COMP 7990 Principles and Practices of Data Analytics

Data Security and Privacy

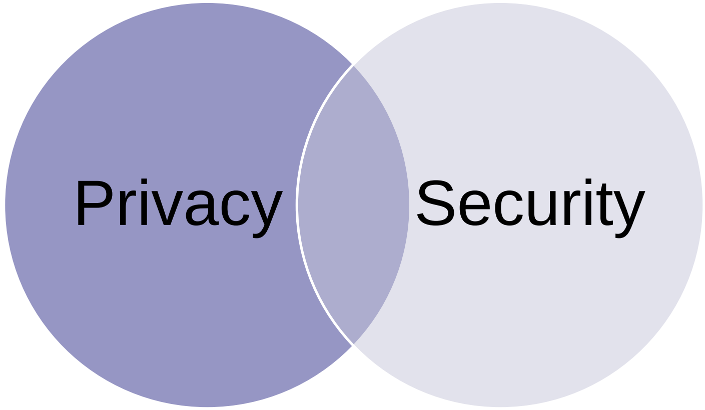
Dr. Eric Zhang

Content

- Security
  - CIA triad
  - Basics of Security
  - Case Study: SQL Injection
- Privacy
  - Legal aspect and compliance
  - Privacy in different processes of data mining

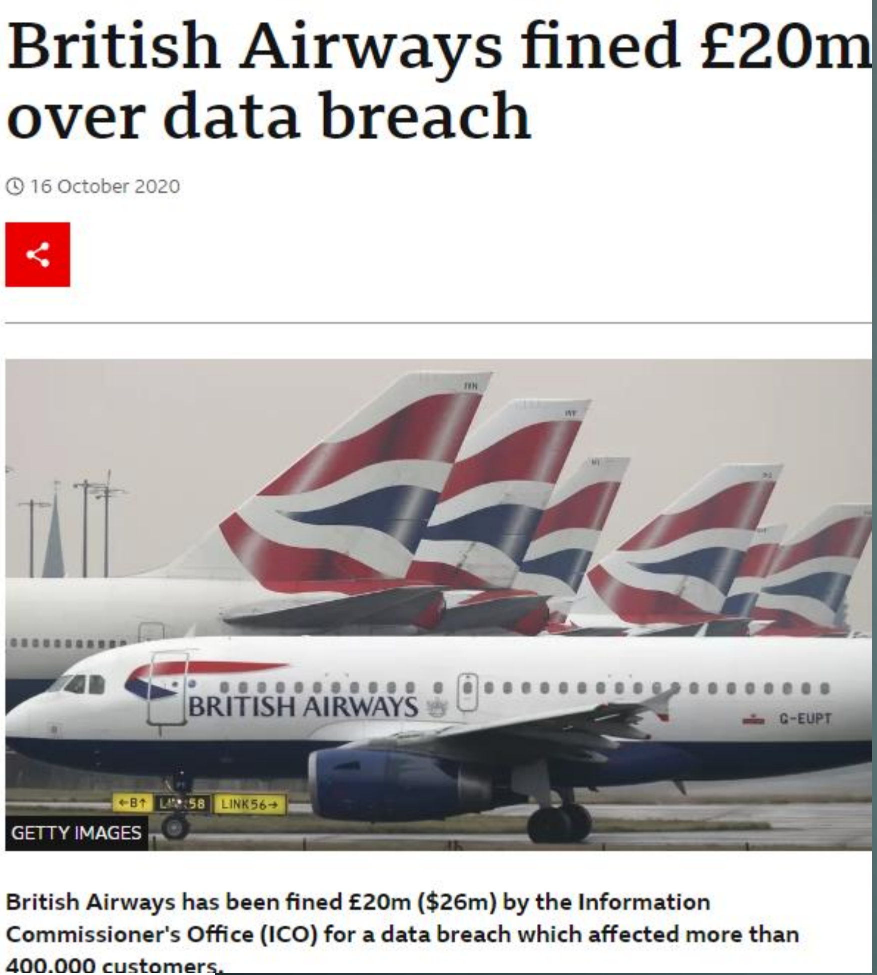
Data Security and Privacy

- ❑ Data Security and Privacy are two very similar but different concepts
- ❑ They are both very important...



Data Breach

- A major European airline suffered a GDPR reportable breach. The breach was reportedly caused by payment application security vulnerabilities exploited by attackers, who harvested more than 400,000 customer payment records. The airline was fined 20 million pounds as a result by the privacy regulator.



British Airways fined £20m over data breach

16 October 2020

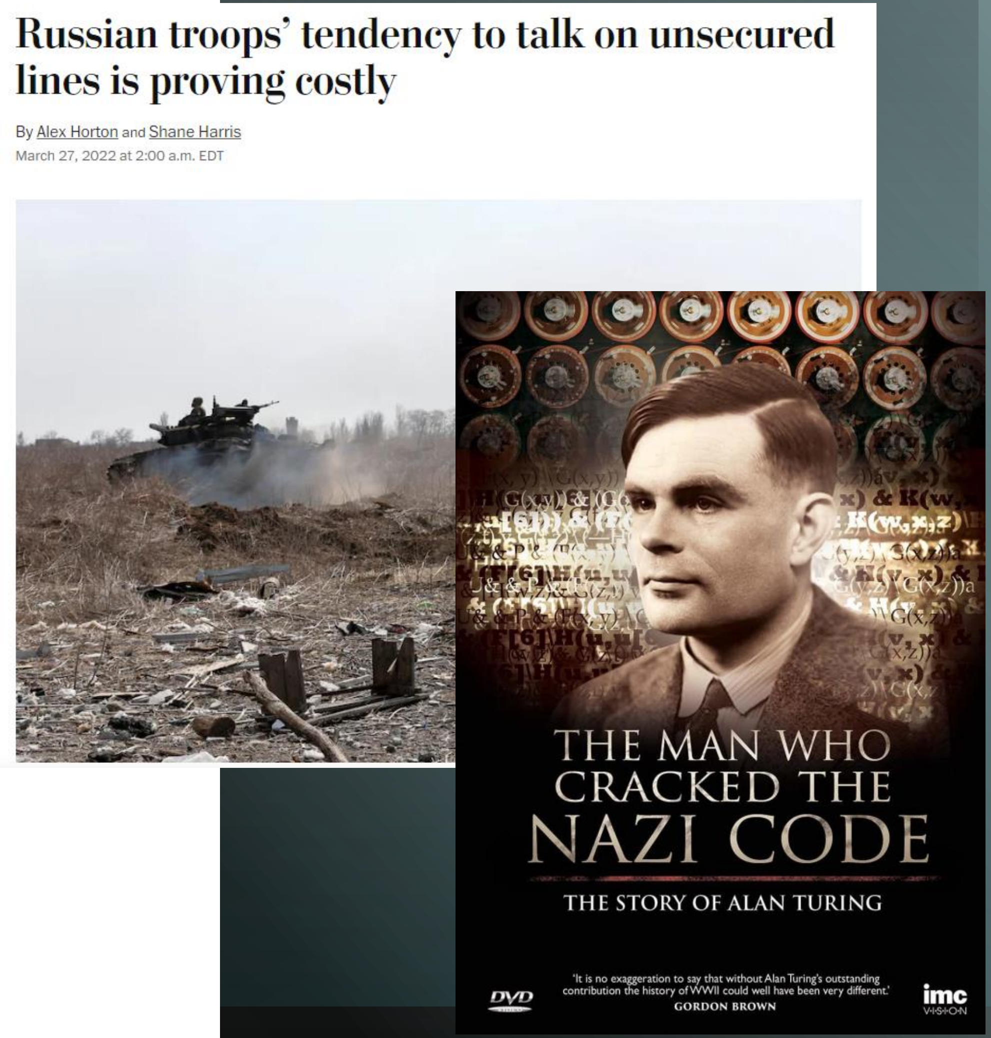
GETTY IMAGES

British Airways has been fined £20m ($26m) by the Information Commissioner's Office (ICO) for a data breach which affected more than 400,000 customers.

image ref: BBC news

Communication Hacked

- Information security plays an important role in modern war



Russian troops' tendency to talk on unsecured lines is proving costly

By Alex Horton and Shane Harris

March 27, 2022 at 2:00 a.m. EDT

THE MAN WHO CRACKED THE NAZI CODE
THE STORY OF ALAN TURING
'It is no exaggeration to say that without Alan Turing's outstanding contribution the history of WWII could well have been very different.'
GORDON BROWN
DVD
imc VISION

image ref: imc vision; Washington post

Service taken down

- Network attacks may cause web service down and company/organization will not be able to continue their business.
- On Feb. 28, 2018, GitHub, a platform for software developers, was hit with a DDoS attack that clocked in at 1.35 terabits per second and lasted for roughly 20 minutes.

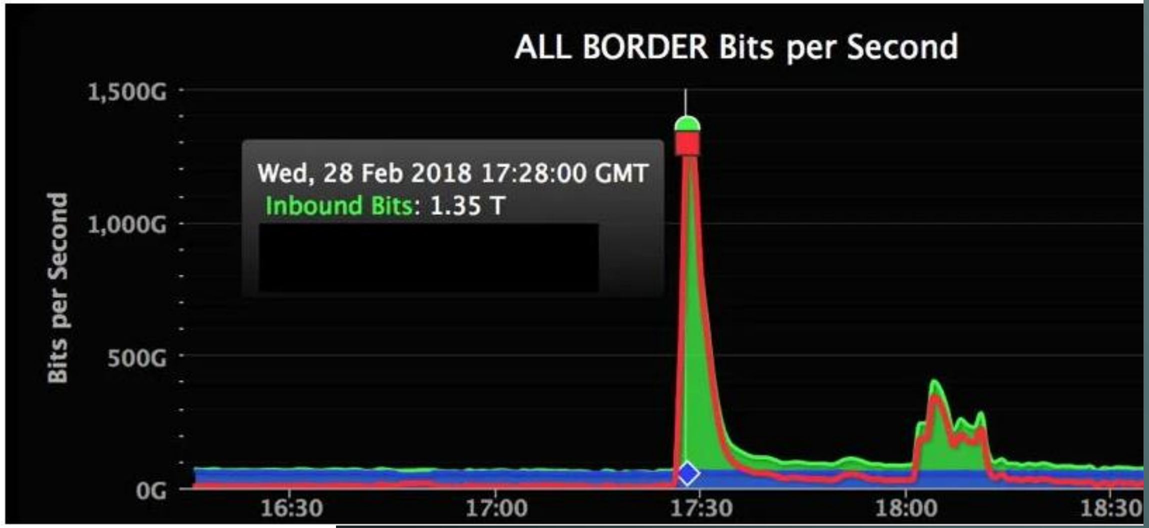
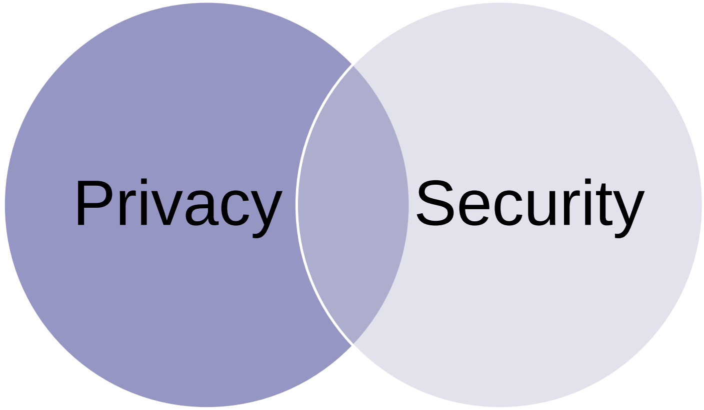


image ref: https://thehackernews.com/

Data Security and Privacy

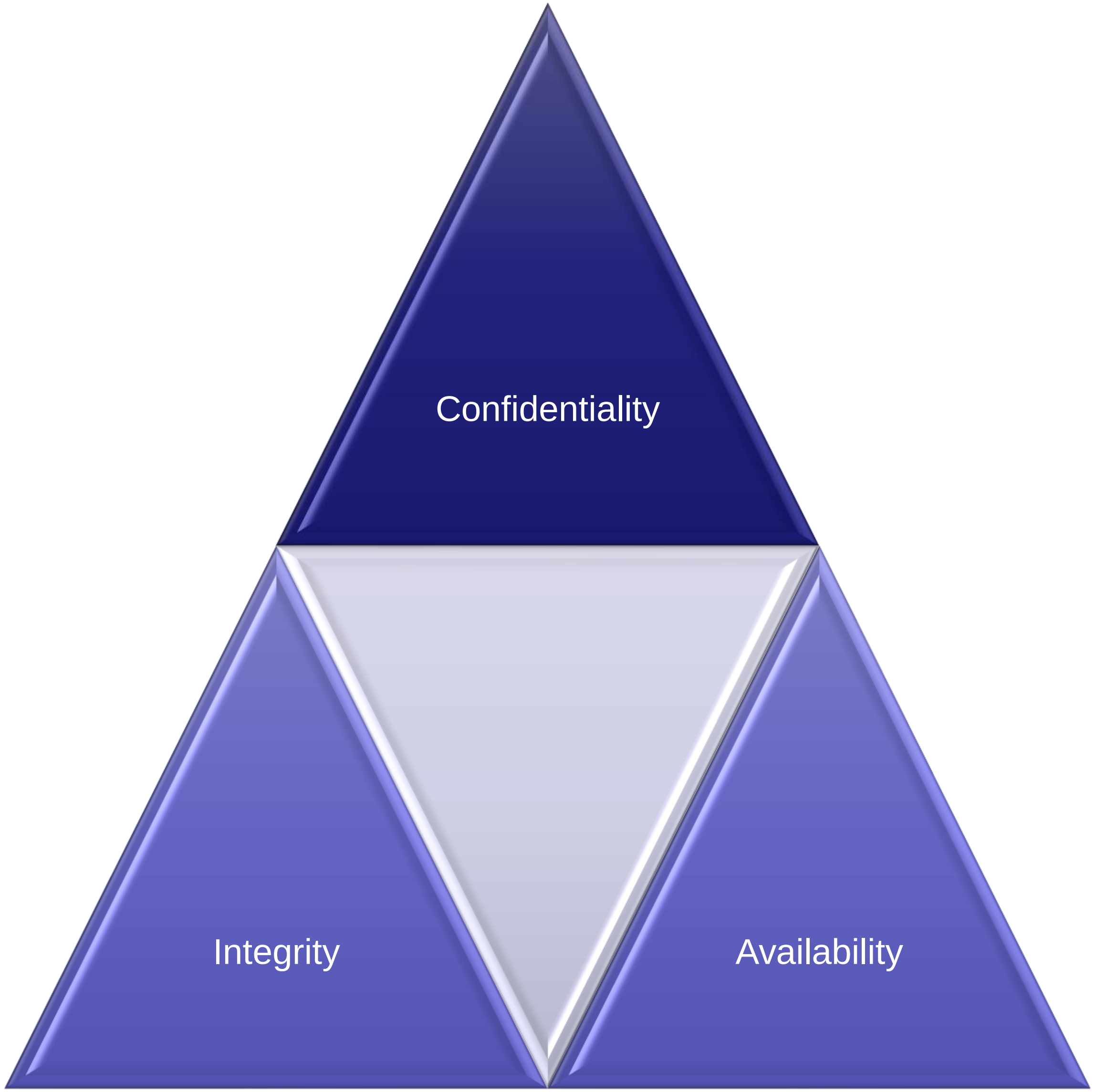
- ❑ Data Security
  - An umbrella term for ensuring data’s **confidentiality**, **integrity**, and **availability**.
    - Storing company documents
    - Make sure browser is connected to a legit website
    - Prevent network jamming attacks
- ❑ Data Privacy
  - Addressing the collection, access, usage, storage of (private/public) data of an individual.
  - Examples of invading privacy:
    - Transfer all customer’s phone number to a marketing company.
    - Print all IG photos of a classmate and stick them on the classroom wall.
    - Leak customer’s credit card info.



Data Security

Security

- ❑ Security is supported by the three core pillars - known as CIA triad.
- ❑ To supports these pillars, it involves many components:
  - ❑ Cryptography (math)
  - ❑ Software security
  - ❑ Software testing
  - ❑ IT auditing
  - ❑ System control
  - ❑ Network security
  - ❑ User educations
  - ❑ ...



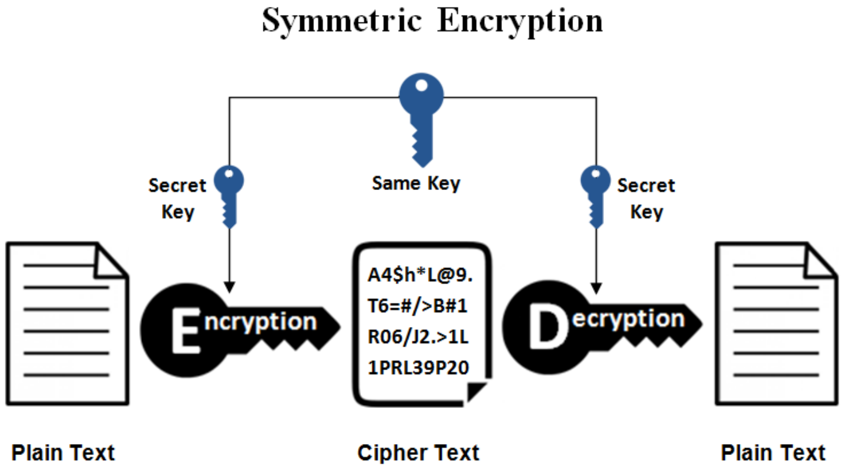
Security - Confidentiality

- ❑ To ensure only the intended person/entity reads the data.

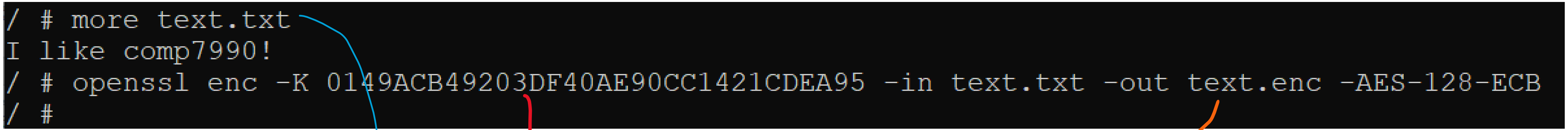
- ❑ Encryption – transform the data into ciphertext with a *key*.
  - ❑ Symmetric key encryption: same key for encryption/decryption
  - ❑ Asymmetric key encryption (aka public key encryption): different key for encryption/decryption

Confidentiality - Symmetric Encryption

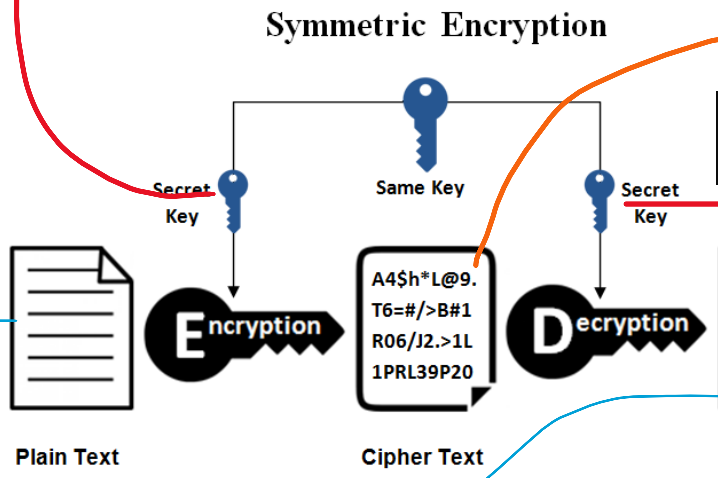
- ❑ Encryption is a process of transforming a plaintext into a ciphertext.
- ❑ Involve a ***key***, shared by the sender and the receiver.
- ❑ In principle the size of a key determine the security
- ❑ Assume the key is 30-bits long. It takes 2^{30} operations for a computer to *brute force* it.
- ❑ If the key is 40-bits long. It takes 2^{40} operations to brute force it.
  - ❑ 10 bits bring 1000 times efforts!
- ❑ A modern symmetric key encryption requires 256-bits key!

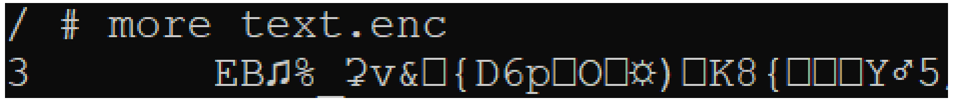


Confidentiality - Symmetric Encryption

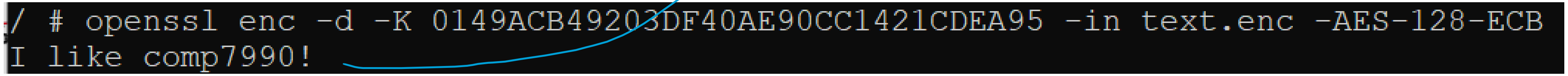


```
/ # more text.txt
I like comp7990!
/ # openssl enc -K 0149ACB49203DF40AE90CC1421CDEA95 -in text.txt -out text.enc -AES-128-ECB
/ #
```





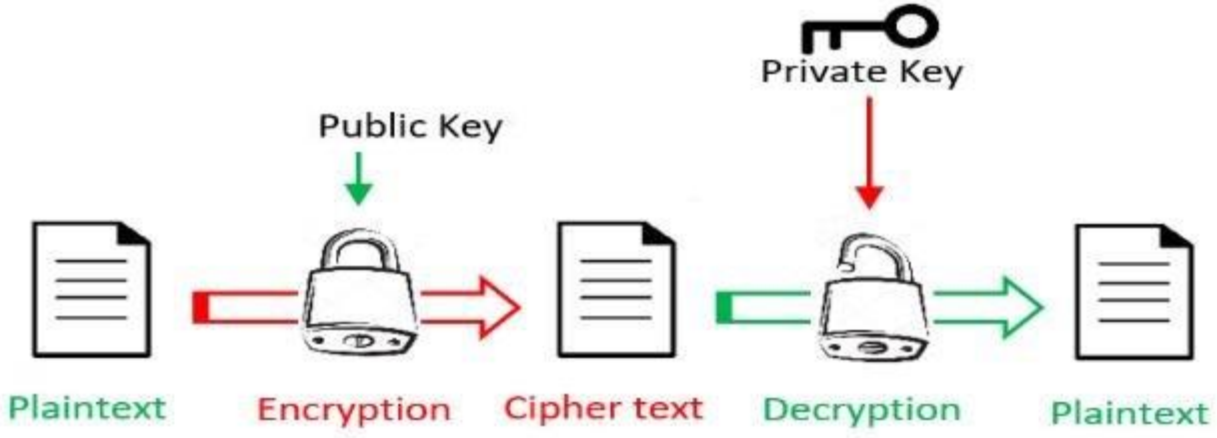
```
/ # more text.enc
3          EB♫%_Ɂv&□{D6p□O□☼)□K8{□□□Y♂5
```



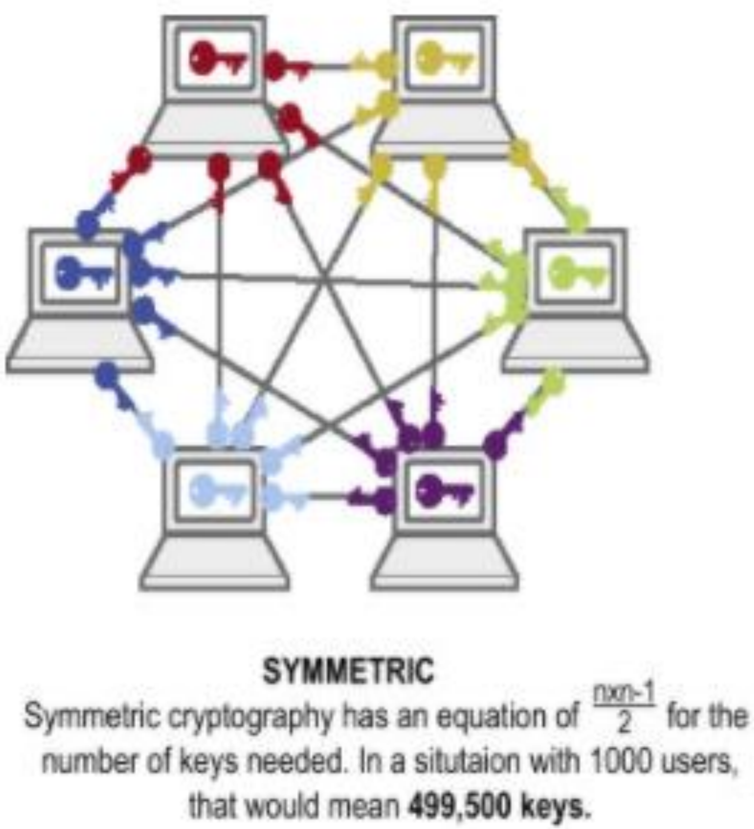
```
/ # openssl enc -d -K 0149ACB49203DF40AE90CC1421CDEA95 -in text.enc -AES-128-ECB
I like comp7990!
```

Confidentiality - Asymmetric Encryption

- ❑ Encryption is a process of transforming a plaintext into a ciphertext.
- ❑ Involve ***a public key and a private key***
- ❑ A public key is used to encrypt a plaintext to a ciphertext
- ❑ But the ciphertext cannot be decrypted using a public key.
  - ❑ It needs a private key to decrypt the message
- ❑ Anyone can encrypt, only the receiver can decrypt (not even the sender!)
- ❑ The length of the private key is recommended to have 256 bits more.



Confidentiality - Key management



SYMMETRIC

Symmetric cryptography has an equation of $\frac{nxn-1}{2}$ for the number of keys needed. In a situtaion with 1000 users, that would mean **499,500 keys.**

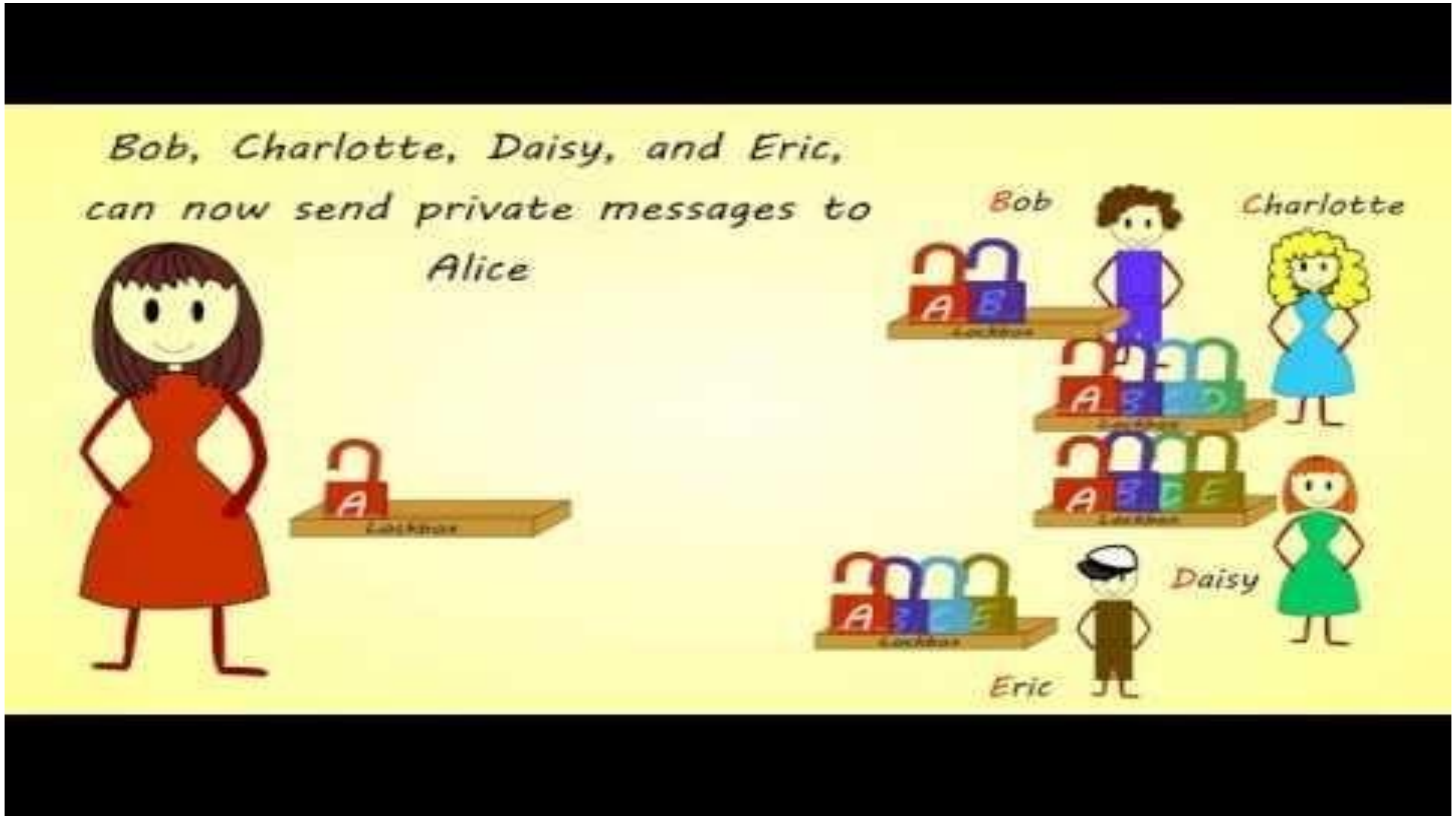

Bob, Charlotte, Daisy, and Eric,
can now send private messages to
Alice
Bob
Charlotte
Daisy
Eric
A
A B
A B C D
A B D E
A B C E
Lockbox

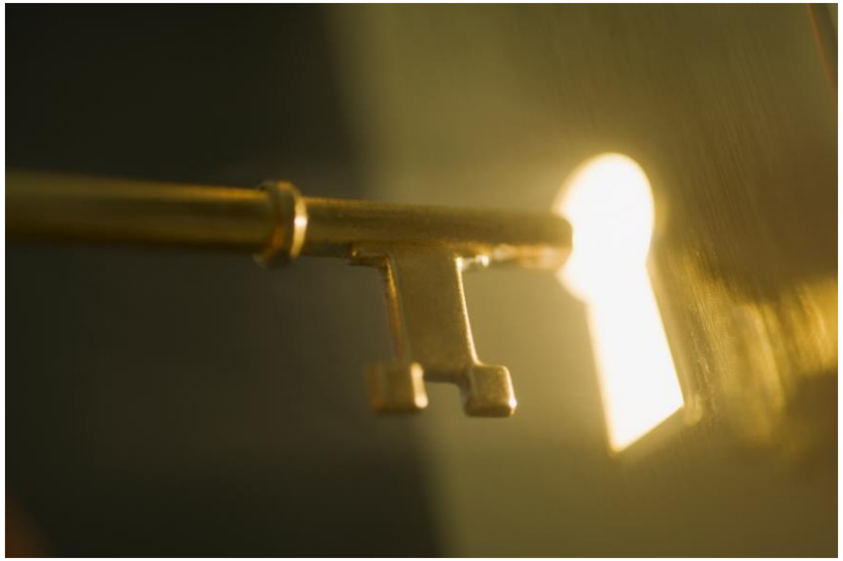
image credit: https://doi.org/10.1016/B978-0-12-818427-1.00011-2.

image credit: https://www.youtube.com/watch?v=E5FEqGYLL0o

Security - Confidentiality

- ❑ To ensure only the intended person/entity reads the data.
- ❑ Not just about encryption - permit a person to access data depend on his/her real identity.

- ❑ Authentication – to ensure the identity of a person
  - ❑ Password
  - ❑ Biometric – FaceID, Fingerprint
  - ❑ Devices – phone (SMS, Duo Mobile)



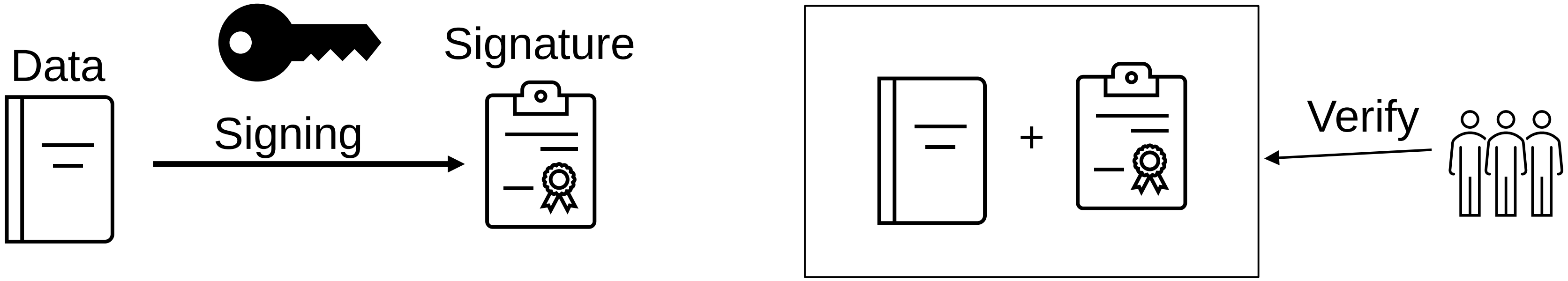
Security - Integrity

❑ To ensure the data has not been modified, inserted, removed.

❑ **Digital Signature** – only the *key* holder can produce a valid signature for data.

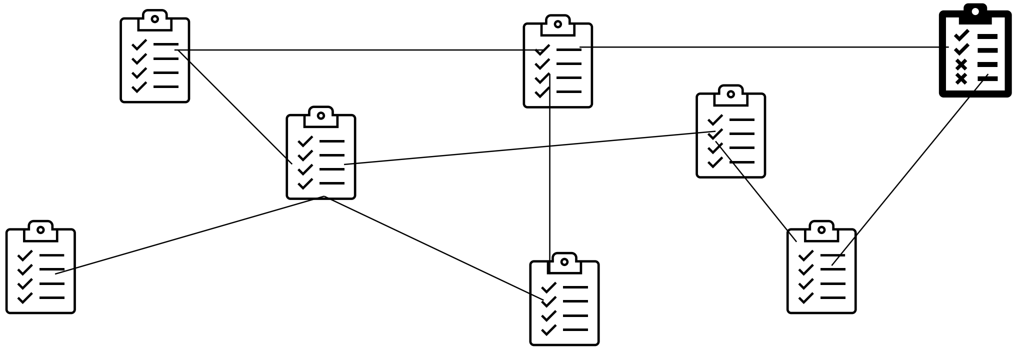
❑ The signature can be verified by any people - anyone can confirm that

1. The signature is associated with the data; and
2. The signature is produced by the key holder



Security - Integrity

- ❑ To ensure the data has not been modified, inserted, removed.
- ❑ **Blockchain** – decentralized ledger to store public data. Resistant against single node failure/attack
- ❑ Nodes in the network share the same set of transactions. If the data on a node is tampered by attackers, the record will be inconsistent and detected.

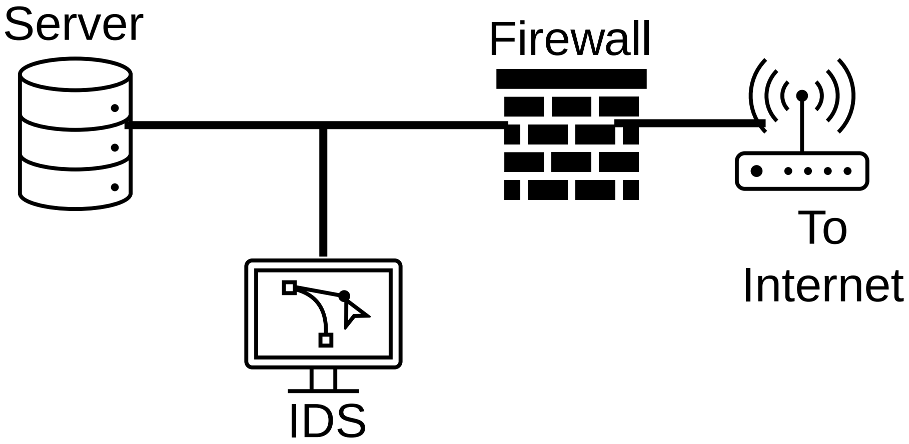


Inconsistent record would not spread to other nodes

Security - Availability

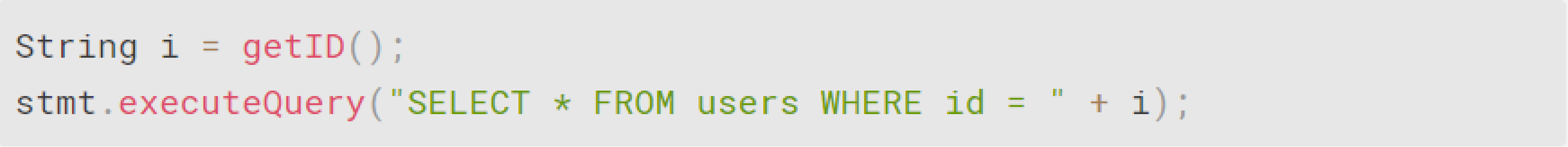
- ❑ To ensure the availability of data and services.

- ❑ **Firewall** and **Intrusion Detection System** (IDS):
  - ❑ Firewall – Installed on the border of the network, allows/denies packets that match pre-defined rules.
    - ❑ e.g. deny external traffic visit port 22
  - ❑ IDS – Attached to the network to listen and analysis the traffic, discovers if there is any attack patterns.
    - ❑ e.g. issue warning if inbound traffic increase 1000 times.



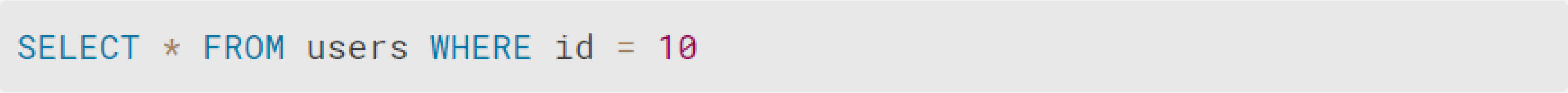
Security – Case Study SQL injection

- SQL injection is a database attack by exploiting the entanglement of instructions and data. Consider the following piece of code:



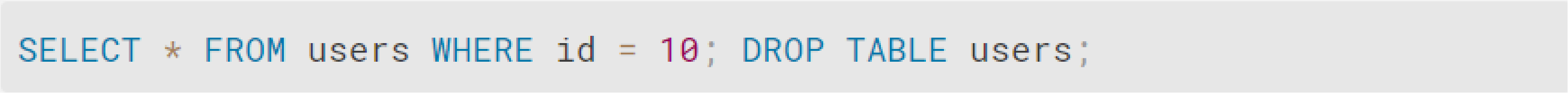
```
String i = getID();
stmt.executeQuery("SELECT * FROM users WHERE id = " + i);
```

- The equivalent SQL would be (assume a user inputs 10):



```
SELECT * FROM users WHERE id = 10
```

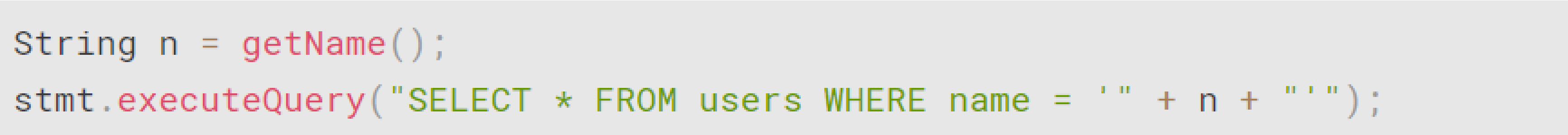
- If the user inputs 10 ; DROP TABLE users; for the ID instead:



```
SELECT * FROM users WHERE id = 10; DROP TABLE users;
```

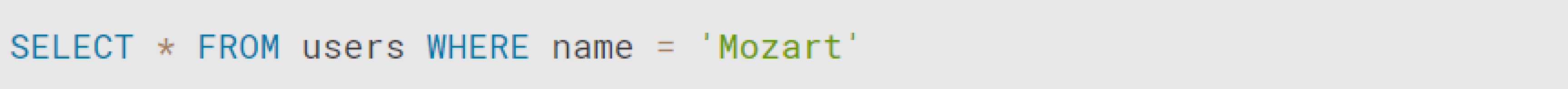
Security – Case Study SQL injection

- ❑ Another example:



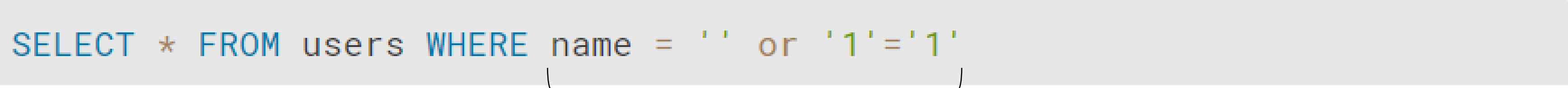
```
String n = getName();
stmt.executeQuery("SELECT * FROM users WHERE name = '" + n + "'");
```

- ❑ The equivalent SQL statement will be



```
SELECT * FROM users WHERE name = 'Mozart'
```

- ❑ What if the user enters ' or '1'='1 , the SQL statement will be



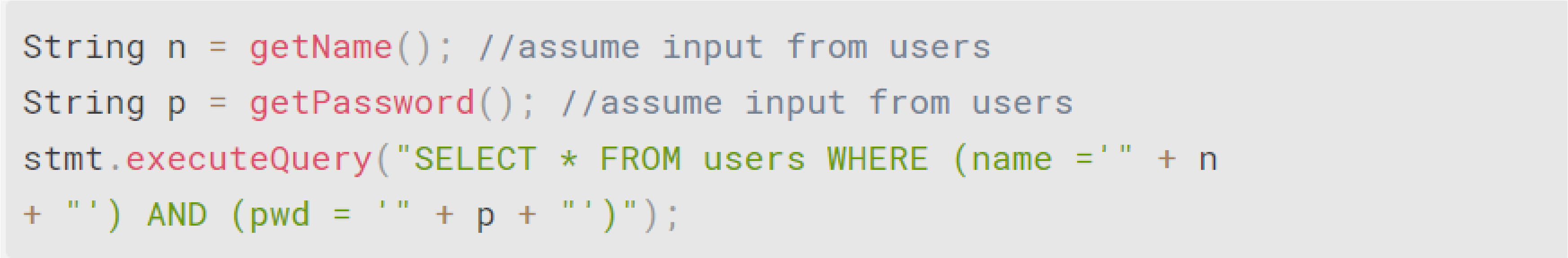
```
SELECT * FROM users WHERE name = '' or '1'='1'
```

Always true

This will dump the entire database!

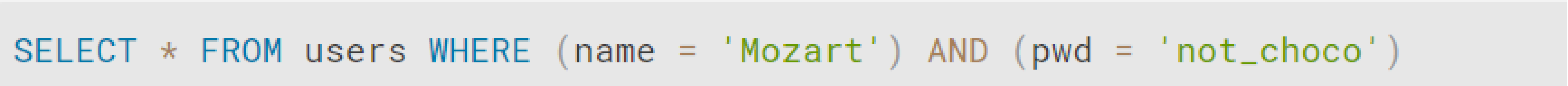
Security – Case Study SQL injection

- ❑ Another example:



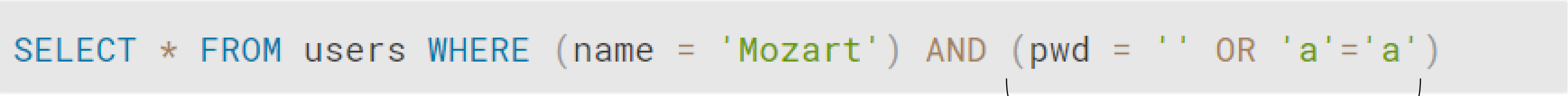
```
String n = getName(); //assume input from users
String p = getPassword(); //assume input from users
stmt.executeQuery("SELECT * FROM users WHERE (name ='" + n
+ "') AND (pwd = '" + p + "')");
```

- ❑ The equivalent SQL statement will be



```
SELECT * FROM users WHERE (name = 'Mozart') AND (pwd = 'not_choco')
```

- ❑ User enters ' OR 'a'='a for password, the SQL statement will be



```
SELECT * FROM users WHERE (name = 'Mozart') AND (pwd = '' OR 'a'='a')
```

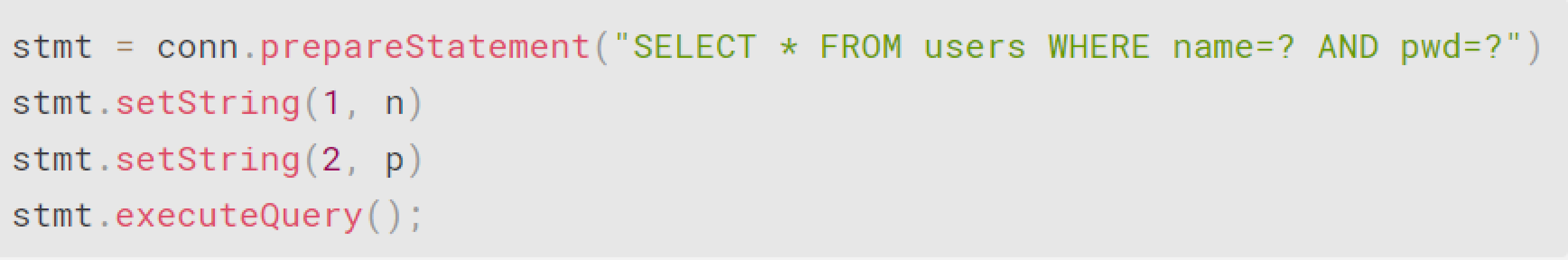
Always true

Will return data even the password is unknown!

https://www.codingame.com/playgrounds/154/sql-injection-demo/sql-injection

Security – Solutions to SQL injection

1. Permanent fix: use *prepareStatement* to avoid SQL injection



```
stmt = conn.prepareStatement("SELECT * FROM users WHERE name=? AND pwd=?")
stmt.setString(1, n)
stmt.setString(2, p)
stmt.executeQuery();
```

❑ The program will be able to separate instruction and data. Data will not be executed.

2. When updating program is not possible: use IDS to detect the attack.

❑ Detect if users submit the string containing ' " & | to a specific webpage.

Data Privacy

Privacy Preserving: Who?

❑ Government / public agencies

- The Centers for Disease Control want to identify disease outbreaks
- Insurance companies have data on disease incidents, seriousness, patient background, etc.
- But can/should they release this information?

❑ Industry Collaborations / Trade Groups

- An industry trade group may want to identify best practices to help members
- But some practices are trade secrets

❑ Multinational Corporations

- A company would like to mine its data for globally valid results
- But national laws may prevent transborder data sharing

❑ Public use of private data

- Data mining enables research studies of large populations
- But these populations are reluctant to release personal information

Sources of Constraints

- ❑ Regulatory requirements

- ❑ Contractual constraints
  - – Posted privacy policy
  - – Corporate agreements

- ❑ Secrecy concerns
  - – Secrets whose release could jeopardize plans
  - – Public Relations – “bad press”

Regulatory Constraint

Primarily national laws

- European Union
- US HIPAA rules
- Many others: (www.privacyexchange.org)

Example: European Union Data Protection Directives

- ❑ Directive 95/46/EC
  - – Passed European Parliament 24 October 1995
  - – The goal is to ensure the free flow of information
    - *Must preserve privacy needs of member states*
  - – Effective October 1998
- ❑ Effect
  - – Provides guidelines for member state legislation
    - Not directly enforceable
  - – Forbids sharing data with states that don’t protect privacy
    - Non-member states must provide adequate protection,
    - Sharing must be for “allowed use”
    - Contract Enforcements ensure adequate protection

EU 95/46/EC: Meeting the Rules

- Personal data is any information that can be traced directly *or indirectly* to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn’t violate privacy

EU 95/46/EC: Meeting the Rules

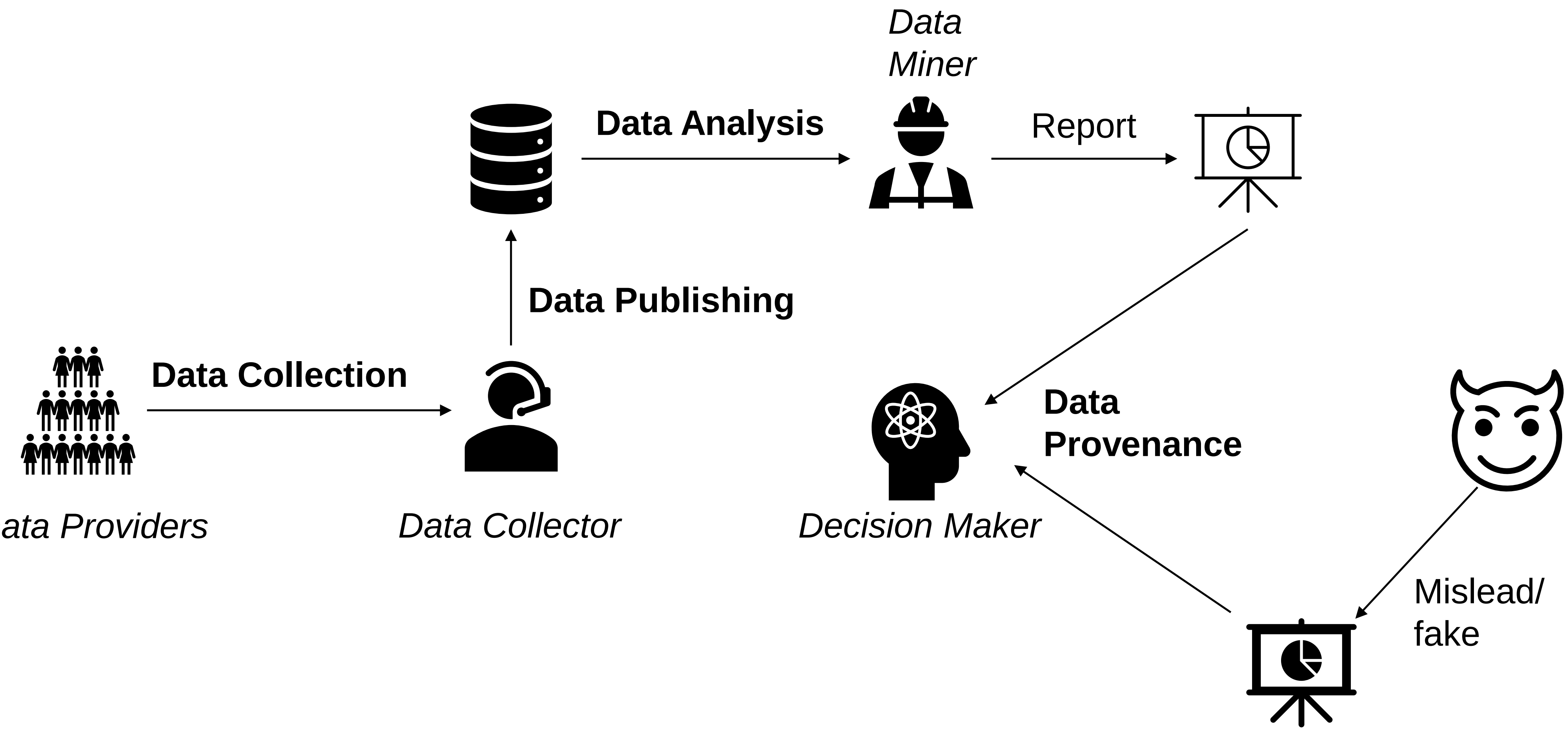
- ❑ Some uses specifically be proscribed
  - – Can’t reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life
- ❑ Must make data available to subjects
  - – Allowed to object to such use
  - – Must give advance notice/right to refuse direct marketing use
- ❑ Limits use for automated decisions (e.g., creditworthiness)
  - – A person can opt-out of automated decision making
  - – Onus on processor to show use is legitimate and safeguards in place to protect person’s interests
  - – Logic involved in decisions must be available to the affected person

Data Mining Process

The 4 type of users in Data Mining process

- **Data Provider:** the user who owns some data that are desired by the data mining task.
- **Data Collector:** the user who collects data from data providers and then publish the data to the data miner.
- **Data Miner:** the user who performs data mining tasks on the data.
- **Decision Maker:** the user who makes decisions based on the data mining results in order to achieve certain goals

Data Mining Process



Data Miner

Data Analysis

Report

Data Publishing

Data Collection

Data Provenance

Data Providers

Data Collector

Decision Maker

Mislead/
fake

Privacy concerns in each steps

- ❑ **Data Collection:** Data provider controls the sensitivity of the data he provides. Data collector maximizes the quality/accuracy of data.
- ❑ **Data Publishing:** Guarantee that the modified data contain no sensitive information but still preserve high utility.
- ❑ **Data Analysis:** extract useful information from data in a privacy-preserving manner.
- ❑ **Data Provenance:** make sure the mining results are credible.

Concerns in Data Collection

- ❑ The major concern of a data provider is whether he can control the sensitivity of the data he provides to others.

- ❑ On one hand, the provider should be able to make his very private data inaccessible to the data collector.

- ❑ On the other hand, if the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensations for the possible loss in privacy.

- ❑ Data Collector wishes to increase the quality of the data collected.

Approaches to Privacy Protection

- ❑ Limit the Access
- ❑ Trade Privacy for Benefit
- ❑ Provide False Data
- ❑ Indirect Questions

Limit the Access

❑ Anti-Tracking

When browsing the Internet, a user can utilize an anti-tracking mechanism to block the trackers from collecting the cookies.

- Privacy Browsing/Incognito mode
- Tor Browser

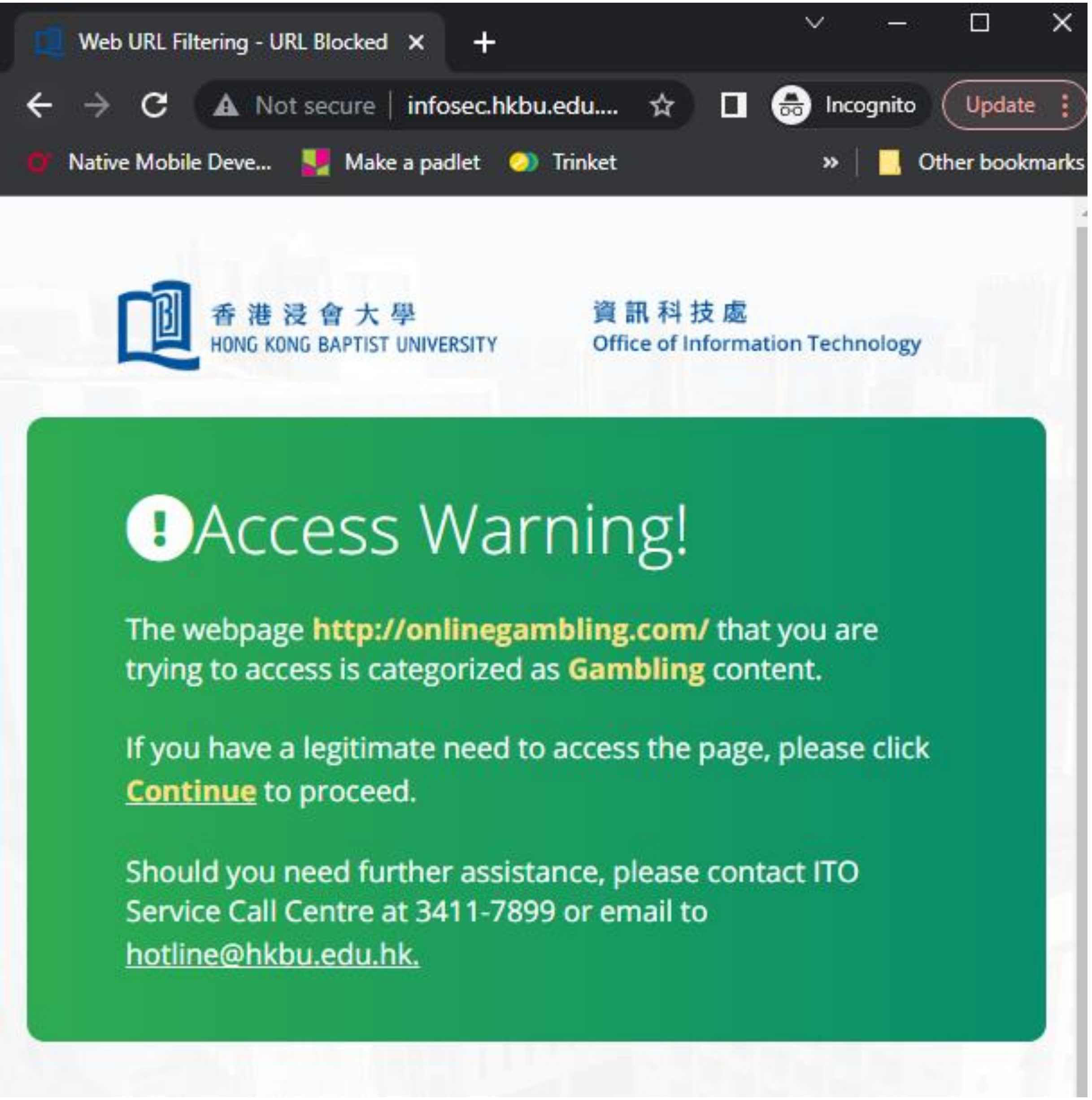
❑ Advertisement and script blockers

❑ Encryption tools

To make sure private online communication between two parties cannot be intercepted by third parties, a user can utilize encryption tools, such as ***TorChat.***

❑ Some anti-virus and anti-malware tools

Tor browser provides extra privacy -- my network admin does not know where am I going.

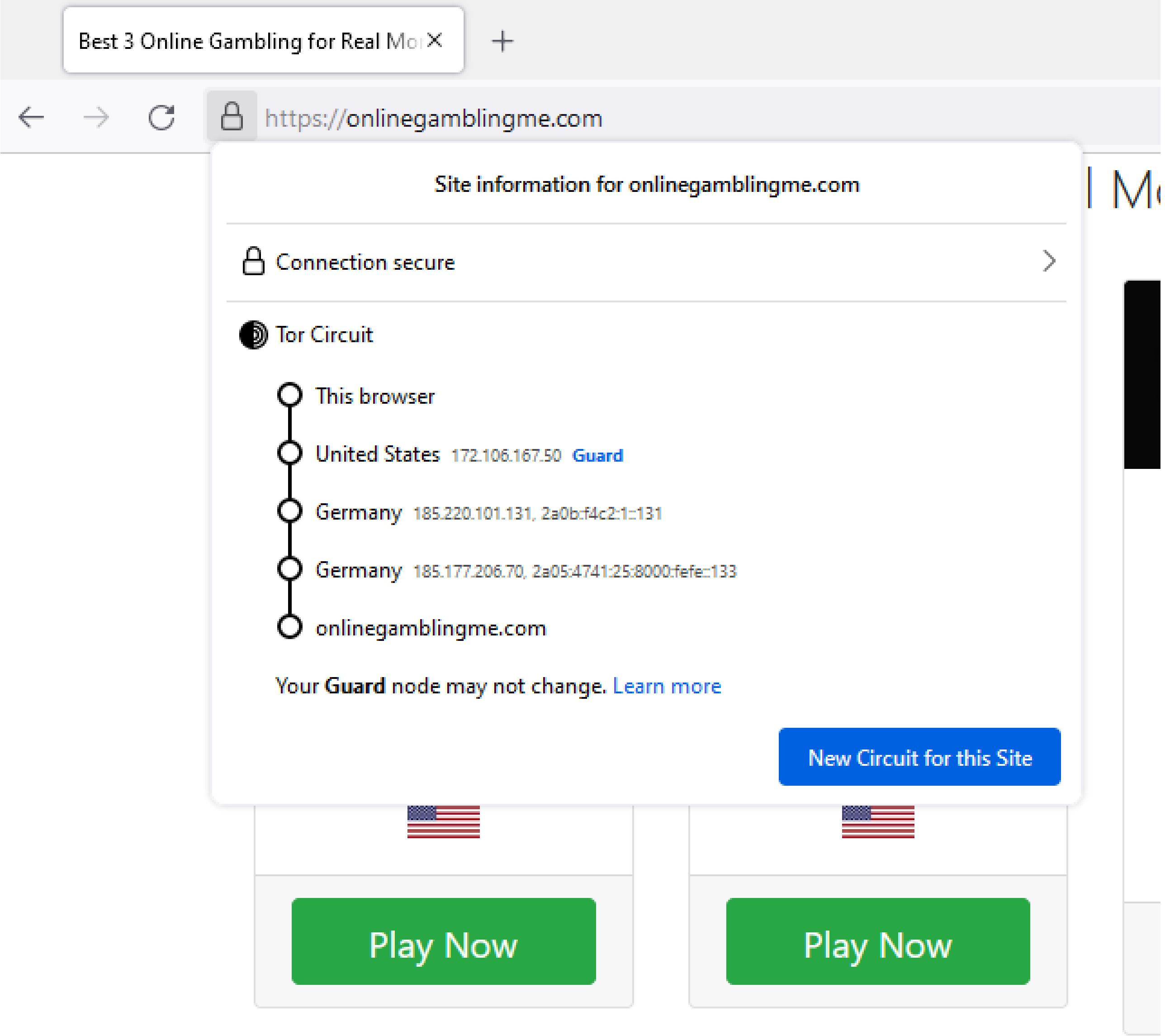

Web URL Filtering - URL Blocked
Not secure | infosec.hkbu.edu....
Incognito
Update
Native Mobile Deve...
Make a padlet
Trinket
Other bookmarks
香港浸會大學
HONG KONG BAPTIST UNIVERSITY
資訊科技處
Office of Information Technology

Access Warning!

The webpage http://onlinegambling.com/ that you are trying to access is categorized as Gambling content.

If you have a legitimate need to access the page, please click Continue to proceed.

Should you need further assistance, please contact ITO Service Call Centre at 3411-7899 or email to hotline@hkbu.edu.hk.


Best 3 Online Gambling for Real Mo
https://onlinegamblingme.com
Play Now
Play Now

Site information for onlinegamblingme.com

Connection secure

Tor Circuit

This browser

United States 172.106.167.50 Guard

Germany 185.220.101.131, 2a0b:f4c2:1::131

Germany 185.177.206.70, 2a05:4741:25:8000:fefe::133

onlinegamblingme.com

Your **Guard** node may not change. Learn more

New Circuit for this Site

Trade Privacy for Benefit

- The data provider may be willing to hand over some of his private data in exchange for certain benefits. Such as better services or monetary rewards.
  - e.g. Octopus rewarding scheme

- The data provider needs to know how to negotiate with the data collector so that he will get enough compensation for any possible loss of privacy.

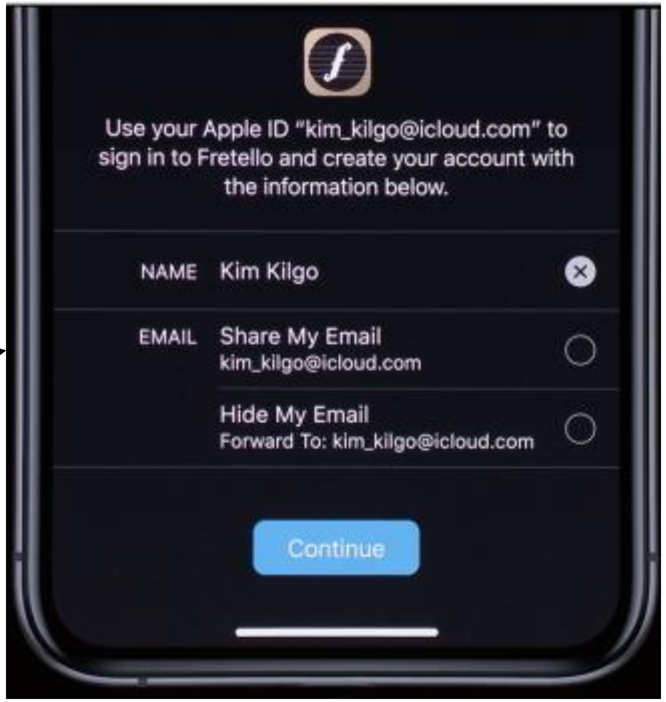
Provide False Data

Internet users cannot completely stop the unwanted access to their personal information. So instead of trying to limit the access, the data provider can provide false information to those untrustworthy data collectors.

□ Using "sockpuppets" to hide one's true activities

□ Using a fake identity to create phony information.

□ Using security tools to mask one's identity.



Use your Apple ID “kim_kilgo@icloud.com” to sign in to Fretello and create your account with the information below.

NAME Kim Kilgo

EMAIL ○ Share My Email
kim_kilgo@icloud.com

○ Hide My Email
Forward To: kim_kilgo@icloud.com

https://9to5mac.com/

Indirect Questioning – Improve the quality of survey

Data providers may reluctant to answer sensitive questions (morality, health, politics).

- Nonresponse rate increase
- Social desirability bias – tends to answering questions in the way that is socially acceptable rather than the truth.

Indirect questioning can improve the data quality by increasing the **perceived privacy** of the respondents.

- Unmatched count technique (UCT)
- Network scale-up technique (NST)
- Nonrandomized response technique (NRRT)
- Randomized response technique (RRT)

In-class experiments

- We are going to conduct experiments **in class** using these four methods and see if these methods are usable.
- To mimic the real situation, you may provide false answer or do not answer to any questions, if you are not comfortable with it.
- Provided that we cannot trace your identity by any functions of Mentimeter, which is the application that we are going to use.

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.1994.tb01578.x

Indirect Questioning - UCT

Unmatched Count Technique (UCT) divides respondents into two groups A and B. Group A has a list of *k* non-sensitive activity. The list in Group B has one additional sensitive activity. Respondents answer **how many** activities they have been engaged.

A: k non-sensitive activities

Mean value = *a*

B: Same list, with one more sensitive activity

Mean value = *b*

Estimate prevalence of the sensitive activity = b - a

UCT estimates higher than direction questioning (15.9 vs 8.7%; 80% higher) in a study of 10th grade students about sex with someone of the same gender.

https://www.menti.com/
Enter the code: 29677748

Experiment - UCT

- Instruction:
  - If your student ID is an odd number (ends with 1,3,5,7,9) answer with List A
  - If your student ID is an even number (ends with 2,4,6,8,0) answer with List B

How many of the following activities did you engage this month?

List A	List B
Post on IG/ Weibo	Post on IG/ Weibo
Swimming	Swimming
Hiking	Hiking
Drinking bubble tea (珍珠奶茶)	Drinking bubble tea (珍珠奶茶)
Shopping	Shopping
	Homework (cheating)

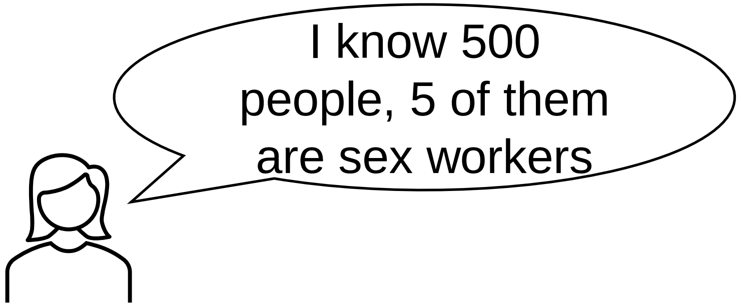
Result

	List A	List B
n		
0		
1		
2		
3		
4		
5		
6		
Total		
Mean		

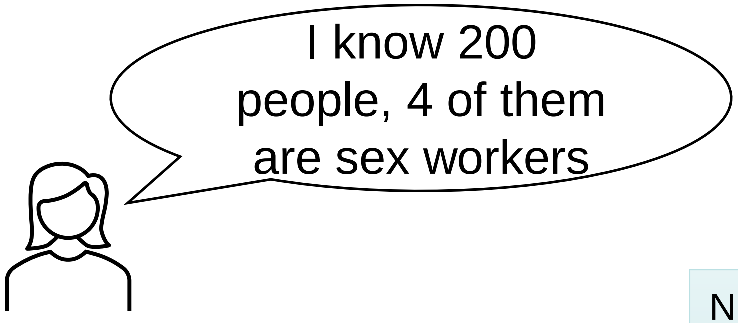
- Prevalence = %
- Respond rate: / = %

Indirect Questioning - NST

Network Scale-up Technique (NST) asks the respondent of how many people he/she knows who are engaging in sensitive activity.



k_1 = 500, s_1 = 5



k_2 = 200, s_2 = 4

NST has been used in studies like deaths in earthquake, drug uses, HIV prevalence, sexial crime against youth and female…

Estimate prevalence of the sensitive activity = $\frac{\sum s_i}{\sum k_i}$

Limitations:

1. Hard to define you know someone in the way you know their hidden side.
2. People’s social networks may not represent the population.

Experiment NST

Q1. How many classmates you know in your undergraduate?

Q2. How many of them had cheated in their final year?

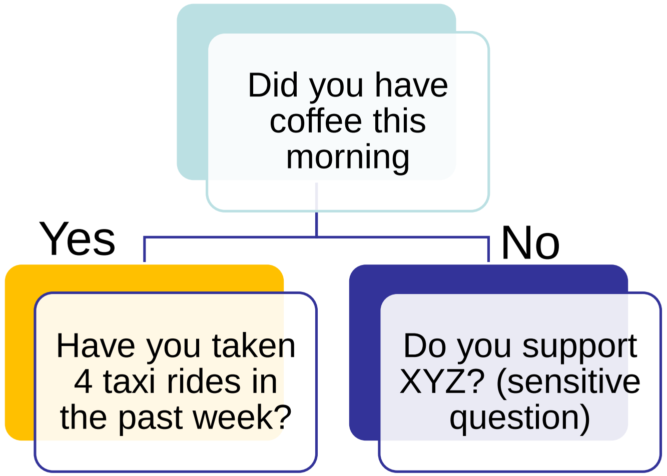
Historical Result

Q1	Q2
Total: 3373	Total: 711

- Estimated prevalence: 711/3373 = 21%
- 8 students refused to respond.
- 1 students gave wrong answer.

Indirect Questioning - NRRT

Nonrandomized response technique (NRRT) asks the respondent two set of questions according to a decision tree. The respondent needs to reply the answer of the second question but not the first one.

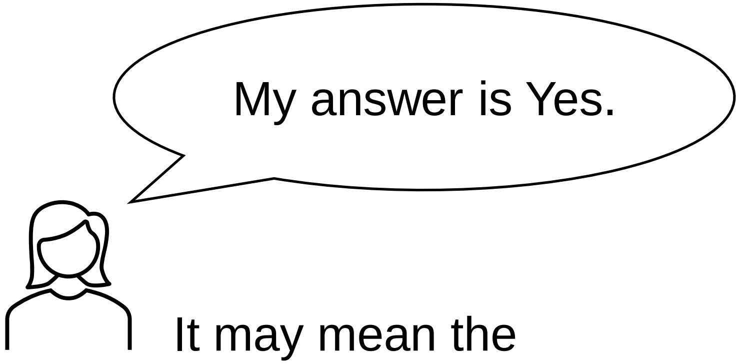

Did you have coffee this morning

Yes

No

Have you taken 4 taxi rides in the past week?

Do you support XYZ? (sensitive question)



It may mean the respondent have taken 4 taxi rides or support XYZ.

Let c = prob. of drink coffee in the morning, t = prob of. take more than 4 taxi rides, s = prob of support XYZ, P = ratio of answer Yes

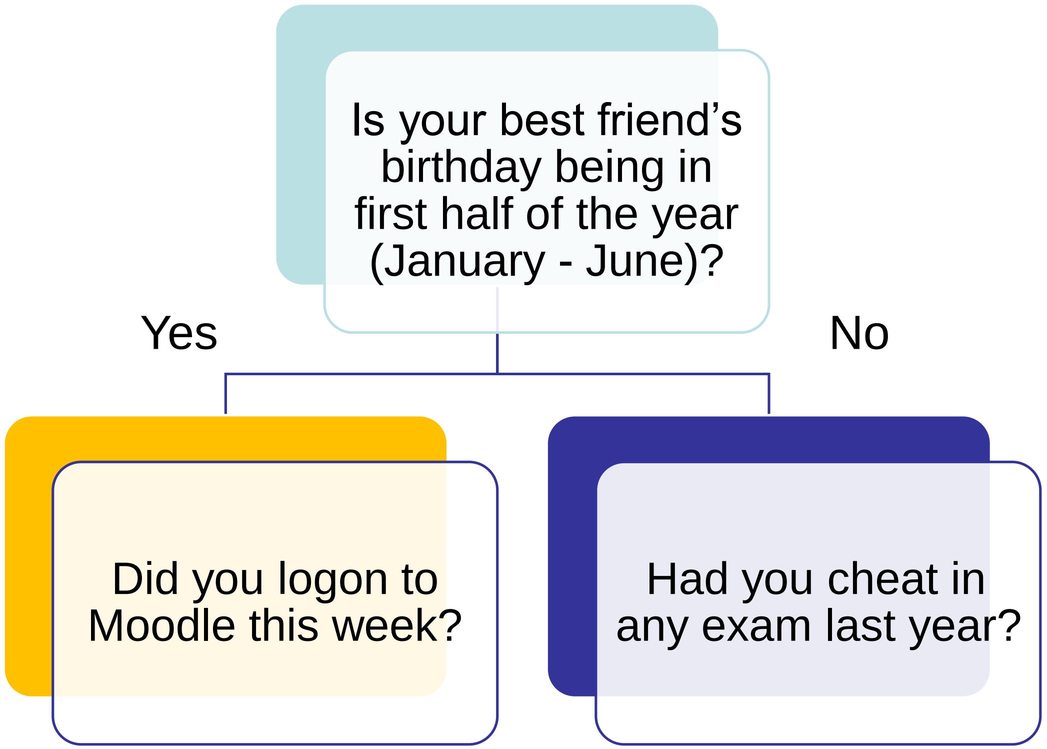
$$P = \mathrm{c} \cdot t + (1 - c)s$$

$$s = (P - c \cdot t)/(1 - c)$$

https://www.menti.com/
Enter the code: 69633416

Experiment with NRRT

- Based on the answer of 1st question, find the 2nd question and answer it on Mentimeter



Is your best friend’s birthday being in first half of the year (January - June)?

Yes

No

Did you logon to Moodle this week?

Had you cheat in any exam last year?

Historical Result

	Probability
c	50%
t	95%
P	/ = %
s	%

Respond rate / = %

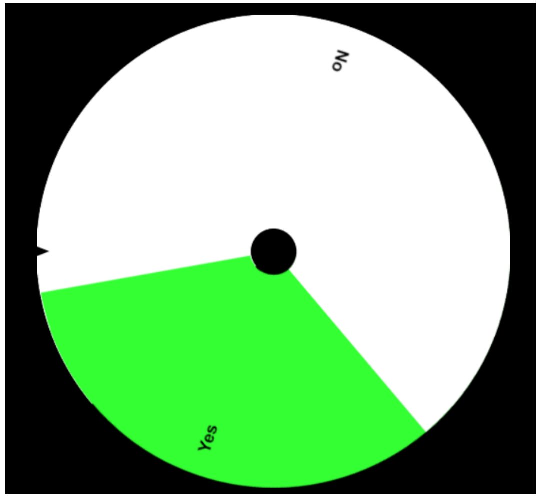
Let c = prob. of bday being in 1st half of year, t = prob of. logon to Moodle, s = prob of cheat, P = ratio of answer Yes

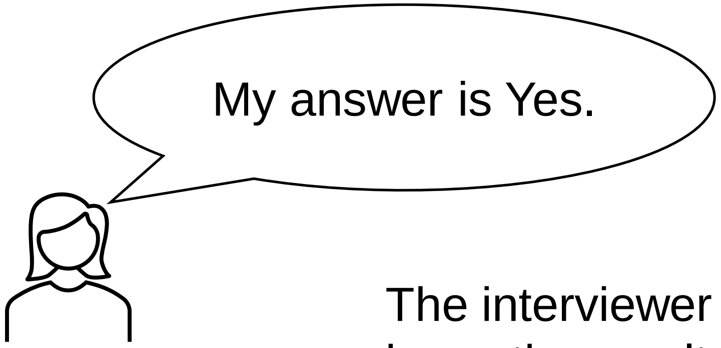
$$P = \mathrm{c} \cdot t + (1 - c)s$$

$$s = (P - \mathrm{c} \cdot t)/(1 - c)$$

Indirect Questioning - RRT

Randomized response technique (RRT) introduce randomness to the respondent's answering process. For example, spin a wheel privately. Answer "Yes" if the result of spin is "Yes". Answer truthfully if the result of spin is "No".





My answer is Yes.

The interviewer does not know the result of the spin. It could mean HIV+ or HIV-

Let P = ratio of answer "Yes", s = prob of HIV+, θ = prob of result of spin is "Yes"

$$P = \theta + (1 - \theta)s$$

$$s = (P - \theta)/(1 - \theta)$$

https://www.menti.com/
Enter the code: 49452167

Experiment

- https://wheeldecide.com/index.php?c1=Yes&c2=No&col=gw&time=5&weights=1,2
- Depends on the result of the spin,
  - If the result is “Yes”, answer “Yes”
  - If the result is “No”, answer truthfully whether you had cheated in any exam last year

Result

n	
Number of Yes	
θ	33.3%
P	/ = %

Estimated Prevalence: ____ %

Let P = ratio of answer Yes, s = prob of cheat, θ = prob of result of spin is “Yes”

$$P = \theta s + (1-\theta)(1-s)$$

$$s = (P-\theta)/(1-\theta)$$

Discussion

- Name the advantages/disadvantages for each method.
- Tell us which one you feel it is most effective (able to make respondents to answer honestly).

Concern in Data Publishing

- Protect the privacy before publishing.

- Make sure that sufficient utility of the data can be retained after the modification; otherwise, collecting the data will be a wasted effort.

- The data modification process adopted by data collector, with the goal of preserving privacy and utility simultaneously, is usually called ***privacy preserving data publishing*** (PPDP).

Basics of PPDP

PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. Each record consists of the following 4 types of attributes:

- ❑ Explicit Identifier (EI): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile no.
- ❑ Quasi-identifier (QI): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
- ❑ Sensitive Data (SD): Attributes that contain confidential information about the record owner, such as disease and salary.
- ❑ Non-sensitive Data (NSD): Data that are not sensitive for the given context

Example of data sensitivities

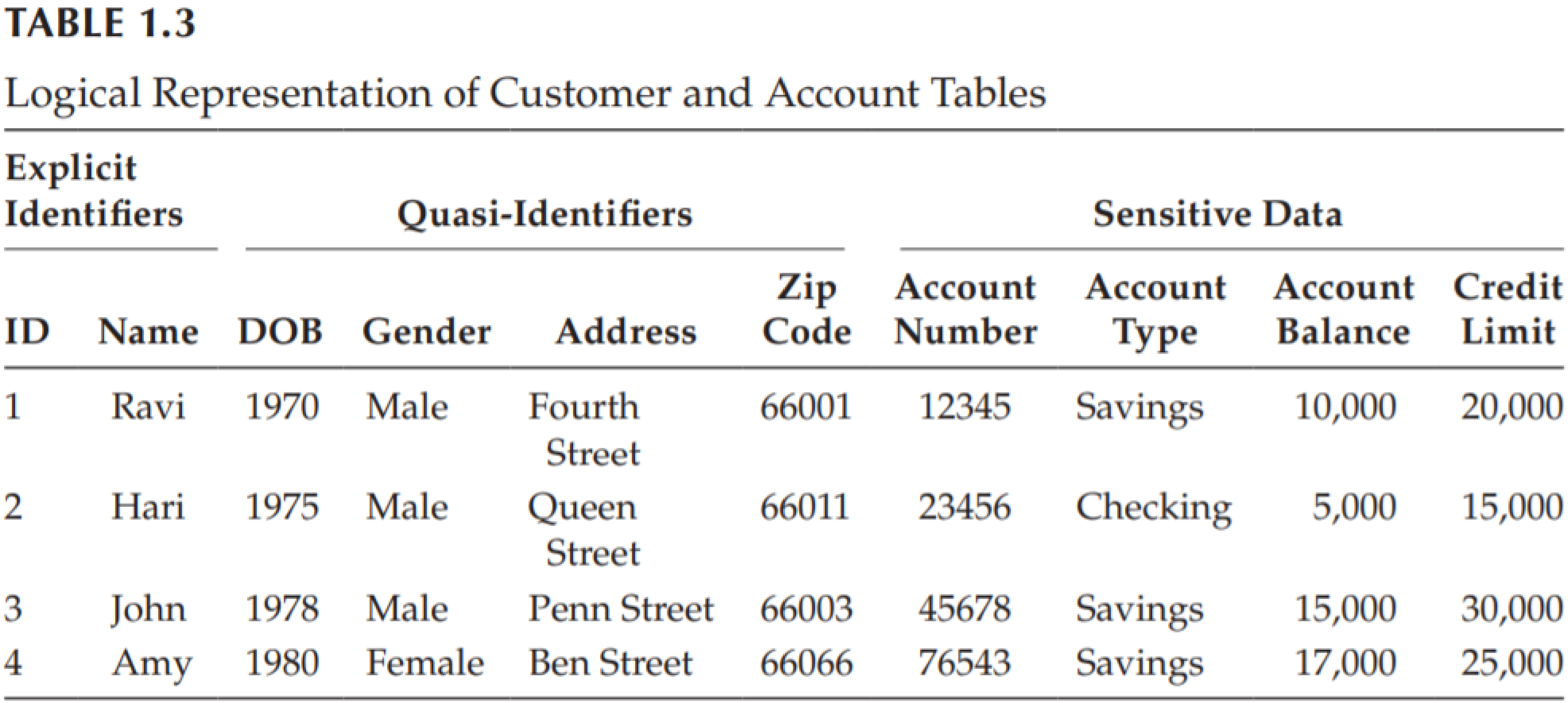


TABLE 1.3

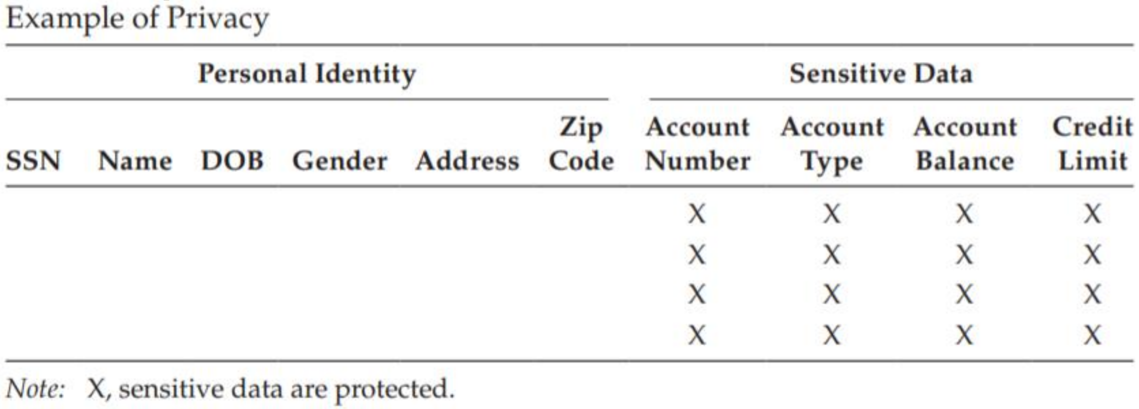
Logical Representation of Customer and Account Tables

Explicit Identifiers		Quasi-Identifiers				Sensitive Data			
ID	**Name**	**DOB**	**Gender**	**Address**	**Zip Code**	**Account Number**	**Account Type**	**Account Balance**	**Credit Limit**
1	Ravi	1970	Male	Fourth Street	66001	12345	Savings	10,000	20,000
2	Hari	1975	Male	Queen Street	66011	23456	Checking	5,000	15,000
3	John	1978	Male	Penn Street	66003	45678	Savings	15,000	30,000
4	Amy	1980	Female	Ben Street	66066	76543	Savings	17,000	25,000

From [4]

Privacy and Anonymity

Privacy can be defined as we have knowledge of a person’s identity but not of an associated personal fact.

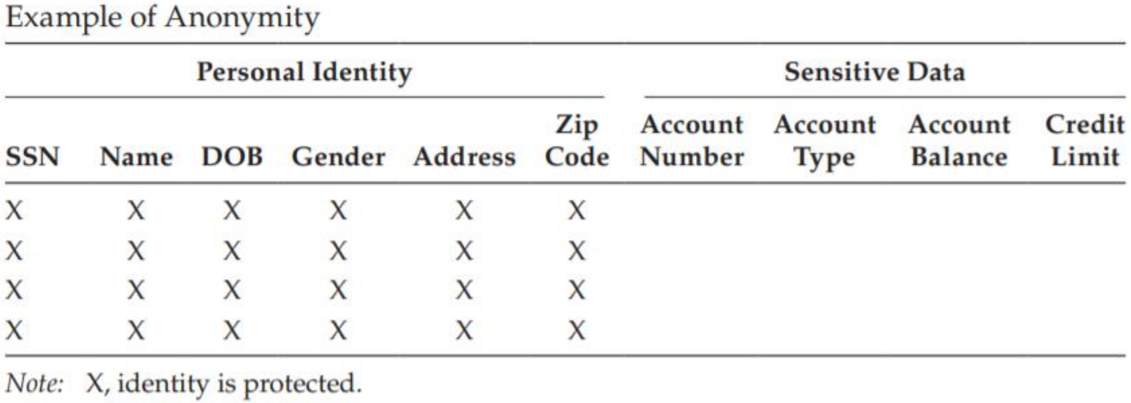


Example of Privacy

Personal Identity						Sensitive Data			
SSN	Name	DOB	Gender	Address	Zip Code	Account Number	Account Type	Account Balance	Credit Limit
						X	X	X	X
						X	X	X	X
						X	X	X	X
						X	X	X	X

Note: X, sensitive data are protected.

Anonymity can be defined as we have knowledge of a personal fact, but not of the associated person’s identity.



Example of Anonymity

Personal Identity						Sensitive Data			
SSN	Name	DOB	Gender	Address	Zip Code	Account Number	Account Type	Account Balance	Credit Limit
X	X	X	X	X	X				
X	X	X	X	X	X				
X	X	X	X	X	X				
X	X	X	X	X	X				

Note: X, identity is protected.

From [4]

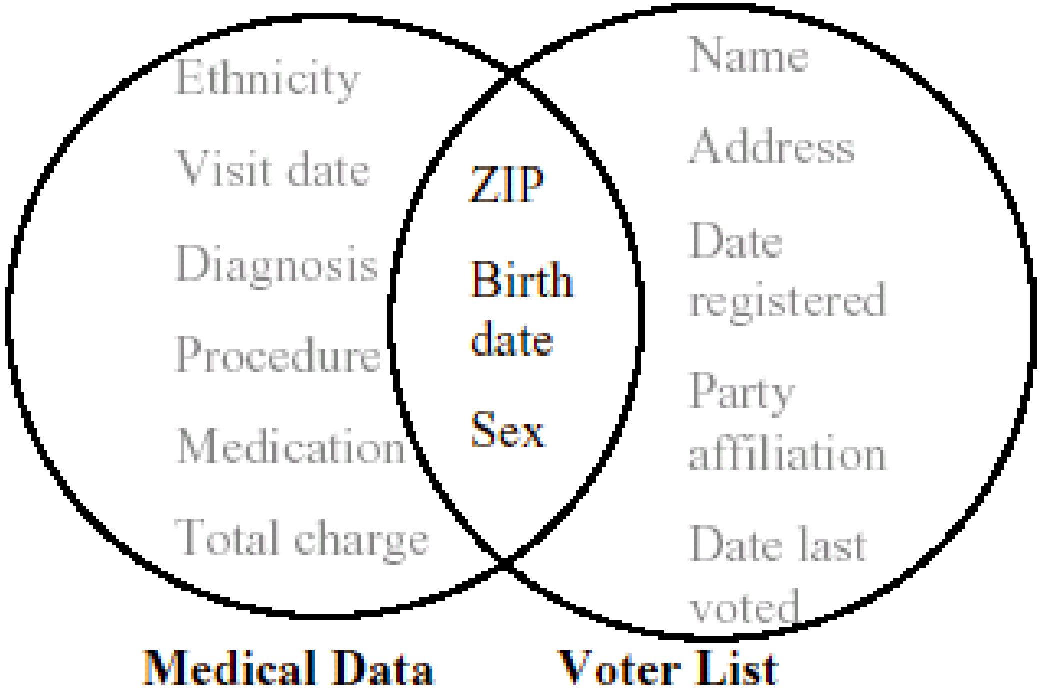
Masking EI isn't enough – Data Re-Identification

Group Insurance Commission (GIC) , who purchases health insurance for state employees, collected patient specific data and share the anonymized data to researchers.

A researcher purchased the voter registration list for Cambridge Massachusetts

69% unique on postal code and birth date

87% unique with all three



From [5]

Balancing Data Privacy and Utility

Privacy preservation should also ensure utility of data. By anonymizing the data, EI are completely masked out, QI is de-identified by applying a *transformation* function, and SD is left in its original form.

QI and SD have a strong correlation in between. If the correlation is lost, the data may not be useful for any purpose.

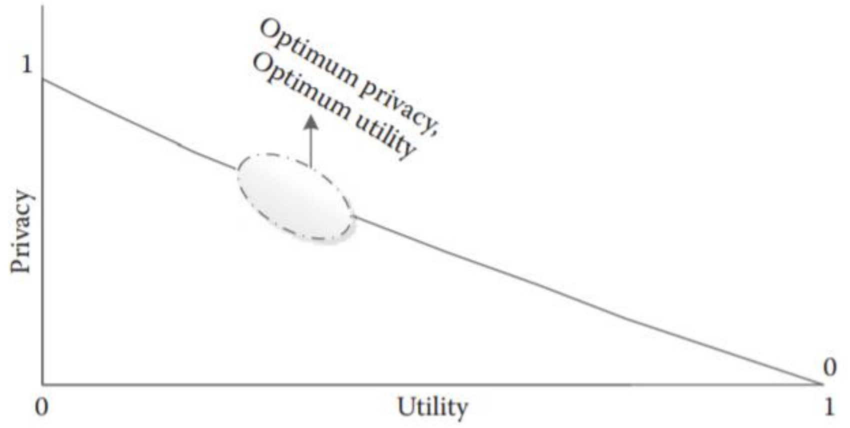


FIGURE 1.5
Privacy versus utility map.

from [4]

HIPAA defines 18 attributes (including name, social security number, phone, ***admission date***) as personal identifiable information which needs to be completely anonymized in data analysis.

- Impossible to analyze the efficacy of the treatment if admission date is anonymized!

Anonymization Operations

- ❑ Generalization. This operation replaces some values with a parent value in the taxonomy of an attribute.
- ❑ Suppression. This operation replaces some values with a special value (e.g. an asterisk '*'), indicating that the replaced values are not disclosed.
- ❑ Anatomization. Publish QI and SD in two separate tables.
- ❑ Permutation. This operation de-associates the relationship of attributes by partitioning a set of data records into groups and shuffling their sensitive values within each group.
- ❑ Perturbation. This operation replaces the original data values with some synthetic data values.

k-anonymity

- k-anonymity is to modify the values of quasi-identifiers in original data table, so that every tuple in the anonymized table is indistinguishable from at least k−1 other tuples along the quasi-identifiers.

- The anonymized table is called a k-anonymous table.

- If a table satisfies k-anonymity and the adversary only knows the quasi-identifier values of the target individual, then the probability that the target’s record being identified by the adversary will not exceed 1/k.

k-anonymity – original table

Explicit Identifier		*Quasi-identifier*				*Sensitive Data*	
Emp ID	**Name**	**Gender**	**DOB**	**Address**	**Education**	**Years of Experience**	**Salary**
000101	Alex	M	15-07-1973	Shatin	Doctorate	20	35000
000102	Bob	M	20-11-1975	Tuen Mum	Masters	17	28000
000103	Carol	F	12-12-1977	Tai Po	Masters	18	26000
000104	Daisy	F	08-07-1974	Mongkok	Doctorate	20	38000
000105	Ernest	M	17-06-1985	Mongkok	Graduate	12	10000
000106	Fred	M	05-07-1980	Kowloon Tong	Graduate	10	9000
000107	Gina	F	01-02-1981	Yuen Long	Masters	15	18000
000108	Henry	M	03-01-1978	North Point	Masters	18	22000
000109	Isaac	M	10-11-1981	Tsing Yi	Graduate	20	15000
000110	Joan	F	18-12-1982	Sheung Wan	Doctorate	15	32000
000111	Keith	M	22-10-1982	Tin Hau	Masters	12	14000
000112	Larry	M	25-11-1979	Shatin	Masters	14	16000

k-anonymity – 2-Anonymous

EI should be completely masked or transformed before shared

Table shuffled

Hidden	*Explicit Identifier*			*Quasi-identifier*				*Sensitive Data*	
(Emp ID)	Transformed Emp ID	Name		Gender	Age	Address	Education	Years of Experience	Salary
000107	5	*	🟨	F	30-40	*	PG	15	18000
000110	2	*	🟨	F	30-40	*	PG	15	32000
000105	10	*	🟧	M	30-40	*	PG	12	10000
000106	4	*	🟧	M	30-40	*	PG	10	9000
000109	20	*	🟧	M	30-40	*	UG	20	15000
000111	8	*	🟧	M	30-40	*	UG	12	14000
000104	17	*	🟦	F	40-50	*	PG	20	38000
000103	16	*	🟦	F	40-50	*	PG	18	26000
000102	11	*	🟫	M	40-50	*	PG	17	28000
000108	9	*	🟫	M	40-50	*	PG	18	22000
000101	22	*	⬛	M	40-50	*	UG	20	35000
000112	18	*	⬛	M	40-50	*	UG	14	16000

Generalization

Suppression

k-anonymity – 4-Anonymous

Hidden	*Explicit Identifier*			*Quasi-identifier*				*Sensitive Data*	
(Emp ID)	Transformed Emp ID	Name		Gender	Age	Address	Education	Years of Experience	Salary
000107	5	*	🟨	F	Any	*	PG	15	18000
000110	2	*	🟨	F	Any	*	PG	15	32000
000105	10	*	🟧	M	30-40	*	UG	12	10000
000106	4	*	🟧	M	30-40	*	UG	10	9000
000109	20	*	🟧	M	30-40	*	UG	20	15000
000111	8	*	🟧	M	30-40	*	UG	12	14000
000104	17	*	🟨	F	Any	*	PG	20	38000
000103	16	*	🟨	F	Any	*	PG	18	26000
000102	11	*	🟫	M	40-50	*	Any	17	28000
000108	9	*	🟫	M	40-50	*	Any	18	22000
000101	22	*	🟫	M	40-50	*	Any	20	35000
000112	18	*	🟫	M	40-50	*	Any	14	16000

k-anonymity

- Record linkage: As most QI attributes are also present in external data sources, the anonymization technique should prevent the linking of a record owner’s QI attribute to these external data sources.

- Utility of the transformed data: Nonperturbative techniques such as generalization preserve the truth in the data table

- Protection of outlier records: It is difficult to mask outlier records even when techniques such as additive noise are added.

- The correlation/association between QI and SD are preserved and protected.

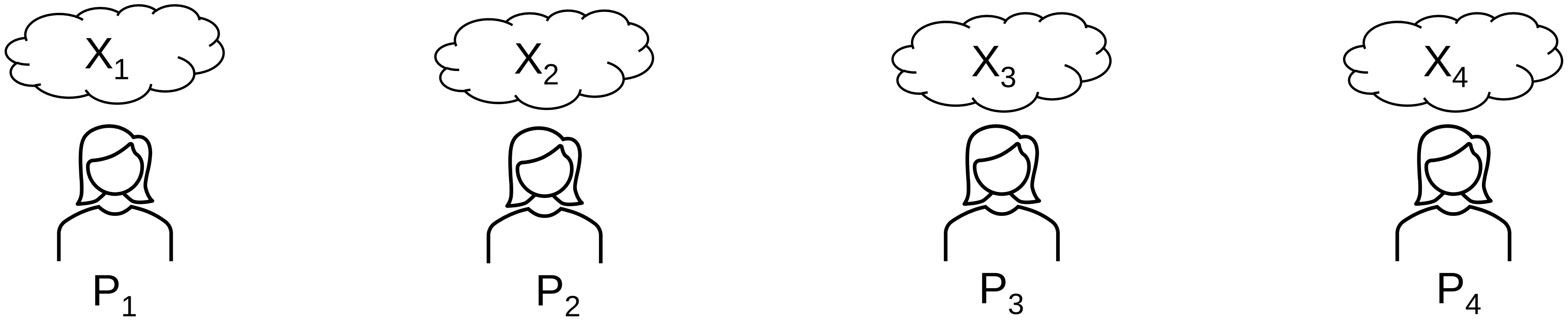
Concern in Data Analysis

- The primary concern of data miner is how to prevent sensitive information from appearing in the mining results.

- To perform a privacy-preserving data mining, the data miner usually needs to modify the data he got from the data collector.

- Even better, the data miner does not have the access to data but mining through collaboration with other parties (MPC).

Secure Multi-party Computation (MPC)



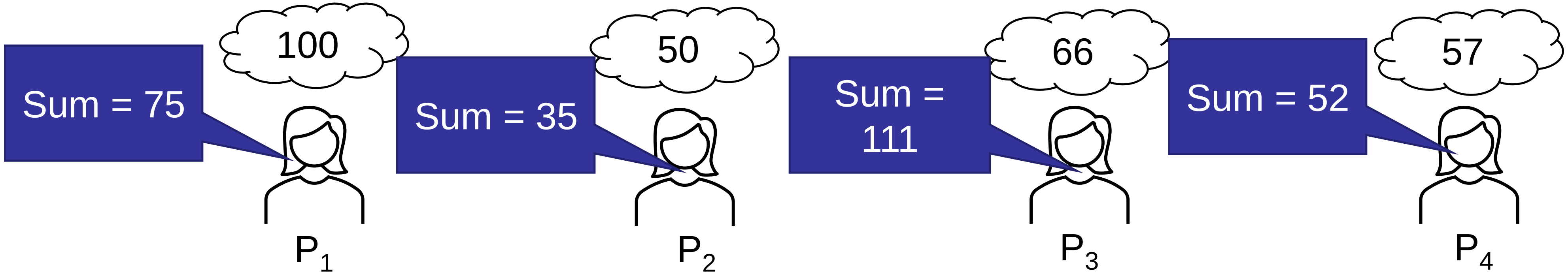
- For a number of participants $P_1,P_2,\ldots,P_m$, each has a private data, $X_1,X_2,\ldots,X_m$. The participants want to compute the value of a public function f on m variables at the point $X_1,X_2,\ldots,X_m$.
- A MPC protocol is called secure, if at the end of the computation, no participant knows anything except his own data and the results of the global calculation.

Secure Multi-party Computation (MPC)

Finding sum:

- Each participant splits its value and sends it to individual participant
- Yield their sum

	P_1	P_2	P_3	P_4
$X_1 = 100$	20	30	40	10
$X_2 = 50$	20	-30	30	30
$X_3 = 66$	15	15	21	15
$X_4 = 57$	20	20	20	-3



Secure Multi-party Computation (MPC)

MPC can also support functions like counting, finding max/min, solving regression over a larger data set (assume each participant hold a part of the data set).

Pros:

- high quality result while data miner has no extra access to data owned by other participants

Cons:

- Require collaborations from other participants
- More overheads
- Security rely on the honesty/vigilant of other participants

Concern in Data Provenance

❑ how to prevent unwanted disclosure of sensitive mining results

❑ how to evaluate the credibility of the received mining results

Approaches to Privacy Protection

- Legal measures. For example, making a contract with the data miner to forbid the miner from disclosing the mining results to a third party.

- The decision maker can utilize methodologies from data provenance, credibility analysis of web information, or other related research fields.

Data Provenance

- If the decision maker does not get the data mining results directly from the data miner, he would want to know how the results are delivered to him and what kind of modification may have been applied to the results, so that he can determine whether the results can be trusted.

- Data provenance refers to the information that helps determine the derivation history of the data, starting from the original source

- With such information, people can better understand the data and judge the credibility of the data.

Web Information Credibility

Because of the lack of publishing barriers, the low cost of dissemination, and the lax control of quality, credibility of web information has become a serious issue.

5 ways Internet users to differentiate false information from the truth:

1. Authority: the real author of false information is usually unclear.

2. Accuracy: false information does not contain accurate data

3. Objectivity: false information is often prejudicial.

4. Currency: for false information, the data about its source, time and place of its origin is incomplete, out of date, or missing.

5. Coverage: false information usually contains no effective links to other information online.

Conclusion

- ❑ Security: different measures to ensure data confidentiality, data integrity, and data availability
  - SQL Injection: a threat to database security. Avoided by proper programming practice
- ❑ Privacy: different measures to ensure personal information (either private or public) is not disclosed
  - Data collections
  - Data publishing
  - Data analysis
  - Data provenance

References

1. Lei Xu , Chunxiao Jiang , Jian Wang, Jain Yuan and Yong Ren, Information Security in Big Data-Privacy and Data Mining, Access, IEEE, 2014, 2: 1149-1176

2. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques.San Mateo, CA, USA: Morgan Kaufmann, 2006.

3. Evrim Oral, Surveying Sensitive Topics with Indirect Questioning, Statistical Methodologies, 2019.

4. Nataraj Venkataramanan, Ashwin Shriram, Data Privacy Principles and Practice, CRC Press, 2017.

5. L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.