Semi-supervised Learning of Compact Document Representations with Deep Networks

Marc'Aurelio Ranzato

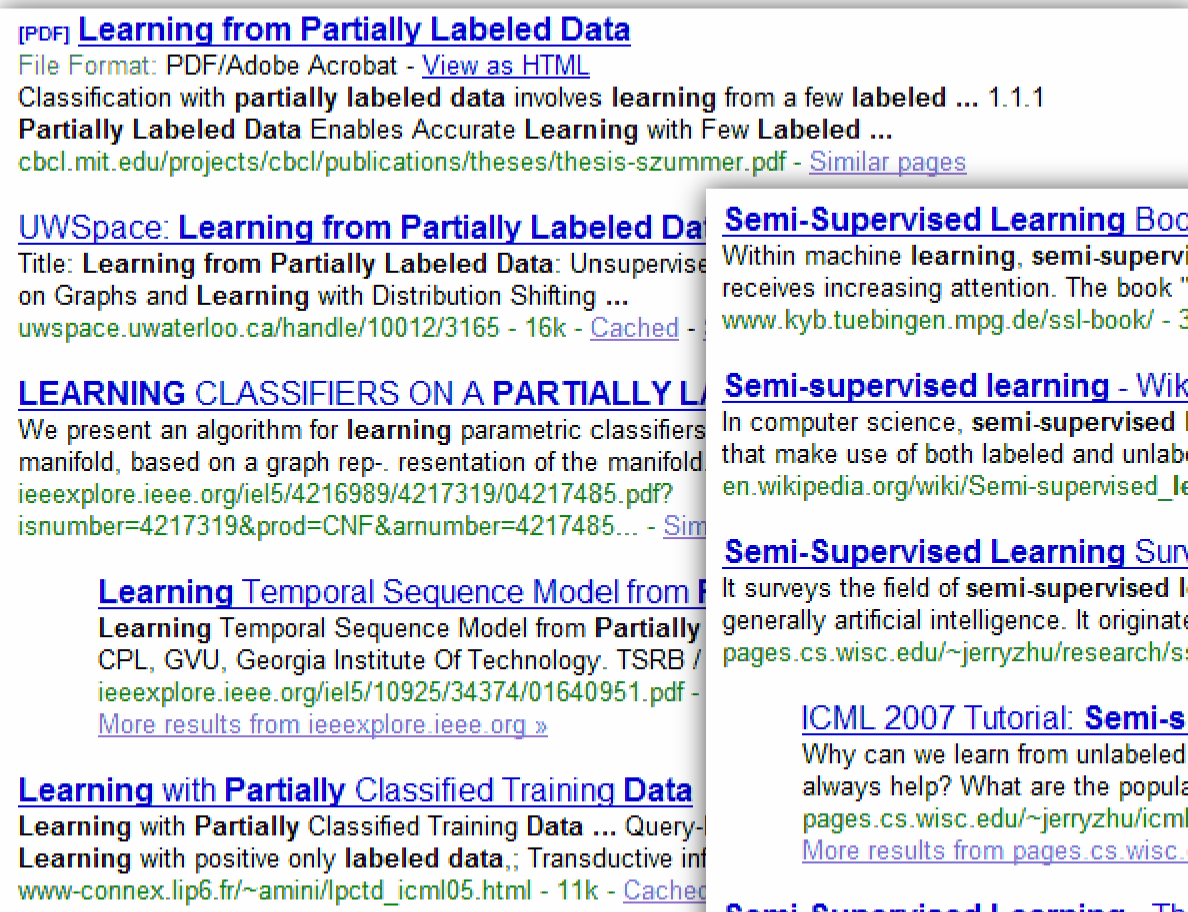
New York University
NY, USA

Martin Szummer

Microsoft Research
Cambridge, UK

Synonymous queries give different results

Query: learning from partially labeled data


[PDF] Learning from Partially Labeled Data
File Format: PDF/Adobe Acrobat - View as HTML
Classification with **partially labeled data** involves **learning** from a few **labeled** ... 1.1.1 **Partially Labeled Data** Enables Accurate **Learning** with Few **Labeled** ...
cbcl.mit.edu/projects/cbcl/publications/theses/thesis-szummer.pdf - Similar pages

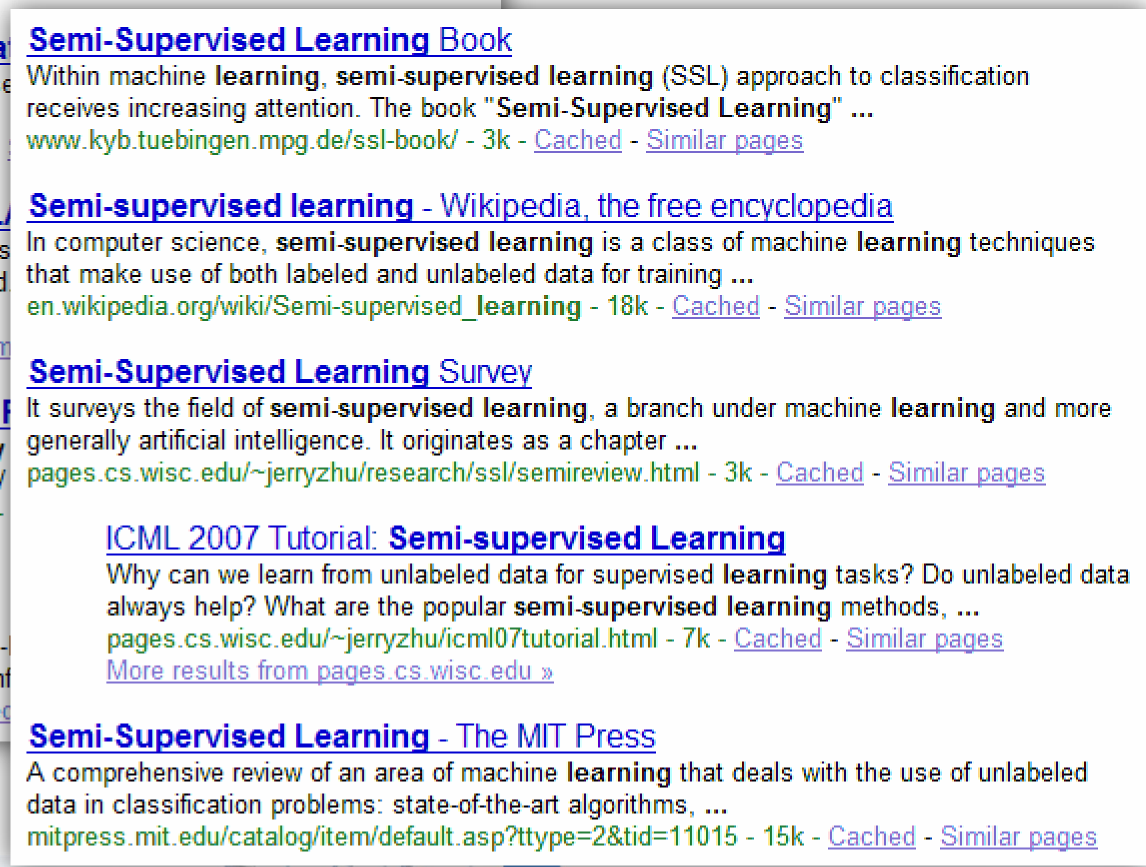
UWSpace: Learning from Partially Labeled Dat
Title: **Learning from Partially Labeled Data**: Unsupervise
on Graphs and **Learning** with Distribution Shifting ...
uwspace.uwaterloo.ca/handle/10012/3165 - 16k - Cached -

LEARNING CLASSIFIERS ON A PARTIALLY L
We present an algorithm for **learning** parametric classifiers
manifold, based on a graph rep-. resentation of the manifold
ieeexplore.ieee.org/iel5/4216989/4217319/04217485.pdf?
isnumber=4217319&prod=CNF&arnumber=4217485... - Sim

Learning Temporal Sequence Model from P
Learning Temporal Sequence Model from **Partially**
CPL, GVU, Georgia Institute Of Technology. TSRB /
ieeexplore.ieee.org/iel5/10925/34374/01640951.pdf -
More results from ieeexplore.ieee.org »

Learning with Partially Classified Training Data
Learning with **Partially** Classified Training **Data** ... Query-
Learning with positive only **labeled data**,; Transductive inf
www-connex.lip6.fr/~amini/lpctd_icml05.html - 11k - Cached

Query: semi-supervised learning


Semi-Supervised Learning Book
Within machine **learning**, **semi-supervised learning** (SSL) approach to classification receives increasing attention. The book "**Semi-Supervised Learning**" ...
www.kyb.tuebingen.mpg.de/ssl-book/ - 3k - Cached - Similar pages

Semi-supervised learning - Wikipedia, the free encyclopedia
In computer science, **semi-supervised learning** is a class of machine **learning** techniques that make use of both labeled and unlabeled data for training ...
en.wikipedia.org/wiki/Semi-supervised_learning - 18k - Cached - Similar pages

Semi-Supervised Learning Survey
It surveys the field of **semi-supervised learning**, a branch under machine **learning** and more generally artificial intelligence. It originates as a chapter ...
pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html - 3k - Cached - Similar pages

ICML 2007 Tutorial: Semi-supervised Learning
Why can we learn from unlabeled data for supervised **learning** tasks? Do unlabeled data always help? What are the popular **semi-supervised learning** methods, ...
pages.cs.wisc.edu/~jerryzhu/icml07tutorial.html - 7k - Cached - Similar pages
More results from pages.cs.wisc.edu »

Semi-Supervised Learning - The MIT Press
A comprehensive review of an area of machine **learning** that deals with the use of unlabeled data in classification problems: state-of-the-art algorithms, ...
mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11015 - 15k - Cached - Similar pages

Better Representations

Goal: Capture document or query topics to handle synonymy and semantics, while remaining

- Compact
  - Forward index is stored in RAM
  - 40 billion index size : each representation bit costs 5 Gb of RAM

OR:

- Sparse
  - Can then be fit in inverted index
  - Example: document represented by its words

Better Representations

Goal: Capture document or query topics to handle synonymy and semantics, while remaining

❖ Compact

- … index is stored in RAM
- 40 … ts 5 Gb of RAM

Can deep networks fit the bill?

OR:

❖ Sparse

- Can then be fit in inverted index
- Example: document represented by its words

Distributed Representations in information retrieval

- ❖ Exponential Family Harmoniums [Welling 2004]
- ❖ Rate Adapting Poisson Model [Gehler et al 2006]

Shallow

- ❖ Neural probabilistic language model [Bengio, et al 2003]
- ❖ Deep Belief Nets
  - Semantic Hashing [Salakhutdinov & Hinton 2007]

    Binary code, achieved by adding noise during training

Deep

Computational Efficiency

- Neural networks computational cost for train and test:

  linear in number of layers (depth)

  quadratic in #units in adjacent layers (width)

- Deep & narrow often cheaper than shallow & wide

Exploit both labeled and unlabeled documents

❖ Unsupervised pretraining, then supervised finetuning

only unsupervised only supervised

but for a given task, how do we ensure pretraining gets us to the right region in space?

Inject label information early:

❖ Semi-supervised training of the bottleneck layer

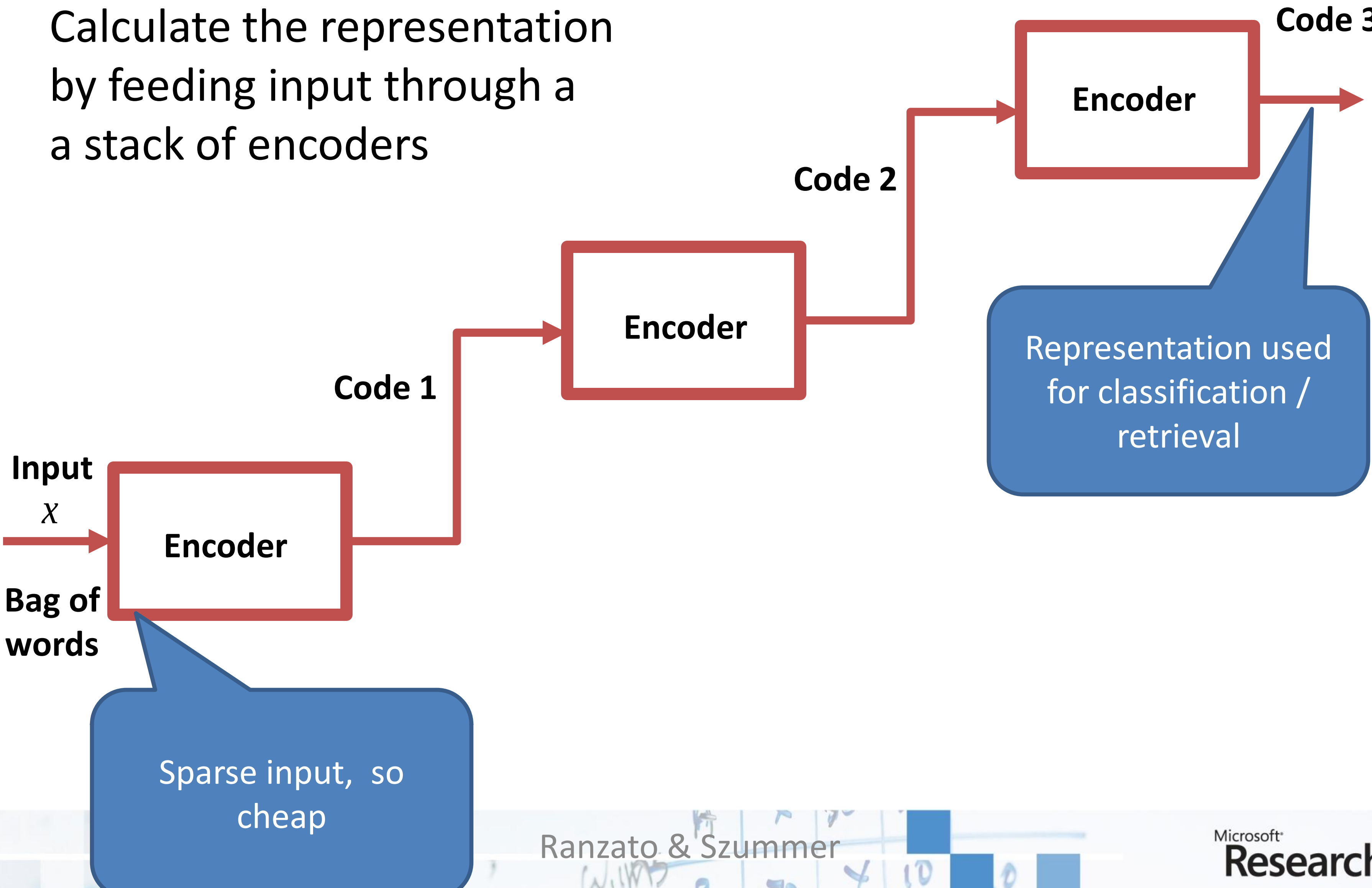
❖ Semi-supervised training of all layers

Outline

- Learning Representations of Text Documents
- Model and Learning Algorithm
- Experiments
  - Visualization
  - Classification
  - Retrieval

Our model: Deep Semi-Supervised Encoder

Calculate the representation by feeding input through a a stack of encoders

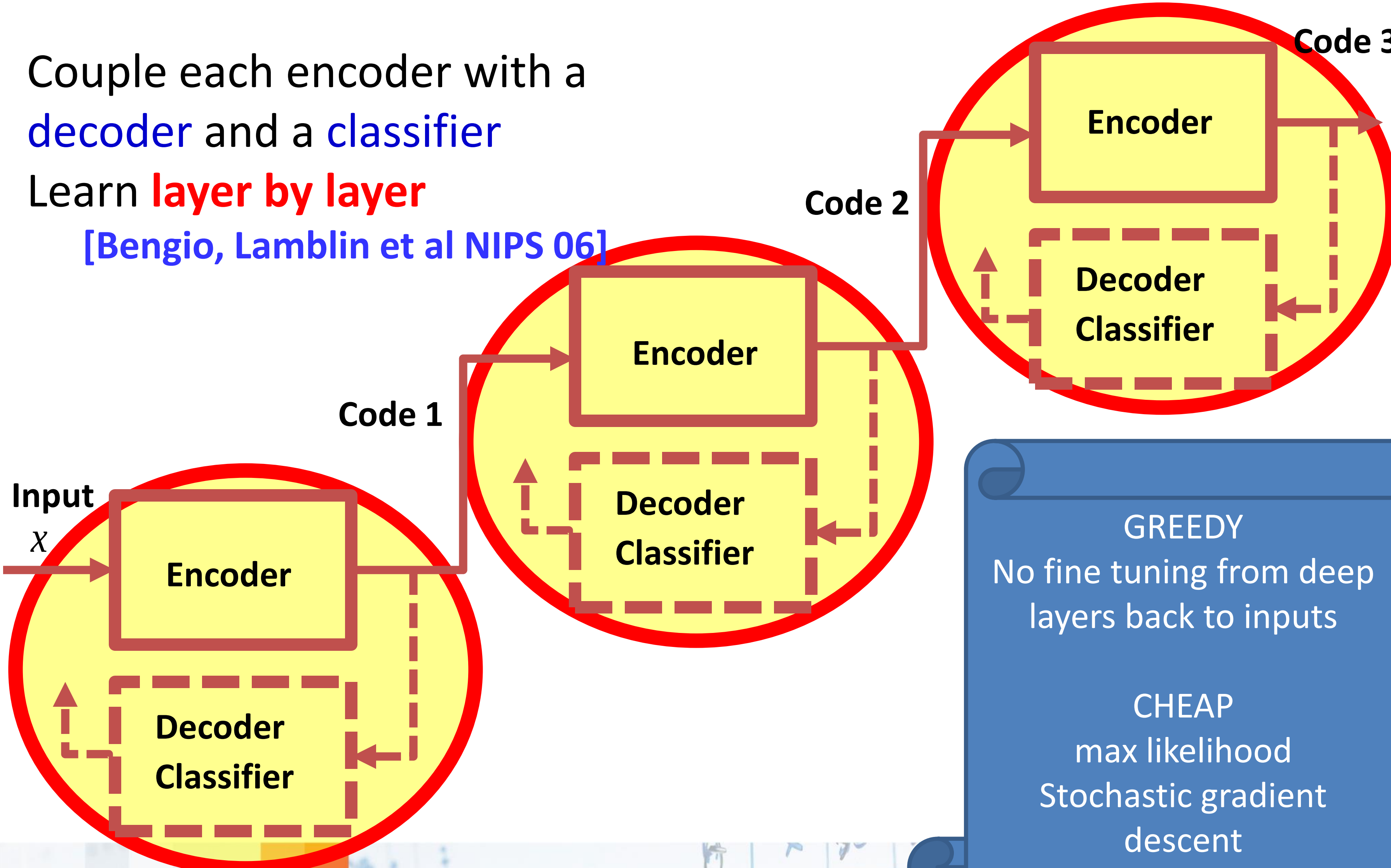


Semi-supervised Greedy Learning

Couple each encoder with a decoder and a classifier

Learn **layer by layer**

[Bengio, Lamblin et al NIPS 06]



GREEDY
No fine tuning from deep layers back to inputs

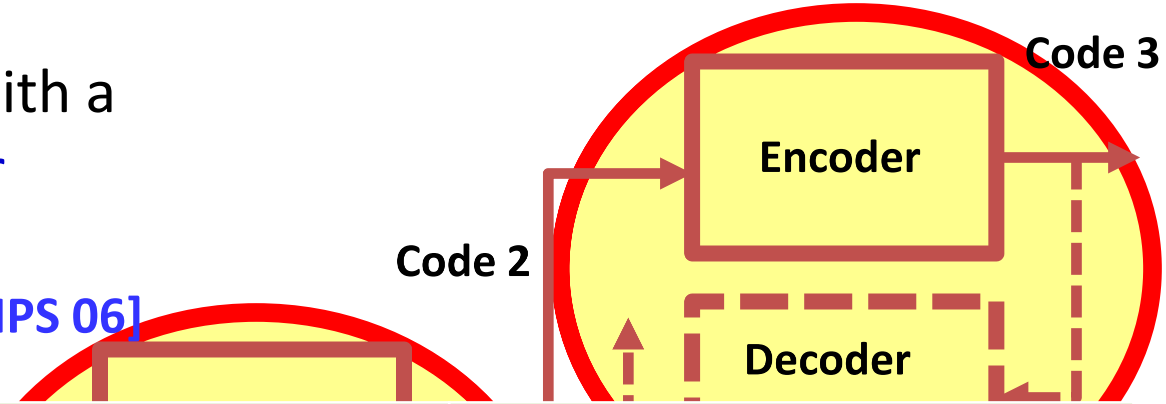
CHEAP
max likelihood
Stochastic gradient descent

Semi-supervised Greedy Learning

Couple each encoder with a decoder and a classifier

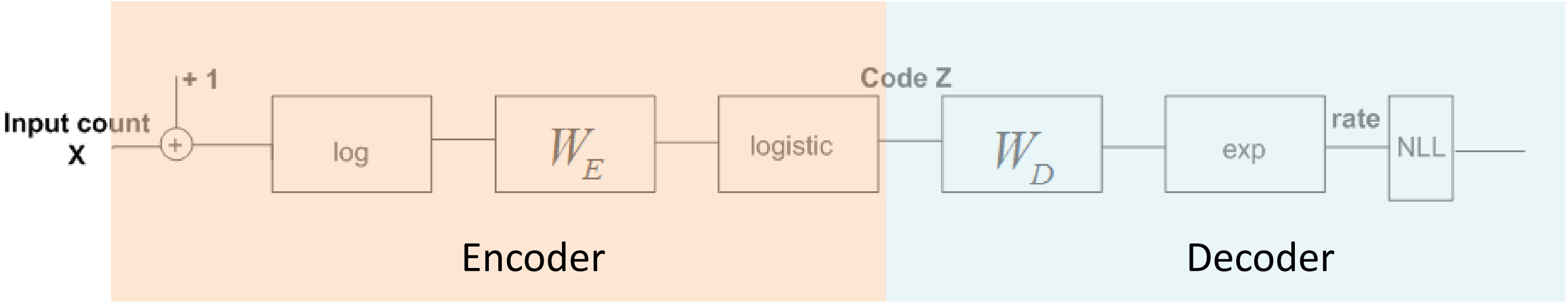
Learn **layer by layer**

[Bengio, Lamblin et al NIPS 06]



	Semi-supervised Deep Auto Encoder	Deep Belief Net
Training	semi-supervised training layer-by-layer no fine-tuning of whole net	1) unsupervised pre-training layer-by-layer 2) supervised fine-tune whole net
Training objectives	single objective: combine reconstruction and classification likelihoods	1) pretraining: model P(x) 2) fine-tuning: model P(y|x)
	deterministic (backprop), 4 epochs	1) pretraining: sampling-based 2) fine-tuning: backprop >> 4 epochs

Model: 1st stage



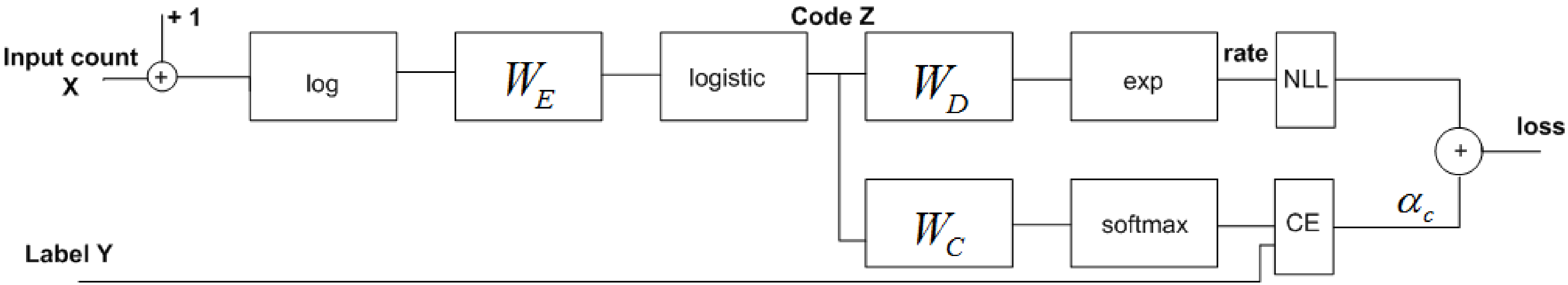
- Model the input count vector with a conditional Poisson distrib.

$$\text{Decoder}: \quad x \sim Poiss(\lambda), \quad \lambda = \beta \exp(W_D z + b_D)$$

- The encoder and the decoder mirror each other

$$\text{Encoder}: \quad z = \text{logistic}\,(W_E \log(x+1) + b_E)$$

Model: 1st stage



- Model the input count vector with a conditional Poisson distrib.

$$\text{Decoder}: \quad x \sim Poiss(\lambda), \quad \lambda = \beta \exp(W_D z + b_D)$$

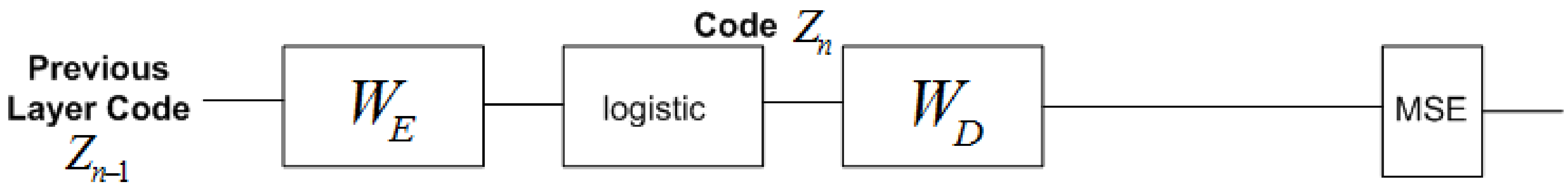
- The encoder and the decoder mirror each other

$$\text{Encoder}: \quad z = \text{logistic}\,(W_E \log(x+1) + b_E)$$

- Objective: reconstruct the input AND predict the label (if available)

$$L = E_R + \alpha_C E_C$$

Model: higher stages



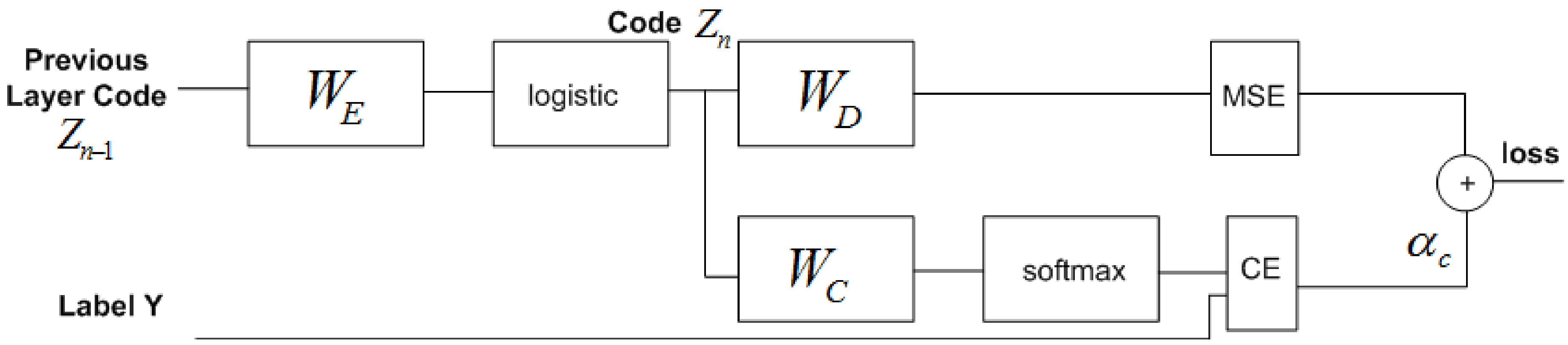
- Model the input vector with a conditional Gaussian distribution

$$x \sim N(W_D Z + b_D, \sigma)$$

- The encoder and the decoder mirror each other

$$Z = \text{logistic}\,(W_E X + b_E)$$

Model: higher stages



- Model the input vector with a conditional Gaussian distribution

$$x \sim N(W_D Z + b_D, \sigma)$$

- The encoder and the decoder mirror each other

$$Z = \text{logistic}\,(W_E X + b_E)$$

- The code has to be able to reconstruct the input as well as to predict the label, if available.

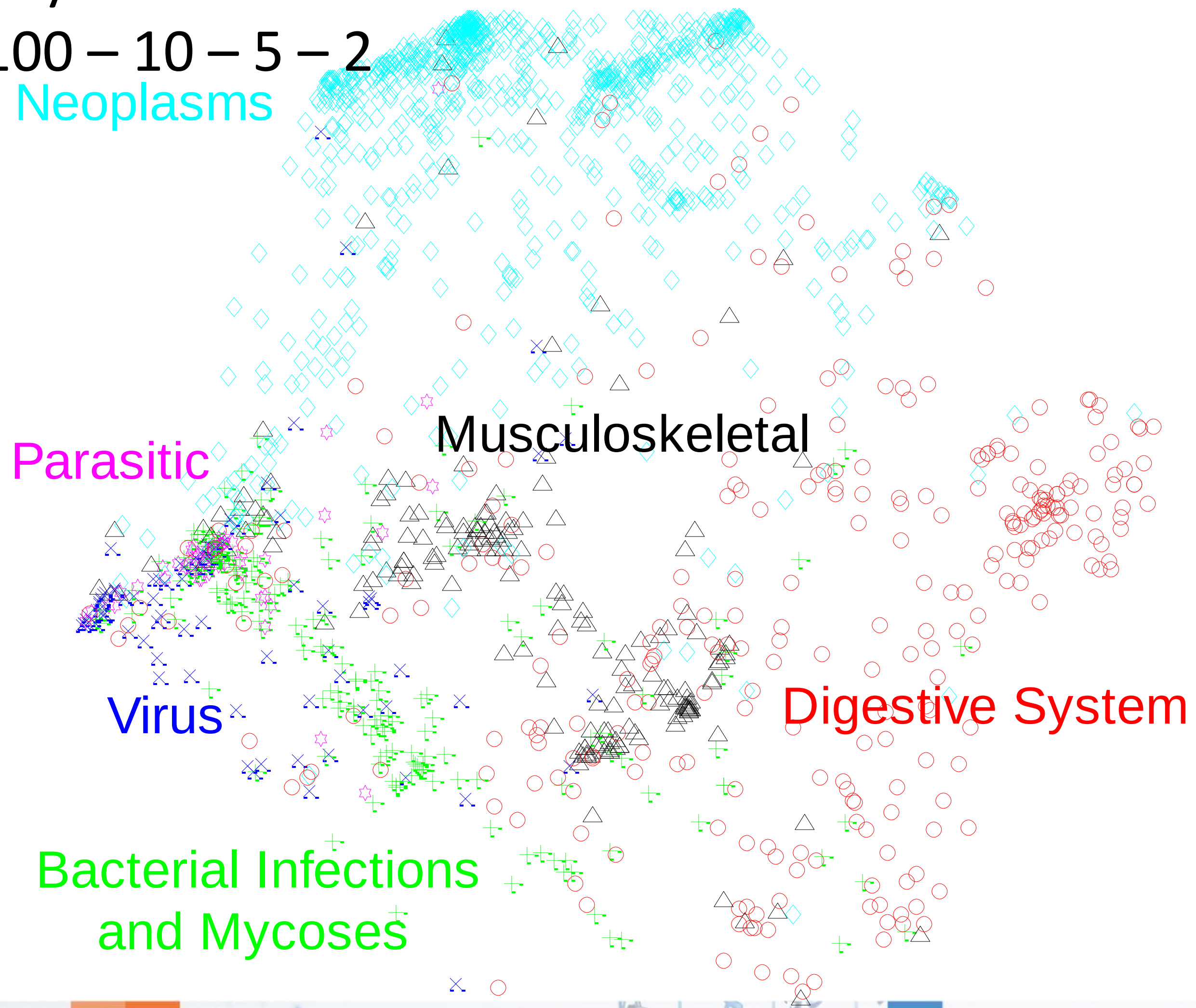
$$L = E_R + \alpha_C E_C$$

- Parameter learning: min L w.r.t. the parameters by stochastic gradient descent

Visualization of codes on Ohsumed corpus

4 hidden layers

30689 – 100 – 10 – 5 – 2

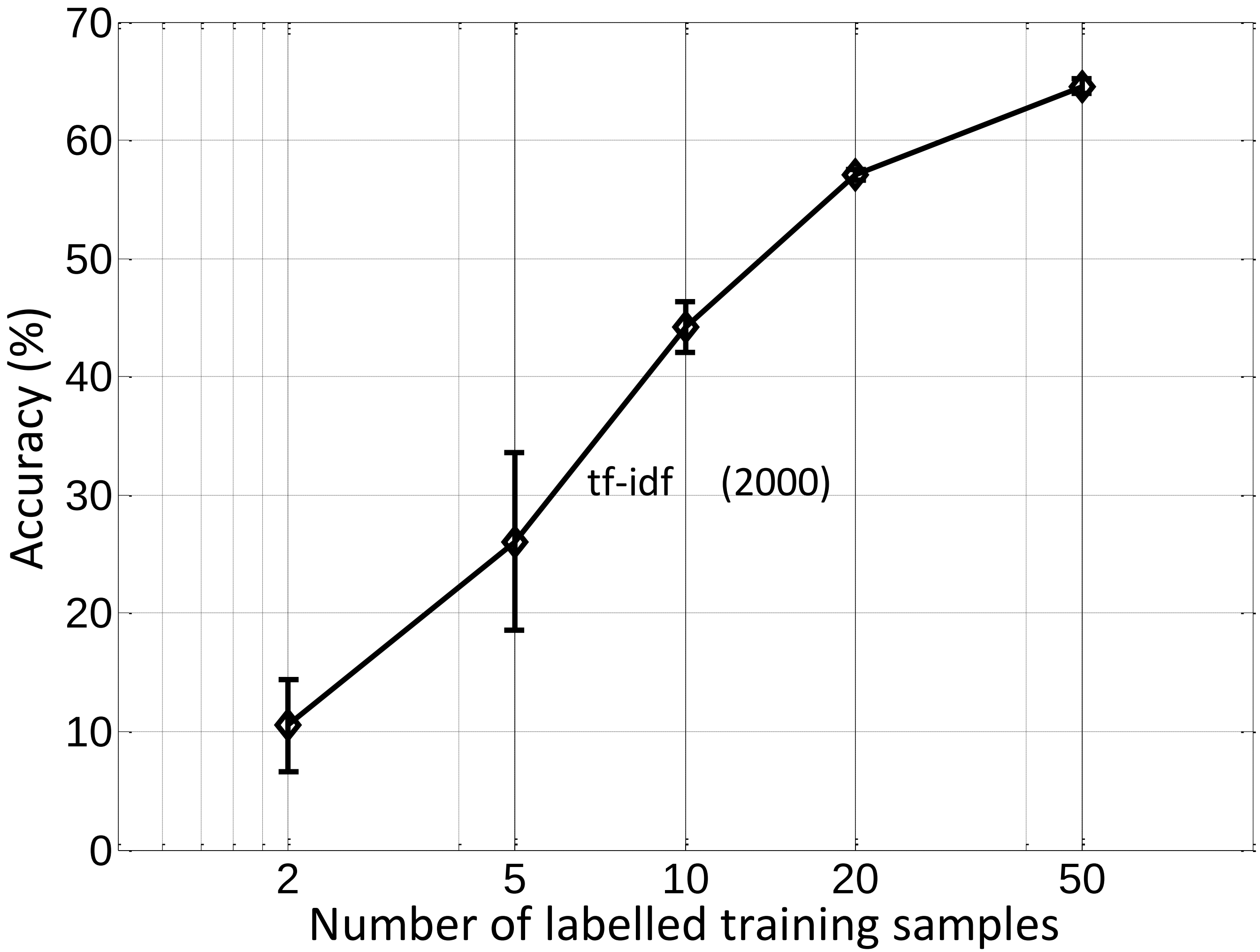


Word neighbors in code space

Neighboring word stems to a given word in the 7-dimensional feature space to which documents of Reuters are mapped after learning. (2000 – 200 – 100 – 7)

Word Stems	Neighboring Word Stems
livestock	beef, meat, pork, cattle
port	ship, vessel, freight
plantat	coffee, cocoa, rubber, palm
barrel	oil, crude, opec, refineri
lend	rate, debt, bond, downgrad

Classification of partially labeled documents



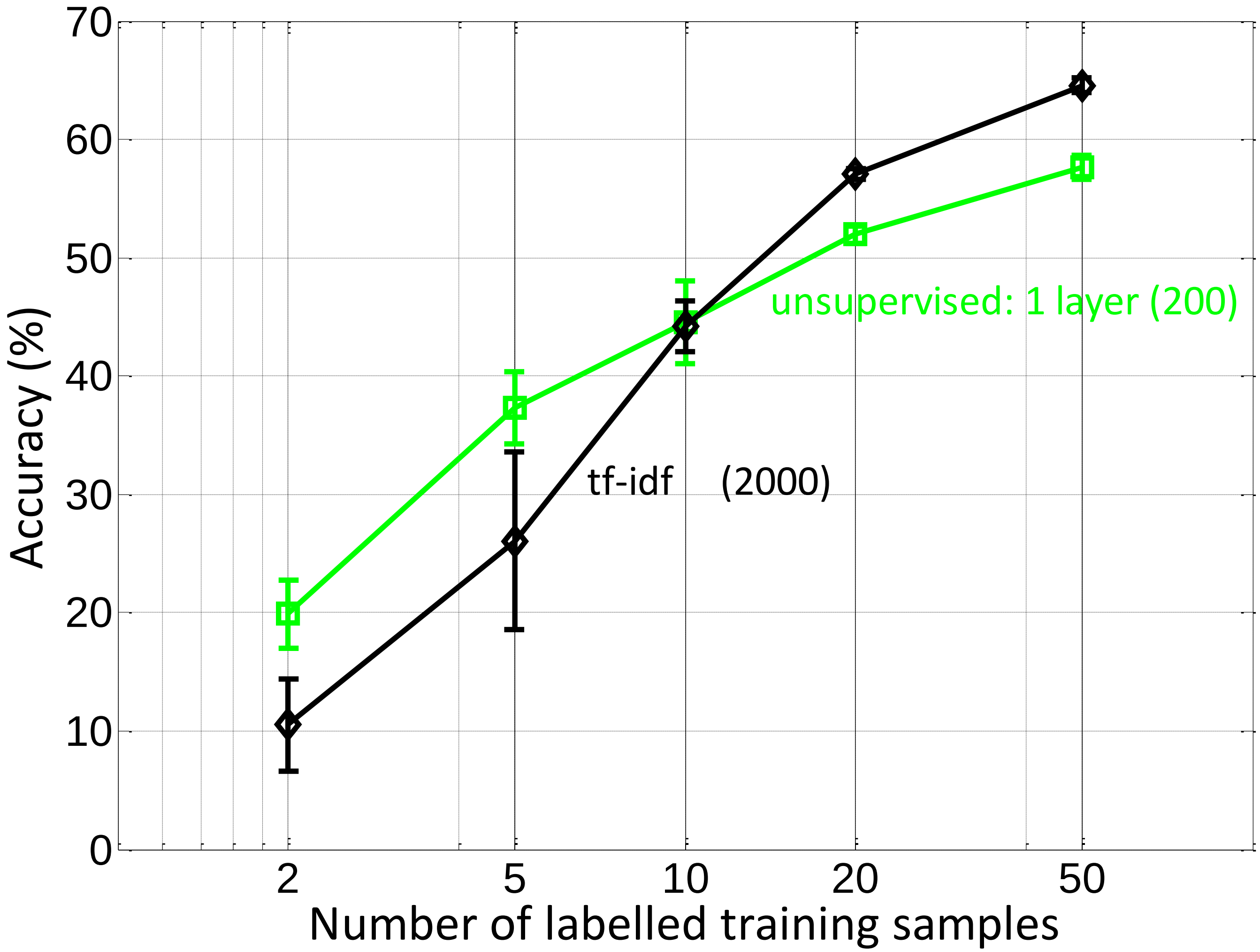
- 20 newsgroups data
- Gaussian SVM is applied to all representations

  (but we could reuse the top-level classifier instead of training SVM)

Labelled samples used for

1) training SVM
2) in semi-supervised representations

Classification of partially labeled documents



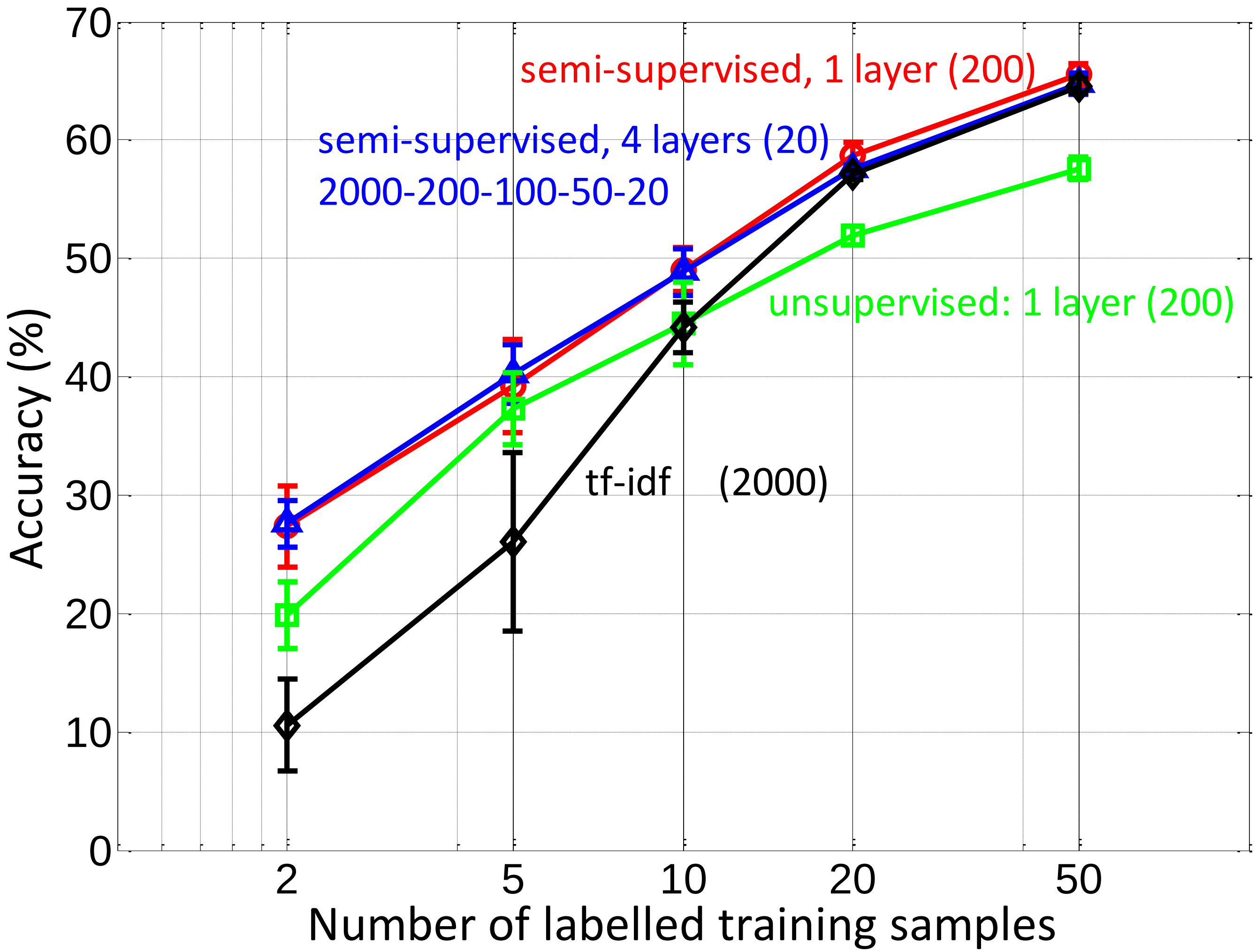
- 20 newsgroups data
- Gaussian SVM is applied to all representations

  (but we could reuse the top-level classifier instead of training SVM)

Labelled samples used for

1) training SVM
2) in semi-supervised representations

Classification of partially labeled documents



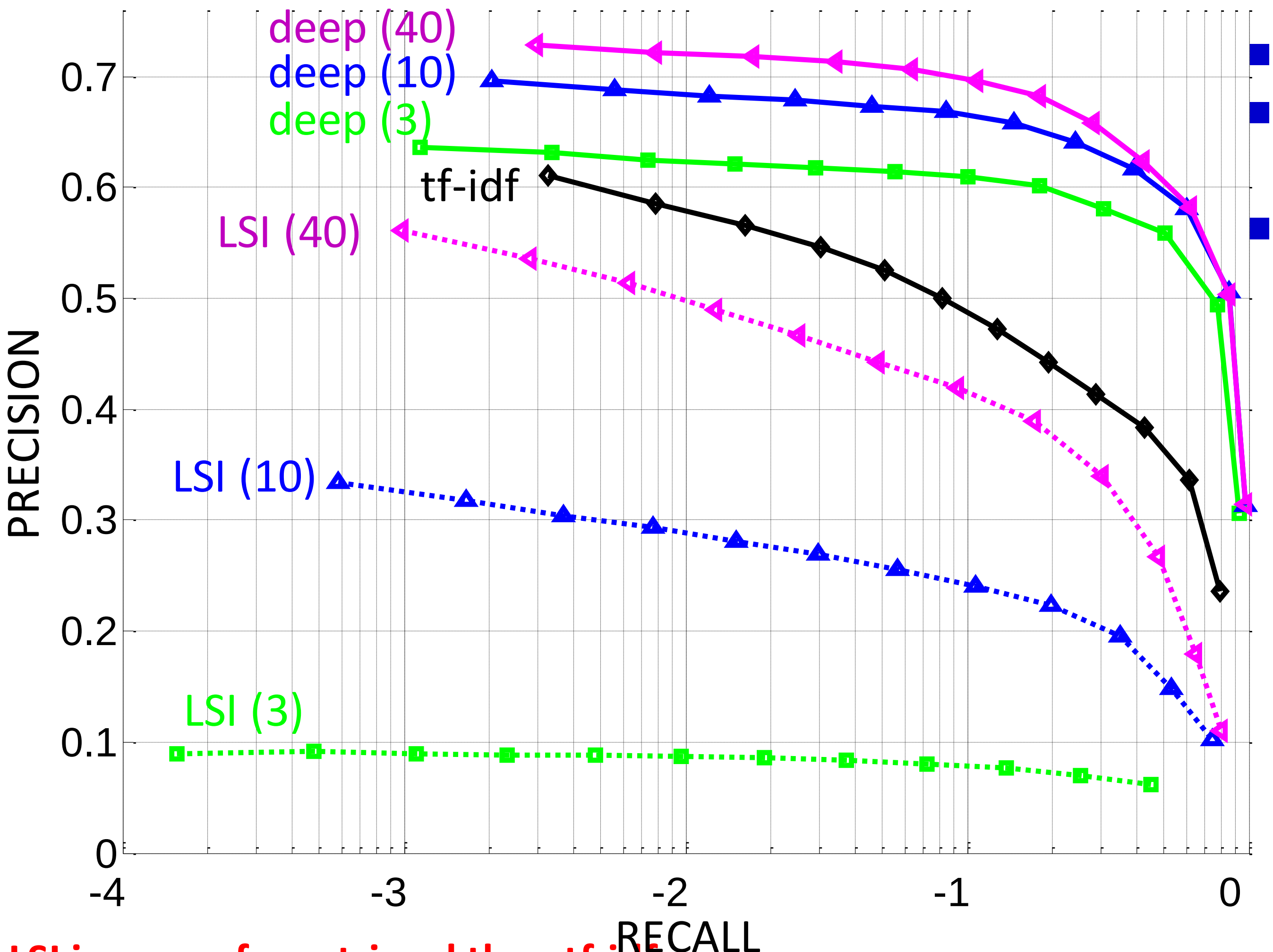
- 20 newsgroups data
- Gaussian SVM is applied to all representations

(but we could reuse the top-level classifier instead of training SVM)

Labelled samples used for

1) training SVM
2) in semi-supervised representations

Deep vs Linear (LSI & TF-IDF)

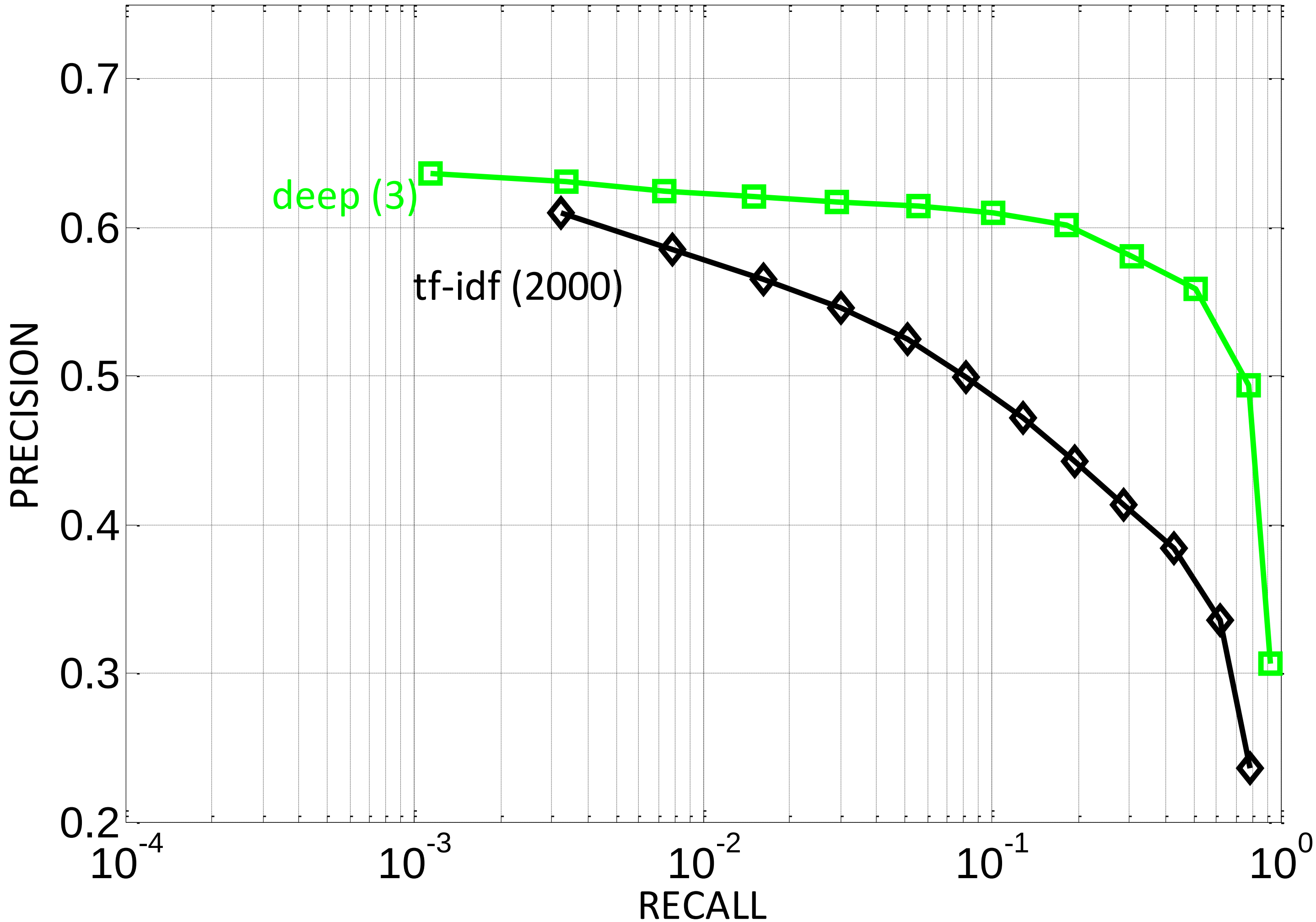


- Reuters-21578
- deep model:
  100-{40,10}-{2,3}
- The deep nonlinear model greatly outperforms a (shallow) linear model.

LSI is worse for retrieval than tf-idf

A deep nonlinear model is better than a linear one!

Deep vs Shallow



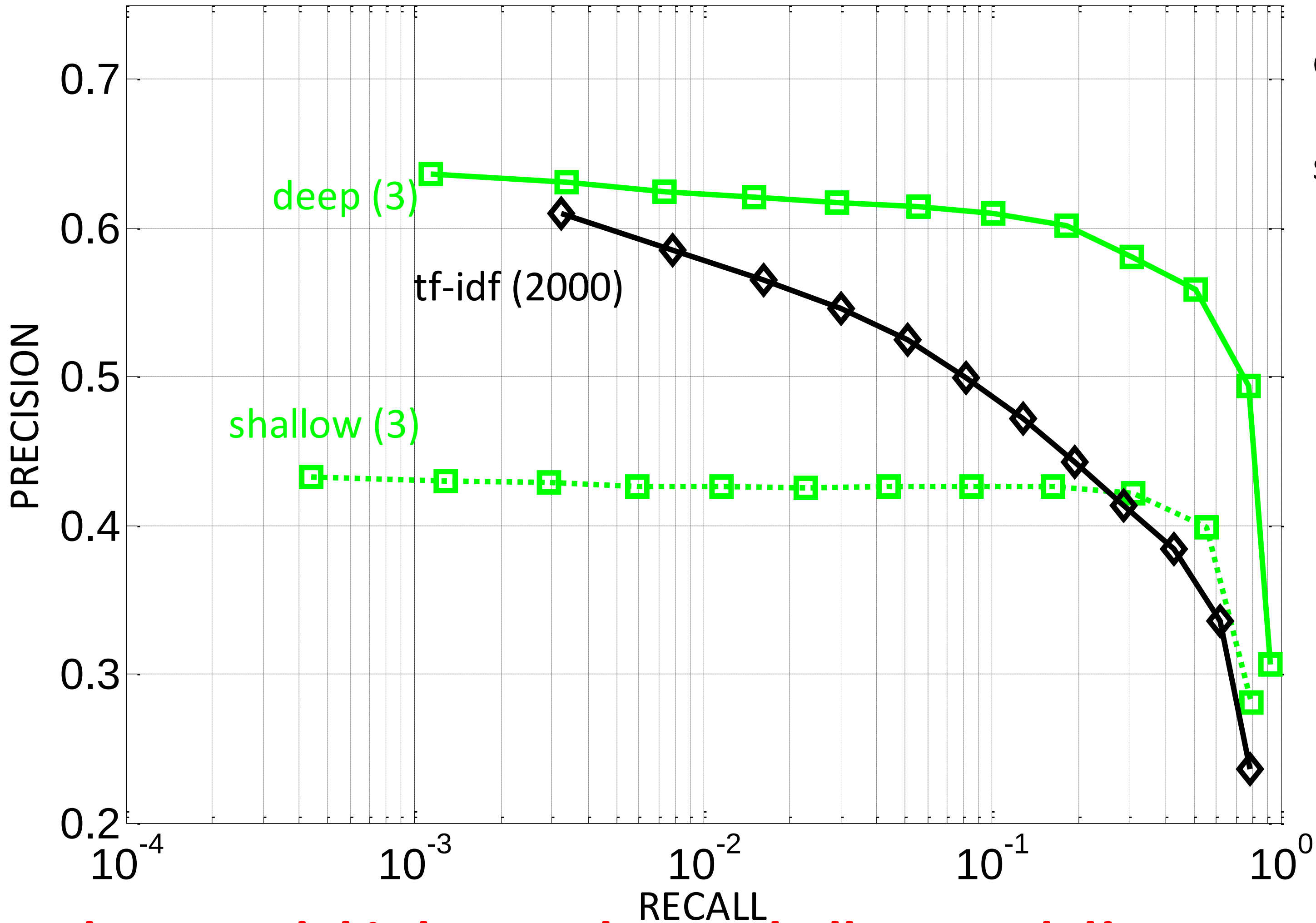
deep: 4 layers

shallow: 1 layer

Reuters-21578

Microsoft Research

Deep vs Shallow



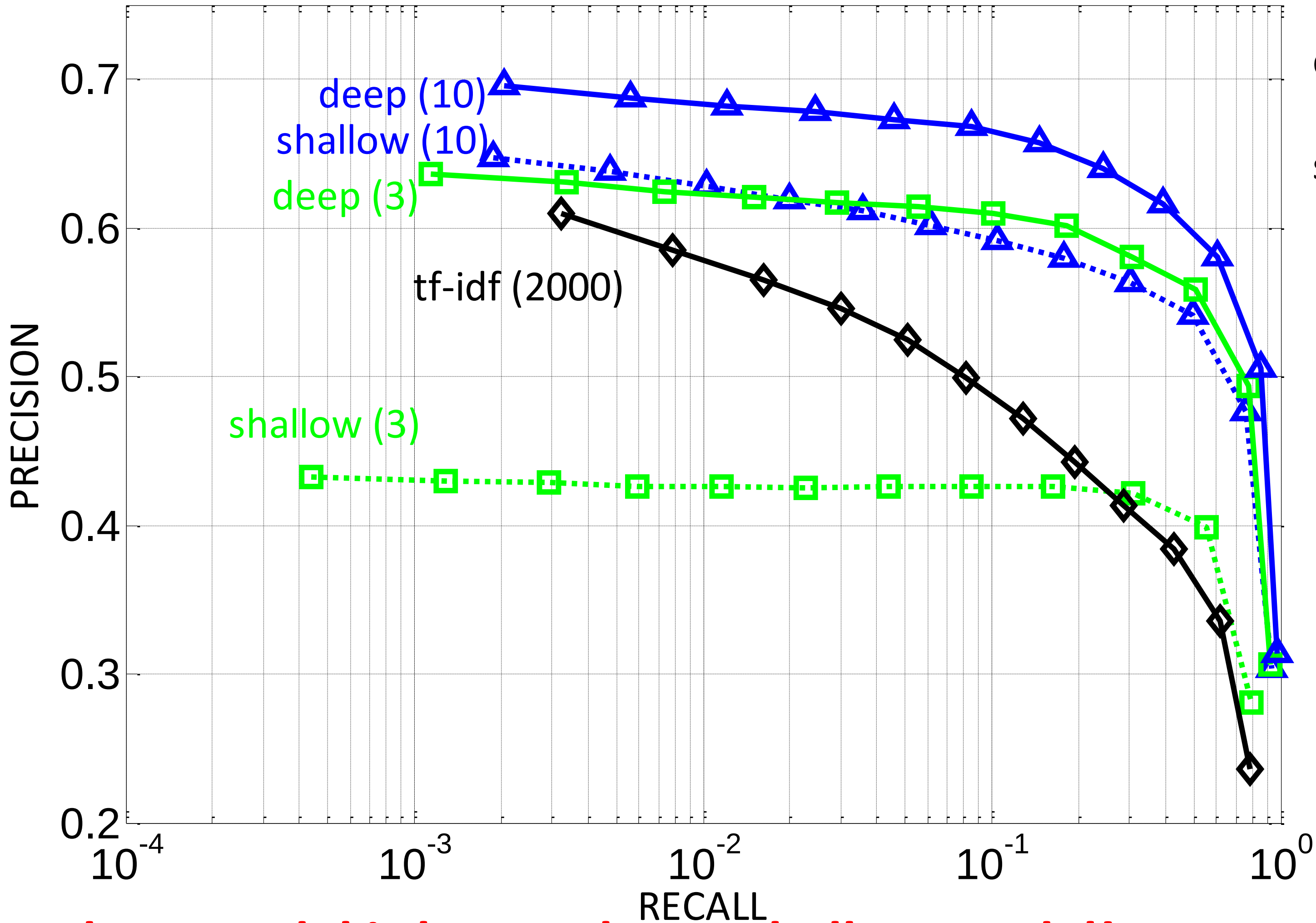
deep: 4 layers

shallow: 1 layer

Reuters-21578

A deep model is better than a shallow model!

Deep vs Shallow



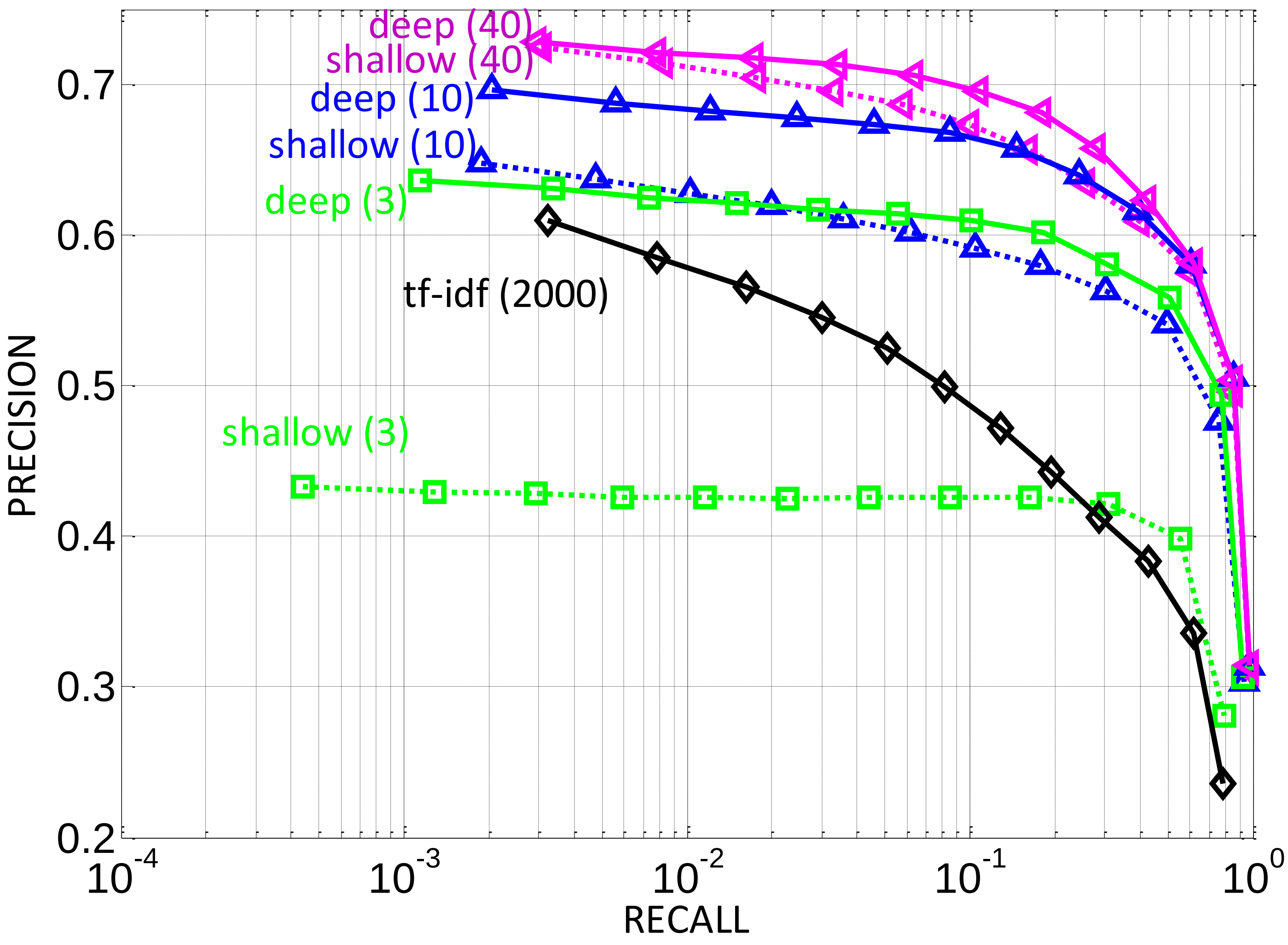
deep: 4 layers

shallow: 1 layer

Reuters-21578

A deep model is better than a shallow model!

Deep vs Shallow

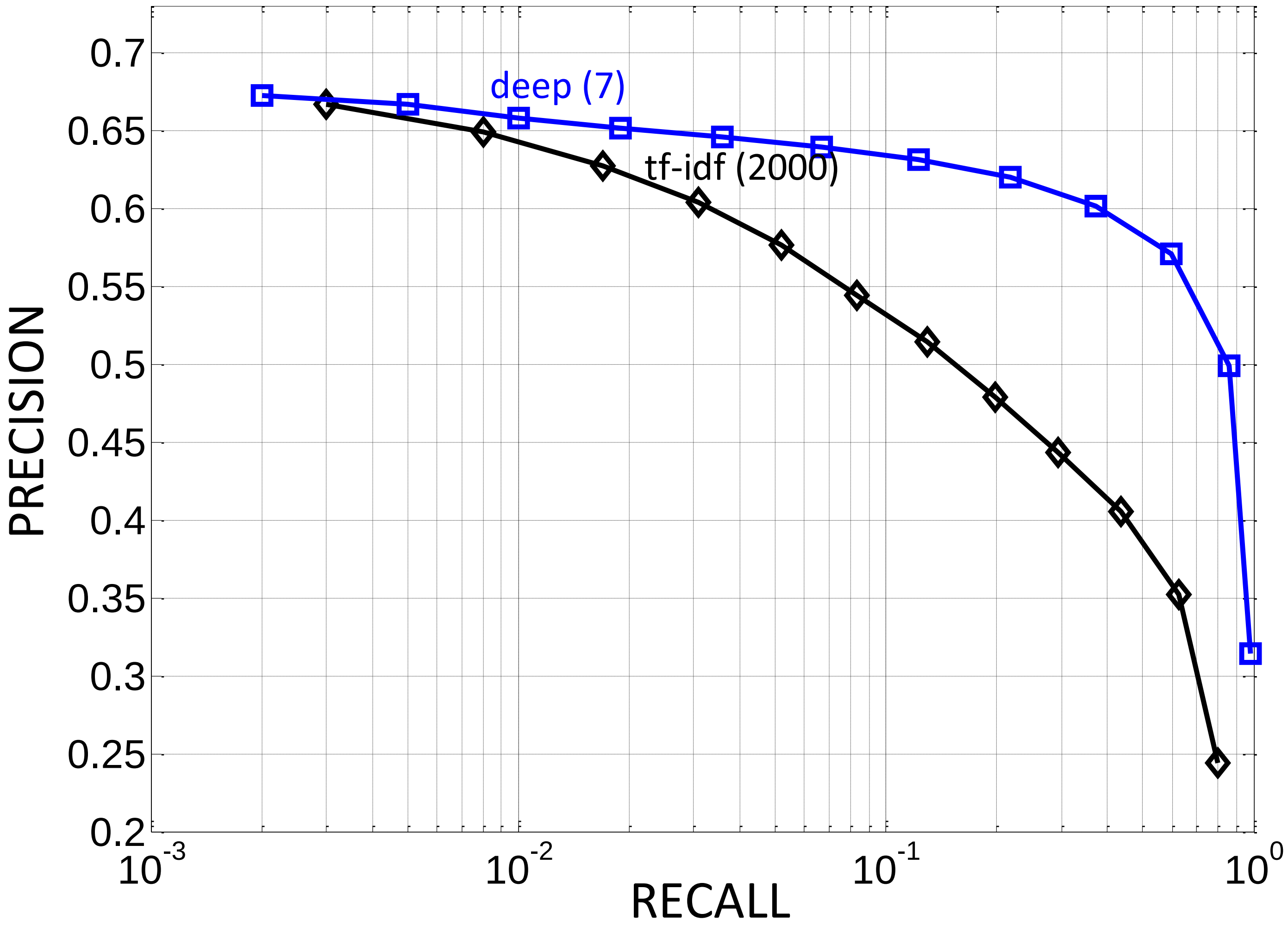


deep: 4 layers

shallow: 1 layer

Reuters-21578

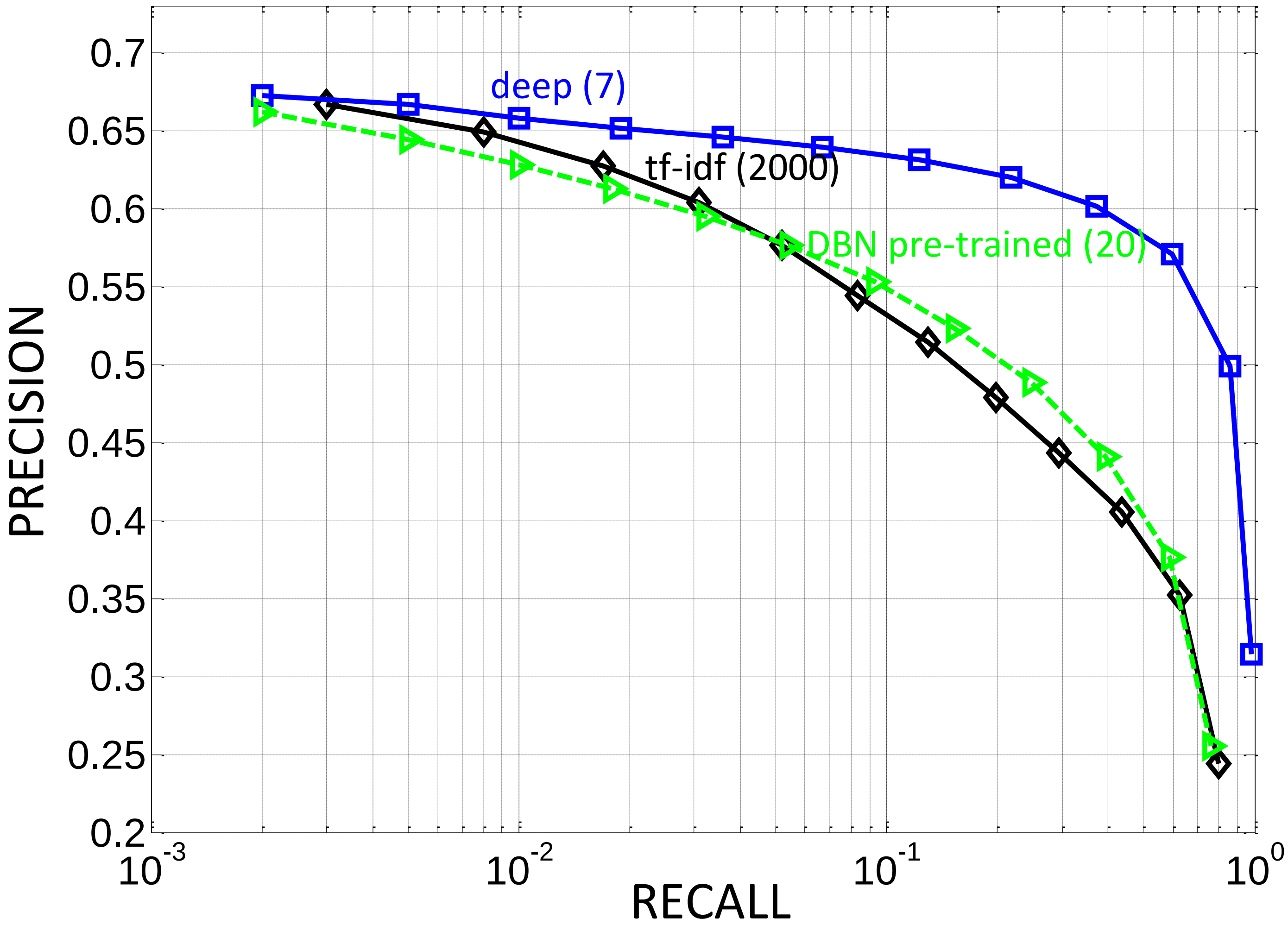
Deep vs DBN vs SESM



deep (7):
2000-200-100-7

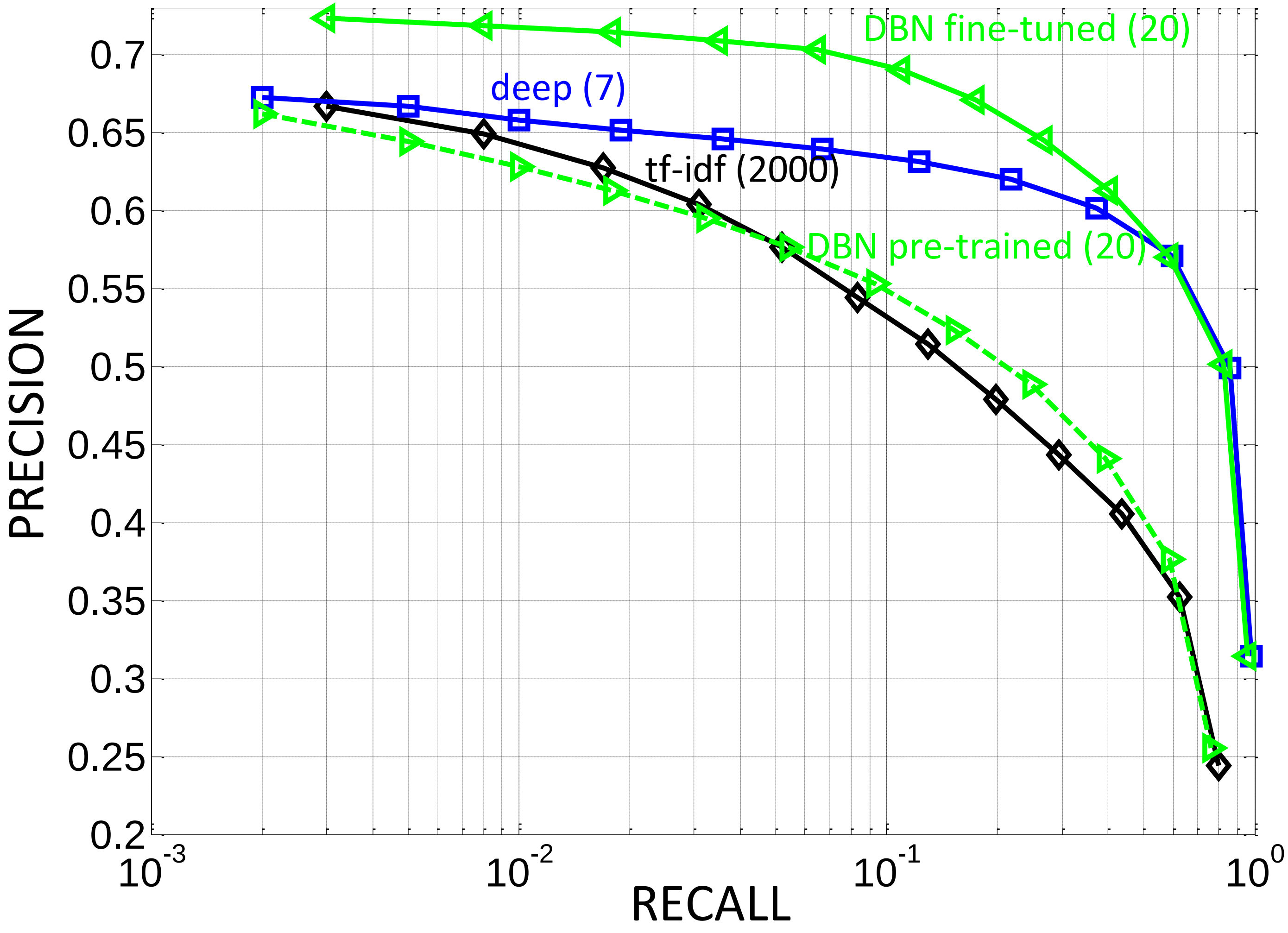
Reuters-21578

Deep vs DBN vs SESM

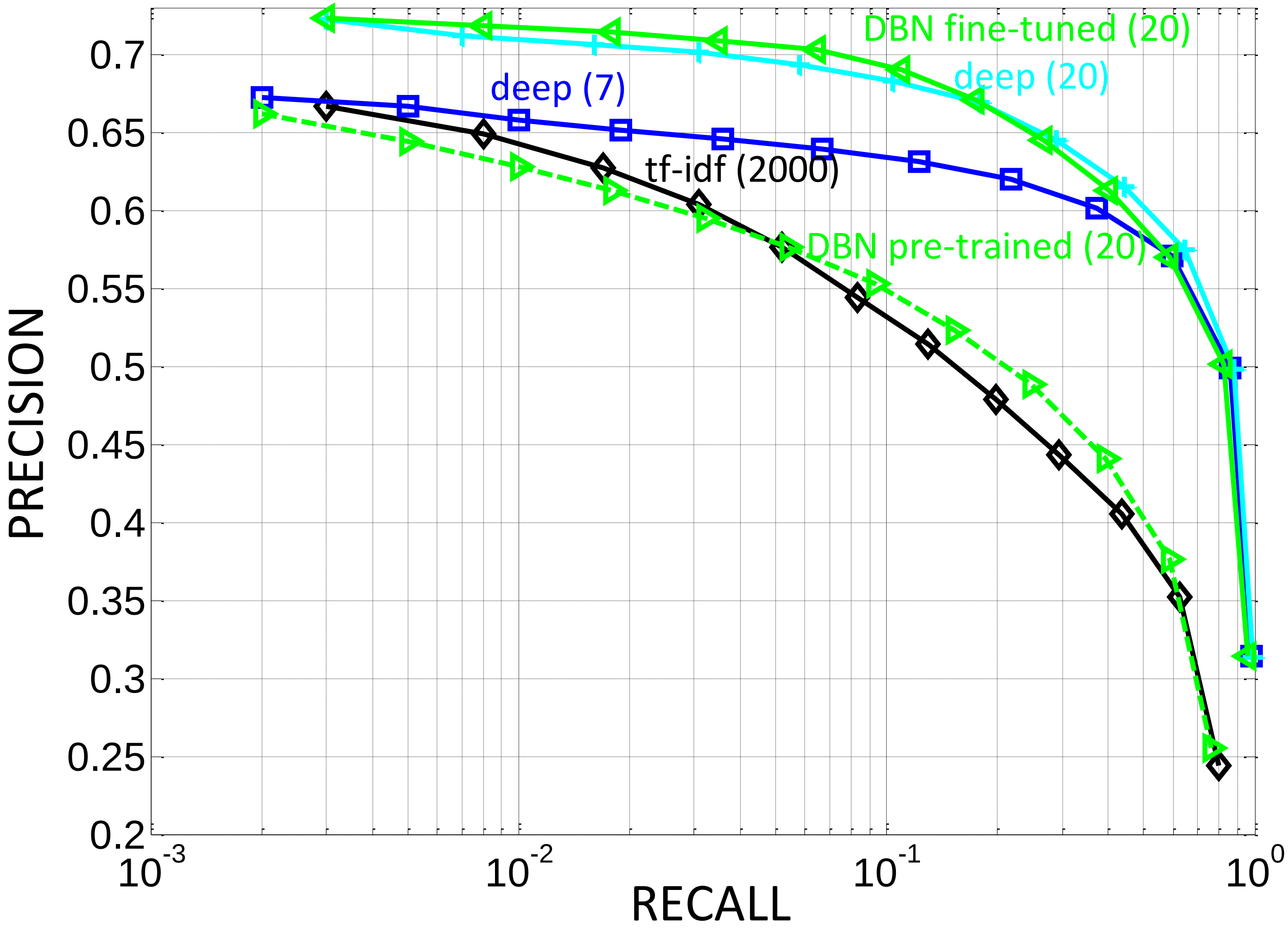


deep (7):
2000-200-100-7

Deep vs DBN vs SESM

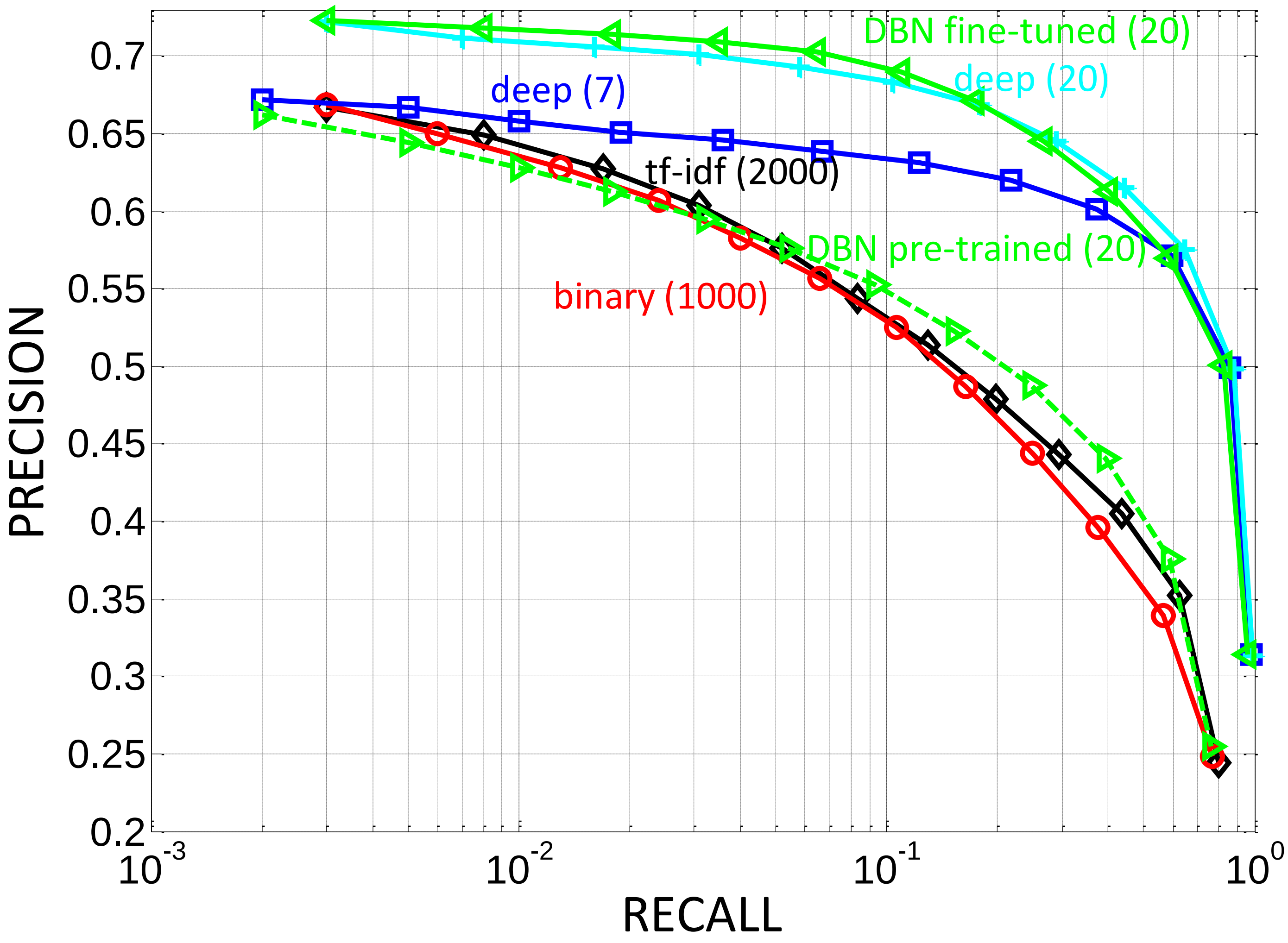


Deep vs DBN vs SESM



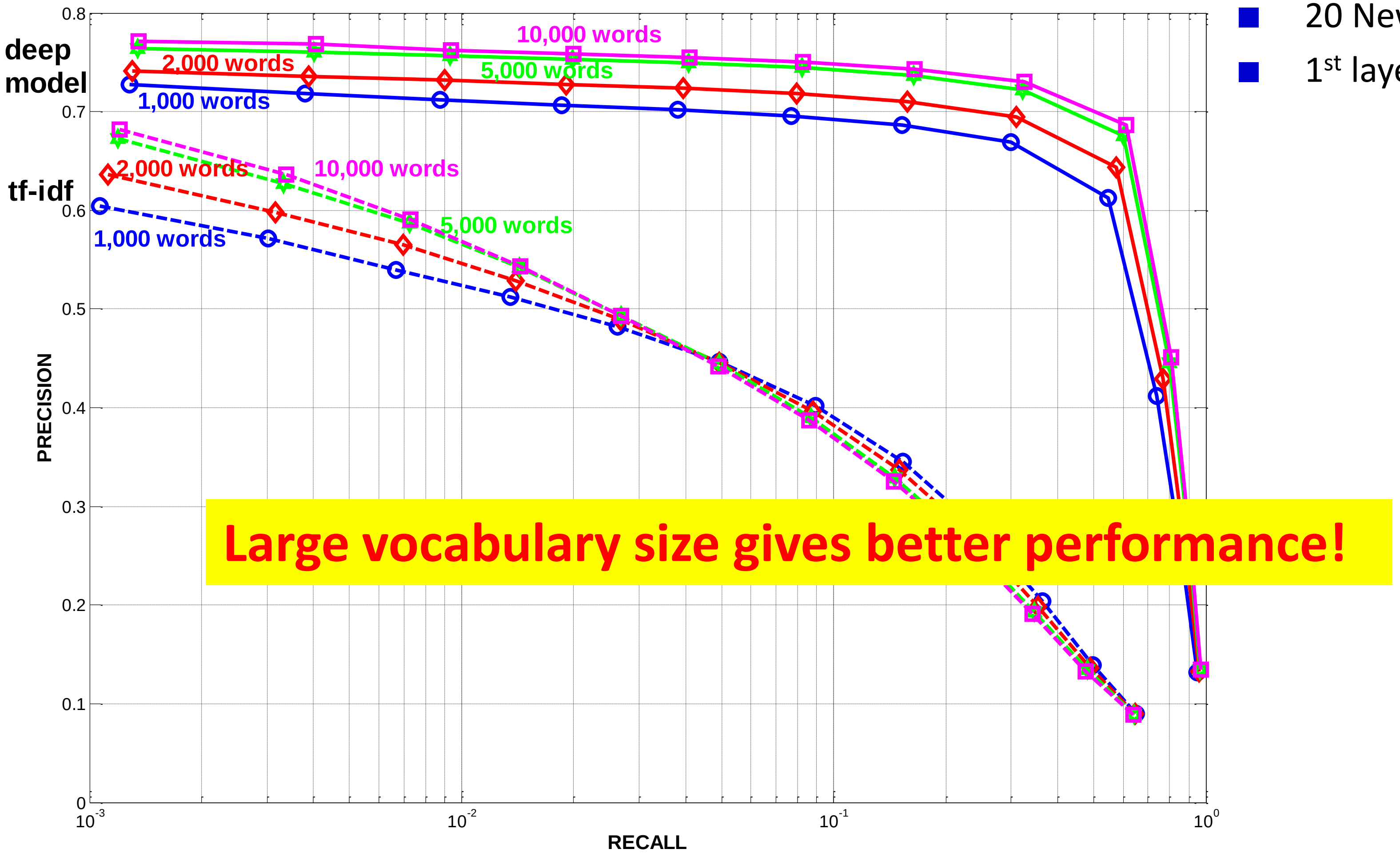
deep (20) & DBN:
2000-200-100-20

Deep vs DBN vs SESM



deep (20) & DBN:
2000-200-100-20

Vocabulary size



- 20 Newsgroups
- 1st layer: 200 units

Summary

- **Deep Semi-supervised auto-encoders**
  - **Efficient inference**
  - **Efficient semi-supervised learning**
  - **Compact** and informative features
- Semi-supervised vs Unsupervised
  - supervision helps
- Deep vs shallow
  - deep is needed to create very compact representations

- Autoencoders can give competitive accuracy to Deep Belief Nets.
  Autoencoders possibly train faster

- Can be integrated in a larger system whose parameters are updated by gradient descent (e.g. a ranker)

Perspectives

- Beyond bag of words
  - Proximity models
  - Language models
  - Linguistic information: Part of speech, grammar, clicks
- Binary representations
- Sparse codes: could be used in the inverted index

Thank you!