Lecture 3

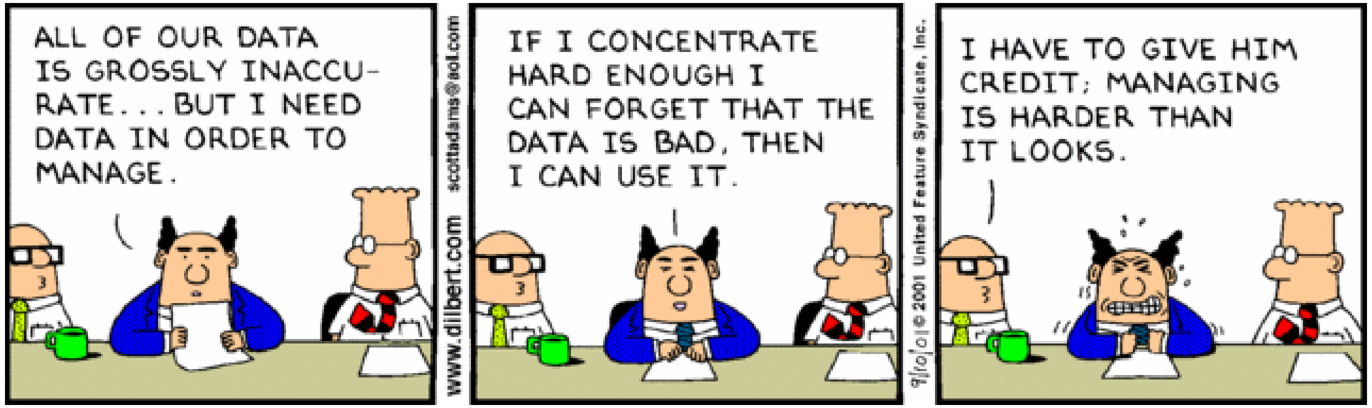
Date Preprocessing

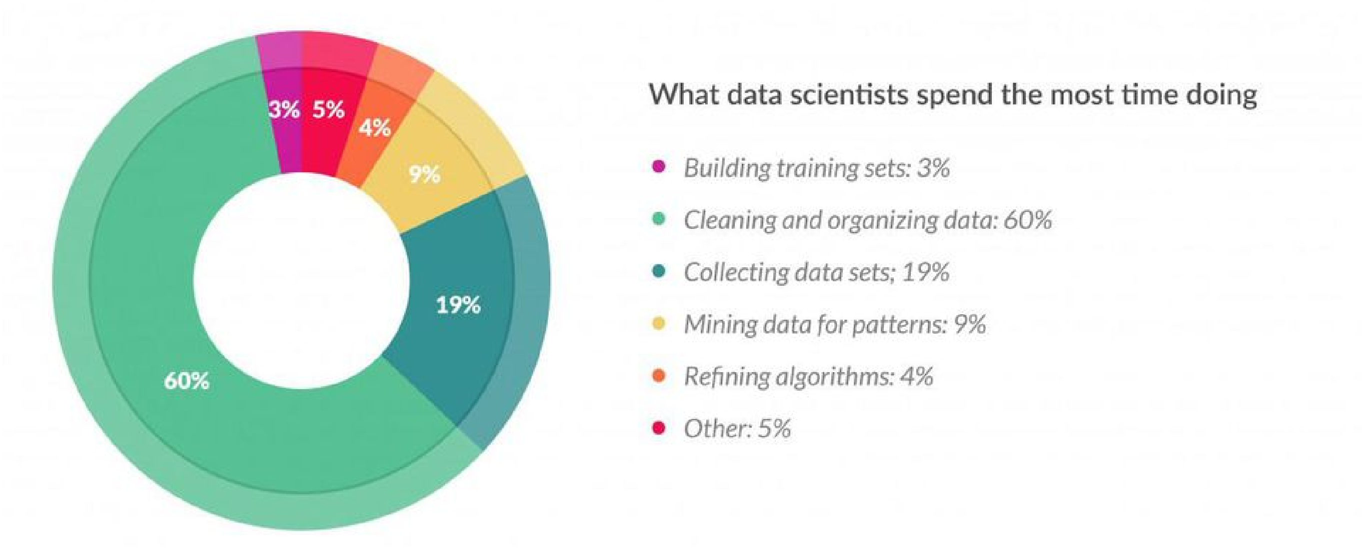
Ashish Dandekar

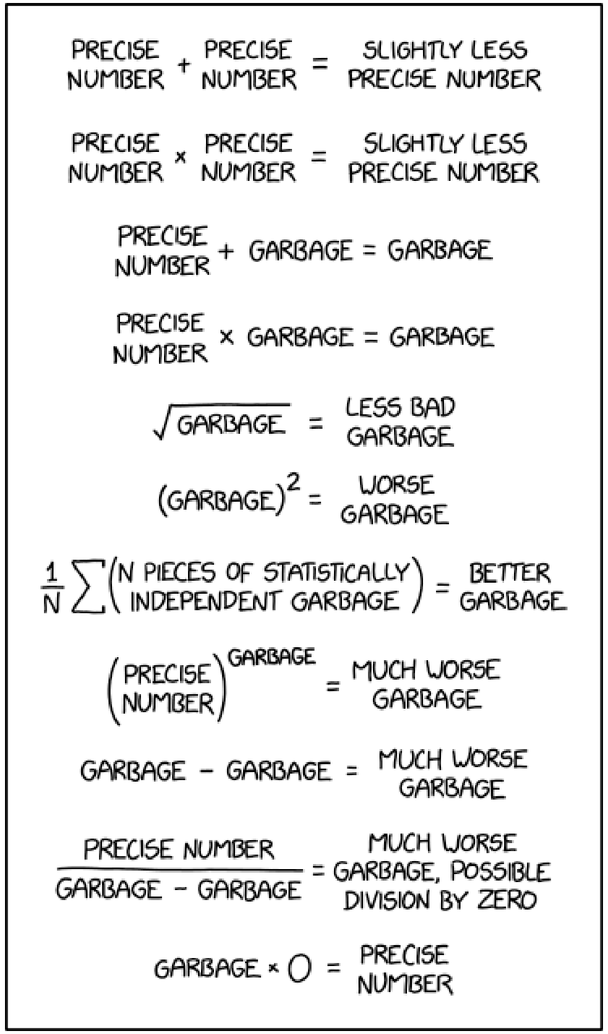
Lecture Overview

Descriptive Statistics
Probability
Hypothesis Testing

Why is it necessary?







Data Quality

Quality of Data

There are three metrics to assess the quality of data.

Accuracy.
Affected by the presence of erroneous data.

Completeness.
Affected by lacking features or values.

Consistency.
Affected by inconsistent aggregation of datasets.

https://developer.ibm.com/technologies/data-science/articles/

Missing Values

Common causes

- Attributes are not collected.
- Attributes are not applicable.

Handling missing values

- Remove datapoints with missing values.
- Remove attributes with missing values.
- *Data Imputation*

Age	Education	Married	Income	Credit Approval
23	Masters	No	75k	Yes
N.A.	Bachelors	Yes	50k	No
26	Masters	No	N.A.	Yes
41	PhD	No	95k	Yes
55	Masters	Yes	80k	No

Noisy Data

Common causes

- Faulty sensor readings.
- Data entry errors.
- Data transmission errors.
- Data format inconsistencies.
- Data unit inconsistencies.

Ways to handle

- Exploratory data analysis
- Use of the central statistical tendencies

Outliers

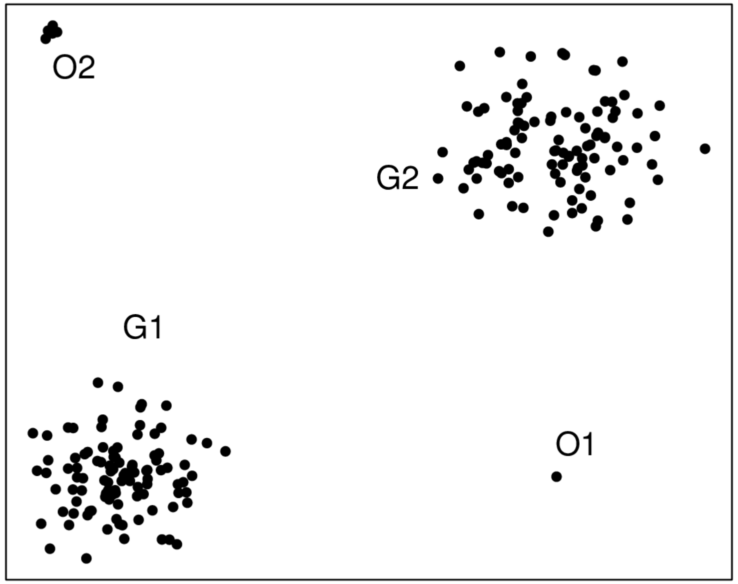
Data points that are *considerably* different than other data points.

Outliers are noise

- Negatively impact the analysis.
- Remove or use methods that subdue their effect.

Outliers are target

- They are the target of the learning.
- For instnce: Fraud detection, anomaly detection.



Data Preprocessing

Introduction

Why to do it?

- Improve data quality.
- Needed for the anaytics model.
- Remove complexity from data for the ease of analysis.

Typical steps

- Data cleaning
- Data reduction
- Data transformation

Why did deep learning become so popular?

Deep learning eliminates the need of data preprocessing to a large extent. We will learn more about this in Week 9.

Data cleaning

Improve data quality

- Remove or fill missing quality
- Identify and remove outliers
- Identify and remove/merge duplicates
- Correct errors and inconsistencies

Data cleaning requires inputs from domain experts.

Data reduction

Reducing the number of datapoints

- Sampling
- Commonly used for preliminary analysis
- Commonly used when dataset extremely large

Reducing the number of attributes

- Removing irrelevant attributes
- Dimensionality reduction

Reducing the number of attributes values

- Aggregation or generalisation
- Binning and smoothing

Data Aggregation

- Introduction
- Data cleaning
- ❯ Data reduction
- Data transformation

Data Aggregation

- Changing granularity of the numerical data
- Generalising the values of the categorical data

	id	age	gender	height
0	0	18393	2	168
1	1	20228	1	156
2	2	18857	1	165
3	3	17623	2	169
4	4	17474	1	156

	id	age	gender	height
0	0	50.0	2	168
1	1	55.0	1	156
2	2	51.0	1	165
3	3	48.0	2	169
4	4	47.0	1	156

Binning

- Sort the data
- Split data into equal bins
- Replace every datapoint with the average value of the respective bin

55 57 59 60 64 65 65 66 67 67 67 68 68 70 70 70 ...

55 57 59 60 64	65 65 66 67 67	67 68 68 70 70	70 ...
59 59 59 59 59	66 66 66 66 66	69 69 69 69 69	72 ...

Data Transformation

Feature Construction

- Creating features that are more meaning ful for analyses.
- Reduce dimensions by getting rid off unnecessary features.

Units Sold	Selling Cost	Production Cost
3	6	2
2	3	3
1	4	5
2	9	5

Units Sold	Profit/Unit
3	4
2	0
1	-1
2	4

Data Transformation

Normalisation

Min-max normalization

$$x_i^{weight} = \frac{x_i^{weight} - min(x^{weight})}{max(x^{weight}) - min(x^{weight})}$$

weight
62.0
85.0
64.0
82.0
56.0

weight
0.273684
0.394737
0.284211
0.378947
0.242105

$$x_i^{weight} = \frac{x_i^{weight} - \mu^{weight}}{\sigma^{weight}}$$

Standardization
(z-score normalization)

weight
-0.847867
0.749826
-0.708937
0.541431
-1.264657

Min-max normalisation

Transforms data in the range $[0, 1]$.

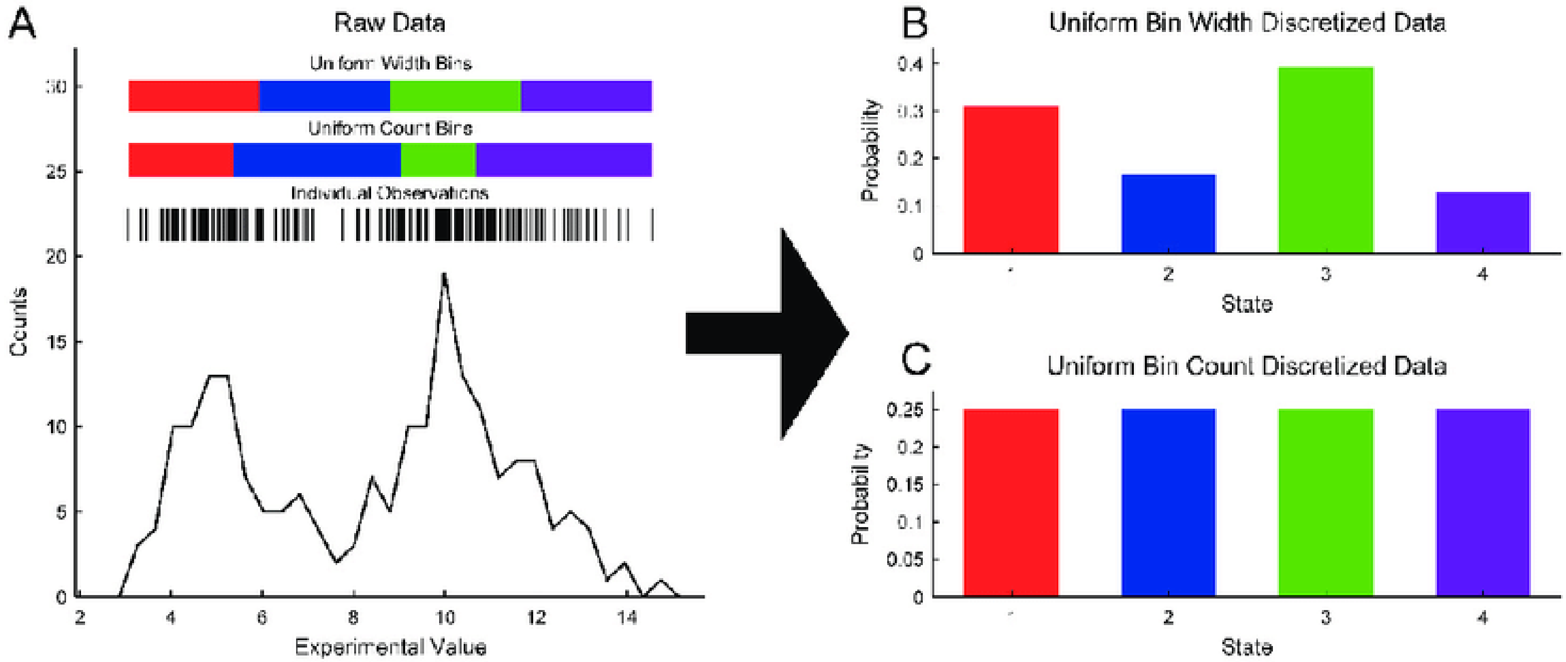
Z normalisation (Standardisation)

Transforms data in the range $[-\infty, \infty]$.

Data Transformation

Discretisation

- It is used to convert numerical data to categorical data
- It is used to convert a regression problem to a classification problem



Data Transformation

Introduction
Data cleaning
Data reduction
❯ Data transformation
Feature construction
Normalisation
Discretisation
Encoding

Encoding

It is used to convert categorical data to numerical data data

Data

City
A
B
C

Ordinal Coding

City_Code
1
2
3

One-hot Encoding

C_A	C_B	C_C
1	0	0
0	1	0
0	0	1

Dummy Variables

	C_1	C_2
A	1	0
B	0	1
C	0	0

Linear algebra (quick review)

Linear algebra review

Let's assume that $\mathbf{x} \in \mathbb{R}^d$, $A \in \mathbb{R}^{n \times d}$, and $B \in \mathbb{R}^{d \times k}$.

Linear combination

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ are d-dimensional vectors. The linear combination of these vectors for some scalars a_is is defines as

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + ... + a_n\mathbf{x}_n$$

Eigenvalue equation

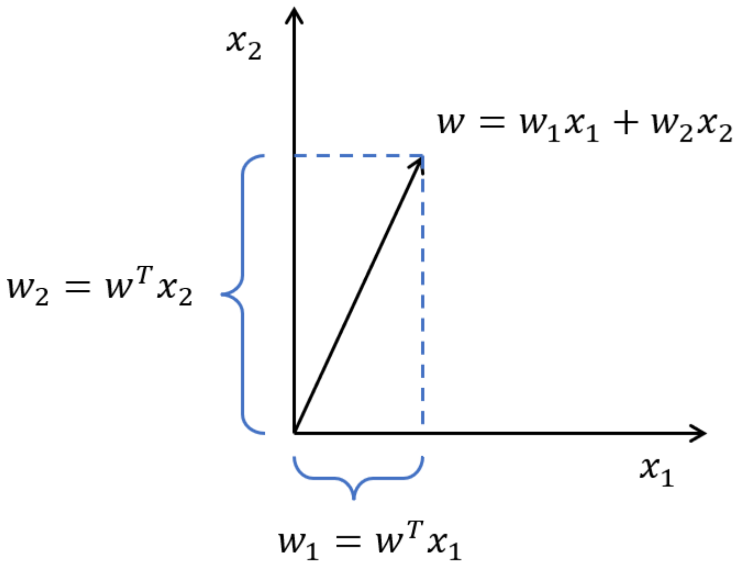
$$A\mathbf{x} = \lambda\mathbf{x}$$

Scalars	Vectors	Matrices
$\mathbf{x}^T\mathbf{x} \in \mathbb{R}$	$A\mathbf{x} \in \mathbb{R}^n$ $B^T\mathbf{x} \in \mathbb{R}^k$	$A^T \in \mathbb{R}^{d \times n}$ $A^TA \in \mathbb{R}^{d \times d}$

$\mathbf{y}$	$d\mathbf{y}/d\mathbf{x}$
$A\mathbf{x}$	A^T
$\mathbf{x}^TA$	A
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^TA\mathbf{x}$	$(A + A^T)\mathbf{x}$

Interpretation

What is the meaning of $w^T x$?



Vector-vector multiplication provides us the magnitude of projection of one vector on the other one.

What is the meaning of $W^T x$?

$$\begin{pmatrix} \text{———} & w_1^T & \text{———} \\ \text{———} & w_2^T & \text{———} \\ & \vdots & \\ \text{———} & w_d^T & \text{———} \end{pmatrix} x = \begin{bmatrix} w_1^T x \\ w_2^T x \\ \vdots \\ w_d^T x \end{bmatrix}$$

Matrix-vector multiplication takes the vector a new vector space spanned by the columns of the matrix.

Bag of Words

Bag of words

How to mathematically represent the following statements?

- I like apples.
- I love oranges.
- I like bananas.

Let's denote the words by ids.

I	like	love	apples	oranges	bananas
1	2	3	4	5	6

Vectorized versions

Word	Vector
I	$[1, 0, 0, 0, 0, 0]$
like	$[0, 1, 0, 0, 0, 0]$
apples	$[0, 0, 0, 1, 0, 0]$
I like apples.	$[1, 1, 0, 1, 0, 0]$

Question

How will you find if `I like apples` contains the word `oranges`?

Bag of Words

Review
Interpretation
❯ Bag of Words

What is a document?

Document is a vector in the vector space spanned by the words.

A document: `(I like apples. I love oranges. I like bananas.)` will be represented as: $[3, 2, 1, 1, 1, 1]$

Question

How will you interpret the following matrix?

$$W = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 2 \end{bmatrix}$$

Question

How will you interpret $W^T d$, where d is a document vector?

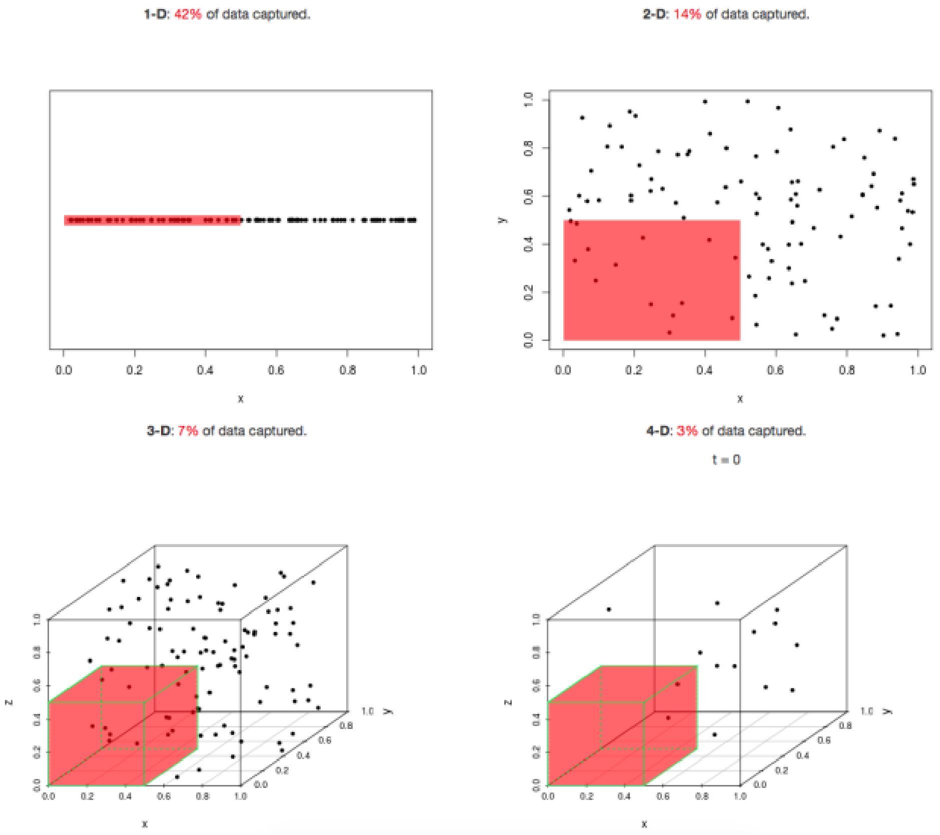
Dimensionality Reduction

Curse of Dimensionality

Higher number of dimensions lead to sparser data!

Skewed Distances

- Datapoints tend to never be close together.
- It tends to be difficult spot outliers.
- Points that are similar in lower dimensions might not be similar in higher dimensions.

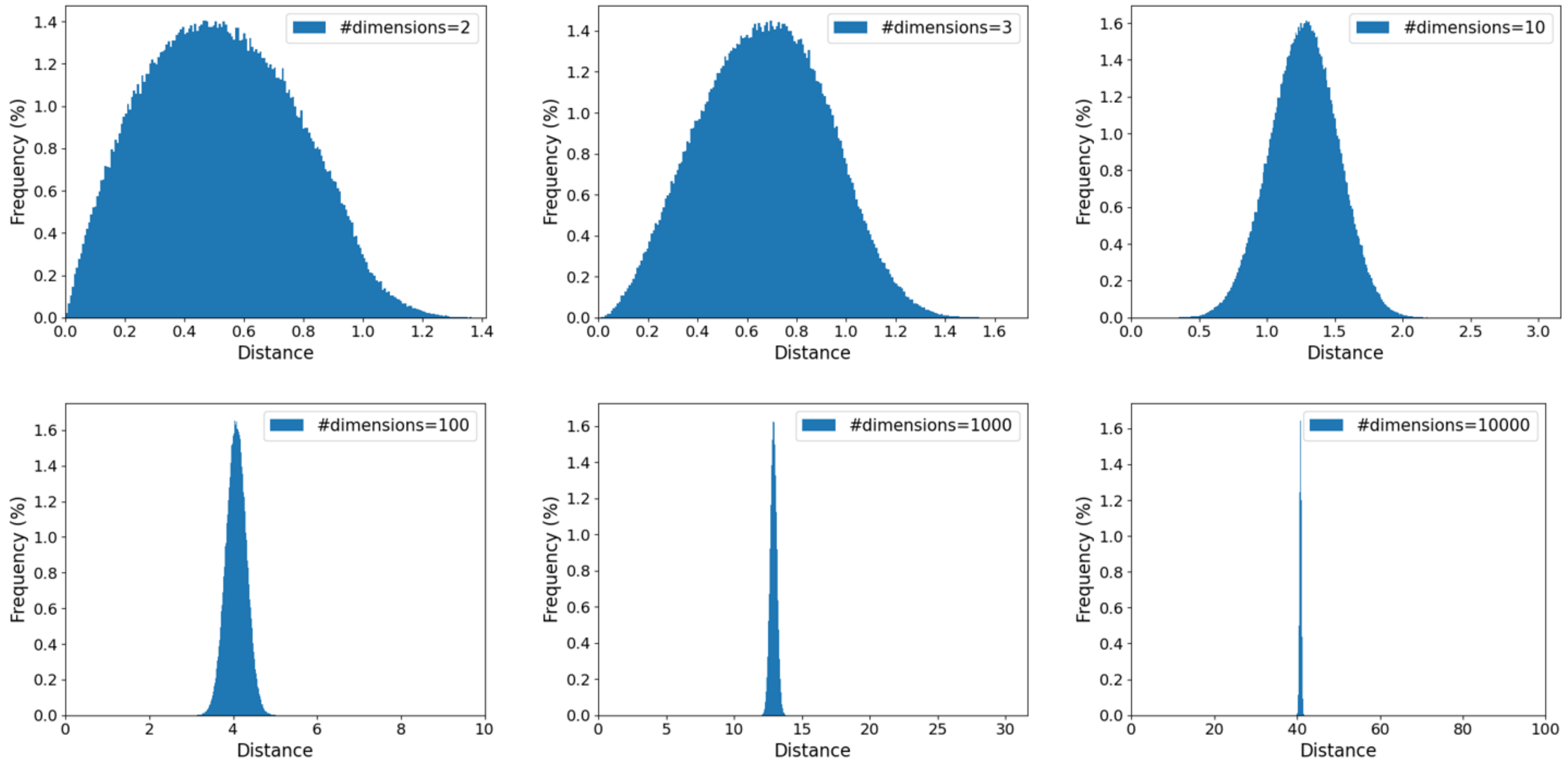


Source: https://eranraviv.com/

Motivation

- **Motivation**
- Heuristic based
- PCA

Distribution of pairwise distances between 1,000 random data points and different number of dimensions.



Heuristic based methods

Missing value ratio.
Remove features with large missing values.

Low variance filter.
Remove features that do not significantly change.

High correlation filter.
Remove features that are strongly correalted with each other.

F1	F2	F3	F4
3	5	3	6
	5	4	8
1	5	6	13
	6	5	10
	5	2	4
	5	2	3

Principle Component Analysis.

Dimensionality reduction *through linear transformations.*

Data Representation

Vectors

We represent datapoints as d-dimenstional vectors, i.e. $x_i \in \mathbb{R}^d$.

Matrices

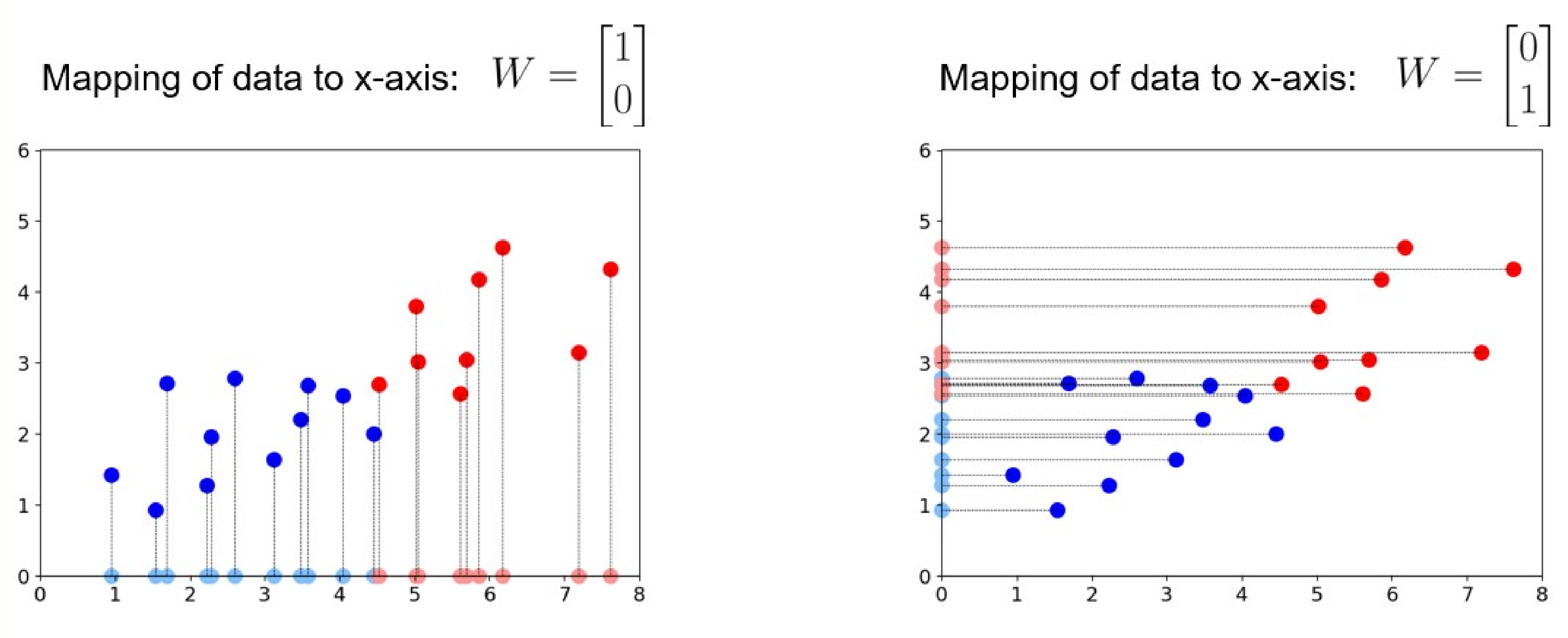
We represent dataset as $n \times d$-dimenstional matrices, i.e. $X \in \mathbb{R}^{n \times d}$.

$$\overset{X}{\begin{bmatrix} n \times d \end{bmatrix}} \overset{W}{\begin{bmatrix} d \times k \end{bmatrix}} = \overset{X'}{\begin{bmatrix} n \times k \end{bmatrix}}$$

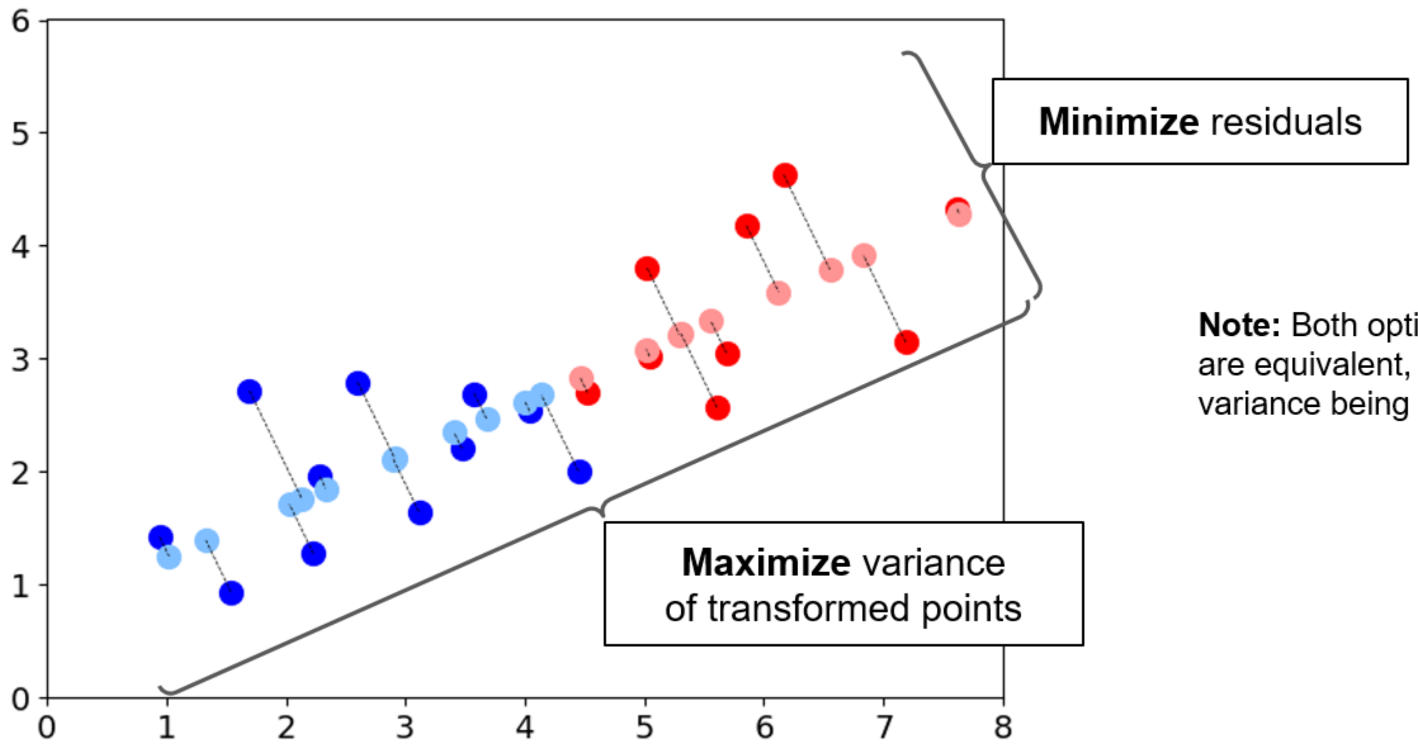
How to find W?

Intuition

Which of the following is a better transformation?



Intuition



Note: Both optimization objectives are equivalent, with maximizing the variance being easier to handle

We want to choose the direction that provides the maximum variance!

Derivation

NOT part of any assessment!

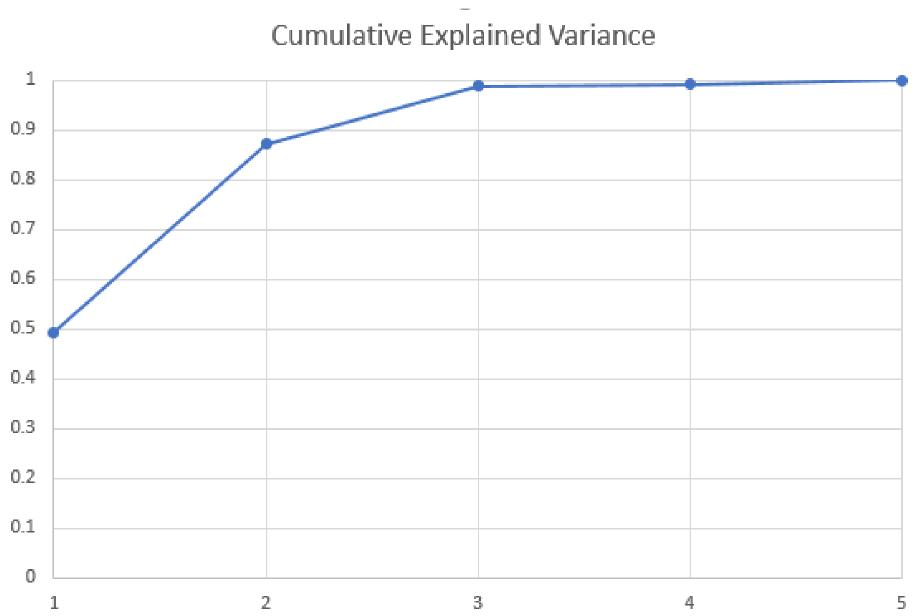
Let w_1 represents the direction of maximum variance. Thus we want to find

$$
\begin{aligned}
w_1 &= \arg\max_w \frac{1}{n} \sum_i (w^T x_i - 0)^2 && \text{(mean centered)} \\
&= \arg\max_w \frac{1}{n} // Xw //^2 \\
&= \arg\max_w \frac{1}{n} w^T (X^T X) w \\
&= \arg\max_w w^T C w && \text{(covariance matrix)} \\
&= \arg\max_w \frac{w^T C w}{w^T w} && \text{(Rayleigh's coefficient (max-eigenvalue))}
\end{aligned}
$$

Choosing k

How to choose $k < d$?

- Sort the eigenvalues in the descending order.
- Explained variance of the j^{th} eigenvalue is $\lambda_j / \sum_i \lambda_i$
- Choose top-k dimesions that explain *most* of the data.



How to code it?

Python code

```
# Let X denote the data

# Standardizing the data
from sklearn.preprocessing import StandardScaler
X = StandardScaler().fit_transform(X)

# Computing PCA
from sklearn.decomposition import PCA
pca = PCA(0.9) # explains 90% of the data
pca.fit(X)

# print the value of k
pca.n_components_

# transform the data
X_lower = pca.transform(X)
```

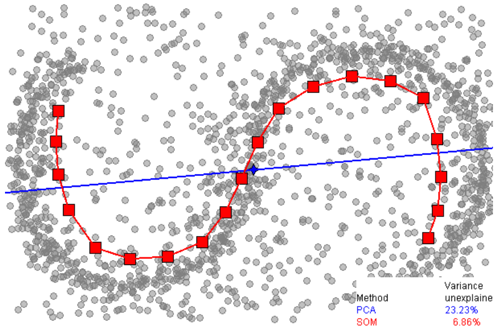
Summary

Pros

- Reduces dimensions without the domain knowledge about the individual dimensions.
- Significantly reduces the amount data.
- Helps visualising high dimensional data.

Cons

- Loss of semantics.
- Only captures linear correlations.



- Does not take into account data labels.

Summary

Summary

Data Quality

- Missing Values
- Noisy Data
- Outliers

Date Preprocessing

- Data Cleaning
- Data Reduction
- Data Transformation

Primer on the Linear Algebra

Dimensionality Reduction

- Heuristic based techniques
- Principal Component Analysis