Variational Approximation of Long-Span Language Models for LVCSR

Anoop Deoras[§], Tomáš Mikolov[§§]
Stefan Kombrink[§§], Martin Karafiát[§§]
Sanjeev Khudanpur[§]

[§] HLTCOE and CLSP, Johns Hopkins University, USA
[§§] Speech@FIT, Brno University of Technology, CZ

May 25th 2011, IEEE- ICASSP 2011

Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline

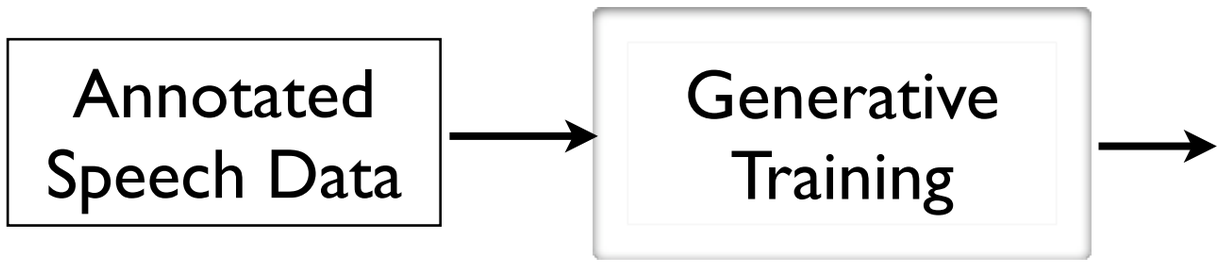
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline

Annotated Speech Data →

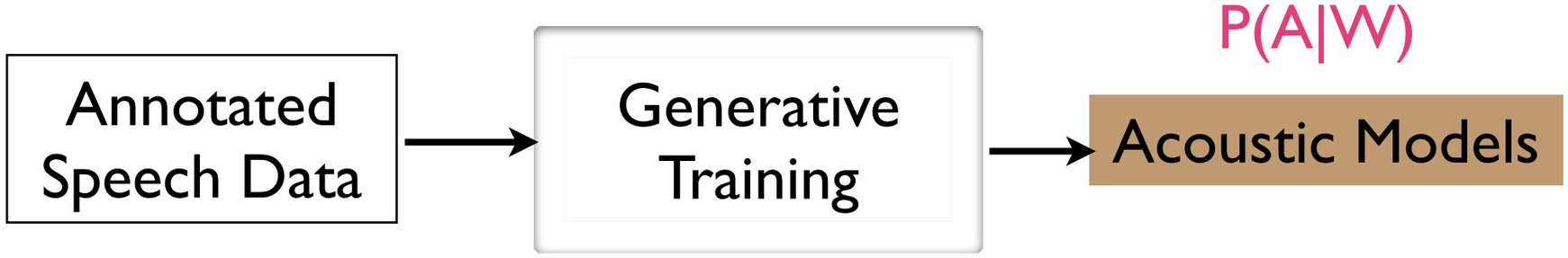
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



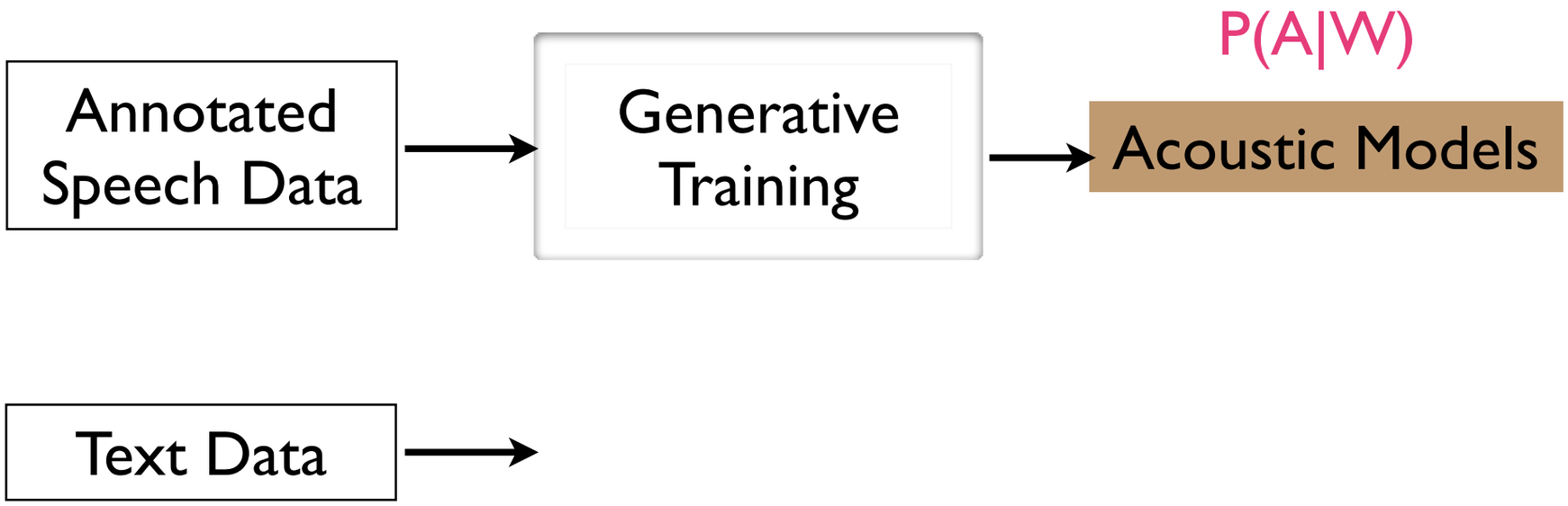
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



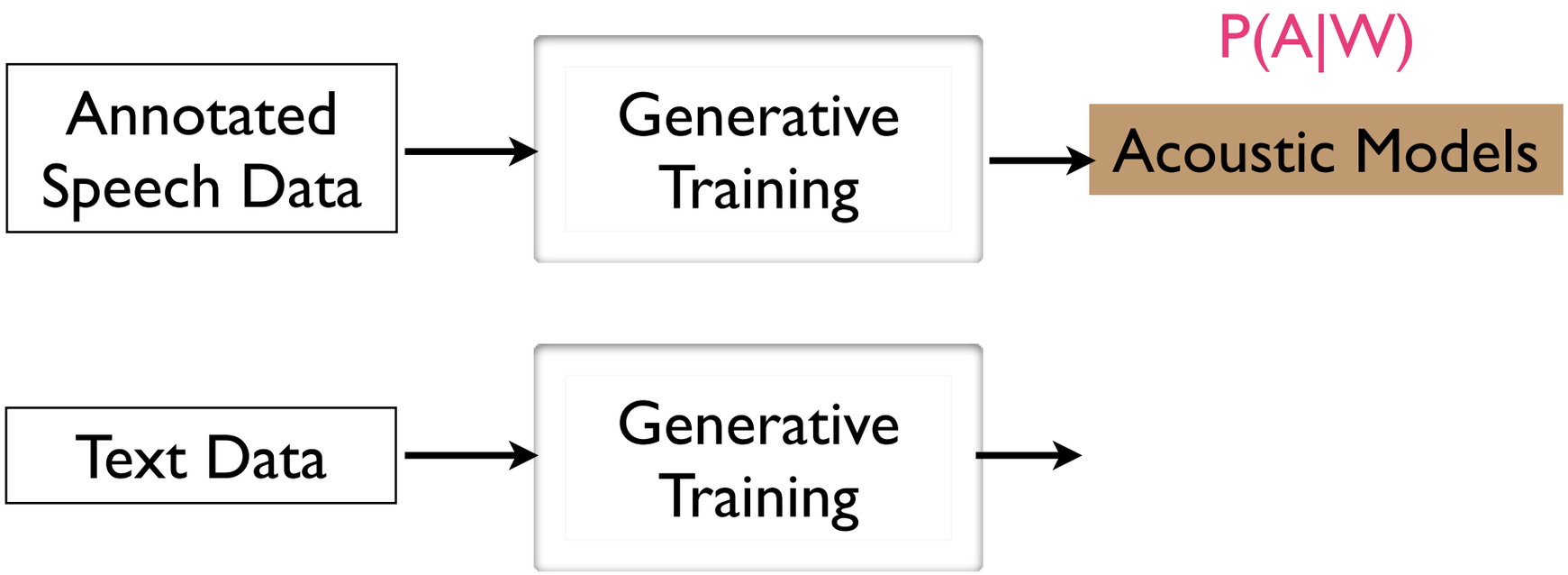
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



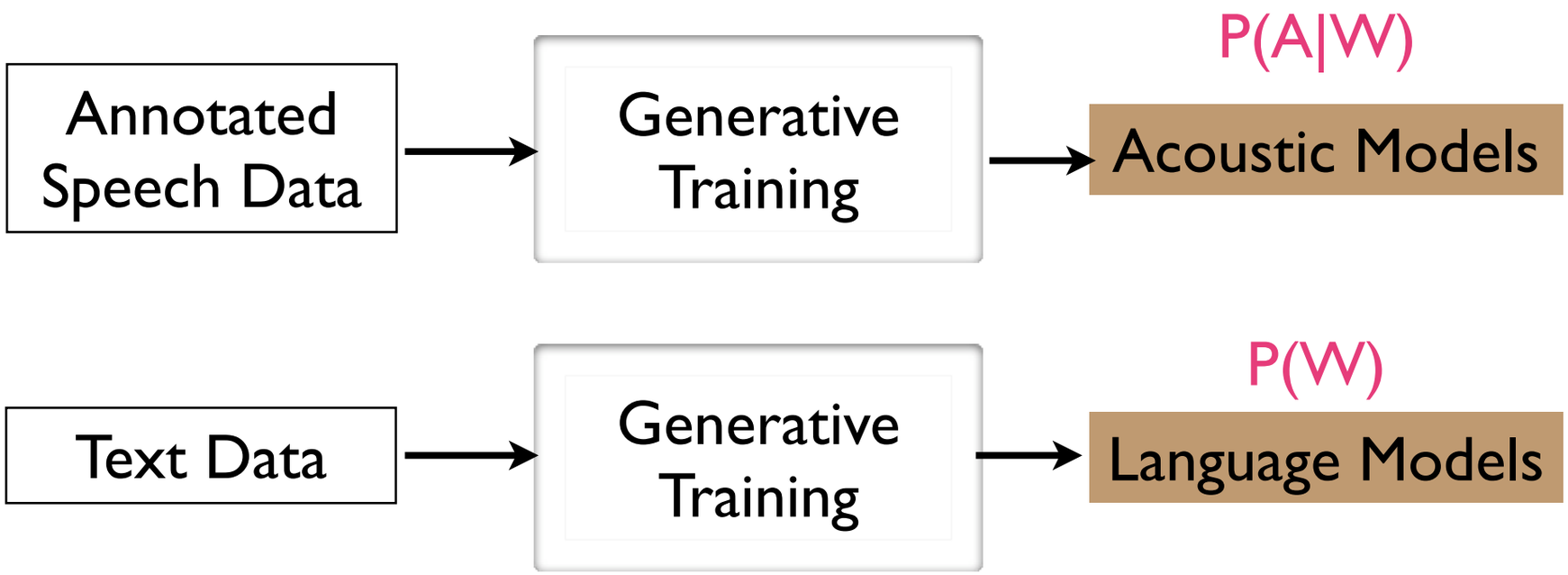
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



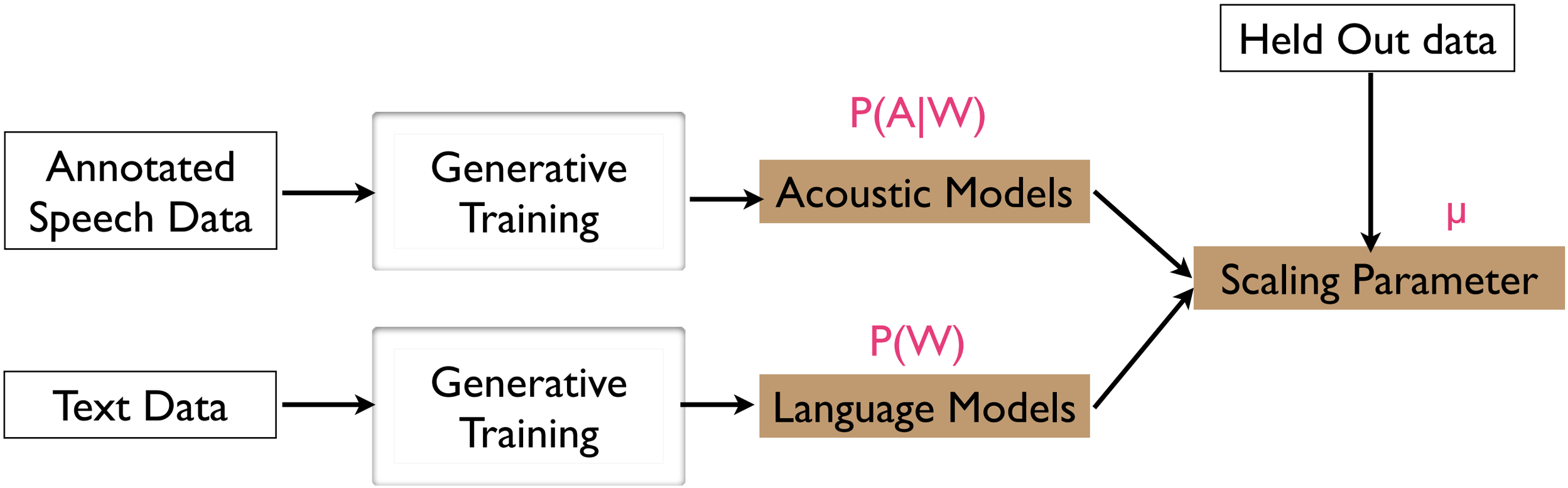
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



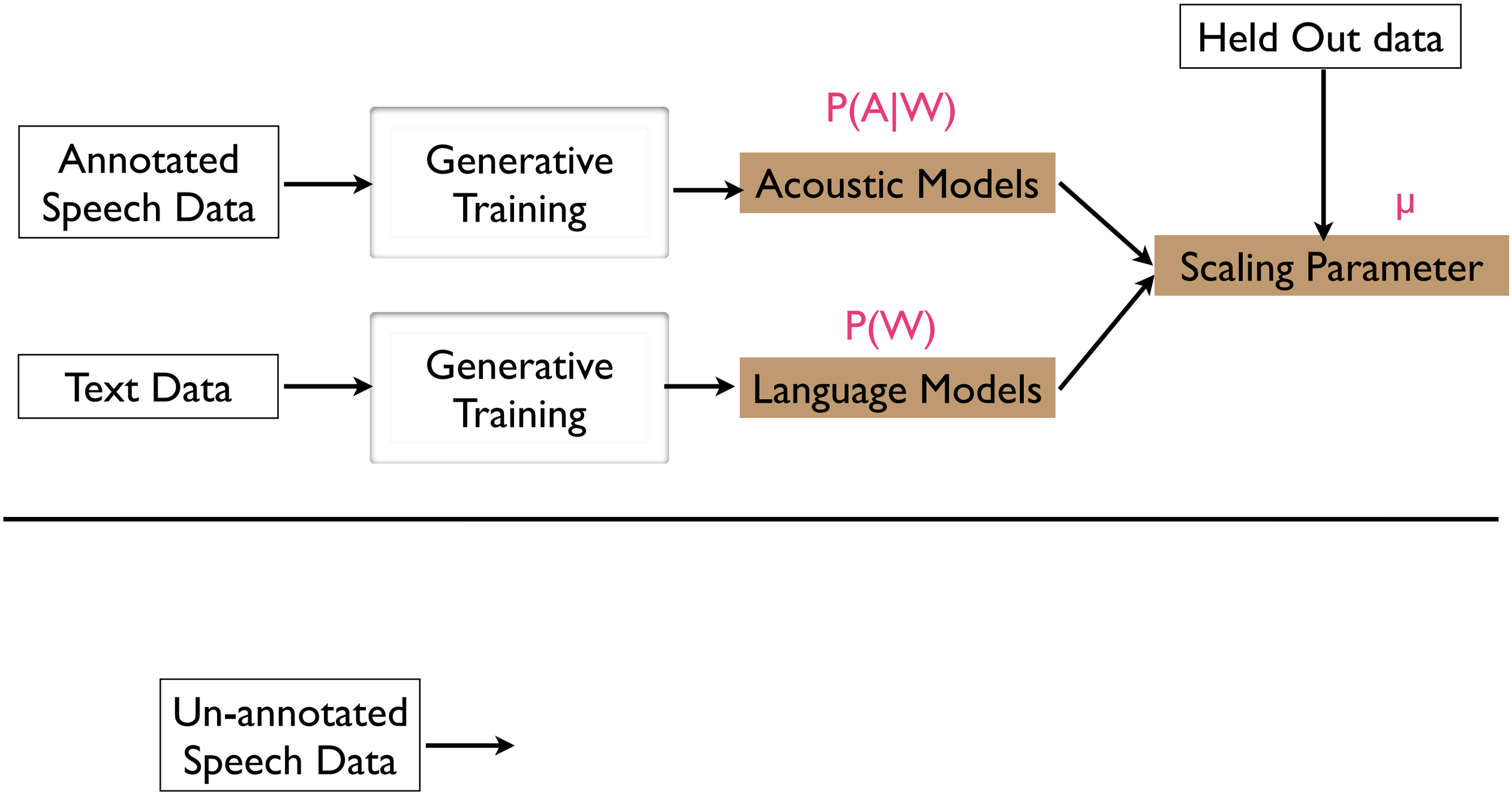
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



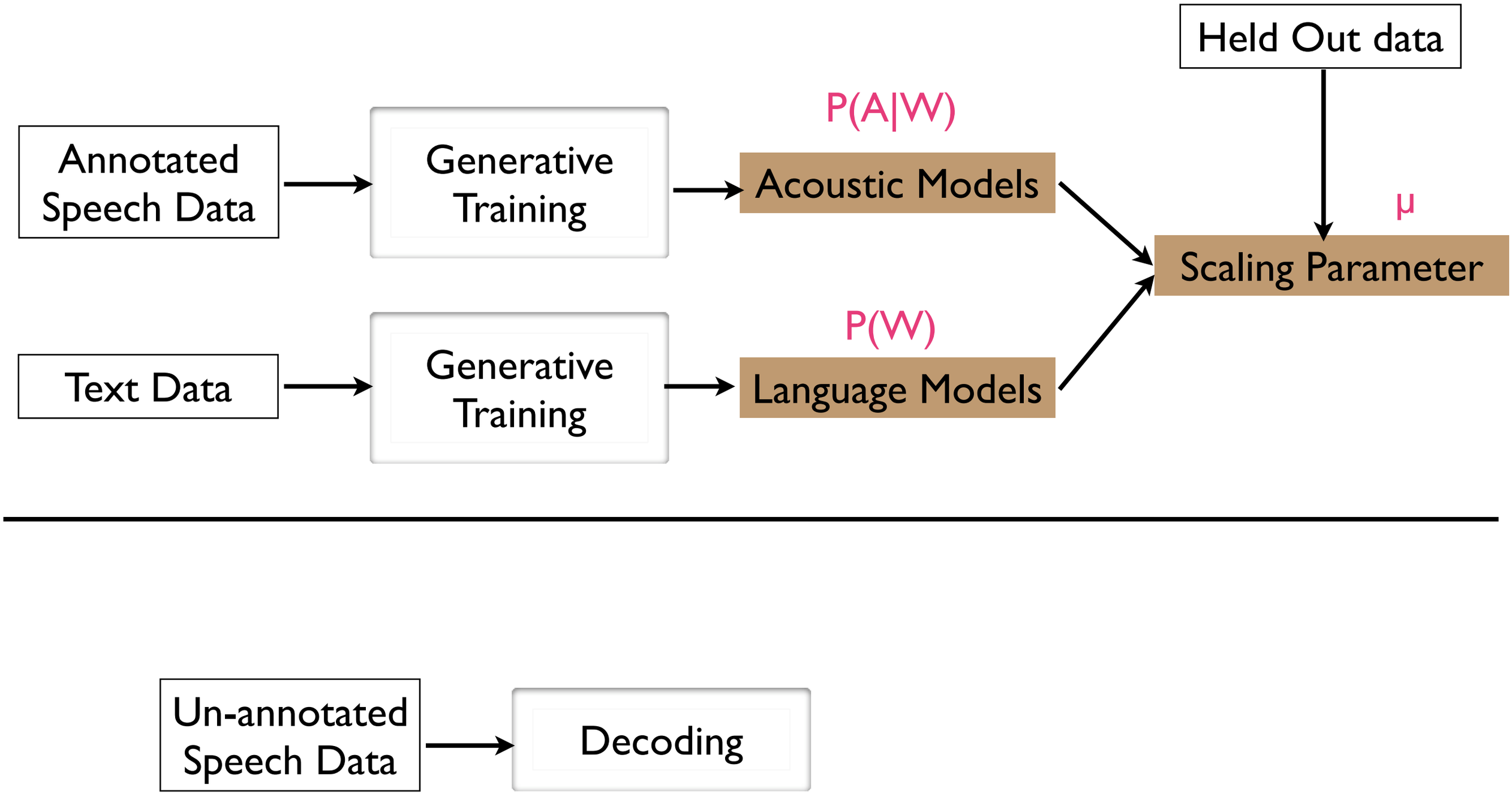
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



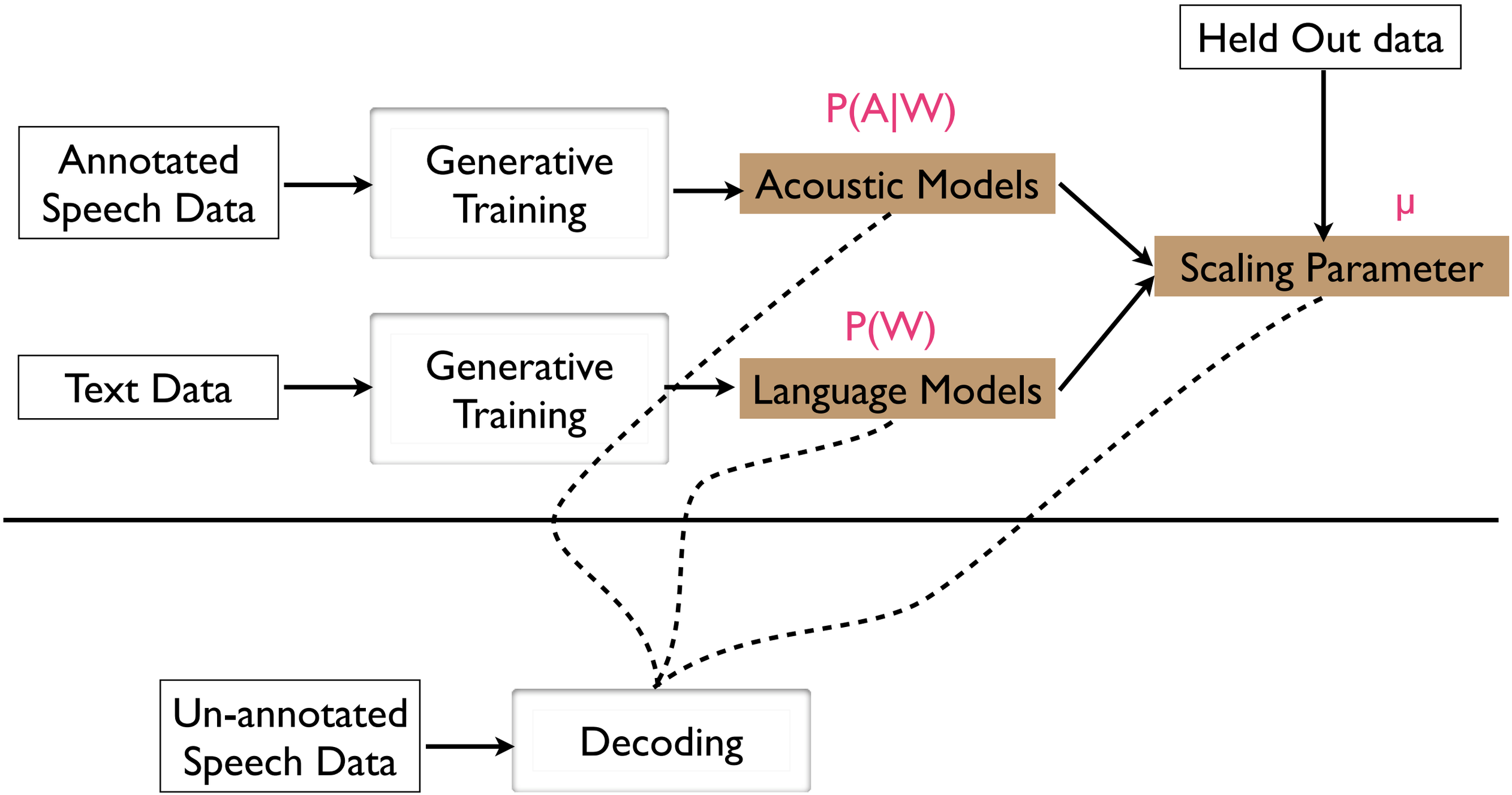
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



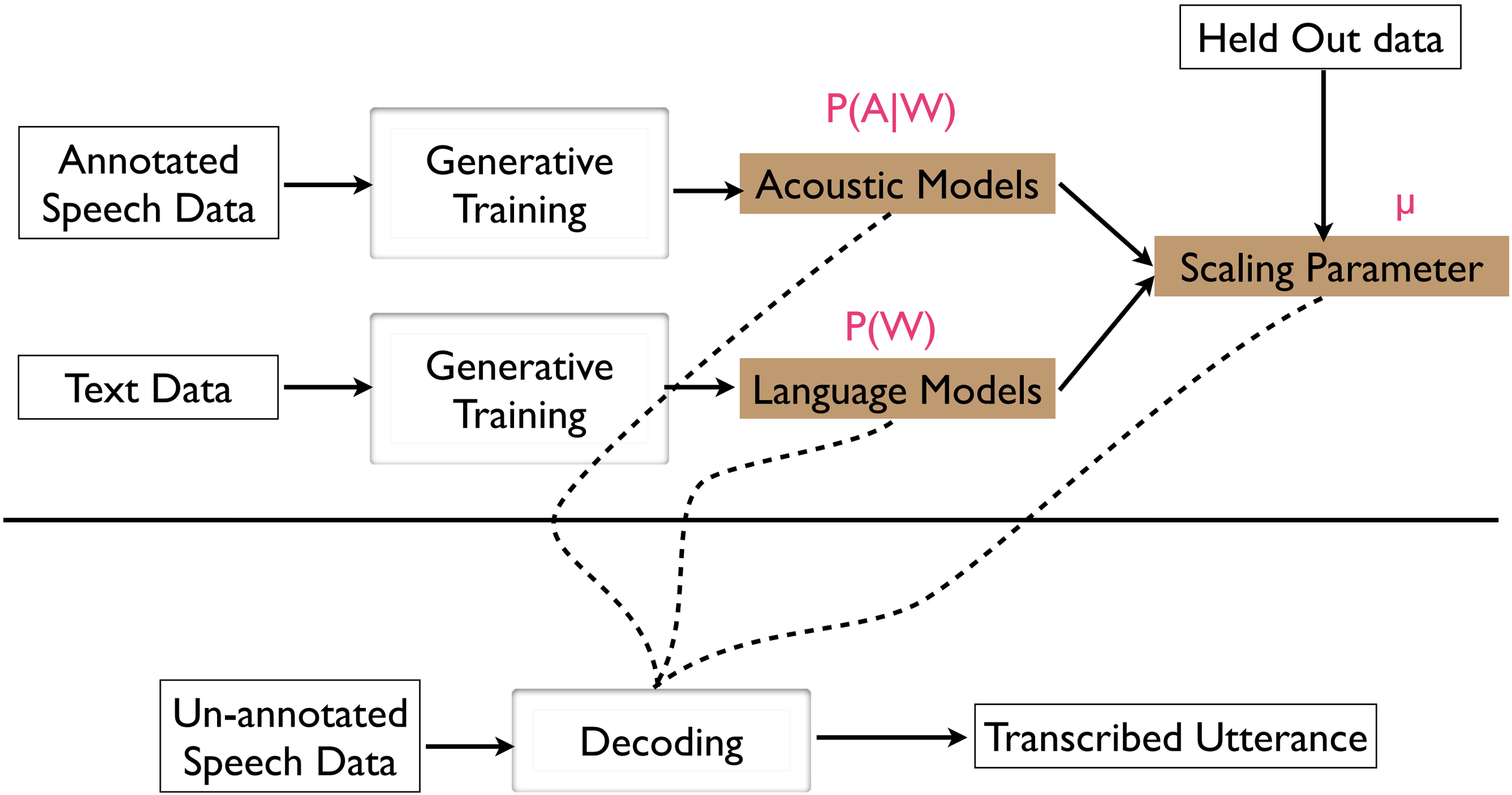
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



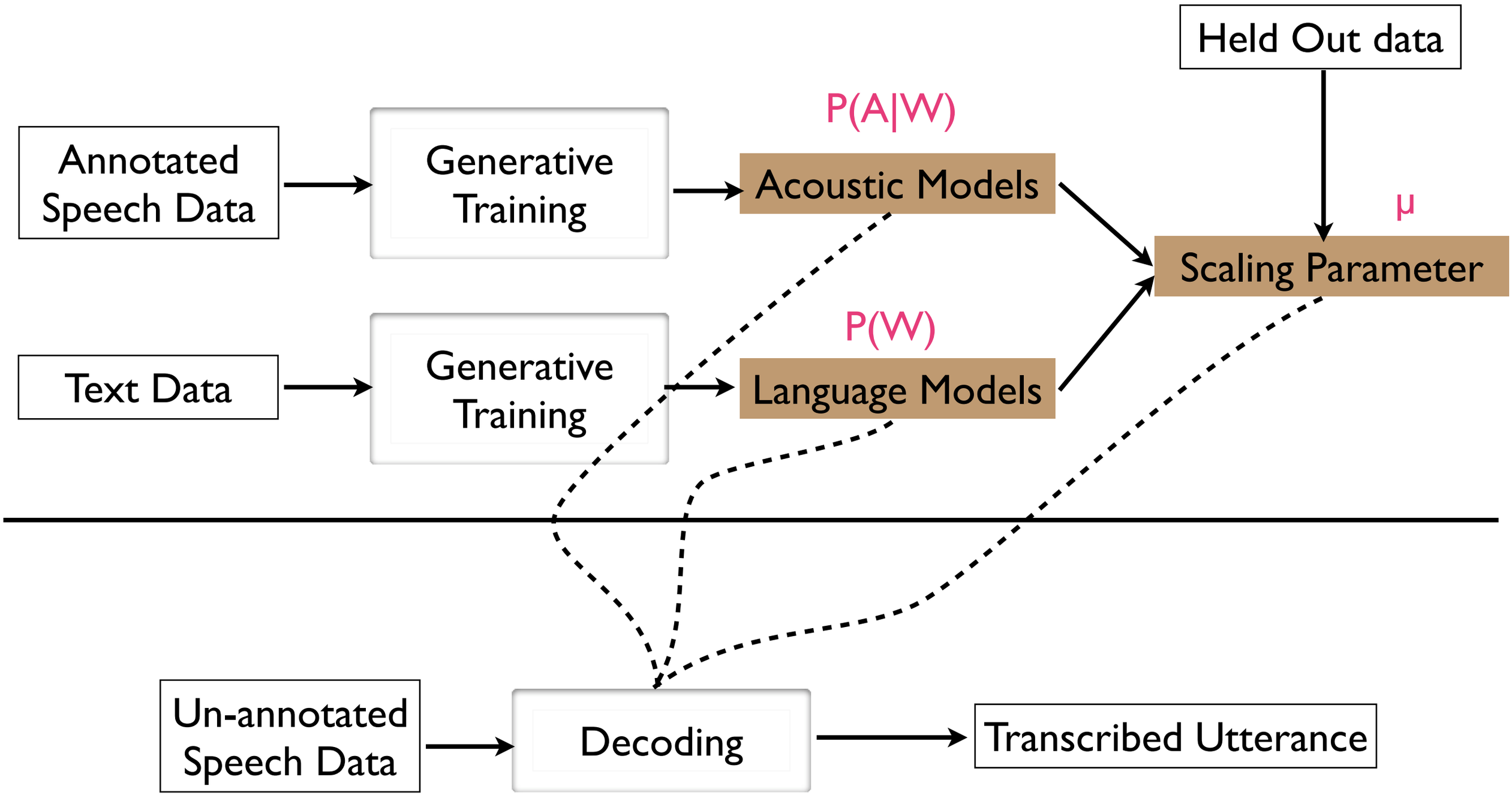
Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



Thanks to Zhifei Li for this template

Statistical Speech Recognition Pipeline



$$W^* = \arg\max_W P(A|W)^\mu P(W)$$

Language Models

Assign a probability distribution P(W) to any word string W.

P(W) is obtained using chain rule and then approximated using Markov assumption:

$$P(W) = \prod_{i=1}^{M} P(w_i|w_{i-1},\ldots,w_1) \approx \prod_{i=1}^{M} P(w_i|\phi(w_{i-1},\ldots,w_1))$$

Language Models

Assign a probability distribution P(W) to any word string W.

P(W) is obtained using chain rule and then approximated using Markov assumption:

{word history}

$$P(W) = \prod_{i=1}^{M} P(w_i|w_{i-1}, \ldots, w_1) \approx \prod_{i=1}^{M} P(w_i|w_{i-1}, \ldots, w_{i-n+1})$$

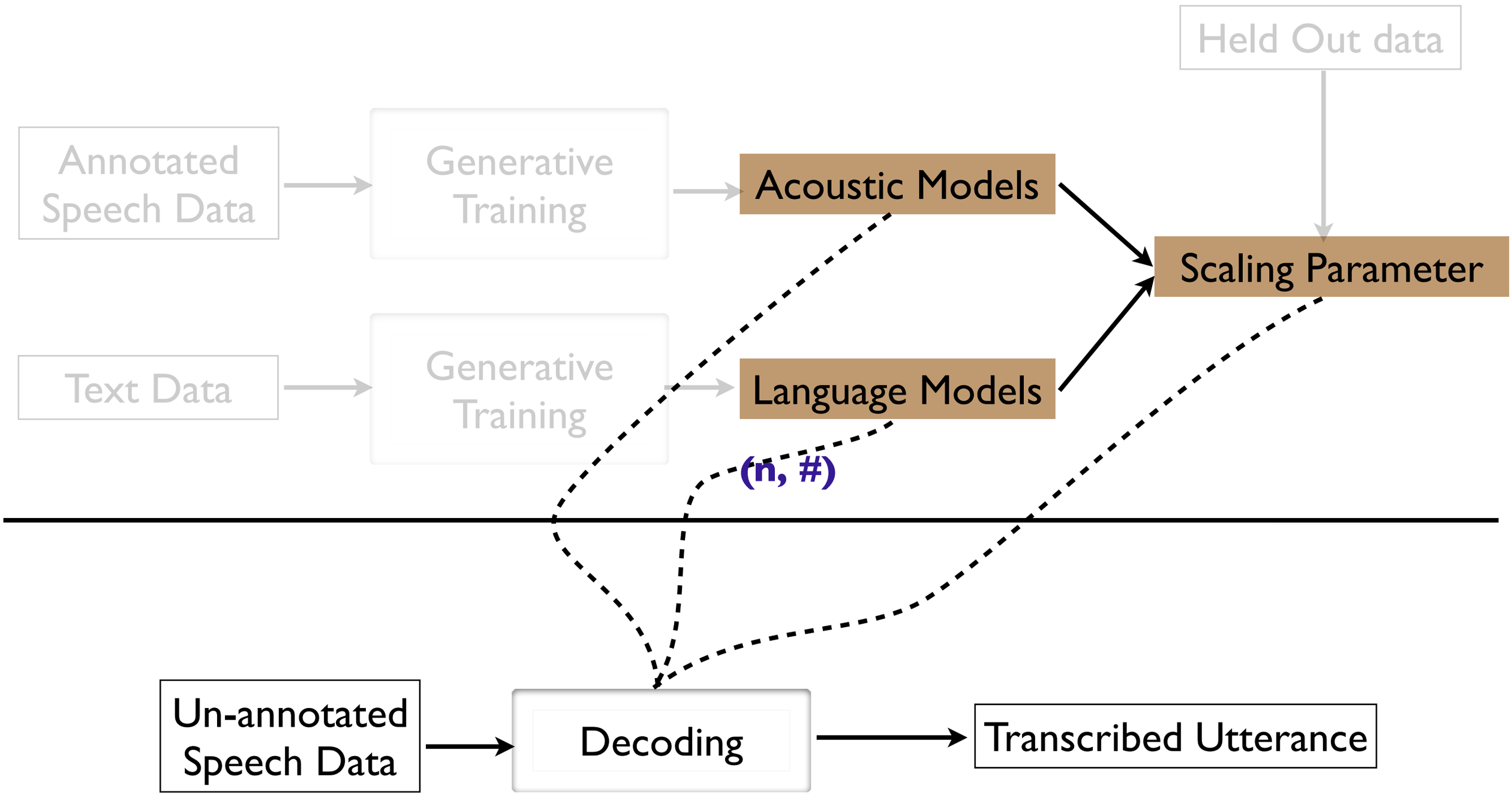
n-gram LM

n: Order of LM

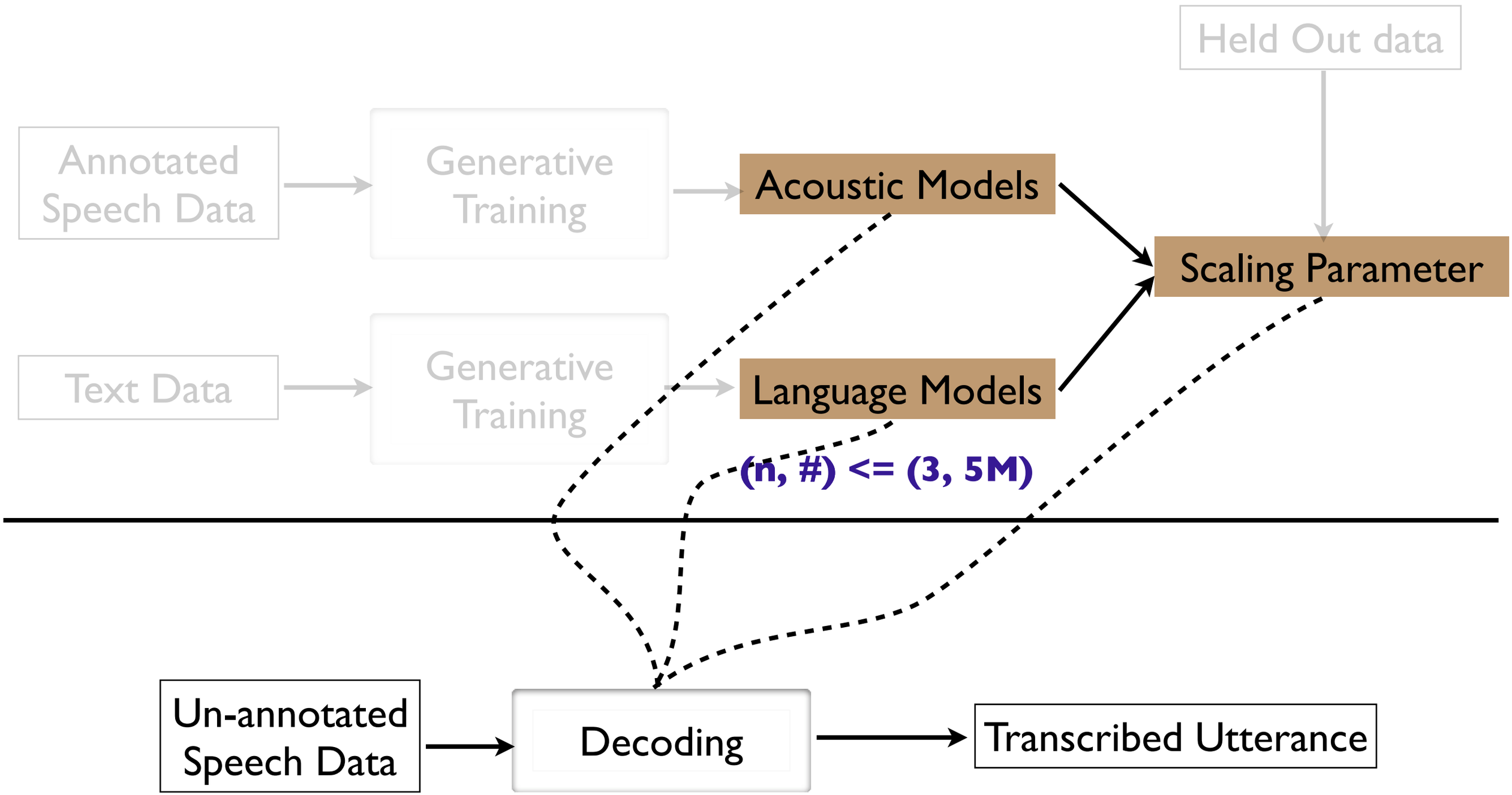
#{word, history}: Number of n-grams

Size of LM (n, #)

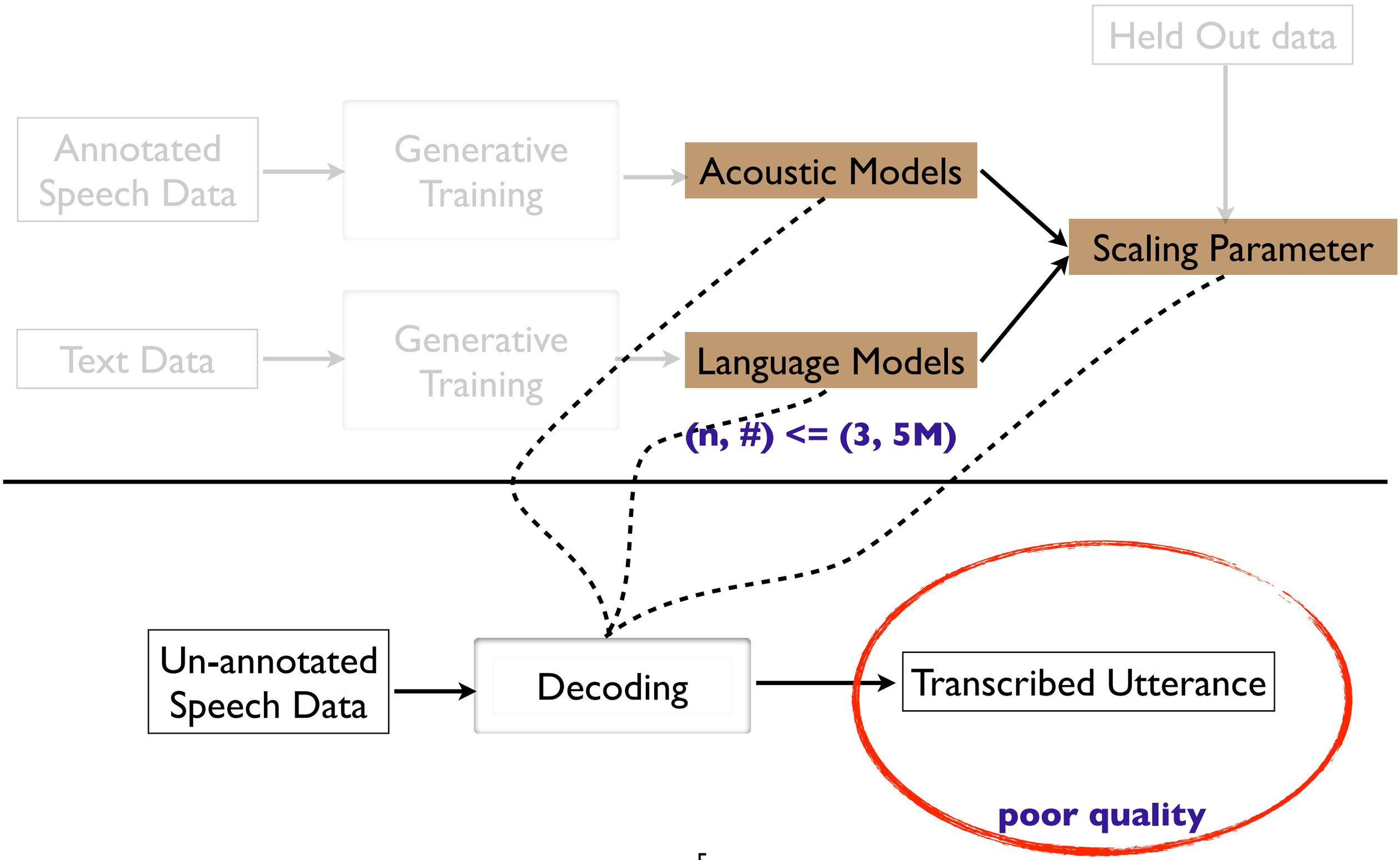
Statistical Speech Recognition Pipeline



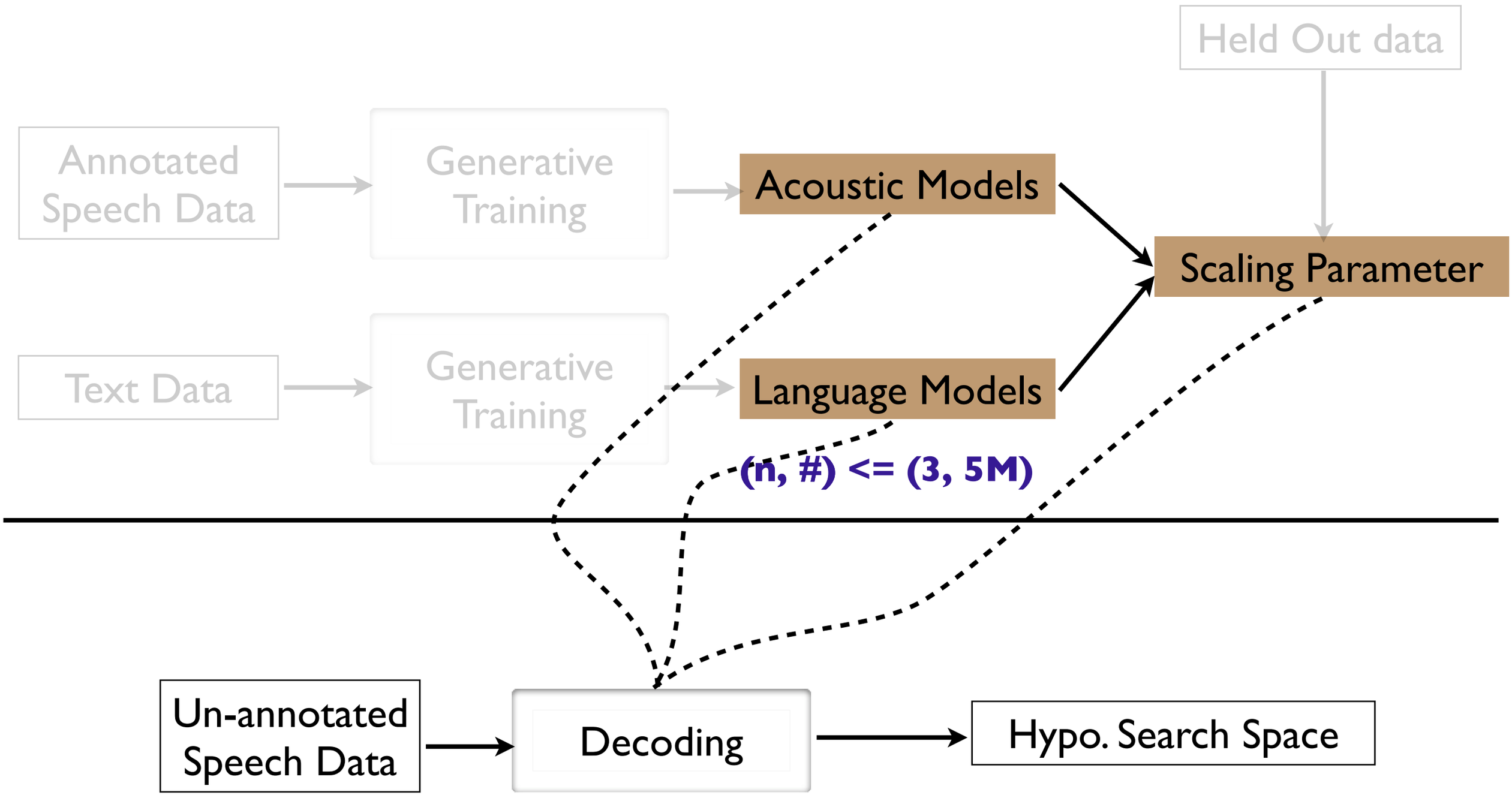
Statistical Speech Recognition Pipeline



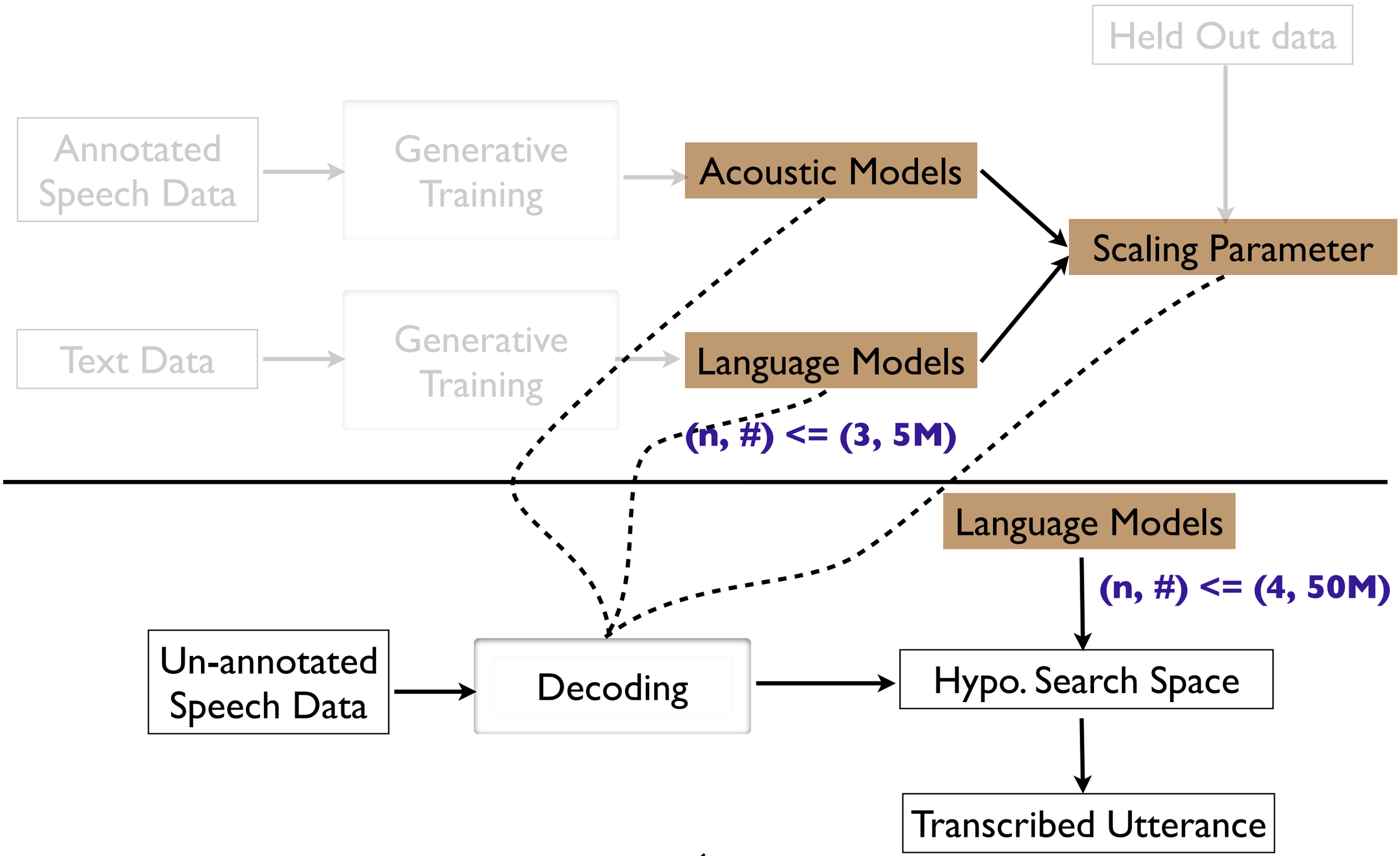
Statistical Speech Recognition Pipeline



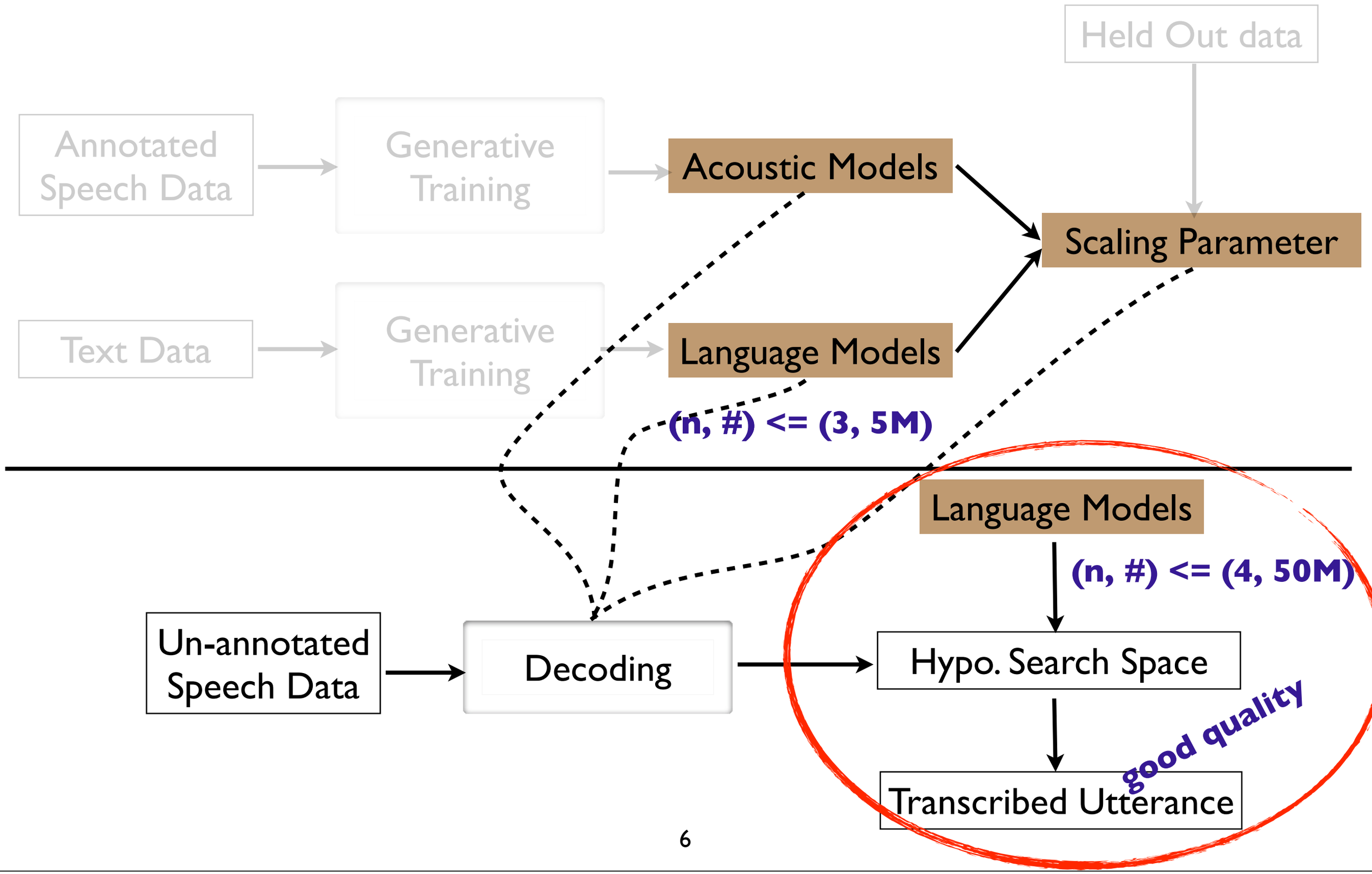
Statistical Speech Recognition Pipeline

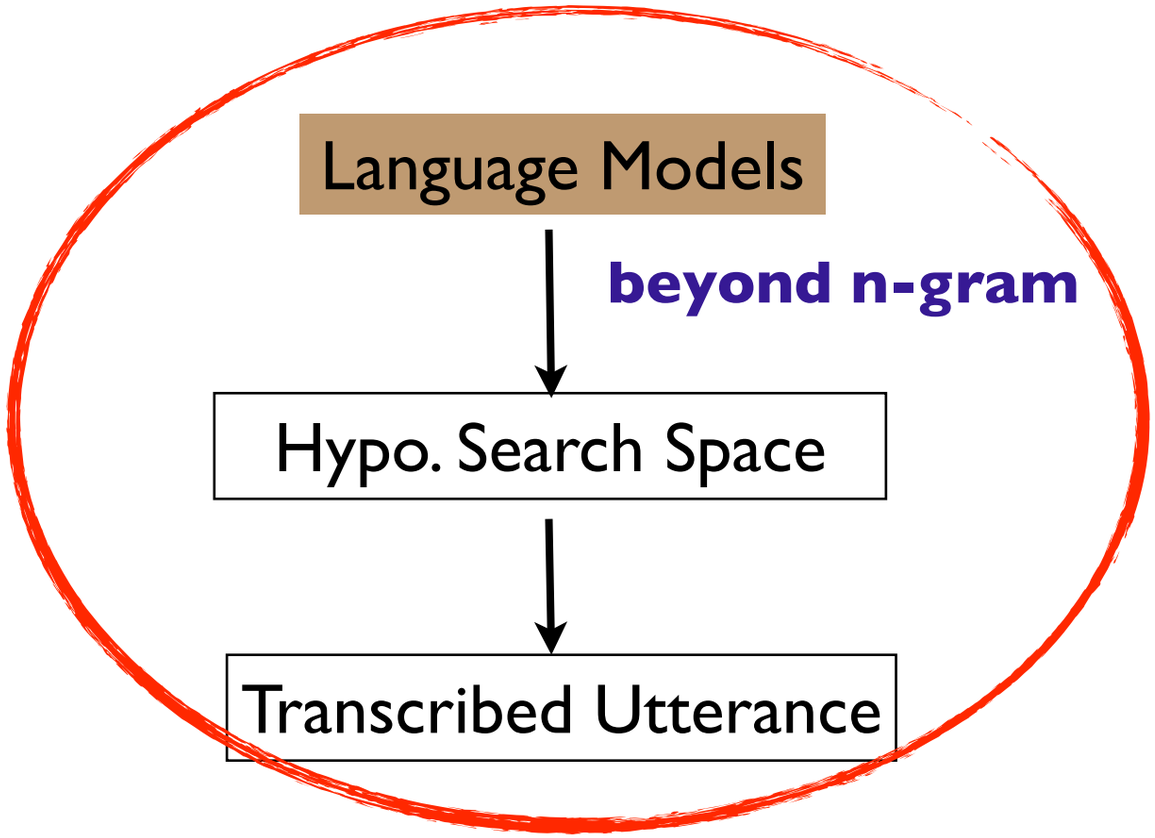


Statistical Speech Recognition Pipeline



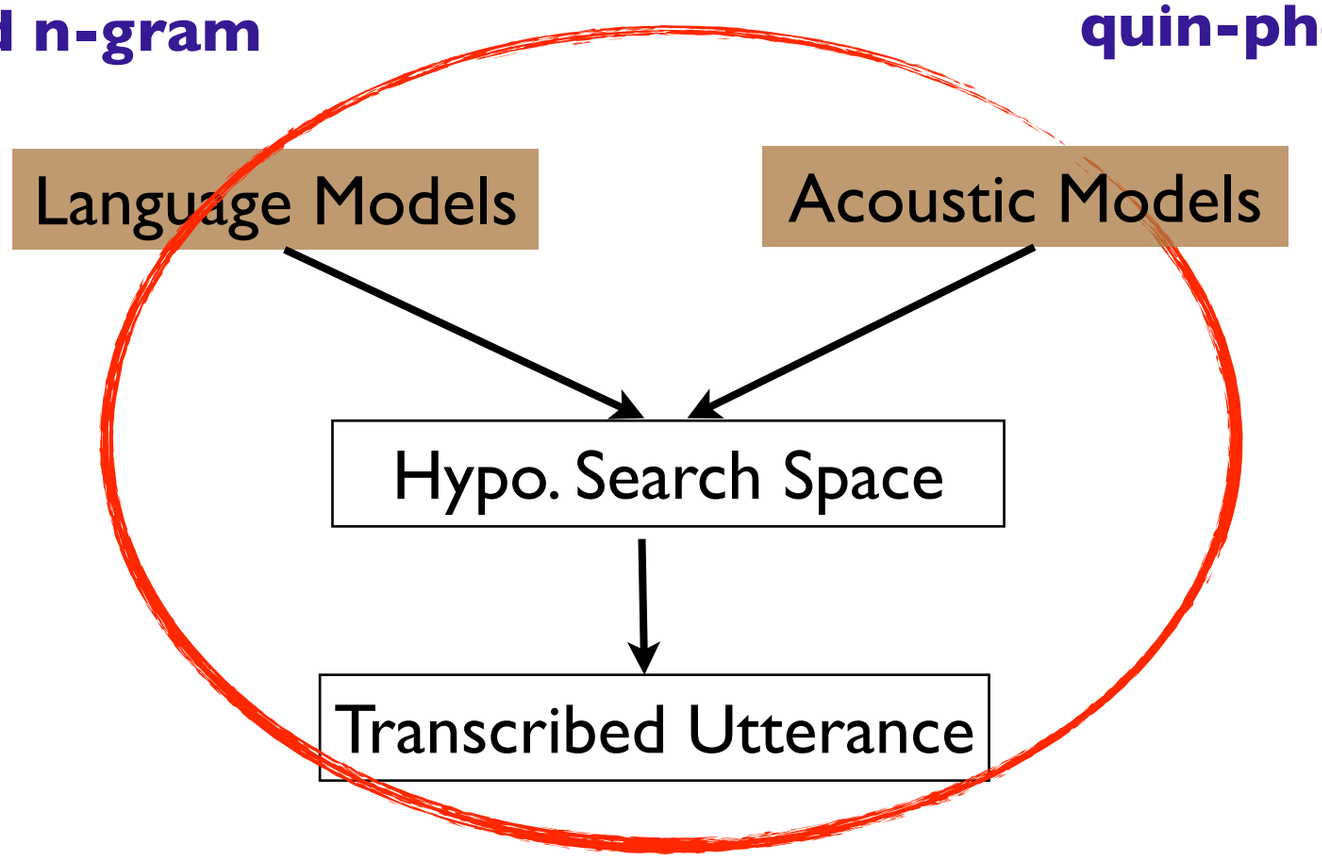
Statistical Speech Recognition Pipeline



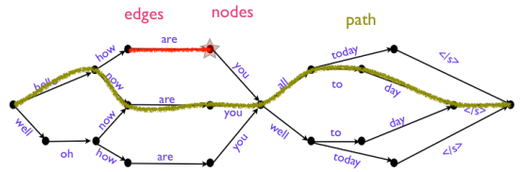

Language Models
beyond n-gram
Hypo. Search Space
Transcribed Utterance

beyond n-gram

quin-phones

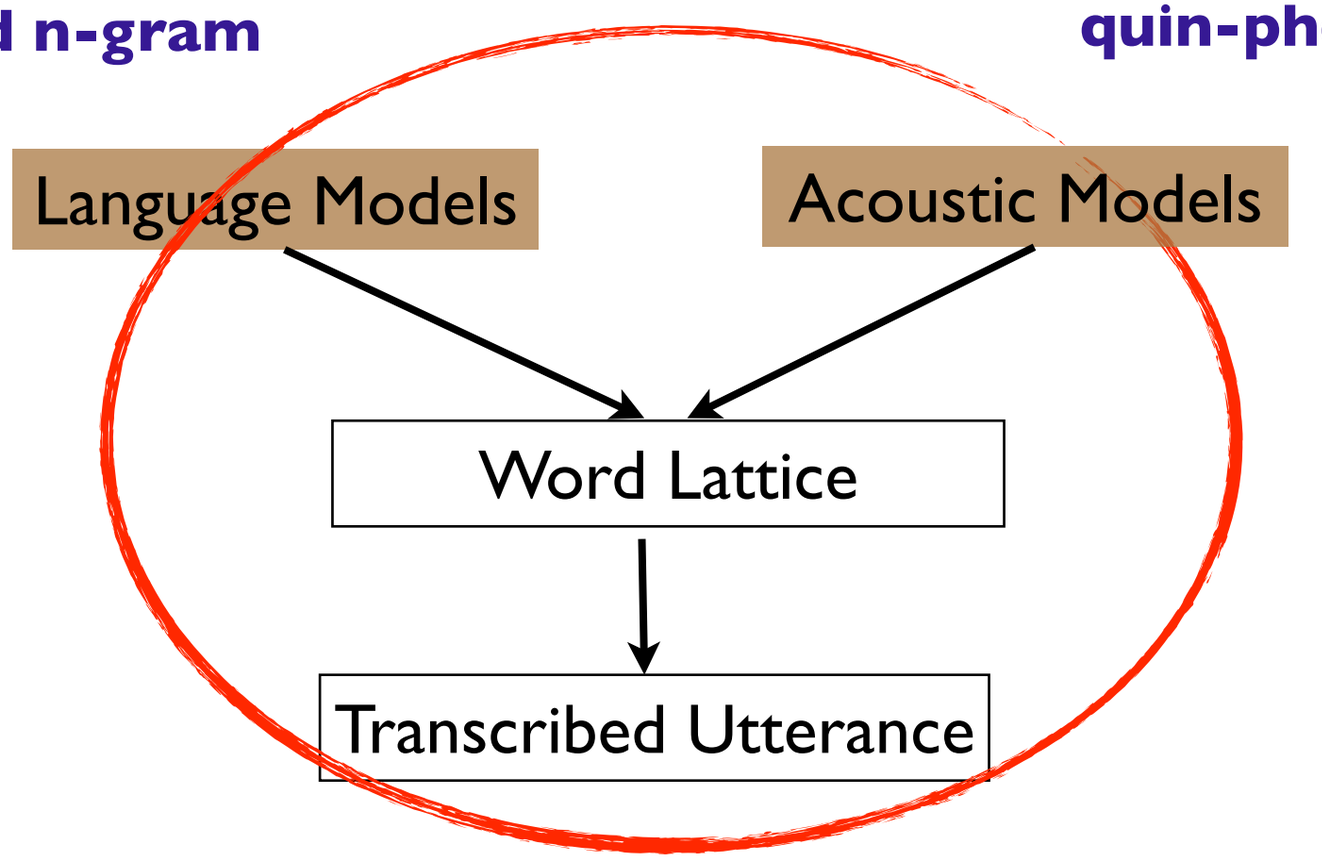


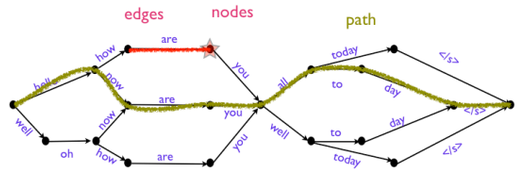
Re-scoring Decoding Framework



beyond n-gram

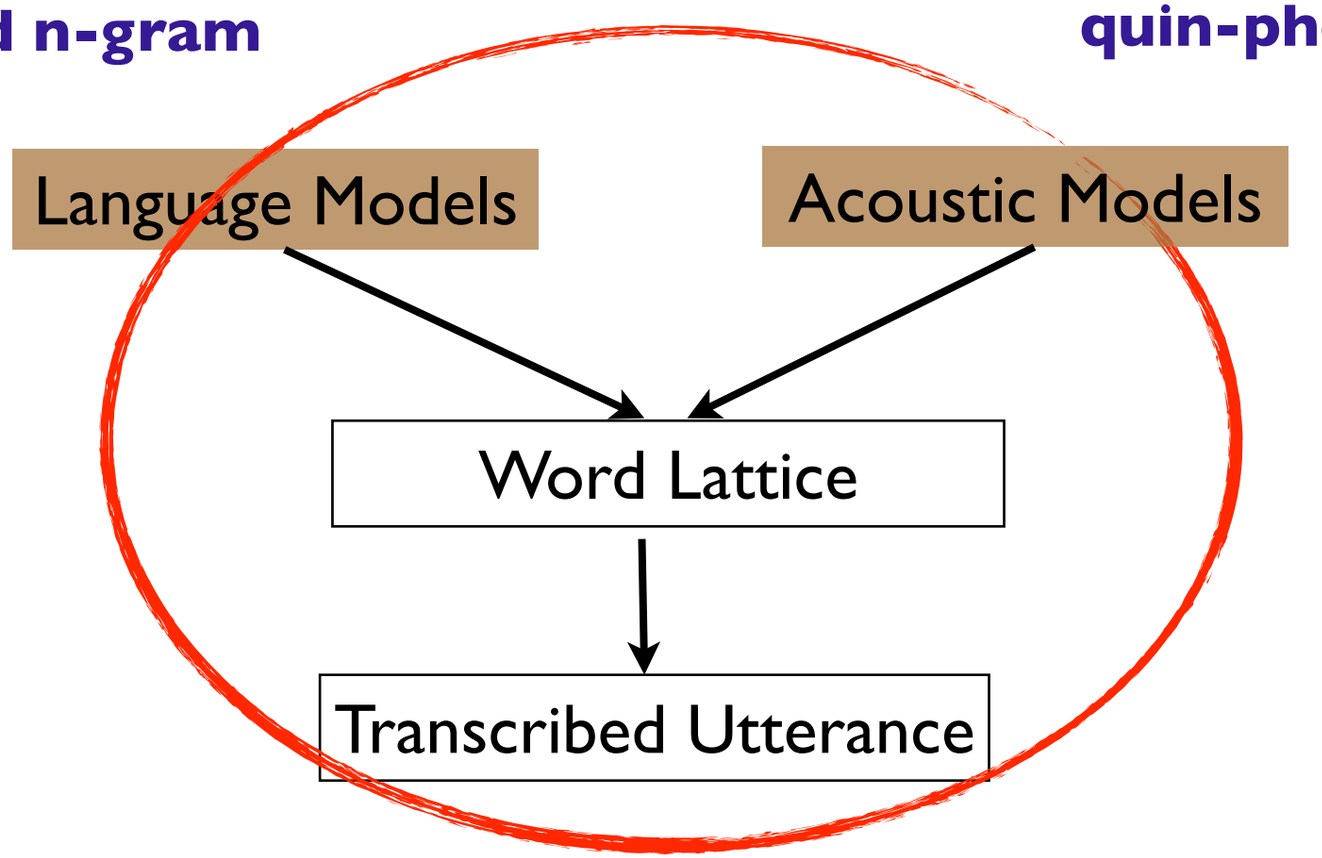
quin-phones



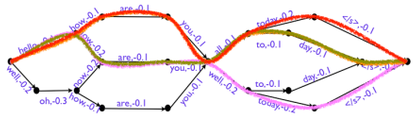


beyond n-gram

quin-phones



- Encodes exponential number of hypotheses (L)
- Deploying long-span models not always possible



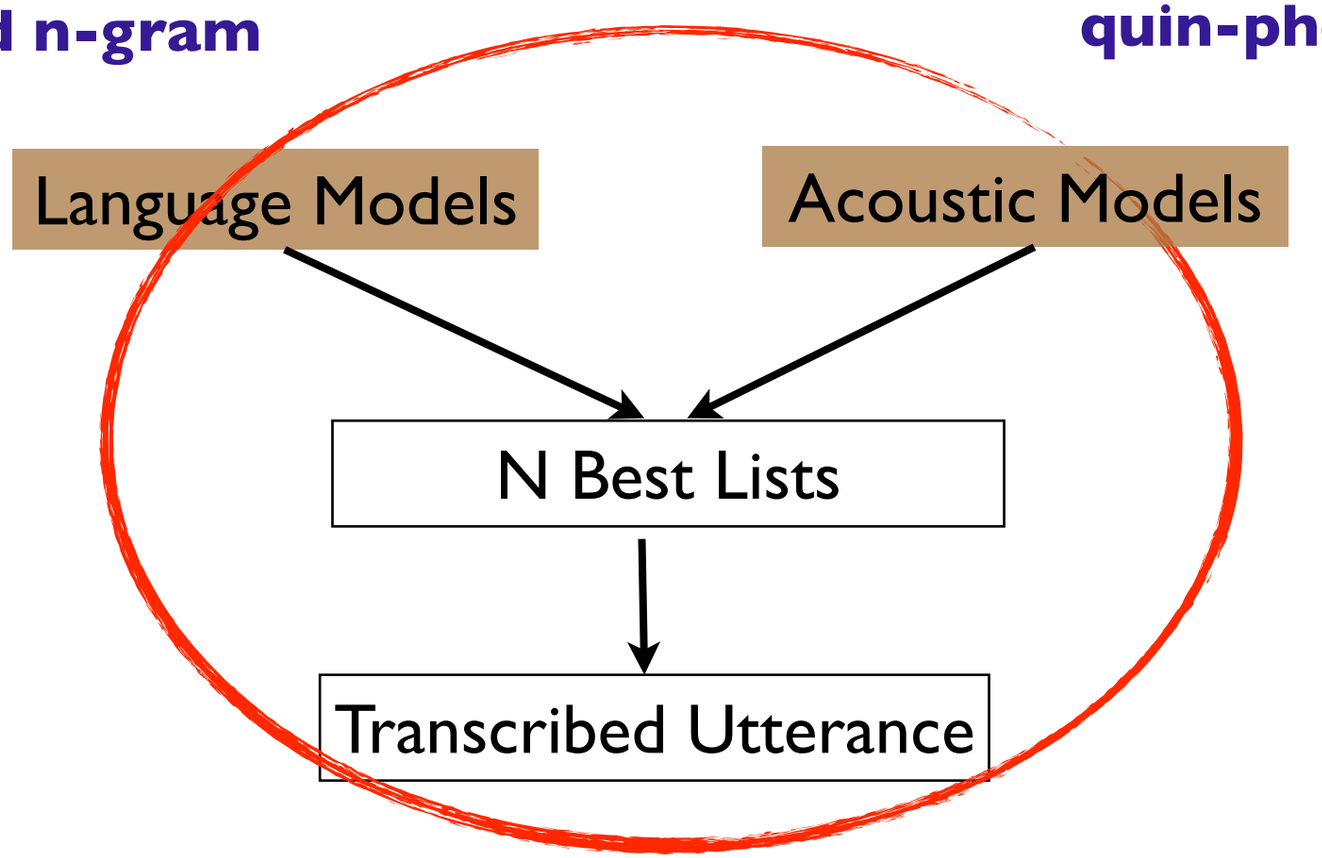
1. hello how are you all to day
2. hello now are you all to day
3. hello how are you all today

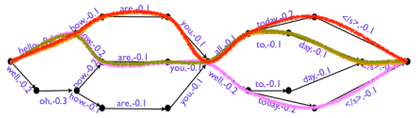
...

N. hello now are you well today

beyond n-gram

quin-phones





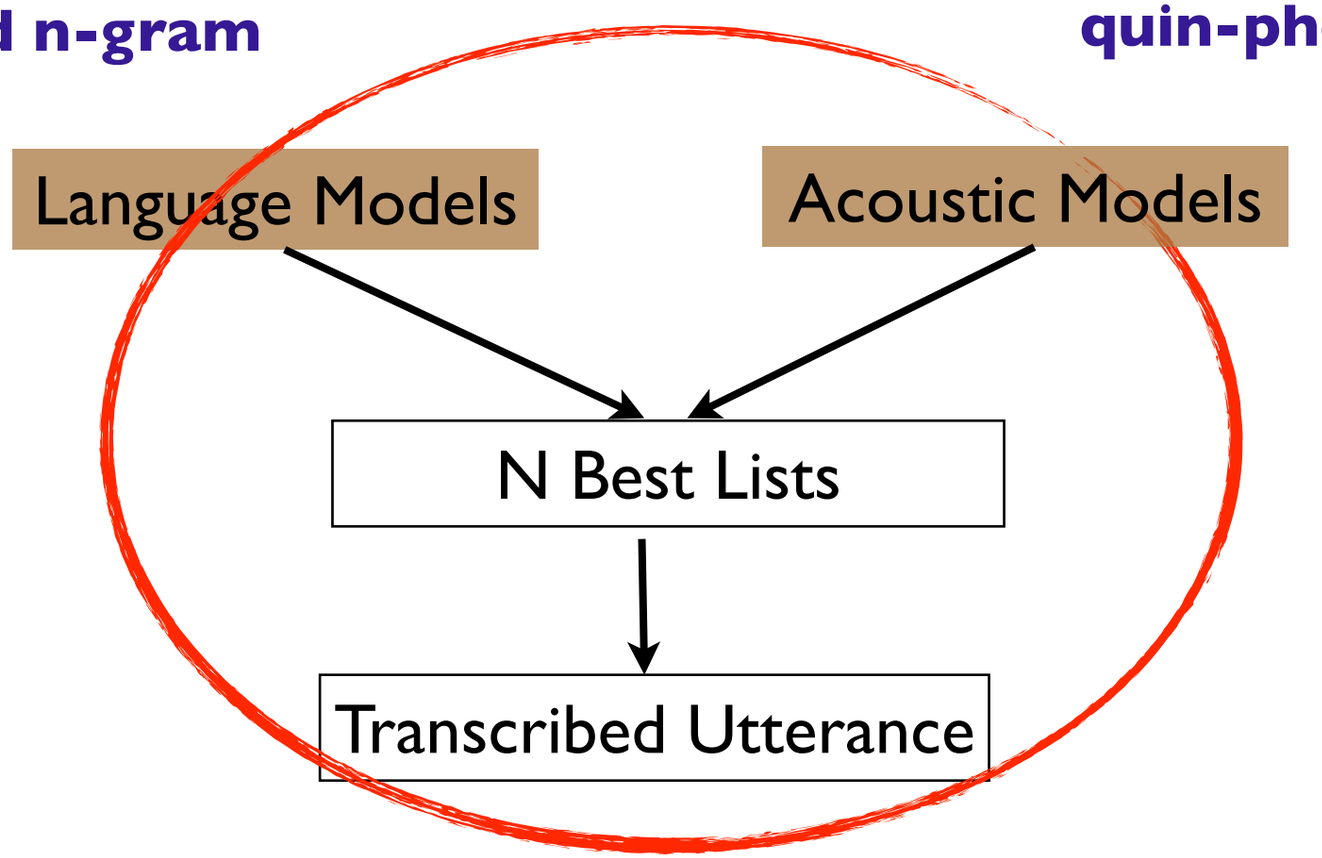
1. hello how are you all to day
2. hello now are you all to day
3. hello how are you all today

...

N. hello now are you well today

beyond n-gram

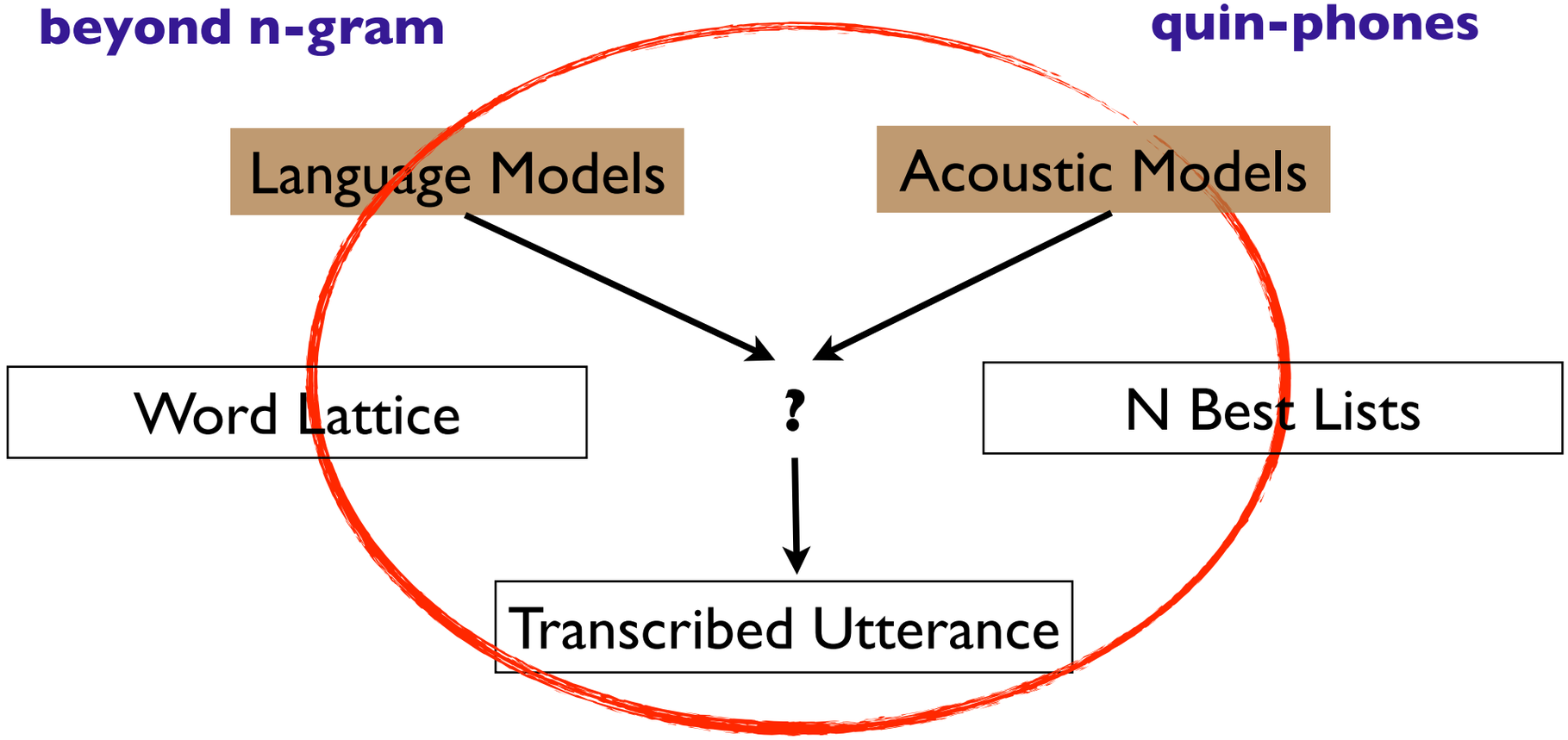
quin-phones



- Long-Span Models can be easily deployed
- N << L
- NBests are **biased** towards baseline models

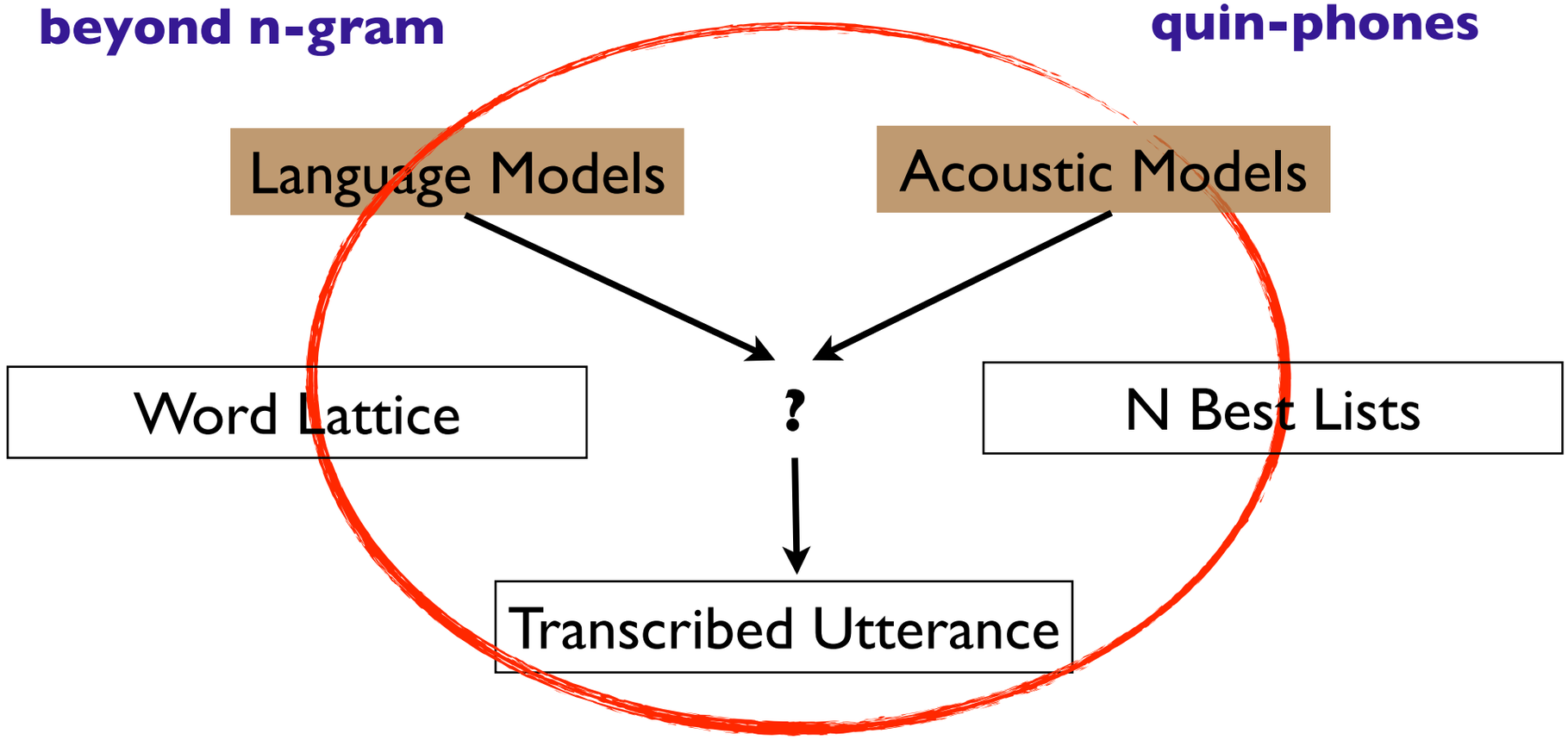
beyond n-gram

quin-phones



beyond n-gram

quin-phones



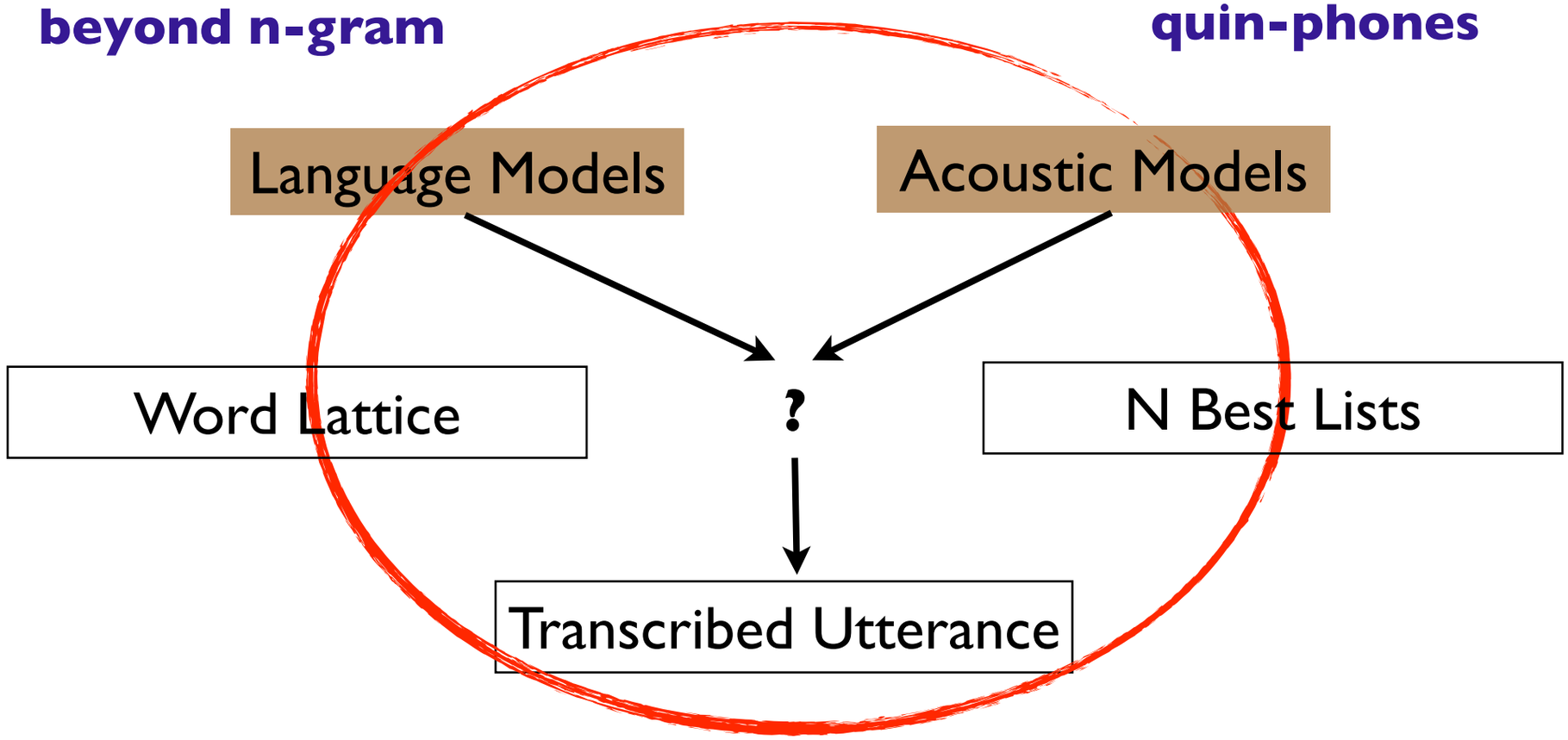
Large Search Space

Local Models

Less Biased

beyond n-gram

quin-phones



Large Search Space

Local Models

Less Biased

Limited Search Space

Long Span Models

More Biased

Large Search Space

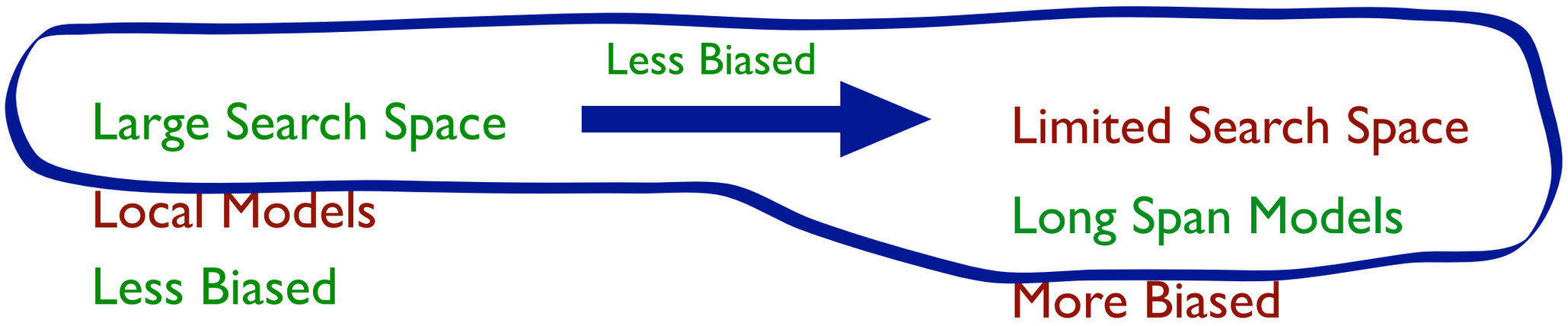
Local Models

Less Biased

Limited Search Space

Long Span Models

More Biased

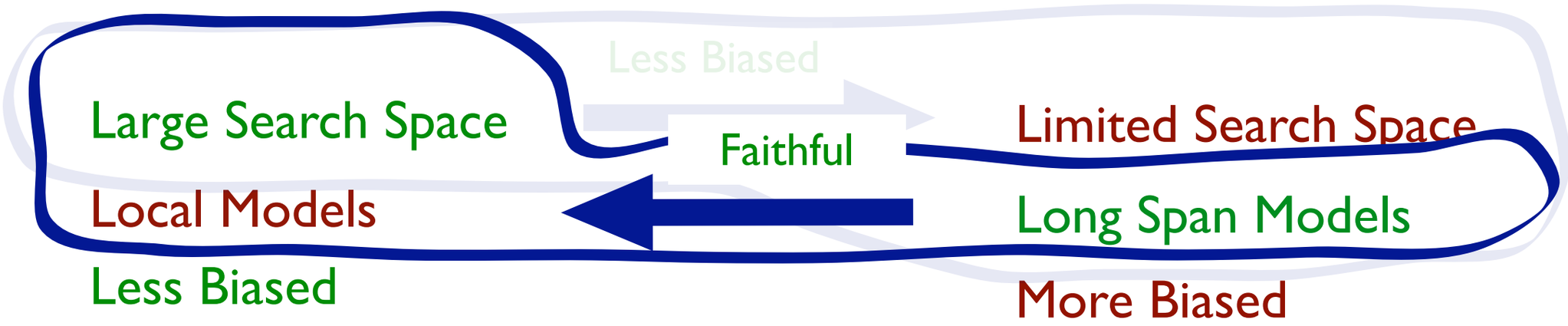


Iterative Decoding for Re-scoring

(Deoras et.al ASRU 09, Rastrow et al ICASSP 2011)

Iterative Decoding for Re-scoring

(Deoras et.al ASRU 09, Rastrow et al ICASSP 2011)

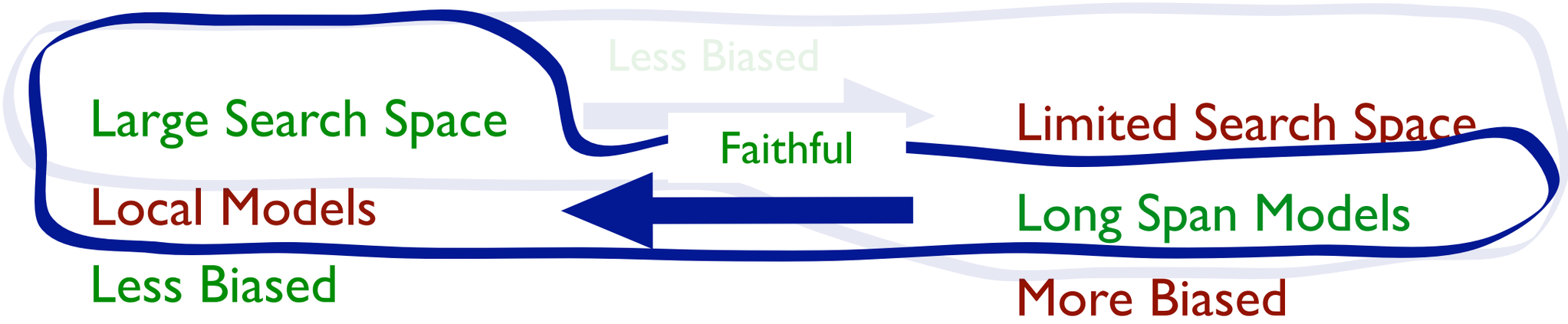


Iterative Decoding for Re-scoring

(Deoras et.al ASRU 09, Rastrow et al ICASSP 2011)

Approximation of Long Span Models

(Deoras et.al. ICASSP 11)



Iterative Decoding for Re-scoring

(Deoras et.al ASRU 09, Rastrow et al ICASSP 2011)

Approximation of Long Span Models

(Deoras et.al. ICASSP 11)

This Talk

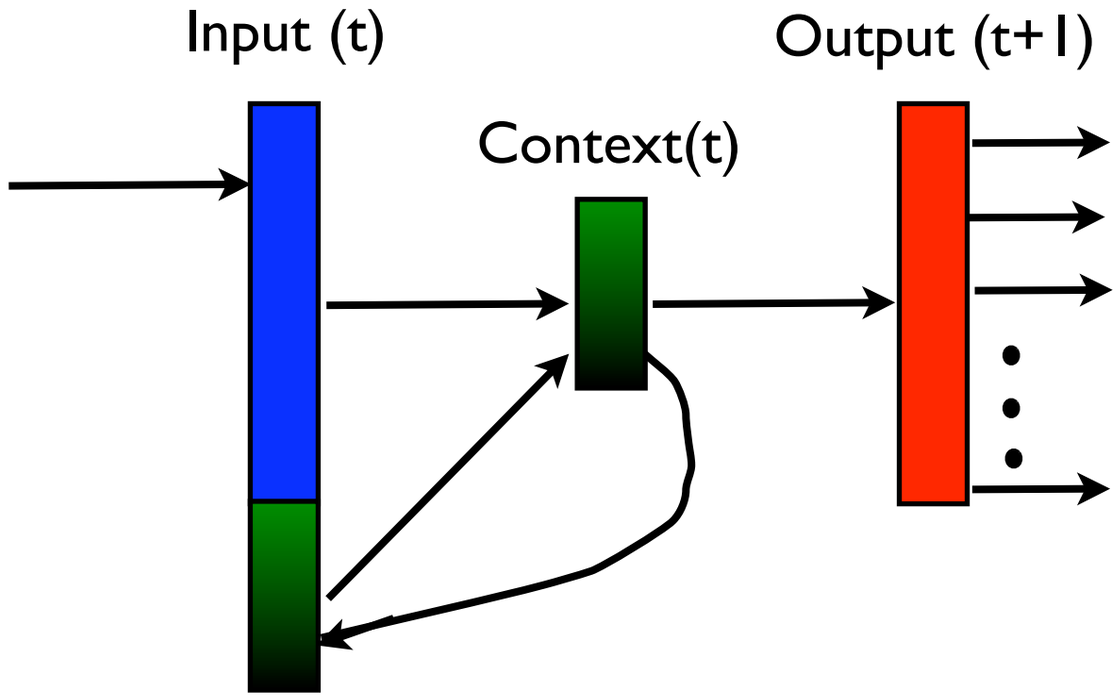
Outline

Introduction and Motivation ✓

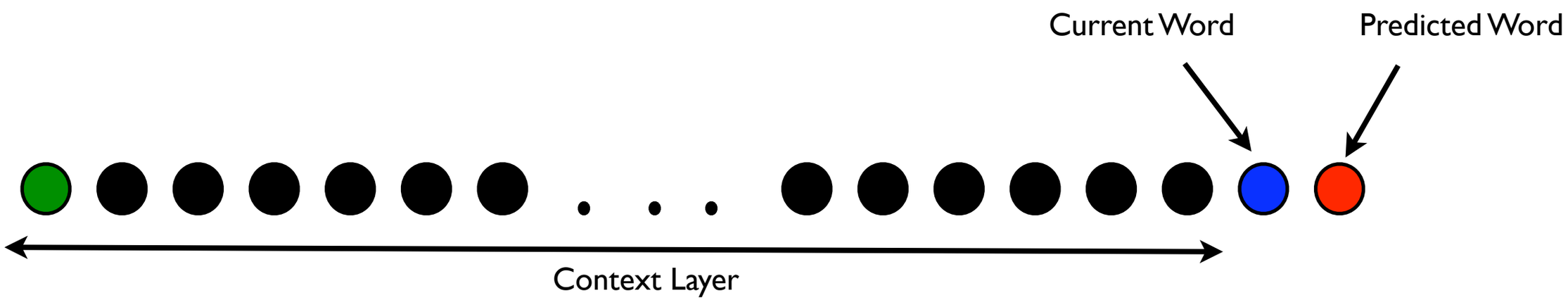
Variational Approximations

- RNN: Long-Span Models
- Variational Approximation Framework
- Experiments and Discussions
- Conclusion

Long Span Models: RNNs

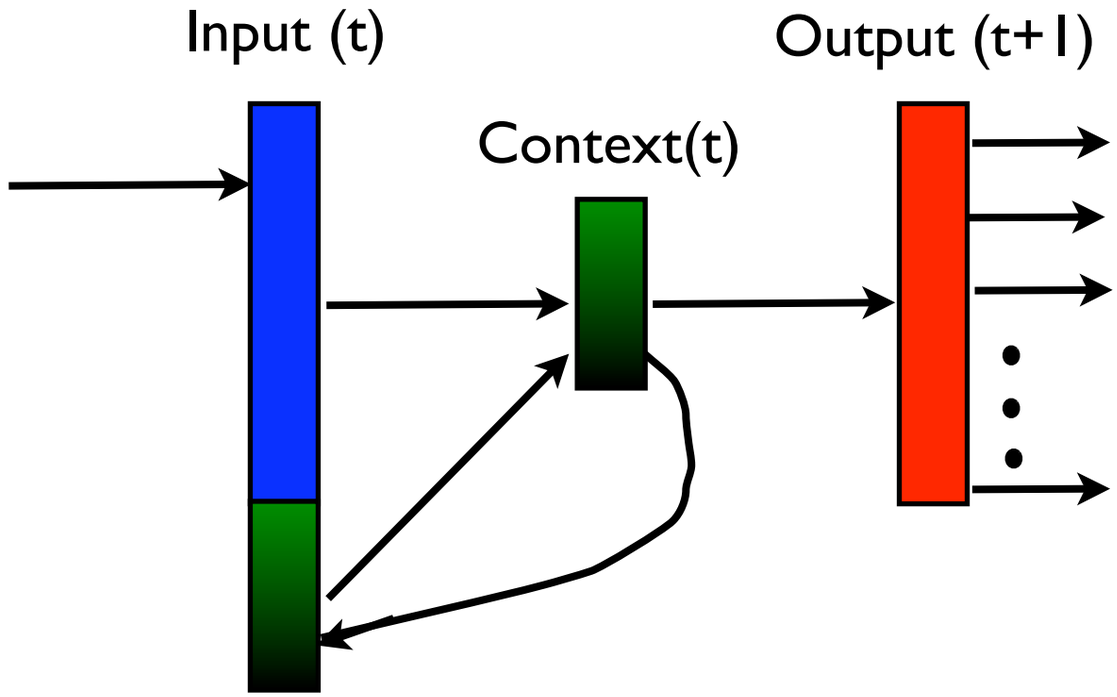


Mikolov et.al. talked about it in previous lecture



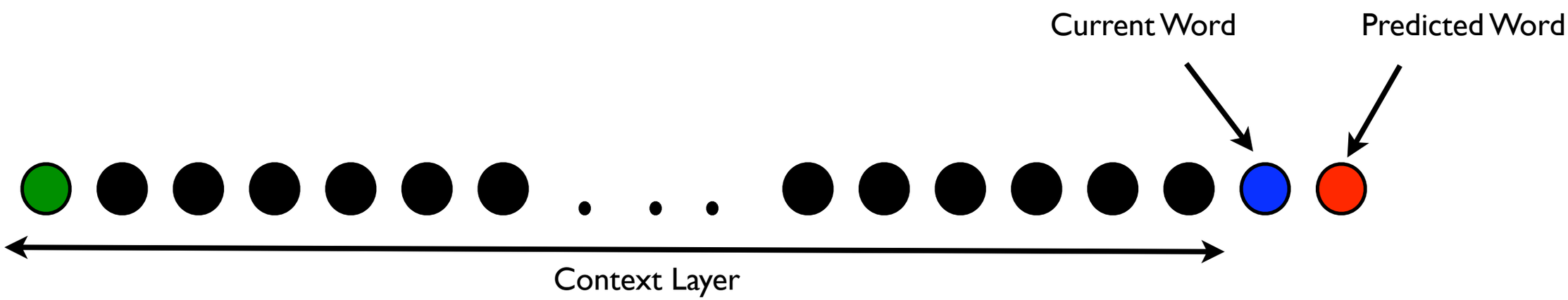
Unlike Backoff models, RNNs compute whole probability distribution at every time step.

Long Span Models: RNNs



KN(5g) : **141**
RNN : **102**

Mikolov et.al. talked about it in previous lecture



Unlike Backoff models, RNNs compute whole probability distribution at every time step.

Outline

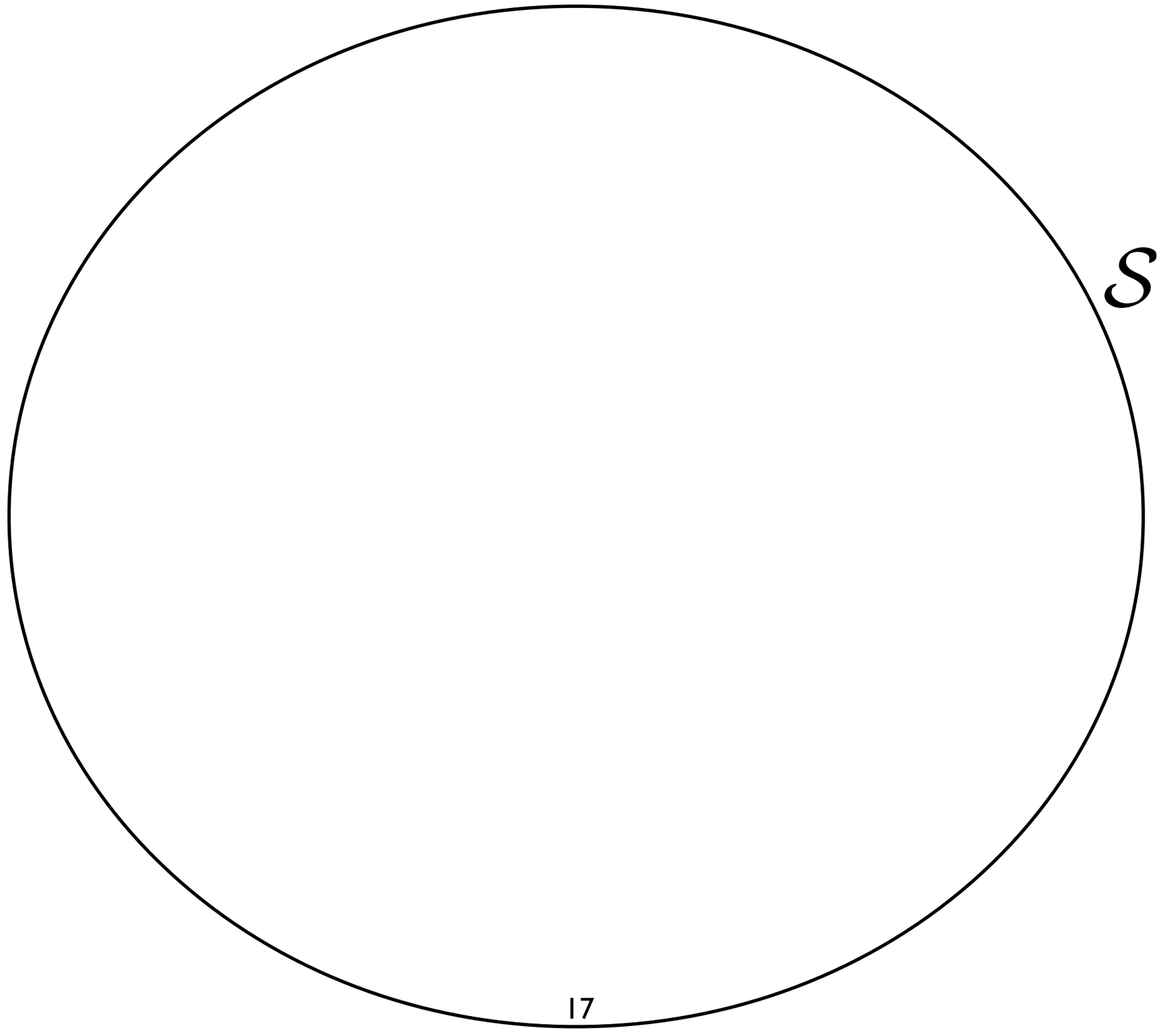
Introduction and Motivation ✓

Variational Approximations

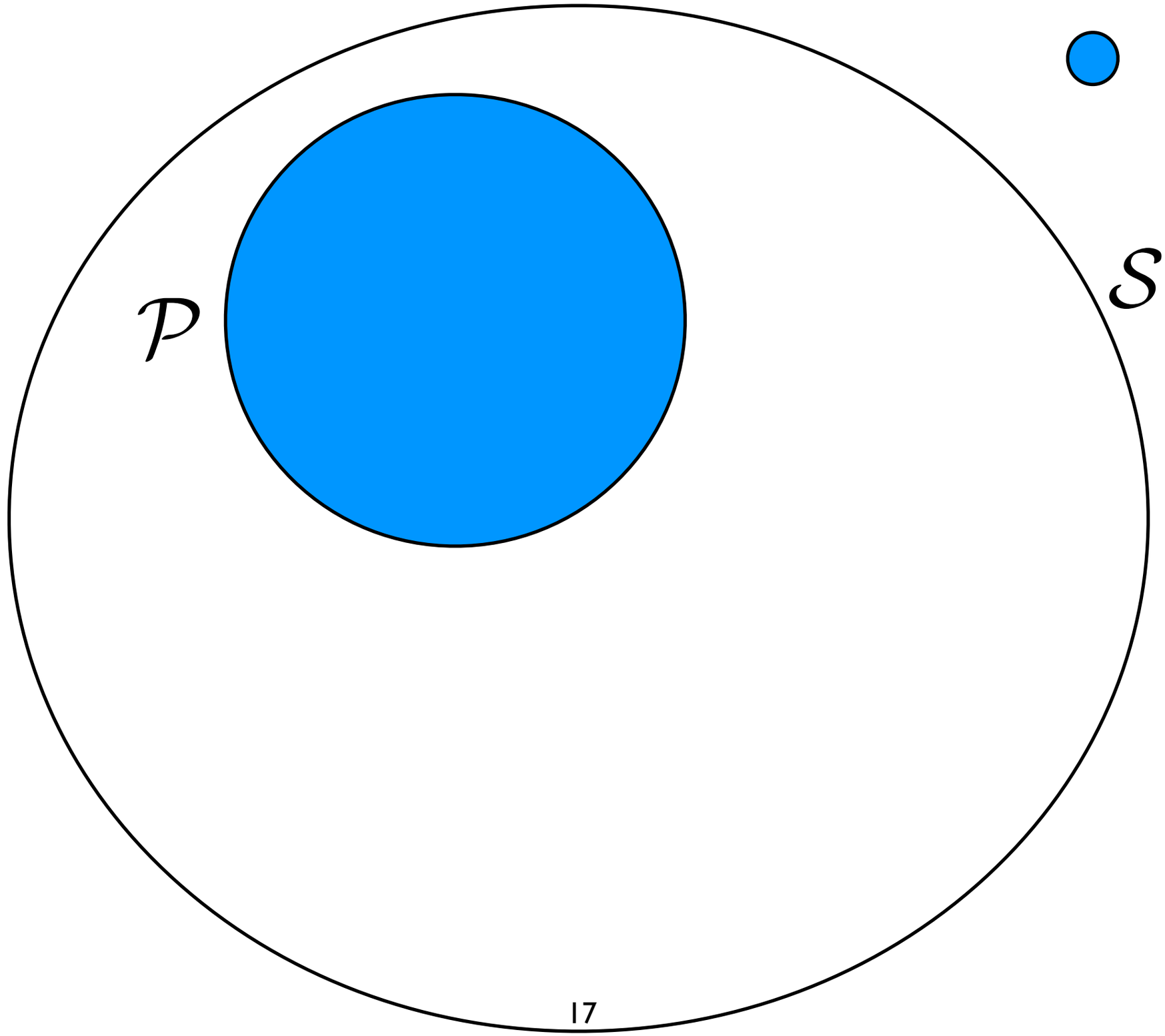
- RNN: Long-Span Models ✓
- Variational Approximation Framework
- Experiments and Discussions
- Conclusion

Variational Approximation

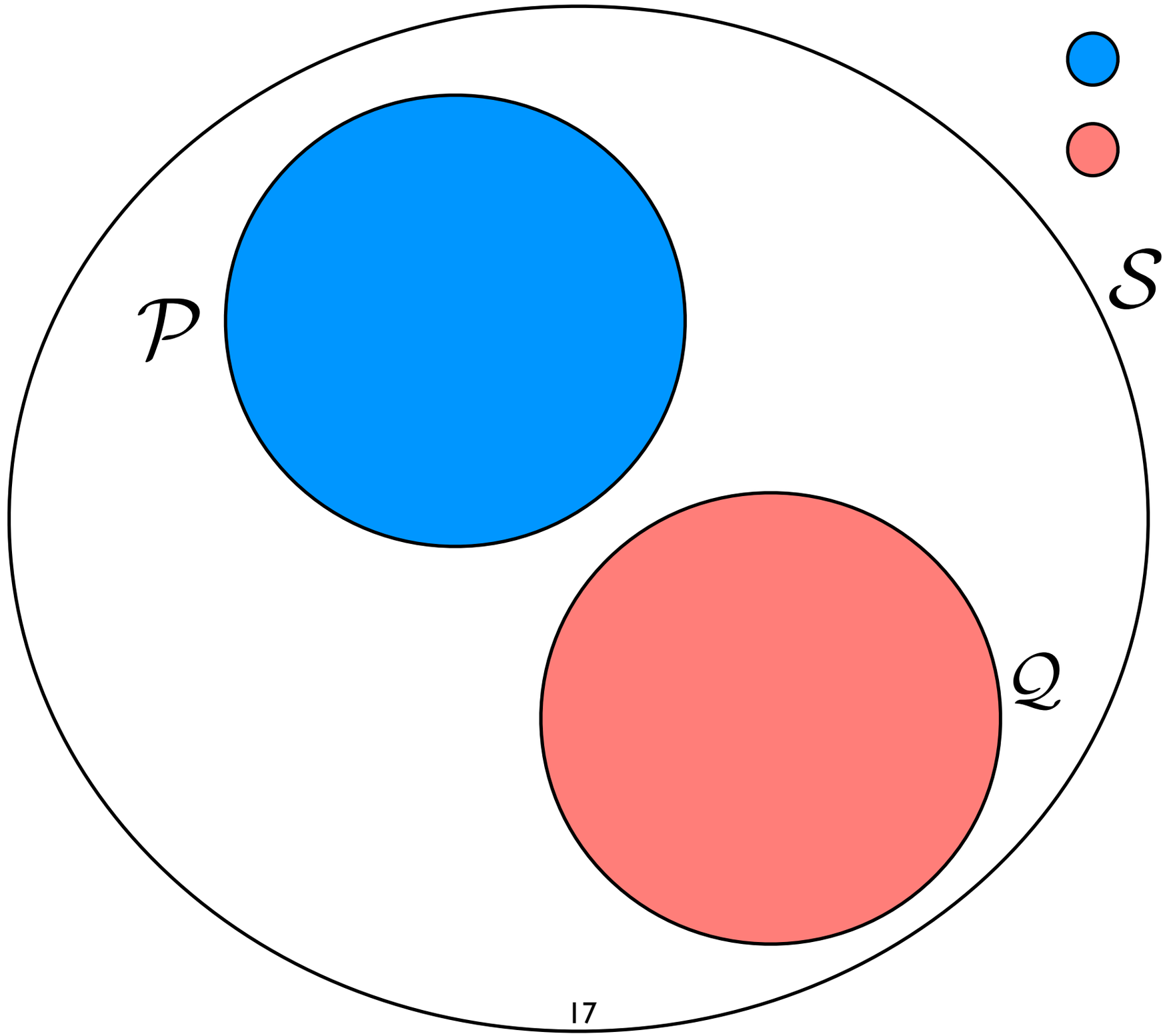
Variational Approximation



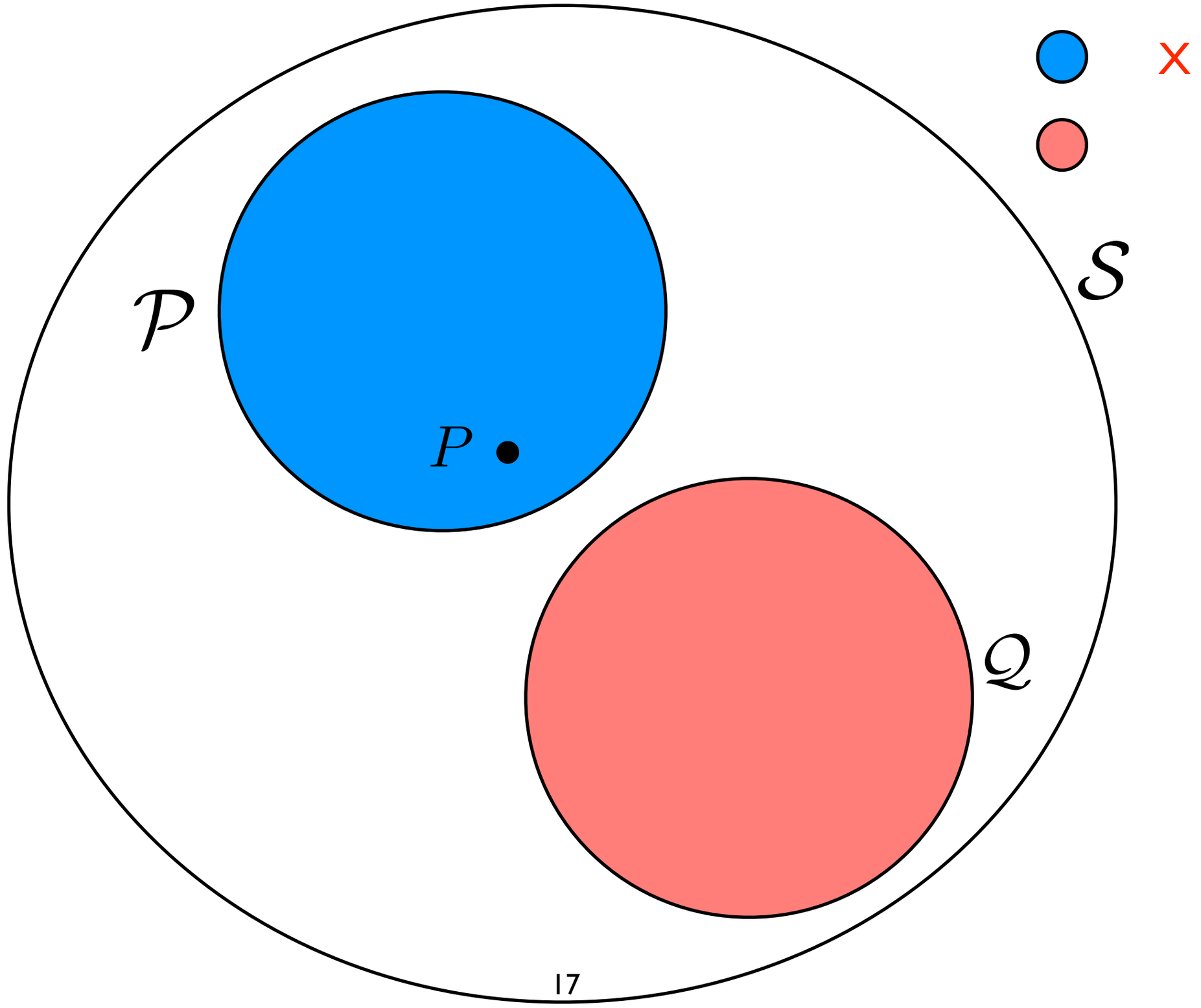
Variational Approximation



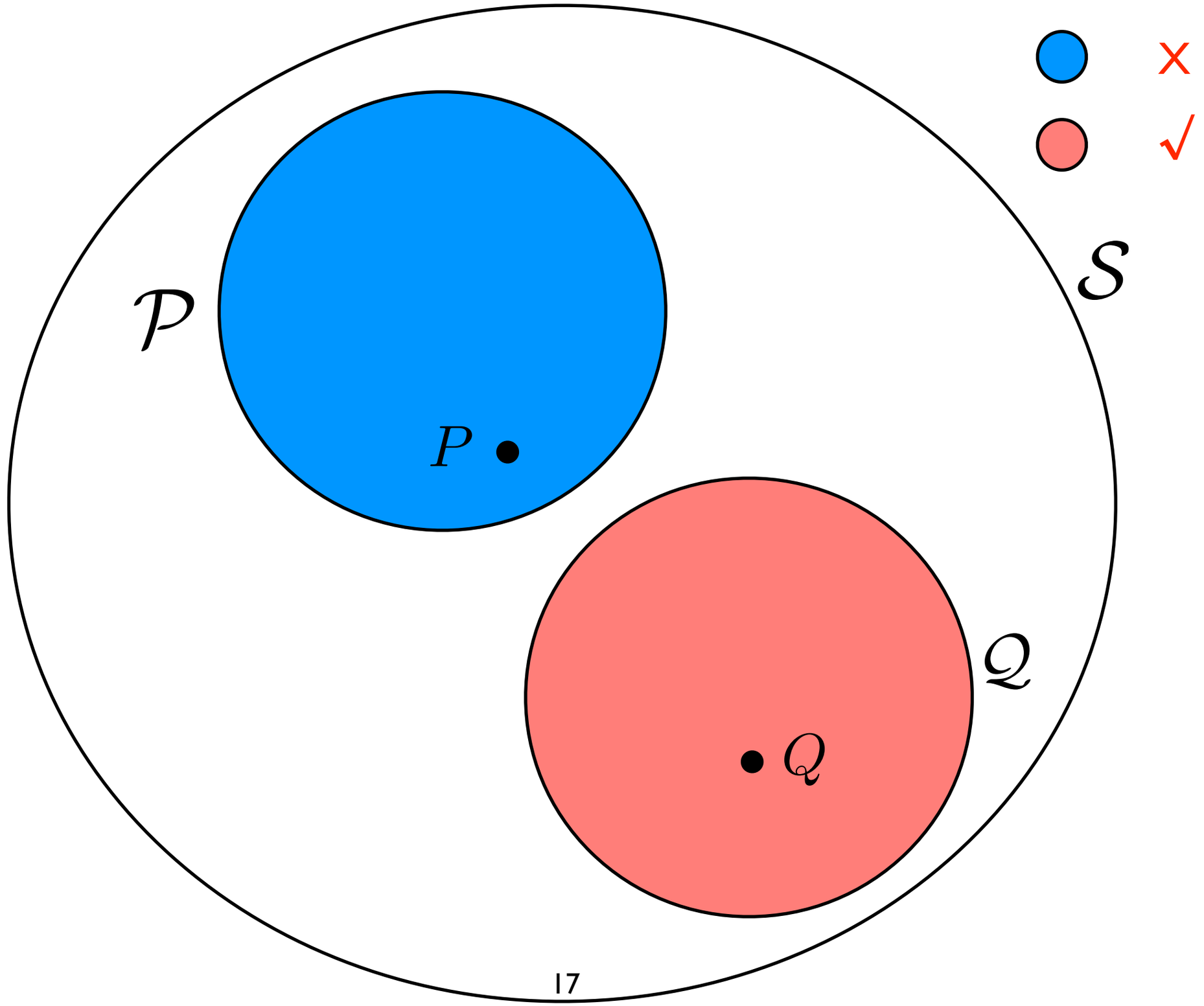
Variational Approximation



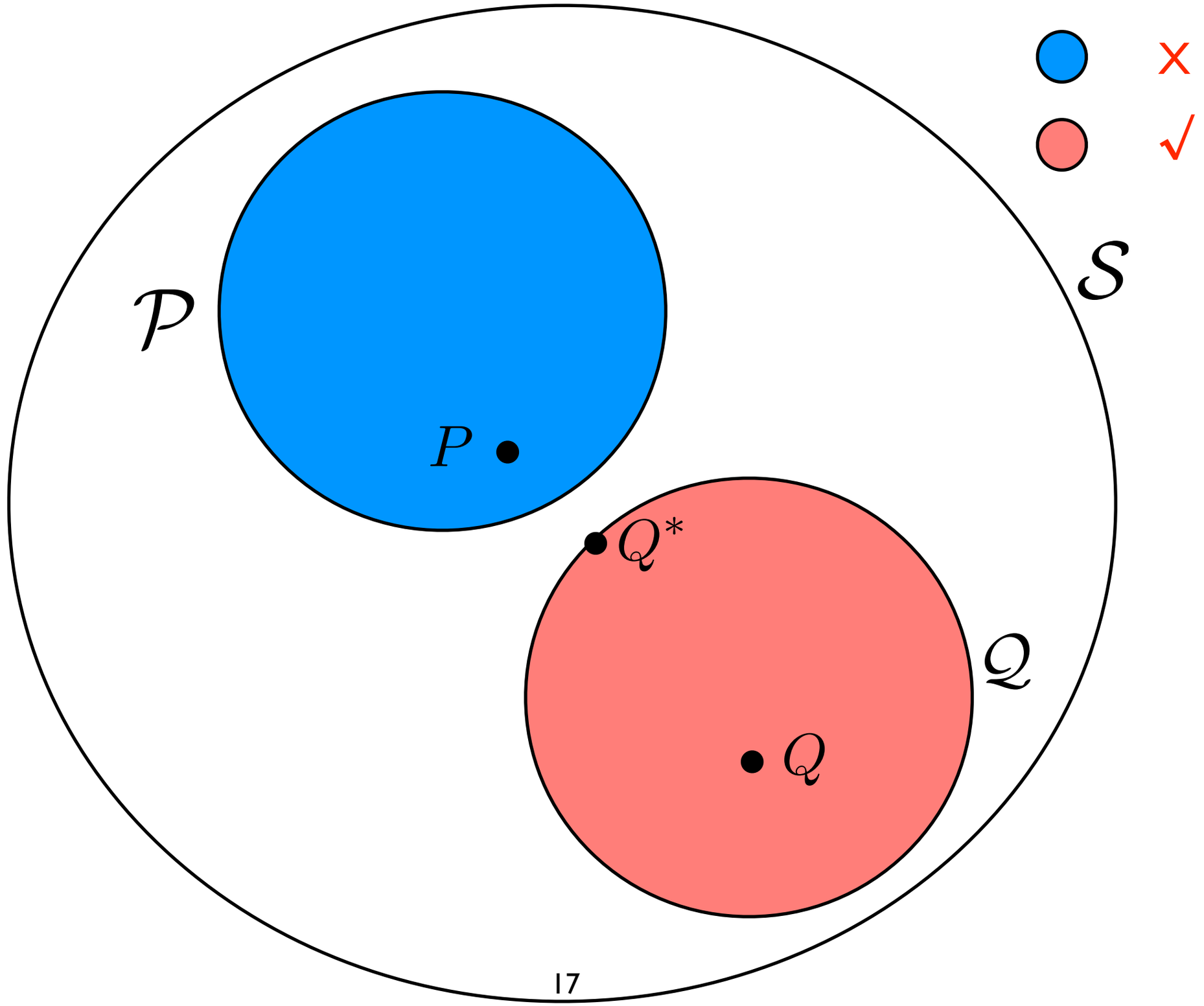
Variational Approximation



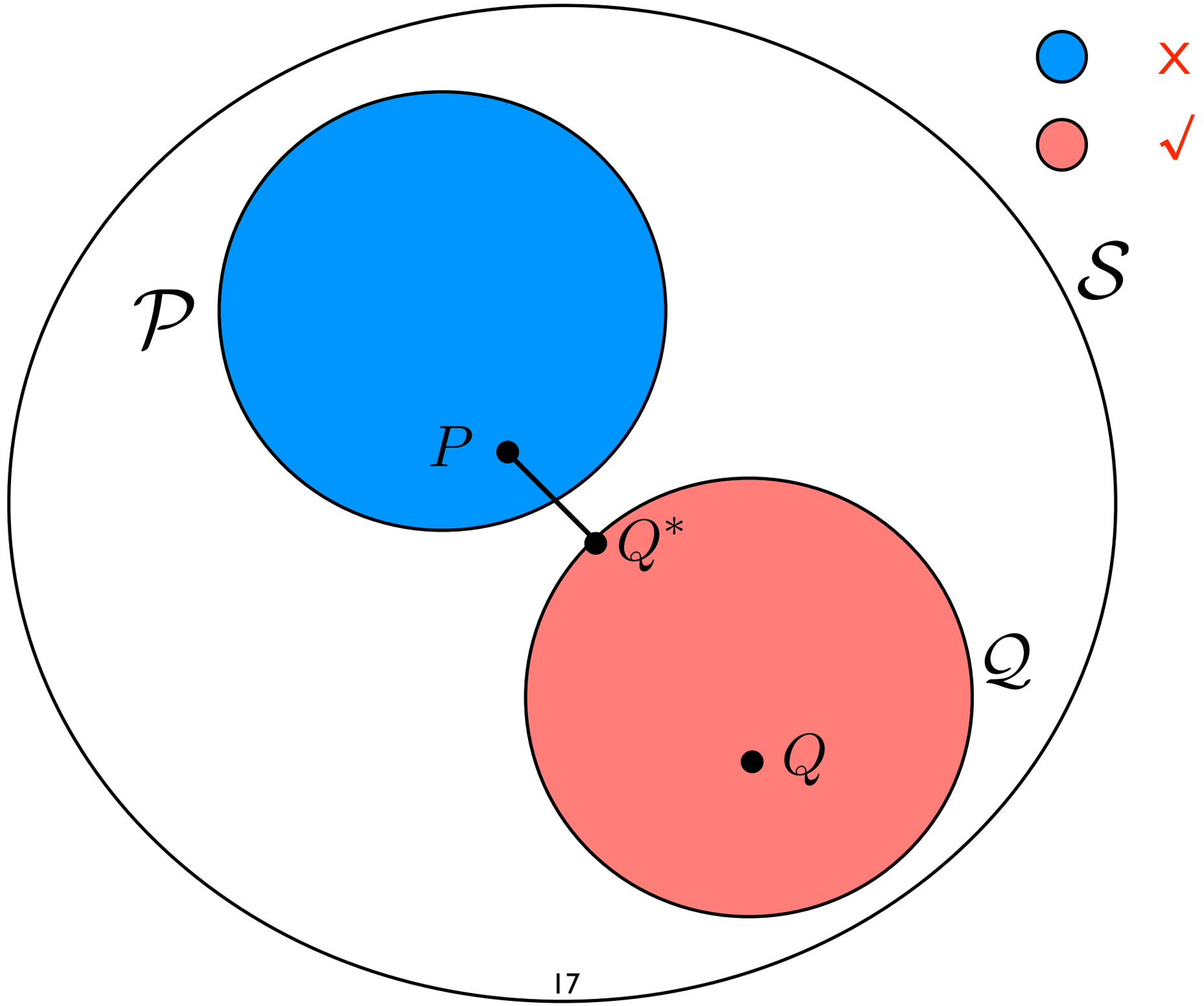
Variational Approximation



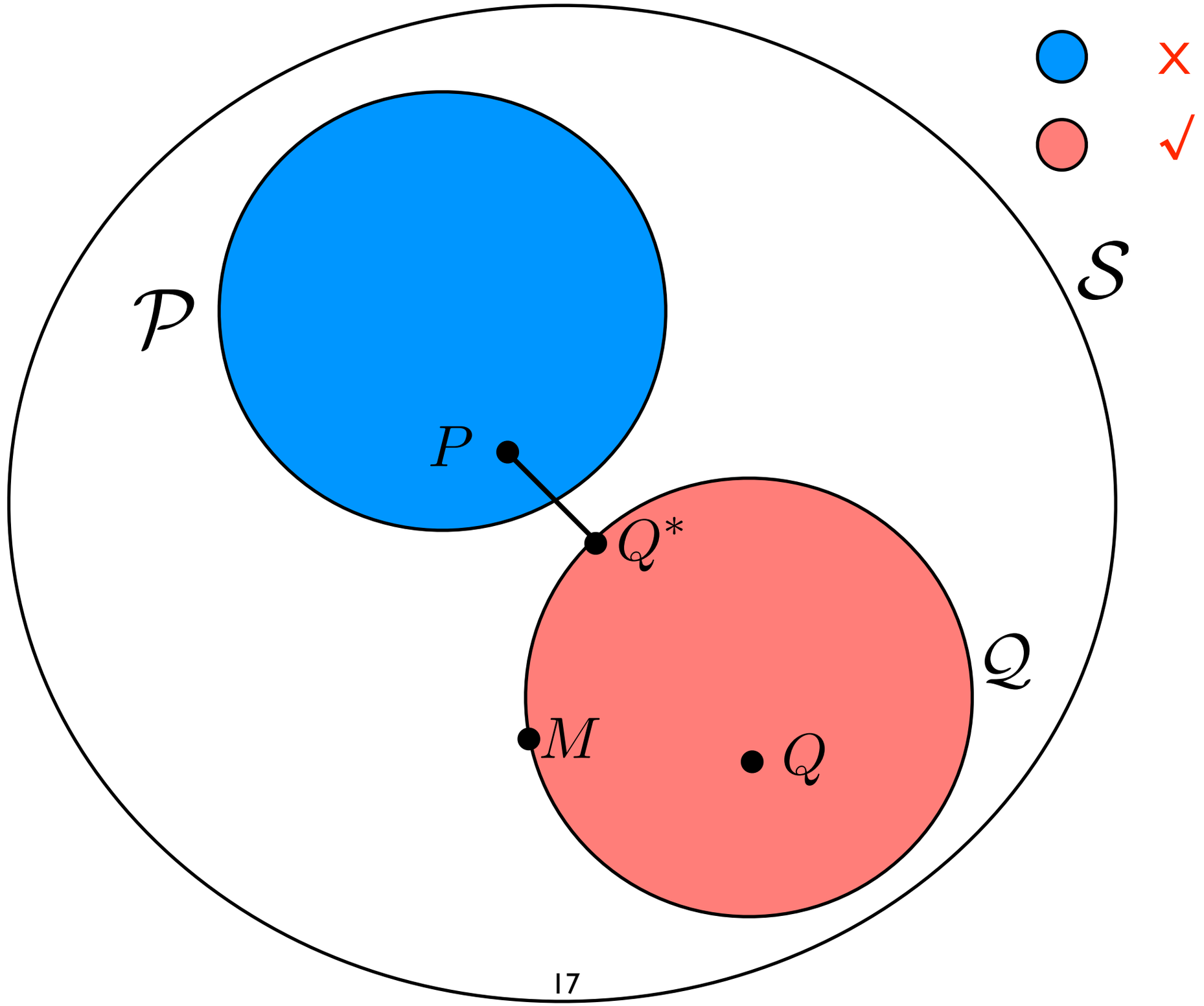
Variational Approximation

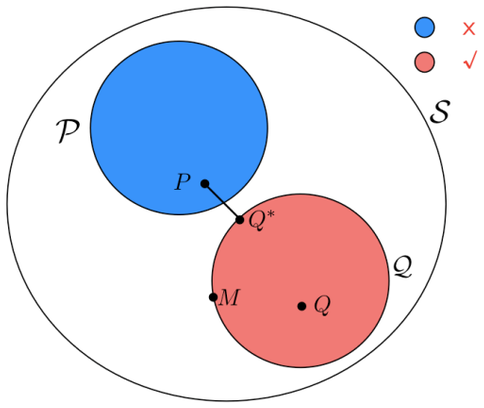


Variational Approximation

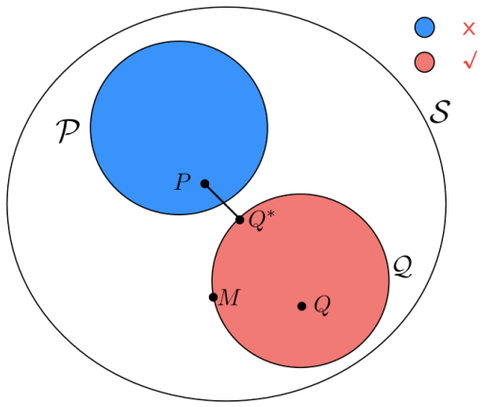


Variational Approximation





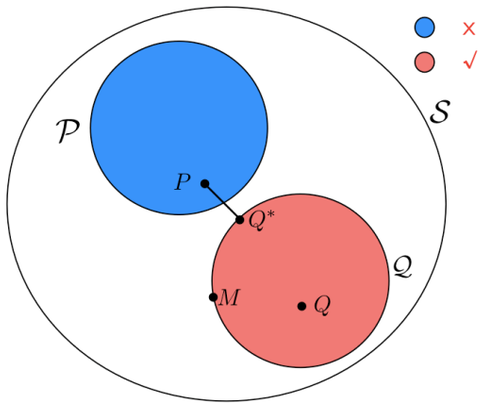
Variational Approximation



Variational Approximation

- Given a long-span model **P**, we want to do speech recognition.

P = Structured LM, Recurrent NN, Random Forest etc ..



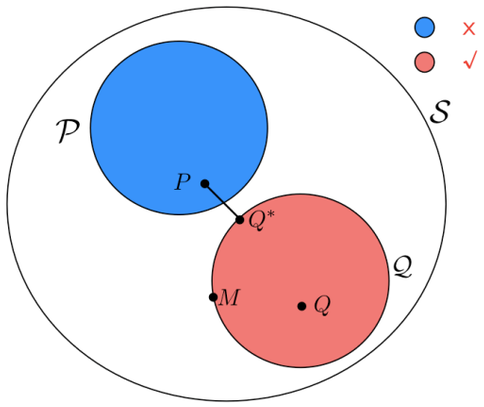
Variational Approximation

- Given a long-span model **P**, we want to do speech recognition.

P = Structured LM, Recurrent NN, Random Forest etc ..

- Decode with **M** and then do N-best re-scoring with **P**

- **M** and **P** are far



Variational Approximation

- Given a long-span model **P**, we want to do speech recognition.

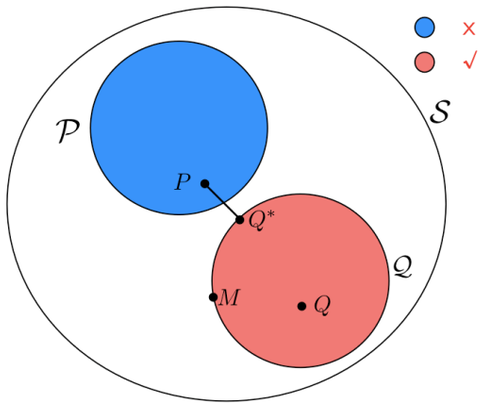
P = Structured LM, Recurrent NN, Random Forest etc ..

- Decode with **M** and then do N-best re-scoring with **P**

- **M** and **P** are far

- Find a *tractable* substitute, **Q**, and do speech recognition with it.

Q may belong to, say, Finite State Machine family.



Variational Approximation

- Given a long-span model **P**, we want to do speech recognition.

P = Structured LM, Recurrent NN, Random Forest etc ..

- Decode with **M** and then do N-best re-scoring with **P**

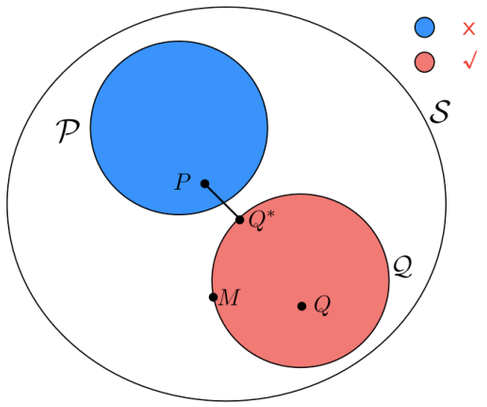
- **M** and **P** are far

- Find a *tractable* substitute, **Q**, and do speech recognition with it.

Q may belong to, say, Finite State Machine family.

- Decode with **Q** and then do N-best re-scoring with **P**

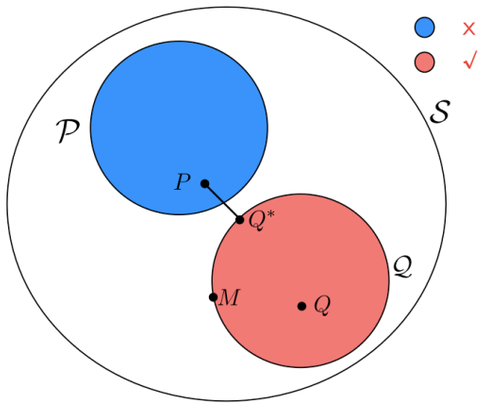
- **Q** and **P** are close



Variational Approximation

- Decode with **M** and then do N-best re-scoring with **P**

- Decode with **Q** and then do N-best re-scoring with **P**

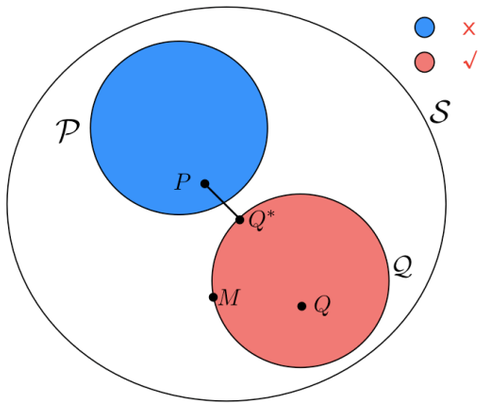


Variational Approximation

- Decode with **M** and then do N-best re-scoring with **P**

0001. I am not the Oracle
0002. I am not the Oracle
0003. I am not the Oracle
.....
0098. I am not the Oracle
0099. I am not the Oracle
0100. I am not the Oracle
.....
1996. I am not the Oracle
1997. I am not the Oracle
1998. I am not the Oracle
1999. I am not the Oracle
2000. I am the Oracle

- Decode with **Q** and then do N-best re-scoring with **P**



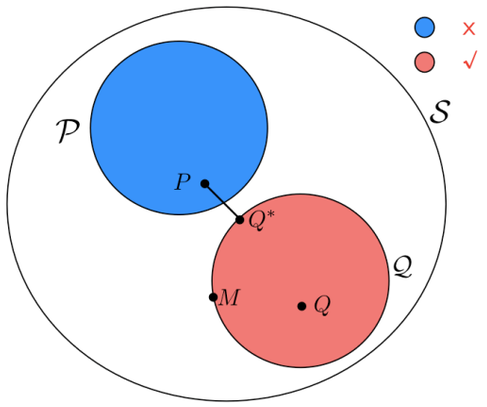
Variational Approximation

- Decode with **M** and then do N-best re-scoring with **P**

0001. I am not the Oracle
0002. I am not the Oracle
0003. I am not the Oracle
.....
0098. I am not the Oracle
0099. I am not the Oracle
0100. I am not the Oracle
.....
1996. I am not the Oracle
1997. I am not the Oracle
1998. I am not the Oracle
1999. I am not the Oracle
2000. I am the Oracle

- Decode with **Q** and then do N-best re-scoring with **P**

0001. I am not the Oracle
0002. I am not the Oracle
0003. I am not the Oracle
.....
0098. I am not the Oracle
0099. I am not the Oracle
0100. I am the Oracle



Variational Approximation

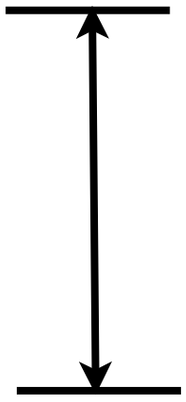
- Decode with **M** and then do N-best re-scoring with **P**

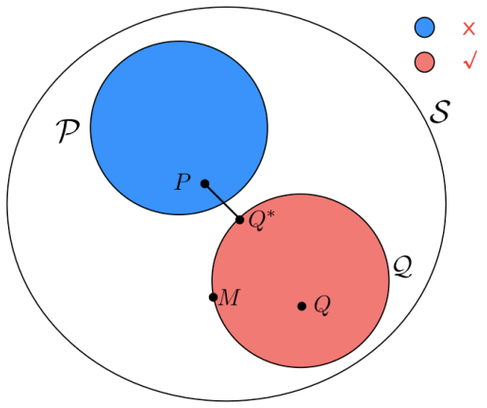
0001. I am not the Oracle
0002. I am not the Oracle
0003. I am not the Oracle
.....
0098. I am not the Oracle
0099. I am not the Oracle
0100. I am not the Oracle
.....
1996. I am not the Oracle
1997. I am not the Oracle
1998. I am not the Oracle
1999. I am not the Oracle
2000. I am the Oracle

- Decode with **Q** and then do N-best re-scoring with **P**

0001. I am not the Oracle
0002. I am not the Oracle
0003. I am not the Oracle
.....
0098. I am not the Oracle
0099. I am not the Oracle
0100. I am the Oracle

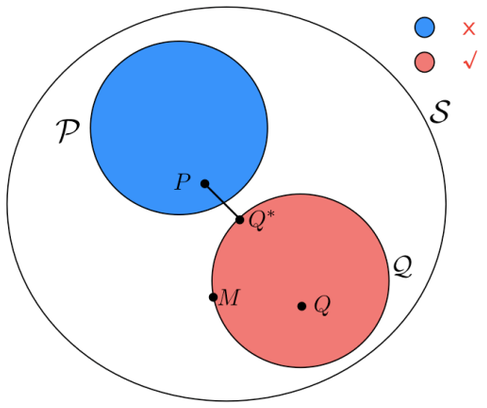
Save Effort





Variational Approximation

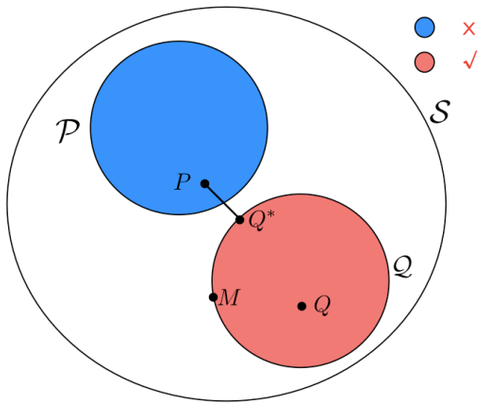
$$Q^* \quad = \quad \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q)$$



Variational Approximation

$$Q^* \quad = \quad \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q)$$

- Restrict the family to n-grams and find a solution in this family.

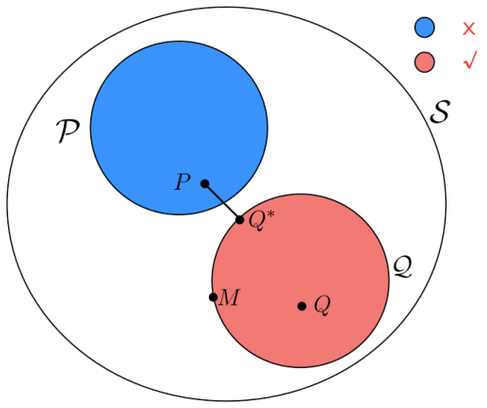


Variational Approximation

$$Q^* \quad = \quad \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q)$$

- Restrict the family to n-grams and find a solution in this family.

- Under some mild conditions, the solution is the **marginalized** version of the long span model.



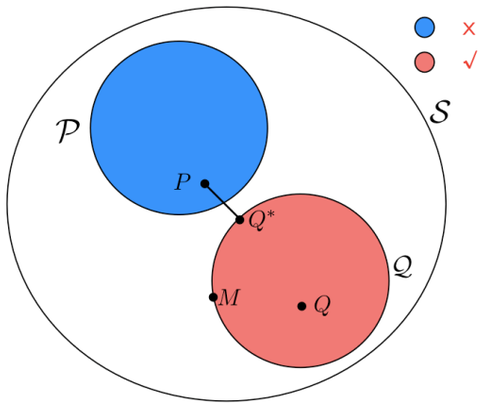
Variational Approximation

$$Q^* \quad = \quad \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q)$$

- Restrict the family to n-grams and find a solution in this family.

- Under some mild conditions, the solution is the **marginalized** version of the long span model.

- Marginalization is extremely difficult for long context models.



Variational Approximation

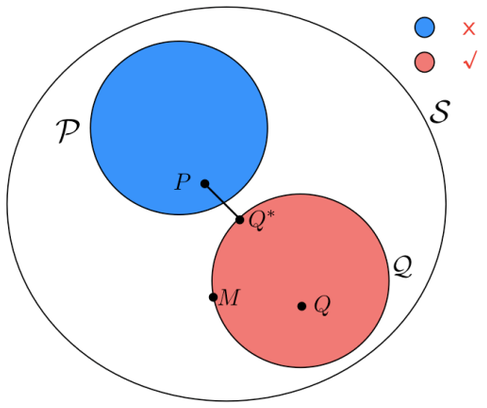
$$Q^* \quad = \quad \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q)$$

- Restrict the family to n-grams and find a solution in this family.

- Under some mild conditions, the solution is the **marginalized** version of the long span model.

- Marginalization is extremely difficult for long context models.

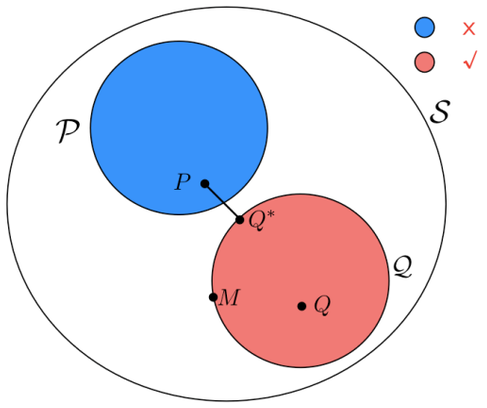
- We obtain solution via **sampling** techniques.



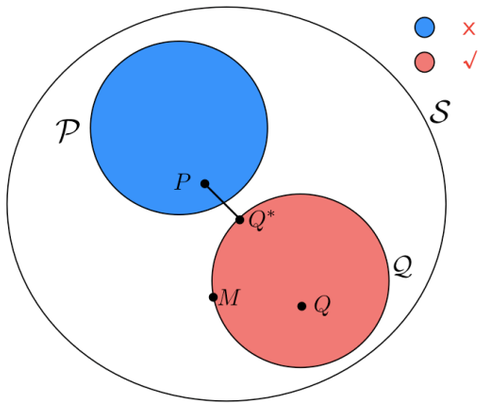
Variational Approximation

$$
\begin{aligned}
Q^* &= \arg\min_{Q \in \mathcal{Q}} \mathrm{D}(P \| Q) \\
&= \arg\min_{Q \in \mathcal{Q}} \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \\
&= \arg\max_{Q \in \mathcal{Q}} \sum_{x \in \mathcal{X}} P(x) \log Q(x)
\end{aligned}
$$

ML Solution

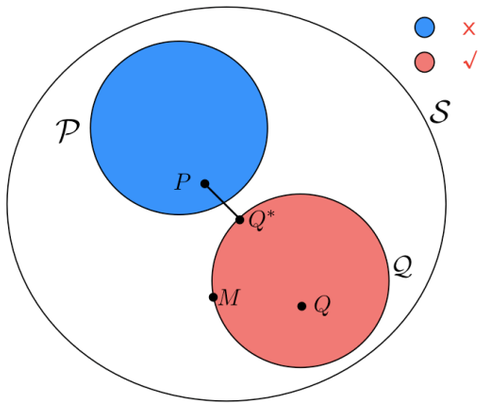


Variational Approximation



Variational Approximation

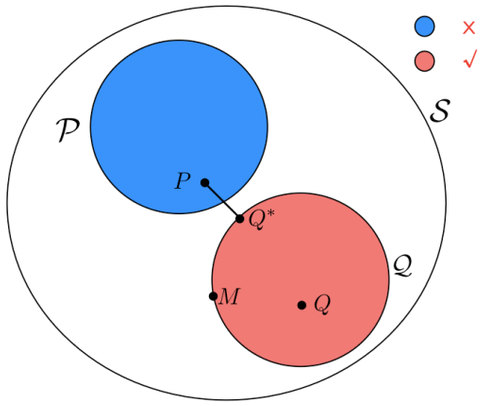
Recipe:



Variational Approximation

Recipe:

1. Get a sophisticated long-span model, **P**

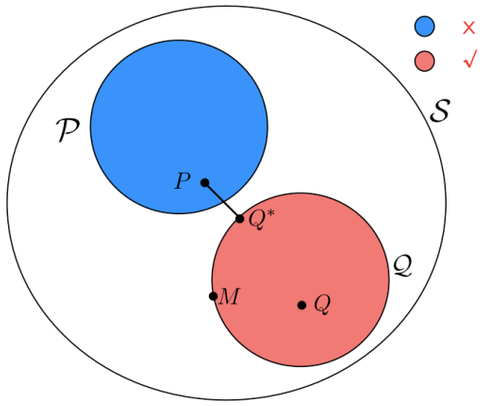


Variational Approximation

Recipe:

1. Get a sophisticated long-span model, **P**

2. Decide on n-gram family.



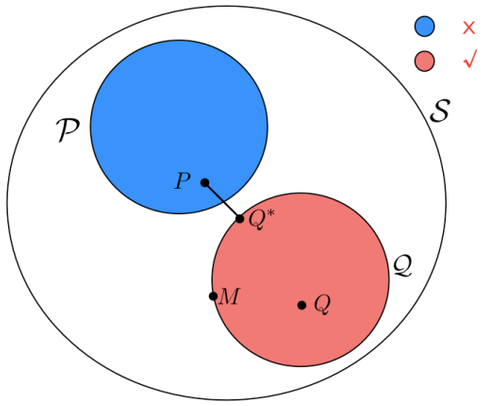
Variational Approximation

Recipe:

1. Get a sophisticated long-span model, **P**

2. Decide on n-gram family.

3. "Synthesize" a huge corpus using **P**



Variational Approximation

Recipe:

1. Get a sophisticated long-span model, **P**

2. Decide on n-gram family.

3. "Synthesize" a huge corpus using **P**

4. Estimate **Q** using Maximum Likelihood.

Q belongs to n-gram family of distributions.

$$\lim_{n\to\infty} \lim_{L\to\infty} KL(P||Q^*) = 0$$

Outline

Introduction and Motivation ✓

Variational Approximations

- RNN: Long-Span Models ✓
- Variational Approximation Framework ✓
- Experiments and Discussions
- Conclusion

Experiments and Results

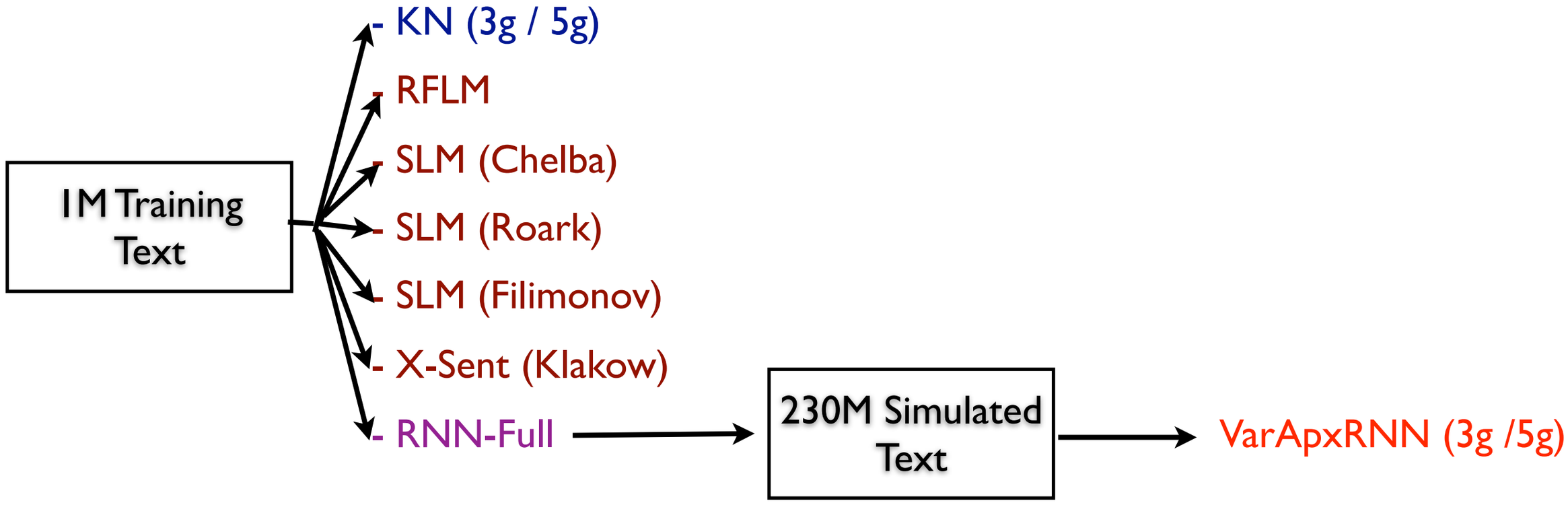
Penn Tree Bank Corpus (sections 00-24)

1M word token for training

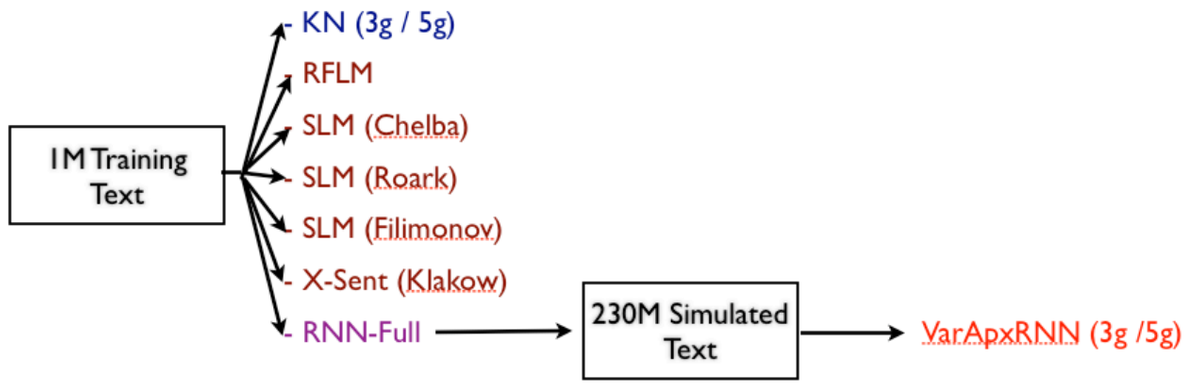
0.2M word token for testing

Top **10K** most frequent words for Vocabulary

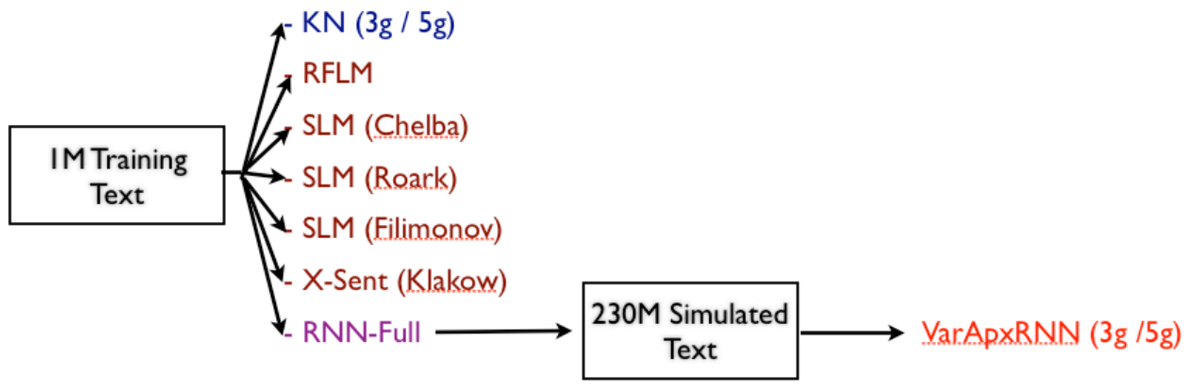
1. Perplexity Experiments



Experiments and Results

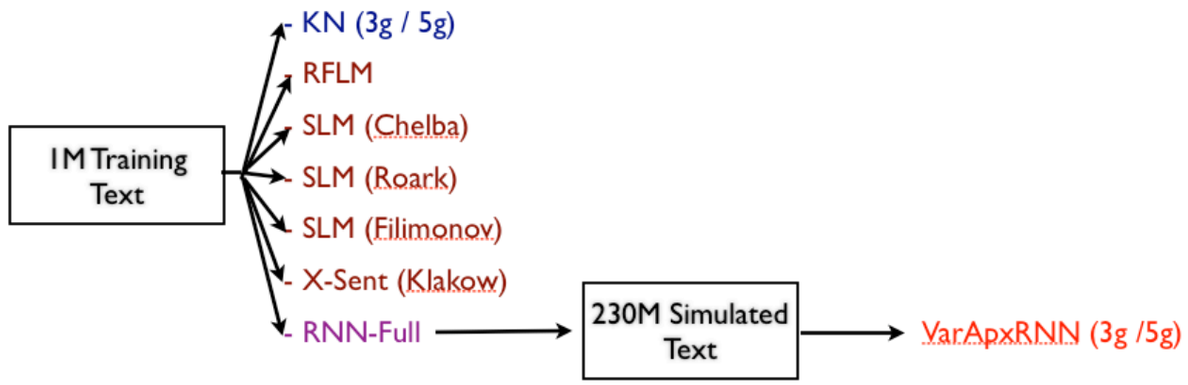


Experiments and Results



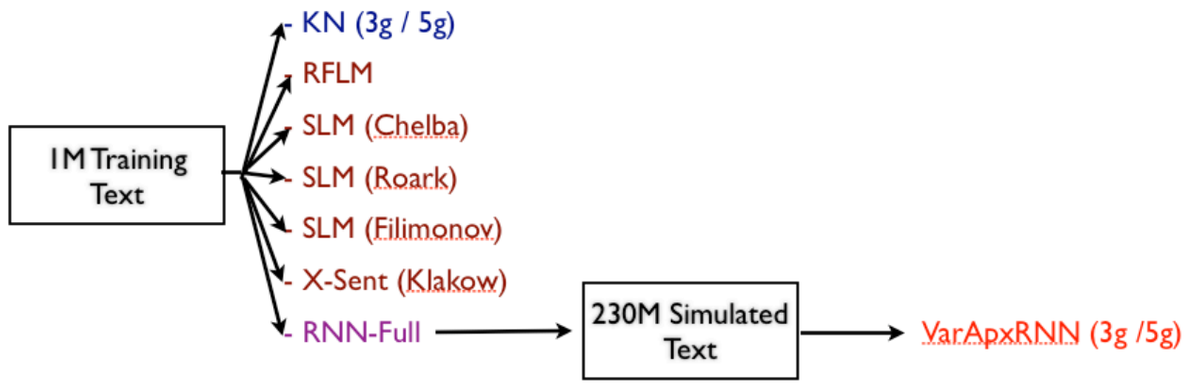
Setup	PPL	Setup	PPL
KN (3g)	148	Random Forest (Xu)	132
VarApxRNN (3g)	152	-	-
VarApx+KN (3g)	**124**	-	-

Experiments and Results



Setup	PPL	Setup	PPL
KN (3g)	148	Random Forest (Xu)	132
VarApxRNN (3g)	152	-	-
VarApx+KN (3g)	**124**	-	-
KN (5g)	141	SLM (Chelba)	149
VarApxRNN (5g)	140	SLM (Roark)	137
VarApx+KN (5g)	**120**	SLM (Filimonov)	125

Experiments and Results



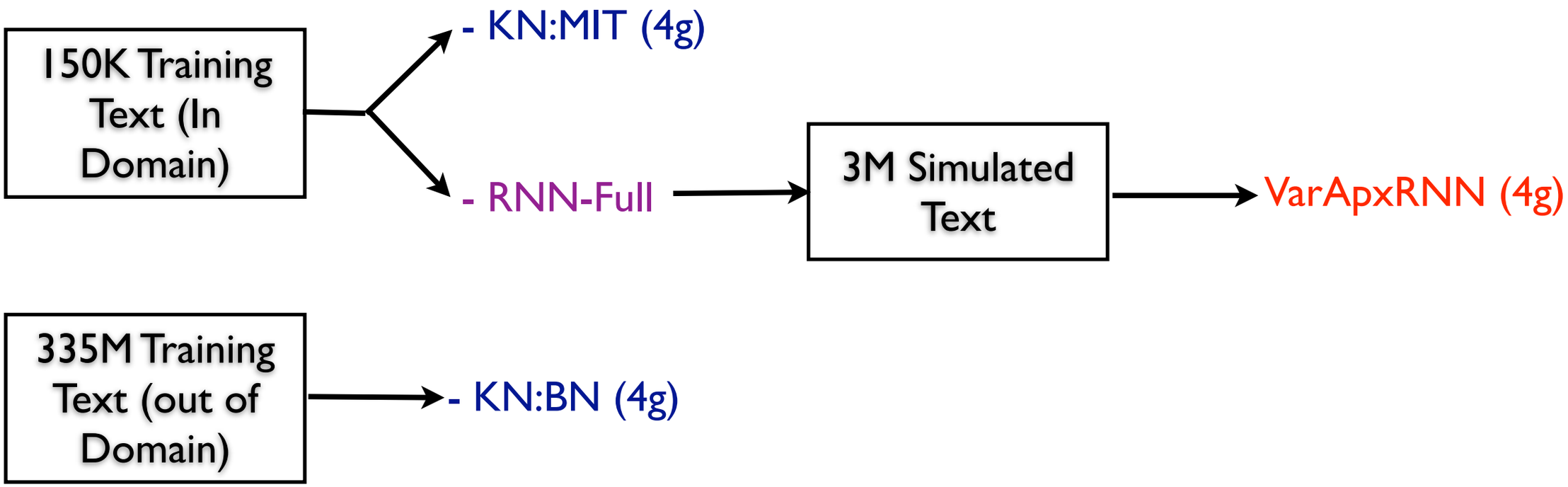
Setup	PPL	Setup	PPL
KN (3g)	148	Random Forest (Xu)	132
VarApxRNN (3g)	152	-	-
VarApx+KN (3g)	**124**	-	-
KN (5g)	141	SLM (Chelba)	149
VarApxRNN (5g)	140	SLM (Roark)	137
VarApx+KN (5g)	**120**	SLM (Filimonov)	125
VarApx+KN + Cache	111	X-Sent (Momtazi)	118
RNN-Full	102	-	-

Experiments and Results

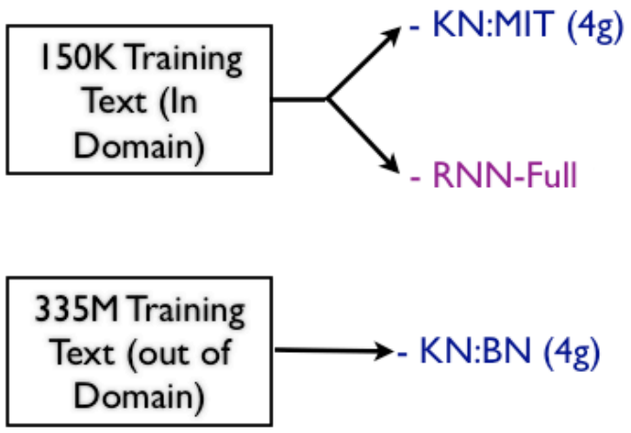
In Domain: MIT Lectures

Out of Domain: BN

2. ASR: Adaptation Experiments



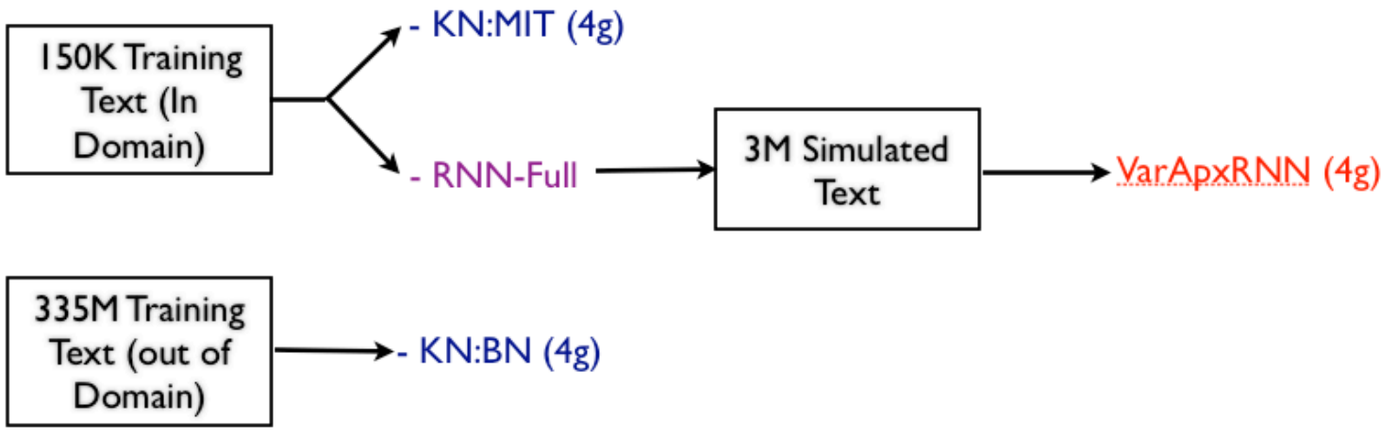
Experiments and Results



Setup	Set 1	Set 2
KN:MIT+BN (4g) decoding	24.7	22.4
+ RNN-Full rescoring (100 best)	24.1	22.4
+ RNN-Full rescoring (2000 best)	23.8	21.6
Oracle (2000 best)	17.9	15.5

Baseline

Experiments and Results



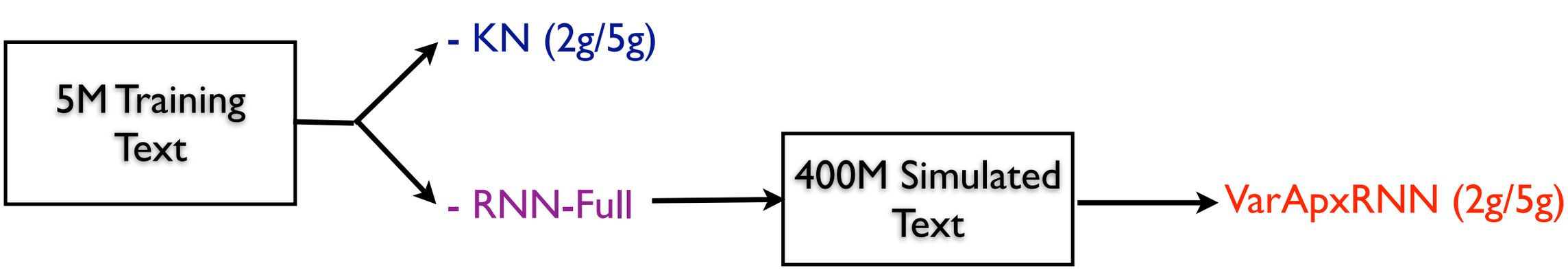
Setup	Set 1	Set 2
KN:MIT+BN (4g) decoding	24.7	22.4
+ RNN-Full rescoring (100 best)	24.1	22.4
+ RNN-Full rescoring (2000 best)	23.8	21.6
Oracle (2000 best)	17.9	15.5
VarApx+KN (4g) decoding	24.3	22.2
+ RNN-Full rescoring (100 best)	23.8	21.7
+ RNN-Full rescoring (2000 best)	23.6	21.5
Oracle (2000 best)	17.5	15.1

Baseline

Proposed

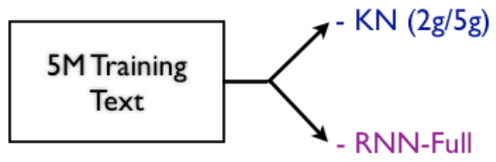
Experiments and Results

2. ASR: CTS and Meeting Recognition



Experiments and Results

2. ASR: CTS and Meeting Recognition

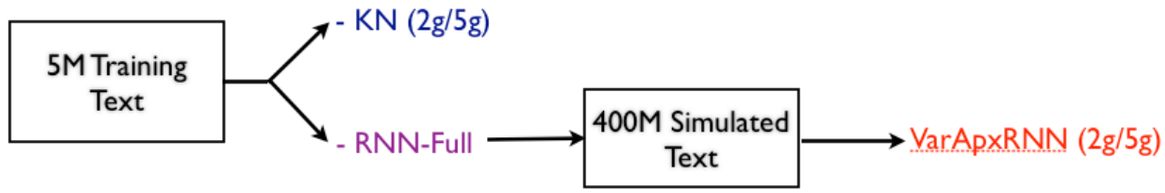


Setup	eval01	rt07s
GT (2g) Decoding	30.3	33.7
+ KN (5g) Lattice Rescoring	**28.0**	**32.4**
+ RNN-Full rescoring (100 best)	27.1	30.8
+ RNN-Full rescoring (1000 best)	26.5	30.5
Oracle (1000 best)	19.5	21.3

Baseline

Experiments and Results

2. ASR: CTS and Meeting Recognition



Setup	eval01	rt07s	
GT (2g) Decoding	30.3	33.7	
+ KN (5g) Lattice Rescoring	**28.0**	**32.4**	**Baseline**
+ RNN-Full rescoring (100 best)	27.1	30.8	
+ RNN-Full rescoring (1000 best)	26.5	30.5	
Oracle (1000 best)	19.5	21.3	
VarApx+GT (2g) Decoding	30.1	33.3	
+ VarApx+KN (5g) Lattice Rescoring	**27.2**	**31.7**	**Proposed**
+ RNN-Full rescoring (100 best)	27.0	30.6	
+ RNN-Full rescoring (1000 best)	**26.5**	**30.4**	
Oracle (1000 best)	19.5	21.0	

Outline

Introduction and Motivation ✓

Variational Approximations

- RNN: Long-Span Models ✓
- Variational Approximation Framework ✓
- Experiments and Discussions ✓
- Conclusion

Conclusion and Future Work

Conclusion and Future Work

1. *n*-gram approximation of long-span LMs yield greater accuracy and allow their easy integration into decoders.

Conclusion and Future Work

1. *n*-gram approximation of long-span LMs yield greater accuracy and allow their easy integration into decoders.

2. RNN LM improves significantly over n-grams with increasing data (forthcoming work), calling for an investigation of more powerful tractable approximations.

Thank you

Questions ?

Bigger Training data experiment

Setup	WER(rt04)	PPL(rt04)
KN:BN Decoding (4g)	14.10	172.3
VarApx+KN Decoding (4g)	13.5	159.6
RNN-Full	12.1	111

Original Training data is **400M** word tokens (Broadcast News).

Sampled about **1B** word tokens.

Pruned the variational model so as to be comparable and usable in decoders.