GENERATIVE LEARNING ALGORITHMS IN SUPERVISED LEARNING

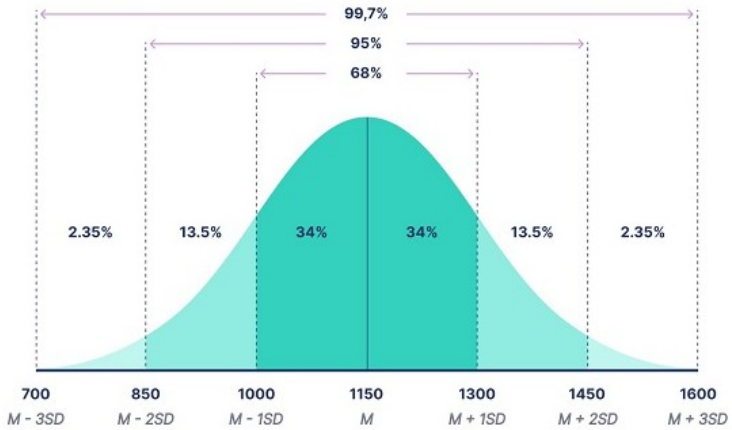
Nhu Thuy Nguyen Thi

University of Information Technology, VNU-HCM

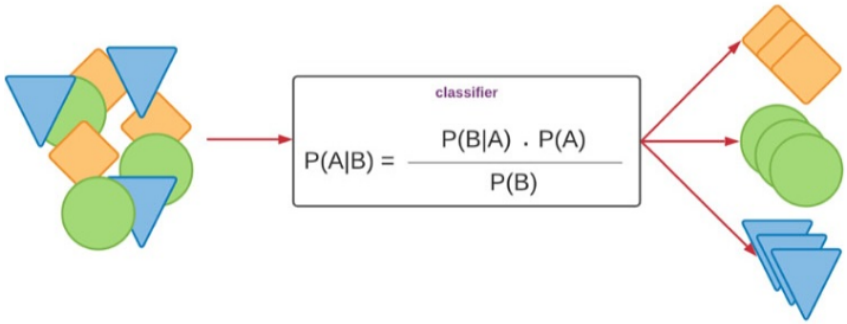
November 13rd, 2025

Generative Algorithms

- Gaussian Discriminant Analysis (GDA)

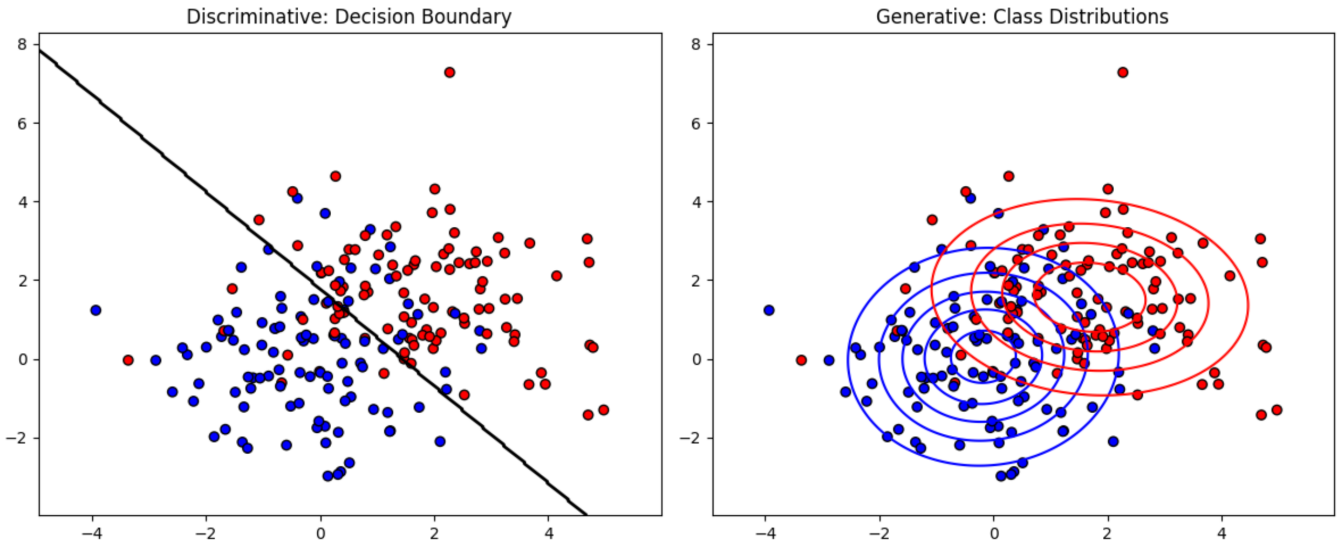


- Naive Bayes (NB)



Gaussian Discriminant Analysis

- Discriminative Learning Algorithms: model the **conditional distribution p(y | x; θ)**
- Generative Learning Algorithms: modeling the **joint probability distribution** $p(x, y)$



Gaussian Discriminant Analysis

Multivariate Normal Distribution

- **Notation**: $\mathcal{N}(\mu, \mathbf{\Sigma})$.
- **Probability Density Function** (PDF)

$$p(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

- d: Dimensionality of the space/data ($\mathbf{x} \in \mathbb{R}^d$).
- $|\mathbf{\Sigma}|$: The determinant of the covariance matrix.
- $\mathbf{x} - \mu$: Distance from the mean.
- $\mathbf{\Sigma}^{-1}$: The **Precision Matrix**.

Gaussian Discriminant Analysis

Let $\mathbf{X} = (X_1, X_2, \ldots, X_D)^T$ be a random vector with the expected value vector (mean vector) $\boldsymbol{\mu} = E[\mathbf{X}]$.
The covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$:

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \tag{1}$$

The element at position (i, j) of the matrix:

$$\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j) = E[(X_i - \boldsymbol{\mu}_i)(X_j - \boldsymbol{\mu}_j)] \tag{2}$$

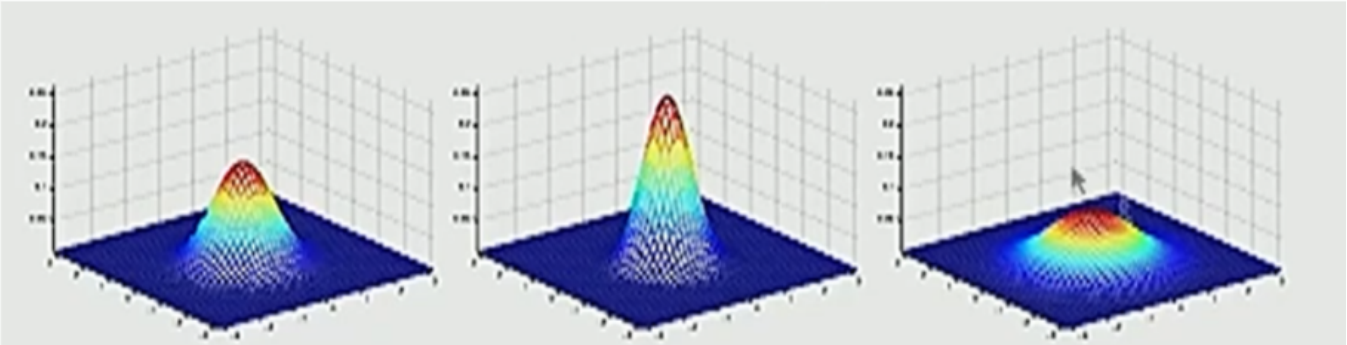
Gaussian Discriminant Analysis

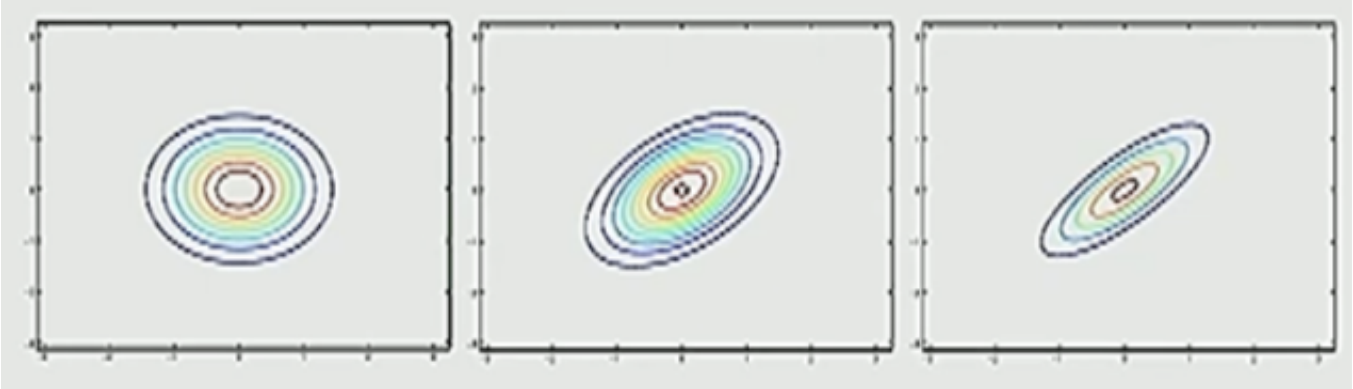
1. **Symmetry:**
   - $\mathbf{\Sigma} = \mathbf{\Sigma}^T$ or $\mathbf{\Sigma}_{ij} = \mathbf{\Sigma}_{ji} \quad \forall i, j$.
   - This property follows directly from $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$.
2. **Positive Semi-Definite (PSD):**
   - For any vector $\mathbf{z} \in \mathbb{R}^D$, the following condition must hold:

$$\mathbf{z}^T \mathbf{\Sigma} \mathbf{z} \geq 0 \tag{3}$$

   - This property ensures that the variance of any linear combination of the random variables is non-negative, since $\mathbf{z}^T \mathbf{\Sigma} \mathbf{z} = \text{Var}(\mathbf{z}^T \mathbf{X})$.

Gaussian Discriminant Analysis





Gaussian Discriminant Analysis

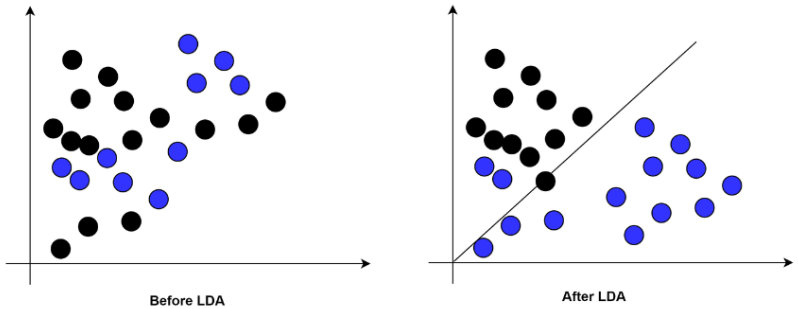
GDA Model

- The prior distribution $p(y)$:

$$y \sim \text{Bernoulli}(\phi), \text{ where } \phi = p(y = 1).$$

The features $\mathbf{x}$ given y are assumed to follow a **Multivariate Normal Distribution** ($\mathcal{N}$):

- **Class 0:** $\mathbf{x}|y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$
- **Class 1:** $\mathbf{x}|y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$



Gaussian Discriminant Analysis

The density of this distribution is given by the formula:

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \qquad \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Model PDF of the Multivariate Normal Distribution to conditional probabilities as:

- $p(\mathbf{x} \mid y = 0)$ uses $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$.
- $p(\mathbf{x} \mid y = 1)$ uses $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}$.

Note: Both classes share the same covariance matrix ($\boldsymbol{\Sigma}$).

Gaussian Discriminant Analysis

Likelihood function:

$$L(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = P\left((\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)}); \phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right)$$

- $\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}$: These are the **parameters** of the GDA model that must be estimated:
  - ϕ: The prior probability $P(y = 1)$.
  - $\boldsymbol{\mu}_0$: The mean vector of Class 0.
  - $\boldsymbol{\mu}_1$: The mean vector of Class 1.
  - $\boldsymbol{\Sigma}$: The **shared** Covariance matrix for both classes (the LDA assumption).
- $(\mathbf{x}^{(i)}, y^{(i)})$: The i-th training example, where $\mathbf{x}^{(i)}$ is the feature vector and $y^{(i)}$ is the class label (0 or 1).

$$L(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \prod_{i=1}^{m} P(\mathbf{x}^{(i)}, y^{(i)}; \phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \qquad (4)$$

Gaussian Discriminant Analysis

- **Equation to Solve:**

$$\nabla \ell(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \mathbf{0}$$

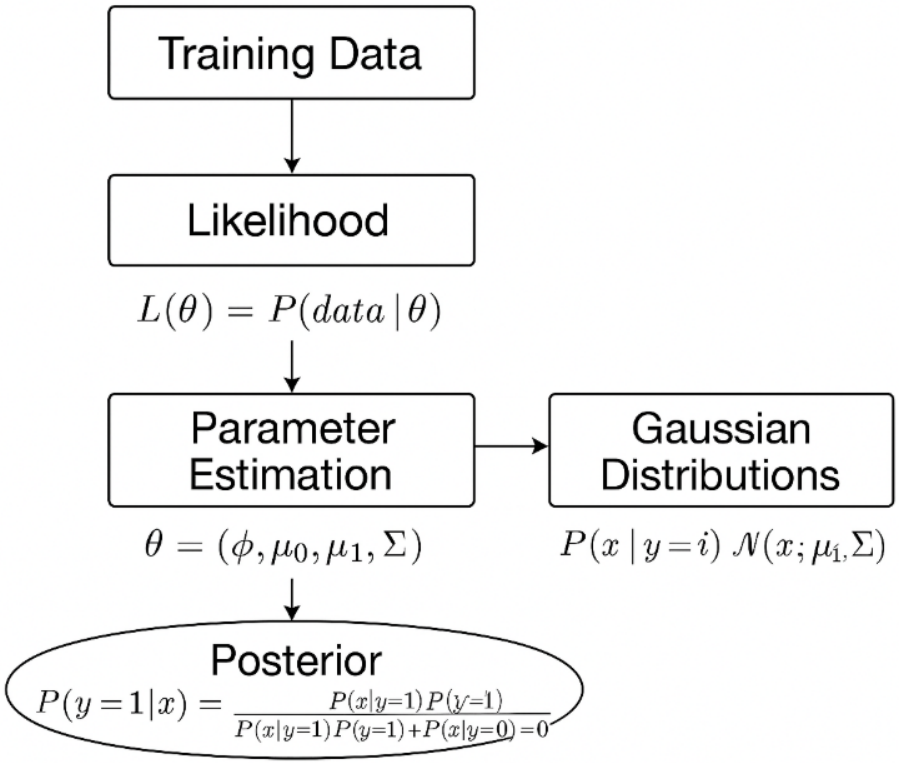
- **Partial Derivatives Components:** To solve the zero-gradient vector:

$$\frac{\partial \ell}{\partial \phi} = 0, \quad \frac{\partial \ell}{\partial \boldsymbol{\mu}_0} = \mathbf{0}, \quad \frac{\partial \ell}{\partial \boldsymbol{\mu}_1} = \mathbf{0}, \quad \frac{\partial \ell}{\partial \boldsymbol{\Sigma}} = \mathbf{0}$$

Modeled Components: P(y=0), P(y=1),
P(x | $y = 0$), $P(\mathbf{x} \mid y = 1)$

$$\Longrightarrow \boxed{P(y \mid \mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}}$$

Gaussian Discriminant Analysis

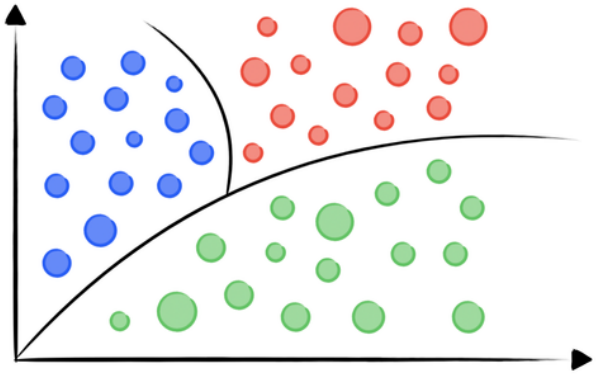


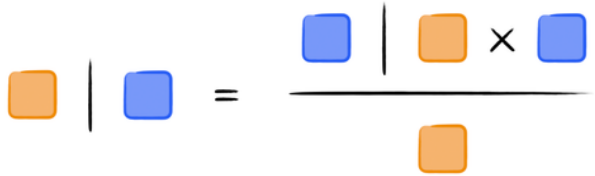
1. **Efficiency**
2. **Generative Modeling**

Naive Bayes

The Naive Bayes algorithm models the probabilities of discrete random variables:

- Conditional Independence
- Laplace Smoothing
- Event Models for Text Classification include:
  - Bernoulli Naive Bayes
  - Multinomial Naive Bayes





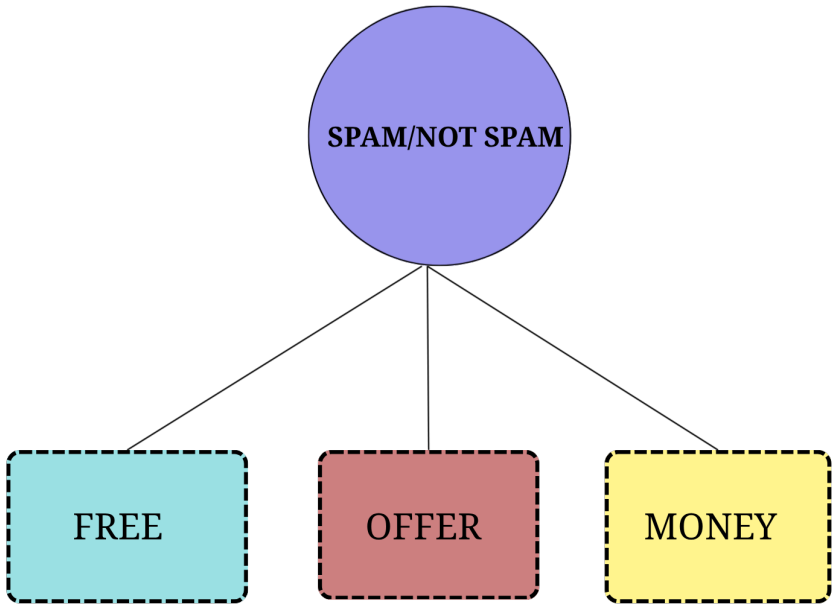
Conditional Independence

Definition: Two random variables A and B are said to be **conditionally independent** given variable C if:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

Notation:

$$A \perp B \mid C$$



Conditional Independence

1. **Expressing Conditional Probability:**

$$p(x_1, \ldots, x_d \mid y) = p(x_1 \mid y)p(x_2 \mid y, x_1)p(x_3 \mid y, x_1, x_2) \cdots p(x_d \mid y, x_1, \ldots$$

2. **Applying the Naive Bayes Assumption:**

$$p(x_j \mid y, x_1, \ldots, x_{j-1}) \approx p(x_j \mid y)$$

3. **The Simplified Formula:**

$$p(x_1, \ldots, x_d \mid y) = \prod_{j=1}^{d} p(x_j \mid y)$$

Laplace Smoothing

$$p(y \mid \mathbf{x}) = \frac{\left(\prod_{j=1}^{d} p(x_j \mid y)\right) p(y)}{\sum_{y' \in Y} \left(\prod_{j=1}^{d} p(x_j \mid y')\right) p(y')}$$

- $p(\mathbf{x} \mid y)$ into the product $\prod_{j=1}^{d} p(x_j \mid y)$. The detailed formula for $y = 1$ becomes:

$$p(y = 1 \mid \mathbf{x}) = \frac{\left(\prod_{j=1}^{d} p(x_j|y{=}1)\right)p(y{=}1)}{\left(\prod_{j=1}^{d} p(x_j|y{=}1)\right)p(y{=}1)+\left(\prod_{j=1}^{d} p(x_j|y{=}0)\right)p(y{=}0)}$$

$$\Longrightarrow \textbf{Zero Probability Problem}$$

Laplace Smoothing

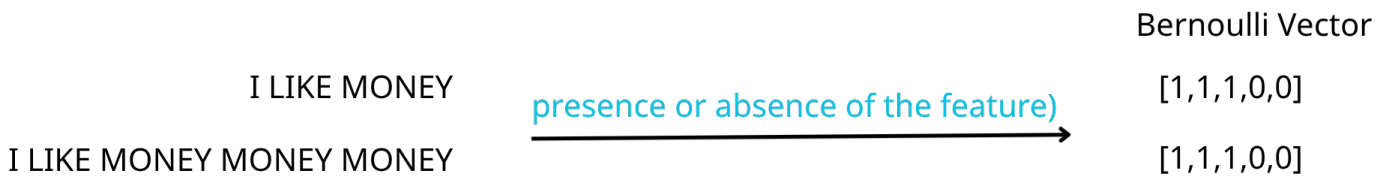
The probability estimate for parameter ϕ_j (e.g., $P(y = j)$ or $P(x_k \mid y = j)$):

$$\phi_j = \frac{\alpha + \sum_{i=1}^{n} \mathbb{1}\{z^{(i)} = j\}}{k\alpha + n}$$

Event Models for Text Classification

- **Bernoulli Event Model Assumption:**
  - Each word is treated as a **binary feature** (present or absent).
  - **Frequency is ignored**; only presence/absence matters.
- **Probability Model:**

$$P(\mathbf{x} \mid y) = \prod_{j=1}^{V} [P(w_j \mid y)]^{x_j} [1 - P(w_j \mid y)]^{1-x_j}$$



Event Models for Text Classification

- **Multinomial Event Model Assumption:**
  - A document is a sequence of word events drawn from a probability distribution.
  - The **frequency** (count) of word occurrences is highly important.
- **Probability Model:**

$$P(\mathbf{x} \mid y) = \frac{\left(\sum_j x_j\right)!}{\prod_j x_j!} \prod_{j=1}^{V} [P(w_j \mid y)]^{x_j}$$

THANK YOU!