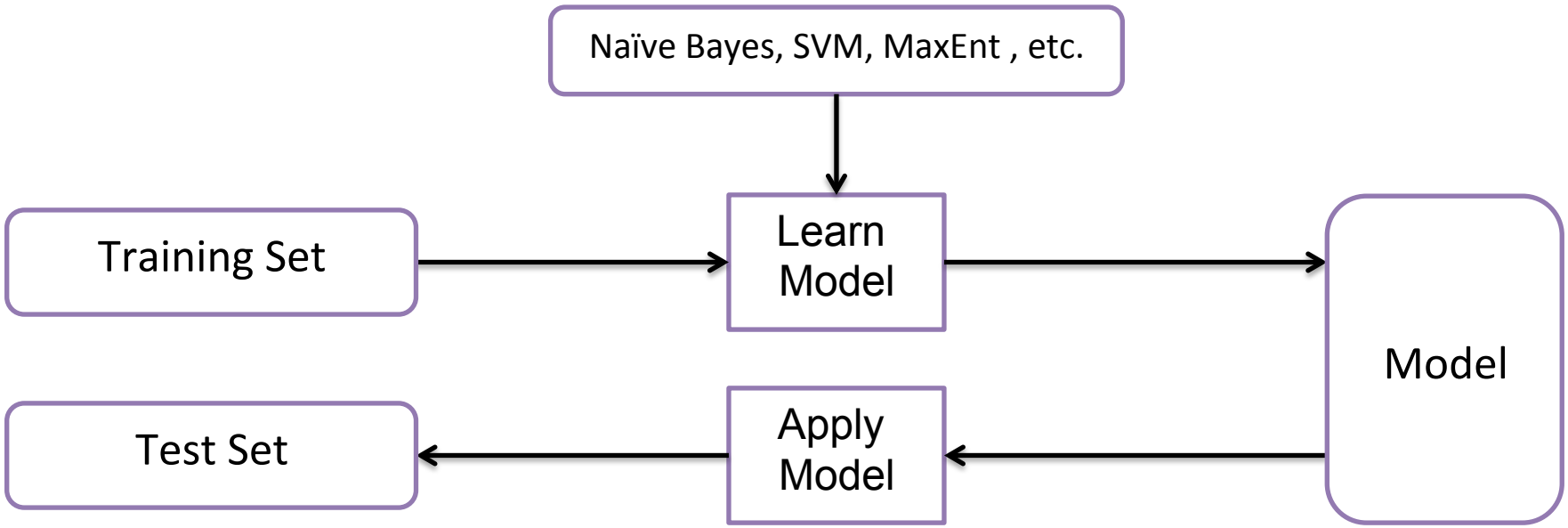
Domain Adaptation for Classification

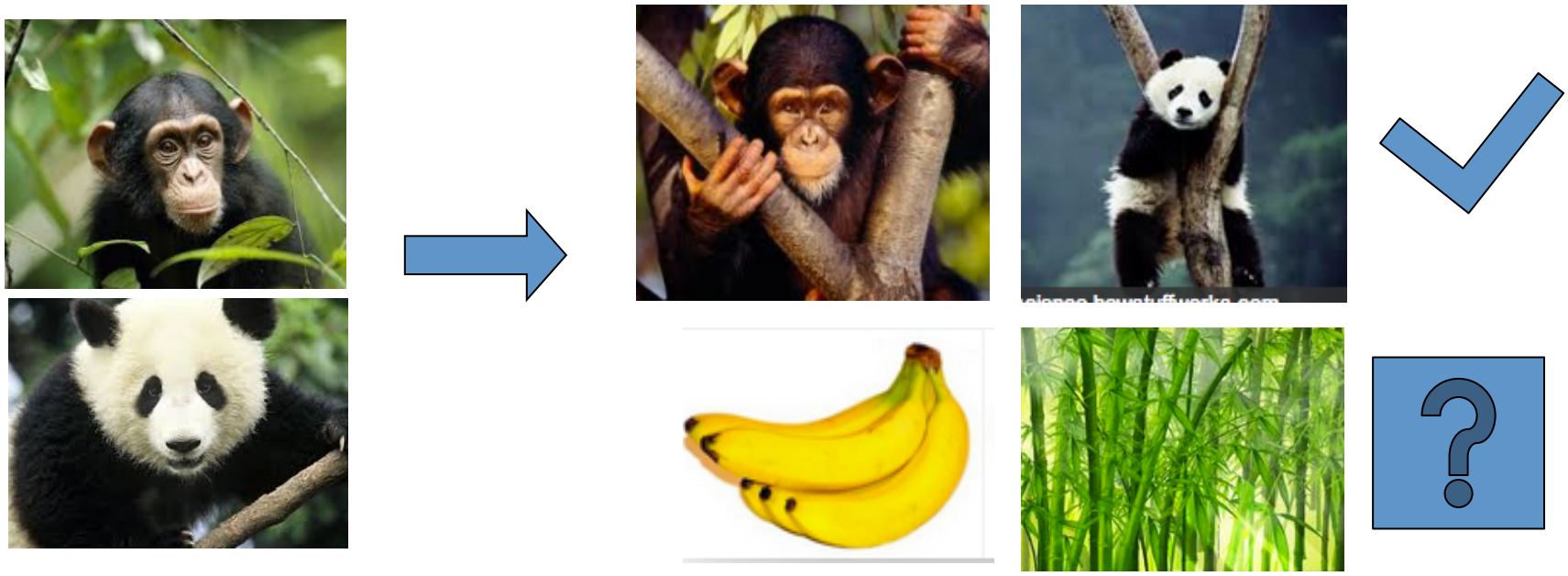
Chenghua Lin
Dept. of Computing Science
University of Aberdeen

Traditional Classification



An Important Assumption

- Training and unseen (test) data
  - in the same feature space
  - follow the same distribution



Training data

Unseen data

In the Real World …

- More often than **NOT**, the aforementioned assumptions do not hold
- May not have sufficient training data in the domain of interest (target)
- Rather, we have a lot of labelled data available in other domains (source)
- Source domain data exhibit different distributions or feature space to the target domain
- Labelling cost is expensive

Some Early Attempts

- Aue and Gamon(2005): *"Explore customizing sentiment classifiers to new domains"*
- Data
  - Movie review data (**movie**)
    - 1000 positive and 1000 negative reviews from movie databases
  - Book review data (**book**)
    - 1000 positive and 1000 negative book reviews from the web
  - Product Support Services web survey data (**pss**)
    - 2564 examples of positive feedback and 2371 examples of negative feed- back
  - Knowledge Base web survey data (**kb**)
    - 6035 examples of "bad" feedback and 6285 examples of "good" feedback.

The Cross-Domain Issues

(Aue and Gamon 2005)

- Sentiment classifier trained on one domain do not generalize well to other domains

	movie	book	kb	pss
movie	**90.45**	70.29	57.59	61.36
book	72.08	**79.42**	59.28	66.59
kb	57.1	58.62	**77.34**	81.42
pss	52.16	55.33	70.48	**83.73**

Table 1: Best results of SVM classifiers within and across domains.

kb: Knowledge base web survey data
pss: product support services web survey data

Cross-Domain Classification

- Strategies
  1. Mixing available labelled data from all source domains → trained a single classifier **(data fusion)**
  2. Similar to (1), but limiting the feature set to those observed in the target domains **(data fusion + feature engineering)**
  3. Train a single classifier for each domain using labelled data → ensemble classifiers **(classifier fusion)**
  4. NaiveBayes-EM: trained on **in-domain labelled** data (small amounts) + **in-domain unlabelled** data (large amounts)

Data fusion

Target domain	Training domains	accuracy
Movies	books, kb, pss	72.89
Books	movie, kb, pss	64.58
Kb	movie, book, pss	63.92
Pss	movie, book, kb	74.88

Data fusion + feature engineering

Target Domain	All_data	Target domain features only
movies	72.89	**59.11**
books	64.58	64.19
kb	63.92	**70.98**
pss	74.99	75.26

NaiveBayes-EM

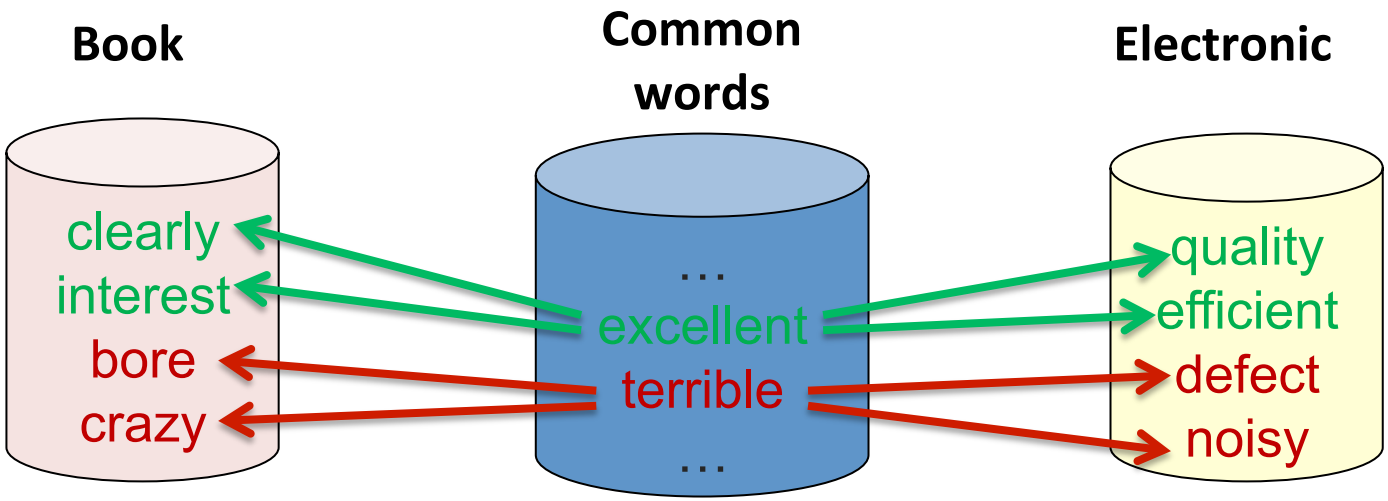
One type of semi-supervised learning

Target domain	Base-line	amount of labeled data		
		50	100	200
movies	72.89	**61.67**	**79.56**	**77.44**
books	64.58	62.48	**71.08**	**76.55**
Kb	63.92	**65.84**	**68.1**	**73.86**
Pss	74.88	**81.79**	**80.75**	**82.39**

Cross-domain Classification

- Domain adaptation
  - One type of transfer learning
  - Source and target tasks are the same, but the domains are different (e.g. feature space)
  - Aim to improve the target predictive function using the knowledge from the source domain
- Situation
  - Plentiful labelled data in the *source* domain, but very few or no labelled data in the *target* domain
  - Example: transferring knowledge from Books reviews → detecting the polarity of Kitchen reviews

Linking Features between Source and Target Domains



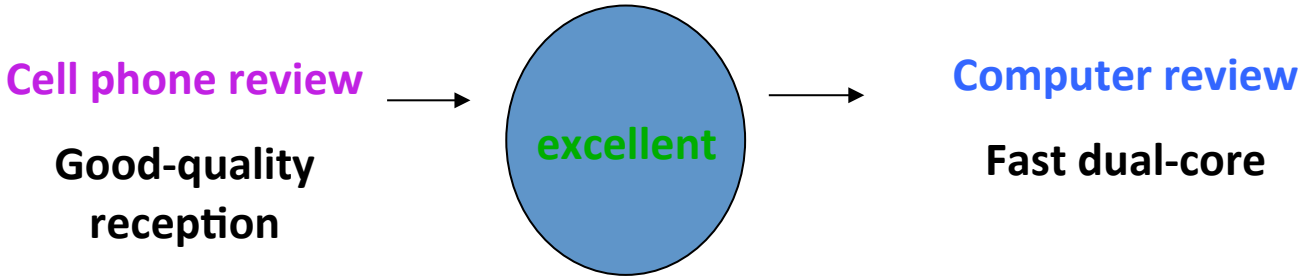
Structural Correspondence Learning
(Blitzer et al., 2007)

- Select pivot features
  - Occur frequently on both source and target domains
  - Chose based on mutual information with labelled source data
- Use pivot features to align other features
  - Mask and predict pivot features from other features
  - Train N linear predictors, one for each pivot feature
  - Each pivot predictor implicitly aligns non-pivot features from **source** & **target** domains
- Dimensionality reduction using singular value decomposition (SVD)
- Linear classifier training

Structural Correspondence Learning

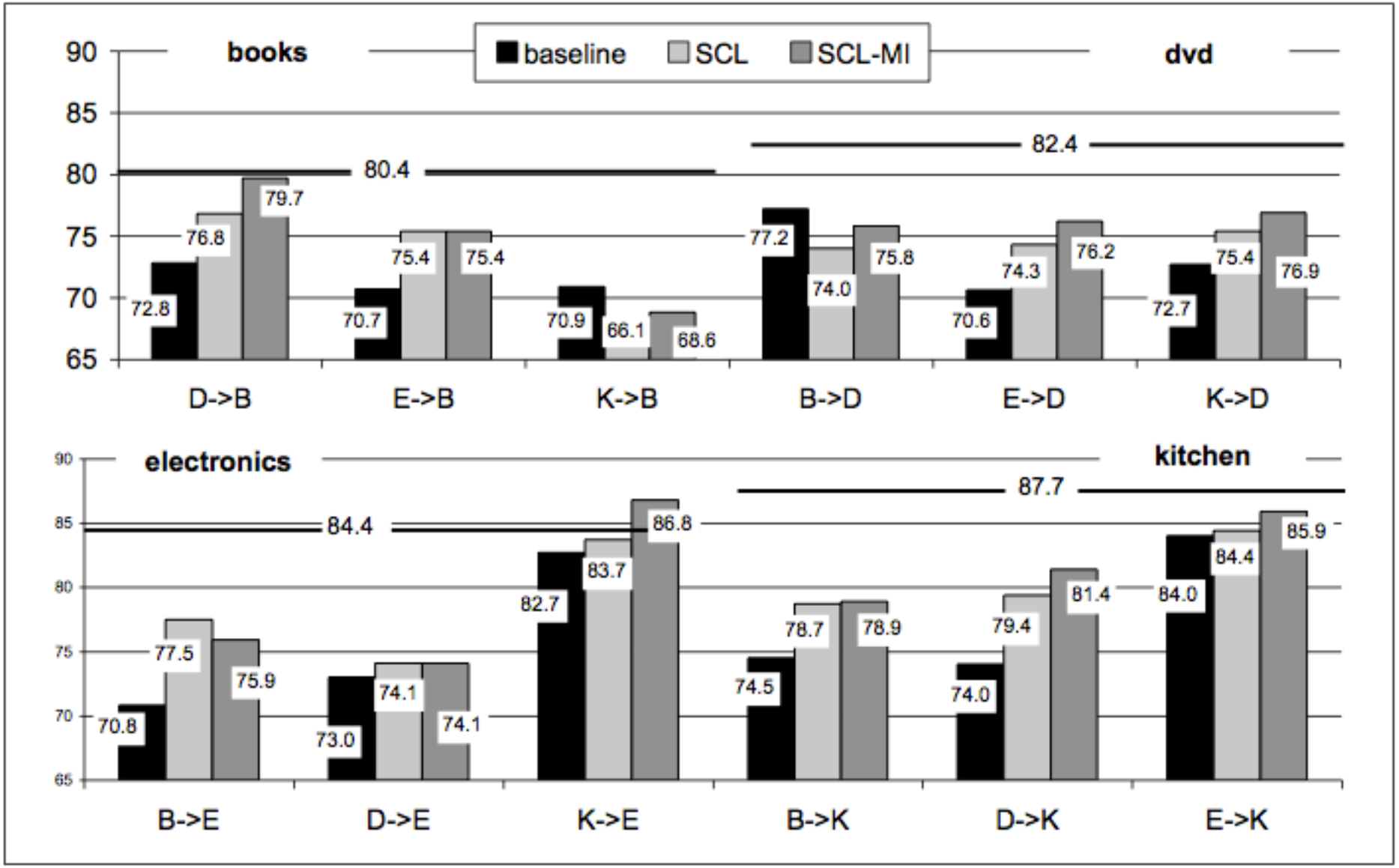
- Example pivot features

SCL-MI, not SCL
a_must a_wonderful loved_it
weak don't_waste awful
highly_recommended and_easy



Datasets

- Multi-domain sentiment (MDS) dataset
  - Products reviews for books, DVDs, electronics, and kitchen appliance
  - 1000 positive and 1000 negative movie reviews for each domain


books
baseline
SCL
SCL-MI
dvd
90
85
80
75
70
65
80.4
82.4
72.8
76.8
79.7
70.7
75.4
75.4
70.9
66.1
68.6
77.2
74.0
75.8
70.6
74.3
76.2
72.7
75.4
76.9
D->B
E->B
K->B
B->D
E->D
K->D
electronics
kitchen
84.4
87.7
70.8
77.5
75.9
73.0
74.1
74.1
82.7
83.7
86.8
74.5
78.7
78.9
74.0
79.4
81.4
84.0
84.4
85.9
B->E
D->E
K->E
B->K
D->K
E->K

Summary

- Supervised classifier is domain dependent
- Target domain refers to the domain of our interest
- We may not always have enough training data in the target domain, but we may have a lot of labelled data available in other domains (source)
- Source domain data exhibit different distributions or feature space to the target domain
- Improve the target predictive function using the knowledge from the source domain