# 732A54/TDDE31

## Big Data Analytics

### Introduction

Huanyu Li

Slides based on previous slides from Patrick Lambrix

LINKÖPING UNIVERSITY

# Teachers

- Examiner: Huanyu Li, HCS, IDA
- Lectures: Huanyu Li,
  Christoph Kessler, SAS, IDA
  Johan Alenlöv, STIMA, IDA
- Labs: Huanyu Li (RDB, BDA1, BDA2),
  Felix Ramnelöv (BDA1, BDA2),
  Johan Alenlöv (BDA3)
- Director of studies: Anders Fröberg, HCS, IDA

# Course overview

- Parallel algorithms for processing Big Data (lectures + lab)
- Introduction to BDA labs (exercise sessions)
- Databases for Big Data (lectures, some topics are covered in labs)
- Machine Learning for Big Data (lectures + lab)

- Visit to National Supercomputer Centre – organization ongoing

# ILOs (Intended Learning Outcomes)

- Collect and store Big Data in a distributed computer environment
- Perform basic queries to a database operating on a distributed file system
- Account for basic principles of parallel computations
- Use MapReduce concept to parallelize common data processing algorithms
- Account for how standard machine learning models should be modified in order to process Big Data
- Use tools for machine learning for Big Data

# Examination

- Written exam
  - Dates:
    - 2026-06-02, 08:00-12:00
    - 2026-08-18, 14:00-18:00
    - 2027-01-07, 08:00-12:00
- Labs
  - RDB for 732A54
  - BDA1 – Spark
  - BDA2 – Spark SQL
  - BDA3 – ML with Spark

# Practical Information

- Sign up for labs via webreg (in pairs)
  - Deadline: **April 7th**
    - https://www.ida.liu.se/webreg3/732A54-2026-1/LAB/
    - https://www.ida.liu.se/webreg3/TDDE31-2026-1/LAB/
- Hand in labs via email and GitLab repositories
- A GitLab repository will be assigned to each group on **April 8th**

- BDA labs require special access to NSC resources
  - Fill out the form in time
  - Resources are only guaranteed during the course period (until 2026-07-01)
  - Strong recommendation to hand in as soon as possible

# Changes w.r.t. last year

- Labs
  - lab compendium for better understanding

# Introduction

# Data, Information or Knowledge?

- Data
  - 36, 38, 39, 36, 35.5
- Information
  - Tom's body temperature over the last hour was 36, 38, 39, 36 and 35.5 degrees
- Knowledge
  - If a person has the body temperature 39, then that person has a fever
- Data consists of known facts recorded in some form, carring implicit meaning

How is the data stored?
How is the data accessed?

# Storing data in various formats and ways

- Text
- Semi-structured data
- Structured data
- Data models
- Knowledge bases
  - Rules + Facts

# Storing data in various formats and ways

- Text
- Semi-structured data
- **Structured data**
- **Data models**
- Knowledge bases
  - Rules + Facts

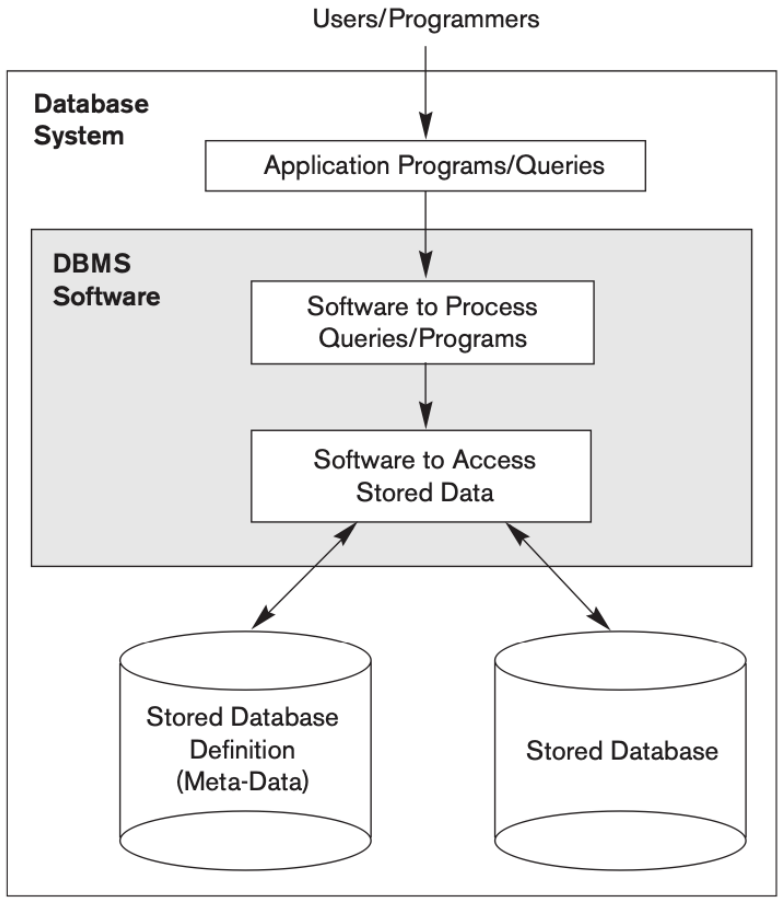
# Data and Data Storage

- Database / Data source
- One (of several) ways to store data in electronic format
- Used in everyday life: bank, hotel, reservations, library search, shopping

# Databases / Data sources

- Database management systems (DBMS):
  - a collection of programs to create and maintain a database
- Database system = database + DBMS
  - MySQL, Oracle, PostgreSQL, etc.

# Databases / Data sources




Elmasri, R. & Navathe, S.B. (2016). *Fundamentals of Database Systems* (7th ed.). Pearson. Chapter 1.

# What information is stored?

- Model the information
  - Entity-Relationship model (ER)
  - Unified Modeling Language (UML)

# Entity-Relationship model (ER)

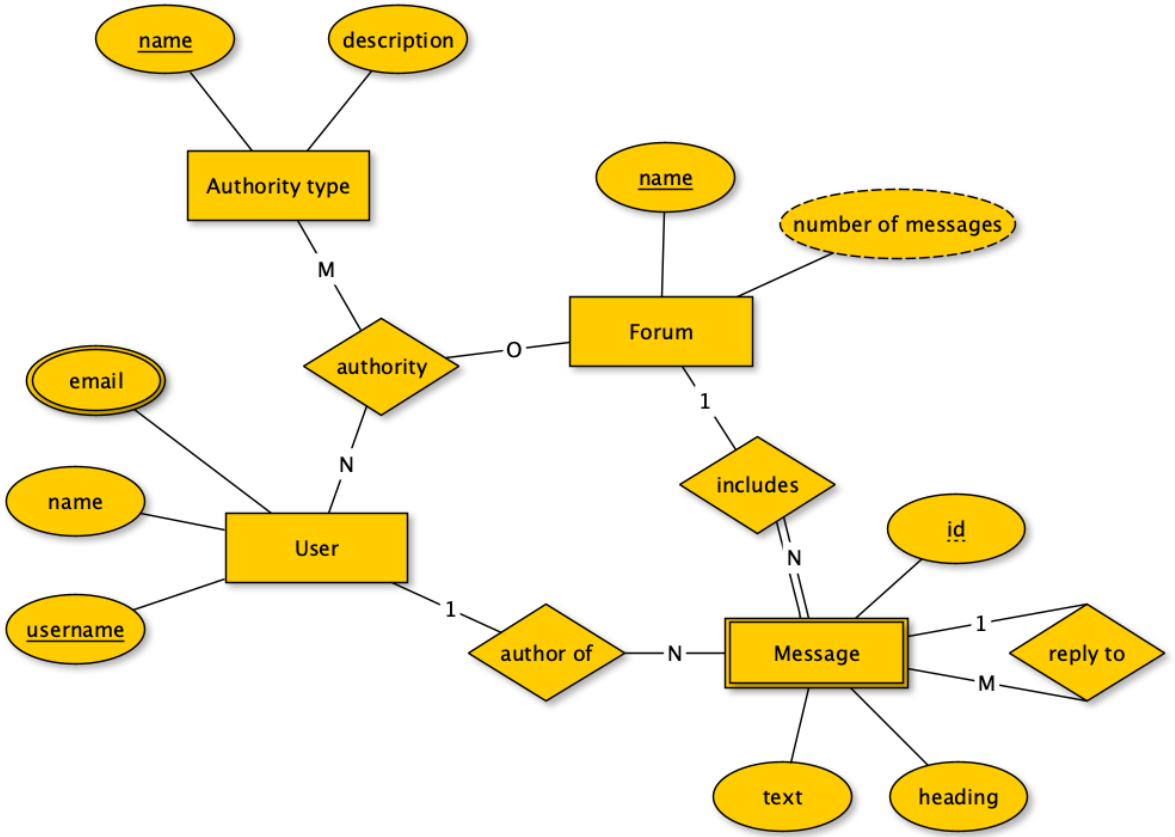
- What information is stored in ER?
  - entities and attributes
  - entity types
  - key attributes
  - relationships
  - cardinality constraints on relationships

- EER: Enhanced Entity-Relationship
  - sub-types

# Example: Web forum

- Multiple web forums in the same database
- Each forum has a unique name
- The messages in each forum should be stored along with heading, text, author, and commented message
- Users should be stored along with usernames and names
- Different authority types should be stored along with a description of each (e.g., admin, reader, writer, …)
- A user can have different authority types for different forums

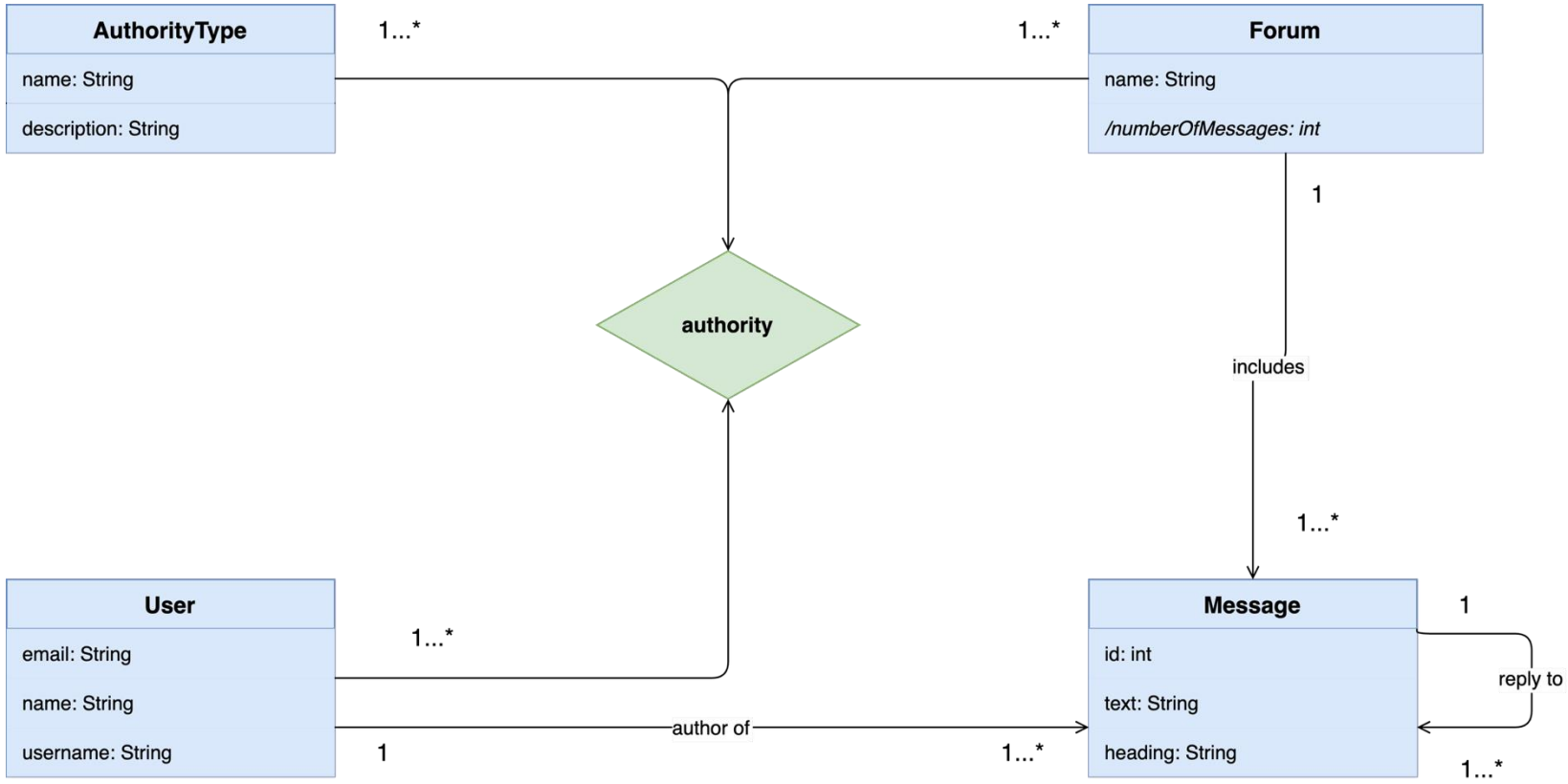
# Example: Web forum - EER



# Unified Modeling Language (UML)

- What information is stored in UML?
  - Use case diagrams
  - Class diagrams for representing (data) structures
  - Activity diagrams
  - …

# Example: Web forum



# How is the data accessed?

- Relational databases – SQL
- Data is stored as ”tables”
- Data is accessed via SQL queries (Structured Query Language)

# Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, temporal, multimedia, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases

# Evolution of Database Technology

- 2000s:
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
  - NoSQL databases
- 2010s:
  - Big data
  - NoSQL databases, graph databases
  - Knowledge graphs

# Evolution of Database Technology

- 2020s:
  - Graph databases
  - Cloud-native databases
  - Vector databases
  - Data spaces and data lakes
  - …

# Storing data in various formats and ways

- Text
- Semi-structured data
- **Structured data**
- **Data models**
- Knowledge bases
  - Rules + Facts

# Knowledge bases

- Facts:
  - *hasSymptom(patient_1, fever),*
  - *hasSymptom(patient_1, cough),*
  - *bodyTemperature(patient_1, 39.2)*
  - *bodyTemperature(patient_2, 38)*
- Rules
  - *hasFever(X) :- bodyTemperature(X, T) AND T > 38.5*
  - *suspect_flu(X) :- hasSymptom(X, fever) AND hasSymptom(X, cough) AND hasFever(X)*
- Queries
  - *hasFever(patient_1)* → *true*
  - *suspect_flu(patient_1)* → *true*
  - *suspect_flu(patient_2)* → *false*

# Interested in more

- 732A57/TDDD12/TDDD37/TDDD81 Database Technology
  - relational databases
- TDDD43 Advanced data models and databases
  - (semi-structured data, knowledge base)

# Data Analytics

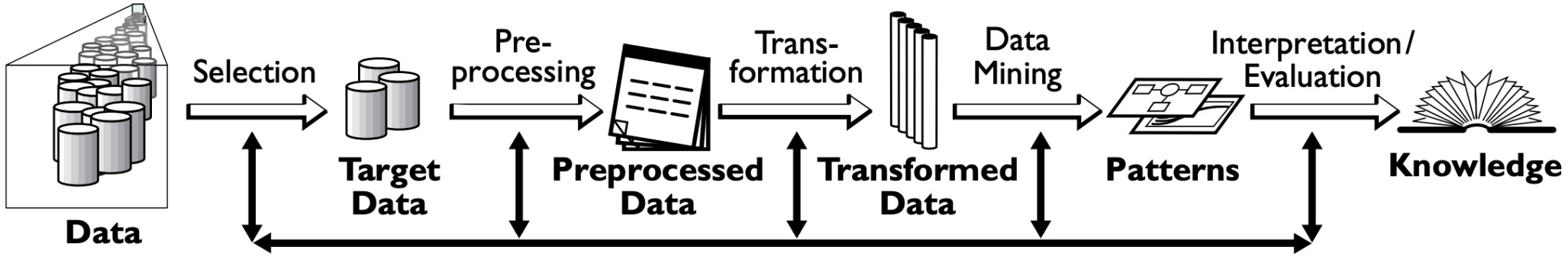
- Discovery, interpretation and communication of meaningful patterns in data

# Why Analytics?

- The Explosive Growth of Data
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digitalization, …

- *"We are drowning in information, but starving for knowledge!"*
  - By John Naisbitt in his 1982 book "Megatrends: Ten New Directions Transforming Our Lives"

# Knowledge Discovery (in Database) Process



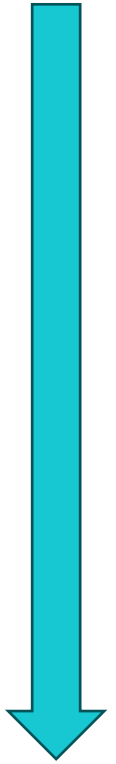

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

# Data Analytics – Oracle Data Mining manual

- Classification
- Regression
- Clustering
- Association
- Anomaly detection
- Feature extraction and creation
- Time Series
- https://docs.oracle.com/en/database/oracle/oracle-database/19/dmcon/index.html

# Data Analytics – IBM

- What is happening? — Descriptive
  - Discovery and explanation
- Why did it happen? — Diagnostic
  - Reporting, analysis, content analytics
- What could happen? — Predictive
  - Predictive analytics and modeling
- What action should I take? — Prescriptive
  - Decision management
- What did I learn, what is best? — Cognitive



# Example 1: Market analysis and management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Cross-market analysis
  - Find associations/co-relations between product sales
  - Predict based on such associations
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers

# Example 2: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications:
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests

# Data Mining – What kinds of patterns?

- Concept/class description:
  - Characterization: summarizing the data of the class under study in general terms
    - E.g. Characteristics of customers spending more than 10000 sek per year
  - Discrimination: comparing target class with other (contrasting) classes
    - E.g. Compare the characteristics of products that had a sales increase to products that had a sales decrease last year

# Data Mining – What kinds of patterns?

- Frequent patterns, association, correlations
  - Frequent itemset
  - Frequent sequential pattern
  - Frequent structured pattern

  - *E.g. buy(X, "Cream cheese") → buy(X, "Butter") [support=3.1%, confidence=75%]*
  - *confidence*: if X buys cream cheese, then there is 75% chance that X buys butter
  - *support*: of all transactions under consideration 3.1% showed that cream cheese and butter were bought together
  - *E.g. Age(X, "20..29") and income(X, "20k..29k") → buys(X, "iphone") [support=2%, confidence=60%]*

# Data Mining – What kinds of patterns?

- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction.
  - The derived model is based on analyzing training data – data whose class labels are known.
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values

# Data Mining – What kinds of patterns?

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster customers to find target groups for marketing
  - Maximizing intra-class similarity & minimizing inter-class similarity
- Outlier analysis of the data
  - Outlier: Data object that does not comply with the general behavior
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation

# Interested in more

- 732A95/TDDE01 Introduction to machine learning
- 732A75/TDDD41 Advanced data mining / Data mining – clustering and association analysis

# Big Data

- So large data that it becomes difficult to process it using a ‘traditional’ system

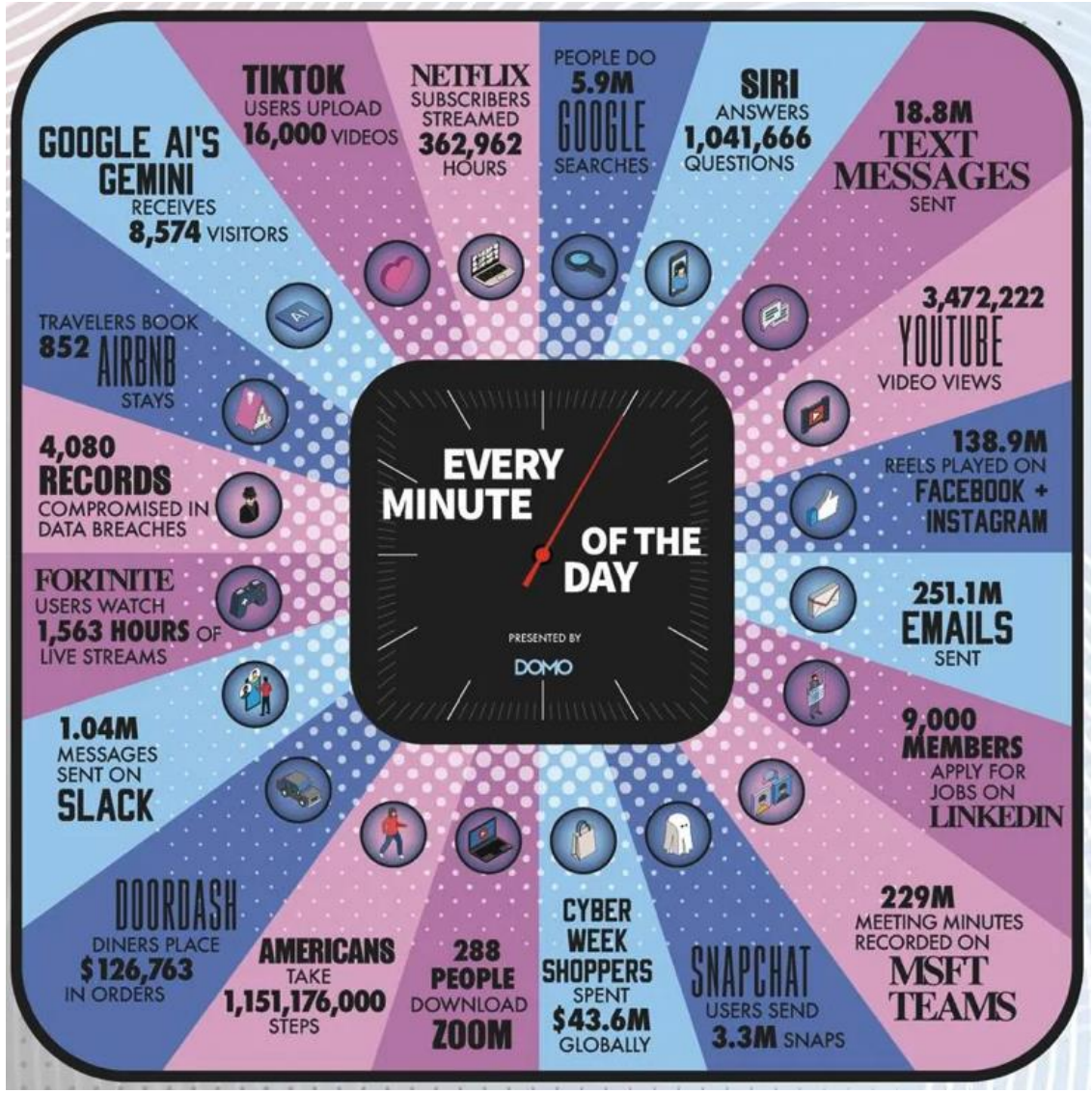
> *“Big data involves data whose volume, diversity, and complexity requires new techniques, algorithms, and analyses to extract valuable (hidden) knowledge”*
>
> *-- Triguero, I., & Galar, M. (2023).*

# Big Data – 3Vs

- Volume
  - size of the data

# Big Data – 3Vs

- Volume
  - 500 hours of video uploaded to YouTube per minute (2019)
  - European Organization for Nuclear Research (CERN)
    - process on average 1 petabyte (1 million GB) of data per day during one run on the Large Hadron Collider (LHC)
  - ICA's annual net sales were around 145 billion SEK in 2023
    - 400-500 million transactions per year with an average grocery around 300-400 SEK
  - Airbus A350 generates up to 1TB of data per flight across 400, 000 sensor parameters (2025)
  - SMS, e-mail, internet, social media




https://bondhighplus.com/blog/what-happens-in-an-internet-minute/

| Year | Emails Sent | Google Searches | YouTube Views | TikTok Views | Online Spend | App Downloads | Instagram Activity |
|---|---|---|---|---|---|---|---|
| 2025 | 231M+ | 6.3M+ | 3.47M | 625M videos watched | $43.6M+ | 174K | 66K photos shared |
| 2024 | 231M | 6.3M | 3.47M | 625M videos watched | $43.6M (peak minute) | 174K | 66K photos shared |
| 2023 | 231M | 5.9M | 3.67M | 167M videos watched | $1.5M+ (est.) | 174K | 694M songs streamed |
| 2022 | 231M | 5.9M | 508 hrs uploaded | 1B+ interactions | $443K (Amazon only) | 437K transfers | 527K photos shared |
| 2021 | 225M (est.) | 5.7M (est.) | 5.0M (est.) | 167M (est.) | $1.3M (est.) | 450K (est.) | 800K (est.) scrolling |
| 2020 | 190M | 4.1M | 4.7M | N/A | $1.1M | 400K | 694K scrolling |
| 2019 | 188M | 3.8M | 4.5M | N/A | $996K | 390K | 347K scrolling |
| 2018 | 187M | 3.7M | 4.3M | N/A | $862K | 375K | 174K scrolling |
| 2017 | 156M | 3.5M | 4.1M | N/A | $751K | 342K | 46K posts uploaded |
| 2016 | 150M | 2.4M | 2.78M | N/A | $203K | 51K | 527K photos shared |
| 2015 | 204M | 2M+ | 1.3M | N/A | $83K | 47K | 3,000 uploads |

https://bondhighplus.com/blog/what-happens-in-an-internet-minute/

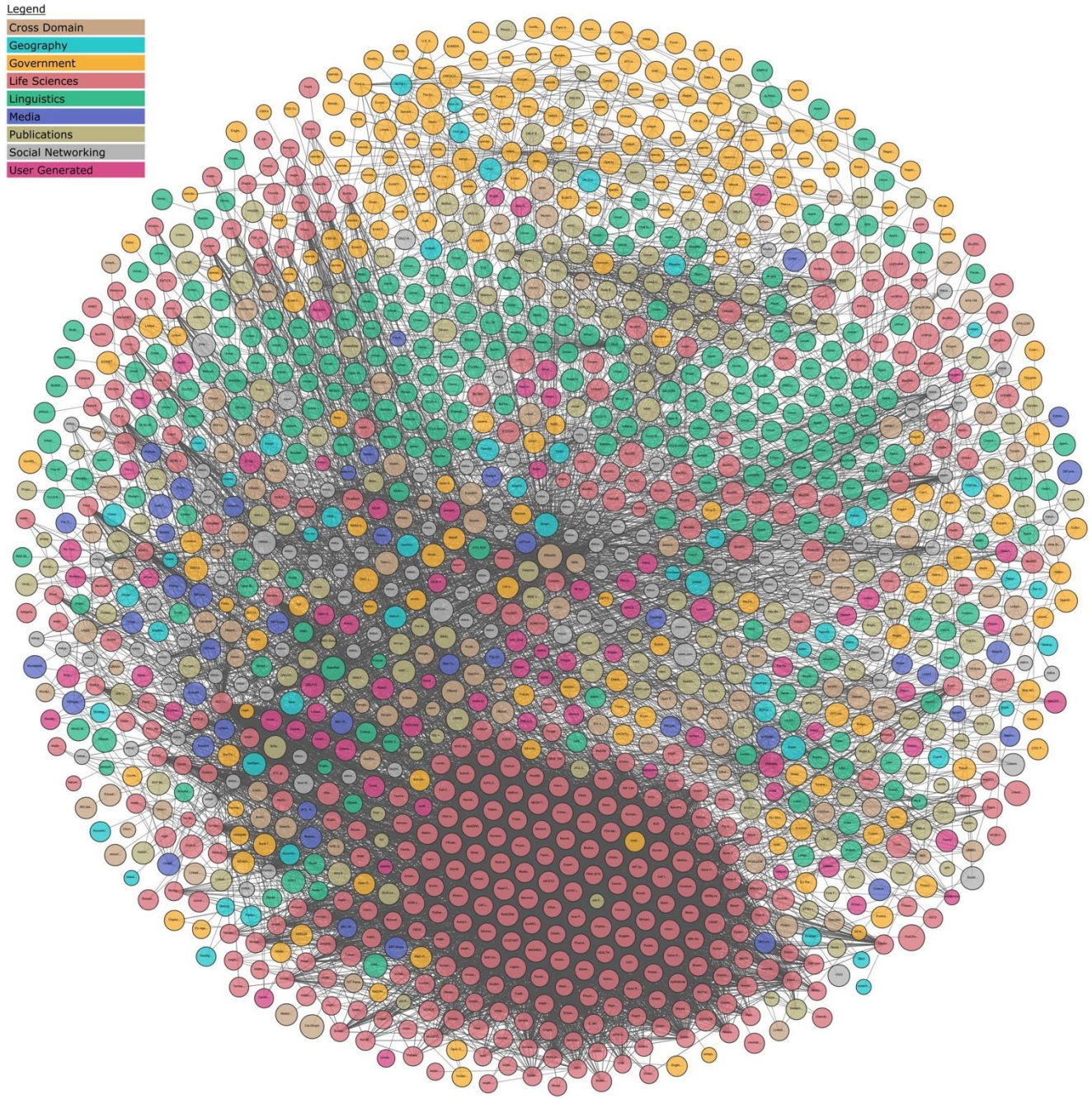
# Big Data – 3Vs

- Volume
  - size of the data
- Variety
  - different formats and structure of data
    - text, semi-structured data, structured data, various databases, knowledge bases

# European data (**data.europa.eu**)

- Formats (# datasets)
  - WMS (485 753)
  - WFS (354 389)
  - CSV (325 494)
  - JSON (85 542)
  - PDF (85 215)
  - HTML (84 429)
  - ZIP (73 996)
  - Excel XLSX (67 408)

- Linked Open Data Cloud
- https://lod-cloud.net/



# Big Data – 3Vs

- Volume
  - size of the data
- Variety
  - different formats and structure of data
    - text, semi-structured data, structured data, various databases, knowledge bases
- Velocity
  - speed of generation and processing of data

# Big Data – 3Vs

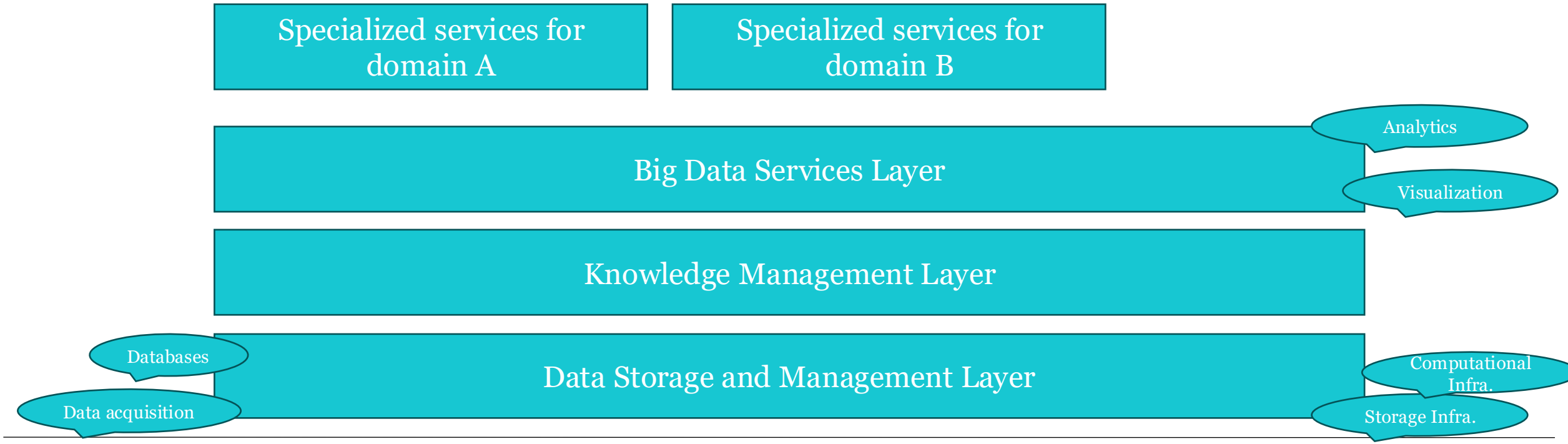
- Velocity
  - Sensor data from (IoT) devices
    - thousands of readings per second
  - Financial market
    - millions of prices per second across global exchanges
  - Social networks
    - X processes hundreds of thousands of posts per minute

# Big Data – other Vs

- Variability
  - inconsistency of the data
- Veracity
  - quality of the data
- Value
  - benefit/value from analyzing the data
- ….

# BDA system architecture

- Layered architecture with many different aspects



# BDA system architecture

- Large amounts of data, distributed environment
  - Unstructured and semi-structured data
  - Not necessarily a schema
  - Heterogeneous
  - Streams
  - Varying quality

Data Storage and Management Layer

# Data Storage and management in this course

- Data storage
  - NoSQL databases
  - OLTP vs OLAP
    - Online Transaction Processing and Online Analytical Processing
  - Horizontal scalability
  - consistency, availability and partition tolerance
- Data management
  - Hadoop Distributed File System

Data Storage and Management Layer

# BDA system architecture

- Semantic technologies
- Integration
- Knowledge acquisition

Knowledge Management Layer

# Knowledge management in this course

- Not a specific focus
- More in TDDD43 for semantic and integration approaches

Knowledge Management Layer

# BDA system architecture

- Analytics services for Big Data

# Big Data Services in this course

- Big data versions of
  - analytics
  - data mining
  - machine learning
- Apply methods to extract (hidden) knowledge or insights in big datasets

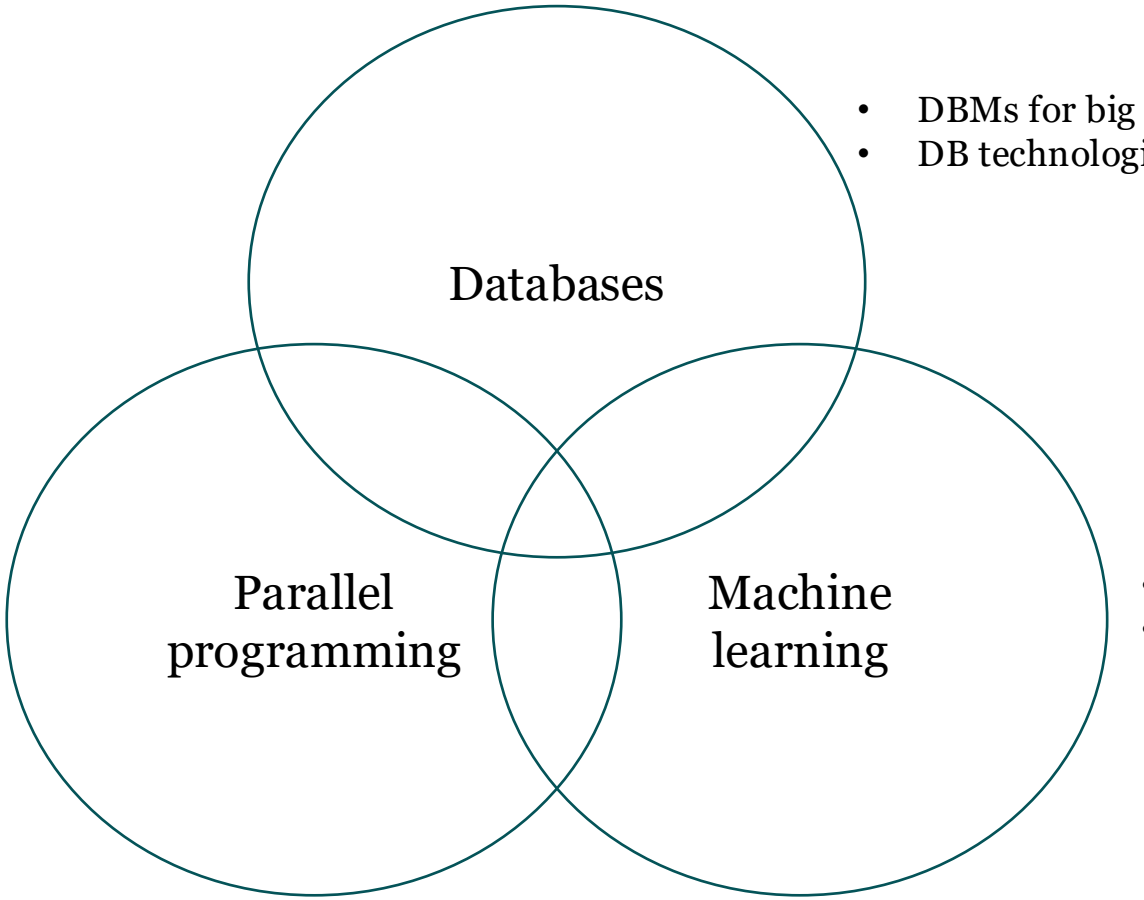
Big Data Services Layer

# Course overview

- Parallel algorithms for processing Big Data (lectures + lab + exercise session)
  - MapReduce, Spark, etc.
- Databases for Big Data (lectures + lab)
- Machine Learning for Big Data (lectures + lab)

- Visit to National Supercomputer Centre – organization ongoing

# Course overview




- DBMs for big data
- DB technologies for data analytics

- High Performance Computing
- Parallel computing
- MapReduce and Spark
- Cluster and Cluster management

- Machine learning with MapReduce
- Machine learning with Spark

# ILOs (Intended Learning Outcomes)

- Collect and store Big Data in a distributed computer environment
- Perform basic queries to a database operating on a distributed file system
- Account for basic principles of parallel computations
- Use MapReduce concept to parallelize common data processing algorithms
- Account for how standard machine learning models should be modified in order to process Big Data
- Use tools for machine learning for Big Data

# Summary

- Big Data
- Big Data System Architecture
- Focuses of this course

LINKÖPING UNIVERSITY