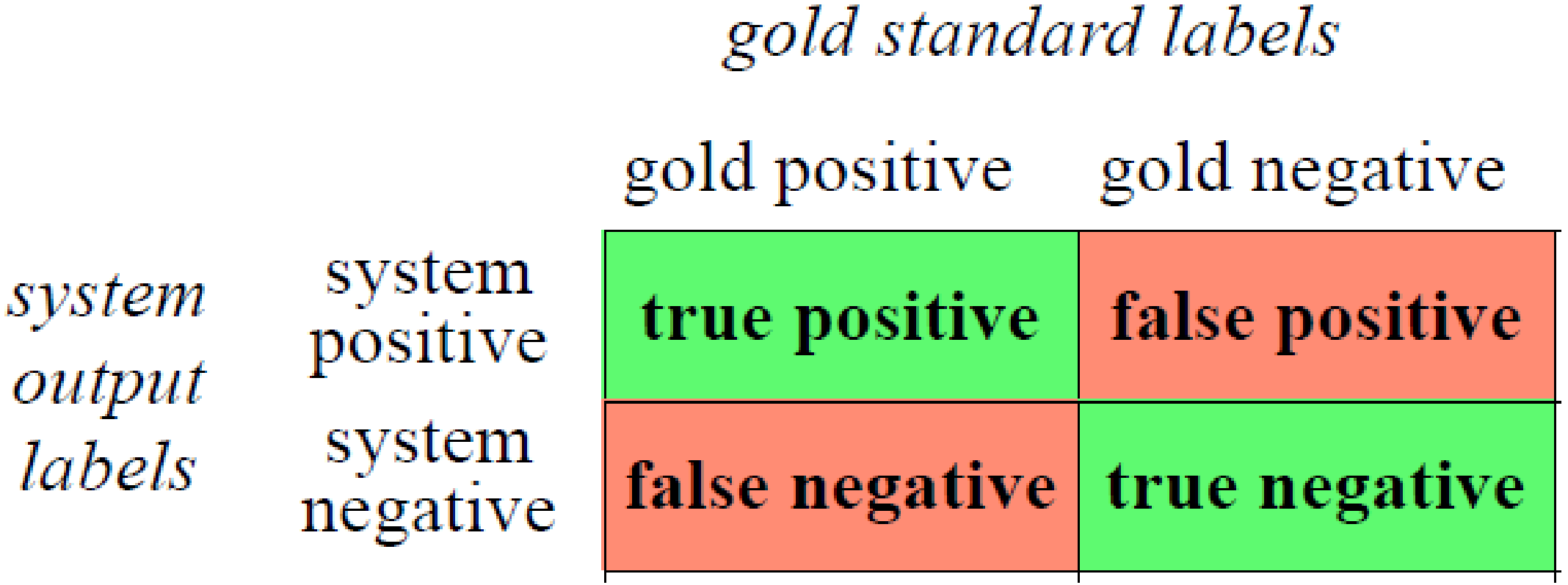
Text Classification Evaluation

Lecture 7

Text Classification and Naïve Bayes

Precision, Recall, and the F measure

The 2-by-2 contingency table



Accuracy

- The **accuracy** of a classifier: the fraction of these classifications that are correct

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + Flase\ Positive + Flase\ Negative}$$

Why not just use accuracy?

- How to build a 99.9999% accurate classifier on a low budget….

Why not just use accuracy?

- How to build a 99.9999% accurate classifier on a low budget….

	Gold Positive	Gold Negative
System Positive	0	0
System Negative	1	99

Why not just use accuracy?

- How to build a 99.9999% accurate classifier on a low budget….

	Gold Positive	Gold Negative
System Positive	1	1
System Negative	0	98

Precision and Recall

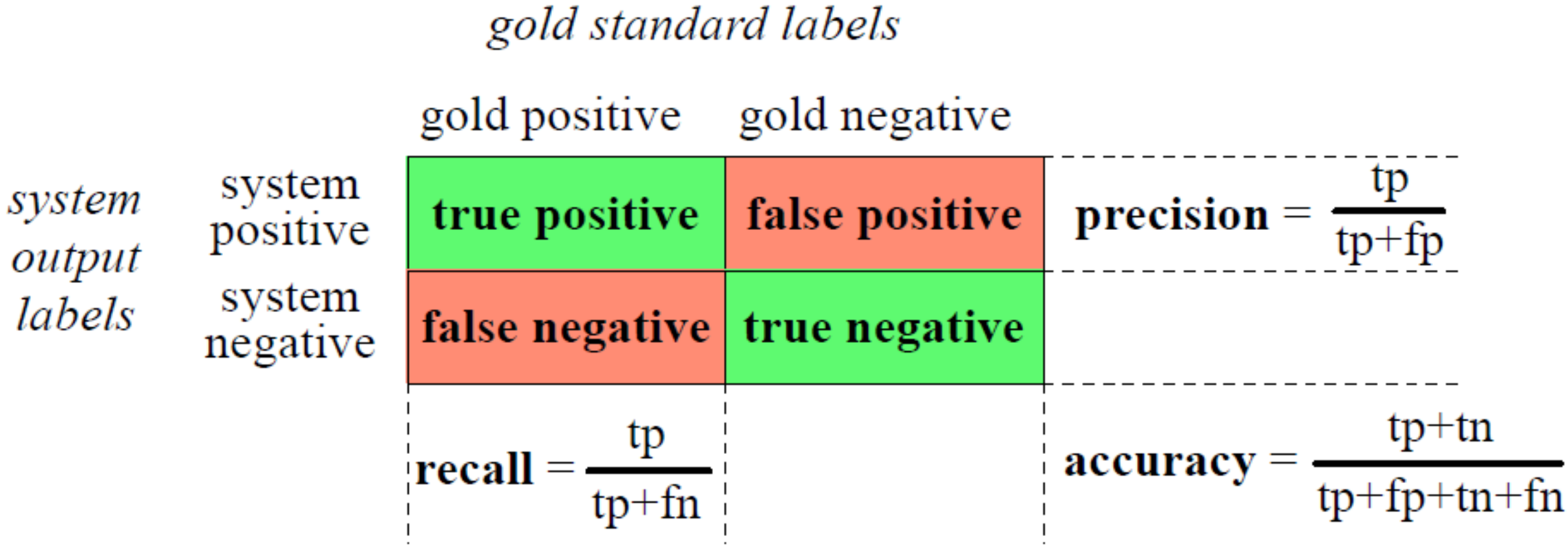
- **Precision**
- Precision measures the percentage of the items that the system labeled as positive that are in fact positive

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- **Recall**
- Recall measures the percentage of items actually present in the input that were correctly identified by the system.

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Precision and Recall



- What will be precision and recall of the low budget 99.99% accurate classifier ?

Classifier 1

	Gold Positive	Gold Negative
System Positive	25	45
System Negative	10	20

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Classifier 2

	Gold Positive	Gold Negative
System Positive	10	30
System Negative	25	35

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Classifier 1

	Gold Positive	Gold Negative
System Positive	25	45
System Negative	10	20

Precision = 25 / (25+45)
= 0.36

Recall = 25 / (25+10)
= 0.71

Accuracy = 25 + 20 / (25+10+45+20)
= 0.45

Classifier 2

	Gold Positive	Gold Negative
System Positive	10	30
System Negative	25	35

Precision = 10 / (10+30)
= 0.25

Recall = 10 / (25+10)
= 0.29

Accuracy = 10+35 / (10+35+30+25)
= 0.45

Precision and Recall

- In some applications recall is more important
- In some applications precision is more important
- In some applications both recall and precision are more important

A Combined F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

A Combined F-Measure

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$):
  - $F = 2PR/(P+R)$

More Than Two Classes: Sets of binary classifiers

- One-of or multinomial classification
  - Classes are mutually exclusive: each document in exactly one class
  - Document d belongs to the one class with maximum score

Evaluation:
Classic Reuters-21578 Data Set

- Most (over)used data set, 21,578 docs (each 90 types, 200 toknens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Reuters Text Categorization data set (**Reuters-21578)** document

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Confusion matrix c

- For each pair of classes <c_1,c_2> how many documents from c_1 were incorrectly assigned to c_2?
  - $c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Per class evaluation measures

Recall:

Fraction of docs in class *i* classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:

Fraction of docs assigned class *i* that are actually about class *i*:

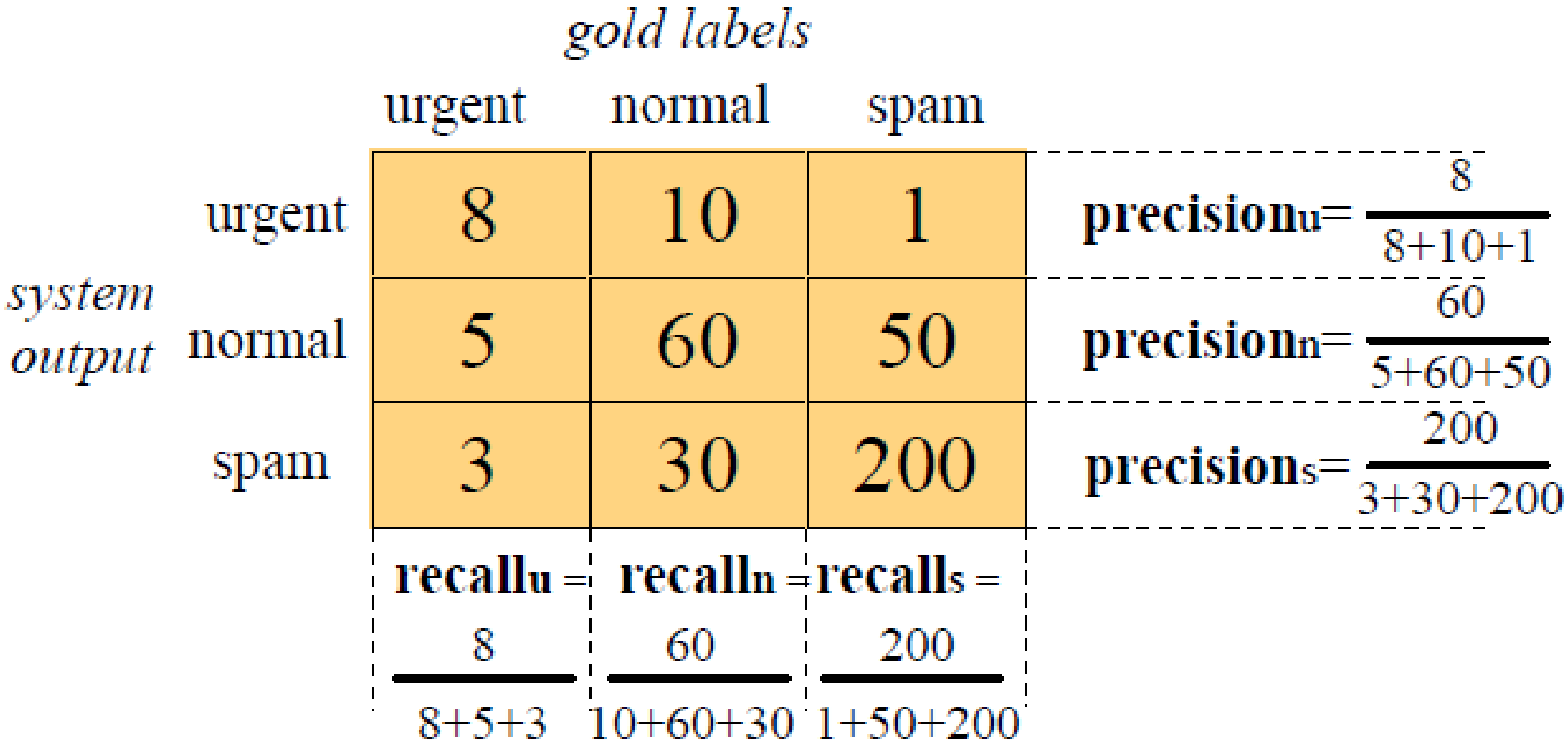
$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 - error rate)

Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Confusion matrix for Multiclass



Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging**: Compute performance for each class, then average.
- **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.

Micro- vs. Macro-Averaging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

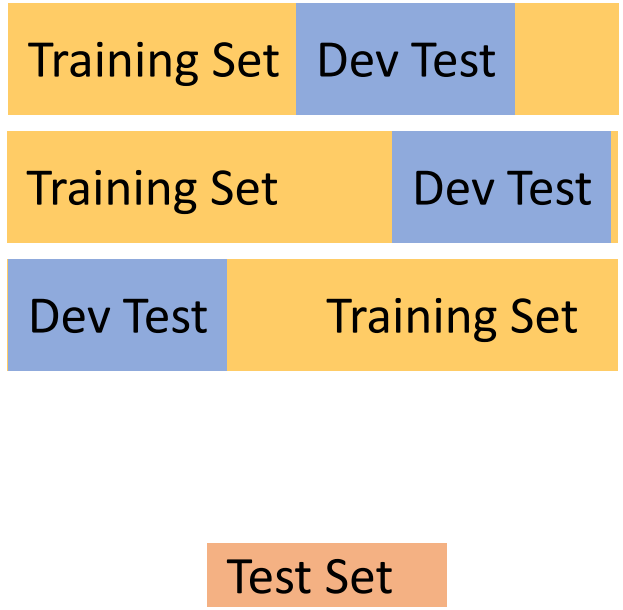
	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

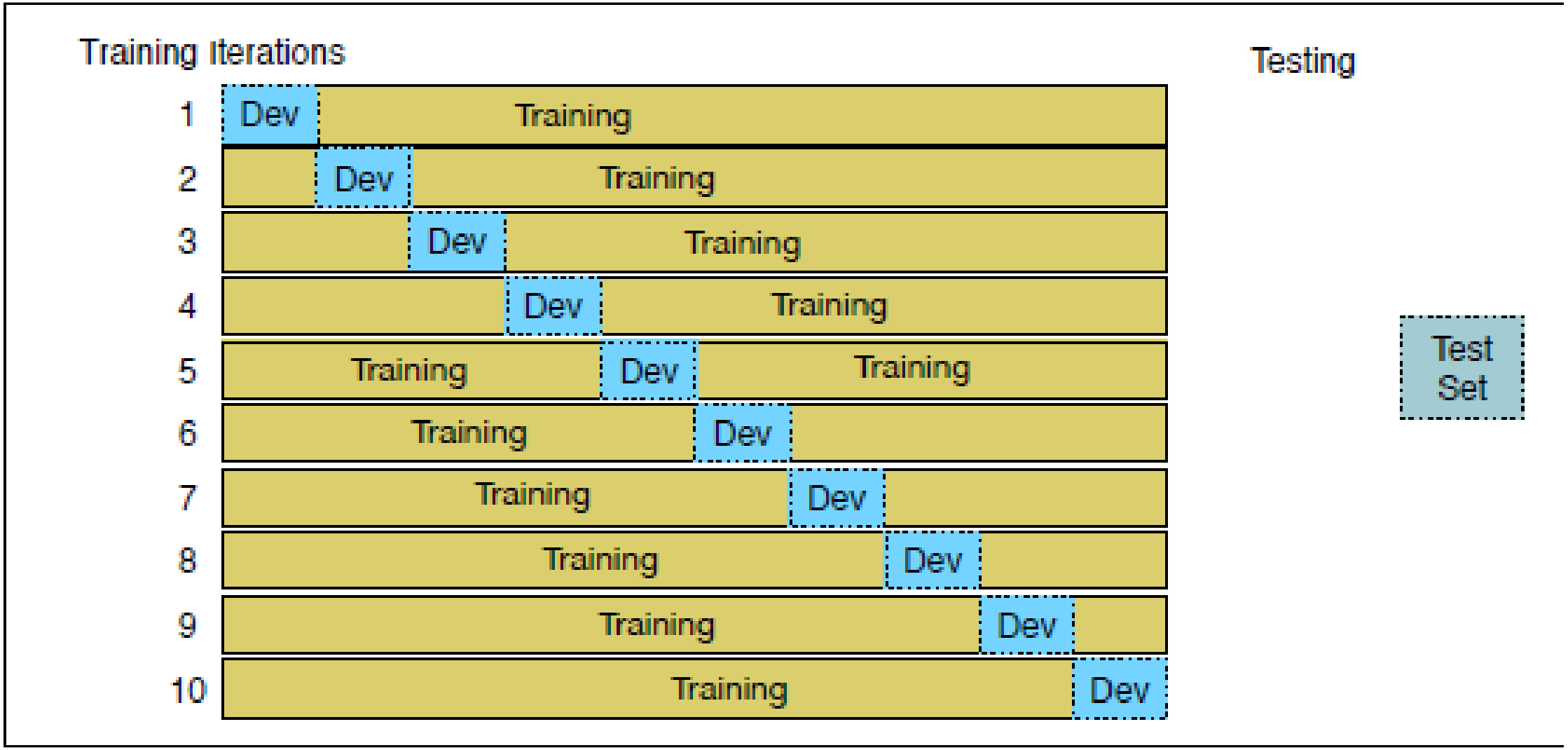
$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Development Test Sets and Cross-validation

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting ('tuning to the test set')
  - more conservative estimate of performance
- Cross-validation over multiple splits
    - Handle sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance



Development Test Sets and Cross-validation



Text Classification and Naïve Bayes

Text Classification: Practical Issues

The Real World

- Gee, I'm building a text classifier for real, now!
- What should I do?

No training data?
Manually written rules

If (wheat or grain) and not (whole or bread) then

Categorize as grain

- Need careful crafting
  - Human tuning on development data
  - Time-consuming: 2 days per class

Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a “high-bias” algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …

A reasonable amount of data?

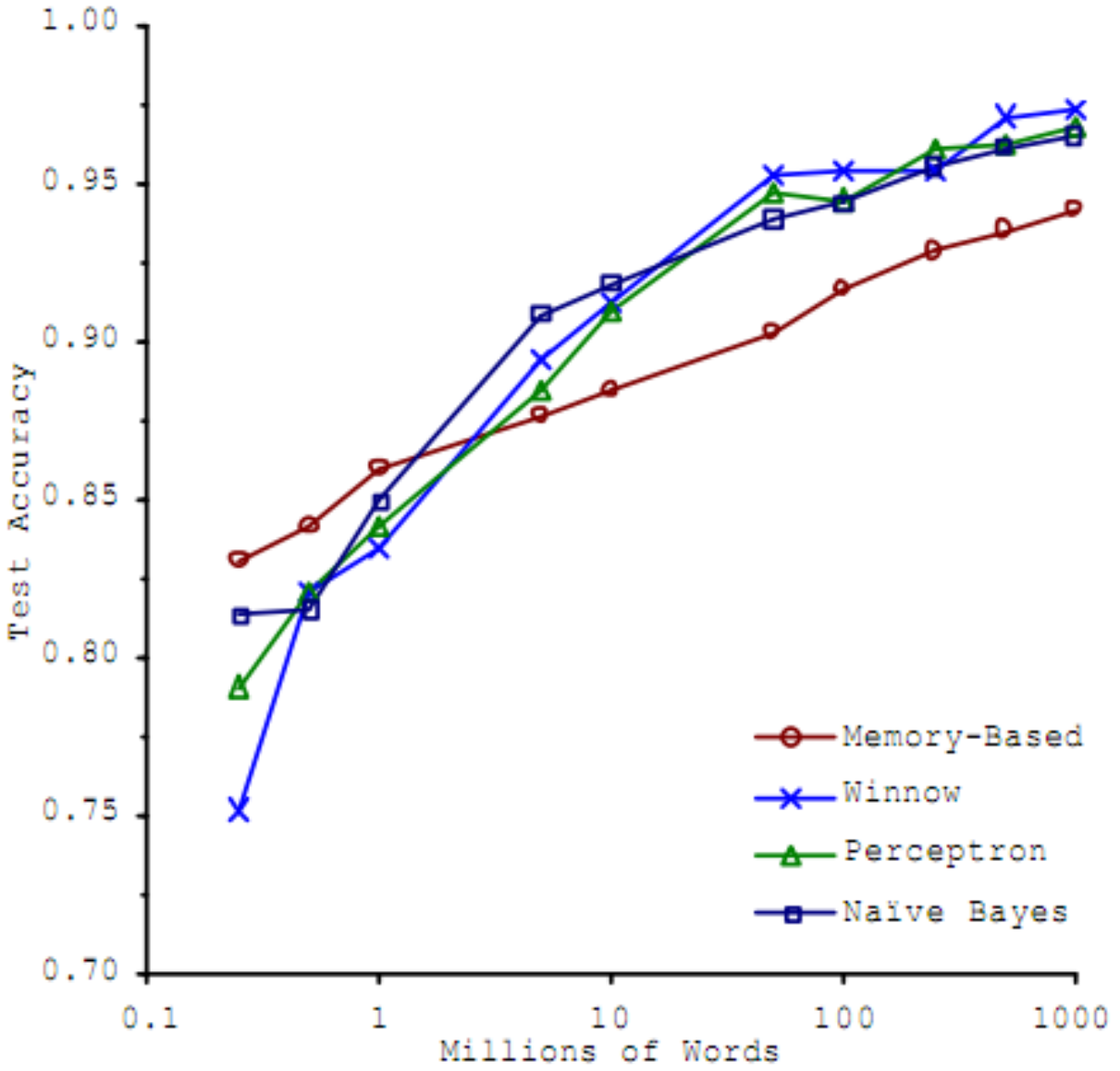
- Perfect for all the clever classifiers
  - SVM
  - Regularized Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
  - Management likes quick fixes

A huge amount of data?

- Can achieve high accuracy!
- At a cost:
  - SVMs (train time) or kNN (test time) can be too slow
  - Regularized logistic regression can be somewhat better
- So Naïve Bayes can come back into its own again!

Accuracy as a function of data size

- With enough data
  - Classifier may not matter



Brill and Banko on spelling correction

Real-world systems generally combine:

- Automatic classification
- Manual review of uncertain/difficult/"new" cases

How to tweak performance

- Domain-specific features and weights: *very* important in real performance
- Sometimes need to collapse terms:
  - Part numbers, chemical formulas, …
  - But stemming generally doesn't help
- Upweighting: Counting a word as if it occurred twice:
  - title words (Cohen & Singer 1996)
  - first sentence of each paragraph (Murata, 1999)
  - In sentences that contain title words (Ko *et al,* 2002)

Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Very good in domains with many equally important features

- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification

Reference

- Speech and Language Processing By **Dan Jurafsky and James H. Martin (3rd Edition)**

  Chapter 4 Naive Bayes and Sentiment Classification

  https://web.stanford.edu/~jurafsky/slp3/4.pdf