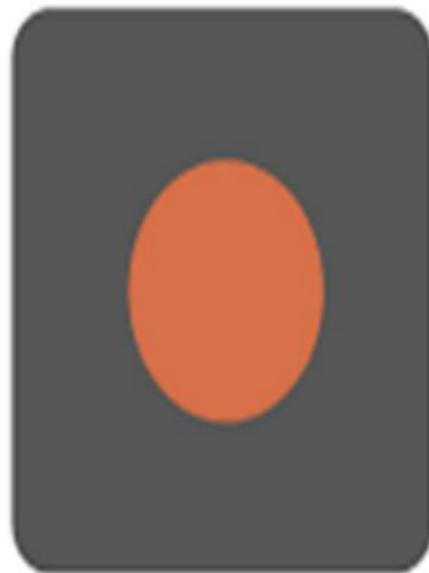


IMP Note to Self



**Start
Recording**



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE

MODULE # 1 : INTRODUCTION

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

- 1 COURSE LOGISTICS
- 2 FUNDAMENTALS OF DATA SCIENCE
- 3 DATA SCIENCE REAL WORLD APPLICATIONS
- 4 DATA SCIENCE CHALLENGES
- 5 DATA SCIENCE TEAMS
- 6 SOFTWARE ENGINEERING FOR DATA SCIENCE
- 7 FURTHER READING

COURSE OBJECTIVES

- CO1** Gain basic understanding of the role of Data Science in various scenarios in the real-world of business, industry and government.
- CO2** Understand various roles and stages in a Data Science Project and ethical issues to be considered.
- CO3** Explore the processes, tools and technologies for collection and analysis of structured and unstructured data.
- CO4** Appreciate the importance of techniques like data visualization, storytelling with data for the effective presentations of the outcomes with the stakeholders.
- CO5** Understand techniques of preparing real-world data for data analytics.
- CO6** Implement data analytic techniques for discovering interesting patterns from data.

COURSE STRUCTURE

M1 Introduction to Data Science

M2 Data Analytics

M3 Data and Data Models

M4 Data Wrangling

M5 Feature Engineering

M6 Classification and Prediction

M7 Association Analysis

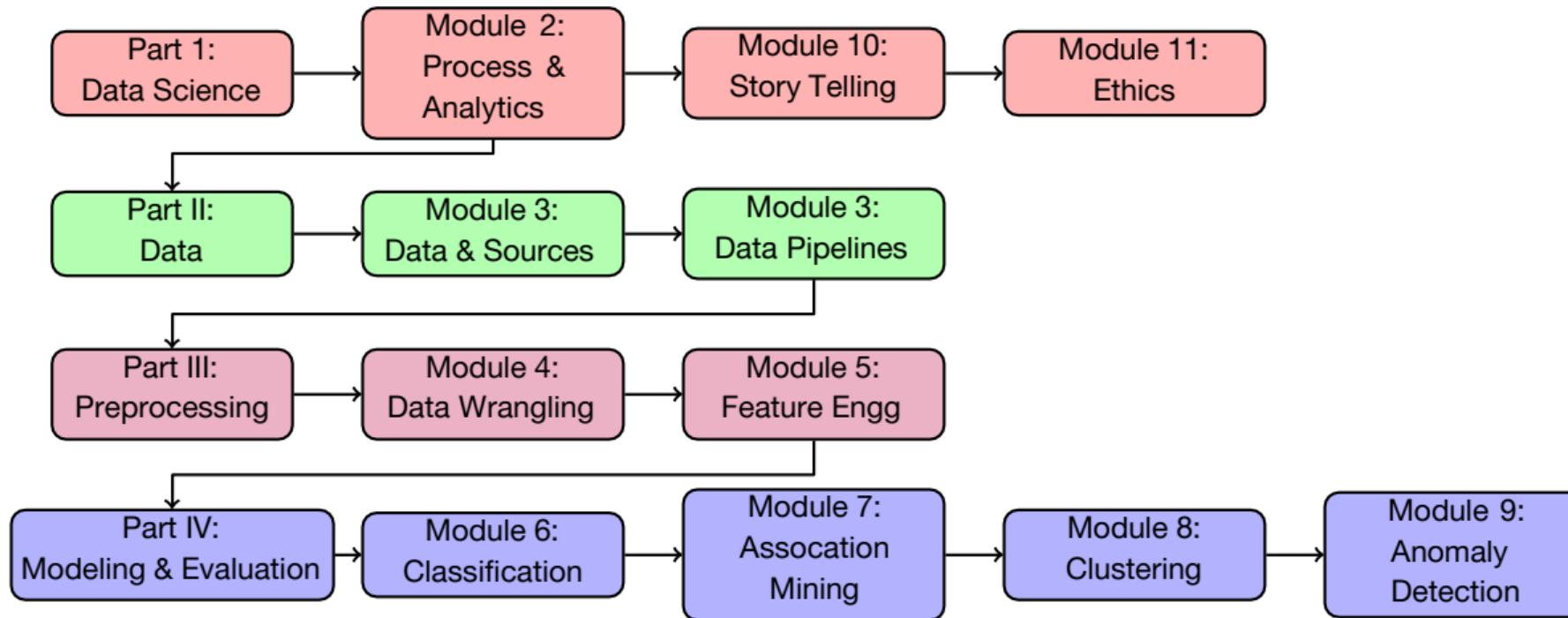
M8 Clustering

M9 Anomaly Detection

M10 Storytelling with Data

M11 Ethics for Data Science

MODULES OVERVIEW



TEXT BOOKS

- | | |
|-------------|---|
| T1
Kumar | Introduction to Data Mining, by Tan, Steinbach and Vipin |
| T2 | Introducing Data Science by Cielen, Meysman and Ali |
| T3 | Storytelling with Data, A data visualization guide for business professionals, by Cole, Nussbaumer Knaflic; Wiley |
| T4 | Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 |

REFERENCE BOOKS

- R1 The Art of Data Science by Roger D Peng and Elizabeth Matsui
- R2 Ethics and Data Science by DJ Patil, Hilary Mason, Mike Loukides
- R3 Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas
- R4 KDD, SEMMA and CRISP-DM: A Parallel Overview , Ana Azevedo and M.F. Santos, IADS-DM, 2008

EVALUATION SCHEDULE

No	Name	Type	Duration	Weight	Remarks
EC1	Quiz I	Online	1 hr	5%	Average of both quizzes
	Quiz II	Online	1 hr	5%	
	Assignment Part I	Online	4 weeks	10%	Sum of both Assignments
	Assignment Part II	Online	4 weeks	15%	
	Mid-sem	Online	As announced	30%	
EC3	Compre-sem Regular	Online	As announced	40%	

LEARNING PLATFORM

Most relevant and up to date info on Canvas

- Handout
- Schedule for Webinar, Quiz, and Assignments.
- Session Slide Deck
- Demo Lab Sheets
- Quiz-I, Quiz-II
- Assignment I, Assignment II

The video recording will be available in Lecture delivery platform.

PLATFORM / DATASET

- Platform
 -) Python / Jupyter Notebook / Google Colab
- Dataset
 -) Datasets as we deem appropriate.
- Webinar
 -) 4 webinars
 -) Either Lab modules will be explained or numerical problems will be solved.
 -) As per schedule

TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

WHY DATA SCIENCE?

- "Data Science is the sexiest job in the 21st century" – IBM.
- Data Science is one of the fastest growing fields in the world.
- According to the U.S. Bureau of Labor Statistics, 11.5 million new jobs will be created by the year 2026.
- Even with COVID-19 situation, and the amount of shortage in talent, there might not be a dip in data science as a career option.

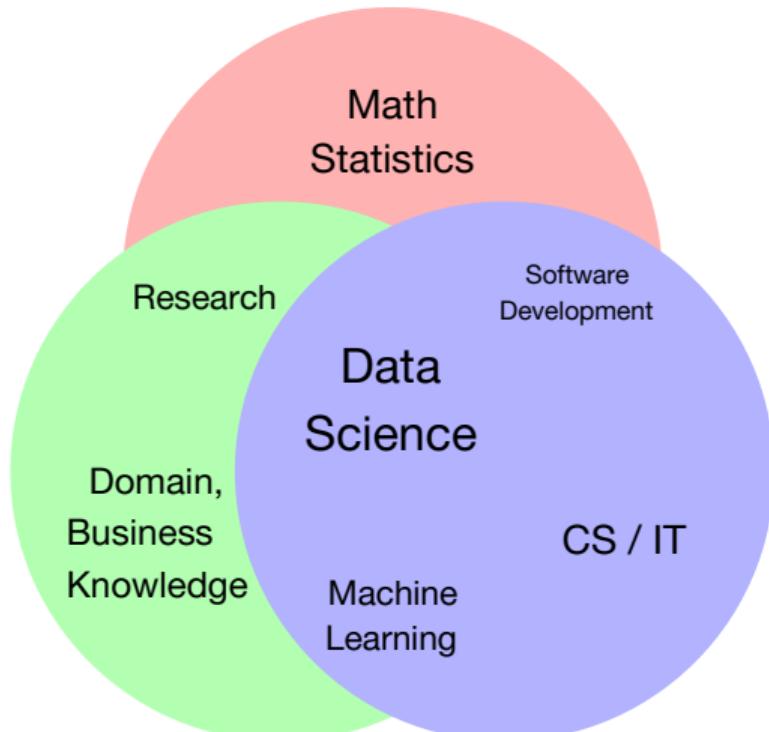
WHY DATA SCIENCE?

- In India, the average salary of a data scientist as of January 2020 is Rs.10L/yr. – Glassdoor, 2020.
- The increase in data science as a career choice in 2020 will also see the rise in its various job roles.
 - Data Engineer
 - Data Administrator
 - Machine Learning Engineer
 - Statistician
 - Data and Analytics Manager

DATA SCIENCE

- Data Science is a study of data.
- Data Science is an art of uncovering insights and trends that are hiding behind the data.
- Data Science helps to translate data into a story. The story telling helps in uncovering insights. The insights help in making decision or strategic choices.
- Data Science is the process of using data to understand different things.
 - Requires a major effort of preparing, cleaning, scrubbing, or standardizing the data.
 - Algorithms are then applied to crunch pre-processed data.
 - This process is iterative and requires analysts' awareness of the best practices.
 - The most important aspect of data science is interpreting the results of the analysis in order to make decisions.

DATA SCIENCE – MULTIPLE DISCIPLINES



NEED OF DATA SCIENCE

- Data deluge, tons of data.
- Powerful algorithms.
- Open software and tools.
- Computational speed, accuracy and cost.
- Data storage in terms of capacity and cost.

DATA SCIENCE, AI AND ML

■ Artificial Intelligence

- › AI involves making machines capable of mimicking human behavior, particularly cognitive functions like facial recognition, automated driving, sorting mail based on postal code.

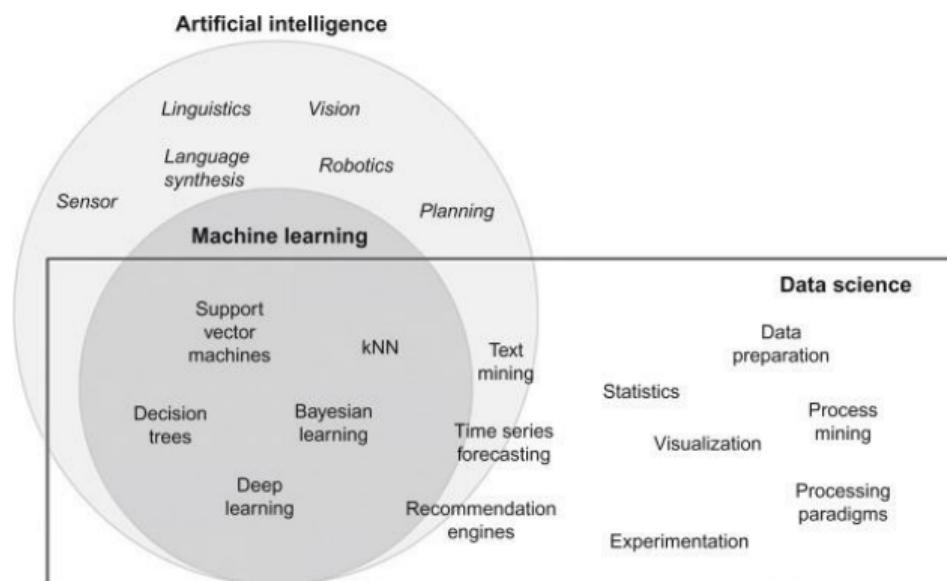
■ Machine Learning

- › Considered a sub-field of or one of the tools of AI.
- › Involves providing machines with the capability of learning from experience.
- › Experience for machines comes in the form of data.

■ Data Science

- › Data science is the application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics to uncover insights from data to enable better decision marking.

DATA SCIENCE, AI AND ML



<https://www.sciencedirect.com/topics/physics-and-astronomy/artificial->

TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

USE CASES OF DATA SCIENCE



DATA SCIENCE IN FACEBOOK

Social Analytics

- Utilizes quantitative research to gain insights about the social interactions of among people.
- Makes use of deep learning, facial recognition, and text analysis.
- In facial recognition, it uses powerful neural networks to classify faces in the photographs.
- In text analysis, it uses “DeepText” to understand people’s interest and aligns photographs with texts.
- It uses deep learning for targeted advertising.
- Using the insights gained from data, it clusters users based on their preferences and provides them with the advertisements that appeal to them.

DATA SCIENCE IN AMAZON

Improving E-Commerce Experience

- Personalized recommendation
 - Predictive analytics (a personalized recommender system) to increase customer satisfaction.
 - Purchase history of customers, other customer suggestions, and user ratings are analyzed to recommend products.
- Anticipatory shipping model
 - Predict the products that are most likely to be purchased by its users.
 - Analyzes pattern of customer purchases and keeps products in the nearest warehouse which the customers may utilize in the future.

DATA SCIENCE IN AMAZON – CONTD...

Improving E-Commerce Experience

- Price discounts
 - Using parameters such as the user activity, order history, prices offered by the competitors, product availability, etc., Amazon provides discounts on popular items and earns profits on less popular items.
- Fraud Detection
 - Detect fraud sellers and fraudulent purchases.
- Improving Packaging Efficiency
 - Optimize packaging of products in warehouses and increases efficiency of packaging lines through the data collected from the workers.

DATA SCIENCE IN UBER

Improving Rider Experience

- Uber maintains large database of drivers, customers, and several other records.
- Makes extensive use of Big Data and crowdsourcing to derive insights and provide best services to its customers.
- Dynamic pricing
 - Use of big Data and data science to calculate fares based on specific parameters.
 - Uber matches customer profile with the most suitable driver and charges them based on the time it takes to cover the distance rather than the distance itself.
 - The time of travel is calculated using algorithms that make use of data related to traffic density and weather conditions.
 - When the demand is higher (more riders) than supply (less drivers), the price of the ride goes up.

DATA SCIENCE IN BANK OF AMERICA

Improving Customer Experience

- Erica – a virtual financial assistant (BoA)
 - Erica serves as a customer advisor to over 45 million users around the world.
 - Erica makes use of Speech Recognition to take customer inputs.
- Fraud detection
 - Uses data science and predictive analytics to detect frauds in payments, insurance, credit cards, and customer information.
- Risk modeling
 - Use data science for risk modeling to regulate financial activities.
- Customer segmentation
 - Segment their customers in the high-value and low-value segments.
 - Data scientists make use of clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

DATA SCIENCE IN AIRBNB

Improving Customer Experience

- Providing better search results
 - Uses big data of customer and host information, homestays and lodge records, and website traffic.
 - Uses data science to provide better search results to its customers and find compatible hosts.
- Detecting bounce rates
 - Use of demographic analytics to analyze bounce rates from their websites.
- Providing ideal lodgings and localities
 - Uses knowledge graphs where the user's preferences are matched with the various parameters to provide ideal lodgings and localities.

DATA SCIENCE IN SPOTIFY

Improving Customer Experience and recommendation

- Providing better music streaming experience
 - Provide personalized music recommendations.
 - Uses over 600 GBs of daily data generated by the users to build its algorithms to boost user experience.
- Improving experience for artists and managers
 - Spotify for Artists application allows the artists and managers to analyze their streams, fan approval and the hits they are generating through Spotify's playlists.

DATA SCIENCE IN SPOTIFY... CONTD..

- Spotify uses data science to gain insights about which universities had the highest percentage of party playlists and which ones spent the most time on it.
- "Spotify Insights" publishes information about the ongoing trends in the music.
- Spotify's Niland, an API based product, uses machine learning to provide better searches and recommendations to its users.
- Spotify analyzes listening habits of its users to predict the Grammy Award Winners.

APPLICATIONS OF DATA SCIENCE



APPLICATIONS OF DATA SCIENCE



TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

DATA SCIENCE CHALLENGES

Data science challenges can be categorized as:

- Data related
- Organization related
- Technology related
- People related
- Skill related

CHALLENGES IN DATA SCIENCE

- Complexity of Data Reality
- Identifying the problem
- Access to right data – Data quantity
- Data Cleansing – Data quality - Data Security
- Granularity, Consistency Availability of Data
- Lack of domain expertise
- Cognitive Bias
- Content and Source Bias

COGNITIVE BIAS

- Cognitive Biases are the distortions of reality because of the lens through which we view the world.
- Each of us sees things differently based on our preconceptions, past experiences, cultural, environmental, and social factors. This doesn't necessarily mean that the way we think or feel about something is truly representative of reality.

TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

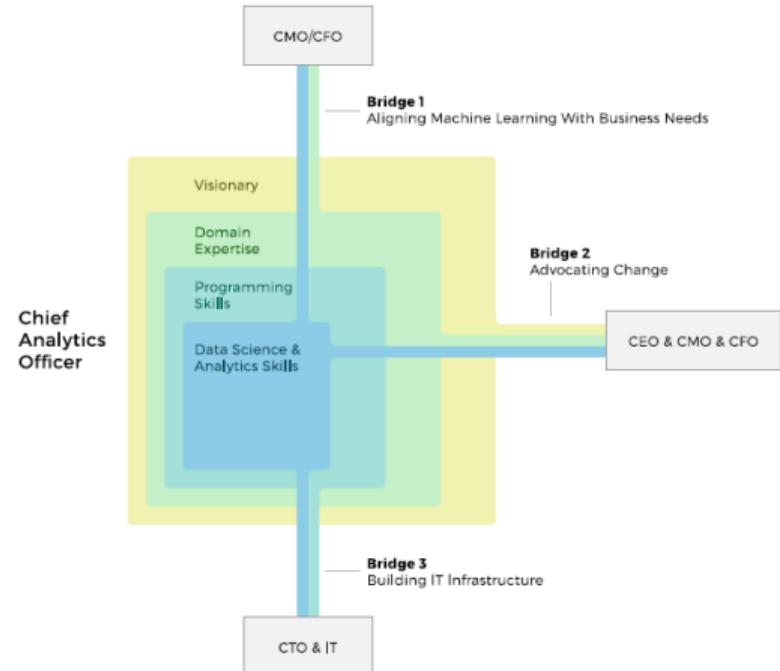
6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

ROLES IN DATA SCIENCE TEAM [1/6]

[1] Chief Analytics Officer / Chief Data Officer

- ⑤ CAO, a “business translator,” bridges the gap between data science and domain expertise acting both as a visionary and a technical lead.
- ⑤ Preferred skills: data science and analytics, programming skills, domain expertise, leadership and visionary abilities.



<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

ROLES IN DATA SCIENCE TEAM [2/6]

[2] Data analyst

- The data analyst role implies proper data collection and interpretation activities.
- An analyst ensures that collected data is relevant and exhaustive while also interpreting the analytics results.
- May require data analysts to have visualization skills to convert alienating numbers into tangible insights through graphics. (eg: IBM or HP)
- Preferred skills: R, Python, JavaScript, C/C++, SQL

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

ROLES IN DATA SCIENCE TEAM [3/6]

3 Business analyst

-) A business analyst basically realizes a CAO's functions but on the operational level.
-) This implies converting business expectations into data analysis.
-) If your core data scientist lacks domain expertise, a business analyst bridges this gulf.
-) Preferred skills: data visualization, business intelligence, SQL.

4 Data scientist

-) A data scientist is a person who solves business tasks using machine learning and data mining techniques.
-) The role can be narrowed down to data preparation and cleaning with further model training and evaluation.
-) Preferred skills: R, SAS, Python, Matlab, SQL, noSQL, Hive, Pig, Hadoop, Spark

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

ROLES IN DATA SCIENCE TEAM [4/6]

Job of a data scientist is often divided into two roles

[4A] Machine Learning Engineer

- › A machine learning engineer combines software engineering and modeling skills by determining which model to use and what data should be used for each model.
- › Probability and statistics are also their forte.
- › Training, monitoring, and maintaining a model.
- › Preferred skills: R, Python, Scala, Julia, Java

[4B] Data Journalist

- › Data journalists help make sense of data output by putting it in the right context.
- › Articulating business problems and shaping analytics results into compelling stories.
- › Present the idea to stakeholders and represent the data team with those unfamiliar with statistics.
- › Preferred skills: SQL, Python, R, Scala, Carto, D3, QGIS, Tableau

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

ROLES IN DATA SCIENCE TEAM [5/6]

5 Data architect

- › Working with Big Data.
- › This role is critical to warehouse the data, define database architecture, centralize data, and ensure integrity across different sources.
- › Preferred skills: SQL, noSQL, XML, Hive, Pig, Hadoop, Spark

6 Data engineer

- › Data engineers implement, test, and maintain infrastructural components that data architects design.
- › Realistically, the role of an engineer and the role of an architect can be combined in one person.
- › Preferred skills: SQL, noSQL, Hive, Pig, Matlab, SAS, Python, Java, Ruby, C++, Perl

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

ROLES IN DATA SCIENCE TEAM [6/6]

[7] Application/data visualization engineer

- This role is only necessary for a specialized data science model.
- An application engineer or other developers from front-end units will oversee end-user data visualization.
- Preferred skills: programming, JavaScript (for visualization), SQL, noSQL.

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

DATA SCIENTIST

Stitch Fix's Michael Hochster defines two types of data scientists:

- Type A stands for Analysis
 - This person is a statistician that makes sense of data without necessarily having strong programming knowledge.
 - Type A data scientists perform data cleaning, forecasting, modeling, visualization, etc.
- Type B stands for Building
 - These folks use data in production.
 - They're excellent good software engineers with some statistics background who build recommendation systems, personalization use cases, etc.

<https://www.alteXsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/>

SKILLSET FOR A DATA SCIENTIST

PROGRAMMING: Most fundamental of a data scientist's skill set. Programming improves your statistics skills, helps you "analyze large datasets" and gives you the ability to create your own tools.

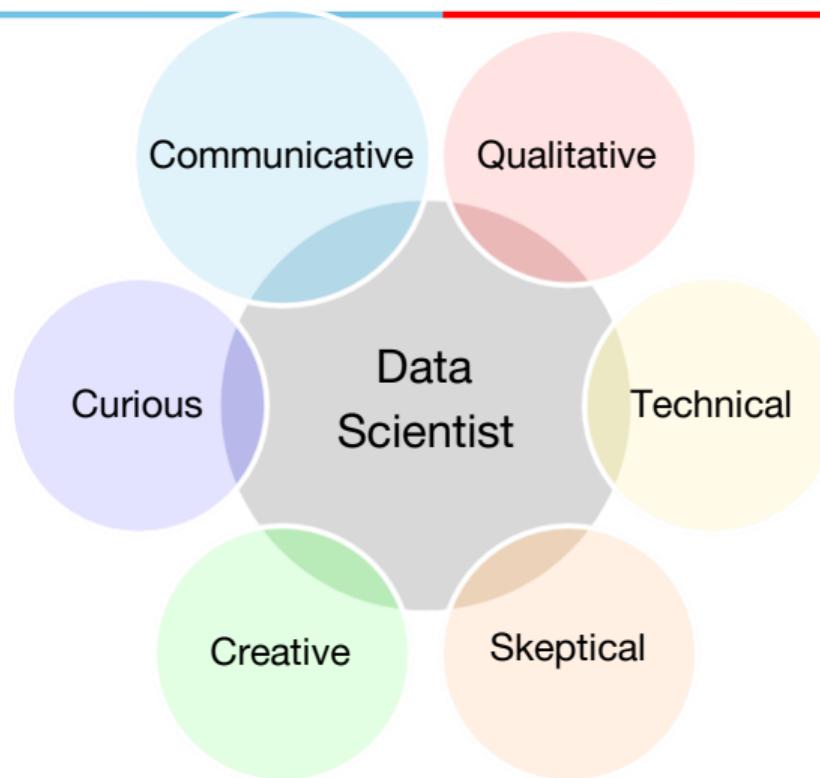
QUANTITATIVE ANALYSIS: Improve your ability to run experimental analysis, scale your data strategy and help you implement machine learning.

PRODUCT INTUITION: Understanding products will help you perform quantitative analysis. It will also help you predict system behavior, establish metrics and improve debugging skills.

COMMUNICATION: Strong communication skills will help you "leverage all of the previous skills listed."

TEAMWORK: It requires being selfless, embracing feedback and sharing your knowledge with your team.

SKILLS REQUIRED FOR A DATA SCIENTIST



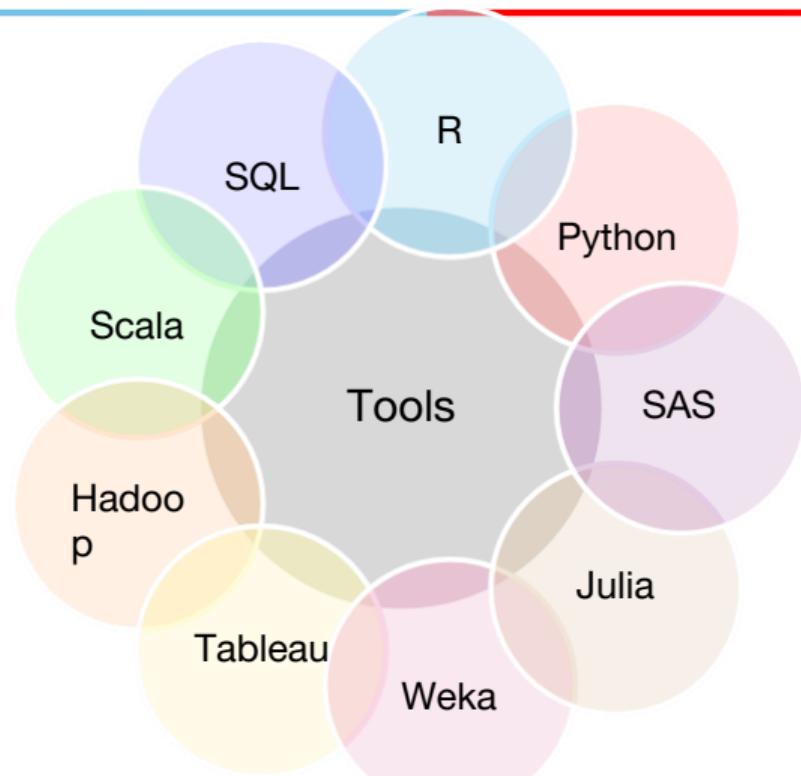
SKILLSET OF A DATA SCIENTIST

NECESSARY AND PREFERRED DATA SCIENCE SKILLS		
Analytics	R/SAS	necessary
Coding	R, Python, Java, C/C++	necessary
Databases	SQL, NoSQL (MongoDB, CouchDB, Cassandra, MemcacheDB, etc.)	necessary
Big Data Processing	Hadoop, Spark, Flink	preferred
Algorithms and Models	Regression models, Hidden Markov models, Support Vector Machines, Dimensionality Reduction algorithms, Ensemble algorithms, Decision Trees, Clustering	necessary
Frameworks and Libraries	TensorFlow, Theano, CNTK, scikit-learn, Caffe, Spark MLlib, etc.	preferred
Domain knowledge	Understanding of company goals, industry fundamentals, business problems, finding new ways to leverage data	preferred
Other	Intellectual curiosity, communication and presentation skills	preferred

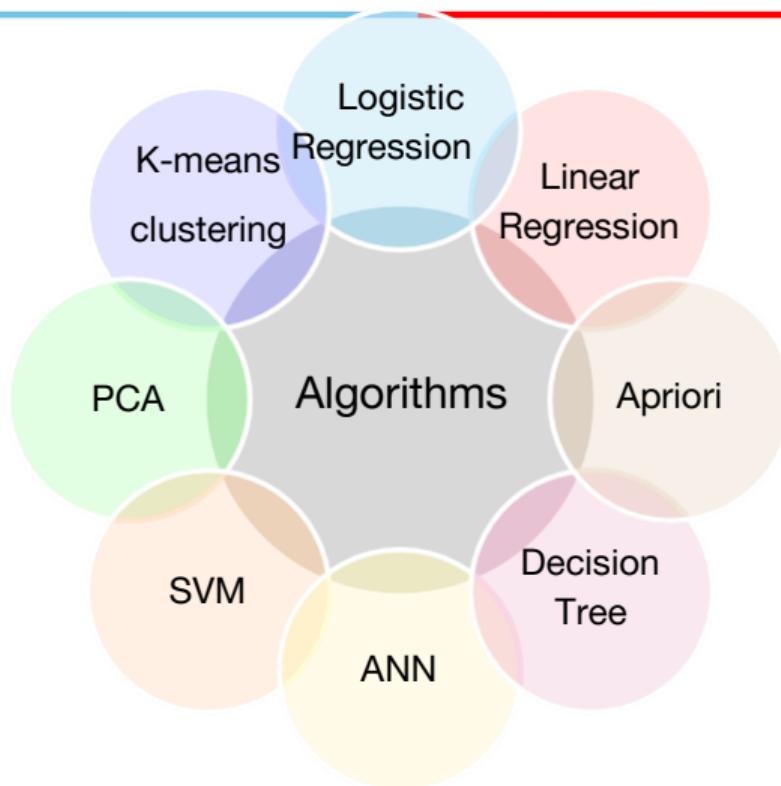


altesoft
software engineering

TOOLS AVAILABLE TO A DATA SCIENTIST



ALGORITHMS FOR A DATA SCIENTIST



DATA SCIENCE TEAM BUILDING

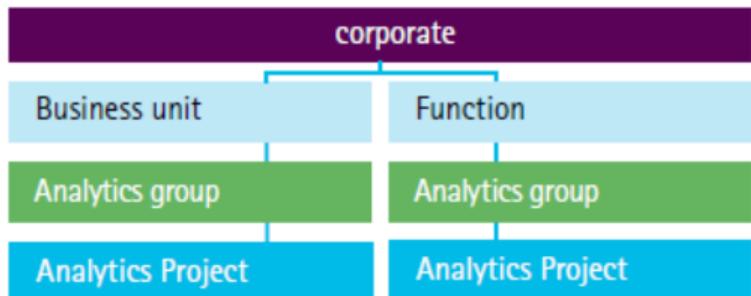
- Get to know each other for better communication
- Foster team cohesion and teamwork
- Encourage collaboration to boost team productivity and performance.

ORGANISATION OF DATA SCIENCE TEAM

[1] Decentralized

-) Data scientists report into specific business units (ex: Marketing) or functional units (ex: Product Recommendations) within a company.
-) Resources allocated only to projects within their silos with no view of analytics activities or priorities outside their function or business unit.
-) Analytics are scattered across the organization in different functions and business units.
-) Little to no coordination
-) Drawback – lead to isolated teams

Decentralized

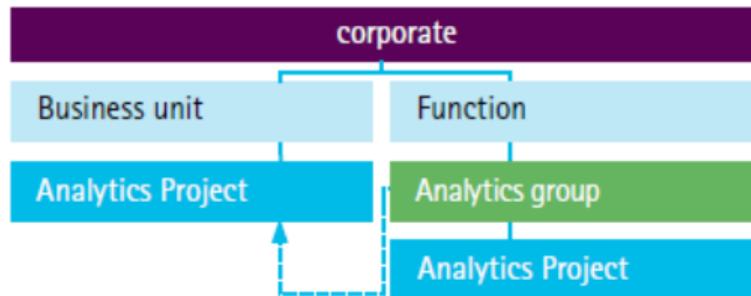


ORGANISATION OF DATA SCIENCE TEAM

[2] Functional

- Resource allocation driven by a functional agenda rather than an enterprise agenda.
- Analysts are located in the functions where the most analytical activity takes place, but may also provide services to rest of the corporation.
- Little coordination

Functional

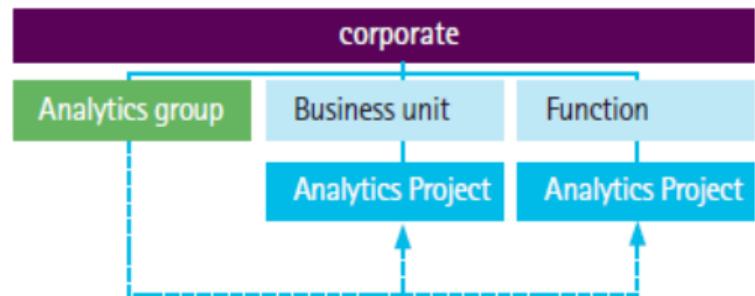


ORGANISATION OF DATA SCIENCE TEAM

[3] Consulting

- Resources allocated based on availability on a first-come first-served basis without necessarily aligning to enterprise objectives
- Analysts work together in a central group but act as internal consultants who charge “clients” (business units) for their services
- No centralized coordination

Consulting

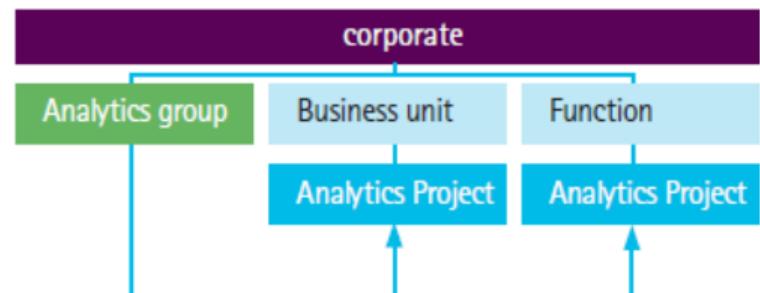


ORGANISATION OF DATA SCIENCE TEAM

[4] Centralized

- Data scientists are members of a core group, reporting to a head of data science or analytics.
- Stronger ownership and management of resource allocation and project prioritization within a central pool.
- Analysts reside in central group, where they serve a variety of functions and business units and work on diverse projects.
- Coordination by central analytic unit
- Challenge – Hard to assess and meet demands for incoming data science projects. (esp in smaller teams)

Centralized

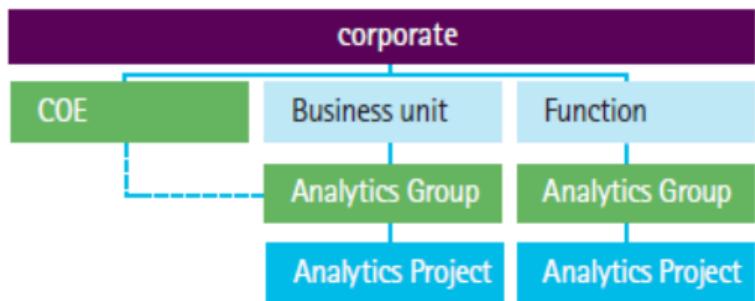


ORGANISATION OF DATA SCIENCE TEAM

[5] Center of Excellence

-) Better alignment of analytics initiatives and resource allocation to enterprise priorities without operational involvement.
-) Analysts are allocated to units throughout the organization and their activities are coordinated by a central entity.
-) Flexible model with right balance of centralized and distributed coordination.

Center of Excellence



ORGANISATION OF DATA SCIENCE TEAM

[6] Federated

-) Same as “Center of Excellence” model with need-based operational involvement to provide SME support.
-) A centralized group of advanced analysts is strategically deployed to enterprise-wide initiatives.
-) Flexible model with right balance of centralized and distributed coordination.

Federated

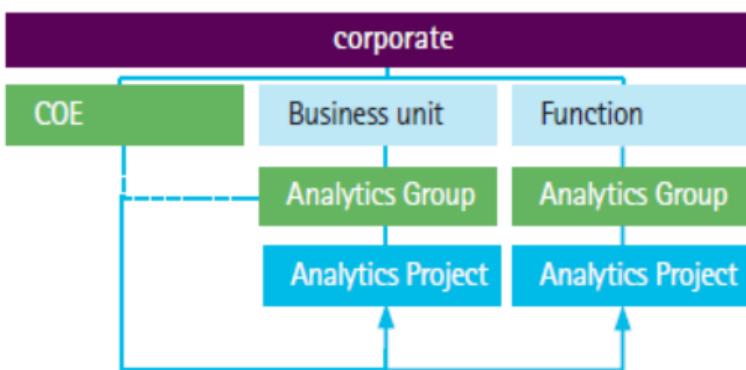


TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

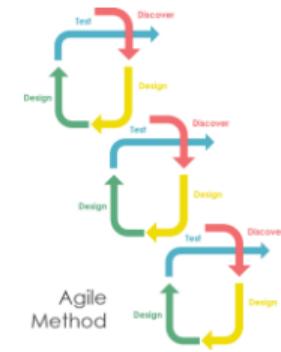
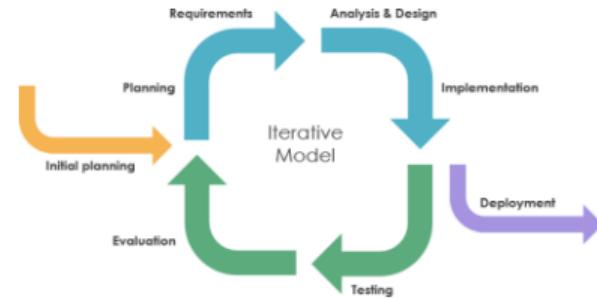
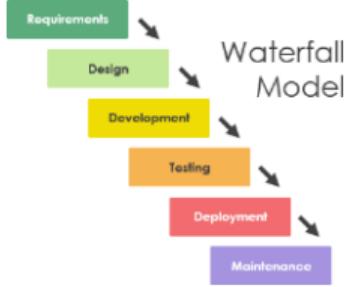
6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

SOFTWARE ENGINEERING

In general,

- Software engineering is an engineering discipline that is concerned with all aspects of software production.
 - Software includes computer programs, all associated documentation, and configuration data that are needed for software to work correctly.
 - Waterfall model, Iterative models, Agile models



DATA SCIENCE PROCESS



DATA SCIENCE VS. SOFTWARE ENGINEERING

Data Science	Software Engineering
<p>Data science involves analyzing huge amounts of data, with some aspects of programming and development.</p>	<p>Software engineering focuses on creating software that serves a specific purpose.</p>
<p>Uses a methodology involving various phases beginning from requirements specification through model deployment to better decision making.</p>	<p>Uses a methodology involving various phases beginning from requirements specification through software deployment into production.</p>

DATA SCIENCE VS. SOFTWARE ENGINEERING

Data Science	Software Engineering
Involves collecting and analyzing data	Concerned with creating useful applications
Data scientists utilize the ETL (Extract, Transform, Load) process	Software engineers use the SDLC process
More process-oriented	Uses frameworks like Waterfall, Agile, and Spiral
Data scientists use tools like Amazon S3, MongoDB, Hadoop, and MySQL	Software engineers use tools like Rails, Django, Flask, and Vue.js
Skills include machine learning, statistics, and data visualization	Skills are focused on coding languages

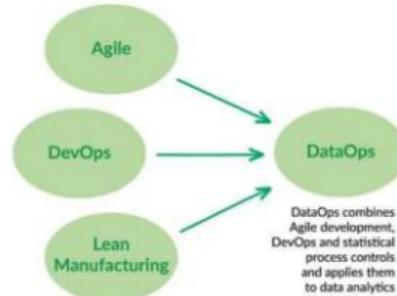
DATAOPS

DATAOPS AS DEFINED BY GARTNER

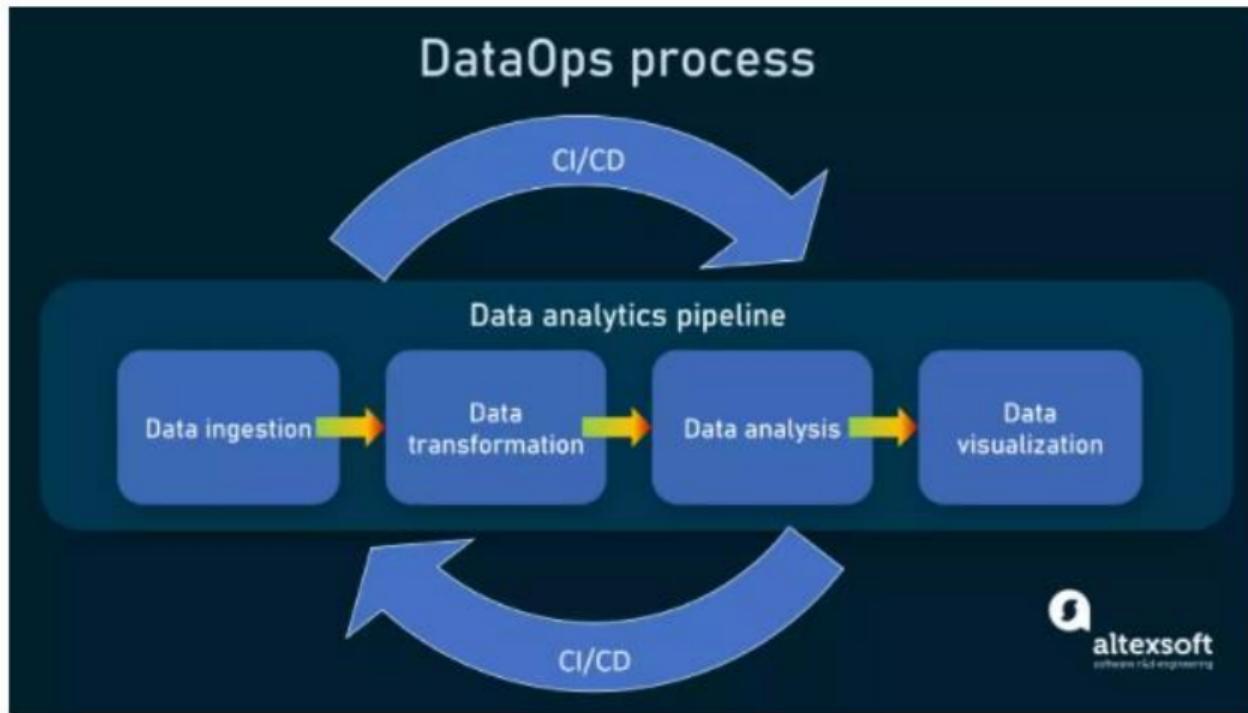
DataOps is a collaborative data management practice, really focused on improving communication, integration, and automation of data flow between managers and consumers of data within an organization.

DATAOPS

- DataOps applies Agile development, DevOps and lean manufacturing to data analytics development and operations.
- Agile governs analytics development.
- DevOps optimizes code verification, builds and delivery of new analytics.
- Lean-manufacturing tool, statistical process control (SPC) orchestrates, monitors and validates the data factory.



DATAOPS



DATAOPS

Data analytics pipeline

- 1 Data ingestion – Data, extracted from various sources, is explored, validated, and loaded into a downstream system.
- 2 Data transformation – Data is cleansed and enriched. Initial data models are designed to meet business needs.
- 3 Data analysis – produce insights using different data analysis techniques.
- 4 Data visualization/reporting – Data insights are represented in the form of reports or interactive dashboards.

<https://www.alteXsoft.com/blog/dataops->

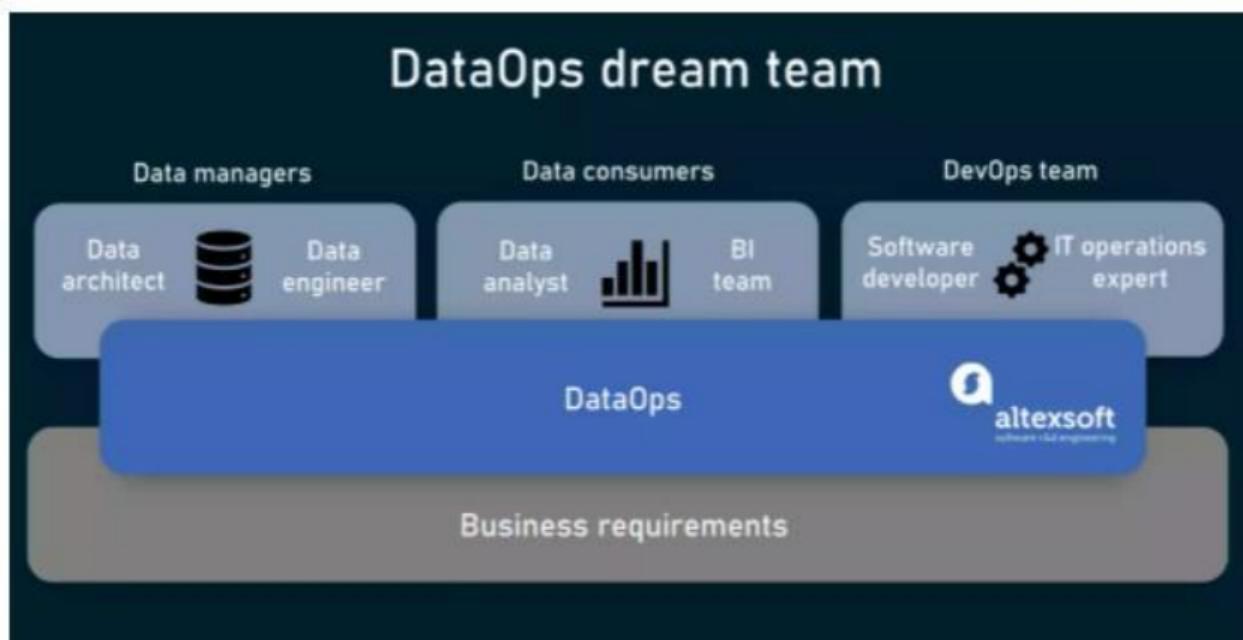
DATAOPS

DataOps puts data pipelines into a CI/CD paradigm.

- Development – involve building a new pipeline, changing a data model or redesigning a dashboard.
- Testing – checking the most minor update for data accuracy, potential deviation, and errors.
- Deployment – moving data jobs between environments, pushing them to the next stage, or deploying the entire pipeline in production.
- Monitoring – allows data professionals to identify bottlenecks, catch abnormal patterns, and measure adoption of changes.
- Orchestration – automates moving data between different stages, monitoring progress, triggering autoscaling, and operations related to data flow management.

<https://www.alteXsoft.com/blog/dataops->

DATAOPS TEAM

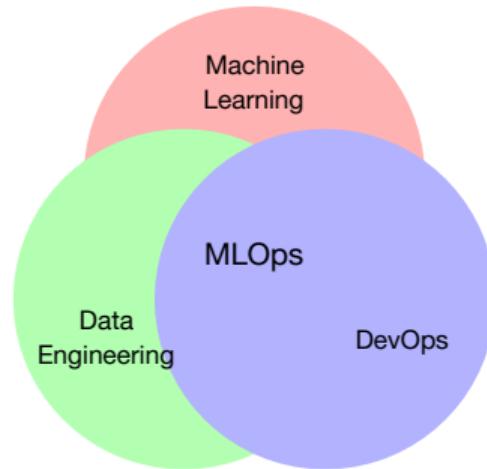


TECHNOLOGIES TO RUN DATAOPS

- [Git](#) for version control
- [Jenkins](#) for CI/CD practices
- [Docker](#) for containerization and [Kubernetes](#) for managing containers
- [Tableau](#) for data visualizations
- [Apache Airflow](#) for data pipeline tools
- Automated testing and monitoring tools
- DataOps Platforms
 -) [DataKitchen](#)
 -) [Saagie](#)
 -) [StreamSets](#)

MLOPS

MLOps is an ML engineering culture and practice that aims at unifying ML system development (Dev) and ML system operation (Ops).

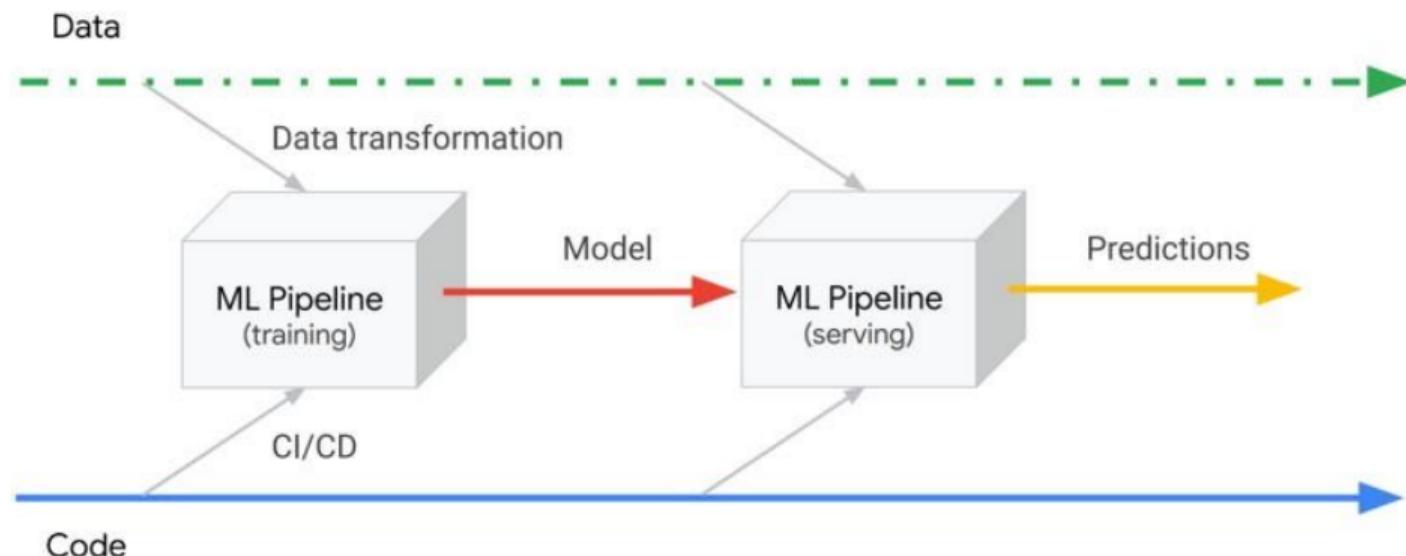


MLOPS

- Real challenge isn't building an ML model, but building an integrated ML system and to continuously operate it in production.
- To deploy and maintain ML systems in production reliably and efficiently.
- Automating continuous integration (CI), continuous delivery (CD), and continuous training (CT) for machine learning (ML) systems.
- Frameworks
 -) Kubeflow and Cloud Build
 -) Amazon AWS MLOps
 -) Microsoft Azure MLOps

<https://ml-ops.org/content/mlops->

MLOPS



MLOPS

- Same data transformations but different implementations .
e.g training pipeline usually runs over batch files that contain all features, while the serving pipeline often runs online and receives only part of the features in the requests
- Two pipelines are consistent, so code reuse and data reuse.
- Each trained model need to tied to the exact versions of code, data and hyperparameters that were used.

DATAOPS AND MLOPS

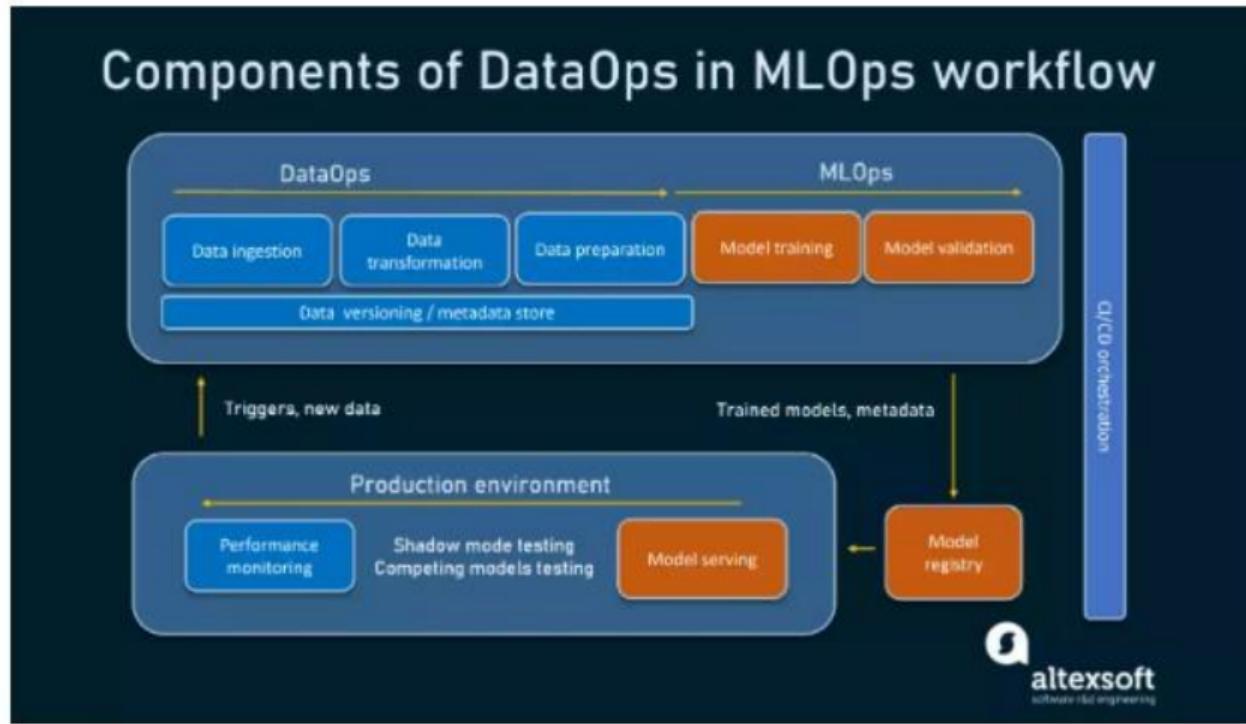


TABLE OF CONTENTS

1 COURSE LOGISTICS

2 FUNDAMENTALS OF DATA SCIENCE

3 DATA SCIENCE REAL WORLD APPLICATIONS

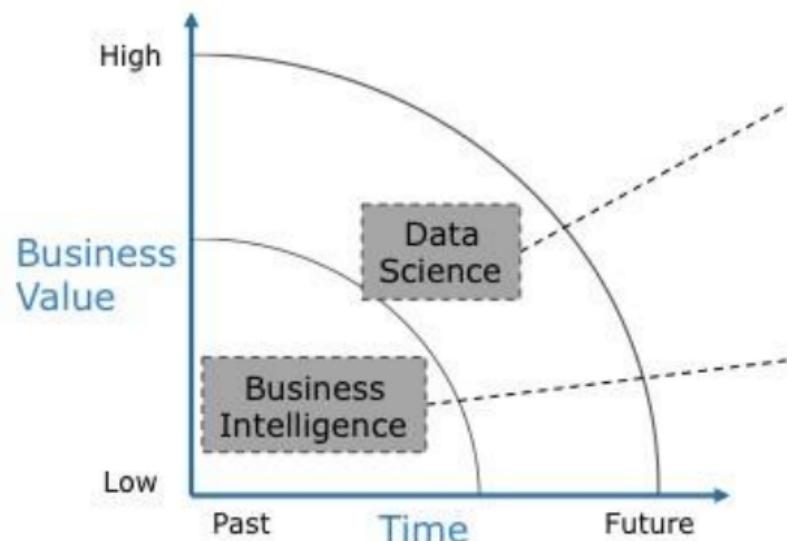
4 DATA SCIENCE CHALLENGES

5 DATA SCIENCE TEAMS

6 SOFTWARE ENGINEERING FOR DATA SCIENCE

7 FURTHER READING

DATA SCIENCE VS. BUSINESS INTELLIGENCE



Data Science

- Predictive analysis
- Prescriptive analysis
- Why...? What will...?
- What should I do...?

Business Intelligence

- Descriptive analysis
- Standard reporting
- What happened?

DATA SCIENCE VS. BUSINESS INTELLIGENCE

	Data Science	Business Intelligence
Perspective	Looking forward	Looking backward
Analysis	Predictive Explorative	Descriptive Comparative
Data	Same data, New analysis Listens to data Distributed	New Data, Same analysis Speaks for data Warehoused
Scope	Specific to business question	Unlimited
Expertise	Data scientist	Business analyst
Deliverable	Insight or story	Table or report
Applicability	Future, correction for influences	Historic, confounding factors

DATA SCIENTIST VS. BUSINESS ANALYST

Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analysis	Retrospective, Descriptive	Predictive, Prescriptive

DATA SCIENCE VS. STATISTICS

	Data Science	Statistics
Type of problem	Semi structured or unstructured	Well structured
Inference model	Explicit inference	No inference
Analysis Objective	Need not be well formed	Well formed objective
Type of Analysis	Explorative	Confirmative
Data collection	Data collection is not linked to the objective	Data collected based on the objective
Size of dataset	Large Heterogeneous	Small Homogeneous
Paradigm	Theory and heuristic (deductive & inductive)	Theory based (deductive)

REFERENCES

- Introducing Data Science by Cielen, Meysman and Ali
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- <https://data-flair.training/blogs/data-science-use-cases/> <https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/>
- <https://www.visual-paradigm.com/guide/software-development-process/> [what-is-a-software-process-model/](https://www.visual-paradigm.com/guide/software-development-process/)
- Building an Analytics-Driven Organization, Accenture
-

REFERENCES

- [https://www.altexsoft.com/blog/datascience/
how-to-structure-data-science-team-key-models-and-roles/](https://www.altexsoft.com/blog/datascience/how-to-structure-data-science-team-key-models-and-roles/)
- [https://www.cio.com/article/3217026/
what-is-a-data-scientist-a-key-data-analytics-role-and-a-lucrative-career.html](https://www.cio.com/article/3217026/what-is-a-data-scientist-a-key-data-analytics-role-and-a-lucrative-career.html)
- <https://atlan.com/what-is-dataops/>
-

THANK YOU

IMP Note to Self





BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 2 : DATA ANALYTICS

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

1 ANALYTICS

2 DATA ANALYTICS

3 DATA ANALYTICS METHODOLOGIES

- CRISP-DM
- Big Data Life-cycle
- SEMMA
- SMAM

4 FURTHER READING

DEFINITION OF ANALYTICS – DICTIONARY

OXFORD Analytics is the systematic computational analysis of data or statistics.

CAMBRIDGE Analytics is a process in which a computer examines information using mathematical methods in order to find useful patterns.

DICTIONARY.COM Analytics is the analysis of data, typically large sets of business data, by the use of mathematics, statistics, and computer software.

Analytics is treated as both a noun and a verb.

DEFINITION OF ANALYTICS – WEBSITES

ORACLE Analytics is the process of discovering, interpreting, and communicating significant patterns in data and using tools to empower your entire organization to ask any question of any data in any environment on any device.

E DUREKA Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain.

INFORMATICA Data analytics is the pursuit of extracting meaning from raw data using specialized computer systems.

GOALS OF DATA ANALYTICS

- To predict something
 -) whether a transaction is a fraud or not
 -) whether it will rain on a particular day
 -) whether a tumour is benign or malignant
- To find patterns in the data
 -) finding the top 10 coldest days in the year
 -) which pages are visited the most on a particular website
 -) finding the most searched celebrity in a particular year
- To find relationships in the data
 -) finding similar news articles
 -) finding similar patients in an electronic health record system
 -) finding related products on an e-commerce website
 -) finding similar images
 -) finding correlation between news items and stock prices

TABLE OF CONTENTS

1 ANALYTICS

2 DATA ANALYTICS

3 DATA ANALYTICS METHODOLOGIES

- CRISP-DM
- Big Data Life-cycle
- SEMMA
- SMAM

4 FURTHER READING

DATA ANALYTICS

- Data analysis is defined as a process of cleaning, transforming, and modelling data to discover useful information for business decision-making.
- 4 different types of analytics
 - 1 Descriptive Analytics
 - 2 Diagnostic Analytics
 - 3 Predictive Analytics
 - 4 Prescriptive Analytics

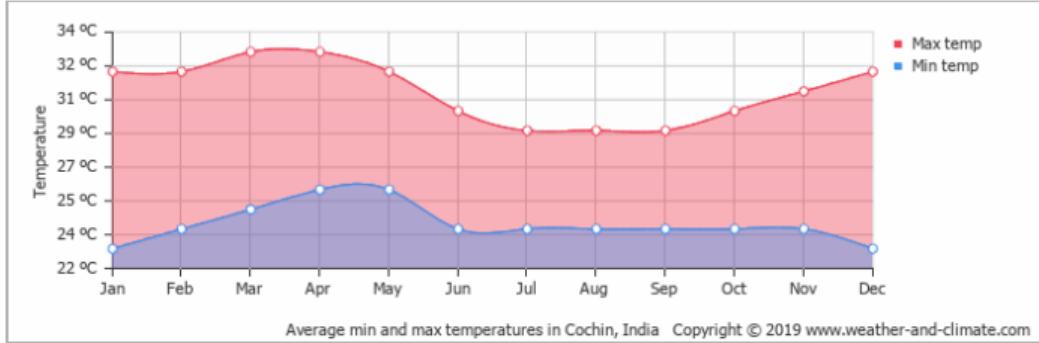
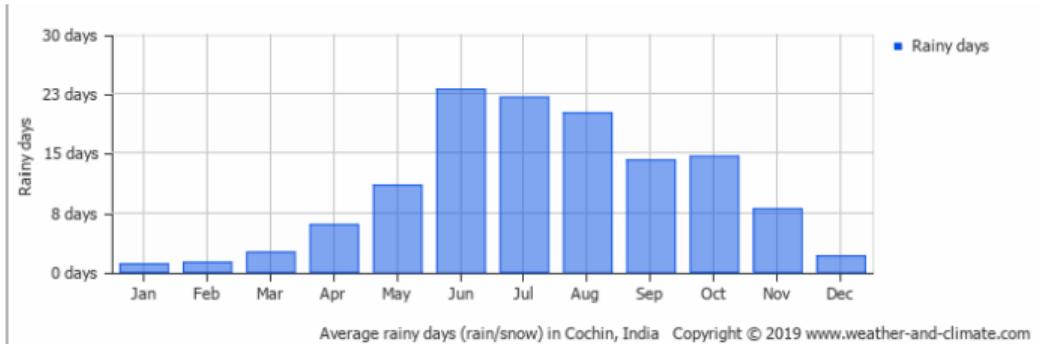
DATA ANALYTICS



DESCRIPTIVE ANALYTICS

- Answers the question of **what happened**.
- Summarize past data usually in the form of dashboards.
- **Insights into the past**.
- Also known as **statistical analysis**.
- Raw data from multiple data sources.

DESCRIPTIVE ANALYTICS EXAMPLE



DESCRIPTIVE ANALYTICS

■ Techniques:

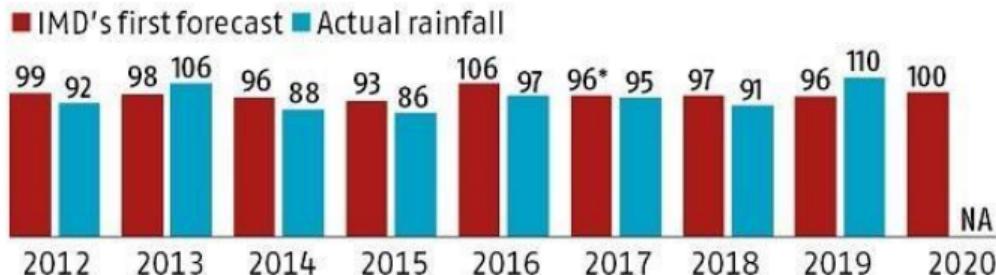
-) Descriptive Statistics - histogram, correlation
-) Data Visualization
-) Exploratory Analysis

PREDICTIVE ANALYTICS

- Answers the question of **what is likely to happen**.
- **Predict future trends.**
- Being able to predict allows one to make better decisions.
- Analysis based on machine or deep learning.
- Accuracy of the forecasting or prediction highly depends on data quality and stability of the situation.

PREDICTIVE ANALYTICS EXAMPLE

MONSOON FORECAST As % of long-period average



*Updated forecast on May 24; NOTE: All the forecasts are with a model error of plus and minus 5%. The first forecast is issued in April every year

LPA is the average rainfall received in the last 50 years, estimated to be 887 mm

Source: IMD

PREDICTIVE ANALYTICS

■ Techniques / Algorithms:

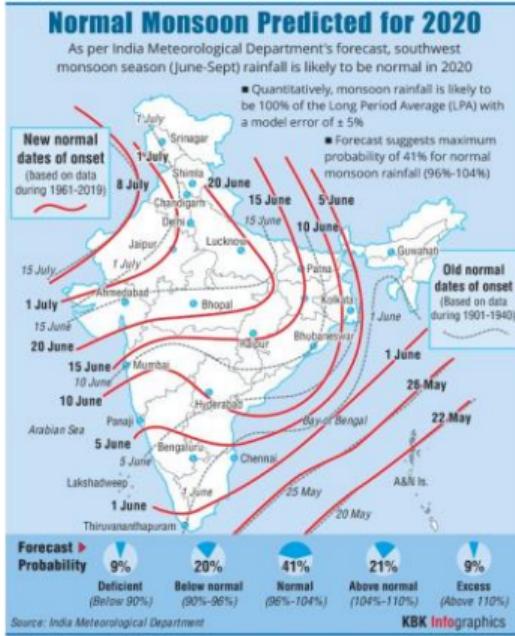
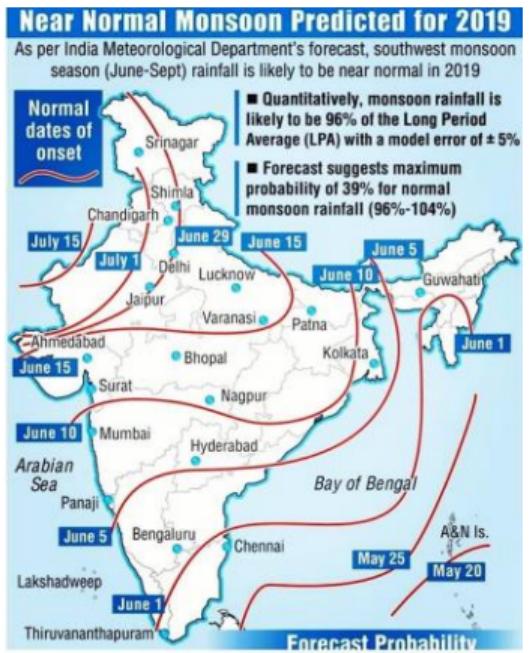
-) Regression
-) Classification
-) ML algorithms like Linear regression, Logistic regression, SVM
-) Deep Learning techniques

DIAGNOSTIC ANALYTICS

- Answers the question of **why** something happened.
- Gives in-depth insights into data.
- Identify relationship between data and identify patterns of behaviour.

DIAGNOSTIC ANALYTICS EXAMPLE

What is the effect of global warming in the Southwest monsoon?



DIAGNOSTIC ANALYTICS

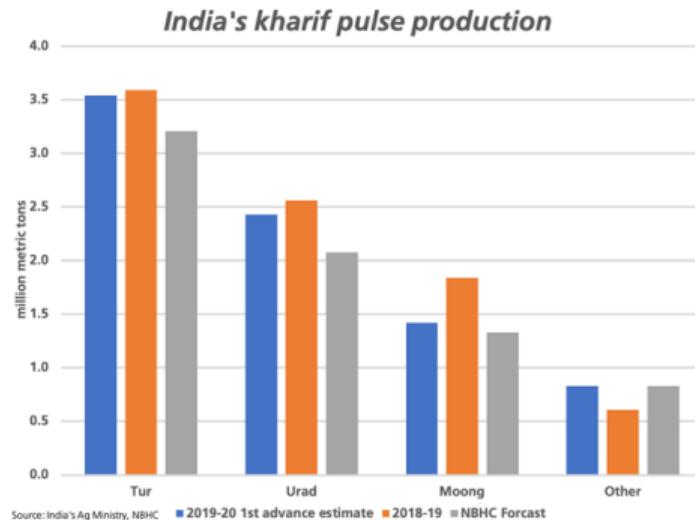
- Pattern recognition to identify patterns.
- Linear / Logistic regression to identify relationship.
- Neural Network
- Deep Learning techniques

PREScriptive A NALYTICS

- Answers the question of **what might happen**.
- **Data-driven decision making and corrective actions, recommendations and suggestions**
- Prescribe what action to take to eliminate a future problem or take full advantage of a promising trend.
- Need historical internal data and external information like trends.
- Analysis based on machine or deep learning, business rules.
- Use of AI to improve decision making.

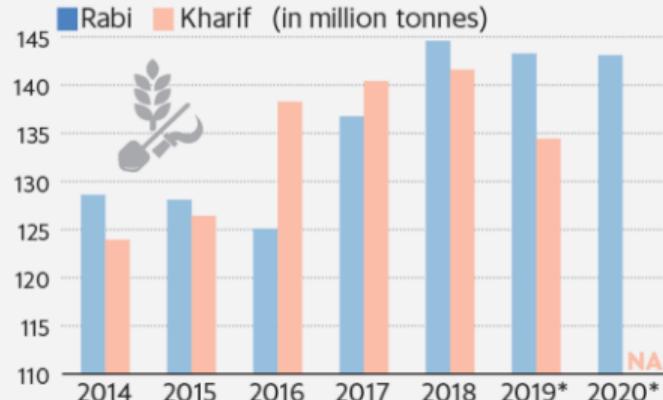
PREScriptive ANALYTICS EXAMPLE

How can we improve the crop production?



Food for thought

Erratic monsoon likely to take a toll on kharif crop.

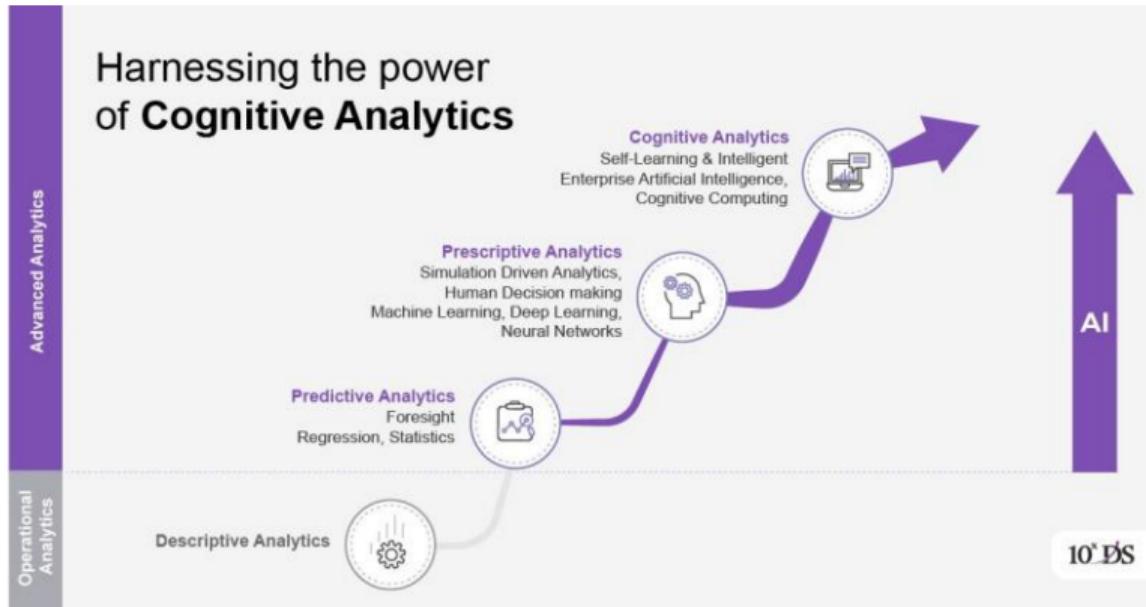


Note: Estimate for Kharif of 2019 are Crisil Research estimates, 2019-20 rabi production is government target

Source: Ministry of agriculture

COGNITIVE ANALYTICS

Cognitive Analytics - What Don't I Know?



COGNITIVE ANALYTICS

- Next level of Analytics
- Human cognition is based on the context and reasoning.
- Cognitive systems mimic how humans reason and process.
- Cognitive systems analyse information and draw inferences using probability.
- They continuously learn from data and reprogram themselves.
- According to one source:
"The essential distinction between cognitive platforms and artificial intelligence systems is that you want an AI to do something for you. A cognitive platform is something you turn to for collaboration or for advice."

<https://interestingengineering.com/cognitive-computing-more-human-than-artificial-intelligence>

COGNITIVE ANALYTICS

- Involves Semantics, AI, Machine learning, Deep Learning, Natural Language Processing, and Neural Networks.
- Simulates human thought process to learn from the data and extract the hidden patterns from data.
- Uses all types of data: audio, video, text, images in the analytics process.
- Although this is the top tier of analytics maturity, Cognitive Analytics can be used in the prior levels.
- According to Jean Francois Puget:
"It extends the analytics journey to areas that were unreachable with more classical analytics techniques like business intelligence, statistics, and operations research."

<https://www.ecapitaladvisors.com/blog/analytics-maturity/>

<https://www.xenonstack.com/insights/what-is-cognitive-analytics/>

TABLE OF CONTENTS

1 ANALYTICS

2 DATA ANALYTICS

3 DATA ANALYTICS METHODOLOGIES

- CRISP-DM
- Big Data Life-cycle
- SEMMA
- SMAM

4 FURTHER READING

DATA ANALYTICS METHODOLOGIES

- Use standard methodology to ensure a good outcome.

- 1 CRISP-DM
- 2 Big Data Life-cycle
- 3 SEMMA
- 4 SMAM

NEED FOR A STANDARD PROCESS

- Framework for recording experience.
 -) Allows projects to be replicated
- Aid to project planning and management.
- “Comfort factor” for new adopters
 -) Demonstrates maturity of Data Mining
 -) Reduces dependency on “stars”
- Encourage best practices and help to obtain better results.

DATA SCIENCE METHODOLOGY

10 Questions the process aims to answer

■ Problem to Approach

- 1 What is the problem that you are trying to solve?
- 2 How can you use data to answer the questions?

■ Working with Data

- 3 What data do you need to answer the question?
- 4 Where is the data coming from? Identify all Sources. How will you acquire it?
- 5 Is the data that you collected representative of the problem to be solved?
- 6 What additional work is required to manipulate and work with the data?

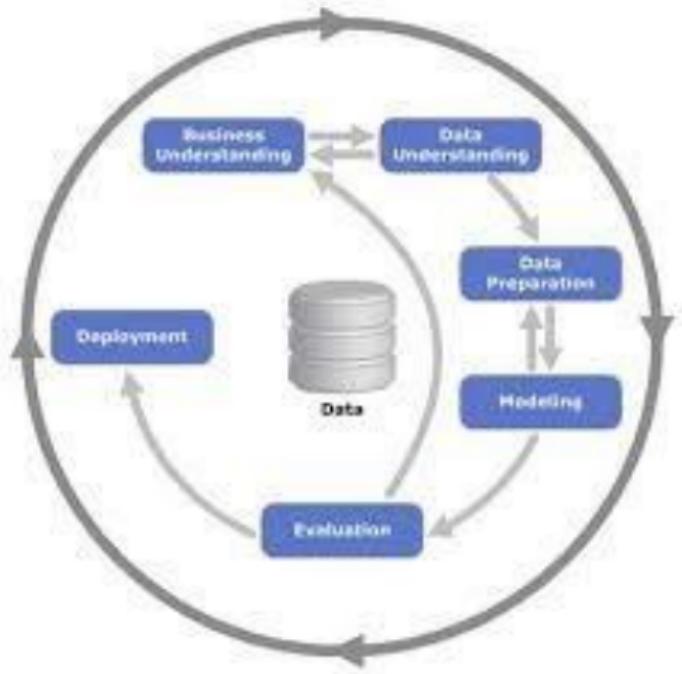
■ Delivering the Answer

- 7 In what way can the data be visualized to get to the answer that is required?
- 8 Does the model used really answer the initial question or does it need to be adjusted?
- 9 Can you put the model into practice?
- 10 Can you get constructive feedback into answering the question?

CRISP-DM

- Cross Industry Standard Process for Data Mining
- conceived around 1996
- 6 high-level phases
- Used in IBM SPSS Modeler tool
- Iterative approach to the development of analytical models.

CRISP-DM Phases



CRISP-DM PHASES

■ Business Understanding

-) Understand project objectives and requirements.
-) Data mining problem definition.

■ Data Understanding

-) Initial data collection and familiarization.
-) Identify data quality issues.
-) Identify initial obvious results.

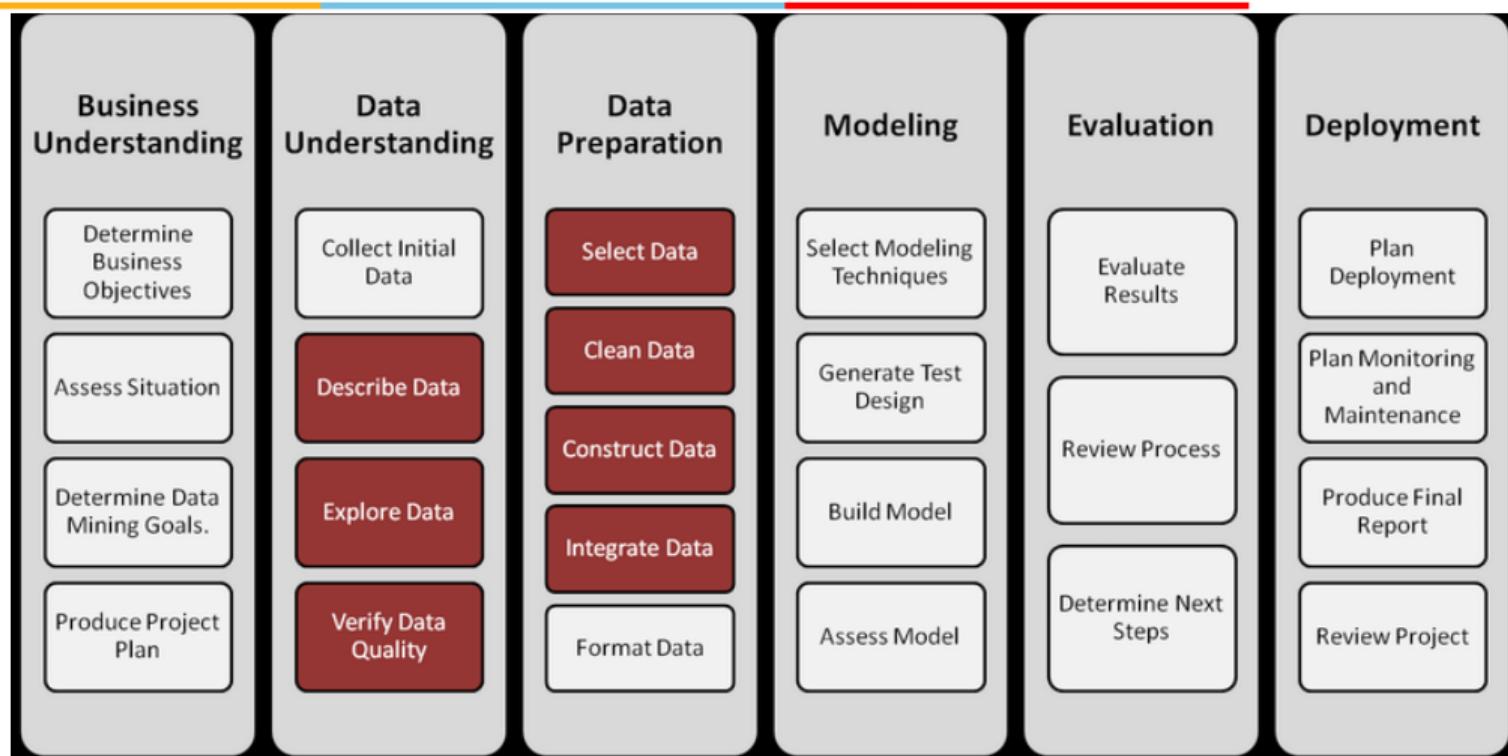
■ Data Preparation

-) Record and attribute selection.
-) Data cleansing.

CRISP-DM PHASES

- Modeling
 -) Run the data mining tools.
- Evaluation
 -) Determine if results meet business objectives.
 -) Identify business issues that should have been addressed earlier.
- Deployment
 -) Put the resulting models into practice.
 -) Set up for continuous mining of the data.

CRISP-DM PHASES AND TASKS



WHY CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills.
- CRISP-DM provides a uniform framework for
 -) guidelines.
 -) experience documentation.
- CRISP-DM is flexible to account for differences.
 -) Different business/agency problems.
 -) Different data

BIG DATA LIFE-CYCLE

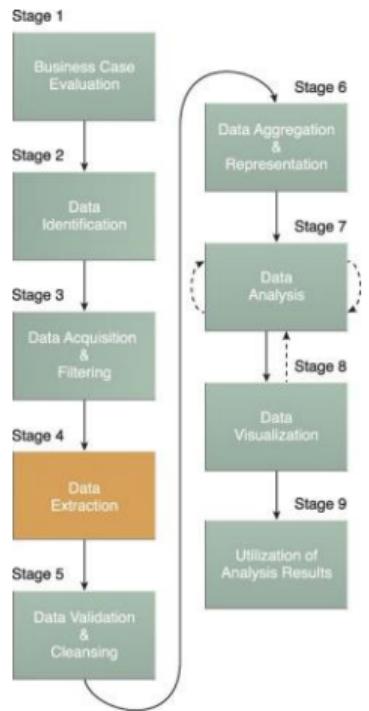
- Data Acquisition
 -) Acquiring information from a rich and varied data environment.
- Data Awareness
 -) Connecting data from different sources into a coherent whole, including modeling content, establishing context, and insuring search-ability.
- Data Analytics
 -) Using contextual data to answer questions about the state of your organization.
- Data Governance
 -) Establishing a framework for providing for the provenance, infrastructure and disposition of that data.

BIG DATA LIFE-CYCLE

-
- Phase 1: Foundations
 - Phase 2: Acquisition
 - Phase 3: Preparation
 - Phase 4: Input and Access
 - Phase 5: Processing
 - Phase 6: Output and Interpretation
 - Phase 7: Storage
 - Phase 8: Integration
 - Phase 9: Analytics and Visualization
 - Phase 10: Consumption
 - Phase 11: Retention, Backup, and Archival
 - Phase 12: Destruction

PS: Some phases may overlap and can be done in parallel.

BIG DATA LIFE-CYCLE



BIG DATA LIFE-CYCLE

■ Phase 1: Foundations

-) Understanding and validating data requirements, solution scope, roles and responsibilities, data infrastructure preparation, technical and non-technical considerations, and understanding data rules in an organization.

■ Phase 2: Data Acquisition

-) Data Acquisition refers to collecting data.
-) Data sets can be obtained from various sources, both internal and external to the business organizations.
-) Data sources can be in
 - 2 structured forms such as transferred from a data warehouse, a data mart, various transaction systems.
 - 2 semi-structured sources such as Weblogs, system logs.
 - 2 unstructured sources such as media files consisting of videos, audios, and pictures.

BIG DATA LIFE-CYCLE

■ Phase 3: Data Preparation

-) Collected data (Raw Data) is rigorously checked for inconsistencies, errors, and duplicates.
-) Redundant, duplicated, incomplete, and incorrect data are removed.
-) The objective is to have clean and useable data sets.

■ Phase 4: Data Input and Access

-) Data input refers to sending data to planned target data repositories, systems, or applications.
-) Data can be stored in CRM (Customer Relationship Management) application, a data lake or a data warehouse.
-) Data access refers to accessing data using various methods.
-) NoSQL is widely used to access big data.

BIG DATA LIFE-CYCLE

■ Phase 5: Data Processing

-) Processing the raw form of data.
-) Convert data into a readable format giving it the form and the context.
-) Interpret the data using the selected data analytics tools such as Hadoop MapReduce, Impala, Hive, Pig, and Spark SQL.
-) Data processing also includes activities
 - 2 Data annotation - refers to labeling the data.
 - 2 Data integration - aims to combine data existing in different sources, and provide a unified view of data to the data consumers.
 - 2 Data representation - refers to the way data is processed, transmitted, and stored.
 - 2 Data aggregation - aims to compile data from databases to combined data-sets to be used for data processing.

BIG DATA LIFE-CYCLE

■ Phase 6: Data Output and Interpretation

-) In the data output phase, the data is in a format which is ready for consumption by the business users.
-) Transform data into usable formats such as plain text, graphs, processed images, or video files.
-) This phase is also called the **data ingestion**.
-) Common Big Data ingestion tools are Sqoop, Flume, and Spark streaming.
-) Interpreting the ingested data requires analyzing ingested data and extract information or meaning out of it to answer the questions related to the Big Data business solutions.

BIG DATA LIFE-CYCLE

■ Phase 7: Data Storage

-) Store data in designed and designated storage units.
-) Storage infrastructure can consist of storage area networks (SAN), network-attached storage (NAS), or direct access storage (DAS) formats.

■ Phase 8: Data Integration

-) Integration of stored data to different systems for various purposes.
-) Integration of data lakes with a data warehouse or data marts.

■ Phase 9: Data Analytics and Visualization

-) Integrated data can be useful and productive for data analytics and visualization.
-) Business value is gained in this phase.

BIG DATA LIFE-CYCLE

■ Phase 10: Data Consumption

-) Data is turned into information ready for consumption by the internal or external users, including customers of the business organization.
-) Data consumption require architectural input for policies, rules, regulations, principles, and guidelines.

■ Phase 11: Retention, Backup, and Archival

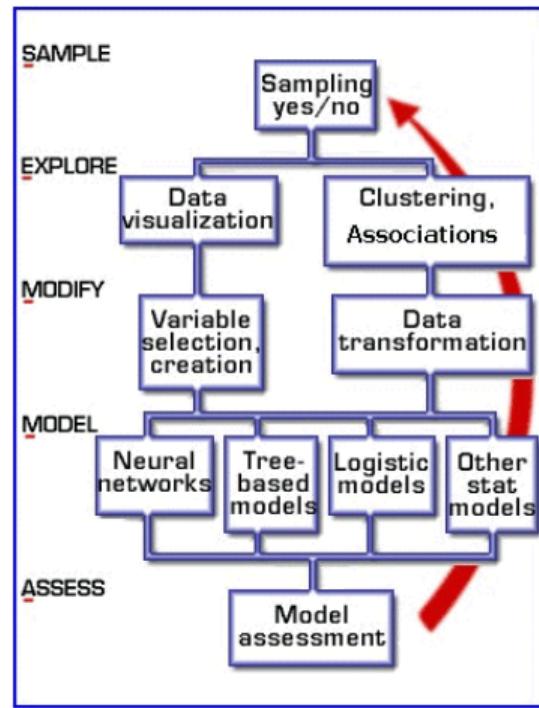
-) Use established data backup strategies, techniques, methods, and tools.
-) Identify, document, and obtain approval for the retention, backup, and archival decisions.

■ Phase 12: Data Destruction

-) There may be regulatory requirements to destruct a particular type of data after a certain amount of times.
-) Confirm the destruction requirements with the data governance team in business organizations.

SEMMA

- SAS Institute
- Sample, Explore, Modify, Model, Assess
- 5 stages



SEMMA STAGES

1 Sample

-) Sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.
-) Optional stage

2 Explore

-) Exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

3 Modify

-) Modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

SEMMA STAGES

1 Model

- Modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

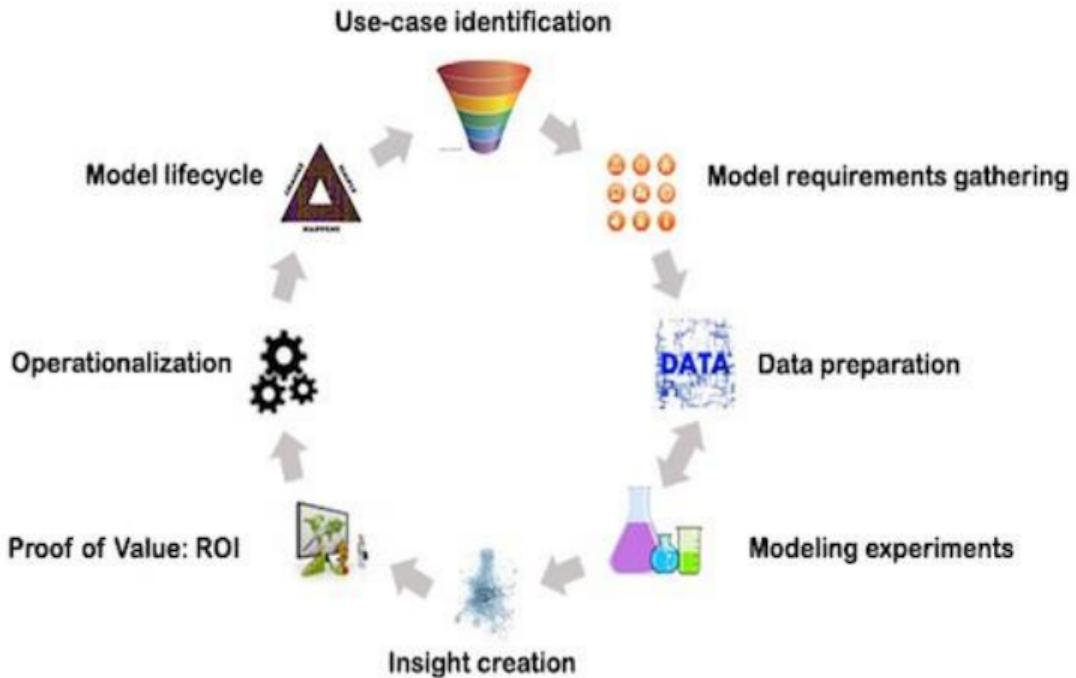
2 Assess

- Assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

SEMMA

- “SEMMA is not a data mining methodology but rather a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining.
- Enterprise Miner can be used as part of any iterative data mining methodology adopted by the client. Naturally steps such as formulating a well defined business or research problem and assembling quality representative data sources are critical to the overall success of any data mining project.
- SEMMA is focused on the model development aspects of data mining.”

- Standard Methodology for Analytics Models



SMAM PHASES

Phase	Description
Use-case identification	Selection of the ideal approach from a list of candidates
Model requirements gathering	Understanding the conditions required for the model to function
Data preparation	Getting the data ready for the modeling
Modeling experiments	Scientific experimentation to solve the business question
Insight creation	Visualization and dash-boarding to provide insight
Proof of Value: ROI	Running the model in a small scale setting to prove the value
Operationalization	Embedding the analytical model in operational systems
Model life-cycle	Governance around model lifetime and refresh

TABLE OF CONTENTS

1 ANALYTICS

2 DATA ANALYTICS

3 DATA ANALYTICS METHODOLOGIES

- CRISP-DM
- Big Data Life-cycle
- SEMMA
- SMAM

4 FURTHER READING

DESCRIPTIVE ANALYTICS – EXAMPLE #1

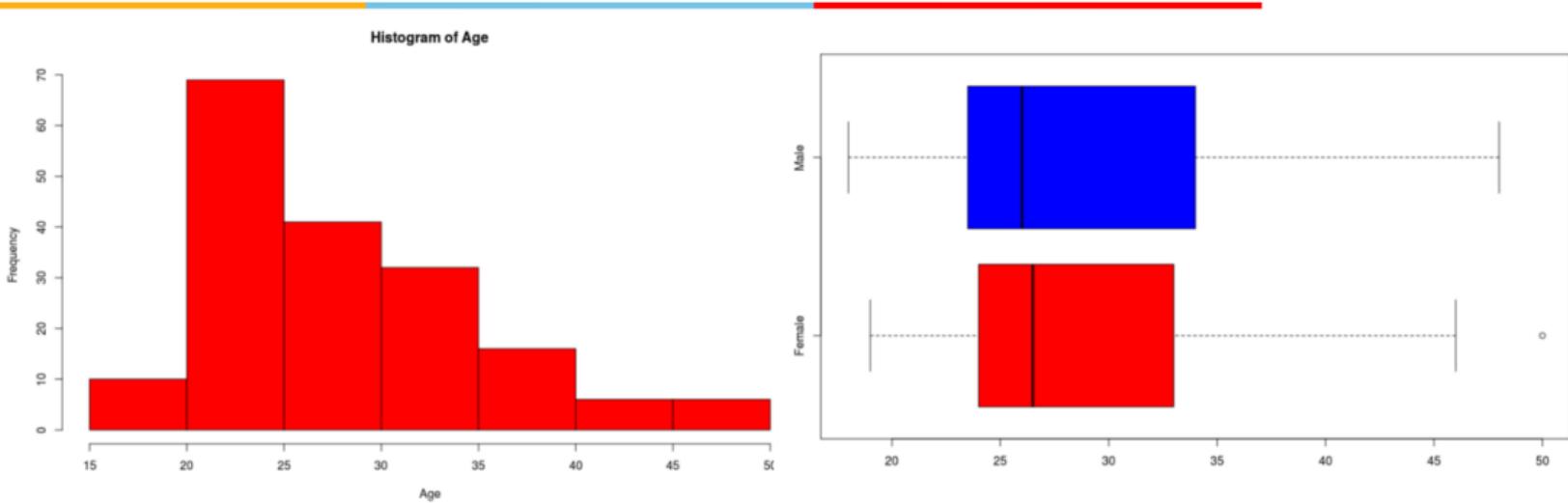
Problem Statement :

“Market research team at Aqua Analytics Pvt. Ltd is assigned a task to identify profile of a typical customer for a Digital fitness band that is offered by Titanic Corp. The market research team decides to investigate whether there are differences across the usage patterns and product lines with respect to customer characteristics”

Data captured

- Gender
- Age (In years)
- Education (In years)
- Relationship Status (Single or Partnered)
- Annual Household income
- Average number of times customer tracks activity each week
- Number of miles customer expect to walk each week
- Self-rated fitness on a scale 1-5 where 1 is poor shape and 5 is excellent.
- Models of the product purchased - IQ75, MZ65, DX87

DESCRIPTIVE ANALYTICS – EXAMPLE #1



Product	Usage	Gender										Totals
		Female	Male									
DX87	2				1	2	16	3	9	2	5	2
IQ75	3	13	6	19	16	7	15	1	1			60
MZ65	4	7	7	14	17	5	7	3				60
Totals	5	20	13	33	36	14	38	7	10	2	5	2
	6											180
	7											

Most of the customers use it 3–4 times a week. No female consumer has ever used IQ75 and MZ65 more than 5 time a week.

DIAGNOSTIC ANALYTICS – EXAMPLE #1

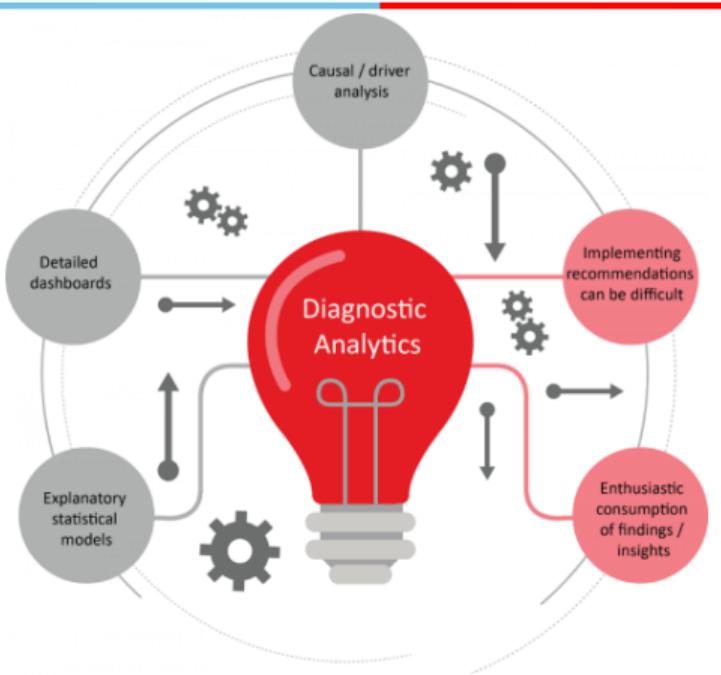
Problem Statement :

“During the 1980s General Electric was selling different products to its customers such as light bulbs, jet engines, windmills, and other related products. Also, they separately sell parts and services this means they would sell you a certain product you would use it until it needs repair either because of normal wear and tear or because it’s broken. And you would come back to GE and then GE would sell you parts and services to fix it. Model for GE was focusing on how much GE was selling, in sales of operational equipment, and in sales of parts and services. And what does GE need to do to drive up those sales?”

<https://medium.com/parrotai/>

[understand-data-analytics-framework-with-a-case-study-in-the-business-world-15bfb421028d](https://medium.com/parrotai/understand-data-analytics-framework-with-a-case-study-in-the-business-world-15bfb421028d)

DIAGNOSTIC ANALYTICS – EXAMPLE #1



<https://www.sganalytics.com/blog/change-management-analytics-adoption/>

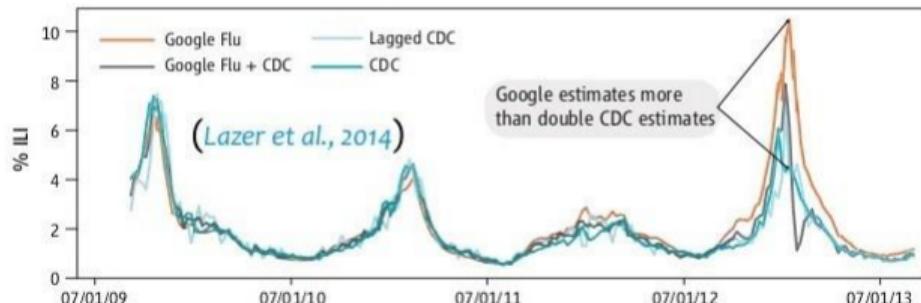
PREDICTIVE ANALYTICS – EXAMPLE #1

- Google launched **Google Flu Trends (GFT)**, to collect predictive analytics regarding the outbreaks of flu. It's a great example of seeing big data analytics in action.
- So, did Google manage to predict influenza activity in real-time by aggregating search engine queries with this big data and adopting predictive analytics?
- Even with a wealth of big data analytics on search queries, GFT **overestimated** the prevalence of flu by over 50% in 2012-2013 and 2011-2012.
- They matched the search engine terms conducted by people in different regions of the world. And, when these queries were compared with traditional flu surveillance systems, Google found that the predictive analytics of the flu season pointed towards a correlation with higher search engine traffic for certain phrases.

PREDICTIVE ANALYTICS – EXAMPLE #1

Google Flu Trends: Failure

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon \quad (\text{Ginsberg et al., 2009})$$



The estimates of the online Google Flu Trends tool were approx. **two times larger** than the ones from the CDC in 2012/13

<https://www.slideshare.net/VasileiosLampos/>

usergenerated-content-collective-and-personalised-inference-tasks

PREDICTIVE ANALYTICS – EXAMPLE # 2

- Colleen Jones applied predictive analytics to FootSmart (a niche online catalog retailer) on a content marketing product. It was called the **FootSmart Health Resource Center (FHRC)** and it consisted of articles, diagrams, quizzes and the like.
- On analyzing the data around **increased search engine visibility**, FHRC was found to help **FootSmart** reach more of the **right kind of target customers**.
- They were receiving more traffic, primarily consisting of people that cared about foot health conditions and their treatments.
- FootSmart decided to push more content at FHRC and also improve its merchandising of the product.
- The result of such informed data-driven decision making?
A 36% increase in weekly sales.

<https://www.footsmart.com/pages/health-resource-center>

PREDICTIVE ANALYTICS – EXAMPLE # 2

Predictive Policing (Self study)

- [https://www.brennancenter.org/our-work/research-reports/
predictive-policing-explained](https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained)
- <https://www.youtube.com/watch?v=YxvyeaL7NEM>

PREScriptive ANALYTICS – EXAMPLE #1

- A health insurance company analyses its data and determines that many of its diabetic patients also suffer from retinopathy.
- With this information, the provider can now use predictive analytics to get an idea of how many more ophthalmology claims it might receive during the next year.
- Then, using prescriptive analytics, the company can look at scenarios where the reimbursement costs for ophthalmology increases, decreases, or holds steady. These scenarios then allow them to make an informed decision about how to proceed in a way that's both cost-effective and beneficial to their customers.
- Analysing data on patients, treatments, appointments, surgeries, and even radiologic techniques can ensure hospitals are properly staffed, the doctors are devising tests and treatments based on probability rather than gut instinct, and the facility can save costs on everything from medical supplies to transport fees to food budgets.

PREScriptive ANALYTICS – EXAMPLE #2

- Whenever you go to Amazon, the site recommends dozens and dozens of products to you. These are based not only on your previous shopping history (reactive), but also based on what you've searched for online, what other people who've shopped for the same things have purchased, and about a million other factors (proactive).
- Amazon and other large retailers are taking deductive, diagnostic, and predictive data and then running it through a prescriptive analytics system to find products that you have a higher chance of buying.
- Every bit of data is broken down and examined with the end goal of helping the company suggest products you may not have even known you wanted.

<https://accent-technologies.com/2020/06/18/examples-of-prescriptive-analytics/>

HEALTHCARE ANALYTICS – CASE STUDY

Self study

- [https://integratedmp.com/
4-key-healthcare-analytics-sources-is-your-practice-using-them/](https://integratedmp.com/4-key-healthcare-analytics-sources-is-your-practice-using-them/)
- <https://www.youtube.com/watch?v=olpuyn6kemg>

REFERENCES

- Big Data Analytics - A Hands-on Approach by Arshdeep Bahga & Vijay Madisetti
- <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-project.html>
- <https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>
- <https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jn8bbjjm1a2.htm&docsetVersion=14.3&locale=en>
- <http://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/>
- <https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html>
- <https://medium.com/illumination-curated/big-data-lifecycle-management-629dfe16b78d>
- <https://www.esadeknowledge.com/view/7-challenges-and-opportunities-in-data-based-decision-making-193560>

THANK YOU



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 3 : DATA

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

DATA

- Data is a collection of data objects and their attributes.
- The type of data determines which tools and techniques can be used to analyze the data.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

DATA

- Data is a collection of data objects and their attributes.
- An attribute is a property or characteristic of an object.

Examples: eye color of a person, temperature

- Attribute is also known as variable, field, characteristic, or feature.
- A collection of attributes describe an object.
- Object is also known as record, point, case, sample, entity, or instance.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ATTRIBUTE / FEATURE

- An attribute is a property or characteristic of an object.
 - ▶ eye color of a person, temperature
- Attribute is also known as variable, field, characteristic, or feature.
- The values used to represent an attribute may have properties that are not properties of the attribute itself.
 - ▶ Average age of an employee may have a meaning , whereas it makes no sense to talk about the average employee ID.

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

ATTRIBUTE / FEATURE

- The type of an attribute should tell us what properties of the attribute are reflected in the values used to measure it.
 - For the age attribute, the properties of the integers used to represent age are very much the properties of the attribute. Even so, ages have a maximum while integers do not.
 - The ID attribute is distinct. The only valid operation for employee IDs is to test whether they are equal.

Attributes

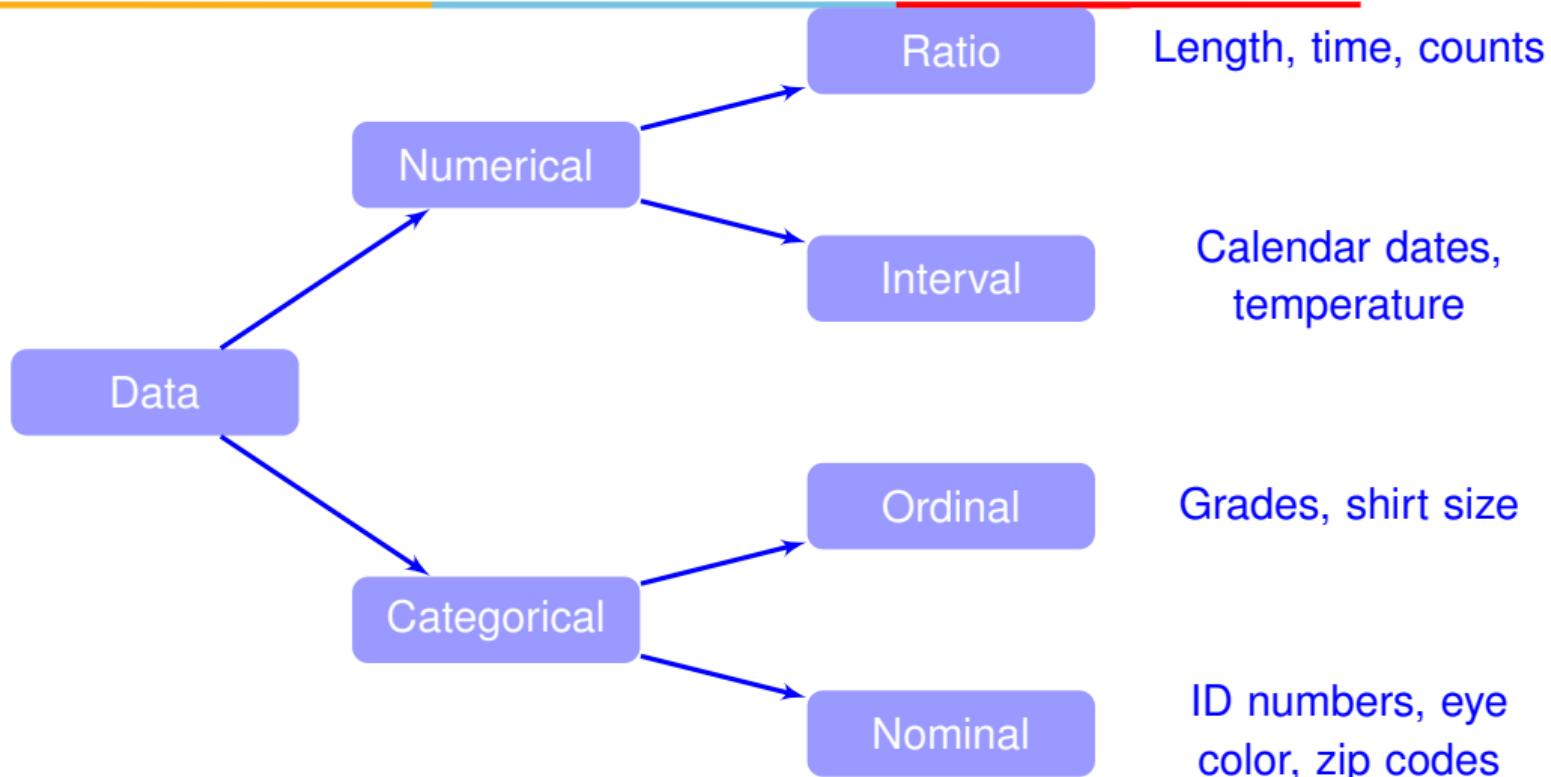
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

PROPERTIES OF ATTRIBUTES

- Specify the type of an attribute by identifying the properties of numbers that correspond to underlying properties of the attribute.
- Properties include
 - ▶ Distinctiveness $=, \neq$
 - ▶ Order $<, >, \geq, \leq$
 - ▶ Addition $+, -$
 - ▶ Multiplication $*, /$
- Based on these properties, we define four types of attributes: nominal, ordinal, interval, and ratio.
- Each attribute type possesses all of the properties and operations of the attribute types above it.

TYPES OF ATTRIBUTES



TYPES OF ATTRIBUTES

Levels of Measurement

Nominal	Ordinal	Interval	Ratio
"Eye color"	"Level of satisfaction"	"Temperature"	"Height"
Named	Named	Named	Named
	Natural order	Natural order	Natural order
		Equal interval between variables	Equal interval between variables
			Has a "true zero" value, thus ratio between values can be calculated

TYPES OF ATTRIBUTES

Attribute Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, χ^2 test
	Ordinal The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
	Ratio For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

ATTRIBUTES AND TRANSFORMATIONS

Attribute Type	Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values
	Ordinal	An order-preserving change of values, i.e., $new_value = f(old_value)$, where f is a monotonic function.
Numeric (Quantitative)	Interval	$new_value = a * old_value + b$, a and b constants.
	Ratio	$new_value = a * old_value$

ATTRIBUTES BY THE NUMBER OF VALUES

- Discrete Attribute

- ▶ only a finite or countable infinite set of values.
- ▶ zip codes, counts, or the set of words in a collection of documents
- ▶ Often represented as integer variables.
- ▶ Note: binary attributes are a special case of discrete attributes

- Continuous Attribute

- ▶ Real numbers as attribute values.
- ▶ temperature, height, or weight
- ▶ Continuous attributes are typically represented as floating-point variables.

- Asymmetric Attribute

- ▶ only presence a non-zero attribute value-is considered.
- ▶ For a specific student, an attribute has a value of 1 if the student took the course associated with that attribute and a value of 0 otherwise
- ▶ Asymmetric binary attributes.

TYPES OF ATTRIBUTES EXAMPLE

Identify the types of attributes in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good

TYPES OF ATTRIBUTES EXAMPLE

Identify the types of attributes in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good
Nominal	Ratio	Nominal	Nominal	Ratio	Ordinal

TYPES OF ATTRIBUTES EXAMPLE

Identify whether the attribute is discrete and continuous in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good

TYPES OF ATTRIBUTES EXAMPLE

Identify whether the attribute is discrete and continuous in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good
Discrete	Continuous	Discrete	Discrete	Continuous	Discrete

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

TYPES OF DATA-SETS

① Structured data

- ▶ Data containing a defined data type, format and structure.
- ▶ Example: transaction data, online analytical processing , OLAP data cubes, traditional RDBMS, CSV file and spreadsheets.

② Semi structured data

- ▶ Textual data file with discernible pattern that enables parsing
- ▶ Example: XML data file, HTML of a web page

③ Quasi structured data

- ▶ Textual data with erratic data format that can be formatted with effort, tools and time
- ▶ Example: Web click-stream data

④ Unstructured data

- ▶ Data that has no inherent structure.
- ▶ Example: text document, PDF, images and video, email

STRUCTURED DATA

RDBMS 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

SEMI-STRUCTURED DATA

JSON Data

```
JSON Object —→ {  
    "company": "mycompany",  
    "companycontacts": {  
        "phone": "123-123-1234",  
        "email": "myemail@domain.com"  
    },  
    "employees": [ ← JSON Array  
        {  
            "id": 101,  
            "name": "John",  
            "contacts": [  
                "email1@employee1.com",  
                "email2@employee1.com"  
            ]  
        },  
        {  
            "id": 102, ← Number Value  
            "name": "William",  
            "contacts": null ← Null Value  
        }  
    ]  
}
```

Annotations:

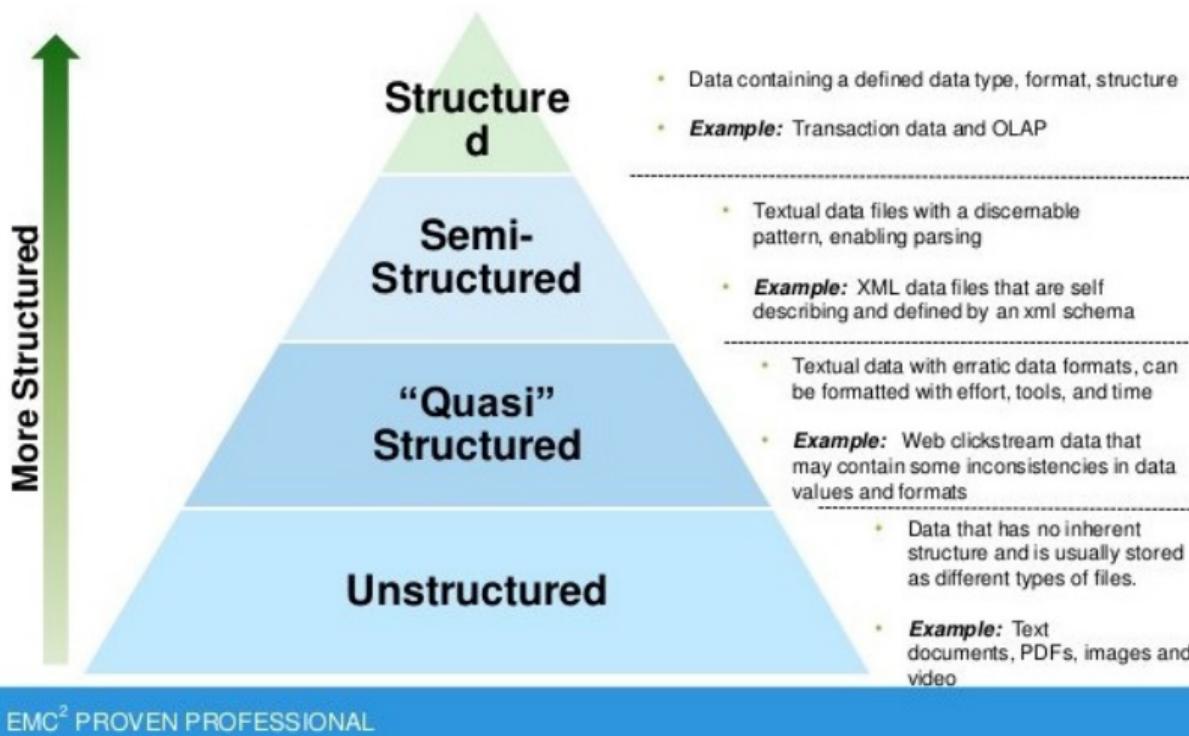
- String Value**: Points to the string value "mycompany".
- Object Inside Object**: Points to the nested object "companycontacts".
- JSON Array**: Points to the array "employees".
- Array Inside Array**: Points to the inner array "contacts" within an employee object.
- Number Value**: Points to the number value 102.
- Null Value**: Points to the null value assigned to the "contacts" field of the second employee object.

QUASI-STRUCTURED DATA

Web Click-Stream

Timestamp										IP Address		URL		
Home	/ user /	sandbox /	Omniture.0.tsv.gz											
Registered User SWID (if logged in)														
View As														
Binary														
Stop preview														
Download														
View File Location														
Refresh														
1331799426 2012-03-15 01:17:06 2860005755985467733 461168763118657821 FAS-2.8-AS3										IP Address		URL		
N 0	99.122.210.248	1	0	10	http://www.acme.com/SH55126545/VD5517036									
4 {7AAB8415-E803-3C5D-7100-E362D7F67CA7}						U en-us,en;q=0.5	516	575	1366	Y				
N Y 2 0 304	sbcglobal.net	15/2/2012 4:16:0	4 240 45	41	10002,00									
011,10020,00007 Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6														
48 0 2 3 0	homestead	usa	528 f1	0	0									
0			0											
Geocoded IP Address										0		WPLG		

TYPES OF DATA-SETS

EMC² PROVEN PROFESSIONAL

DATA-SETS

① Public data

- ▶ Data that has been collected and preprocessed for academic or research purposes and made public.
- ▶ <https://archive.ics.uci.edu/>

② Private data

- ▶ Data that is specific to an organization.
- ▶ Privacy rules like IT Act 2000 and GDPR applies.

RECORD DATA

- Record data – flat file (CSV), RDBMS
- Transaction data – set of items – banking, retail, e-commerce
- Data Matrix – record data with only numeric attributes. – SPSS data matrix
- Sparse Data Matrix – binary asymmetric data. 0/1 entries.
- Document term matrix – Frequency of terms that appears in documents

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Lead	Projection of y Lead	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

team	coach	play	bill	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

(d) Document-term matrix.

ORDERED DATA EXAMPLE

- Sequential data or temporal data – Record data + time. Eg: Money transfer transaction in Banking
- Sequence data – Positions instead of time stamp. Eg: DNA sequence bases (G, T, A, C)
- Time series data – temporal autocorrelation

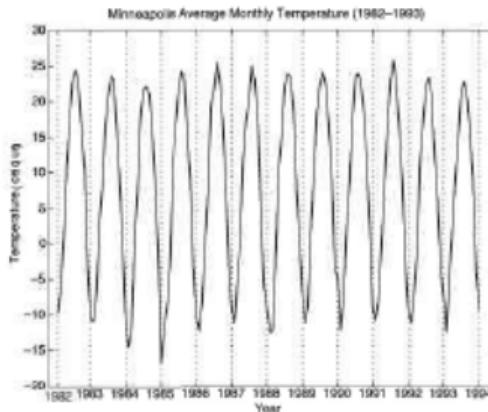
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

(a) Sequential transaction data.

GGTTCGGCCTTCAGCCCCGGCG
 CGCAGGGCCCGCCCCGGCGCGTC
 GAGAAGGGCCCGCCTGGCGGGCG
 GGGGGAGGCAGGGCCGCCGAGC
 CCAACCGAGTCCGACCAAGGTGCC
 CCTCTGCTGCCCTAGACCTGA
 GCTCATTAGGCGGCAGCGGACAG
 GCCAAGTAGAACACGCGAAGCGC
 TGGGCTGCCTGCTGCGACCAAGGG

(b) Genomic sequence data.



(c) Temperature time series.

TEXT DATA

- Text is considered as 1-D data
- Eg: Email body, PDF document, word document

AUDIO DATA

- Audio is considered as 1-D time series data
- Eg: Speech, Music

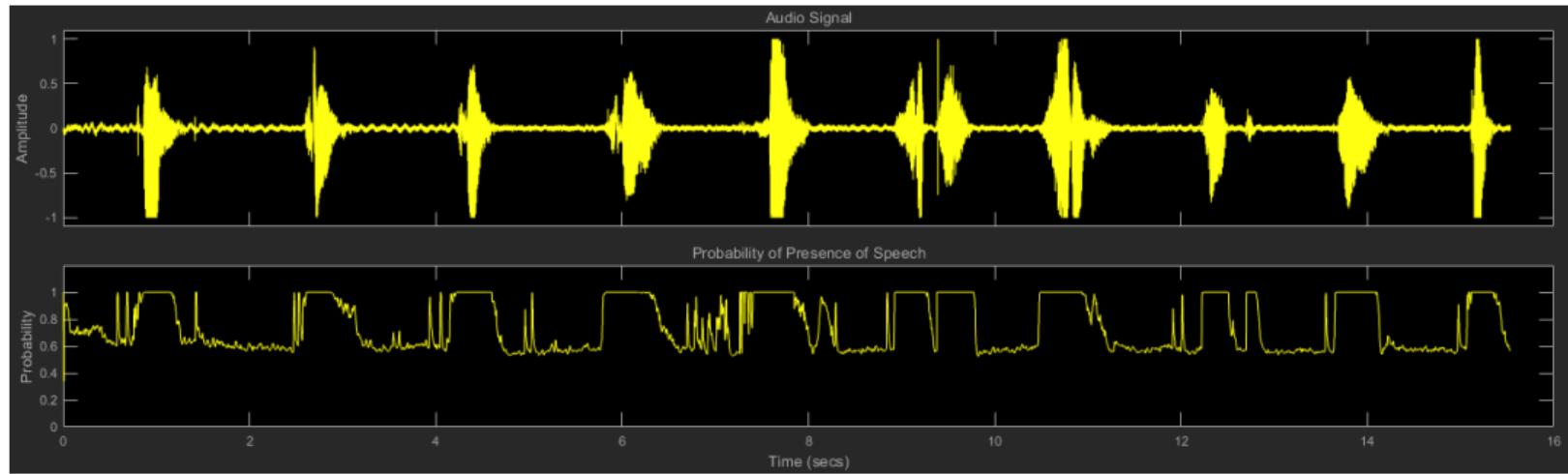


IMAGE DATA

- Images are considered as 2-D data in Euclidean space
- Digital Images are stored in a matrix or grid form where the intensity or colour information is stored in the (x,y) position.
- Black and white image – intensity is represented as 0 and 1 respectively
- Greyscale image – intensity is represented as an integer between 0 and 255. 0 represents black, grey is 125 and 255 is white.
- Colour image – contains 3 bands or channels – Red, Green and Blue – each colour is represented as an integer between 0 and 255.

DIGITAL GRayscale IMAGE

Pixel intensities = $I(x, y)$



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
3	127	243	255	155	33	226	52	2	0	10	13	232	255	255	36
6	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

<https://mozanunal.com/2019/11/img2sh/>

DIGITAL COLOUR IMAGE

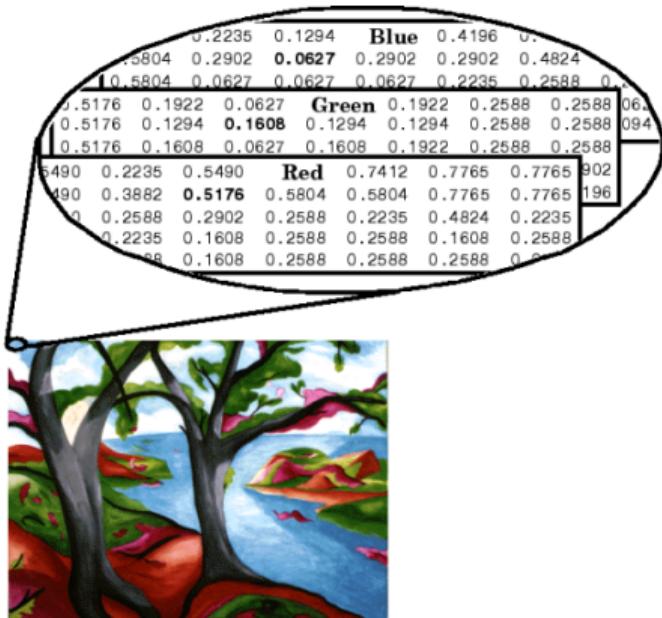


Colour Image

This image is composed of many colors and almost all colors can be generated from the three primary colors- **Red, Green, and Blue**. We can say that each colored image is composed of these three colors or 3 channels- Red, Green, and Blue-

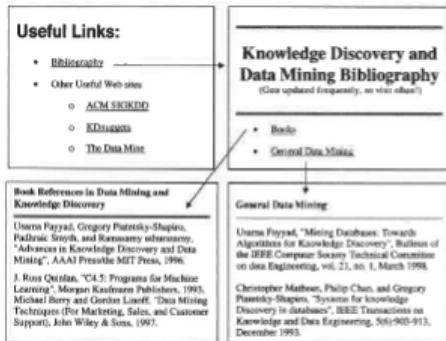


DIGITAL COLOUR IMAGE

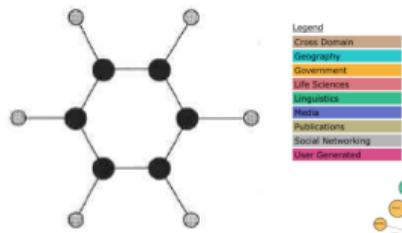


GRAPH DATA EXAMPLE

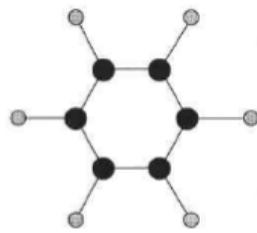
- Data with relationships among objects – Web pages
- Data with objects as graphs – chemical compound



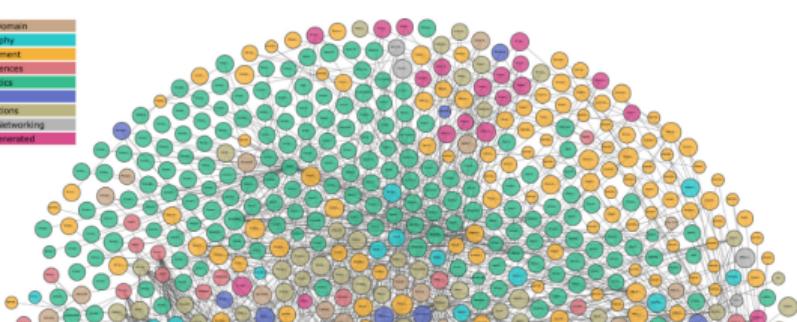
(a) Linked Web pages.



(b) Benzene molecule.



The **Linked Open Data Cloud**



<https://lod-cloud.net/>

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

DATA QUALITY INDEX

UNIQUENESS

- Existence of unique values for a specific data attribute within a table
- **Example:** Data attribute which has duplicated values will not have the highest score on uniqueness dimension

CONSISTENCY

- Logical coherence within data of a system that free them from contradiction
- **Example:** 'Order Fulfilment Date' should be after the 'Order Creation Date'

INTEGRITY

- Existence of data values in reference table(s) from different system(s)
- **Example:** 'Product ID' values should exist in the Product reference table



COMPLETENESS

- Existence of values in a specific data attribute (data field)
- **Example:** Data attribute with missing values is not complete

TIMELINESS

- Degree to which data is representative of current business conditions (updated and available)
- **Example:** A plan price change not updated on the day it was issued creates a breach of timeliness

CONFORMITY

- Data are valid if it conforms to the syntax (format, type, range) of its definition
- **Example:** 'Landline Number' should be numeric with 8 digits

DATA QUALITY ISSUES

- Noise and outliers
 - ▶ Noise is a random error or variance in a measured data object.
 - ▶ Data objects with behaviors that are very different from expectation are called outliers or anomalies.
- Inaccurate data
 - ▶ Inaccurate data – data having incorrect attribute values
 - ▶ Caused by, faulty data collection instruments, human or computer errors occurring at data entry, users may purposely submit incorrect data values, errors in data transmission
- Inconsistent data
 - ▶ inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).

DATA QUALITY ISSUES

- Missing data
 - ▶ Data that is not filled / available intentionally or otherwise.
 - ▶ Attributes of interest may not always be available, such as customer information for sales transaction data.
 - ▶ Some data were not considered important at the time of entry.
 - ▶ Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions.
- Duplicate data
- Orphaned data
- Text encoding errors
- Data that is biased

EXAMPLE: DATA QUALITY ISSUES

Find the issues in the given data.

Name	Age	Date of Birth	Course ID	CGPA
Amy	24	01-Jan-1995	CS 104	7.4
Ben	23	Dec-01-1996	CS 102	7.5
Cathy	25	01-Nov-1994		6.7
Diana	24	Oct-01-1995	CS 104	7.9
Ben	23	Dec-01-1996	CS 102	7.5
Eden	24		CS 103	87.5
Fischer		01-01-1959	CS 105	7.0

EXAMPLE: DATA QUALITY ISSUES

- Missing data – age, date of birth, course ID
- Inconsistent data – date of birth
- Duplicate data – Ben is duplicated
- Data Conformity – CGPA = 87.5

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

FORMAL DATA MODELS

- **Model is something we construct to help us understand the real world.**
- One key goal of formal modelling is to develop a precise specification of your question and how your data can be used to answer that question.
- **Formal models** allow you to identify clearly what you are trying to infer from data and what form the relationships between features of the population take.

GENERAL FRAMEWORK FOR MODELLING

- Apply the basic epicycle of analysis to the formal modelling portion of data analysis.
 - ① Setting expectations.
 - ★ Develop a primary model that represents your best sense of what provides the answer to your question. This model is chosen based on whatever information you have currently available.
 - ② Collecting Information.
 - ★ Create a set of secondary models that challenge the primary model in some way.
 - ③ Revising expectations.
 - ★ If our secondary models are successful in challenging our primary model and put the primary model's conclusions in some doubt, then we may need to adjust or modify the primary model to better reflect what we have learned from the secondary models.

STATISTICAL MODEL

- A statistical model serves two key purposes in a data analysis,
 - ▶ quantitative summary of data.
 - ▶ impose a specific structure on the population from which the data were sampled.
- **A statistic is any summary of the data.**
- The sample mean, median, the standard deviation, the maximum, the minimum, and the range are statistics.

MODELS AS EXPECTATIONS

- A statistical model must impose some structure on the data.
- **A statistical model provides a description of how the world works and how the data were generated.**
- The model is essentially an expectation of the relationships between various factors in the real world and in your dataset.
- Mimics the Population behavior, realized through Sample of data.
- A statistical model allows for some randomness in generating the data.

DATA MODEL - CASE STUDY

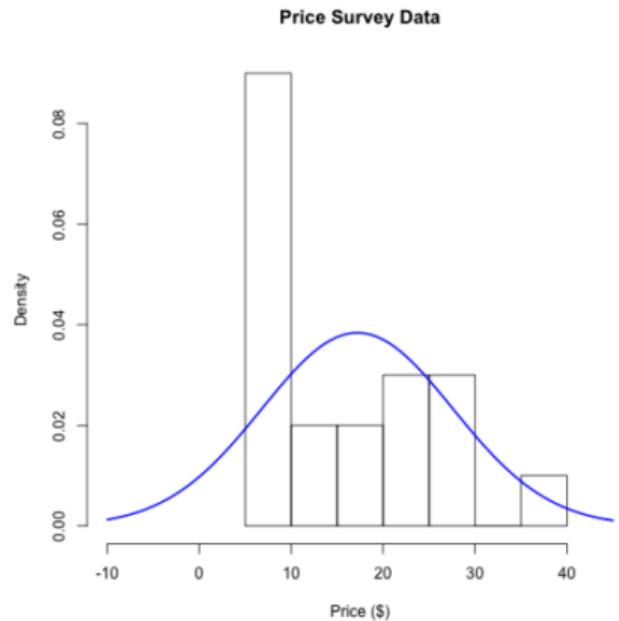
- Conduct a survey of 20 people to ask them how much they'd be willing to spend on a product you're developing.
- The survey response

25, 20, 15, 5, 30, 7, 5, 10, 12, 40, 30, 30, 10, 25, 10, 20, 10, 10, 25, 5

- What do the data say?
- Note: The example is hypothetical, generally we select higher sample size for modelling.

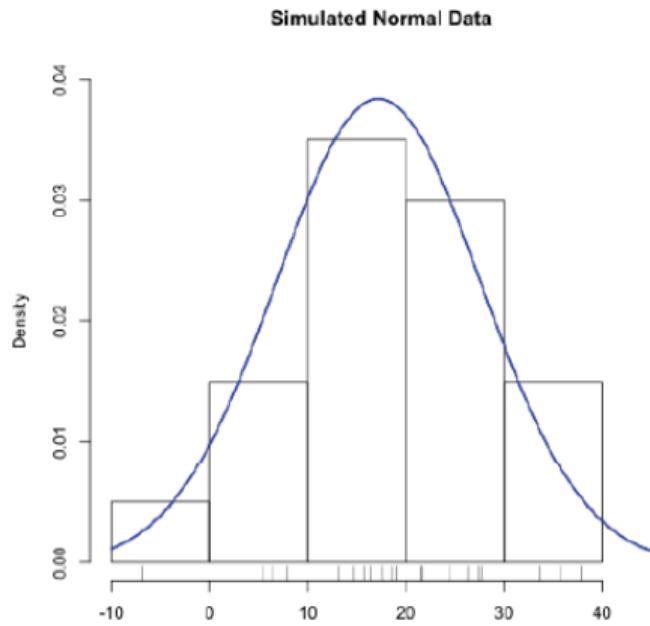
STEP 1: SETTING EXPECTATIONS

- The sample data represents the overall population likely to purchase the product.
- Mean - \$17.2 and Standard Deviation - \$10.39
- Expectation under the Normally distributed data (Normal model) is that the distribution of prices that people are willing to pay.
- According to the model, about 68% of the population would be willing to pay somewhere between \$6.81 and \$27.59 for this new product.



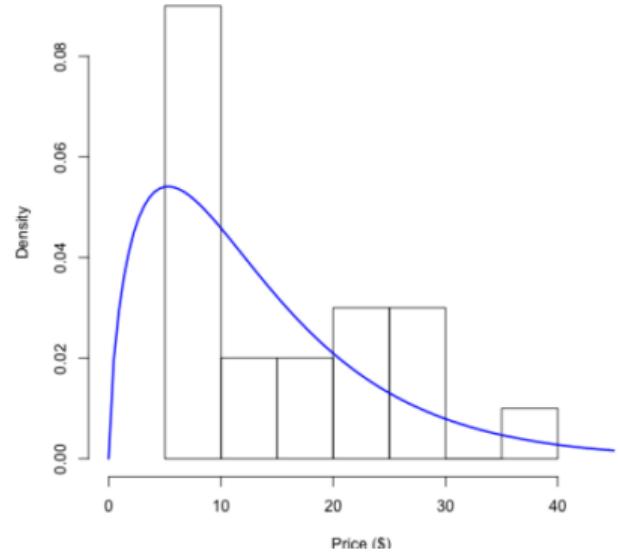
STEP 2: COMPARING MODEL EXPECTATIONS WITH REALITY

- Given the parameters, our expectation under the Normal model is that the distribution of prices that people are willing to pay looks like a bell-shaped curve.
- E.g. Normal curve on top of the histogram of the 20 data points of the amount people say they are willing to pay. The histogram has a large spike around 10.
- Normal distribution allows for negative values on the left-hand side of the plot, but there are no data points in that region of the plot.



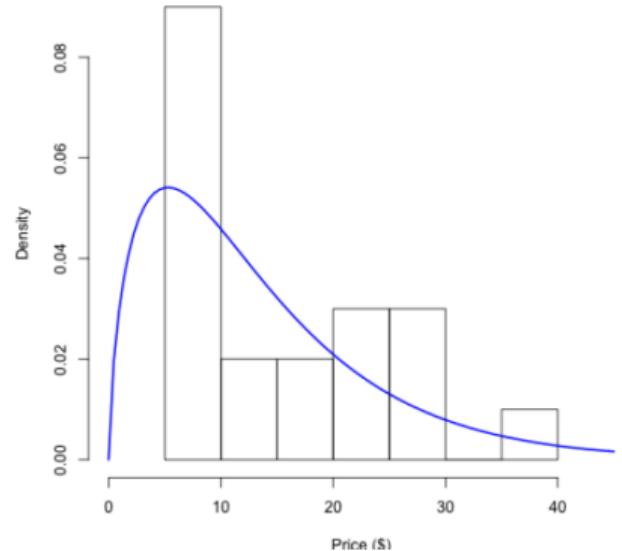
STEP 3: REFINING OUR EXPECTATIONS

- When the model and the data don't match very well.
 - Get a different model.
 - Get different data.
 - Do both.
- E.g. Choose a different statistical model to represent the population, the Gamma distribution, which has the feature that it only allows positive.



STEP 3: REFINING OUR EXPECTATIONS

- Normal vs Gamma Distribution – which model to choose?
- Qn 1: What percentage of the population is willing to pay atleast \$30 for this product?
 - ▶ Normal distribution – 11% would pay \$30 or more
 - ▶ Gamma distribution – 7% would pay \$30 or more
- Based on which model suits your problem at hand, choose the appropriate model.



DEVELOPING A BENCHMARK MODEL

- The goal is to develop a benchmark model that serves us as a baseline, upon we'll measure the performance of a better and more attuned algorithm.
- Benchmarking requires experiments to be comparable, measurable, and reproducible.

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

CLASS OR CONCEPT DESCRIPTIONS

- **Class or Concept Descriptions describe individual classes and concepts in summarized, concise, and yet precise terms.**
- Concept descriptions can be derived using
 - ① data characterization, by summarizing the data of the class under study
 - ② data discrimination, by comparison of the target class with one or a set of comparative classes
 - ③ both data characterization and discrimination.

DATA CHARACTERIZATION

- **Data characterization is a summarization of the general characteristics or features of a target class of data.**
- Methods for data characterization
 - ▶ data summaries based on statistical measures and plots
 - ▶ data cube-based OLAP roll-up operation
 - ▶ attribute-oriented induction technique
- Output of data characterization
 - ▶ bar charts, curves, multidimensional data cubes, and multidimensional tables.
 - ▶ generalized relations or in rule form called **characteristic rules**.

DATA DISCRIMINATION

- **Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.**
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
- The methods used for data discrimination are similar to those used for data characterization.
- Output presentation
 - ▶ Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

ASSOCIATION ANALYSES

- **Frequent patterns are patterns that occur frequently in data.**
- Many kinds of frequent patterns
 - ▶ A **frequent itemset** refers to a set of items that often appear together in a transactional data set. E.g: milk and bread
 - ▶ **A frequently occurring subsequence is a (frequent) sequential pattern.** Eg: customers tend to purchase first a laptop, followed by a digital camera, and then a memory card.
 - ▶ A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. **If a substructure occurs frequently, it is called a (frequent) structured pattern.**
 - ▶ Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

PREDICTION ANALYSES

- The term **prediction** refers to both **numeric prediction** and **class label prediction**.
- Classification and regression may need to be preceded by **relevance analysis**, which attempts to identify attributes that are significantly relevant to the classification and regression process.

CLASSIFICATION ANALYSES

- **Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.**
- The models are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known).
- The model is used to predict the class label of objects for which the class label is unknown.
- The derived model may be represented in as classification rules (i.e., IF-THEN rules), decision trees, mathematical formulae, or neural networks, naive Bayesian classification, support vector machines, and k-nearest-neighbor classification.
- Classification predicts categorical (discrete, unordered) labels.

REGRESSION ANALYSES

- **Regression models continuous-valued functions.**
- Regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
- Regression analysis is a statistical methodology that is most often used for numeric prediction.
- Regression also encompasses the identification of distribution trends based on the available data.

CLUSTER ANALYSIS

- Clustering analyzes data objects without consulting class labels.
- Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
- clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.
- Each cluster so formed can be viewed as a class of objects, from which rules can be derived.
- Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

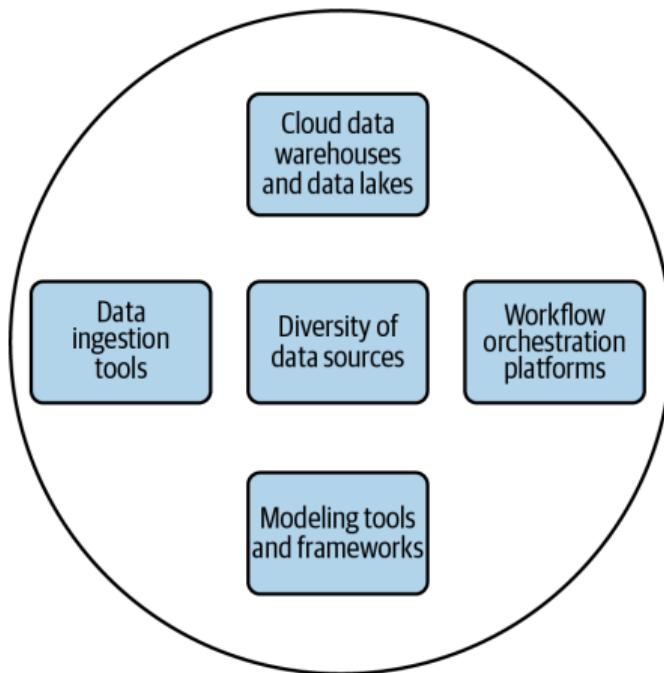
DATA PIPELINE STAGES

- **Data pipelines are sets of processes that move and transform data from various sources to a destination where new value can be derived.**
- In their simplest form, pipelines may extract only data from one source such as a REST API and load to a destination such as a SQL table in a data warehouse.
- In practice, data pipelines consist of multiple steps including data extraction, data preprocessing, data validation, and at times training or running a machine learning model before delivering data to its final destination.
- **Data engineers** specialize in building and maintaining the data pipelines.

WHY BUILD DATA PIPELINES?

- For every dashboard and insight that a data analyst generates and for each predictive model developed by a data scientist, there are data pipelines working behind the scenes.
- A single dashboard, or a single metric may be derived from data originating in multiple source systems.
- Data pipelines extract data from sources and load them into simple database tables or flat files for analysts to use. Raw data is refined along the way to clean, structure, normalize, combine, aggregate, and anonymize or secure it.

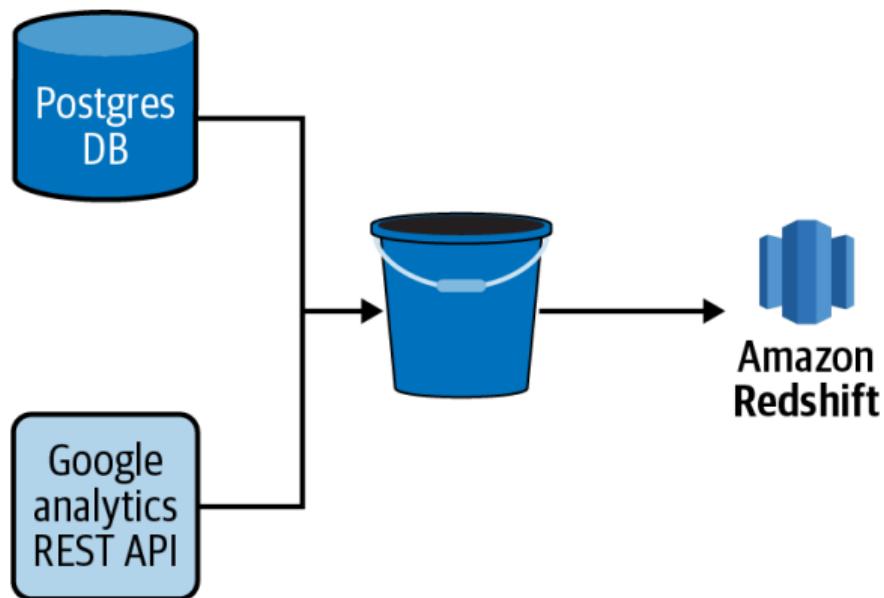
DIVERSITY OF DATA SOURCES



DATA INGESTION

- The term data ingestion refers to extracting data from one source and loading it into another.
- Ingestion Interface and data structure
 - ▶ A database behind an application, such as a Postgres or MySQL database or NoSQL database
 - ▶ JSON from a REST API
 - ▶ A stream processing platform such as Apache Kafka
 - ▶ A shared network file system or cloud storage bucket containing logs, comma-separated value (CSV) files, and other flat files
 - ▶ Semi-structured log data
 - ▶ A data warehouse or data lake
 - ▶ Data in HDFS or HBase database
- Data ingestion is traditionally both the extract and load steps of an ETL or ELT process

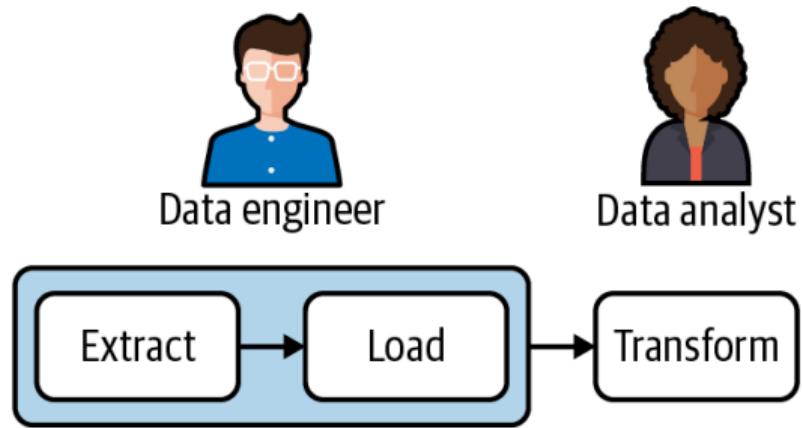
SIMPLE PIPELINE



ETL AND ELT

- E– extract step
 - ▶ gathers data from various sources in preparation for loading and transforming.
- L – load step
 - ▶ brings either the raw data (in the case of ELT) or the fully transformed data (in the case of ETL) into the final destination.
 - ▶ load data into the data warehouse, data lake, or other destination.
- T – transform step
 - ▶ raw data from each source system is combined and formatted in a such a way that it's useful to analysts, visualization tools

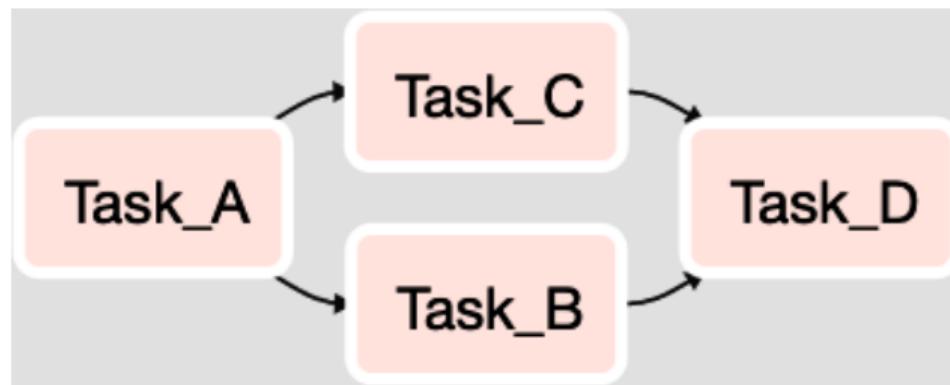
ELT PIPELINE



ORCHESTRATING PIPELINES

- Orchestration ensures that the steps in a pipeline are run in the correct order and that dependencies between steps are managed properly.
- Pipeline steps (tasks) are always directed, meaning they start with a task or multiple tasks and end with a specific task or tasks. This is required to guarantee a path of execution.
- Pipeline graphs must also be acyclic, meaning that a task cannot point back to a previously completed task.
- Pipelines are implemented as DAGs (Directed Acyclic Graphs).
- Orchestration tool – Apache Airflow

ORCHESTRATION DAG



VARIOUS DATA SOURCE

The screenshot shows the Tableau Connect interface. On the left, there's a sidebar with navigation links like 'Search for Data', 'Tableau Server', 'To a File' (with options for Microsoft Excel, Text file, JSON file, Microsoft Access, PDF file, Spatial file, Statistical file, More...), and 'To a Server' (with options for MySQL, Oracle, Amazon Redshift, IBM DB2, More...). The main area is titled 'Connect' and features a search bar at the top. Below the search bar is a grid of data source names, grouped into four columns. The sources listed are:

Column 1	Column 2	Column 3	Column 4
Action Matrix	Firebird 3	MemSQL	SAP Sybase ASE
Action Vector	Google Ads	Microsoft Analysis Services	SAP Sybase IQ
Alibaba AnalyticDB for MySQL	Google Analytics	Microsoft PowerPivot	ServiceNow ITSM
Alibaba Data Lake Analytics	Google BigQuery	Microsoft SQL Server	SharePoint Lists
Alibaba MaxCompute	Google Cloud SQL	MonetDB	Snowflake
Amazon Athena	Google Drive	MongoDB BI Connector	Spark SQL
Amazon Aurora for MySQL	Google Sheets	MySQL	Splunk
Amazon EMR Hadoop Hive	Hortonworks Hadoop Hive	OData	Teradata
Amazon Redshift	IBM BigInsights	OneDrive	Teradata OLAP Connector
Anaplan	IBM DB2	Oracle	TIBCO Data Virtualization
Apache Drill	IBM PDA (Netezza)	Oracle Eloqua	Vertica
Aster Database	Impala	Oracle Essbase	Web Data Connector
Azure Synapse Analytics	Intuit QuickBooks Online	Pivotal Greenplum Database	
Box	Kognito	PostgreSQL	Other Databases (JDBC)
Cloudera Hadoop	Kyvos	Presto	Other Databases (ODBC)
Databricks	LinkedIn Sales Navigator	Progress OpenEdge	
Denodo	MapR Hadoop Hive	Qubole Presto	
Dropbox	MariaDB	Salesforce	
Esri ArcGIS Server	Marketo	SAP HANA	
Exasol	MarkLogic	SAP NetWeaver Business Warehouse	

VARIOUS DATA SOURCE

Actian Matrix	Firebird 3
Actian Vector	Google Ads
Alibaba AnalyticDB for MySQL	Google Analytics
Alibaba Data Lake Analytics	Google BigQuery
Alibaba MaxCompute	Google Cloud SQL
Amazon Athena	Google Drive
Amazon Aurora for MySQL	Google Sheets
Amazon EMR Hadoop Hive	Hortonworks Hadoop Hive
Amazon Redshift	IBM BigInsights
Anaplan	IBM DB2
Apache Drill	IBM PDA (Netezza)
Aster Database	Impala
Azure Synapse Analytics	Intuit QuickBooks Online
Box	Kognitio
Cloudera Hadoop	Kyvos
Databricks	LinkedIn Sales Navigator
Denodo	MapR Hadoop Hive
Dropbox	MariaDB
Esri ArcGIS Server	Marketo
Exasol	MarkLogic

VARIOUS DATA SOURCE

MemSQL	SAP Sybase ASE
Microsoft Analysis Services	SAP Sybase IQ
Microsoft PowerPivot	ServiceNow ITSM
Microsoft SQL Server	SharePoint Lists
MonetDB	Snowflake
MongoDB BI Connector	Spark SQL
MySQL	Splunk
OData	Teradata
OneDrive	Teradata OLAP Connector
Oracle	TIBCO Data Virtualization
Oracle Eloqua	Vertica
Oracle Essbase	Web Data Connector
Pivotal Greenplum Database	
PostgreSQL	Other Databases (JDBC)
Presto	Other Databases (ODBC)
Progress OpenEdge	
Qubole Presto	
Salesforce	
SAP HANA	
SAP NetWeaver Business Warehouse	

TABLE OF CONTENTS

- 1 DATA
- 2 DATA-SETS
- 3 DATA QUALITY
- 4 DATA MODELS
- 5 ANALYSIS IN DATA SCIENCE
- 6 DATA PIPELINES AND PATTERNS
- 7 FURTHER READING

DATABASE DATA

- **Database management system (DBMS)**, consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.

RDBMS DATA

- **A relational database (RDBMS)** is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.

DATA WAREHOUSE

- **A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.**
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- Data in a data warehouse is structured and optimized for reporting and analysis queries.

TRANSACTIONAL DATA

- Each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
- A transaction includes a unique transaction identity number (trans ID) and a list of the items making up the transaction.
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

DATA LAKES

- **A data lake is where data is stored, but without the structure or query optimization of a data warehouse.**
- It will contain a high volume of data as well as a variety of data types.
- It is not optimized for querying such data in the interest of reporting and analysis.
- Eg: a single data lake might contain a collection of blog posts stored as text files, flat file extracts from a relational database, and JSON objects containing events generated by sensors in an industrial system.

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - The Art of Data Science by Roger D Peng and Elizabeth Matsui (R1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)
 - On Being a Data Skeptic Publisher(s): O'Reilly Media, Inc. ISBN: 9781449374310

THANK YOU




BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE
MODULE # 4
Febin.A.Vahab
IDS Course Team
BITS Pilani





The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.



TABLE OF CONTENTS

-  [STATISTICAL DESCRIPTIONS OF DATA SELF STUDY](#)

-  [DATA PREPARATION](#)

-  [DATA AGGREGATION, SAMPLING](#)

-  [DATA SIMILARITY & DISSIMILARITY MEASURE](#)

5

6

7

8

9



Statistical Descriptions of Data

- Measuring the Central Tendency
- Measuring the Dispersion of Data
- Boxplot Analysis



MEASURES OF CENTRAL TENDENCY

- Gives an idea of the central tendency of the data.
- Measures of central tendency include the mean, median, mode, and midrange.
- Let x_1, x_2, \dots, x_N be the set of N observed values or observations for numeric attribute X . Assume X is sorted in increasing order.



MEAN

- Common and effective numeric measure of the "center" of a set of data is the **(arithmetic) mean**.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

- Weighted average

-) Sometimes, each value x_i in a set may be associated with a weight w_i .
 -) The **weights** reflect the significance, importance, or occurrence frequency attached to their respective values.

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \dots + w_Nx_N}{N} = \frac{\sum_{i=1}^N w_i x_i}{N}$$

- Issue: Mean is sensitive to extreme (e.g., outlier) values.
- Issue: For **skewed (asymmetric) data**, a better measure of the center of data is the median.



BITS Pilani, Deemed to be University under Section 3 of UGC Act, 1956



MEDIAN

- If N is odd, then the median is the middle value of the ordered set.
- If N is even, then the median is not unique; it is the two middlemost values and any value in between.
- If X is a numeric attribute, the median is taken as the average of the two middlemost values.



MODE

- Mode for a set of data is the value that occurs most frequently in the set.
- Mode can be determined for qualitative and quantitative attributes.
- Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**.

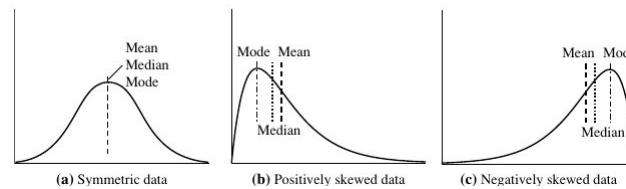


SYMMETRIC DATA AND SKEWED DATA

- In a **unimodal frequency curve with perfect symmetric data distribution**, the mean, median, and mode are all at the same center value.

$$\text{mean} - \text{mode} \approx 3(\text{mean} - \text{median})$$

- Data in most real applications are not symmetric.
 - Positively skewed - the mode occurs at a value that is smaller than the median.
 - Negatively skewed - the mode occurs at a value greater than the median.





MIDRANGE

- Average of minimum and maximum values.

$$\text{midrange} = \frac{\min + \max}{2}$$



EXAMPLE

$$X = [30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110]$$

$$\text{mean} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = 58$$

$$\text{median} = \frac{52 + 56}{2} = 54$$

$$\text{mode} = 52, 70$$

$$\text{midrange} = \frac{30 + 110}{2} = 70$$



DATA DISPERSION MEASURES

- Range
- Quartiles, and interquartile range
- Five-number summary and boxplots
- Variance and standard deviation



RANGE

- The range of the set is the difference between the largest and smallest values.

$$\text{range} = \max - \min$$



QUANTILES

- Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.
- The k^{th} q -quantile for a given data distribution is the value x such that at most k/q of the data values are less than x and at most $(q - k)/q$ of the data values are more than x , where k is an integer such that $0 < k < q$.
- There are $q - 1$ q -quantiles.



QUARTILES OR PERCENTILES

- Three data points that split the data distribution into four equal parts
- Each part represents one-fourth of the data distribution.
- Q_1 is the 25th percentile and Q_3 is the 75th percentile
- Quartiles give an indication of a distribution's center, spread, and shape



INTERQUARTILE RANGE (IQR)

- Distance between the first and third quartiles
- Measure of spread that gives the range covered by the middle half of the data.

$$IQR = Q3 - Q1$$

- Identifying outliers as values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.



FIVE-NUMBER SUMMARY

- The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3 , and the smallest and largest individual observations.
- Written in the order

Five – number Summary = [Minimum, Q1, Median, Q3, Maximum]



Exercise

Find the outlier in the following data using Inter-Quartile Range.

Data = 10, 2, 11, 15, 11, 14, 13, 17, 12, 22, 14, 11.

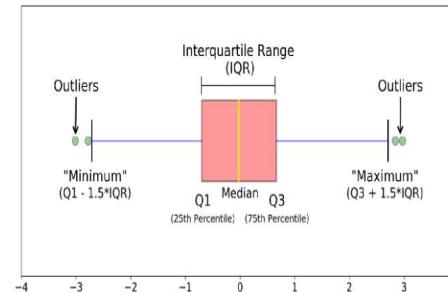
1. Sort : 10, 11, 11, 11, 12, 12, 13, 14, 14, 15, 17, 22
2. Median: $(12+13)/2=12.5=Q_2$
3. $Q_1=11$ (25^{th} percentile)
4. $Q_3=14.5$ (75^{th} percentile)
5. $IQR=Q_3-Q_1=3.5$
6. $\text{Min}=Q_1-1.5IQR=5.75$
7. $\text{Max}=Q_3+1.5IQR=19.75$

Outlier=22



MEASURING THE DISPERSION OF DATA BOXPLOT ANALYSIS

- **Five-number summary** of a distribution
 - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
 - Data is represented with a box
 - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
 - The median is marked by a line within the box
 - Whiskers: two lines outside the box extended to Minimum and Maximum
 - Outliers: points beyond a specified outlier threshold, plotted individually





VARIANCE

- Variance and standard deviation indicate how spread out a data distribution is.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2,$$



STANDARD DEVIATION

- Standard deviation σ of the observations is the square root of the variance σ^2 .
- A low standard deviation means that the data observations tend to be very close to the mean.
- A high standard deviation indicates that the data are spread out over a large range of values.
- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.



EXAMPLE

$$X = [30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110]$$

$$Q_1 = 47 \quad \text{3rd value}$$

$$Q_2 = 52 \quad \text{6th value}$$

$$Q_3 = 63 \quad \text{9th value}$$

$$IQR = 63 - 47 = 16$$

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \dots + 110^2) - 58^2 \approx 379.17$$

$$\sigma = \sqrt{379.17} \approx 19.47$$



TABLE OF CONTENTS

 [STATISTICAL DESCRIPTIONS OF DATA SELF STUDY](#)

 [DATA PREPARATION](#)

 [DATA AGGREGATION, SAMPLING](#)

 [DATA SIMILARITY & DISSIMILARITY MEASURE](#)

5

6

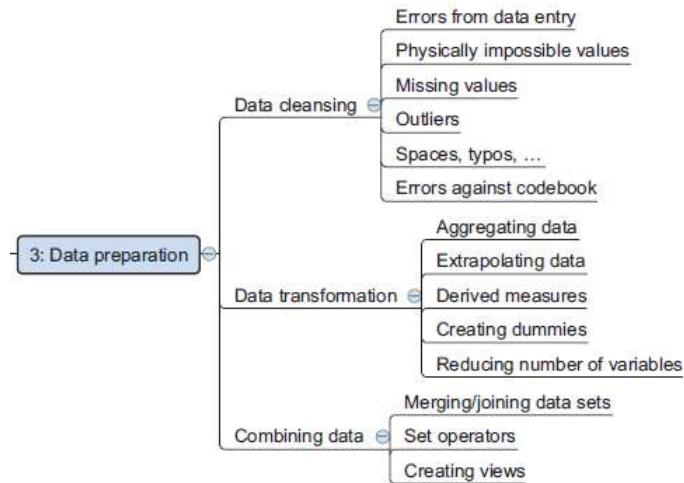
7

8

9



DATA PREPARATION





DATA CLEANSING

- Focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.
- Two types of errors
 -) Interpretation error
 - Age < 100
 - Height of a person is less than 7feet.
 - Price is positive.
 -) Inconsistencies between data sources or against your company's standardized values.
 - Female and F
 - Feet and meter
 - Dollars and Pounds



DATA CLEANSING

Error description	Possible solution
<i>Errors pointing to false values within one data set</i>	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
<i>Errors pointing to inconsistencies between data sets</i>	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation



DATA CLEANSING

- Errors from data entry
 -) Cause
 - Typos
 - Errors due to lack of concentration
 - Machine or hardware failure
 -) Detection
 - Frequency table
 -) Correction
 - Simple assignment statements
 - If-then-else rules
- White-spaces and typos
 -) Remove leading and trailing white-spaces.
 -) Change case of the alphabets from upper to lower.



DATA CLEANSING

- Physically impossible values
 -) Examples
 - 2 Age < 100
 - 2 Height of a person is less than 7feet.
 - 2 Price is positive.
 -) If-then-else rules
- Outliers
 -) Use visualization techniques like box plots or scatter plots.
 -) Use statistical summary with minimum and maximum values.
 -) Identifying outliers as values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.



DATA CLEANSING

■ Missing values

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to null	Easy to perform	Not every modeling technique and/or implementation can handle null values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute



MISSING VALUES

- Ignore the tuple.
 -) Used when the class label is missing in a classification task.
 -) Not very effective, unless the tuple contains several attributes with missing values.
 -) Poor technique when the percentage of missing values per attribute varies considerably.
 -) By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand.
- Fill in the missing value manually.
 -) Time consuming.
 -) May not be feasible given a large data set with many missing values.



MISSING VALUES

- Use a global constant to fill in the missing value.
 -) Replace all missing attribute values by the same constant such as a label like "Unknown" or -1.
 -) If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.
- Use a measure of central tendency for the attribute.
 -) Central tendency indicates the "middle" value of a data distribution. E.g., mean or median
 -) For normal (symmetric) data distributions, the mean can be used.
 -) Skewed data distribution should employ the median.



MISSING VALUES

- Use the attribute mean or median for all samples belonging to the same class as the given tuple.
 -) For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple.
 -) If the data distribution for a given class is skewed, the median value is a better choice.
- Use the most probable value to fill in the missing value.
 -) This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.
 -) For example, using the other customer attributes in the data set, we may construct a decision tree to predict the missing values for income.
 -) Most popular strategy.



DATA CLEANSING

- Deviations from code-book

-) A code book is a description of your data. It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means.
 -) Discrepancies between the code-book and the data should be corrected.

- Different units of measurement

-) Pay attention to the respective units of measurement.
 -) Simple conversion can rectify.

- Different levels of aggregation

-) Data set containing data per week versus one containing data per work week.
 -) Data summarization will fix it.



NOISY DATA

- Noise is a random error or variance in a measured variable. Outliers may represent noise.
- Noisy data can be removed by using smoothing techniques.
 -) Binning
 - Smoothing by bin means
 - Smoothing by bin medians
 -) Regression
 -) Outlier Analysis
 -) Concept hierarchies are a form of data discretization that can also be used for data smoothing.
 - 2 For example: A concept hierarchy for price may map real price values into three categories: inexpensive, moderately priced, and expensive.



4	8	15	21	21	24	25	28	34
Binning freq = 3								
① <u>smoothing by bin mean</u> <u>bin median</u> <u>bin boundaries</u>								
Bin 1 : 4, 8, 15 \Rightarrow	9	9	9	8	8	8	4, 4, 15	
Bin 2 : 21, 21, 24 \Rightarrow	22	22	22	21	21	21	21, 21, 24	
Bin 3 : - 25, 28, 34 \Rightarrow	29	29	29	28	28	28	25, 25, 34	



TABLE OF CONTENTS

 [STATISTICAL DESCRIPTIONS OF DATA SELF STUDY](#)

 [DATA PREPARATION](#)

 [DATA AGGREGATION, SAMPLING](#)

 [DATA SIMILARITY & DISSIMILARITY MEASURE](#)

5

6

7

8

9



COMBINING DATA

- Two operations to combine information from different data sets.

-) Joining

- Enriching an observation from one table with information from another table.

- Requires primary keys or candidate keys.

- Use views to virtually combine data.

-) Appending or stacking

- Adding the observations of one table to those of another table.

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Region
John Doe	NY
Jackie Qi	NC

DATA
JOIN

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

DATA
APPEND



DATA AGGREGATION

- Aggregation is combining two or more objects into a single object.
 -) Consider a data set consisting of transactions (data objects) recording the daily sales of products in various store locations (Minneapolis, Chicago, Paris, ...) for different days over the course of a year.
 -) One way to aggregate transactions for this data set is to replace all the transactions of a single store with a single store-wide transaction.
 -) This reduces
 - The hundreds or thousands of transactions that occur daily at a specific store to a single daily transaction.
 - The number of data objects is reduced to the number of stores.



DATA AGGREGATION

- To create the aggregate transaction that represents the sales of a single store or date.
- Quantitative attributes, such as price, are typically aggregated by taking a sum or an average.
- Qualitative attribute, such as item description, can either be omitted or summarized as the set of all the items that were sold at that location.
- The data in the table can also be viewed as a multidimensional array, where each attribute is a dimension.
 -) Aggregation is the process of eliminating attributes (such as the type of item) or reducing the number of values for a particular attribute (e.g., reducing the possible values for date from 365 days to 12 months).
 -) Commonly used in Online Analytical Processing (OLAP).



DATA AGGREGATION

■ Advantages

-) Require less memory and processing time.
-) Provides a high-level view of the data instead of a low-level view.
-) the behavior of groups of objects or attributes is often more stable than that of individual objects or attributes.

■ Disadvantage

-) potential loss of interesting details.

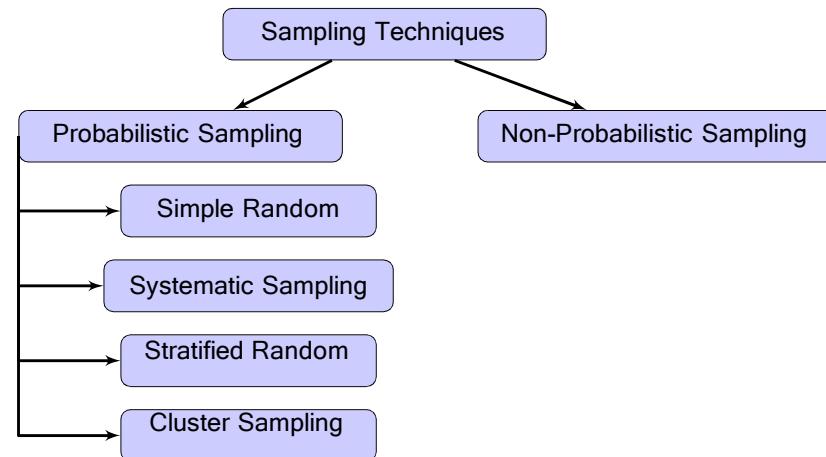


DATA SAMPLING

- A process by which representative samples are selected from a well defined population is known as sampling.
- Sampling is a technique used for selecting a subset of the data objects to be analyzed.
- The motivations for sampling in statistics and data mining are often different.
 -) Statisticians use sampling because obtaining the entire set of data of interest is too expensive or time consuming.
 -) Data miners sample because it is too expensive or time consuming to process all the data.
- In some cases, using a sampling algorithm can reduce the data size to the point where a better, but more expensive algorithm can be used.



SAMPLING TECHNIQUES





PROBABILISTIC SAMPLING TECHNIQUES

PROBABILISTIC SAMPLING means that every item in the population has an equal chance of being included in sample.

SIMPLE RANDOM SAMPLING means that every case of the population has an equal probability of inclusion in sample.

- Eg: Randomly picking mango from a basket of fruits.
- Sampling without replacement
- Sampling with replacement

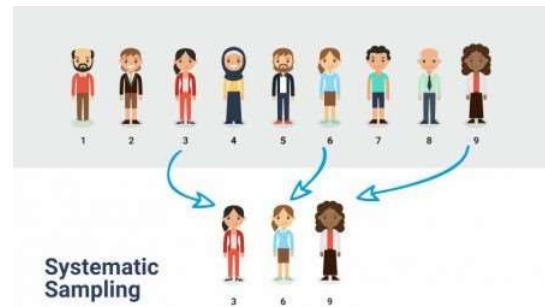




PROBABILISTIC SAMPLING TECHNIQUES

SYSTEMATIC SAMPLING is where every nth case after a random start is selected.

Eg: Picking every 5th fruit from a basket of fruits.





PROBABILISTIC SAMPLING TECHNIQUES

STRATIFIED SAMPLING is where the population is divided into strata and a random sample is taken from each strata.

Eg: One mango, one orange, one banana from a basket of fruits.

Two versions of stratified sampling:

- Equal numbers of objects are drawn from each group even though the groups are of different sizes.
- The number of objects drawn from each group is proportional to the size of that group.

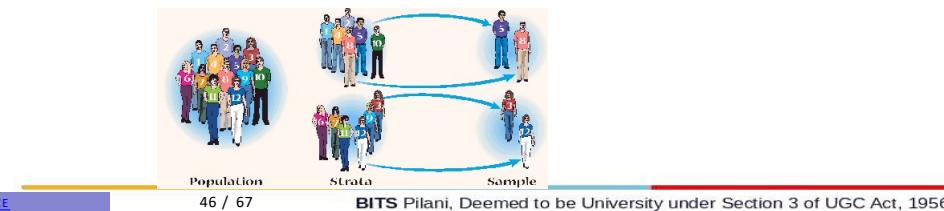




TABLE OF CONTENTS

 [STATISTICAL DESCRIPTIONS OF DATA SELF STUDY](#)

 [DATA PREPARATION](#)

 [DATA AGGREGATION, SAMPLING](#)

 [DATA SIMILARITY & DISSIMILARITY MEASURE](#)

5

6

7

8

9



MEASURES OF PROXIMITY

- Similarity and dissimilarity measures are measures of proximity.
- A **similarity measure** for two objects, i and j, will typically return the value 1 if they are identical and 0 if the objects are unalike.
- The higher the similarity value, the greater the similarity between objects.
- A **dissimilarity measure** returns a value of 0 if the objects are the same.
- The higher the dissimilarity value, the more dissimilar the two objects are.



MEASURING DATA SIMILARITY AND DISSIMILARITY

- **Similarity**
 - Numerical measure of how alike two data objects are
 - Value is higher when objects are more alike
 - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity



MEASURING DATA SIMILARITY AND DISSIMILARITY

Various proximity measures

- Data Matrix versus Dissimilarity Matrix
- Proximity Measures for Nominal Attributes
- Proximity Measures for Ordinal Attributes
- Proximity Measures for Numeric Data
- Proximity Measures for Binary Attributes
- Symmetric Binary Attributes
- Asymmetric Binary Attributes
- Proximity Measures for Mixed Types
- Cosine Similarity



DATA MATRIX AND DISSIMILARITY MATRIX

Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} \theta & & & \\ d(2,1) & \theta & & \\ d(3,1) & d(3,2) & \theta & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots & \theta \end{bmatrix}$$

PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES



Categorical Attribute

Attribute '*Color*' can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

- **Simple matching**

- m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Similarity can be computed as:

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}$$



PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES-NOMINAL

- The similarity matrix considering **only color attribute** is shown below
- The dissimilarity matrix is as shown below:
 - 0 = identical; 1 = dissimilar
- m: # of matches, p: total # of attributes

	m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$
d(2,1)	0	$\frac{1-0}{1} = 1$						
d(3,1)	0	$\frac{1-0}{1} = 1$	d(3,2)	0	$\frac{1-0}{1} = 1$			
d(4,1)	1	$\frac{1-1}{1} = 0$	d(4,2)	0	$\frac{1-0}{1} = 1$	d(4,3)	0	$\frac{1-0}{1} = 1$

Object	Color	Position	Size
1	R	L	L
2	B	C	M
3	G	C	M
4	R	L	H

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



PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES-NOMINAL

- Dissimilarity matrix considering both the categorical attributes (i.e. color and position).
- The dissimilarity matrix is as shown below:
- 0 = identical; 1 = dissimilar
- m : # of matches, p : total # of attributes

	m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$
$d(2,1)$	0	$\frac{2-0}{2} = 1$												
$d(3,1)$	0	$\frac{2-0}{2} = 1$	$d(3,2)$	1	$\frac{2-1}{2} = 0.5$									
$d(4,1)$	2	$\frac{2-2}{2} = 0$	$d(4,2)$	0	$\frac{2-0}{2} = 1$	$d(4,3)$	0	$\frac{2-0}{2} = 1$						
$d(5,1)$	0	$\frac{2-0}{2} = 1$	$d(5,2)$	0	$\frac{2-0}{2} = 1$	$d(5,3)$	1	$\frac{2-1}{2} = 0.5$	$d(5,4)$	0	$\frac{2-0}{2} = 1$			
$d(6,1)$	0	$\frac{2-0}{2} = 1$	$d(6,2)$	0	$\frac{2-0}{2} = 1$	$d(6,3)$	1	$\frac{2-1}{2} = 0.5$	$d(6,4)$	0	$\frac{2-0}{2} = 1$	$d(6,5)$	2	$\frac{2-2}{2} = 0$

Object	Color	Position	Size
1	R	L	L
2	B	C	M
3	G	C	M
4	R	L	H
5	G	R	L
6	G	R	H

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & & & & \\ 2 & 1 & 0 & & & \\ 3 & 1 & 0.5 & 0 & & \\ 4 & 0 & 1 & 1 & 0 & \\ 5 & 1 & 1 & 0.5 & 1 & 0 \\ 6 & 1 & 1 & 0.5 & 1 & 0 \end{bmatrix}$$

PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES- ORDINAL



- Values of ordinal attribute follow a sequence (ordering)
 - E.g. Grade = {A, B, C}, where A > B > C
 - Size attribute: Small < Medium < Large < X-Large

Suppose,

- f is an ordinal attribute and the set of values of $f = \{a_1, a_2, \dots, a_M\}$
- M = the number of possible states that the ordinal attribute can have
- Let M values of f are ordered in ascending order as $a_1 < a_2 < \dots < a_M$
- Let i th attribute value a_i be ranked as i, $i=1,2,\dots, M_f$

PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES- ORDINAL



- **Step 1)**

- The value of f for the i th object is x_{if} , and f has M_f ordered states, representing the ranking $1, \dots, M_f$
 - Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \dots, M_f\}$

- **Step 2)**

- Since each ordinal attribute can have a different number of states, it is often necessary to map the range of each attribute onto $[0.0, 1.0]$ so that each attribute has equal weight
 - We perform such data normalization by replacing the rank r_{if} of the i th object in the f th attribute by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- **Step 3)**

- Dissimilarity can then be computed using any of the distance measures used for numeric attributes, using z_{if} to represent the f value for the i th object

PROXIMITY MEASURE FOR CATEGORICAL ATTRIBUTES- ORDINAL



- Here, the attribute **Test** has three states:
 - Fair, Good and Excellent, so $M_f = 3$
- Step 1:
 - The four attribute values are assigned the ranks 3, 1, 2, and 3 respectively
- Step 2:
 - Normalizes the ranking by mapping ranks to [0.0, 1.0]
 - $\text{Norm}(\text{Rank } 3) = \frac{3-1}{3-1} = 1.0$; $\text{Norm}(\text{Rank } 2) = \frac{2-1}{3-1} = 0.5$; $\text{Norm}(\text{Rank } 1) = \frac{1-1}{3-1} = 0.0$
- Step 3:
 - Using Euclidean distance, a dissimilarity matrix is obtained as shown
- Therefore, students 1 and 2 are most dissimilar, as are students 2 and 4

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Student	Test	Rank	Norm
1	Excellent	3	1.0
2	Fair	1	0.0
3	Good	2	0.5
4	Excellent	3	1.0

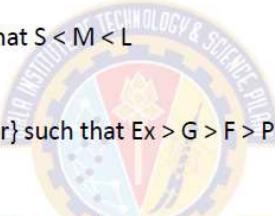
$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 0 & & \\ 3 & 1.0 & 0 & \\ 4 & 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$



PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES- ORDINAL

- Consider the following set of records, where each record is defined by two ordinal attributes
 - Size={Small, Medium, Large} such that S < M < L
 - Let's say
 - S = 1, M = 2, L = 3
 - Quality = {Excellent, Good, Fair, Poor} such that Ex > G > F > P
 - Let's say
 - Ex = 4, G = 3, F = 2, & P = 1
- Normalized values of Size {Small, Med, Large} = {0.0, 0.5, 1.0}:
 - High = $\frac{3-1}{3-1} = 1$; Low = $\frac{1-1}{3-1} = 0$; Medium = $\frac{2-1}{3-1} = 0.5$;
- Normalized values of {Ex, G, F, P} = {1.0, 0.67, 0.33, 0.0}:
 - Ex = $\frac{4-1}{4-1} = 1.0$; G = $\frac{3-1}{4-1} = 0.67$; F = $\frac{2-1}{4-1} = 0.33$; P = $\frac{1-1}{4-1} = 0.0$

Object	Size	Quality
A	S (0.0)	G (0.67)
B	L (1.0)	Ex (1.0)
C	L (1.0)	P (0.0)
D	M (0.5)	F (0.33)



PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES-ORDINAL



Object	Size	Quality
A	S (0.0)	G (0.67)
B	L (1.0)	Ex (1.0)
C	L (1.0)	P (0.0)
D	M (0.5)	F (0.33)

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1.053 & 0 & \\ 3 & 1.204 & 1.0 & 0 \\ 4 & 0.605 & 0.836 & 0.599 \\ & & & 0 \end{bmatrix}$$

	<i>Distance Measure</i>		<i>Distance Measure</i>
$d(2,1)$	$\sqrt{(1.0 - 0.0)^2 + (1.0 - 0.67)^2} = \sqrt{1.0 + 0.1089}$ $= \sqrt{1.1089} = 1.053$	$d(3,1)$	$\sqrt{(1.0 - 0.0)^2 + (0.0 - 0.67)^2} = \sqrt{1.0 + 0.4489}$ $= \sqrt{1.4489} = 1.204$
$d(3,2)$	$\sqrt{(1.0 - 1.0)^2 + (0.0 - 1.0)^2} = \sqrt{0.0 + 1.0} = \sqrt{1.0}$ $= 1.0$	$d(4,1)$	$\sqrt{(0.5 - 0.0)^2 + (0.33 - 0.67)^2} = \sqrt{0.25 + 0.1156}$ $= \sqrt{0.3656} = 0.605$
$d(4,2)$	$\sqrt{(0.5 - 1.0)^2 + (0.33 - 1.0)^2} = \sqrt{0.25 + 0.4489}$ $= \sqrt{0.6989} = 0.836$	$d(4,3)$	$\sqrt{(1.0 - 0.5)^2 + (0.33 - 0.0)^2} = \sqrt{0.25 + 0.1089}$ $= \sqrt{0.3589} = 0.599$



Exercise-CANVAS Discussion

Object	Color	Position	Size	Rank	Value
1	R	L	L	1	0.0
2	B	C	M	2	0.5
3	G	C	M	2	0.5
4	R	L	H	3	1.0
5	G	R	L	1	0.0
6	G	R	H	3	1.0

Calculate the dissimilarity matrix for the ordinal attributes



Exercise-CANVAS Discussion

Ordinal Attribute

- Dissimilarity matrix considering the ordinal attribute (Size)
 - 0 = identical; 1 = dissimilar
- Here, the attribute Size has three states:
 - Low, Med, & High, So $M_f = 3$
- Let's say
 - Low = 1, Med = 2, and High = 3 (any arbitrary values can be used)
- Normalized values of {Low, Med, High} = {0.0, 0.5, 1.0}:
 - For High $\frac{3-1}{3-1} = 1$; for Low $\frac{1-1}{3-1} = 0$; for Medium $\frac{2-1}{3-1} = 0.5$;

	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $
d(2,1)	$ 0.5 - 0.0 = 0.5$				
d(3,1)	$ 0.5 - 0.0 = 0.5$	$d(3,2) 0.5 - 0.5 = 0.0$			
d(4,1)	$ 1.0 - 0.0 = 1.0$	$d(4,2) 1.0 - 0.5 = 0.5$	$d(4,3) 1.0 - 0.5 = 0.5$		
d(5,1)	$ 0.0 - 0.0 = 0.0$	$d(5,2) 0.0 - 0.5 = 0.5$	$d(5,3) 0.0 - 0.5 = 0.5$	$d(5,4) 0.0 - 1.0 = 1.0$	
d(6,1)	$ 1.0 - 0.0 = 1.0$	$d(6,2) 1.0 - 0.5 = 0.5$	$d(6,3) 1.0 - 0.5 = 0.5$	$d(6,4) 0.0 - 0.0 = 0.0$	$d(6,5) 1.0 - 0.0 = 1.0$

Object	Color	Position	Size	Rank	Value
1	R	L	L	1	0.0
2	B	C	M	2	0.5
3	G	C	M	2	0.5
4	R	L	H	3	1.0
5	G	R	L	1	0.0
6	G	R	H	3	1.0

$$d = \begin{bmatrix} 1 & 0 & & & & \\ 2 & 0.5 & 0 & & & \\ 3 & 0.5 & 0 & 0 & & \\ 4 & 1 & 0.5 & 0.5 & 0 & \\ 5 & 0 & 0.5 & 0.5 & 1 & 0 \\ 6 & 1 & 0.5 & 0.5 & 0 & 1 \end{bmatrix}$$

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$



Exercise-CANVAS Discussion

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Calculate the dissimilarity matrix and similarity matrix for the ordinal attributes



Exercise-CANVAS Discussion

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Calculate the dissimilarity matrix and similarity matrix for the ordinal attributes

- There are three states for *test-2*: *fair*, *good*, and *excellent*, that is, $n= 3$.
- **Step 1:** we replace each value for *test-2* by its rank, the four objects are assigned the ranks 3, 1, 2, and 3, respectively.
- **Step 2 :** normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0.
- **Step 3:**use the Euclidean distance to calculate the dissimilarity

Dissimilarity Matrix

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

Similarity Matrix, $s=1-d$

	1	2	3	4
1	1			
2	0	1		
3	0.5	0.5	0	
4	1	0	0.5	1

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:



Dissimilarity of Numeric Data: Euclidean & Manhattan Dist.

- **Euclidean Distance** (i.e., straight line or "as the crow flies")

– The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **Manhattan (or city block) distance**

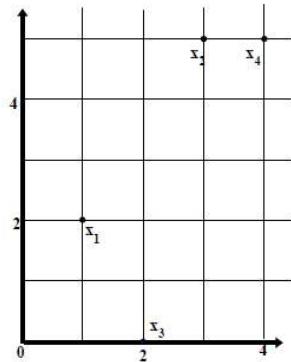
– It is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks)

– The Manhattan distance between objects i and j is defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Euclidean Distance



point	attribute1	attribute2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

$$\begin{aligned} & \text{Euclidean distance between } x_1 \text{ & } x_2 \\ &= \sqrt{(3 - 1)^2 + (5 - 2)^2} \\ &= \sqrt{4 + 9} = \sqrt{13} = 3.61 \end{aligned}$$

Dissimilarity Matrix
(with Euclidean Distance)

	x_1	x_2	x_3	x_4
x_1	0			
x_2		3.61		
x_3			5.1	0
x_4			4.24	0

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:



Manhattan Distance

Manhattan distance between x_1 & x_2

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

$$d(x_2, x_1) = |3 - 1| + |5 - 2| = 2 + 3 = 5$$

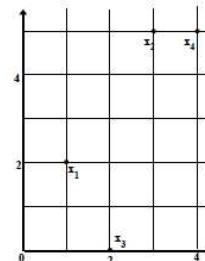
$$d(x_3, x_1) = |2 - 1| + |0 - 2| = 1 + 2 = 3$$

$$d(x_3, x_2) = |2 - 3| + |0 - 5| = 1 + 5 = 6$$

Dissimilarity Matrix

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:



Dissimilarity of Numeric Data: Minkowski Distance

- Minkowski distance is a generalization of the Euclidean and Manhattan distances
- It is defined as

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

- Where
 - $i = x_{i1}, x_{i2}, \dots, x_{ip}$ and $j = x_{j1}, x_{j2}, \dots, x_{jp}$ are two objects described by p numeric attributes
 - h is a real number such that $h \geq 1$
 - It is also called as L_h norm
 - When $h = 1$, it represents the Manhattan distance (i.e., L_1 norm)
 - When $h = 2$, it represents the Euclidean distance (i.e., L_2 norm)

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:



Dissimilarity of Numeric Data: Supremum Distance

- The **supremum distance** is a generalization of the Minkowski distance for $h = \infty$
 - also referred to as L_{\max} , L_∞ norm, and as the **Chebyshev distance**
- To compute it, we find the attribute f that gives the maximum difference in values between the two objects
- This difference is the supremum distance, defined more formally as:

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}| \right)^{1/h} = \max_f |x_{if} - x_{jf}|$$

– The L^∞ norm is also known as the *uniform norm*



PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Supremum

Supremum Distance between x_1 & x_2

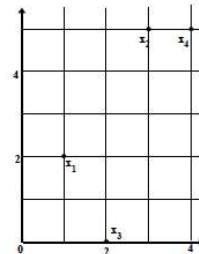
$$d(i,j) = \max_f |x_{if} - x_{jf}|$$

$$\begin{aligned} d(x_2, x_1) &= \text{Max}(|3 - 1|, |5 - 2|) = 3 \\ d(x_3, x_1) &= \text{Max}(|2 - 1|, |0 - 2|) = 2 \\ d(x_3, x_2) &= \text{Max}(|2 - 3|, |0 - 5|) = 5 \end{aligned}$$

Dissimilarity Matrix

L ₀	x ₁	x ₂	x ₃	x ₄
x ₁	0			
x ₂		0		
x ₃			0	
x ₄				0

point	attribute1	attribute2
x ₁	1	2
x ₂	3	5
x ₃	2	0
x ₄	4	5





-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - The Art of Data Science by Roger D Peng and Elizabeth Matsui (R1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)
 - On Being a Data Skeptic Publisher(s): O'Reilly Media, Inc. ISBN: 9781449374310

THANK YOU



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 4 : DATA WRANGLING(CONTD...)

IDS Course Team
BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

1 DATA SIMILARITY & DISSIMILARITY MEASURE

2 3 VISUALIZATION TECHNIQUES FOR DATA EXPLORATORY ANALYSIS

3 HANDLING NUMERIC DATA

4 MANAGING CATEGORICAL ATTRIBUTES DEALING WITH TEXTUAL DATA

5

6

7

MEASURES OF PROXIMITY

- Similarity and dissimilarity measures are measures of proximity.
- A **similarity measure** for two objects, i and j, will typically return the value 1 if they are identical and 0 if the objects are unlike.
- The higher the similarity value, the greater the similarity between objects.
- A **dissimilarity measure** returns a value of 0 if the objects are the same.
- The higher the dissimilarity value, the more dissimilar the two objects are.

MEASURING DATA SIMILARITY AND DISSIMILARITY

Various proximity measures

- Data Matrix versus Dissimilarity Matrix
- Proximity Measures for Nominal Attributes
- Proximity Measures for Binary Attributes
- Symmetric Binary Attributes
- Asymmetric Binary Attributes
- Proximity Measures for Ordinal Attributes
- Proximity Measures for Numeric Data
- Proximity Measures for Mixed Types
- Cosine Similarity

DATA MATRIX AND DISSIMILARITY MATRIX

Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} 0 \\ d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Categorical Attribute

Attribute '*Color*' can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

- **Simple matching**

- m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Similarity can be computed as:

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}$$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Categorical Attribute

- The similarity matrix considering **only color attribute** is shown below
- The dissimilarity matrix is as shown below:
 - 0 = identical; 1 = dissimilar
- m: # of matches, p: total # of attributes



	m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$
d(2,1)	0	$\frac{1-0}{1} = 1$						
d(3,1)	0	$\frac{1-0}{1} = 1$	d(3,2)	0	$\frac{1-0}{1} = 1$			
d(4,1)	1	$\frac{1-1}{1} = 0$	d(4,2)	0	$\frac{1-0}{1} = 1$	d(4,3)	0	$\frac{1-0}{1} = 1$

Object	Color	Position	Size
1	R	L	L
2	B	C	M
3	G	C	M
4	R	L	H

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

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Categorical Attribute

- Dissimilarity matrix considering both the categorical attributes (i.e. **color** and **position**).
- The dissimilarity matrix is as shown below:
 - $- 0 = \text{identical}; 1 = \text{dissimilar}$
- m : # of matches, p : total # of attributes

	m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$		m	$\frac{p-m}{p}$	
$d(2,1)$	0	$\frac{2-0}{2} = 1$										
$d(3,1)$	0	$\frac{2-0}{2} = 1$	$d(3,2)$	1	$\frac{2-1}{2} = 0.5$							
$d(4,1)$	2	$\frac{2-2}{2} = 0$	$d(4,2)$	0	$\frac{2-0}{2} = 1$	$d(4,3)$	0	$\frac{2-0}{2} = 1$				
$d(5,1)$	0	$\frac{2-0}{2} = 1$	$d(5,2)$	0	$\frac{2-0}{2} = 1$	$d(5,3)$	1	$\frac{2-1}{2} = 0.5$	$d(5,4)$	0	$\frac{2-0}{2} = 1$	
$d(6,1)$	0	$\frac{2-0}{2} = 1$	$d(6,2)$	0	$\frac{2-0}{2} = 1$	$d(6,3)$	1	$\frac{2-1}{2} = 0.5$	$d(6,4)$	0	$\frac{2-0}{2} = 1$	$d(6,5)$
										2	$\frac{2-2}{2} = 0$	

Object	Color	Position	Size
1	R	L	L
2	B	C	M
3	G	C	M
4	R	L	H
5	G	R	L
6	G	R	H

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & & & & \\ 2 & 1 & 0 & & & \\ 3 & 1 & 0.5 & 0 & & \\ 4 & 0 & 1 & 1 & 0 & \\ 5 & 1 & 1 & 0.5 & 1 & 0 \\ 6 & 1 & 1 & 0.5 & 1 & 0 \end{bmatrix}$$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Ordinal Attribute

- Values of ordinal attribute follow a sequence (ordering)
 - E.g. Grade = {A, B, C}, where A >B >C
 - E.g., Size attribute: Small < Medium < Large < X-Large
- Suppose, f is an ordinal attribute and the set of values of f
 - $f = \{a_1, a_2, \dots, a_M\}$
 - M = the number of possible states that the ordinal attribute can have
- Let M values of f are ordered in ascending order as
 - $a_1 < a_2 < \dots < a_M$
 - Let i^{th} attribute value a_i be ranked as i , $i=1,2,\dots, M_f$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Ordinal Attribute

- **Step 1)**

- The value of f for the i th object is x_{if} , and f has M_f ordered states, representing the ranking $1, \dots, M_f$
- Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \dots, M_f\}$

- **Step 2)**

- Since each ordinal attribute can have a different number of states, it is often necessary to map the range of each attribute onto $[0.0, 1.0]$ so that each attribute has equal weight
- We perform such data normalization by replacing the rank r_{if} of the i th object in the f th attribute by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- **Step 3)**

- Dissimilarity can then be computed using any of the distance measures used for numeric attributes, using z_{if} to represent the f value for the i th object

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Ordinal Attribute

- Here, the attribute **Test** has three states:

– Fair, Good and Excellent, so $M_f = 3$

- Step 1:

– The four attribute values are assigned the ranks 3, 1, 2, and 3 respectively

- Step 2:

– Normalizes the ranking by mapping ranks to [0.0, 1.0]

$$\text{Norm}(\text{Rank } 3) = \frac{3-1}{3-1} = 1.0; \text{Norm}(\text{Rank } 2) = \frac{2-1}{3-1} = 0.5; \text{Norm}(\text{Rank } 1) = \frac{1-1}{3-1} = 0.0$$

- Step 3:

– Using Euclidean distance, a dissimilarity matrix is obtained as shown

- Therefore, students 1 and 2 are most dissimilar, as are students 2 and 4

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Student	Test	Rank	Norm
1	Excellent	3	1.0
2	Fair	1	0.0
3	Good	2	0.5
4	Excellent	3	1.0

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 0 & & \\ 3 & 1.0 & 0 & \\ 4 & 0.5 & 0.5 & 0 \\ 1 & 0 & & \\ 4 & 1.0 & 0.5 & \\ 3 & 0.5 & 0 & \\ 2 & 0.5 & 0.5 & 0 \\ 1 & 0 & 1.0 & \\ 3 & 0.5 & 0.5 & 0 \\ 4 & 0 & 0.5 & 1.0 \\ 2 & 0 & 0.5 & 1.0 \end{bmatrix}$$

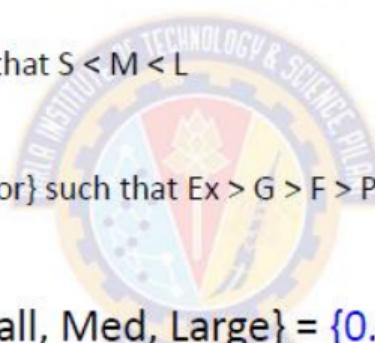
	$ a_i - a_j $		$ a_i - a_j $		$ a_i - a_j $
d(2,1)	$ 1.0 - 0.0 = 1.0$				
d(3,1)	$ 0.5 - 1.0 = 0.5$	d(3,2)	$ 0.5 - 0.0 = 0.5$		
d(4,1)	$ 1.0 - 1.0 = 0.0$	d(4,2)	$ 1.0 - 0.0 = 1.0$	d(4,3)	$ 1.0 - 0.5 = 0.5$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Ordinal Attribute

- Consider the following set of records, where each record is defined by two ordinal attributes

- Size = {Small, Medium, Large} such that S < M < L
 - Let's say
 - S = 1, M = 2, L = 3
- Quality = {Excellent, Good, Fair, Poor} such that Ex > G > F > P
 - Let's say
 - Ex = 4, G = 3, F = 2, & P = 1



Object	Size	Quality
A	S (0.0)	G (0.67)
B	L (1.0)	Ex (1.0)
C	L (1.0)	P (0.0)
D	M (0.5)	F (0.33)

- Normalized values of Size {Small, Med, Large} = {0.0, 0.5, 1.0}:

$$\text{High} = \frac{3-1}{3-1} = 1; \text{Low} = \frac{1-1}{3-1} = 0; \text{Medium} = \frac{2-1}{3-1} = 0.5;$$

- Normalized values of {Ex, G, F, P} = {1.0, 0.67, 0.33, 0.0}:

$$Ex = \frac{4-1}{4-1} = 1.0; G = \frac{3-1}{4-1} = 0.67; F = \frac{2-1}{4-1} = 0.33; P = \frac{1-1}{4-1} = 0.0$$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Ordinal Attribute

Object	Size	Quality
A	S (0.0)	G (0.67)
B	L (1.0)	Ex (1.0)
C	L (1.0)	P (0.0)
D	M (0.5)	F (0.33)

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1.053 & 0 & \\ 3 & 1.204 & 1.0 & 0 \\ 4 & 0.605 & 0.836 & 0.599 & 0 \end{bmatrix}$$

<i>Distance Measure</i>		<i>Distance Measure</i>	
$d(2,1)$	$\sqrt{(1.0 - 0.0)^2 + (1.0 - 0.67)^2} = \sqrt{1.0 + 0.1089} = \sqrt{1.1089} = 1.053$	$d(3,1)$	$\sqrt{(1.0 - 0.0)^2 + (0.0 - 0.67)^2} = \sqrt{1.0 + 0.4489} = \sqrt{1.4489} = 1.204$
$d(3,2)$	$\sqrt{(1.0 - 1.0)^2 + (0.0 - 1.0)^2} = \sqrt{0.0 + 1.0} = \sqrt{1.0} = 1.0$	$d(4,1)$	$\sqrt{(0.5 - 0.0)^2 + (0.33 - 0.67)^2} = \sqrt{0.25 + 0.1156} = \sqrt{0.3656} = 0.605$
$d(4,2)$	$\sqrt{(0.5 - 1.0)^2 + (0.33 - 1.0)^2} = \sqrt{0.25 + 0.4489} = \sqrt{0.6989} = 0.836$	$d(4,3)$	$\sqrt{(1.0 - 0.5)^2 + (0.33 - 0.0)^2} = \sqrt{0.25 + 0.1089} = \sqrt{0.3589} = 0.599$

EXERCISE-CANVAS DISCUSSION

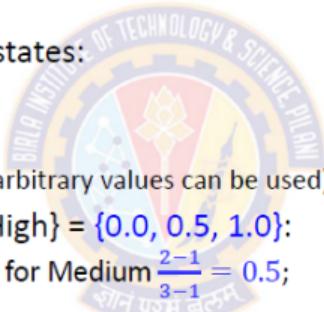
Object	Color	Position	Size	Rank	Value
1	R	L	L	1	0.0
2	B	C	M	2	0.5
3	G	C	M	2	0.5
4	R	L	H	3	1.0
5	G	R	L	1	0.0
6	G	R	H	3	1.0

Calculate the dissimilarity matrix for the ordinal attributes

EXERCISE-CANVAS DISCUSSION

Ordinal Attribute

- Dissimilarity matrix considering the ordinal **attribute (Size)**
 - 0 = identical; 1 = dissimilar
- Here, the attribute **Size** has three states:
 - Low, Med, & High, So $M_f = 3$
- Let's say
 - Low = 1, Med = 2, and High = 3 (any arbitrary values can be used)
- Normalized values of {Low, Med, High} = {0.0, 0.5, 1.0}:
 - For High $\frac{3-1}{3-1} = 1$; for Low $\frac{1-1}{3-1} = 0$; for Medium $\frac{2-1}{3-1} = 0.5$;



	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $	$ a_i - a_j $
d(2,1)	$ 0.5 - 0.0 = 0.5$				
d(3,1)	$ 0.5 - 0.0 = 0.5$	$d(3,2) 0.5 - 0.5 = 0.0$			
d(4,1)	$ 1.0 - 0.0 = 1.0$	$d(4,2) 1.0 - 0.5 = 0.5$	$d(4,3) 1.0 - 0.5 = 0.5$		
d(5,1)	$ 0.0 - 0.0 = 0.0$	$d(5,2) 0.0 - 0.5 = 0.5$	$d(5,3) 0.0 - 0.5 = 0.5$	$d(5,4) 0.0 - 1.0 = 1.0$	
d(6,1)	$ 1.0 - 0.0 = 1.0$	$d(6,2) 1.0 - 0.5 = 0.5$	$d(6,3) 1.0 - 0.5 = 0.5$	$d(6,4) 0.0 - 0.0 = 0.0$	$d(6,5) 1.0 - 0.0 = 1.0$

Object	Color	Position	Size	Rank	Value
1	R	L	L	1	0.0
2	B	C	M	2	0.5
3	G	C	M	2	0.5
4	R	L	H	3	1.0
5	G	R	L	1	0.0
6	G	R	H	3	1.0

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & & & & \\ 2 & 0.5 & 0 & & & \\ 3 & 0.5 & 0 & 0 & & \\ 4 & 1 & 0.5 & 0.5 & 0 & \\ 5 & 0 & 0.5 & 0.5 & 1 & 0 \\ 6 & 1 & 0.5 & 0.5 & 0 & 1 & 0 \end{bmatrix}$$

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

EXERCISE-CANVAS DISCUSSION

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Calculate the dissimilarity matrix and similarity matrix for the ordinal attributes

EXERCISE-CANVAS DISCUSSION

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Calculate the dissimilarity matrix and similarity matrix for the ordinal attributes

- There are three states for *test-2*: *fair*, *good*, and *excellent*, that is, $n= 3$.
- **Step 1:** we replace each value for *test-2* by its rank, the four objects are assigned the ranks 3, 1, 2, and 3, respectively.
- **Step 2 :** normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0.
- **Step 3:** use the Euclidean distance to calculate the dissimilarity

Dissimilarity Matrix

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

Similarity Matrix, $s=1-d$

	1	2	3	4
1	1			
2	0	1		
3	0.5	0.5	0	
4	1	0	0.5	1

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

- A contingency table for binary data

		Object <i>j</i>		
		1	0	sum
Object <i>i</i>	1	<i>q</i>	<i>r</i>	<i>q+r</i>
	0	<i>s</i>	<i>t</i>	<i>s+t</i>
	sum	<i>q+s</i>	<i>r+t</i>	<i>p</i>

- where q is the number of attributes that equal 1 for both objects i and j ,
- r is the number of attributes that equal 1 for object i but equal 0 for object j ,
- s is the number of attributes that equal 0 for object i but equal 1 for object j ,
- t is the number of attributes that equal 0 for both objects i and j .
- The total number of attributes is p , where $p = q+r+s+t$.

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

Symmetric binary attributes

- For symmetric binary attributes, each state is equally valuable
- Dissimilarity that is based on symmetric binary attributes is called **symmetric binary dissimilarity**
- If objects i and j are described by symmetric binary attributes, then the dissimilarity between i and j is

Contingency Table for Binary Attributes

		Object j		sum
		1	0	
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
	sum	$q + s$	$r + t$	p

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

- Asymmetric binary attributes
 - For asymmetric binary attributes, the two states are not equally important
 - E.g., the *positive* (1) and *negative* (0) outcomes of a disease test
 - Given two asymmetric binary attributes, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match)
 - Therefore, such binary attributes are often considered "monary" (having one state)
 - The dissimilarity based on these attributes is called **asymmetric binary dissimilarity**
 - Here, the number of negative matches, t , is considered unimportant and is thus ignored in the following computation:

$$d(i, j) = \frac{r + s}{q + r + s}$$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

- Measurement of similarity
 - We can measure the difference between two binary attributes based on the notion of similarity instead of dissimilarity
 - For example, the **asymmetric binary similarity** between the objects i and j can be computed as
 - $$\text{sim}(i, j) = 1 - d(i, j) = 1 - \frac{r + s}{q + r + s} = \frac{q}{q + r + s}$$
 - The coefficient $\text{sim}(i, j)$ above is called the **Jaccard coefficient**

PROXIMITY MEASURE FOR BINARY ATTRIBUTES-

EXERCISE

Symmetric Binary Attributes

- Consider the following two dataset, where objects are defined with symmetric binary attributes.

- Gender = {M, F}
- Food = {V, N}
- Caste = {H, M}
- Education = {E, M}
- Hobby = {T, C}
- Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Ahmed	M (1)	V (0)	M (1)	E (1)	C (1)	N (0)
Surekha	M (1)	N (1)	H (0)	M (0)	T (0)	N (0)
Mahesh	F (0)	N (1)	H (0)	E (1)	C (1)	Y (1)

Contingency Table (Ahmed & Surekha)

$$d(Surekha, Ahmed) = \frac{r + s}{q + r + s + t} = \frac{1 + 3}{1 + 1 + 3 + 1} = \frac{4}{6} = 0.67$$

Object	Ahmed			
		1	0	Sum
Surekha	1	1	1	2
	0	3	1	4
	Sum	4	2	6

PROXIMITY MEASURE FOR BINARY ATTRIBUTES-

EXERCISE

Symmetric Binary Attributes

- Jaccard's Coefficient Example

Fruit	Sphere	Sweet	Sour	Crunchy
Apple	Yes	Yes	Yes	Yes
Banana	No	Yes	No	No

Fruit	Sphere	Sweet	Sour	Crunchy
Apple	1	1	1	1
Banana	0	1	0	0

Contingency Table (Banana & Apple)

Object	Apple			
		1	0	Sum
Banana	1	1	3	4
	0	0	0	0
	Sum	1	3	4

$$d(Apple, Banana) = \frac{r + s}{q + r + s + t}$$
$$= \frac{3 + 0}{1 + 3 + 0 + 0} = 0.75$$

$$\text{Jaccard Coeff} = \frac{q}{q + r + s} = \frac{1}{1 + 3 + 0} = 0.25$$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES- EXERCISE

Asymmetric Binary Attributes

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Name	G	F	C	T-1	T-2	T-3	T-4
Jack	1	1	0	1	0	0	0
Mary	0	1	0	1	0	1	0
Jim	1	1	1	0	0	0	0

- Gender is a symmetric attribute (not included in the calculation)
- The remaining attributes are asymmetric binary
- Let the values Y and P = 1, and the value N = 0

Object	Mary		
	1	0	Sum
Jack	1	2 (q)	0 (r)
	0	1 (s)	3 (t)
	Sum	3	3

Contingency Table (Jack & Mary)

$$d(Jack, Mary) = \frac{r+s}{q+r+s} = \frac{0+1}{2+0+1} = 0.33$$

$$Jaccard\ Coeff = \frac{q}{q+r+s} = \frac{2}{2+0+1} = 0.67$$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES- SUMMARY

- Distance measure for symmetric binary variables,
dissimilarity: $d(i, j) = \frac{r + s}{q + r + s + t}.$
- Distance measure for asymmetric binary variables,
dissimilarity: $d(i, j) = \frac{r + s}{q + r + s}.$
- Similarity between Asymmetric binary values is given by
Jaccard coefficient:

$$sim(i, j) = \frac{q}{q + r + s} = 1 - d(i, j).$$

EXERCISE

Suppose that a patient record table contains the attributes *name*, *gender*, *fever*, *cough*, *test-1*, *test-2*, *test-3*, and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary. Compute the dissimilarity matrix for asymmetric binary attributes

Relational Table Where Patients Are Described by Binary Attributes

<i>name</i>	<i>gender</i>	<i>fever</i>	<i>cough</i>	<i>test-1</i>	<i>test-2</i>	<i>test-3</i>	<i>test-4</i>
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

EXERCISE

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

$$d(\text{Jack}, \text{Jim}) = \frac{1+1}{1+1+1} = 0.67,$$

$$d(\text{Jack}, \text{Mary}) = \frac{0+1}{2+0+1} = 0.33,$$

$$d(\text{Jim}, \text{Mary}) = \frac{1+2}{1+1+2} = 0.75.$$

		Jim	
		1	0
Jack	1	$q = 1$	$r = 1$
	0	$s = 1$	$t = 3$

		Mary	
		1	0
Jack	1	$q = 2$	$r = 0$
	0	$s = 1$	$t = 3$

		Jim	
		1	0
Mary	1	$q = 1$	$r = 1$
	0	$s = 2$	$t = 2$

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Dissimilarity of Numeric Data: Euclidean & Manhattan Dist.

- **Euclidean Distance** (i.e., straight line or "as the crow flies")

– The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **Manhattan (or city block) distance**

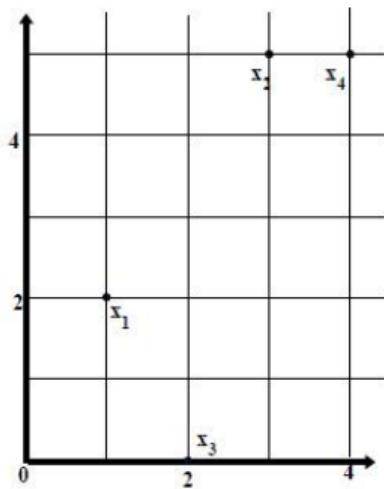
– It is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks)

– The Manhattan distance between objects i and j is defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Euclidean Distance



point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Euclidean distance between x_1 & x_2

$$= \sqrt{(3 - 1)^2 + (5 - 2)^2}$$

$$= \sqrt{4 + 9} = \sqrt{13} = 3.61$$

Dissimilarity Matrix
(with Euclidean Distance)

	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Manhattan Distance

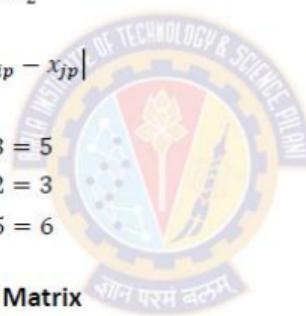
Manhattan distance between x_1 & x_2

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

$$d(x_2, x_1) = |3 - 1| + |5 - 2| = 2 + 3 = 5$$

$$d(x_3, x_1) = |2 - 1| + |0 - 2| = 1 + 2 = 3$$

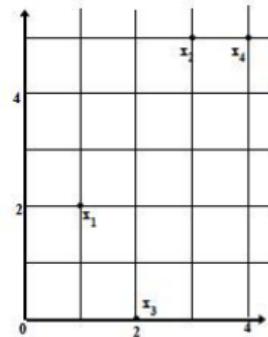
$$d(x_3, x_2) = |2 - 3| + |0 - 5| = 1 + 5 = 6$$



Dissimilarity Matrix

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



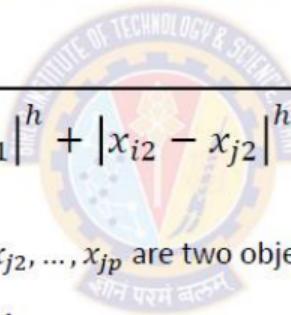
PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Dissimilarity of Numeric Data: Minkowski Distance

- Minkowski distance is a generalization of the Euclidean and Manhattan distances
- It is defined as

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

- Where
 - $i = x_{i1}, x_{i2}, \dots, x_{ip}$ and $j = x_{j1}, x_{j2}, \dots, x_{jp}$ are two objects described by p numeric attributes
 - h is a real number such that $h \geq 1$
 - It is also called as L_h norm
 - When $h = 1$, it represents the Manhattan distance (i.e., L_1 norm)
 - When $h = 2$, it represents the Euclidean distance (i.e., L_2 norm)



PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Dissimilarity of Numeric Data: Supremum Distance

- The **supremum distance** is a generalization of the Minkowski distance for $h = \infty$
 - also referred to as L_{\max} , L_∞ norm, and as the **Chebyshev distance**
- To compute it, we find the attribute f that gives the maximum difference in values between the two objects
- This difference is the supremum distance, defined more formally as:

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}| \right)^{1/h} = \max_f |x_{if} - x_{jf}|$$

- The L^∞ norm is also known as the *uniform norm*

PROXIMITY MEASURES FOR NUMERIC ATTRIBUTES:

Supremum

Supremum Distance between x_1 & x_2

$$d(i, j) = \max_f |x_{if} - x_{jf}|$$

$$d(x_2, x_1) = \text{Max} (|3 - 1|, |5 - 2|) = 3$$

$$d(x_3, x_1) = \text{Max} (|2 - 1|, |0 - 2|) = 2$$

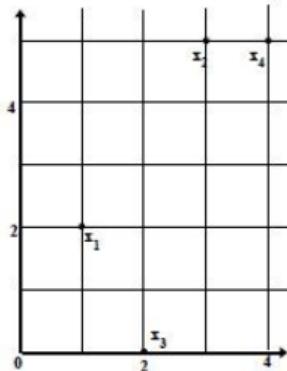
$$d(x_3, x_2) = \text{Max} (|2 - 3|, |0 - 5|) = 5$$



Dissimilarity Matrix

L_{ij}	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

point	attribute1	attribute2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

- Process all attribute types together, performing a single analysis
- Combine the different attributes into a single dissimilarity matrix, bringing all of the meaningful attributes onto a common scale of the interval [0.0, 1.0]
- Suppose that the data set contains p attributes of mixed type
- The dissimilarity $d(i, j)$ between objects i and j is defined as:

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

- Where,

- the indicator

- $\delta_{ij}^{(f)} = 0$, if either

- (1) x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or j), OR
 - (2) $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary

- Otherwise

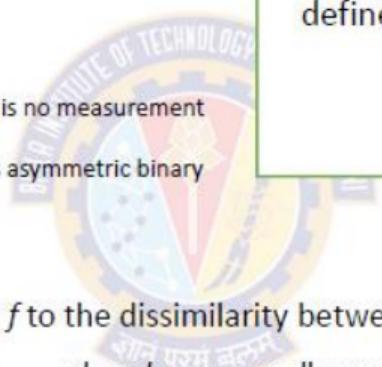
- $\delta_{ij}^{(f)} = 1$

- $d_{ij}^{(f)}$ is the contribution of attribute f to the dissimilarity between i and j . This is computed as:

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, where h runs over all non-missing objects for attribute f
 - If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$
 - If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if}-1}{M_f-1}$, and treat z_{if} as numeric

- The dissimilarity between two objects is defined as

$$d(i,j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$



PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

- First calculate distance (dissimilarity) matrices for each of the attributes separately
- We have already calculated the dissimilarity measures for Color and Quality
- Now, we will calculate for Quantity

Object	Color (Categorical)	Quality (Ordinal)	Quantity (Numeric)
1	R	Excellent	475
2	B	Fair	10
3	G	Good	1000
4	R	Excellent	500

- For numeric attribute like Quantity, we have to normalize the values so that the values can be mapped [0.0, 1.0]

$$d(i, j) = \frac{x_{if} - x_{jf}}{\max - \min}$$

$$d = \begin{matrix} & \text{Color} & & \text{Quality} & & \text{Quantity} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 1.0 & 0 & 2 & 1.0 & 0 & 2 & 0 \\ 3 & 1 & 0 & 1 & 3 & 0.5 & 0.5 & 3 & 0 \\ 4 & 0 & 1 & 1 & 4 & 0 & 1.0 & 0.5 & 4 & 0 \end{bmatrix} & d = & \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 1.0 & 0 & 2 & 1.0 & 0 & 2 & 0 \\ 3 & 1 & 0 & 1 & 3 & 0.5 & 0.5 & 3 & 0 \\ 4 & 0 & 1 & 1 & 4 & 0 & 1.0 & 0.5 & 4 & 0 \end{bmatrix} & d = & \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 3 & 0 & 0 & 0 & 3 & 0 & 0 & 3 & 0 \\ 4 & 0 & 0 & 0 & 4 & 0 & 0 & 4 & 0 \end{bmatrix} \end{matrix}$$

PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

	$\frac{ x_{if} - x_{jf} }{max - min}$		$\frac{ x_{if} - x_{jf} }{max - min}$		$\frac{ x_{if} - x_{jf} }{max - min}$
d(2,1)	$\frac{ 10 - 475 }{1000 - 10} = 0.470$				
d(3,1)	$\frac{ 1000 - 475 }{1000 - 10} = 0.530$	d(3,2)	$\frac{ 1000 - 10 }{1000 - 10} = 1.0$		
d(4,1)	$\frac{ 500 - 475 }{1000 - 10} = 0.025$	d(4,2)	$\frac{ 500 - 10 }{1000 - 10} = 0.4949$	d(4,3)	$\frac{ 500 - 1000 }{1000 - 10} = 0.505$

Object	Color (Categorical)	Quality (Ordinal)	Quantity (Numeric)
1	R	Excellent	475
2	B	Fair	10
3	G	Good	1000
4	R	Excellent	500



Quantity

$$d = \begin{bmatrix} 1 & 0 \\ 2 & 0.470 & 0 \\ 3 & 0.530 & 1.0 & 0 \\ 4 & 0.025 & 0.4949 & 0.505 & 0 \end{bmatrix}$$

PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

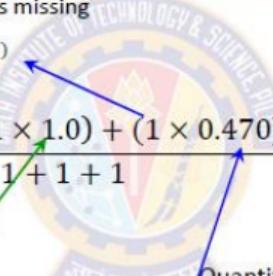
Mixed Attributes

- $\delta_{ij}^{(f)} = 1$, because
 - None of the attribute values in the objects is missing
 - No attribute has zero value
- Sample calculation for $d(2,1)$

$$d(i,j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

Object	Color (Categorical)	Quality (Ordinal)	Quantity (Numeric)
1	R	Excellent	475
2	B	Fair	10
3	G	Good	1000
4	R	Excellent	500

$$d(2,1) = \frac{(1 \times 1) + (1 \times 1.0) + (1 \times 0.470)}{1 + 1 + 1} = \frac{2.470}{3} = 0.823$$


 Color Quality Quantity

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & & \\ 2 & 1 & 0 & \\ 3 & 1 & 1 & 0 \\ 4 & 0 & 1 & 1 \end{bmatrix} \quad d = \begin{bmatrix} 1 & 0 & & \\ 2 & 1.0 & 0 & \\ 3 & 0.5 & 0.5 & 0 \\ 4 & 0 & 1.0 & 0.5 \end{bmatrix} \quad d = \begin{bmatrix} 1 & 0 & & \\ 2 & 0.470 & 0 & \\ 3 & 0.530 & 1.0 & 0 \\ 4 & 0.025 & 0.4949 & 0.505 \end{bmatrix}$$

PROXIMITY MEASURES FOR MIXED TYPE ATTRIBUTES

Mixed Attributes

	$\frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$		$\frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$
d(2,1)	$\frac{(1 \times 1) + (1 \times 1) + (1 \times 0.470)}{1 + 1 + 1} = 0.823$		
d(3,1)	$\frac{(1 \times 1) + (1 \times 0.5) + (1 \times 0.530)}{1 + 1 + 1} = 0.677$	d(3,2)	$\frac{(1 \times 1) + (1 \times 0.5) + (1 \times 1.0)}{1 + 1 + 1} = 0.833$
d(4,1)	$\frac{(1 \times 0) + (1 \times 0) + (1 \times 0.025)}{1 + 1 + 1} = 0.0083$	d(4,2)	$\frac{(1 \times 1) + (1 \times 1) + (1 \times 0.4949)}{1 + 1 + 1} = 0.832$
d(4,3)	$\frac{(1 \times 1) + (1 \times 0.5) + (1 \times 0.505)}{1 + 1 + 1} = 0.668$		

Overall Dissimilarity Matrix

સુધી પરમાણુમાં

$$d = \begin{bmatrix} 1 & 0 & 0.823 & 0.0083 \\ 2 & 0 & 0 & 0.677 \\ 3 & 0.823 & 0 & 0.833 \\ 4 & 0.0083 & 0.677 & 0 \end{bmatrix}$$

Color

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

Quality

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & & \\ 2 & 1.0 & 0 & \\ 3 & 0.5 & 0.5 & 0 \\ 4 & 0 & 1.0 & 0.5 \end{bmatrix}$$

Quantity

$$d = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & & \\ 2 & 0.470 & 0 & \\ 3 & 0.530 & 1.0 & 0 \\ 4 & 0.025 & 0.4949 & 0.505 \end{bmatrix}$$

COSINE SIMILARITY

Cosine similarity is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

Where $\|x\|$ is given by $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$.

EXERCISE

Suppose that x and y are the first two term-frequency vectors in That is, $x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and $y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$. How similar are x and y ? Compute the cosine similarity between the two vectors.

$$\begin{aligned}x \cdot y &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\&\quad + 0 \times 0 + 0 \times 1 = 25\end{aligned}$$

$$\|x\| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$\|y\| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$sim(x, y) = 0.94$$

TABLE OF CONTENTS

1 DATA SIMILARITY & DISSIMILARITY MEASURE

2 3 VISUALIZATION TECHNIQUES FOR DATA EXPLORATORY ANALYSIS

3 HANDLING NUMERIC DATA

4 MANAGING CATEGORICAL ATTRIBUTES

5 DEALING WITH TEXTUAL DATA

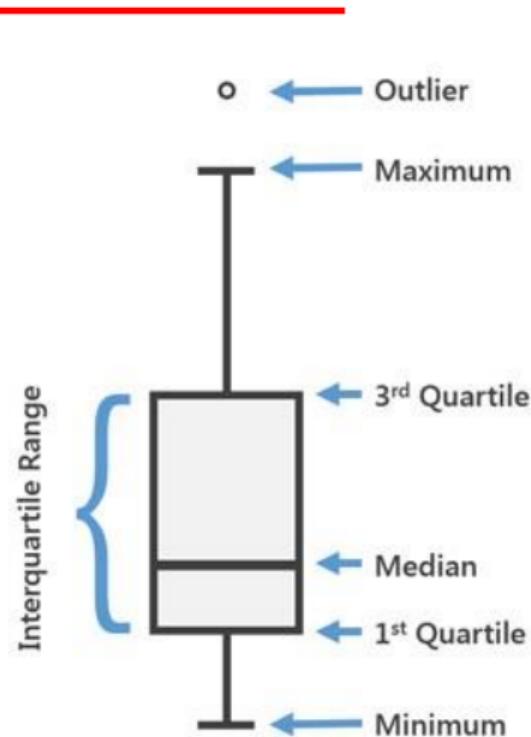
6

7

8

BOXPLOT

- A boxplot incorporates the five-number summary.
- The ends of the box are at the quartiles.
- The box length is the interquartile range.
- The median is marked by a line within the box.
- The whiskers outside the box extend to the Minimum and Maximum observations.
- Computed in $O(n \log n)$ time.



HISTOGRAM

- Graphical method for summarizing the distribution of an attribute, X .

- If X is nominal

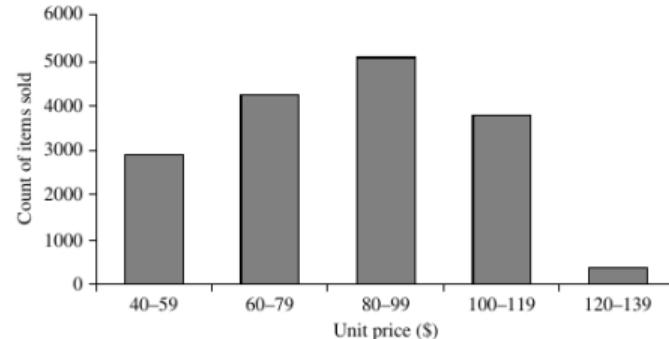
- **Bar chart**

- A vertical bar is drawn for each known value of X .
 - The height of the bar indicates the frequency of that X value.

- If X is numeric

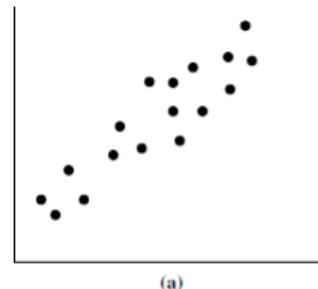
- **Histogram**

- The range of values for X is partitioned into disjoint consecutive **subranges or buckets or bins**.
 - The range of a bucket is known as the **width**.
 - The buckets are of equal width.

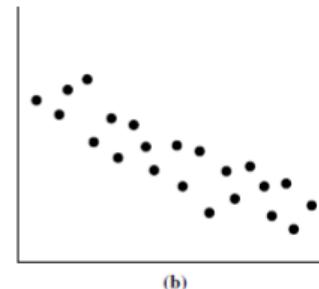


SCATTERPLOT

- Determine if there appears to be a relationship, pattern, or trend between two numeric attributes.
- Provide a visualization of bi-variate data to see clusters of points and outliers, or correlation relationships.
- Correlations can be positive, negative, or null (uncorrelated).



(a)



(b)

Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

DEMO CODE

Visualization.ipynb

TABLE OF CONTENTS

1 DATA SIMILARITY & DISSIMILARITY MEASURE

2 3 VISUALIZATION TECHNIQUES FOR DATA EXPLORATORY ANALYSIS

3 HANDLING NUMERIC DATA

4 MANAGING CATEGORICAL ATTRIBUTES

5 DEALING WITH TEXTUAL DATA

6

7

8

HANDLING NUMERIC DATA

Techniques are

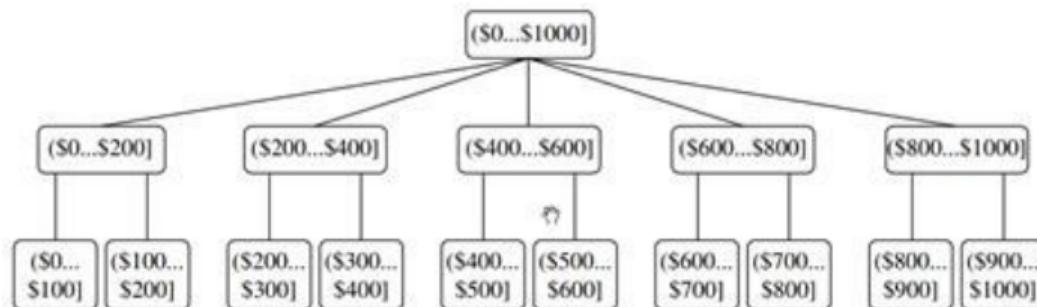
- Discretization – Convert numeric data into discrete categories
- Binarization – Convert numeric data into binary categories
- Normalization – Scale numeric data to a specific range
- Smoothing
 - create an approximating function that attempts to capture important patterns in the data, while leaving out noise
 - random method, simple moving average, random walk, simple exponential, and exponential moving average (Will learn in ISM)

DISCRETIZATION

- Convert continuous attribute into a discrete attribute.
- Discretization involves converting the raw values of a numeric attribute (e.g., age) into
 - interval labels (e.g., 0–10, 11–20, etc.)
 - conceptual labels (e.g., youth, adult, senior)
- Discretization Process
 - The raw data are replaced by a smaller number of interval or concept labels.
 - This simplifies the original data and makes the mining more efficient.
 - Concept hierarchies are also useful for mining at multiple abstraction levels.

CONCEPT HIERARCHY

- Divide the range of a continuous attribute into intervals.
- Interval labels can then be used to replace actual data values.
- The labels, in turn, can be recursively organized into higher-level concepts.
- This results in a concept hierarchy for the numeric attribute.



A concept hierarchy for the attribute *price*, where an interval $(\$X \dots \$Y]$ denotes the range from $\$X$ (exclusive) to $\$Y$ (inclusive).

DISCRETIZATION TECHNIQUES

Discretization techniques can be categorized based on how the discretization is performed.

■ Supervised vs. Unsupervised discretization

-) If the discretization process uses class information, then we say it is supervised discretization. Otherwise, it is unsupervised.

■ Top-down discretization or Splitting

-) The process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range. Then the process repeats recursively on the resulting intervals.

■ Bottom-up discretization or Merging

-) The process starts by considering all of the continuous values as potential split-points. Removes some by merging neighborhood values to form intervals. Then recursively applies this process to the resulting intervals.

DISCRETIZATION TECHNIQUES

- Unsupervised discretization
 - Binning [Equal-interval, Equal-frequency] (Top-down split)
 - Histogram analysis (Top-down split)
 - Clustering analysis (Top-down split or Bottom-up merge)
 - Correlation analysis (Bottom-up merge)
- Supervised discretization
 - Entropy-based discretization (Top-down split)

UNSUPERVISED DISCRETIZATION

- Class labels are ignored.
- The best number of bins k is determined experimentally.
- User specifies the **number of intervals** and/or **how many data points** to be included in any given interval.
- Use Binning methods.

DISCRETIZATION BY BINNING METHODS

1 Equal Width (distance) binning

- Each bin has equal width.

$$\text{width} = \text{interval} = \frac{\max - \min}{\# \text{bins}}$$

- Highly sensitive to outliers.
- If outliers are present, the width of each bin is large, resulting in skewed data.

2 Equal Depth (frequency) binning

- Specify the number of values that have to be stored in each bin.
- Number of entries in each bin are equal.
- Some values can be stored in different bins.

BINNING EXAMPLE

Discretize the following data into 3 discrete categories using binning technique.

70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81, 53, 56, 57, 63, 66, 67, 67, 67, 68, 69, 70, 70.

BINNING EXAMPLE

Original Data	53, 56, 57, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81			
Method		Bin1	Bin 2	Bin 3
Equal Width	width= $81-53 = 28$ $28/3 = 9.33$	$[53, 62) =$ 53, 56, 57	$[62, 72) =$ 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70	$[72, 81] =$ 72, 73, 75, 75, 76, 76, 78, 79, 80, 81
Equal Depth	depth = $24 / 3 = 8$	53, 56, 57, 63, 66, 67, 67, 67	68, 69, 70, 70, 70, 70, 72, 73	75, 75, 76, 76, 78, 79, 80, 81

DEMO CODE

Binning.ipynb

DISCRETIZATION BY HISTOGRAM ANALYSIS

- Histogram analysis is an unsupervised discretization technique because it does not use class information.
- Histograms use binning to approximate data distributions and are a popular form of data reduction.
- A histogram for an attribute, X, partitions the data distribution of X into disjoint subsets, referred to as buckets or bins.
- If each bucket represents only a single attribute–value/frequency pair, the buckets are called singleton buckets.
- Often, buckets represent continuous ranges for the given attribute.
- The histogram analysis algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy.

DISCRETIZATION BY HISTOGRAM ANALYSIS

1 Equal Width Histogram

- The values are partitioned into equal size partitions or ranges.

2 Equal Frequency Histogram

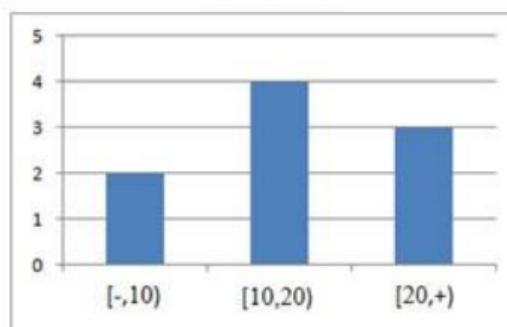
- The values are partitioned such that each partition contains the same number of data objects.

- Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28

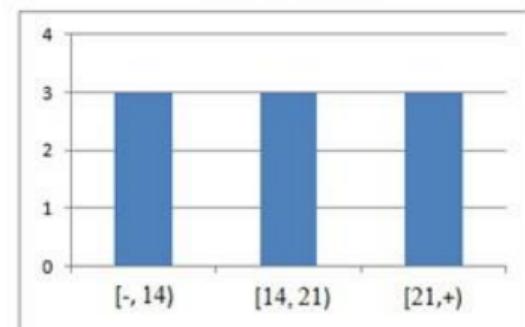
- Equal width**

- Bin 1: 0, 4 [-,10)
- Bin 2: 12, 16, 16, 18 [10,20)
- Bin 3: 24, 26, 28 [20,+)

Equal width



Equal frequency



VARIABLE TRANSFORMATION

- Variable transformation involves changing the values of an attribute.
- For each object (tuple), a transformation is applied to the value of the variable for that object.
 - 1 Simple functional transformations
 - 2 Normalization

SIMPLE FUNCTIONAL TRANSFORMATION

- For this type of variable transformation, a simple mathematical function is applied to each value individually.
- For a variable x , simple transformations include
 - x^k , $\log x$, e^x , \sqrt{x} , $\frac{1}{x}$, $\sin x$, $|x|$
- In statistics, variable transformations, especially $\log x$, \sqrt{x} , $\frac{1}{x}$, are often used to transform data that does not have a Gaussian (normal) distribution into data that does. While this can be important, other reasons often take precedence in data mining.
- Eg: Transfer of data bytes may be represented in the the \log_{10} transformation.

SIMPLE FUNCTIONAL TRANSFORMATION

- Variable transformations should be applied with caution since they change the nature of the data.
- For instance, the transformation $\frac{1}{x}$ reduces the magnitude of values that are 1 or larger, but increases the magnitude of values between 0 and 1.
- To understand the effect of a transformation, it is important to ask questions such as:
 - Does the order need to be maintained?
 - Does the transformation apply to all values, especially negative values and 0?
 - What is the effect of the transformation on the values between 0 and 1?

NORMALIZATION

- Normalizing the data attempts to give all attributes an equal weight.
- The goal of standardization or normalization is to make an entire set of values have a particular property.
- Normalization is particularly useful for:
 - classification algorithms involving neural networks.
 - 2 normalizing the input values for each attribute in the training tuples will help speed up the learning phase.
 - distance measurements such as nearest-neighbor classification and clustering.
 - 2 normalization helps prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes).

WHY FEATURE SCALING?

- Features with bigger magnitude **dominate** over the features with smaller magnitudes.
- Good practice to have all variables within a similar scale.
- Euclidean distances are **sensitive** to feature magnitude.
- Gradient descent converges faster when all the variables are in the similar scale.
- Feature scaling helps **decrease the time** of finding support vectors.

WHY FEATURE SCALING?

- For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income) from out-weighing attributes with initially smaller ranges (e.g., binary attributes).

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes	F Age	Outcome
6	148	72	35	0	34	0.627	50	1
1	85	66	29	0	27	0.351	31	0
8	183	64	0	0	23	0.672	32	1
1	89	66	23	94	28	0.167	21	0
0	137	40	35	168	43	2.288	33	1
5	116	74	0	0	26	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35	0.134	29	0
2	197	70	45	543	31	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	38	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27	1.441	57	0
1	189	60	23	846	30	0.398	59	1

ALGORITHMS SENSITIVE TO FEATURE MAGNITUDE

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-Means Clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

NORMALIZATION

- Scale the feature magnitude to a standard range like $[0, 1]$ or $[-1, +1]$ or any other. Techniques
 - Min-Max normalization
 - z-score normalization
 - Normalization by decimal scaling
- Impact of outliers in the data ???

MIN-MAX SCALING

- Min-max scaling squeezes (or stretches) all feature values to be within the range of $[0, 1]$.
- Min-Max normalization preserves the relationships among the original data values.
- It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for X .

$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{for range } [0, 1]$$

$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)} (new_{\max} - new_{\min}) + new_{\min} \quad \text{for range } [new_{\min}, new_{\max}]$$

MIN-MAX NORMALIZATION

Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively. The new range is [0.0,1.0]. Apply min-max normalization to value of \$73,600.

$$\begin{aligned}\hat{x} &= \frac{x - \min(x)}{\max(x) - \min(x)} (new_{max} - new_{min}) + new_{min} \\ &= \frac{73600 - 12000}{98000 - 12000} (1.0 - 0.0) + 0.0 \\ &= 0.716\end{aligned}$$

Z-SCORE NORMALIZATION

- In z-score normalization (or zero-mean normalization), the values for an attribute, x , are normalized based on the mean $\mu(x)$ and standard deviation $\sigma(x)$ of x .
- The resulting scaled feature has a mean of 0 and a variance of 1.
- New range is $[-3\sigma, +3\sigma]$.

$$\hat{x} = \frac{x - \mu(x)}{\sigma(x)}$$

- z-score normalization is useful when the actual minimum and maximum of attribute X are unknown, or when there are outliers that dominate the min-max normalization.

Z-SCORE NORMALIZATION

Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively. Apply z-score normalization to value of \$73,600.

$$\begin{aligned}\hat{x} &= \frac{x - \mu(x)}{\sigma(x)} \\ &= \frac{73600 - 54000}{16000} \\ &= 1.225\end{aligned}$$

DECIMAL NORMALIZATION

- Normalizes by moving the decimal point of values of attribute x .
- The number of decimal points moved depends on the maximum absolute value of x .
- New range is $[-1, +1]$.

$j = \text{smallest integer such that } \max(|\hat{x}|) < 1$

$$\hat{x} = \frac{x}{10^j}$$

DECIMAL NORMALIZATION

Example 1		
CGPA	Formula	Normalized CGPA
2	$2/10$	0.2
3	$3/10$	0.3
Example 2		
Bonus	Formula	Normalized Bonus
450	$450/1000$	0.45
310	$310/1000$	0.31
Example 3		
Salary	Formula	Normalized Salary
48000	$48000/100000$	0.48
67000	$67000/100000$	0.67

DEMO CODE

Normalization.ipynb

TABLE OF CONTENTS

1 DATA SIMILARITY & DISSIMILARITY MEASURE

2 3 VISUALIZATION TECHNIQUES FOR DATA EXPLORATORY ANALYSIS

3 HANDLING NUMERIC DATA

4 MANAGING CATEGORICAL ATTRIBUTES

5 DEALING WITH TEXTUAL DATA

6

7

8

BINARIZATION

- Binarization maps a continuous or categorical attribute into **one or more binary attributes.**
- Must maintain **ordinal relationship.**
- Algorithms that find association patterns require that the data be in the form of binary attributes.
E.g., Apriori algorithm, Frequent Pattern (FP) Growth algorithm

BINARIZATION TECHNIQUES

- One-hot encoding
- Label Encoding
- Ordinal Encoding
- Binary Encoding

ONE-HOT ENCODING

- Encode each categorical variable with a set of Boolean variables which take values 0 or 1, indicating if a category is present for each observation.
- One binary attribute for each categorical value.
- Advantages
 -) Makes no assumption about the distribution or categories of the categorical variable .
 -) Keeps all the information of the categorical variable .
 -) Suitable for linear models.
- Disadvantages
 -) Expands the feature space.
 -) Does not add extra information while encoding.
 -) Many dummy variables may be identical, introducing redundant information .
 -) Number of resulting attributes may become too large.
- In multi-class classification, the class label is converted using one-hot encoding.

ONE-HOT ENCODING EXAMPLE

- Assume an ordinal attribute for representing service of a restaurant:
 $(Awful < Poor < OK < Good < Great)$ requires 5 bits to maintain the ordinal relationship.

Service Quality	X1	X2	X3	X4	X5
Awful	0	0	0	0	1
Poor	0	0	0	1	0
OK	0	0	1	0	0
Good	0	1	0	0	0
Great	1	0	0	0	0

ON-HOT ENCODING EXAMPLE

id	color
1	red
2	blue
3	green
4	blue



id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

LABEL ENCODING

- Replace the categories by digits from 1 to n (or 0 to $n - 1$, depending the implementation), where n is the number of distinct categories of the variable.
- The categories are arranged in ascending order and the numbers are assigned.
- Advantages
 - Straightforward to implement.
 - Does not expand the feature space.
 - Work well enough with tree based algorithms.
- Disadvantages
 - Does not add extra information while encoding.
 - Not suitable for linear models.
 - Does not handle new categories in test set automatically.
- Used for features which have multiple values into domain. eg: colour, protocol types

LABEL ENCODING EXAMPLE

- Assume an ordinal attribute for representing service of a restaurant: (Awful, Poor, OK, Good, Great)

Service Quality	Integer Value
Awful	0
Poor	1
OK	2
Good	3
Great	4

DEMO CODE

Encoding.ipynb

TABLE OF CONTENTS

1 DATA SIMILARITY & DISSIMILARITY MEASURE

2 3 VISUALIZATION TECHNIQUES FOR DATA EXPLORATORY ANALYSIS

3 HANDLING NUMERIC DATA

4 MANAGING CATEGORICAL ATTRIBUTES

5 DEALING WITH TEXTUAL DATA

6

7

8

STEPS INVOLVED IN THE TEXTUAL DATA PROCESSING

- Removing special characters, changing the case (up-casing and down-casing).
- Tokenization – process of discretizing words within a document.
- Creating Document Vector or Term Document Matrix.
- Filtering Stop Words
- Lexical Substitution
- Stemming / Lemmatization

TOKENIZATION

- Document – In the text mining context, each sentence is considered a distinct document.
- Token – Each word is called a token.
- Tokenization – The process of discretizing words within a document is called tokenization.

DOCUMENT VECTOR OR TERM DOCUMENT MATRIX

- Create a matrix where each column consists of a token and the cells show the counts of the number of times a token appears.
- Each token is now an attribute in standard data science parlance and each document is an example (record).
- Unstructured raw data is now transformed into a format that is recognized by machine learning algorithms for training.
- The matrix / table is referred to as Document Vector or Term Document Matrix (TDM)
- As more new statements are added that have little in common, we end up with a very sparse matrix.
- We could also choose to use the term frequencies (TF) for each token instead of simply counting the number of occurrences.

TERM DOCUMENT MATRIX – EXAMPLE

Document 1	This is a book on data mining
Document 2	This book describes data mining and text mining using RapidMiner

Table 9.1 Building a Matrix of Terms From Unstructured Raw Text

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Table 9.2 Using Term Frequencies Instead of Term Counts in a TDM

	This	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	$1/7 = 0.1428$	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0
Document 2	$1/10 = 0.1$	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

TDM, Term document matrix.

STOP WORDS

- There are common words such as "a," "this," "and," and other similar terms. They do not really convey specific meaning.
- Most parts of speech such as articles, conjunctions, prepositions, and pronouns need to be filtered before additional analysis is performed.
- Such terms are called **stop words**.
- Stop word filtering is usually the second step that follows immediately after tokenization.
- The document vector gets reduced significantly after applying standard English stop word filtering.

STOP WORDS

- Domain specific terms might also need to be filtered out.
 - For example, if we are analyzing text related to the automotive industry, we may want to filter out terms common to this industry such as "car," "automobile," "vehicle," and so on.
- This is generally achieved by creating a separate dictionary where these context specific terms can be defined and then term filtering can be applied to remove them from the data.

LEXICAL SUBSTITUTION

- **Lexical substitution** is the process of finding an alternative for a word in the context of a clause.
- It is used to align all the terms to the same term based on the field or subject which is being analyzed.
- This is especially important in areas with specific jargon, e.g., in clinical settings.
- Example: common salt, NaCl, sodium chloride can be replaced by NaCl.
- Domain specific

STEMMING

- Stemming is usually the next process step following term filtering.
- Words such as "recognized," "recognizable," or "recognition" may be encountered in different usages, but contextually they may all imply the same meaning.
- The root of all these highlighted words is "recognize."
- The conversion of unstructured text to structured data can be simplified by reducing terms in a document to their basic stems, because only the occurrence of the root terms has to be taken into account.
- This process is called stemming.

PORTR STEMMING

- The most common stemming technique for text mining in English is the **Porter Stemming method**.
- Porter stemming works on a set of rules where the basic idea is to remove and/or replace the suffix of words.
 - Replace all terms which end in 'ies' by 'y,' such as replacing the term "anomalies" with "anomaly."
 - Stem all terms ending in "s" by removing the "s," as in "algorithms" to "algorithm."
- While the Porter stemmer is extremely efficient, it can make mistakes that could prove costly.
 - "arms" and "army" would both be stemmed to "arm," which would result in somewhat different contextual meanings.

LEMMATIZATION

- Lemmatization converts a word to its root form, in a more grammatically sensitive way.
 - While both stemming and lemmatization would reduce "cars" to "car," lemmatization can also bring back conjugated verbs to their unconjugated forms such as "are" to "be."
- Lemmatization uses POS Tagging (Part of Speech Tagging) heavily.
- POS Tagging is the process of attributing a grammatical label to every part of a sentence.
 - Eg: "Game of Thrones is a television series."
 - POS Tagging:

```
{“game”：“NN”}, {“of”：“IN”}, {“thrones”：“NNS”}, {“is”：“VBZ”}, {“a”：“DT”},  
 {“television”：“NN”}, {“series”：“NN”}
```

where: NN = noun, IN = preposition, NNS = noun in its plural form, VBZ = third-person singular verb, and DT = determiner.

DEMO CODE

NLP.ipynb

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



INTRODUCTION TO DATA SCIENCE MODULE # 5 : FEATURE ENGINEERING

IDS Course Team

BITS Pilani

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

1 FEATURE ENGINEERING

2 FEATURE SELECTION

3 FILTER METHODS

- Pearson's Correlation Coefficient
- Chi-Squared Statistic
- Information Theory Metrics
- Gini Index

4 WRAPPER METHODS

5 EVALUATION OF FEATURE SELECTION

6 FEATURE ENGINEERING FOR TEXT

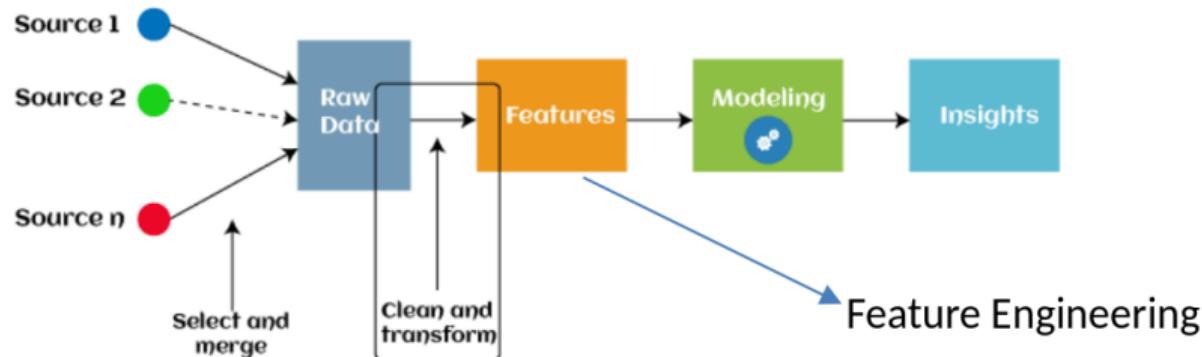
FEATURES

- Feature is a **property** of an object under study.
- Features are the **basic building blocks** of datasets.

Building Area	Common Area	Type of Flooring	Distance From Bus Depot	Sale Price per square feet
11345	350	Marble	16503.22	6,715
2000	1334	Vitrified Tiles	16321.19	3,230
2544	924	Wood Vitrified Tiles	15619.92	6,588

FEATURE ENGINEERING

- Feature Engineering is the process of **selecting and extracting useful, predictive features** from data.
- The goal is to create a set of features that **best represent** the information contained in the data, producing a simpler model that **generalizes well** to future observations.



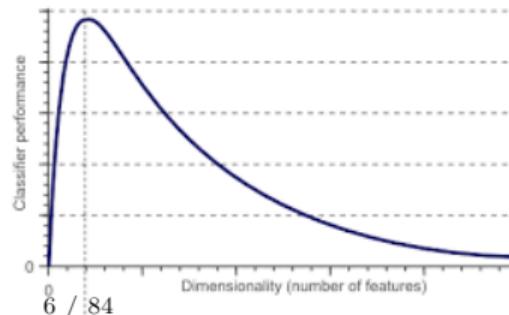
MOTIVATION FOR FEATURE ENGINEERING

HUGHES PHENOMENON

Given fixed number of data points, performance of a regressor or a classifier first increases but later decreases as the number of dimensions of the data increases.

Reasons for this phenomenon

- Redundant Features
- Correlation between features
- Irrelevant Features



FEATURE CREATION

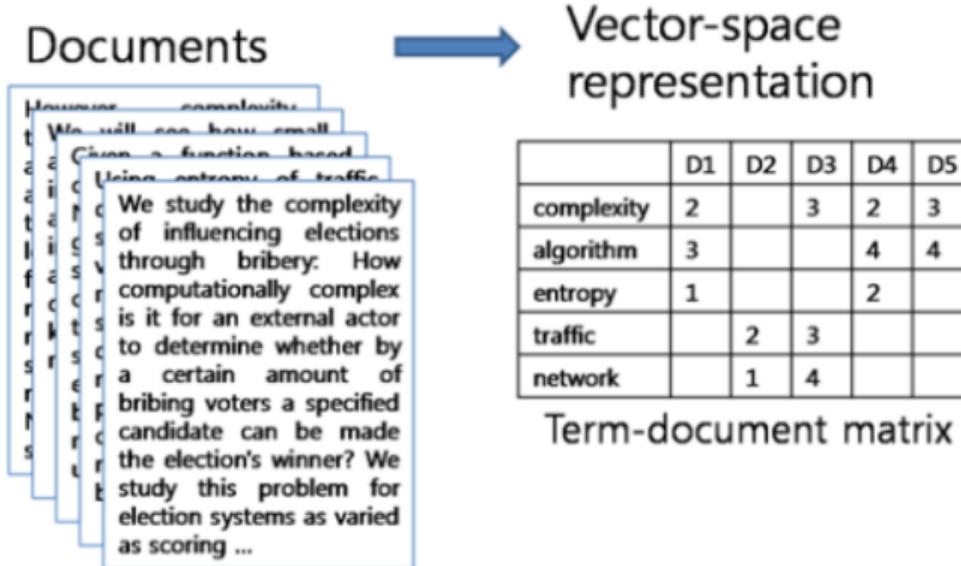
- Create new attributes that can capture important information in a dataset much more efficiently than the original attributes.
- Two general methodologies:
 - ▶ Feature Extraction
 - ▶ Feature Construction
 - ★ Create dummy features
 - ★ Create derived features

FEATURE EXTRACTION

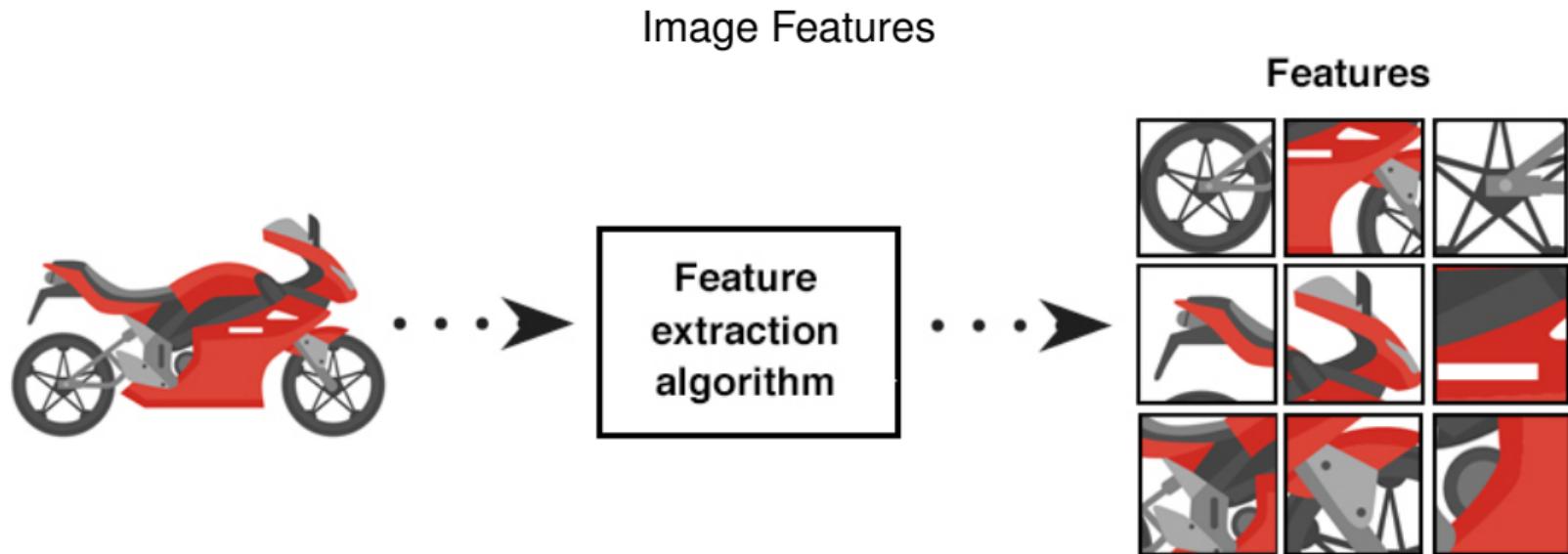
- Machine learning algorithms operate on a numeric feature space, expecting input as a two-dimensional array where **rows are instances and columns are features**.
- To perform machine learning on unstructured data like images and text, we need to transform the data into vector representations such that we can apply numeric machine learning.
- This process is called **feature extraction** or vectorization.
- Mostly rely on domain knowledge
 - ▶ Fourier Transform
 - ▶ Wavelet Transform
 - ▶ Scale-Invariant Feature Transform (SIFT) for images
 - ▶ Vector space transformation for text (TF-IDF)

FEATURE EXTRACTION

Bag of Words

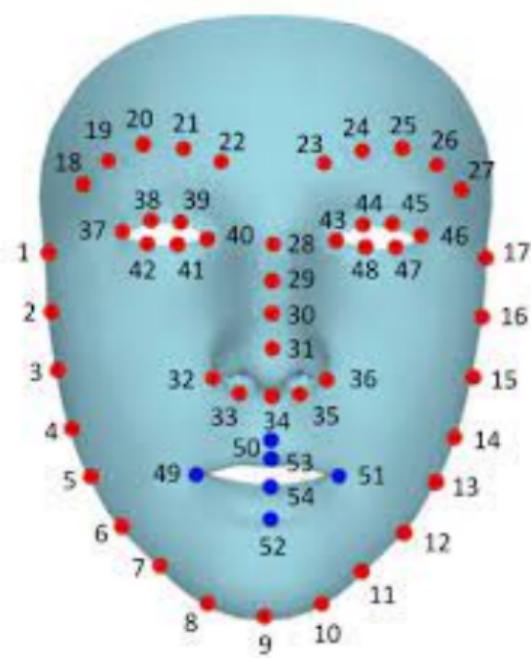


FEATURE EXTRACTION



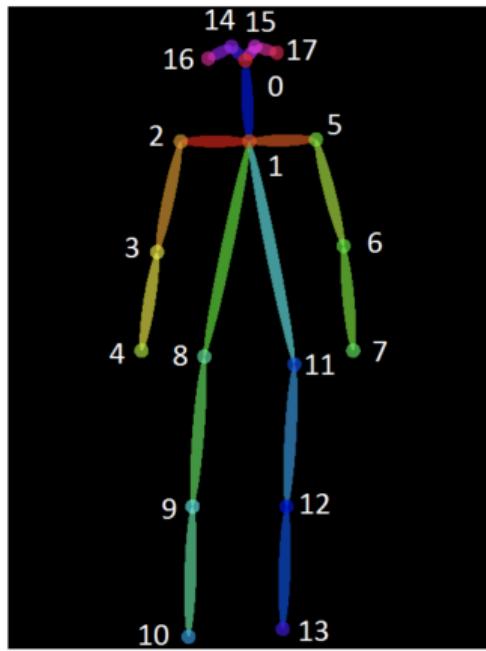
FEATURE EXTRACTION

Facial Landmarks



FEATURE EXTRACTION

Human Pose Estimation



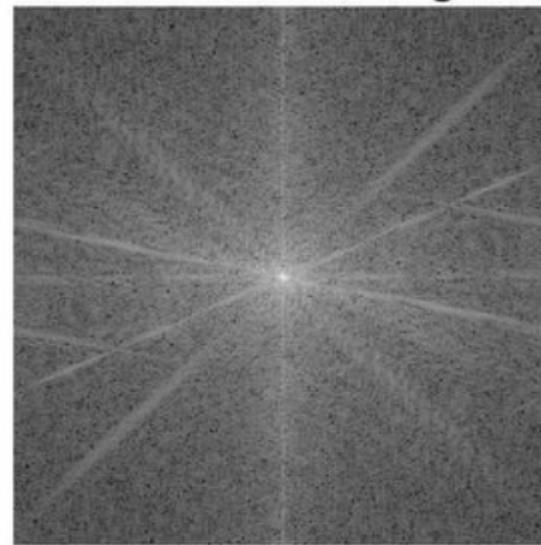
FEATURE EXTRACTION

Mapping Data to New Space – Fourier Transform on Images

Original Image



Transformed Image



FEATURE CONSTRUCTION

- Create dummy features
 - ▶ Often used to convert categorical variable to into numerical variables.
 - ▶ Use one-hot encoding or label encoding.

State (nominal scale)	State (Label encoding)
Maharashtra	3
Tamil Nadu	4
Delhi	0
Karnataka	2
Gujarat	1
Uttar Pradesh	5

FEATURE CONSTRUCTION

Customer ID	Gender	Payment Method
C001	Female	Online banking
C002	Male	Online banking
C003	Female	Credit card
C004	Male	Debit Card

Customer ID	Gender	Online banking	Credit card	Debit Card
C001	Female	1	0	0
C002	Male	1	0	0
C003	Female	0	1	0
C004	Male	0	0	1

FEATURE CONSTRUCTION

- Create derived features
- Involves creating a new feature using data from existing features
- Mostly rely on domain knowledge
- Eg: Calculating price per sqft

Area	Price (Rs)	Price/Sft (Rs)
1800	81,00,000	4500
2000	78,00,000	3900
1550	65,10,000	4200
2400	1,15,20,000	4800
3500	1,22,50,000	3500
2800	1,45,60,000	5200

FEATURE CONSTRUCTION

Customer ID	Gender	Session Begin	Session End
C001	Female	15-06-2019 10:30	15-06-2019 11:15
C002	Male	13-06-2019 08:00	13-06-2019 08:03
C003	Female	02-06-2019 16:25	02-06-2019 18:35
C004	Male	01-06-2019 11:20	01-06-2019 13:00

Customer ID	Gender	Session Duration
C001	Female	45
C002	Male	3
C003	Female	125
C004	Male	100

TABLE OF CONTENTS

1 FEATURE ENGINEERING

2 FEATURE SELECTION

3 FILTER METHODS

- Pearson's Correlation Coefficient
- Chi-Squared Statistic
- Information Theory Metrics
- Gini Index

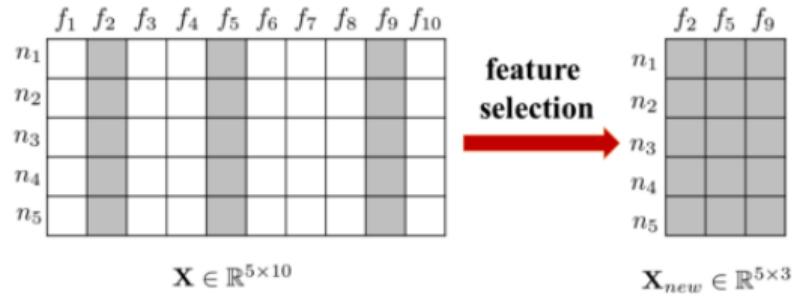
4 WRAPPER METHODS

5 EVALUATION OF FEATURE SELECTION

6 FEATURE ENGINEERING FOR TEXT

FEATURE SELECTION

- Feature selection is the process of identifying relevant and important features from irrelevant or redundant features.
- It intends to select a subset of attributes or features that makes the most meaningful contribution to a machine learning activity.



FACTORS AFFECTING FEATURE SELECTION

- Feature Relevance
 - ▶ In supervised algorithms, it is important for each feature to contribute towards the class label, otherwise it is irrelevant.
 - ▶ Need to determine : Strongly relevant, Moderately relevant and Weakly relevant features.
 - ▶ In case of unsupervised algorithms, there is no labelled data. During the grouping process, the algorithm identifies the irrelevant features.
- Feature Redundancy
 - ▶ A feature may contribute to information that is similar to the information contributed by one or more features.
 - ▶ All features having potential redundancy are candidates for rejection in the final feature subset.
 - ▶ If two features X_1, X_2 are highly correlated, then the two features become redundant features since they have same information in terms of correlation measure.

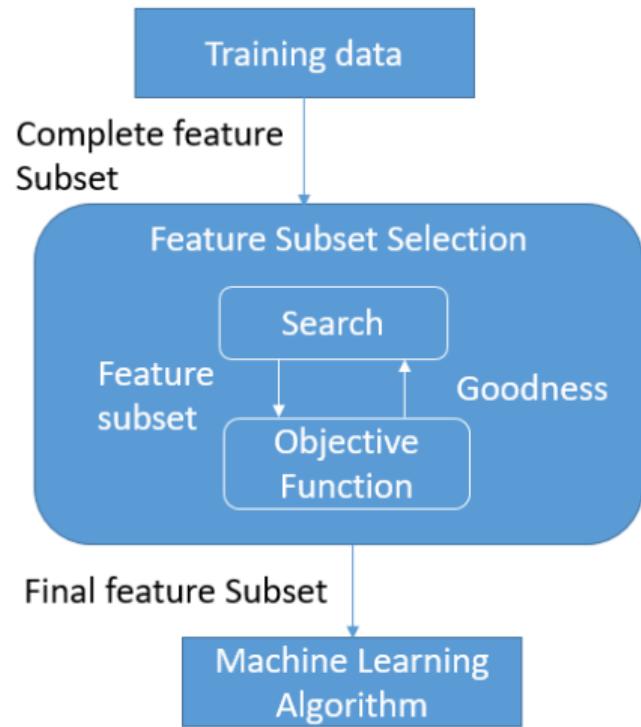
FEATURE SUBSET SELECTION

- Given: D initial set of features $F = \{f_1, f_2, f_3, \dots, f_D\}$ and target class label T .
- Find: Minimum subset $F' = \{f'_1, f'_2, f'_3, \dots, f'_M\}$ that achieves maximum classification performance where $F' \subseteq F$.
- There are 2^D possible subsets.
- Need a criteria to decide which subset is the best:
 - Classifier based on these M features has the lowest probability of error of all such classifiers.
- Evaluating 2^D possible subsets is time consuming and expensive.
- Use heuristics to reduce the search space.

STEPS IN FEATURE SELECTION

Feature selection is an optimization problem having the following steps:

- Step1: Search the space of all possible features.
- Step2: Pick the optimal subset using an objective function.



FEATURE SELECTION APPROACHES

- Unsupervised: Filter Methods
 - ▶ Use only features/predictor variables.
 - ▶ Select the features that have the most information.
- Supervised: Wrapper Methods
 - ▶ Train using the selected subset.
 - ▶ Estimate error on the validation set .
- Embedded Methods
 - ▶ Feature selection is done while training the model.
 - ▶ Example: Lasso (L1) Regularization and Decision Tree

TABLE OF CONTENTS

- 1 FEATURE ENGINEERING
- 2 FEATURE SELECTION
- 3 FILTER METHODS
 - Pearson's Correlation Coefficient
 - Chi-Squared Statistic
 - Information Theory Metrics
 - Gini Index
- 4 WRAPPER METHODS
- 5 EVALUATION OF FEATURE SELECTION
- 6 FEATURE ENGINEERING FOR TEXT

FILTER METHODS

- The Predictive power of individual feature is evaluated.
- Rank each feature according to some uni-variate metric and select the highest ranking features.
- Compute a score for each feature.
- The score should reflect the discriminative power of each feature.
- Advantages
 - ▶ Fast
 - ▶ Provides generically useful feature set.
- Disadvantages
 - ▶ Cause higher error than wrapper methods.
 - ▶ A feature that is not useful by itself can be very useful when combined with others. Filter methods can miss it.

FILTER METHODS

Algorithm

Given Input: large feature set F .

- ① Identify candidate subset $S \subseteq F$.
- ② While ! stop_criterion()
 - ① Evaluate utility function J using S .
 - ② Adapt S .
- ③ Return S .

TYPES OF FILTERS

- Correlation-based
 - ▶ **Pearson correlation**
 - ▶ Spearman rank correlation
 - ▶ Kendall concordance
- Statistical/probabilistic independence metrics
 - ▶ **Chi-square statistic**
 - ▶ F-statistic
 - ▶ Welch's statistic
- Information-theoretic metrics
 - ▶ **Mutual Information (Information Gain)**
 - ▶ Gain Ratio
- Others
 - ▶ **Gini index**
 - ▶ Fisher score
 - ▶ Cramer's V

WHICH FILTER ?

How do I pick the right filter ?

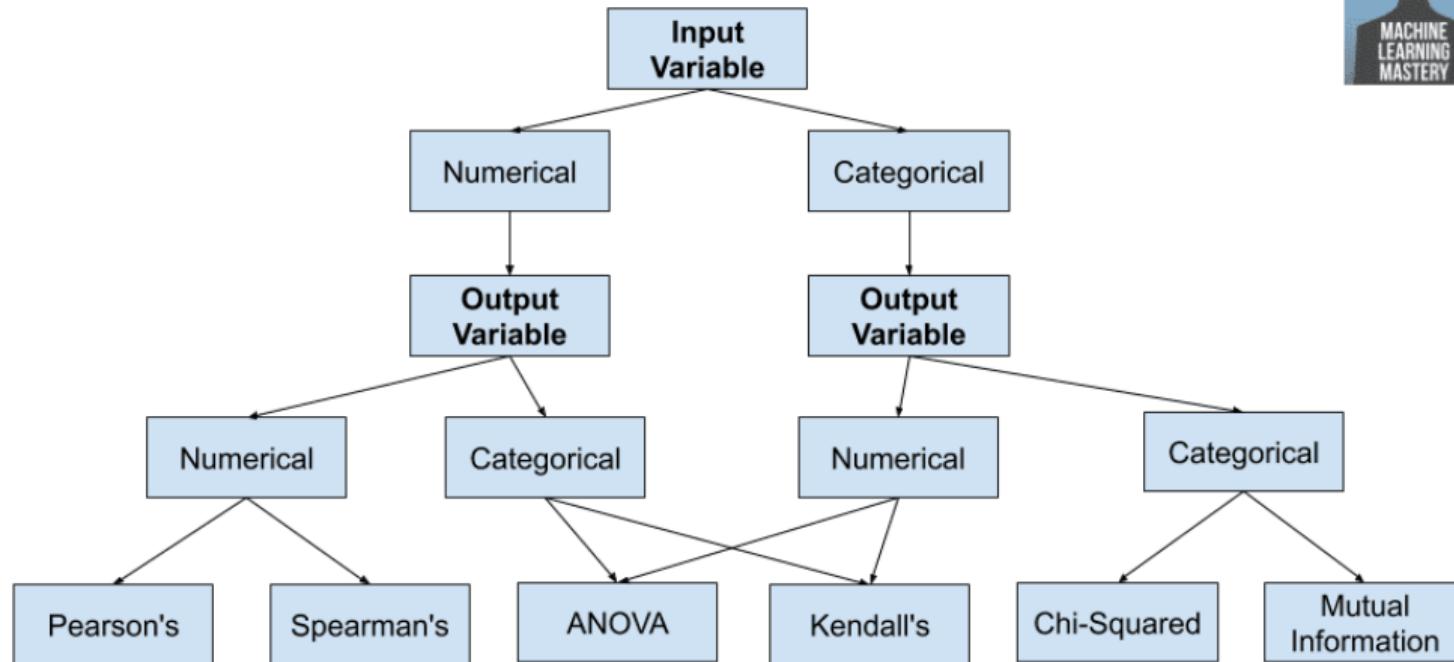
- Type of variables/targets (continuous, discrete, categorical).
- Class distribution
- Degree of non-linearity / feature interaction.

NO FREE LUNCH THEOREM

No Free Lunch theorem states that there is **no universal model** that works best for every problem.

WHICH FILTER ?

How to Choose a Feature Selection Method



PEARSON'S CORRELATION COEFFICIENT

- Used to measure the strength of association between **two continuous features**.
- Both positive and negative correlation are useful.
- We use Pearson Correlation to compute the correlation matrix or heat map.

Steps

- ① Compute the Pearson's Correlation Coefficient for each feature.
- ② Sort according the score.
- ③ Retain the highest ranked features, discard the lowest ranked.

Limitation

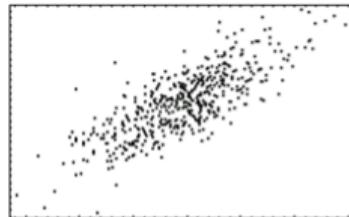
- Pearson assumes all features are **independent**.
- Pearson identifies only **linear** correlations
 - ▶ Positive linear relationship – In children, as the height increases, weight also increases.
 - ▶ Negative linear relationship – If the vehicle increases its speed, the time taken to travel decreases.

PEARSON'S CORRELATION COEFFICIENT

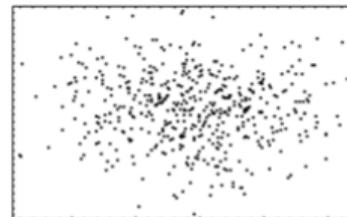
Feature: $x_k = \{x_k^{(1)}, \dots, x_k^{(N)}\}^T$

Target: $y = \{y^{(1)}, \dots, y^{(N)}\}^T$

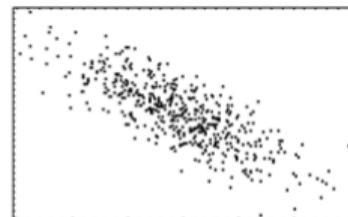
$$r(x, y) = \frac{\sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{(x^{(i)} - \bar{x})^2} \sqrt{(y^{(i)} - \bar{y})^2}}$$



$r = +0.5$



$r = 0.0$



$r = -0.5$

INTERPRETATION OF THE PEARSON CORRELATION

- $-1 \leq r_{A,B} \leq +1$
- If $r_{A,B} > 0$
 - ▶ A and B are positively correlated.
 - ▶ The values of A increase as the values of B increase.
 - ▶ The higher the value, the stronger the correlation (i.e., the more each attribute implies the other).
- If $r_{A,B} < 0$
 - ▶ A and B are negatively correlated.
 - ▶ The values of one attribute increase as the values of the other attribute decrease.
- If $r_{A,B} = 0$
 - ▶ A and B are independent and there is no correlation between them.
- If $r_{A,B} = -1 \text{ or } +1$
 - ▶ linear fit is perfect: all data points lie on one line.
- Use scatter plot for visualizing.

PEARSON'S CORRELATION EXAMPLE

Check whether sale of ice creams and sun glasses are related?

Ice cream sale	Sun glasses sale
A	B
20	30
10	5
23	29
5	10

PEARSON'S CORRELATION EXAMPLE

A	B	$A - \bar{A}$	$(A - \bar{A})^2$	$B - \bar{B}$	$(B - \bar{B})^2$	$(A - \bar{A})(B - \bar{B})$
20	30	5.5	30.25	11.5	132.25	63.25
10	5	-4.5	20.25	-13.5	182.25	60.75
23	29	8.5	72.25	10.5	110.25	89.25
5	10	-9.5	90.25	-8.5	72.25	80.75
58	74		213		497	294

PEARSON'S CORRELATION EXAMPLE

$$\bar{A} = \frac{58}{4} = 14.5$$

$$\bar{B} = \frac{74}{4} = 18.5$$

$$\sigma_A = \sqrt{\frac{213}{4}} = 7.29$$

$$\sigma_B = \sqrt{\frac{497}{4}} = 11.15$$

$$r_{A,B} = \frac{\sum(A - \bar{A})(B - \bar{B})}{n * \sigma_A * \sigma_B} = \frac{294}{4 * 7.29 * 11.15} = 0.9 \approx 1$$

= So positively correlated.

DEMO CODE

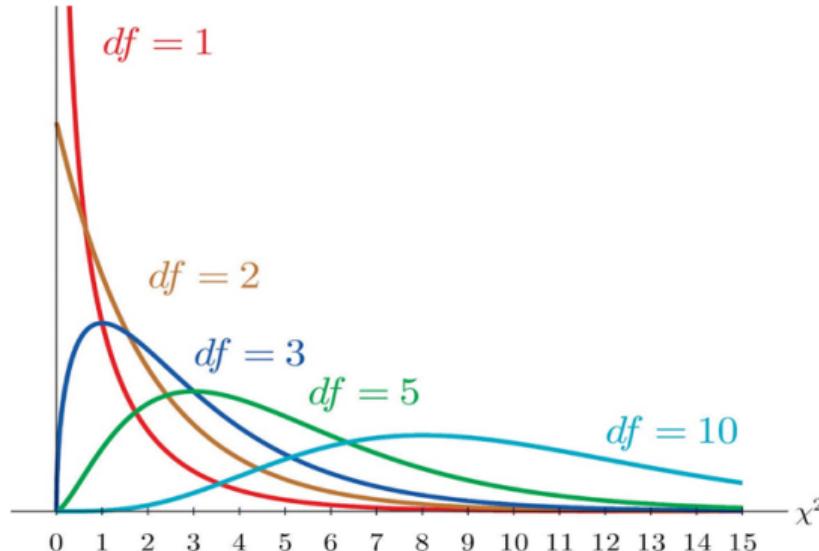
PearsonExample.py

CorrelationCoefficient.ipynb

PearsonCorrelation Covid Data.ipynb

χ^2 STATISTIC

- Chi-square test of independence allow us to see whether or not two **categorical variables** are related or not.
- The probability density function for the χ^2 distribution with r degrees of freedom (df) .



χ^2 STATISTIC EXAMPLE

Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table below:

Gender	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

χ^2 STATISTIC EXAMPLE

- To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.
- Step 1: Define the Hypothesis
 - ▶ **H0:** There is no link between gender and political party preference. [Null Hypothesis]
 - ▶ **HA:** There is a link between gender and political party preference. [Alternate Hypothesis]

χ^2 STATISTIC EXAMPLE

- Step 2: Calculate expected values

$$\text{Expected Value} = \frac{\text{RowTotal} \times \text{ColumnTotal}}{\text{Total number of observations}}$$

Gender	Republican	Democrat	Independent	Total
Male	$\frac{200*240}{440} = 109$	$\frac{200*130}{440} = 59$	$\frac{200*50}{440} = 23$	200
Female	$\frac{220*240}{440} = 120$	$\frac{220*130}{440} = 65$	$\frac{220*50}{440} = 25$	220
Total	240	130	50	440

χ^2 STATISTIC EXAMPLE

- Step 3: Calculate $\frac{(O-E)^2}{E}$ for each cell in the table.

Gender	Republican	Democrat	Independent
Male	$\frac{(100-109)^2}{109} = 0.74$	$\frac{(70-59)^2}{59} = 2.05$	$\frac{(30-23)^2}{23} = 2.13$
Female	$\frac{(140-120)^2}{120} = 3.33$	$\frac{(60-65)^2}{65} = 0.38$	$\frac{(20-25)^2}{25} = 1$

- Step 4: Calculate the χ^2 statistic.

$$\chi^2 = 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1 = 9.837$$

χ^2 STATISTIC EXAMPLE

- Step 5: From the table, critical value = 5.991 ($df = 2$, $alpha = 0.05$)
- Since Calculated value of $\chi^2 >$ Critical Value.
- H_0 is rejected ; H_A is accepted.
- Interpretation: There is sufficient evidence to say that there is a link between the gender and political party preference.

Critical values of the Chi-square distribution with d degrees of freedom

d	Probability of exceeding the critical value						
	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1
© 2013 Sinauer Associates, Inc.

χ^2 STATISTIC EXAMPLE

A group of customers were classified in terms of personality (introvert, extrovert or normal) and in terms of color preference (red, yellow or green) with the purpose of seeing whether there is an association (relationship) between personality and color preference. Data was collected from 400 customers and presented in the $3(\text{rows}) \times 3(\text{cols})$ contingency table below.

Observed Counts		Colors		
Personality		Red	Yellow	Green
Introvert		11	5	1
Extrovert		8	6	8
Normal		3	10	12
Total		22	21	21
				64

χ^2 STATISTIC EXAMPLE

Step 1:

- Set up hypotheses and determine level of significance.
- Null hypothesis(H_0): Color preference is independent of personality.
- Alternative hypothesis(H_A): Color preference is dependent on personality .
- Level of significance: specifies the probability of error. Generally it is set as 5%.

$$\alpha = 0.05$$

- Assume that H_0 is always true unless the evidence portraits something else in which case we will reject H_0 and accept H_A .

χ^2 STATISTIC EXAMPLE

Step 2:

- Compute the expected count.

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Expected Counts Personality	Colors			
	Red	Yellow	Green	Total
Introvert	5.8	5.6	5.6	17
Extrovert	7.6	7.2	7.2	22
Normal	8.6	8.2	8.2	25
Total	22	21	21	64

χ^2 STATISTIC EXAMPLE

Step 3:

- Compute the Chi-Squared Statistic.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$\chi^2 = \frac{(11 - 5.8)^2}{5.8} + \frac{(5 - 5.6)^2}{5.6} + \dots + \frac{(12 - 8.2)^2}{8.2} = 14.5$$

χ^2 STATISTIC EXAMPLE

Step 4

- Compute degrees of freedom.

$$df = (r - 1)(c - 1)$$

r is the number of categories in one variable and c is the number of categories in the other.

- $df = (3 - 1) \times (3 - 1) = 4$ (contingency table)

Step 5

- From the table, critical value = 9.488 ($df = 4$, $alpha = 0.05$)
- Since Calculated value of $\chi^2 >$ Critical Value of χ^2 H_0 is rejected ; H_A is accepted.
- Interpretation: There is sufficient evidence to say that Color Preference depends on the Personality.

DEMO CODE

ChiSquareGeneral.ipynb

ChiSquareCovidExample.ipynb

INFORMATION THEORY METRICS

- Information-theoretic concepts can only be applied to **discrete variables**.
- For continuous feature values, some data discretization techniques are required beforehand.

INFORMATION GAIN

- Information Gain $IG(X, Y)$ is a measure of the mutual independence between two features X and Y.
- Measures non-linear dependencies.

$$\begin{aligned} IG(X, Y) &= \text{Entropy}(Y) - \text{ConditionalEntropy}(Y|X) \\ &= H(Y) - H(Y|X) \end{aligned}$$

$$H(Y) = - \sum_{y \in Y} p(y) \log_2 p(y)$$

$$H(Y|X) = - \sum_{y \in Y} P(Y = y | X = x) \log_2 P(Y = y | X = x)$$

$$= - \sum_{x \in X, y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)}$$

INFORMATION GAIN

- Information Gain is symmetric.

$$IG(X, Y) = IG(Y, X)$$

- Higher Information Gain; better prediction of Y given X .
- $IG(X, Y) = 0$ if X and Y are independent.
- Biased towards the features having large number of discrete values.

INFORMATION GAIN

Compute the Information Gain for the attribute Travel Cost wrt Transport Mode.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

INFORMATION GAIN

- Step 1: Compute the Entropy of Transport Mode.

Transport Mode		
Bus	Car	Train
4	3	3

$$H(\text{Transport}) = H(4, 3, 3)$$

$$\begin{aligned} &= -\frac{4}{10} \log_2 \frac{4}{10} - \frac{3}{10} \log_2 \frac{3}{10} - \frac{3}{10} \log_2 \frac{3}{10} \\ &= 1.571 \end{aligned}$$

INFORMATION GAIN

- Step 2: Compute the Entropy of target given one feature.

Feature	Transport Mode		
	Bus	Train	Car
Cheap	4	1	0
Expensive	0	0	3
Standard	0	2	0

$$H(\text{Transport}|\text{Cost}) = H(5, 3, 2)$$

$$= -\frac{5}{10} \left(\frac{4}{5} \log_2 \frac{4}{5} + \frac{1}{5} \log_2 \frac{1}{5} \right) - \frac{3}{10} \left(\frac{3}{3} \log_2 \frac{3}{3} \right) - \frac{2}{10} \left(\frac{2}{2} \log_2 \frac{2}{2} \right)$$

$$= 0.36$$

INFORMATION GAIN

- Step 3: Compute the information gain.

$$\begin{aligned}IG(\text{Transport}|\text{Cost}) &= H(4, 3, 3) - (H(5, 3, 2) \\&= 1.571 - 0.36 \\&= 1.211\end{aligned}$$

DEMO CODE

InformationGainCovidData.ipynb

GINI INDEX

- Gini index minimizes the probability of misclassification.
- Used in CART (Classification and Regression Tree) algorithms.

$$Gini = 1 - \sum_{i=1}^K p_k^2$$

where p_k denotes the proportion of instances belonging to class k .

- Higher Gini Index; better prediction of Y given X .

GINI INDEX

Compute the Gini Index for the feature Travel Cost wrt Transport Mode.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

GINI INDEX

- Step 1: Compute the Gini Index for each value of the feature.

$$Gini(\text{Transport} | \text{Cost} = \text{Cheap}) = 1 - (0.8^2 + 0.2^2) = 0.32$$

$$Gini(\text{Transport} | \text{Cost} = \text{Expensive}) = 1 - (1^2 + 0) = 0$$

$$Gini(\text{Transport} | \text{Cost} = \text{Standard}) = 1 - (1^2 + 0) = 0$$

- Step 2: Compute the Gini Index for feature.

$$Gini(\text{Transport} | \text{Cost}) = \frac{5}{10} * 0.32 + \frac{3}{10} * 0 + \frac{2}{10} * 0 = 0.16$$

- Gini index < 0.2 represents **perfect** equality between features transport mode and travel cost.

GINI INDEX INTERPRETATION

- Gini index < 0.2 represents **perfect** equality.
- Gini index between 0.2 and 0.3 represent **relative** equality.
- Gini index between 0.3 and 0.4 represent **adequate** equality.
- Gini index between 0.4 and 0.5 represent big gap.
- Gini index > 0.5 represent severe gap.
- 0 represents perfect equality.
- 1 represents perfect inequality.

TABLE OF CONTENTS

- ① FEATURE ENGINEERING
- ② FEATURE SELECTION
- ③ FILTER METHODS
 - Pearson's Correlation Coefficient
 - Chi-Squared Statistic
 - Information Theory Metrics
 - Gini Index
- ④ WRAPPER METHODS
- ⑤ EVALUATION OF FEATURE SELECTION
- ⑥ FEATURE ENGINEERING FOR TEXT

WRAPPER METHODS

- Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset.
- The feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset.
- It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.
- The wrapper methods usually result in better predictive accuracy than filter methods.

WRAPPER METHODS

- Greedy Based algorithms.
- Performance of the method depends on the machine learning models chosen.
- Sequential feature selection algorithm add or remove one feature at a time based on the classifier performance until a desired criterion is met.
- Two methods
 - ▶ Sequential Forward Selection(SFS)
 - ▶ Sequential Backward Selection(SBS)
- Advantages
 - ▶ Highest performance
- Disadvantages
 - ▶ Computationally expensive
 - ▶ Memory intensive

WRAPPER METHODS TYPES

- Forward selection
 - ▶ starts with one predictor and adds more iteratively.
 - ▶ At each subsequent iteration, the best of the remaining original predictors are added based on performance criteria.
 - ▶ *SequentialFeatureSelector class from mlxtend*
- Backward elimination
 - ▶ starts with all predictors and eliminates one-by-one iteratively.
 - ▶ One of the most popular algorithms is Recursive Feature Elimination (RFE) which eliminates less important predictors based on feature importance ranking.
 - ▶ *RFE class from sklearn*

SEQUENTIAL FORWARD SELECTION

Data: Input: $Y = \{y_1, y_2, \dots, y_d\}$

Result: Output: $X_k = \{x_j | j = 1, 2, \dots, k; x_j \in Y\}$, where $k = (0, 1, 2, \dots, d)$, $k < d$

Initialization: $X_0 = \emptyset$, $k = 0$;

while $k = \text{desired set of features}$ **do**

$x^* = \arg \max J(X_k + x)$, where $x \in Y - X_k$;

$X_{k+1} = X_k + x^*$;

$k = k + 1$;

end

SFS EXAMPLE – WINE DATA

	feature_idx	cv_scores	avg_score	feature_names
1	(6,)	[0.8426966292134831]	0.842697	(flavanoids,)
2	(6, 9)	[0.9438202247191011]	0.94382	(flavanoids, color_intensity)
3	(6, 8, 9)	[0.949438202247191]	0.949438	(flavanoids, proanthocyanins, color_intensity)
4	(0, 6, 8, 9)	[0.9662921348314607]	0.966292	(alcohol, flavanoids, proanthocyanins, color_i...)
5	(0, 5, 6, 8, 9)	[0.9775280898876404]	0.977528	(alcohol, total_phenols, flavanoids, proanthoc...)
6	(0, 5, 6, 7, 8, 9)	[0.9719101123595506]	0.97191	(alcohol, total_phenols, flavanoids, nonflavan...)
7	(0, 2, 5, 6, 7, 8, 9)	[0.9662921348314607]	0.966292	(alcohol, ash, total_phenols, flavanoids, nonf...

SEQUENTIAL BACKWARD SELECTION

Data: Input: $Y = \{y_1, y_2, \dots, y_d\}$

Result: Output: $X_k = \{x_j | j = 1, 2, \dots, k; x_j \in Y\}$, where $k = (0, 1, 2, \dots, d)$, $k < d$

Initialization: $X_0 = Y$, $k = d$;

while $k = \text{desired set of features}$ **do**

$x^- = \arg \max J(X_k - x)$, where $x \in X_k$;

$X_{k-1} = X_k - x^-$;

$k = k - 1$;

end

SBS EXAMPLE – WINE DATA

	feature_idx	cv_scores	avg_score	feature_names
13	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)	[0.8258426966292135]	0.825843	(alcohol, malic_acid, ash, alcalinity_of_ash, ...)
12	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	[0.9269662921348315]	0.926966	(alcohol, malic_acid, ash, alcalinity_of_ash, ...)
11	(0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11)	[0.9550561797752809]	0.955056	(alcohol, malic_acid, ash, alcalinity_of_ash, ...)
10	(0, 2, 3, 5, 6, 7, 8, 9, 10, 11)	[0.9662921348314607]	0.966292	(alcohol, ash, alcalinity_of_ash, total_phenol...)
9	(0, 2, 3, 6, 7, 8, 9, 10, 11)	[0.9719101123595506]	0.97191	(alcohol, ash, alcalinity_of_ash, flavanoids, ...)
8	(0, 2, 3, 6, 7, 8, 10, 11)	[0.9719101123595506]	0.97191	(alcohol, ash, alcalinity_of_ash, flavanoids, ...)
7	(0, 2, 3, 6, 7, 8, 11)	[0.9775280898876404]	0.977528	(alcohol, ash, alcalinity_of_ash, flavanoids, ...)
6	(0, 2, 3, 6, 8, 11)	[0.9775280898876404]	0.977528	(alcohol, ash, alcalinity_of_ash, flavanoids, ...)

EMBEDDED METHODS

- Embedded methods combine the qualities of filter and wrapper methods.
- Implemented by algorithms that have their own built-in feature selection methods.
- The most common embedded technique are the tree algorithm's like RandomForest, ExtraTree, and so on.

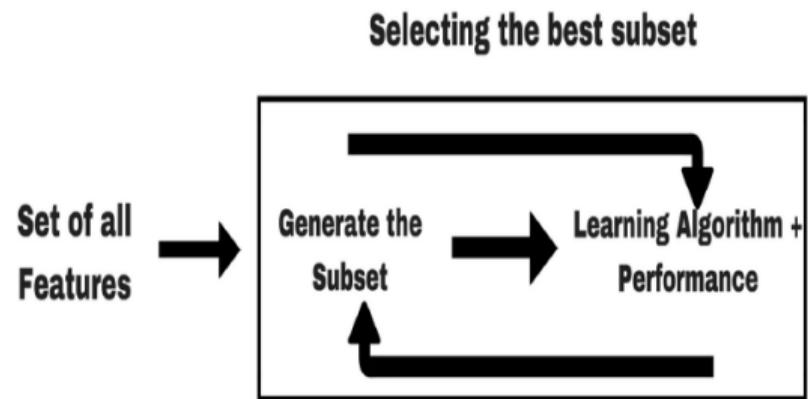


TABLE OF CONTENTS

1 FEATURE ENGINEERING

2 FEATURE SELECTION

3 FILTER METHODS

- Pearson's Correlation Coefficient
- Chi-Squared Statistic
- Information Theory Metrics
- Gini Index

4 WRAPPER METHODS

5 EVALUATION OF FEATURE SELECTION

6 FEATURE ENGINEERING FOR TEXT

EVALUATION OF FEATURE SELECTION

- Feature weighting:
 - ▶ given a desired feature number k , rank features according to the feature scores, and then return the top- k .
- Feature subset selection:
 - ▶ directly return the obtained feature subset (cannot specify beforehand)

EVALUATION OF FEATURE SELECTION

- Supervised feature selection
 - ① Divide data into training and testing set.
 - ② Perform feature selection to obtain selected features.
 - ③ Obtain the training and testing data on the selected features.
 - ④ Feed into a classifier.
 - ⑤ Obtain the classification performance on (e.g., F1, AUC).

The higher the classification performance, the better the selected features are.

EVALUATION OF FEATURE SELECTION

- Unsupervised feature selection
 - ① Perform feature selection on data to obtain selected features.
 - ② Obtain new data on the selected features.
 - ③ Perform clustering.
 - ④ Compare the obtained clustering with the ground truth.
 - ⑤ Obtain clustering evaluation results (e.g., NMI).

The higher the clustering performance, the better the selected features are.

TABLE OF CONTENTS

- ① FEATURE ENGINEERING
- ② FEATURE SELECTION
- ③ FILTER METHODS
 - Pearson's Correlation Coefficient
 - Chi-Squared Statistic
 - Information Theory Metrics
 - Gini Index
- ④ WRAPPER METHODS
- ⑤ EVALUATION OF FEATURE SELECTION
- ⑥ FEATURE ENGINEERING FOR TEXT

N-GRAMS

- There are families of words in the spoken and written language that typically go together. Grouping such terms, called **n-grams**, and analyzing them statistically can present new insights.
- The final pre-processing step typically involves forming these n-grams and storing them in the document vector.
- Algorithms providing n-grams become computationally expensive and the results become huge so in practice the amount of "n" will vary based on the size of the documents and the corpus.

EXAMPLE

- This is a sentence.

Unigram		This	is	a	sentence
Bigram		This is	is a	a sentence	
Trigram		This is a	is a sentence		

- The Margarita pizza does not taste bad.

Unigram		The	pizza	does	not	taste	bad
Bigram		The pizza	pizza does	does not	not taste	taste	bad
Trigram		The pizza does	pizza does not	does not taste	not taste	bad	

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

- Consider a web search problem where the user types in some keywords and the search engine extracts all the documents (essentially, web pages) that contain these keywords.
- How does the search engine know which web pages to serve up?
- In addition to using network rank or page rank, the search engine also runs some form of text mining to identify the most relevant web pages.
 - ▶ Example, the user types in the following keywords: "RapidMiner books that describe text mining."
- In this case, the search engines run on the following basic logic:
 - ▶ Give a high weight-age to those keywords that are relatively rare.
 - ▶ Give a high weight-age to those web pages that contain a large number of instances of the rare keywords.

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

- Term Frequency (TF)
 - Refers to the ratio of the number of times a keyword appears in a given document, n_k (where k is the keyword), to the total number of terms in the document, n .

$$TF(t, d) = \begin{cases} 0 & \text{if } freq(t, d) = 0 \\ 1 + \log(1 + \log(freq(t, d))) & \text{otherwise} \end{cases}$$

- Inverse Document Frequency (IDF)
 - Refers to the ratio of the total number of documents N , to the number of documents that contain the keyword k , N_k .

$$IDF(t) = \log \left(\frac{1 + |d|}{|d_t|} \right)$$

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY

- Term Frequency-Inverse Document Frequency (TF-IDF)

$$TF - IDF = TF * IDF$$

- Typically, TF-IDF scores for every word in the set of documents is calculated in the preprocessing stage.

EXAMPLE

For the given documents,

- D1: An apple is a fruit, which is red in colour.
 - D2: An orange fruit is orange in colour.
 - D3. A kiwi fruit is green coloured fruit.
- A. Remove stop words and lemmatize.
- B. Construct the term frequency matrix.
- C. Compute the TF-IDF matrix.

EXAMPLE

- Remove stop words
 - ▶ D1: apple fruit red colour
 - ▶ D2: orange fruit orange colour
 - ▶ D3: kiwi fruit green coloured fruit
- Lemmatize
 - ▶ D1: apple fruit red colour
 - ▶ D2: orange fruit orange colour
 - ▶ D3. kiwi fruit green colour fruit
- Construct the term frequency matrix.

	Apple	Fruit	Red	Colour	Orange	Kiwi	Green
D1	1	1	1	1	0	0	0
D2	0	1	0	1	2	0	0
D3	0	2	0	1	0	1	1

EXAMPLE

- $TF(t, d) = 1 + \log(1 + \log(freq(t, d)))$

	Apple	Fruit	Red	Colour	Orange	Kiwi	Green
D1	1	1	1	1	0	0	0
D2	0	1	0	1	1.114	0	0
D3	0	1.114	0	1	0	1	1

- $IDF(t) = \log\left(\frac{1+|d|}{|d_t|}\right)$

Apple	Fruit	Red	Colour	Orange	Kiwi	Green
0.602	0.124	0.602	0.124	0.602	0.602	0.602

$$\text{Apple} = \log((1 + 3)/1) = 0.602$$

$$\text{Fruit} = \log((1 + 3)/3) = 0.124$$

EXAMPLE

- $TF - IDF = TF * IDF$

	Apple	Fruit	Red	Colour	Orange	Kiwi	Green
D1	0.602	0.125	0.602	0.125	0	0	0
D2	0	0.125	0	0.125	0.670	0	0
D3	0	0.138	0	0.125	0	0.602	0.602

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



INTRODUCTION TO DATA SCIENCE MODULE # 6 : CLASSIFICATION

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

- 1 CLASSIFICATION AND PREDICTION
- 2 CLASSIFICATION
- 3 DECISION TREE ALGORITHM
- 4 OCCAM's RAZOR
- 5 EVALUATION OF CLASSIFICATION TECHNIQUES

LEARNING EXPERIENCE

During classification or prediction, a model is generated by observing or learning from historical data. This learning can be two types.

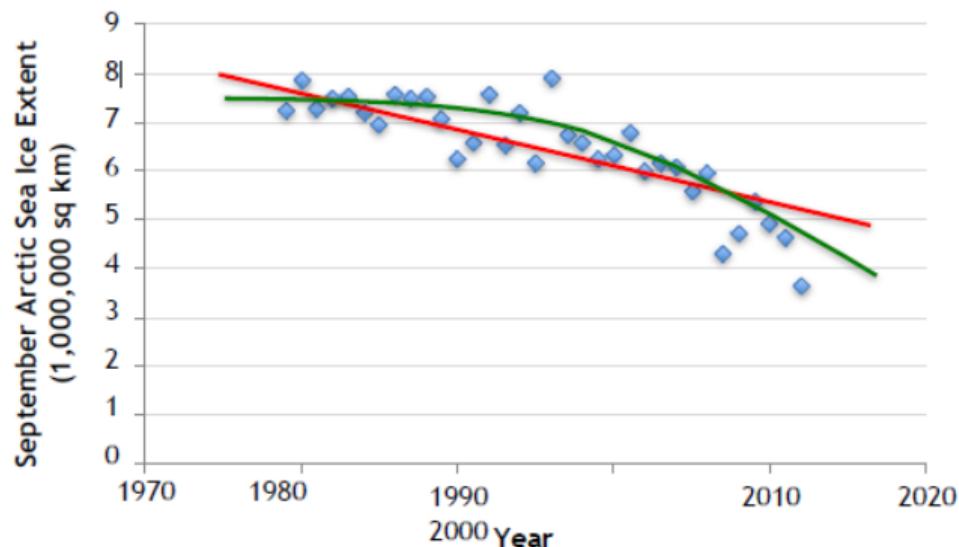
- Supervised (inductive) learning
 - ▶ Given: training data, desired outputs (labels)
- Unsupervised learning
 - ▶ Given: training data (without desired outputs)

SUPERVISED LEARNING

- Desired output is already known.
- Given $(x_1, y_1); (x_2, y_2); \dots; (x_n, y_n)$.
- A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples.
- Learn a function $f(x)$ to predict y given x .
- Example: pattern association, classification, regression.

SUPERVISED LEARNING

- Regression
 - y is real valued (continuous).

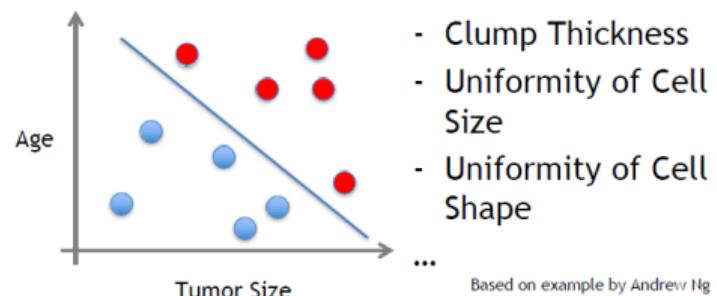
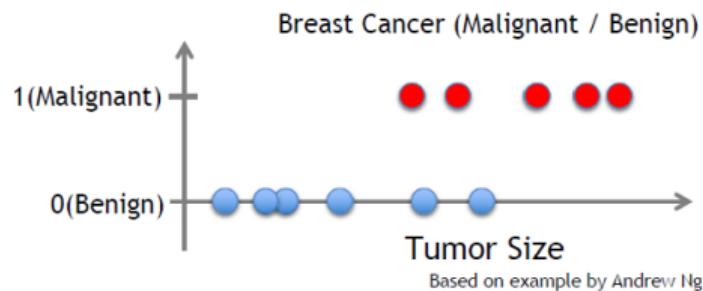


Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

SUPERVISED LEARNING

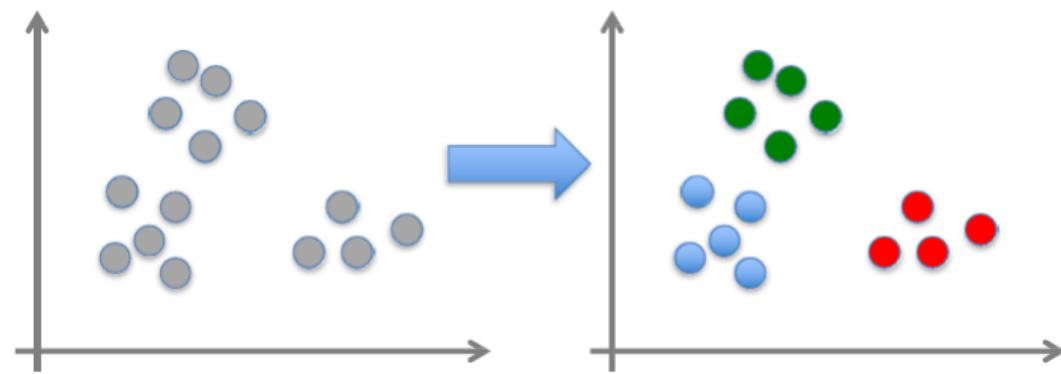
- Classification

- y is discrete (integer).



UNSUPERVISED LEARNING

- No target outputs.
- Given x_1, x_2, \dots, x_n .
- The goal is to group similar units close together in certain areas of the value range.
- Output the hidden structure begin the input x .
- Example: clustering



TRAIN / VALIDATE / TEST SETS

- Training set
 - ▶ Approx 70 to 90% of the **actual** dataset is used for training the algorithm.
 - ▶ Used to learn the parameters of the model.
- Validation set
 - ▶ Approx 10 to 20% of the **training** dataset is used for validating the algorithm.
 - ▶ Used to tune the parameters of the model.
- Testing set
 - ▶ Approx 10 to 20% of the **actual** dataset is used for testing the algorithm.
 - ▶ Used to test against new data.



TABLE OF CONTENTS

- 1 CLASSIFICATION AND PREDICTION
- 2 CLASSIFICATION
- 3 DECISION TREE ALGORITHM
- 4 OCCAM's RAZOR
- 5 EVALUATION OF CLASSIFICATION TECHNIQUES

CLASSIFICATION

- Predict categorical (discrete, unordered) class labels.
- Learn of a mapping function, $y = f(x)$ that can predict the associated class label y of a given tuple X .
- In general, mapping is represented in the form of classification rules, decision trees, or mathematical formulae.

CLASSIFICATION ALGORITHMS

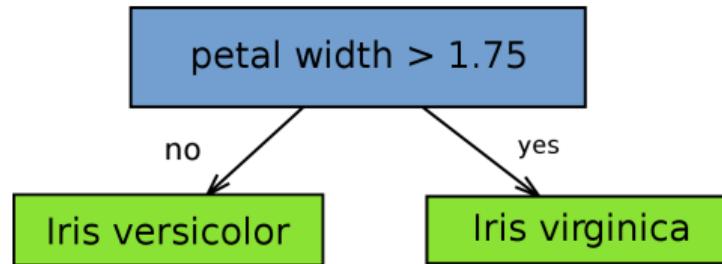
- Decision Tree
- Random Forest [Discuss in ML course]
- Logistic Regression [Discuss in ML course]
- Naive Bayes Classifier [Discuss in ML course]
- Support Vector Machine [Discuss in ML course]
- Neural Network [Discuss in DL course]

TABLE OF CONTENTS

- 1 CLASSIFICATION AND PREDICTION
- 2 CLASSIFICATION
- 3 DECISION TREE ALGORITHM
- 4 OCCAM's RAZOR
- 5 EVALUATION OF CLASSIFICATION TECHNIQUES

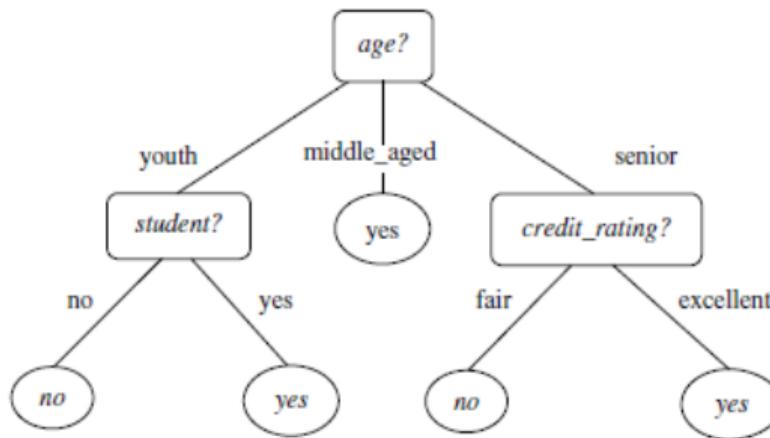
DECISION TREE

- Learning of decision trees from class-labelled training data.
- Decision Tree
 - ▶ Flowchart-like tree structure.
 - ▶ The topmost node in a tree is the root node.
 - ▶ Each internal node denotes a test on an attribute.
 - ▶ Each branch represents an outcome of the test.
 - ▶ Each leaf node holds a class label.



DECISION TREE

- Given a tuple for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree.
- A path is traced from the root to a leaf node, which holds the class prediction for that tuple.



DECISION TREE CLASSIFIERS

Advantages

- Decision trees can easily be converted to classification rules.
- Does not require any domain knowledge or parameter setting.
- Decision trees can handle multidimensional data.
- Simple, fast, good accuracy

Applications

- Medicine
- Manufacturing and production
- Financial analysis
- Astronomy
- Molecular biology

DECISION TREE GENERATION

2 Steps

- Tree construction
 - ▶ At start, all the training examples are at the root.
 - ▶ Partition examples recursively based on selected attributes.
- Tree pruning
 - ▶ Identify and remove branches that reflect noise or outliers.

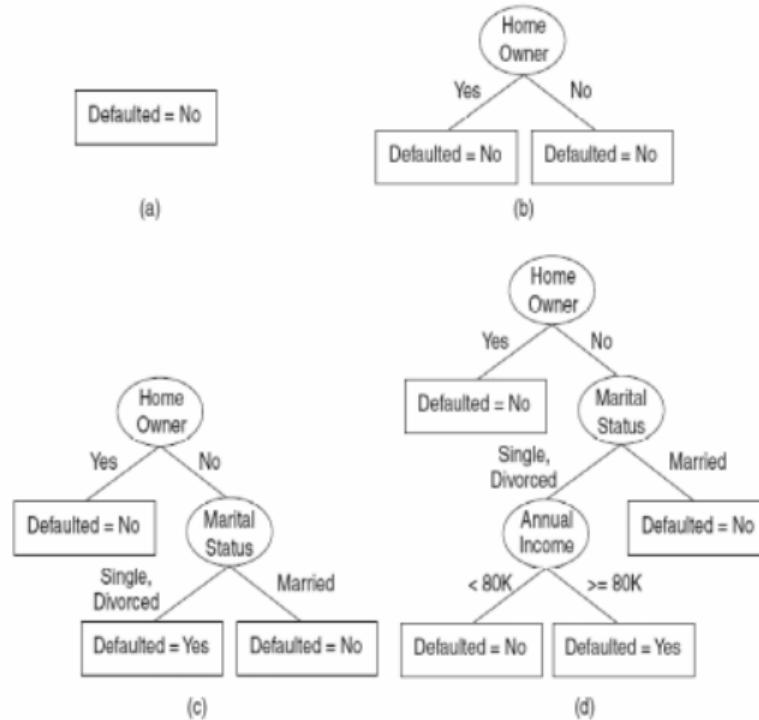
DECISION TREE CONSTRUCTION: HUNT'S ALGORITHM

- Construct a tree T from a training set D .
- If all the records in D belong to class C or if D is sufficiently pure, then the node is a leaf node and assigned class label C .
- Purity of a node is defined as the probability of corresponding class.
- If an attribute A does not partition D in a sufficiently pure manner, then choose another attribute A' and partition D according to A' values.
- Recursively construct tree and sub-trees until
 - ▶ All leaf nodes satisfy the minimum purity threshold.
 - ▶ Tree cannot be further split.
 - ▶ Maximum depth of tree is achieved.

DECISION TREE CONSTRUCTION EXAMPLE

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

binary categorical continuous class

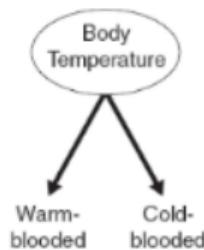


DESIGN DECISIONS

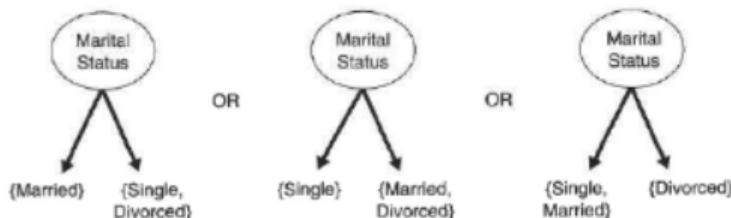
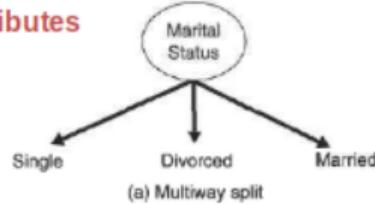
- How should the training examples be split?
- How should the splitting procedure stop?

DESIGN DECISIONS - SPLITTING METHODS EXAMPLE

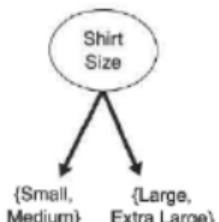
Binary Attributes



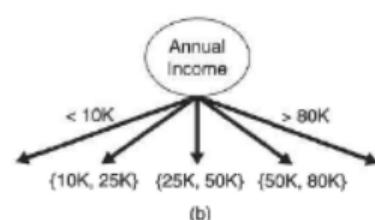
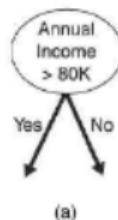
Nominal Attributes



Ordinal Attributes



Continuous Attributes



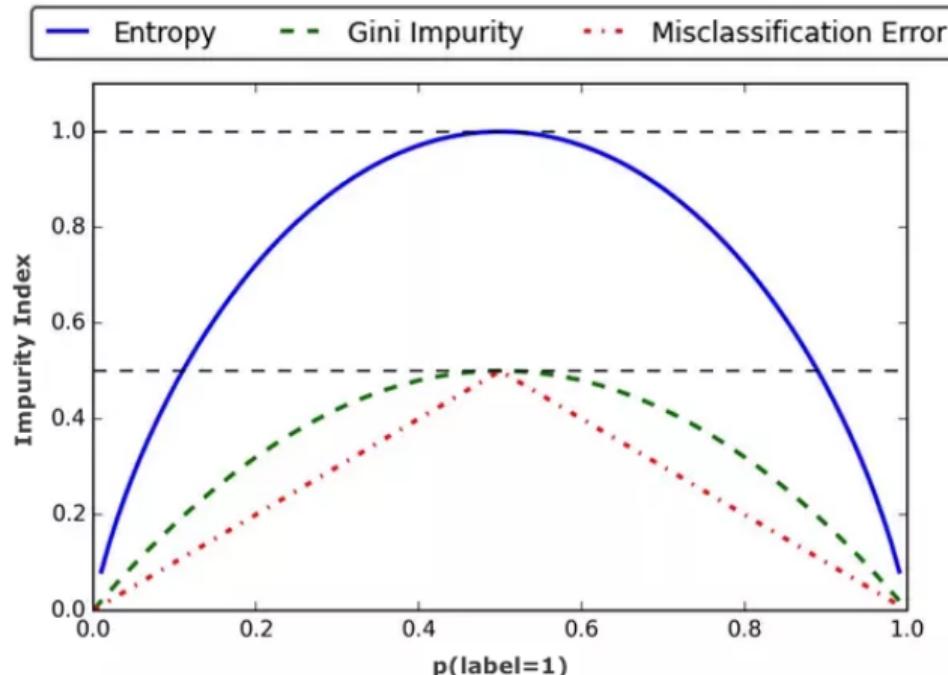
DESIGN DECISION - SELECTING THE BEST SPLIT

- $p(c | t)$: fraction of examples associated with node t belonging to class c .
- Best split is selected based on the degree of impurity of the child nodes.
- Class distribution (0,1) has highest purity.
- Class distribution (0.5,0.5) has lowest purity (highest impurity).
- Intuition: high purity \rightarrow small value of impurity measures \rightarrow better split
- Splitting criteria can be derived using most informative attribute obtained using the following metrics.

$$\text{Entropy}(t) = - \sum_{c=1}^C p(c | t) \log p(c | t)$$

$$\text{Gini}(t) = 1 - \sum_{c=1}^C [p(c | t)]^2$$

COMPARING IMPURITY MEASURES FOR BINARY CLASSIFICATION



DESIGN DECISION - SELECTING THE BEST SPLIT

Node N_1	Count
Class = 0	0
Class = 1	6

$$Gini = 1 - (0/6)^2 - (6/6)^2 = 0$$

$$Entropy = -(0/6) \log_2(0/6) - (6/6) \log_2(6/6) = 0$$

$$Error = 1 - \max[0/6, 6/6] = 0$$

Node N_2	Count
Class = 0	1
Class = 1	5

$$Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

$$Entropy = -(1/6) \log_2(1/6) - (5/6) \log_2(5/6) = 0.65$$

$$Error = 1 - \max[1/6, 5/6] = 0.167$$

Node N_3	Count
Class = 0	3
Class = 1	3

$$Gini = ?$$

$$Entropy = ?$$

$$Error = ?$$

DECISION TREE ALGORITHMS

- Iterative Dichotomiser 3 (ID 3)
 - ▶ Entropy based criteria
 - ▶ Gives an exhaustive decision tree.
 - ▶ Categorical inputs are handled.
- C 4.5
 - ▶ Entropy based criteria
 - ▶ Handle missing data.
 - ▶ Categorical and continuous inputs are handled.
 - ▶ Uses Tree Pruning to addresses over-fitting problem of ID 3.
- CART (Classification and Regression Tree)
 - ▶ Gini Index is used.
 - ▶ Categorical and continuous inputs are handled.

CHARACTERISTICS OF DECISION TREE INDUCTION

- Non-parametric approach
- Computationally inexpensive, even with large training set
- Easy to interpret
- Accuracy is comparable to other classifiers
- Robust to noise, with methods to prevent overfitting
- Immune to presence of redundant or irrelevant attributes
- Splits using single attribute at a time i.e rectilinear decision boundaries
- Limits decision tree representation for modelling complex relationships among continuous attributes.
- Tree pruning strategies has more effect on the performance of decision trees rather than choice of impurity measure.

ID 3 ALGORITHM

- ① Construct entropy for dataset.
- ② For every attribute
 - ① Calculate entropy for all categorical values.

$$\text{Entropy } H_x = - \sum_{\forall x \in X} P(x) \log_2 P(x)$$

- ② Take average information entropy for the current attribute.

$$\text{Information Entropy } H_{Y|X} = - \sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) \log_2 P(y|x)$$

- ③ Calculate gain for the current attribute.

$$\text{Information gain } IG_A = H_D - H_{D|A}$$

- ④ Pick the highest gain attribute.
- ⑤ Repeat until the desired tree is obtained.

DECISION TREE - INFORMATION GAIN

Construct a decision tree for the given dataset using Information gain.

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

DECISION TREE - INFORMATION GAIN

Step 1: Calculate entropy of the target.

Play Golf	
Yes	No
9	5

$$\begin{aligned}H(D) &= \text{Entropy}(PlayGolf) \\&= - \sum_{\forall x \in X} P(x) \log_2 P(x) \\&= \frac{-9}{14} \log \frac{9}{14} + \frac{-5}{14} \log \frac{5}{14} \\&= 0.94\end{aligned}$$

DECISION TREE - INFORMATION GAIN

Step 2 : Calculate entropy and Information gain of each attribute.

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5

$$\text{Entropy}(\text{Sunny}) = \frac{-2}{5} \log \frac{2}{5} + \frac{-3}{5} \log \frac{3}{5} = 0.971$$

$$\text{Entropy}(\text{Overcast}) = \frac{-4}{4} \log \frac{4}{4} + \frac{-0}{4} \log \frac{0}{4} = 0$$

$$\text{Entropy}(\text{Rainy}) = \frac{-3}{5} \log \frac{3}{5} + \frac{-2}{5} \log \frac{2}{5} = 0.971$$

DECISION TREE - INFORMATION GAIN

$$H(D|A) = - \sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) \log_2 P(y|x)$$

$$\begin{aligned} H(PlayGolf|Outlook) &= \frac{5}{14} * 0.971 + 0 + \frac{5}{14} * 0.971 \\ &= 0.693 \end{aligned}$$

$$\begin{aligned} Gain(Outlook) &= H(D) - H(D|A) \\ &= 0.94 - 0.693 \\ &= 0.247 \end{aligned}$$

DECISION TREE - INFORMATION GAIN

Step 2b

		Play Golf		
		Yes	No	
Temperature	Hot	2	2	4
	Mild	4	2	6
	Cool	3	1	4

$$\text{Entropy}(\text{Hot}) = \frac{-2}{4} \log \frac{2}{4} + \frac{-2}{4} \log \frac{2}{4} = 1$$

$$\text{Entropy}(\text{Mild}) = \frac{-4}{6} \log \frac{4}{6} + \frac{-2}{6} \log \frac{2}{6} = 0.92$$

$$\text{Entropy}(\text{Cool}) = \frac{-3}{4} \log \frac{3}{4} + \frac{-1}{4} \log \frac{1}{4} = 0.81$$

DECISION TREE - INFORMATION GAIN

$$\begin{aligned}H(\text{PlayGolf} | \text{Temperature}) &= \frac{4}{14} * 1 + \frac{6}{14} * 0.92 + \frac{4}{14} * 0.81 \\&= 0.911\end{aligned}$$

$$\begin{aligned}\text{Gain}(\text{Temperature}) &= H(D) - H(D|A) \\&= 0.94 - 0.911 \\&= 0.029\end{aligned}$$

DECISION TREE - INFORMATION GAIN

Step 2c

		Play Golf		
		Yes	No	
Humidity	High	3	4	7
	Normal	6	1	7

$$\text{Entropy}(\text{High}) = \frac{-3}{7} \log \frac{3}{7} + \frac{-4}{7} \log \frac{4}{7} = 0.985$$

$$\text{Entropy}(\text{Normal}) = \frac{-6}{7} \log \frac{6}{7} + \frac{-1}{7} \log \frac{1}{7} = 0.59$$

$$H(\text{PlayGolf}|\text{Humidity}) = \frac{7}{14} * 0.985 + \frac{7}{14} * 0.59 = 0.785$$

$$\text{Gain}(\text{Humidity}) = 0.94 - 0.785 = 0.155$$

DECISION TREE - INFORMATION GAIN

Step 2d

		Play Golf		
		Yes	No	
Windy	False	6	2	8
	True	3	3	6

$$\text{Entropy}(\text{False}) = \frac{-6}{8} \log \frac{6}{8} + \frac{-2}{8} \log \frac{2}{8} = 0.811$$

$$\text{Entropy}(\text{True}) = \frac{-3}{6} \log \frac{3}{6} + \frac{-3}{6} \log \frac{3}{6} = 1$$

$$H(\text{PlayGolf}|\text{Windy}) = \frac{8}{14} * 0.811 + \frac{6}{14} * 1 = 0.892$$

$$\text{Gain}(\text{Windy}) = 0.94 - 0.892 = 0.048$$

DECISION TREE - INFORMATION GAIN

Step 2e: Compare Information Gain

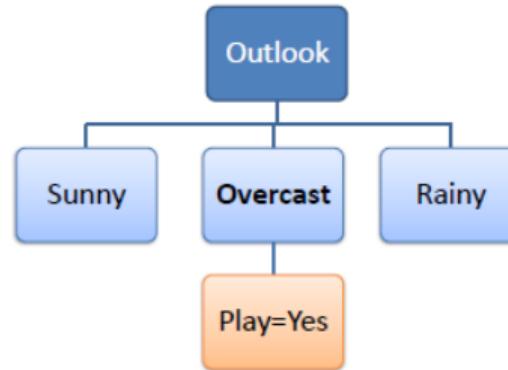
Attribute	Gain
Outlook	0.247
Temperature	0.029
Humidity	0.155
Windy	0.048

Outlook has the max gain.

DECISION TREE - INFORMATION GAIN

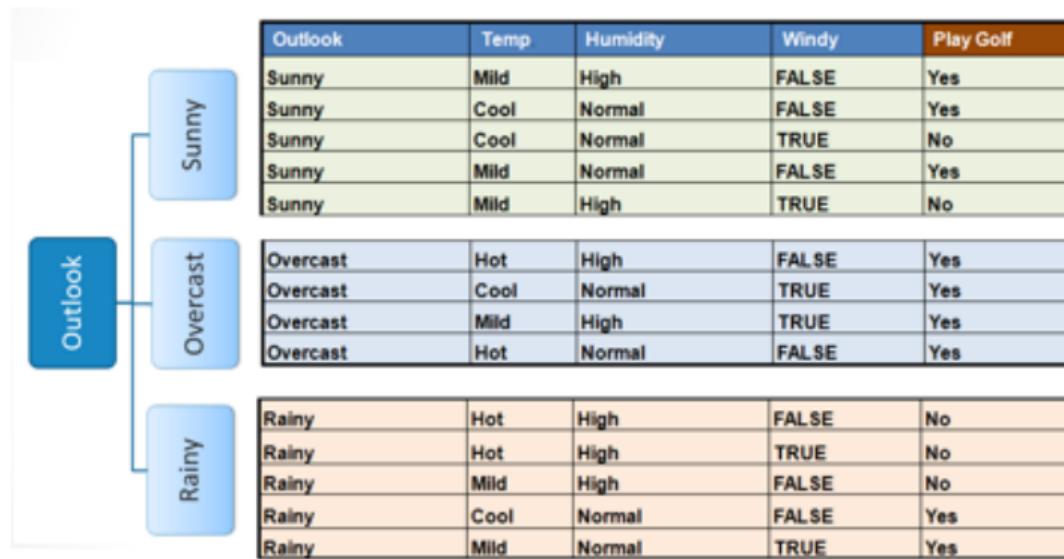
Step 3:

- Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.
- Outlook has the max gain. So choose Outlook as the root decision point.
- A branch with entropy of 0 is a leaf node.
- A branch with entropy more than 0 needs further splitting.



DECISION TREE - INFORMATION GAIN

Step 4: Repeat



DECISION TREE - INFORMATION GAIN

Step 4a: Consider only rows which have Outlook = Sunny.

$$\text{Entropy}(\text{Sunny}) = \frac{-2}{5} \log \frac{2}{5} + \frac{-3}{5} \log \frac{3}{5} = 0.971$$

$$\text{Entropy}(\text{Windy} = \text{False}) = \frac{-3}{3} \log \frac{3}{3} + \frac{0}{3} \log \frac{0}{3} = 0$$

$$\text{Entropy}(\text{Windy} = \text{True}) = \frac{-0}{2} \log \frac{0}{2} + \frac{-2}{2} \log \frac{2}{2} = 0$$

$$H(\text{PlayGolf}|\text{Sunny}, \text{Windy}) = \frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$$

$$\text{Gain}(\text{PlayGolf}|\text{Sunny}, \text{Windy}) = 0.971 - 0 = 0.971$$

Windy has max the max gain. Choose Windy as the next decision point.

DECISION TREE - INFORMATION GAIN

Step 4b: Consider only rows which have Outlook = Rainy.

$$\text{Entropy}(\text{Rainy}) = \frac{-2}{5} \log \frac{2}{5} + \frac{-3}{5} \log \frac{3}{5} = 0.971$$

$$\text{Entropy}(\text{Humidity} = \text{Normal}) = \frac{-3}{3} \log \frac{3}{3} + \frac{0}{3} \log \frac{0}{3} = 0$$

$$\text{Entropy}(\text{Humidity} = \text{High}) = \frac{-0}{2} \log \frac{0}{2} + \frac{-2}{2} \log \frac{2}{2} = 0$$

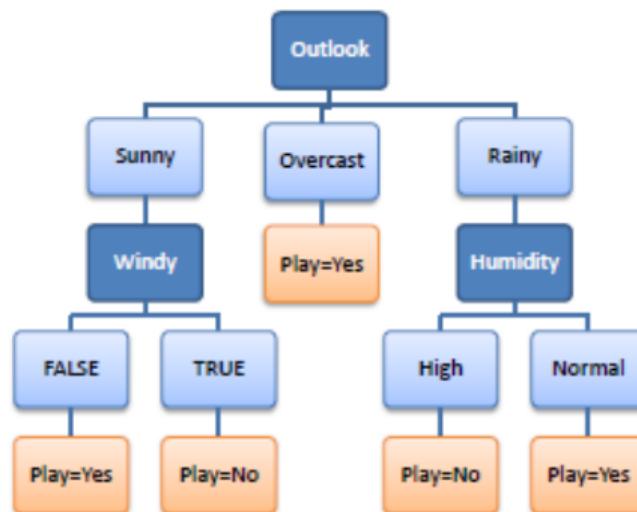
$$H(\text{PlayGolf} | \text{Rainy}, \text{Humidity}) = \frac{3}{5} * 0 + \frac{2}{5} * 0 = 0$$

$$\text{Gain}(\text{PlayGolf} | \text{Rainy}, \text{Humidity}) = 0.971 - 0 = 0.971$$

Humidity has the max gain. Choose Humidity as the next decision point.

DECISION TREE - INFORMATION GAIN

Step 5: Draw the final decision tree.



CART ALGORITHM

- ① Construct Gini index for dataset.
- ② For every attribute
 - ① Calculate Gini index for all categorical values.

$$Gini_x = 1 - \sum_{\forall x \in X} P(x)^2$$

- ② Calculate Gini gain for the current attribute.
- ③ Pick the highest Gini gain attribute.
- ④ Repeat until the desired tree is obtained.

PROBLEMS WITH INFORMATION GAIN

- Natural bias of information gain is that it favours attributes with many possible values.
- The problem is that the partition is too specific, too many small classes are generated.
- We need to look at alternative measures.

DECISION TREE - GINI INDEX

Construct a decision tree for the given dataset using Gini Index.

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

DECISION TREE - GINI INDEX

Step 1: Calculate the Gini index for data-set.

Play Golf	
Yes	No
9	5

$Gini(PlayGolf)$

$$= 1 - \sum_{\forall x \in X} P(x)^2$$

$$= 1 - \left(\frac{-9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$= 0.46$$

DECISION TREE - GINI INDEX

Step 2a : Calculate entropy and Gini gain of each attribute.

Outlook	Play Golf			Computation	Gini Index
	Yes	No	Total		
Sunny	3	2	5	$1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2$	0.48
Overcast	4	0	4	$1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$	0
Rainy	2	3	5	$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$	0.48
Wtd Gini				$\frac{5}{14} * 0.48 + 0 + \frac{5}{14} * 0.48$	0.342

DECISION TREE - GINI INDEX

Step 2b : Calculate entropy and Gini gain of each attribute.

Temperature	Play Golf			Computation	Gini Index
	Yes	No	Total		
Hot	2	2	4	$1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$	0.5
Mild	4	2	6	$1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$	0.44
Cool	3	1	4	$1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$	0.375
Wtd Gini				$\frac{4}{14} * 0.5 + \frac{6}{14} * 0.44 + \frac{4}{14} * 0.375$	0.438

DECISION TREE - GINI INDEX

Step 2c : Calculate entropy and Gini gain of each attribute.

Humidity	Play Golf			Computation	Gini Index
	Yes	No	Total		
High	3	4	7	$1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$	0.489
Normal	6	1	7	$1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2$	0.245
Wtd Gini				$\frac{7}{14} * 0.489 + \frac{7}{14} * 0.245$	0.367

DECISION TREE - GINI INDEX

Step 2d : Calculate entropy and Gini gain of each attribute.

Windy	Play Golf			Computation	Gini Index
	Yes	No	Total		
False	6	2	8	$1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2$	0.375
True	3	3	6	$1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2$	0.5
Wtd Gini				$\frac{8}{14} * 0.375 + \frac{6}{14} * 0.5$	0.428

DECISION TREE - GINI INDEX

Step 2e: Compare Gini Gain

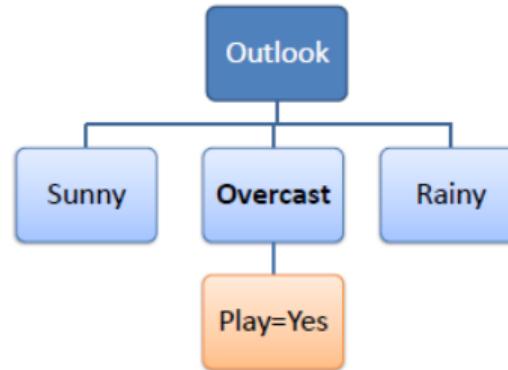
Attribute	Wtd Gini Index	Gini Gain
Outlook	0.342	$0.46 - 0.342 = 0.118$
Temperature	0.438	$0.46 - 0.438 = 0.022$
Humidity	0.367	$0.46 - 0.367 = 0.093$
Windy	0.44	$0.46 - 0.428 = 0.032$

Outlook has the max Gini gain.

DECISION TREE - GINI INDEX

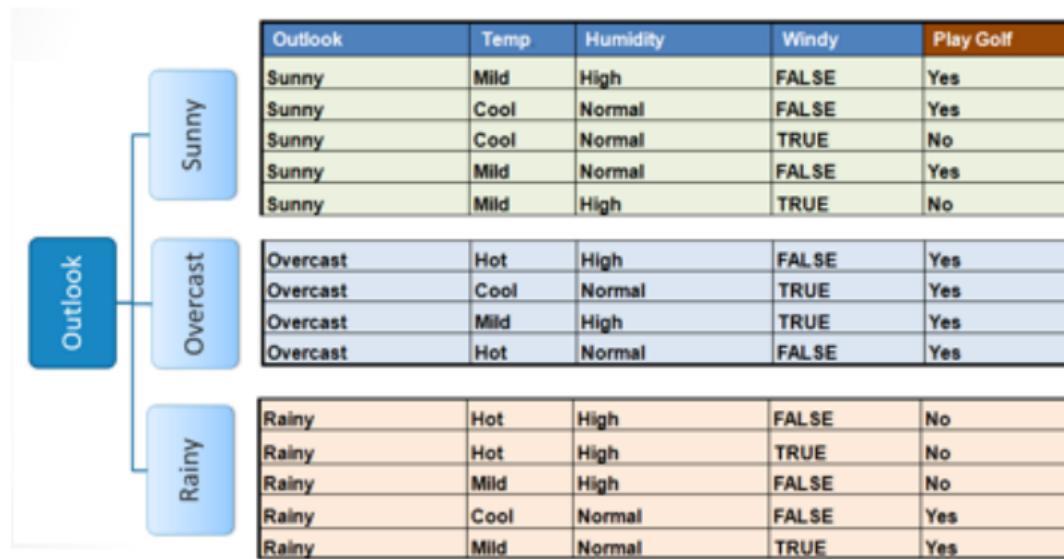
Step 3:

- Choose attribute with the largest Gini gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.
- Outlook has the max Gini gain. So choose Outlook as the root decision point.
- A branch with Gini index of 0 is a leaf node.
- A branch with Gini index more than 0 needs further splitting.



DECISION TREE - GINI INDEX

Step 4: Repeat



DECISION TREE - GINI INDEX

Step 4a: Consider only rows which have Outlook = Sunny.

$$Gini(Sunny) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\begin{aligned} Gini(Humidity) &= \left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 + 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \right] / 2 \\ &= 0.53 \end{aligned}$$

$$Gini(Windy) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

Windy has max Gini gain.

Choose Windy as the next decision point.

DECISION TREE - GINI INDEX

Step 4b: Consider only rows which have Outlook = Rainy.

$$Gini(Rainy) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\begin{aligned} Gini(Windy) &= \left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 + 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \right] / 2 \\ &= 0.53 \end{aligned}$$

$$Gini(Humidity) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

Humidity has max Gini gain.

Choose Humidity as the next decision point.

DECISION TREE - GINI INDEX

Step 5: Draw the final decision tree.

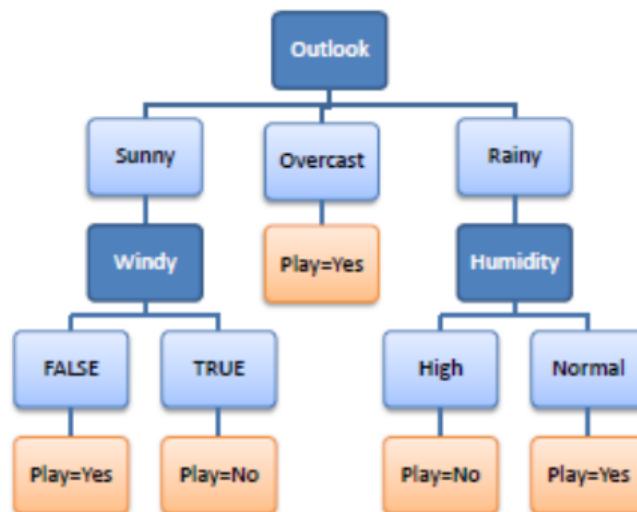


TABLE OF CONTENTS

- 1 CLASSIFICATION AND PREDICTION
- 2 CLASSIFICATION
- 3 DECISION TREE ALGORITHM
- 4 OCCAM's RAZOR
- 5 EVALUATION OF CLASSIFICATION TECHNIQUES

OCCAM'S RAZOR



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."¹

1

¹Courtesy:www.phdcomics.com

OCCAM'S RAZOR

PRINCIPLE OF OCCAM'S RAZOR

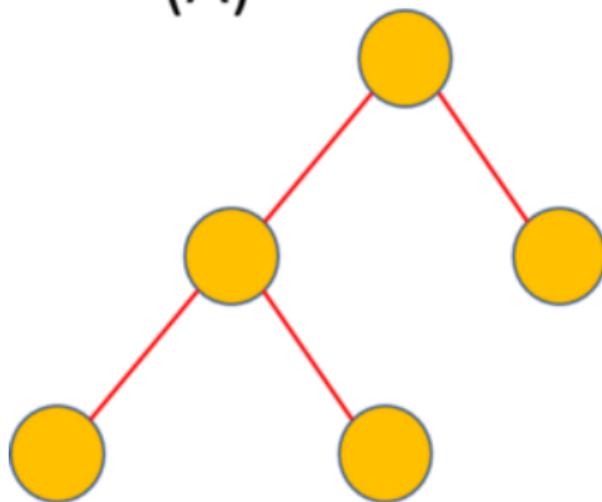
Simpler trees are better.

How to achieve?

- Early stopping for learning decision trees
 - ▶ Limit tree depth: Stop splitting after a certain depth. Stop tree building when depth = max depth.
 - ▶ Minimum node size: Do not split an intermediate node which contains too few data points.
- Pruning: Simplify the tree after the learning algorithm terminates
 - ▶ Simple measure of complexity of tree is number of leaf nodes. Limit the number of leaf nodes.

OCCAM'S RAZOR

(A)



(B)

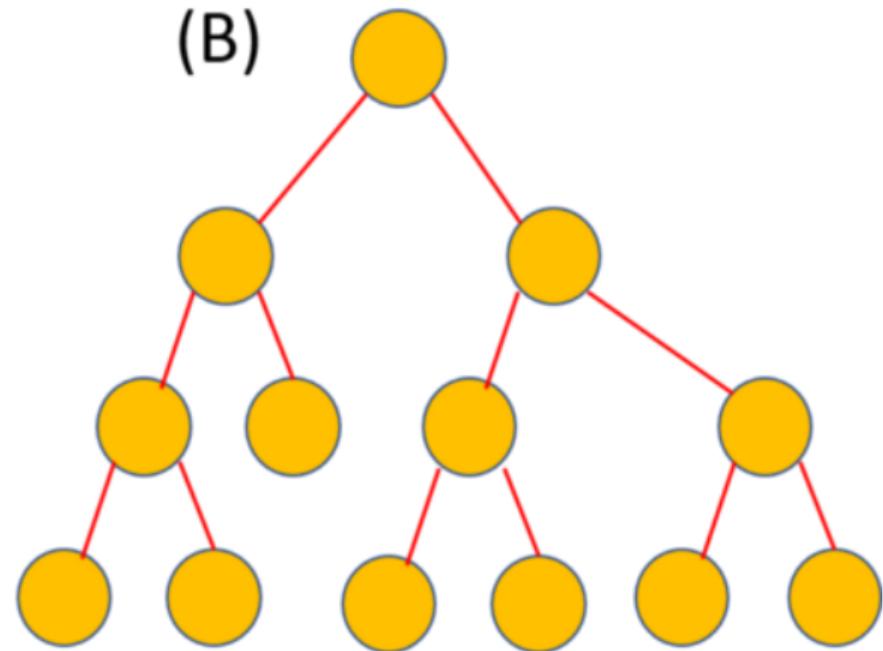


TABLE OF CONTENTS

- 1 CLASSIFICATION AND PREDICTION
- 2 CLASSIFICATION
- 3 DECISION TREE ALGORITHM
- 4 OCCAM's RAZOR
- 5 EVALUATION OF CLASSIFICATION TECHNIQUES

EVALUATION OF CLASSIFIERS

- Evaluation tools to measure how well a classifier has performed.
 - Compare the performances of multiple classifiers.
- ① Confusion Matrix
 - ② Accuracy
 - ③ Precision
 - ④ Recall
 - ⑤ Receiver Operating Characteristic (ROC) Curves
 - ⑥ Area under the curve (AUC)

EVALUATION OF CLASSIFIERS

- ① True positives $TP \approx \text{large}$:
 - ▶ Positive tuples that were correctly labeled by the classifier.
- ② True negatives $TN \approx \text{large}$:
 - ▶ Negative tuples that were correctly labeled by the classifier.
- ③ False positives $FP \approx 0$:
 - ▶ Negative tuples that were incorrectly labeled as positive.
- ④ False negatives $FN \approx 0$:
 - ▶ Positive tuples that were mislabeled as negative.

CONFUSION MATRIX

<p>True positives (TP) (images that contain people)</p> 	<p>Predicted negatives (FN) (images predicted not to contain people)</p> 
<p>Predicted positives (FP) (images predicted to contain people)</p> 	<p>True negatives (TN) (images that do not contain people)</p> 

CONFUSION MATRIX

- Evaluate the performance of a classifier by using a **confusion matrix**.
- **Confusion matrix** is a specific table that allows visualization of the performance of a classifier.
- TP and TN tells when the classifier is getting things right.
- FP and FN tell us when the classifier is getting things wrong.

Actual Class	Predicted Class		Total
	Positive	Negative	
Positive	True Positives (TP)	False Negatives (FN)	P
Negative	False Positives (FP)	True Negatives (TN)	N

CONFUSION MATRIX EXAMPLE

Actual Class	Predicted Class		Total
	Subscribed	Not Subscribed	
Subscribed	45	5	50
Not Subscribed	2	48	50
Total	47	53	100

ACCURACY

- Accuracy or Overall Success Rate of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.
- It is defined as the sum of TP and TN divided by total number of examples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

- A good classifier has a high accuracy score.

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

ISSUES WITH ACCURACY

- Consider a 2-class problem
 - Number of Class NO examples = 990
 - Number of Class YES examples = 10
 - This is called Imbalanced class.
 - Application of imbalance class is detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)
- If a model predicts everything to be class NO, accuracy is $990/1000 = 99\%$
 - This is misleading because this trivial model does not detect any class YES example.

Actual Class	Predicted Class	
	Yes	No
Yes	0	10
No	0	990

ERROR RATE

- Error Rate or Misclassification Rate

$$\text{Error Rate} = 1 - \text{Accuracy}$$

TRUE POSITIVE RATE (TPR)

- True Positive Rate (TPR) shows the percentage of positive instances the classifier correctly identified.
- Also called Recall, Sensitivity.

$$TPR = \frac{TP}{TP + FN}$$

- A model that produces no false negatives has a recall of 1.0.

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

TRUE NEGATIVE RATE (TPR)

- True Negative Rate (TNR) shows the percentage of negative instances the classifier correctly identified.
- Also called Specificity.

$$TNR = \frac{TN}{FP + TN}$$

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

PRECISION

- Precision is the percentage of instances marked positive that really are positive.
- Also called Positive Predictive Value (PPV).

$$\text{Precision} = \frac{TP}{TP + FP}$$

- A model that produces no false positives has precision of 1.0.

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

FALSE POSITIVE RATE (FPR)

- False Positive Rate (FPR) shows the percentage of negative instances the classifier marked as positive.
- Also called False Alarm Rate, Fall-out or Type I Error Rate.

$$FPR = \frac{FP}{FP + TN}$$

$$FPR = 1 - specificity$$

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

FALSE NEGATIVE RATE (FNR)

- False Negative Rate (FNR) shows the percentage of positive instances the classifier marked as negative.
- Also called Miss Rate or Type II Error Rate.

$$FNR = \frac{FN}{TP + FN}$$

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

F-SCORE

- F-score is the harmonic mean of precision and recall.

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

$$= \frac{2TP}{2TP + FP + FN}$$

Actual Class	Predicted Class	
	Yes	No
Yes	TP	FN
No	FP	TN

- F-score of 1.0 indicates perfect precision and recall.

WHICH CLASSIFIER IS BETTER? NO SKEW

Actual Class	Predicted Class	
	T1	Yes
Yes	50	50
No	1	99

Actual Class	Predicted Class	
	T2	Yes
Yes	99	1
No	10	90

Actual Class	Predicted Class	
	T3	Yes
Yes	99	1
No	1	99

WHICH CLASSIFIER IS BETTER? MEDIUM SKEW

Actual Class	Predicted Class	
	T1	Yes
Yes	50	50
No	10	990

Actual Class	Predicted Class	
	T2	Yes
Yes	99	1
No	100	900

Actual Class	Predicted Class	
	T3	Yes
Yes	99	1
No	10	990

WHICH CLASSIFIER IS BETTER? HIGH SKEW

Actual Class	Predicted Class	
	T1	Yes
Yes	50	50
No	100	9900

Actual Class	Predicted Class	
	T2	Yes
Yes	99	1
No	1000	9000

Actual Class	Predicted Class	
	T3	Yes
Yes	99	1
No	100	9900

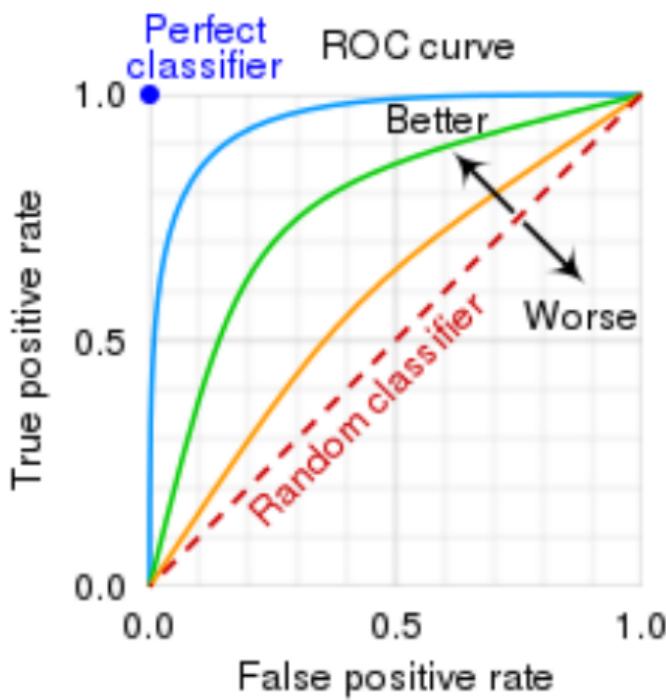
WHICH MODEL TO USE?

	False Positive Rate	False Negative Rate
Model 1	41%	3%
Model 2	5%	25%

- Mistakes have different costs:
 - ▶ Disease Screening – LOW FN Rate
 - ▶ Spam filtering – LOW FP Rate
- Conservative vs Aggressive settings:
 - ▶ The same application might need multiple tradeoffs.

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

- ROC curve evaluates the performance of a classifier based on TP and FP.
- ROC curve is a graph plotted using FPR on the horizontal axis and TPR on the vertical axis.
- (TPR, FPR)
 - ▶ (0,0) – classify everything as negative.
 - ▶ (1,1) – classify everything as positive.
 - ▶ (1,0) – ideal(top left corner)
- Diagonal line when $\text{TPR} = \text{FPR}$. Obtained by random guessing.

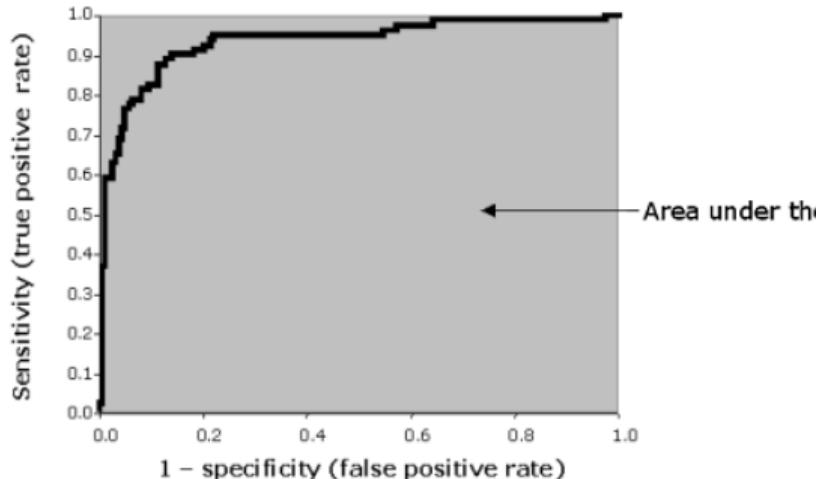


ROC CURVE

- ROC curve can be used to select a threshold for a classifier, which maximizes the true positives and in turn minimizes the false positives.
- ROC Curves help determine the exact trade-off between the true positive rate and false-positive rate for a model using different measures of probability thresholds.
- ROC curves are more appropriate to be used when the observations present are balanced between each class.

AREA UNDER THE CURVE (AUC)

- AUC is calculated by measuring the area under the ROC curve.
- Higher AUC scores mean the classifier performs better.
- The score can range from 0.5 to 1.0.



CONSTRUCTING ROC CURVE

Instance	$P(+ A)$	True Class
1	0.95	Y
2	0.93	Y
3	0.87	N
4	0.85	N
5	0.85	N
6	0.85	Y
7	0.76	N
8	0.53	Y
9	0.43	N
10	0.25	Y

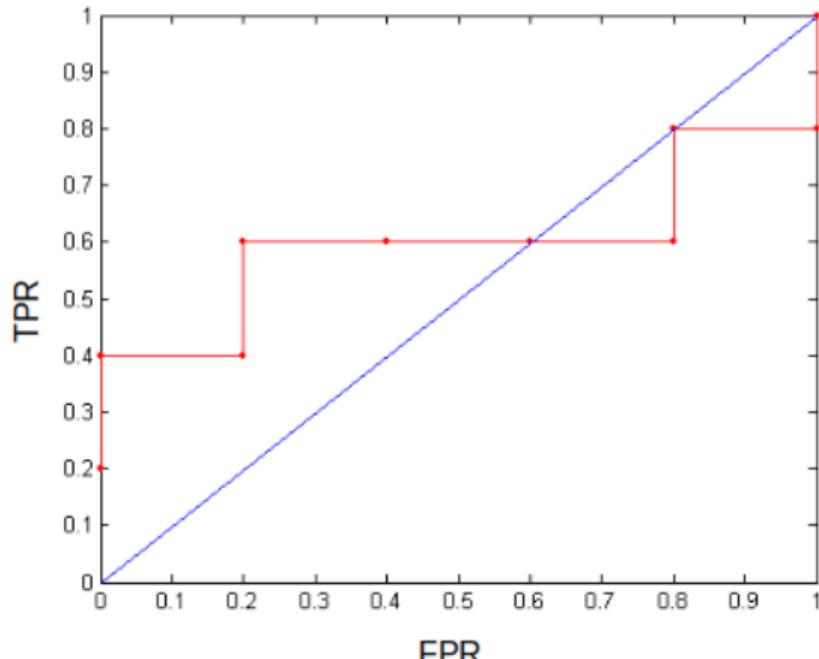
- Use classifier that produces probability for each test instance $P(+|A)$.
- Sort the instances according to $P(+|A)$ in decreasing order.
- Apply threshold at each unique value of $P(+|A)$.
- Count the number of TP, FP, TN, FN at each threshold.
 $\text{TP rate, TPR} = \text{TP}/(\text{TP}+\text{FN})$
 $\text{FP rate, FPR} = \text{FP}/(\text{FP} + \text{TN})$

CONSTRUCTING ROC CURVE

Class	Y	N	Y	N	N	N	Y	N	Y	Y	
Threshold \geq	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

CONSTRUCTING ROC CURVE

ROC Curve:



-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 7 : ASSOCIATION ANALYSIS

IDS Course Team

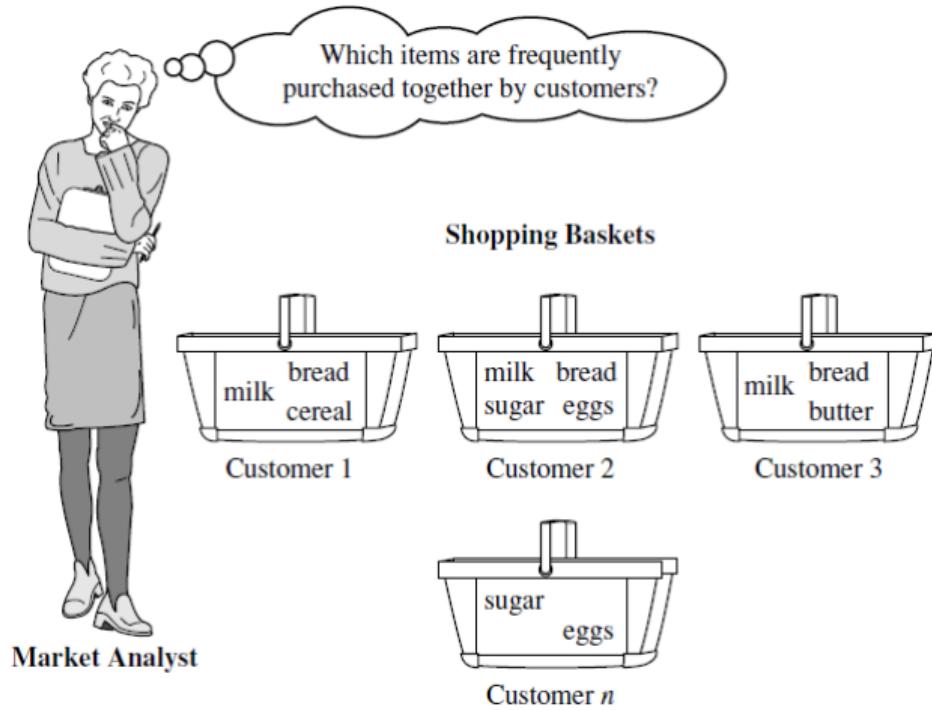
BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

- 1 ASSOCIATION ANALYSIS
- 2 APRIORI ALGORITHM
- 3 FP GROWTH ALGORITHM
- 4 PATTERN EVALUATION METHODS

MARKET BASKET ANALYSIS



MARKET BASKET ANALYSIS

- Frequent item-set mining.
- Analyze customer buying habits by finding associations between the different items that customers place in their shopping baskets.
- The discovery of interesting correlation relationships among huge amounts of business transaction records can help in business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.

FREQUENT PATTERNS

- Frequent patterns are patterns that appear frequently in a data set.
- Frequent item-set
 - ▶ A set of items that appear frequently together in a transaction data set is a frequent item-set.
 - ▶ Example – milk and bread
- Frequent sequential pattern
 - ▶ A sub-sequence that occurs frequently in a shopping history database is a frequent sequential pattern.
 - ▶ Example – buying first a PC, then a digital camera, and then a memory card
- Frequent structured pattern
 - ▶ Substructures can be sub-graphs, sub-trees, or sub-lattices combined with item-sets or sub-sequences.
 - ▶ If a substructure occurs frequently, it is called a frequent structured pattern.

MARKET BASKET ANALYSIS

- Each **item** can be represented by a Boolean variable representing the presence or absence of it.
- Each **basket** can then be represented by a Boolean vector of values assigned to these variables.
- The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.
- The **frequent patterns** can be represented in the form of **association rules**.

ASSOCIATION RULE

- An **association rule** is an implication of the form

$$A \implies B$$

where $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, $A \neq \phi$, $B \neq \phi$ and $A \cap B = \phi$

- The rule $A \implies B$ holds in the transaction set D with **support s**.
- The rule $A \implies B$ has **confidence c** in the transaction set D .
- Rules that satisfy both a minimum support threshold (*min_sup*) and a minimum confidence threshold (*min_conf*) are called strong.

SUPPORT OF A RULE

- Support s is the percentage of transactions in D that contain $A \cup B$.

$$\text{support}(A \implies B) = P(A \cup B)$$

- If 80 % of all transactions contain bread,

$$\text{support}\{\text{bread}\} = 0.8$$

- If 60 % of all transactions contain bread and butter,

$$\text{support}\{\text{bread, butter}\} = 0.6$$

CONFIDENCE OF A RULE

- The **confidence** c is the percentage of transactions in D containing A that also contain B .
- Confidence is the percentage of transaction that contain both A and B out of all the transactions that contain A .

$$\text{confidence}(A \implies B) = \frac{\text{SupportCount}(A \cup B)}{\text{SupportCount}(A)}$$

- Confidence is defined as the measure of certainty associated with each discovered rule.
- Confidence cannot tell if a rule contains implication of the relationship or if the rule is purely coincidental.

FREQUENT ITEM-SET

- A set of items is referred to as an item-set.
- An item-set that contains k items is a k -item-set.
- The occurrence frequency of an item-set is the number of transactions that contain the item-set. This is also known as the frequency, support count, or count of the item-set.
- If the support of an item-set \mathcal{I} satisfies a prespecified minimum support threshold then \mathcal{I} is a frequent item-set. It is denoted by \mathcal{L}_k .

CLOSED FREQUENT ITEM-SET

- An item-set X is **closed** in a data set D if there exists no proper super-item-set such that Y has the same support count as X in D .
Every item of X is contained in Y but there is at least one item of Y that is not in X .
- An item-set X is a **closed frequent item-set** in set D if X is both closed and frequent in D .
- An item-set X is a **maximal frequent item-set** (or max-item-set) in a data set D if X is frequent, and there exists no super-item-set Y such that $X \subset Y$ and Y is frequent in D .

ASSOCIATION RULE MINING

- Two step process
 - ① Find all frequent item-sets.

$$\text{frequency}(\text{item-sets}) > \text{min-sup}$$

- ② Generate strong association rules from the frequent item-sets

$$A \implies B \text{ with } \text{min-sup} \text{ and } \text{min-conf}$$

- Two approaches
 - ① Apriori Algorithm
 - ② FP Growth

FREQUENT ITEM-SETS GENERATION

- Brute-force approach
 - ▶ Each itemset in the lattice is a candidate frequent itemset.
 - ▶ Count the support of each candidate by scanning the database.
 - ▶ Match each transaction against every candidate.
 - ▶ Complexity is $O(NMw)$ –Expensive since $M = 2^d$.

FREQUENT ITEM-SETS GENERATION STRATEGIES

- Reduce the number of candidates M
 - ▶ Complete search: $M = 2^d$
 - ▶ Use pruning techniques to reduce M .
- Reduce the number of transactions N
 - ▶ Reduce size of N as the size of itemset increases.
 - ▶ Used by DHP and vertical-based mining algorithms
- Reduce the number of comparisons NM
 - ▶ Use efficient data structures to store the candidates or transactions.
 - ▶ No need to match every candidate against every transaction

TABLE OF CONTENTS

- 1 ASSOCIATION ANALYSIS
- 2 APRIORI ALGORITHM
- 3 FP GROWTH ALGORITHM
- 4 PATTERN EVALUATION METHODS

APRIORI ALGORITHM

- Algorithm uses prior knowledge of frequent item-set properties.
- Apriori employs an iterative approach known as a level-wise search, where k -item-sets are used to explore $k + 1$ -item-sets.
- Uses Apriori property.
- Assume minimum support threshold as 0.5.
- Iteration 1
 - ▶ Identify frequent 1-item-sets.
 - ▶ Prune or discard item-sets that have $support < 0.5$ (appear in less than 50% transactions).
- Iteration 2
 - ▶ Identify frequent 2-item-sets.
 - ▶ Prune or discard item-sets that have $support < 0.5$ (appear in less than 50% transactions).

APRIORI PROPERTY

- Apriori property states that if an item set is frequent then any subset of the frequent item-set must also be frequent.
- If $support\{a, b\} = 0.6$ then $support\{a\} = 0.6$ and $support\{b\} = 0.6$.
- Also known as [download closure property](#).

APRIORI ALGORITHM

- Let C_k be the set of candidate k item-sets.
- Let L_k be the set of k item set that satisfy minimum support.
- Let D be transaction data set.
- Let δ be minimum support threshold.
- Let N indicate Max length of an item set could reach.
- Apriori iteratively computes frequent item-sets L_{k+1} based on L_k .

APRIORI ALGORITHM

```
 $L_1 \leftarrow$  Frequent 1-item-set ;  
 $k \leftarrow 2$  ;  
while  $L_{k-1} \neq \phi$  do  
     $Temp \leftarrow$  candidateItemSet( $L_{k-1}$ ) ;  
     $C_k \leftarrow$  frequencyOfItemSet ( $Temp$ ) ;  
     $L_{k+1} \leftarrow$  candidates in  $C_{k+1}$  that satisfy  $\delta$ ;  
end  
 $k \leftarrow k + 1$ ;  
return  $\cup_k L_k$ 
```

GENERATING ASSOCIATION RULES

- Association rules can be generated as follows:
 - ▶ For each frequent item-set b , generate all nonempty subsets of b .
 - ▶ For every nonempty subset a of b , output the rule

$$a \implies (b - a) \text{ if } \frac{\text{sup_count}(b)}{\text{sup_count}(a)} \geq \text{min_conf}$$
- Because the rules are generated from frequent item-sets, each one automatically satisfies the minimum support.
- Strong association rules satisfy both minimum support and minimum confidence.

APRIORI ALGORITHM EXAMPLE

Apply Apriori algorithm for finding frequent item-sets in the following transaction dataset. Using the generated frequent item-sets mine the association rules. Assume that the minimum support count required is 2 and minimum confidence threshold is 70%.

TID	List of Items
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

APRIORI ALGORITHM EXAMPLE

- Step 1: Scan the transaction dataset D for count of each candidate and generate C_1 .
- Step 2: Compare candidate support count with minimum support count $min_sup = 2$ and generate L_1 .



Step 1: C_1

Item-set	Count
$\{ I_1 \}$	6
$\{ I_2 \}$	7
$\{ I_3 \}$	6
$\{ I_4 \}$	2
$\{ I_5 \}$	2

Step 2: L_1

Item-set	Count
$\{ I_1 \}$	6
$\{ I_2 \}$	7
$\{ I_3 \}$	6
$\{ I_4 \}$	2
$\{ I_5 \}$	2

APRIORI ALGORITHM EXAMPLE

- Step 3: Generate C_2 candidates from L_1 .
- Step 4: Scan D for count of each candidate and generate C_2 .
- Step 5: Compare candidate support count with $min_sup = 2$ and generate L_2 .

Step 3: C_2

Item-set
$\{ I_1, I_2 \}$
$\{ I_1, I_3 \}$
$\{ I_1, I_4 \}$
$\{ I_1, I_5 \}$
$\{ I_2, I_3 \}$
$\{ I_2, I_4 \}$
$\{ I_2, I_5 \}$
$\{ I_3, I_4 \}$
$\{ I_3, I_5 \}$
$\{ I_4, I_5 \}$

Step 4: C_2

Item-set	Count
$\{ I_1, I_2 \}$	4
$\{ I_1, I_3 \}$	4
$\{ I_1, I_4 \}$	1
$\{ I_1, I_5 \}$	2
$\{ I_2, I_3 \}$	4
$\{ I_2, I_4 \}$	2
$\{ I_2, I_5 \}$	2
$\{ I_3, I_4 \}$	0
$\{ I_3, I_5 \}$	1
$\{ I_4, I_5 \}$	0



Step 5: L_2

Item-set	Count
$\{ I_1, I_2 \}$	4
$\{ I_1, I_3 \}$	4
$\{ I_1, I_5 \}$	2
$\{ I_2, I_3 \}$	4
$\{ I_2, I_4 \}$	2
$\{ I_2, I_5 \}$	2



APRIORI ALGORITHM EXAMPLE

- Join:

$$\begin{aligned}C_3 &= L_2 \bowtie L_2 \\&= \{\{l_1, l_2\}, \{l_1, l_3\}, \{l_1, l_5\}, \{l_2, l_3\}, \{l_2, l_4\}, \{l_2, l_5\} \\&\quad \bowtie \{l_1, l_2\}, \{l_1, l_3\}, \{l_1, l_5\}, \{l_2, l_3\}, \{l_2, l_4\}, \{l_2, l_5\}\} \\&= \{\{l_1, l_2, l_3\}, \{l_1, l_2, l_5\}, \{l_1, l_3, l_5\}, \\&\quad \{l_2, l_3, l_4\}, \{l_2, l_3, l_5\}, \{l_2, l_4, l_5\}\}\end{aligned}$$

- Prune using the Apriori property: All nonempty subsets of a frequent item-set must also be frequent.
 - The 2-item subsets of $\{l_1, l_2, l_3\}$ are $\{l_1, l_2\}, \{l_1, l_3\}, \{l_1, l_5\}$. All 2-item subsets of $\{l_1, l_2, l_3\}$ are members of L_2 . Therefore, keep $\{l_1, l_2, l_3\}$ in C_3 .

APRIORI ALGORITHM EXAMPLE

- Prune using the Apriori property.
 - ▶ The 2-item subsets of $\{I_1, I_2, I_5\}$ are $\{I_1, I_2\}, \{I_1, I_5\}, \{I_2, I_5\}$. All 2-item subsets of $\{I_1, I_2, I_5\}$ are members of L_2 . Therefore, keep $\{I_1, I_2, I_5\}$ in C_3 .
 - ▶ The 2-item subsets of $\{I_1, I_3, I_5\}$ are $\{I_1, I_3\}, \{I_1, I_5\}, \{I_3, I_5\}$. $\{I_3, I_5\}$ is not a member of L_2 . So remove $\{I_1, I_3, I_5\}$ from C_3 .
 - ▶ The 2-item subsets of $\{I_2, I_3, I_4\}$ are $\{I_2, I_3\}, \{I_2, I_4\}, \{I_3, I_4\}$. $\{I_3, I_4\}$ is not a member of L_2 . So remove $\{I_2, I_3, I_4\}$ from C_3 .
 - ▶ The 2-item subsets of $\{I_2, I_3, I_5\}$ are $\{I_2, I_3\}, \{I_2, I_5\}, \{I_3, I_5\}$. $\{I_3, I_5\}$ is not a member of L_2 . So remove $\{I_2, I_3, I_5\}$ from C_3 .
 - ▶ The 2-item subsets of $\{I_2, I_4, I_5\}$ are $\{I_2, I_4\}, \{I_2, I_5\}, \{I_4, I_5\}$. $\{I_4, I_5\}$ is not a member of L_2 . So remove $\{I_2, I_4, I_5\}$ from C_3 .
- Therefore, $C_3 = \{I_1, I_2, I_3\}$ and $\{I_1, I_2, I_5\}$ after pruning.

APRIORI ALGORITHM EXAMPLE

- Step 6: Generate C_3 candidates from L_2 .
- Step 7: Scan D for count of each candidate and generate C_3 .
- Step 8: Compare candidate support count with $min_sup = 2$ and generate L_3 .

Step 6: C_3		Step 7: C_3		Step 8: L_3	
Item-set		Item-set	Count	Item-set	Count
{ I_1, I_2, I_3 }		{ I_1, I_2, I_3 }	2	{ I_1, I_2, I_3 }	2
{ I_1, I_2, I_5 }		{ I_1, I_2, I_5 }	2	{ I_1, I_2, I_5 }	2

- Step 9: Generate a candidate set of 4-item-sets,

$$C_4 = L_3 \bowtie L_3 = \{ I_1, I_2, I_3, I_5 \}$$

- Step 10: C_4 gets pruned because its subset $\{ I_2, I_3, I_5 \}$ is not frequent. Thus, $C_4 = \emptyset$, and the algorithm terminates.

ASSOCIATION RULE MINING USING APRIORI ALGORITHM

EXAMPLE

- Step 1: Frequent item-sets are $X_1 = \{I_1, I_2, I_3\}$ and $X_2 = \{I_1, I_2, I_5\}$
- Step 2: Association rules that can be generated from X_1 .

Step 2:

Rule	Confidence
$\{I_1, I_2\} \Rightarrow \{I_3\}$	$c = 2/4 = 50\%$
$\{I_1, I_3\} \Rightarrow \{I_2\}$	$c = 2/4 = 50\%$
$\{I_2, I_3\} \Rightarrow \{I_1\}$	$c = 2/4 = 50\%$
$\{I_1\} \Rightarrow \{I_2, I_3\}$	$c = 2/6 = 33\%$
$\{I_2\} \Rightarrow \{I_1, I_3\}$	$c = 2/7 = 29\%$
$\{I_3\} \Rightarrow \{I_1, I_2\}$	$c = 2/6 = 33\%$

ASSOCIATION RULE MINING USING APRIORI ALGORITHM

EXAMPLE

- Step 3: Association rules that can be generated from X_2 .

Step 3:

Rule	Confidence
$\{I_1, I_2\} \Rightarrow \{I_5\}$	$c = 2/4 = 50\%$
$\{I_1, I_5\} \Rightarrow \{I_2\}$	$c = 2/2 = 100\%$
$\{I_2, I_5\} \Rightarrow \{I_1\}$	$c = 2/2 = 100\%$
$\{I_1\} \Rightarrow \{I_2, I_5\}$	$c = 2/6 = 33\%$
$\{I_2\} \Rightarrow \{I_1, I_5\}$	$c = 2/7 = 29\%$
$\{I_5\} \Rightarrow \{I_1, I_2\}$	$c = 2/2 = 100\%$

ASSOCIATION RULE MINING USING APRIORI ALGORITHM

EXAMPLE

- Step 4: Strong association rules have minimum confidence threshold more than 70%

Step 4:

Rule	Confidence
$\{I_1, I_5\} \Rightarrow \{I_2\}$	$c = 2/2 = 100\%$
$\{I_2, I_5\} \Rightarrow \{I_1\}$	$c = 2/2 = 100\%$
$\{I_5\} \Rightarrow \{I_1, I_2\}$	$c = 2/2 = 100\%$

TABLE OF CONTENTS

- 1 ASSOCIATION ANALYSIS
- 2 APRIORI ALGORITHM
- 3 FP GROWTH ALGORITHM
- 4 PATTERN EVALUATION METHODS

FP GROWTH ALGORITHM

- Frequent Pattern Growth
- Divide and conquer approach
- Step 1
 - ▶ Compress the database representing frequent items into a **frequent pattern tree**, or FP-tree.
 - ▶ FP-tree retains the item-set association information.
- Step 2
 - ▶ Divide the compressed database into a set of conditional databases, each associated with one frequent item or **pattern fragment**.
 - ▶ Mine each database separately.
- Reduces the size of the data sets and the patterns to be searched.

FP-TREE CONSTRUCTION

$FPtree(D, min_sup)$

- ① Scan the transaction database D once.
- ② Collect F , the set of frequent items, and the support of each frequent item.
- ③ Sort F in support-descending order as L , the list of frequent items.
- ④ Create the root of an FP-tree T and label it as *null*.
- ⑤ For each transaction $Trans$ in D do the following:
 - ① Select the frequent items in $Trans$.
 - ② Sort them according to the order of F .
 - ③ Let the sorted frequent-item list in $Trans$ be $[p|P]$, where p is the first element and P is the remaining list.
 - ④ Call $insert_tree([p|P], T)$.

FP-TREE CONSTRUCTION

insert_tree([p|P], T)

- ① If T has a child N such that $N.item_name = p.item_name$
increment the count of N by 1.
- ② else
create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same $item_name$ via the node-link structure.
- ③ If P is nonempty,
call *insert_tree([p|P], T)* recursively.

FP-TREE MINING

FPgrowth(tree, α)

- ① If Tree contains a single path P then
- ② For each combination β of the nodes in the path P
 - ① Generate pattern $\beta \cup \alpha$ with support_count = minimum support count of nodes in β
- ③ else for each a_i in the header of tree
 - ① Generate pattern $\beta = a_i \cup \alpha$ with support_count = a_i support_count
 - ② construct β 's conditional pattern base and then β 's conditional FP tree $tree_\beta$
 - ③ if $tree_\beta \neq \phi$ then
call *FPgrowth(tree $_\beta$, β)*

ASSOCIATION RULE MINING USING FP GROWTH

Apply FP growth algorithm for finding frequent item-sets and generating the association rules in the following transaction dataset. Assume that the minimum support count required is 2.

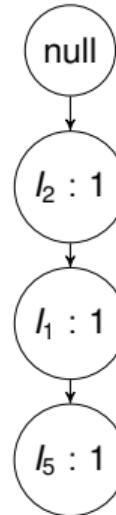
TID	List of Items
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

ASSOCIATION RULE MINING USING FP GROWTH

- Step 1: Generate frequent 1-item-sets and arrange them in descending order of support count.

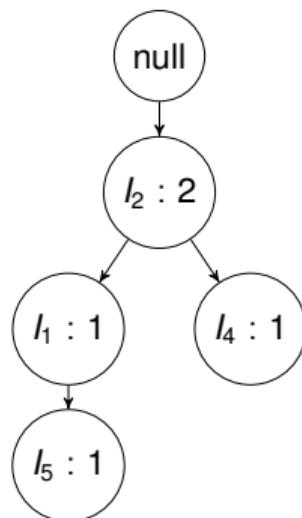
L	Item	support
$\{I_2\}$		7
$\{I_1\}$		6
$\{I_3\}$		6
$\{I_4\}$		2
$\{I_5\}$		2

- Step 2: Construct FP-tree.
- For transaction $T_1 : \{I_2, I_3, I_5\}$

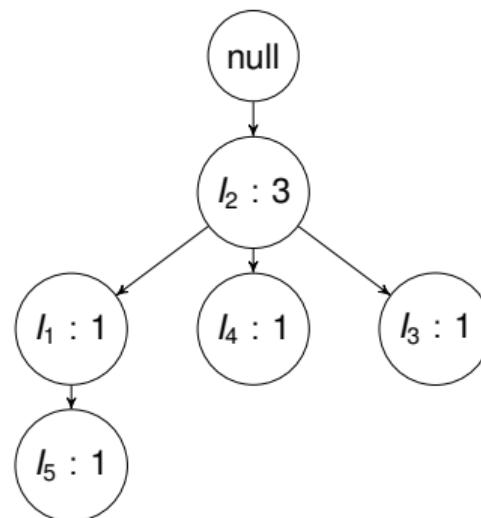


ASSOCIATION RULE MINING USING FP GROWTH

For transaction $T_2 : \{l_2, l_4\}$

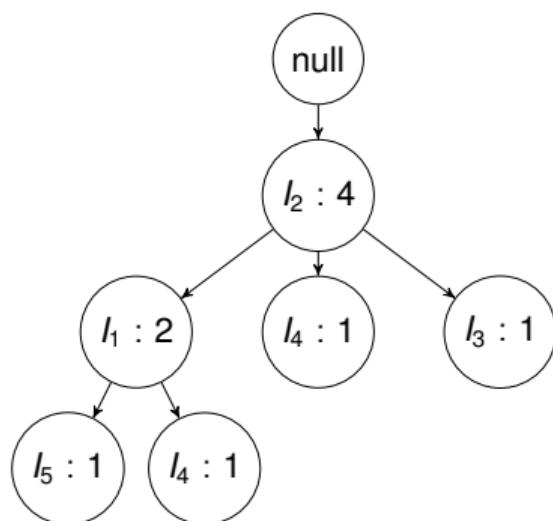


For transaction $T_3 : \{l_2, l_3\}$

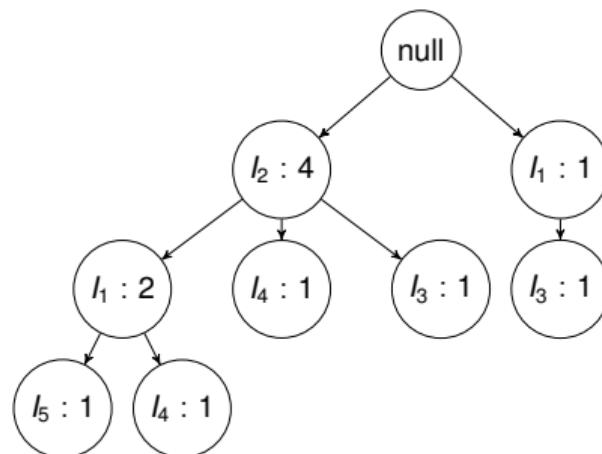


ASSOCIATION RULE MINING USING FP GROWTH

For transaction $T_4 : \{l_2, l_1, l_4\}$

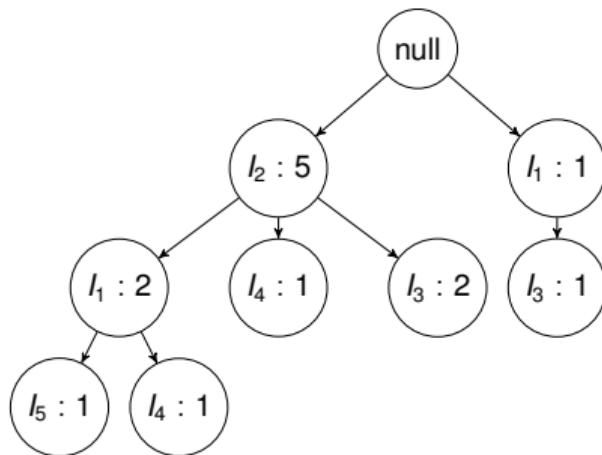


For transaction $T_5 : \{l_1, l_3\}$

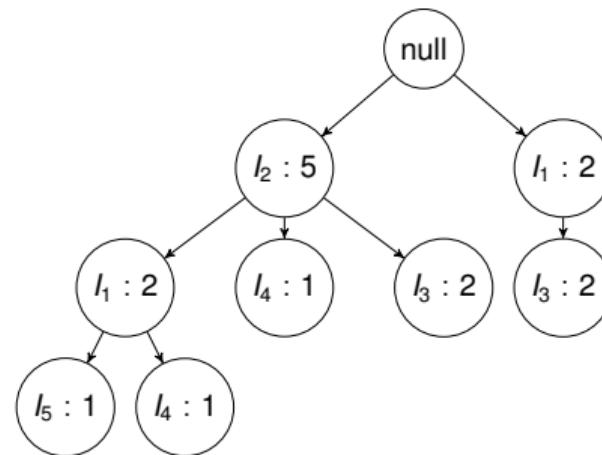


ASSOCIATION RULE MINING USING FP GROWTH

For transaction $T_6 : \{l_2, l_3\}$

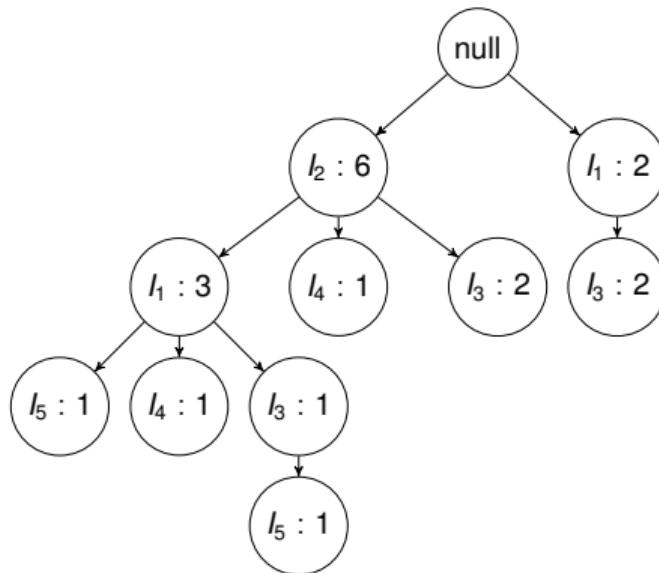


For transaction $T_7 : \{l_1, l_3\}$

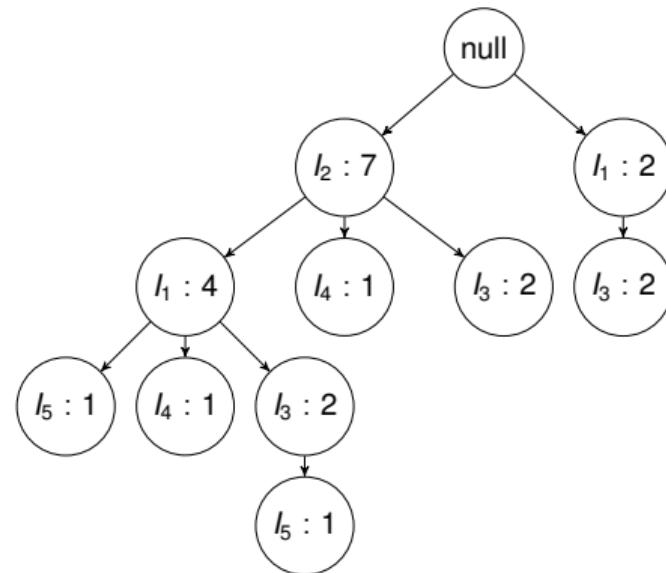


ASSOCIATION RULE MINING USING FP GROWTH

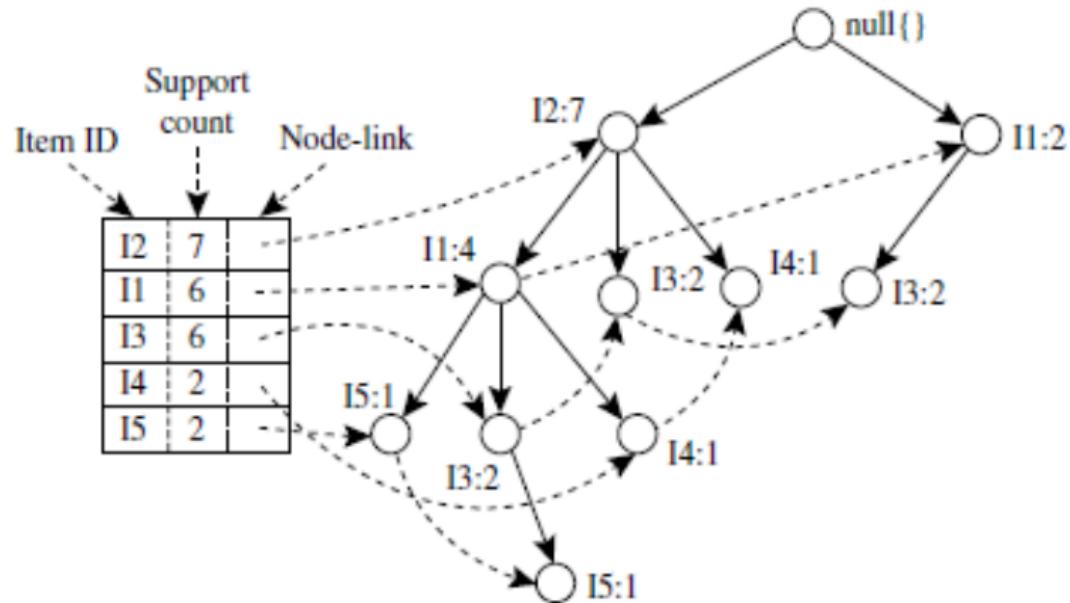
For transaction $T_8 : \{I_2, I_1, I_3, I_5\}$



For transaction $T_9 : \{I_2, I_1, I_3\}$



ASSOCIATION RULE MINING USING FP GROWTH

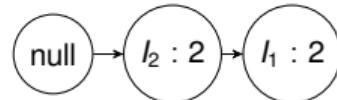


ASSOCIATION RULE MINING USING FP GROWTH

- Step 3: Construct conditional pattern base for each frequent 1-item-set. Start with last item in L i.e. I_5 .

For transaction I_5

- Paths involving $I_5 \rightarrow \{I_2, I_1, I_5\} : 1; \{I_2, I_1, I_3, I_5\} : 1$
Prefix paths involving $I_5 \rightarrow \{I_2, I_1\} : 1; \{I_2, I_1, I_3\} : 1$
- Construct I_5 conditional FP tree.
 I_3 is ignored as its support count is $1 < \text{min_support}$.

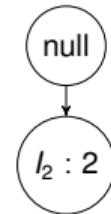


- Generate combinations of frequent pattern for I_5 .
 $FP(I_5) \rightarrow \{I_2, I_5\} : 2; \{I_1, I_5\} : 2; \{I_2, I_1, I_5\} : 2$

ASSOCIATION RULE MINING USING FP GROWTH

For transaction I_4

- ① Paths involving $I_4 \rightarrow \{I_2, I_1, I_4\} : 1; \{I_2, I_4\} : 1$
Prefix paths involving $I_4 \rightarrow \{I_2, I_1\} : 1; \{I_2\} : 1$
- ② Construct I_4 conditional FP tree.
 I_1 is ignored as its support count is 1, which is less than min_support.

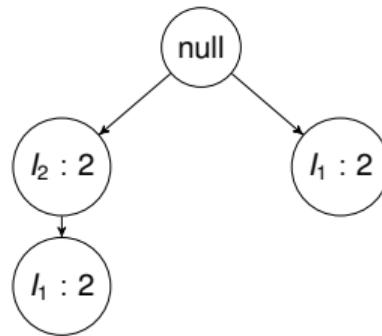


- ③ Generate combinations of frequent pattern for I_4 .
 $FP(I_4) \rightarrow \{I_2, I_4\} : 2$

ASSOCIATION RULE MINING USING FP GROWTH

For transaction I_3

- ① Paths involving $I_3 \rightarrow \{I_2, I_1, I_3\} : 2; \{I_2, I_3\} : 2; \{I_1, I_3\} : 2$
Prefix paths involving $I_3 \rightarrow \{I_2, I_1\} : 1; \{I_2\} : 2; \{I_1\} : 2$
- ② Construct I_3 conditional FP tree.

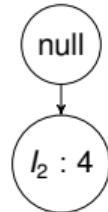


- ③ Generate combinations of frequent pattern for I_3 .
 $FP(I_3) \rightarrow \{I_2, I_3\} : 4; \{I_1, I_3\} : 4; \{I_2, I_1, I_3\} : 2$

ASSOCIATION RULE MINING USING FP GROWTH

For transaction I_1

- ① Paths involving $I_1 \rightarrow \{I_2, I_1\} : 4$; $\{I_1\} : 2$
Prefix paths involving $I_1 \rightarrow \{I_2\} : 4$
- ② Construct I_1 conditional FP tree.



- ③ Generate combinations of frequent pattern for I_1 .
 $FP(I_1) \rightarrow \{I_2, I_1\} : 4$

ASSOCIATION RULE MINING USING FP GROWTH

- Step 4: Mining FP-tree by creating conditional sub pattern bases.

Item	Conditional Pattern Base	Frequent pattern generated
I_5	$\{I_2, I_1\} : 1;$ $\{I_2, I_1, I_3\} : 1$	$\{I_2, I_5\} : 2; \{I_1, I_5\} : 2;$ $\{I_2, I_1, I_5\} : 2$
I_4	$\{I_2, I_1\} : 1; \{I_2\} : 1$	$\{I_2, I_4\} : 2$
I_3	$\{I_2, I_1\} : 2;$ $\{I_2\} : 2; \{I_1\} : 2$	$\{I_2, I_3\} : 4; \{I_1, I_3\} : 4;$ $\{I_2, I_1, I_3\} : 2$
I_1	$\{I_2\} : 4$	$\{I_2, I_1\} : 4$

ASSOCIATION RULE MINING USING FP GROWTH

- Step 5: Generate association rules from frequent item-sets.

	Frequent pattern	Rules generated	Confidence	Rule acceptance
FP1	$\{I_2, I_5\}$	$I_5 \rightarrow I_2$	$2/2 = 1 > 0.7$	Accepted
FP2	$\{I_1, I_5\}$	$I_5 \rightarrow I_1$	$2/2 = 1 > 0.7$	Accepted
FP3	$\{I_2, I_1, I_5\}$	$I_5 \rightarrow I_1 I_2$ $I_1 I_5 \rightarrow I_2$ $I_2 I_5 \rightarrow I_1$	$2/2 = 1 > 0.7$ $2/2 = 1 > 0.7$ $2/2 = 1 > 0.7$	Accepted Accepted Accepted
FP4	$\{I_2, I_4\}$	$I_4 \rightarrow I_2$	$2/2 = 1 > 0.7$	Accepted

ASSOCIATION RULE MINING USING FP GROWTH

FP1	$\{I_2, I_5\}$	$I_2 \rightarrow I_5$	$2/7 = 0.28 < 0.7$	rejected
FP2	$\{I_1, I_5\}$	$I_1 \rightarrow I_5$	$2/6 = 0.33 \nmid 0.7$	rejected
FP3	$\{I_2, I_1, I_5\}$	$I_2 \rightarrow I_1 I_5$	$2/7 = 0.28 < 0.7$	rejected
		$I_1 \rightarrow I_2 I_5$	$2/6 = 0.33 \nmid 0.7$	rejected
		$I_2 I_1 \rightarrow I_5$	$2/4 = 0.5 < 0.7$	rejected
FP4	$\{I_2, I_4\}$	$I_2 \rightarrow I_4$	$2/7 = 0.28 < 0.7$	rejected
FP5	$\{I_2, I_3\}$	$I_2 \rightarrow I_3$	$4/7 = 0.6 < 0.7$	rejected
		$I_3 \rightarrow I_2$	$4/6 = 0.67 < 0.7$	rejected
FP6	$\{I_1, I_3\}$	$I_1 \rightarrow I_3$	$4/6 = 0.67 < 0.7$	rejected

ASSOCIATION RULE MINING

		$I_3 \rightarrow I_1$	$4/6 = 0.67 < 0.7$	rejected
FP7	$\{I_2, I_1, I_3\}$	$I_2 \rightarrow I_1 I_3$	$2/7 = 0.28 < 0.7$	rejected
		$I_1 \rightarrow I_2 I_3$	$2/6 = 0.33 < 0.7$	rejected
		$I_3 \rightarrow I_2 I_1$	$2/6 = 0.33 < 0.7$	rejected
		$I_1 I_3 \rightarrow I_2$	$2/4 = 0.5 < 0.7$	rejected
		$I_2 I_3 \rightarrow I_1$	$2/4 = 0.5 < 0.7$	rejected
		$I_2 I_1 \rightarrow I_3$	$2/4 = 0.5 < 0.7$	rejected
FP8	$\{I_2, I_1\}$	$I_2 \rightarrow I_1$	$4/7 = 0.6 < 0.7$	rejected
		$I_1 \rightarrow I_2$	$4/6 = 0.67 < 0.7$	rejected

APRIORI VS FP GROWTH

Apriori	FP Growth
Pattern Generation	
Generates pattern by pairing the items.	Generates pattern by constructing a FP tree.
Candidate Generation	
Uses candidate generation	No candidate generation
Process	
Slower	Faster
Exponential increase in process run-time as number of item-sets increases.	Linear increase in process run-time as number of item-sets increases.
Memory Usage	
Saved in memory	Compact version is saved.

TABLE OF CONTENTS

- 1 ASSOCIATION ANALYSIS
- 2 APRIORI ALGORITHM
- 3 FP GROWTH ALGORITHM
- 4 PATTERN EVALUATION METHODS

PATTERN EVALUATION METHODS

- Support
- Confidence
- Correlation Rules - Lift

LIFT

- The occurrence of item-set A is independent of the occurrence of item-set B if $P(A \cup B) = P(A)P(B)$; otherwise, item-sets A and B are dependent and correlated as events.

$$\text{lift}(A, B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)}$$

- If $\text{lift}(A, B) < 1$, then the occurrence of A is negatively correlated with the occurrence of B .
- If $\text{lift}(A, B) > 1$, then A and B are positively correlated, meaning that the occurrence of A implies the occurrence of B .
- If $\text{lift}(A, B) = 1$, then A and B are independent and there is no correlation between them.

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 8 : CLUSTERING

IDS Course Team

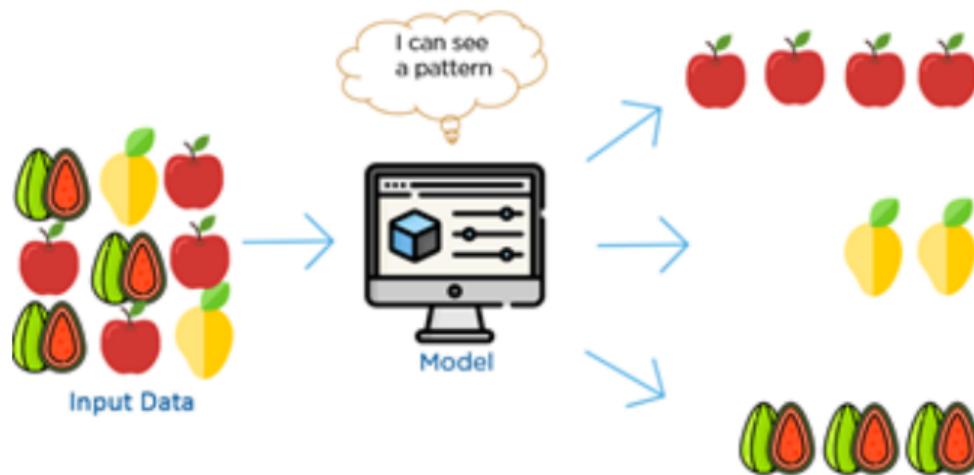
BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

- 1 CLUSTERING ANALYSIS - CONCEPTS
- 2 PARTITIONING METHODS
- 3 K-MEANS ALGORITHM
- 4 HIERARCHICAL CLUSTERING
- 5 DENSITY BASED CLUSTERING
- 6 EVALUATION OF CLUSTERING ALGORITHMS

CLUSTERING

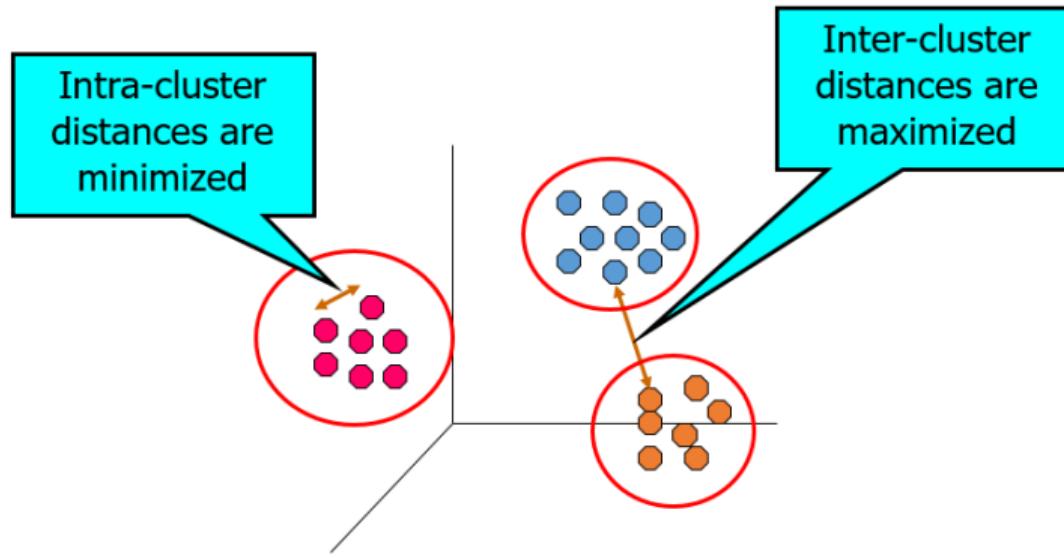


- Clustering is useful in that it can lead to the discovery of previously unknown groups within the data.

Image credit: Hunter Heidenreich

CLUSTERING

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.

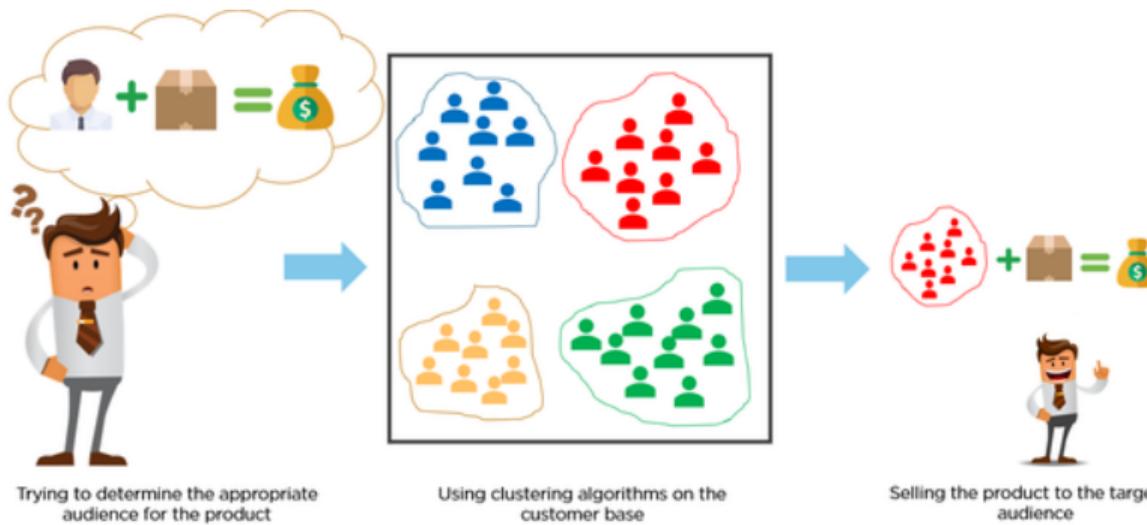


CLUSTERING

- Clustering uses **unsupervised learning** because the class label information is not present.
- Clustering is a form of **learning by observation**.

APPLICATIONS OF CLUSTERING

Market Segmentation



APPLICATIONS OF CLUSTERING

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs. This facilitates the development of business strategies for enhanced customer relationship management.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- Earthquake studies: Observed earth quake epicenters should be clustered along continent faults.

APPLICATIONS OF CLUSTERING

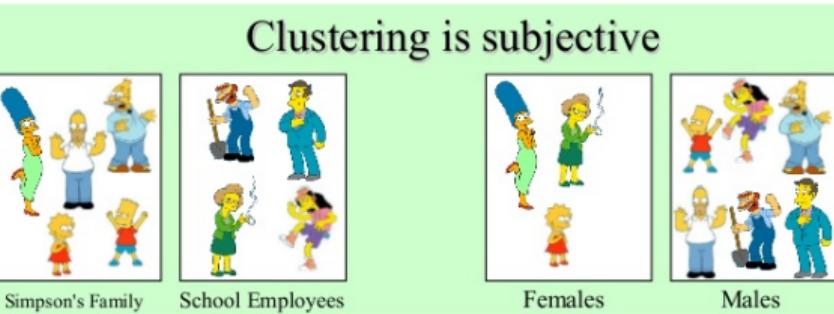
- Image Segmentation
- Web Search: Organize the search results into groups and present the results in a concise and easily accessible way.
- Identify distribution of data
- Clustering as a preprocessing step: attribute subset selection, characterization
- Outlier detection

CLUSTERING IS SUBJECTIVE

What is a natural grouping among these objects?



Clustering is subjective

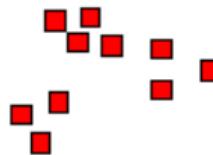


Simpson's Family	School Employees	Females	Males
			

NOTION OF A CLUSTER CAN BE AMBIGUOUS



How many clusters?



Two Clusters



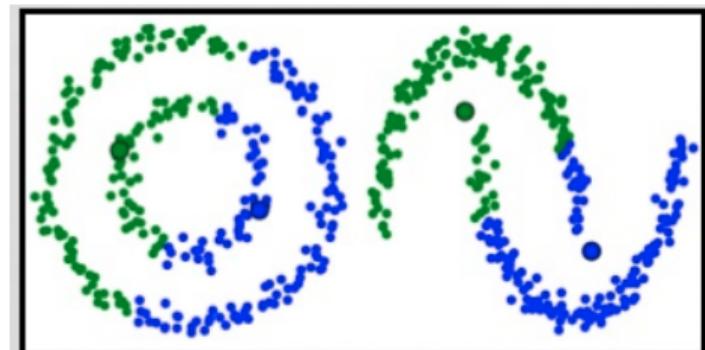
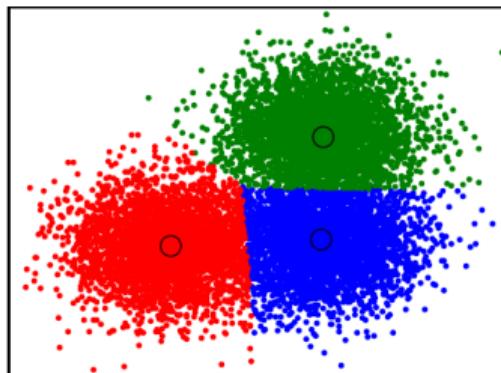
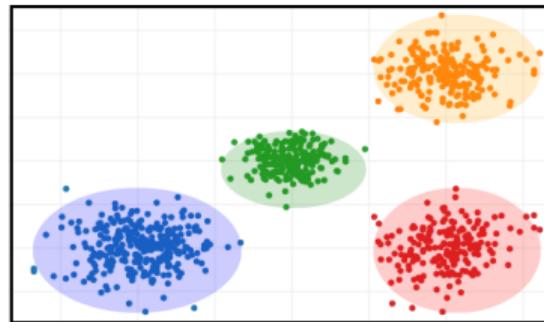
Four Clusters



Six Clusters



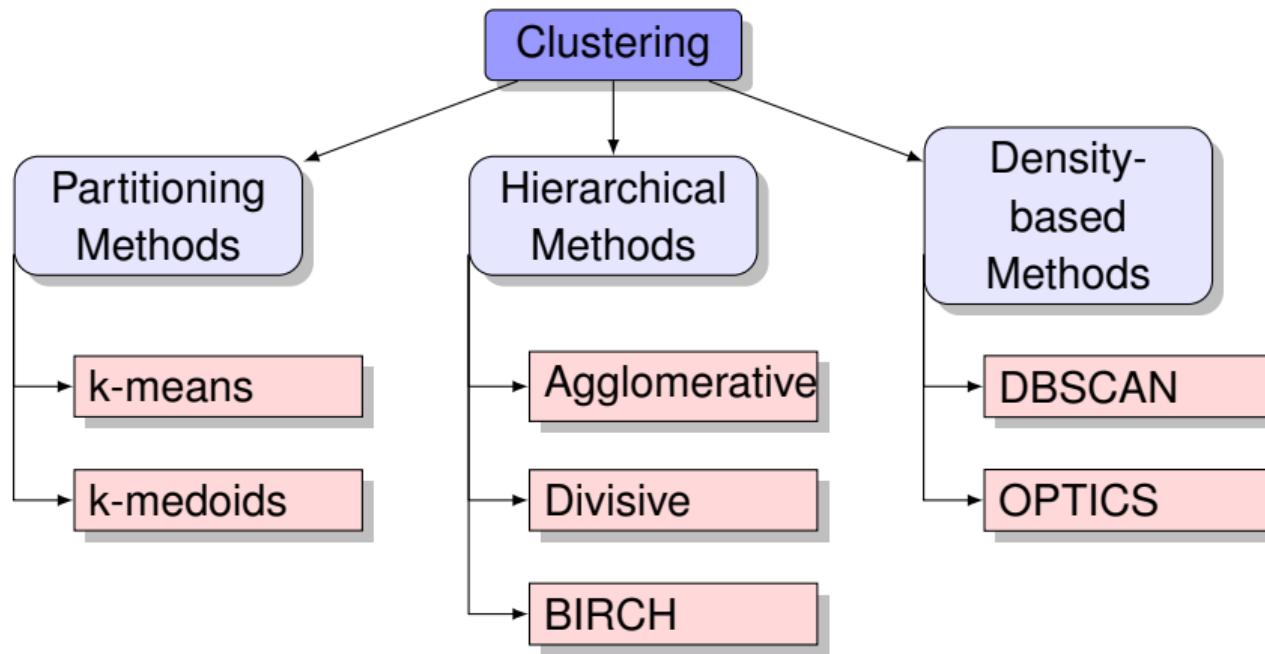
TYPES OF CLUSTERS



CLUSTERING METHODS

- Partitioning criteria - No hierarchy or hierarchical partitioning
- Separation of clusters - mutually exclusive or not
- Similarity measure - distance based, density based and contiguity based.
- Clustering space - High dimensional vs low dimensional space

CLUSTERING METHODS



COMPARISON OF CLUSTERING METHODS

Method	Characteristics
Partitioning	<ul style="list-style-type: none"> ● Find mutually exclusive clusters of spherical shape methods ● Distance-based ● May use mean or medoid to represent cluster center ● Effective for small- to medium-size data sets
Hierarchical	<ul style="list-style-type: none"> ● Clustering is a hierarchical decomposition methods ● Cannot correct erroneous merges or splits ● Other techniques like micro-clustering or object “linkages”
Density-based	<ul style="list-style-type: none"> ● Arbitrarily shaped clusters methods ● Clusters are dense regions of objects in space that are separated by low-density regions ● Cluster density: Each point must have a minimum number of points within its “neighborhood”. ● May filter out outliers

TABLE OF CONTENTS

- 1 CLUSTERING ANALYSIS - CONCEPTS
- 2 PARTITIONING METHODS
- 3 K-MEANS ALGORITHM
- 4 HIERARCHICAL CLUSTERING
- 5 DENSITY BASED CLUSTERING
- 6 EVALUATION OF CLUSTERING ALGORITHMS

PARTITIONING METHODS

- Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$.
- Each group must contain at least one object.
- Iterative Relocation Technique to improve the partitioning.
- Distance based methods.
- Clustering Algorithms – k-means, k-medoids or PAM (Partition around medoids):
- Cluster shape – Spherical shaped exclusive clusters
- Cluster size – Small to medium cluster

SIMILARITY MEASURE –NUMERICAL ATTRIBUTES

Euclidean distance $d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$

Manhattan distance $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{in} - x_{jn}|$

Minkowski distance $d(i, j) = [|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{in} - x_{jn}|^p]^{1/p}$

$$\begin{aligned}\text{Cosine Similarity } s(x, y) &= \frac{x \cdot y}{\|x\| \|y\|} \\ &= \frac{\sum_{k=1}^n x_k y_k}{\sqrt{x_1^2 + \dots + x_p^2} \sqrt{y_1^2 + \dots + y_p^2}}\end{aligned}$$

TABLE OF CONTENTS

- 1 CLUSTERING ANALYSIS - CONCEPTS
- 2 PARTITIONING METHODS
- 3 K-MEANS ALGORITHM
- 4 HIERARCHICAL CLUSTERING
- 5 DENSITY BASED CLUSTERING
- 6 EVALUATION OF CLUSTERING ALGORITHMS

K-MEANS CONCEPTS

- Centroid-Based Technique
- Centroid is the mean of the objects or points in the cluster.
- k number of clusters. The centroids are C_1, \dots, C_k and $C_i \cap C_j = \emptyset$.
- Objective function aims for high intracluster similarity and low intercluster similarity.
- The difference between an object $p \in C_i$ and cluster center c_i , is measured by $dist(p, c_i)$ the Euclidean distance between two points x and y .
- The quality of cluster C_i is measured by the within cluster variation.

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, c_i)^2$$

K-MEANS ALGORITHM

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

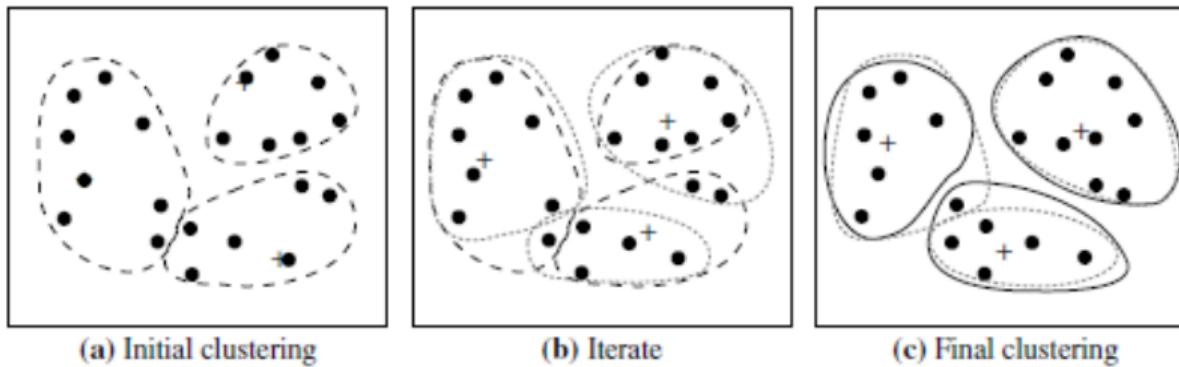
- k : the number of clusters,
- D : a data set containing n objects.

Output: A set of k clusters.

Method:

- (1) arbitrarily choose k objects from D as the initial cluster centers;
- (2) **repeat**
- (3) (re)assign each object to the cluster to which the object is the most similar,
 based on the mean value of the objects in the cluster;
- (4) update the cluster means, that is, calculate the mean value of the objects for
 each cluster;
- (5) **until** no change;

K-MEANS ALGORITHM

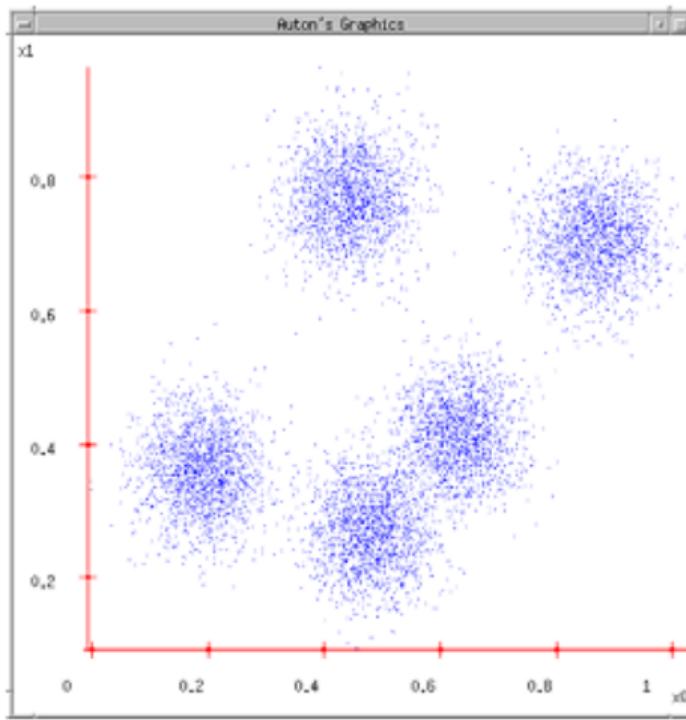


Clustering of a set of objects using the k -means method; for (b) update cluster centers and reassign objects accordingly (the mean of each cluster is marked by a +).

K-MEANS EXAMPLE (SOURCE: A. MOORE)

K-means

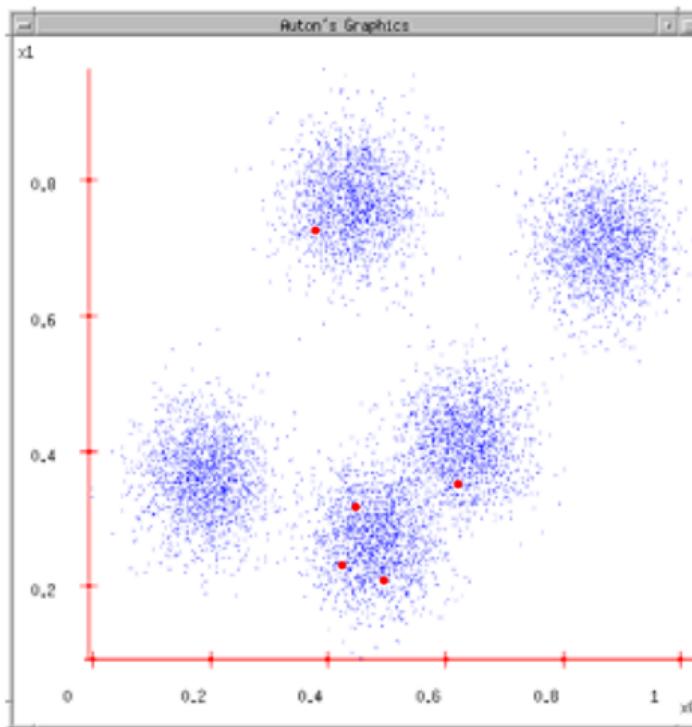
1. Ask user how many clusters they'd like.
(e.g. $k=5$)



K-MEANS EXAMPLE (SOURCE: A. MOORE)

K-means

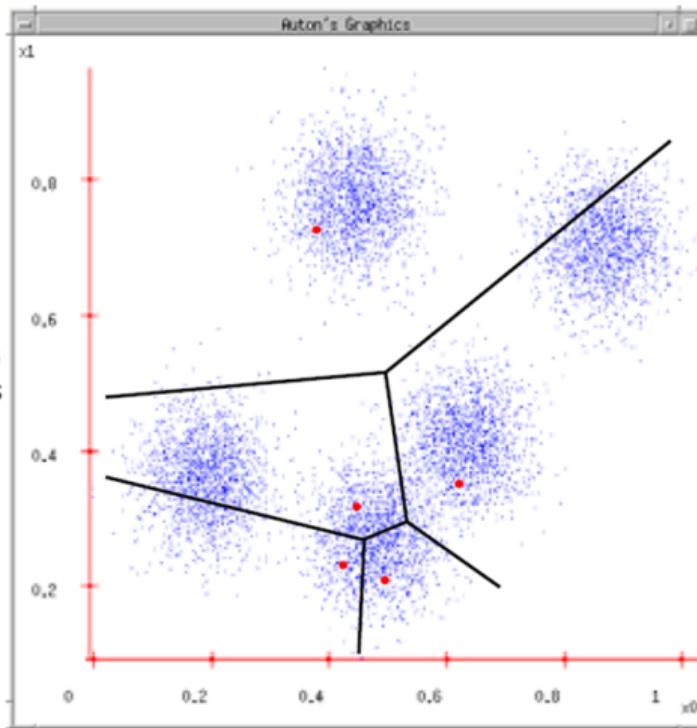
1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations



K-MEANS EXAMPLE (SOURCE: A. MOORE)

K-means

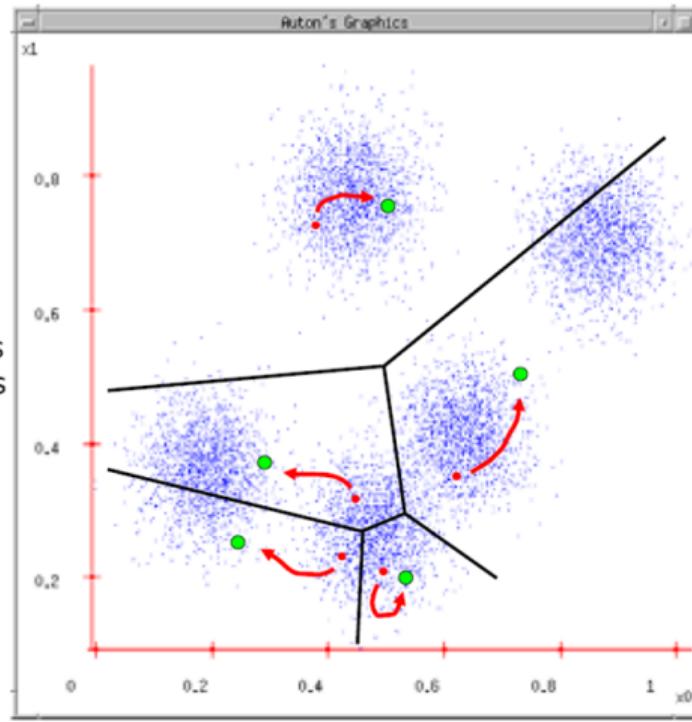
1. Ask user how many clusters they'd like.
(e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



K-MEANS EXAMPLE (SOURCE: A. MOORE)

K-means

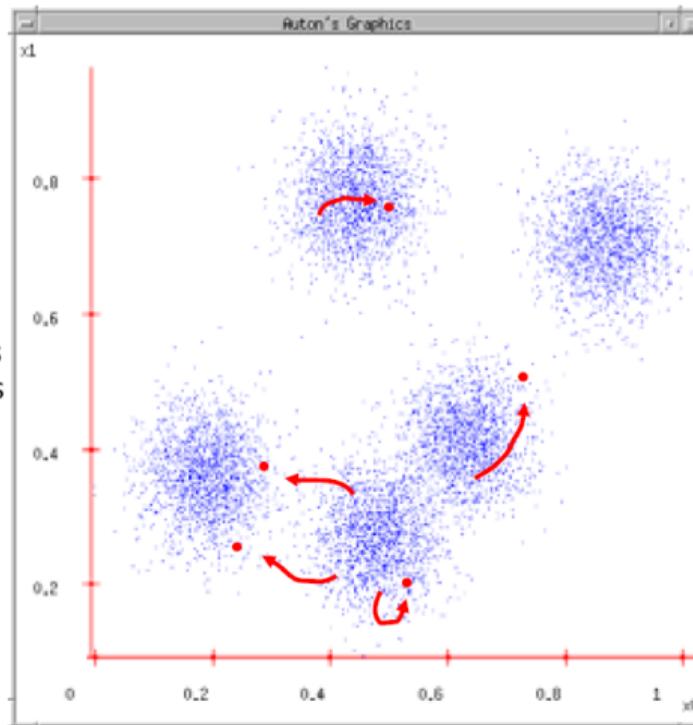
1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-MEANS EXAMPLE (SOURCE: A. MOORE)

K-means

1. Ask user how many clusters they'd like.
(e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



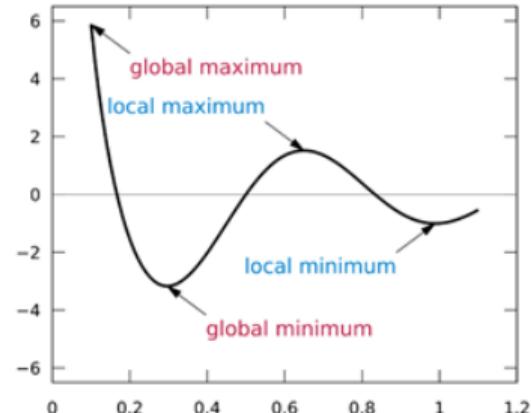
K-MEANS STRENGTHS & WEAKNESS

Strengths

- Compact and separate clusters
- Complexity is $O(nktd)$.
- Applied on Numeric data
- Relatively scalable
- Efficient in processing large datasets.

Weakness

- NP hard problem
- No guaranteed convergence to global optimum.
- Results depend on the initial random selection of cluster centers.



K-MEANS PROBLEM

Use the k-means algorithm and Euclidean distance to cluster the following 5 examples into 2 clusters: $A = (0, 1)$; $B(3, 0)$; $C(2, 4)$; $D(2, 1)$; $E(3, 5)$

STEP 1 : Compute the Euclidean distance matrix

	A	B	C	D	E
A	0	3.16	3.61	2	5
B		0	4.12	1.414	5
C			0	3	1.414
D				0	4.12
E					0

STEP 2 : Randomly choose two cluster centroids.

Let A and C be the cluster centroids.

STEP 3 : Cluster 1: { A, B, D } and Cluster 2: { C, E }

K-MEANS PROBLEM

STEP 4 : Compute the Centroids,

$$\text{Cluster1} = \text{mean of A,B,D} = \frac{0+3+2}{3}, \frac{1+0+1}{3} = (1.66, 0.66)$$

$$\text{Cluster 2} = \text{mean of C,E} = \frac{2+3}{2}, \frac{4+5}{2} = (2.5, 4.5)$$

STEP 5 : Repeat; Compute the Euclidean distance matrix to the cluster centers.

	A	B	C	D	E	C1	C2
A	0	3.16	3.61	2	5		
B		0	4.12	1.414	5		
C			0	3	1.414		
D				0	4.12		
E					0		
C1	1.69	1.49	3.35	0.48	4.54	0	
C2	4.30	4.52	0.70	3.53	0.70	3.93	0

K-MEANS PROBLEM

STEP 6 : Cluster 1: { A, B, D }

Cluster 2: { C, E }

STEP 7 : Recompute the Centroids,

$$\text{Cluster1} = \text{mean of } A, B, D = \frac{0+3+2}{3}, \frac{1+0+1}{3} = (1.66, 0.66)$$

$$\text{Cluster 2} = \text{mean of } C, E = \frac{2+3}{2}, \frac{4+5}{2} = (2.5, 4.5)$$

STEP 8 : Stop as converged or no change is seen.

K-MEANS - INITIAL CENTROIDS

- Multiple runs – Helps, but probability is not favourable.
- Sample and use hierarchical clustering to determine initial centroids.
- Select more than k initial centroids and then select among these initial centroids, the most widely separated ones.
- Post-processing

DETERMINING THE NUMBER OF CLUSTERS

- ① Set the number of clusters to about $\sqrt{\frac{n}{2}}$ for a data set of n points.
- ② Use elbow method.
 - ▶ Elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. A heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters.
- ③ Use cross validation technique.
 - ▶ Compare the overall quality measure with respect to different values of k , and find the number of clusters that best fits the data.

ELBOW METHOD - NUMERICAL PROBLEM

For the following dataset, using different values of $k = 2; 3; 4; 5; 6$ determine the best values of k using Elbow method.

$$D = \{2; 10; 12; 4; 25; 3; 30; 20; 11\}$$

Assumptions: Manhattan distance measure.

k=2: Assume initial centroids as 3 and 10.

$$C_1 = \{3; 2; 4\}$$

$$C_2 = \{10; 12; 25; 30; 20; 11\}$$

$$\begin{aligned} \text{Variance} &= (3 - 3)^2 + (3 - 2)^2 + (3 - 4)^2 + (10 - 10)^2 + (12 - 10)^2 \\ &\quad + (11 - 10)^2 + (20 - 10)^2 + (25 - 10)^2 + (30 - 10)^2 \\ &= 732 \end{aligned}$$

ELBOW METHOD - NUMERICAL PROBLEM

k=3: Assume initial centroids as 3 ; 10 and 25.

$$C_1 = \{3; 2; 4\}$$

$$C_2 = \{10; 12; 11\}$$

$$C_3 = \{25; 30; 20\}$$

$$\begin{aligned}Var &= (3 - 3)^2 + (3 - 2)^2 + (3 - 4)^2 + (10 - 10)^2 + (12 - 10)^2 \\&\quad + (11 - 10)^2 + (20 - 25)^2 + (25 - 25)^2 + (30 - 25)^2 \\&= 57\end{aligned}$$

ELBOW METHOD - NUMERICAL PROBLEM

k=4: Assume initial centroids as 3 ; 10; 20 and 30.

$$C_1 = \{3; 2; 4\}$$

$$C_2 = \{10; 12; 11\}$$

$$C_3 = \{25; 20\}$$

$$C_4 = \{30\}$$

$$\begin{aligned}Var &= (3 - 3)^2 + (3 - 2)^2 + (3 - 4)^2 + (10 - 10)^2 + (12 - 10)^2 \\&\quad + (11 - 10)^2 + (20 - 25)^2 + (25 - 25)^2 + (30 - 30)^2 \\&= 32\end{aligned}$$

ELBOW METHOD - NUMERICAL PROBLEM

k=5: Assume initial centroids as 3 ; 10; 12; 20 and 25.

$$C_1 = \{3; 2; 4\}$$

$$C_2 = \{10\}$$

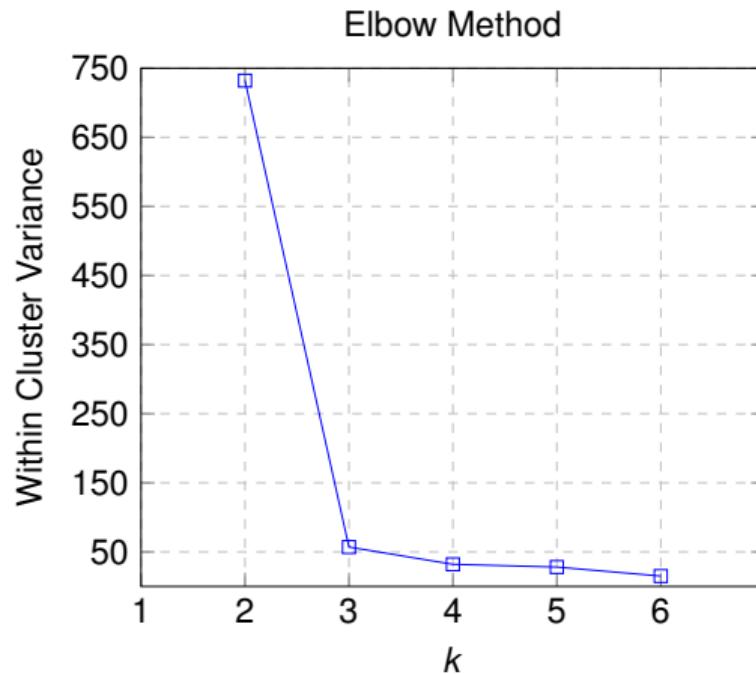
$$C_3 = \{12; 11\}$$

$$C_4 = \{20\}$$

$$C_5 = \{25; 30\}$$

$$\begin{aligned}Var &= (3 - 3)^2 + (3 - 2)^2 + (3 - 4)^2 + (10 - 10)^2 + (12 - 12)^2 \\&\quad + (11 - 12)^2 + (20 - 20)^2 + (25 - 25)^2 + (30 - 25)^2 \\&= 28\end{aligned}$$

ELBOW METHOD - NUMERICAL PROBLEM



Using the elbow method, $k = 3$ will yield better clustering results.

K-MEANS - PRE-PROCESSING AND POST-PROCESSING

Pre-processing

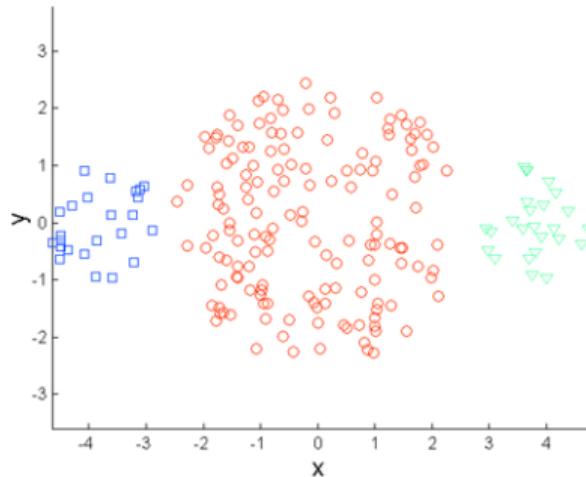
- Normalize the data
- Eliminate outliers

Post-processing

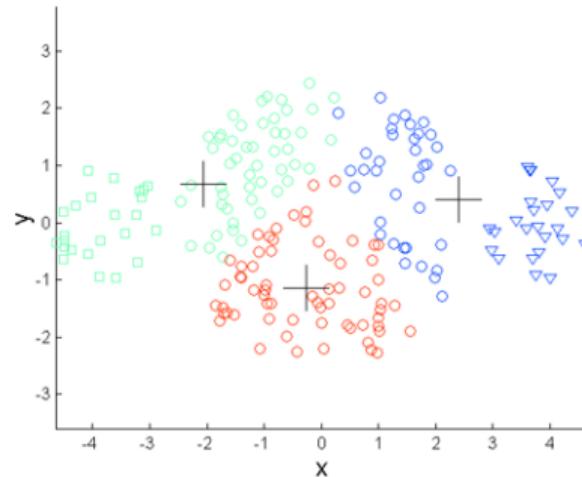
- Eliminate small clusters that may represent outliers.
- Split 'loose' clusters, i.e., clusters with relatively high SSE.
- Merge clusters that are 'close' and that have relatively low SSE.

K-MEANS LIMITATIONS: DIFFERING SIZES

Issue:



Original Points

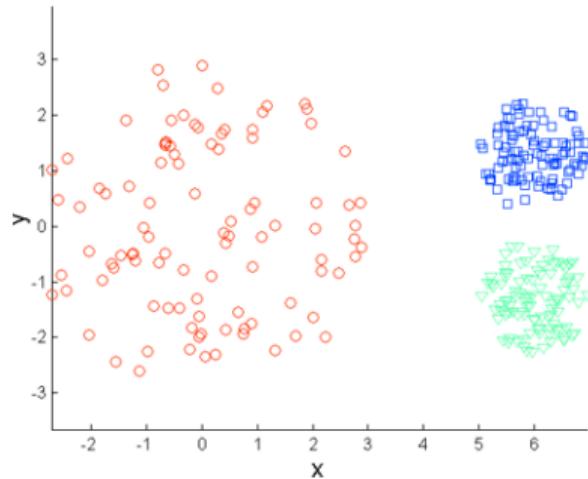


K-means (3 Clusters)

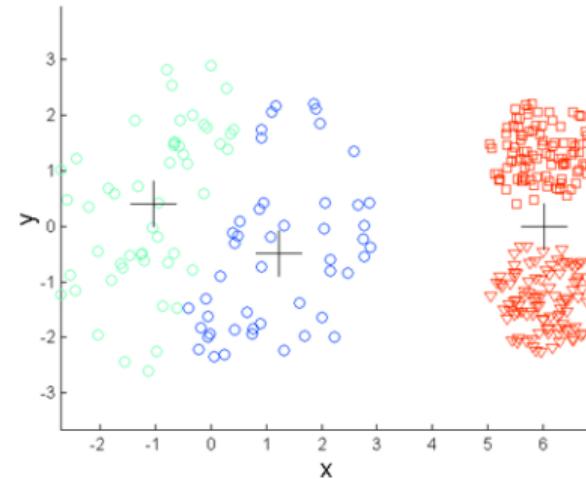
Solution: Use many clusters. Find parts of clusters. Put them together.

K-MEANS LIMITATIONS: DIFFERING DENSITY

Issue:



Original Points

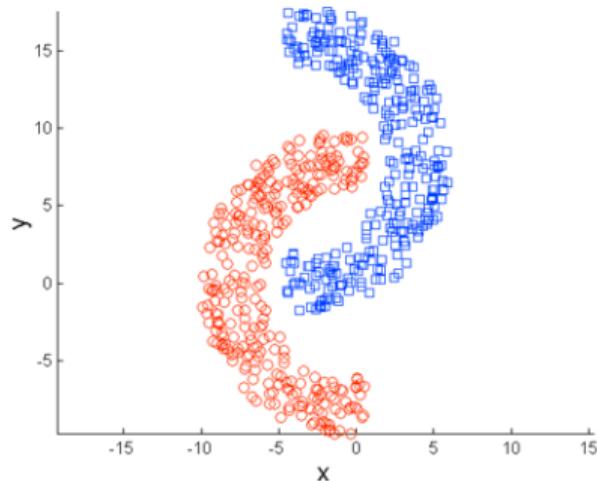


K-means (3 Clusters)

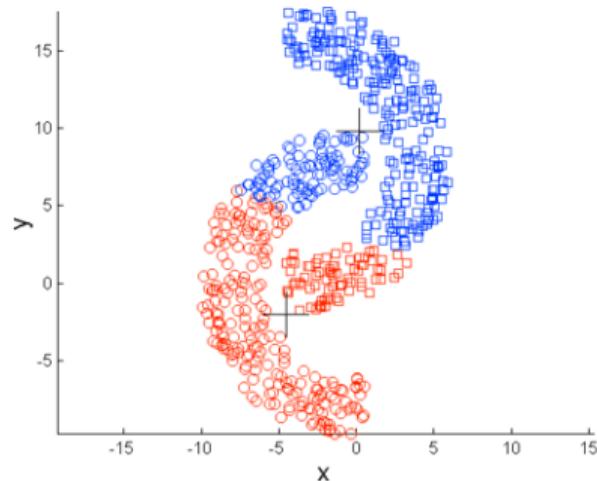
Solution: Use many clusters. Find parts of clusters. Put them together.

K-MEANS LIMITATIONS: NON-GLOBULAR SHAPES

Issue:



Original Points



K-means (2 Clusters)

Solution: Use many clusters. Find parts of clusters. Put them together.

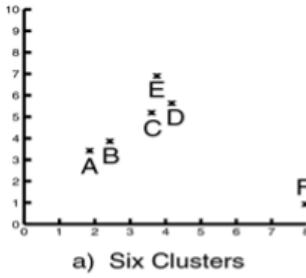
TABLE OF CONTENTS

- 1 CLUSTERING ANALYSIS - CONCEPTS
- 2 PARTITIONING METHODS
- 3 K-MEANS ALGORITHM
- 4 HIERARCHICAL CLUSTERING
- 5 DENSITY BASED CLUSTERING
- 6 EVALUATION OF CLUSTERING ALGORITHMS

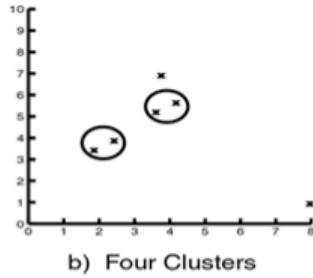
HIERARCHICAL CLUSTERING

- Creates a hierarchical decomposition of the given set of data objects.
- A set of nested clusters organized as a hierarchical tree.
- Agglomerative approach, also called the bottom-up approach.
- Divisive approach, also called the top-down approach.
- Distance-based or density-based methods.
- Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone.
- Hierarchy of clusters
- Clustering Algorithms – BIRCH, AGNES, DIANA
- Cluster shape – Spherical shaped clusters
- Cluster size – Small to medium cluster, can be scaled to large data-set.

HIERARCHICAL CLUSTERING



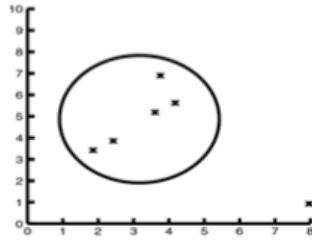
a) Six Clusters



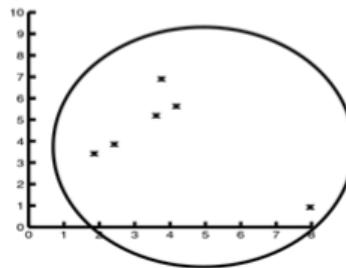
b) Four Clusters



c) Three Clusters



d) Two Clusters

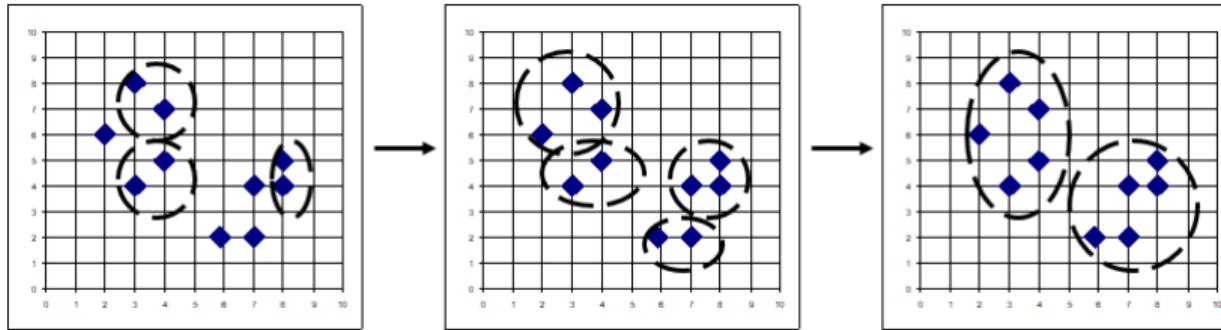


e) One Cluster

⁰Source: Yücel Saygin

AGGLOMERATIVE CONCEPTS

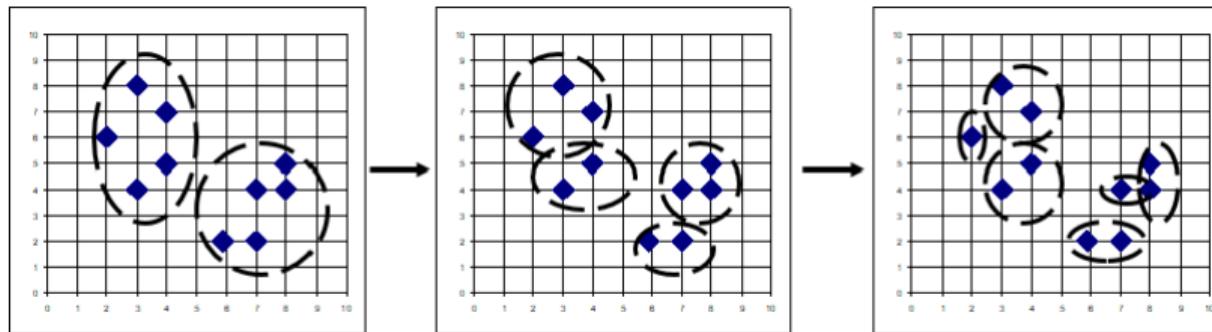
- Bottom-up strategy.



⁰Source: Yücel Saygin

DIVISIVE CONCEPTS

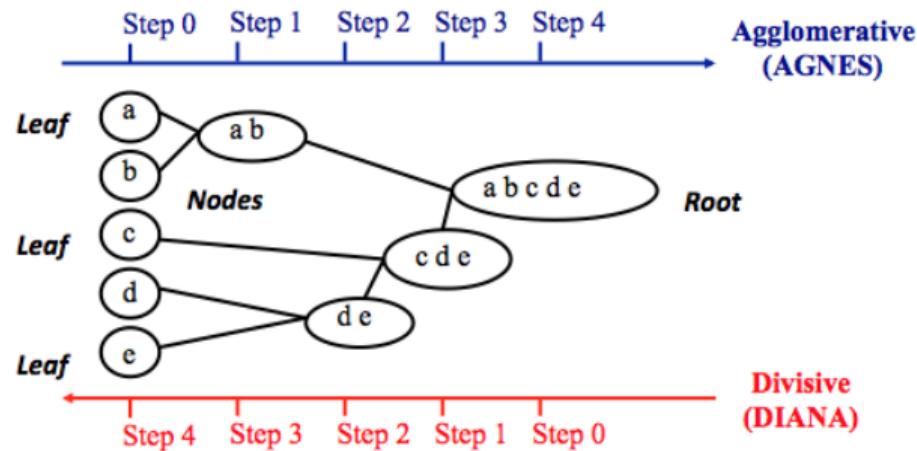
- Top-down strategy.



⁰Source: Yücel Saygin

AGGLOMERATIVE VERSUS DIVISIVE CLUSTERING

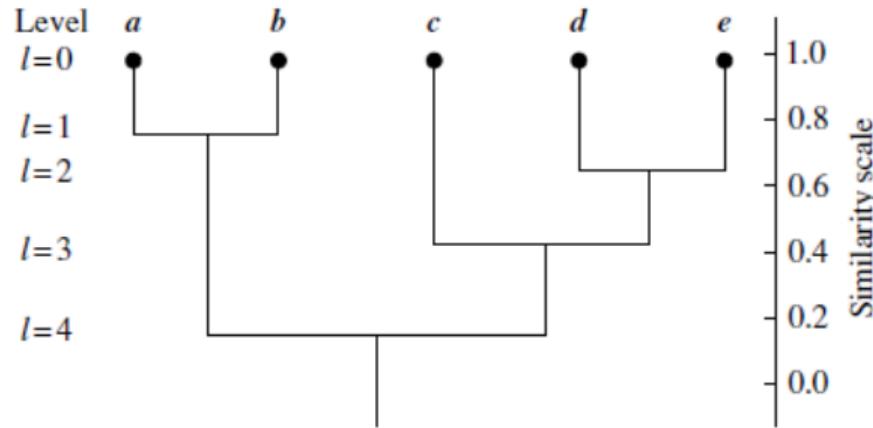
Agglomerative clustering – AGNES – AGglomerative NESting



Divisive clustering – DIANA – Dlvisive ANAlysis

DENDOGRAM

- A dendrogram is a tree structure used to represent the process of hierarchical clustering.
- It shows step-by-step grouping of objects an agglomerative or a divisive method.



AGGLOMERATIVE ALGORITHM

Input:

$D = \{t_1, t_2, \dots, t_n\}$ // Set of elements
 A // Adjacency matrix showing distance between elements.

Output:

DE // Dendrogram represented as a set of ordered triples.

Agglomerative Algorithm:

$d = 0;$

$k = n;$

$K = \{\{t_1\}, \dots, \{t_n\}\};$

$DE = \{< d, k, K >\};$ // Initially dendrogram contains each element in its own cluster.

repeat

$oldk = k;$

$d = d + 1;$

$A_d =$ Vertex adjacency matrix for graph with threshold distance of $d;$

$< k, K > = NewClusters(A_d, D);$

if $oldk \neq k$ **then**

$DE = DE \cup < d, k, K >;$ // New set of clusters added to dendrogram.

until $k = 1$

NEAREST NEIGHBOR CLUSTERING ALGORITHM

- Uses minimum distance measure.

$$dist_{min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

- If the clustering process is terminated when the distance between nearest clusters exceeds a user-defined threshold, it is called a single-linkage algorithm.

FARTHEST NEIGHBOR CLUSTERING ALGORITHM

- Uses maximum distance measure.

$$dist_{max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

- If the clustering process is terminated when the maximum distance between nearest clusters exceeds a user-defined threshold, it is called a complete-linkage algorithm.

AVERAGE LINKAGE CLUSTERING ALGORITHM

- Uses average distance measure.

$$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

LINKAGE MEASURES

- Single linkage

$$dist_{min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

- Complete linkage

$$dist_{max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

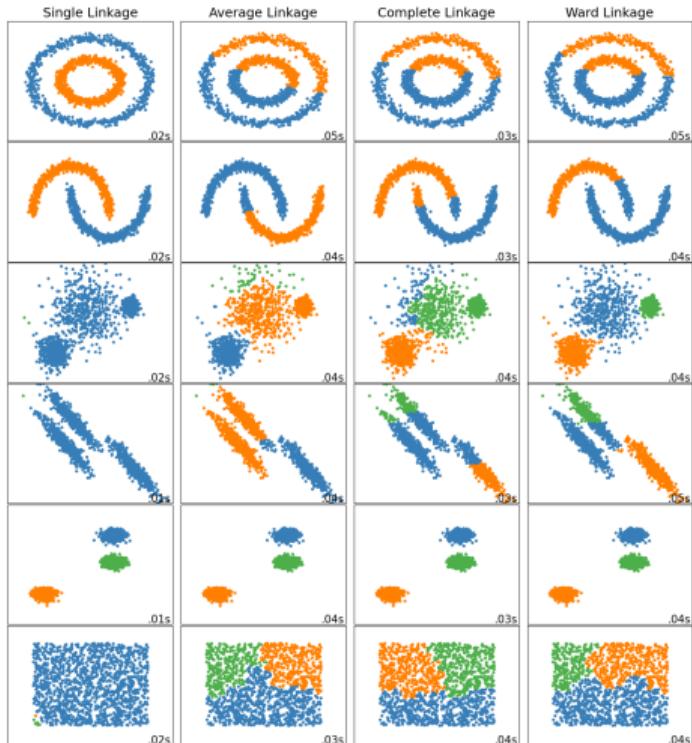
- Average linkage

$$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p_i \in C_i, p_j \in C_j} |p_i - p_j|$$

LINKAGE MEASURES

- Single linkage
 - ▶ fast
 - ▶ perform well on non-globular data
 - ▶ performs poorly in the presence of noise
- Complete linkage
 - ▶ perform well on cleanly separated globular clusters
 - ▶ overly sensitive to outliers or noisy data
- Average linkage
 - ▶ perform well on cleanly separated globular clusters
 - ▶ handle categorical as well as numeric data.

LINKAGE MEASURES



SINGLE LINKAGE PROBLEM

Cluster the following points $A = (2, 2)$; $B(5, 8)$; $C(2, 4)$; $D(4, 3)$; $E(3, 5)$ using single linkage hierarchical clustering algorithm.

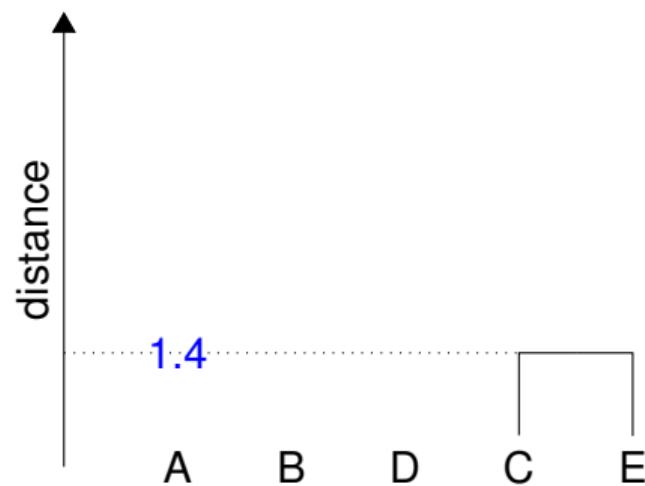
STEP 1 : Compute the Euclidean distance matrix

	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

STEP 2 : Find the minimum distance for within cluster similarity. $d_{min} = 1.4$ for the points C and E. So group them together.

SINGLE LINKAGE PROBLEM

STEP 3 : Draw the dendrogram.



SINGLE LINKAGE PROBLEM

STEP 4 : Recompute the distance matrix using single linkage or minimum distance.

$$\text{Distance}(A, CE) = \min(AC, AE) = \min(2.0, 3.2) = 2.0$$

$$\text{Distance}(B, CE) = \min(BC, BE) = \min(5.0, 3.6) = 3.6$$

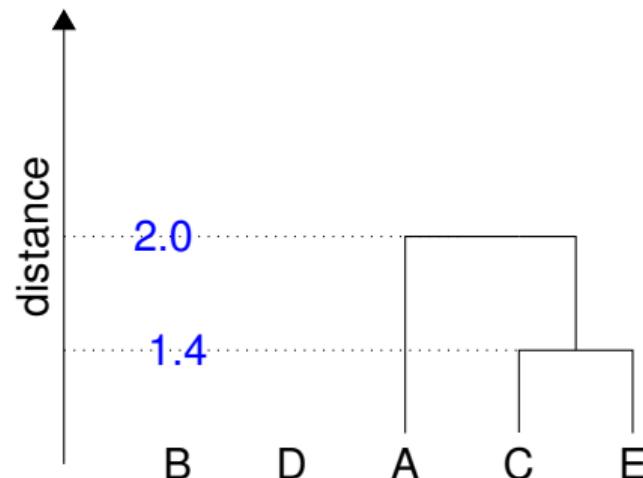
$$\begin{aligned}\text{Distance}(D, CE) &= \min(CD, DE) \\ &= \min(2.2, 2.2) = 2.2\end{aligned}$$

	A	B	D	CE
A	0	6.7	2.2	2.0
B		0	5.1	3.6
D			0	2.2
CE				0

SINGLE LINKAGE PROBLEM

STEP 5 : Find the minimum distance for within cluster similarity. $d_{min} = 2.0$ for the points A and CE. So group them together.

STEP 6 : Draw the dendrogram.



SINGLE LINKAGE PROBLEM

STEP 7 : Recompute the distance matrix using single linkage or minimum distance.

$$\begin{aligned} \text{Distance}(B, ACE) &= \min(AB, BC, BE) \\ &= \min(6.7, 5.0, 3.6) = 3.6 \end{aligned}$$

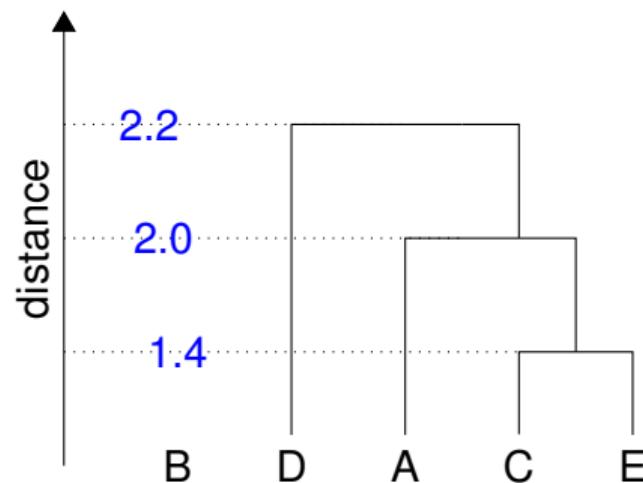
$$\begin{aligned} \text{Distance}(D, ACE) &= \min(AD, CD, DE) \\ &= \min(2.2, 2.2, 2.2) = 2.2 \end{aligned}$$

	B	D	ACE
B	0	5.1	3.6
D		0	2.2
ACE			0

STEP 8 : Find the minimum distance for within cluster similarity. $d_{min} = 2.2$ for the points D and ACE. So group them together.

SINGLE LINKAGE PROBLEM

STEP 9 : Draw the dendrogram.



SINGLE LINKAGE PROBLEM

STEP 10 : Recompute the distance matrix using single linkage or minimum distance.

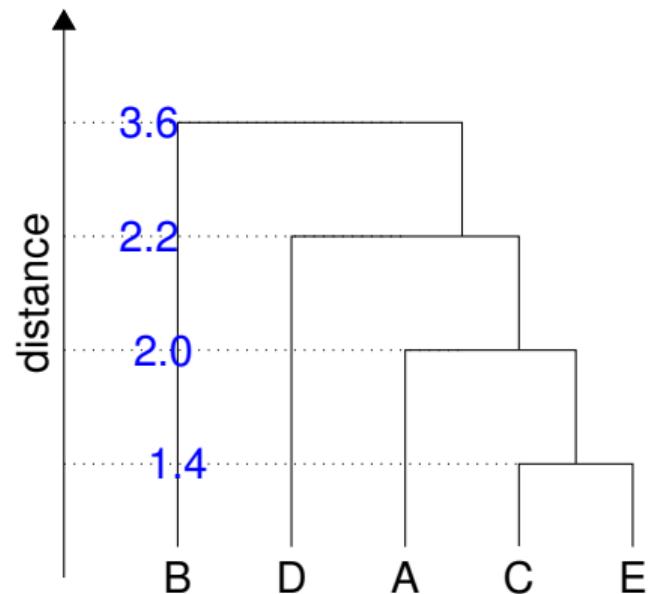
$$\begin{aligned} \text{Distance}(B, DACE) &= \min(AB, BC, BD, BE) \\ &= \min(6.7, 5.0, 5.1, 3.6) = 3.6 \end{aligned}$$

	B	DACE
B	0	3.6
DACE		0

STEP 11 : Merge all clusters at $d = 3.6$.

SINGLE LINKAGE PROBLEM

STEP 12 : Draw the dendrogram.



ADDITIONAL PROBLEMS

- Cluster the following points $A = (2, 2)$; $B(5, 8)$; $C(2, 4)$; $D(4, 3)$; $E(3, 5)$ using average linkage hierarchical clustering algorithm.

COMPLETE LINKAGE PROBLEM

Cluster the following points $A = (2, 2); B(5, 8); C(2, 4); D(4, 3); E(3, 5)$ using complete linkage hierarchical clustering algorithm.

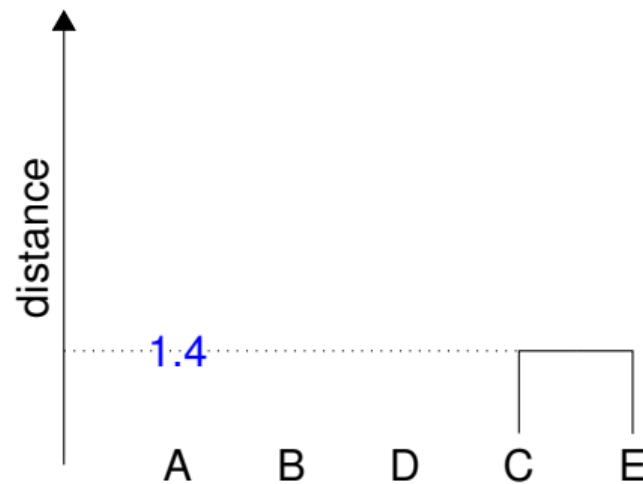
STEP 1 : Compute the Euclidean distance matrix

	A	B	C	D	E
A	0	6.7	2.0	2.2	3.2
B		0	5.0	5.1	3.6
C			0	2.2	1.4
D				0	2.2
E					0

STEP 2 : Find the minimum distance for within cluster similarity. $d_{min} = 1.4$ for the points C and E. So group them together.

COMPLETE LINKAGE PROBLEM

STEP 3 : Draw the dendrogram.



COMPLETE LINKAGE PROBLEM

STEP 4 : Recompute the distance matrix using complete linkage or maximum distance.

$$\text{Distance}(A, CE) = \max(AC, AE) = \max(2.0, 3.2) = 3.2$$

$$\text{Distance}(B, CE) = \max(BC, BE) = \max(5.0, 3.6) = 5.0$$

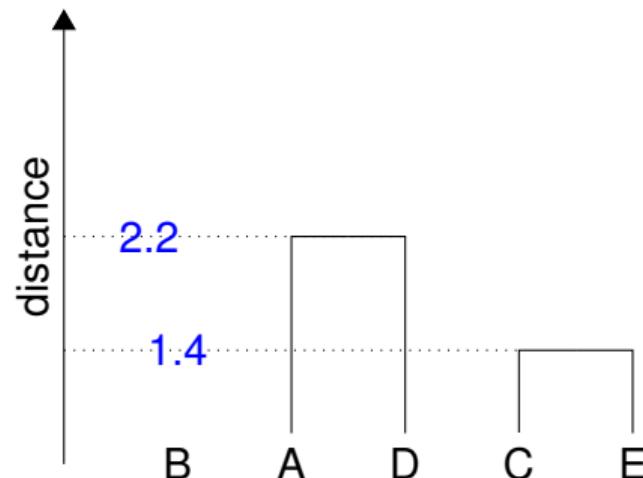
$$\text{Distance}(D, CE) = \max(CD, DE) = \max(2.2, 2.2) = 2.2$$

	A	B	D	CE
A	0	6.7	2.2	3.2
B		0	5.1	5.0
D			0	2.2
CE				0

COMPLETE LINKAGE PROBLEM

STEP 5 : Find the minimum distance for within cluster similarity. $d_{min} = 2.2$ for the points A and D. So group them together.

STEP 6 : Draw the dendrogram.



COMPLETE LINKAGE PROBLEM

STEP 7 : Recompute the distance matrix using complete linkage or maximum distance.

$$\text{Distance}(B, AD) = \max(AB, BD) = \max(6.7, 2.2) = 6.7$$

$$\text{Distance}(B, CE) = \max(BC, BE) = \max(5.0, 3.6) = 5.0$$

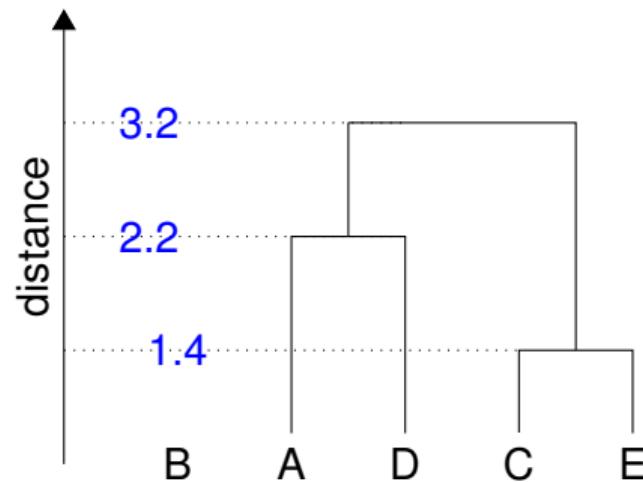
$$\begin{aligned}\text{Distance}(AD, CE) &= \max(AC, AE, DC, DE) \\ &= \max(2.0, 3.2, 2.2, 2.2) = 3.2\end{aligned}$$

	B	AD	CE
B	0	6.7	5.0
AD		0	3.2
CE			0

COMPLETE LINKAGE PROBLEM

STEP 8 : Find the minimum distance for within cluster similarity. $d_{min} = 3.2$ for the points AD and CE. So group them together.

STEP 9 : Draw the dendrogram.



COMPLETE LINKAGE PROBLEM

STEP 10 : Recompute the distance matrix using complete linkage or maximum distance.

$$\begin{aligned} \text{Distance}(B, ADCE) &= \max(AB, BC, BD, BE) \\ &= \max(6.7, 5.0, 5.1, 3.6) = 6.7 \end{aligned}$$

	B	ADCE
B	0	6.7
ADCE		0

STEP 11 : Merge all clusters at $d = 6.7$.

COMPLETE LINKAGE PROBLEM

STEP 12 : Draw the dendrogram.

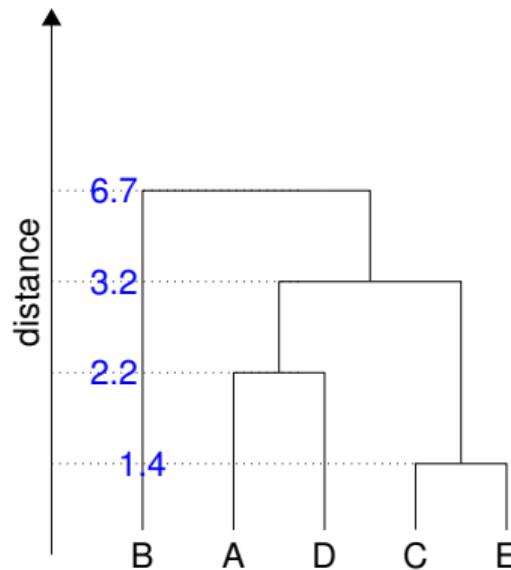
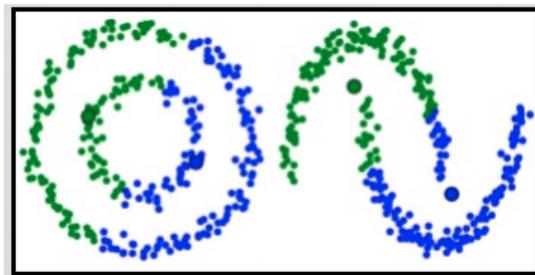


TABLE OF CONTENTS

- 1 CLUSTERING ANALYSIS - CONCEPTS
- 2 PARTITIONING METHODS
- 3 K-MEANS ALGORITHM
- 4 HIERARCHICAL CLUSTERING
- 5 DENSITY BASED CLUSTERING
- 6 EVALUATION OF CLUSTERING ALGORITHMS

SPHERICAL-SHAPED CLUSTERS



- Partitioning and hierarchical methods have difficulty finding clusters of **arbitrary shape** such as the "S" shape and oval clusters.
- Use density based clustering methods.

Notion of Density

- Clusters are considered as dense regions in the data space, separated by sparse regions.

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- The density of an object o can be measured by the number of objects close to o .
- DBSCAN finds core objects and then connects core objects and their neighborhoods to form dense regions as clusters.

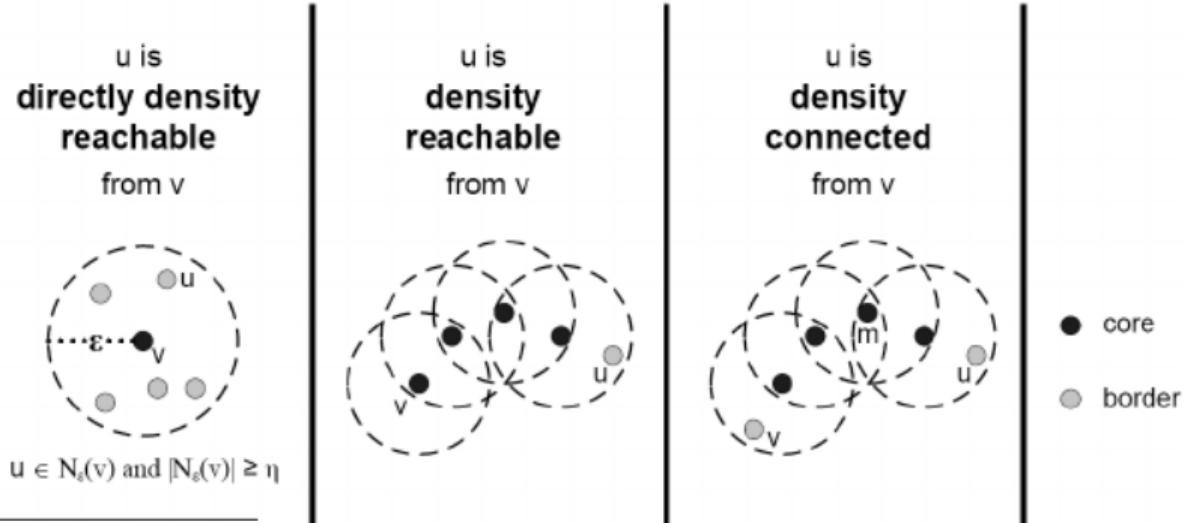
DBSCAN CONCEPTS

- ① Density of a neighborhood
 - ▶ Number of objects in the neighborhood.
- ② ϵ neighborhood of an object o
 - ▶ ϵ neighborhood is the space within a radius centered at o .
 - ▶ ϵ is a hyper-parameter.
- ③ Core object
 - ▶ Core objects have dense neighborhoods.
- ④ MinPts
 - ▶ specifies the density threshold of dense regions.
 - ▶ An object is a **core object** if the ϵ -neighborhood of the object contains at least MinPts objects.
 - ▶ Hyper-parameter

DBSCAN CONCEPTS

⑤ Directly density reachable (DDR)

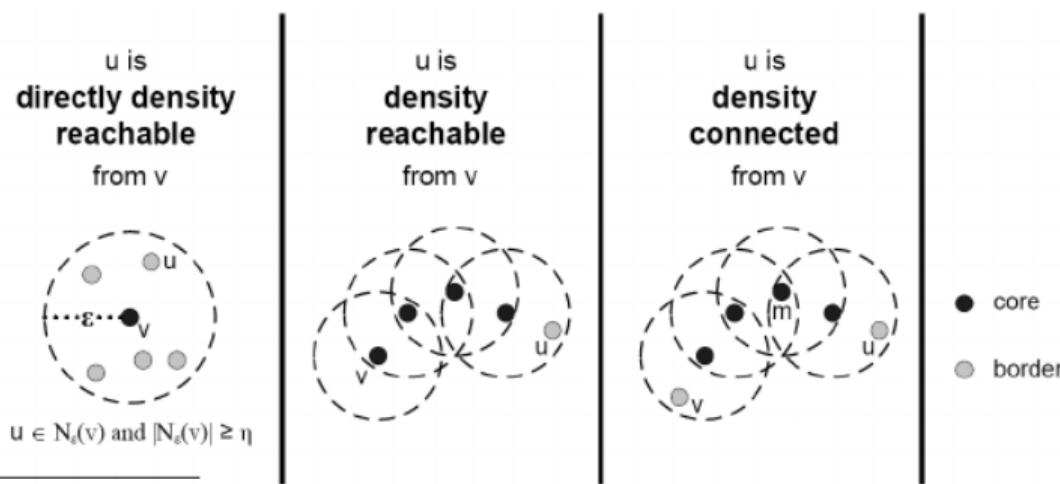
- For a core object q and an object p , p is **directly density-reachable** from q (with respect to ϵ and MinPts) if p is within the ϵ -neighborhood of q .



DBSCAN CONCEPTS

⑥ Density reachable (DR)

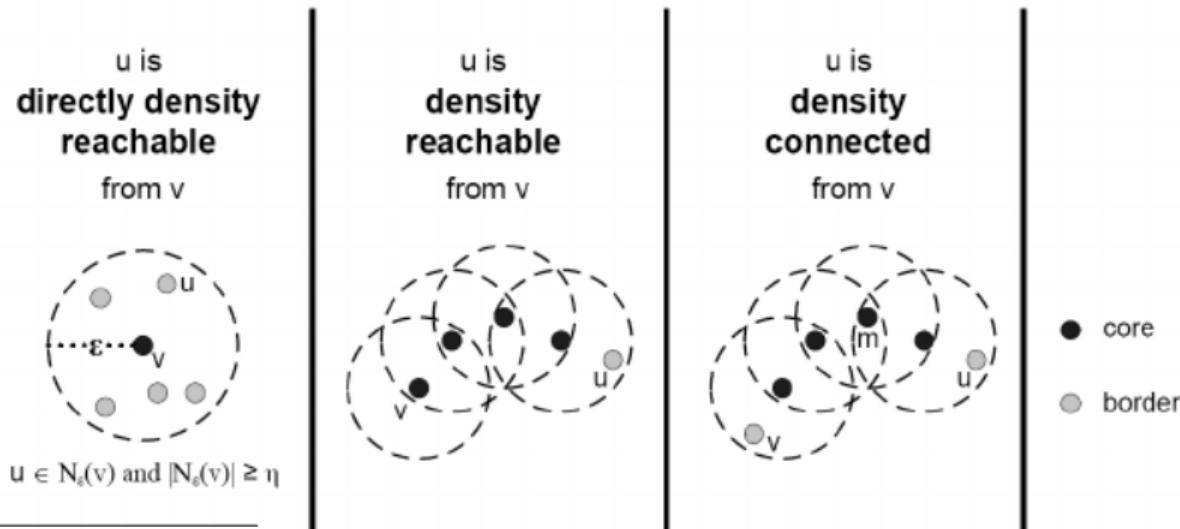
- Object p is **density-reachable** from q (with respect to ϵ and MinPts) if there is a chain of objects p_1, \dots, p_n , such that $p_1 = q$ and $p_n = p$ and p_{i+1} is directly density-reachable from p_i .



DBSCAN CONCEPTS

⑦ Density Connected (DC)

- Two objects $p_1, p_2 \in D$ are **density-connected** (with respect to ϵ and MinPts) if there is an object $q \in D$ such that both p_1 and p_2 are density reachable from q .



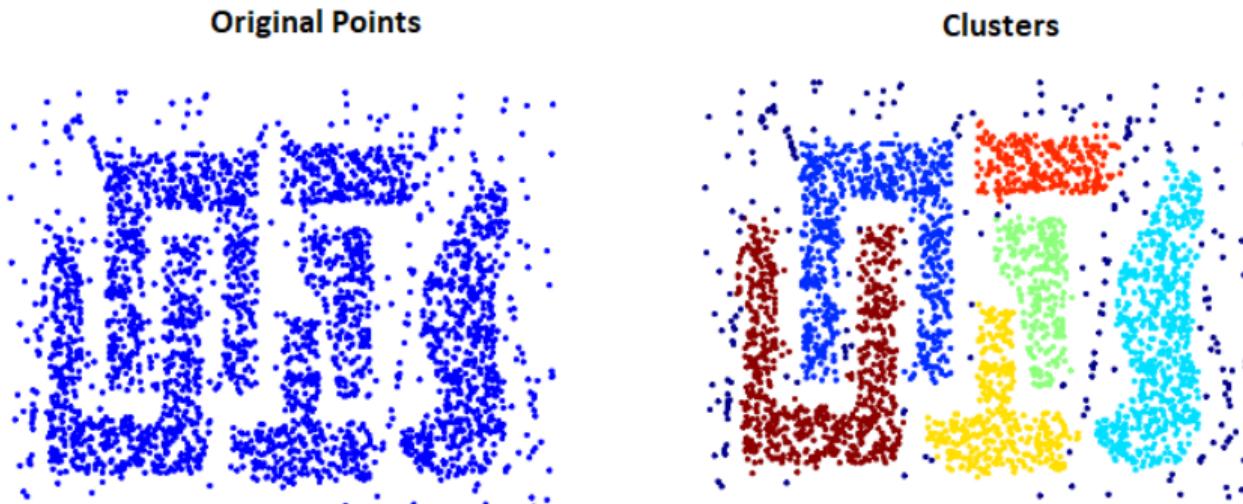
© Joerg Laessig

DBSCAN CONCEPTS

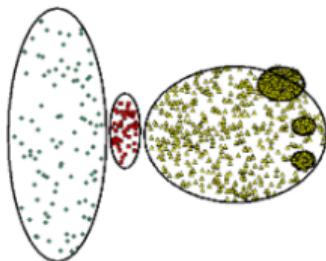
⑧ Density based cluster

- ▶ A set of points form a cluster $C \subset D$ if
 - ① for any two objects $o_1, o_2 \in C$, o_1 and o_2 are density connected.
 - ② there does not exist an object $o \in C$ and another object $o' \in (D - C)$ such that o and o' are density connected.

DBSCAN SUCCEEDS

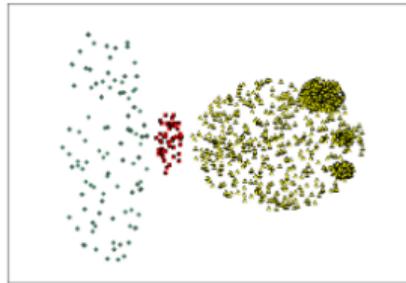


DBSCAN FAILS

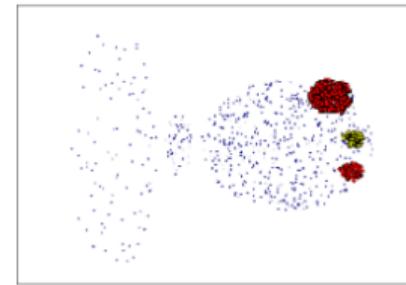


Original Points

- Cannot handle varying densities
- sensitive to parameters—hard to determine the correct set of parameters



($\text{MinPts}=4$, $\text{Eps}=9.92$)



($\text{MinPts}=4$, $\text{Eps}=9.75$)

DBSCAN NUMERICAL PROBLEM

Use the DBSCAN algorithm to cluster the following:

$A = (2, 2); B(3, 1); C(3, 4); D(5, 3); E(10, 14)$.

Take $\text{eps} = 3$ and $\text{MinPts} = 4$

Step 1: Compute the Euclidean distance matrix.

	A	B	C	D	E
A	0	1.4	2.2	3.2	14.4
B		0	3.0	2.8	14.7
C			0	2.2	12.2
D				0	12.1
E					0

DBSCAN NUMERICAL PROBLEM

Step 2: Compute eps neighbors and density of each point.

eps Neighbor	density
$N(A) = \{B, C\}$	$d(A) = 3$
$N(B) = \{A, C, D\}$	$d(B) = 4$
$N(C) = \{A, B, C\}$	$d(C) = 4$
$N(D) = \{B, C\}$	$d(D) = 3$
$N(E) = \{\}$	$d(E) = 1$

DBSCAN NUMERICAL PROBLEM

Step 3: Iterate over each core point.

Point B *Unvisited* = {A, C, D, E}

Visited = {B}

$N(B)$ = {A, C, D}

C_1 = {B}

$p' = A$; *Visited* = {B, A}

Unvisited = {C, D, E}

$N(B) = N(B) \cup N(A)$

 = {A, C, D} \cup {B, C}

 = {A, B, C, D}

$C_1 = \{B, A\}$

DBSCAN NUMERICAL PROBLEM

$p' = C$; *Visited* = {B, A, C}

Unvisited = {D, E}

$$\begin{aligned}N(B) &= N(B) \cup N(C) \\&= \{A, B, C, D\} \cup \{A, B, D\} \\&= \{A, B, C, D\}\end{aligned}$$

$C_1 = \{B, A, C\}$

$p' = D$; *Visited* = {B, A, C, D}

Unvisited = {E}

$$\begin{aligned}N(B) &= N(B) \cup N(D) \\&= \{A, B, C, D\} \cup \{B, C\} \\&= \{A, B, C, D\}\end{aligned}$$

$C_1 = \{B, A, C, D\}$

DBSCAN NUMERICAL PROBLEM

Step 4: Unvisited has only point E, which is not a core point. All others points are visited.
So stop the iterations.

$$\text{Core - pts} = \{B, C\}$$

$$\text{Noise} = \{E\}$$

$$\text{Clusters } C_1 = \{B, A, C, D\}$$

TABLE OF CONTENTS

- ① CLUSTERING ANALYSIS - CONCEPTS
- ② PARTITIONING METHODS
- ③ K-MEANS ALGORITHM
- ④ HIERARCHICAL CLUSTERING
- ⑤ DENSITY BASED CLUSTERING
- ⑥ EVALUATION OF CLUSTERING ALGORITHMS

CLUSTER EVALUATION TASKS

- Assessing clustering tendency
 - ▶ Clustering analysis on a data set is meaningful only when there is a non-random structure in the data.
 - ▶ Assess whether a non-random structure exists in the data.
- Determining the number of clusters in a data set
 - ▶ Estimate the number of clusters even before a clustering algorithm is used to derive detailed clusters.
- Measuring clustering quality
 - ▶ Measure how well the clusters relate to the data set.
 - ▶ Measure how well the clusters match the ground truth, if such truth is available.
 - ▶ Methods to score clustering and compare two sets of clustering results on the same data set.

1. ASSESSING CLUSTERING TENDENCY

- Clustering tendency assessment determines whether a given data set has a non-random structure, which may lead to meaningful clusters.
- Clustering requires non-uniform distribution of data.
- The Hopkins Statistic is a spatial statistic that tests the spatial randomness of a variable as distributed in a space.

1. ASSESSING CLUSTERING TENDENCY

Calculate Hopkins Statistic.

- ① Sample n points, p_1, \dots, p_n uniformly from data set D .
- ② For each point p_i , find the distance x_i between p_i and its nearest neighbor in D .

$$x_i = \min_{v \in D} \{dist(p_i, D)\}$$

- ③ Sample n points, q_1, \dots, q_n uniformly from data set D .
- ④ For each point q_i , find the distance y_i between q_i and its nearest neighbor in $D - \{q_i\}$.

$$y_i = \min_{v \in D, v \neq q_i} \{dist(q_i, D)\}$$

1. ASSESSING CLUSTERING TENDENCY

- ⑤ Calculate Hopkins Statistic H as

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$

Interpretation of Hopkins Statistic H

- If D is uniformly distributed, $H \approx 0.5$.
- If D is highly clustered, $H \approx 0$.
- If $H > 0.5$ then D may not have statistically significant clusters.
- If D cannot be clustered, $H \approx 1.0$.

HOPKINS STATISTIC - NUMERICAL PROBLEM

Determine whether the following dataset has clustering tendency.

$$D = \{2; 10; 12; 4; 25; 3; 30; 20; 11\}$$

Step 1: Randomly sample 5 points from D with replacement.

$$\text{Sample} = \{2; 12; 25; 30; 11\}$$

Find out which points are nearest to each point in the sample and sum all those nearest distances.

$$\begin{aligned}\sum x_i &= d(2, 3) + d(12, 11) + d(25, 20) + d(30, 25) + d(11, 10) \\ &= 1 + 1 + 5 + 5 + 1 = 13\end{aligned}$$

HOPKINS STATISTIC - NUMERICAL PROBLEM

Step 2: Randomly sample 5 points from D but do not place them back (without replacement).

$$S1 = \{12\} \quad D = \{2; 10; 4; 25; 3; 30; 20; 11\}$$

$$y_i = d(12, 11) = 1$$

$$S2 = \{30\} \quad D = \{2; 10; 4; 25; 3; 20; 11\}$$

$$y_i = d(30, 25) = 5$$

$$S3 = \{4\} \quad D = \{2; 10; 25; 3; 20; 11\}$$

$$y_i = d(4, 3) = 1$$

$$S4 = \{11\} \quad D = \{2; 10; 25; 3; 20\}$$

$$y_i = d(11, 10) = 1$$

$$S5 = \{3\} \quad D = \{2; 10; 25; 20\}$$

$$y_i = d(3, 2) = 1$$

$$\sum y_i = 1 + 5 + 1 + 1 + 1 = 9$$

HOPKINS STATISTIC - NUMERICAL PROBLEM

Step 3: Compute Hopkins Statistic.

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$
$$= 0.4$$

The data can be clustered, but will not get highly distinct clusters.

2. DETERMINING THE NUMBER OF CLUSTERS

- ① Set the number of clusters to about $\sqrt{\frac{n}{2}}$ for a data set of n points.
- ② Use elbow method.
 - ▶ Elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. A heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters.
- ③ Use cross validation technique.
 - ▶ Compare the overall quality measure with respect to different values of k , and find the number of clusters that best fits the data.

3. MEASURING CLUSTERING QUALITY

① Extrinsic methods

- ▶ Ground truth is available.
- ▶ Compare the clustering against the group truth and measure.
- ▶ Extrinsic methods are also known as supervised methods.

② Intrinsic methods

- ▶ Ground truth is unavailable
- ▶ Evaluate the goodness of a clustering by considering how well the clusters are separated.
- ▶ Intrinsic methods are also known as unsupervised methods.

3A. EXTRINSIC METHODS

Four essential criteria for clustering quality $Q(C, C_g)$.

- ① Cluster homogeneity
 - ▶ the more pure the clusters in a clustering are, the better the clustering.
- ② Cluster completeness
 - ▶ clustering should assign objects belonging to the same category (according to ground truth) to the same cluster.
- ③ Rag bag
 - ▶ putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag.
- ④ Small cluster preservation
 - ▶ splitting a small category into pieces is more harmful than splitting a large category into pieces.

3A. EXTRINSIC METHODS

BCubed Metrics

- Precision of an object indicates how many other objects in the same cluster belong to the same category as the object.
- Recall of an object rejects how many objects of the same category are assigned to the same cluster.
- Let $D = \{o_1; \dots; o_n\}$.
- Let the ground truth be $L(o_i)$.
- Let the cluster id be $C(o_i)$.
- Correctness of the relation between two objects o_i and o_j .

3A. EXTRINSIC METHODS

$$\text{Correctness}(o_i, o_j) = \begin{cases} 1 & \text{if } L(o_i) = L(o_j) \implies C(o_i) = C(o_j) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{BCubed precision} = \frac{\sum_{i=1}^n \frac{\sum_{o_j; i \neq j; C(o_i) = C(o_j)} \text{Correctness}(o_i, o_j)}{||\{o_j | i \neq j; C(o_i) = C(o_j)\}||}}{n}$$

$$\text{BCubed recall} = \frac{\sum_{i=1}^n \frac{\sum_{o_j; i \neq j; L(o_i) = L(o_j)} \text{Correctness}(o_i, o_j)}{||\{o_j | i \neq j; L(o_i) = L(o_j)\}||}}{n}$$

3B. INTRINSIC METHODS

- Evaluate a clustering by examining
 - ▶ how well the clusters are separated.
 - ▶ how compact the clusters are.
- Use Silhouette Coefficient
- The value of the silhouette Coefficient is between -1 and 1.
- The value of $a(o)$ reflects the compactness of the cluster to which object o belongs.
The smaller the value, the more compact the cluster.
- The value of $b(o)$ captures the degree to which o is separated from other clusters.
The larger $b(o)$ is, the more separated o is from other clusters.
- When the silhouette Coefficient value of o approaches 1, the cluster containing o is compact and o is far away from other clusters. (preferred)

3B. INTRINSIC METHODS

- Calculate $a(o) = \text{average distance between } o \text{ and all other objects in the cluster to which } o \text{ belongs.}$
- Calculate $b(o) = \text{minimum average distance from } o \text{ and all clusters to which } o \text{ does not belong.}$

$$a(o) = \frac{\sum_{o' \in C_i; o \neq o'} dist(o, o')}{|C_i| - 1}$$

$$b(o) = \min_{c_j: 1 \leq k \leq k; j \neq i} \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|}$$

$$\text{Silhouette Coefficient } s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

SILHOUETTE COEFFICIENT - NUMERICAL PROBLEM

Using the data on distances given below, compute Silhouette Coefficient for each point, for each clusters and the overall clustering. Cluster 1 contains $[p_1; p_2]$ and Cluster 2 contains $[p_3; p_4]$.

	p_1	p_2	p_3	p_4
p_1	0			
p_2	0.1	0		
p_3	0.65	0.70	0	
p_4	0.55	0.60	0.30	0

Step s: Compute $a(o)$ and $b(o)$ for all points.

$a(o) = \text{avg distance between } p_i \text{ and all other points in } C_1$.

$b(o) = \text{avg distance between } p_i \text{ and all other points in } C_2$.

Compute Silhouette Coefficient.

SILHOUETTE COEFFICIENT - NUMERICAL PROBLEM

For p_1 : $a = 0.1$

$$b = \frac{0.65 + 0.55}{2} = 0.6$$

$$s = \frac{b - a}{\max(a, b)} = \frac{0.6 - 0.1}{\max(0.1, 0.6)} = 0.833$$

For p_2 : $a = 0.1$

$$b = \frac{0.7 + 0.6}{2} = 0.65$$

$$s = \frac{b - a}{\max(a, b)} = \frac{0.65 - 0.1}{\max(0.1, 0.65)} = 0.846$$

For p_3 : $a = 0.3$

$$b = \frac{0.65 + 0.7}{2} = 0.675$$

$$s = \frac{b - a}{\max(a, b)} = \frac{0.675 - 0.3}{\max(0.3, 0.675)} = 0.555$$

SILHOUETTE COEFFICIENT - NUMERICAL PROBLEM

For p_4 : $a = 0.3$

$$b = \frac{0.55 + 0.6}{2} = 0.575$$

$$s = \frac{b - a}{\max(a, b)} = \frac{0.575 - 0.3}{\max(0.3, 0.575)} = 0.478$$

$$\text{For } C_1 : s = \frac{0.833 + 0.846}{2} = 0.839$$

$$\text{For } C_2 : s = \frac{0.555 + 0.478}{2} = 0.516$$

$$\text{Overall clustering: } s = \frac{0.839 + 0.516}{2} = 0.677$$

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



INTRODUCTION TO DATA SCIENCE MODULE # 9 : ANOMALY DETECTION

IDS Course Team

BITS Pilani

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

1 ANOMALY DETECTION

ANOMALY DETECTION

- **Anomaly detection** (also known as outlier detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called **outliers or anomalies**.
- Anomaly detection tries to capture the **exceptional cases** that deviate substantially from the majority patterns.

CLUSTERING VS ANOMALY DETECTION

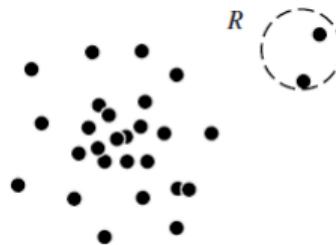
- Clustering finds the majority patterns in a data set and organizes the data accordingly.
- Anomaly detection tries to capture those exceptional cases that deviate substantially from the majority patterns.

ANOMALY DETECTION - APPLICATIONS

- Fraud Detection
- Intrusion Detection
- Ecosystem Disturbances
- Public Health
- Medical care
- Industry damage detection
- Sensor / video network surveillance

ANOMALY

- An **anomaly or outlier** is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.



OUTLIER VS NOISE

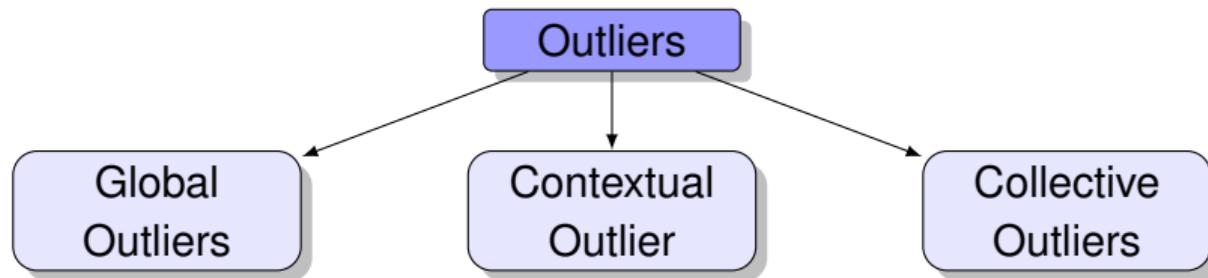
Outlier

- An **outlier** is a data object that deviates significantly from the rest of the objects.
- Outliers are interesting because they are suspected of not being generated by the same mechanisms as the rest of the data.

Noise

- **Noise** is a random error or variance in a measured variable. Noise is not interesting.
- Noise should be removed before outlier detection.

TYPES OF OUTLIERS



TYPES OF OUTLIERS

① Global Outliers or point anomalies

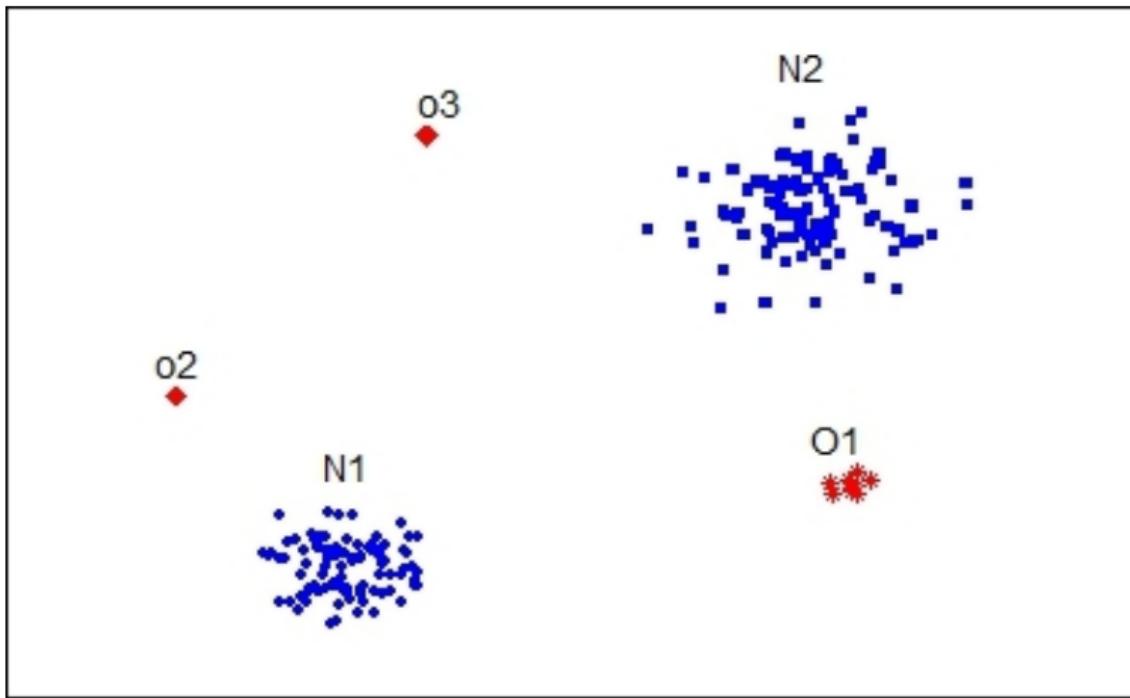
- ▶ Object is O_g if it deviates significantly from the rest of the data set.
- ▶ Simplest type of outliers.
- ▶ Eg: Intrusion detection - suspected victim of hacking.
- ▶ Issue: Find an appropriate measurement of deviation.
- ▶ Contextual outlier with empty contextual attributes.

TYPES OF OUTLIERS

② Contextual Outliers or conditional outliers

- ▶ Object is O_c if it deviates significantly based on specific context of the objects.
- ▶ An object in a data set is a local outlier if its density significantly deviates from the local area in which it occurs.
- ▶ Contextual attributes - define the object's context.
- ▶ Behavioral attributes - define the characteristics of objects. Used to evaluate the object as an outlier.
- ▶ Issue: Context outlier detection requires background information to determine contextual attributes and contexts.

TYPES OF OUTLIERS



TYPES OF OUTLIERS

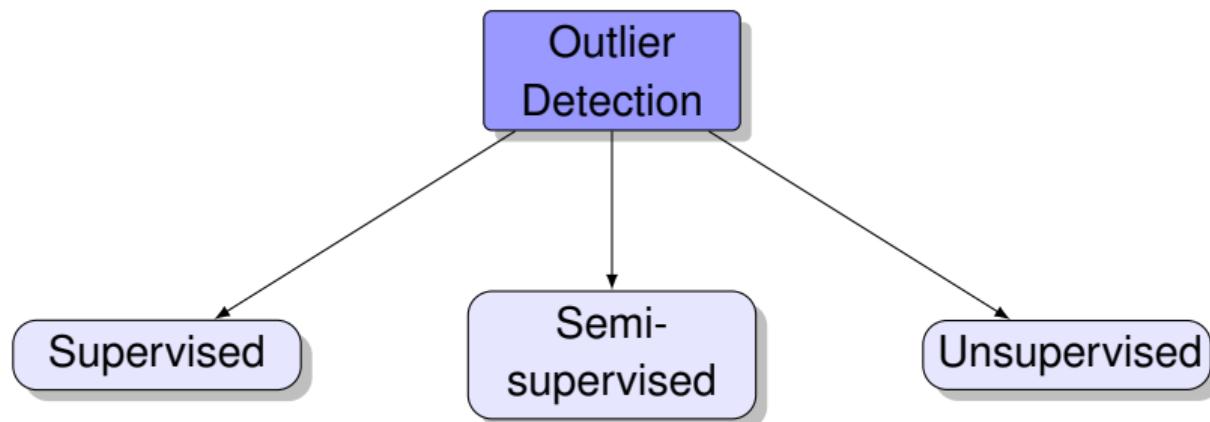
⑤ Collective Outliers

- ▶ a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. The individual data objects may not be outliers.
- ▶ Eg: Denial of service package from several computers. So consider them as a whole.
- ▶ Collective outlier detection requires background information to model the relationship among objects to find groups of outliers.

CHALLENGES OF ANOMALY DETECTION

- ① Modeling normal objects and outliers effectively.
- ② Application-specific outlier detection, impossible to develop a universally applicable outlier detection method.
- ③ Handling noise in outlier detection. Noise and missing data may “hide” outliers and reduce the effectiveness of outlier detection
- ④ Understandability, understand why the detected objects are outliers.

ANOMALY DETECTION METHODS



ANOMALY DETECTION METHODS

① Supervised Methods

- ▶ Modeling outlier detection as a **classification problem**.
 - ★ Samples examined by domain experts used for training and testing.
- ▶ Methods for Learning a classifier for outlier detection effectively:
 - ★ Model normal objects and report those not matching the model as outliers
 - ★ Model outliers and treat those not matching the model as normal.
- ▶ Challenges
 - ★ **Imbalanced classes**, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers.
 - ★ Catch as many outliers as possible, i.e., **recall** is more important than accuracy (i.e., not mislabeling normal objects as outliers)

ANOMALY DETECTION METHODS

② Unsupervised Methods

- ▶ The normal objects are somewhat “clustered.”
- ▶ An unsupervised outlier detection method expects that normal objects follow a pattern or form multiple groups.
- ▶ An outlier is expected to occur far away in feature space from any of those groups of normal objects.
- ▶ Weakness: Cannot detect collective outlier effectively.
 - ★ Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area.
- ▶ Many clustering methods can be adapted for unsupervised methods.
 - ★ Find clusters, then outliers: not belonging to any cluster.
 - ★ Problem 1: Hard to distinguish noise from outliers.
 - ★ Problem 2: Costly since first clustering: but far less outliers than normal objects .
 - ★ Newer methods: tackle outliers directly.

ANOMALY DETECTION METHODS

⑤ Semi-supervised methods

- ▶ Only a small set of the normal and/or outlier objects are labeled, but most of the data are unlabeled.
- ▶ In many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both.
- ▶ Semi-supervised outlier detection: Regarded as applications of semi-supervised learning.
- ▶ If some labeled normal objects are available
 - ★ Use the labeled examples and the proximate unlabeled objects to train a model for normal objects.
 - ★ Those not fitting the model of normal objects are detected as outliers.
- ▶ If only some labeled outliers are available, a small number of labeled outliers may not cover the possible outliers well.
 - ★ To improve the quality of outlier detection, one can get help from models for normal objects learned from unsupervised methods.

MINING CONTEXTUAL OUTLIERS

Transform into Conventional Outlier Detection

- If the contexts can be clearly identified, transform it to conventional outlier detection.
 - ① Identify the context of the object using the contextual attributes.
 - ② Calculate the outlier score for the object in the context using a conventional outlier detection method.
- Steps:
 - ① locate c's context,
 - ② compare c with the other customers in the same group, and
 - ③ use a conventional outlier detection method

MINING CONTEXTUAL OUTLIERS

Modeling Normal Behavior with Respect to Contexts

- In some applications, one cannot clearly partition the data into contexts.
 - ▶ Eg: If a customer suddenly purchased a product that is unrelated to those she recently browsed, it is unclear how many products browsed earlier should be considered as the context.
- Model the “normal” behavior with respect to contexts.
 - ① Using a training data set, train a model that predicts the expected behavior attribute values with respect to the contextual attribute values.
 - ② An object is a contextual outlier, if its behavior attribute values significantly deviate from the values predicted by the model.
- Using a prediction model that links the contexts and behavior, these methods avoid the explicit identification of specific contexts.

MINING COLLECTIVE OUTLIERS

On the set of "Structured objects"

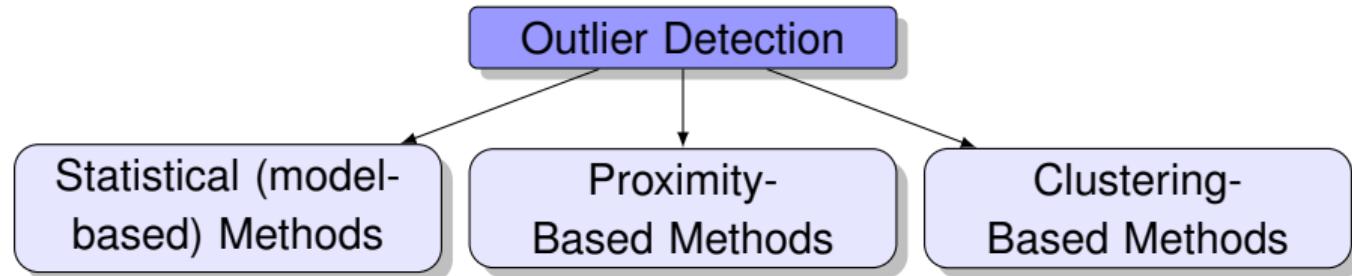
- Collective outlier - objects as a group deviate from the entire data.
- Need to examine the structure of the data set, i.e, the relationships between multiple data objects.
- Each of these structures is inherent to its respective type of data.
 - ▶ For temporal data (such as time series and sequences) explore the structures formed by time, which occur in segments of the time series or sub-sequences.
 - ▶ For spatial data, explore local areas.
 - ▶ For graph and network data, we explore sub-graphs.
- Difference from the contextual outlier detection: the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.

MINING COLLECTIVE OUTLIERS

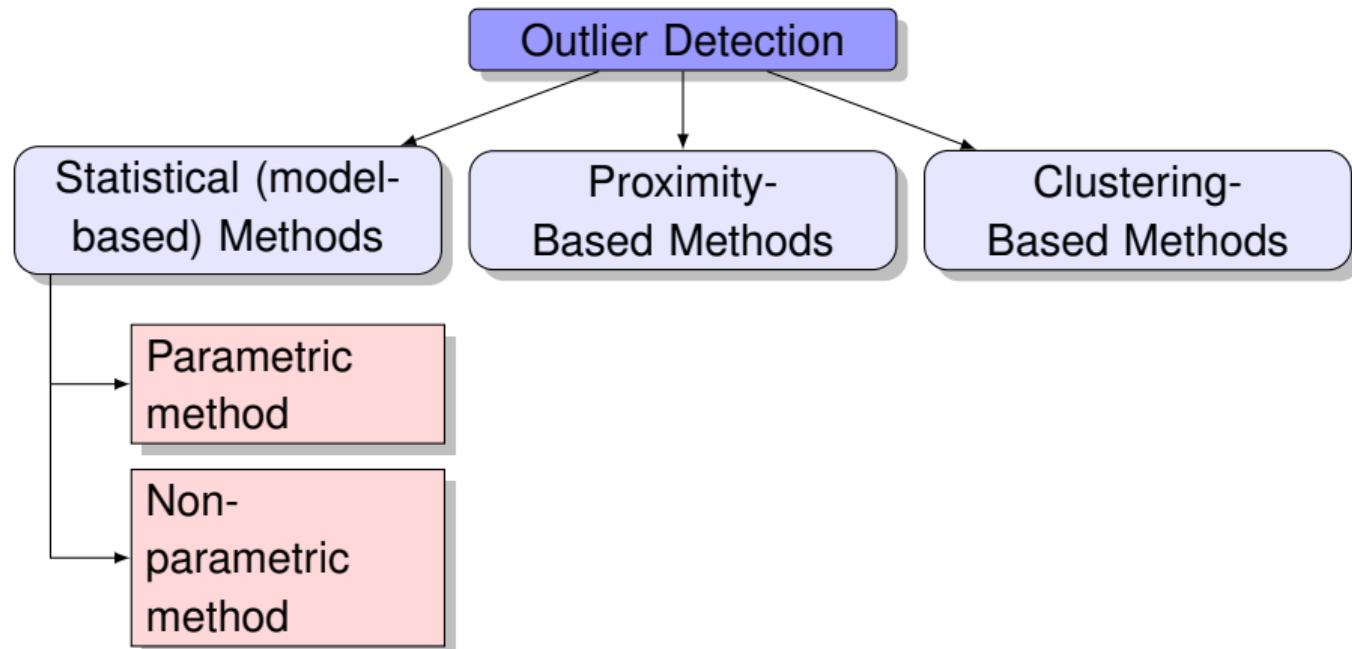
On the set of "Structured Objects"

- Two categories of Collective outlier detection methods
 - ① Reduce the problem to conventional outlier detection.
 - ★ Identify structure units, treat each structure unit (e.g., sub-sequence, time series segment, local area, or sub-graph) as a data object, and extract features.
 - ★ Then outlier detection on the set of “structured objects” constructed as such using the extracted features.
 - ② Models the expected behavior of structure units directly.
- Collective outlier detection is subtle due to the challenge of exploring the structures in data.
 - ▶ The exploration typically uses heuristics, and thus may be application dependent.
 - ▶ The computational cost is often high due to the sophisticated mining process.

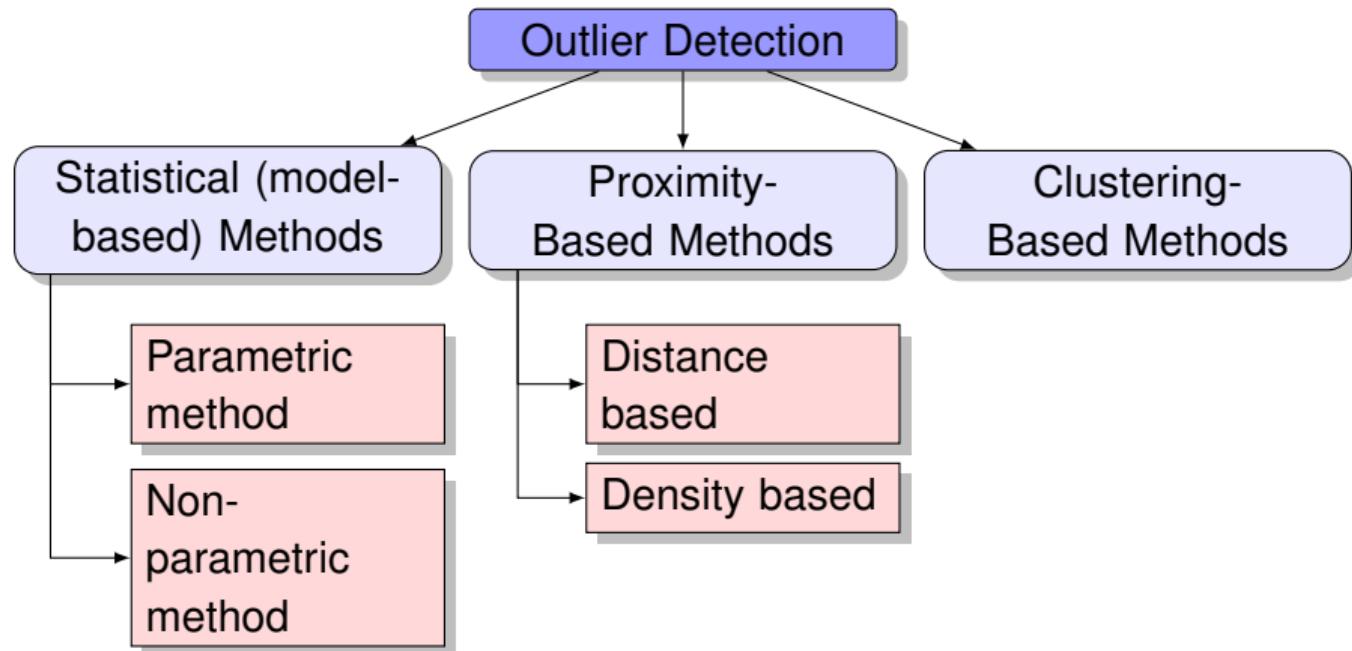
ANOMALY DETECTION METHODS IN THIS COURSE



ANOMALY DETECTION METHODS IN THIS COURSE

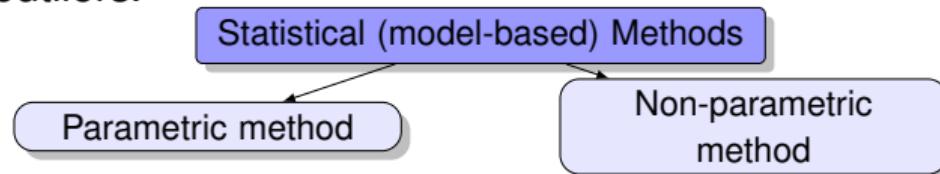


ANOMALY DETECTION METHODS IN THIS COURSE



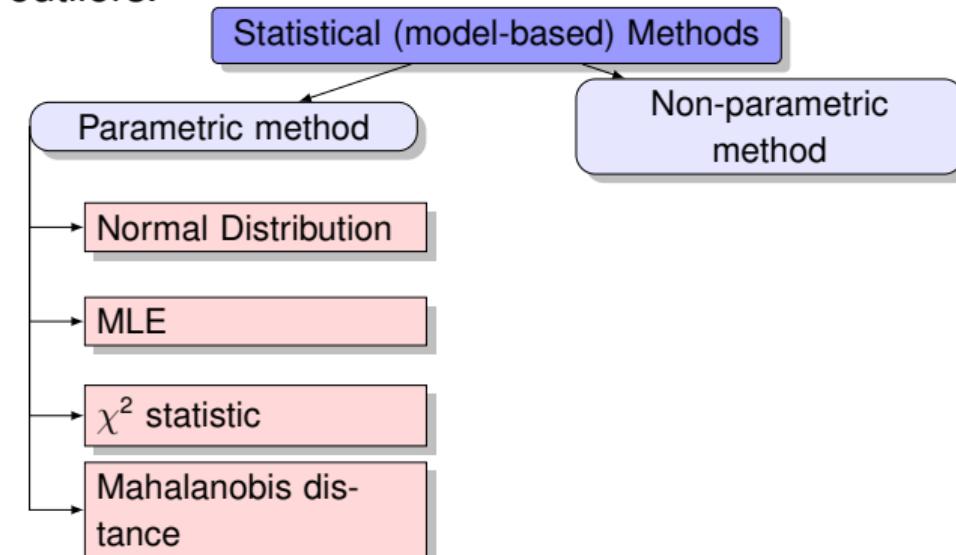
1. STATISTICAL OR MODEL-BASED METHODS

- Statistical approaches assume that the objects in a data set are generated by a stochastic process.
- Normal data objects are generated by a statistical model. Data that do not follow the model are outliers.



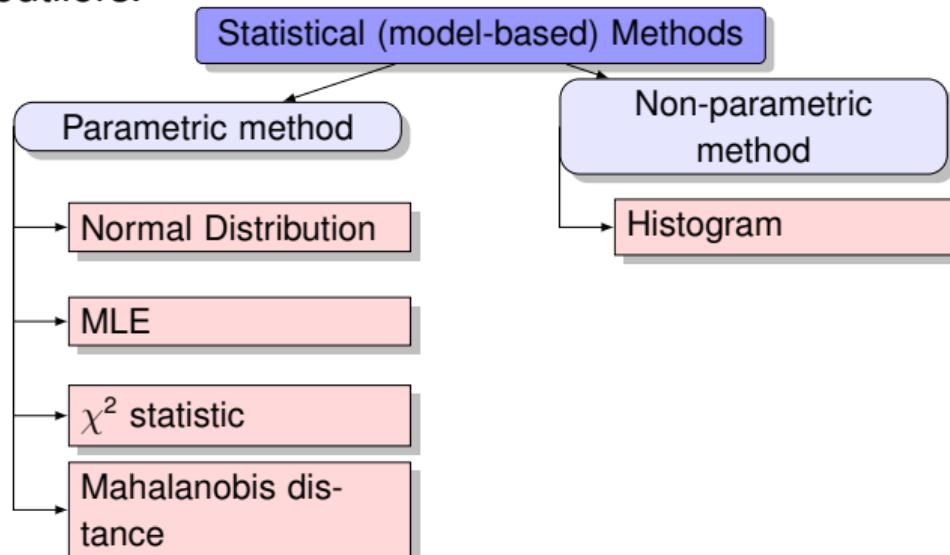
1. STATISTICAL OR MODEL-BASED METHODS

- Statistical approaches assume that the objects in a data set are generated by a stochastic process.
- Normal data objects are generated by a statistical model. Data that do not follow the model are outliers.



1. STATISTICAL OR MODEL-BASED METHODS

- Statistical approaches assume that the objects in a data set are generated by a stochastic process.
- Normal data objects are generated by a statistical model. Data that do not follow the model are outliers.



1. STATISTICAL OR MODEL-BASED METHODS

1 Parametric Model

- ▶ A parametric method assumes that the normal data objects are generated by a parametric distribution with parameter Θ .
- ▶ The probability density function of the parametric distribution $f(x, \Theta)$ gives the probability that object x is generated by the distribution.
- ▶ The smaller this value, the more likely x is an outlier.

1. STATISTICAL OR MODEL-BASED METHODS

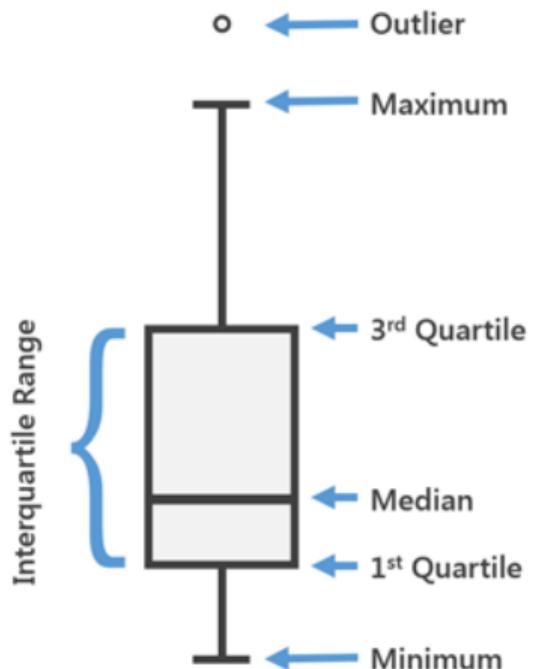
1 Parametric Model

- ▶ A parametric method assumes that the normal data objects are generated by a parametric distribution with parameter Θ .
- ▶ The probability density function of the parametric distribution $f(x, \Theta)$ gives the probability that object x is generated by the distribution.
- ▶ The smaller this value, the more likely x is an outlier.

2 Non-parametric Model

- ▶ A nonparametric method does not assume an a priori statistical model.
- ▶ Nonparametric method tries to determine the model from the input data.

1.1.1 PARAMETRIC MODEL – BOXPLOT



IQR EXAMPLE

Find the outlier in the following data using Inter-Quartile Range.

Data = 10, 12, 11, 15, 11, 14, 13, 17, 12, 22, 14, 11.

- 1 Arrange in order
- 2 Find half-way point
- 3 Find Q_1
- 4 Find Q_3
- 5 Find IQR
- 6 Find Minimum value
- 7 Find Maximum value

IQR EXAMPLE

Find the outlier in the following data using Inter-Quartile Range.

Data = 10, 12, 11, 15, 11, 14, 13, 17, 12, 22, 14, 11.

- 1 Arrange in order
- 2 Find half-way point
- 3 Find Q_1
- 4 Find Q_3
- 5 Find IQR
- 6 Find Minimum value
- 7 Find Maximum value

10, 11, 11, 11, 12, 12, 13, 14, 14, 15, 17, 22

$$Q_2 = \frac{12+13}{2} = 12.5$$

$$Q_1 = \frac{11+11}{2} = 11.0$$

$$Q_3 = \frac{14+15}{2} = 14.5$$

$$IQR = Q_3 - Q_1 = 3.5$$

$$\text{Min} = Q_1 - 1.5 * IQR = 5.75$$

$$\text{Max} = Q_3 + 1.5 * IQR = 19.75$$

So point 22 is an outlier.

1.1.2 DISCORDANCY TEST

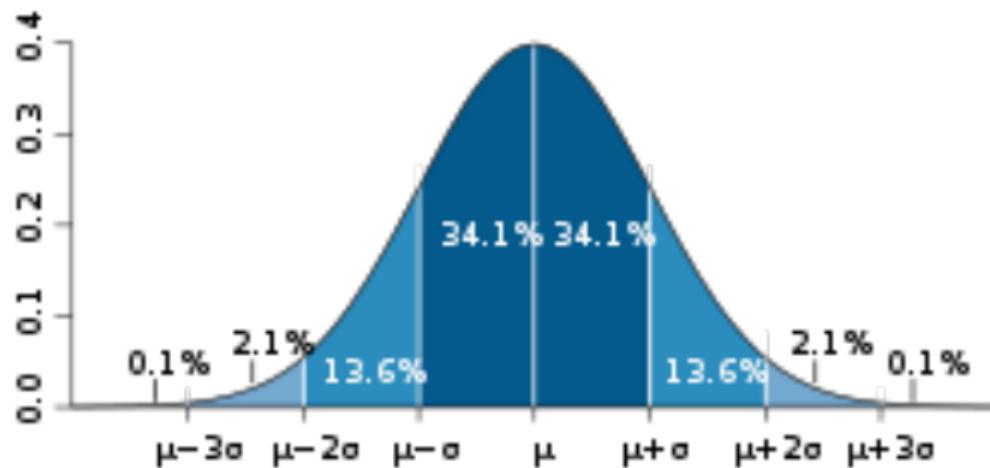
- The statistical distribution-based approach identifies outliers with respect to the model using a discordancy test.
- A statistical discordancy test examines first a working hypothesis.
- A **working hypothesis**, H , is a statement that the entire data set of n objects comes from an initial distribution model, F .

$$H : o_i \in F, \quad \text{where} \quad i = 1, 2, \dots n$$

- The hypothesis is retained if there is no statistically significant evidence supporting its rejection.
- A **discordancy test** verifies whether an object, o_i , is significantly large (or small) in relation to the distribution F .
- The result is very much dependent on which model F is chosen because o_i may be an outlier under one model and a perfectly valid value under another.

1.1.3 NORMAL OR GAUSSIAN CURVE

Detecting outliers using a statistical (Gaussian) model.



The objects that fall outside the $\mu + 3\sigma$ are considered as outliers.

1.1.4 UNI-VARIATE OUTLIERS USING MLE

- Data involving only one attribute or variable are called **uni-variate data**.
- Assume that data are generated from a normal distribution.
- Learn the parameters of the normal distribution μ and σ using Maximum Likelihood Method (MLE).
- **Identify the points with low probability as outliers.**
- Use Maximum Likelihood Method to estimate μ and σ .
- Maximize the log-likelihood function.

1.1.4 UNI-VARIATE OUTLIERS USING MLE

$$\begin{aligned}
 \ln \mathcal{L}(\mu, \sigma^2) &= \sum_{i=1}^n \ln \{(x_i | (\mu, \sigma^2))\} \\
 &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\
 \hat{\mu} &= \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\
 \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2
 \end{aligned}$$

MLE EXAMPLE

Find the outlier in the following data using MLE.

Data = 24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4.

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 28.61$$

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \approx \sqrt{2.29} \approx 1.51$$

$$3\hat{\sigma} = 3 * 1.51 = 4.53$$

$$\hat{\mu} + 3\hat{\sigma} = 28.61 + 4.53 = 33.14$$

$$\hat{\mu} - 3\hat{\sigma} = 28.61 - 4.53 = 24.08$$

The point 24.0 is outside the range $\hat{\mu} \pm 3\hat{\sigma}$, hence an outlier.

1.1.5 MULTIVARIATE OUTLIER DETECTION

Mahalanobis distance

- For an object, x , in the data set, the Mahalanobis distance from x to the mean \bar{x} , with S^{-1} as the covariance matrix is given by

$$MDist(x, \bar{x}) = (x - \bar{x})^T S^{-1} (x - \bar{x})$$

- Calculate the mean vector from the multivariate data set.
- For each object x , calculate $MDist(x, \bar{x})$.
- Detect outliers in the transformed uni-variate data set, $MDist(x, \bar{x})$.
- If $MDist(x, \bar{x})$ is determined to be an outlier, then x is regarded as an outlier as well.

1.1.6 MULTIVARIATE OUTLIER DETECTION

χ^2 distance

- For an object, x , in the data set, the χ^2 distance from x to the mean \bar{x} is

$$\chi^2 = \sum_{i=1}^n \frac{(x_i - E_i)^2}{E_i}$$

- E_i is the mean of i -th dimension among all the objects.
- If χ^2 -statistic is large, the object is an outlier.

1.2 HISTOGRAM

1 Histogram construction

- ▶ Construct a histogram using the input data.
- ▶ Specify the type of histogram (e.g., equal width or equal depth).
- ▶ Specify parameters like the number of bins in the histogram or the size of each bin.

2 Outlier detection

- ▶ If the object falls in one of the histogram's bins, the object is regarded as normal.
Otherwise, it is considered an outlier.
- ▶ Use the histogram to assign an outlier score to the object. Object's outlier score be the inverse of the volume of the bin in which the object falls

HISTOGRAM – NUMERICAL PROBLEM

Mr. Ben wants to make an investment in the stock market. He has shortlisted below stocks and wants to know the frequency of the prices. Plot histogram to show the distribution.

SI No	Stock Price
A	190
B	250
C	171
D	690
E	500
F	301
G	722
H	100
I	1000
J	310
K	800
L	500
M	200
N	510

HISTOGRAM - NUMERICAL PROBLEM

Step 1:

$$\text{Number of bins} = \sqrt{n} = \sqrt{14} = 4$$

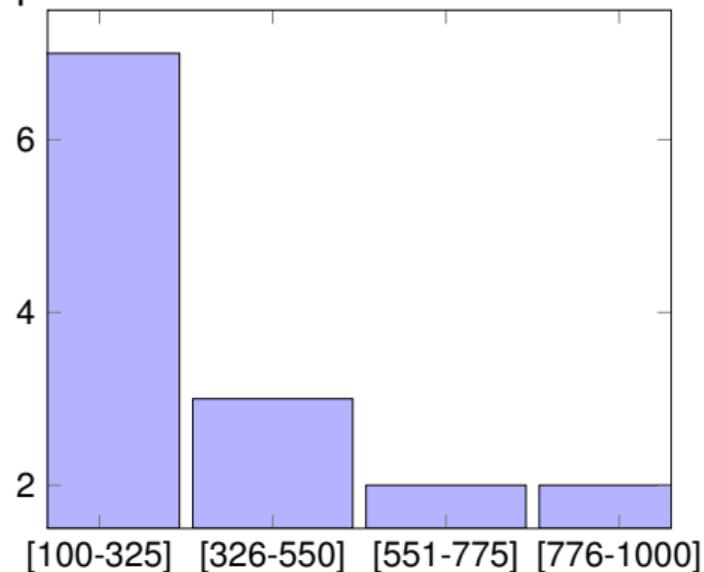
$$\text{Bin size} = \frac{\max - \min}{\sqrt{n}} = \frac{1000 - 100}{4} = 225$$

Step 2:

Bin No	Bin	Values	Frequency
1	[100, 325]	190, 250, 171, 301, 100, 310, 200	7
2	[326, 550]	500, 500, 510	3
3	[551, 775]	690, 722	2
4	[776, 1000]	1000, 800	2

HISTOGRAM - NUMERICAL PROBLEM

Step 3: Draw the histogram plot.



2. PROXIMITY-BASED METHODS

- Proximity-based methods assume that an object is an outlier if the nearest neighbors of the object are **far away** in feature space.
- Assumption of proximity-based approach: **Proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set.**
- Two types of proximity-based outlier detection methods.
 - ① **Distance-based outlier detection:**
An object o is an outlier if its neighborhood does not have enough other points.
 - ② **Density-based outlier detection:**
An object o is an outlier if its density is relatively much lower than that of its neighbors.

2.1 DISTANCE-BASED OUTLIER DETECTION

- For each object o , examine the number of other objects in the r -neighborhood of o , where r is a user-specified distance threshold.
- An object o is an outlier if most (taking π as a fraction threshold) of the objects in D are far away from o , i.e., not in the r -neighborhood of o .
- An object o is a $DB(r, \pi)$ outlier if

$$\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$$

- Equivalently, one can check the distance between o and its k -th nearest neighbor o_k , where $k = \lceil \pi \|D\| \rceil$.

o is an outlier if $dist(o, o_k) > r$

2.1 DISTANCE-BASED OUTLIER DETECTION

Algorithm: Distance-based outlier detection.

Input:

- a set of objects $D = \{o_1, \dots, o_n\}$, threshold r ($r > 0$) and π ($0 < \pi \leq 1$);

Output: $DB(r, \pi)$ outliers in D .

Method:

```
for i = 1 to n do
    count ← 0
    for j = 1 to n do
        if i ≠ j and dist(oi, oj) ≤ r then
            count ← count + 1
            if count ≥ π · n then
                exit {oi cannot be a DB(r, π) outlier}
            endif
        endif
    endfor
    print oi {oi is a DB(r, π) outlier according to (Eq. 12.10)}
endfor;
```

2.1 DISTANCE-BASED OUTLIER DETECTION

Limitations:

- Distance-based outliers, such as $DB(r\pi)$, are just one type of outliers.
- Distance-based outlier detection takes a global view of the data set.
- $DB(r\pi)$ outlier, for example, is far (as quantified by parameter r) from at least $(1 - \pi) \times 100\%$ of the objects in the data set. In other words, an outlier as such is remote from the majority of the data.
- To detect distance-based outliers, we need two global parameters, r and π , which are applied to every outlier object.
- Many real-world data sets demonstrate a more complex structure, where objects may be considered outliers with respect to their local neighborhoods, rather than with respect to the global data distribution.

PROXIMITY BASED OUTLIER DETECTION

k-nearest neighbor approach

Consider distance matrix given below. Assume that outlier score of object is given by average of distances to first k-nearest neighbors. Determine the outlier from given dataset with 2-nearest neighbor approach.

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

Step 1:

$k = 2$; For all objects determine k nearest neighbor $N(x, k)$.

$$N(A, 2) = \{B, C\}$$

$$N(B, 2) = \{A, C\}$$

$$N(C, 2) = \{B, D\}$$

$$N(D, 2) = \{C, E\}$$

$$N(E, 2) = \{C, D\}$$

PROXIMITY BASED OUTLIER DETECTION

k-nearest neighbor approach

Step 2: Determine the outlier score for each object.

$$OS(p) = \frac{d(p, x_1) + \dots + d(p, x_k)}{k}$$

$$OS(A) = \frac{1 + 4}{2} = 2.5$$

$$OS(B) = \frac{1 + 2}{2} = 1.5$$

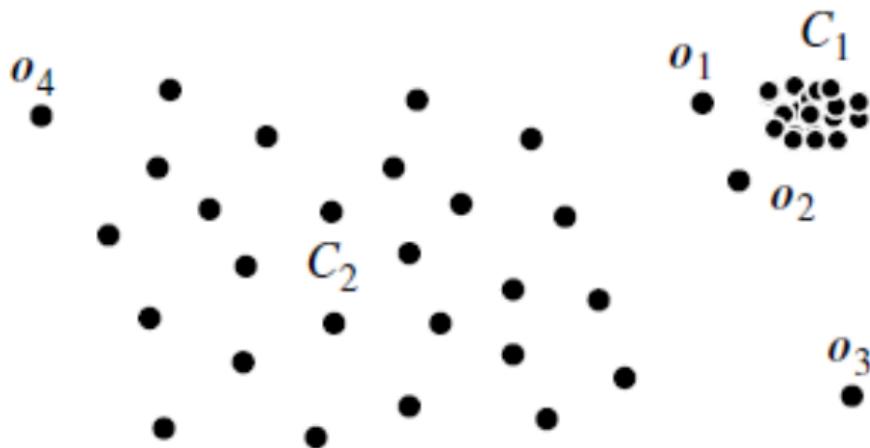
$$OS(C) = \frac{2 + 3}{2} = 2.5$$

$$OS(D) = \frac{3 + 4}{2} = 3.5$$

$$OS(E) = \frac{4 + 4}{2} = 4$$

As outlier score of object E is highest, so object E is termed as outlier.

2.2 DENSITY-BASED OUTLIER DETECTION



In Fig., O_1 and O_2 are local outliers to C_1 , O_3 is a global outlier, but O_4 is not an outlier. However, proximity-based clustering cannot find O_1 and O_4 as outlier (e.g., compared with O_3).

2.2 DENSITY-BASED OUTLIER DETECTION

- Intuition: The density around an outlier object is significantly different from the density around its neighbors.
- Method: Use the **relative density** of an object against its neighbors as the indicator of the degree of the object being outliers.
- Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution.

RELATIVE DENSITY BASED OUTLIERS

Assume that outlier score of an object is relative density around an object. Using relative density outlier score algorithm with given distance matrix, determine the outlier. Assume k-nearest neighbors = 2.

Step 1:

$k = 2$; For all objects determine k nearest neighbor $N(x, k)$.

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

$$N(A, 2) = \{B, C\}$$

$$N(B, 2) = \{A, C\}$$

$$N(C, 2) = \{B, D\}$$

$$N(D, 2) = \{C, E\}$$

$$N(E, 2) = \{C, D\}$$

RELATIVE DENSITY BASED OUTLIERS

Step 2: For all objects, determine density $\text{density}(x, k)$ using k-NN.

$$\text{density}(x, k) = \left(\frac{\sum_{y \in N(x, k)} \text{dist}(x, y)}{|N(x, k)|} \right)^{-1}$$

$$\text{density}(A, 2) = \left(\frac{1 + 4}{2} \right)^{-1} = \frac{2}{5} = 0.4$$

$$\text{density}(B, 2) = \left(\frac{1 + 2}{2} \right)^{-1} = \frac{2}{3} = 0.67$$

$$\text{density}(C, 2) = \left(\frac{2 + 3}{2} \right)^{-1} = \frac{2}{5} = 0.4$$

RELATIVE DENSITY BASED OUTLIERS

$$\text{density}(D, 2) = \left(\frac{3+4}{2} \right)^{-1} = \frac{2}{7} = 0.28$$

$$\text{density}(E, 2) = \left(\frac{4+4}{2} \right)^{-1} = \frac{2}{8} = 0.25$$

Step 3: Determine outlier score using average relative density (ard).

$$OS(x) = ard(x, k) = \frac{\text{density}(x, k)}{\sum_{y \in N(x, k)} \frac{\text{density}(y, k)}{|N(x, k)|}}$$

RELATIVE DENSITY BASED OUTLIERS

$$OS(A) = ard(A, 2) = \frac{0.4}{\frac{0.67+0.4}{2}} = 0.747$$

$$OS(B) = ard(B, 2) = \frac{0.67}{\frac{0.4+0.4}{2}} = 1.675$$

$$OS(C) = ard(C, 2) = \frac{0.4}{\frac{0.67+0.28}{2}} = 0.842$$

$$OS(D) = ard(D, 2) = \frac{0.28}{\frac{0.4+0.25}{2}} = 0.861$$

$$OS(E) = ard(E, 2) = \frac{0.25}{\frac{0.4+0.28}{2}} = 0.735$$

When using average relative density, lowest outlier score is to be considered. As outlier score of E is lowest, E is determined as outlier in this case.

DENSITY BASED OUTLIER DETECTION

Inverse Density Approach

Assume that outlier score of an object is inverse of density around an object. Using inverse density definition, determine an outlier for the objects specified in the following distance matrix. Consider 3 nearest neighbors for density determination.

	A	B	C	D	E
A	0	1	4	5	7
B	1	0	2	6	8
C	4	2	0	3	4
D	5	6	3	0	4
E	7	8	4	4	0

Step 1:

$k = 3$; For all objects determine k nearest neighbor $N(x, k)$.

$$N(x, k) = \text{neighbors}$$

$$N(A, 3) = \{B, C, D\}$$

$$N(B, 3) = \{A, C, D\}$$

$$N(C, 3) = \{B, D, E\}$$

$$N(D, 3) = \{C, E, A\}$$

$$N(E, 3) = \{C, D, A\}$$

DENSITY BASED OUTLIER DETECTION

Inverse Density Approach

Step 2: For all objects, determine density $\text{density}(x, k)$ using k-NN.

$$\text{density}(x, k) = \left[\frac{\sum_{y \in N(x, k)} \text{dist}(x, y)}{|N(x, k)|} \right]^{-1}$$

$$\text{density}(A, 3) = \left(\frac{1 + 4 + 5}{3} \right)^{-1} = \frac{3}{10} = 0.3$$

$$\text{density}(B, 3) = \left(\frac{1 + 2 + 6}{3} \right)^{-1} = \frac{3}{9} = 0.33$$

DENSITY BASED OUTLIER DETECTION

Inverse Density Approach

$$\text{density}(C, 3) = \left(\frac{2 + 3 + 4}{3} \right)^{-1} = \frac{3}{9} = 0.33$$

$$\text{density}(D, 3) = \left(\frac{3 + 4 + 5}{3} \right)^{-1} = \frac{3}{12} = 0.25$$

$$\text{density}(E, 3) = \left(\frac{4 + 4 + 7}{3} \right)^{-1} = \frac{3}{15} = 0.2$$

DENSITY BASED OUTLIER DETECTION

Inverse Density Approach

Step 3: For all objects, determine outlier score using inverse density.

$$OS(x) = \frac{1}{\text{density}(x, k)}$$

$$OS(A) = \frac{1}{\text{density}(A, 3)} = \frac{1}{0.3} = 3.33$$

$$OS(B) = \frac{1}{\text{density}(B, 3)} = \frac{1}{0.33} = 3.03$$

$$OS(C) = \frac{1}{\text{density}(C, 3)} = \frac{1}{0.33} = 3.03$$

DENSITY BASED OUTLIER DETECTION

Inverse Density Approach

$$OS(D) = \frac{1}{\text{density}(D, 3)} = \frac{1}{0.25} = 4$$

$$OS(E) = \frac{1}{\text{density}(E, 3)} = \frac{1}{0.3} = 5$$

As outlier score of E is highest, E is determined as outlier in this case.

3. CLUSTERING-BASED METHODS

- Clustering-based methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.
- Clustering is an expensive data mining operation.

-
- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T1)
 - Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 (T4)

THANK YOU



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE MODULE # 10 : STORYTELLING WITH DATA

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

Agenda



- Structuring the story
- Narrative Structure
- Story Crafting Strategies
- Step by Step Storytelling



Art of Storytelling

Magic of Story

- ❑ A good story
 - ✓ At start, grabs attention, takes on journey, evokes emotion
 - ✓ In middle, one gets indulged into it, looking for finish
 - ✓ At end, gets transferred to long term memory

Structuring the story

The constructs of a story

- Clear beginning (setup)
- Middle (conflict)
- End (resolution)

Structuring the story (cont...)

The constructs of a story - Beginning

- Introduce the plot
- Build context for audience
- Describe in form of Imbalance – balance – solution
- Communicate for audience, not for yourself

Structuring the story (cont...)

The constructs of a story - Middle

- Bulk of the part of story – full of twists and turns
- Make information more specific for audience
- Introduce problem and ways to resolve it
- Convince audience of need of action

Structuring the story (cont...)

The constructs of a story - End

- Climax of the story
- End with a call to action
- Recap the problem and resulting need for action

Story telling

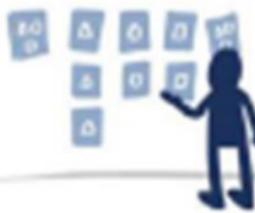
STEP 1:
UNDERSTAND
the CONTEXT

WHO is your audience?
WHAT do you need them to do?
HOW will data help make your point?

ARTICULATE
your BIG IDEA



CREATE a
STORYBOARD



- brainstorm
- edit
- get feedback

Storytelling Context

WHERE to BEGIN?

① **WHO**

is your
AUDIENCE?
BE SPECIFIC!



What is their
relationship to **YOU**?
What motivates them?
What keeps them
up at night?

② **WHAT**

DO YOU NEED
them to do?
BE EXPLICIT!



Don't assume
they will
connect the dots!

③ **HOW**

WILL DATA HELP
make your point?
BE DISCERNING!



What data
will act as
evidence for
the case?

Narrative Structure

Narration of a story

- Important ingredient of successful storytelling
- Words – spoken , written or both
- Important to grab audience attention
- Has to have a logical order that makes sense

Narrative Structure (cont...)

Narrative flow

- ❑ How to convey the message through a proper order
 - ❑ Busy audience ?
 - ❑ New audience?
 - ❑ High level overview or more detailed oriented?
 - ❑ Credibility ?
 - ❑ Collaborative process?

Narrative Structure (cont...)

Narrative flow

- Order Chronologically
 - ✓ Natural way of ordering
 - ✓ Tell the story in order in which events happened
 - ✓ Work well if need to establish credibility
- Lead with the ending
 - ✓ Start with call to action
 - ✓ Unwind the story in backward manner
 - ✓ Work well if trust is already established

Narrative Structure (cont...)

Narration Types

- Spoken
 - Live presentation
 - Advantages
 - ✓ Words/Visuals resonate your voice
 - ✓ Audience can get doubts clarified immediately
 - Challenge
 - ✓ Unpredictable audience
 - ✓ Distraction by slides in presentation if too much is placed on it



Narrative Structure (cont...)

Narration Types

- ❑ Written
 - ❑ Report, Sliduements, Presentation decks
 - ❑ Advantages
 - ✓ Written evidences of story
 - ✓ Audience has control where to focus
 - ❑ Challenge
 - ✓ No live doubt clearing sessions
 - ✓ Need to provide very careful attention to the content
 - ✓ Content should be self explanatory



Story Crafting Strategies

Strategies

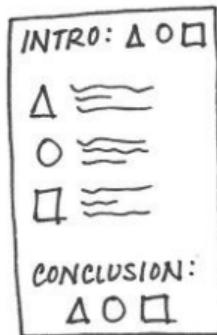
- Power of repetition
- Horizontal and Vertical Logic
- Reverse storyboarding
- Fresh perspective



Story Crafting Strategies (cont...)

Strategies - Repetition

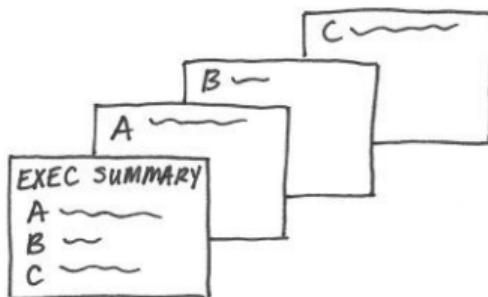
- If repeated multiple times, gets into long term memory
- Repetition — thrice
 - ✓ Once – Introduction of story - executive summary
 - ✓ Twice – actual story content – main content
 - ✓ Thrice – Conclusion – main topics / actions



Story Crafting Strategies (cont...)

Strategies – Horizontal Logic

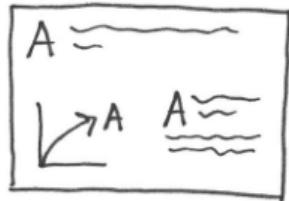
- Read slide titles and together they represent story
- Slide titles needs to be actions
- Helpful to test whether the story is clearly coming in deck



Story Crafting Strategies (cont...)

Strategies – Vertical Logic

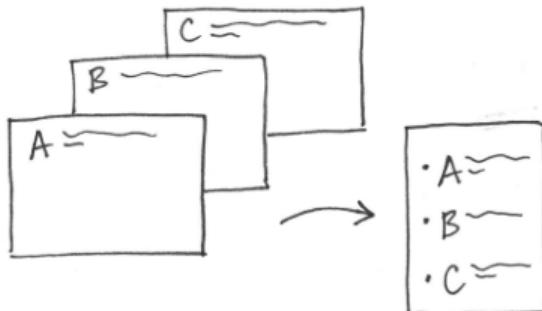
- All information is present on same slide
- Content reinforces title and vice-a-versa
- Decision to eliminate is more important than to keep it



Story Crafting Strategies (cont...)

Strategies – Reverse Storyboarding

- Take final communication, flip through it and write down main point from each page
- Nice way to check horizontal logic
- Resulting list looks like a storyboard or outline
- Help to rearrange, add , delete the pieces around



Story Crafting Strategies (cont...)

Strategies – Fresh Perspective

- Look at communication through audience lens
- Getting feedback from third person is more important
- Ask them about – what attracted them, what is important, what are the questions in their mind
- Help to find out where to concentrate during the iterations



Storytelling Strategies

THREE
MINUTE
STORY



Storytelling Strategies

BIG IDEA*

*from Nancy Duarte
(Resonate)



A SINGLE SENTENCE that...

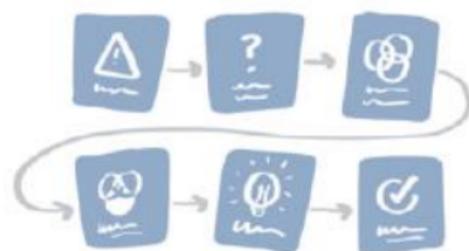
the "So what?" boiled down further

- ① articulates your point of view
- ② conveys what is at stake
- ③ is one complete sentence

Storytelling Strategies

STORY BOARDing

UPFRONT PLANNING to CREATE STRUCTURE



STICKY NOTES help to...

avert attachment
to work done
on computer

force concise
articulation

easily
rearrange
the flow

1
BRAINSTORM

2
EDIT

3
GET FEEDBACK

Summary-Lessons of Storytelling



Storytelling

- ✓ Structuring the Story
 - Beginning , middle and end
 - Plot, twist and conclusion
- ✓ Narrative Structure
 - Ordering – chronological , lead with ending
 - Types – Spoken , Written
- ✓ Story Crafting Strategies
 - Repetition
 - Horizontal and Vertical logic
 - Reverse Storyboarding
 - Fresh Perspective

CASE STUDY 1

Scenario

Start-up Company – Product Price Determination

- Start-up developing a new consumer product
- Need to price the product
- Price variations of other similar products available

Storytelling Lessons : Beginning

Understand the context

- Get Robust understanding of the context
- Who** : VP of product , primary decision maker in determining product price
- What**: Understand how competitors' pricing has changed over time and recommend a price range
- How**: With help of retail price over time for competitor products

Storytelling Lessons : Beginning

Tell a Story

“Competitive Landscape – Pricing “

Storytelling Lessons : Beginning

Tell a Story

In the next **5 minutes...**

OUR GOAL:

- 1** Understand **how prices have changed over time** in the competitive landscape.
- 2** Use this knowledge to **inform the pricing of our product**.

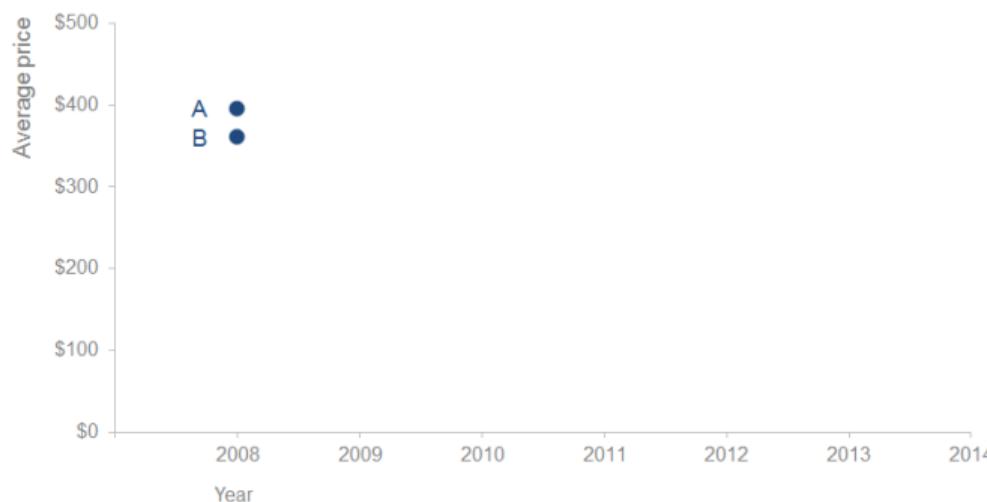
We will end with a **specific recommendation**.

Storytelling Lessons : Middle (cont...)

Tell a Story

Products A and B were launched in 2008 at price points of \$360+

Retail price over time

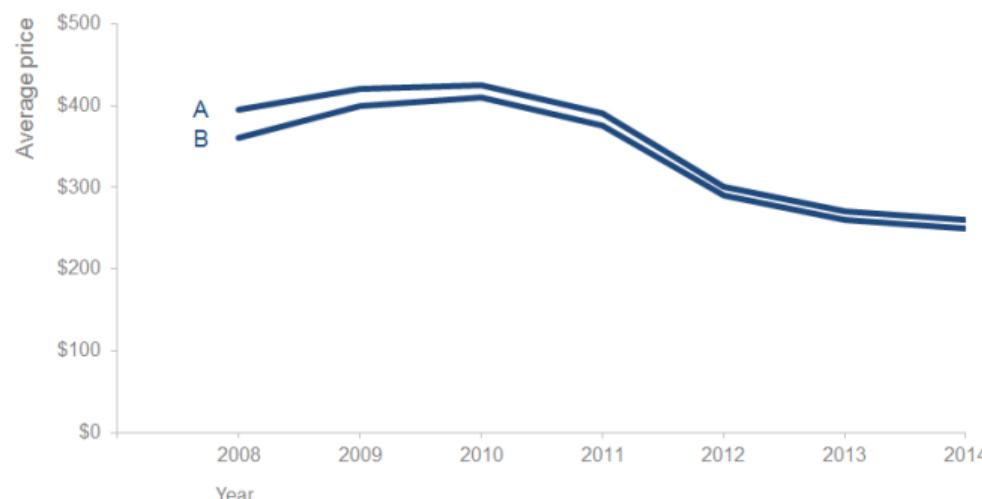


Storytelling Lessons : Middle (cont...)

Tell a Story

They have been priced similarly over time, with B consistently slightly lower than A

Retail price over time

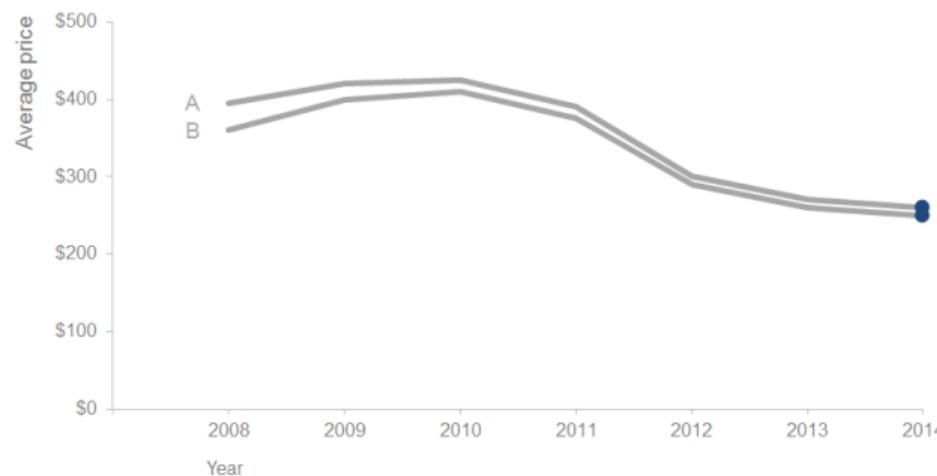


Storytelling Lessons : Middle (cont...)

Tell a Story

In 2014, Products A and B were priced at \$260 and \$250, respectively

Retail price over time

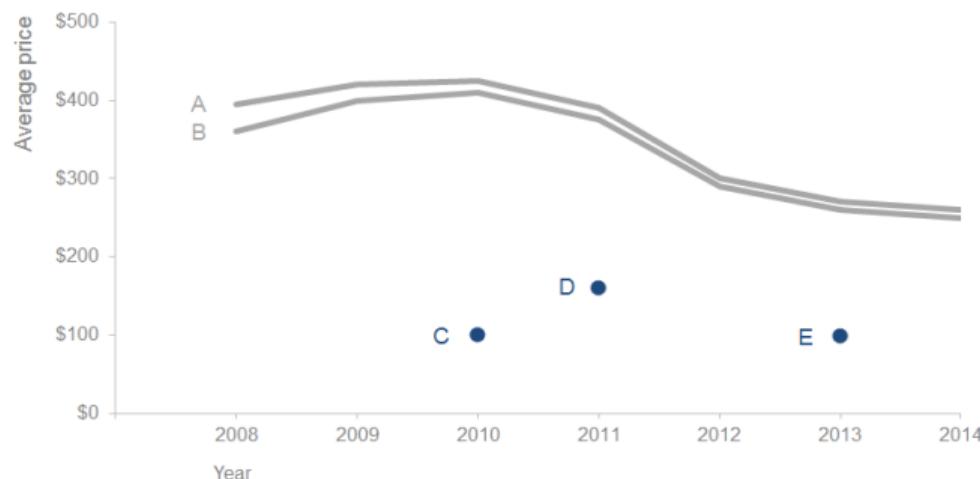


Storytelling Lessons : Middle (cont...)

Tell a Story

Products C, D, and E were each introduced later
at **much lower price points...**

Retail price over time

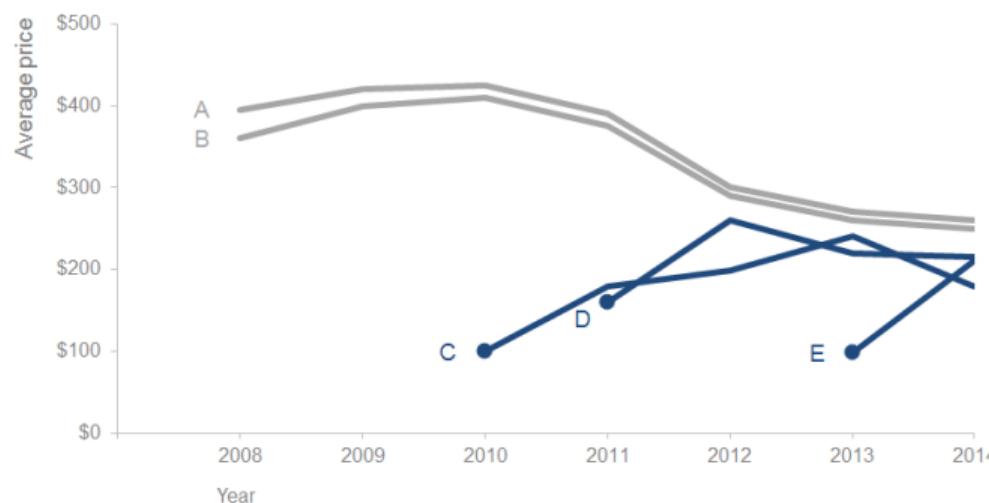


Storytelling Lessons : Middle (cont...)

Tell a Story

...but all have **increased in price** since their respective launches

Retail price over time

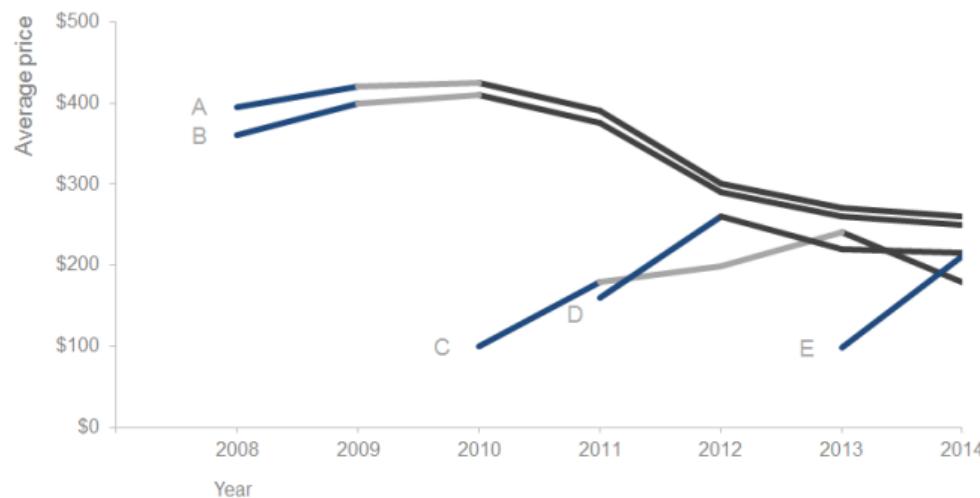


Storytelling Lessons : Middle (cont...)

Tell a Story

In fact, with the launch of a new product in this space, we tend to see an **initial price increase**, followed by a **decrease** over time

Retail price over time

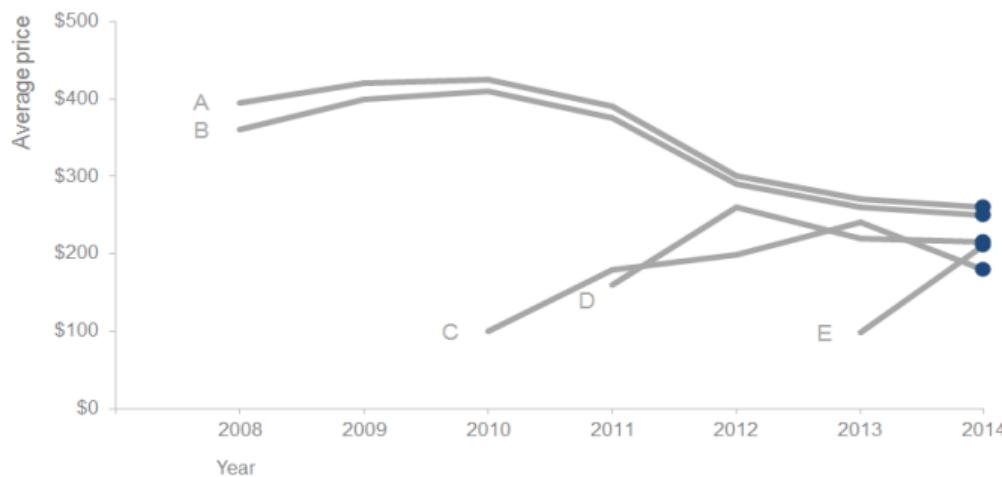


Storytelling Lessons : Middle (cont...)

Tell a Story

As of 2014, retail prices have converged, with an **average retail price of \$223**, ranging from a low of \$180 (C) to a high of \$260 (A)

Retail price over time

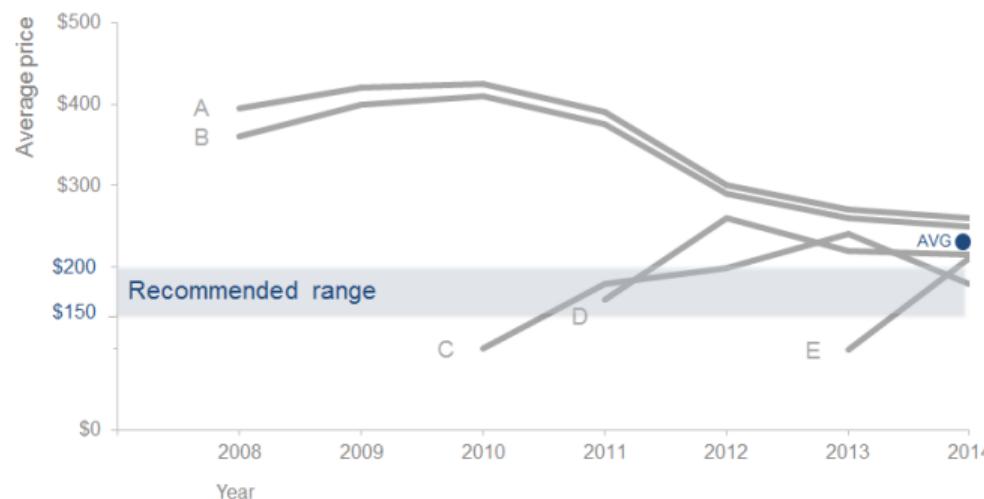


Storytelling Lessons : Ending

Tell a Story

To be competitive, we recommend introducing our product *below the \$223 average price point in the \$150-\$200 range*

Retail price over time



CASE STUDY 2

Scenario 2

Imagine you are a fourth grade science teacher. You just wrapped up an experimental pilot summer learning program on science that was aimed at giving kids exposure to the unpopular subject. You surveyed the children at the onset and end of the program to understand whether and how perceptions toward science changed. You believe the data shows a great success story. You would like to continue to offer the summer learning program on science going forward.

Identify the Context.(WHO,WHAT and HOW)

WHO

We want to communicate to is **THE BUDGET COMMITTEE**, which controls the funding we need, to continue the program.

WHAT

Demonstrate the success of the program and ask for a specific funding amount to continue to offer it

HOW

Use the data collected via survey at the onset and end of the program to illustrate the increase in positive perceptions of science before and after the pilot summer learning program.

Storytelling Context - Beginning

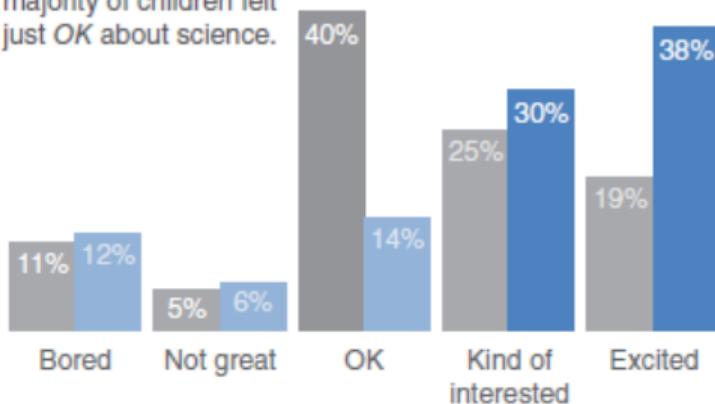
- **Who:** The budget committee that can approve funding for continuation of the summer learning program.
- **What:** The summer learning program on science was a success; please approve budget of \$X to continue.
- **How:** Illustrate success with data collected through the survey conducted before and after the pilot program.

Storytelling - Middle

Pilot program was a success

How do you feel about science?

BEFORE program, the majority of children felt just *OK* about science.



AFTER program,
more children
were *Kind of
interested &
Excited* about
science.

Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Storytelling - 3 -minute Story – Ending

3-minute story: A group of us in the science department were brainstorming about how to resolve an ongoing issue we have with incoming fourth-graders. It seems that when kids get to their first science class, they come in with this attitude that it's going to be difficult and they aren't going to like it. It takes a good amount of time at the beginning of the school year to get beyond that. So we thought, what if we try to give kids exposure to science sooner? Can we influence their perception? We piloted a learning program last summer aimed at doing just that. We invited elementary school students and ended up with a large group of second- and third-graders. Our goal was to give them earlier exposure to science in hopes of forming positive perception. To test whether we were successful, we surveyed the students before and after the program. We found that, going into the program, the biggest segment of students, 40%, felt just "OK" about science, whereas after the program, most of these shifted into positive perceptions, with nearly 70% of total students expressing some level of interest toward science. We feel that this demonstrates the success of the program and that we should not only continue to offer it, but also to expand our reach with it going forward.

Storytelling - Big Idea – Ending

The pilot summer learning program was successful at improving students' perceptions of science and, because of this success, we recommend continuing to offer it going forward; please approve our budget for this program.

References



Knafllic, Cole. Storytelling With Data: A Data Visualization Guide for Business Professionals, Wiley, © 2015

- Chapter 1, Chapter 7 and Chapter 8

THANK YOU





BITS Pilani
Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE SESSION # 13 : ETHICS FOR DATA SCIENCE

IDS Course Team
BITS Pilani

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

TABLE OF CONTENTS

1 ETHICS FOR DATA SCIENCE

2 FIVE C'S IN DATA

BIAS AND FAIRNESS IN DATA

FAIRNESS

Fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics.

BIAS

Bias is inclination or prejudice for or against one person or group, especially in a way considered to be unfair.

SOME EXAMPLES OF BIAS IN DATA

GENDER BIAS Biases present in the word embedding (i.e. which words are closer to she than to he, etc.) trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. Most of these biases are implicit and hard to recognize.

Extreme *she* occupations

- | | | |
|-----------------|-----------------------|------------------------|
| 1. homemaker | 2. nurse | 3. receptionist |
| 4. librarian | 5. socialite | 6. hairdresser |
| 7. nanny | 8. bookkeeper | 9. stylist |
| 10. housekeeper | 11. interior designer | 12. guidance counselor |

Extreme *he* occupations

- | | | |
|----------------|-------------------|----------------|
| 1. maestro | 2. skipper | 3. protege |
| 4. philosopher | 5. captain | 6. architect |
| 7. financier | 8. warrior | 9. broadcaster |
| 10. magician | 11. fighter pilot | 12. boss |

SOME EXAMPLES OF BIAS IN DATA CONTD...

AGE BIAS Training a facial recognition system and then using photos from Instagram to test it may underperform for real world data as majority of Instagram users are between the ages of 18 and 35.

EVALUATION BIAS Data used to test a model doesn't accurately represent the real world.

HUMAN BIAS Having prior knowledge of the problem you're trying to solve can help you to select relevant features during modeling, introduces human bias but can often speed up or improve the modeling process.

TYPES OF BIAS

1 Data

- ① Representation bias
- ② Aggregation bias
- ③ Measurement bias
- ④ Linking bias
- ⑤ Omitted variable bias

2 Human review

- ① Historical bias
- ② Behavioural bias
- ③ Population bias
- ④ Temporal bias
- ⑤ Social bias
- ⑥ Self-Selection Bias

3 AI / ML

- ① Algorithmic bias
- ② User Interaction Bias
- ③ Popularity bias
- ④ Emergent bias
- ⑤ Evaluation bias

TYPES OF BIAS IN DATA

REPRESENTATION BIAS arises from how we sample from a population during data collection process. Eg: Lack of geographical diversity in datasets like ImageNet results in demonstrable bias towards Western cultures.

AGGREGATION BIAS (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. Eg: HbA1c levels, that are widely used to diagnose and monitor diabetes, differ in complex ways across genders and ethnicities. A model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population.

MEASUREMENT OR REPORTING BIAS arises from how we choose, utilize, and measure particular features. Eg: Ethnicity bias observed in recidivism risk prediction tool COMPAS.

TYPES OF BIAS IN DATA

SAMPLING BIAS is similar to representation bias, and it arises due to nonrandom sampling of subgroups.

LINKING BIAS arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users. Eg: Social networks can be biased toward low-degree nodes when only considering the links in the network and not considering the content and behavior of users in the network.

OMITTED VARIABLE BIAS occurs when one or more important variables are left out of the model. Eg: Churn prediction tool is not considering, appearance of a new strong competitor. (omitted variable)

TYPES OF BIAS IN HUMAN REVIEW

HISTORICAL BIAS is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection. Eg: Bias can be found in a 2018 image search result in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were women, leading to biased search results towards male CEOs.

POPULATION BIAS arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population. Eg: Women more likely to use Pinterest, Facebook, Instagram, while men being more active in Reddit or Twitter.

SELF-SELECTION BIAS is a subtype of the selection or sampling bias in which subjects of the research select themselves. Eg: Opinion polls, where the most enthusiastic users are more likely to complete the poll.

TYPES OF BIAS IN HUMAN REVIEW

SOCIAL BIAS happens when others' actions affect our judgment. Eg: We want to rate an item with a low score, but when influenced by other high ratings, we change our scoring.

BEHAVIORAL BIAS arises from different user behavior across platforms, contexts, or different datasets. Eg: differences in emoji representations among platforms can result in different reactions and behavior from people and sometimes even leading to communication errors.

TEMPORAL BIAS arises from differences in populations and behaviors over time. Eg: In Twitter where people start using a hashtag at some point to capture attention, then continue the discussion without using the hashtag.

CONTENT PRODUCTION BIAS arises from structural, lexical, semantic, and syntactic differences in the contents generated by users. Eg: differences in use of language across different gender, age groups, across and within countries.

HOW TO TEST FOR BIAS IN DATA?

- Association tests

Eg: Unwarranted associations (UA) framework by Tramer et.al.

<https://arxiv.org/pdf/1510.02377.pdf>

- Perturbation tests

Eg: FairML

[urlhttps://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html](https://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html)

HOW TO FIX THE BIAS?

Three different strategies to reduce and even eliminate biases:

PRE-PROCESSING eliminating any sources of unfairness in the data before the algorithm is formulated.

IN-PROCESSING making fairness adjustments as part of the process by which algorithm is constructed.

POST-PROCESSING after the algorithm is applied, its performance is adjusted to make it fairer.

TABLE OF CONTENTS

1 ETHICS FOR DATA SCIENCE

2 FIVE C'S IN DATA

FIVE C's IN DATA

- 5 guidelines to establish trust while building data products:
 - ① Consent
 - ② Clarity
 - ③ Consistency
 - ④ Control and Transparency
 - ⑤ Consequences and harm
- Europe's General Data Protection Regulation (GDPR) rules
- Indian Information Technology (IT) Act 2000

FIVE C's IN DATA

- Consent
 - ▶ Agreement about what data is being collected and how data will be used.
 - ▶ In data science, the user either accepts the terms or they don't get access. It is binary and it is non-negotiable.

- Clarity
 - ▶ Users must have clarity about
 - ★ what data they are providing
 - ★ what is going to be done with the data
 - ★ any consequences of how their data is used
 - ▶ Users frequently don't understand how that data could be used. Eg: Tweets are public, and can be collected and used for research or may be sold.

FIVE C's IN DATA

- Consistency and Trust
 - ▶ Trust requires consistency over time.
 - ▶ An organization can expose data intentionally or unintentionally.
 - ▶ Failing to safeguard customer data breaks trust. Safeguarding data means consistency over time.
- Control and Transparency
 - ▶ Users have no effective control over how their data is used. They are given all-or-nothing choices.
 - ▶ Data privacy rights shifting to give users greater control of their data.
- Consequences and harm
 - ▶ Data that is being collected could lead to unforeseen consequences.
 - ▶ Children's Online Privacy Protection Act (COPPA) protects children and their data.
 - ▶ Genetic Information Non-discrimination Act (GINA) 2008 in response to genetic testing

5 PRINCIPLES OF DATA ETHICS

- Ownership

- ▶ Every individual has ownership over their personal information.
- ▶ Obtain consent are through signed written agreements, digital privacy policies that ask users to agree to a company's terms and conditions, and pop-ups with checkboxes that permit websites to track users' online behavior with cookies.
- ▶ Never assume a customer is OK with you collecting their data; always ask for permission to avoid ethical issues.

- Transparency

- ▶ Users have a right to know how you plan to collect, store, and use it.
- ▶ When gathering data, exercise transparency.

5 PRINCIPLES OF DATA ETHICS

- Privacy

- ▶ A customer may consent to collect, store, and analyze their personally identifiable information (PII), that doesn't mean they want it publicly available.
- ▶ Data security methods that help protect privacy include dual-authentication password protection and file encryption.
- ▶ De-identifying a dataset when all pieces of PII are removed, leaving only anonymous data.

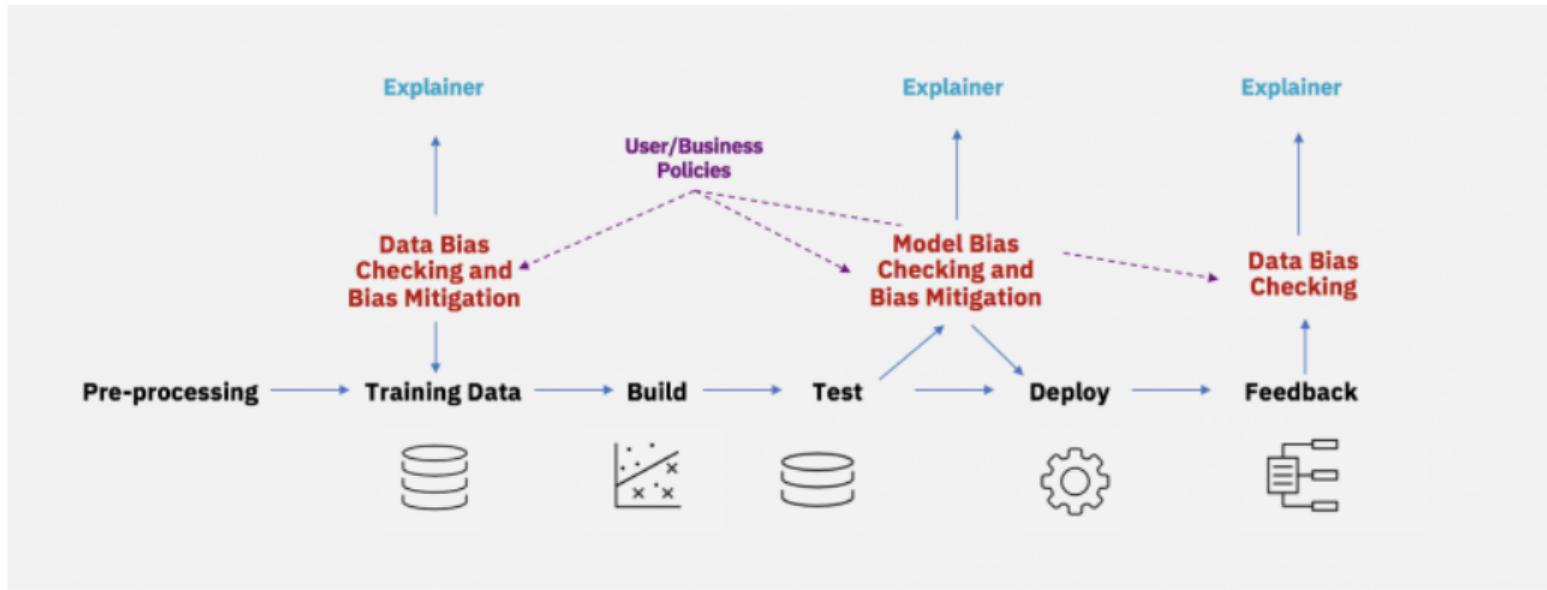
- Intention

- ▶ Strive to collect the minimum viable amount of data.

- Outcomes

- ▶ Even when intentions are good, the outcome of data analysis can cause inadvertent harm to individuals or groups of people. This is called a disparate impact.

IBM RESEARCH AI FAIRNESS



<https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>

-
- Ethics and Data Science by DJ Patil, Hilary Mason, Mike Loukidesr (R2)
 - Bias in Data: <https://arxiv.org/pdf/1908.09635.pdf>
 - Ethics in Data <https://online.hbs.edu/blog/post/data-ethics>

THANK YOU