

Frustratingly Easy Domain Adaptation

Hal Daumé III

School of Computing
University of Utah

me@hal3.name



Problem

- “was trained on”
- My tagger ~~expects~~ data like:

Source Domain

But the unknown culprits, who had access to some of the company's computers for an undetermined period...

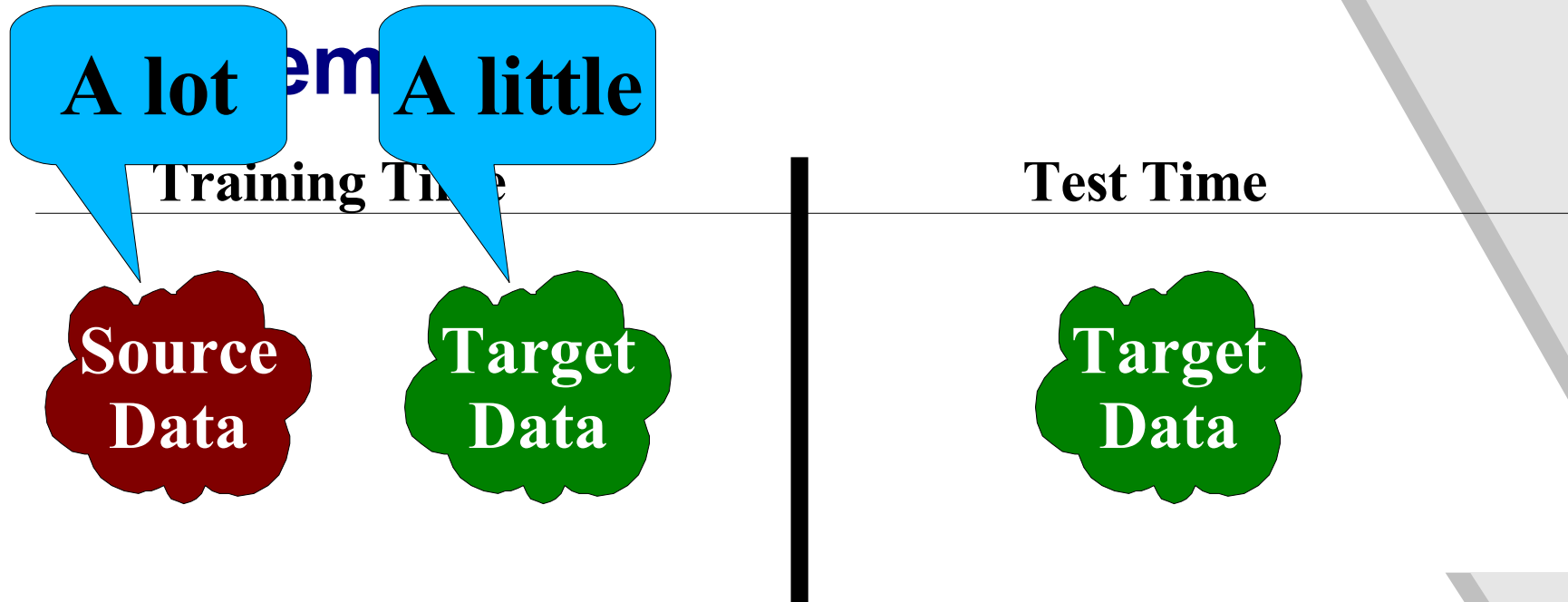
Target Domain

- ...but then I give it data like:

you know it is it's pretty much general practice
now you know

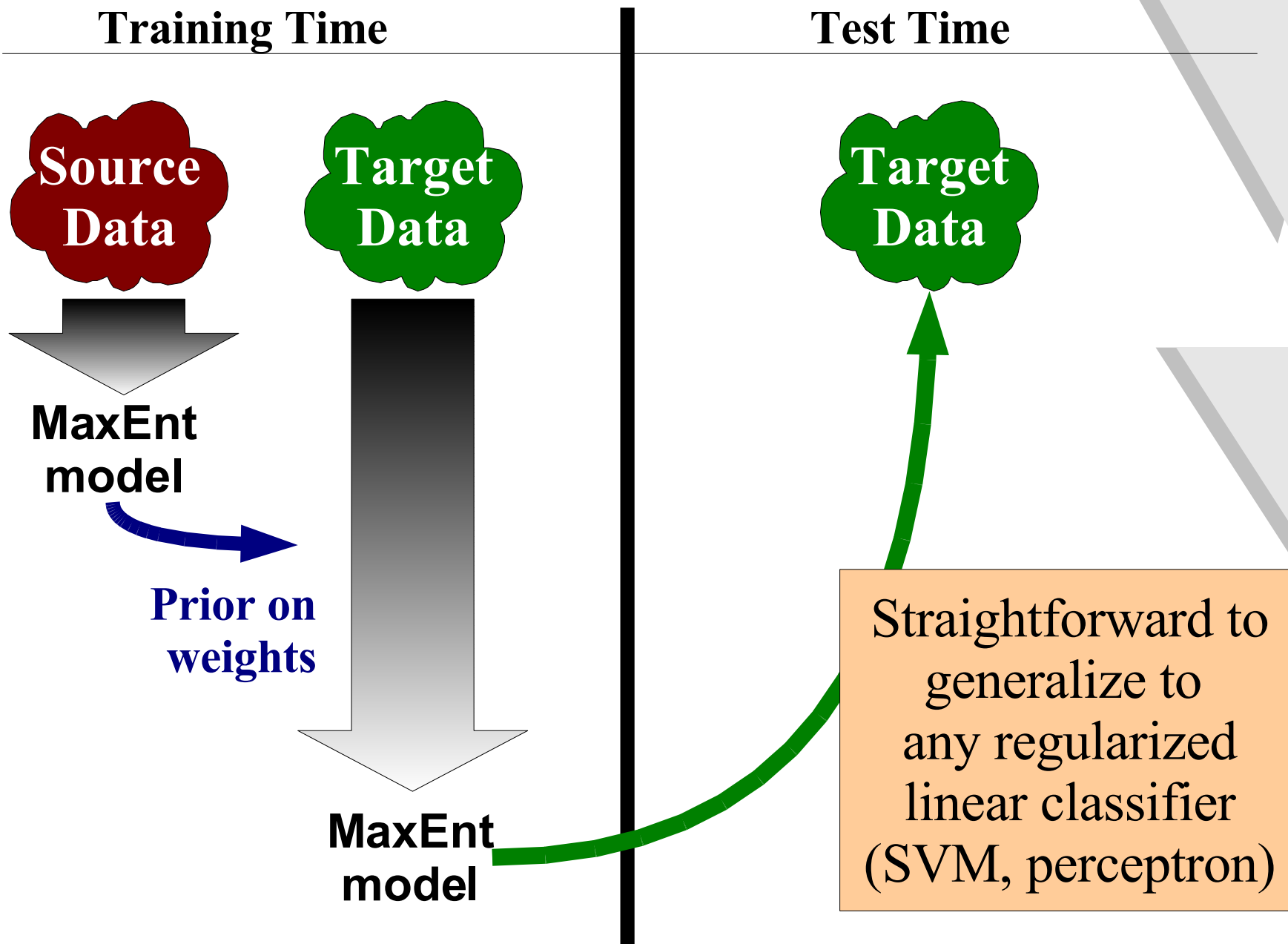
Solutions...

- “*LDC Solution*” -- Annotate more data!
 - **Pros:** will give us good models
 - **Cons:** Too expensive, wastes old effort, no fun
- “*NLP Solution*” -- Just use our news model on non-news
 - **Pros:** Easy
 - **Cons:** Performs poorly, no fun
- “*ML Junkie Solution*” -- Build new learning algorithms
 - **Pros:** Often works well, fun
 - **Cons:** Often hard to implement, computationally expensive
- “*Our Solution*” – Preprocess the data
 - **Pros:** Works well, easy to implement, compu'lly cheap
 - **Cons:** Theoretical justification

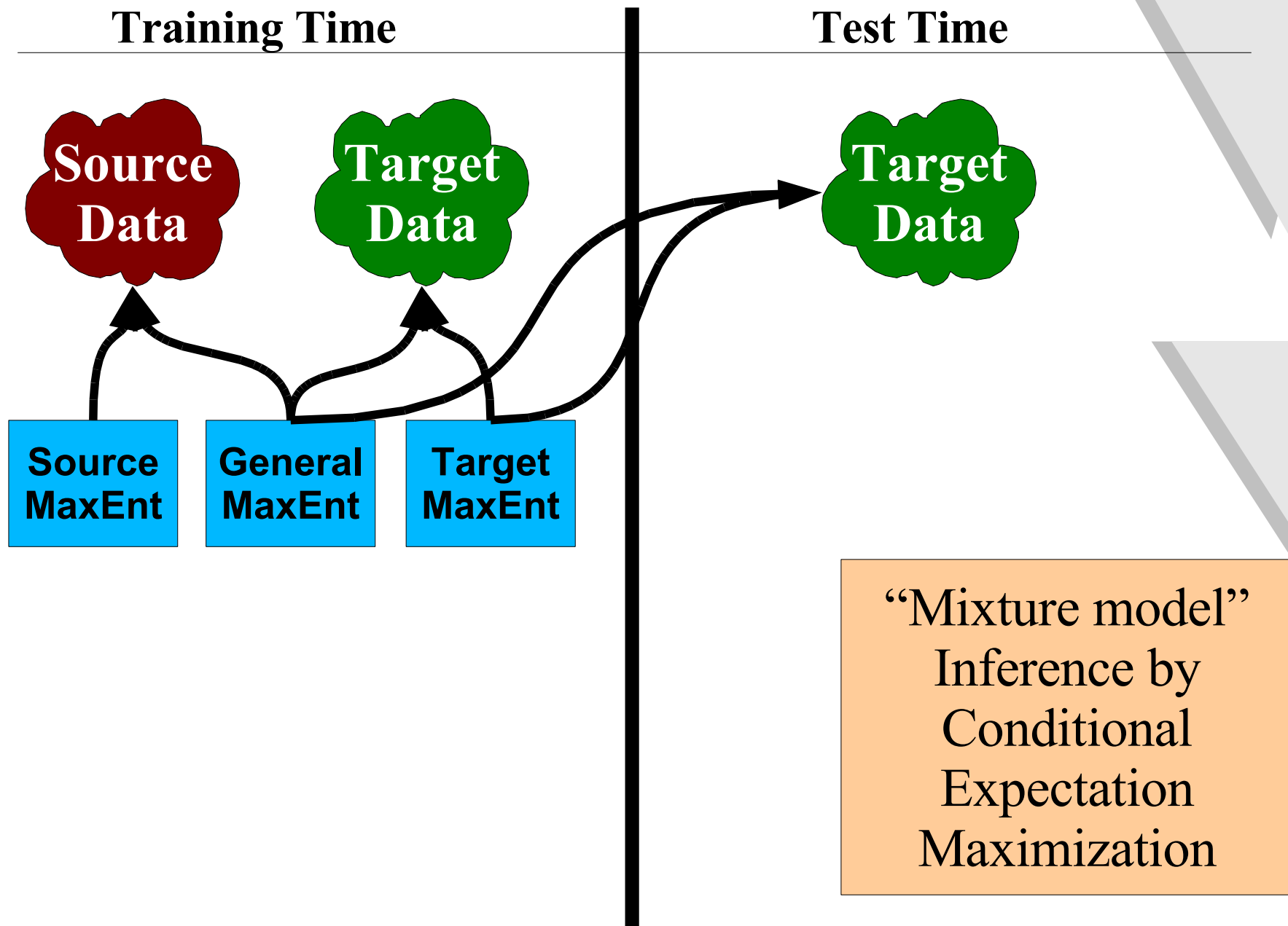


We assume all data is *labeled*.
If you only have unlabeled target
data, talk to John Blitzer

Prior Work – Chelba and Acero



Prior Work – Daumé III and Marcu



“MONITOR” versus “THE”

News domain:

“MONITOR” is a **verb**
“THE” is a **determiner**

Technical domain:

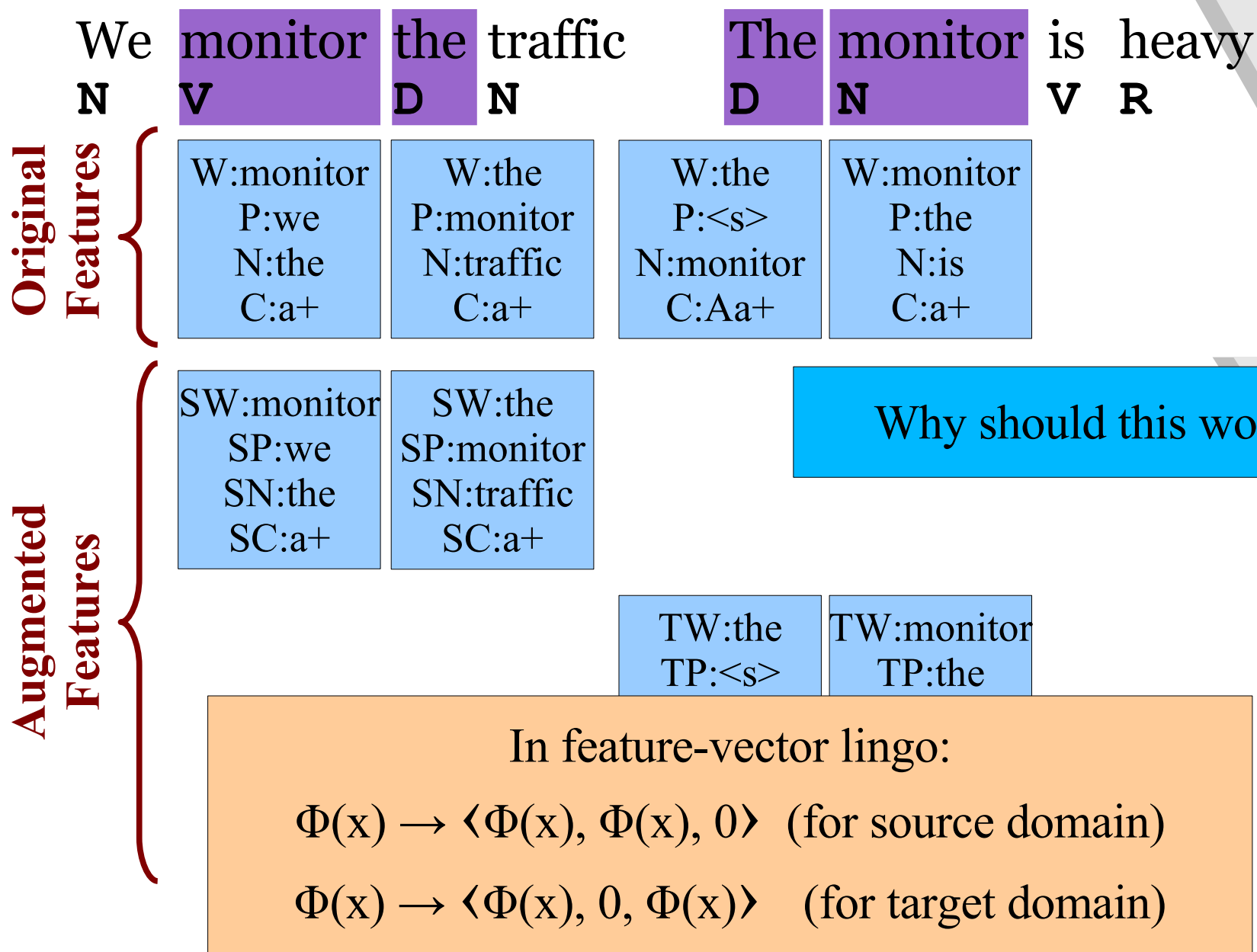
“MONITOR” is a **noun**
“THE” is a **determiner**

Key Idea:

Share some features (“the”)
Don't share others (“monitor”)

(and let the *learner* decide which are which)

Feature Augmentation



A Kernel Perspective

In feature-vector lingo:

$$\Phi(\mathbf{x}) \rightarrow \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}), 0 \rangle \quad (\text{for source domain})$$

$$\Phi(\mathbf{x}) \rightarrow \langle \Phi(\mathbf{x}), 0, \Phi(\mathbf{x}) \rangle \quad (\text{for target domain})$$

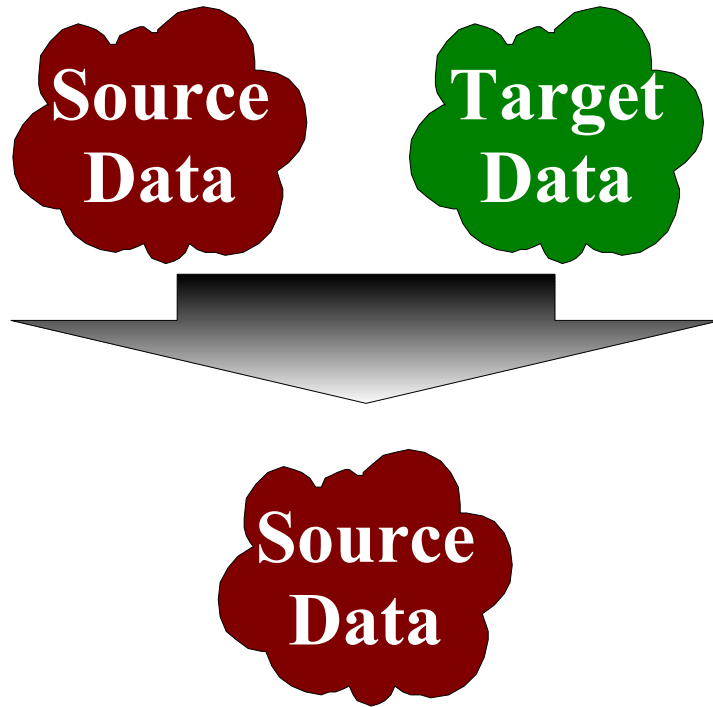
$$K^{\text{aug}}(\mathbf{x}, \mathbf{z}) = \begin{cases} 2K(\mathbf{x}, \mathbf{z}) & \text{if } \mathbf{x}, \mathbf{z} \text{ from same domain} \\ K(\mathbf{x}, \mathbf{z}) & \text{otherwise} \end{cases}$$

Experimental Setup

- Lots of data sets:
 - ACE: Named entity recognition (6 domains, ~20% target size)
 - CoNLL: Named entity recognition (2 domains, 10% target size)
 - PubMed: POS tagging (2 domains, 12% target size)
 - CNN: recapitalization (2 domains, 2% target size)
 - Treebank: Chunking (3 or 10 domains, 2% to 32% target size)
 - **Always:** 75% train, 12.5% dev, 12.5% test
- Lots of baselines...
- Evaluation metric: Hamming loss (+McNemar)
- Sequence labeling using SEARN

Obvious Approach 1: SrcOnly

Training Time

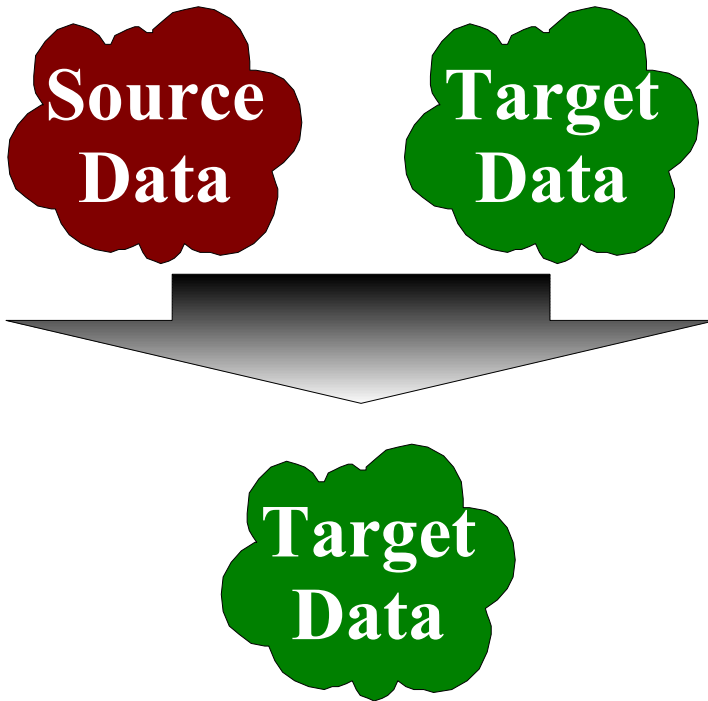


Test Time



Obvious Approach 2: TgtOnly

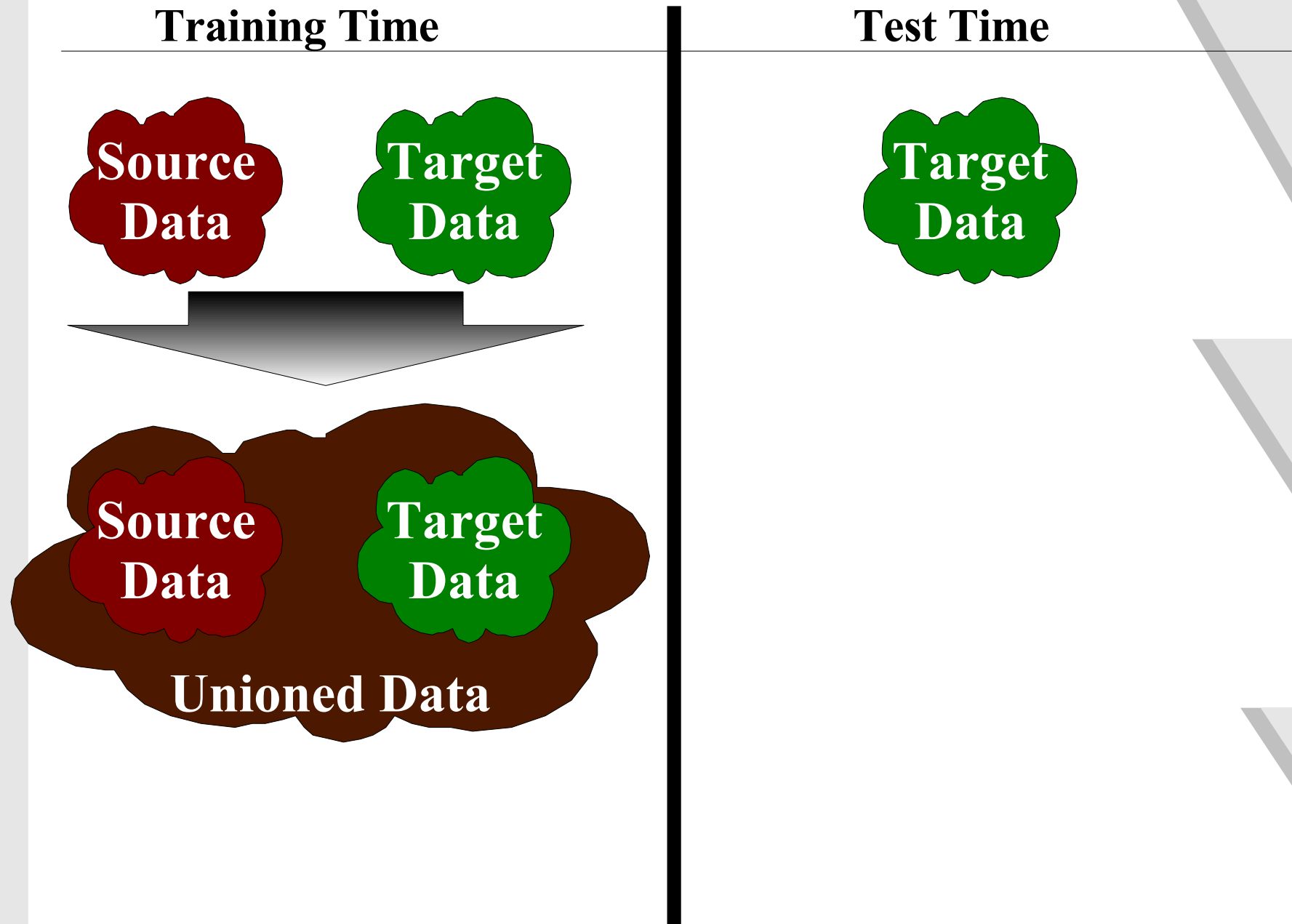
Training Time



Test Time



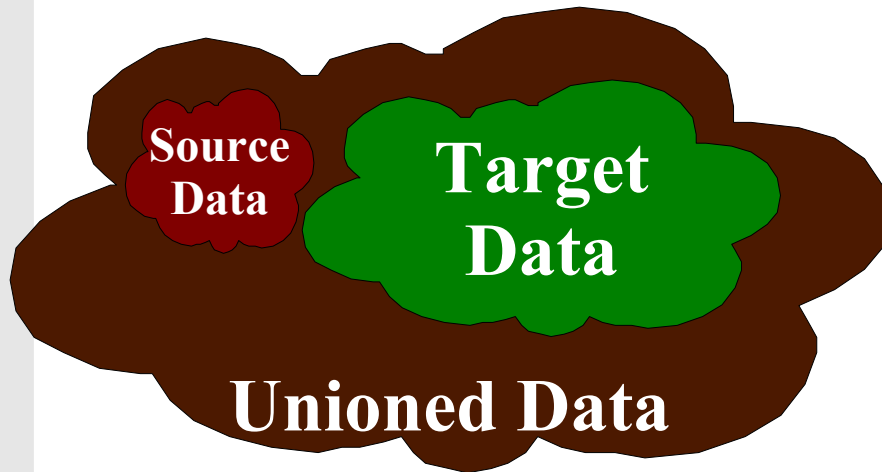
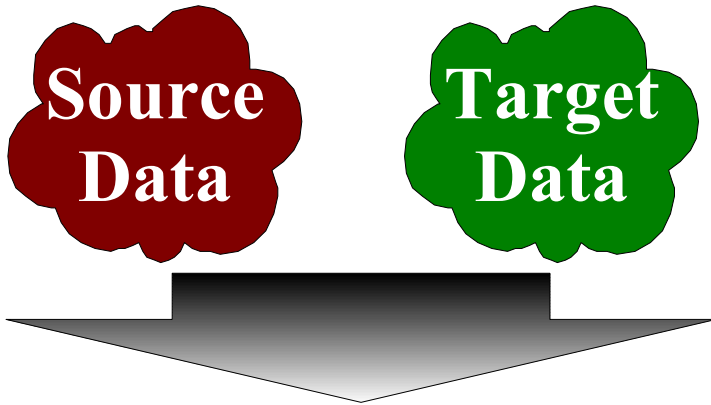
Obvious Approach 3: All



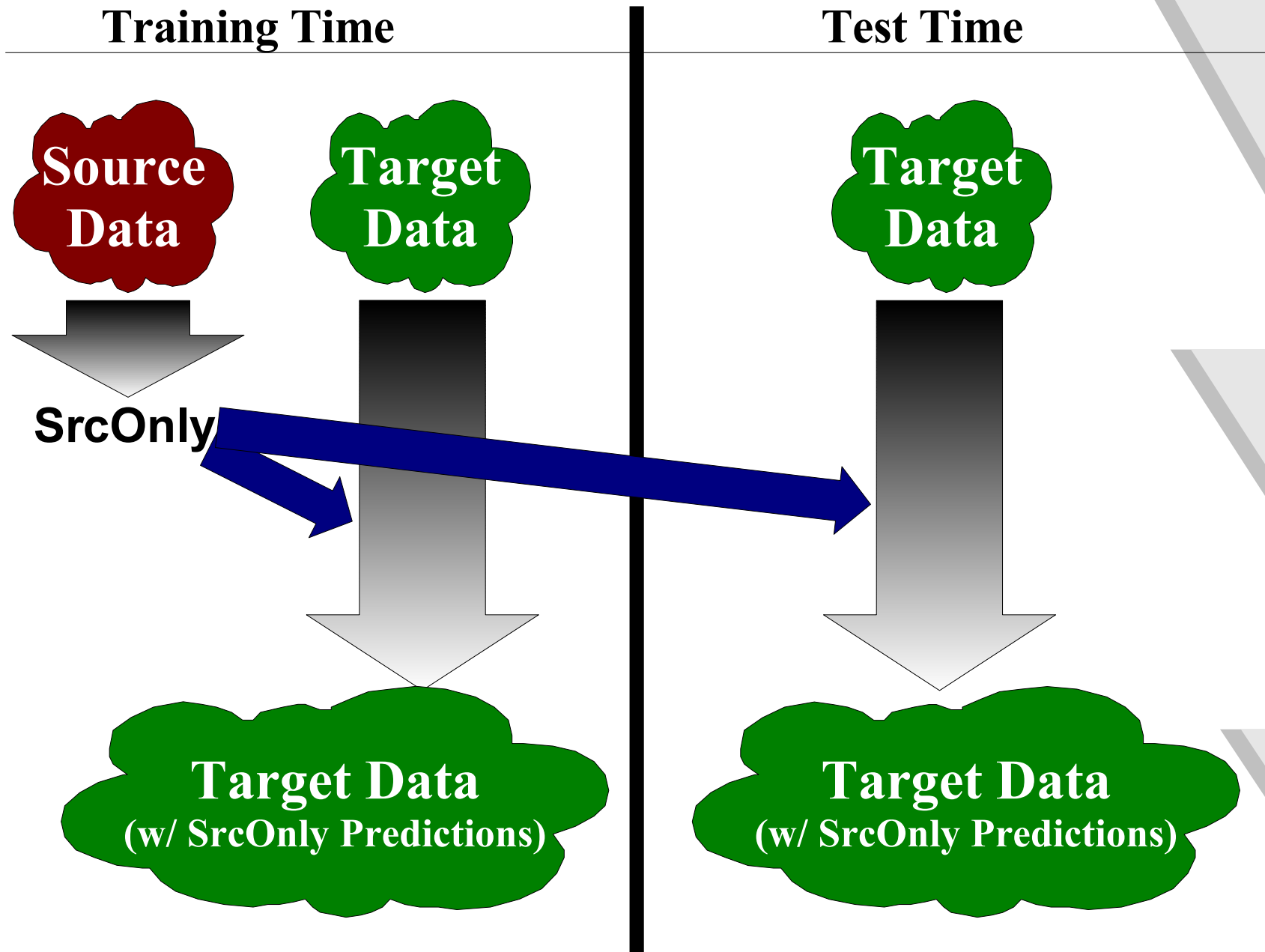
Obvious Approach 4: Weighted

Training Time

Test Time

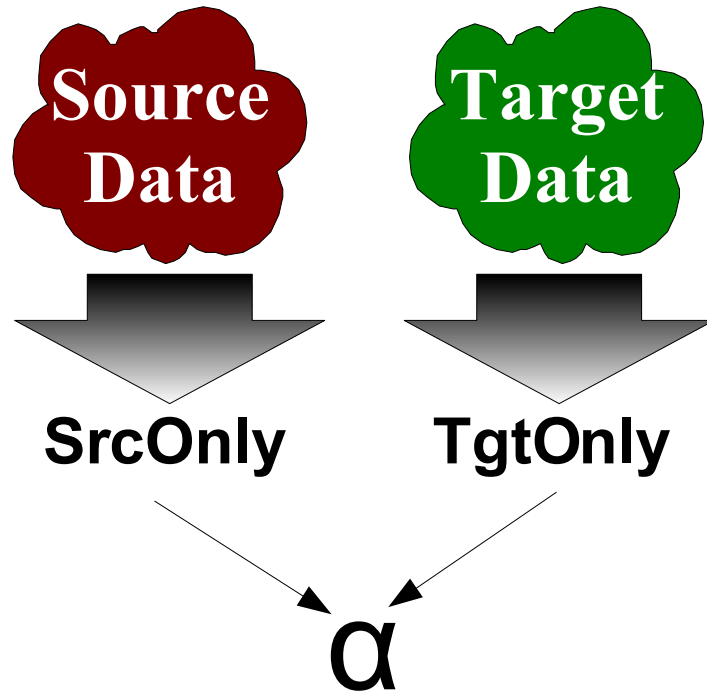


Obvious Approach 5: Pred

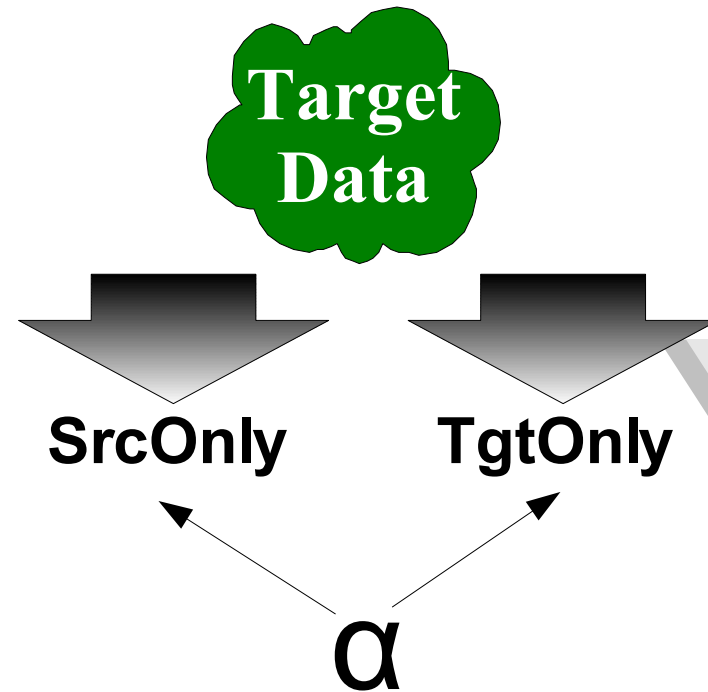


Obvious Approach 6: LinInt

Training Time



Test Time



Results – Error Rates

Task	Dom	SrcOnly	TgtOnly	Baseline	Prior	Augment
ACE- NER	bn	4.98	2.37	2.11 (pred)	2.06	1.98
	bc	4.54	4.07	3.53 (weight)	3.47	3.47
	nw	4.78	3.71	3.56 (pred)	3.68	3.39
	wl	2.45	2.45	2.12 (all)	2.41	2.12
	un	3.67	2.46	2.10 (linint)	2.03	1.91
	cts	2.08	0.46	0.40 (all)	0.34	0.32
CoNLL	tgt	2.49	2.95	1.75 (wgt/li)	1.89	1.76
PubMed	tgt	12.02	4.15	3.95 (linint)	3.99	3.61
CNN	tgt	10.29	3.82	3.44 (linint)	3.35	3.37
	wsj	6.63	4.35	4.30 (weight)	4.27	4.11
	swbd3	15.90	4.15	4.09 (linint)	3.60	3.51
Tree bank- Chunk	br-cf	5.16	6.27	4.72 (linint)	5.22	5.15
	br-cg	4.32	5.36	4.15 (all)	4.25	4.90
	br-ck	5.05	6.32	5.01 (prd/li)	5.27	5.41
	br-cl	5.66	6.60	5.39 (wgt/prd)	5.99	5.73
	br-cm	3.57	6.59	3.11 (all)	4.08	4.89
	br-cn	4.60	5.56	4.19 (prd/li)	4.48	4.42
	br-cp	4.82	5.62	4.55 (wgt/prd/li)	4.87	4.78
	br-cr	5.78	9.13	5.15 (linint)	6.71	6.30
Treebank-	brown	6.35	5.75	4.72 (linint)	4.72	4.65

Discussion

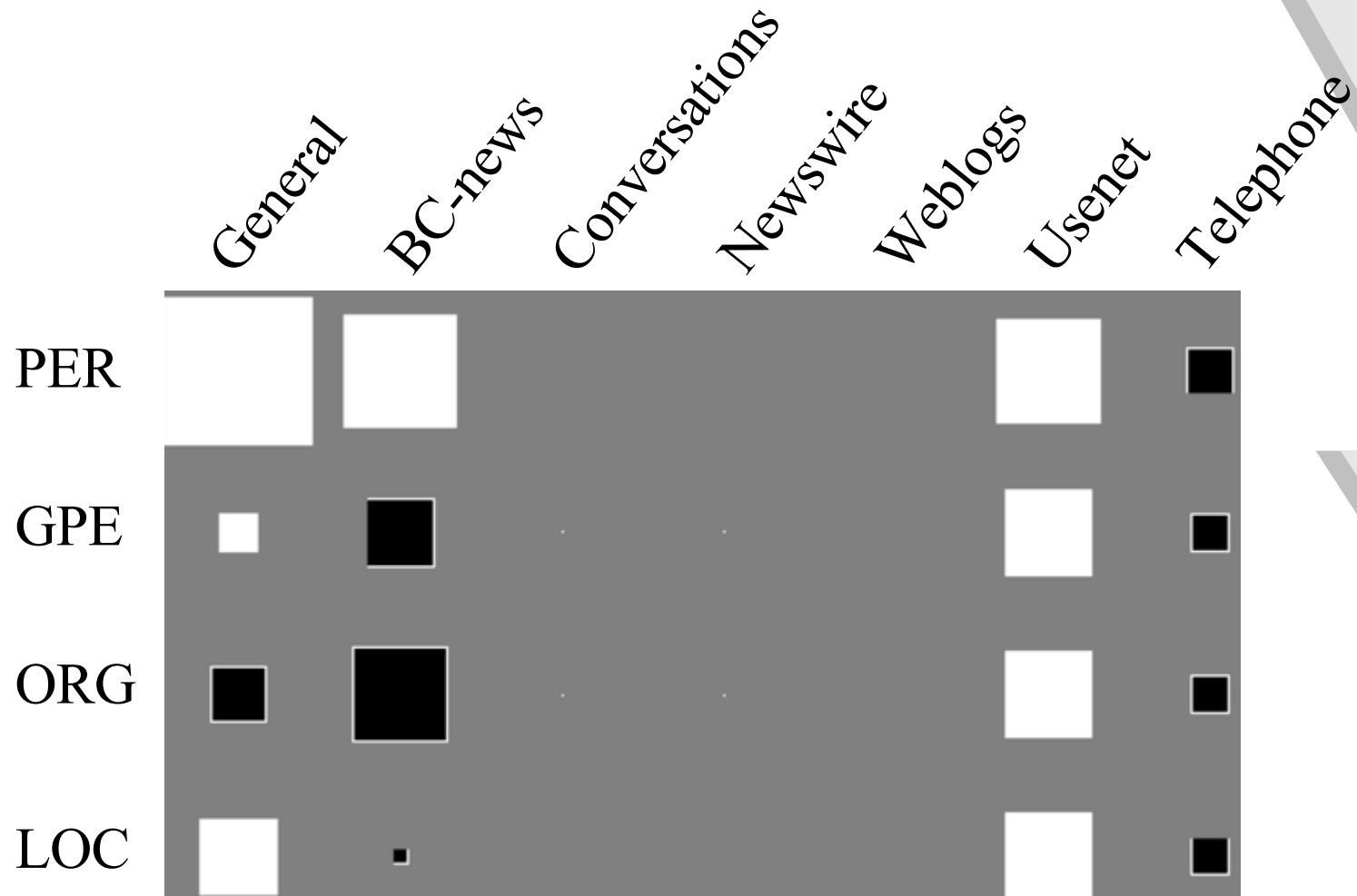
- What's good?
 - Works well (if $T < S$), applicable to any classifier
 - Easy to implement –
10 lines of Perl: <http://hal3.name/easyadapt.pl.gz>
 - Very fast – leverages any classifier
- What could perhaps be slightly better maybe?
 - Theory – why should this help?
 - Unannotated target data?

Thanks!
Questions?

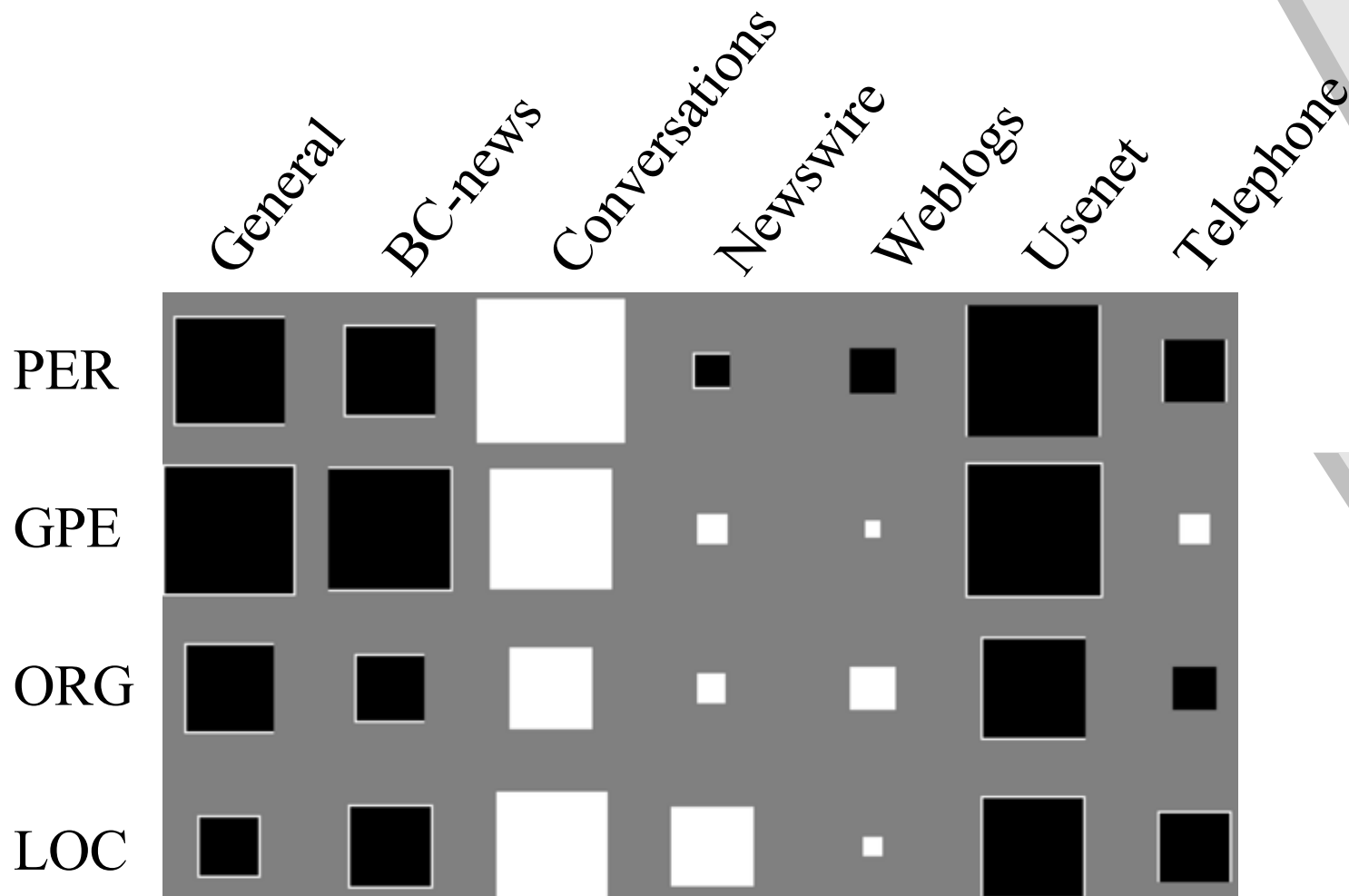
State of Affairs

	Perf.	Impl.	Speed	Generality
Baselines (Numerous)	Bad	Good	Good	Good
Prior (Chelba+ Acero)	Good	Okay	Good	Okay
MegaM (Daume+ Marcu)	Great	Terrible	Terrible	Okay
Proposed approach	Very Good	Great	Good	Great

Hinton Diagram: /bush/ on ACE-NER



Hinton Diagram: /P=the/ on ACE-NER



“the Iraqi people”

“the Pentagon”

“the Bush (advisors|cabinet|..)”

“the South”