Obliviator Reveals the Cost of *Nonlinear Guardedness* in Concept Erasure

Ramin Akbari*

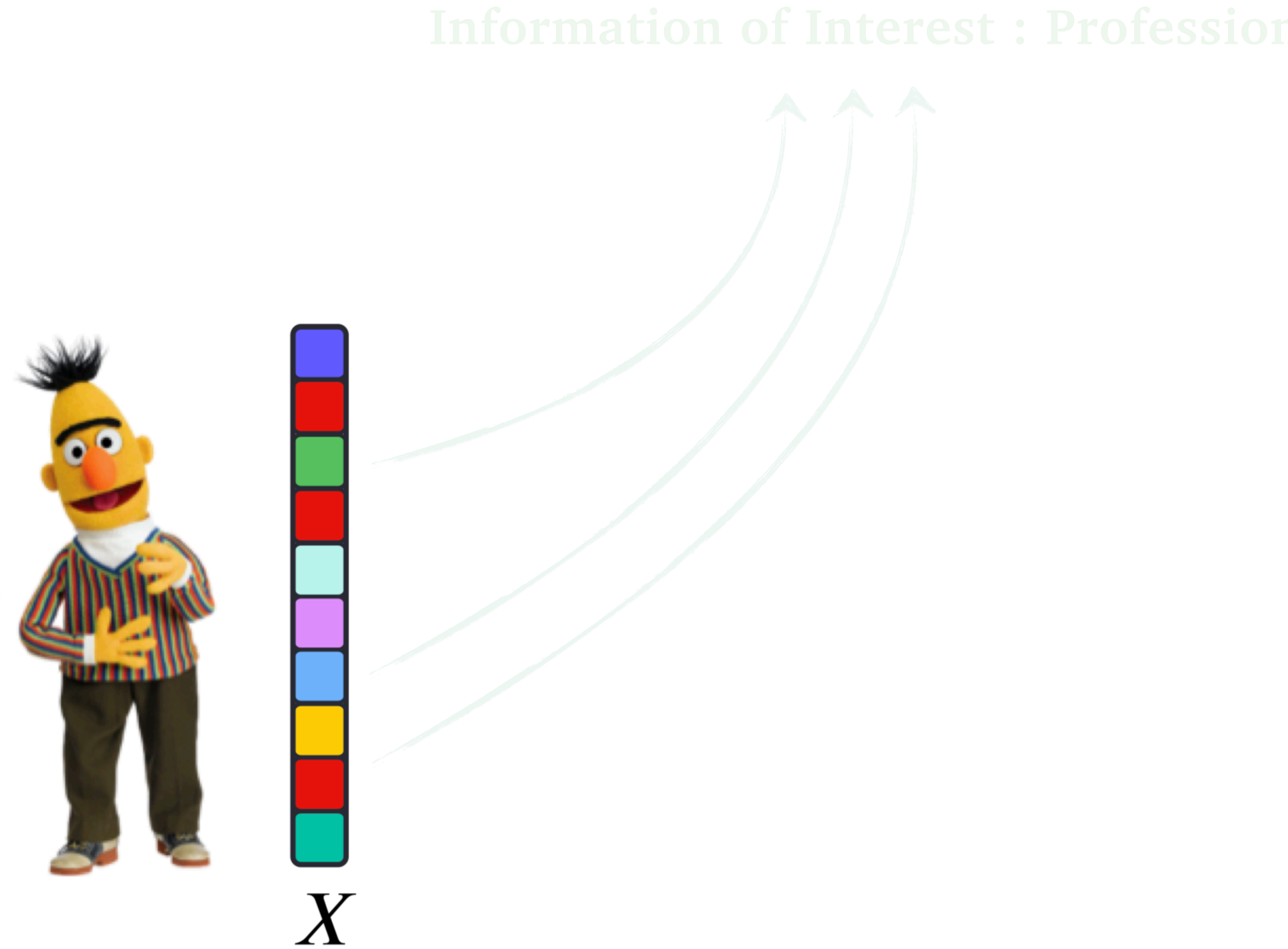
Milad Afshari*

Vishnu Naresh Boddeti

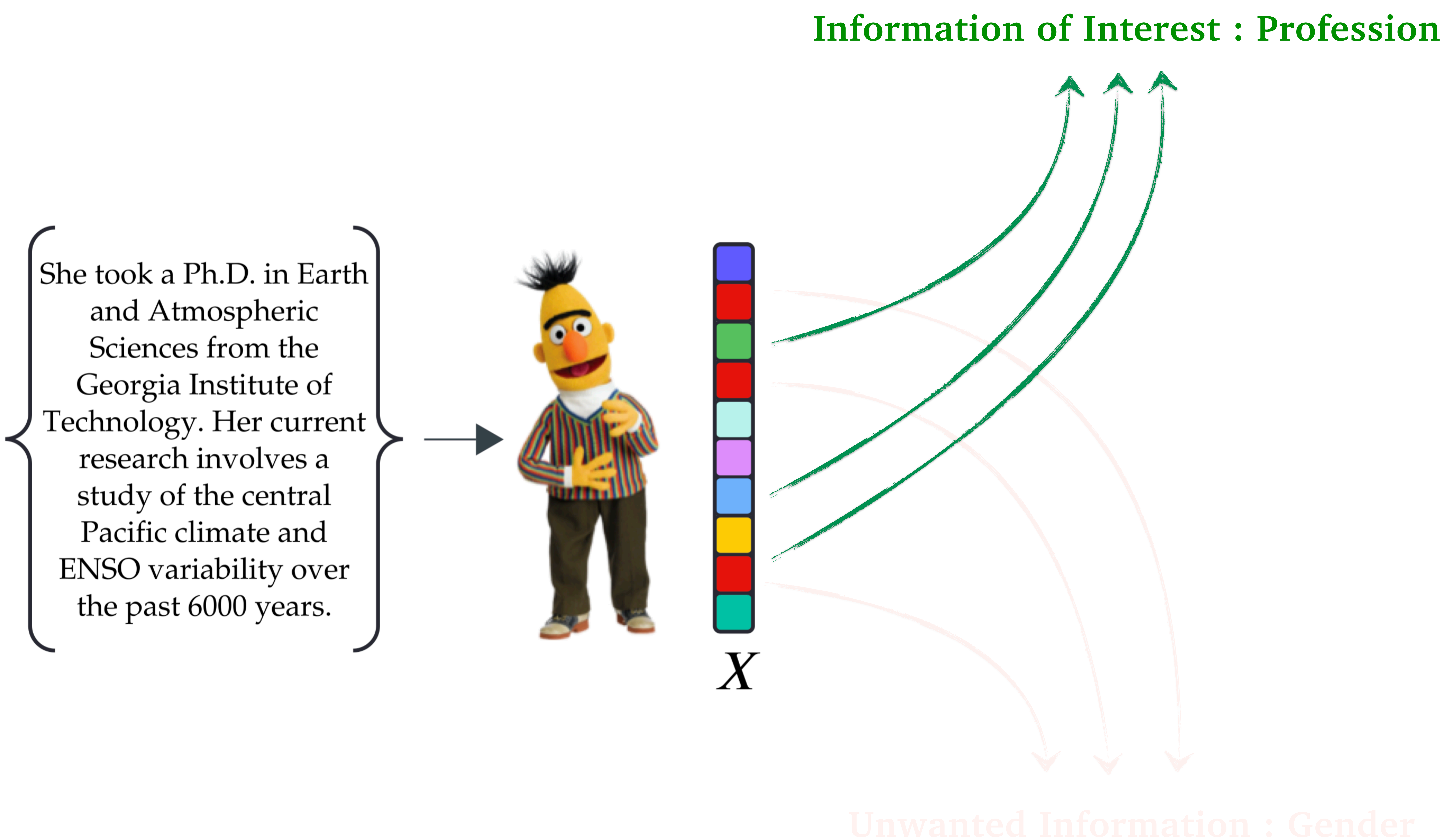
*Equal Contribution

Why Do We Need Erasure ?

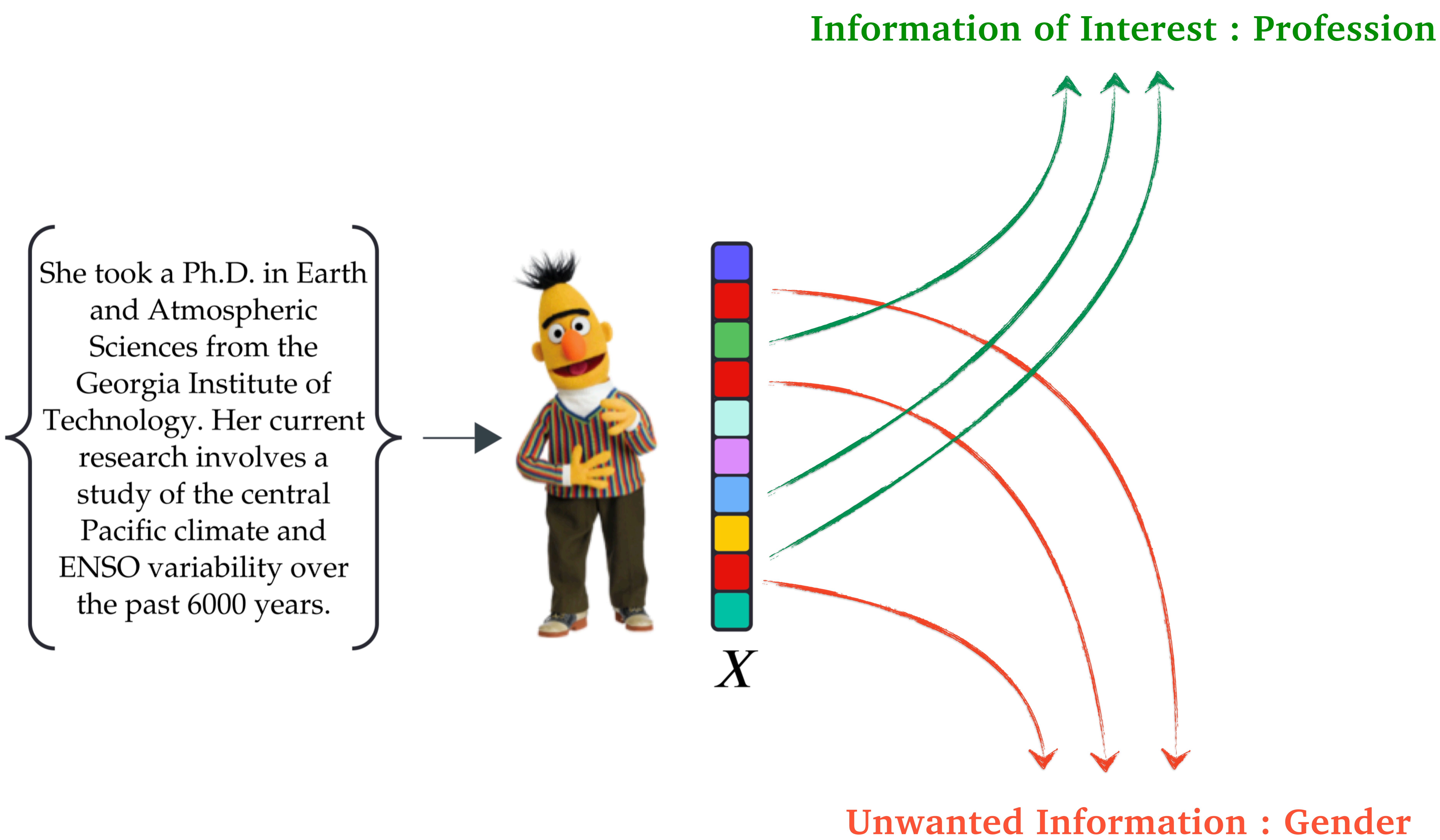
She took a Ph.D. in Earth and Atmospheric Sciences from the Georgia Institute of Technology. Her current research involves a study of the central Pacific climate and ENSO variability over the past 6000 years.



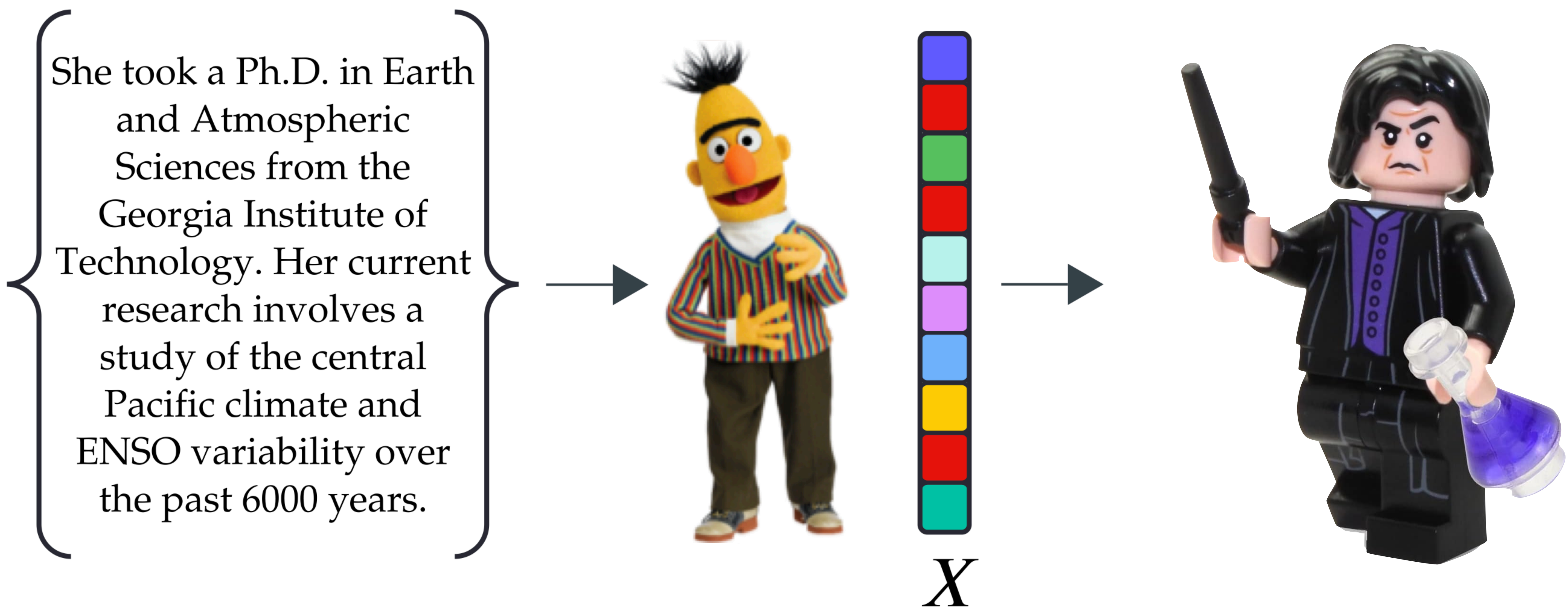
Why Do We Need Erasure ?



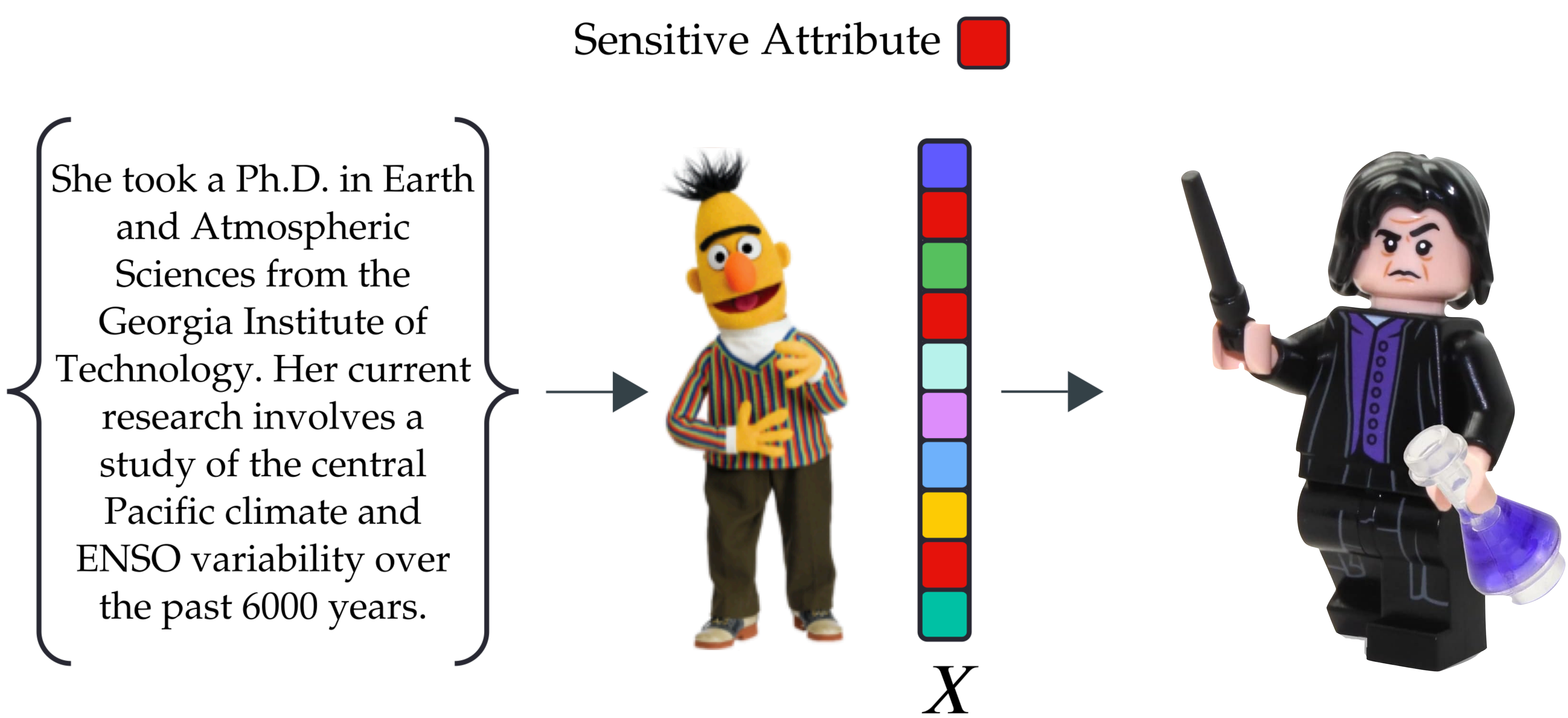
Why Do We Need Erasure ?



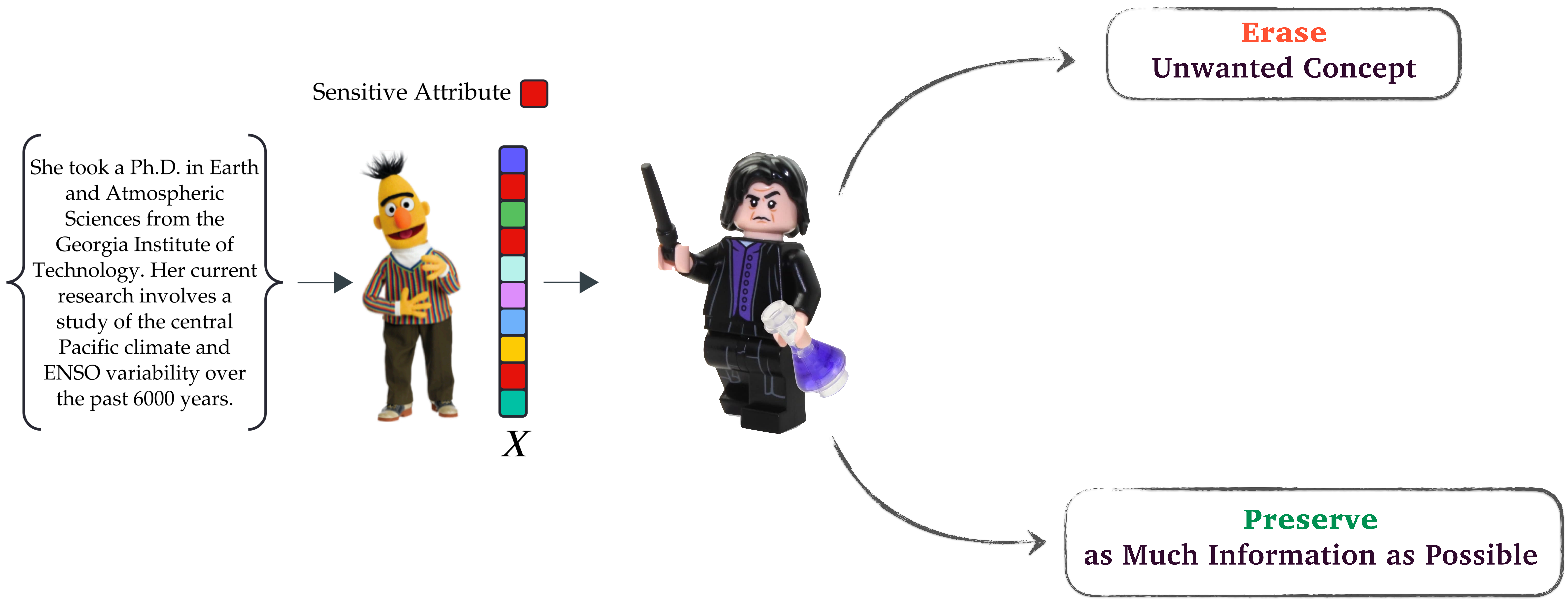
What Is The Ultimate Goal of Concept Erasure ?



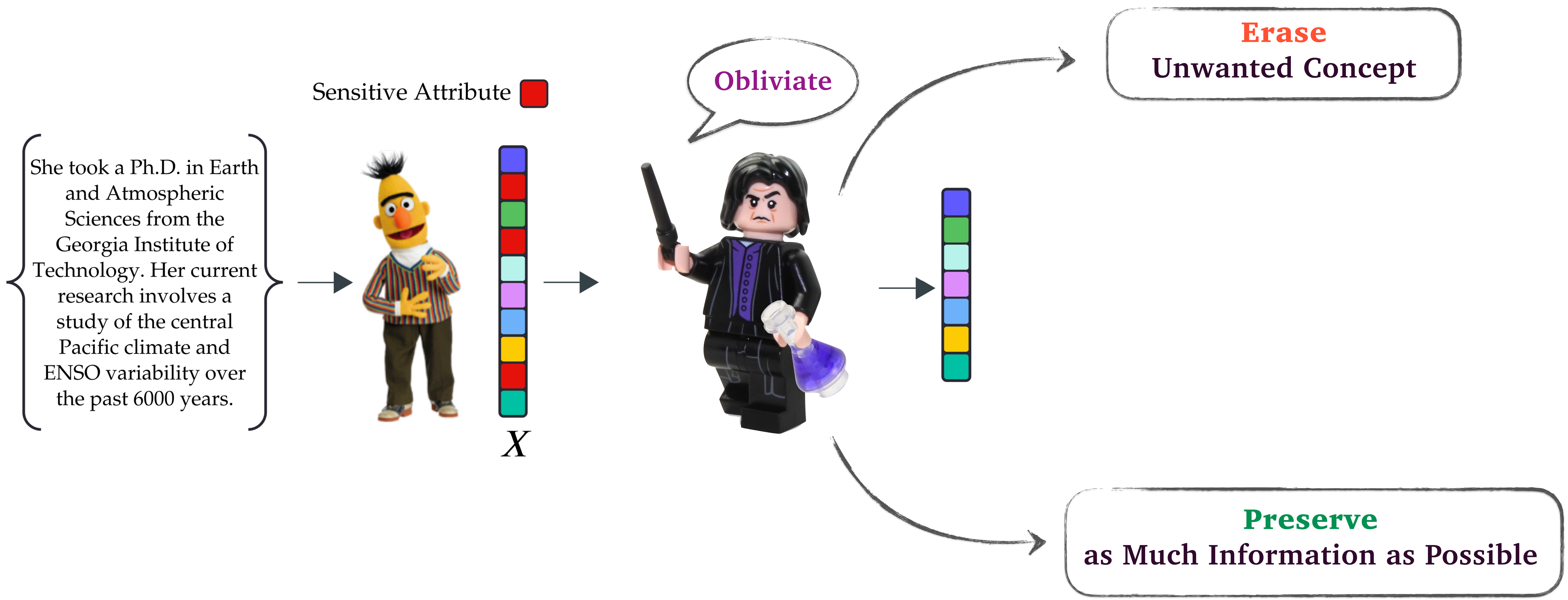
What Is The Ultimate Goal of Concept Erasure ?



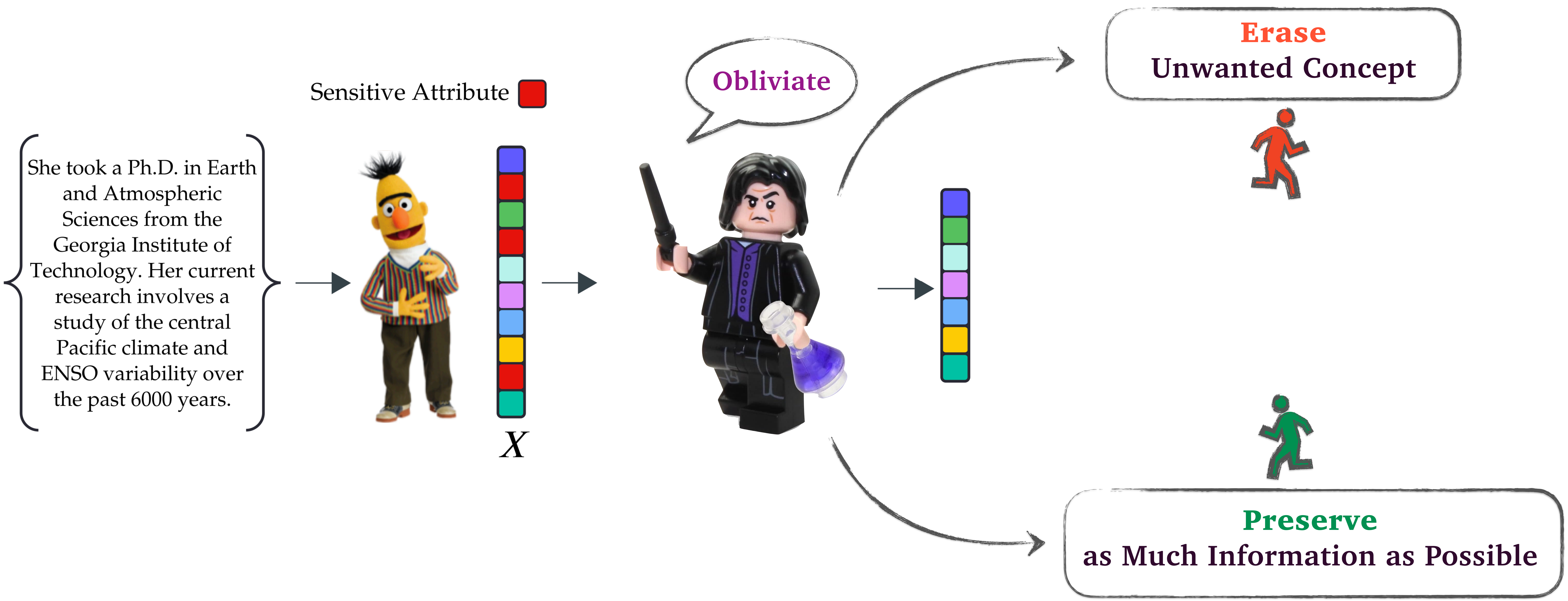
What Is The Ultimate Goal of Concept Erasure ?



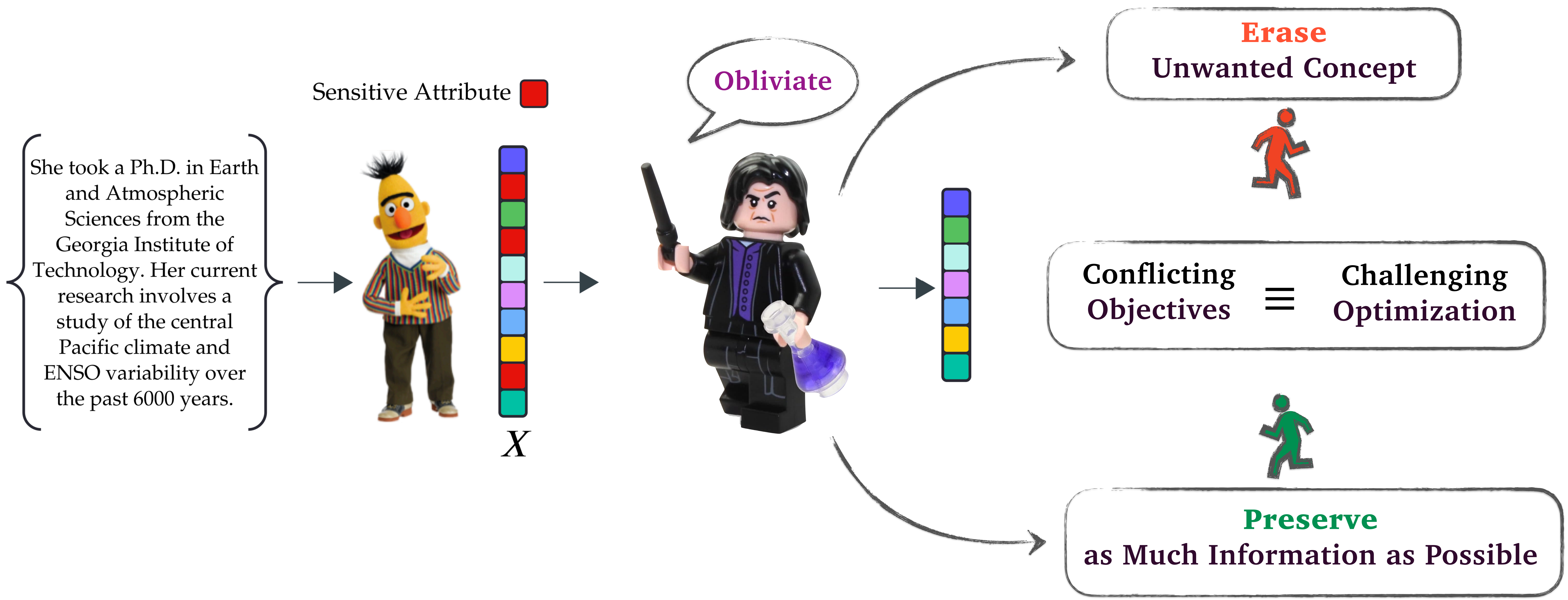
What Is The Ultimate Goal of Concept Erasure ?



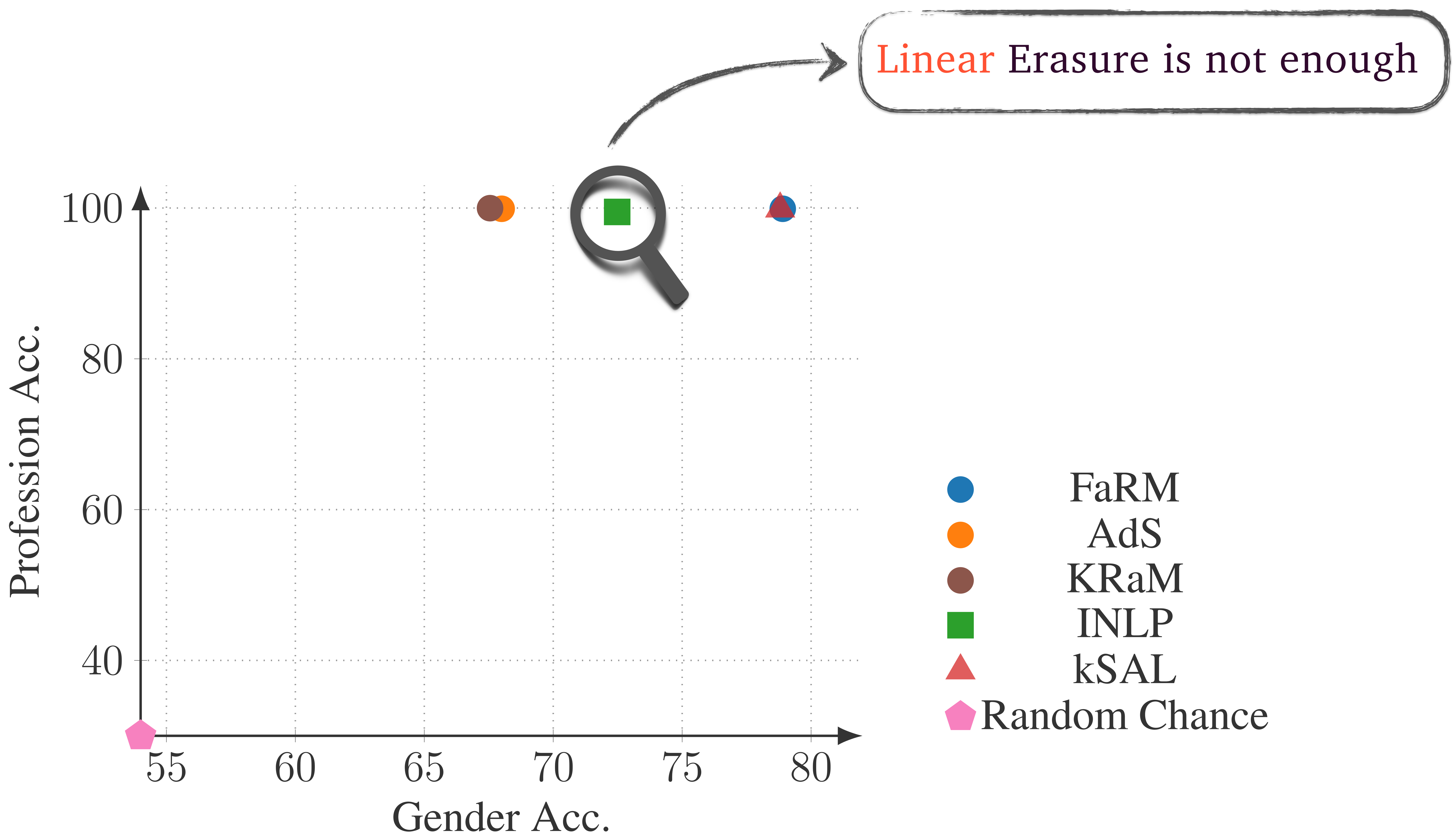
What Is The Ultimate Goal of Concept Erasure ?



What Is The Ultimate Goal of Concept Erasure ?



How Effective Are Current Methods In Concept Erasure ?



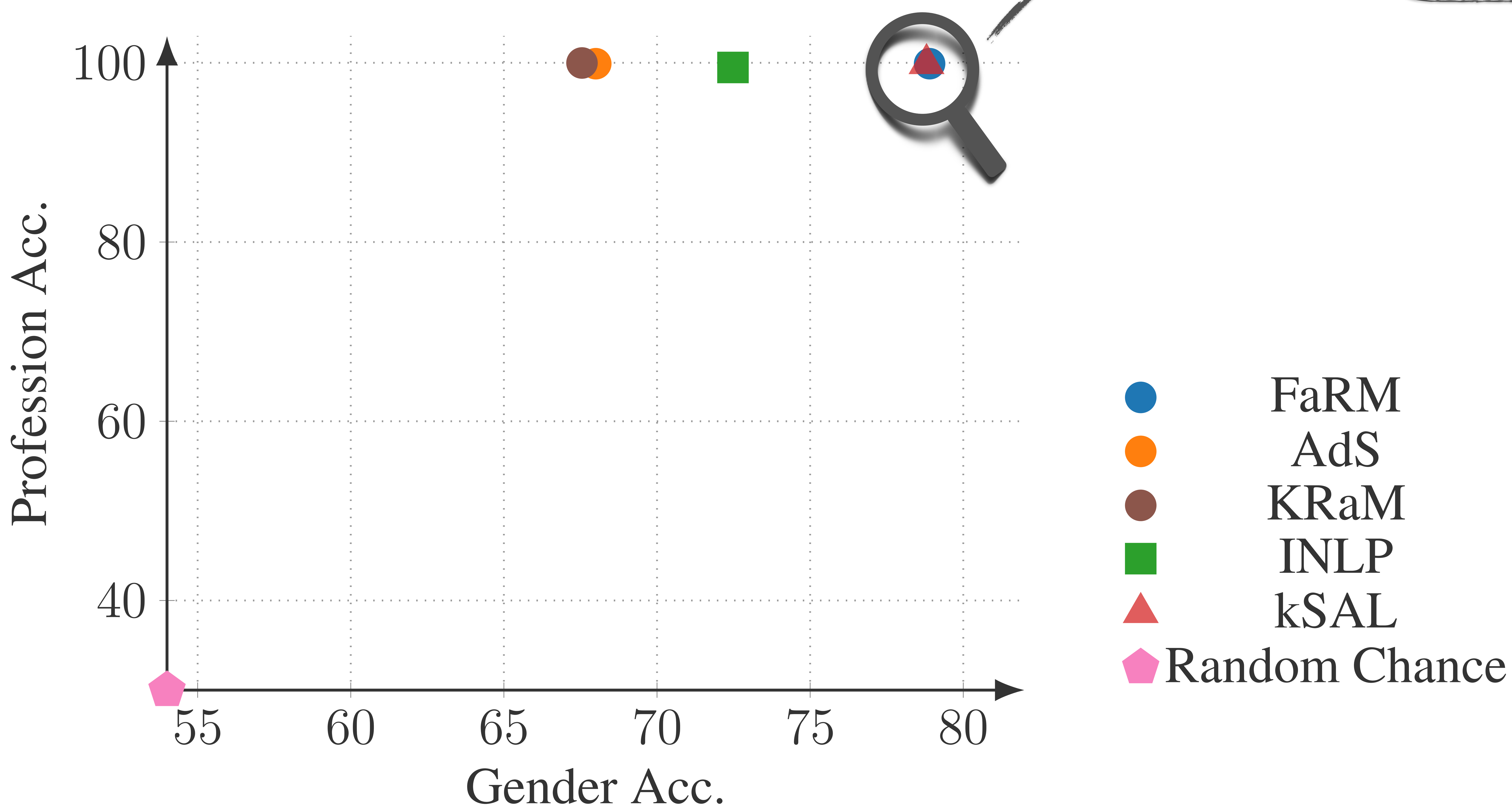
Bias in Bios

Utility : Profession

Unwanted : Gender

How **Effective** Are **Current** Methods In Concept Erasure ?

Some Nonlinear methods
did Worse than Linear method

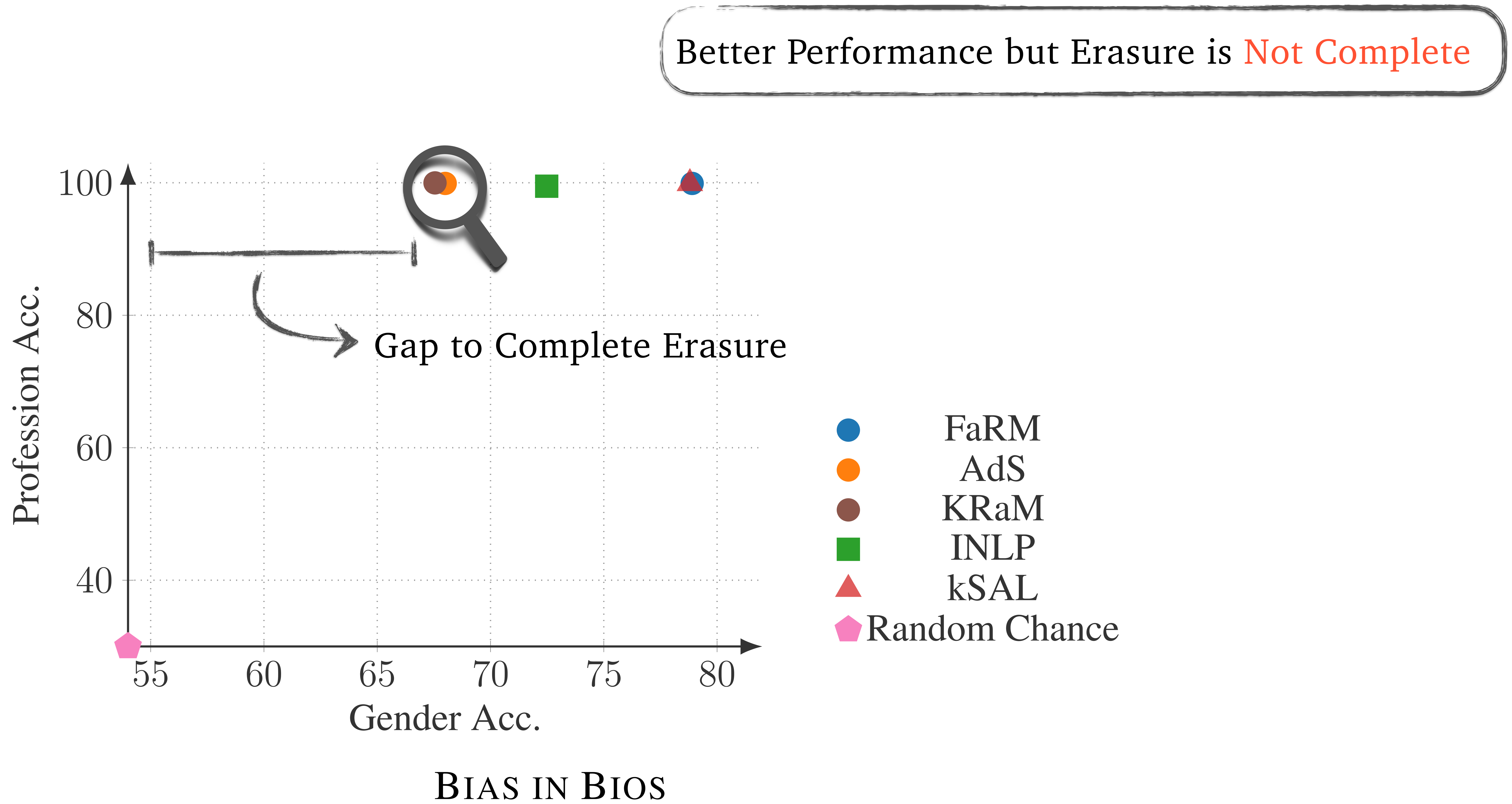


Bias in Bios

Utility : Profession

Unwanted : Gender

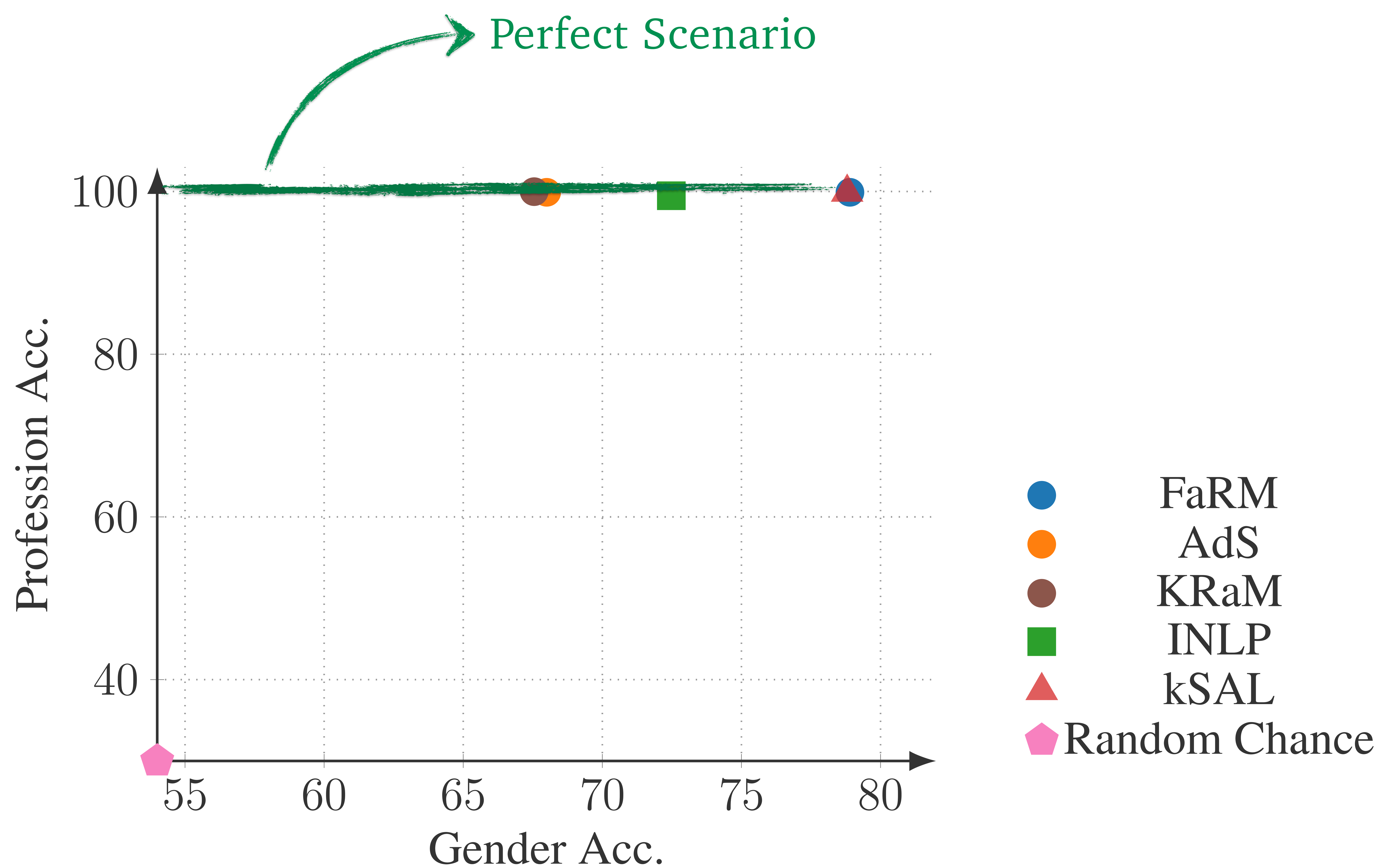
How Effective Are Current Methods In Concept Erasure ?



Utility : Profession

Unwanted : Gender

What Is The Complete Utility-Erasure Trade-off Profile ?

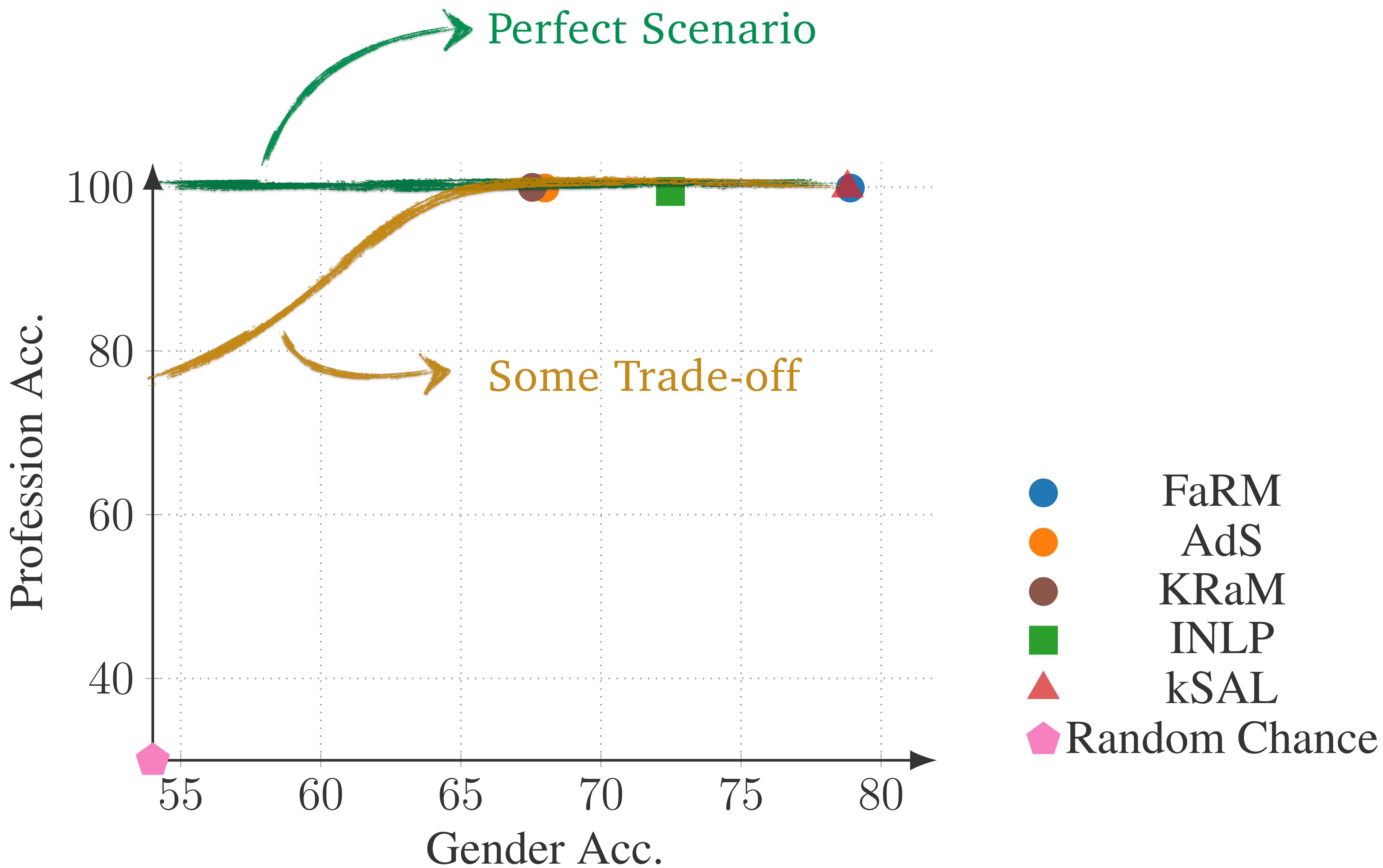


BIAS IN BIOS

Utility : Profession

Unwanted : Gender

What Is The Complete Utility-Erasure Trade-off Profile ?

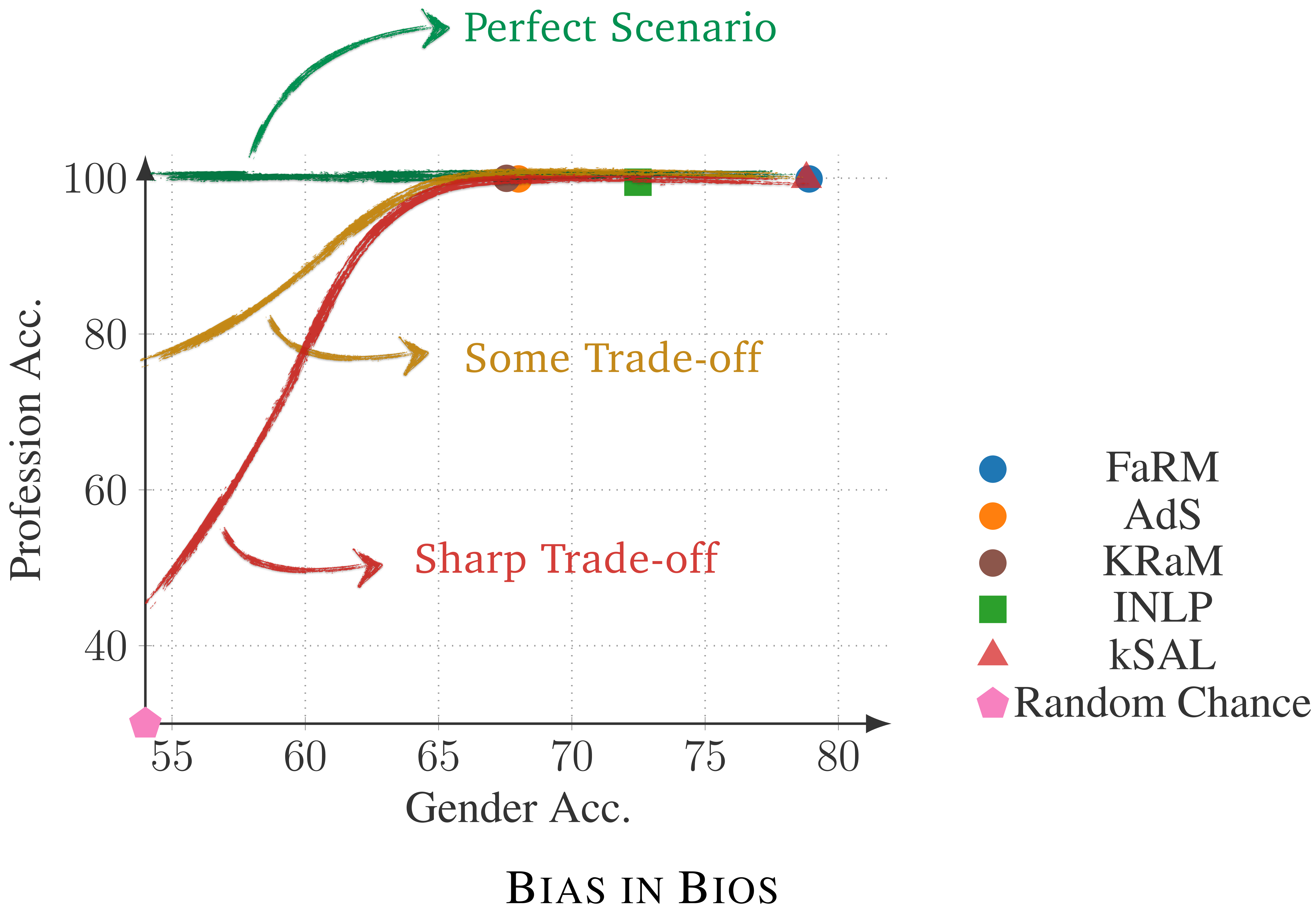


Bias in Bios

Utility : Profession

Unwanted : Gender

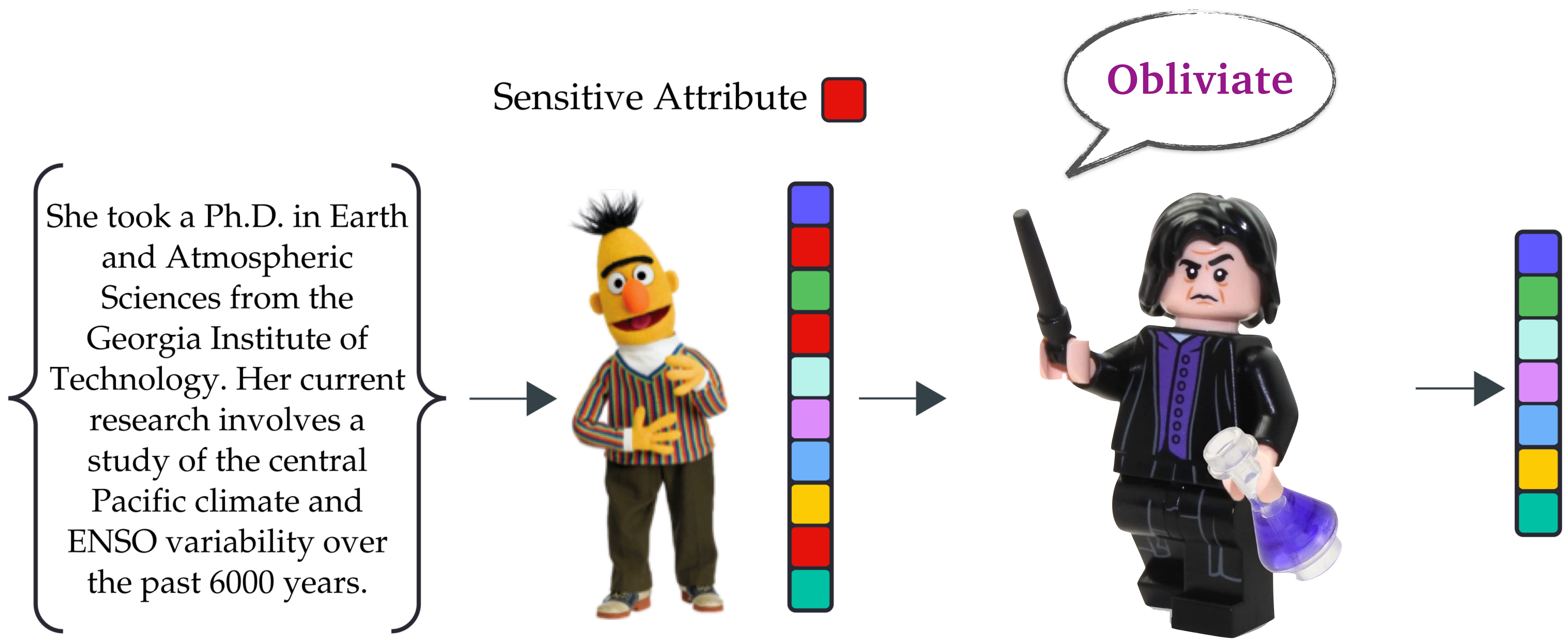
What Is The Complete Utility-Erasure Trade-off Profile ?



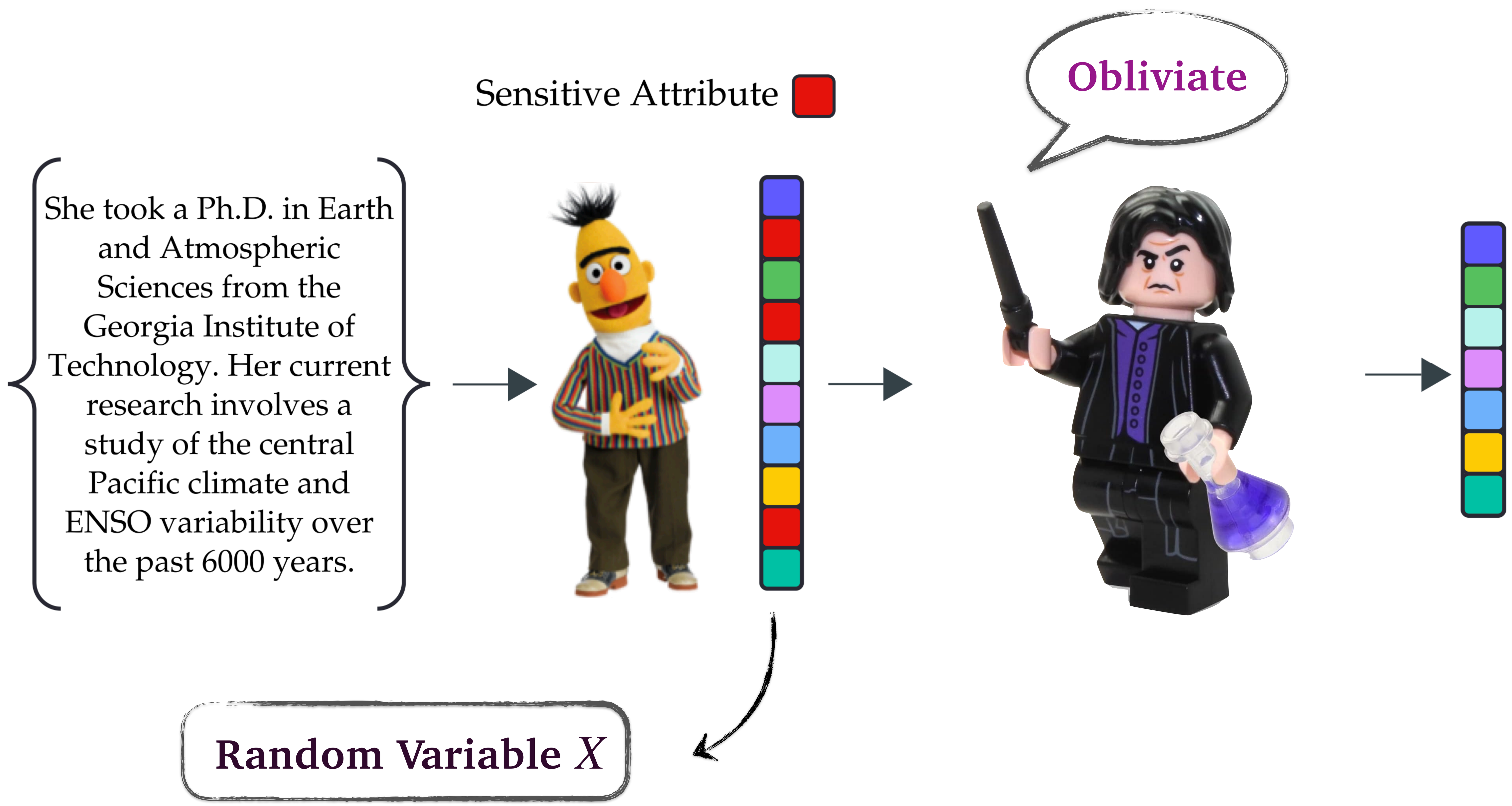
Utility : Profession

Unwanted : Gender

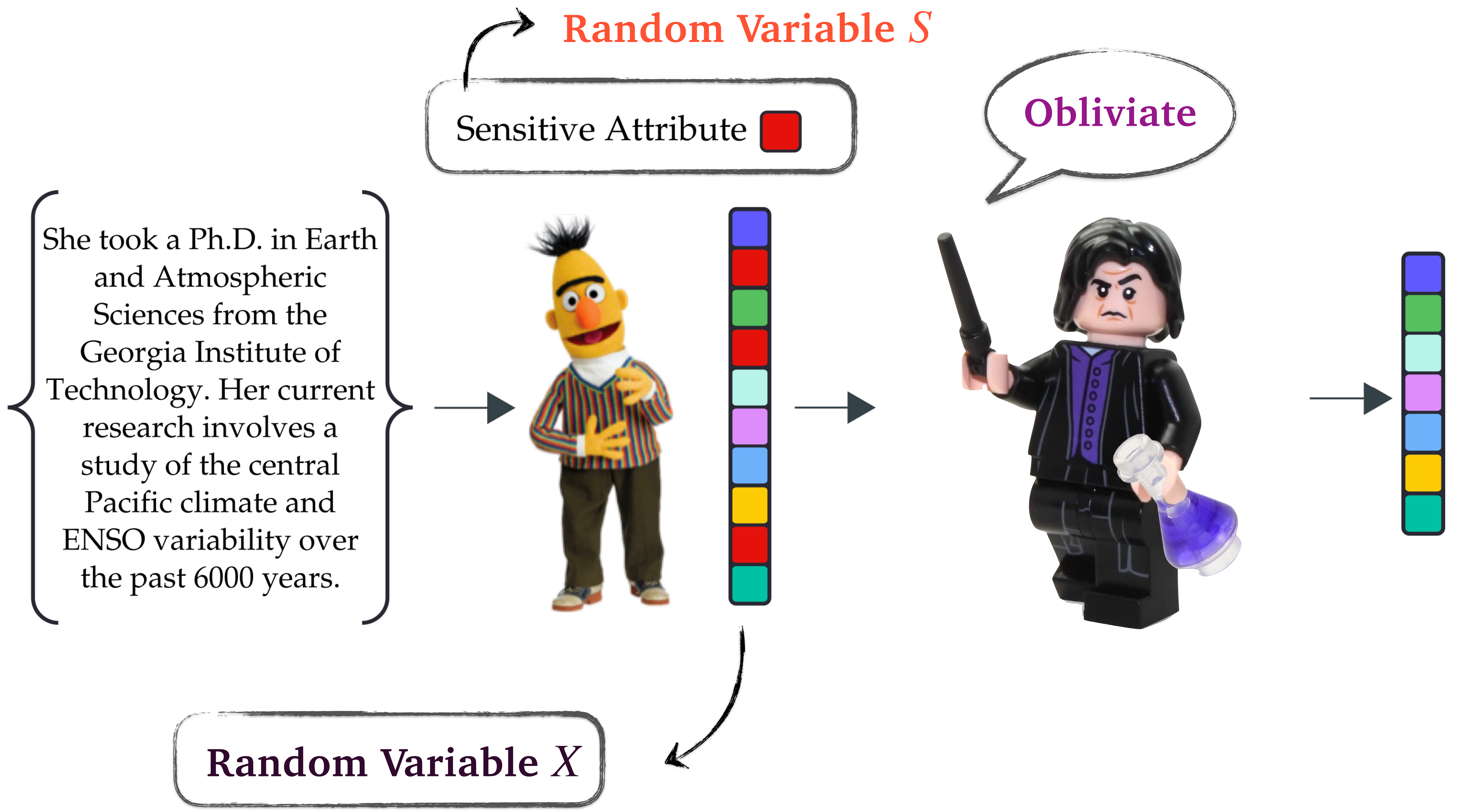
How Do We Formulate Concept Erasure ?



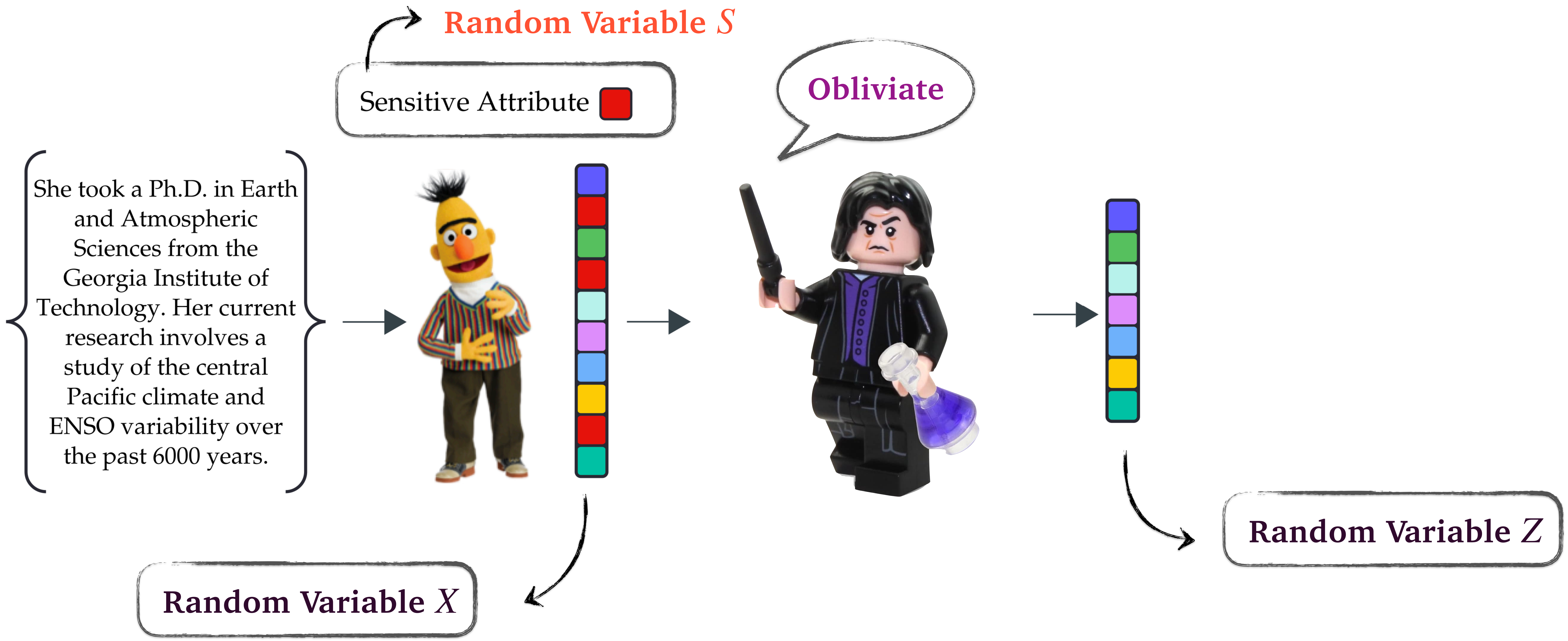
How Do We Formulate Concept Erasure ?



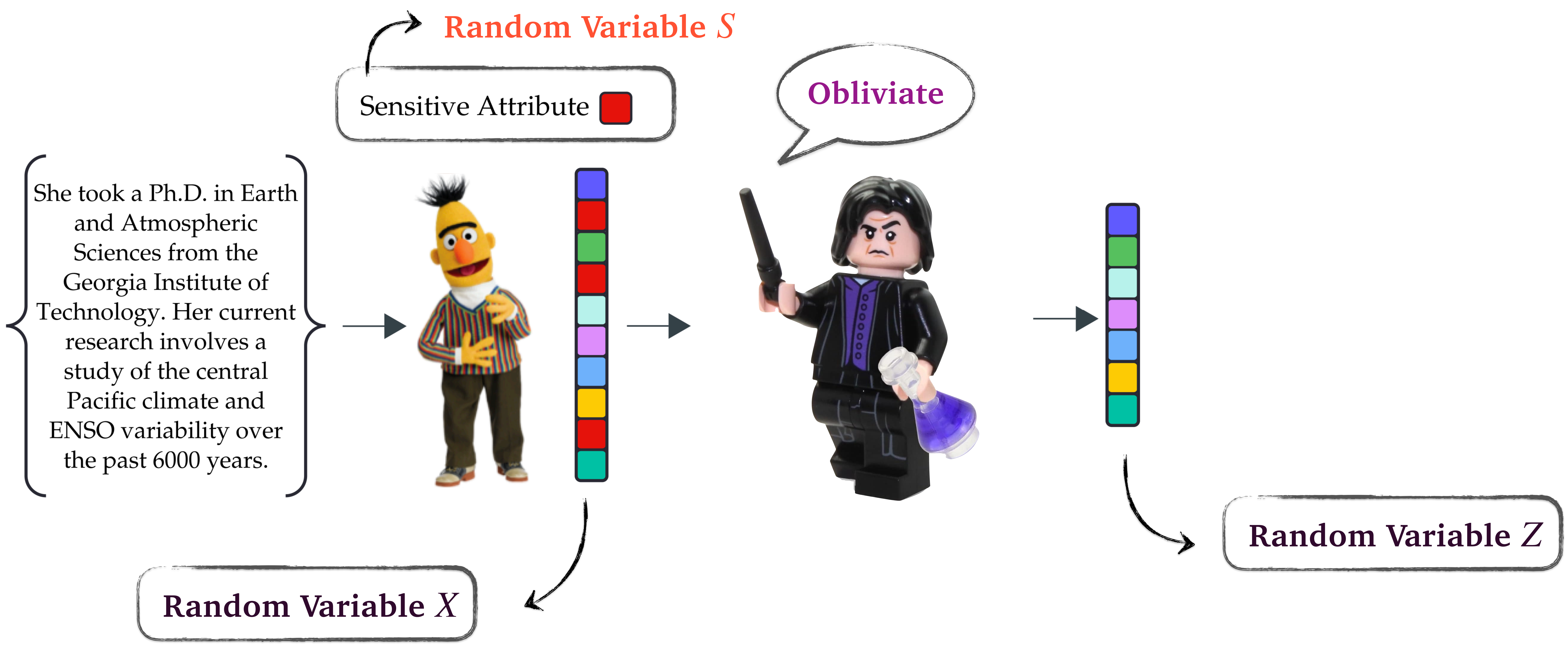
How Do We Formulate Concept Erasure ?



How Do We Formulate Concept Erasure ?



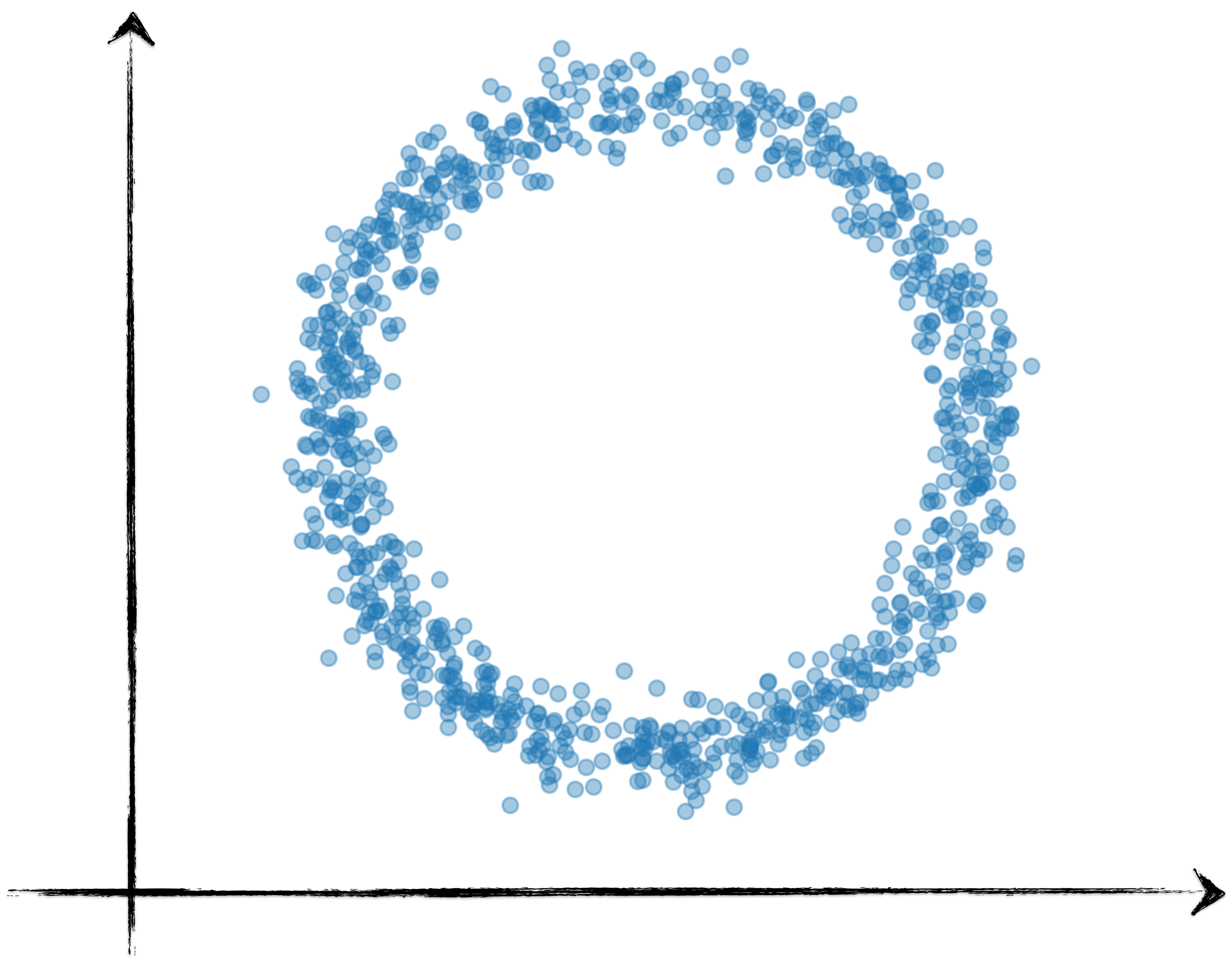
How Do We Formulate Concept Erasure ?



Complete Erasure $\iff Z \perp\!\!\!\perp S$

Why Linear Statistical Dependence Is Not Enough ?

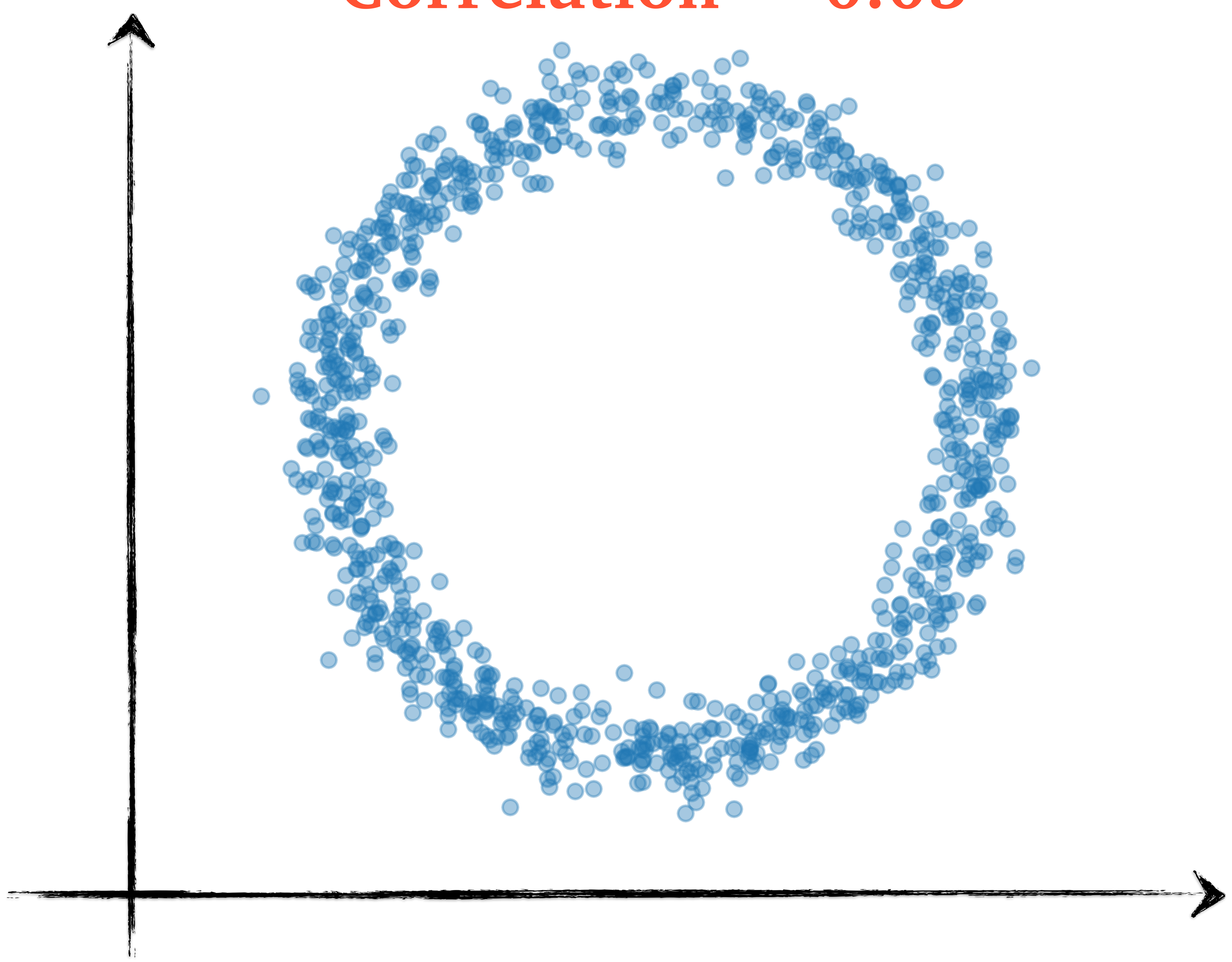
$$\textbf{correlation} = \frac{\textbf{Cov}(X, S)}{\sigma_X \; \sigma_S} = \frac{\mathbb{E}\left[(X - \mu_X)(S - \mu_S)\right]}{\sigma_X \; \sigma_S}$$



Why Linear Statistical Dependence Is Not Enough ?

$$\textbf{correlation} = \frac{\textbf{Cov}(X, S)}{\sigma_X \, \sigma_S} = \frac{\mathbb{E}\left[(X - \mu_X)(S - \mu_S)\right]}{\sigma_X \, \sigma_S}$$

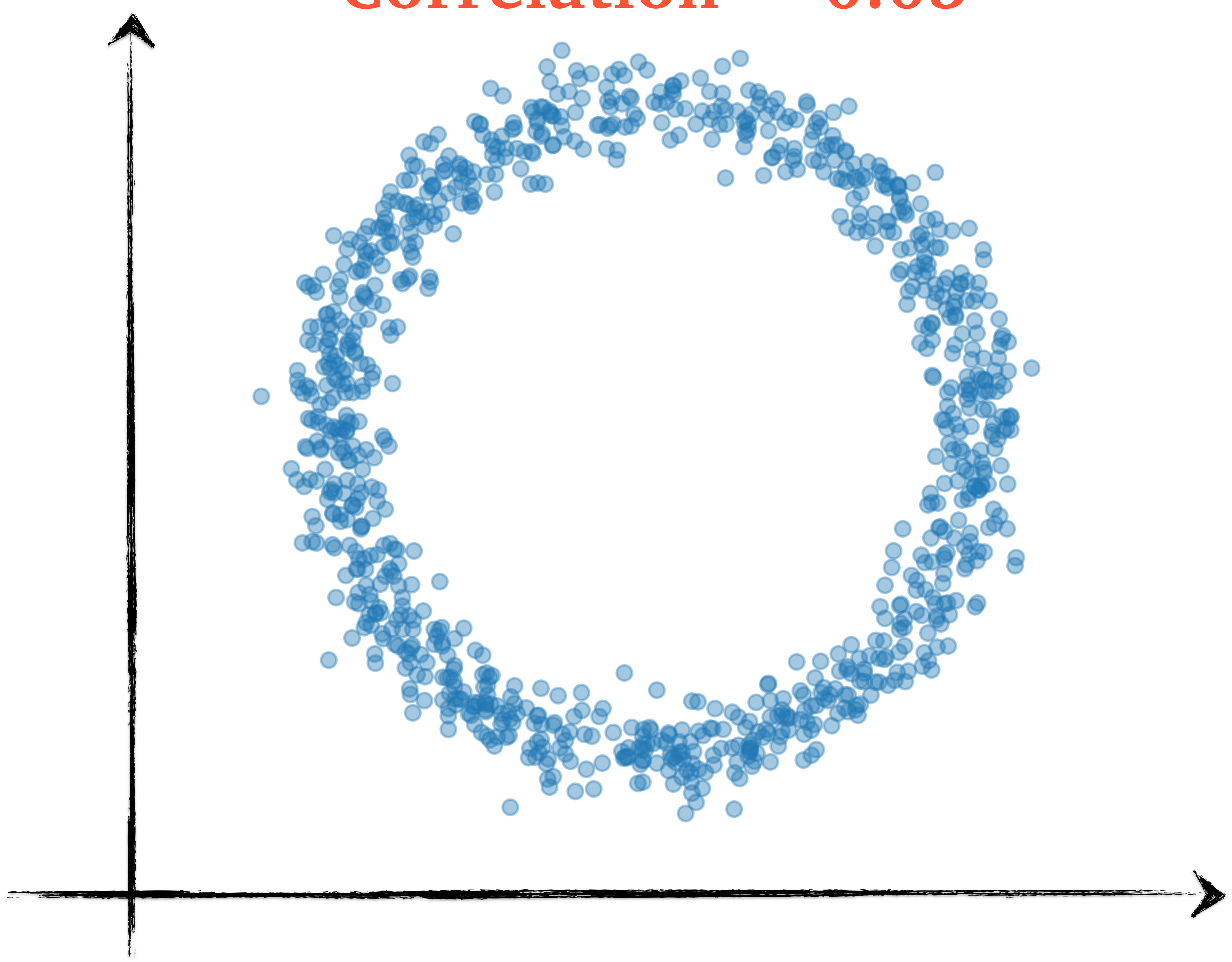
Correlation = 0.03



Why Linear Statistical Dependence Is Not Enough ?

$$\text{correlation} = \frac{\mathbf{Cov}(X, S)}{\sigma_X\ \sigma_S} = \frac{\mathbb{E}\left[(X-\mu_X)(S-\mu_S)\right]}{\sigma_X\ \sigma_S}$$

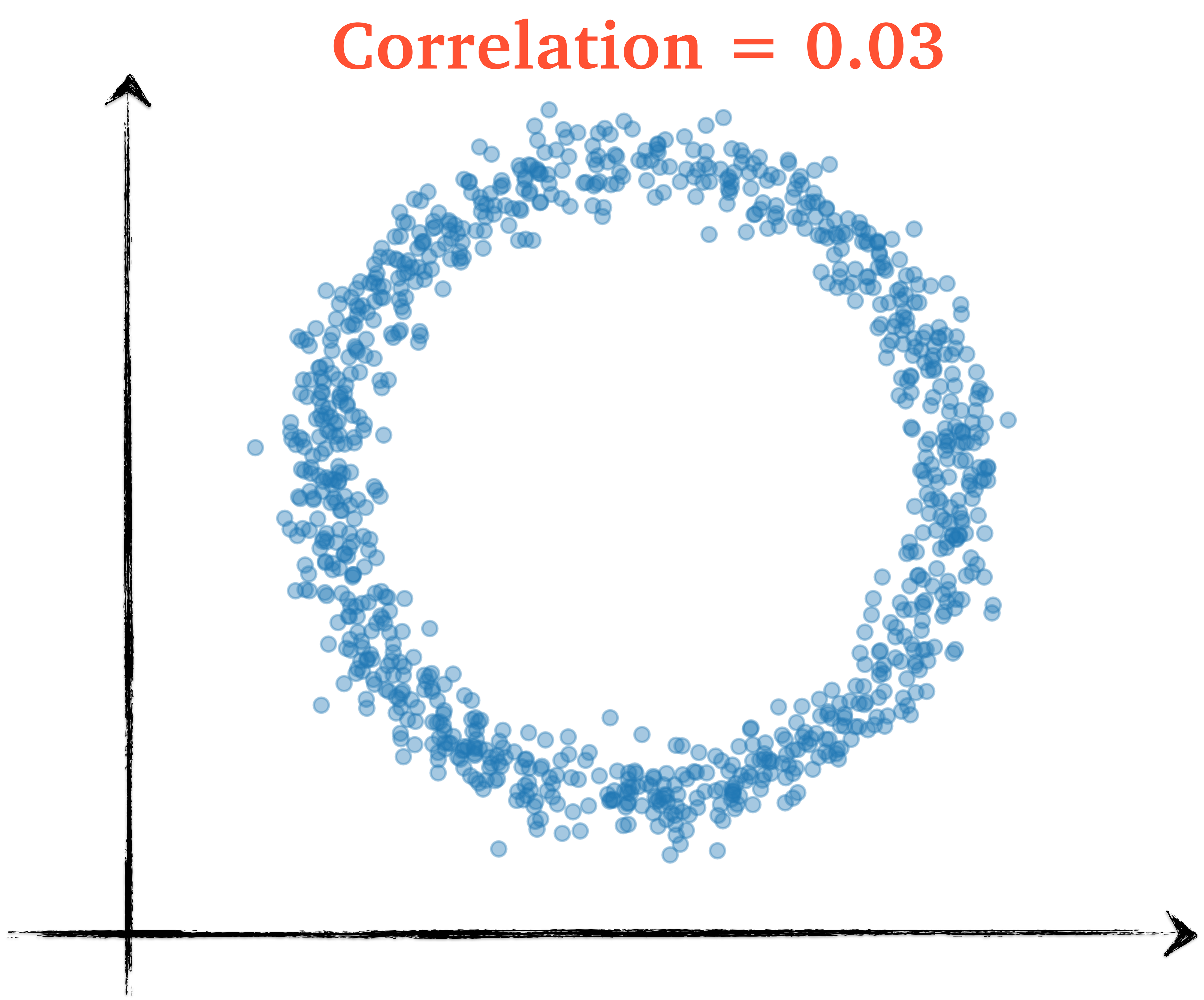
Correlation = 0.03



Independence → Correlation = 0

Why Linear Statistical Dependence Is Not Enough ?

$$\text{correlation} = \frac{\mathbf{Cov}(X, S)}{\sigma_X \, \sigma_S} = \frac{\mathbb{E}\left[(X - \mu_X)(S - \mu_S)\right]}{\sigma_X \, \sigma_S}$$

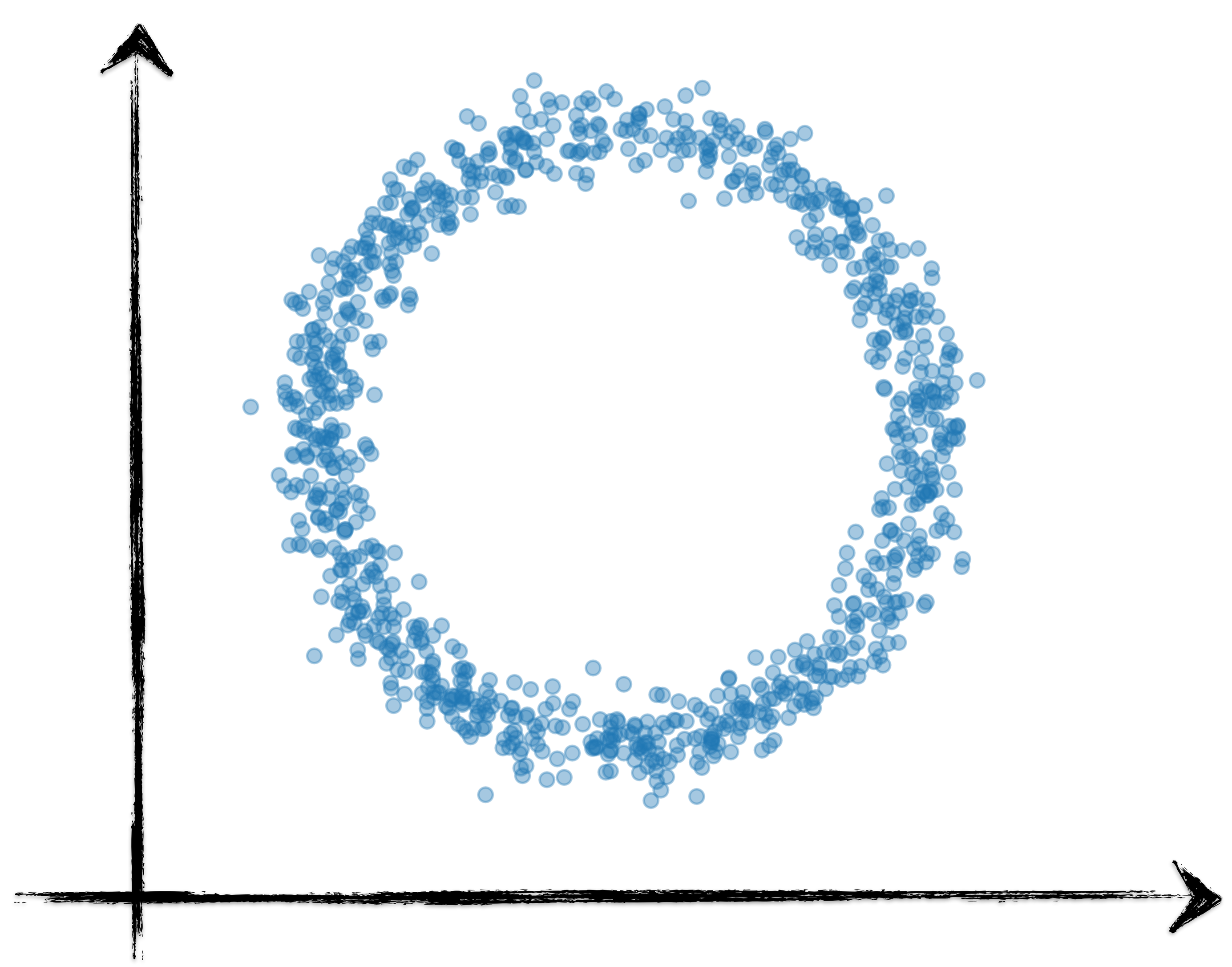


Independence $\rightarrow$ Correlation = 0

Correlation = 0 $\not\rightarrow$ Independence

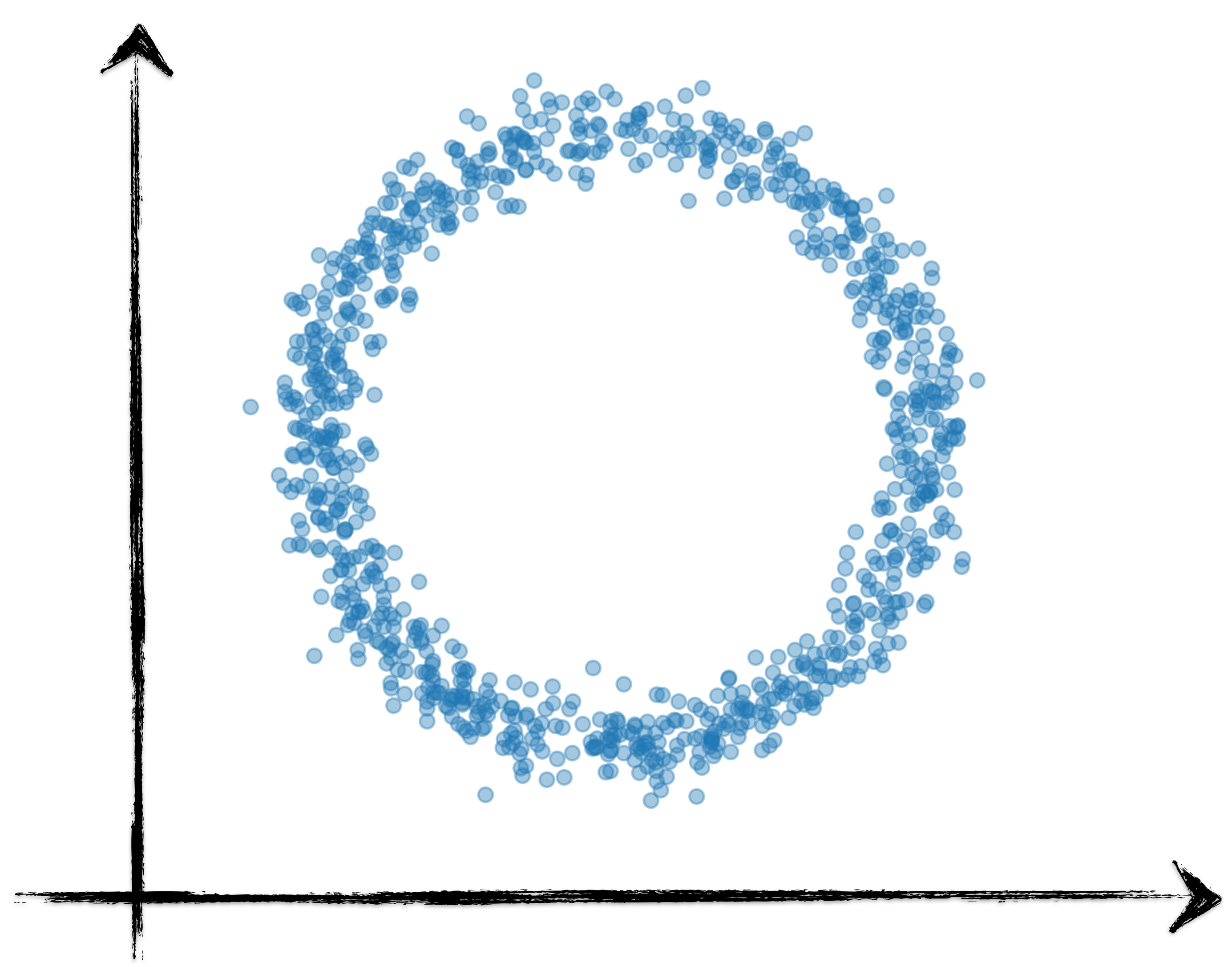
How Do We Capture Nonlinear Statistical Dependency ?

Correlation = 0.03



How Do We Capture Nonlinear Statistical Dependency ?

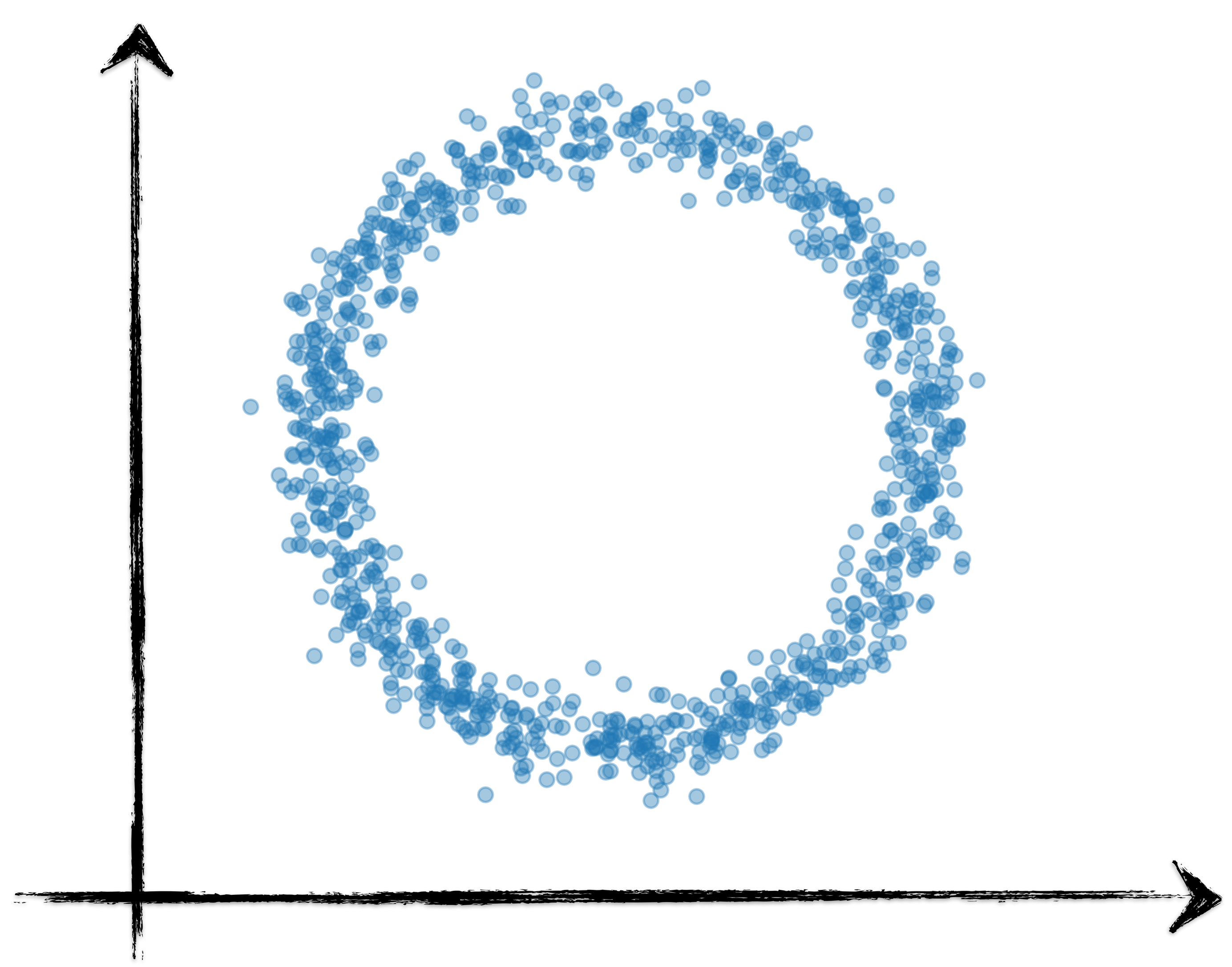
Correlation = 0.03



Transform Nonlinear Dependency between X and S

How Do We Capture Nonlinear Statistical Dependency ?

Correlation = 0.03

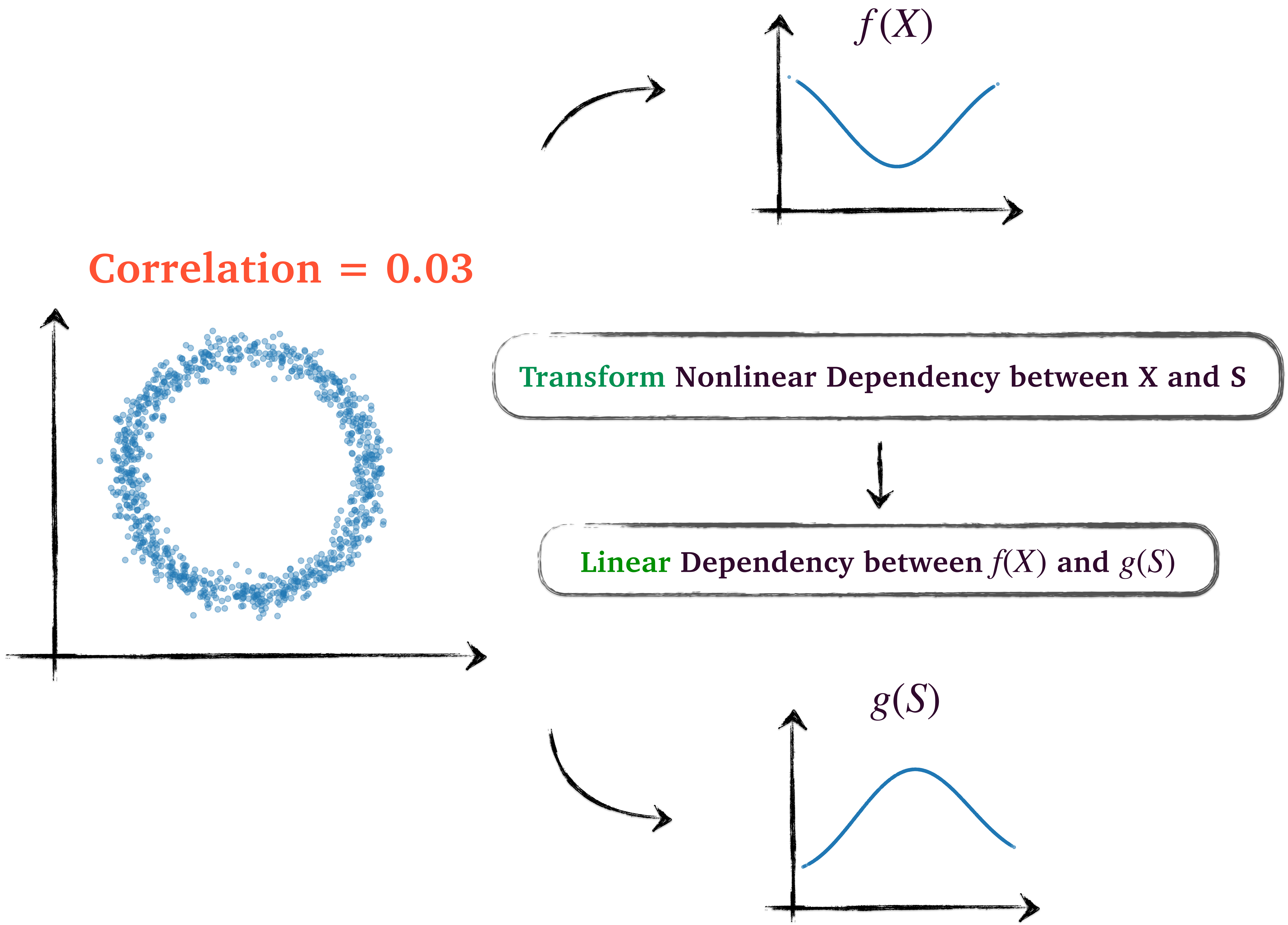


Transform Nonlinear Dependency between X and S

↓

Linear Dependency between $f(X)$ and $g(S)$

How Do We Capture Nonlinear Statistical Dependency ?



How Do We Capture Nonlinear Statistical Dependency ?

How Do We Formalize a Nonlinear Statistical Dependency Test ?

Let $\mathscr{F}$ and $\mathscr{G}$ be a characteristic RKHS

How Do We Formalize a Nonlinear Statistical Dependency Test ?

Let $\mathscr{F}$ and $\mathscr{G}$ be a characteristic RKHS

Regularization Constraint

$$\sup_{g \in \mathscr{G}} \sup_{f \in \mathscr{F}} \mathbf{Cov}(f(X), g(S)) = \langle g, \mathbb{C}\mathbf{ov} f \rangle_{\mathscr{G}} \qquad s.t \qquad \|f\|_{\mathscr{F}} = \|g\|_{\mathscr{G}} = 1$$

How Do We Formalize a Nonlinear Statistical Dependency Test ?

Let $\mathscr{F}$ and $\mathscr{G}$ be a characteristic RKHS

Regularization Constraint

$$\sup_{g \in \mathscr{G}} \sup_{f \in \mathscr{F}} \mathbf{Cov}(f(X), g(S)) = \langle g, \mathbb{C}\mathbf{ov} f \rangle_{\mathscr{G}} \quad s.t \quad \|f\|_{\mathscr{F}} = \|g\|_{\mathscr{G}} = 1$$

$\|\mathbb{C}\mathbf{ov}\|^2_{HS}$ Hilbert-Schmidt Norm of Covariance Operator Indicates Existence of Such Functions

How Do We Formalize a Nonlinear Statistical Dependency Test ?

Let $\mathscr{F}$ and $\mathscr{G}$ be a characteristic RKHS

Regularization Constraint

$$\sup_{g \in \mathscr{G}} \sup_{f \in \mathscr{F}} \mathbf{Cov}(f(X), g(S)) = \langle g, \mathbb{C}\mathbf{ov} f \rangle_{\mathscr{G}} \quad s.t \quad \|f\|_{\mathscr{F}} = \|g\|_{\mathscr{G}} = 1$$

$\|\mathbb{C}\text{ov}\|_{HS}^{2}$ Hilbert-Schmidt Norm of Covariance Operator Indicates Existence of Such Functions

$\|\mathbb{C}\text{ov}\|_{HS}^{2} = 0 \iff$ Statistical Independence

How Do We Formalize a Nonlinear Statistical Dependency Test ?

Let $\mathscr{F}$ and $\mathscr{G}$ be a characteristic RKHS

Regularization Constraint

$$\sup_{g \in \mathscr{G}} \sup_{f \in \mathscr{F}} \mathbf{Cov}(f(X), g(S)) = \langle g, \mathbb{C}\mathbf{ov} f \rangle_{\mathscr{G}} \quad s.t \quad \|f\|_{\mathscr{F}} = \|g\|_{\mathscr{G}} = 1$$

$\|\mathbb{C}\mathrm{ov}\|_{HS}^2$ Hilbert-Schmidt Norm of Covariance Operator Indicates Existence of Such Functions

$\|\mathbb{C}\mathrm{ov}\|_{HS}^2 = 0 \iff$ Statistical Independence

The Correct Proxy To Impose Statistical Independence

How Do We Formulate Erasure Using Statistical Independence ?

$$Z_\theta = \varepsilon(X; \boldsymbol{\theta}) \quad \rightarrow$$

How Do We Formulate Erasure Using Statistical Independence ?

$$Z_\theta = \varepsilon(X; \boldsymbol{\theta}) \quad \rightarrow$$

Z_θ → Random Variable After Erasure

S → Random Variable Represent Unwanted Label

Y → Random Variable Represent Utility Label

How Do We Formulate Erasure Using Statistical Independence ?

$$Z_\theta = \varepsilon(X; \boldsymbol{\theta}) \quad \rightarrow$$

$Z_\theta \quad \rightarrow$ **Random Variable After Erasure**

$S \quad \rightarrow$ **Random Variable Represent Unwanted Label**

$Y \quad \rightarrow$ **Random Variable Represent Utility Label**

$$\inf_\theta \quad \mathbf{HSIC}(Z_\theta, S) - \mathbf{HSIC}(Z_\theta, Y)$$

Minimize Statistical Dependency

A Proxy to Preserve Utility Information

How Do We Formulate Erasure Using Statistical Independence ?

$$Z_\theta = \varepsilon(X; \boldsymbol{\theta}) \quad \rightarrow$$

$Z_\theta \quad \rightarrow$ Random Variable After Erasure

$S \quad \rightarrow$ Random Variable Represent Unwanted Label

$Y \quad \rightarrow$ Random Variable Represent Utility Label

$$\inf_\theta \quad \mathbf{HSIC}(Z_\theta, S) - \mathbf{HSIC}(Z_\theta, Y)$$

Conflicting Objectives $\equiv$ Challenging Optimization

Minimize Statistical Dependency

A Proxy to Preserve Utility Information

How Do We Solve This Optimization ?

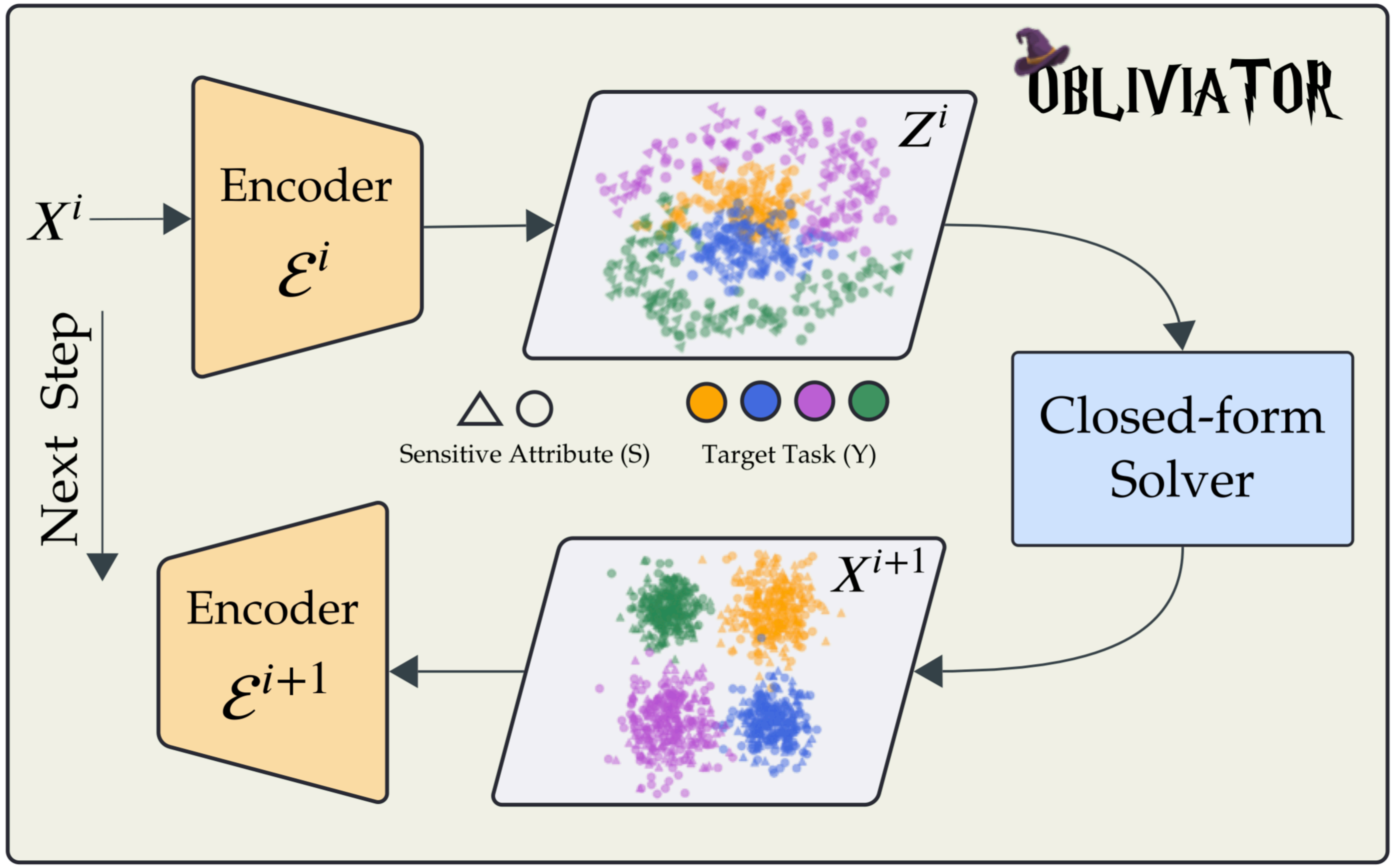
Single-Shot Optimization Leads to Poor Solutions

How Do We Solve This Optimization ?

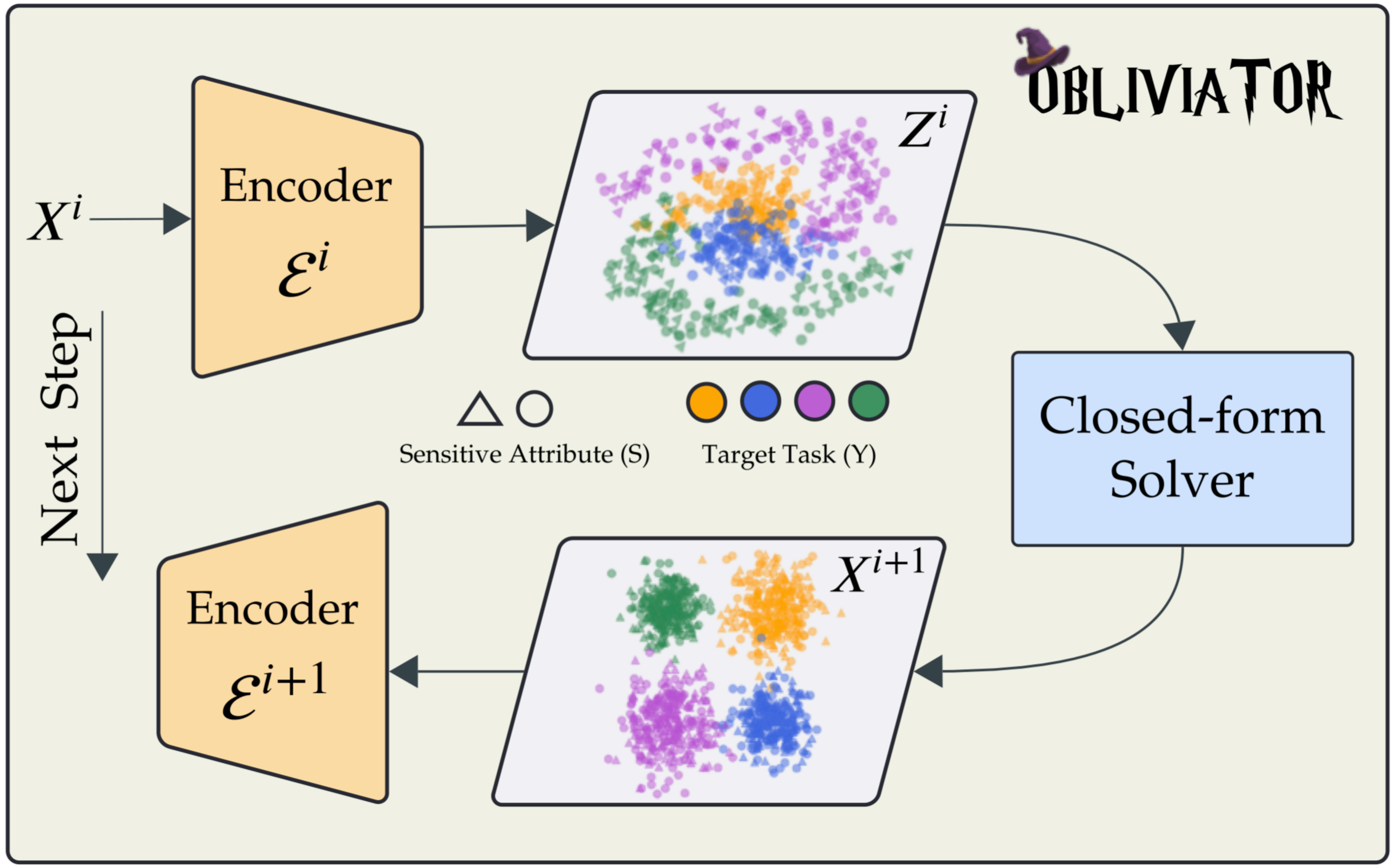
Single-Shot Optimization Leads to Poor Solutions

To Make Erasure Smoother We Propose an Iterative approach

How Do We Solve This Optimization ?

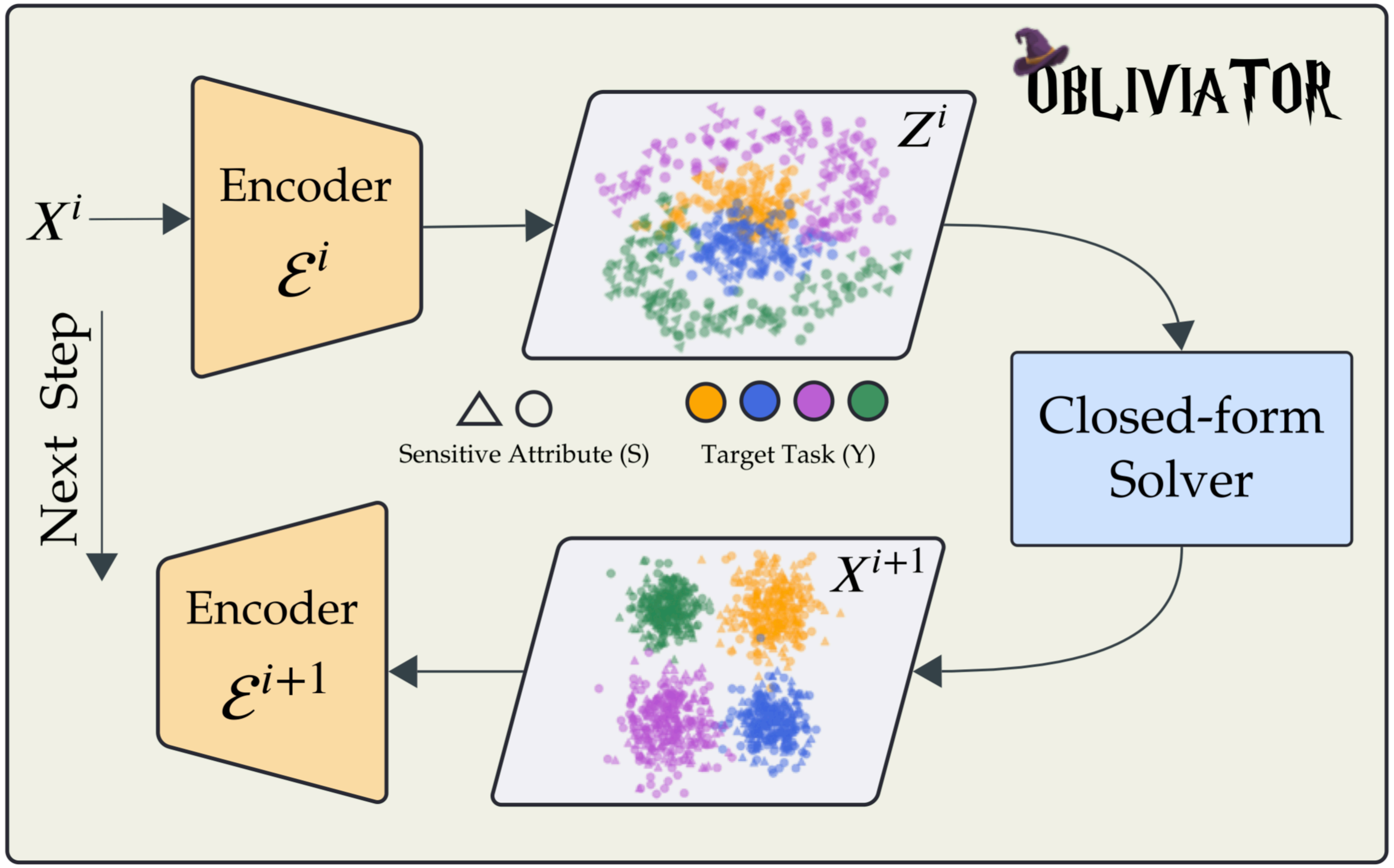


How Do We Solve This Optimization ?



Step One

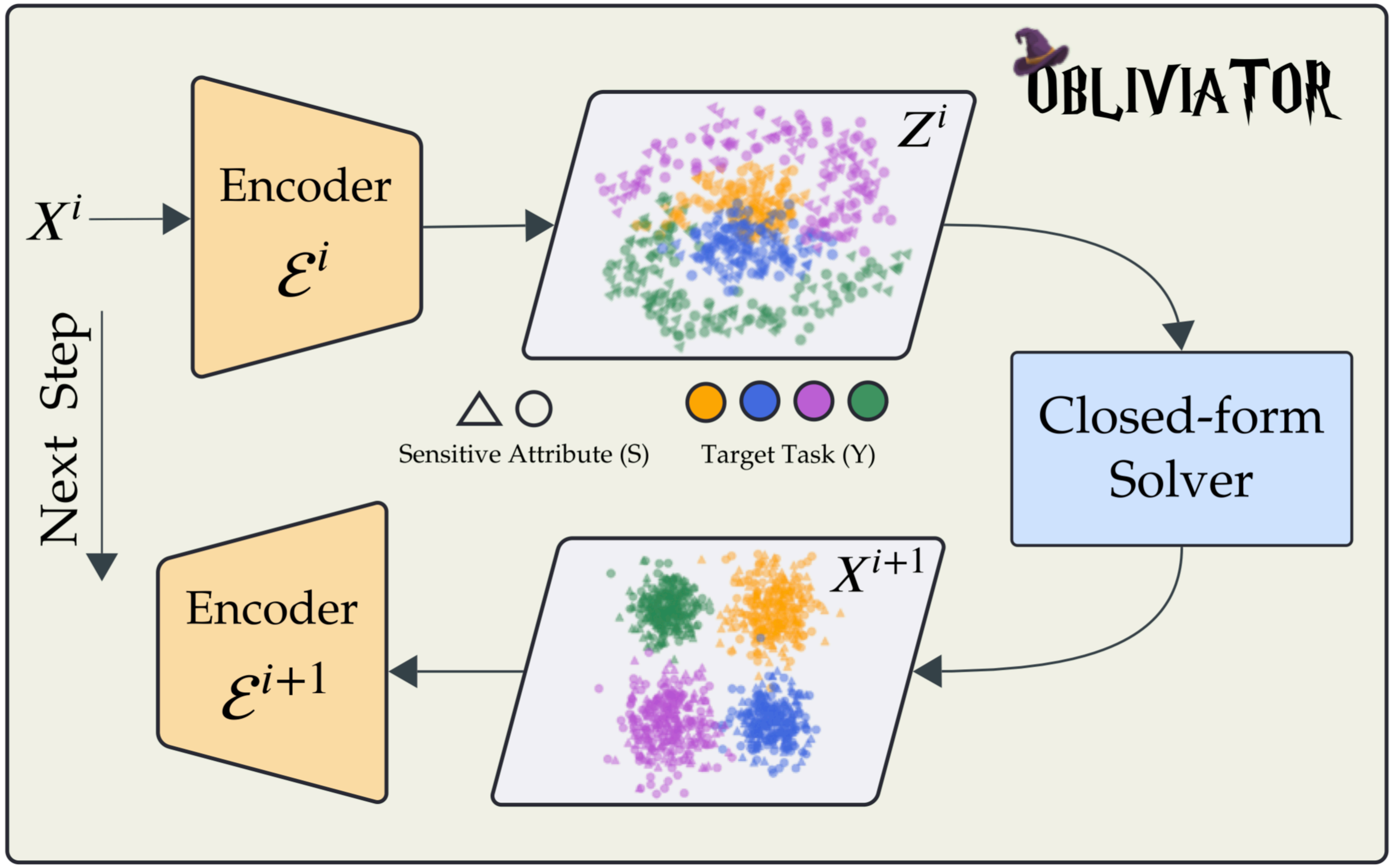
How Do We Solve This Optimization ?



Step One

Erase and Preserve via guarding function $Z^i_\theta = \varepsilon^i(X^i; \boldsymbol{\theta}^i)$

How Do We Solve This Optimization ?

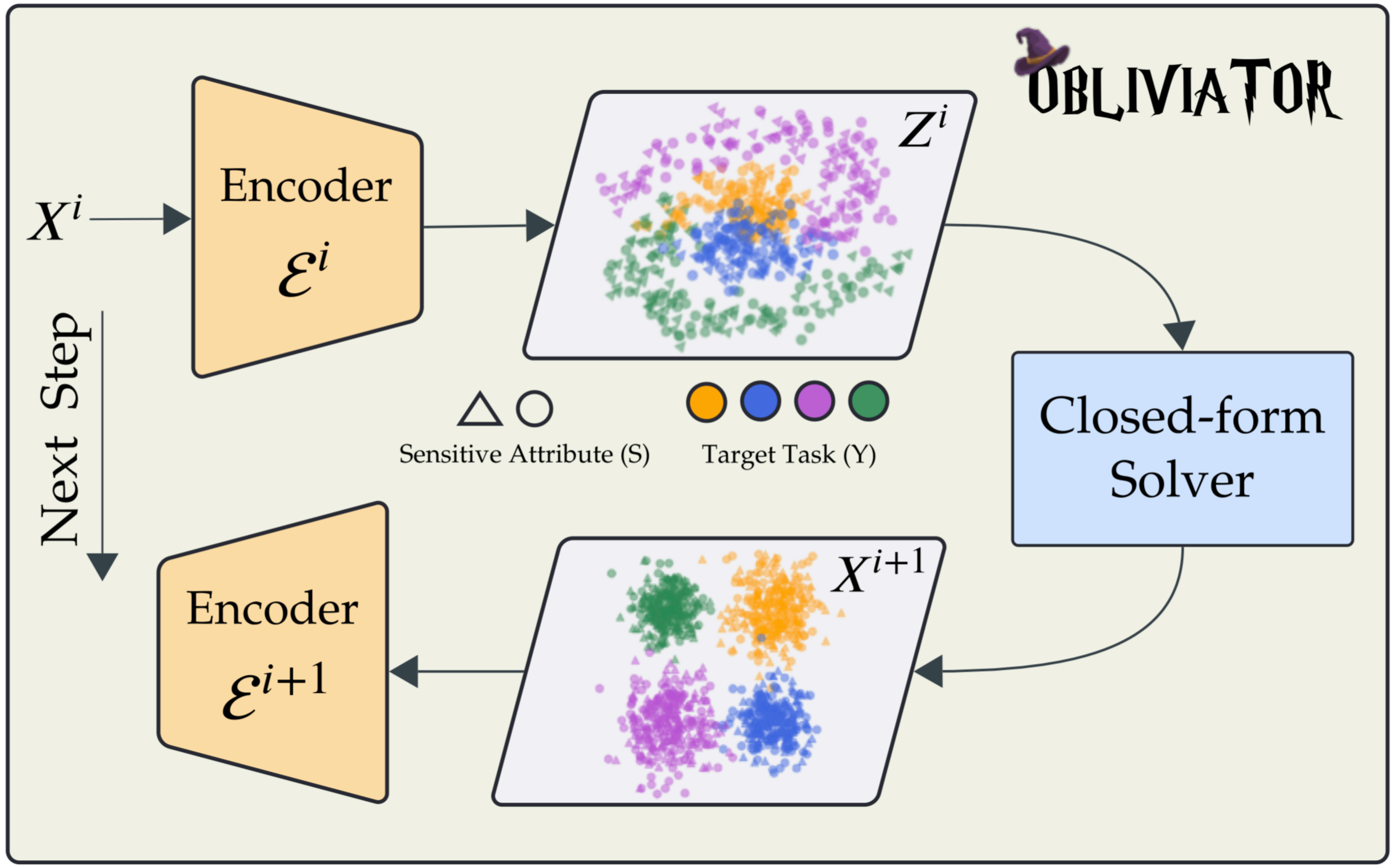


Step One

Erase and Preserve via guarding function $Z^i_\theta = \varepsilon^i(X^i; \boldsymbol{\theta}^i)$

$$\inf_{\theta} \quad \mathbf{HSIC}(Z_\theta, S) - \left[\mathbf{HSIC}(Z^i_\theta, Y) + \mathbf{HSIC}(Z^i_\theta, X) + \mathbf{HSIC}(Z^i_\theta, X^i)\right]$$

How Do We Solve This Optimization ?



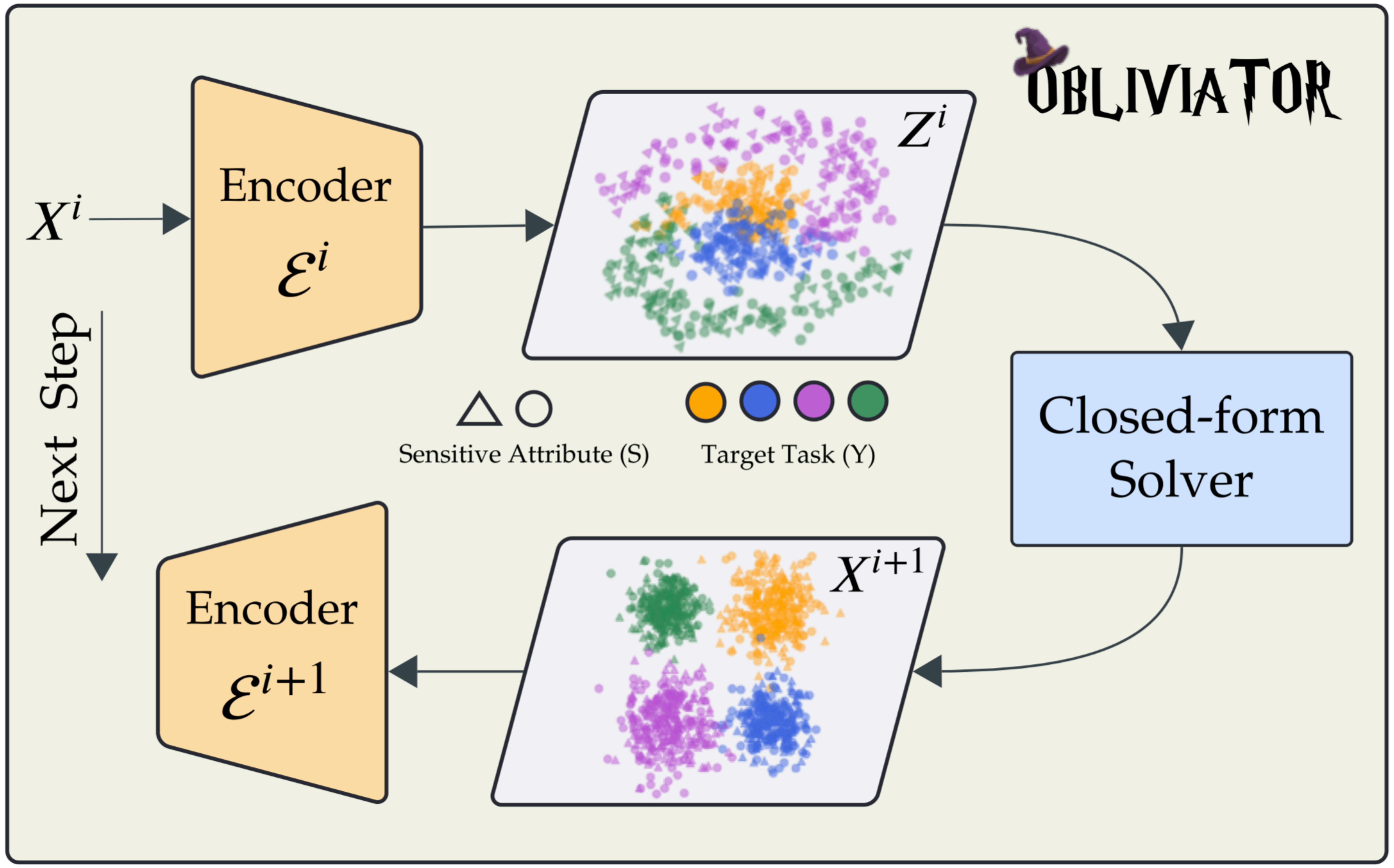
Step One

Erase and Preserve via guarding function $Z^i_\theta = \varepsilon^i(X^i; \boldsymbol{\theta}^i)$

$$\inf_\theta \quad \mathbf{HSIC}(Z_\theta, S) - \boxed{\left[\mathbf{HSIC}(Z^i_\theta, Y) + \mathbf{HSIC}(Z^i_\theta, X) + \mathbf{HSIC}(Z^i_\theta, X^i)\right]}$$

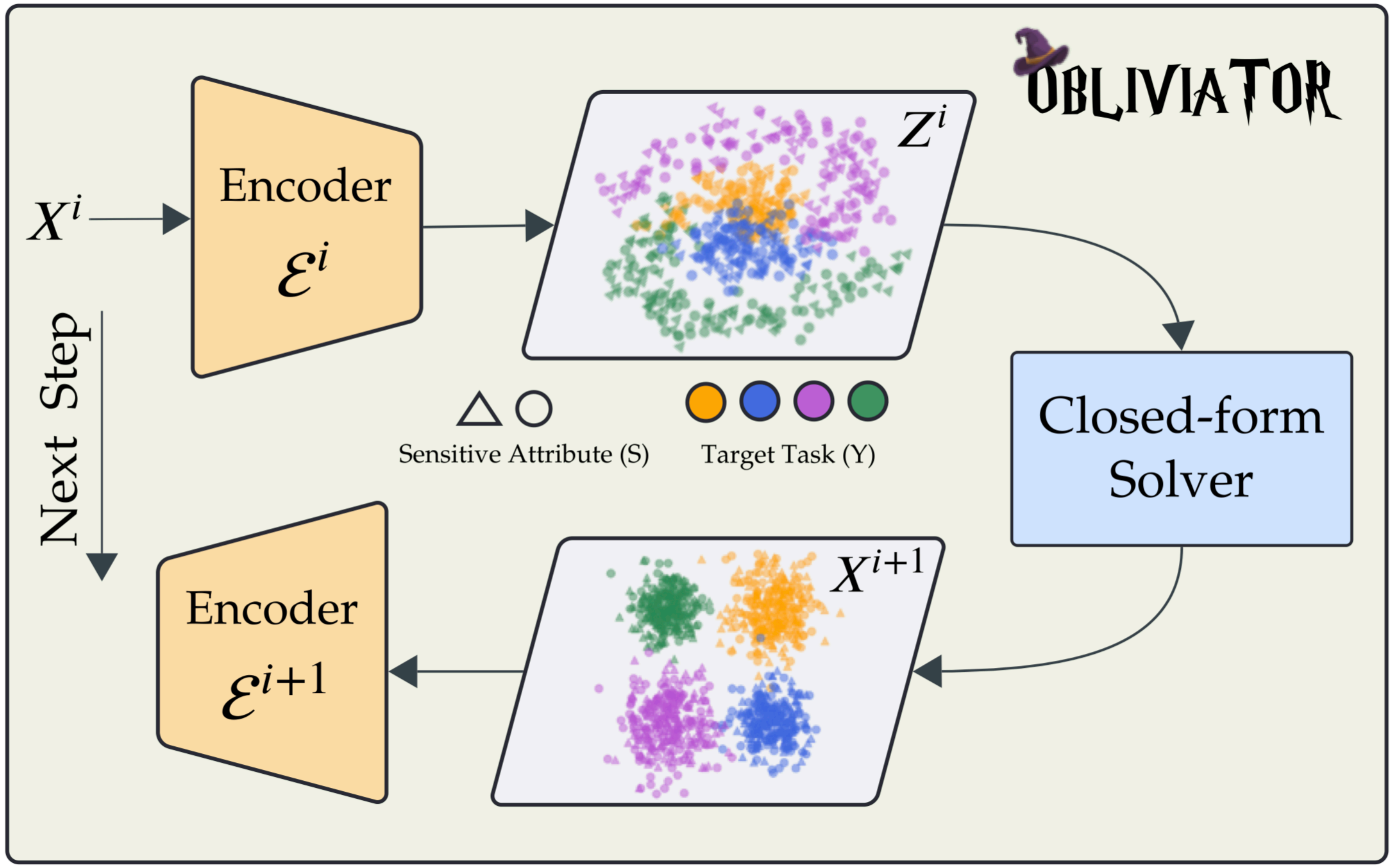
More Information-Preserving Loss

How Do We Solve This Optimization ?



Step Two

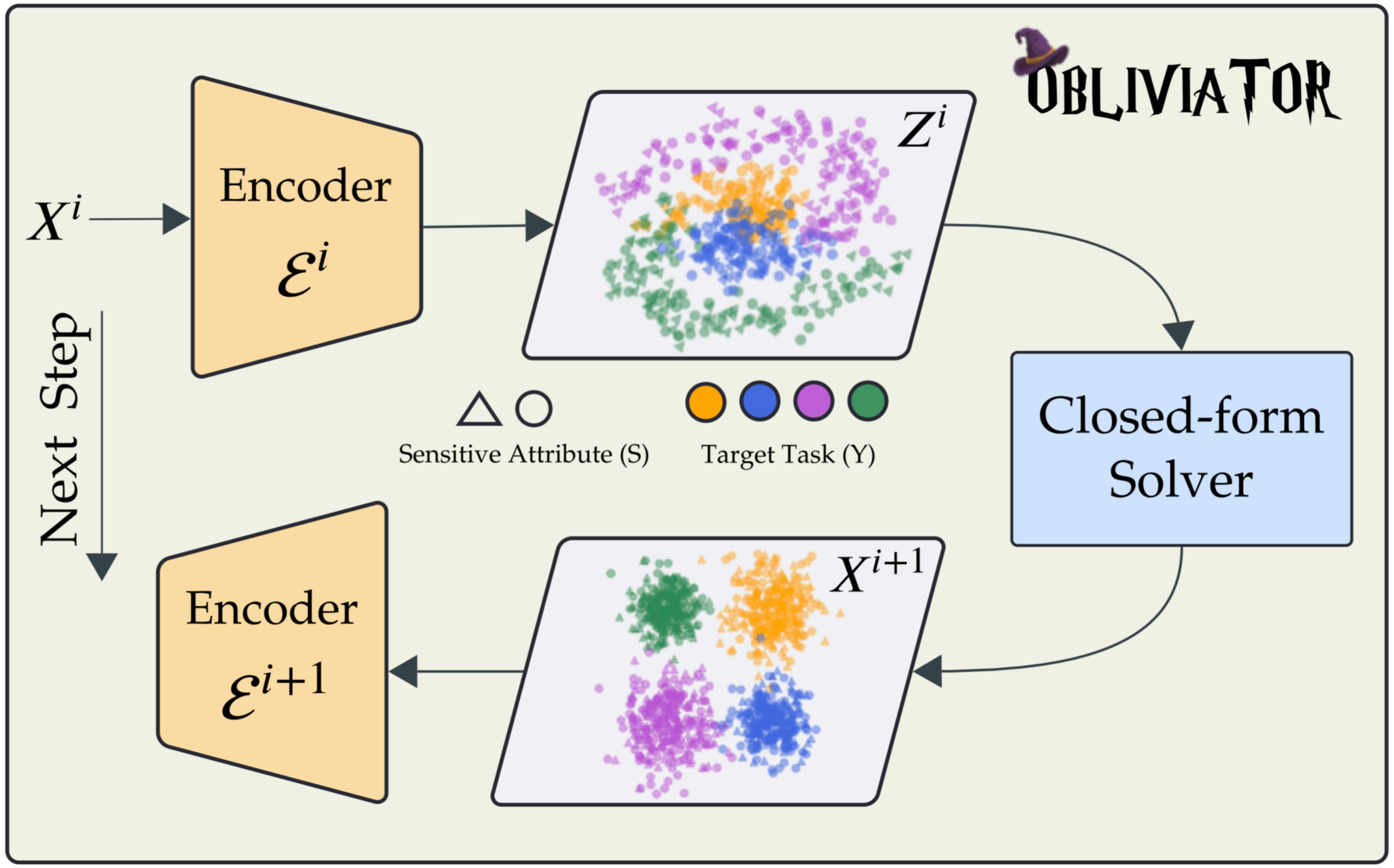
How Do We Solve This Optimization ?



Step Two

Separate Useful Information from Unwanted Concepts:

How Do We Solve This Optimization ?



Step Two

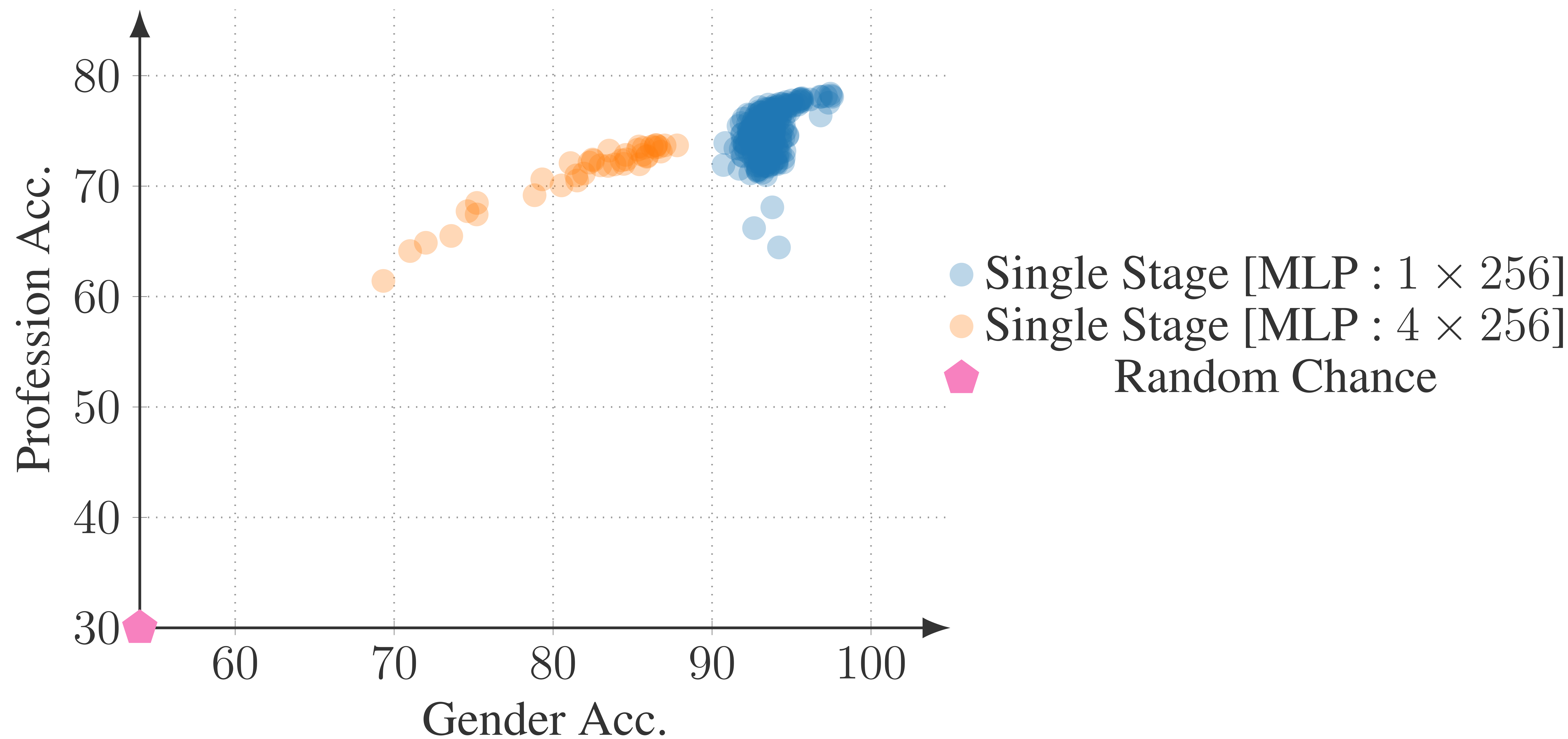
Separate Useful Information from Unwanted Concepts:

$$\sup_{\{g_a \in \mathcal{G}_a\}} \sup_{f \in \mathcal{F}} \mathbf{Cov}^2(f(Z^i_\theta), g_y(Y)) + \mathbf{Cov}^2(f(Z^i_\theta), g_{x^i}(X^i)) + \mathbf{Cov}^2(f(Z^i_\theta), g_x(X))$$

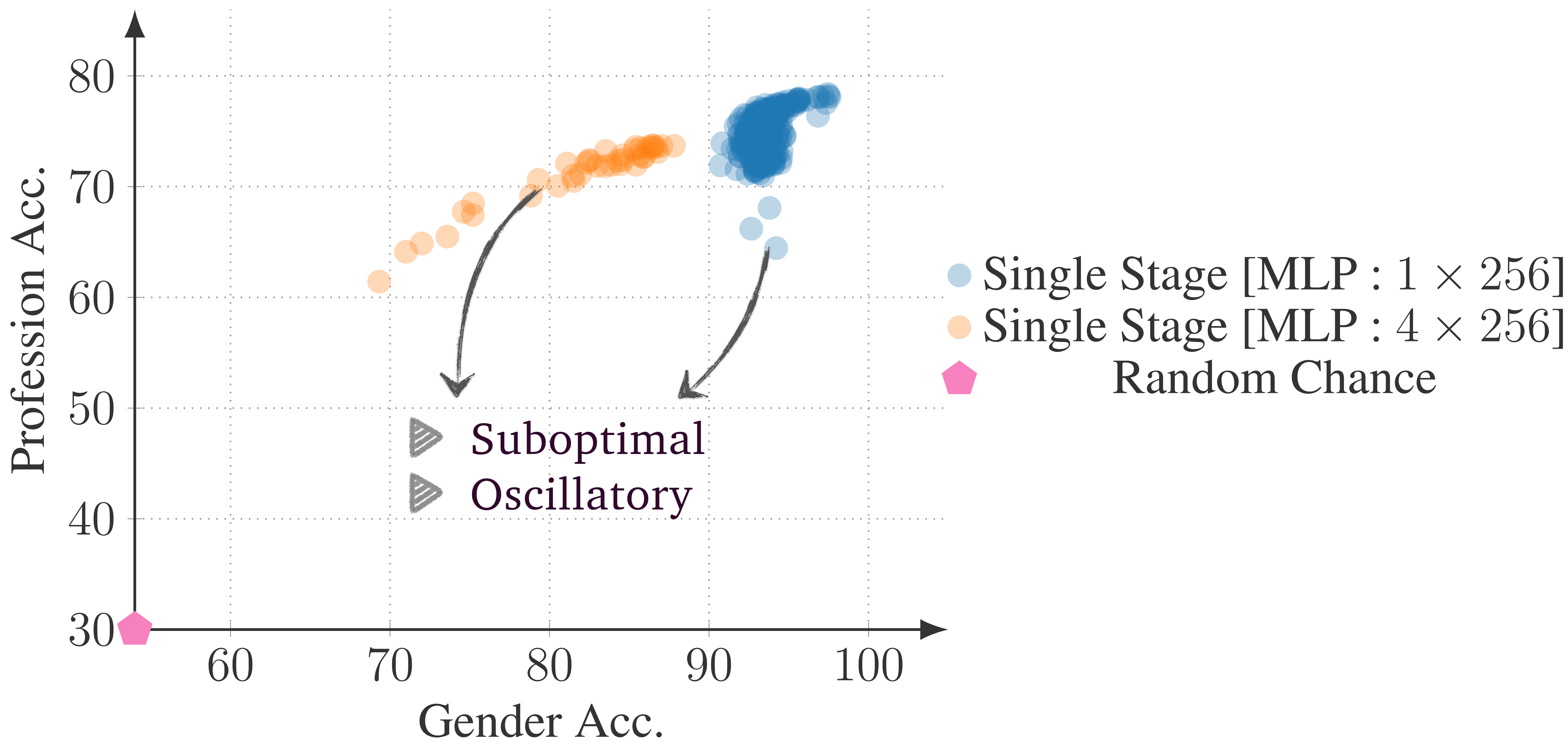
$$s.t \quad \sup_{g_s \in \mathcal{G}_s} \mathbf{Cov}(f(Z^i_\theta), g_s(S)) = 0 \quad \|f\|_{\mathcal{F}} = \|g_a\|_{\mathcal{G}_a} = 1$$

Single Stage Vs. Iterative Erasure

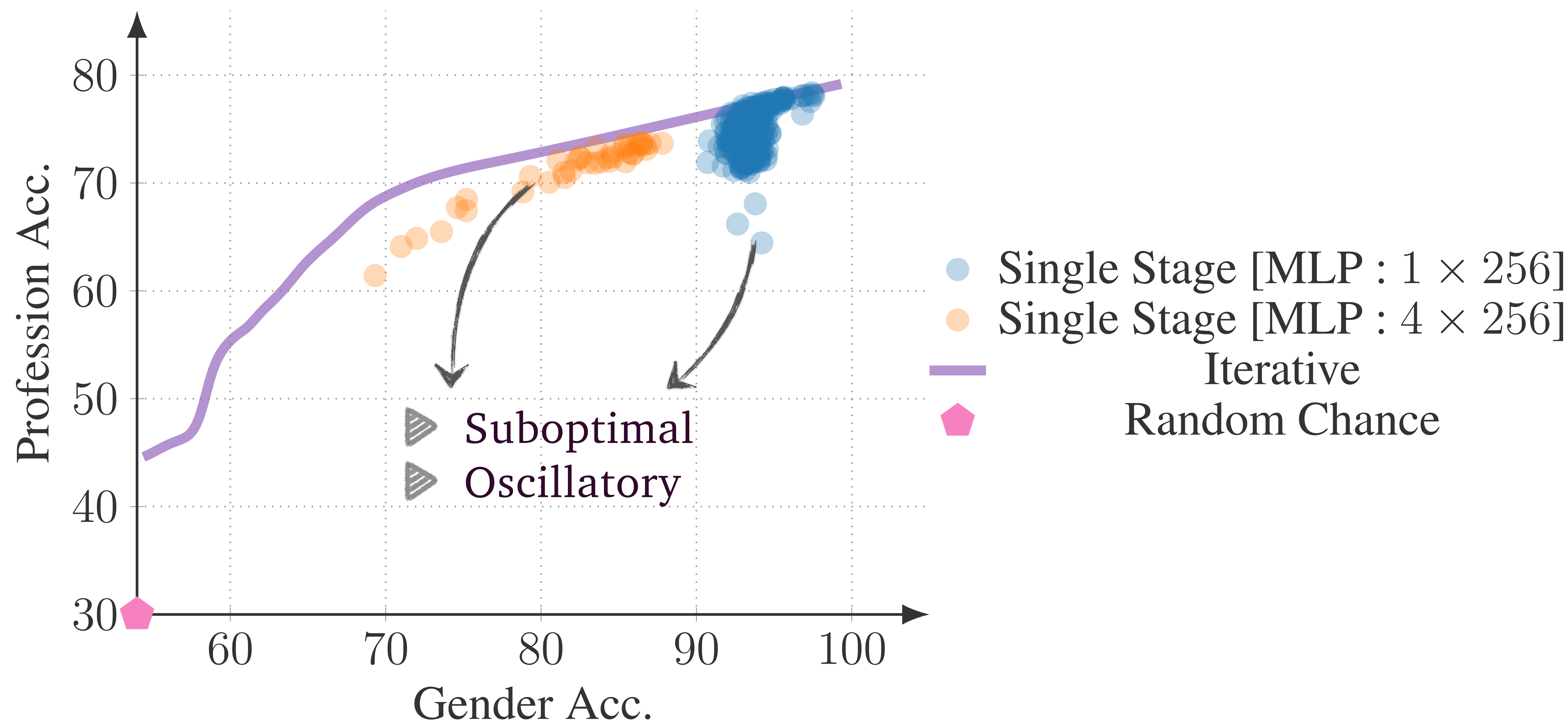
Single Stage Vs. Iterative Erasure



Single Stage Vs. Iterative Erasure

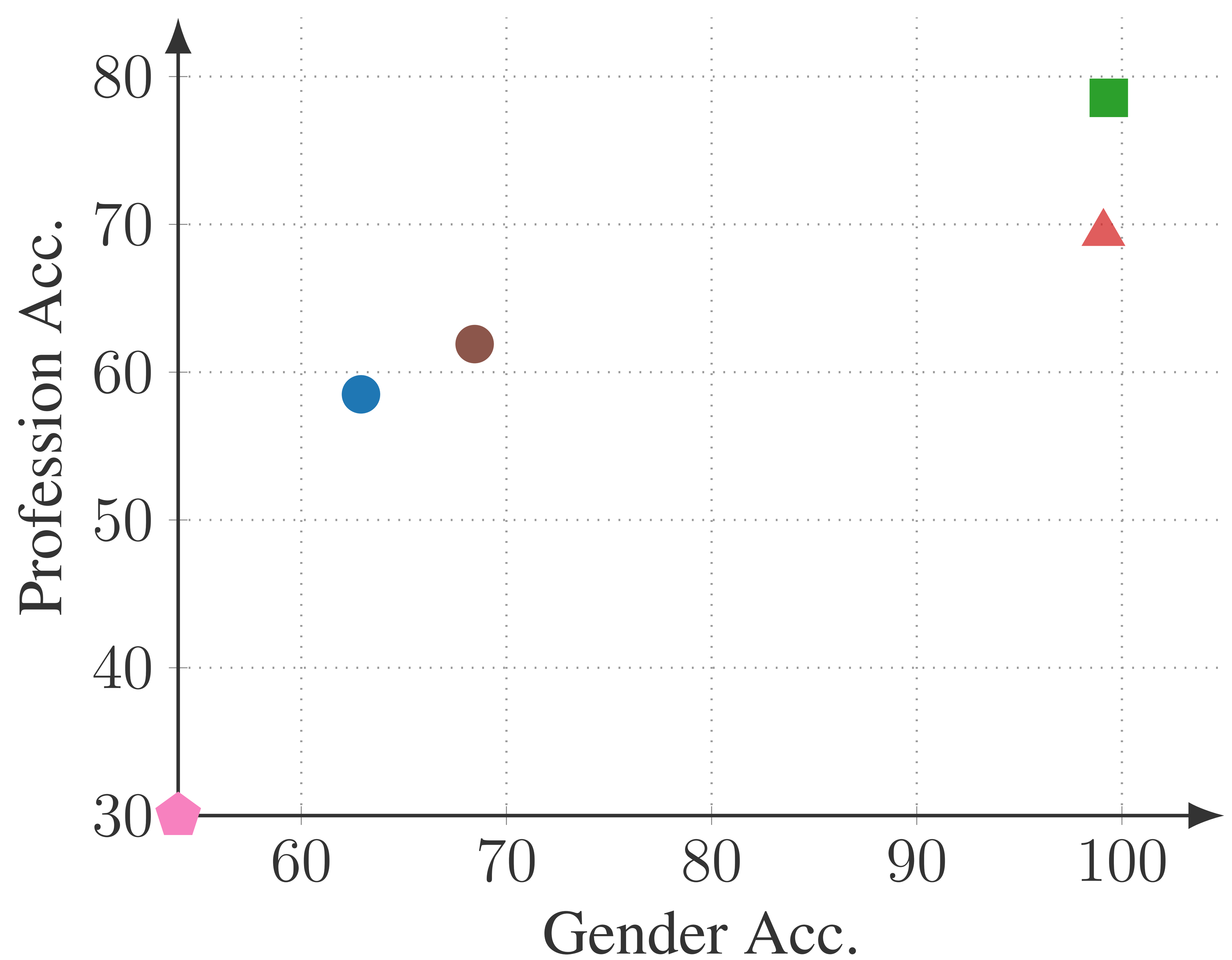


Single Stage Vs. Iterative Erasure

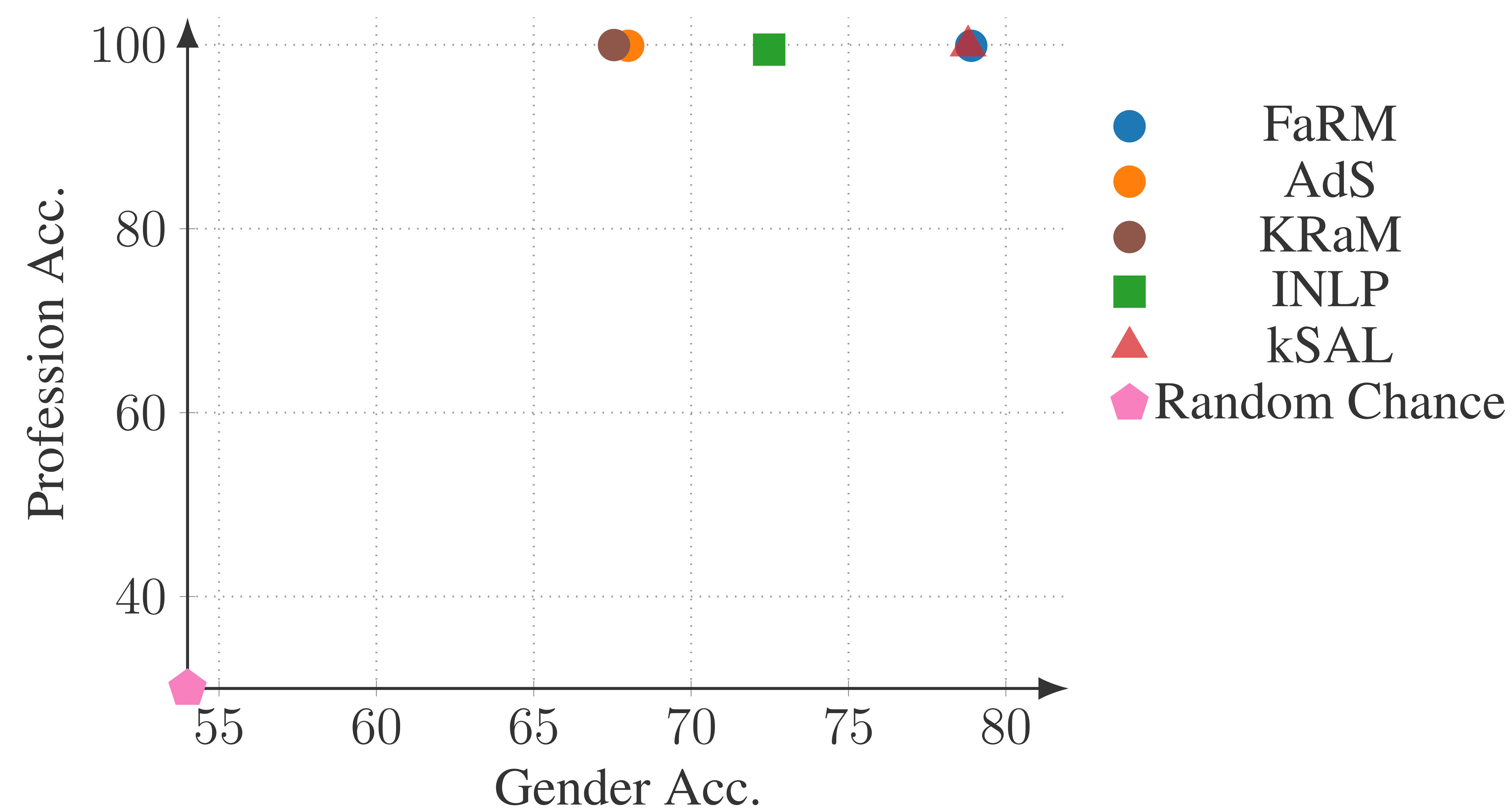


What does Obliviator Reveal about Concept Erasure ?

Obliviator Reveals : Achieving Nonlinear Guardedness



Bias in Bios



Bias in Bios

Representation: Frozen

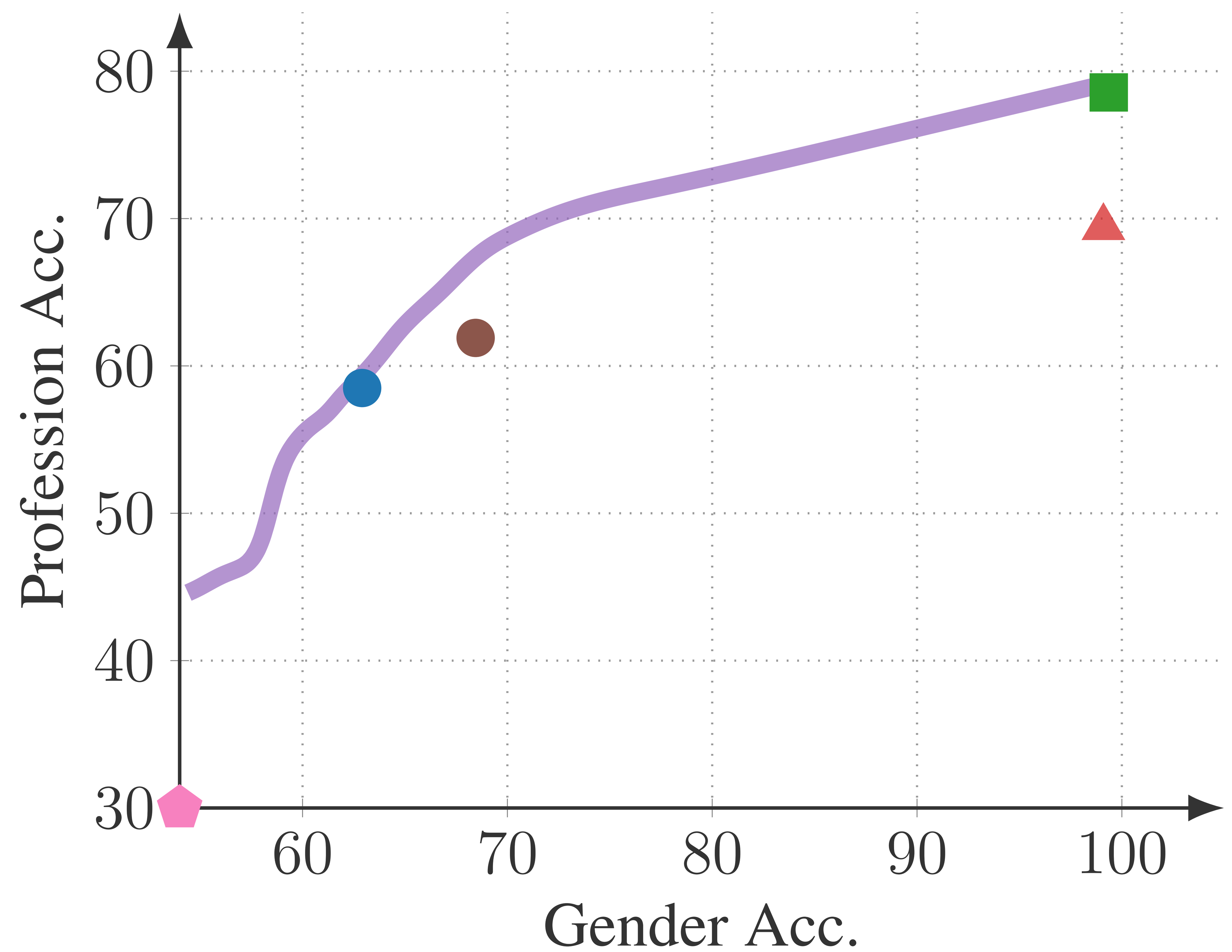
Representation: Finetuned

PLM : BERT

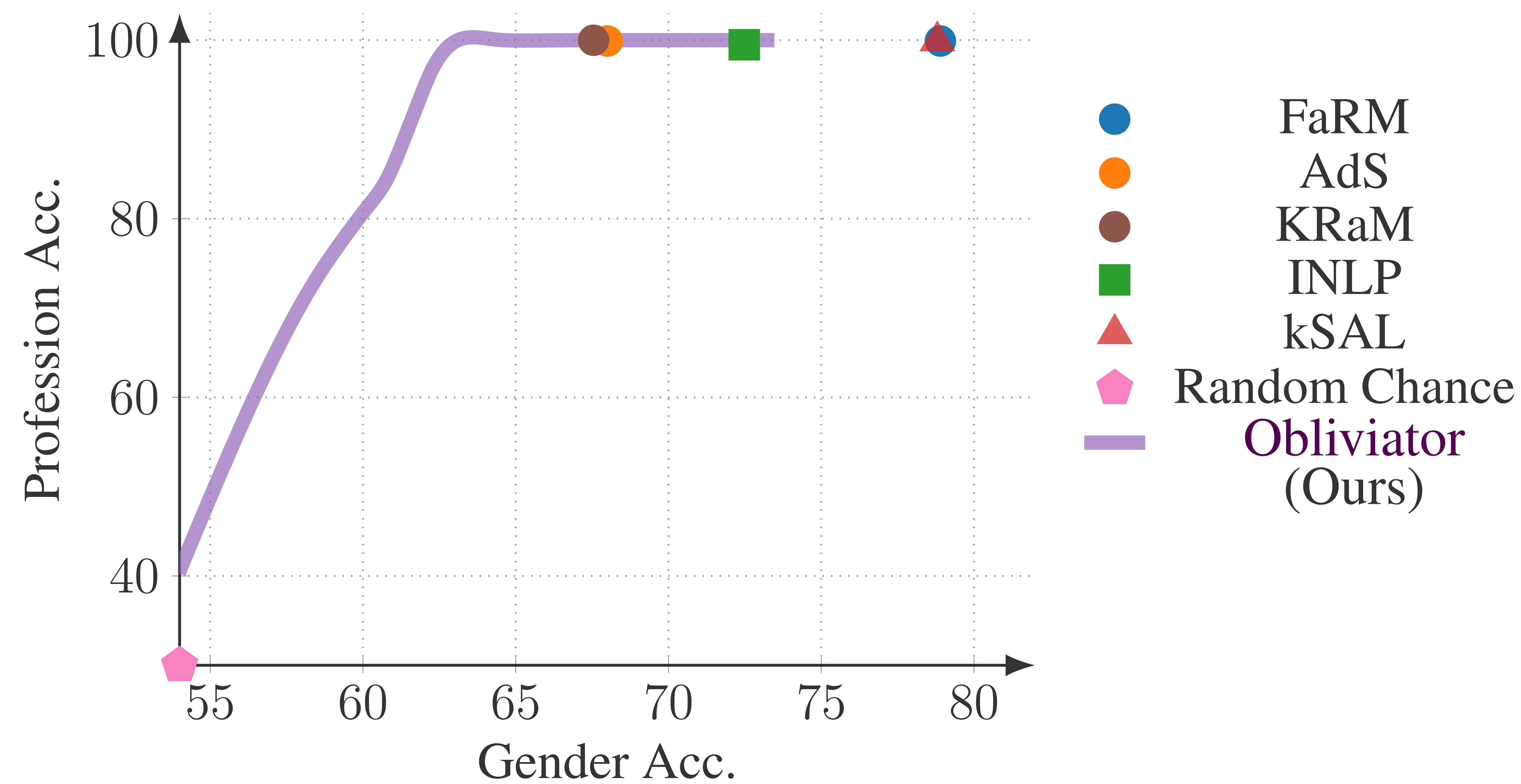
Utility : Profession

Unwanted : Gender

Obliviator Reveals : Achieving Nonlinear Guardedness



Bias in Bios



Bias in Bios

Representation: Frozen

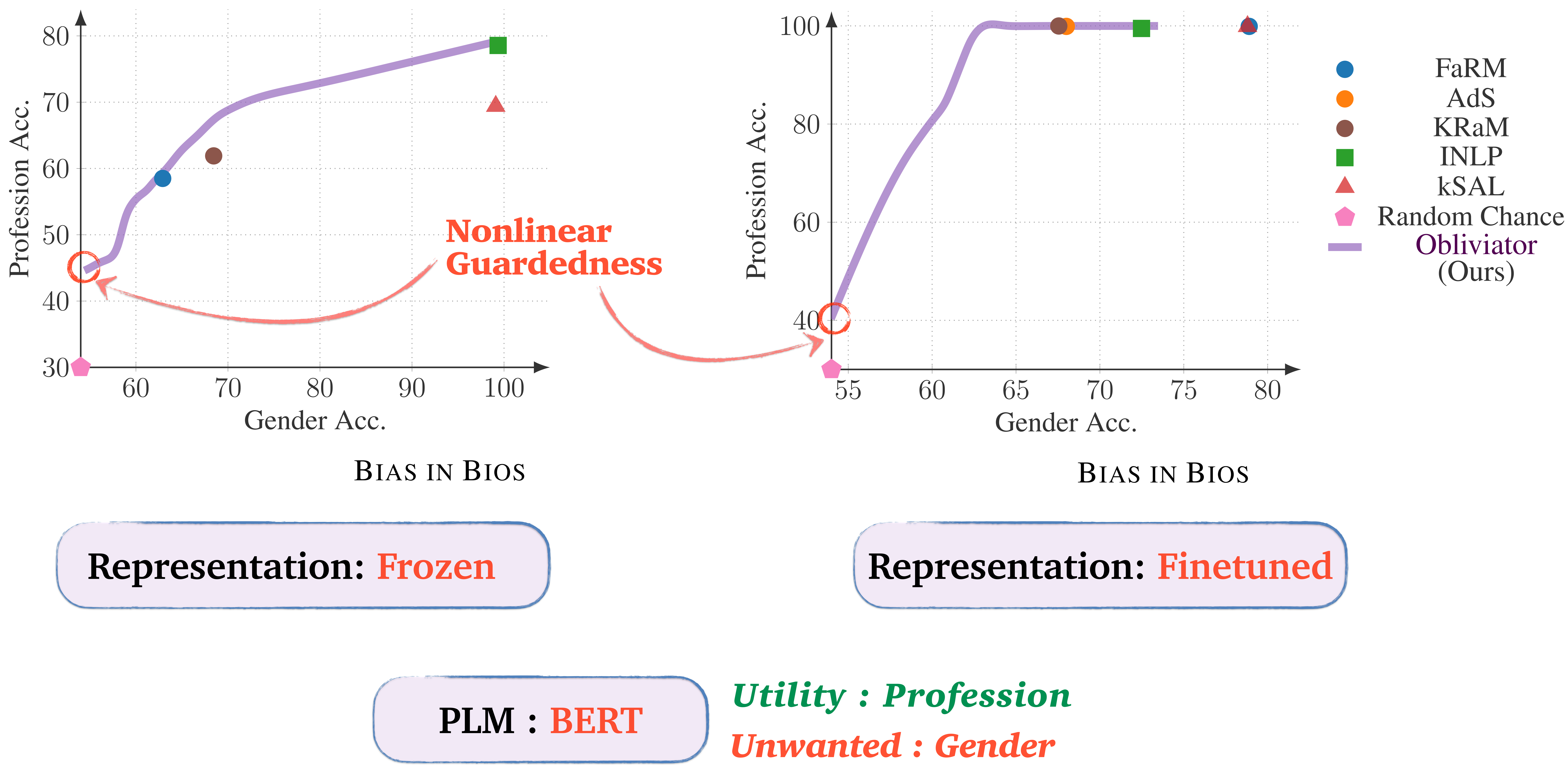
Representation: Finetuned

PLM : BERT

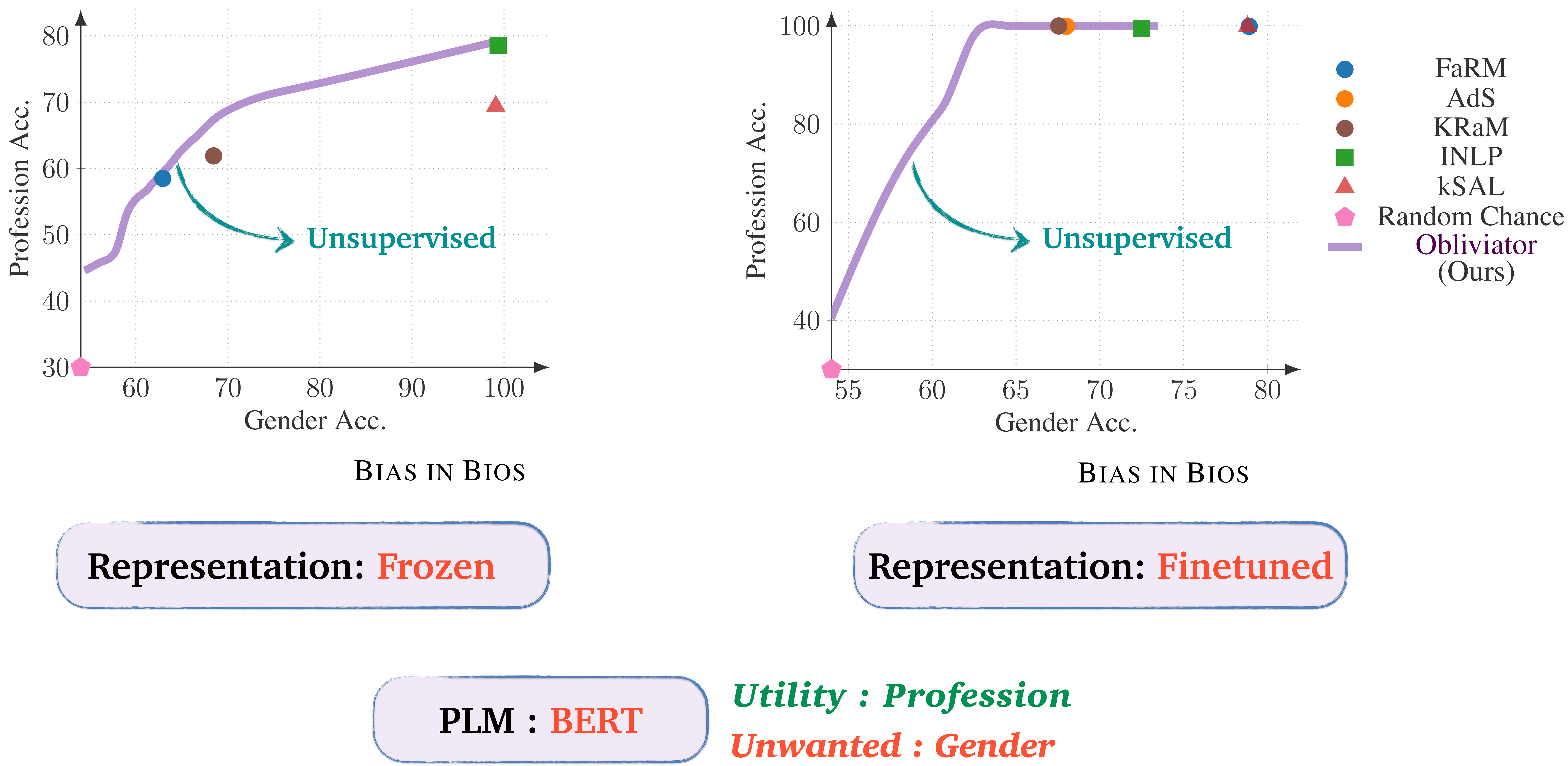
Utility : Profession

Unwanted : Gender

Obliviator Reveals : Achieving Nonlinear Guardedness



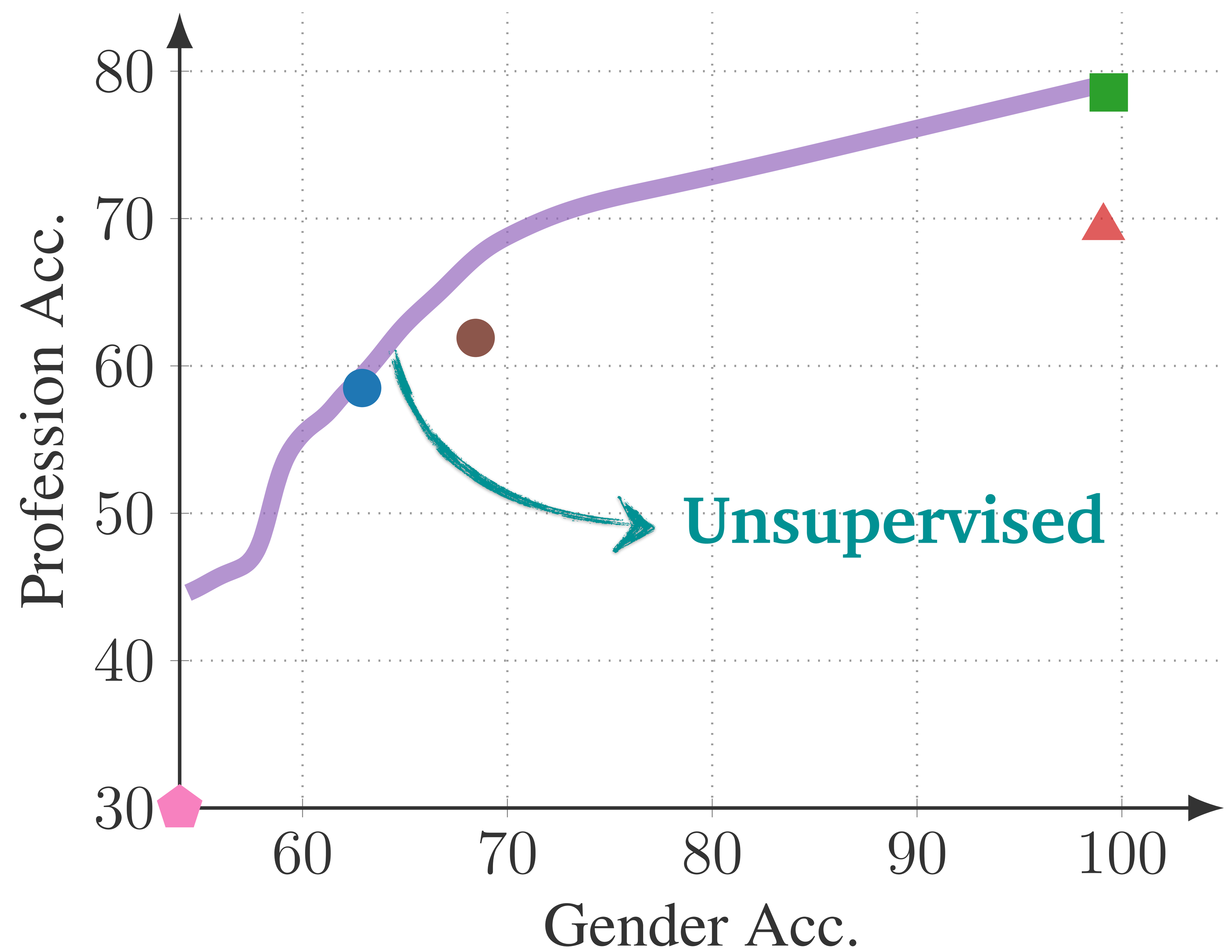
Obliviator Reveals : Effect of Target Task Labels on Erasure



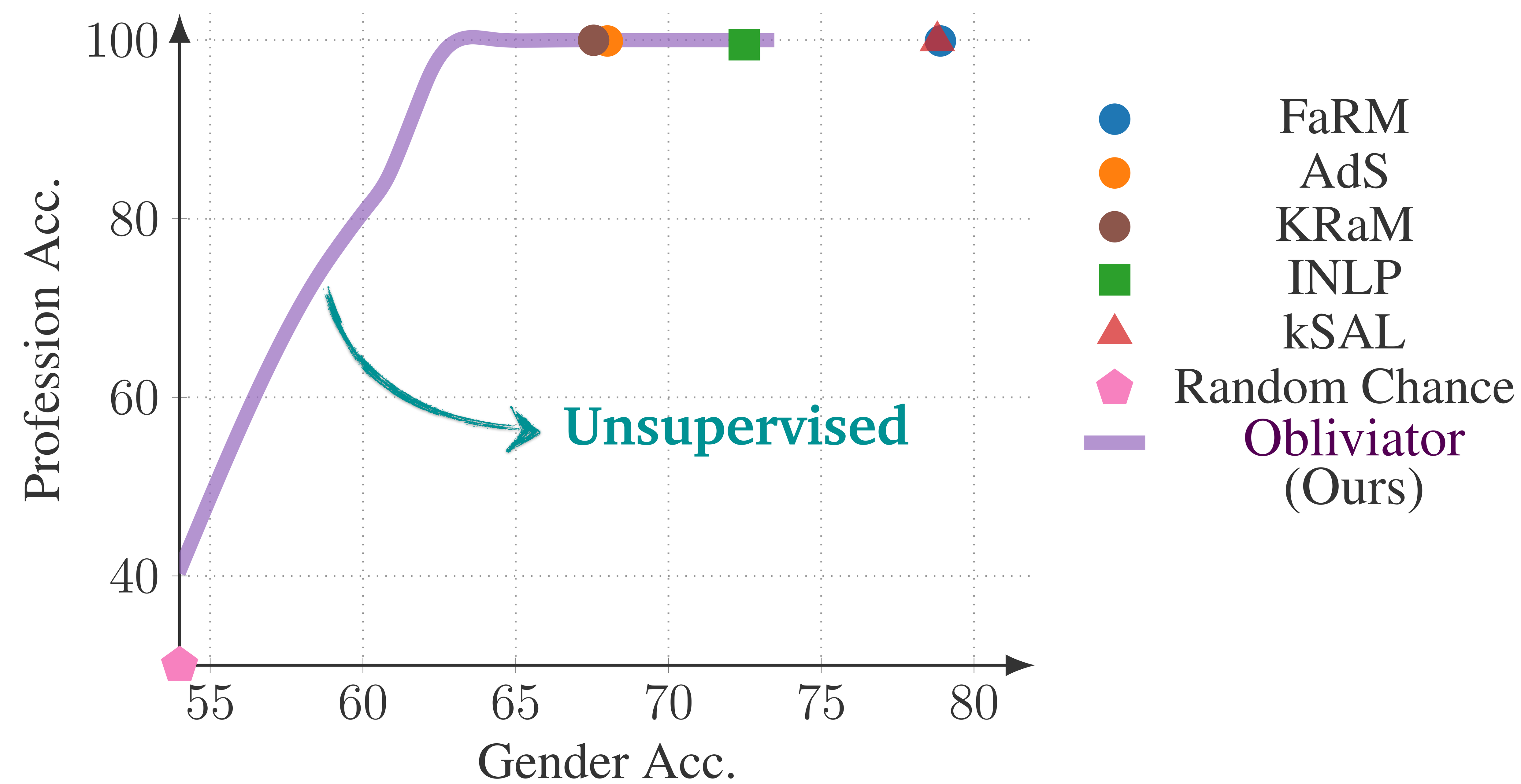
Obliviator Reveals : Effect of Target Task Labels on Erasure

Unsupervised

$$\inf_{\theta} \quad \mathrm{HSIC}(Z_\theta, S) - \left[\cancel{\mathrm{HSIC}(Z^i_\theta, Y)} + \mathrm{HSIC}(Z^i_\theta, X) + \mathrm{HSIC}(Z^i_\theta, X^i)\right]$$



BIAS IN BIOS



BIAS IN BIOS

Representation: Frozen

Representation: Finetuned

PLM : BERT

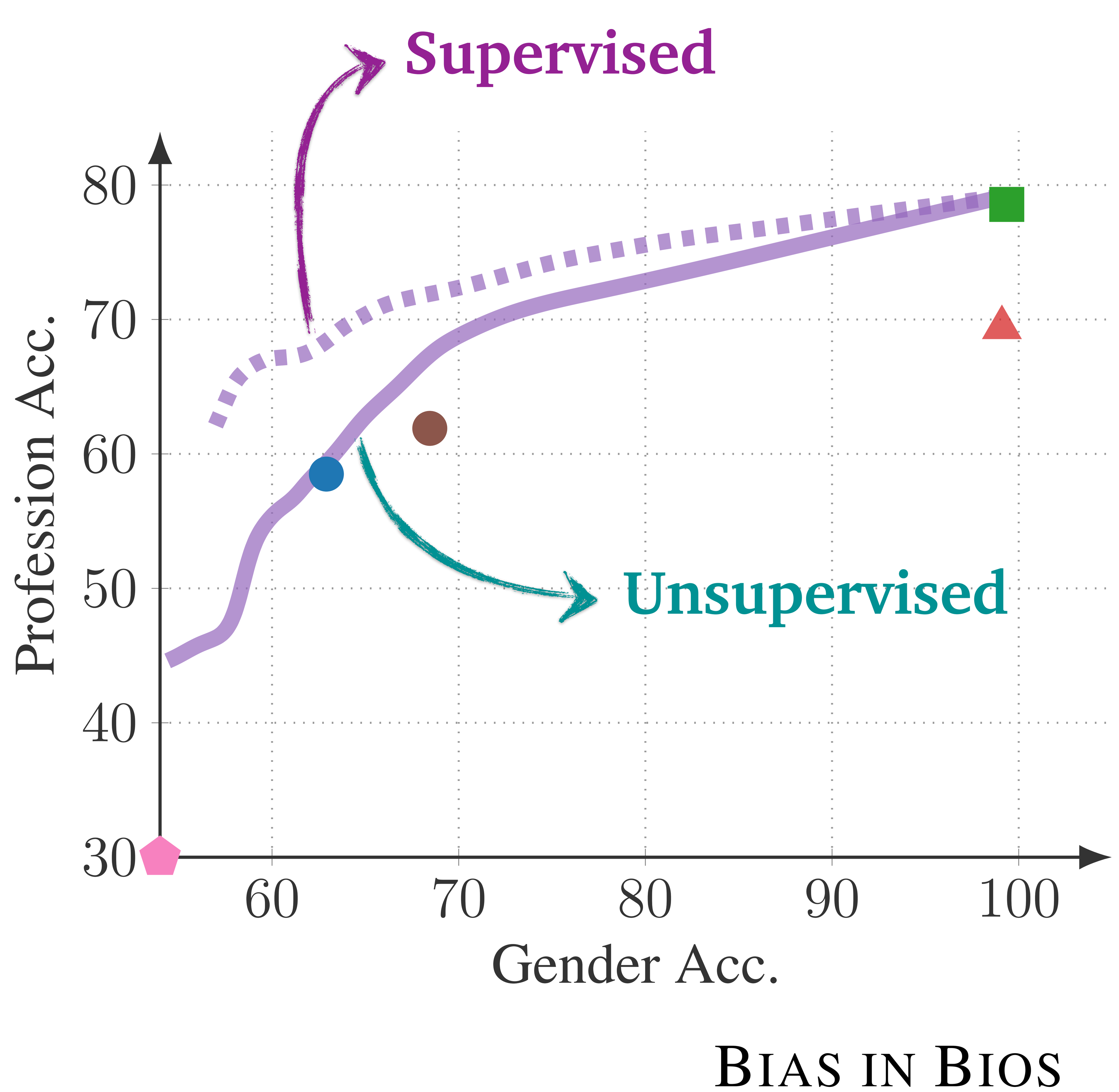
Utility : Profession

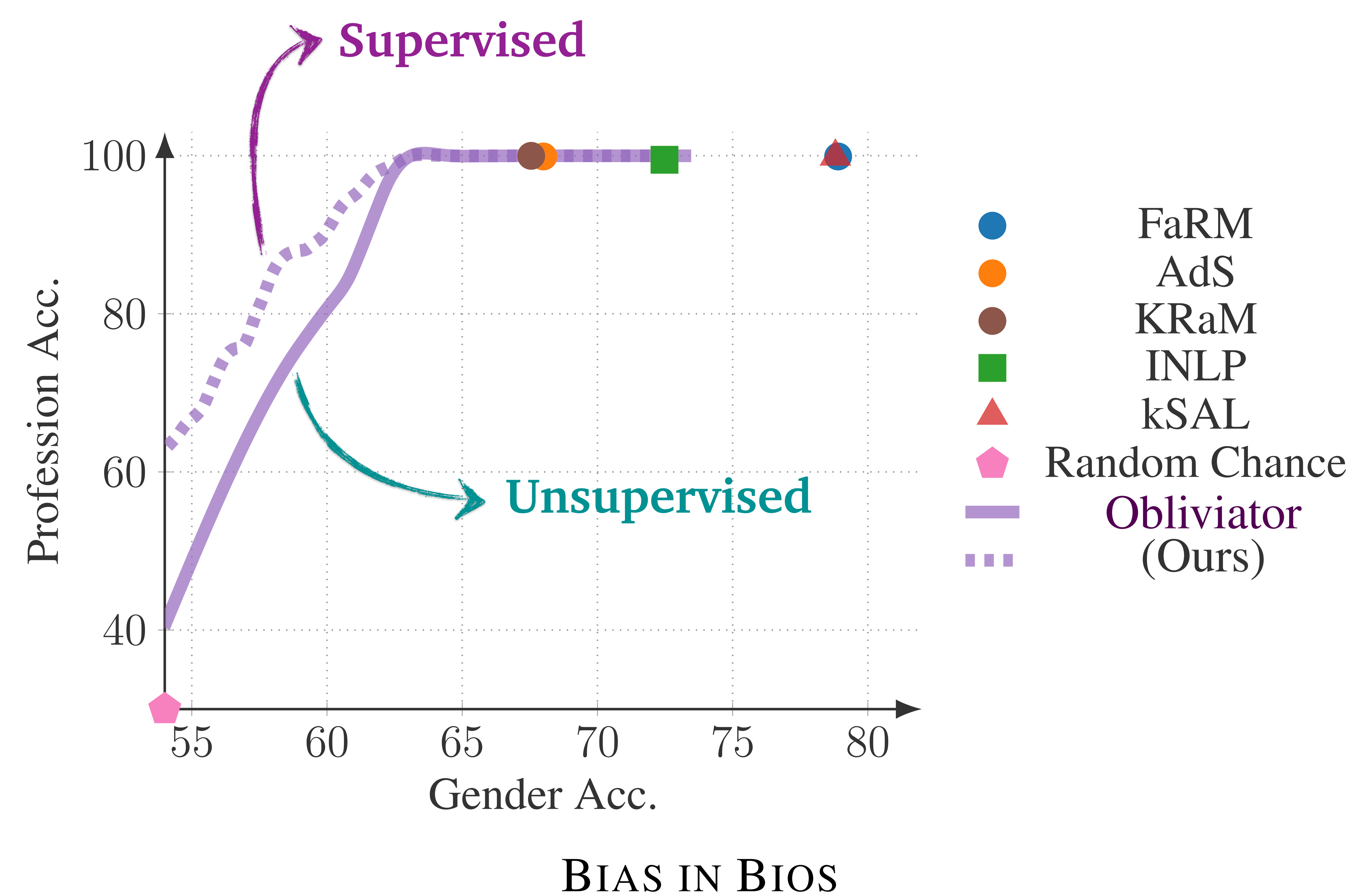
Unwanted : Gender

Obliviator Reveals : Effect of Target Task Labels on Erasure

Supervised

$$\inf_{\theta} \quad \mathrm{HSIC}(Z_\theta, S) - \left[\mathrm{HSIC}(Z_\theta^i, Y) + \mathrm{HSIC}(Z_\theta^i, X) + \mathrm{HSIC}(Z_\theta^i, X^i)\right]$$





Representation: Frozen

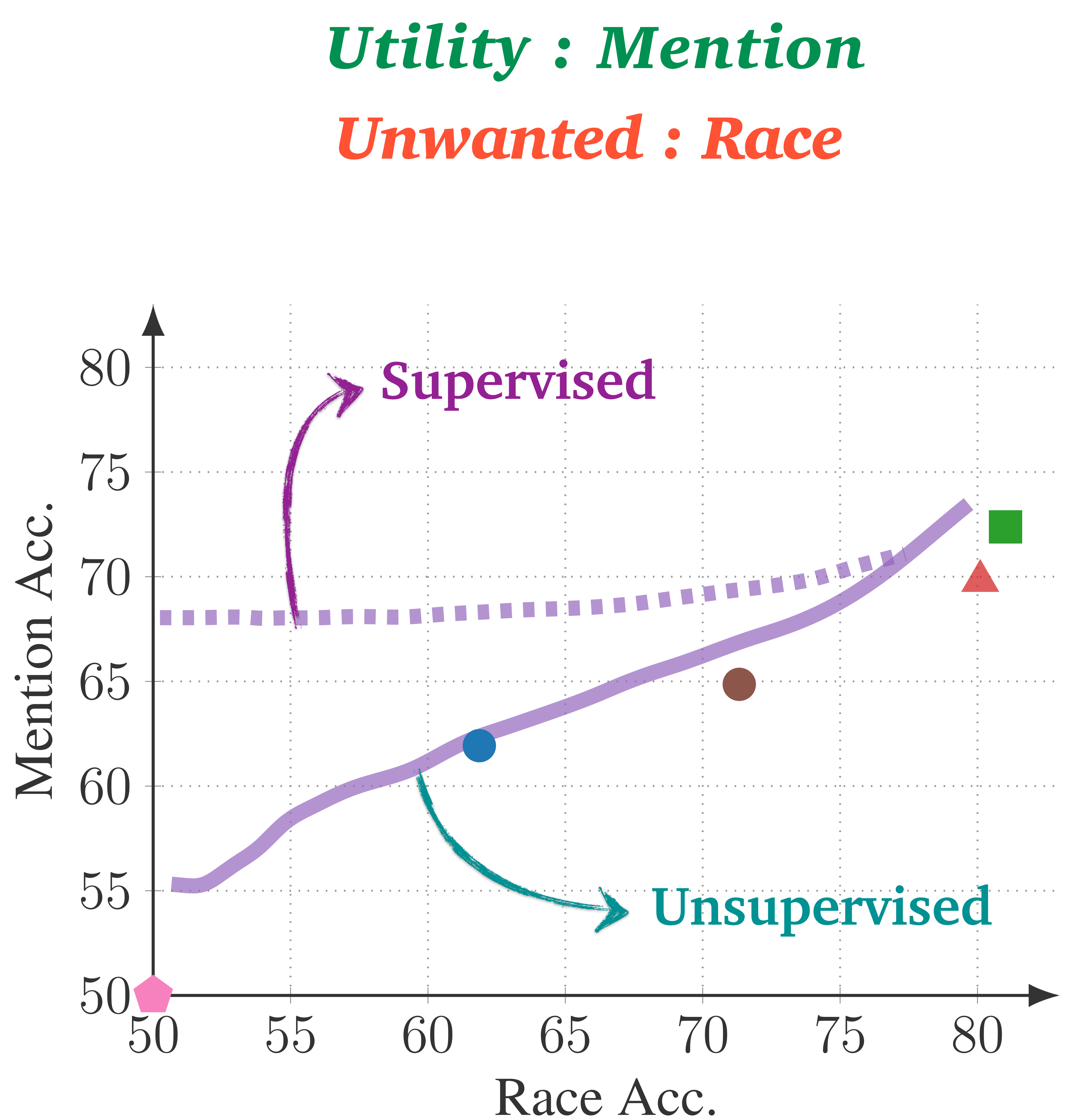
Representation: Finetuned

PLM : BERT

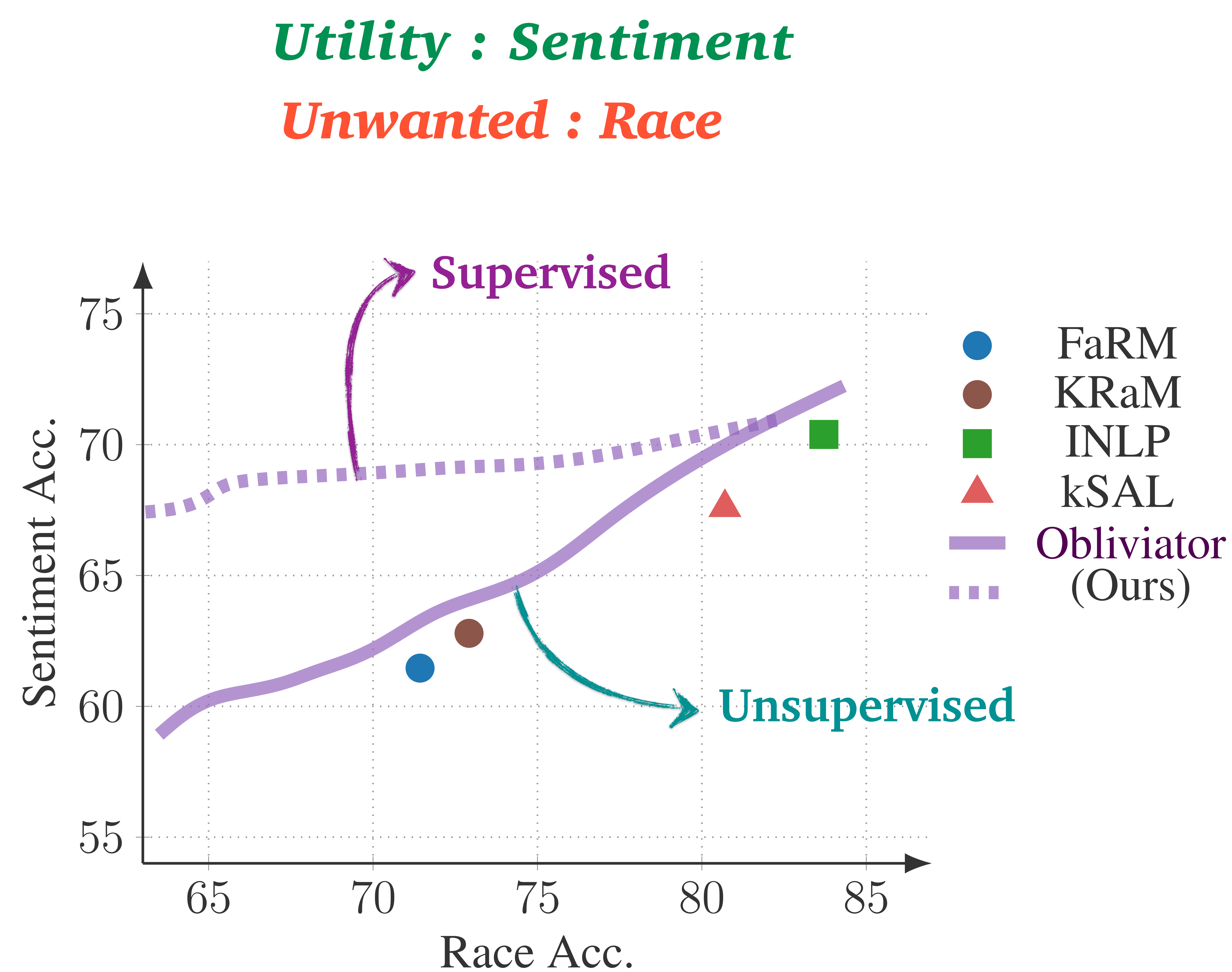
Utility : Profession

Unwanted : Gender

Obliviator Reveals : Effect of Target Task Labels on Erasure



(a) DIAL-MENTION

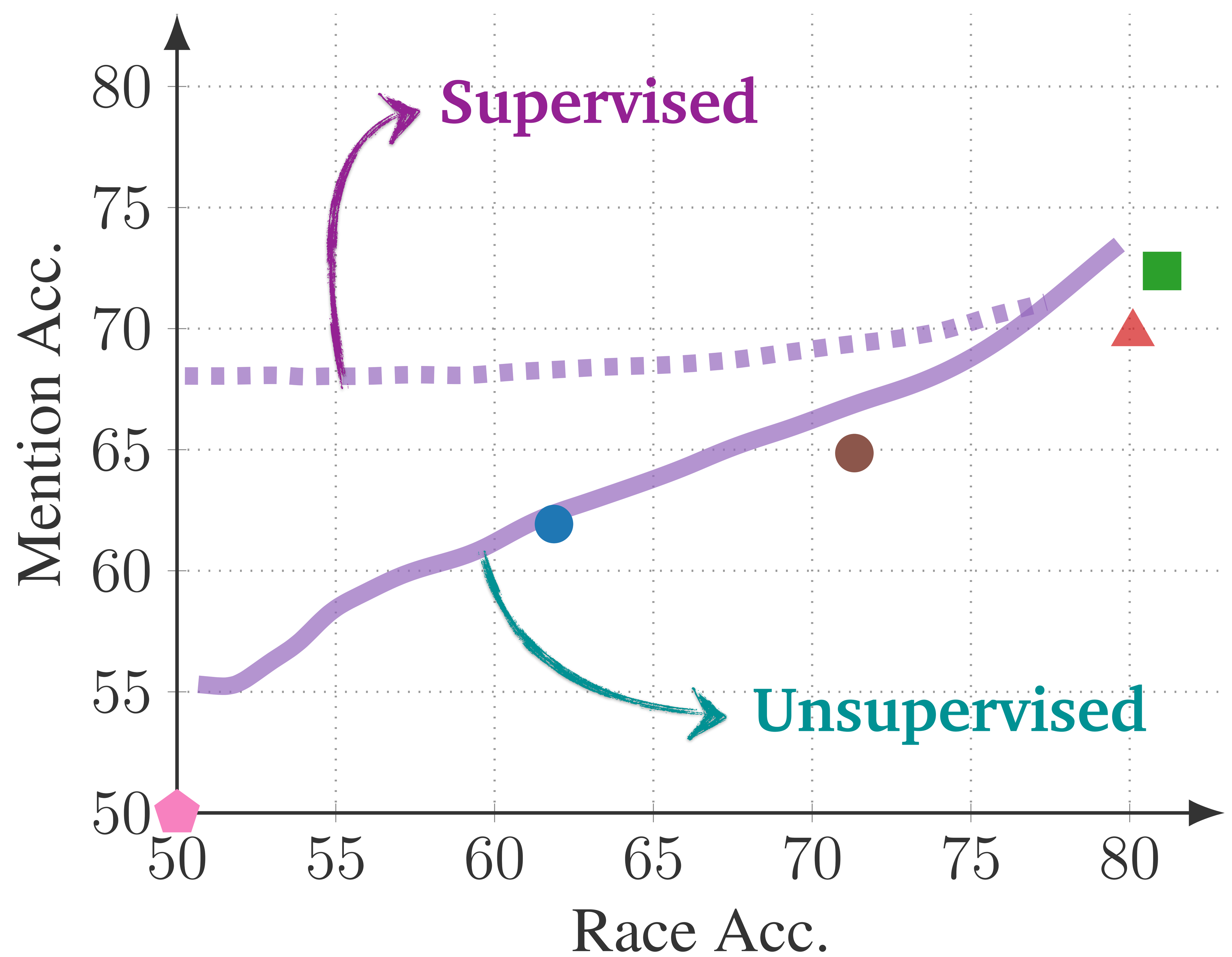


(b) DIAL-SENTIMENT

Representation: Frozen | **PLM : BERT**

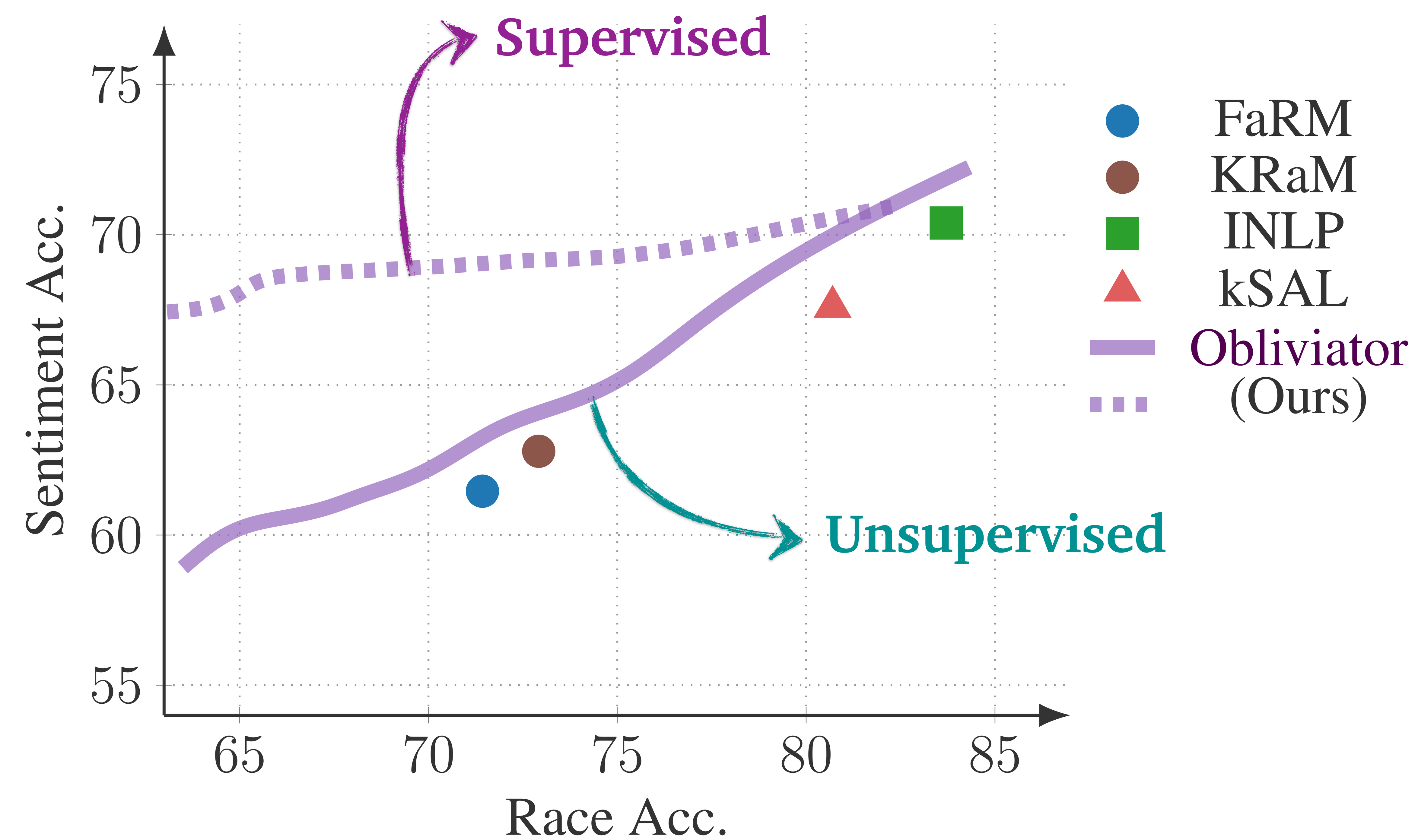
Obliviator Reveals : Effect of Target Task Labels on Erasure

Utility : Mention
Unwanted : Race



(a) DIAL-MENTION

Utility : Sentiment
Unwanted : Race

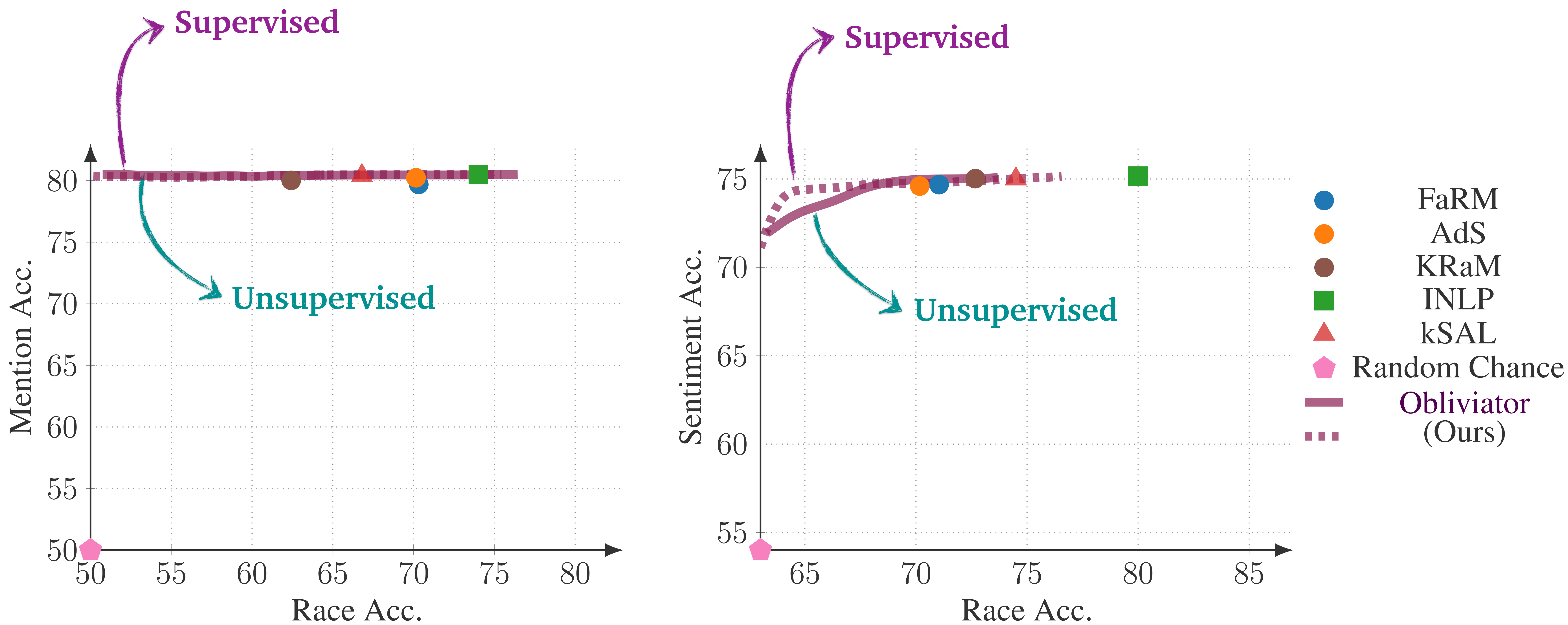


(b) DIAL-SENTIMENT

Representation: Frozen

PLM : BERT

Obliviator Reveals : Effect of Target Task Labels on Erasure

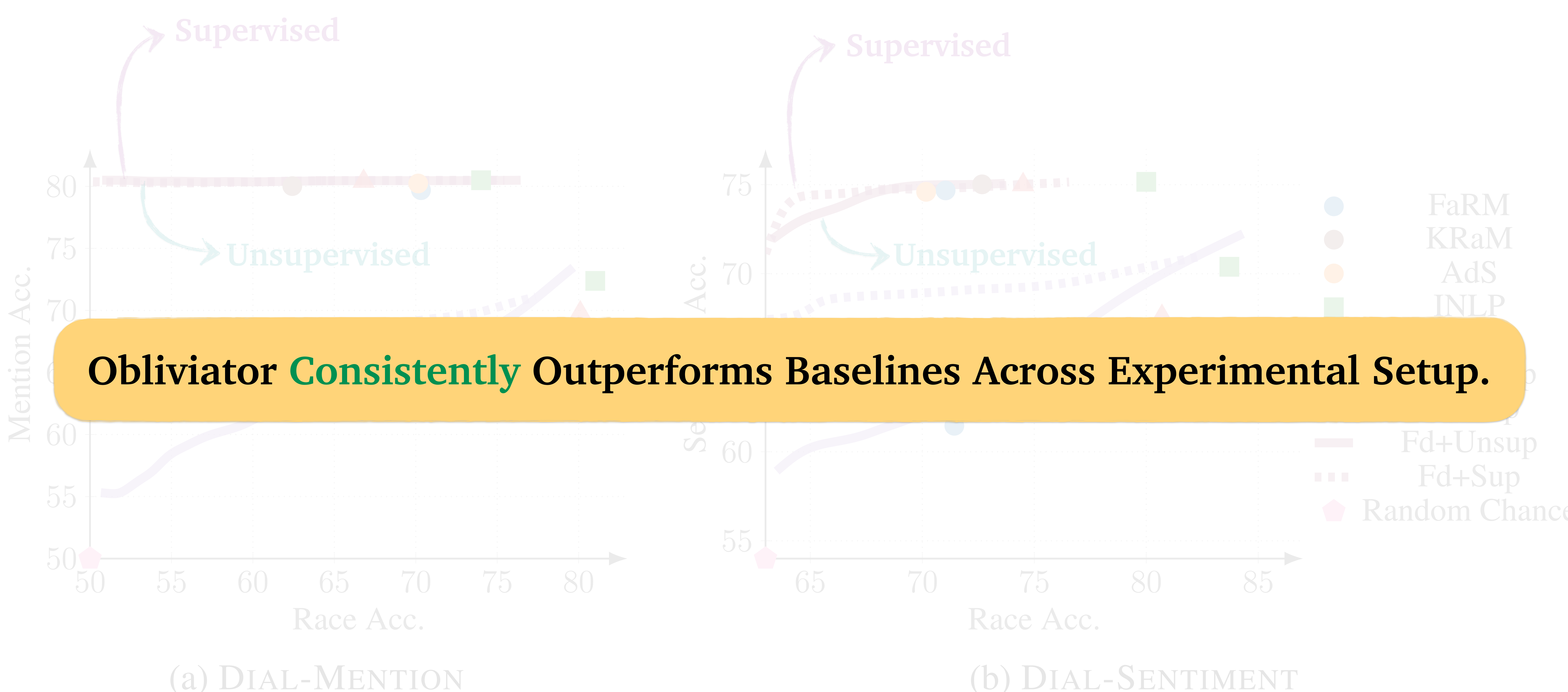


(a) DIAL-MENTION

(b) DIAL-SENTIMENT

Representation: Finetuned | **PLM : BERT**

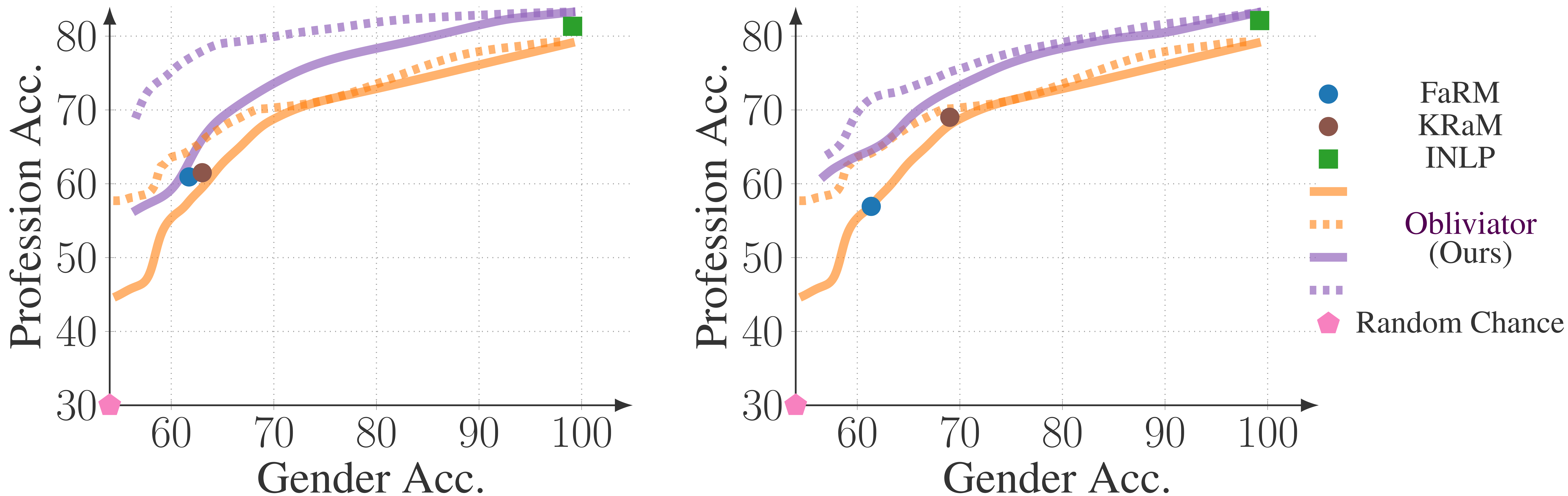
Obliviator Reveals : Effect of Target Task Labels on Erasure



Obliviator Consistently Outperforms Baselines Across Experimental Setup.

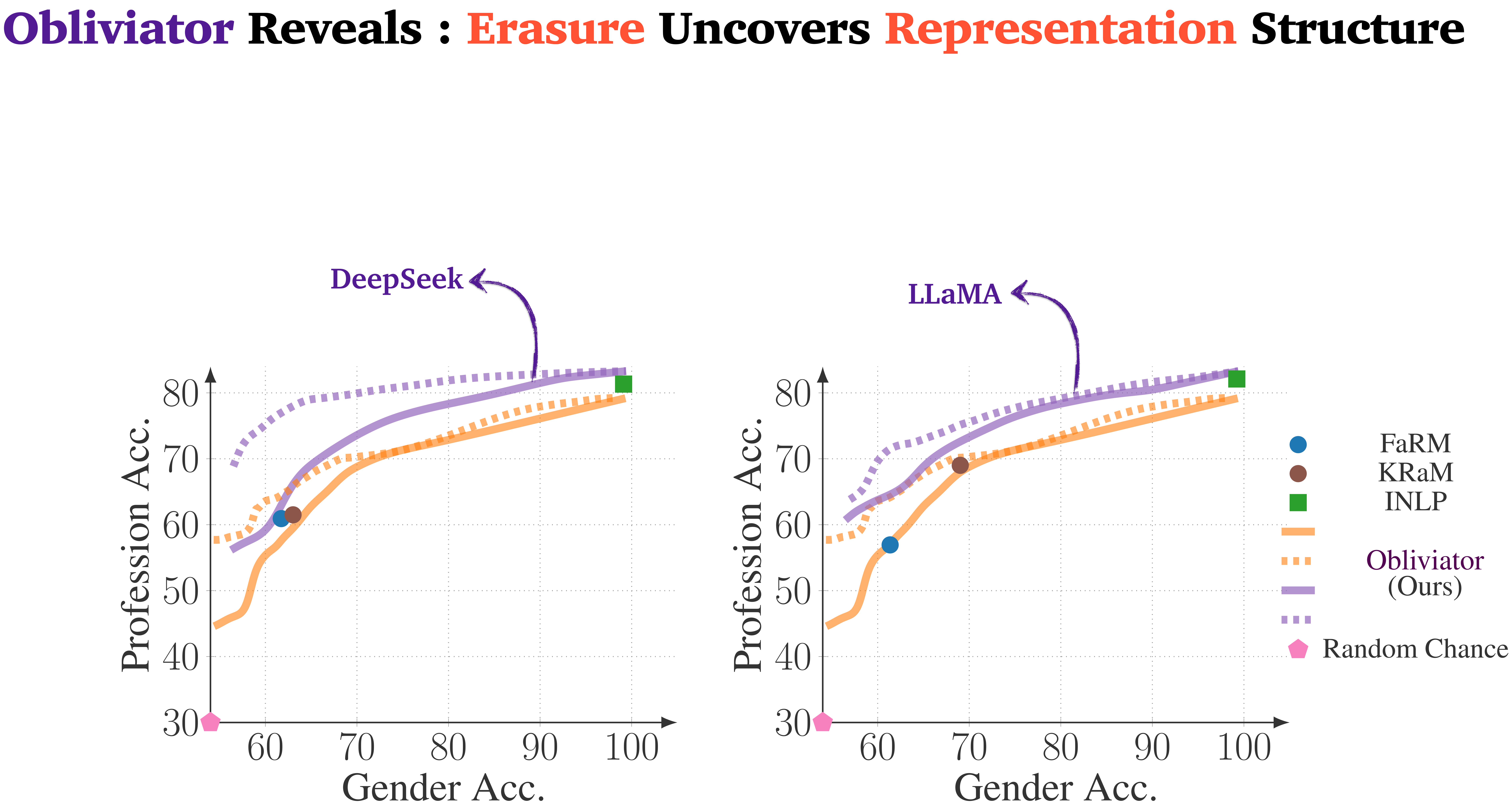
Obliviator Reveals : Erasure Uncovers Representation Structure

Obliviator Reveals : Erasure Uncovers Representation Structure



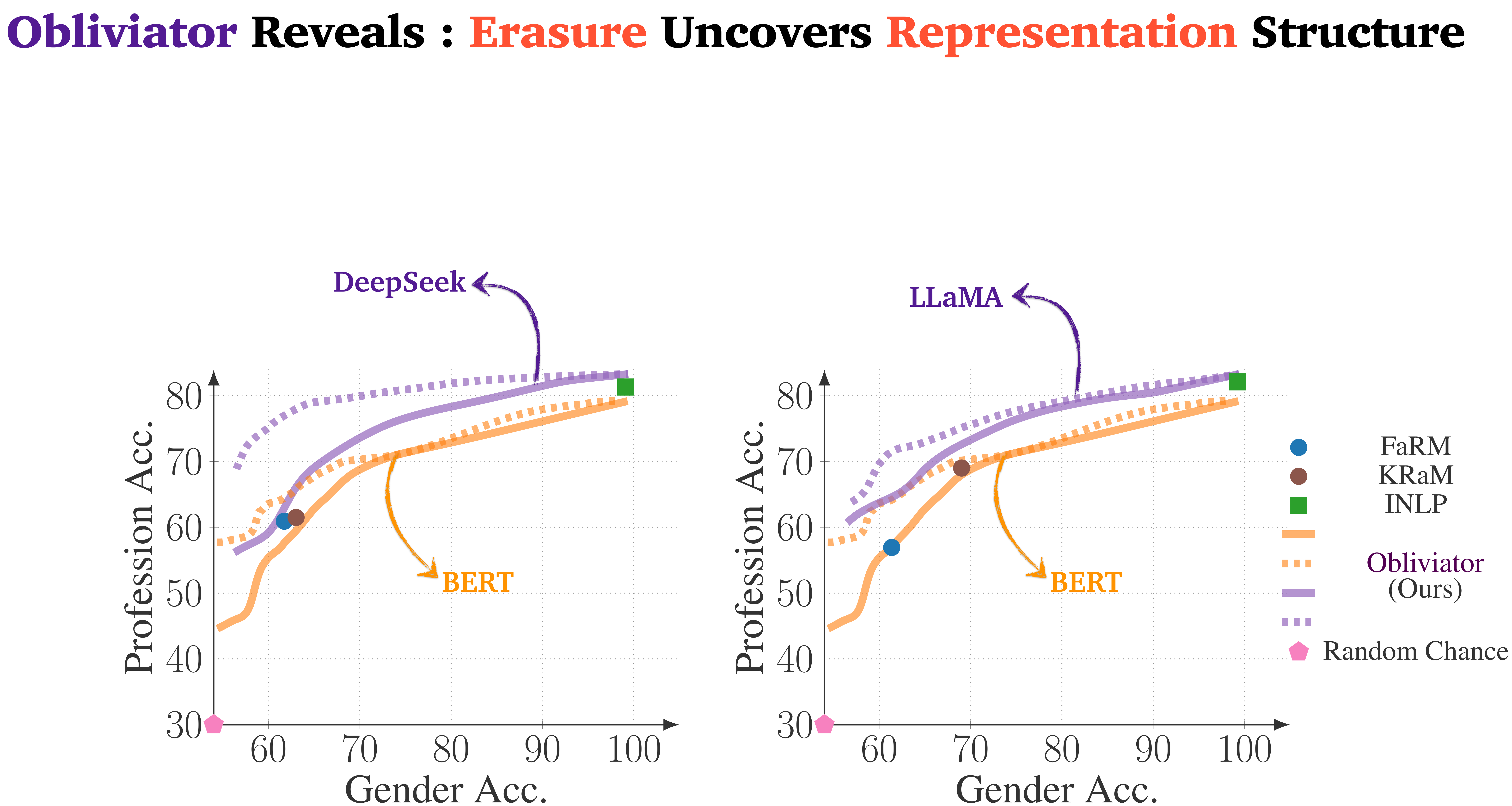
(a) DeepSeek Representations

(b) LLaMA Representations



(a) DeepSeek Representations

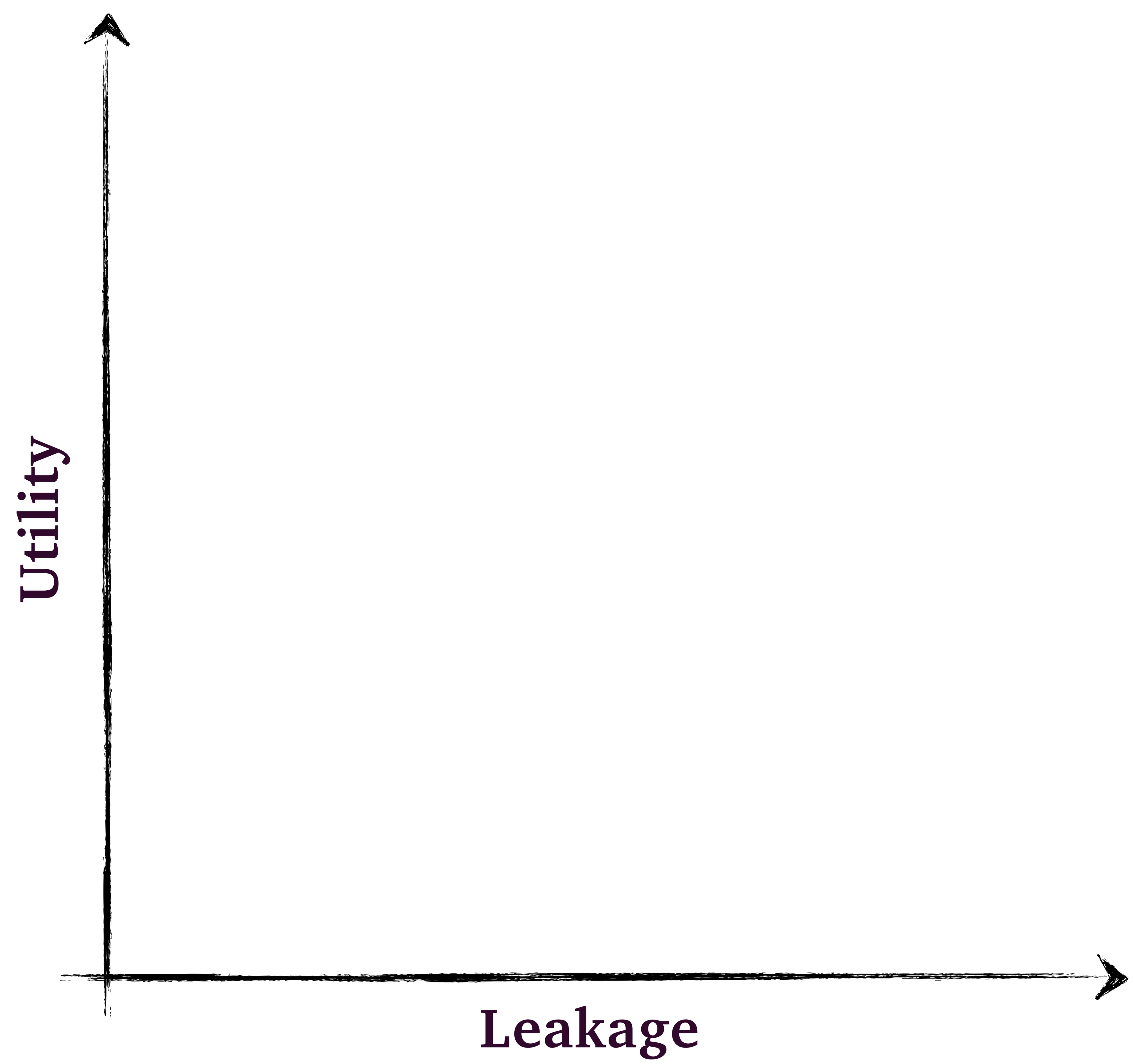
(b) LLaMA Representations



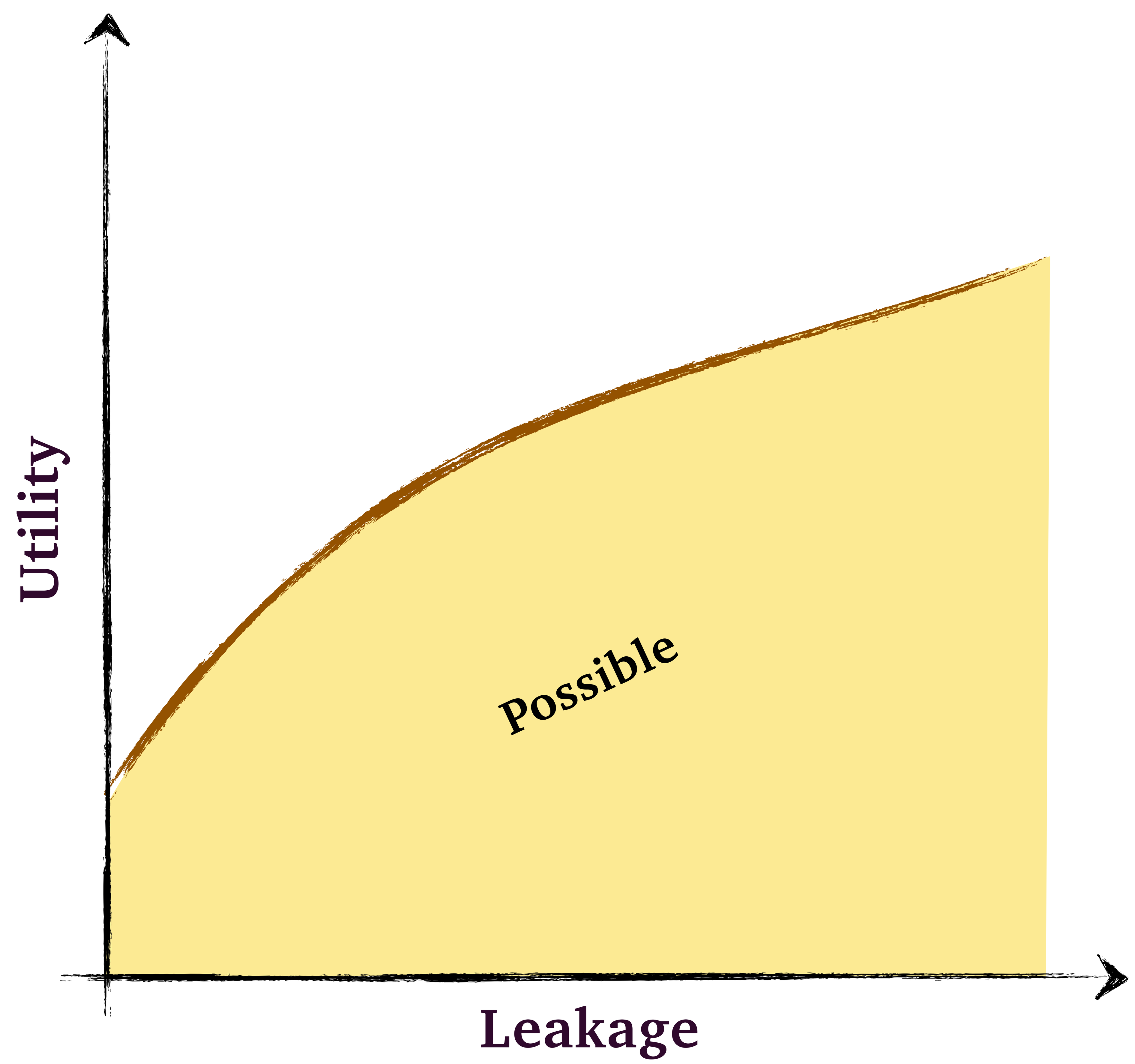
(a) DeepSeek Representations

(b) LLaMA Representations

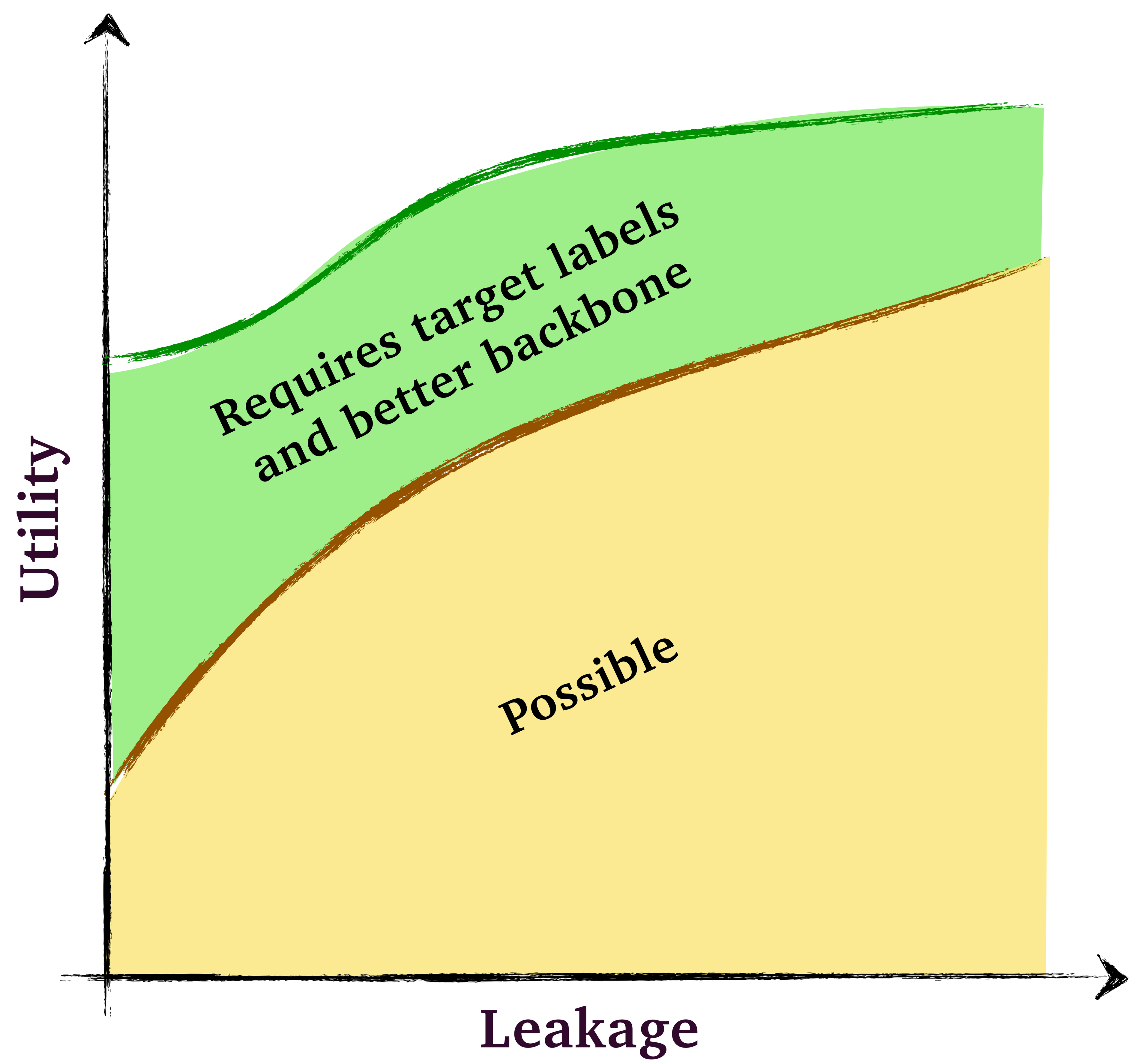
Obliviator Reveals : Utility–Erasure Trade-off



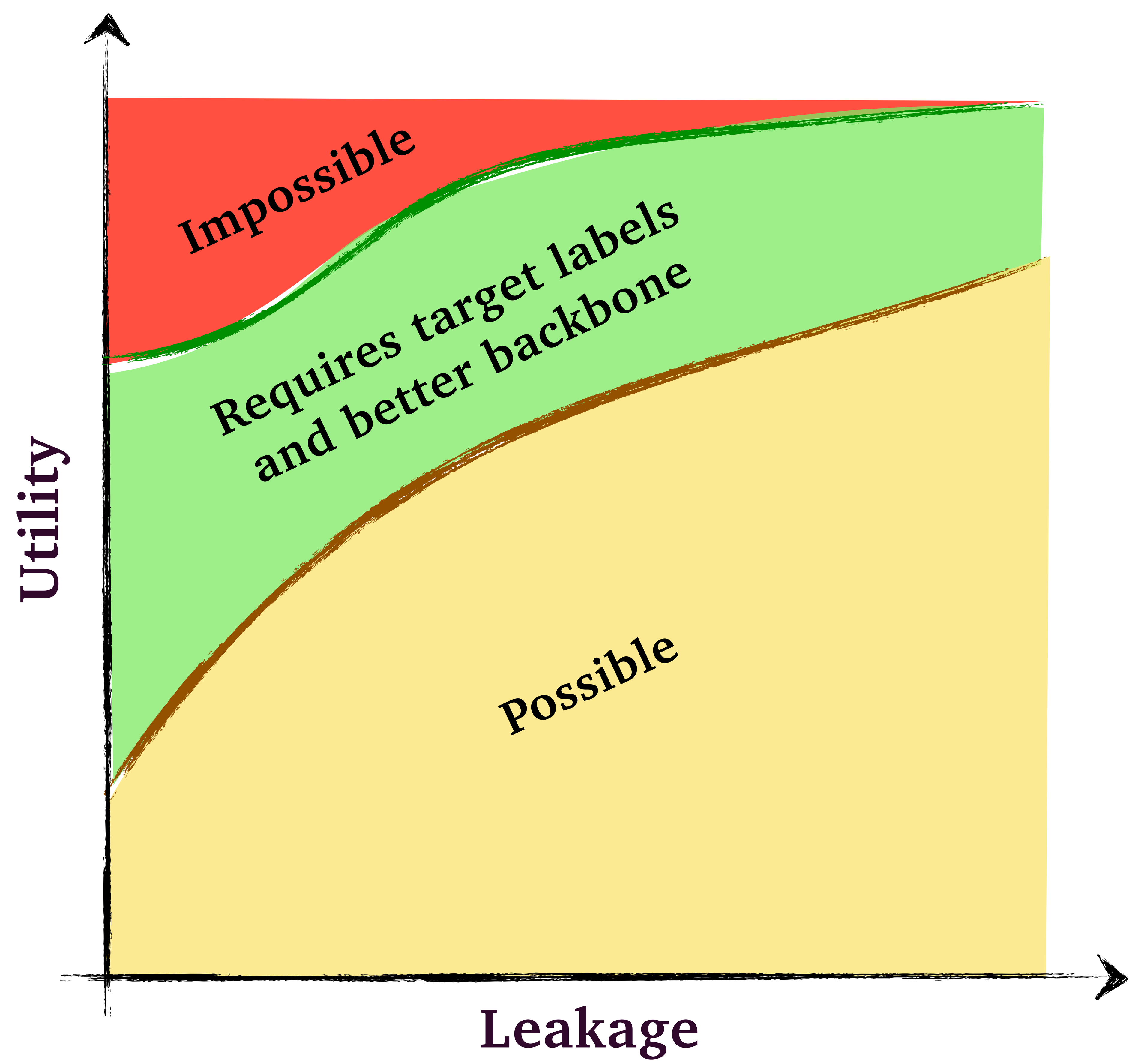
Obliviator Reveals : Utility–Erasure Trade-off



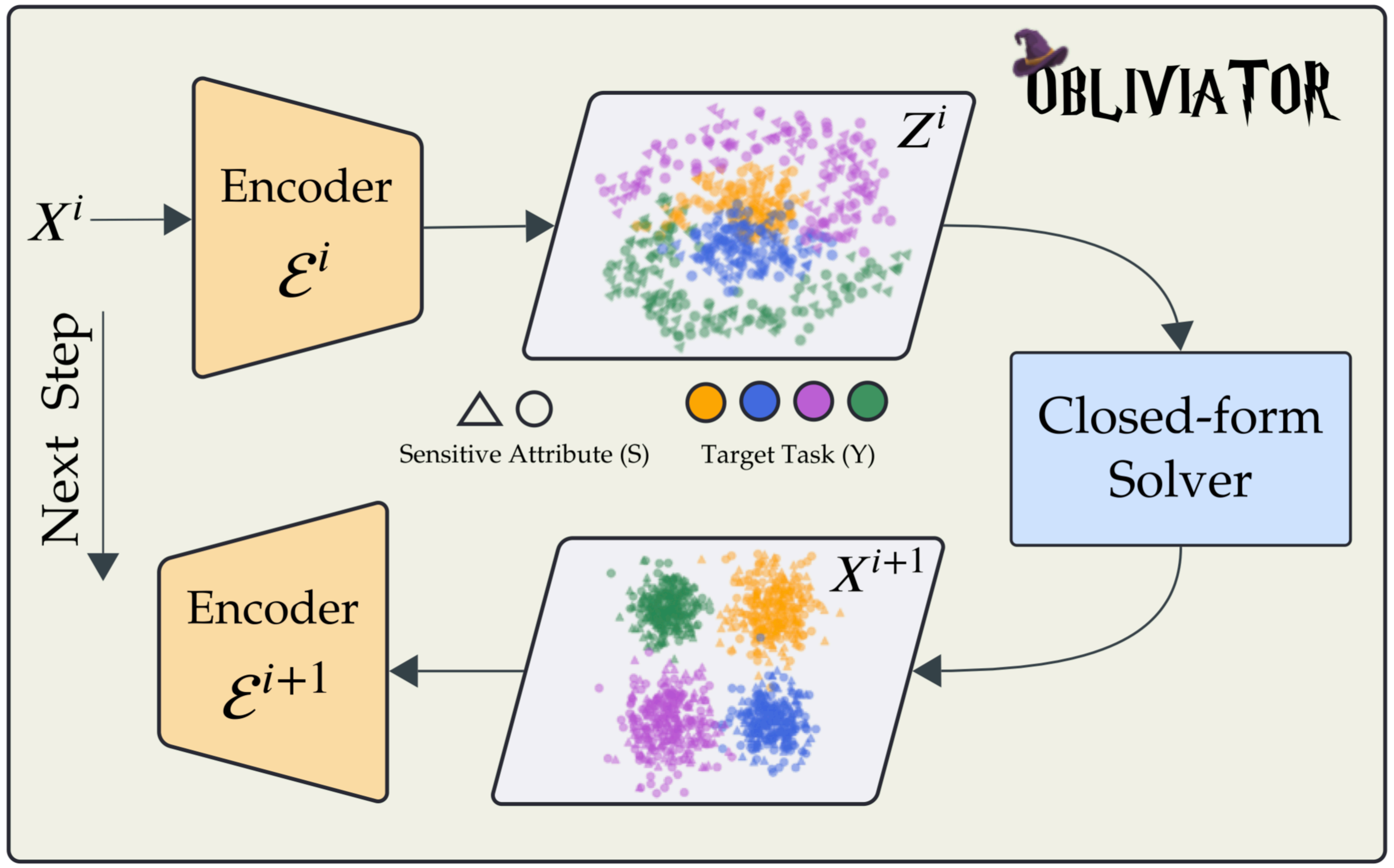
Obliviator Reveals : Utility–Erasure Trade-off



Obliviator Reveals : Utility–Erasure Trade-off



Obliviator At a Glance



- ***SOTA Utility-Erasure Trade-off***
- ***Achieves Nonlinear Guardedness***
- ***Computationally Efficient***
- ***Fine-Control over Erasure***