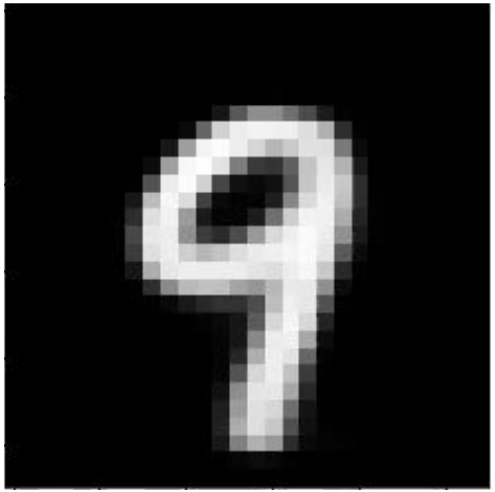
Variational Autoencoders (VAE)

Team 4:
Gopher Knowledge!

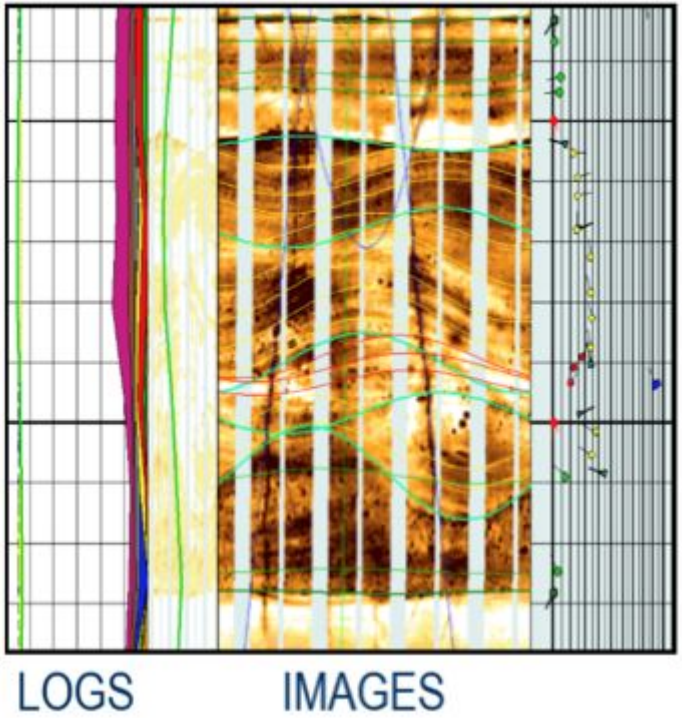


Motivation

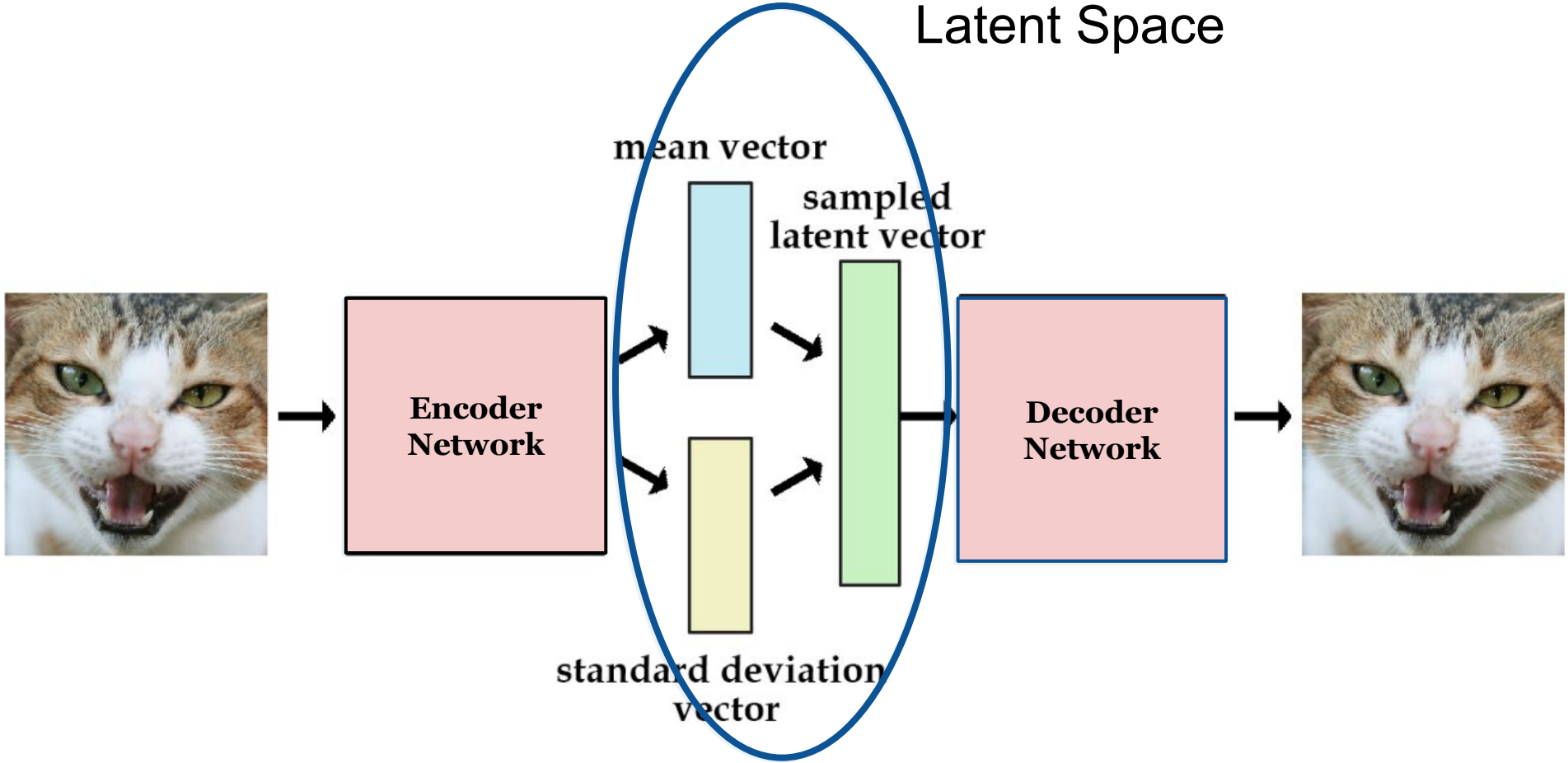
- Traditional classification requires *labelled* data, which is expensive or impossible.

- Classification via clustering is possible, but difficult to extract meaning

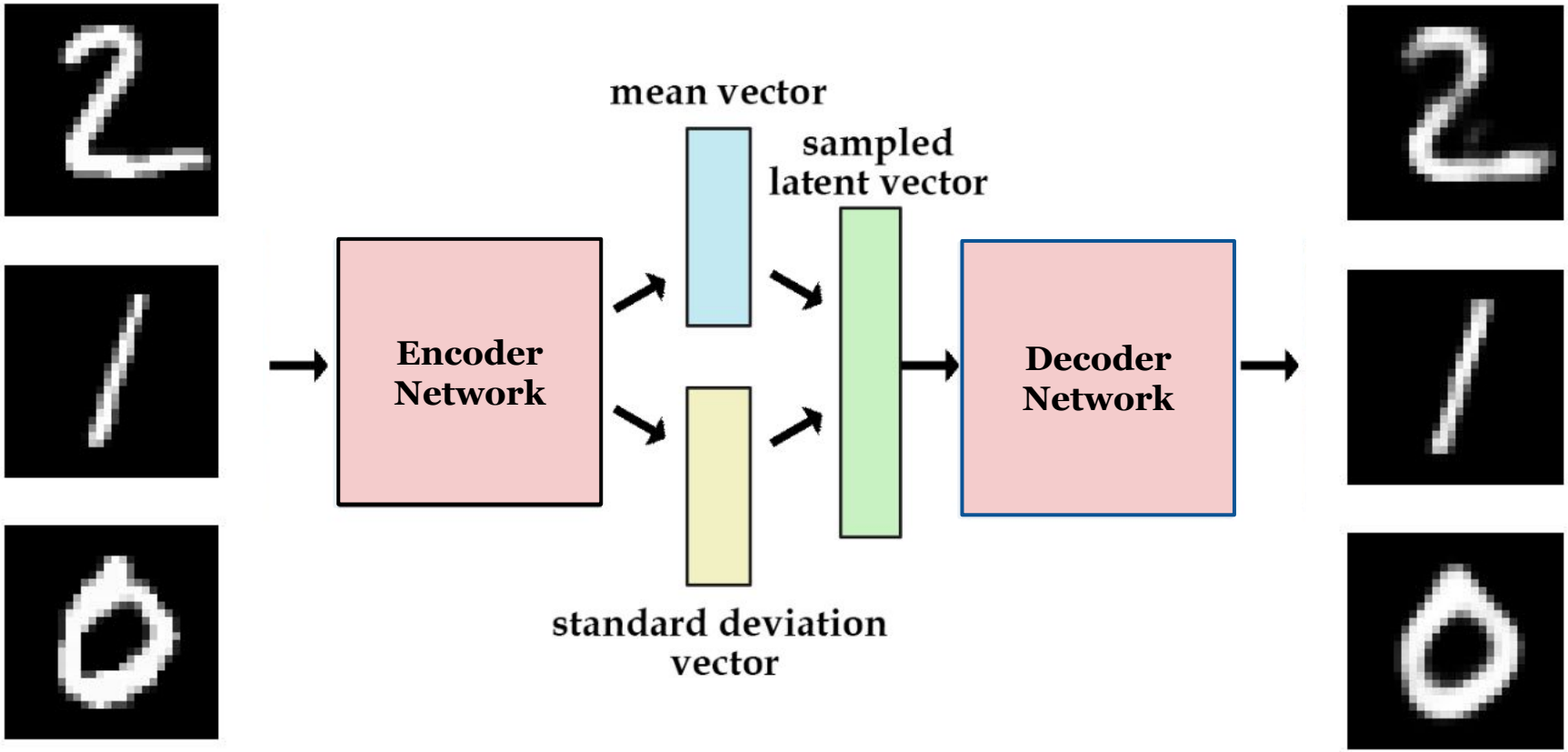
- We want to find underlying, meaningful structure for how data is generated.



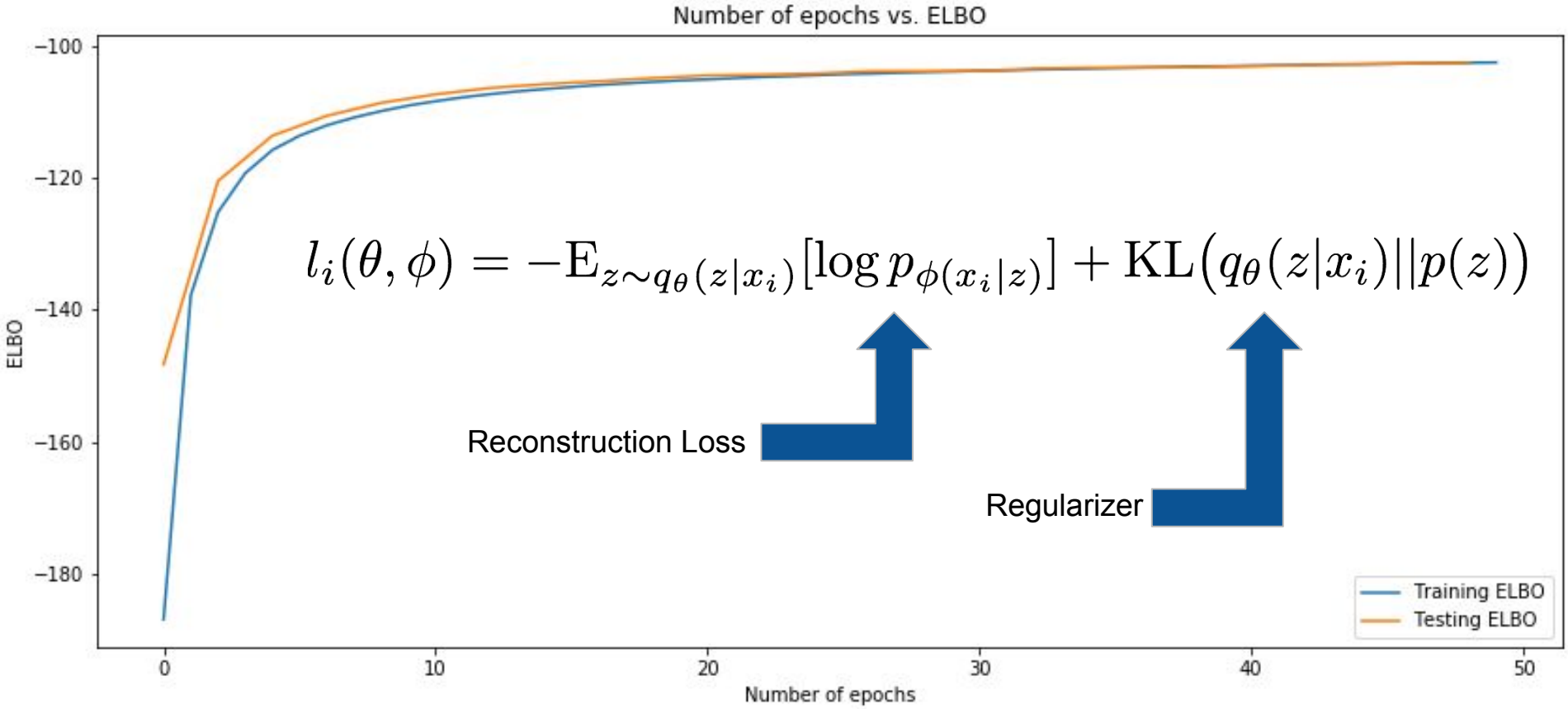
Structure of a VAE



Example


mean vector
sampled
latent vector
Encoder
Network
Decoder
Network
standard deviation
vector

Training the VAE

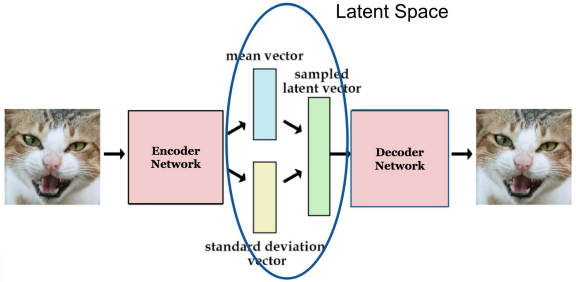


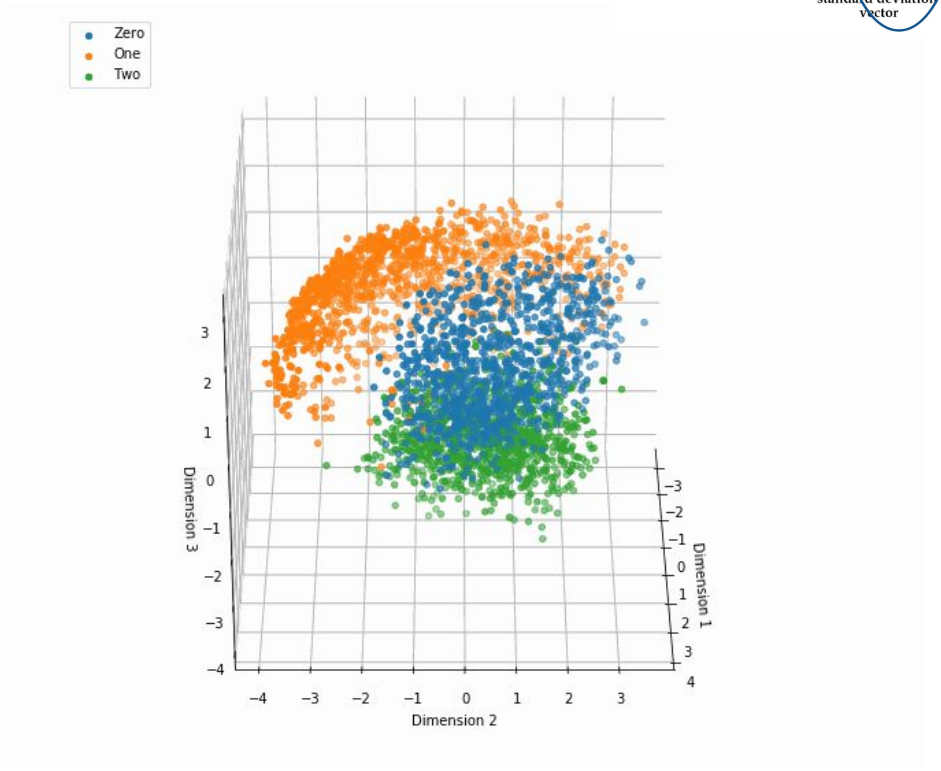
$$l_i(\theta, \phi) = -\mathrm{E}_{z \sim q_\theta(z|x_i)}[\log p_{\phi(x_i|z)}] + \mathrm{KL}\big(q_\theta(z|x_i)||p(z)\big)$$

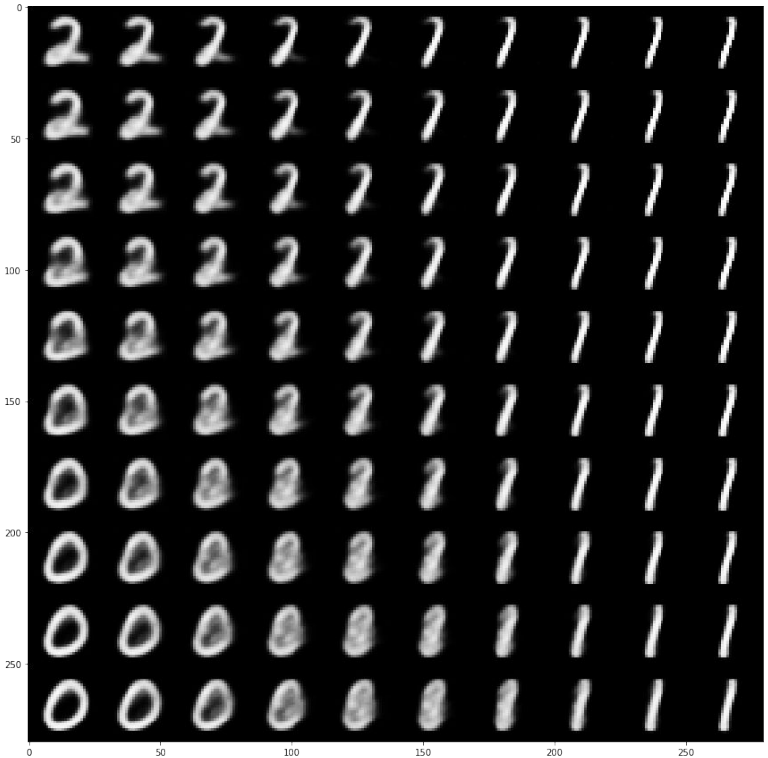
Reconstruction Loss

Regularizer

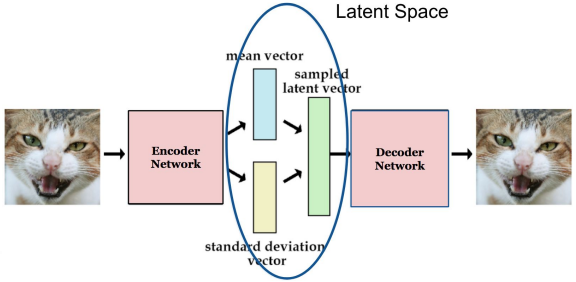
Latent Space

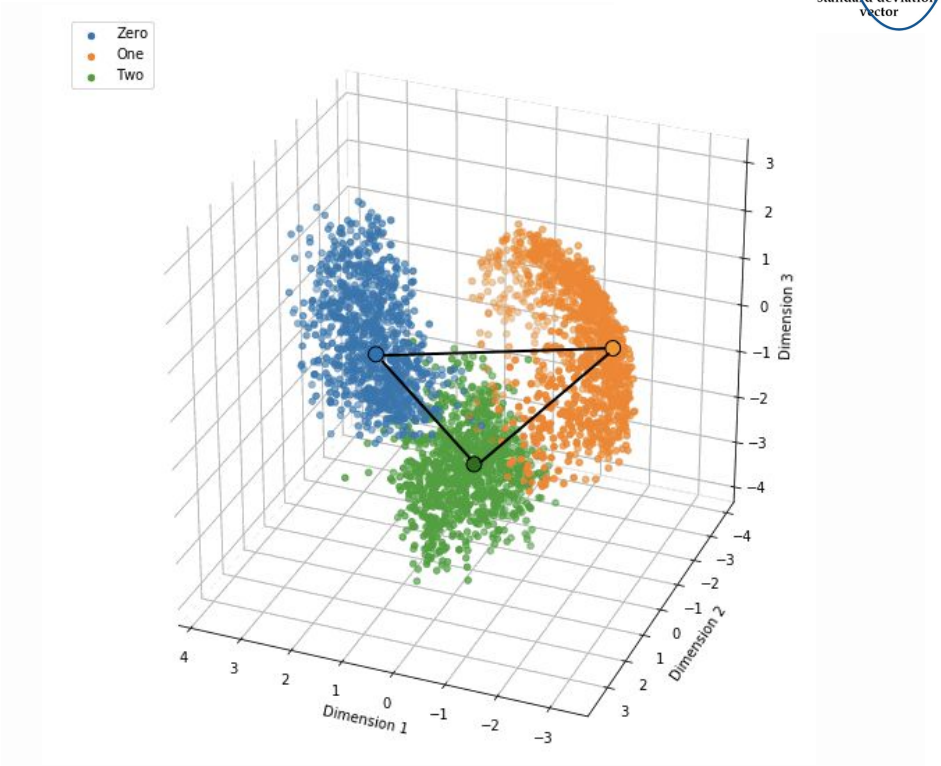


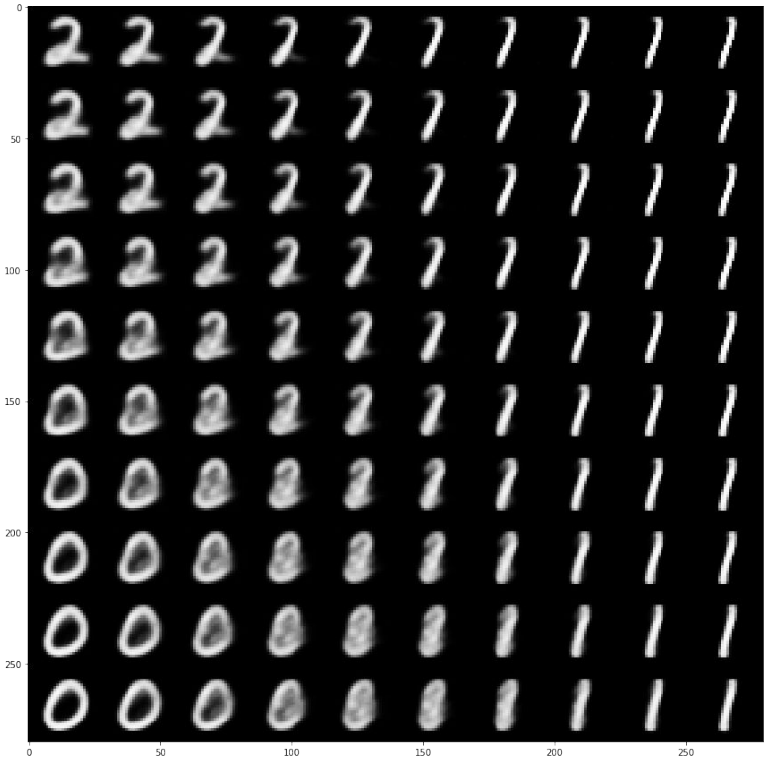




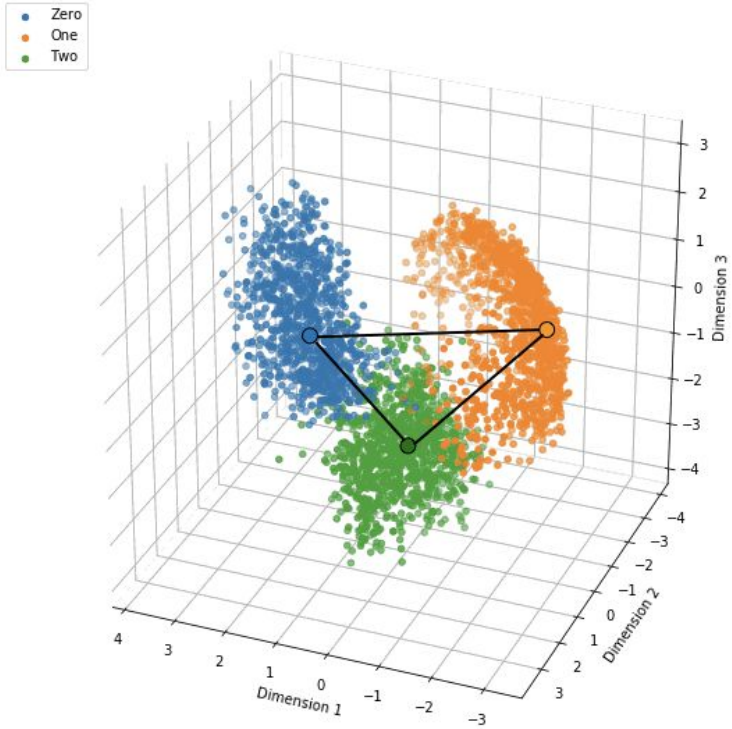
Latent Space

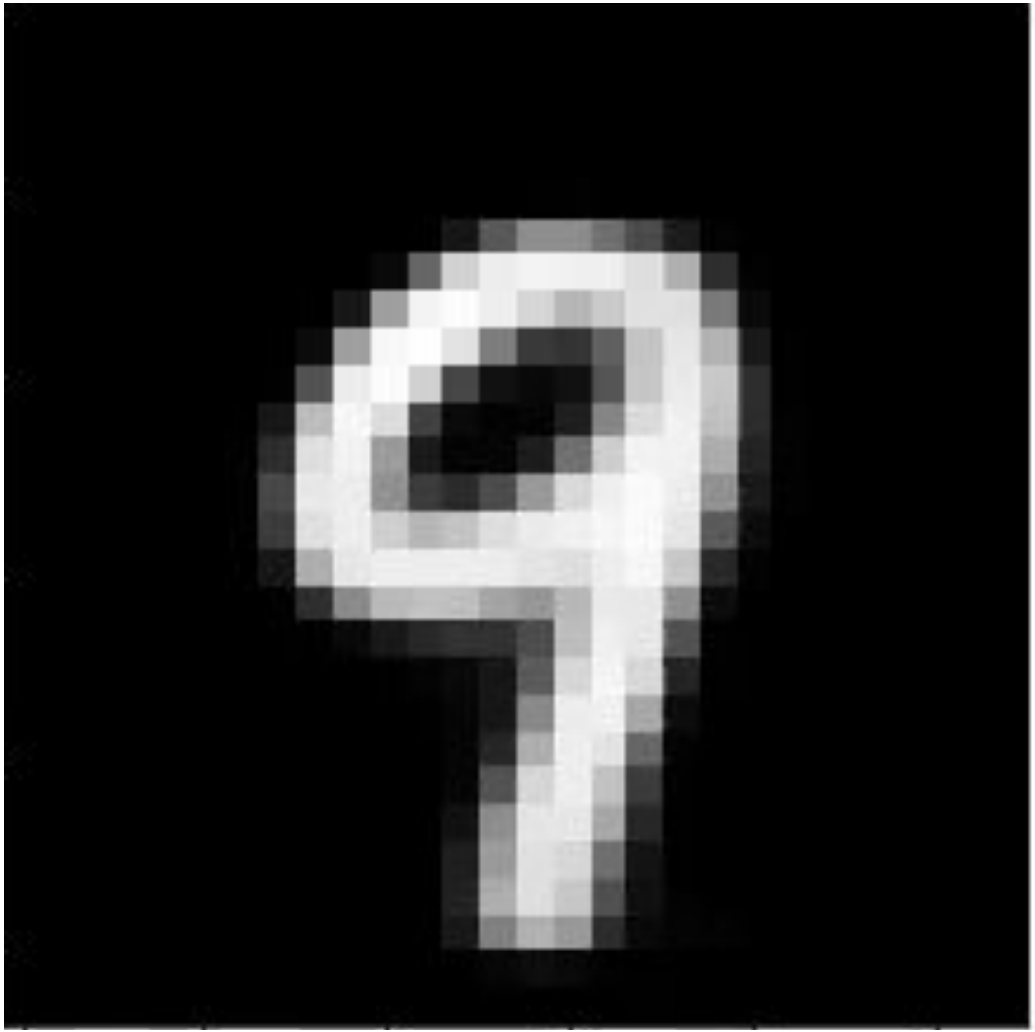




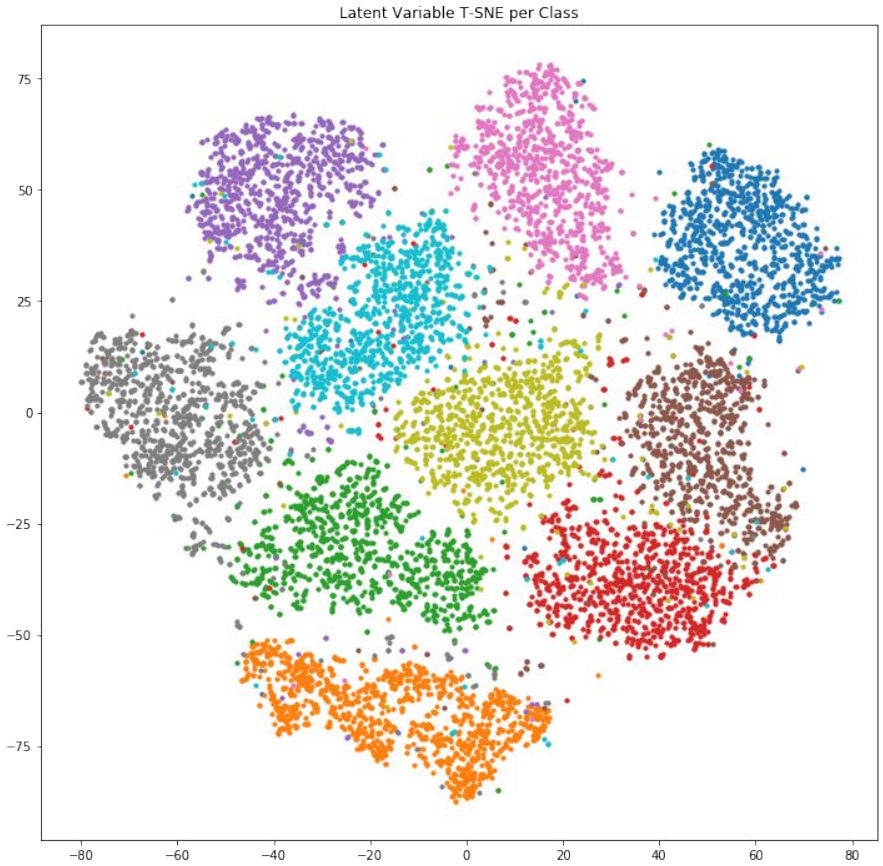


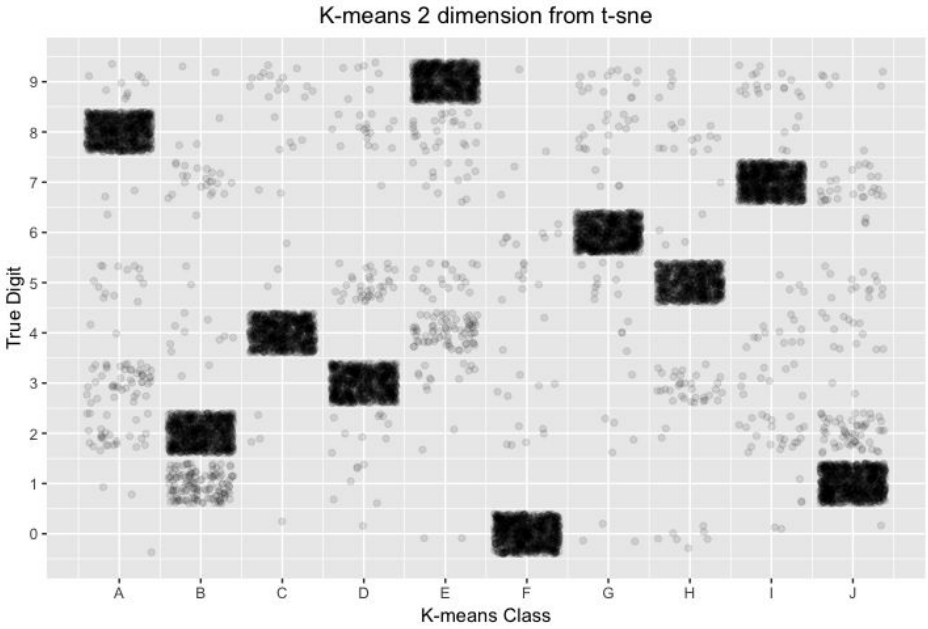
A path through latent space





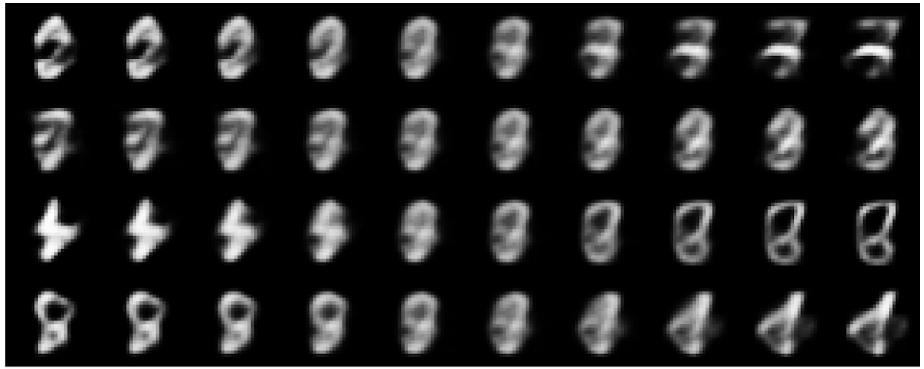
Current methods: clustering latent space





Accuracy on 50-dim latent space	~ 72 %
Accuracy after dimension reduction (t-SNE)	~ 91 %

Dimension Requirements



20-dimensional latent space

50-dimensional latent space

Latent Dirichlet Allocation (LDA)

Topics

```
gene      0.04
dna       0.02
genetic   0.01
...
```

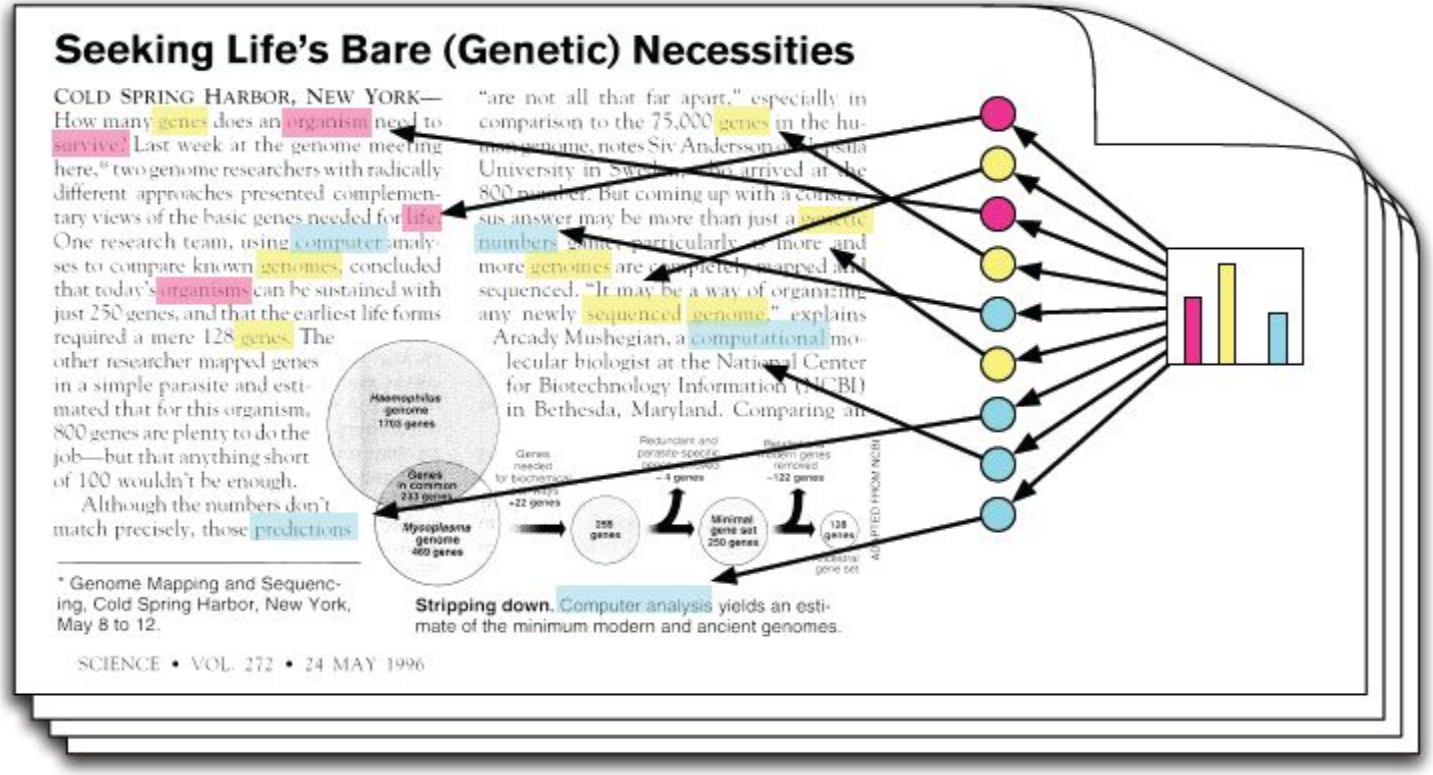
```
life      0.02
evolve    0.01
organism  0.01
...
```

```
brain     0.04
neuron    0.02
nerve     0.01
...
```

```
data      0.02
number    0.02
computer  0.01
...
```

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers. Since particularly more and more genomes are completely mapped and sequenced, "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

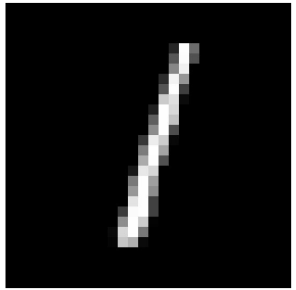
SCIENCE • VOL. 272 • 24 MAY 1996

LDA on Images

Topic 1: 10%
Topic 2: 70%
Topic 3: 20%

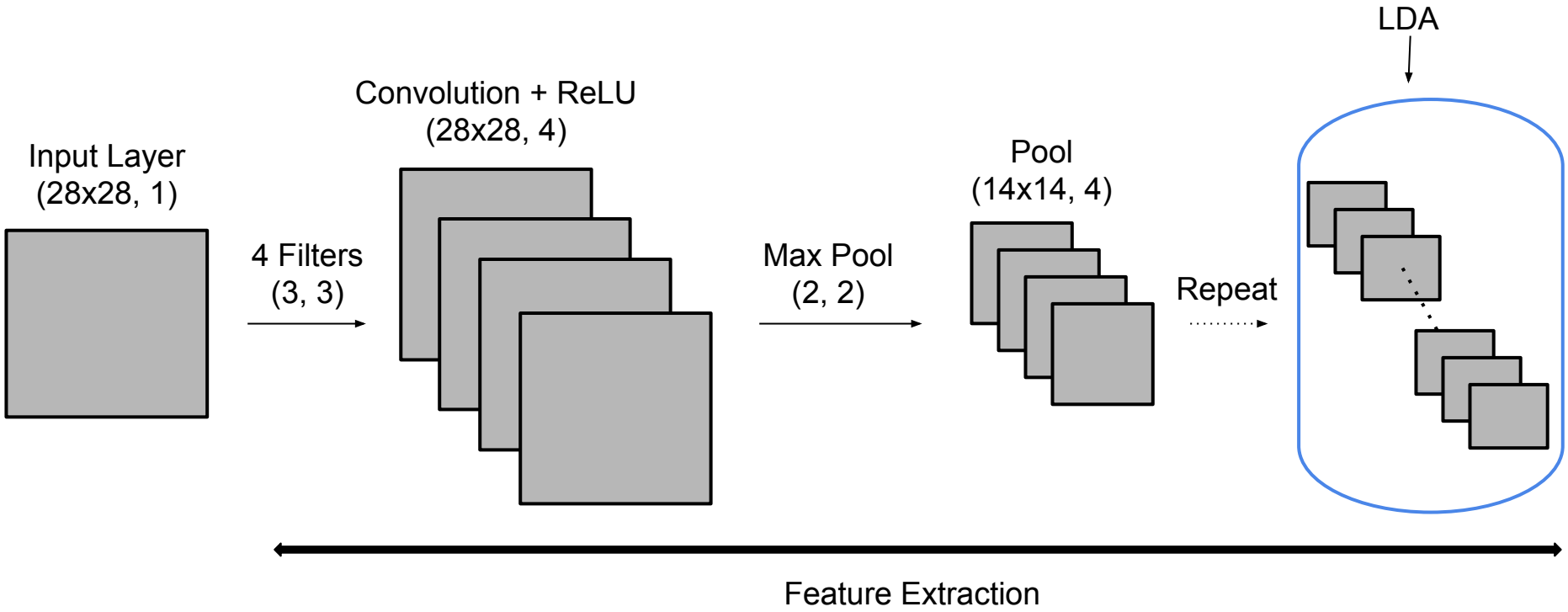
20% Word 1
35% Word 2
15% Word 3
30% Word 4

Digit 0: 05%
Digit 1: 90%
Digit 2: 05%

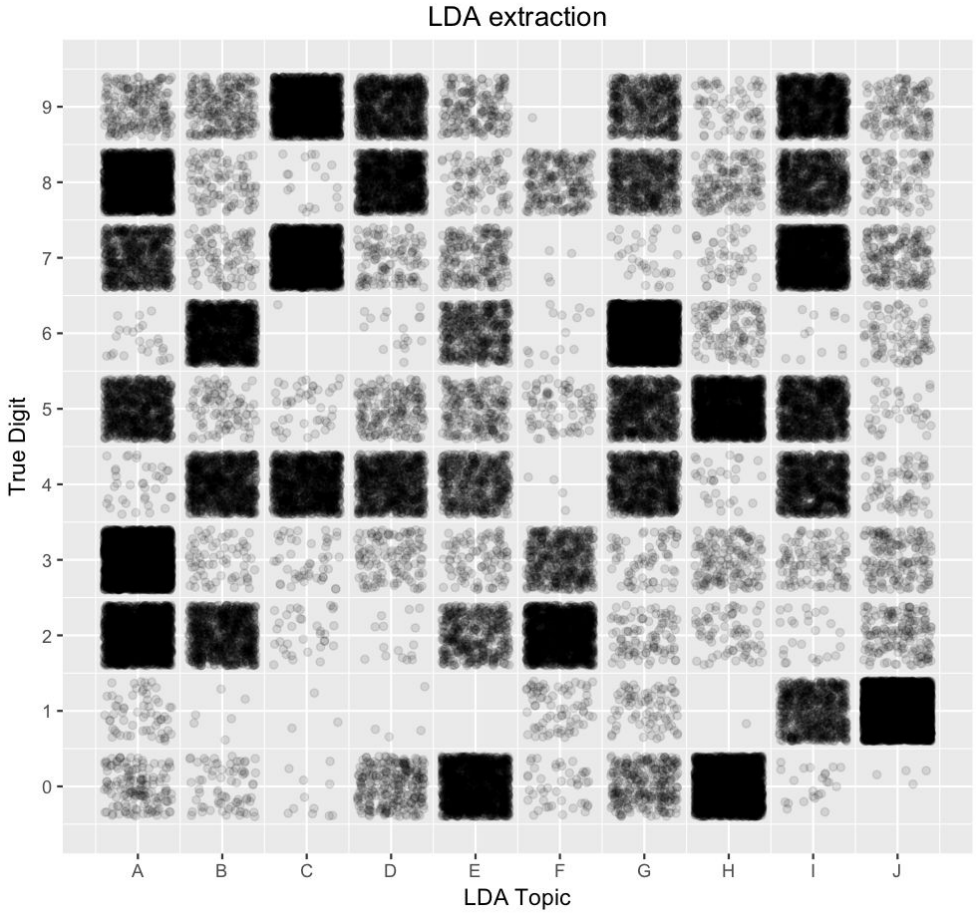


20% Feature 1
35% Feature 2
15% Feature 3
30% Feature 4

Convolutional Neural Networks (CNN)

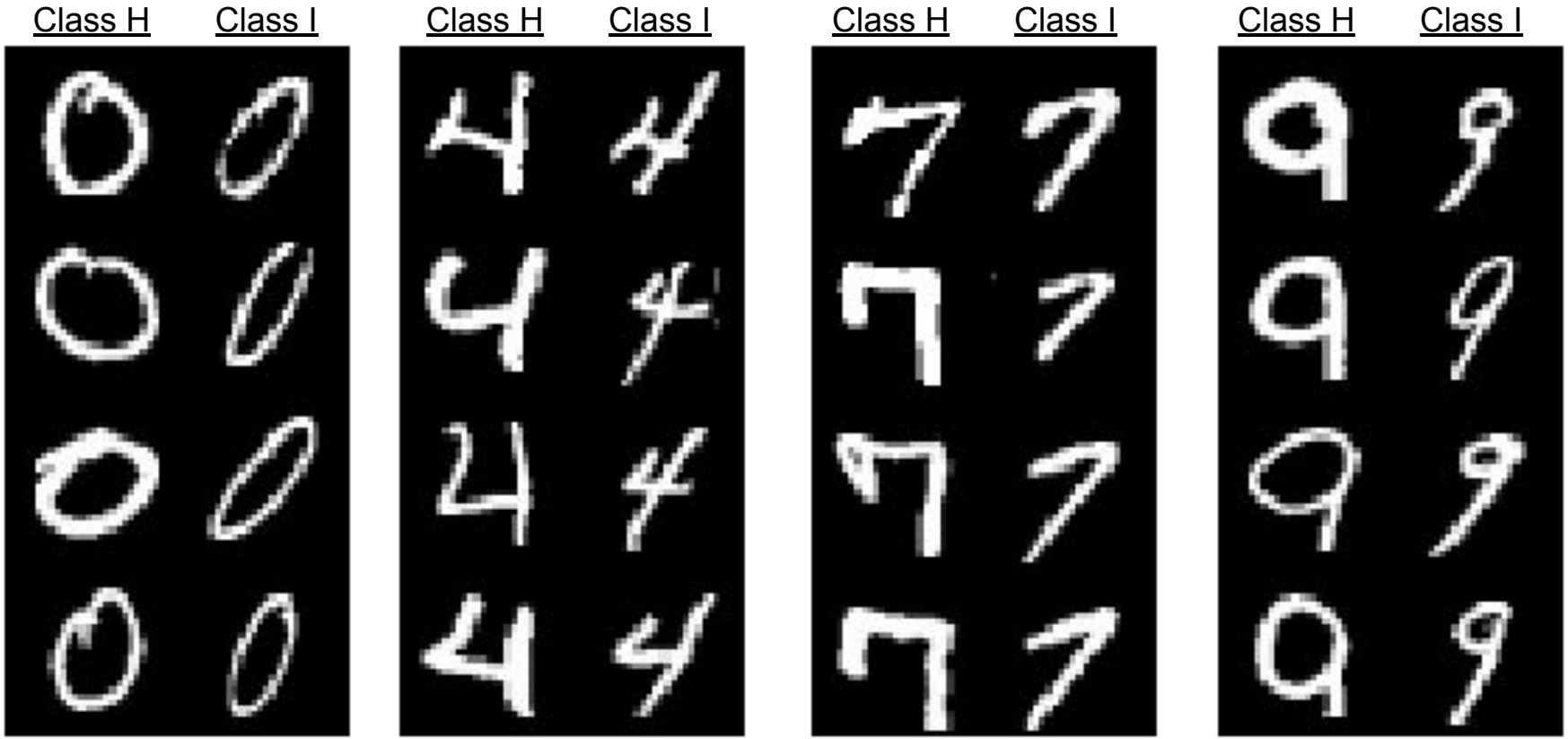


LDA Results

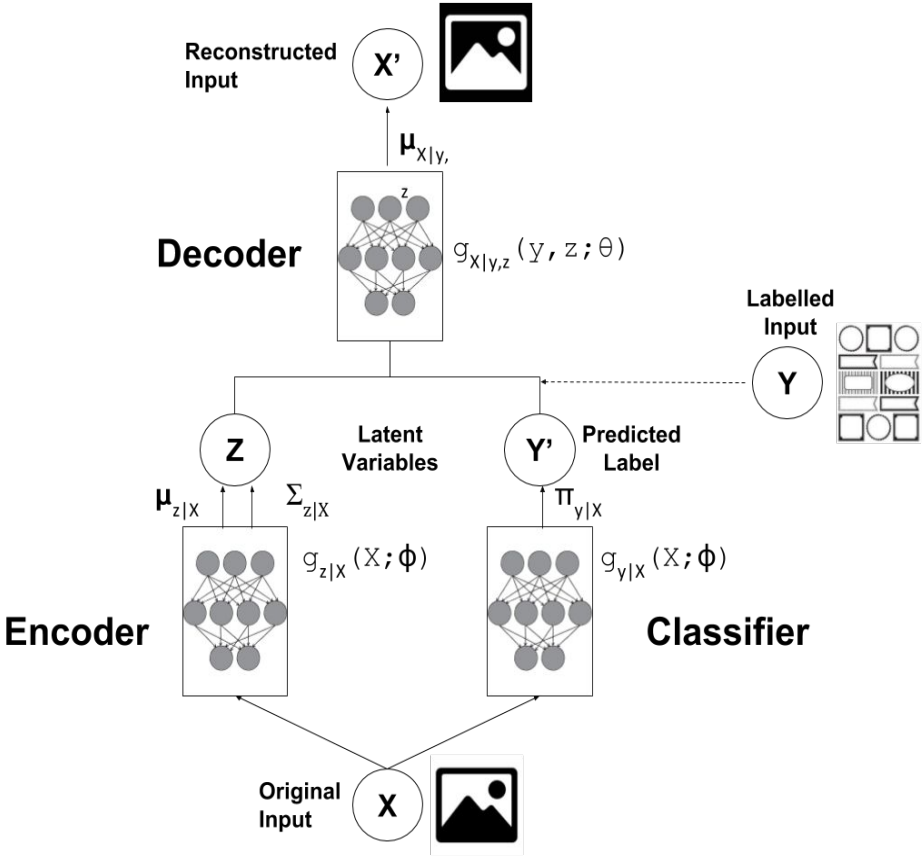

LDA extraction
True Digit
9
8
7
6
5
4
3
2
1
0
A
B
C
D
E
F
G
H
I
J
LDA Topic

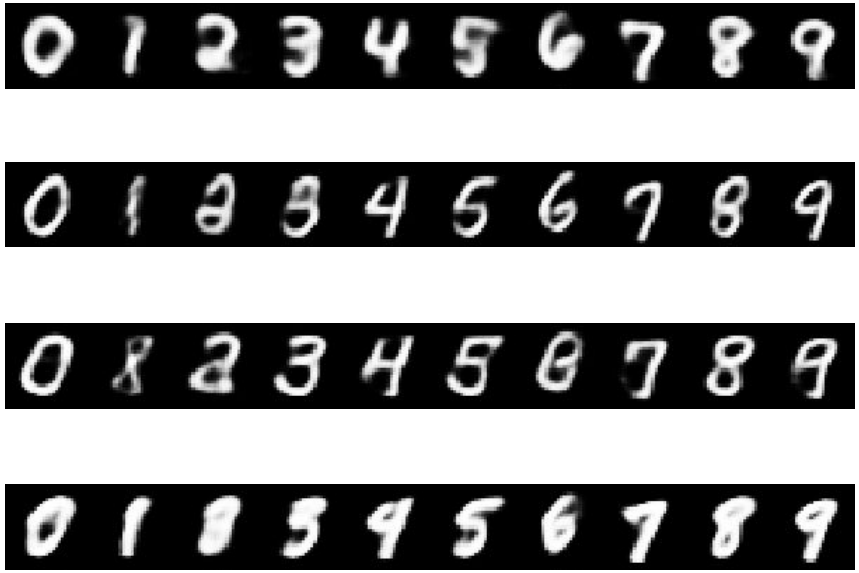

Class H
Class I

LDA Class Groupings

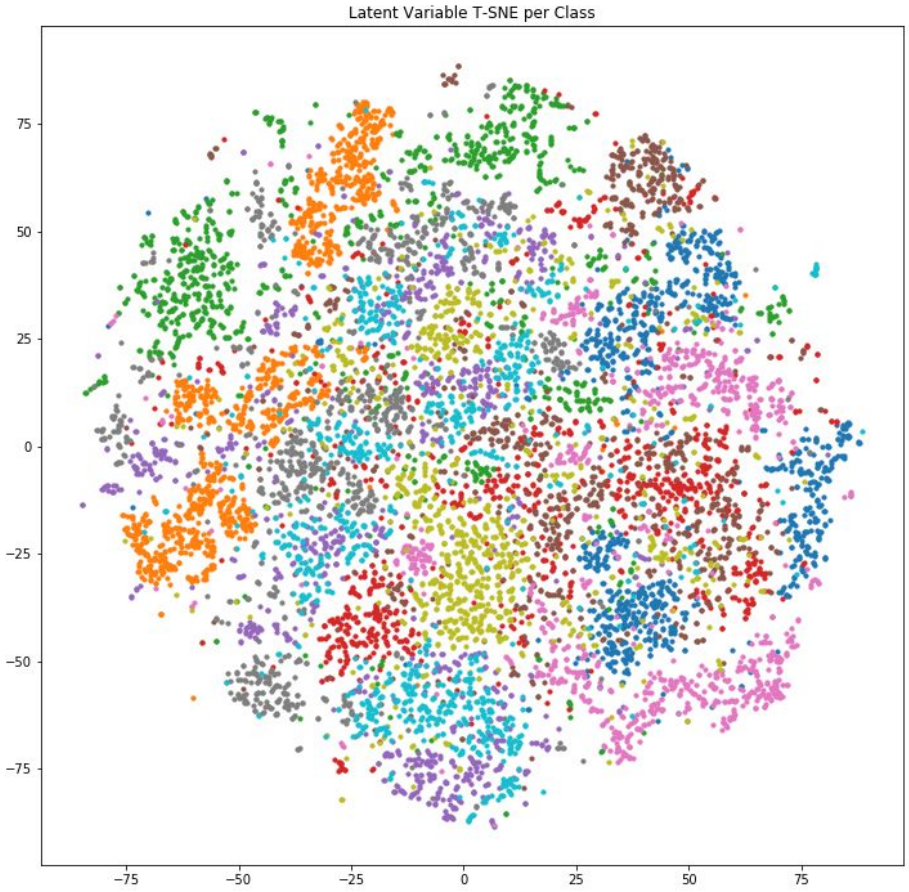


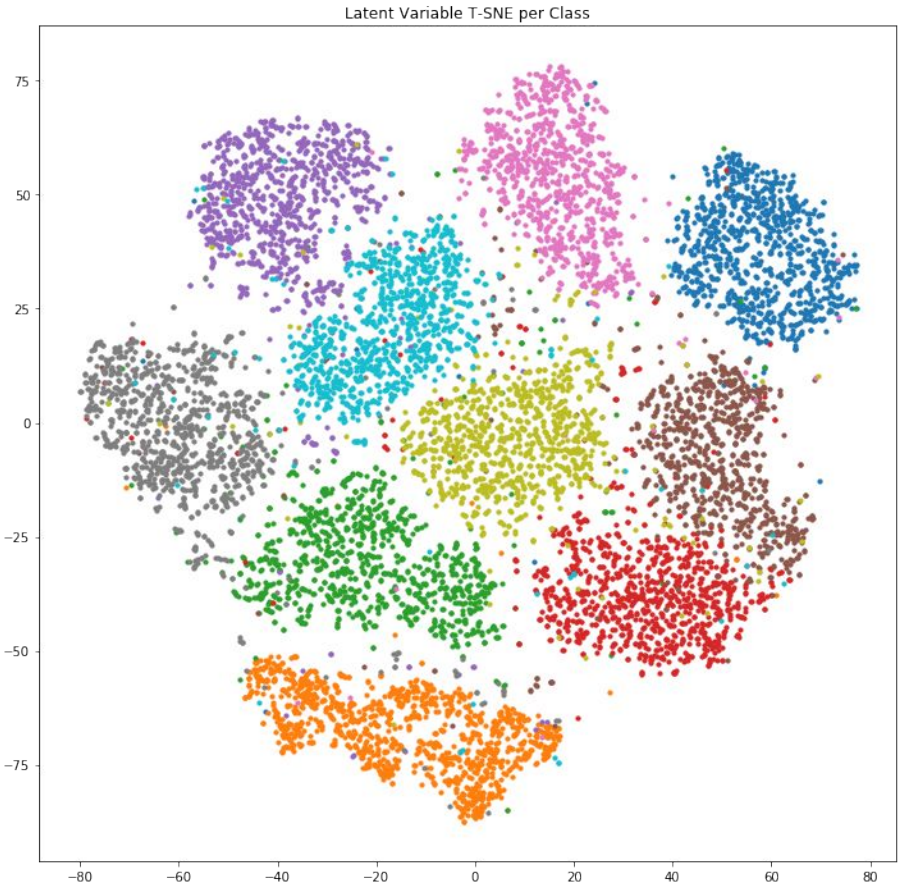
Semi-Supervised Learning





Semi-Supervised Learning



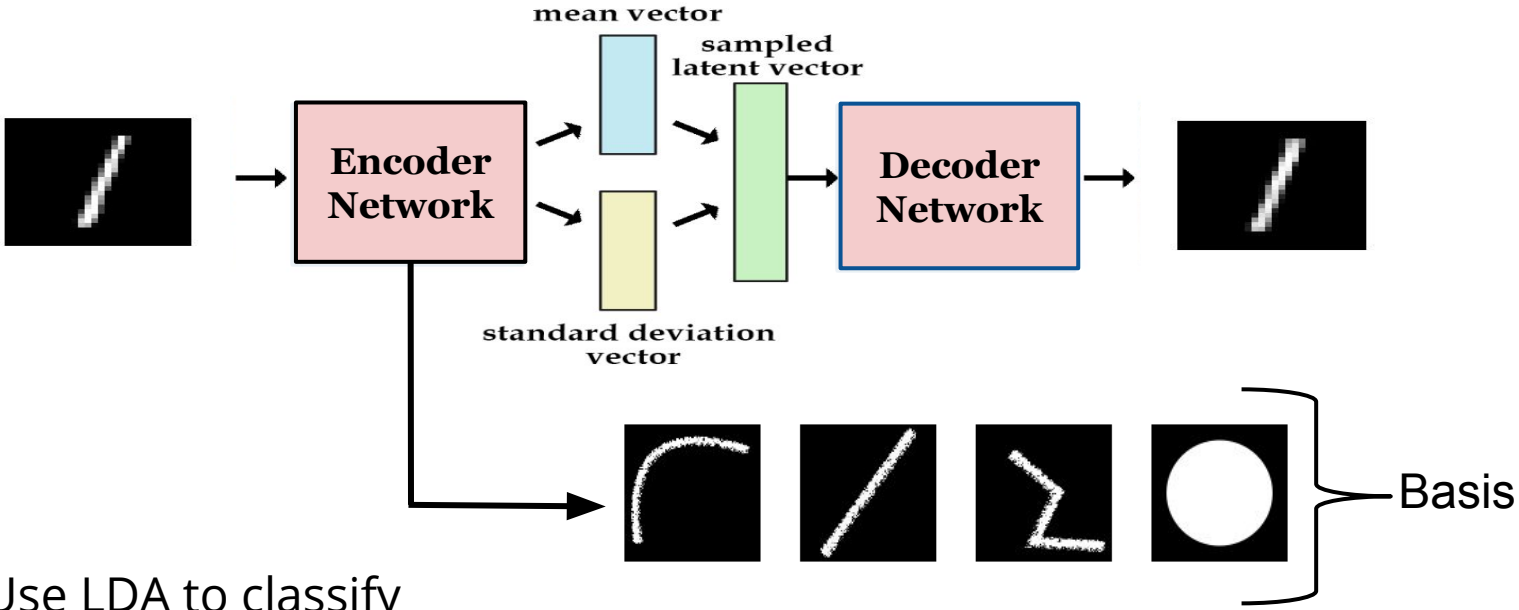


Results:

Number of labeled samples:	200	800	1600	3000
Validation Accuracy:	0.7522	0.9007	0.9380	0.9494
Test Accuracy:	0.7380	0.8968	0.9321	0.9427

Future Work

Step 1: Learn a basis of features from the VAE encoder



Step 2: Use LDA to classify

Topics:
Digits

Documents:
Images

Words:

References

Papers:

- Auto - Encoding Variational Bayes, Diederik P.Kingma, Max Welling
- Latent Dirichlet Allocation, David M.Blei, Andrew Y. Ng, Michael I. Jordan
- Categorical Reparameterization With Gumbel - Softmax, Eric Jang, Shixaing Gu, Ben Poole

Codes:

- https://github.com/uber/pyro/blob/dev/examples/vae/vae.py
- https://pypi.org/project/lda/
- http://bjlkeng.github.io/posts/semi-supervised-learning-with-variational-autoencoders/

Team Gopher Knowledge:

- Hua Chen, University of Delaware
- Aaron Cohen, Indiana University
- Mingchang Ding, University of Delaware
- Melanie Jensen, Tulane University
- Christopher Miller, University of California, Berkeley
- Michael Ramsey, University of Colorado

Special Thanks To:

- Irfan Bulu, Schlumberger-Doll Research
- Our “baby gopher” Gabriel
- TAs Binh Tang and Chris Browne
- Directors and administrations of IMA, especially Ben, Dan, Fadil, Thomas, Georgia, and Katherine