JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Lecture 5

BU.330.775 Machine Learning

Minghong Xu, PhD.
Associate Professor

Review

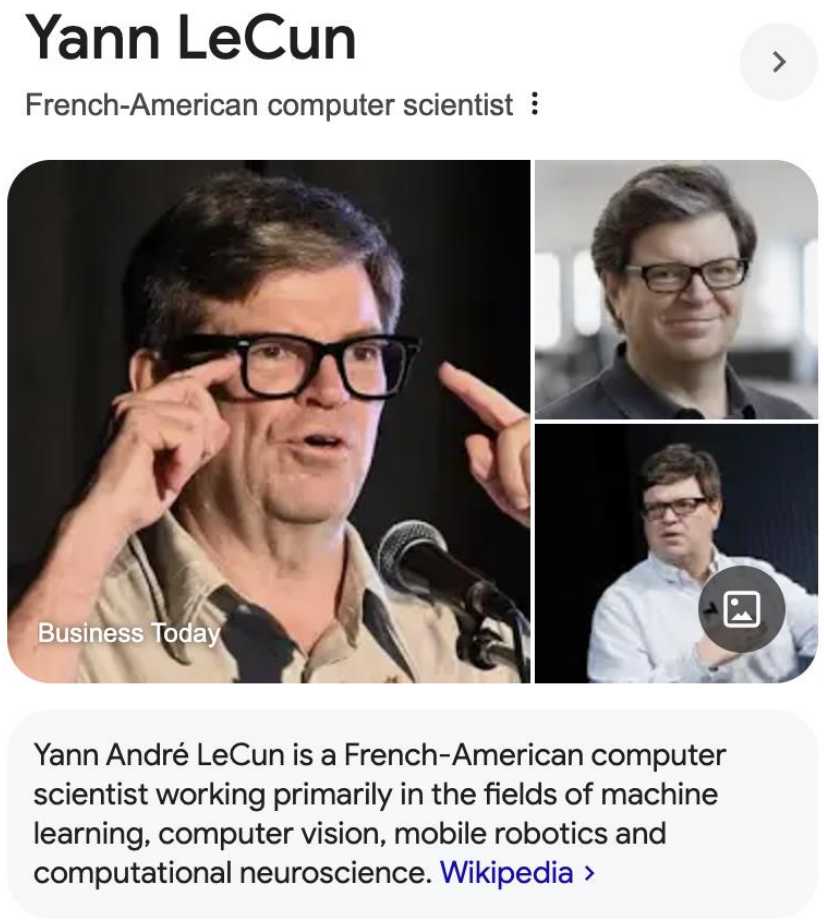
- Evaluate model performance
  - Accuracy, confusion matrix
  - Tradeoff between precision and recall
  - False positive vs false negative
- Another competition next week

Today's Agenda

- Unsupervised Machine Learning, Part I
  - Dimensionality Reduction
  - PCA
- Hands on practice using breast cancer dataset and MNIST
  - A lot of Python exercise today

Value of Unsupervised

- Yann LeCun: “if intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake”



Training and “Testing” in Unsupervised

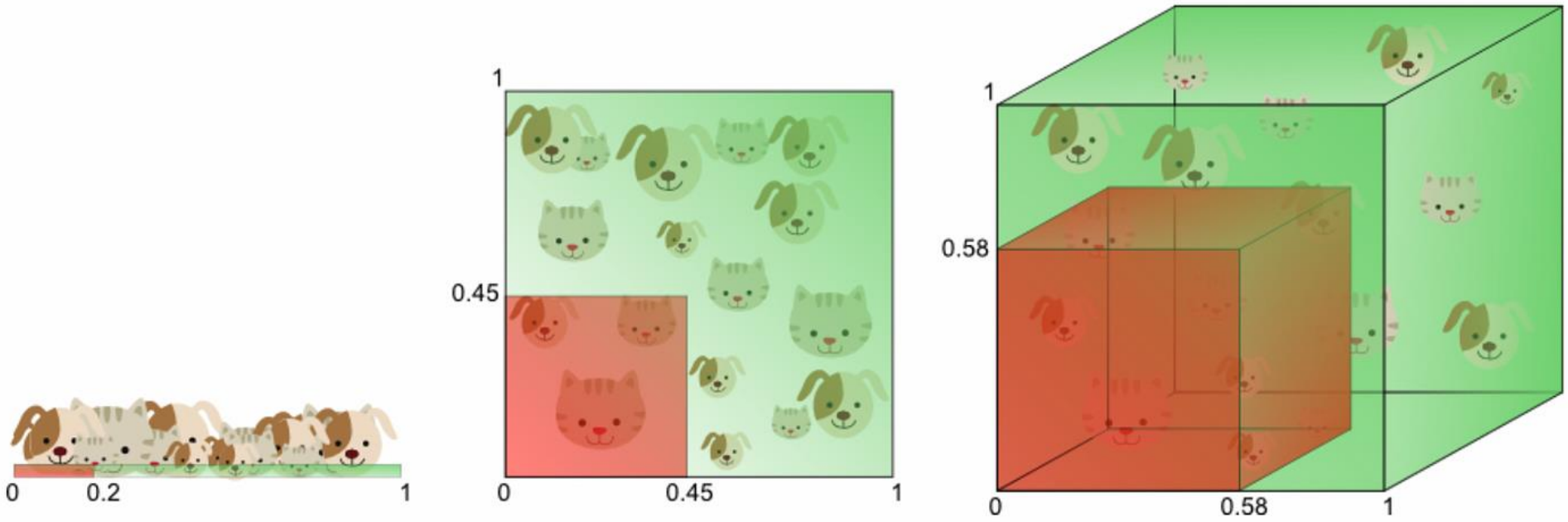
- Unsupervised approaches usually used in an **exploratory** setting
- Training phase is to expose the model to data
  - Explore patterns, clusters, relationships
  - But not related with label
- traditional “testing” not exist
  - *If I do not have correct/suggested answers, can I test you?*
  - Alternative evaluations exist
    - Internal: use some metrics to access the quality
    - External: if labels are available, compare to that
    - Generalization: apply to a separate dataset if the patterns learnt still hold

Unsupervised Learning

- **Dimensionality reduction**
  - **Visualization**
  - **Factor analysis (Finance)**
  - **Natural language processing**
  - **Gene sequencing**
- Clustering
  - Product recommendations
  - Customer segmentation
  - Targeted marketing
  - Medical diagnostics
- Association Rule (in Cloud Computing course)

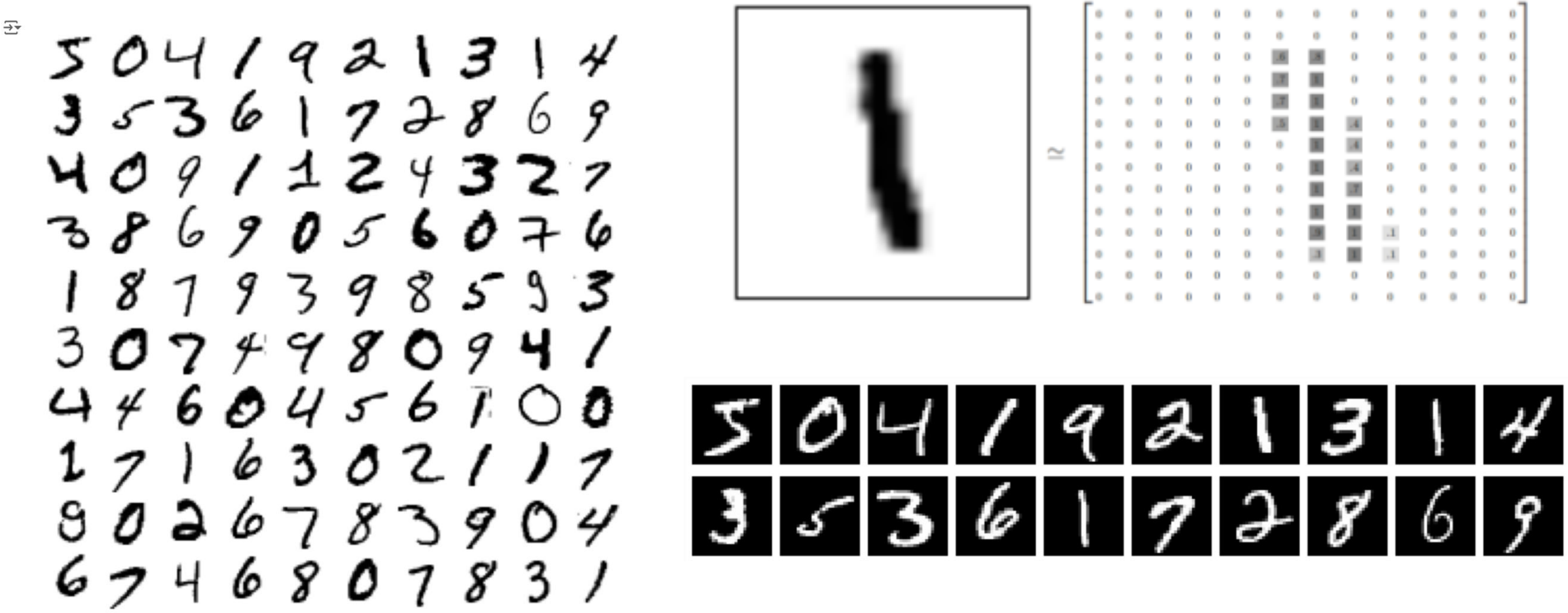
Curse of Dimensionality

- As the dimensionality of the feature space increases, the number of configurations can grow exponentially
- High dimensions: lot of features



https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

MNIST Example



Another Image Example

- Downsampling or subsampling
  - Two neighboring pixels are often highly correlated
  - Merge two into a single pixel (average/max) does not lose much information

Issues of High Dimension

- More parameters to learn
- Slow training process
- Data points become more sparsely distributed
- Bad efficiency in computation
- Difficulty to find a good solution

Notes

How to Solve

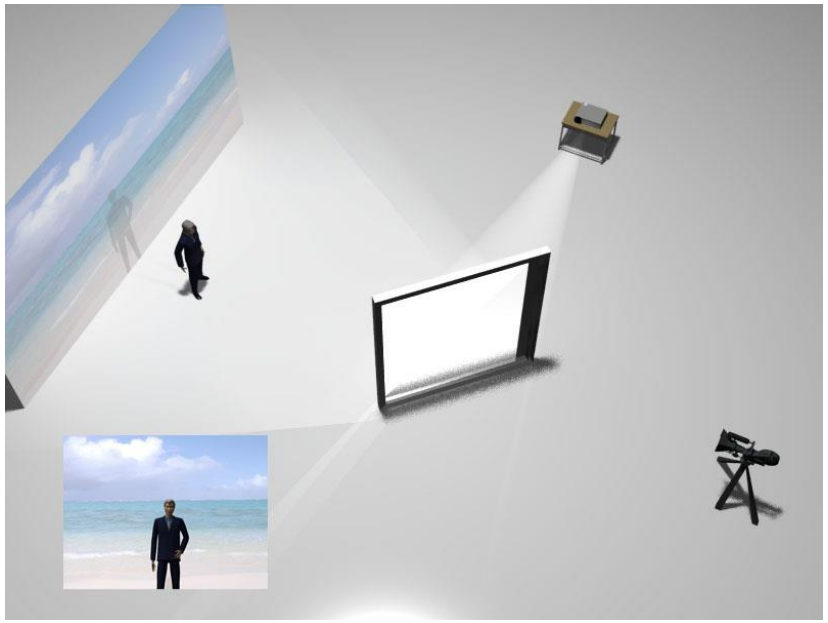
- Reduce the dimensionality!

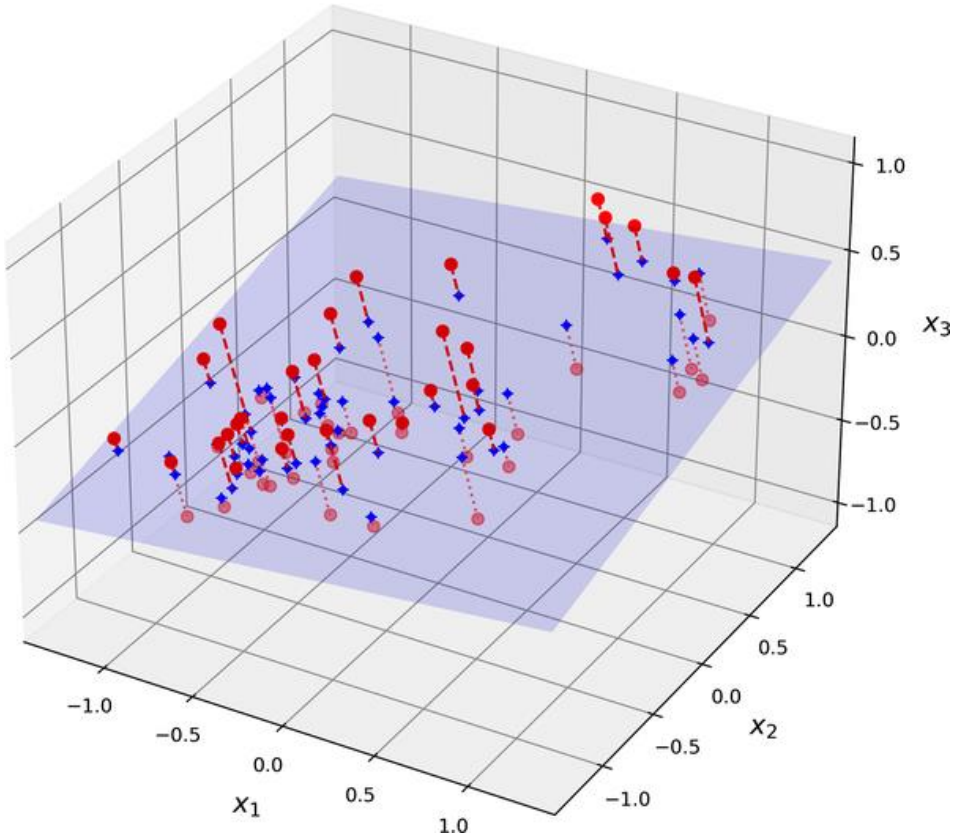
- Speed up training
- Better visualization
  - Reduce to two dimensions
  - For communication with non-data scientists

- **Caution**: reducing dimensionality does cause some information loss
  - Train first with original data before dimensionality reduction

Projection

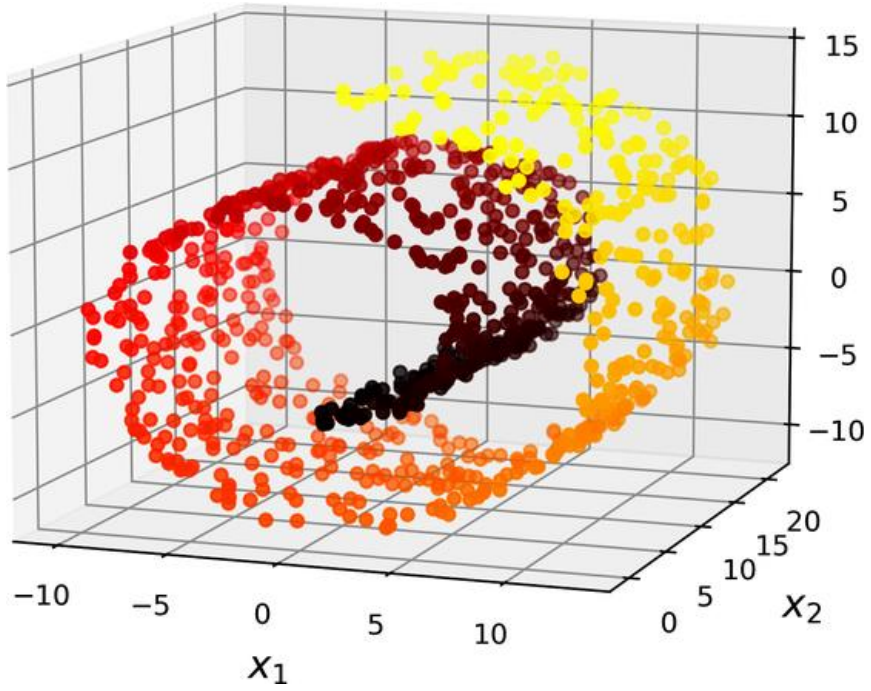
- Project every instance perpendicularly onto a subspace
- *Information loss?*

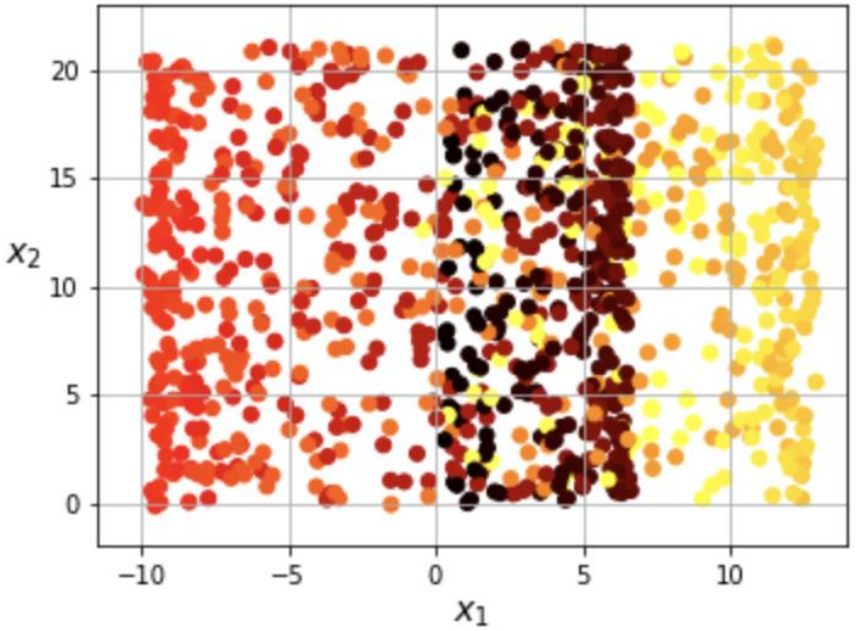




Issue of Projection

- Subspace may twist and turn



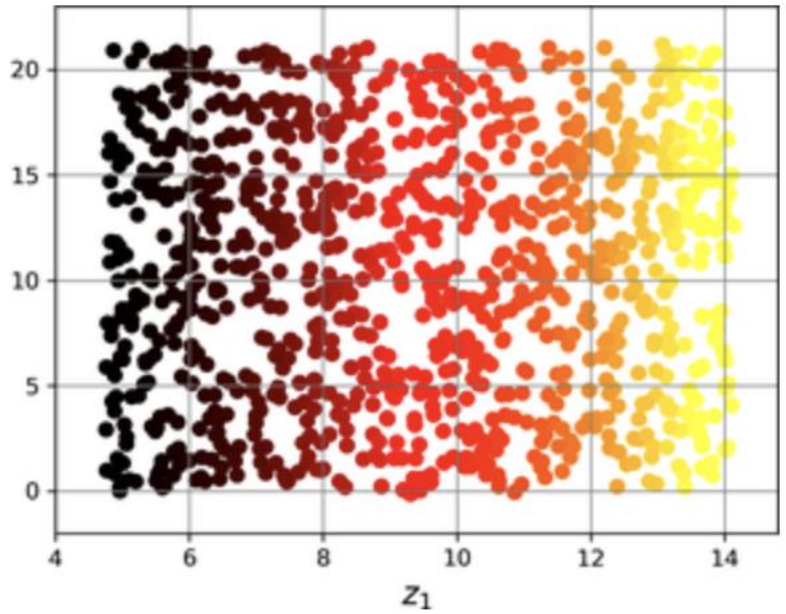


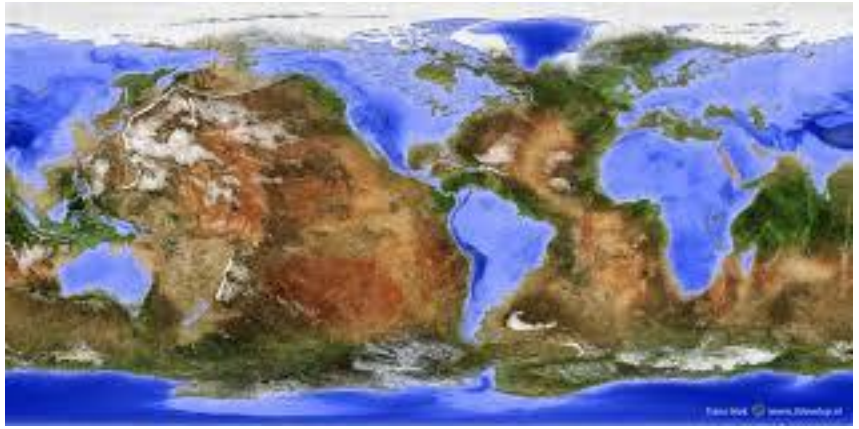
Squashed

- https://scikit-learn.org/dev/modules/generated/sklearn.datasets.make_swiss_roll.html

Manifold Learning

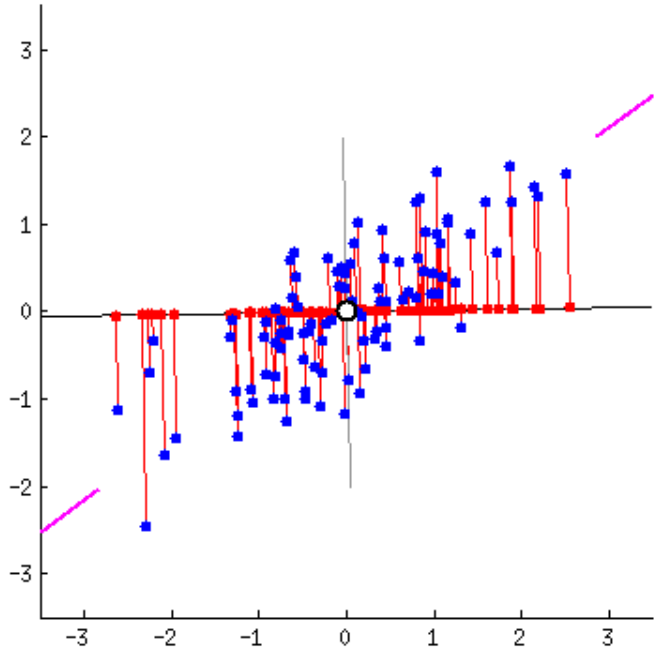
- “Unrolling”





Principal Component Analysis (PCA)

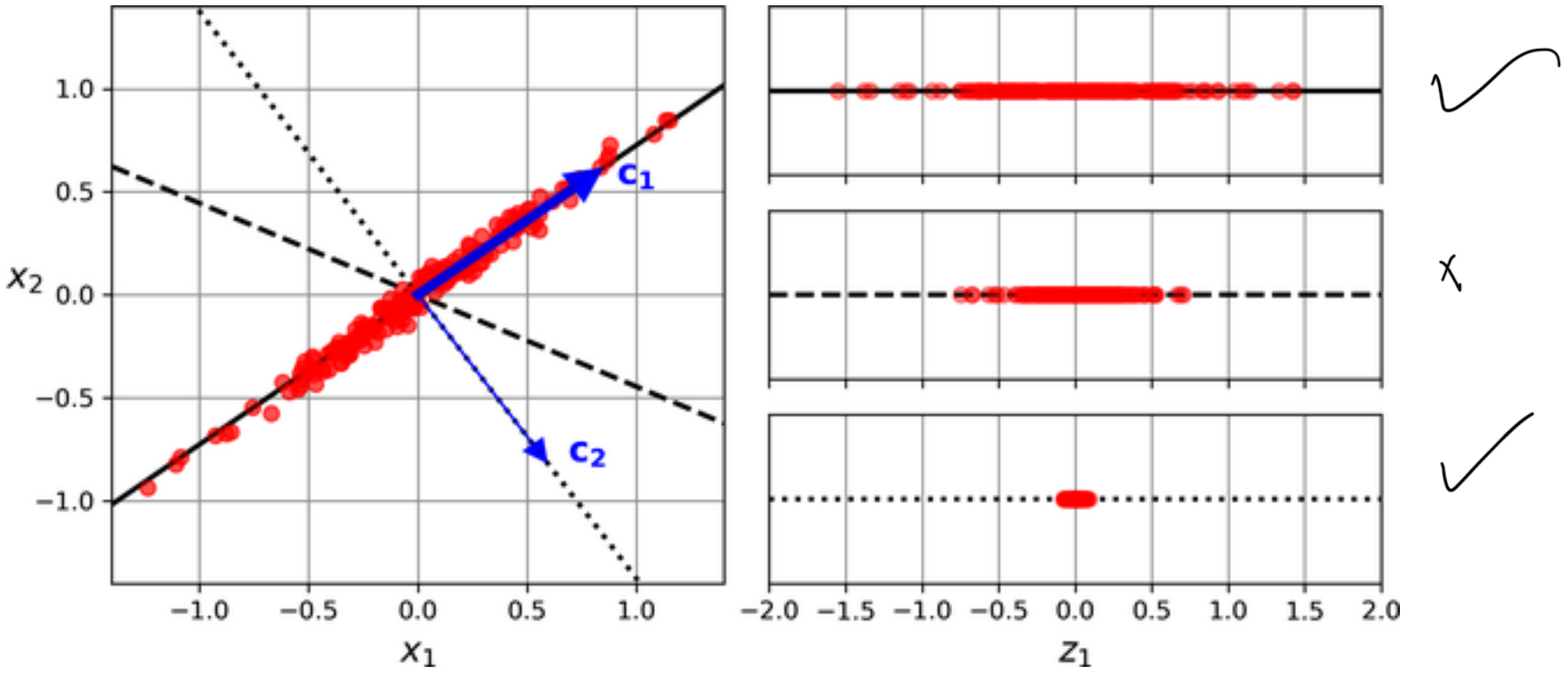
- Most popular dimensionality reduction algorithm
- A projection method



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

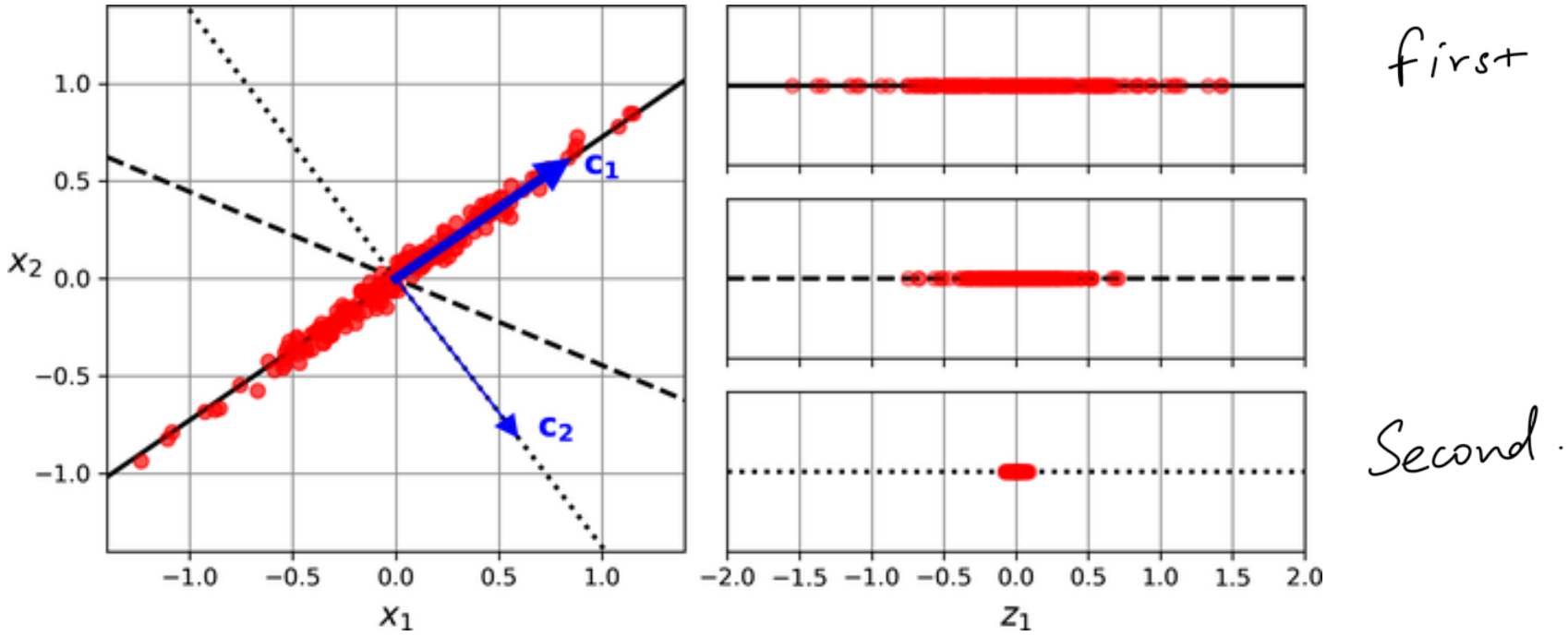
Choose the Right Hyperplane

- Hyperplane: reduced dimensional plane of the original space



Principal Components

- First principal component: the axis that preserve the maximum variance
- *How about the second principal component?*



Orthogonal: Uncorrelated

- *What is the issue if two features are highly correlated?*
- Second axis orthogonal to the first, preserving the largest amount of remaining variance
- Third axis orthogonal to the previous axes, preserving the largest amount of remaining variance
- …
- i^{th} principal component: i^{th} such axis

- Each principal component: linear combination of original features

PCA

- Project the training set onto a lower-dimensional hyperplane
- Select only a subset of features which are important for explaining the data
- New features are statistically uncorrelated
- **Explained variance ratio**: indicate the portion of dataset's variance that lie along each principal component
  - An internal evaluation metric
- Choose right number of dimensions: add up to a sufficient portion of variance, such as 95%

Scaling Before PCA

- Feature scaling is required before PCA

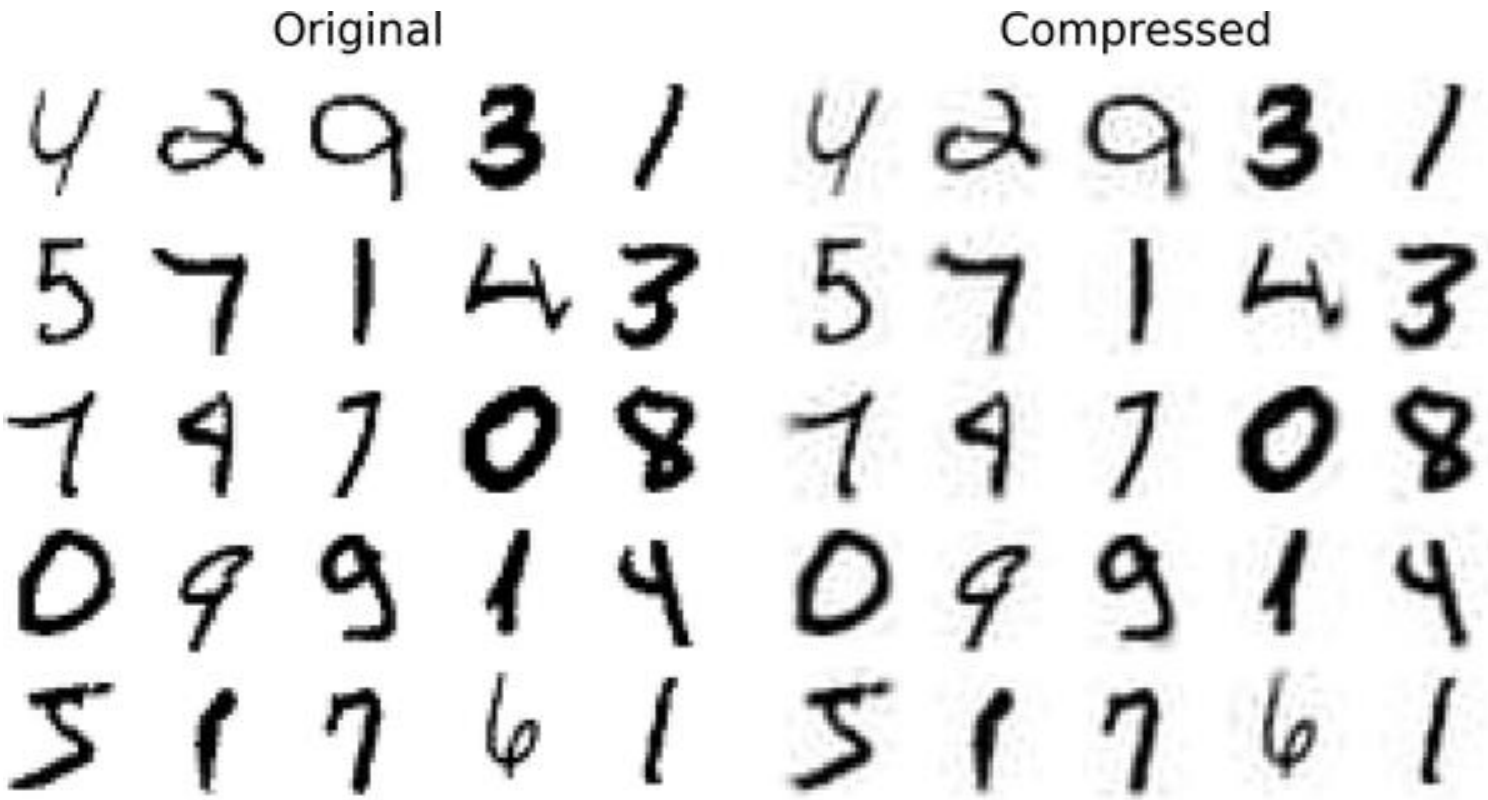
Notes

- *Why?*
- PCA is based on variance
- Different scales can skew results

and critical step.

PCA for Compression

- Apply PCA to MNIST preserving 95% of variance

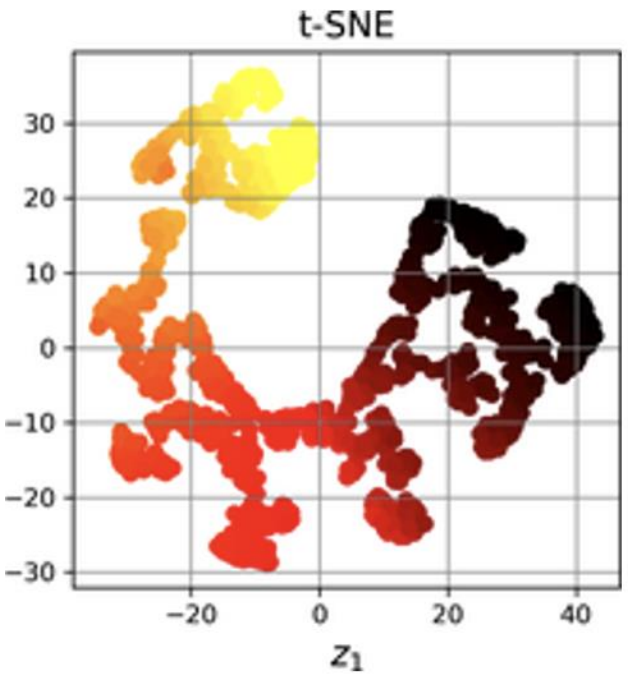


Issue of PCA

- Can be very slow for high-dimensional datasets
- Some speed-up implementations:
  - Randomized PCA: find approximation instead
  - Incremental PCA: use mini-batch approach
  - Random Projection: use random linear projection instead of PCA

t-SNE

- t-distributed stochastic neighbor embedding (t-SNE)
- A manifold learning method
- Keep similar instances close and dissimilar instances apart
- Mostly used for visualization
  - Visualize clusters of instances in high-dimensional space



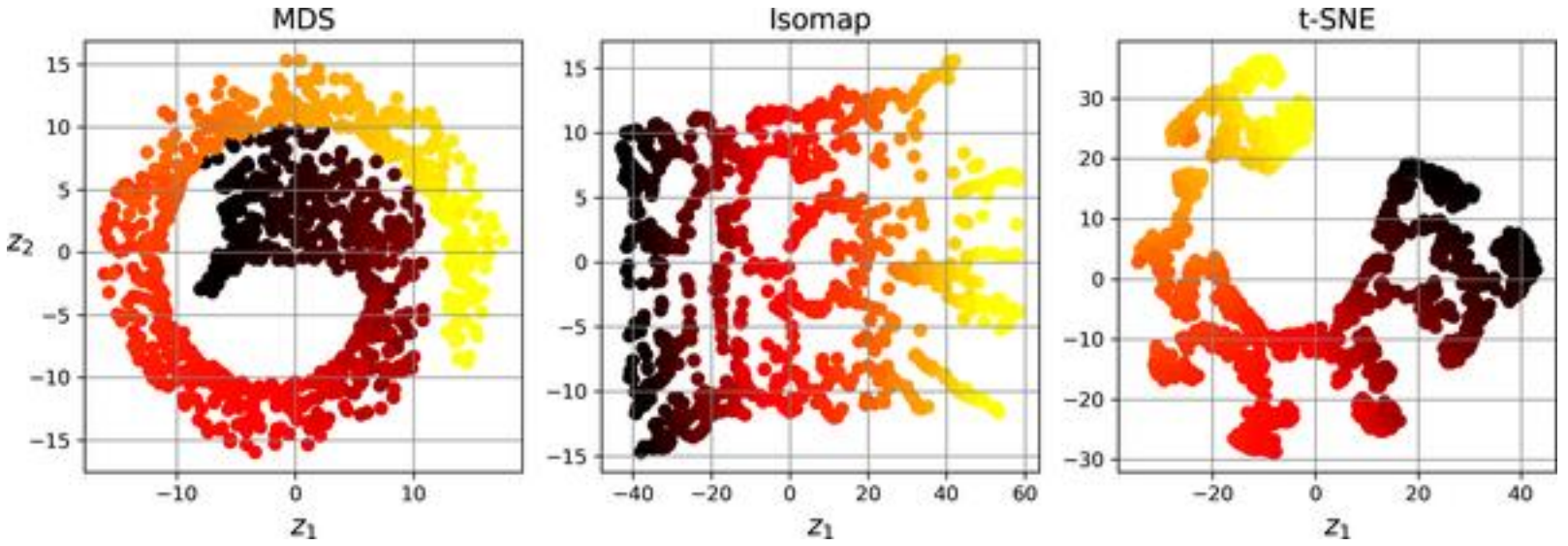
t-SNE Steps (Optional)

1. Feature scaling: ensure features are on a similar scale
2. Compute similarities in high-dimensional space
3. Initialize points in low-dimensional space
4. Repeat until convergence
   - Compute similarities in low-dimensional space
   - Minimize differences between two similarities
   - Update low-dimensional points

» https://distill.pub/2016/misread-tsne/

Other Techniques (Optional)

- Locally linear embedding (LLE): nonlinear dimensionality reduction, a manifold learning technique
- Multidimensional scaling (MDS): preserve distances between instances
- Isometric Mapping (Isomap): preserve geodesic distances
- Linear discriminant analysis (LDA): keep classes as far as possible



Scenarios and Choices

Scenario	Choices
Data is linear	PCA, LDA
Data is non-linear	t-SNE, Isomap
Visualization	t-SNE
Feature engineering	LDA, PCA
Interpretation	PCA
Large dataset (computational efficiency)	PCA
Robust to noise	PCA

Feature Engineering

- **Feature extraction**: transform raw data into informative features, and reduce dimensionality
  - Numeric data: mean, variance, standard deviation, histogram…
  - Text data: tf-idf, word embedding (in AI Essentials)
- **Feature selection**: choose a subset of existing features without transforming them
- **Feature creation**: generate new features instead of reducing
  - Categorical data encoding
- *PCA: feature extraction or feature selection?*

PCA is features extraction.

Feature Selection

- Univariate statistics
  - Statistically significant relationship between each feature and target
  - Select features with highest confidence
  - Also known as analysis of variance (ANOVA)
- Model-based selection
  - Use supervised machine learning model to judge the importance, may be different from the final supervised modeling
  - Common: decision trees, regularized linear regressions
- Iterative selection
  - A series of models are built
  - Start with zero and add features, or
  - Start with all and remove features

Lab 5

- PCA on breast cancer dataset
- PCA on MNIST
- PCA vs t-SNE for visualization
  - On a smaller hand-written digit dataset

Next Week

- Unsupervised Machine Learning, Part II
- Machine Learning Competition
  - Supervised + unsupervised
  - Wireless presentation instructions will be sent

Happy Thanksgiving and
Happy Machine Learning