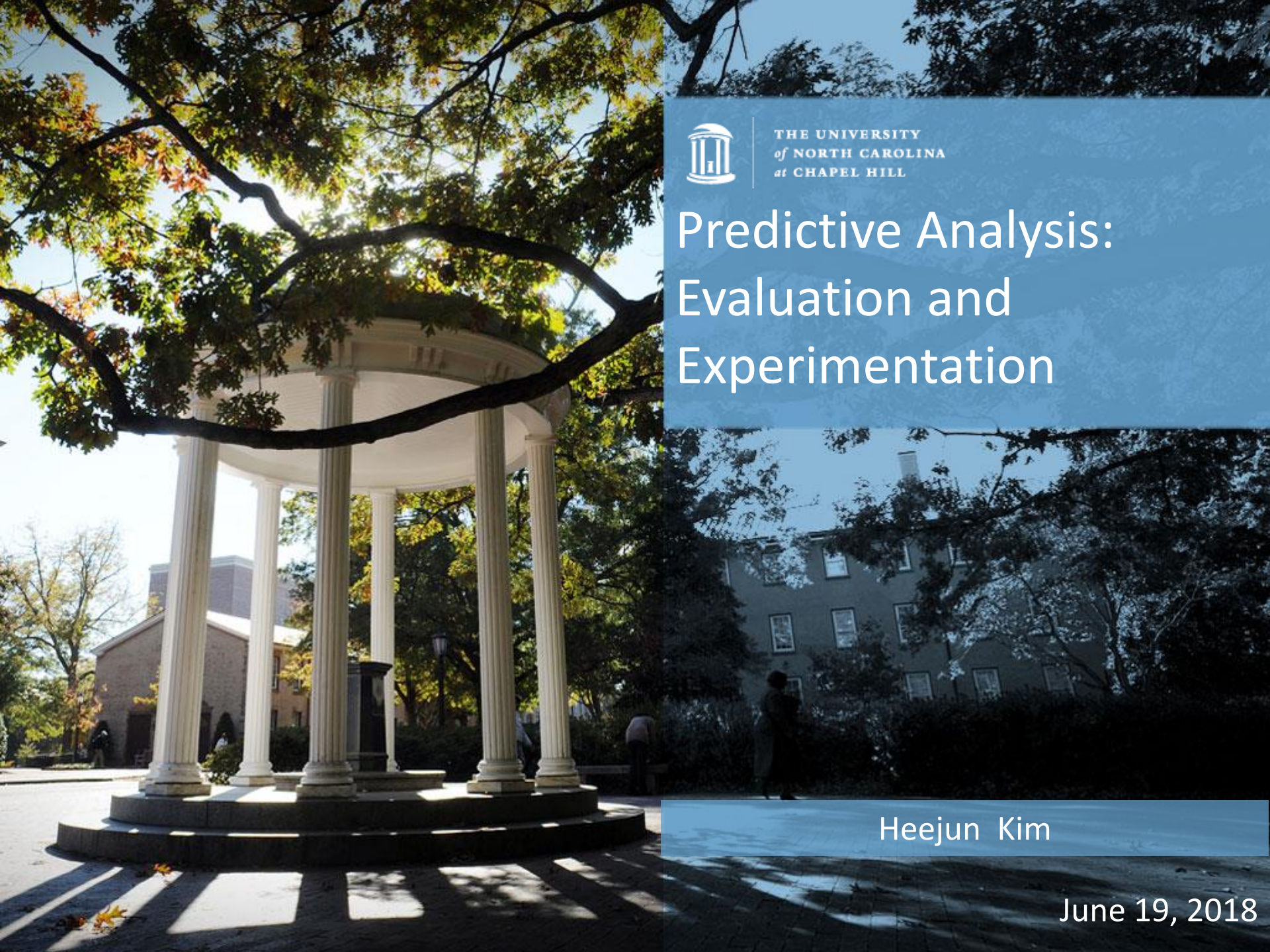


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Predictive Analysis: Evaluation and Experimentation

Heejun Kim

June 19, 2018

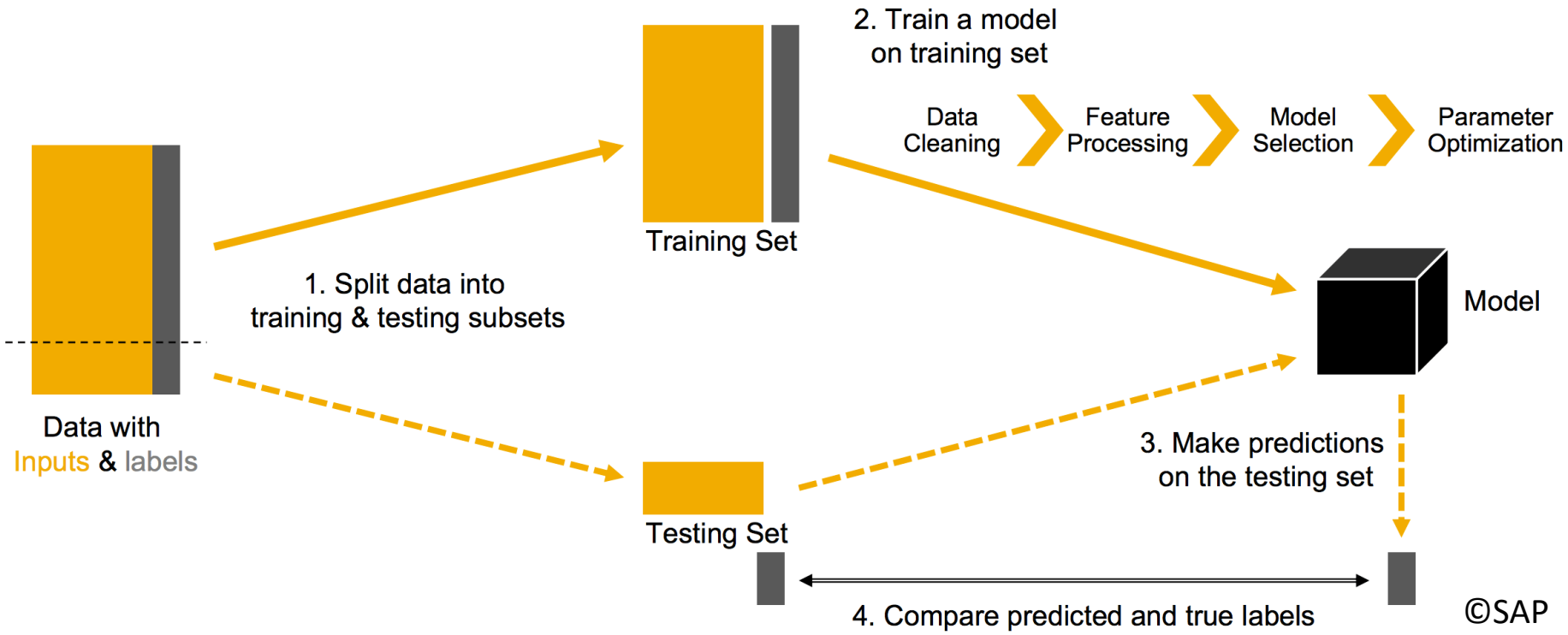
Evaluation and Experimentation

- **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



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

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

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

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

Evaluation Metrics
(1) accuracy

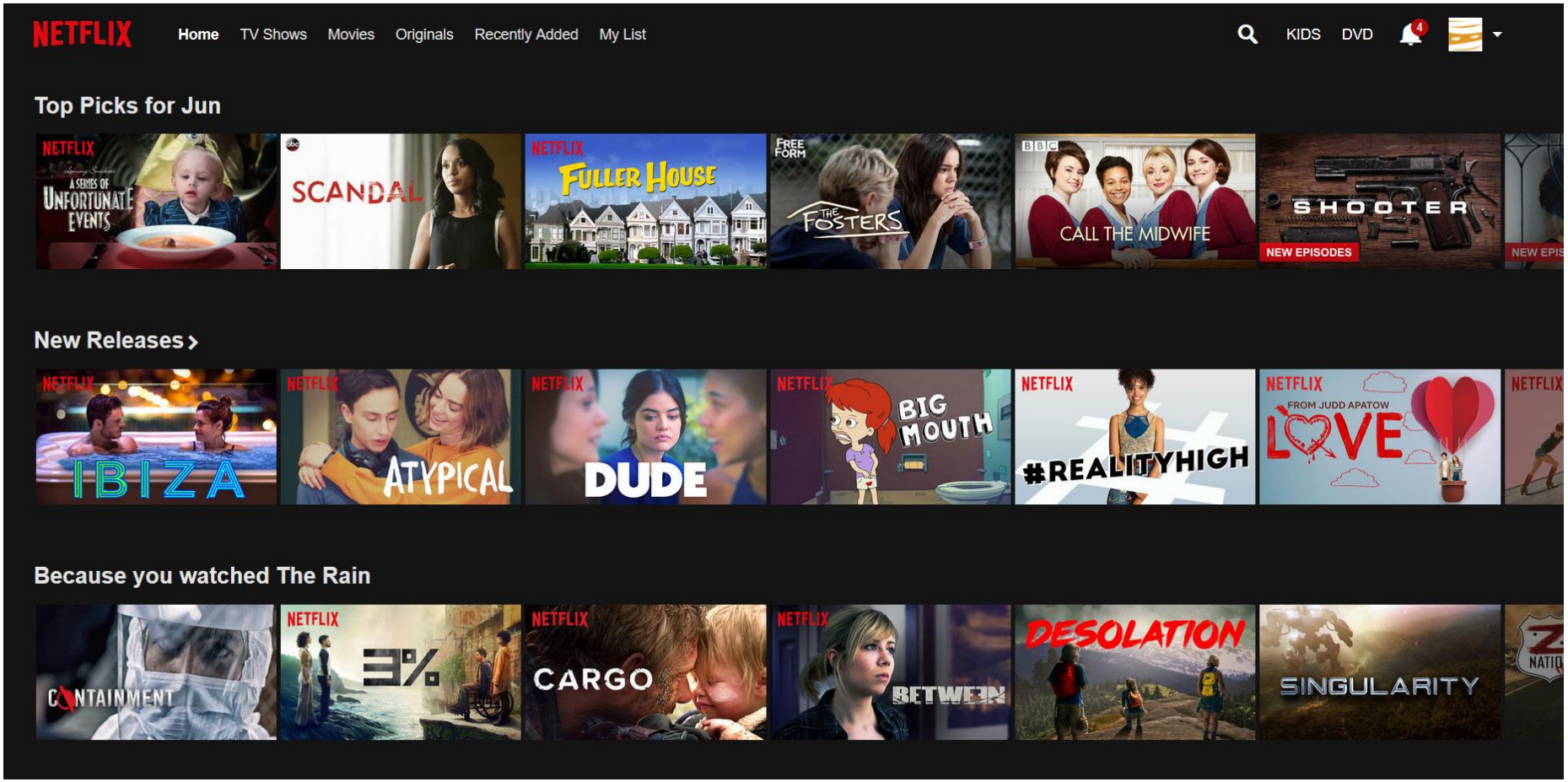
- What assumption(s) does accuracy make?

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

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

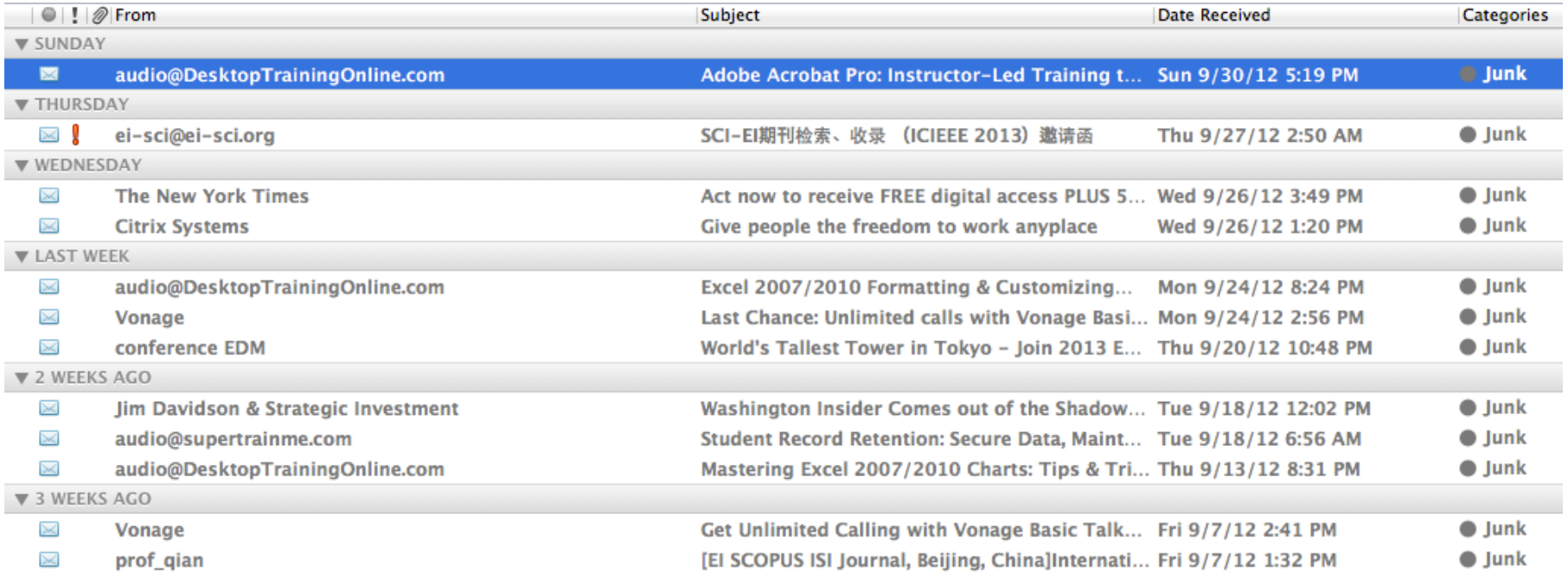
Evaluation Metrics

- 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

Evaluation Metrics

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

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

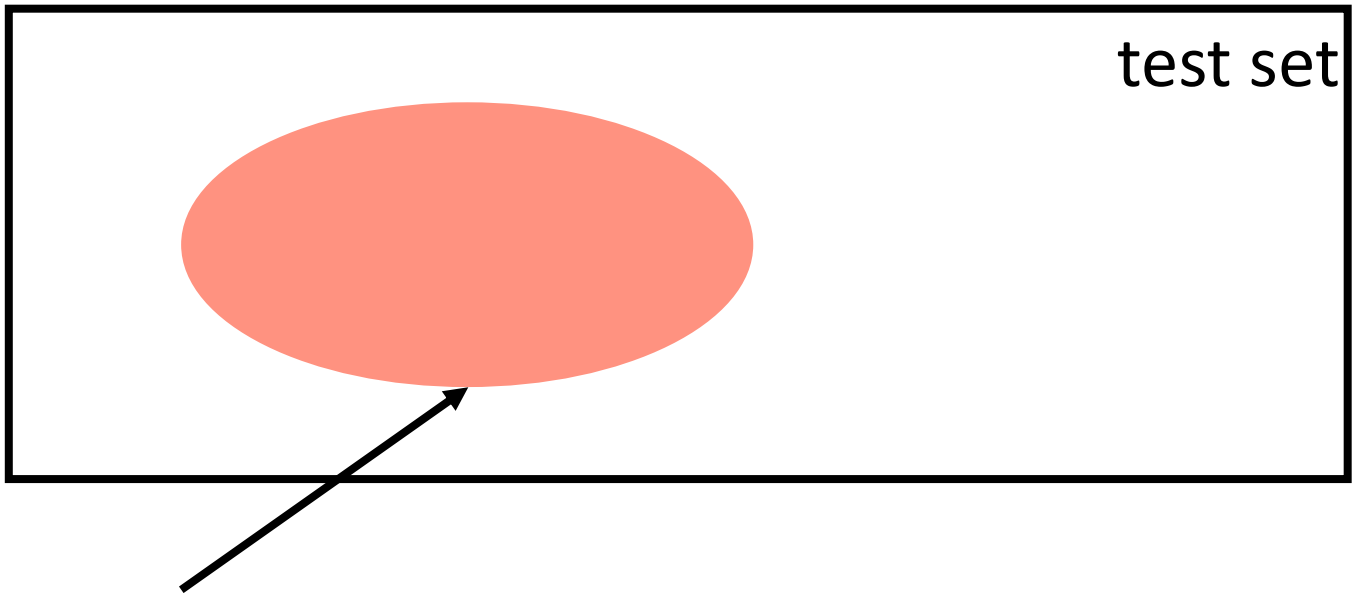
- For a given class **C**:
  - precision: the percentage of positive predictions that are truly positive
  - recall: the percentage of true positives that are correctly predicted positive

Evaluation Metrics

(2) precision and (3) recall

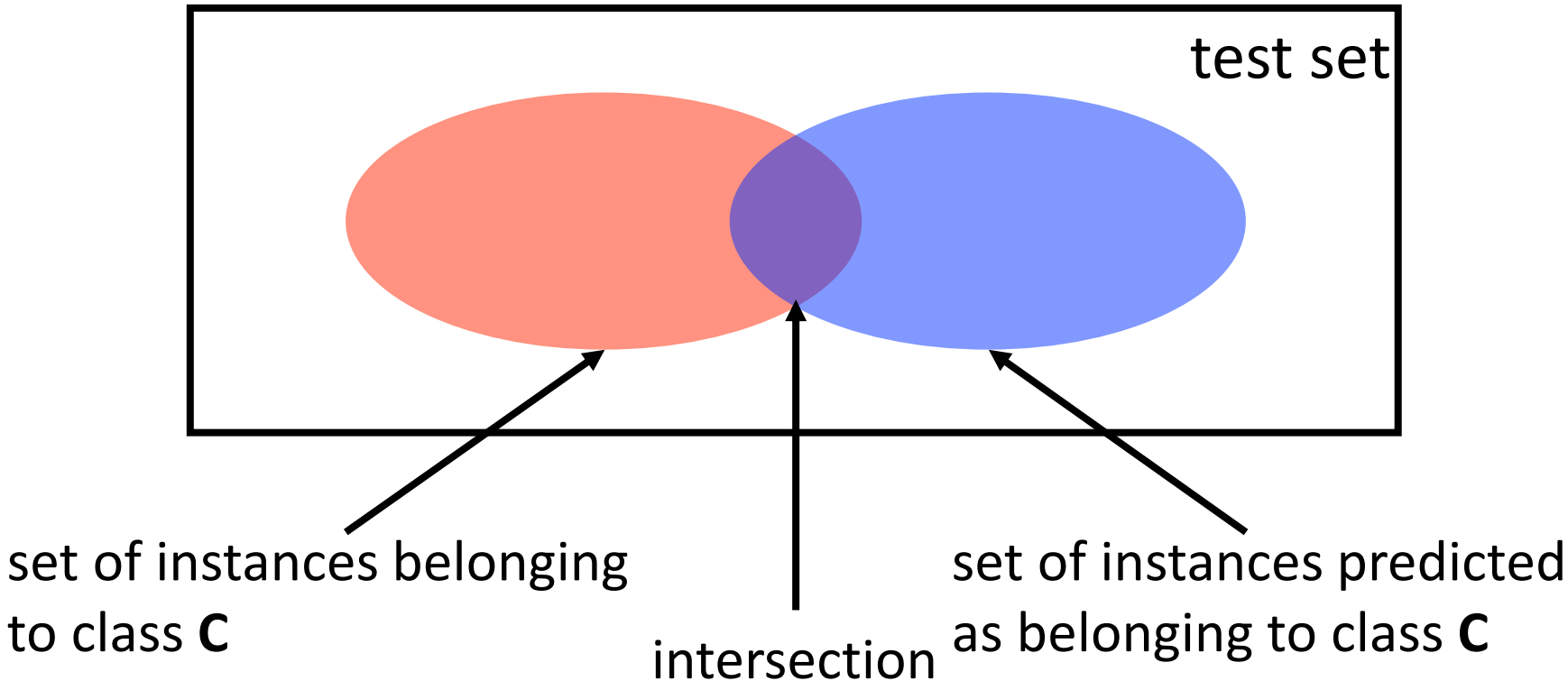
Evaluation Metrics

(2) precision and (3) recall



Evaluation Metrics

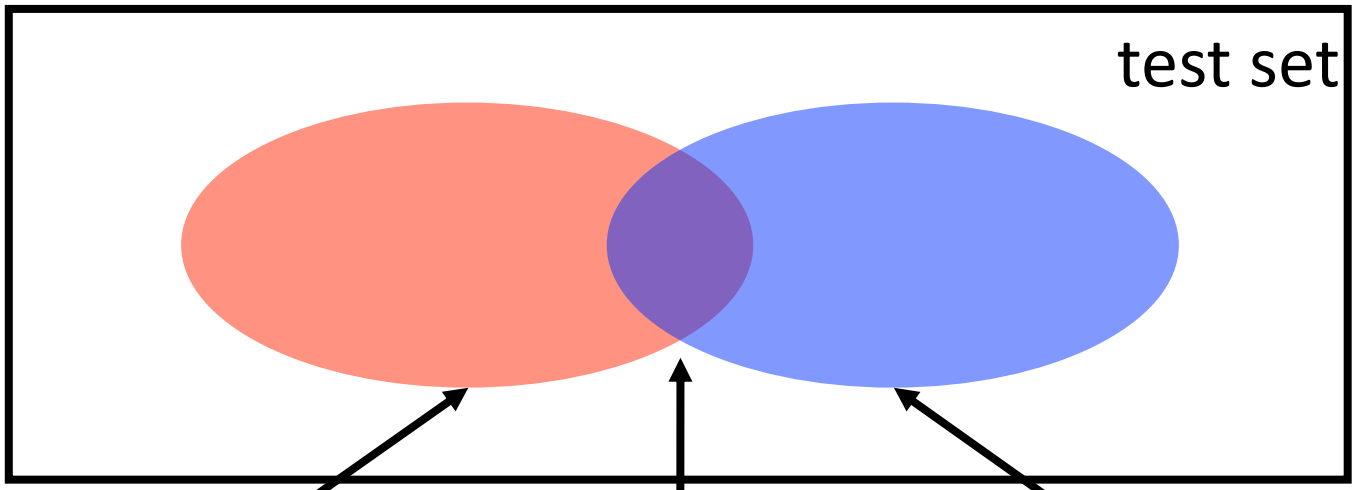
(2) precision and (3) recall



Evaluation Metrics

(2) precision and (3) recall

- Precision (the percentage of positive predictions that are truly positive) = ?



set of instances belonging to class **C**

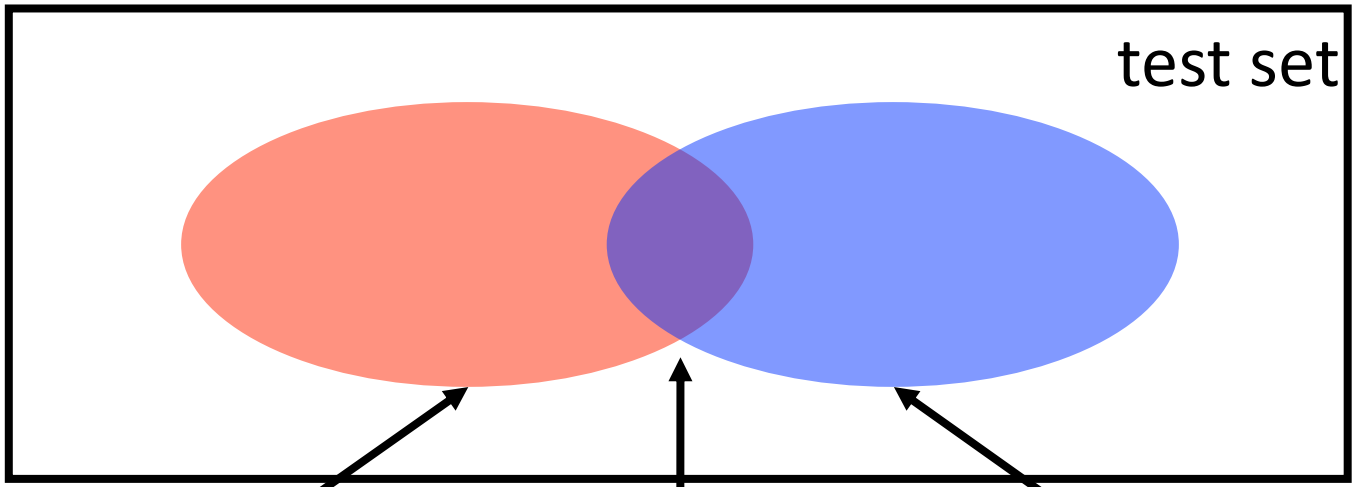
intersection

set of instances predicted as belonging to class **C**

Evaluation Metrics

(2) precision and (3) recall

- Recall (the percentage of true positives that are correctly predicted positive) = ?



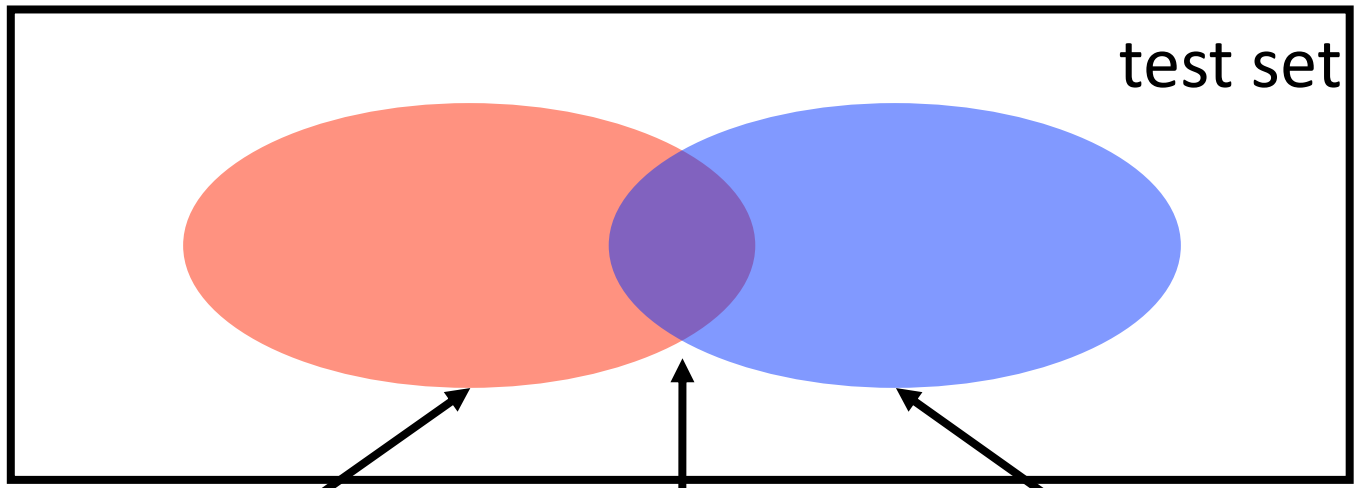
set of instances belonging to class **C**

intersection

set of instances predicted as belonging to class **C**

Evaluation Metrics

(2) precision and (3) recall



set of instances belonging to class **C**

intersection

set of instances predicted as belonging to class **C**

Evaluation Metrics

(2) precision

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

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

For a positive class, the percentage of correct predictions that are truly positive

Evaluation Metrics

(3) recall

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

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

For a positive class, the percentage of true positives that are correctly predicted positive

Evaluation Metrics

prevision vs. recall

		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

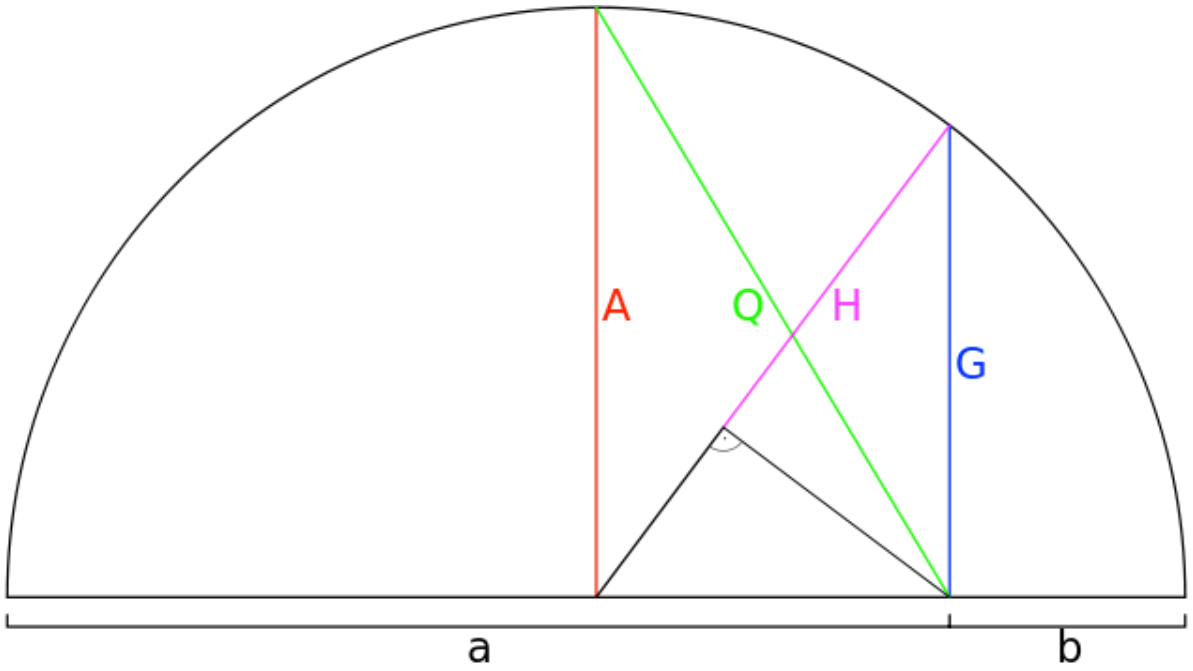
$$\mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$$

Evaluation Metrics

(4) f-measure

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

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

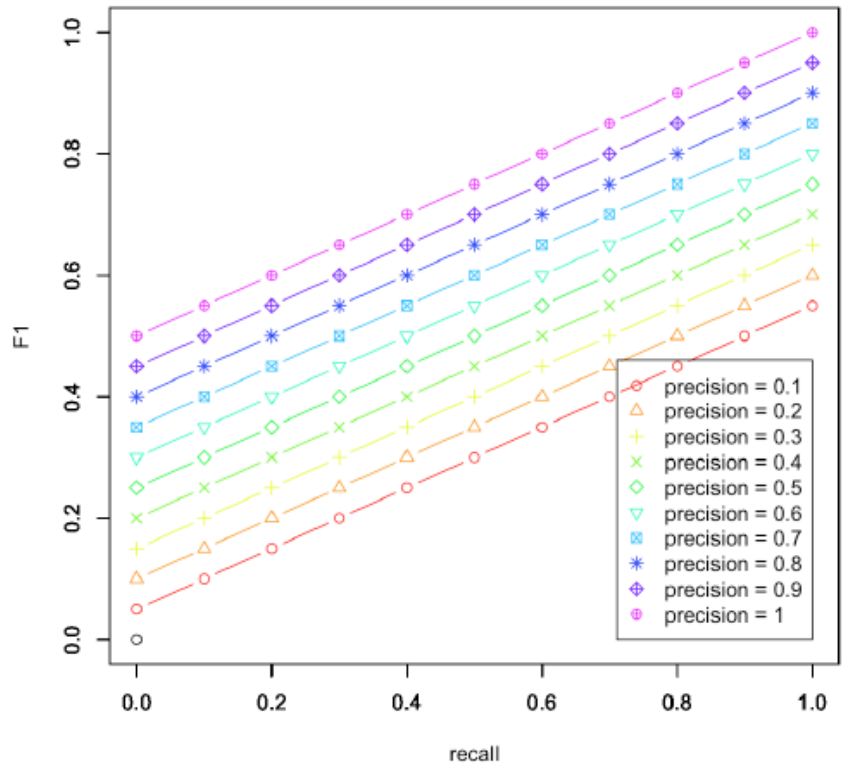


Evaluation Metrics

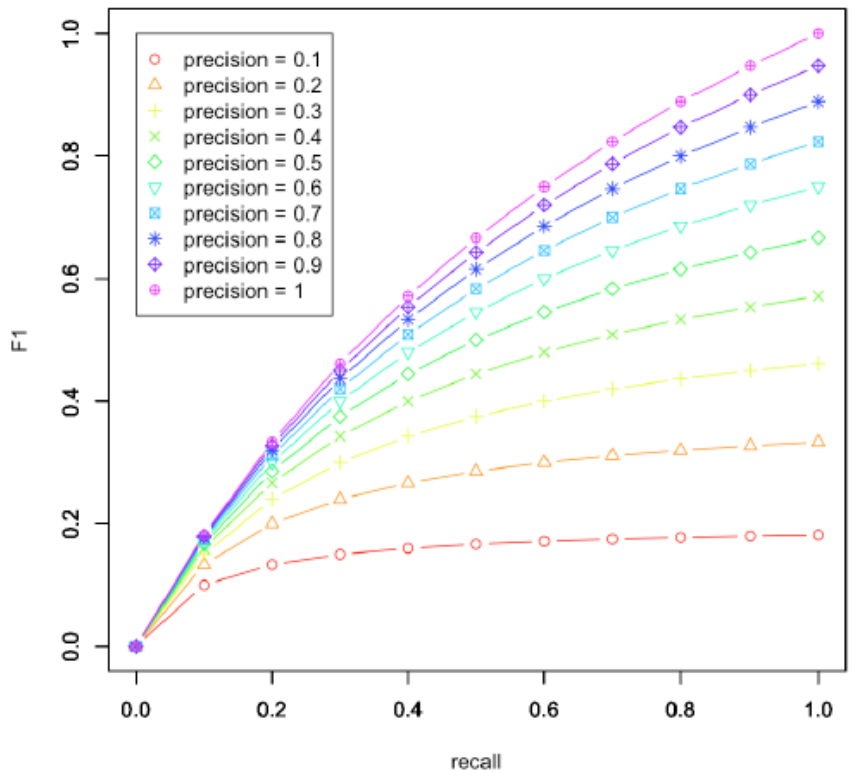
(4) f-measure

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

Arithmetic mean



Harmonic mean



(slide courtesy of Ben Carterette)

Evaluation Metrics

(5) precision-recall curves

- F-measure: assumes that the “end users” care <u>equally</u> 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:

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

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

- Otherwise, predict negative (NEG)

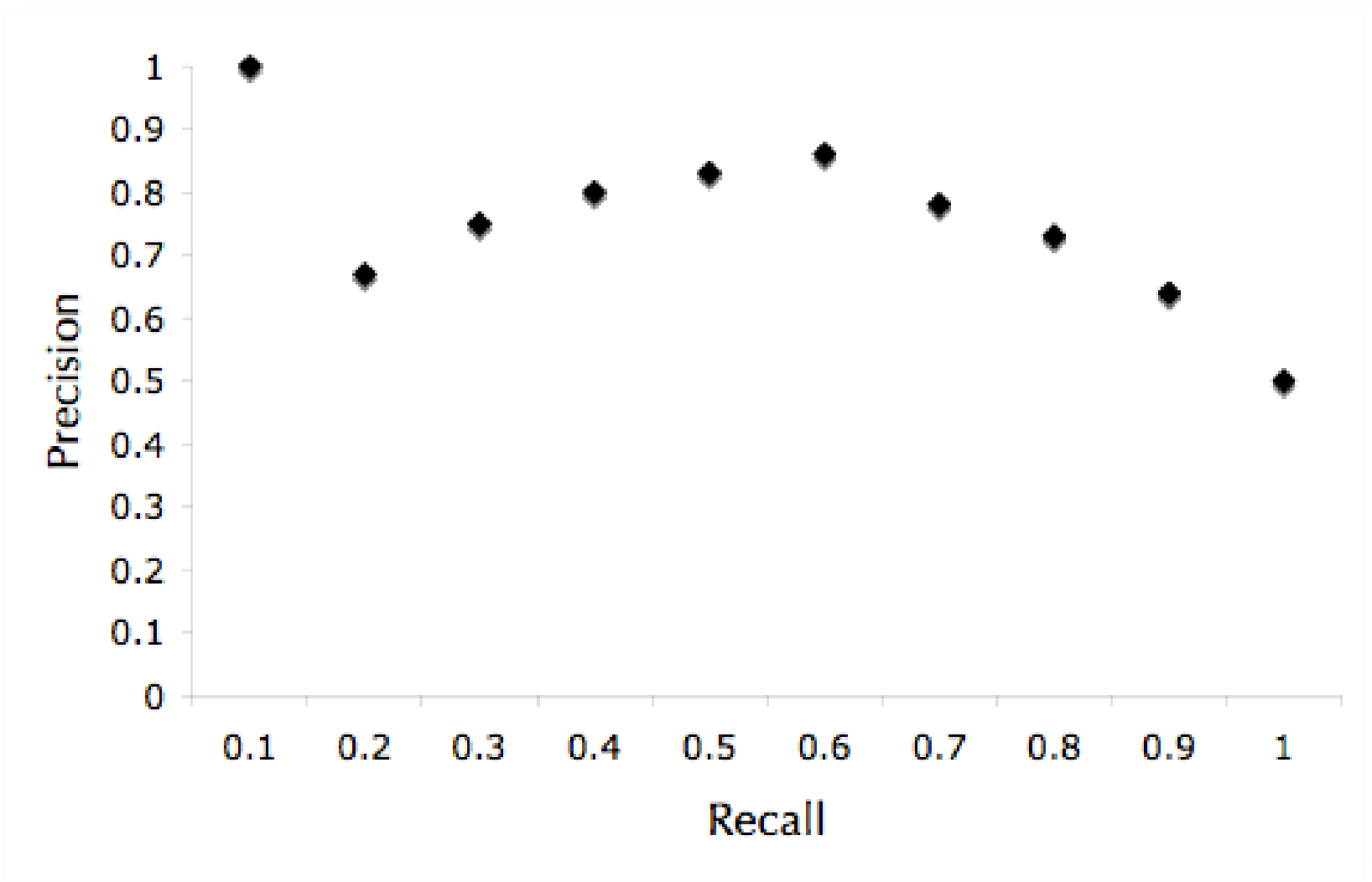
Evaluation Metrics

(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

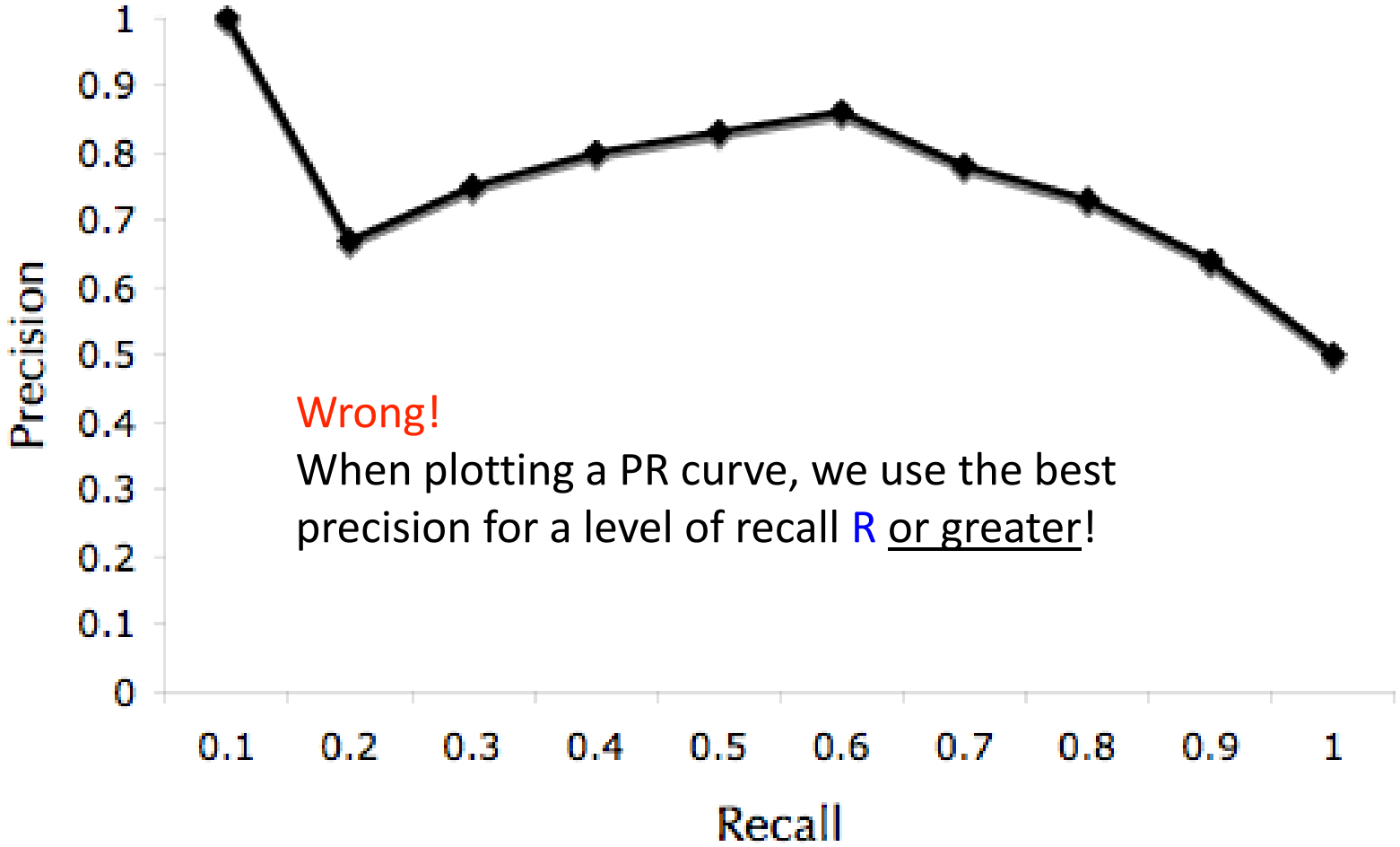
Evaluation Metrics

(5) precision-recall curves



Evaluation Metrics

(5) precision-recall curves

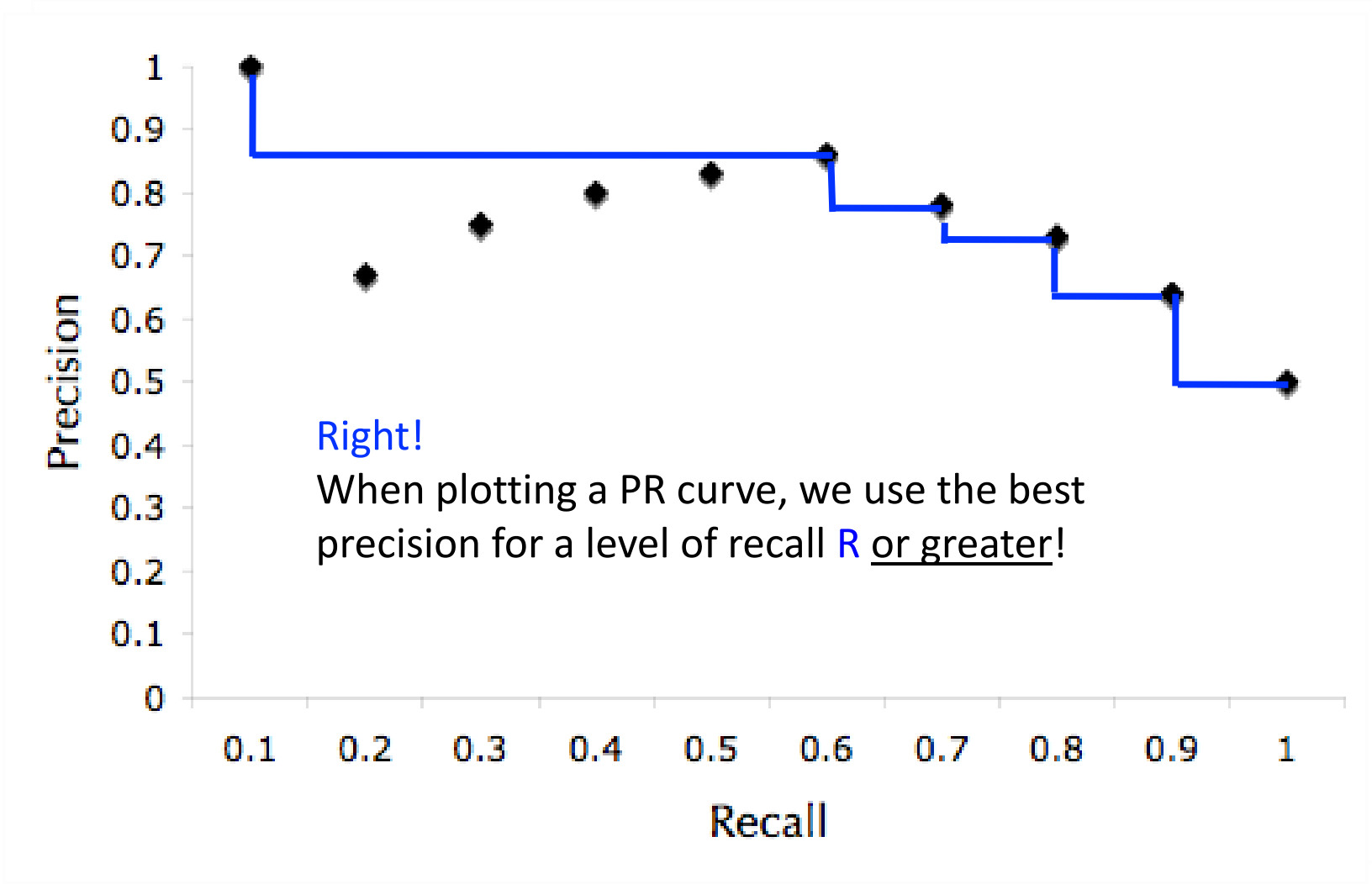


Wrong!

When plotting a PR curve, we use the best precision for a level of recall R or greater!

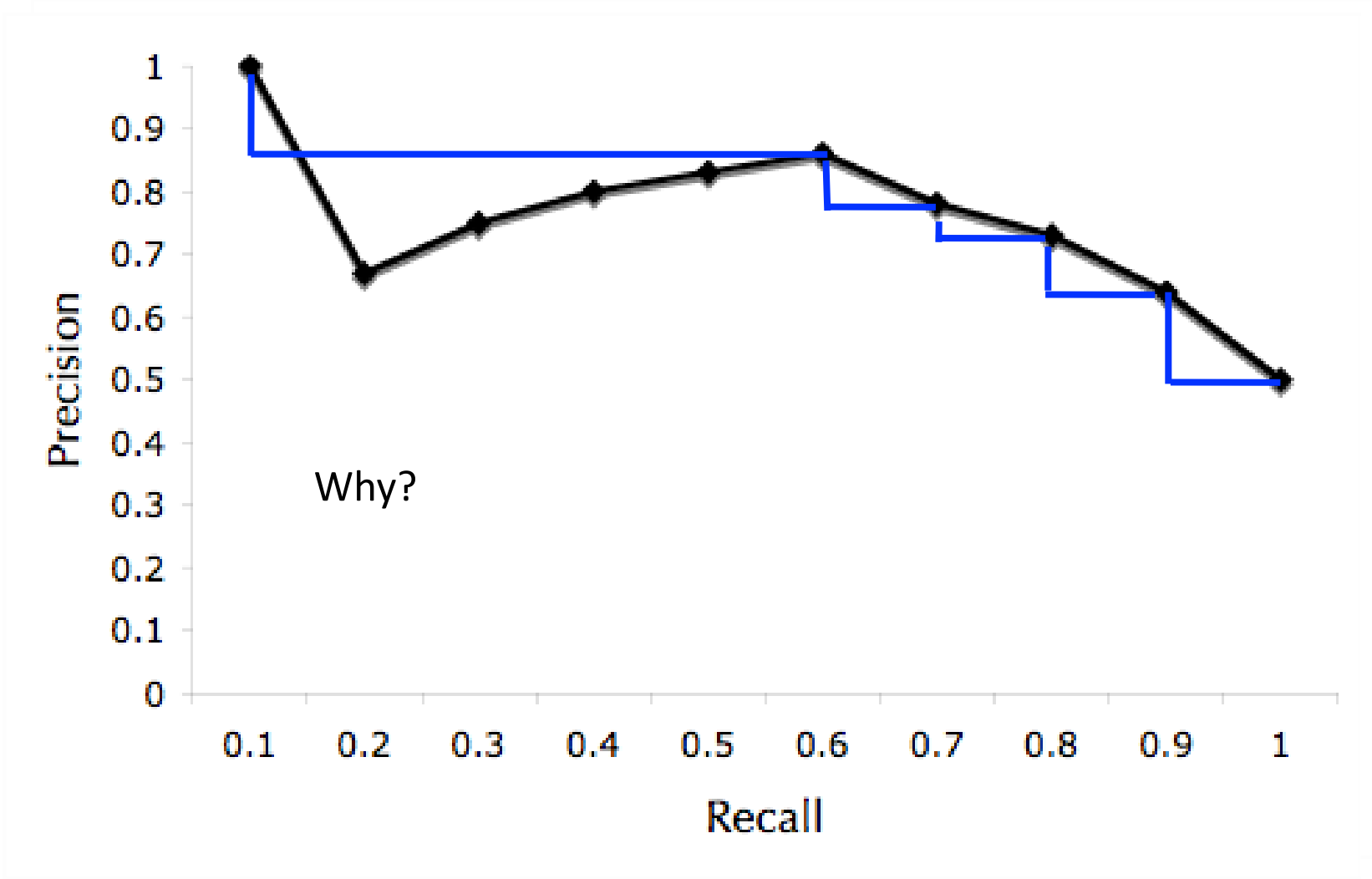
Evaluation Metrics

(5) precision-recall curves



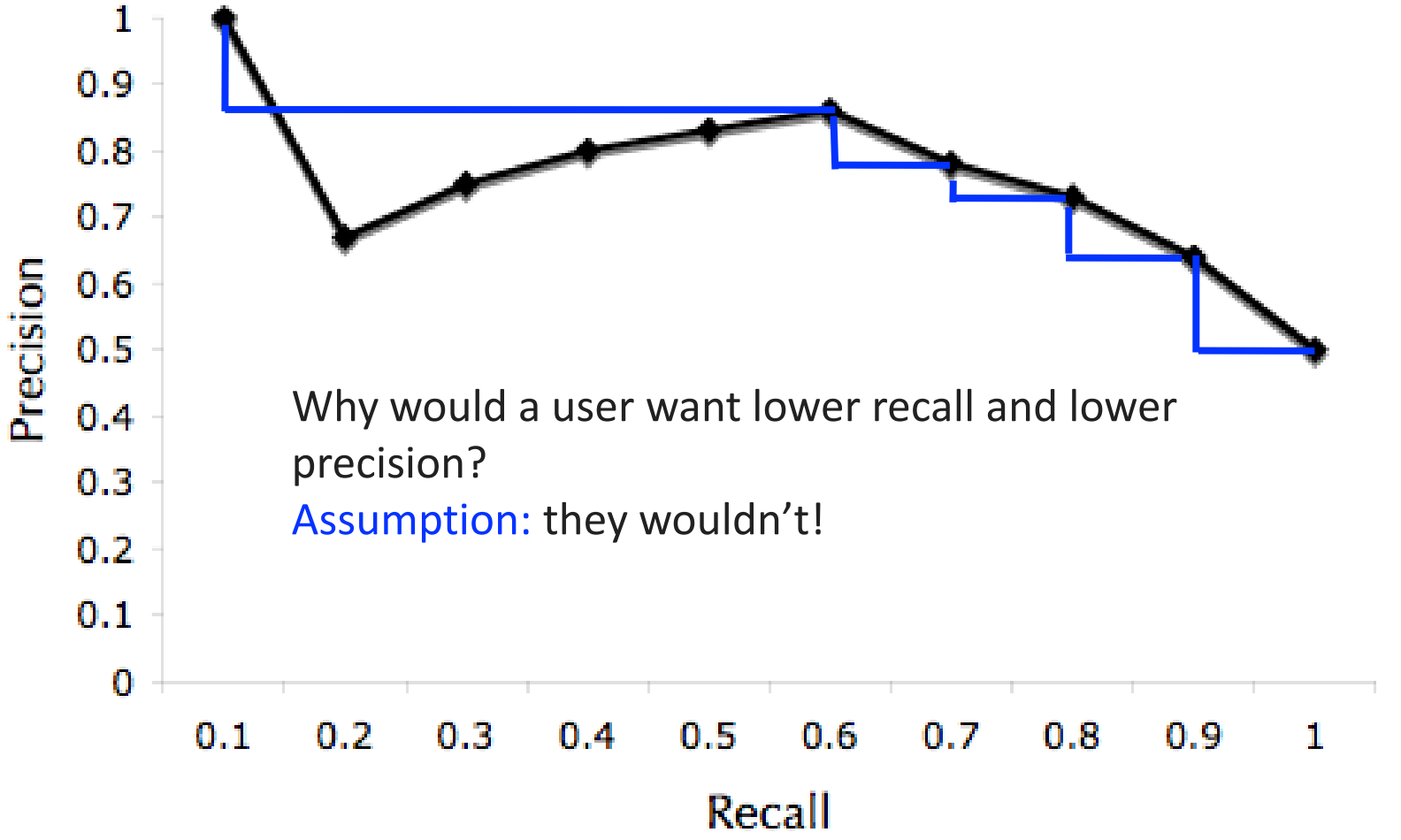
Evaluation Metrics

(5) precision-recall curves



Evaluation Metrics

(5) precision-recall curves

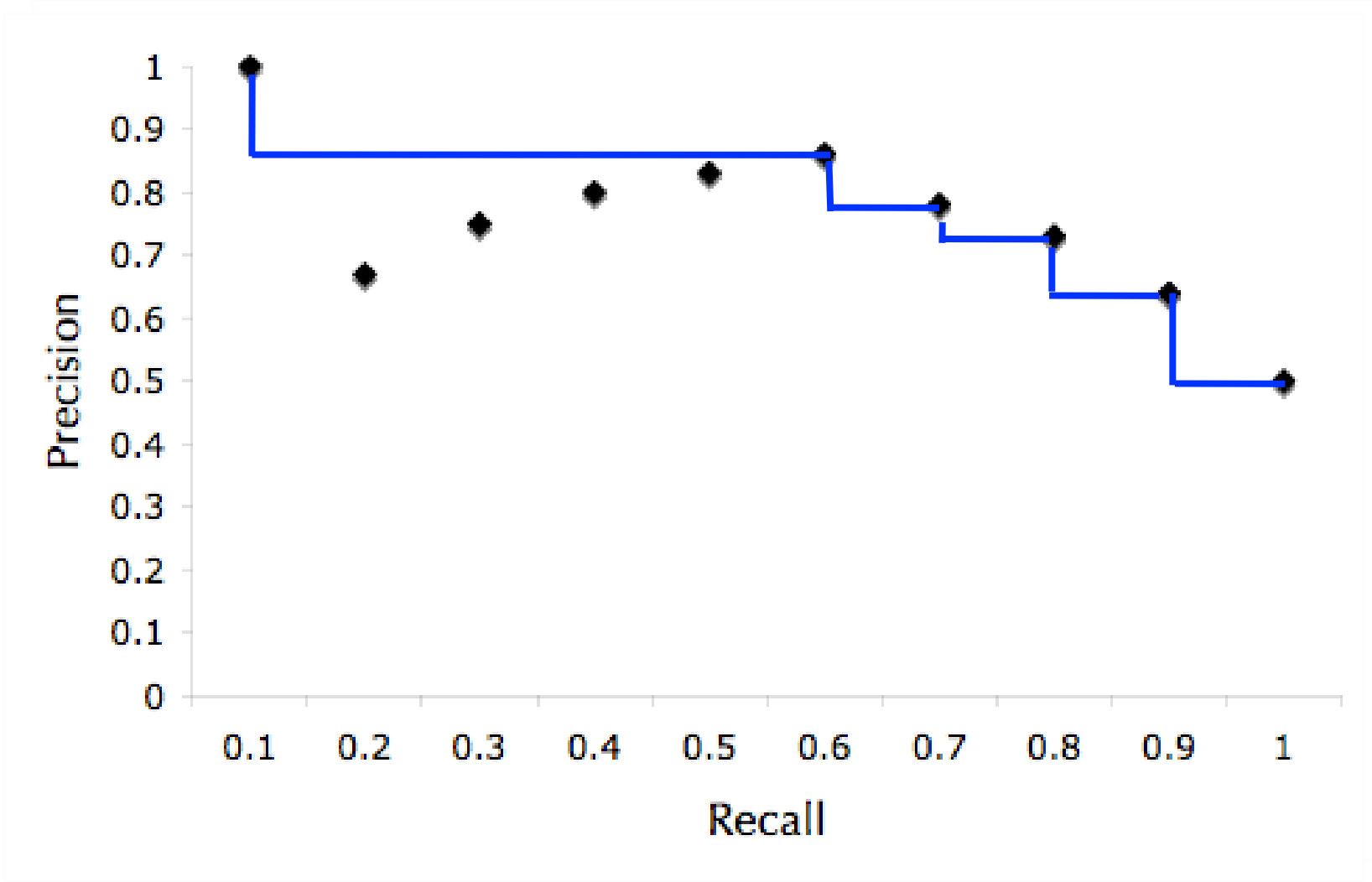


Why would a user want lower recall and lower precision?

Assumption: they wouldn't!

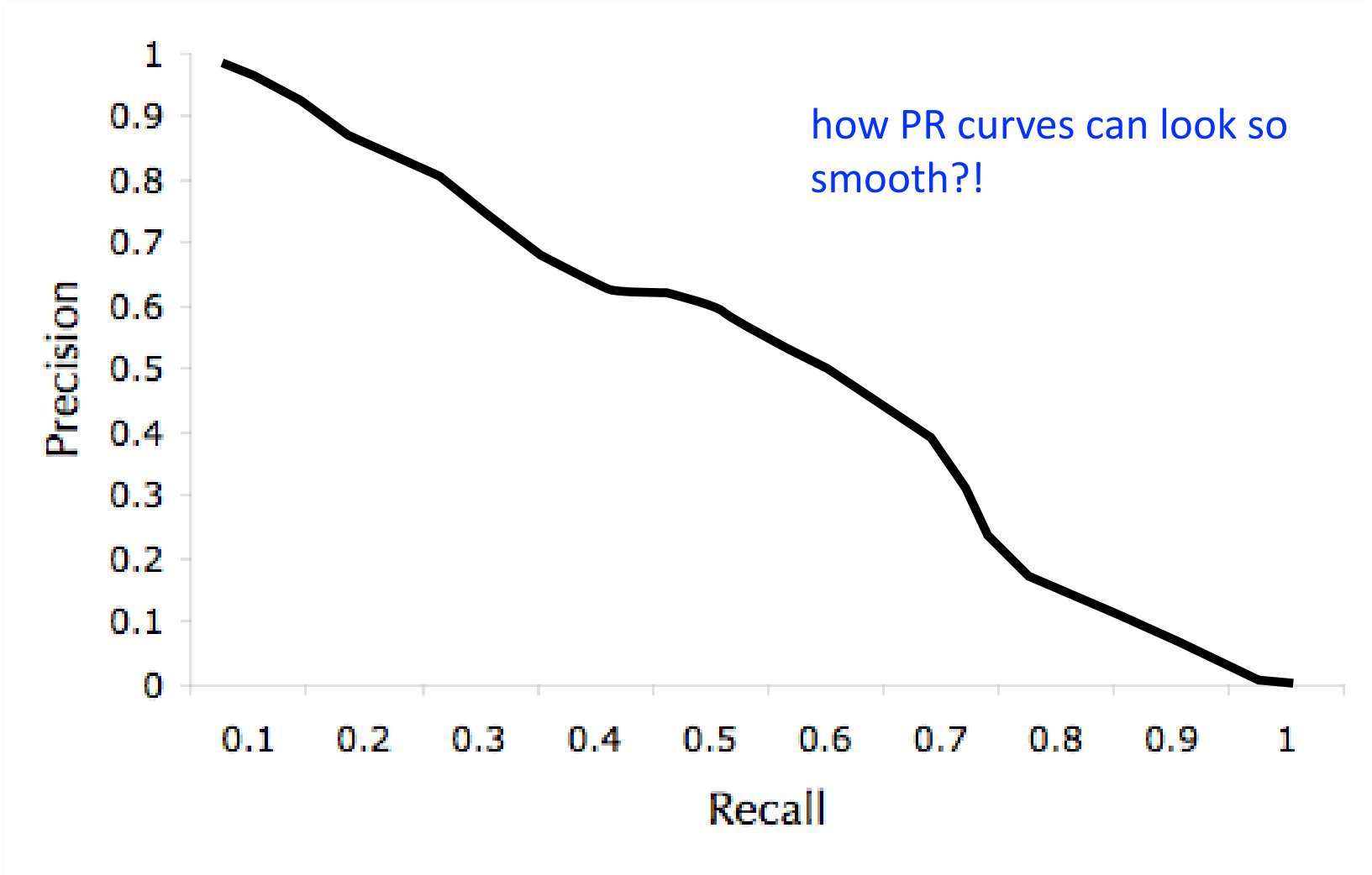
Evaluation Metrics

(5) precision-recall curves



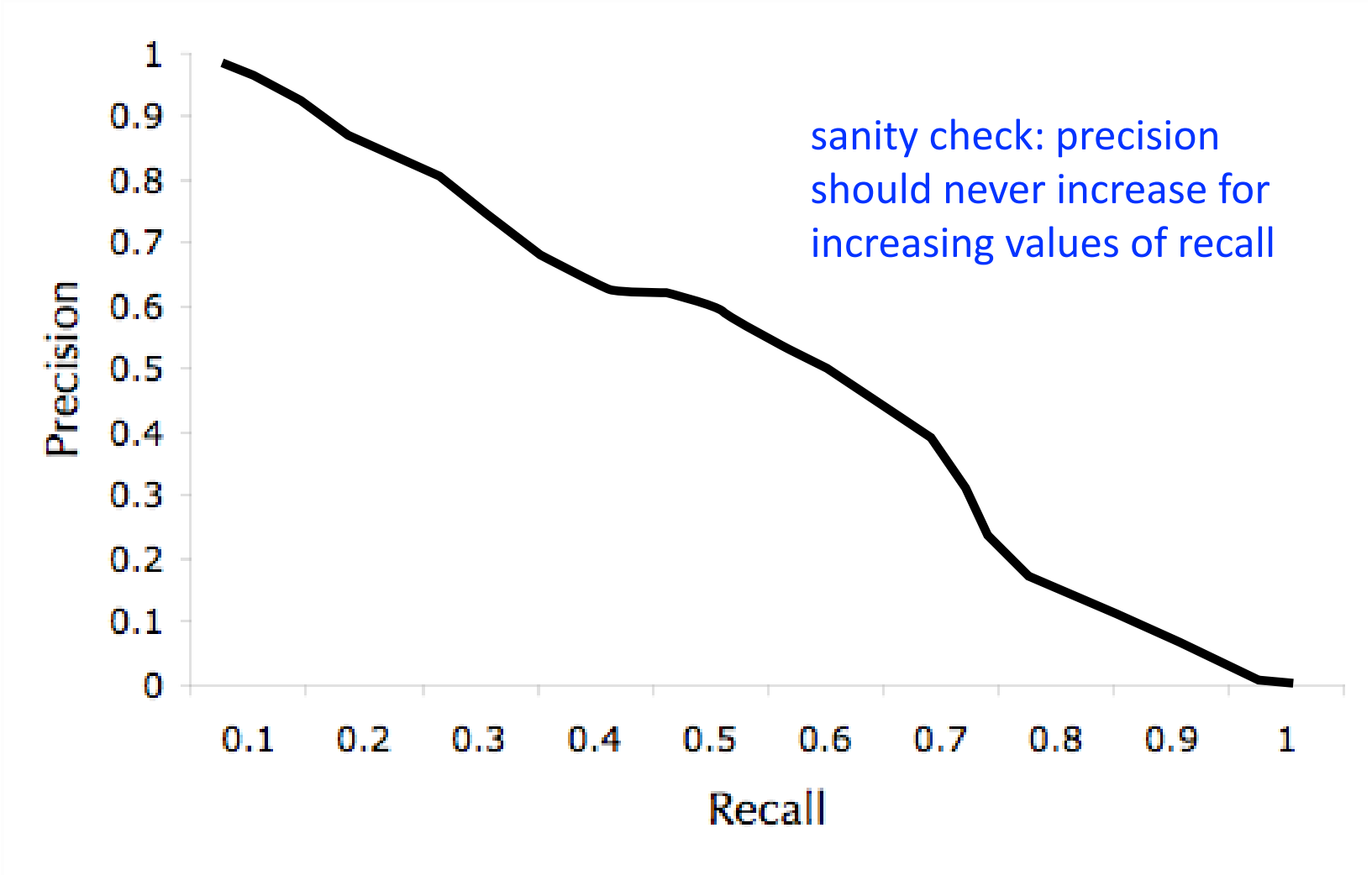
Evaluation Metrics

(5) precision-recall curves



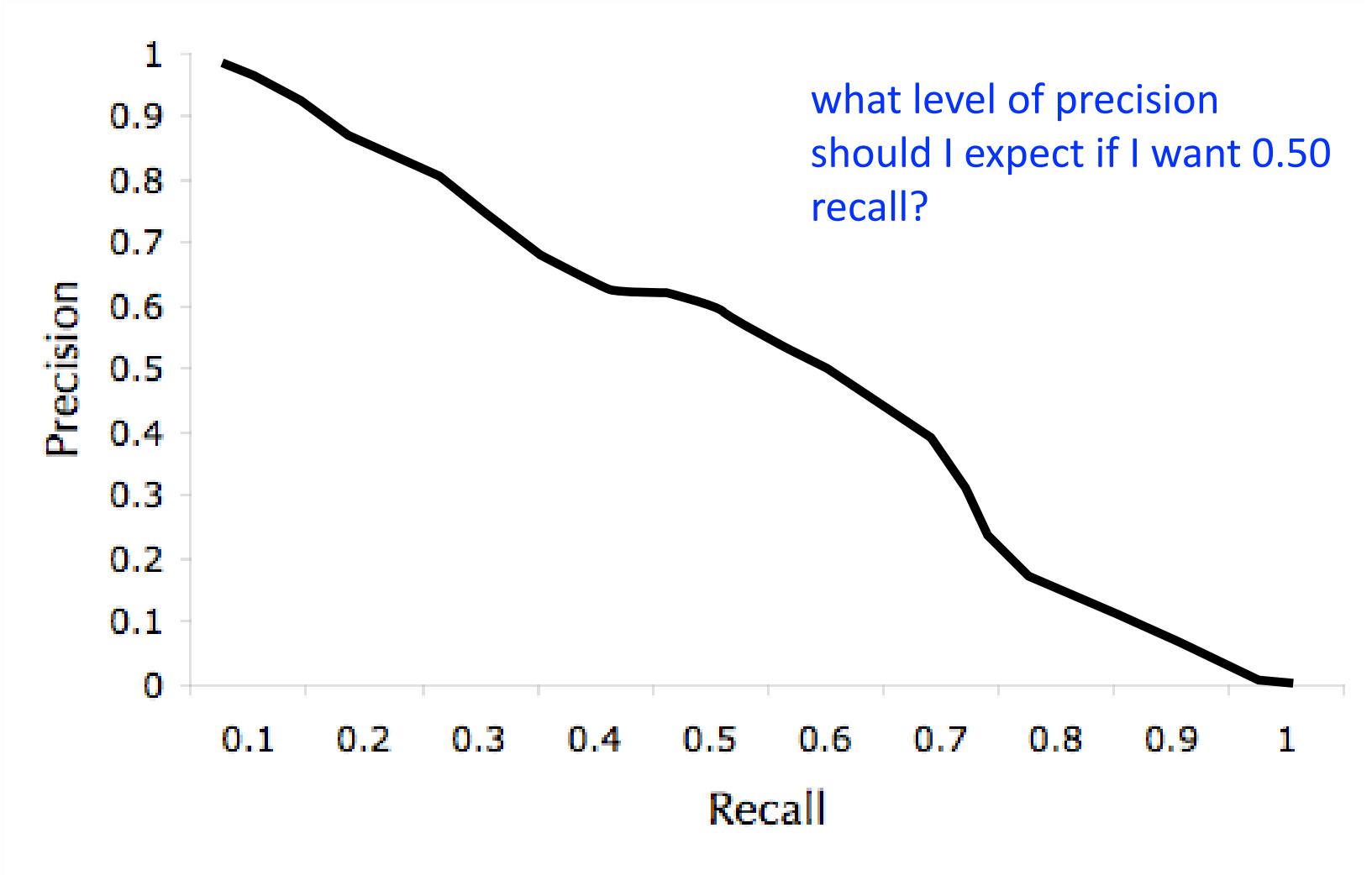
Evaluation Metrics

(5) precision-recall curves



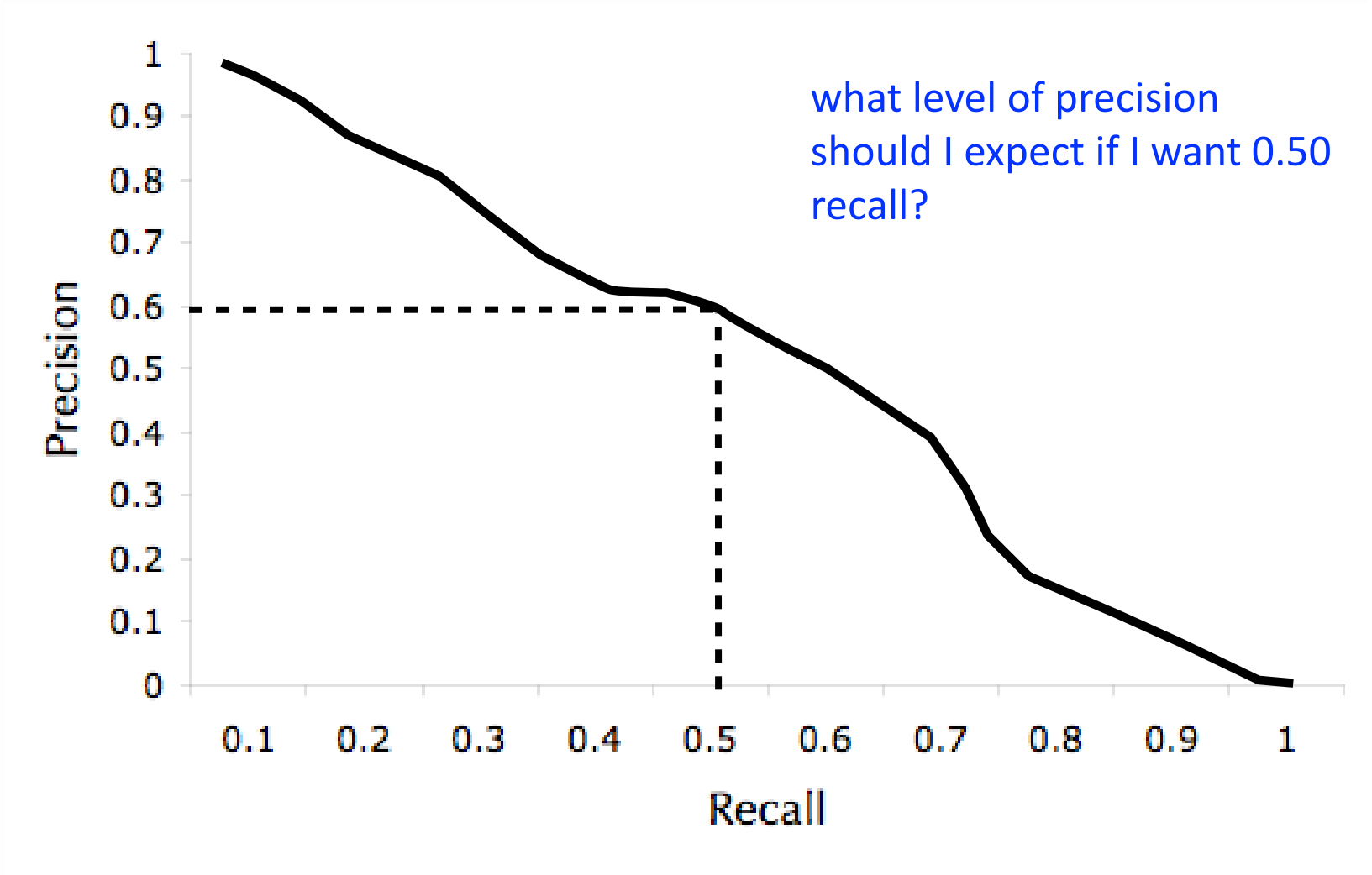
Evaluation Metrics

(5) precision-recall curves



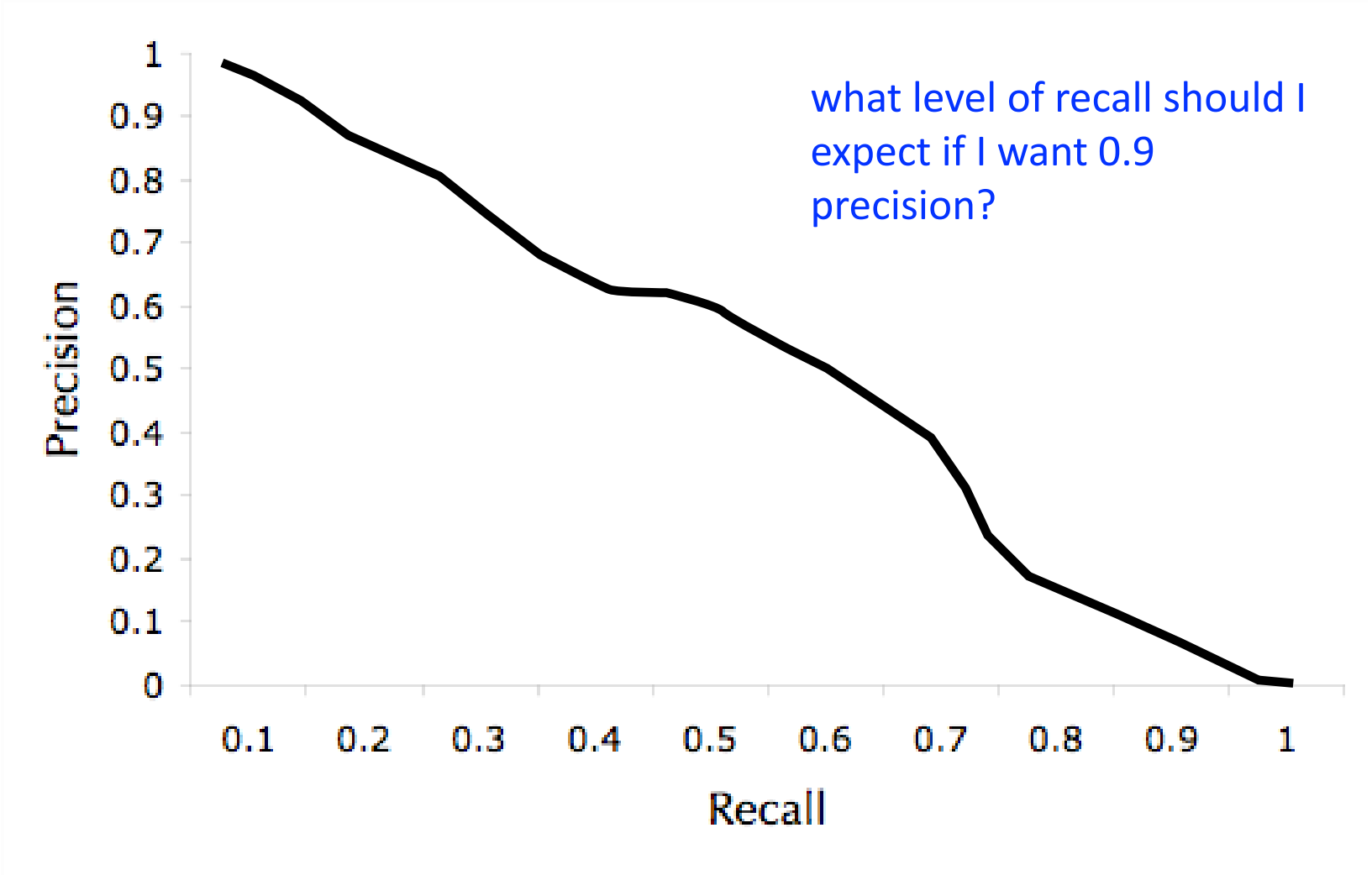
Evaluation Metrics

(5) precision-recall curves



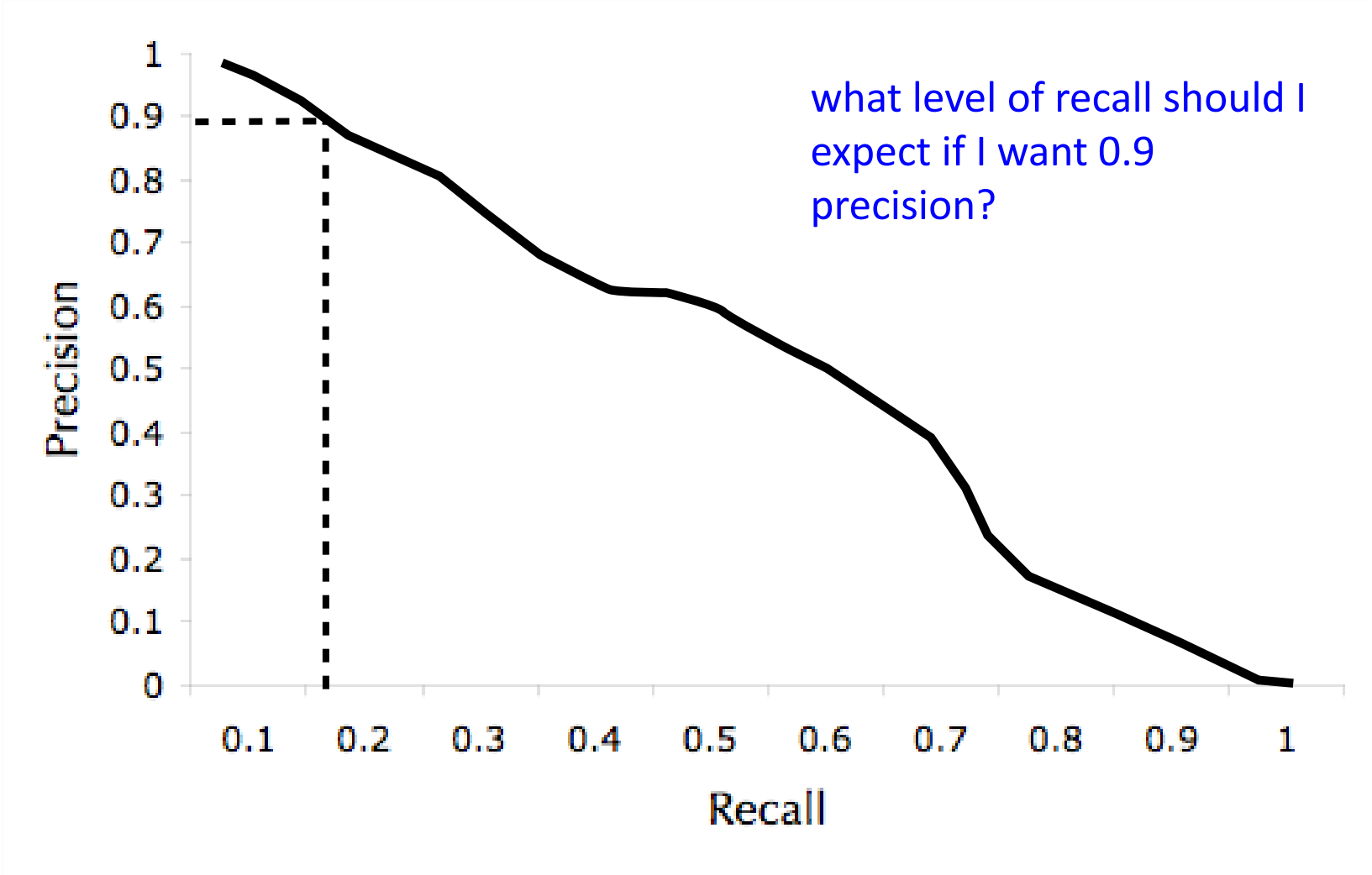
Evaluation Metrics

(5) precision-recall curves



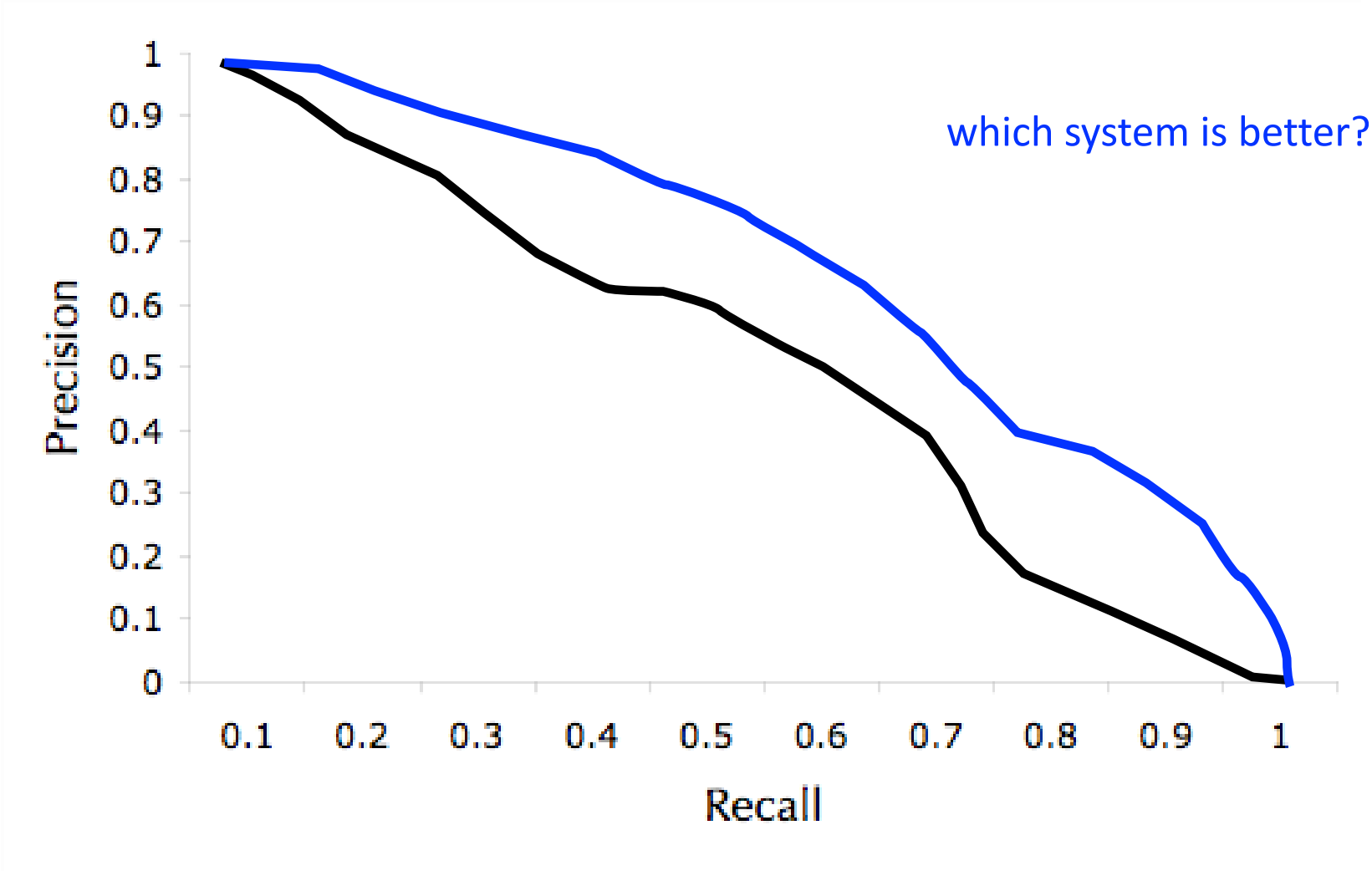
Evaluation Metrics

(5) precision-recall curves



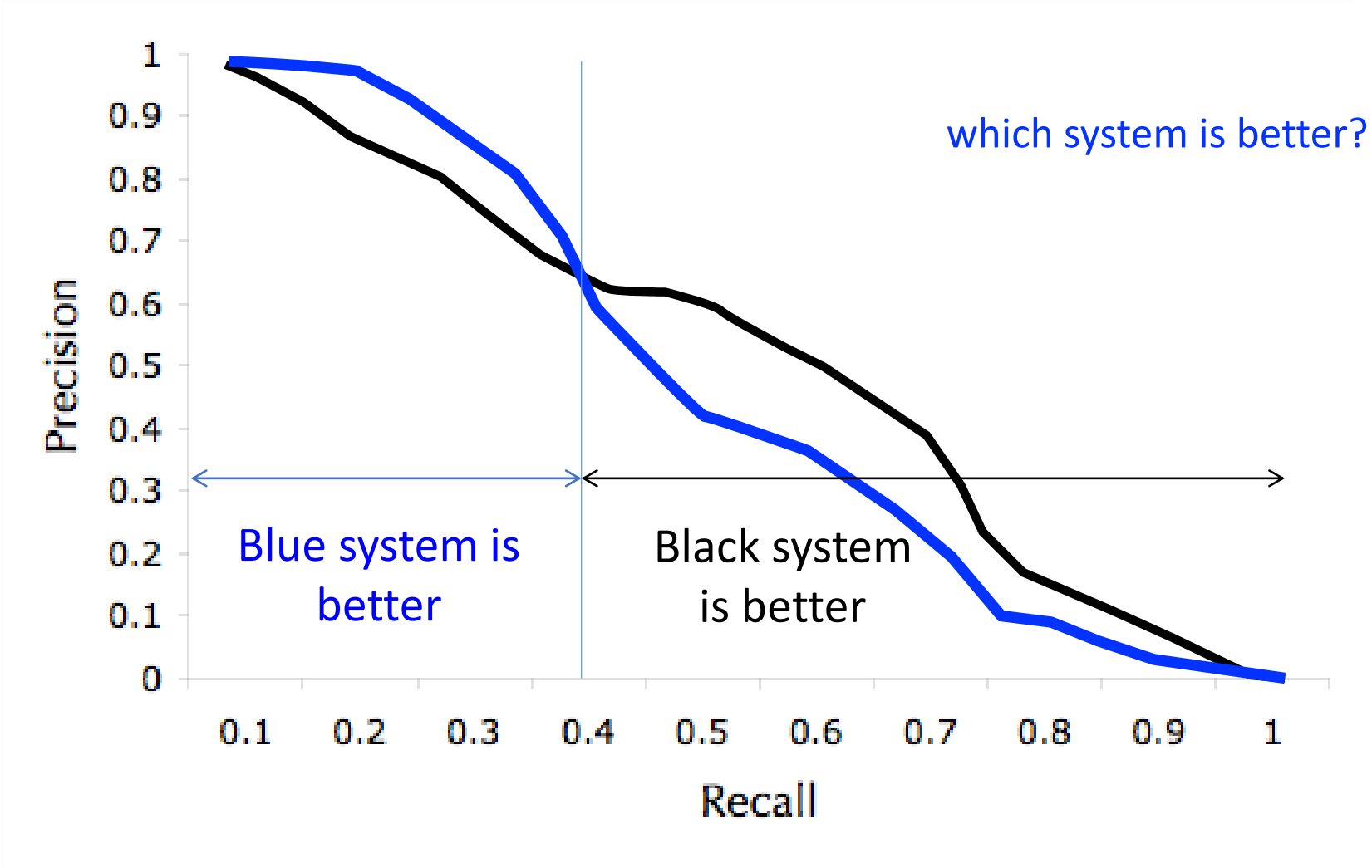
Evaluation Metrics

(5) precision-recall curves



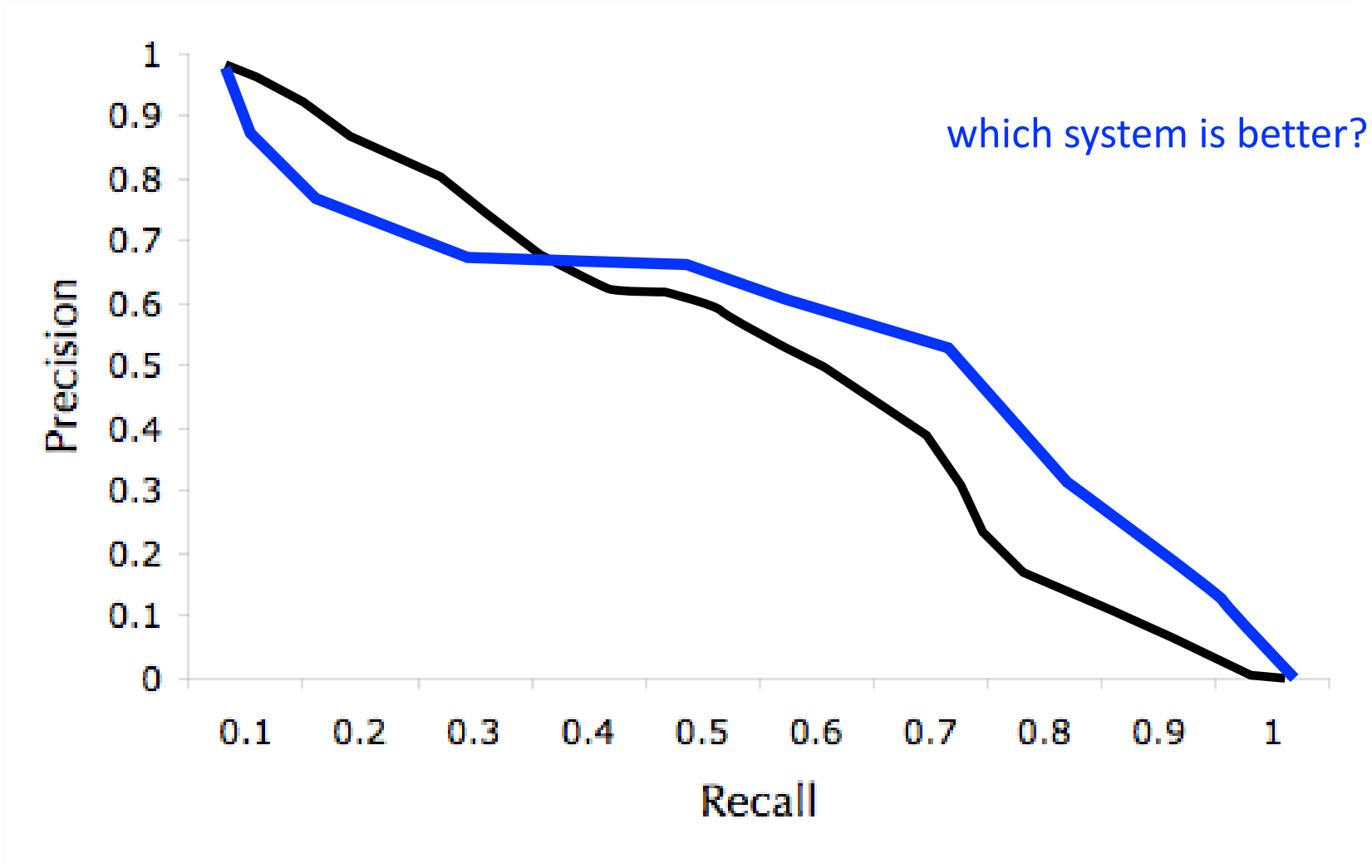
Evaluation Metrics

(5) precision-recall curves



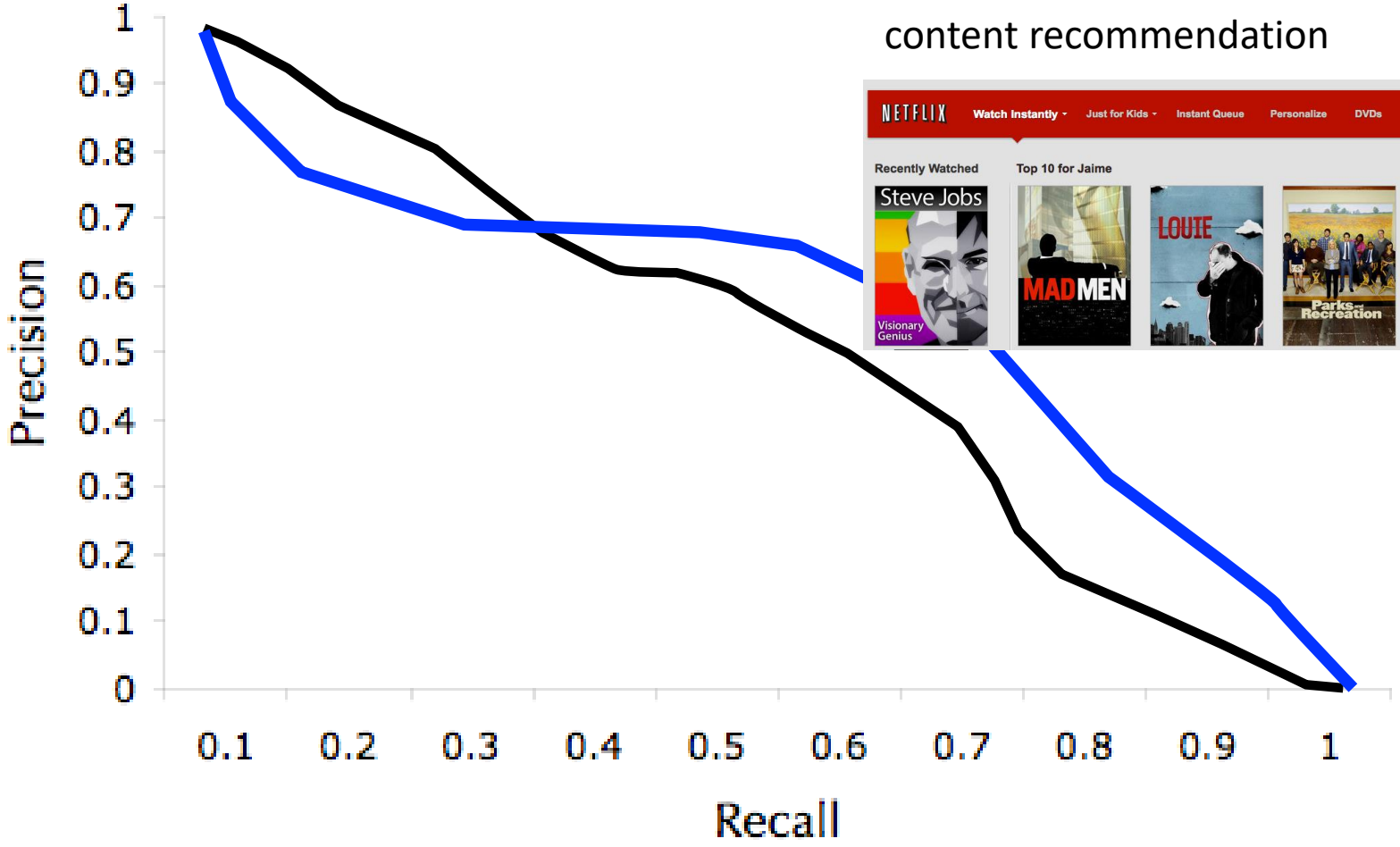
Evaluation Metrics

(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

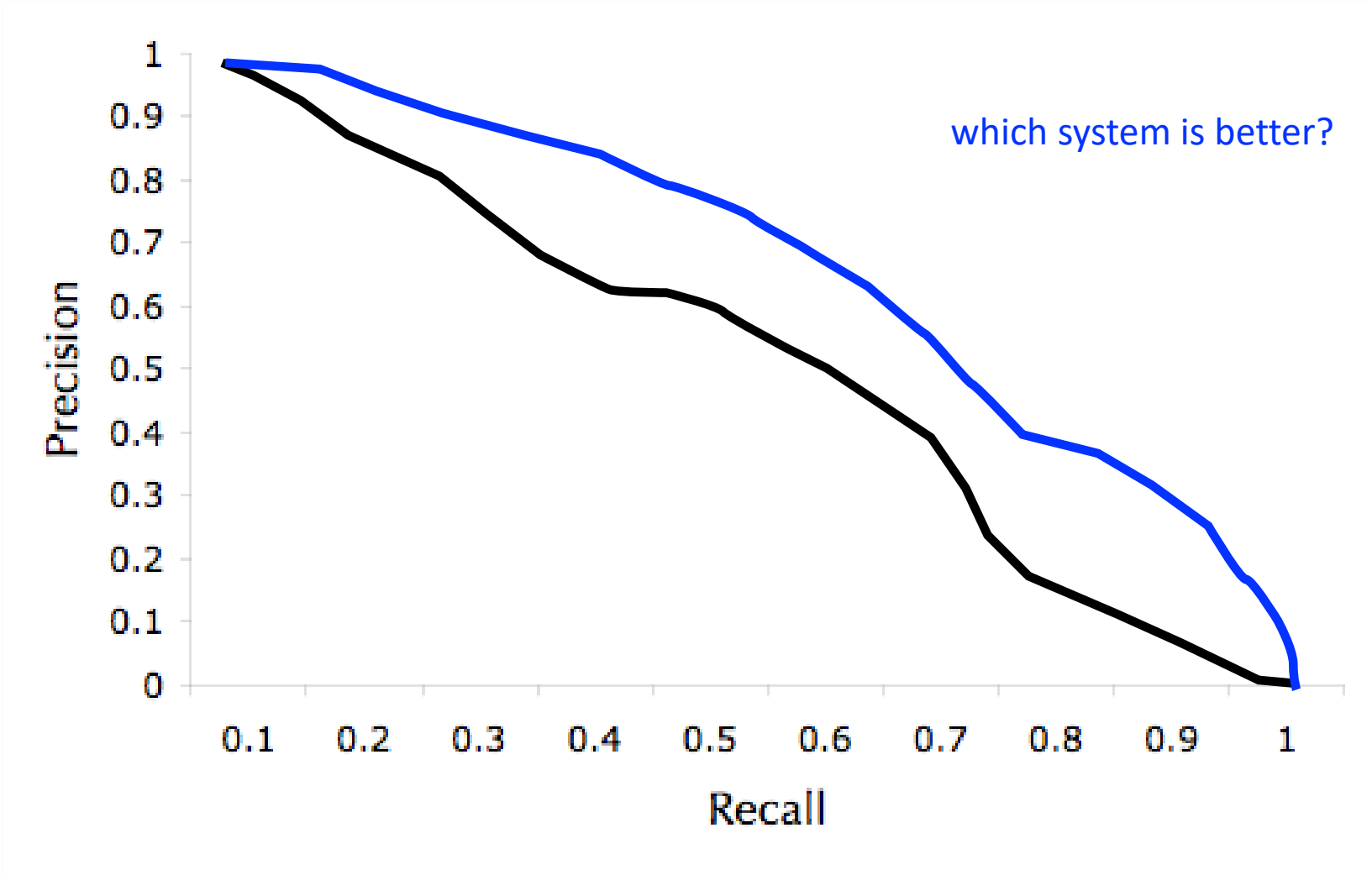
- PR curves show different precision-recall operating points (or **trade-off** points)
- How many false positives will I have to sift through for a desired level of recall?
- How many true positives will I have to miss for a desired level of precision?

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



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		

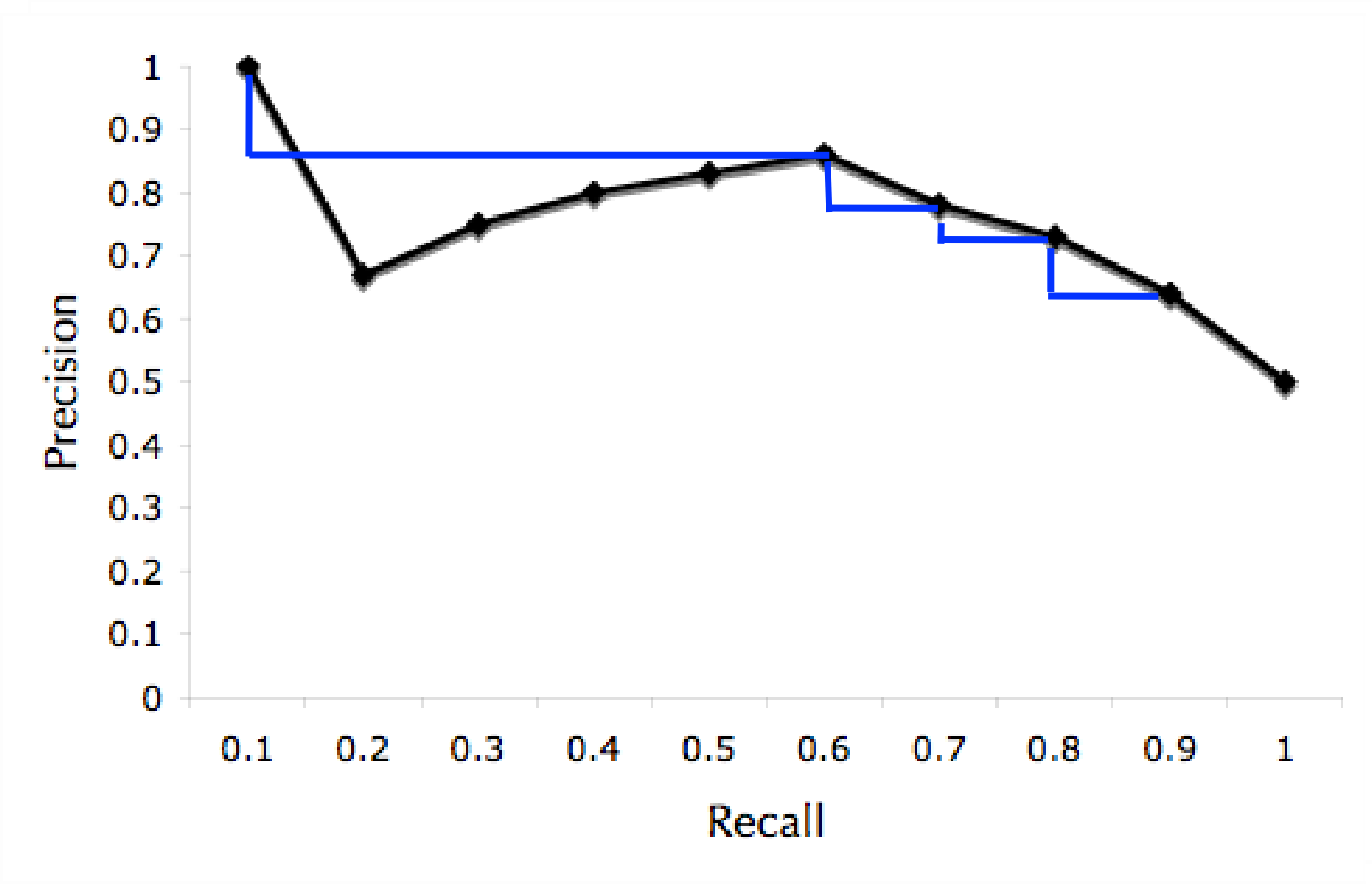
Evaluation Metrics

(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	

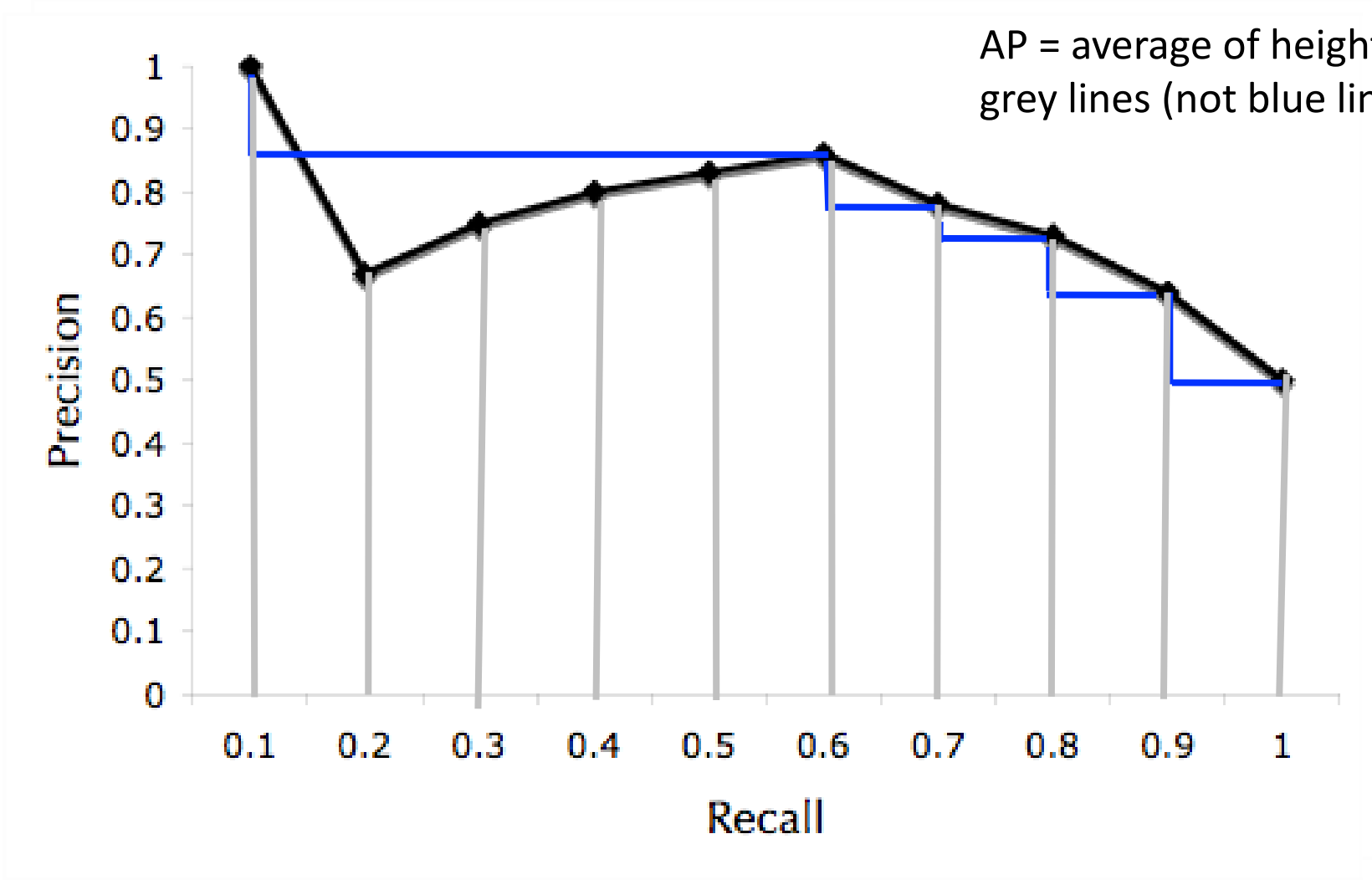
Evaluation Metrics

(6) average precision



Evaluation Metrics

(6) average precision



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 Precisions

Evaluating numeric prediction

- Same strategies: independent test set, cross-validation, significance tests, etc.
- Difference: error measures
- Actual target values: $a_1\ a_2\ \ldots a_n$
- Predicted target values: $p_1\ p_2\ \ldots\ p_n$
- Most popular measure: *mean-squared error*

$$\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}$$

- Easy to manipulate mathematically

(slide courtesy of Witten et al., 2017)

Other measures

- The *root mean-squared error* :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

- The *mean absolute error* is less sensitive to outliers than the mean-squared error:

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$$

- Sometimes *relative* error values are more appropriate (e.g. 10% for an error of 50 when predicting 500)

(slide courtesy of Witten et al., 2017)

Correlation coefficient

- Measures the *statistical correlation* between the predicted values and the actual values

$$\frac{S_{PA}}{\sqrt{S_P S_A}}, \text{ where } S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1}, S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1},$$
$$S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1} \text{ (here, } \bar{a} \text{ is the mean value over the test data)}$$

- Scale independent, between –1 and +1
- Good performance leads to large values!

(slide courtesy of Witten et al., 2017)

Evaluation and Experimentation

- Evaluation Metrics
- **Cross-Validation**
- Significance Tests

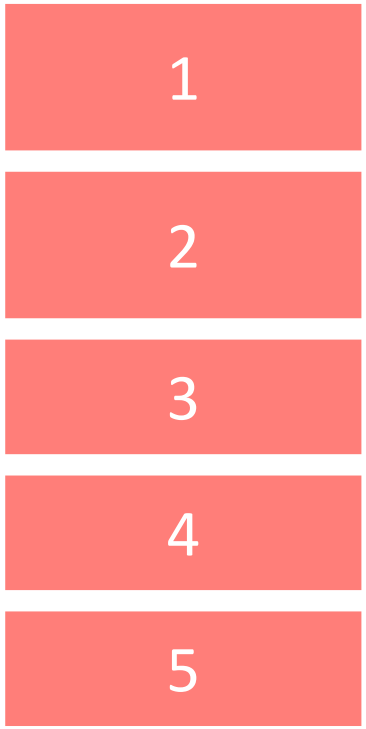
N-fold cross-validation

1. divide the data into N sets of instances
2. use the union of N-1 sets to find the best parameter values
3. measure performance (using the best parameters) on the held-out set
4. do steps 2-3 N times
5. average performance across the N held-out sets

Cross-Validation

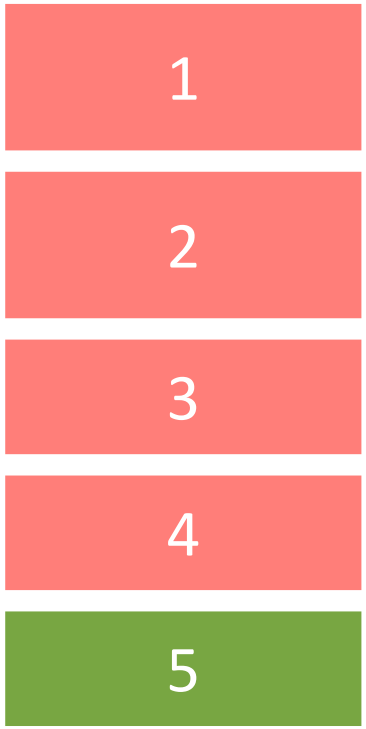
Cross-Validation

- Split the data into N = 5 folds



Cross-Validation

- For each fold, find the parameter value that maximizes performance on the union of N - 1 folds and test (using this parameter value) on the held-out fold.

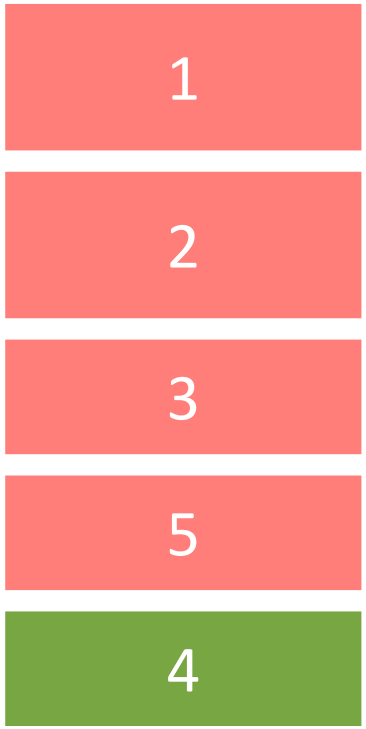


K = 5

F = 0.50

Cross-Validation

- For each fold, find the parameter value that maximizes performance on the union of N - 1 folds and test (using this parameter value) on the held-out fold.

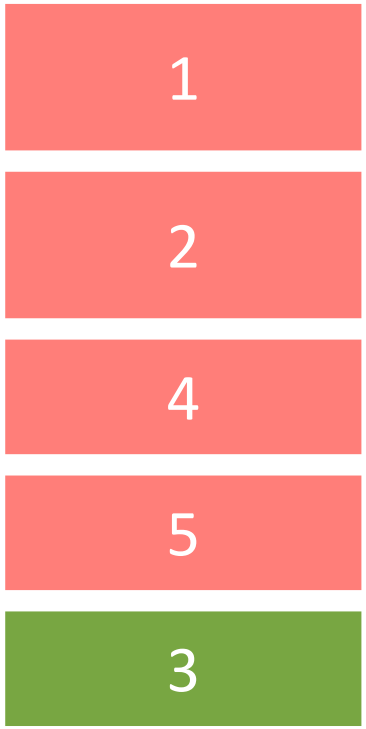


K = 6

F = 0.60

Cross-Validation

- For each fold, find the parameter value that maximizes performance on the union of N - 1 folds and test (using this parameter value) on the held-out fold.

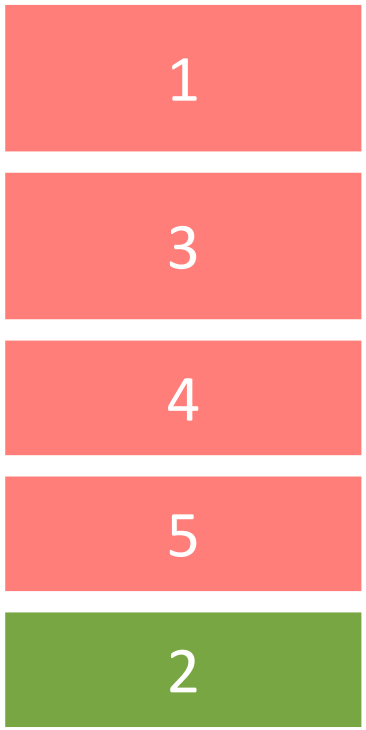


K = 6

F = 0.70

Cross-Validation

- For each fold, find the parameter value that maximizes performance on the union of N - 1 folds and test (using this parameter value) on the held-out fold.

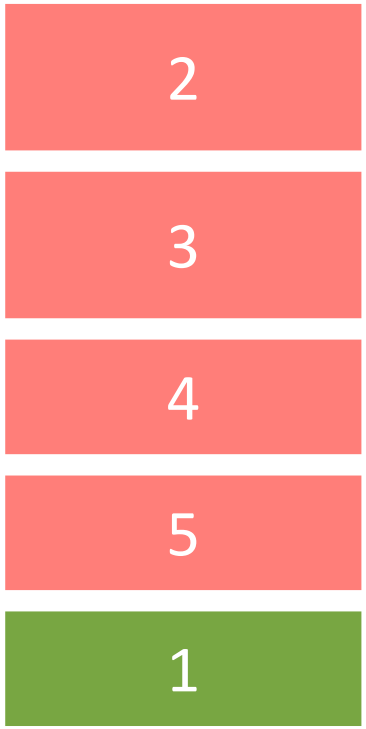


K = 4

F = 0.60

Cross-Validation

- For each fold, find the parameter value that maximizes performance on the union of N - 1 folds and test (using this parameter value) on the held-out fold.

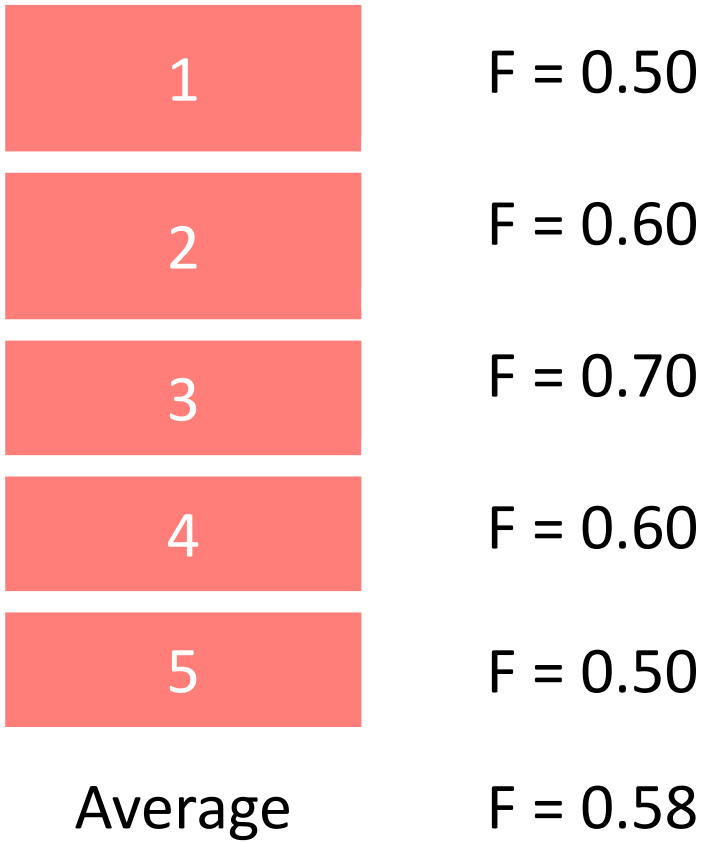


K = 5

F = 0.50

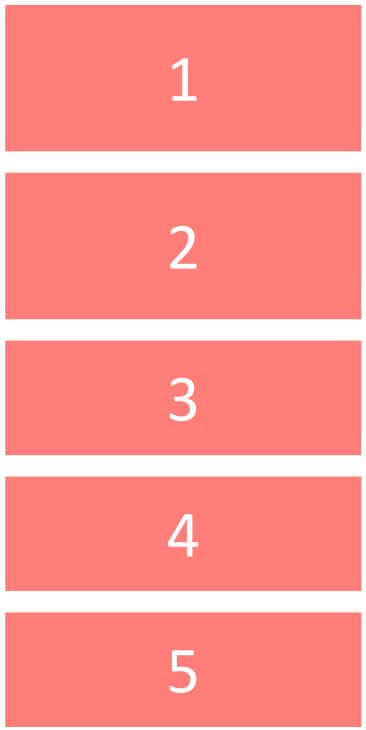
Cross-Validation

- Average the performance across held-out folds



Cross-Validation

- Average the performance across held-out folds



F = 0.50

F = 0.60

F = 0.70

F = 0.60

F = 0.50

Average F = 0.58

Advantages and Disadvantages?

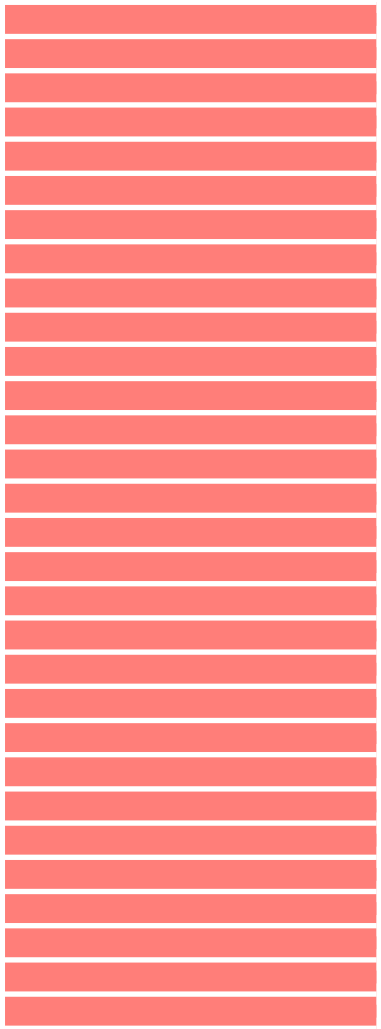
N-Fold Cross-Validation

- Advantage
  - multiple rounds of generalization performance.
- Disadvantage
  - ultimately, we'll tune parameters on the whole dataset and send our system into the world.
  - a model trained on 100% of the data should perform better than one trained on 80%.
  - thus, we may be underestimating the model's performance!

Leave-One-Out Cross-Validation

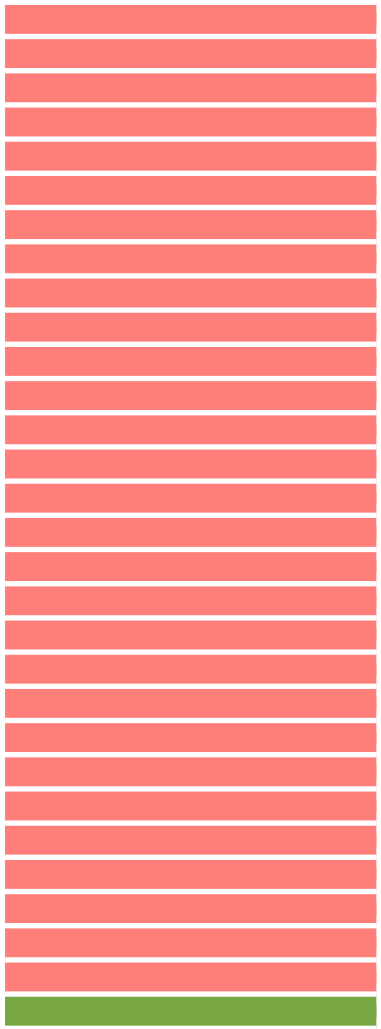
Leave-One-Out Cross-Validation

- Split the data into N folds of 1 instance each



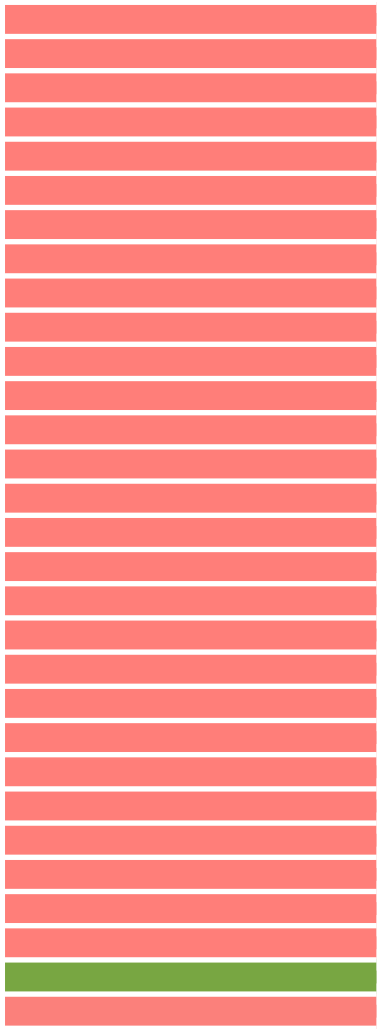
Leave-One-Out Cross-Validation

- For each instance, find the parameter value that maximize performance on for the other instances and and test (using this parameter value) on the held-out instance.



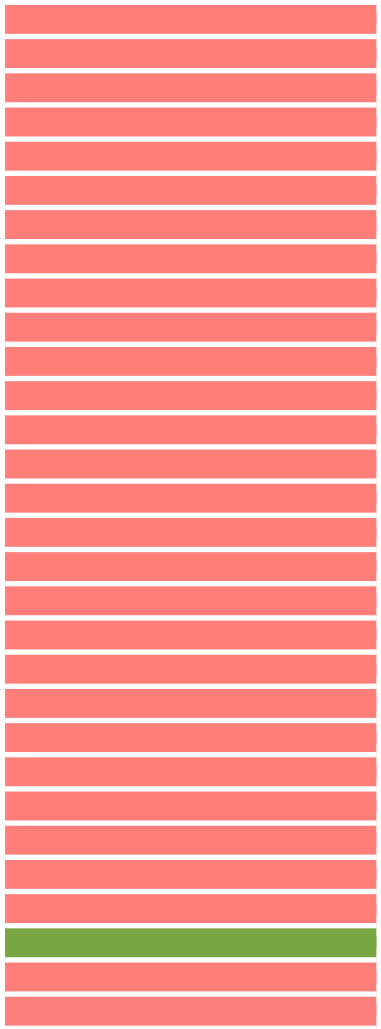
Leave-One-Out Cross-Validation

- For each instance, find the parameter value that maximize performance on for the other instances and and test (using this parameter value) on the held-out instance.



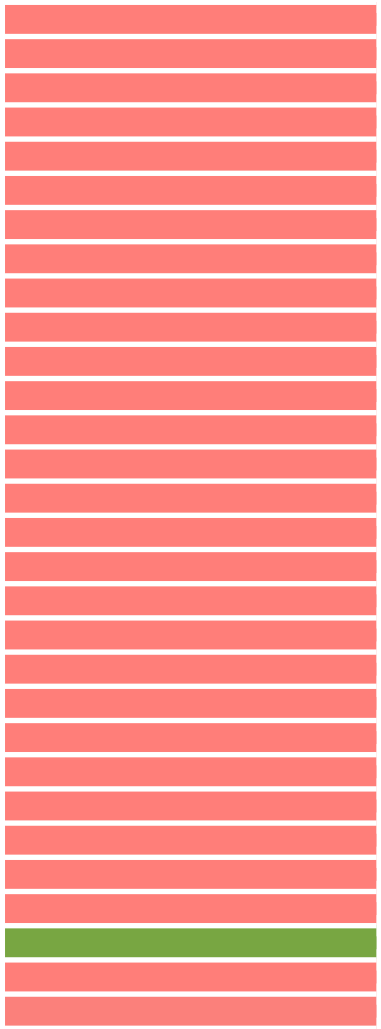
Leave-One-Out Cross-Validation

- For each instance, find the parameter value that maximize performance on for the other instances and and test (using this parameter value) on the held-out instance.
- And so on ...
- Finally, average the performance for each held-out instance



Leave-One-Out Cross-Validation

- For each instance, find the parameter value that maximize performance on for the other instances and and test (using this parameter value) on the held-out instance.
- And so on ...
- Finally, average the performance for each held-out instance



Advantages and Disadvantages?

Leave-One-Out Cross-Validation

- Advantages
  - multiple rounds of generalization performance.
  - each training fold is as similar as possible to the one we will ultimately use to tune parameters before sending the system out into the world.
- Disadvantage
  - our estimate of generalization performance may be artificially high
  - we are likely to try lots of different things and pick the one with the best “generalization” performance
  - still indirectly over-training to the dataset (sigh...)

Evaluation and Experimentation

- Evaluation Metrics
- Cross-Validation
- **Significance Tests**

Comparing Systems

- Train and test both systems using 10-fold cross validation
- Use the same folds for both systems
- Compare the difference in average performance across held-out folds

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Significance Tests
motivation

- Why would it be risky to conclude that System B is better System A?
- Put differently, what is it that we're trying to achieve?

Significance Tests
motivation

- In theory: that the average performance of System B is greater than the average performance of System A for all possible test sets.
- However, we don't have all test sets. We have a sample
- And, this sample may favor one system vs. the other!

Significance Tests
definition

- A significance test is a statistical tool that allows us to determine whether a difference in performance reflects a true pattern or just random chance

Significance Tests
ingredients

- Test statistic: a measure used to judge the two systems (e.g., the difference between their average F-measure)
- Null hypothesis: no “true” difference between the two systems
- P-value: take the value of the observed test statistic and compute the probability of observing a statistical summary (e.g., sample mean difference between two compared groups) that is large (or larger) under the null hypothesis

Significance Tests
ingredients

- If the p-value is large, we cannot reject the null hypothesis
- That is, we cannot claim that one system is better than the other
- If the p-value is small ($p<0.05$), we can reject the null hypothesis
- That is, the observed test-statistic is not due to random chance

Comparing Systems

- P-value: the probability of observing a difference **equal to or greater than** 0.07 under the null hypothesis (i.e., the systems are actually equal).

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Fisher's Randomization Test
procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

Step 1: for each fold, flip a coin and if it lands 'heads', flip the result between System A and B

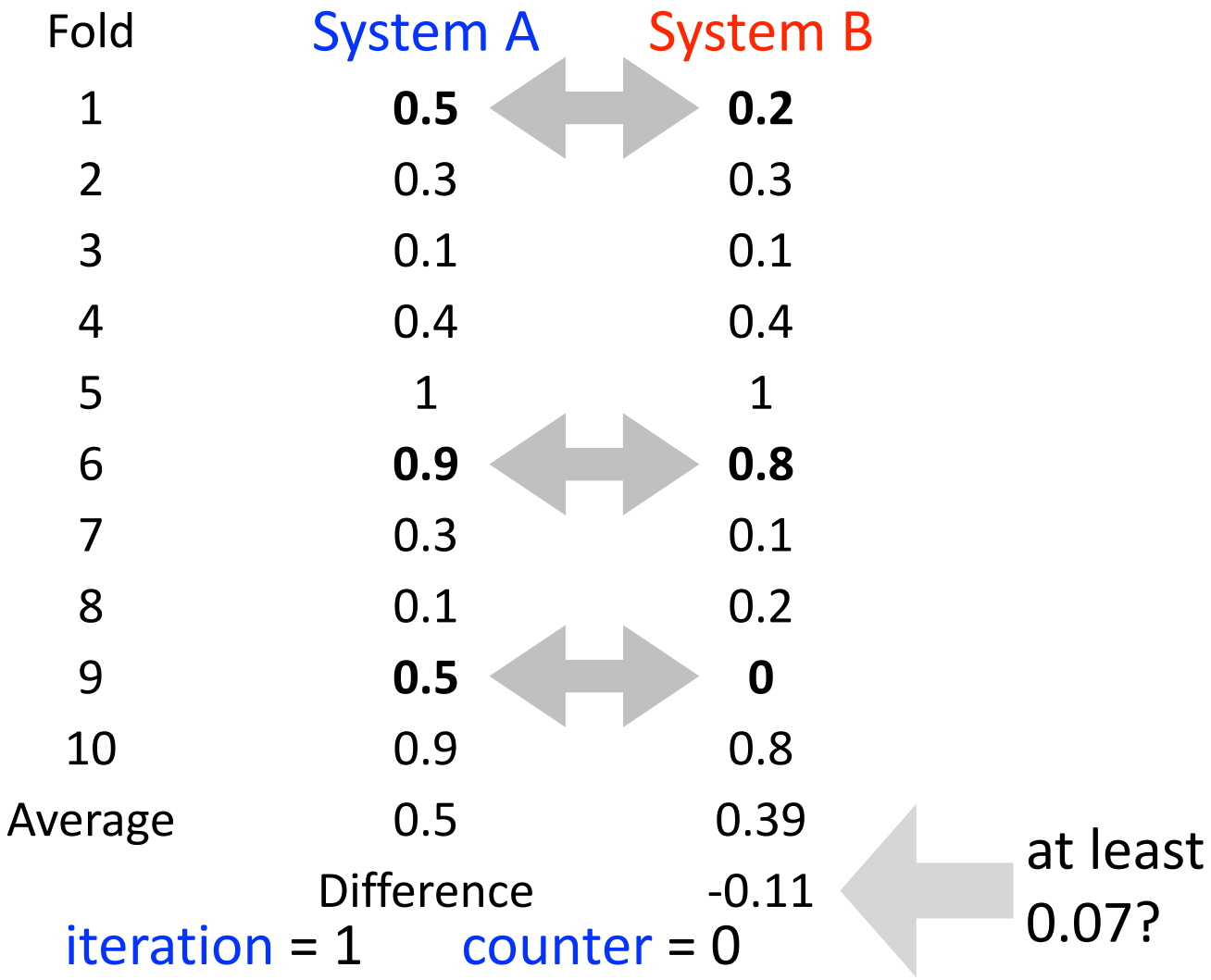
Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter

- **Output:** counter / N

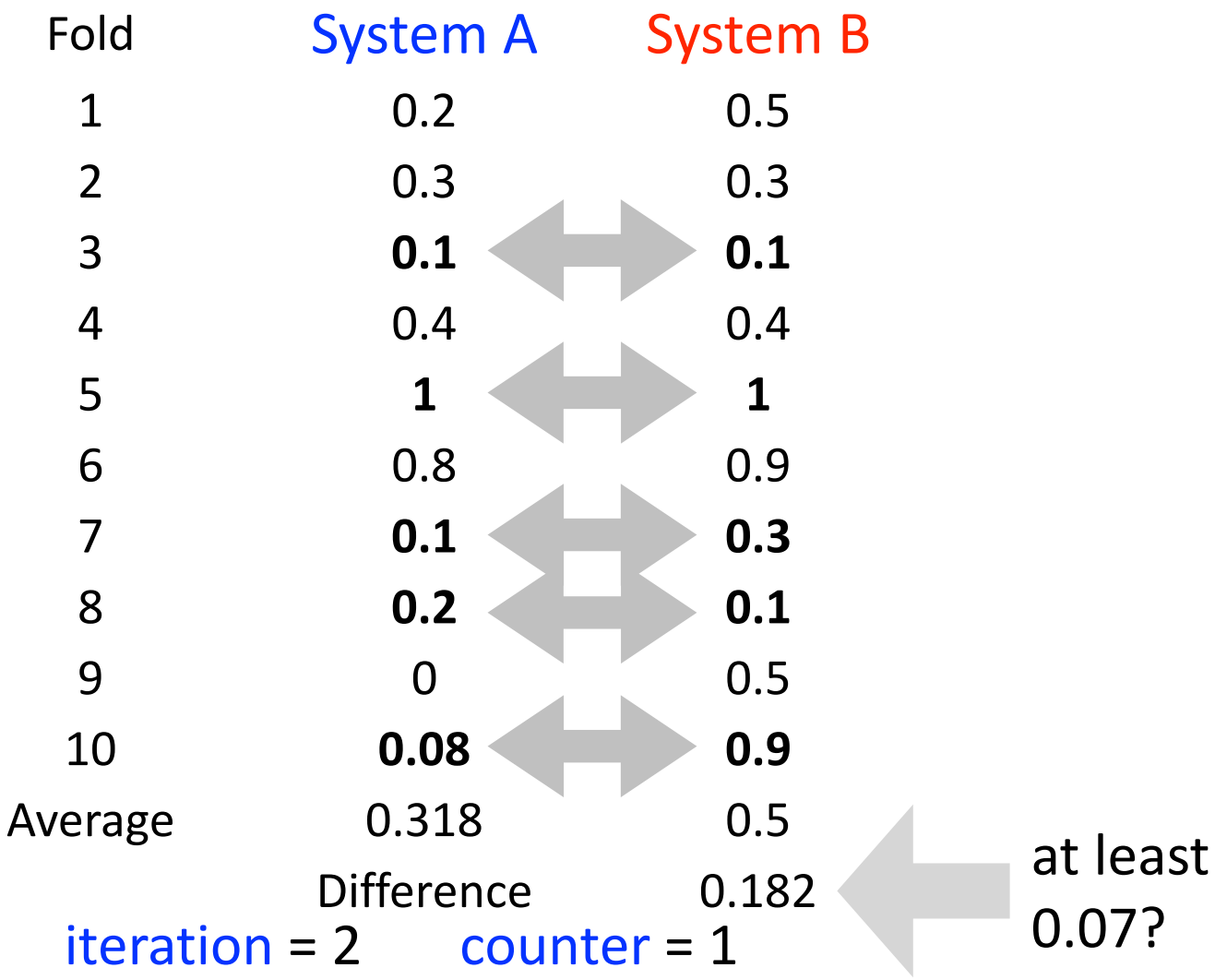
Fisher's Randomization Test

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

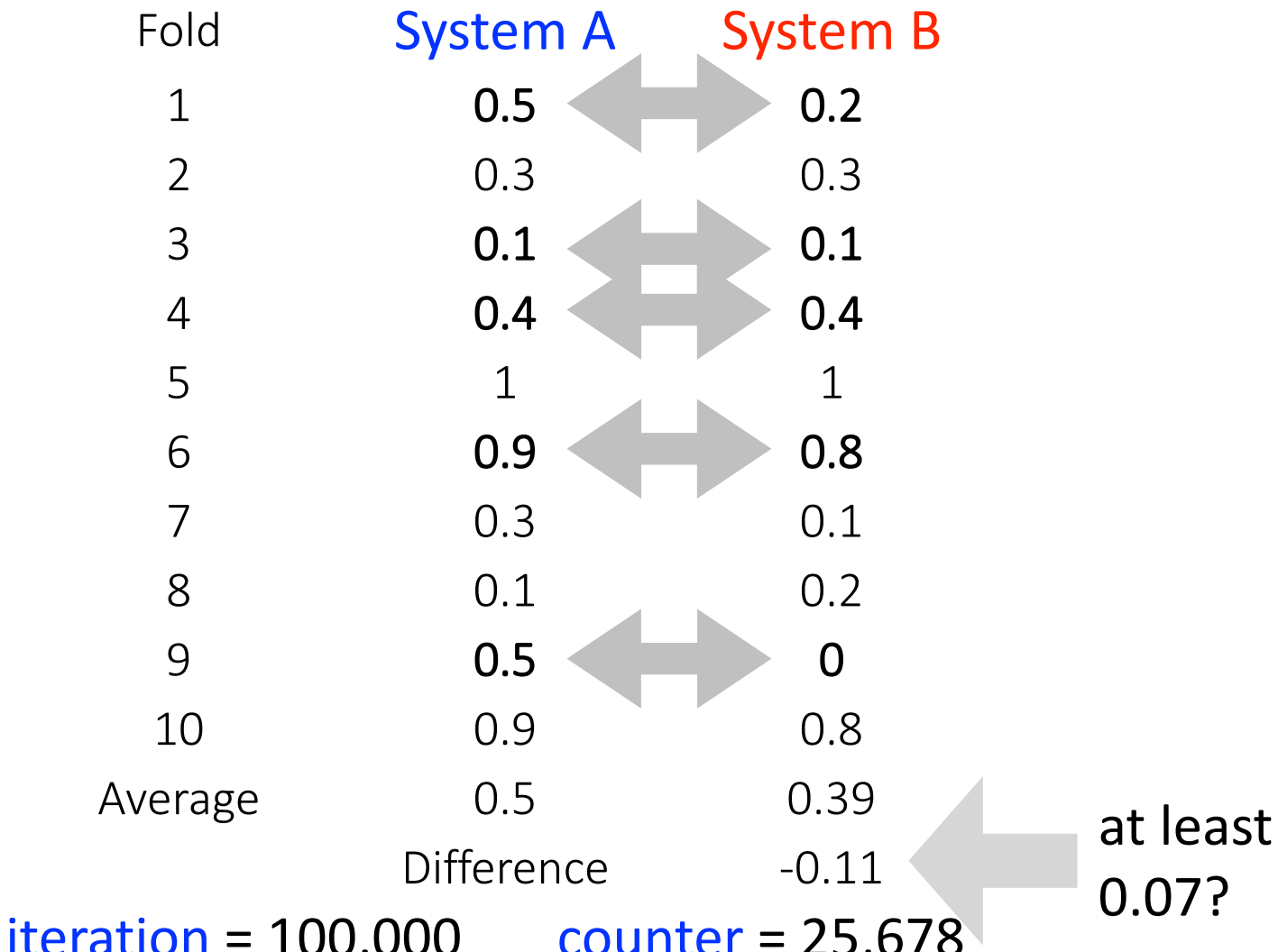
Fisher's Randomization Test



Fisher's Randomization Test



Fisher's Randomization Test



Fisher's Randomization Test
procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

Step 1: for each query, flip a coin and if it lands 'heads', flip the result between System A and B

Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter

- **Output:** counter / N = (25,678/100,000) = 0.25678

Fisher's Randomization Test

- Under the null hypothesis, the probability of observing a value of the test statistic of 0.07 or greater is about 0.26.
- Because $p > 0.05$, we cannot confidently say that the value of the test statistic is not due to random chance.
- A difference between the average F-measure values of 0.07 is not significant

Fisher's Randomization Test
procedure

- **Inputs:** counter = 0, N = 100,000
- Repeat N times:

Step 1: for each query, flip a coin and if it lands 'heads', flip the result between System A and B

Step 2: see whether the test statistic is equal to or greater than the one observed and, if so, increment counter

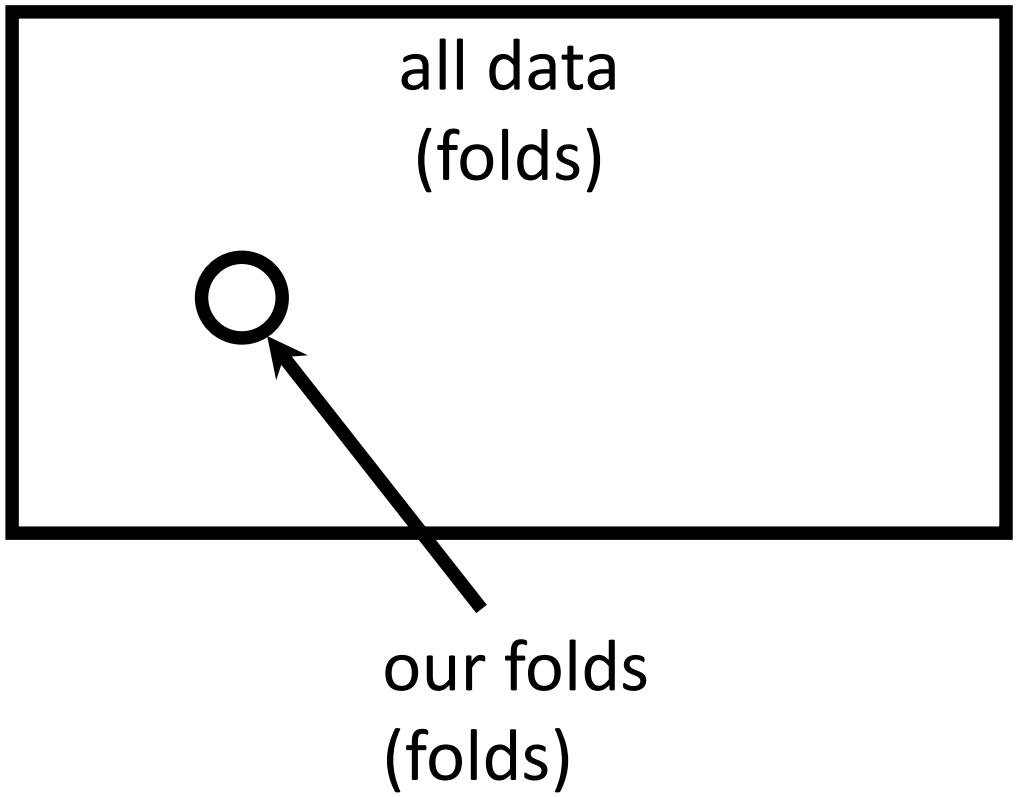
- **Output:** counter / N = (25,678/100,00) = 0.25678

This is a one-tailed test (B > A).
We can modify it to be a two-tailed test (|B| > A)

Bootstrap-Shift Test

motivation

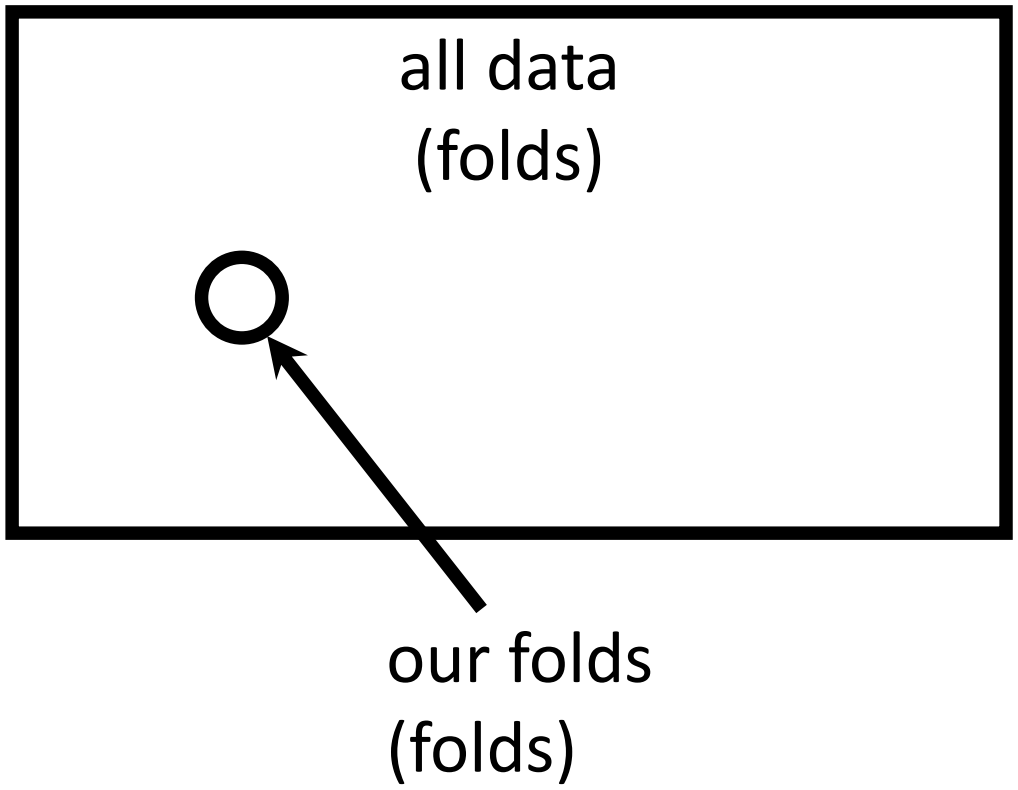
- Our sample is a representative sample of all data



Bootstrap-Shift Test

motivation

- If we sample (with replacement) from our sample, we can generate a new representative sample of all data



Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute average of numbers in T
- **Step 4:** reduce every number in T by average and put in T'
- **Output:** % of numbers in T' greater than or equal to the observed test statistic

Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute average of numbers in T
- **Step 4:** reduce every number in T by average and put in T'
- **Output:** % of numbers in T greater than or equal to the observed test statistic

Bootstrap-Shift Test

Fold	System A	System B
1	0.2	0.5
2	0.3	0.3
3	0.1	0.1
4	0.4	0.4
5	1	1
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.41	0.48
	Difference	0.07

Bootstrap-Shift Test

Fold	System A	System B	sample
1	0.2	0.5	**0**
2	0.3	0.3	**1**
3	0.1	0.1	**2**
4	0.4	0.4	**2**
5	1	1	**0**
6	0.8	0.9	**1**
7	0.3	0.1	**1**
8	0.1	0.2	**1**
9	0	0.5	**2**
10	0.9	0.8	**0**

iteration = 1

Bootstrap-Shift Test

Fold	System A	System B
2	0.3	0.3
3	0.1	0.1
3	0.1	0.1
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
9	0	0.5
Average	0.25	0.35
	Difference	**0.1**

iteration = 1

T = {0.10}

Bootstrap-Shift Test

Fold	System A	System B	sample
1	0.2	0.5	**0**
2	0.3	0.3	**0**
3	0.1	0.1	**3**
4	0.4	0.4	**2**
5	1	1	**0**
6	0.8	0.9	**1**
7	0.3	0.1	**1**
8	0.1	0.2	**1**
9	0	0.5	**1**
10	0.9	0.8	**1**

T = {0.10}

iteration = 2

Bootstrap-Shift Test

Fold	System A	System B
3	0.1	0.1
3	0.1	0.1
3	0.1	0.1
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
9	0	0.5
10	0.9	0.8
Average	0.32	0.36
	Difference	**0.04**

iteration = 2

T = {0.10, 0.04}

Bootstrap-Shift Test

Fold	System A	System B
1	0.2	0.5
1	0.2	0.5
4	0.4	0.4
4	0.4	0.4
4	0.4	0.4
6	0.8	0.9
7	0.3	0.1
8	0.1	0.2
8	0.1	0.2
10	0.9	0.8
Average	0.38	0.44
	Difference	**0.06**

iteration = 100,000

T = {0.10,
0.04,
……,
0.06}

Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute average of numbers in T
- **Step 4:** reduce every number in T by average and put in T'
- **Output:** % of numbers in T' greater than or equal to the observed test statistic

Bootstrap-Shift Test
procedure

- For the purpose of this example, let's assume N = 10.

T = {**0.10**,
0.04,
0.21,
0.20,
0.13,
0.09,
0.22,
0.07,
0.03,
0.11}

Step 3

T'= {**-0.02**,
-0.08,
0.09,
0.08,
0.01,
-0.03,
0.10,
-0.05,
-0.09,
-0.01}

Step 4

Average = 0.12

Bootstrap-Shift Test

procedure

- **Inputs:** Array T = {}, N = 100,000
- Repeat N times:

  Step 1: sample 10 folds (with replacement) from our set of 10 folds (called a subsample)

  Step 2: compute test statistic associated with new sample and add to T

- **Step 3:** compute average of numbers in T
- **Step 4:** reduce every number in T by average
- **Output:** % of numbers in T' greater than or equal to the observed test statistic

Bootstrap-Shift Test

procedure

- **Output:** (3/10) = **0.30**

T = {0.10,
0.04,
0.21,
0.20,
0.13,
0.09,
0.22,
0.07,
0.03,
0.11}

Step 3

T'= {-0.02,
-0.08,
0.09,
0.08,
0.01,
-0.03,
0.10,
-0.05,
-0.09,
-0.01}

Step 4

Average = 0.12

Bootstrap-Shift Test
procedure

- For the purpose of this example, let's assume N = 10.

T = {0.10,
0.04,
0.21,
0.20,
0.13,
0.09,
0.22,
0.07,
0.03,
0.11}

We modify it to be a two-tailed test.

Step 3

Average = 0.12

Step 4

T'= {-0.02,
-0.08,
0.09,
0.08,
0.01,
-0.03,
0.10,
-0.05,
-0.09,
-0.01}

Significance Tests

summary

- Significance tests help us determine whether the outcome of an experiment signals a “true” trend
- The null hypothesis is that the observed outcome is due to random chance (sample bias, error, etc.)
- There are many types of tests
- Parametric tests: assume a particular distribution for the test statistic under the null hypothesis
- Non-parametric tests: make no assumptions about the test statistic distribution under the null hypothesis
- The randomization and bootstrap-shift tests make no assumptions, are robust, and easy to understand

Any Questions?

Exploratory Analysis: Clustering

Next Class