privacy dynamics

Privacy Plus Utility

Preserving Data Insights with State-of-the-Art Privacy Protection

WILL THOMPSON, DIRECTOR OF ENGINEERING

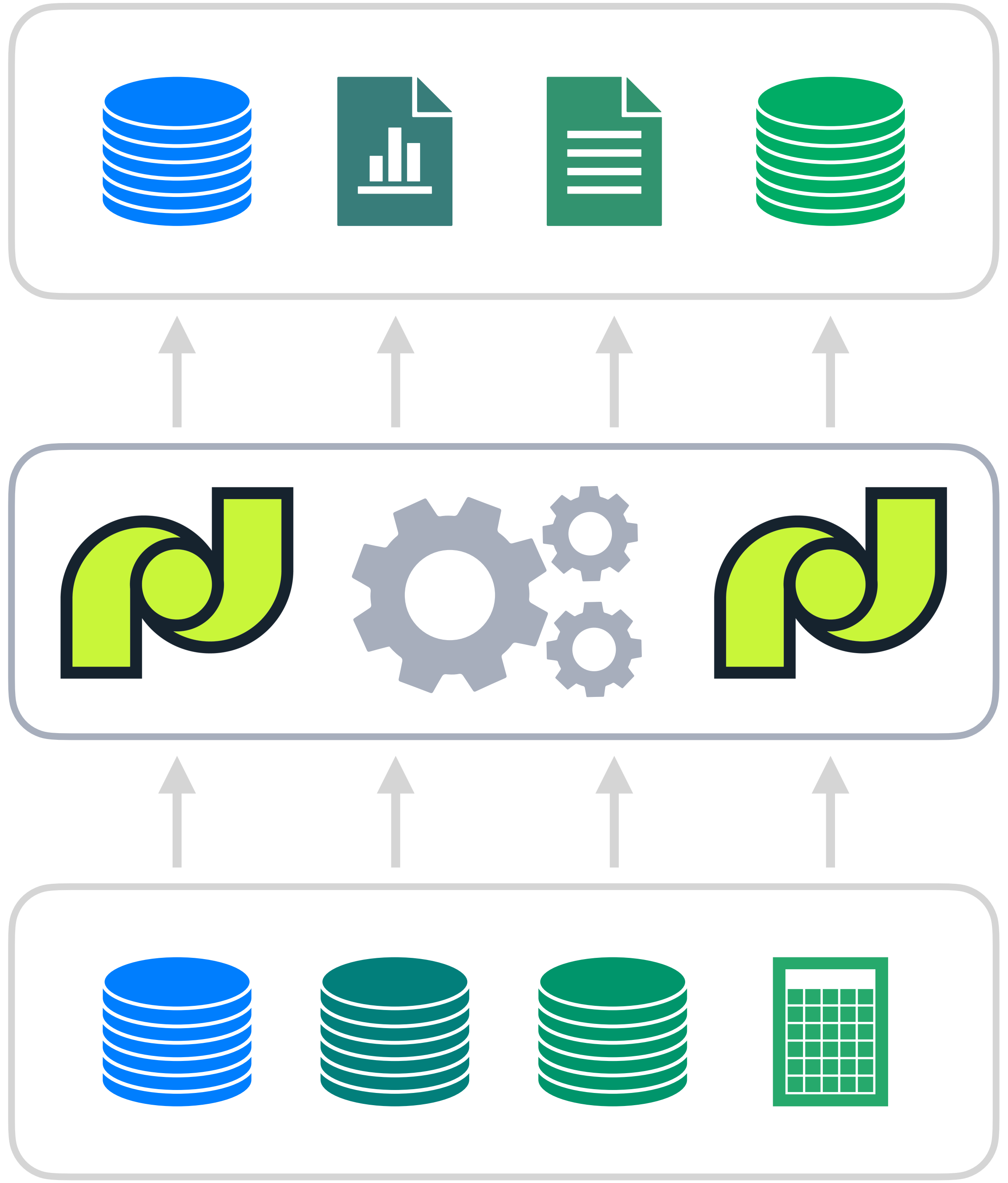
MARCH 24, 2022

Product Goals

🎚️ Privacy switch for the modern data stack

📤 Dataset sharing

- Data scientist/engineer-focused workflows
- Varying degrees of trust between 3rd parties
- Analysts want to use their own analytics tooling



What is data privacy?

This talk

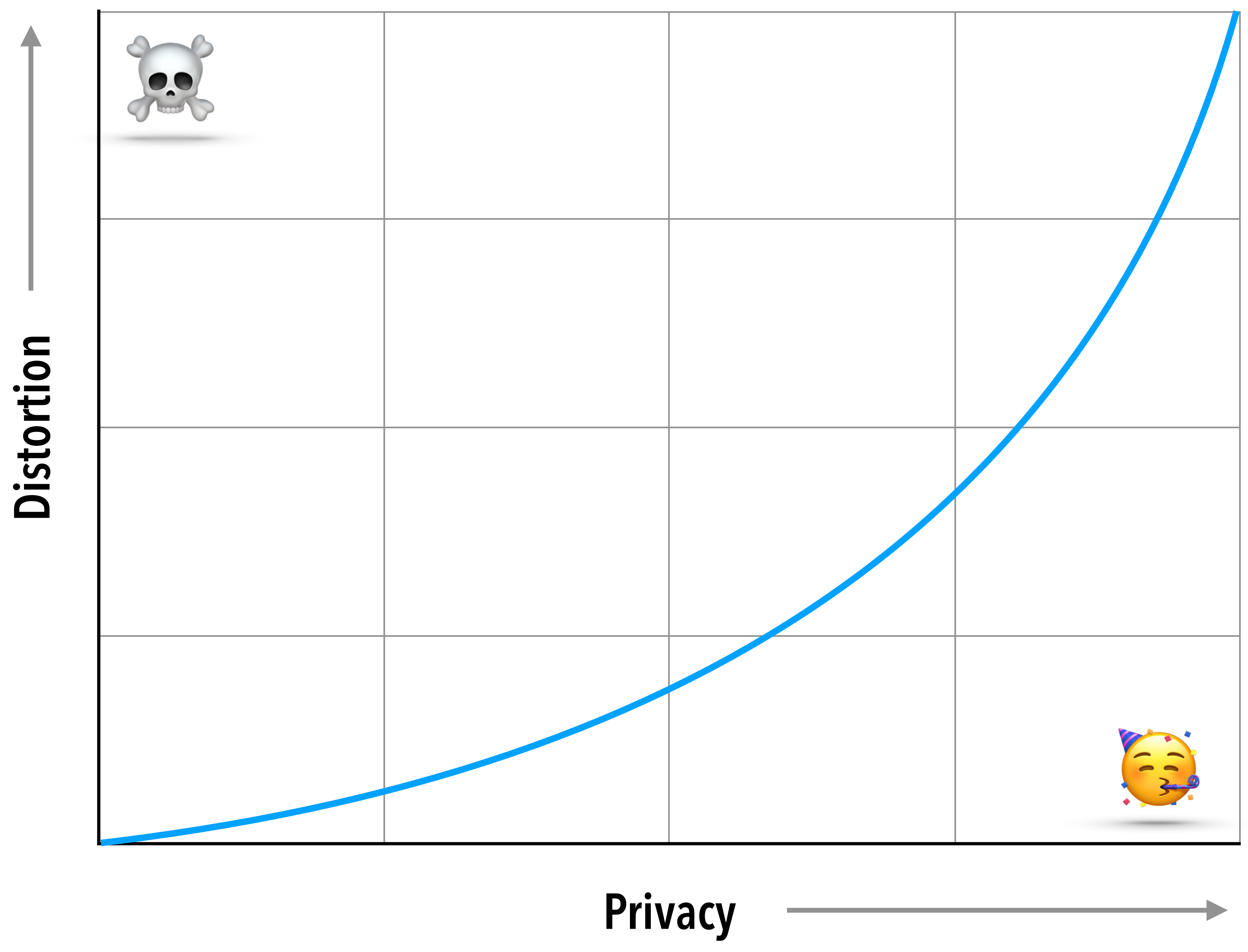
Data release

Protecting identities of individuals represented in data, i.e. *not* data security or governance.

Concepts

- **Pseudonymization**
  Remove or replace direct identifiers (DIDs), e.g. name, address, phone number
- **Re-identification**
  Use indirect/quasi-identifier (QIDs) - e.g. age, zipcode, gender -or personal attributes to match an individual in an external dataset or learn new info using inference attacks.
- **Anonymization (de-identification)**
  Change QIDs or personal attribute values to mitigate risk.

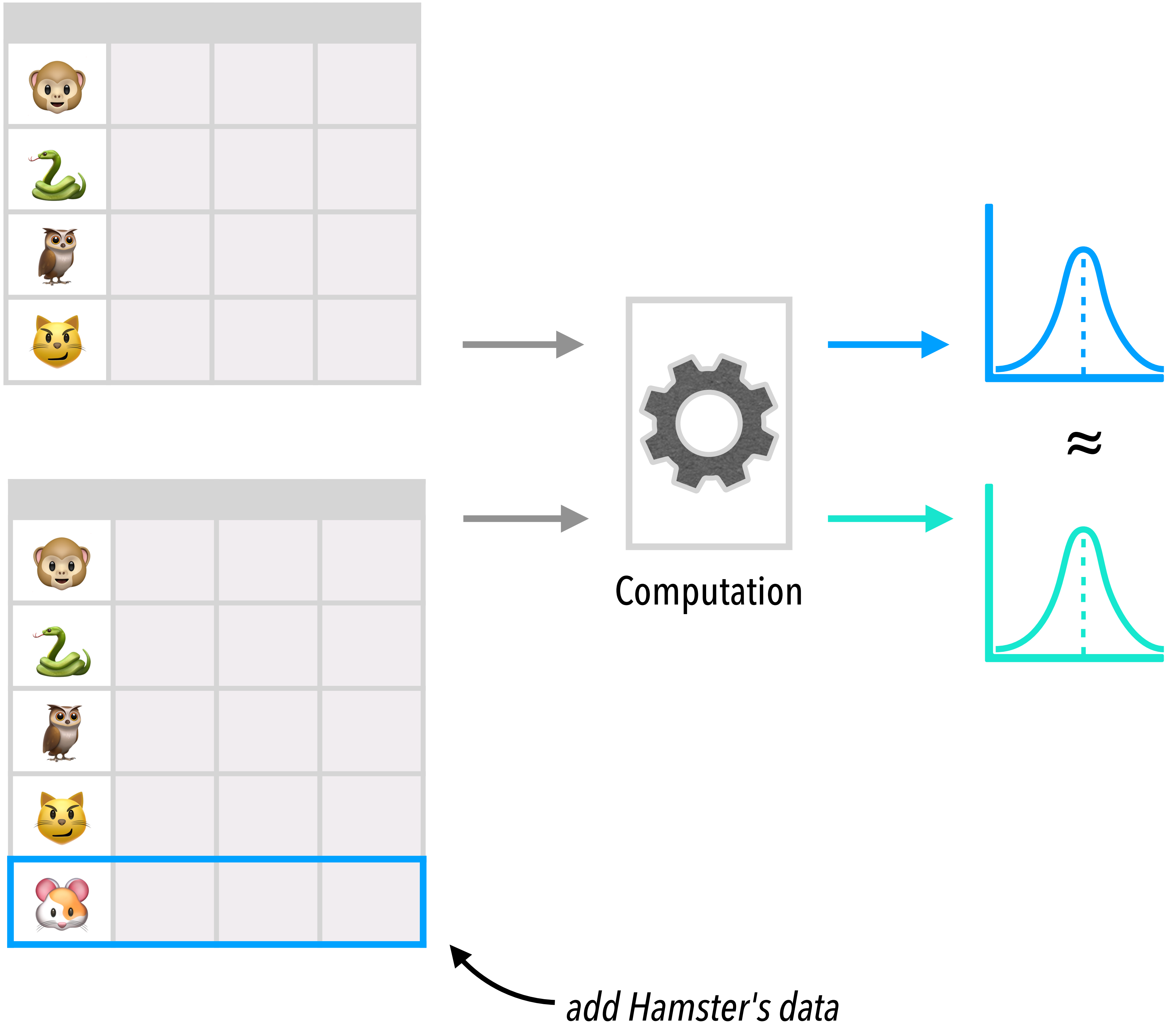
Privacy vs Utility



METHODS EXPLORED

Global Differential Privacy

- 😎 Indistinguishability of computation output when input differs by one individual's data
- ≈ Differentially private output is roughly the same, with or without Hamster's data
- 📶 ε (epsilon) measures "how roughly"
- 🤏 Smaller ε is more private



METHODS EXPLORED

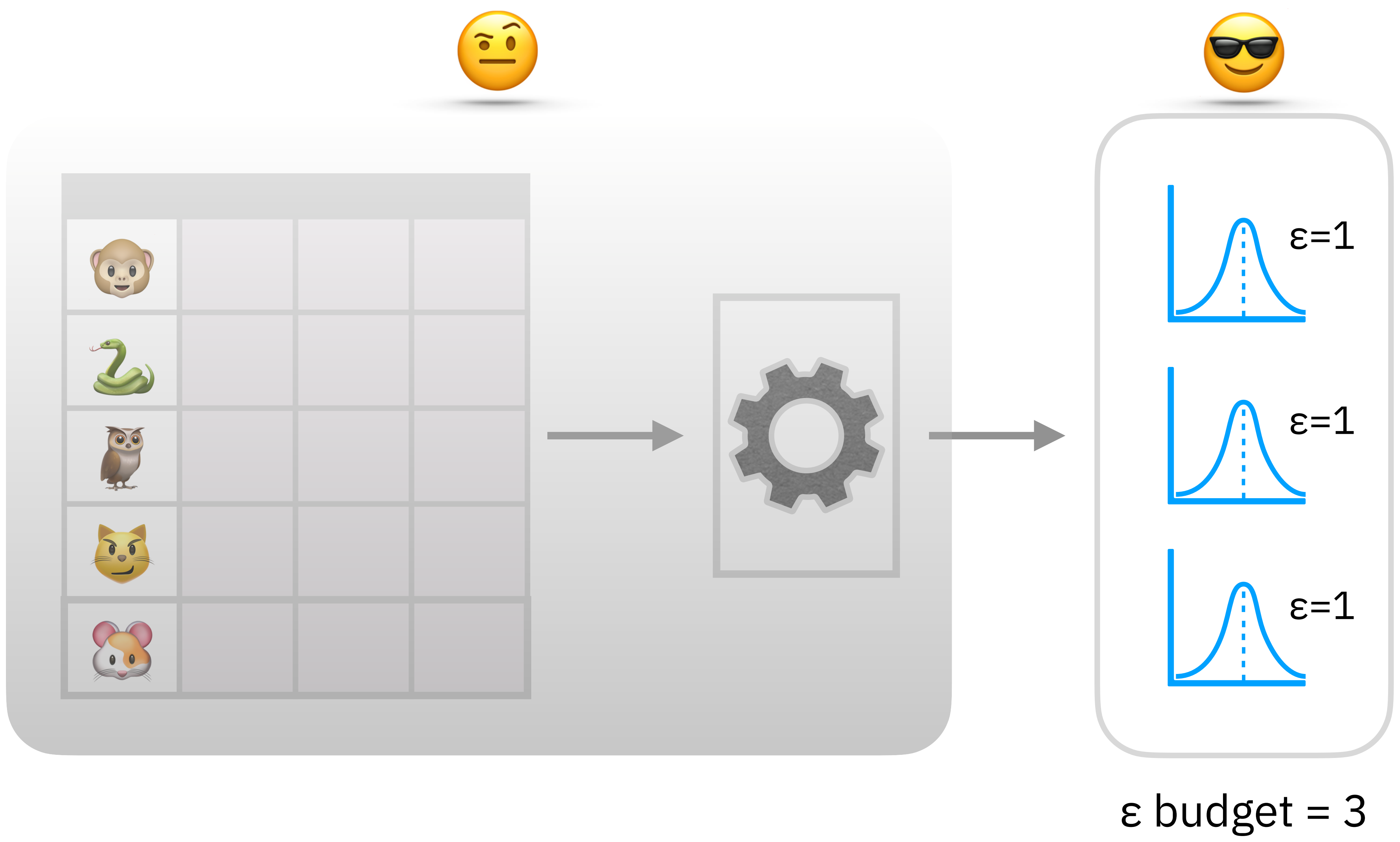
Global Differential Privacy

🎯 Only adds noise to a single statistic	High utility
🛡️ Strong guarantee on total information loss	ε is an upper-bound / worst-case
🧩 Composable	ε is cumulative across multiple releases.
🔮 Makes no assumptions about attacker	Attacker's motives or background knowledge don't affect privacy guarantee

METHODS EXPLORED

Global Differential Privacy

- 🧑‍💻 Analysts use centralized DP system
- 🔐 Centralized DP system requires trust
- Protects statistics, not datasets
- ✋ Bounded ε: each query contributes to "privacy budget"

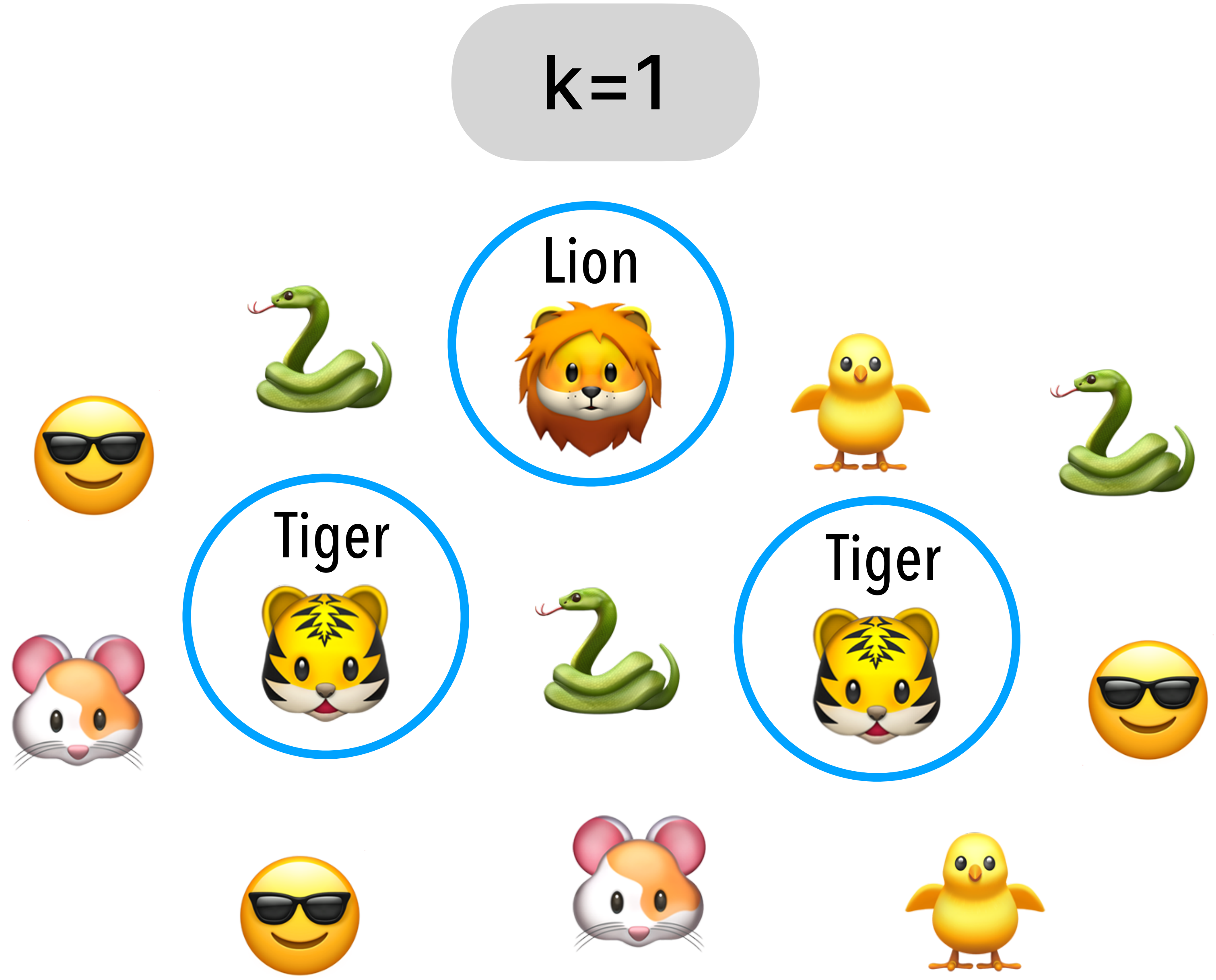


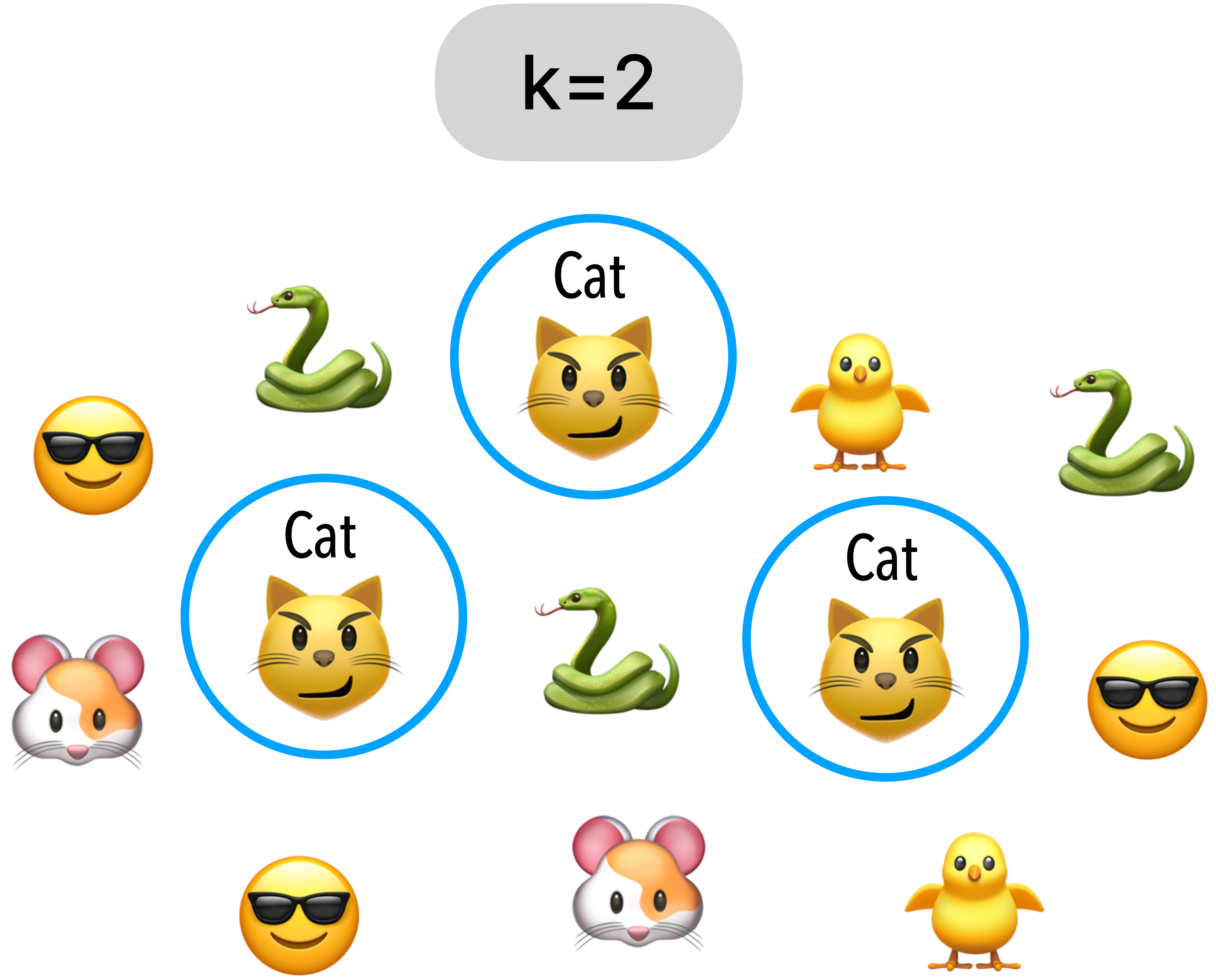
METHODS EXPLORED

K-Anonymity

What is k?

Each record's quasi-identifiers match at least k-1 other records





METHODS EXPLORED

K-Anonymity

k=1

Age	Zipcode	Sex	Hispanic	Condition
39	78745	male	no	seizure
39	78745	male	no	wheezing
37	78704	male	yes	obesity
38	78745	male	no	C.H.F.
37	78704	male	yes	chest pain
37	78745	female	yes	fever
37	78745	female	yes	fever
38	78745	female	yes	newborn
38	78745	female	yes	vomiting
37	78701	female	no	hypertension
38	78701	male	no	pneumonia
38	78701	male	no	fever

k=2

Age	Zipcode	Sex	Hispanic	Condition
30-39	78745	male	no	seizure
30-39	78745	male	no	wheezing
37	78704	male	yes	obesity
30-39	78745	male	no	C.H.F.
37	78704	male	yes	chest pain
37	787**	female	*	fever
37	787**	female	*	fever
38	78745	female	yes	newborn
38	78745	female	yes	vomiting
37	787**	female	*	hypertension
38	78701	male	no	pneumonia
38	78701	male	no	fever

METHODS EXPLORED

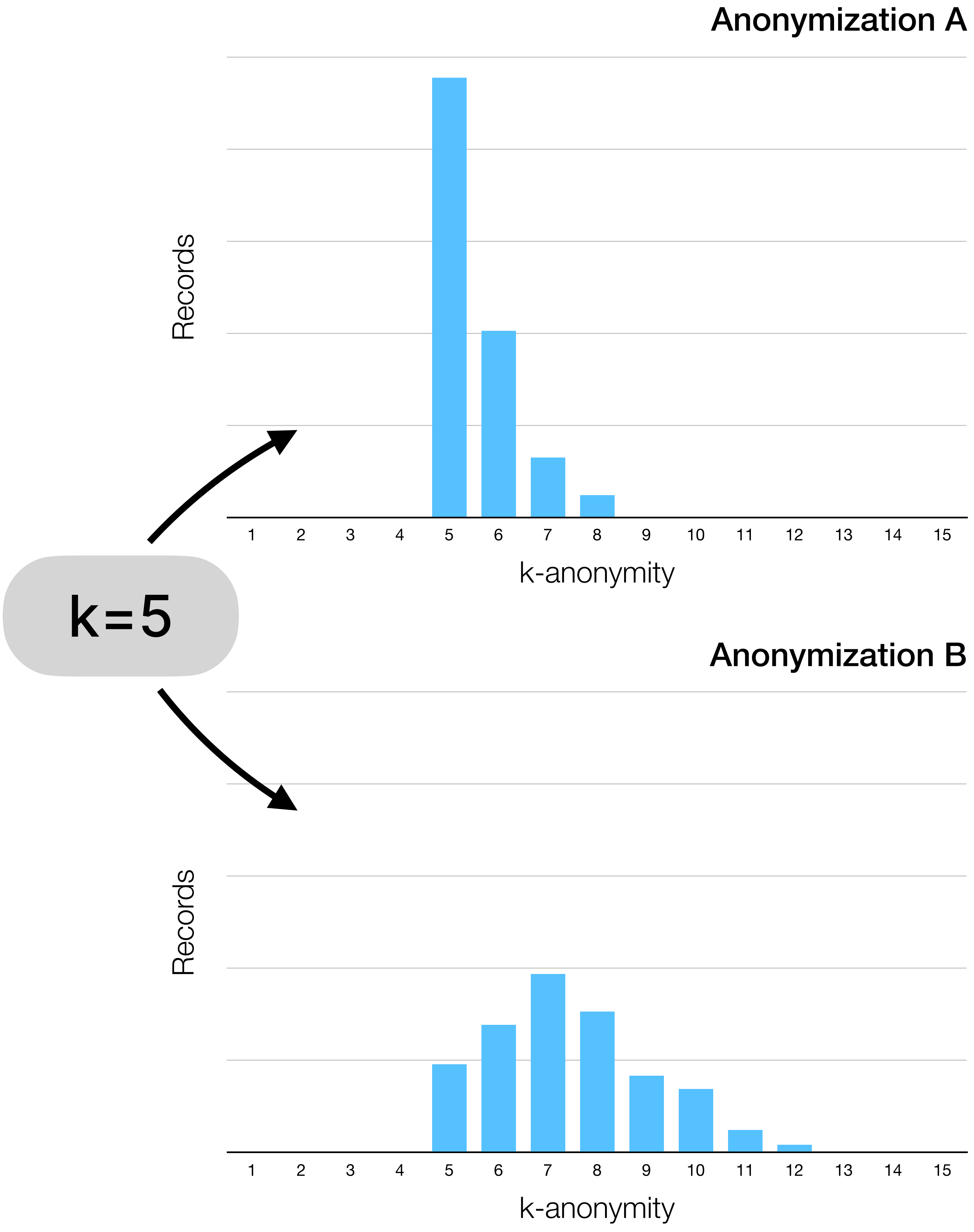
K-Anonymity

📋 Protects whole datasets	Data can easily be shared
🕵️ Directly addresses re-identification / linking attacks	Individuals "blend" with other individuals, providing plausible deniability
🎯 Only generalize/suppress values needed to achieve k-target	Minimizes information loss, good utility

METHODS EXPLORED

K-Anonymity

- K-anonymity is only a threshold metric
- 🎲 Precise re-identification risk is more complex
  - Depends on an attack model
  - Probabilistic
- ≇ Not composable
- ⌛ Computationally expensive optimization algorithms



METHODS EXPLORED

Local Differential Privacy

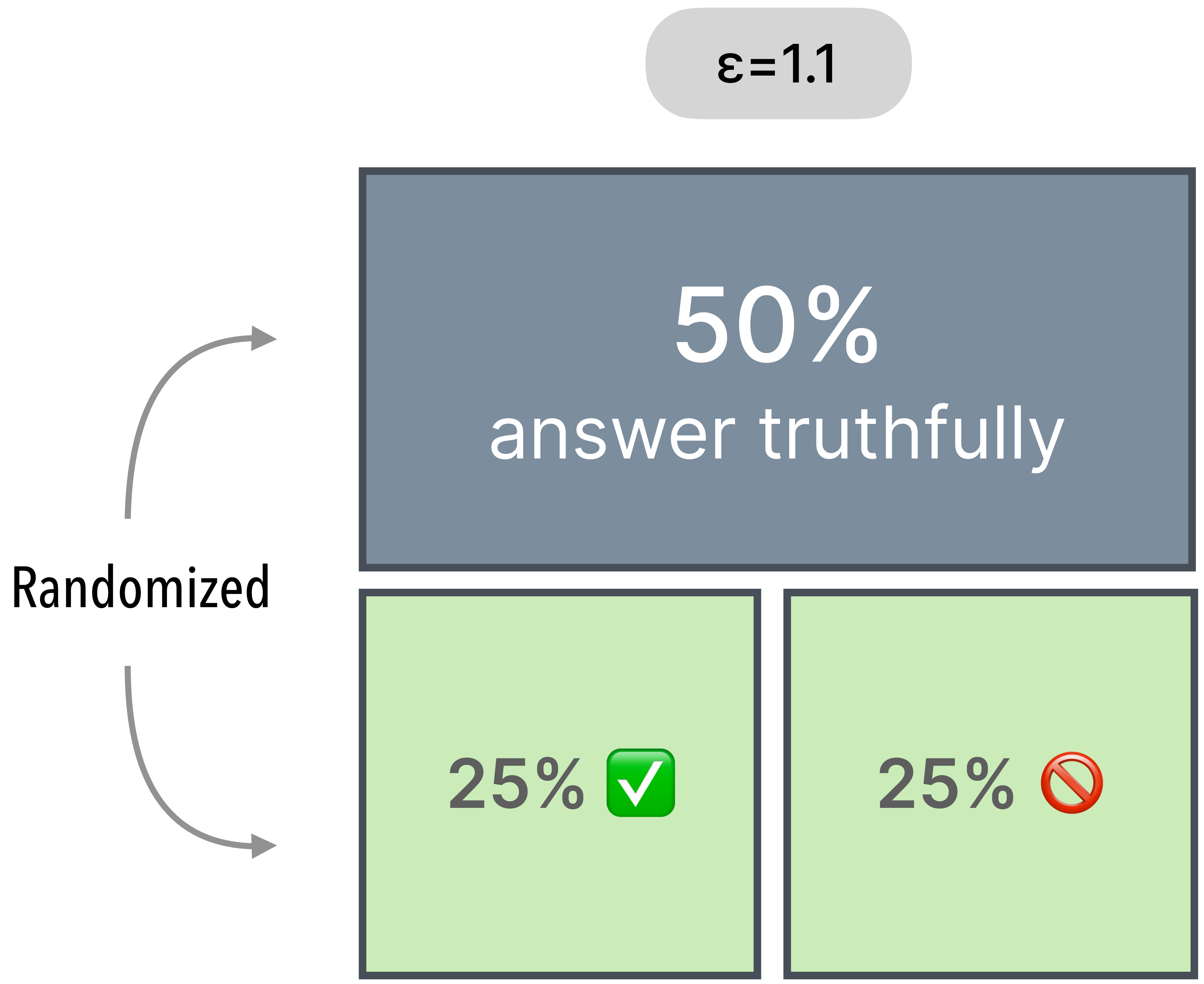
- 🔀 Randomized response
- 🗒️ Survey interview anonymity
- 😎 $F(x)$ is ε-differentially private if

$$\frac{P[F(x) = S]}{P[F(x') = S]} \leq e^{\epsilon}$$

$$\frac{P[\text{Correct answer}]}{P[\text{Incorrect answer}]} \leq e^{\epsilon}$$

$$\frac{0.75}{0.25} \leq e^{\epsilon}$$

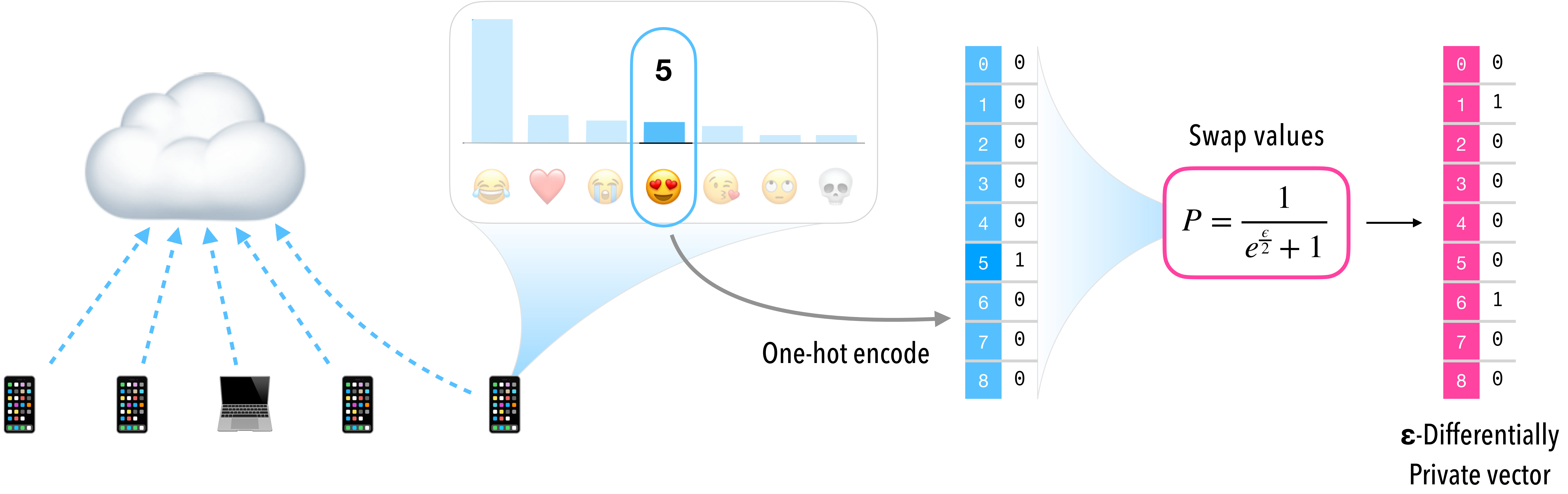
$$3 \leq e^{\epsilon}$$



METHODS EXPLORED

Local Differential Privacy

🍎 Apple emoji histograms



METHODS EXPLORED

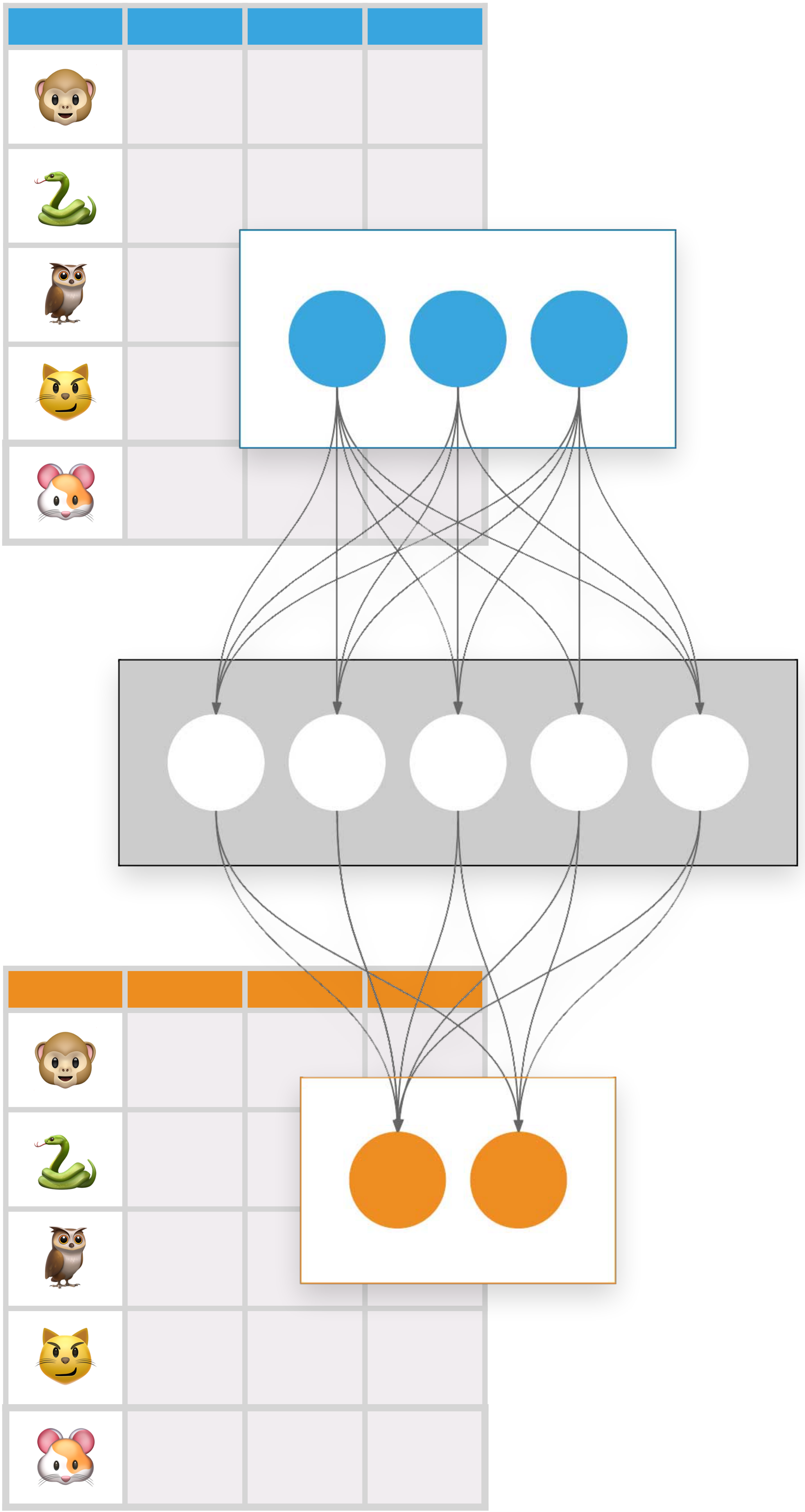
Local Differential Privacy

- 📋 Protects whole datasets (like k-anonymity)
- 🛡️ Strong privacy guarantees (like global DP)
- 🧩 Composable ε (like global DP)
- ⚙️ Hard to reconcile ε with re-id risk
  - Re-id risk models not well established
  - Re-id risk may be small, even with large ε
- 😥 Typically orders of magnitude more utility loss vs global DP

METHODS EXPLORED

Synthetic data

- 🧠 Learning model is trained on unprotected data
- 📈 Model captures statistical properties of original data
- 👶 Model produces new dataset that "behaves like" original



METHODS EXPLORED

Synthetic data

⚠️ Synthetic does not equate to private	• Models can be attacked • Synthesized data can be attacked • *Noise still needed to protect synthetic data*
👀 Privacy-utility tradeoff doesn't outperform other methods	Privacy gain / utility loss is hard to predict
⌛ Model training phase is computationally expensive	Impractical for large or highly dynamic data
🤞 Re-id risk assessment models are promising	Potential for increased utility when addressing re-id risk

PRIVACY DYNAMICS

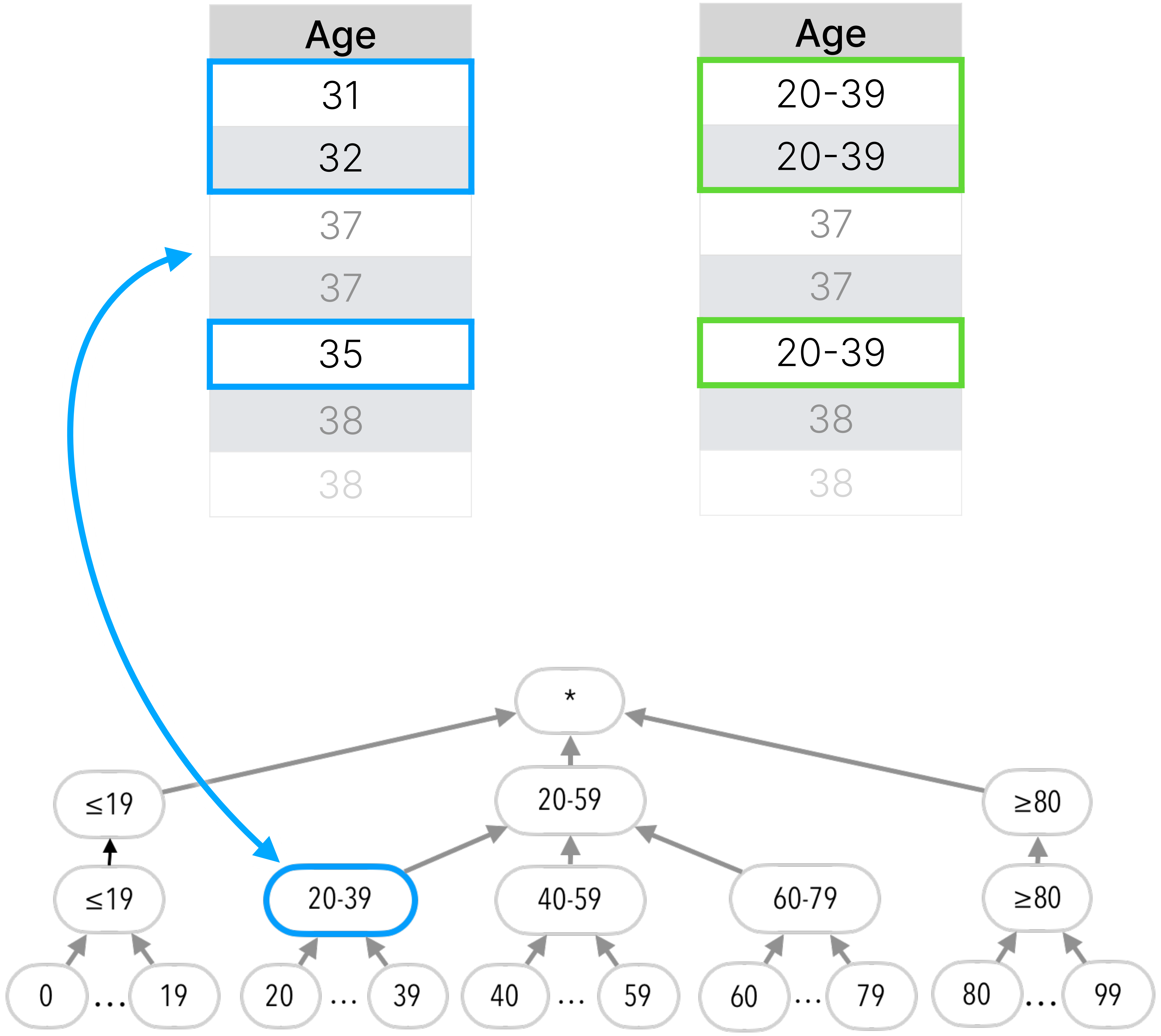
Elimination Round

	💜	💔
Global differential privacy	High utility, Strong privacy	Interactive model
K-anonymity	Good utility, Reasonable privacy	Expensive compute, mixed types
Local differential privacy	Strong privacy	Low utility, Hard to quantify re-id risk
Synthetic data	TBD	Even more expensive compute

Microaggregation

🏛️ Classical K-Anonymity

- Optimizes for predefined generalization hierarchy
- Constraints of hierarchy limit precision
- Generalization results in mixed type data
  - Numeric values mixed with category values
  - Categories mixed with other categories



Age generalization hierarchy

METHODS EXPLORED

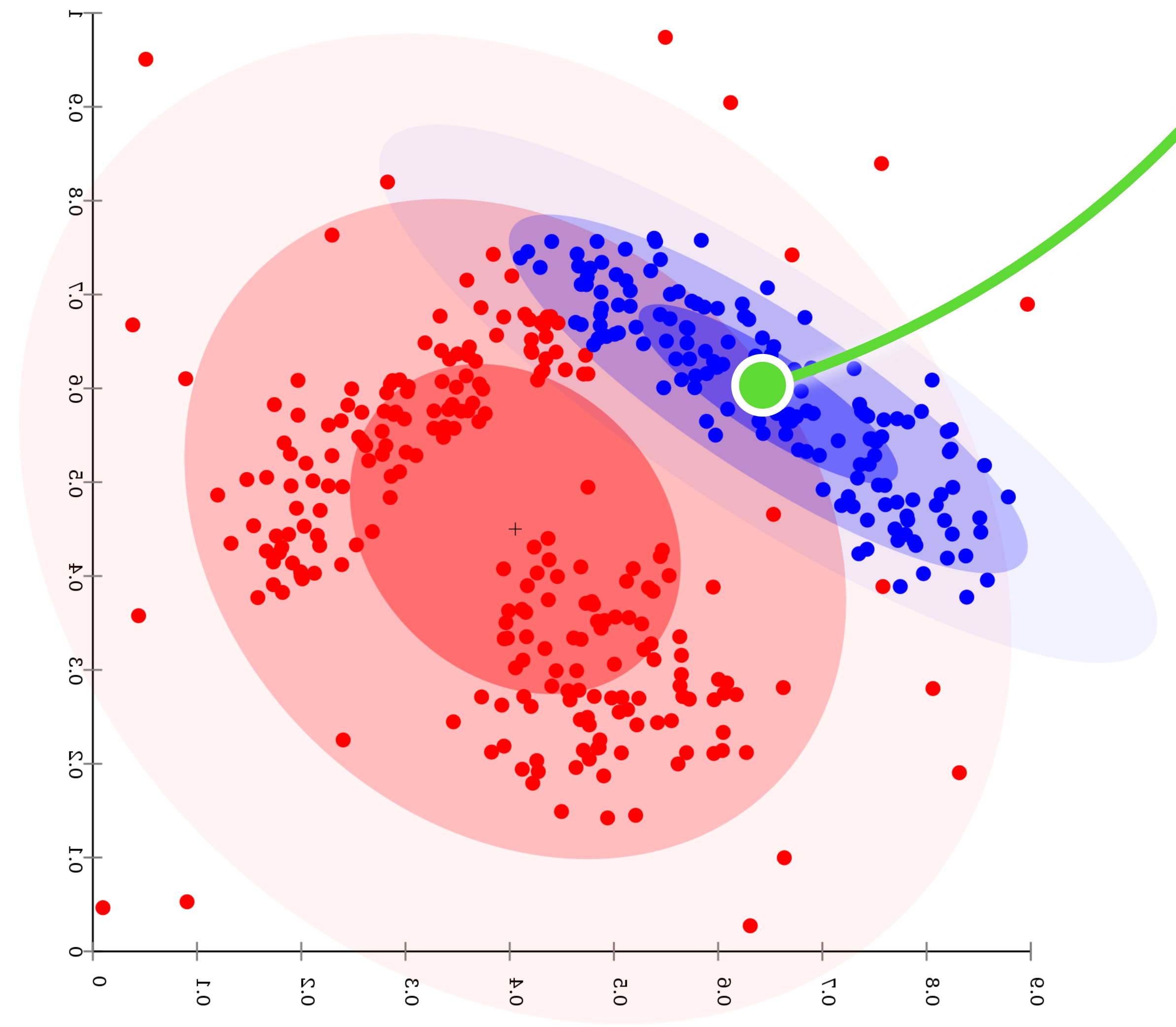
Microaggregation

👾 Microaggregation

- Compute k-sized similar clusters
- Hierarchy-free generalization can publish "cluster center"

Age
31
32
37
37
35
38
38

Age
31-35
31-35
37
37
31-35
38
38



METHODS EXPLORED

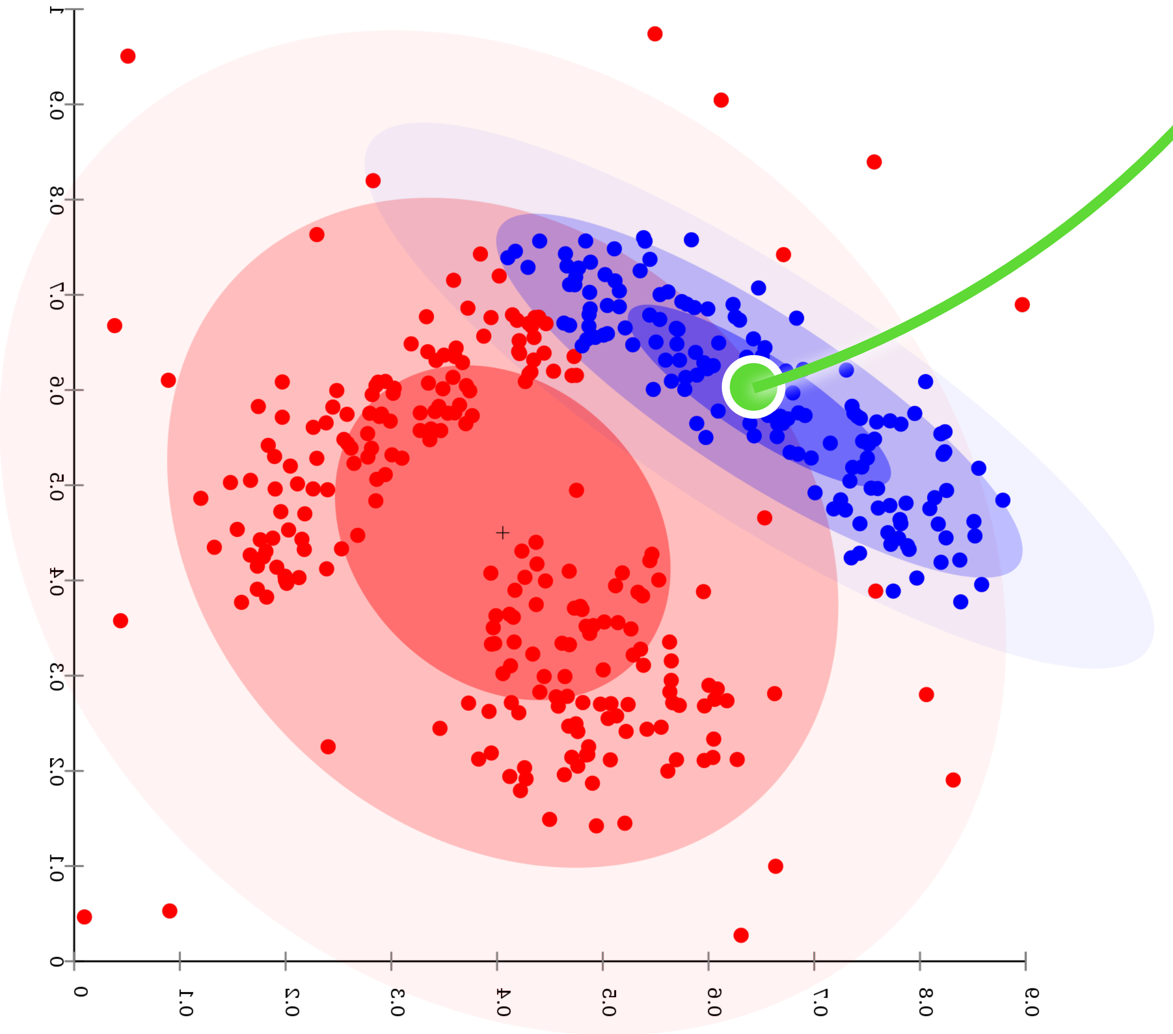
Microaggregation

🔀 Perturbation: data can change

- Maintain data semantics for downstream analysis
- More precisely target cluster center with median/mode
- Target geometric/geographical center
- Avoid suppression

Age
31
32
37
37
35
38
38

Age
32
32
37
37
32
38
38

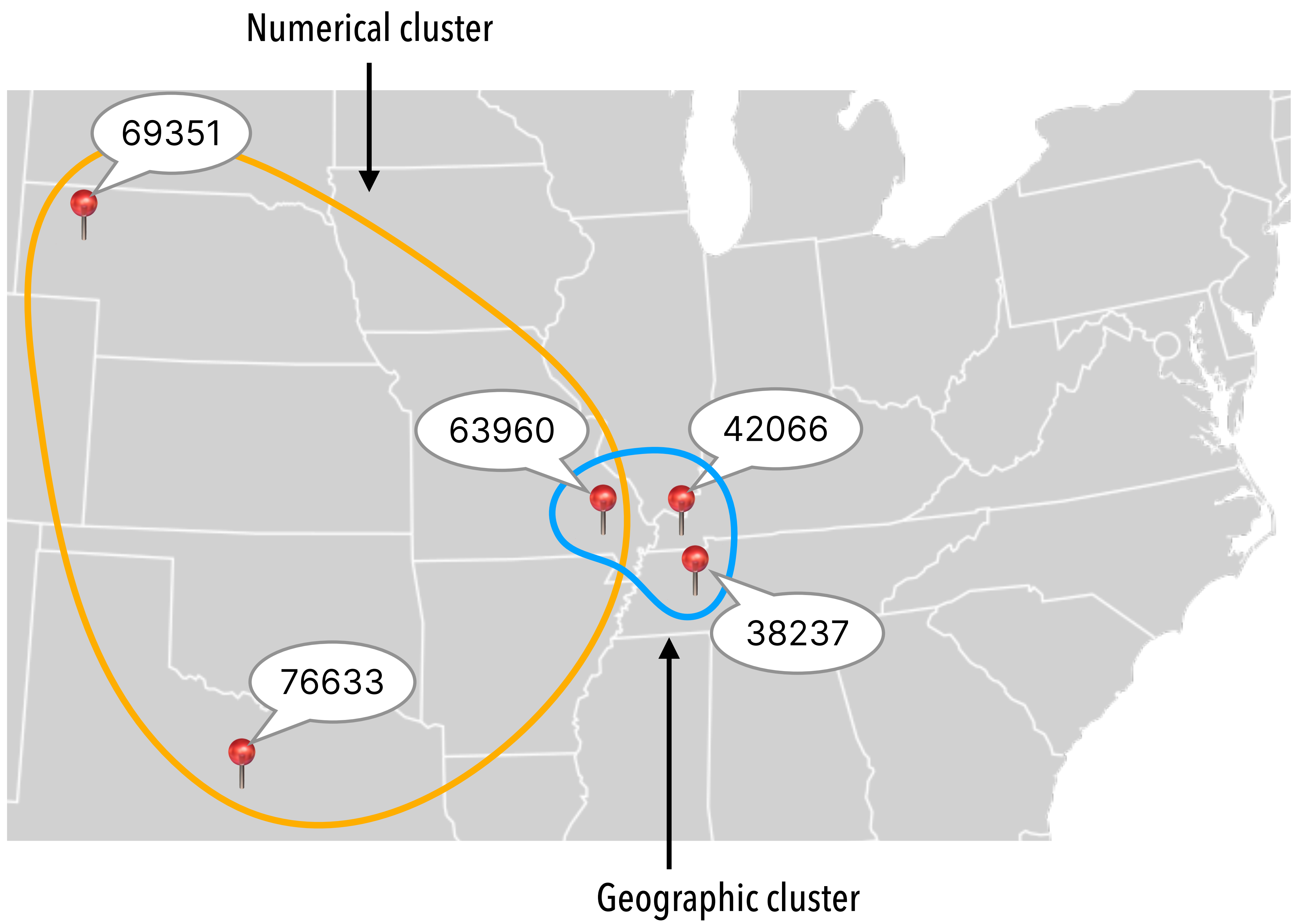


METHODS EXPLORED

Microaggregation

🔀 Perturbation: data can change

- Maintain data semantics for downstream analysis
- More precisely target cluster center with median/mode
- Target geometric/geographical center
- Avoid suppression



METHODS EXPLORED

Microaggregation

🔀 Perturbation: data can change

- Maintain data semantics for downstream analysis
- More precisely target cluster center with median/mode
- Target geometric/geographical center
- Avoid suppression

Sex	Sex	Sex
M	*	M
M	*	M
M	M	M
F	F	F
F	*	M
M	M	M
M	M	M
Unprotected	Suppressed	Mode

Conclusion

- Privacy Dynamics found microaggregation to offer balanced privacy and utility for data sharing
- Every data privacy method presents tradeoffs
- Most appropriate method depends many factors:
  - Sensitivity of content
  - Size of dataset
  - Expected analysis
  - Audience size and trust
  - *More*

Thank you

WILL@PRIVACYDYNAMICS.IO