3. Privacy

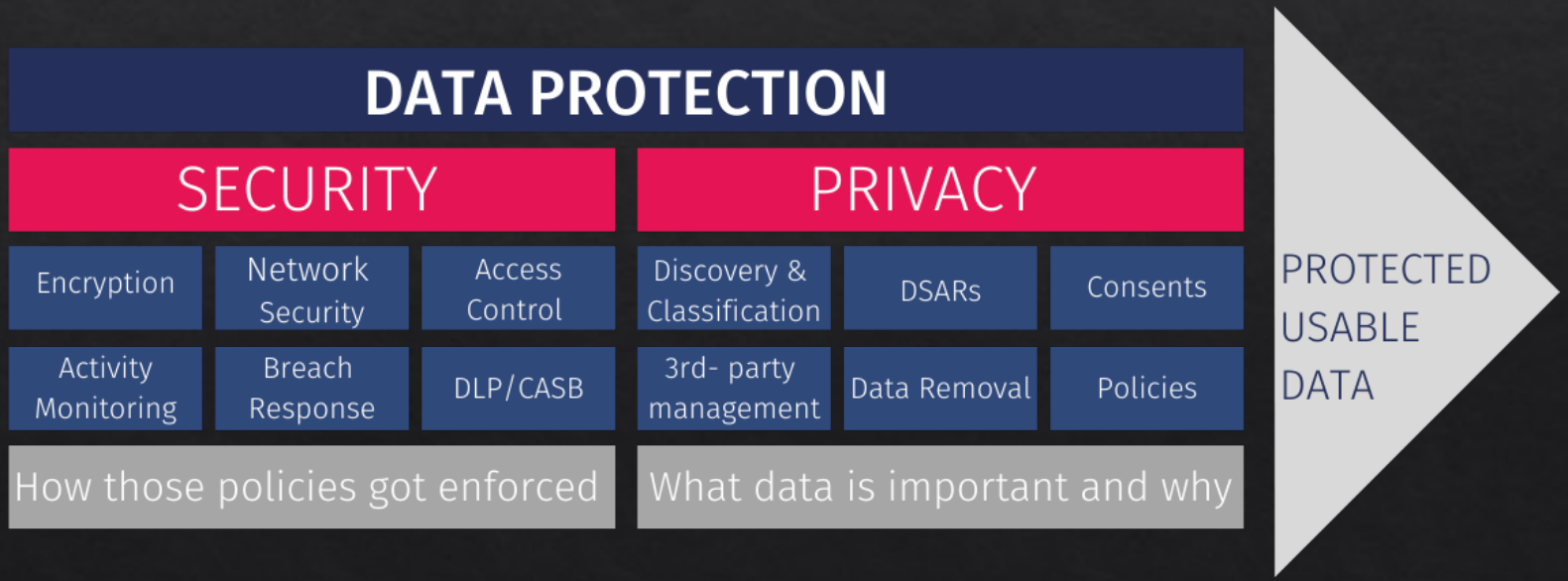
Jin Hong

jin.hong@uwa.edu.au

Privacy – what is it?

Privacy vs. security

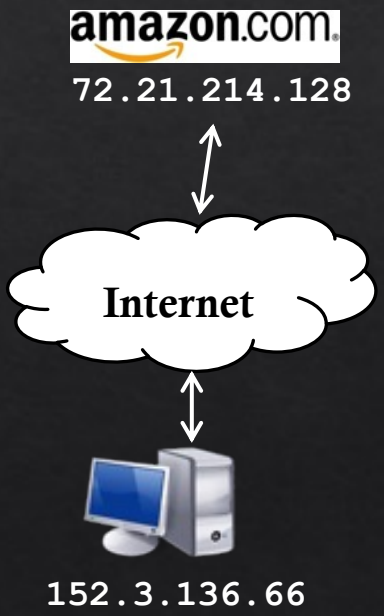
- Privacy – management of (your) data
- Security – preventing misuse of (your) data



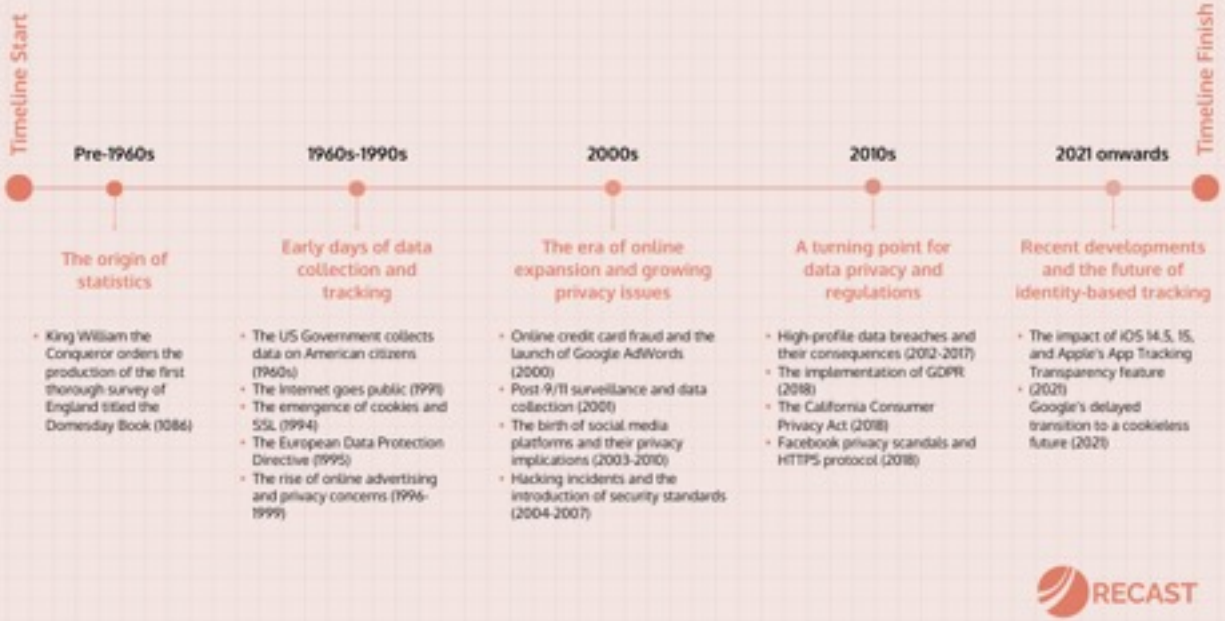
Privacy-sensitive information

Tracking on the web

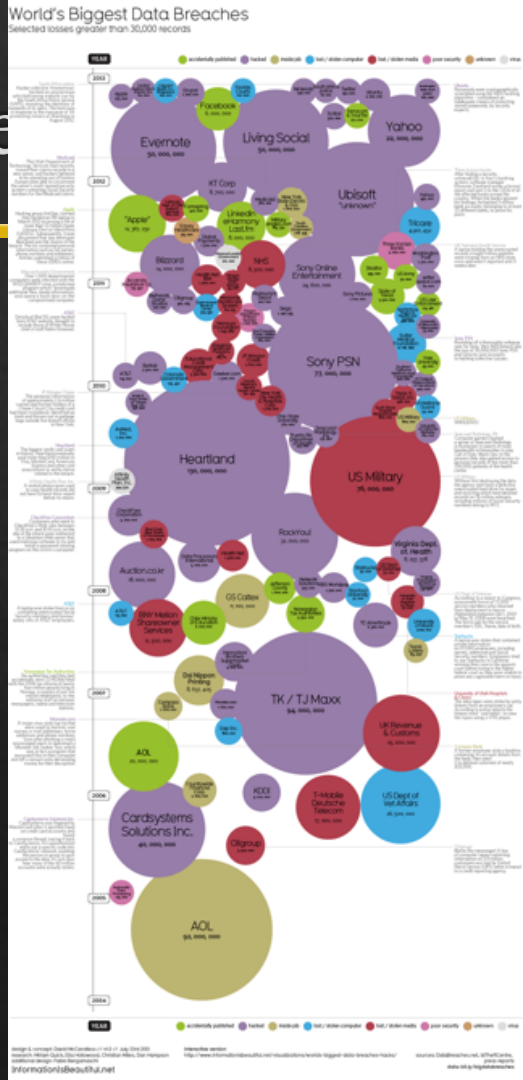
- IP address
  - Number identifying your computer on the Internet
  - Visible to site you are visiting
  - Not always permanent
- Cookies
  - Text stored on your computer by site
  - Sent back to site by your browser
  - Used to save prefs, shopping cart, etc.
  - Can track you even if IP changes



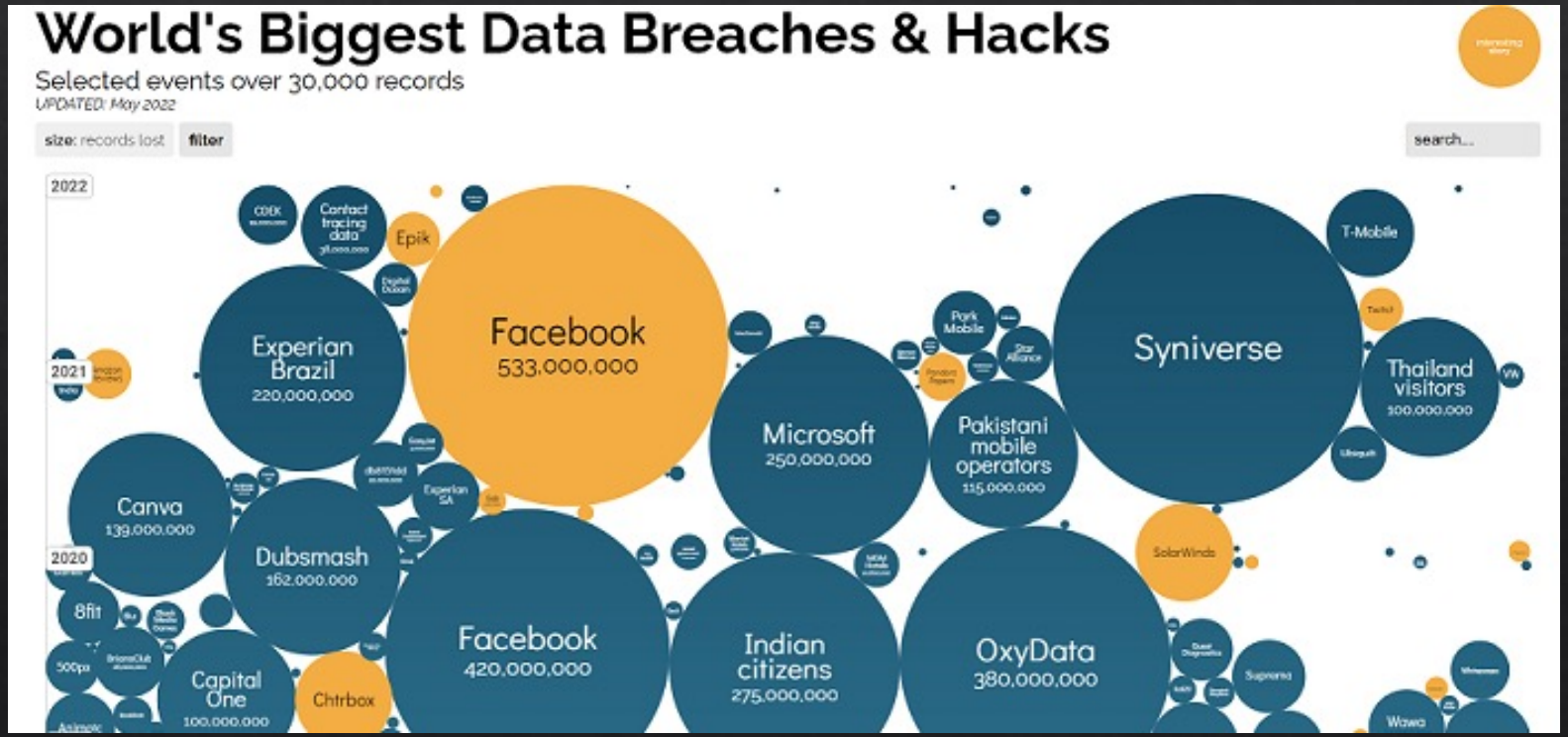
The Evolution of Privacy: A Timeline of Events Reshaping Identity-Based Tracking and Analytics



Data breach, data breach everywhere



Data breach, data breach everywhere!



Find Australian data breach here: https://www.webberinsurance.com.au/data-breaches-list

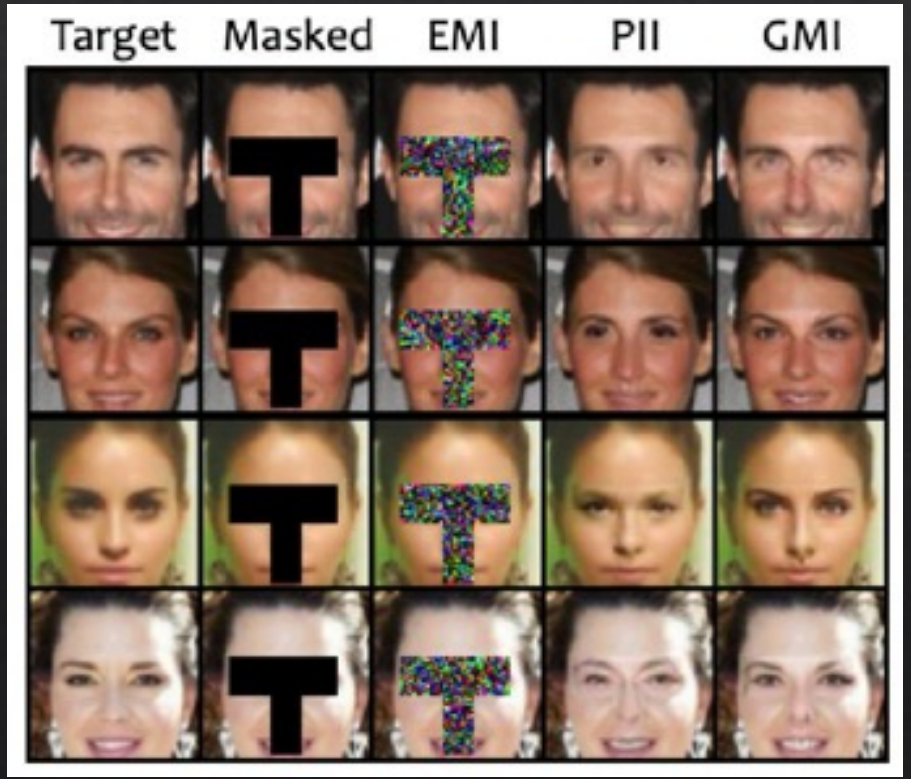
Threat: leaking sensitive information

- If we know Alice who is 55 years old female, what can we tell?

Name	Age	Gender	Zip Code	Smoker	Diagnosis
*	60–70	Male	191**	Y	Heart disease
*	60–70	Female	191**	N	Arthritis
*	60–70	Male	191**	Y	Lung cancer
*	60–70	Female	191**	N	Crohn's disease
*	60–70	Male	191**	Y	Lung cancer
*	***50–60***	***Female***	191**	N	HIV
*	50–60	Male	191**	Y	Lyme disease
*	50–60	Male	191**	Y	Seasonal allergies
*	***50–60***	***Female***	191**	N	Ulcerative colitis

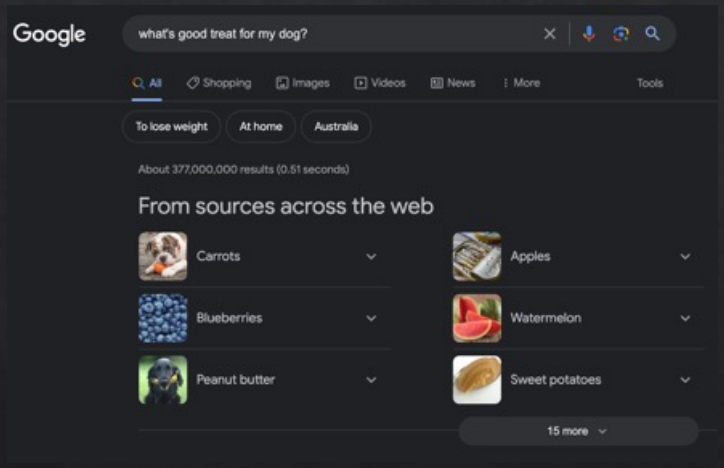
Threat: reidentification

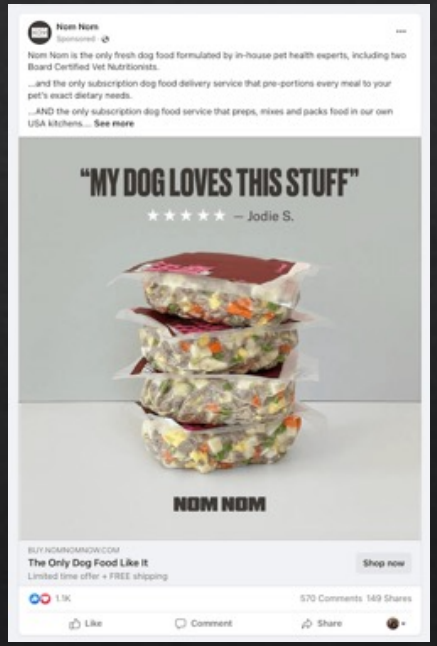
- ML can be used to reconstruct images (and other types of data)
- Col1: original
- Col2: prompt
- Col4: guess from public data
- Col5: reconstruction using ML



Source: Zhang et al., "The secret revealer: Generative model-inversion attacks against deep neural networks." https://arxiv.org/abs/1911.07135

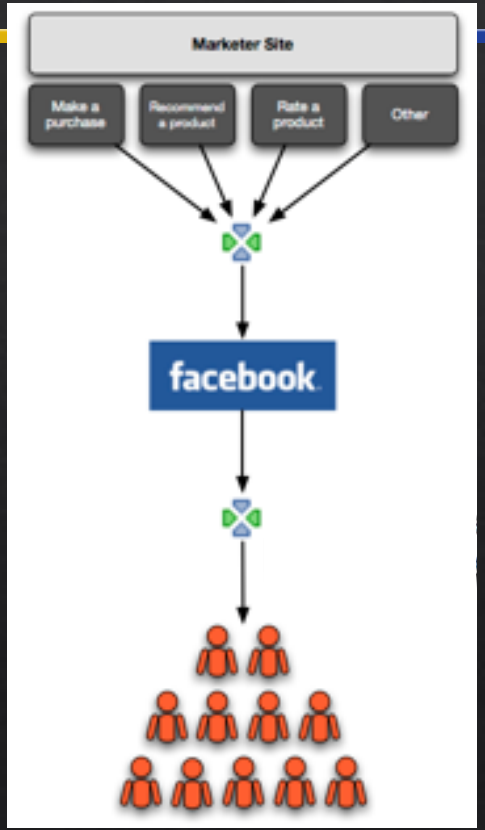
Threat: collusion among services





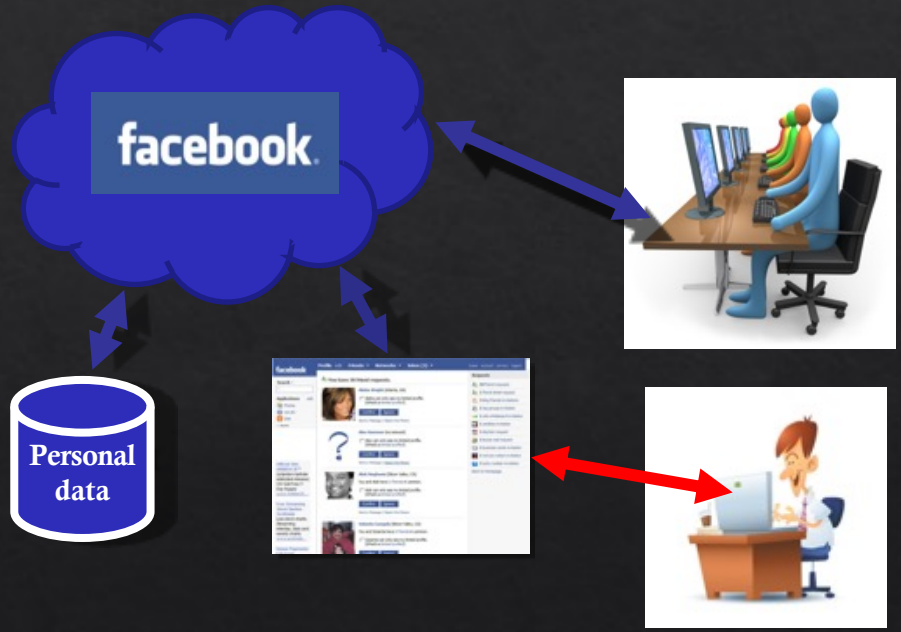
Threat: collusion among services

- Your search may not be private!
  - Analytic trackers will observe your search patterns and match with others to find products to recommend
  - The trackers can also see where the traffic is going
    - i.e., they may not be able to observe the content of your search, but knows where the traffic is going to
  - Trackers may collaborate to "enhance" user experience



Social networks

- Pros
  -
  -
- Cons
  -
  -



Centralized structure

Alternatives?

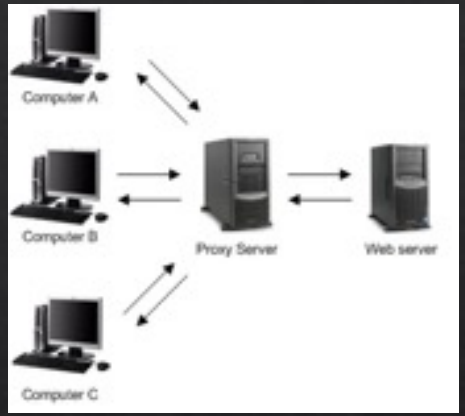
- Anonymisation
- Encryption
- Decentralization

Alternatives?

- Anonymisation
- Encryption
- Decentralization

Anonymisation

- Hide identity, remove identifying info
- Proxy server: connect through a third party to hide IP
- Health data released for research purposes: remove name, address, etc



Anonymisation techniques

- Data masking/pseudonyms
- Generalisation (e.g., K-anonymity)
- Perturbation (e.g., differential privacy)

Data masking and pseudonyms

- Simple to implement
  - Data masking – remove "sensitive" data from the record before sharing
  - Pseudonyms – replace "sensitive" data from the record before sharing
- However, these are more susceptible to deanonymisation attacks
  - E.g., re-identification attack

Threat: deanonymisation

- Netflix Prize dataset, released 2006
- 100,000,000 (private) ratings from 500,000 users
- Competition to improve recommendations
  - i.e., if user X likes movies A,B,C, will also like D
- Anonymised: username replaced by a number

Threat: deanonymisation

- Problem: can combine “private” ratings from Netflix with public reviews from IMDB to identify users in dataset
- May expose embarrassing info about members…

Threat: deanonymisation

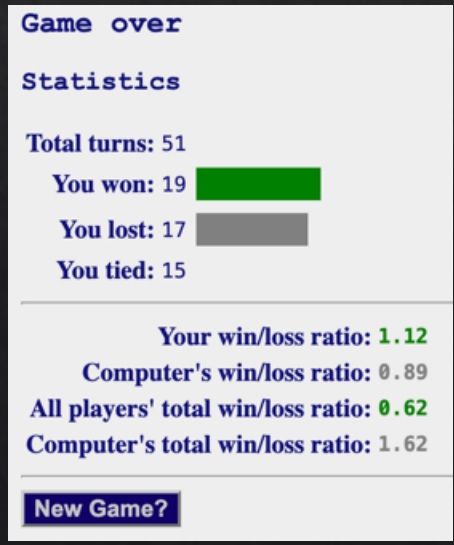
User	Movie	Rating
1234	Rocky II	3/5
1234	The Wizard	4/5
1234	The Dark Knight	5/5
…		
1234	**Girls Gone Wild**	5/5

User	Movie	Rating
dukefan	The Wizard	8/10
dukefan	The Dark Knight	10/10
dukefan	Rocky II	6/10
…		

Threat: deanonymisation

- Lesson: cannot always anonymise data simply by removing identifiers
- Vulnerable to aggregating data from multiple sources/networks

- Humans are (usually) predictable
  - E.g., try Rock-paper-scissors vs AI
    - https://www.essentially.net/rsp/
  - E.g., try …



Generalisation

- You can transform the data to be more generalised
  - i.e., it is more difficult to distinguish data in records
- Aggregation is one type of generalisation technique
  - E.g., combine multiple data together
  - E.g., aggregate the minutes-electricity meter readings to hourly interval
- K-anonymity is also a generalisation approach
  - This is a broader approach, multiple methods could be adopted to achieve k-anonymity
  - E.g., create groups from provided data points
  - E.g., age groups instead of using specific age values

Generalisation: Limitation

- You lose precision/details
- Analysis relying on the precision of the data cannot function so well
  - E.g., activity recognition
  - E.g., https://github.com/adrienpetralia/ApplianceDetectionBenchmark
  - E.g., https://dl.acm.org/doi/10.1145/3486611.3486650

Perturbation

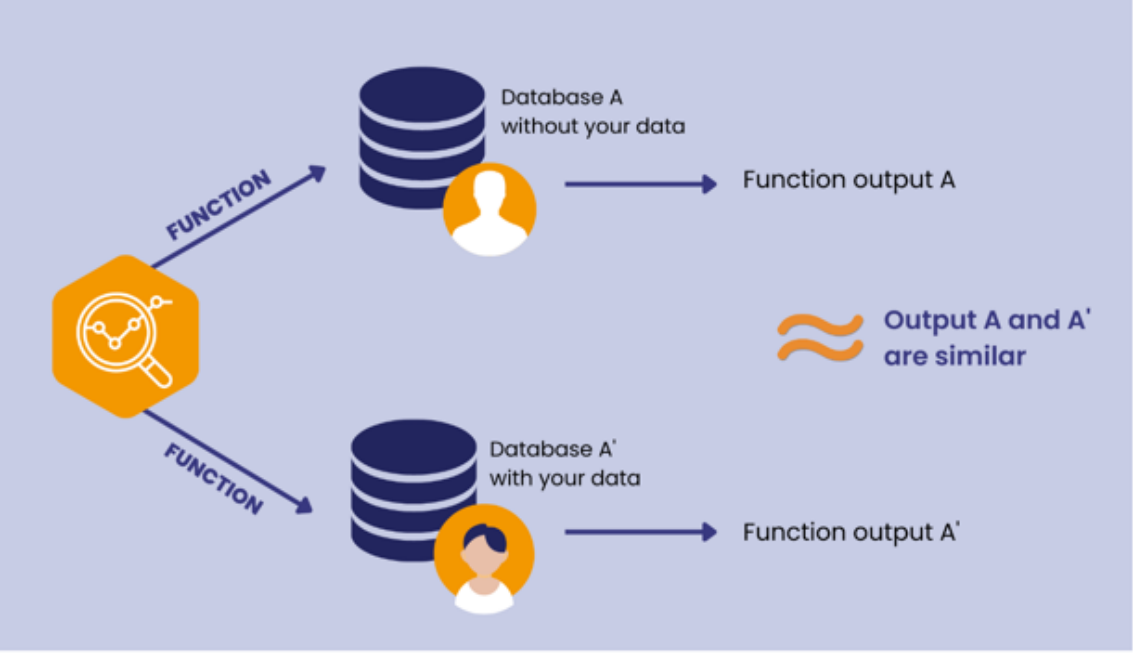
- This approach is to modify the content of the data
- Altering the original data may hinder the use of dataset
  - E.g., billing vs additive perturbation
  - E.g., activity recognition vs random perturbation

- Random perturbation is not that bad for the mean of the population (statistically speaking)
- But it does increase the variance a lot, which limits its usability in practice
  - But we won't go into statistical details here…

- To address this issue, one approach is to use differential privacy (DP)

Differential Privacy

- Remember this from CITS1003?



Differential Privacy

X: The data *universe*.

$D \subset X$: The dataset (one element per person)

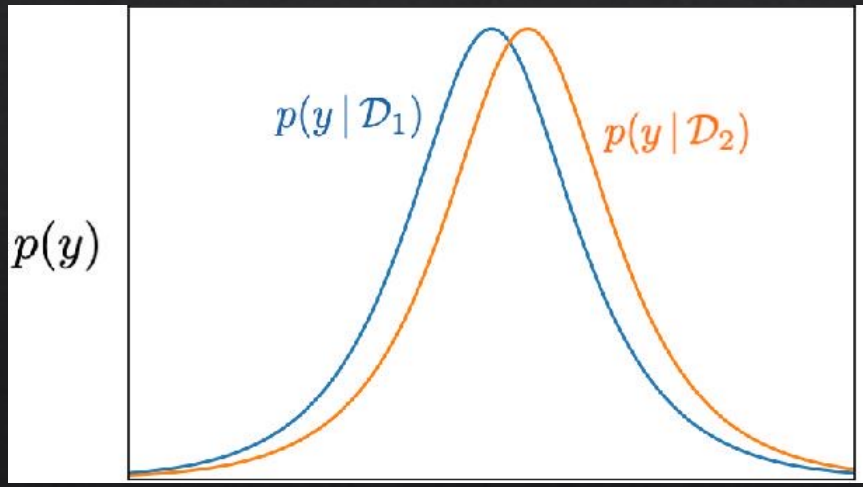
Definition: Two datasets $D, D' \subset X$ are *neighbors* if they differ in the data of a single individual.

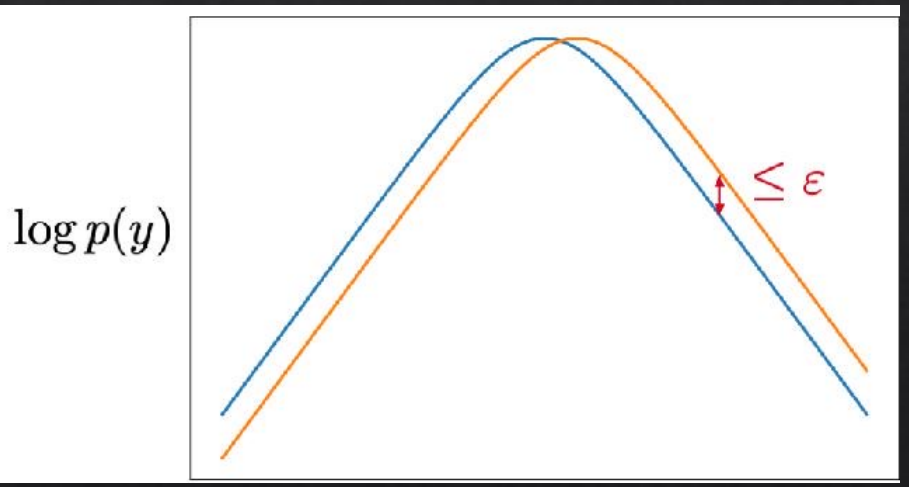
Definition: An algorithm M is ϵ-differentially private if for all pairs of neighboring datasets D, D', and for all outputs x:

$$\Pr[M(D) = x] \leq (1 + \epsilon) \Pr[M(D') = x]$$

Differential Privacy

- What does it look like?





Differential Privacy

- A commonly used method to achieving DP is to use Laplace mechanism.
- It adds noise to the output of a function (not the data), and the amount of noise depends on the sensitivity specified.
- Other approaches include randomisation and other perturbation methods.

Differential Privacy: What is ϵ?

- Perspective taken by theory: Pick ϵ, prove (or evaluate) accuracy.
- Realities of practice: Hard accuracy requirements.
  - Find the smallest level of ϵ consistent with accuracy targets.

Differential Privacy: Limitations

- Noise in the data still reduces the utility of the data
  - But in this case, we know the error is bounded by epsilon
- The way DP is applied in different scenarios are inconsistent
  - The setup of noise addition, privacy budget, etc. are all dependent on the person(s) who set it up

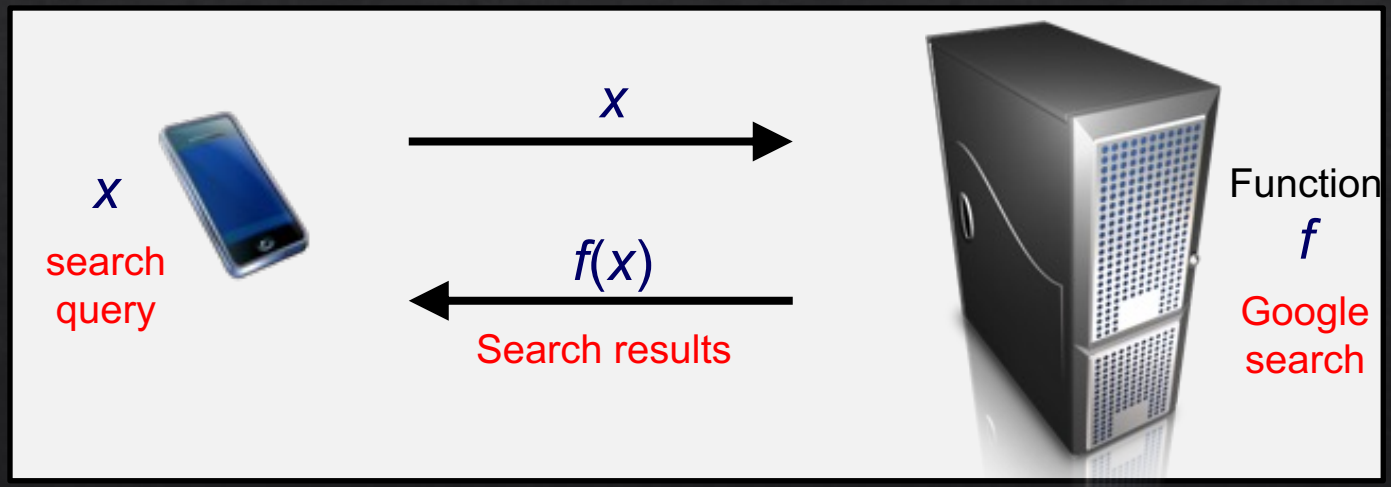
Alternatives?

- Anonymisation
- Encryption
- Decentralization

Cryptographic

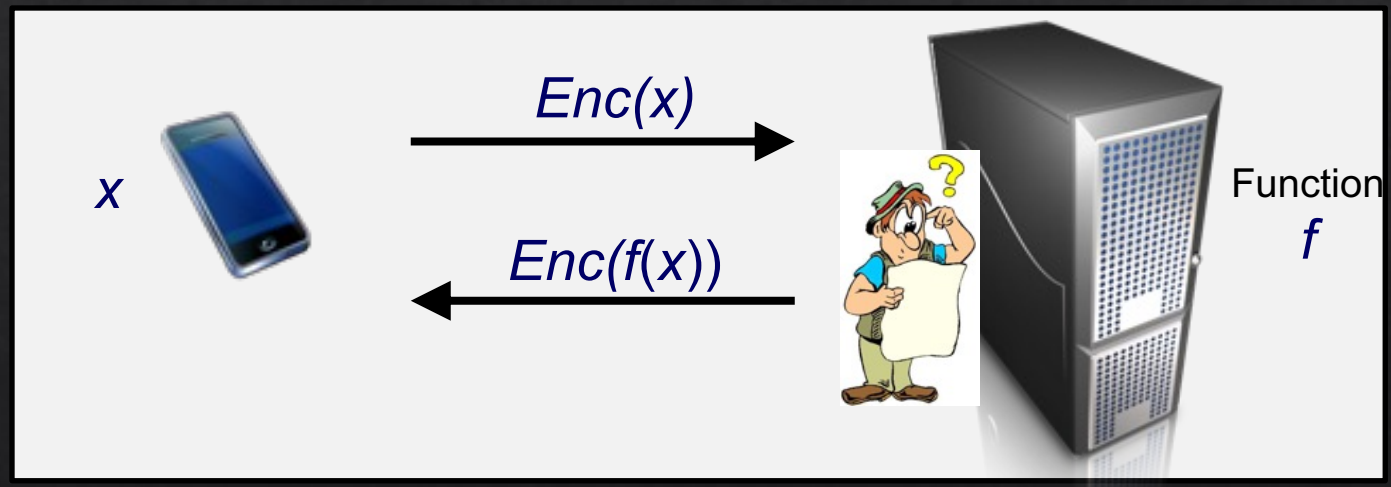
- Homomorphic encryption (HE)
- Multiparty encryption (MPE)
- Zero-knowledge proofs (ZKP)
- …

Homomorphic Encryption (HE)



WANT PRIVACY!

Homomorphic Encryption (HE)



WANT PRIVACY!

Homomorphic Encryption (HE)

Some people noted the algebraic structure in RSA…

$$E(m_1) = {m_1}^e \qquad E(m_2) = {m_2}^e$$

Ergo … $E(m_1) \times E(m_2)$

$$= {m_1}^e \times {m_2}^e$$

$$= (m_1 \times m_2)^e$$

$$= E(m_1 \times m_2)$$

Multiplicative Homomorphism

$$E(m_1) \times E(m_2) = E(m_1 \times m_2)$$

Computing on Encrypted Data

What people really wanted was the ability to do arbitrary computing on encrypted data…

… and this required the ability to compute *both* sums *and* products …

… on the same encrypted data set!

Computing on Encrypted Data

Why SUMs and PRODUCTs?

SUM = XOR

0 XOR 0	0
1 XOR 0	1
0 XOR 1	1
1 XOR 1	0

PRODUCT = AND

0 AND 0	0
1 AND 0	0
0 AND 1	0
1 AND 1	1

Computing on Encrypted Data

Because {XOR, AND} is Turing-complete …

… any function is a combination of XOR and AND gates

XOR

0 XOR 0	0
1 XOR 0	1
0 XOR 1	1
1 XOR 1	0

AND

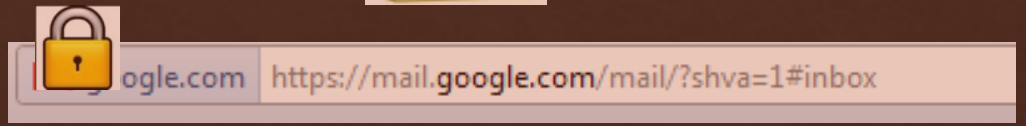
0 AND 0	0
1 AND 0	0
0 AND 1	0
1 AND 1	1

Computing on Encrypted Data

Fully-Homomorphic Encryption!

Amazing Applications:

Private Cloud Computing



Delegate *arbitrary processing* of data

without giving away *access* to it

Homomorphic Encryption (HE)

- You can utilise existing libraries!
  - Don't reinvent the wheel
    - Unless it's a new improved type of wheel
  - They have different capabilities and performances, so choose ones that are fit for the purpose
  - Continuously researched area, so stay tuned

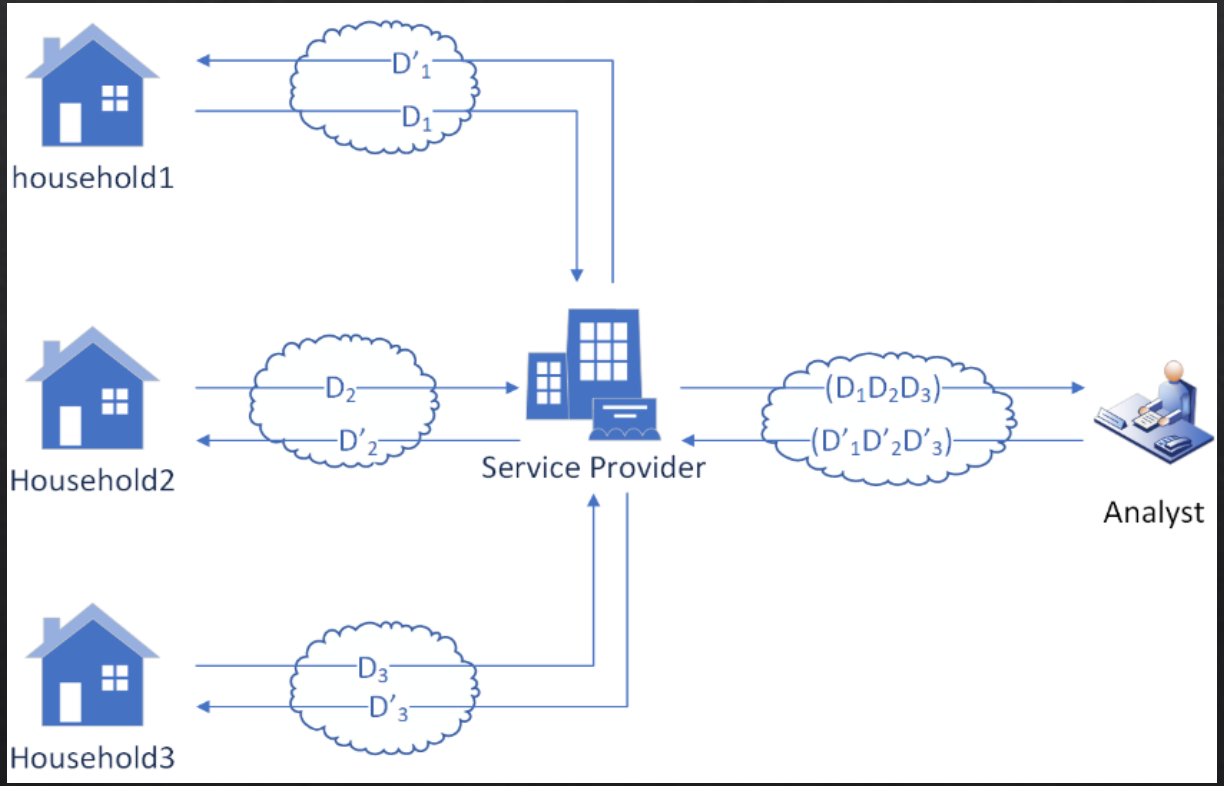
FHE libraries

Name	Developer	BGV[19]	CKKS[37]	BFV[22]	FHEW[34]	CKKS Bootstrapping[46]	TFHE[35]	Description
HElib[46]	IBM	Yes	Yes	No	No	No	No	BGV scheme with the GHS optimizations.
Microsoft SEAL[47]	Microsoft	Yes	Yes	Yes	No	No	No	
OpenFHE	Duality Technologies, Samsung Advanced Institute of Technology [kr], Intel, MIT, University of California, San Diego and others.	Yes	Yes	Yes	Yes	Yes	Yes	Successor to PALISADE.
PALISADE[48]	New Jersey Institute of Technology, Duality Technologies, Raytheon BBN Technologies, MIT, University of California, San Diego and others.	Yes	Yes	Yes	Yes	No	Yes	General-purpose lattice cryptography library. Predecessor of OpenFHE.
HEAAN[49]	Seoul National University	No	Yes	No	No	Yes	No	
FHEW[34]	Leo Ducas and Daniele Micciancio	No	No	No	Yes	No	No	
TFHE[35]	Ilaria Chillotti, Nicolas Gama, Mariya Georgieva and Malika Izabachene	No	No	No	No	No	Yes	
FV-NFLlib[50]	CryptoExperts	No	No	Yes	No	No	No	
NuFHE[51]	NuCypher	No	No	No	No	No	Yes	Provides a GPU implementation of TFHE.
REDcuFHE[52]	TwC Group	No	No	No	No	No	Yes	A multi-GPU implementation of TFHE.
Lattigo[53]	EPFL-LDS, Tune Insight	Yes	Yes	Yes	No	Yes[54]	No	Implementation in Go along with their distributed variants[55] enabling Secure multi-party computation.
TFHE-rs[56]	Zama	No	No	No	No	No	Yes	Rust implementation of TFHE-extended. Supporting boolean, integer operation and univariate function evaluation (via Programmable Bootstrapping[57]).

Homomorphic Encryption (HE)

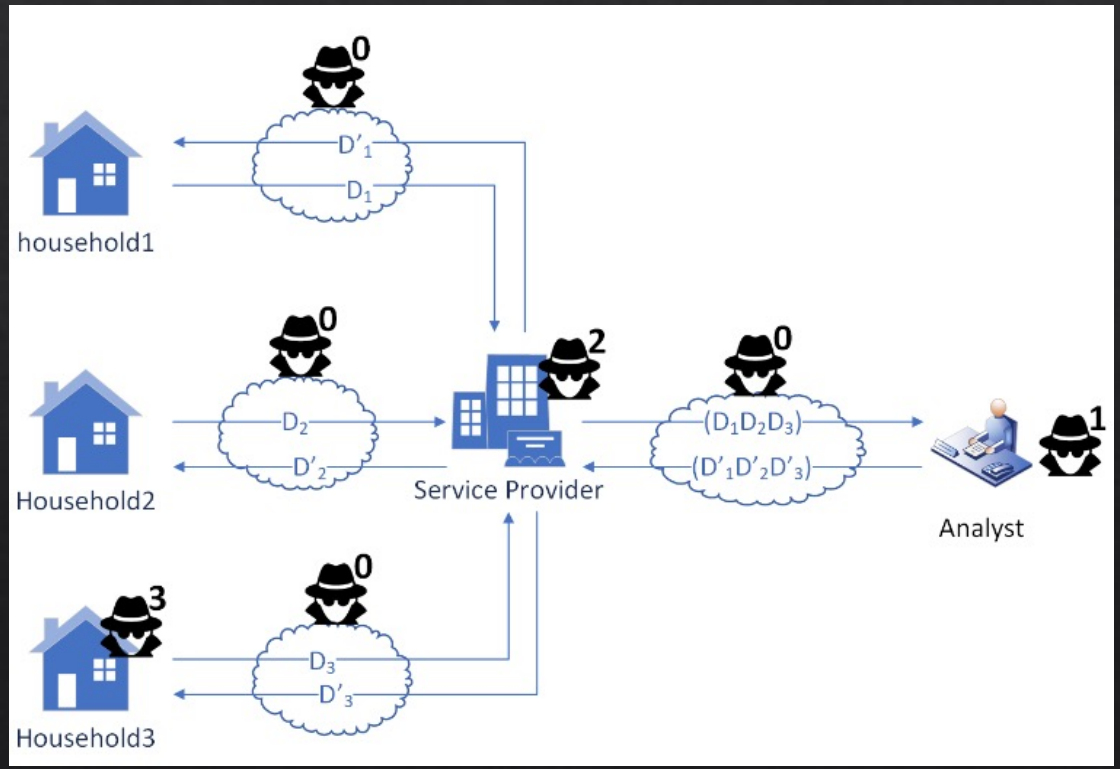
- From this point, it gets very "mathy" to truly understand how HE works.
- But this isn't a math unit so we won't go into details but rather look at how to utilise this.

HE in SMS


household1
D'1
D1
Household2
D2
D'2
Service Provider
(D1D2D3)
(D'1D'2D'3)
Analyst
Household3
D3
D'3

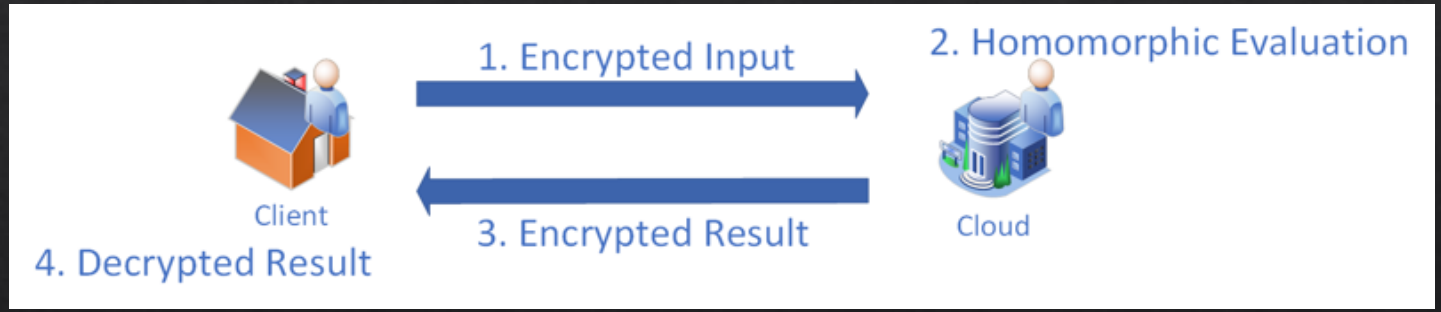
HE in SMS

- Privacy issues:
  - Data eavesdropping
    - Activity recognition
    - Appliance inference
    - Membership inference attack
  - Data tampering



HE in SMS

- Solution?
  - Use HE!



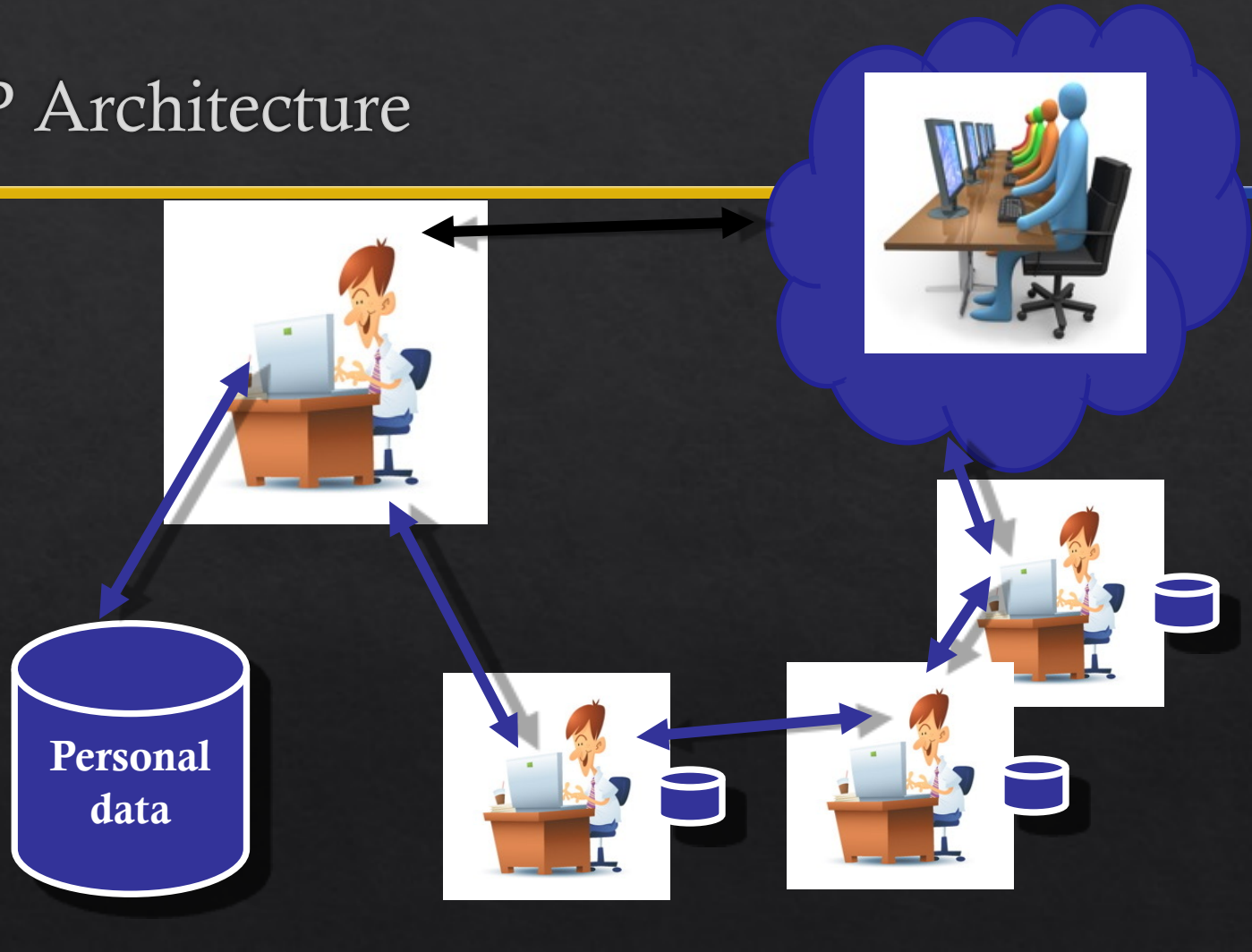
Key size	Plaintext	Ciphertext	Encryption	Decryption
1024	6.87MB	287.64MB	216.87s	68.63s
2048	6.87MB	527.17MB	1152.98s	357.17s
3072	6.87MB	754.62MB	3111.14s	993.80s

Selective Encryption

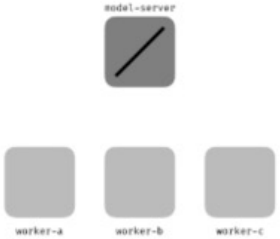
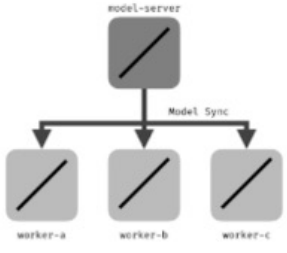
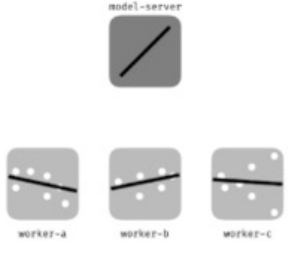
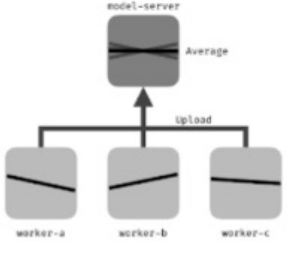
Alternatives?

- Anonymisation
- Encryption
- Decentralization

P2P Architecture



Federated Learning

Step 1	Step 2	Step 3	Step 4
			
Central server chooses a statistical model to be trained	Central server transmits the initial model to several nodes	Nodes train the model locally with their own data	Central server pools model results and generate one global mode without accessing any data

Decentralization: pros and cons

Decentralization: Takeaways

- Decentralized network structure can enhance privacy
- Difficult to achieve true anonymity
- Fine-grained control over data can help
  - Tension with usability

Conclusion

- Various approaches to provide privacy to users online
- Privacy-preserving techniques have evolved over time, but their adoption is very slow due to complexity and costs
- A typical performance/cost vs security/privacy trade-off exists for considerations
- No one technique outperforms the others – requires careful consideration of the choice
- Being aware of the latest advances in the area is key to staying private and safe

References

- CPS 96 – Duke University
- CSC2515 – University of Toronto
- Vinod Vaikuntanathan for HE contents