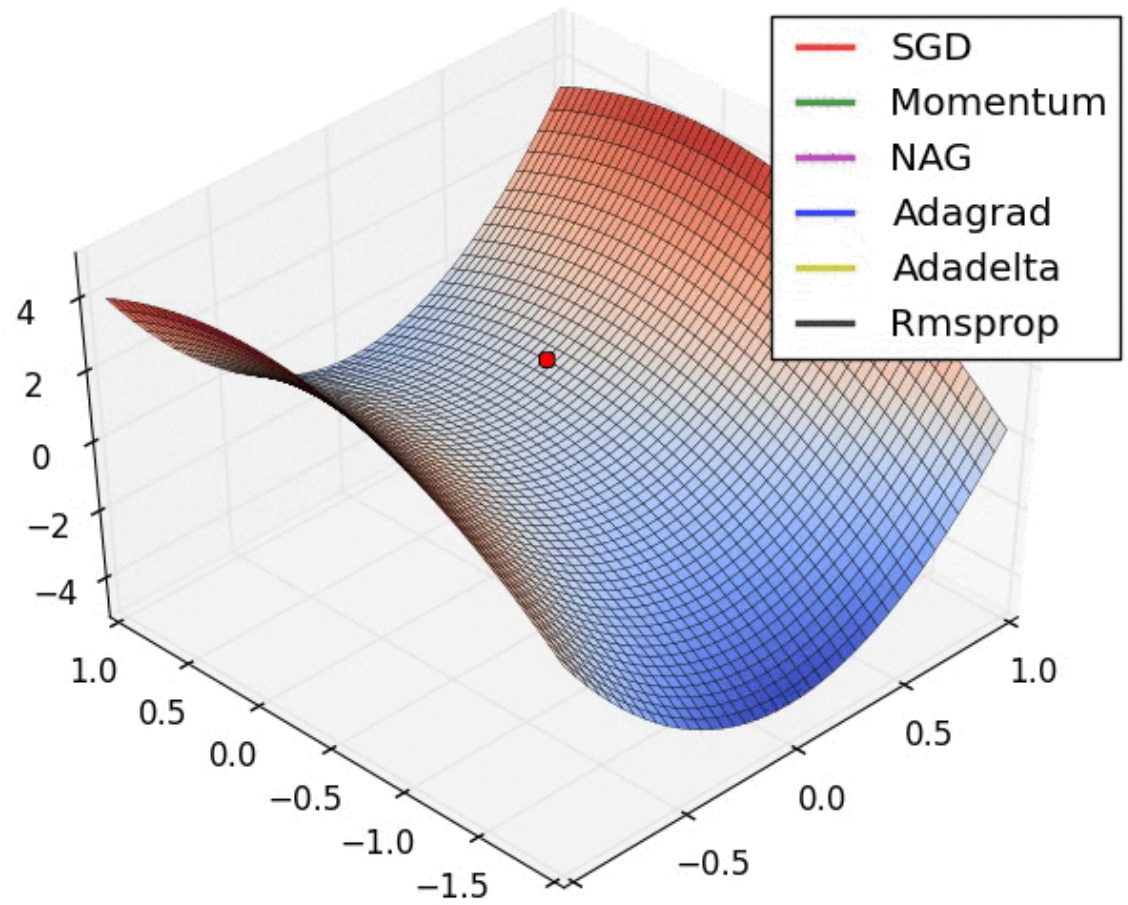
Lecture 5: Learning Representations

Olexandr Isayev

Department of Chemistry, CMU

olexandr@cmu.edu


SGD
Momentum
NAG
Adagrad
Adadelta
Rmsprop
4
2
0
−2
−4
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0
0.5
0.0
−0.5

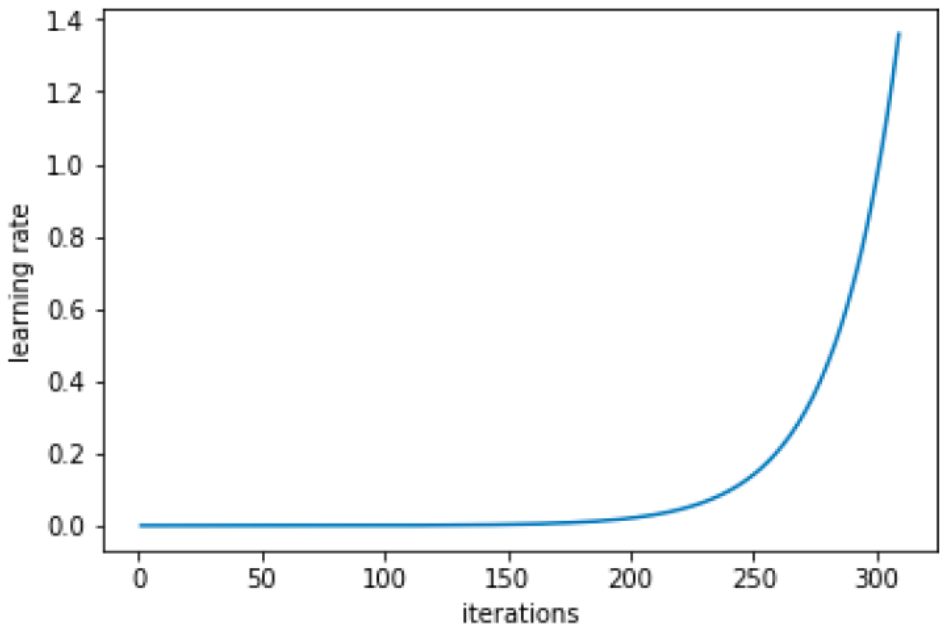
Which optimizer should we use?

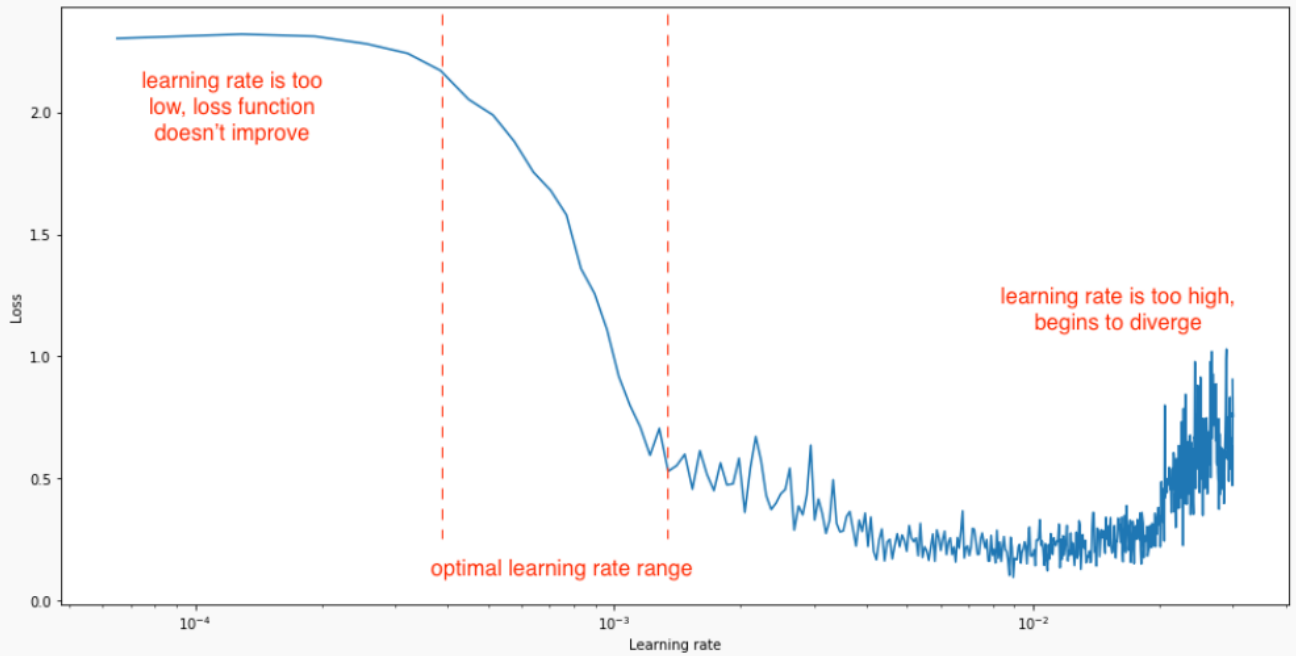
Adam *works well in practice and outperforms other Adaptive techniques.*

There are newer flavors of Adam: **NAdam**, **AdamW**…

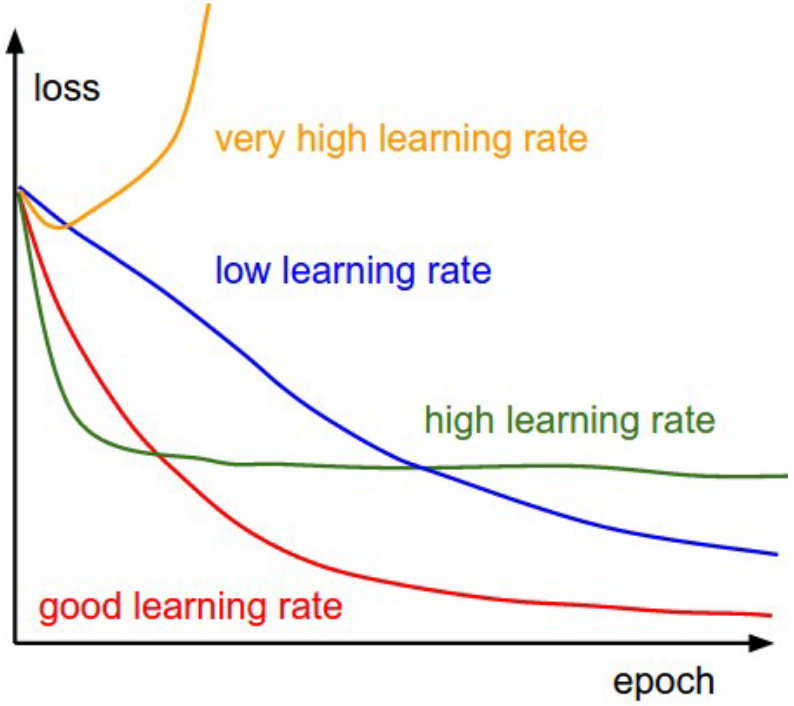
If your input data is sparse then methods such as **SGD,NAG and momentum** are inferior and perform poorly. **For sparse data sets one should use one of the *adaptive learning-rate* methods**. An additional benefit is that we won't need to adjust the learning rate but likely achieve the best results with the default value.

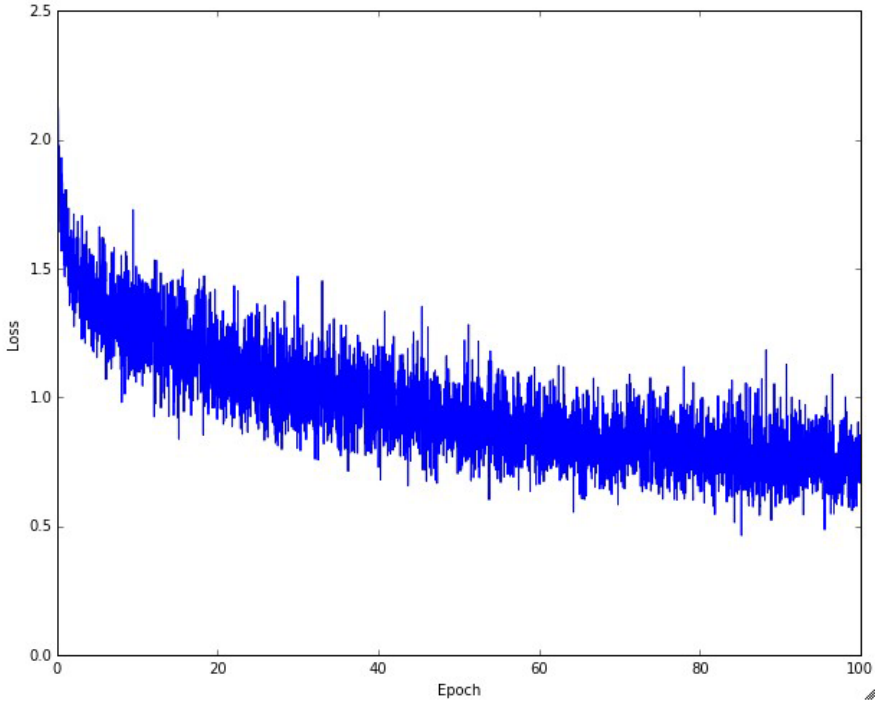
If one wants fast convergence and train a deep Neural Network Model or a highly complex Neural Network then **Adam or any other Adaptive learning rate techniques** should be used because they outperforms every other optimization algorithms.


learning rate
iterations
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
50
100
150
200
250
300


learning rate is too low, loss function doesn't improve
optimal learning rate range
learning rate is too high, begins to diverge
Loss
Learning rate
0.0
0.5
1.0
1.5
2.0
10^{-4}
10^{-3}
10^{-2}

Optimal learning rate: Monitor loss function

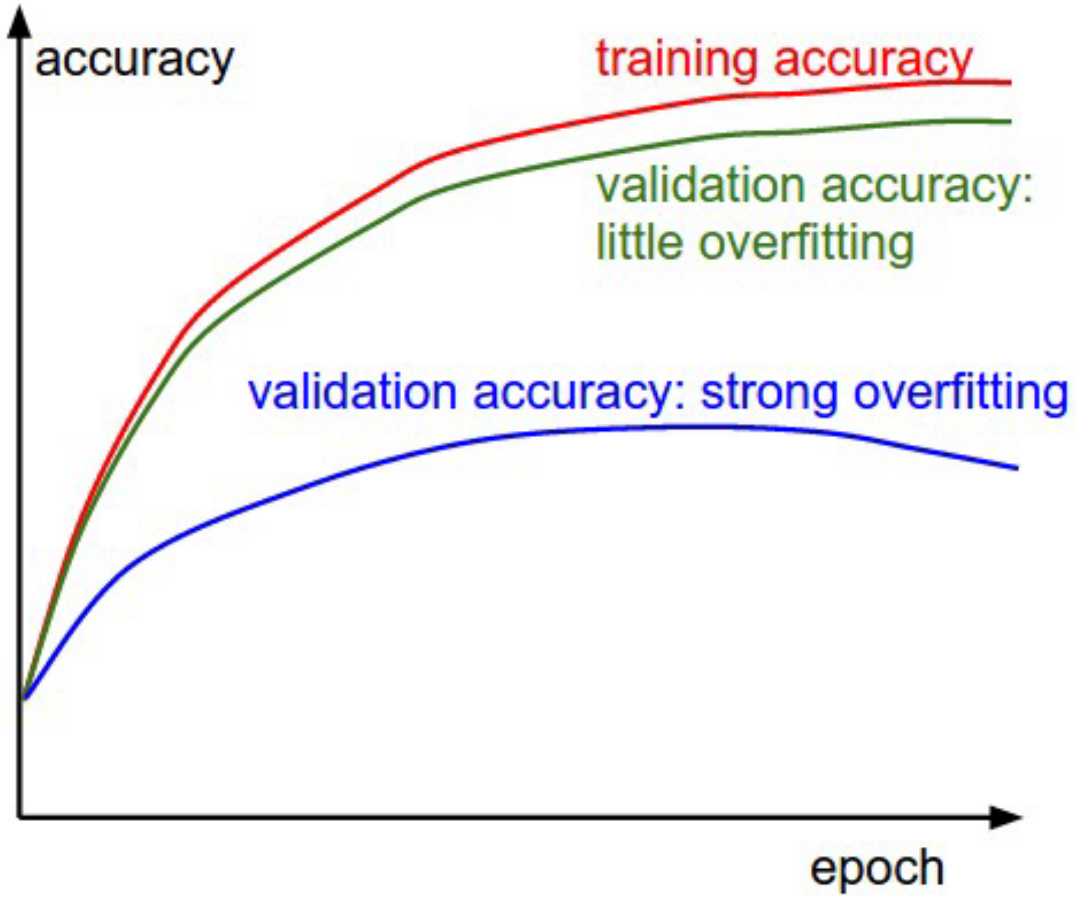




Left: A cartoon depicting the effects of different learning rates. With low learning rates the improvements will be linear. With high learning rates they will start to look more exponential. Higher learning rates will decay the loss faster, but they get stuck at worse values of loss (green line). This is because there is too much "energy" in the optimization and the parameters are bouncing around chaotically, unable to settle in a nice spot in the optimization landscape. **Right:** An example of a typical loss function over time.

From: https://cs231n.github.io/

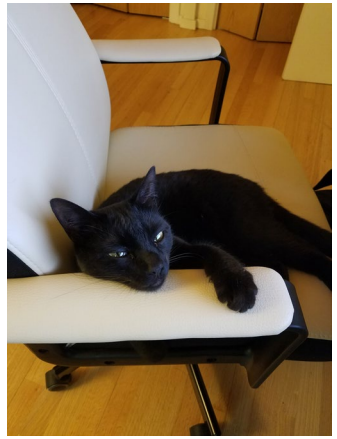
Train/Val accuracy

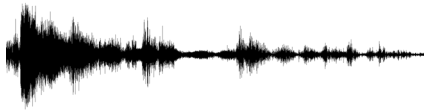


The gap between the training and validation accuracy indicates the amount of overfitting. Two possible cases are shown in the diagram on the left. The blue validation error curve shows very small validation accuracy compared to the training accuracy, indicating strong overfitting (note, it's possible for the validation accuracy to even start to go down after some point). When you see this in practice you probably want to increase regularization (stronger L2 weight penalty, more dropout, etc.) or collect more data. The other possible case is when the validation accuracy tracks the training accuracy fairly well. This case indicates that your model capacity is not high enough: make the model larger by increasing the number of parameters.

From: https://cs231n.github.io/

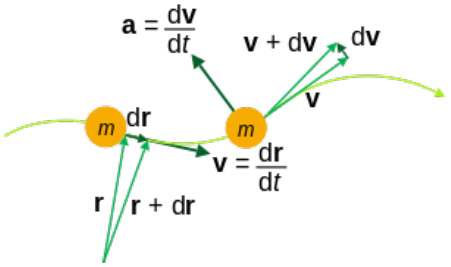
Things...

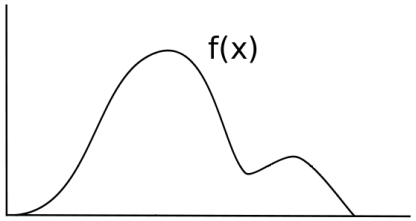


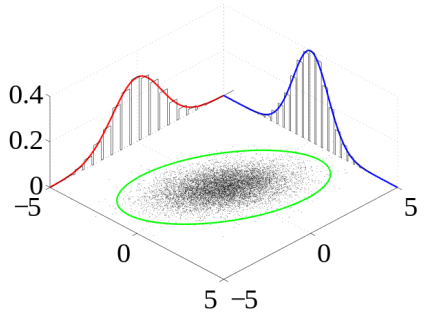


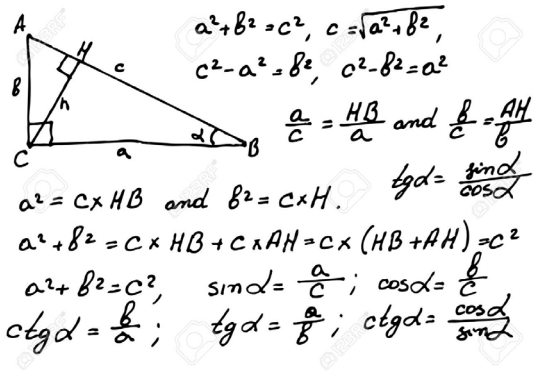
My heart beats as if the world is dropping,
you may not feel the love but i do its a heart
breaking moment of your life. enjoy the times
that we have, it might not sound good but
one thing it rhymes it might not be romantic
but i think it is great,the best rhyme i've ever
heard.

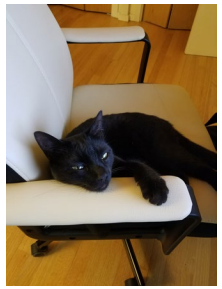
Our Knowledge...

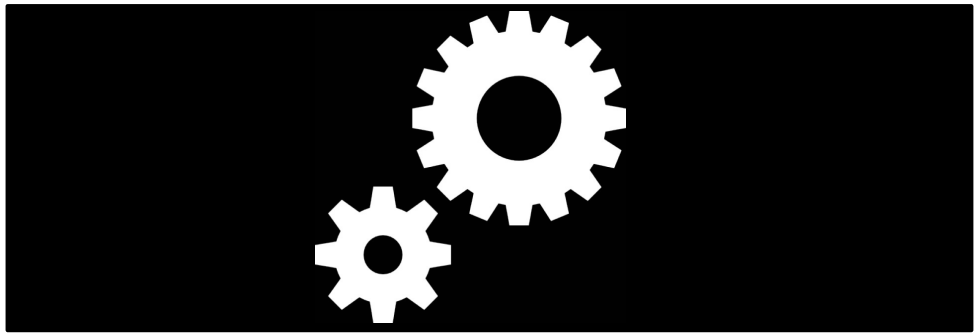












“Transcript”

I dare not speak of what I have done. Such twisted thoughts overtook my mind and now I am sorry to say that *I have done the deed.* I have murdered the King of Scotland, King Duncan.

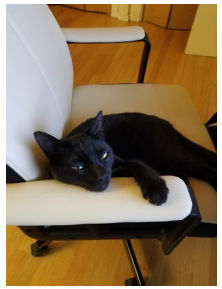
After naming me the *Worthy Thane of Cawdor,* this is how I repay him. I have betrayed him in the most unimaginable way a person possibly could, and I've been disloyal to him, just like I have to Banquo, whom I have lost as a dear friend. I wish that I had never done such a treacherous thing, as I am afraid that *I shall sleep no more*.

I was waiting anxiously for my Lady to sound the bell that called me to do the deed. But before she did, a symbol of the supernatural appeared before my eyes. *The dagger of the mind* captured me and the handle was to my hand yet I couldn't grasp it, but *I could see thee still*. I knew not whether to follow or to discard it from my eyes, but the *false creation* remained.

As I stepped closer to Duncan's room, I thought that I would panic and freeze, but when I got nearer, a sickening thought made me feel like I was doing the right thing! As soon as I heard the bell I knew that it was the bell summoning me.

I heard him pleading as the dagger pierced through his skin,

Macbeth was guilty.



I dare not speak of what I have done. Such twisted thoughts overtook my mind and now I am sorry to say that *I have done the deed.* I have murdered the King of Scotland, King Duncan.

After naming me the *Worthy Thane of Cawdor,* this is how I repay him. I have betrayed him in the most unimaginable way a person possibly could, and I've been disloyal to him, just like I have to Banquo, whom I have lost as a dear friend. I wish that I had never done such a treacherous thing, as I am afraid that *I shall sleep no more*.

I was waiting anxiously for my Lady to sound the bell that called me to do the deed. But before she did, a symbol of the supernatural appeared before my eyes. *The dagger of the mind* captured me and the handle was to my hand yet I couldn't grasp it, but *I could see thee still*. I knew not whether to follow or to discard it from my eyes, but the *false creation* remained.

As I stepped closer to Duncan's room, I thought that I would panic and freeze, but when I got nearer, a sickening thought made me feel like I was doing the right thing! As soon as I heard the bell I knew that it was the bell summoning me.

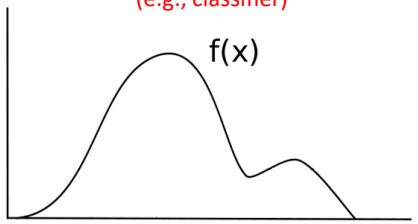
I heard him pleading as the dagger pierced through his skin,

Representation

[81 20 84 64 58 39 17 54 72 15]

Mathematical Model

(e.g., classifier)



I dare not speak of what I have done. Such twisted thoughts overtook my mind and now I am sorry to say that *I have done the deed.* I have murdered the King of Scotland, King Duncan.

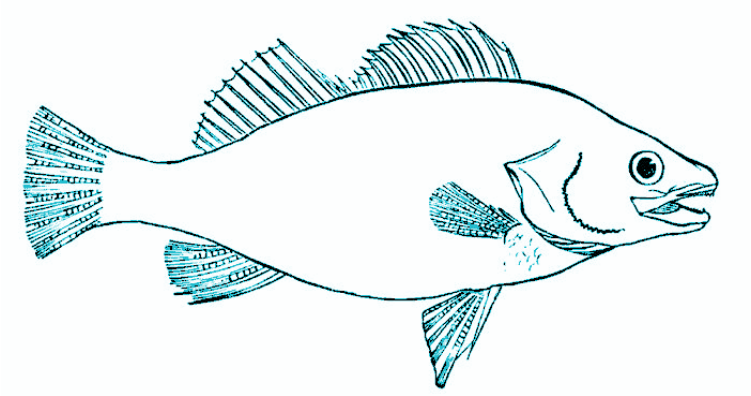
After naming me the *Worthy Thane of Cawdor,* this is how I repay him. I have betrayed him in the most unimaginable way a person possibly could, and I've been disloyal to him, just like I have to Banquo, whom I have lost as a dear friend. I wish that I had never done such a treacherous thing, as I am afraid that *I shall sleep no more*.

I was waiting anxiously for my Lady to sound the bell that called

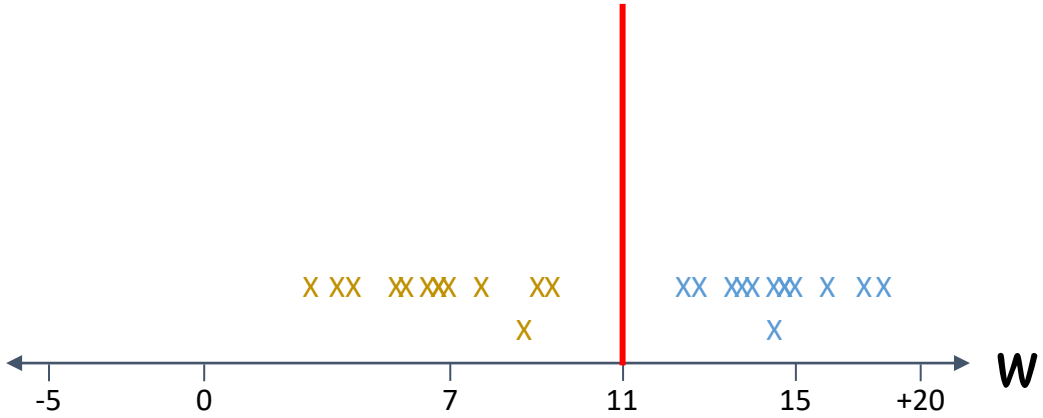
Macbeth was guilty.

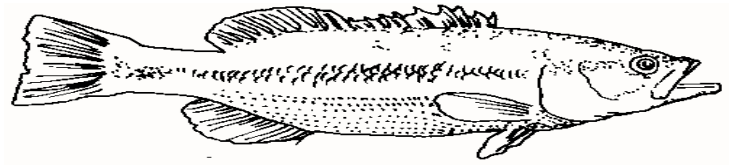
As I stepped closer to Duncan's room, I thought that I would panic and freeze, but when I got nearer, a sickening thought made me feel like I was doing the right thing! As soon as I heard the bell I knew that it was the bell summoning me.

I heard him pleading as the dagger pierced through his skin,

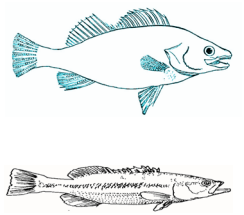


~12 lbs





~8 lbs



Representation

Weight (*w*)

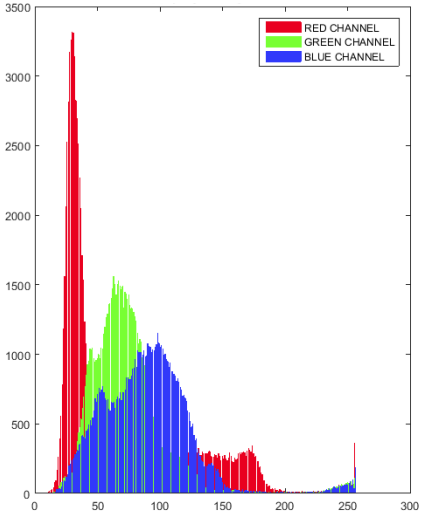
Mathematical Model (Classifier)

w>11

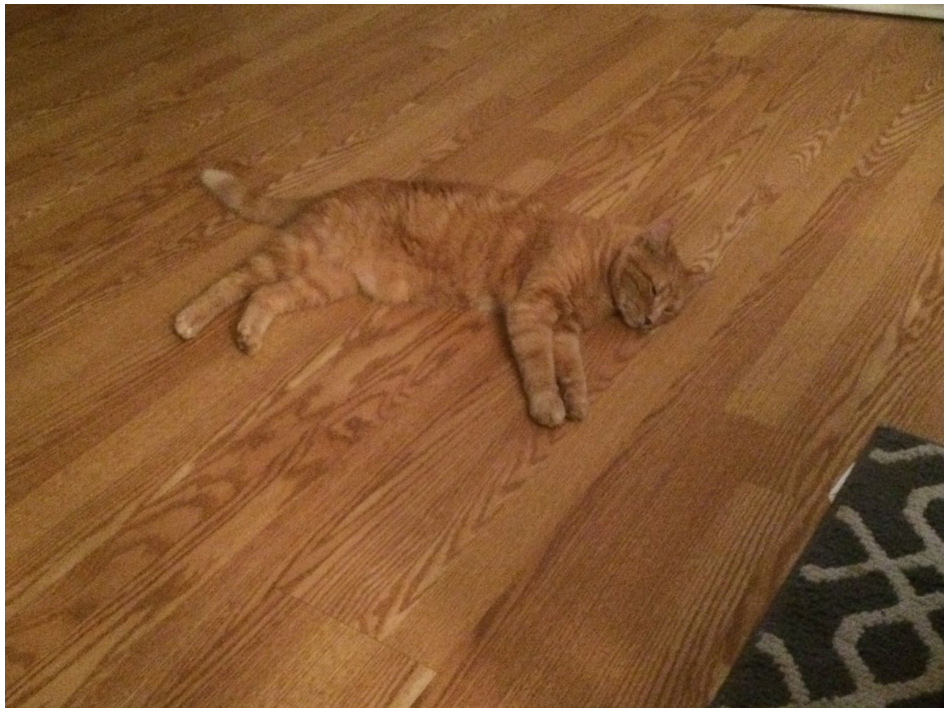
Type A

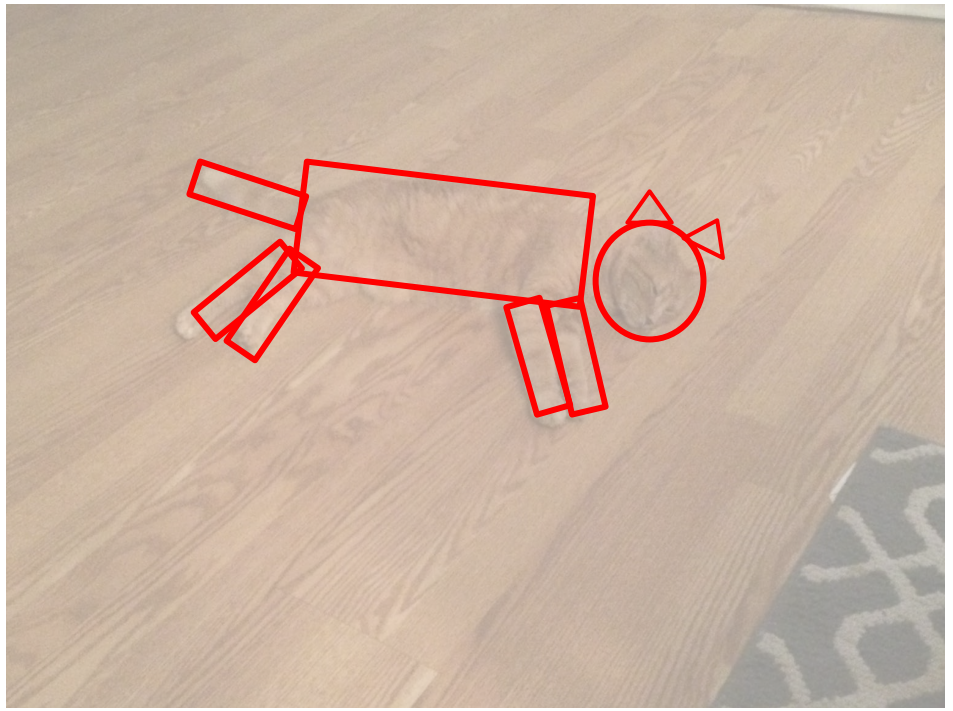
Type B

Represent these cats for a cat detector!

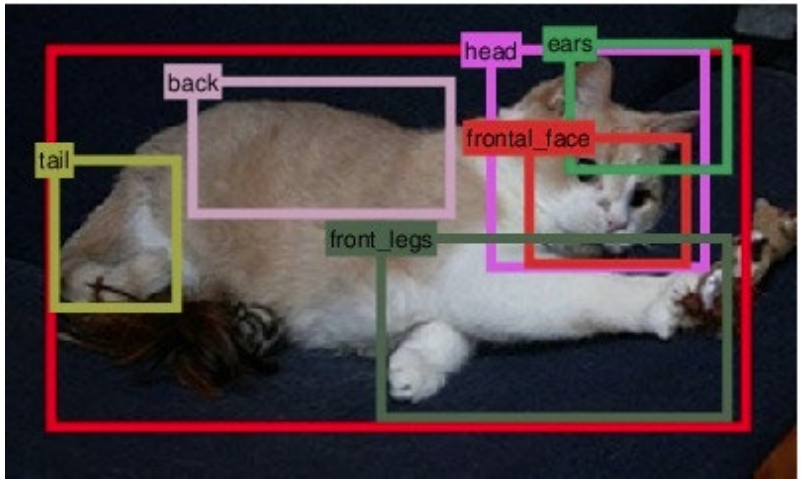


Represent these cats for a cat detector! (II)

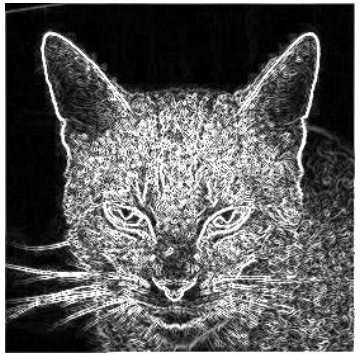


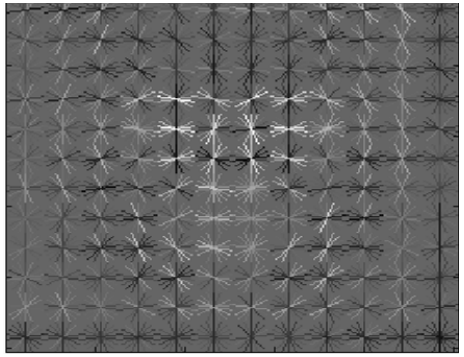


Represent these cats for a cat detector! (III)

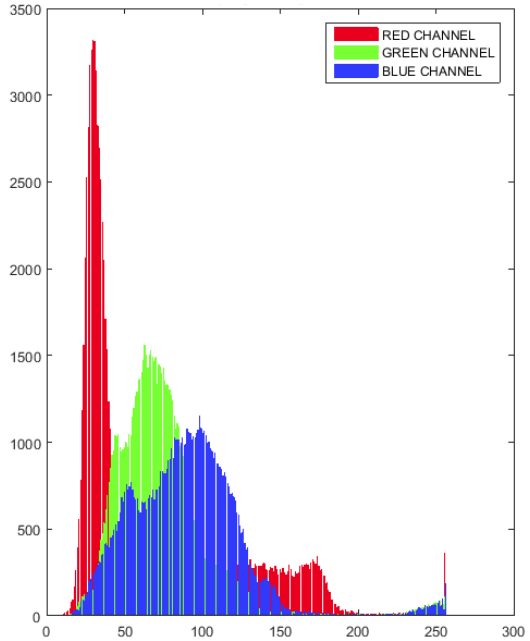


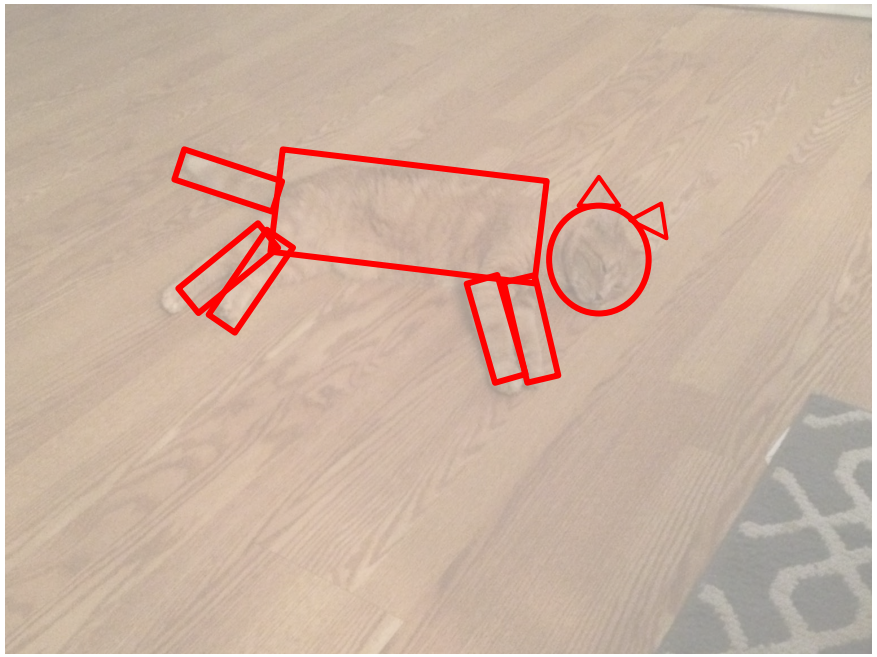
Represent these cats for a cat detector! (IV)





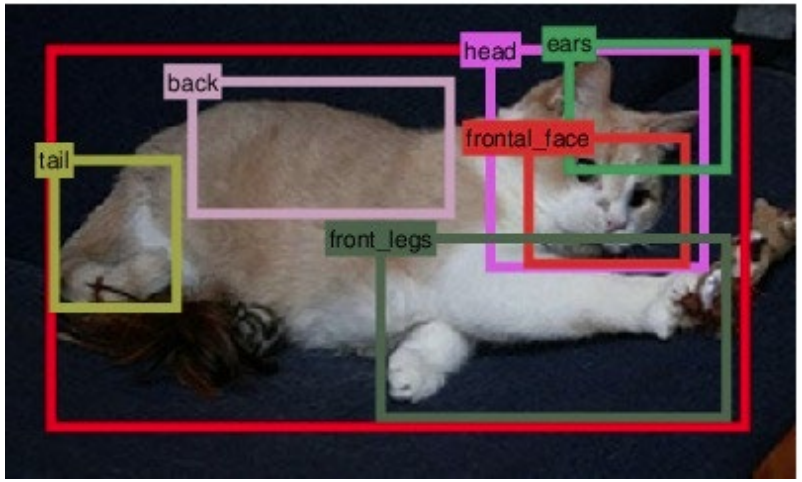
Color Histograms

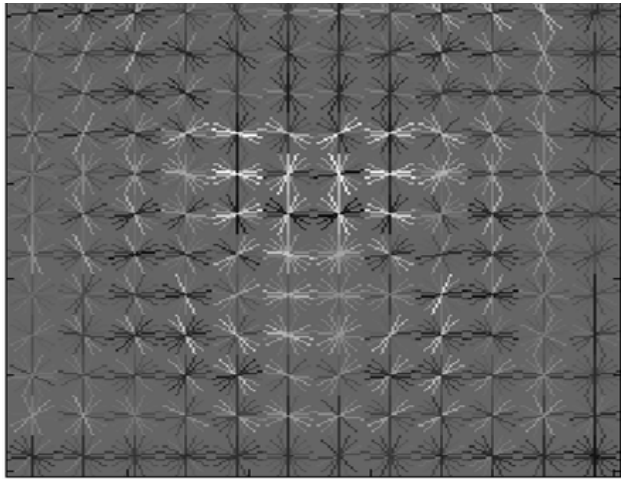




Models based Shapes

Deformable Part based Models (DPM)





Histogram of Gradients (HOG)

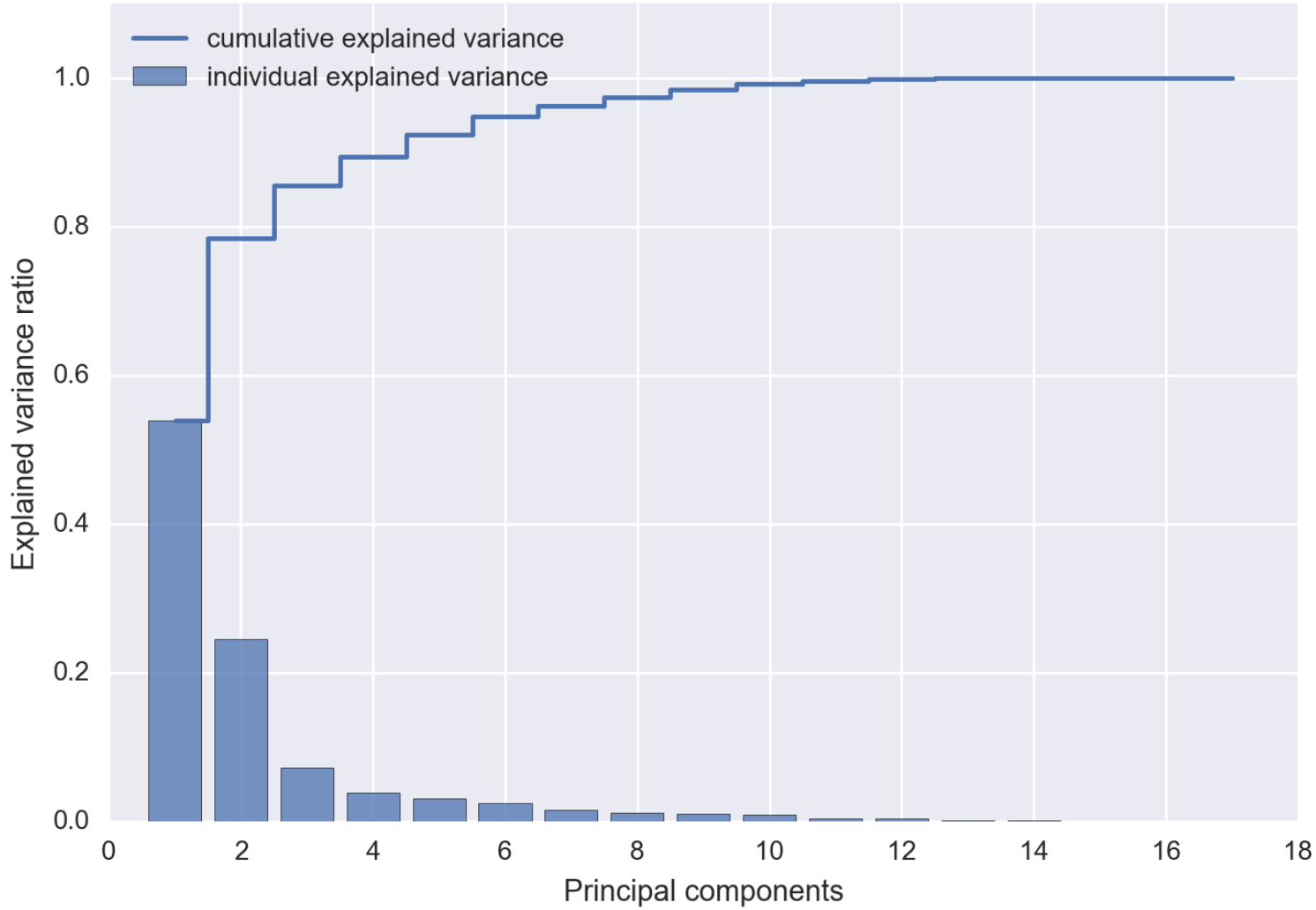
Felzenszwalb et al., 2010.
Dalal and Triggs, 2005.
Beis and Lowe, 1997.

Learning Representations

Definition of PCA

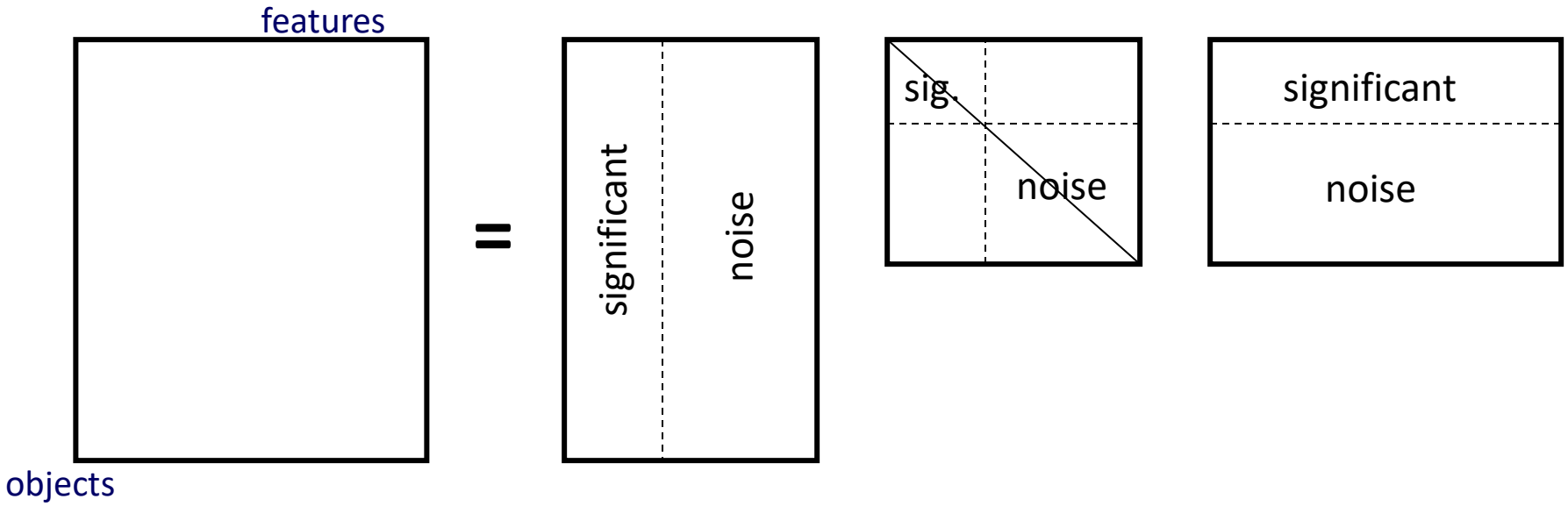
PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on

Explained Variance

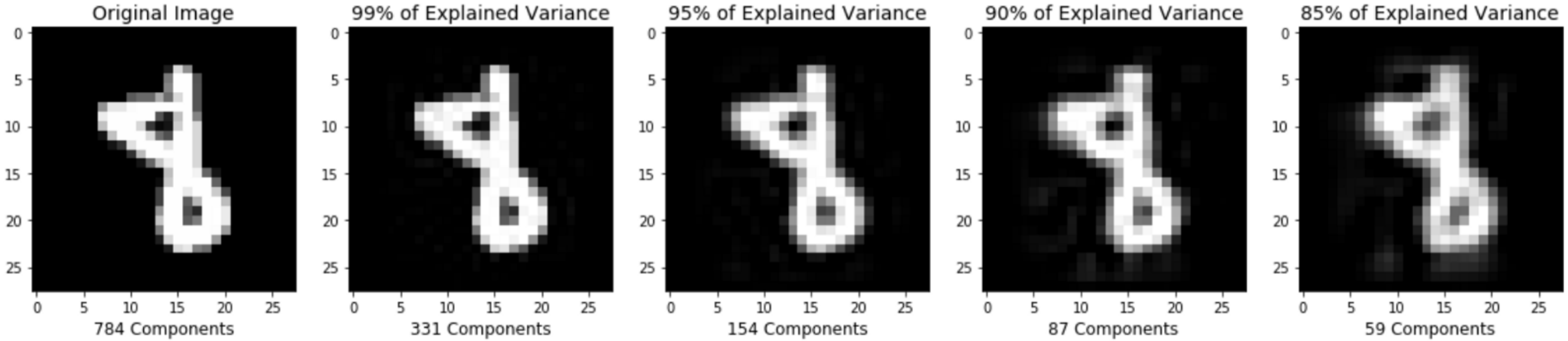


SVD/PCA and Rank-k approximations

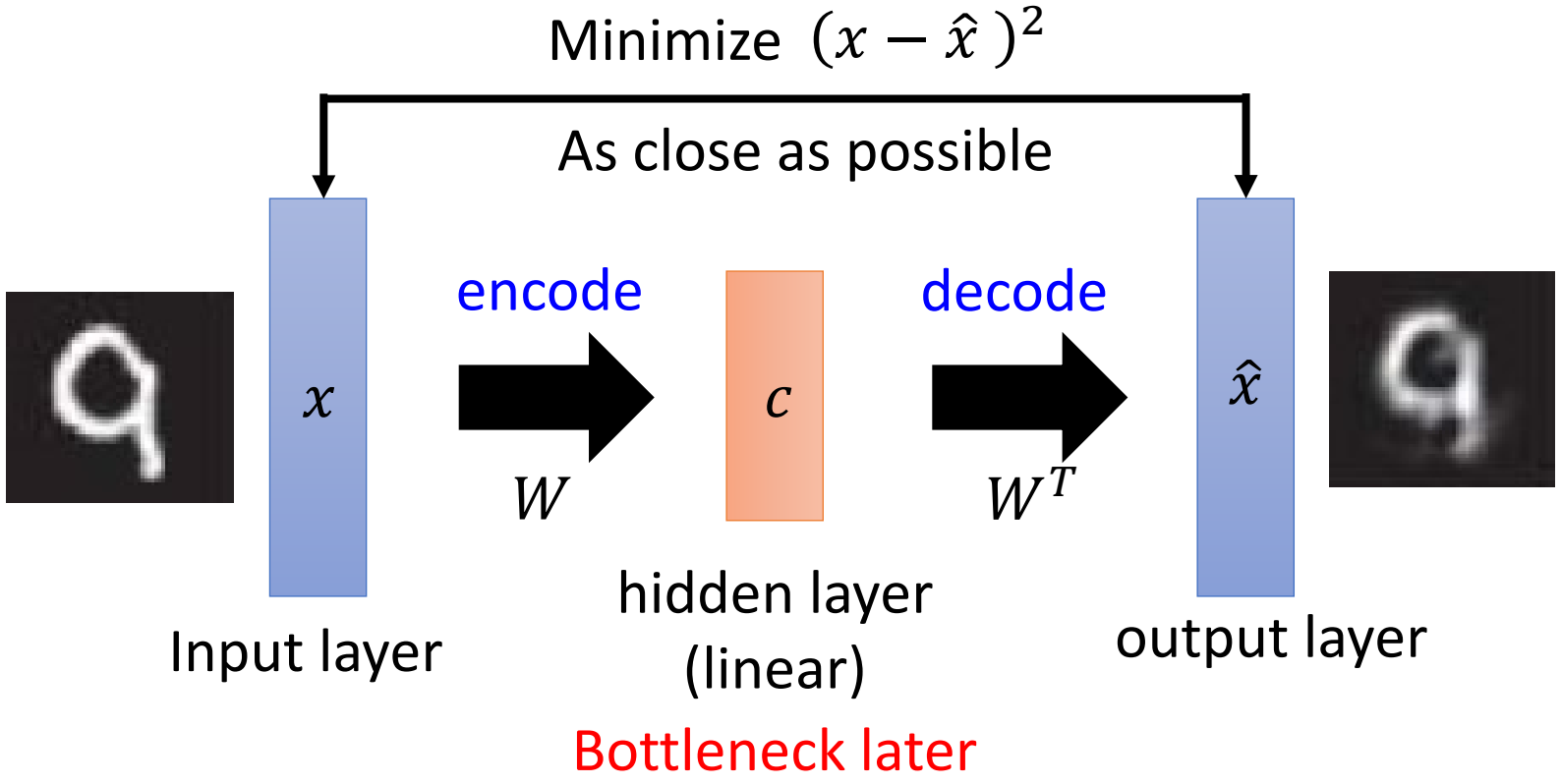
$$\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T$$



Data Compression

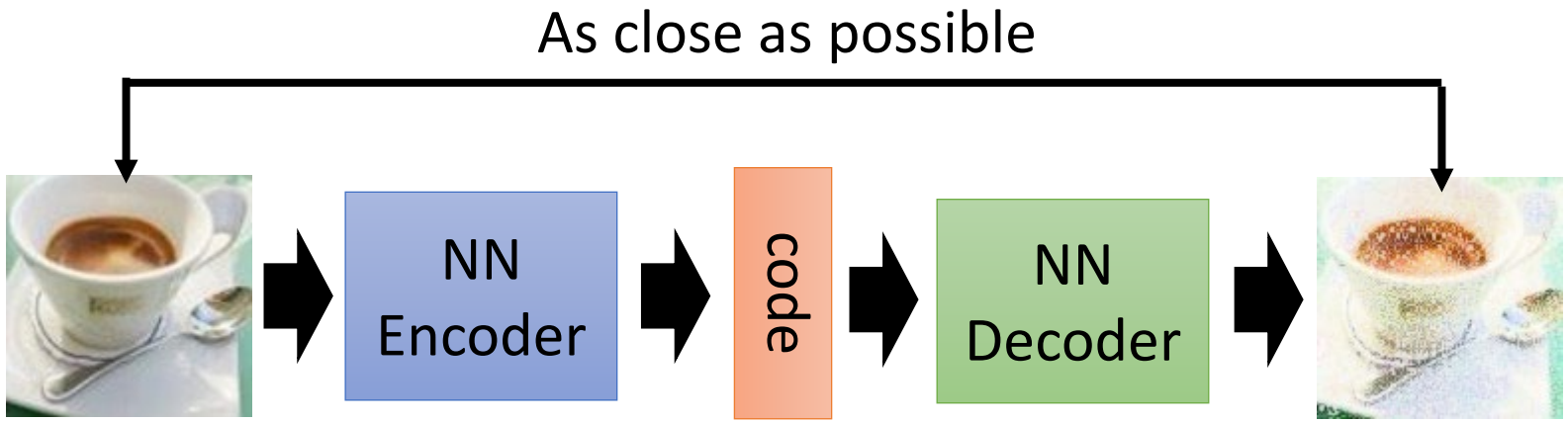


Recap: PCA

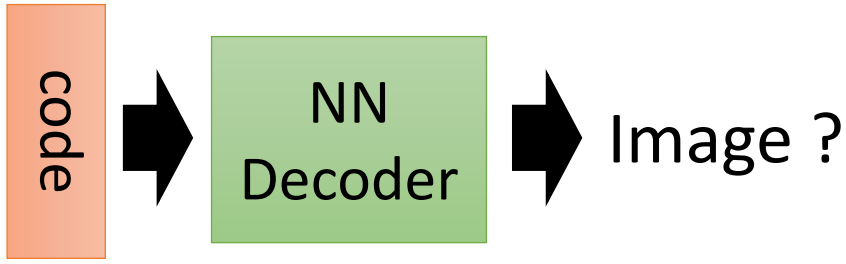


Output of the hidden layer is the code

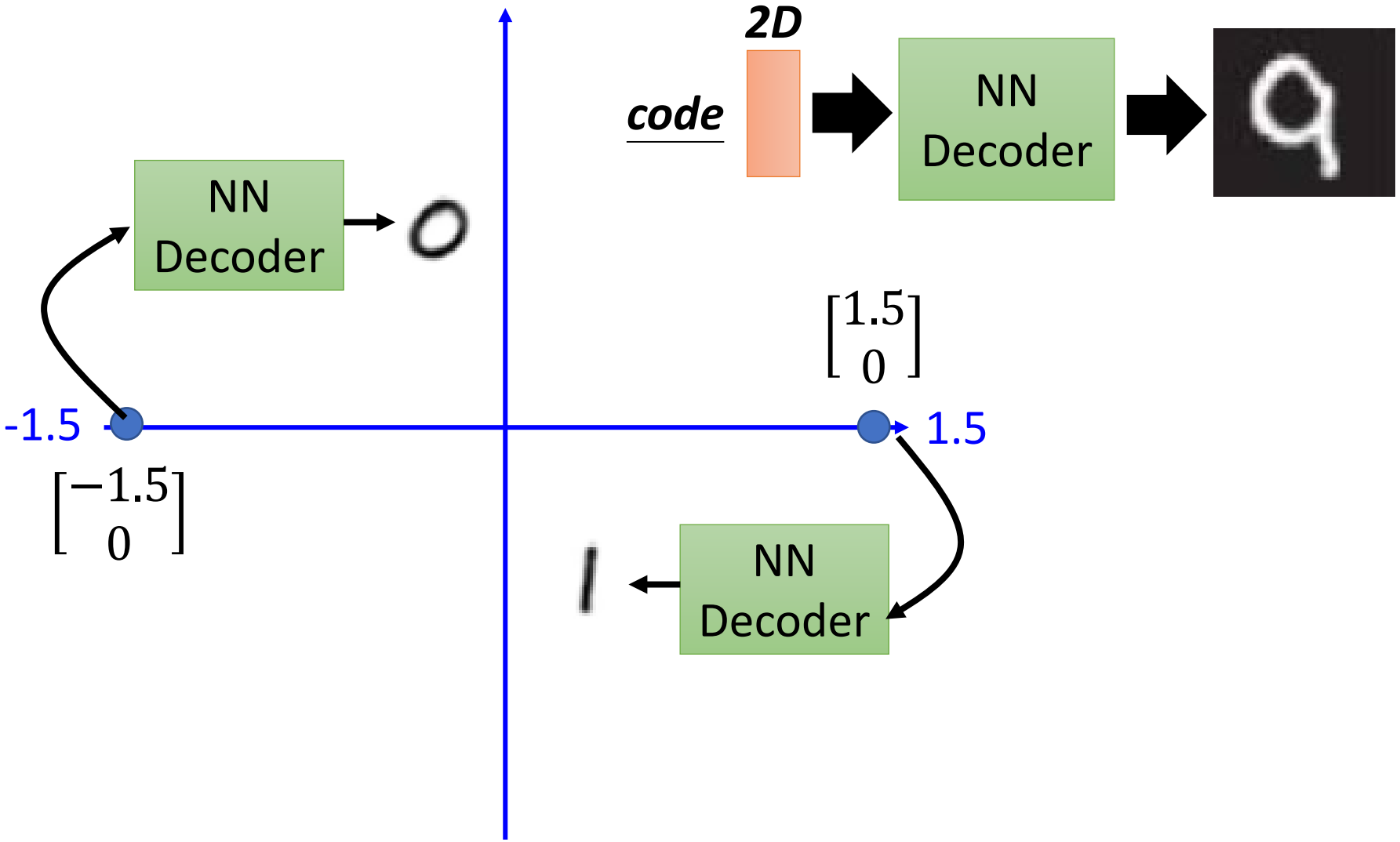
Auto-encoder



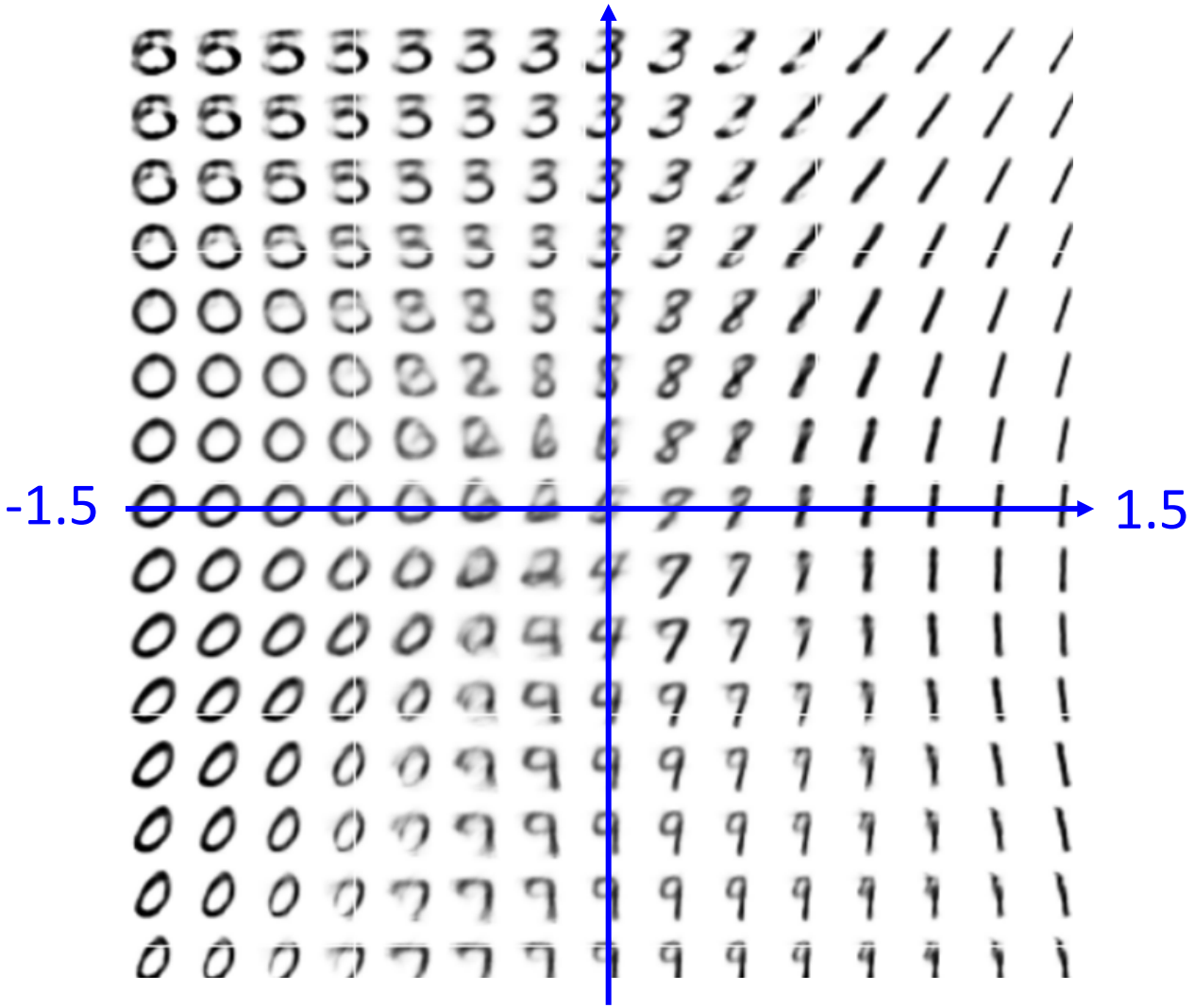
Randomly generate a vector as code



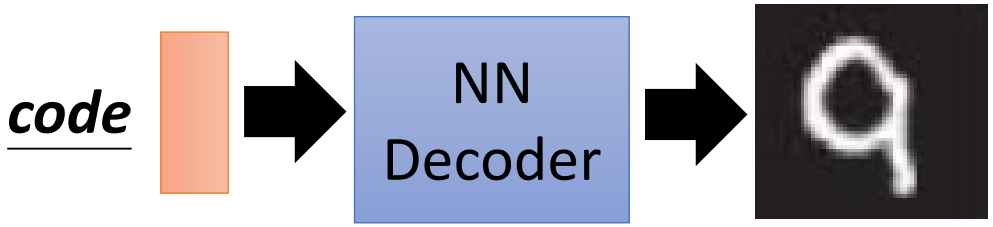
Auto-encoder



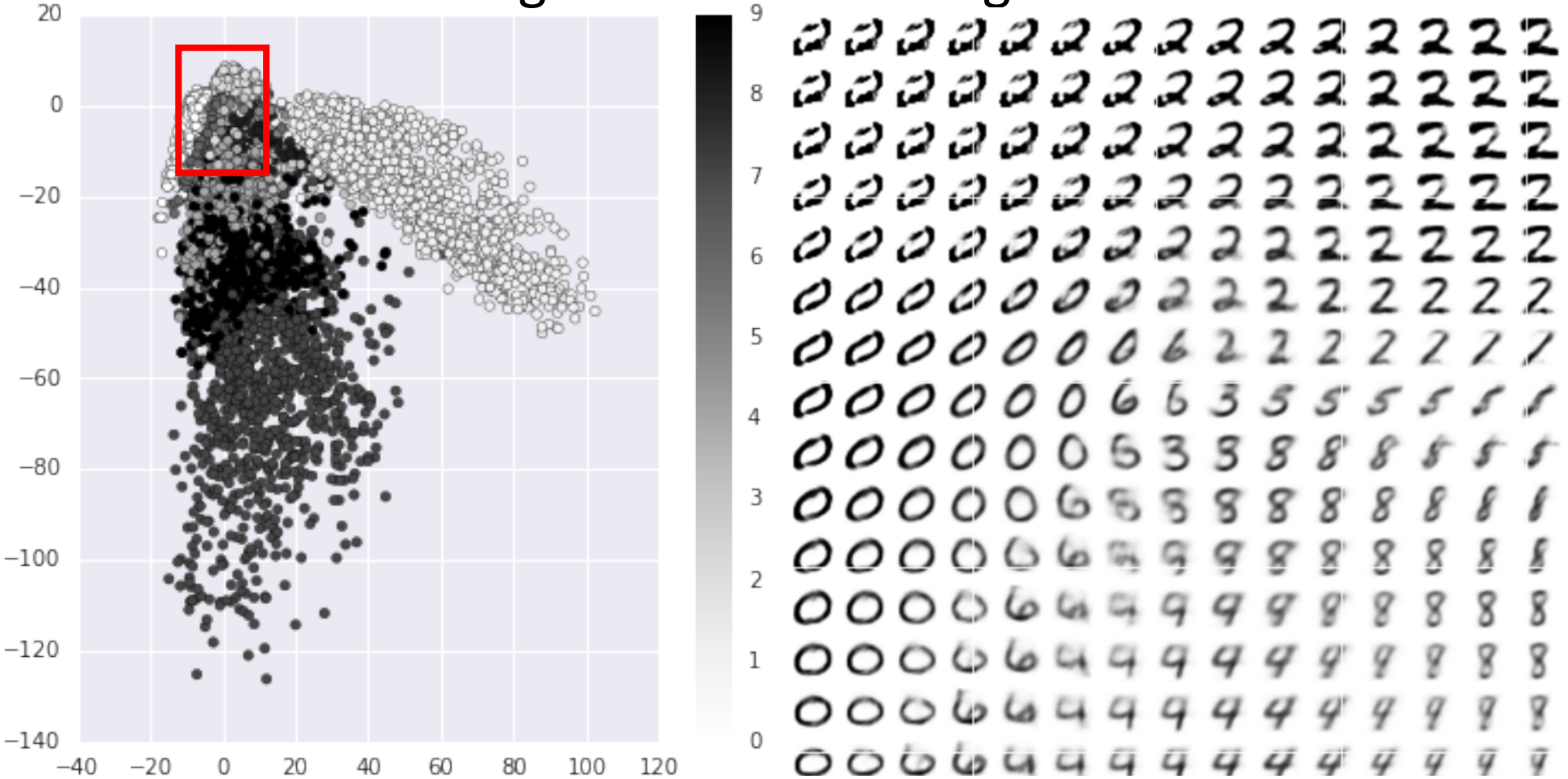
Auto-encoder



Next .....

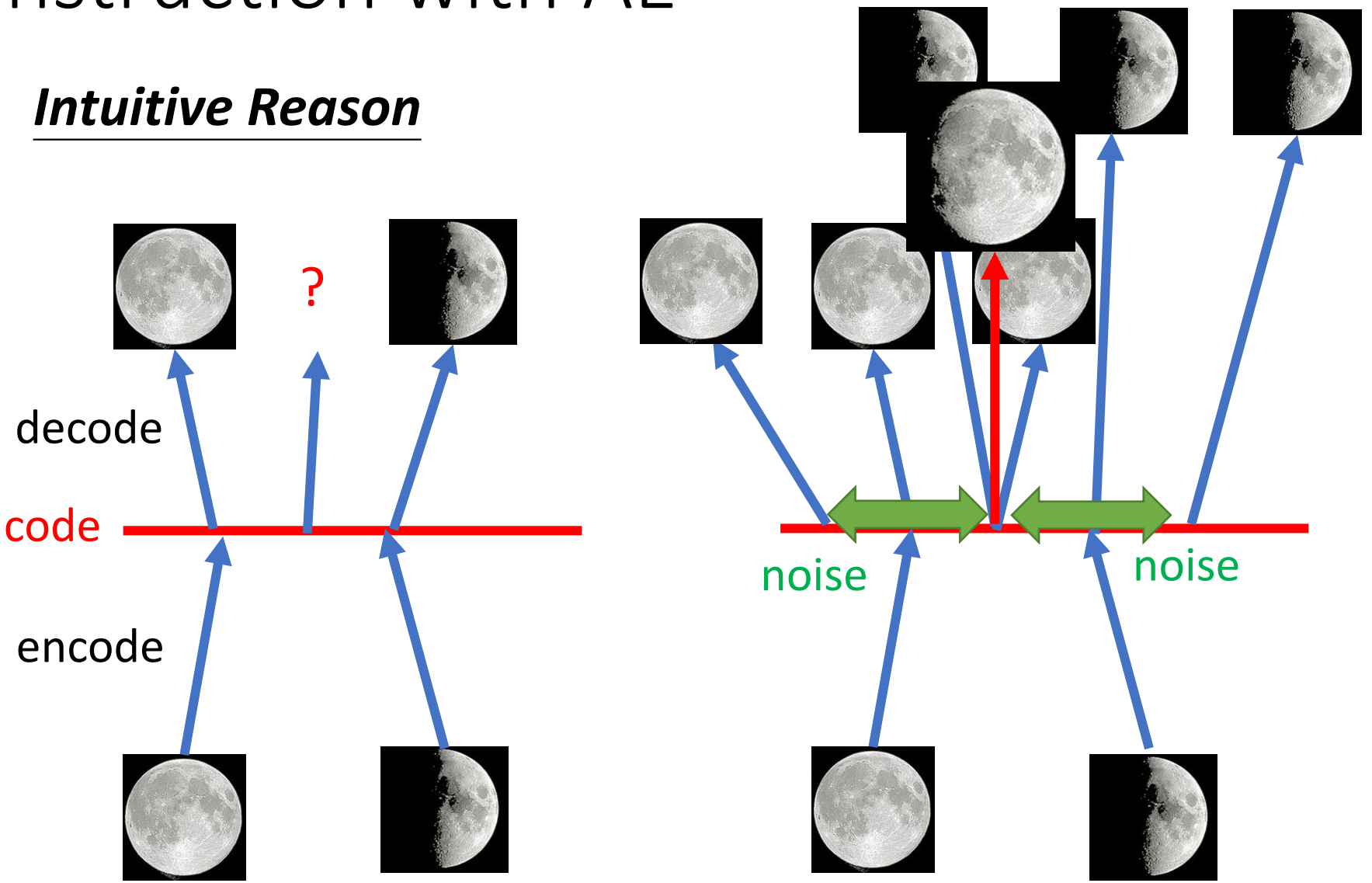


- Can we use decoder to generate something?

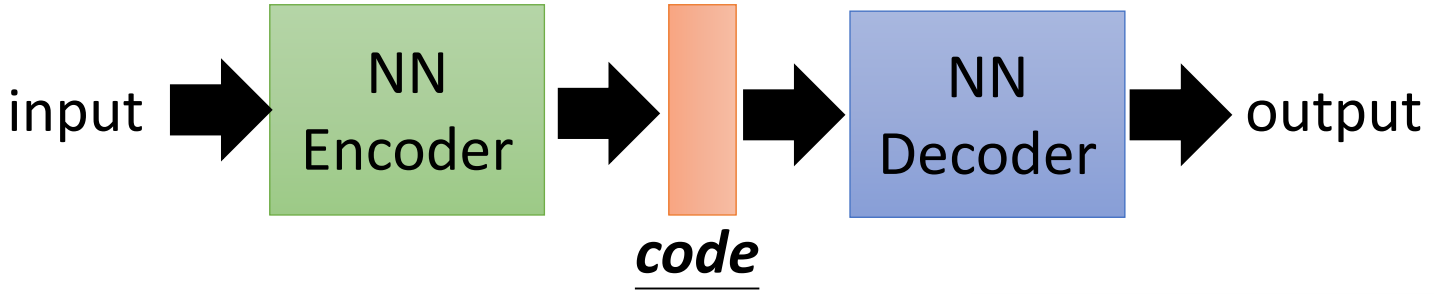


Reconstruction with AE

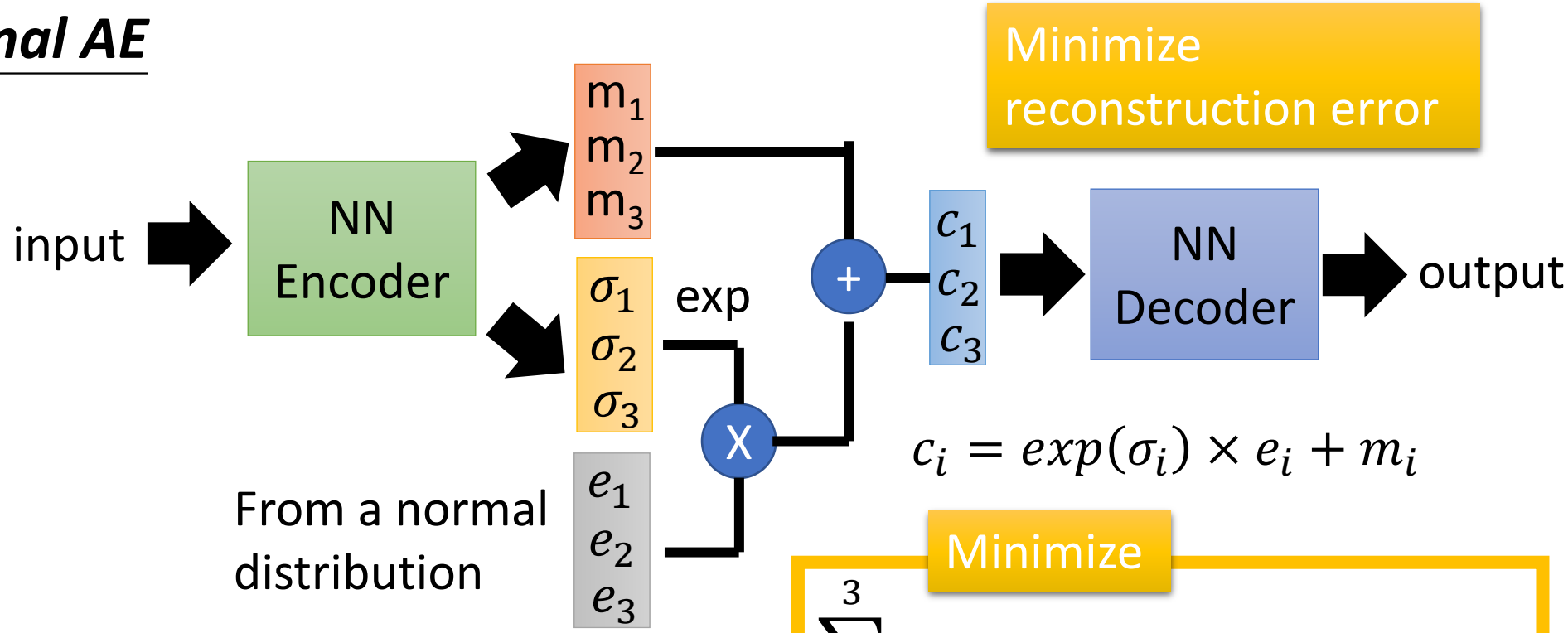
Intuitive Reason



Auto-encoder

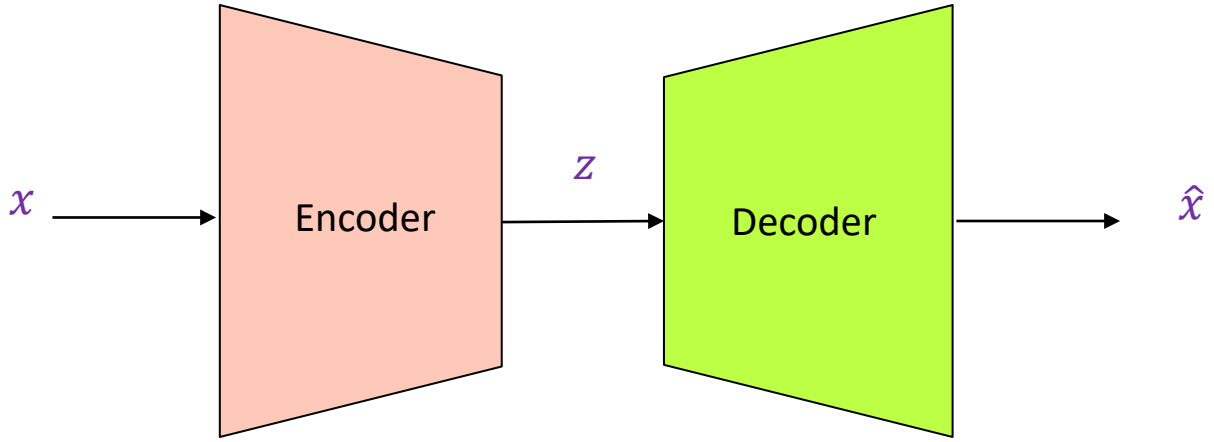


Variational AE



Variational autoencoders: Overview

- Probabilistic formulation based on *variational Bayes* framework
- At training time, jointly learn *encoder* and *decoder* by maximizing (a bound on) the data likelihood
- At test time, discard encoder and use decoder to sample from the learned distribution



D. Kingma and M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

Variational autoencoders: Overview

- Probabilistic generative model of the data distribution:

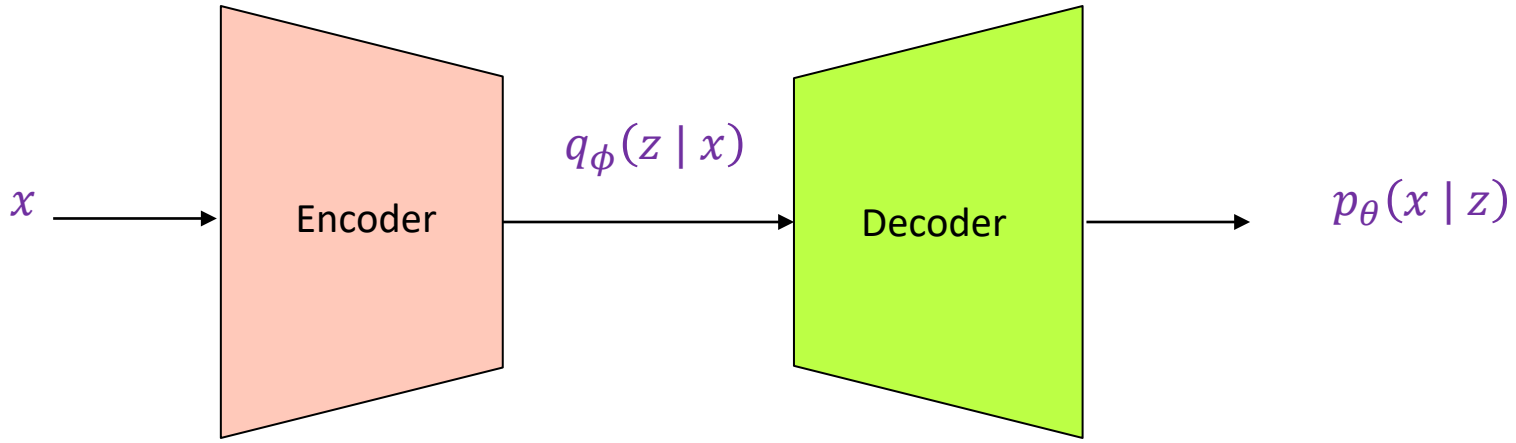
Sample z from *prior* $p(z)$ (usually Gaussian)

Sample x from *conditional* $p(x|z)$

Try to approximate the conditional with neural network

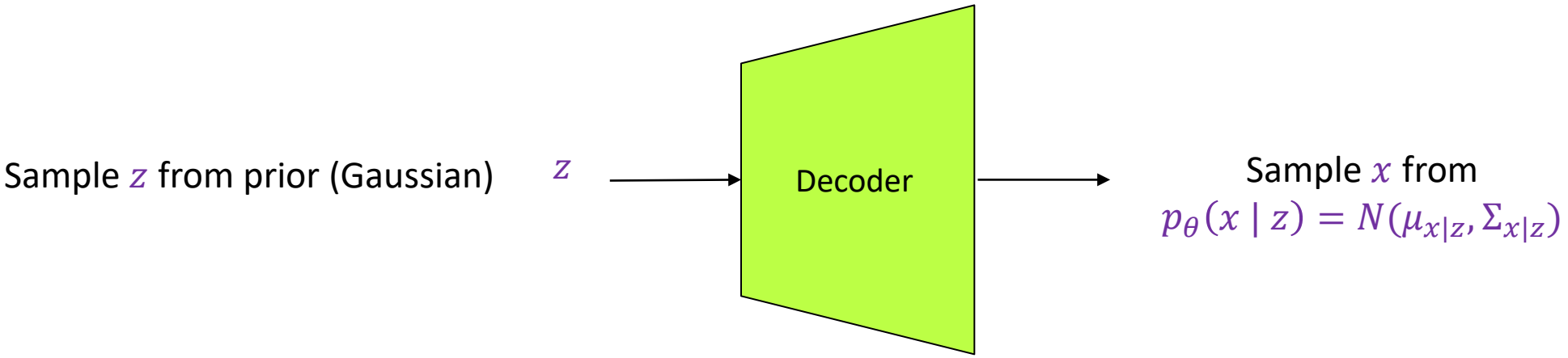
Variational autoencoders: Overview

- At training time, jointly learn *encoder* and *decoder*
- **Encoder:** given inputs x, output $q_\phi(z \mid x)$
  - Specifically, output mean and (diagonal) covariance, or $\mu_{z|x}$ and $\Sigma_{z|x}$, so that $q_\phi(z \mid x) = N(\mu_{z|x}, \Sigma_{z|x})$
- **Decoder:** given z, output $p_\theta(x \mid z)$
  - Specifically, output $\mu_{x|z}$ and $\Sigma_{x|z}$ so that $p_\theta(x \mid z) = N(\mu_{x|z}, \Sigma_{x|z})$
- **Training objective:** (a bound on) data likelihood under the model



Variational autoencoders: Overview

- At test time, discard encoder and use decoder to sample from $p_\theta(x \mid z) = N(\mu_{x|z}, \Sigma_{x|z})$



Variational autoencoders: Generating data

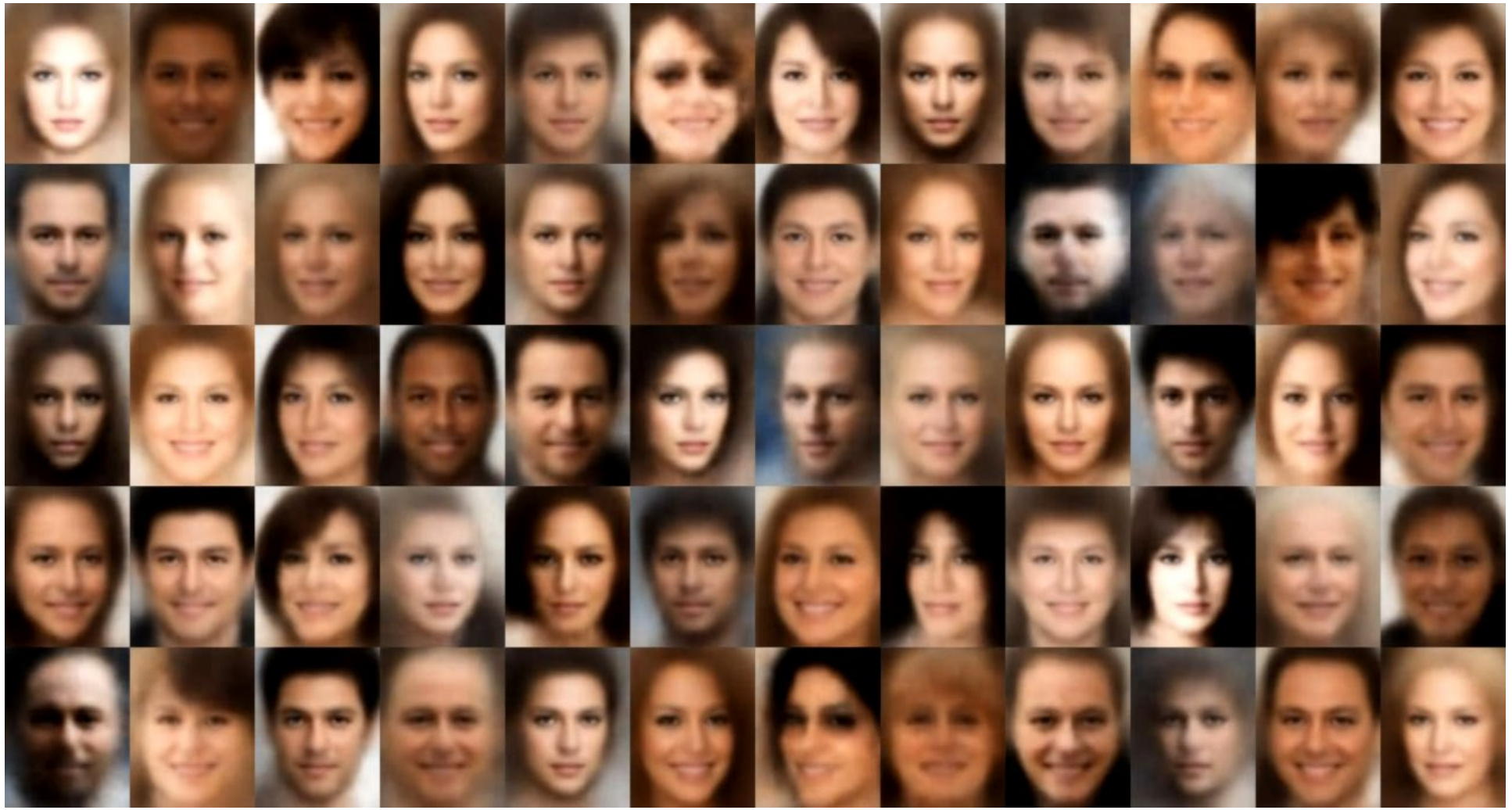
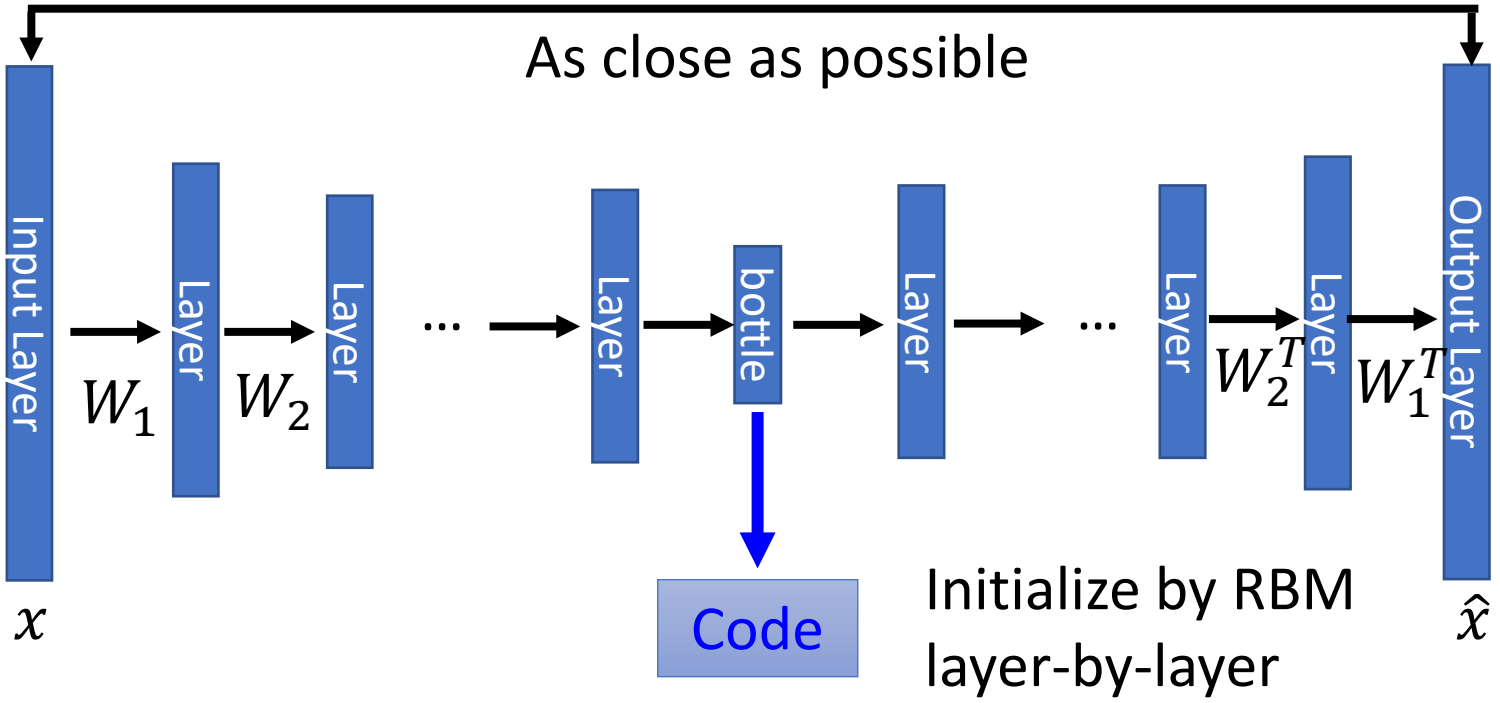


Image source

Deep Auto-encoder

Symmetric is not necessary.

- Of course, the auto-encoder can be deep



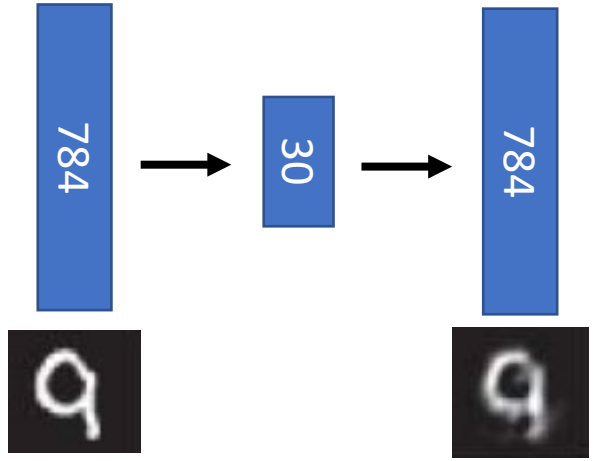
Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

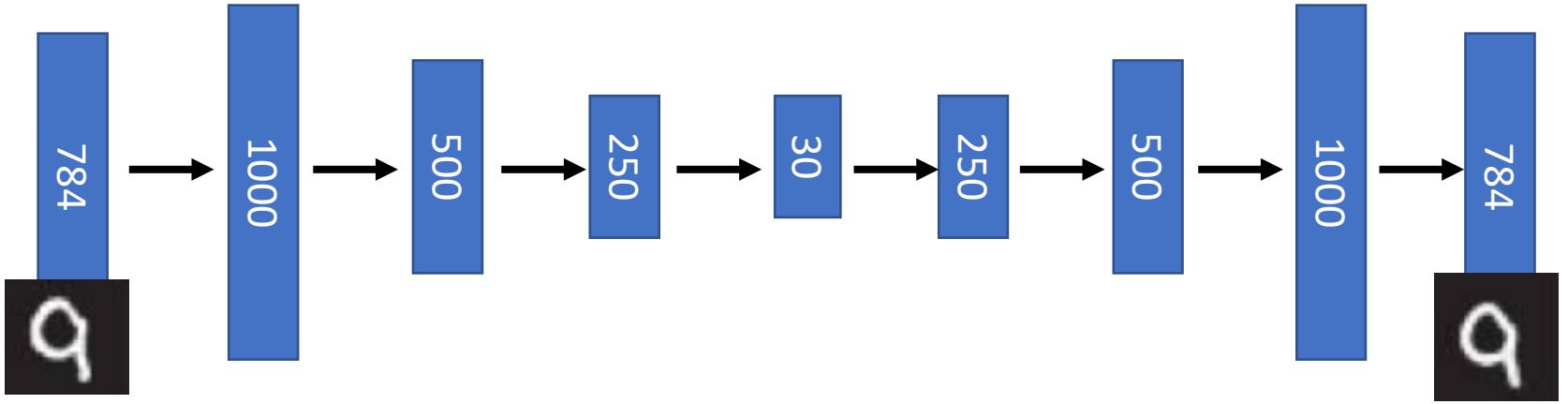
Deep Auto-encoder

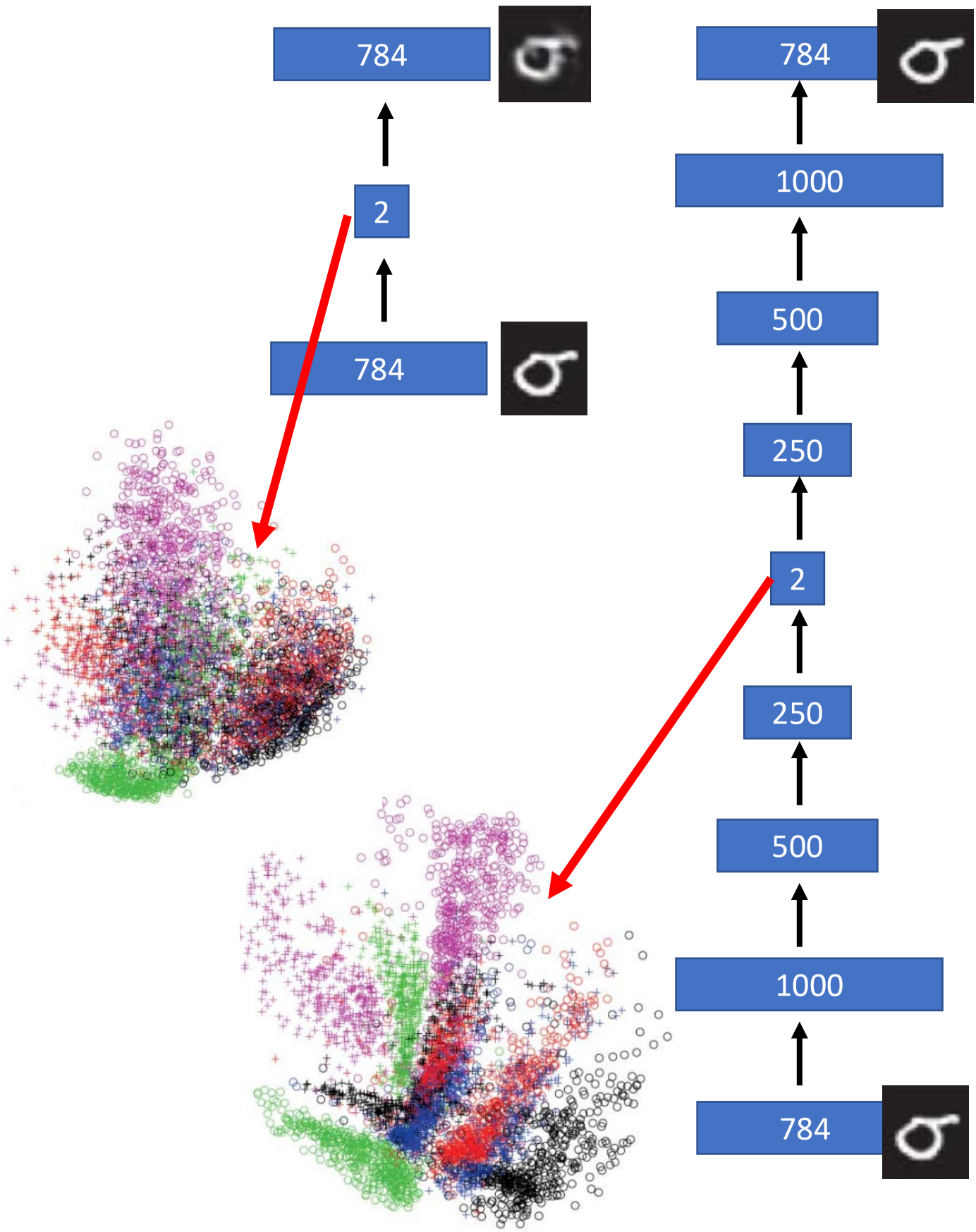
Original Image

PCA

Deep Auto-encoder





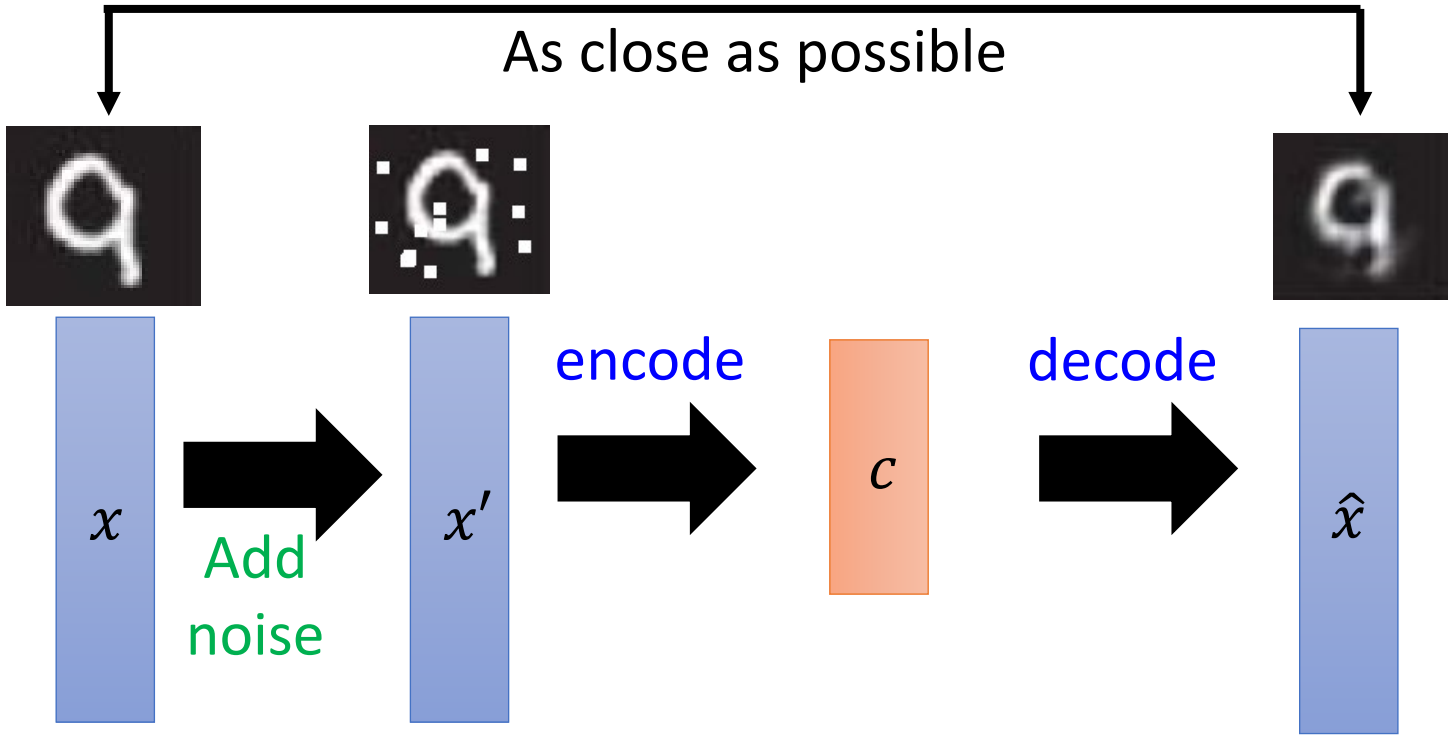

784
2
784
784
1000
500
250
2
250
500
1000
784

Auto-encoder

More: Contractive auto-encoder

Ref: Rifai, Salah, et al. "Contractive auto-encoders: Explicit invariance during feature extraction. " *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

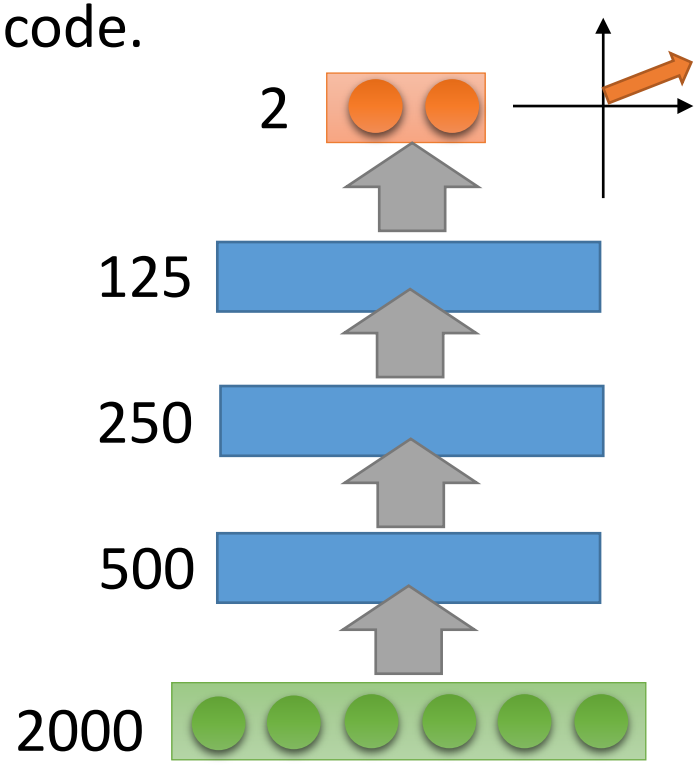
- De-noising auto-encoder



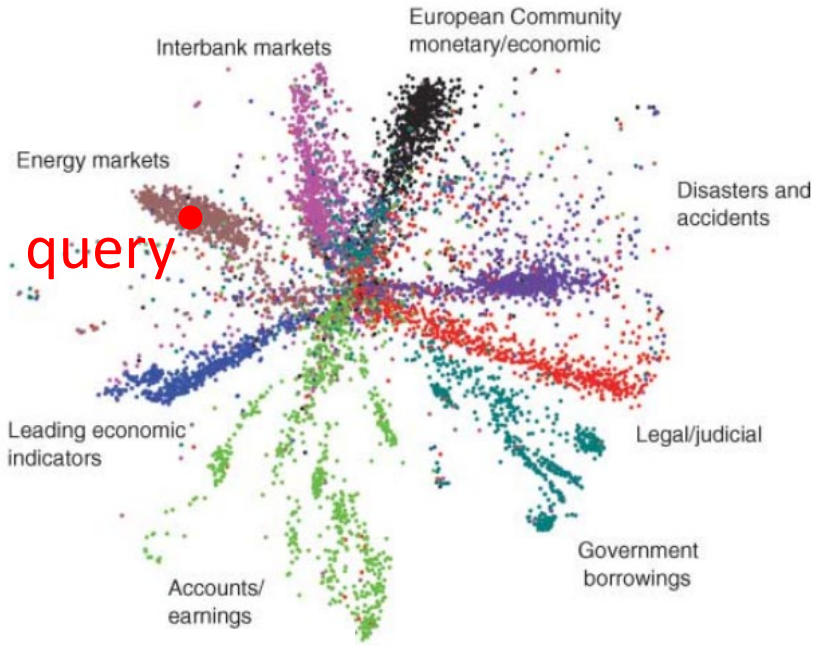
Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *ICML,* 2008.

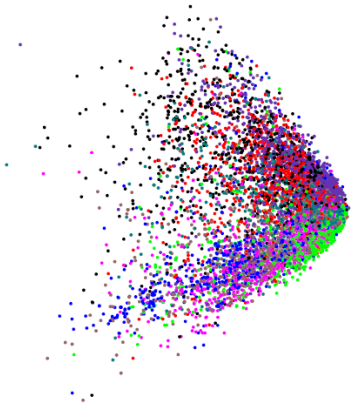
Auto-encoder – Text Retrieval

The documents talking about the same thing will have close code.



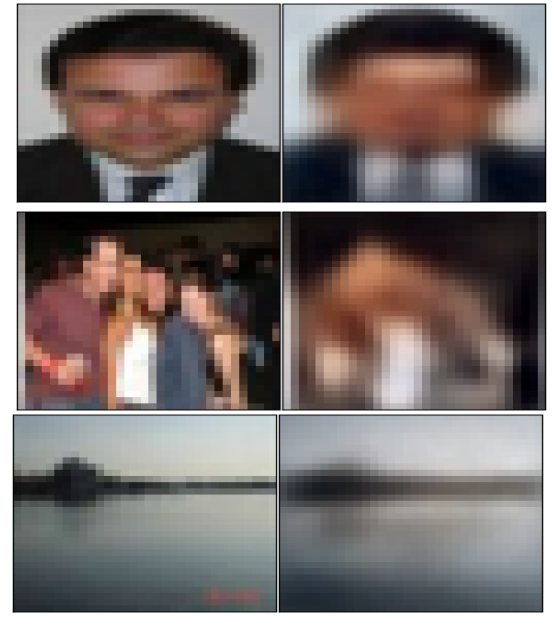
Bag-of-word
(document or query)

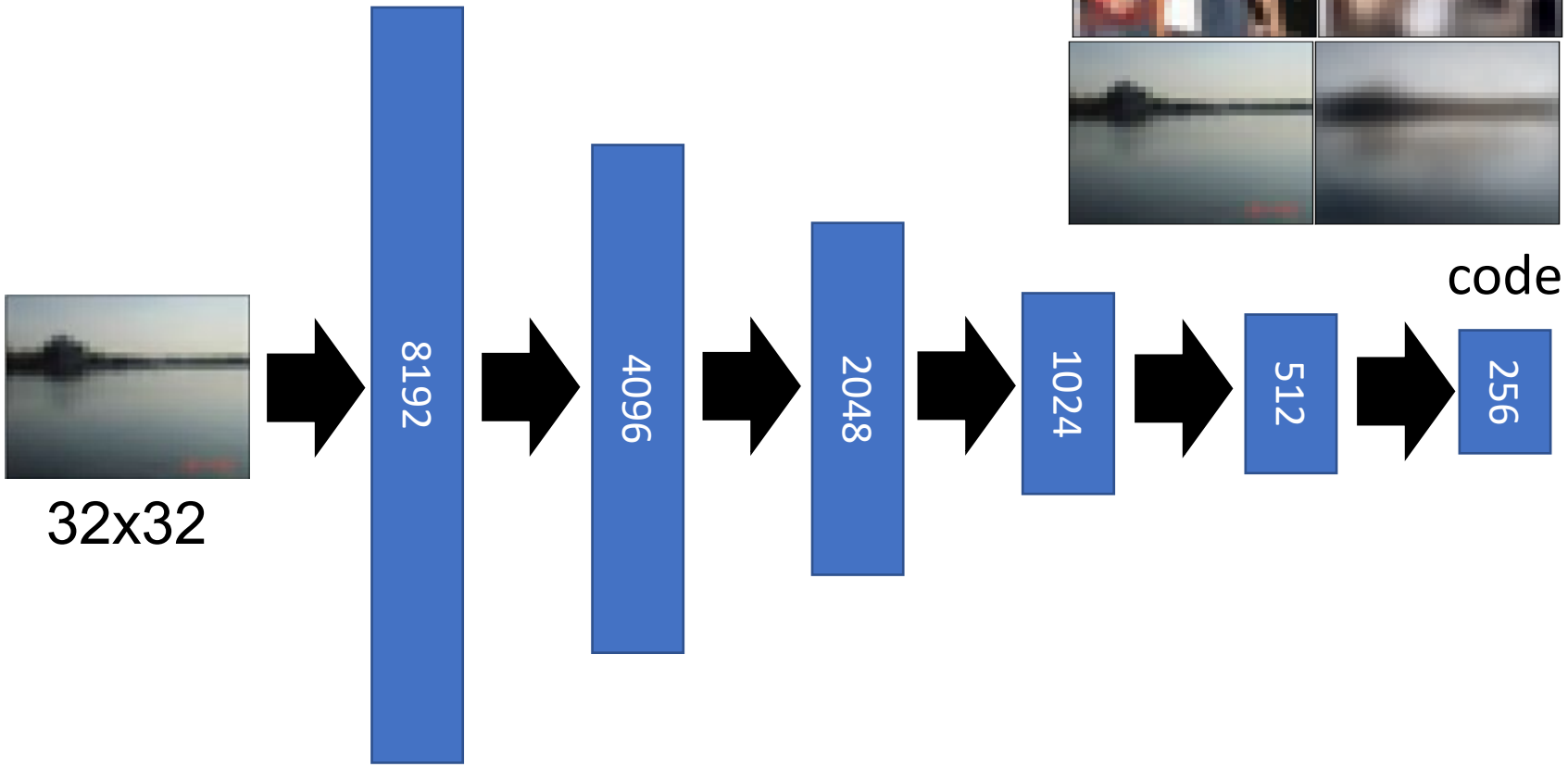




PCA: project documents to 2 latent topics

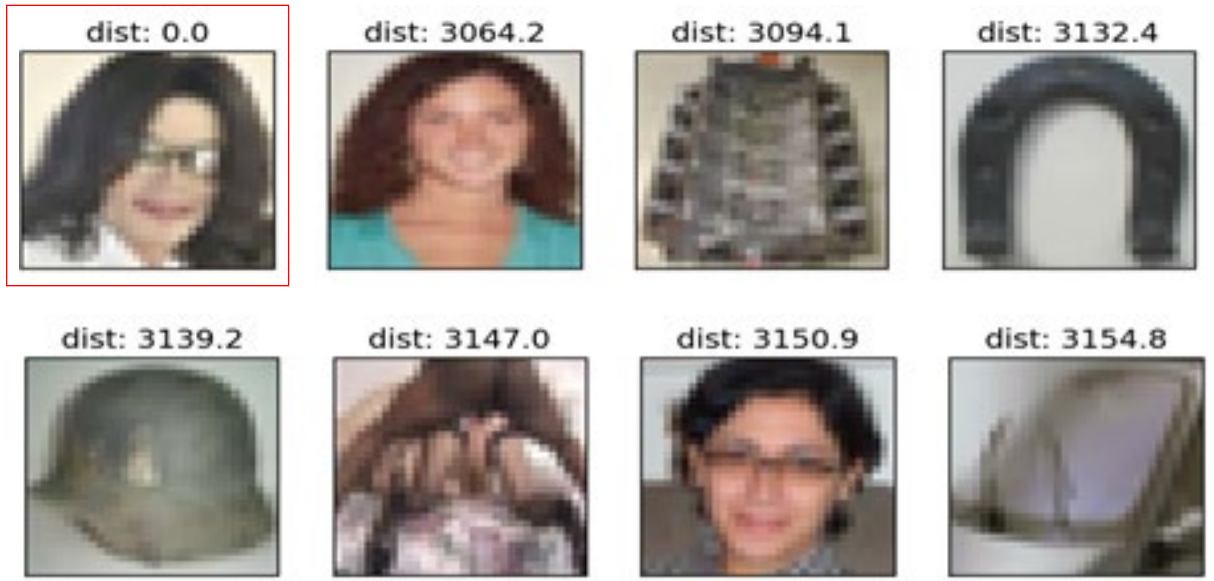
Auto-encoder – Similar Image Search





(crawl millions of images from the Internet)

Retrieved using Euclidean distance in pixel intensity space



retrieved using 256 codes

Embeddings

Embeddings

An embedding is a mapping of a discrete — categorical — variable to a vector of continuous numbers.

In the context of neural networks, embeddings are *low-dimensional, learned* continuous vector representations of discrete variables.

Neural network embeddings are useful because they can *reduce the dimensionality* of categorical variables and *meaningfully represent* categories in the transformed space.

Embeddings

Neural network embeddings have 3 primary purposes:

- Finding nearest neighbors in the embedding space. These can be used to make recommendations based on user interests or cluster categories.
- As input to a machine learning model for a supervised task.
- For visualization of concepts and relations between categories.

What is the simplest embedding?

- One-hot encoding
- Simplest label encoding method. Not learned representation
- Sparse
- Suboptimal representation:
  - For high-cardinality variables — those with many unique categories — the dimensionality of the transformed vector becomes unmanageable.
  - The mapping is completely uninformed: “similar” categories are not placed closer to each other in embedding space.

Learning Embeddings

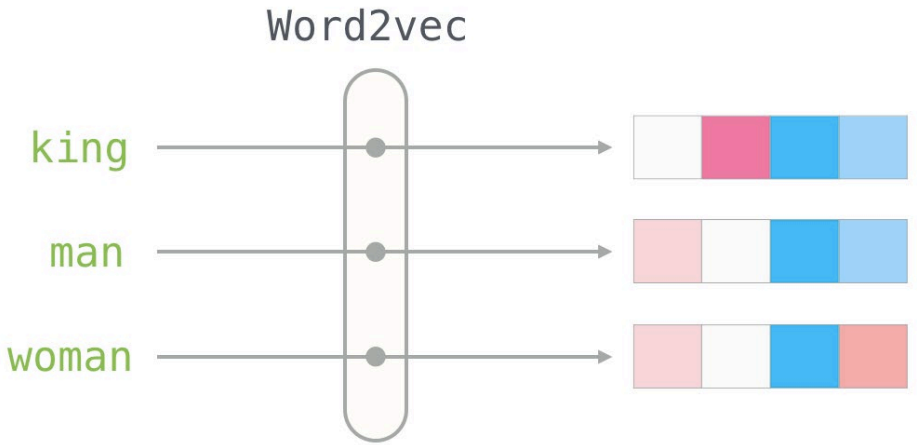
The main issue with one-hot encoding is that the transformation does not rely on any supervision.

We can greatly improve embeddings by ***learning*** them using a neural network on a supervised task.

The resulting embedded vectors are representations of categories where similar categories — relative to the task — are closer to one another.

Word2Vec

Distributed vector representation for words



Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

Word2Vec Assumptions

The more often two words co-occur, the closer their vectors will be

Two words have close meanings if their local neighborhoods are similar

Maps of words (trained on the same dataset) should be similar for each language => can build translators!

Word2Vec training & Sliding window

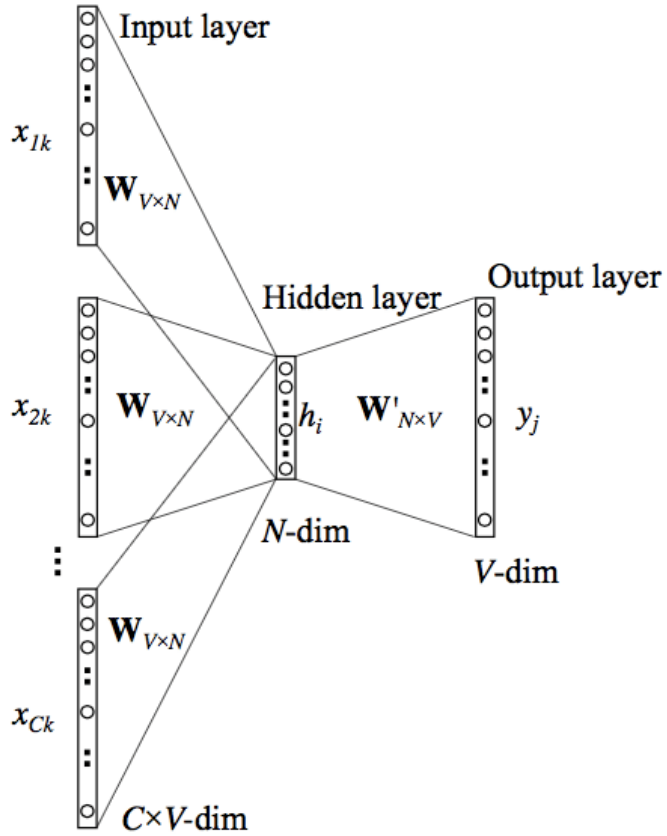
#1	natural Xᴋ	language Y(c=1)	processing Y(c=2)	and	machine	learning	is	fun	and	exciting	#1
#2	natural Y(c=1)	language Xᴋ	processing Y(c=2)	and Y(c=3)	machine	learning	is	fun	and	exciting	#2
#3	natural Y(c=1)	language Y(c=2)	processing Xᴋ	and Y(c=3)	machine Y(c=4)	learning	is	fun	and	exciting	#3
#4	natural	language Y(c=1)	processing Y(c=2)	and Xᴋ	machine Y(c=3)	learning Y(c=4)	is	fun	and	exciting	#4
#5	natural	language	processing Y(c=1)	and Y(c=2)	machine Xᴋ	learning Y(c=3)	is Y(c=4)	fun	and	exciting	#5
#6	natural	language	processing	and Y(c=1)	machine Y(c=2)	learning Xᴋ	is Y(c=3)	fun Y(c=4)	and	exciting	#6
#7	natural	language	processing	and	machine Y(c=1)	learning Y(c=2)	is Xᴋ	fun Y(c=3)	and Y(c=4)	exciting	#7
#8	natural	language	processing	and	machine	learning Y(c=1)	is Y(c=2)	fun Xᴋ	and Y(c=3)	exciting Y(c=4)	#8
#9	natural	language	processing	and	machine	learning	is Y(c=1)	fun Y(c=2)	and Xᴋ	exciting Y(c=3)	#9
#10	natural	language	processing	and	machine	learning	is	fun Y(c=1)	and Y(c=2)	exciting Xᴋ	#10

#	Token	#1			#2				#3					#4					#5				
0	natural	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2	processing	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
3	and	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
4	machine	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
5	learning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
6	is	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	fun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	exciting	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		Xᴋ	Y(c=1)	Y(c=2)	Xᴋ	Y(c=1)	Y(c=2)	Y(c=3)	Xᴋ	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	Xᴋ	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	Xᴋ	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)

#	Token	#6					#7					#8					#9				#10		
0	natural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	processing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	and	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
4	machine	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	learning	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	is	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
7	fun	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
8	exciting	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0
		Xᴋ	Y(c=1)	Y(c=2)	(c=3	Y(c=4)	Xᴋ	Y(c=1)	(c=2	Y(c=3)	Y(c=4)	Xᴋ	Y(c=1)	(c=2	Y(c=3)	Y(c=4)	Xᴋ	Y(c=1)	(c=2	Y(c=3)	Xᴋ	Y(c=1)	Y(c=2)

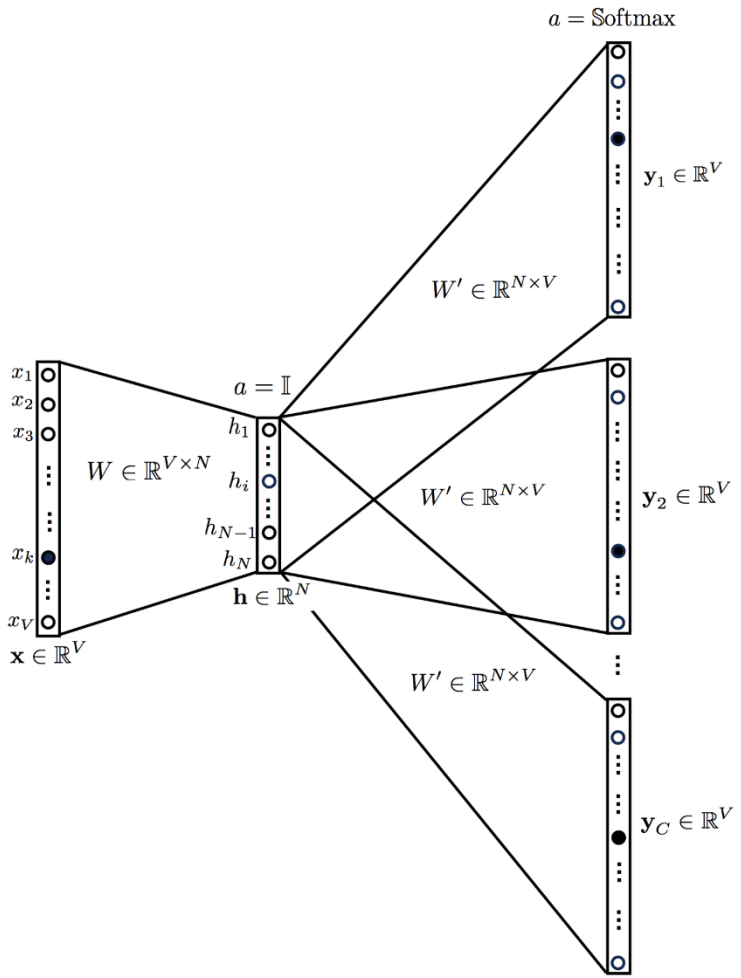
derekchia.com

Continuous BoW (CBOW) Model



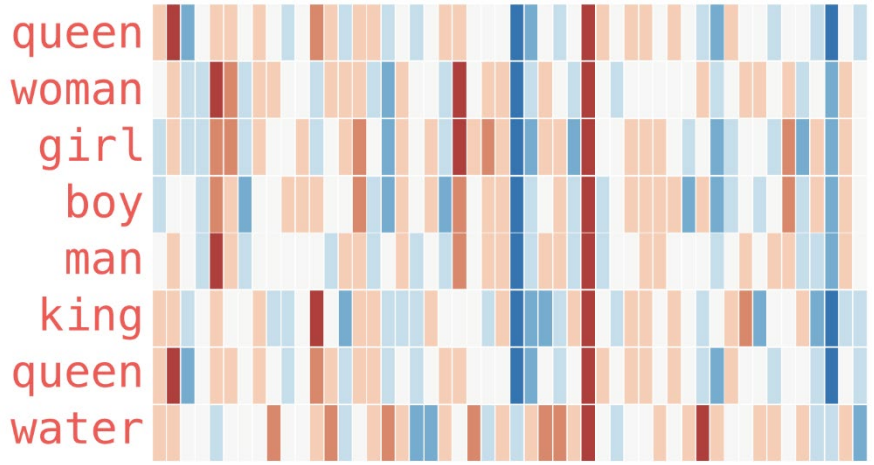
The **CBOW** model architecture tries to predict the current target word (the center word) based on the source context words (surrounding words).

Skip-gram (SG) Model

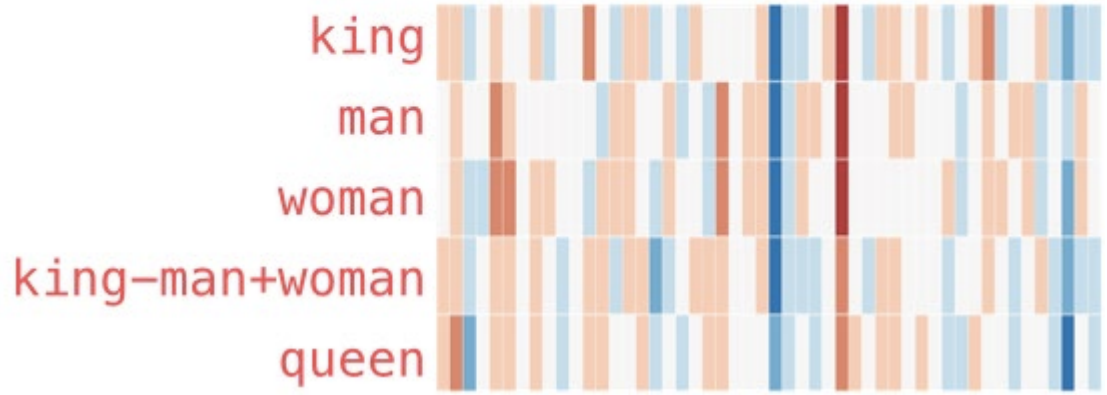


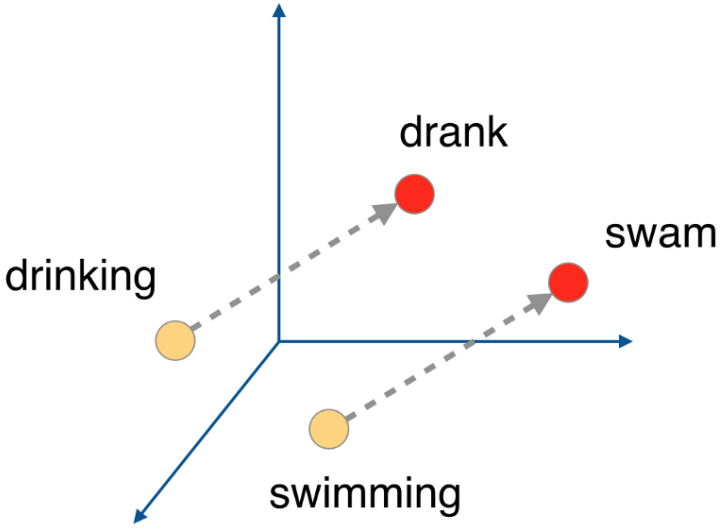
GS it takes every word in a large corpora (we will call it the focus word) and also takes one-by-one the words that surround it within a defined ‘window’ to then feed a neural network that after training will predict the probability for each word to actually appear in the window around the focus word.

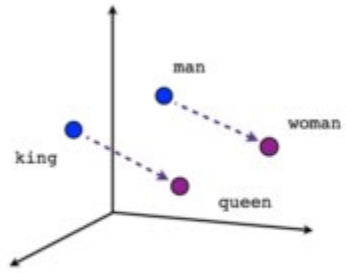
Vector Operations



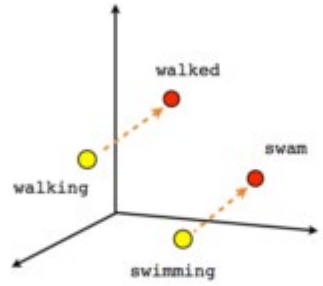
king − man + woman ~= queen



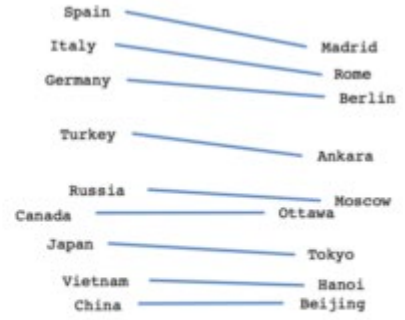




Male-Female

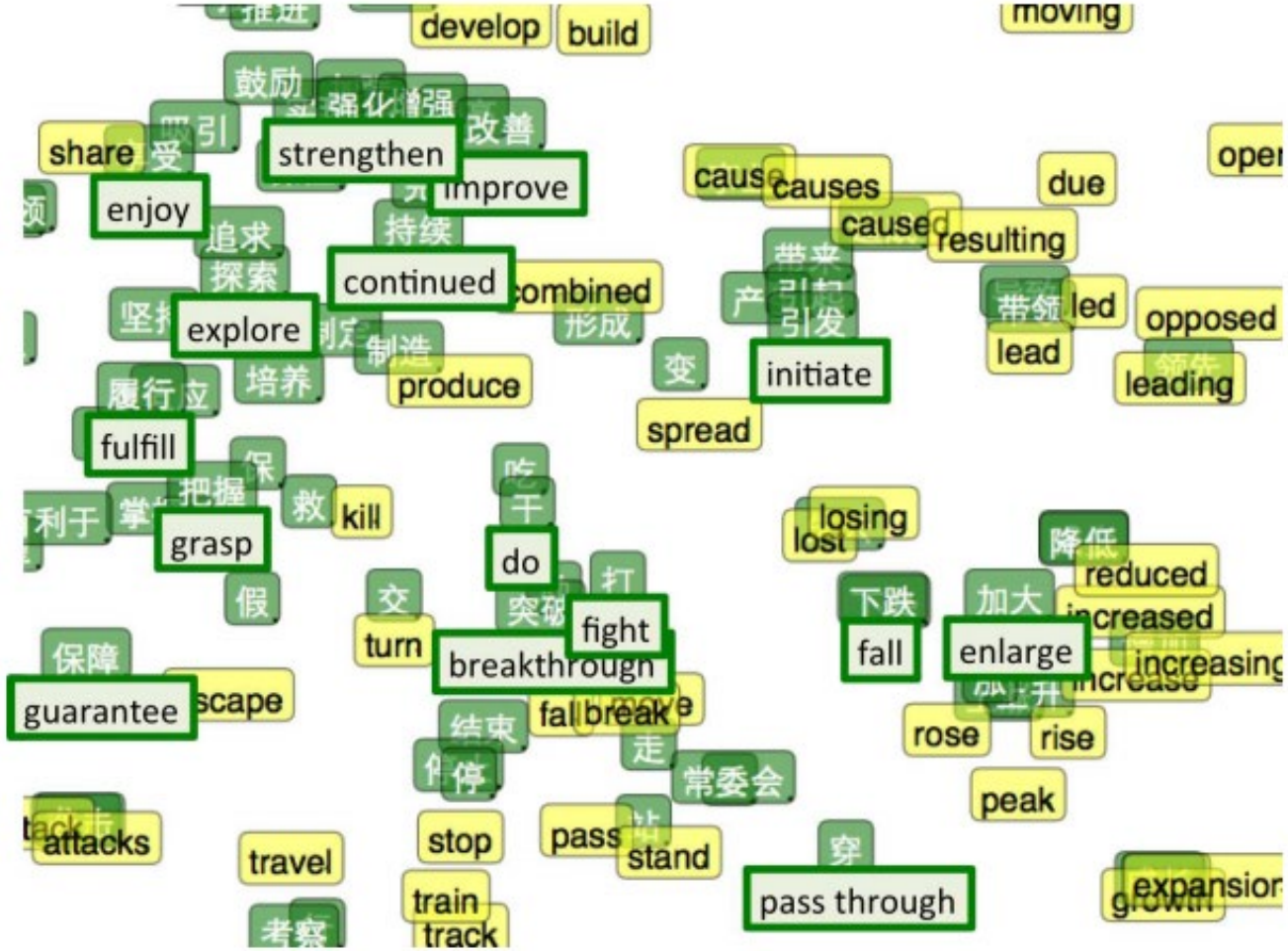


Verb tense



Country-Capital

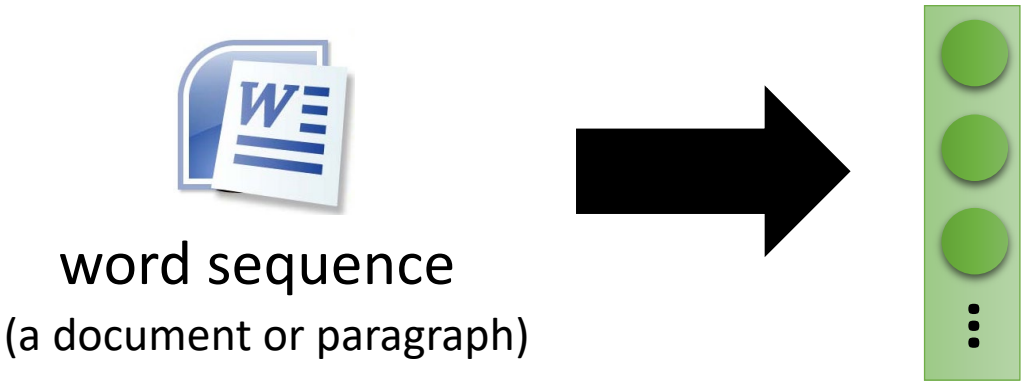
Multi-lingual Embedding

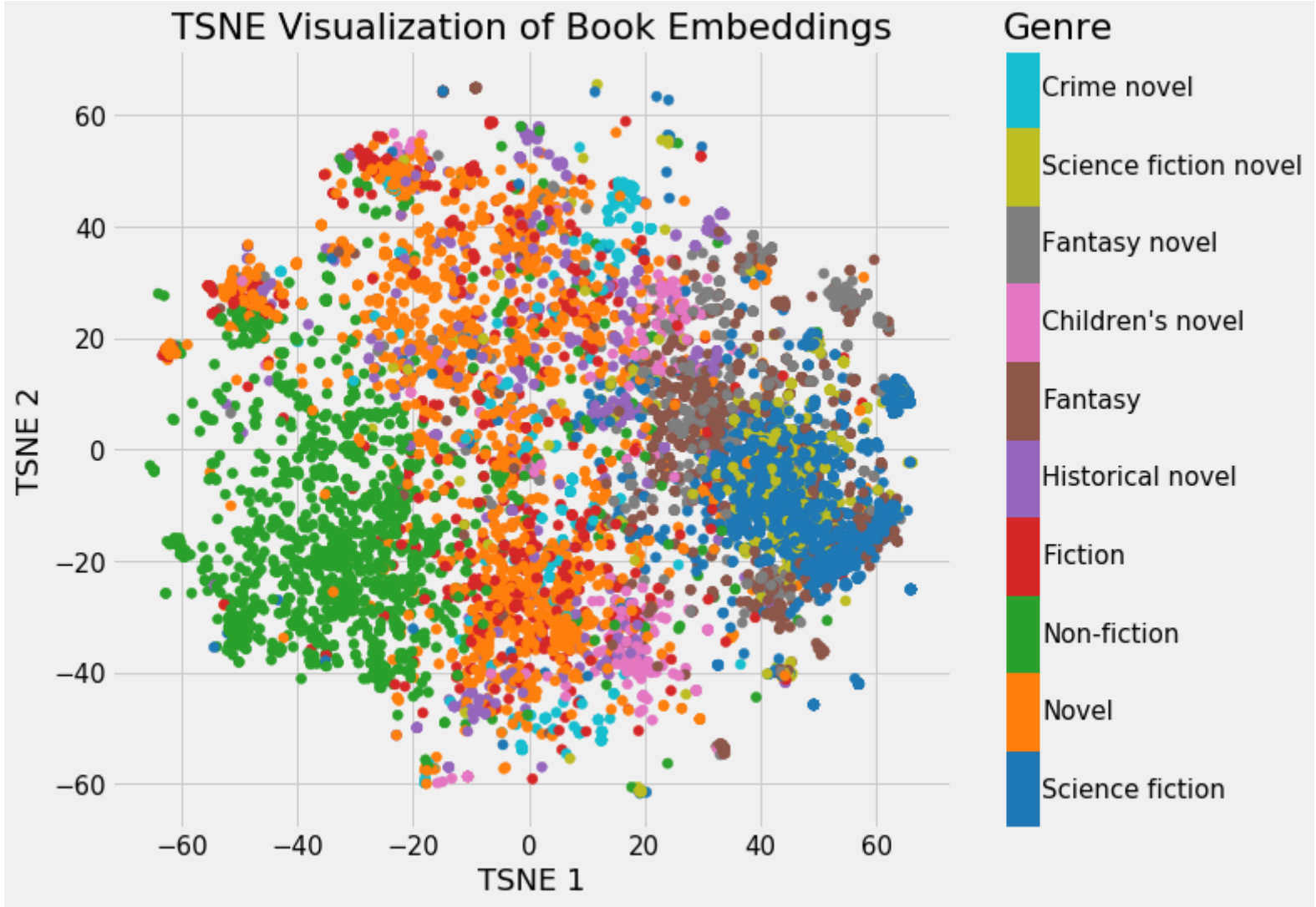


Bilingual Word Embeddings for Phrase-Based Machine Translation, Will Zou, Richard Socher, Daniel Cer and Christopher Manning, EMNLP, 2013

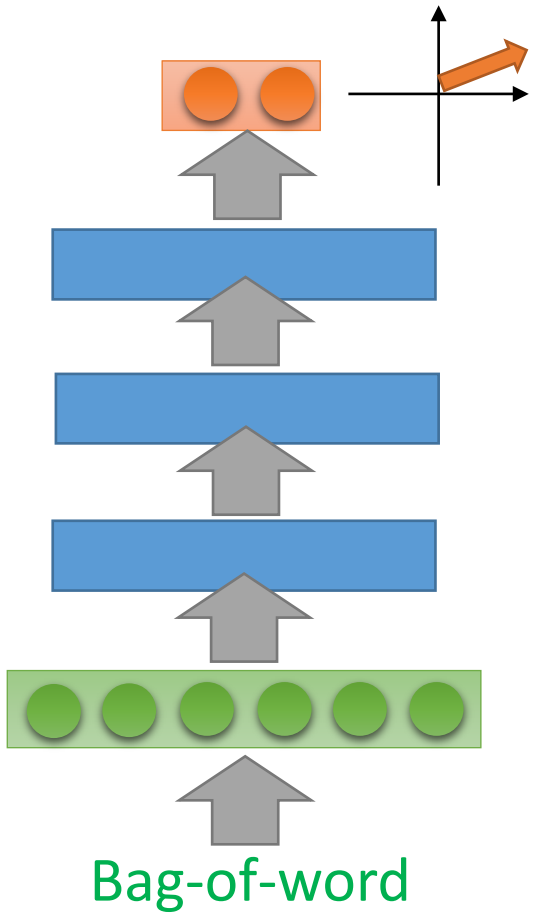
Document Embedding

- word sequences with different lengths → the vector with the same length
  - The vector representing the meaning of the word sequence
  - A word sequence can be a document or a paragraph

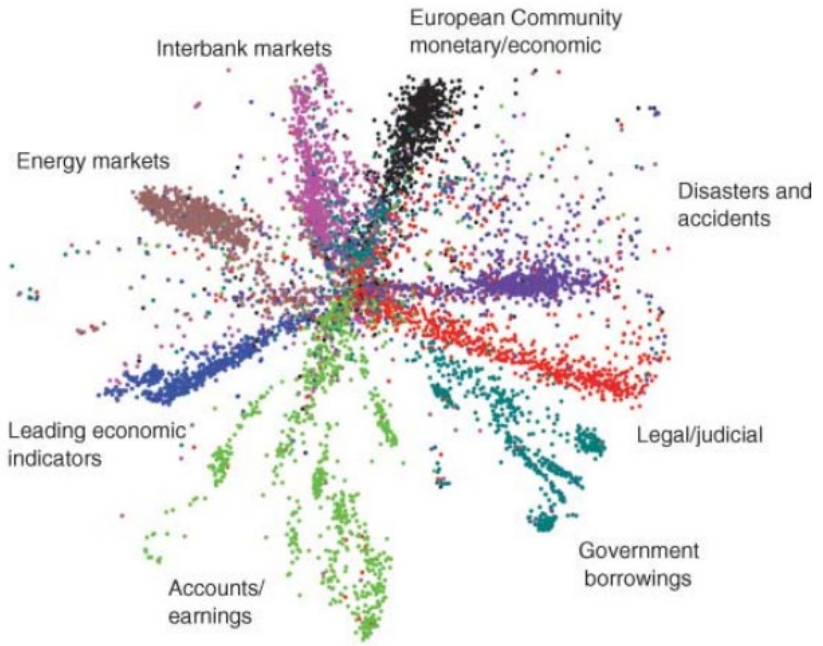



TSNE Visualization of Book Embeddings
TSNE 2
TSNE 1
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
Genre
Crime novel
Science fiction novel
Fantasy novel
Children's novel
Fantasy
Historical novel
Fiction
Non-fiction
Novel
Science fiction

Semantic Embedding

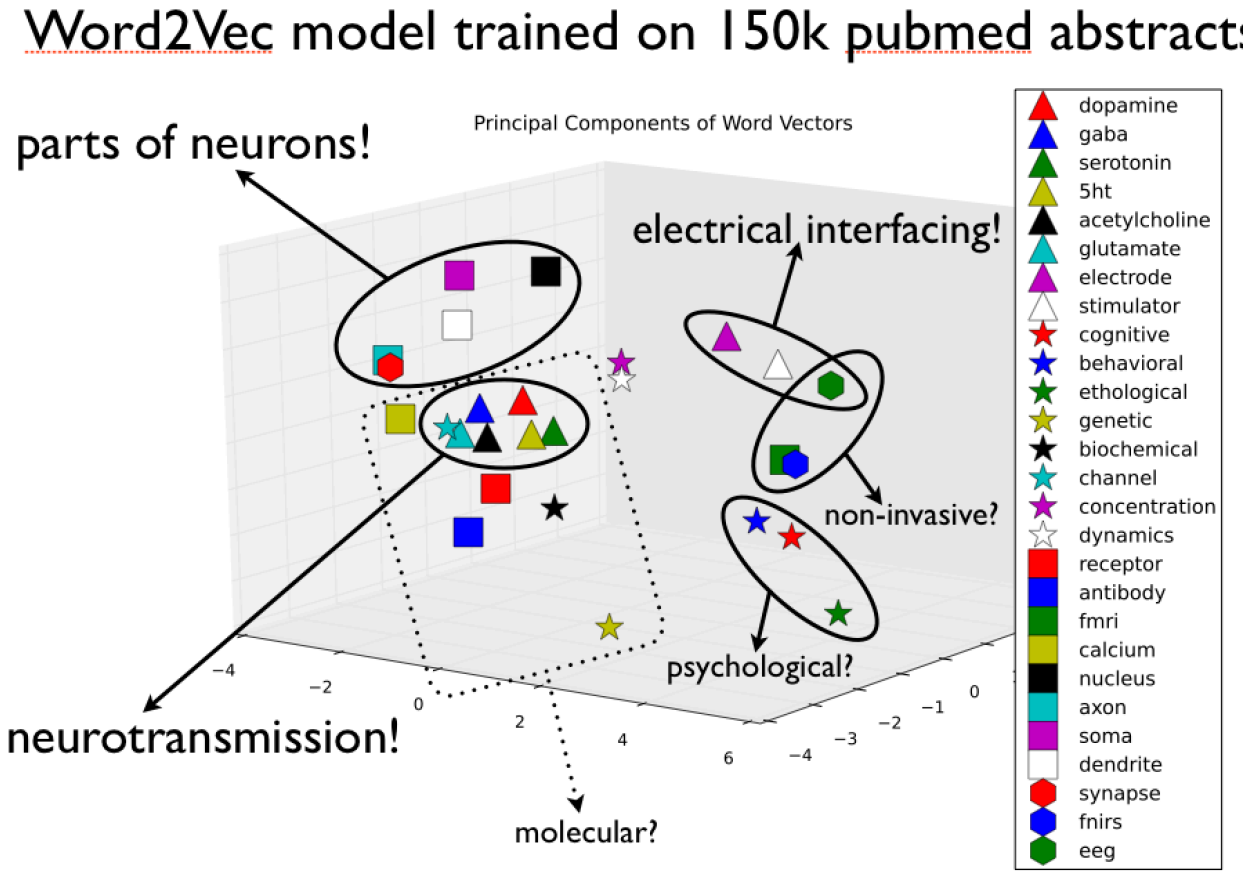


Bag-of-word



Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

Semantic space: Pubmed abstract analysis

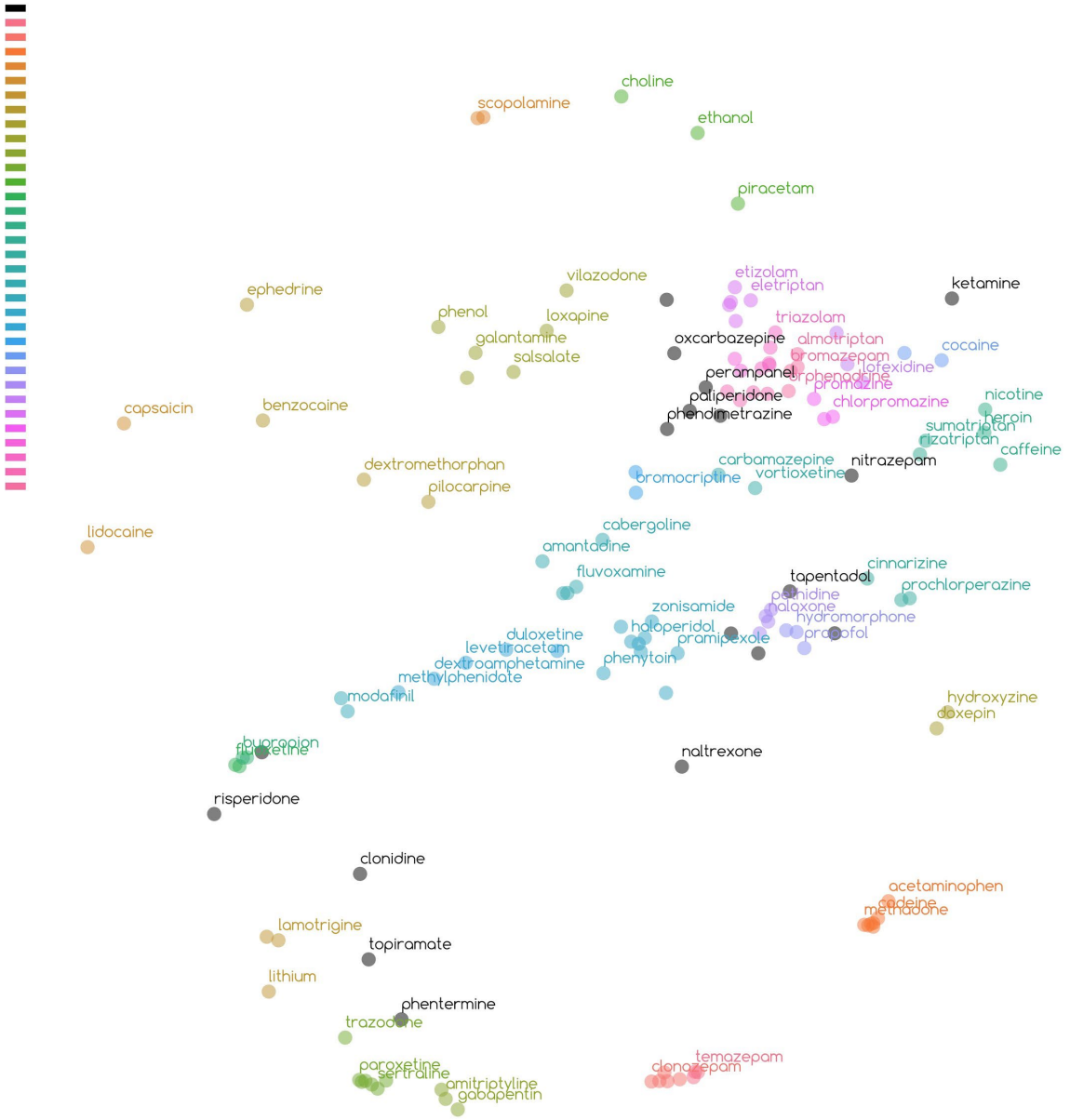


Credit to: http://web.mit.edu/amarbles/www/neuro_word2vec.html

Drug clustering in semantic similarity space

- Sources of user comments:

webmd.com
patient.info
drugs.com
amazon.com
askapatient.com dailystrength.org

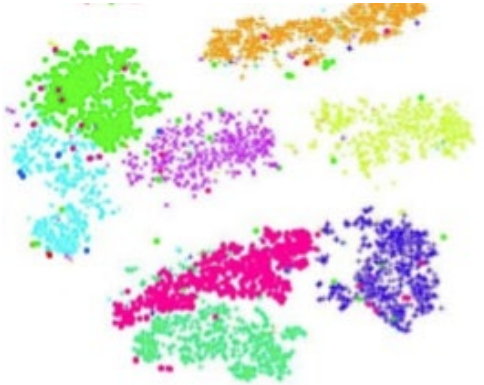


Tutubalina et al, 2020

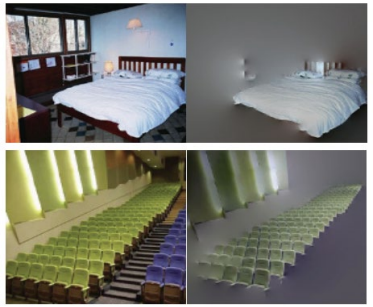
Understanding and Probing Representations

Understanding Representations

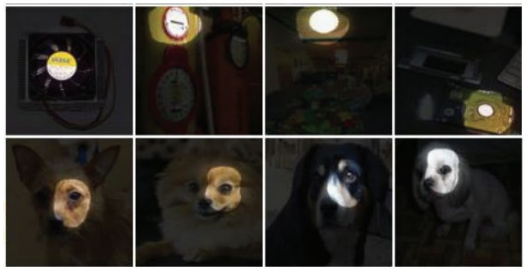
- Why do we need this?
- Tools:
  - Through the window of output domain: e.g. t-SNE



  - Through the window of input domain: e.g. Minimal Image.

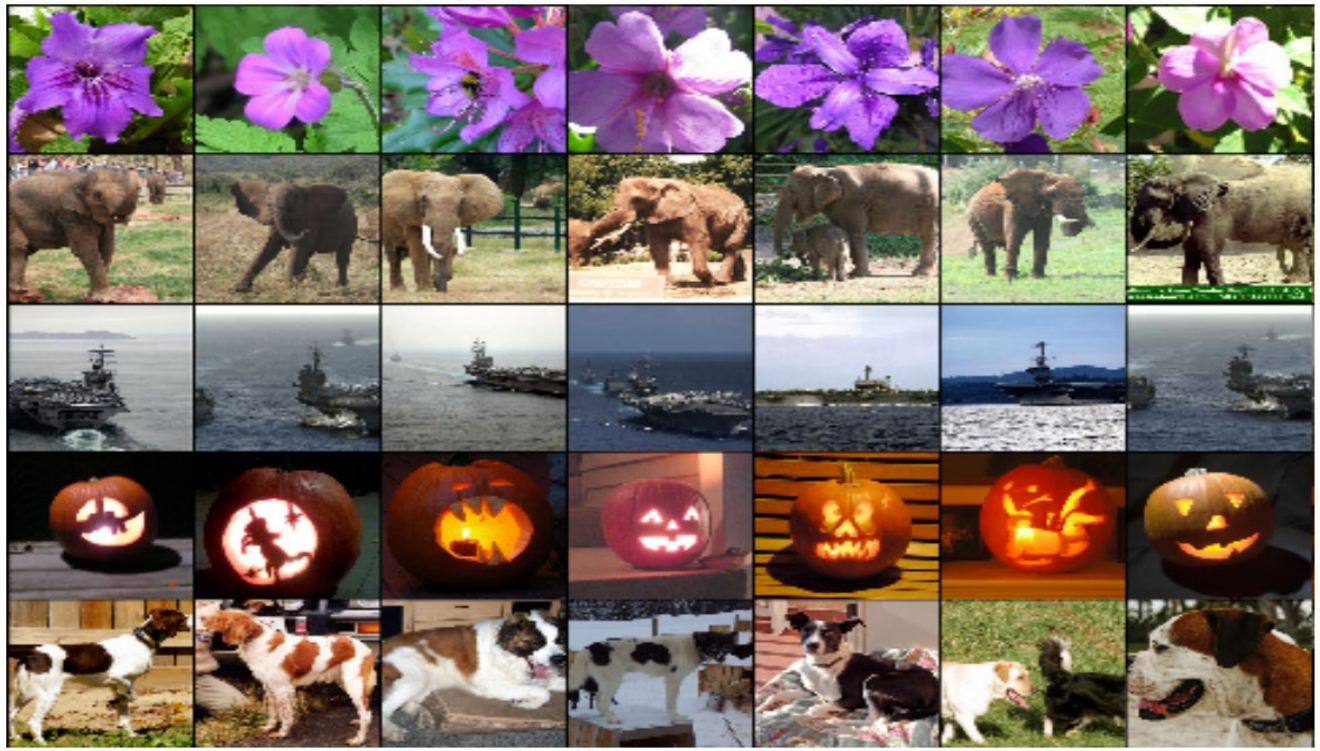


  - Through the window of computation elements: e.g. Receptive fields



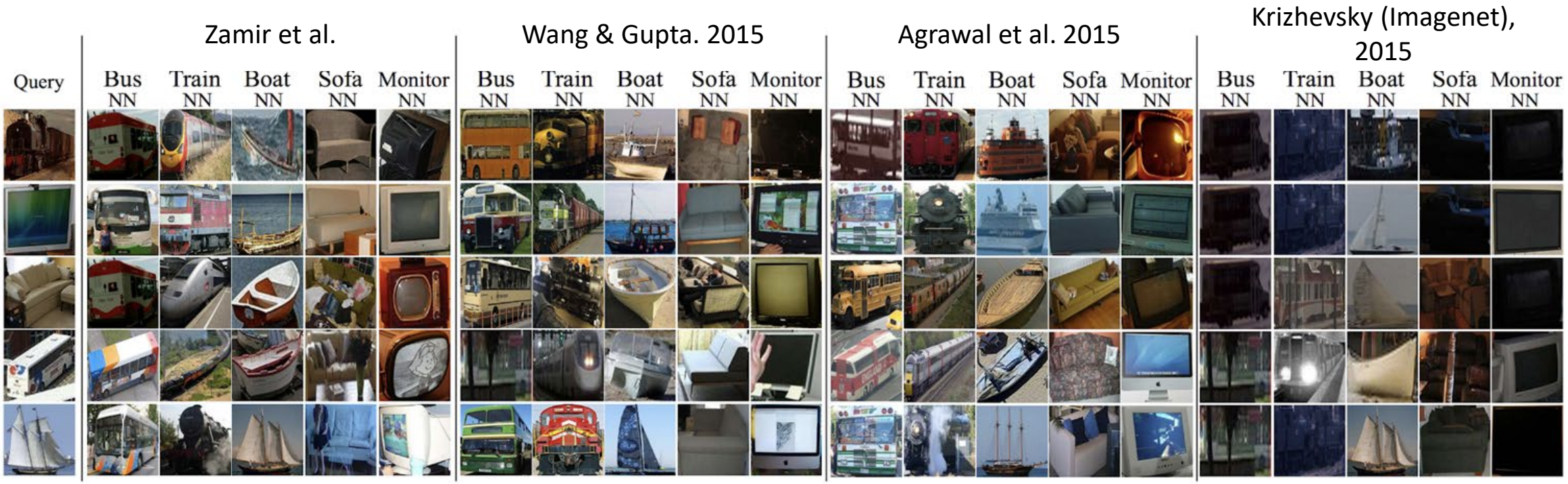
Nearest neighbors in full dimensional space

Query | Nearest Neighbors



Krizhevsky et al. 2012.
Deng et al. 2009.

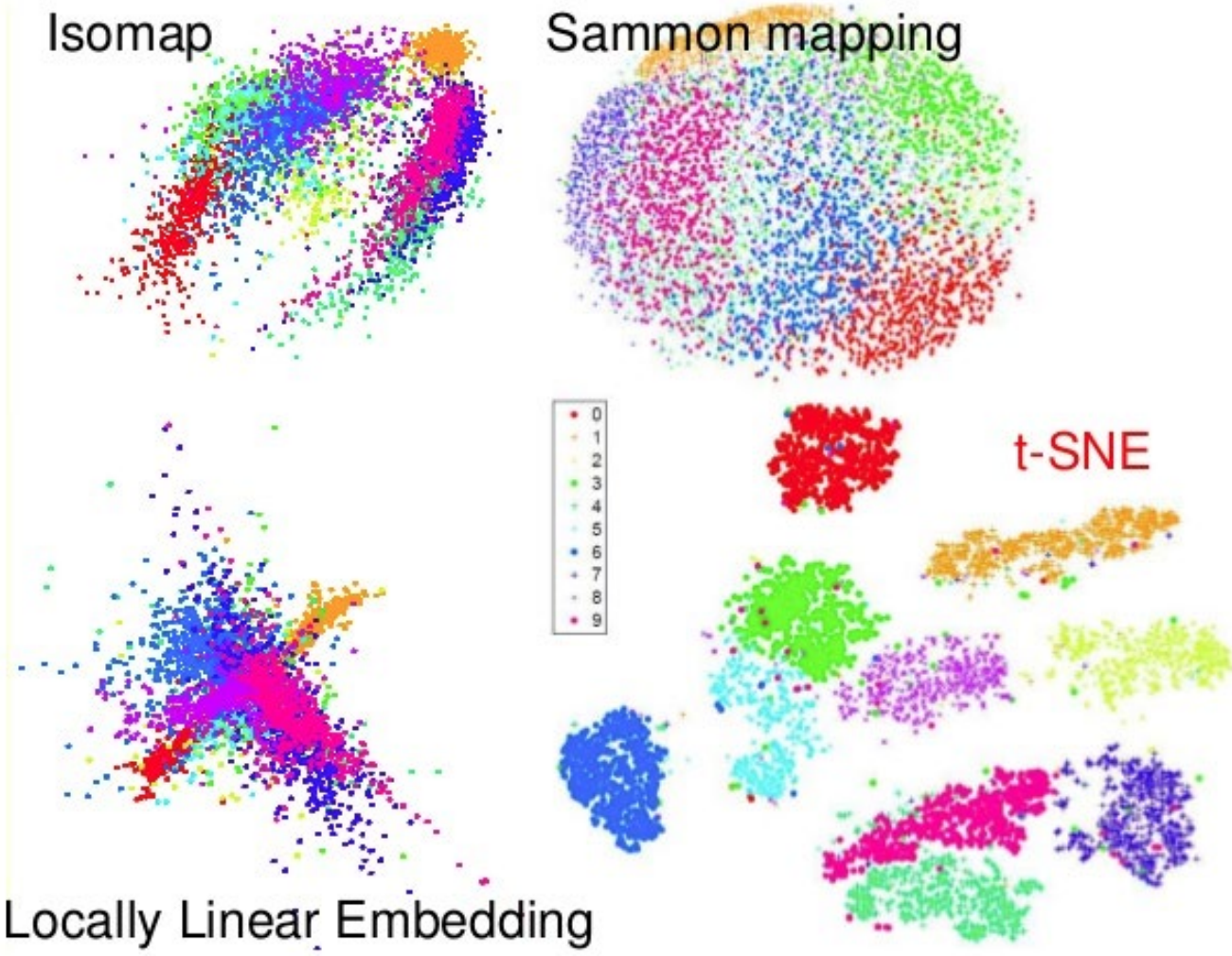
Nearest neighbors in full dimensional space



Zamir et al. 2016

Low-dimensional embeddings

- 6000 MNIST Digits
  - tSNE
  - Isomap
  - Sammon M
  - LLE

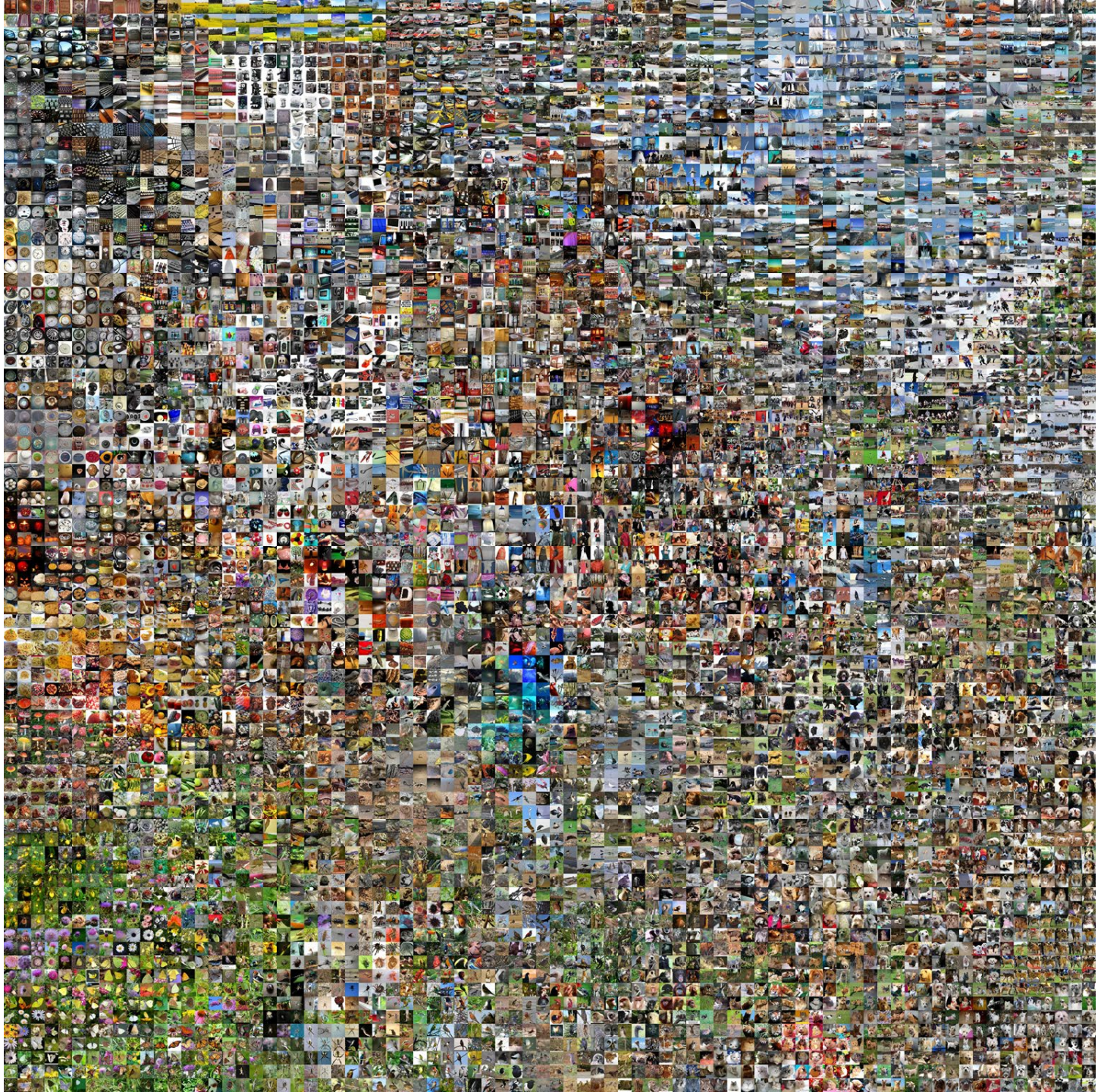


Van der Maaten & Hinton. 2008

Low-dimensional embeddings

- tSNE

Van der Maaten & Hinton. 2008



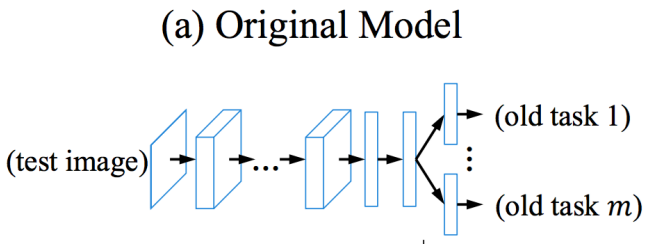
Van der Maaten & Hinton. 2008

Methods of Mixing Representations

Mixing Representations

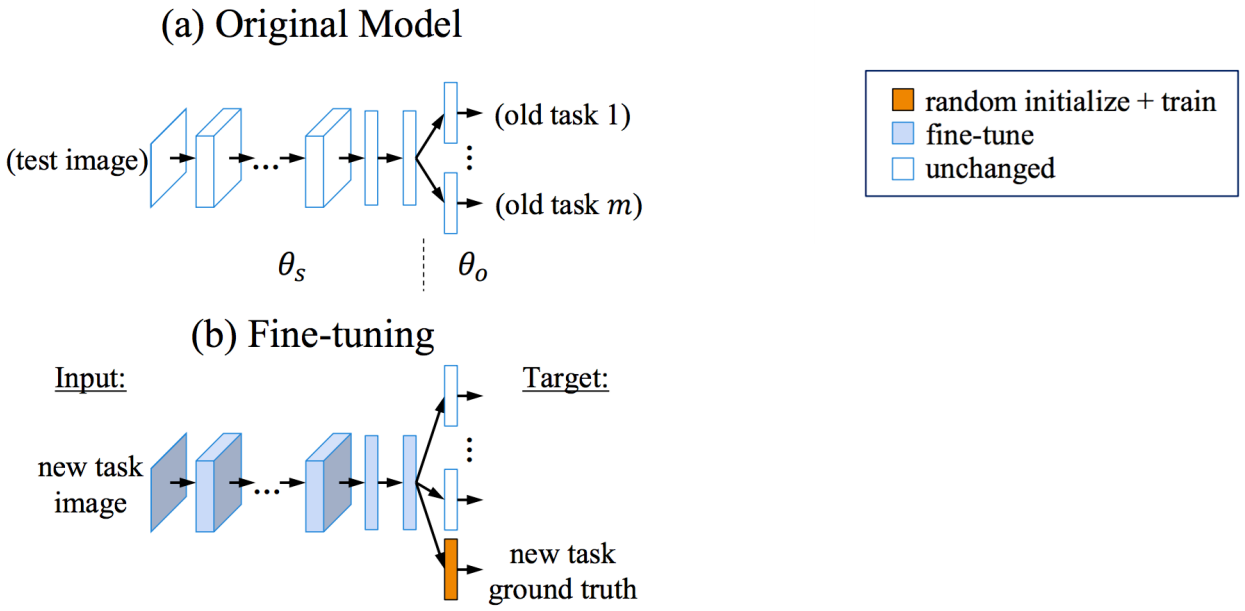
- **Sometimes you wish to mix two/multiple tasks/representations**
  - To expand
  - To transfer information or labeled data across tasks
  - To form a multi-task representation
  - To form a better single-task representation

Mixing Representations - How?



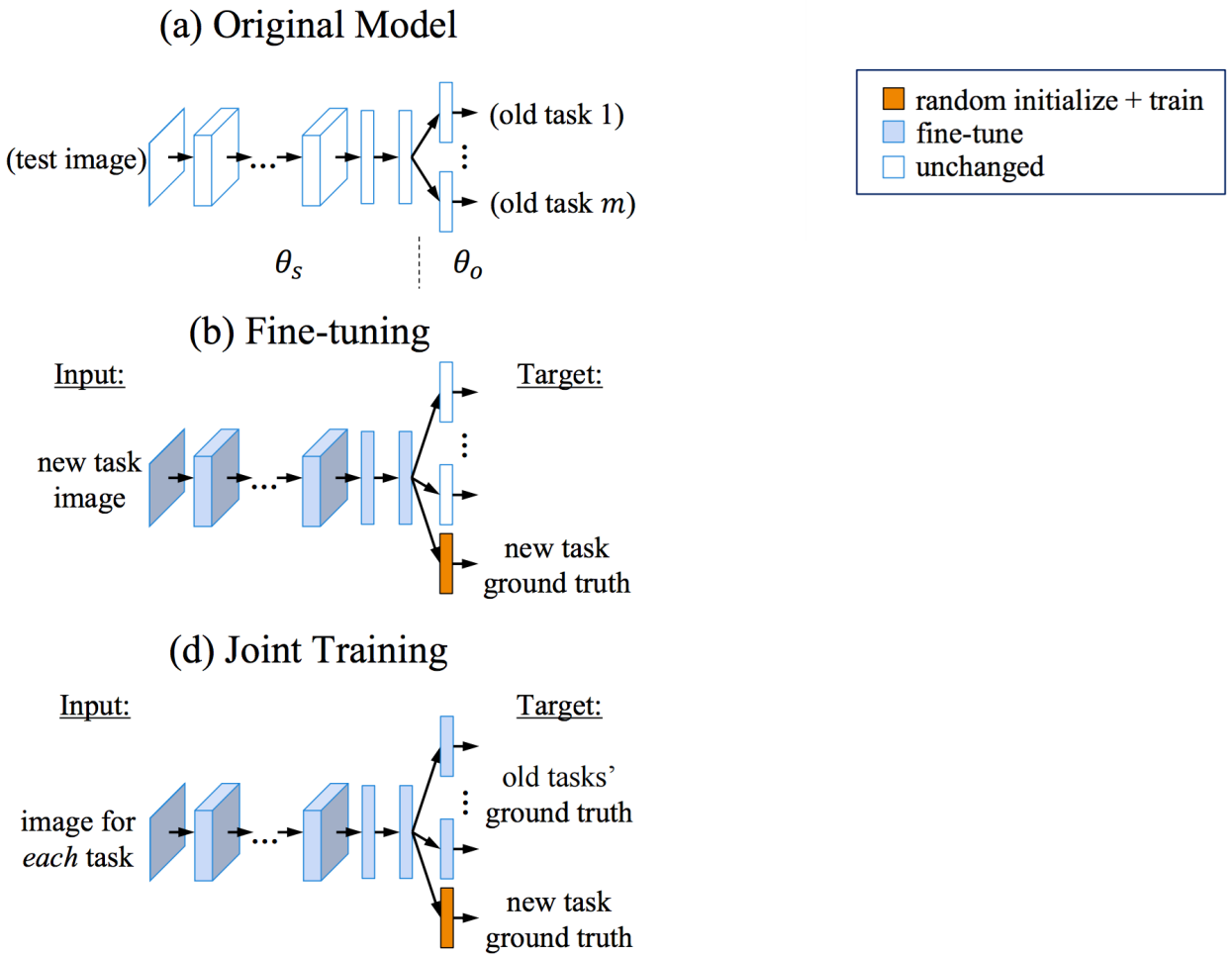
Li & Hoiem. 2016.

Mixing Representations - fine tuning



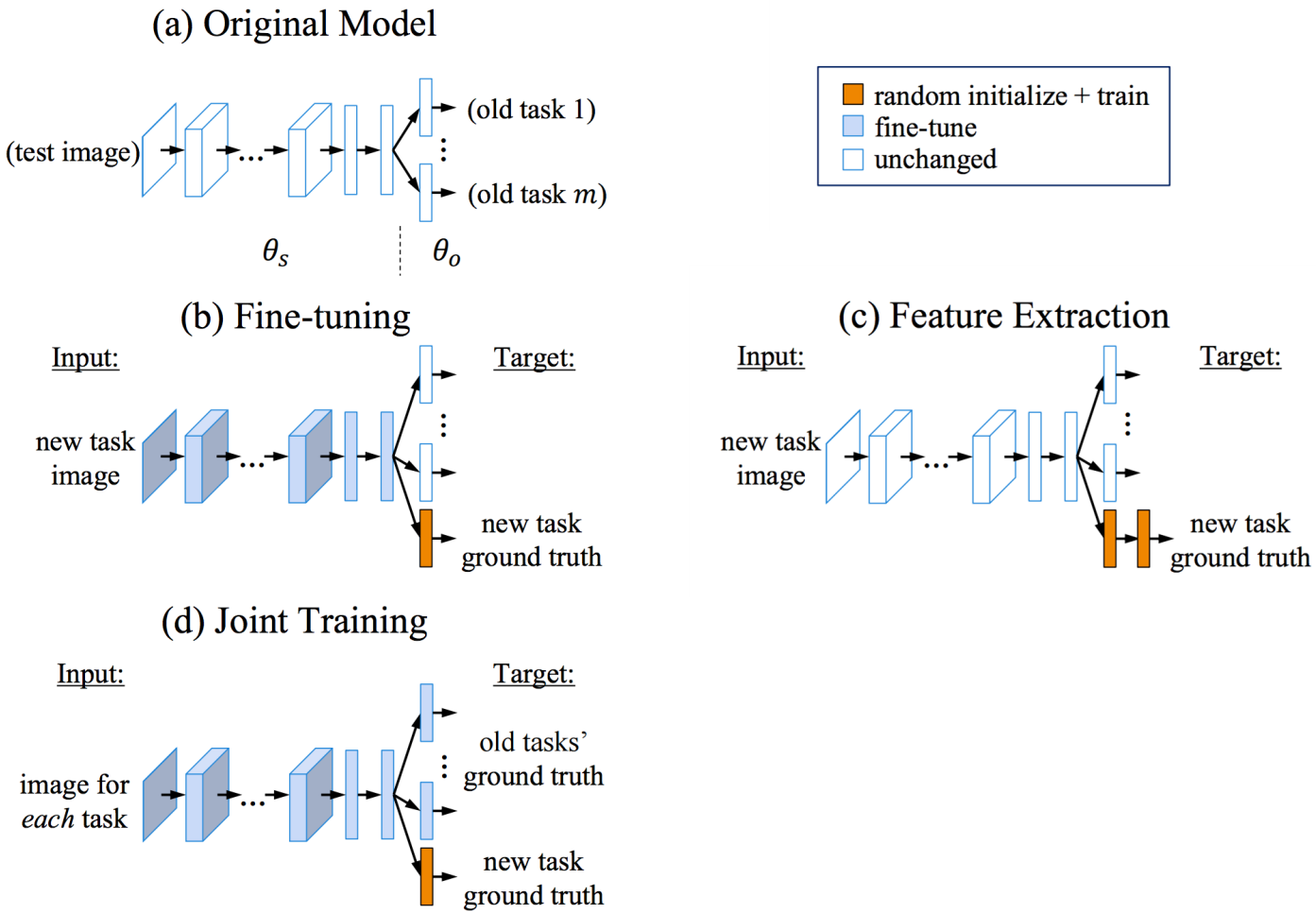
Li & Hoiem. 2016.

Mixing Representations - joint (multi-task) training



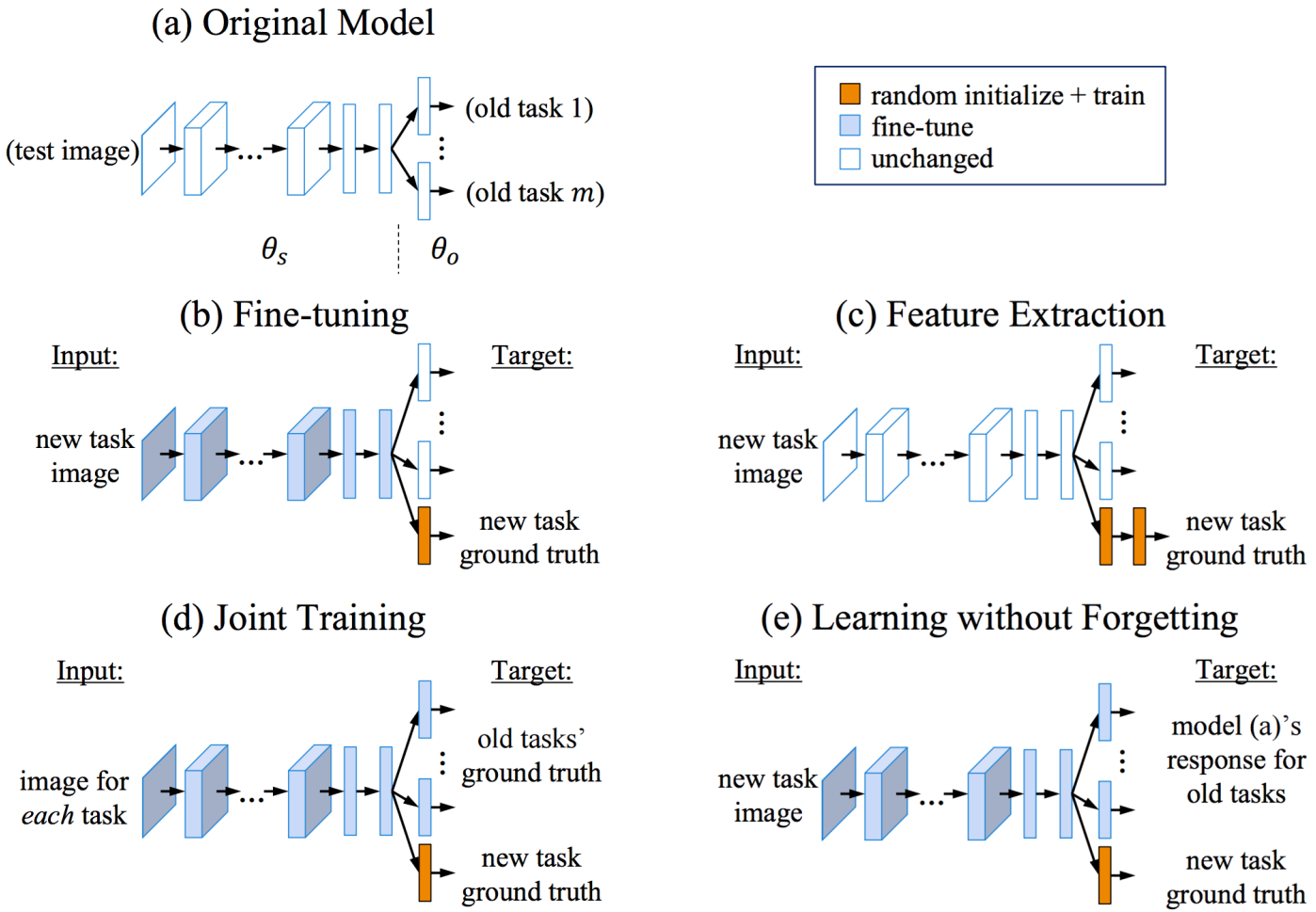
Li & Hoiem. 2016.

Mixing Representations - feature extraction



Li & Hoiem. 2016.

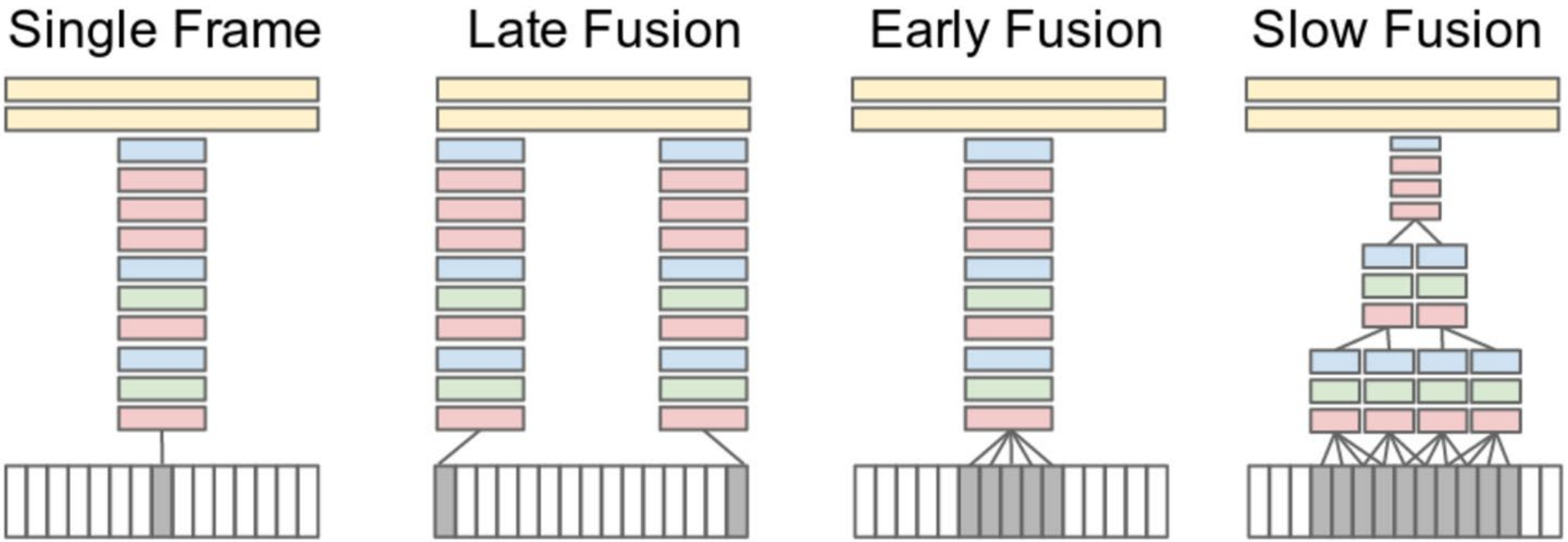
Mixing Representations - LwF



Li & Hoiem ECCV'16

Li & Hoiem. 2016.

Video processing - frame fusion



Karpathy et al. 2015.