

# SECURITY IN SOCIAL NETWORKS

R DEVIKA

AP II/CSE/SOC

# TEXT BOOKS

- BhavaniThuraisingham, SatyenAbrol, Raymond Heatherly, Murat Kantarcioglu, VaibhavKhadilkar, Latifur Khan, Analyzing and Securing Social Network
- Fadi Al-Turjman, B.D. Deebak, Security in IoT Social Networks, Academic Press, 1st Edition, 2020.al Networks, Auerbach Publications, 1st Edition, 2020.

# What is a Social Network?

- A Social Network Site is a Web-based service that allows individuals to:
  - construct a public or semi-public profile within a bounded system;
  - articulate a list of other users with whom they share a connection;
  - view and traverse their list of connections and those made by others within the system;
  - increasing their social capital.
- We will not consider other kinds of social networks that don't rely on a social network provider.



# Social Media Users Statistics 2023:



1. There are **4.9 billion social media users** in the world as of 2023.
2. It is forecasted that there will be **5.85 billion** social media users worldwide **by 2027**.
3. **Facebook** is the **biggest social media platform** in terms of user base. It has 2.91 billion users as of 2023.

# Introduction

- OSNs
- online applications that allow their users to connect by means of various link type
- a way for users to interact, reflecting the social networks or social relations among people, for example, those who share interests and/or activities
- social media has demonstrated exponential growth, making it the most popular activity platform on the WWW
- social media such as Twitter and Facebook

# Two major technologies in OSN

## i) data mining technologies

Analyzing these networks and extracting useful information

- such as location, demographics, and sentiments of the network participants,

## (ii) security and privacy technologies

- privacy of the participants of the network
- provide controlled access to the information posted

social media and social networks interchangeably

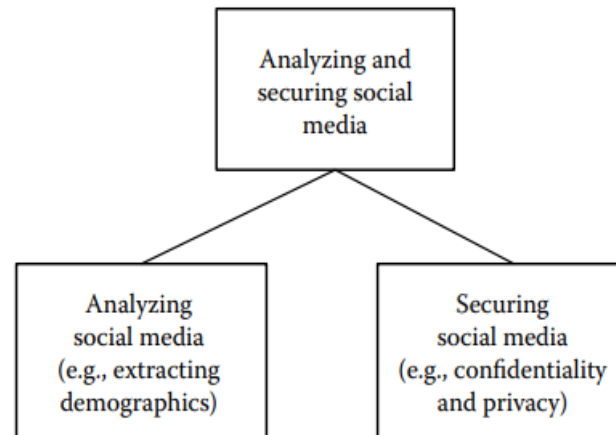
data analytics and data mining interchangeably

# social networks

- networks for users to collaborate and share information
- shared information
  - in the form of text, images, audio, video, and animation, among others.

# ANALYZING SOCIAL NETWORKS

- Carried out manually using statistical techniques
  - determining the friends of individuals,
  - as well as leaders of groups, and
  - analyzing the sentiments of the OSN members and extracting their information

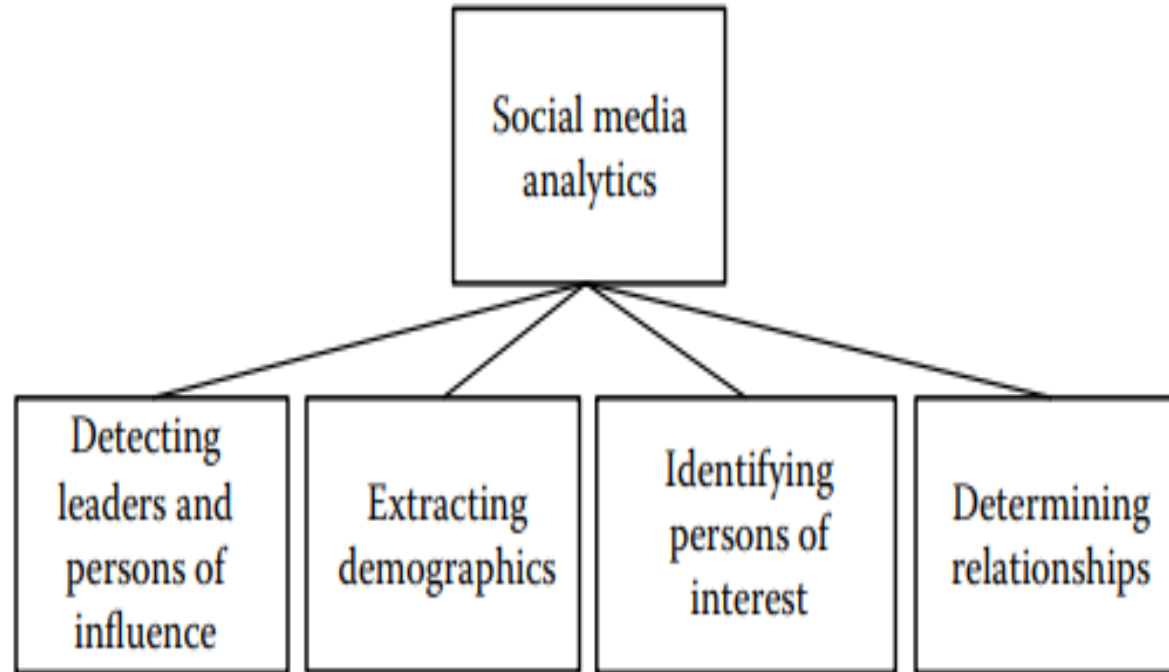




- trends in social networking sites
  - real time and location based
- location-based social media services
  - Privacy and Safety.
    - Gowalla
    - knowing the location of the user makes it easier for spammers to attack the user in a more personalized manner.
  - Trustworthiness
    - companies use social media
      - Companies worldwide, from Burger King to Ford, are using social media to target customers.
  - Advertising and Marketing
    - OSN act as a customer relationship management tool
    - OSNs connect people at low cost;
    - OSNs for advertising in the form of banners and text ads

- Social networking companies have shifted their focus from building relationships to **identifying temporal and spatial patterns** in messages.
- Twitter
  - feature whereby users can find location-based trending topics
    - it only covers users who mention their locations explicitly and,
    - the topic of search is limited to the trending topics.

# Various types of social media analytics approaches



# SURVEY OF SOCIAL NETWORKS

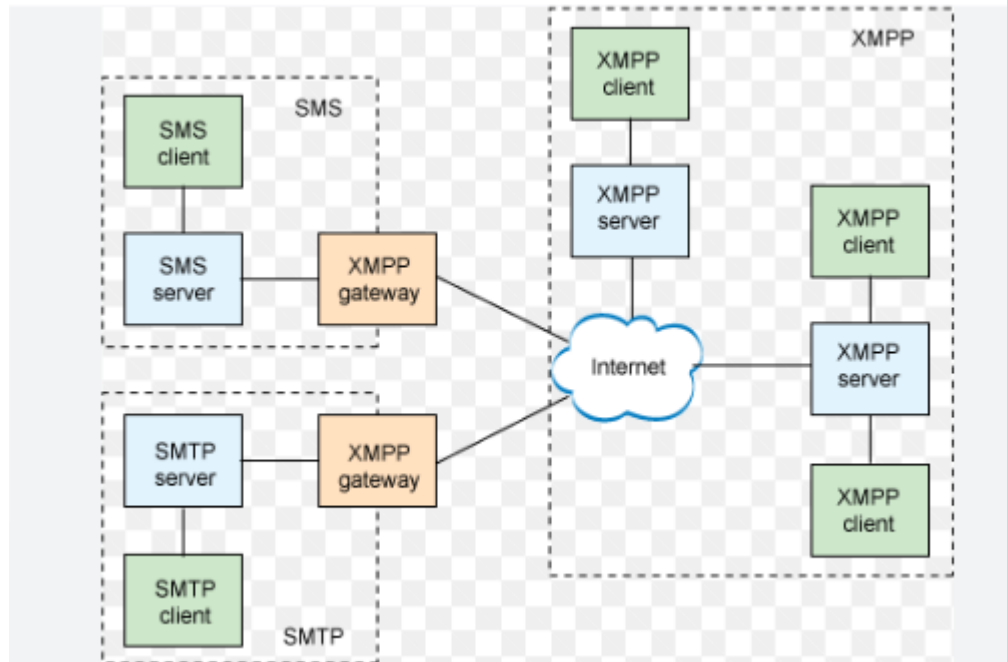
- Classmates
  - By linking together people from the same school and class year
- Bolt (<http://bolt3.com/bolt-social-network/>)
  - e-cards, music sharing, and Internet radio
- AsianAvenue (<http://www.asianave.com/>)
  - a feature allowing members to view their page's history
  - who has viewed their site, making a user's browsing history public, not private
- Gapyear (<https://www.gapyear.com/>)
  - community information and connections for planning a youth's year off
  - Travel advice, volunteer jobs, destinations, and general backpacking information

- Fotki
  - is a photo-sharing social network
  - Fotki has grown as a Web 2.0 service engine
    - support video sharing and audio comments on uploaded content
    - supports many languages
    - provides unlimited storage to paid, premium members
- Open Diary
  - online blogging community
  - free e-mail accounts
  - a continuous supply of public journals from all seven continents
  - security breach

# DETAILS OF FOUR POPULAR SOCIAL NETWORKS

- Facebook
  - Timeline & friends update
    - Work and Education,
    - Family and Relationships
    - Living, Health and Wellness,
    - and Milestones and Experiences
  - Social Experiments
  - security and Privacy- Secure browsing (HTTPS) encrypts your connection to Facebook
- Google+.
  - Circles.
  - Pages
    - profiles for businesses, organizations, publications, or other entities that are not related to a single individual
  - Other services
    - Games, messenger, photo editing and saving, mobile upload and diagnostics, apps, calendars, and video streaming
  - Hangouts- conference call solution / create instant webcasts—like Skype;

# Hangouts is based on many technologies,



- XMPP, Jingle, RTP, ICE, STUN, and SRTP
- Extensible Messaging and Presence Protocol(xmpp)
  - **XMPP lets users' devices send messages asynchronously**, meaning you can send multiple messages in a row without waiting for a response, and two users don't have to be online at the same time in order to message each other
  - decentralized client-server architecture

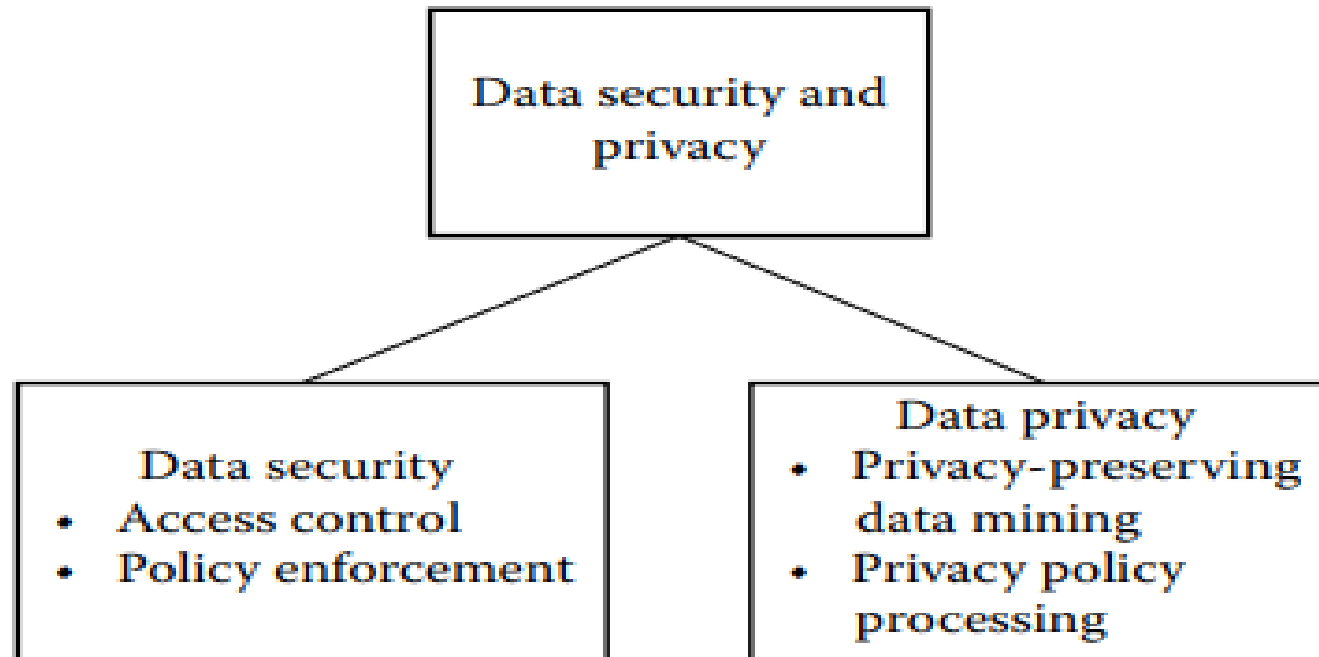
- The Real-time Transport Protocol(RTP) is a network protocol for delivering audio and video over IP networks
- Interactive Connectivity Establishment (ICE) NAT
  - to find ways for two computers to talk to each other as directly as possible in peer-to-peer networking
  - **detects messages and notifications**[STUN].
  - encryption, confidentiality, message authentication, and replay protection to your transmitted audio and video traffic[SRTP].
- Advt
  - To reduce transmission lag to below 100 milliseconds
- Disadvt
  - Cloud computational requirements for Hangouts sessions are immense.



- Twitter
  - 140 Character Limit.
- LinkedIn
  - aims to be professional and make professional connections in all of its social networks
  - Users cannot upload their resumes directly
  - User adds skills and work history to their profile

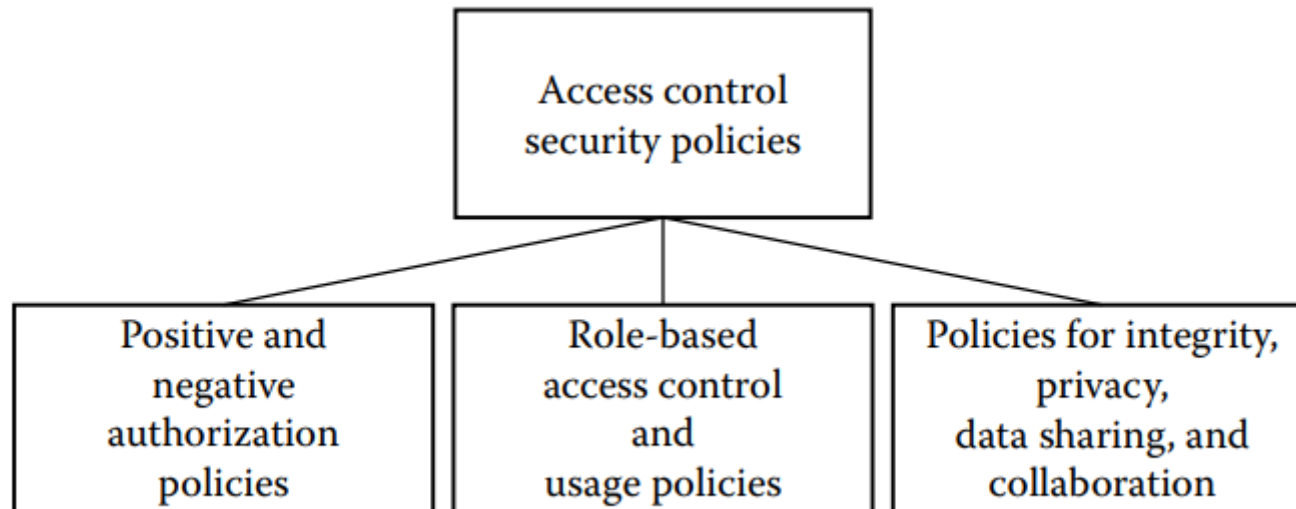
# Data Security, Privacy and Data Mining Techniques

- SECURITY POLICIES



# Access Control Policies

- first examined for operating systems.
- Access could be read access or write access.
  - Write access could include access to modify, append, or delete.



# Authorization-Based Access Control Policies

- **Positive authorization:** user John is granted access to relation EMP
- **Negative authorization:** John does not have access to relation EMP
- **Conflict resolution:**
  - a rule that grants John read access to relation EMP
  - a rule that does not grant John read access to the salary attribute in EMP
  - a system enforces the least privilege rule
    - John has access to EMP except for the salary values.
- **Strong and weak authorization:**
  - Example, if John is granted access to EMP - strong authorization rule
  - where John is not granted access to salary attribute- a weak authorization rule

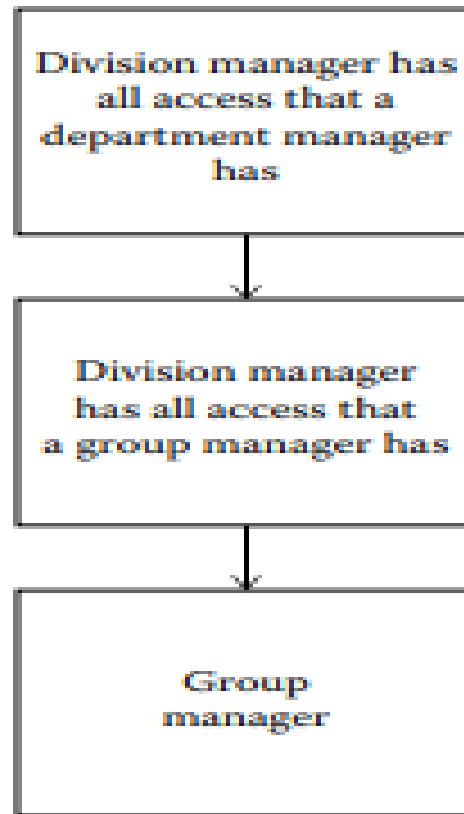
# Access Control Policies [Cont'd]

- **Propagation of authorization rules:**
  - John has read access to relation EMP, then does it automatically mean that John has read access to every element in EMP
- **Special rules**
  - Content- and context-based rule
  - event-based constraints,
  - Example
    - Content: John has read access to tuples only in DEPT D100.
    - Context : John does not have read access to names and salaries taken together;
    - Event : after the election, John has access to all elements in relation EMP.
- **Consistency and completeness of rules**
  - Big challenges to ensure

# Role-Based Access Control[RBAC]

- a very popular access control policy
  - now implemented in commercial systems
    - commercial systems, including Trusted Oracle
- The idea
  - to grant access to users **depending on their roles and functions.**
  - a type of authorization policy that depends on the user role and the activities that go with the role.

# The role hierarchy

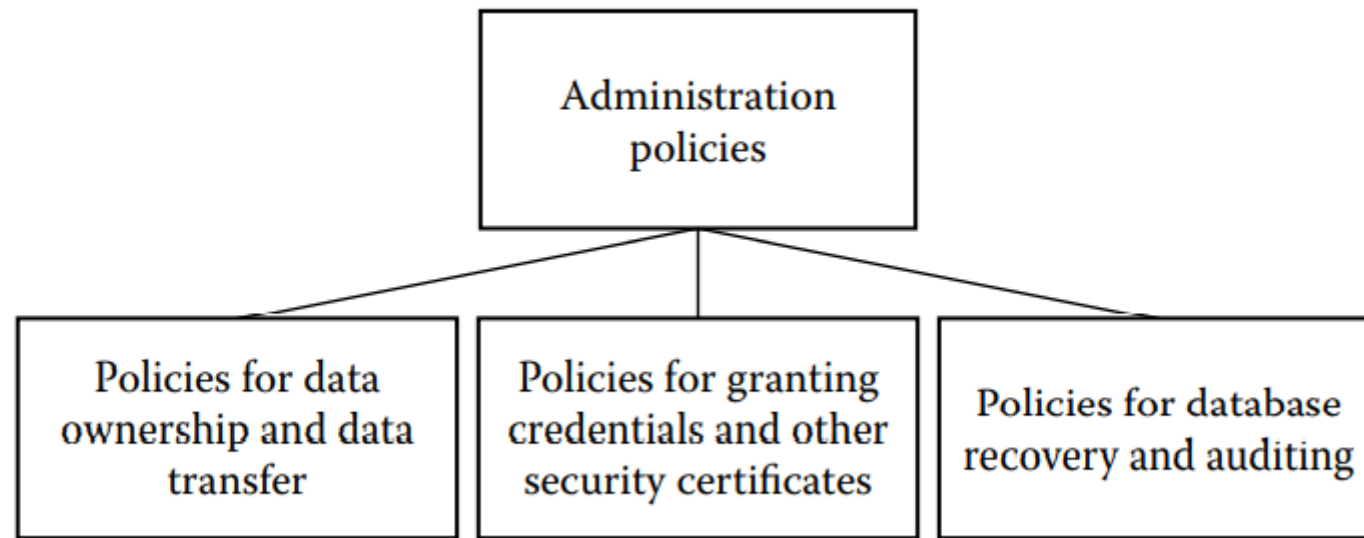


# Administration Policies

- Administration duties
  - keeping the data current,
  - making sure the metadata is updated
  - whenever the data is updated,
  - ensuring recovery from failures and related activities
- Database Administrator (DBA)-the data-related issues
- System Security Officer (SSO)-security-related issues
- get more complicated in distributed environments
  - web environment-multiple parties involved in distributing documents,
    - including the owner, the publisher, and the users requesting the data



# Administration policies



# Identification and Authentication

- Identification
  - users must identify themselves with their user ID and password.
- Authentication
  - the system must then match the user ID with the password to ensure that this is indeed the person he is purporting to be
- Hackers can break into the system and get the passwords of users
- Solution to password-based scheme:
  - Biometrics techniques
  - Face recognition
  - voice recognition techniques

# Auditing a Database System

- Databases are audited for multiple purposes
  - To keep track of the number of queries posed,
  - The number of updates made,
  - The number of transactions executed,
  - The number of times the secondary storage is accessed so that the system can be designed more efficiently.
- **Audited for security purposes**
  - Have any of the access control rules been bypassed by releasing information to the users?
  - Has the inference problem occurred?
    - Way to infer or derive sensitive data from non-sensitive data
  - Has privacy been violated?
  - Have there been unauthorized intrusion

- database may be mined
- to detect any abnormal patterns or behaviors
- credit card fraud and identity theft.

- Views can be used as security
  - Letting users access data through the view, without granting the users permissions to directly access the underlying base tables of the view
  - **Can limit access to only selected columns of the base table.**
  - Can provide value-based security for the information in a table
  - WHERE clause
- **Two types** of database views
  - dynamic views and static views
  - Dynamic views can contain data from one or two tables and automatically include all of the columns from the specified table or tables.

- **Design Flexibility**
- Improved Security

# Views for Security

the DBA could form views and grant access to the views

EMP			
SS#	Ename	Salary	D#
1	John	20K	10
2	Paul	30K	20
3	Mary	40K	20
4	Jane	20K	20
5	Bill	20K	10
6	Larry	20K	10
1	Michelle	30K	20

Rules:

John has read access to V1.

John has write access to V2.

V1. View EMP (D# = 20)		
SS#	Ename	Salary
2	Paul	30K
3	Mary	40K
4	Jane	20K
1	Michelle	30K

V2. View EMP (D# = 10)		
SS#	Ename	Salary
1	John	20K
5	Bill	20K
6	Larry	20K

# Policy Enforcement and Related Issues

- SQL: data definition and data manipulation for relational system
- SQL has **GRANT** and **REVOKE** constructs for specifying grant and revoke access to users
- Example 1:
  - Consider the situation where John does not have read access to names and salaries in EMP taken together, but he can read names and salaries separately. One could specify this in SQL-like language as follows


```
GRANT JOHN READ
EMP.SALARY
GRANT JOHN READ
EMP.NAME
NOT GRANT JOHN READ
Together (EMP.NAME, EMP.SALARY) .
```



- Example 2:
- to grant John read access to the employees who earn less than 30K

```
GRANT JOHN READ  
EMP  
Where EMP.SALARY < 30K
```

# Query Modification

- to modify the query based on the constraints.
- Ex:
  - Consider a query by John to retrieve all tuples from EMP. Suppose that John only has read access to all the tuples where the salary is less than 30K and the employee is not in the Security department. [The attributes of EMP are, say, name, salary, age, and department]
  - Then, the query is
  - Select \* from EM  Select \* from EMP Where salary <30k and Dept is not security

# Discretionary Security and Database Functions

## **Secure database functions:**

Query processing: enforce access control rules during query processing; inference control; consider security constraints for query optimization.

Transaction management: check whether security constraints are satisfied during transaction execution.

Storage management: develop special access methods and index strategies that take into consideration the security constraints.

Metadata management: enforce access control on metadata. Ensure that data is not released to unauthorized individuals by releasing the metadata.

Integrity management: ensure that integrity of the data is maintained while enforcing security.

# DATA PRIVACY

- protecting sensitive information of individuals
  - what information is to be released about him or her

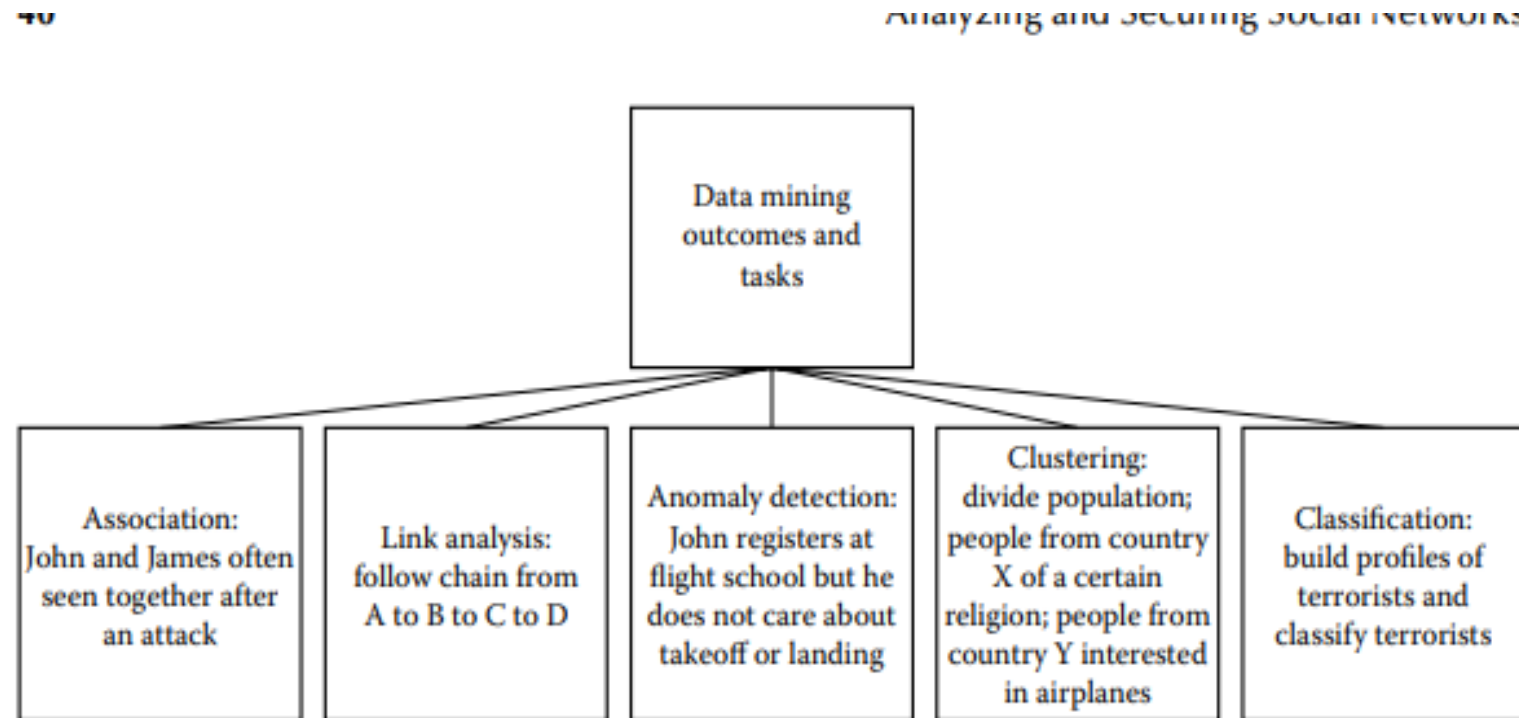
# Data mining techniques

R Devika

APII/CSE/SOC

- predictive and descriptive.
- Predictive tasks
  - essentially predict whether an item belongs to a class or not.
  - most prominent predictive tasks is classification
  - Markov model, decision trees, artificial neural networks (ANNs), support vector machines (SVMs), association rule mining (ARM),
- Descriptive tasks
  - in general extract patterns from the examples
  - include making associations and forming clusters.

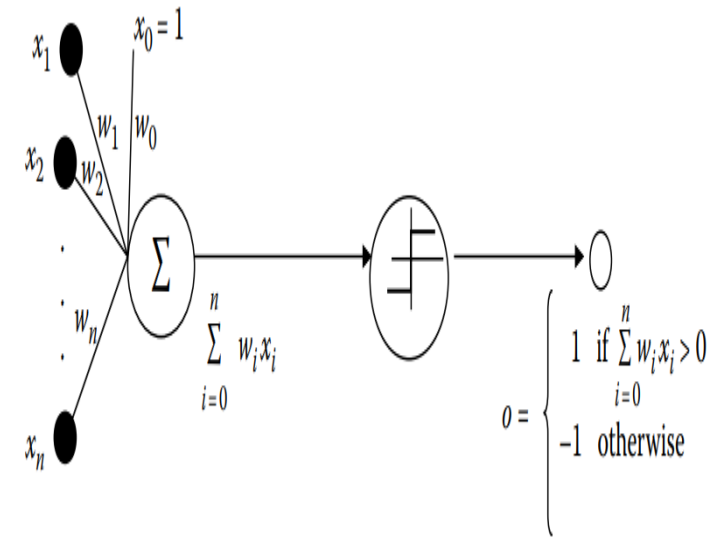
# OVERVIEW OF DATA MINING TASKS AND TECHNIQUE



**FIGURE 4.1** Data mining tasks.

# ARTIFICIAL NEURAL NETWORKS

- robust classification technique
- real-valued, discrete-valued, and vector-valued functions
- learning process of ANN involves adjustments to the node weights
- used in many areas
  - interpreting visual scenes,
  - speech recognition,
  - and learning robot control strategies.
- a simple neuron unit
  - a perceptr
- forecasting the hotel occupancy rate base on
  - tweets

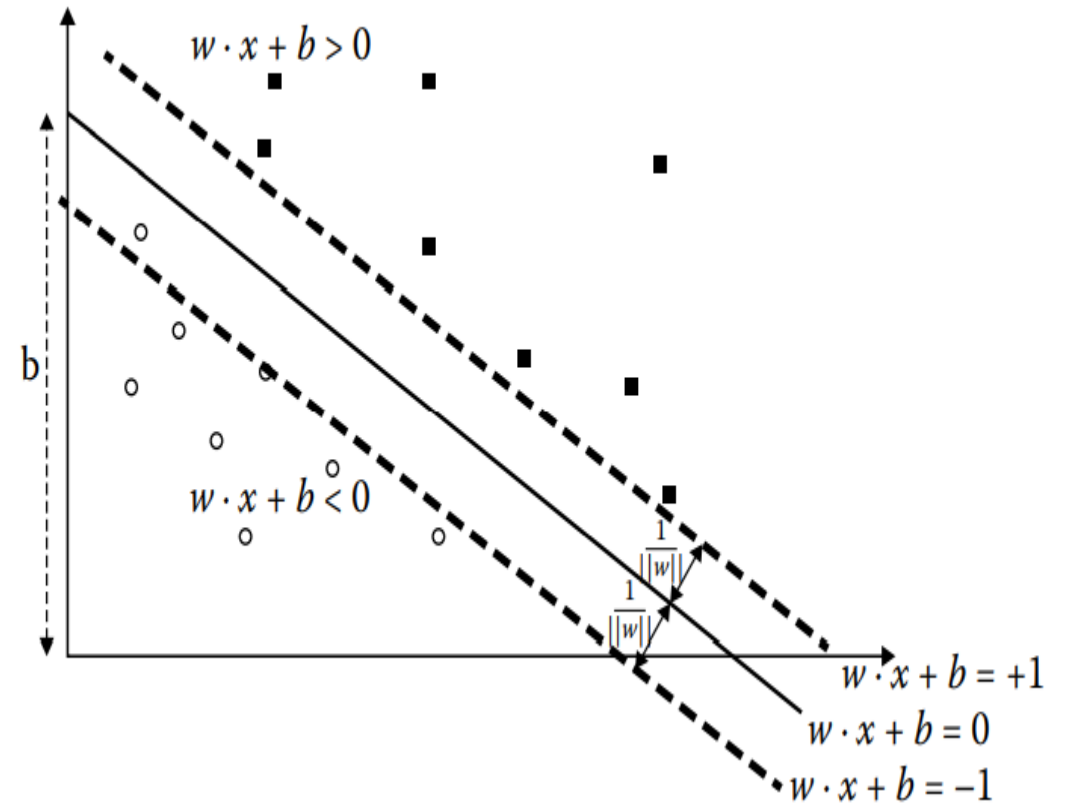




- The information obtained by analyzing social networks could be especially valuable for many applications
  - online advertisement targeting
  - personalized recommendation ,
  - viral marketing ,
  - social healthcare ,
  - social influence analysis
  - and academic network analysis

# SUPPORT VECTOR MACHINE

- Hyperplane classifier
- linear separability
  - margin maximization
  - kernels
  - N training data points
  - higher speed and better performance



# MARKOV MODEL

- predictive models
  - it predicts the next action depending on the result of previous action or actions
- Markov chains
  - the sequences of web pages visited by surfers-fed as input
- zeroth-order Markov model
  - the unconditional probability of the state
- first-order Markov model- $\{P_1, P_2\}, \{P_2, P_3\}, \dots, \{P_{k-1}, P_k\}$ .
  - page-to-page transitional probabilities
  - the next action is predicted based on only the last action performed
- Kth-order Markov model
  - probability of user visit to  $P_k$ th page given its previous  $k - 1$  page visits

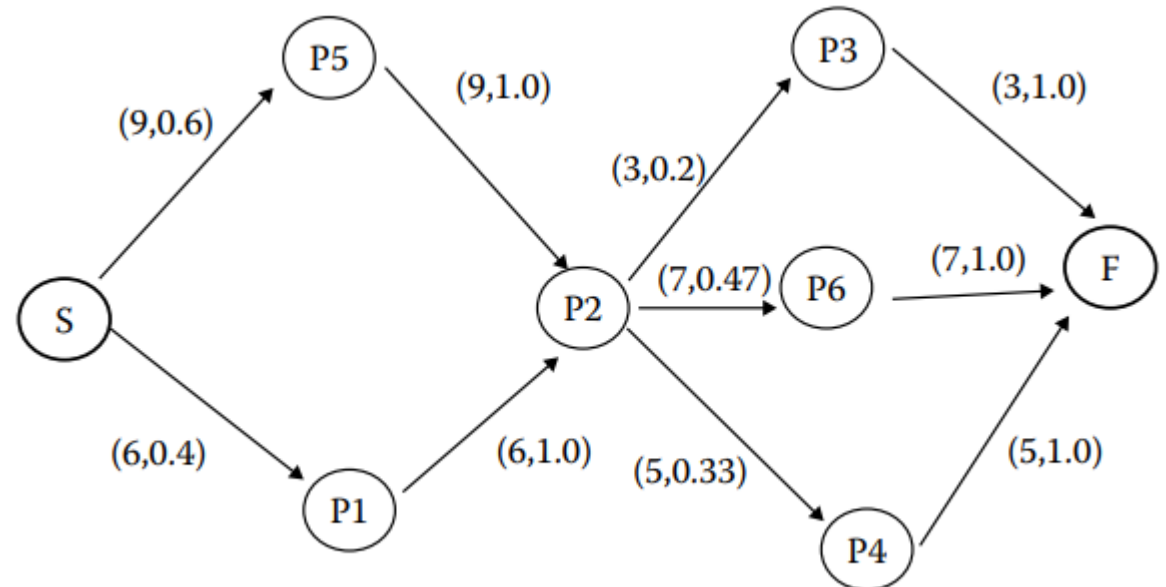
# Illustrative example of different orders of Markov model

- six web pages, P1
- , P2, P3, P4, P5, and P6
- States S and F correspond to the initial and final states
- first number is the frequency of that transition
- second number is the transition probability

**TABLE 4.1**

**Collection of User Sessions and Their Frequencies**

Session	Frequency
P1,P2,P4	5
P1,P2,P6	1
P5,P2,P6	6
P5,P2,P3	3



- For example,
- the transition probability of the transition (P2 to P3) is 0.2
- the number of times users traverse from page 2 to page 3 is 3,
- the number of times page 2 is visited is 15
- (i.e.,  $0.2 = 3/15$ )

# ASSOCIATION RULE MINING (ARM)

- finds the relationships among item sets based on their co-occurrence in the transactions
- discovers the frequent patterns (regularities) among those items sets.
- Example
  - what are the items purchased together in a superstore?
- Social network analysis
  - an analysis of Twitter data might reveal that users who tweet about a particular topic are also likely to tweet about other related topics, which could inform the identification of groups or communities within the network

- **Apriori algorithm**
- **FP-Growth algorithm**
- **ECLAT algorithm**
- **Metrics for Evaluating Association Rules**

$$\text{Support}(\{X\} \rightarrow \{Y\}) = \frac{\text{Transactions containing both } X \text{ and } Y}{\text{Total number of transactions}}$$

$$\text{Confidence}(\{X\} \rightarrow \{Y\}) = \frac{\text{Transactions containing both } X \text{ and } Y}{\text{Transactions containing } X}$$

$$\text{Lift}(\{X\} \rightarrow \{Y\}) = \frac{(\text{Transactions containing both } X \text{ and } Y) / (\text{Transactions containing } X)}{\text{Fraction of transactions containing } Y}$$

# Implementing Association Rules in Python

- PyCaret



# Image mining

R DEVIKA

AP II /CSE/SOC

# Image mining

- huge amounts of digital images are generated and saved every day
- to find images of specific entities from a database
- to find an image pattern
  - helpful to understand the relationships **between high-level semantic concepts**/descriptions and **low-level visual features**.
- Applications of digital image
  - commercial and news media photo libraries
  - scientific and nonphotographic image databases,
  - and medical image databases.
- content-based image retrieval (CBIR)
  - retrieving relevant images
  - computes relevance
    - similarity of visual content
      - image features
        - color histograms, textures, shapes, and spatial layout.
    - visual similarity is not semantic similarity
- semantic gap is the major problem- (CBIR)
  - a “red ball” image of a “red rose.

# Annotation of images with keywords

- a keyword-based query interface
- way to publish an image data repository
- Image annotation
  - practice of assigning labels to an image
- major problem
  - most of images are not annotated.
    - Laborious AND error-prone
    - subjective process to manually annotate a large collection of images

# Image mining and analytics

- important for social media as the members post numerous images.
- Database
  - WELL DEFINED-structured data-number
  - Ill defined-unstructured data-
    - Example
      - images, audio, and video are data with ill-defined semantics.
- Feature Selection
  - images are represented by derived data or features
    - color, texture, and shape
  - Many of these features have multivalues
    - color histogram and moment description
  - Dimensionality of derived image data is usually very high.
    - generate as many features as possible
    - not aware which feature is more relevant.
    - noise.
      - irrelevant or duplicated information(feature)
      - curse of dimensionality.
- Research topic
  - optimal subset of features.
    - A feature selection approach to find optimal feature subsets for the network intrusion detection system
    - An effective machine learning based technique for cardiac disease prediction with optimal feature subset selection
    - Optimal Features Selection for Designing a Fault Diagnosis System

- *pre-process the images*
- *extract the features*
- *cluster the images* on similarity,
- and *evaluate for the optimal number of clusters*

- **Pre-processing of images.**

- *comparable* in *color*, *value range*, and *image size*.

**1.colorscales:** Conversion of the image into e.g. grayscale (2-D) or color (3-D).

**2.scale:** Normalize all pixel values between the minimum and maximum range of [0, 255].

**3.dim:** Resize each image to make sure that the number of features is the same

- **Extraction of image features.**
  - using the pixel value information
  - many approaches to extract features
    - **Principal component analysis (PCA)**
      - reduce dimensionality and extract Principal Components
    - **Histogram of Oriented Gradients (HOG)**
      - to the direction and orientation of edges from image data

# Automatic Image Annotation

- object recognition
  - trying to recognize objects in an image
  - generate descriptions for the image according to semantics of the objects.
- To produce accurate and complete semantic descriptions for an image
  - store descriptions in an image database
- Singular value decomposition or principal component analysis
  - Reduce the more number of feature into less feature
    - But fail to consider feature selection/ feature weight.
- **Image Annotation Tool using OpenCV**



# Image Classification

- important area- medical domain
- diagnostic aid in a real-world clinical setting
- **Convolutional Neural Networks (CNNs)**
- **kNN k-nearest neighbour**

## Image Classification-steps <sup>[1]</sup>

- **Digital Data:-** An image is captured by using *digital camera* or any *mobile phone camera*.



## Image Classification-steps <sup>[2]</sup>

- **Pre-processing:-** *Improvement* of the image data.



*Normalized image*



*Gray-Scale image*



*Resize image*



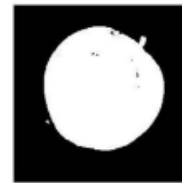
*Noise removal*



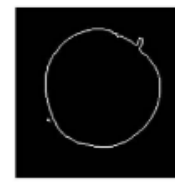
*Contrast enhancement*



*Binary image*



*Complemented Binary image*



*Boundary image*

## Image Classification-steps <sup>[2]</sup>

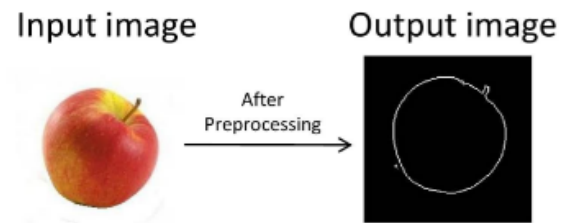
- **Feature Extraction:-** The process of *measuring* or *calculating* or *detecting* the features from the image samples.

The most common two types of feature extraction are:-

- ✓ Geometric feature extraction
- ✓ Color feature extraction

## Image Classification-steps <sup>[1]</sup>

- **Selection of training data:-** Selection of the *particular attribute* which best describes the pattern.









### Train Data Set

Name	Major	Minor	Area	Perimeter	Red #	Yellow #
'AP.jpg'	[175.4774]	[169.6791]	[23328]	[604.9848]	[1267]	[621]

# Image Classification-steps <sup>[1]</sup>

- **Decision and Classification:-** Categorizes detected objects into predefined classes by using suitable method that compares the image patterns with the target patterns.
- **Classification Output:-**

Sampled Image						
Area or Size	1200000	900000	800000	1100000	600000	700000
Colour Intensity	80	90	91	82	100	95
Grading	Grade A	Grade B	Grade B	Grade A	Grade C	Grade C

Cloud computing

# Cloud computing

- service-oriented computing
  - services are being outsourced
  - handle massive amounts of data.
  - delivers computing as a service
    - hardware services, systems services, data services, and storage services.
- Google
  - MapReduce framework
  - Apache's Hadoop Distributed File System
- characterized by massive scalability and new Internet-driven economics
  - stock market data
  - weather and related data.

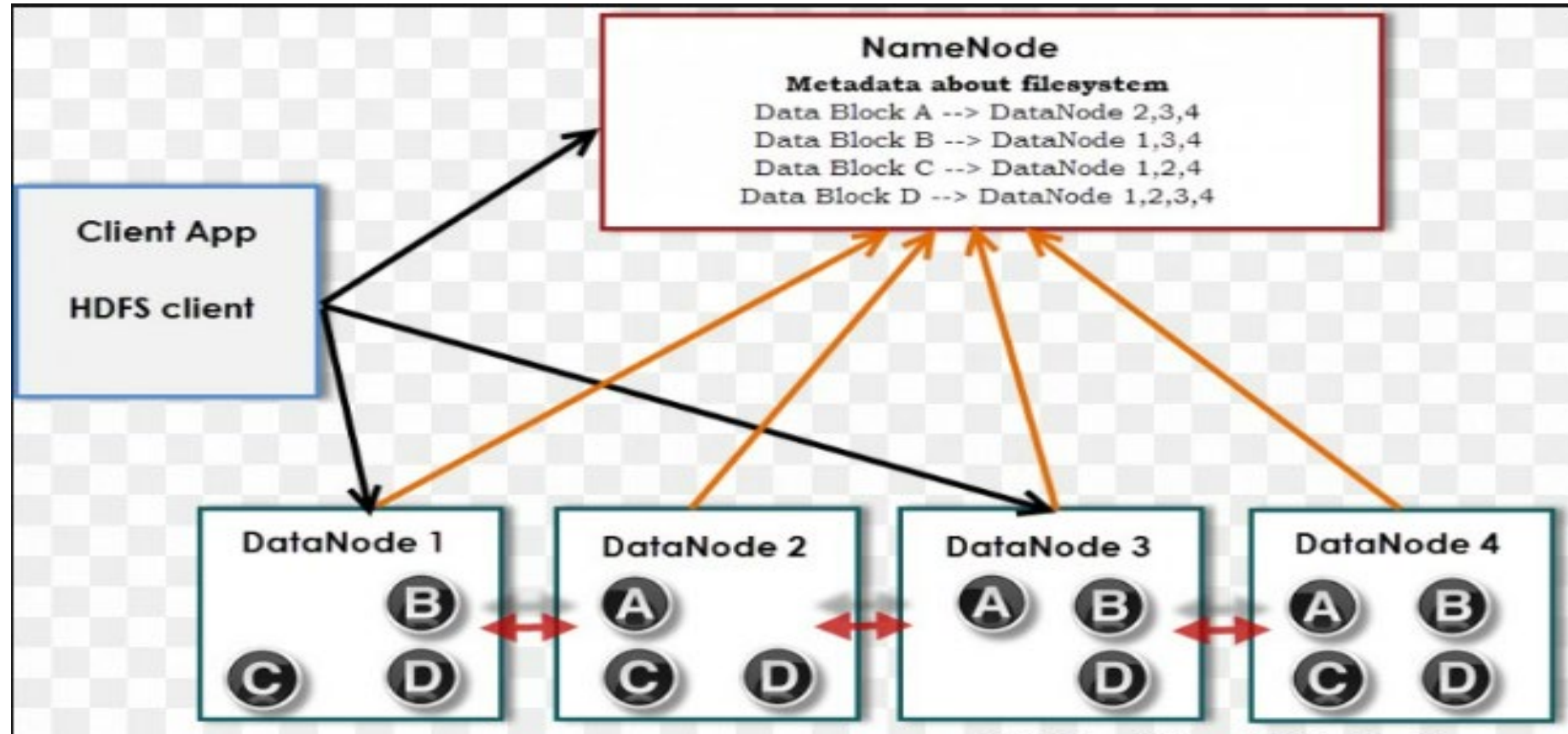


- HDFS to store all of the customer data in a single, centralized location
- data can then be processed in parallel across multiple nodes in the HDFS cluster
- allowing the company to quickly and efficiently analyze the data
- HDFS is a highly scalable and robust file system that is well-suited for big data.

- For example, consider a weather data company that collects and stores data from thousands of weather stations around the world. The data is in the form of temperature readings, precipitation levels, wind speed, etc. collected every hour. The company wants to analyze this data to understand climate patterns and predict future weather patterns.
- In this case, the company can use HDFS to store all the weather data in a single, centralized location. The data can be divided into blocks and distributed across multiple nodes in the HDFS cluster, allowing for parallel processing. The company can then use a tool such as Apache Hive or Apache Pig to analyze the data and extract insights. The parallel processing capabilities of HDFS help to speed up the analysis process and allow the company to process the large amounts of data in a timely manner.

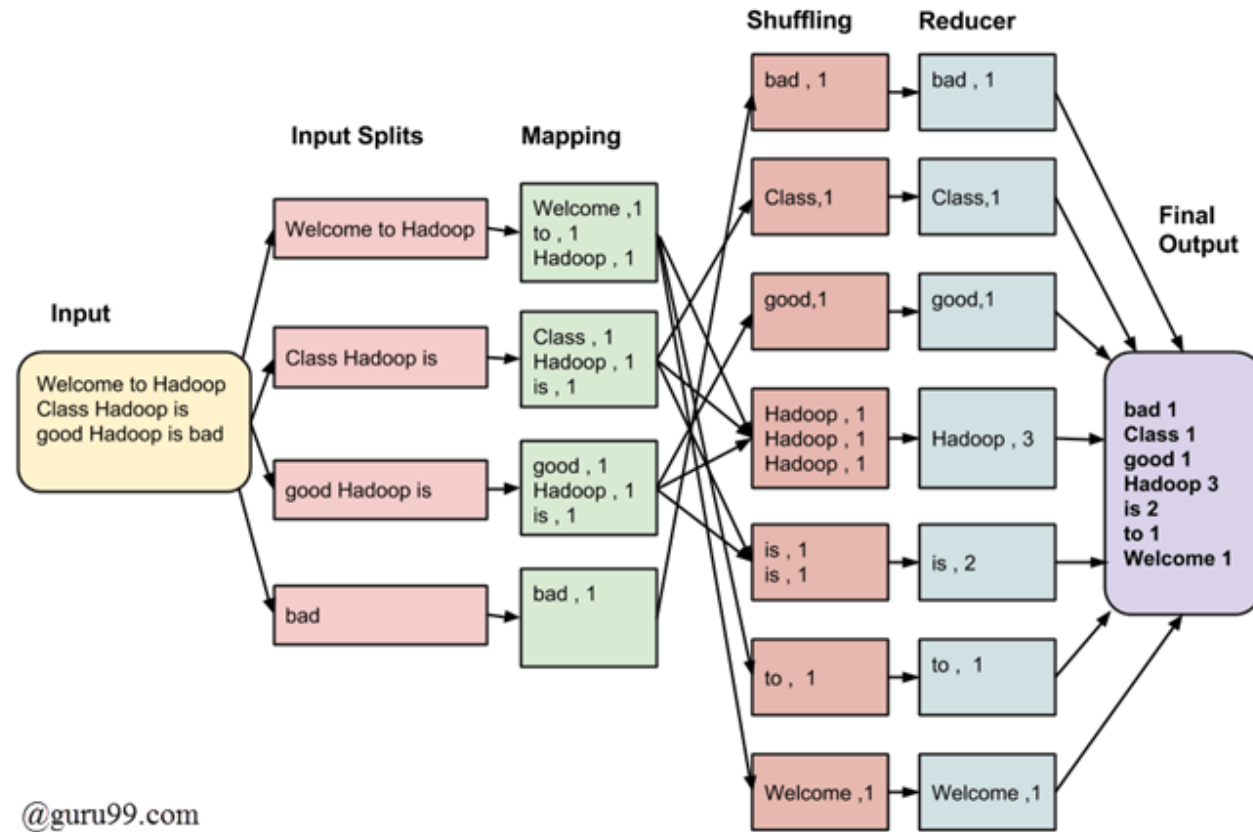
# to find trends and patterns in customer behavior.

- For example, consider a scenario where a company has multiple branches and generates huge amounts of sales data every day. This data needs to be processed, aggregated and stored for further analysis. Using a traditional file system to store and process this data would become slow and unwieldy as the data grows.
- In this scenario, HDFS can be used to store and process the sales data. The data can be stored as large files in HDFS and can be divided into smaller blocks and distributed across multiple nodes in a cluster. The parallel processing capability of HDFS allows the company to quickly process and aggregate the sales data. The data is also replicated across multiple nodes, providing fault tolerance and ensuring that the data is always available, even in the case of node failure.



# map reduce with example

- For example, consider a scenario where a company wants to count the number of occurrences of each word in a large collection of documents. Using MapReduce, the task can be divided into two phases: the Map phase and the Reduce phase
- For example, consider a scenario where a company wants to analyze the clickstream data of its website to determine the most popular products.



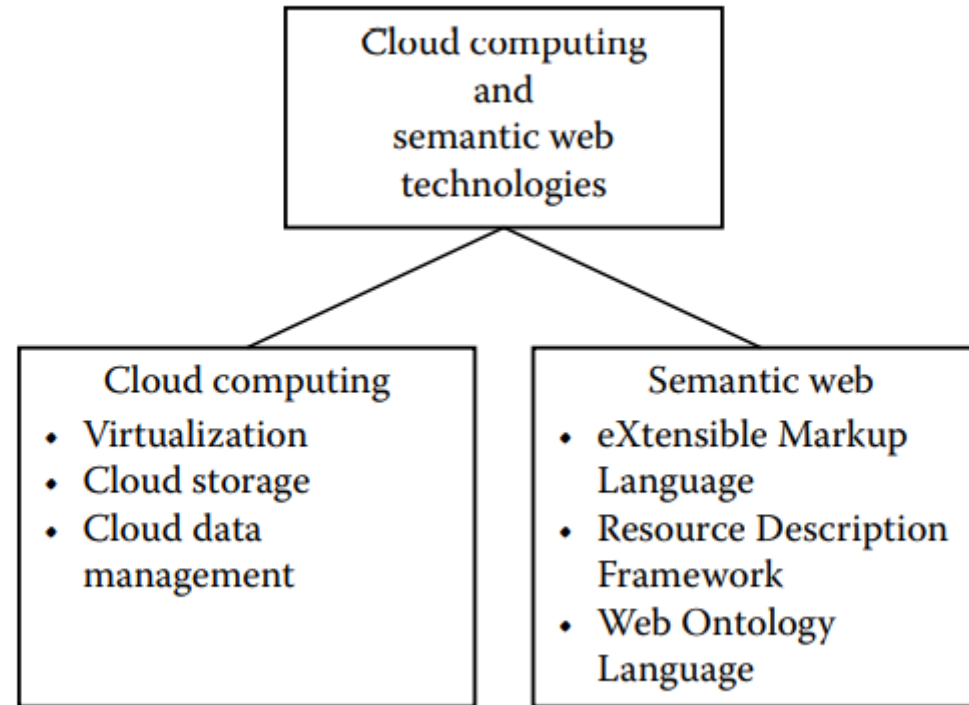
- Suppose you have a large dataset of log files from a website and you want to count the number of times each unique IP address appears in the logs
- In the Map stage, the map function would take each log file as input and extract the IP address from each record. It would then emit a key-value pair where the key is the IP address and the value is 1.
- In the Reduce stage, the reduce function would receive a list of values for each IP address and add them together to get the total count. The output would be a set of key-value pairs where each key is an IP address and the corresponding value is the total count of that IP address in the logs.

# Difference between MapReduce and HDFS

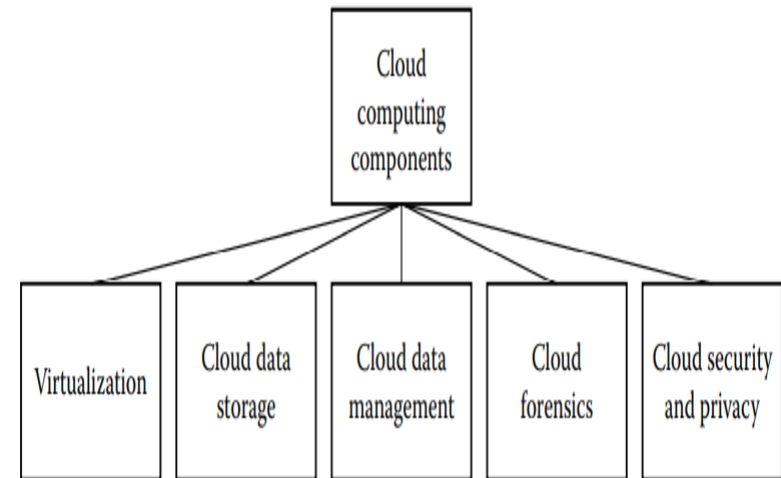
- In summary, HDFS provides the storage layer for Hadoop, while MapReduce provides the processing layer for Hadoop. HDFS stores the data, and MapReduce processes the data stored in HDFS. They work together to provide a complete solution for processing and storing large amounts of data.



# Cloud computing and semantic web technologies



- the delivery of computing resources and services over the internet
- allowing users to access data and applications from anywhere with an internet connection.
- operates on centralized data centers
- don't require real-time data processing in data transmission.



- cloud Deployment Models

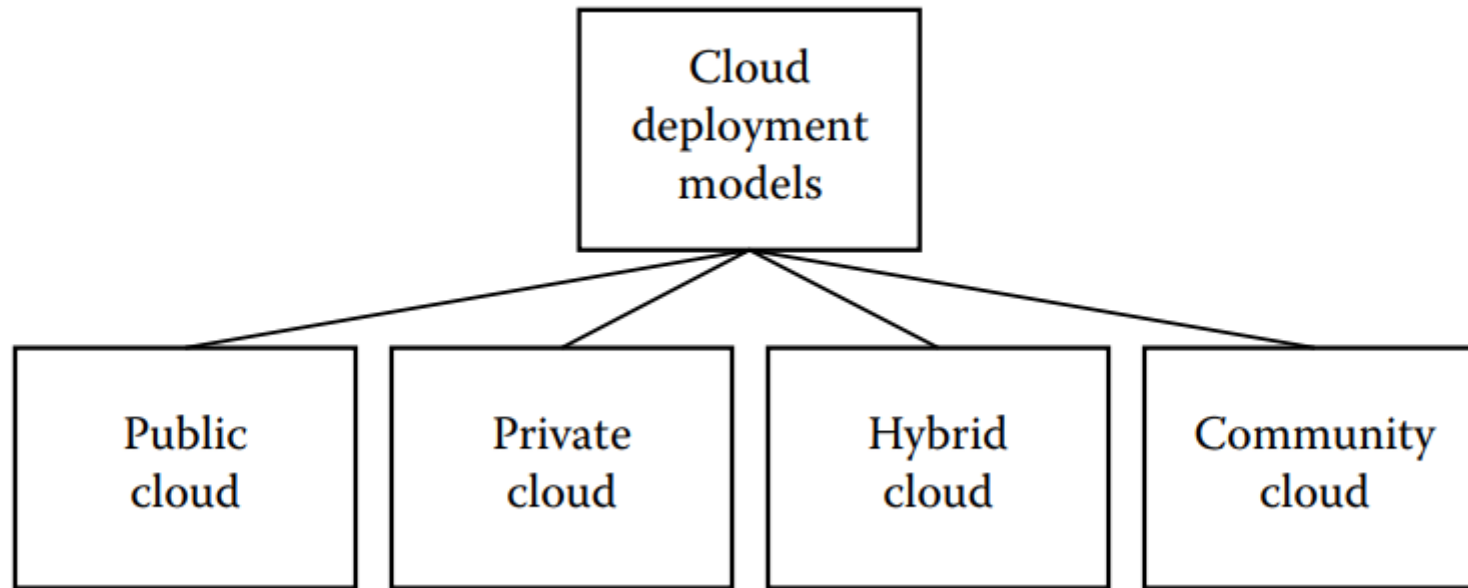
- public cloud-offered on a pay-per-use basis, where clients only pay for what they use.
  - Examples of public cloud providers are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)
- community cloud
- hybrid cloud
- private cloud

- service Models

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)-This includes components such as operating systems, databases, web servers, middleware, and other tools necessary for application development and deployment.
- Software as a Service (SaaS)
- Data as a Service (DaaS).

- (IaaS)
  - a collection of hardware and networks for use by the general public or organizations
- (PaaS)
  - systems software such as operating systems (OSs) and execution environments.
- DaaS
  - provides data to the cloud users

# cloud deployment models



# Virtualization

- creating something virtual and not actual
  - hardware, software, memory, or data.
- Hardware virtualization
  - virtual machine monitor, also known as the hypervisor
  - VMware and XEN
- types of virtualization
  - OS level virtualization
  - storage virtualization
  - data virtualization
  - database virtualization
- also incorporated into embedded systems and mobile phones

- VMware ESXi: A type 1 hypervisor that provides enterprise-class virtualization for servers.
- Microsoft Hyper-V: A type 1 hypervisor that is included with Windows Server and provides virtualization capabilities for both Windows and Linux virtual machines.
- Oracle VirtualBox: A type 2 hypervisor that is widely used for desktop virtualization and development.
- KVM (Kernel-based Virtual Machine): An open-source type 1 hypervisor that is integrated into the Linux kernel and supports virtualization for both Linux and Windows virtual machines.
- Xen: An open-source type 1 hypervisor that is widely used for cloud computing and supports a wide range of operating systems, including Linux, Windows, and Solaris.

# OS level virtualization

- to run multiple instances of Linux on a single physical server.
- Docker and LXC (Linux Containers) are two popular technologies for OS-level virtualization.
- to create, deploy, and run applications inside containers



# storage virtualization

- physical storage devices from multiple sources are aggregated into a virtual storage pool. This pool can then be managed as a single entity, and the virtualized storage resources can be allocated, managed, and reconfigured dynamically
- Some benefits of storage virtualization include improved storage utilization, better data management, increased data availability and disaster recovery, and easier scalability of storage resources.
- Storage Area Networks (SANs), Network Attached Storage (NAS) systems, and software-defined storage (SDS)

# data virtualization

- Virtual databases, such as those provided by vendors like Informatica, Denodo, and Red Hat.
- the creation of a virtualized view of data
- goal of data virtualization is to provide a unified access to diverse data sources, improving data integration and agility while reducing data duplication and the need for data movement.
- With data virtualization, the company can access and combine this data into a single, virtualized data layer, without having to physically move or copy the data into a central repository.

# DATABASE VIRTUALIZATION

- Oracle Virtual Private Database (VPD)
- Microsoft SQL Server database mirroring
- IBM DB2 pureScale
- MySQL Cluster
- PostgreSQL Postgres-XC.

# Cloud Storage and Data Management

- storage managers
- A single object (e.g., the entire video database of a customer) may be stored in multiple locations.
- Each location may store objects for multiple customers.
- Advantages
  - need not purchase expensive storage devices.
  - Data could be placed anywhere
  - Maintenance such as backup and recovery
    - serious security concerns with respect to storing data

a cloud database manager.

- a Virtual Machine Image must be purchased.
- example is the Amazon Relational Database Service (<http://aws.amazon.com/rds/>)

# semantic web

- collection of technologies
- to produce machine-understandable web pages.
- to enhance the current Web with a layer of meaning, allowing machines to understand the content and context of the data on the web



# SEMANTIC WEB

R Devika

AP II/CSE/SOC



# SEMANTIC WEB

- current web technologies
  - integration of information from a syntactic point of view
  - a lot to be done to handle the different semantics of various systems and applications
  - “human-in-the-loop”
- to make the web more intelligent
  - to integrate disparate information sources
  - alleviate humans from the burden of having
- One **needs machine-understandable web pages** and the use of **ontologies for information integration**

# SEMANTIC WEB (cont'd)

- Highly intelligent and sophisticate
  - needs little or no human intervention to carry out tasks
    - scheduling appointments
    - coordinating activities
    - searching for complex documents
- Languages are contributing a lot toward developing the semantic web

- ontology matching
- intelligent agents
- markup languages

## Examples of Intelligent Agents

- Assistant agent in MS Office



- Computer viruses



- Trading agents



- Characters in computer games



- Web spiders



# Illustrates the layered technology stack for the semantic web

Trust
SWRL
OWL
RDF
XML
Foundations

# XML

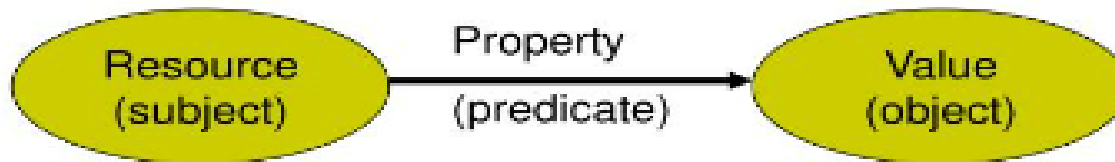
- is an extensible markup language specified
- needed due to the limitations of HTML (Hypertext Markup Language) and complexities of SGML (Standard Generalized Markup Language)
- designed to make the interchange of structured documents over the Internet easier
  - used for exchanging data between applications
  - designed for data exchange
- used to be Document Type Definitions (DTDs)
  - define the role of each element of text in a formal model
- limited support for linking and reasoning about data.
- flexible format for exchanging structured data

# RDF-Resource Description Framework

- designed for representing and linking data on the Web
- a standardized way of describing resources and their relationships
- data is highly interconnected and can be processed and understood by machines
- for representing and linking data on the Web
- **three pairwise disjoint infinite sets of terms**
  - set U of URI references, the set L of literals the set B of blanks
  - The set U ∪ L of names is called the vocabulary
  - ordered triple (s, p, o) as the subject, predicate, and object of a triple  $s \xrightarrow{p} o$ .
  - subject–predicate–object expressions (e.g., "Bob is 35", or "Bob knows John").

# RDF (cont'd)

- Triple
  - A Resource (Subject) is anything that can have a URI: URIs or blank nodes
  - A Property (Predicate) is one of the features of the Resource: URIs
  - A Property value (Object) is the value of a Property, which can be literal or another resource: URIs, literal, blank nodes

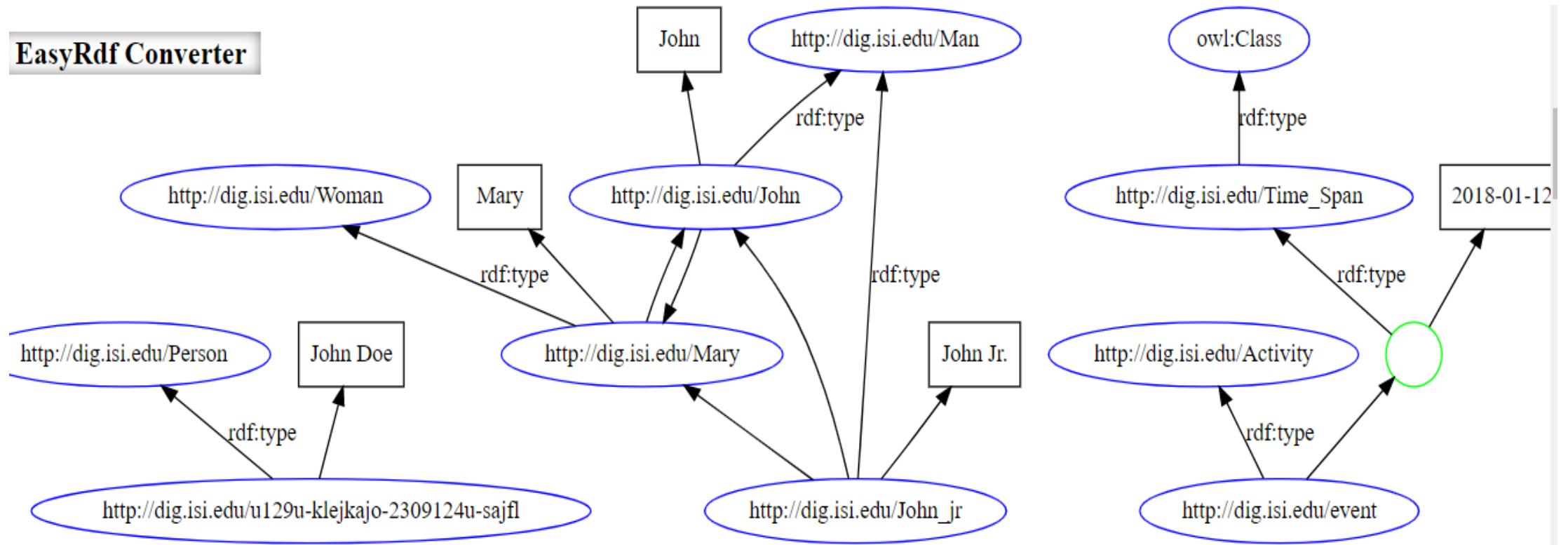


Literals can be the object of an RDF statement, but cannot be the subject or the predicate

Subject	Predicate	Object
-----		
<JohnDoe>	<hasName>	"John Doe"
<JohnDoe>	<hasAge>	"30"
<JohnDoe>	<hasOccupation>	"Software Engineer"
<JohnDoe>	<livesIn>	<NewYork>
<NewYork>	<hasName>	"New York"
<NewYork>	<isLocatedIn>	<USA>
<USA>	<hasName>	"United States of America"

# RDF Data Visualization

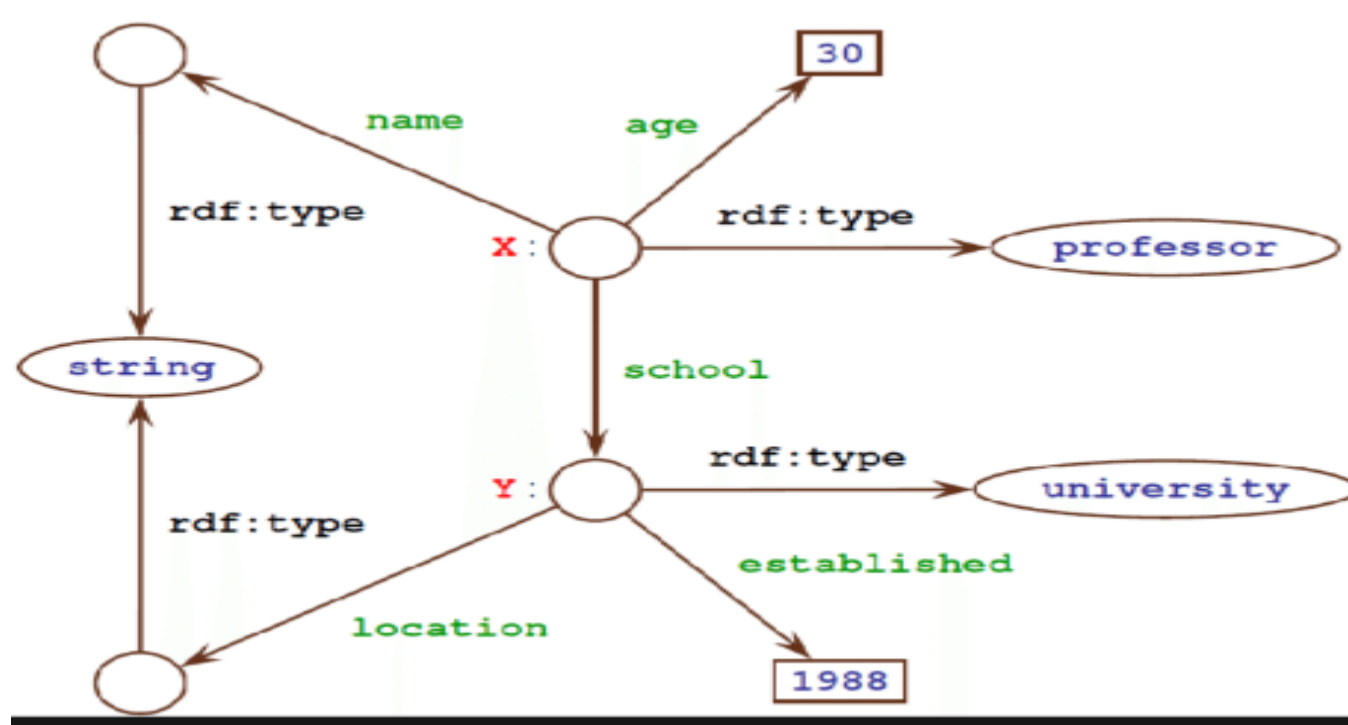
## EasyRdf Converter



suggested [rEd](#) and [Vizio](#) but these require manual drawing. You can also convert from [rEd](#) to [PDF](#) and [SkE](#). Someone mentioned [Griff](#) but that requires installation of [AlloraGraph](#) (transcript



# Example for RDF



# Xml

```
<person>
  <name>John Doe</name>
  <age>35</age>
  <address>
    <street>123 Main St</street>
    <city>Anytown</city>
    <state>CA</state>
    <zip>12345</zip>
  </address>
</person>
```

# Equivalent representation of the same information in RDF

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.org/people/JohnDoe>
  foaf:name "John Doe" ;
  foaf:age "35"^^xsd:integer ;
  foaf:based_near [
    foaf:name "Anytown, CA 12345" ;
    foaf:address [
      foaf:street "123 Main St" ;
      foaf:locality "Anytown" ;
      foaf:region "CA" ;
      foaf:postalCode "12345" ;
    ]
  ] .
```

# Here is a simple XML example for social network analysis

[person example](#)[xml]

- Person [example](#) [rdf]

# OWL -Web Ontology Language

- is built on top of RDF
- a more expressive data model
- more powerful language for modeling
  - to describe the types of resources and relationships that exist in a domain,
- provides a way to reason about these resources and relationships

# OWL has three sublanguages

- OWL Lite
- OWL DL
- and OWL Full
- variety of application
  - knowledge management
  - semantic web
  - linked data

Here's an example in OWL that defines a class for a student and properties for their name, age, and major

- [example](#)

# How you can run an OWL ontology using the Pellet reasoner

- Write your OWL ontology in the OWL syntax and save it as a file with a .owl extension.
- Download and install the Pellet reasoner
- Start the Pellet reasoner
  - `java -jar pellet.jar`
- Load your OWL ontology into the Pellet reasoner
  - `load <path to your owl file>.owl`
- Ask the reasoner to perform reasoning tasks
  - to determine all instances of a particular class,
  - or to find all subclasses of a given class.
- Analyze the results.

# SWRL (Semantic Web Rule Language)

- rule language for the Semantic Web.
- to define rules that can be used to infer new knowledge from existing knowledge represented in RDF and OWL ontologies.
- SWRL rules are written in a syntax
  - production rule systems, such as those used in expert systems
    - computer programs that use a set of rules to make decisions and solve problem
  - Antecedent
    - a set of conditions that must be true
  - a consequent
    - (the result of the rule if the antecedent is true).
    - Query: John has Mary as a parent and Mary has Bill as a brother then John has Bill as an uncle.
    - Ans:  $\text{hasParent}(\text{?x1}, \text{?x2}) \wedge \text{hasBrother}(\text{?x2}, \text{?x3}) \Rightarrow \text{hasUncle}(\text{?x1}, \text{?x3})$
    - decision support systems, data integration systems, and knowledge management systems.



# Example

- [example of a SWRL rule written in OWL](#)

# Semantic Web Security

- Data protection
  - the exchange of data between various systems, which raises concerns regarding data protection and privacy
- Data authenticity
  - can be easily manipulated
    - increases the risk of malicious actors injecting false information into the system
- Data integrity
  - any corruption or alteration of data can affect the accuracy of results.
- Access control:
  - the sharing of data between multiple systems, which requires proper access control mechanisms to ensure that sensitive data is not disclosed to unauthorized entities.

- RDF (Resource Description Framework) is a data model for linked data
  - integrated with cloud technology
    - a scalable and decentralized infrastructure for storing and accessing RDF data
  - integration can be done through various cloud services
  - Cloud storage services
    - Amazon S3 or Google Cloud Storage for storing large RDF datasets
  - Cloud databases
    - Amazon Neptune or Google Cloud Bigtable for querying and managing RDF data
  - Cloud platforms
    - Amazon Web Services or Google Cloud Platform for hosting RDF applications and services

# Advantage of integrating RDF with cloud technology

- ability to store and process large amounts of data
- high availability
- scalability
- accessibility to RDF data