

# Predictive Analysis: Evaluation and Experimentation

Jaime Arguello  
INLS 613: Text Data Mining  
[jarguell@email.unc.edu](mailto:jarguell@email.unc.edu)

# Tips for Final Project

- Focus on a task that is interesting and useful from a theoretical or practical perspective
- Explore different feature representations rather than different learning algorithms (e.g., feature ablation analysis)
- Assume that a learning algorithm (or clustering algorithm) will capitalize on the most obvious and uninteresting type of evidence

# Evaluation and Experimentation

- Evaluation Metrics
- Cross-Validation
- Significance Tests

# Evaluation

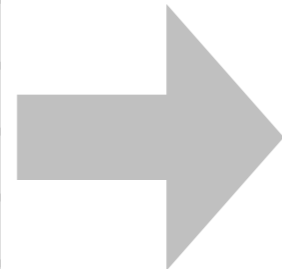
- **Predictive analysis:** training a model to make predictions on previously unseen data
- **Evaluation:** using previously unseen labeled data to estimate the quality of a model's predictions on new data
- **Evaluation Metric:** a measure that summarizes the quality of a model's predictions

# Predictive Analysis

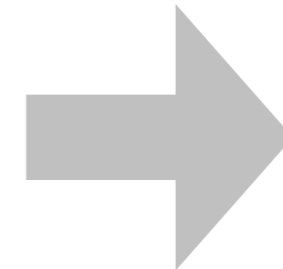
training

w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>	w <sub>4</sub>	w <sub>5</sub>	w <sub>6</sub>	w <sub>7</sub>	w <sub>8</sub>	w <sub>9</sub>	w <sub>10</sub>	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

labeled examples



machine learning algorithm

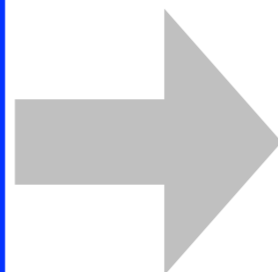


model

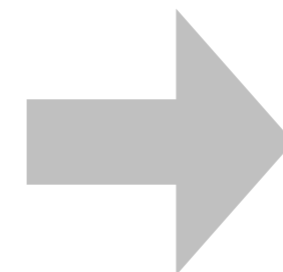
testing

w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>	w <sub>4</sub>	w <sub>5</sub>	w <sub>6</sub>	w <sub>7</sub>	w <sub>8</sub>	w <sub>9</sub>	w <sub>10</sub>	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

new, labeled examples



model



w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>	w <sub>4</sub>	w <sub>5</sub>	w <sub>6</sub>	w <sub>7</sub>	w <sub>8</sub>	w <sub>9</sub>	w <sub>10</sub>	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

predictions

# Evaluation Metrics

- There are many different metrics
- Different metrics make different assumptions about what end users care about
- Choosing the most appropriate metric is important!

# Evaluation Metrics

## (1) accuracy

- **Accuracy:** percentage of correct predictions

		true	
		pos	neg
predicted	pos	a	b
	neg	c	d

$$A = \frac{(a + d)}{(a + b + c + d)}$$

# Evaluation Metrics

## (1) accuracy

- **Accuracy:** percentage of correct predictions

		true		
		pos	neut.	neg
predicted	pos	a	b	c
	neut.	d	e	f
	neg	g	h	i

$$A = \frac{(a + e + i)}{(a + b + c + d + e + f + g + h + i)}$$



# Evaluation Metrics

## (1) accuracy

- What assumption(s) does accuracy make?

		true		
		pos	neut.	neg
predicted	pos	a	b	c
	neut.	d	e	f
	neg	g	h	i

$$A = \frac{(a + e + i)}{(a + b + c + d + e + f + g + h + i)}$$

# Evaluation Metrics

- Content recommendation: relevant vs. non-relevant



# Evaluation Metrics

- Email spam filtering: spam vs. ham

From	Subject	Date Received	Categories
▼ SUNDAY			
audio@DesktopTrainingOnline.com	Adobe Acrobat Pro: Instructor-Led Training t...	Sun 9/30/12 5:19 PM	Junk
▼ THURSDAY			
ei-sci@ei-sci.org	SCI-EI期刊检索、收录 (ICIEEE 2013) 邀请函	Thu 9/27/12 2:50 AM	Junk
▼ WEDNESDAY			
The New York Times	Act now to receive FREE digital access PLUS 5...	Wed 9/26/12 3:49 PM	Junk
Citrix Systems	Give people the freedom to work anyplace	Wed 9/26/12 1:20 PM	Junk
▼ LAST WEEK			
audio@DesktopTrainingOnline.com	Excel 2007/2010 Formatting & Customizing...	Mon 9/24/12 8:24 PM	Junk
Vonage	Last Chance: Unlimited calls with Vonage Basi...	Mon 9/24/12 2:56 PM	Junk
conference EDM	World's Tallest Tower in Tokyo - Join 2013 E...	Thu 9/20/12 10:48 PM	Junk
▼ 2 WEEKS AGO			
Jim Davidson & Strategic Investment	Washington Insider Comes out of the Shadow...	Tue 9/18/12 12:02 PM	Junk
audio@supertrainme.com	Student Record Retention: Secure Data, Maint...	Tue 9/18/12 6:56 AM	Junk
audio@DesktopTrainingOnline.com	Mastering Excel 2007/2010 Charts: Tips & Tri...	Thu 9/13/12 8:31 PM	Junk
▼ 3 WEEKS AGO			
Vonage	Get Unlimited Calling with Vonage Basic Talk...	Fri 9/7/12 2:41 PM	Junk
prof_qian	[EI SCOPUS ISI Journal, Beijing, China]Internati...	Fri 9/7/12 1:32 PM	Junk

# Evaluation Metrics

- Product reviews: positive vs. negative vs. neutral

twitter



# Evaluation Metrics

- Text-based Forecasting: buy vs. sell vs. hold

twitter



# Evaluation Metrics

- Health monitoring system: alarm vs. no alarm



# Evaluation Metrics

## (1) accuracy

- What assumption(s) does accuracy make?
- It assumes that all prediction errors are equally bad
- Oftentimes, we care more about one class than the others
- If so, the class of interest is usually the minority class
- We are looking for the “needles in the haystack”
- In this case, accuracy is not a good evaluation metric
- There are metrics that provide more insight into per-class performance

# Evaluation Metrics

(2) precision and (3) recall

- For a given class **C**:
  - ▶ **precision**: the percentage of positive predictions that are truly positive
  - ▶ **recall**: the percentage of true positives that are correctly predicted positive



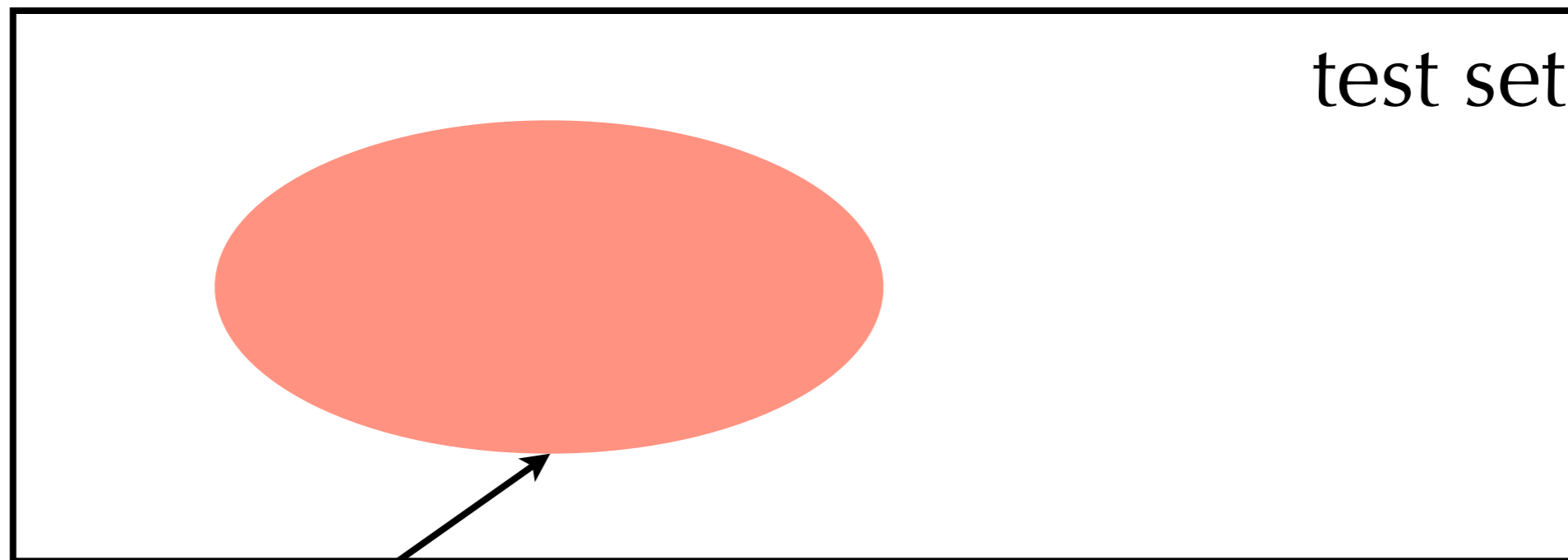
# Evaluation Metrics

(2) precision and (3) recall



# Evaluation Metrics

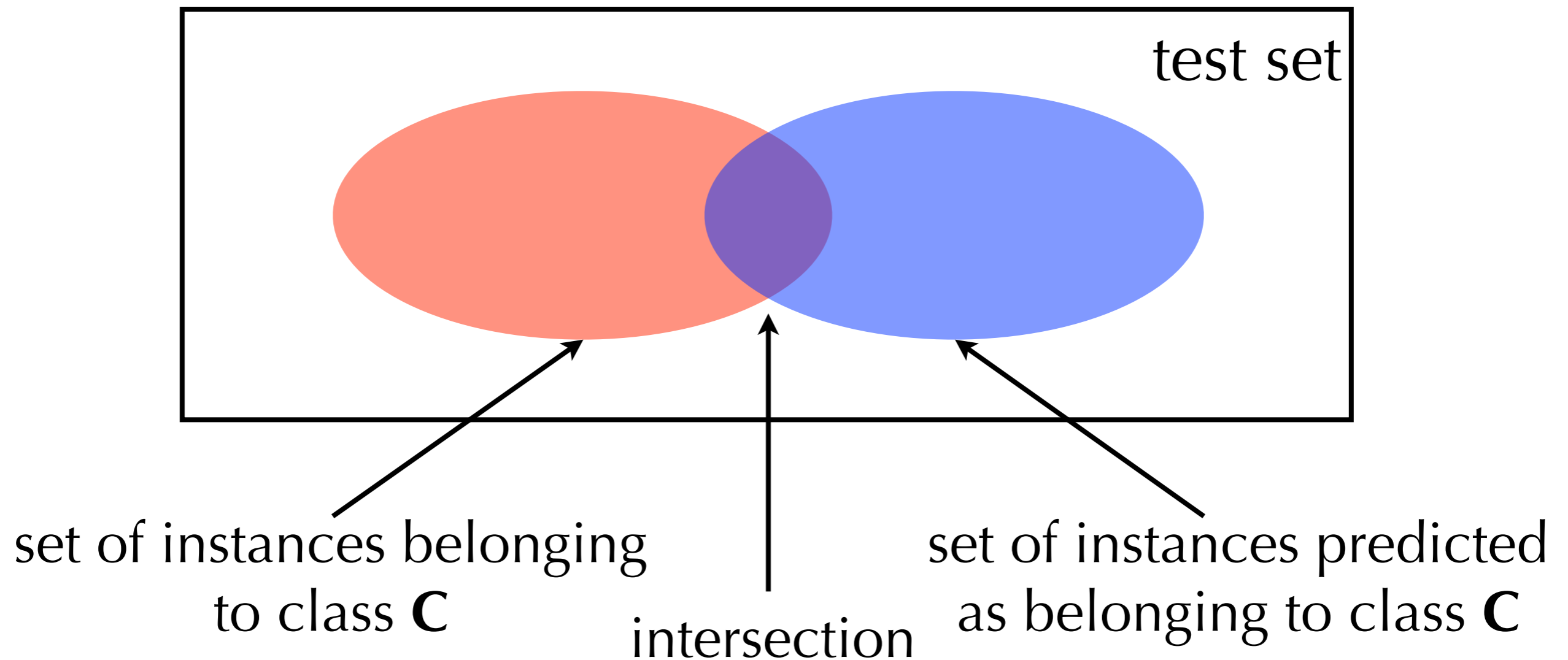
(2) precision and (3) recall



set of instances belonging  
to class **C**

# Evaluation Metrics

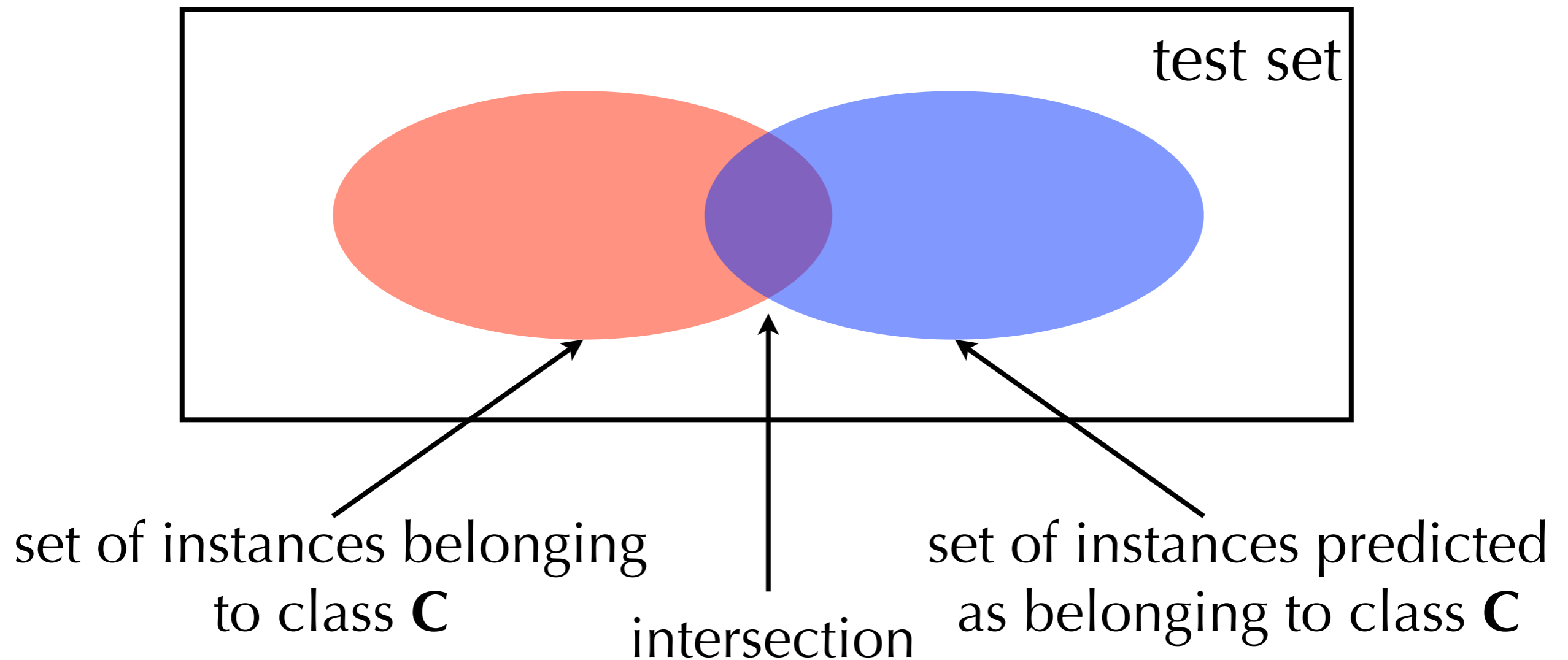
(2) precision and (3) recall



# Evaluation Metrics

(2) precision and (3) recall

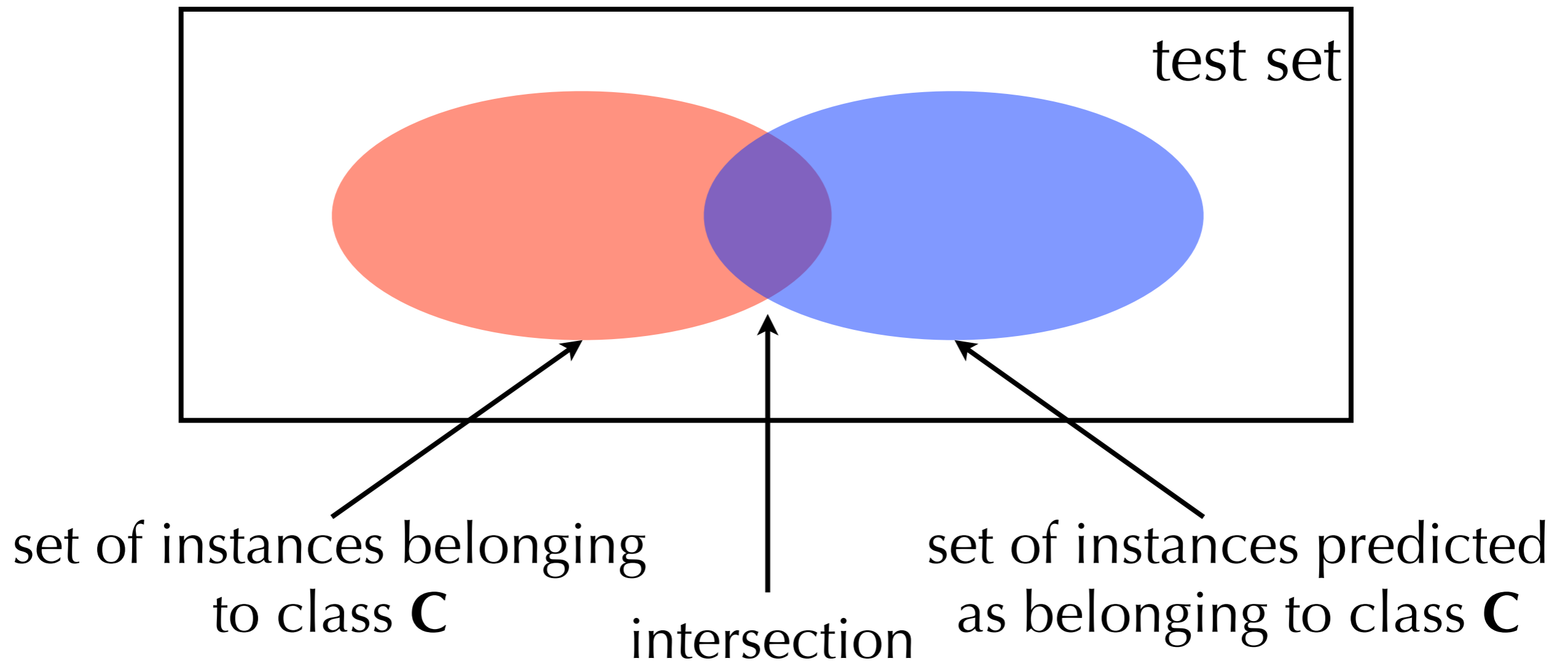
- Precision = ?



# Evaluation Metrics

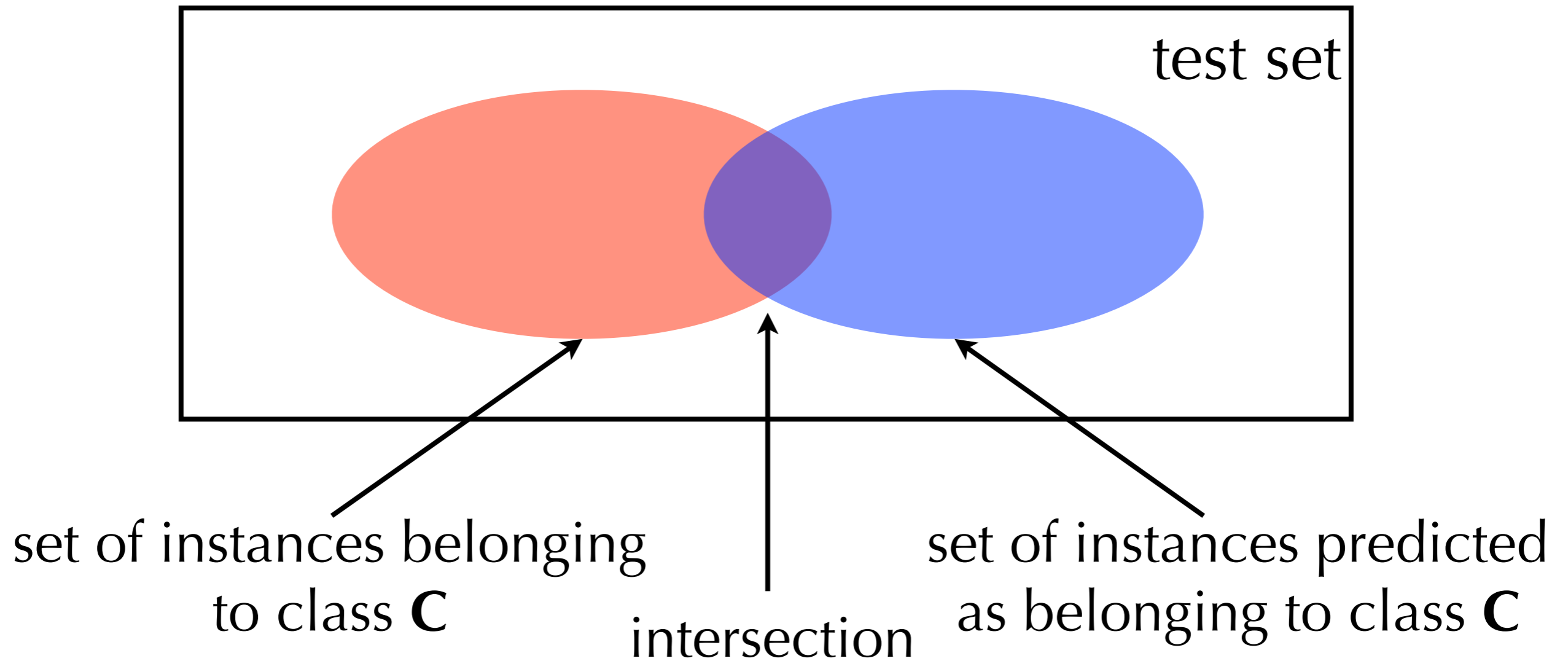
(2) precision and (3) recall

- Recall = ?



# Evaluation Metrics

(2) precision and (3) recall



# Evaluation Metrics

## (2) precision

predicted

		true		
		pos	neut.	neg
predicted	pos	a	b	c
	neut.	d	e	f
	neg	g	h	i

$$\mathcal{P}_{\text{positive}} = \frac{a}{a + b + c}$$

# Evaluation Metrics

(3) recall

predicted

		true		
		pos	neut.	neg
predicted	pos	a	b	c
	neut.	d	e	f
	neg	g	h	i

$$\mathcal{R}_{\text{positive}} = \frac{a}{a + d + g}$$



# Evaluation Metrics

precision vs. recall

predicted

		true		
		pos	neut.	neg
predicted	pos	a	b	c
	neut.	d	e	f
	neg	g	h	i

# Evaluation Metrics

## (4) f-measure

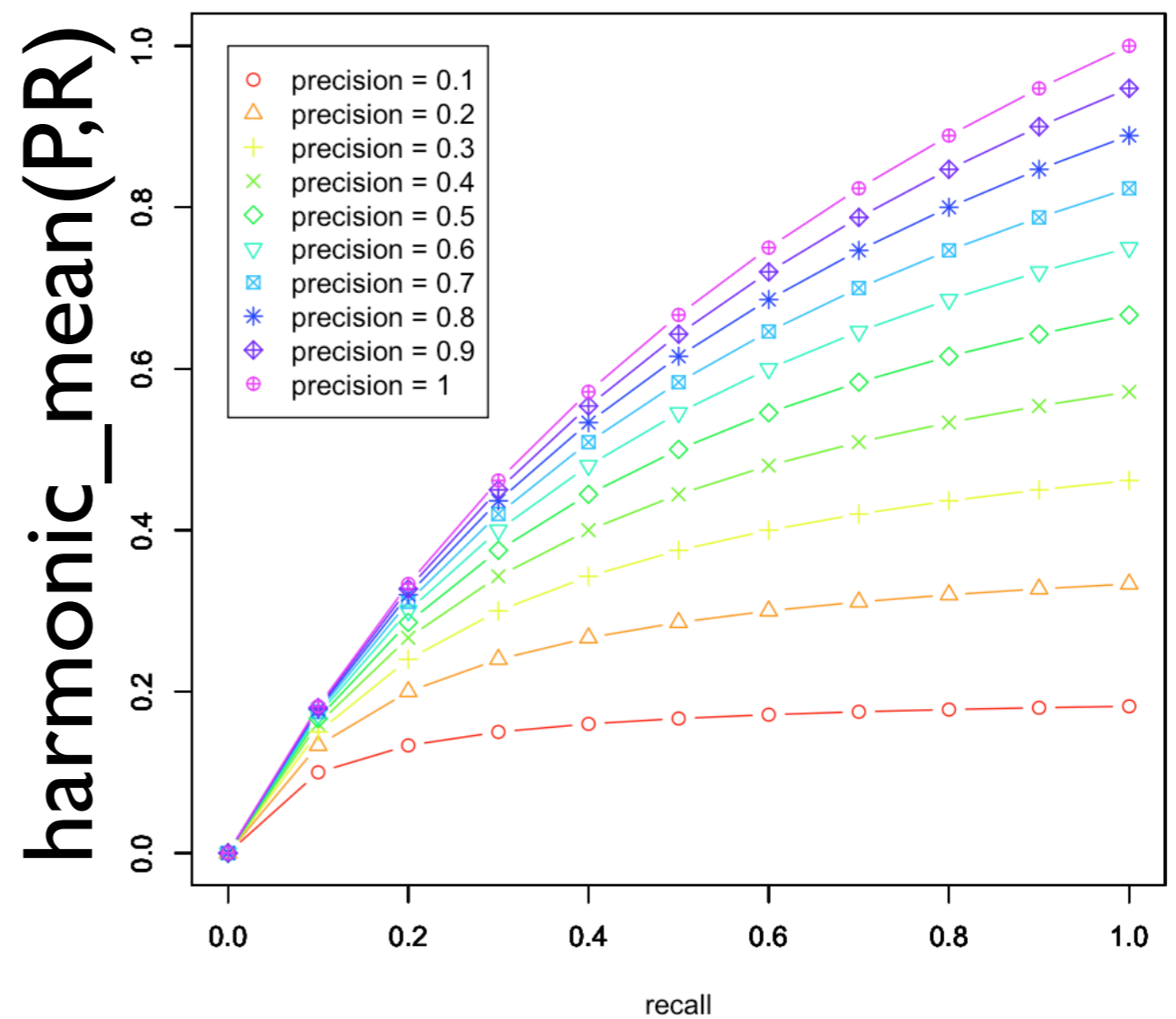
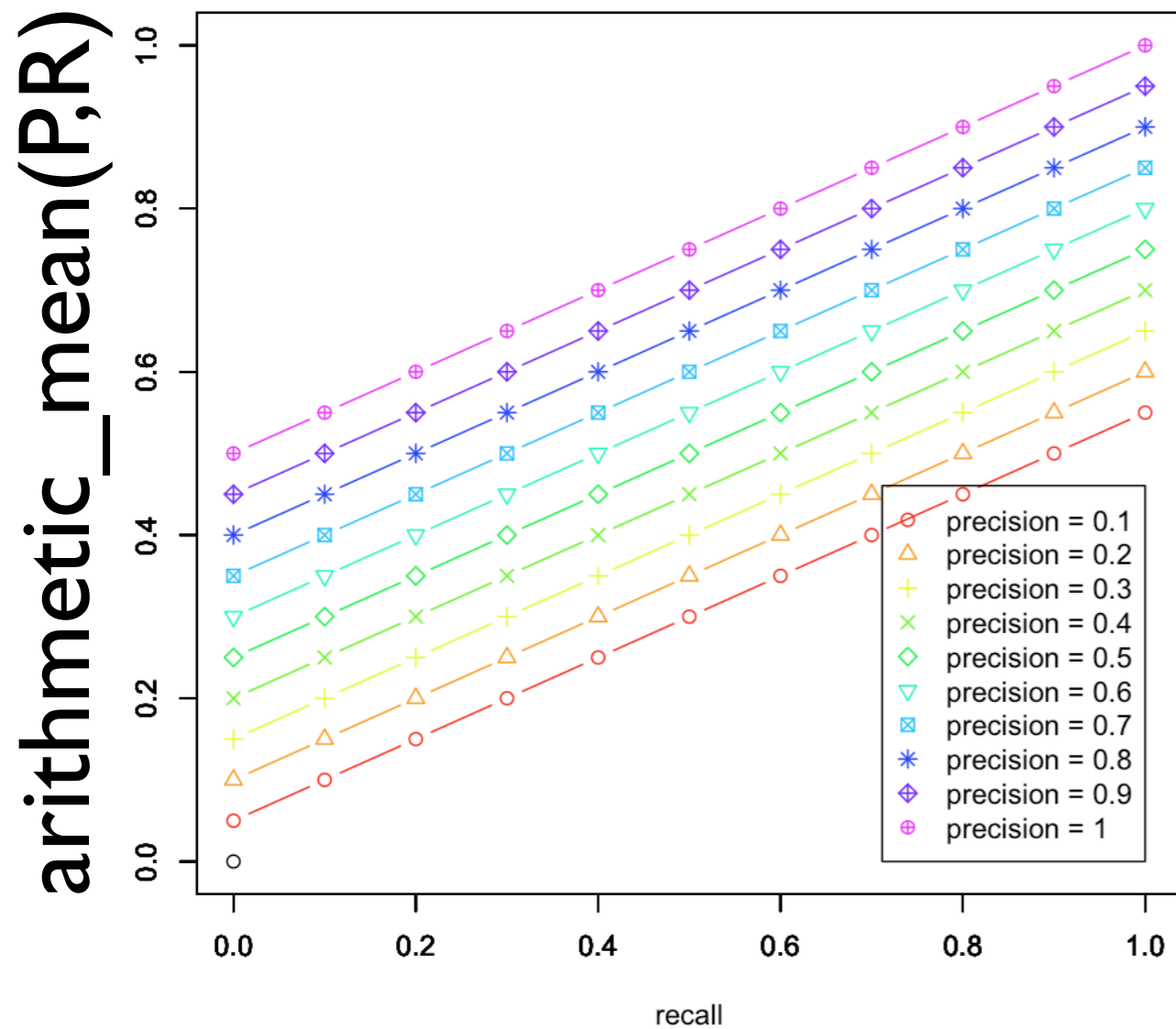
- **F-measure**: the harmonic (not arithmetic) mean of precision and recall

$$\mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$$

# Evaluation Metrics

## (4) f-measure

- **F-measure:** the harmonic (not arithmetic) mean of precision and recall



(slide courtesy of Ben Carterette)

# Evaluation Metrics

## (5) precision-recall curves

- **F-measure:** assumes that the “end users” care equally about precision and recall



# Evaluation Metrics

## (5) precision-recall curves

- Most machine-learning algorithms provide a prediction confidence value
- The prediction confidence value can be used as a threshold in order to trade-off precision and recall

# Evaluation Metrics

## (5) precision-recall curves

- Remember Naive Bayes classification?
- Given instance  $D$ , predict positive (**POS**) if:

$$P(POS|D) \geq P(NEG|D)$$

- Otherwise, predict negative (**NEG**)

# Evaluation Metrics

## (5) precision-recall curves

- Remember Naive Bayes classification?
- Given instance  $D$ , predict positive (**POS**) if:

$$P(\text{POS}|D) \geq P(\text{NEG}|D)$$

- Otherwise, predict negative (**NEG**)

this value can be used as a threshold for classification into the **POS** class

# Evaluation Metrics

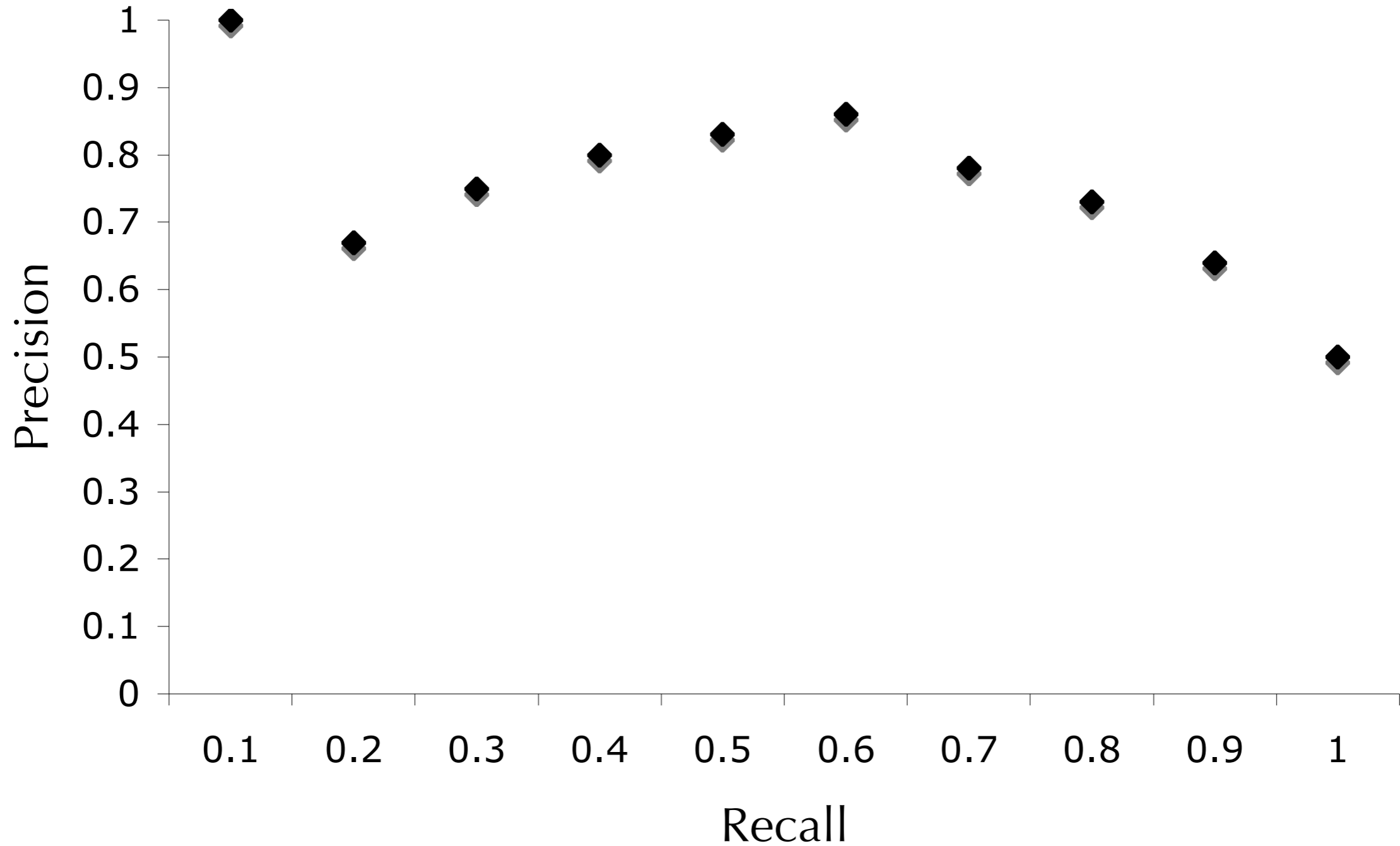
## (5) precision-recall curves

rank (K)	ranking	P(POS D)	P@K	R@K
1	█	0.99	1.00	0.10
2		0.87	0.50	0.10
3	█	0.84	0.67	0.20
4	█	0.83	0.75	0.30
5	█	0.77	0.80	0.40
6	█	0.63	0.83	0.50
7	█	0.58	0.86	0.60
8		0.57	0.75	0.60
9	█	0.56	0.78	0.70
10		0.34	0.70	0.70
11	█	0.33	0.73	0.80
12		0.25	0.67	0.80
13		0.21	0.62	0.80
14	█	0.15	0.64	0.90
15		0.14	0.60	0.90
16		0.14	0.56	0.90
17		0.12	0.53	0.90
18		0.08	0.50	0.90
19		0.01	0.47	0.90
20	█	0.01	0.50	1.00



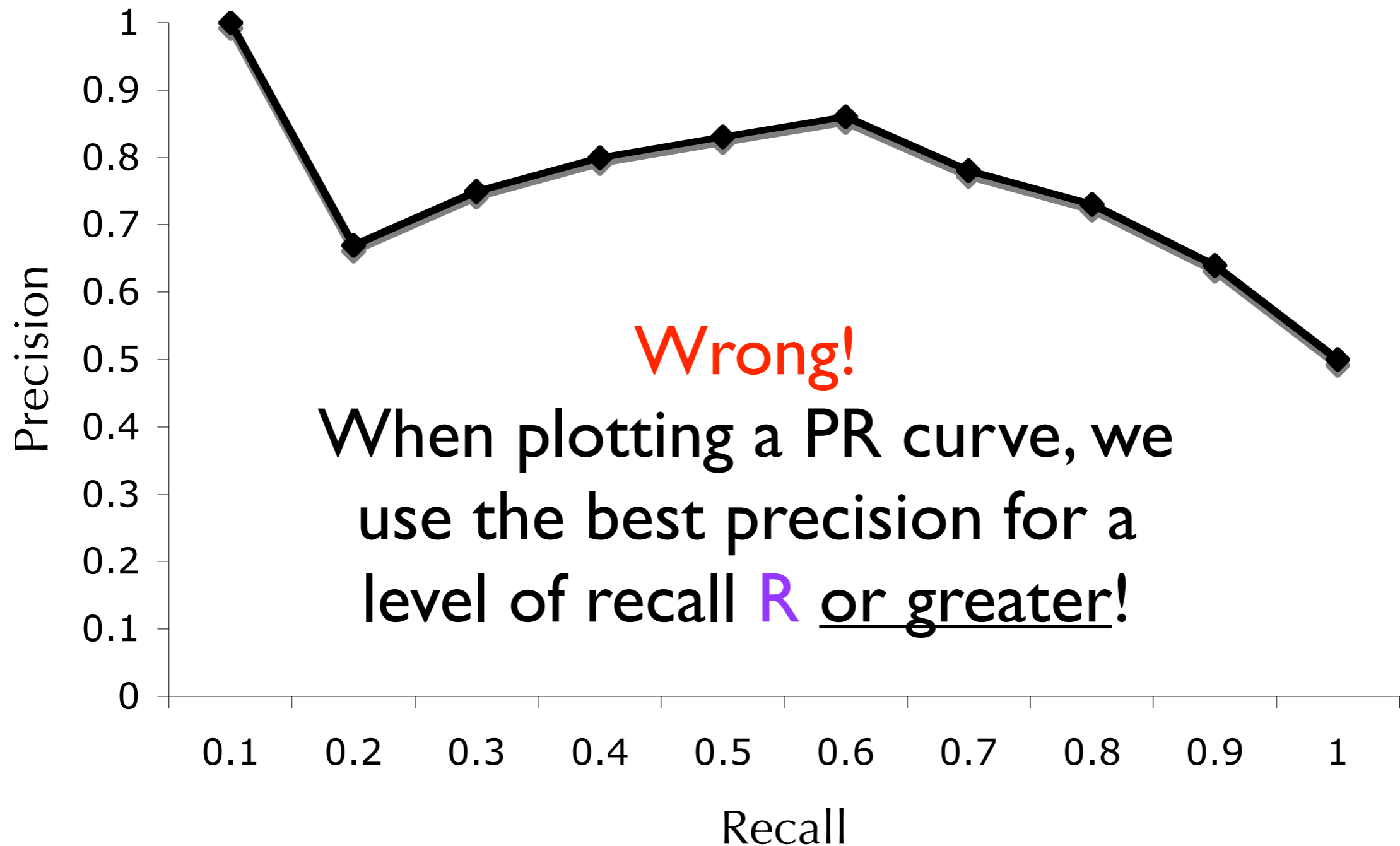
# Evaluation Metrics

(5) precision-recall curves



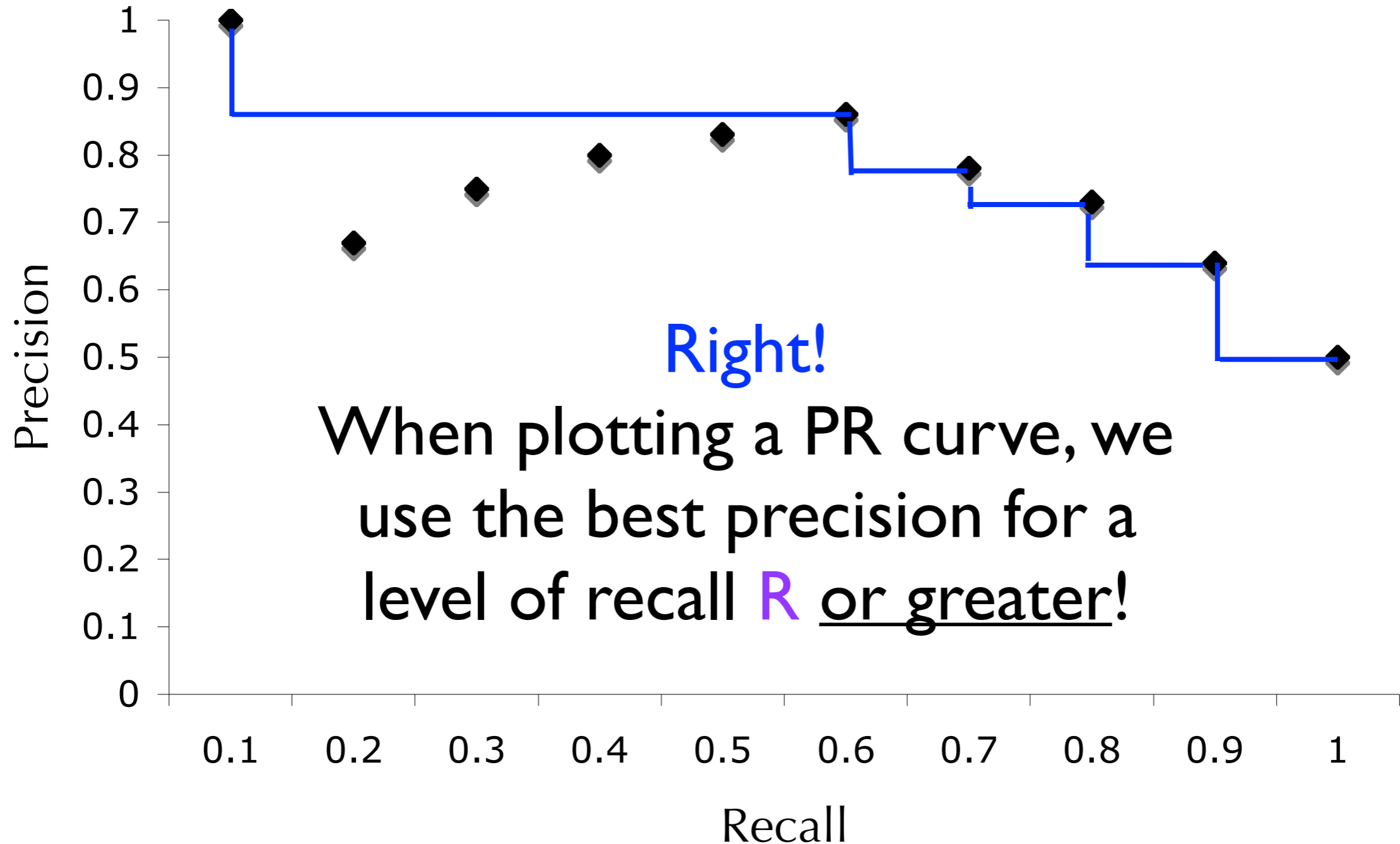
# Evaluation Metrics

(5) precision-recall curves



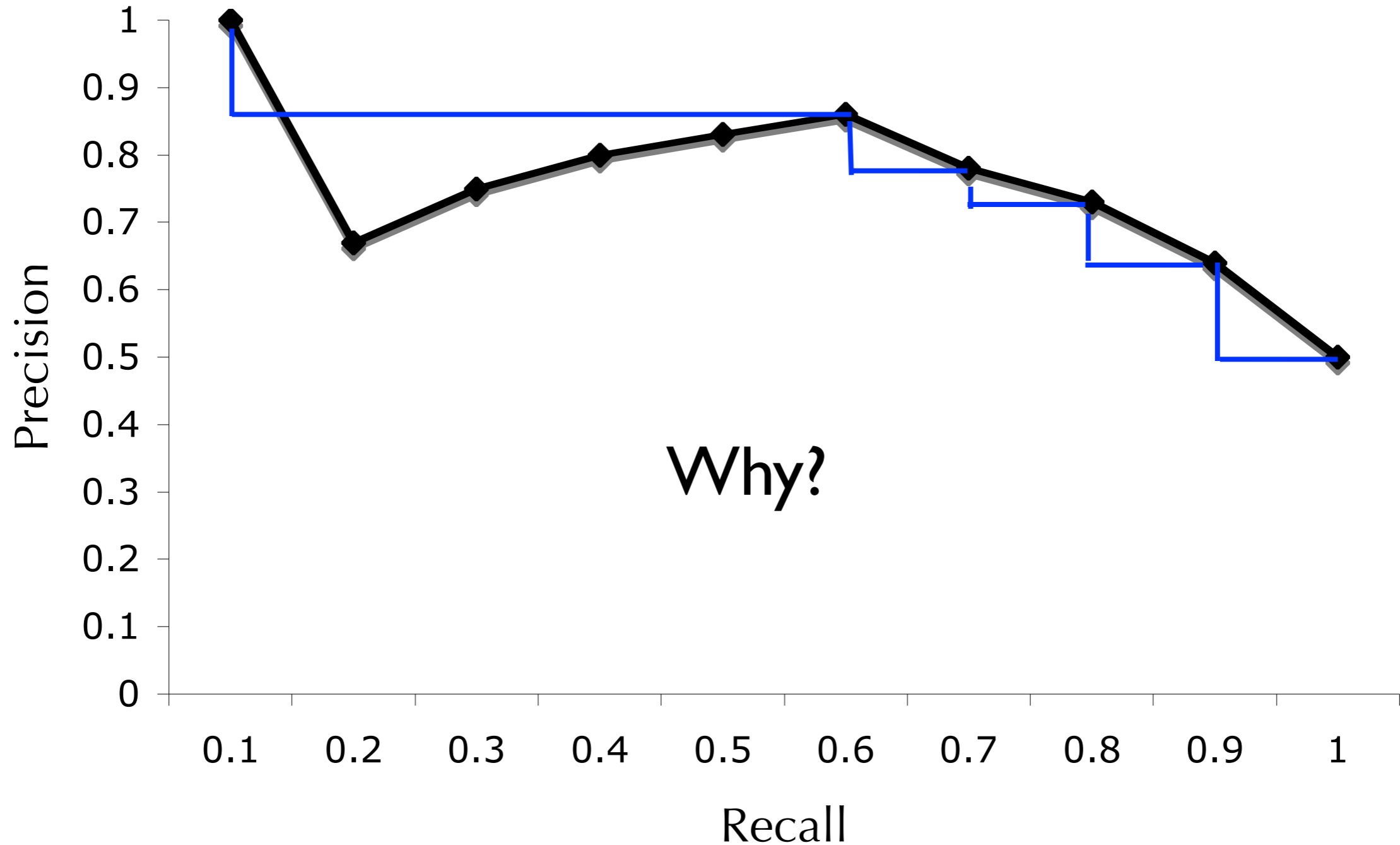
# Evaluation Metrics

(5) precision-recall curves



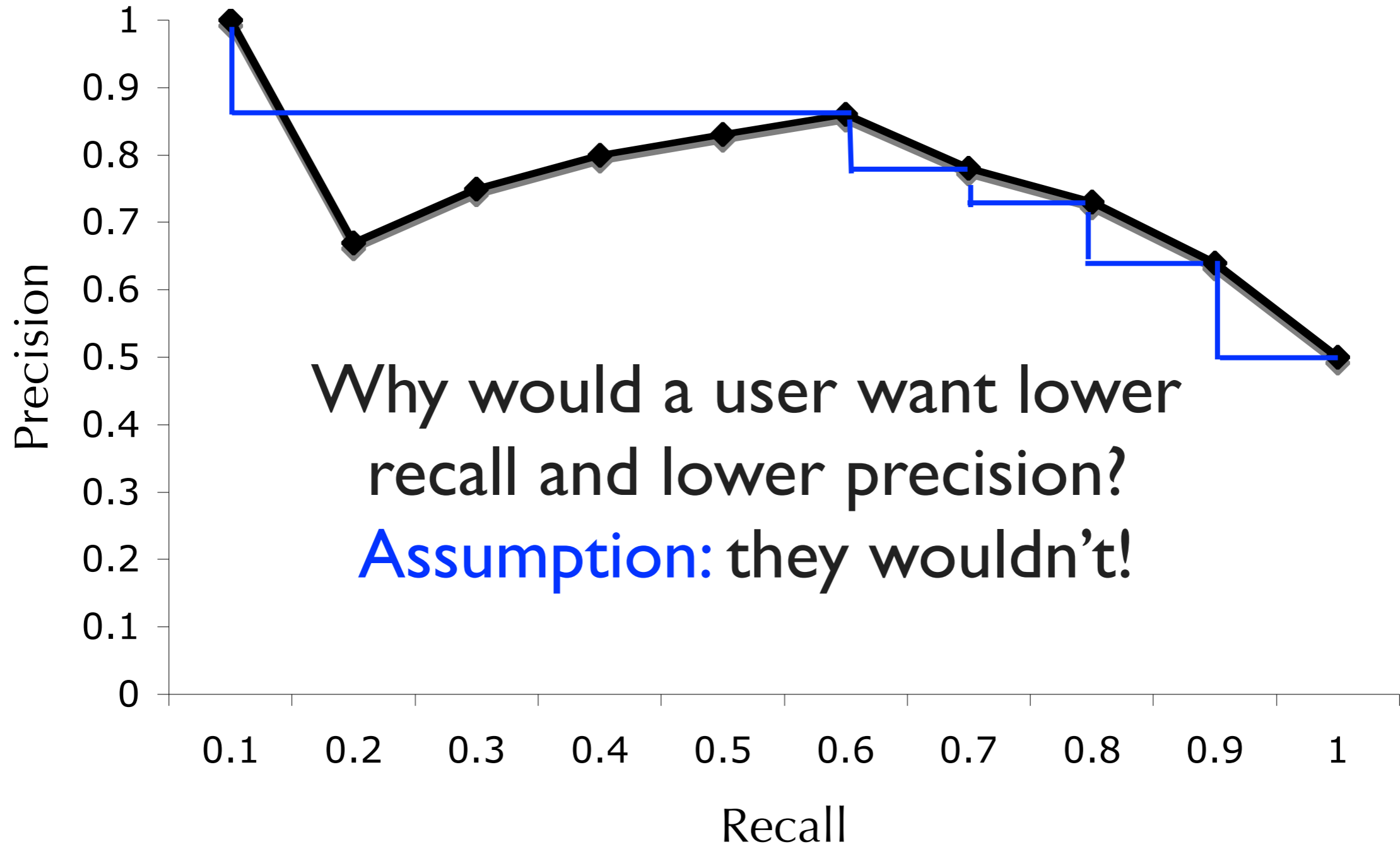
# Evaluation Metrics

(5) precision-recall curves



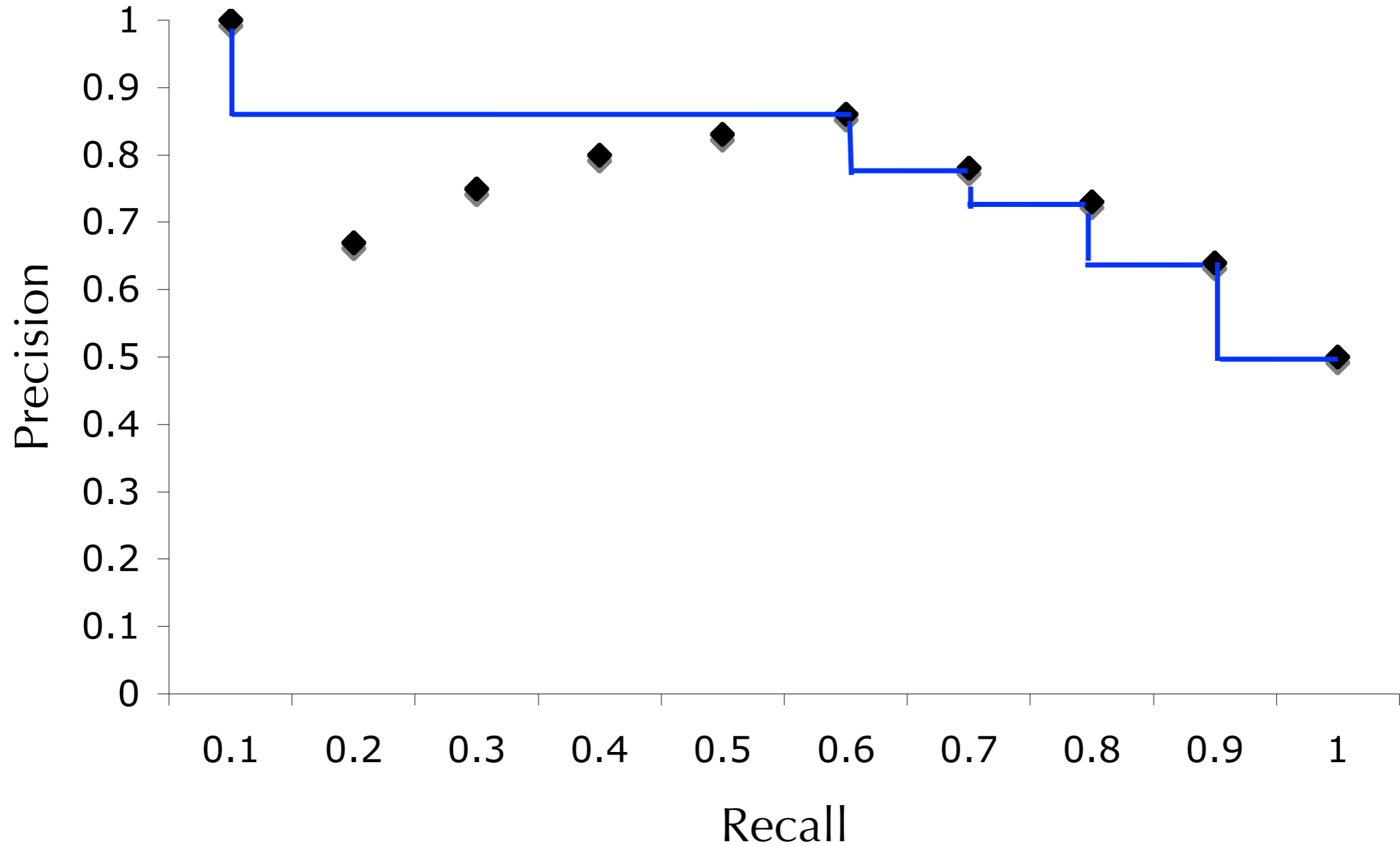
# Evaluation Metrics

(5) precision-recall curves



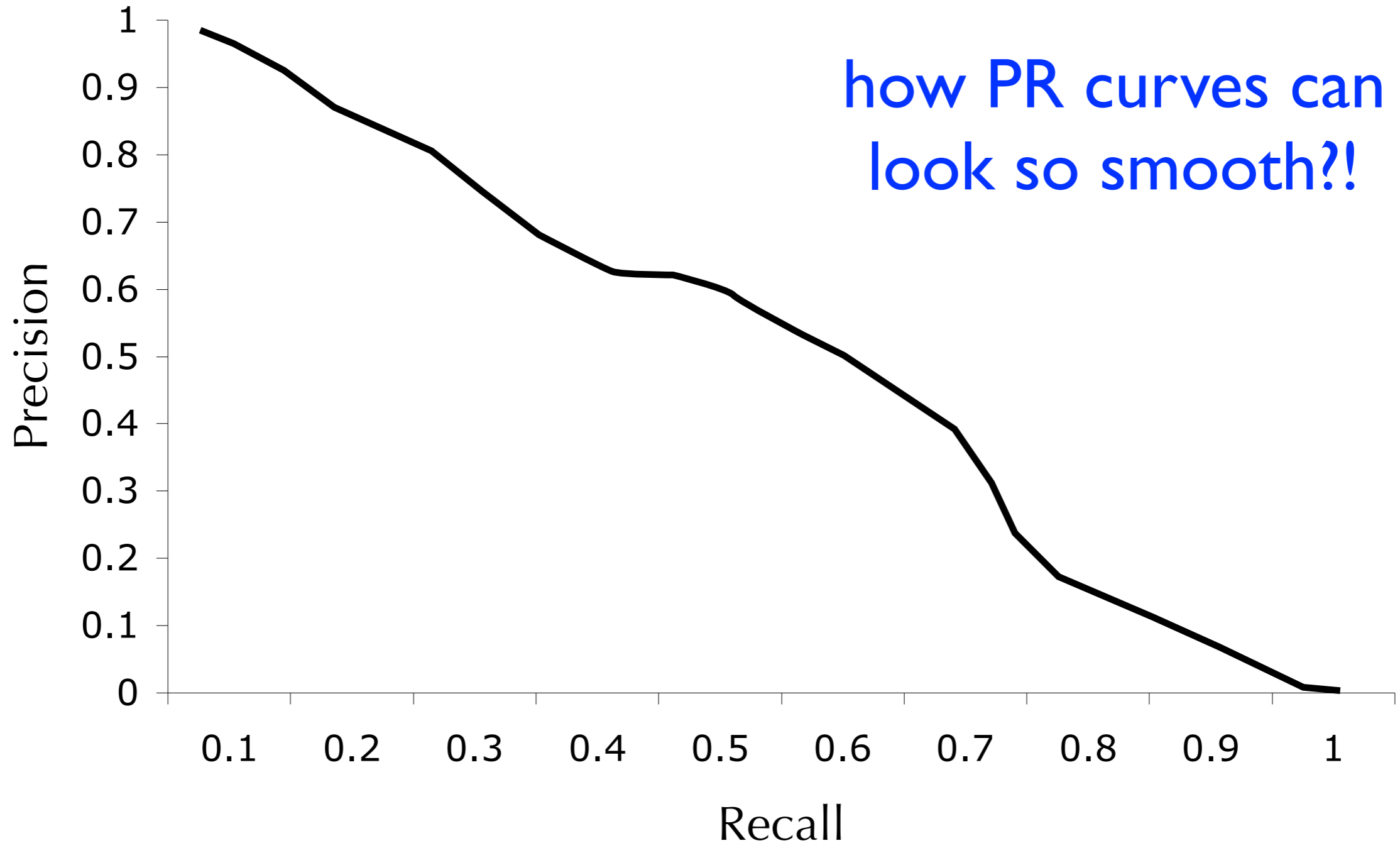
# Evaluation Metrics

(5) precision-recall curves



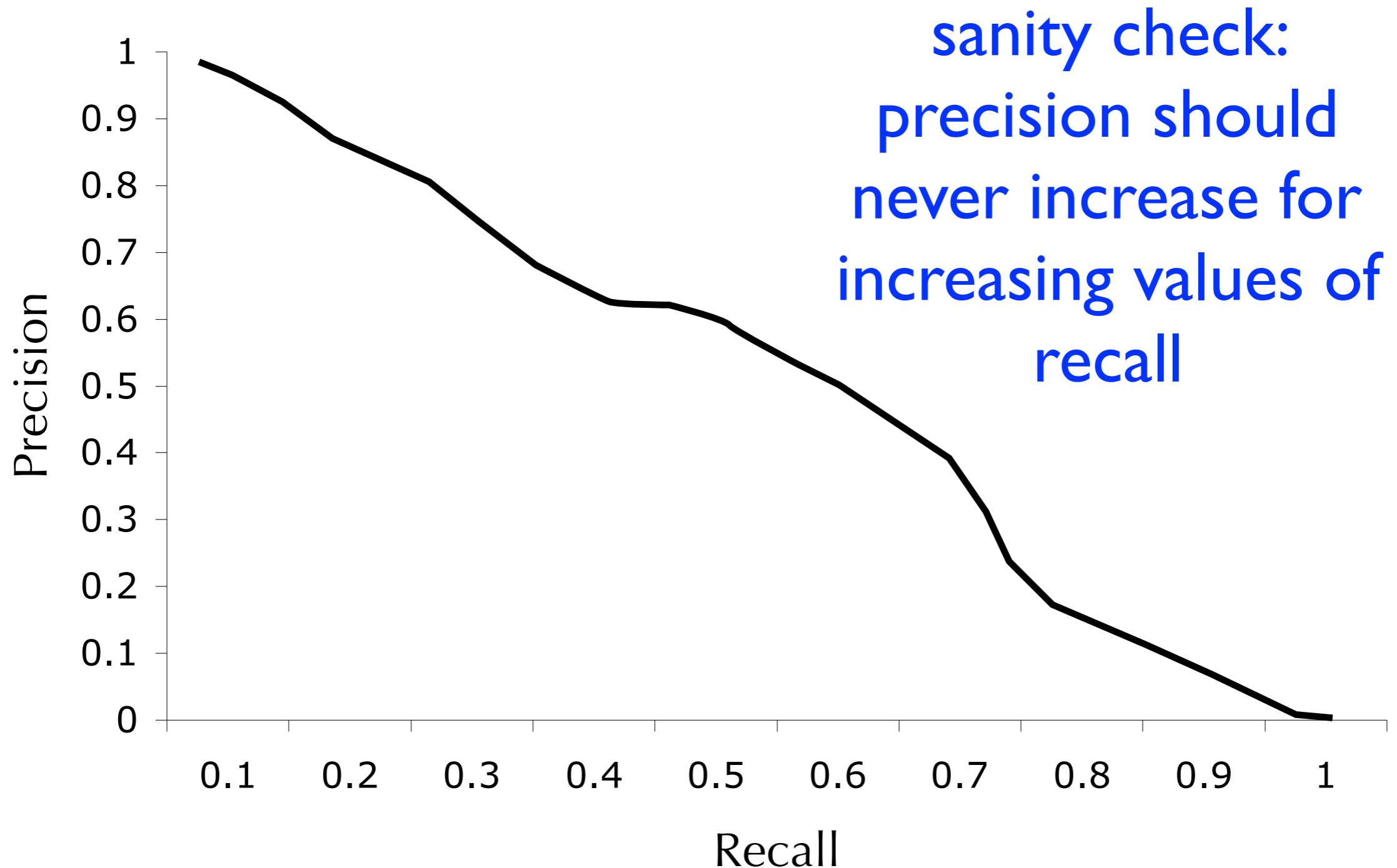
# Evaluation Metrics

(5) precision-recall curves



# Evaluation Metrics

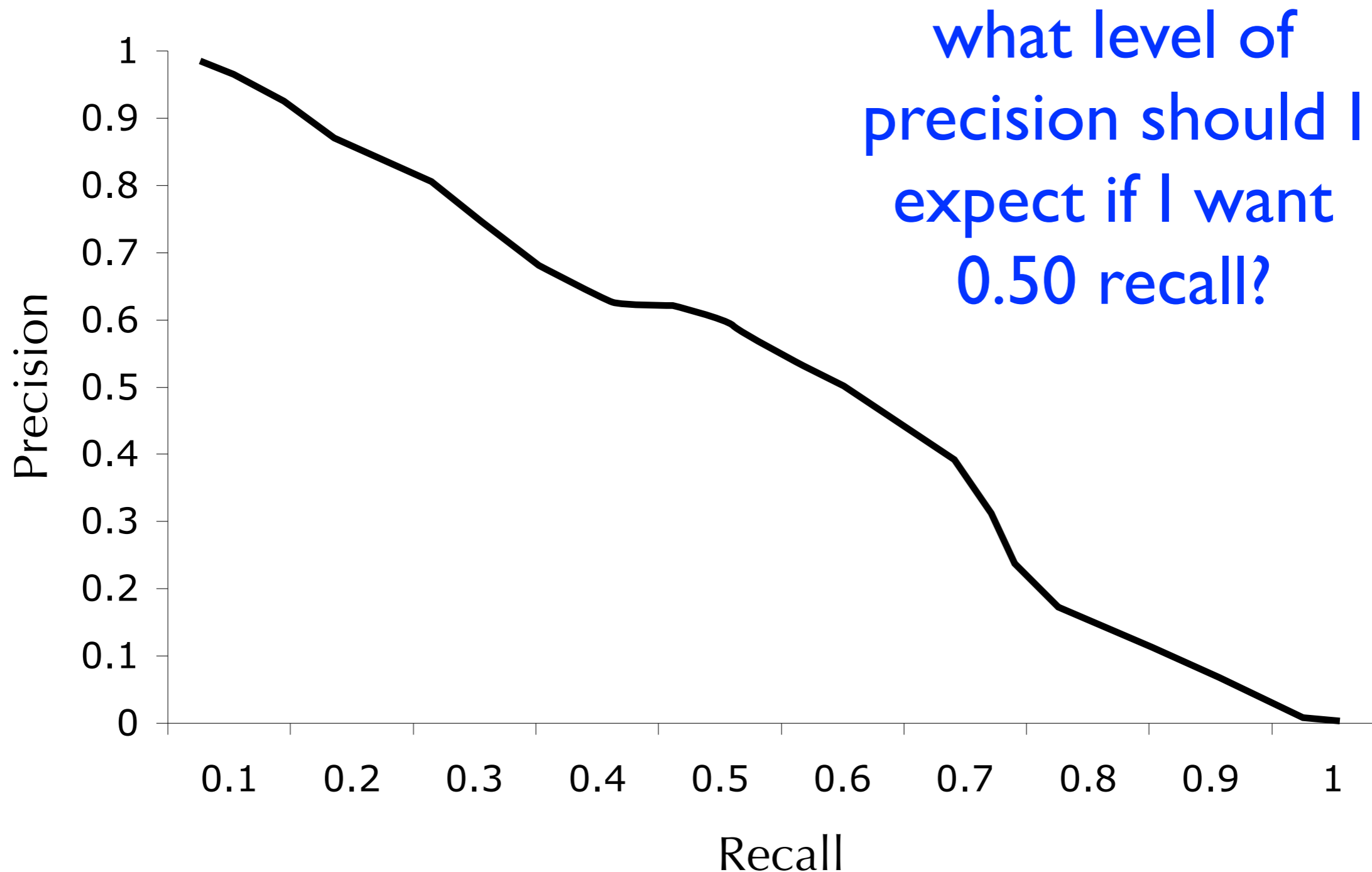
(5) precision-recall curves





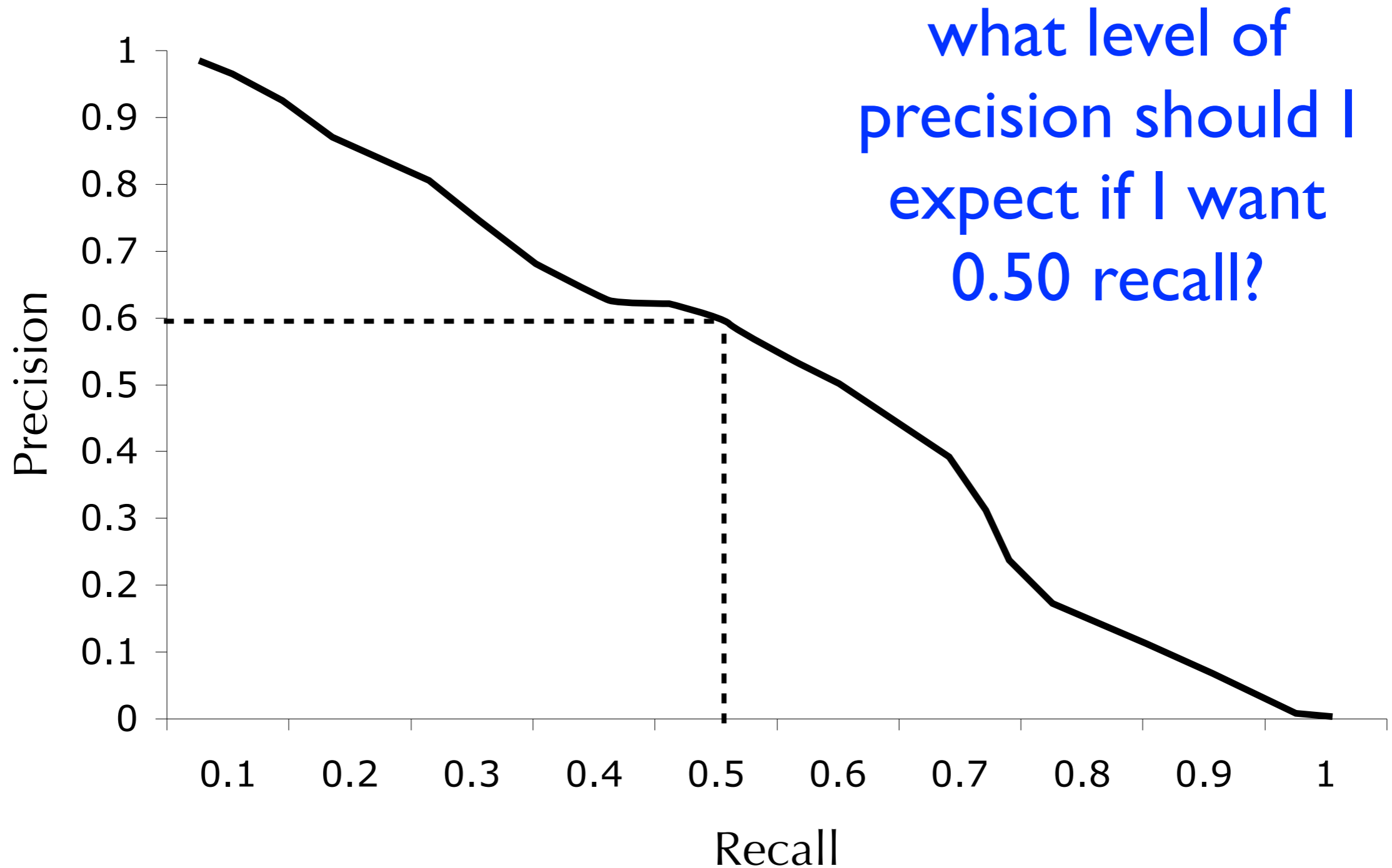
# Evaluation Metrics

(5) precision-recall curves



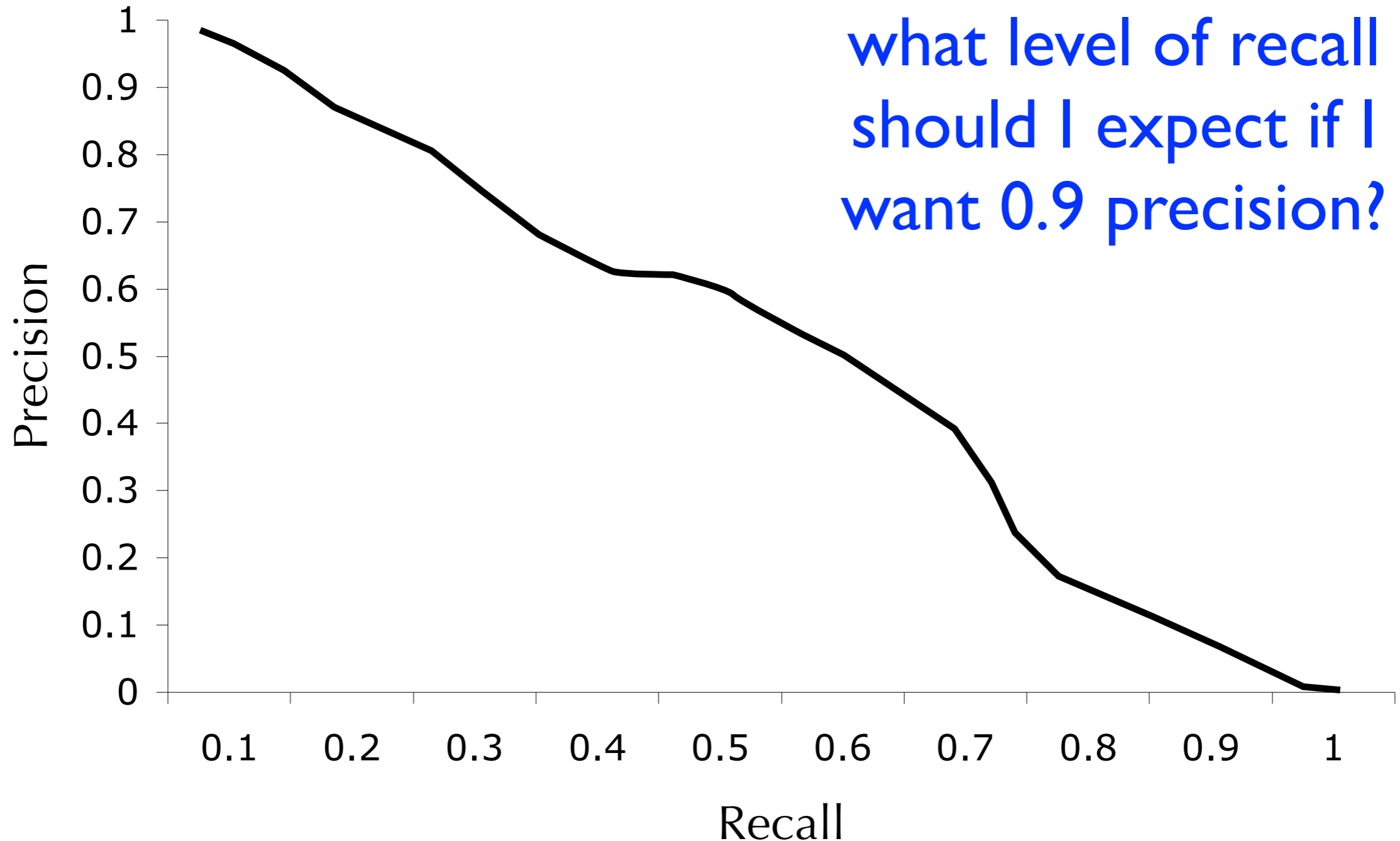
# Evaluation Metrics

## (5) precision-recall curves



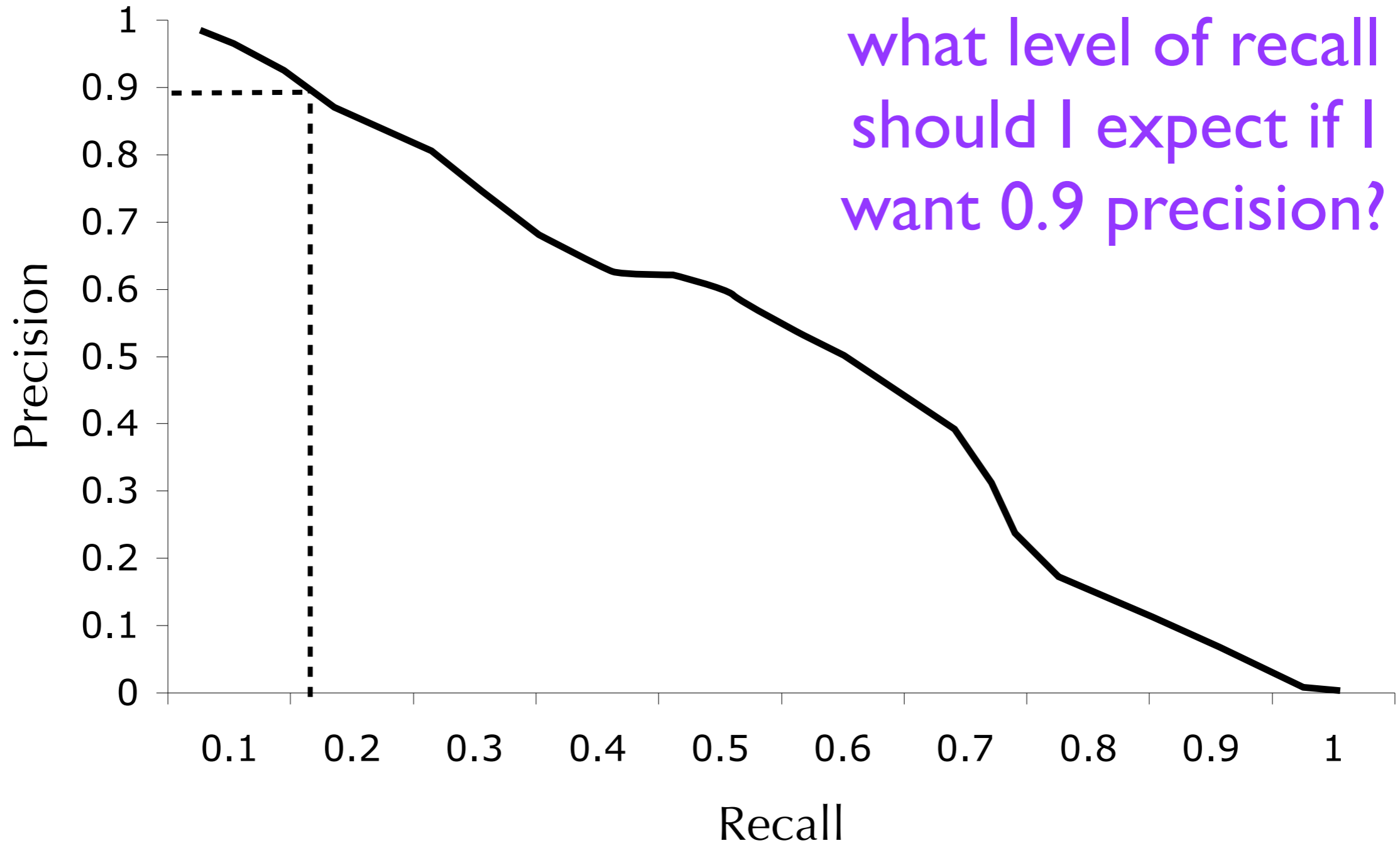
# Evaluation Metrics

(5) precision-recall curves



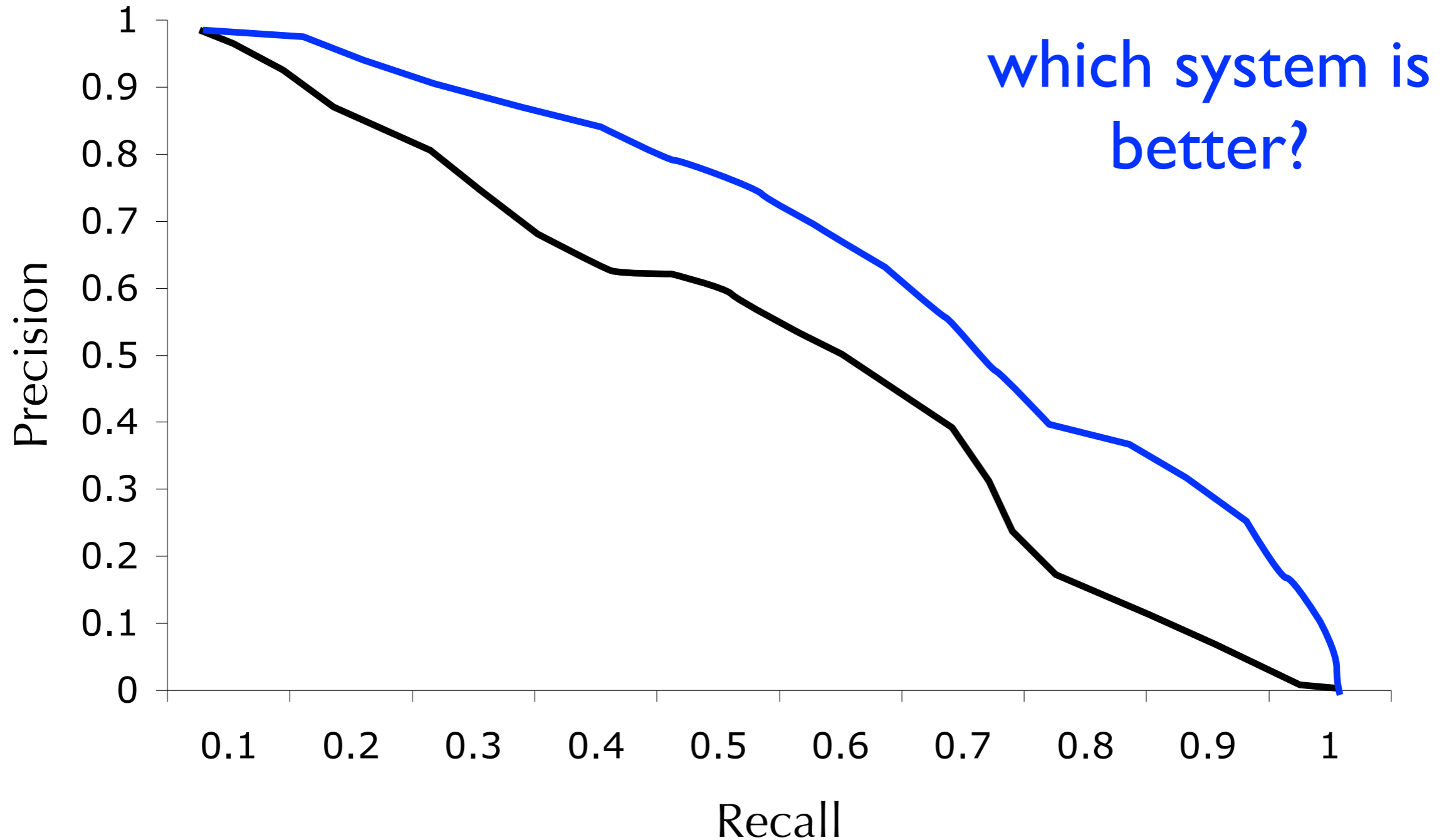
# Evaluation Metrics

## (5) precision-recall curves



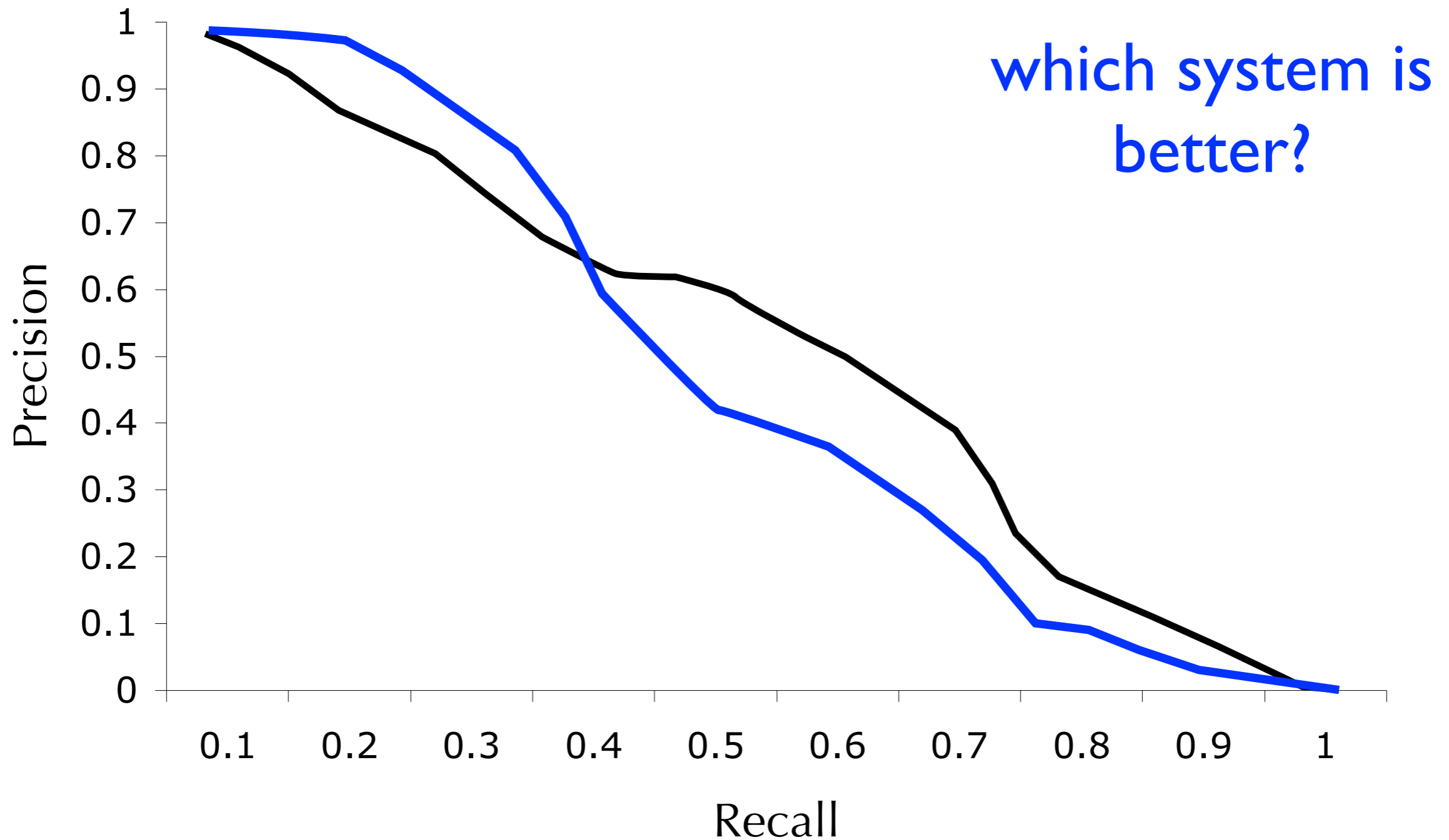
# Evaluation Metrics

(5) precision-recall curves



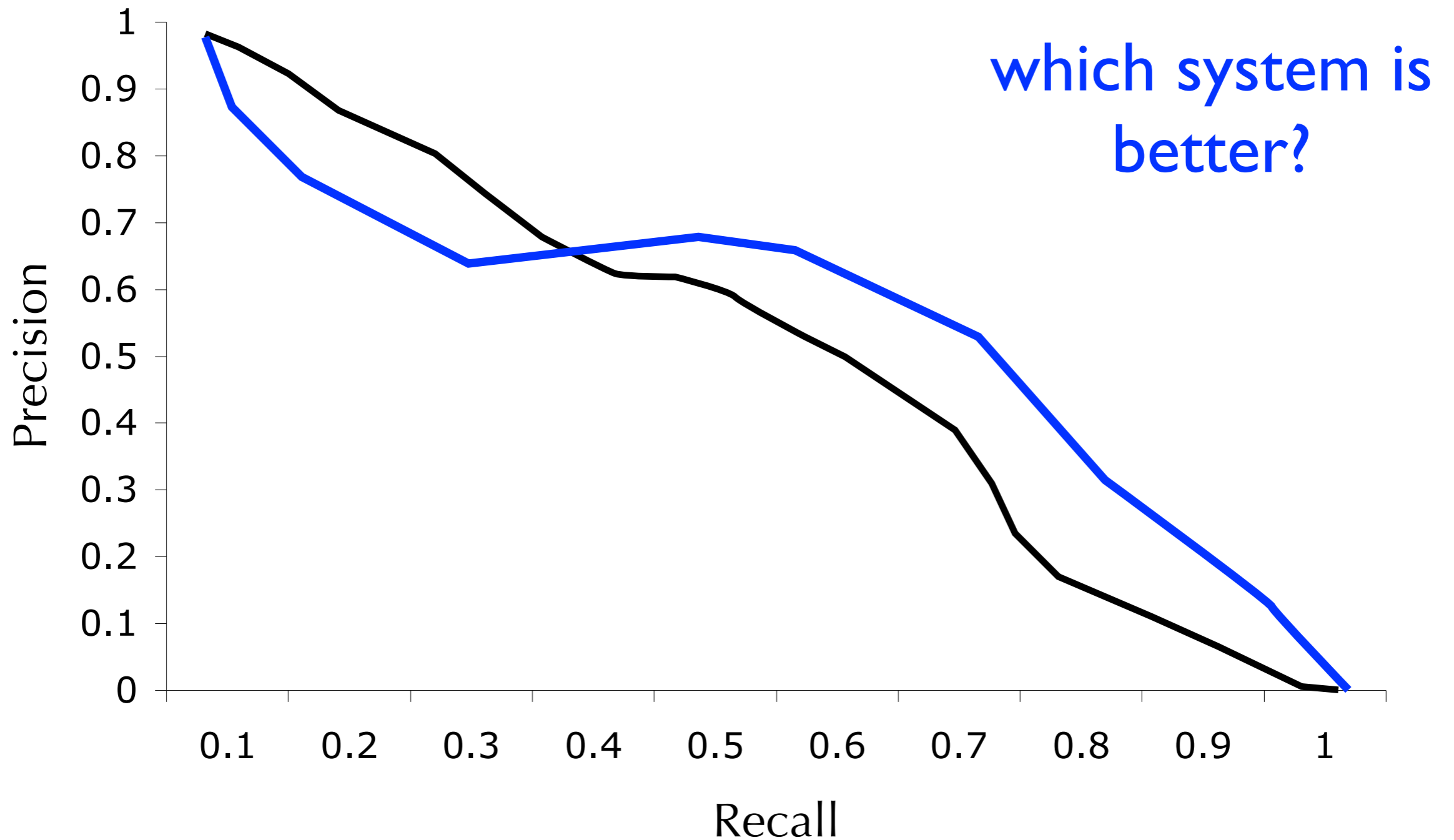
# Evaluation Metrics

(5) precision-recall curves



# Evaluation Metrics

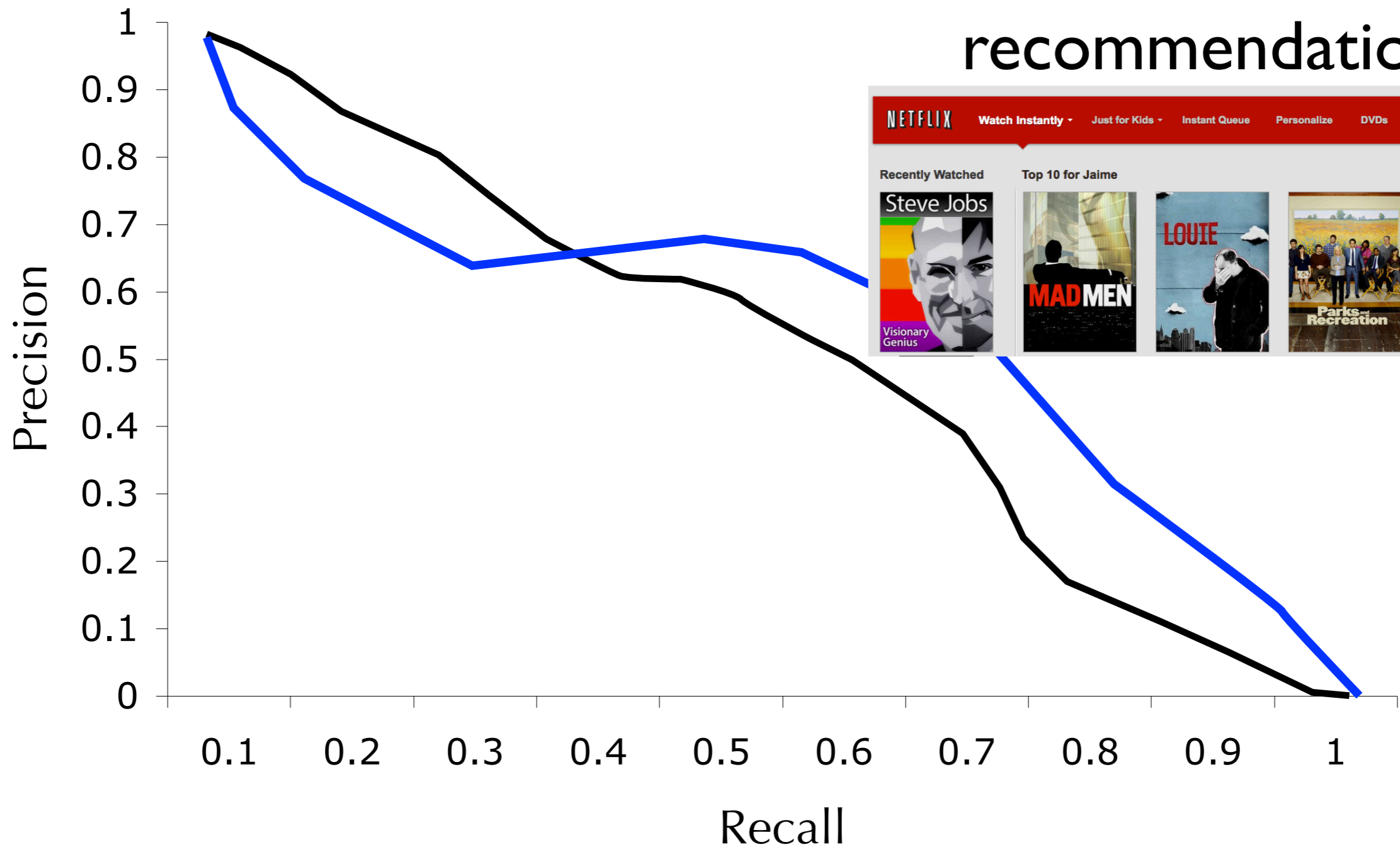
(5) precision-recall curves



# Evaluation Metrics

(5) precision-recall curves

content  
recommendation



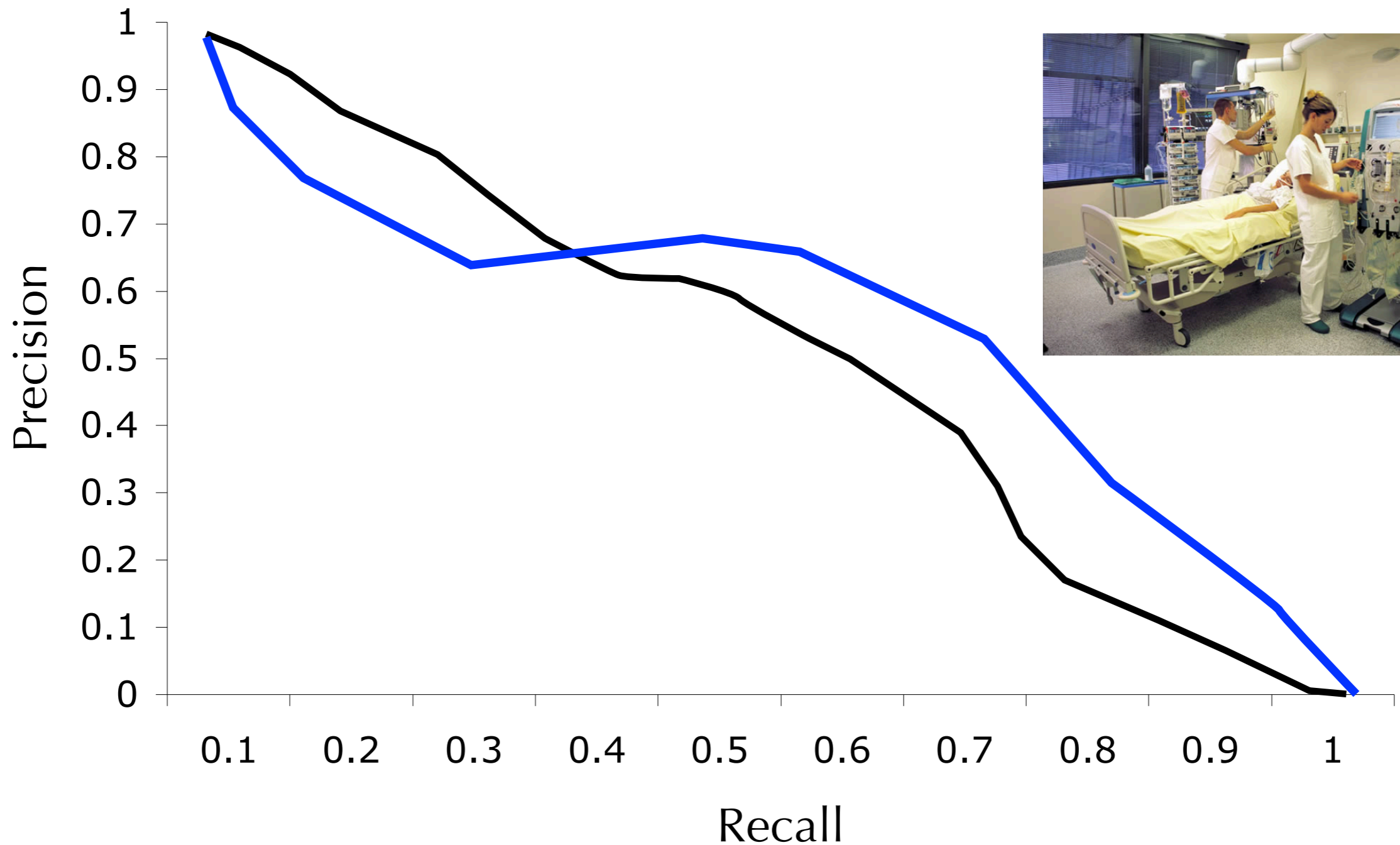
PR curves for 'relevant'



# Evaluation Metrics

(5) precision-recall curves

health monitoring



PR curves for 'alarm'

# Evaluation Metrics

## (5) precision-recall curves

- PR curves show different precision-recall operating points (or trade-off points)
- How many false positives will I have to sift through for a desired level of recall?
- How many cases will I have to miss for a desired level of precision?

# Evaluation Metrics

## (6) average precision

- In some situations, we may want to summarize the quality of a PR curve using a single number
  - ▶ when comparing across lots of different models or feature representations
- **Average precision:** proportional (not equal) to the area under the PR curve

# Evaluation Metrics

## (6) average precision

- Average Precision
  1. Sort instances by descending order of confidence value
  2. Go down the ranking, and measure  $P@K$  where recall increases
  3. Take the average of all  $P@K$  values where recall increases

# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87		
3		0.84	0.67	0.20
4		0.83	0.75	0.30
5		0.77	0.80	0.40
6		0.63	0.83	0.50
7		0.58	0.86	0.60
8		0.57		
9		0.56	0.78	0.70
10		0.34		
11		0.33	0.73	0.80
12		0.25		
13		0.21		
14		0.15	0.64	0.90
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01	0.50	1.00
Average Precision			0.76	

# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87	1.00	0.20
3		0.84	1.00	0.30
4		0.83	1.00	0.40
5		0.77	1.00	0.50
6		0.63	1.00	0.60
7		0.58	1.00	0.70
8		0.57	1.00	0.80
9		0.56	1.00	0.90
10		0.34	1.00	1.00
11		0.33		
12		0.25		
13		0.21		
14		0.15		
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01		
		Average Precision	1.00	

# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87	1.00	0.20
3		0.84	1.00	0.30
4		0.83	1.00	0.40
5		0.77	1.00	0.50
6		0.63	1.00	0.60
7		0.58	1.00	0.70
8		0.57	1.00	0.80
9		0.56	1.00	0.90
10		0.34		
11		0.33	0.91	1.00
12		0.25		
13		0.21		
14		0.15		
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01		
Average Precision			0.99	

# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87	1.00	0.20
3		0.84	1.00	0.30
4		0.83	1.00	0.40
5		0.77	1.00	0.50
6		0.63	1.00	0.60
7		0.58	1.00	0.70
8		0.57	1.00	0.80
9		0.56	1.00	0.90
10		0.34	1.00	1.00
11		0.33		
12		0.25		
13		0.21		
14		0.15		
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01		
		Average Precision	1.00	



# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87		
3		0.84	0.67	0.20
4		0.83	0.75	0.30
5		0.77	0.80	0.40
6		0.63	0.83	0.50
7		0.58	0.86	0.60
8		0.57	0.88	0.70
9		0.56	0.89	0.80
10		0.34	0.90	0.90
11		0.33	0.91	1.00
12		0.25		
13		0.21		
14		0.15		
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01		
	Average Precision		0.85	

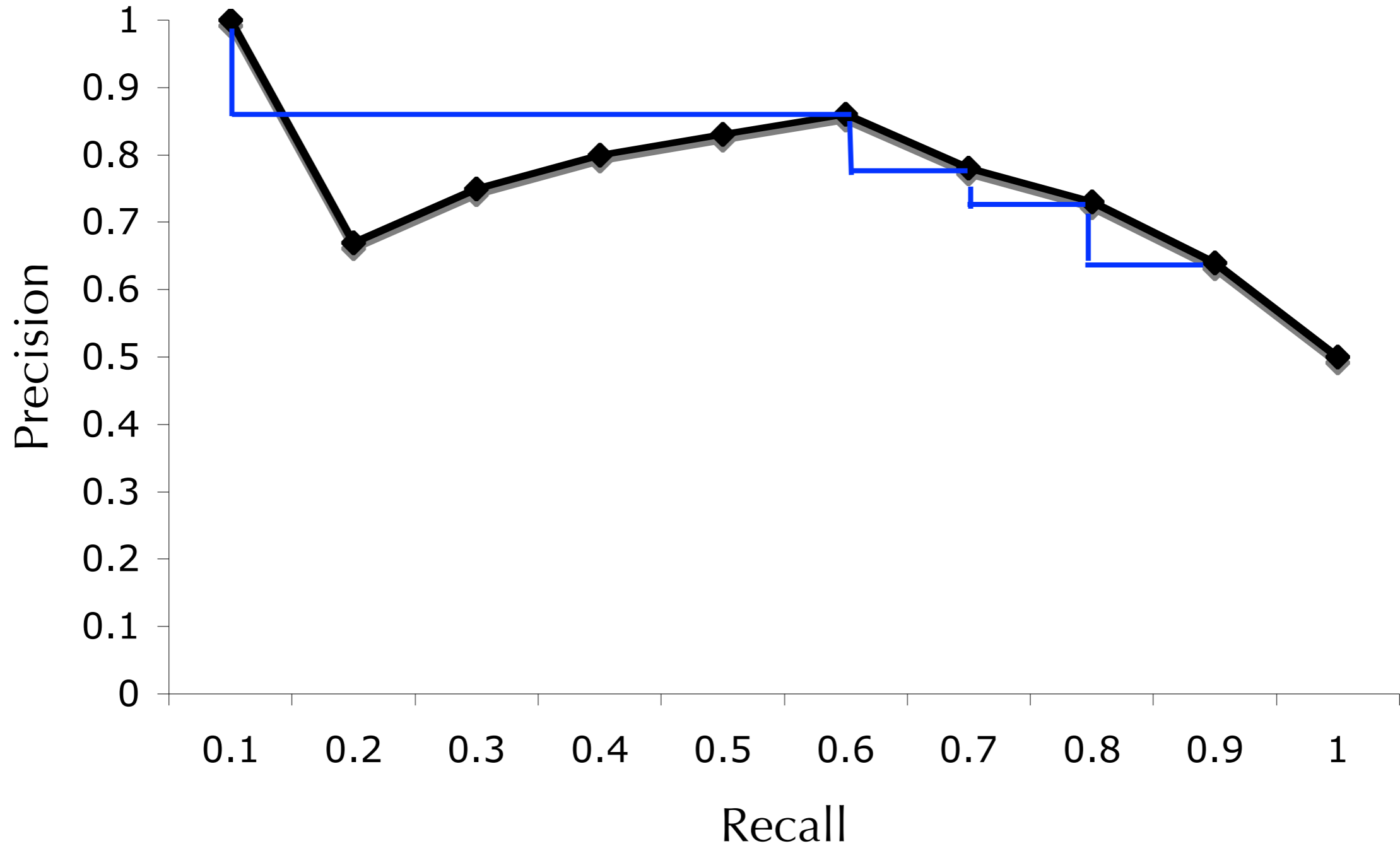
# Evaluation Metrics

## (6) average precision

rank (K)	ranking	P(POS D)	P@K	R@K
1		0.99	1.00	0.10
2		0.87		
3		0.84	0.67	0.20
4		0.83	0.75	0.30
5		0.77	0.80	0.40
6		0.63	0.83	0.50
7		0.58	0.86	0.60
8		0.57		
9		0.56	0.78	0.70
10		0.34		
11		0.33	0.73	0.80
12		0.25		
13		0.21		
14		0.15	0.64	0.90
15		0.14		
16		0.14		
17		0.12		
18		0.08		
19		0.01		
20		0.01	0.50	1.00
Average Precision			0.76	

# Evaluation Metrics

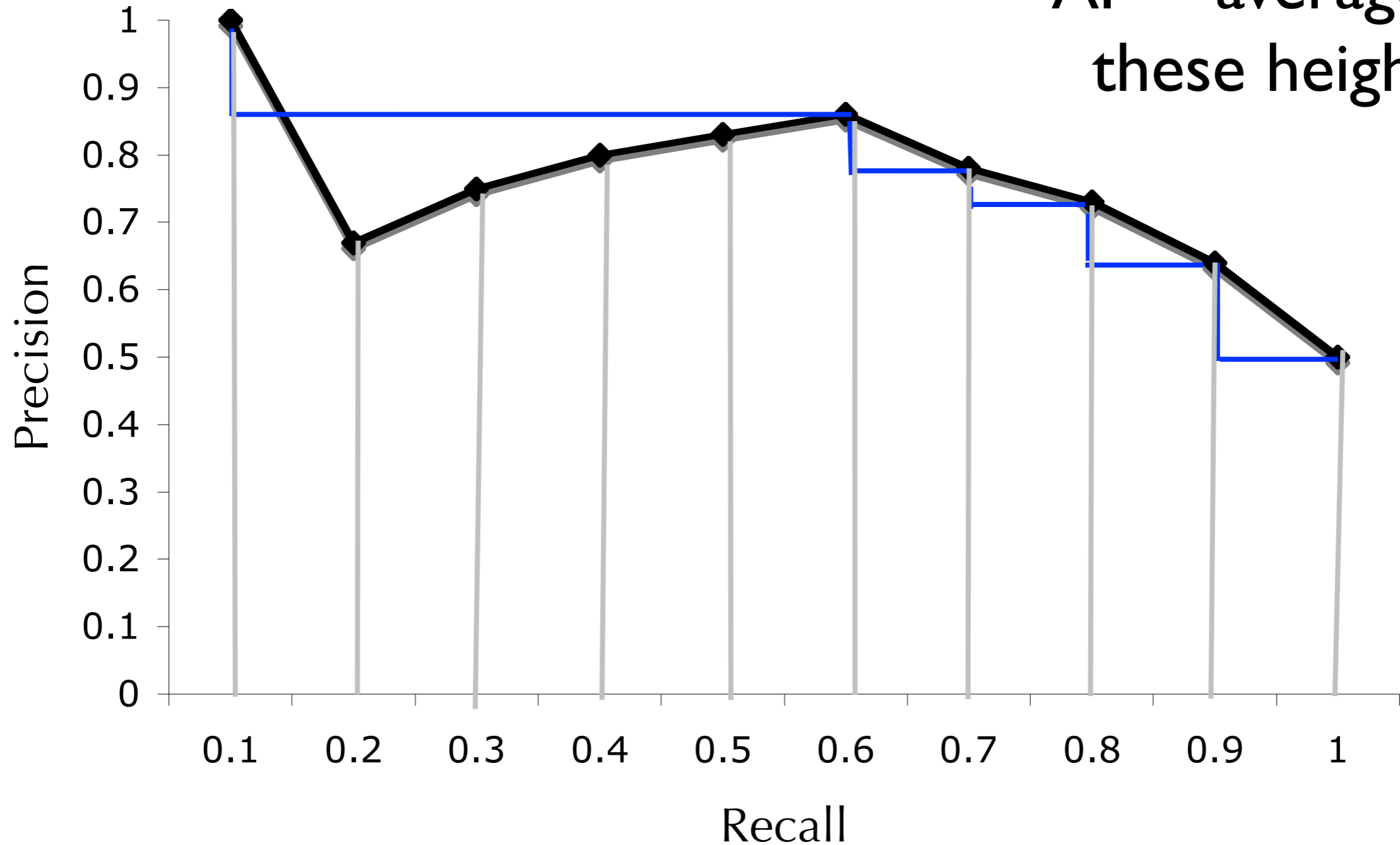
(6) average precision



# Evaluation Metrics

(6) average precision

AP = average of these heights



# Evaluation Metrics

## (6) average precision

- Average precision is proportional to the area under the PR curve
- It punishes high-confident mistakes more severely than low-confident mistakes

# Evaluation Metrics

- Accuracy
- Precision
- Recall
- F-measure (or F1 measure)
- PR curves (not a metric, but rather a way to show different PR operating points)
- Average Precision