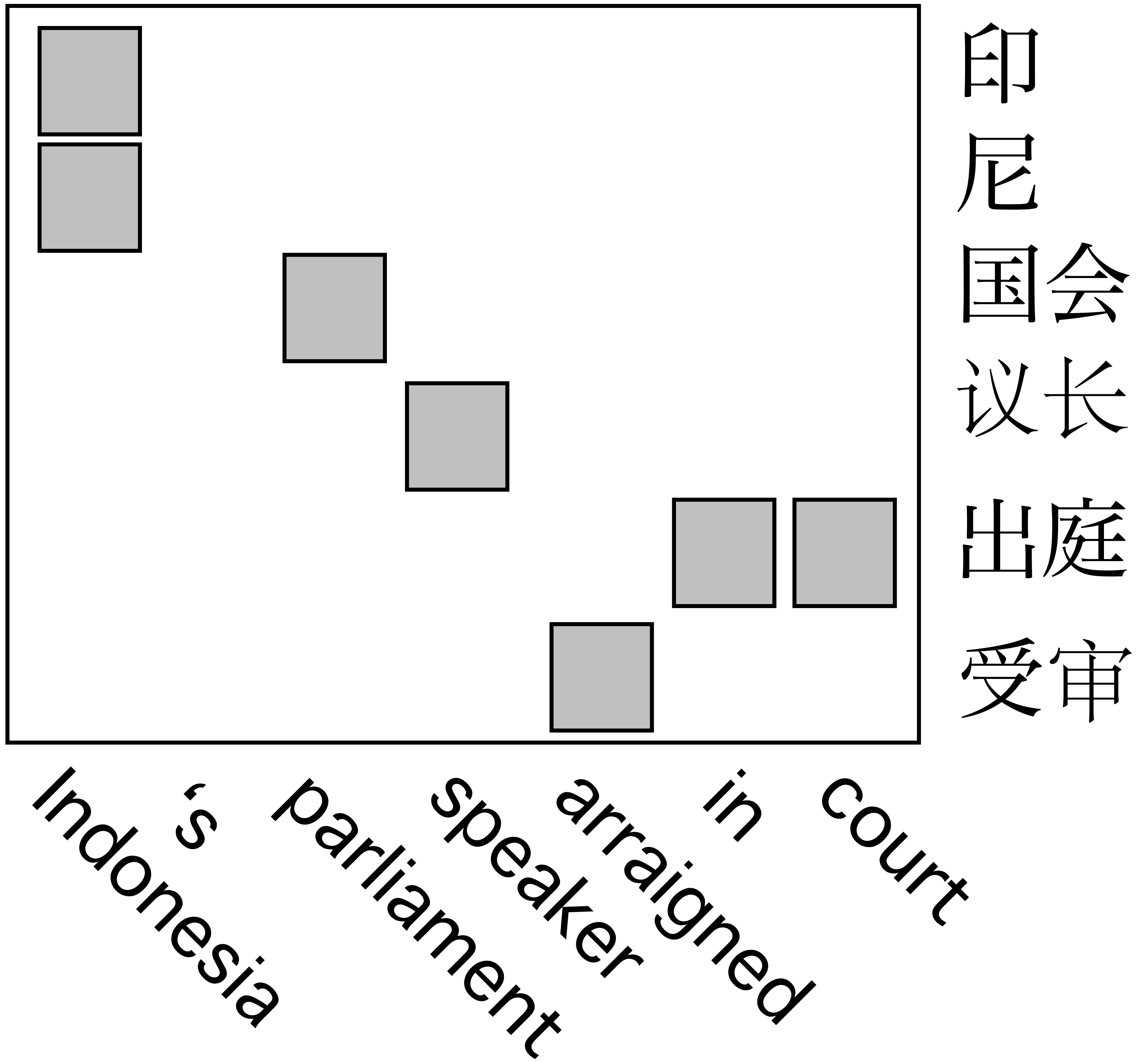
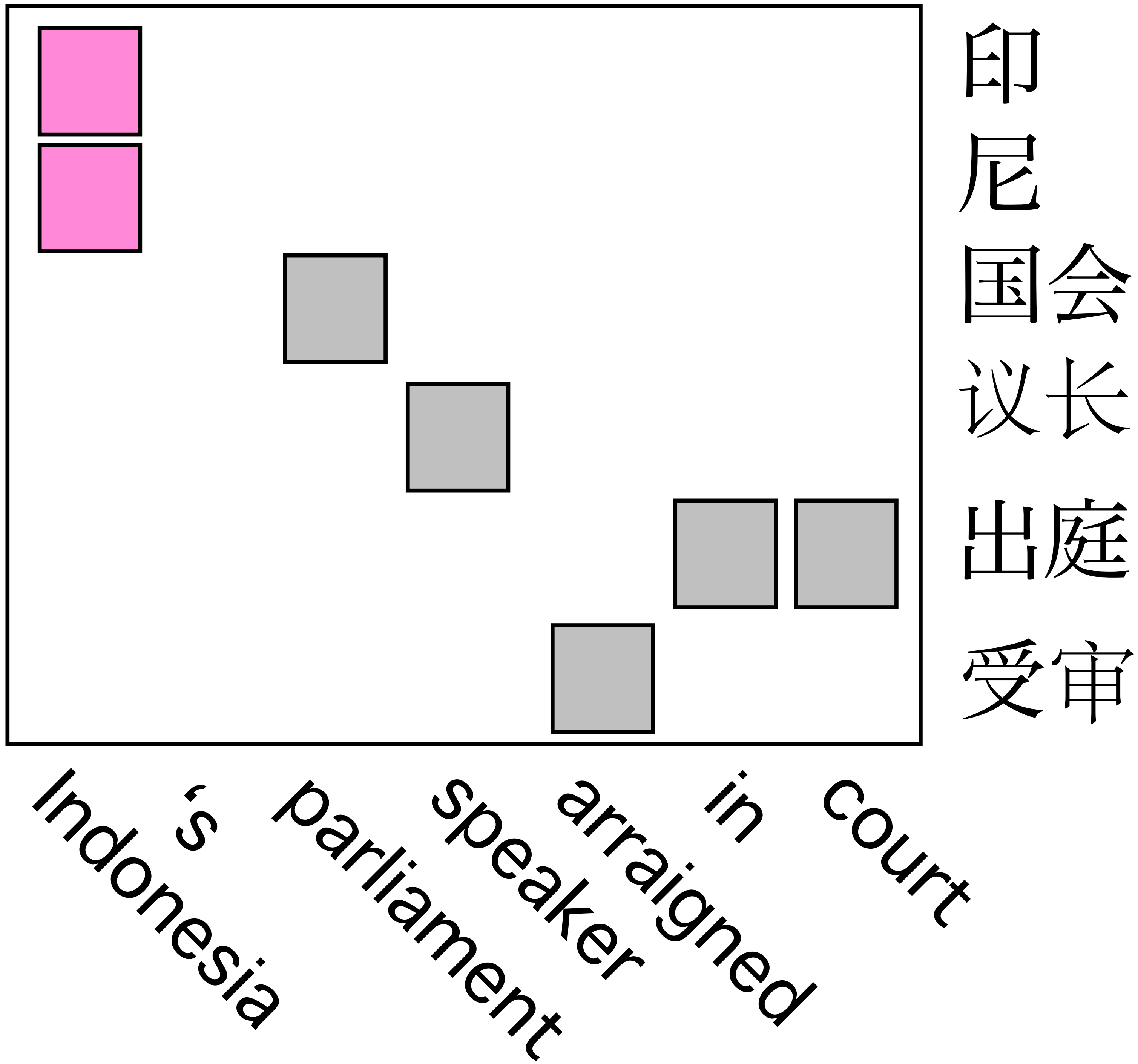
Better Word Alignment with Supervised ITG Models

Aria Haghighi, John Blitzer,
John DeNero, and Dan Klein
UC Berkeley

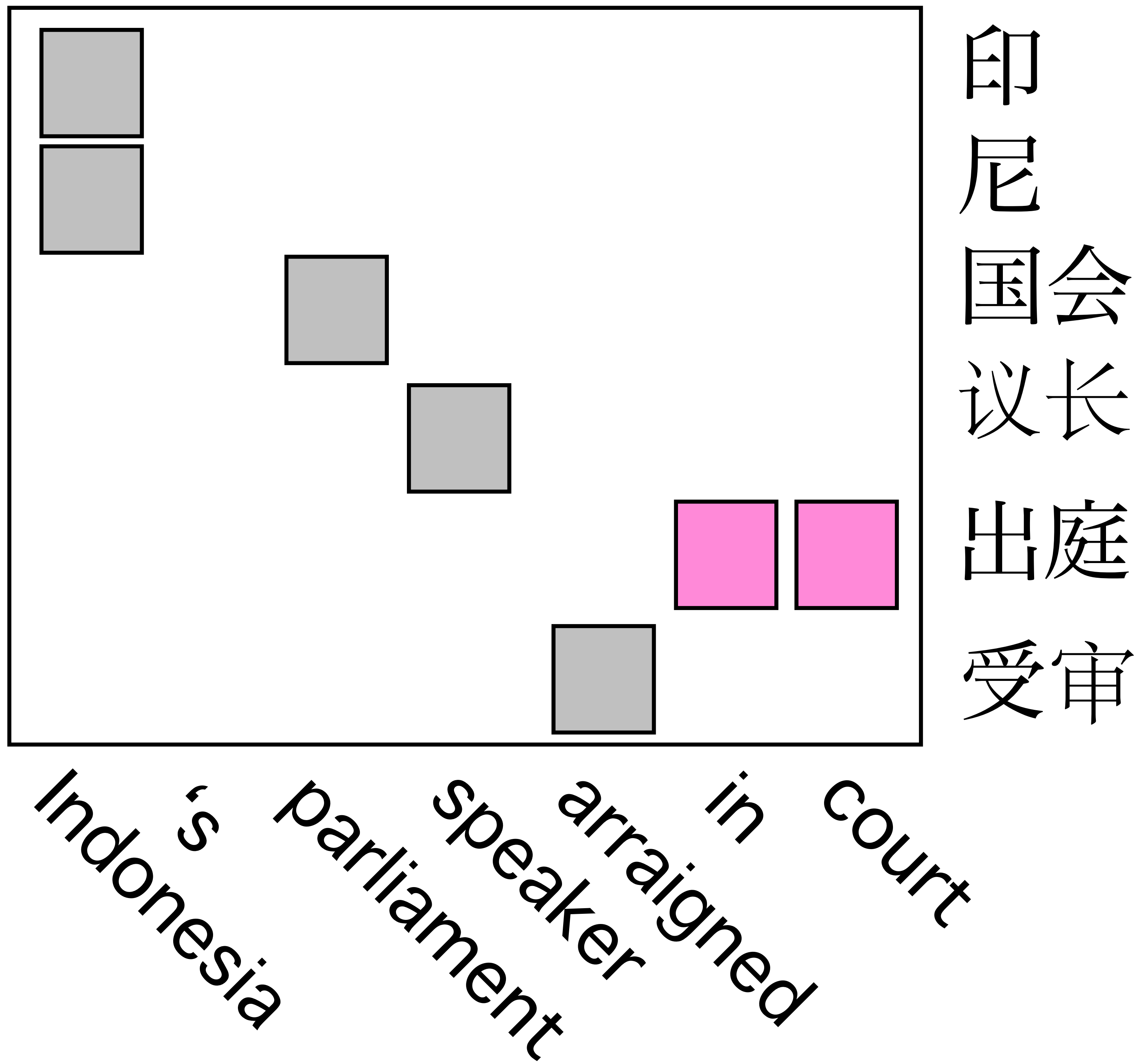
Word Alignment



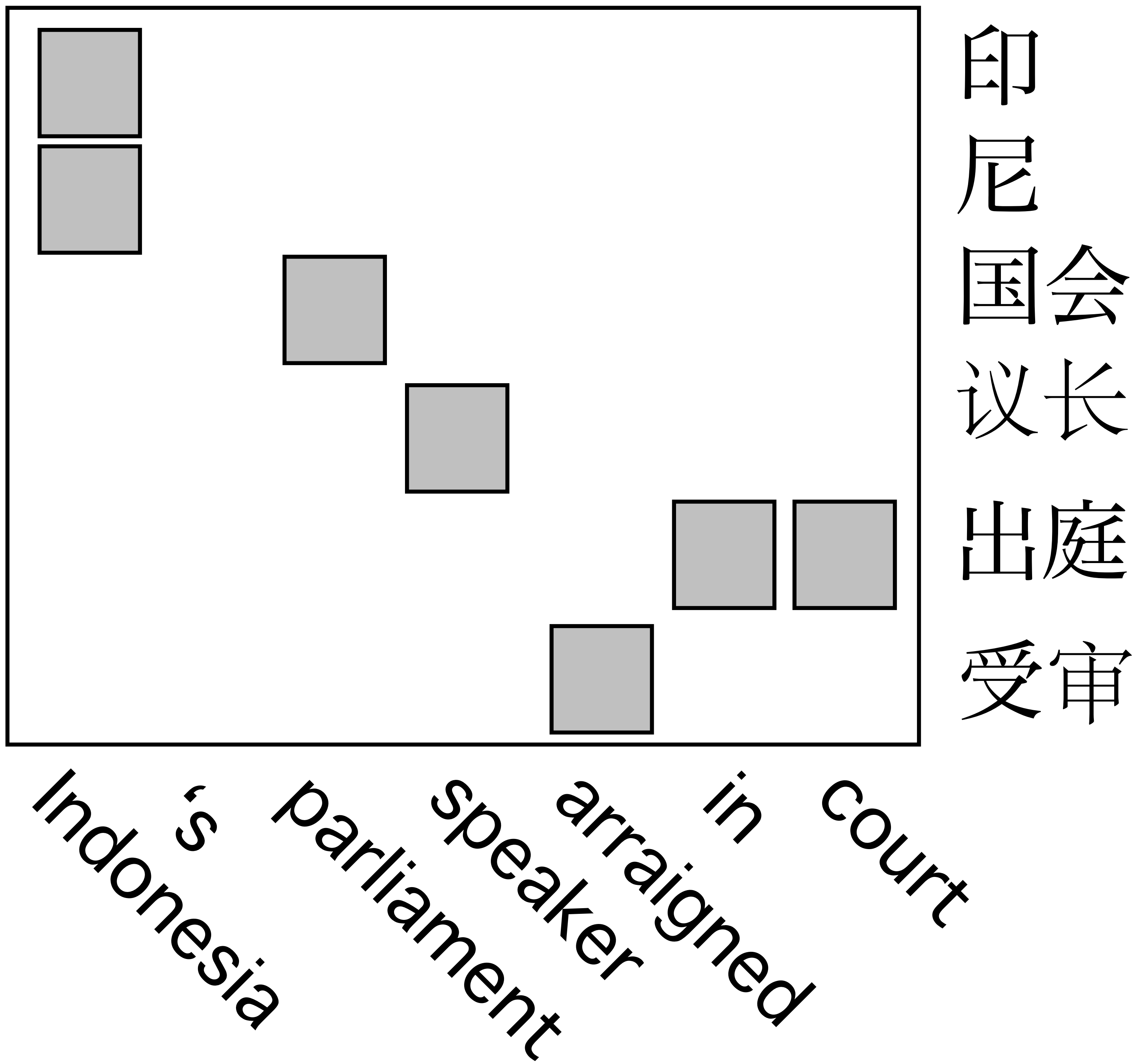
Word Alignment



Word Alignment



Word Alignment



Experimental Setup

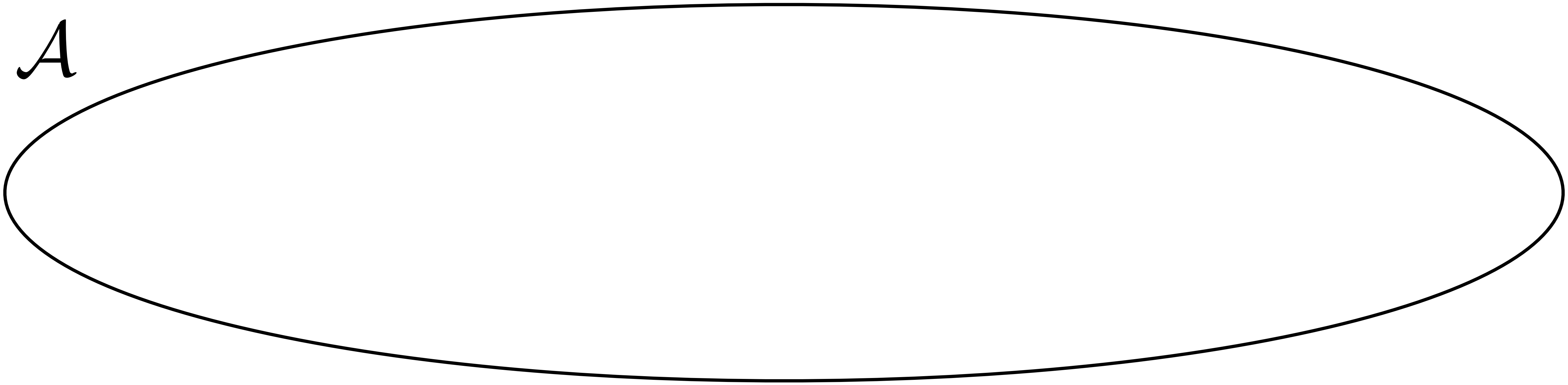
Experimental Setup

- 2002 Chinese-English NIST data
  - 150 labeled training examples
  - 191 test examples

Experimental Setup

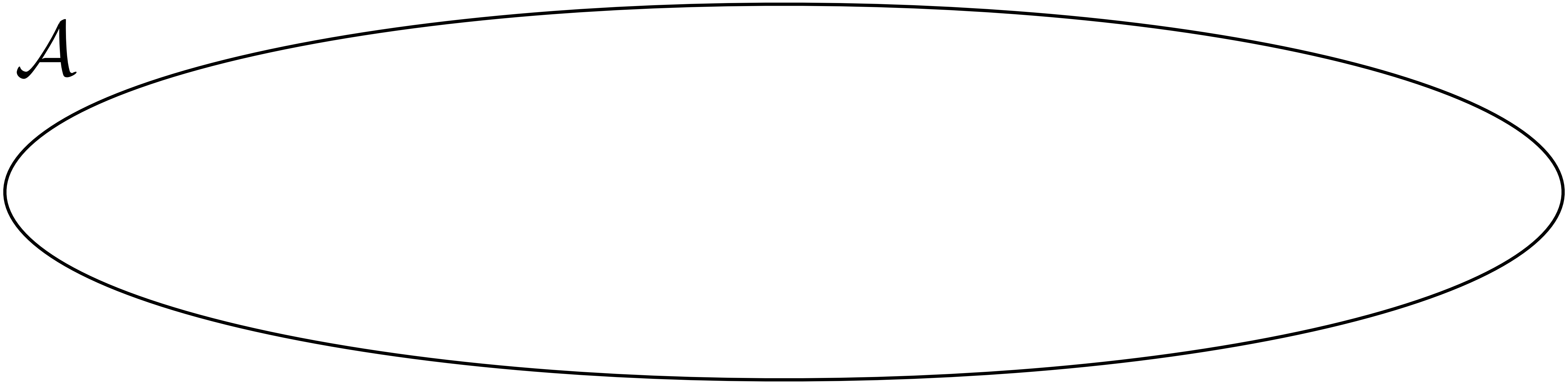
- 2002 Chinese-English NIST data
  - 150 labeled training examples
  - 191 test examples
- Evaluation
  - Prec, Recall over gold alignments
  - BLEU end-to-end

Alignment Families



Alignment Families

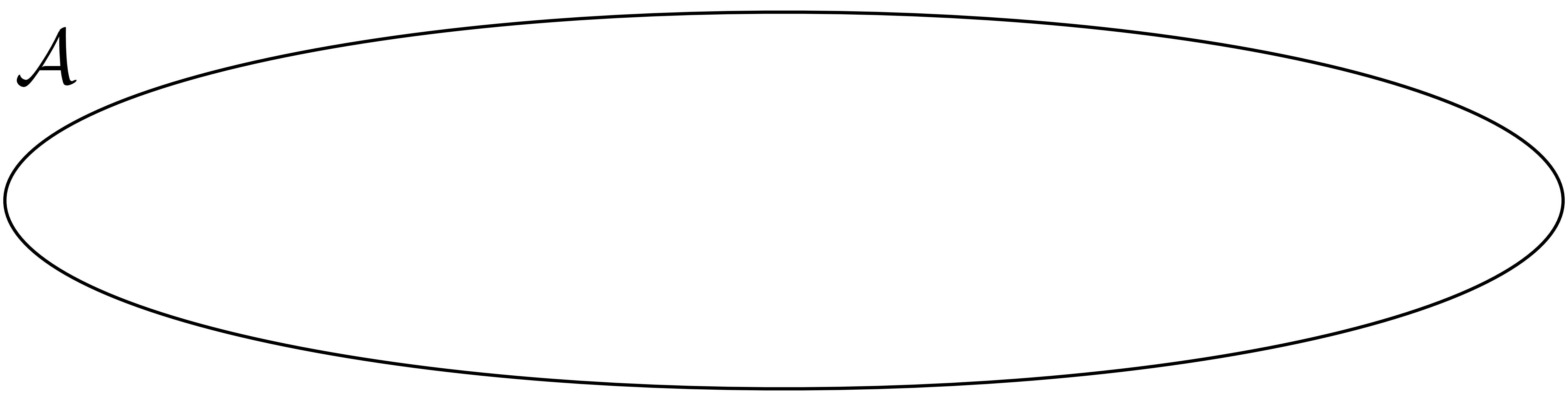
<u>Optimal AER</u>



Alignment Families

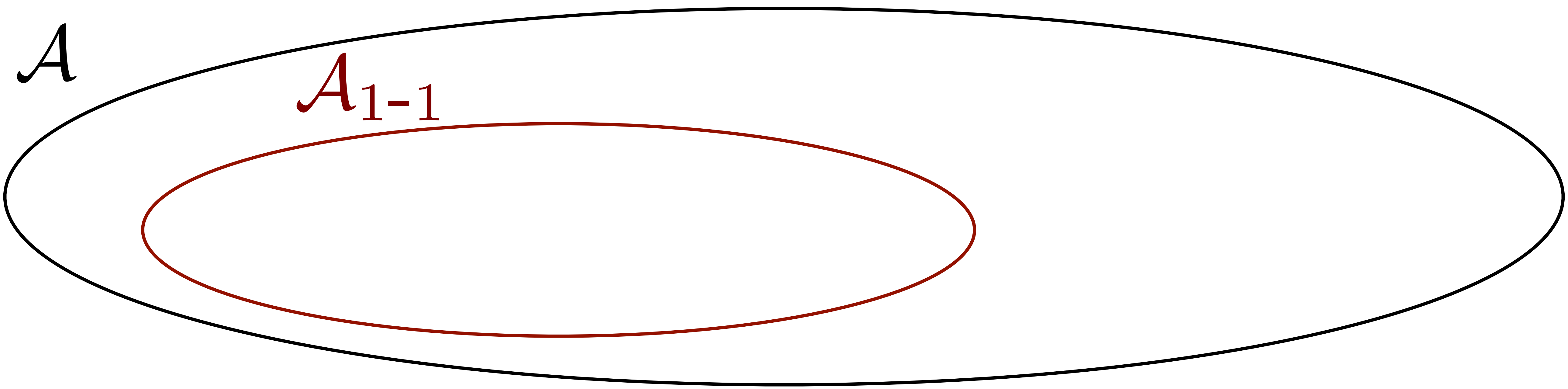
	Optimal AER
$\mathcal{A}$	0.0

$\mathcal{A}$



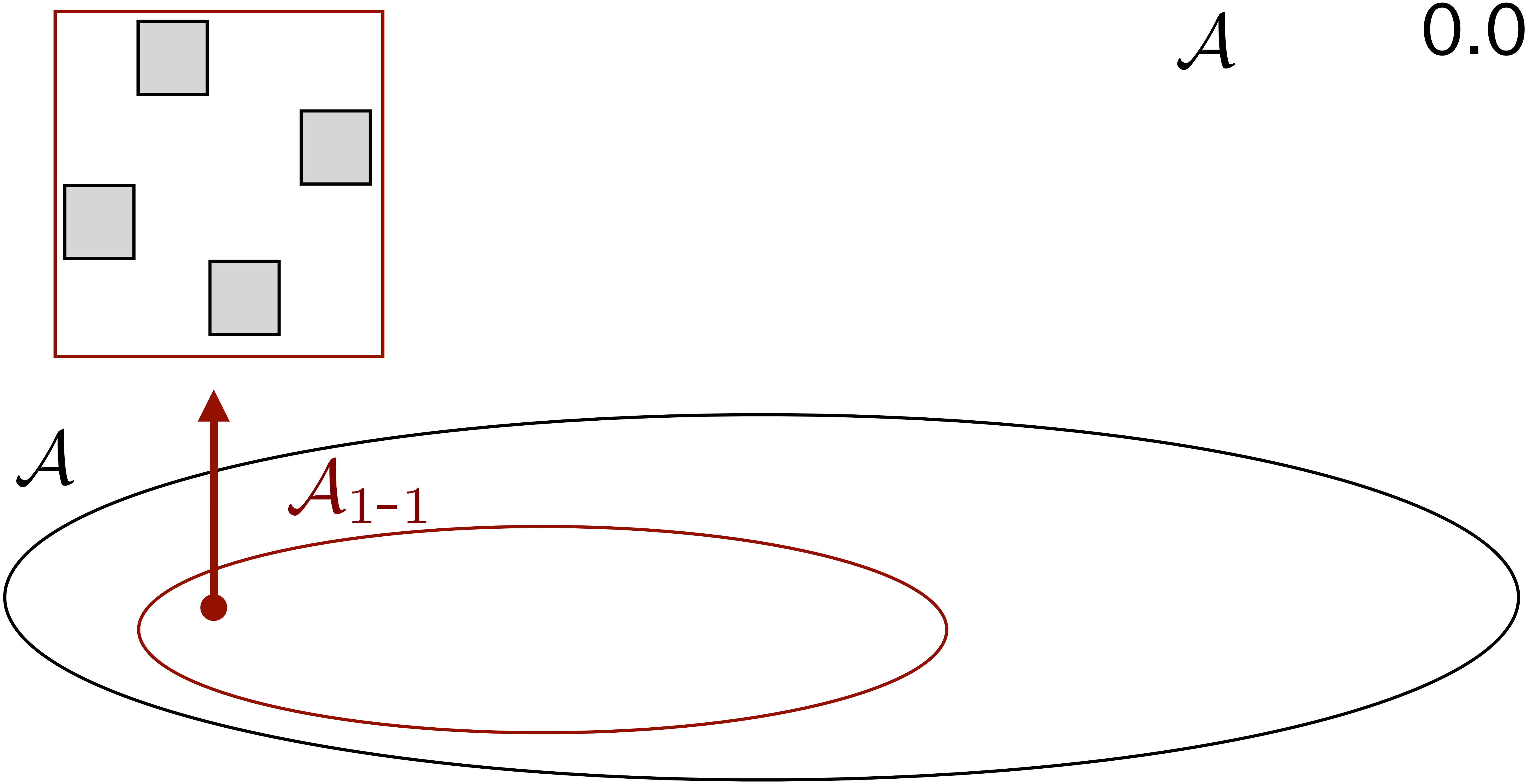
Alignment Families

Optimal AER	
$\mathcal{A}$	0.0



Alignment Families

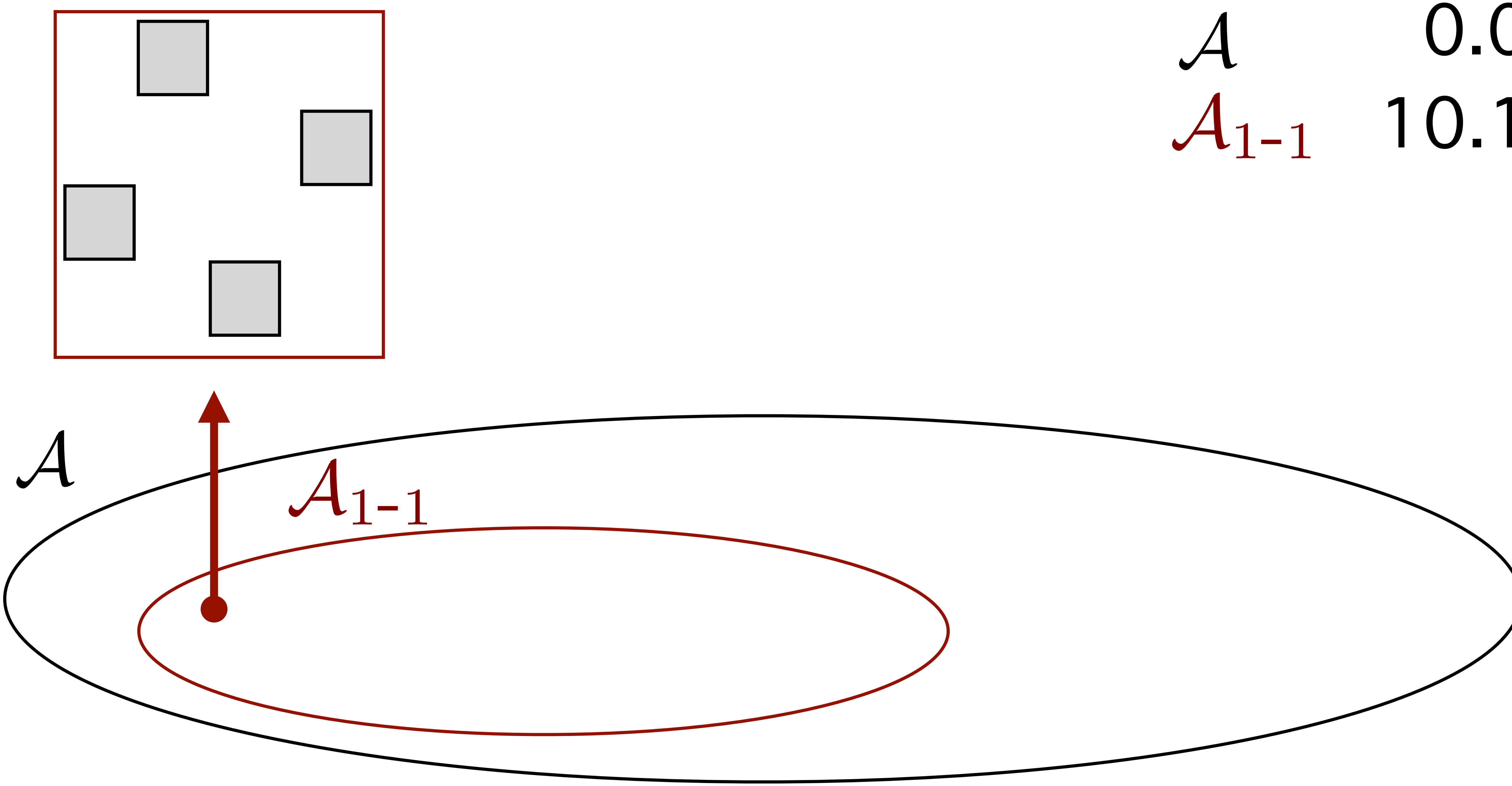
Optimal AER	
$\mathcal{A}$	0.0



Alignment Families

<u>Optimal AER</u>

$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1



Alignment Families

Alignment Families

Inversion Transduction Grammar (ITG)

[Wu, '97]

Alignment Families

Inversion Transduction Grammar (ITG)

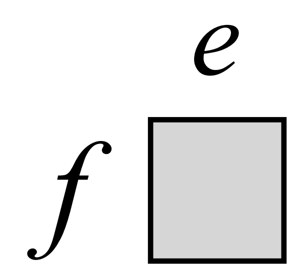
$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$

[Wu, '97]

Alignment Families

Inversion Transduction Grammar (ITG)

$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$

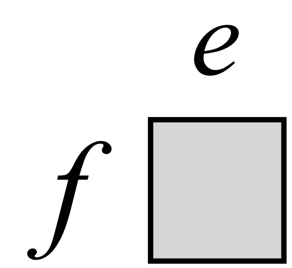


[Wu, '97]

Alignment Families

Inversion Transduction Grammar (ITG)

$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$



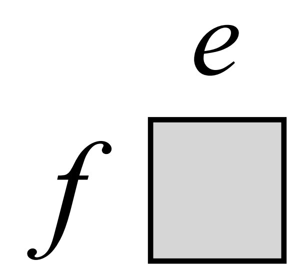
$$X \rightarrow X^{(L)} X^{(R)}$$

[Wu, ’97]

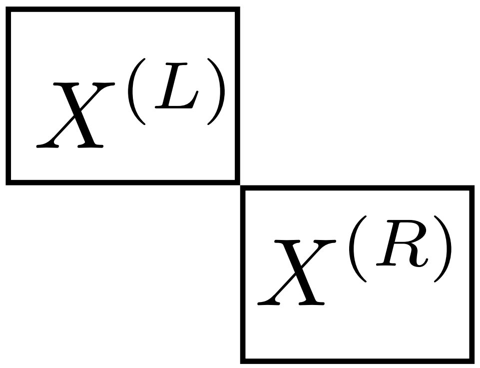
Alignment Families

Inversion Transduction Grammar (ITG)

$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$



$$X \rightarrow X^{(L)} X^{(R)}$$

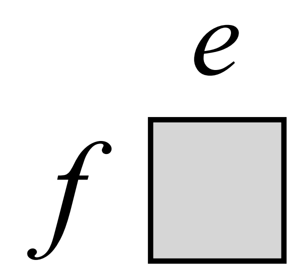


[Wu, ’97]

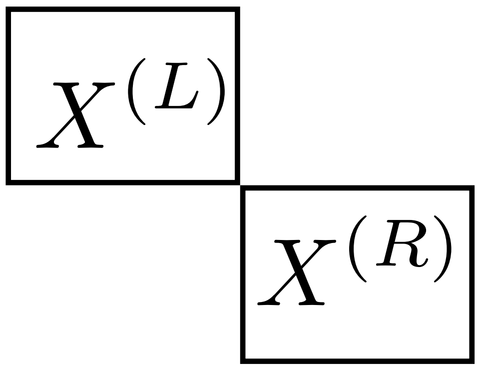
Alignment Families

Inversion Transduction Grammar (ITG)

$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$



$$X \rightarrow X^{(L)} X^{(R)}$$



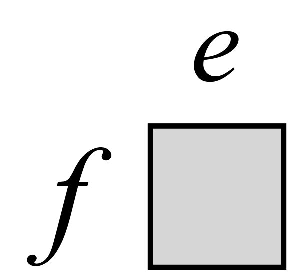
$$X \rightsquigarrow X^{(L)} X^{(R)}$$

[Wu, ’97]

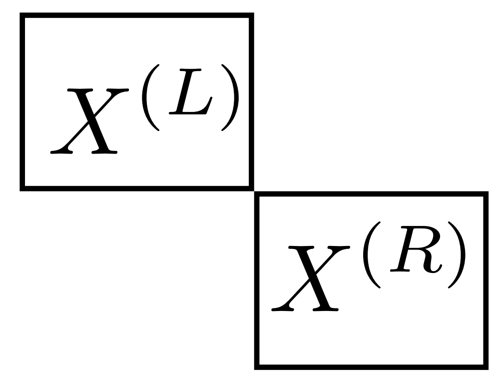
Alignment Families

Inversion Transduction Grammar (ITG)

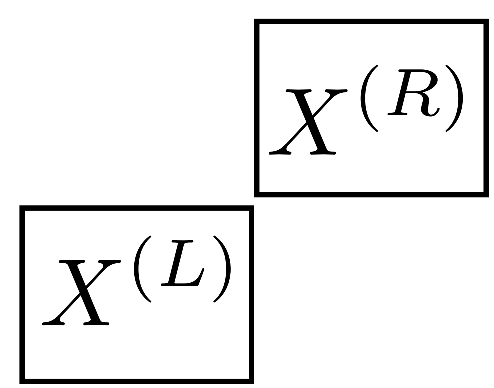
$$X \rightarrow \langle f, e \rangle, \langle f, \epsilon \rangle, \langle \epsilon, e \rangle$$



$$X \rightarrow X^{(L)} X^{(R)}$$

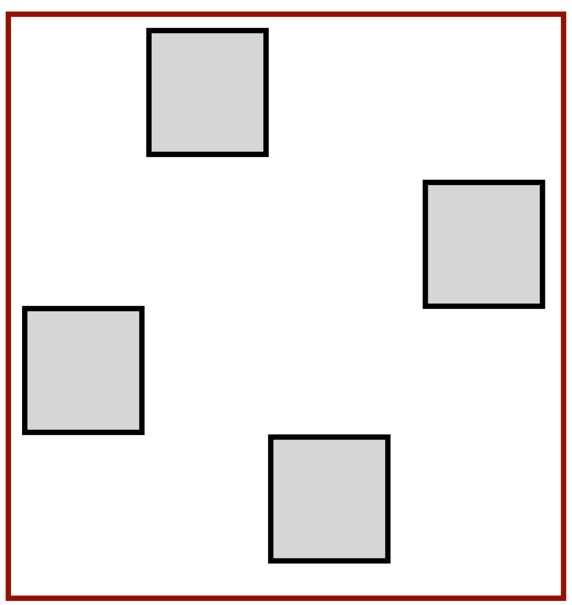


$$X \rightsquigarrow X^{(L)} X^{(R)}$$

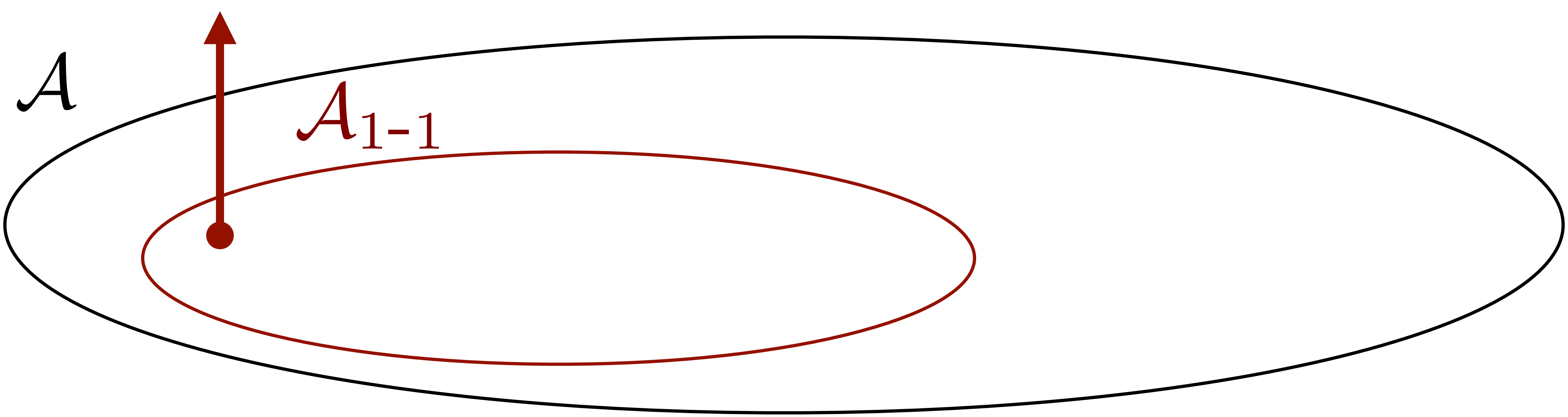


[Wu, ’97]

Alignment Families

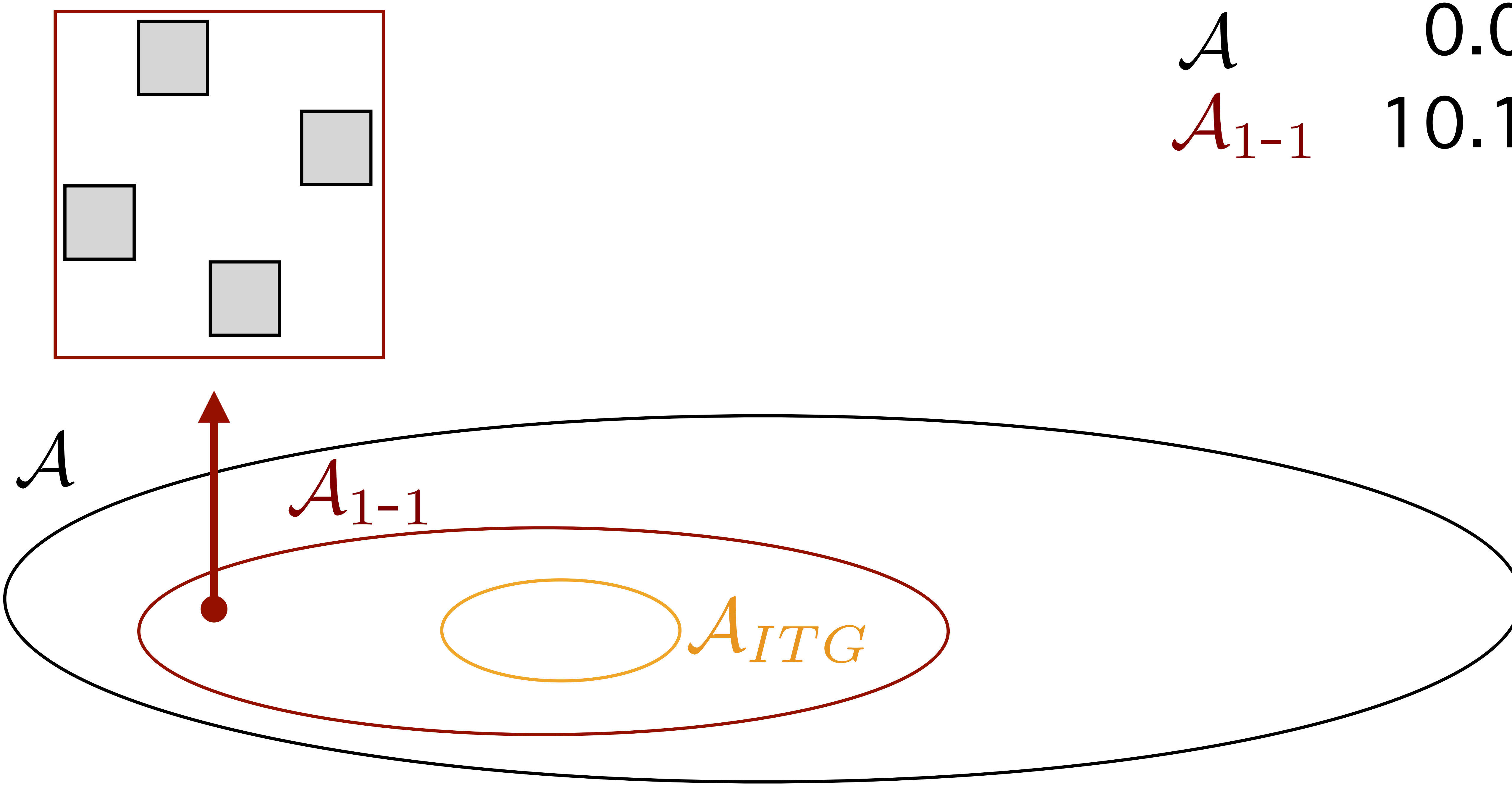


Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1

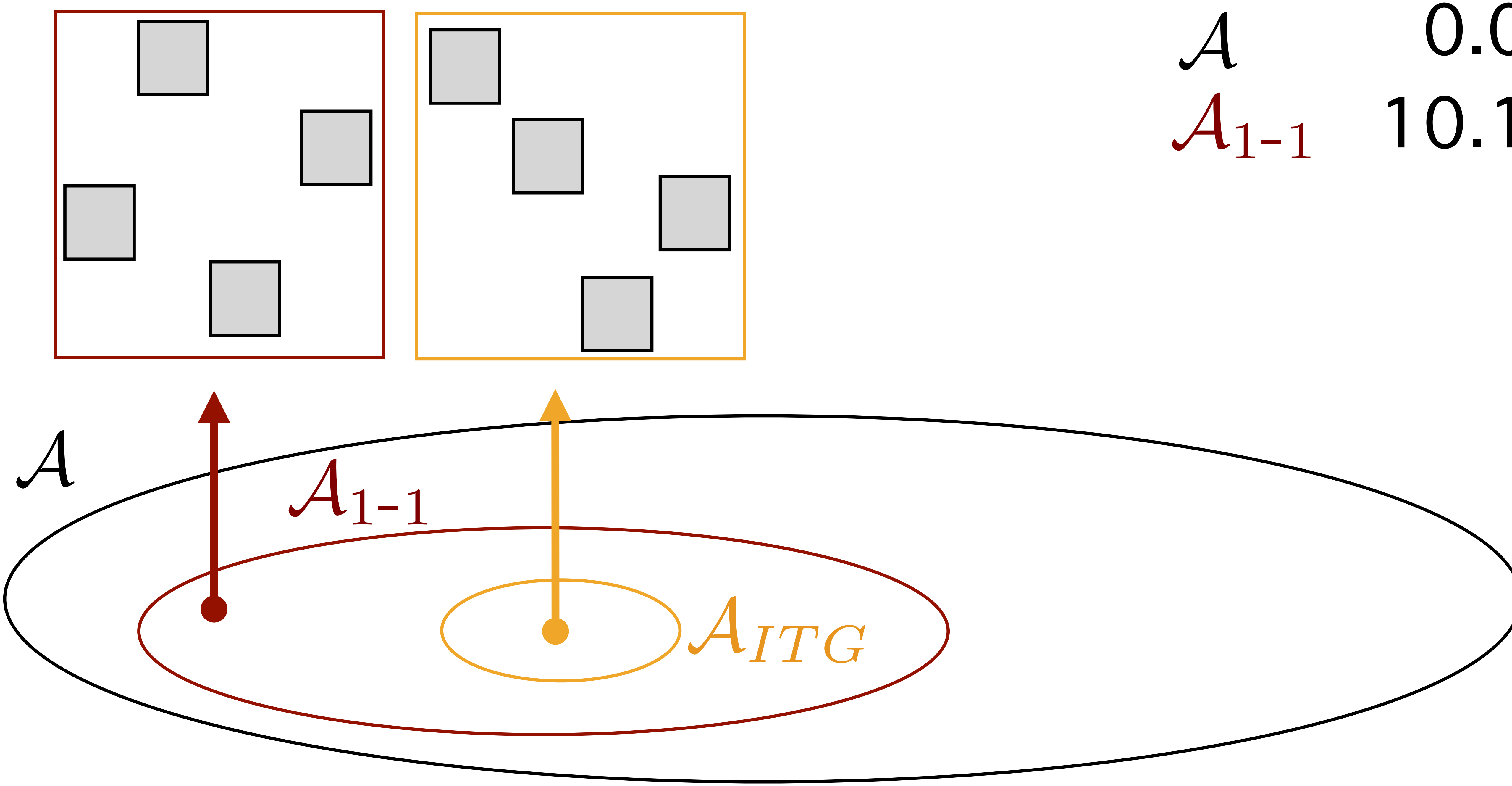


Alignment Families

Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1

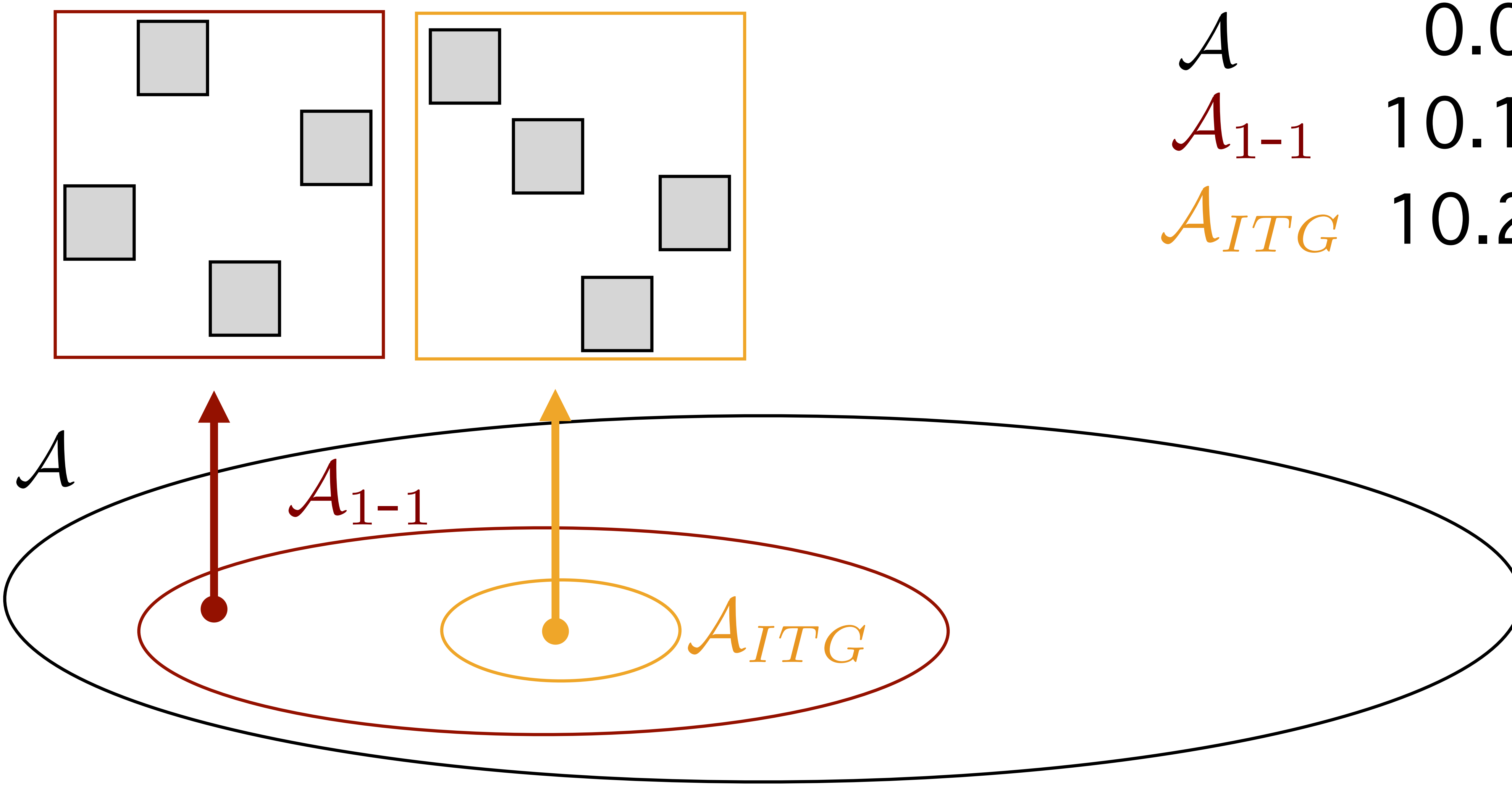


Alignment Families



Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1

Alignment Families



Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2

Alignment Families

Alignment Families

Block ITG (BITG)

Alignment Families

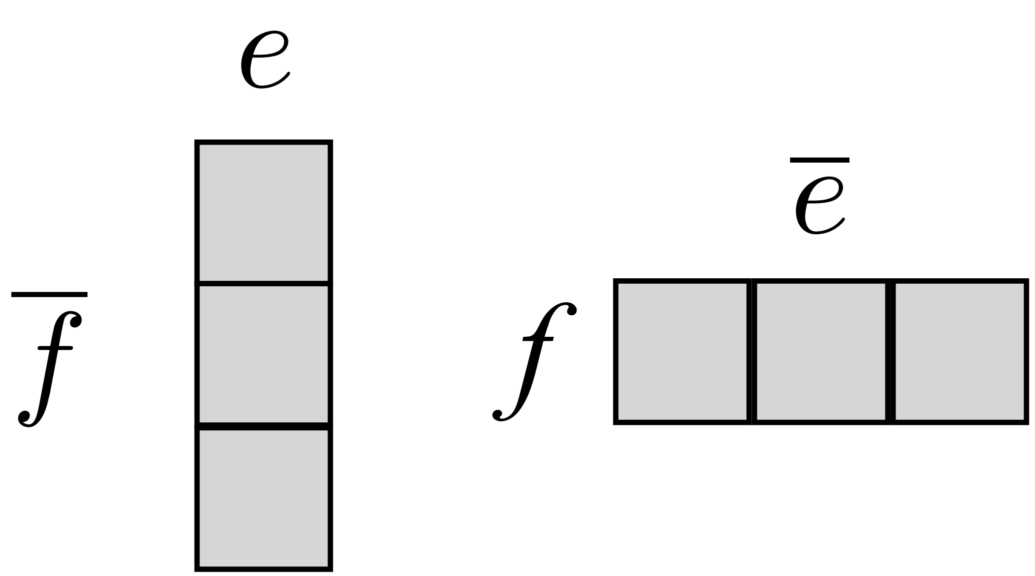
Block ITG (BITG)

$$X \rightarrow \langle \overline{f}, e \rangle, \langle f, \overline{e} \rangle$$

Alignment Families

Block ITG (BITG)

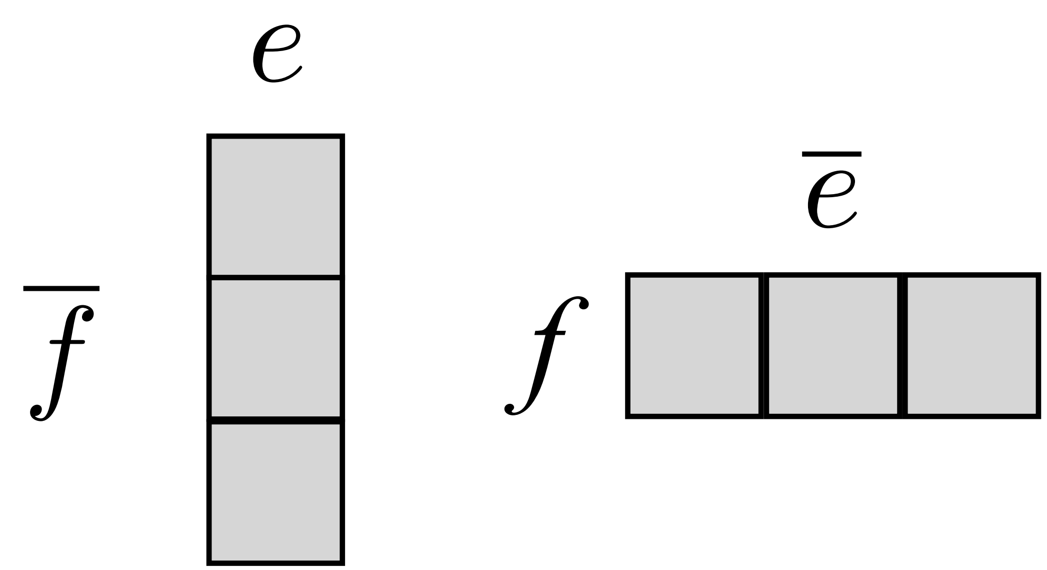
$$X \rightarrow \langle \overline{f}, e \rangle, \langle f, \overline{e} \rangle$$



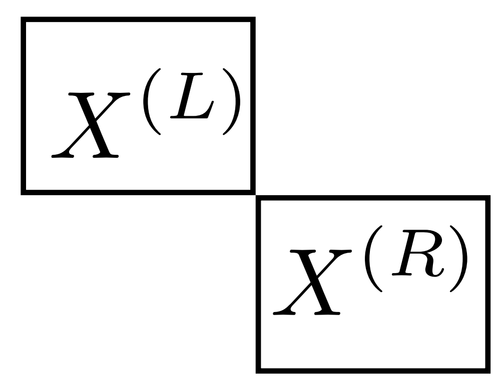
Alignment Families

Block ITG (BITG)

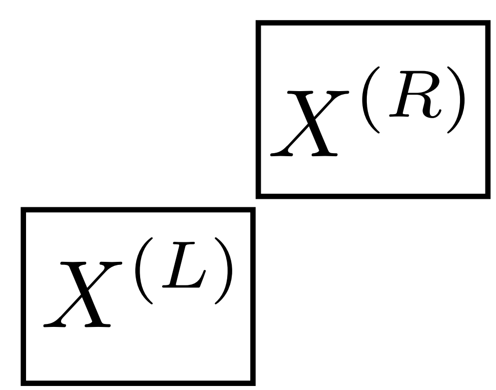
$$X \rightarrow \langle \overline{f}, e \rangle, \langle f, \overline{e} \rangle$$



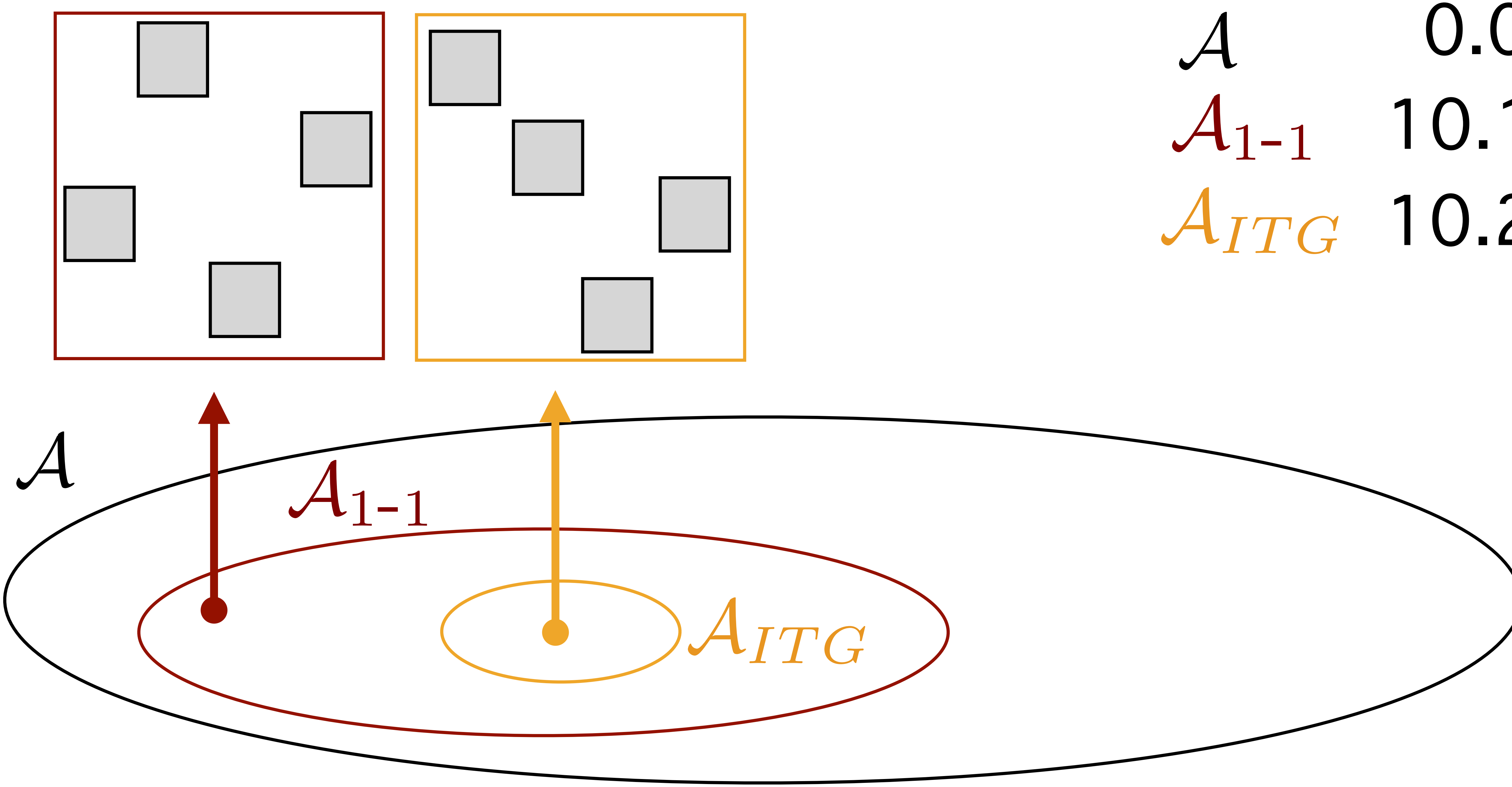
$$X \rightarrow X^{(L)} X^{(R)}$$



$$X \rightsquigarrow X^{(L)} X^{(R)}$$

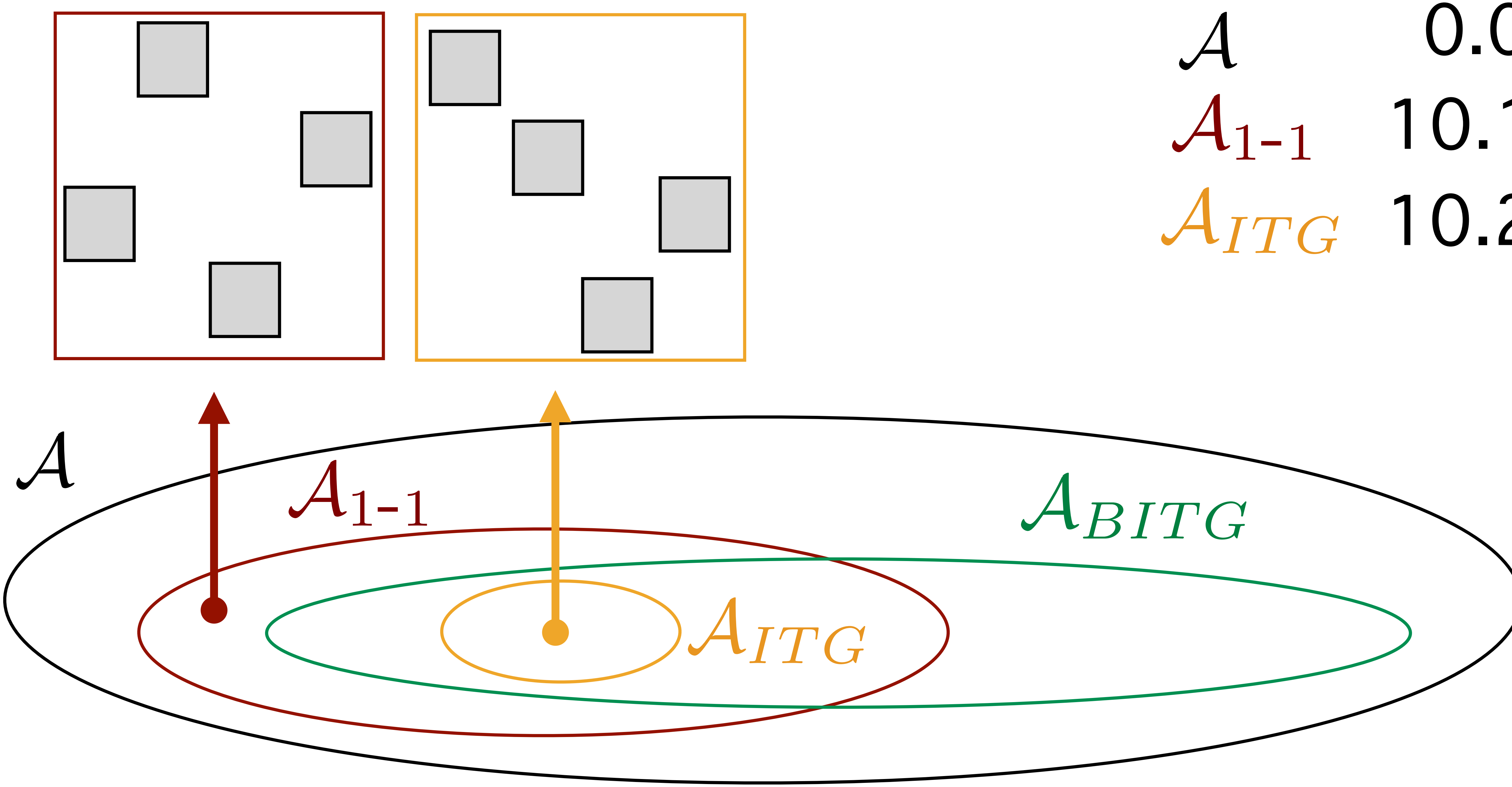


Alignment Families



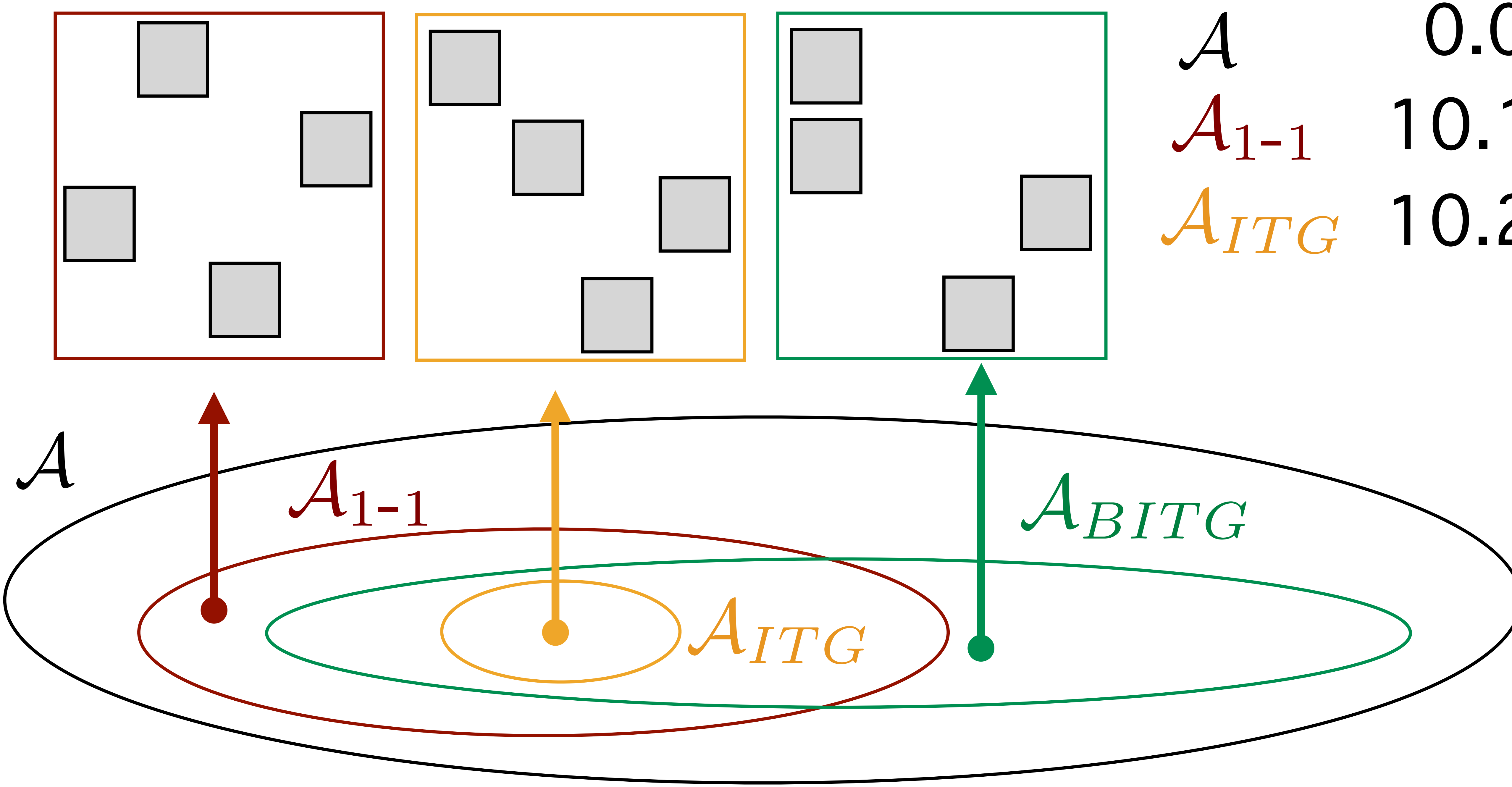
	Optimal AER
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2

Alignment Families



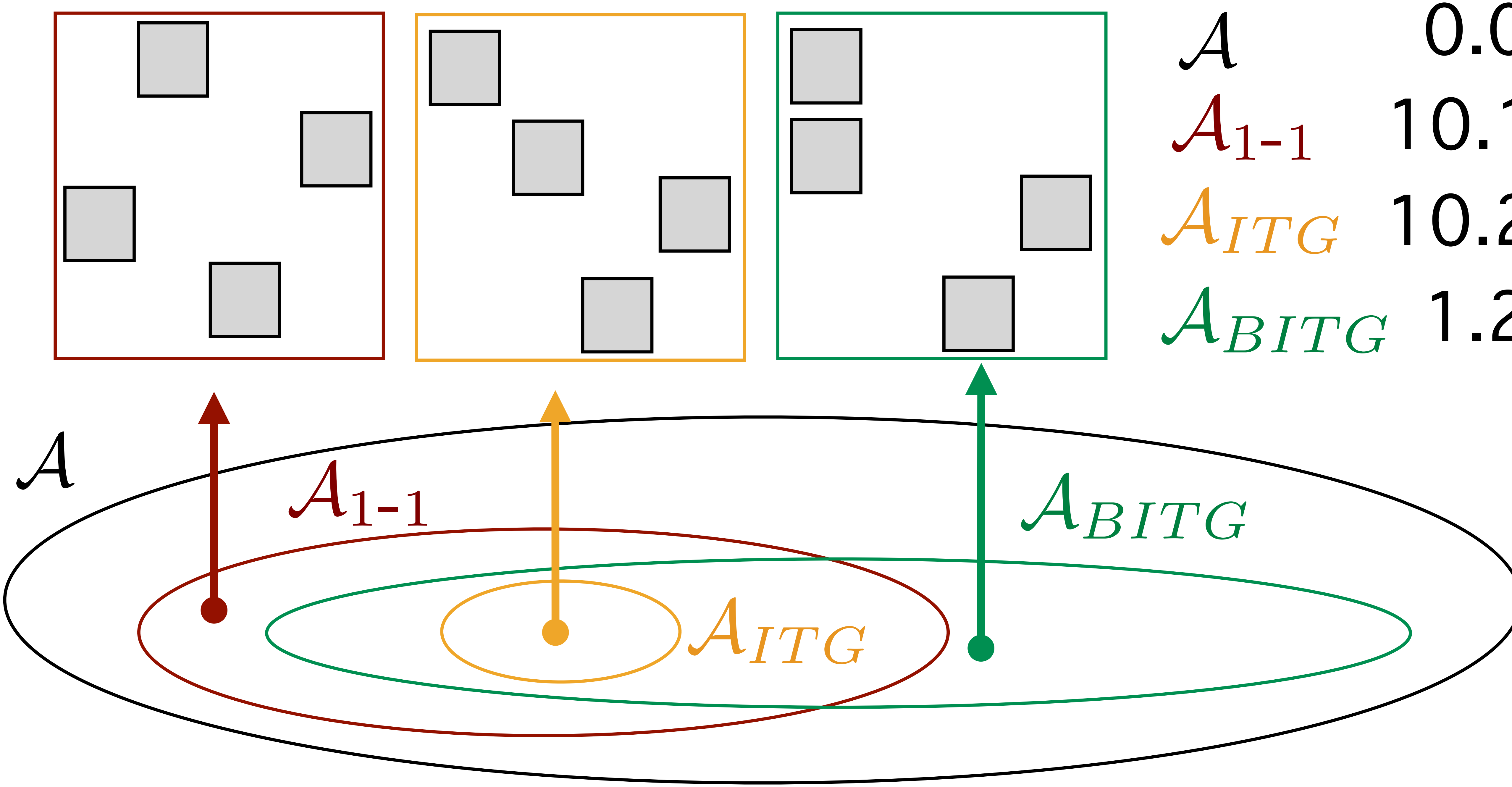
Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2

Alignment Families



Optimal AER	
$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2

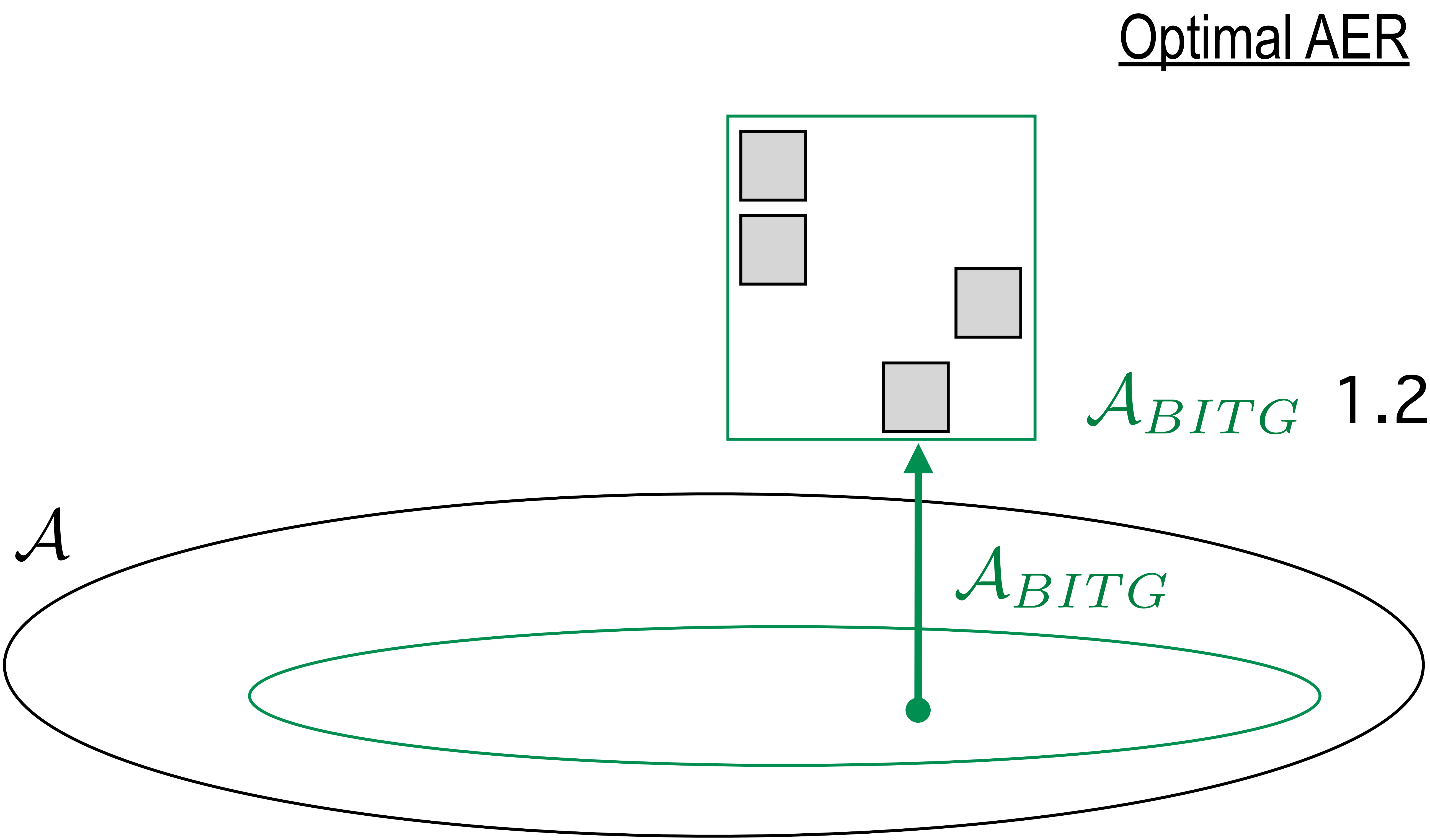
Alignment Families



Optimal AER

$\mathcal{A}$	0.0
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2
$\mathcal{A}_{BITG}$	1.2

Alignment Families



Supervised Data

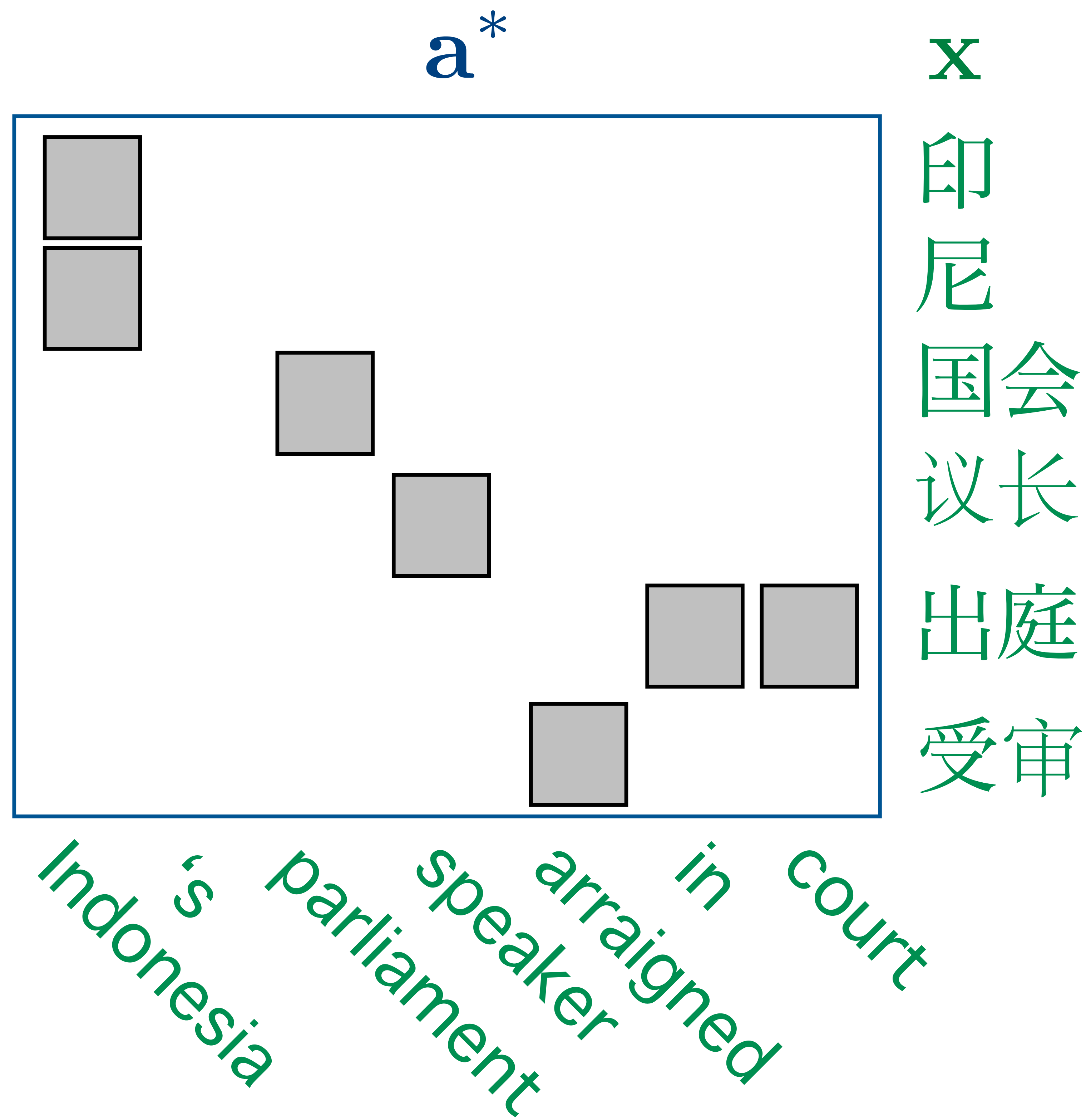
Supervised Data

x

印
尼
国会
议长
出庭
受审

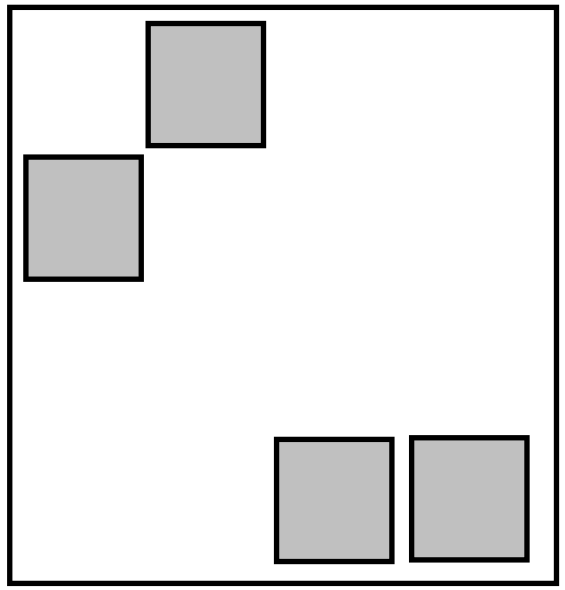
Indonesia ‘s parliament speaker arraigned in court

Supervised Data



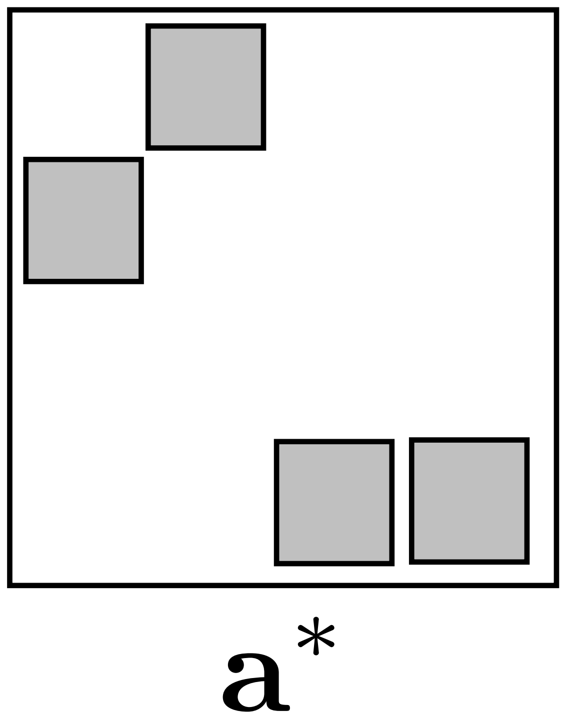
Loss Function

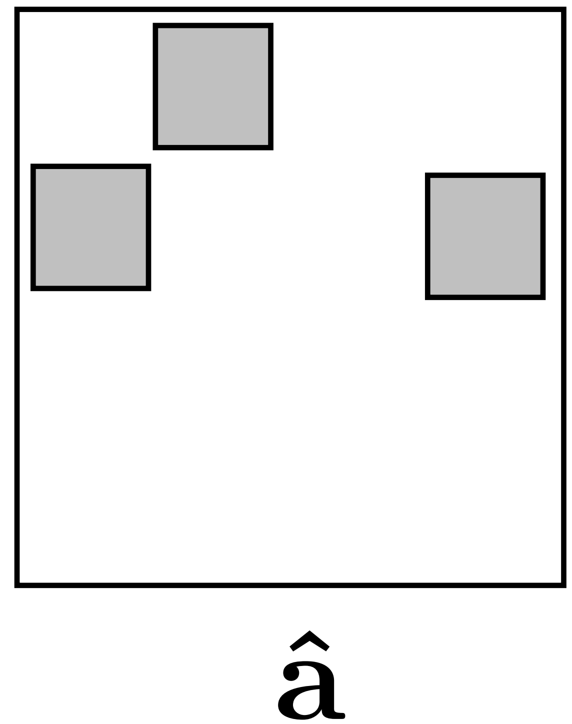
Loss Function



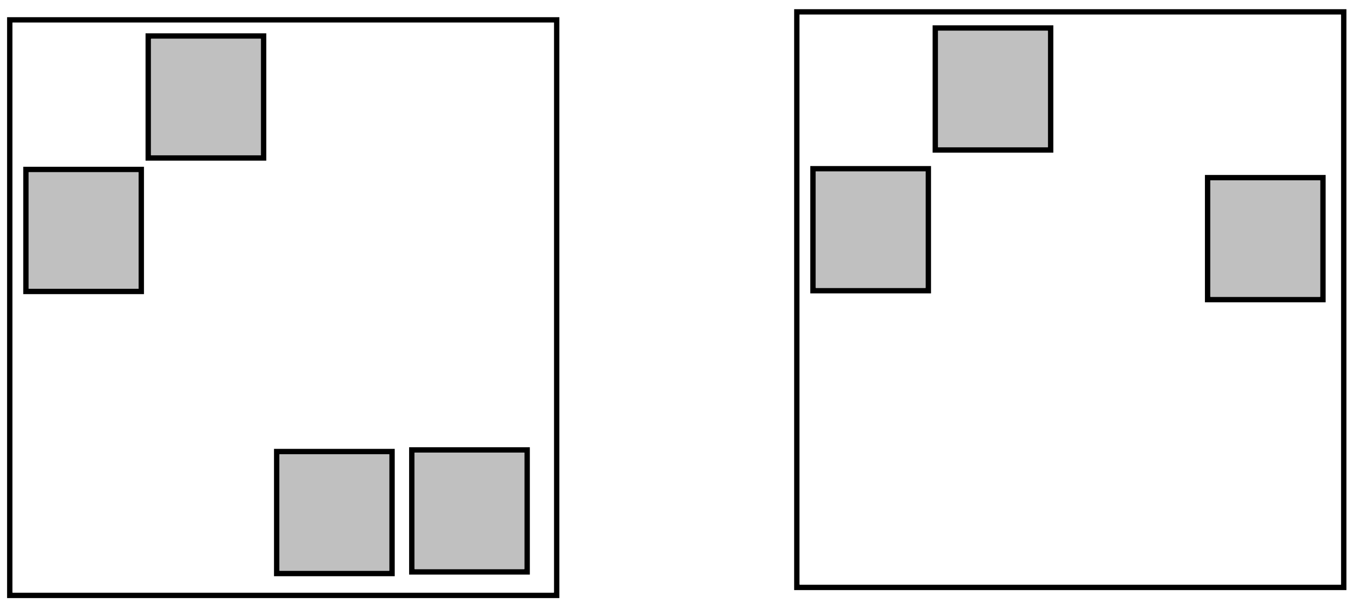
$\mathbf{a}^*$

Loss Function





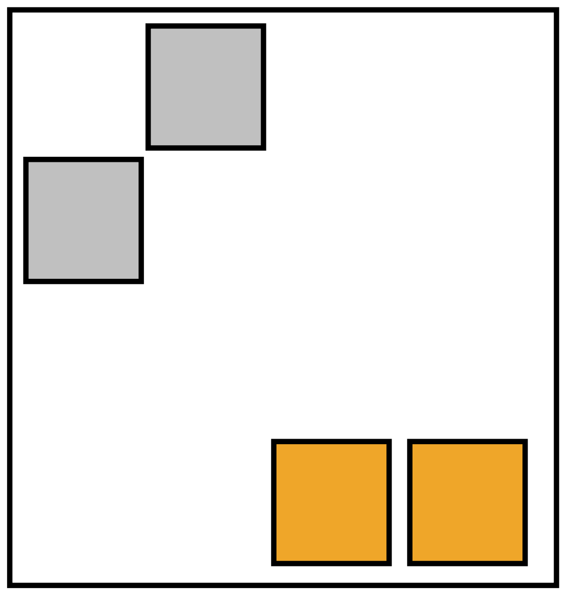
Loss Function



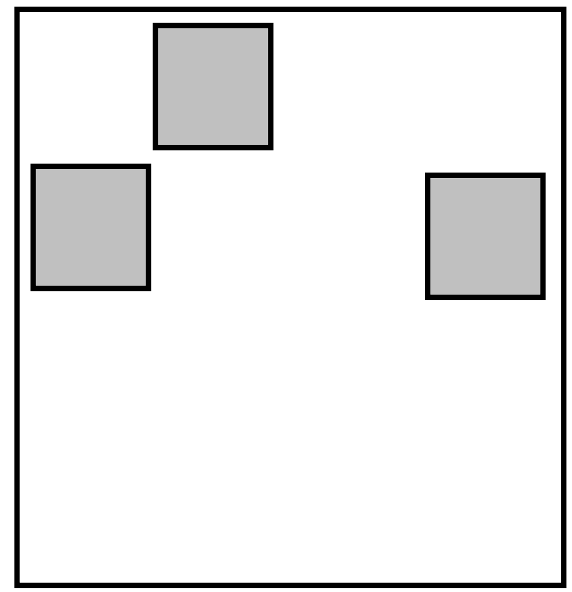
$\mathbf{a}^*$ $\hat{\mathbf{a}}$

$$\begin{aligned} L(\mathbf{a}^*, \mathbf{a}) \quad &= \quad \#\text{ of missing sure alignments} + \\ &\quad\; \#\text{ of incorrect alignments} \end{aligned}$$

Loss Function



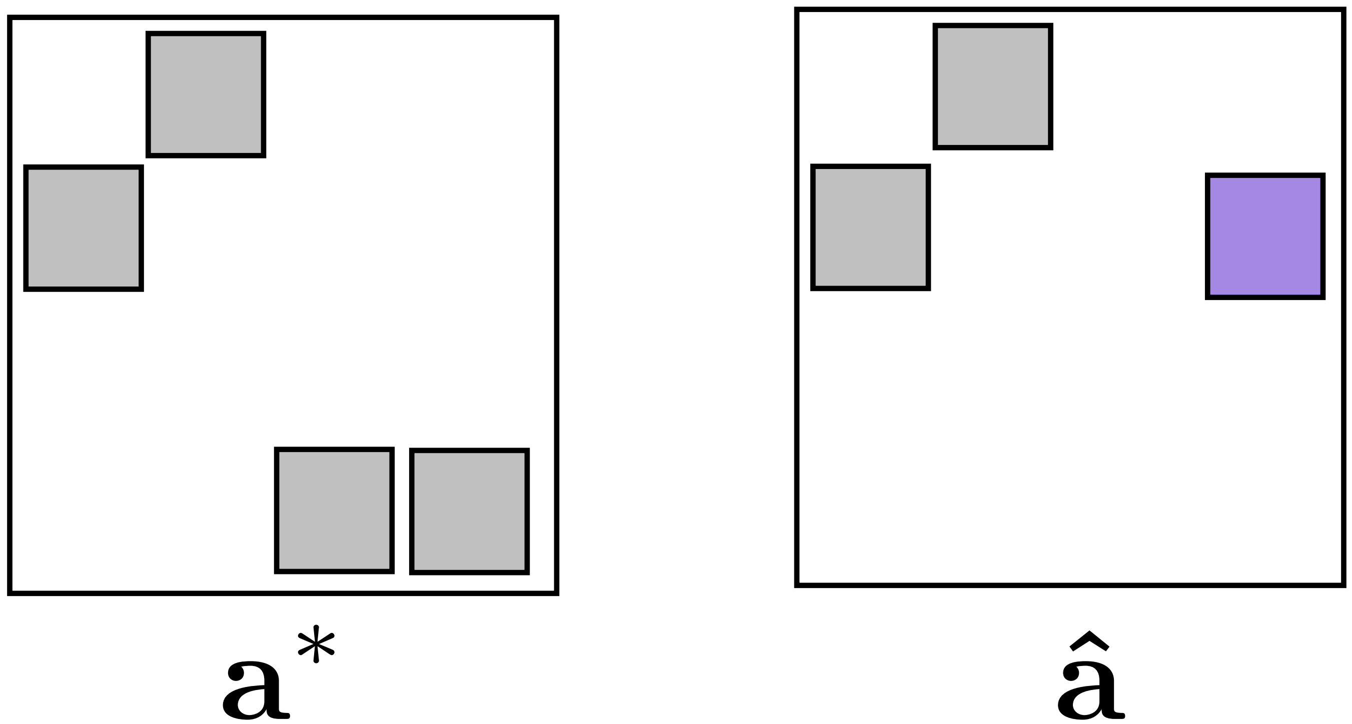
$\mathbf{a}^*$



$\hat{\mathbf{a}}$

$$L(\mathbf{a}^*, \mathbf{a}) \quad = \quad \#\ \text{of missing sure alignments} + \\ \#\ \text{of incorrect alignments}$$

Loss Function

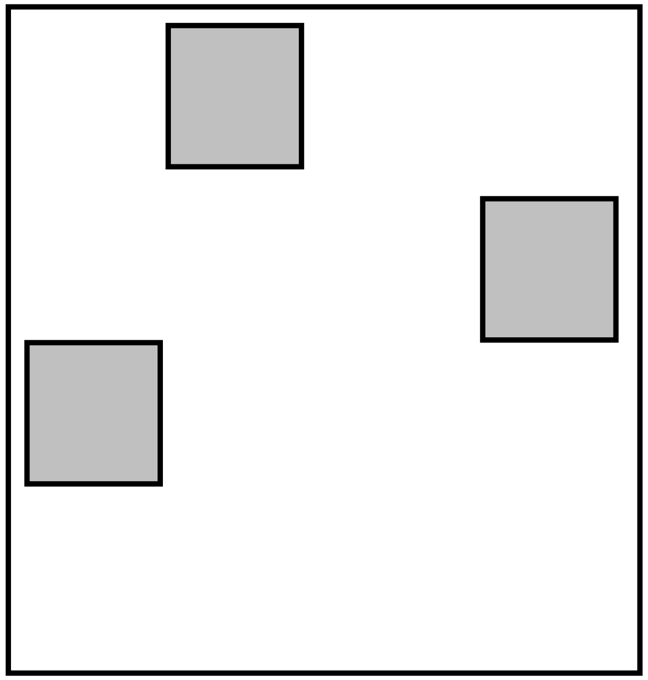


$$L(\mathbf{a}^*, \mathbf{a}) \quad = \quad \#\text{ of missing sure alignments} + \#\text{ of incorrect alignments}$$

Linear Model

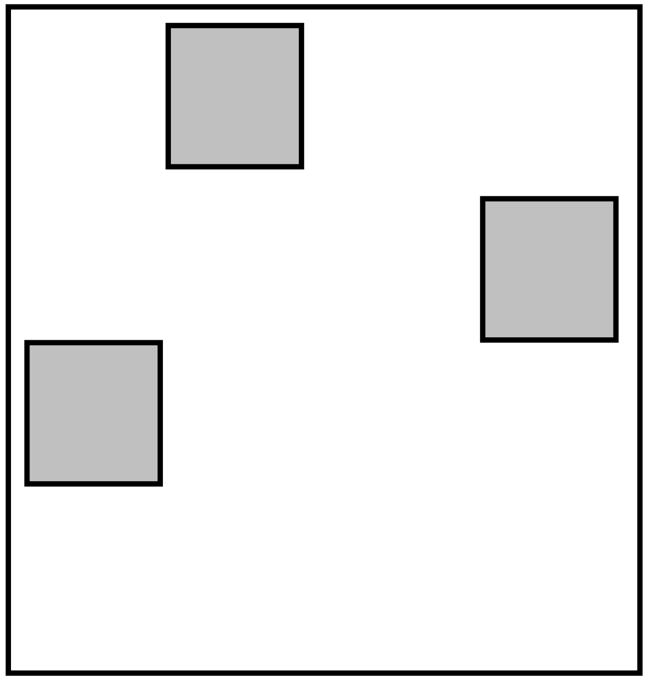
Linear Model

$$\mathbf{a} \in \mathcal{A}'$$



Linear Model

$\mathbf{a} \in \mathcal{A}'$

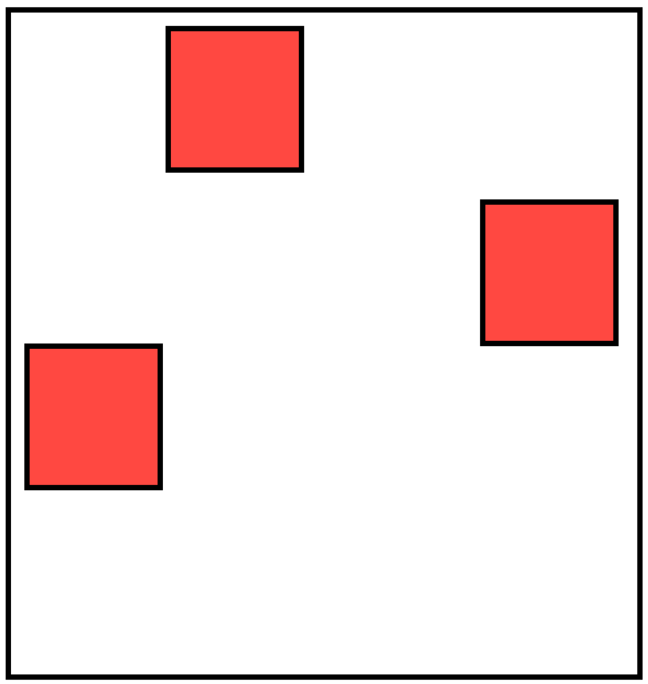


Score

$$\mathbf{w}^T \phi(\mathbf{a})$$

Linear Model

$\mathbf{a} \in \mathcal{A}'$



Score

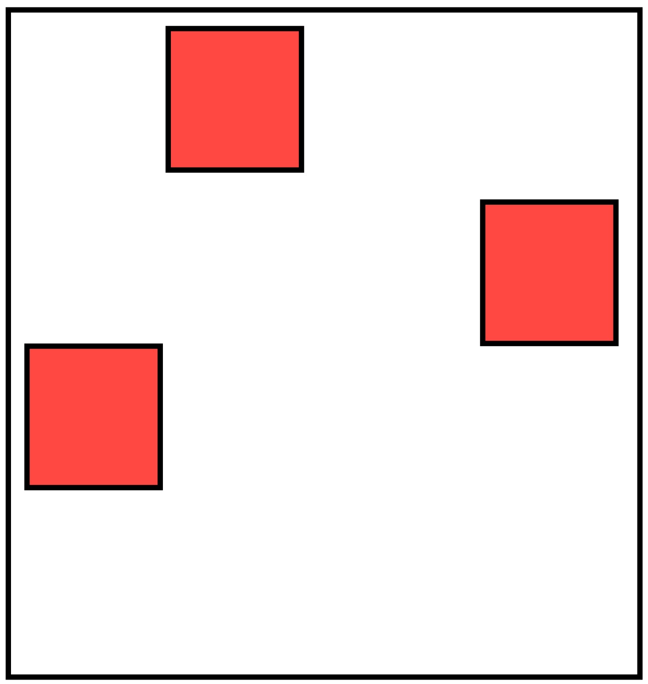
$$\mathbf{w}^T \phi(\mathbf{a})$$

Features

$$\phi(\mathbf{a}) = \sum_{(i,j) \in \mathbf{a}} \phi_{ij}$$

Linear Model

$$\mathbf{a} \in \mathcal{A}'$$



Score

$$\mathbf{w}^T \phi(\mathbf{a})$$

Features

$$\phi(\mathbf{a}) = \sum_{(i,j) \in \mathbf{a}} \phi_{ij}$$

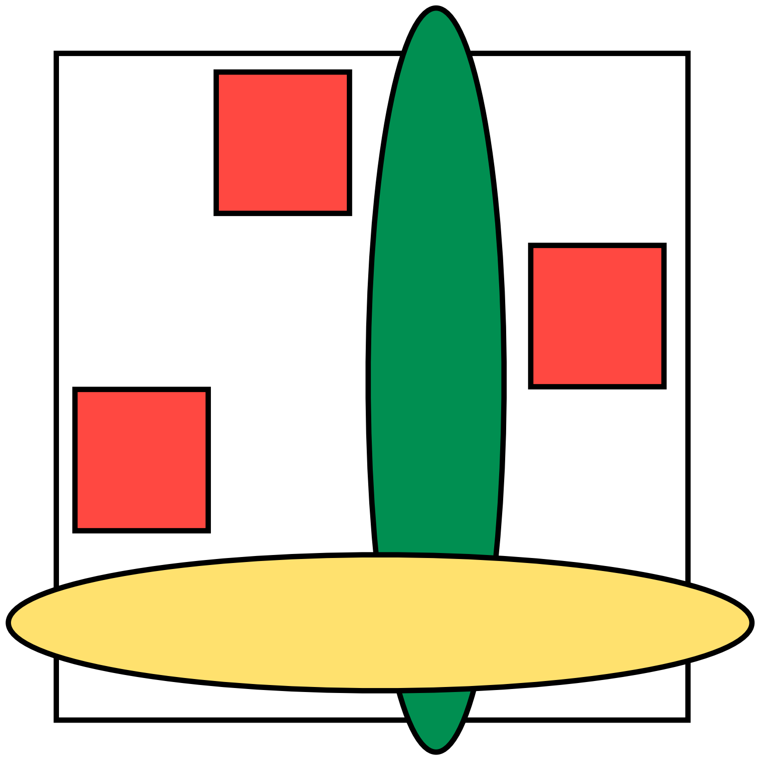
ϕ_{13} = { PartialDictMatch(*Indonesia*, 印)=true,

Dice(*Indonesia*, 印) = 0.85,

LinearDistance = 2, ...}

Linear Model

$$\mathbf{a} \in \mathcal{A}'$$



<u>Score</u>

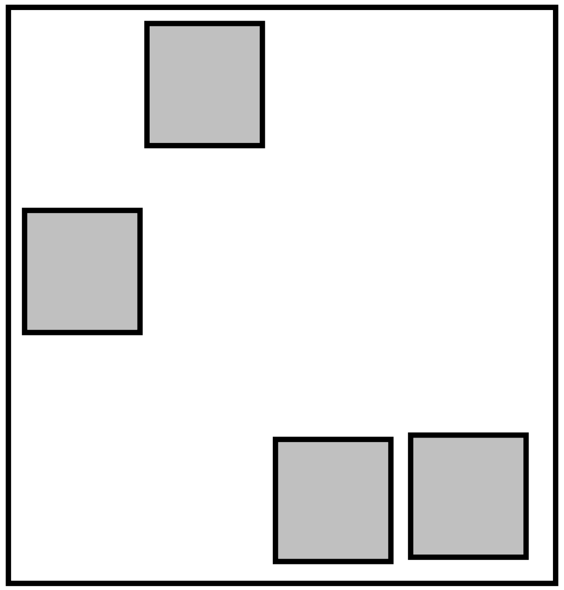
$$\mathbf{w}^T \phi(\mathbf{a})$$

<u>Features</u>

$$\phi(\mathbf{a}) = \sum_{(i,j)\in\mathbf{a}} \phi_{ij} + \sum_{i\notin\mathbf{a}} \phi_{i\epsilon} + \sum_{j\notin\mathbf{a}} \phi_{\epsilon j}$$

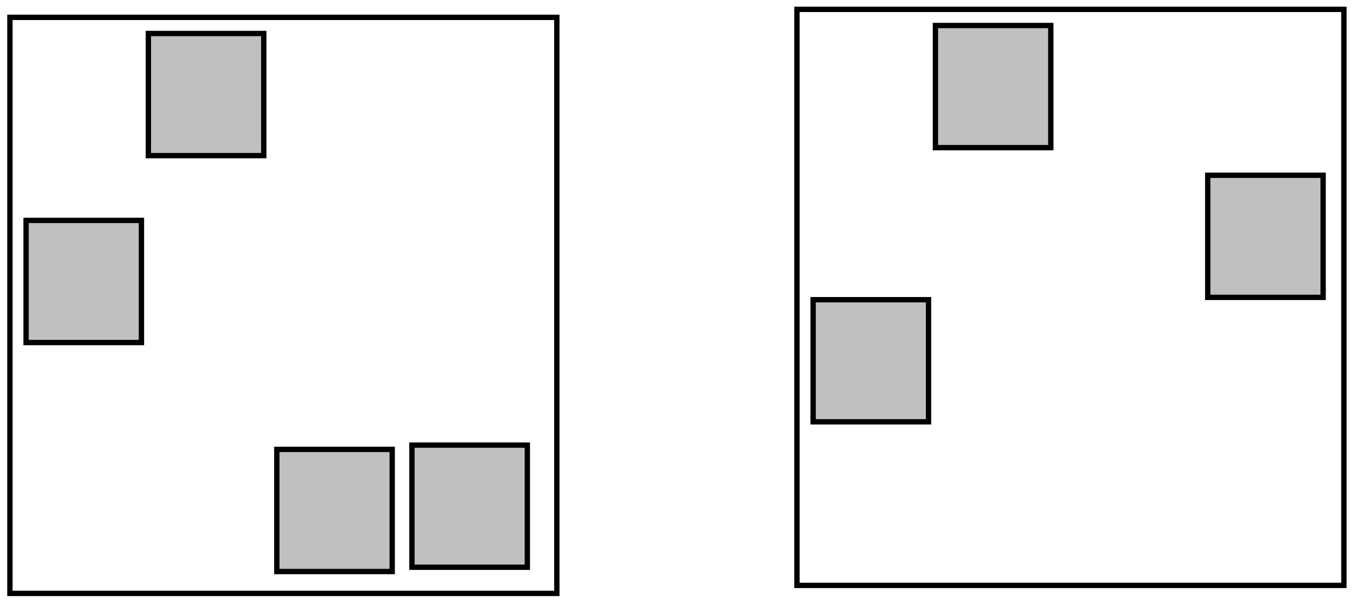
Margin Training

Margin Training



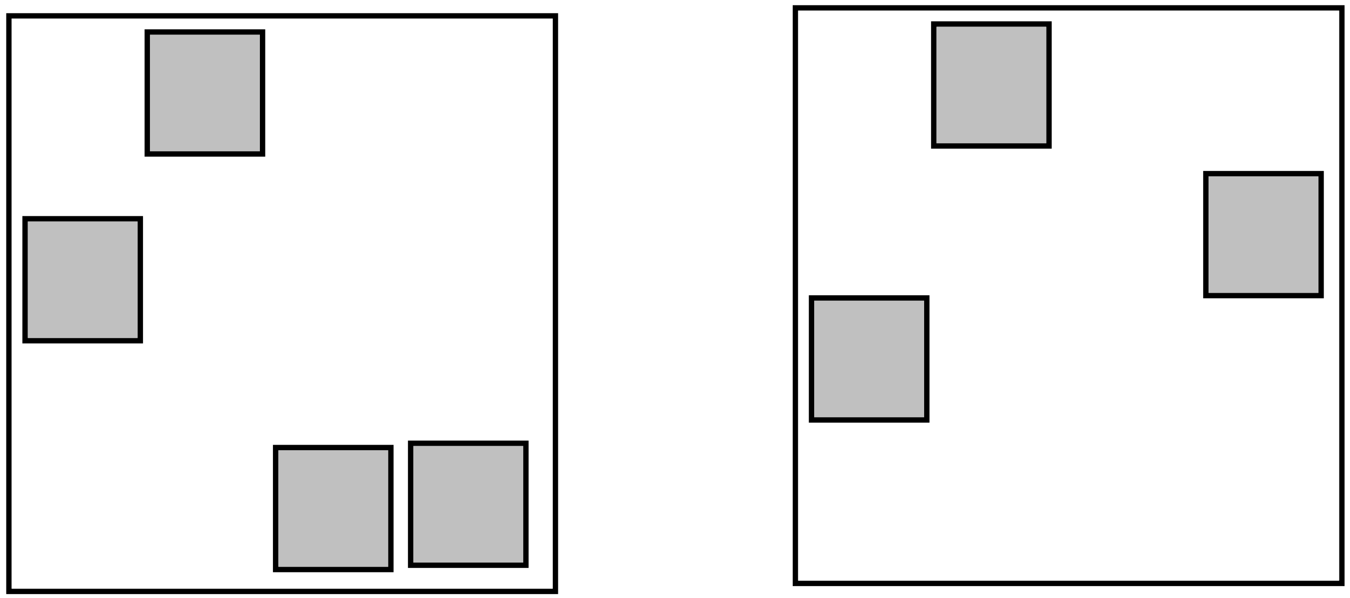
$\mathbf{a}^*$

Margin Training



$\mathbf{a}^*$ $\qquad$ $\hat{\mathbf{a}}$

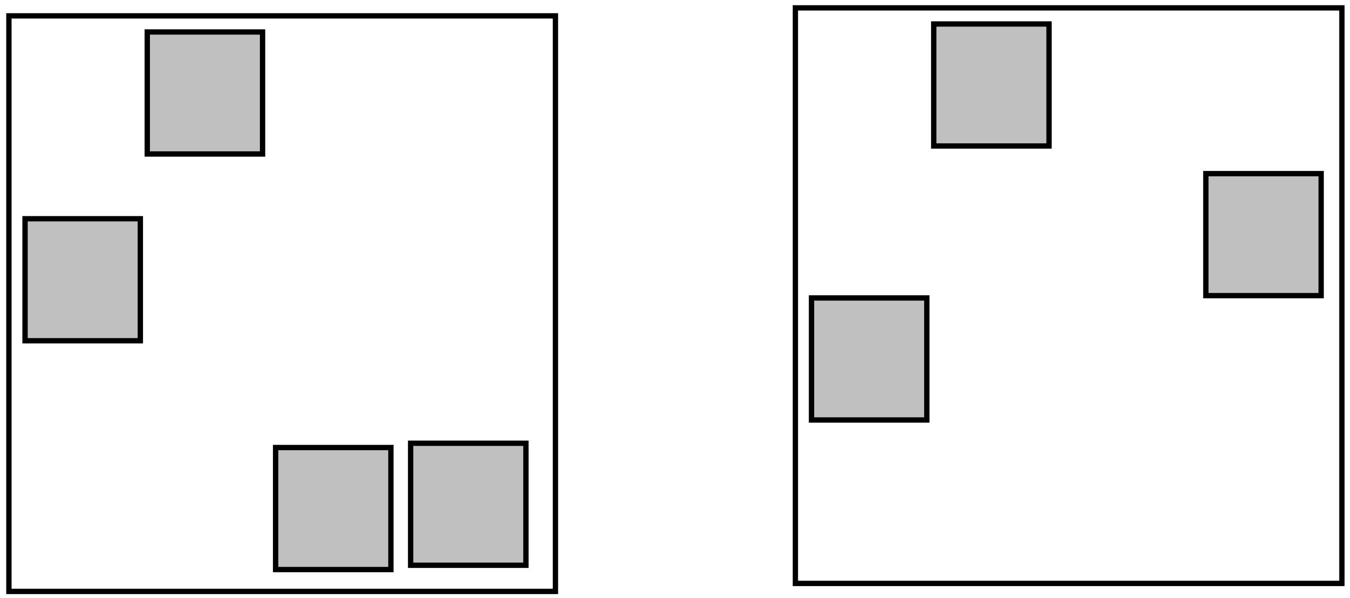
Margin Training



$\mathbf{a}^*$ $\hat{\mathbf{a}}$

$$\arg\max_{\mathbf{a}\in\mathcal{A}'}\left(\mathbf{w}^T\phi(\mathbf{a})\right)$$

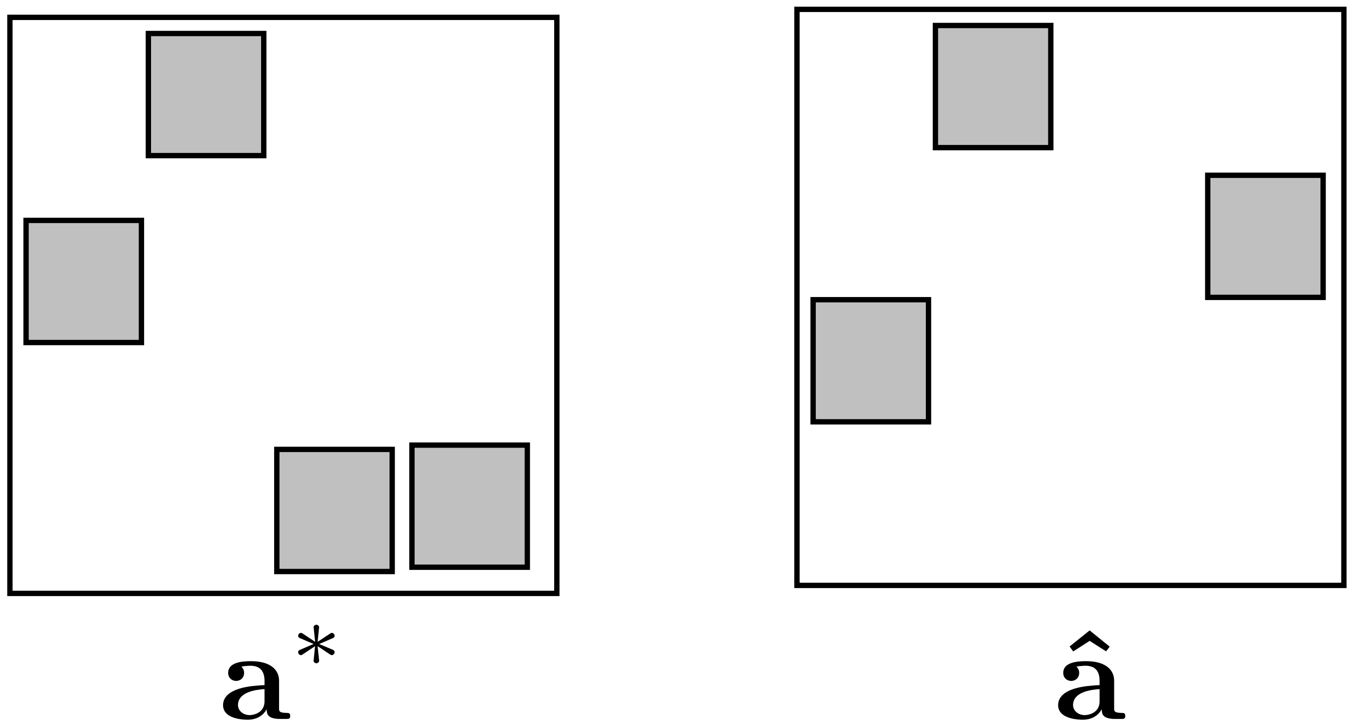
Margin Training



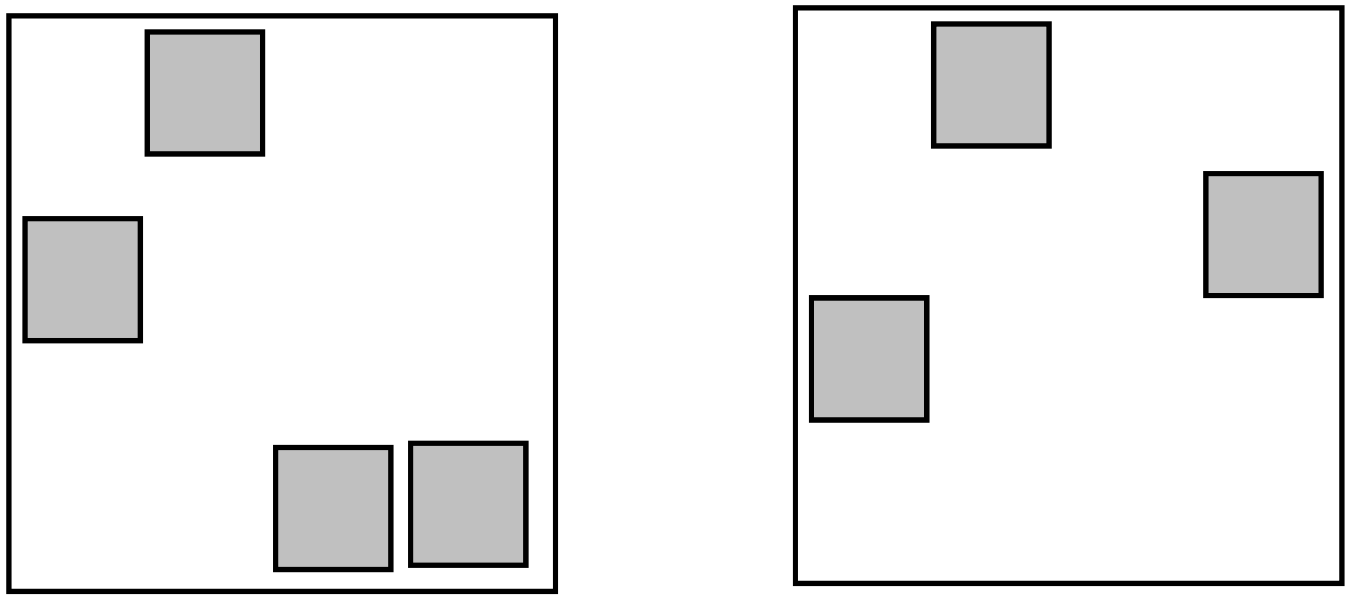
$\mathbf{a}^*$ $\hat{\mathbf{a}}$

$$\arg\max_{\mathbf{a}\in\mathcal{A}'}\left(\mathbf{w}^T\phi(\mathbf{a}) + \lambda L(\mathbf{a}^*, \mathbf{a})\right)$$

Margin Training



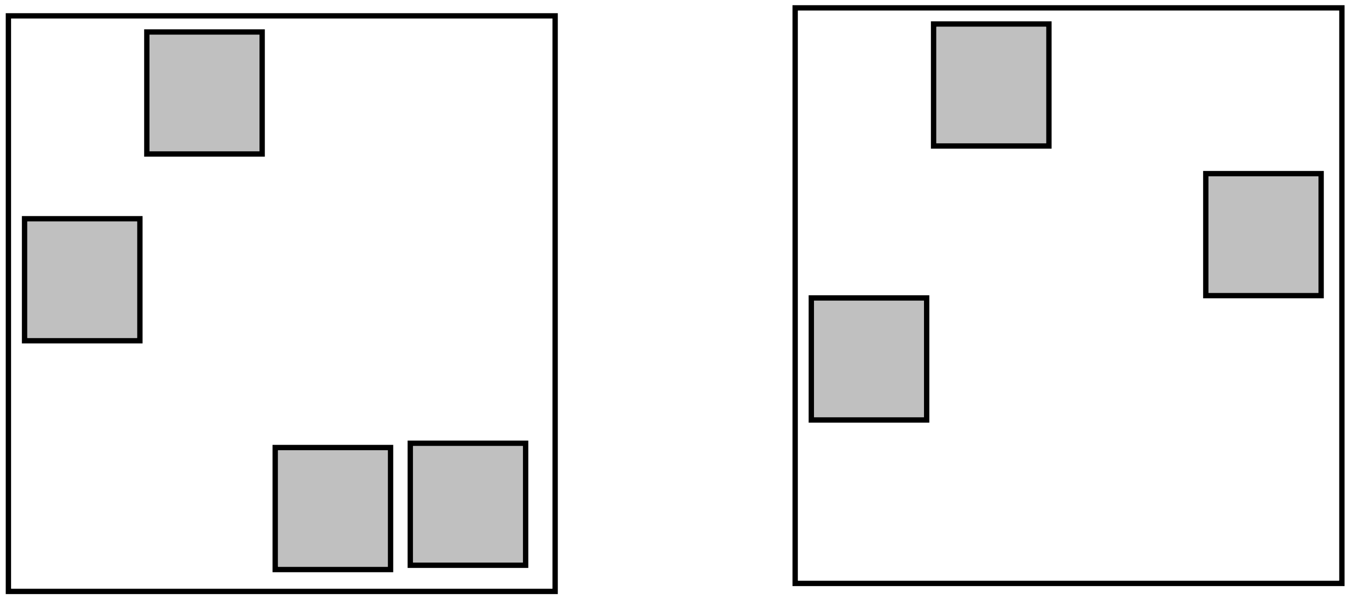
Margin Training



$\mathbf{a}^*$ $\hat{\mathbf{a}}$

$$\mathbf{w'}^T \phi(\mathbf{a}^*) \geq \mathbf{w'}^T \phi(\hat{\mathbf{a}}) + L(\mathbf{a}^*, \hat{\mathbf{a}})$$

Margin Training

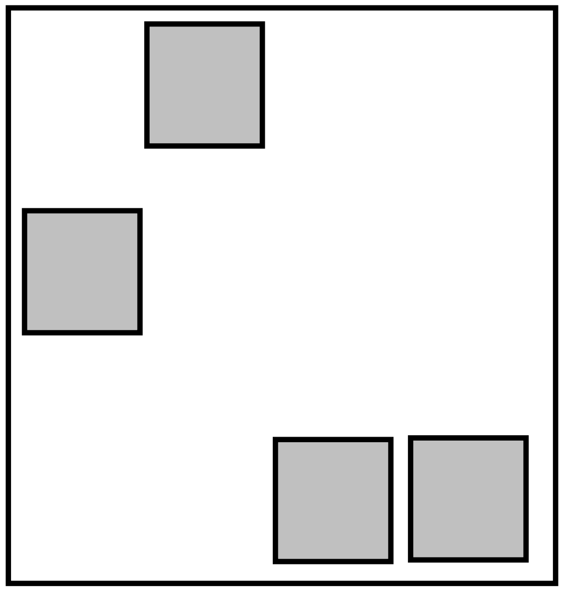


$\mathbf{a}^*$ $\hat{\mathbf{a}}$

$$\min_{\mathbf{w}'} \|\mathbf{w}' - \mathbf{w}\|^2$$

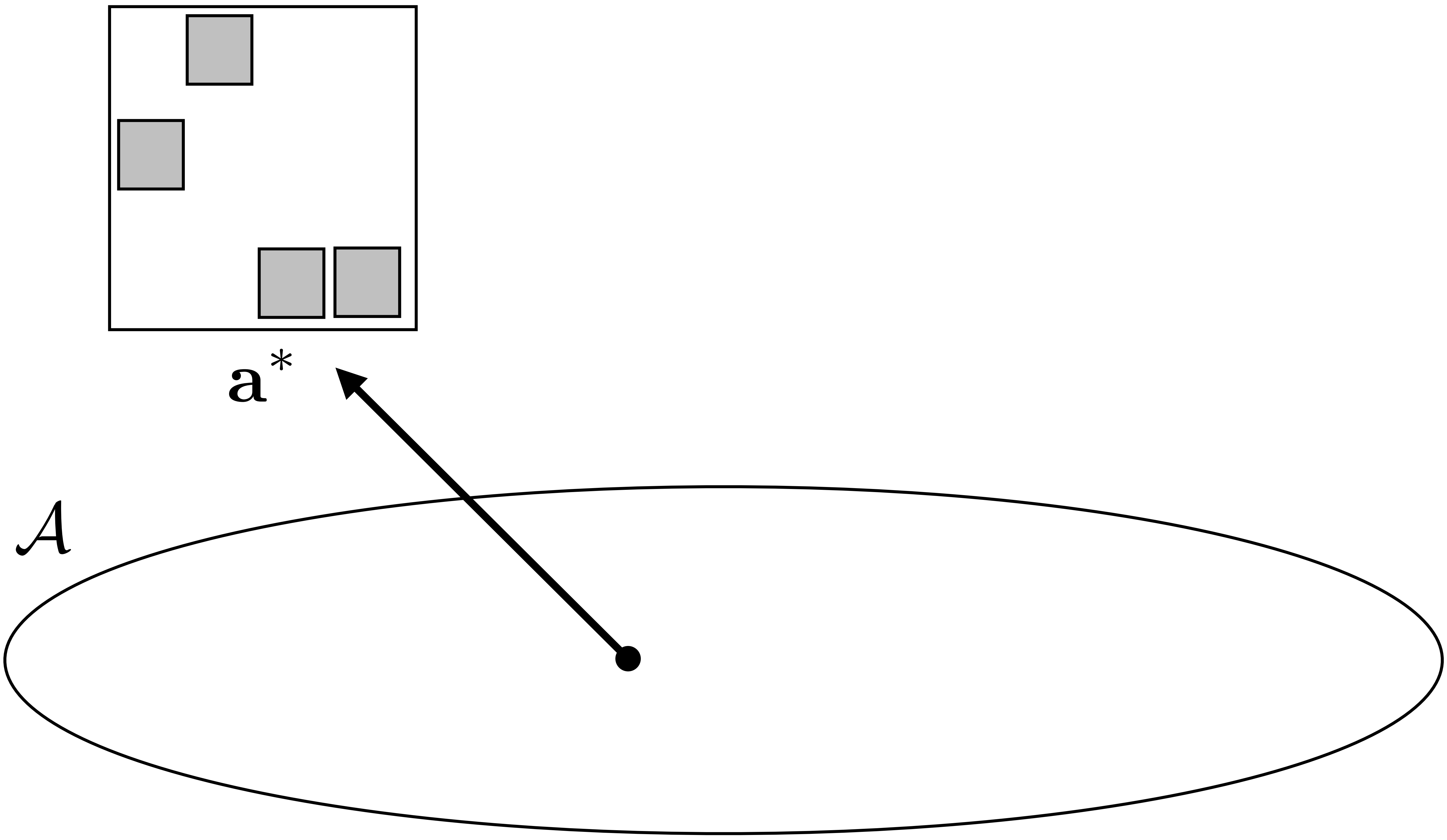
$$\mathbf{w}'^T \phi(\mathbf{a}^*) \geq \mathbf{w}'^T \phi(\hat{\mathbf{a}}) + L(\mathbf{a}^*, \hat{\mathbf{a}})$$

Oracle Projection

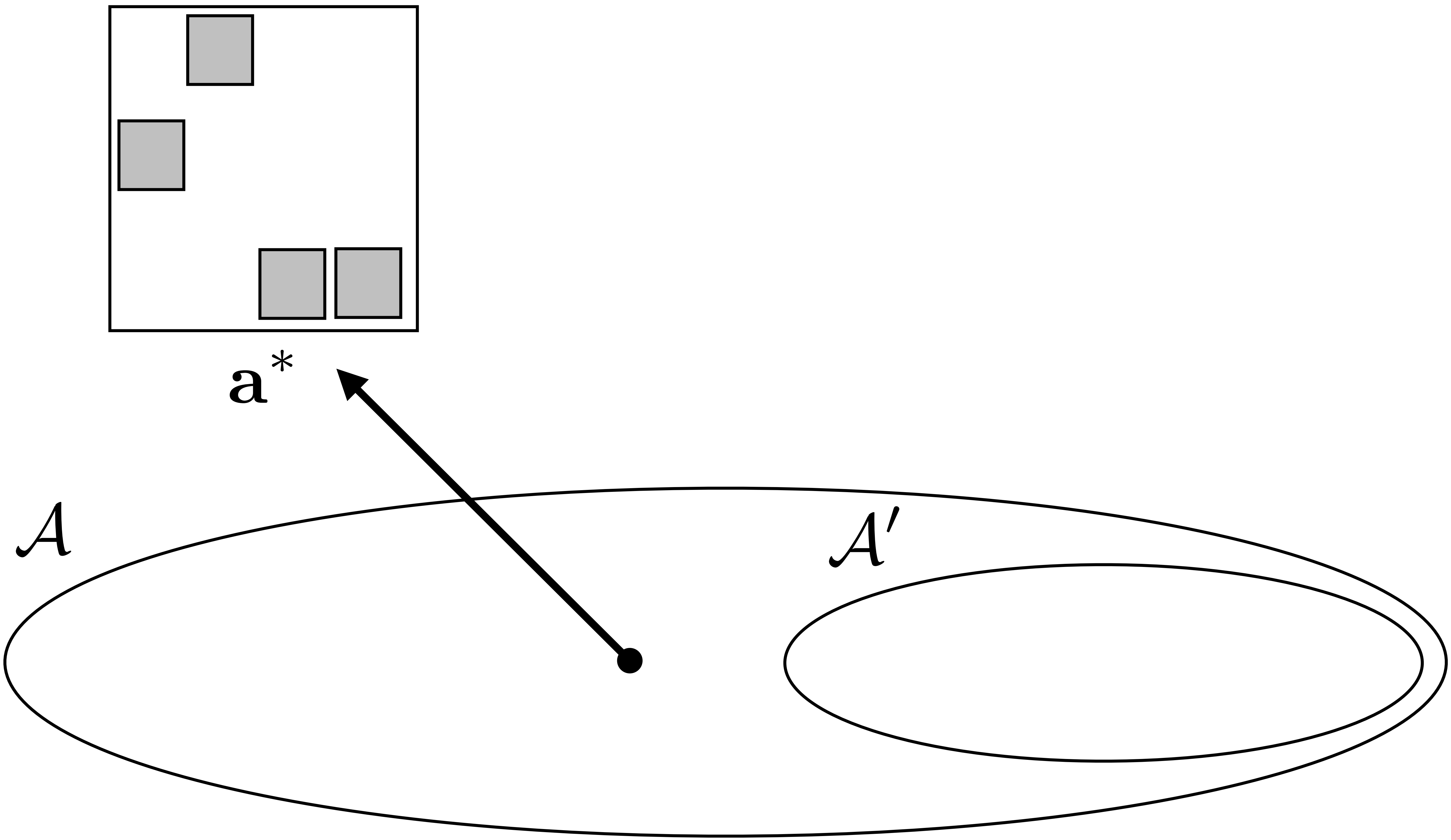


$\mathbf{a}^*$

Oracle Projection

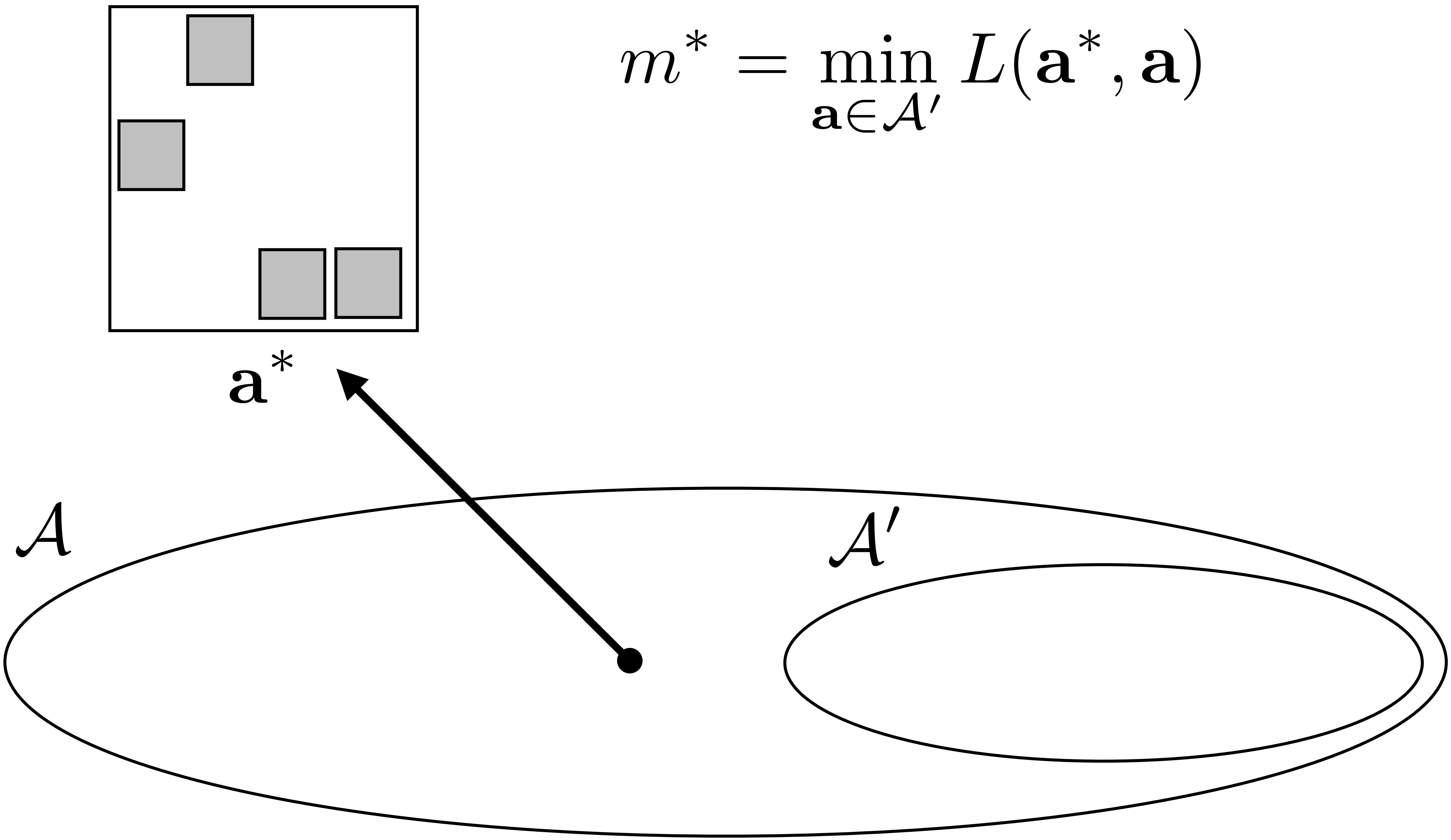


Oracle Projection

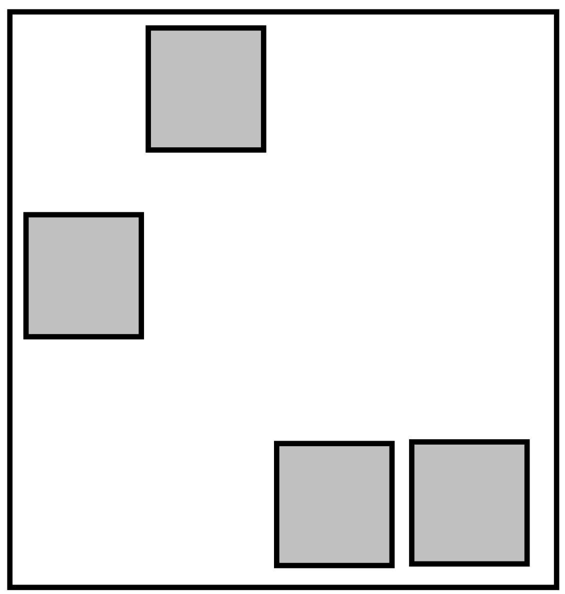


Oracle Projection

$$m^* = \min_{\mathbf{a} \in \mathcal{A}'} L(\mathbf{a}^*, \mathbf{a})$$



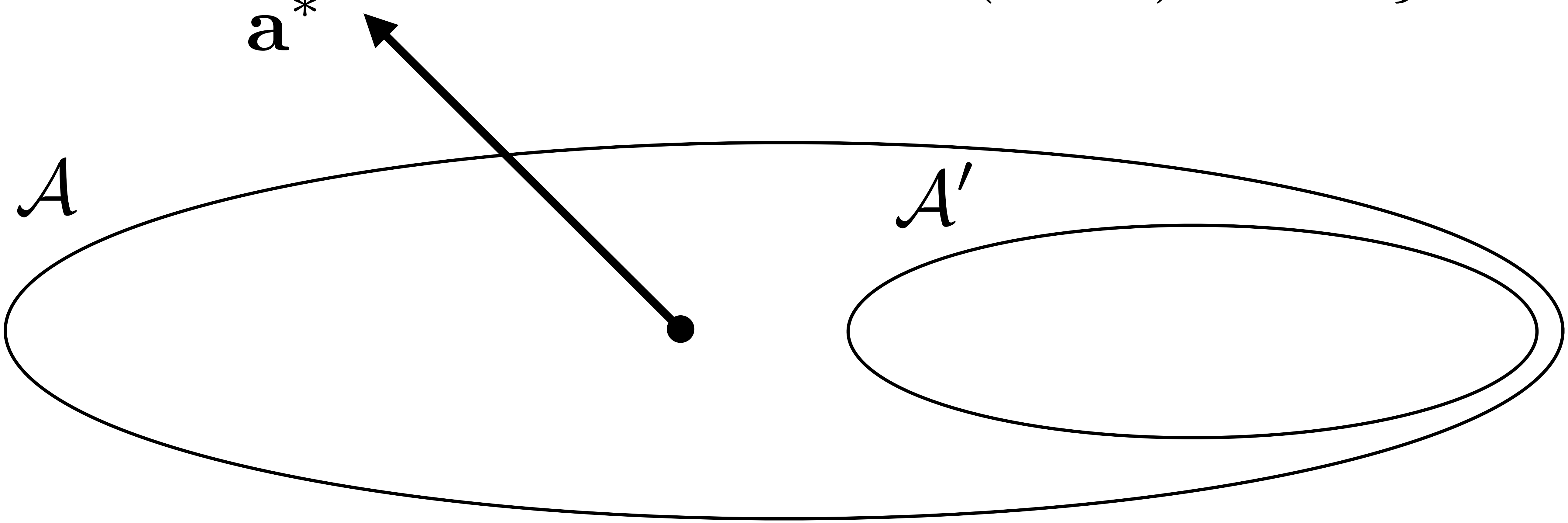
Oracle Projection



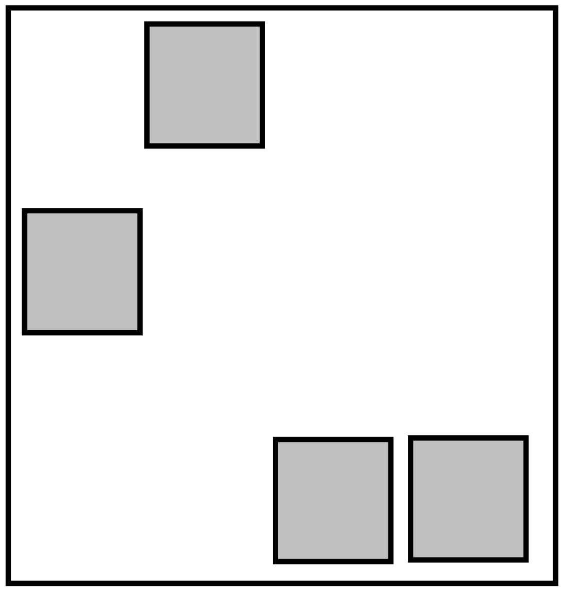
$\mathbf{a}^*$

$$m^* = \min_{\mathbf{a} \in \mathcal{A}'} L(\mathbf{a}^*, \mathbf{a})$$

$$\mathcal{M}(\mathbf{a}^*) = \{\mathbf{a} \in \mathcal{A}' \text{ s.t. } L(\mathbf{a}^*, \mathbf{a}) = m^*\}$$



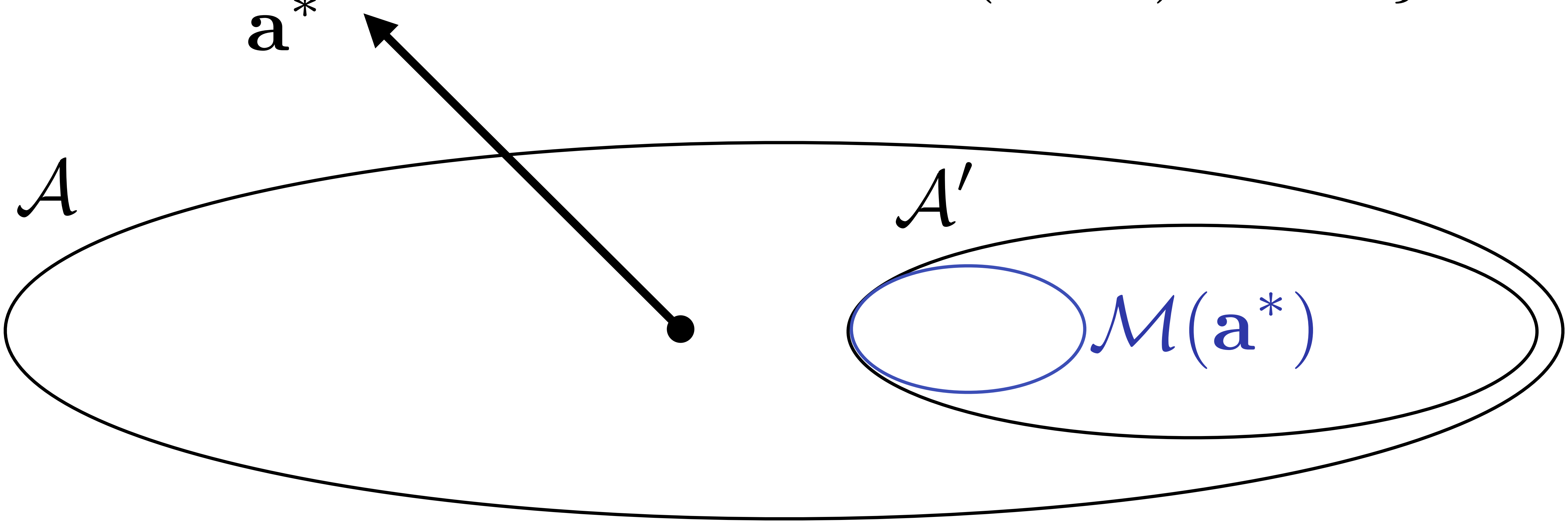
Oracle Projection



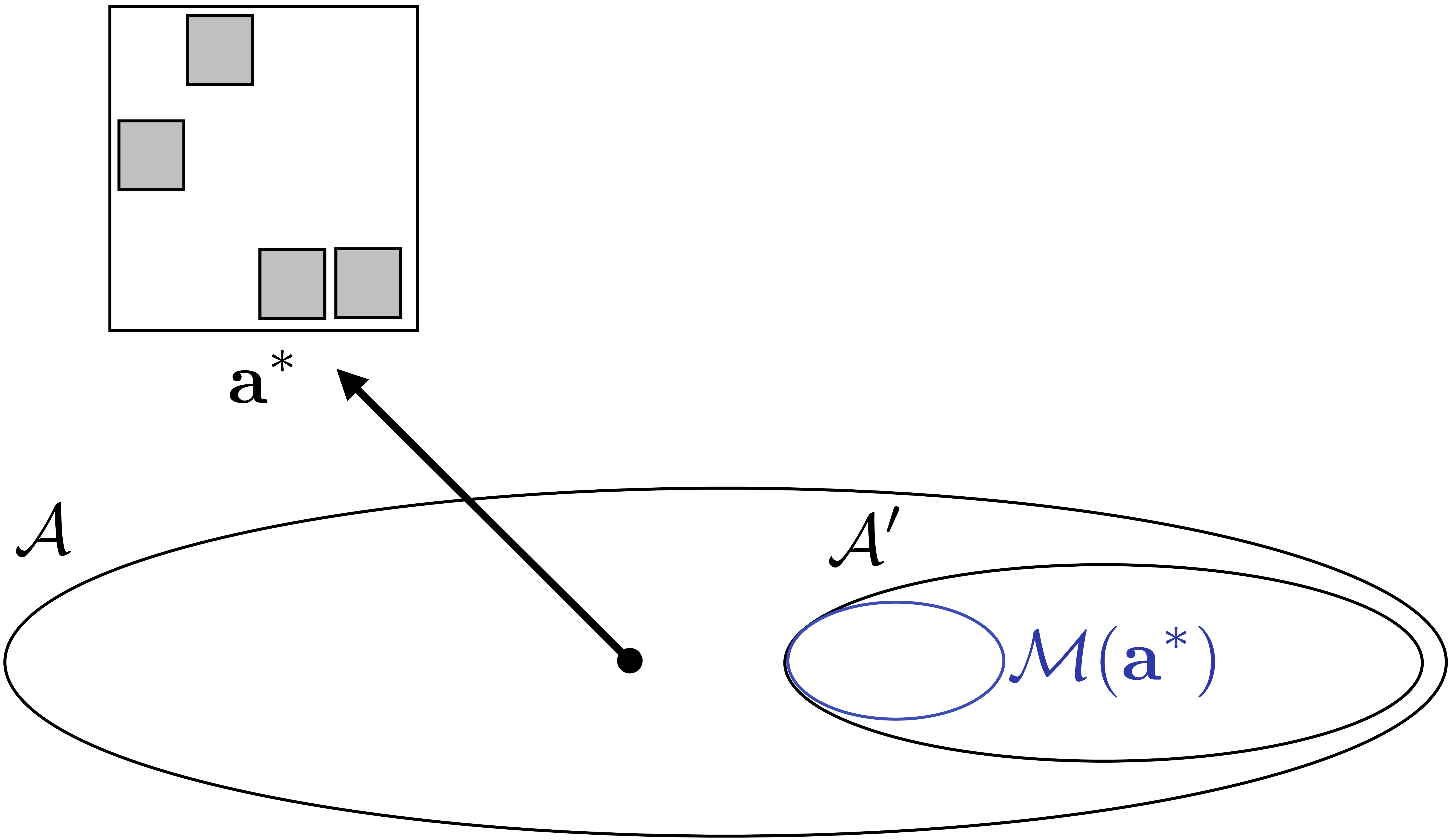
$\mathbf{a}^*$

$$m^* = \min_{\mathbf{a} \in \mathcal{A}'} L(\mathbf{a}^*, \mathbf{a})$$

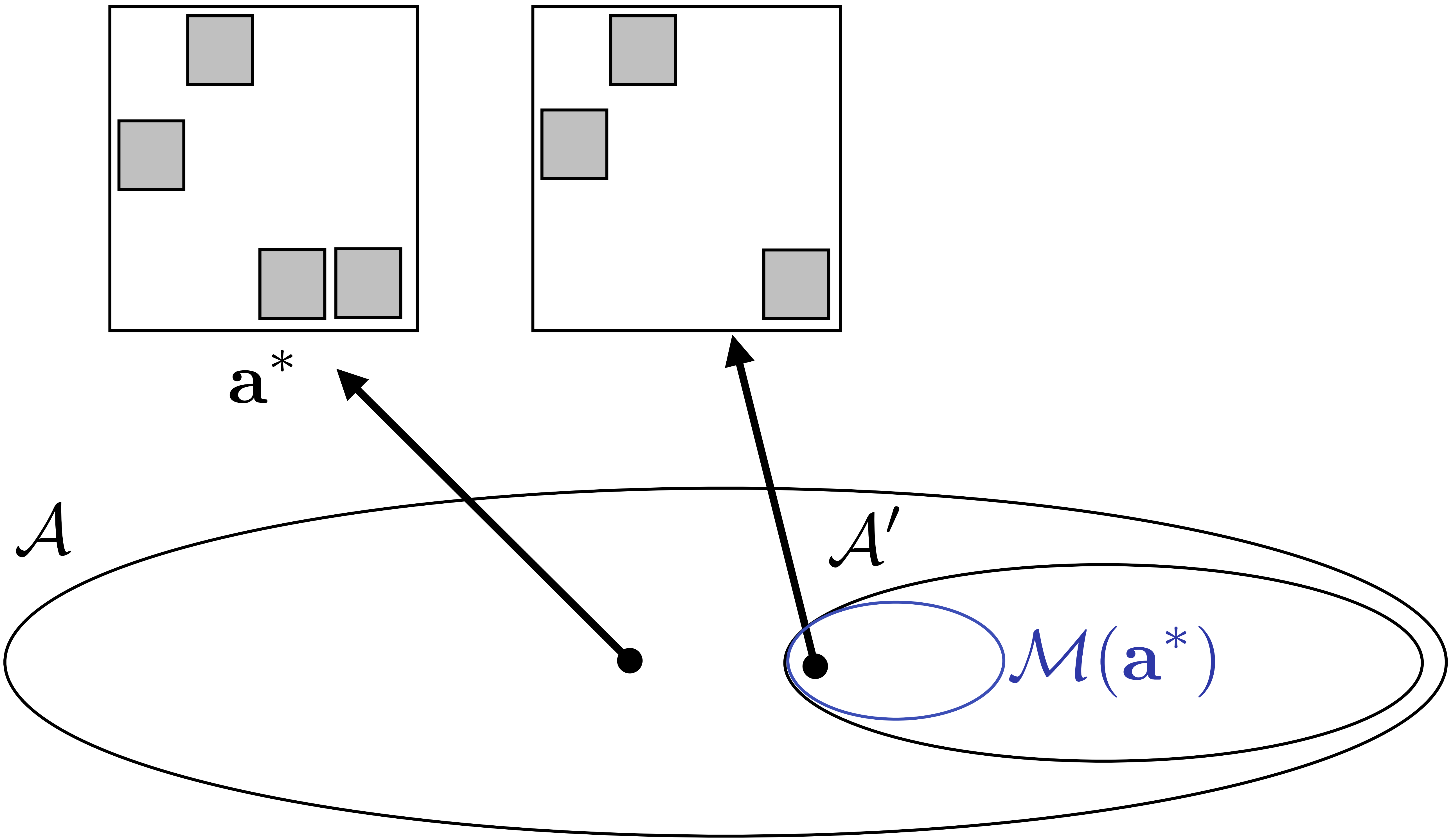
$$\mathcal{M}(\mathbf{a}^*) = \{\mathbf{a} \in \mathcal{A}' \text{ s.t. } L(\mathbf{a}^*, \mathbf{a}) = m^*\}$$



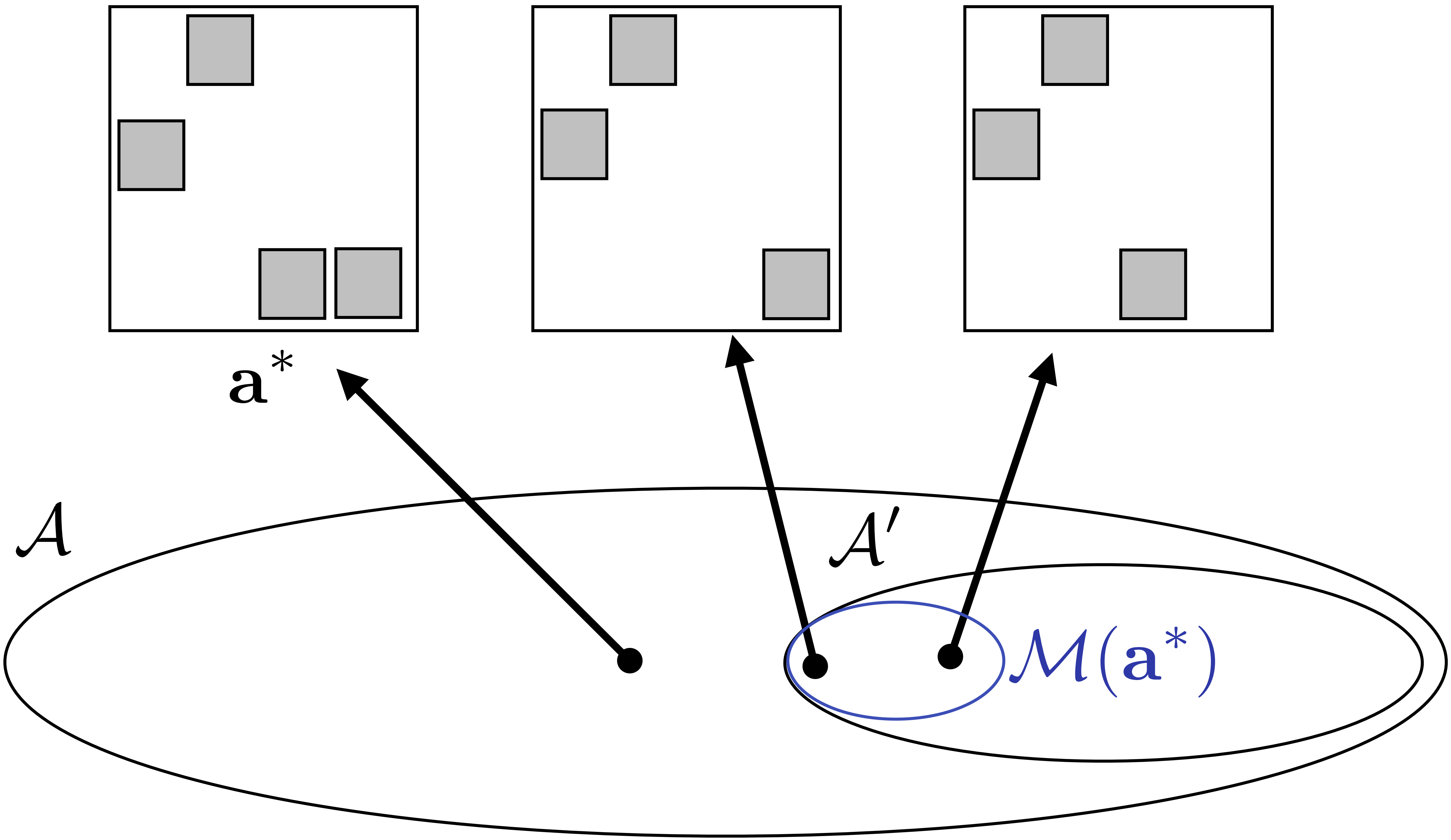
Oracle Projection



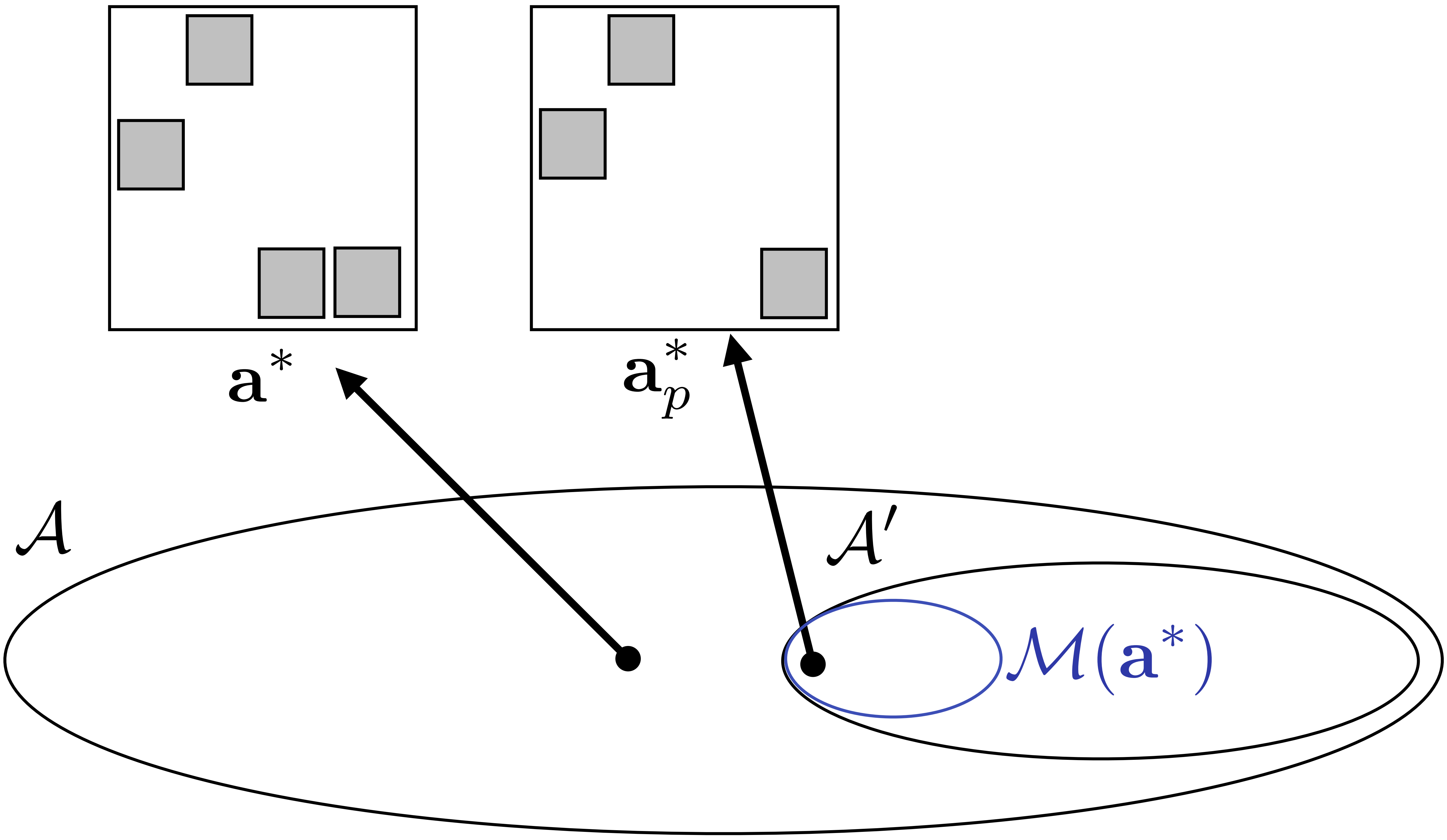
Oracle Projection



Oracle Projection



Oracle Projection



Learning Alignments

1-1 ITG BITG

Learning Alignments

1-1 ITG BITG

- MIRA Trained

Learning Alignments

1-1 ITG BITG

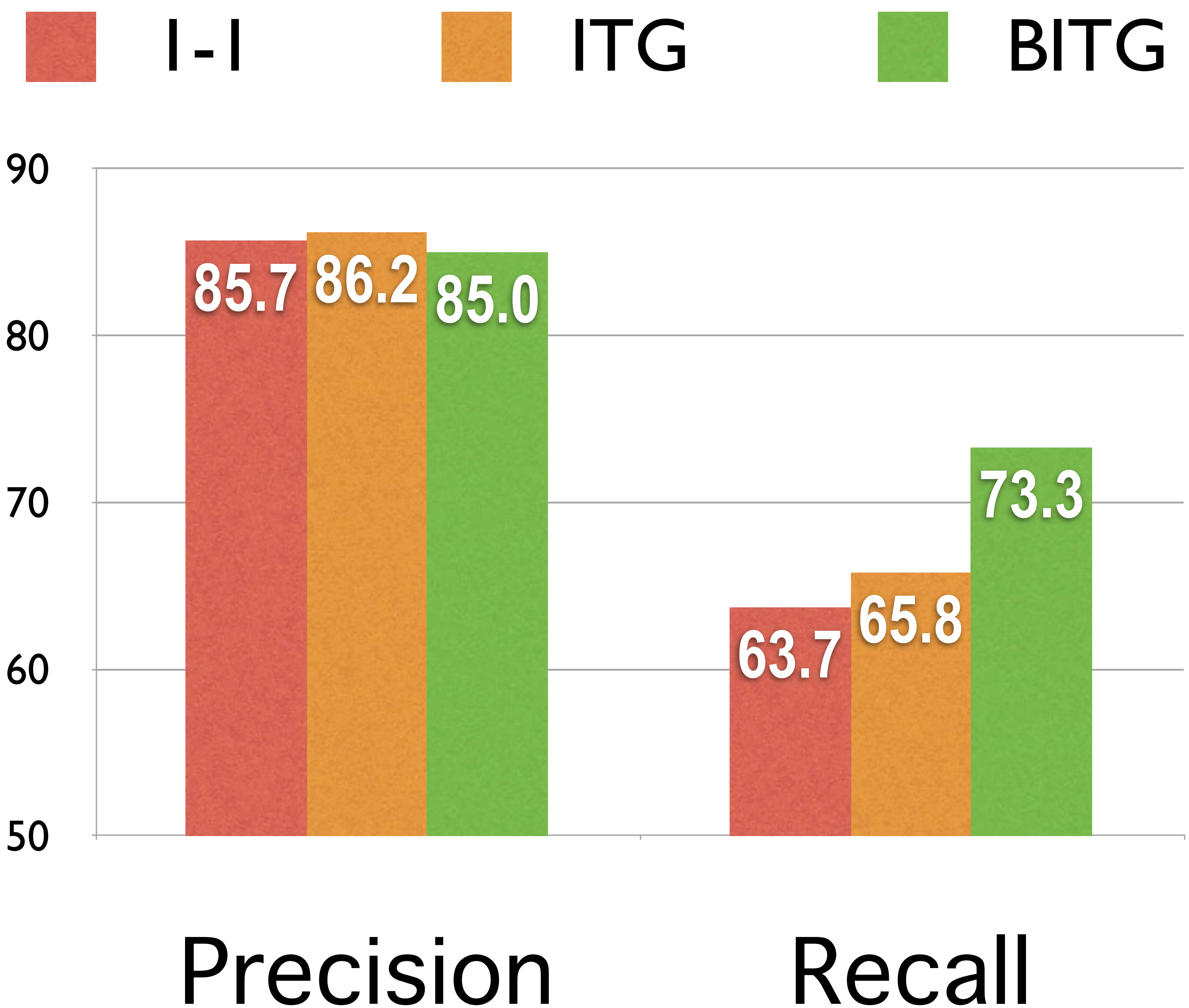
- MIRA Trained
- Viterbi Inference

Learning Alignments

1-1 ITG BITG

- MIRA Trained
- Viterbi Inference
- Simple Features
  - Dice
  - Lexical
  - Distance
  - Dictionary

Learning Alignments



- MIRA Trained
- Viterbi Inference
- Simple Features
  - Dice
  - Lexical
  - Distance
  - Dictionary

Likelihood Training

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$\mathcal{A}_{1\text{-}1}$

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$\mathcal{A}_{1\text{-}1}$

$\boxed{\#P}$

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$\mathcal{A}_{1\text{-}1}$

$$\boxed{\cancel{\#P}}$$

Likelihood Training

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T\phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a}\in\mathcal{A}'} \exp\{\mathbf{w}^T\phi(\mathbf{a})\}$$

$\mathcal{A}_{\text{1-1}}$ $\quad$ $\mathcal{A}_{ITG}$

$\cancel{\#P}$

Likelihood Training

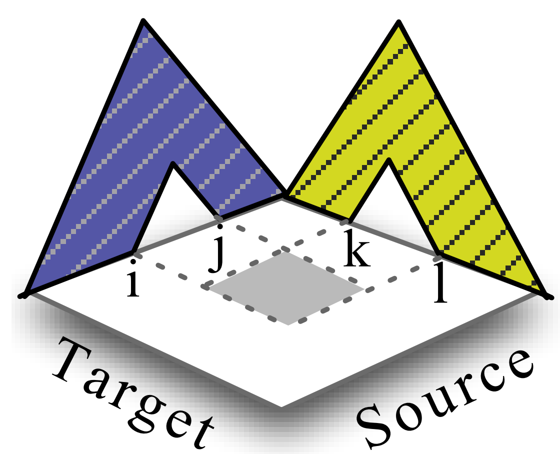
$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$\mathcal{A}_{1\text{-}1}$

~~#P~~

$\mathcal{A}_{ITG}$



Likelihood Training

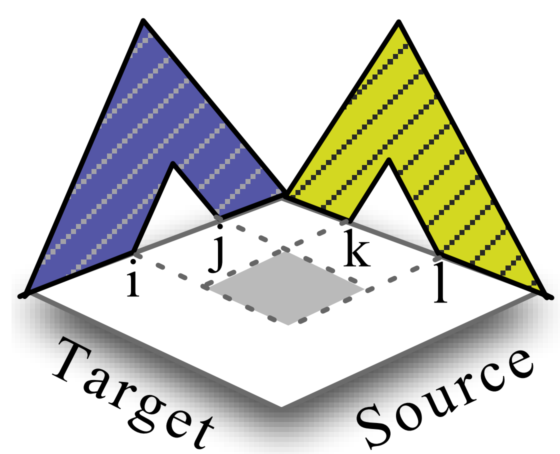
$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$Z_{\mathbf{w}}(\mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}'} \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

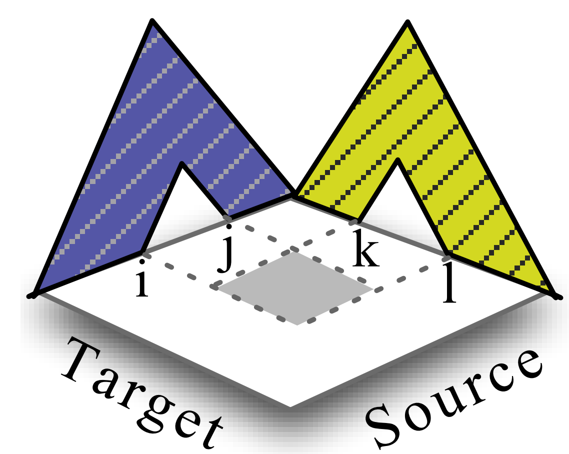
$\mathcal{A}_{1\text{-}1}$

~~#P~~

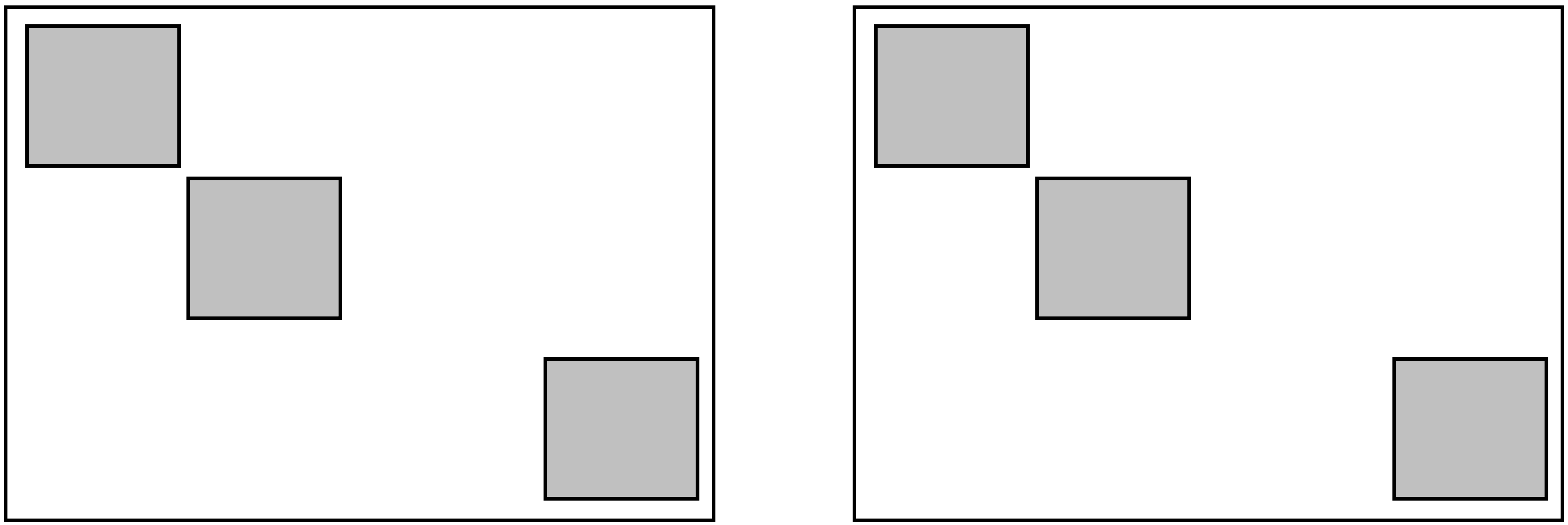
$\mathcal{A}_{ITG}$



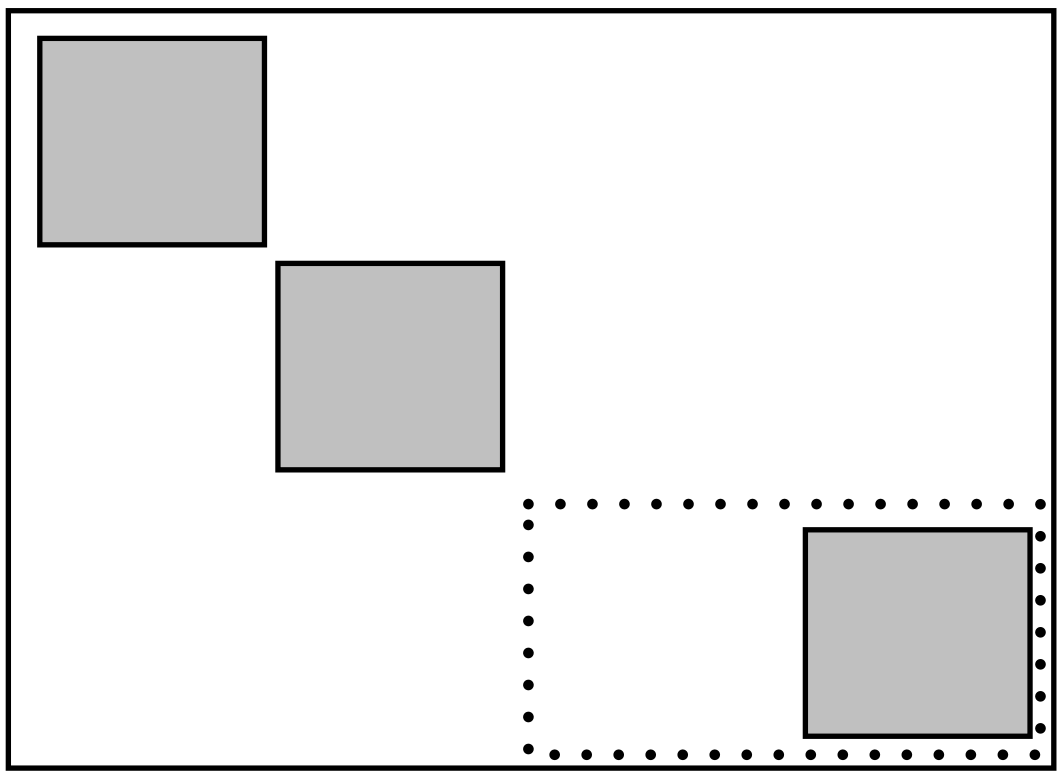
$\mathcal{A}_{BITG}$

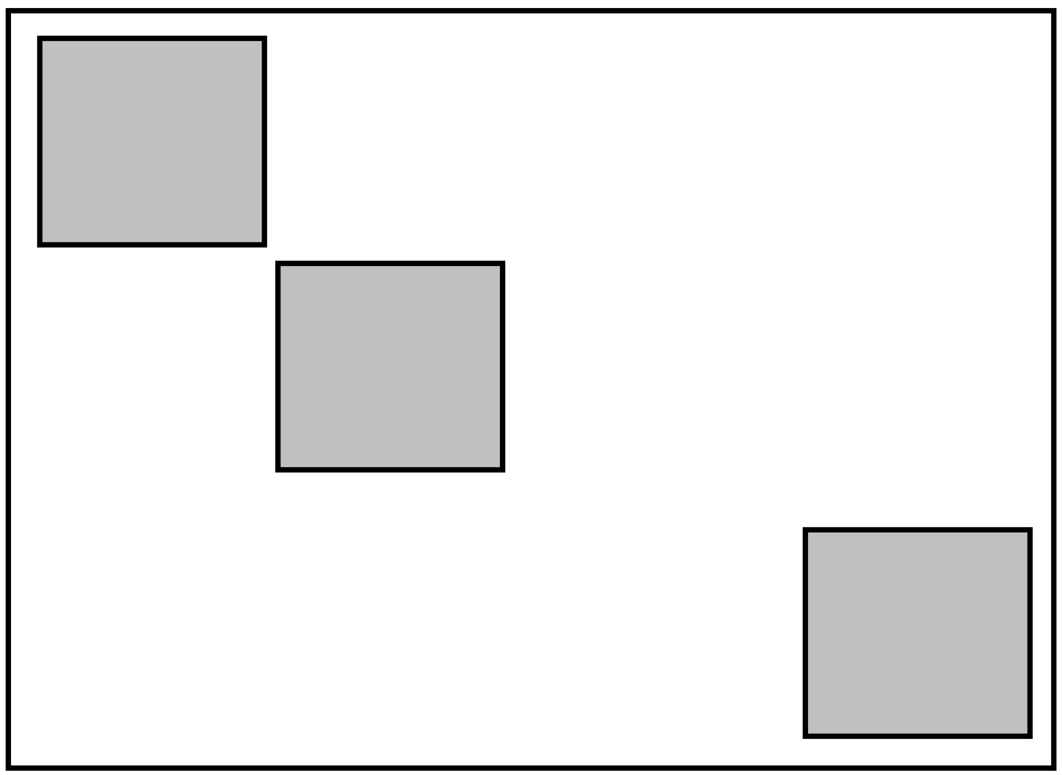


Derivations vs. Alignments

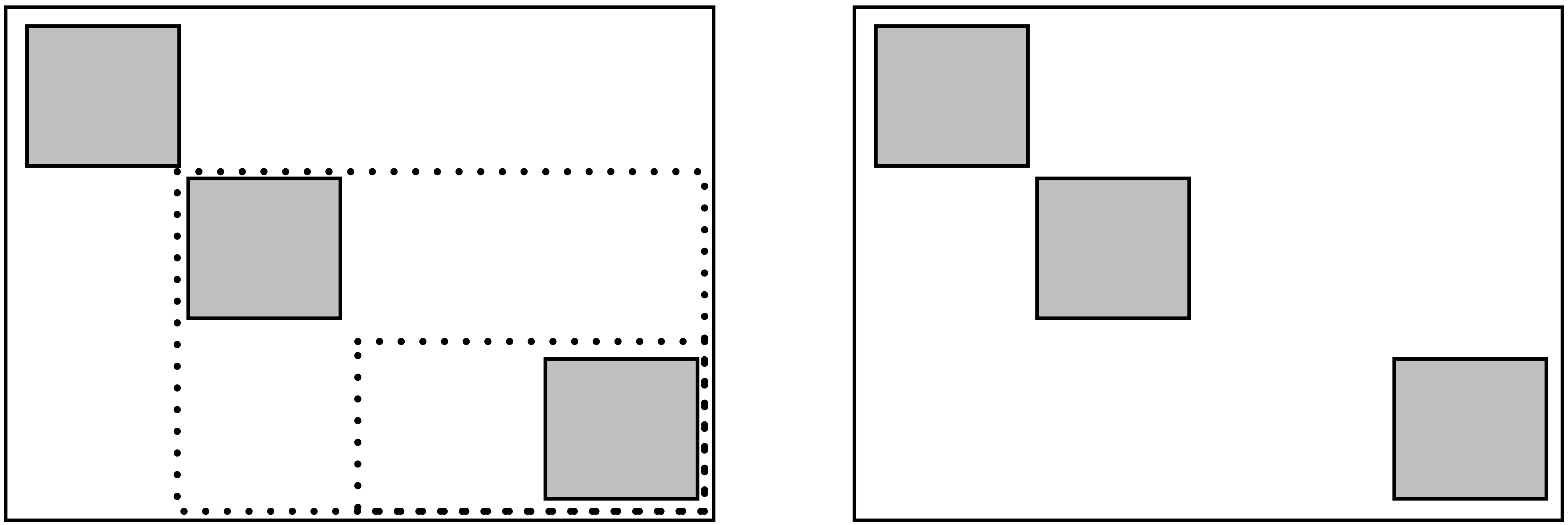


Derivations vs. Alignments

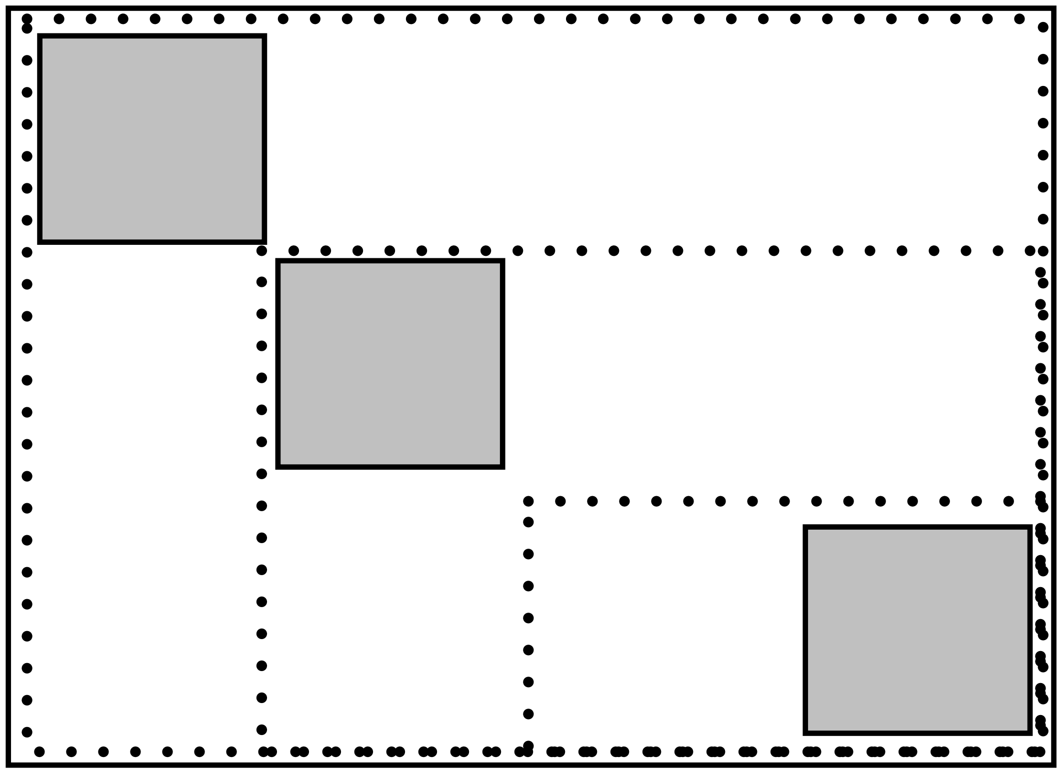


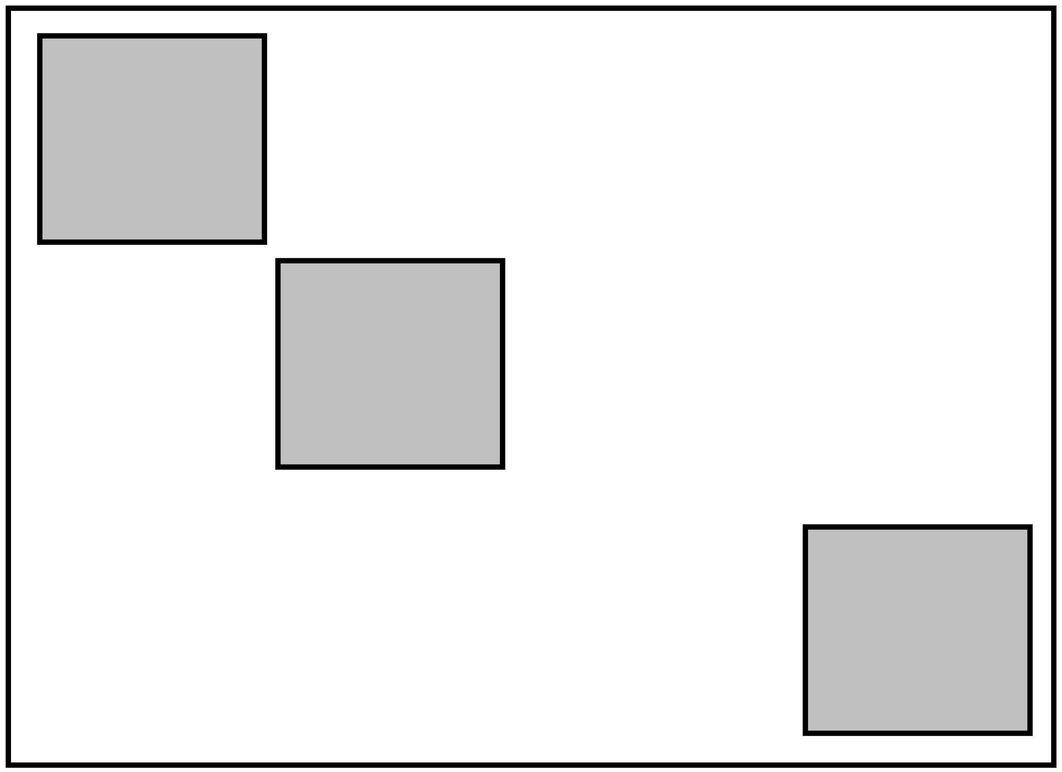


Derivations vs. Alignments

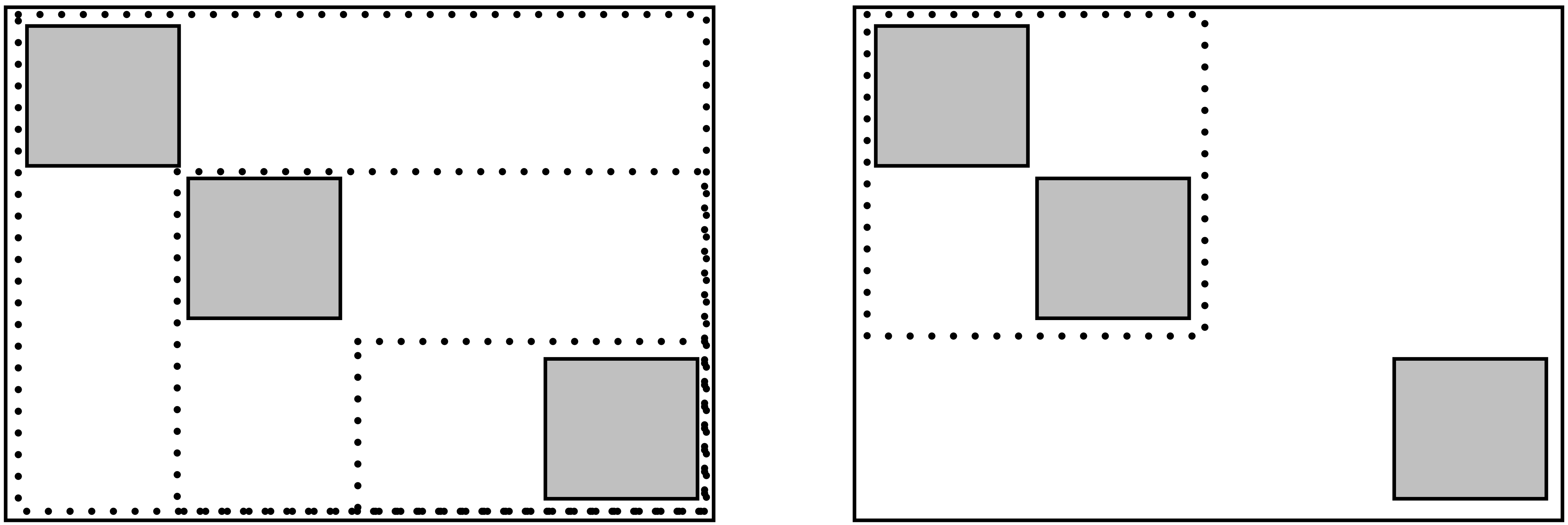


Derivations vs. Alignments

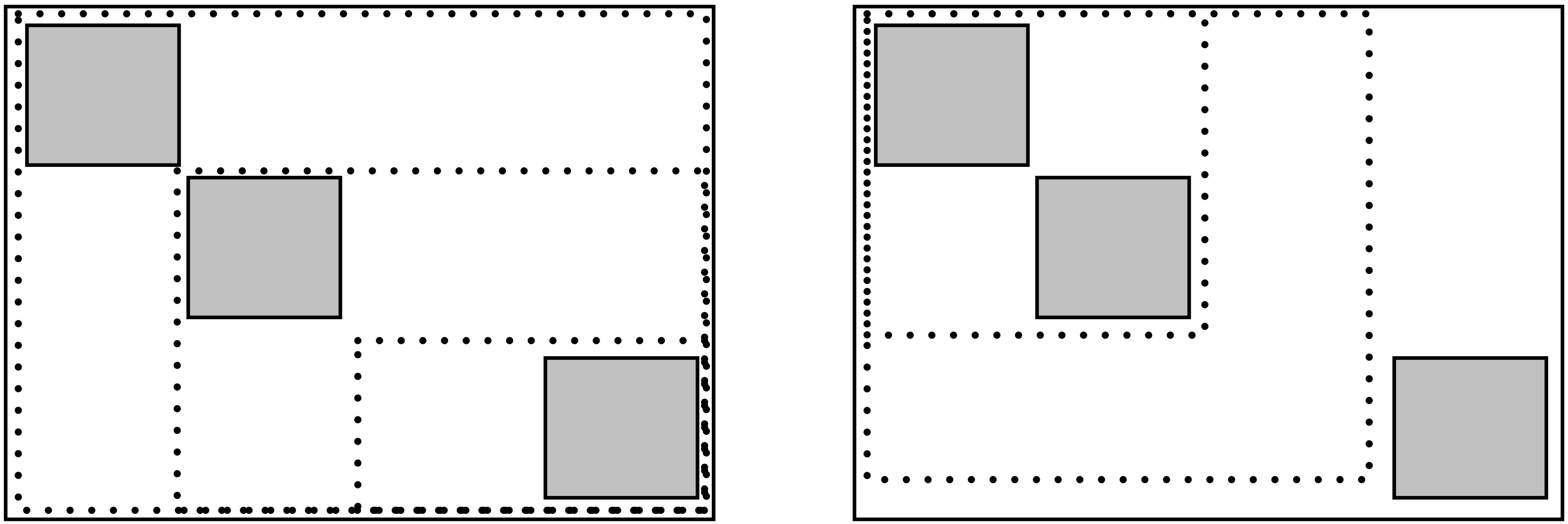




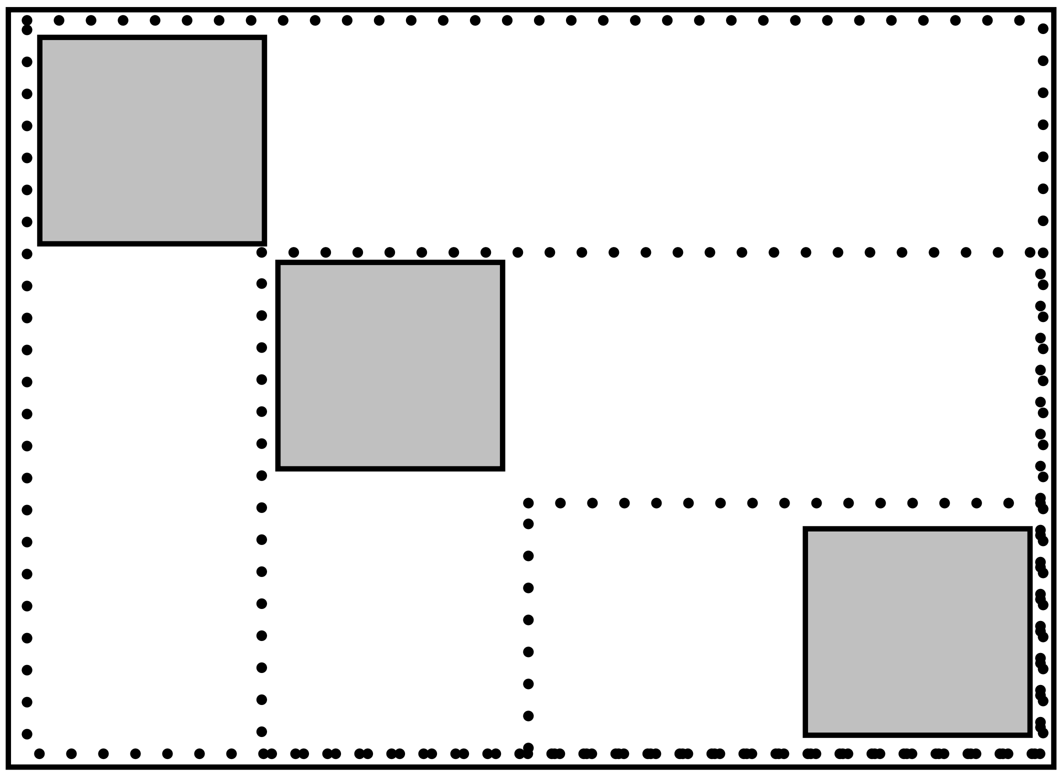
Derivations vs. Alignments

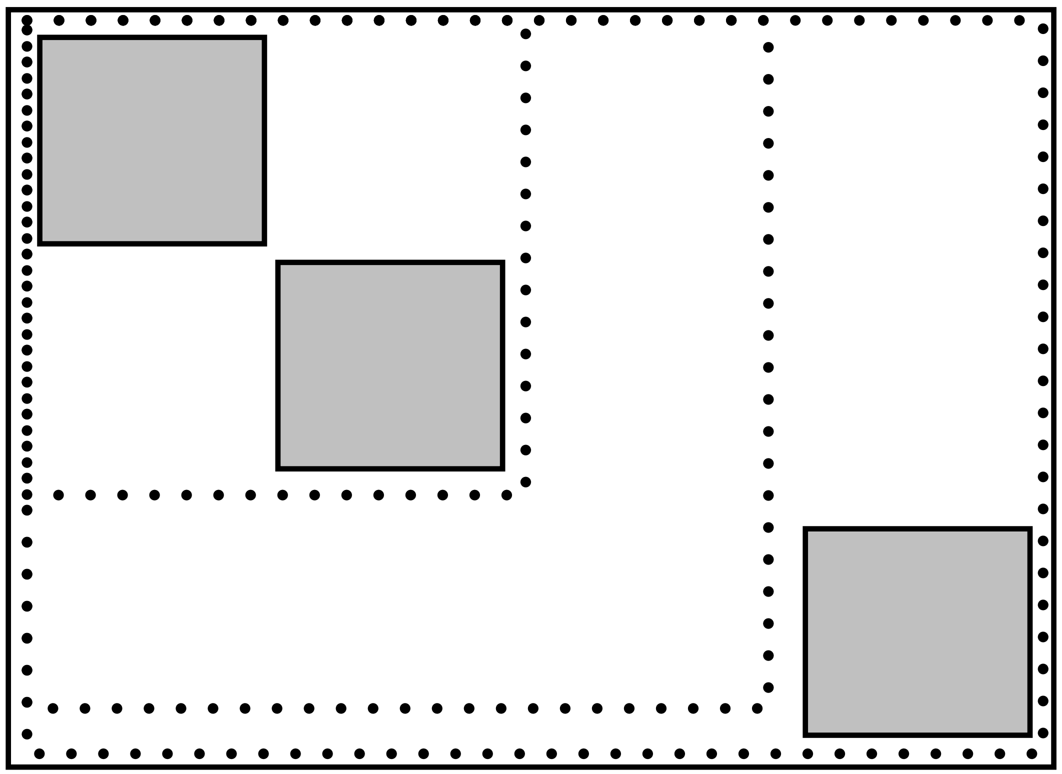


Derivations vs. Alignments

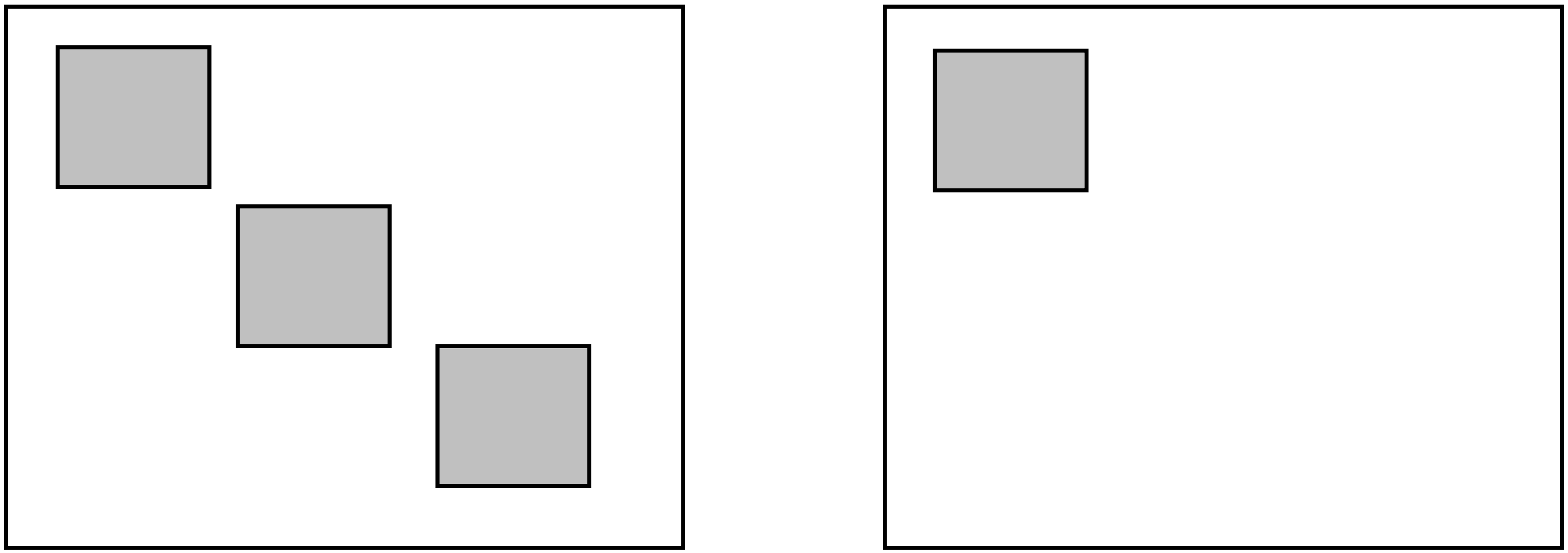


Derivations vs. Alignments



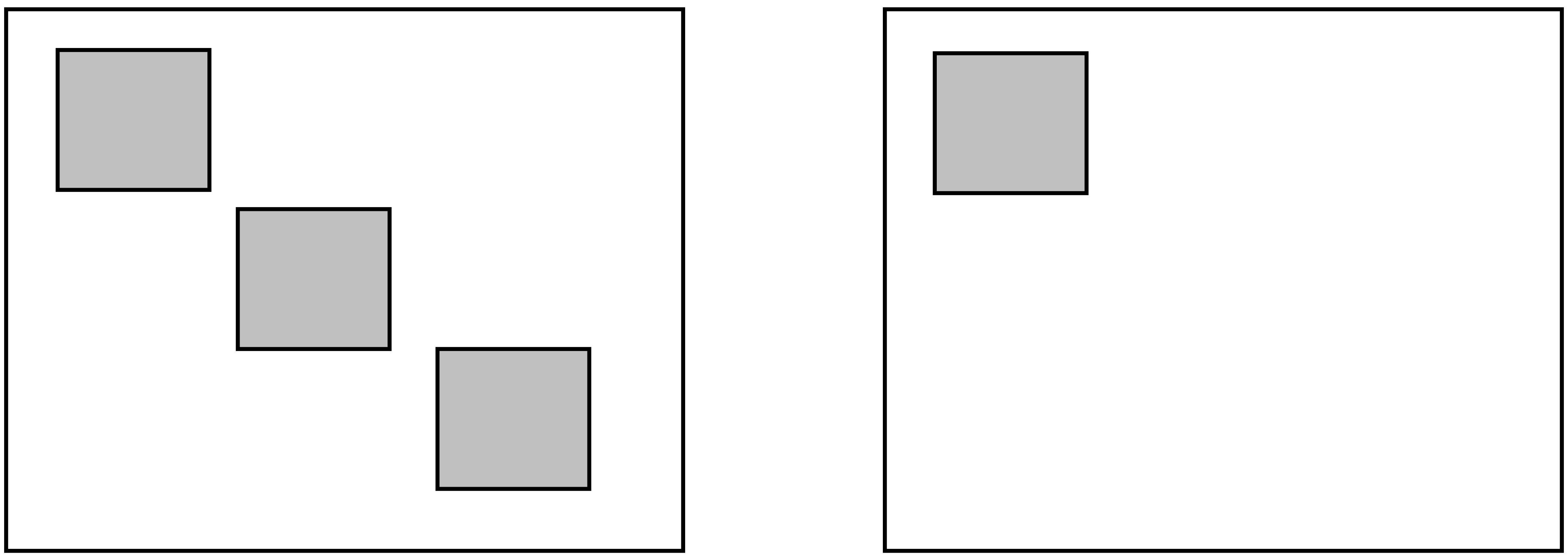


ITG Normal Form



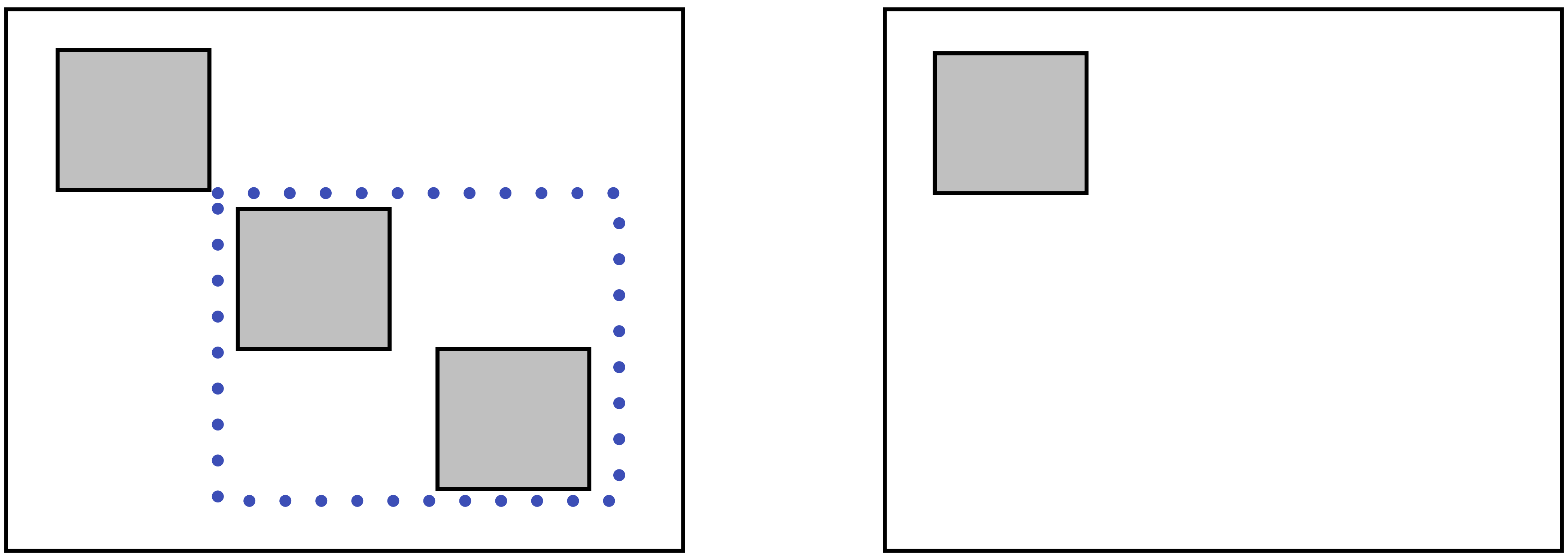
ITG Normal Form

N-ary Productions



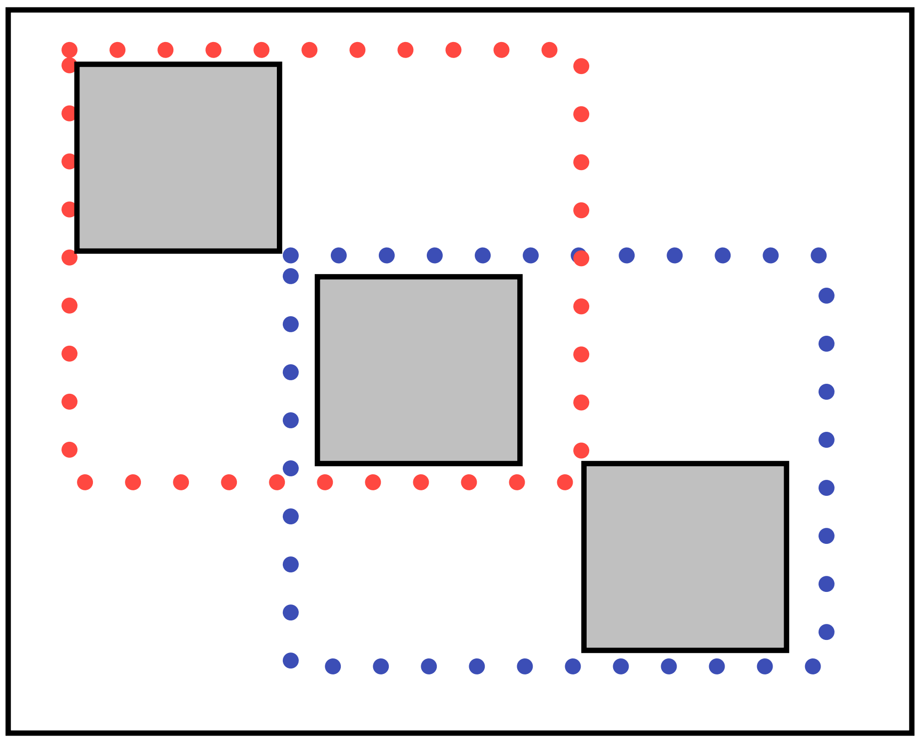
ITG Normal Form

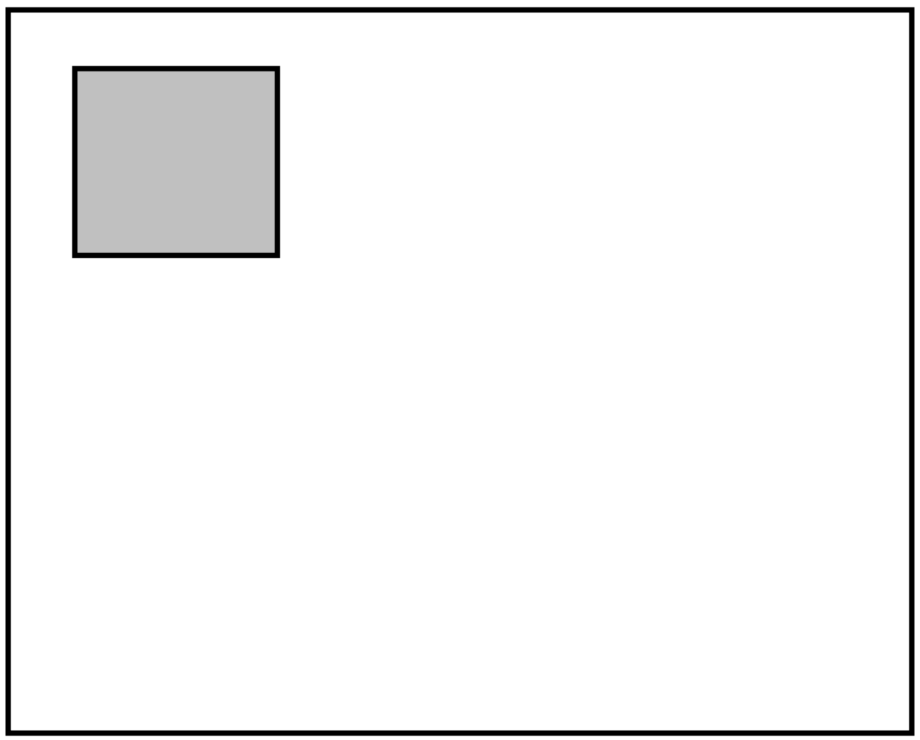
N-ary Productions



ITG Normal Form

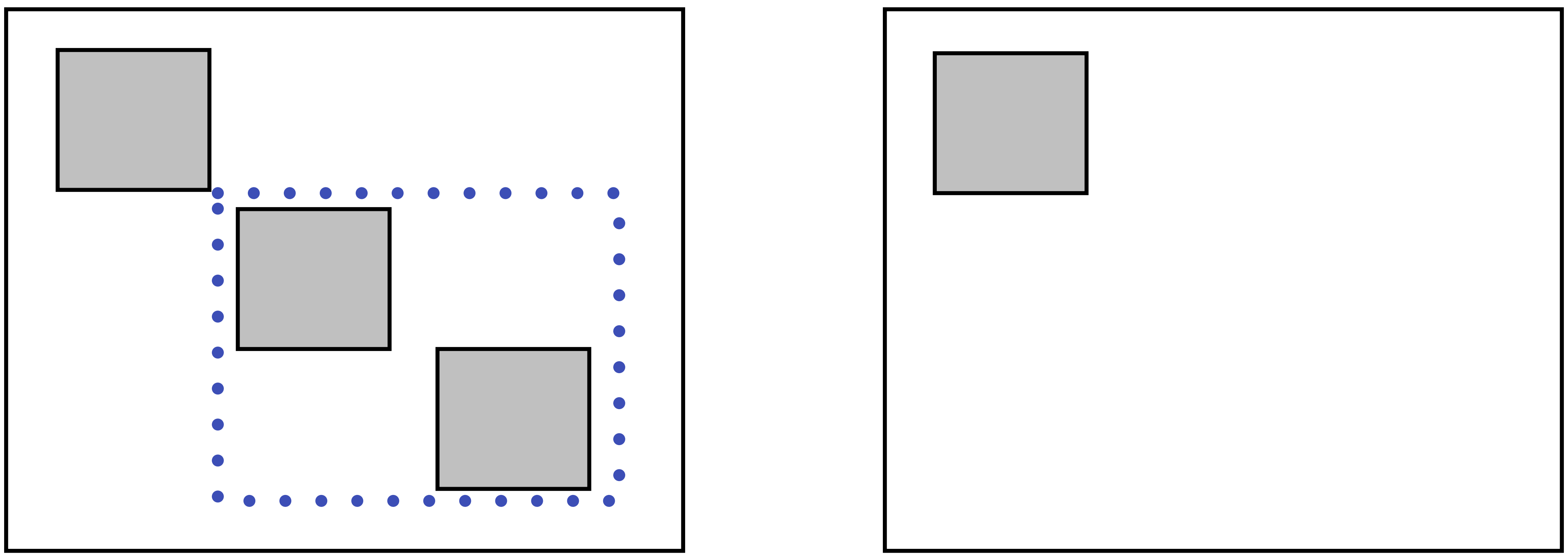
N-ary Productions





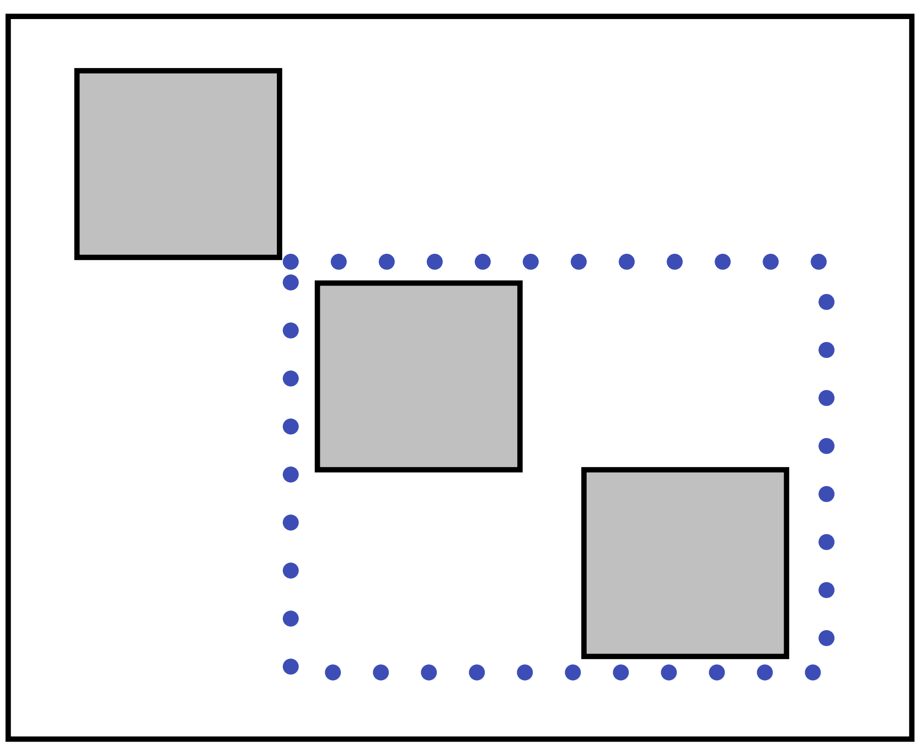
ITG Normal Form

N-ary Productions

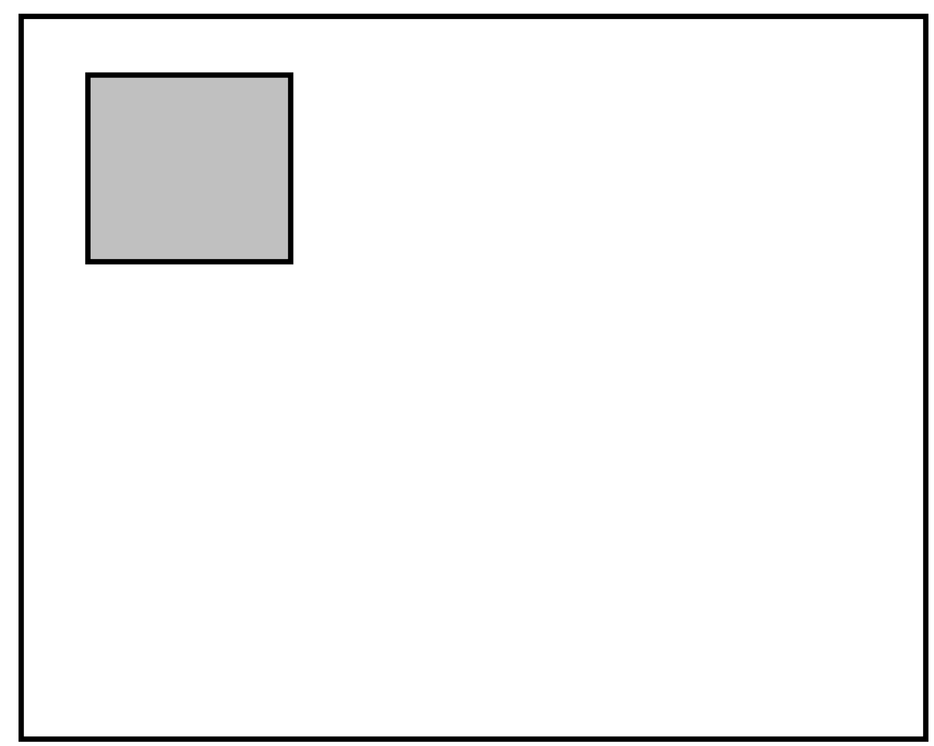


ITG Normal Form

N-ary Productions

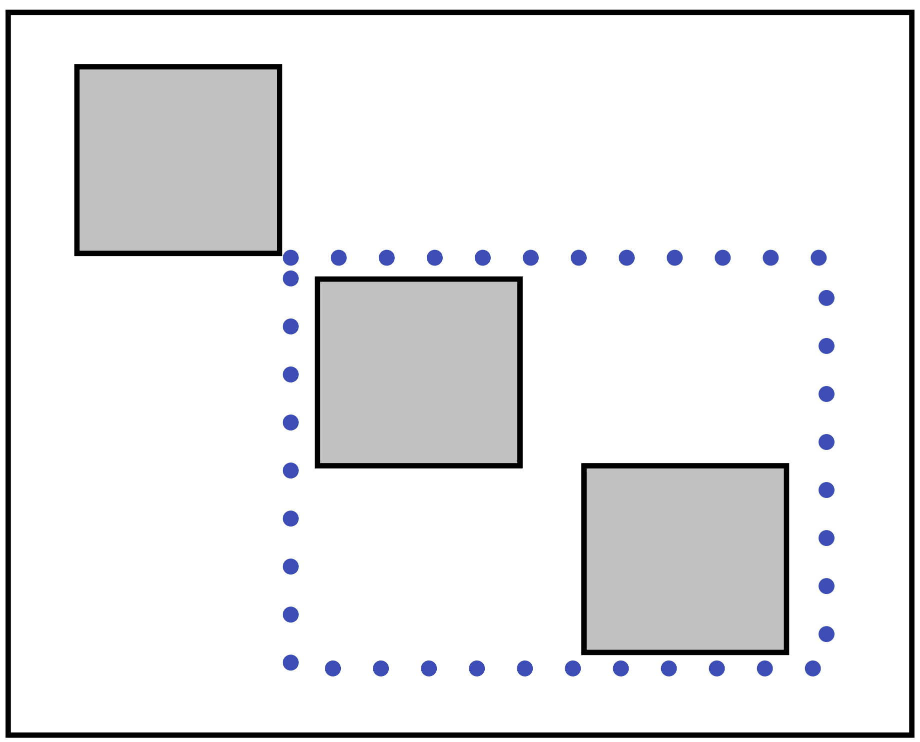


Null Attachment

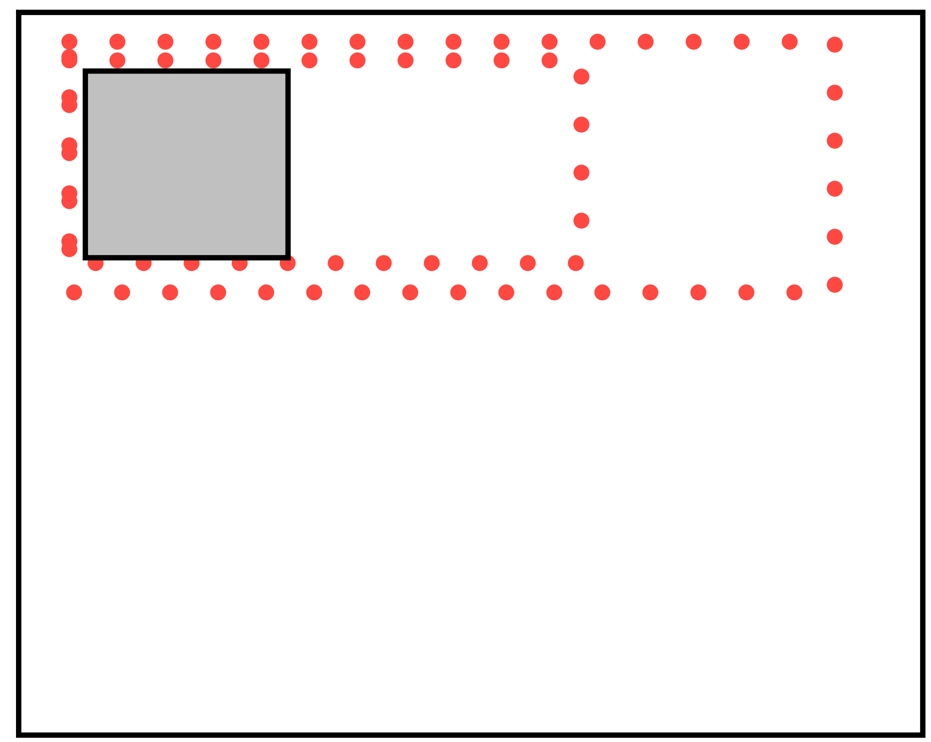


ITG Normal Form

N-ary Productions

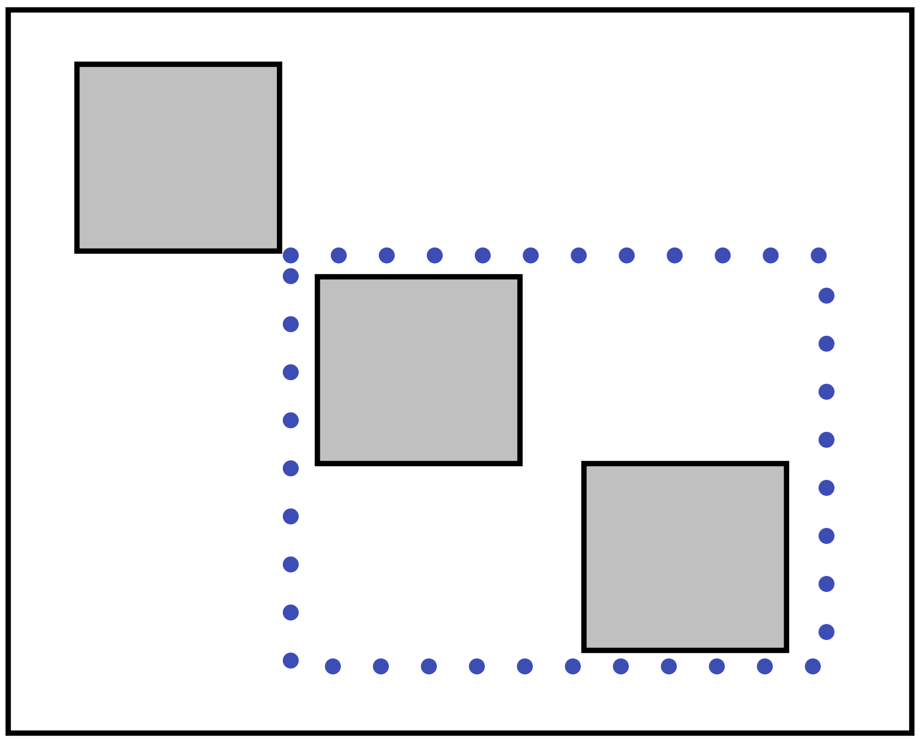


Null Attachment

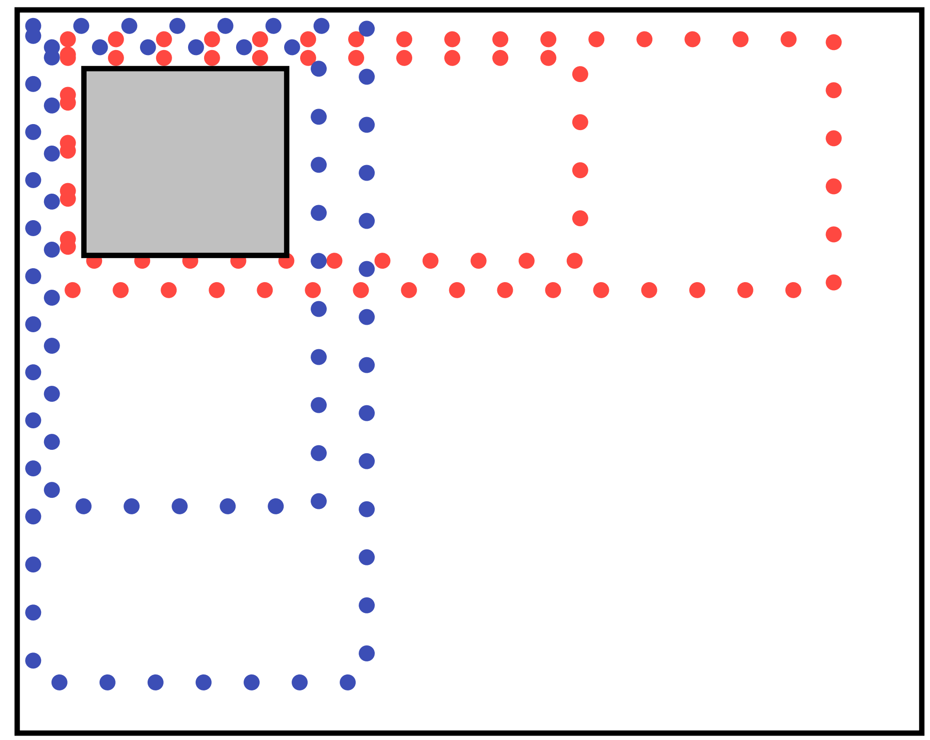


ITG Normal Form

N-ary Productions

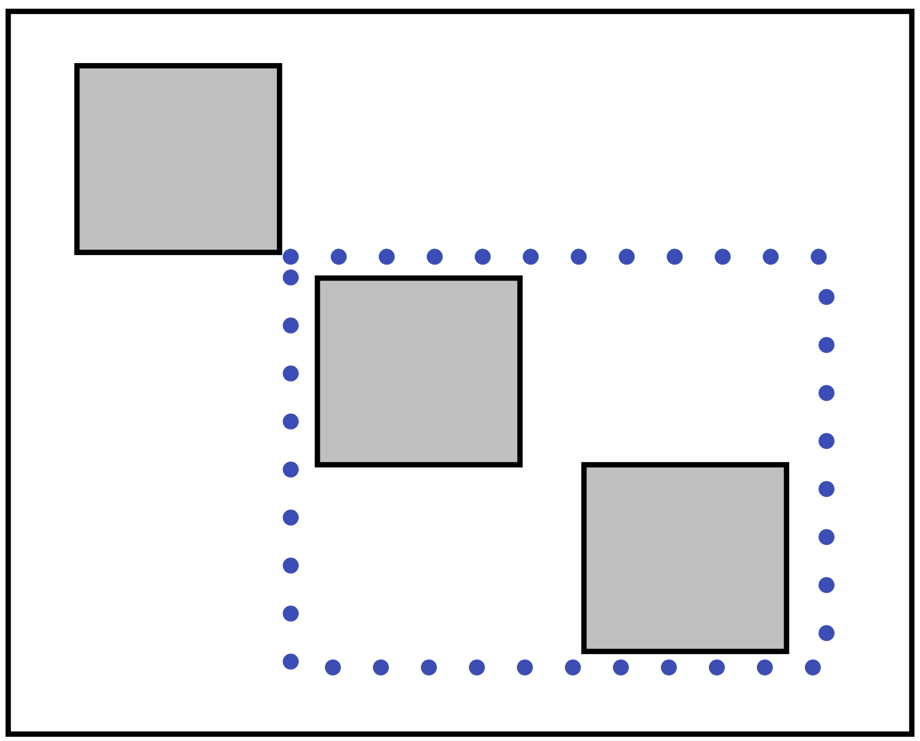


Null Attachment

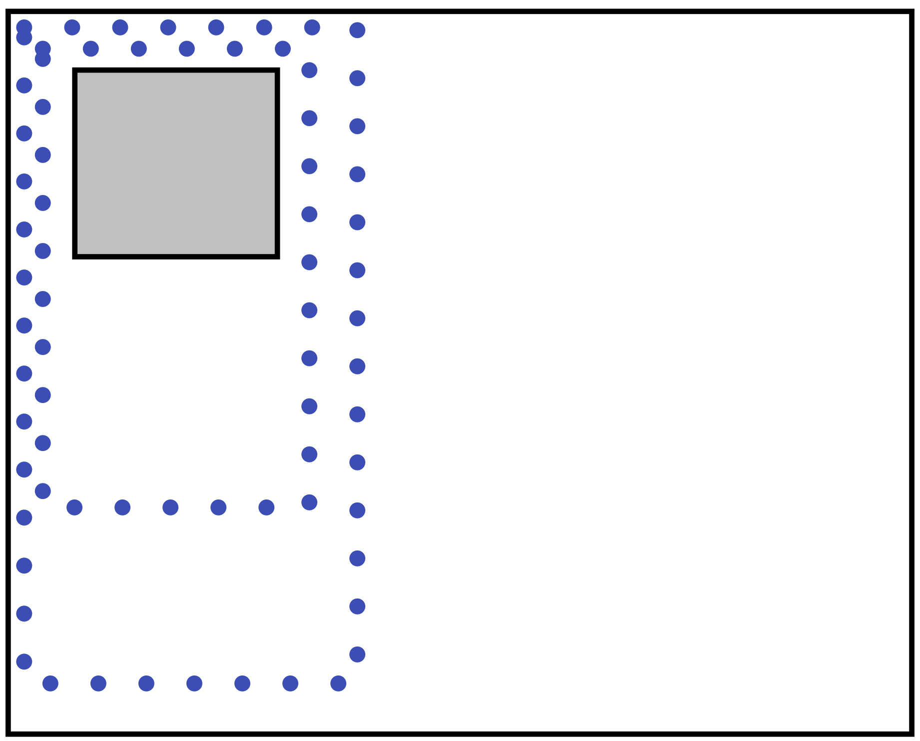


ITG Normal Form

N-ary Productions



Null Attachment



Oracle Projection

Oracle Projection

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

Oracle Projection

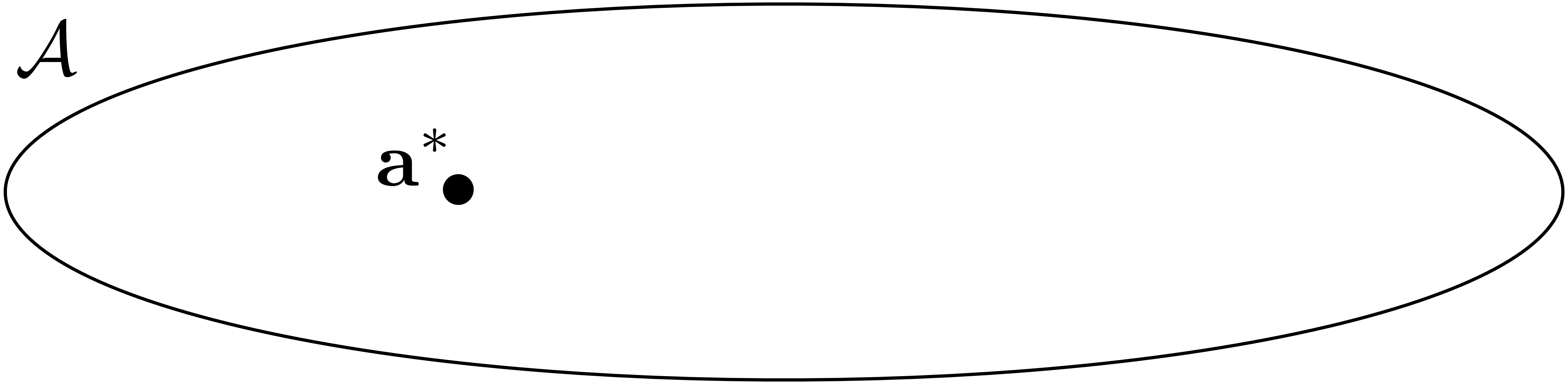
$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$\max_{\mathbf{w}} \sum_i \log P_{\mathbf{w}}(\mathbf{a}_i^* | \mathbf{x}_i)$$

Oracle Projection

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

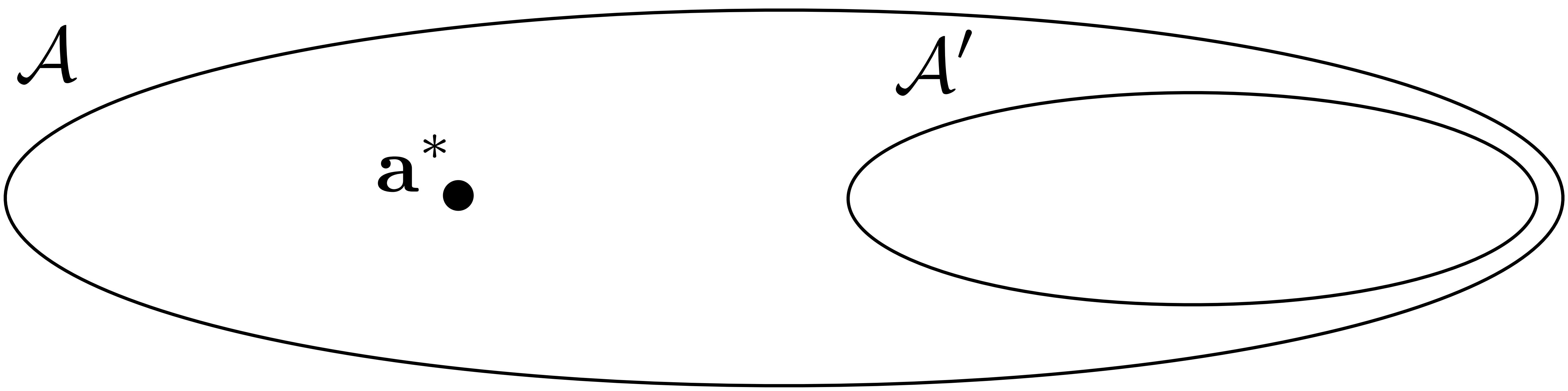
$$\max_{\mathbf{w}} \sum_i \log P_{\mathbf{w}}(\mathbf{a}_i^*|\mathbf{x}_i)$$



Oracle Projection

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

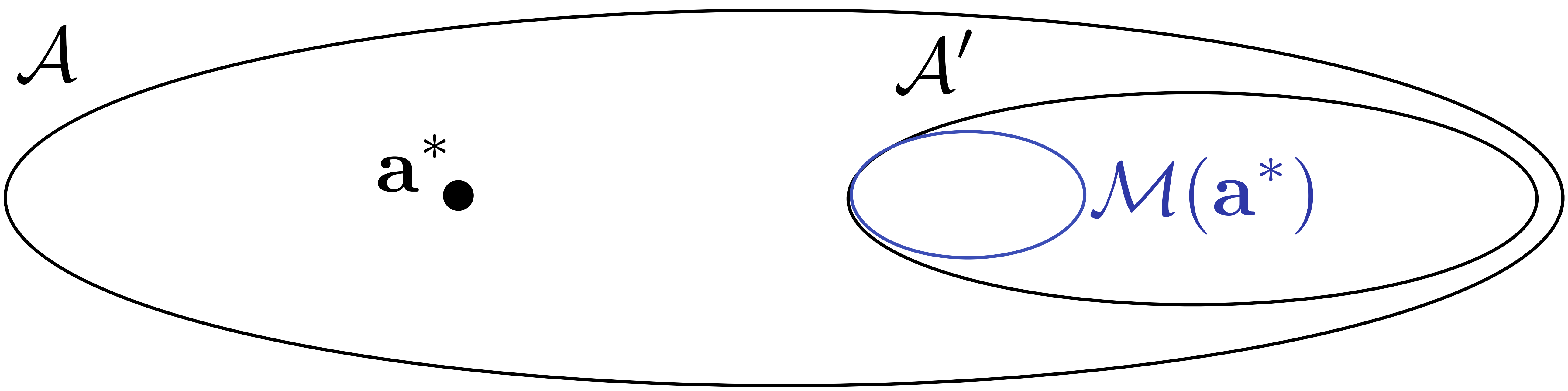
$$\max_{\mathbf{w}} \sum_i \log P_{\mathbf{w}}(\mathbf{a}_i^* | \mathbf{x}_i)$$



Oracle Projection

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

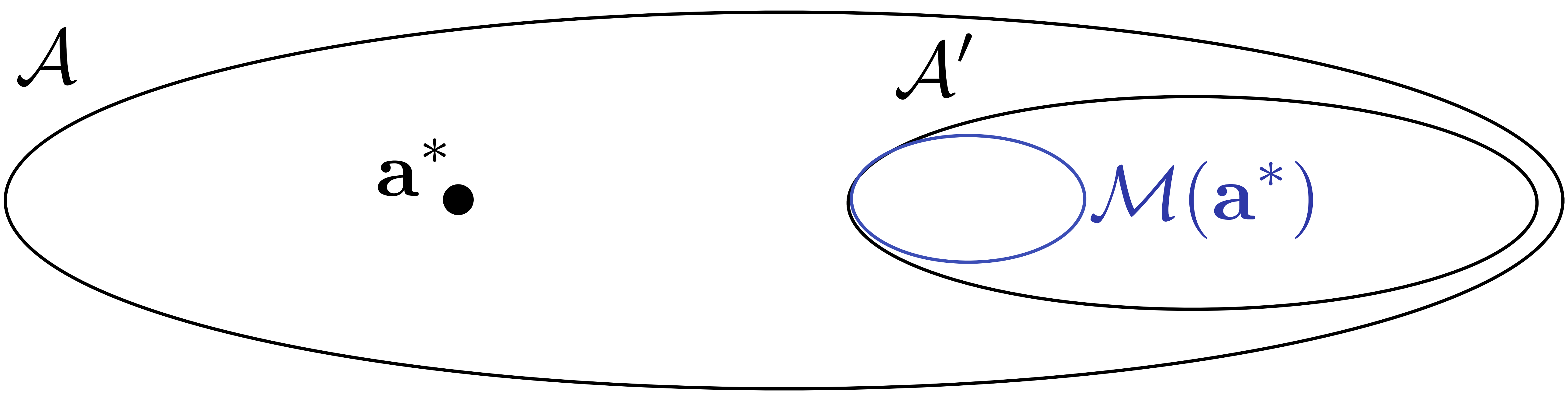
$$\max_{\mathbf{w}} \sum_i \log P_{\mathbf{w}}(\mathbf{a}_i^*|\mathbf{x}_i)$$



Oracle Projection

$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) \propto \exp\{\mathbf{w}^T \phi(\mathbf{a})\}$$

$$\max_{\mathbf{w}} \sum_i \log P_{\mathbf{w}}(\mathcal{M}(\mathbf{a}_i^*)|\mathbf{x}_i)$$



Learning Alignments

Margin

Likelihood

Learning Alignments

Margin

Likelihood

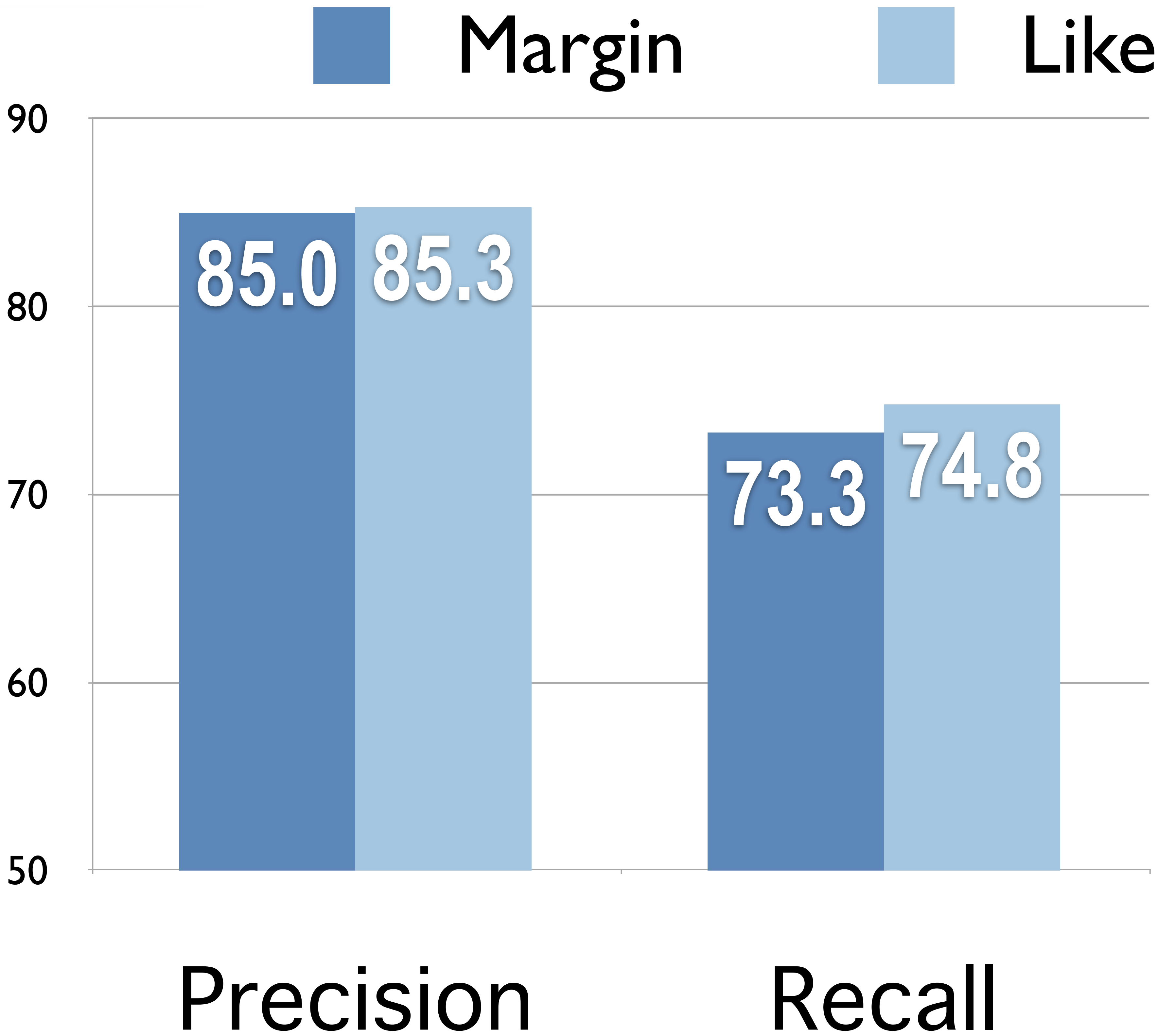
- Viterbi Inference

Learning Alignments

Margin | Likelihood

- Viterbi Inference
- Simple Features
  - Dice
  - Lexical
  - Distance
  - Dictionary

Learning Alignments

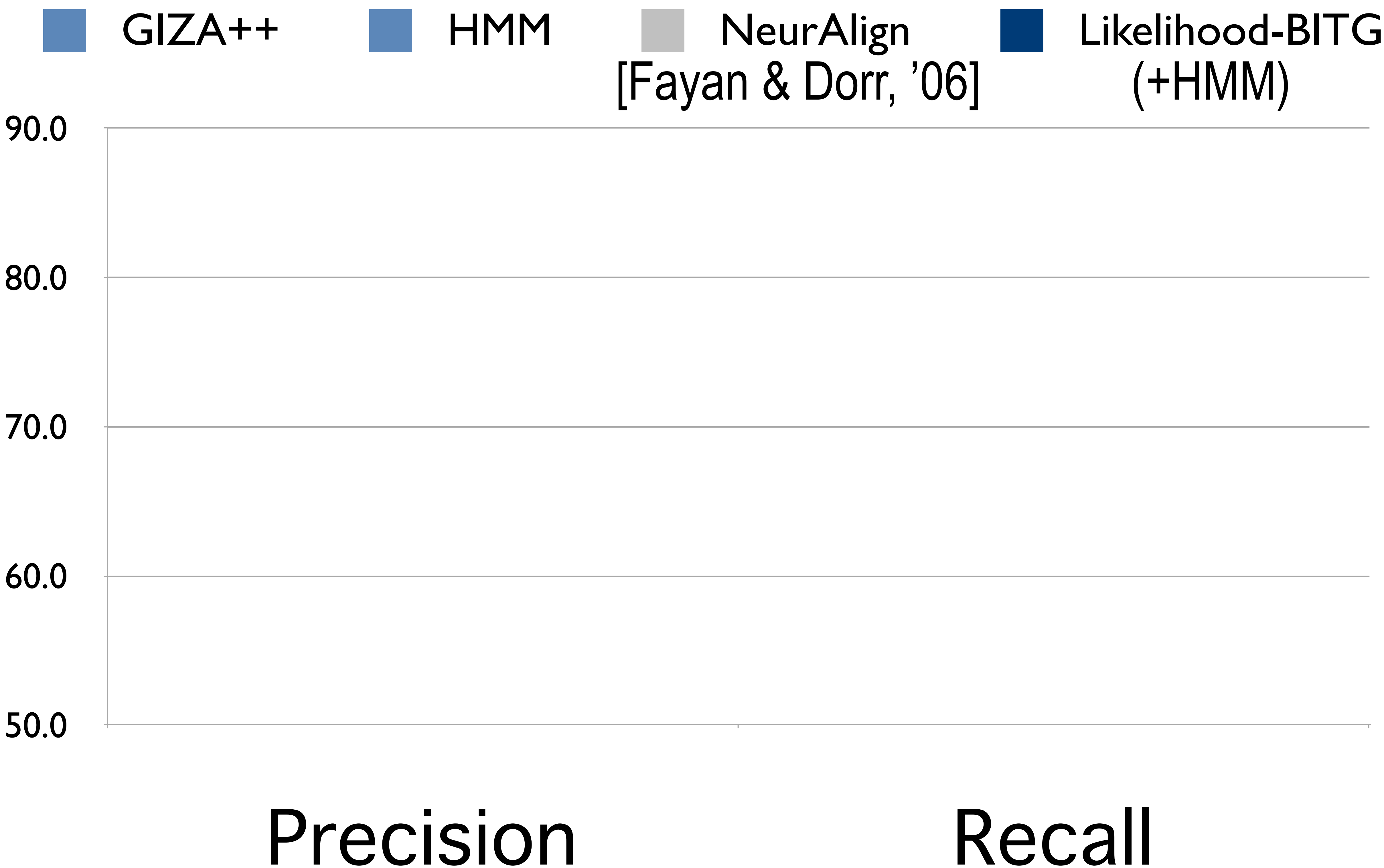


- Viterbi Inference
- Simple Features
  - Dice
  - Lexical
  - Distance
  - Dictionary

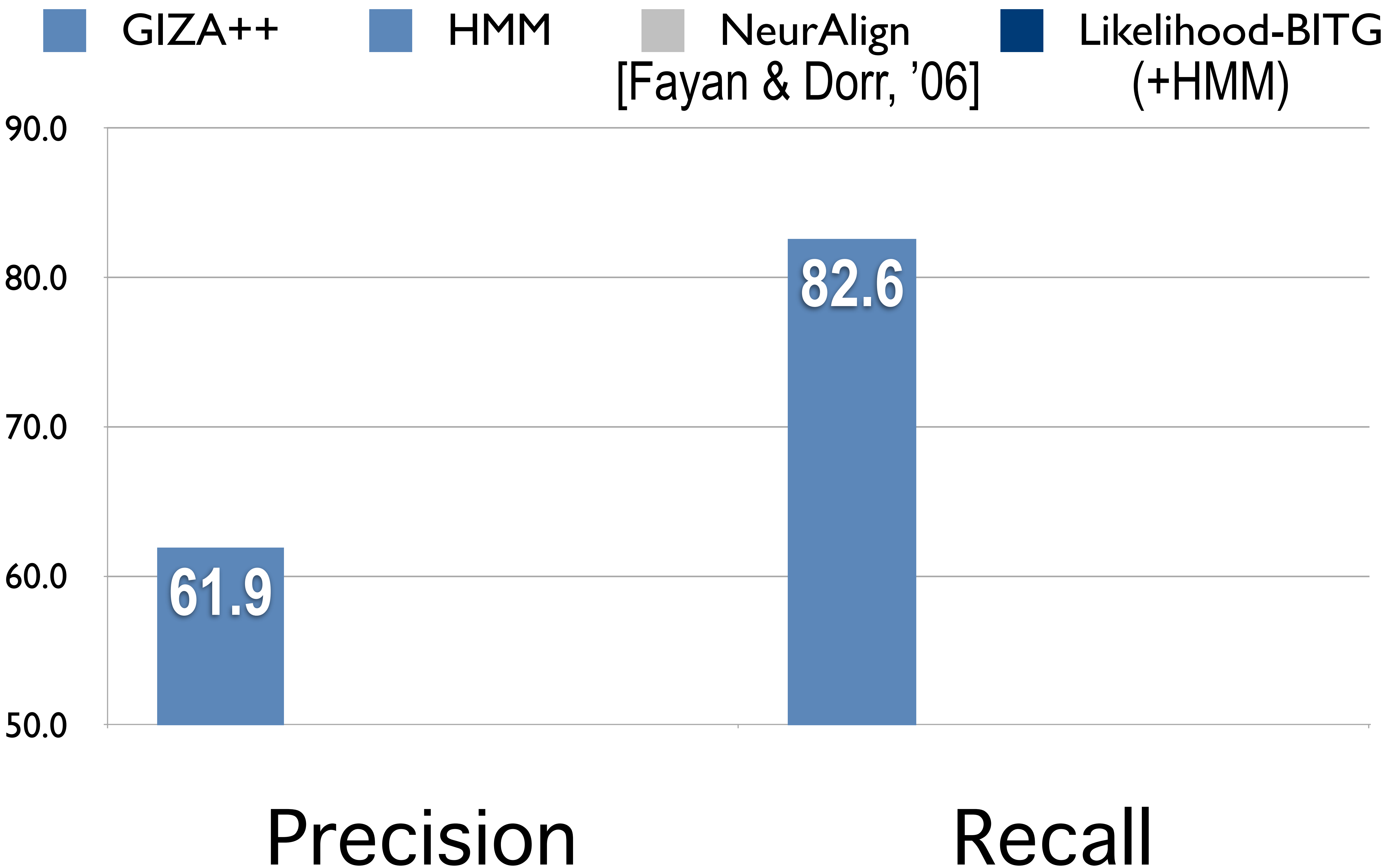
Learning Alignments

GIZA++ HMM NeurAlign [Fayan & Dorr, ’06] Likelihood-BITG (+HMM)

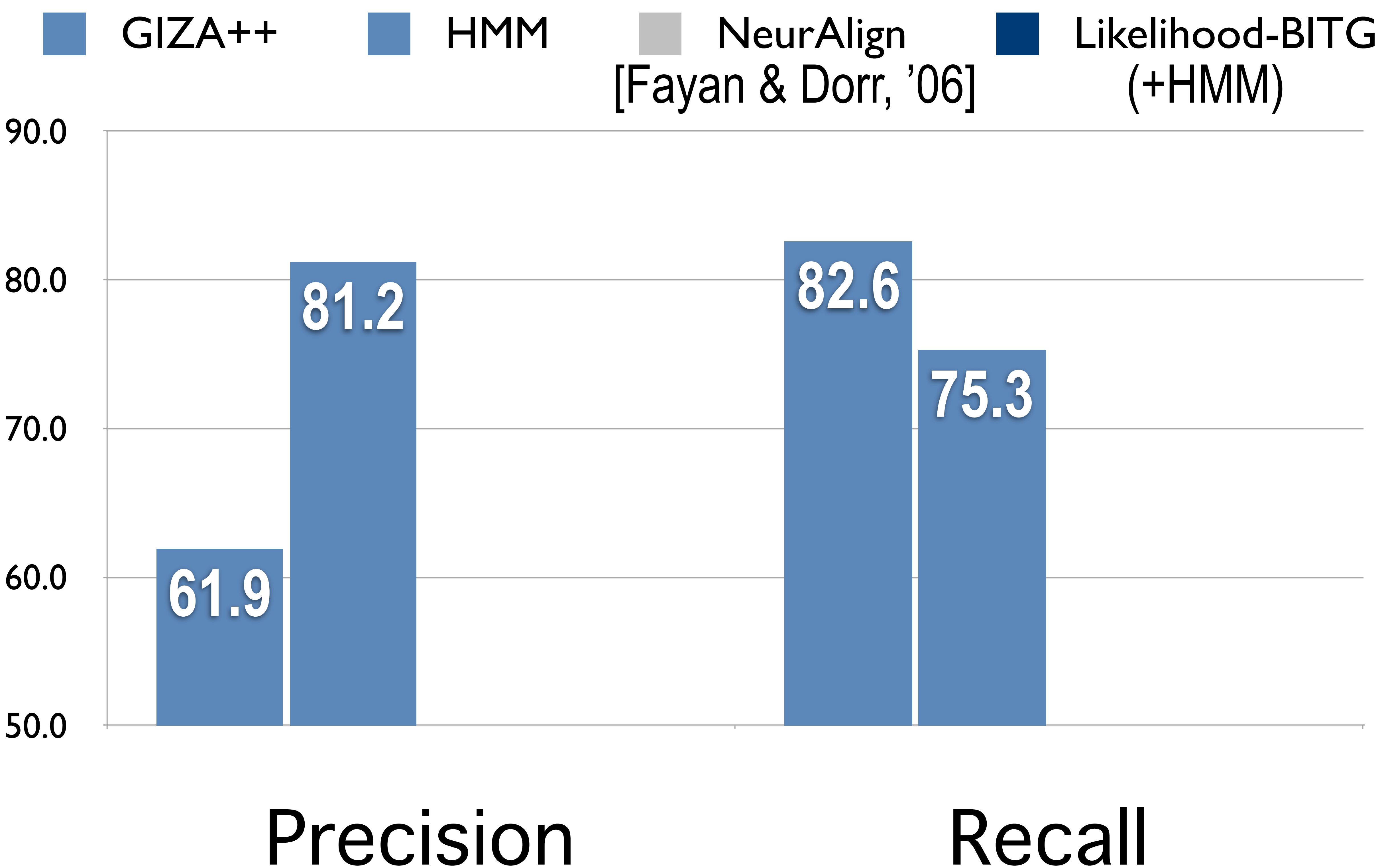
Learning Alignments



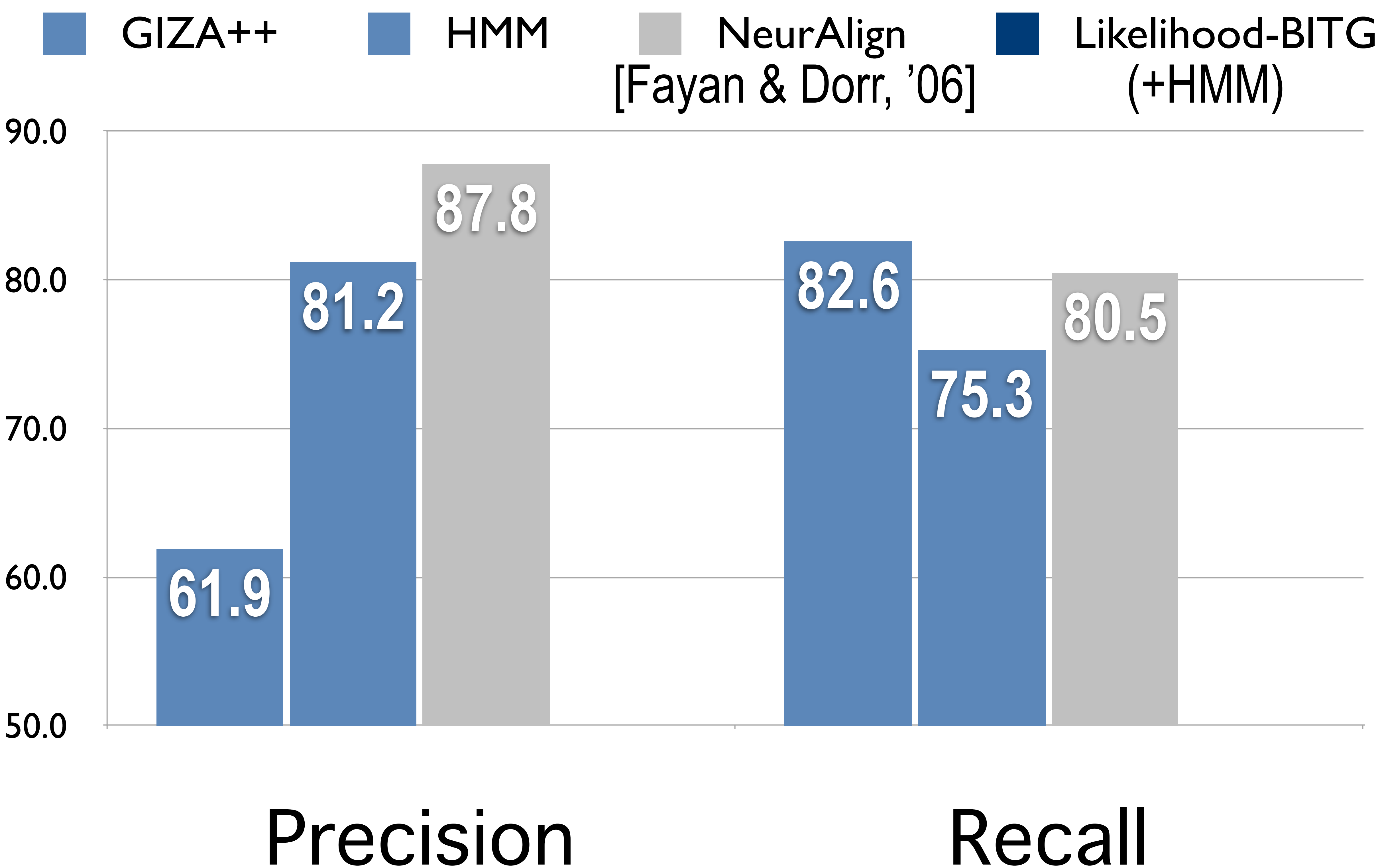
Learning Alignments



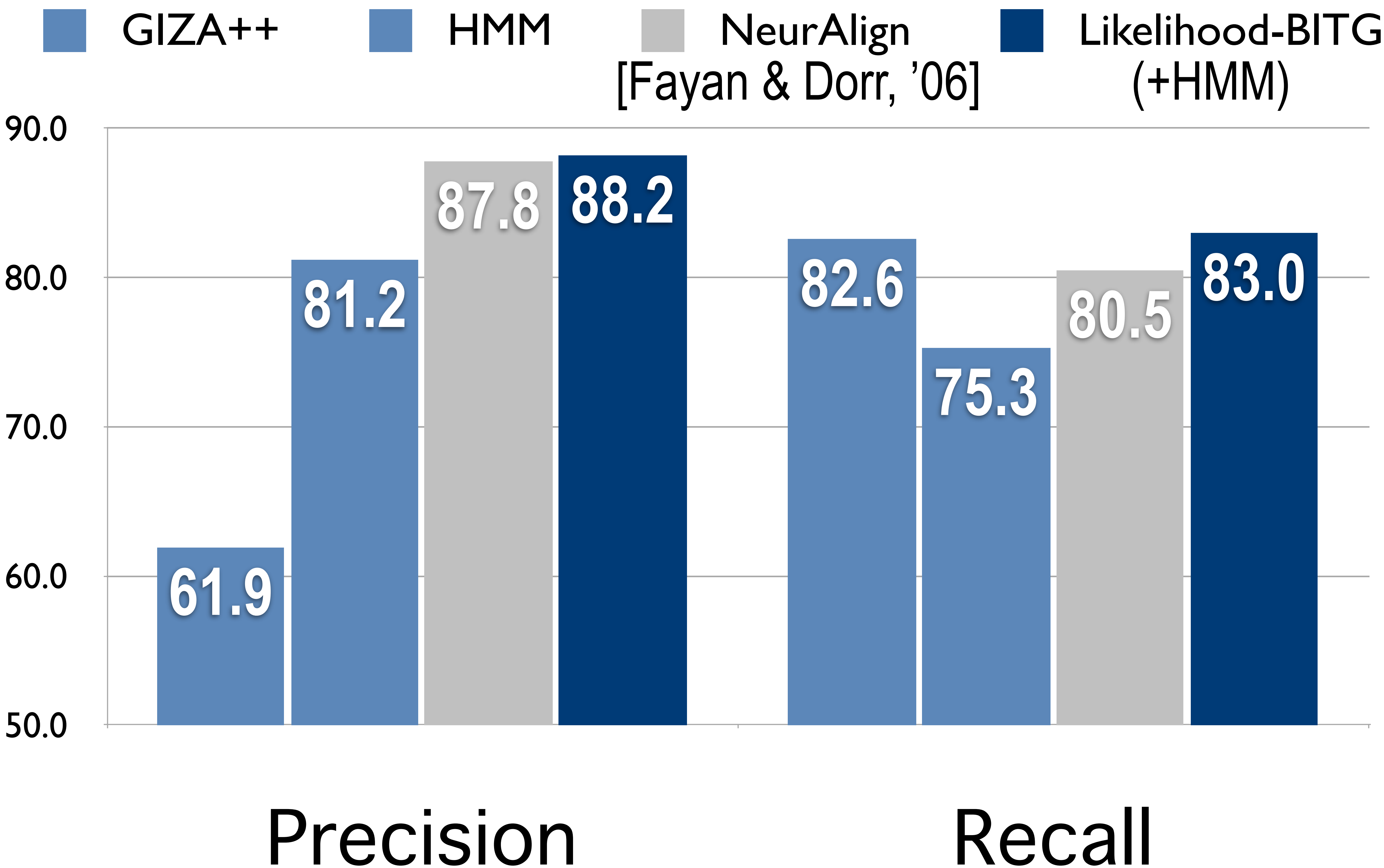
Learning Alignments



Learning Alignments



Learning Alignments



End-to-End Experiments

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]
- Data: FBIS 100k sents max length 40

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]
- Data: FBIS 100k sents max length 40
- LM: 5-gram on Eng. Gigaword Xinhua

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]
- Data: FBIS 100k sents max length 40
- LM: 5-gram on Eng. Gigaword Xinhua
- Tuning: 300 sents. of NIST MT04 test

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]
- Data: FBIS 100k sents max length 40
- LM: 5-gram on Eng. Gigaword Xinhua
- Tuning: 300 sents. of NIST MT04 test
- Test: NIST 2005 Chinese-English

End-to-End Experiments

- Decoder: JosHUa [Li et. al, 2009]
- Data: FBIS 100k sents max length 40
- LM: 5-gram on Eng. Gigaword Xinhua
- Tuning: 300 sents. of NIST MT04 test
- Test: NIST 2005 Chinese-English

Alignments	Recall	Prec	BLEU
GIZA++	**84**	62	23.22
HMM	77	79	23.05
LL-BITG	83	**81**	**24.32**

Conclusions

Conclusions

- Blocks are important, ITG tractable

Conclusions

- Blocks are important, ITG tractable
- Normal form and oracle projection allow for likelihood training

Conclusions

- Blocks are important, ITG tractable
- Normal form and oracle projection allow for likelihood training
- Word alignment improvements yield BLEU improvements

Conclusions

- Blocks are important, ITG tractable
- Normal form and oracle projection allow for likelihood training
- Word alignment improvements yield BLEU improvements

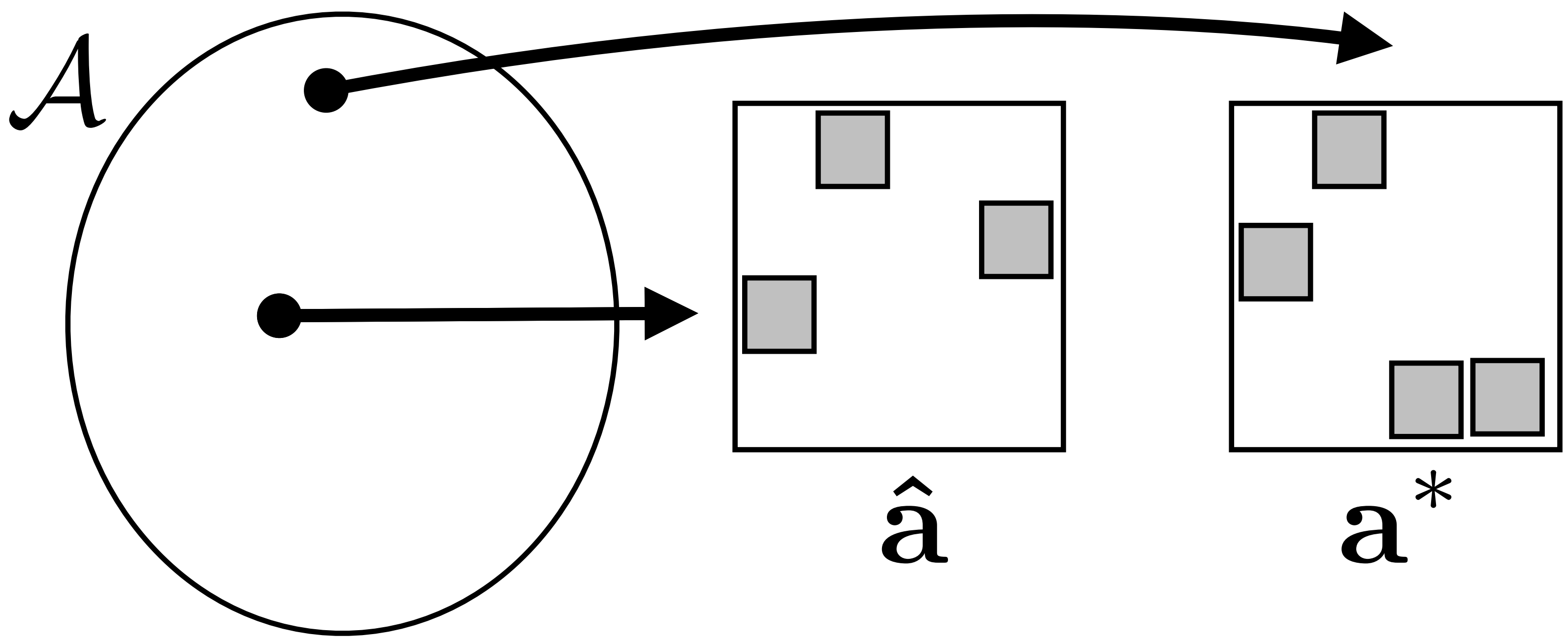
Conclusions

- Blocks are important, ITG tractable
- Normal form and oracle projection allow for likelihood training
- Word alignment improvements yield BLEU improvements
- Software available @ nlp.cs.berkeley.edu

Thanks!

http://nlp.cs.berkeley.edu

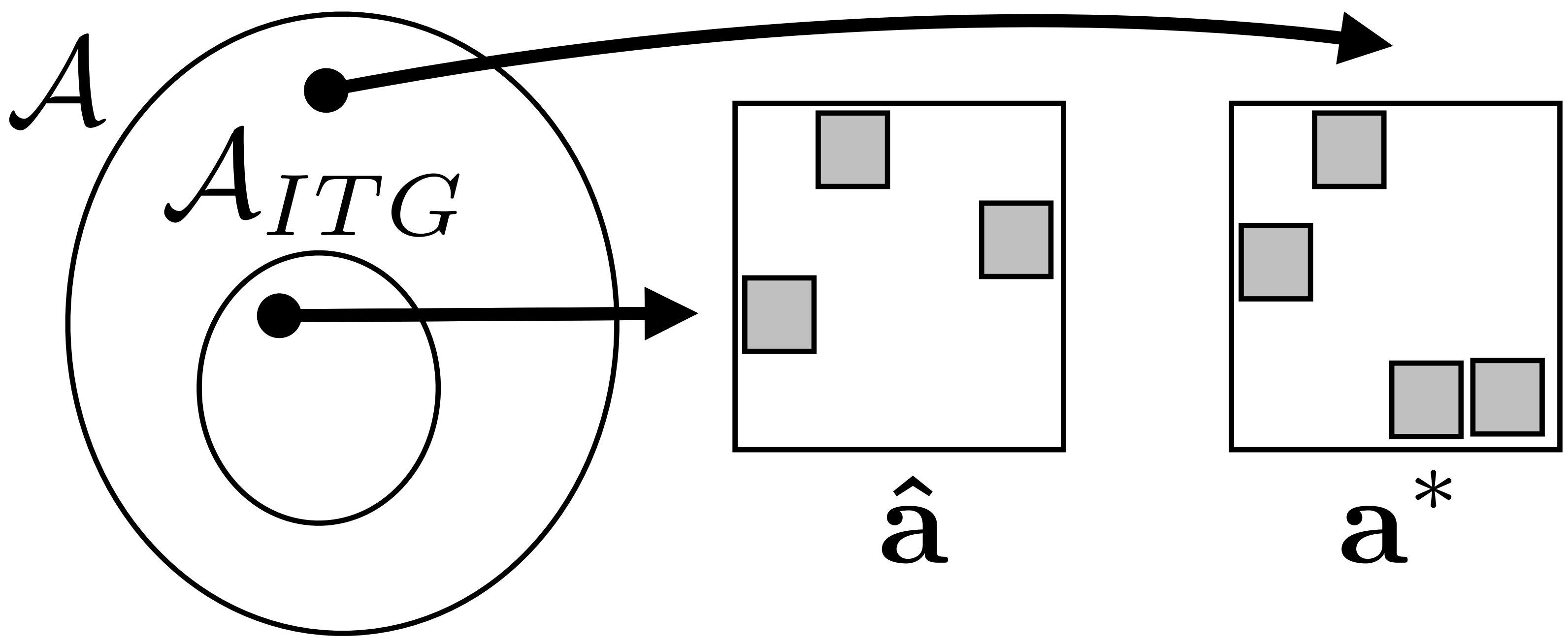
Likelihood Criterion



$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) = \frac{s_{\mathbf{w}}(\mathbf{a})}{\sum_{\mathbf{a}' \in \mathcal{A}} s_{\mathbf{w}}(\mathbf{a}')}$$

$$\max_{\mathbf{w}} \sum_{(x,\mathbf{a}^*) \in \mathcal{D}} \log P(\mathbf{a}^*|\mathbf{x})$$

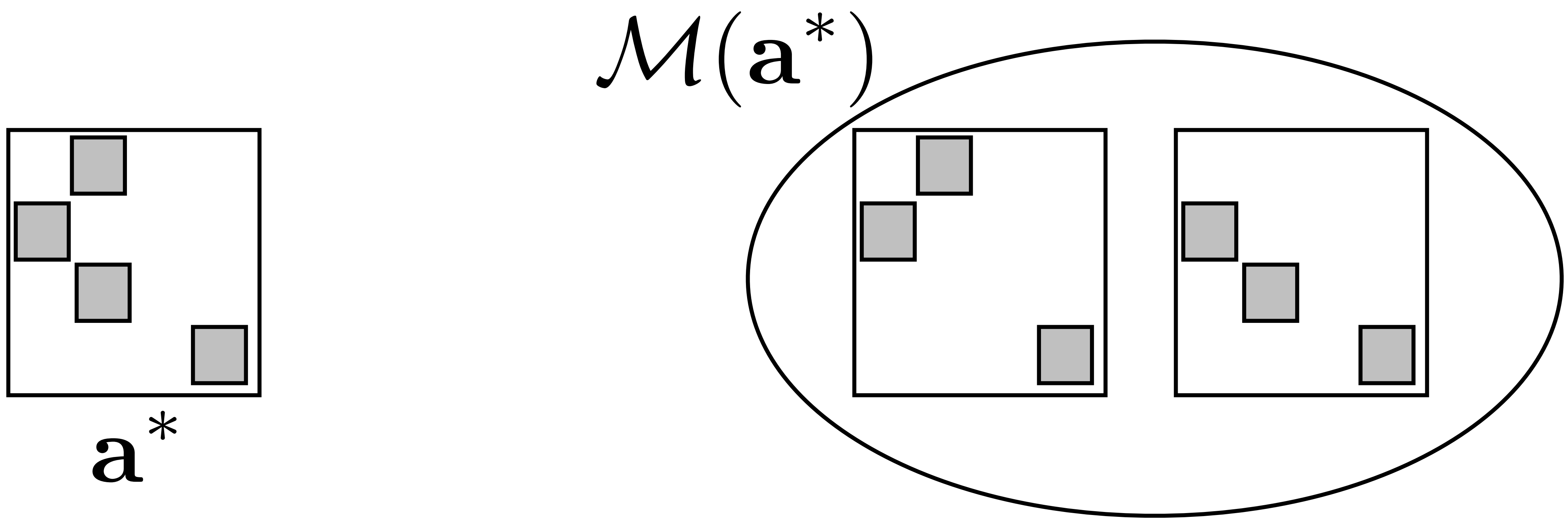
Likelihood Criterion



$$P_{\mathbf{w}}(\mathbf{a}|\mathbf{x}) = \frac{s_{\mathbf{w}}(\mathbf{a})}{\sum_{\mathbf{a}' \in \mathcal{A}_{ITG}} s_{\mathbf{w}}(\mathbf{a}')}$$

$$\max_{\mathbf{w}} \sum_{(x,\mathbf{a}^*) \in \mathcal{D}} \log P(\mathbf{a}^*|\mathbf{x})$$

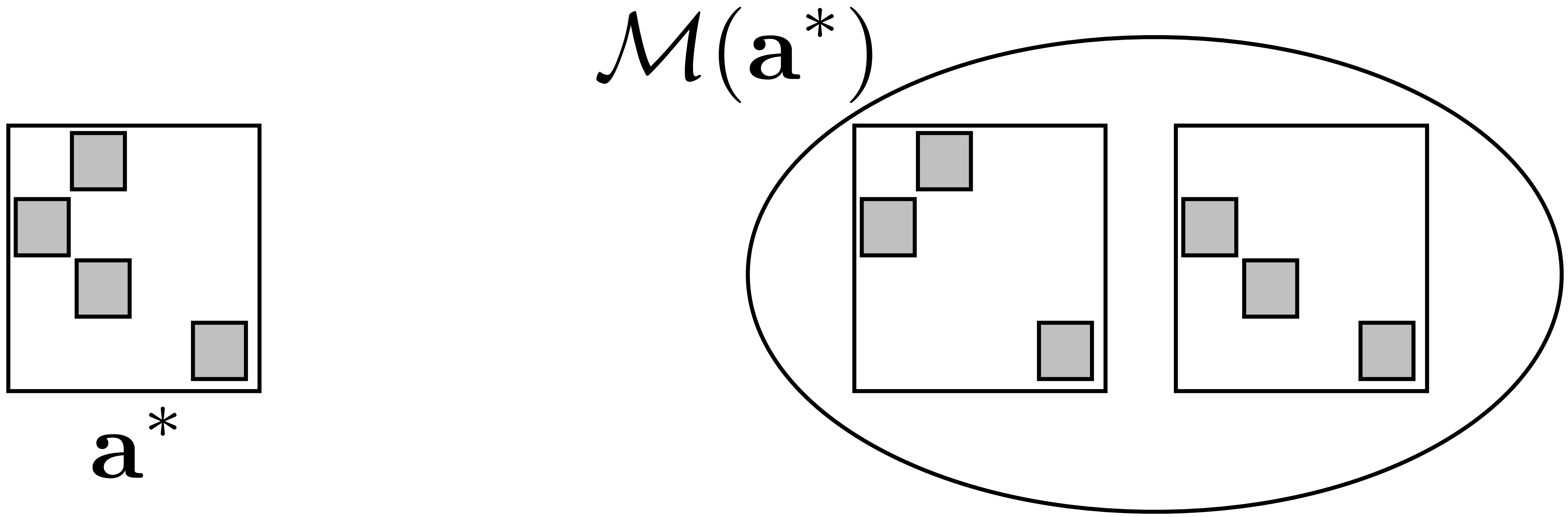
Likelihood Criterion



$$m^* = \min_{\mathbf{a} \in \mathcal{A}_{ITG}} L(\mathbf{a}^*, \mathbf{a})$$

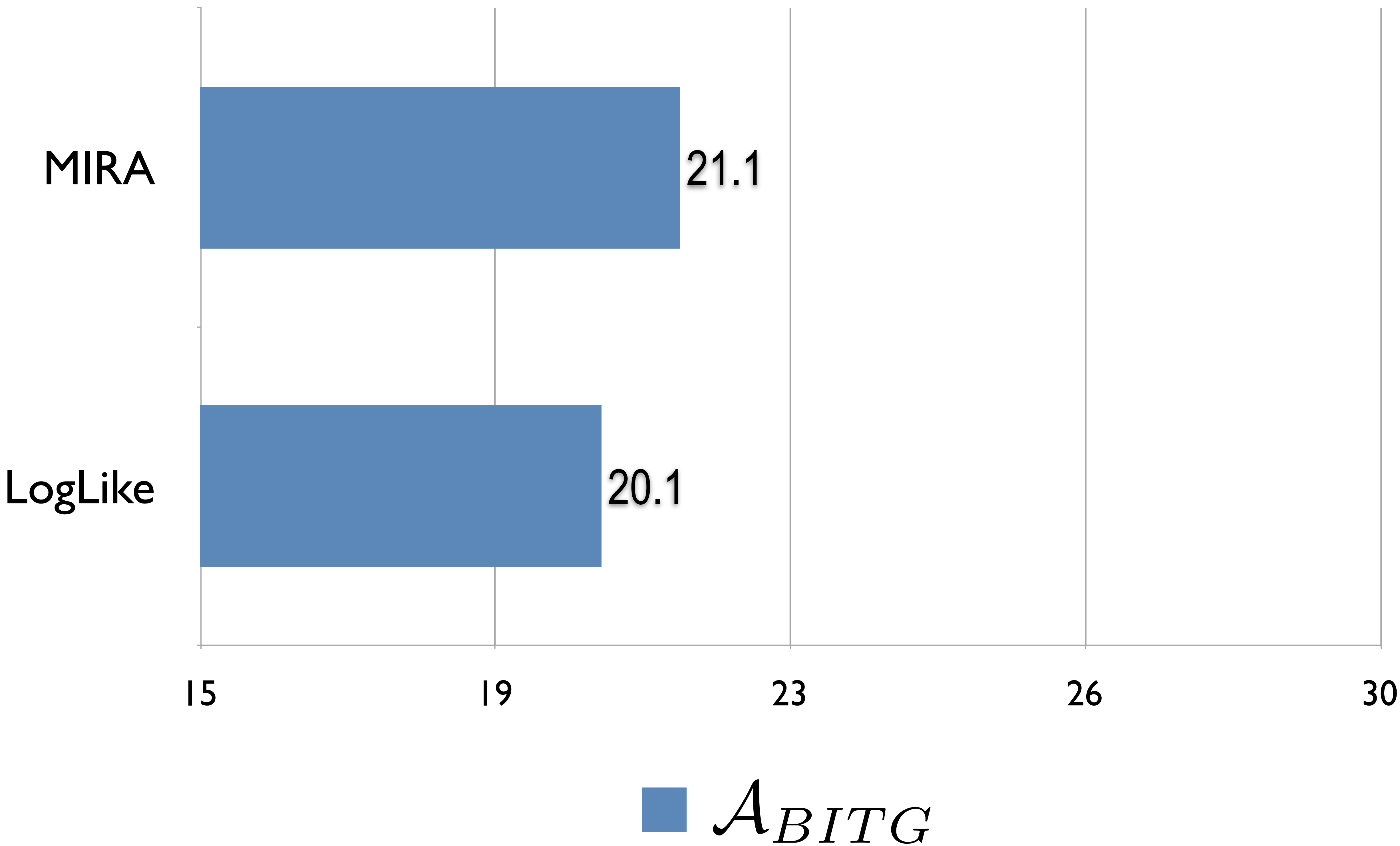
$$\mathcal{M}(\mathbf{a}^*) = \{\mathbf{a} \in \mathcal{A}_{ITG} \text{ s.t. } L(\mathbf{a}^*, \mathbf{a}) = m^*\}$$

Likelihood Criterion



$$\max_{\mathbf{w}} \sum_{(x,\mathbf{a}^*)\in\mathcal{D}} \log P(\mathcal{M}(\mathbf{a}^*)|\mathbf{x})$$

Likelihood Criterion



Adding External Features

Adding Joint HMM posteriors from DeNero et. al. (2007)

	MIRA									Likelihood					
	1-1			ITG			BITG			BITG-S			BITG-N		
Features	P	R	AER	P	R	AER	P	R	AER	P	R	AER	P	R	AER
Dice, dist, blcks, dict, lex	85.7	63.7	26.8	**86.2**	65.8	25.2	85.0	73.3	21.1	85.7	73.7	20.6	85.3	**74.8**	**20.1**
+HMM	90.5	69.4	21.2	**91.2**	70.1	20.3	90.2	80.1	15.0	87.3	82.8	14.9	88.2	**83.0**	**14.4**

TODO: Keynote Chart

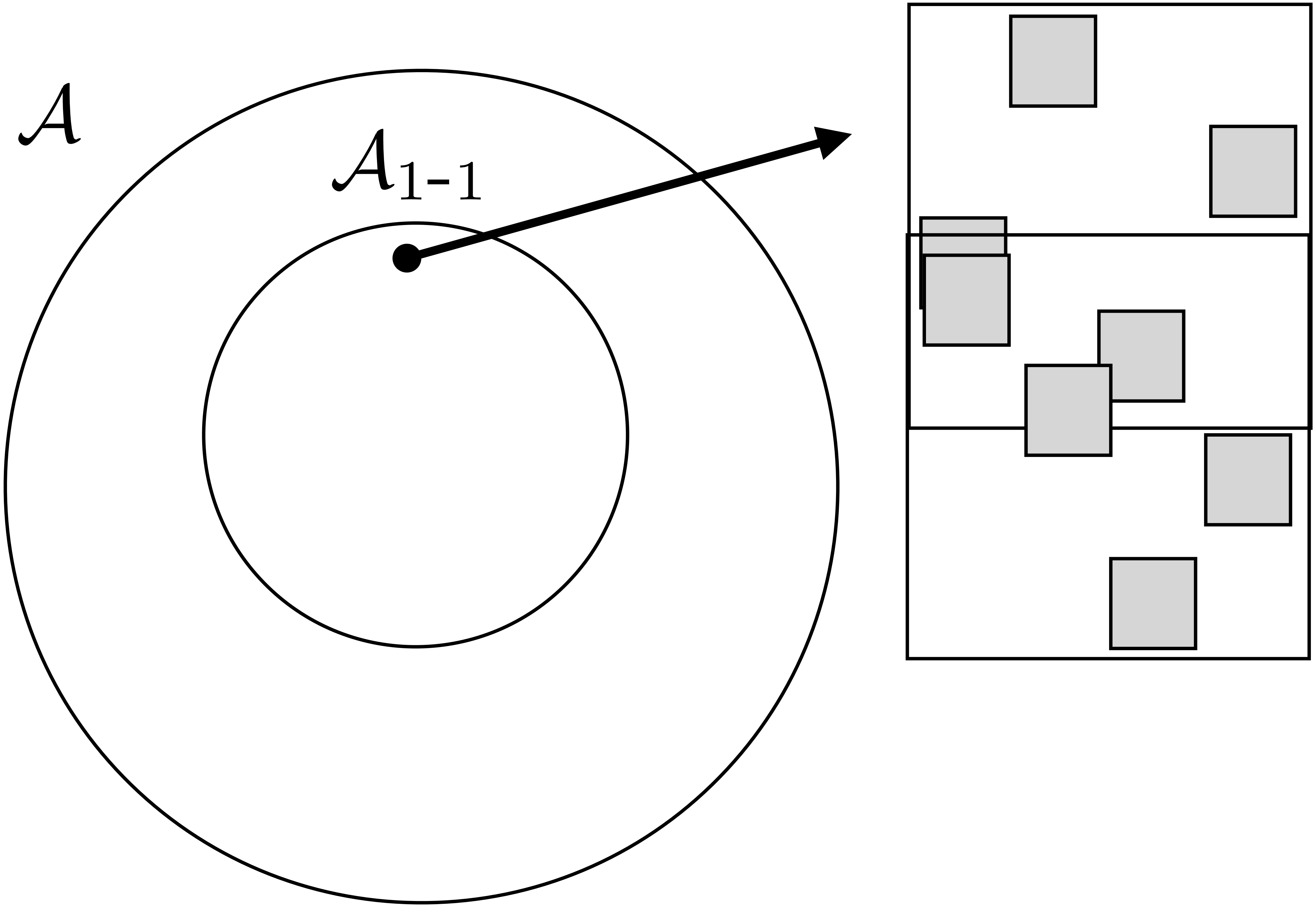
End-to-End Results

Using JosHUa decoder (HIERO)

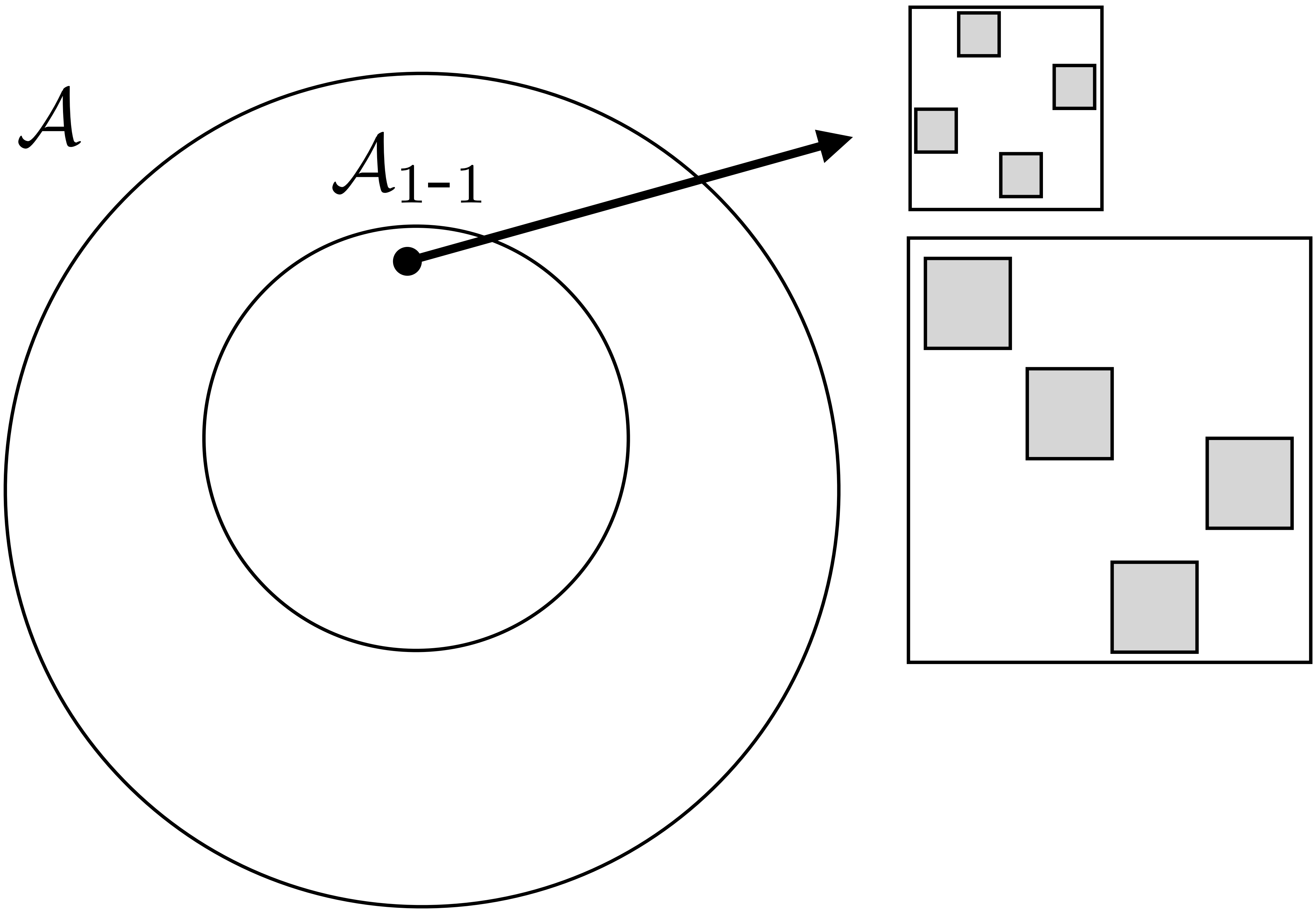
Alignments			Translations	
Model	Prec	Rec	Rules	BLEU
GIZA++	62	84	1.9M	23.22
Joint HMM	79	77	4.0M	23.05
Viterbi ITG	90	80	3.8M	24.28
Posterior ITG	81	83	4.2M	**24.32**

TODO: Keynote Chart

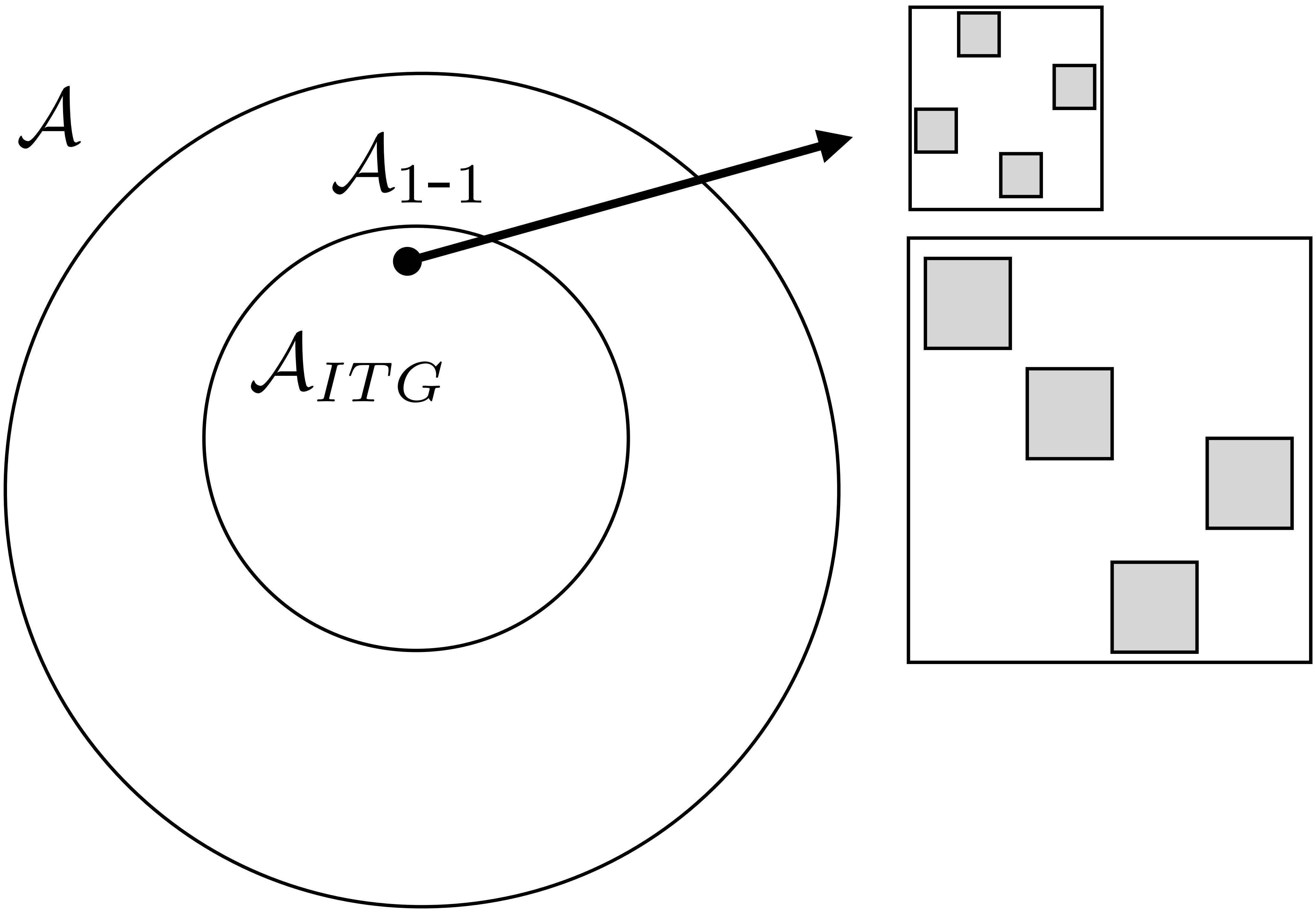
Alignment Families



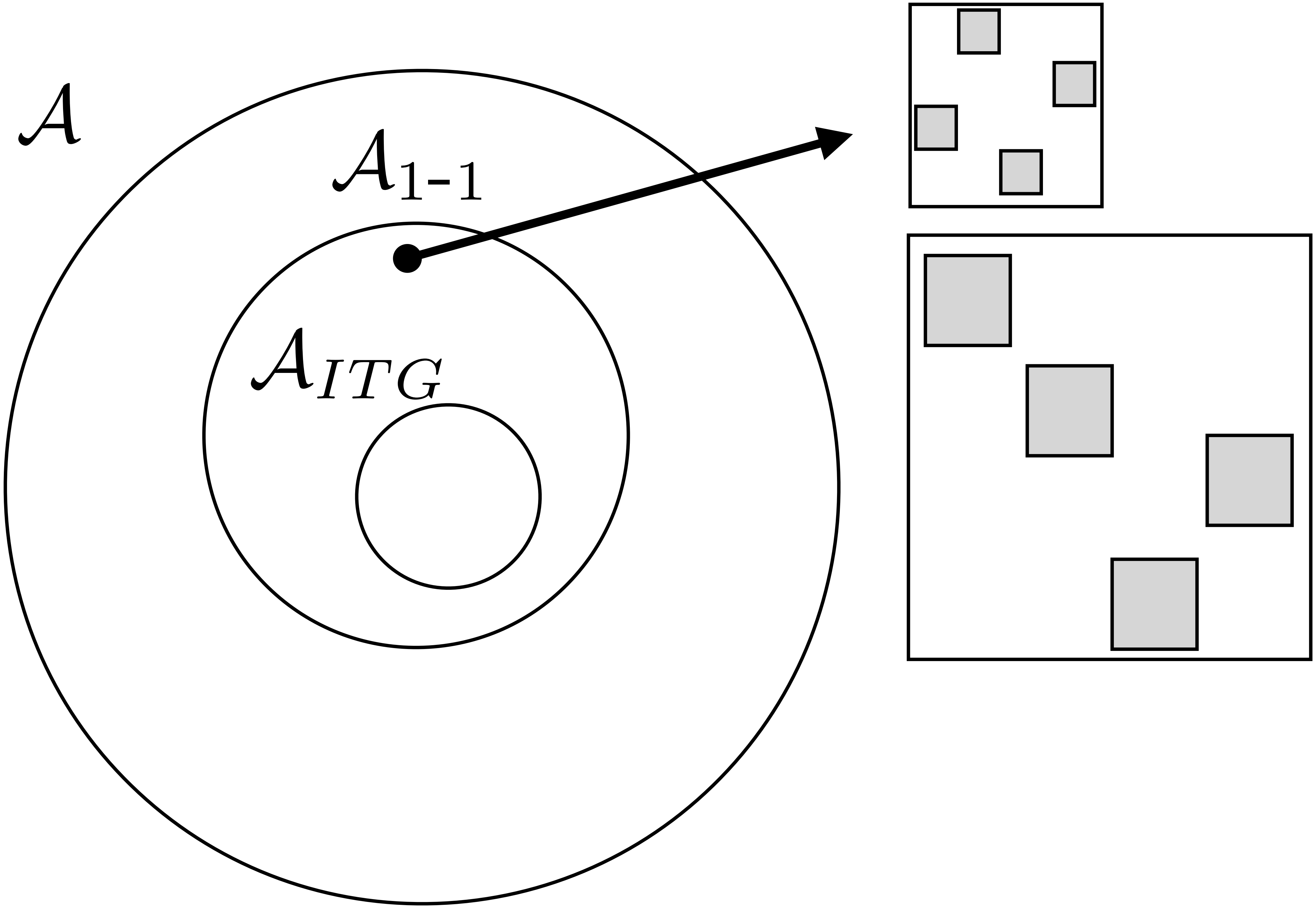
Alignment Families



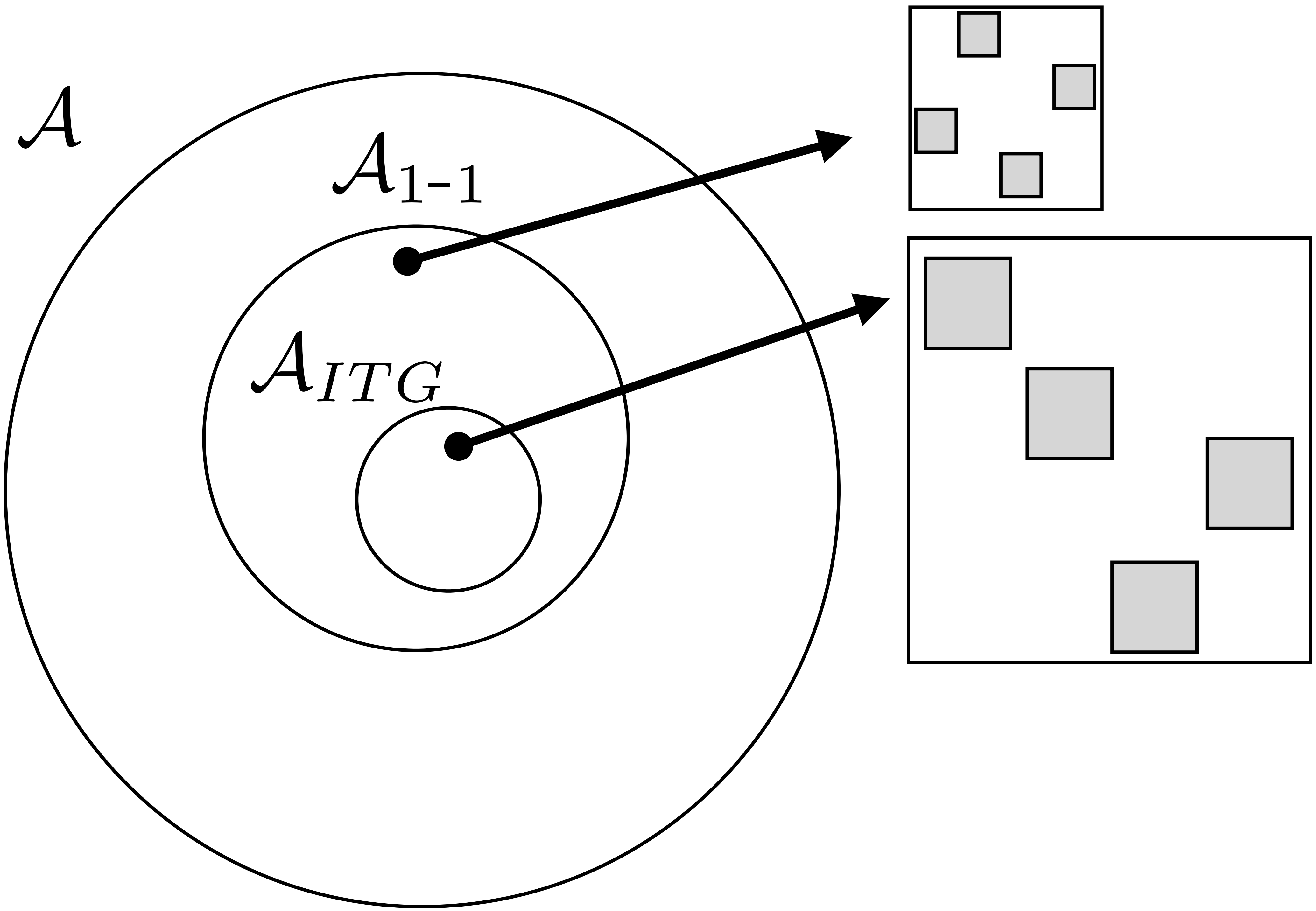
Alignment Families



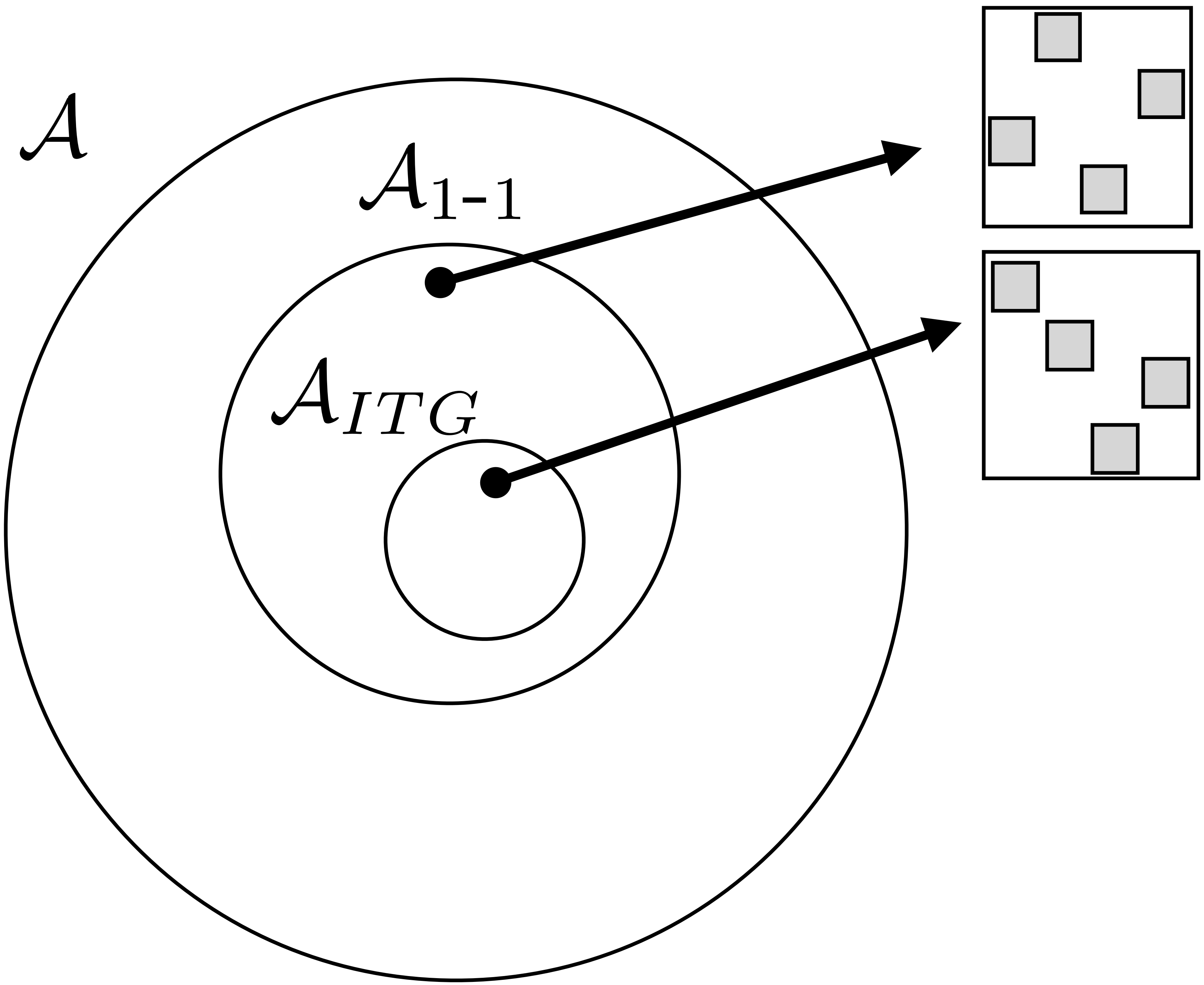
Alignment Families



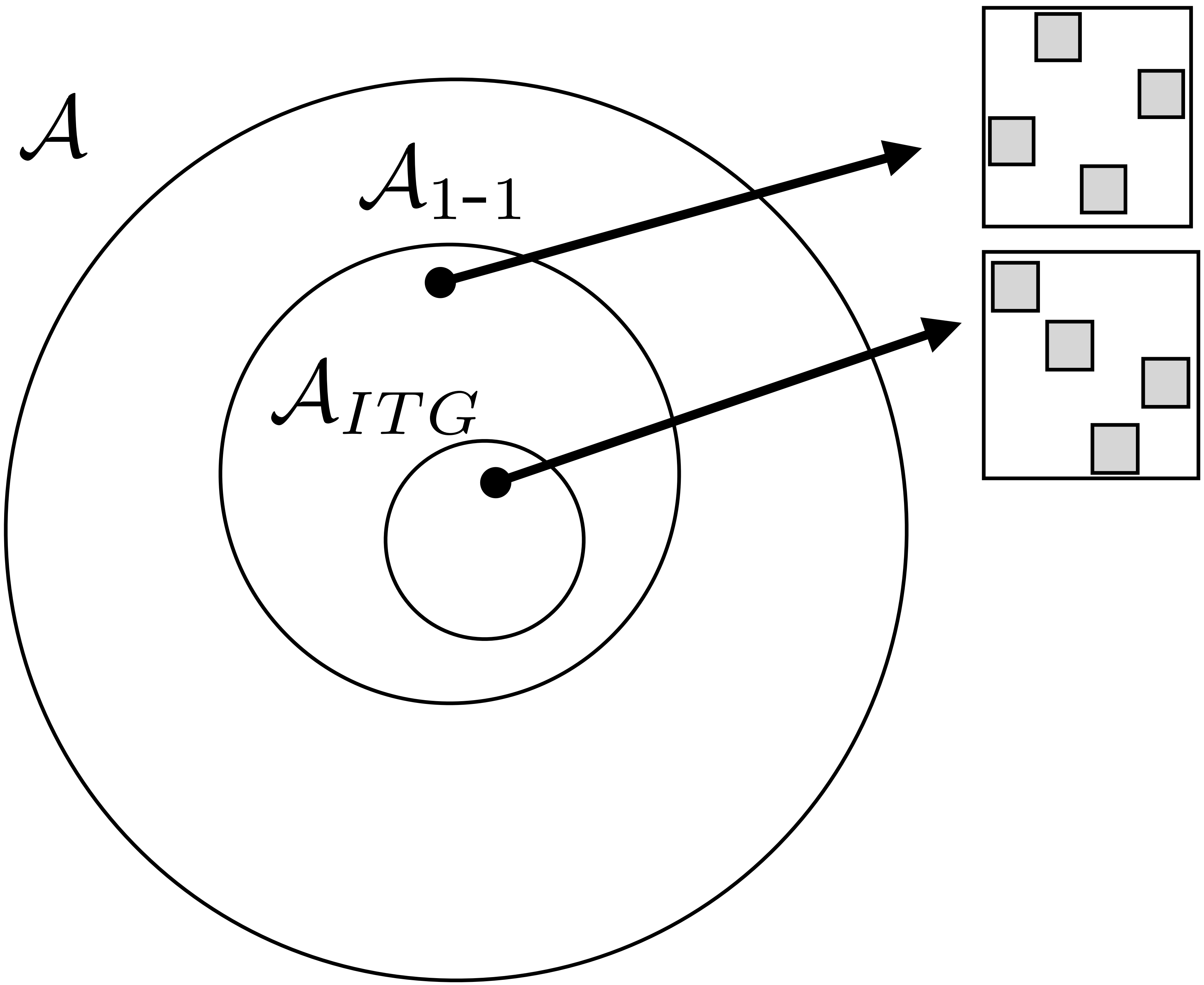
Alignment Families



Alignment Families

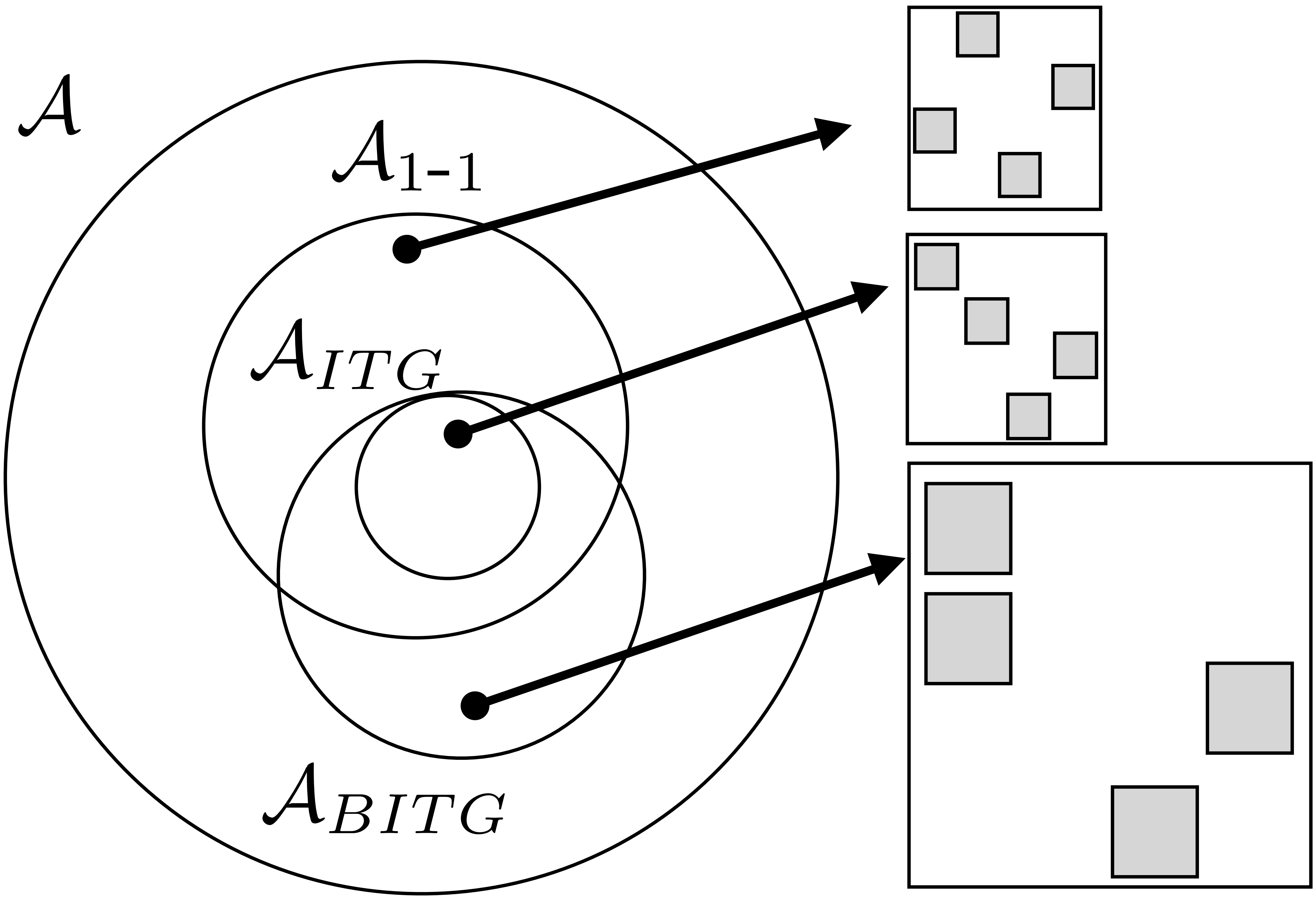


Alignment Families

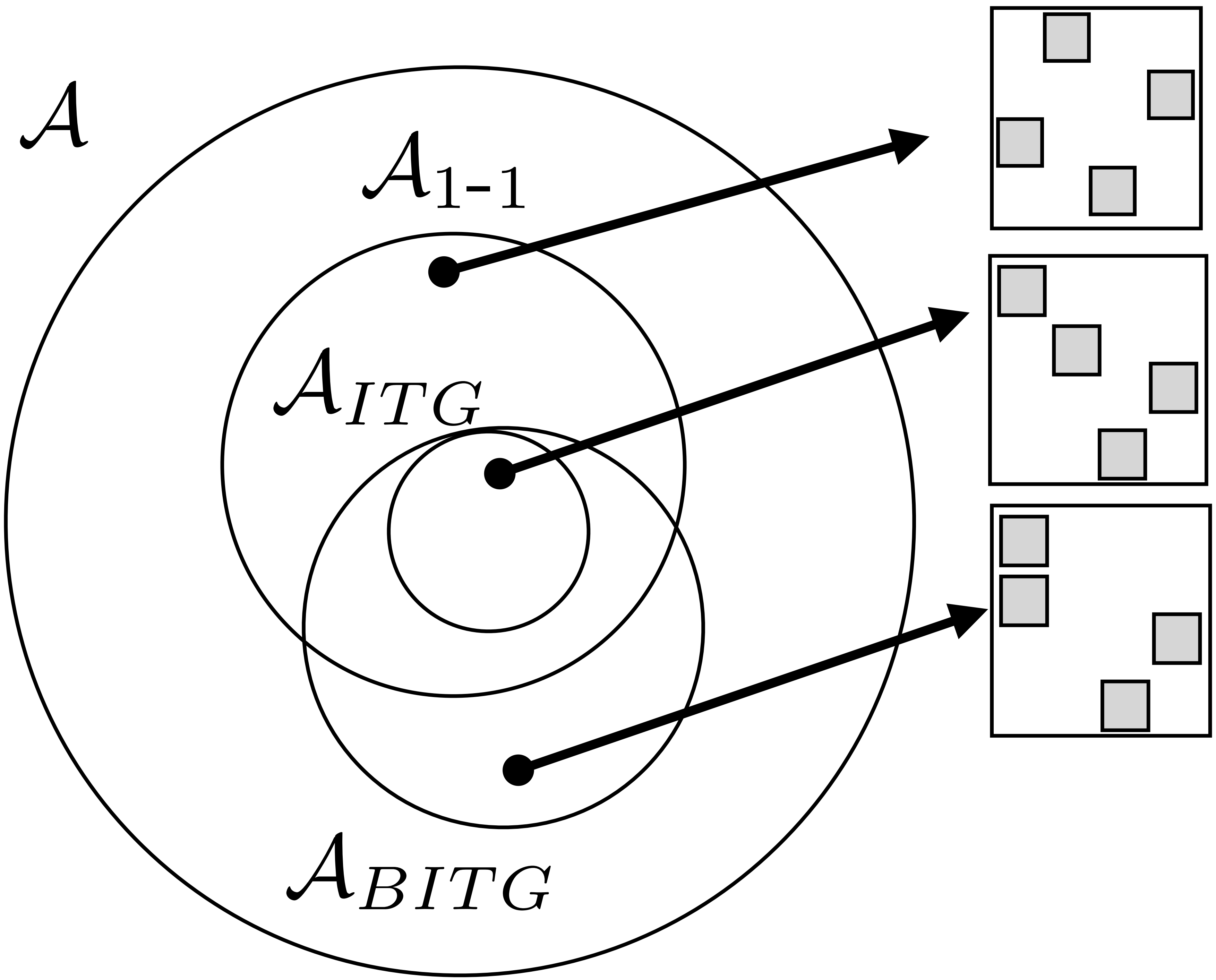


Optimal AER	
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2

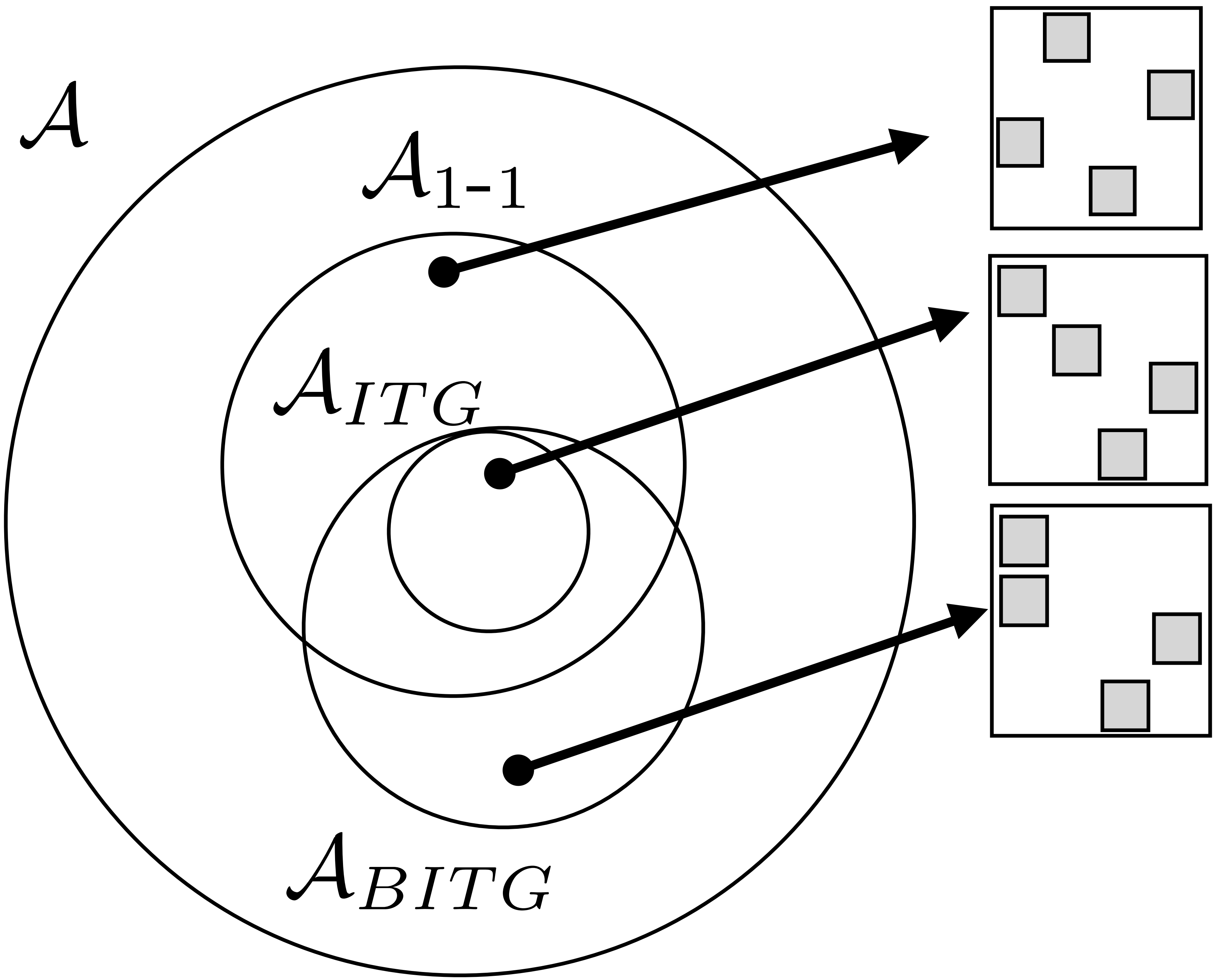
Alignment Families



Alignment Families



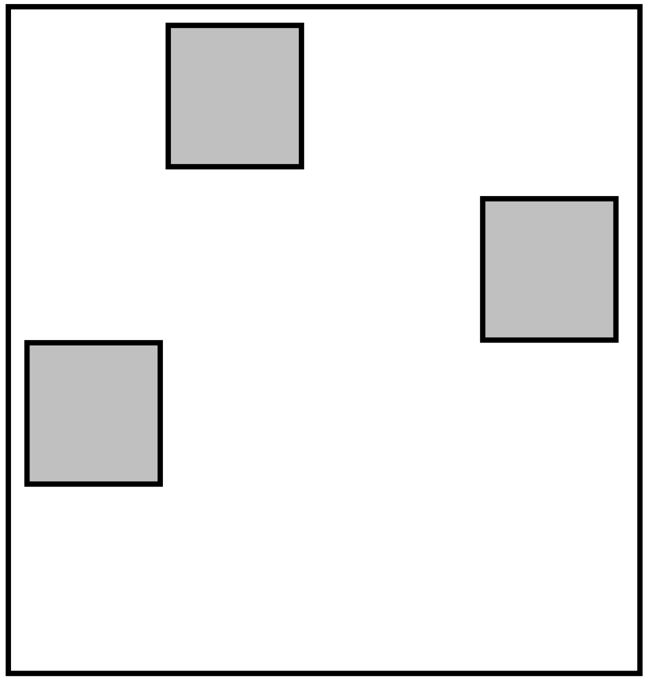
Alignment Families



Optimal AER	
$\mathcal{A}_{1\text{-}1}$	10.1
$\mathcal{A}_{ITG}$	10.2
$\mathcal{A}_{BITG}$	1.2

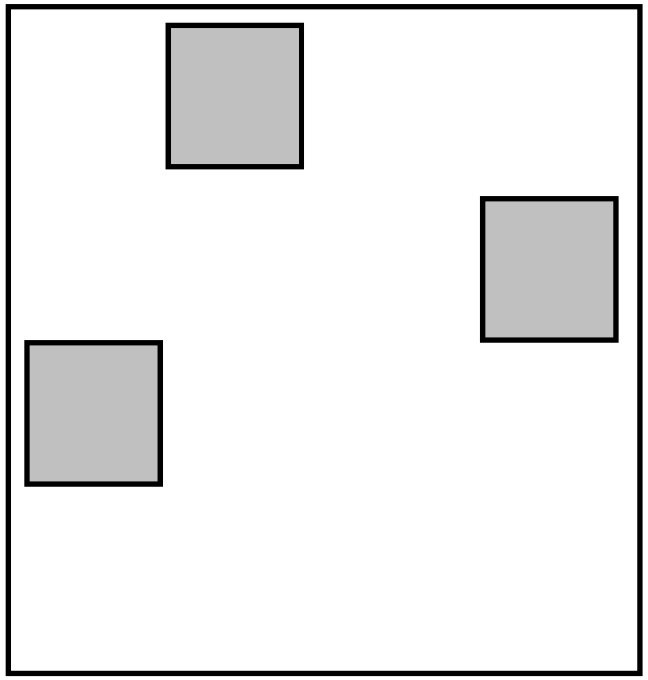
Learning Alignments

$$\mathbf{a} \in \mathcal{A}$$



Learning Alignments

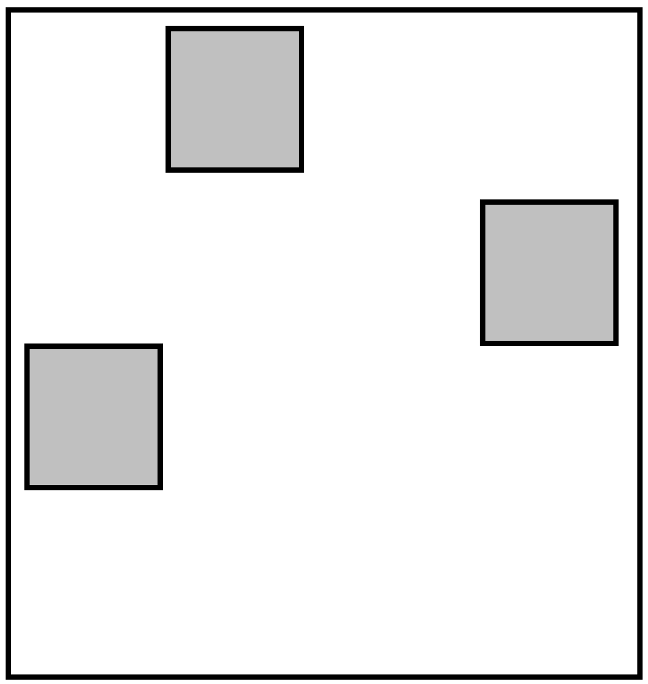
$$\mathbf{a} \in \mathcal{A}$$



$$s(\mathbf{a}) = \mathbf{w}^T \phi(\mathbf{a})$$

Learning Alignments

$$\mathbf{a} \in \mathcal{A}$$



Score

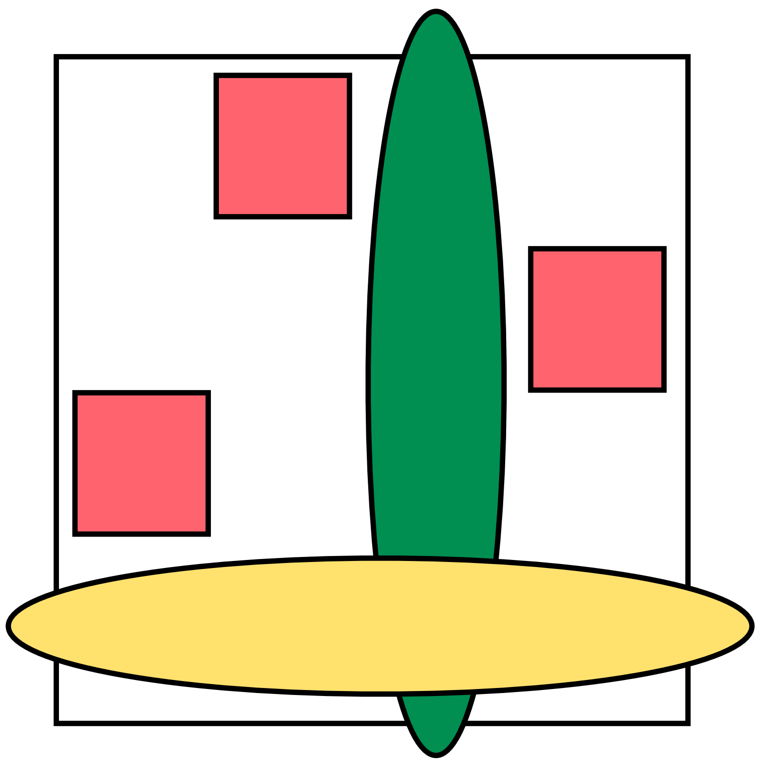
$$\mathbf{w}^T \phi(\mathbf{a})$$

Features

$$\phi(\mathbf{a}) = \sum_{(i,j)\in\mathbf{a}} \phi_{ij} + \sum_{i\notin\mathbf{a}} \phi_{i\epsilon} + \sum_{j\notin\mathbf{a}} \phi_{\epsilon j}$$

Learning Alignments

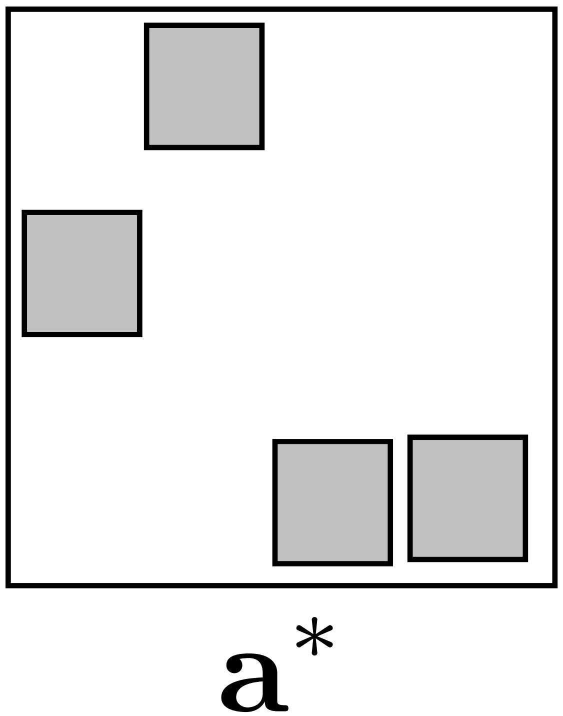
$$\mathbf{a} \in \mathcal{A}$$

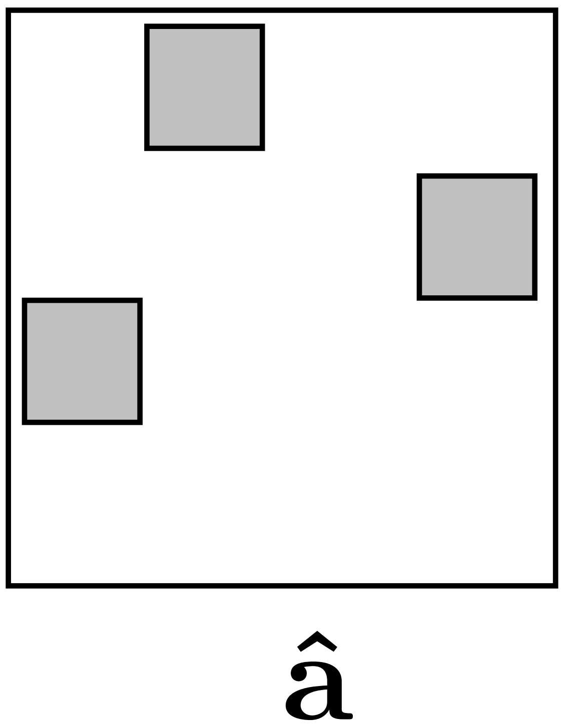


$$s_{\mathbf{w}}(\mathbf{a}) = \mathbf{w}^T \phi(\mathbf{a})$$

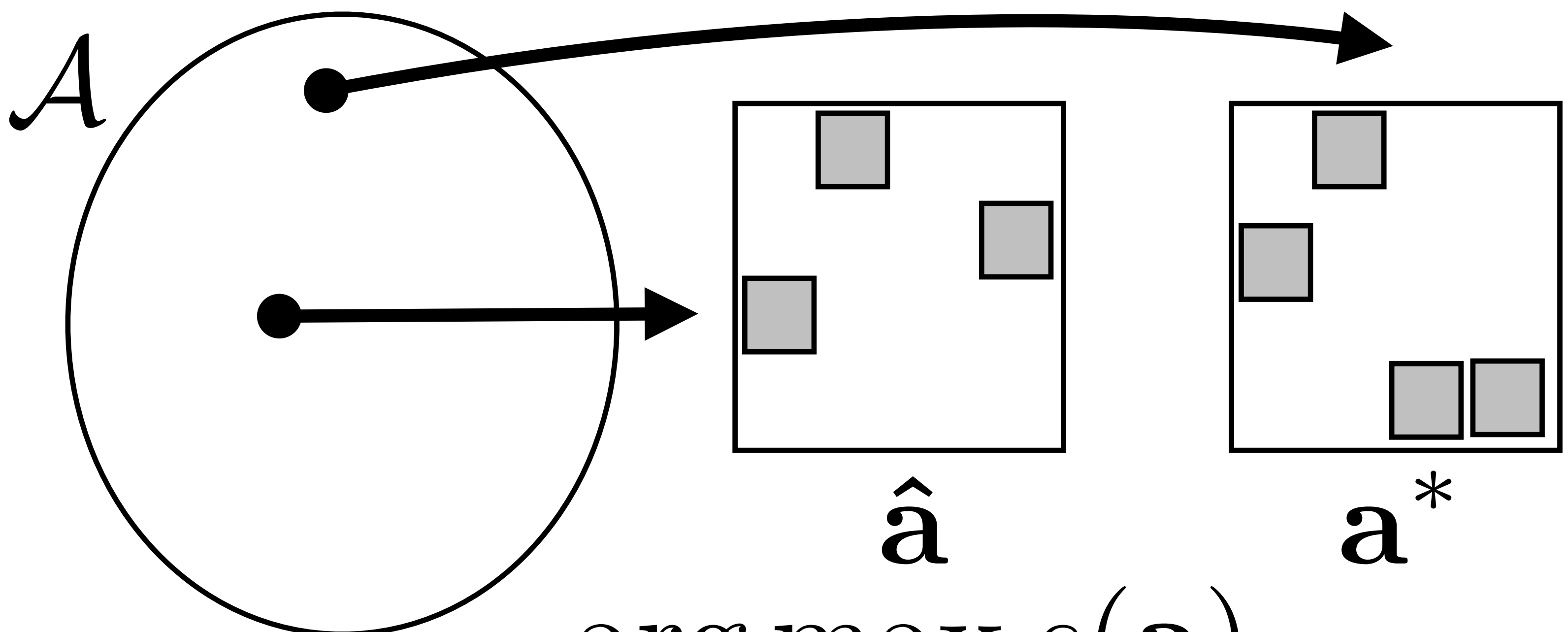
$$\phi(\mathbf{a}) = \sum_{(i,j)\in\mathbf{a}} \phi_{ij} + \sum_{i\notin\mathbf{a}} \phi_{i\epsilon} + \sum_{j\notin\mathbf{a}} \phi_{\epsilon j}$$

MIRA: Margin Criterion





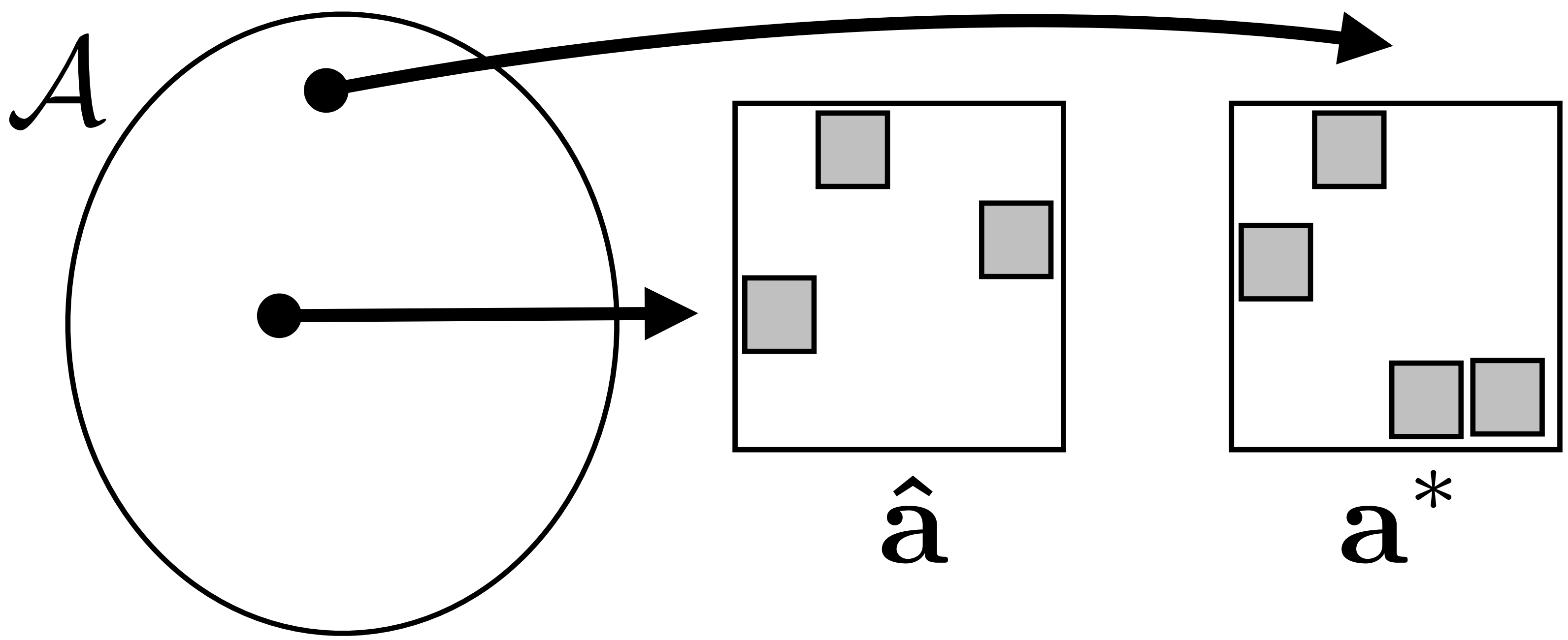
MIRA: Margin Criterion



$$\arg\max_{\mathbf{a} \in \mathcal{A}} s(\mathbf{a})$$

$$s_{\mathbf{w}}(\mathbf{a}^*) \geq s_{\mathbf{w}}(\hat{\mathbf{a}}) + L(\mathbf{a}^*, \hat{\mathbf{a}})$$

MIRA: Margin Criterion

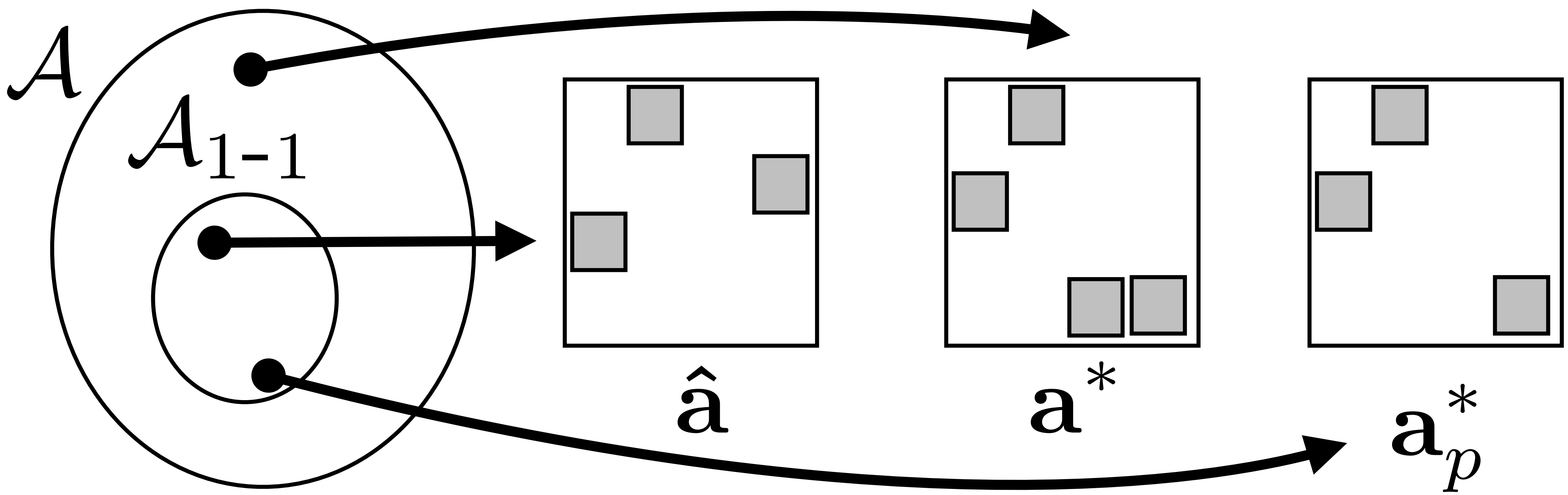


$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_t\|^2$$

$$\text{s.t. } \hat{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{A}} s_{\mathbf{w}_t}(\mathbf{a})$$

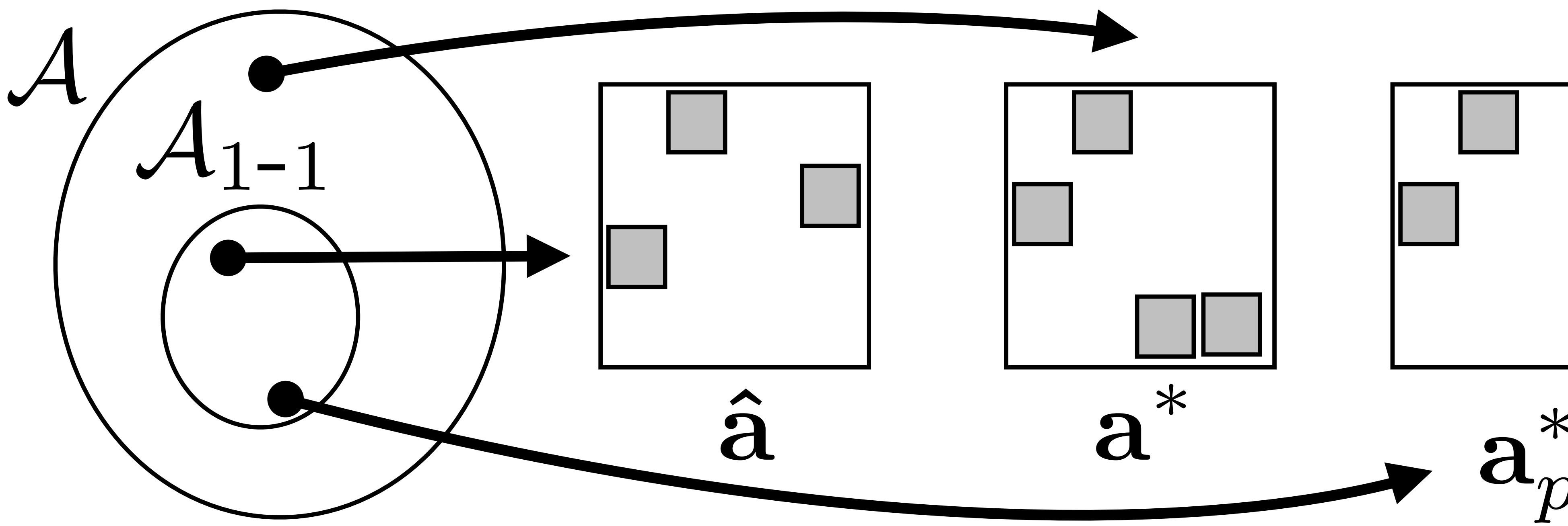
$$s_{\mathbf{w}_{t+1}}(\mathbf{a}^*) \geq s_{\mathbf{w}_{t+1}}(\hat{\mathbf{a}}) + L(\mathbf{a}^*, \hat{\mathbf{a}})$$

MIRA: Margin Criterion



$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{A}_{1\text{-}1}} s_{\mathbf{w}_t}(\mathbf{a}) + \lambda L(\mathbf{a}_p^*, \mathbf{a})$$

MIRA: Margin Criterion



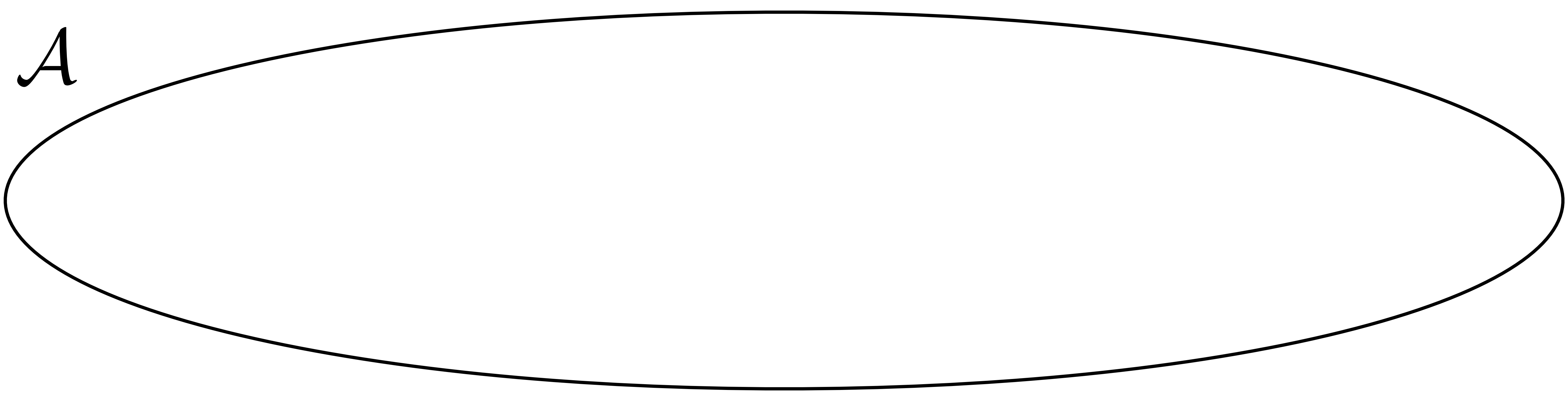
$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}_t\|^2$$

$$\text{s.t. } \hat{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{A}_{1\text{-}1}} s_{\mathbf{w}_t}(\mathbf{a}) + \lambda L(\mathbf{a}_p^*, \mathbf{a})$$

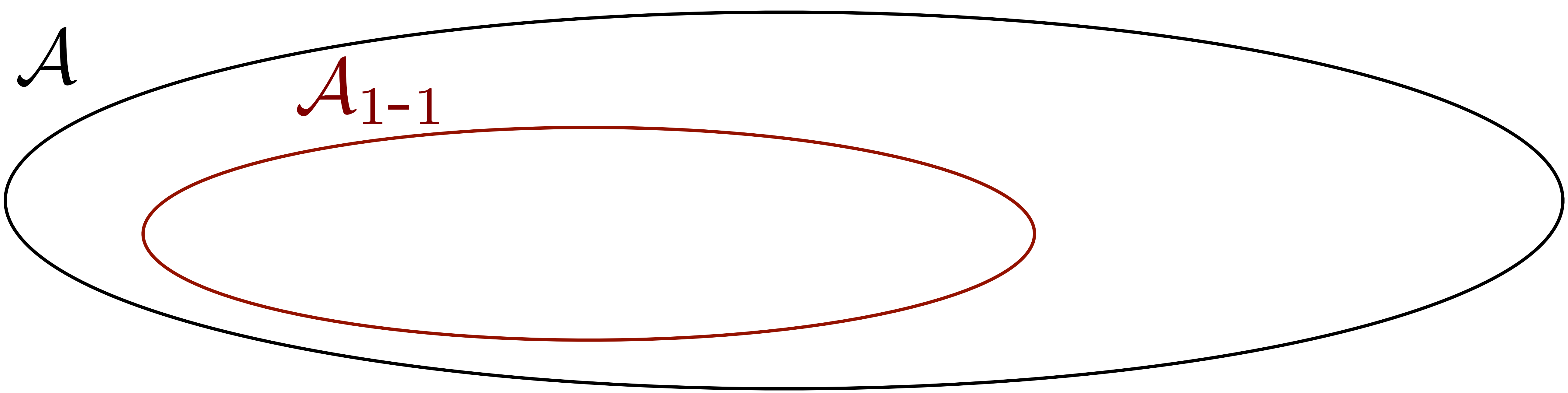
$$s_{\mathbf{w}_{t+1}}(\mathbf{a}^*) \geq s_{\mathbf{w}_{t+1}}(\hat{\mathbf{a}}) + L(\mathbf{a}_p^*, \hat{\mathbf{a}})$$

Alignment Families

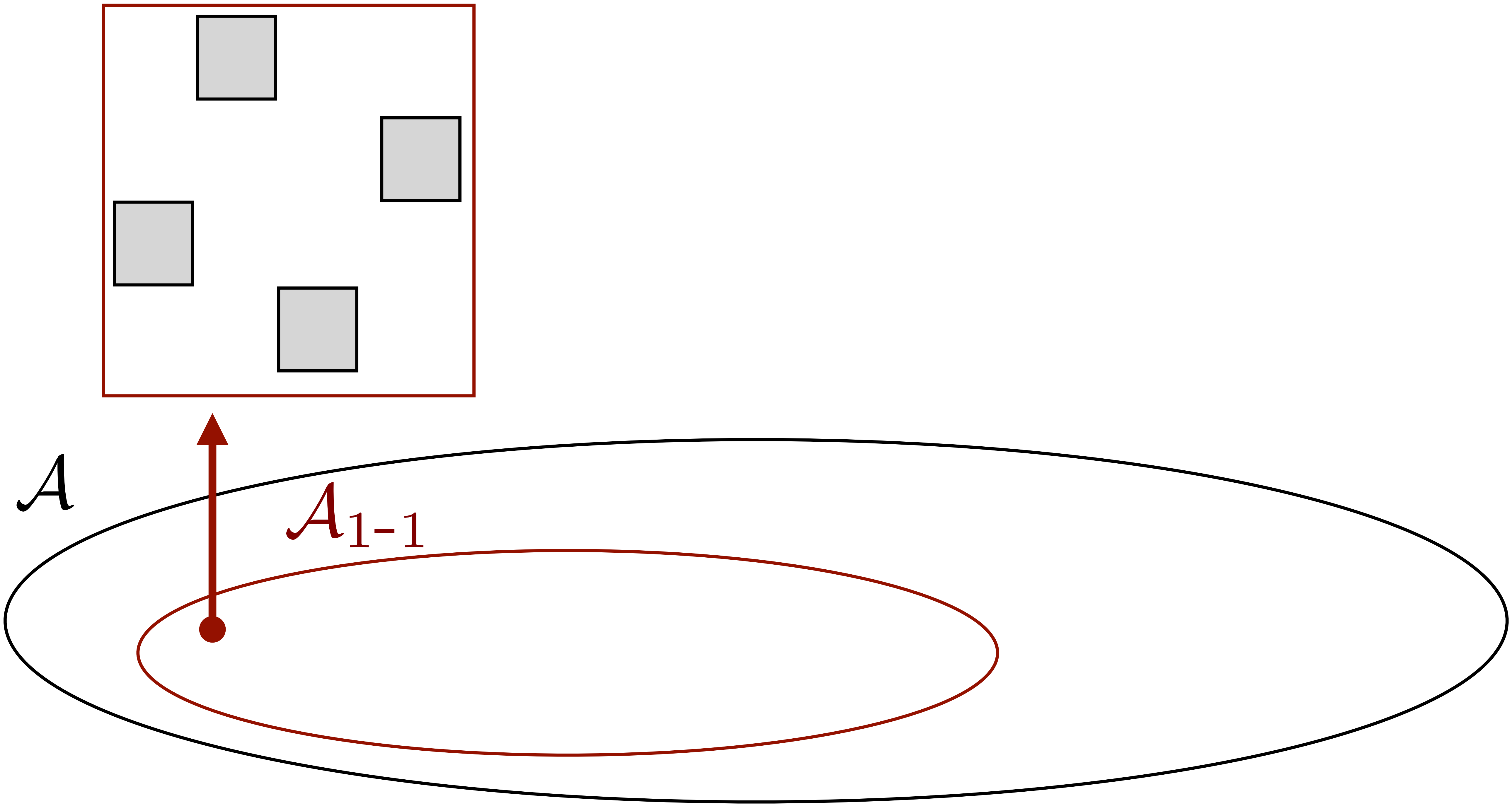
$\mathcal{A}$



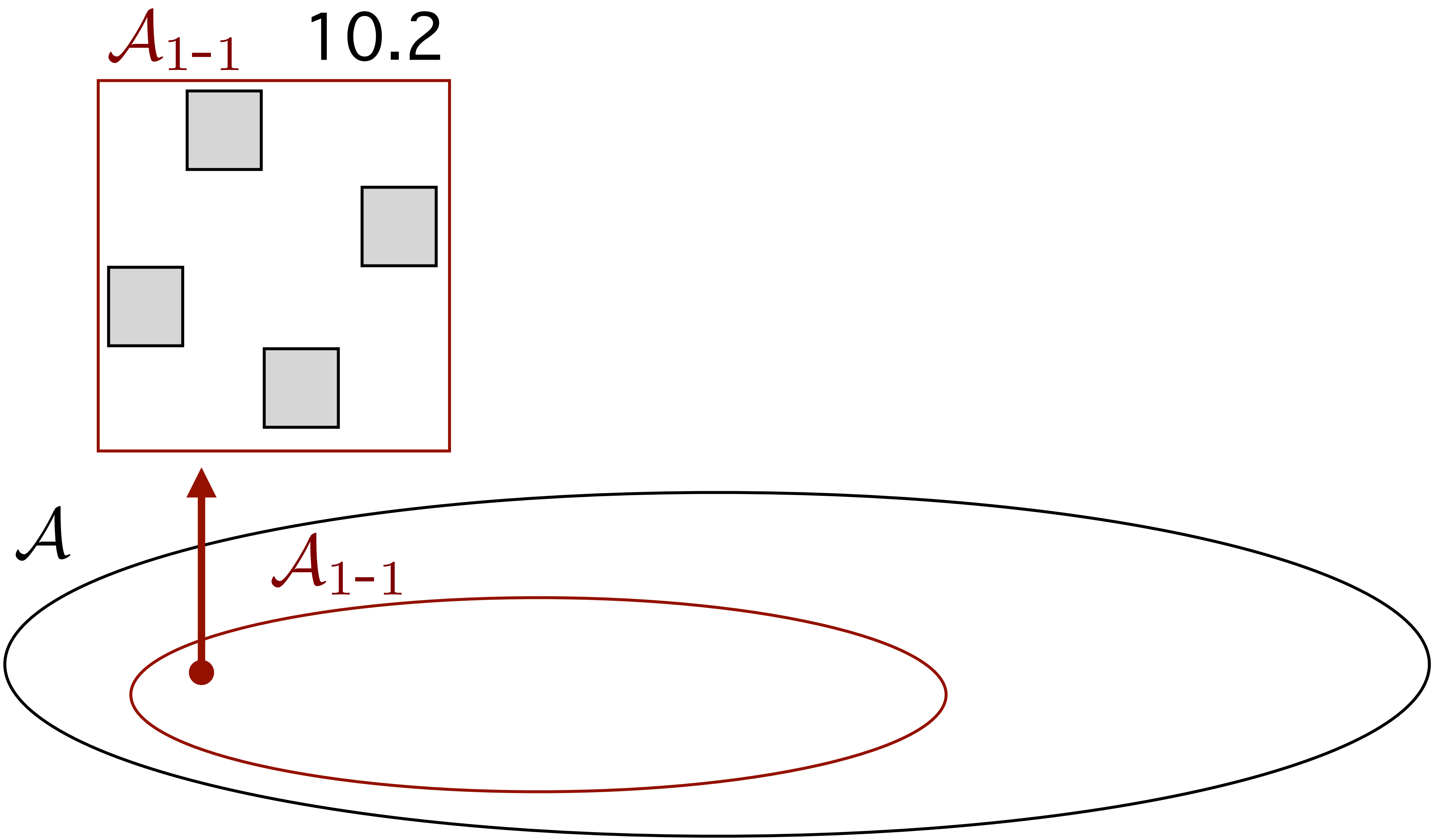
Alignment Families



Alignment Families



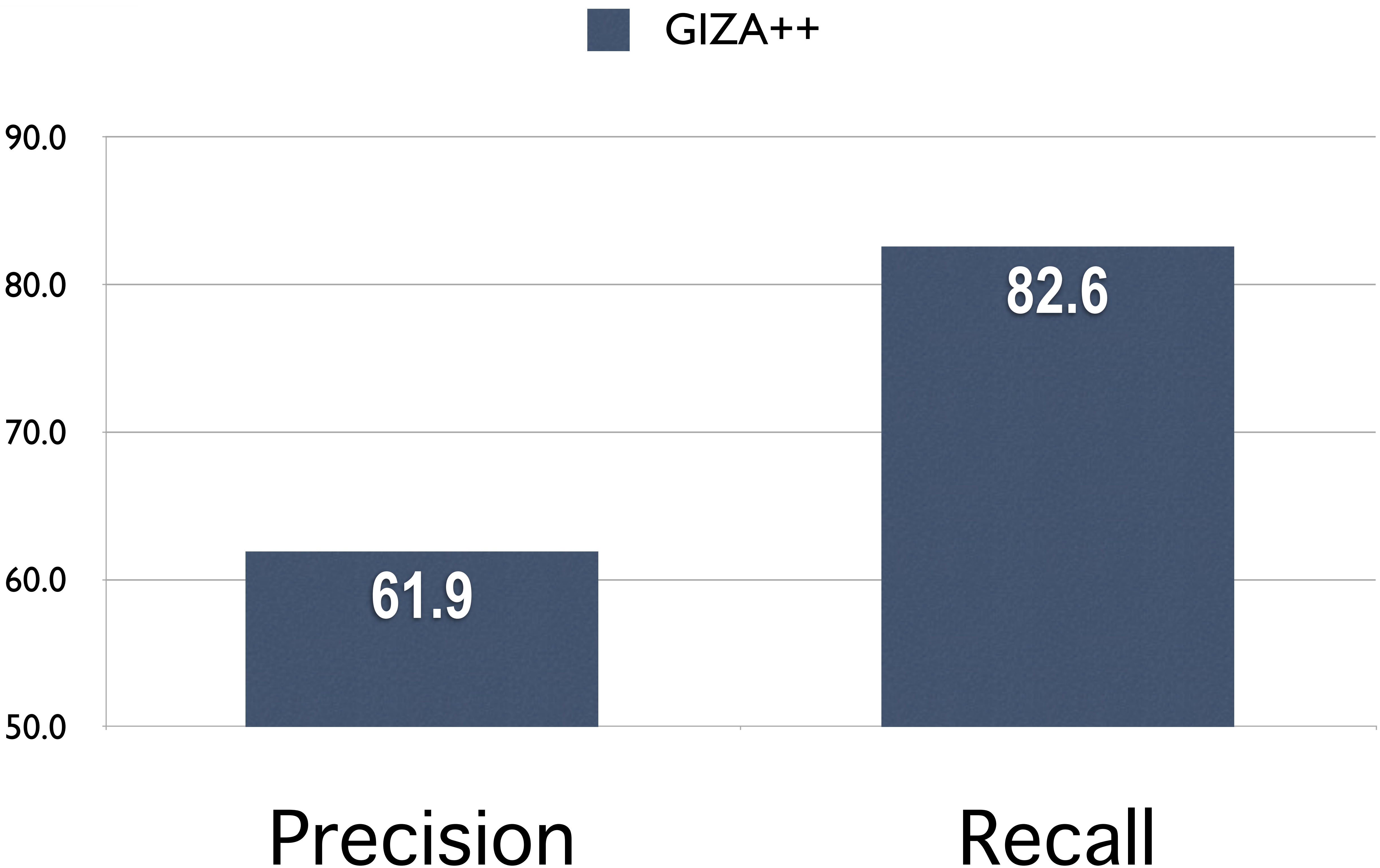
Alignment Families



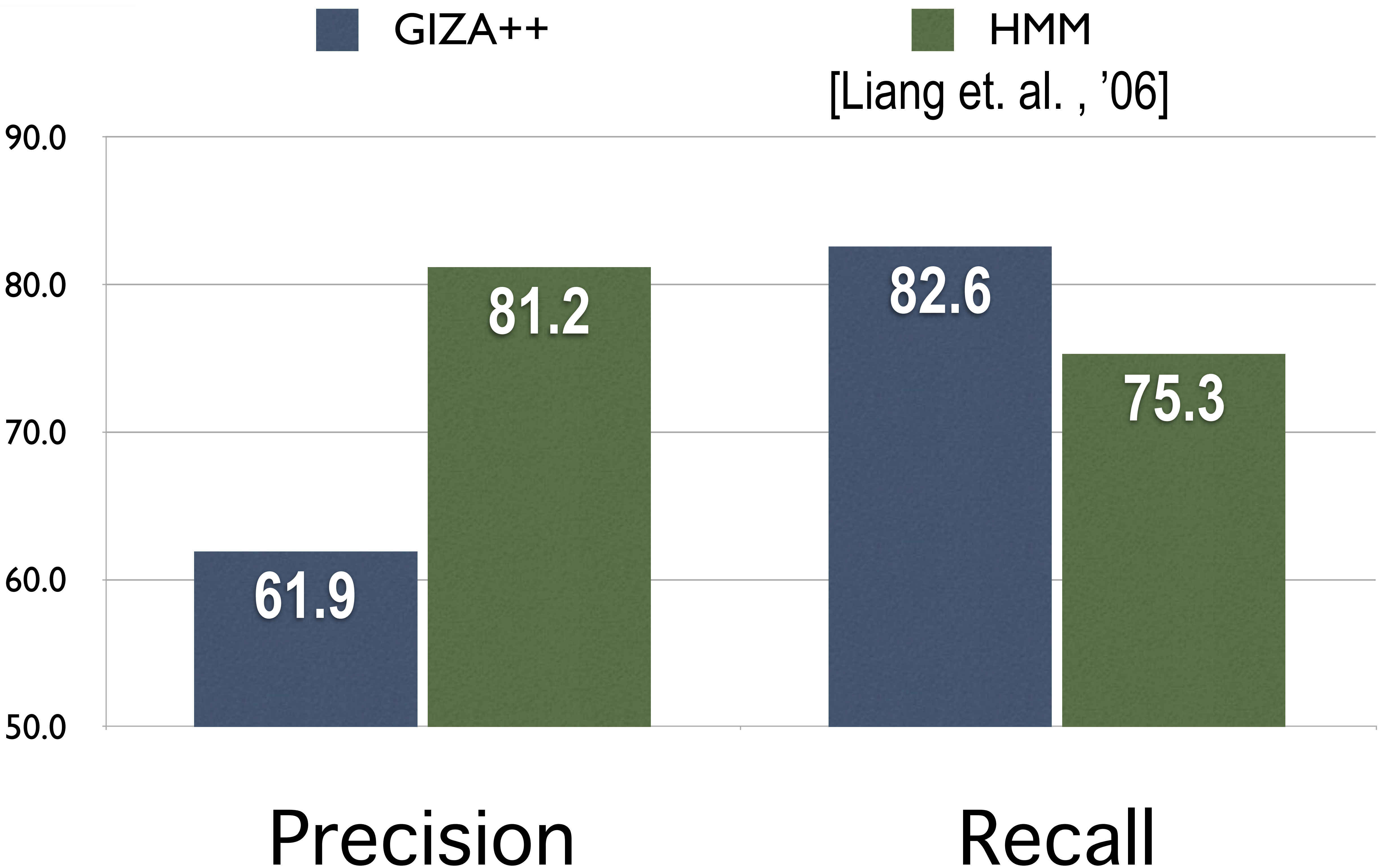
Results

- GIZA++

Results

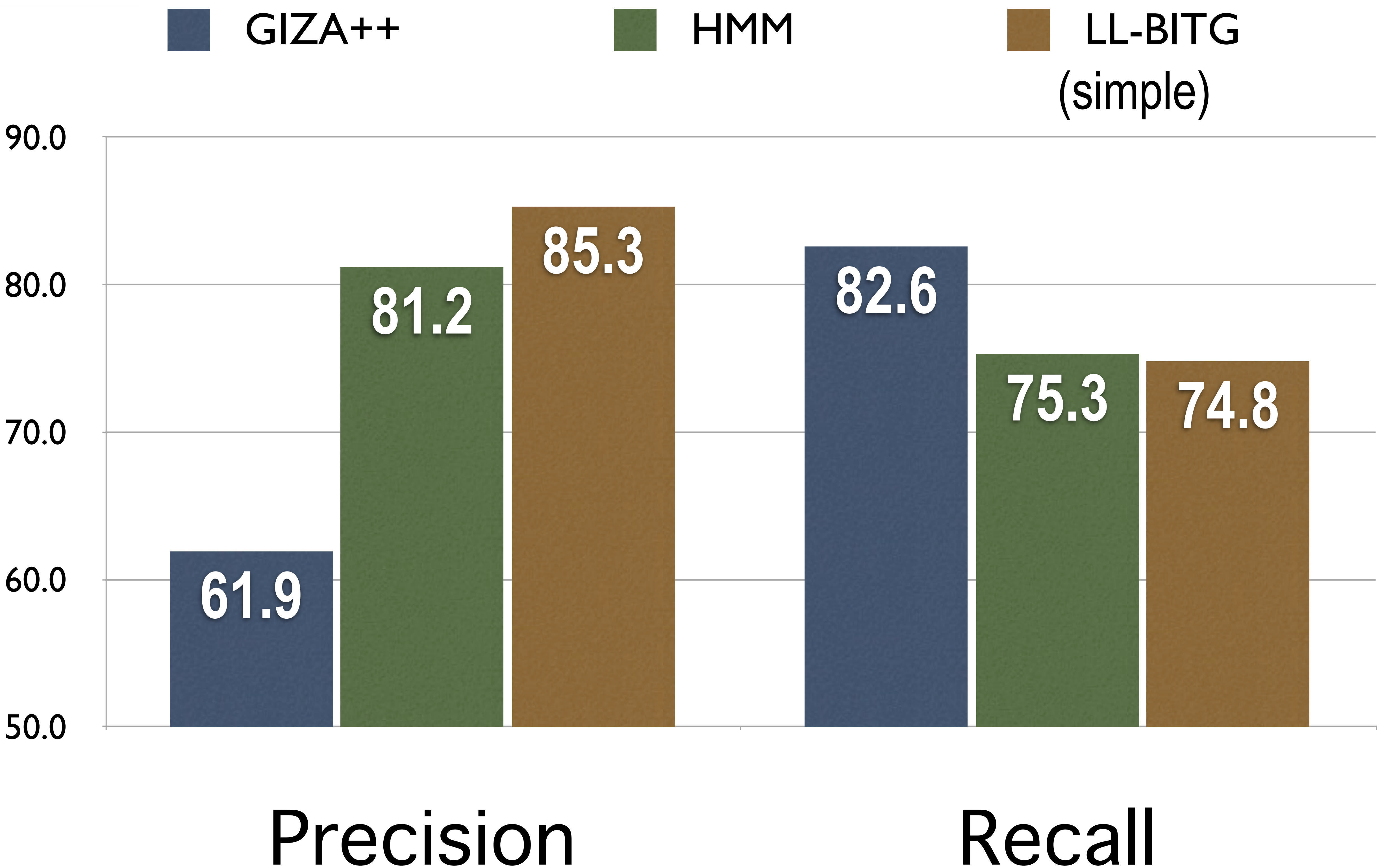


Results



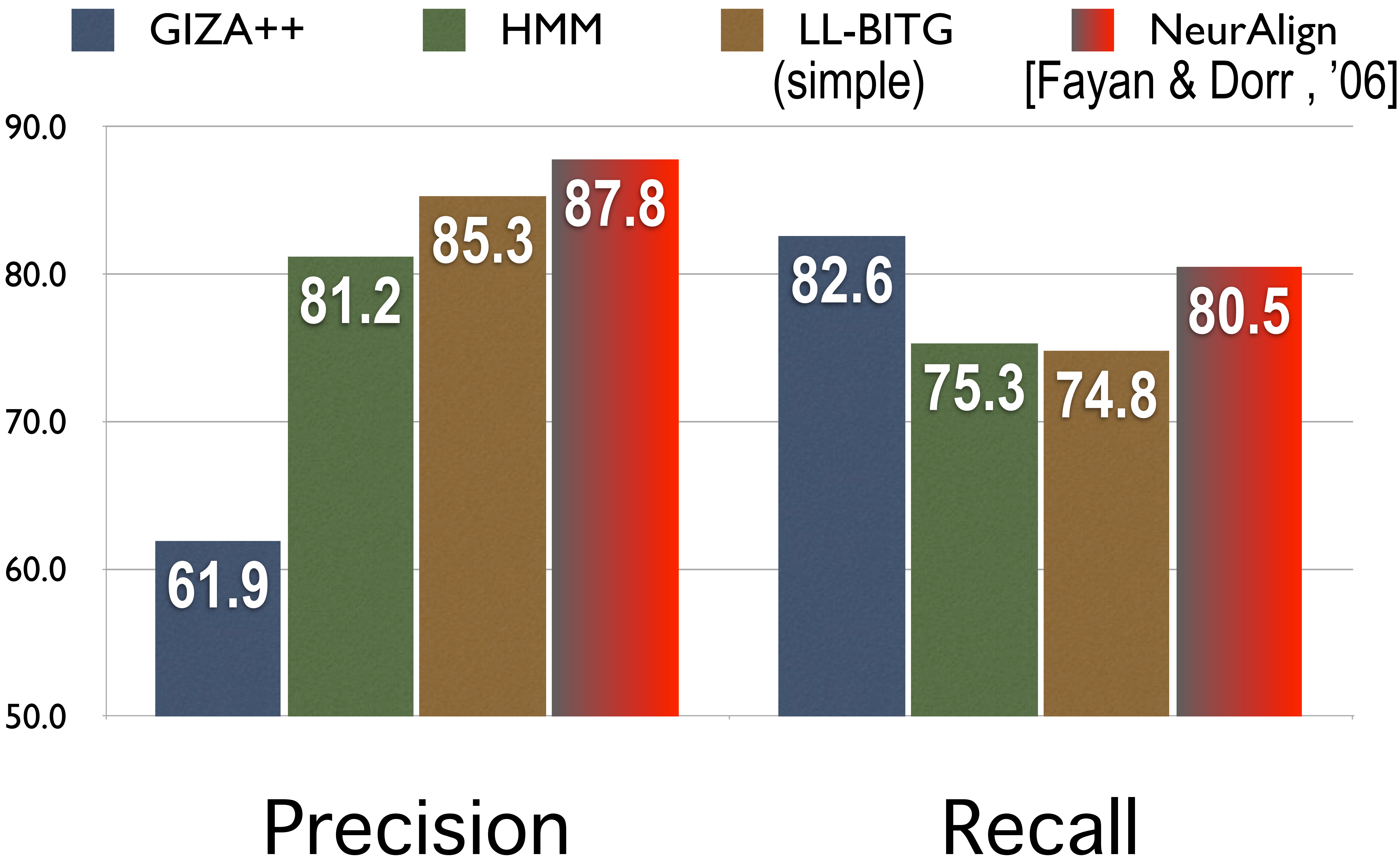
GIZA++
HMM
[Liang et. al. , '06]
90.0
80.0
70.0
60.0
50.0
61.9
81.2
82.6
75.3
Precision
Recall

Results

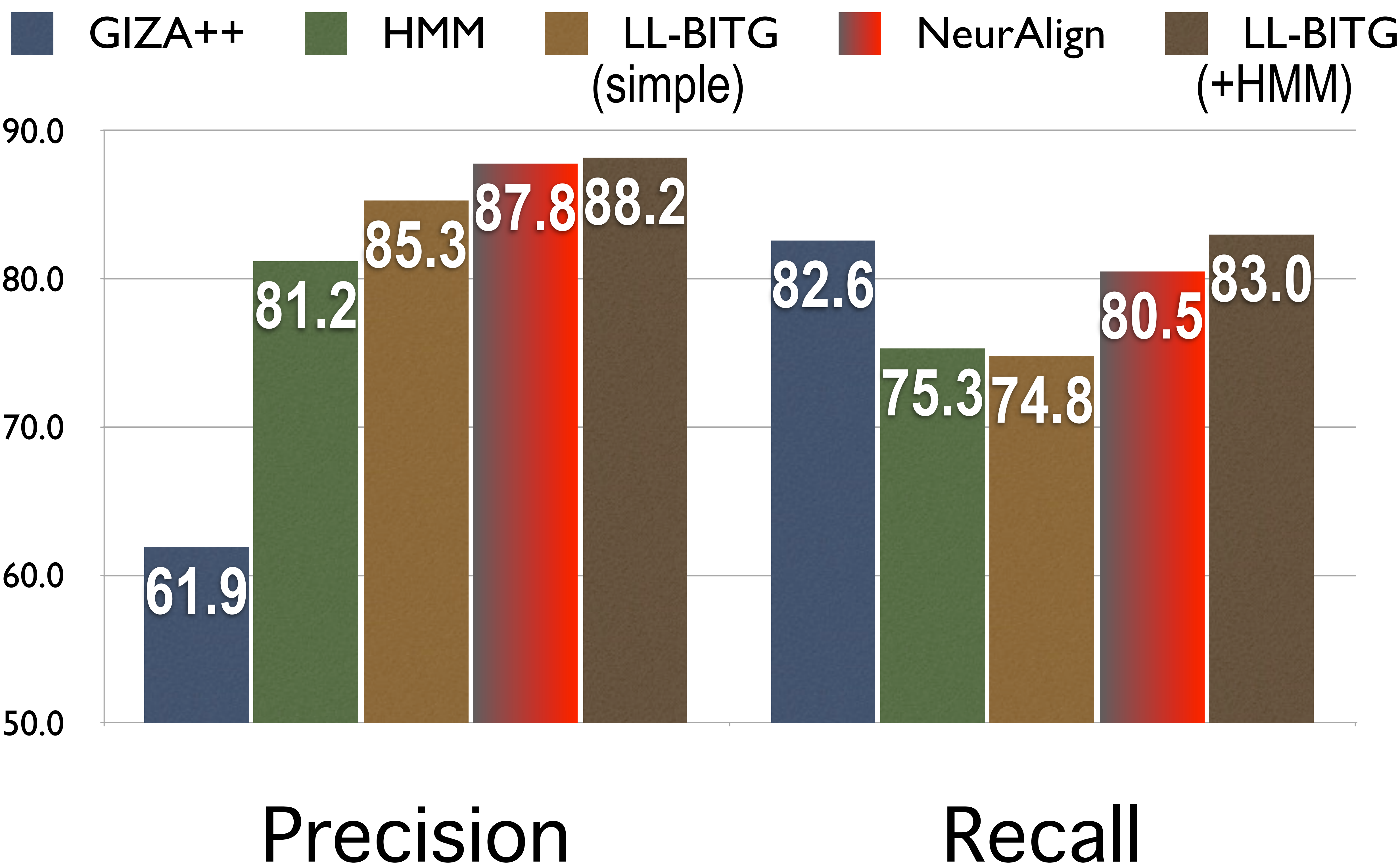


GIZA++
HMM
LL-BITG (simple)
90.0
80.0
70.0
60.0
50.0
61.9
81.2
85.3
82.6
75.3
74.8
Precision
Recall

Learning Alignments

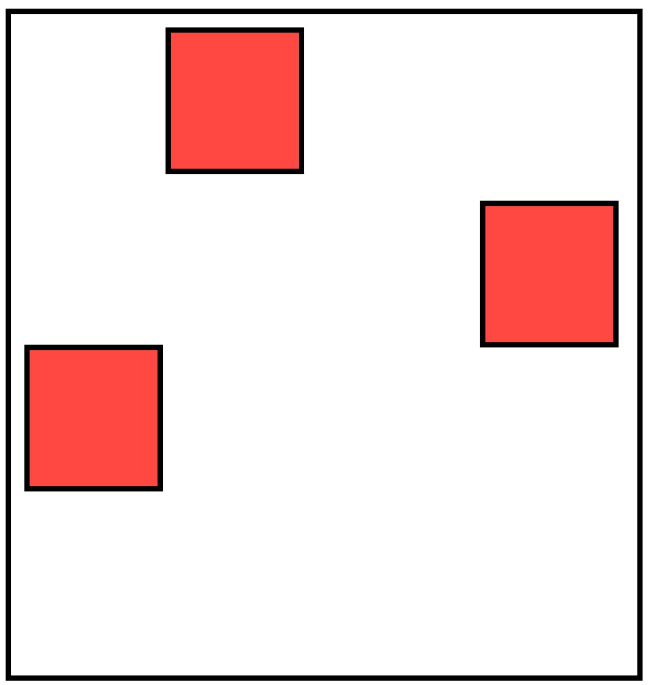


Learning Alignments



Linear Model

$$\mathbf{a} \in \mathcal{A}'$$



Score

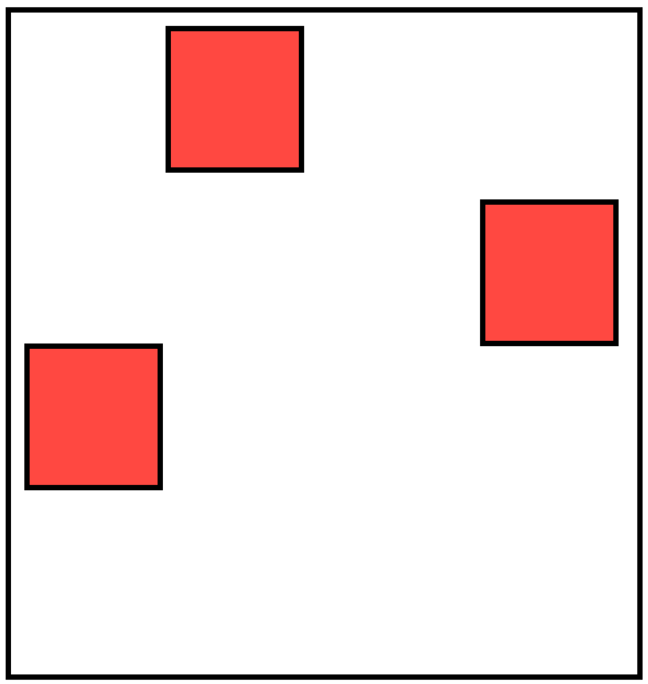
$$\mathbf{w}^T \phi(\mathbf{a})$$

Features

$$\phi(\mathbf{a}) = \sum_{(i,j)\in\mathbf{a}} \phi_{ij}$$

Linear Model

$$\mathbf{a} \in \mathcal{A}'$$



Score

$$\mathbf{w}^T \phi(\mathbf{a})$$

Features

$$\phi(\mathbf{a}) \quad = \quad \sum_{(i,j)\in\mathbf{a}} \phi_{ij}$$

{ Dice(i,j), Dictionary(i,j), LogFreqDiff(i,j),... }