ONLINE AND OFFLINE DATA PRIVACY

SPIROS ANTONATOS

antonat (class99)

APRIL 2022

DEFINITION OF PRIVACY

a state in which one is not observed or disturbed by other people

the state of being free from public attention

Information **privacy** is the right to have **some** control over how your personal information is collected and used

Data privacy is focused on the use and governance of personal data—things like putting policies in place to ensure that consumers' personal information is being collected, shared and used in appropriate ways.

What is Privacy (iapp.org)

HOW MANY OF YOU USE FACEBOOK?

OR INSTAGRAM? OR WHATSAPP? OR VIBER? OR GMAIL? OR YOUTUBE? OR AMAZON?.....
OR JUST CALLING FROM YOUR MOBILE?

WHEN I LAST ATTENDED THIS CLASS..

…approximately 2005

1 years old

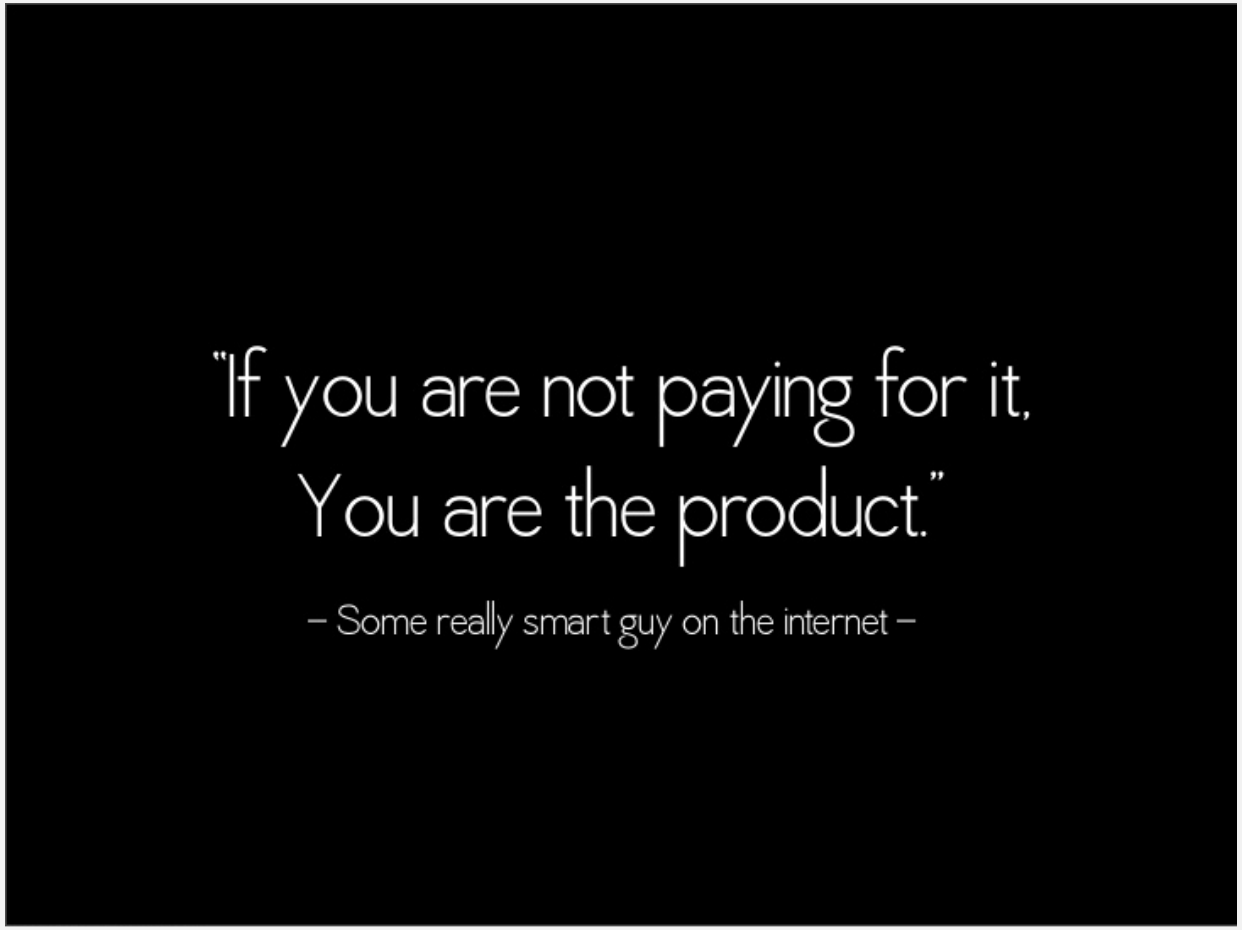
2 years old

2 years old

< 1 years old

WHEN IS THE LAST TIME YOU GAVE SOMETHING FOR FREE?

AND HOW VALUABLE WAS IT?



IT IS ALL ABOUT SELLING THE DATA

- You interact with a site
- Your activities generate a certain profile
- Your profile is shared to an advertising network
- Sites get money when you click on an advertisement
- Your data can be used for generating insights for better services (what is the most visited location in town?)

HOW MANY OF YOU HAVE YOU READ THE TERMS OF SERVICE BEFORE PRESSING "I ACCEPT"?

"A Deloitte survey of 2,000 consumers in the U.S found that 91% of people consent to legal terms and services conditions without reading them. For younger people, ages **18-34** the rate is even higher with 97% agreeing to conditions before reading."*

Deloitte study: 91 percent of Americans agree to terms of service without reading (businessinsider.com)

WHAT KIND OF ACTIVITIES CHARACTERIZE ME?

- Location data – Cell towers, WiFi
- Demographics – Profile page information
- Queries on search engines - Google, Bing
- Your social network updates – Facebook, Instagram
  - Not only yours but also your friends'
- Your purchases - Amazon, Ebay, credit card history
- Your watched videos - YouTube

I AM OLD SCHOOL! I HAVE A FLIP PHONE!

- Even if you have an old phone
- Even if you do not text or call
- Still you update your location with your provider every few minutes – this is called paging

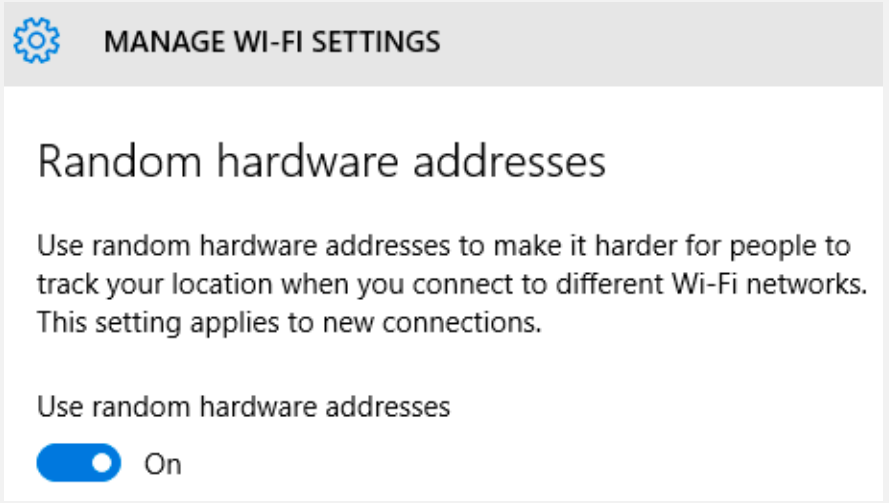
Just four points of reference, with fairly low spatial and temporal resolution, was enough to uniquely identify 95 percent of 1.5 million phones

https://www.nature.com/articles/srep01376

WHAT OTHER IDENTIFIERS ARE THERE?

- IP Address
- MAC Address
- Browser capabilities
- Cookies
- Other HW identifiers

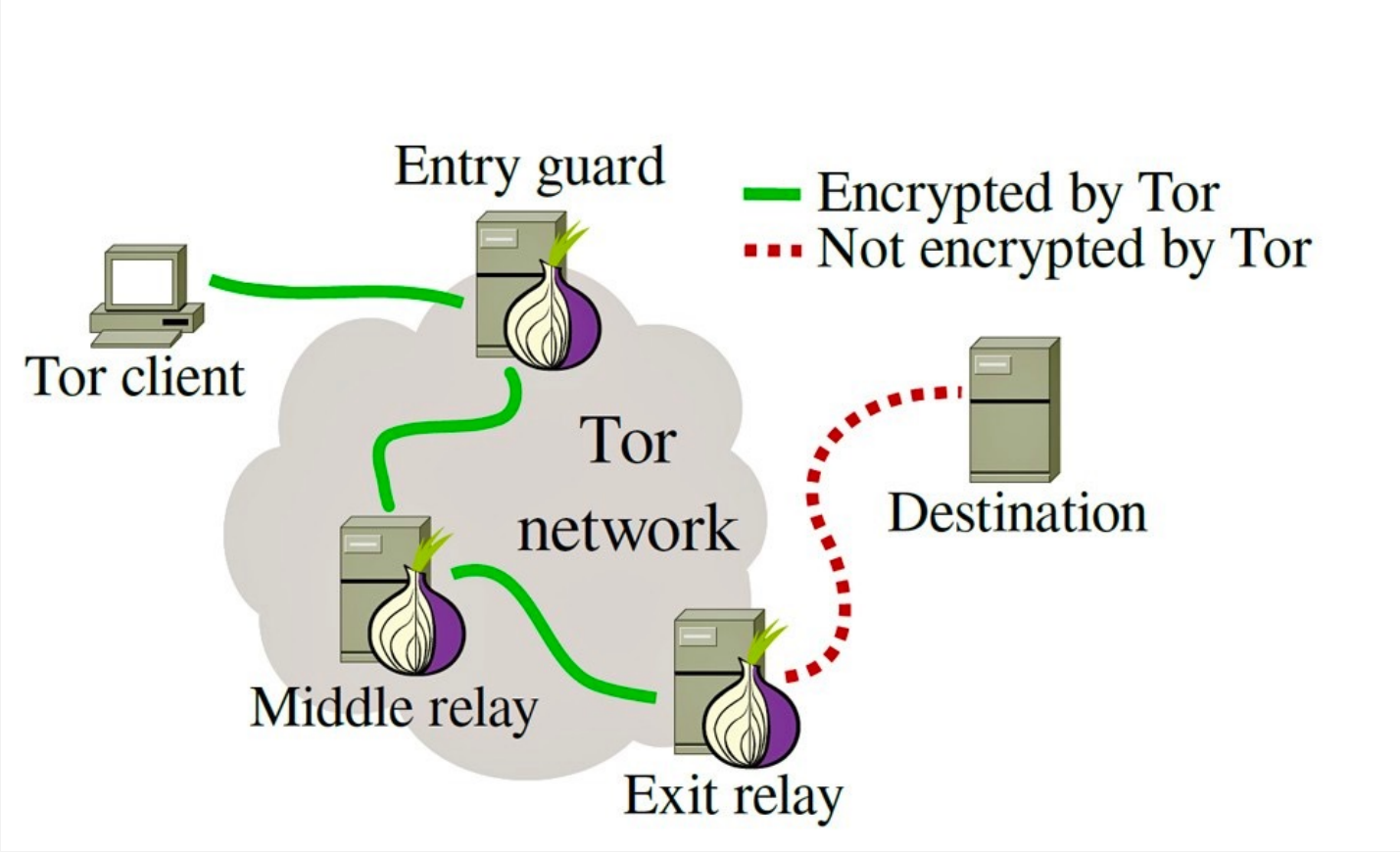
Try: https://amiunique.org/fp



WHAT CAN I DO TO PROTECT MY PRIVACY?

- Tor
- Use services only when necessary
- Login-less./ Cookie-less activities (incognito)
- Do not overshare

TOR



- Traffic is proxied through Tor nodes
- Multiple layers of encryptions
- Each node peels off one layer and forwards the message
- Requires a Tor client installation
- Tor client strips out identifying information
- Within Tor a node can either know who you are or where you are going but not both

Warning: DNS traffic is not relayed through Tor*

*This might change with the adoption of DNS-over-HTTPS

- https://www.torproject.org

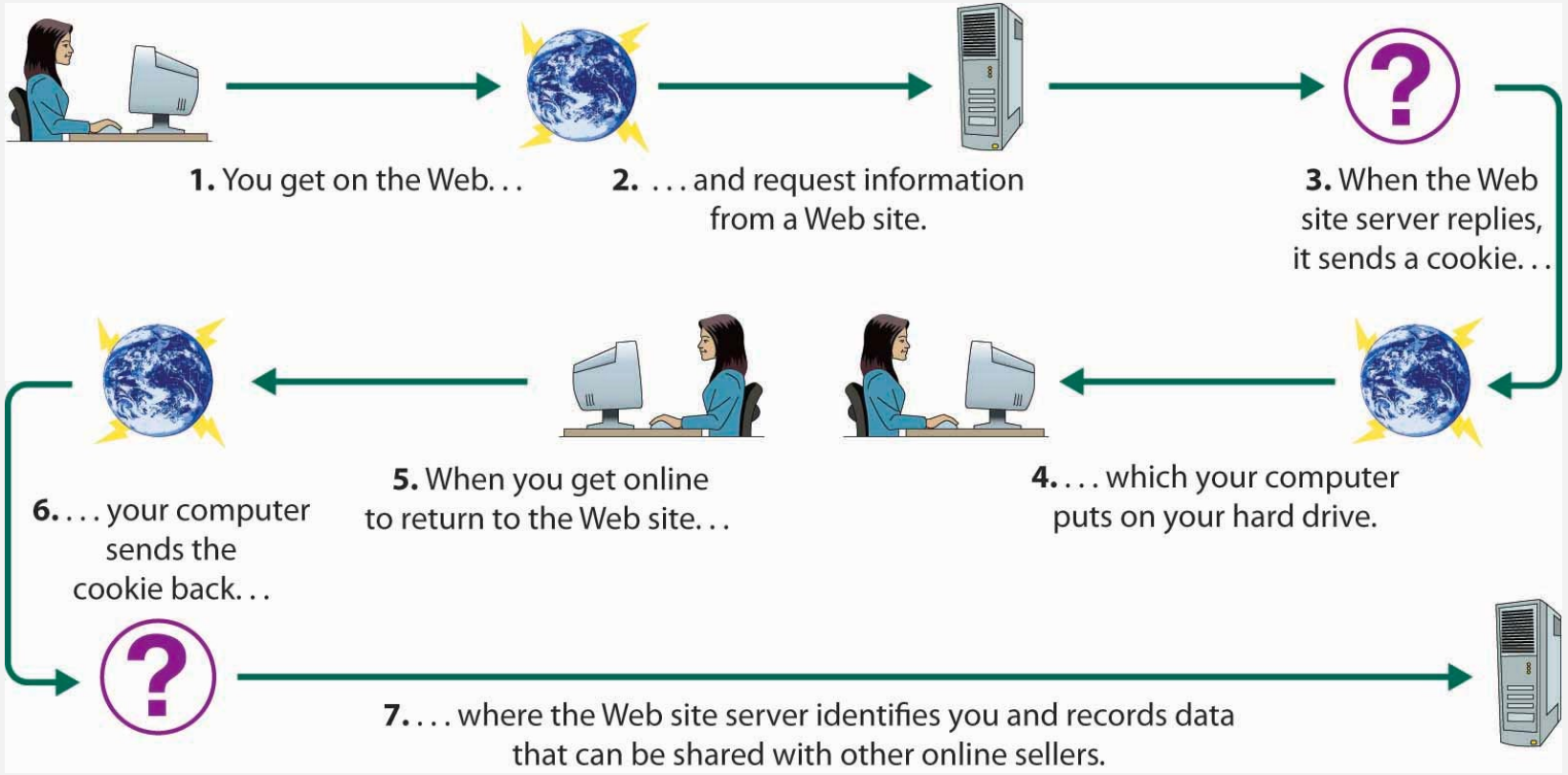
VPN AND TOR CANNOT PROTECT ME?

- VPN only changes your IP address
  - The rest of your traffic remains the same
- The VPN service still knows who you are
- Numerous attacks against Tor
  - Timing attacks*
  - Browser-based attacks**
  - Rogue nodes
  - Traffic correlation

* Preventing Active Timing Attacks in Low-Latency Anonymous Communication | SpringerLink

** Browser-Based Attacks on Tor | SpringerLink

FEWER COOKIES



- Block third party cookies
- Install addons like Privacy Badger or NoScript
- Do not accept cookies when prompted
- Interesting development: Google to ‘phase out’ third-party cookies in Chrome, but not for two years

DO NOT OVERSHARE

- Roughly ~1% of tweets include geolocation*
- Full list of OSINT techiques: OSINT Techniques - Home

*World Tweeting Tendencies in Real-time: geolocalized tweets


ALWAYS HOPE, THERE IS

GDPR

[Infosecurity.US - https://infosecurity.us - Becca Rick's 'How Paypal Shares Your Data](https://infosecurity.us)

The legislation aims to give back to individuals located in the EU control over their Personal Data and simplify the regulatory environment for international business.

May 25, 2018

Global Impact

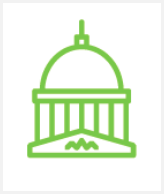
4% or €20M

Potential penalty for non-compliance

5 Key General Data Protection Regulation Obligations

Rights of EU Data Subjects

Security of Personal Data



Consent

Accountability of Compliance

Data Protection by Design and by Default

WHEN DOES THE GDPR APPLY

The GDPR territorially applies to processing of Personal Data by:

- organizations having establishments located in the EEA
  - The nationality of the data subject or where the data is being provided from is irrelevant
  - No matter if the organization is acting as a Controller or a Processor

- organizations located outside the EEA offering goods or services to individuals in an EEA Member State
  - No matter if the goods and services are offered for payment or for free
  - Whether or not goods and services are offered to individuals in the EEA has to be assessed based on the actual circumstances (e.g., language, currency, shipment, top level domain)
  - The nationality of the data subject is irrelevant; only a data subject's physical location is relevant

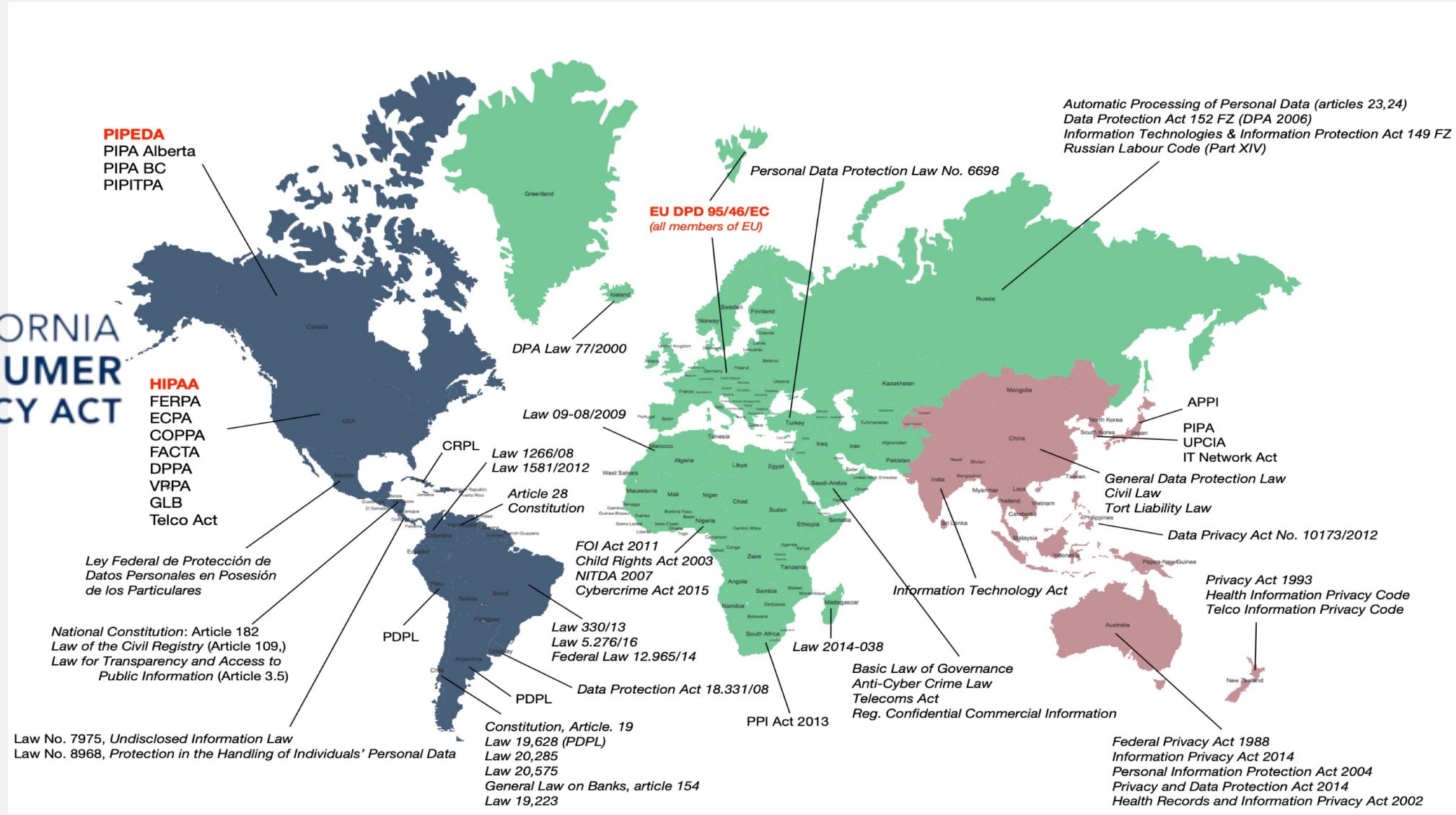
- organizations located outside the EEA monitoring individuals' behaviors taking place in the EEA
  - e.g., tracking of individuals on the internet, profiling an individual
  - The nationality of the data subject is irrelevant; only a data subject's physical location when being monitored is relevant

- organizations located outside the EEA if EEA Member State law applies by virtue of public international law
  - e.g., a member state's diplomatic mission or consular post, airplanes or ships registered under the flag of a member State of the EEA

WHAT IT MEANS FOR YOU

- No more vague terms of use, no more tiny fonts
- You have the right to be forgotten
- You can ask the companies for the full set of data they have regarding you
- You can ask the companies with whom they share the data and why
- You can stop them from sharing your data through consent

THE WORLDWIDE DATA PROTECTION/PRIVACY LANDSCAPE



CALIFORNIA CONSUMER PRIVACY ACT

DATA PRIVACY

WHAT IS DATA PRIVACY?

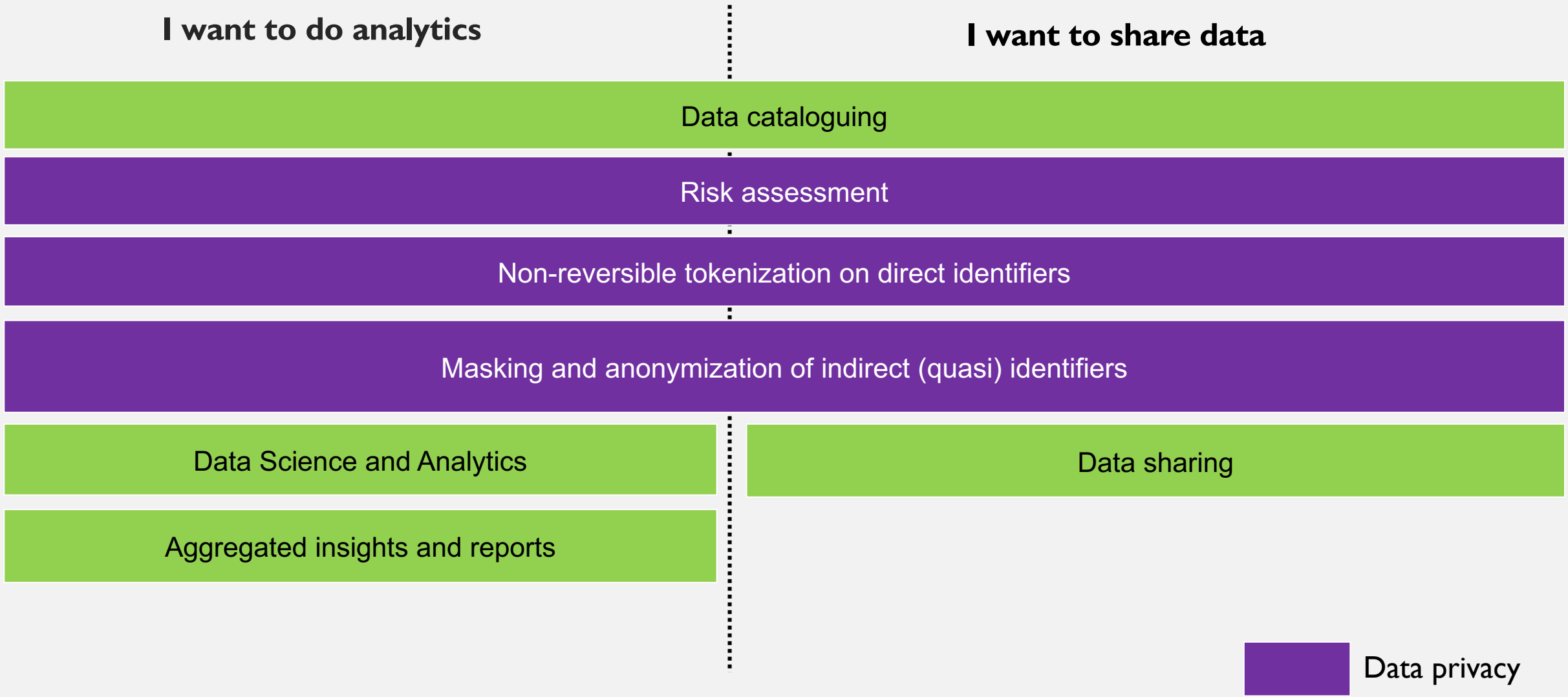
- A methodology to anonymize data sufficiently
- Applied when there is no need to know the identity of the users
- Mainly for performing analytics and sharing data

WHAT WE PROTECT AGAINST

Singling out	Linkability	Inference
Definitions*		
• "Singling out" occurs where it is possible to **distinguish** the data relating to one individual from all other information in a dataset. This may be because information relating to one individual has a **unique value**	Any linking of identifiers in a data set will make it more likely that an individual is identifiable. A major risk factor which may lead to the identification of individuals from anonymised data is the risk of data from one or more other sources being combined or matched with the anonymised data.	• In some cases, it may be possible to infer a link between two pieces of information in a set of data, even though the information is not expressly linked. This may occur, for example, if a dataset contains statistics regarding the seniority and pay of the employees of a company.

* https://www.dataprotection.ie/docs/Anonymisation-and-pseudonymisation/1594.htm

DATA PRIVACY: A USE-CASE BASED APPROACH



DIRECT AND INDIRECT IDENTIFIERS

PID	Name	Address	SSN	Birth	Gender	ZIP	Marital status	A1	A2	…
0	Maria	10 NY E. Avenue	473-57-8129	09/64	Female	**94139**	**Divorced**	…	…	…
1	Jenny	5 Brighton Street	472-37-1179	09/64	Female	94138	Divorced	…	…	…
2	Nick	12 Doyle Ave.	423-67-8159	04/64	Male	94138	Widow	…	…	…
3	Tom	154 West End Av.	123-50-8129	04/64	Male	94139	Married	…	…	…
4	John	93 Somers Str.	533-53-8178	**03/63**	**Male**	**94139**	Married	…	…	…
5	Bob	35 University Av.	832-39-8146	**03/63**	**Male**	**94138**	Married	…	…	…
6	Sarah	63 Mirror Street	664-38-8138	09/64	Female	94141	Married	…	…	…
7	Eleni	67 Common Av.	727-37-8194	09/61	Female	94141	Married	…	…	…
8	Dave	65 Main Str.	312-89-8389	05/61	Male	94138	Single	…	…	…
9	Thomas	84 Main Ave.	289-78-1899	05/61	Male	94138	Single	…	…	…

Direct identifiers

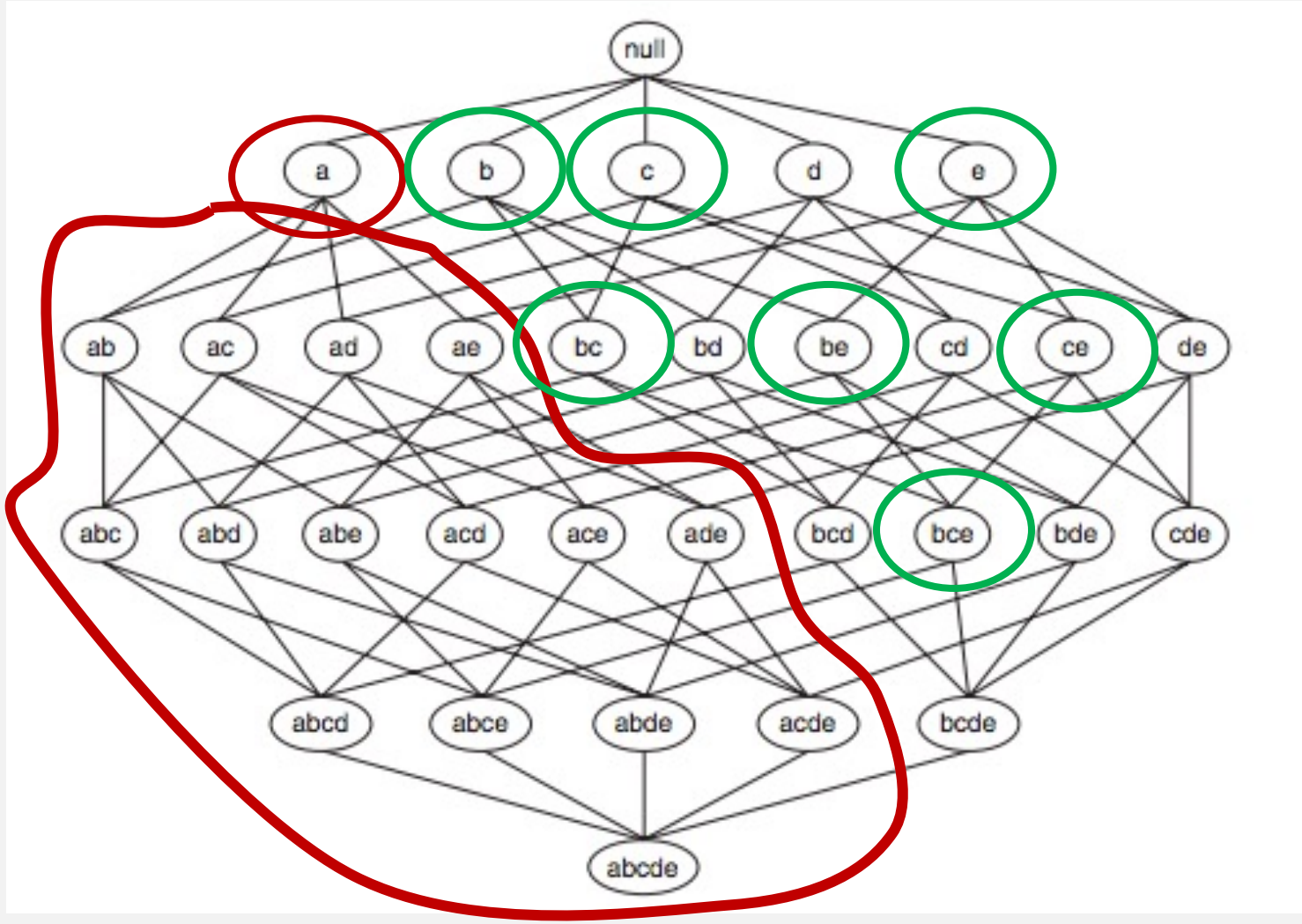
Single columns that contain unique values

Quasi-identifiers

Combination of columns that contain unique values

Other information

HOW DO WE DISCOVER IDENTIFIERS?



- N columns means 2^N combinations
- If a column has unique values, then all combinations of that column has unique values
- If a combination of columns has no unique values, then all subsets have no unique values

Arvid Heise, Jorge-Arnulfo Quiané-Ruiz, Ziawasch Abedjan, Anja Jentzsch, and Felix Naumann. 2013. Scalable discovery of unique column combinations. Proc. VLDB Endow. 7, 4 (December 2013), 301–312. DOI:https://doi.org/10.14778/2732240.2732248

FPVI: A scalable method for discovering privacy vulnerabilities in microdata. Aris Gkoulalas-Divanis and Stefano Braghin and Spyridon Antonatos. *IEEE International Smart Cities Conference (ISC2), pp. 1-8*, 2016

WHAT IS TOKENIZATION

- A token is a surrogate value used in place of an underlying sensitive value.
- The ***Tokenisation*** of PI is called ***Pseudonymization***
- **Token attributes** –the physical structure of the token. This includes
  - The length of the token, character encoding (ASCII, UTF-8, UTF-16, EBSDIC, etc.), general character set (Latin -1, …), suppressed characters such as '110O', special formats, leading zeros in numeric fields, uniqueness of the token
- **Token utility** –what you can do with the token
  - Irreversible (1- way) vs reversible (2-way) tokens, Application semantics preserving, Key evolving tokens,
  - Domain aware tokens
- **Token security** –the security value of the token
  - The security strength of the underlying algorithm, key length, HSM use, key management security

WHAT IS MASKING, WHEN AND HOW IT IS USED

- Replaces original values with fictionalized ones
- Includes techniques like generalization, binning, offsetting, randomization, redaction and many more
- No overlap with tokenization

Example:

user123@Hotmail.com -> abcde@unknown.com

- **Data masking properties**
- **Utility-preserving:** ensures that the fictionalized values maintain useful information encoded in the original value
- E.g. domain names in URLs/e-mail addresses, chapter/categories in ICD, location proximity etc.
- **Compound data masking**: extracts relationships among data fields and ensures that the fictionalized values preserve these relationships
- E.g. an entry and discharge date in a hospital record, a city and a country field etc.
- **Consistent data masking**: ensures that all instances of an original value will be replaced by the same fictionalized value

LINKING RISK EXAMPLE

Case: an attacker can link the data with other public/private datasets

For example, demographics tables can be linked with census datasets, transactions can be linked with merchant logs

Question: how do I discover datasets that are linkable?

Goal: Assess the linking risk through dataset linking

Gender	ZIP code	Date of birth
Male	12345	1985-07-04
Female	67890	2001-01-10

My dataset

Only 3 records linked, we have 33% chance to find the real identity

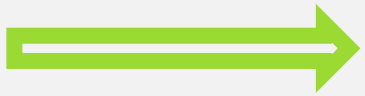
Name	Gender	ZIP	Date of birth
John Doe	Male	12345	1985-07-04
Jack Doe	Male	12345	1985-07-04
Peter Doe	Male	12345	1985-07-04
Name1	Female	67890	2001-01-10
Name 2	Female	67890	2001-01-10
…			
Name 1000	Female	67890	2001-01-10

This record links with 1000 records, so we have a 0.1% chance to recover the real name

WHAT IS K-ANONYMITY, WHEN AND HOW IT IS USED

Age	Sex	Disease
20	M	HIV
23	F	HIV
25	M	Obesity
27	F	HIV
28	F	Cancer
29	F	Obesity

Attributes we need to protect



Age	Sex	Disease
[20-25]	M	HIV
[20-25]	M	Obesity
[23-27]	F	HIV
[23-27]	F	HIV
[28-29]	F	Cancer
[28-29]	F	Obesity

This table holds individual values

Very easy to find an invidivual from the raw data

Anonymization process ensures that data are indistinguishable under certain conditions on their identifiable attributes (quasi-identifiers)

We transform data in such a way that each pair of age and sex values appears at least 2 times

Risk -> 1/2

We can generalize the approach to create groups of N elements

Risk -> 1/N

WHAT IS WRONG WITH THIS ANONYMIZATION APPROACH?

Age	Sex	Disease
20-25	M	HIV
20-25	M	Obesity
23-27	F	HIV
23-27	F	HIV
28-29	F	Cancer
28-29	F	Obesity

This group is vulnerable to inference attacks

HOW CAN WE SOLVE THE INFERENCE PROBLEM?

Age	Sex	Disease
20-25	M	HIV
20-25	M	Obesity
23-29	F	HIV
23-29	F	HIV
23-29	F	Cancer
23-29	F	Obesity

This group is now extended and contains 3 distinct diseases so we only have 33% chance to infer the diseases

A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24

DIFFERENTIAL PRIVACY

- Individual privacy preserved
- Population trends still observable
- No assumptions on attacker
- Privacy is future proof

A SIMPLE EXAMPLE OF DIFFERENTIAL PRIVACY: WHO IS FAN OF JUSTIN BIEBER?

Participant	Actual answer		Noisy answer
A	0	→	1
B	0	→	0
C	1	→	0
D	1	→	1
⋮	⋮		⋮
Z	1	→	0
Total	17	→	16

Published data

- Individual values are not reliable
- No way to reconstruct originals
- Aggregate statistics still representative

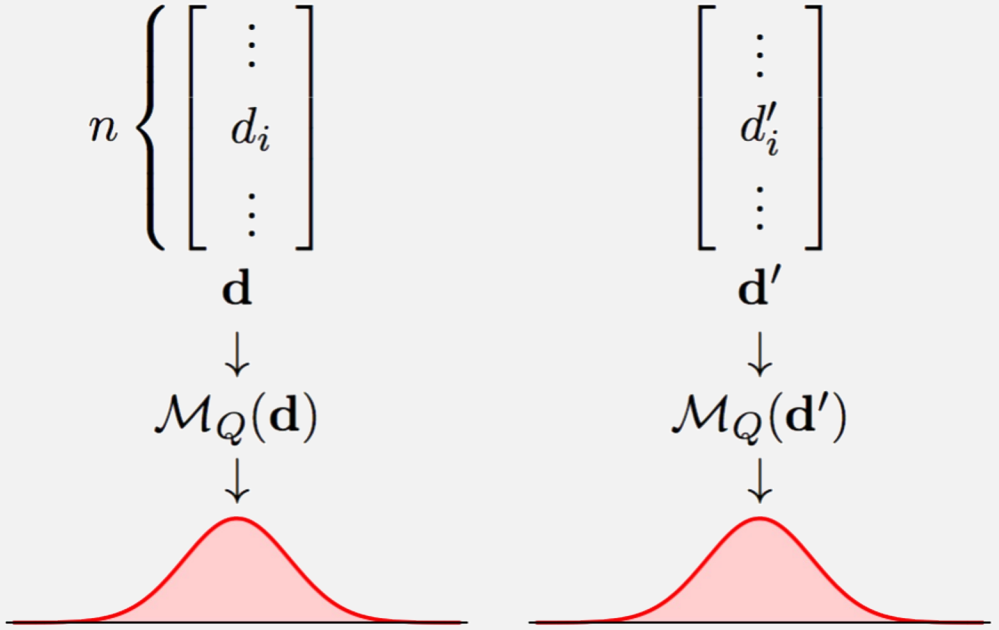
Model parameters control privacy/accuracy tradeoff

ϵ-DIFFERENTIAL PRIVACY

Extending ϵ-Indistinguishability

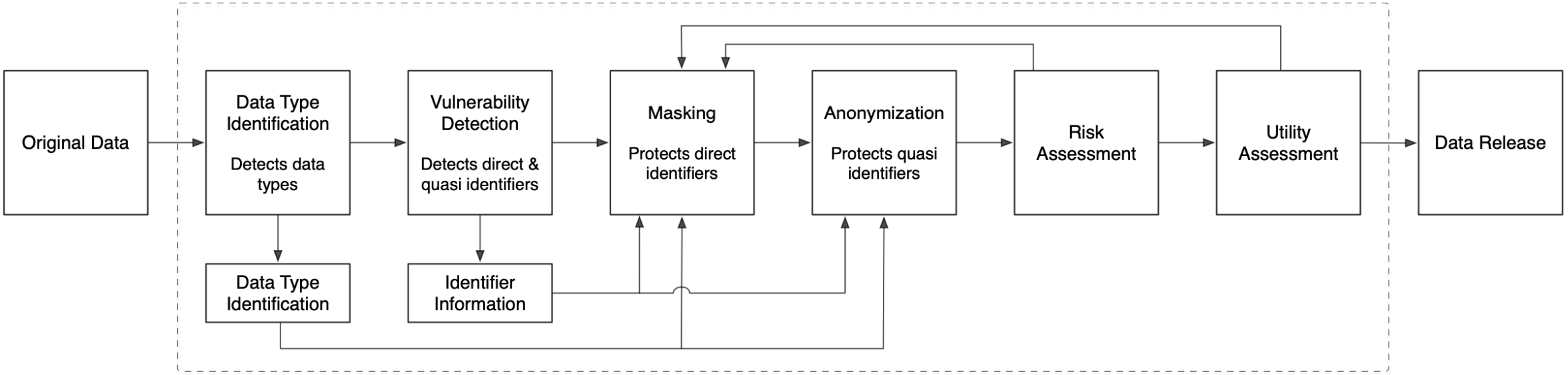
- ϵ-Differential privacy is a specific application of ϵ-indistinguishability
- Statistical indistinguishability between two datasets differing in one record
- Can be applied to:
  - Query answers (data mining, PPDM), or
  - Entire dataset (data publishing, PPDP)
- Noise must be large enough to hide presence/absence of the most extreme values
- Ensures plausible deniability
- Gives a standardised privacy measurement

In a nutshell



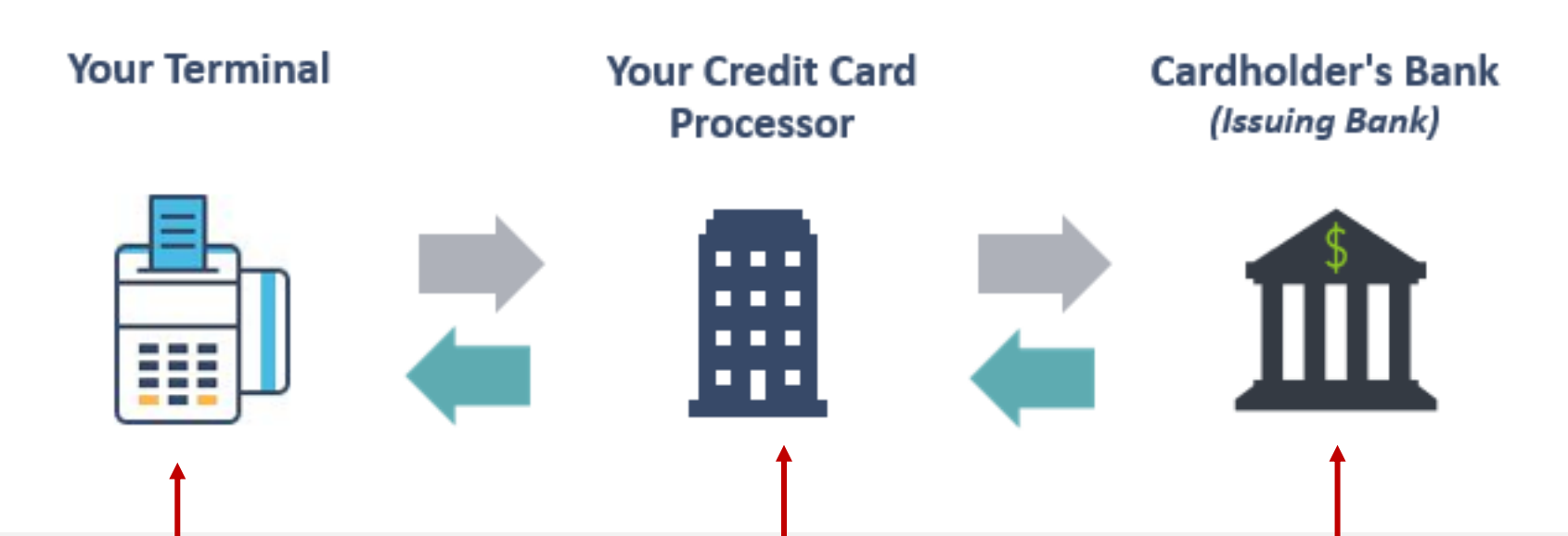
$$\mathbb{P}\left(\mathcal{M}_Q(\boldsymbol{d}) \in A\right) \leq e^{\epsilon}\mathbb{P}\left(\mathcal{M}_Q(\boldsymbol{d}') \in A\right) + \delta$$
$$\forall A \subseteq Range(Q)$$

A DATA PRIVACY PIPELINE



S. Antonatos, S. Braghin, N. Holohan, Y. Gkoufas and P. Mac Aonghusa, "PRIMA: An End-to-End Framework for Privacy at Scale," *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris, France, 2018

A REAL WORLD EXAMPLE: DATA TRUST (1/2)



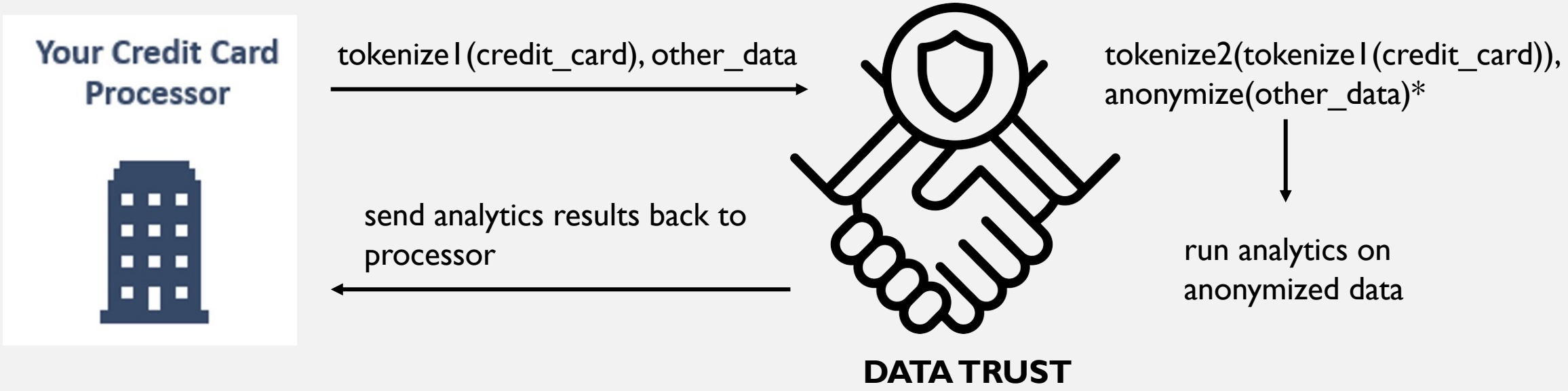
Your credit card number is personal information

Your credit card processor only knows your credit card number

Your bank has all the details for you including the information on how to reach you and ask for consent

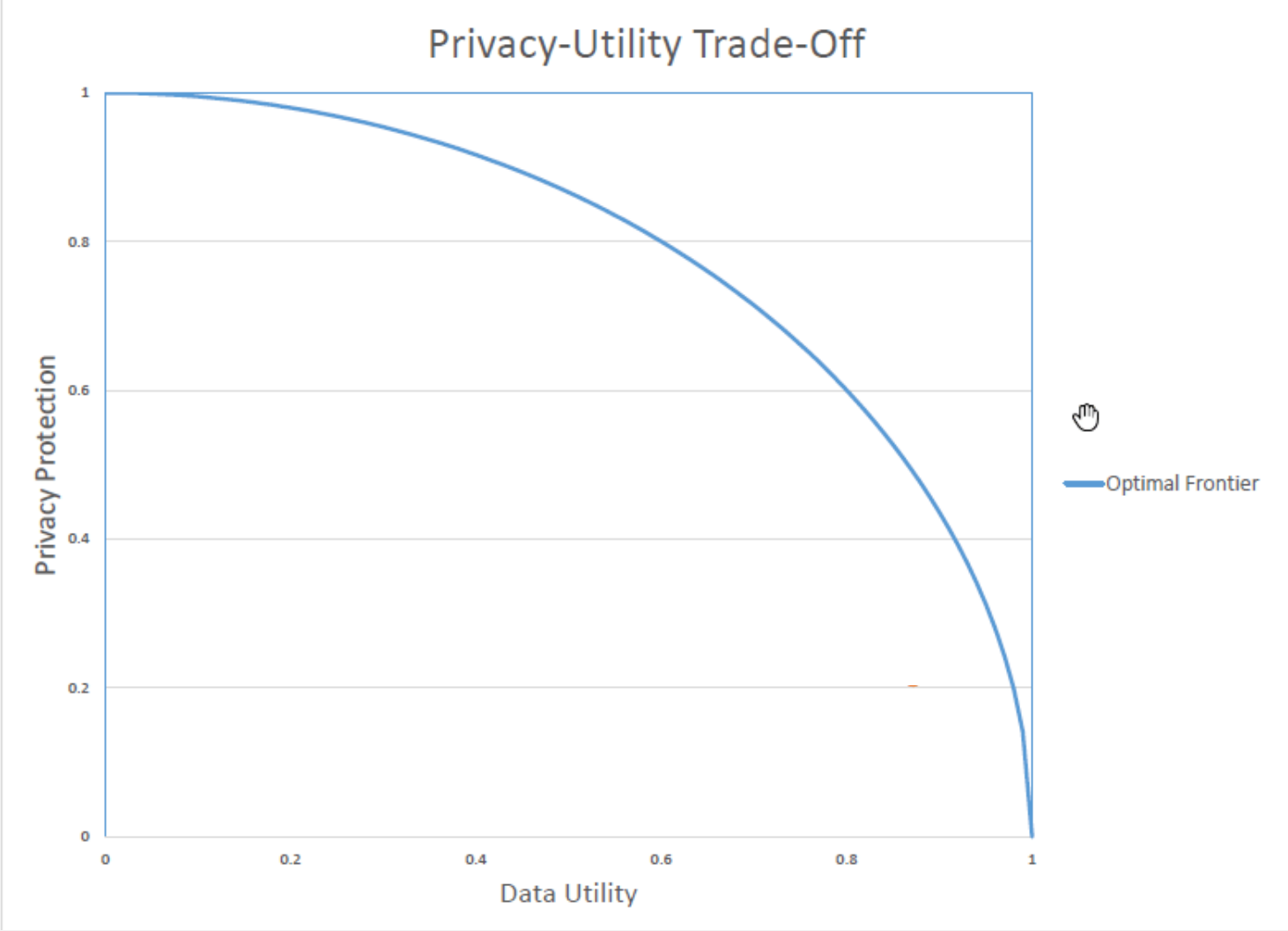
Q: How will the credit card processors run analytics if they cannot reach you and ask for your consent?
A: The **GDPR** does not apply to anonymised information.

A REAL WORLD EXAMPLE: DATA TRUST (2/2)



* GDPR has no technical directives for anonymization techniques. Data Trust entity ensures that anonymization is sufficient (no PII, minimum linking risk)

YOU CAN'T BEAT THE CURVE



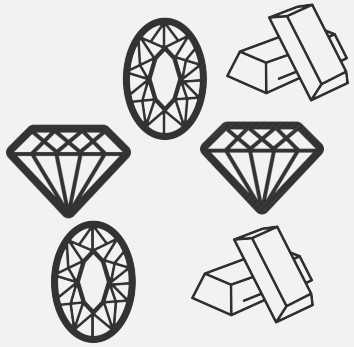
OPEN RESEARCH QUESTIONS

- How do we provide anonymization to unstructured data with statistical guarantees?
  - NLP + named entity extraction is not enough
  - How about images, video and audio?
- Longitudinal data are hard to get anonymized (at scale)
- Impact of anonymization on ML/AI
- Risk-utility measurement tools still at trial and error mode

SECURE COMPUTATIONS

HOMOMORPHIC ENCRYPTION

Alice runs a jewelry store

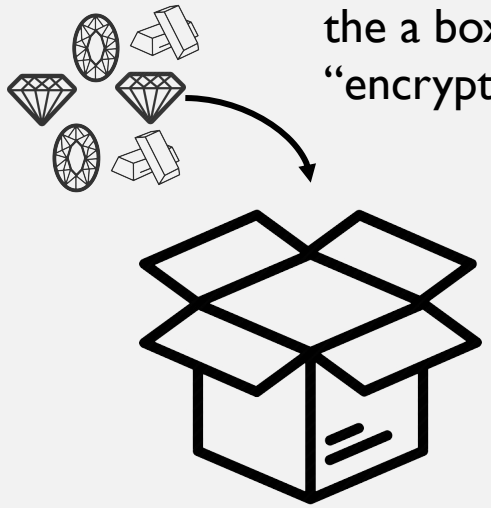


The raw material worth millions of dollars

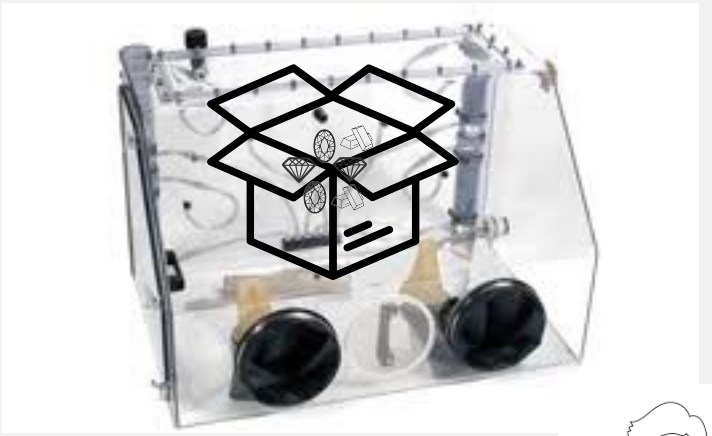
How can we make jewelry without personnel having access to the raw material?

Introducing the "FHE glove box" that allows jewel crafters to work without having direct access to the raw material

HOMOMORPHIC ENCRYPTION



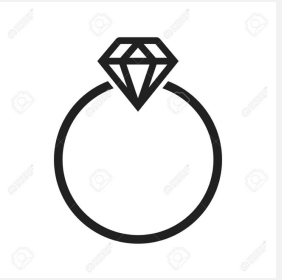
Step 1.
Put the jewels inside the a box in an "encrypted" form



Step 2. Jewelcrafter works through the glovebox using only encrypted material and makes a ring

Step 3. Alice opens the glove box using her decryption key and takes the real (decrypted) ring

Step 4. Profit



Evolution of Homomorphic Encryption
From Research to Reality

While the concept of fully homomorphic encryption (FHE) was first envisioned in the 1970s, researchers and academia have made significant progress in the last decade to develop the algorithms, use-cases, and toolkits that will help drive FHE into the business world in the near future.

2009 ---------- 2020 ---------->

Plausibility phase → **Improvement phase** → **Early adoption and usability** → **Future: Production phase**

Breakthrough
- FHE demonstrated, but too slow for practical use

Refining Algorithm
- New and improved variations of FHE

Compute-Power
- Exponential growth in industry compute performance

Use-Cases
- Demonstrating specific types of analysis using FHE (banking, genetics, etc.)

Education & Tools
- Instructions and toolkits for developers to experiment

Standards
- Industry-wide definition of standards for FHE (in progress)

Early Adoption
- SDKs and consumable services for end-users
- Enterprises start to build and test FHE prototypes apps

Ecosystem
- Pan-industry collaboration and partnerships

Real-World Applications
- Organizations begin using FHE applications

First FHE Scheme (2009)
IBM researcher demonstrates first working algorithm for FHE

GitHub Library (2012)
IBM publishes first FHE library on GitHub for community collaboration

Speed Jump: 1 Million times faster (2013)
An operation taking 1 second without encryption was shown to take 12 days with FHE (1 million times faster than previously)

Analysis of Human Genome <1 hour (2015)
IBM Researchers perform FHE analysis of human genomes in less than 1 hour

First Industry Consortium (2017)
Homomorphic Encryption Standardization group formed to develop community standards for security, API, and apps

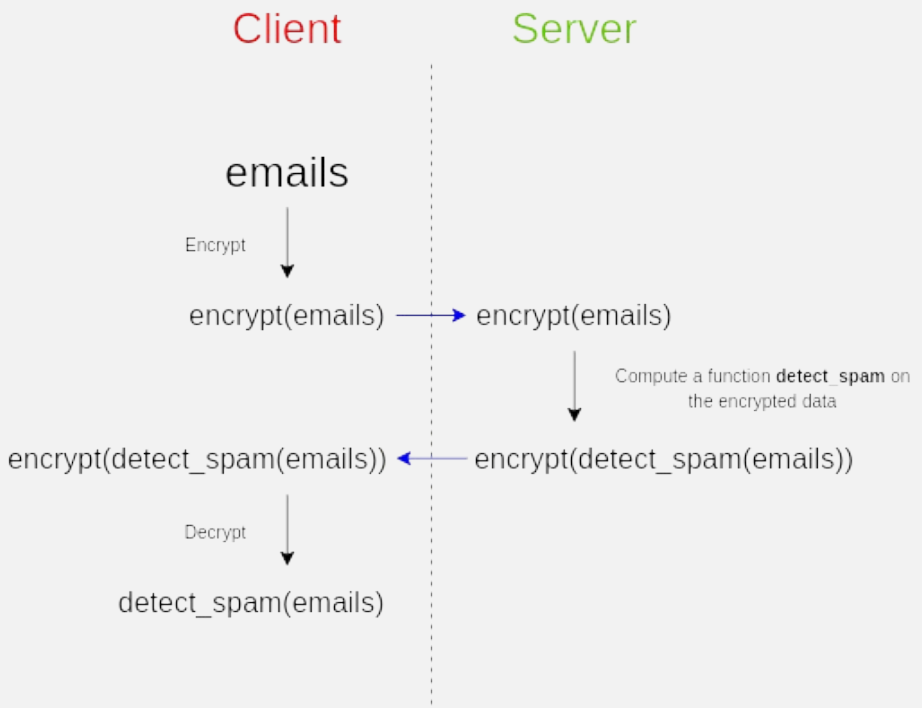
Today: IBM Homomorphic Encryption Services (2020)
Testing platform + expert support for businesses to experiment with FHE

FHE becomes Mainstream
Following successful early adoption, FHE can become a new industry standard for processing sensitive data

[IBM News Room - IBM Prepares Clients for Next Generation Encryption Technology]

HE USE CASES

- Private Search
- Computation on Cloud
- Encrypted databases



HE HAS GONE MAINSTREAM

- GitHub - data61/python-paillier: A library for Partially Homomorphic Encryption in Python
- GitHub - microsoft/SEAL: Microsoft SEAL is an easy-to-use and powerful homomorphic encryption library.
- GitHub - IBM/fhe-toolkit-linux: IBM Fully Homomorphic Encryption Toolkit For Linux
- GitHub - n1analytics/javallier: A Java library for Paillier partially homomorphic encryption.

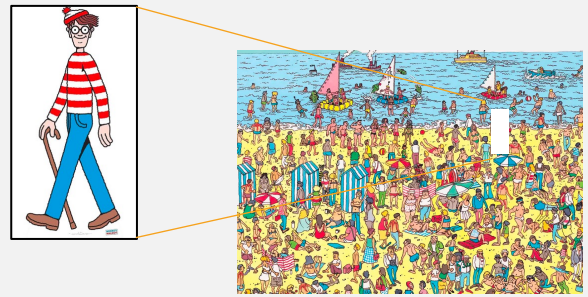
Still orders of magnitude slower than plaintext operations!
Still works for one pair of keys - additional users means additional copies of encrypted data
Secure the key!!!!

ZERO KNOWLEDGE PROOF

Alice and Bob play "Where's Waldo?". Alice claims to have found Waldo but Bob does not believe her

Bob draws a pattern at the back of the picture, more like a watermark or signature. This is to ensure that Alice will not work on another puzzle image

Alice enters an empty room with only a pair of scissors and the image



Alice crops the part of the picture that contains Waldo and exits the room while destroying the rest of the picture

Bob verifies that a) this is Waldo and b) his signature is at the back side.

APPLICATIONS OF ZKP

- **Ownership**: prove that you own a key
- **Account balance**: prove that you have enough funds without revealing your available balance
- **Membership**: prove that you are part of a group without revealing your identity
- **Sealed bid auctions**: prove who won without revealing any bid

STAY SAFE & THANK YOU

RESOURCES

- Latanya Sweeney. 2002. K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557–570.
- A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24
- Dwork C. (2011) Differential Privacy. In: van Tilborg H.C.A., Jajodia S. (eds) Encyclopedia of Cryptography and Security. Springer, Boston, MA.
- Craig Gentry. A fully homomorphic encryption scheme. https://crypto.stanford.edu/craig/craig-thesis.pdf
- Wu, H., & Wang, F. (2014). A survey of noninteractive zero knowledge proof system and its applications. The Scientific World Journal, 2014, 560484.