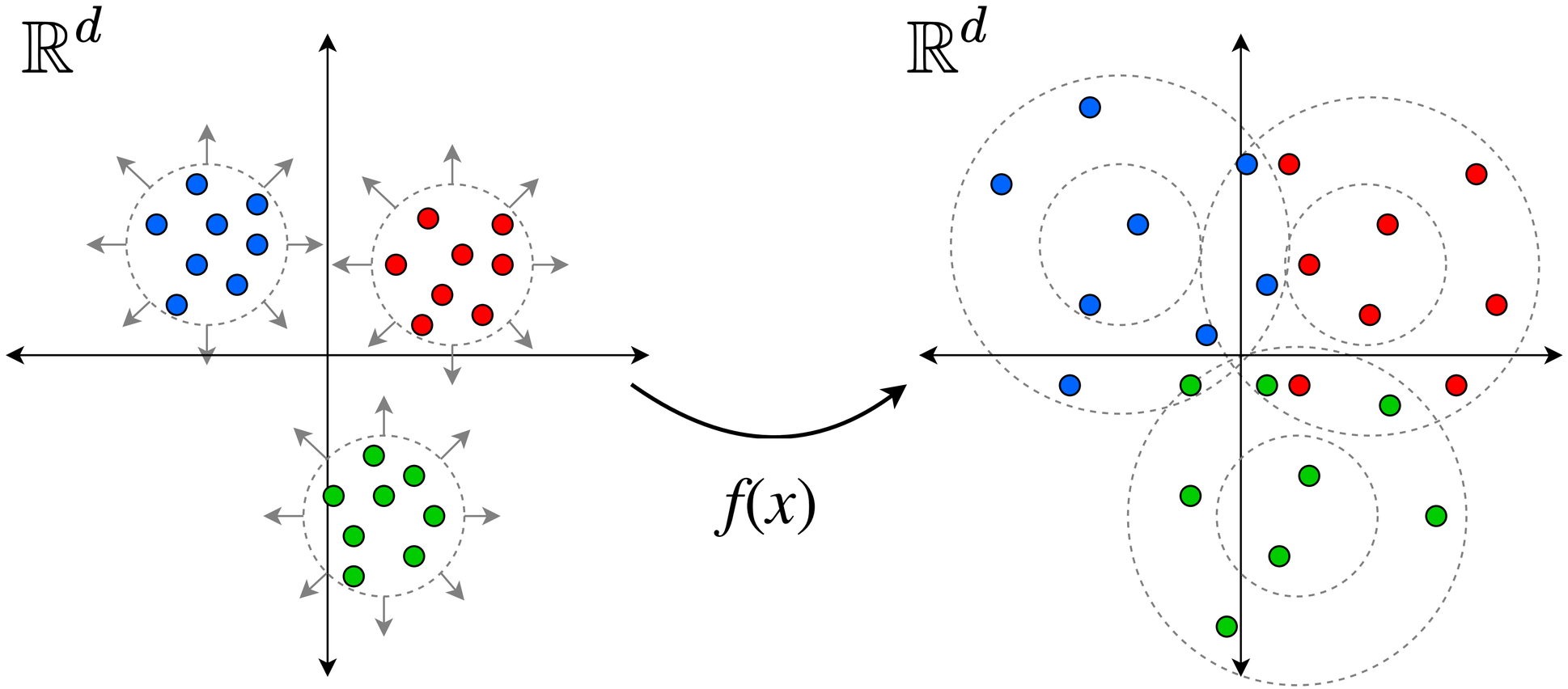
Fundamental Limits of Perfect Concept Erasure

Perfect **E**rasure **F**unctions (PEF, AISTATS 2025)

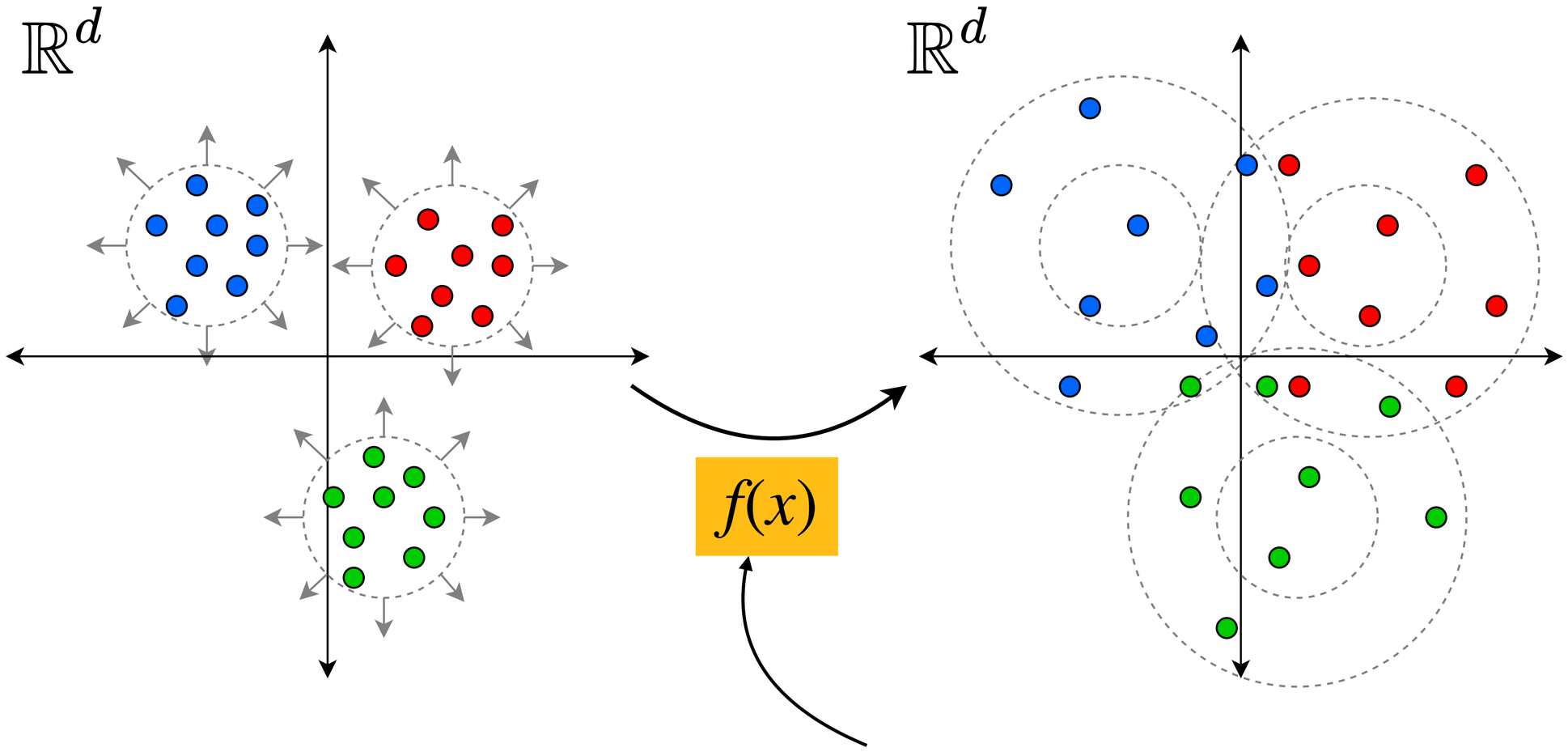
Concept Erasure PEF

Motivation



Concept Erasure PEF

Motivation



Can we analytically derive the perfect erasure function?

Setup & Key Assumptions

X : Input Representations (e.g., text representations)

Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

Setup & Key Assumptions

X : Input Representations (e.g., text representations)

Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

- Markov Property: $A \rightarrow X \overset{f}{\rightarrow} Z$

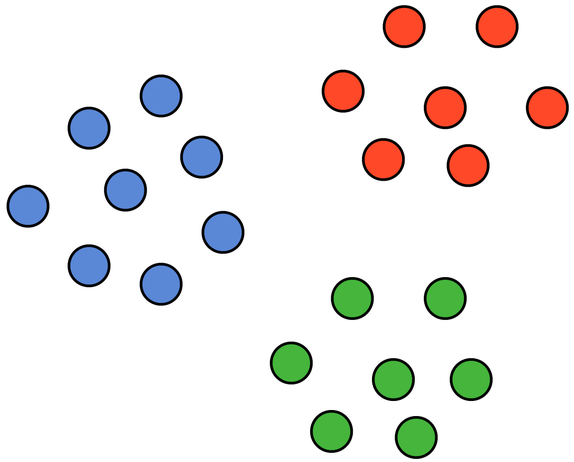
Setup & Key Assumptions

X : Input Representations (e.g., text representations)

Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

- Markov Property: $A \rightarrow X \xrightarrow{f} Z$
- Support sets ($\mathscr{X}, \mathscr{Z}, \mathscr{A}$) are finite



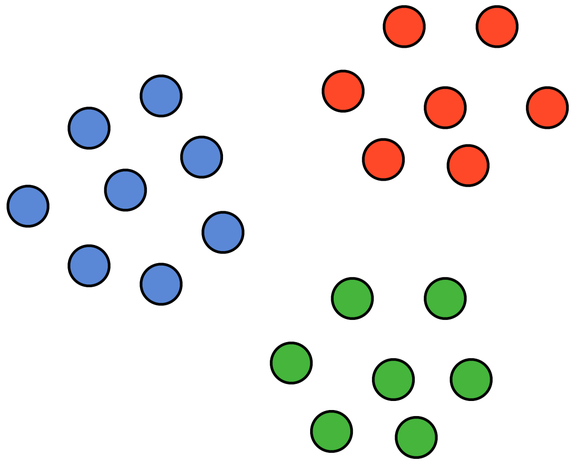
Setup & Key Assumptions

X : Input Representations (e.g., text representations)

Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

- Markov Property: $A \rightarrow X \xrightarrow{f} Z$
- Support sets ($\mathcal{X}, \mathcal{Z}, \mathcal{A}$) are finite
- $|\mathcal{X}| > |\mathcal{A}|$



Perfect Erasure

X : Input Representations (e.g., text representations)

Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

$$\max_{f} I(Z;X) \text{ subject to } I(Z;A) = 0.$$

Utility: mutual information with original representations

Privacy: mutual information with concept variable

Perfect Erasure

X : Input Representations (e.g., text representations)

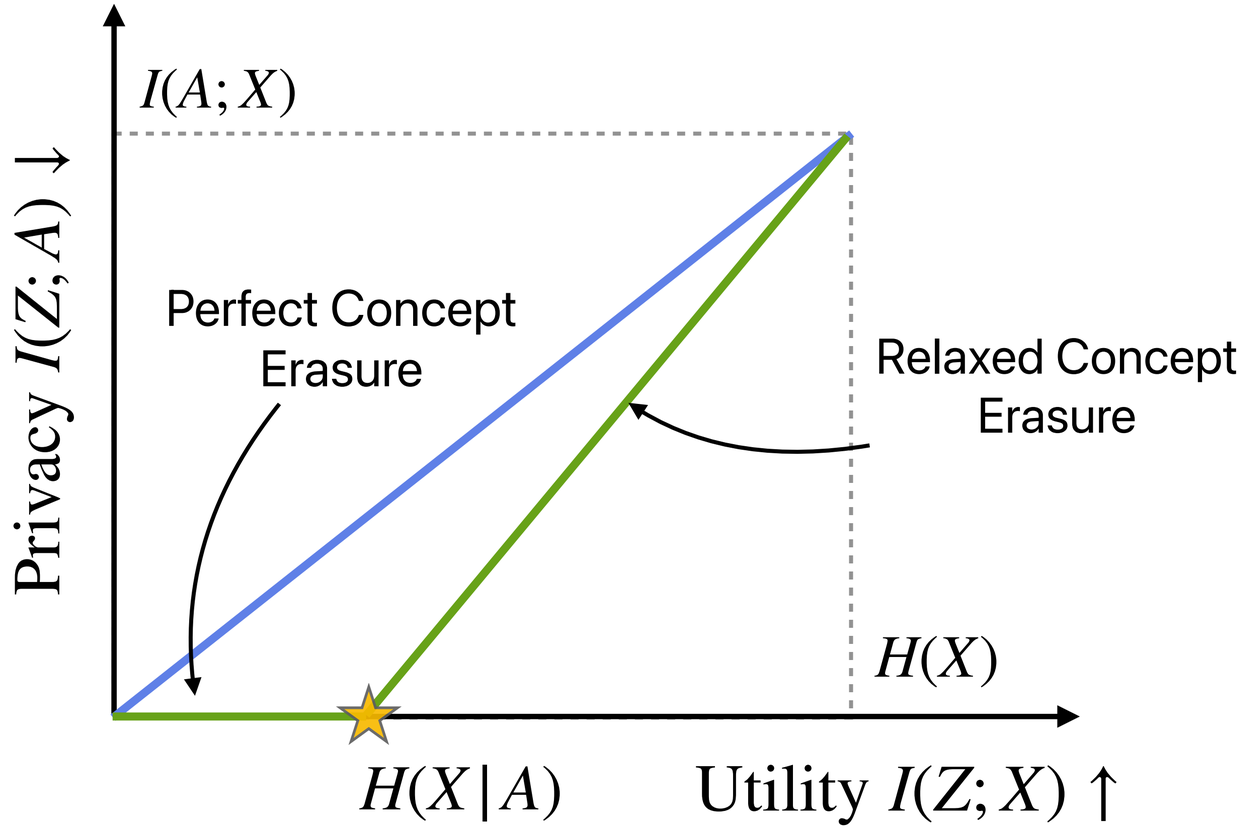
Z : Erased Representations (post erasure, $Z = f(X)$)

A : Categorical Concept (e.g., gender)

$$\max_{f} I(Z; X) \text{ subject to } I(Z; A) = 0.$$

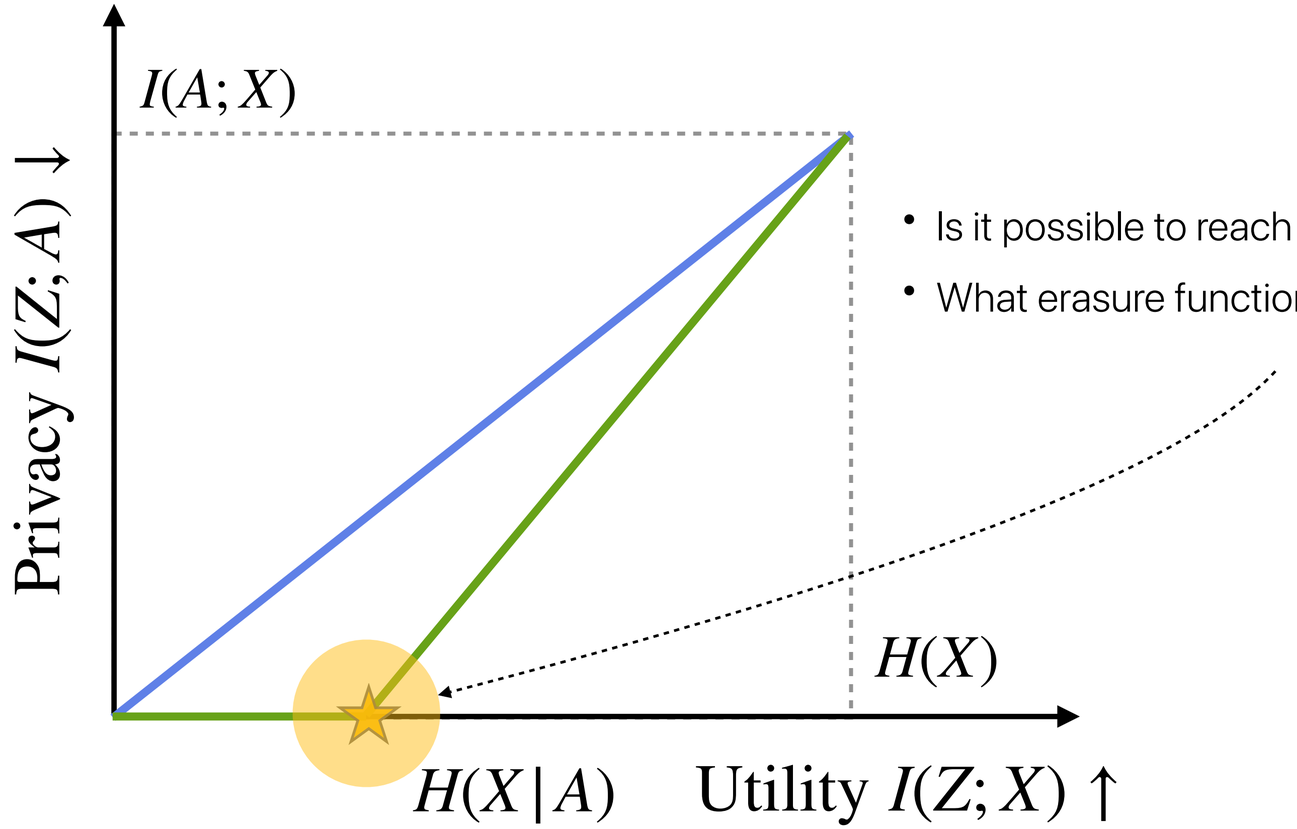
Optimize for f

Privacy Funnel



[Calmon et al., 2017] Principal Inertia Components and its Applications.

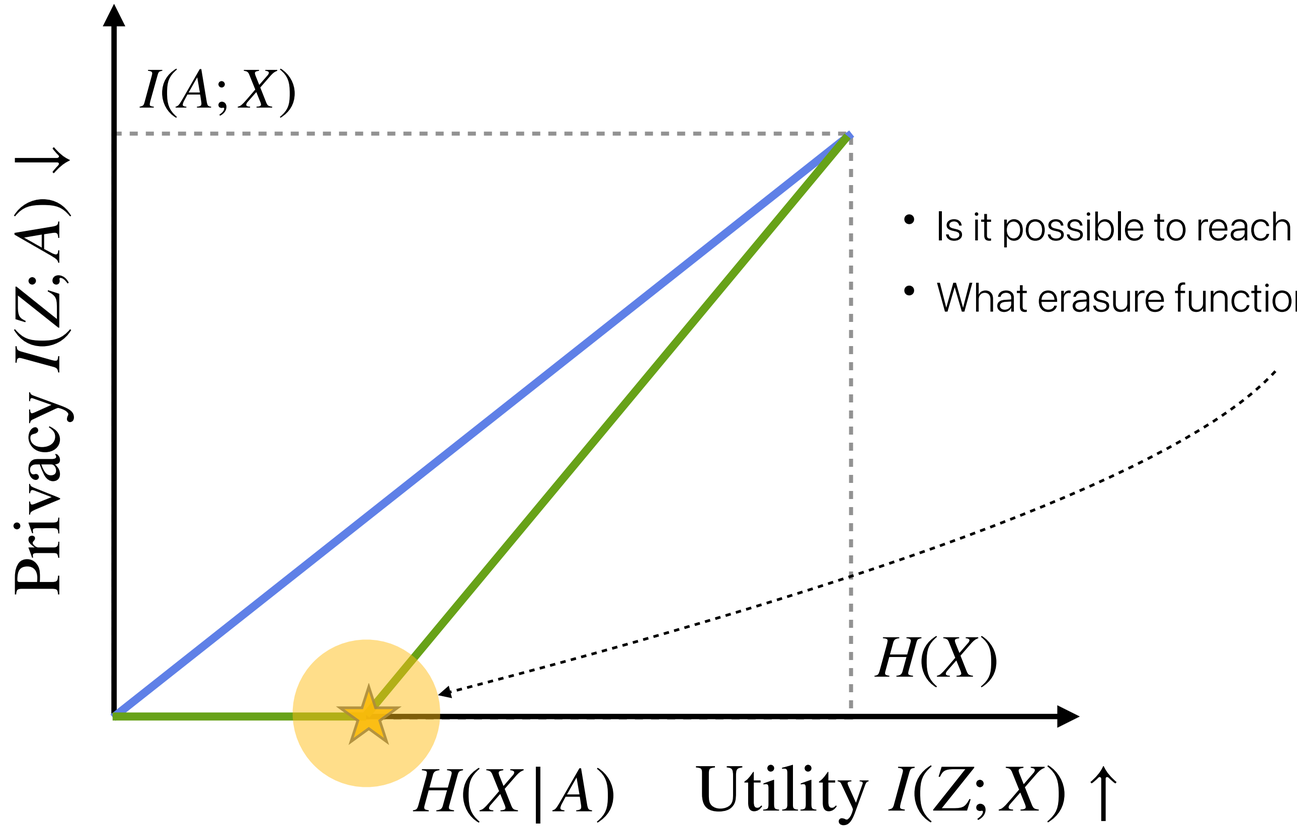
Privacy Funnel



- Is it possible to reach this point for any X, A?
- What erasure function f achieves this point?

[Calmon et al., 2017] Principal Inertia Components and its Applications.

Privacy Funnel



- Is it possible to reach this point for any X, A?
- What erasure function f achieves this point?

[Calmon et al., 2017] Principal Inertia Components and its Applications.

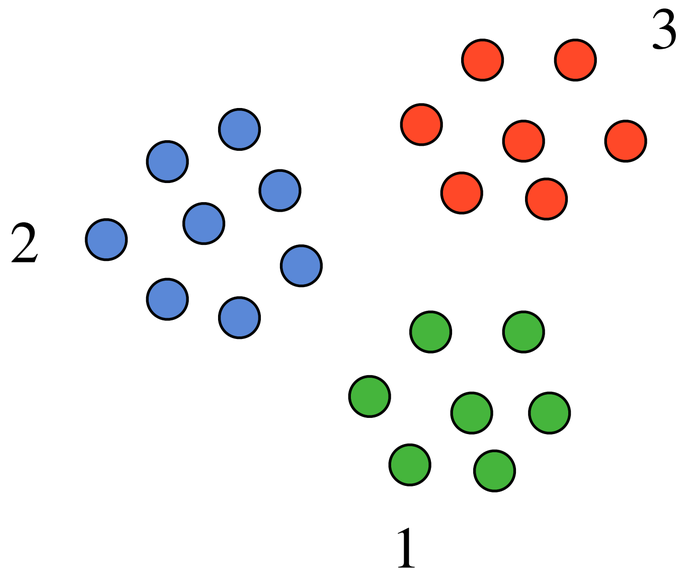
Main Result: Feasibility

Perfect concept erasure is feasible if and only if (i, j):

(a) Support sizes of concept groups are same, $|\mathcal{X}_i| = |\mathcal{X}_j|$

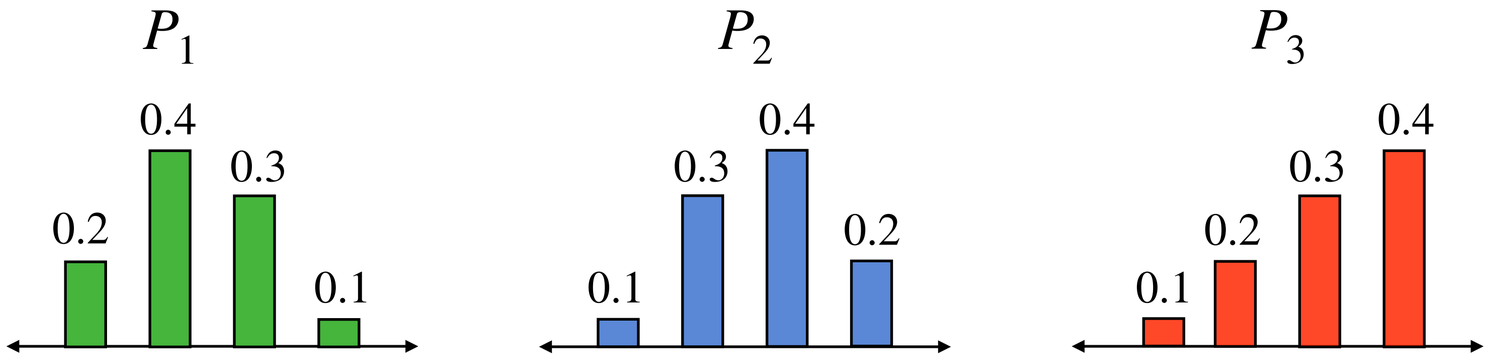
(b) Distribution of representations are permutations, $P(\mathcal{X}_i) = \sigma(P(\mathcal{X}_j))$

Main Result: Feasibility



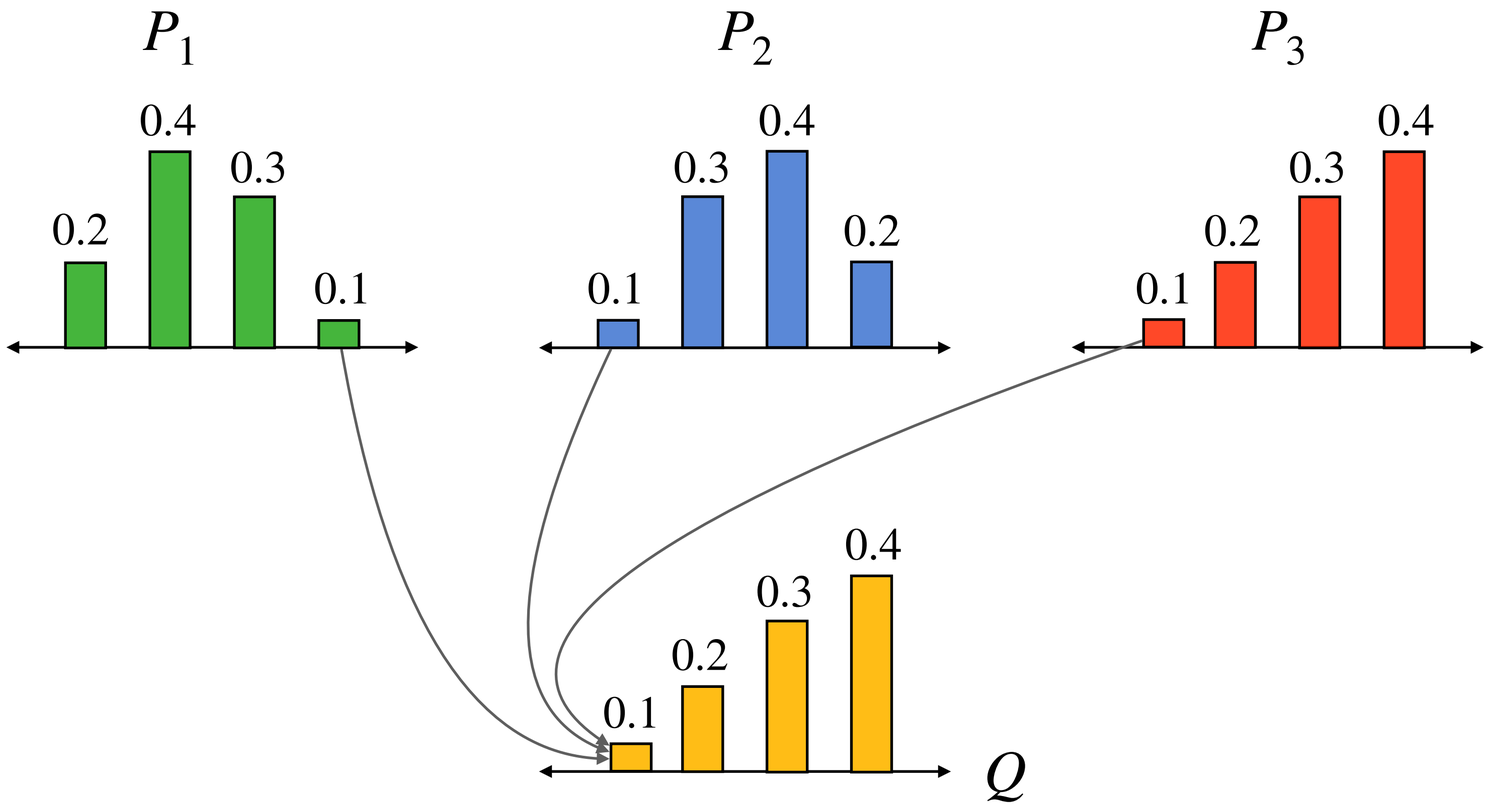
Concept Erasure PEF

Main Result: Feasibility

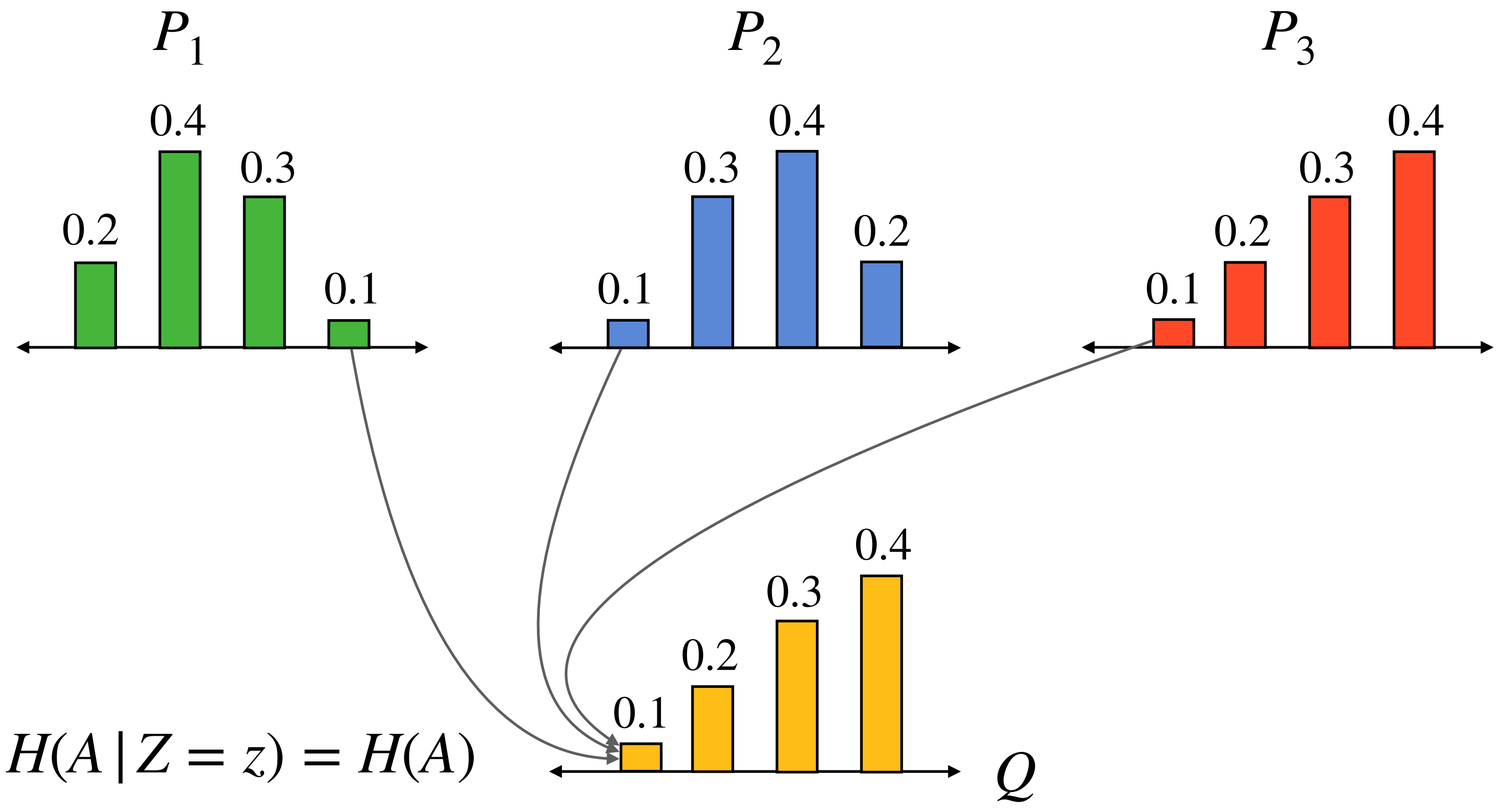


Distributions are permutations of each other, $P(\mathcal{X}_i) = \sigma(P(\mathcal{X}_j))$

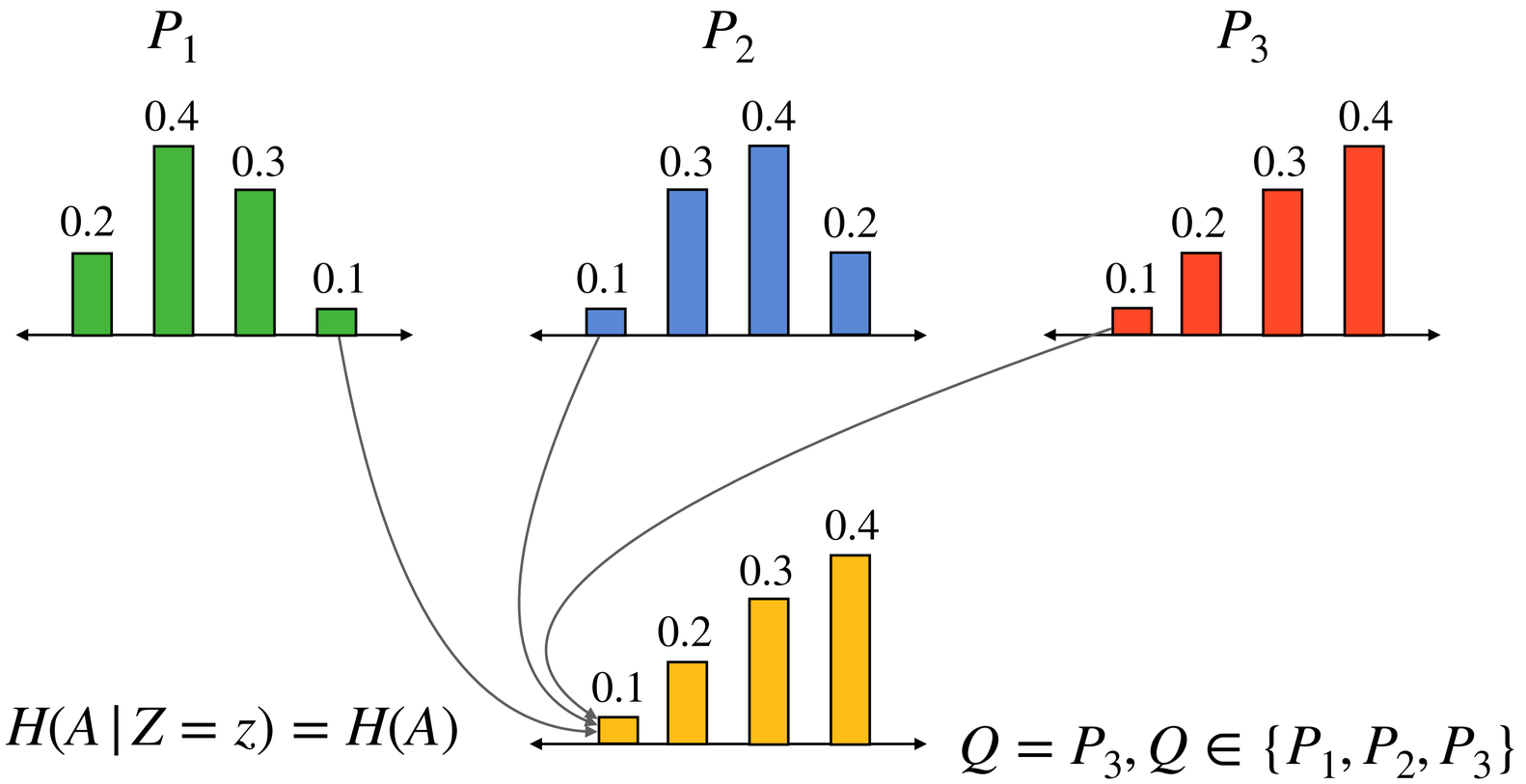
Main Result: Erasure function



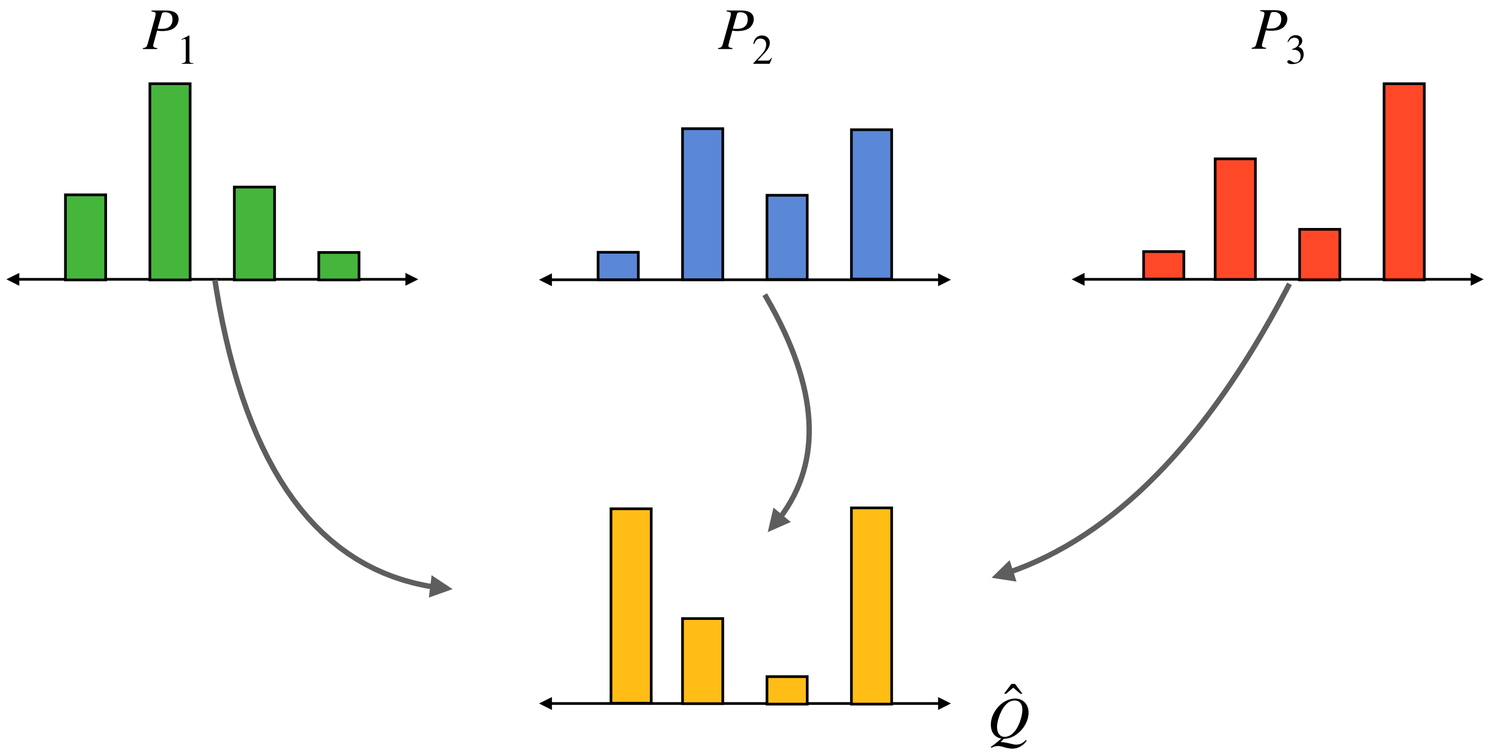
Main Result: Erasure function



Main Result: Erasure function



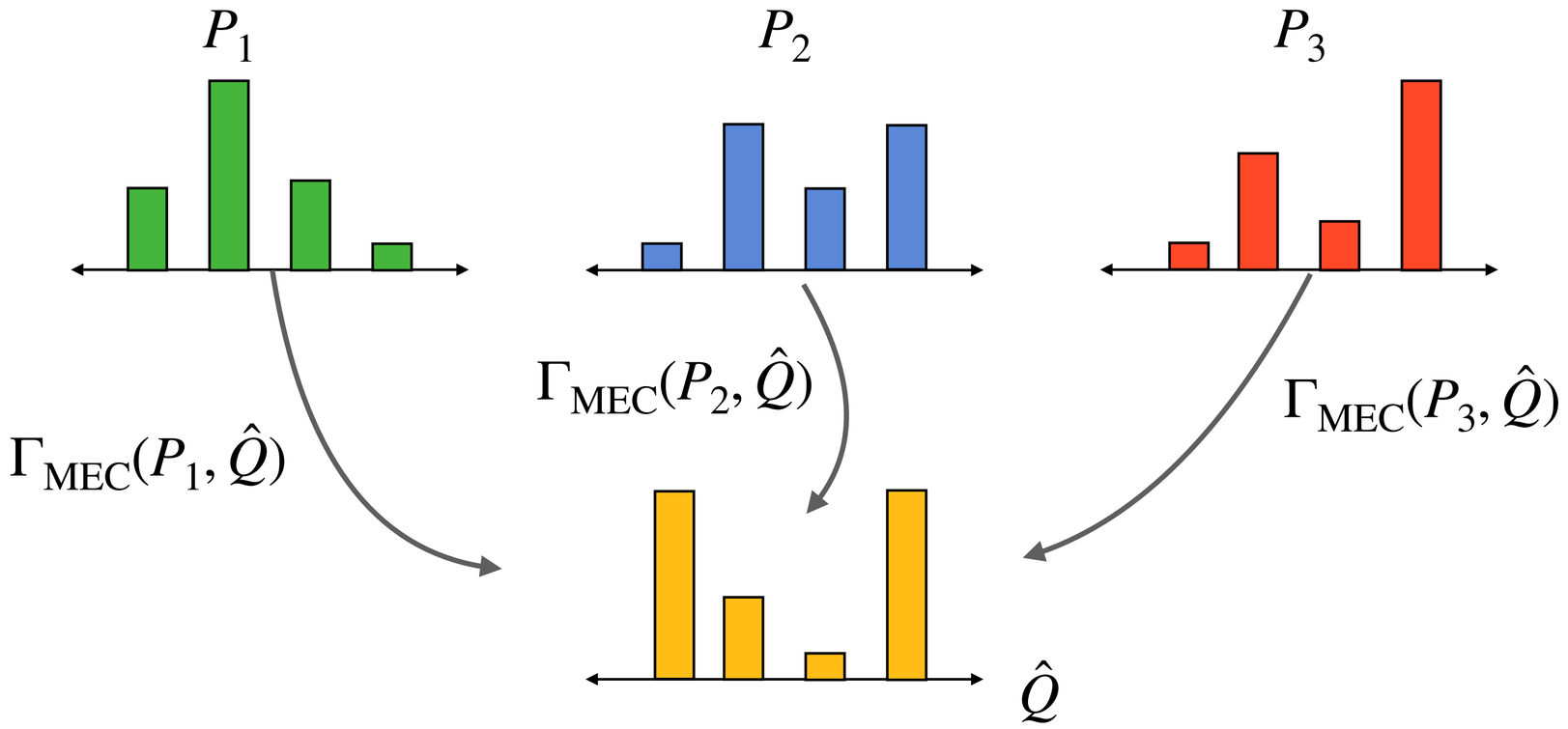
Main Result: Unequal Distributions



Concept Erasure PEF

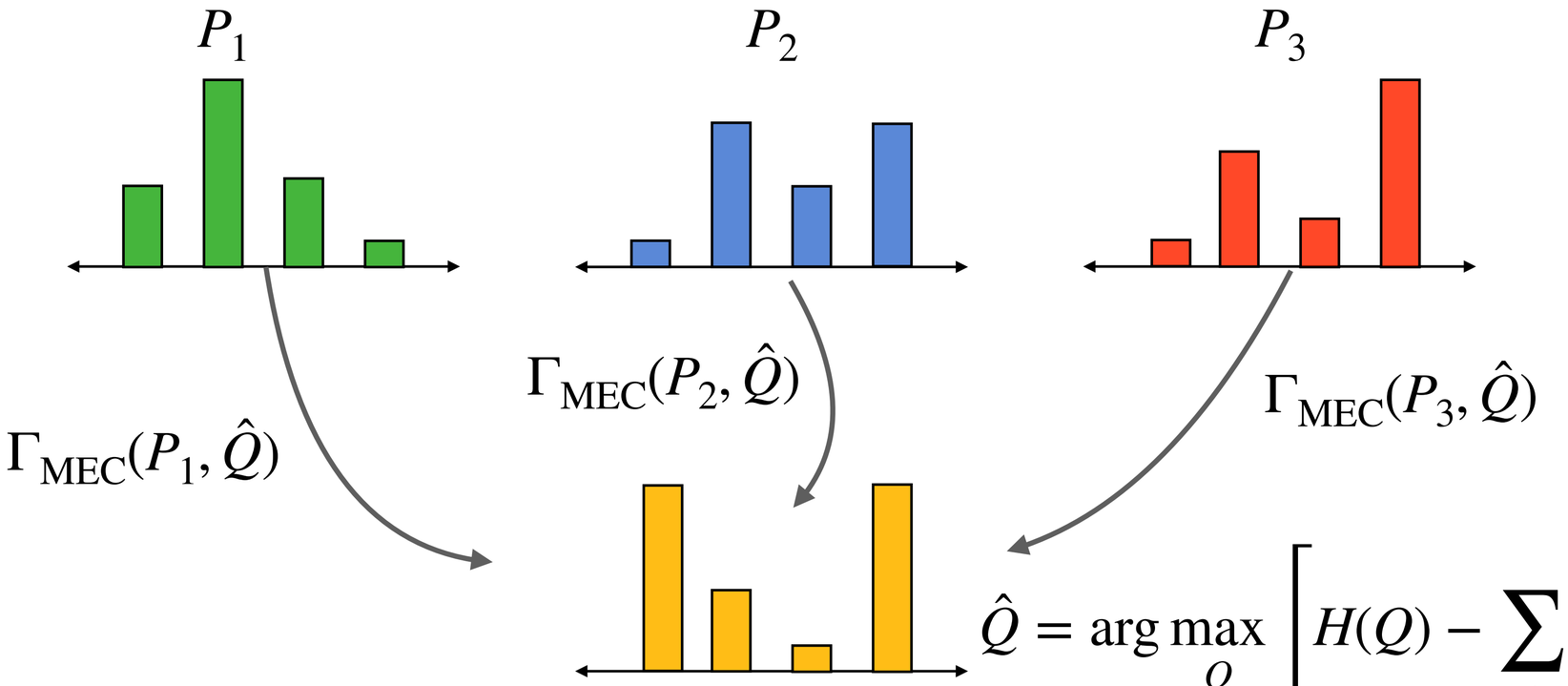
Main Result: Unequal Distributions

$\Gamma_{\text{MEC}}(\cdot, \cdot)$ is the minimum entropy coupling (minimizes $H(P_i, Q)$)



Main Result: Unequal Distributions

$\Gamma_{\text{MEC}}(\cdot, \cdot)$ is the minimum entropy coupling (minimizes $H(P_i, Q)$)

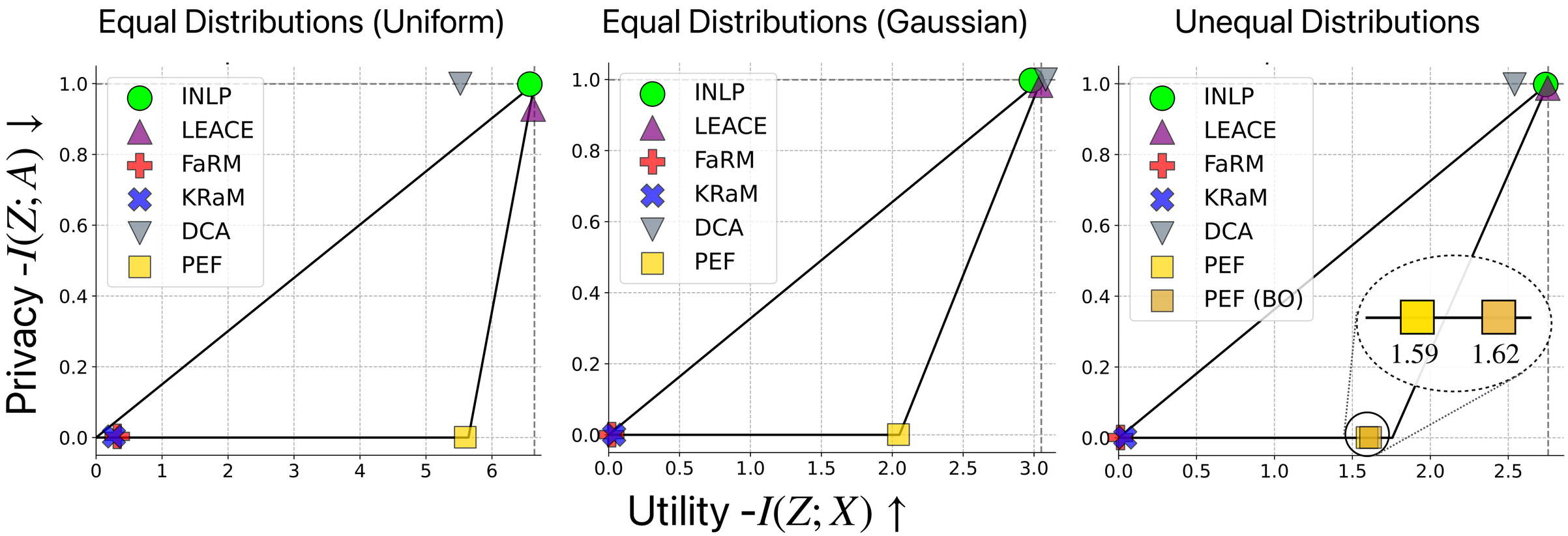


$$\hat{Q} = \arg\max_{Q} \left[H(Q) - \sum_i p(a_i) H_{min}(P_i, Q) \right]$$

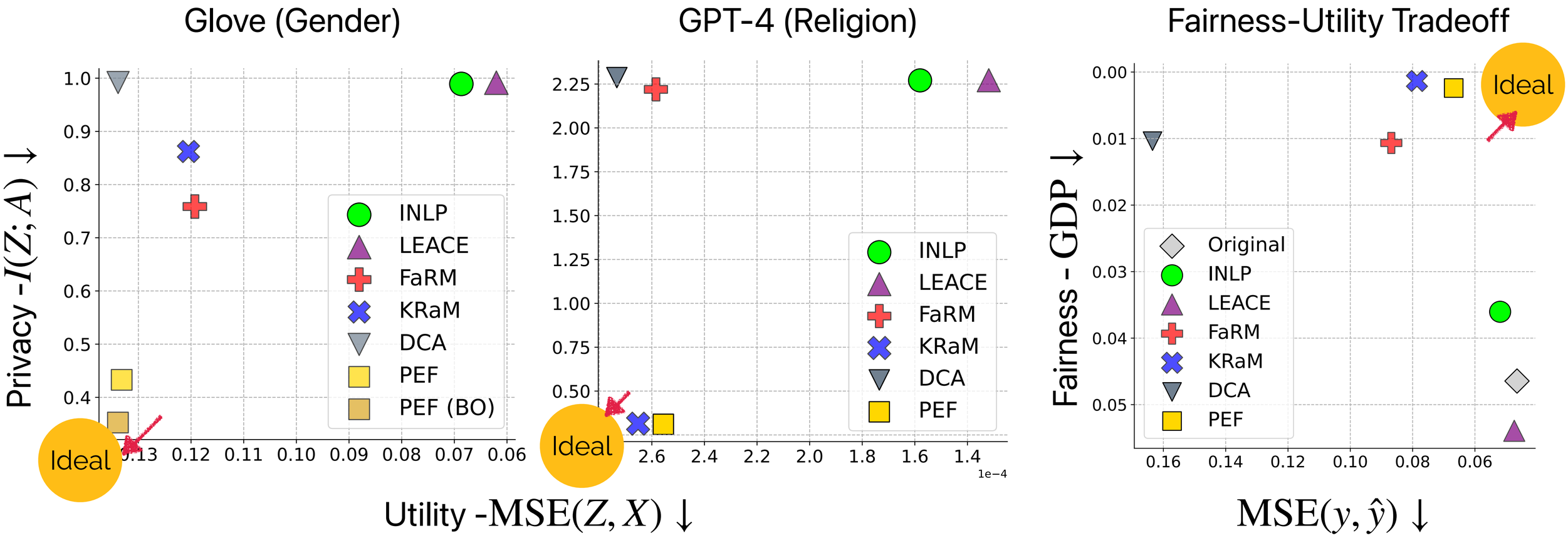
Experimental Setting

- Experiments using synthetic and real-world representations
- PEF is able to achieve the theoretical guarantees empirically
- **Toxicity classification**: Erasure helps improve fairness in text classification using GPT-4 representations

Experimental Results (Synthetic)



Experimental Results (Real-world)



Takeaways

[T1] PEF We derive the fundamental limits and data constraints for perfect erasure 

[T2] PEF achieves perfect erasure under mild assumptions 

[T3] PEF is effective in real-world scenarios outperforming existing techniques