PostgreSQL, Planning, PostGIS, Partitioning, PaaS, Permissions and now....

Patterns and Packages in PostgreSQL for Privacy Preservation

15 November 2019, Sydney

www.2019.pgdu.org

mantaq10
Atif Rahman

I was like her
According to Pearson-R
We were both outliers

- Data Engineering
- ML Pipelines
- Herding Cats

NDB Breach Notifications April 2018 – March 2019

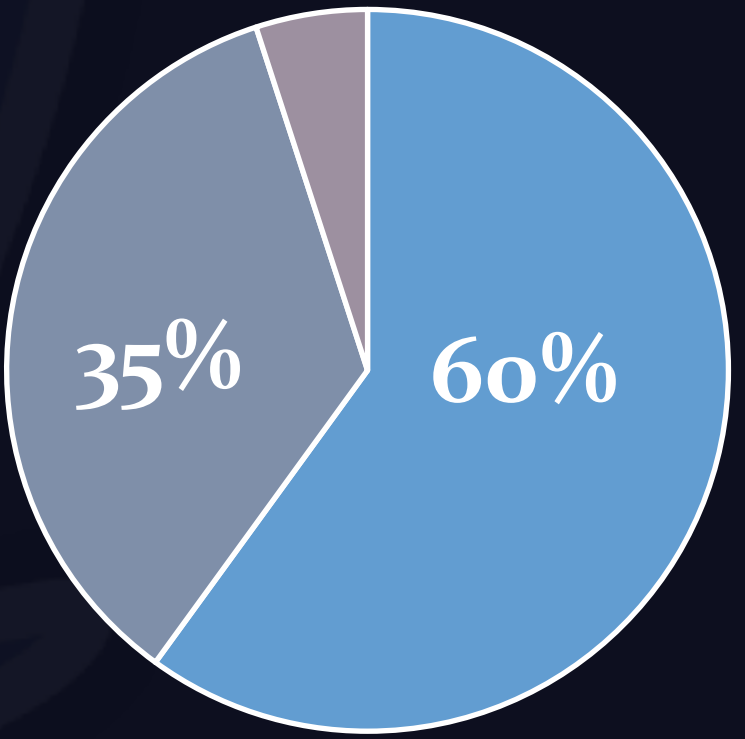
964

OAIC Report 2019

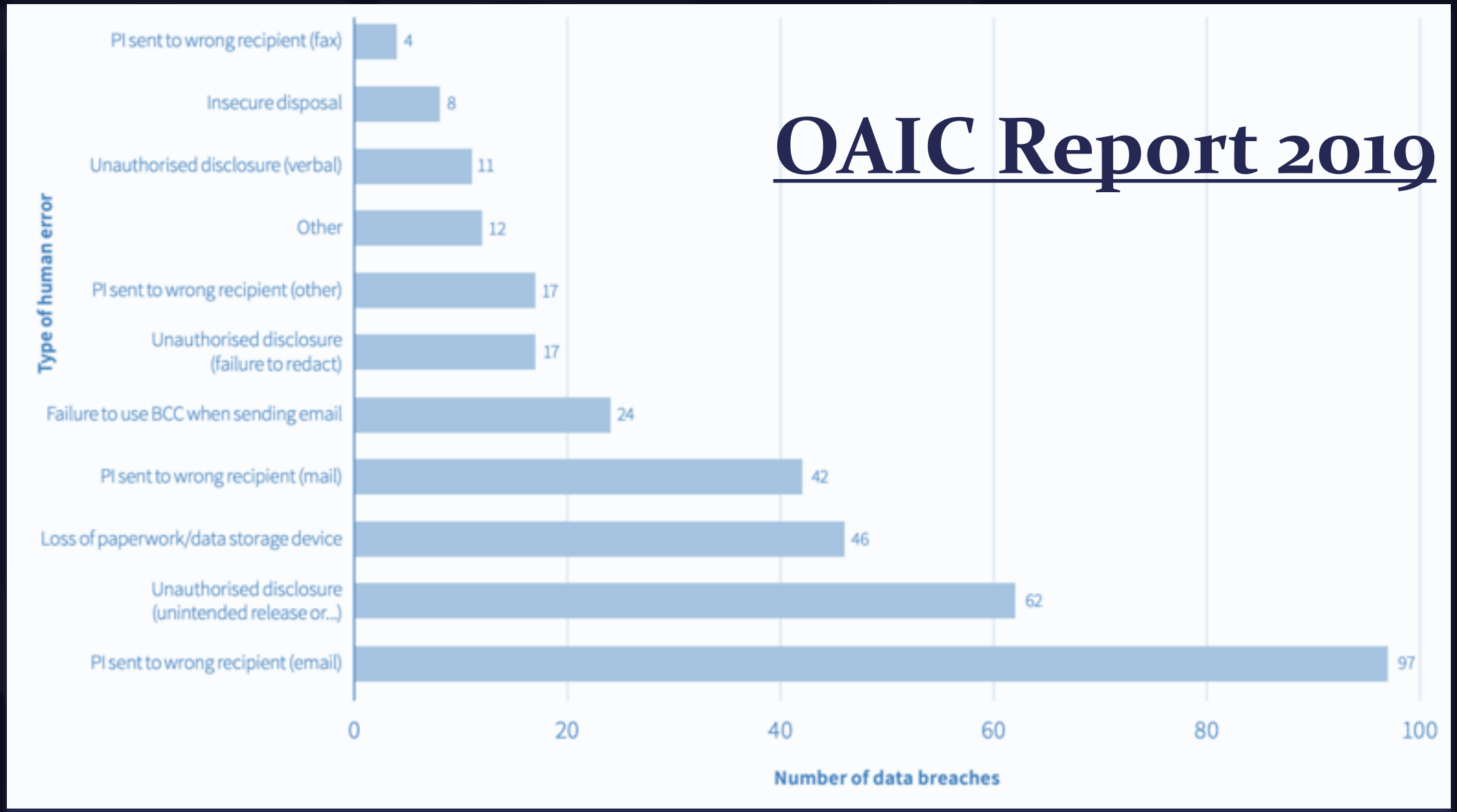
Human Error

55% Healthcare

41% Financial



OAIC Report 2019



You can have **security**
but not necessarily **privacy**

Security → Protection

Binary

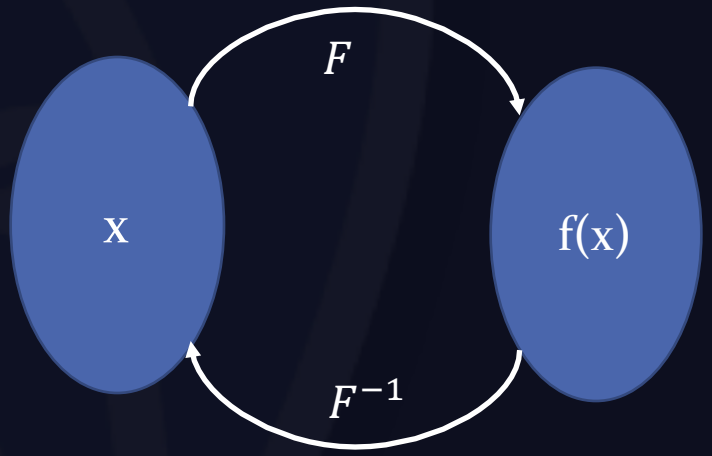
Privacy → Usage

Contextual

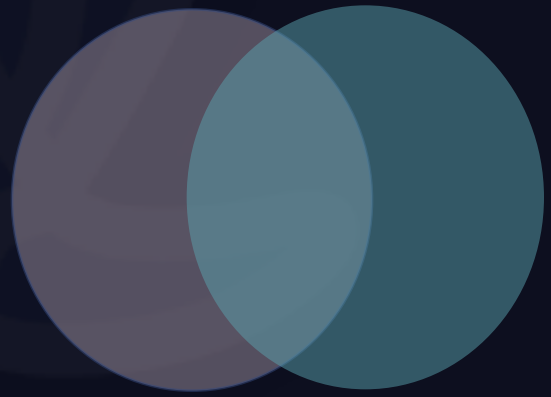
ISO/IEC 29100:2011: Privacy Framework

Privacy Guarantees

1. De-Identification (Record Keys (PK, FK, SK))
2. Re-Identification (Brute Force & Decryption)
3. Re-Identification (Record Linkage * Math)
4. Ethical Computing (Permissibility & Compliance)



Loss-less Functions
vs
Lossy Functions



PII and Attribute Augmentation

"Homomorphic encryption schemes are often repackaging vulnerabilities (practical chosen-ciphertext attacks) as features." – The Internet

Record Linkage

"87% of the U.S. population is uniquely identified by date of birth, gender, postal code." *Latanya Sweeney (k-anonymity)*

"Decreasing the precision of the data, or perturbing it statistically, makes re-identification gradually harder at a substantial cost to utility". *Chris Culnane, Benjamin Rubinstein, Vanessa Teague @UniMelb*

Privacy vs Utility Trade-off

Bleeding Edge
Cutting Edge
Established

DP SM HE AN

Privacy Guarantee

Better Utility

SM: Secure Multiparty Computing DP: Differential Privacy HE: Homomorphic Encryption AN: Anonymisation

1. AN: (Pseudo)Anonymisation

- REPLACEMENT
- SUPRESSION
- PERTURBATION
- GENERALISATION

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	SARAH CONNOR	12-06-1962	JB Vet	63456	12
112	PAMELA LANDY	18-10-1971	FBI	54367	45

(reversible or random)

REPLACEMENT

(PG String Functions)
(PGAnonymizer)

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	MIKE OBAMA	13-07-1982	JB Vet	63456	12
112	BRUCE LEE	19-11-1991	FBI	54367	45

1. AN: (Pseudo) Anonymisation

- REPLACEMENT
- SUPRESSION
- PERTURBATION
- GENERALISATION

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	SARAH CONNOR	12-06-1962	JB Vet	63456	12
112	PAMELA LANDY	18-10-1971	FBI	54367	45

SUPRESSION

(Wildcard or Removal)
- 18 PII Attributes

(PG String Functions)
(PGAnonymizer)

ID	NAME	EMPLOYER	ZIPCODE	FK_SHOP
101	M*** ****A	JB Vet	63456	12
112	B**** **E	FBI	54367	45

1. AN: (Pseudo) Anonymisation

- REPLACEMENT
- SUPRESSION
- PERTURBATION
- GENERALISATION

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	SARAH CONNOR	12-06-1962	JB Vet	63456	12
112	PAMELA LANDY	18-10-1971	FBI	54367	45

(Additive Noise)
(PDF)
(Data Imputation)

PERTURBATION

(PGAnonymizer)
(Google DP)
(Uber DP)

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	SARAH CONNOR	12-07-1958	JB Vet	64532	12
112	PAMELA LANDY	18-11-1973	FBI	57843	45

1. AN: (Pseudo)Anonymisation

- REPLACEMENT
- SUPRESSION
- PERTURBATION
- GENERALISATION

ID	NAME	DOB	EMPLOYER	ZIPCODE	FK_SHOP
101	SARAH CONNOR	12-06-1962	JB Vet	63456	12
112	PAMELA LANDY	18-10-1971	FBI	54367	45

(K-Anonymity or Masking)

GENERALISATION

(PGAnonymizer)
(PG Aggregate Functions)

ID	NAME	DOB	EMPLOYER	σ_ZIPCODE	FK_SHOP
101	SARAH CONNOR	1960s	JB Vet	0.37	12
112	PAMELA LANDY	1970s	FBI	-0.99	45

Privacy vs Utility Trade-off

Bleeding Edge
Cutting Edge
Established

DP SM HE AN

Privacy Guarantee

Better Utility

SM: Secure Multiparty Computing DP: Differential Privacy HE: Homomorphic Encryption AN: Anonymisation

Differential Privacy

Database with Ned in it

Statistical Properties

The Oracle

Perturbations (Noise)

Private Database.

Not sure if Ned is there anymore

- Works on the Data itself, not on the management environment
- Considerably fast compared to encryption techniques.
- Quantum Safe (ish)

Differential Privacy on PostgreSQL

https://github.com/google/differential-privacy

Count
Sum
Mean
Variance
Standard deviation
Order statistics (including min, max, and median)

Laplace Functions for UDFs

Privacy Loss

- Epsilon & Delta
- Risk Score for every attribute used for a particular person
- Risk Score for total number of records with similar values
- (rule of thumb) k = 11

HE: Homomorphic Encryption

Ability to apply computations on encrypted data!

- Malleable
- Performance
- Operators

- Partial HE
- Full HE

- BFV
- BGV
- CKKS

Microsoft SEAL
PALISADE
HELib
HEAAN
TFHE

Trade-Offs | Categories | Schemes | Libraries

Privacy vs Utility Trade-off

Bleeding Edge
Cutting Edge
Established

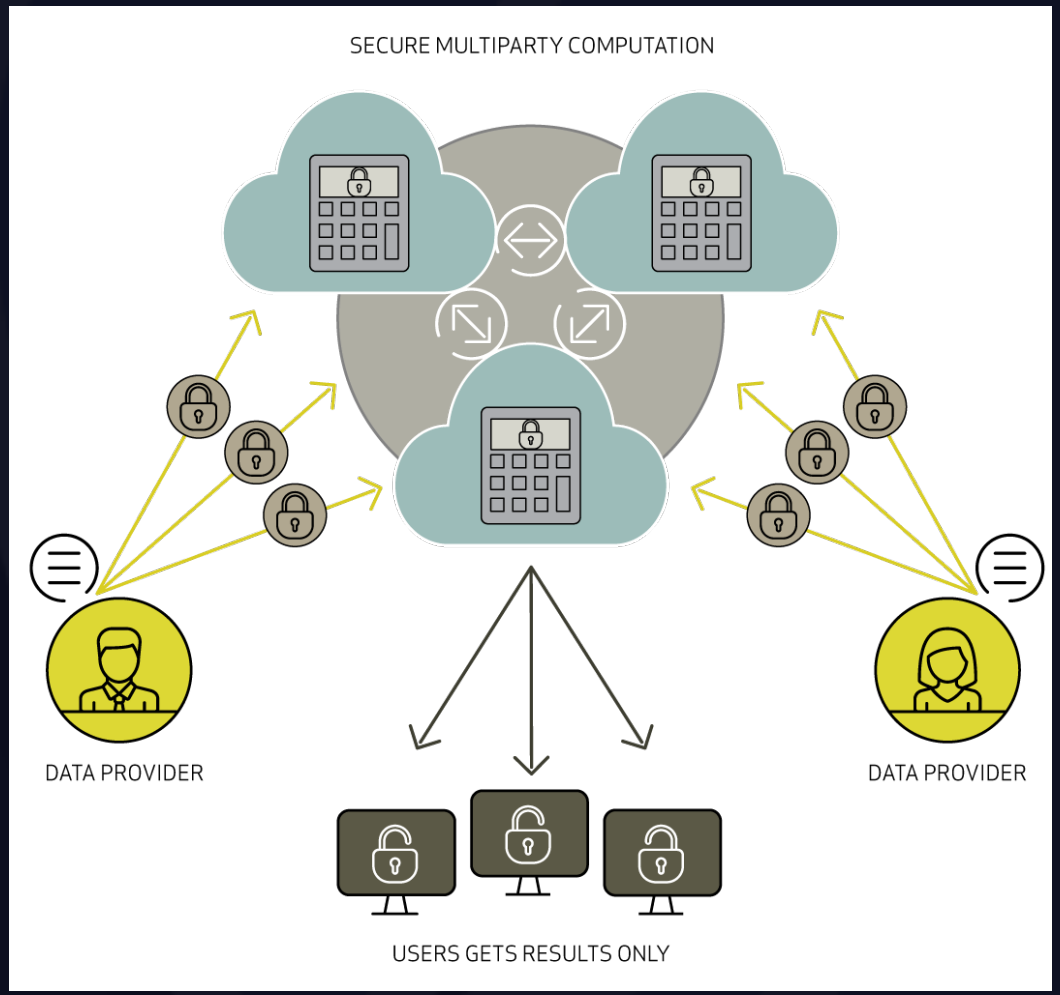
DP SM HE AN

Privacy Guarantee

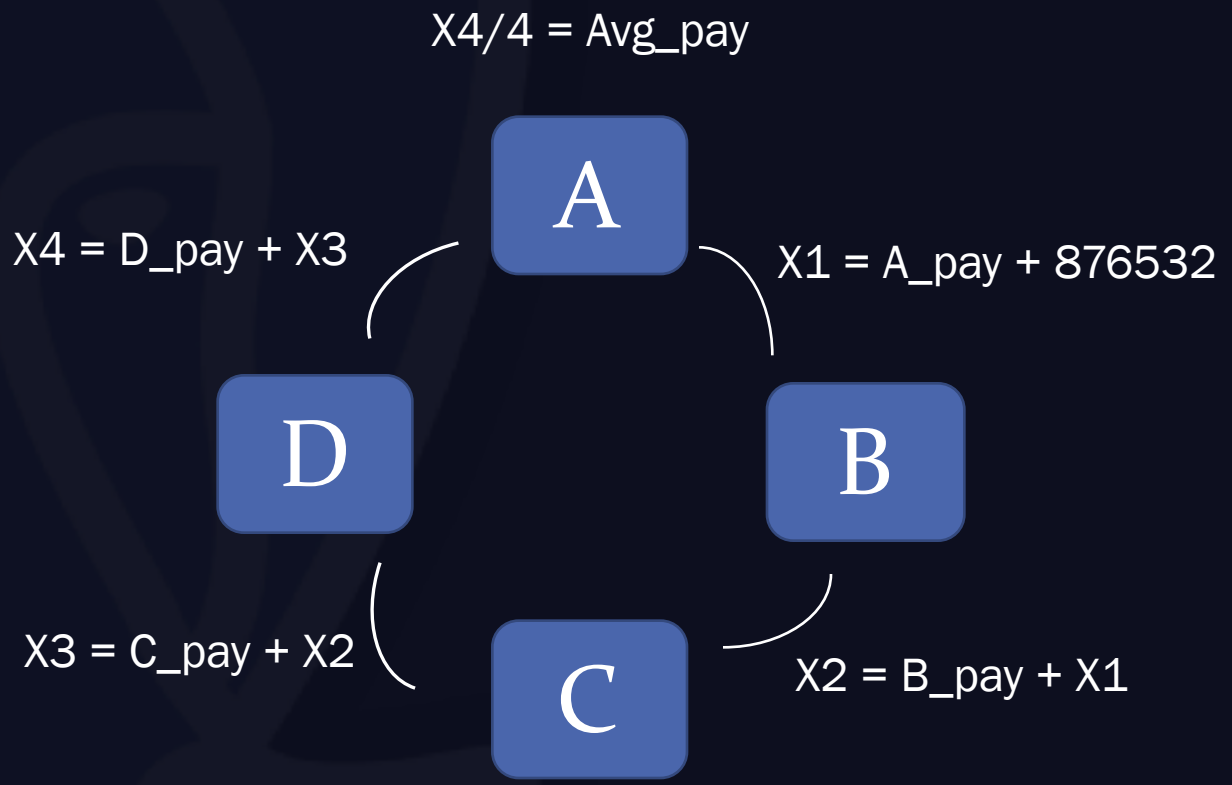
Better Utility

SM: Secure Multiparty Computing DP: Differential Privacy HE: Homomorphic Encryption AN: Anonymisation

Secure Multi-party Computing

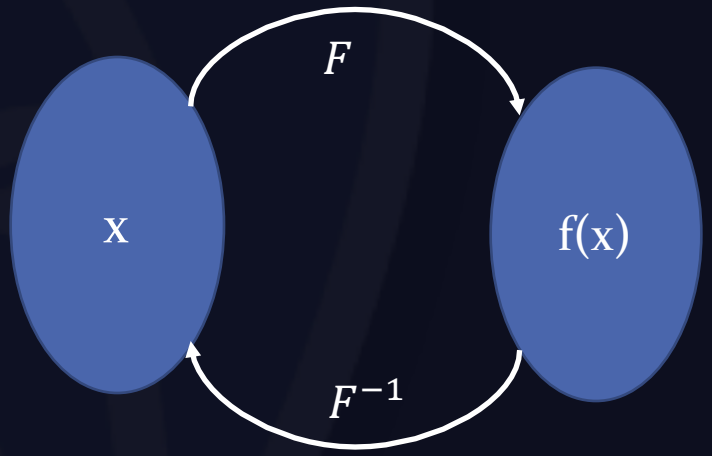


K-Anonymity

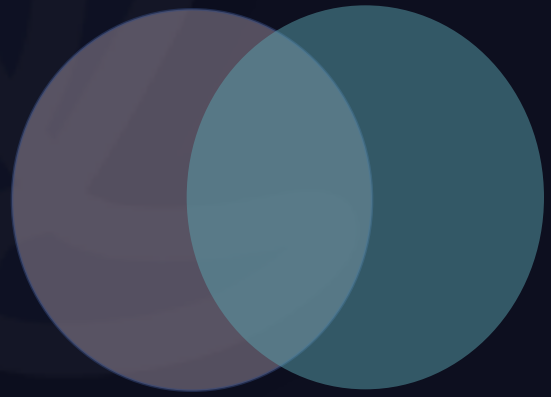


Privacy Guarantees

1. De-Identification (Record Keys (PK, FK, SK))
2. Re-Identification (Brute Force & Decryption)
3. Re-Identification (Record Linkage & Math)
4. Ethical Computing (Permissibility & Compliance)



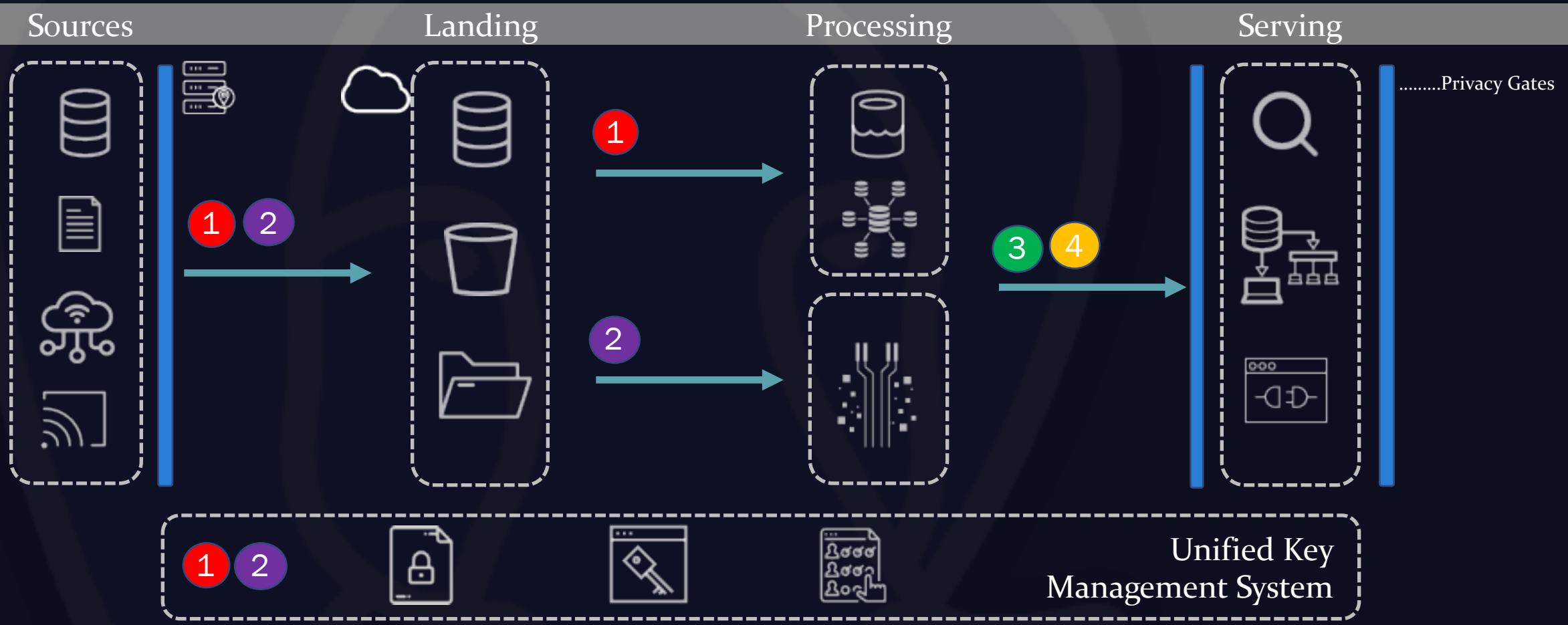
Loss-less Functions
vs
Lossy Functions



PII and Attribute Augmentation

"Homomorphic encryption schemes are often repackaging vulnerabilities (practical chosen-ciphertext attacks) as features." – The Internet

Typical Data Pipelines



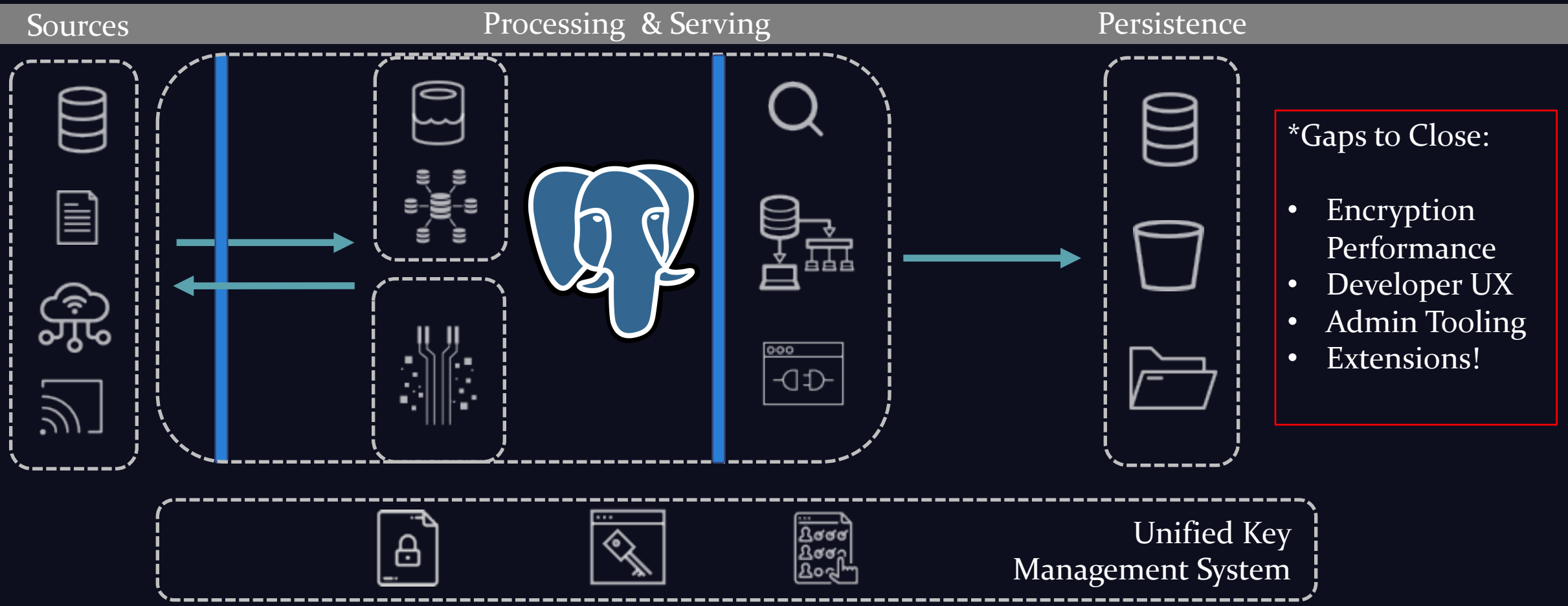
De-Identification (Record Keys (PK, FK, SK))

Re-Identification (Brute Force & Decryption)

Re-Identification (Record Linkage)

Ethical Computing (Permissibility & Compliance)

Emerging Data Architecture (Data Fabrics) [HTAP = OLTP + OLAP]



Key Takeaways

Securing your database doesn't guarantee data privacy.

There are trade-offs between privacy and utility

You can provision privacy controls within PostgreSQL

PostgreSQL fits emerging (data) architecture patterns

Atif is pledging to build an extension, he needs my help!

Questions