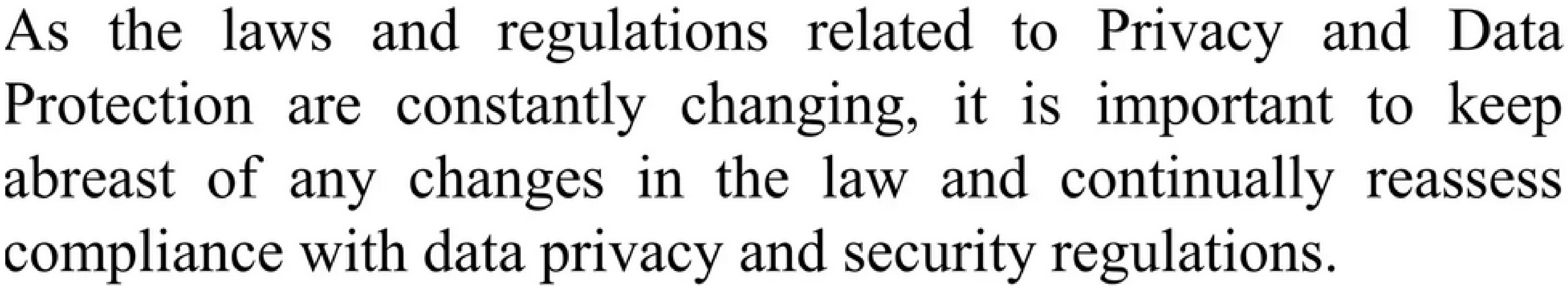
Data Security and Privacy
DSE 3258

L10 –Introduction to Data Privacy

Introduction to Data Privacy

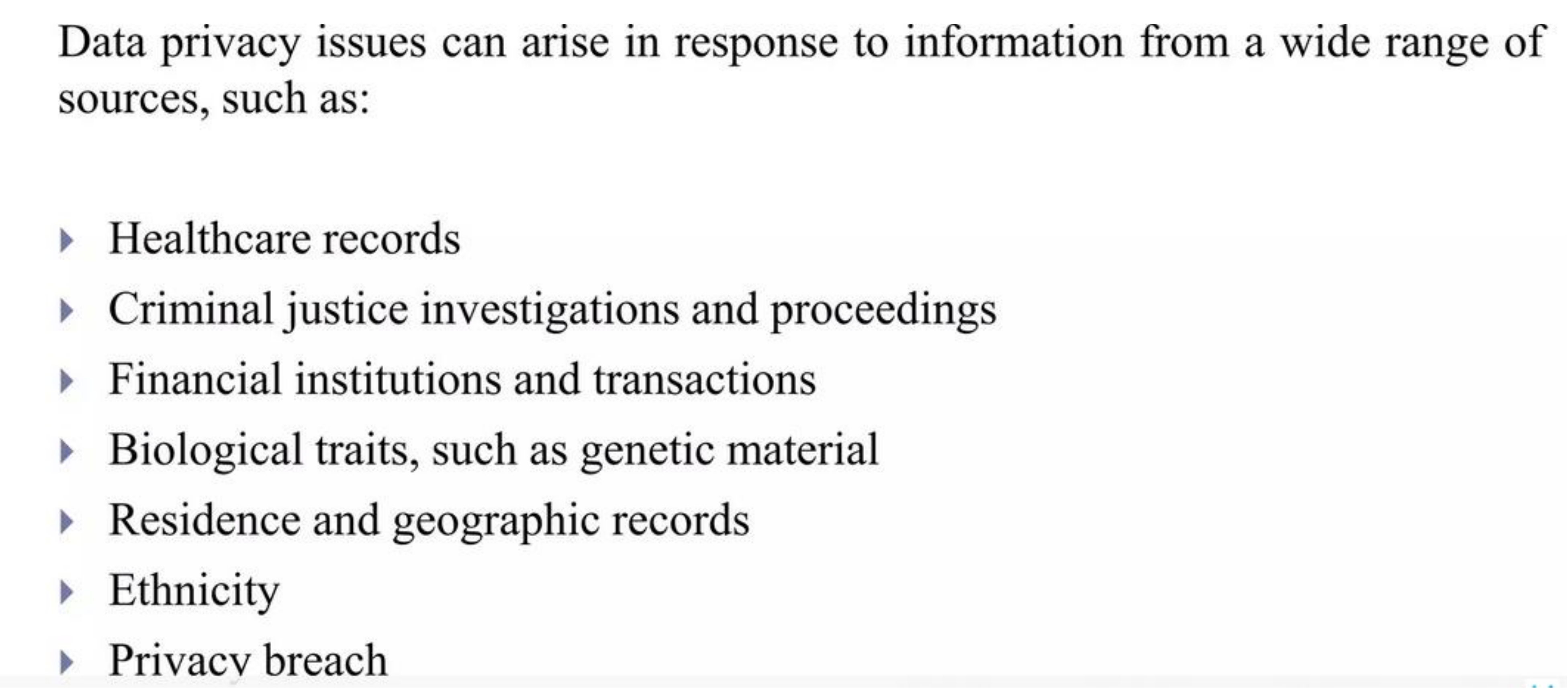
- **What is data Privacy?**
- Data privacy, also called information privacy, is an aspect of data protection that addresses the proper storage, access, retention, immutability and security of sensitive data.
- Data privacy is typically associated with the proper handling of personal data or personally identifiable information (PII), such as names, addresses, Social Security numbers and credit card numbers. However, the idea also extends to other valuable or confidential data, including financial data, intellectual property and personal health information.
- Data privacy is not a single concept or approach. Instead, it's a discipline involving rules, practices, guidelines and tools to help organizations establish and maintain required levels of privacy compliance.

Introduction to Data Privacy



As the laws and regulations related to Privacy and Data Protection are constantly changing, it is important to keep abreast of any changes in the law and continually reassess compliance with data privacy and security regulations.

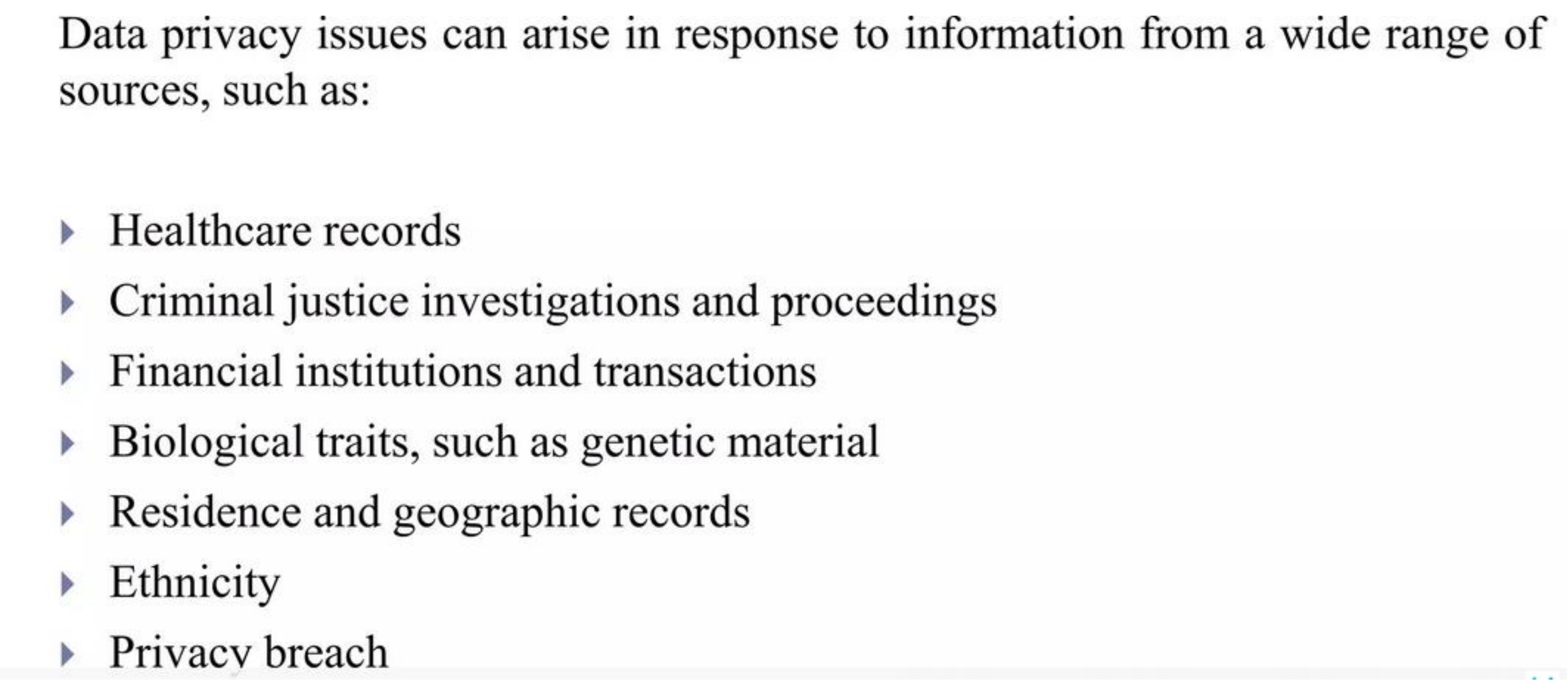
Introduction to Data Privacy



Data privacy issues can arise in response to information from a wide range of sources, such as:

- Healthcare records
- Criminal justice investigations and proceedings
- Financial institutions and transactions
- Biological traits, such as genetic material
- Residence and geographic records
- Ethnicity
- Privacy breach

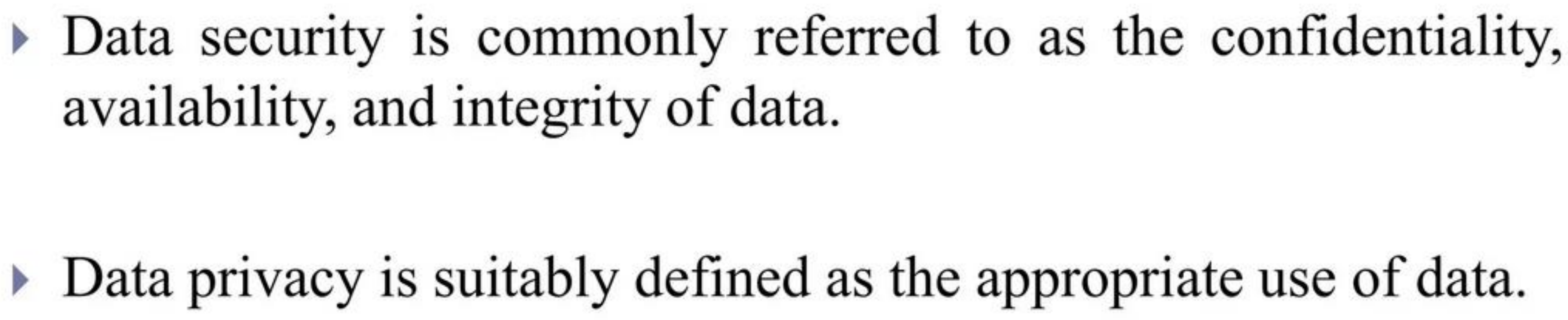
Introduction to Data Privacy



Data privacy issues can arise in response to information from a wide range of sources, such as:

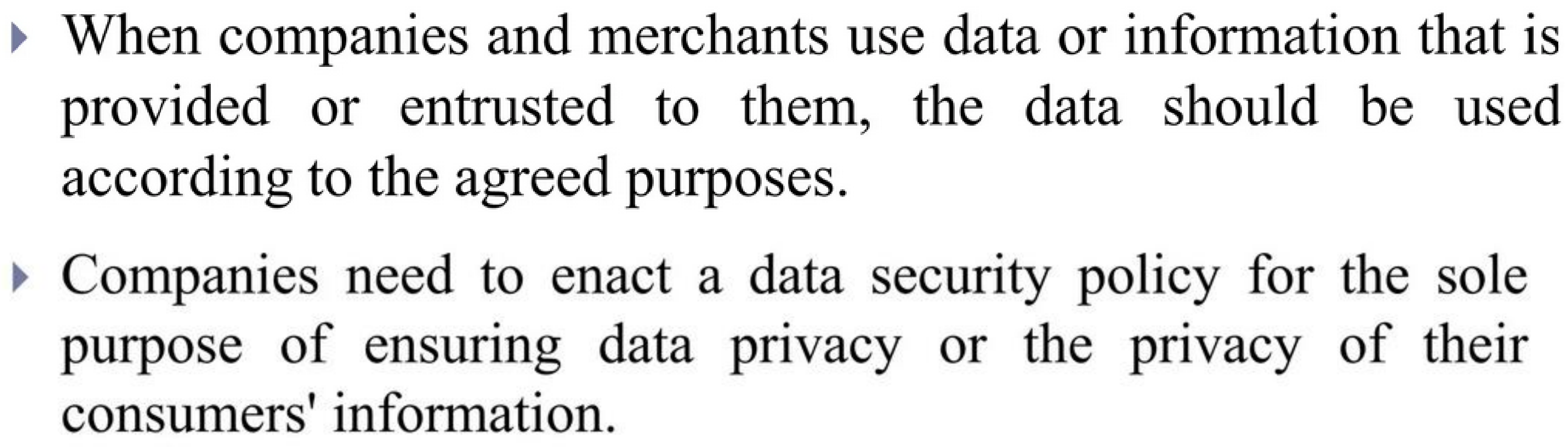
- Healthcare records
- Criminal justice investigations and proceedings
- Financial institutions and transactions
- Biological traits, such as genetic material
- Residence and geographic records
- Ethnicity
- Privacy breach

Data Security Vs Privacy



- Data security is commonly referred to as the confidentiality, availability, and integrity of data.

- Data privacy is suitably defined as the appropriate use of data.



- When companies and merchants use data or information that is provided or entrusted to them, the data should be used according to the agreed purposes.
- Companies need to enact a data security policy for the sole purpose of ensuring data privacy or the privacy of their consumers' information.

Types of privacy attacks

- In privacy-related attacks, the goal of an adversary is to gain knowledge that was not intended to be shared.
- There are several types based on different domains.
- **Linkage Attacks**
  - Linkage attacks use more than one data source and link them together to re-identify individuals or to gain more information to identify individuals. In general, linkage attacks are successful when attackers have an auxiliary data source that connects easily with another dataset.
- **Singling Out Attacks**
  - Singling out attacks work by singling out an individual in a public release and attempting to gather more information about them via the same dataset or via other sources. These attacks can also be performed in reverse, by bringing information about an individual to a released dataset and attempting to deduce whether this person is included and can be identified.

Types of privacy attacks

- **Membership Inference Attack**
  - In a membership inference attack, the attacker tries to learn if a person was a member of the training data.
- **Inferring Sensitive Attributes**
  - Membership inference attacks can be generalized to describe group attributes of the training data population, called an *attribute privacy attack*. In this case, the attacker wants to learn things about the underlying population and uses the same technique to test theories about types of people who might be represented in the training dataset. This attack reveals sensitive group details.

Types of privacy attacks

- **Identity theft:** This involves stealing someone's personal information, such as their name, date of birth, Social Security number, and financial information, with the intent of assuming their identity or using the information for fraudulent purposes.
- Identity theft is committed in many different ways and its victims are typically left with damage to their credit, finances, and reputation.
- To find such information, they may search the hard drives of stolen or discarded computers; hack into computers or computer networks of organizations and corporations; access computer-based public records; use information-gathering malware to infect computers; browse social networking sites; or use deceptive emails or text messages.

Types of privacy attacks

- **Data breaches**: This occurs when an unauthorized person gains access to sensitive information that is stored in a database or computer system. This can result in the exposure of personal information, such as credit card numbers, passwords, and other sensitive data.
- **Phishing :** Phishing refers to an attempt to steal sensitive information, typically in the form of usernames, passwords, credit card numbers, bank account information or other important data in order to utilize or sell the stolen information. This involves tricking individuals into providing their personal information by posing as a legitimate company or organization. Phishing attacks are often carried out through email, social media, or other forms of communication.

Types of privacy attacks

- **Malware attacks**: This involves infecting a computer or device with malicious software that can steal personal information, monitor user activity, or perform other harmful actions.
- **Eavesdropping**: This involves intercepting and listening to private conversations or communications, such as phone calls or text messages, without the knowledge or consent of the individuals involved
- **Password cracking (also called, password hacking)** is an attack vector that involves hackers attempting to crack or determine a password. Password hacking uses a variety of programmatic techniques and automation using specialized tools.
- **Surveillance:** This involves monitoring an individual's activities, either physically or digitally, without their knowledge or consent. This can include tracking their location, online activity, or other behaviors.

Data linking and profiling

- **Data profiling :** is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of data quality issues, risks, and overall trends.
- **Data profiling** is the systematic process of determining and recording the characteristics of data sets. We can also think of it as building a metadata catalog that summarizes the essential characteristics.
  - This can include information such as demographic data, buying habits, online activity, and other personal information.
  - Profiling can be used for a variety of purposes, such as marketing, advertising, and personalization of services.

Uses of data profiling

- **Query optimization**
  - ✓ Data profiling provides information on the characteristics of a database, such as rows, columns, average values, and more.
  - ✓ Statistics about each database helps to estimate the query design, considerations, and implementation plan. As a result, we can optimize your queries for better performance.
- **Data integration**
  - ✓ To integrate multiple datasets, we first need to understand the datasets and their relationships. This is crucial to understand how to link datasets, what's the best way to link them, do we need to take into account different conventions such as name or unit of measurement, and so on.
- **Scientific data management**
  - ✓ Before importing raw data into your databases, it's important to understand the nature of that data. That's where data profiling can help. After profiling these datasets, you can develop a plan to extract that data and adopt the appropriate schema.
- **Data analytics**
  - ✓ Any analysis or data mining starts with data profiling. Data profiling gives an initial high-level understanding of the dataset and its characteristics so that you can choose the right algorithms.

https://atlan.com/data-profiling-101/

Uses of data profiling

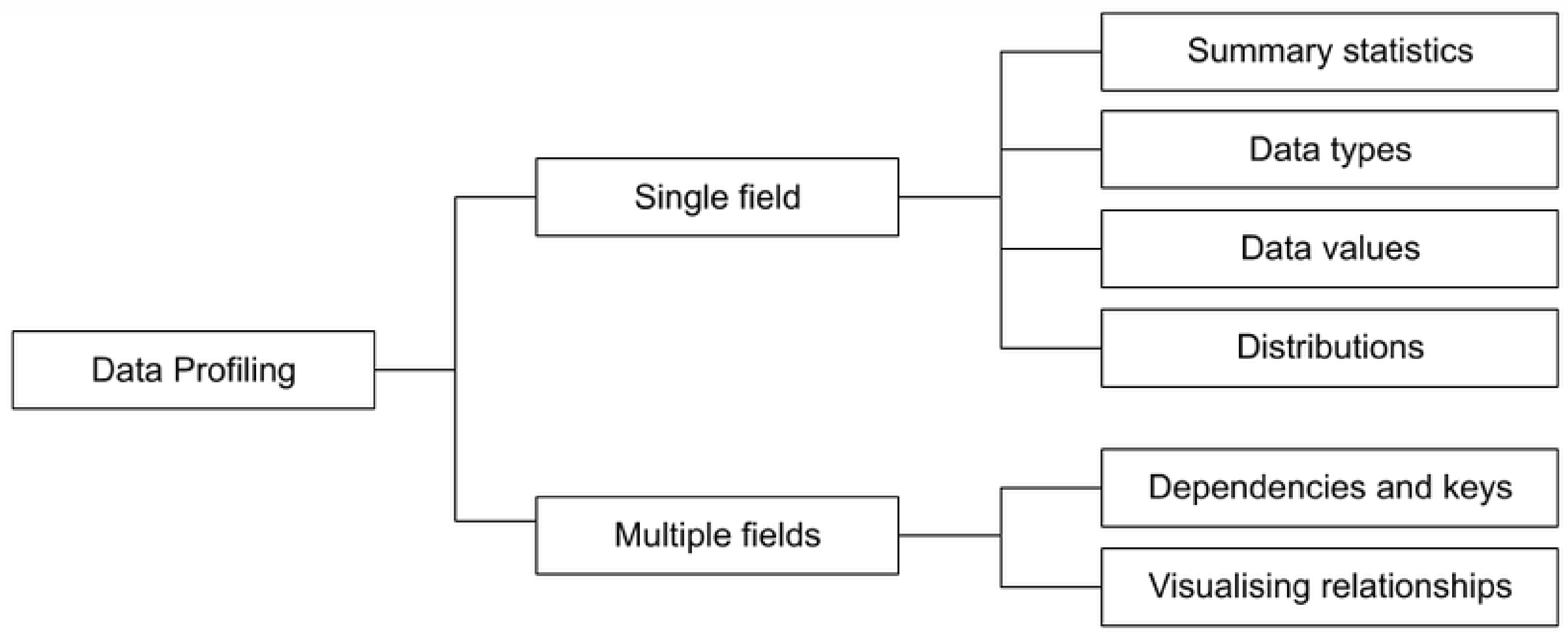
- **Project management**
  - ✓ Taking data-driven decisions requires a solid understanding of the data you have, and the information you need for the project. With data profiling, you can take stock of your data, its quality, completeness, and credibility. You can also determine whether you have all the data you need to make your project work.
- **Data discovery**
  - ✓

    Having data available to be used broadly across an organization requires that data be easily accessible, searchable, and understandable. Data profiling can help by enabling you to compile the metadata needed, along with descriptive summaries and metrics for better context.

https://atlan.com/data-profiling-101/

data profiling: Techniques

- According to Felix Naumann's data profiling can be done using single or multiple fields.



Data Profiling

Single field

Summary statistics

Data types

Data values

Distributions

Multiple fields

Dependencies and keys

Visualising relationships

https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/folien/SS17/DP_02_ProfilingIntro.pdf

- ■ Single column profiling
  - ✓Most basic form of data profiling
  - ✓ Assumption: All values are of same type
  - ✓Assumption: All values have some common properties
    - ✓ – That are to be discovered
  - ✓Often part of the basic statistics gathered by DBMS
- ■ Multicolumn profiling
  - ✓ Discover joint properties
  - ✓Discover dependencies

Single field profiling is the most basic form of profiling that assumes all fields are of the same type and share common properties. This type of profiling helps you discover:

- **Summary statistics**: This includes count of data and mathematical aggregations such as maximum, minimum, and mean values.
- **Data types**: This involves determining whether the data is categorical, continuous, and exhibits any patterns. Simple data types include strings, numbers, and timestamps, whereas more complex types include XML
- **Data values**: This means identifying the characteristics and patterns in data values. Examples include address fields, cities, ID strings, and more. Profiling data values also helps you assess your data against known business rules.
- **Distributions**: Visualizing data distribution is useful in spotting outliers. For categorical data, you can see counts per category. Meanwhile, for numerical data, you can plot histograms and note characteristics like skewness, presence of outliers etc..

Multi-field profiling explores the relationship between fields to discover:

- **Inclusion dependencies, keys, and functional dependencies**: With profiling, you can find out if the values in one field are a subset of values in other fields.
- **Visualize numerical relationships**: Profiling helps explore the relationships between numerical fields using pair plots, cross-correlation heat maps, or tables of correlations between fields. These visualizations provide a quick overview of the relationships each data set has with other assets.

Data profiling tools and algorithms

IBM InfoSphere Information Analyzer

□ http://www.ibm.com/software/data/infosphere/information-analyzer/

■ Oracle Enterprise Data Quality

□ http://www.oracle.com/us/products/middleware/data-integration/enterprise-data-quality/overview/index.html

■ Talend Data Quality

□ http://www.talend.com/products/data-quality

■ Ataccama DQ Analyzer

□ http://www.ataccama.com/en/products/dq-analyzer.html

■ SAP BusinessObjects Data Insight and SAP BusinessObjects Information Steward

□ http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/datainsight/index.epx

□ http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/information-steward/index.epx

■ Informatica Data Explorer

□ http://www.informatica.com/us/products/data-quality/data-explorer/

■ Microsoft SQL Server Integration Services Data Profiling Task and Viewer

□ http://msdn.microsoft.com/en-us/library/bb895310.aspx

■ Trillium Software Data Profiling

□ http://www.trilliumsoftware.com/home/products/data-profiling.aspx

■ CloverETL Data Profiler

□ http://www.cloveretl.com/products/profiler

■ OpenRefine

□ http://www.openrefine.org

Data linking

- **Data linking** refers to the process of connecting two or more sets of data that were previously thought to be unrelated.
    - Can involve combining data from different sources, such as databases, social media, and online activity, to create a more comprehensive picture of an individual's behavior or preferences.
- Data linking is the process of collating information from different sources in order to create a more valuable and helpful data set. The linking of information about the same person or entity from disparate sources allows, among other things, the construction of a chronological sequence of events. This information is of immense value at the policy level to derive meaningful decisions.

https://www.spotfire.com/glossary/what-is-data-linking

Ways to link data sets

1. Unique identifier

A unique identifier is available on each data set that establishes the links between these data sets. It is also called deterministic or exact linking because the unique identifiers either match completely, or do not at all.

2. Linkage key

- ✓When a unique identifier is not available, or there isn’t enough quality in the data to rely on, another approach is used called linkage key.
- ✓The linkage key works like a substitute for the unique identifier in this method.
- ✓This key is created using information like name and address available on both data sets.
- ✓ These linkage keys maintain the privacy of the person or entity as the key is used in place of the name and address.

Ways to link data sets

- **3. Probabilistic linking**
- This is another style of data linking, and it is used when a unique identifier is unavailable. It is based on the probability that the pair of records, taken from one data set, refers to the same entity or person. In this method, advanced data linking software is used to obtain accurate results.
- **4. Statistical linking**
- This technique combines records similar to the entity but not necessarily the same person or organization. This kind of data linking may not give the most accurate results but does provide a pattern or trend from the given information or statistics.

Access control models

- Access control is the process of controlling who has access to resources or information within an organization's computer systems.
- Access control mechanisms are designed to ensure that only authorized individuals are able to access resources, while preventing unauthorized access.