Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

Knowledge Discovery and Data Mining 1

(Data Mining Algorithms 1)

Winter Semester 2022/23

Further Topics

Agenda

Data Privacy

Situation

- Huge volume of data is collected
- From a variety of devices and platforms (e.g. Smartphones, Wearables, Social Networks, Medical systems)
- Capturing human behaviors, locations, routines, activities and affiliations
- Providing an opportunity to perform data analytics

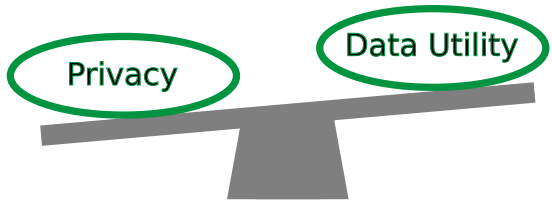
Data Abuse is inevitable

- It compromises individual's privacy
- Or breaches the security of an institution

Data Privacy

- These privacy concerns need to be mitigated
- They have prompted huge research interest to *protect data*
- But,

$$\text{Strong Privacy Protection} \implies \text{Poor Data Utility}$$
$$\text{Good Data Utility} \implies \text{Weak Privacy Protection}$$



Challenge

Find a good trade-off between Data Utility and Privacy

Data Privacy

Objectives of Privacy Preserving Data Mining

- Ensure data privacy
- Maintain a good trade-off between data utility and privacy

Paradigms

- k-Anonymity
- l-Diversity
- Differential Privacy

Linkage Attack

Method

Different public records can be linked to it to breach privacy

Hospital Records

Private	*Public*			
Name	Sex	Age	Zip	Disease
Alice	F	29	52062	Breast Cancer
Janes	F	27	52064	Breast Cancer
Jones	M	21	52066	Lung Cancer
Frank	M	35	52072	Heart Disease
Ben	M	33	52078	Fever
Betty	F	37	52080	Nose Pains

Public Records from Sport Club

Public				
Name	Sex	Age	Zip	Sport
Alice	F	29	52062	Tennis
Theo	M	41	52066	Golf
John	M	24	52062	Soccer
Betty	F	37	52080	Tennis
James	M	34	82066	Soccer

k-Anonymity

k-Anonymity

Privacy paradigm for protecting data records before *publication*

Three kinds of attributes:

1. *Key Attributes*: Uniquely identifiable attributes (e.g., Name, Social Security Number, Telephone Number)
2. *Quasi-identifier*: Groups of attributes that can be combined with external data to uniquely re-identify an individual (e.g. (Date of Birth, Zip Code, Gender))
3. *Sensitive Attributes*: An attacker should not be able to combine these with the key attributes. (e.g. Disease, Salary, Habit, Location etc.)

k-Anonymity

Attention

Hiding key attributes alone does not guarantee privacy.

An attacker may be able to break the privacy by combining the quasi-identifiers from the data with those from publicly available information.

Definition: k-Anonymity

Given a set of quasi-identifiers in a database table, the database table is said to be *k-Anonymous*, if the sequence of records in each quasi-identifier exists at least k times.

Ensure privacy by *Suppression* or *Generalization* of quasi-identifiers.

k-Anonymity: Suppression

Suppression

Accomplished by replacing a part or the entire attribute value by placeholder, e.g. "?" (= generalization)

Example

- Suppress Postal Code: 52062 $\mapsto$ 52???
- Suppress Gender: Male $\mapsto$?; Female $\mapsto$?

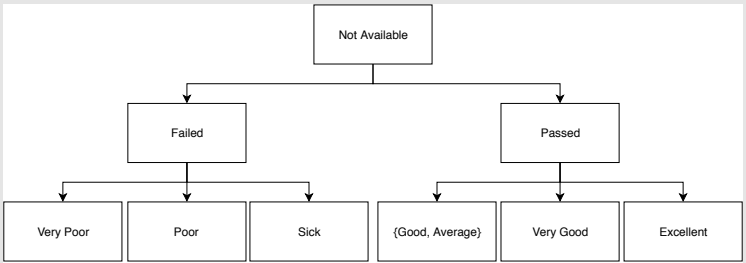
k-Anonymity: Generalization

Generalization

Accomplished by aggregating values from fine levels to coarser resolution using generalisation hierarchy.

Example

Generalize exam grades:



Shortcomings: Background Knowledge Attack

Background Knowledge Attack

Lack of diversity of the sensitive attribute values (homogeneity)

Example

- *Background Knowledge*: Alice is (i) 29 years old and (ii) female
- *Homogeneity*: All 2*-aged females have Breast Cancer.
  $\Longrightarrow$ Alice has BC!

Release

Quasi Identifier			*Sensitive*
Sex	Age	Zip	Disease
F	2?	520??	Breast Cancer
F	2?	520??	Breast Cancer
M	2?	520??	Lung Cancer
M	3?	520??	Heart Disease
M	3?	520??	Fever
F	3?	520??	Nose Pains

This led to the creation of a new privacy model called *l*-diversity

l-Diversity

Distinct l-Diversity

An quasi-identifier is l-diverse, if there are at least l different values. A dataset is l-diverse, if all QIs are l-diverse.

Example

Not "diverse"

Quasi Identifier	Sensitive
QI 1	Headache
QI 1	Headache
QI 1	Headache
QI 2	Cancer
QI 2	Cancer

2-diverse

Quasi Identifier	Sensitive
QI 1	Headache
QI 1	Cancer
QI 1	Headache
QI 2	Headache
QI 2	Cancer

l-Diversity

Other Variants

- *Entropy l-Diversity*: For each equivalent class, the entropy of the distribution of its sensitive values must be at least l
- *Probabilistic l-Diversity*: The most frequent sensitive value of an equivalent class must be at most $1/l$

Limitations

- Not necessary at times
- Difficult to achieve: For large record size, many equivalent classes will be needed to satisfy l-Diversity
- Does not consider the distribution of sensitive attributes

Background Attack Assumption

- k-Anonymity and l-Diversity make assumptions about the adversary
- They at times fall short of their goal to prevent data disclosure
- There is another privacy paradigm which does not rely on background knowledge, called *Differential Privacy*

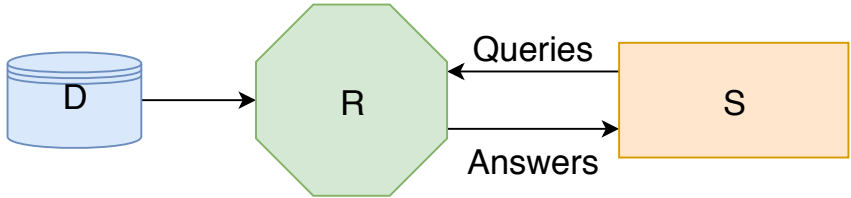
Differential Privacy

Core Idea

Privacy through data perturbation.

- The addition or removal of one record from a database should not reveal any information to an adversary, i.e. your *presence* or *absence* does not reveal or leak any information.
- Use a randomization mechanism to perturb query results of `count`, `sum`, `mean` functions, as well as other statistical query functions.

Differential Privacy



Definition

A randomized mechanism $R(x)$ provides ϵ-*differential privacy* if for any two databases D_1 and D_2 that differ on at most one element, and all outputs $S \subseteq Range(R)$

$$\frac{Pr[R(D_1) \in S]}{Pr[R(D_2) \in S]} \leq \exp(\epsilon)$$

ϵ is a parameter called *privacy budget/level*.

Data Perturbation

Data perturbation is achieved by noise addition.

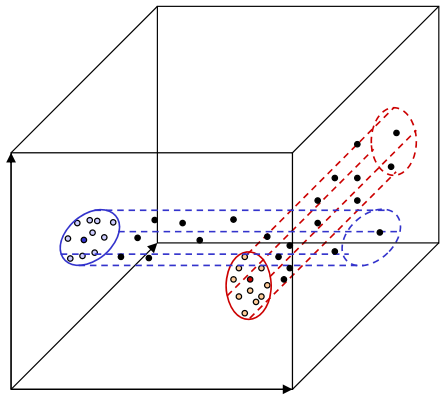
Some Kinds of Noise

- Laplacian noise
- Gaussian noise
- Exponential Mechanism

Agenda

High-Dimensional Data

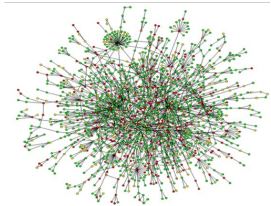
- Challenges:
  - *Curse of dimensionality*: distances become more and more similar
  - Datasets become sparse.
  - Expensive distance measures
  - Degeneration of index structures
  - Unintuitive properties in high dimensions.
- Tasks
  - Feature Selection
  - Feature Reduction / Metric Learning
  - Clustering in High-Dimensional Spaces



$\rightsquigarrow$ Knowledge Discovery in Databases 2 (summer)

Graph Data

- Graphs are everywhere!
  - Chemical data analysis, proteins
  - Biological pathways/networks
  - Program control flow, traffic flow
  - Web graph, social network analysis
- Typical tasks
  - Measure similarity between graphs
  - Find frequent patterns in graphs
  - Generate "realistic" synthetic graphs
  - Identify groups in social networks
  - Integrate additional information



$\rightsquigarrow$ Knowledge Discovery in Databases 2 (summer)

Further Machine Learning Methods

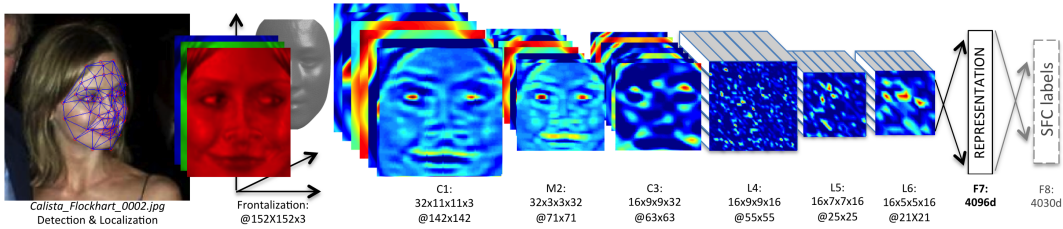


Image Source: Taigman, et al. "Deepface: Closing the gap to human-level performance in face verification." CVPR'14.

- Graphical Models
- Generative Models
- Neural Networks
- Deep Learning

$\rightsquigarrow$ Machine Learning (summer), Deep Learning and Artificial Intelligence (winter)

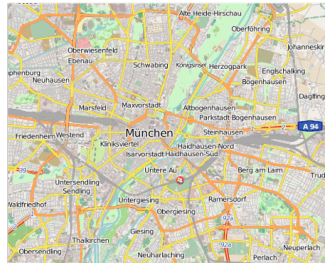
Decision Making / Planning

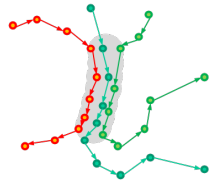
- Setting:
  - Agents are in some environment, observe, and have to take actions that influence the environment.
- Methods:
  - Deterministic/Stochastic Planning
  - A^*-Search
  - Model-Free Reinforcement Learning
  - Q-Learning
  - Adversarial Search (e.g. Alpha-Beta Pruning)

$\rightsquigarrow$ Deep Learning (winter), AI for Games (summer)

Spatial Data

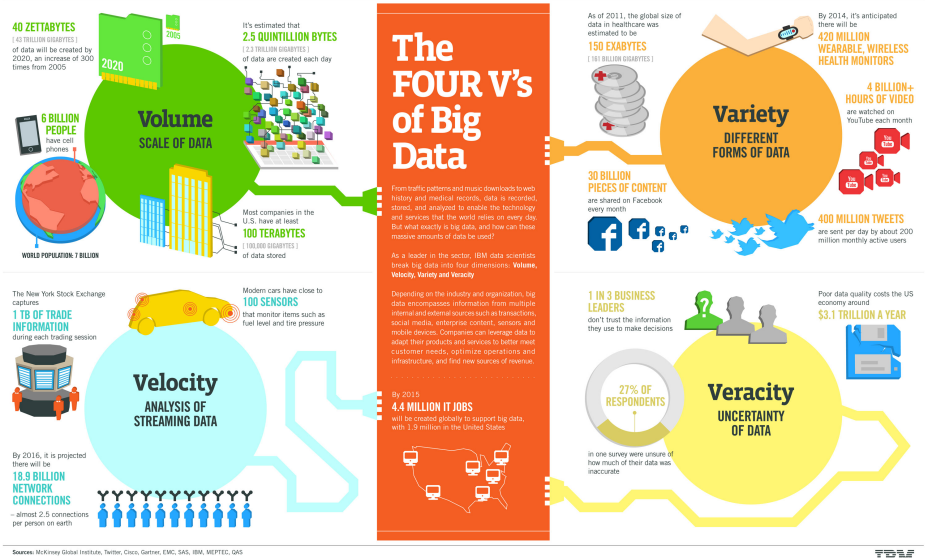
- Mining spatial data
  - Spatial clustering, outlier detection, prediction, rule mining, ...
- Spatial data management
  - Process spatial queries without scanning the whole database
  - Spatial index structures: BSP-tree, R-tree, Quad-tree, ...
- Mining trajectory data
  - Similarity models for trajectories
  - Trajectory compression
  - Mining patterns in trajectories (encounters, flocks, ...)





$\rightsquigarrow$ AI for Games (summer)

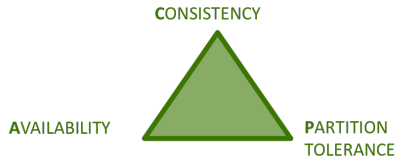
Big Data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Big Data Management

- Vertical scaling limited and expensive
  ⇝ Distributed storage
- NoSQL databases
  - Redis
  - MongoDB
  - Cassandra
  - Neo4J
- Distributed file systems
  - GFS (Google)
  - HDFS (Hadoop)
  - S3 (Amazon)

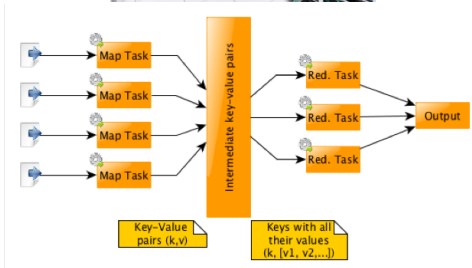
https://www.greentree.com/latest-news/avoiding-cumulus-congestus



⇝ Big Data Management and Analytics (winter)

Distributed Data Processing

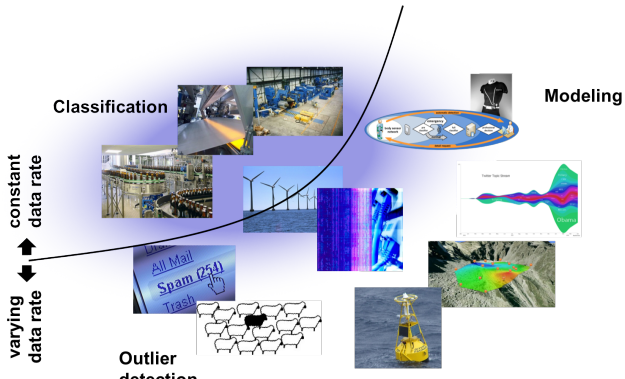
- Processing and analyzing big data
- Map-Reduce: Programming model for distributed processing of large datasets
  - Algorithms are specified as sequences of map and reduce functions
  - Programs are automatically parallelized and executed on a cluster
  - System is tolerant to hardware faults
- Frameworks
  - Apache Spark (batch processing)
  - Apache Flink (stream processing)



$\rightsquigarrow$ Big Data Management and Analytics (winter)

Stream Data

- Data objects arrive over time in a continuous data stream
- Challenges
  - Infinite stream
  - Limited time and memory
  - Evolving distribution
  - Varying data rates
  - Concept drift
- Typical tasks
  - Sampling and buffering
  - Stream statistics
  - Aging mechanisms



⇝ Big Data Management and Analytics (winter)

Seminars, Practicals, Theses

Dive deeper into specific topics and get hands-on experience:

- Master Seminar "Recent Developments in Data Science" (summer and winter)
- Master Practical "Big Data Science" (summer)
- Master Practical "Applied Reinforcement Learning" (summer)
- Individual Bachelor and Master Theses

Conclusion

Best wishes for your exams, and see you at further opportunities!