Dimensionality Reduction & Visualization

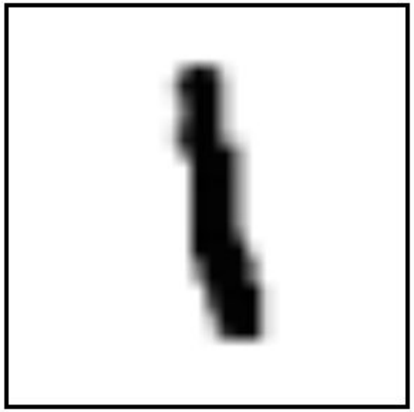
Data Mining & Analytics

Prof. Zach Pardos

INFO 254/154: Spring ’19

Dimensionality Reduction examples

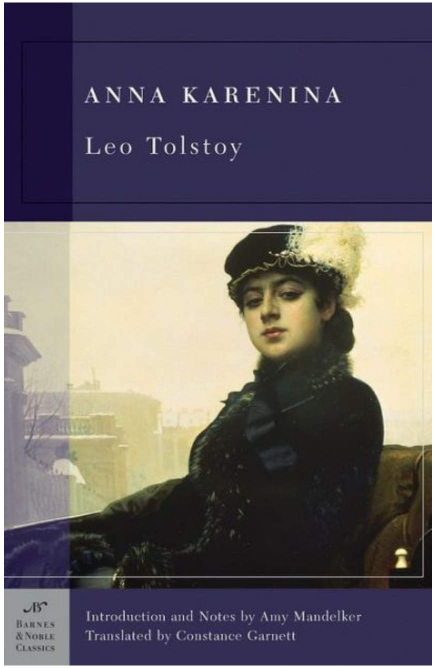
Images



$$\simeq \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .6 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .7 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9 & 1 & .1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & 1 & .1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \simeq [\ 0.23\ 0.51\ ]$$

Dimensionality Reduction examples

Text


ANNA KARENINA
Leo Tolstoy
Introduction and Notes by Amy Mandelker
Translated by Constance Garnett
BARNES & NOBLE CLASSICS

≃ [0 0 0 1 1 0 0 1 0 0 0 … 1 0] ≃ [0.42 -0.93]

Dimensionality Reduction examples

Collaborative filtering (item ratings)

[Transformers 2 (5 star), Fight club (5 star), Trainspotting (4 star), Particle Fever (5 star) ...] $\simeq$ [0 1 0 … 5 5 4 5 … 0 0 0 0 3] $\simeq$ [0.2 -0.31 0.50]

Dimensionality Reduction examples

Arbitrary sequences

[play_video1, answer_quiz2_correct, ...] ≃ [[1 0 0 0] [0 0 1 0] [0 1 0 0]...] ≃ [-0.85 0.67]

Dimensionality Reduction

Purposes for reduction

- Human Visualization to better understand the data, model, and domain
- Lower the number of parameters of a model by decreasing the input size
- Comparing similarity of input with respect to the most “important” features

Dimensionality Reduction

Techniques for dimensionality reduction

- Principal Component Analysis (PCA) - linear
- Autoencoder - nonlinear (neural network)
- t-Stochastic Neighbour Embedding (t-SNE) - nonlinear (used in lab 6b)

Feed-forward neural network

Autoencoder

Input size = 3
Output size = 3
Hidden size = 2

X = O =

0.20	0.10	-0.89

W_{xh} =

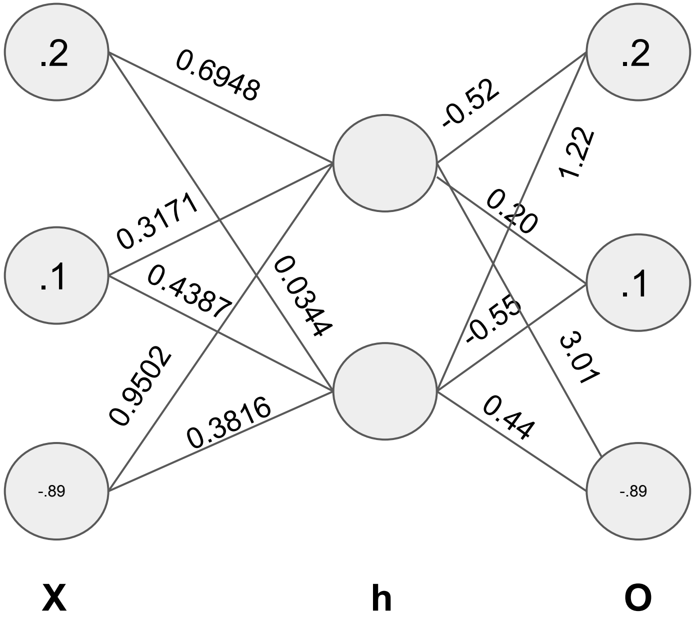
0.6948	0.0344
0.3171	0.4387
0.9502	0.3816

h =

0.3171	0.4387

W_{ho} =

-0.52	0.20	3.01
1.22	-0.55	0.44



Main distinctive features of an Autoencoder

- Input is the same as the output
- Center hidden layer dimension < input
- Network is symmetric

Dimensionality Reduction examples

Collaborative filtering (item ratings)

[Transformers 2 (5 star), Fight club (5 star), Trainspotting (4 star), Particle Fever (5 star) ...] ≃ [0 1 0 … 5 5 4 5 … 0 0 0 0 3] ≃ [0.2 -0.31 0.50]

Dimensionality Reduction for Recommendation

Training deep Autoencoders for collaborative filtering

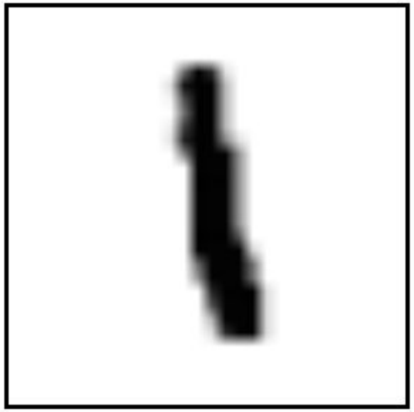
Oleksii Kuchaiev & Boris Ginsburg

NVIDIA.

[video link]

Dimensionality Reduction examples

Images



$$\simeq \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .6 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .9 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \simeq [\ 0.23\ 0.51\ ]$$

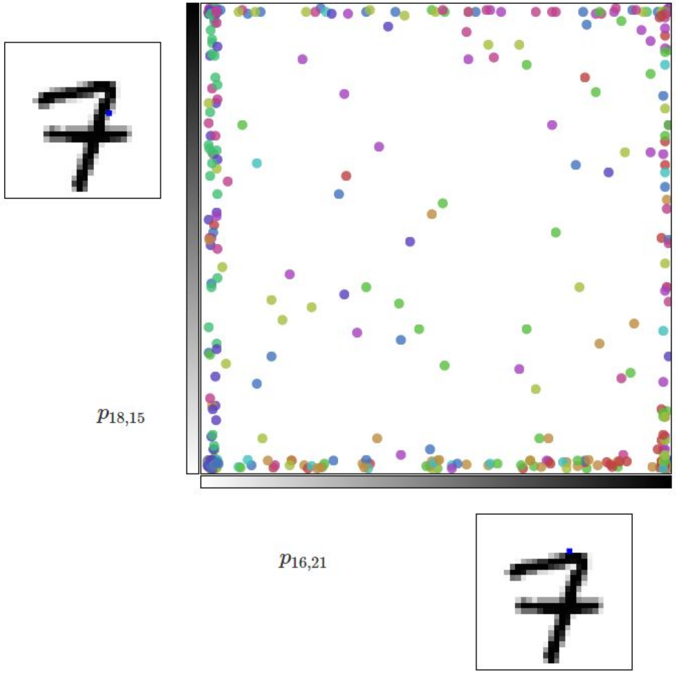
MNIST dataset

- 70,000 example images of digitized handwritten digits
- Digits are 0-9
- Each image is 28x28 (784 dimensions)

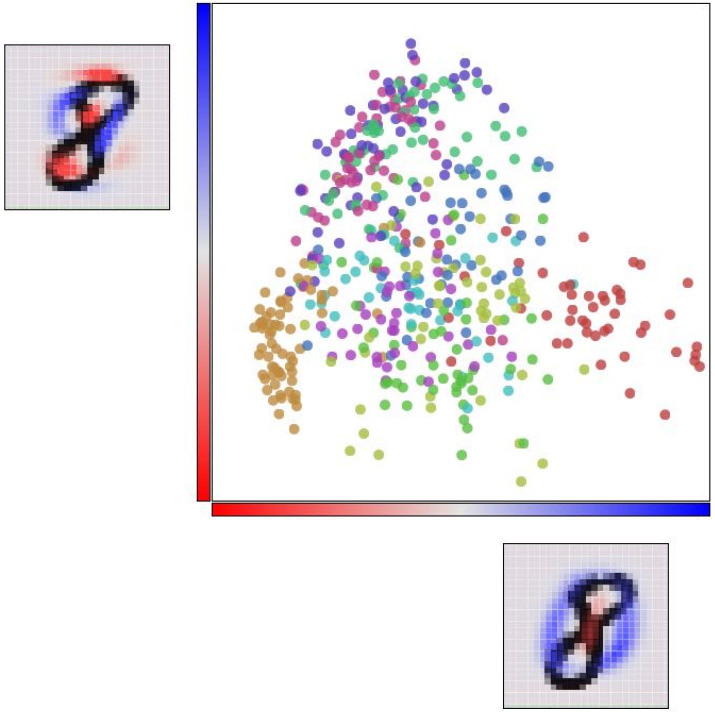
[Interactive tutorial]

Dimensionality Reduction examples

Using two arbitrary pixels to reduce to 2D



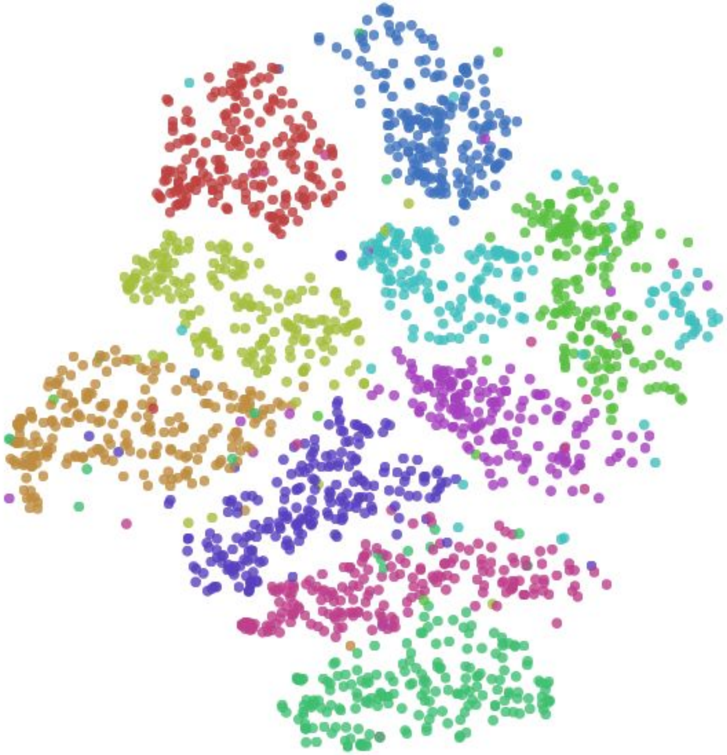
Using PCA to reduce to 2D



[Interactive tutorial]

Dimensionality Reduction examples

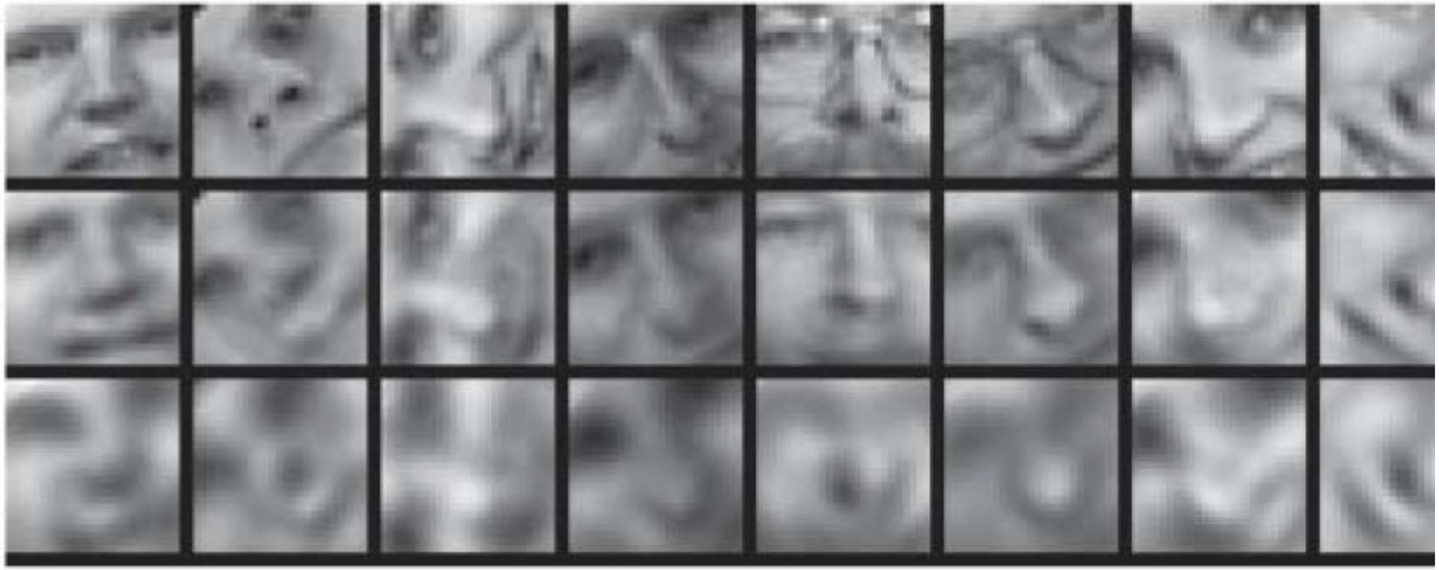
Using t-SNE to reduce to 2D

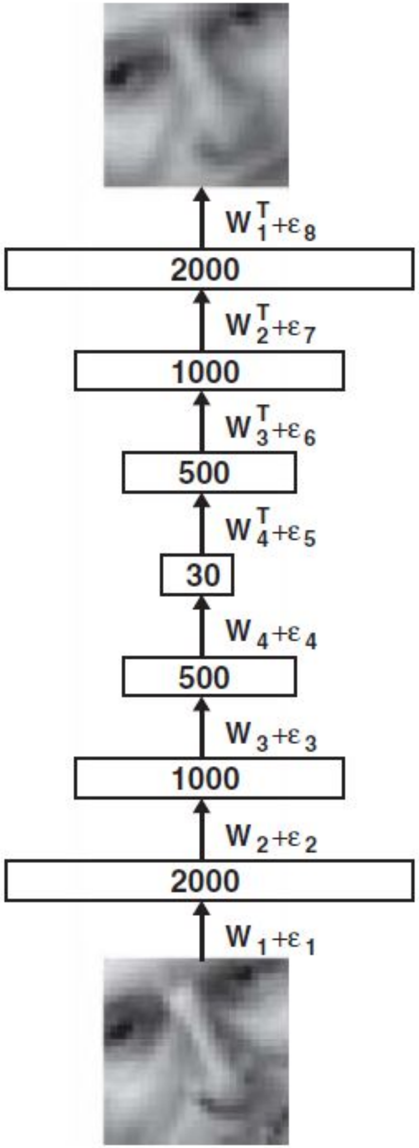


[Interactive tutorial]

Dimensionality Reduction examples

Images



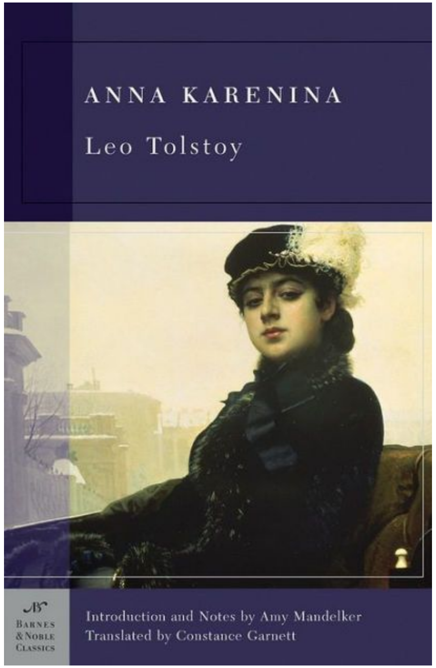


- Reconstructing images of faces
- First row is original, second row is Autoencoder, third row is PCA (both with size 30 dimensionality reduction)

[Hinton & Salakhutdinov, 2006]

Dimensionality Reduction examples

Text

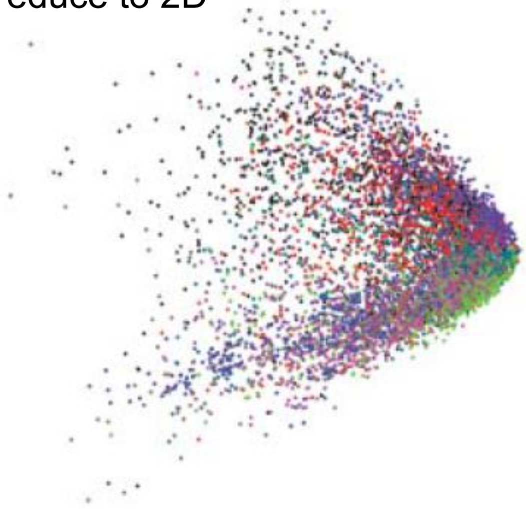

ANNA KARENINA
Leo Tolstoy
Introduction and Notes by Amy Mandelker
Translated by Constance Garnett
BARNES & NOBLE CLASSICS

$\simeq$ [0 0 0 1 1 0 0 1 0 0 0 … 1 0] $\simeq$ [0.42 -0.93]

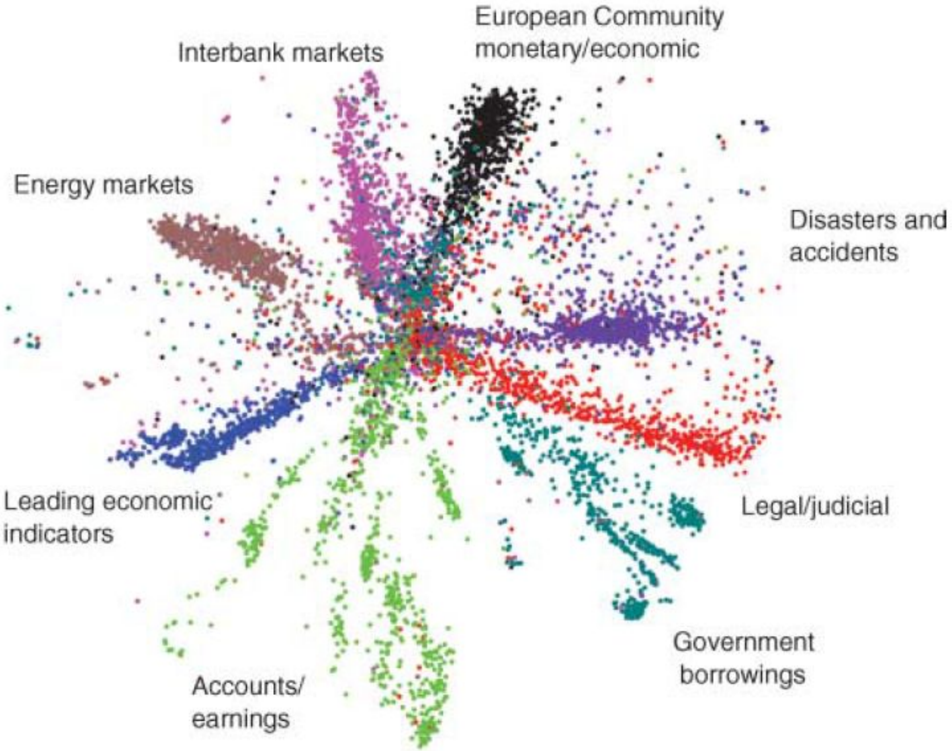
Dimensionality Reduction examples

Text

Using LSA/SVD (similar to PCA)
to reduce to 2D



Using Autoencoder to reduce to 2D



- 800,000 newswire stories
- Pre-categorized (colors)

[Hinton & Salakhutdinov, 2006]

Dimensionality Reduction

Purposes for reduction

- Human Visualization to better understand the data, model, and domain
- Lower the number of parameters of a model by decreasing the input size
- Comparing similarity of input with respect to the most “important” features

Dimensionality Reduction

What is considered “truth” with respect to a visualization?

Dimensionality Reduction example

- Arizona State University Online Course content, dimensionality reduced based on 778 student sequences through the online course

[video1, quiz2, video1, quiz2, video 3. .]
student's course pathway

[[1 0 0 0] [0 0 1 0] [0 1 0 0]...]
student's one-hot pathway

Video 1 [-0.85 0.67]
Quiz 2 [0.30 0.99]
Video 3 [-0.24 -0.55]
t-SNE reduced skip-gram embedding of course pages

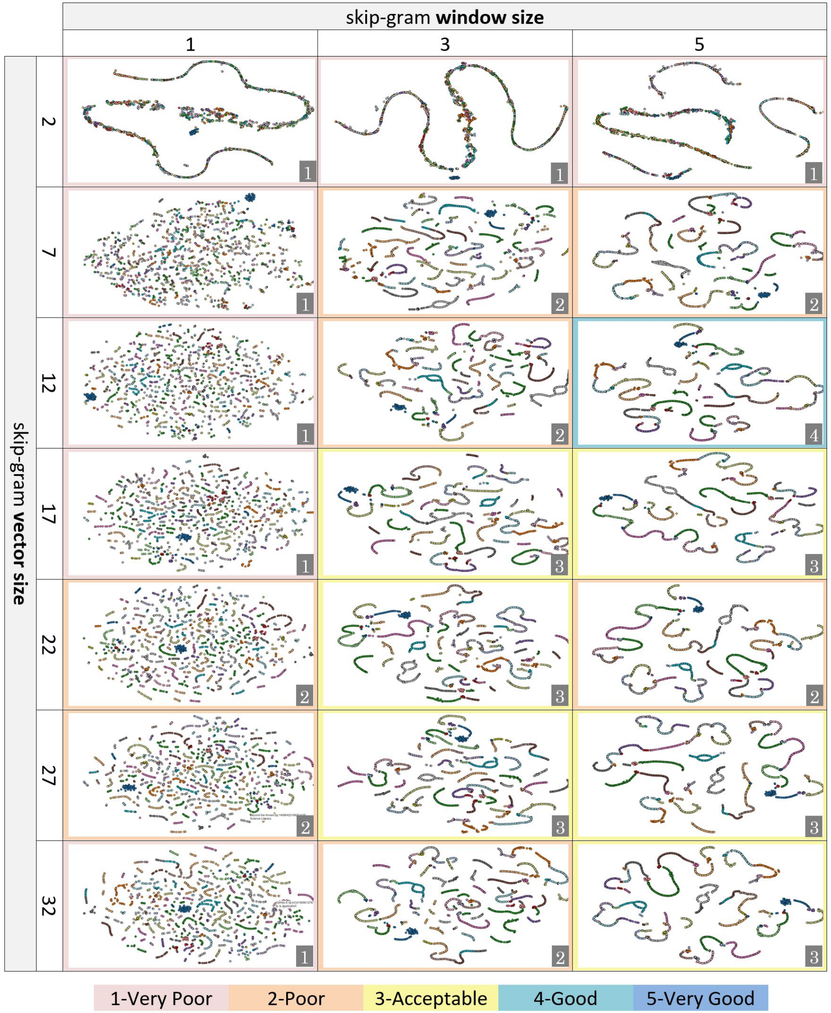


Figure 4: Skip-gram t-SNE visualizations with a variety of window sizes (1-5) and vector sizes (2-32) with instructor usefulness ratings shown in the lower right corner. Each of the 21 scatter plots depict the complete set of elements in the course and their relationship to one another. Colours of plot points represent the lesson within the course to which each point belongs. Useful plots were those which depicted hypothesized and plausibly explainable relationships between course elements.

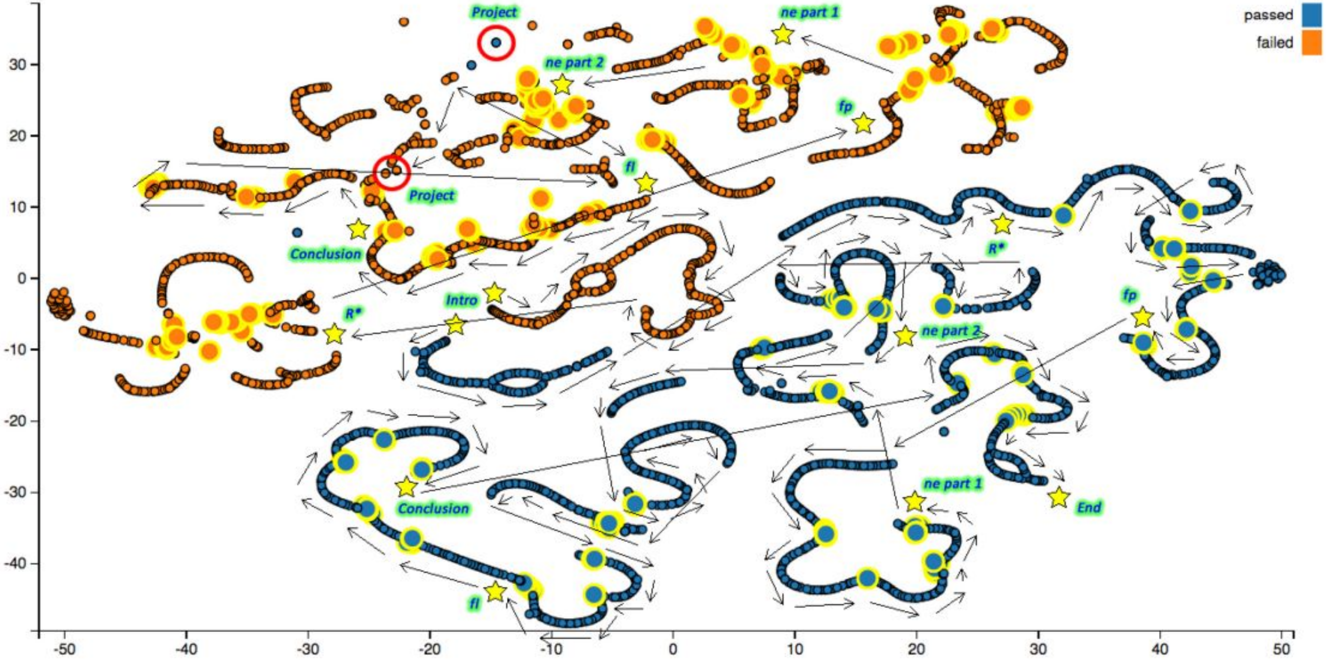
21 versions of the visualization shown to the instructor based on different skip-gram parameters. Instructor is serving here as a validation for model selection.

[Pardos & Horodyskyj (arXiv)]

Dimensionality Reduction examples

Arbitrary sequences

[video1, quiz2, video1, quiz2, video 3. .] ≃ [[1 0 0 0] [0 0 1 0] [0 1 0 0]...] ≃ Video 1 [-0.85 0.67]
Quiz 2 [0.30 0.99]
Video 3 [-0.24 -0.55]



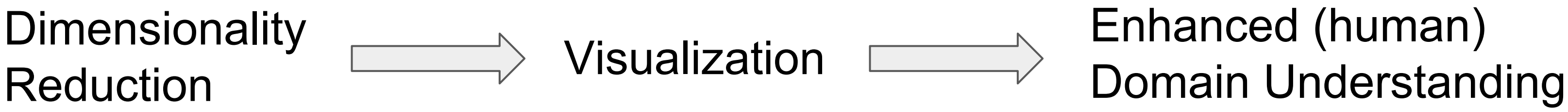
[Pardos & Horodyskyj (arXiv)]

- Course content dimensionality reduction based on 778 student sequences through the online course, with a separate embedding learned for student who passed vs failed the course.

Dimensionality Reduction

Purposes for reduction

- Visualizing to better understand the structure of your data
  - Can also be used to help understand phenomena in your domain



Dimensionality Reduction → Visualization → Enhanced (human) Domain Understanding

Course Logistics

- Quizzes Resume (2 remain)
  - this Thursday (word2vec/autoencoder)
  - and Thursday of next week (Recurrent Neural Networks)
- Labs Resume (1 required, 1 extra credit)
  - this Thursday (Lab 6b, Dimensionality Reduction/Visualization)
  - and Thursday of next week (extra credit - Recurrent Neural Networks)
- Final project
  - should represents approximately 2 labs worth of effort PER MEMBER
  - teamwork constitutes 15% of final project grade (no single teams allowed)
  - presentation time will be limited
    - make sure to rehearse
    - provide extra slides you want us to see in an appendix section (non-presented)