

2.3 Classifier Evaluation

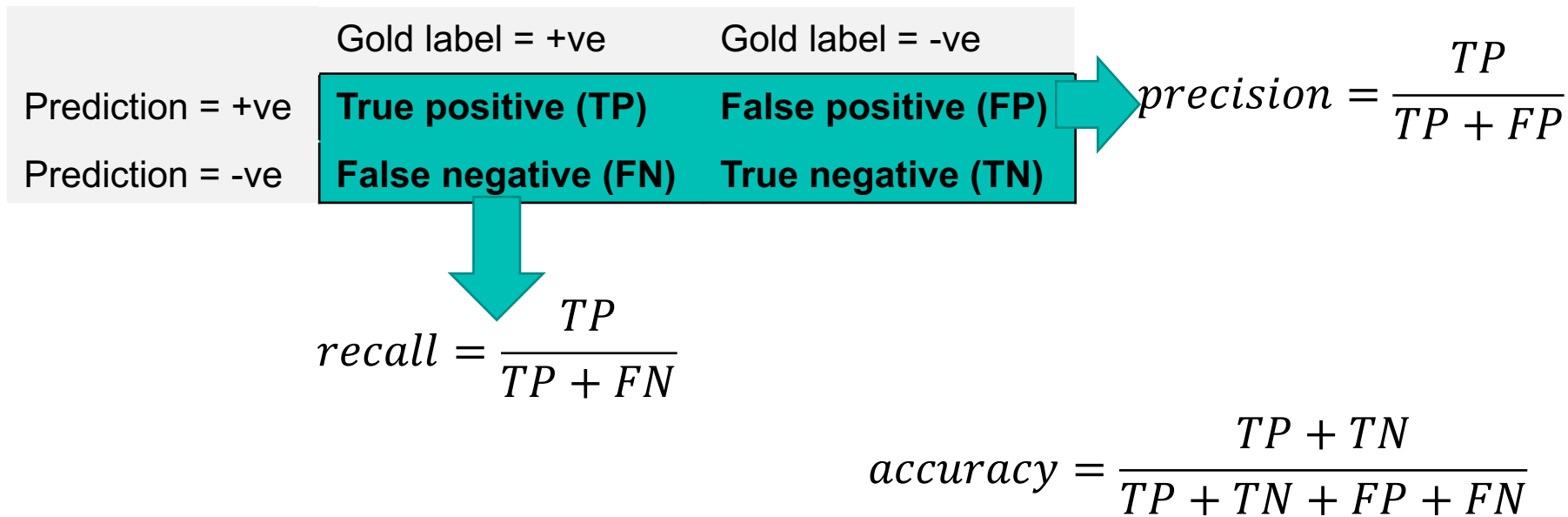
Edwin Simpson

Department of Computer Science,
University of Bristol, UK.

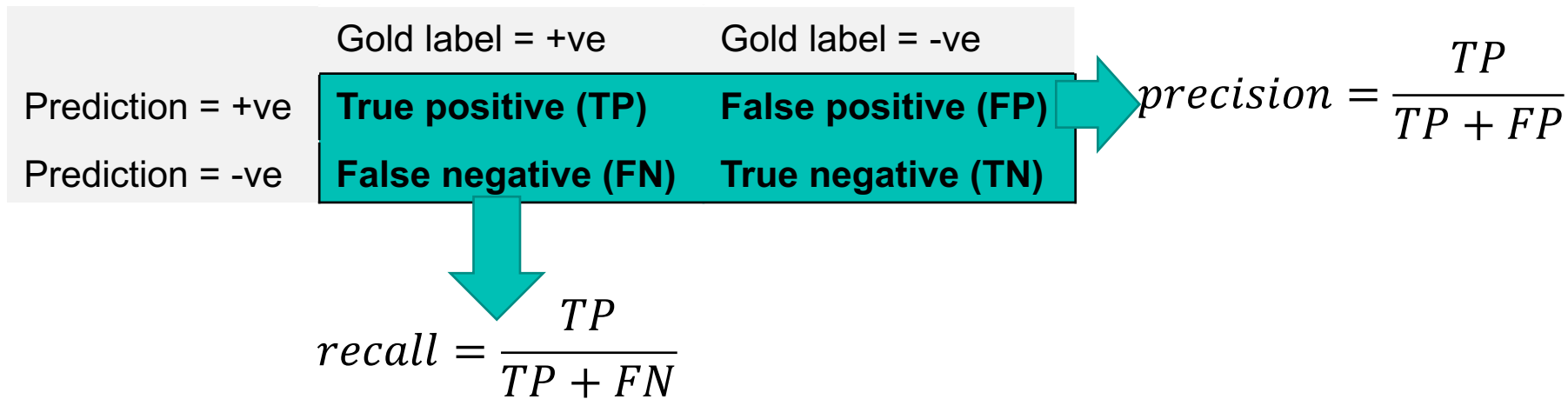
Classifier Evaluation

- Use **gold labels** to compute performance metrics:
 - Provided by expert annotators;
 - Or crowdsourced data that is then curated by a developer.
- **Training set:** gold labels are used to train the classifier;
- **Development set** (validation or dev set): assess performance to tune classifier parameters and design;
- **Test set:** kept blind during development then used to compare different systems.

Classifier Metrics





Classifier Metrics



$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Classifier Metrics

| | Gold label = +ve | Gold label = -ve |
|------------------|----------------------------|----------------------------|
| Prediction = +ve | True positive (TP) | False positive (FP) |
| Prediction = -ve | False negative (FN) | True negative (TN) |

| | | | |
|---|-------------------------------|---|------------------------------------|
|  | $recall = \frac{TP}{TP + FN}$ |  | $specificity = \frac{TN}{TN + FP}$ |
|---|-------------------------------|---|------------------------------------|

Multiclass Classifier Metrics

- Micro F1:
 - Sum up counts of TP, FP and FN across all classes;
 - More frequent classes will dominate.
- Care is needed when using aggregated metrics as they can hide information about the nature of errors.

| | Gold label = A | Gold label = B | Gold label = C |
|----------------|----------------|----------------|----------------|
| Prediction = A | 234 | 1 | 8 |
| Prediction = B | 12 | 139 | 90 |
| Prediction=C | 13 | 7 | 300 |

TP = 234+139+300

Multiclass Classifier Metrics

- Macro F1:
 - Mean of F1 scores across classes;
 - Gives equal weight to under-represented classes.

| | Gold label = A | Gold label = B | Gold label = C |
|----------------|----------------|----------------|----------------|
| Prediction = A | 234 | 1 | 8 |
| Prediction = B | 12 | 139 | 90 |
| Prediction=C | 13 | 7 | 300 |

precision =
 $234/(234+1+8)$

recall =
 $234/(234+12+13)$

Summary

- Precision, recall and F1 scores are frequently used to assess classifier performance
- To evaluate a classifier, it's important to examine contingency tables or confusion matrices as well as aggregate metrics like F1