SECURITY IN COMPUTING, FIFTH EDITION

Databases and Big Data

Objectives for Chapter 7

- Basic database terminology and concepts
- Security requirements for databases
- Implementing access controls in databases
- Protecting sensitive data
- Data mining and big data

Database Terms

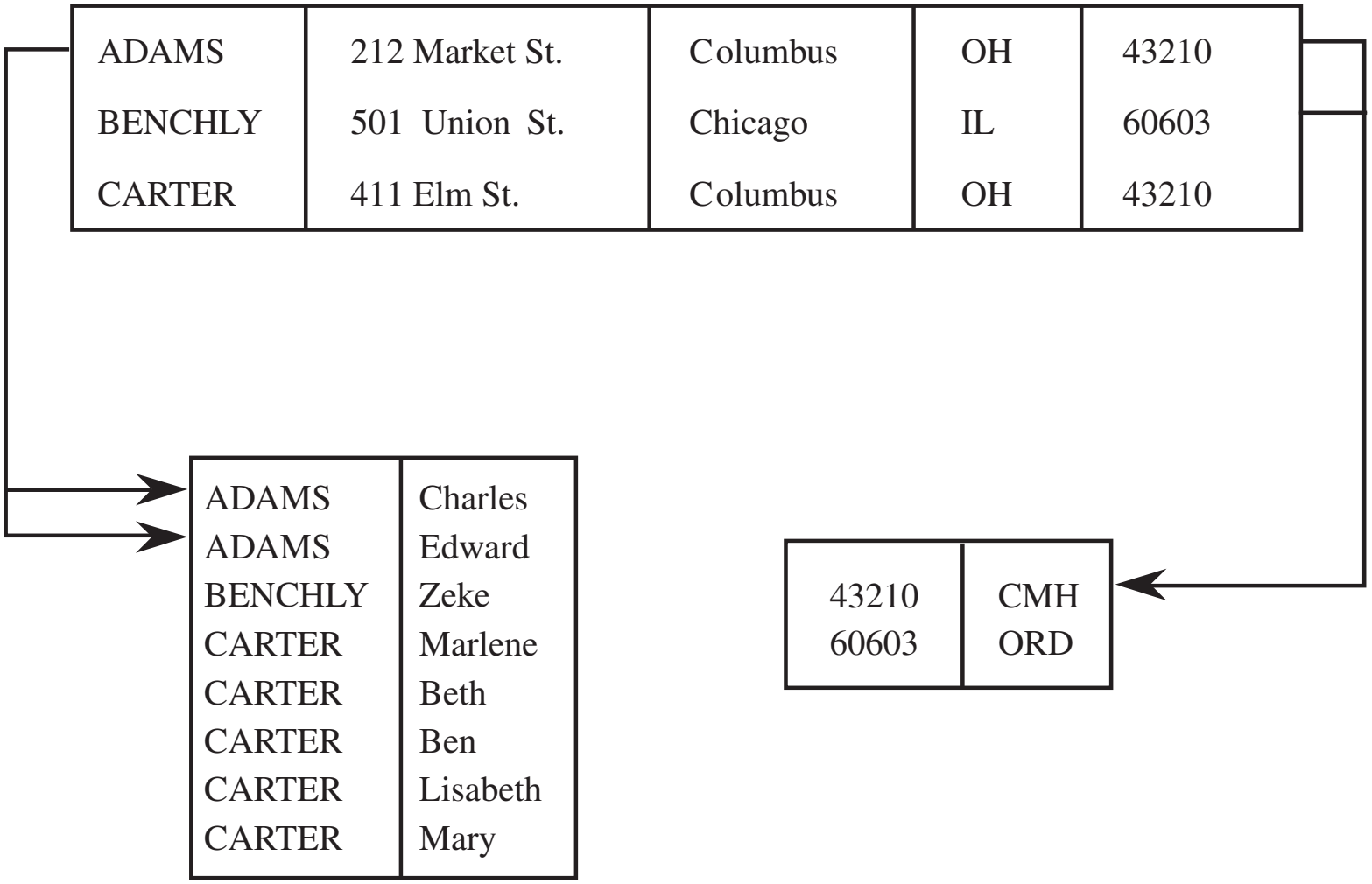
- Database administrator
- Database management system (DBMS)
- Record
- Field/element
- Schema
- Subschema
- Attribute
- Relation

Database Terms

- What is a database?
  - A collection of data and a set of rules that organize the data by specifying certain relationships among the data
- Database administrator
  - Person who defines the rules that organize the data and controls who should have access to what parts of the data
- Database management system (DBMS)
  - The system through which users interact with the database
- Record
  - One related group of data

- Field/element
  - Elementary data items that make up a record (e.g., name, address, city)
- Schema
  - Logical structure of a database
- Subschema
  - The portion of a database a given user has access to
- Attribute
  - A column in a database

Database Example



ADAMS	212 Market St.	Columbus	OH	43210
BENCHLY	501 Union St.	Chicago	IL	60603
CARTER	411 Elm St.	Columbus	OH	43210

ADAMS	Charles
ADAMS	Edward
BENCHLY	Zeke
CARTER	Marlene
CARTER	Beth
CARTER	Ben
CARTER	Lisabeth
CARTER	Mary

43210	CMH
60603	ORD

Schema Example

Name	First	Address	City	State	Zip	Airport
ADAMS	Charles	212 Market St.	Columbus	OH	43210	CMH
ADAMS	Edward	212 Market St.	Columbus	OH	43210	CMH
BENCHLY	Zeke	501 Union St.	Chicago	IL	60603	ORD
CARTER	Marlene	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Beth	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Ben	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Lisabeth	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Mary	411 Elm St.	Columbus	OH	43210	CMH

Queries

- A query is a command that tells the database to retrieve, modify, add, or delete a field or record
- The most common database query language is SQL
- The result of executing a query is a subschema

Example SQL Query

- `SELECT ZIP='43210'`

Name	First	Address	City	State	Zip	Airport
ADAMS	Charles	212 Market St.	Columbus	OH	43210	CMH
ADAMS	Edward	212 Market St.	Columbus	OH	43210	CMH
CARTER	Marlene	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Beth	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Ben	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Lisabeth	411 Elm St.	Columbus	OH	43210	CMH
CARTER	Mary	411 Elm St.	Columbus	OH	43210	CMH

Advantages of Databases

- A database is a single collection of data, stored and maintained at one central location
- People and application have access to it as needed
- Implementation may involve some other physical storage
  - At a local or remote location
- The users are unaware of the physical arrangements
  - Only see a unified logical arrangement

Advantages of Databases

- Shared access
  - many users can use one common, centralized set of data
- Controlled access
  - Only authorized users are allowed to view or to modify data values
- Minimal redundancy
  - Individual users do not have to collect and maintain their own sets of data
- Data consistency
  - A change to a data value affects all users of the data value
- Data integrity
  - Data values are protected against accidental or malicious undesirable changes

Database Security

- Why is security required?
- Databases used by many entities, such as:
  - banks, large retailers, and law enforcement
- Therefore, data needs to be protected
  - It's confidentiality, integrity and availability to its users

Database Security Requirements

- Physical integrity
- Logical integrity
- Element integrity
- Auditability
- Access control
- User authentication
- Availability

Reliability and Integrity

- Reliability: in the context of databases, reliability is the ability to run for long periods without failing
- Database integrity: concern that the database as a whole is protected against damage
- Element integrity: concern that the value of a specific data element is written or changed only by authorized users
- Element accuracy: concern that only correct values are written into the elements of a database

Database Update

- Concern: What if the database system fails in the middle of an update?
  - Leaving the database in a partially updated and inconsistent state
- Solution: two-phase update

Two-Phase Update

- Phase 1: Intent
  - DBMS does everything it can, other than making changes to the database, to prepare for the update
    - Collects records, opens files, locks out users, makes calculations
  - DBMS commits by writing a commit flag to the database
- Phase 2: Write
  - DBMS completes all write operations
  - DBMS removes the commit flag
- If the DBMS fails during either phase 1 or phase 2, it can be restarted and repeat that phase without causing harm

Other Database Security Concerns

- Error detection and correction codes to protect data integrity
- For recovery purposes, a database can maintain a change log, allowing it to repeat changes as necessary when recovering from failure
- Databases use locks and atomic operations to maintain consistency
  - Writes are treated as atomic operations
  - Records are locked during write so they cannot be read in a partially updated state

Sensitive Data

- Inherently sensitive
  - Passwords, locations of weapons
- From a sensitive source
  - Confidential informant
- Declared sensitive
  - Classified document, name of an anonymous donor
- Part of a sensitive attribute or record
  - Salary attribute in an employment database
- Sensitive in relation to previously disclosed information
  - An encrypted file combined with the password to open it

Preventing Disclosure

- Keeping records from being dumped out of the database is not sufficient to actually prevent disclosure.
- There are many ways to deduce the content of a database listed on this slide
  - all of them must be considered when protecting sensitive database information.
- To apply the appropriate protection mechanisms, it is important to understand:
  - the range of possible contents of each attribute
  - the data available to potential attackers

Types of Disclosures

- Exact data
- Bounds
- Negative result
- Existence
- Probable value
- Direct inference
- Inference by arithmetic
- Aggregation
- Hidden data attributes
  - File tags
  - Geotags

Inference

- **Inference** is a way to infer or derive sensitive data from non-sensitive data.
- The inference problem is a subtle vulnerability in database security
- Example: a database may only return the number of entities who have a certain condition
  - If query is focused enough, return of zero would confirm a certain patient does not have this condition

Inference

- **Inference by Arithmetic**: Using statistics to determine sensitive information about users
  - Example: A database may only return the mean of the sensitive values
    - And not the actual values
  - An attacker may create a query which returns one or very few results
    - ‘select female with illness A who lives on Mamaroneck road and is 56 years old’
    - In this case, the attacker may learn the exact information

Aggregation

- Building sensitive results from less sensitive inputs
- Aggregation allows searches in parallel through databases
  - And combining the results

Aggregation

- Example: looking for crime suspects:
  - Find out who ad a motive for committing the crime,
  - when the crime was committed
  - who had alibis covering that time
  - who had the skills
- Using database queries to parallelize this scenario:
  - Create a list of possible suspects,
  - a list with possible motive, and
  - a list of capable persons
- The intersection of these lists is a single person
  - the police have their prime suspect

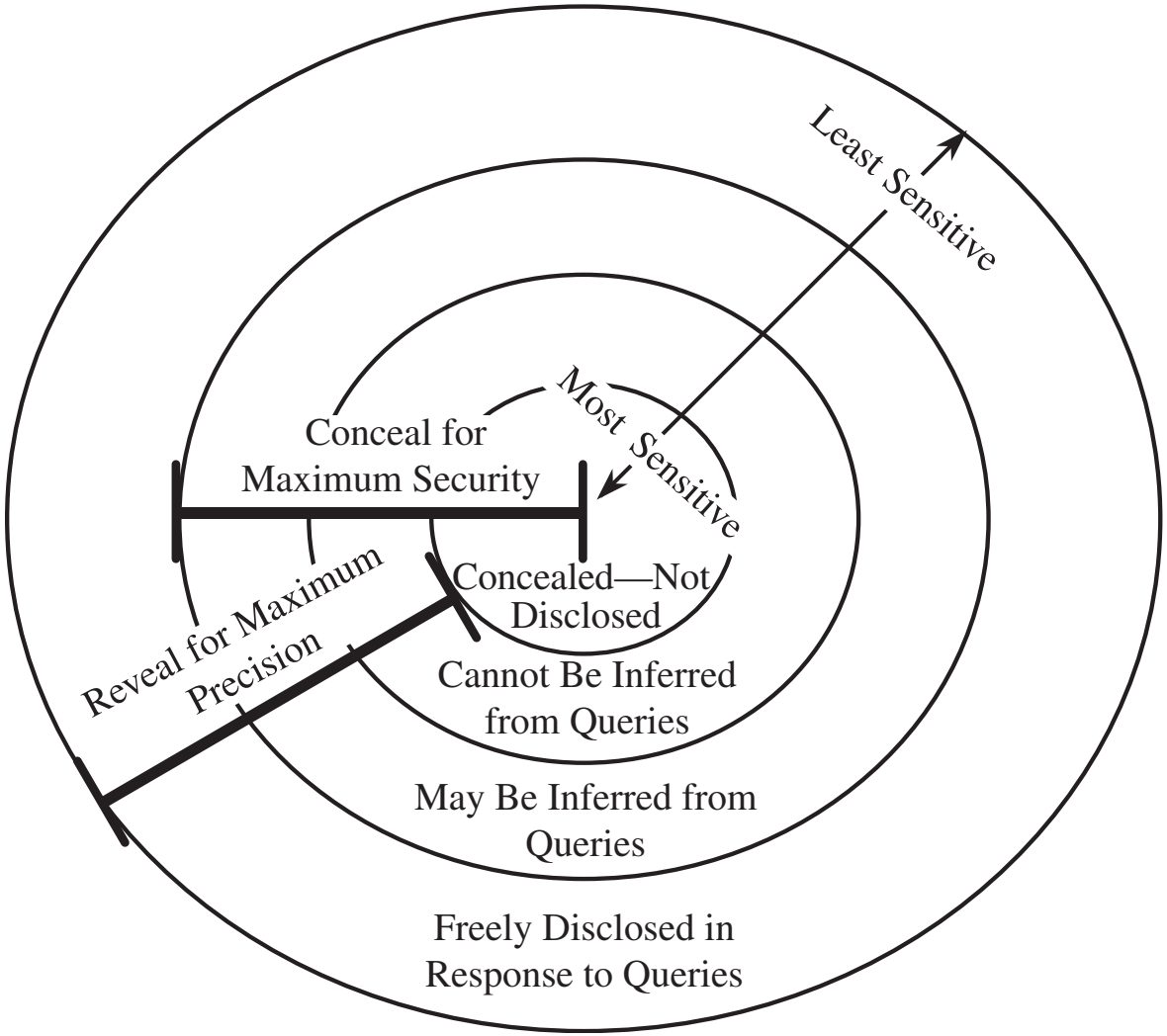
Preventing Disclosure

- Suppress obviously sensitive information
- Keep track of what each user knows based on past queries
- Disguise the data

Security vs. Precision

- Precise, complete, and consistent responses to queries against sensitive information make it more likely that the sensitive information will be disclosed

Security vs. Precision



Suppression Techniques

- Limited response suppression
  - Eliminates certain low-frequency elements from being displayed
- Combined results
  - Ranges, rounding, sums, averages
- Random sample
- Blocking small sample sizes
- Random data perturbation
  - Randomly add or subtract a small error value to/from actual values
- Swapping
  - Randomly swapping values for individual records while keeping statistical results the same

Data Suppression

- Less complex data makes for simpler inference and therefore is more likely to require suppression.
- The disclosure prevention must be balanced against the database requirements
  - as the loss of precision and completeness may make the database unusable

Summary

- Database security requirements include:
  - Physical integrity
  - Logical integrity
  - Element integrity
  - Auditability
  - Access control
  - User authentication
  - Availability

DATA MINING AND BIG DATA

Big Data

- Collection of massive amounts of data
  - Often collected from different resources
    - Often not intended to be databases or structured as such
- Big data comes from sources outside the company
  - Unlike “small data”, who may be generated solely by the organization’s own internal systems
- Big data can come from:
  - Social media as well as video and audio recordings.
  - Government databases, market analytics, and customer reports
  - Etc.

Big Data

- Example: the set of all index entries for search engines
  - A search engine may report it has a few million pages answering your query
  - However, the usefulness of the millionth link may not be high
  - Most users find what they want in the first few results
    - Or redo the query with a different question

Data Mining

- Concept related to Big Data
- More data are being collected and saved than ever before
  - cost per megabyte of storage has fallen significantly
  - Networks and the Internet allow sharing of databases by people
    - in ways previously unimagined

Data Mining

- Data Mining allows searching for meaningful information in massive amounts of data
  - Through intelligent analyzing and querying of the data
  - In a largely automated way

https://www.lynda.com/SPSS-tutorials/Essential-Elements-Predictive-Analytics-Data-Mining/578072-2.html

Data Mining

- People and programs that search and sift datasets to derive data
  - Isn't that what databases are for?
- Data mining implies searching for patterns and connections that were previously unknown
  - and perhaps even unpredictable.
- Example: left-handed people are more likely to prefer fried eggs to poached eggs
  - Found in a study that used data mining techniques

The Human Face of Big Data

- PBS: The Human Face of Big Data

Data Mining

- Data mining uses different techniques to discover patterns and relations on large datasets
  - Such as statistics, machine learning, mathematical models, pattern recognition
- The size and value of the datasets present an important security and privacy challenge, as the consequences of disclosure are naturally high

Data Mining

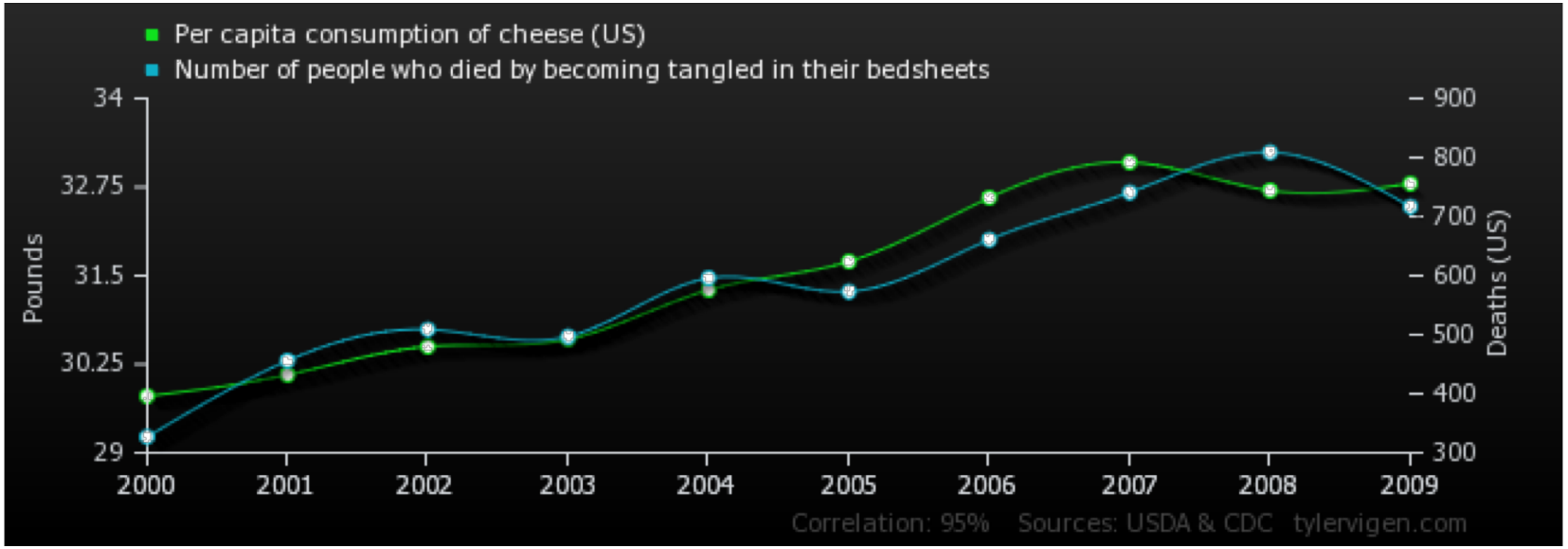
- Functions that can be performed:
  - Association: one event often goes with another),
  - Sequences: one event often leads to another),
  - Classification: events exhibit patterns, for example, coincidence)
  - Clustering: some items have similar characteristics)
  - Forecasting:past events foretell future ones)

Data Mining and Databases

- Generally, database queries are manual
  - Data mining is more automatic
- Data mining presents probable relationships
- Humans interpret the output of data mining algorithms
  - For example, not all relationships are causal
    - More variables may be needed to understand cause
    - Some may be random

Data Mining

- Example: correlation found between
  - Per Capital Consumption of Cheese (US)
  - Number of people who died by becoming tangled in their bedsheets



http://tylervigen.com/view_correlation?id=7

Data Mining and security

- Can support analysis of security data
  - Data mining is widely used to analyze system data
    - for example, audit logs, to identify patterns related to attacks
    - Finding precursors to an attack can help develop good prevention tools and techniques
    - Detecting actions associated with an attack can help pinpoint vulnerabilities
      - to control any damage that may have occurred

Data Mining and security

- How does data mining affects the CIA Triad?
  - Confidentiality, Integrity and Availability
- Confidentiality includes:
  - Privacy of individual's data
  - proprietary and commercially sensitive data
  - protecting the value of intellectual property
- How to control what we disclose/derive?

Data Mining and security

- Integrity: need to ensure correctness
  - incorrect data are both useless and potentially damaging
- Availability: considerations relates to both performance and structure
  - Combining databases not originally designed to be combined affects results
    - whether results can be obtained in a timely manner
      - or even at all

Privacy and Sensitivity

- Goal of data mining is summary results
  - not individual data items
  - However, still poses a risk to individual privacy
    - Due to inference and aggregation issues
- Companies, organizations and governments can also be affected
  - By results of data correlation and aggregation

The Human Face of Big Data

- PBS: The Human Face of Big Data

Data Mining Challenges

- Correcting mistakes in data
  - What happens when data is moved to more databases before the original database can be corrected?
    - Need for correction may not be disclosed
    - Open challenge
- Example: a government's intelligence service collects data on suspicious activities
  - Names of suspicious persons are foreign, written in a different alphabet.
  - When transformed into the government's alphabet, the transformation is irregular
    - Different agents may spell names differently

Data Mining Challenges

- Preserving privacy
  - Anonymization of the database can help protect privacy
    - But only to a limited amount
  - More data can help identify individuals
    - More data terms reduce the number of persons matching all attributes
    - Example: Who is the cancer patient living on Mamaroneck Rd., aged 55, in a household with two cats, subscribing to *Money* magazine, who makes frequent telephone calls to Canada.

Data Mining Challenges

- Granular access control
  - Access control is often performed in a coarse way
    - E.g., all users that have a certain role get same permissions
- Secure data storage
  - Data may be collected globally and through cloud providers
    - Where security details are largely unknown to users
- Transaction logs
- Real-time security monitoring

Data Mining Challenges

- Many challenges remain open, partially solved or solved in certain data mining packages.
- As data mining platforms evolve, these features will mature

Summary

- There are many subtle ways for sensitive data to be inadvertently disclosed
- There is no single answer for prevention
- Data mining and big data have numerous open security and privacy challenges

Questions?