Transfer Learning for NLP

Sebastian Ruder
Research Scientist, AYLIEN
PhD Candidate, Insight Centre, Dublin

Agenda

1. What is Transfer Learning?

2. Why Transfer Learning now?

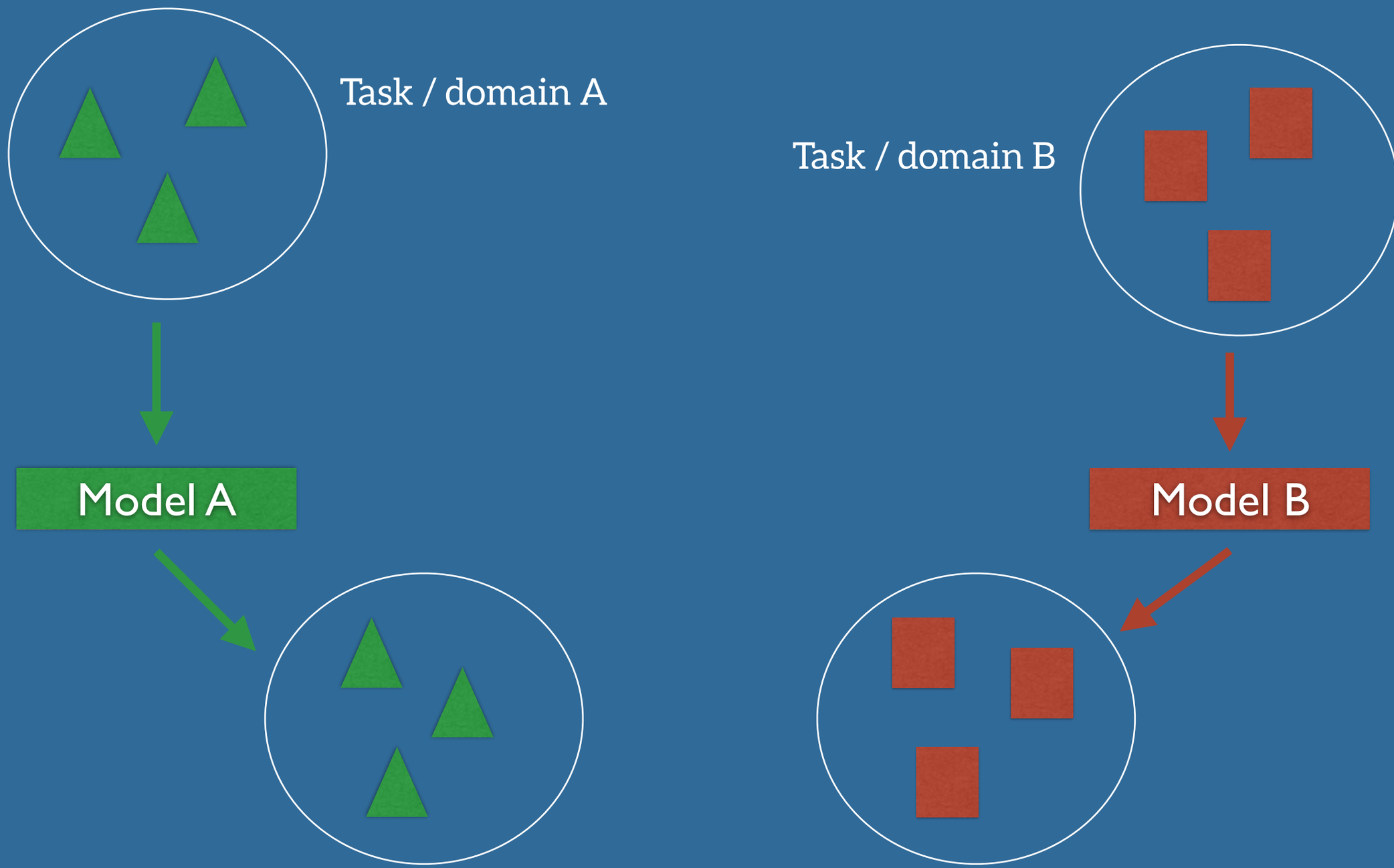
3. Transfer Learning in practice

4. Transfer Learning for NLP

5. Current research

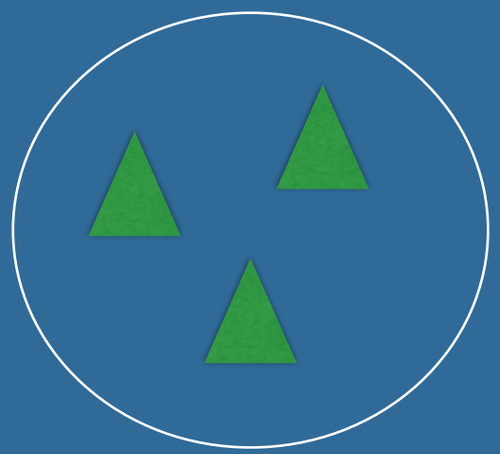
What is Transfer Learning?

Traditional ML



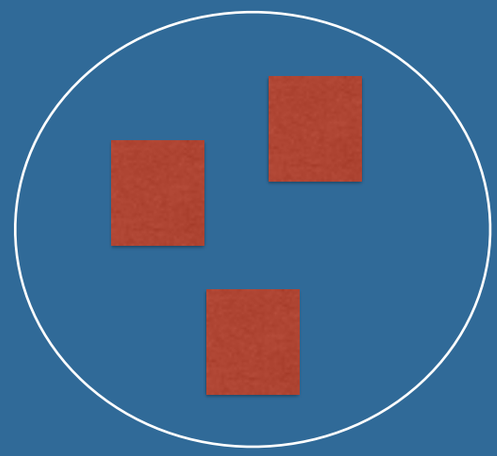
What is Transfer Learning?

Traditional ML



Task / domain A

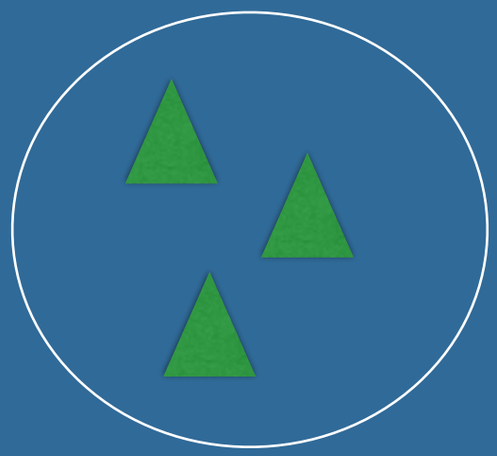
Task / domain B

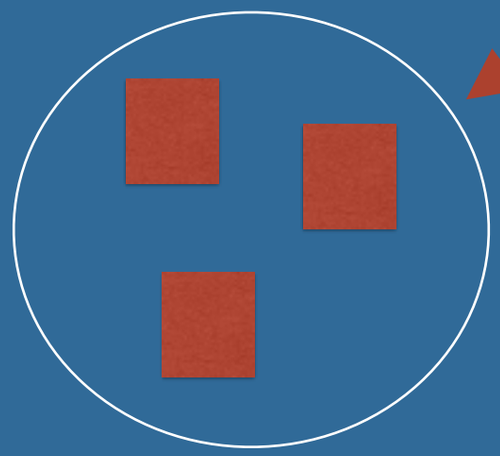


Training and evaluation on the same task or domain.

Model A

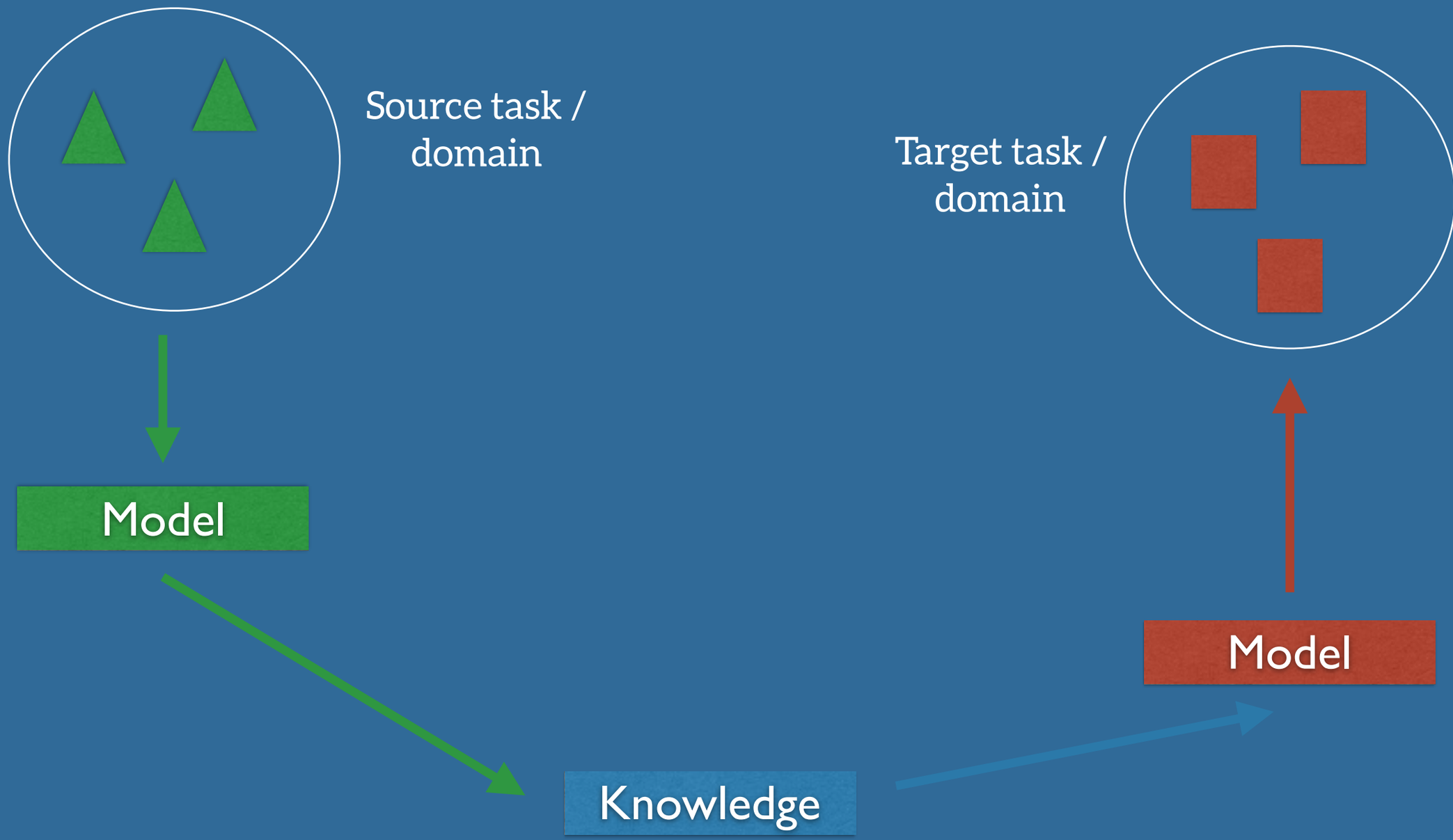
Model B





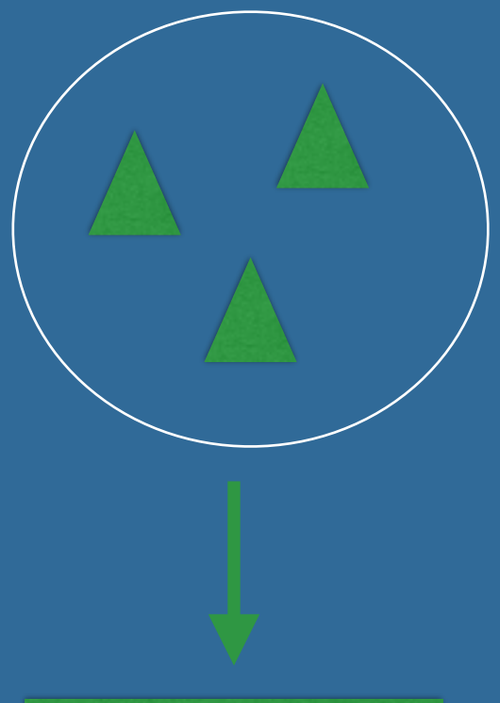
What is Transfer Learning?

Transfer learning



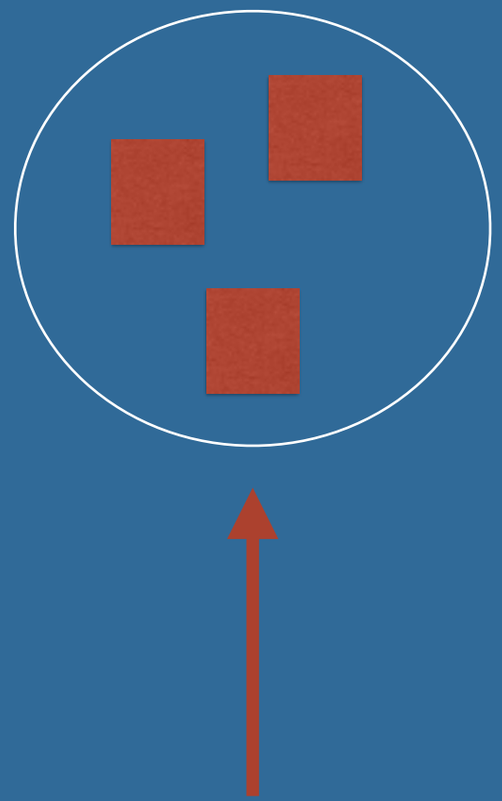
What is Transfer Learning?

Transfer learning



Source task / domain

Model

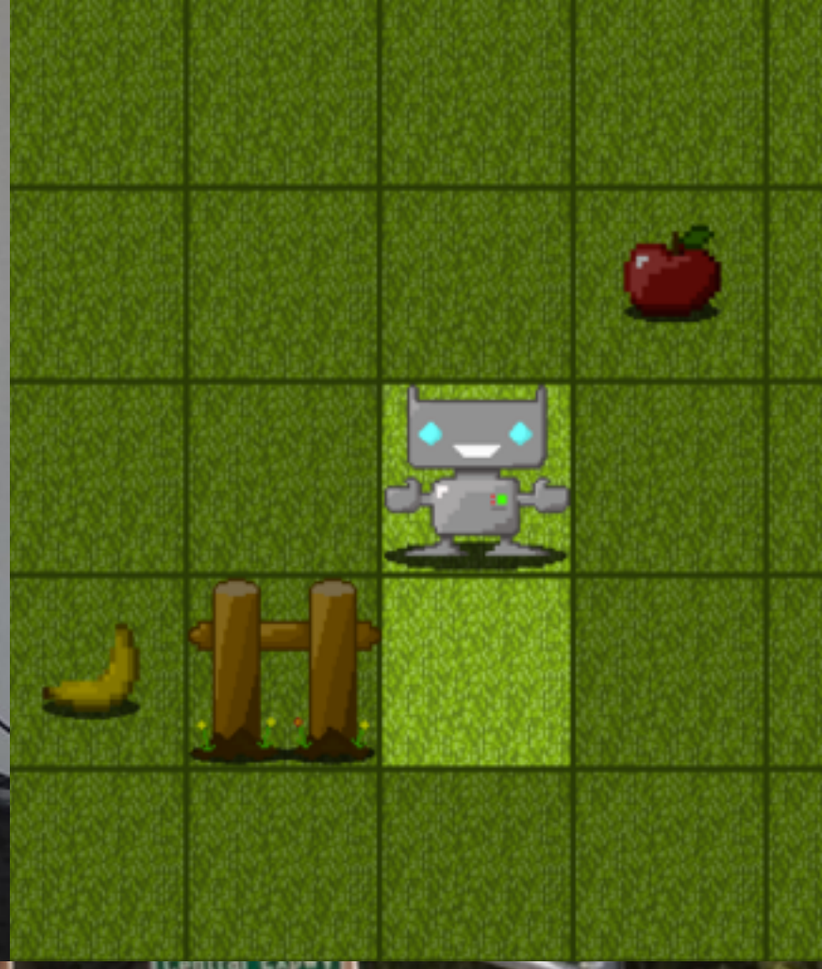


Target task / domain

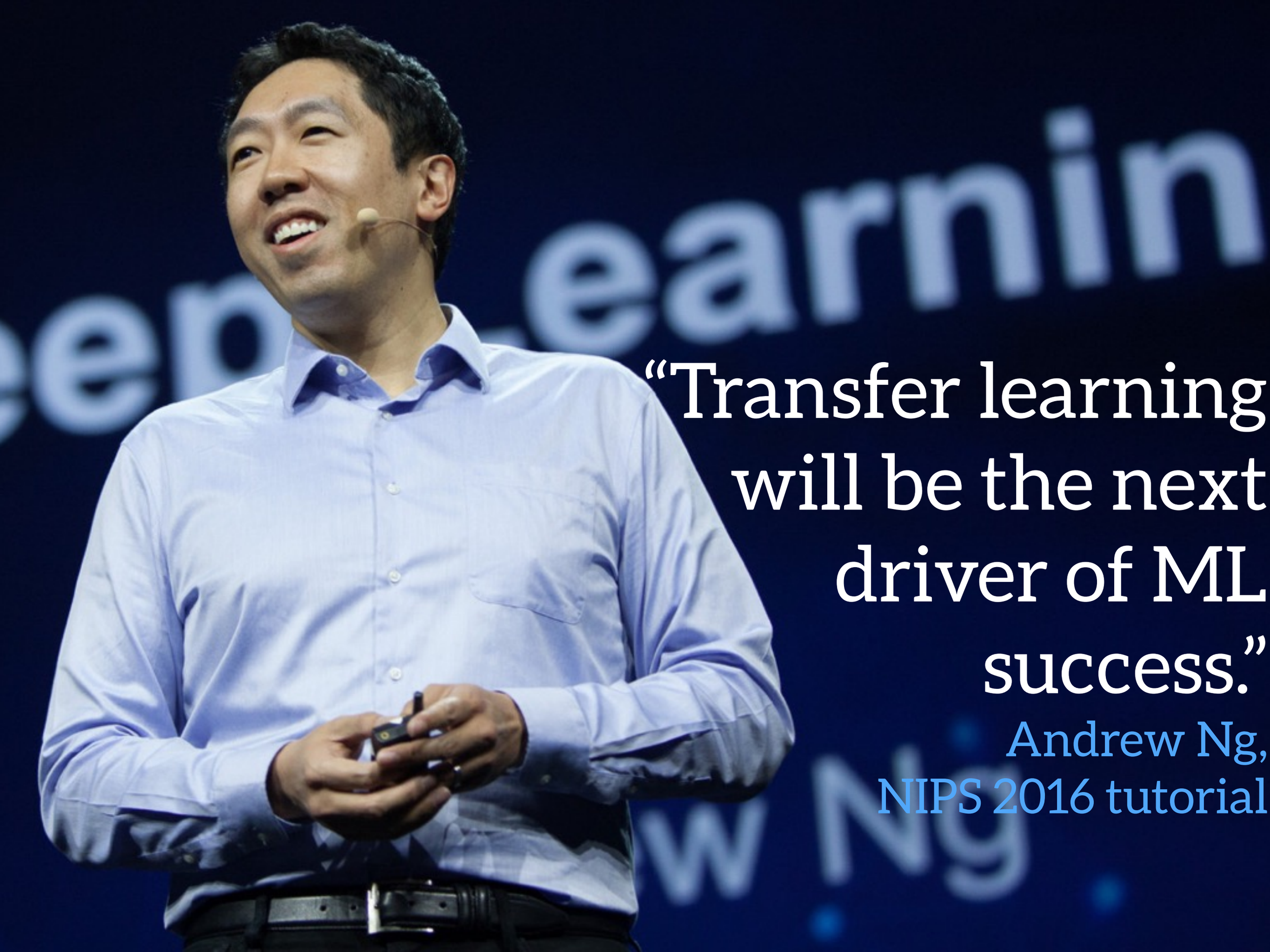
Model

Storing knowledge gained solving one problem and applying it to a different but related problem.

Knowledge

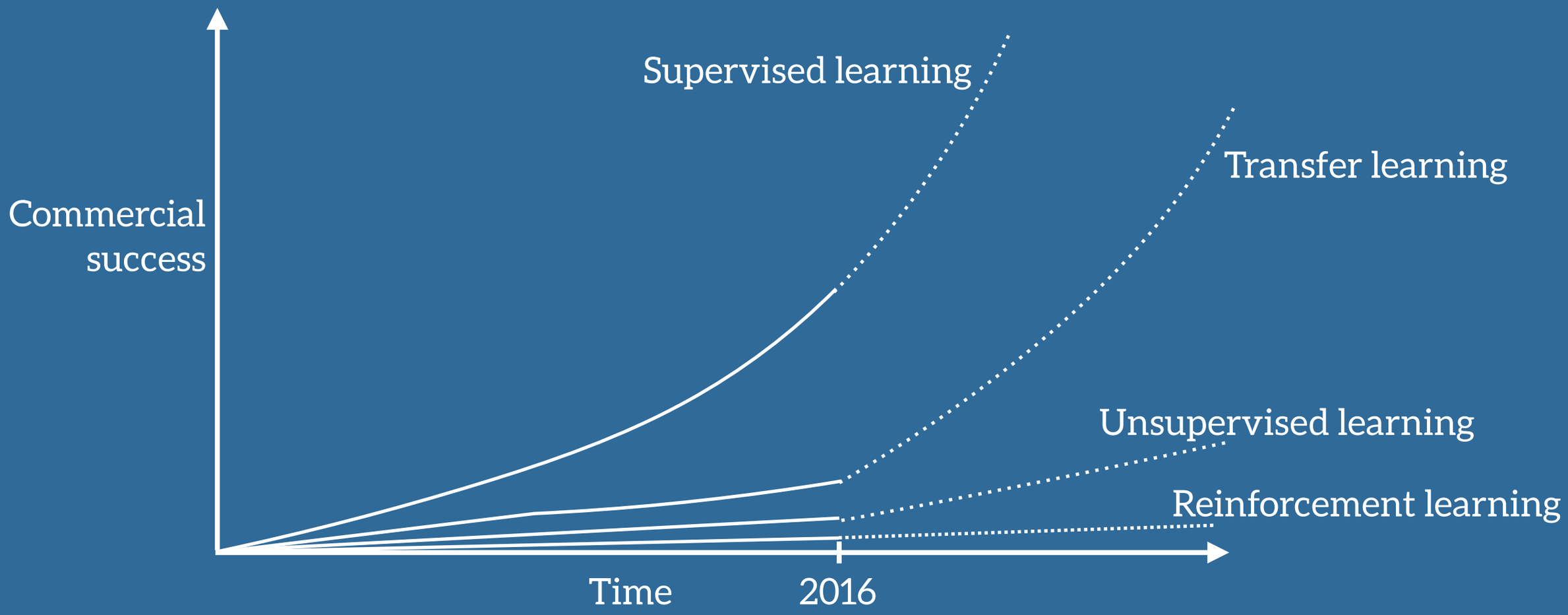



RED LIGHT
VIOLATION
$336
MINIMUM FINE
6YCT869
Google
self-driving car


“Transfer learning will be the next driver of ML success.”
Andrew Ng,
NIPS 2016 tutorial

Why Transfer Learning now?

Drivers of ML success in industry



- Andrew Ng, NIPS 2016 tutorial

Why Transfer Learning now?

1. Learn very accurate input-output mapping
2. Maturity of ML models
   - Computer vision (5% error on ImageNet)
   - Automatic speech recognition (3x faster than typing, 20% more accurate[1])
3. Large-scale deployment & adoption of ML models
   - Google's NMT System[2]

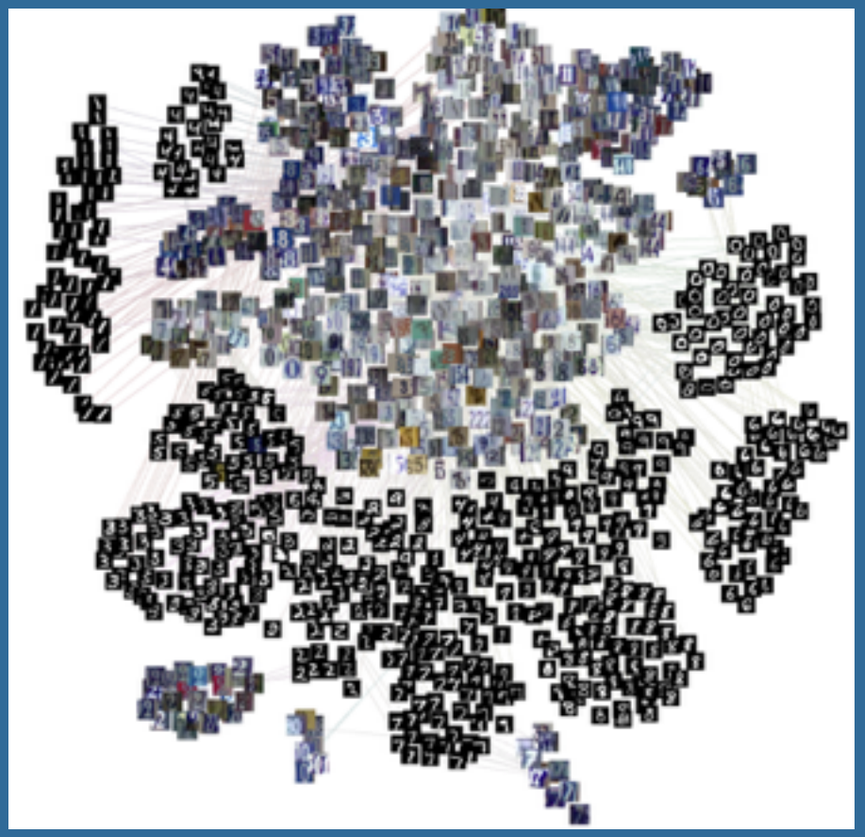
⟷ Huge reliance on labeled data
Novel tasks / domains without (labeled) data

[1] Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. (2016). Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices. arXiv preprint arXiv:1608.07323.
[2] Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144.

Transfer Learning in practice

Computer vision

- Train new model on features of large model trained on ImageNet[3]
- Train model to confuse source and target domains[4]
- Train model on domain-invariant representations[5,6]



[3] Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 512–519.
[4] Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. Proceedings of the 32nd International Conference on Machine Learning., 37.
[5] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain Separation Networks. NIPS 2016.
[6] Sener, O., Song, H. O., Saxena, A., & Savarese, S. (2016). Learning Transferrable Representations for Unsupervised Domain Adaptation. NIPS 2016.

Transfer Learning for NLP

A more technical definition

- Task $\mathcal{T}$ and domain $\mathcal{D}$
- Domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where
  - $\mathcal{X}$: feature space, e.g. BOW representations
  - $P(X)$: e.g. distribution over terms in documents
- Task $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$ where
  - $\mathcal{Y}$: label space, e.g. true/false labels
  - $P(Y|X)$: learned mapping from samples to labels
- Transfer learning:
  Learning when $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$

Transfer Learning for NLP

Transfer scenarios

1. $P(X_S) \neq P(X_T)$: Different topics, text types, etc.

2. $\mathcal{X}_S \neq \mathcal{X}_T$: Different languages.

3. $P(Y_S|X_S) \neq P(Y_T|X_T)$: Unbalanced classes.

4. $\mathcal{Y}_S \neq \mathcal{Y}_T$: Different tasks.

Transfer Learning for NLP

Transfer scenarios

1. $P(X_S) \neq P(X_T)$: Different topics, text types, etc.

2. $\mathcal{X}_S \neq \mathcal{X}_T$: Different languages.

3. $P(Y_S|X_S) \neq P(Y_T|X_T)$: Unbalanced classes.

4. $\mathcal{Y}_S \neq \mathcal{Y}_T$: Different tasks.

Transfer Learning for NLP

Transfer scenarios

1. $P(X_S) \neq P(X_T)$: Different topics, text types, etc.

2. $\mathcal{X}_S \neq \mathcal{X}_T$: Different languages.

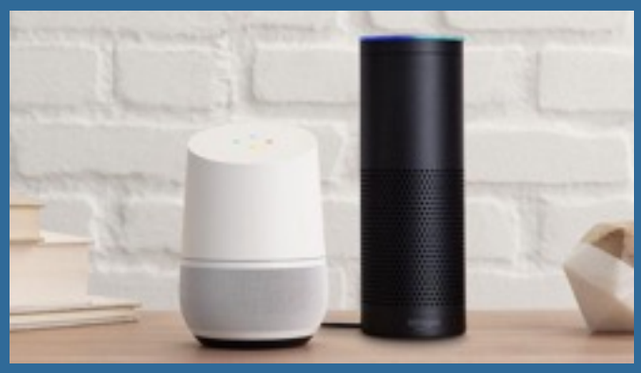
Domain adaptation

3. $P(Y_S|X_S) \neq P(Y_T|X_T)$: Unbalanced classes.

4. $\mathcal{Y}_S \neq \mathcal{Y}_T$: Different tasks.

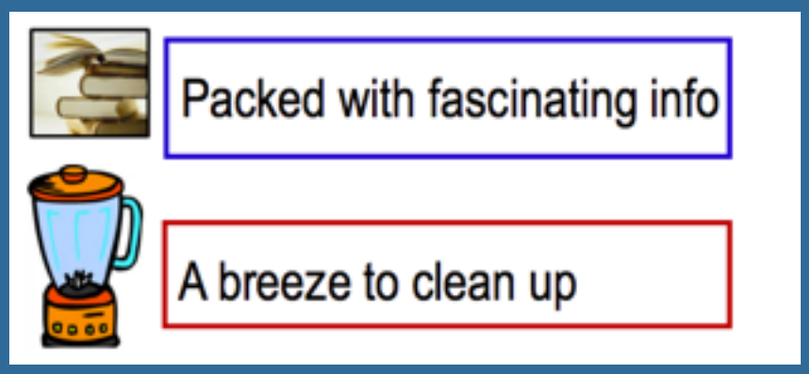
Transfer Learning for NLP

Training and test distributions are different.



Different text types.

Different accents/ages.



Different topics/categories.

→ Performance drop or even collapse is inevitable.

Transfer Learning for NLP

Current status

- Not as straightforward as in CV
  - No universal deep features
- However: “Simple” transfer through word embeddings is pervasive
- History of research for task-specific transfer, e.g. sentiment analysis, POS tagging leveraging NLP phenomena such as structured features, sentiment words, etc.
- Few research on transfer between tasks
- More recently: research on representations

Current research

Transfer learning challenges in real-world scenarios

1. One-to-one adaptation is rare and many source domains are generally available.

2. Models need to be adapted frequently as conditions change, new data becomes available, etc.

3. Target domains may be unknown or no target domain data may be available.

Current research

Two recent examples

1. Sebastian Ruder, Barbara Plank. (2017). **Learning to select data for transfer learning with Bayesian Optimization**.

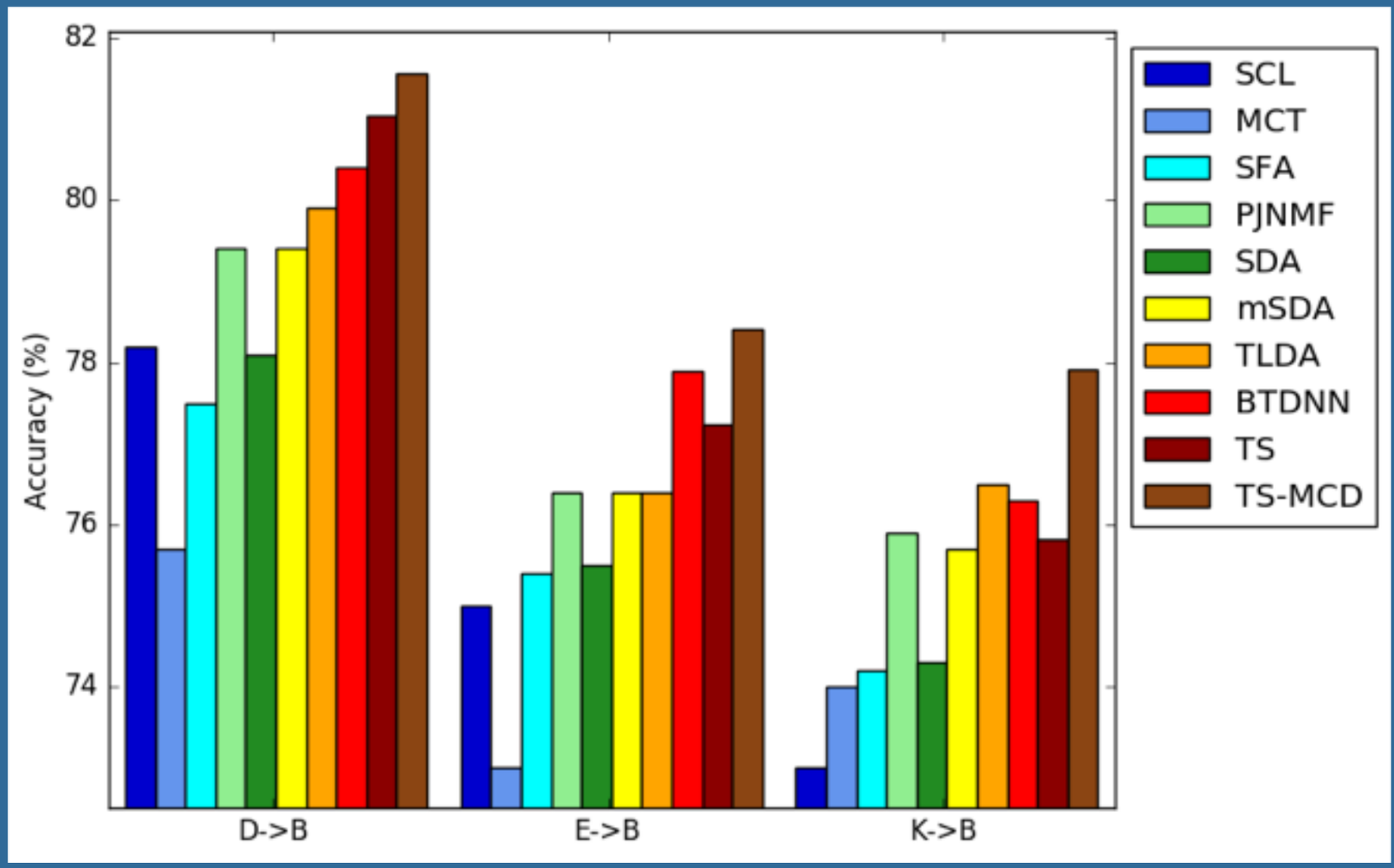
2. Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, Anders Søgaard. (2017). **Sluice networks: Learning what to share between loosely related tasks**. arXiv preprint arXiv:1705.08142

Current research

Learning to select data for transfer learning I

Why data selection for transfer learning?

- Knowing what data is relevant directly impacts downstream performance.
- E.g. sentiment analysis: stark performance differences across source domains.



Current research

Learning to select data for transfer learning II

- Existing approaches
  i. study domain similarity metrics in isolation;
  ii. use only similarity;
  iii. focus on a single task.

- Intuition: Different tasks and domains demand a different notion of similarity.

- Idea: Learn the similarity metric.

Current research

Learning to select data for transfer learning III

- Treat similarity metric as black-box function; use Bayesian Optimisation (BayesOpt).

- Learn linear data selection metric $\mathcal{S}$:
  $\mathcal{S} = \phi(X) \cdot w^\intercal$ where $\phi(X) \in \mathbb{R}^{n \times l}$ are data selection features for each training example, l is # of features, and $w \in \mathbb{R}^l$ are weights learned by BayesOpt.

- Use $\mathcal{S}$ to rank all training examples; select top n for training.

Current research

Learning to select data for transfer learning IV

Features

- **Similarity**: Jensen-Shannon divergence, Rényi divergence, Bhattacharyya distance, cosine similarity, Euclidean distance, variational distance.
- **Representations**: Term distributions, topic distributions, word embeddings.
- **Diversity**: # of word types, type-token ratio, entropy, Simpson's index, Rényi entropy, quadratic entropy

Current research

Learning to select data for transfer learning V

- **Tasks**: sentiment analysis, part-of-speech (POS) tagging, dependency parsing

- **Datasets and models**:
  i. *Sentiment analysis*: Amazon reviews dataset (Blitzer et al., 2006); linear SVM.
  ii. *POS tagging*: SANCL 2012 shared task; Structured Perceptron, SotA BiLSTM tagger (Plank et al., 2016).
  iii. *Dependency parsing*: SotA BiLSTM parser (Kiperwasser and Goldberg, 2016).

Current research

Learning to select data for transfer learning VI

Feature set	Book	DVD	Electronics	Kitchen
Random	71.17	70.51	76.75	77.94
JS divergence (examples)	72.51	68.21	76.41	77.47
JS divergence (domain)	75.26	73.74	72.60	80.01
Similarity + diversity (term dists)	76.20	77.60	**82.67**	**84.98**
Similarity + diversity (topic dists)	**77.16**	**79.00**	81.92	84.92

- Learned measures outperform baselines across all tasks and domains.
- Competitive with SotA domain adaptation.
- Diversity complements similarity.
- Measures transfer across models, domains, tasks.

Current research

Learning what to share for multi-task learning I

What if the target domain is unknown?

- Have a model that works well across most domains (Ben-David et al., 2007).
- Many ways to do this:
  - Data augmentation (mainly in computer vision)
  - Regularisation
    i. Norms, e.g. ℓ_1 (lasso), ℓ_2, group lasso, etc.
    ii. Dropout; noise;
    iii. Multi-task learning (MTL)

Current research

Learning what to share for multi-task learning I

What if the target domain is unknown?

- Have a model that works well across most domains (Ben-David et al., 2007).
- Many ways to do this:
  - Data augmentation (mainly in computer vision)
  - Regularisation
    i. Norms, e.g. ℓ_1 (lasso), ℓ_2, group lasso, etc.
    ii. Dropout; noise;
    iii. Multi-task learning (MTL)

Current research

Learning what to share for multi-task learning I

What if the target domain is unknown?

- Have a model that works well across most domains (Ben-David et al., 2007).
- Many ways to do this:
  - Data augmentation (mainly in computer vision)
  - Regularisation
    i. Norms, e.g. ℓ_1 (lasso), ℓ_2, group lasso, etc.
    ii. Dropout; noise;
    iii. Multi-task learning (MTL)

$\rightarrow$ Helps transfer knowledge across tasks

Current research

Learning what to share for multi-task learning II

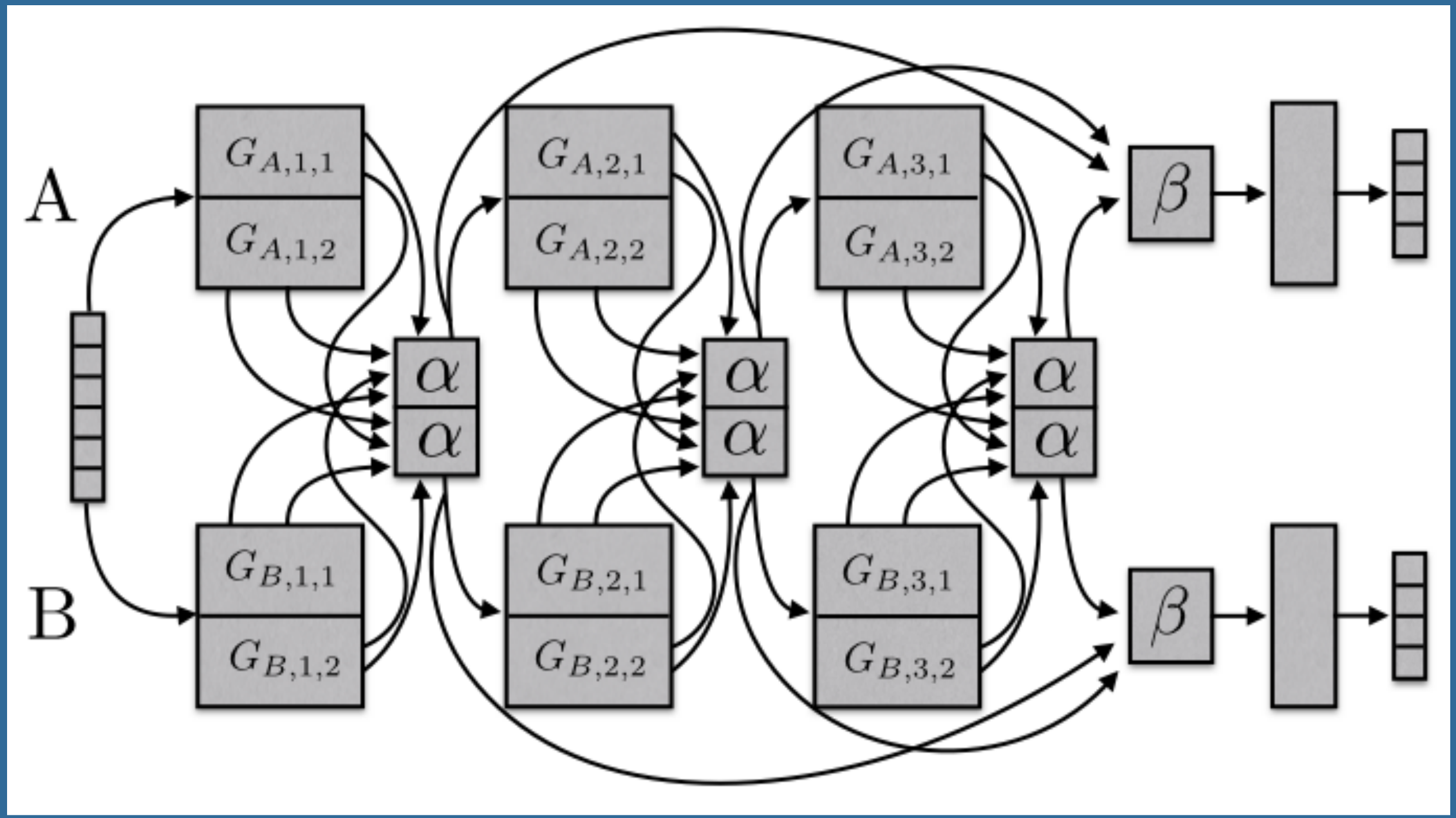
- Existing MTL approaches require amount of sharing to be manually specified; hard to estimate if sharing leads to improvements.

- Propose **Sluice Networks**, a general MTL framework.

- 3 parts:
  i. α parameters that control sharing across tasks;
  ii. β parameters that control sharing across layers;
  iii. orthogonality constraint to enforce subspaces.

Current research

Learning what to share for multi-task learning III



Current research

Learning what to share for multi-task learning IV

Experiments

- **Data**: OntoNotes 5.0; 7 domains.
- **Tasks**: chunking, NER, and semantic role labeling (main task) and POS tagging (auxiliary task).
- **Baselines**: single-task model, hard parameter sharing (Caruana, 1998), low supervision (Søgaard and Goldberg, 2016), cross-stitch networks (Misra et al., 2016).
- **Results**: Sluice networks outperform baselines on both in-domain and out-of-domain data.

Current research

Learning what to share for multi-task learning V

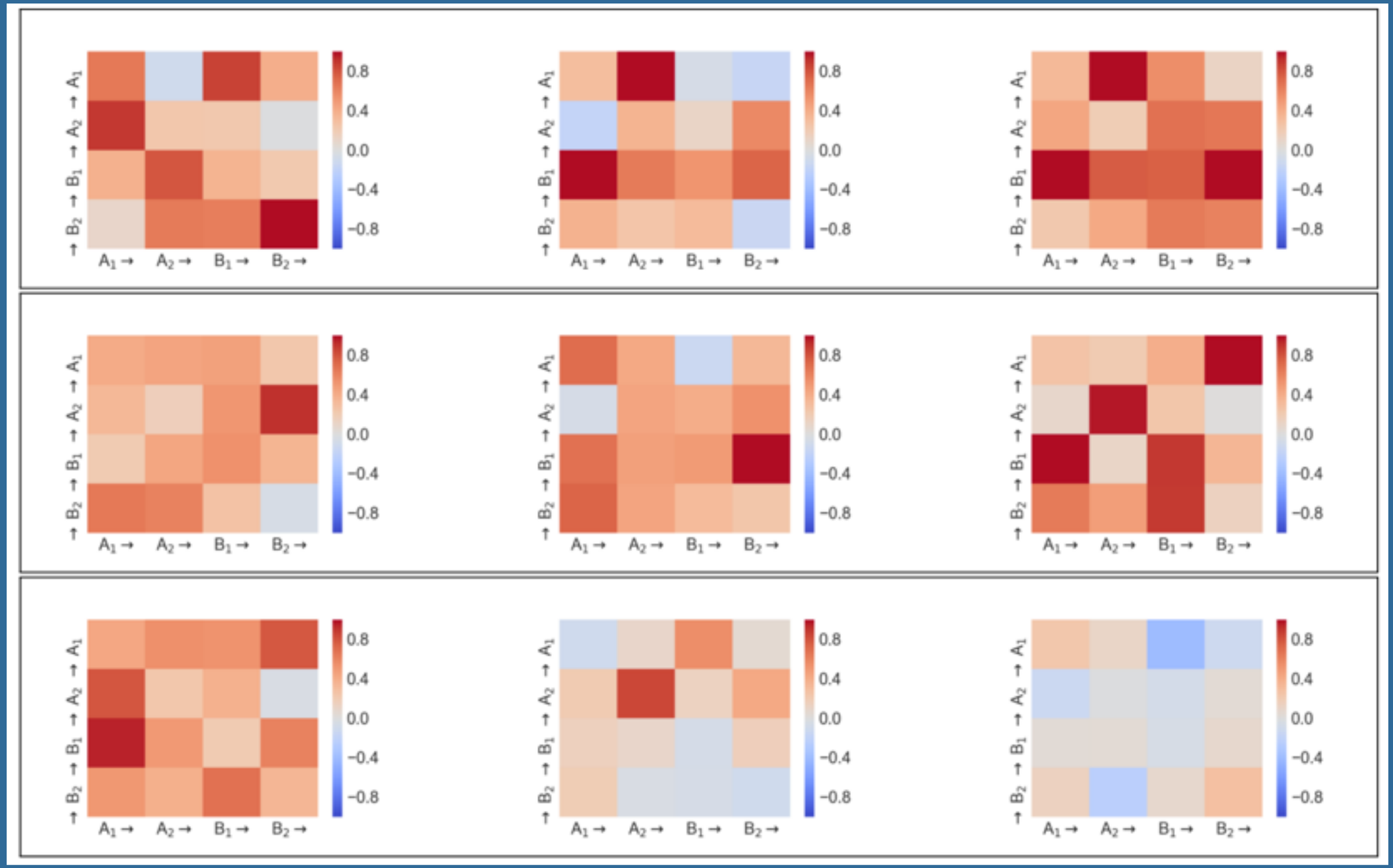
Analysis

- **Task properties and performance**
  - Gains are higher with less training data;
  - Sluice networks learn to share more with more variance in data.

- **Ablation analysis**
  - Learnable α and β values are better.
  - Subspaces help for all domains.
  - Concatenation of layer outputs also works.

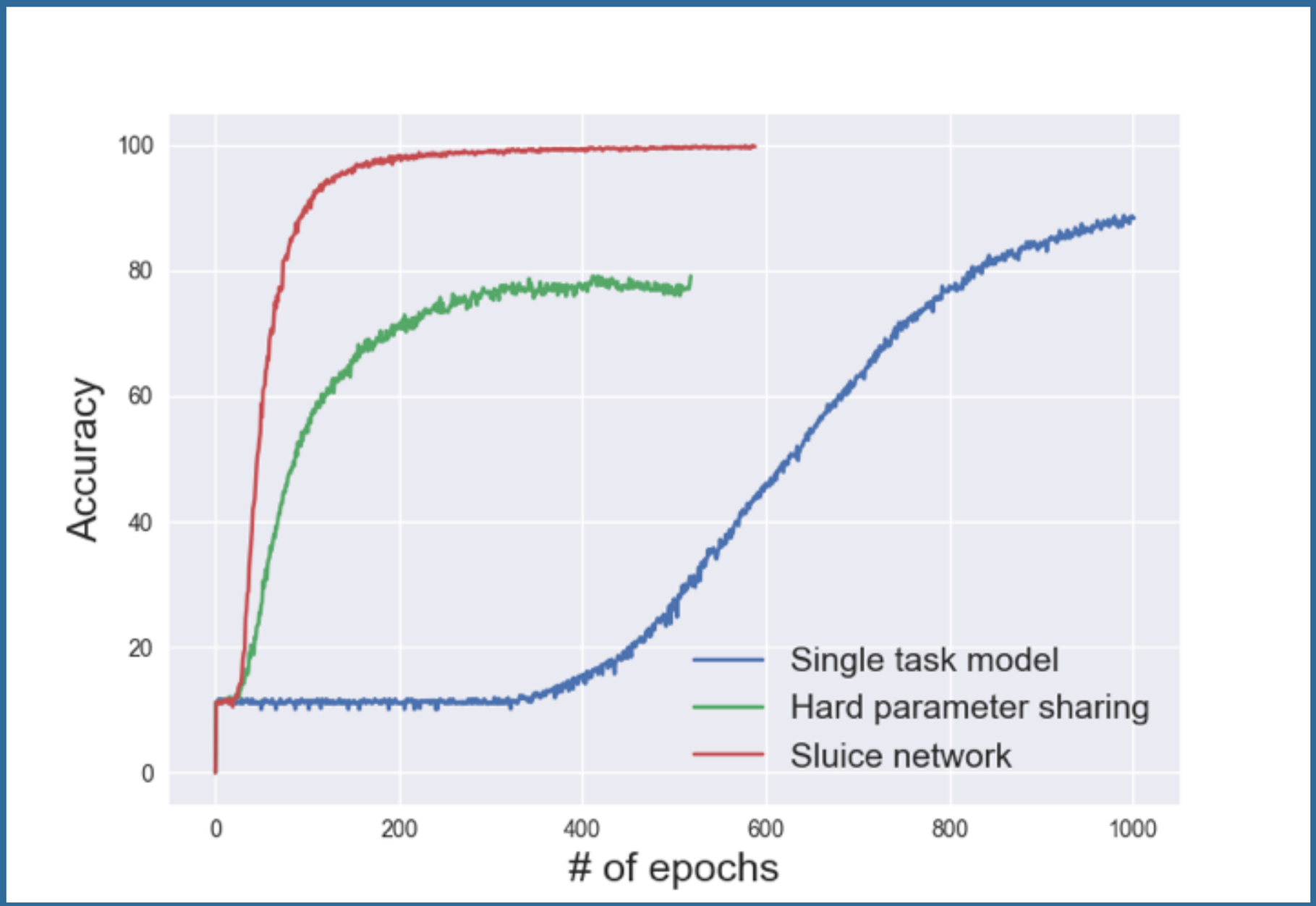
Current research

Learning what to share for multi-task learning VI



Current research

Learning what to share for multi-task learning VI



Conclusion

- Many new models, tasks, and domains.

→ Lots of cool work to do in Transfer Learning for NLP.

- Lots of fundamental unanswered research questions:
  - What is a domain? When are two domains similar?
  - When are two tasks related? …

- When should we use Transfer Learning / MTL? When does it work best?

References

Image credit

- Google Research blog post[11]
- Mikolov, T., Joulin, A., & Baroni, M. (2015). A Roadmap towards Machine Intelligence. arXiv preprint arXiv:1511.08130.
- Google Research blog post[12]

My papers

- Sebastian Ruder, Barbara Plank. (2017). **Learning to select data for transfer learning with Bayesian Optimization**.
- Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, Anders Søgaard. (2017). **Sluice networks: Learning what to share between loosely related tasks**. arXiv preprint arXiv:1705.08142

[11] https://research.googleblog.com/2016/10/how-robots-can-acquire-new-skills-from.html
[12] https://googleblog.blogspot.ie/2014/04/the-latest-chapter-for-self-driving-car.html

Thanks for your attention!

Questions?