Data Mining 2025

Introduction to Data Mining

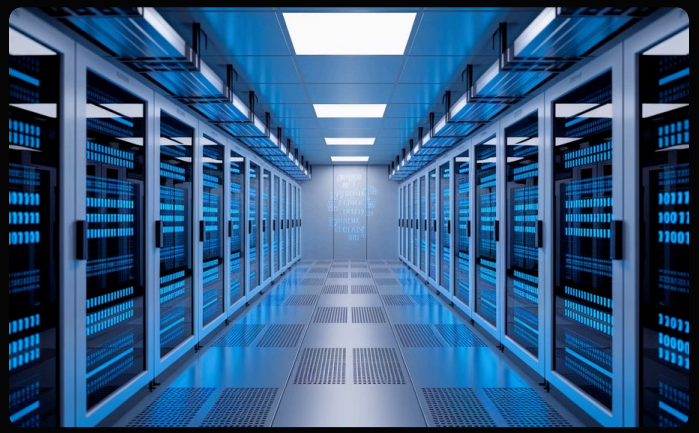
Dept. of Computer Science and Information Engineering

National Cheng Kung University

Kun-Ta Chuang

ktchuang@mail.ncku.edu.tw

What is Data?

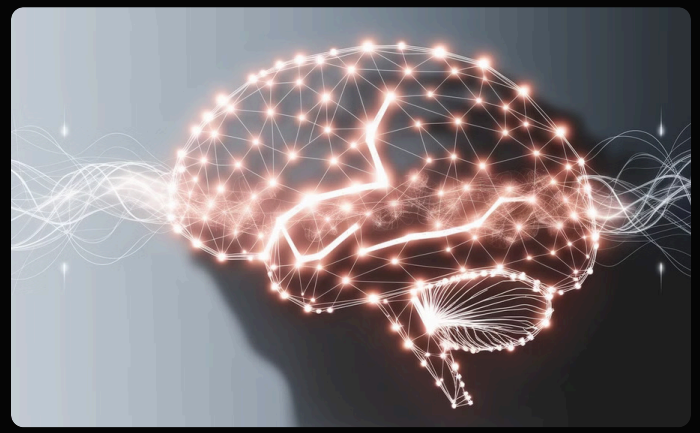


Data: The Raw Material

Raw facts, figures, symbols, and observations in various forms like numbers, text, images, audio, and video - representing the fundamental building blocks of information.

Information: Processed Data

When data is processed and organized into meaningful structures, it becomes information - the first step towards understanding patterns and relationships.

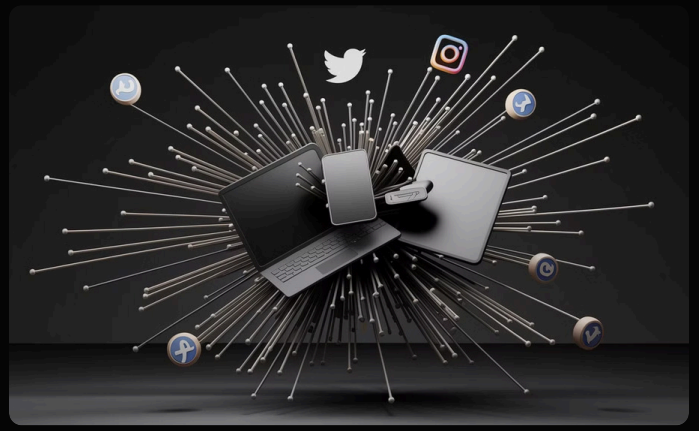


Knowledge: Actionable Insights

The final transformation occurs when information patterns reveal insights and understanding, creating actionable knowledge that drives decision-making.

Understanding data's intrinsic nature is essential to transform it into **actionable insights**.

Data in the Modern World

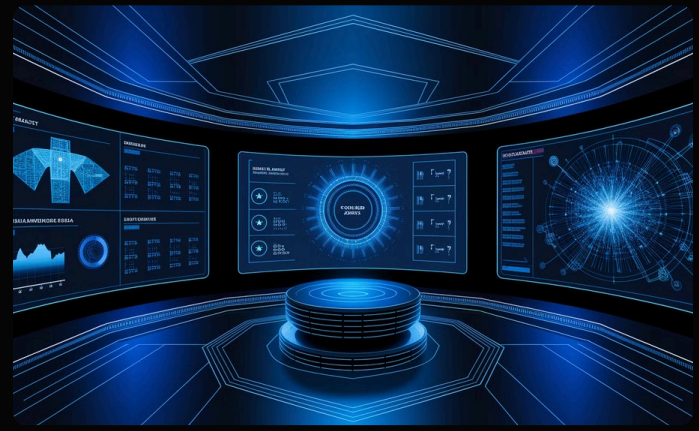


The Data Explosion

Modern digital landscape generates massive volumes of data daily through devices, sensors, web activities, and social media interactions, reaching from terabytes to petabytes.

The 4 V's of Modern Data

- **Volume**: Enormous data quantities
- **Velocity**: Real-time data streams
- **Variety**: Multiple data formats
- **Veracity**: Quality and accuracy



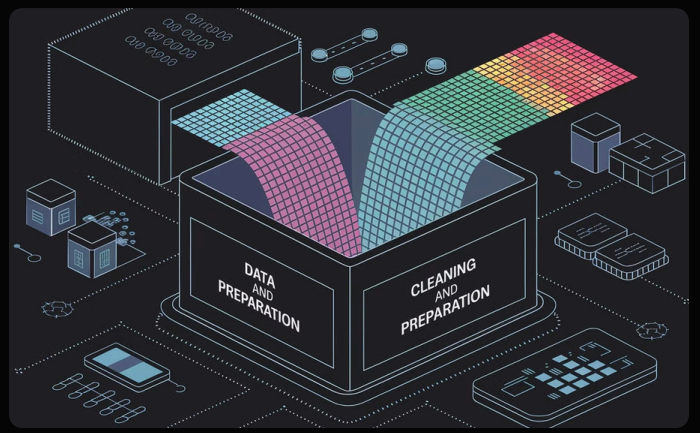
Societal & Business Impact

Data drives innovation across sectors, from scientific breakthroughs to business intelligence and healthcare advances, making it the new digital oil of modern economy.

Why Understanding Data is Crucial for Data Mining

Data Mining Process

Extraction of valuable patterns and trends relies on proper understanding of how raw data transforms into actionable knowledge



Data Quality & Preprocessing

Effective cleaning, integration, and transformation require deep knowledge of data properties to ensure quality results

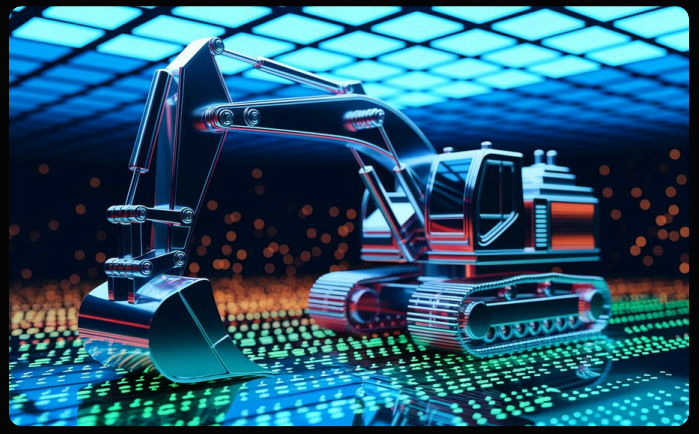
Data Mining Challenges

Handling noisy, incomplete, and heterogeneous data requires careful selection of techniques based on data type and structure

A solid grasp of what data is—and its nuances—is the foundation for designing effective, scalable data mining solutions.

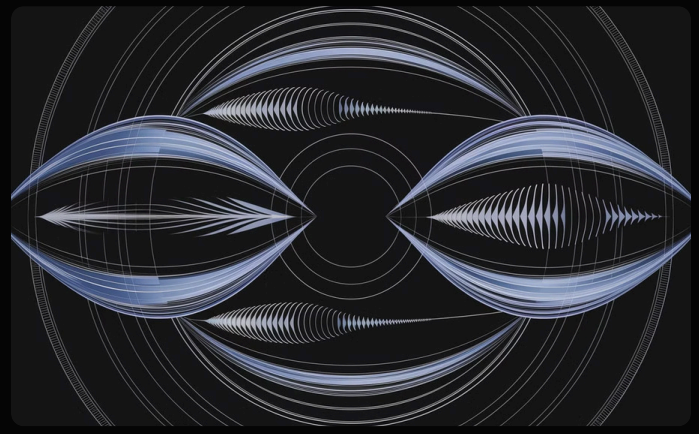
What is Data Mining?

Turning Data into Knowledge



Extracting Knowledge from Data

Data mining is the process of extracting patterns, models, or knowledge from large datasets. It's about discovering insights hidden within vast amounts of data that would be difficult to identify by humans alone.



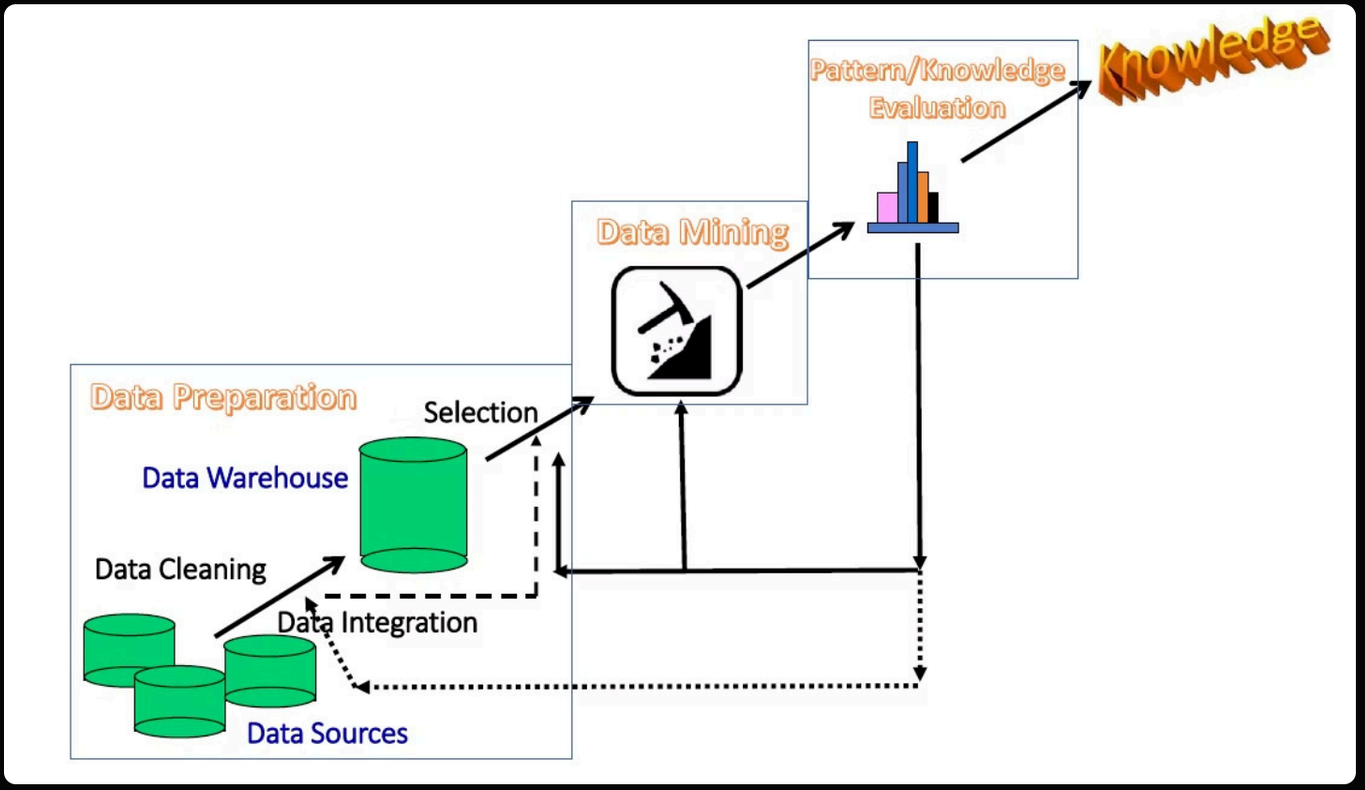
Knowledge Discovery Process

Data mining is a key step in the *knowledge discovery from data* (KDD) process, integrating techniques from databases, statistics, machine learning, and pattern recognition.

Real-World Examples

Google's Flu Trends[1] demonstrates data mining in action, analyzing billions of search queries to predict flu activity up to two weeks faster than traditional methods by identifying patterns in health-related searches.

Process of KDD



Data Preparation

Involves cleaning, integrating, transforming, and selecting data.

Data Mining

Applying intelligent methods to extract patterns and models from data.

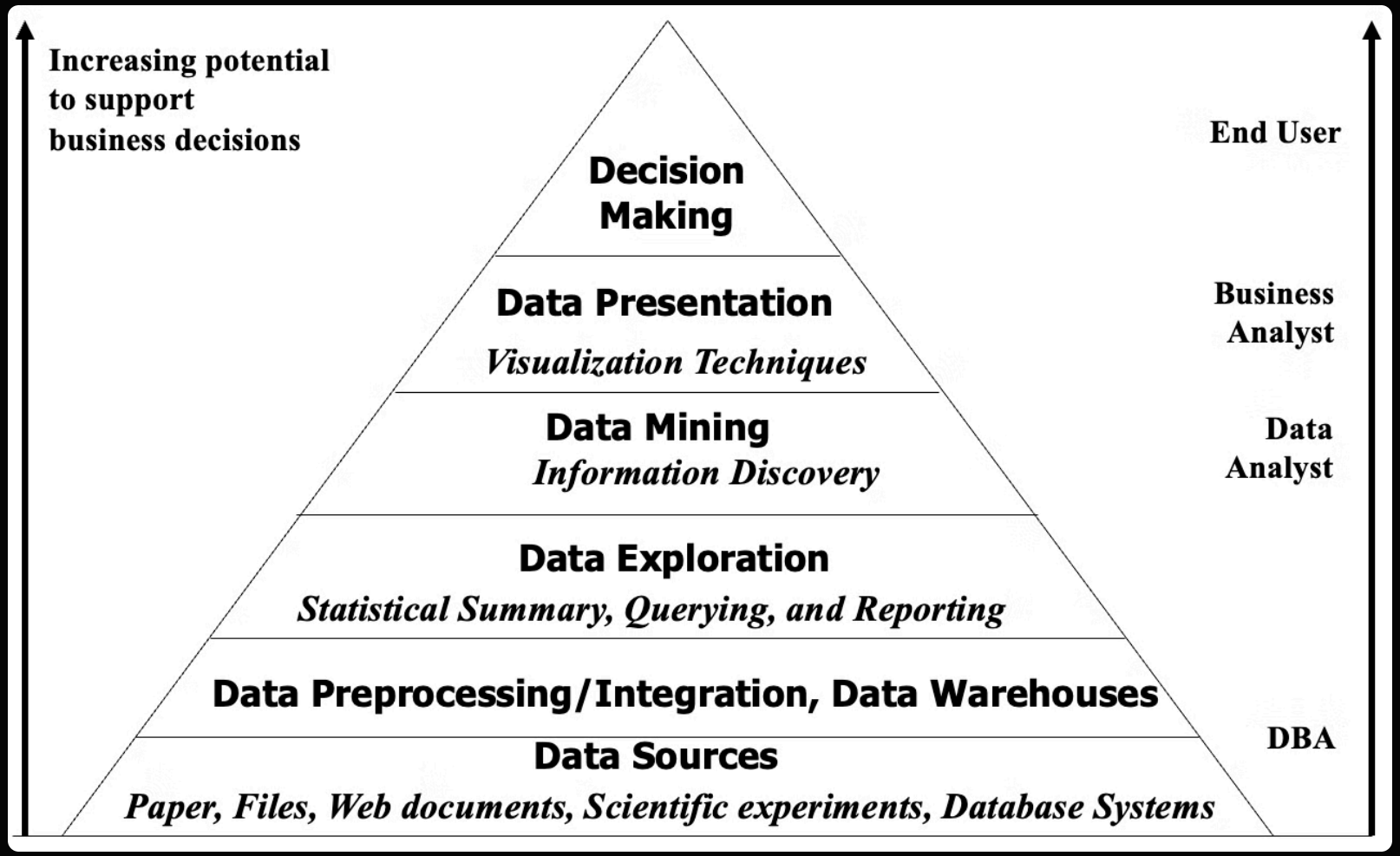
Pattern/Model Evaluation

Identifying valuable patterns based on interestingness measures.

Knowledge Presentation

Presenting mined knowledge through visualization and knowledge representation techniques.

Process of Data Mining in Business Intelligence



Data Types for Data Mining

Structured & Semistructured Data

Data with defined organization and structure:

- Structured data includes tables, data cubes, and matrices (e.g., sales database with product name, price)
- Semistructured data has flexible organization, such as transactions with multiple items and XML with nested tags
- Includes graph or network data, where nodes/links have semantic descriptions

Example: An online shopping site storing products in a relational table, plus user reviews in text form

Unstructured Data

Data without predefined schema:

- Primarily consists of text, images, audio, and video content
- Requires specialized techniques like NLP for text mining and computer vision for images/videos
- Involves complex pattern recognition methods

Example: Customer product reviews and advertisement videos

Data Types for Data Mining (Cont'd)

Data for Different Applications

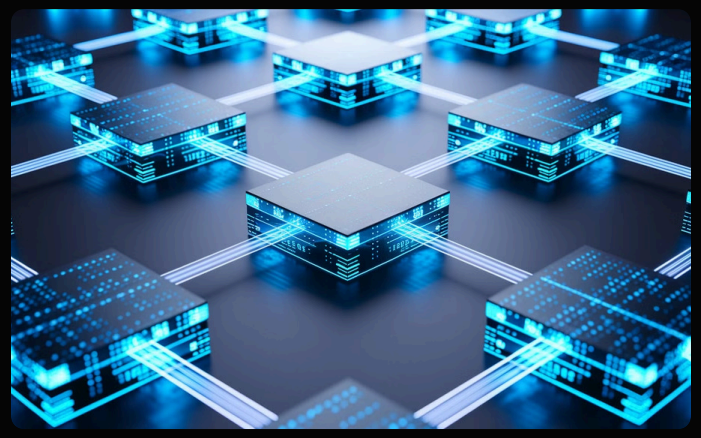
- **Sequence & Time-Series**: DNA/protein sequences vs. shopping transaction logs vs. sensor data
- **Spatial & Temporal**: Maps, geotagged data, moving object trajectories (e.g., GPS logs)
- **Graphs & Networks**: Social networks (Facebook/LinkedIn), communication networks, author-keyword links
- Example: Mining social network data vs. mining road traffic sensor streams—each needs specialized methods

Stored vs. Streaming Data

- **Stored**: Large finite repositories (e.g., data warehouses)
- **Streaming**: Continuous, infinite flow (e.g., live video surveillance, remote sensing)
- Example: Analyzing monthly sales data (stored) vs. monitoring real-time video feed for anomalies (streaming)

Different data types and application requirements demand diverse, specialized data mining techniques.

Data Types for Data Mining (Cont'd)

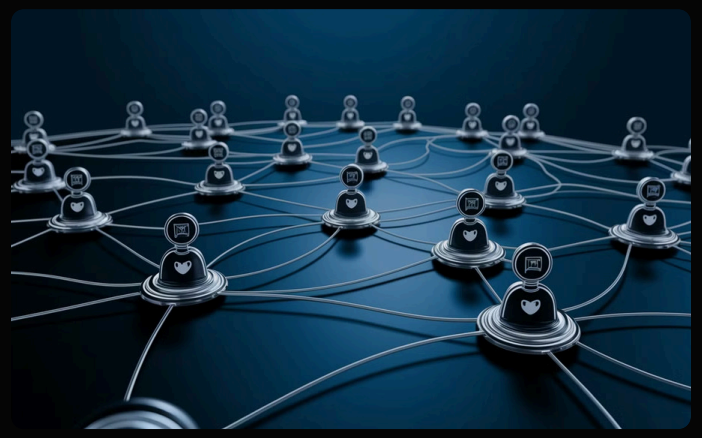


Database-Oriented Data

Includes relational databases, data warehouses, and transactional databases that form the foundation of traditional data mining.

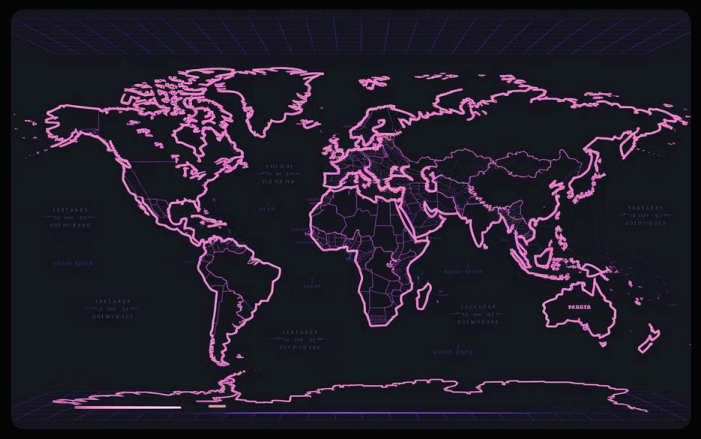
Real-Time and Sequential Data

Data streams, sensor data, time-series data, and sequential data that capture temporal patterns and changes.



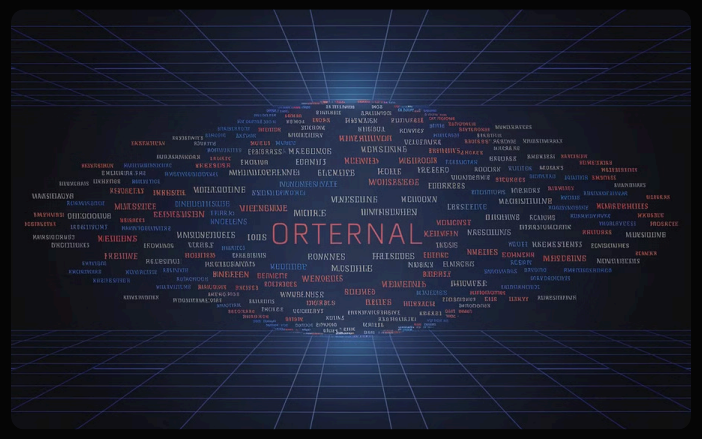
Structured and Network Data

Structure data, graphs, social networks, and multi-linked data showing complex relationships and connections.



Spatial and Multimedia

Spatial data, spatiotemporal data, and multimedia databases requiring specialized analysis techniques.



Text and Web Data

Text databases and World-Wide Web data, including heterogeneous and legacy databases requiring advanced processing.

Mining Various Kinds of Knowledge

Two Broad Categories of Data Mining

Descriptive Data Mining

Characterizes the properties of the data, revealing its inherent structure and trends.

Predictive Data Mining

Induces patterns from data to forecast future events or predict values for unseen data.

Mining Various Kinds of Knowledge (Cont'd)

Multidimensional Data Summarization

Data Summarization

Compresses large datasets into concise summaries, enabling efficient analysis and comparisons.

Multidimensional Data Cubes

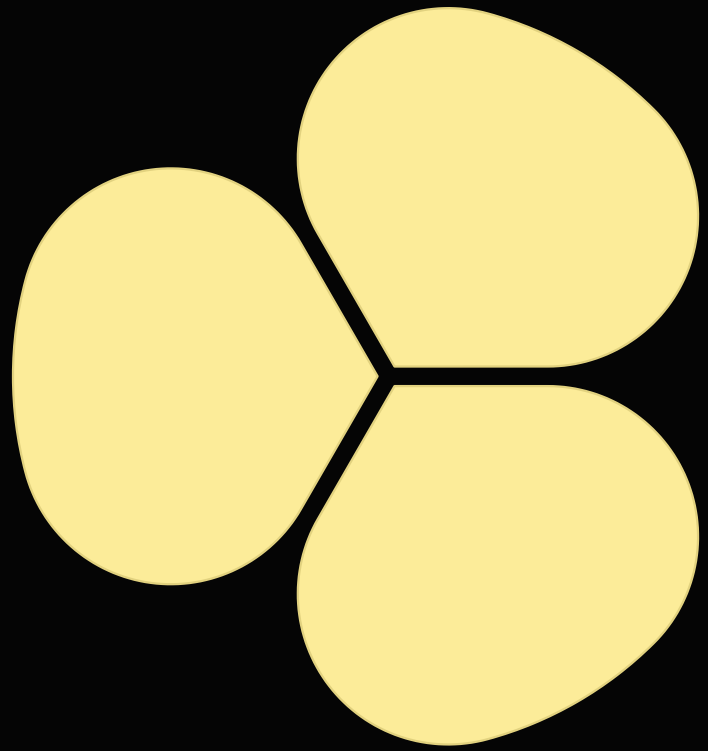
Provide a multidimensional view of data, facilitating interactive exploration and drill-down analysis.

Mining Various Kinds of Knowledge (Cont'd)

Mining Frequent Patterns, Associations, and Correlations

Frequent Itemsets

Groups of items that frequently occur together in transactional data.



Association Rules

Express relationships between itemsets, highlighting patterns of co-occurrence.

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40}K..49K\text{”}) \Rightarrow buys(X, \text{“}laptop\text{”})[sup. = 0.5\%][conf. = 60\%]$$

Correlations

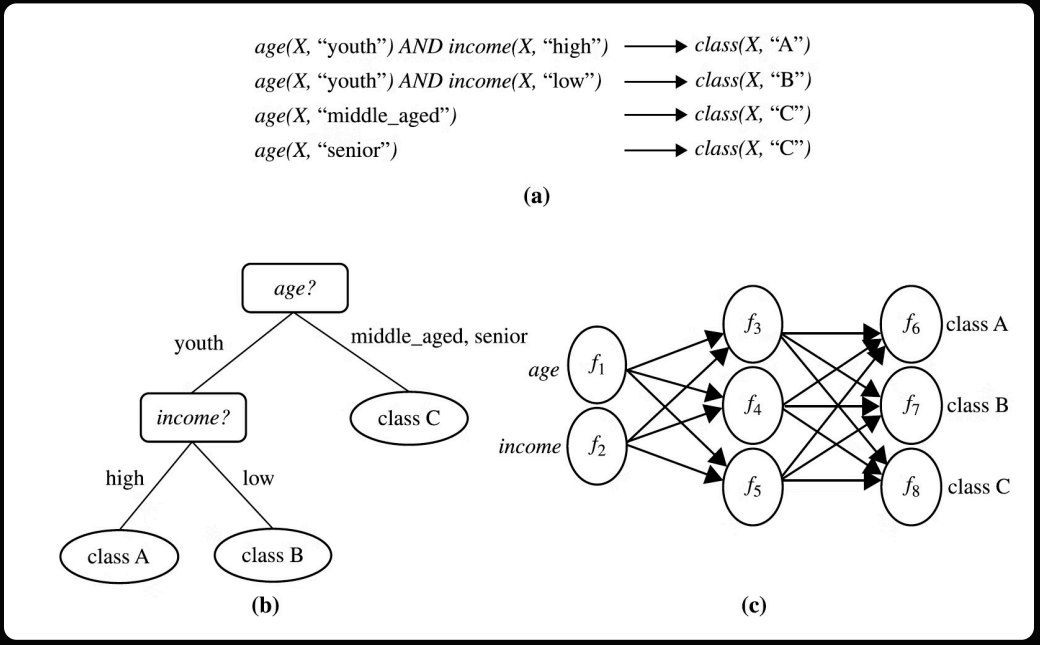
Statistical dependencies between attributes, revealing potential causal relationships.

Mining Various Kinds of Knowledge (Cont'd)

Classification and Regression for Predictive Analysis

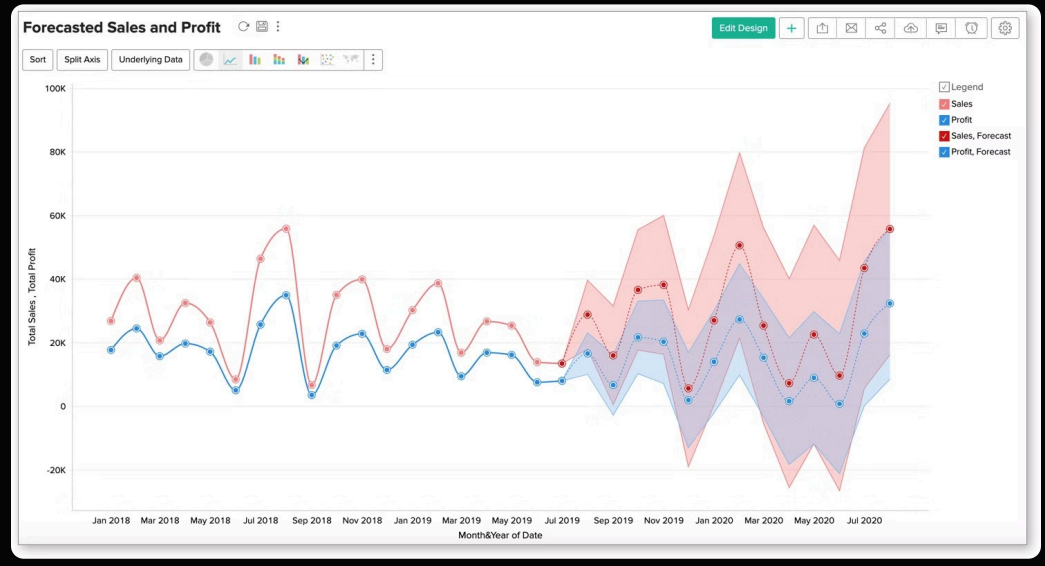
Classification

Predicts categorical labels, assigning data points to predefined classes.



Regression

Predicts continuous values, forecasting numerical outcomes for unseen data.

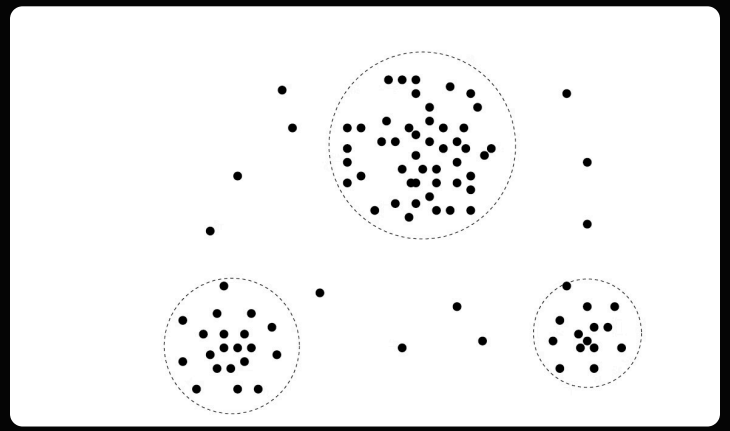


Mining Various Kinds of Knowledge (Cont'd)

Cluster Analysis: Grouping Similar Data

Clustering

Groups data points based on similarity, forming clusters of closely related objects.



Taxonomy Formation

Organizes data into a hierarchical structure of classes, revealing relationships between clusters.

Mining Various Kinds of Knowledge (Cont'd)

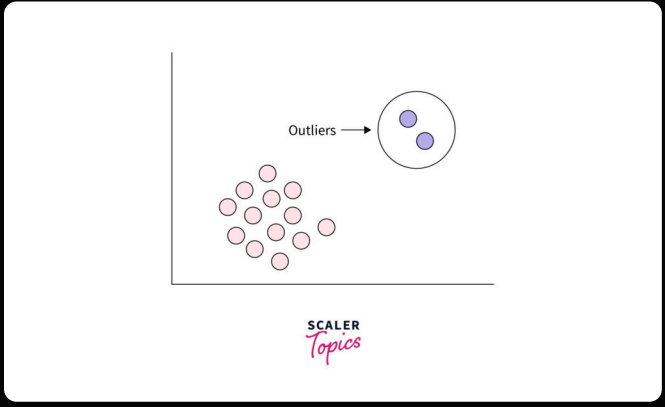
Outlier Analysis: Uncovering Unusual Patterns

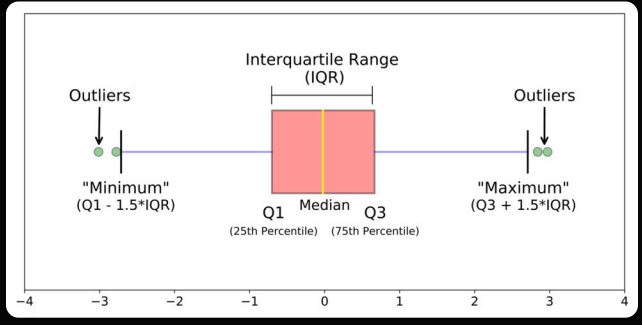
Fraud Detection

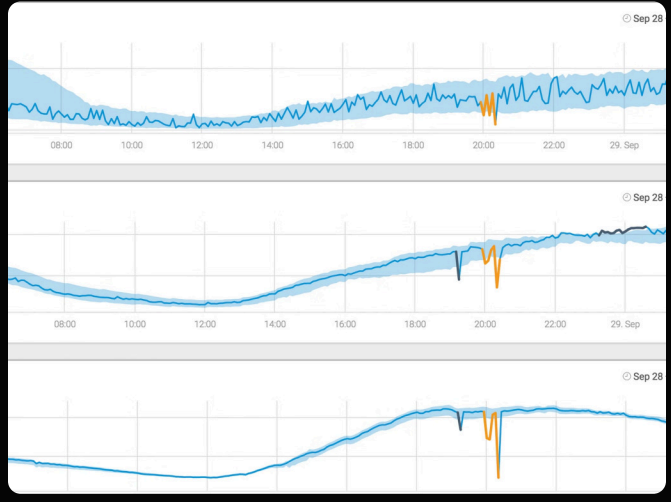
Identifies unusual transactions or behaviors that may indicate fraudulent activity.

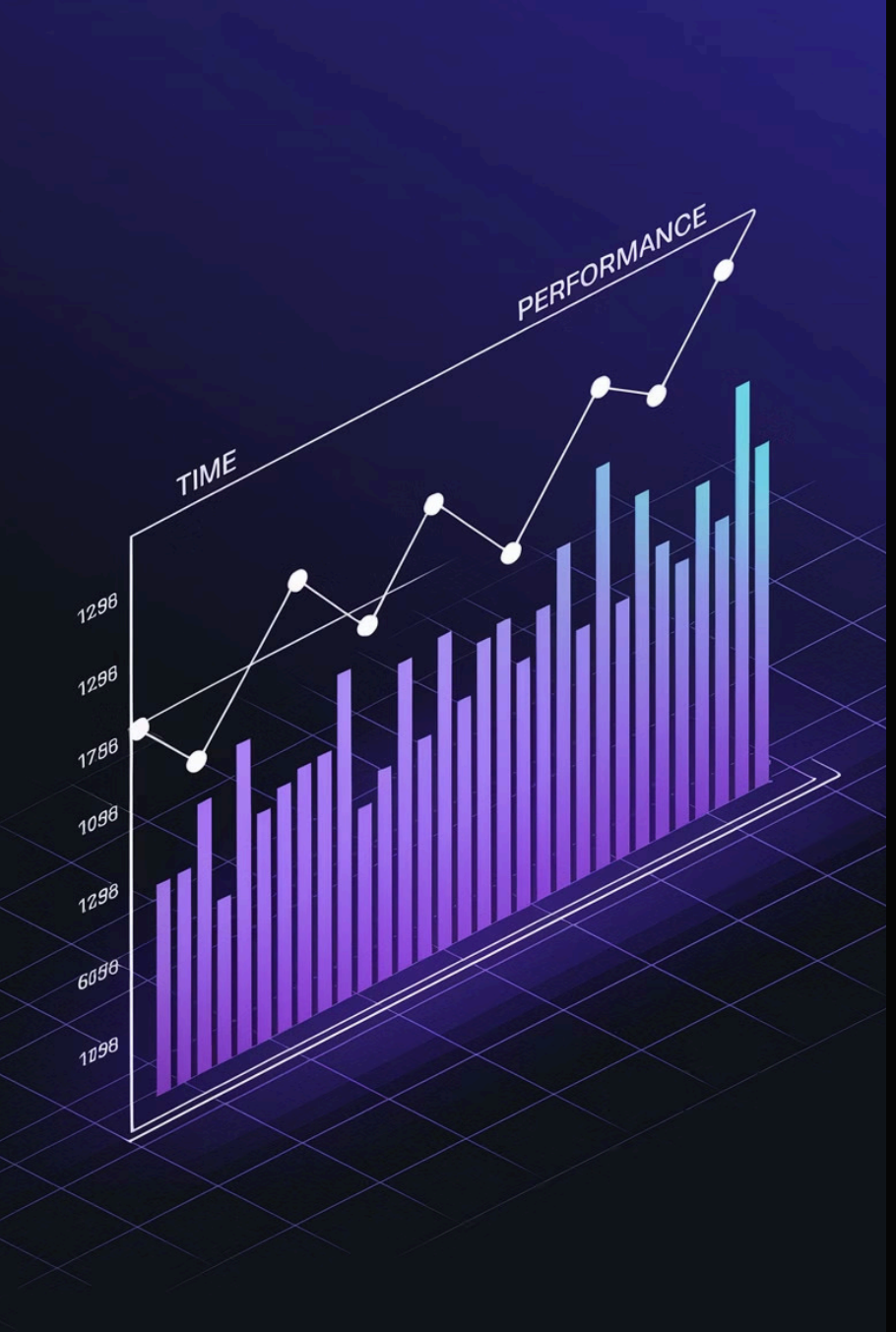
Anomaly Mining

Discovers rare events or data points that deviate from the typical patterns.









Assessing the Interestingness of Mining Results

Understanding the interestingness of mined patterns and models is crucial for guiding data mining and making informed decisions.

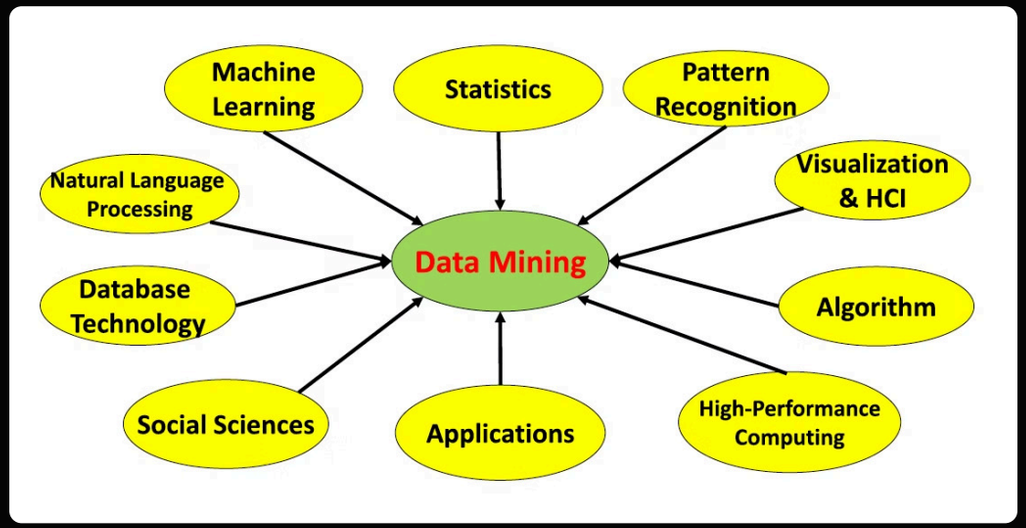
Example Interestingness Metrics

$support(X \Rightarrow Y) = P(X \cup Y)$

$confidence(X \Rightarrow Y) = P(Y|X)$

Data Mining: Confluence of Multiple Disciplines

Data mining serves as a confluence of multiple disciplines, contributing to its success and extensive applications while simultaneously being nurtured by and impacting these fields.



The interdisciplinary nature of data mining research and development contributes significantly to its success, while its applications continue to influence the evolution of these contributing disciplines.

Machine Learning & Statistics

Core analytical methods and statistical foundations for pattern discovery and knowledge extraction

Database Technology

Efficient data storage, retrieval, and management systems for handling massive datasets

Computing & HCI

High-performance computing infrastructure and interactive visualization techniques

Field Sciences

Social sciences, natural language processing, and pattern recognition across various fields

Machine Learning and Data Mining

Machine Learning

Core focus on algorithmic techniques to learn/improve from data (e.g., classification, clustering).

- Supervised learning with labeled data (e.g., recognizing handwritten digits)
- Unsupervised learning with unlabeled data (e.g., grouping images into clusters)

Thinking on **Machine Intelligence**

Data Mining

Tackles large, real-world data with domain-driven needs (e.g., frequent pattern mining, network analysis).

- Handles huge/streaming data with scalable algorithms
- Uses weakly supervised methods for limited labeled data (semisupervised, active learning, transfer learning)

Thnking on **Business Intelligence**

Both are intertwined but data mining goes broader and interdiscipline—targeting big data challenges and application-specific solutions.

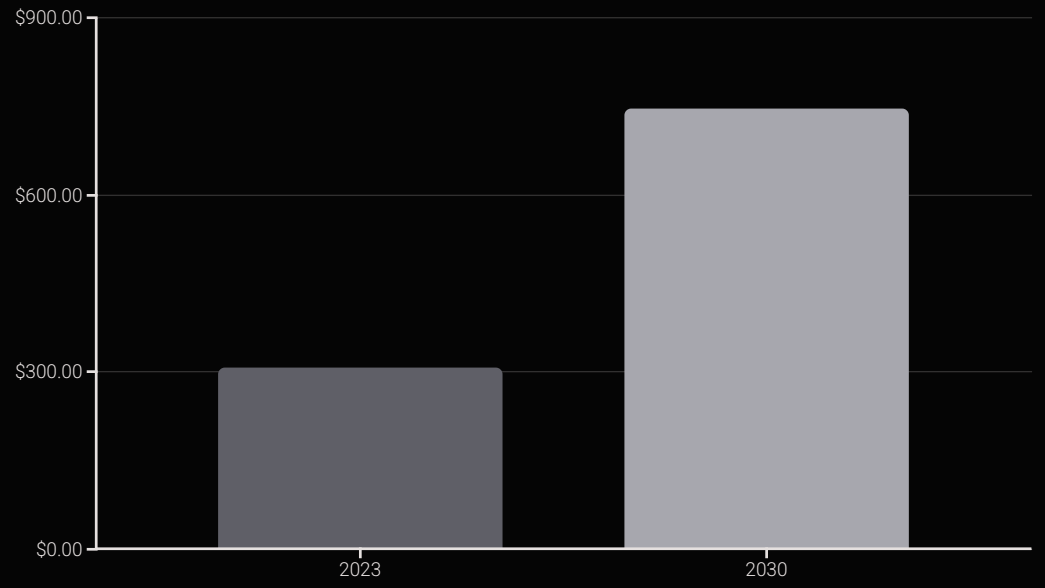
Big Data and Data Mining

Global Big Data Analytics Market Size

350B	12%	40%
Market Size	CAGR	North America
Expected to reach $350 billion by 2025.	Expected compound annual growth rate of 12%.	North America accounts for 40% of the global market.

The global big data analytics market continues to expand. North America leads the market, while the Asia-Pacific region shows the fastest growth. Enterprise demand for data-driven decision making continues to rise.

Big Data and Data Mining (Cont'd)



Kyligence

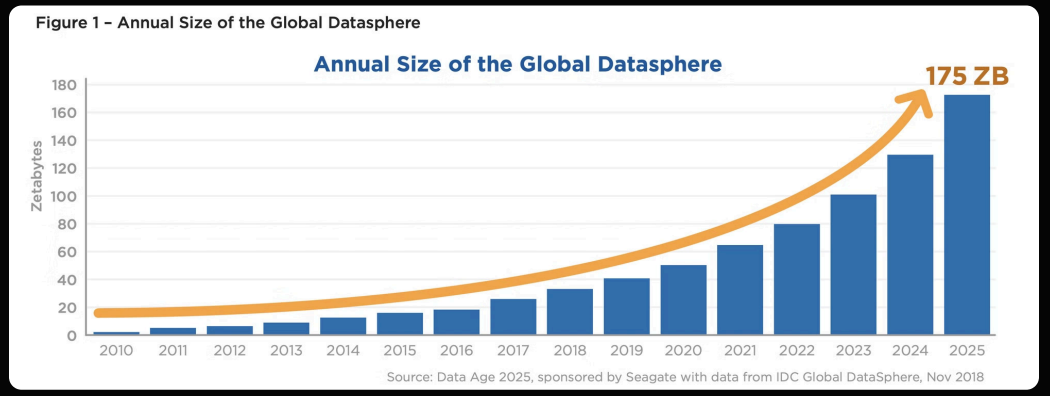
The Top 10 Data and Analytics ...

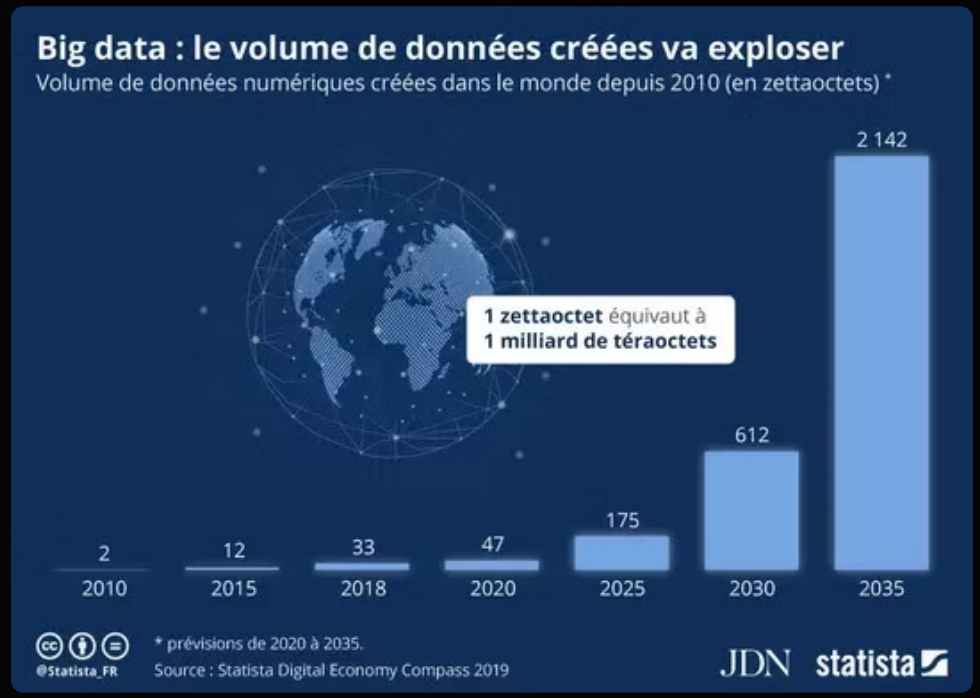
The global big data analytics market is projected to grow from $307.52 billio...

According to statistics, the global big data analytics market is rapidly expanding, projected to grow from approximately $307.52 billion in 2023 to $745.15 billion by 2030. This demonstrates the enormous potential and rapid development trend of the data industry.

Big Data and Data Mining (Cont'd)

Today's big data is small data in the future





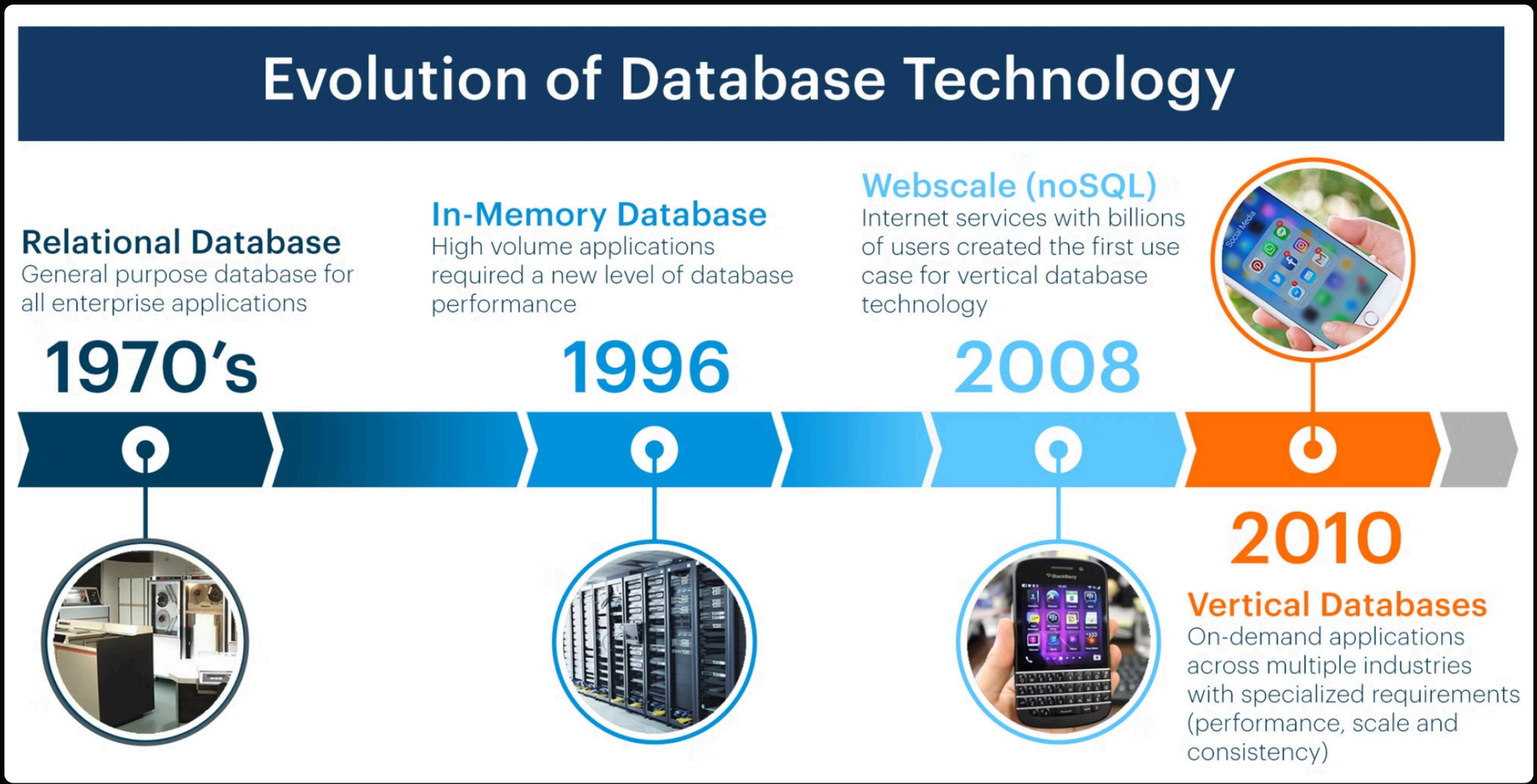
We urgently need a Global Data Convention. Here's why

Global data volume will multiply by 45 between 2020 and 2035

Evolution of Database Technology in the Big Data Era

A historical overview from relational systems to modern scalable solutions

Key milestones: Relational DBs, NoSQL, NewSQL, and hybrid approaches



Relational Databases: The Foundation

1970s – 1990s:

1

- Emergence of the relational model (Edgar F. Codd's work)
- **SQL** standardized and adopted widely (IBM's System R, Oracle, etc.)
- Focus on **ACID** properties (Atomicity, Consistency, Isolation, Durability)

2

Impact:

- Structured query language enabled complex queries
- Laid the groundwork for business applications and data warehousing

Good Ref for DB history, its evoluation and the furture development: "What Goes Around Comes Around... And Around...", **Michael Stonebraker**, Andrew Pavlo , July 2024, ACM SIGMOD Record, Volume 53, Issue 2.
https://sigmodrecord.org/2024/06/30/what-goes-around-comes-around-and-around/

The Big Data Challenge & Changing Needs

2000s Onward:

- Explosion in data volume, variety, and velocity (Internet, social media, IoT)
- Traditional relational DBs struggled with:
  - Horizontal scaling limitations
  - Handling unstructured and semi-structured data

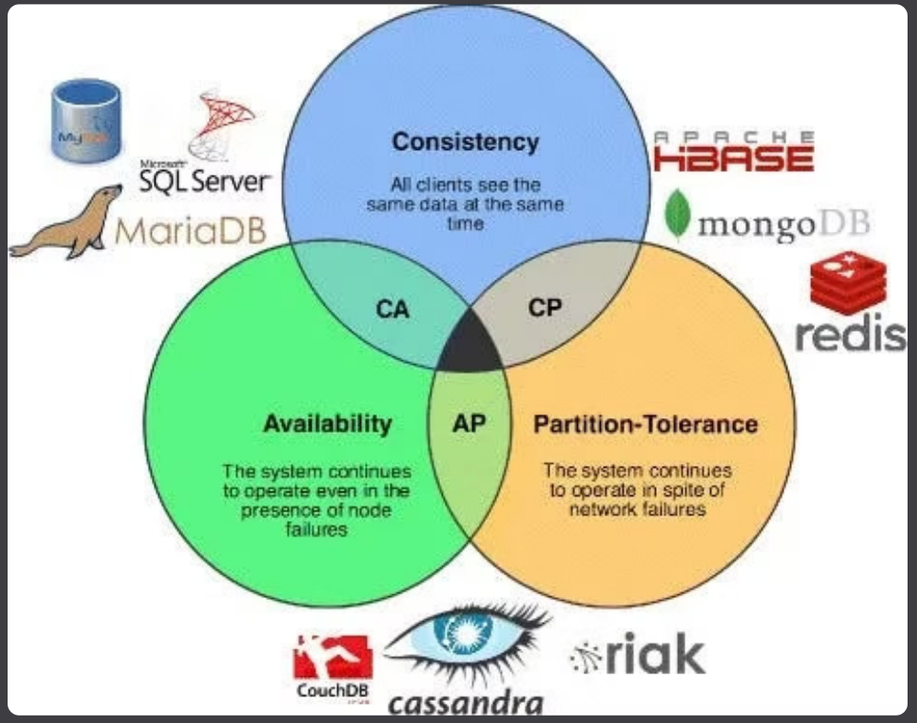
Result:

A demand for new database models that emphasize scalability, flexibility, and distributed computing

NoSQL Databases: Embracing Scalability & Flexibility

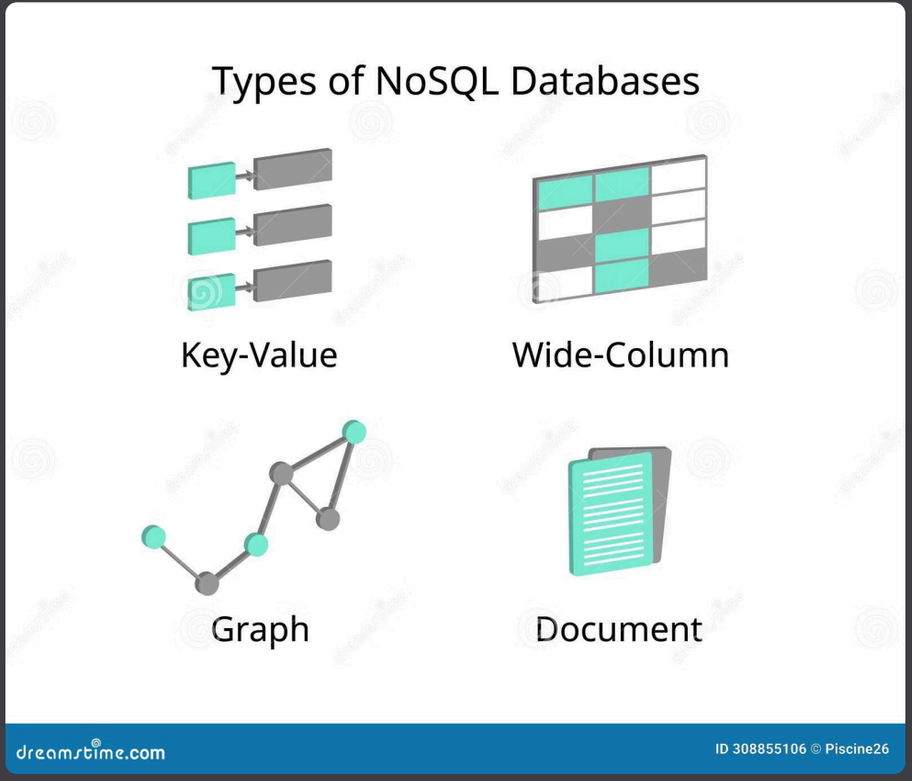
Characteristics:

- **Schema-free** designs and support for unstructured data
- **Horizontal scalability** through distributed architectures
- CAP theorem trade-offs (Consistency, Availability, Partition tolerance)



Popular Types & Examples:

- **Key-Value Stores:** Redis, Riak
- **Document Stores:** MongoDB, CouchDB
- **Column-Family Stores:** Cassandra, HBase
- **Graph Databases:** Neo4j, OrientDB



CAP Theorem: A Brief Overview

The CAP theorem, also known as Brewer's theorem, states that in a distributed data system, you can only achieve two of the following three guarantees simultaneously:

Consistency (C)

Every read receives the most recent write or an error

Availability (A)

Every request receives a response, without guarantee of most recent write

Partition Tolerance (P)

System continues to operate despite network failures

Key Insight: When network partitions (communication failures) occur, a distributed database must choose between ensuring consistency or availability.

Practical Impact:

1. **NoSQL Systems**
   Often favor availability and partition tolerance over strict consistency
2. **Relational and NewSQL Systems**
   Tend to prioritize consistency, sometimes sacrificing availability in distributed environments

NewSQL Databases: Bridging the Gap

Motivation:

Combine the strong consistency and familiar SQL interface of relational DBs with the scalability of NoSQL

Key Features:

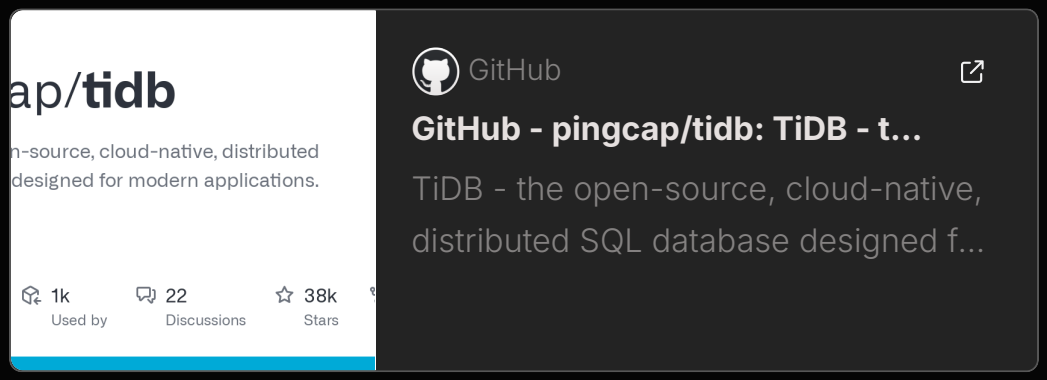
- ACID compliance in distributed environments
- High throughput and low latency

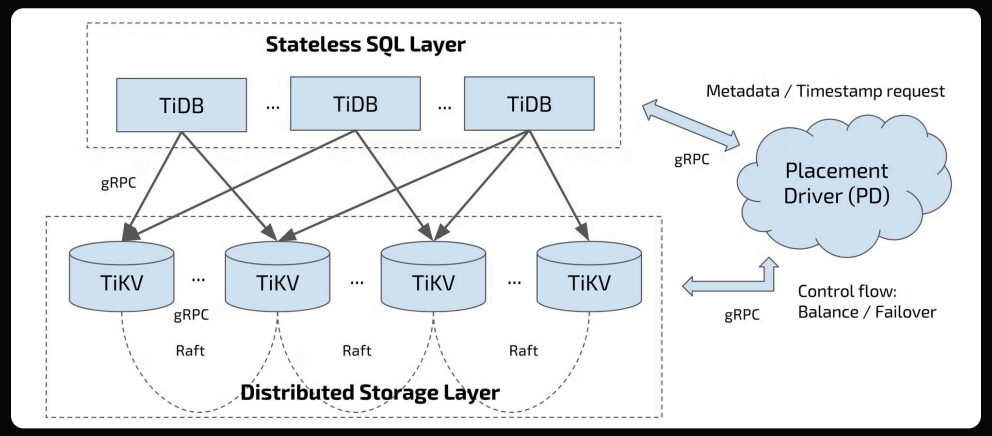
Examples:

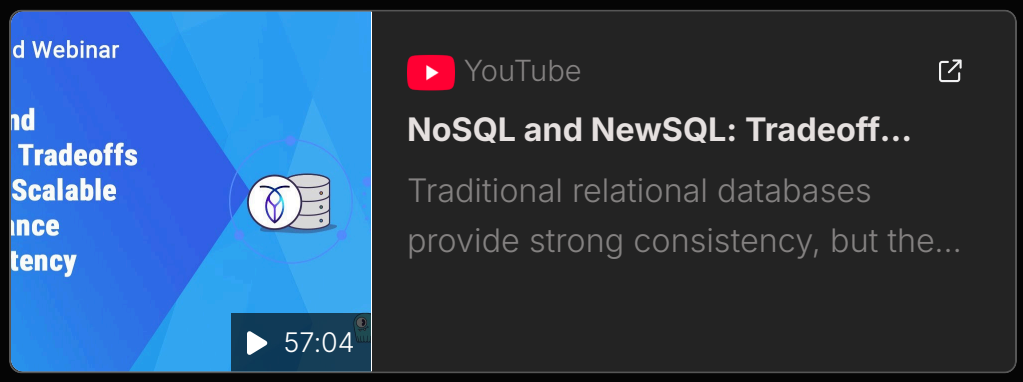
Google Spanner, VoltDB, CockroachDB, TiDB

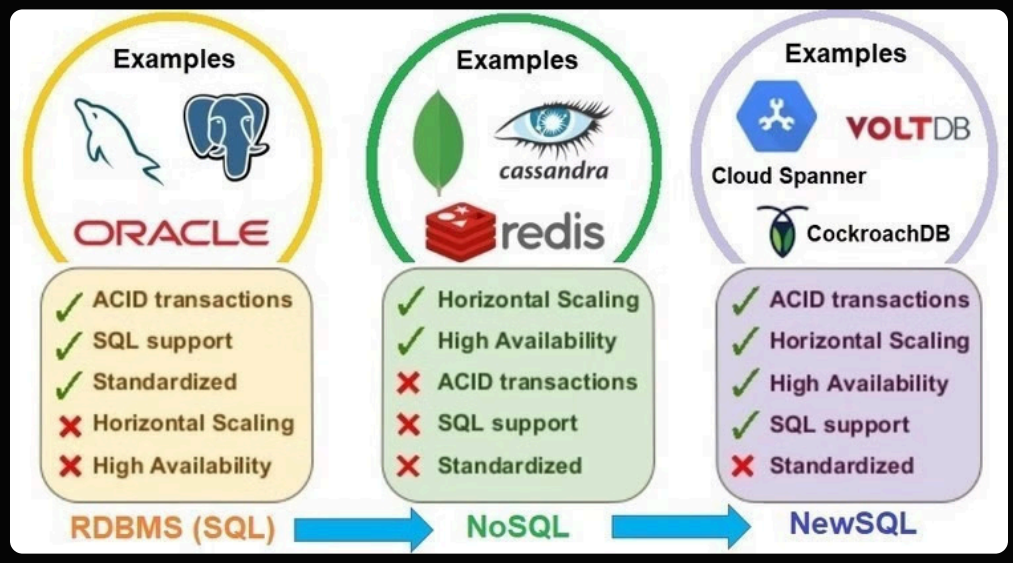
Impact:

Offers a path for organizations needing both reliability and massive scale









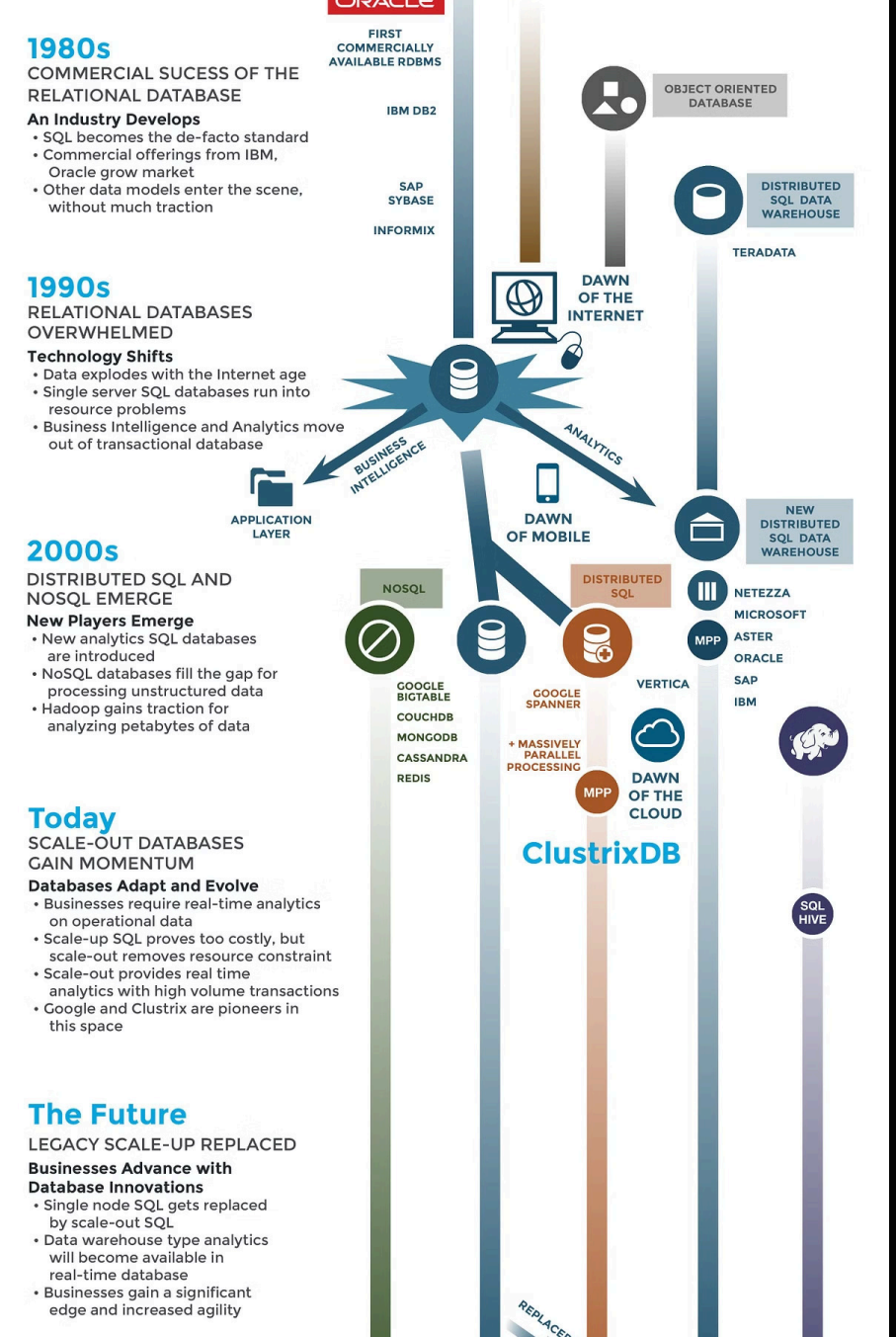
Future Trends of Big Data Technology

Hybrid Solutions:

- Integration of traditional data warehousing with big data platforms (e.g., data lakes)
- Polyglot persistence: using multiple database types based on workload

Future Trends:

- Increased automation in scaling and data integration
- Real-time analytics with stream processing
- Continued innovation blending SQL's expressiveness with NoSQL's performance



Summary of Big Data Technology

1 Key Takeaways:

- Relational DBs built the foundation for structured data management
- Big data challenges drove the emergence of NoSQL for flexible, scalable storage
- NewSQL offers a middle ground with SQL support and distributed scalability

2 Discussion Points:

- How do your current projects balance consistency and scalability?
- Which database model best fits your organization's big data needs?

Brief History of Data Mining Society

1989

IJCAI Workshop on Knowledge Discovery in Databases, leading to the publication "Knowledge Discovery in Databases" by G. Piatetsky-Shapiro and W. Frawley (1991)

1991–1994

AAAI Workshops on Knowledge Discovery in Databases, resulting in "Advances in Knowledge Discovery and Data Mining" by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (1996)

1995–1998

International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)

1997

Launch of Journal of Data Mining and Knowledge Discovery

1998

Establishment of PAKDD and PKDD conferences

Beginning of ACM SIGKDD conferences and SIGKDD Explorations

2001

Launch of SIAM-Data Mining (SDM) and IEEE ICDM conferences

2007

Introduction of ACM Transactions on KDD

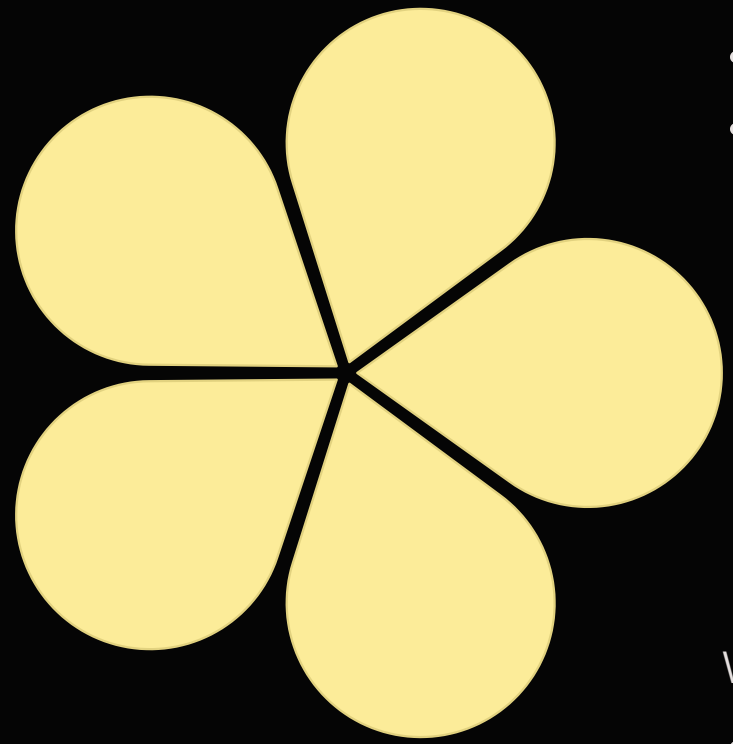
Conferences and Journals on Data Mining

Core Data Mining

- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- IEEE Int. Conf. on Data Mining (ICDM)
- ECML-PKDD & PAKDD
- ACM Int. Conf. on Web Search and Data Mining (WSDM)

Key Journals

- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- ACM Trans. on KDD (TKDD)
- ACM Trans. on Intelligent Systems and Technology (TIST)



Databases & Big Data

- ACM SIGMOD, VLDB, IEEE ICDE
- EDBT, ICDT
- IEEE BigData

Machine Learning & AI

- ICML, NeurIPS, ICLR
- AAAI, IJCAI
- CVPR (Pattern Recognition)

Web, IR & NLP

- WWW, SIGIR, WSDM
- ACL, NAACL, EMNLP
- ACM CIKM

Kaggle: Data Science Community

Collaboration Hub

Connect and work with data scientists and machine learning engineers from around the world

Competitive Platform

Participate in data science competitions to solve real-world problems and showcase your skills

Learning Environment

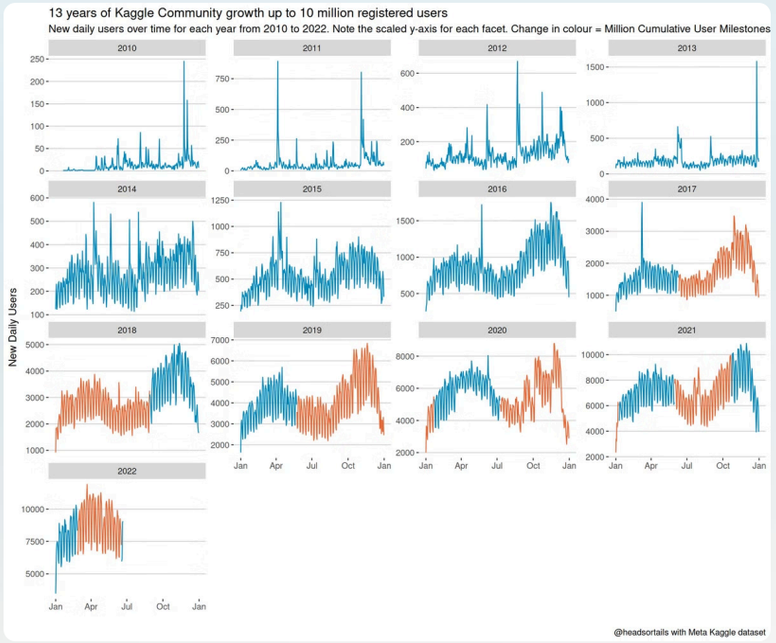
Access educational resources and learn from top data scientists in a supportive community

KDD Cup and then Kaggle

The KDD Cup is an annual data mining and machine learning competition established in 1997. It is held in conjunction with the ACM SIGKDD conference. Participants compete to solve challenging real-world problems using large datasets to advance the field of data mining and knowledge discovery.

NTU has dominated the KDD Cups for six consecutive years.

Championships won include KDD Cup 2008, 2010, 2011, 2012, and 2013.



The History of Kaggle

1

2010: Founded

Launched as a platform for predictive modeling competitions

2

Growth

Rapidly became the hub for data science challenges worldwide

3

Community Impact

Attracted thousands of practitioners and experts

4

2017: Acquisition

Joined Google, expanding resources and global reach

Kaggle

The Advancement of Kaggle Competiti...

Explore and run machine learning code with Kaggle Notebooks | Using data from multiple...

Overview of Data Competition Flow

Problem Statement

Organizers post a real-world problem and dataset

Final Evaluation

Winners determined after rigorous review process

Data Distribution

Participants receive training data to build models

Submission Process

Competitors submit predictions which are automatically scored

Leaderboard

Real-time ranking based on model performance

Detailed Data Competition Workflow

Registration & Setup

Participants sign up and review competition rules

Data Exploration

Download datasets and perform exploratory data analysis

Model Development

Develop and validate predictive models using provided data

Submission & Scoring

Generate predictions and submit via Kaggle interface

Iteration

Refine models based on leaderboard feedback

Questions & Discussion

Open discussion for questions and insights on data mining and preprocessing techniques.