

# **FOUNDATION OF DATA SCIENCE**



## **3. Data Understanding and Preprocessing**

Sushant Chalise

# Material Adaptation



Introduction to Data Mining : Tan, Steinbach, Karpatne, Kumar  
Rajad Shakya  
Assoc. Prof. Arun K Timalina, PhD  
Assoc. Prof. Sanjeeb P Panday, PhD  
Internet

# Syllabus (10hrs)

- Types of data:
  - Structured, unstructured, semi-structured
- Data preprocessing requirements
- Data sources and collection methods
- Data cleaning and preparation

# Syllabus (10hrs)

- Data wrangling and associated tools
- Data enrichment, validation and publishing
- Data transformation and normalization
- Dimensionality reduction linear factor model, principal component analysis(PCA)

# ALL ABOUT DATA

The background is a solid teal color. It features several faint, semi-transparent data visualizations. A large donut chart is positioned in the upper right quadrant. To its right and below are several smaller pie charts of varying sizes. In the bottom right corner, there is a bar chart with four vertical bars of increasing height from left to right.

Sushant Chalise

# Data

- What is data?
- What is information?
- What is knowledge?
- Is there a difference?



# What are these?

23, 22, 23, 24, 30, 28, 27



# What are these?

23, 22, 23, 24, 30, 28, 27

23°C, 22°C, 23°C, 24°C, 30°C, 28°C, 27°C





# What are these?

23, 22, 23, 24, 30, 28, 27

23°C, 22°C, 23°C, 24°C, 30°C, 28°C, 27°C

Maximum Temperature of second week of July

# What are these?

23, 22, 23, 24, 30, 28, 27

23°C, 22°C, 23°C, 24°C, 30°C, 28°C, 27°C

Maximum Temperature of second week of July in Kathmandu

What clothings should I bring when I visit  
Kathmandu in July?

# What is data?

- Is an organization or systematic record of a particular quantity.
- It is the different values of that quantity represented together in a set.
- It might not have significant on itself.
- It is the collection of data objects and its attributes.
  - Attributes - property or characteristics of an object
  - Data objects - a record, point case, sample, instance.
- It is the collection of facts and figures to be used for a specific purpose.
- In, computing it is an information that is translated in a form that is efficient for transfer, movement and processing.

# What is information?

- Information is defined as classified or organized data that has some meaningful value for the user.
- Information is also the processed data used to make decisions and take action.
- Information is the output that results from analyzing, contextualizing, structuring, interpreting the data.
- In a nutshell, data can be a number, symbol, character, word, codes, graphs, etc.
- On the other hand, information is data put into context. Information is utilised by humans in some significant way.
- Knowledge can be to make decisions, forecasts, predict, etc.

Data	Information
Data is unorganised and unrefined facts	Information comprises processed, organised data presented in a meaningful context
Data is an individual unit that contains raw materials which do not carry any specific meaning.	Information is a group of data that collectively carries a logical meaning.
Data doesn't depend on information.	Information depends on data.
Raw data alone is insufficient for decision making	Information is sufficient for decision making
An example of data is a student's test score	The average score of a class is the information derived from the given data.

# Data Classification

- Nominal
- Ordinal
- Interval
- Ratio



These are often referred to as types of attributes.

- Record
  - Data Matrix, Document, Transaction
- Graph
  - WWW, Molecular Structure
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

- Structured
- Semi-Structured
- UnStructured

- Continuous
- Discrete



# Structured Data

- highly organized and easy to process.
- It is typically stored in tabular form, such as in relational databases or spreadsheets
- Each data element is identifiable by a specific data type (such as integer, string, date) and follows a defined schema.

# Structured Data: Characteristics

- Organized: Data is arranged in rows and columns with a predefined structure.
- Data Types: Typically includes numbers, strings, dates, and booleans.
- Queryable: Can be easily queried using SQL (Structured Query Language).
- Ease of Processing: Since the data is well-organized, it is easier to clean, preprocess, and analyze.



# Structured Data



## Demographic Data

- Age
- Current Location
- Email
- Mailing Address
- Name
- Telephone number

## Firmographic Data

- Company Address
- Company Name
- Industry
- Number of Employees
- Revenue

## Behavioral Data

- Email Open Rates
- Product and Service Usage Patterns
- Purchase Patterns
- Social Media Engagement
- Videos and Content Consumed
- Web Activity history

## Transactional Data

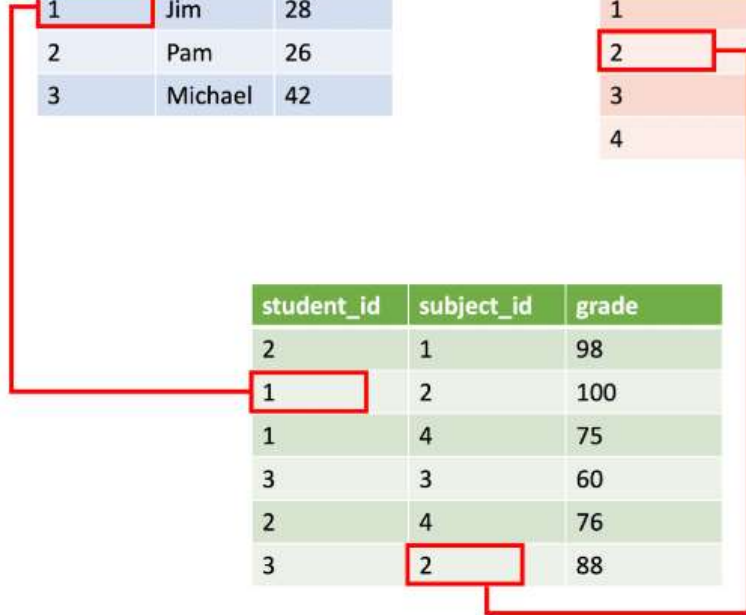
- Credit Card Payments
- Insurance Claims
- Invoices
- Purchase Orders
- Sales Orders
- Shipping Documents

# Structured Data

id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88



# Unstructured Data

- data that does not have a predefined data model or organization.
- difficult to store and analyze using traditional databases and requires advanced techniques for processing and understanding.

# Unstructured Data: Characteristics

- No Fixed Structure: The data does not fit into a tabular format or predefined schema.
- Variety of Formats: Includes text, images, audio, video, social media posts, etc.
- Difficult to Process: Requires more complex tools like natural language processing (NLP) for text data or image processing techniques for visual data.

# Unstructured Data



# Semi-structured Data

- mix between structured and unstructured data.
- It does not conform to the strict schema of structured data
- but has some organizational properties that make it easier to analyze than unstructured data.

# Semi-structured Data: Characteristics

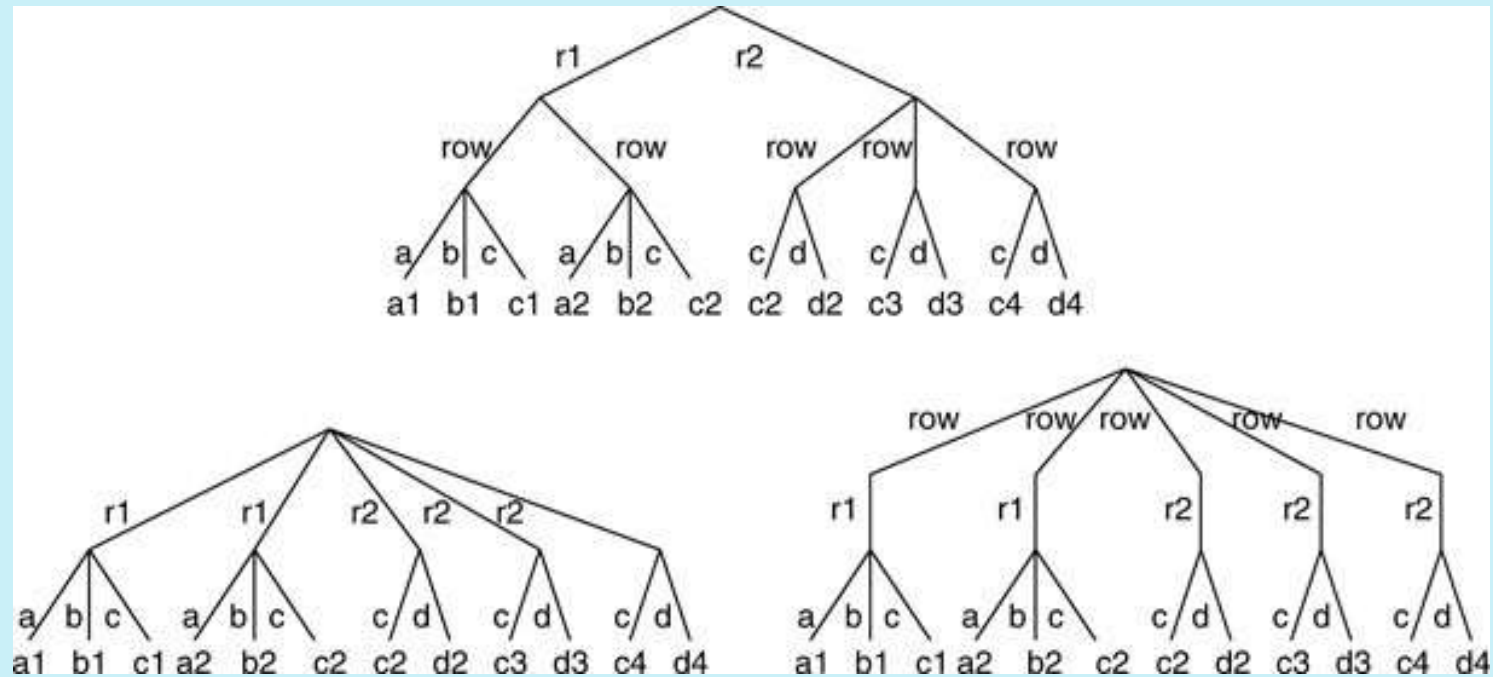
- Partial Structure:
  - Has some structure (e.g., key-value pairs, tags, or metadata) but lacks the rigid structure of rows and columns like structured data.
- Flexible Schema:
  - Can evolve over time and does not require a fixed schema.

# Semi-structured Data: Characteristics

- Easy to Parse:
  - While not tabular, semi-structured data is still easier to parse and process than unstructured data, using formats like JSON, XML, or NoSQL databases.



# Semistructured Data



# At a Glance

## Unstructured data

The university has 5600 students.  
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.  
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

## Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

## Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

# What is Big Data?



# Data preprocessing requirements

Sushant Chalise

# Why do we need Data Preprocessing?



# Why do we need Data Preprocessing?

Data in the real word is dirty.



# Why do we need Data Preprocessing?

Data in the real word is dirty.  
It is inconsistent.



# Why do we need Data Preprocessing?

Data in the real word is dirty.  
It is inconsistent.  
It is noisy.



# Why do we need Data Preprocessing?

Data in the real word is dirty.  
It is inconsistent.  
It is noisy.  
It is incomplete.  
And MOST of all no quality.

# Misleading Statistics



# Misleading Statistics



**"80% OF  
DENTISTS  
RECOMMEND  
COLGATE."**

**BUT THEY DIDN'T NECESSARILY PREFER COLGATE  
OVER OTHER BRANDS. THE AD DOESN'T TELL YOU  
THAT THE DENTISTS IN THE SURVEY WERE  
PERMITTED TO CHOOSE MORE THAN ONE BRAND. BUT  
"GIVEN A CHOICE BETWEEN USING THESE DENTAL  
HYGIENE PRODUCTS (ONE OF WHICH IS COLGATE)  
AND BRUSHING ALONE, 80% OF DENTISTS  
RECOMMEND USING TOOTHPASTE" ISN'T QUITE  
AS CATCHY.**

**CRACKED.COM**

# Data preprocessing requirements

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=" "
- noisy: containing errors or outliers
  - e.g., Salary="-10"
- inconsistent: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
  - e.g., discrepancy between duplicate records

# Data preprocessing requirements

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
  - Duplicate records also need data cleaning

# Data preprocessing requirements

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

# Data preprocessing requirements

- raw data is often incomplete, inconsistent, and in a format that is not suitable for analysis.
- Preprocessing ensures that the data is cleaned, formatted, and transformed into a usable state.
- Before proceeding, it's important to ensure that the data is reliable and free from errors.

# Data preprocessing

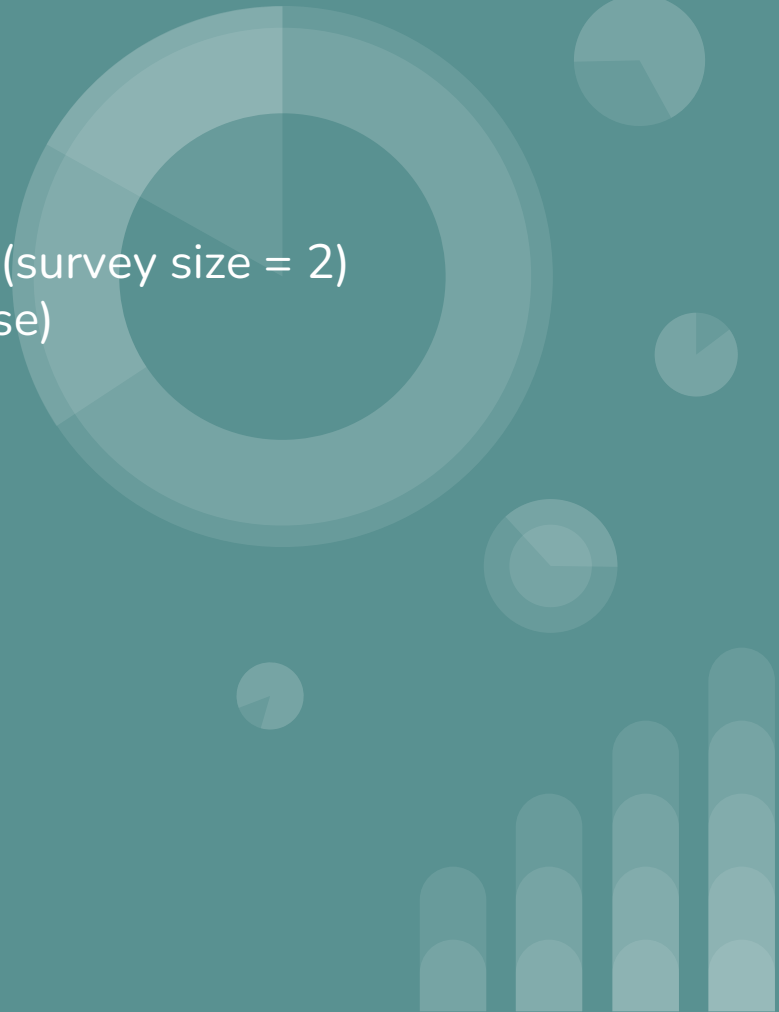
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data





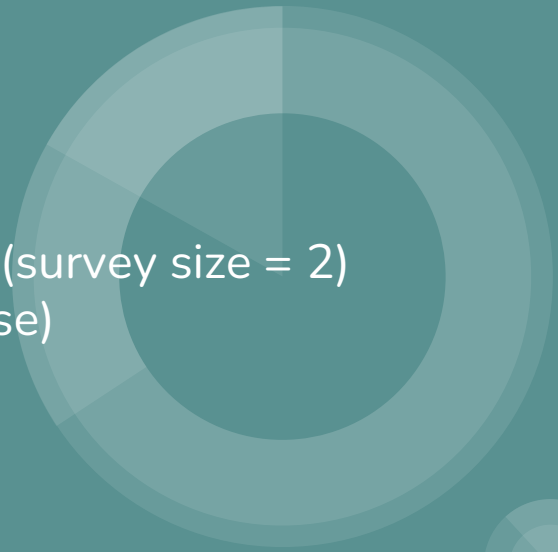
# Misleading Statistics

- 50% of graphic designers have red hair (survey size = 2)
- 50% of all marriages end in divorce (false)



# Misleading Statistics

- 50% of graphic designers have red hair (survey size = 2)
- 50% of all marriages end in divorce (false)



# Data Sources and Collection Methods

Sushant Chalise

# Need of Data Collection

- Before a judge makes a ruling in a court case
- or a general creates a plan of attack,
- they must have as many relevant facts as possible.
- Whether you're in academia, trying to conduct research,
- or part of the commercial sector, thinking of how to promote a new product,
- you need data collection to help you make better choices.

# When you plan to buy a new phone



# When you plan to buy a new phone



# When you plan to buy a new phone



# Primary Data

- Data collected directly from the source. This could include:
  - Surveys (customer feedback, market research).
  - Experiments (controlled observations in laboratories).
  - Sensors and IoT devices (temperature, motion, environmental data).
  - Direct user inputs (forms, interviews).



# Primary Data

- Advantages:
  - Tailored for specific research or project needs.
  - High accuracy and relevance.
- Challenges:
  - Time-consuming and expensive.
  - Requires significant effort for collection.

# Secondary Data

- Pre-existing data collected by someone else for a different purpose.
  - Government databases (census data, economic statistics).
  - Public datasets (Kaggle, UCI Machine Learning Repository).
  - Online repositories (GitHub, open data portals).

# Secondary Data

- Advantages:
  - Cost-effective and readily available.
  - Useful for benchmarking and comparative studies.
- Challenges:
  - May not fit the exact requirements.
  - Risk of outdated or irrelevant data.

# Primary and Secondary Sources of Data

## A) Primary data

- Primary data means first-hand information collected by an investigator.
- It is collected for the first time.
- It is original and more reliable.
- For example, the population census conducted by the government of Nepal after every ten years is primary data.

## B) Secondary data

- Secondary data refers to second-hand information.
- It is not originally collected and rather obtained from already published or unpublished sources.
- For example, the address of a person taken from the telephone directory or the phone number of a company taken from Truecaller are secondary data.

# Other Classification of Data Source

- Internal data: Created by organizational processes, including email marketing, customer profiles, and online activity.
- External data: Derived from outside sources like social media, historical demographic data, and websites.
- Third-party analytics: Provided through analytics platforms like Google Analytics.
- Open data: Free, public-accessible data, like government and health and science data.

# Collection Methods



# How does ACP Pradyuman makes an arrest?



# How does ACP Pradyuman makes an arrest?



## 1. Surveys and Questionnaires



# How does ACP Pradyuman makes an arrest?



## 2. Experiments

# Collection Methods

- Surveys and Questionnaires
- Interviews
- Observations
- Experiments
- Focus Groups



# Other Collection Methods

- Published Sources
- Online Databases
- Government and Institutional Records
- Publicly Available Data
- Past Research Studies



# Collection Methods

- Surveys and Questionnaires:
  - Researchers design structured questionnaires or surveys to collect data from individuals or groups.
  - These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms.
- Interviews:
  - Interviews involve direct interaction between the researcher and the respondent.
  - They can be conducted in person, over the phone, or through video conferencing.
  - Interviews can be structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational).

# Collection Methods

- Observations:
  - Researchers observe and record behaviors, actions, or events in their natural setting.
  - This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention.
- Experiments:
  - Experimental studies involve manipulating variables to observe their impact on the outcome.
  - Researchers control the conditions and collect data to conclude cause-and-effect relationships.
- Focus Groups:
  - Focus groups bring together a small group of individuals who discuss specific topics in a moderated setting.
  - This method helps in understanding the opinions, perceptions, and experiences shared by the participants.

# Data Collection - Primary



# Other Collection Methods

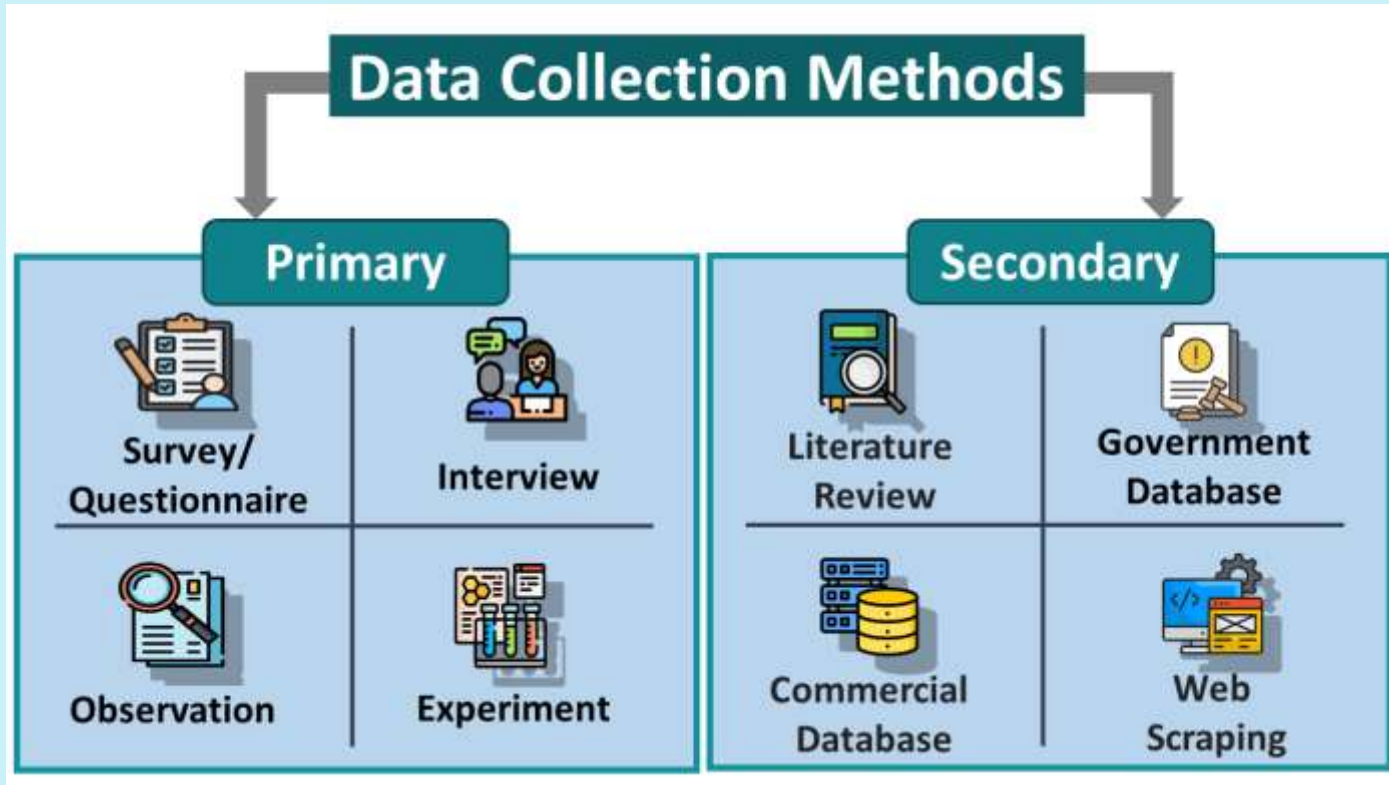
- Published Sources:
  - Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data.
- Online Databases:
  - Numerous online **databases** provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.
- Government and Institutional Records:
  - Government agencies, research institutions, and organizations often maintain databases or records that can be used for research purposes.

# Other Collection Methods

- Publicly Available Data:
  - Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research.
- Past Research Studies:
  - Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyze the data to gain insights or build upon existing knowledge.



# Data Collection Methods



# Other Methods of Collection

- Automated,
- Manual,
- Crowdsourced, etc.



# Automated Collection

- Collecting data using automated tools and technologies.
- Web Scraping:
  - Extracting data from websites using tools like BeautifulSoup or Scrapy.

# Automated Collection

- APIs:
  - Collecting data programmatically from services like Twitter, Google Maps, or OpenWeatherMap.
- Sensors and IoT Devices:
  - Automatic logging of environmental data, fitness tracker logs, etc.

# Automated Collection

- Advantages:
  - Scalable for large datasets.
  - Reduces human error.
- Challenges:
  - Requires technical expertise to set up.
  - Legal and ethical considerations, such as web scraping permissions.

# Manual Collection

- Data collected manually by humans through active involvement.
  - Surveys (Google Forms, Typeform).
  - Interviews or focus groups.
  - Observational studies.

# Manual Collection

- Advantages:
  - Can capture nuanced and qualitative data.
  - Flexible and adaptable to specific needs.
- Challenges:
  - Time-intensive and prone to human error.
  - Limited scalability.

# Crowdsourced Collection

- Leveraging the collective effort of a group to gather data.
  - OpenStreetMap for geospatial data.
  - Public contributions (Wikipedia edits, community polls).



# Crowdsourced Collection

- Advantages:
  - Cost-efficient for large-scale data collection.
  - Promotes diverse perspectives.
- Challenges:
  - May have quality issues due to varied contributor expertise.
  - Requires validation and cleaning.

# Data Cleaning and Preparation

Sushant Chalise

# Data Cleaning

- process of identifying and rectifying errors, inconsistencies, and missing values in a dataset to ensure it is in the right format for analysis.
- goal is to improve data quality and make it ready for use in building machine learning models

# Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, representational, and accessibility

# Data Cleaning and Preparation Includes

- Understanding the Data
- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

# 1: Understanding the Data

- Exploratory Data Analysis (EDA):
  - Visualizing the data using histograms, scatter plots, and box plots to understand distributions and relationships.
  - Analyzing summary statistics (mean, median, standard deviation) to spot outliers and anomalies.

# Anscombe's quartet

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

# Anscombe's quartet

Anscombe's quartet

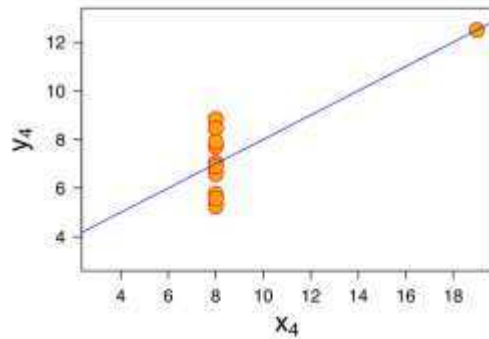
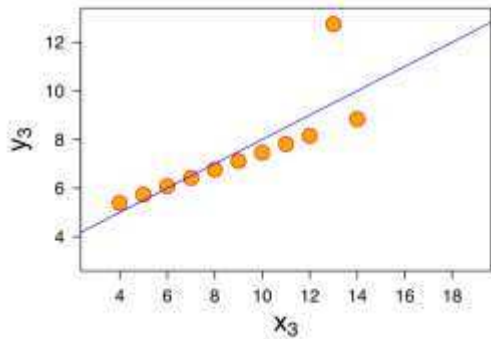
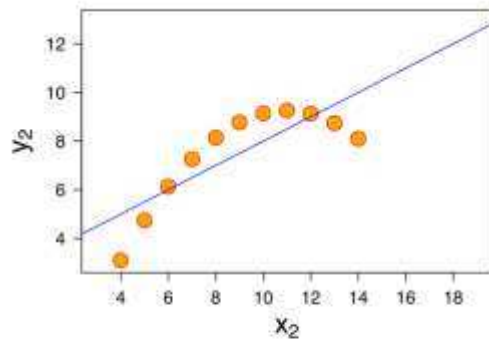
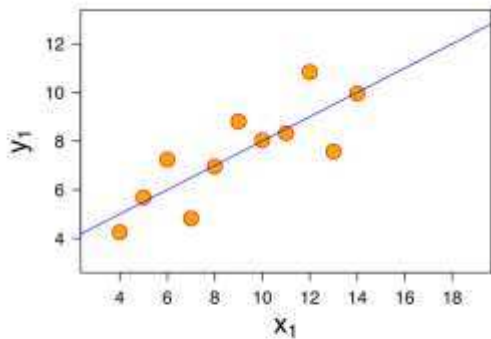
Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: $s_x^2$	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: $s_y^2$	4.125	$\pm 0.003$
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: $R^2$	0.67	to 2 decimal places



# Anscombe's quartet



## 2: Handling Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

## 2: Handling Missing Data

- Removing Missing Data:
  - Drop rows or columns with missing values if the loss of data is minimal and won't affect analysis.
  - done using `dropna()` in pandas or similar methods in other libraries.

## 2: Handling Missing Data

- Imputation:
  - Filling in missing values with the mean, median, or mode of the column (for numerical data).
  - done using `df.fillna(df.mean())` in pandas or similar methods in other libraries.
  - the most probable value: inference-based such as Bayesian formula or decision tree.

### 3: Removing Duplicates

- Duplicate entries can skew results and create bias in analysis.
- Use `drop_duplicates()` to remove rows that are identical across all or specific columns.

## 4: Handling Outliers

- Outliers are data points that significantly differ from other observations in the dataset.
- If outliers are due to data entry errors, remove them.
- Visualizations like box plots or z-scores help identify outliers.

## 5: Handling Inconsistent Data

- Inconsistencies occur when the data is not uniform across the dataset.
- Different formats (e.g., dates in MM-DD-YYYY vs. YYYY-MM-DD).
- Categorical data with inconsistent labels (e.g., Male, M, male).

## 6: Handling Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention



# 6: Handling Noisy Data

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g.,
  - deal with possible outliers)

# Disadvantages

- Analysts may lose out on actionable insights due to incomplete data.
  - This is very common in case where missing observations and outliers are dropped.
  - It may lead to an even bigger problem when automated.
- Some automated data cleaning tools are not very smart and may end up mishandling some observations in the data set.
- It is time- consuming
  - Data cleaning may take a lot of time, especially when dealing with large data.
- The process is very expensive.

# Data wrangling and associated tools

Sushant Chalise

# Data Wrangling

- process of transforming raw data into a more usable format.
- important for ensuring that your data is high quality and well-structured, which is crucial for accurate data analysis.
- Poorly wrangled data can lead to flawed conclusions

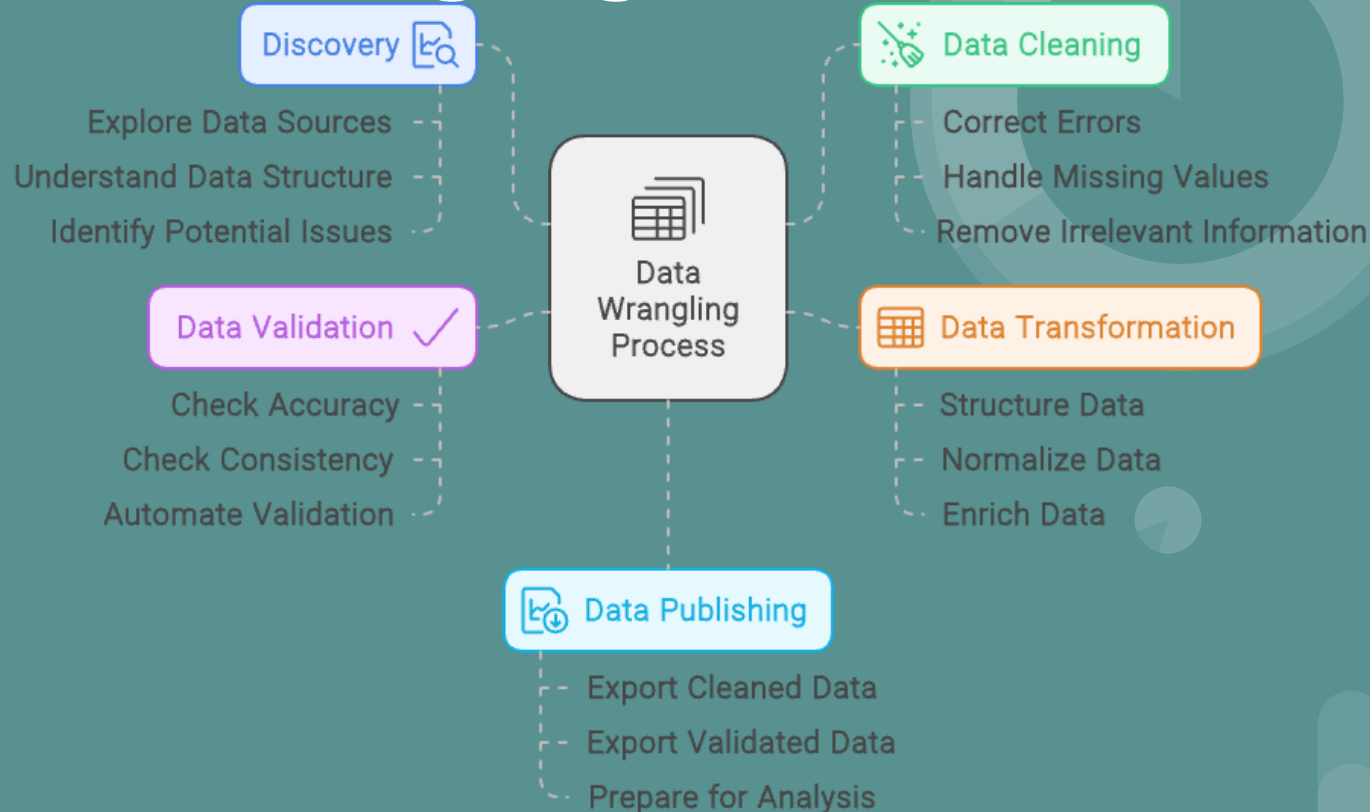
# Data Wrangling

- Data wrangling is the process of transforming raw, messy data into a structured, clean format that can be easily analyzed.
- This process is also referred to as data munging.
- It involves a range of tasks like data cleaning, data transformation, and data enrichment to prepare the data for downstream use.

# Data Wrangling

- Data wrangling is different from data cleaning in that it is a broader process that encompasses additional steps beyond just identifying and fixing issues in the data.
  - While cleaning focuses on removing errors,
  - wrangling aims to reshape and enhance the data.
- Data wrangling transforms raw, unstructured data into clean, structured formats to enable accurate analysis and impactful insights.
- It involves identifying data issues, cleaning, reshaping, and enriching data through techniques like fixing missing values, removing duplicates, standardizing formats, and joining data.

# Data Wrangling Process



# Data Wrangling Process

## DATA WRANGLING PROCESS





# Data Wrangling Tools

- Python.
- [Alteryx APA](#)
- Altair Monarch
- Datameer
- Scrapy
- ParseHub
- Microsoft Power Query
- Tableau Desktop
- Tamr
- Astera, etc.



# Data enrichment, validation and publishing

The background is a solid teal color. It features several faint, semi-transparent data visualization elements. A large donut chart is positioned in the upper right quadrant. To its right and below are several smaller pie charts of varying sizes. In the bottom right corner, there is a bar chart with four vertical bars of increasing height from left to right.

Sushant Chalise

# Data Enrichment

- process of enhancing or improving the quality of existing data by adding additional information from external or internal sources
- to increase the value of the data and to provide deeper insights for analysis.
- Providing more context to the data, making it more informative

# Data Enrichment

- Customer Data: Adding geolocation or demographic information (age, income level) to customer profiles.
- E-commerce: Enriching product data with customer reviews, ratings, and social media sentiment.

# Data Enrichment

- Financial Data: Enriching transaction data with information about stock prices, market trends, or macroeconomic data.
- Social Media Data: Enriching user behavior data (posts, likes) with sentiment analysis or category tagging.

# Data Validation

- process of ensuring that data is accurate, complete, and of high quality before it is used for analysis, reporting, or decision-making.
- Ensuring data is consistent across different datasets or within the same dataset
- Ensuring that data adheres to predefined formats (e.g., correct date format)

# Data Validation

- Email Validation: Checking that email addresses follow the correct syntax
- Date Validation: Ensuring that date fields contain valid dates (e.g., not entering "30th February").
- Numerical Range Validation: Ensuring that values like age or salary fall within expected ranges (e.g.,  $\text{age} > 0$  and  $< 120$ ).

# Types of Data Validation:

- Format Validation: Ensuring data follows a specific format (e.g., phone number format or postal code).
- Range Validation: Ensuring values fall within an acceptable range (e.g., scores between 0-100).
- Consistency Validation: Ensuring data does not contradict itself (e.g., matching customer's age with their date of birth).



# Data Publishing

- process of making data available for sharing, use, or distribution within a specific ecosystem or with the general public.
- involves preparing the data for release, ensuring it is properly formatted, and making it accessible through appropriate channels (e.g., APIs, websites, or reports).

# Data Publishing: Characteristics

- Data Formatting: Publishing data in a format that is easy to use and integrate with other systems (e.g., CSV, JSON, XML).
- Documentation: Providing clear documentation to help users understand the data, how to interpret it, and how it can be used.

# Data Publishing: Characteristics

- Accessibility: Ensuring that the data is easily accessible by users who need it.
- Versioning and Updates: Ensuring that the data is regularly updated and versioned, especially in the case of dynamic datasets (e.g., market data, weather data).

# Data transformation and normalization

The background is a solid teal color. It features several decorative elements: a large, faint pie chart in the upper right quadrant; several smaller, faint pie charts scattered in the upper right and middle right areas; and a faint bar chart in the bottom right corner with four bars of increasing height.

Sushant Chalise

# Data Transformation

- process of converting raw data into a format that is more suitable for analysis or machine learning.
- Ensures that data is in a uniform format, making it easier to analyze and model.
- Helps improve the accuracy and performance of models by converting data into forms that are more suitable for the learning algorithms.

# Categorical Encoding

- Categorical variables (such as Gender or Country) need to be converted to numeric representations.
- One-Hot Encoding: Creates binary columns for each category (e.g., Male=1, Female=0).
- Label Encoding: Assigns an integer value to each category (e.g., Male=0, Female=1).

# Normalization

- process of adjusting the scale of features so that they all have the same range or distribution, especially for numerical data.
- ensures that no single feature dominates the analysis or modeling process due to its larger scale.

# Normalization : Importance

- Improved Performance: Some machine learning algorithms (like KNN, gradient descent-based methods) perform better when data features are on a similar scale.
- Large-scale features can dominate the model, leading to biased results. Normalization helps prevent this.



# Types of Normalization:

- Min-Max Scaling:
  - Rescales the data to fit within a specific range, typically [0, 1].

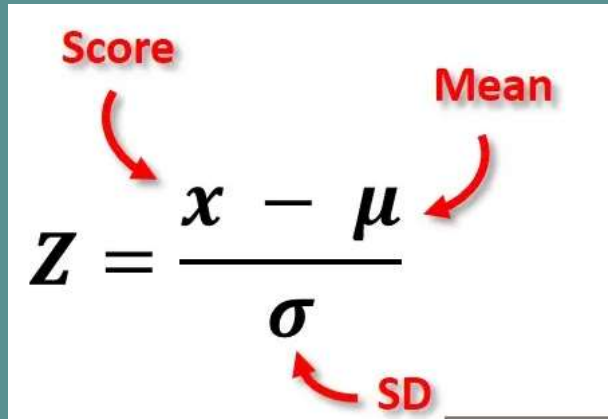
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

# Types of Normalization:

- Min-Max Scaling:
  - If a feature age ranges from 15 to 80, applying min-max scaling will transform these values into a  $[0, 1]$  range, where the minimum value (15) is scaled to 0, and the maximum value (80) is scaled to 1.

# Types of Normalization:

- Z-Score Normalization (Standardization):
  - Standardizes data by transforming it to have a mean of 0 and a standard deviation of 1.



The diagram shows the Z-score formula with red labels and arrows indicating the meaning of each variable:

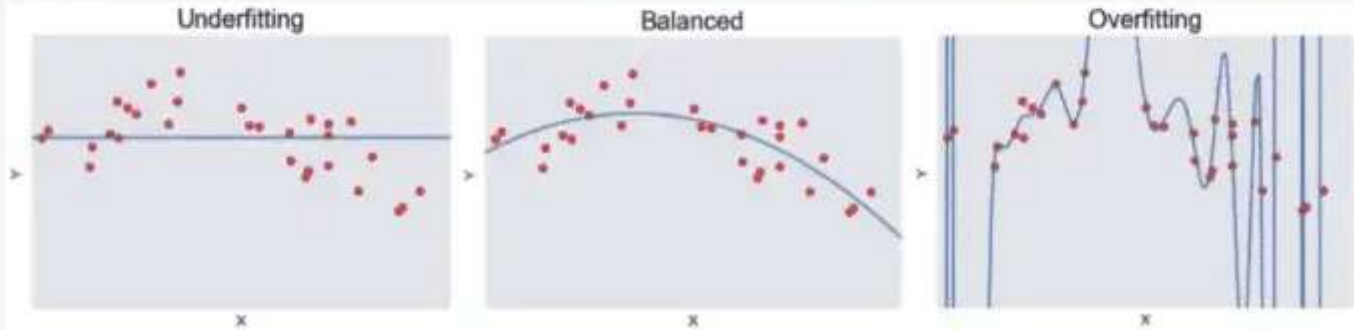
$$Z = \frac{x - \mu}{\sigma}$$

- Score** points to  $x$ .
- Mean** points to  $\mu$ .
- SD** (Standard Deviation) points to  $\sigma$ .

# **Dimensionality reduction linear factor model, principal component analysis (PCA)**

Sushant Chalise

## Underfitting VS Overfitting



jcchouinard.com

# Data Dimensionality

- From a theoretical point of view, increasing the number of features should lead to better performance.
- In practice, the inclusion of more features leads to worse performance (i.e., curse of dimensionality).
- The number of training examples required increases exponentially with dimensionality.

# Curse of Dimensionality

- Eg: Predicting House Pricing
- M1 - 2 Features - Score 1
- M2 - 5 Features - Score 2
- M3 - 10 Features - Score 3
- M4 - 50 Features - Score 4
- M5 - 100 Features - Score 5
- M6 - 1000 Features - Score 6

# Curse of Dimensionality

- Score 1 < Score 2 < Score 3 < Score 4 < Score 5 ?



# Curse of Dimensionality

- Score 1 < Score 2 < Score 3 < Score 4 < Score 5 ?
- Score 1 << Score 2 << Score 3 << Score 4 << Score 5 ?

# Curse of Dimensionality

- Score 1 < Score 2 < Score 3 < Score 4 < Score 5 ?
- Score 1 << Score 2 << Score 3 << Score 4 << Score 5 ?
- After certain threshold say 50 features

# Curse of Dimensionality

- Score 1 < Score 2 < Score 3 < Score 4 < Score 5 ?
- Score 1 << Score 2 << Score 3 << Score 4 << Score 5 ?
- After certain threshold say 50 features
- Score 2 > Score 4
- Score 2 > Score 5

# Why Reduce Dimensionality?

1. Reduces time complexity: Less computation
2. Reduces space complexity: Less parameters
3. Saves the cost of acquiring the feature
4. Simpler models are more robust
5. Easier to interpret; simpler explanation
6. Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions

# Dimensionality Reduction

Significant improvements can be achieved by first mapping (projecting) the data into a lower-dimensional space.


$$x = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} \longrightarrow \text{reduce dimensionality} \longrightarrow y = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} \quad (K \ll N)$$

Dimensionality can be reduced by:

- Combining features using a linear or non-linear transformations.
- Selecting a subset of features (i.e., feature selection).

# Dimensionality Reduction Technique



1. Chi Square
  2. Correlation Matrix
  3. PCA
  4. SVD
  5. Kernel PCA
  6. LDA
- 

# Principal Component Analysis

- Dimensionality reduction implies **information loss**; PCA preserves as much information as possible by **minimizing** the “reconstruction” error:

$$\|x - \hat{x}\|$$

$$x = a_1 v_1 + a_2 v_2 + \cdots + a_N v_N$$

$$\hat{x} = b_1 u_1 + b_2 u_2 + \cdots + b_K u_K$$

- How should we determine the “best” lower dimensional space (i.e., basis  $u_1, u_2, \dots, u_K$ )?

By the “best” eigenvectors of the **covariance** matrix of the data (i.e., corresponding to the “largest” eigenvalues – also called “**principal components**”)

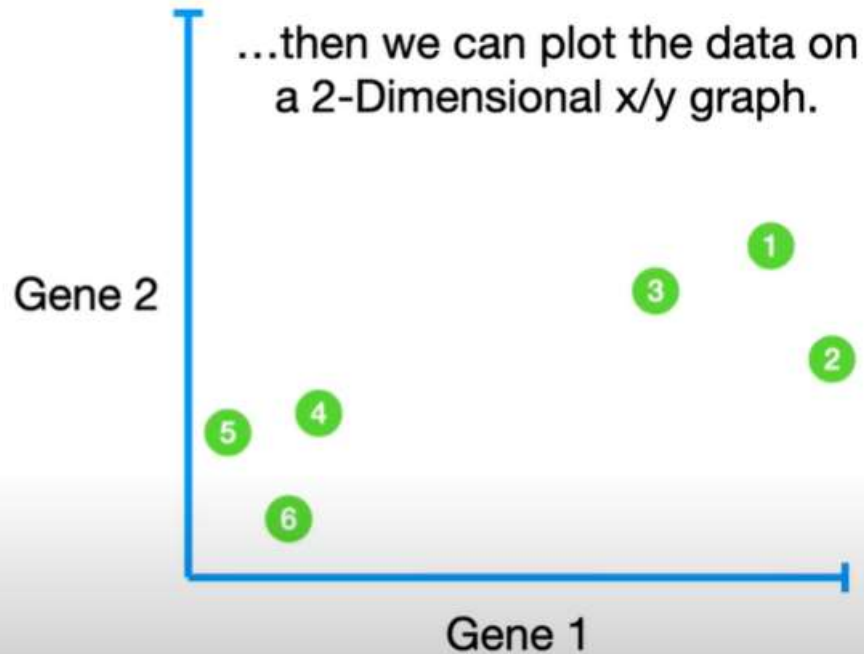
# Example

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



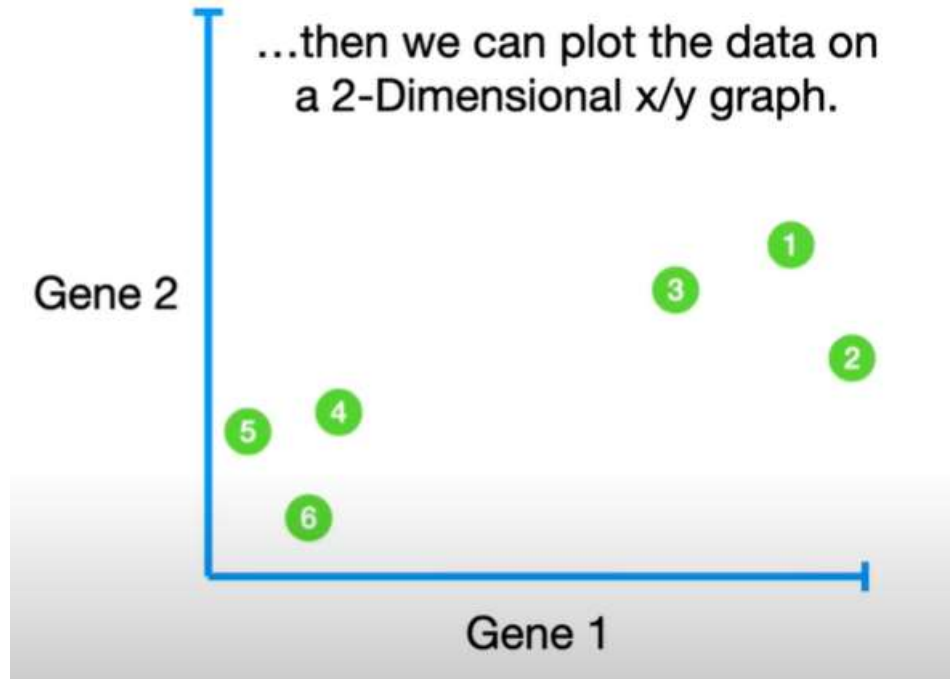
# Example

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



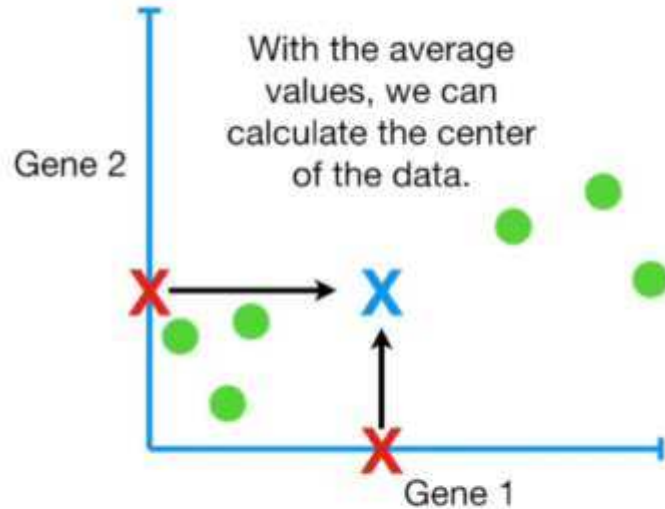
# Example

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1

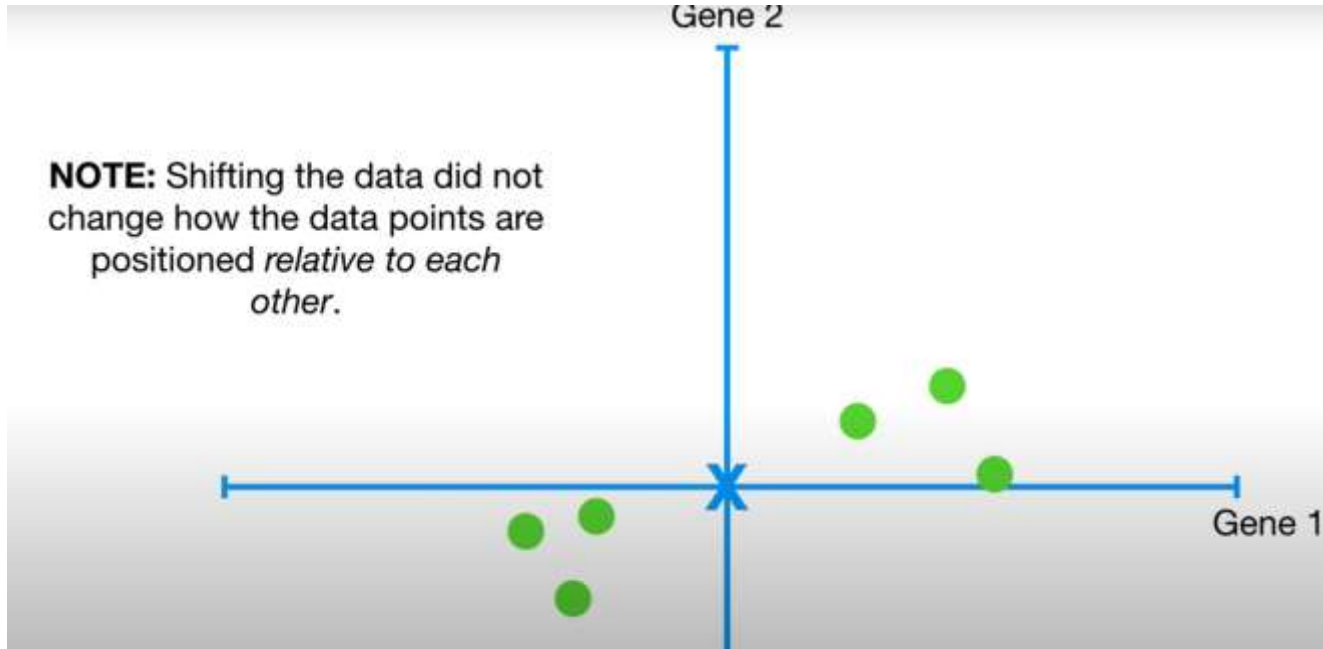


# Example

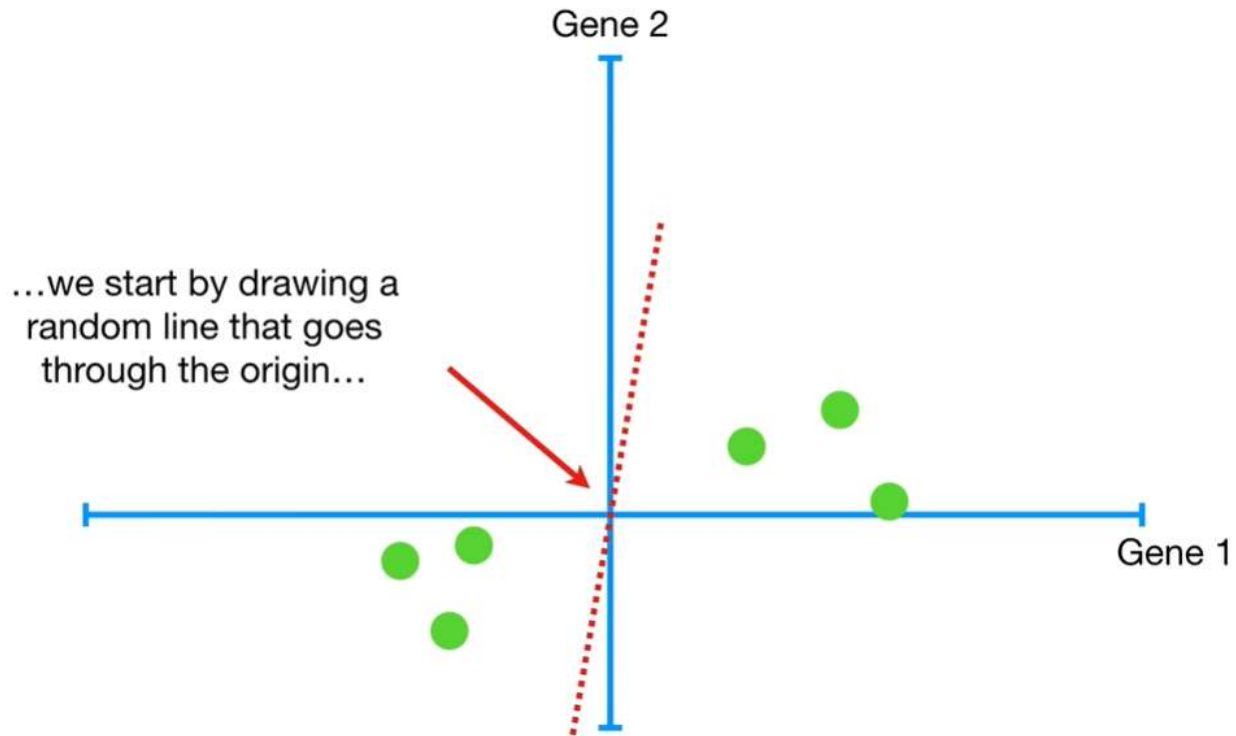
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



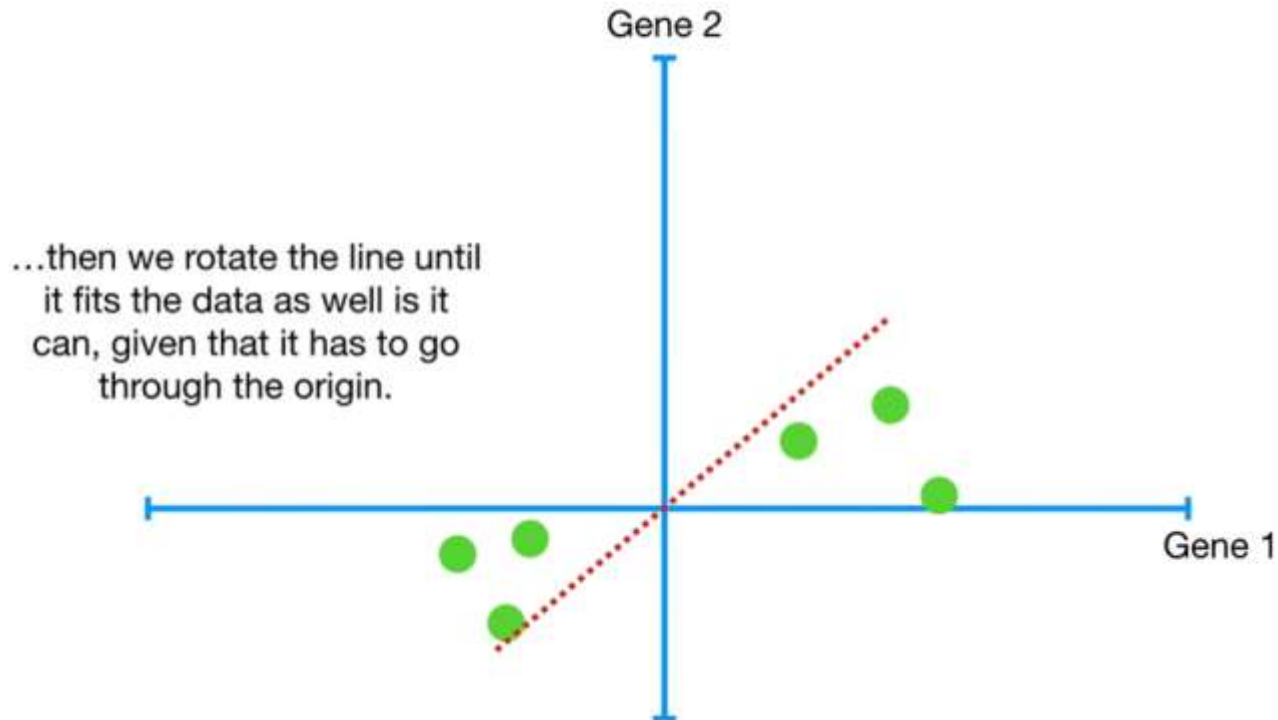
# Example



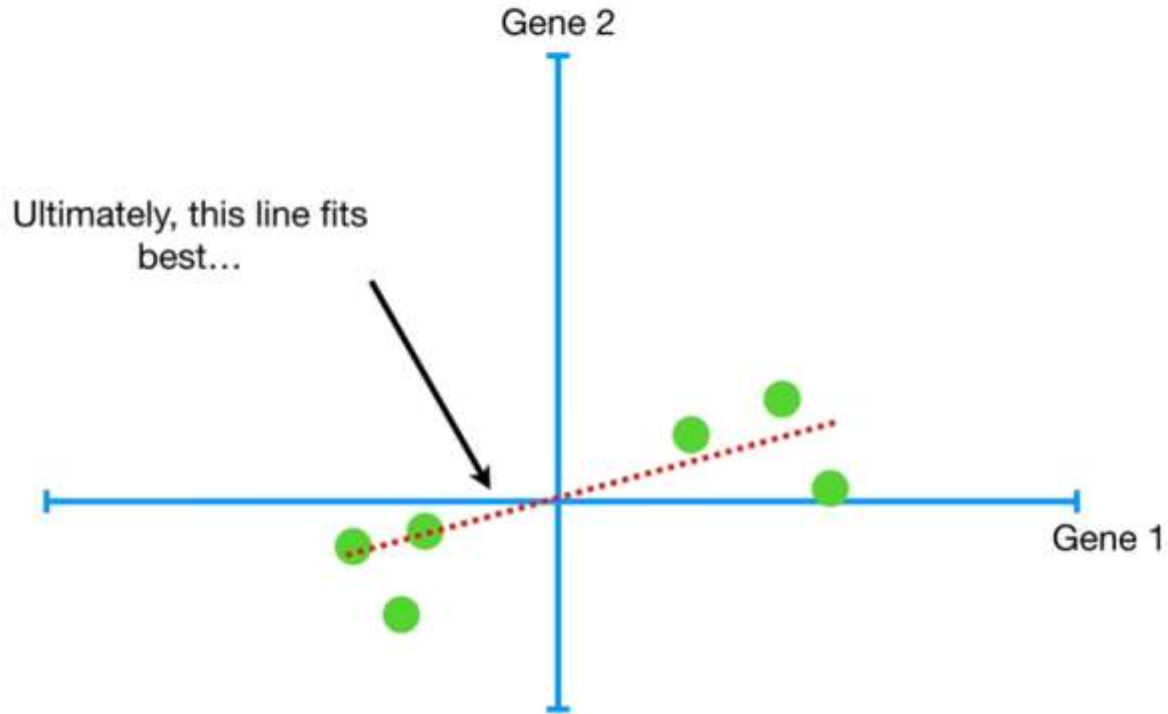
# Example



# Example

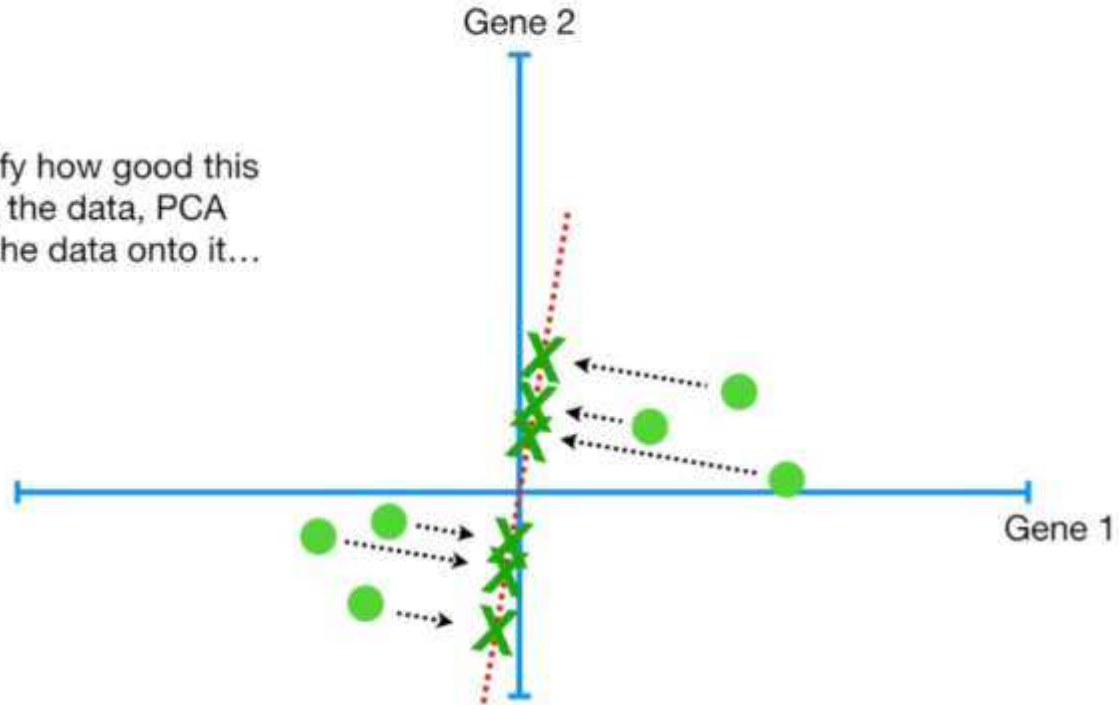


# Example



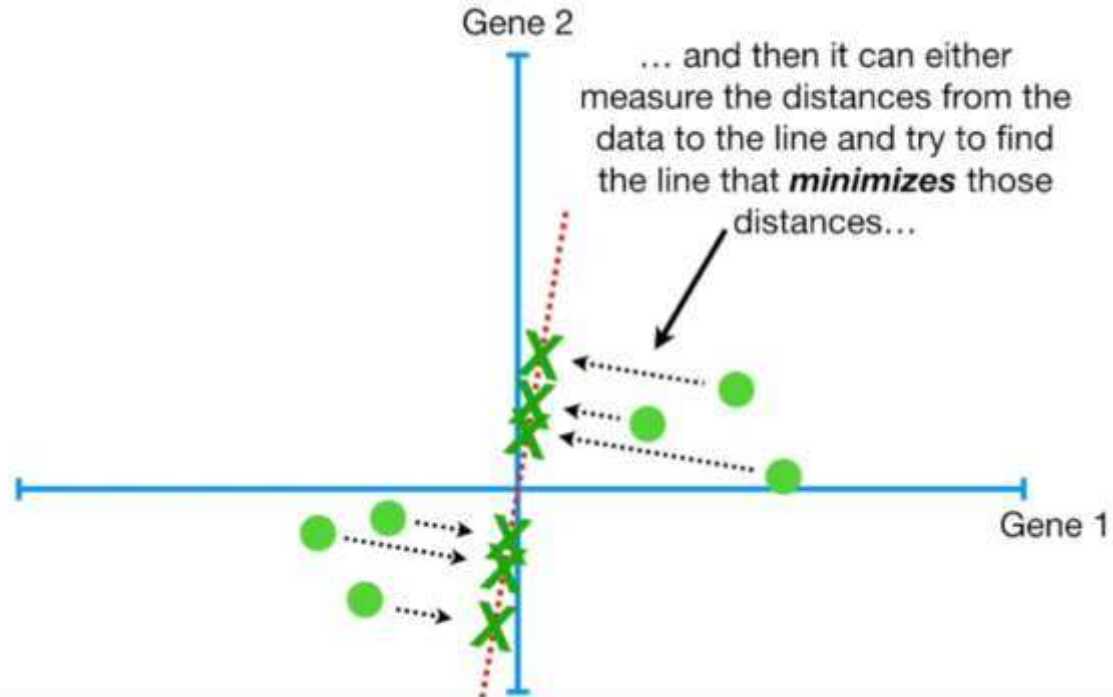
# Example

To quantify how good this line fits the data, PCA projects the data onto it...





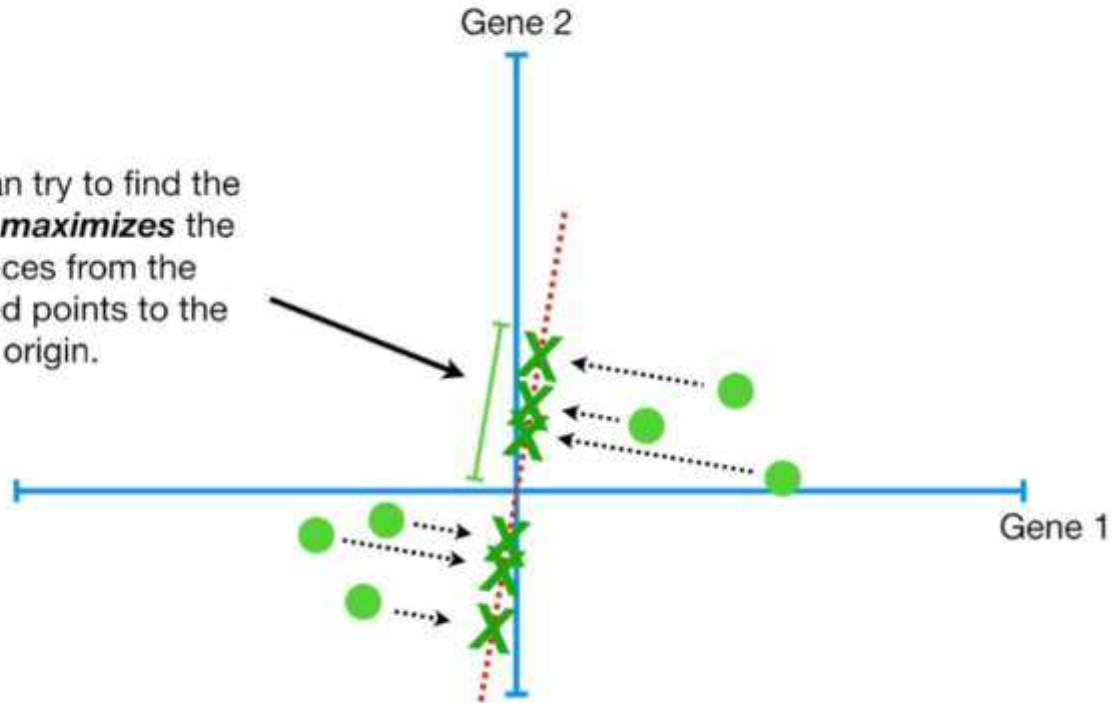
# Example



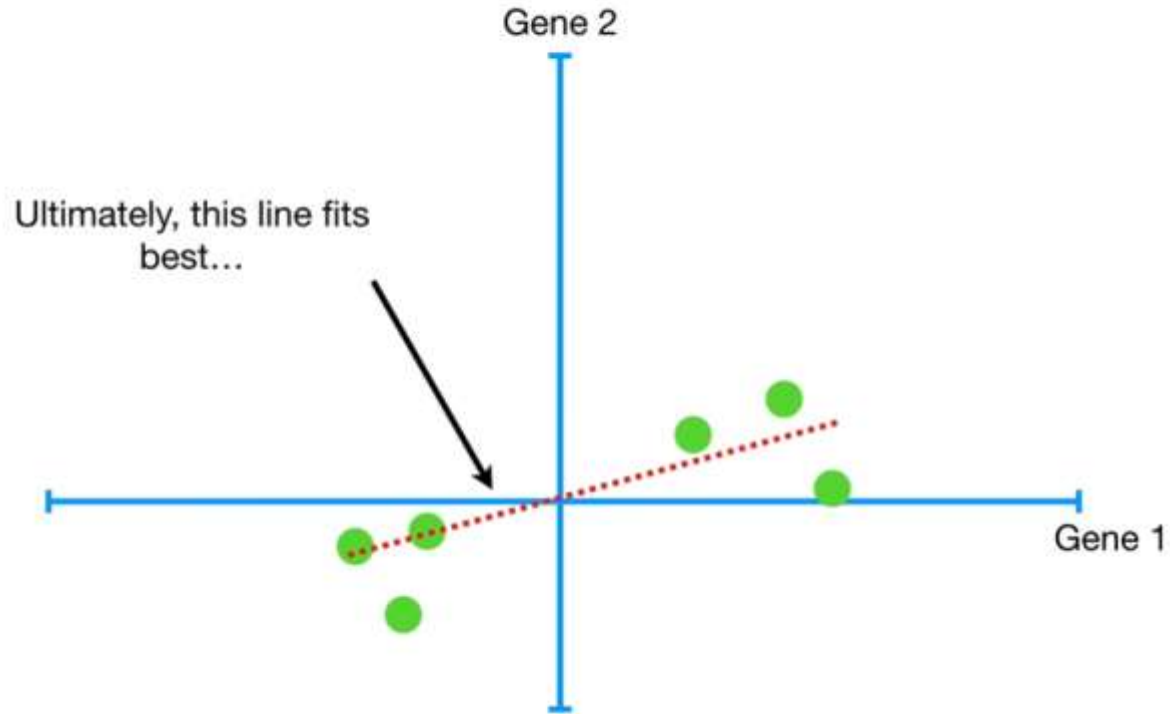
# Example

---

...or it can try to find the line that **maximizes** the distances from the projected points to the origin.

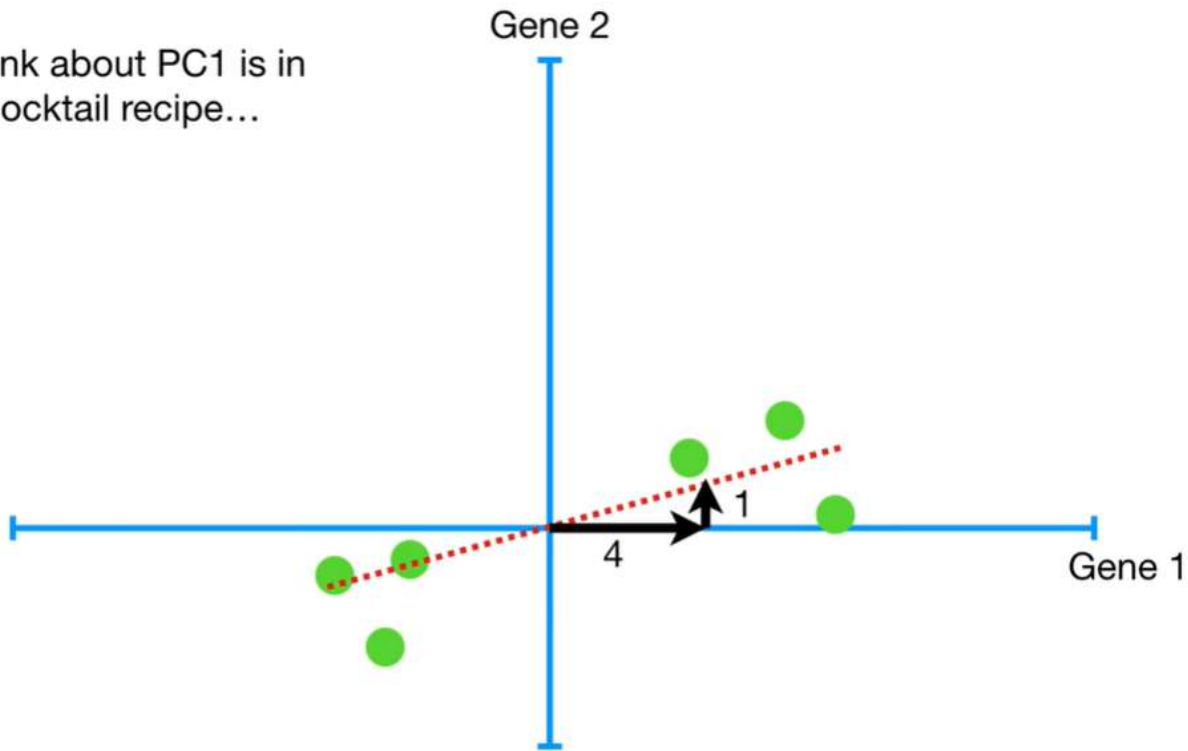


# Example



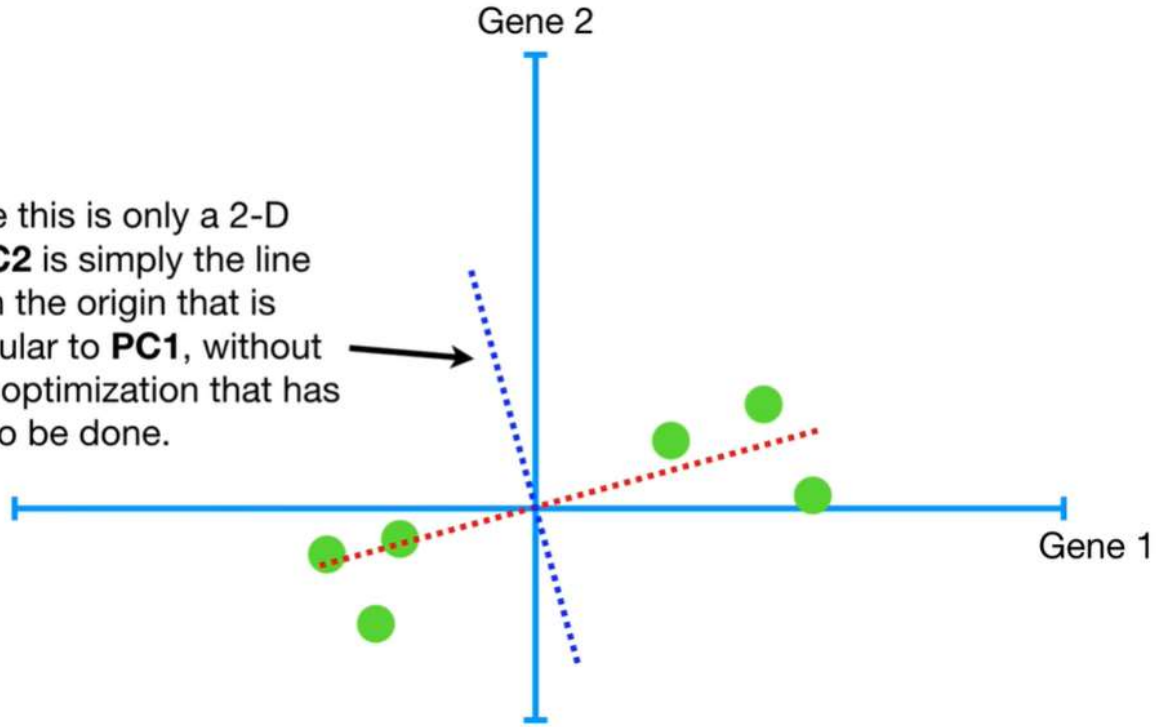
# Example

One way to think about PC1 is in terms of a cocktail recipe...

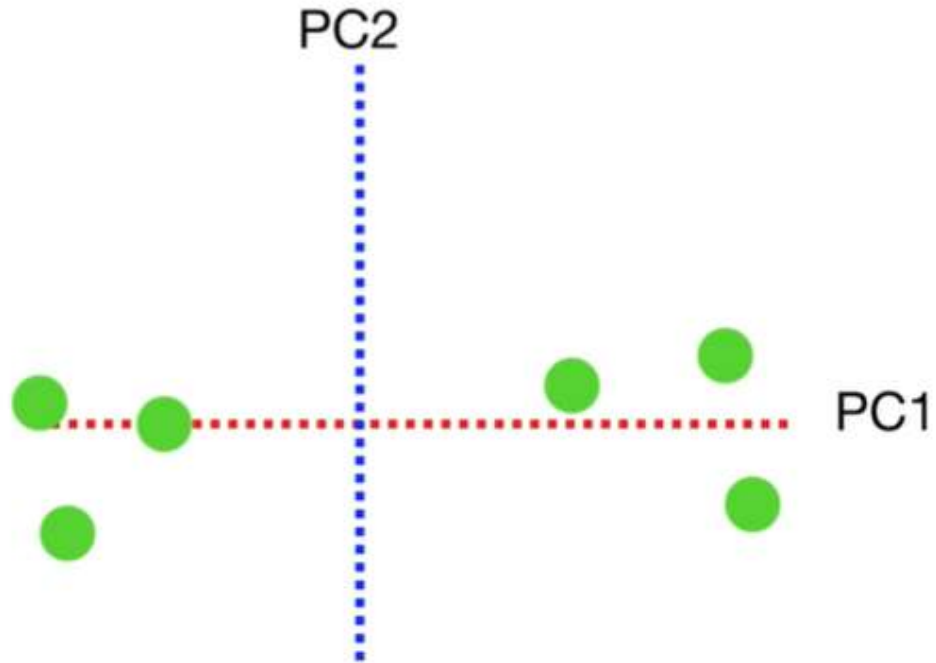


# Example

Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.

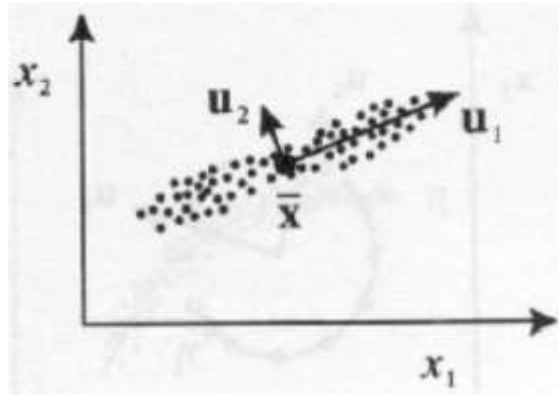


# Example

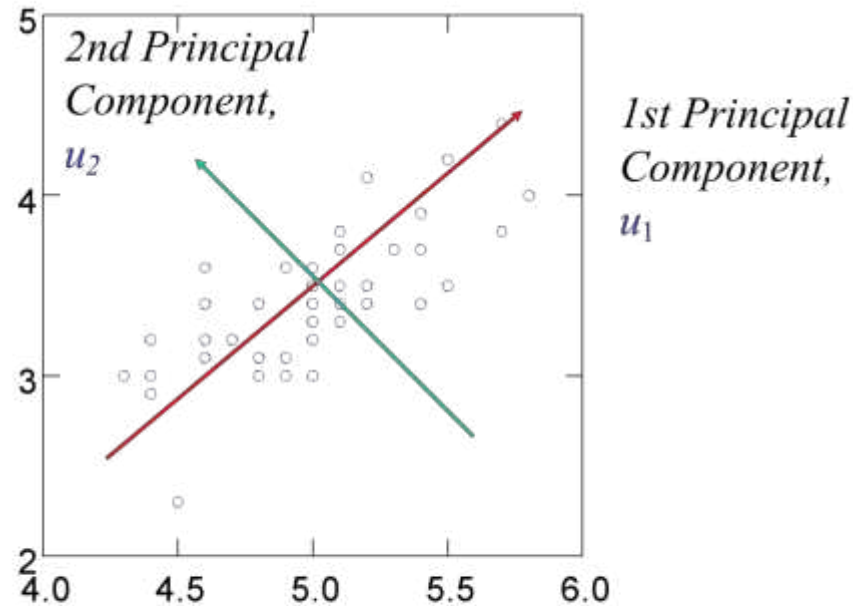


# PCA (Geometric interpretation)

- PCA projects the data along the directions where the data varies **most**.
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the **largest** eigenvalues.
- The magnitude of the eigenvalues corresponds to the **variance** of the data along the eigenvector directions.



# PCA (Geometric interpretation)

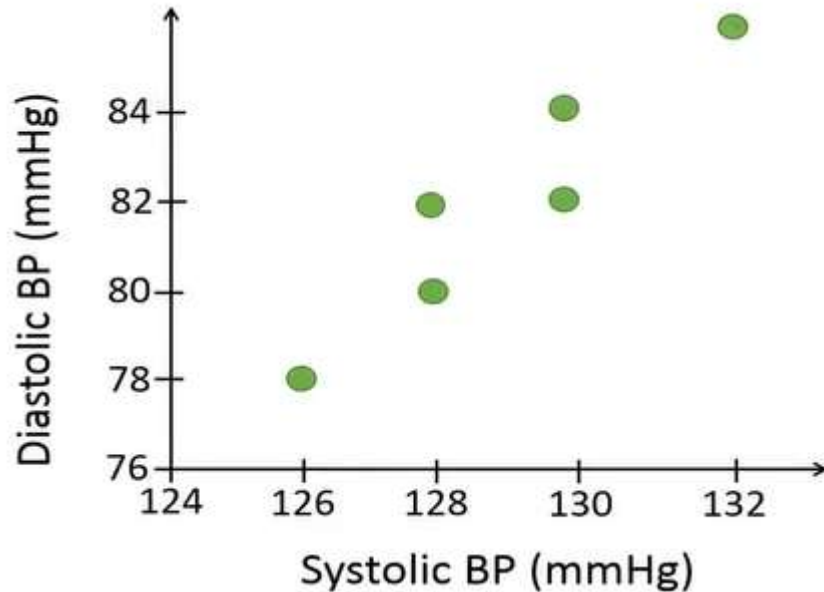


$u_1$  explains as much as possible of original variance in data set

$u_2$  explains as much as possible of remaining variance



# PCA - Example



Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86

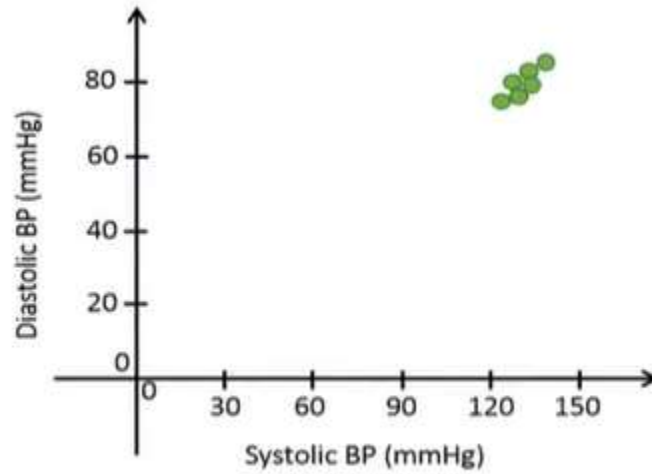


# PCA - Steps

1. Center the data
2. Calculate the covariance matrix
3. Calculate eigenvalues of the covariance matrix
4. Calculate eigenvectors of the covariance matrix
5. Order the eigenvectors
6. Calculate the principal components

# 1. Center the data

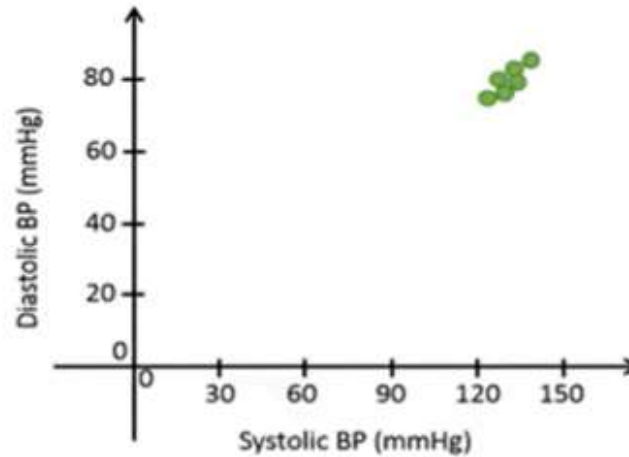
Systolic BP	Diastolic BP
126	78
128	80
128	82
130	82
130	84
132	86



# 1. Center the data

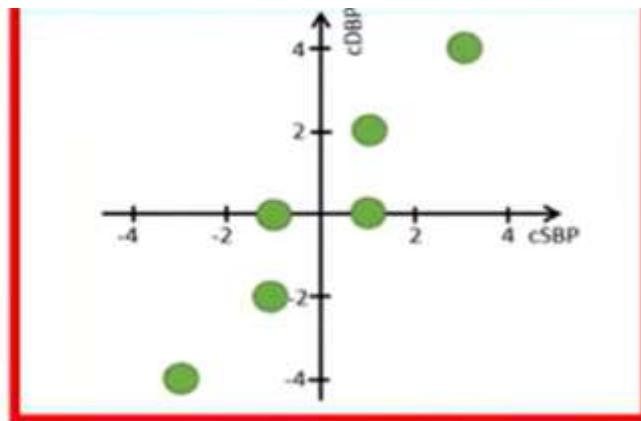
Systolic BP	Diastolic BP
$126 - 129 = -3$	$78 - 82 = -4$
$128 - 129 = -1$	$80 - 82 = -2$
$128 - 129 = -1$	$82 - 82 = 0$
$130 - 129 = 1$	$82 - 82 = 0$
$130 - 129 = 1$	$84 - 82 = 2$
$132 - 129 = 3$	$86 - 82 = 4$

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



# 1. Center the data

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



## 2. Calculate CoVariance Matrix

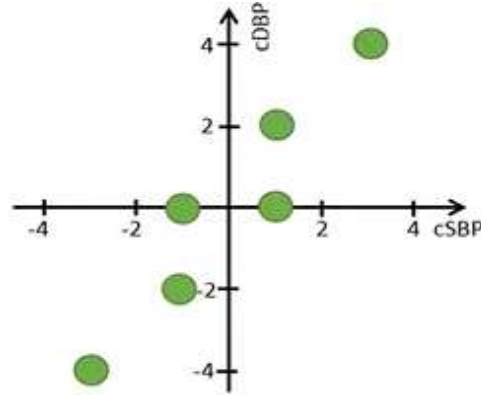
$$\text{var}(\text{cSBP}) = \frac{1}{n-1} \sum_{i=1}^n (\text{cSBP}_i - \overline{\text{cSBP}})^2 = ((-3)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2 + 3^2) / (6-1) = 22 / 5 = 4.4$$

$$\text{var}(\text{cDBP}) = \frac{1}{n-1} \sum_{i=1}^n (\text{cDBP}_i - \overline{\text{cDBP}})^2 = ((-4)^2 + (-2)^2 + 0^2 + 0^2 + 2^2 + 4^2) / (6-1) = 40 / 5 = 8$$

$$\text{cov}(\text{cSBP}, \text{cDBP}) = \frac{1}{n-1} \sum_{i=1}^n (\text{cSBP}_i - \overline{\text{cSBP}}) \cdot (\text{cDBP}_i - \overline{\text{cDBP}}) = \boxed{(-3) \cdot (-4)} + (-1) \cdot (-2) + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot 2 + 3 \cdot 4 / (6-1) = 28 / 5 = 5.6$$

## 2. Calculate CoVariance Matrix

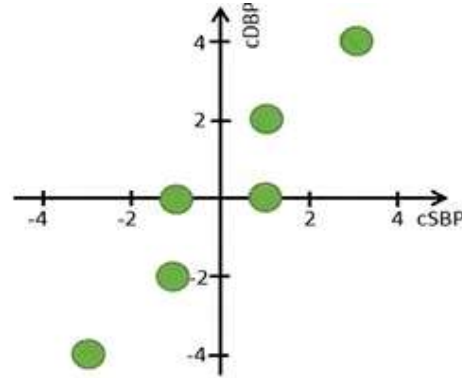
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

### 3. Calculate Eigenvalues of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A - \lambda I| = 0$$

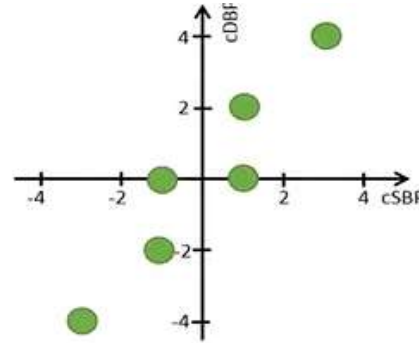
	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\det \begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$



### 3. Calculate Eigenvalues of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$\det|A - \lambda I| = 0$$

$$(4.4 - \lambda)(8.0 - \lambda) - 5.6 \cdot 5.6 = 0$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

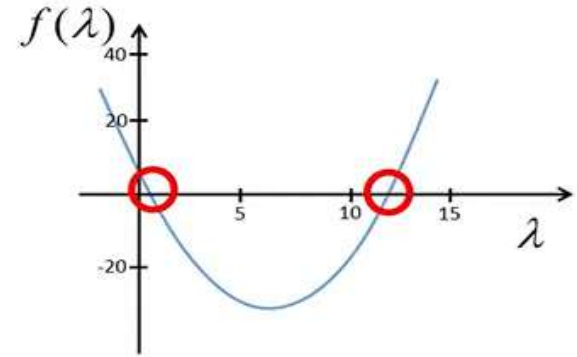
$$\det \begin{bmatrix} (4.4 - \lambda) & 5.6 \\ 5.6 & (8.0 - \lambda) \end{bmatrix} = 0$$

$$3.84 - 12.4\lambda + \lambda^2 = 0$$

### 3. Calculate Eigenvalues of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

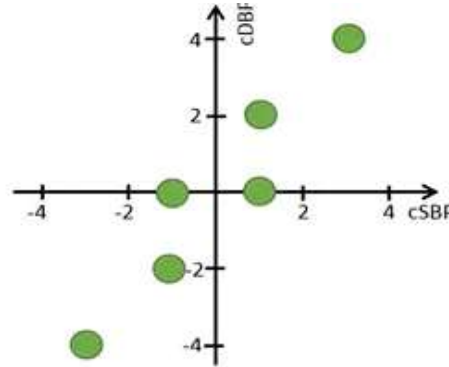


$$3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\lambda_1 = 0.32 \quad \lambda_2 = 12.08$$

## 4. Calculate Eigenvectors of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



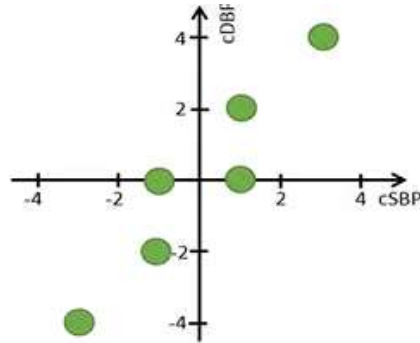
$$A \cdot v = \lambda \cdot v$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \boxed{12.08} \begin{bmatrix} x \\ y \end{bmatrix}$$

## 4. Calculate Eigenvectors of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$A \cdot v = \lambda \cdot v$$

$$4.4x + 5.6y = 12.08x$$

$$5.6x + 8.0y = 12.08y$$

$$5.6y = 7.68x$$

$$5.6x = 4.08y$$

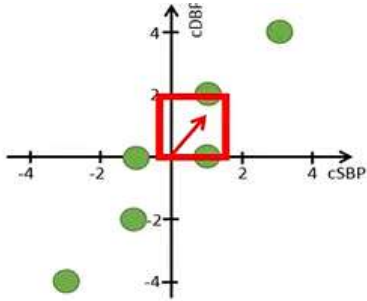
$$y = 1.37x$$

$$1.37x = y$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 12.08 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

## 4. Calculate Eigenvectors of Cov Matrix



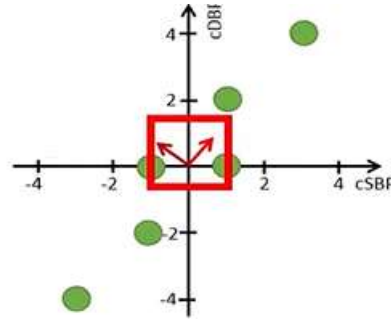
$$v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$$

And Normalize

$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

## 4. Calculate Eigenvectors of Cov Matrix

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4

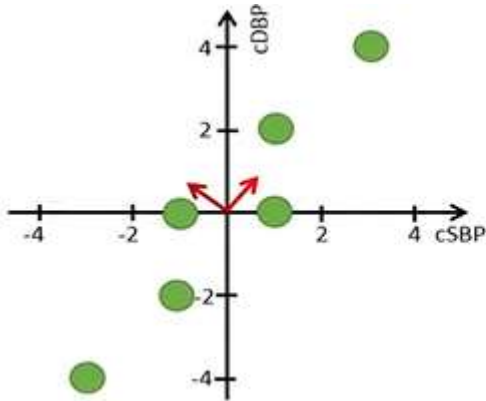


$$v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_2 = 12.08$$

$$v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_1 = 0.32$$

	SBP	DBP
SBP	4.4	5.6
DBP	5.6	8.0

## 5. Order the EigenVectors



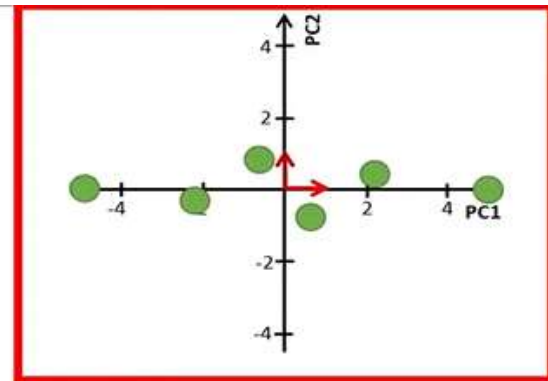
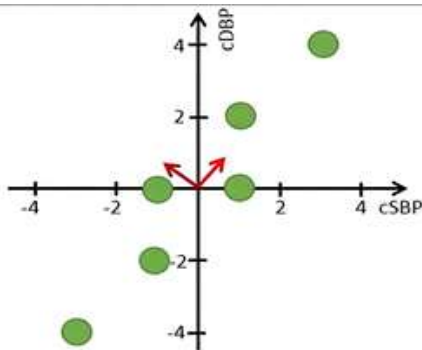
$$v_1 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix} \quad \lambda_1 = 12.08$$

$$v_2 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix} \quad \lambda_2 = 0.32$$

$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

# 6. Find the Principal Component

Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4



$$V = \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix}$$

$$D = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

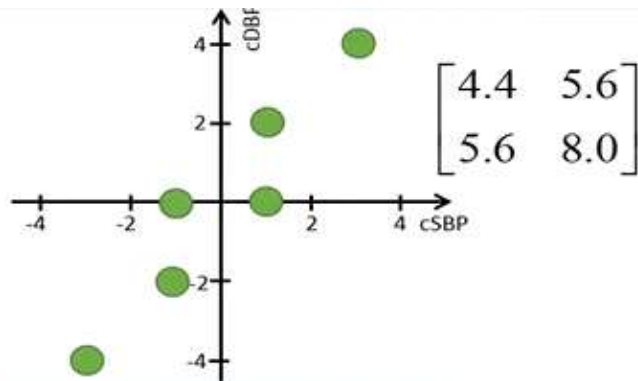
$$DV = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$\cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} =$$

$$\begin{matrix} \text{PC1} & \text{PC2} \\ \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix} \end{matrix}$$

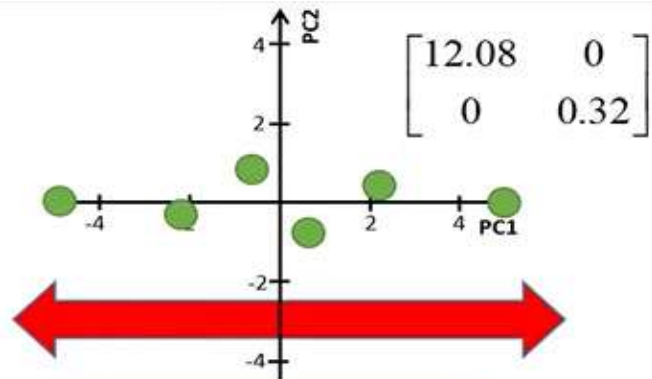


# Interpret the PCA



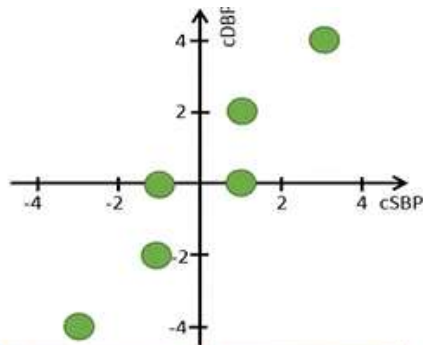
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\% \text{ var} = \frac{12.08}{12.08 + 0.32} = 97.4\%$$



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

# Interpret the PCA



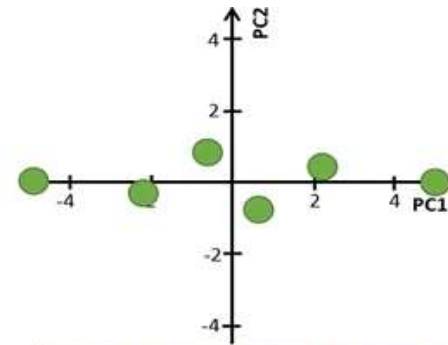
Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

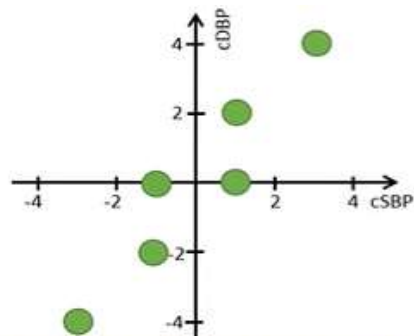
$$PC2 = -0.81 \cdot cSBP + 0.59 \cdot cDBP$$

$$PC1_6 = 0.59 \cdot 3 + 0.81 \cdot 4 = 5$$



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

# Interpret the PCA

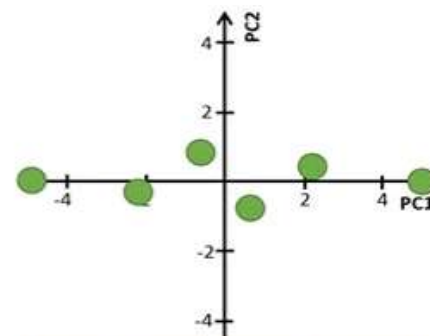


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$\begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0.59 & -0.81 \\ 0.81 & 0.59 \end{bmatrix} = \begin{bmatrix} -5.0 & 0.1 \\ -2.2 & -0.4 \\ -0.6 & 0.8 \\ 0.6 & -0.8 \\ 2.2 & 0.4 \\ 5.0 & -0.1 \end{bmatrix}$$

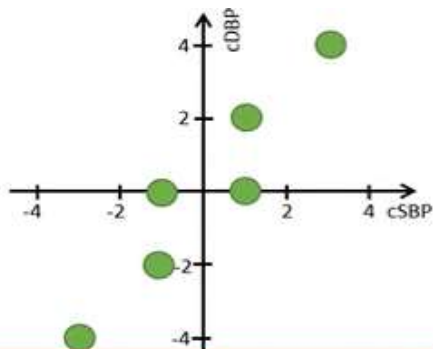
$$PC1 = 0.59 \cdot (SBP - \overline{SBP}) + 0.81 \cdot (DBP - \overline{DBP})$$

$$PC1 = 0.59 \cdot [132 - 129] + 0.81 \cdot [86 - 82] = 5.0$$

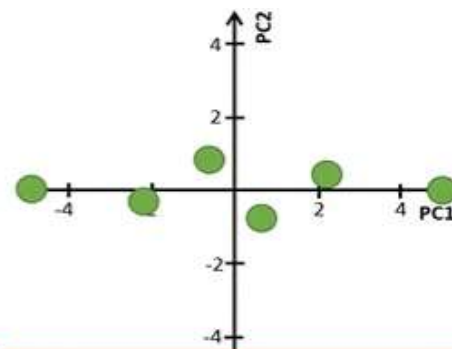


PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

# Interpret the PCA

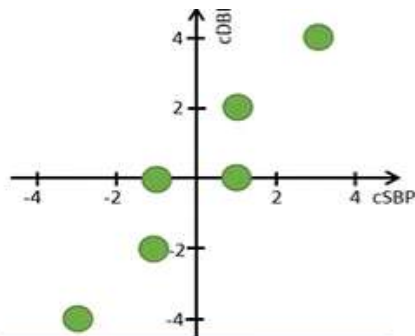


Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0



PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	0.8
0.6	-0.8
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

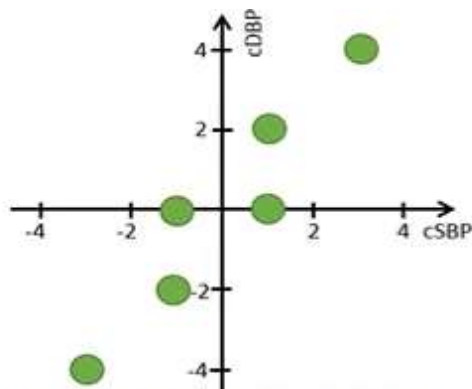
# Interpret the PCA



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

PC1	PC2
-5.0	0.1
-2.2	-0.4
-0.6	1.5
0.6	-0.3
2.2	0.4
5.0	-0.1
Var=12.08	Var=0.32

# Interpret the PCA



Centered SBP	Centered DBP
-3	-4
-1	-2
-1	0
1	0
1	2
3	4
Var=4.4	Var=8.0

$$PC1 = 0.59 \cdot cSBP + 0.81 \cdot cDBP$$

PC1
-5.0
-2.2
-0.6
0.6
2.2
5.0
Var=12.08

# Usages of PCA

PCA is mostly used as a tool for **Compression** and **Simplifying** data for **easier learning** in exploratory data analysis and for making predictive models.

1- Better Perspective and less Complexity

2 - Better visualization

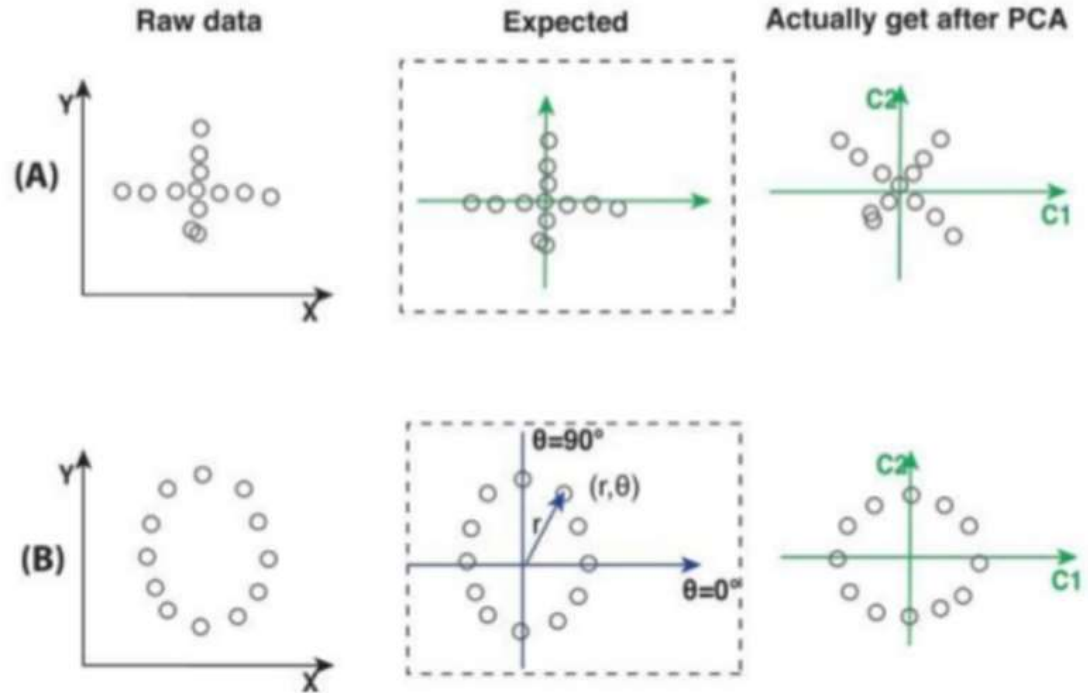
3- Reduce size

4- Different perspective:



# Limitation of PCA

If the data does not follow a multidimensional normal (Gaussian) distribution, PCA may not give the best principal components



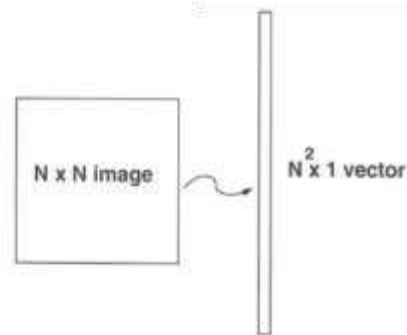


# Application

## Computation of low-dimensional basis (i.e., eigenfaces):

Step 1: obtain face images  $I_1, I_2, \dots, I_M$  (training faces)

(very important: the face images must be centered and of the same size)



Step 2: represent every image  $I_i$  as a vector  $\Gamma_i$

# Example

Normalized face images

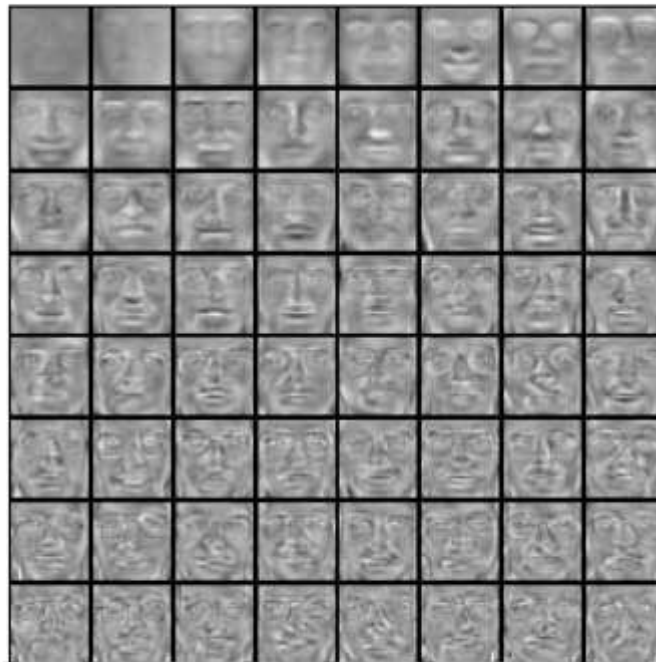


# Example

Mean:  $\mu$



Top eigenvectors:  $u_1, \dots, u_k$



# Example

- Representing faces onto this basis

- Each face (minus the mean)  $\Phi_i$  in the training set can be represented as a linear combination of the best  $K$  eigenvectors:

$$\hat{\Phi}_i - \text{mean} = \sum_{j=1}^K w_j u_j, \quad (w_j = u_j^T \Phi_i)$$

(where  $\|u_j\| = 1$ )

(we call the  $u_j$ 's *eigenfaces*)



Face reconstruction:

# Thank You

Sushant Chalise