CityU
香港城市大學
City University of Hong Kong

Lecture 10:

Data Management

CS5481 Data Engineering

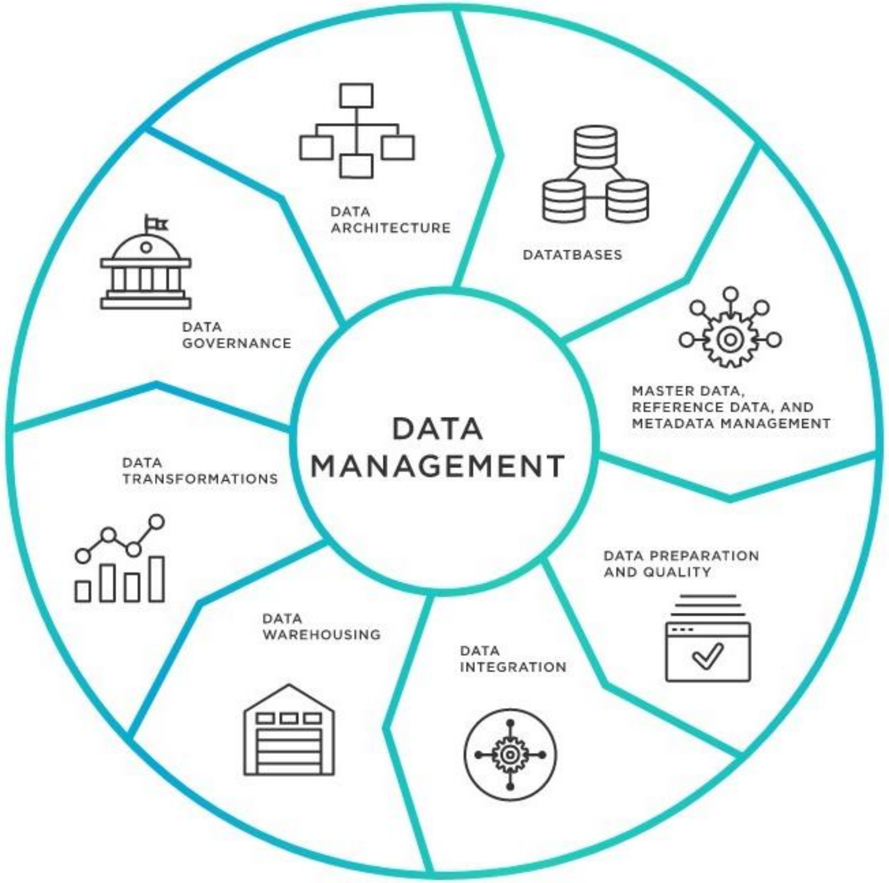
Instructor: Linqi Song

Outline

1. What is Data Management
2. Data Quality
3. Data Security
4. Data Privacy

What is data management?

- Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively.
- As organizations create and consume data at unprecedented rates, data management solutions become essential for making sense of the vast quantities of data.
- Today's leading data management software ensures that reliable, up-to-date data is always used to drive decisions.



Data management techniques (1)

Types of data management: Data management plays several roles in an organization’s data environment, making essential functions easier and less time-intensive. These data management techniques include the following:

- **Data pipelines** enable the automated transfer of data from one system to another.
- **Data preparation** is used to clean and transform raw data into the right shape and format for analysis, including making corrections and combining data sets.
- **ETLs (Extract, Transform, Load)** are built to take the data from one system, transform it, and load it into the organization’s data warehouse.
- **Data catalogs** help manage metadata to create a complete picture of the data, providing a summary of its changes, locations, and quality while also making the data easy to find.

Data management techniques (2)

- **Data warehouses** are places to consolidate various data sources, contend with the many data types businesses store, and provide a clear route for data analysis.
- **Data governance** defines standards, processes, and policies to maintain data security and integrity.
- **Data architecture** provides a formal approach for creating and managing data flow.
- **Data security** protects data from unauthorized access and corruption.
- **Data modeling** documents the flow of data through an application or organization.

Data quality – What is data quality?

- **Data quality** refers to the development and implementation of activities that apply quality management techniques to data in order to ensure the data is fit to serve the specific needs of an organization in a particular context. Data that is deemed fit for its intended purpose is considered high quality data.
- Examples of data quality issues include duplicated data, incomplete data, inconsistent data, incorrect data, poorly defined data, poorly organized data, and poor data security.

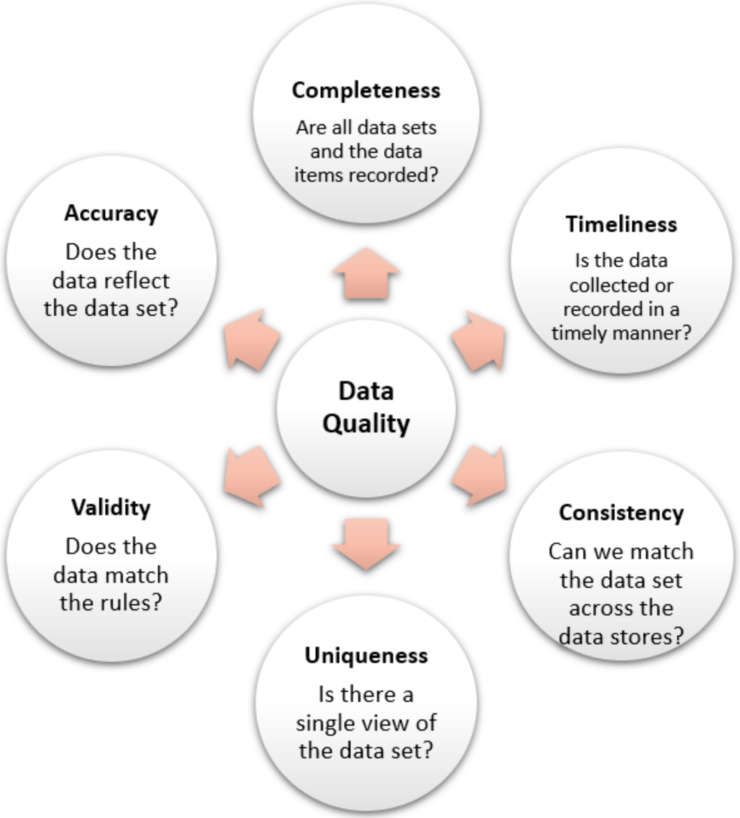
Data quality dimensions (1)

There are six main dimensions of data quality: accuracy, completeness, consistency, validity, uniqueness, and timeliness.

Accuracy: The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.

Completeness: Completeness is a measure of the data's ability to effectively deliver all the required values that are available.

Consistency: Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in difference locations should not conflict with one another.


Completeness
Are all data sets and the data items recorded?
Accuracy
Does the data reflect the data set?
Timeliness
Is the data collected or recorded in a timely manner?
Data Quality
Validity
Does the data match the rules?
Consistency
Can we match the data set across the data stores?
Uniqueness
Is there a single view of the data set?

Data quality dimensions (2)

Validity: Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.

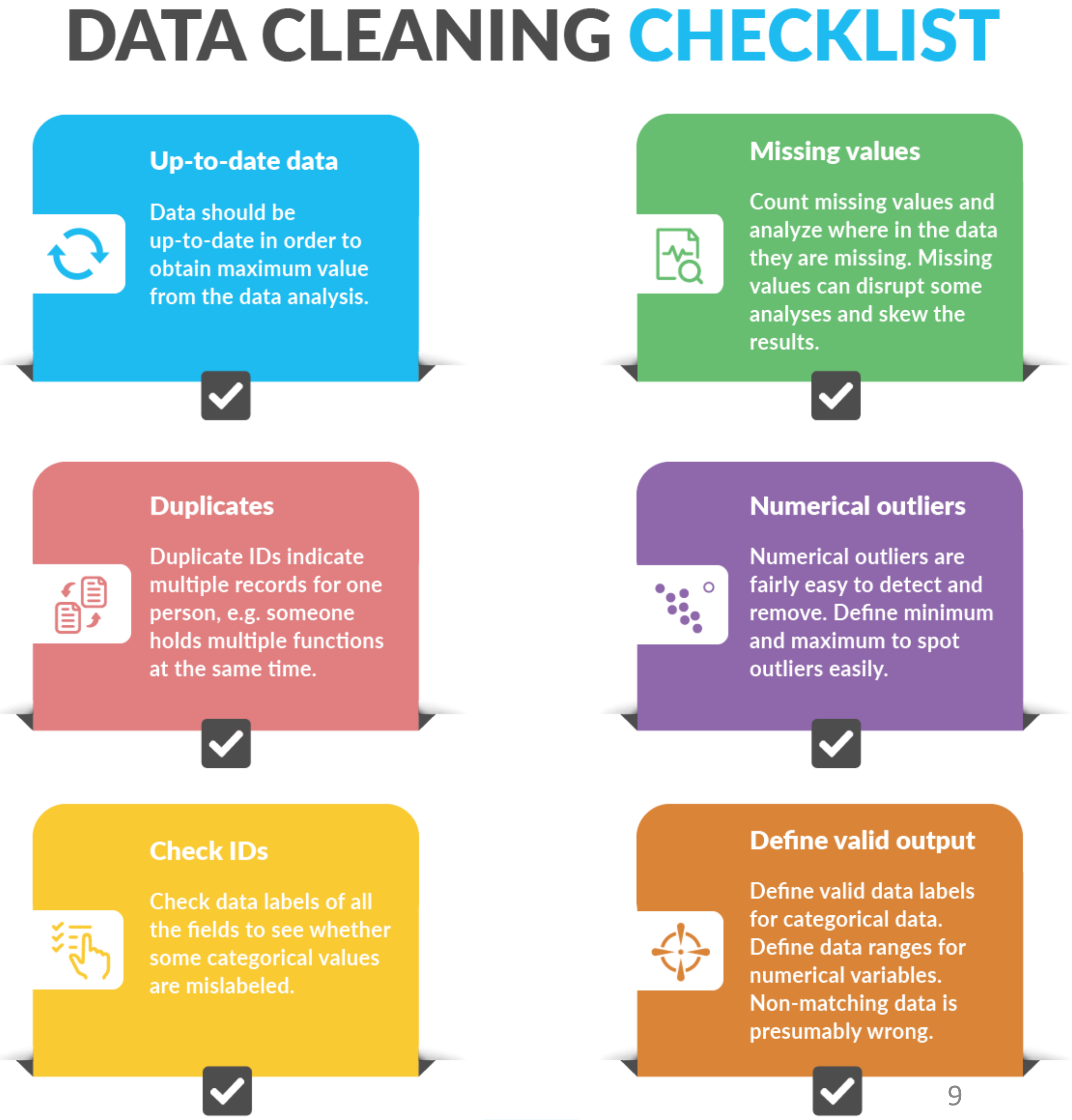
Uniqueness: Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.

Timeliness: Timely data is data that is available when it is required. Data may be updated in real time to ensure that it is readily available and accessible.

How to improve data quality? (1)

Data quality measures can be accomplished with data quality tools, which typically provide data quality management capabilities such as:

- **Data preprocessing (Lecture 3)** – data cleaning, such as handling duplicates, outliers, and missing data, etc.



DATA CLEANING CHECKLIST

Up-to-date data

Data should be up-to-date in order to obtain maximum value from the data analysis.

Missing values

Count missing values and analyze where in the data they are missing. Missing values can disrupt some analyses and skew the results.

Duplicates

Duplicate IDs indicate multiple records for one person, e.g. someone holds multiple functions at the same time.

Numerical outliers

Numerical outliers are fairly easy to detect and remove. Define minimum and maximum to spot outliers easily.

Check IDs

Check data labels of all the fields to see whether some categorical values are mislabeled.

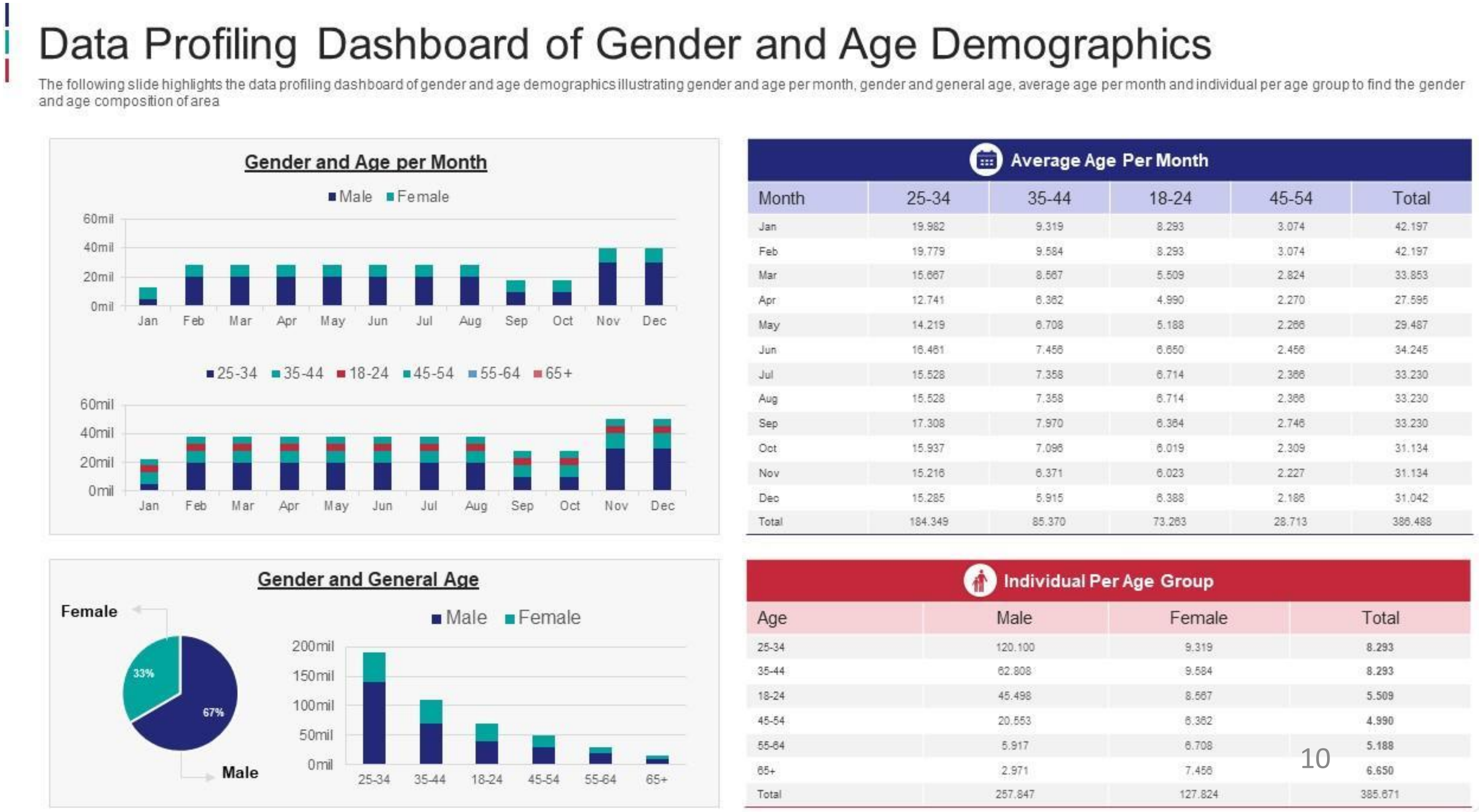
Define valid output

Define valid data labels for categorical data. Define data ranges for numerical variables. Non-matching data is presumably wrong.

How to improve data quality? (2)

- **Data profiling** - the process of examining, analyzing, and creating useful summaries of data.
  - The first step in the data quality improvement process is understanding your data.
  - Data profiling is the initial assessment of the current state of the data sets.

Original data



Data Profiling Dashboard of Gender and Age Demographics

The following slide highlights the data profiling dashboard of gender and age demographics illustrating gender and age per month, gender and general age, average age per month and individual per age group to find the gender and age composition of area

Average Age Per Month

Month	25-34	35-44	18-24	45-54	Total
Jan	19.982	9.319	8.293	3.074	42.197
Feb	19.779	9.584	8.293	3.074	42.197
Mar	15.667	8.567	5.509	2.824	33.853
Apr	12.741	6.362	4.990	2.270	27.595
May	14.219	6.708	5.188	2.266	29.487
Jun	16.461	7.456	6.650	2.456	34.245
Jul	15.528	7.358	6.714	2.366	33.230
Aug	15.528	7.358	6.714	2.366	33.230
Sep	17.308	7.970	6.364	2.746	33.230
Oct	15.937	7.096	6.019	2.309	31.134
Nov	15.216	6.371	6.023	2.227	31.134
Dec	15.285	5.915	6.388	2.186	31.042
Total	184.349	85.370	73.263	28.713	386.488

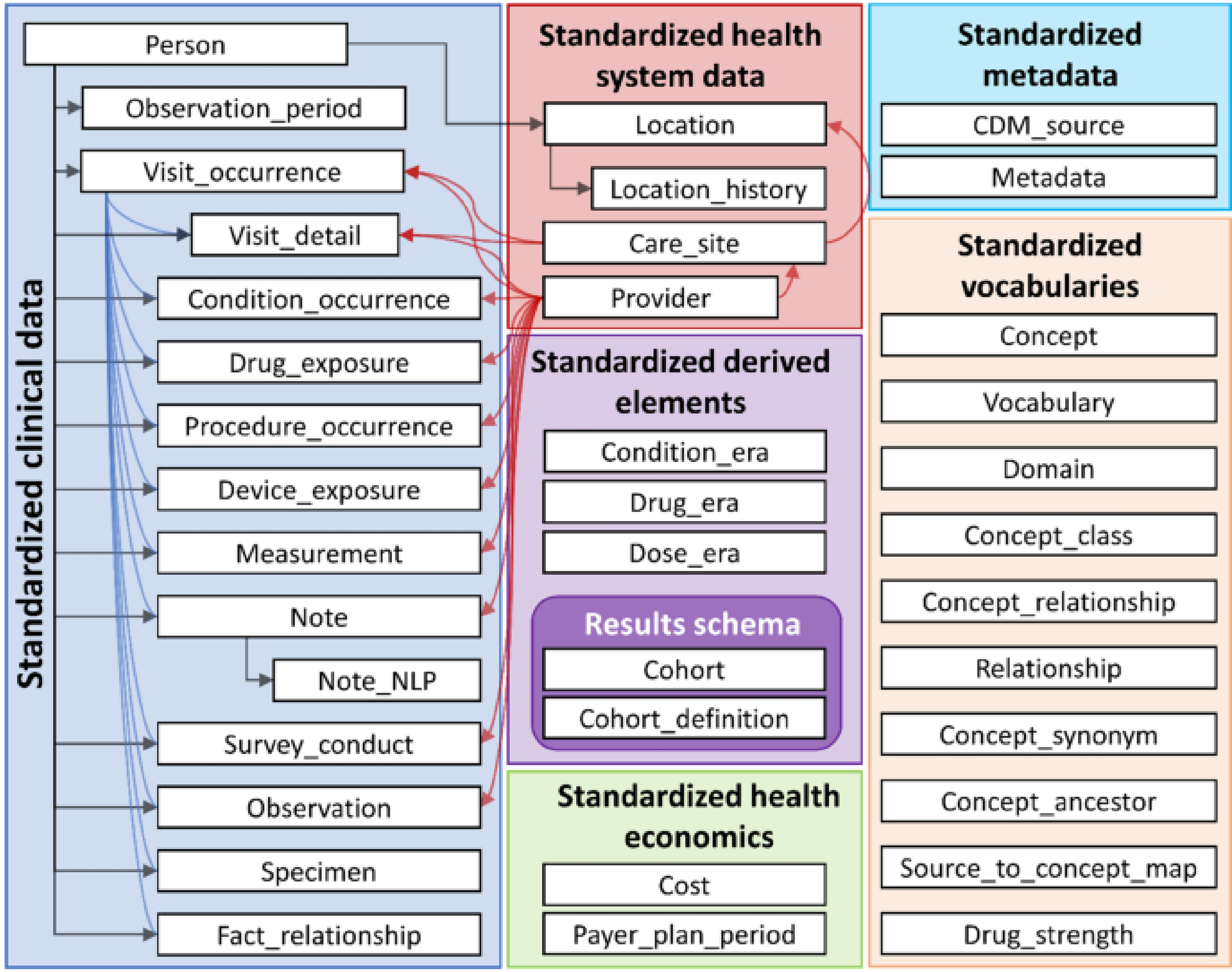
Individual Per Age Group

Age	Male	Female	Total
25-34	120.100	9.319	8.293
35-44	62.808	9.584	8.293
18-24	45.498	8.567	5.509
45-54	20.553	6.362	4.990
55-64	5.917	6.708	5.188
65+	2.971	7.456	6.650
Total	257.847	127.824	385.671

How to improve data quality? (3)

- **Data Standardization** - disparate data sets are conformed to a common data format.

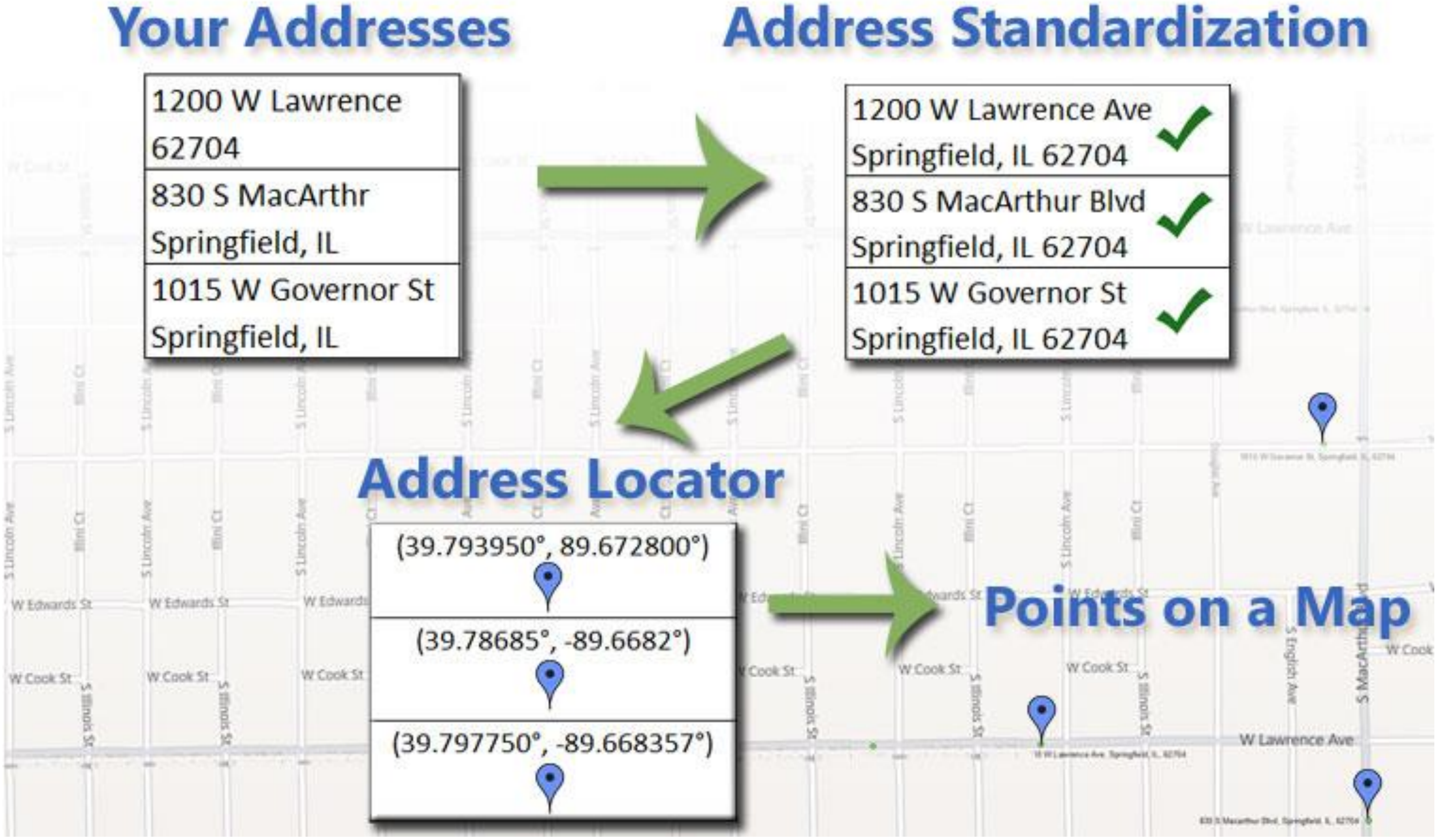
A set of rules and designs.



A healthcare domain data standardization example

How to improve data quality? (4)

- **Geocoding** - The description of a location is transformed into coordinates that conform to U.S. and worldwide geographic standards.



Your Addresses

1200 W Lawrence
62704

830 S MacArthr
Springfield, IL

1015 W Governor St
Springfield, IL

Address Standardization

1200 W Lawrence Ave
Springfield, IL 62704 ✓

830 S MacArthur Blvd
Springfield, IL 62704 ✓

1015 W Governor St
Springfield, IL 62704 ✓

Address Locator

(39.793950°, 89.672800°)

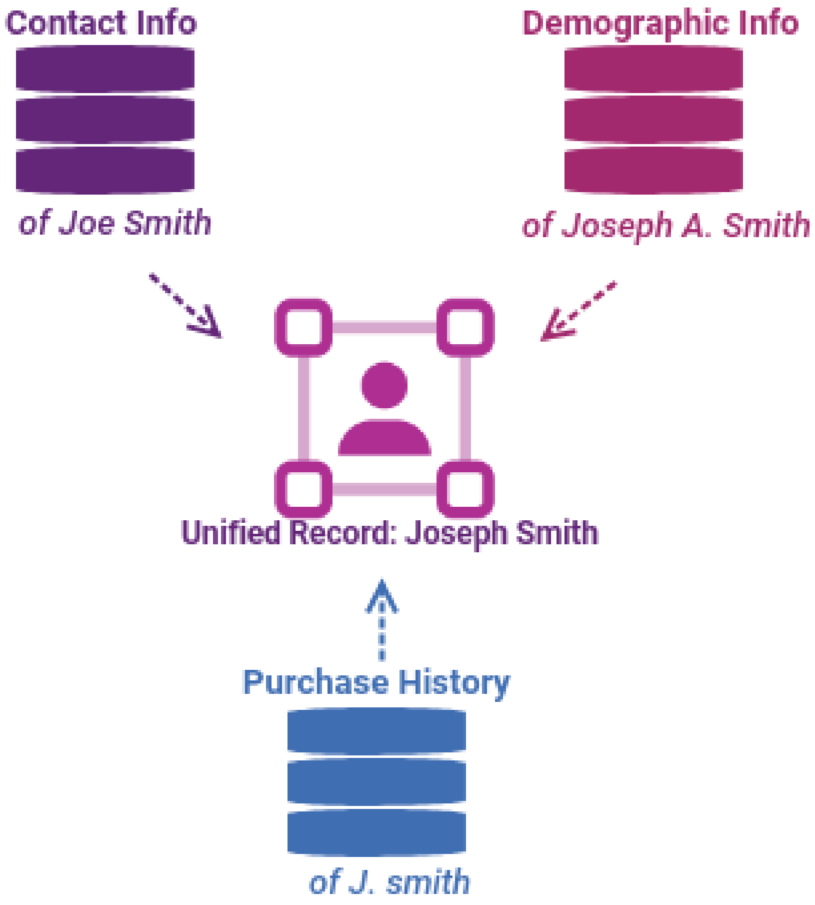
(39.78685°, -89.6682°)

(39.797750°, -89.668357°)

Points on a Map

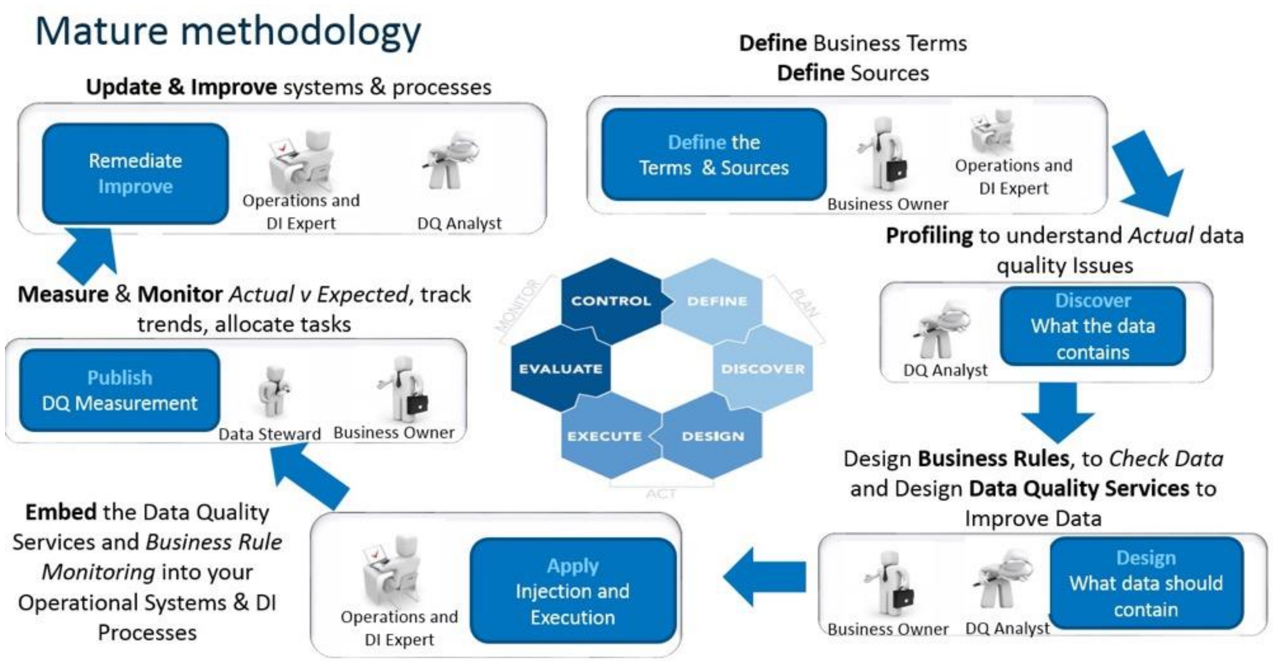
How to improve data quality? (5)

- **Matching or Linking** - identifies and merges matching pieces of information in big data sets. Finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, and databases)



How to improve data quality? (6)

- **Data quality monitoring** - frequent data quality checks are essential. Data quality software in combination with machine learning can automatically detect, report, and correct data variations based on predefined business rules and parameters.



Mature methodology

Define Business Terms
Define Sources

Define the Terms & Sources

Business Owner

Operations and DI Expert

Profiling to understand *Actual* data quality Issues

Discover What the data contains

DQ Analyst

Design **Business Rules**, to *Check Data* and Design **Data Quality Services** to Improve Data

Business Owner

DQ Analyst

Design What data should contain

Apply Injection and Execution

Operations and DI Expert

Embed the Data Quality Services and *Business Rule Monitoring* into your Operational Systems & DI Processes

Measure & **Monitor** *Actual v Expected*, track trends, allocate tasks

Publish DQ Measurement

Data Steward

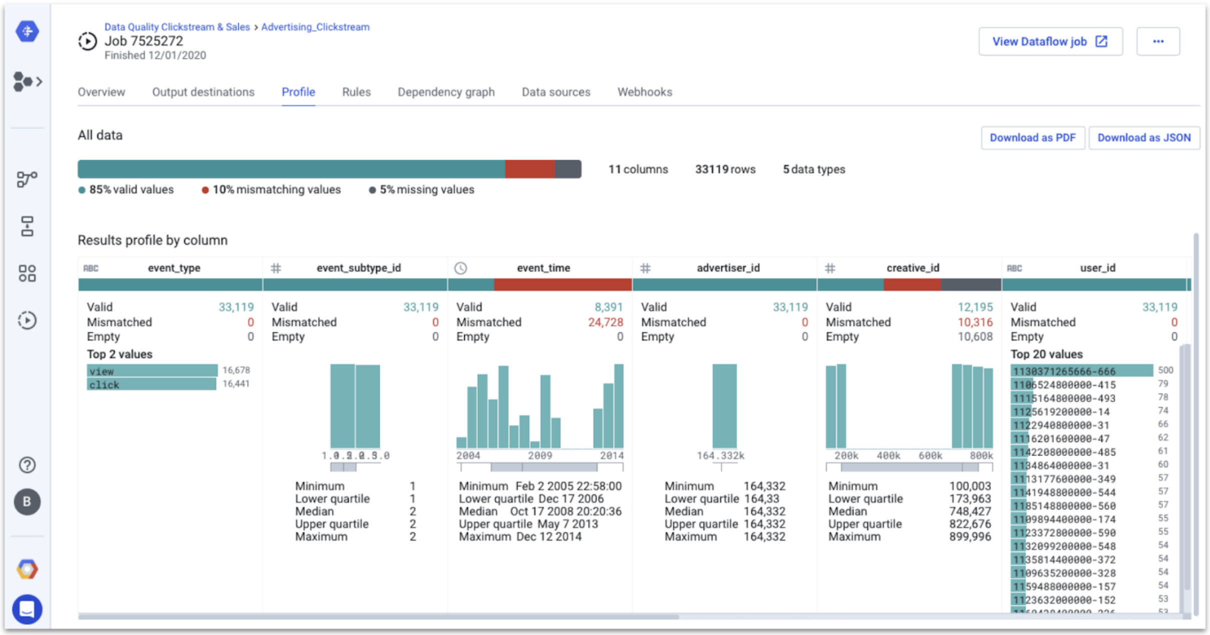
Business Owner

Update & Improve systems & processes

Remediate Improve

Operations and DI Expert

DQ Analyst



Data Quality Clickstream & Sales > Advertising_Clickstream
Job 7525272
Finished 12/01/2020

All data

11 columns 33119 rows 5 data types

● 85% valid values ● 10% mismatching values ● 5% missing values

Results profile by column

	event_type	event_subtype_id	event_time	advertiser_id	creative_id	user_id
Valid	33,119	33,119	8,391	33,119	12,195	33,119
Mismatched	0	0	24,728	0	10,316	0
Empty	0	0	0	0	10,608	0

event_type — Top 2 values

view	16,678
click	16,441

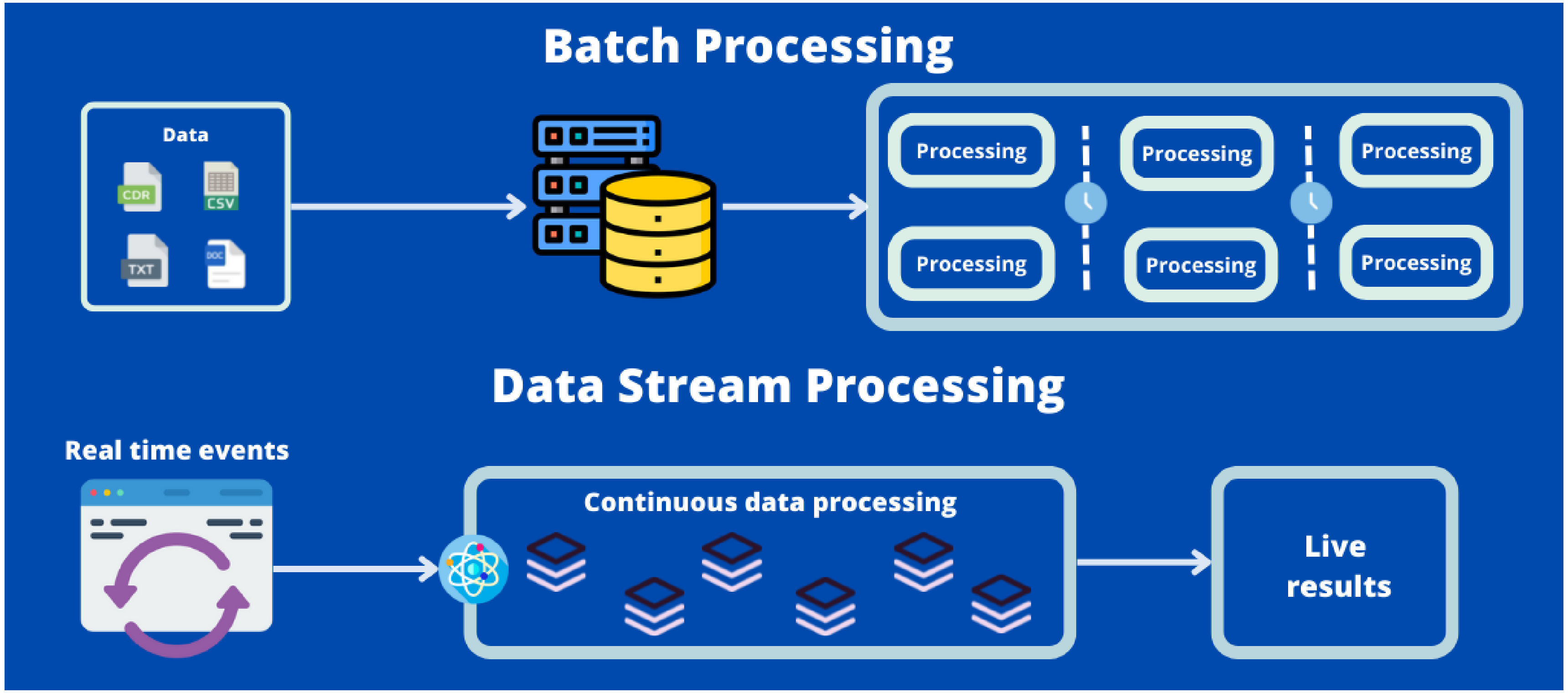
	event_subtype_id	event_time	advertiser_id	creative_id
Minimum	1	Feb 2 2005 22:58:00	164,332	100,003
Lower quartile	1	Dec 17 2006	164,33	173,963
Median	2	Oct 17 2008 20:20:36	164,332	748,427
Upper quartile	2	May 7 2013	164,332	822,676
Maximum	2	Dec 12 2014	164,332	899,996

user_id — Top 20 values

1130371265666-666	500
1106524800000-415	79
1115164800000-493	78
1125619200000-14	74
1122940800000-31	66
1116201600000-47	62
1142208000000-485	61
1134864000000-31	60
1113177600000-349	57
1141948800000-544	57
1185148800000-560	57
1109894400000-174	55
1123372800000-590	55
1132099200000-548	54
1135814400000-372	54
1109635200000-328	54
1159488000000-157	54
1123632000000-152	53

How to improve data quality? (7)

- **Batch and Real time** - Once the data is initially cleansed, an effective data quality framework should be able to deploy the same rules and processes across all applications and data types at scale over time.



What is data security?

- **Data security is the process** of safeguarding digital information throughout its entire life cycle to protect it from corruption, theft, or unauthorized access.
  - It covers everything—hardware, software, storage devices, and user devices; access and administrative controls; and organizations' policies and procedures.
- **Data security uses tools and technologies** that enhance visibility of a company's data and how it is being used. These tools can protect data through processes like data masking, encryption, and redaction of sensitive information. The process also helps organizations streamline their auditing procedures and comply with increasingly stringent data protection regulations.

Data security risks (1)

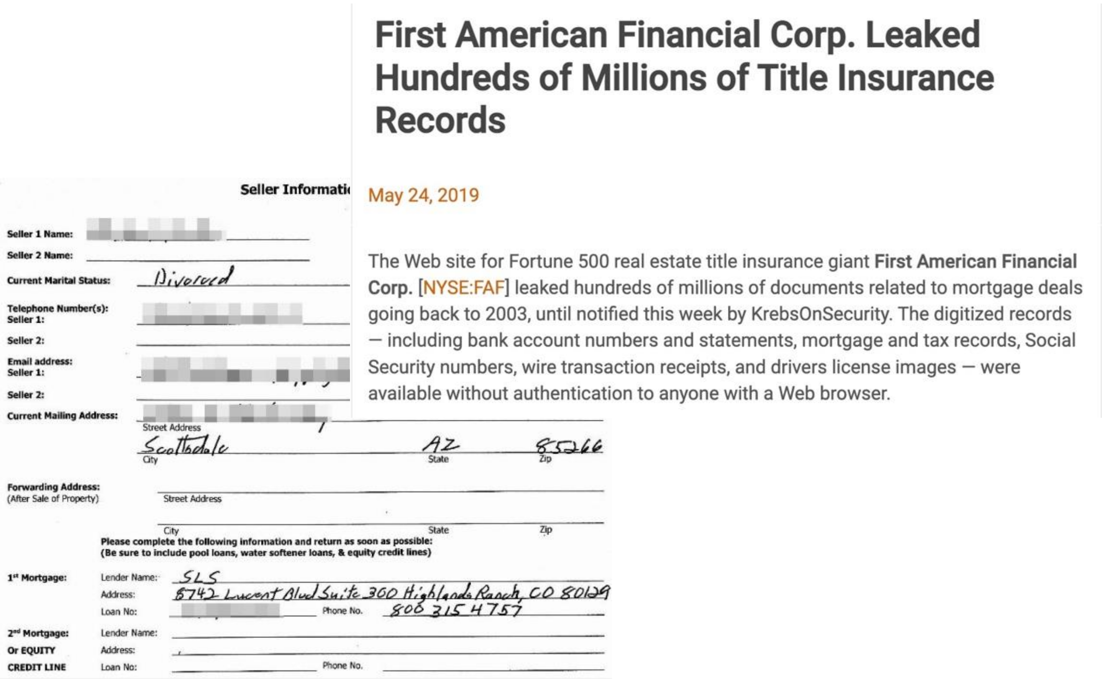
Organizations face an increasingly complex landscape of security threats with cyberattacks being launched by more sophisticated attackers. Some of the biggest risks to data security include:

- **Accidental Data Exposure**

Many data breaches are not a result of hacking but through employees accidentally or negligently exposing sensitive information. Employees can easily lose, share, or grant access to data with the wrong person, or mishandle or lose information because they are not aware of their company’s security policies.

- **Insider Threats**

One of the biggest data security threats to any organization is its own employees. Insider threats are individuals who intentionally or inadvertently put their own organization’s data at risk.



First American Financial Corp. Leaked Hundreds of Millions of Title Insurance Records

May 24, 2019

The Web site for Fortune 500 real estate title insurance giant **First American Financial Corp.** [NYSE:FAF] leaked hundreds of millions of documents related to mortgage deals going back to 2003, until notified this week by KrebsOnSecurity. The digitized records — including bank account numbers and statements, mortgage and tax records, Social Security numbers, wire transaction receipts, and drivers license images — were available without authentication to anyone with a Web browser.

Seller Informati

Seller 1 Name: [redacted]
Seller 2 Name: ____
Current Marital Status: Divorced
Telephone Number(s): Seller 1: [redacted]
Seller 2: ____
Email address: Seller 1: [redacted]
Seller 2: ____
Current Mailing Address: [redacted]
Street Address
Scottsdale AZ 85266
City State Zip

Forwarding Address:
(After Sale of Property)
Street Address
City State Zip

Please complete the following information and return as soon as possible:
(Be sure to include pool loans, water softener loans, & equity credit lines)

1st Mortgage: Lender Name: SLS
Address: 8742 Lucent Blvd Suite 300 Highlands Ranch, CO 80129
Loan No: [redacted] Phone No. 800 315 4757

2nd Mortgage: Or EQUITY CREDIT LINE Lender Name: ____
Address: ____
Loan No: ____ Phone No. ____

Types of Insider Threats

Negligent

Insiders who pose an unintentional threat due to human error and lack of security awareness

Malicious

Current or former employees who abused their access to steal intellectual property for personal gains

Third - Party

Vendors who misuse their access and compromise the security of critical data

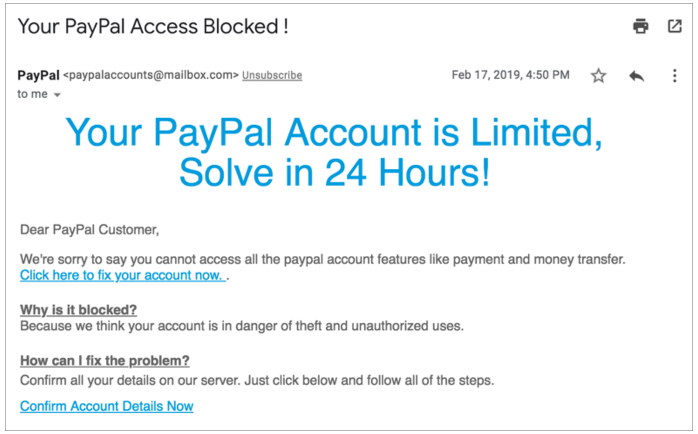
Data security risks (2)

- **Phishing Attacks**

In a phishing attack, a cyber criminal sends messages, typically via email, short message service (SMS), or instant messaging services, that appear to be from a trusted sender. Messages include malicious links or attachments that lead recipients to either download malware or visit a spoofed website that enables the attacker to steal their login credentials or financial information.

- **Malware**

Malicious software is typically spread through email- and web-based attacks. Attackers use malware to infect computers and corporate networks by exploiting vulnerabilities in their software, such as web browsers or web applications. Malware can lead to serious data security events like data theft, extortion, and network damage.



Your PayPal Access Blocked !

PayPal <paypalaccounts@mailbox.com> Unsubscribe Feb 17, 2019, 4:50 PM

to me

Your PayPal Account is Limited, Solve in 24 Hours!

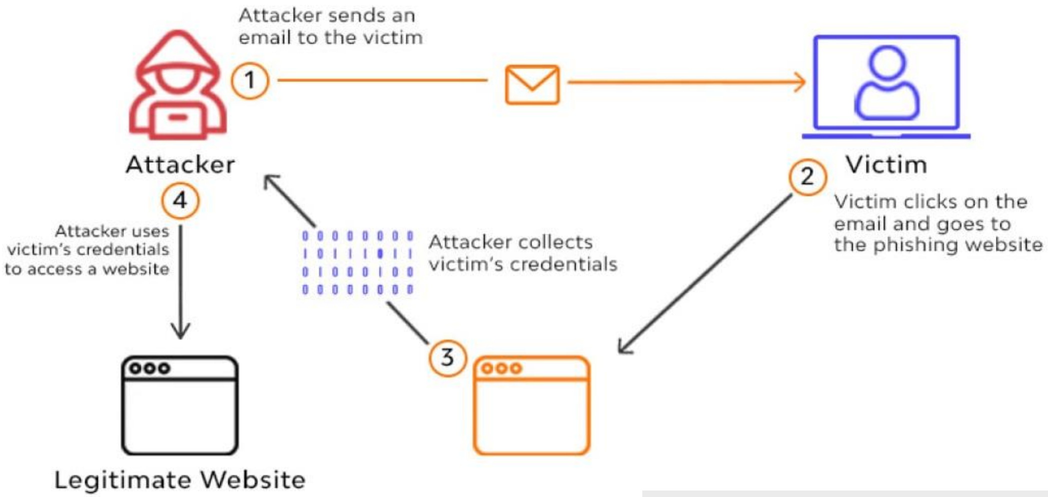
Dear PayPal Customer,

We're sorry to say you cannot access all the paypal account features like payment and money transfer.
Click here to fix your account now. .

Why is it blocked?
Because we think your account is in danger of theft and unauthorized uses.

How can I fix the problem?
Confirm all your details on our server. Just click below and follow all of the steps.

Confirm Account Details Now



Attacker sends an email to the victim

1

Attacker

Victim

2

Victim clicks on the email and goes to the phishing website

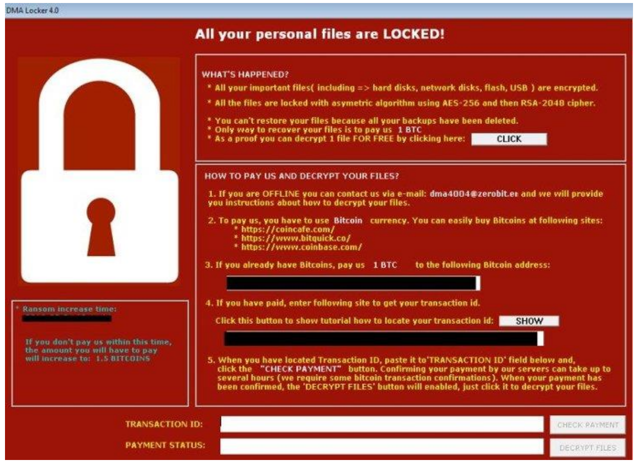
3

Attacker collects victim's credentials

4

Attacker uses victim's credentials to access a website

Legitimate Website



DMA Locker 4.0

All your personal files are LOCKED!

WHAT'S HAPPENED?

- All your important files(including => hard disks, network disks, flash, USB) are encrypted.
- All the files are locked with asymetric algorithm using AES-256 and then RSA-2048 cipher.
- You can't restore your files because all your backups have been deleted.
- Only way to recover your files is to pay us 1 BTC
- As a proof you can decrypt 1 file FOR FREE by clicking here: CLICK

HOW TO PAY US AND DECRYPT YOUR FILES?

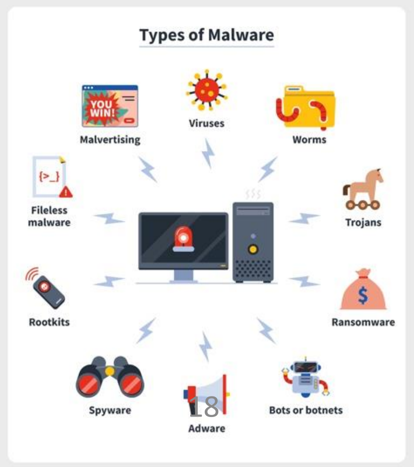
1. If you are OFFLINE you can contact us via e-mail: dma4004@zerobit.er and we will provide you instructions about how to decrypt your files.
2. To pay us, you have to use Bitcoin currency. You can easily buy Bitcoins at following sites:
   - https://coincafe.com/
   - https://www.bitquick.co/
   - https://www.coinbase.com/
3. If you already have Bitcoins, pay us 1 BTC to the following Bitcoin address:
   [redacted]
4. If you have paid, enter following site to get your transaction id.
   Click this button to show tutorial how to locate your transaction id: SHOW
   [redacted]
5. When you have located Transaction ID, paste it to'TRANSACTION ID' field below and, click the "CHECK PAYMENT" button. Confirming your payment by our servers can take up to several hours (we require some bitcoin transaction confirmations). When your payment has been confirmed, the 'DECRYPT FILES' button will enabled, just click it to decrypt your files.

- Ransom increase time:
[redacted]

If you don't pay us within this time, the amount you will have to pay will increase to: 1.5 BITCOINS

TRANSACTION ID: CHECK PAYMENT

PAYMENT STATUS: DECRYPT FILES



Types of Malware

How to achieve data security? (1)

Organizations can use a wide range of data security methods to safeguard their data, devices, networks, systems, and users. Some of the most common types of data security, which organizations should look to combine to ensure they have the best possible strategy, include:

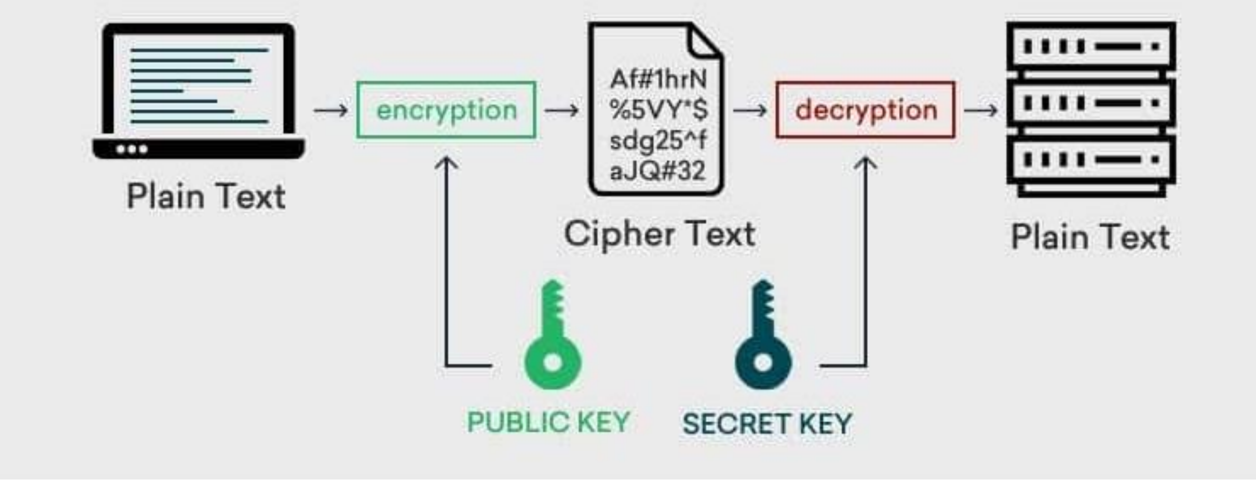
- **Regulation measures:**
  - Monitor user behavior and manage accounts
  - Enforce security policies
  - Provide security awareness training
  - Conduct proactive network monitoring
  - Consistently apply necessary software patches and updates on all systems
  - Implement malware protection to prevent attacks

How to achieve data security? (2)

- **Encryption**

  Data encryption is the use of algorithms to scramble data and hide its true meaning. Encrypting data ensures messages can only be read by recipients with the appropriate decryption key. This is crucial, especially in the event of a data breach, because even if an attacker manages to gain access to the data, they will not be able to read it without the decryption key.

  Data encryption also involves the use of solutions like tokenization, which protects data as it moves through an organization's entire IT infrastructure.



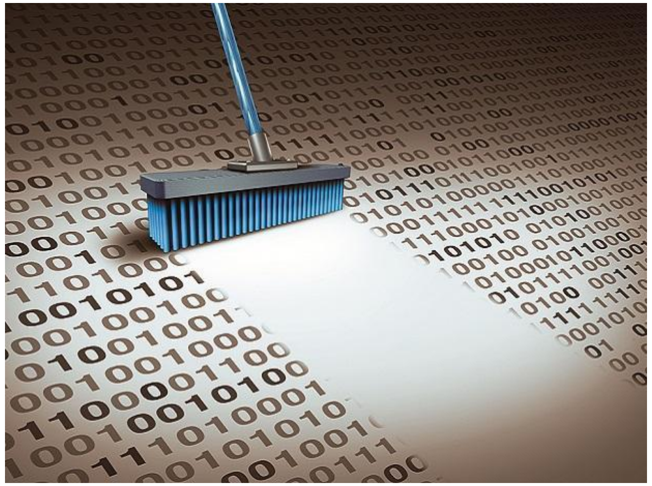
How to achieve data security? (3)

- **Data Erasure**

  There will be occasions in which organizations no longer require data and need it permanently removed from their systems. Data erasure is an effective data security management technique that removes liability and the chance of a data breach occurring.

  This is a software-based technique for **effectively overwriting electronically stored data** with random binary information as per a given standard.

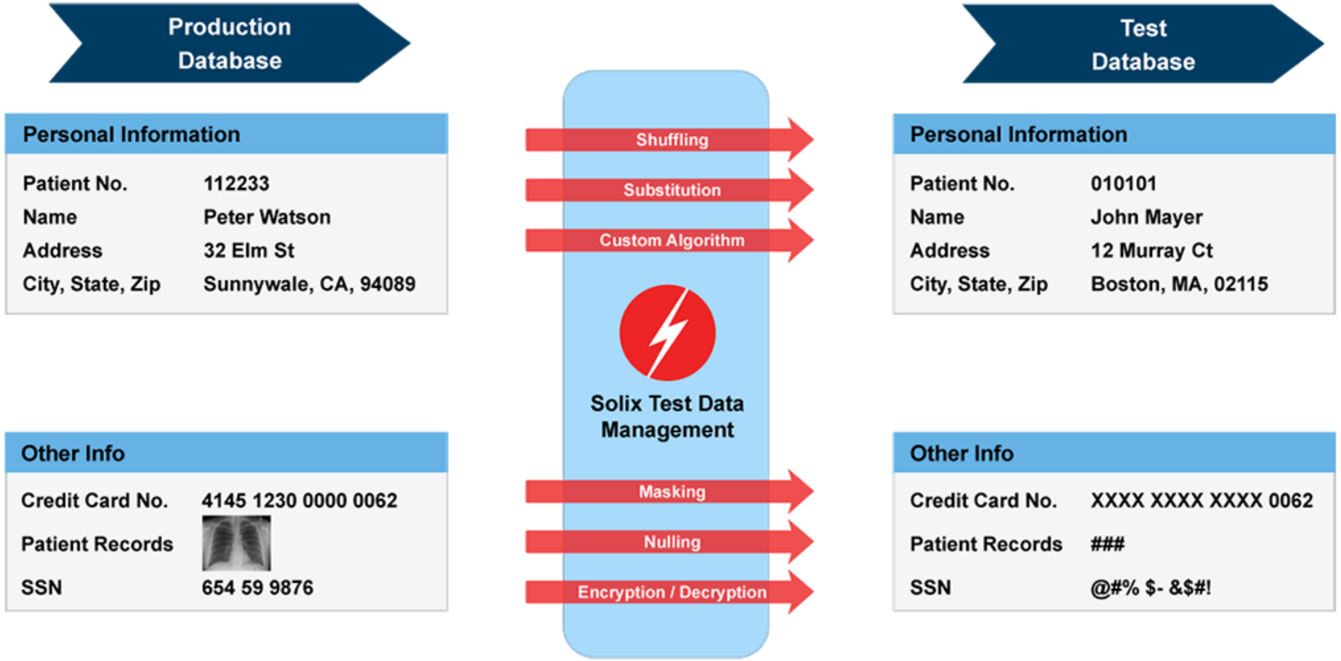
  This is a necessary part of any secure hardware disposal or decommissioning process.



How to achieve data security? (4)

- **Data Masking**

Data masking enables an organization to hide data by obscuring and replacing specific letters or numbers. This process is a form of encryption that renders the data useless should a hacker intercept it. The original message can only be uncovered by someone who has the code to decrypt or replace the masked characters.



Production Database

Personal Information

Patient No.	112233
Name	Peter Watson
Address	32 Elm St
City, State, Zip	Sunnywale, CA, 94089

Other Info

Credit Card No.	4145 1230 0000 0062
Patient Records	
SSN	654 59 9876

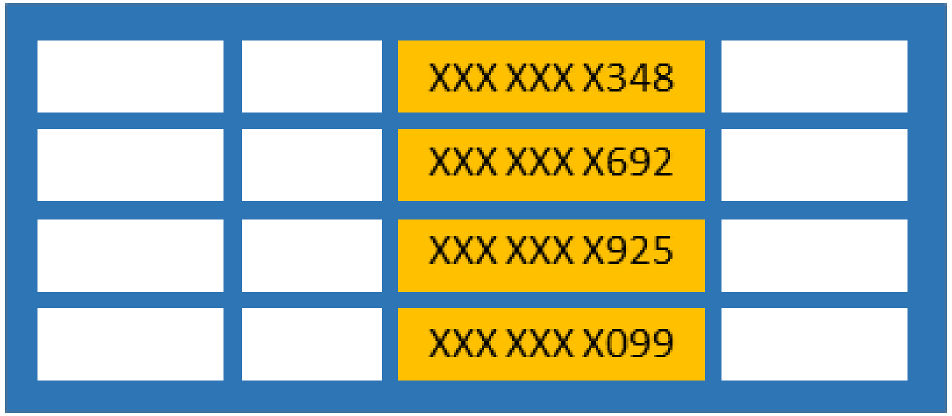
Test Database

Personal Information

Patient No.	010101
Name	John Mayer
Address	12 Murray Ct
City, State, Zip	Boston, MA, 02115

Other Info

Credit Card No.	XXXX XXXX XXXX 0062
Patient Records	###
SSN	@#% $- &$#!

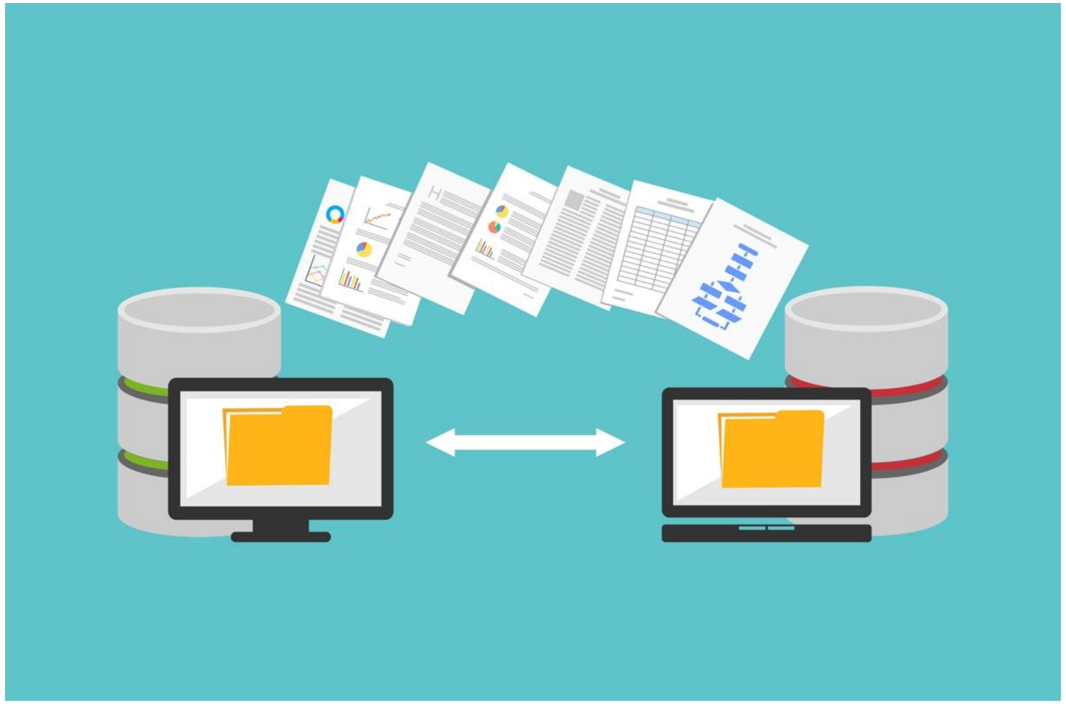


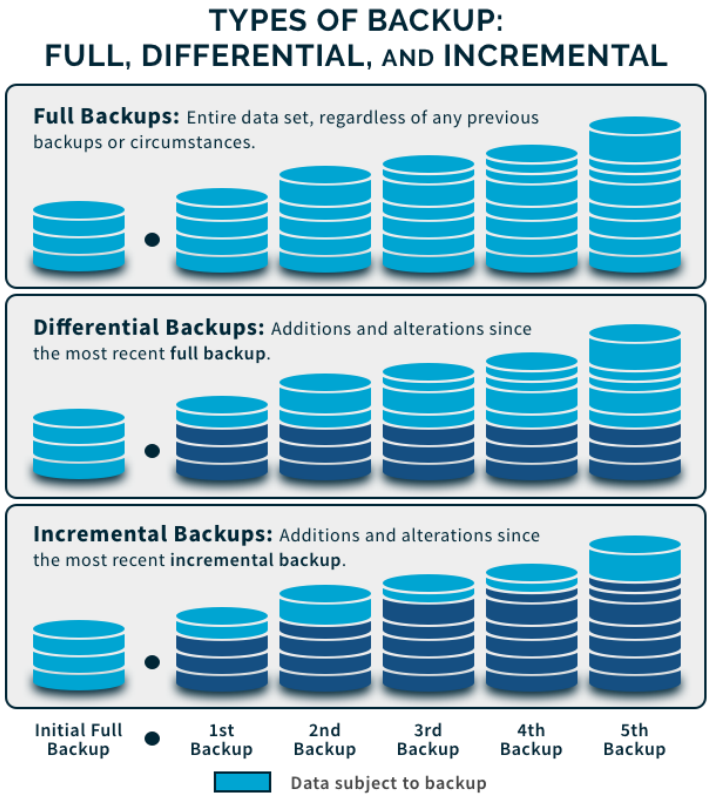
		XXX XXX X348	
		XXX XXX X692	
		XXX XXX X925	
		XXX XXX X099	

How to achieve data security? (5)

- **Data Resiliency**

  Organizations can mitigate the risk of accidental destruction or loss of data by creating backups or copies of their data. Data backups are vital to protecting information and ensuring it is always available. This is particularly important during a data breach or ransomware attack, ensuring the organization can restore a previous backup.





TYPES OF BACKUP:
FULL, DIFFERENTIAL, AND INCREMENTAL

Full Backups: Entire data set, regardless of any previous backups or circumstances.

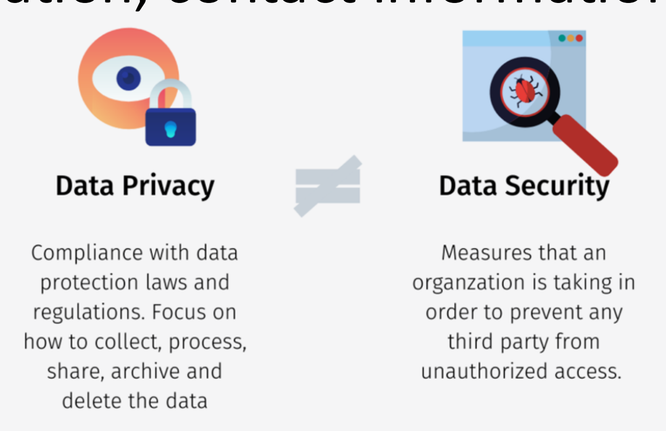
Differential Backups: Additions and alterations since the most recent **full backup**.

Incremental Backups: Additions and alterations since the most recent **incremental backup**.

Data privacy

Data privacy generally means the ability of a person to determine for themselves when, how, and to what extent personal information about them is shared with or communicated to others, or being used.

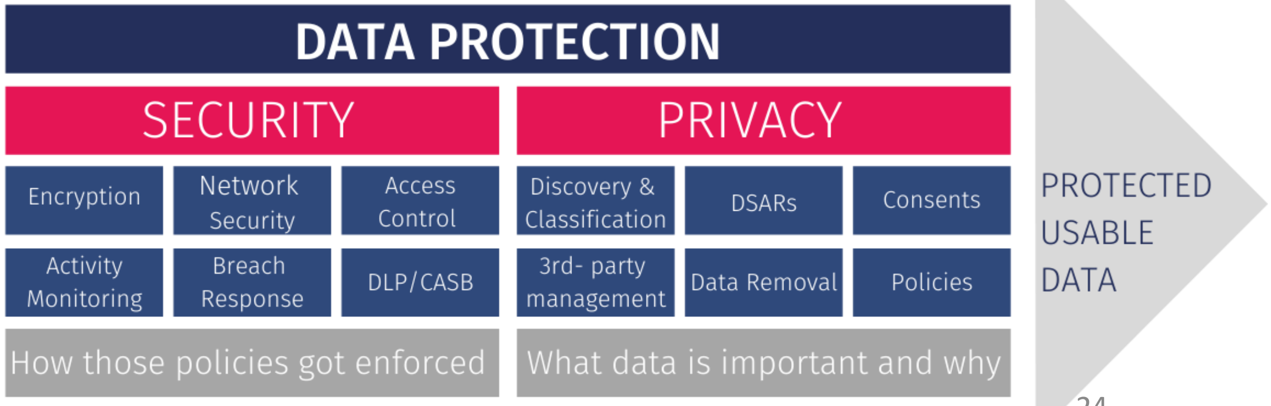
This personal information can be one's name, location, contact information, or online or real-world behavior.



Data Privacy ≠ **Data Security**

Data Privacy: Compliance with data protection laws and regulations. Focus on how to collect, process, share, archive and delete the data

Data Security: Measures that an organzation is taking in order to prevent any third party from unauthorized access.



DATA PROTECTION

SECURITY			PRIVACY		
Encryption	Network Security	Access Control	Discovery & Classification	DSARs	Consents
Activity Monitoring	Breach Response	DLP/CASB	3rd- party management	Data Removal	Policies
How those policies got enforced			What data is important and why		

PROTECTED USABLE DATA

Data privacy’s focus

- Data Privacy or Information privacy is a part of the data protection area that deals with the proper handling of data focusing on compliance with data protection regulations.
- Data Privacy is centered around how data should be collected, stored, managed, and shared with any third parties, as well as compliance with the applicable privacy laws (such as California Consumer Privacy Act- CCPA or General Data Protection Regulation GDPR).

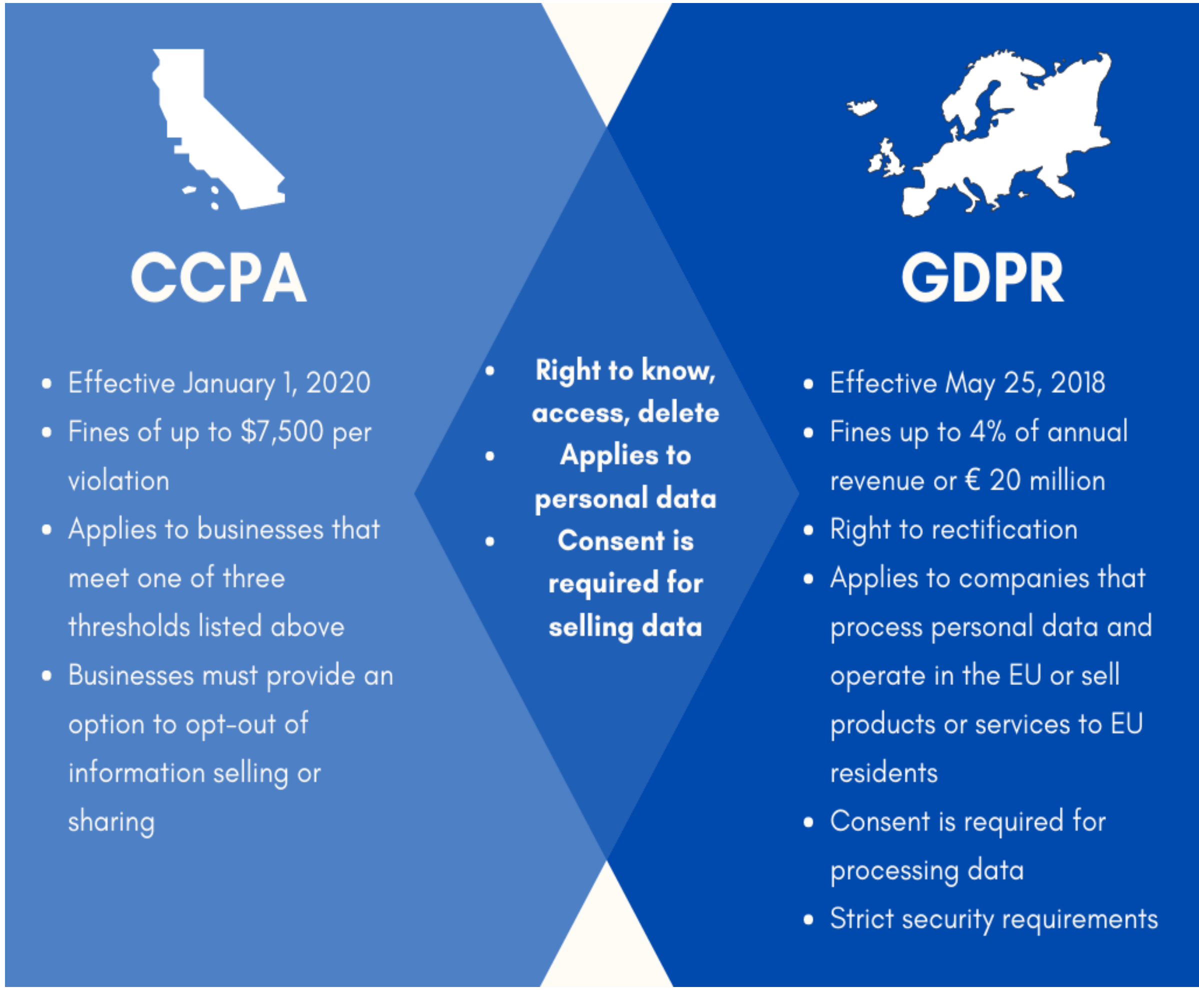
Data privacy regulations

- **General Data Protection Regulations (GDPR) in Europe**

The 2018 GDPR legislation is a piece of law that protects the personal data of European citizens. It aims to increase people's control and privacy rights over their data and places strict controls on how organizations process that information. GDPR ensures that organizations process personal data securely and protect it from unauthorized processing, accidental loss, damage, and destruction.

- **California Consumer Privacy Act (CCPA) in California, US**

The CCPA aims to give California's consumers more control over how businesses collect their personal data. This includes the right to know what information a business has and how it is shared or used, the right to delete that information, the right to opt out of that data being sold to third parties, and the right to avoid discrimination for exercising these CCPA rights. Organizations must provide consumers with notice of their privacy practices.



CCPA

- Effective January 1, 2020
- Fines of up to $7,500 per violation
- Applies to businesses that meet one of three thresholds listed above
- Businesses must provide an option to opt-out of information selling or sharing

- **Right to know, access, delete**
- **Applies to personal data**
- **Consent is required for selling data**

GDPR

- Effective May 25, 2018
- Fines up to 4% of annual revenue or € 20 million
- Right to rectification
- Applies to companies that process personal data and operate in the EU or sell products or services to EU residents
- Consent is required for processing data
- Strict security requirements

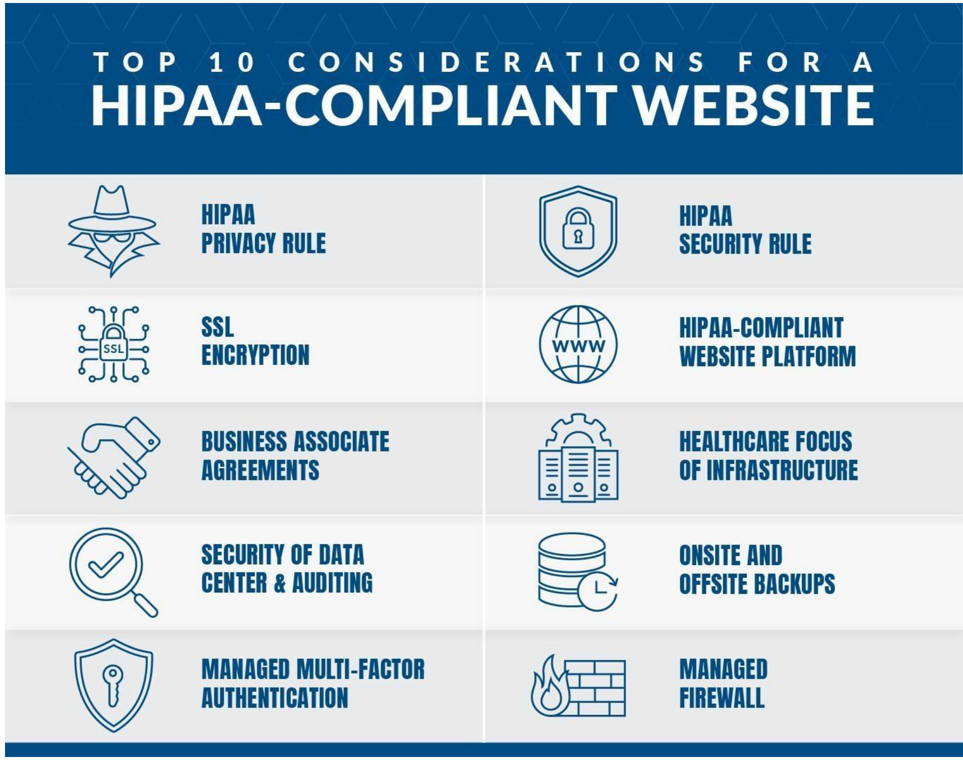
Data privacy regulations

- **Health Insurance Portability and Accountability Act (HIPAA) in US**

HIPAA of 1996 is a United States Act of Congress enacted federal law. It protects patients' health data from being exposed without their consent or knowledge. HIPAA contains a privacy rule, which addresses the disclosure and use of patient information and ensures that data is properly protected. It also has a security rule, which protects all individually identifiable health information that an organization creates, maintains, receives, or transmits electronically.

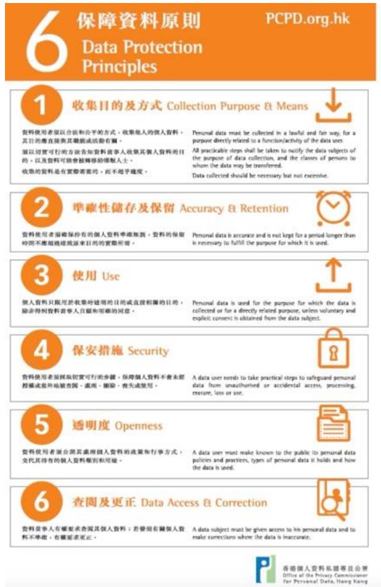
- **Personal Data (Privacy) Ordinance (PDPO) in HK**

The purpose of this 1995 ordinance is to protect the privacy rights of a person in regard to his personal data, i.e., the Data Subject: the information which relates to a living person and can be used to identify that person and it exists in a form in which access or processing is practicable. Examples of data subject protected by this ordinance include name, address, phone number, identity card number, photo, medical record and employment records. The data user, who collects, holds, or process this data is liable for any unlawful or wrongful use of this data.



TOP 10 CONSIDERATIONS FOR A
HIPAA-COMPLIANT WEBSITE

HIPAA PRIVACY RULE	HIPAA SECURITY RULE
SSL ENCRYPTION	HIPAA-COMPLIANT WEBSITE PLATFORM
BUSINESS ASSOCIATE AGREEMENTS	HEALTHCARE FOCUS OF INFRASTRUCTURE
SECURITY OF DATA CENTER & AUDITING	ONSITE AND OFFSITE BACKUPS
MANAGED MULTI-FACTOR AUTHENTICATION	MANAGED FIREWALL



6 保障資料原則 Data Protection Principles PCPD.org.hk

1 收集目的及方式 Collection Purpose & Means

資料使用者須以合法和公平的方式，收集他人的個人資料，其目的應直接與其職能或活動有關。
須以切實可行的方法告知資料當事人收集其個人資料的目的，以及資料可能會被轉移給哪類人士。
收集的資料是有實際需要的，而不超乎適度。

Personal data must be collected in a lawful and fair way, for a purpose directly related to a function/activity of the data user.
All practicable steps shall be taken to notify the data subjects of the purpose of data collection, and the classes of persons to whom the data may be transferred.
Data collected should be necessary but not excessive.

2 準確性儲存及保留 Accuracy & Retention

資料使用者須確保持有的個人資料準確無誤，資料的保留時間不應超過達致原來目的的實際所需。

Personal data is accurate and is not kept for a period longer than is necessary to fulfill the purpose for which it is used.

3 使用 Use

個人資料只限用於收集時述明的目的或直接相關的目的，除非得到資料當事人自願和明確的同意。

Personal data is used for the purpose for which the data is collected or for a directly related purpose, unless voluntary and explicit consent is obtained from the data subject.

4 保安措施 Security

資料使用者須採取切實可行的步驟，保障個人資料不會未經授權或意外地被查閱、處理、刪除、喪失或使用。

A data user needs to take practical steps to safeguard personal data from unauthorised or accidental access, processing, erasure, loss or use.

5 透明度 Openness

資料使用者須公開其處理個人資料的政策和行事方式，交代其持有的個人資料類別和用途。

A data user must make known to the public its personal data policies and practices, types of personal data it holds and how the data is used.

6 查閱及更正 Data Access & Correction

資料當事人有權要求查閱其個人資料；若發現有關個人資料不準確，有權要求更正。

A data subject must be given access to his personal data and to make corrections where the data is inaccurate.

香港個人資料私隱專員公署
Office of the Privacy Commissioner for Personal Data, Hong Kong

- The six data protection principles form the base of the Ordinance.
- Data users must comply with the six data protection principles in the collection, holding, accuracy, retention period, security, privacy policy and access to and correction of personal data.

香港個人資料私隱專員公署
Office of the Privacy Commissioner for Personal Data, Hong Kong

Why is data privacy regulation important?

- **For individuals**

Privacy laws around the world aim to give back individuals control over their data, empowering them to know how their data is being used, by whom and why, giving them control over how their personal data is being processed and used.

Organizations that collect personal data are obligated to respond to those questions and manage personal data in a compliant way. According to Gartner’s predictions for the future of privacy, privacy is today what “organic” or “cruelty-free” was in the past decade.

- **From A Business Perspective**

Businesses can not operate without processing personal data in some way. However, in order to stay compliant, companies now have to manage personal data in a transparent and compliant way, be accountable for personal data they process, and adhere to privacy principles.

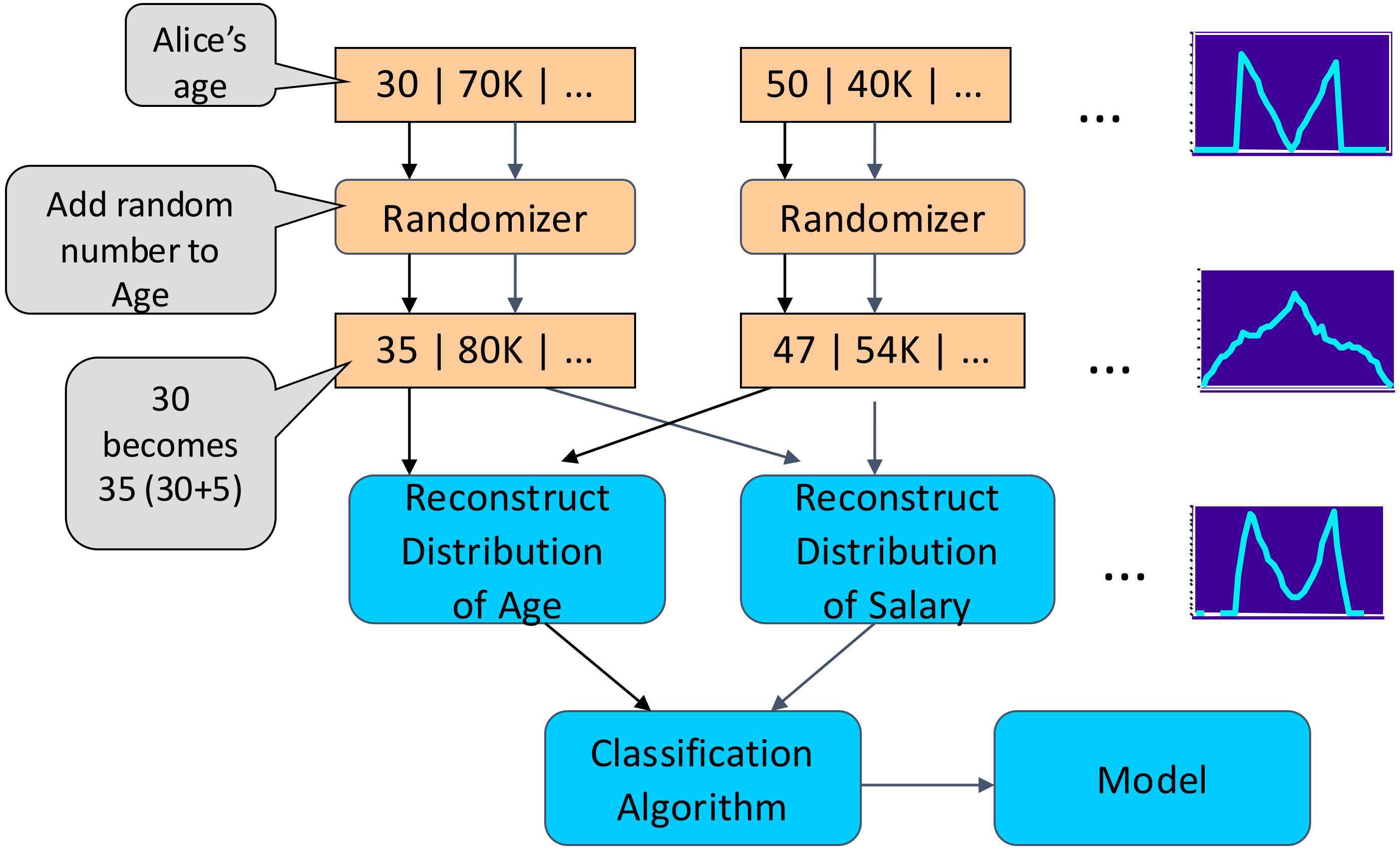
Otherwise, they risk huge regulatory fines, loss of customers’ trust, investor appeal, and data breaches.

Importance of transparency

- In this age of **data economy**, true company value lies in the collected personal data. This means data is an asset worthy of protecting and keeping.
- What companies keep forgetting is that the personal data of individuals processed by the companies are only borrowed.
- Privacy laws enable individuals to exercise their rights, such as the **right to be forgotten**, and in certain circumstances, **individuals can take back ownership of their data.**
- In order for companies to keep the data and keep the trust, they will have to demonstrate transparency by openly communicating how they process and manage personal data.

Privacy preserving methods: data perturbation

- Perturb data with value distortion
    - User provides $x_i + r$ instead of x_i
    - r is a random value
        - Uniform, uniform distribution between $[-\alpha, \alpha]$
        - Gaussian, normal distribution with $\mu = 0$, σ



Alice's age

30 | 70K | ...

50 | 40K | ...

...

Add random number to Age

Randomizer

Randomizer

35 | 80K | ...

47 | 54K | ...

...

30 becomes 35 (30+5)

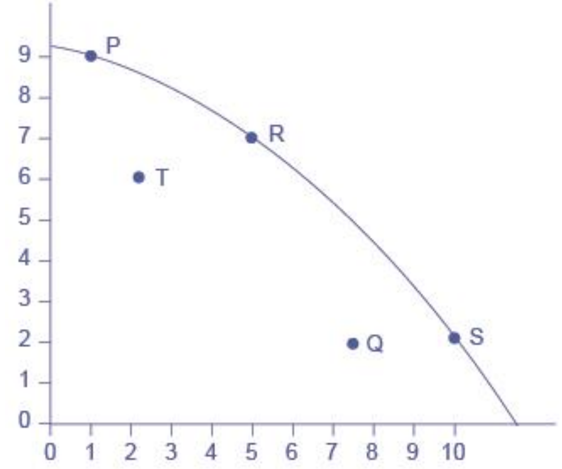
Reconstruct Distribution of Age

Reconstruct Distribution of Salary

...

Classification Algorithm

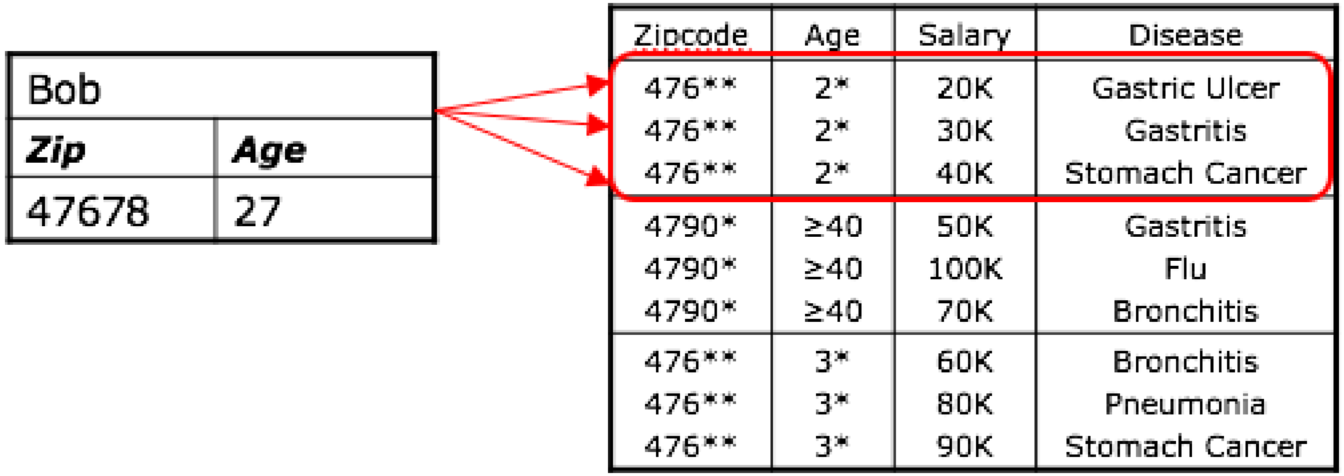
Model



Privacy-performance tradeoff
More noises
higher privacy, lower performance

Privacy preserving methods: anonymity

- A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k−1 other individuals whose information also appear in the release.
- Hiding a person among k of them.



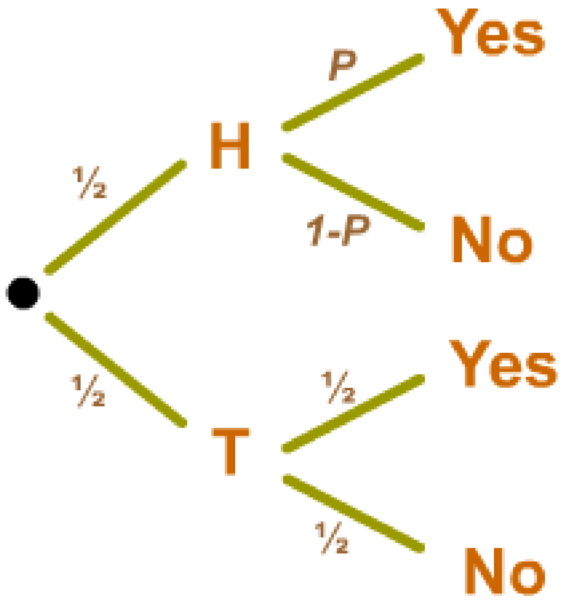
Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Privacy preserving methods: randomized response

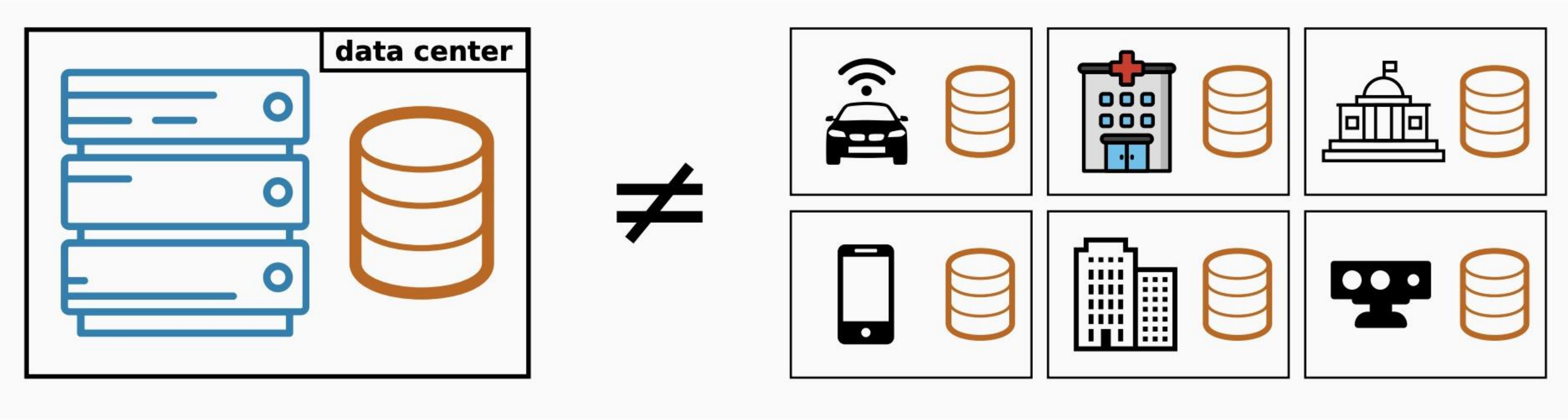
- Randomized response: allows respondents to respond to sensitive issues (such as criminal behavior) while maintaining confidentiality.
  - For example, in an interview, “did you go to the bar last weekend?”.
  - Before they answer, they flip a coin. They are then instructed to answer “yes” or “no” randomly if the coin comes up tails, and truthfully, if it comes up heads.



If for example, we get 60% of “yes” and 40% of “no”, then we kind of know the true “yes” rate.
Solving this equation: 0.25+0.5p = 0.6, 0.25+0.5(1-p) = 0.4

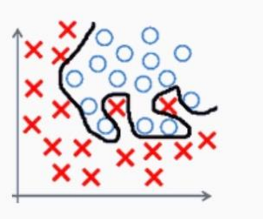
Privacy preserving methods: federated learning (1)

- The standard setting in Machine Learning (ML) considers a centralized dataset processed in a tightly integrated system.
- But in the real world data is often decentralized across many parties.



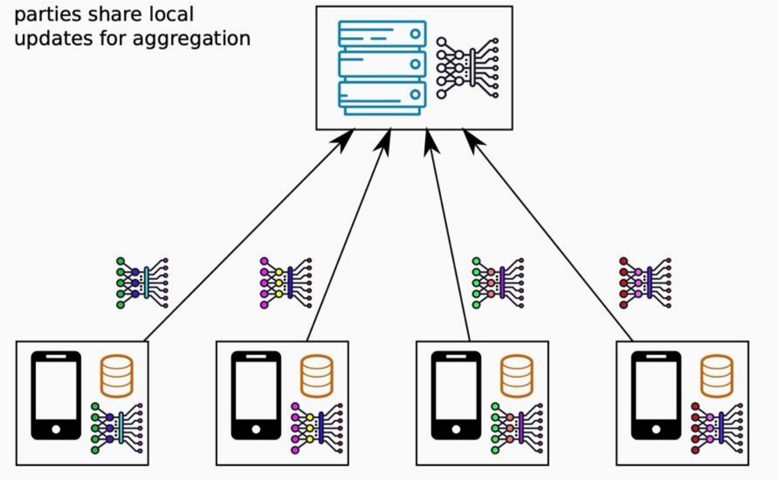
Privacy preserving methods: federated learning (2)

- Sending the data may be too costly
  - Self-driving cars are expected to generate several TBs of data a day
  - Some wireless devices have limited bandwidth/power
- Data may be considered too sensitive
  - We see a growing public awareness and regulations on data privacy
  - Keeping control of data can give a competitive advantage in business and research
- The local dataset may be too small and biased
  - Bad predictive performance (e.g., due to overfitting)
  - Non-statistically significant results (e.g., medical studies)
  - Not representative of the target distribution



Privacy preserving methods: federated learning (3)

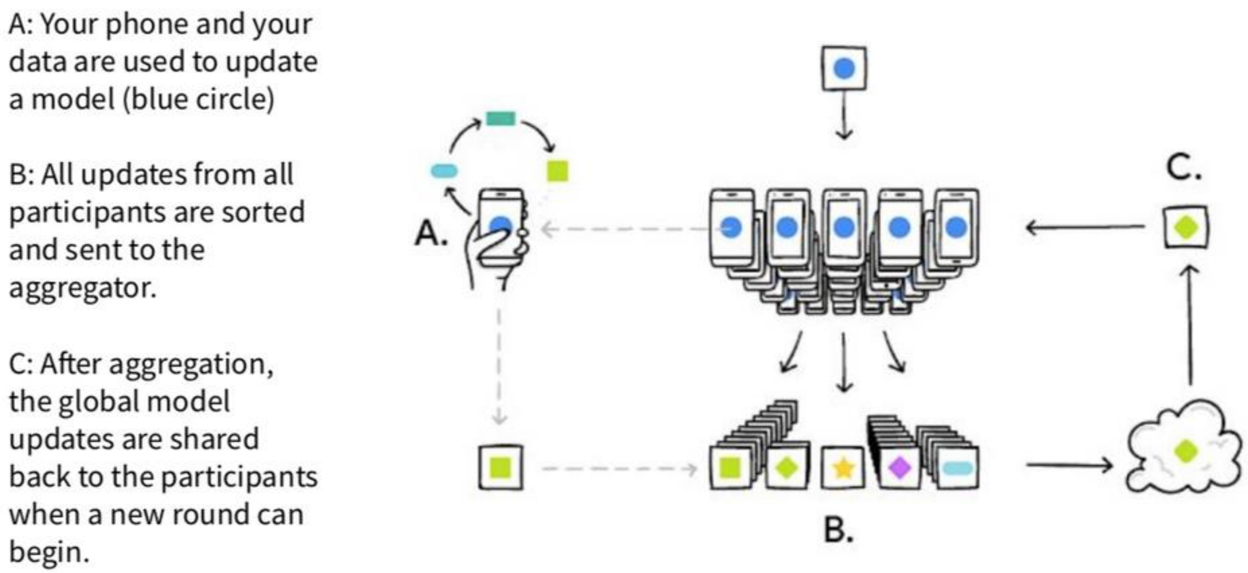
- Federated learning is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them.
- This approach stands in contrast to traditional centralized machine learning techniques where all the local datasets are uploaded to one server, as well as to more classical decentralized approaches which often assume that local data samples are identically distributed.



In 2016, the term FL was first coined by Google researchers

Privacy preserving methods: federated learning (4)

- Federated learning enables multiple agents to build a common, robust machine learning model without sharing data, thus allowing to address critical issues such as data privacy, data security, data access rights and access to heterogeneous data. Its applications are spread over a number of industries including defense, IoT, and pharmaceutics.



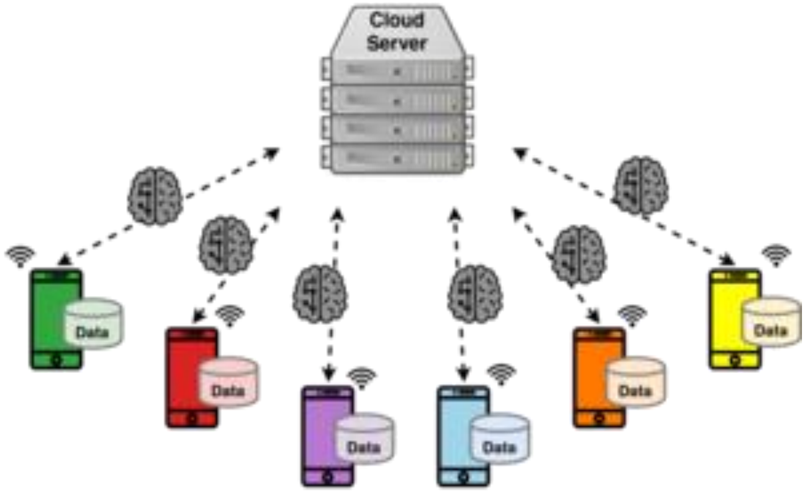
A: Your phone and your data are used to update a model (blue circle)

B: All updates from all participants are sorted and sent to the aggregator.

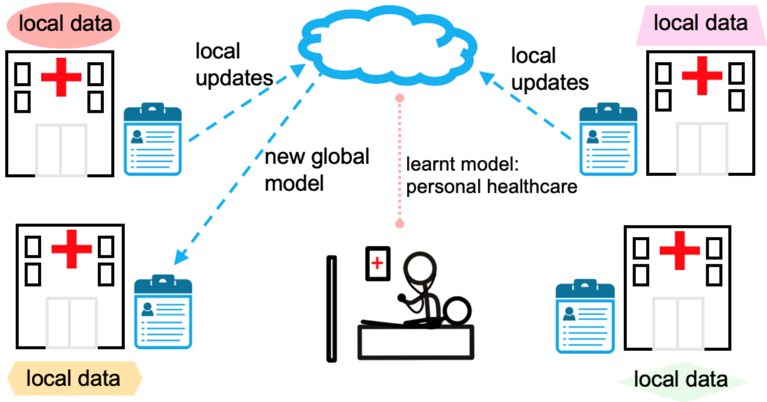
C: After aggregation, the global model updates are shared back to the participants when a new round can begin.

Privacy preserving methods: federated learning (5)

- Cross-device federated learning
  - Smart phones, apps, IoT, edge devices
  - Massive number of parties (up to 10^{10})
  - Small dataset per party (could be size 1)
  - Limited availability and reliability
  - Some parties may be malicious

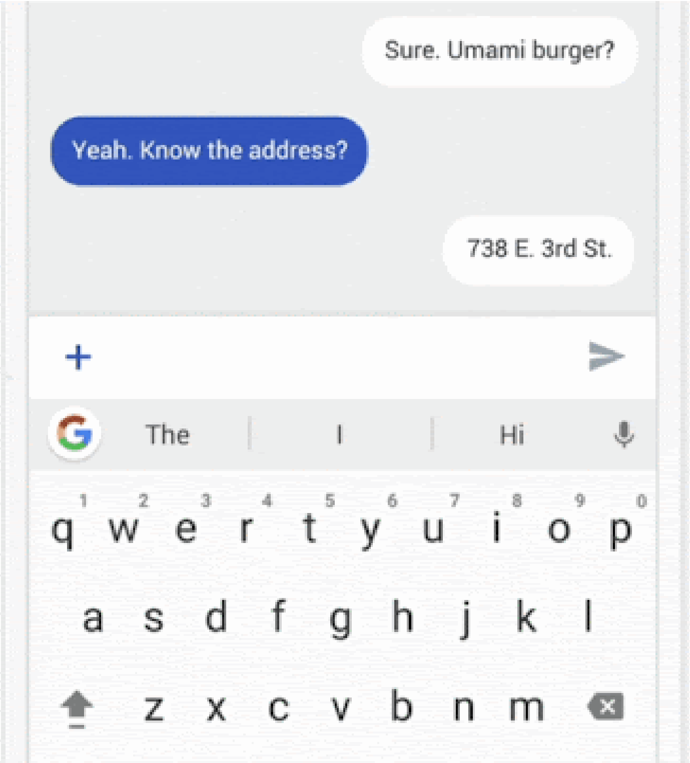


- Cross-silo federated learning
  - Institutions, organizations, hospitals, etc.
  - Small number of parties (2-100)
  - Medium to large dataset per party
  - Reliable parties, almost always available
  - Parties are typically honest



Privacy preserving methods: federated learning (6)

- Examples
  - Google’s Gboard on Android, next typing word prediction federated learning model



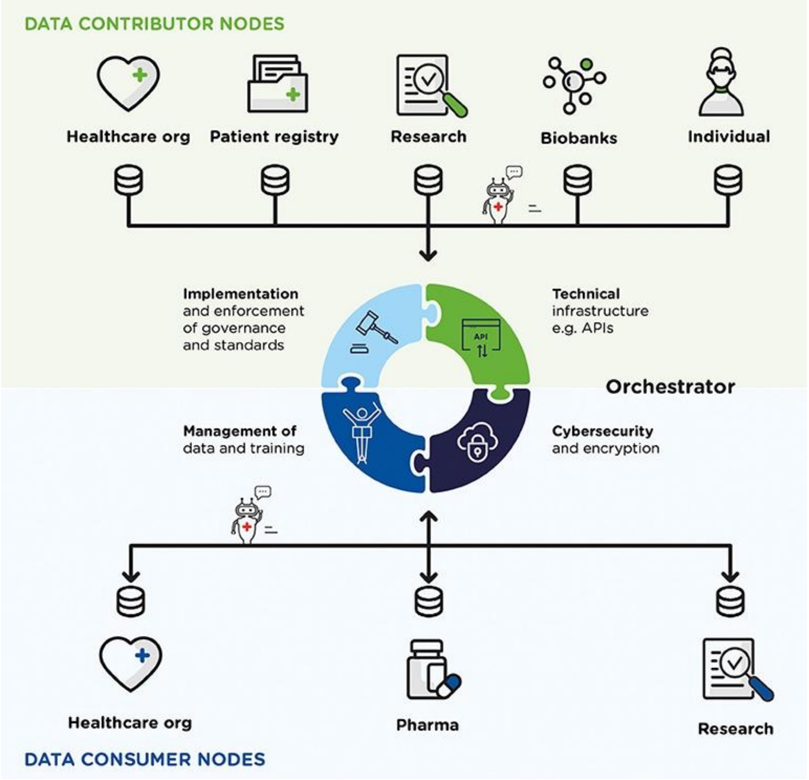
Sure. Umami burger?

Yeah. Know the address?

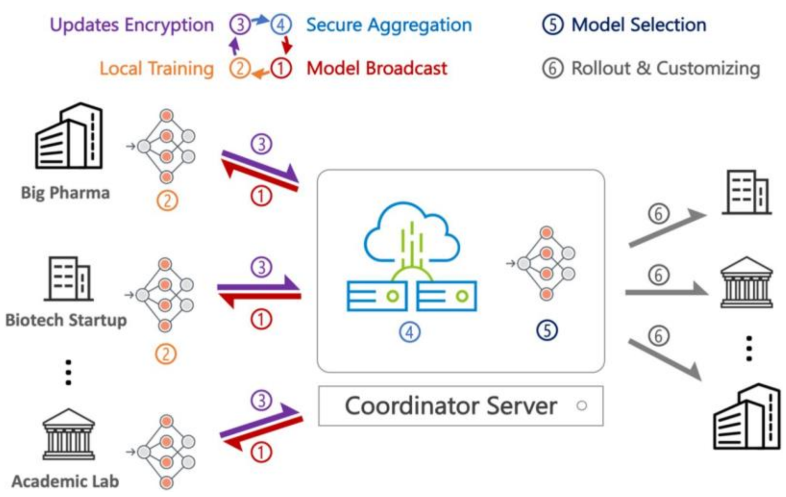
738 E. 3rd St.

Privacy preserving methods: federated learning (7)

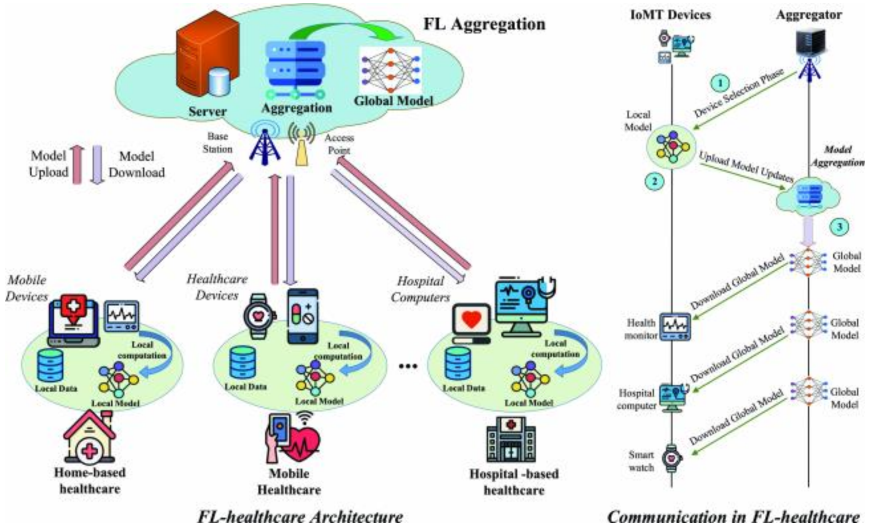
- Examples in healthcare
  - Healthcare data are scarce, sensitive, and distributed
  - Data-driven medicine (e.g., drug discovery: the process by which new candidate medications are discovered).
  - Digital healthcare (digital care programs, technologies with health, healthcare, living, and society to enhance the efficiency of healthcare delivery and to make medicine more personalized and precise).


DATA CONTRIBUTOR NODES
Healthcare org
Patient registry
Research
Biobanks
Individual
Implementation and enforcement of governance and standards
Technical infrastructure e.g. APIs
API
Orchestrator
Management of data and training
Cybersecurity and encryption
Healthcare org
Pharma
Research
DATA CONSUMER NODES

Healthcare data



Drug discovery



Digital healthcare

Privacy preserving methods: federated learning (8)

- Challenges
  - Statistical and resource heterogeneity
  - Personalization
  - Communication efficiency
  - Privacy preserving

Thanks for your attention!

Appendix

1. https://dataprivacymanager.net/5-things-you-need-to-know-about-data-privacy/

2. https://www.fortinet.com/resources/cyberglossary/data-security#:~:text=Data%20security%20is%20the%20process,and%20organizations'%20policies%20and%20procedures.

3. https://www.heavy.ai/technical-glossary/data-quality