Multiple Aspect Ranking Using the Good Grief Algorithm

Benjamin Snyder and Regina Barzilay
MIT

From One Opinion To Many

- Much previous work assumes *one opinion per text*.
  (Turney 2002; Pang et al 2002; Pang & Lee 2005)
- Real texts contain multiple, related, opinions.

> The food was a little greasy, but it was priced pretty well. Our only complaint was the service after our order was taken.

Food Price Service

http://www.we8there.com

Ruby Tuesdays

495 Union St # 2168
Brass Mill Mall
Waterbury, CT 06706
Phone: (203) 757-5509

Google Map

Number of reviews to date: 1

Loris Feldman (11/09/2006)

Add your review!
Email a friend

RATINGS KEY

- *x5 - Excellent*
- *x4 - Above Avg.*
- *x3 - Average*
- *x2 - Below Avg.*
- *x1 - Poor*

Food	
Service	
Price/Value	
Atmosphere	
Overall	
Was it a pleasant experience?	No
How many were seated at your table?	2
Do they accept reservations?	Yes
Would you return?	Not sure
Credit cards accepted?	Yes

Cutting Costs results in cutting customers. I love Ruby Tuesdays' burgers, but I will probably not be craving them as much. They are trying to cut costs. For example, by providing no ketchup with burgers and by giving chips with your take-out burger orders. The chips were a disappointment when expecting fries. The plates have gotten smaller giving each customer a lot less food. Also, the dessert, the chocolate tall cake should have its name changed. It now comes on a small plate, not the big margarita glass, and has barely any sauce and a tiny scoop of ice cream. It is also not heated since it would fall over. With all these changes for the worst, the prices still remain the same. The service has become horrible. One hour and 35 minutes to get two burgers and one dessert. It is sad that they have changed for the worse because I used to love this place.

Submitted by:
Loris Feldman (11/09/2006)
Posted 9 reviews

Multiple Aspect Opinion Analysis

- The Task:

  Predict writer's opinion on a fixed set of aspects (e.g. food, service, price etc) using a fixed scale (e.g. from 1-5).

- Simple Approach:

  Treat each aspect as an independent ranking (rating) problem.

Shortcomings of Independent Ranking

- Multiple opinions in a single text are correlated.
- Real text relates opinions in coherent, meaningful ways.

> The food was a little greasy, but it was priced pretty well. Our only complaint was the service after our order was taken.

Service Food Price

Shortcomings of Independent Ranking

- Multiple opinions in a single text are correlated.
- Real text relates opinions in coherent, meaningful ways.

> The food was a little greasy, **but** it was priced pretty well. **Our only complaint** was the service after our order was taken.

Service **<** Food **<** Price

Shortcomings of Independent Ranking

- Multiple opinions in a single text are correlated.
- Real text relates opinions in coherent, meaningful ways.

The food was a little greasy, **but** it was priced pretty well. **Our only complaint** was the service after our order was taken.

Service **<** Food **<** Price

- Independent ranking fails to model correlations and to exploit discourse cues.

Key Goals

- Build on previous success of simple linear models trained with Perceptron in NLP:
  - Simple, fast training
  - Exact, fast decoding
  - High performance
- Extend framework to tasks with complex label dependencies:
  - Task-specific dependency space
  - Label dependencies sensitive to input features
  - Factorization of label prediction and dependency models

Our Idea: The Model

- Individual ranking model for each aspect.
- Add a "meta-model" which predicts discourse relations between aspects. e.g.,

Service < Food < Price *(order)*

Service = Food = Price *(agreement)*

~[Service = Food = Price] *(disagreement)*

Our Idea: The Model

- Individual ranking model for each aspect.
- Add a "meta-model" which predicts discourse relations between aspects. e.g.,

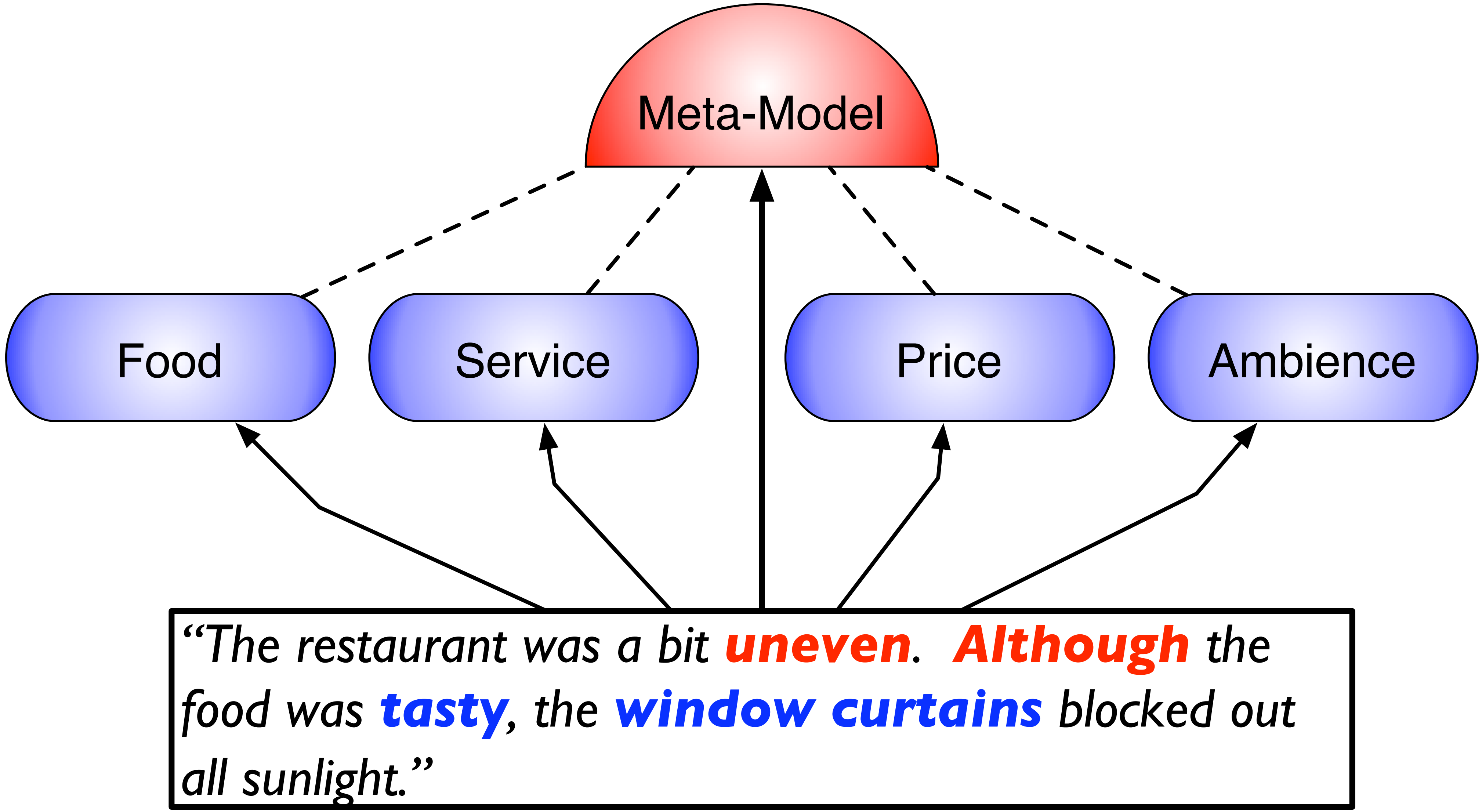
Service < Food < Price *(order)*

$$\left\{ \begin{array}{ll} \text{Service} = \text{Food} = \text{Price} & \textit{(agreement)} \\ \sim[\text{Service} = \text{Food} = \text{Price}] & \textit{(disagreement)} \end{array} \right\}$$

Binary Agreement Model

Our Idea: Learning and Inference

- Resolve differences between individual rankers and meta-model using overall confidence of ***all*** model components.
- Meta-model "glues together" individual ranking decisions in coherent way.
- Optimize individual ranker parameters to operate with meta-model through joint Perceptron updates.


Meta-Model
Food
Service
Price
Ambience
“The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight.”

Basic Ranking Framework

(Crammer & Singer 2001)

Goal:

Assign each input $\mathbf{x} \in \mathbb{R}^n$ a rank in $\{1, ..., k\}$

Model:

weight vector: $\mathbf{w} \in \mathbb{R}^n$

boundaries divide real line into k segments:

$$\mathbf{b} = (b_1, ..., b_{k-1})$$

Rank Decoding

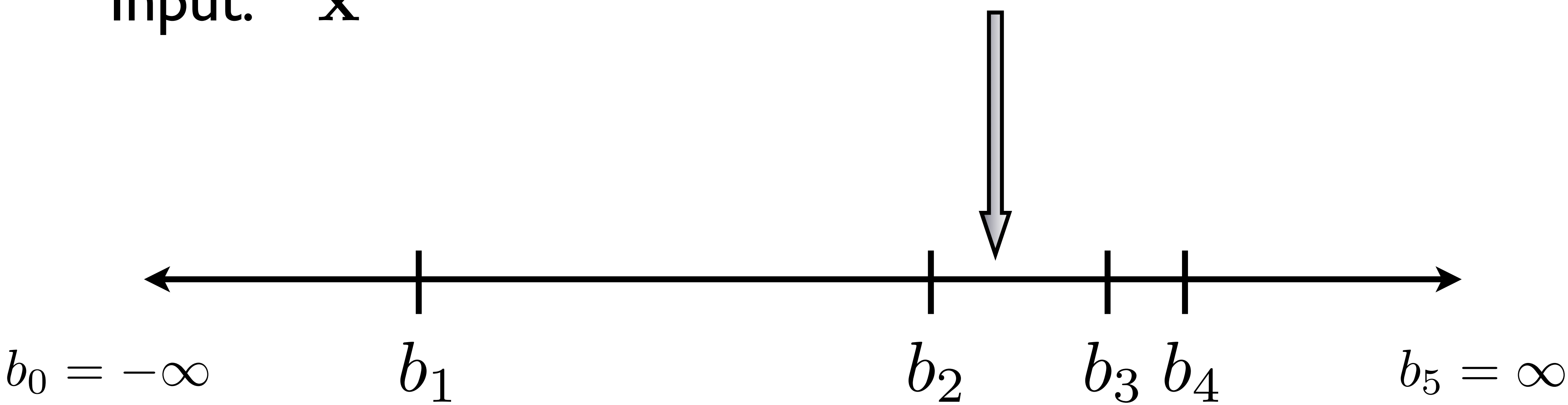
Input: $\mathbf{x}$

Output:

Rank Decoding

Input: $\mathbf{x}$

$score(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$

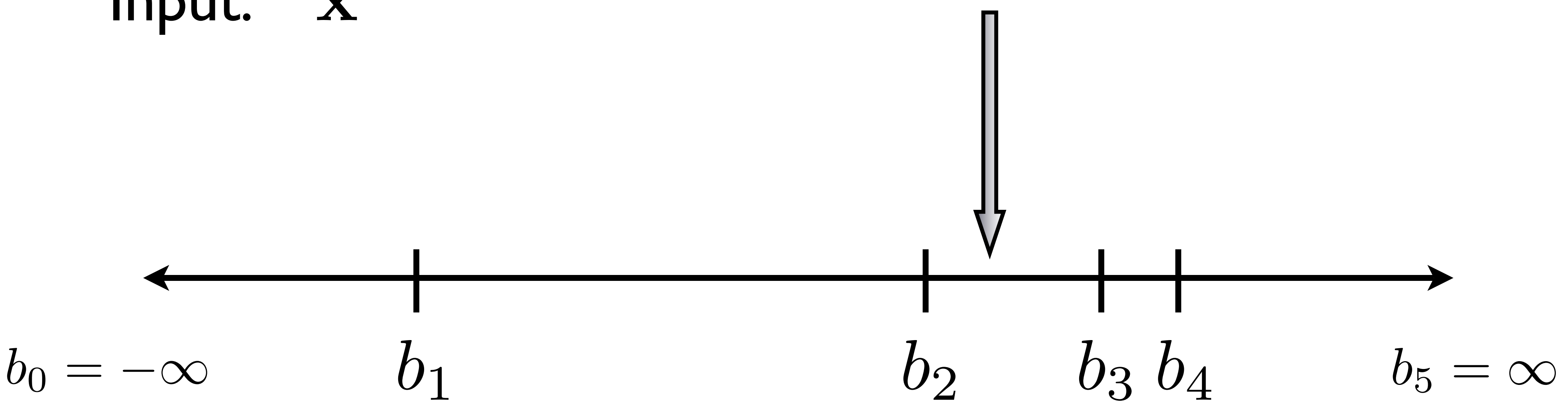


Output:

Rank Decoding

$score(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$

Input: $\mathbf{x}$



Output: $\hat{y} = 3$

Multiple Aspect Ranking

- Each input $\mathbf{x}$ has m aspects.
  - e.g. reviews rate products for different qualities -- food, service, ambience etc.
- Associated with each input $\mathbf{x}$ is a rank *vector* $\mathbf{r}$. (The component $r_i \in \{1, ..., k\}$ is the rank for aspect i.)

$$\mathbf{r} = < 5, 3, 4, 4, 4 >$$

food service ...

Joint Ranking Model

- Ranking model for each aspect i :

  $$(\mathbf{w}[i], \mathbf{b}[i])$$

- Linear agreement model $\mathbf{a} \in \mathbb{R}^n$ to predict unanimity across aspects:

  $$sign(\mathbf{a} \cdot \mathbf{x})$$

Component Models

- Combine predictions of individual rankers and agreement model:
  - Introduce "grief terms" and choose joint rank which minimizes their sum.

Grief of Component Models

$g_i(\mathbf{x}, r_i)$ **:** Measure of *dissatisfaction* of i^{th} aspect ranking model with rank r_i for input $\mathbf{x}$.

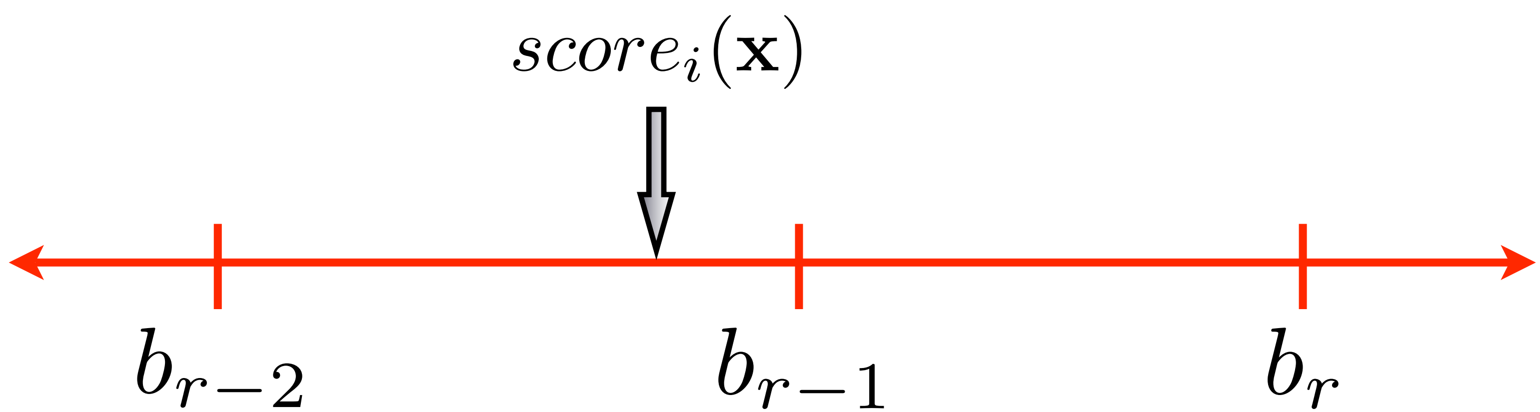
$g_a(\mathbf{x}, \mathbf{r})$ **:** Measure of *dissatisfaction* of agreement model with rank vector $\mathbf{r}$ for input $\mathbf{x}$.

Grief of i^{th} Ranking Model

$$g_i(\mathbf{x}, r) = \min |c|$$

$$\text{s.t.}$$

$$b[i]_{r-1} \leq score_i(\mathbf{x}) + c < b[i]_r$$

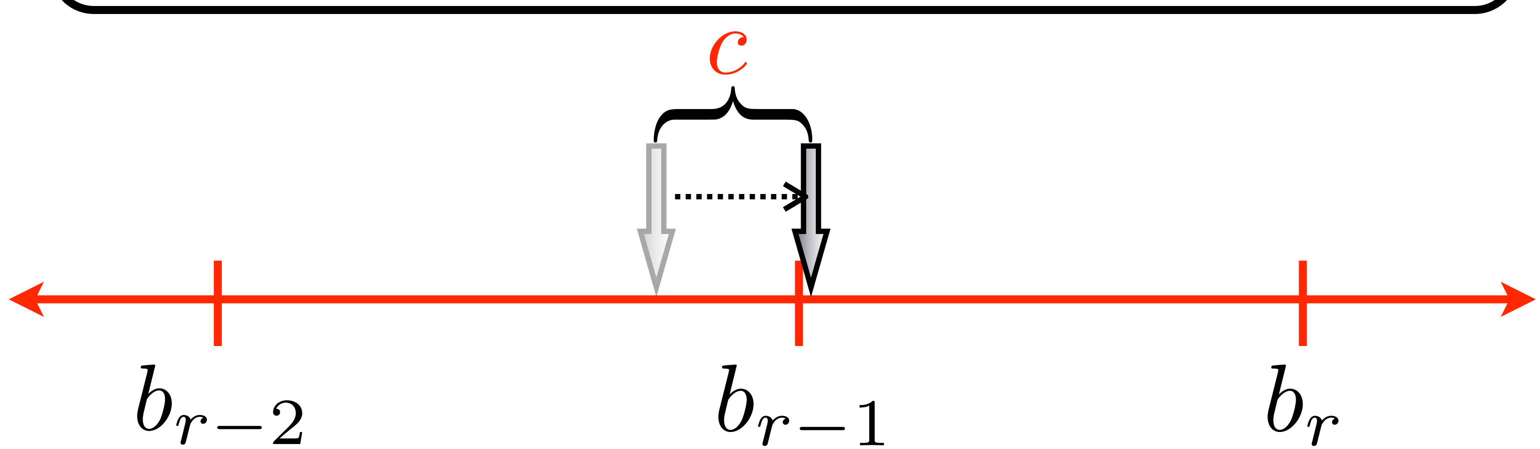


Grief of i^{th} Ranking Model

$$g_i(\mathbf{x}, r) = \min |c|$$

$$\text{s.t.}$$

$$b[i]_{r-1} \leq score_i(\mathbf{x}) + c < b[i]_r$$



Grief of Agreement Model

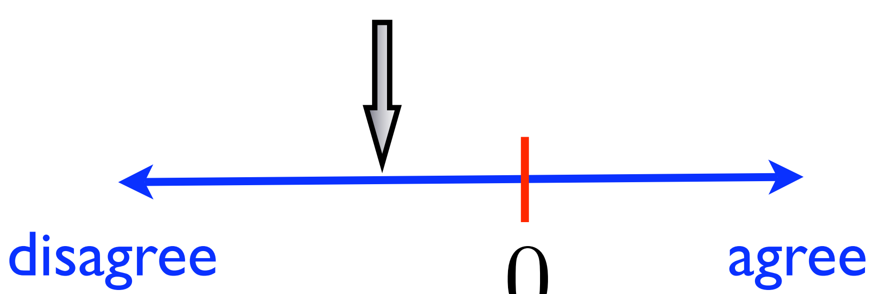
$$g_a(\mathbf{x}, \mathbf{r}) = \min |c|$$

$$\text{s.t.}$$

$$[(\mathbf{a} \cdot \mathbf{x} + c > 0) \ \wedge \ (r_1 = r_2 = \ldots = r_m)]$$

$$\vee$$

$$[(\mathbf{a} \cdot \mathbf{x} + c \leq 0) \ \wedge \ \neg(r_1 = r_2 = \ldots = r_m)]$$



$$r_1 = r_2 = \ldots = r_m$$

Grief of Agreement Model

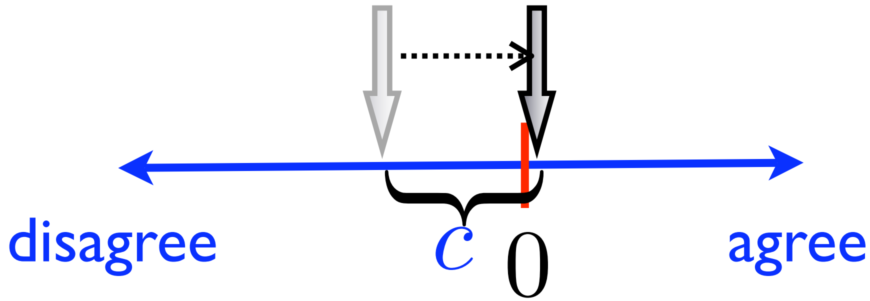
$$g_a(\mathbf{x}, \mathbf{r}) = \min |c|$$

$$\text{s.t.}$$

$$[(\mathbf{a} \cdot \mathbf{x} + c > 0) \ \wedge \ (r_1 = r_2 = \ldots = r_m)]$$

$$\vee$$

$$[(\mathbf{a} \cdot \mathbf{x} + c \leq 0) \ \wedge \ \neg(r_1 = r_2 = \ldots = r_m)]$$



$$r_1 = r_2 = \ldots = r_m$$

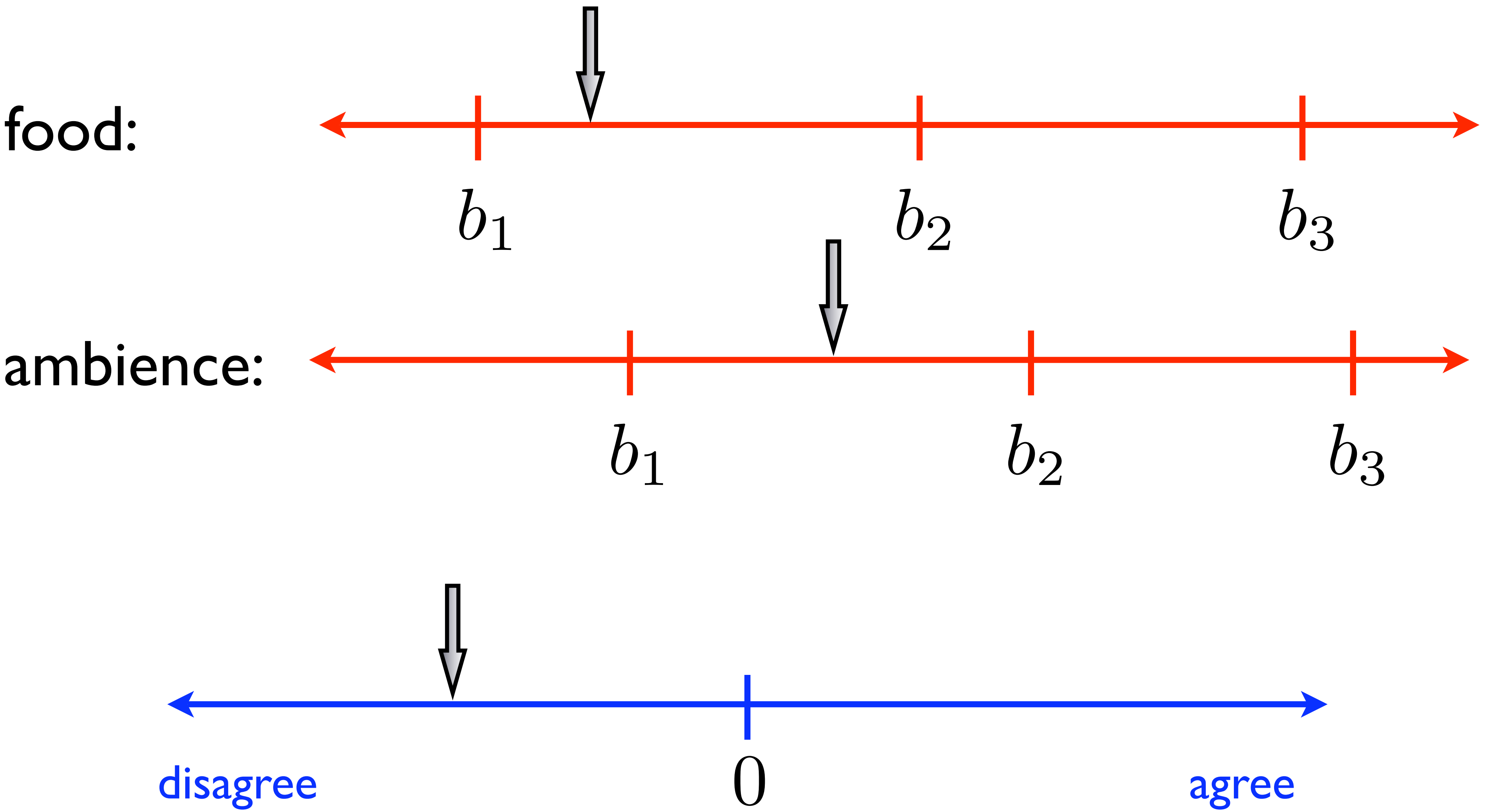
“Good Grief” Decoding

- Select joint rank $\mathbf{r} \in \{1, ..., k\}^m$ which minimizes total grief:

$$H(\mathbf{x}) = \arg\min_{\mathbf{r} \in \mathcal{Y}^m} \left[g_a(\mathbf{x}, \mathbf{r}) + \sum_{i=1}^{m} g_i(\mathbf{x}, r_i) \right]$$

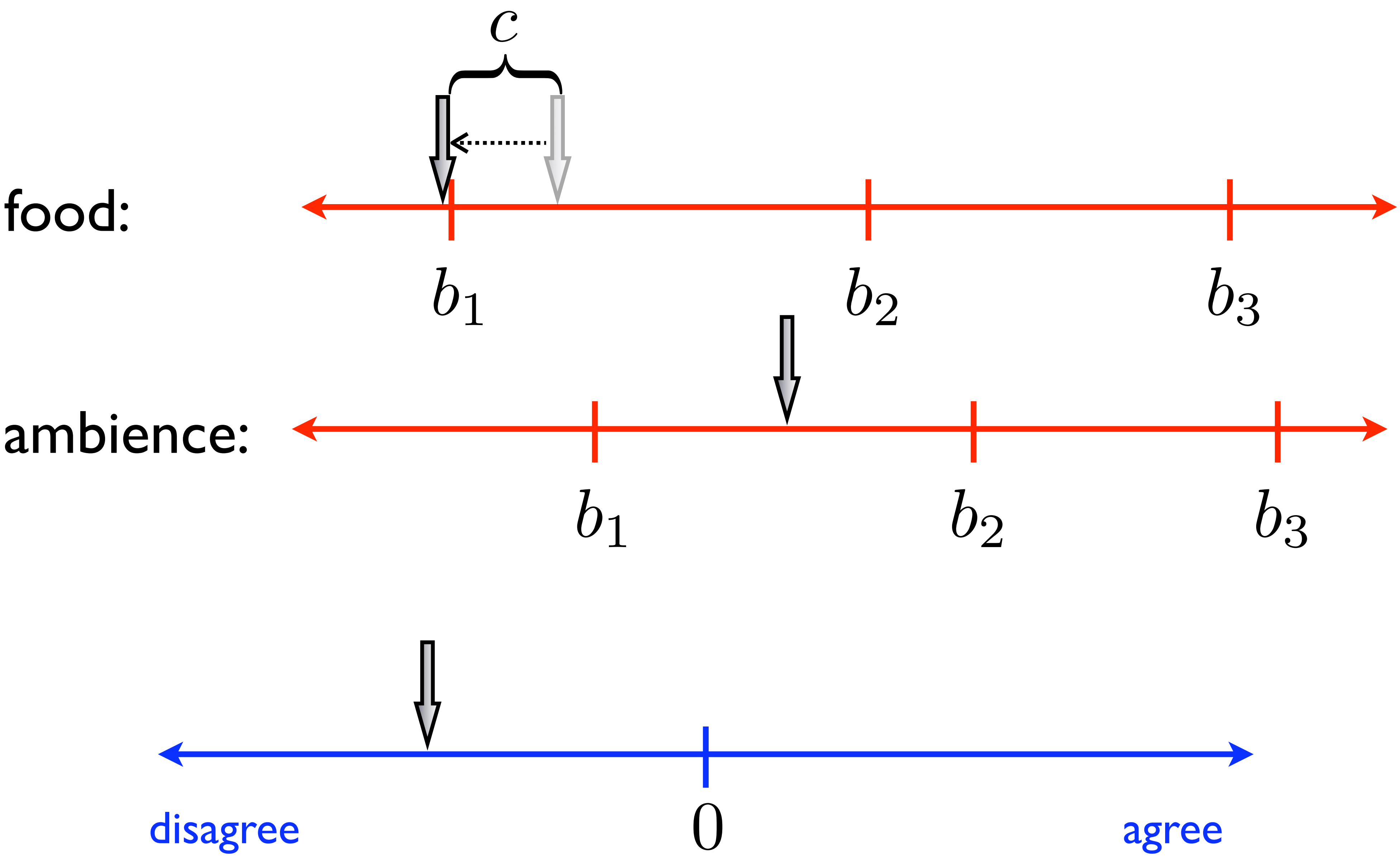
- Exact search is linear in number of aspects: $O(m)$

Disagreement with high confidence



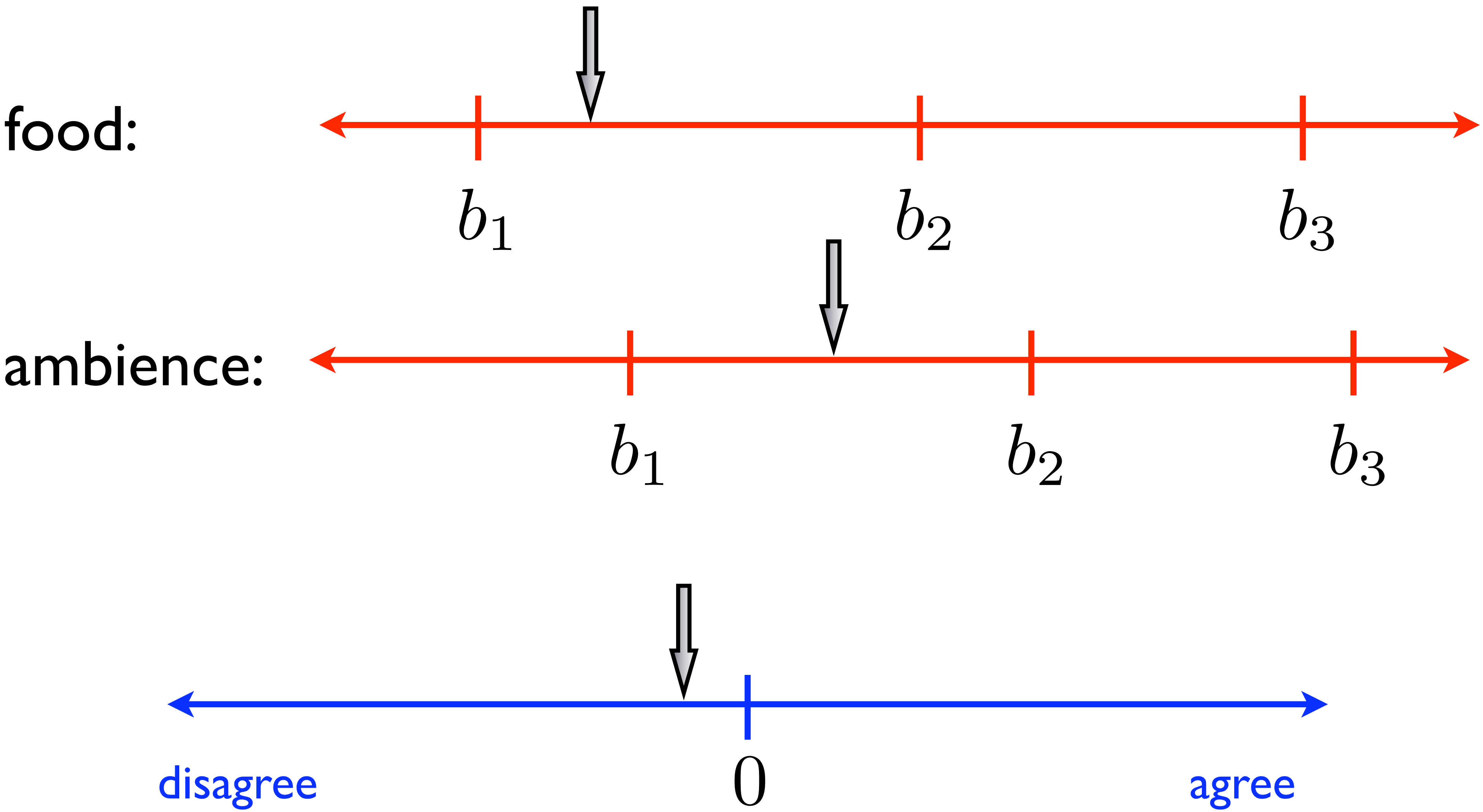
Agreement model

Disagreement with high confidence



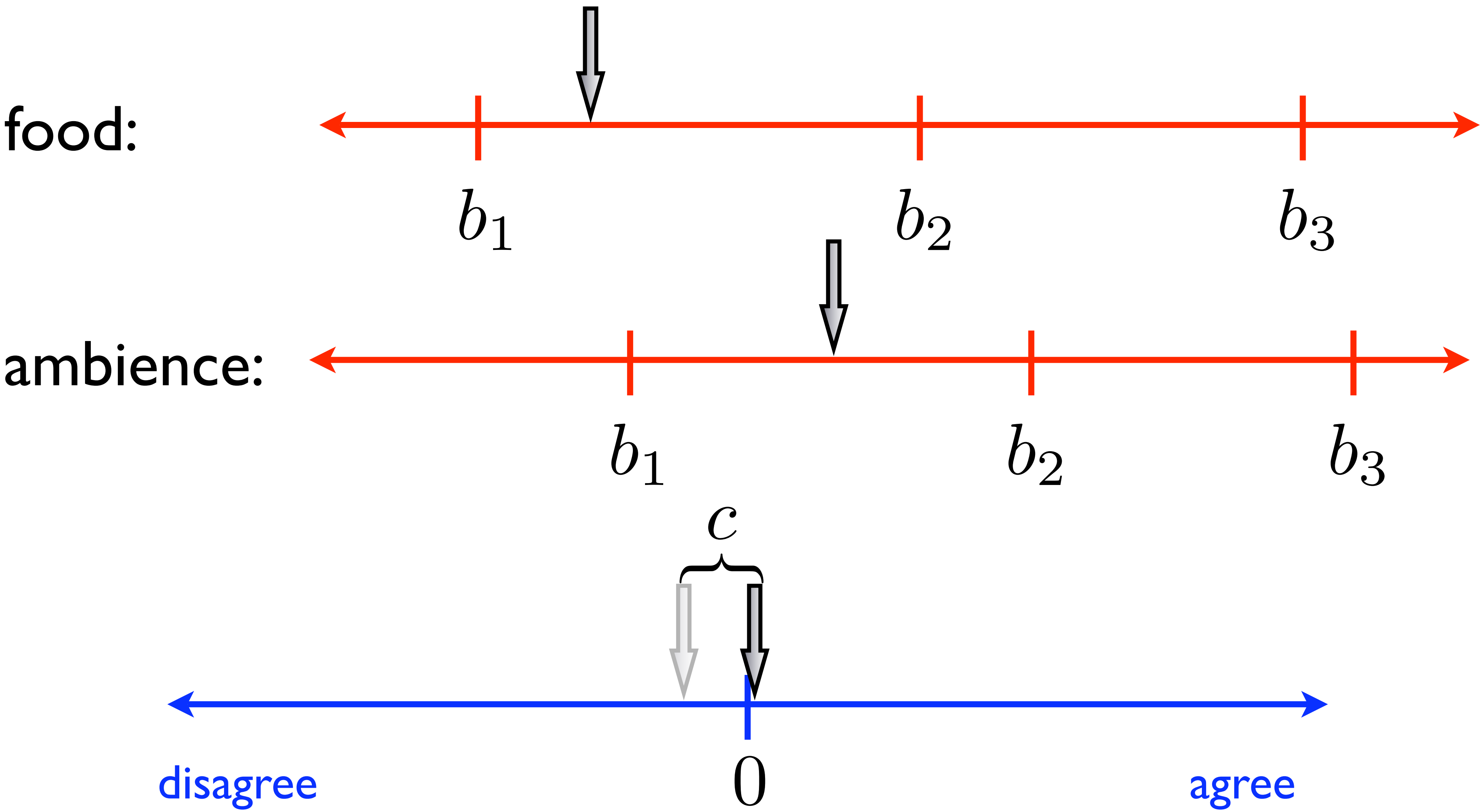
Agreement model

Disagreement with low confidence



Agreement model

Disagreement with low confidence



Agreement model

Joint Learning

Idea: *optimize individual ranker parameters for Good Grief Decoding.*

1. Train agreement model on corpus.
2. Incorporate Grief Minimization into online learning procedure for rankers:
   - Jointly decode each training instance.
   - Simultaneously update all rankers.

Joint Online Learning

Input : $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^T, \mathbf{y}^T)$, Agreement model $\mathbf{a}$, Decoder defintion $H(\mathbf{x})$.
Initialize : Set $\mathbf{w}[i]^1 = 0$, $b[i]^1_1, ..., b[i]^1_{k-1} = 0$, $b[i]^1_k = \infty$, $\forall i \in 1...m$.
Loop : For $t = 1, 2, ..., T$:

1. Get a new instance $\mathbf{x}^t \in \mathbb{R}^n$.
2. Predict $\hat{\mathbf{y}}^t = H(\mathbf{x}; \mathbf{w}^t, \mathbf{b}^t, \mathbf{a})$.
3. Get a new label $\mathbf{y}^t$.
4. For aspect $i = 1, ..., m$:

   If $\hat{y}[i]^t \neq y[i]^t$ update model:

   4.a For $r = 1, ..., k-1$: If $y[i]^t \leq r$ then $y[i]^t_r = -1$ else $y[i]^t_r = 1$.

   4.b For $r = 1, ..., k-1$: If $(\hat{y}[i]^t - r)y[i]^t_r \leq 0$ then $\tau[i]^t_r = y[i]^t_r$ else $\tau[i]^t_r = 0$.

   4.c Update $\mathbf{w}[i]^{t+1} \leftarrow \mathbf{w}[i]^t + (\sum_r \tau[i]^t_r)\mathbf{x}^t$.
   For $r = 1, ..., k-1$ update : $b[i]^{t+1}_r \leftarrow b[i]^t_r - \tau[i]^t_r$.

Output : $H(\mathbf{x}; \mathbf{w}^{T+1}, \mathbf{b}^{T+1}, \mathbf{a})$.

Update rule based on (Crammer & Singer 2001)

Joint Online Learning

Input : $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^T, \mathbf{y}^T)$, Agreement model $\mathbf{a}$, Decoder defintion $H(\mathbf{x})$.
Initialize : Set $\mathbf{w}[i]^1 = 0$, $b[i]^1_1, ..., b[i]^1_{k-1} = 0$, $b[i]^1_k = \infty$, $\forall i \in 1...m$.
Loop : For $t = 1, 2, ..., T$:

1. Get a new instance $\mathbf{x}^t \in \mathbb{R}^n$.
2. Predict $\hat{\mathbf{y}}^t = \boxed{H(\mathbf{x}; \mathbf{w}^t, \mathbf{b}^t, \mathbf{a})}$.
3. Get a new label $\mathbf{y}^t$.
4. For aspect $i = 1, ..., m$:

   If $\hat{y}[i]^t \neq y[i]^t$ update model:

   4.a For $r = 1, ..., k-1$: If $y[i]^t \leq r$ then $y[i]^t_r = -1$ else $y[i]^t_r = 1$.

   4.b For $r = 1, ..., k-1$: If $(\hat{y}[i]^t - r)y[i]^t_r \leq 0$ then $\tau[i]^t_r = y[i]^t_r$ else $\tau[i]^t_r = 0$.

   4.c Update $\mathbf{w}[i]^{t+1} \leftarrow \mathbf{w}[i]^t + (\sum_r \tau[i]^t_r)\mathbf{x}^t$.
   For $r = 1, ..., k-1$ update : $b[i]^{t+1}_r \leftarrow b[i]^t_r - \tau[i]^t_r$.

Output : $H(\mathbf{x}; \mathbf{w}^{T+1}, \mathbf{b}^{T+1}, \mathbf{a})$.

Grief Minimization

Update rule based on (Crammer & Singer 2001)

Joint Online Learning

Input : $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^T, \mathbf{y}^T)$, Agreement model $\mathbf{a}$, Decoder defintion $H(\mathbf{x})$.
Initialize : Set $\mathbf{w}[i]^1 = 0,\ b[i]^1_1, ..., b[i]^1_{k-1} = 0,\ b[i]^1_k = \infty,\ \forall i \in 1...m$.
Loop : For $t = 1, 2, ..., T$:

1. Get a new instance $\mathbf{x}^t \in \mathbb{R}^n$.
2. Predict $\hat{\mathbf{y}}^t = \boxed{H(\mathbf{x}; \mathbf{w}^t, \mathbf{b}^t, \mathbf{a})}$.
3. Get a new label $\mathbf{y}^t$.
4. $\boxed{\text{For aspect } i = 1, ..., m\text{:}}$

   If $\hat{y}[i]^t \neq y[i]^t$ update model:

   4.a For $r = 1, ..., k-1$: If $y[i]^t \leq r$ then $y[i]^t_r = -1$
   else $y[i]^t_r = 1$.

   4.b For $r = 1, ..., k-1$: If $(\hat{y}[i]^t - r)y[i]^t_r \leq 0$ then $\tau[i]^t_r = y[i]^t_r$
   else $\tau[i]^t_r = 0$.

   4.c Update $\mathbf{w}[i]^{t+1} \leftarrow \mathbf{w}[i]^t + (\sum_r \tau[i]^t_r)\mathbf{x}^t$.
   For $r = 1, ..., k-1$ update : $b[i]^{t+1}_r \leftarrow b[i]^t_r - \tau[i]^t_r$.

Grief Minimization

Simultaneous Updates

Output : $H(\mathbf{x}; \mathbf{w}^{T+1}, \mathbf{b}^{T+1}, \mathbf{a})$.

Update rule based on (Crammer & Singer 2001)

Joint Online Learning

Input : $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^T, \mathbf{y}^T)$, Agreement model $\mathbf{a}$, Decoder defintion $H(\mathbf{x})$.
Initialize : Set $\mathbf{w}[i]^1 = 0$, $b[i]^1_1, ..., b[i]^1_{k-1} = 0$, $b[i]^1_k = \infty$, $\forall i \in 1...m$.
Loop : For $t = 1, 2, ..., T$:

1. Get a new instance $\mathbf{x}^t \in \mathbb{R}^n$.
2. Predict $\hat{\mathbf{y}}^t = \boxed{H(\mathbf{x}; \mathbf{w}^t, \mathbf{b}^t, \mathbf{a})}$.
3. Get a new label $\mathbf{y}^t$.
4. $\boxed{\text{For aspect } i = 1, ..., m:}$

If $\hat{y}[i]^t \neq y[i]^t$ update model:

4.a For $r = 1, ..., k-1$: If $y[i]^t \leq r$ then $y[i]^t_r = -1$ else $y[i]^t_r = 1$.

4.b For $r = 1, ..., k-1$: If $(\hat{y}[i]^t - r)y[i]^t_r \leq 0$ then $\tau[i]^t_r = y[i]^t_r$ else $\tau[i]^t_r = 0$.

4.c Update $\mathbf{w}[i]^{t+1} \leftarrow \mathbf{w}[i]^t + (\sum_r \tau[i]^t_r)\mathbf{x}^t$.
For $r = 1, ..., k-1$ update : $b[i]^{t+1}_r \leftarrow b[i]^t_r - \tau[i]^t_r$.

Output : $H(\mathbf{x}; \mathbf{w}^{T+1}, \mathbf{b}^{T+1}, \mathbf{a})$.

Grief Minimization

Simultaneous Updates

Update rule based on (Crammer & Singer 2001)

Feature Representation

- Each review represented as binary feature vector.
  - Features track presence or absence of words and word bigrams.
  - About 30,000 total features.
- Lexical features previously found effective for:
  - Sentiment Analysis (Wiebe 2000; Pang et al 2004)
  - Discourse Analysis (Marcu & Echihabi 2002)

Evaluation

- 4,500 restaurant reviews (www.we8there.com)
- 3,500 / 500 / 500 random split into training, development, and test data.
- Average review length: 115 words.
- Each review ranks restaurant with respect to: food, service, ambience, value, and overall experience on a scale of 1-5.

Average Rank Loss : $$\sum_{t=1}^{T} \left| \hat{r}^t - r^t \right| / T$$

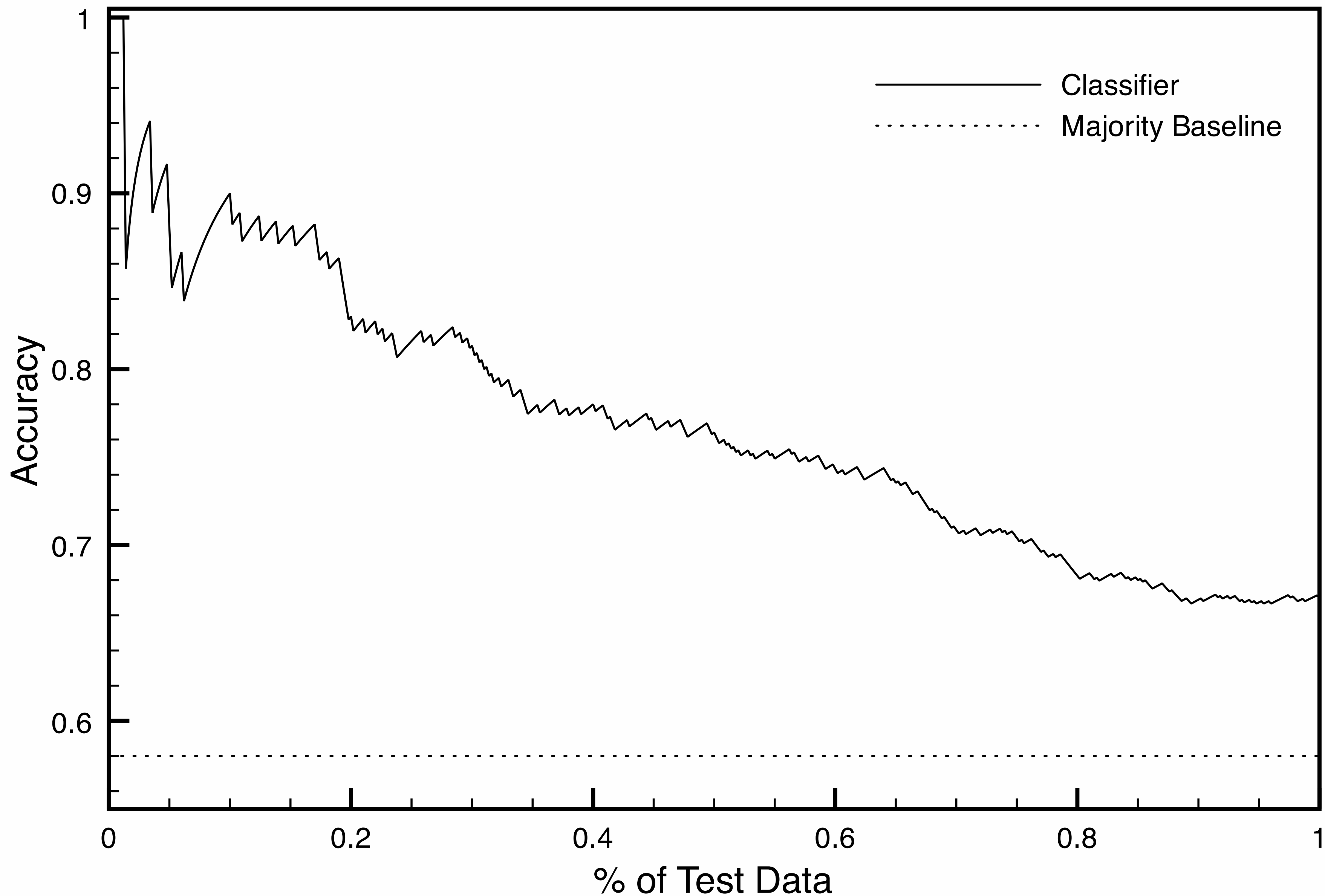
Performance of Agreement Model

- Majority Baseline (disagreement): 58%
- Agreement Model accuracy: 67%
- <u>According to Good Grief Criterion:</u>

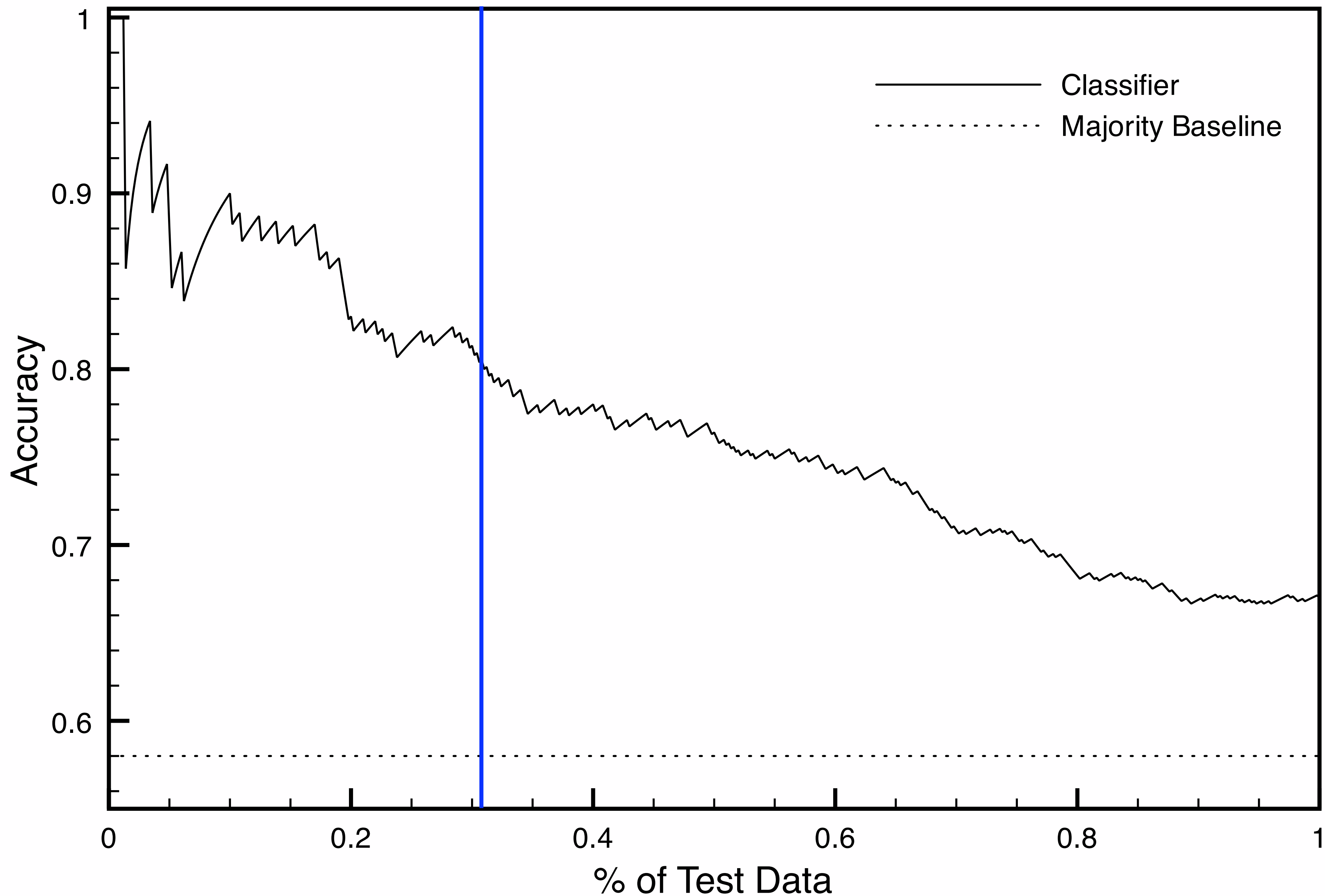
  Raw accuracy not what matters, rather *accuracy as function of confidence*.

  - As confidence goes up, so does accuracy

Accuracy of Agreement Model

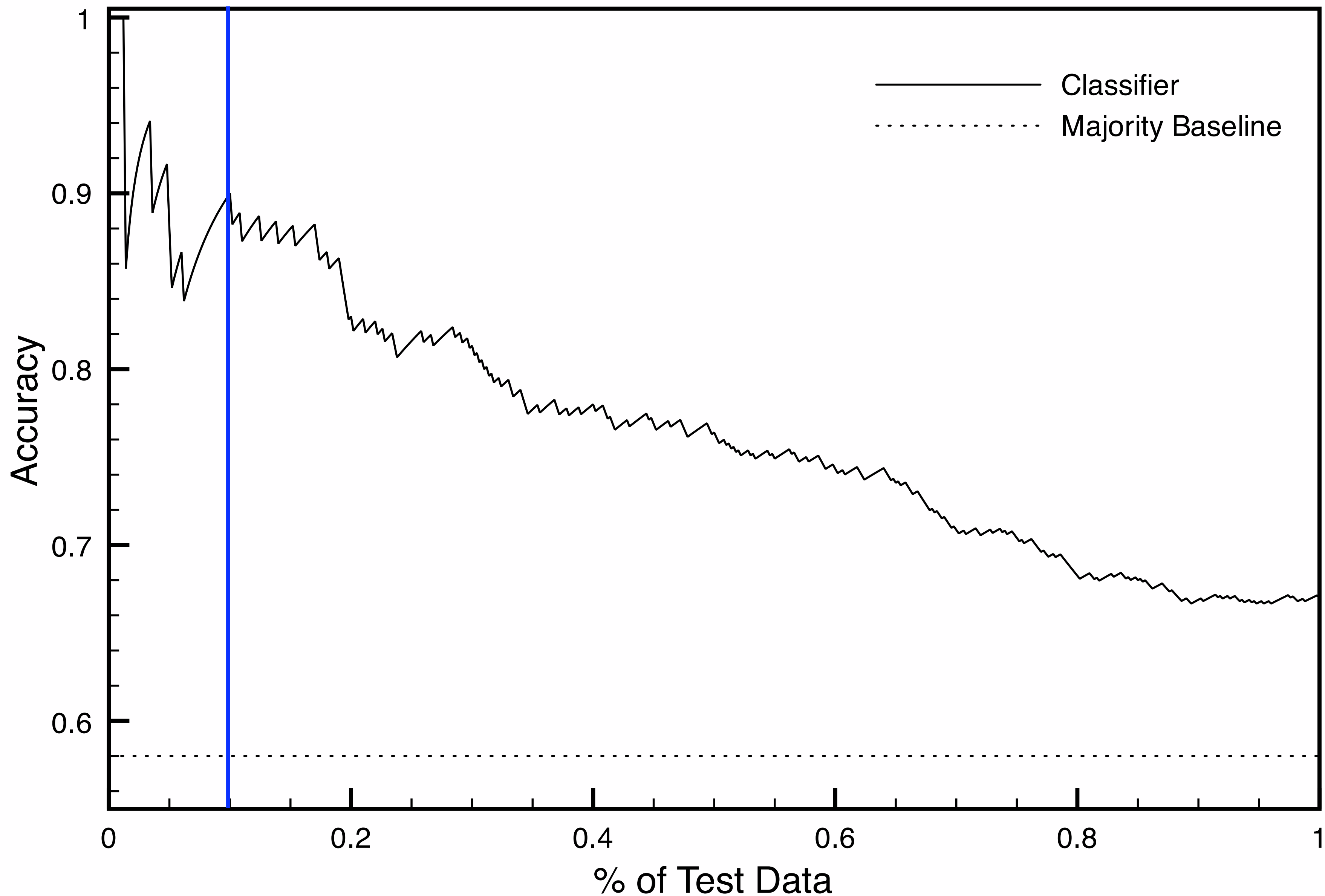


Accuracy of Agreement Model



- 33% of data with highest confidence classified at 80% accuracy.

Accuracy of Agreement Model



- 10% of data with highest confidence classified at 90% accuracy.

Baselines

- PRANK: Independent rankers for each aspect trained using PRank algorithm (Crammer & Singer 2001)
- MAJORITY: $< 5, 5, 5, 5, 5 >$
- SIM: Joint model using cosine similarity between aspects (Basilico & Hofmann 2004)

$$score_i(\mathbf{x}) = \mathbf{w}[i] \cdot \mathbf{x} + \sum_j sim(i,j)(\mathbf{w}[j] \cdot \mathbf{x})$$

- GG DECODE: “Good Grief” decoding but independent training

Average Rank Loss on Test Set

	Food	Service	Value	Atmosphere	Experience	**Total**
MAJORITY	0.848	1.056	1.030	1.044	1.028	1.001
PRANK	0.606	0.676	0.700	0.776	0.618	0.675
SIM	0.562	0.648	0.706	0.798	0.600	0.663
GG DECODE	0.544	0.648	0.704	0.798	0.584*	0.656
GG train+decode	**0.534***	**0.622***	**0.644***	**0.774***	**0.584**	**0.632***

* = Statistically Significant improvement over closest rival using Fisher Sign Test.

Average Rank Loss

	Agreement	Disagreement
PRANK	0.414	0.864
GG TRAIN+DECODE	0.324	0.854

- Cases of Disagreement:
  - 58% of corpus
  - relative reduction in error: 1%
- Cases of Agreement:
  - 42% of corpus
  - relative reduction in error: 22%

Technical Contributions

- <u>Novel framework</u> for tasks with complex label dependencies:
  - simple, fast, exact, and accurate
- <u>Explicit Meta-Model</u>:
  - task-specific dependency spaces
  - features tailored for dependency prediction
  - joint Perceptron updates for label predictors

Conclusions & Future Work

- Applied Good Grief framework to Multiple Aspect Sentiment Analysis:
  - Agreement Model guides aspect rank predictions
- Outperform all baseline models.
- Future Work: apply GG framework to other tasks
  - classification, regression etc
  - more complex label dependency spaces

Data and Code available:

`http://people.csail.mit.edu/bsnyder`

Thank You!

Model Expressivity

Model Expressivity

$\Rightarrow$ $\boxed{\text{Ambience}}$

- <u>Fully Implicit Opinions:</u>

  "Our only complaint was the service after our order was taken"

Model Expressivity

$\Rightarrow$ Ambience 

- <u>Fully Implicit Opinions:</u>

  "Our only complaint was the service after our order was taken"

Model Expressivity

$\Rightarrow$ $\boxed{\text{Ambience}}$

- Fully Implicit Opinions:

  "Our only complaint was the service after our order was taken"

- One Opinion expressed in terms of another:

  "The food was good, but not the ambience"

Model Expressivity

$\Rightarrow$ Ambience 

- Fully Implicit Opinions:

  "Our only complaint was the service after our order was taken"

- One Opinion expressed in terms of another:

  "The food was good, but not the ambience"

Decoding

“The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight.”

Food

Ambience

Agreement

Decoding

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."

Food

Ambience

Agreement

Decoding

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."

Food

Ambience

Agreement

Decoding

“The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight.”

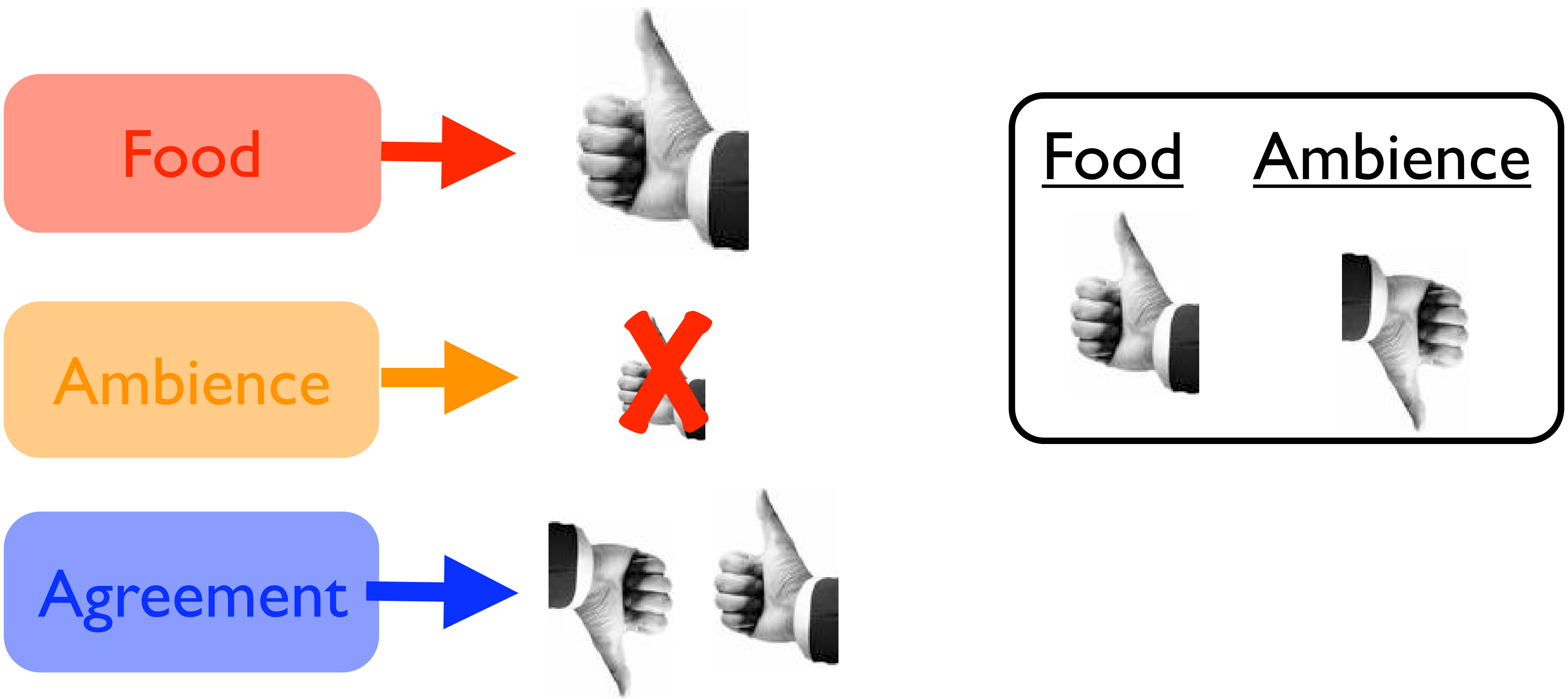
Food

Ambience

Agreement

Decoding

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."



Online Training

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."

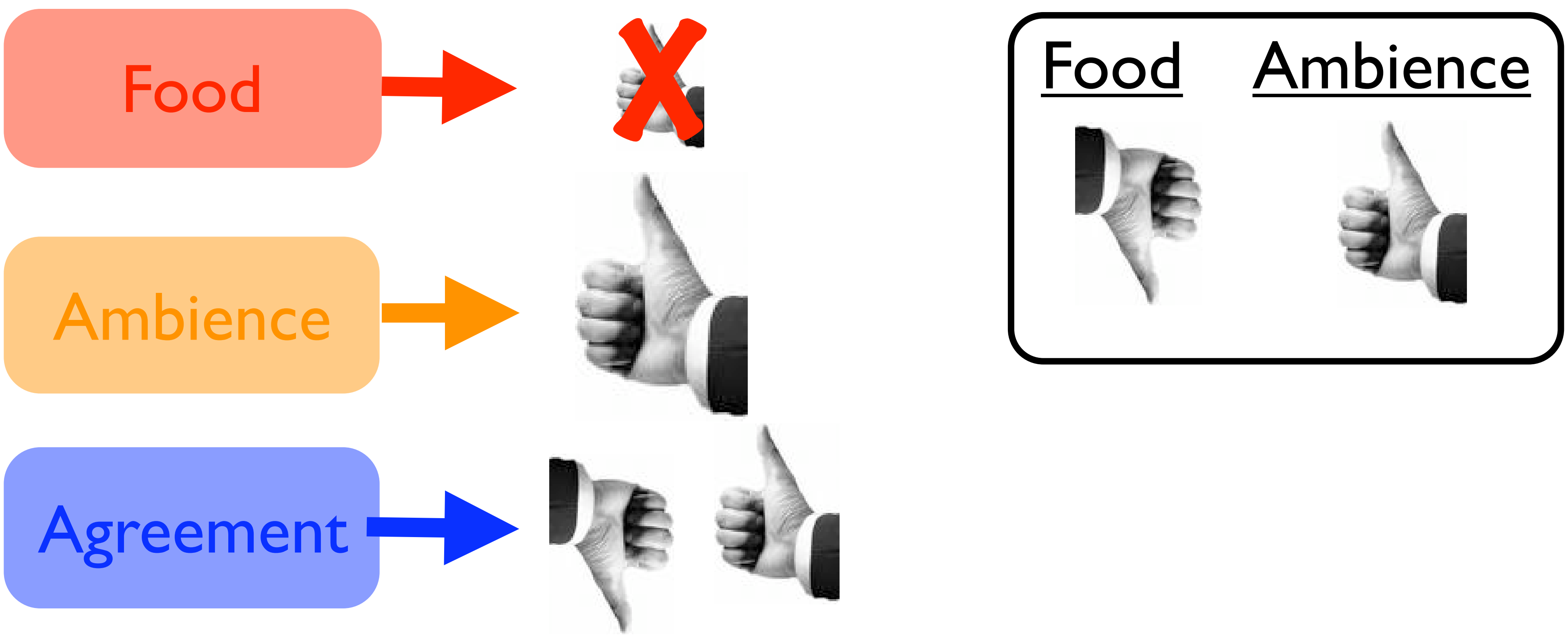
Food

Ambience

Agreement

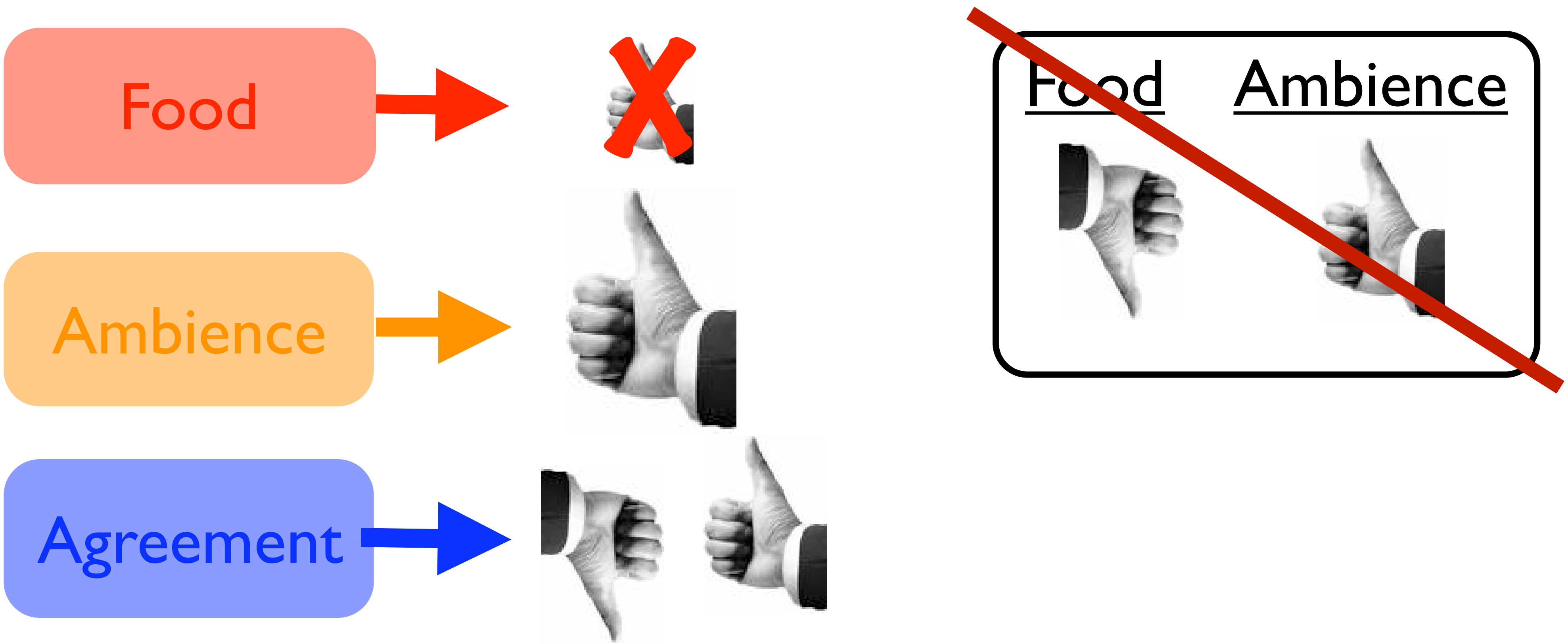
Online Training

“The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight.”



Online Training

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."



Online Training

"The restaurant was a bit uneven. Although the food was tasty, the window curtains blocked out all sunlight."