Annotations, Evaluation, and Generalization

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

- Annotations
  - Chance-adjusted calculation
  - Practical ethics example in ChatGPT
- Evaluation
  - Held-out data and overfitting
  - Classification metrics
  - Statistical testing (J&M 4.9) - hold off until later in course

Cohen's Kappa for IAA

- If some classes predominate, raw agreement rate may be misleading
- Idea: normalize accuracy (agreement) rate such that answering randomly = 0.
  - From psychology / psychometrics / content analysis
- **Chance-adjusted agreement:**

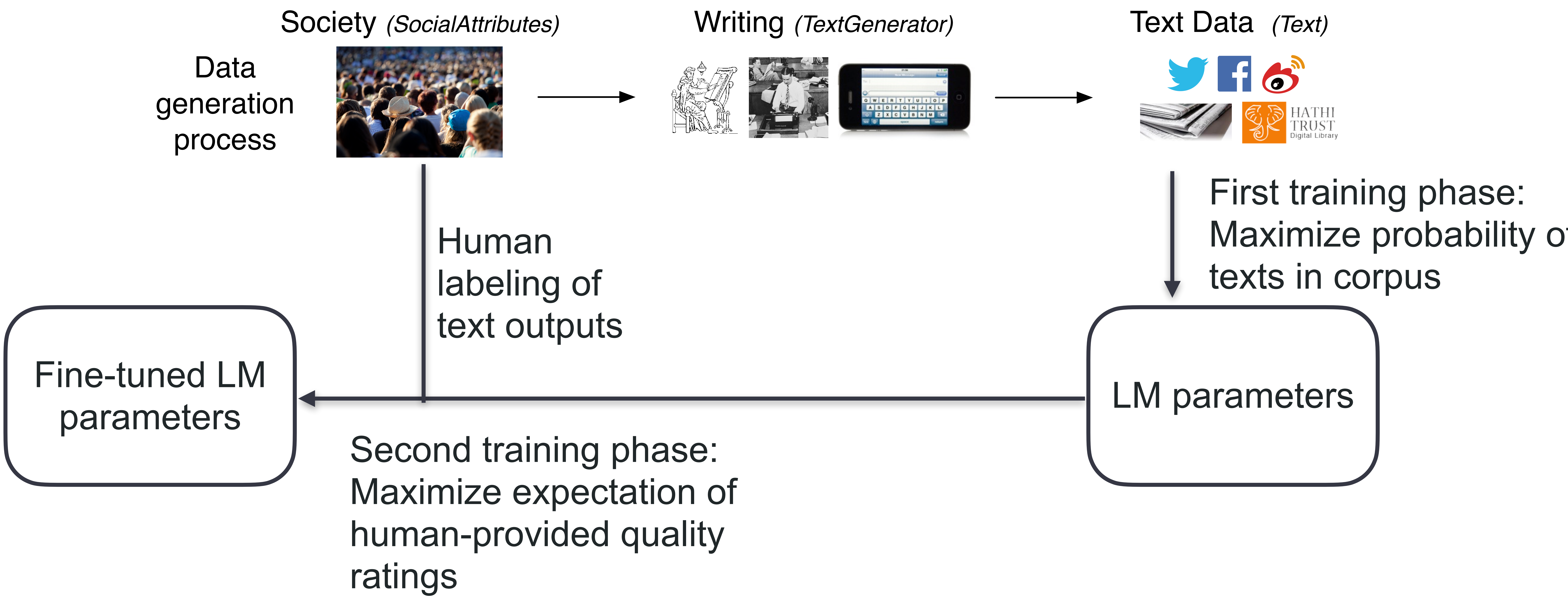
p_o: **o**bserved agreement rate

p_e: **e**xpected (by chance) rate

Other chanced-adjusted metrics: Fleiss, Krippendorff... see reading

When is annotating ethical?

Human labeling is key to ChatGPT



[Ouyang et al., 2022, Taori et al. 2023]

Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisifies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

[Ouyang et al., 2022]

'That Was Torture;' OpenAI Reportedly Relied on Low-Paid Kenyan Laborers to Sift Through Horrific Content to Make ChatGPT Palatable

The laborers reportedly looked through graphic accounts of child sexual abuse, murder, torture, suicide, and, incest.

By **Mack DeGeurin** Published January 18, 2023 | Comments (6) | Alerts

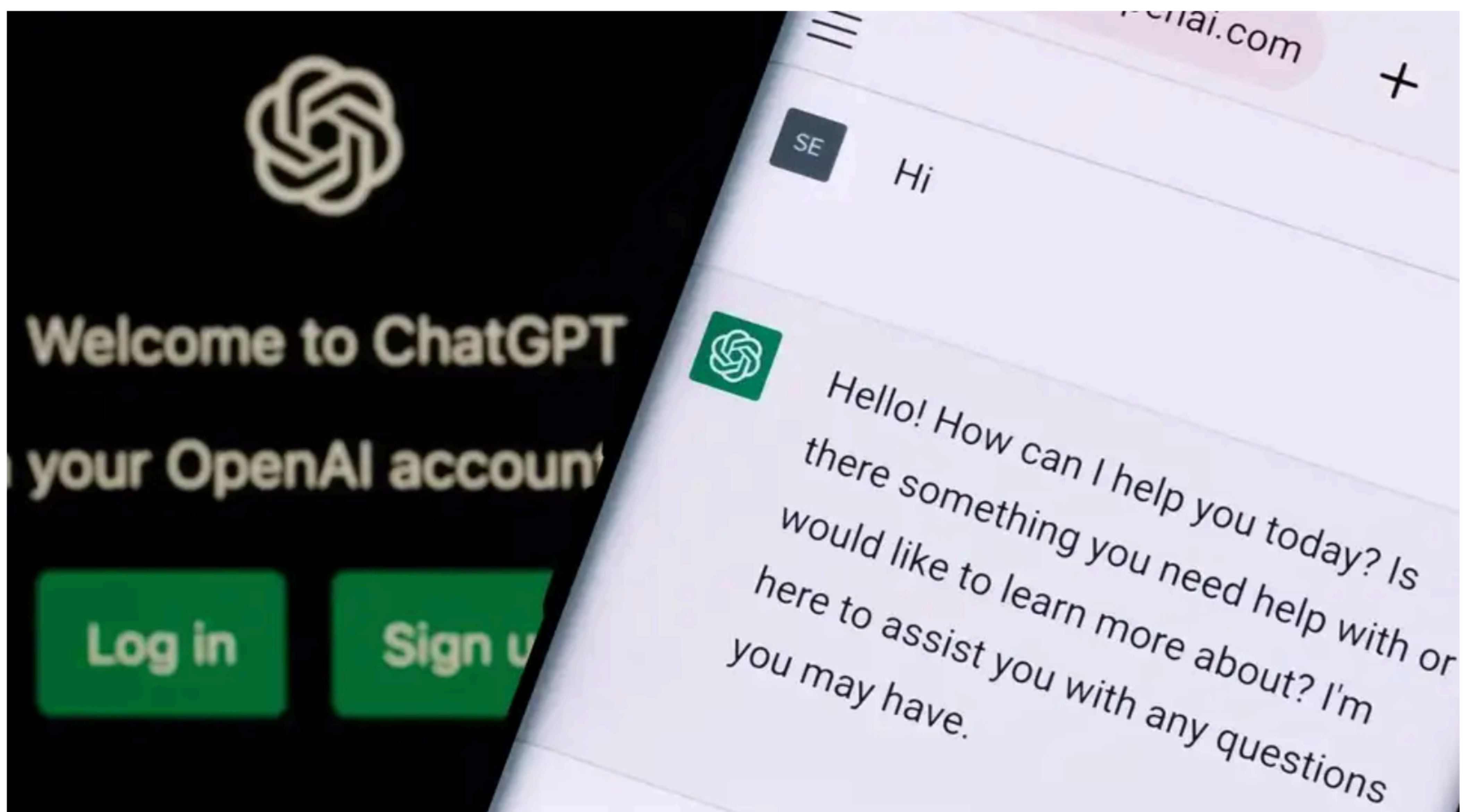
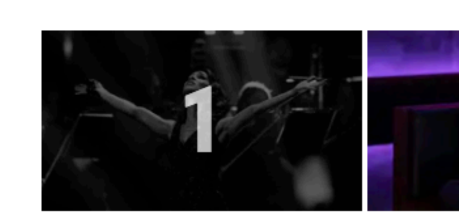


Image: Ascannio (Shutterstock)

You m



Held-out data for evaluation

- How well will my classifier work in the future?
  - Can we look at classifier accuracy on training data?

Held-out data for evaluation

- Need to diagnose how much your model is **overfitting** the training set
- Data splits are key. Some ways to split:
    - Training set -vs- test set
    - Training set -vs- "validation"/"development" set -vs- test set
    - Cross-validation (within training set) -vs- test set

Cross-validation

- Cross-validation (within training set) -vs- test set

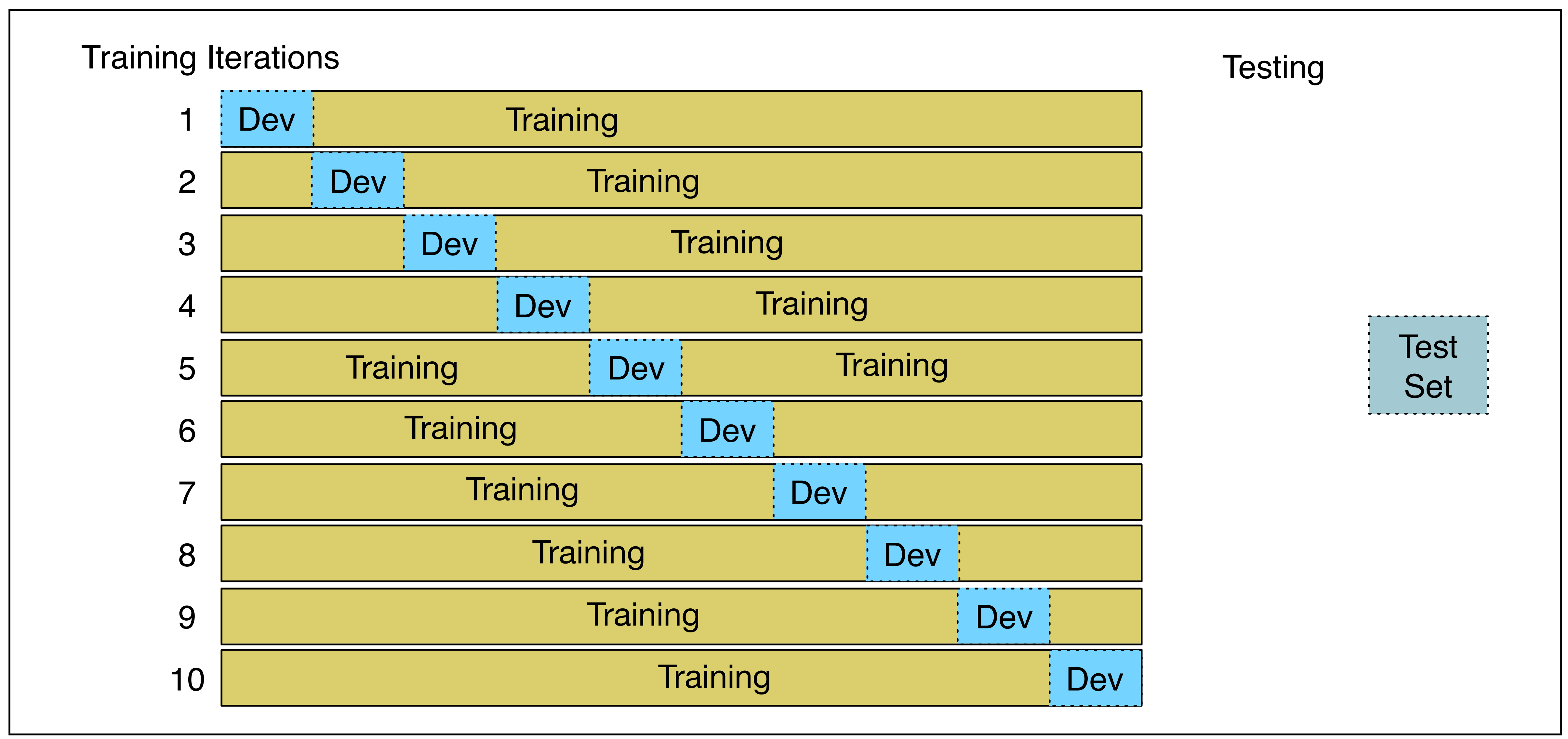


Figure 4.7 10-fold cross-validation

Regularization in Naive Bayes

Regularization in logistic regression

- If "dog" only occurs for class **k**, what weight will it get?
- Consider MLE training:

- Solution: ***regularized*** training for logistic regression

Regularization tradeoffs

- No regularization <--------------> Very strong regularization

Do I have enough labels?

- For training, hundreds to thousands of annotations may be needed for reasonable performance
    - Current work: how to usefully make NLP models with <10 or <100 training examples. "Few-shot learning"
- Exact amounts are difficult to know in advance. Can do a **learning curve** to estimate if more annotations will be useful.

Evaluation metrics

- Accuracy =

		gold standard labels		
		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** = $\frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall = $\frac{\text{tp}}{\text{tp+fn}}$		**accuracy** = $\frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Figure 4.4 A confusion matrix for visualizing how well a binary classification system performs against gold standard labels.

- But do we care about false positives and negatives equally?
- What about rare classes?

Precision, recall, F1

gold standard labels

system output labels		gold positive	gold negative	
	system positive	**true positive**	**false positive**	**precision** = $\frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall = $\frac{\text{tp}}{\text{tp+fn}}$		**accuracy** = $\frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Figure 4.4 A confusion matrix for visualizing how well a binary classification system performs against gold standard labels.

- macro vs. micro F1