Introduction to Transfer Learning

GENERAL FRAMEWORK AND NLP APPLICATIONS

METIS

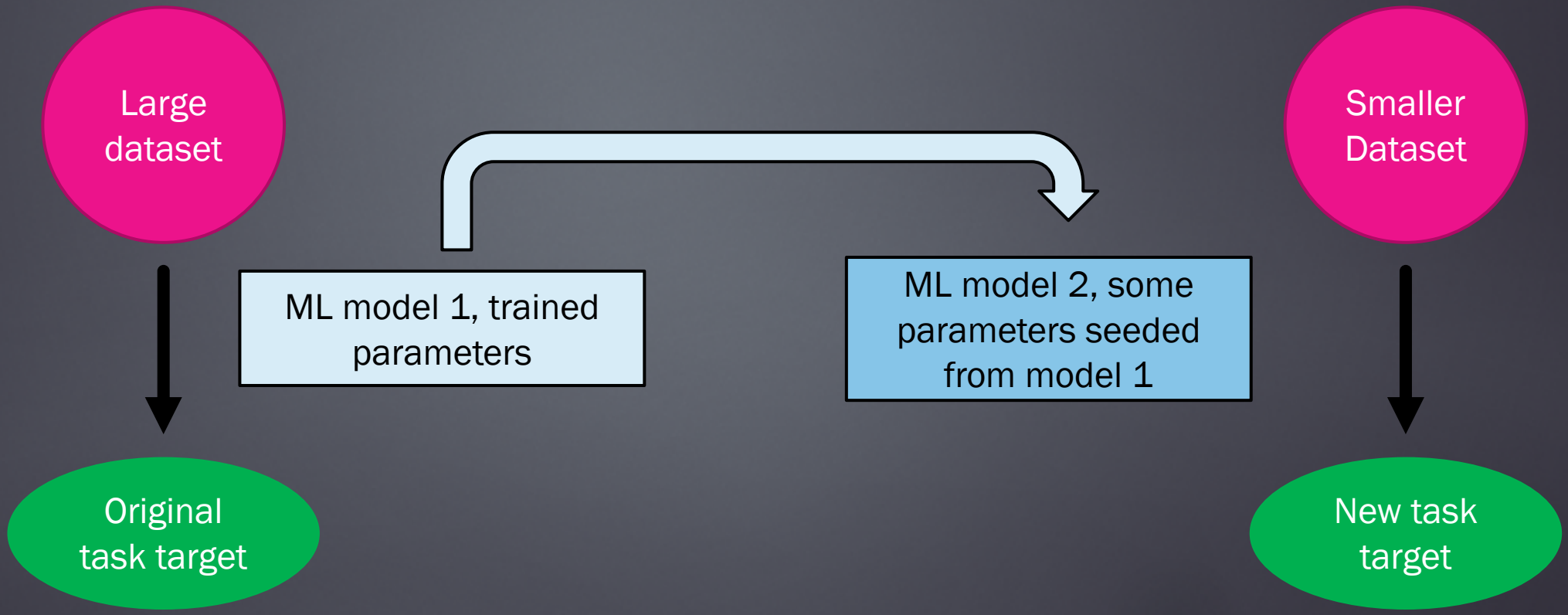
Roadmap

- What is the structure and purpose of transfer learning?
  - Leveraging larger datasets and optimizing training time
- What are examples of how it's applied in NLP?
  - Documents as sequences of pre-trained word embeddings
  - Seeding embedding parameters for text processing neural nets
- Are there examples beyond text?
  - Image recognition

What is transfer learning?

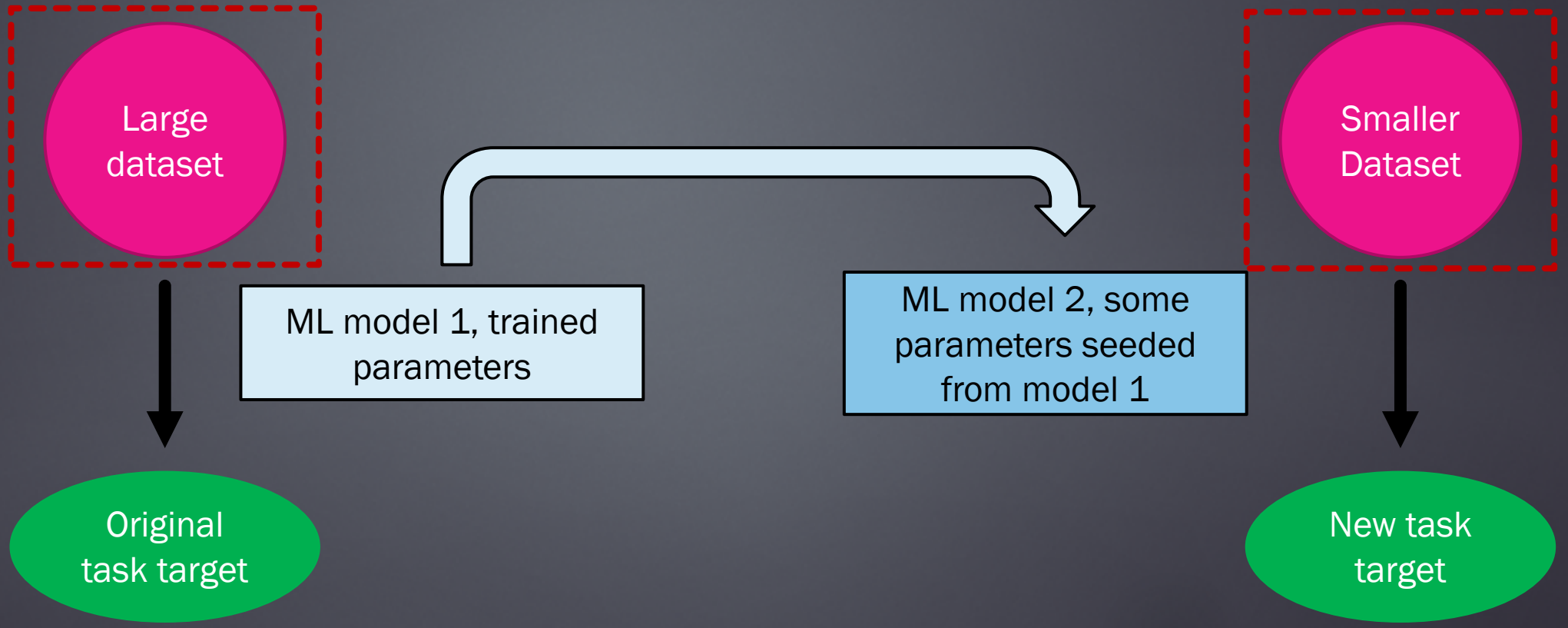
Transfer learning, as a diagram

- In transfer learning, we take parameters from a model learned on one predictive task/dataset as fixed or initial values for parameters of a 2nd model that will learn on a different task. Intuitively, we try to adapt information learned from a broader general task to a related specialist task with less training data.
- Example we'll see: general language modeling -> text classification



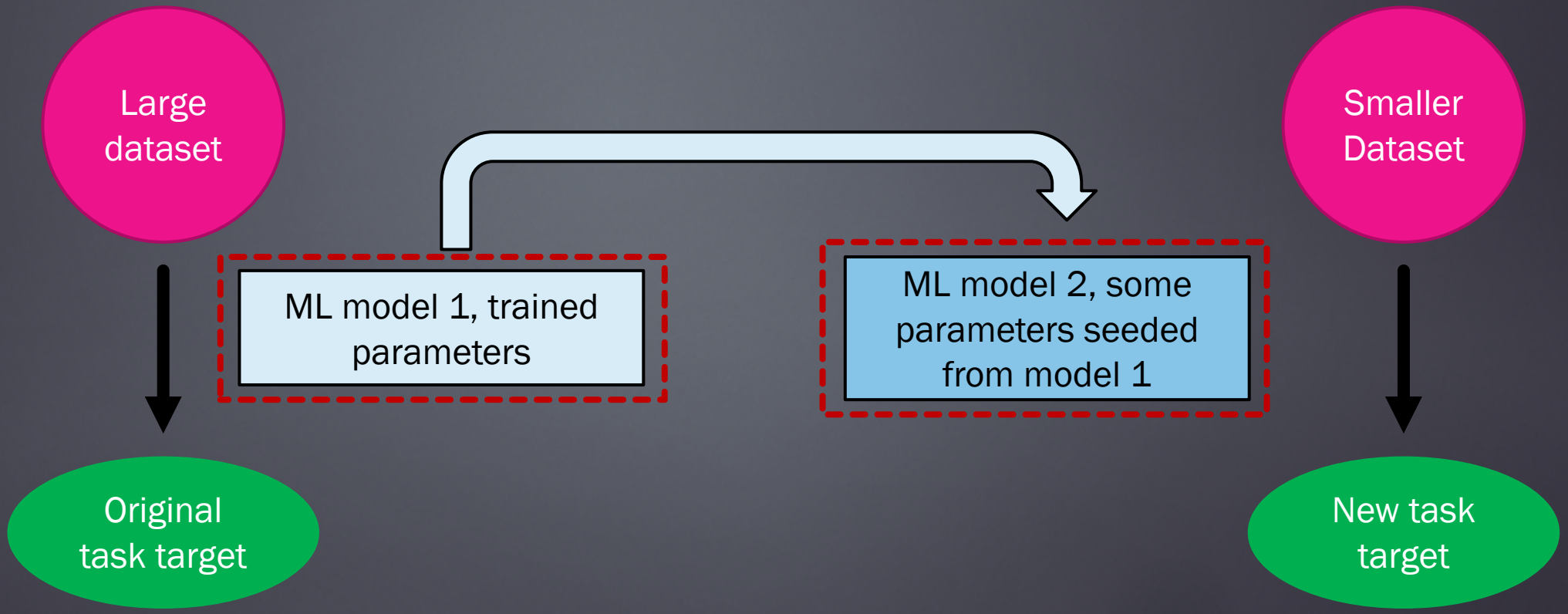
Advantages: dataset size

- If task 1 is sufficiently related to task 2, transfer learning is a means of leveraging a much larger knowledge base when building a predictive model for task 2, likely improving our model quality.



Advantages: training time

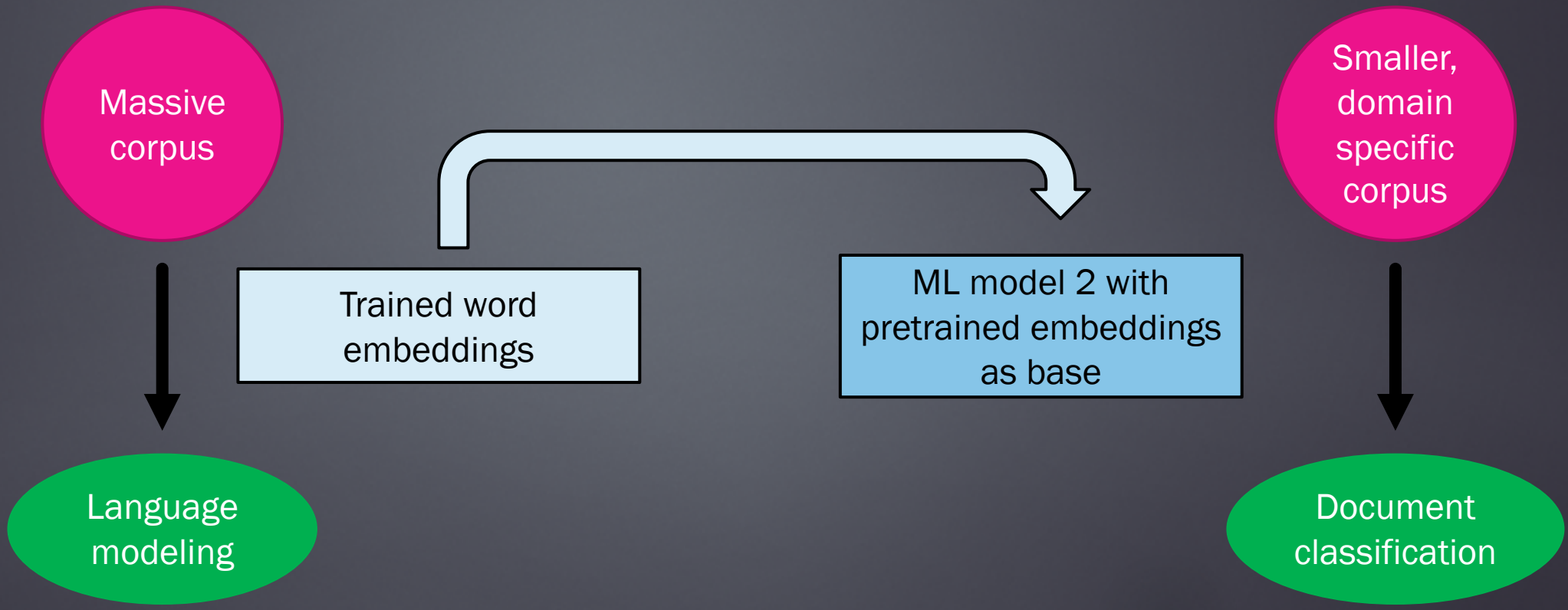
- If in modeling task 1 we've already found parameters that give a strong representation of some of the information contained in task 2, why reinvent the wheel? Using some pre-trained parameters means fewer parameters to learn for task 2, and therefore reduction in training time.



Applying transfer learning to NLP

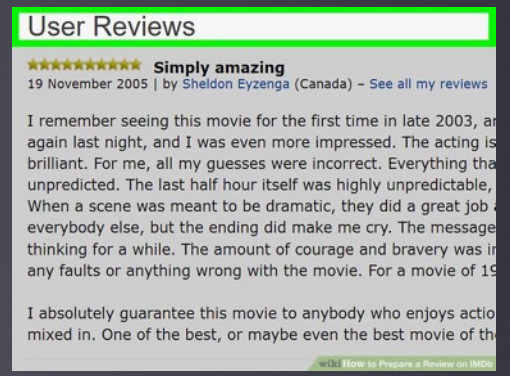
Transfer learning, NLP diagram

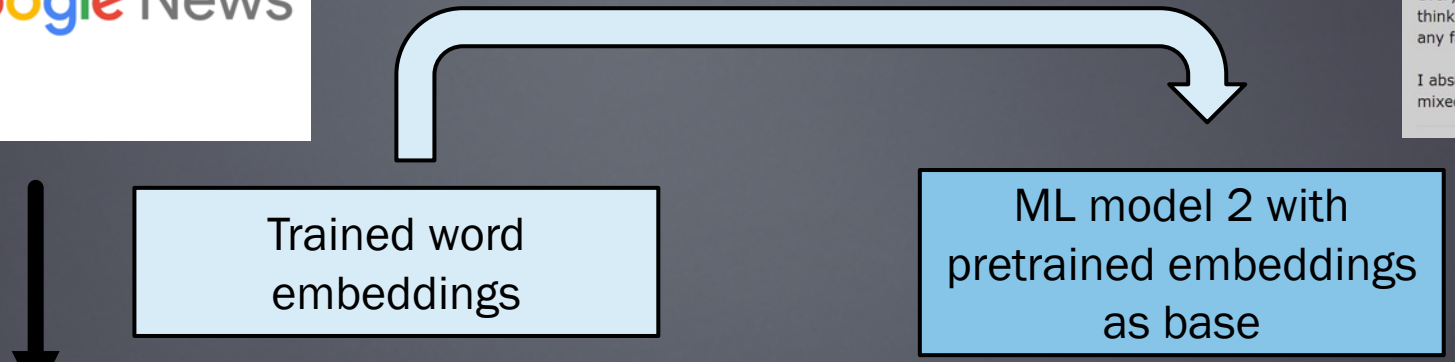
- NLP transfer learning often exploits the fact that there are readily accessible sources of massive text data for general language modeling, e.g. through the internet.
- These massive corpora are used to pre-train word embeddings that the downstream model uses to numerically represent words. Hopefully, the embeddings capture structural information about general language usage that is highly relevant to our domain-specific language usage.



Transfer learning, NLP example

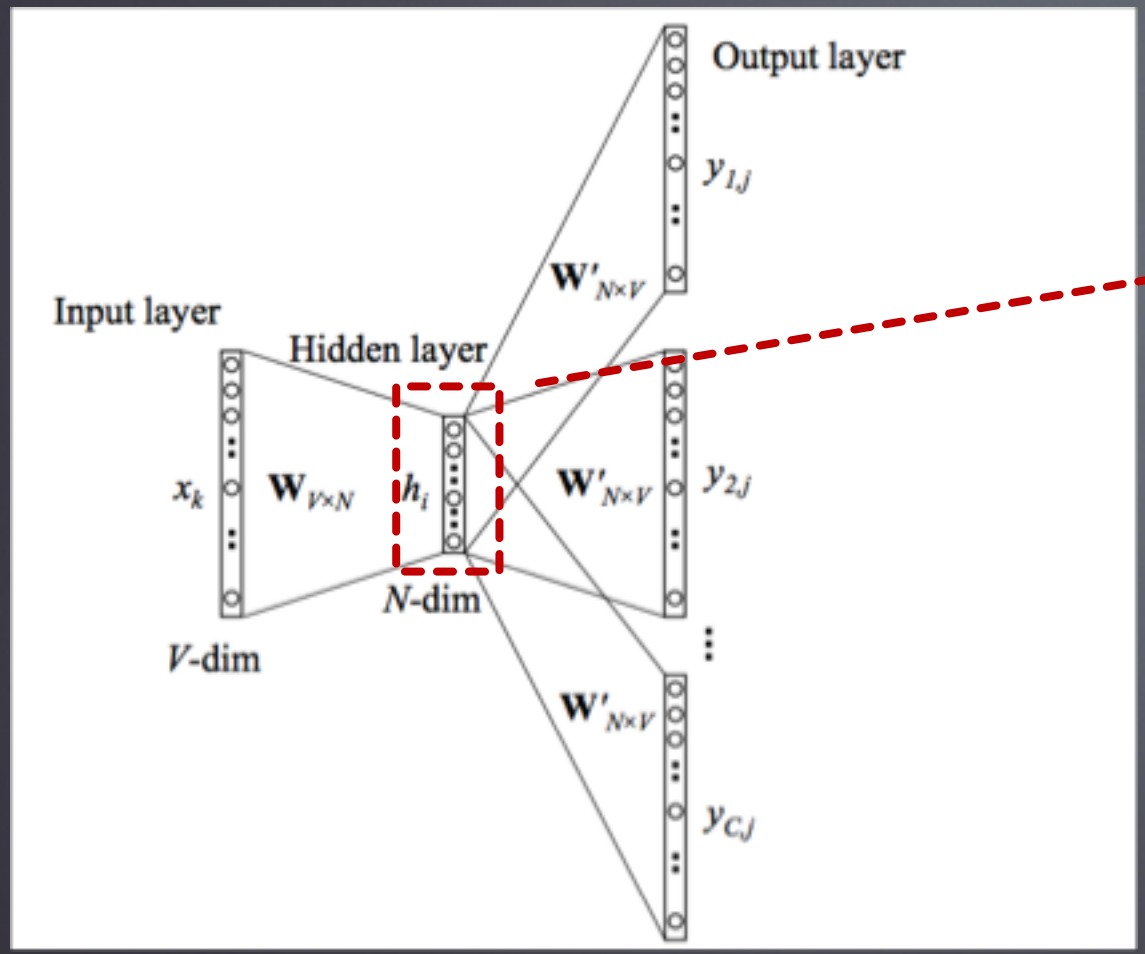
- After word embeddings are pre-trained once, they can be permanently maintained and reused across a wide variety of downstream tasks, leading to huge efficiency gains.
- In practice, it's very common to use publicly available pre-trained embeddings like Google's or Facebook's FastText.





But what does this look like concretely?

Pre-training: learn and extract hidden layer weights of a language modeling neural network



Transfer: represent task-specific documents as a sequence of vectors (the pre-trained weights)

Document: ”Oh hello world”

1.03	0.01	1.00
1.07	3.01	0.03
0.03	0.03	1.10
7.01	1.11	1.11
3.03	5.00	1.09

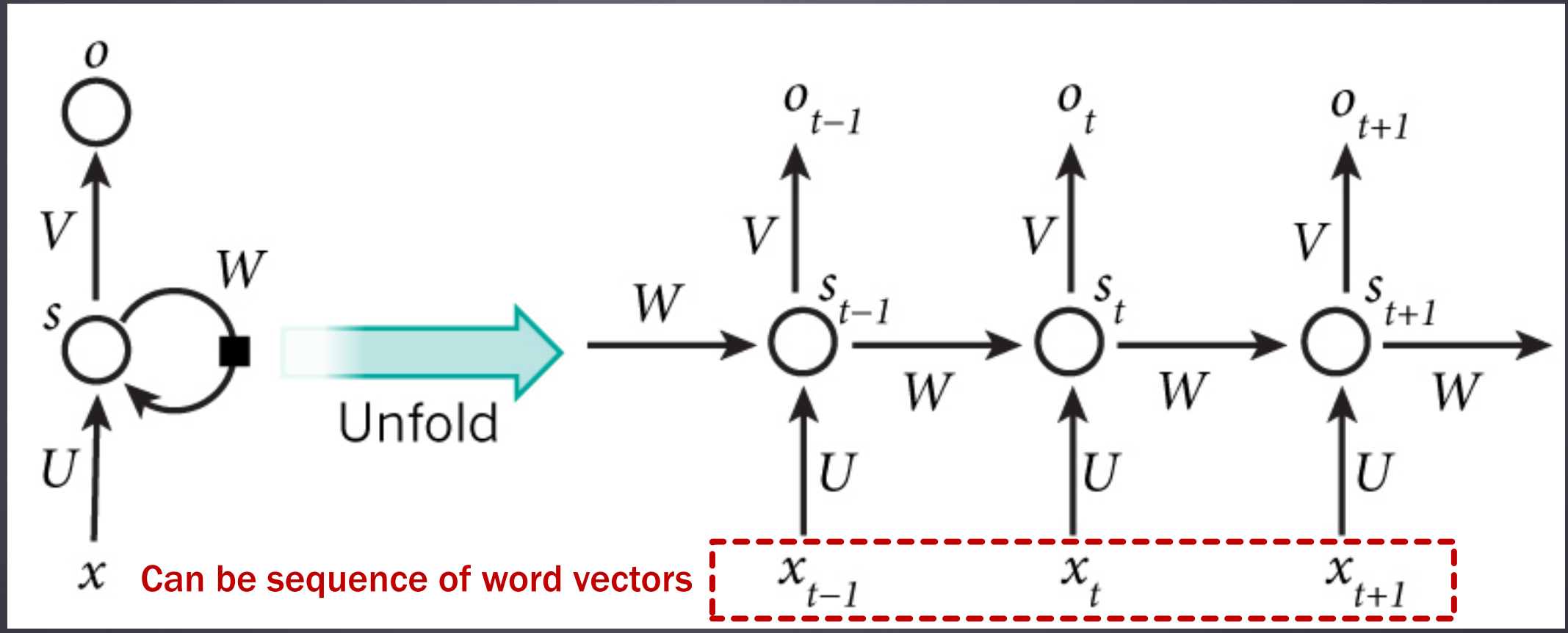
Sequence of N-dim, pretrained word vectors

Model taking sequences as input

Prediction task

How does a model take sequential input?

- Recurrent neural networks can process a sequence step by step for modeling!
- We'll learn more about these; they are very commonly used in text prediction tasks



What if I don't want to use a sequence?

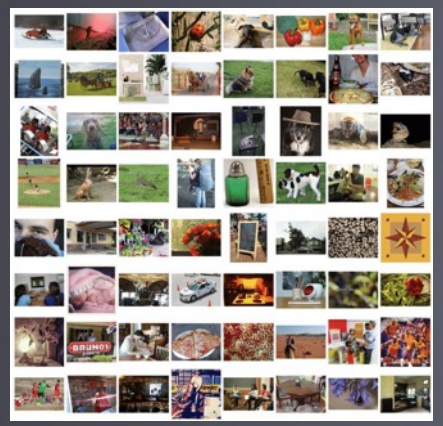
- Less commonly, there might be cases where you'd want to use pre-trained word vectors for a downstream task that doesn't involve a sequence-based model
- Example: unsupervised tasks such as clustering when syntactical similarities of documents are very relevant, since word vectors can capture syntactic relationships much more effectively than count-based vectors; supervised tasks where you desire a more traditional/simpler model
- Approach: summarize/flatten the sequence of word vectors to get a single document-level vector, e.g. simple average

Document: "Oh hello world"

Beyond text, a glance at image transfer learning

Transfer learning, image data example

- Another extremely common use of transfer learning is in computer vision. Neural nets are pre-trained on a massive image classification dataset (e.g. ImageNet with 1000+ object types), and pre-trained weights are transferred to tasks with smaller datasets because they're able to effectively capture generically useful image patterns.



Trained convolutional layers

Neural net with pre-trained convolutional base

Image classification

Image classification

Summary

- Transfer learning is a means of leveraging larger, general datasets in order to improve model performance and training time on tasks with a smaller dataset.
- In NLP, transfer learning typically takes the form of representing task-specific documents as a sequence of pre-trained word embeddings, which are then fed to sequence-based neural networks for predictive modeling.
- Transfer learning is also very prevalent in computer vision tasks.