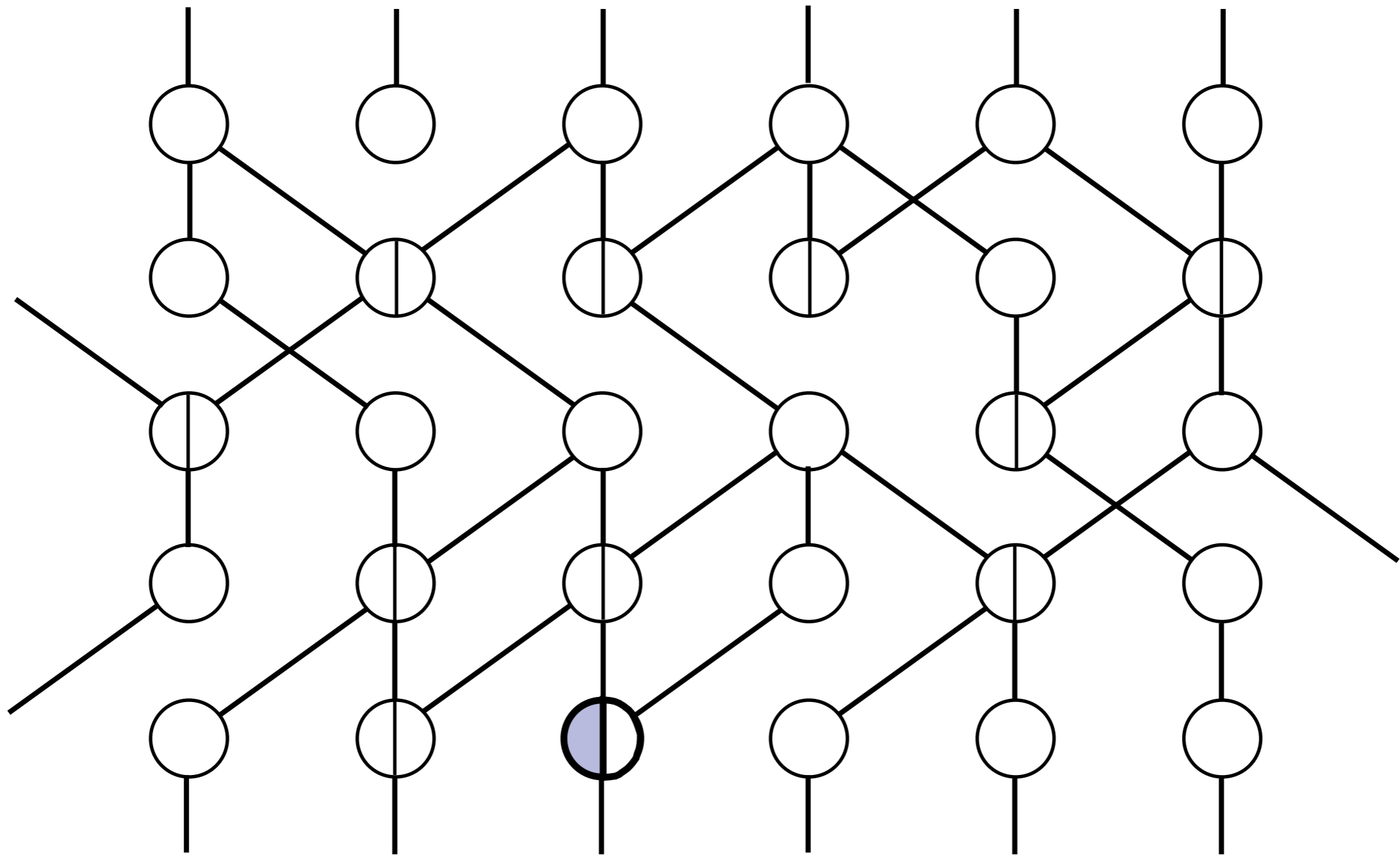


Multiple Languages



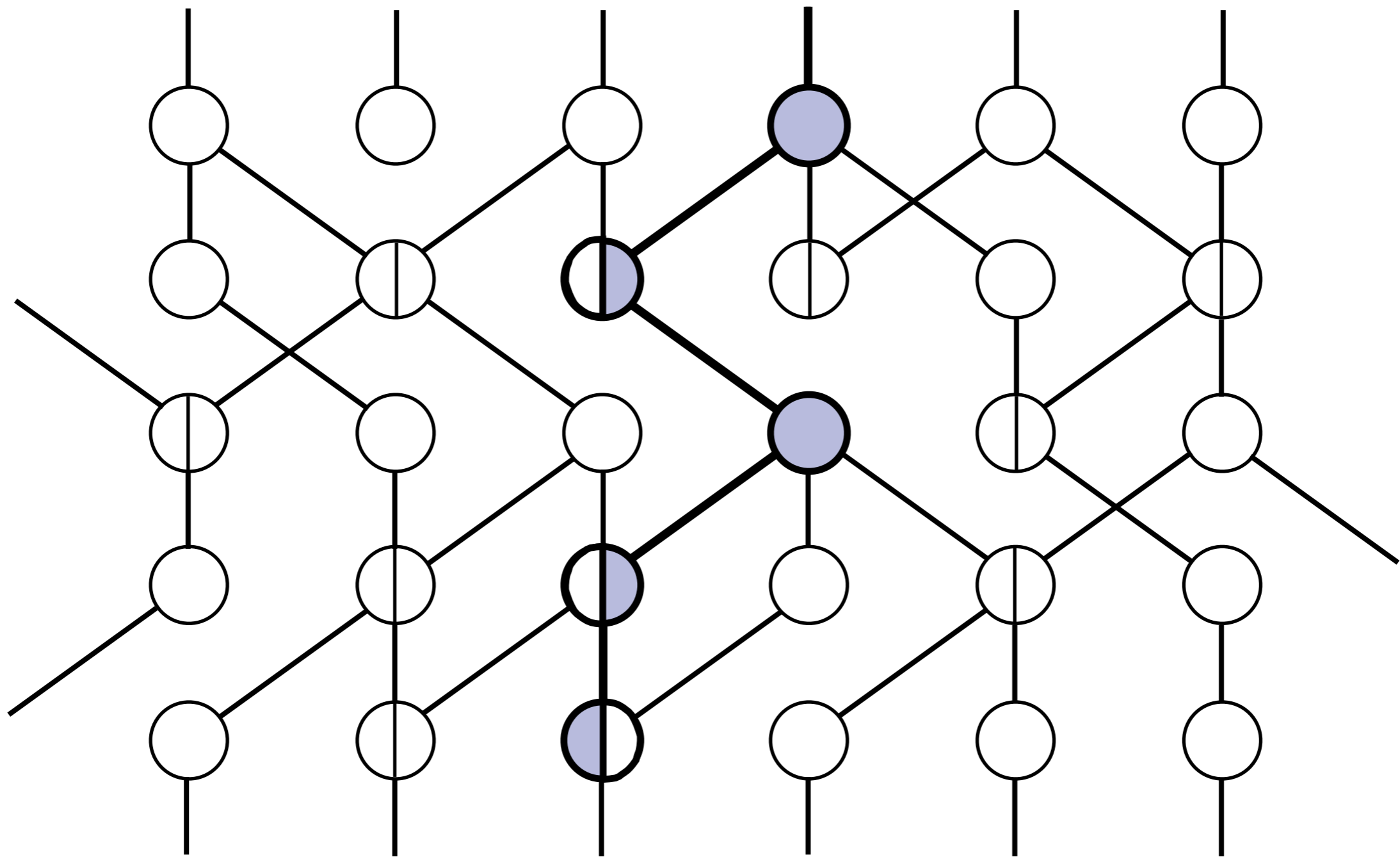


Multiple Languages





Multiple Languages





Distributions over Languages: Discussion



Distributions over Languages: Discussion

- Results look like multiple teacher for low α and single chain for high α .



Distributions over Languages: Discussion

- Results look like multiple teacher for low α and single chain for high α .
- Recall the prior: $p(l) \propto \begin{cases} n_l & \text{you've heard } l \text{ before} \\ \alpha & l \text{ is a new language} \end{cases}$



Distributions over Languages: Discussion

- Results look like multiple teacher for low α and single chain for high α .
- Recall the prior: $p(l) \propto \begin{cases} n_l & \text{you've heard } l \text{ before} \\ \alpha & \textit{l} \text{ is a new language} \end{cases}$
- If you take the limit as $\alpha \rightarrow 0$ you get a single language assumption (i.e. multiple teacher setting).



Distributions over Languages: Discussion

- Results look like multiple teacher for low α and single chain for high α .
- Recall the prior: $p(l) \propto \begin{cases} n_l & \text{you've heard } l \text{ before} \\ \alpha & \textit{l} \text{ is a new language} \end{cases}$
- If you take the limit as $\alpha \rightarrow 0$ you get a single language assumption (i.e. multiple teacher setting).
- If you take the limit as $\alpha \rightarrow \infty$ you get a one language per teacher assumption; learning dynamics are equivalent to single chain.



Conclusion



Conclusion

- The analysis framework of iterated learning with Bayesian agents can be naturally extended to multiple teachers.



Conclusion

- The analysis framework of iterated learning with Bayesian agents can be naturally extended to multiple teachers.
- Our findings interpolate between previous results depending on the learners' belief that multiple languages are present.



Conclusion

- The analysis framework of iterated learning with Bayesian agents can be naturally extended to multiple teachers.
- Our findings interpolate between previous results depending on the learners' belief that multiple languages are present.
- Additional simulations described in the paper confirm these findings.



Thank You

