# Probabilistic Feature Attention as an Alternative to Variables

BRANDON PRICKETT

UNIVERSITY OF MASSACHUSETTS AMHERST

# Overview

1. Introduction
2. My Model (MaxEnt + ***Probablistic Feature Attention***)
3. Identity Generalization
4. Similarity-based Generalization
5. Discussion

# Introduction

# Evidence for variables in phonology

- Variables have been included in theories of phonology for a while (e.g. Halle 1962).
  - In this context, a variable would be any representation that ties together individual tokens in a way that ignores those tokens' individual characteristics.
  - However, more recent models of phonotactics have not included any explicit use of variables (Hayes and Wilson 2008; Pater and Moreton 2012).

# Evidence for variables in phonology

- Variables have been included in theories of phonology for a while (e.g. Halle 1962).
  - In this context, a variable would be any representation that ties together individual tokens in a way that ignores those tokens' individual characteristics.
  - However, more recent models of phonotactics have not included any explicit use of variables (Hayes and Wilson 2008; Pater and Moreton 2012).
- Apparent evidence for variables exists in Hebrew, where stems are not grammatical if their first two consonants are identical (Berent 2013):
  - *simem* 'he intoxicated', but **sisem*
  - This is typically represented with a constraint that includes variables to stand in for the first two segments: **#[α]V[α]*

# Evidence for variables in phonology

- Variables have been included in theories of phonology for a while (e.g. Halle 1962).
  - In this context, a variable would be any representation that ties together individual tokens in a way that ignores those tokens' individual characteristics.
  - However, more recent models of phonotactics have not included any explicit use of variables (Hayes and Wilson 2008; Pater and Moreton 2012).
- Apparent evidence for variables exists in Hebrew, where stems are not grammatical if their first two consonants are identical (Berent 2013):
  - *simem* 'he intoxicated', but **sisem*
  - This is typically represented with a constraint that includes variables to stand in for the first two segments: **#[α]V[α]*
- Berent (2013) argues that the fact that Hebrew speakers generalize this pattern to non-native segments means that variables must be used by the phonological grammar.
  - Additionally, Gallagher (2013) and Moreton (2012) have both showed that participants in artificial language learning studies seemed to be using variables in their phonology.

# My Model

# The base model

- To explore the effects of PFA, I'll be adding it to a fairly standard MaxEnt phonotactic learner, GMECCS (Pater and Moreton 2012, Moreton et al. 2017).
- Following Gluck and Bower (1988), GMECCS uses a constraint set that includes every possible ngram of every possible feature bundle.
  - To ensure that there aren't infinite constraints, the model is limited to the smallest feature set and ngram size necessary to run a particular simulation.
  - For example, if the model was used in a simulation with only four segments, two relevant features, and words of length 1, it would only need 8 constraints:

| | *[+voice] | *[-voice] | *[+cont.] | *[-cont.] | *[+voice, +cont.] | *[+voice, -cont.] | *[-voice, +cont.] | *[-voice, -cont.] |
|---|---|---|---|---|---|---|---|---|
| d | * | | | * | | * | | |
| z | * | | * | | * | | | |
| t | | * | | * | | | | * |
| s | | * | * | | | | * | |

- Following Hayes and Wilson (2008), GMECCS uses gradient descent to find the optimal weights for these constraints.

# Gradient Descent for Phonotactics

**Learning Datum:**

Higher Weights→

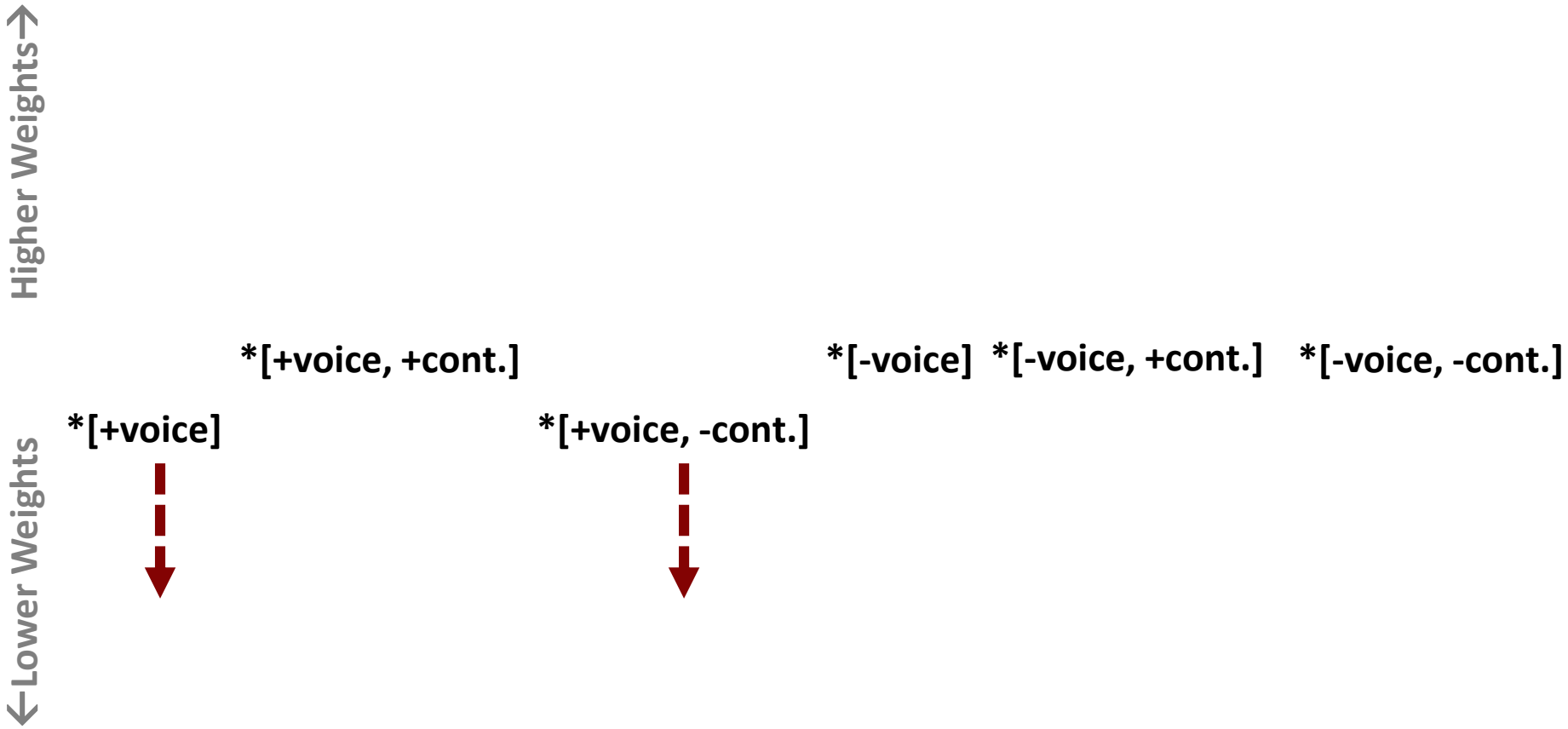
*[+voice] *[+voice, +cont.] *[+voice, -cont.] *[-voice] *[-voice, +cont.] *[-voice, -cont.]

←Lower Weights

# Gradient Descent for Phonotactics

**Learning Datum:**

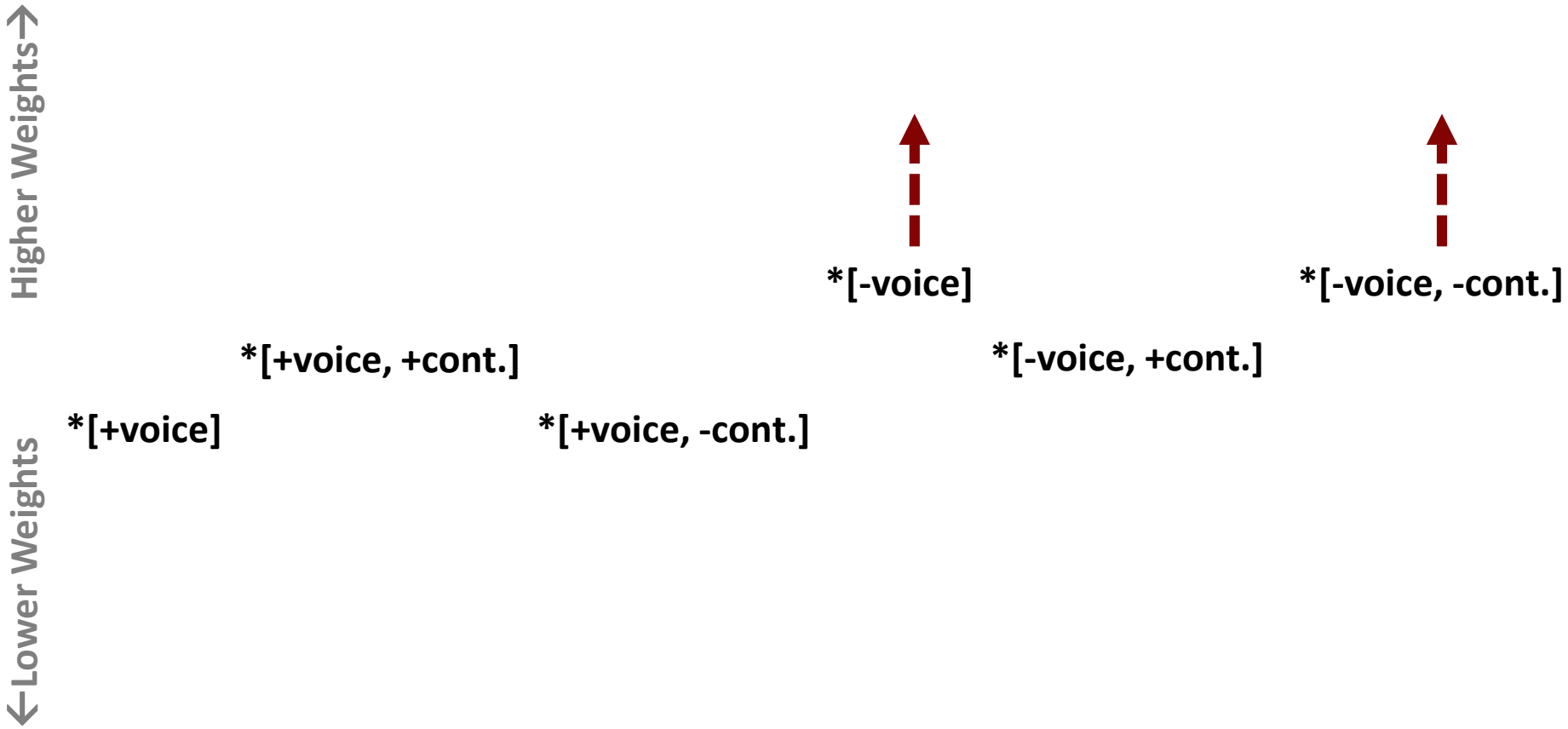
**Pr(d) = 1**



# Gradient Descent for Phonotactics

**Learning Datum:**

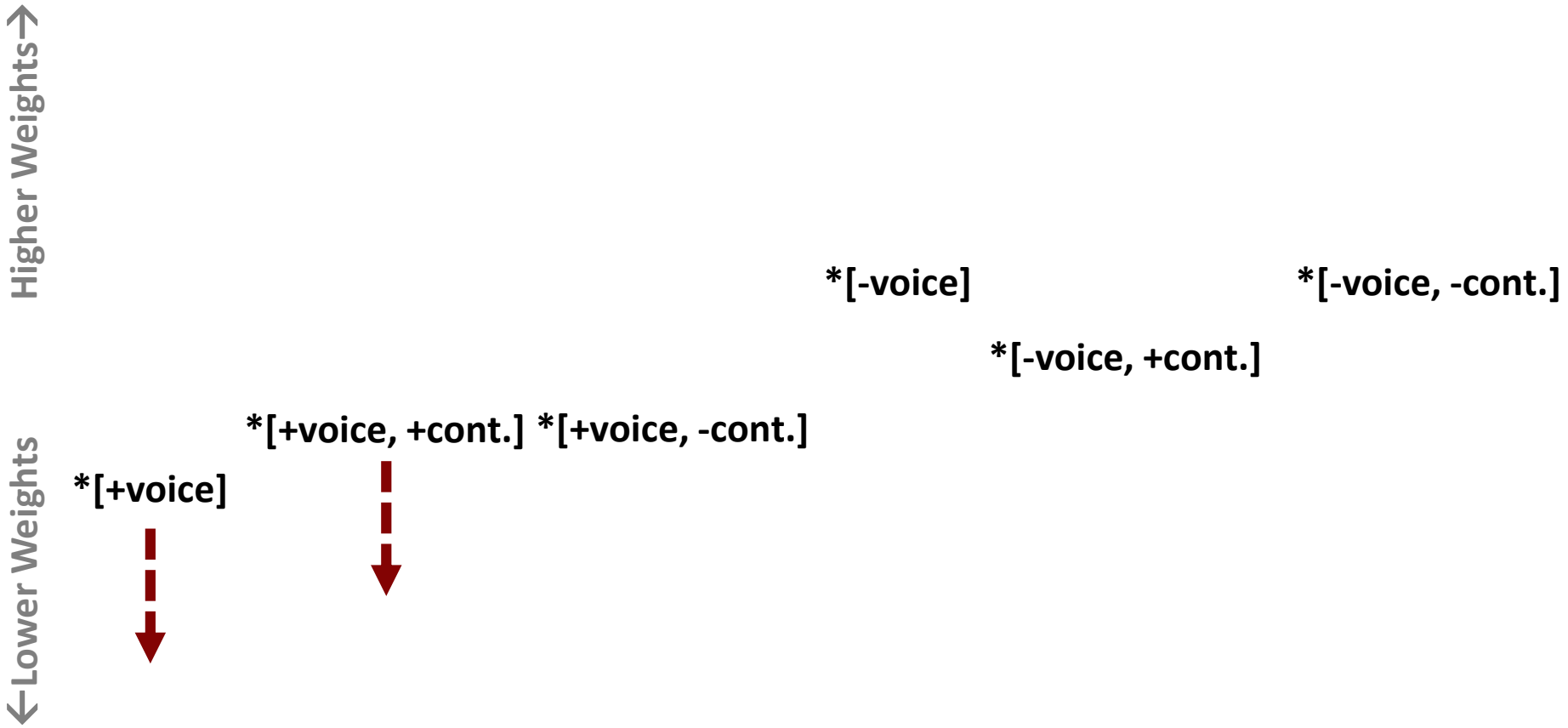
**Pr(t) = 0**



# Gradient Descent for Phonotactics

**Learning Datum:**

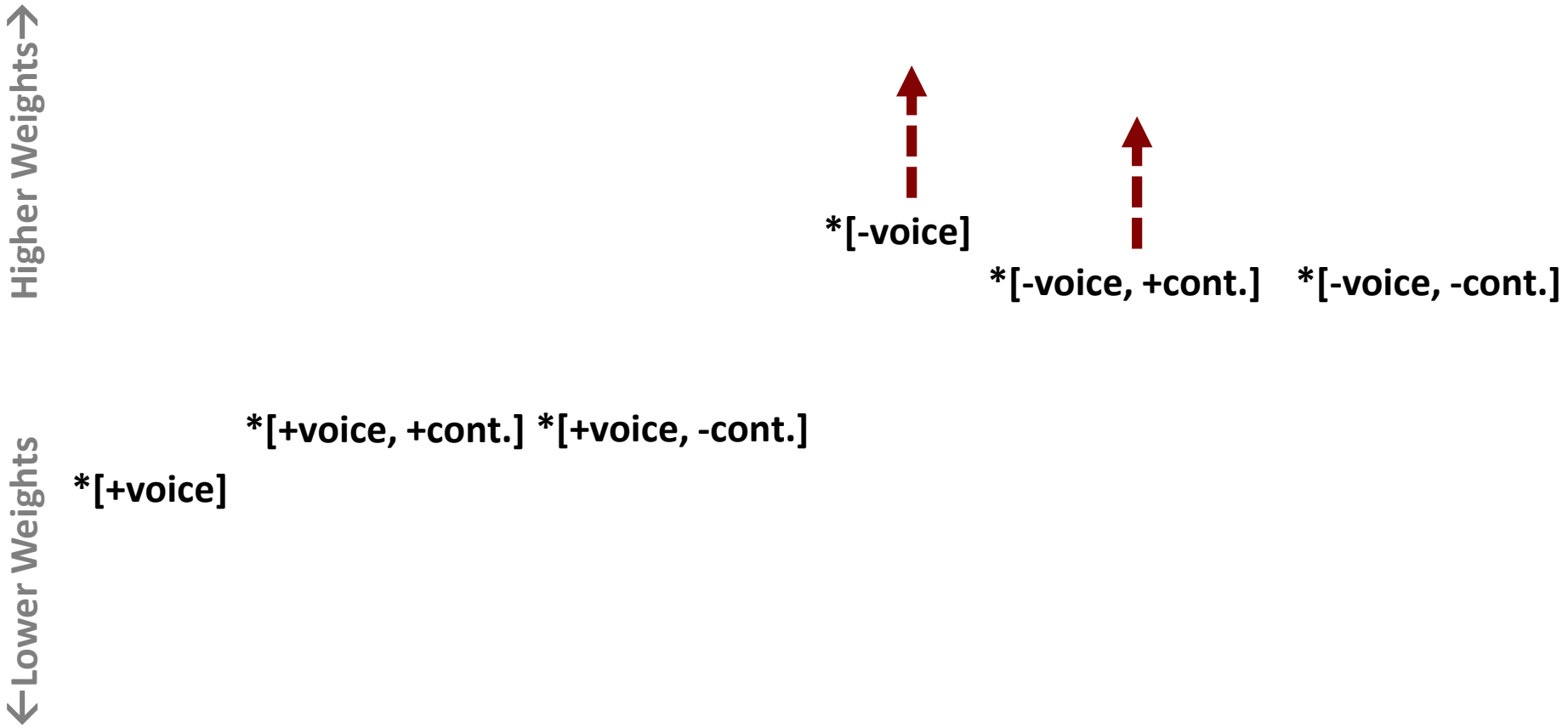
**Pr(z) = 1**



# Gradient Descent for Phonotactics

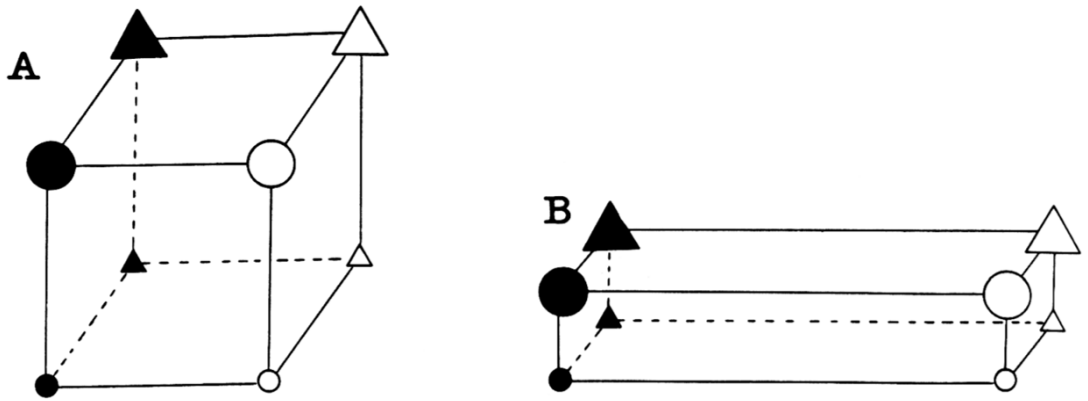
**Learning Datum:**

**Pr(s) = 0**



# Probabilistic Feature Attention (PFA)

- In Probabilistic Feature Attention (PFA), learners only attend to a subset of features for each datum in each iteration of learning.
  - This was inspired by dropout (Srivastava et al. 2014), a mechanism used for training deep neural networks.
  - It's also related to *Selective Feature Attention*, which was proposed by Nosofsky (1986) to explain biases in visual category learning.




**(Figure from Nosofksy 1986)**

- The claims that I'm making with PFA are:
  (1) Language learners don't attend to every phonological feature every time they hear a word.
  (2) This lack of attention creates ambiguity in the learner's input.
  (3) In the face of ambiguity, learners err on the side of assigning constraint violations (i.e ambiguous segments' violation vectors are the union of the violation vectors for the segments that make them up).

# Ambiguity due to PFA

**(Unambiguous segments when all features are attended to.)**

| | *[+voice] | *[-voice] | *[+cont.] | *[-cont.] | *[+voice, +cont.] | *[+voice, -cont.] | *[-voice, +cont.] | *[-voice, -cont.] |
|---|---|---|---|---|---|---|---|---|
| d | * | | | * | | * | | |
| z | * | | * | | * | | | |
| t | | * | | * | | | | * |
| s | | * | * | | | | * | |
| T | | * | * | * | | | * | * |
| D | * | | * | * | * | * | | |
| Δ | * | * | | * | | * | | * |
| Z | * | * | * | | * | | * | |
| ? | * | * | * | * | * | * | * | * |

# Ambiguity due to PFA

**(Ambiguous segments when only [voice] is attended to.)**

| | *[+voice] | *[-voice] | *[+cont.] | *[-cont.] | *[+voice, +cont.] | *[+voice, -cont.] | *[-voice, +cont.] | *[-voice, -cont.] |
|---|---|---|---|---|---|---|---|---|
| d | * | | | * | | * | | |
| z | * | | * | | * | | | |
| t | | * | | * | | | | * |
| s | | * | * | | | | * | |
| T | | * | * | * | | | * | * |
| D | * | | * | * | * | * | | |
| ð | * | * | | * | | * | | * |
| Z | * | * | * | | * | | * | |
| ? | * | * | * | * | * | * | * | * |

# Ambiguity due to PFA

**(Ambiguous segments when only [continuant] is attended to.)**

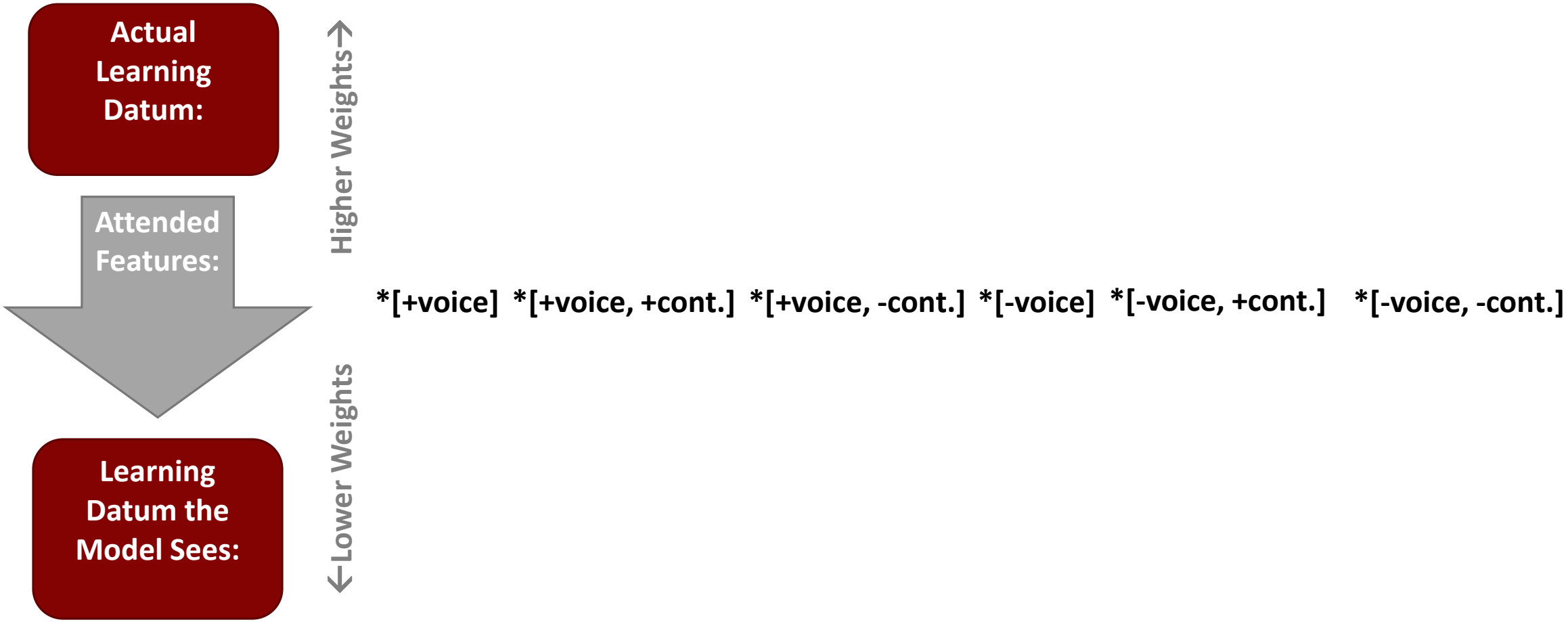
| | *[+voice] | *[-voice] | *[+cont.] | *[-cont.] | *[+voice, +cont.] | *[+voice, -cont.] | *[-voice, +cont.] | *[-voice, -cont.] |
|---|---|---|---|---|---|---|---|---|
| d | * | | | * | | * | | |
| z | * | | * | | * | | | |
| t | | * | | * | | | | * |
| s | | * | * | | | | * | |
| T | | * | * | * | | | * | * |
| D | * | | * | * | * | * | | |
| ʤ | * | * | | * | | * | | * |
| ʒ | * | * | * | | * | | * | |
| ? | * | * | * | * | * | * | * | * |

# Ambiguity due to PFA

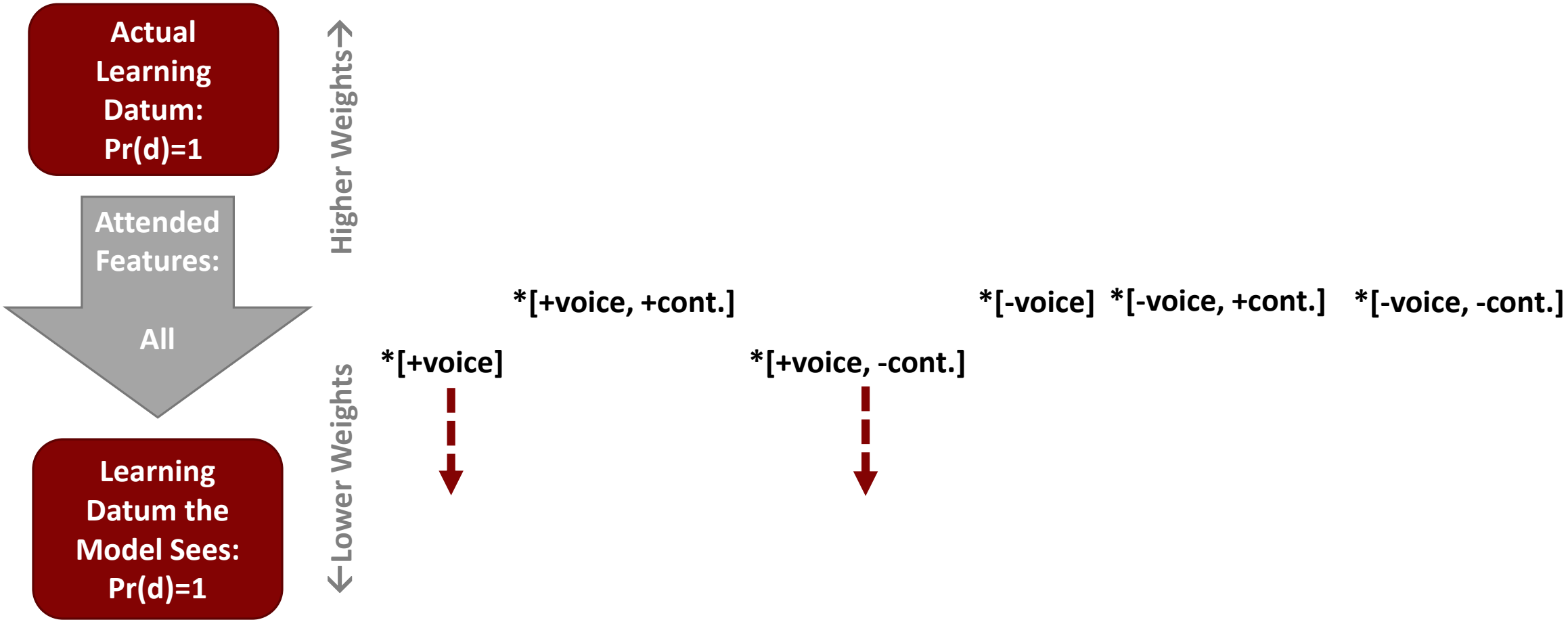
**(Ambiguous segment no features are attended to.)**

| | *[+voice] | *[-voice] | *[+cont.] | *[-cont.] | *[+voice, +cont.] | *[+voice, -cont.] | *[-voice, +cont.] | *[-voice, -cont.] |
|---|---|---|---|---|---|---|---|---|
| d | * | | | * | | * | | |
| z | * | | * | | * | | | |
| t | | * | | * | | | | * |
| s | | * | * | | | | * | |
| T | | * | * | * | | | * | * |
| D | * | | * | * | * | * | | |
| ð | * | * | | * | | * | | * |
| Z | * | * | * | | * | | * | |
| ? | * | * | * | * | * | * | * | * |

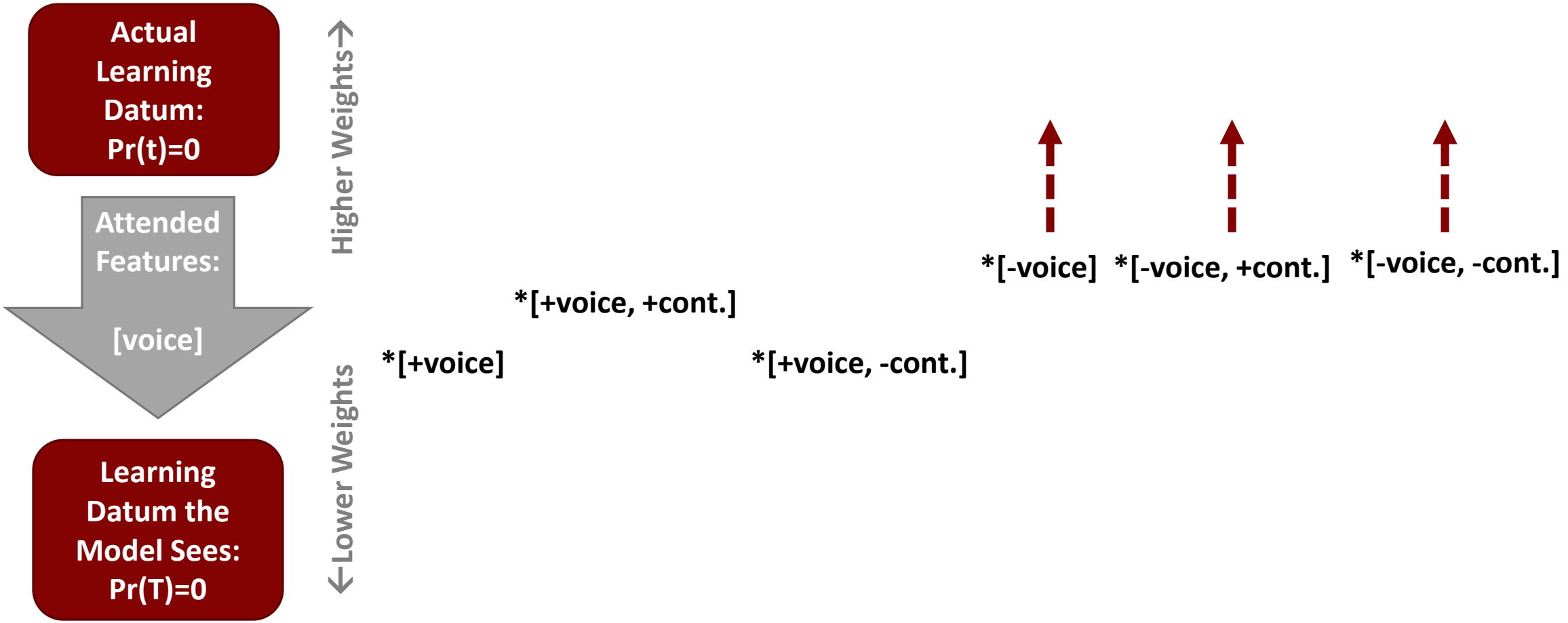
# Learning with PFA



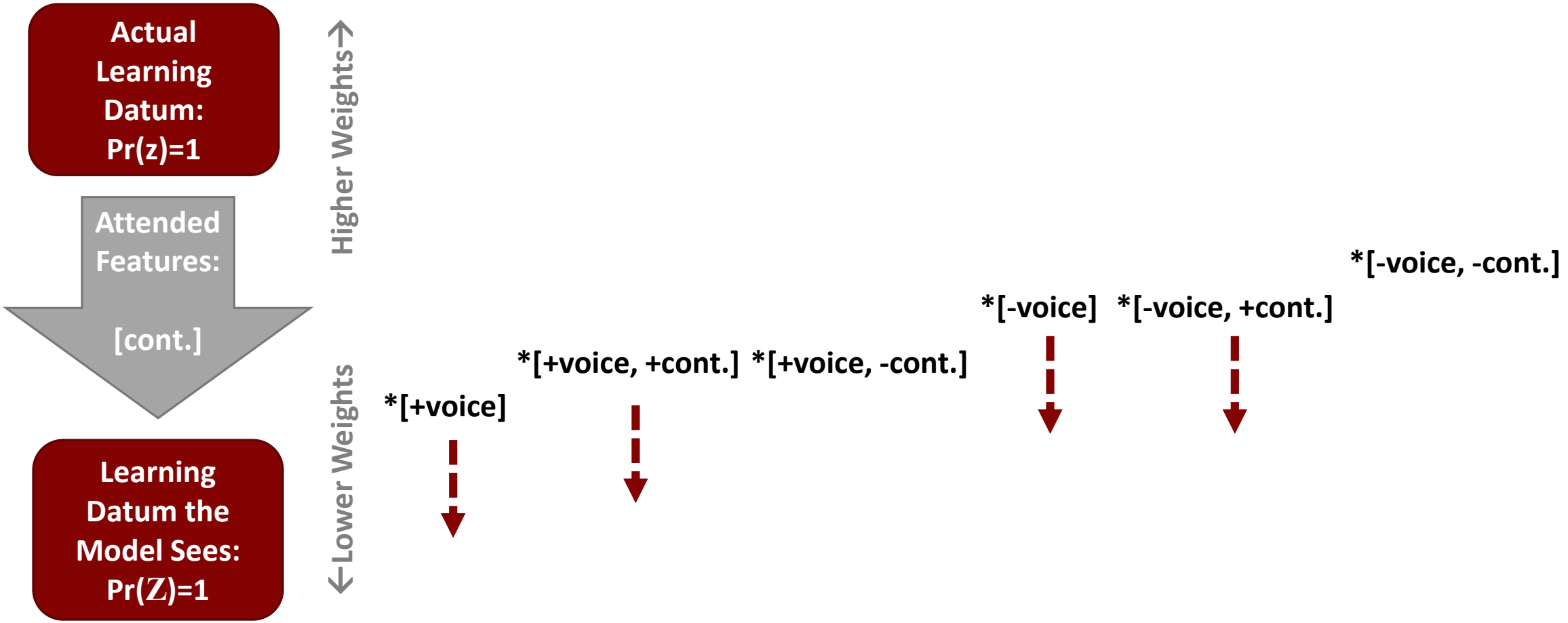
# Learning with PFA



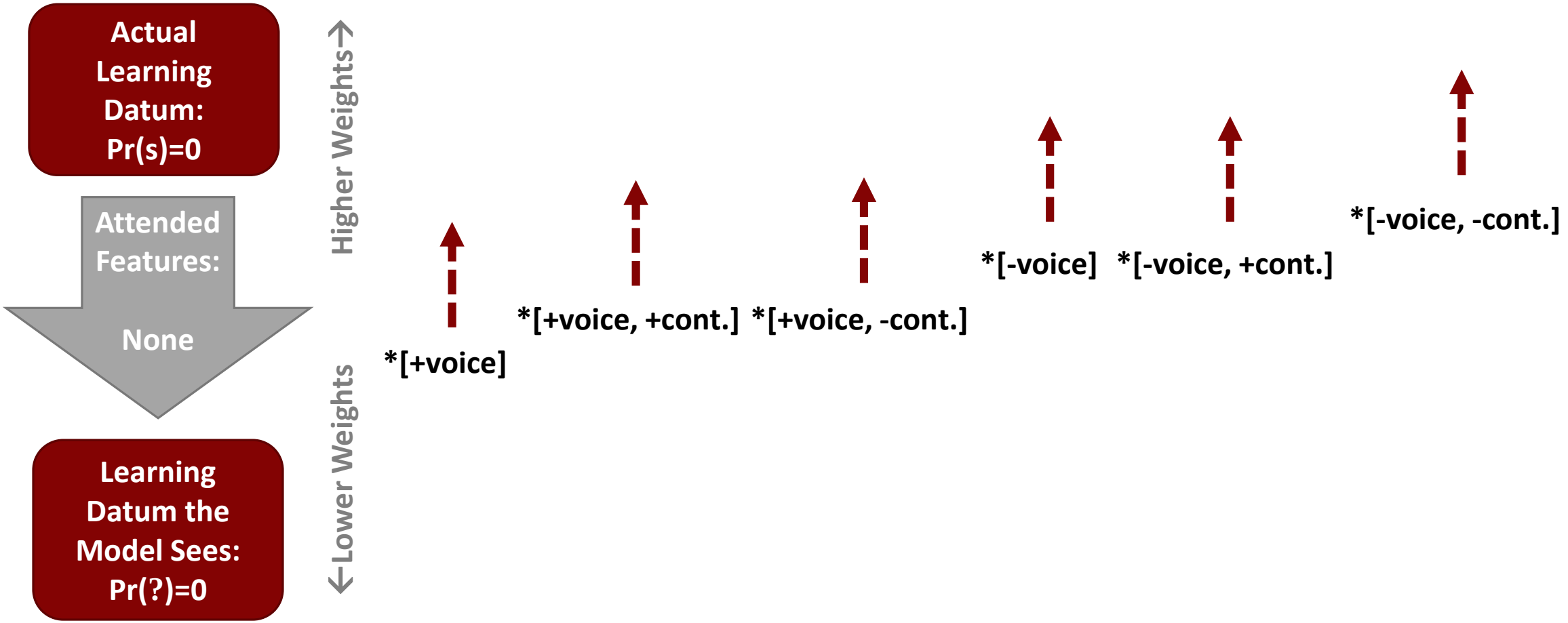
# Learning with PFA



# Learning with PFA

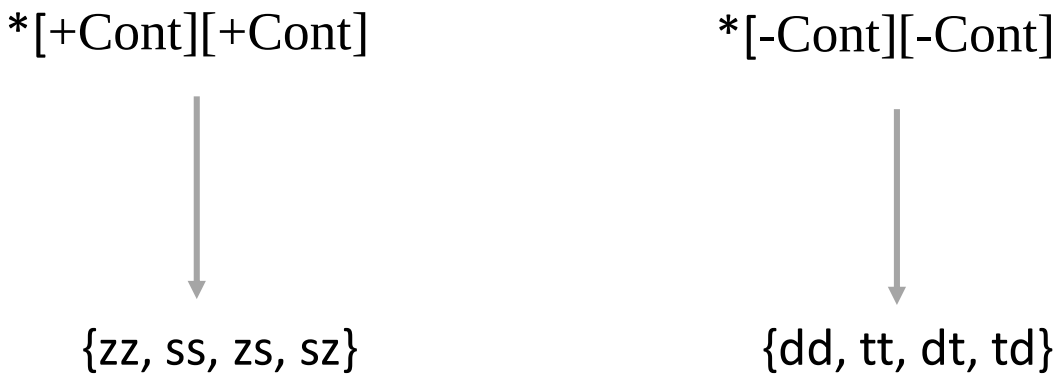


# Learning with PFA



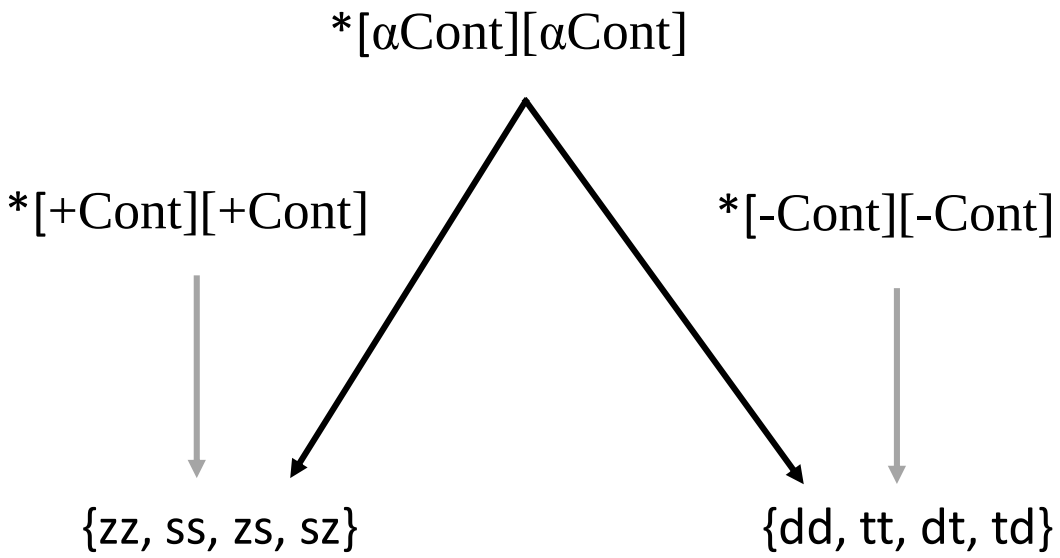
# The relationship between PFA and Variables

- So on a conceptual level, why would we expect PFA to be a viable alternative to variables?
- Because they both perform a similar task of tying together two classes of segments.
- In the figure to the right, this is shown for our simple language, where the only features are [voice] and [continuant] and the only segments are [t], [s], [d], and [z].
- But now, bigrams are allowed and the relevant pattern is dissimilation of the feature [continuant].



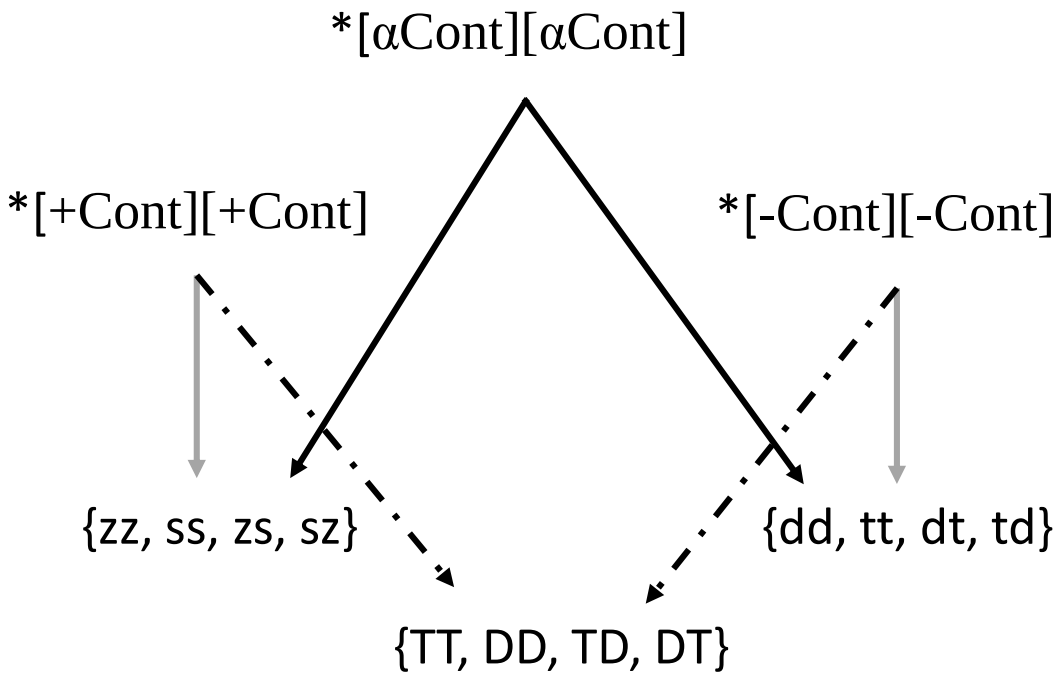
# The relationship between PFA and Variables

- So on a conceptual level, why would we expect PFA to be a viable alternative to variables?
- Because they both perform a similar task of tying together two classes of segments.
- In the figure to the right, this is shown for our simple language, where the only features are [voice] and [continuant] and the only segments are [t], [s], [d], and [z].
- But now, bigrams are allowed and the relevant pattern is dissimilation of the feature [continuant].
  - Variables capture this by creating a constraint that refers to both classes of illegal clusters.



# The relationship between PFA and Variables

- So on a conceptual level, why would we expect PFA to be a viable alternative to variables?
- Because they both perform a similar task of tying together two classes of segments.
- In the figure to the right, this is shown for our simple language, where the only features are [voice] and [continuant] and the only segments are [t], [s], [d], and [z].
- But now, bigrams are allowed and the relevant pattern is dissimilation of the feature [continuant].
  - Variables capture this by creating a constraint that refers to both classes of illegal clusters.
  - PFA captures this by creating ambiguous words that violate the two constraints that refer to the two classes of illegal clusters.



# Identity Generalization

# The Phenomenon

- I'm using the term ***Identity Generalization*** to refer to the human behavior of generalizing a phonological pattern involving words like "dada" and "baba" to words like "gaga".
- Gallagher (2013) showed that humans perform this task in an artificial language learning paradigm.
  - Her subjects were exposed to a phonological alternation that was blocked in words with the shapes [dVdV] and [bVbV].
  - When tested on [gVgV] and [dVgV] sequences, subjects generalized the process to the latter significantly more than the former (which is what a model with variables would predict).
  - In modeling this, I'll frame learning the words that weren't undergoers as a phonotactic learning problem.

# The Phenomenon

- I'm using the term ***Identity Generalization*** to refer to the human behavior of generalizing a phonological pattern involving words like "dada" and "baba" to words like "gaga".
- Gallagher (2013) showed that humans perform this task in an artificial language learning paradigm.
  - Her subjects were exposed to a phonological alternation that was blocked in words with the shapes [dVdV] and [bVbV].
  - When tested on [gVgV] and [dVgV] sequences, subjects generalized the process to the latter significantly more than the former (which is what a model with variables would predict).
  - In modeling this, I'll frame learning the words that weren't undergoers as a phonotactic learning problem.
- See Marcus et al. (1999) and Berent et al. (2014) for similar results in artificial language learning experiments involving reduplicative patterns.

# The Phenomenon

- I'm using the term ***Identity Generalization*** to refer to the human behavior of generalizing a phonological pattern involving words like "dada" and "baba" to words like "gaga".
- Gallagher (2013) showed that humans perform this task in an artificial language learning paradigm.
  - Her subjects were exposed to a phonological alternation that was blocked in words with the shapes [dVdV] and [bVbV].
  - When tested on [gVgV] and [dVgV] sequences, subjects generalized the process to the latter significantly more than the former (which is what a model with variables would predict).
  - In modeling this, I'll frame learning the words that weren't undergoers as a phonotactic learning problem.
- See Marcus et al. (1999) and Berent et al. (2014) for similar results in artificial language learning experiments involving reduplicative patterns.
- As I mentioned previously, this has also been observed in natural language.
  - Hebrew speakers judge nonce words like [tʃetʃem] as worse than words like [metʃetʃ] (Berent 2013).

# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior without variables (Berent et al. 2012; Gallagher 2013).
  - The unigram [d] will gradually acquire more probability than the unigram [g], since it's attested.
  - Constraints penalizing [dg] and [gg] will have the same weights, since both bigrams are unattested.

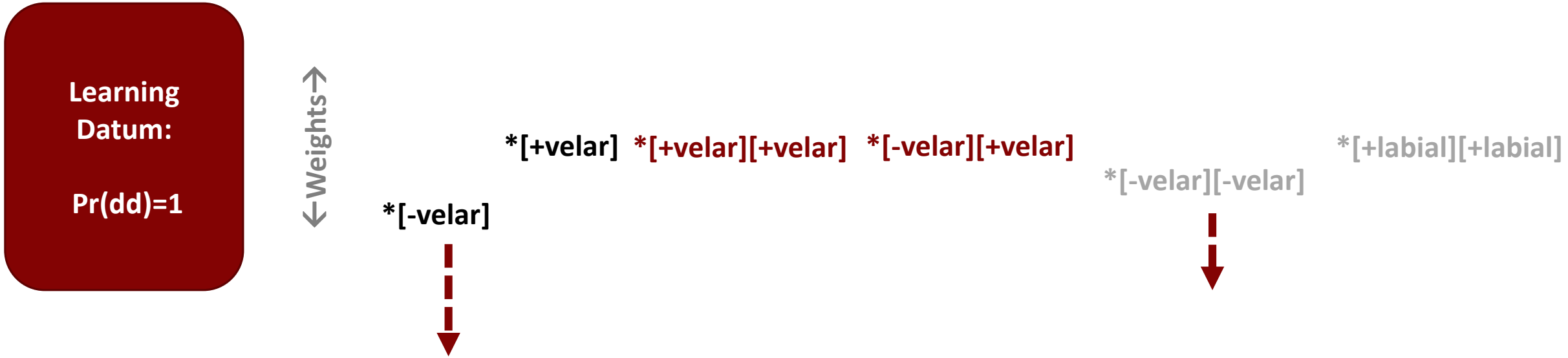
**Learning Datum:**

←Weights→

***[-velar]** ***[+velar]** ***[+velar][+velar]** ***[-velar][+velar]** *[-velar][-velar] *[+labial][+labial]

# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior without variables (Berent et al. 2012; Gallagher 2013).
  - The unigram [d] will gradually acquire more probability than the unigram [g], since it's attested.
  - Constraints penalizing [dg] and [gg] will have the same weights, since both bigrams are unattested.

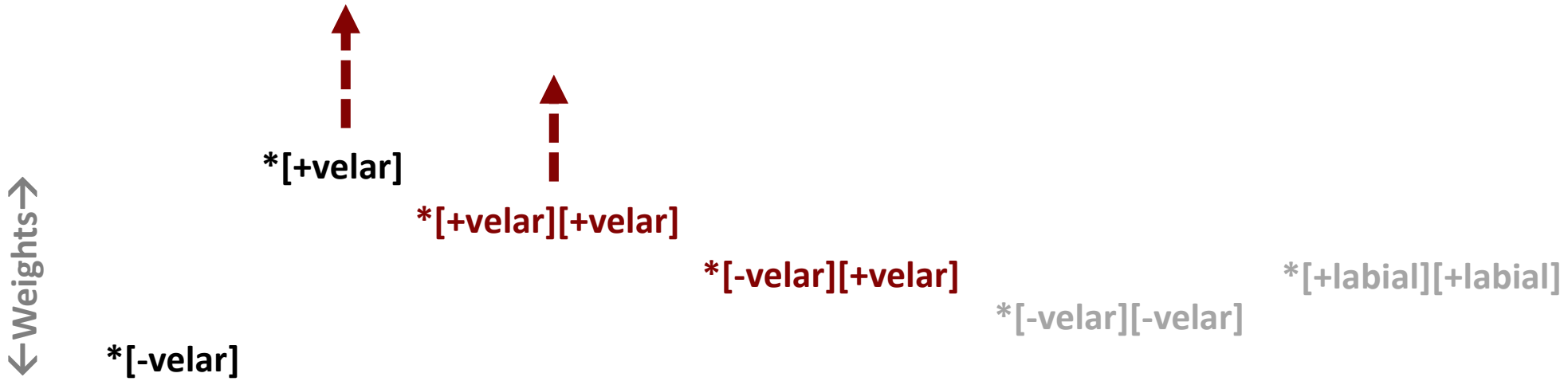


# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior without variables (Berent et al. 2012; Gallagher 2013).
  - The unigram [d] will gradually acquire more probability than the unigram [g], since it's attested.
  - Constraints penalizing [dg] and [gg] will have the same weights, since both bigrams are unattested.

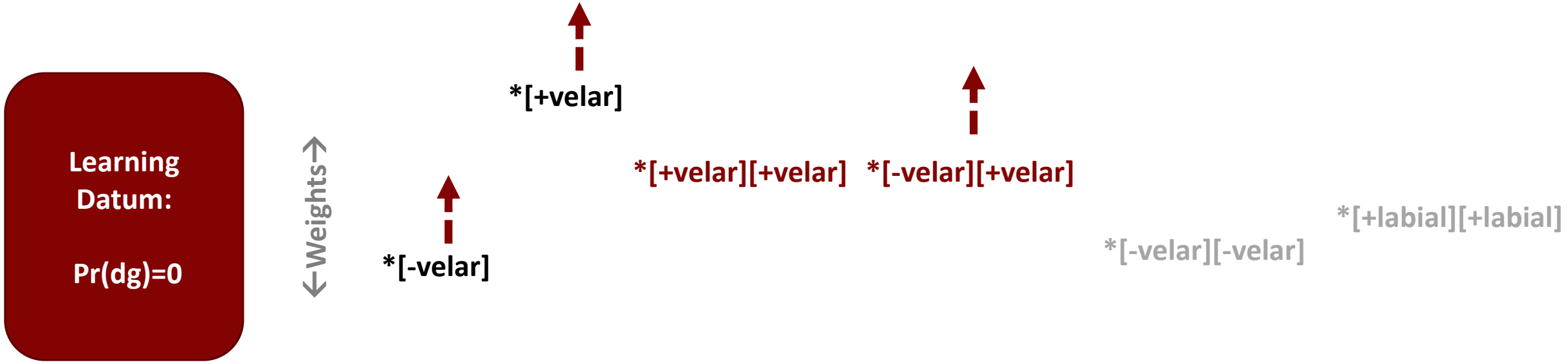
**Learning Datum:**

**Pr(gg)=0**



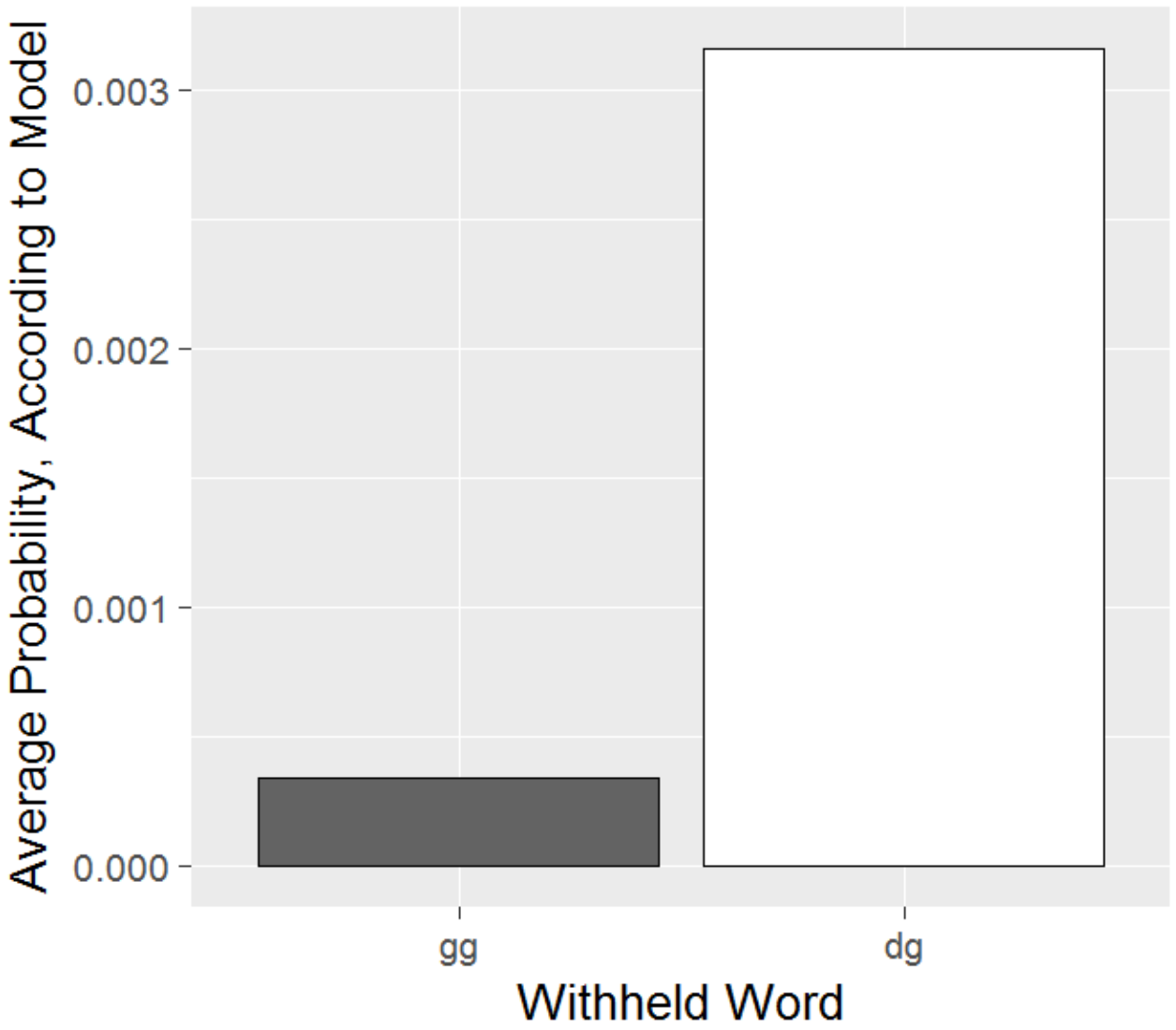
# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior without variables (Berent et al. 2012; Gallagher 2013).
  - The unigram [d] will gradually acquire more probability than the unigram [g], since it's attested.
  - Constraints penalizing [dg] and [gg] will have the same weights, since both bigrams are unattested.



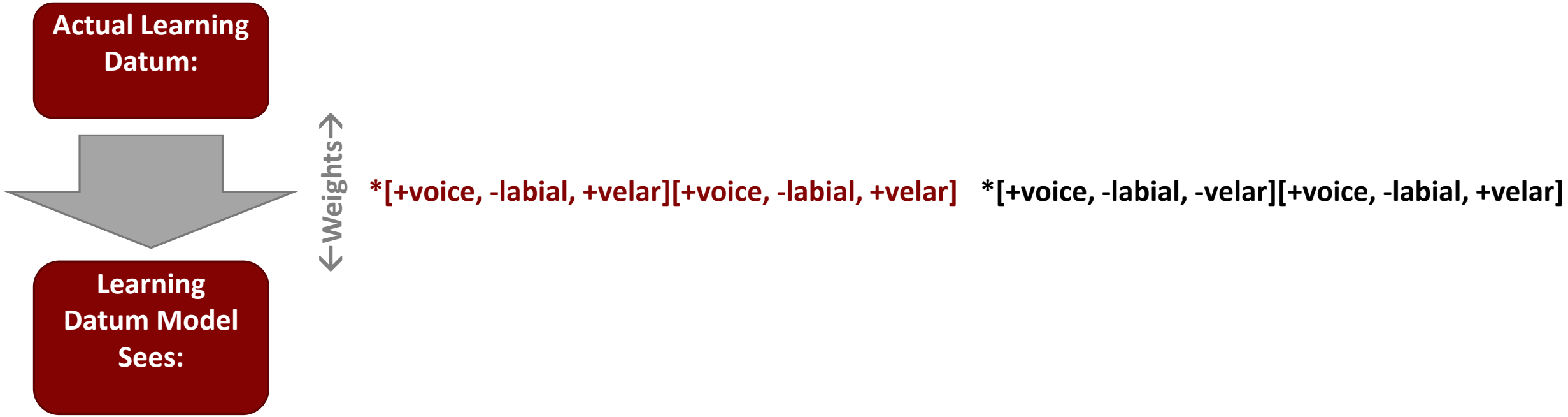
# Results with Vanilla GMECCS

- I replicated these results on the right using a model with no variables or PFA that was given training data based on Gallagher (2013)
- 200 epochs, learning rate of .01, weights initialized at zero, averaged over 25 runs



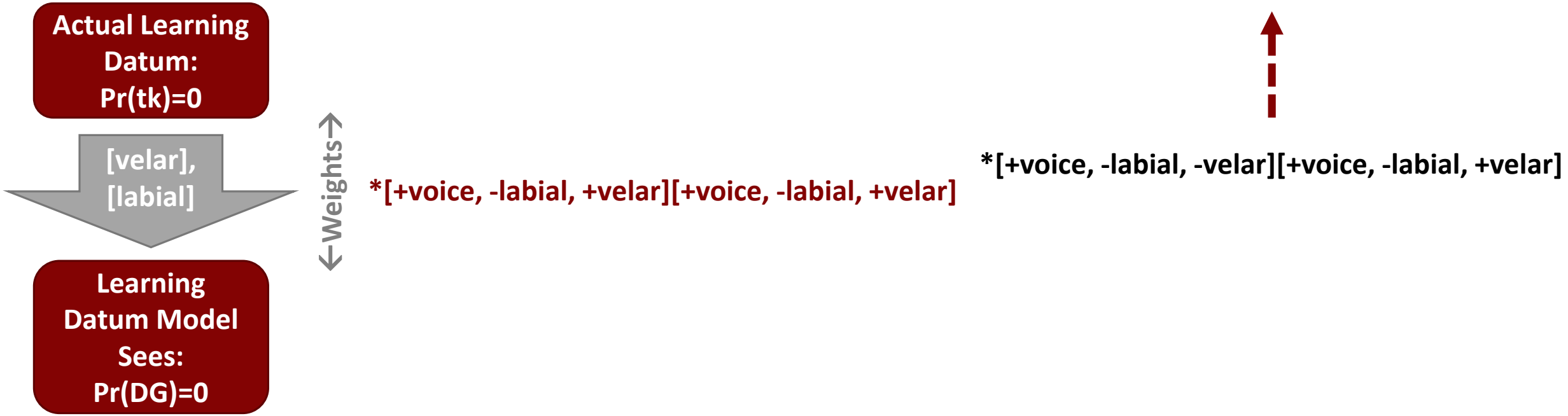
# Modeling with PFA

- Berent et al. (2012) and Gallagher (2013) show that variables successfully model this phenomenon. Can PFA?
- Over the course of learning, **[gg]** bigrams will hold onto more probability than **[dg]** bigrams, because the latter are more likely to become ambiguous with other unattested segments.



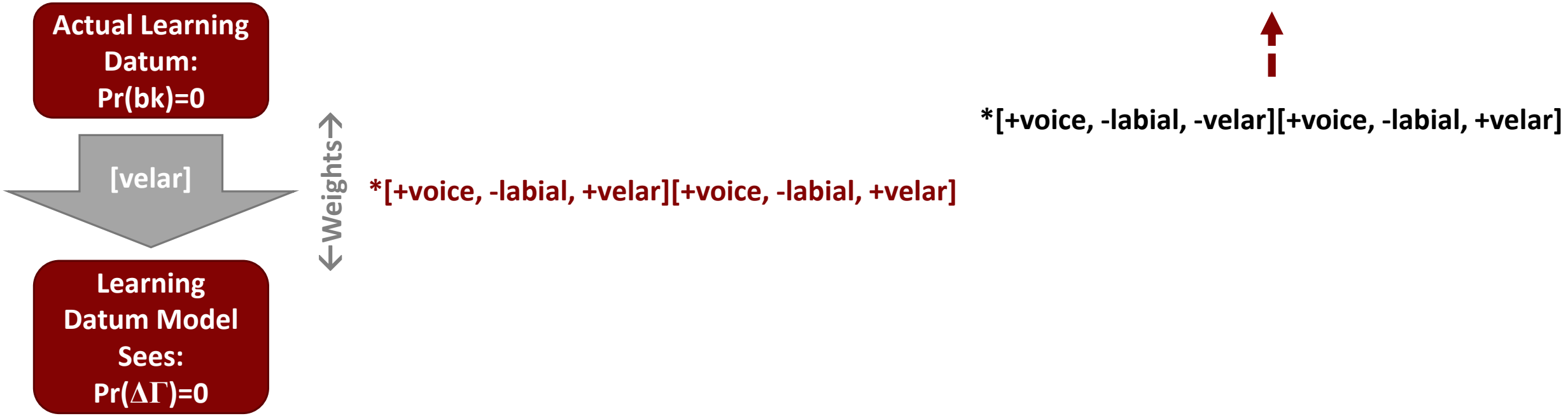
# Modeling with PFA

- Berent et al. (2012) and Gallagher (2013) show that variables successfully model this phenomenon. Can PFA?
- Over the course of learning, **[gg]** bigrams will hold onto more probability than **[dg]** bigrams, because the latter are more likely to become ambiguous with other unattested segments.



# Modeling with PFA

- Berent et al. (2012) and Gallagher (2013) show that variables successfully model this phenomenon. Can PFA?
- Over the course of learning, **[gg]** bigrams will hold onto more probability than **[dg]** bigrams, because the latter are more likely to become ambiguous with other unattested segments.

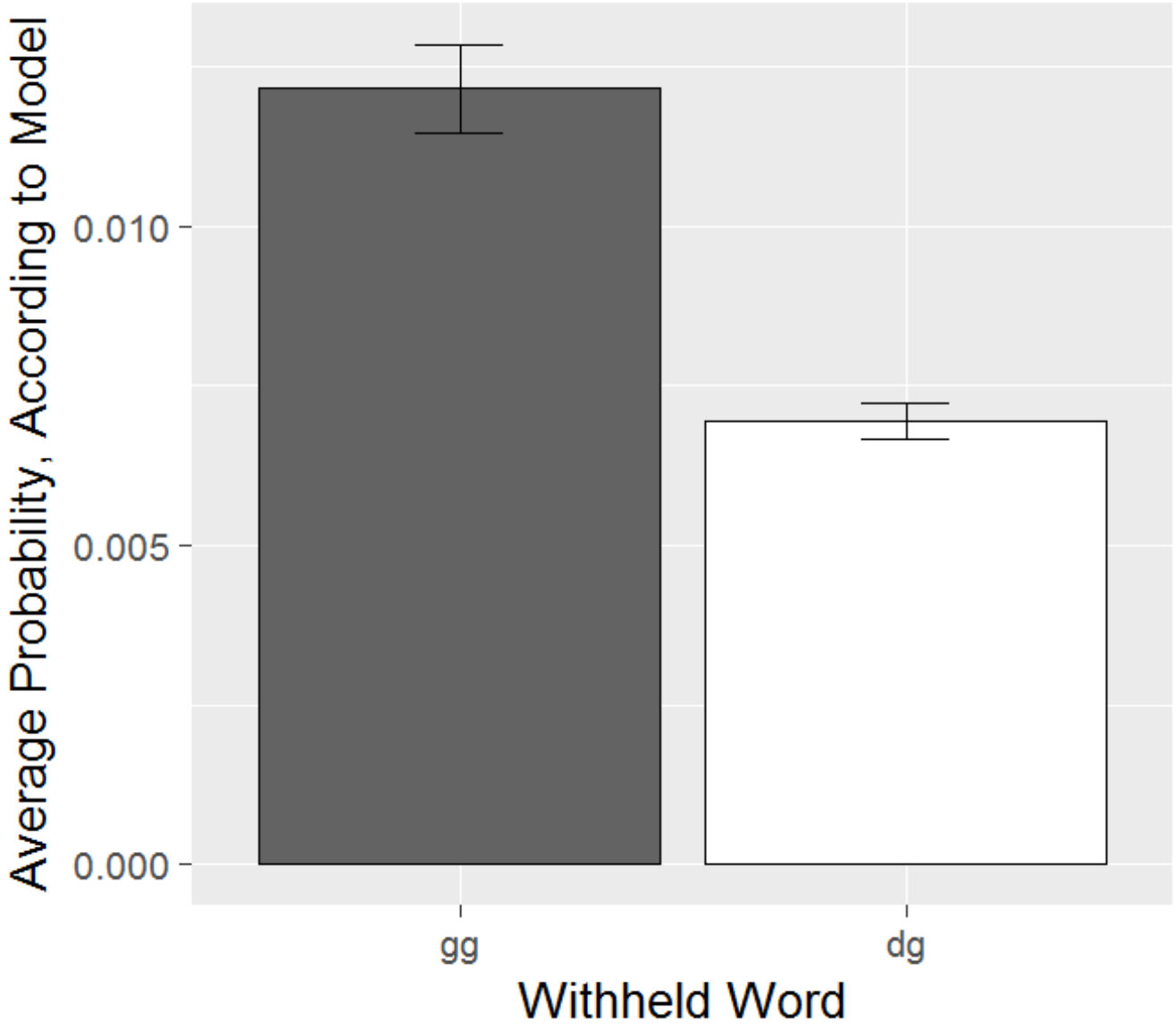


# Unattested bigrams

| [-velar][-velar] | | [+velar][-velar] | [-velar][+velar] | [+velar][+velar] |
|---|---|---|---|---|
| tt | pd | kt | tk | kk |
| td | pp | kd | tg | kg |
| tp | pb | kp | dk | gk |
| tb | bt | kb | **dg** | **gg** |
| dt | bd | gt | pk | |
| dp | bp | gd | pg | |
| db | | gp | bk | |
| pt | | gb | bg | |

# Results with PFA

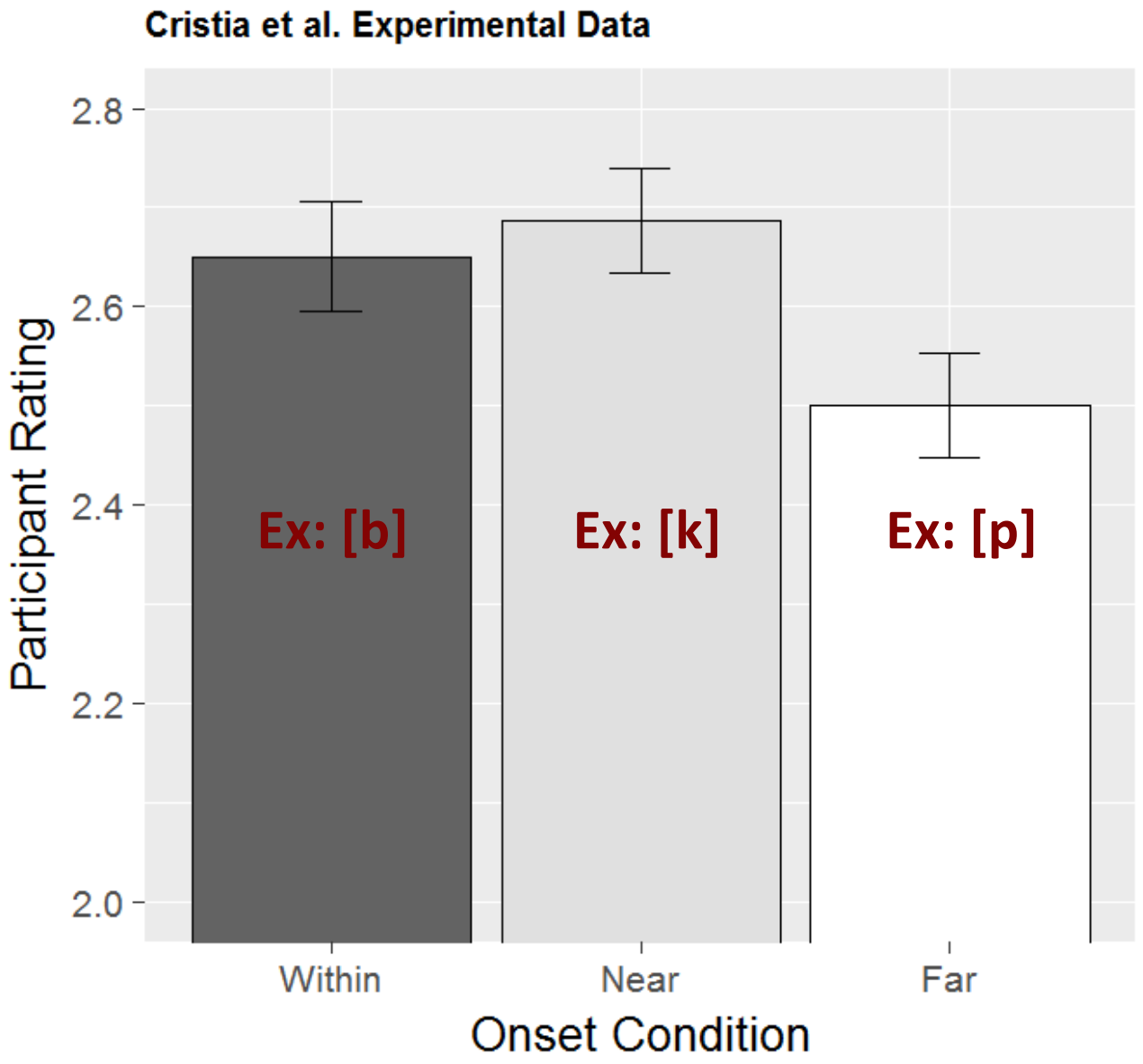
- The figure on the right shows the results for the PFA model trained on the pattern from Gallagher (2013).
- 200 epochs, learning rate of .01, weights initialized at zero, averaged over 25 runs, probability of attending to each feature, each epoch=.25



# Similarity-based Generalization

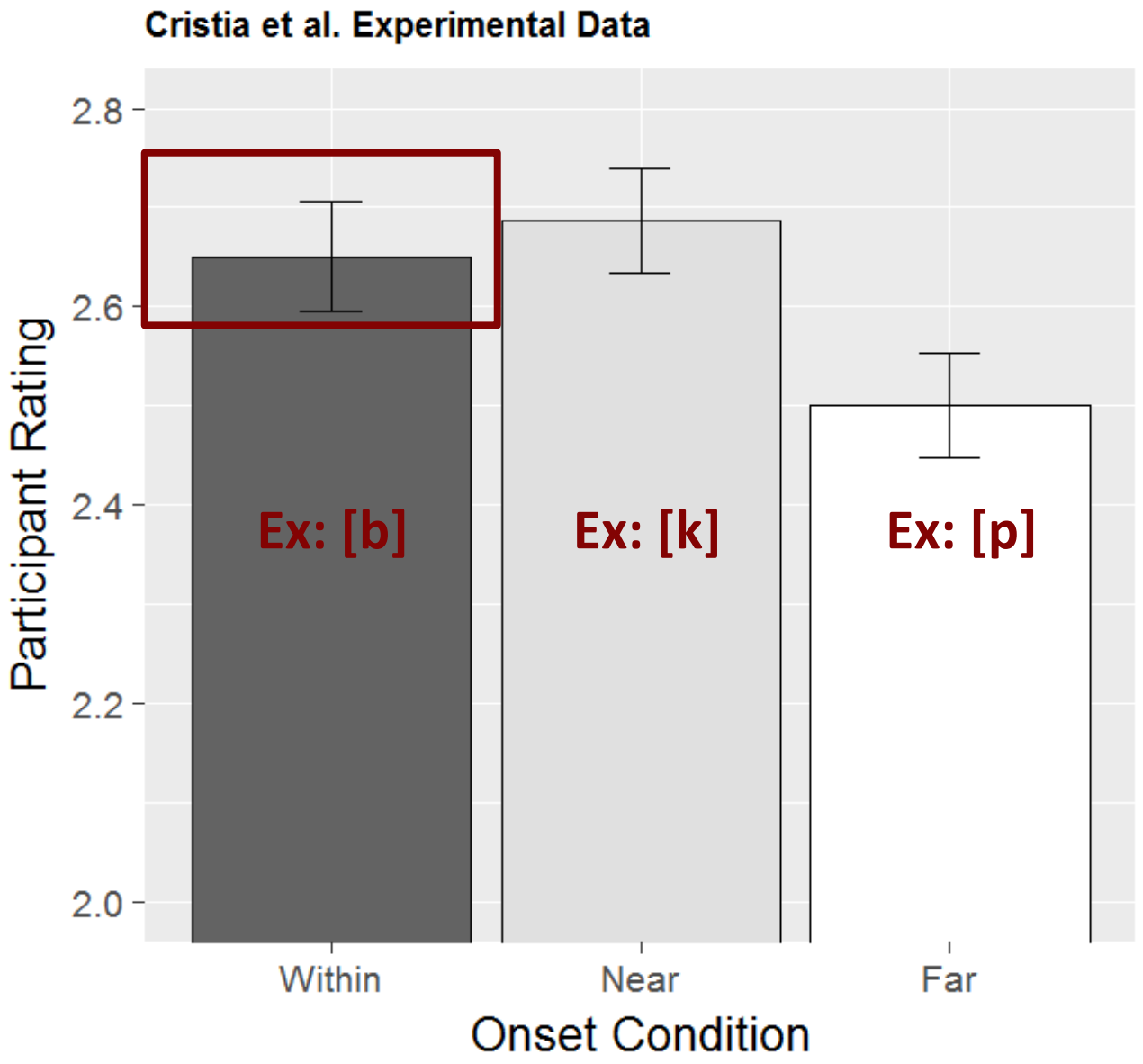
# The Phenomenon

- Cristia et al. (2013) observed ***Similarity-based Generalization*** when training subjects on an onset restriction.
  - For example, in training, subjects might have seen all onsets being [+voice].



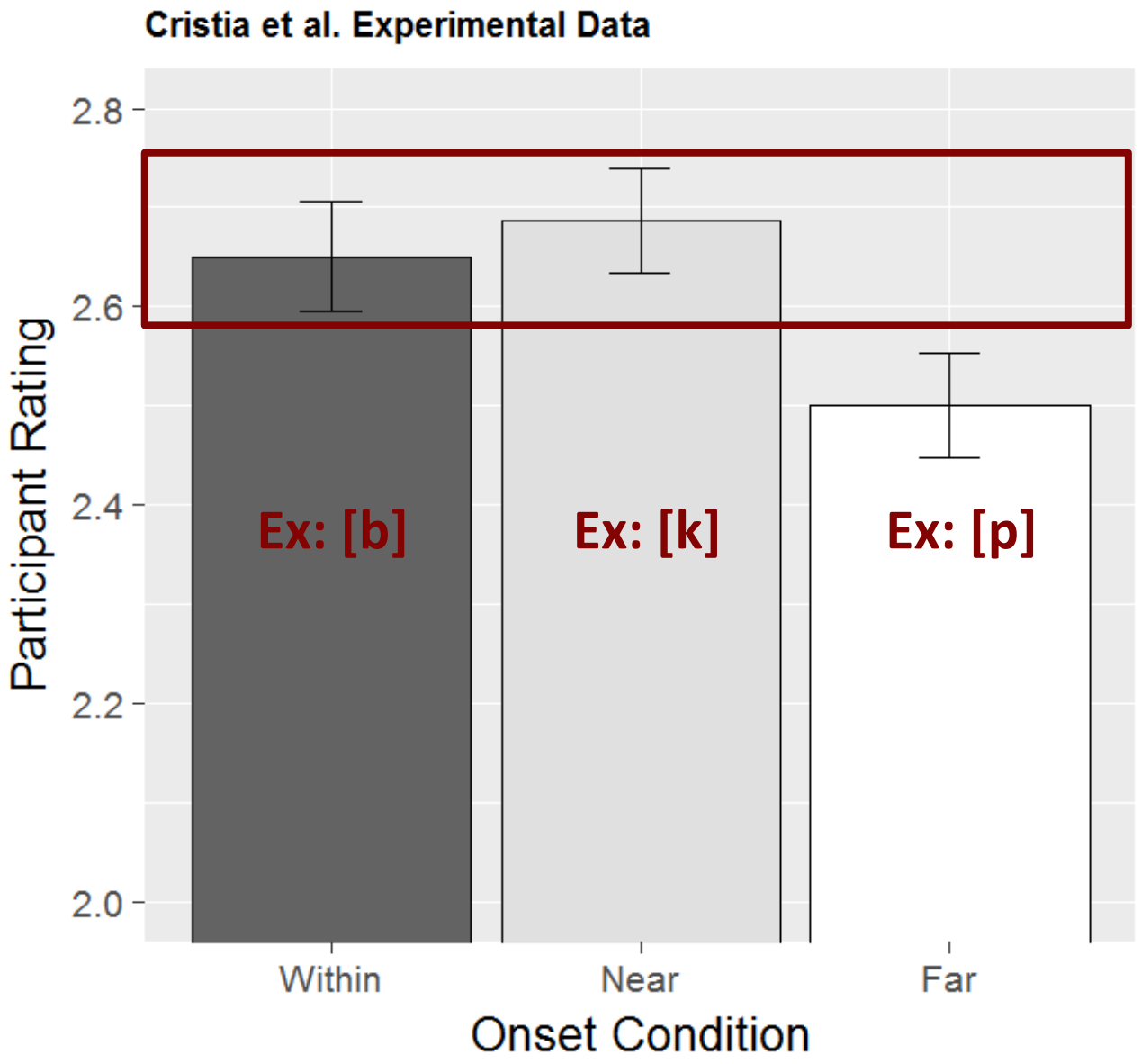
# The Phenomenon

- Cristia et al. (2013) observed ***Similarity-based Generalization*** when training subjects on an onset restriction.
  - For example, in training, subjects might have seen all onsets being [+voice].
  - In testing, subjects generalized to segments **within** the natural class of trained sounds…



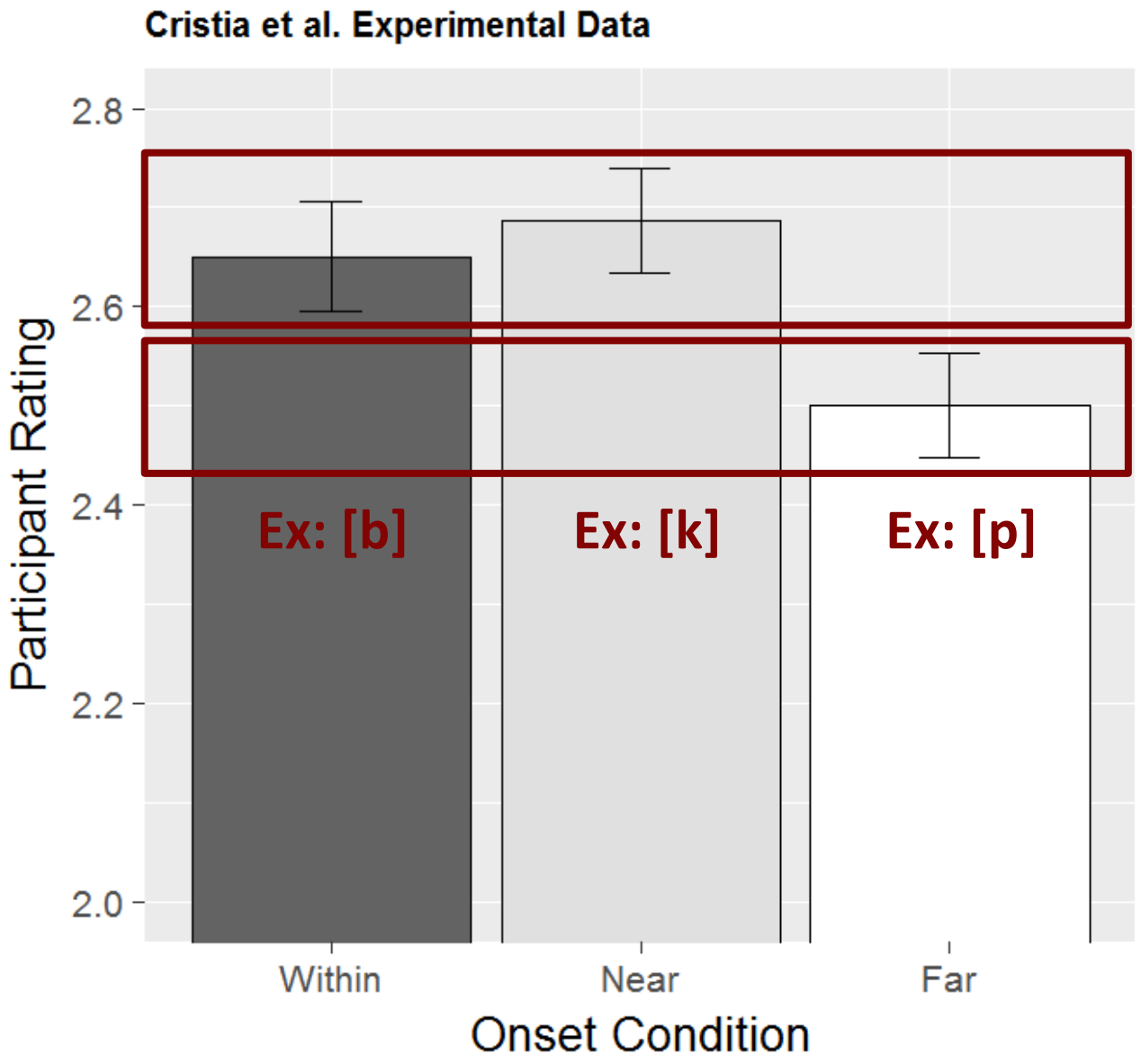
# The Phenomenon

- Cristia et al. (2013) observed ***Similarity-based Generalization*** when training subjects on an onset restriction.
  - For example, in training, subjects might have seen all onsets being [+voice].
  - In testing, subjects generalized to segments **within** the natural class of trained sounds…
  - …And those that were featurally **near** that natural class.



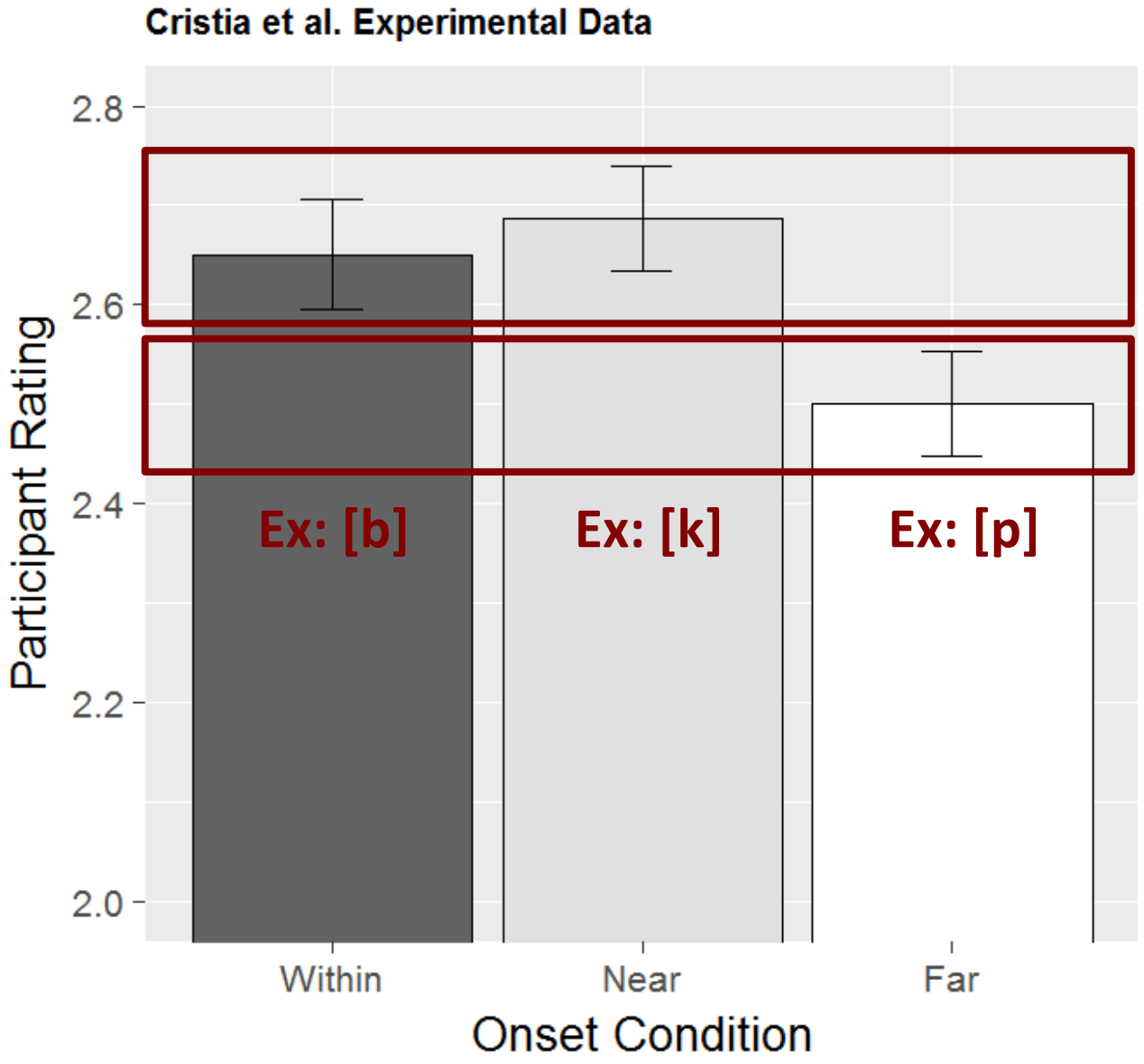
# The Phenomenon

- Cristia et al. (2013) observed ***Similarity-based Generalization*** when training subjects on an onset restriction.
  - For example, in training, subjects might have seen all onsets being [+voice].
  - In testing, subjects generalized to segments **within** the natural class of trained sounds…
  - …And those that were featurally **near** that natural class.



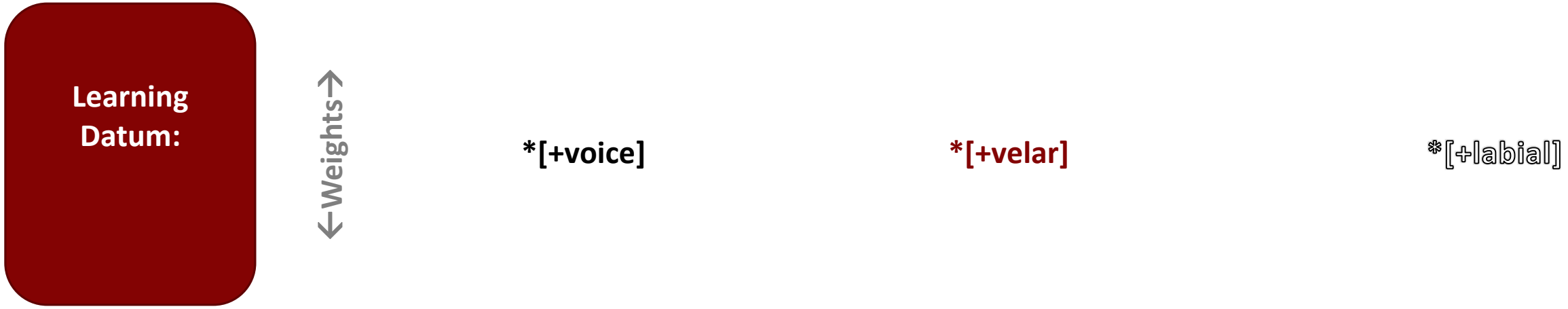
# The Phenomenon

- Cristia et al. (2013) observed ***Similarity-based Generalization*** when training subjects on an onset restriction.
  - For example, in training, subjects might have seen all onsets being [+voice].
  - In testing, subjects generalized to segments **within** the natural class of trained sounds…
  - …And those that were featurally **near** that natural class.
- An example of this in natural language could have happened in the dialect of German historically spoken in Schaffhausen (Mielke 2004).



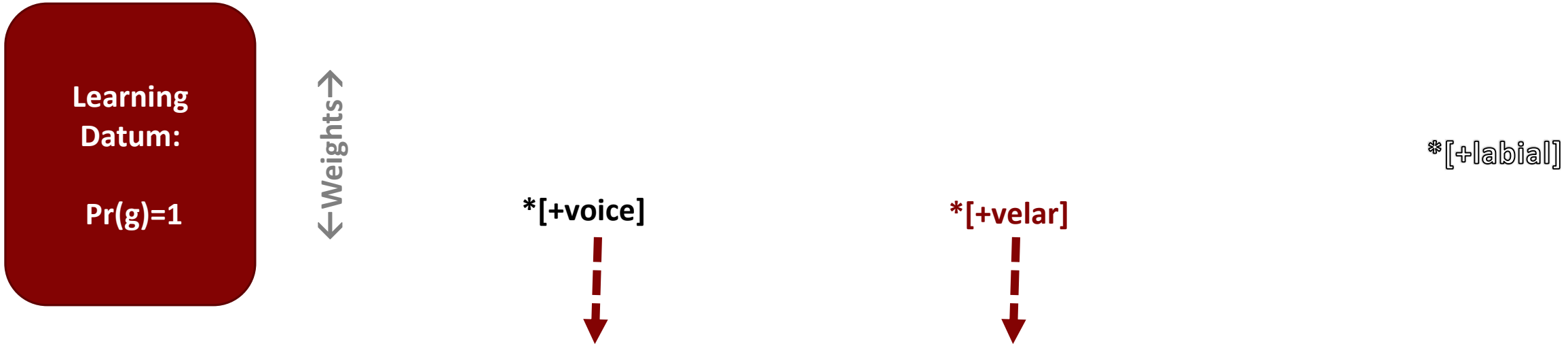
# Modeling with a vanilla MaxEnt model

- Since the model is maximizing the probability of the training data, any segments that are not a part of the training data will lose probability over the course of learning.
  - However, segments outside of natural classes that include the training data will lose probability especially quick, because of constraints that only refer to **single features**.
  - So Attested segments like [g] will get the most probability, followed by Within segments like **[b]**, Near segments like **[k]**, and Far segments like [p], respectively.



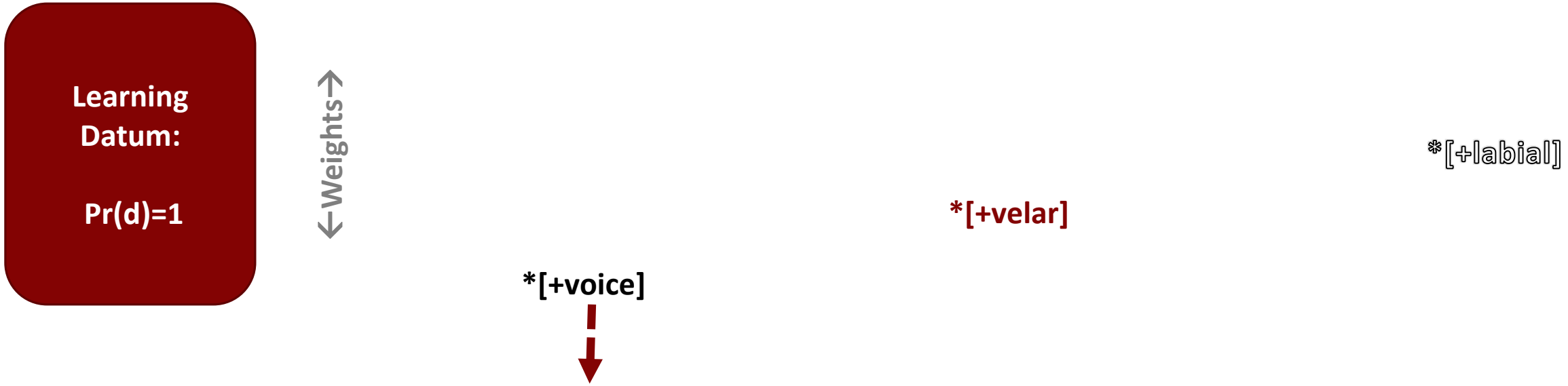
# Modeling with a vanilla MaxEnt model

- Since the model is maximizing the probability of the training data, any segments that are not a part of the training data will lose probability over the course of learning.
  - However, segments outside of natural classes that include the training data will lose probability especially quick, because of constraints that only refer to **single features**.
  - So Attested segments like [g] will get the most probability, followed by Within segments like **[b]**, Near segments like **[k]**, and Far segments like [p], respectively.



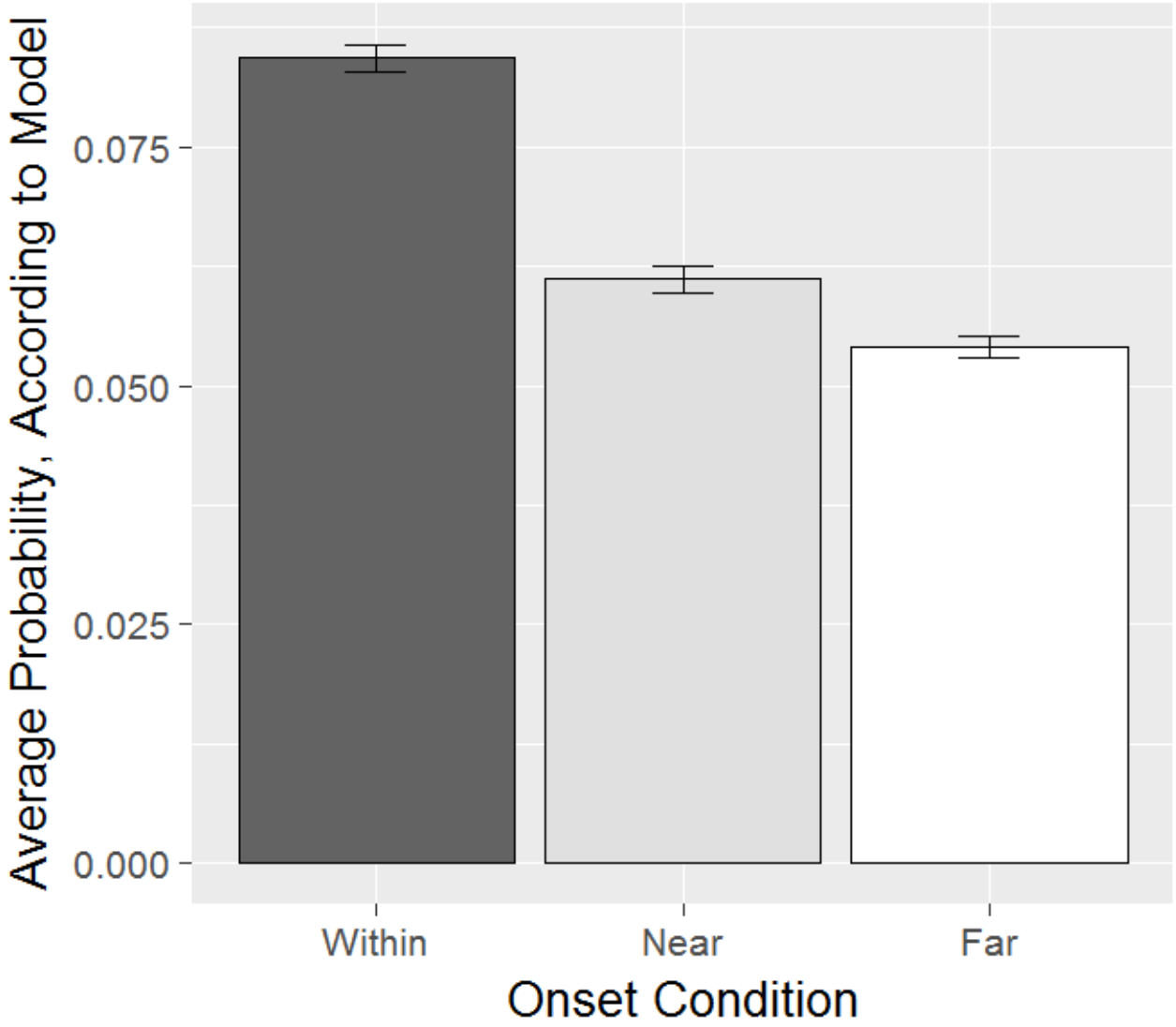
# Modeling with a vanilla MaxEnt model

- Since the model is maximizing the probability of the training data, any segments that are not a part of the training data will lose probability over the course of learning.
  - However, segments outside of natural classes that include the training data will lose probability especially quick, because of constraints that only refer to **single features**.
  - So Attested segments like [g] will get the most probability, followed by Within segments like **[b]**, Near segments like **[k]**, and Far segments like [p], respectively.



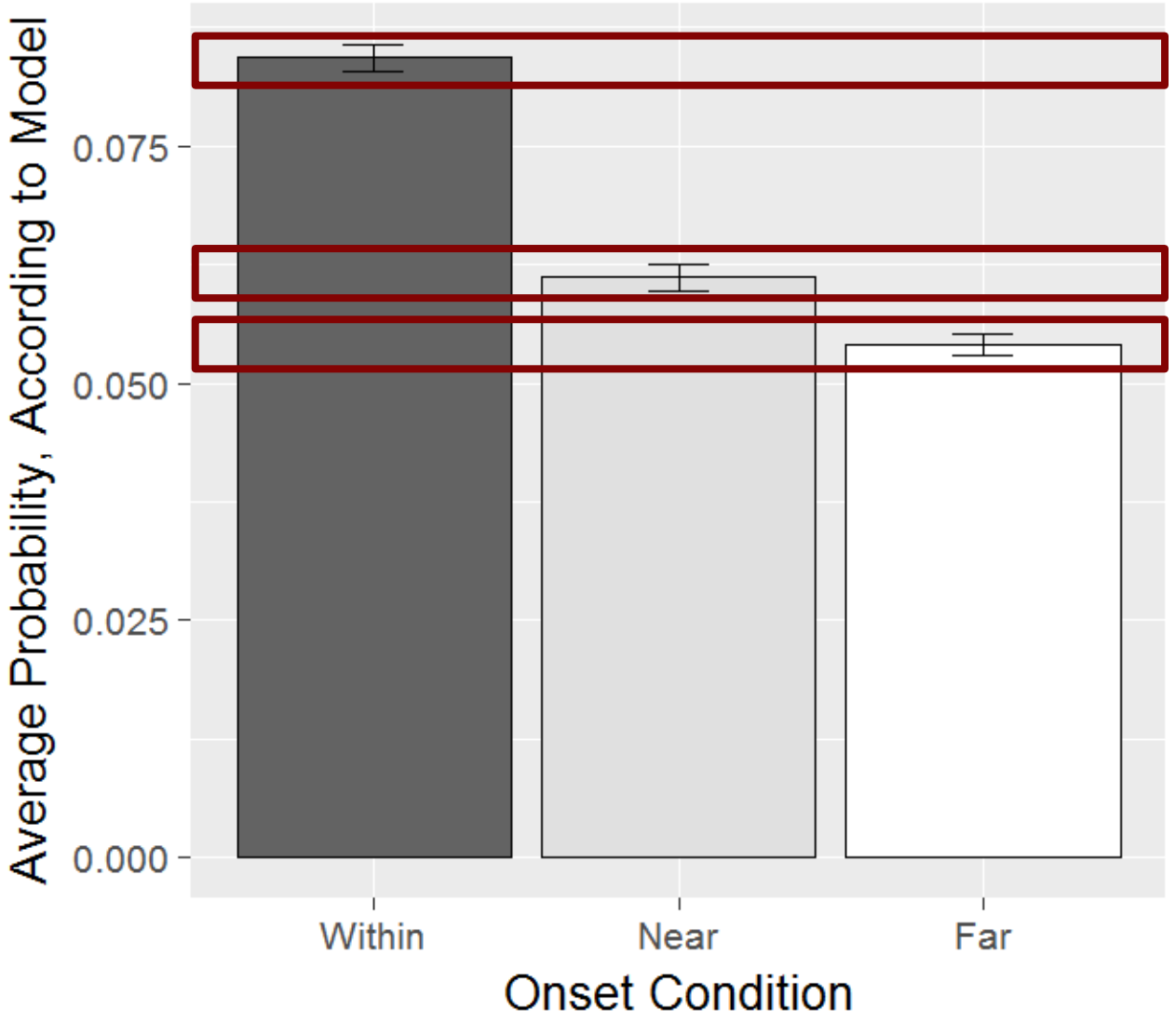
# Results with Vanilla GMECCS

- The results to the right are from the epoch out of 200 that was most correlated with Cristia et al.'s (2013) data.
  - Learning rate of .01, initial weights of 0, averaged over 10 runs per experiment condition.



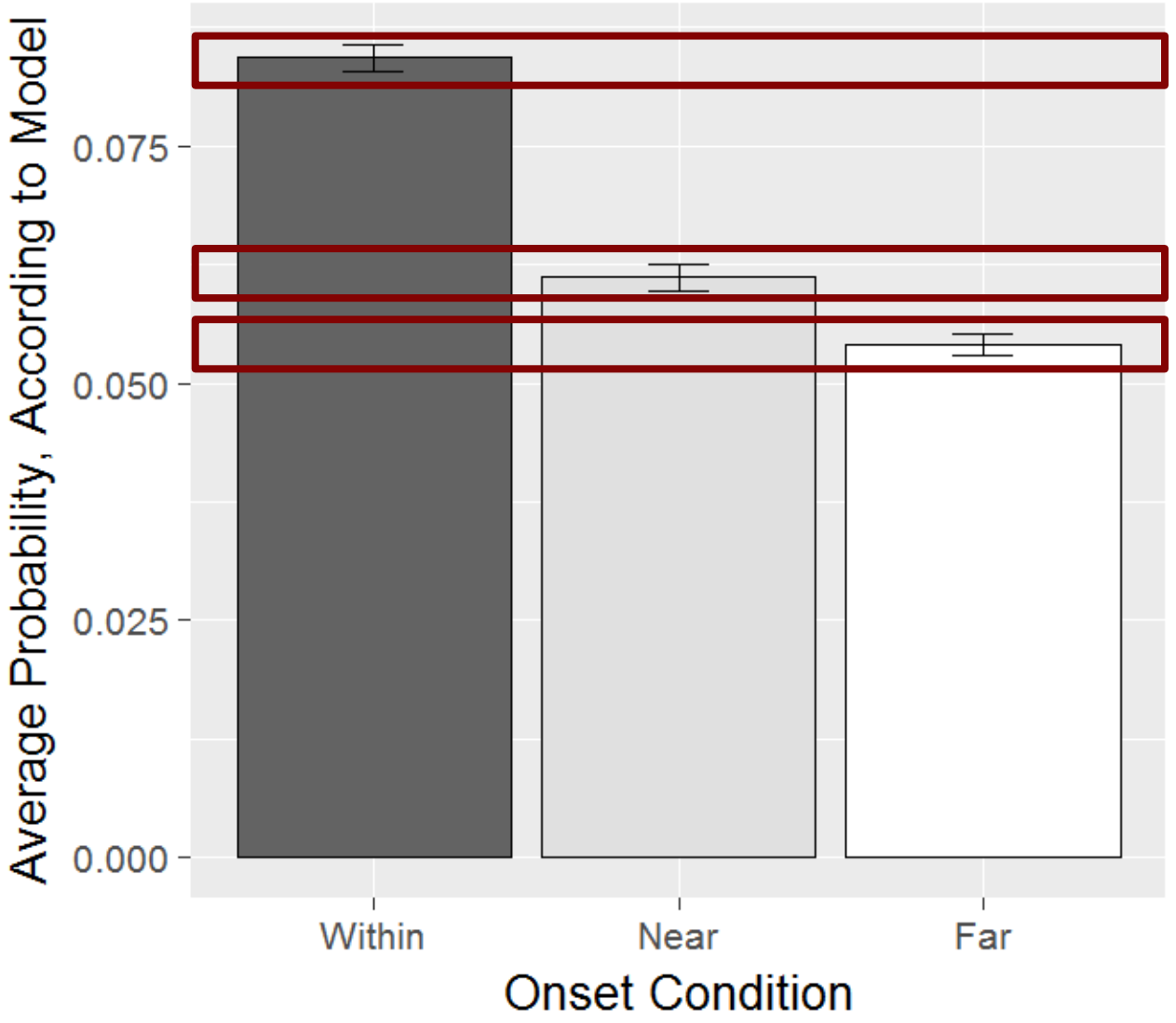
# Results with Vanilla GMECCS

- The results to the right are from the epoch out of 200 that was most correlated with Cristia et al.'s (2013) data.
  - Learning rate of .01, initial weights of 0, averaged over 10 runs per experiment condition.



# Results with Vanilla GMECCS

- The results to the right are from the epoch out of 200 that was most correlated with Cristia et al.'s (2013) data.
  - Learning rate of .01, initial weights of 0, averaged over 10 runs per experiment condition.
- Variables don't make any changes to this behavior, since they have to be restricted to occurring across segments (Moreton 2012).
  - If they could happen within single segments, then relatively weird patterns, like *#[αVoice, αStrident] would be possible.

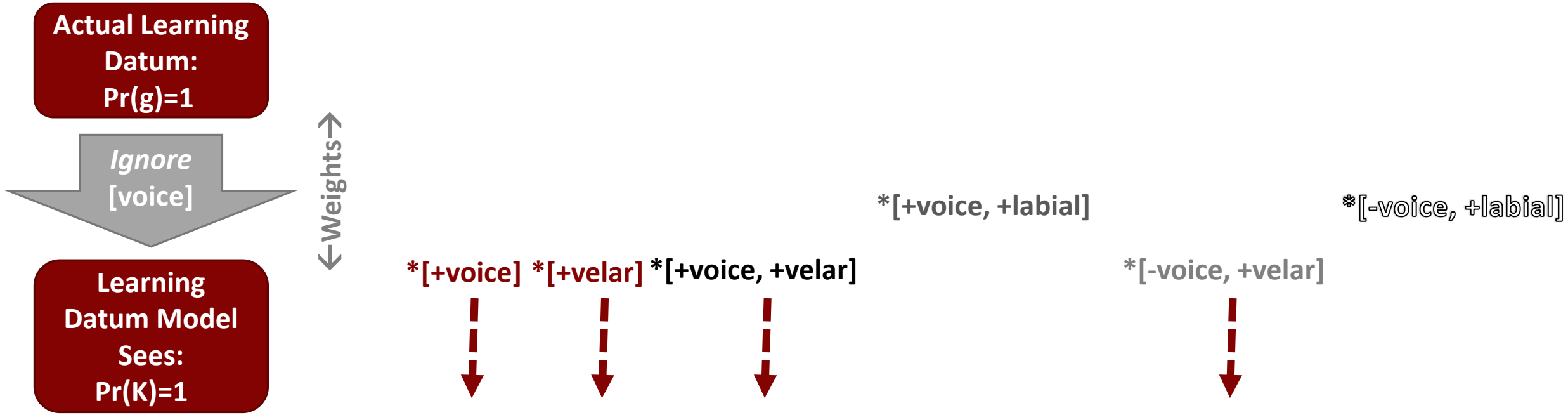


# Modeling with PFA

- Over the course of learning, Near segments like **[k]** are more likely to be ambiguous with Attested segments like [g] than Far segments like [p] are.
- This means that Within segments like **[b]** and Near segments will behave similarly.

# Modeling with PFA

- Over the course of learning, Near segments like [k] are more likely to be ambiguous with Attested segments like [g] than Far segments like [p].
- This means that Within segments like **[b]** and Near segments will behave similarly.

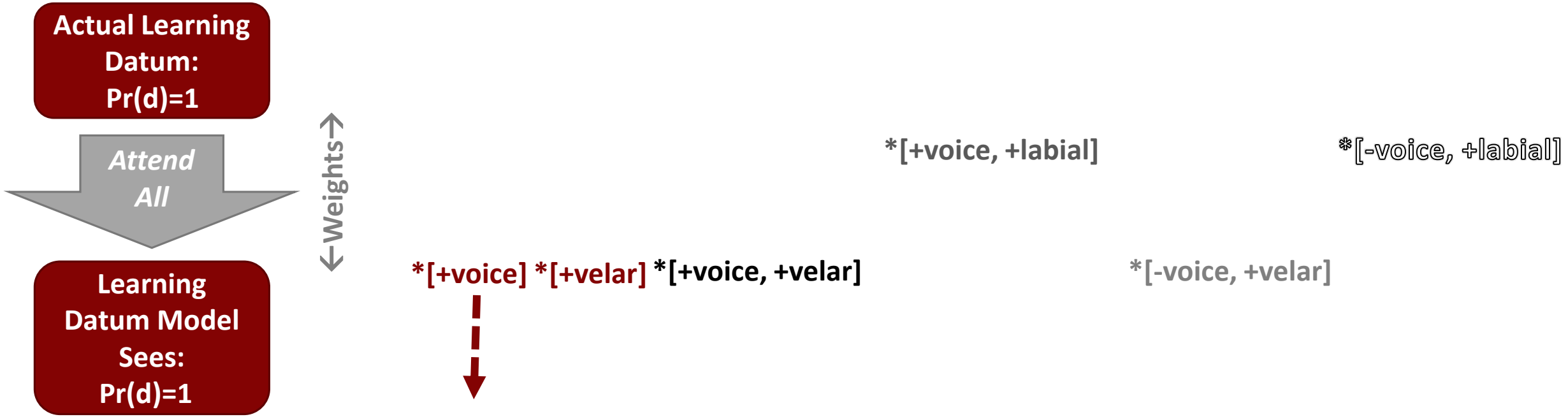


# Modeling with PFA

- Over the course of learning, Near segments like [k] are more likely to be ambiguous with Attested segments like [g] than Far segments like [p].
- This means that Within segments like **[b]** and Near segments will behave similarly.

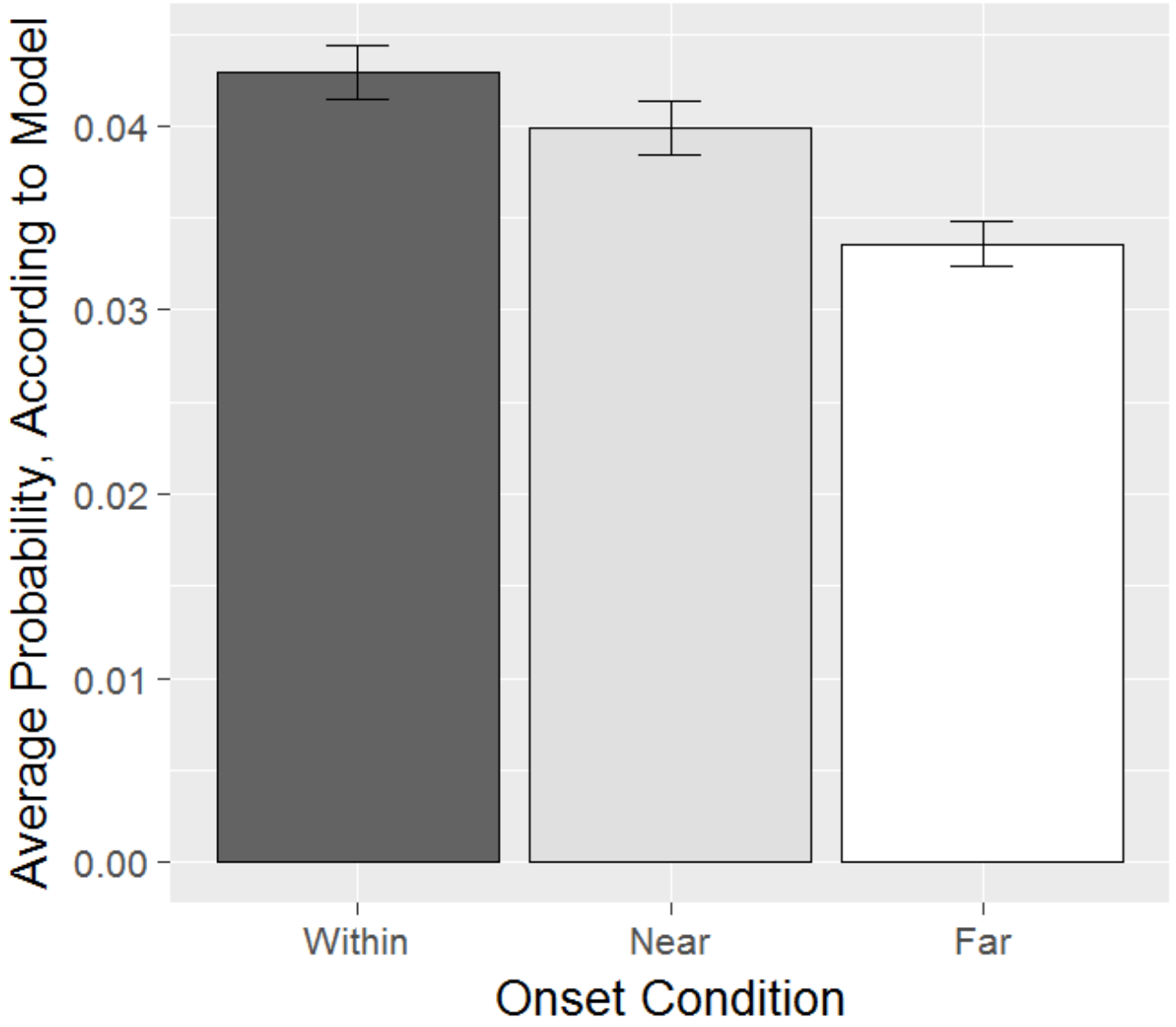
# Modeling with PFA

- Over the course of learning, Near segments like [k] are more likely to be ambiguous with Attested segments like [g] than Far segments like [p].
- This means that Within segments like **[b]** and Near segments will behave similarly.



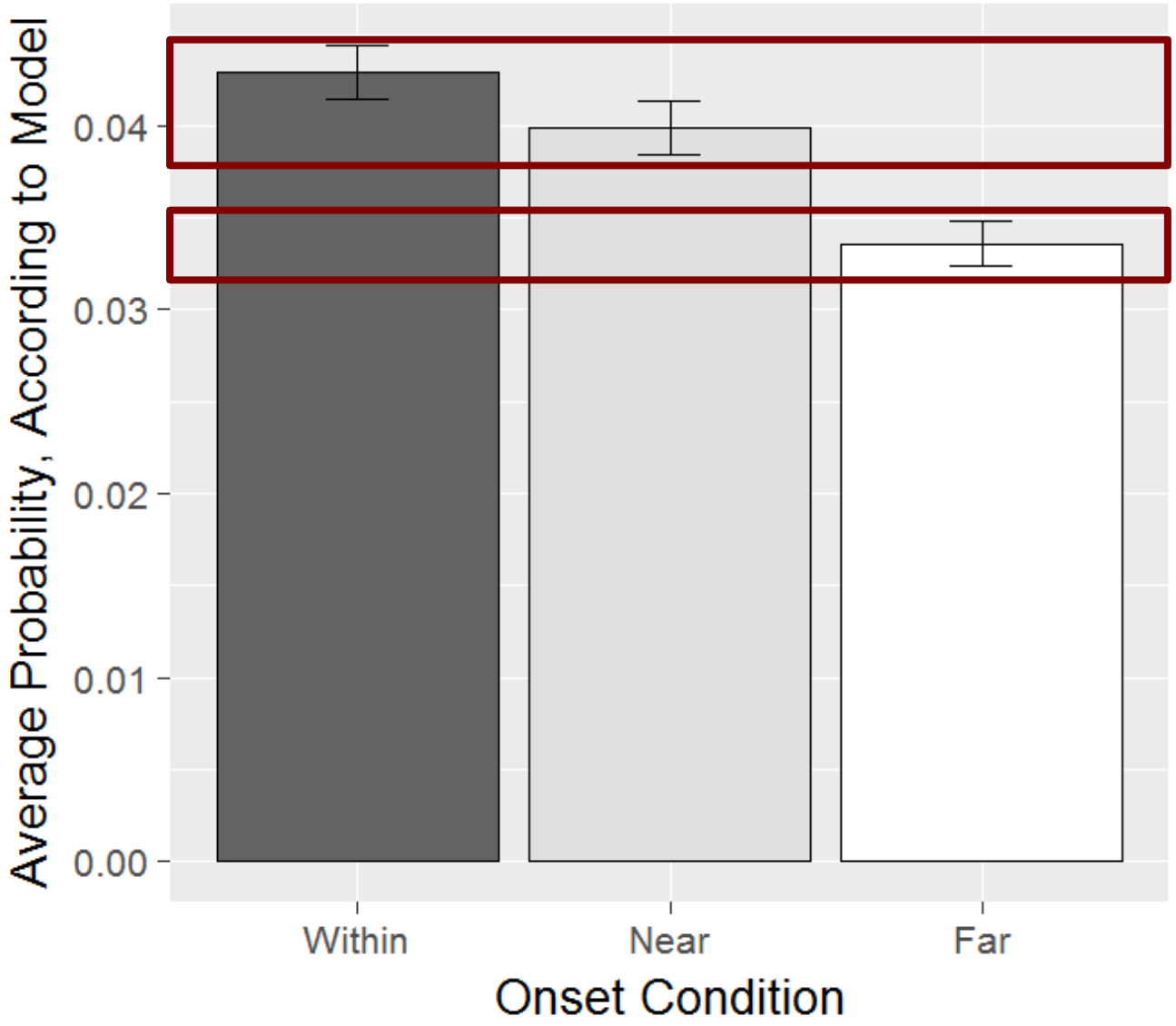
# Modeling with PFA

- The results to the right are from the epoch out of 200 that was most correlated with Cristia et al.'s (2013) data.
  - Learning rate of .01, initial weights of 0, averaged over 10 runs per experiment condition, probability of attending to each feature, each epoch=.25



# Modeling with PFA

- The results to the right are from the epoch out of 200 that was most correlated with Cristia et al.'s (2013) data.
  - Learning rate of .01, initial weights of 0, averaged over 10 runs per experiment condition, probability of attending to each feature, each epoch=.25



# Discussion

# Future work

- This presentation has been focused on generalization, rather than learning biases. Exploring the latter is an important next step.
  - I already have results for Intradimensional Bias (Moreton 2012) and Identity Bias (Gallagher 2013) showing that PFA predicts both.
  - What about other biases that haven't been attributed to variables?

# Future work

- This presentation has been focused on generalization, rather than learning biases. Exploring the latter is an important next step.
  - I already have results for Intradimensional Bias (Moreton 2012) and Identity Bias (Gallagher 2013) showing that PFA predicts both.
  - What about other biases that haven't been attributed to variables?
- What about overpredictions of PFA and variables? Are there any crazy patterns that either theory predicts to be more easily learned?

# Future work

- This presentation has been focused on generalization, rather than learning biases. Exploring the latter is an important next step.
  - I already have results for Intradimensional Bias (Moreton 2012) and Identity Bias (Gallagher 2013) showing that PFA predicts both.
  - What about other biases that haven't been attributed to variables?
- What about overpredictions of PFA and variables? Are there any crazy patterns that either theory predicts to be more easily learned?
- PFA worked in the domain of phonotactics. It'd be interesting to apply it to other domains where identity functions have been shown to be crucial, such as reduplication (Marcus et al. 1999).

# Future work

- This presentation has been focused on generalization, rather than learning biases. Exploring the latter is an important next step.
  - I already have results for Intradimensional Bias (Moreton 2012) and Identity Bias (Gallagher 2013) showing that PFA predicts both.
  - What about other biases that haven't been attributed to variables?
- What about overpredictions of PFA and variables? Are there any crazy patterns that either theory predicts to be more easily learned?
- PFA worked in the domain of phonotactics. It'd be interesting to apply it to other domains where identity functions have been shown to be crucial, such as reduplication (Marcus et al. 1999).
- Simulating natural language data
  - Do children acquiring phonology make mistakes that could distinguish between these two theories?
  - What about diachronic changes like the one in German? Are there more facts about language change that can be explained better by PFA than more standard models of phonotactic learning?
  - And what happens when PFA is applied to larger datasets, like the data used by Berent et al. (2012) to model identity phonotactics in Hebrew?

# Conclusion

- In the past, variables were said to be the only way to properly model Identity Generalization (Marcus 2001; Berent et al. 2012; Berent 2013; Gallagher 2013).
- Here I showed that this isn't the case—by introducing error in how the model's training data is represented, you can get it to generalize and learn in ways that correctly predict human behavior.
- And on top of that, PFA predicts Similarity-based Generalization, giving a unified account of both phenomena.

# References


Berent, I., Dupuis, A., & Brentari, D. (2014). Phonological reduplication in sign language: Rules rule. *Frontiers in psychology*, *5*, 560.

Berent, I., Wilson, C., Marcus, G. F., & Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic inquiry*, *43*(1), 97-119.

Berent, I. (2013). The phonological mind. *Trends in cognitive sciences*, *17*(7), 319-327.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2), 259-285.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577-614.

Gallagher, G. (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, *30*(2), 253-295.

Gallagher, G. (2014). An identity bias in phonotactics: Evidence from Cochabamba Quechua. *Laboratory Phonology*, *5*(3), 337-378.

Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27, 166–195*. doi:10.1016/0749-596X(88)90072-1

Halle, M. (1962). A descriptive convention for treating assimilation and dissimilation. *MIT Research Laboratory of Electronics Quarterly Progress Report 66(XVIII)*, 295–296.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379-440.

Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.

Mielke, J. (2004). *The emergence of distinctive features* (Doctoral dissertation, The Ohio State University).

Moreton, E. (2012). Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165-183.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Pater, J., & Moreton, E. (2014). Structurally biased phonology: complexity in learning and typology. *The EFL Journal*, *3*(2).

# Appendix

# Differences with the Original GMECCS

- There are a few marginal differences between my implementation of GMECCS and the original:
  - Constraints that don't refer to any relevant segments (e.g. [+continuant, +velar]) aren't included.
  - When there are pairs of constraints that are violated by the exact same class of relevant words (e.g. *[+labial, -velar] and *[+labial]) only one is included in the model.
  - I don't include a version of each unigram constraint for each segmental position (e.g. $*[+\text{voice}]_{\text{first segment}}$ and $*[+\text{voice}]_{\text{second segment}}$).
- A more meaningful difference is that I'll be training my model online (1 datum at a time), rather than in batch (where all data are presented at every weight update).
  - This is to demonstrate the effects of Probabilistic Feature Attention in a more believable way, although I've run all of the simulations presented here in batch and get the same results.
- And, of course, I'll be showing results from simulations with and without **Probabilistic Feature Attention** to demonstrate its effects on phonotactic learning.

# Modeling Ident Generalization with Variables

- When MaxEnt models are given variables, they're able to model Identity Generalization (Berent et al. 2012; Gallagher 2013).
  - E.g. adding the constraints *[α][α],*[¬α][α], and *[α][¬α] for the pattern from Gallagher (2013).
  - Since the constraints above aren't violated by any of the words in the training data and *are* violated by a majority of the words outside of the training data, they receive a high weight.

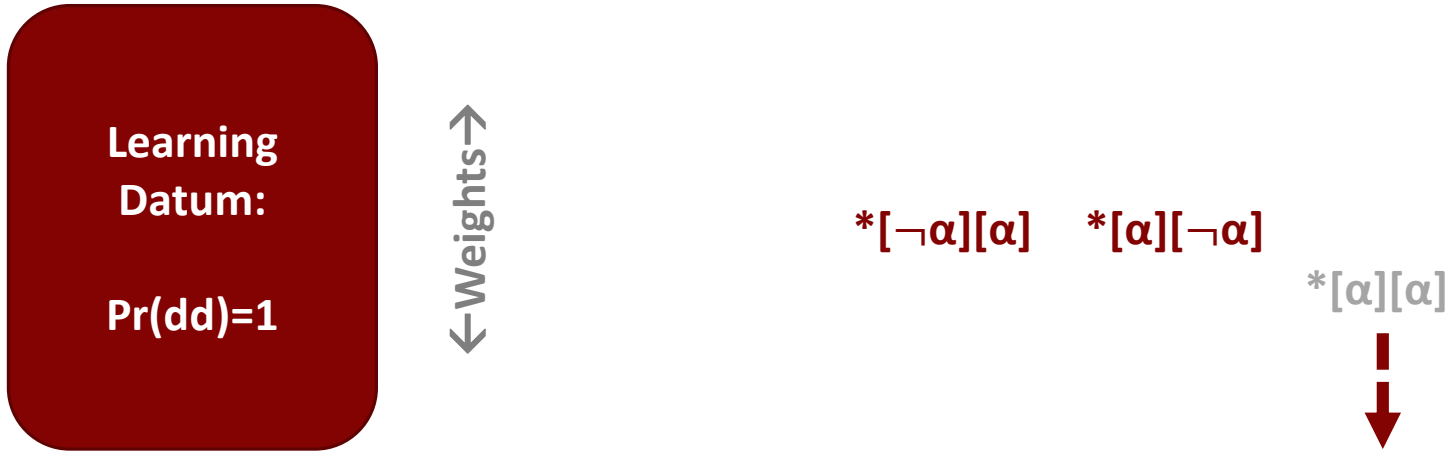
**Learning Datum:**

←Weights→

***[¬α][α]** ***[α][¬α]** *[α][α]

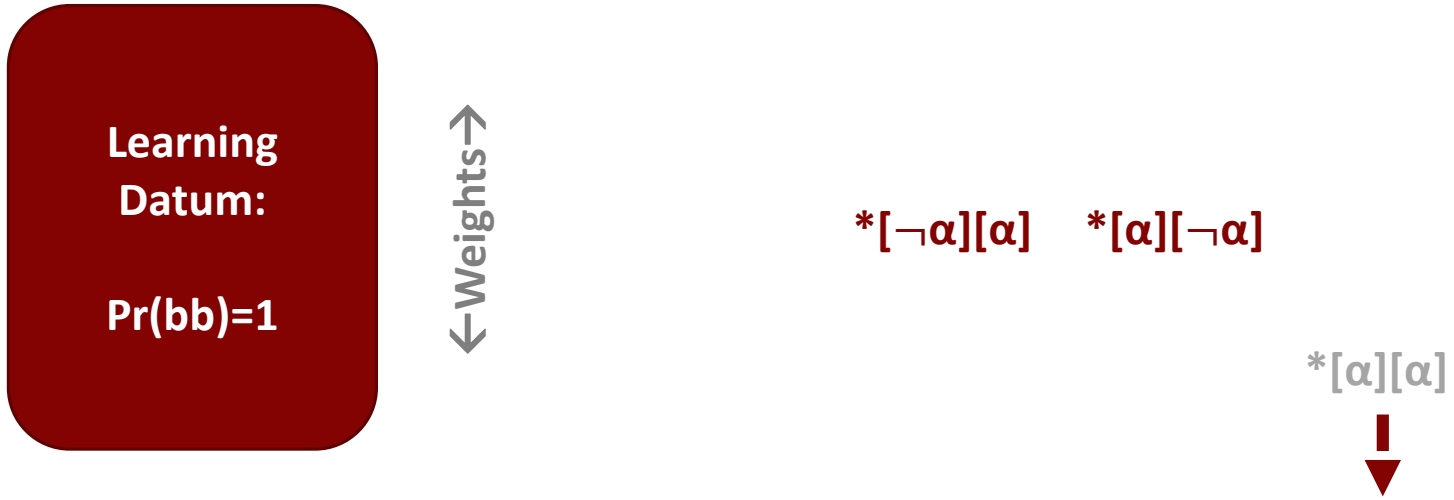
# Modeling Ident Generalization with Variables

- When MaxEnt models are given variables, they're able to model Identity Generalization (Berent et al. 2012; Gallagher 2013).
  - E.g. adding the constraints *[α][α],*[¬α][α], and *[α][¬α] for the pattern from Gallagher (2013).
  - Since the constraints above aren't violated by any of the words in the training data and *are* violated by a majority of the words outside of the training data, they receive a high weight.



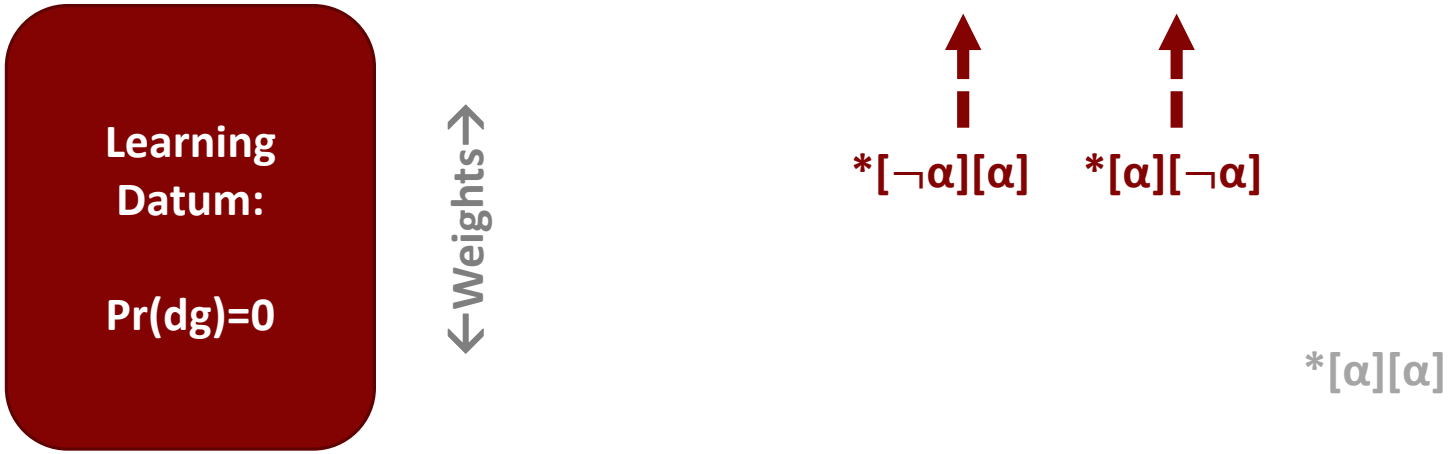
# Modeling Ident Generalization with Variables

- When MaxEnt models are given variables, they're able to model Identity Generalization (Berent et al. 2012; Gallagher 2013).
  - E.g. adding the constraints *[α][α],*[¬α][α], and *[α][¬α] for the pattern from Gallagher (2013).
  - Since the constraints above aren't violated by any of the words in the training data and *are* violated by a majority of the words outside of the training data, they receive a high weight.



# Modeling Ident Generalization with Variables

- When MaxEnt models are given variables, they're able to model Identity Generalization (Berent et al. 2012; Gallagher 2013).
  - E.g. adding the constraints *[α][α],*[¬α][α], and *[α][¬α] for the pattern from Gallagher (2013).
  - Since the constraints above aren't violated by any of the words in the training data and *are* violated by a majority of the words outside of the training data, they receive a high weight.

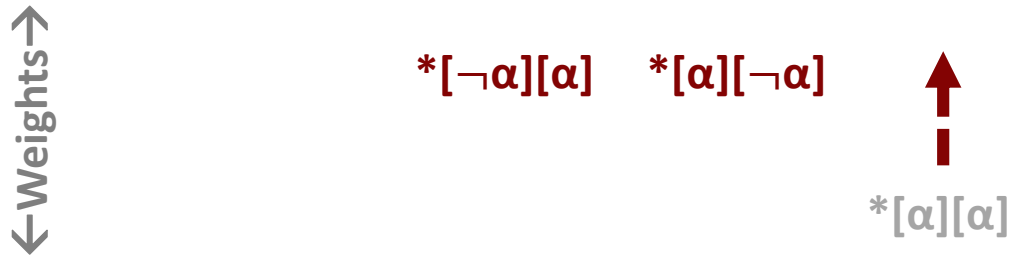


# Modeling Ident Generalization with Variables

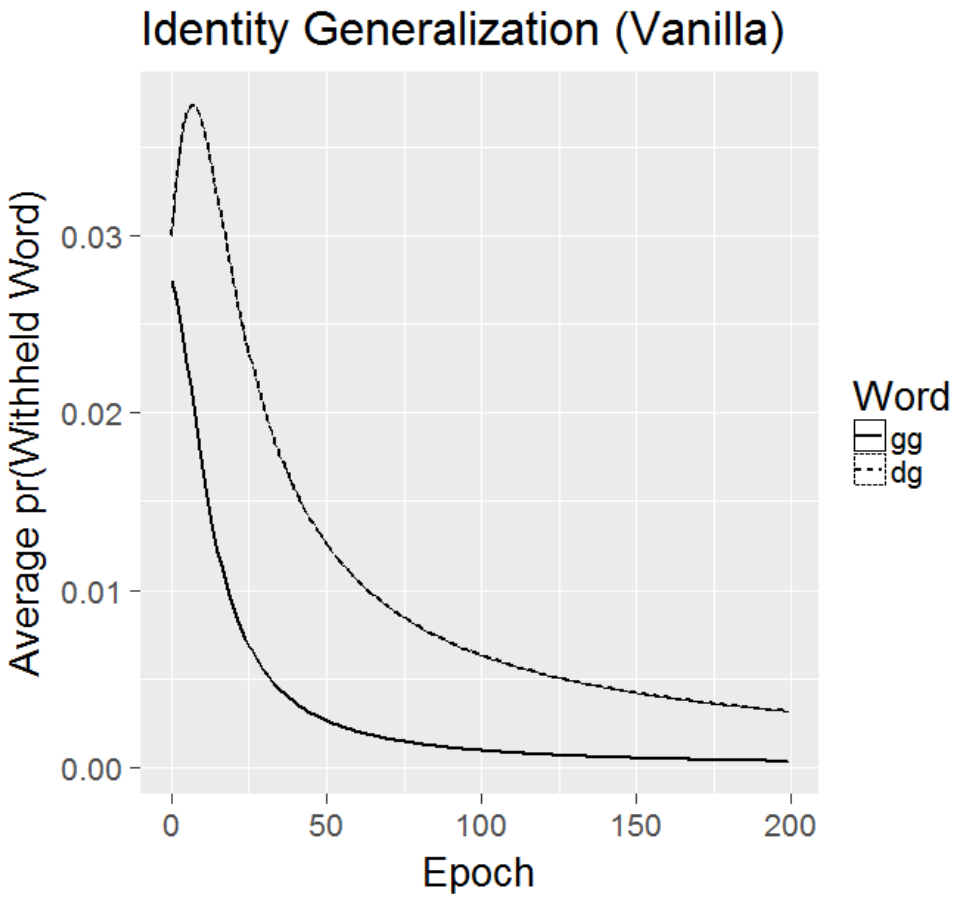
- When MaxEnt models are given variables, they're able to model Identity Generalization (Berent et al. 2012; Gallagher 2013).
  - E.g. adding the constraints *[α][α],*[¬α][α], and *[α][¬α] for the pattern from Gallagher (2013).
  - Since the constraints above aren't violated by any of the words in the training data and *are* violated by a majority of the words outside of the training data, they receive a high weight.

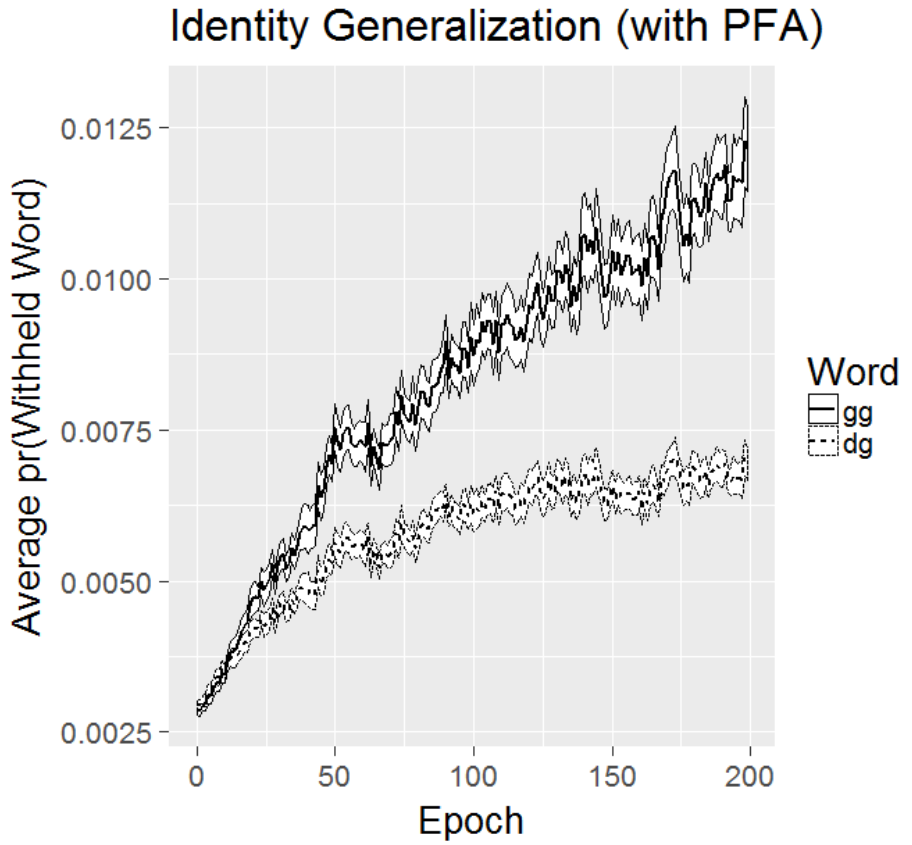
**Learning Datum:**

**Pr(gg)=0**



# Identity Generalization Learning Curve



# But what about different test cases?

- Gallagher (2013) only tested subjects' judgments of [dg] vs. [gg], but could differences in the way the comparison bigram is structured change PFA's results?
- I used [kg] to test this, since the segments in it differ in regards to [voice] instead of [velar]. The feature [voice] is more symmetrical, so one might expect differences in PFA's effects.
- For example, there's a closer number of [-voice][+voice] and [+voice][+voice] bigrams in the unattested data than what we saw for [dg].

| [-voice][-voice] | | [+voice][-voice] | | [-voice][+voice] | | [+voice][+voice] | |
|---|---|---|---|---|---|---|---|
| tt | kk | dt | bk | td | pg | db | gb |
| tp | pk | dp | gt | tb | kd | dg | **gg** |
| tk | kt | dk | gb | tg | kb | bd | |
| pt | kp | bt | gk | pd | **kg** | bg | |
| pp | | bp | | pb | | gd | |

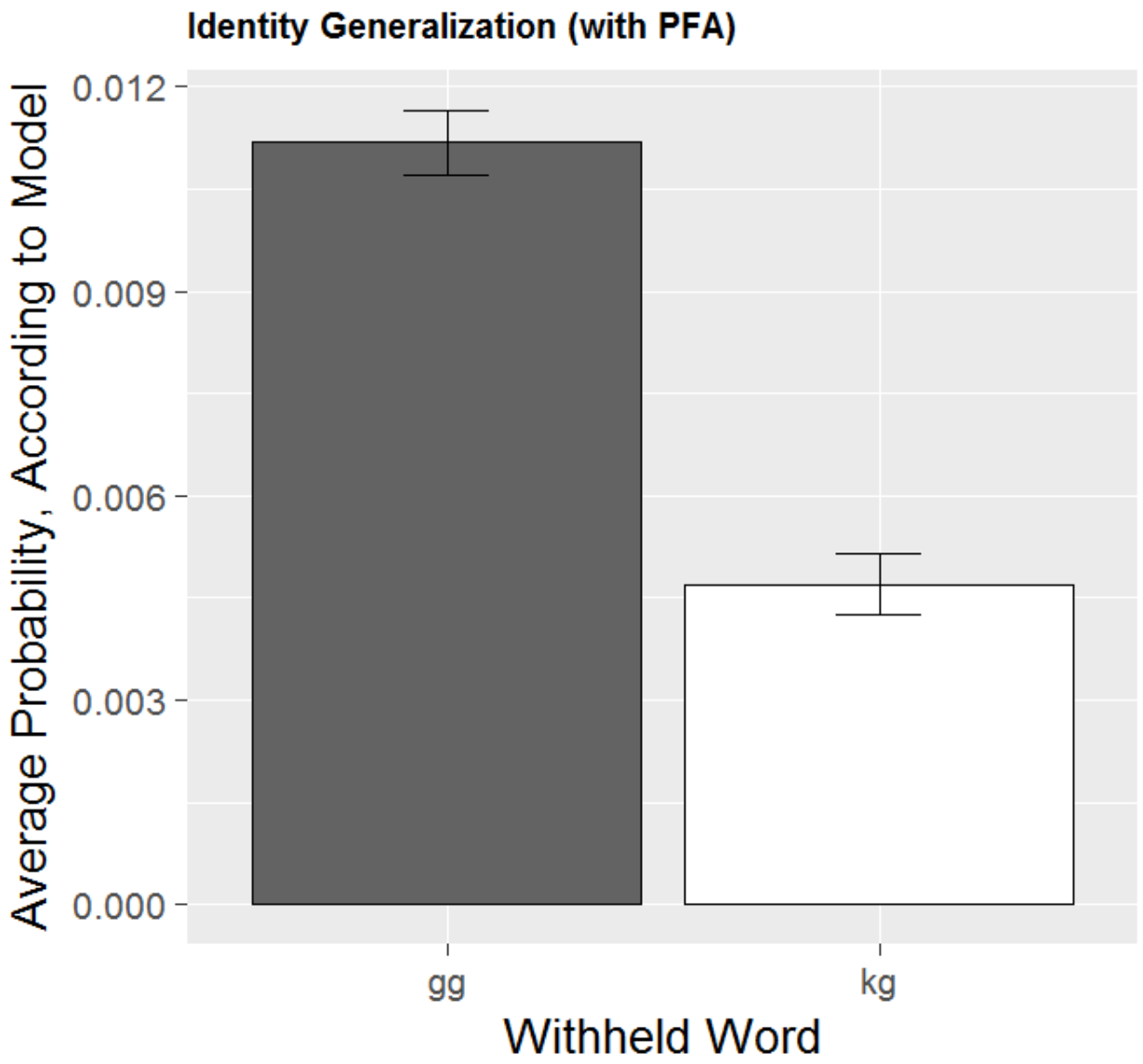
# Why [kg] still isn’t generalized

- So [kg] and [gg] have almost equal similarity to unattested data, but unlike [dg], [kg] has less similarity with attested ones.
- This means that as more probability is pushed onto attested forms, [gg] will get more accidentally pushed onto it than [kg] will.

| [+voice][+voice] |
| --- |
| dd |
| bb |

# [kg] vs. [gg] with PFA

- The results on the right show that even with a [kg] bigram as comparison, the model correctly generalizes to [gg] sequences more.
- 200 epochs, learning rate of .01, initial weights of 0, averaged over 5 runs, probability of attending to each feature, each epoch=.25

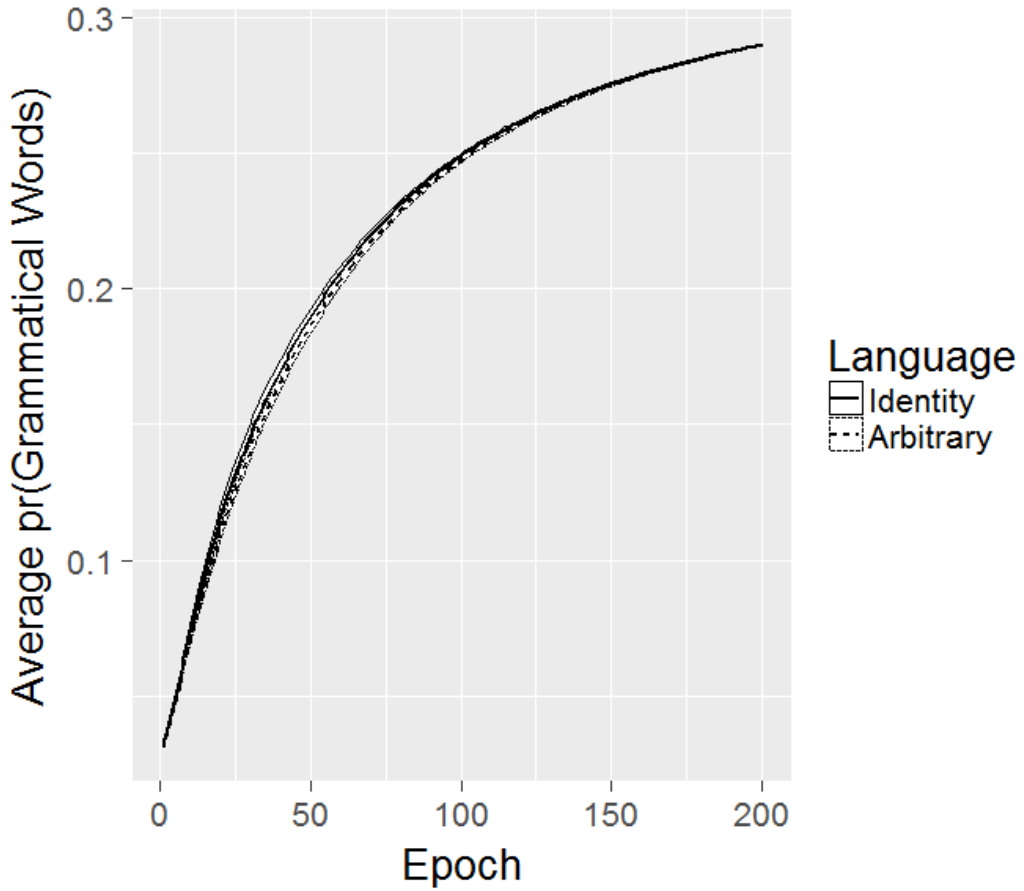


# Identity Bias

- ***Identity Bias*** refers to the fact that humans have been shown to learn categories like {bb, dd, gg} more easily than categories like {bd, dg, gb}.
  - If each category is thought of as a language, this means that humans have a bias toward identity-based phonotactic patterns over arbitrary ones.
  - Identity-based phonotactic patterns are typologically common (Gallagher 2014), strengthening the hypothesis that a bias could be favoring them in language learning.
- Gallagher (2013) showed this using an artificial language learning paradigm similar to the one she used for Identity Generalization.
  - In this experiment, the two categories above were used and participants had to learn a phonological pattern involving the segment in their given category.
  - They were tested on the accuracy of their learning and subjects learning the identity-based language were significantly more accurate.
- See Endress et al. (2007) for an example of this bias affecting non-linguistic learning.

# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior without variables (Gallagher 2013).
  - Since a vanilla model has an equal number of constraints that are violated by the training data in both the Identity-based and Arbitrary languages, there's little difference between their learning curves.
- I replicated these results on the right using a model with no variables or PFA that was given training data based on Gallagher (2013).
  - 200 epochs, learning rate of .01, averaged over 25 runs

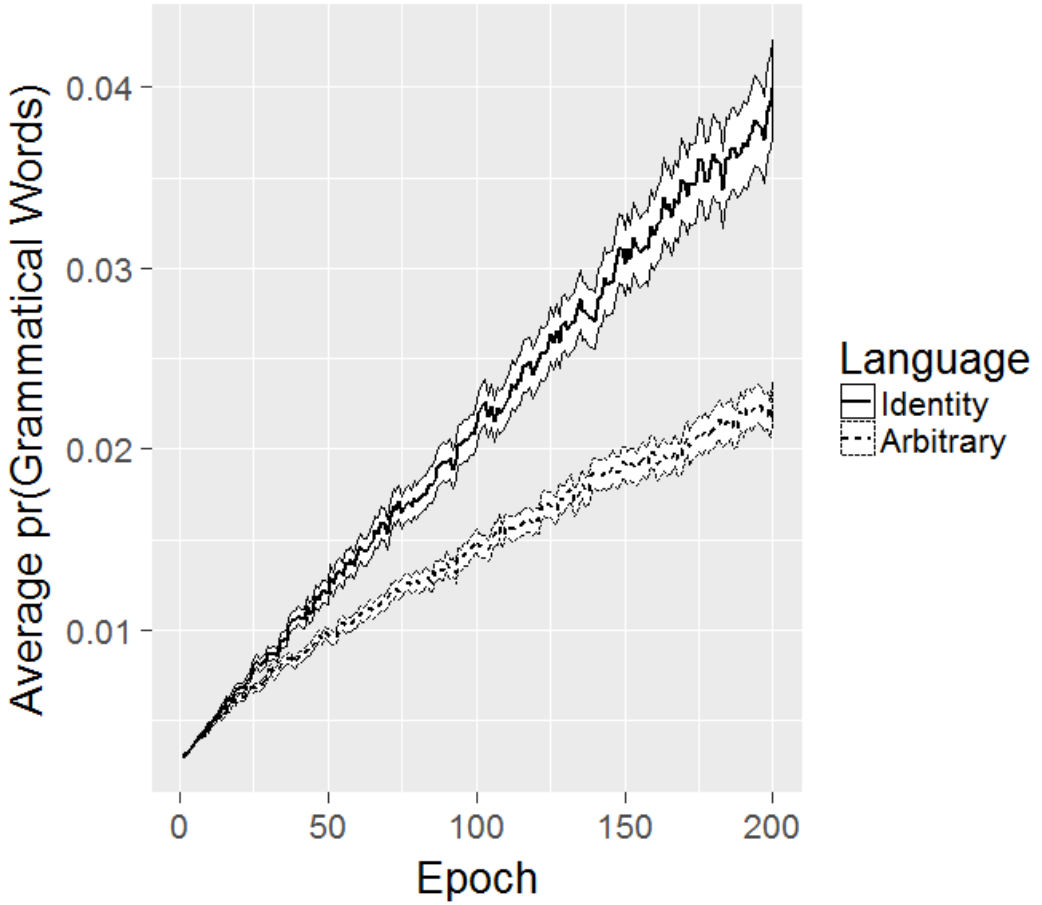


# Modeling with Variables

- When a model is given variables, it's able to model this bias because it has simple, general constraints that can refer to all of the items that aren't in the identity-based language.
  - E.g. *[¬α][α] and *[α][¬α]
  - Since it requires a larger number of constraints to refer to words that aren't in the arbitrary language, that one takes longer to learn (see Pater and Moreton 2012 for more on how and why patterns defined by fewer constraints are easier to learn).
- Gallagher (2013) demonstrated this for an example form Quechua with her identity-based MaxEnt model.

# Modeling with PFA

- PFA correctly predicts Identity Bias.
- Over the course of learning in the identity language, the model is attempting to assign more probability to dVd, bVb, and gVg words than any of their alternatives.
  - Due to PFA, the attested words in this language will often be identical with one another (since they have a large number of shared features that are identical across segments).
  - Additionally, many words will be likely to turn into these when a feature isn't being attended to (e.g. tVd → DVD when voicing is ambiguous).
- The figure on the right shows the results for the PFA model trained on the pattern from Gallagher (2013).
  - 200 epochs, learning rate of .01, averaged over 25 runs, probability of attending to each feature=.25

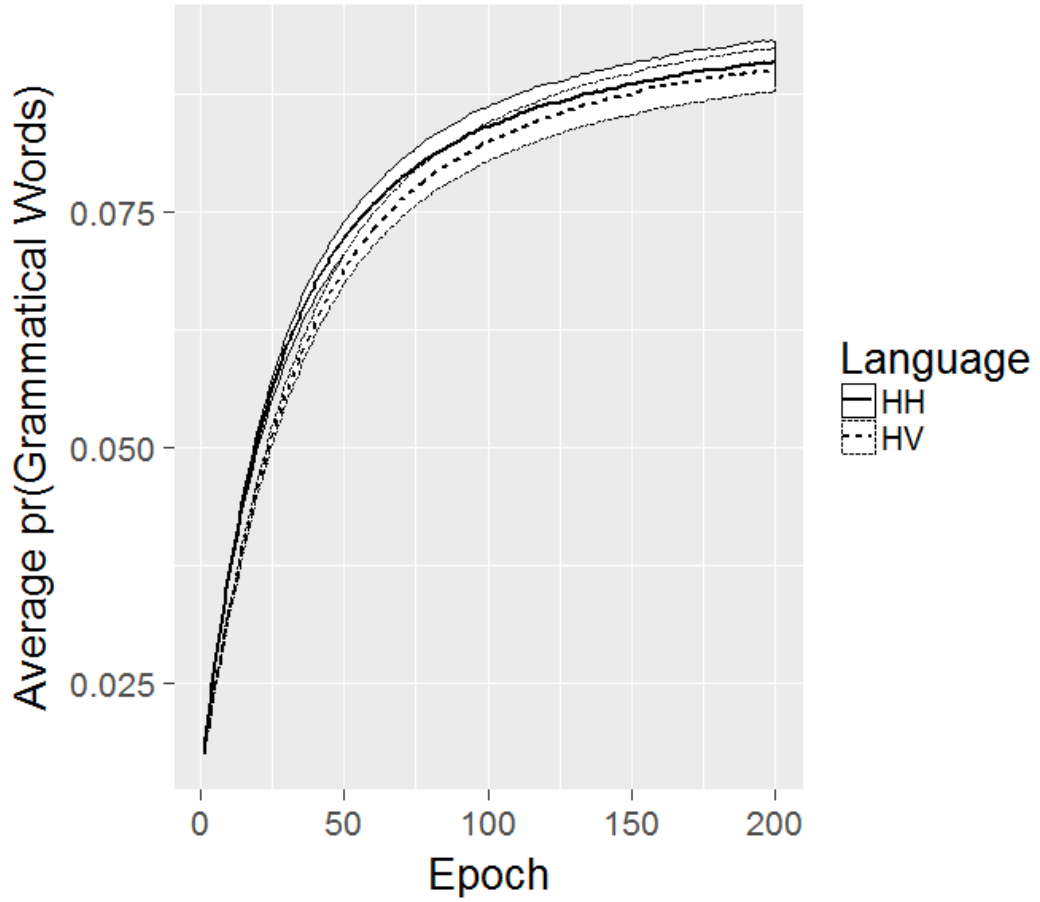


# Intradimensional Bias

- ***Intradimensional Bias*** refers to the fact that humans learn patterns involving the same feature across multiple segments more easily than patterns that involve different features across segments.
  - For example, a vowel harmony pattern is easier to learn than a pattern in which the height of a vowel depends on the voicing of the preceding consonant (Moreton 2008).
- Moreton (2012) showed that a number of different intradimensional phonotactic patterns were easier to learn than minimally different interdimensional ones.
  - The language's that I'll be modeling here are Moreton's (2012) HH and HV languages.
  - In HH, the height of vowels always had to agree.
  - In HV, voiced consonants only occurred after high vowels and voiceless consonants only occurred after low ones.
- This bias also exists in non-linguistic learning (see Moreton 2012:§1.3 for a review).

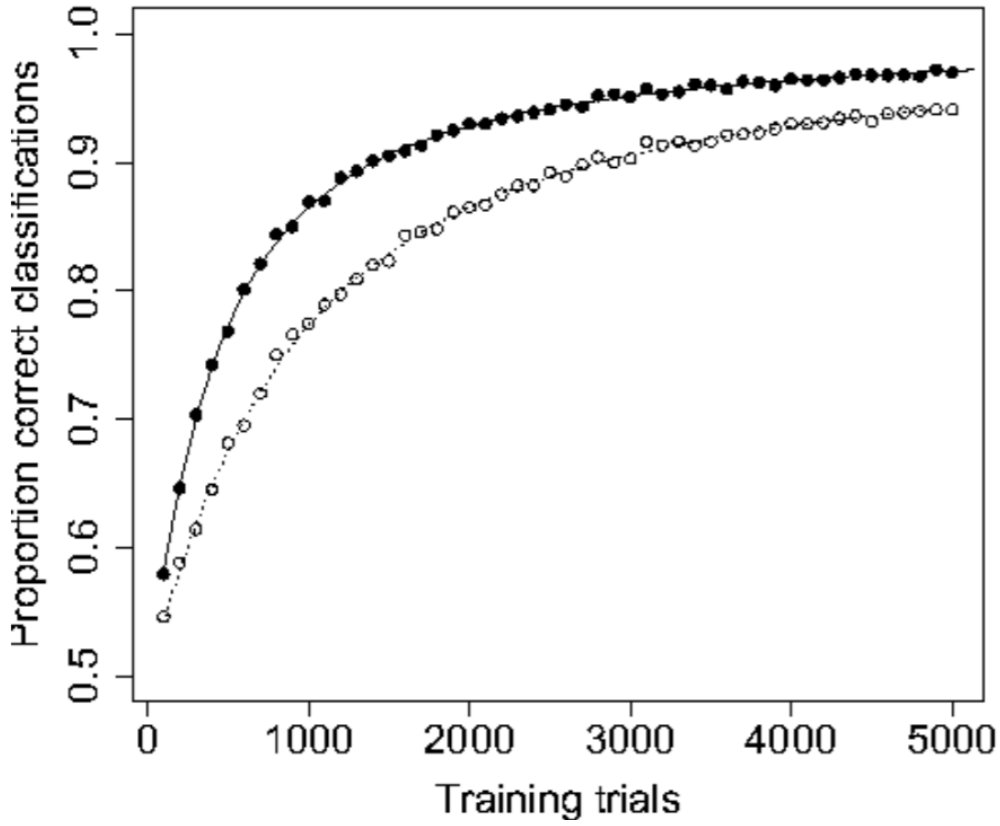
# Modeling with a vanilla MaxEnt model

- Standard phonotactic models (e.g. Hayes and Wilson 2008; Pater and Moreton 2012) can't capture this behavior well without variables (Moreton 2012).
  - Since these models don't have any way of representing intersegmental similarity, they can't capture any facts about the whether a two-segment pattern is inter- or intradimensional.
- I replicated these results on the right using a model with no variables or PFA that was given training data based on Moreton's (2012) experiment.
  - 200 epochs, learning rate of .01, averaged over 25 runs



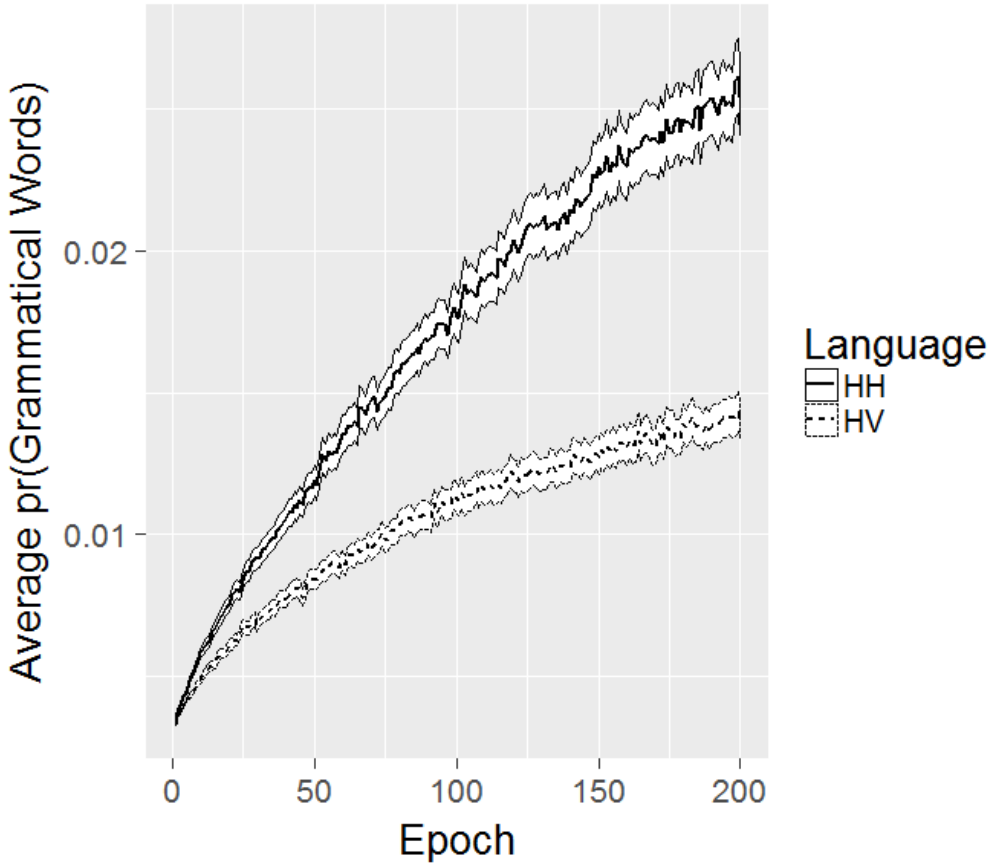
# Modeling with Variables

- When a model is given variables, it's able to model this bias because it can represent intersegmental similarity using variables on the feature values across segments.
  - E.g. the *[αHigh][αHigh] input node used by Moreton's (2012) neural network.
  - This allows intradimensional patterns to be represented using fewer input nodes (or constraints in the case of a MaxEnt model) than their interdimensional counterparts.
- The results from Moreton's (2012) simulations are shown on the right.
  - The Configural Cue Model (Gluck and Bower 1988) was the base neural network that Moreton (2012) added variables to for these simulations.
  - Black dots are the HH pattern, white dots are HV



# Modeling with PFA

- PFA also correctly predicts Intradimensional Bias.
- Over the course of learning, the model is attempting to assign more probability to either HH sequences or HV sequences, depending on the language.
  - Due to PFA, the fewer features that are relevant to a pattern (across and within segments), the less likely ambiguity is to hinder learning of the pattern in a given weight update.
  - E.g. when [high] is not attended to during an update in the HH pattern, any relevant information is gone from the data.
  - But if either [high] or [voice] is not attended to during HV, the pattern will be obscured in the data.
- The figure on the right shows the results for the PFA model trained on the Moreton (2012) pattern.
  - 200 epochs, learning rate of .01, averaged over 25 runs, probability of attending to each feature=.25



# Variables in Psychology

- In psychology, there's been a debate over whether models of cognition should include algebraic variables (Marcus 2001).
- In this context, a variable would be any representation that ties together individual tokens in a way that ignores those tokens' individual characteristics.
  - For example, describing reduplication as the mapping "α → αα", where α stands for any arbitrary stem.
- This is typically used to argue against connectionist models that use variable-free representations.
  - For example, Marcus et al. (1999) showed that recurrent neural networks without explicit variables couldn't model human generalization of a reduplicative pattern.
  - Endress et al. (2007) showed that subjects learning a non-linguistic categorization task seemed to behave in a way that variable-free models couldn't predict.