Toward Robust and Reliable Post-hoc Explanations to Machine Learning Models

Ngoc Bui - Graph and Geometric Deep Learning Group - Sep 2023

Why’s Explainable AI?

Input

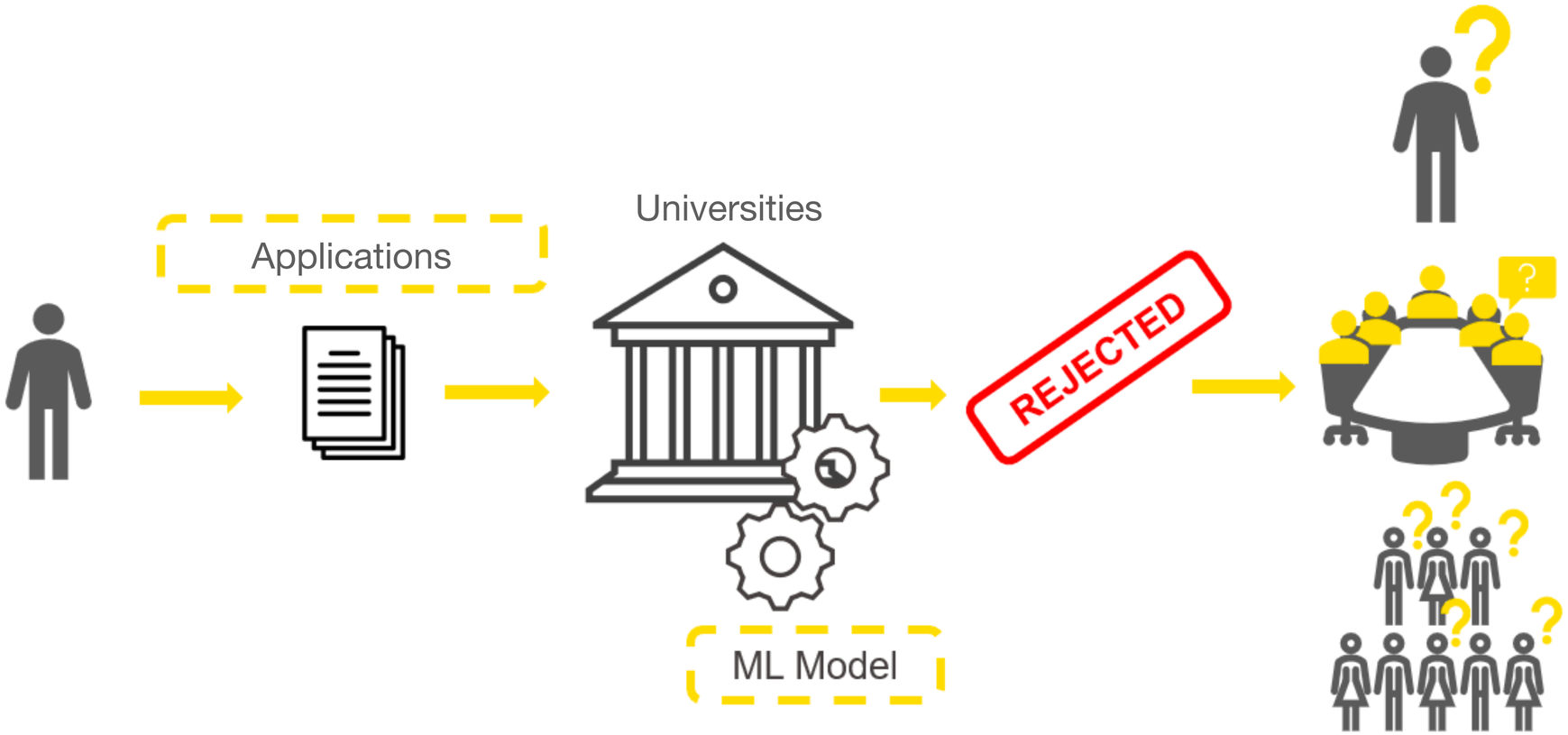
Output

Input is converted into output

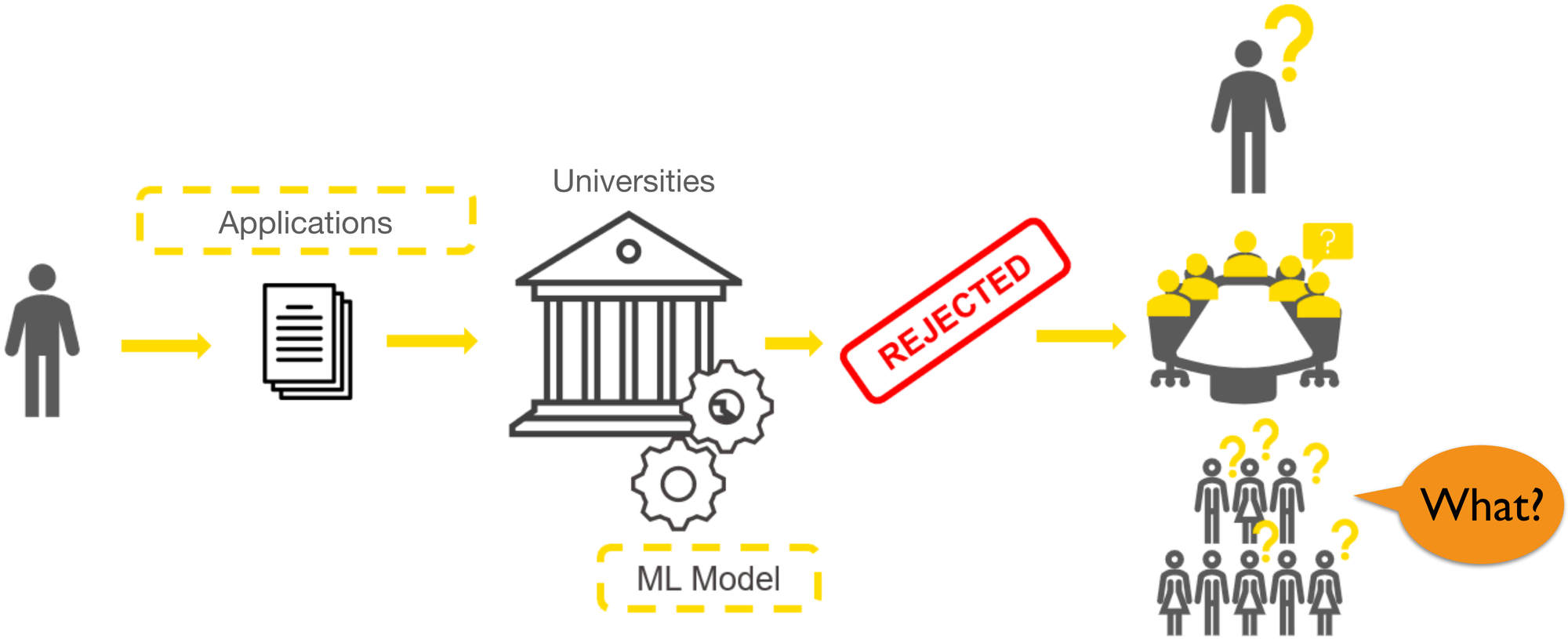
Why’s Explainable AI?

Input is converted into output

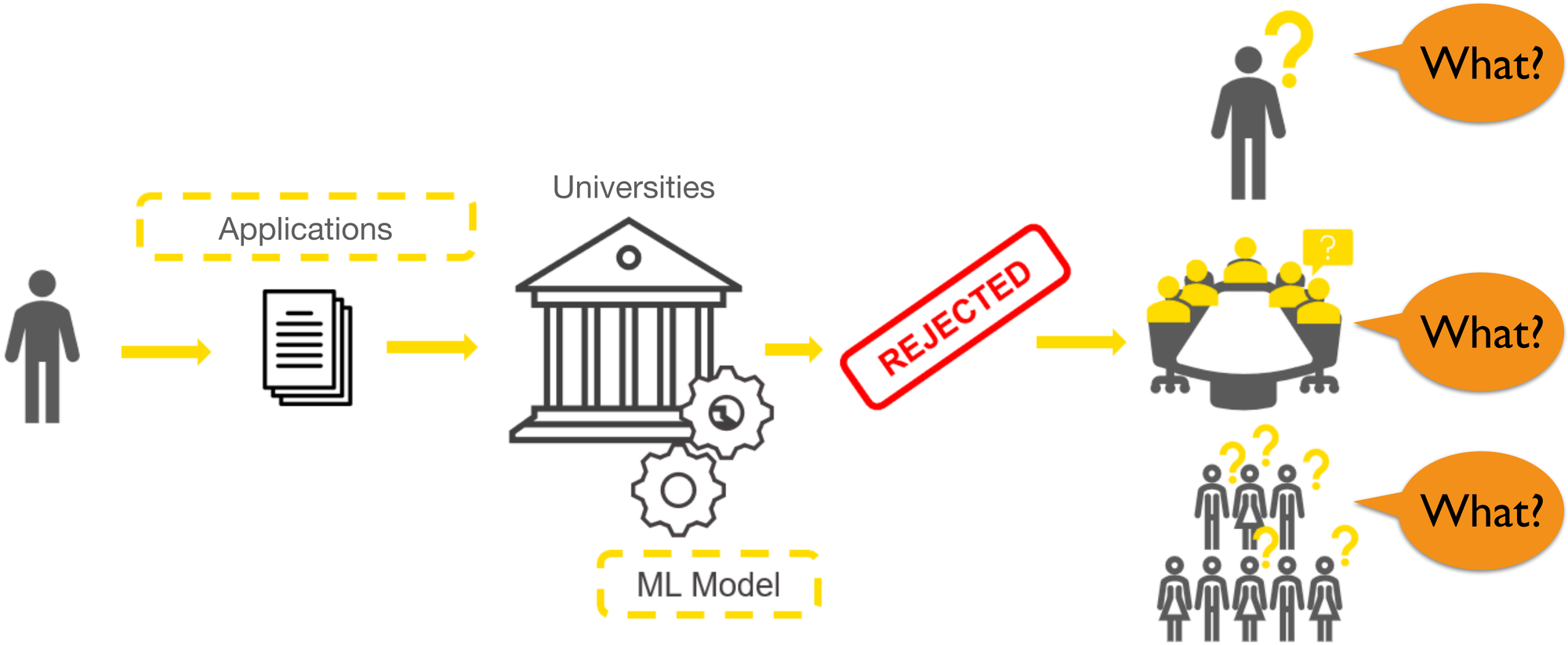
Why’s Explainable AI?



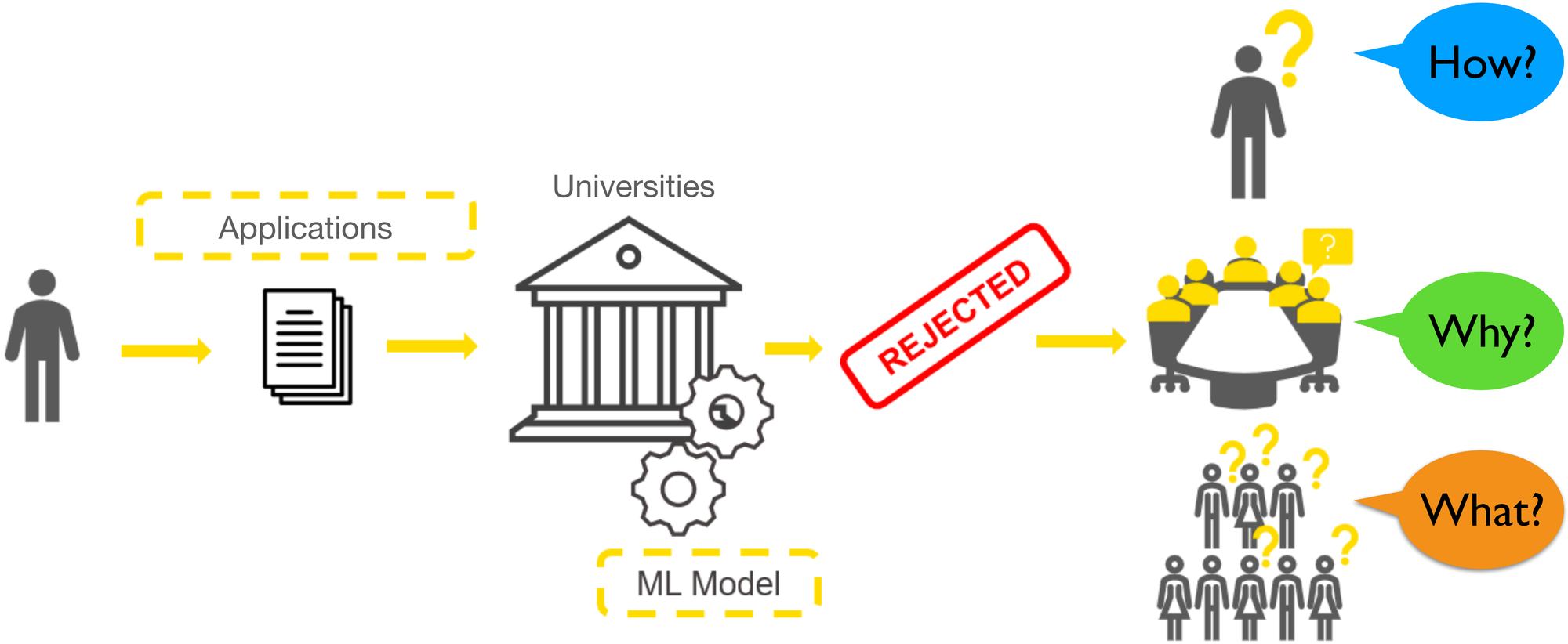
Why's Explainable AI?



Why's Explainable AI?

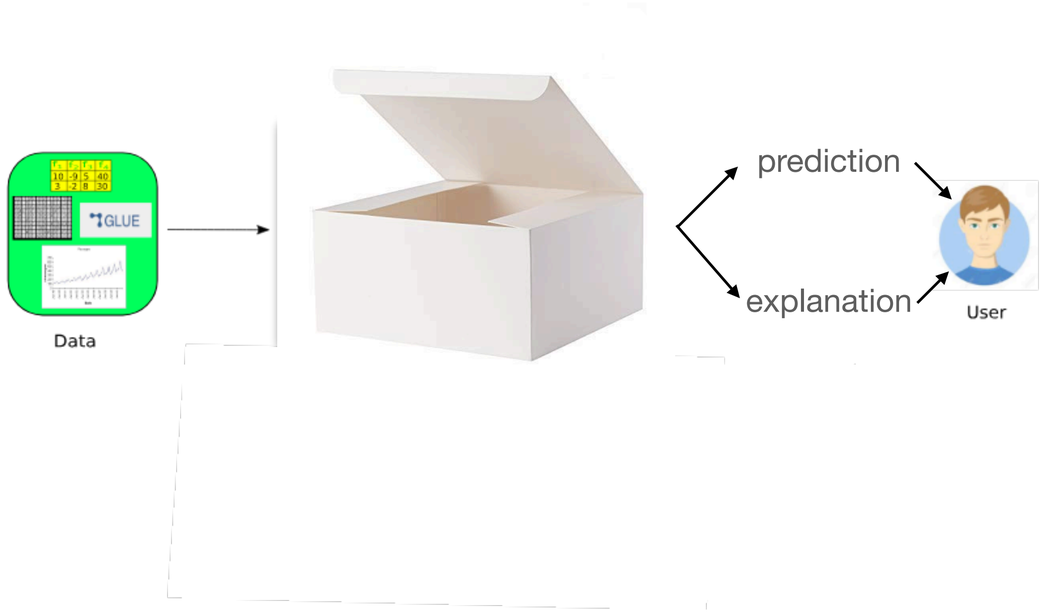


Why's Explainable AI?

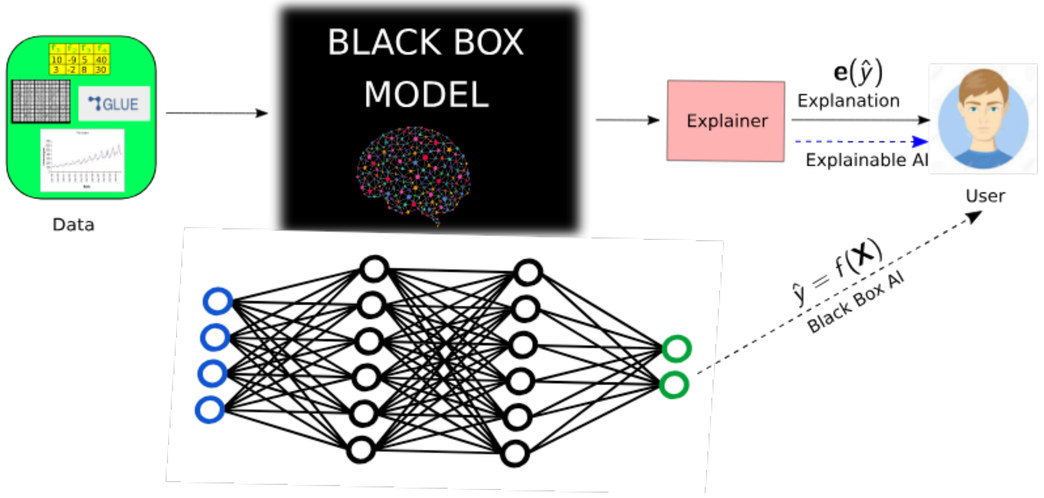


Explanations

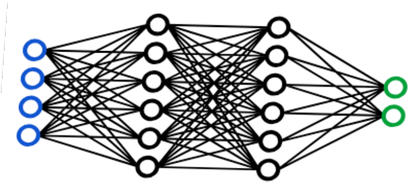
Ante-hoc Explanations



Post-hoc Explanations



Post-hoc Explanations



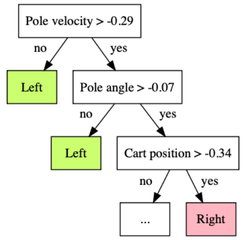
Explainers

Not today's talk

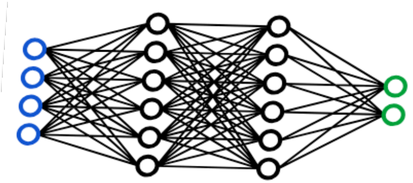
Global Explainers

Local Explainers

Model Distillation



Post-hoc Explanations



Explainers

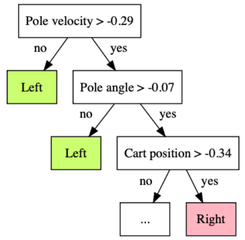
Input instance

Not today's talk

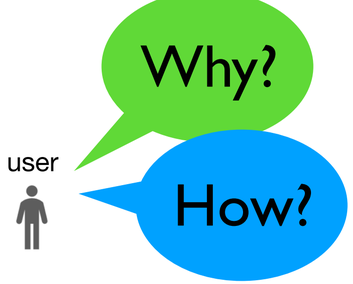
Global Explainers

Local Explainers

Model Distillation



Local Explanations



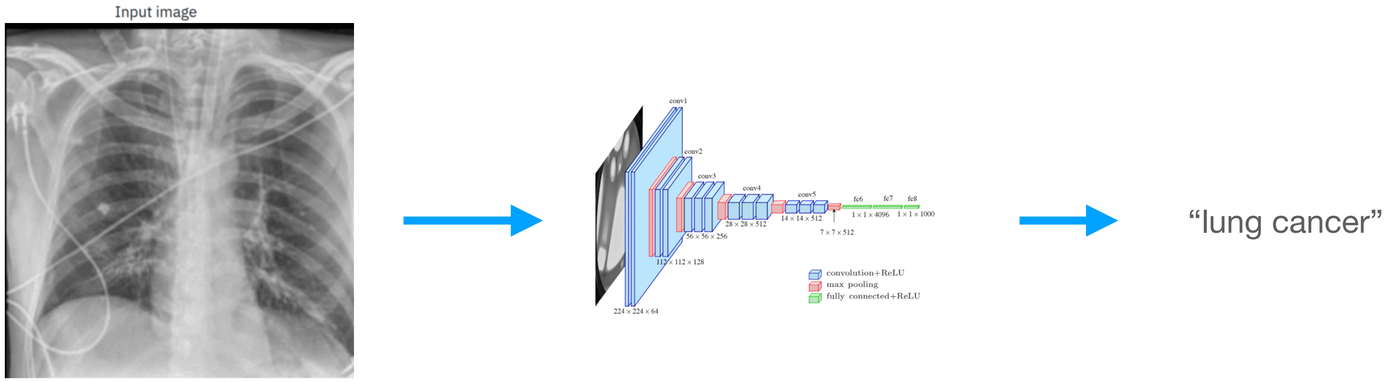
Local Explanations

Why?

user

Which features are important to make such a decision?

Saliency Map (e.g. GradCAM)



“lung cancer”

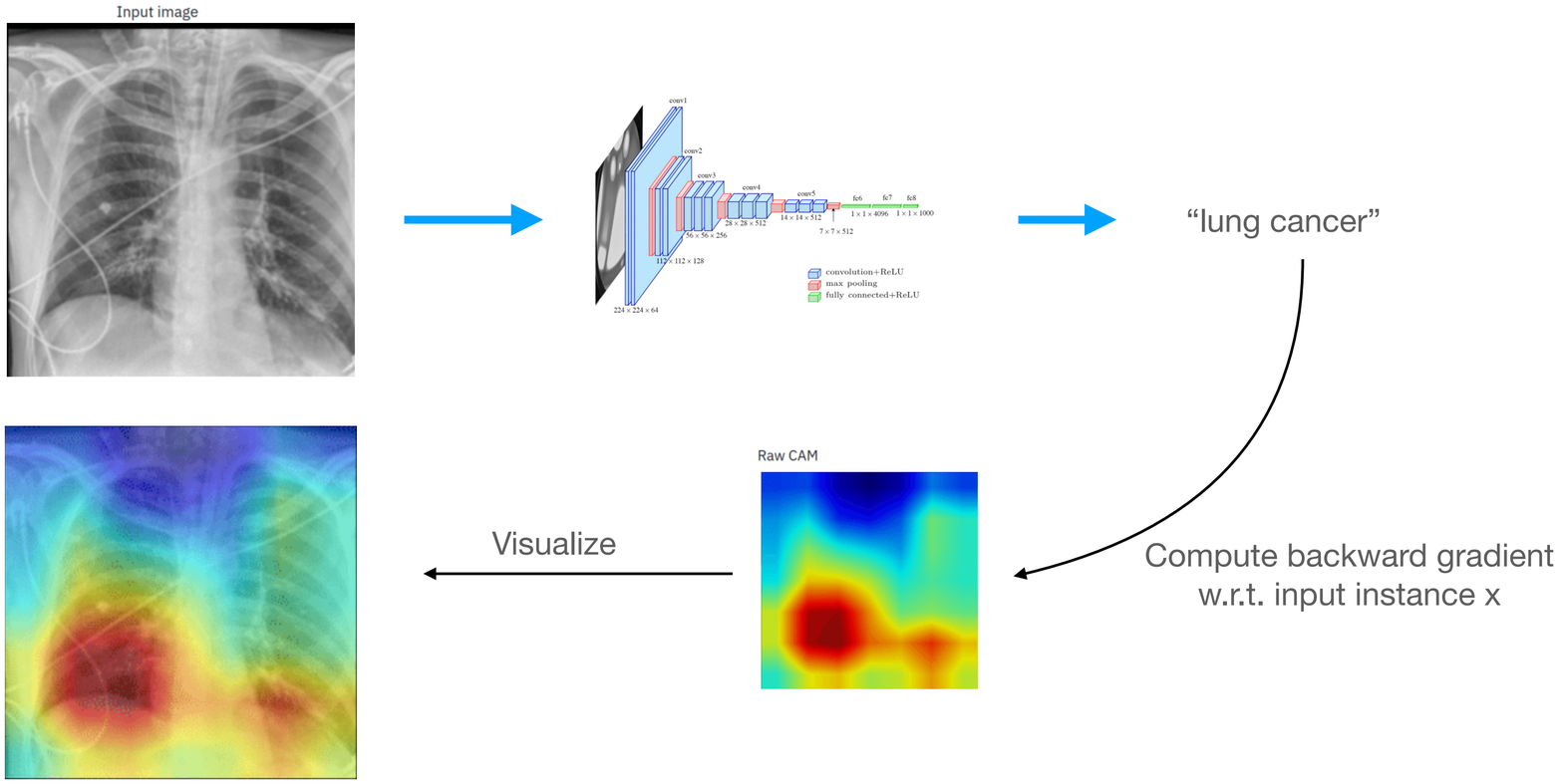
Local Explanations

Why?

user

Which features are important to make such a decision?

Saliency Map (e.g. GradCAM)



Local Explanations

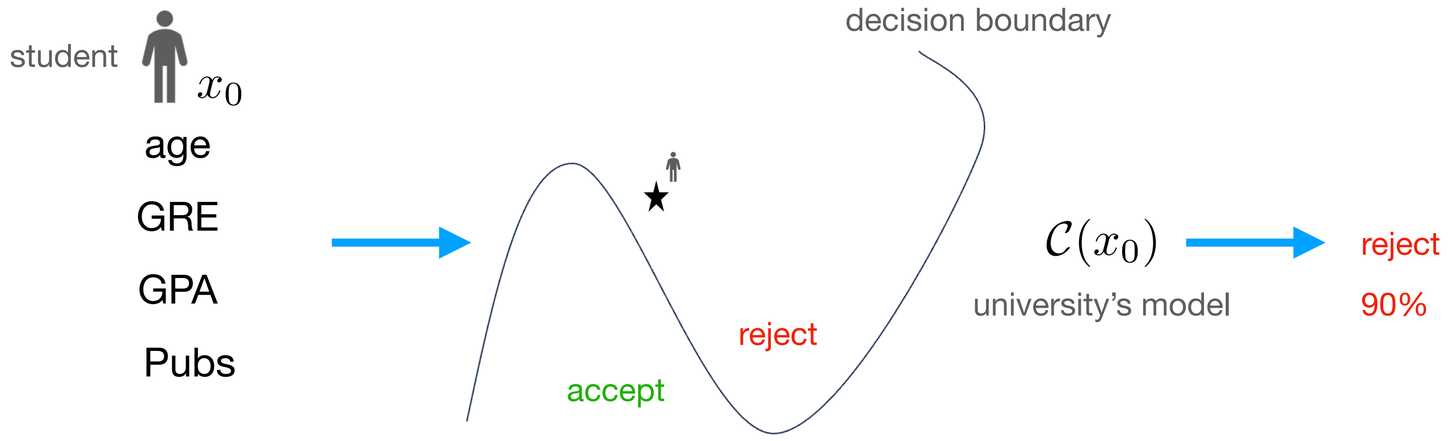
user

Why?

Which features are important to make such a decision?

Saliency Map (e.g. GradCAM)

Feature-based Attribution Methods (e.g. LIME, SHAP)



Local Explanations

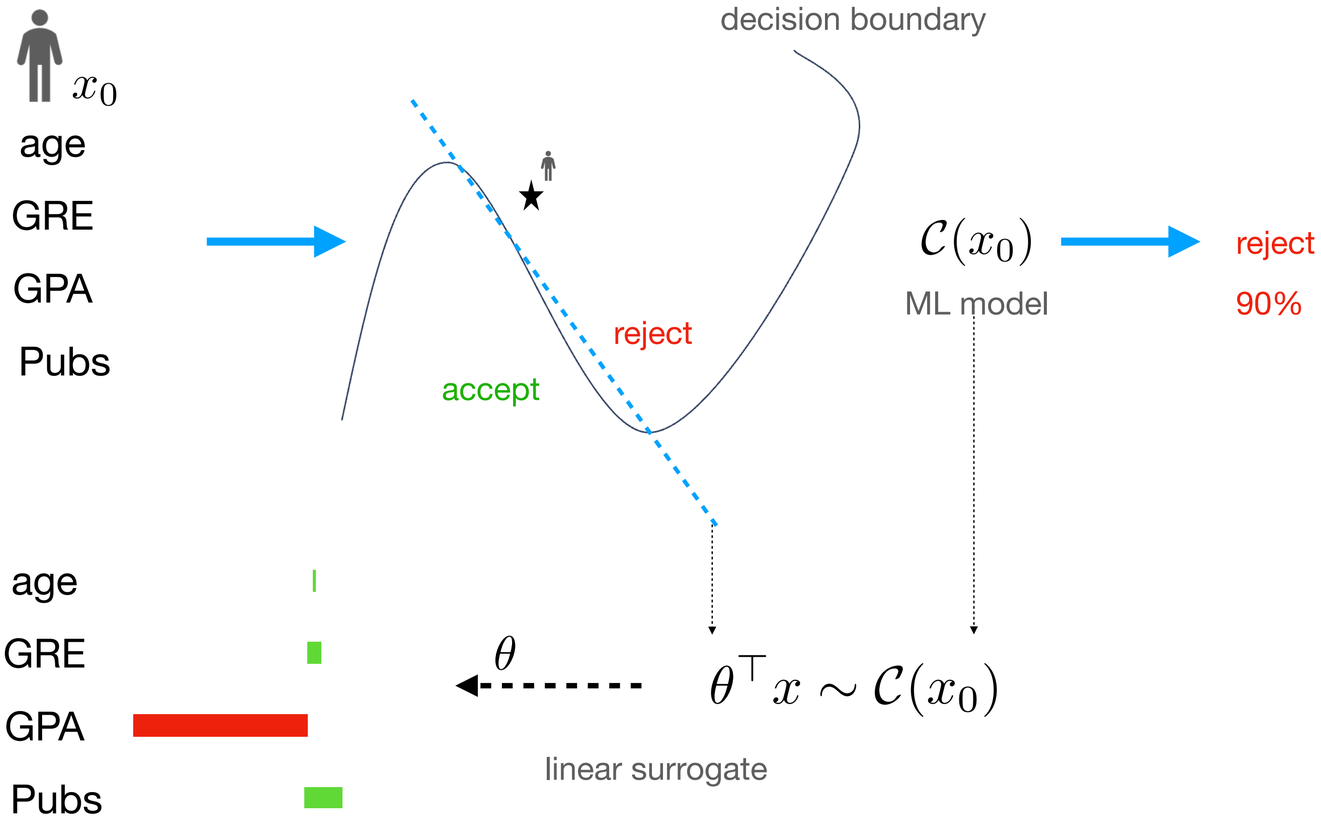
user

Why?

Which features are important to make such a decision?

Saliency Map (e.g. GradCAM)

Feature-based Attribution Methods (e.g. LIME, SHAP)

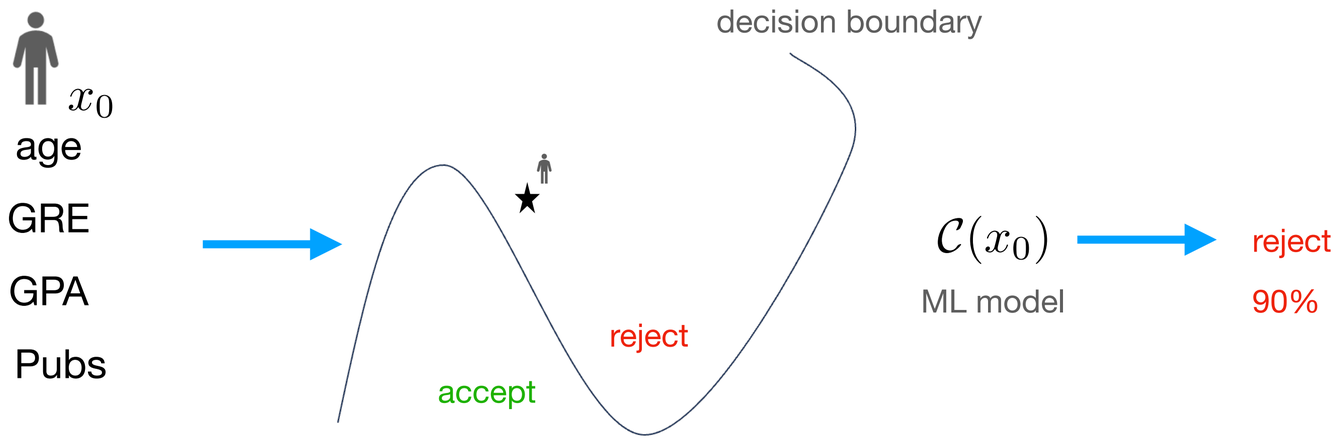


Local Explanations

How?

user

How can I improve to get a favorable result?



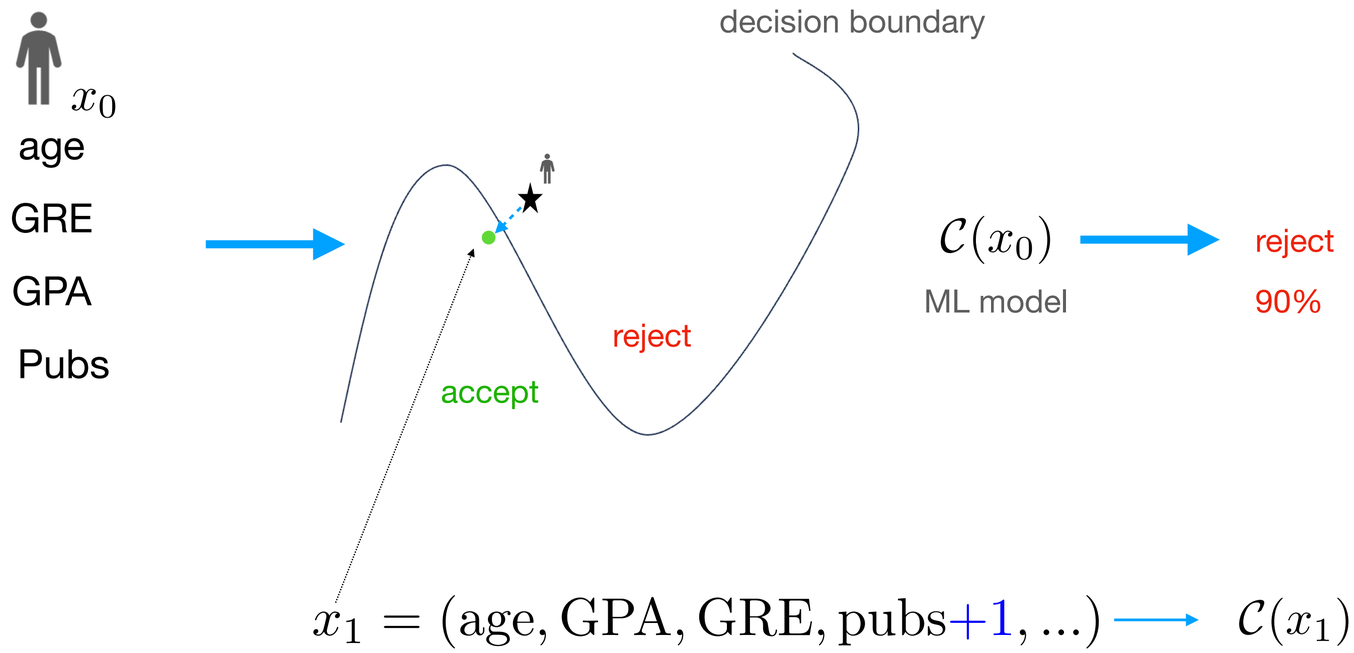
Local Explanations

How?

user

How can I improve to get a favorable result?

Counterfactual Explanations (a.k.a. Algorithmic Recourse):
examples that can flip the model's outcome



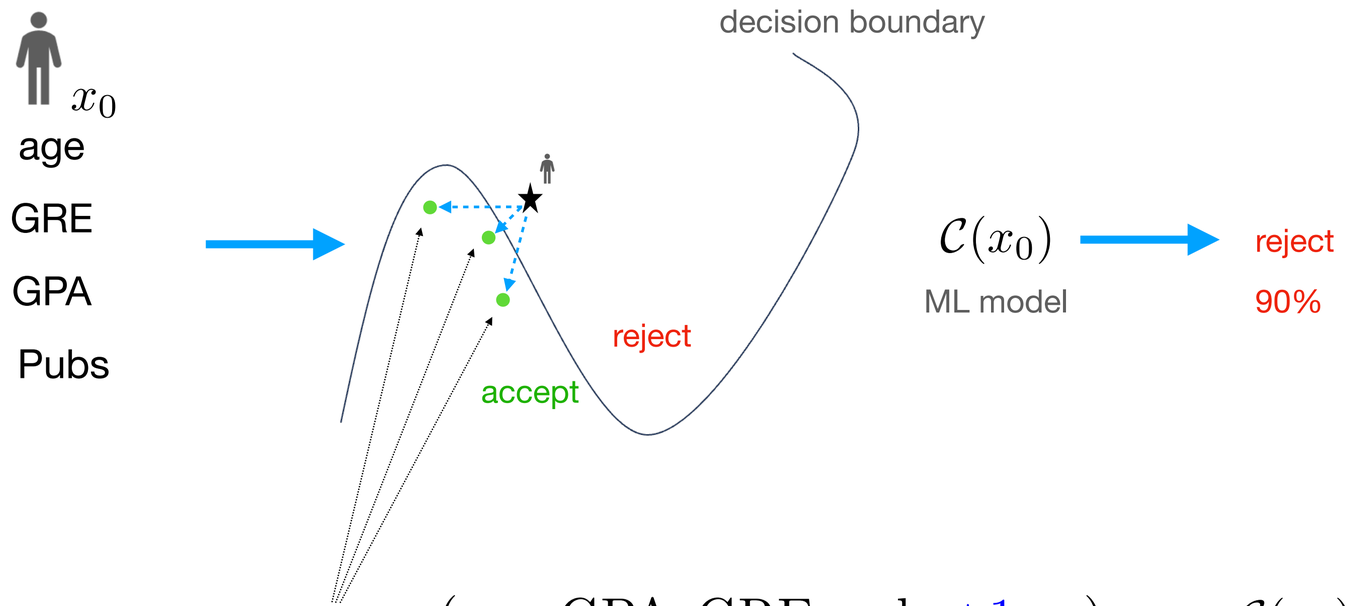
Local Explanations

How?

user

How can I improve to get a favorable result?

Counterfactual Plan: a set of counterfactual explanations



$x_1 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}{+}1, ...) \longrightarrow \mathcal{C}(x_1) \longrightarrow$ accept

$x_2 = (\text{age}, \text{GPA}, \text{GRE}{+}15, \text{pubs}, ...) \longrightarrow \mathcal{C}(x_2) \longrightarrow$ accept

$x_3 = (\text{age}, \text{GPA}{+}0.2, \text{GRE}, \text{pubs}, ...) \longrightarrow \mathcal{C}(x_3) \longrightarrow$ accept

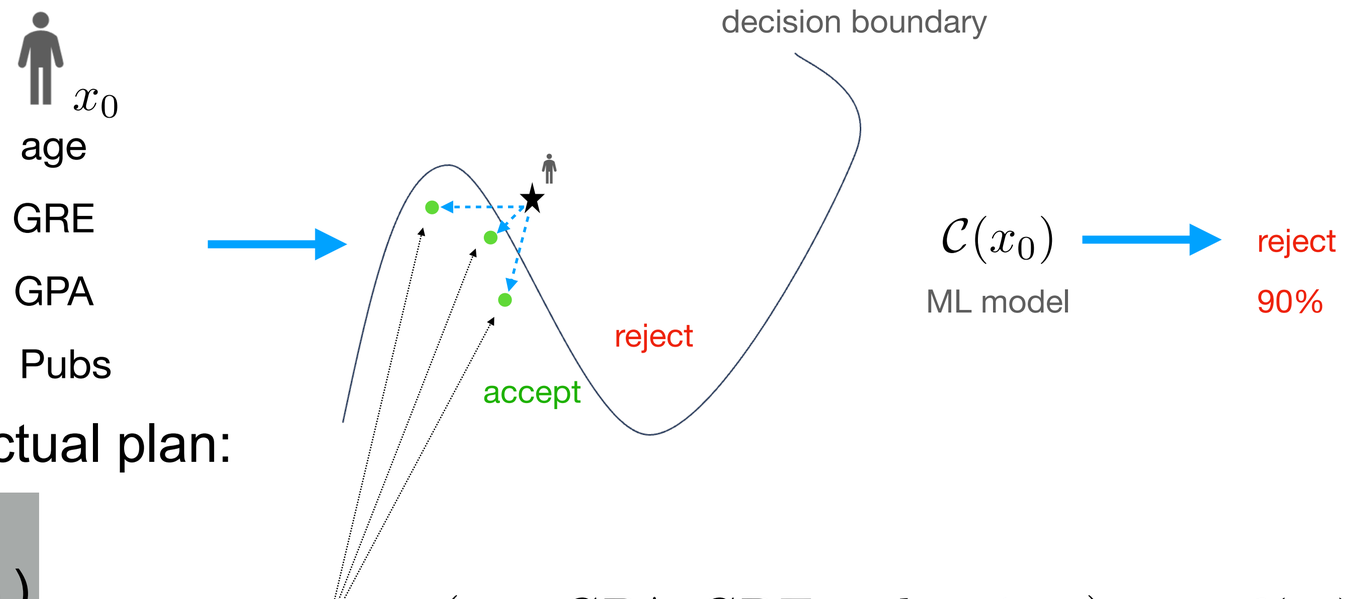
Local Explanations

How?

user

How can I improve to get a favorable result?

Counterfactual Plan: a set of counterfactual explanations



Criteria to construct counterfactual plan:

1. Validity
2. Proximity (closeness to x_0)
3. Diversity
4. Actionability

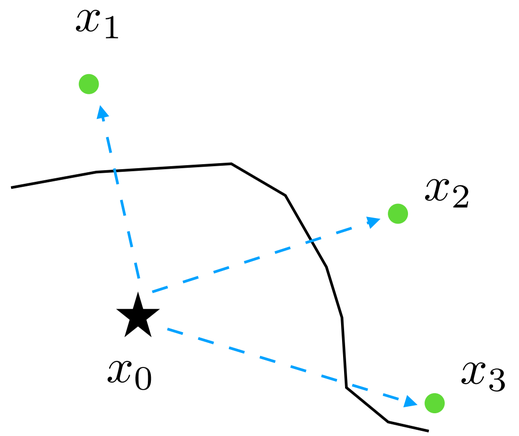
$x_1 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}+1, ...) \longrightarrow \mathcal{C}(x_1) \longrightarrow$ accept

$x_2 = (\text{age}, \text{GPA}, \text{GRE}+15, \text{pubs}, ...) \longrightarrow \mathcal{C}(x_2) \longrightarrow$ accept

$x_3 = (\text{age}, \text{GPA}+0.2, \text{GRE}, \text{pubs}, ...) \longrightarrow \mathcal{C}(x_3) \longrightarrow$ accept

Counterfactual Plan - Construction

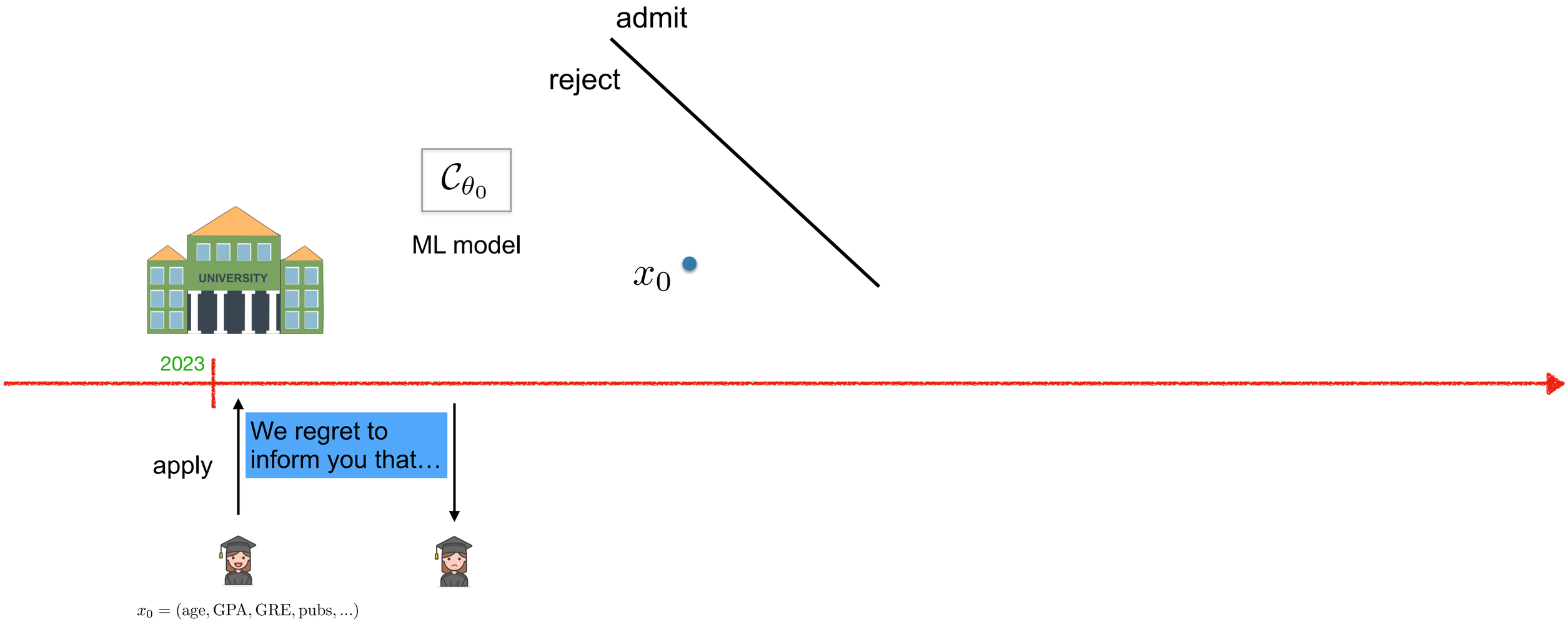
- Mixed-Integer Programming
  - Russell (2019) (for linear models)
    - Similar to Ustun (2018)
    - Promote diversity through diverse sets of constraints
- Multi-objective Evolutionary Algorithm
  - Dandl et al. (2020)
    - Universal but slow
    - Suitable to promote actionability
- Iterative, Gradient-based Optimization
  - Mothilal et al. (2020) (DICE)
    - A weighted sum of three objectives: validity, proximity, and diversity.
    - Solved by projected gradient descent



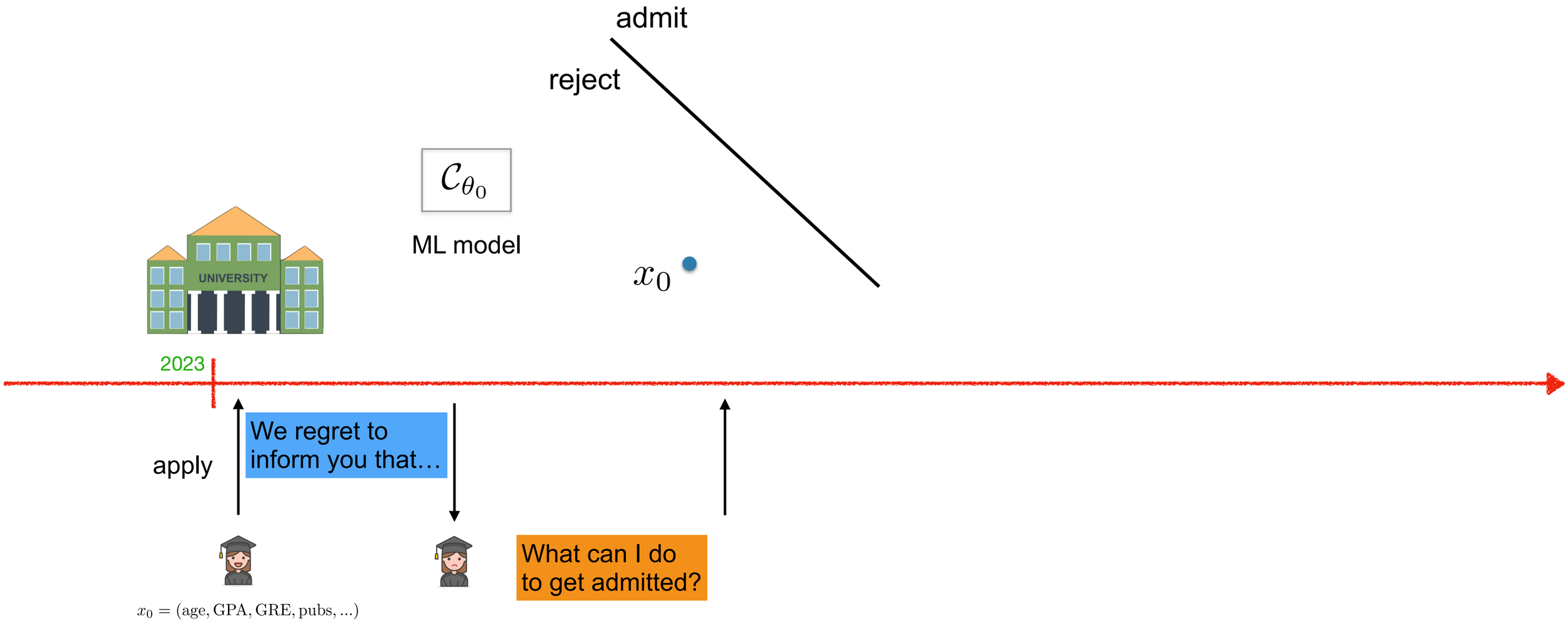
Counterfactual Explanations under Distributional Ambiguity

ICLR'22

Counterfactual Plan under Model Shifts



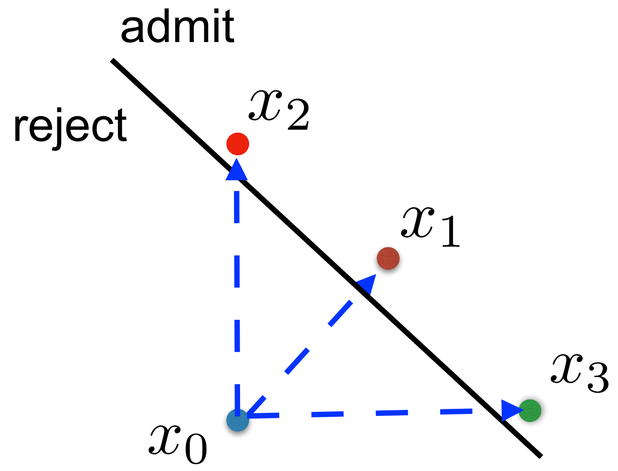
Counterfactual Plan under Model Shifts



Counterfactual Plan under Model Shifts

$\mathcal{C}_{\theta_0}$

ML model



2023

apply

We regret to inform you that…

$x_0 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}, ...)$

What can I do to get admitted?

$\{x_j\}_{j=1,\ldots,J}$

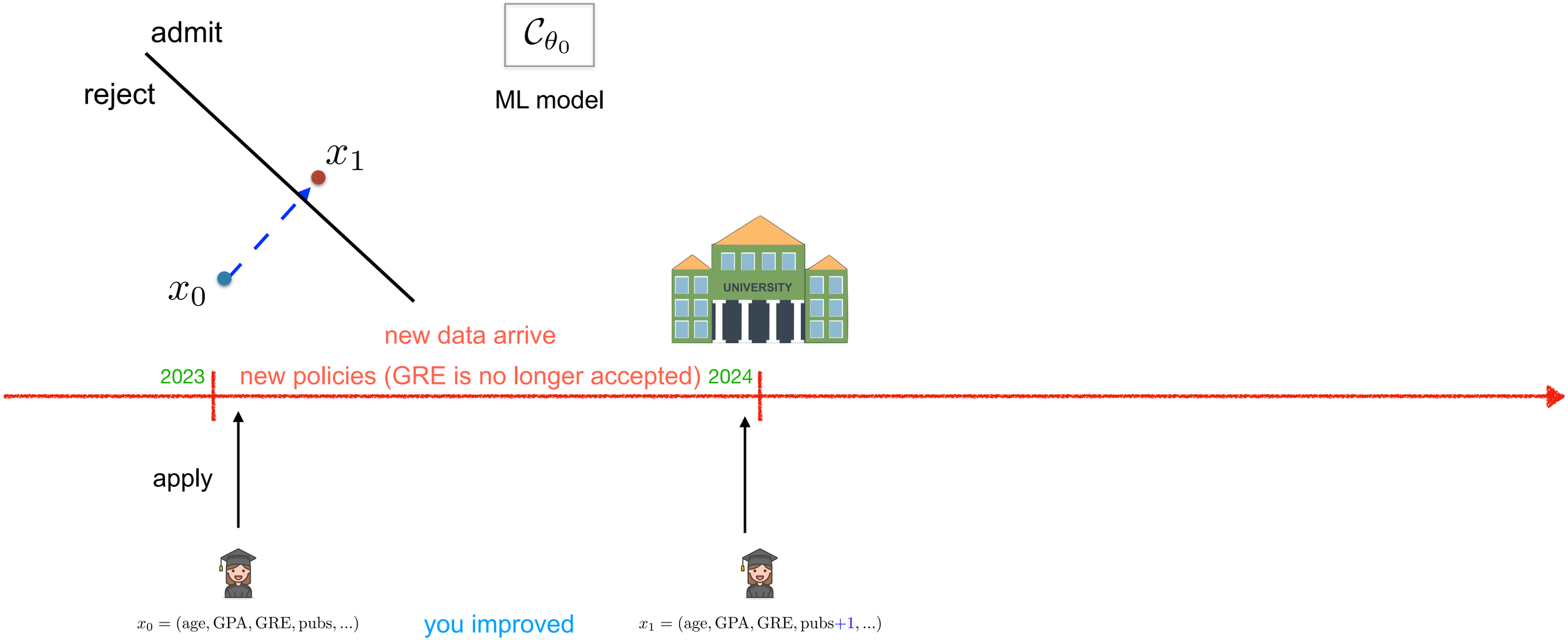
CF Plan

$x_1 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}{+}1, ...)$
$x_2 = (\text{age}, \text{GPA}, \text{GRE}{+}15, \text{pubs}, ...)$
$x_3 = (\text{age}, \text{GPA}{+}0.2, \text{GRE}, \text{pubs}, ...)$

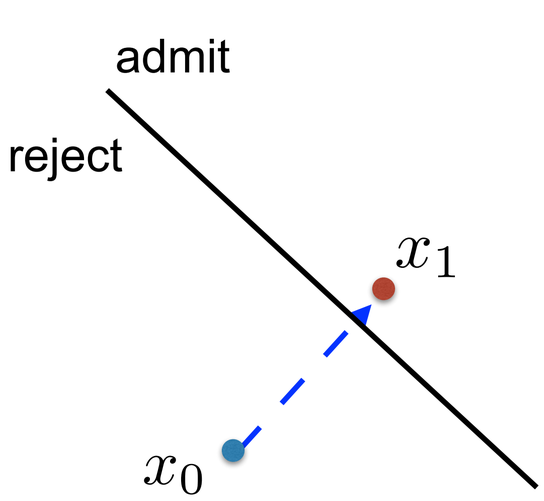
$x_1 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}{+}1, ...)$

Okay I will try to publish one more top-tier paper

Counterfactual Plan under Model Shifts



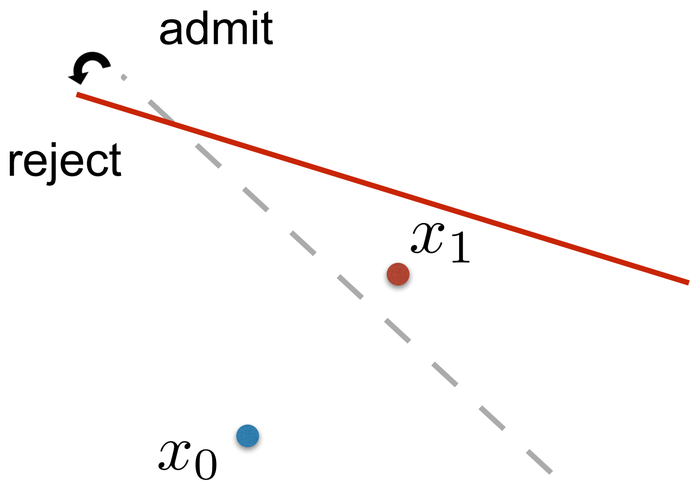
Counterfactual Plan under Model Shifts



$\mathcal{C}_{\theta_0}$ — retraining → $\mathcal{C}_{\tilde{\theta}}$

ML model (2023)

ML model



new data arrive

2023 new policies (GRE is no longer accepted) 2024

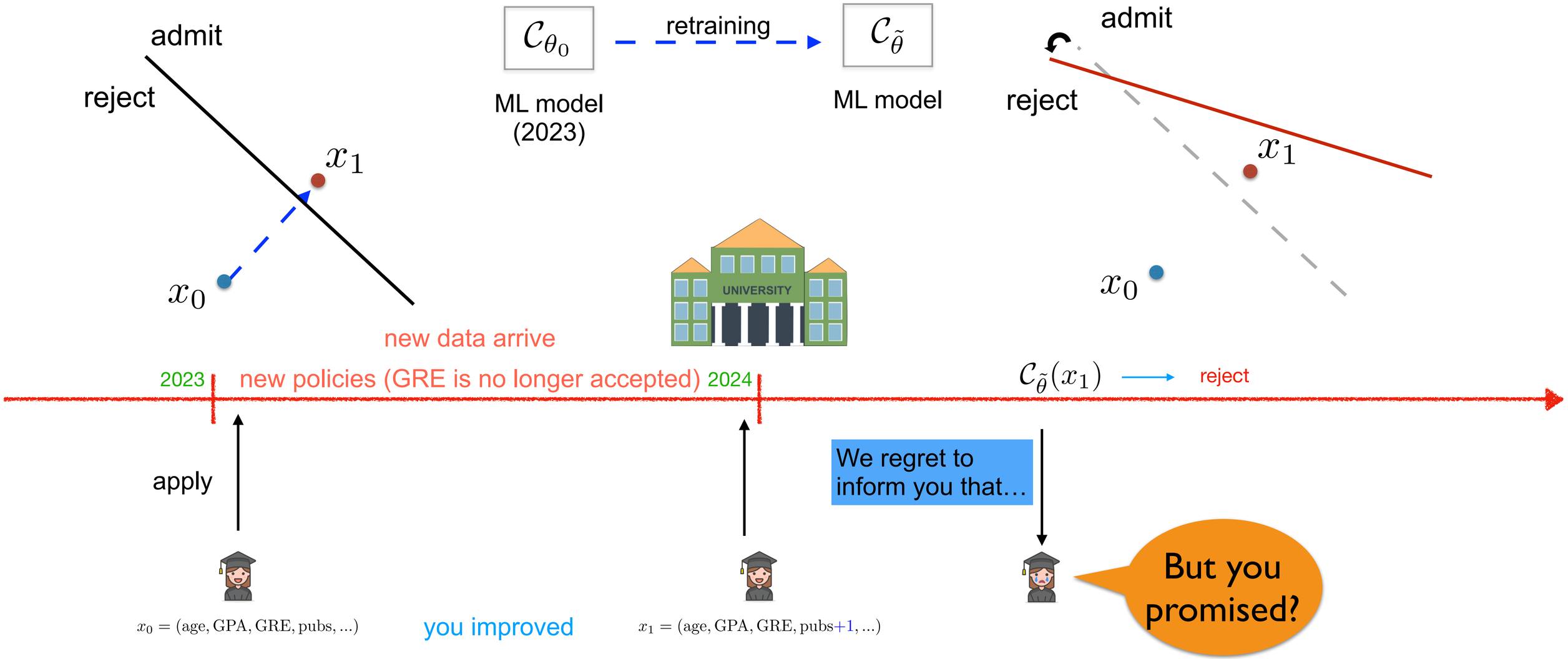
apply

$x_0 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}, ...)$

you improved

$x_1 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}{+}1, ...)$

Counterfactual Plan under Model Shifts



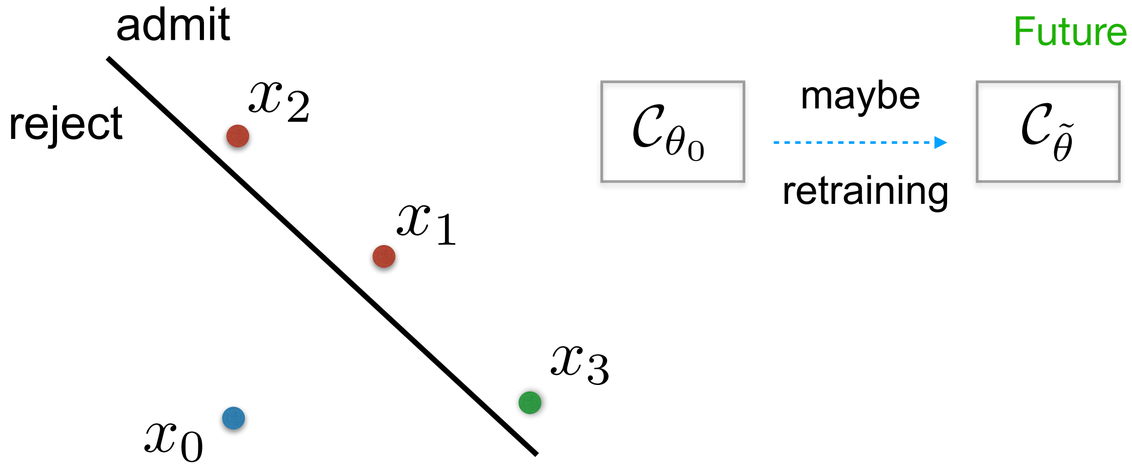
Settings

We are given:

- An ML model (assumed to be linear)

$$\mathcal{C}_{\theta_0}(x) = \begin{cases} 0 \Rightarrow \text{reject (unfavorable)} \\ 1 \Rightarrow \text{accept (favorable)} \end{cases}$$

- A CF plan $\{x_j\}_{j=1,\ldots,J}$



Definition: A plan $\{x_j\}_{j=1,\ldots,J}$ is valid w.r.t. $\tilde{\theta}$ if

$$\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \qquad \forall j = 1, \ldots, J$$

$x_0 = (\text{age}, \text{GPA}, \text{GRE}, \text{pubs}, \ldots)$

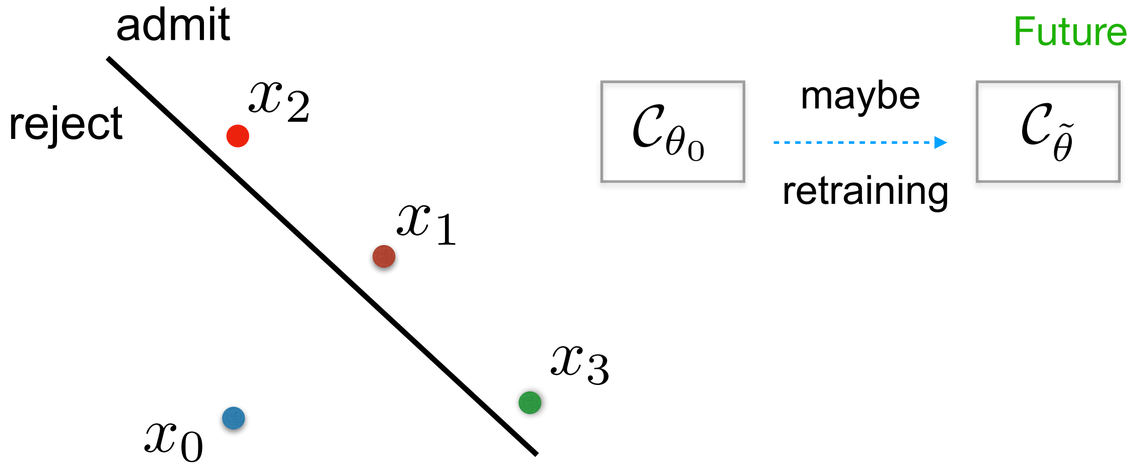
Settings

We are given:

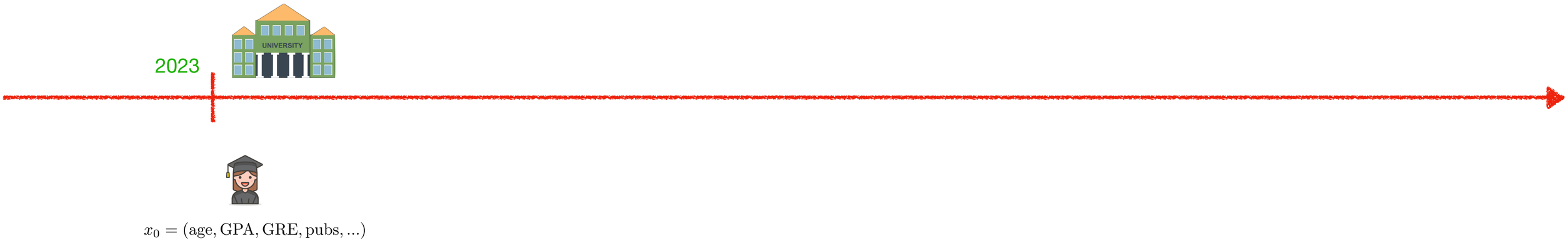
- An ML model (assumed to be linear)

$$\mathcal{C}_{\theta_0}(x) = \begin{cases} 0 \Rightarrow \text{reject (unfavorable)} \\ 1 \Rightarrow \text{accept (favorable)} \end{cases}$$

- A CF plan $\{x_j\}_{j=1,\ldots,J}$



Q1. Given $\{x_j\}$, what is the probability that the plan is valid (w.r.t. $\tilde{\theta}$)?
Q2. Can we construct a counterfactual plan that is robust to model shifts?



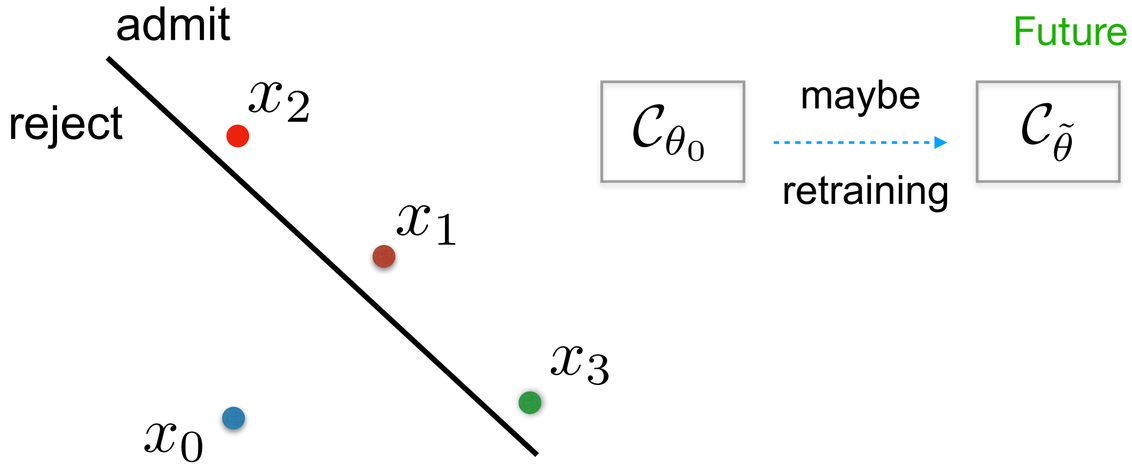
Settings

We are given:

- An ML model (assumed to be linear)

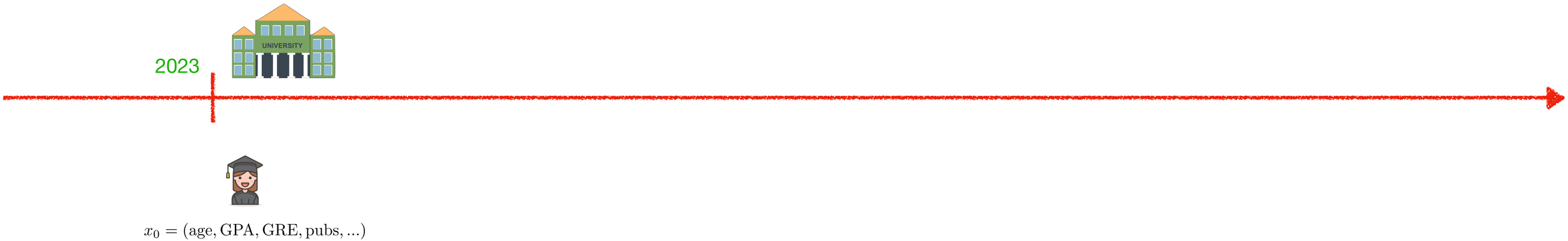
$$\mathcal{C}_{\theta_0}(x) = \begin{cases} 0 \Rightarrow \text{reject (unfavorable)} \\ 1 \Rightarrow \text{accept (favorable)} \end{cases}$$

- A CF plan $\{x_j\}_{j=1,\ldots,J}$

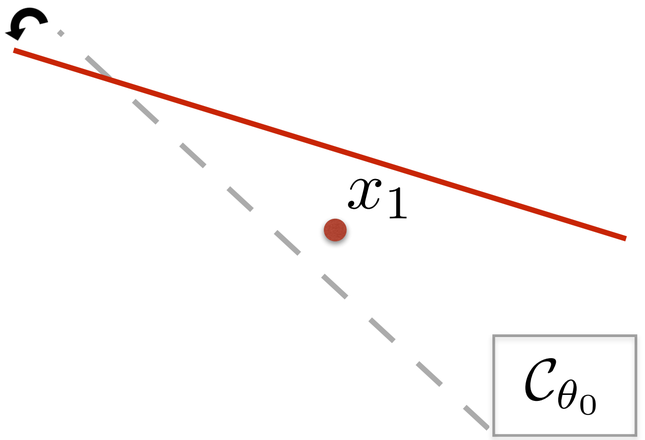


We don't know it

Q1. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?
Q2. How can we construct a counterfactual plan that is robust to model shifts?



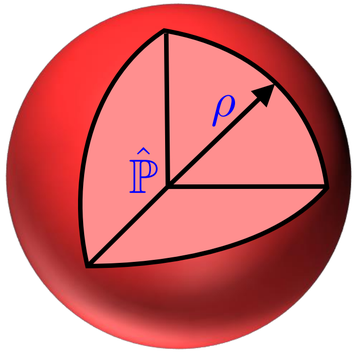
Uncertainty Modeling



Assume that the *future* parameter $\tilde{\theta} \sim \mathbb{P} \in \mathbb{B} = \{\mathbb{P} : \mathbb{G}(\mathbb{P}, \hat{\mathbb{P}}) \leq \rho\}$

Ambiguity set $\mathbb{B}$

- neighborhood around $\hat{\mathbb{P}} \sim (\hat{\mu}, \hat{\Sigma})$
- aims to contain the true distribution $\mathbb{P}$ with a high probability



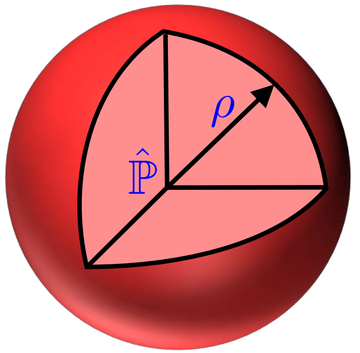
$\hat{\mathbb{P}} \sim (\hat{\mu}, \hat{\Sigma})$: estimated distribution (e.g. $\widehat{\mu} = \theta_0, \widehat{\Sigma} = I$)

$\mathbb{G}(\cdot\ , \cdot)$: distance between 2 distributions

ρ : uncertainty radius (controllable hyperparameter)

Uncertainty Modeling

$$\tilde{\theta} \sim \mathbb{P} \in \mathbb{B} = \{\mathbb{P} : \mathbb{G}(\mathbb{P}, \hat{\mathbb{P}}) \leq \rho\}$$



$\mathbb{G}(\cdot\,,\cdot)$: Gelbrich distance between two distributions

Definition: For $\mathbb{Q}_1 \sim (\mu_1, \Sigma_1)$ and $\mathbb{Q}_2 \sim (\mu_2, \Sigma_2)$

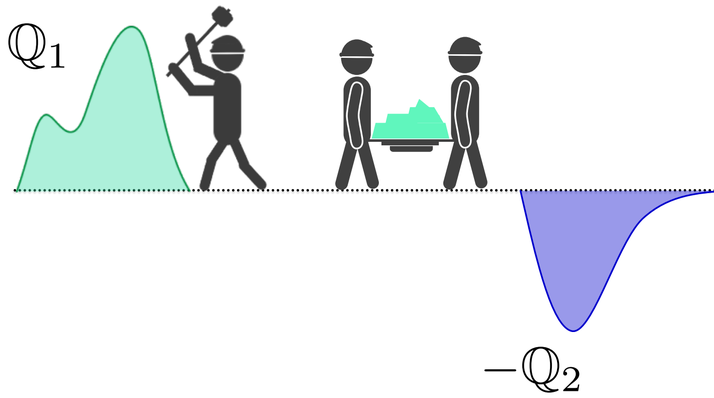
$$\mathbb{G}(\mathbb{Q}_1, \mathbb{Q}_2) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{\frac{1}{2}} \Sigma_1 \Sigma_2^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}$$

Proposition (Connection to Type-2 Wasserstein distance)**:**

For $\mathbb{Q}_1 \sim (\mu_1, \Sigma_1)$ and $\mathbb{Q}_2 \sim (\mu_2, \Sigma_2)$

$$W(\mathbb{Q}_1, \mathbb{Q}_2) \geq \mathbb{G}(\mathbb{Q}_1, \mathbb{Q}_2)$$

$W^2(\mathbb{Q}_1, \mathbb{Q}_2) =$ minimum cost of moving $\mathbb{Q}_1$ to $\mathbb{Q}_2$



Gelbrich distance has a nice connection to Wasserstein distance but is much easier to compute.

Gelbrich (1990)

Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j)$$

True validity

Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity

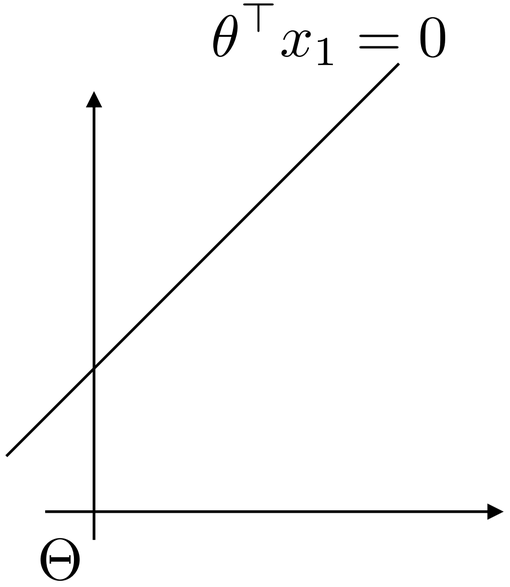
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity



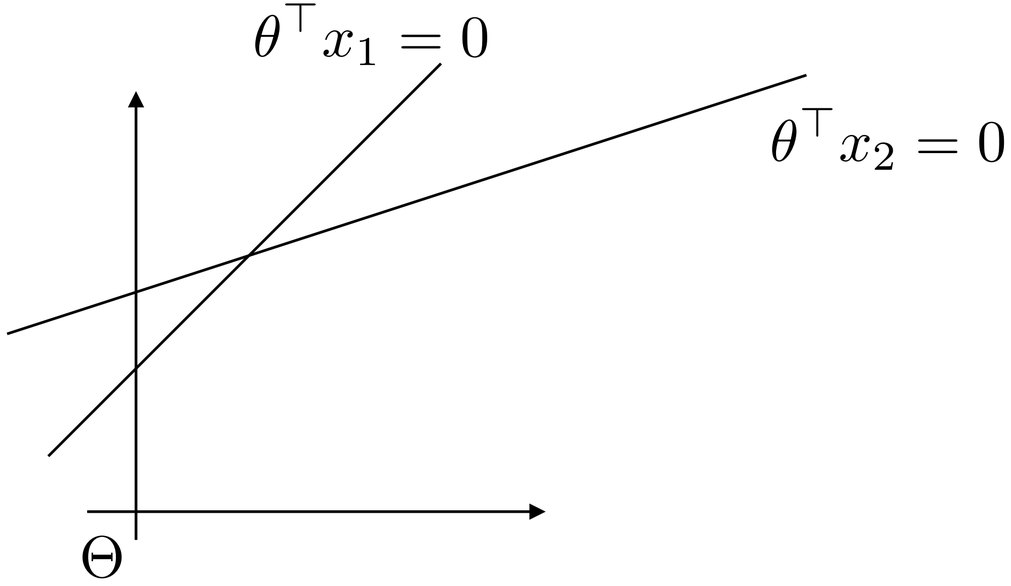
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity



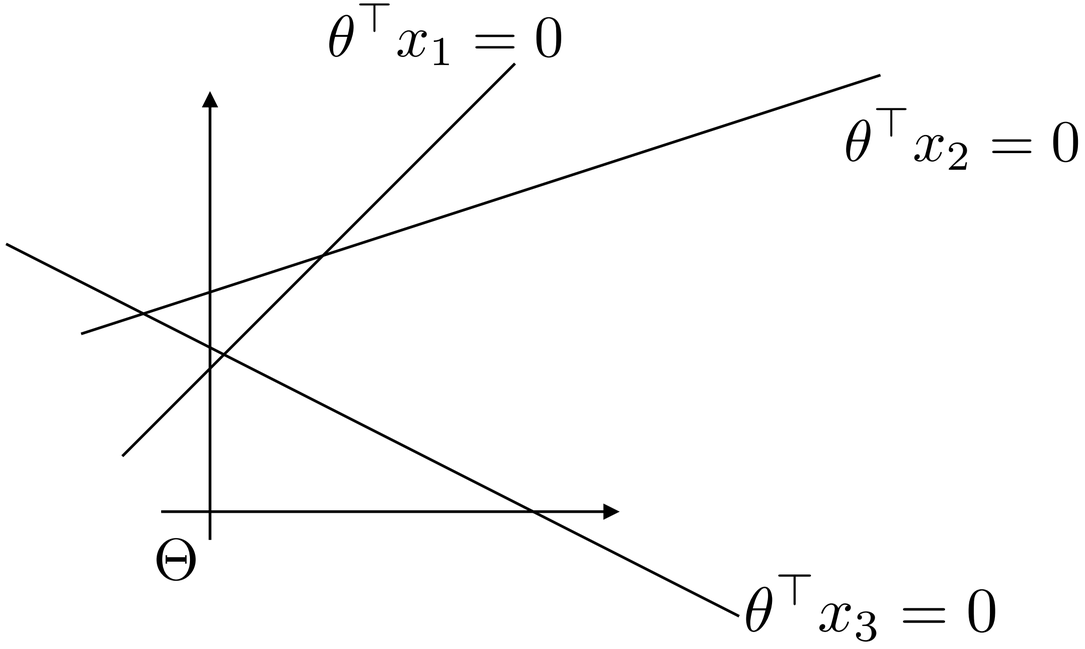
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity



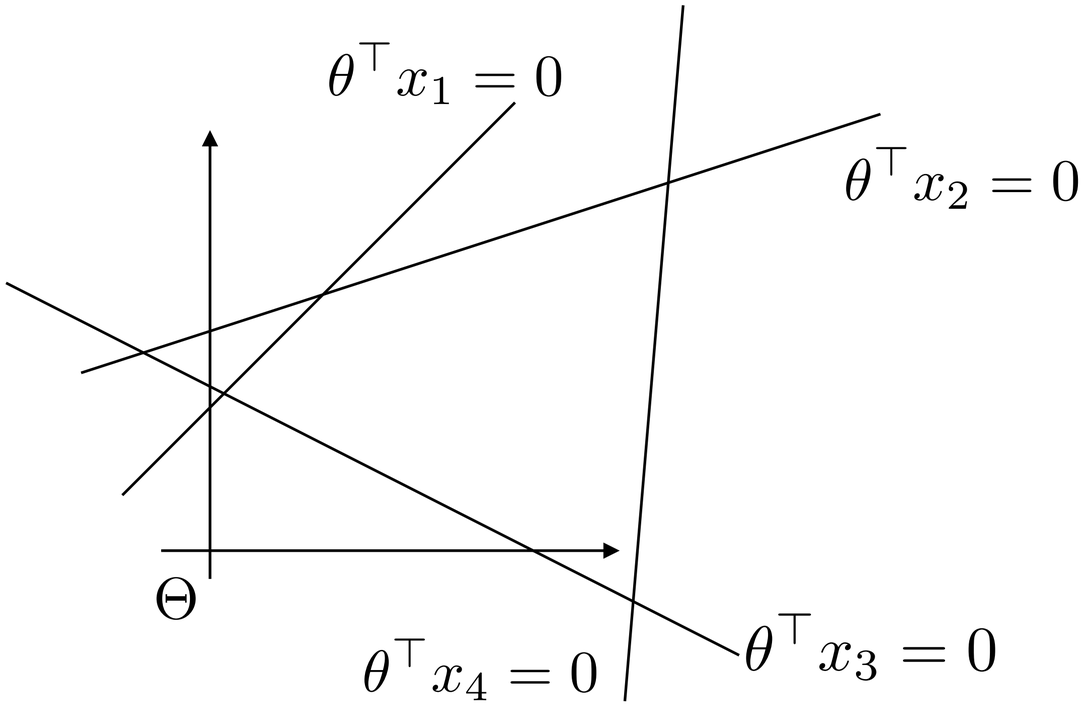
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity

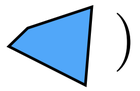


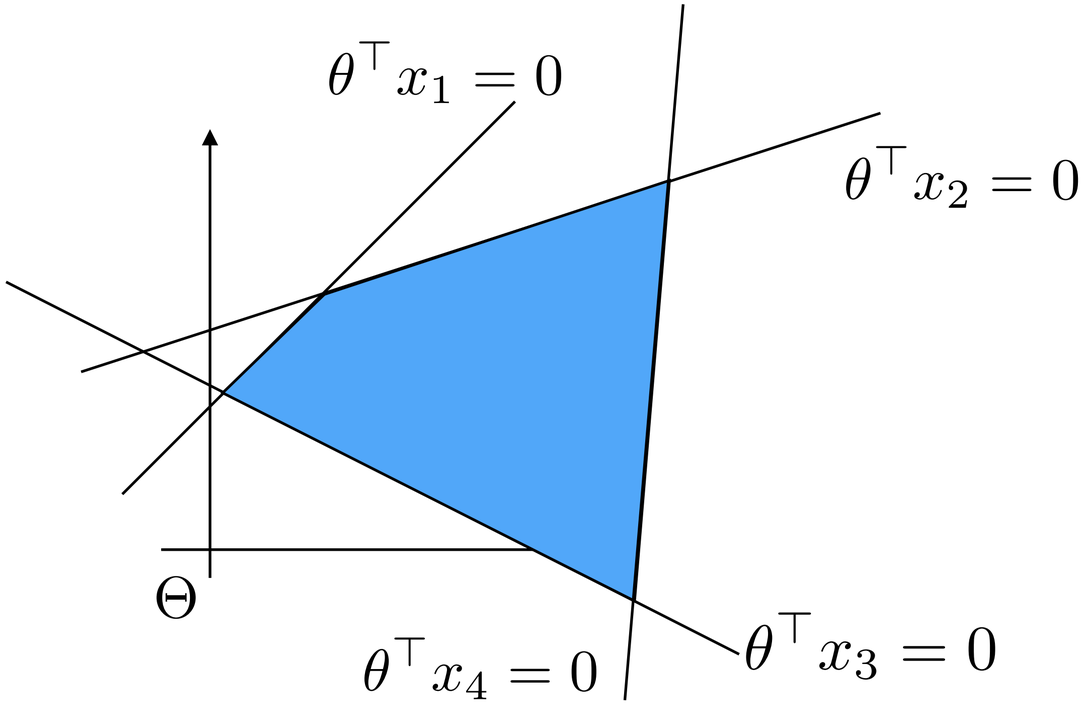
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity $= \mathbb{P}(\tilde{\theta} \in$  $)$



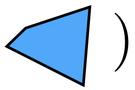
Vandenberghe et al. (2007)

Uncertainty Quantification

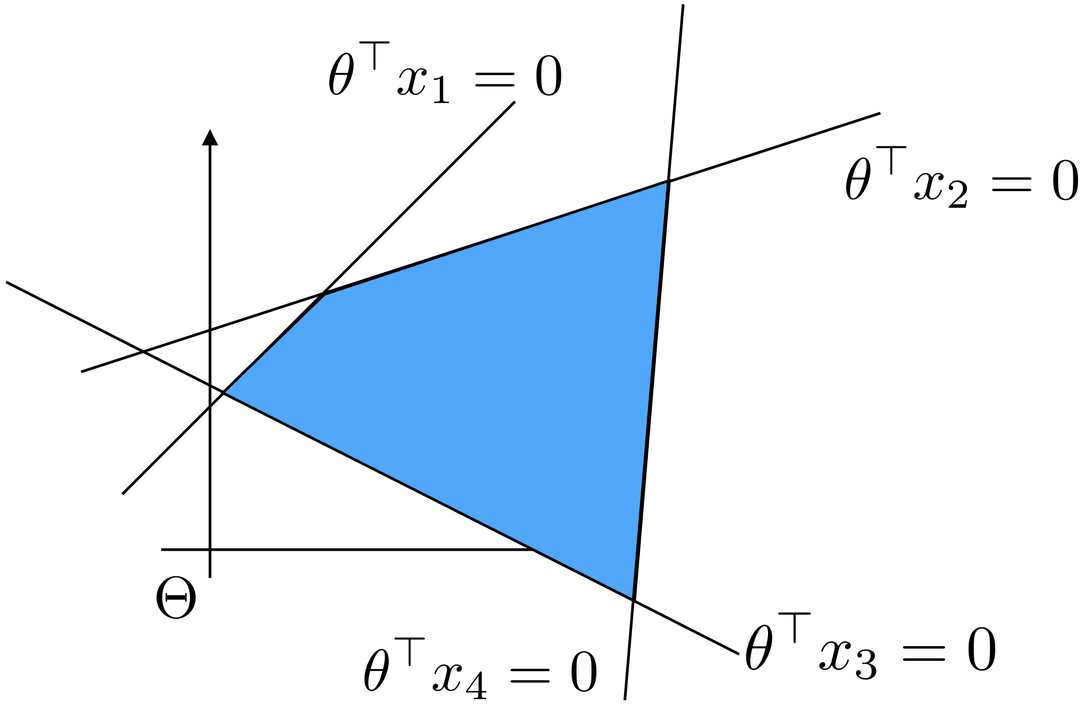
- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

True validity $= \mathbb{P}(\tilde{\theta} \in$  $)$

We don't know



Vandenberghe et al. (2007)

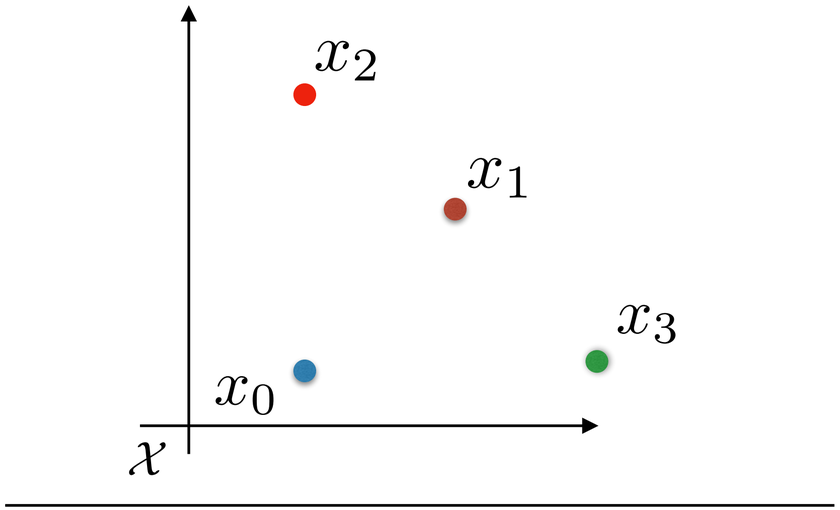
Uncertainty Quantification

- **Q1**. Given $\{x_j\}$, what is the probability that it is a valid plan (w.r.t. $\tilde{\theta}$)?

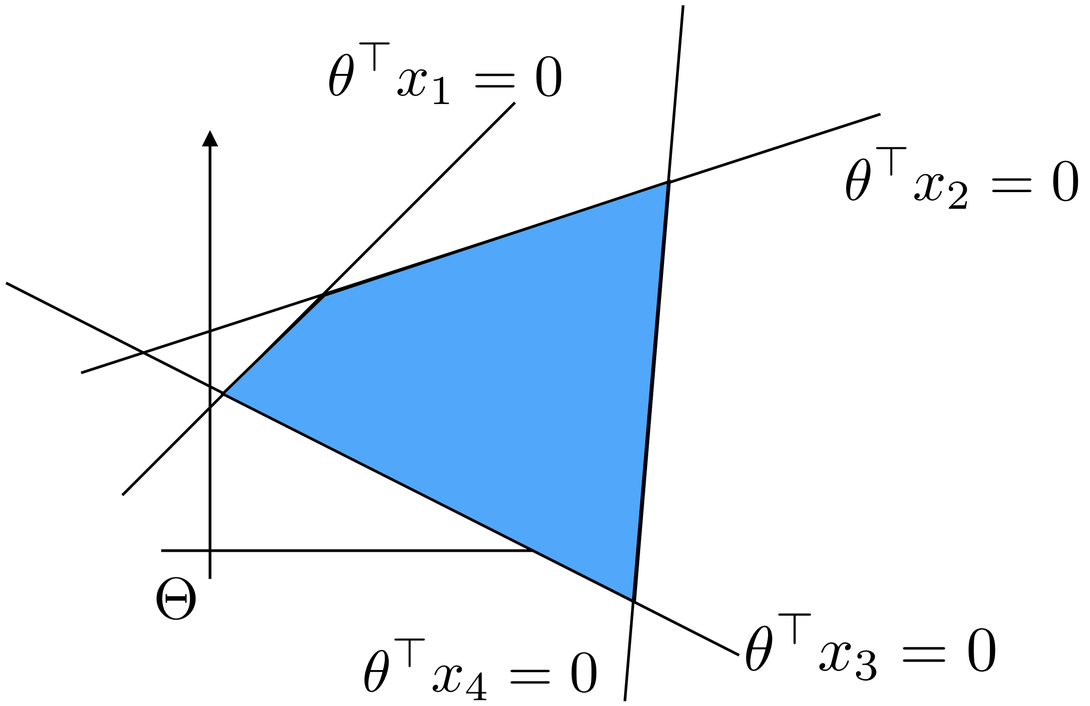
$$\mathcal{C}_\theta(x) = \begin{cases} 0 & \text{if } \theta^\top x < 0 \\ 1 & \text{if } \theta^\top x \geq 0 \end{cases}$$

$$\mathbb{P}(\mathcal{C}_{\tilde{\theta}}(x_j) = 1 \ \ \forall j) \quad = \mathbb{P}(\tilde{\theta}^\top x_j \geq 0 \ \ \forall j)$$

$$\text{True validity} \qquad = \mathbb{P}(\tilde{\theta} \in \ \blacktriangleleft \ )$$



$\sim$

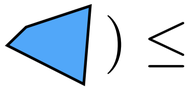
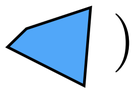
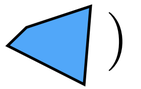


Vandenberghe et al. (2007)

Uncertainty Quantification

Assume that the *future* parameter $\tilde{\theta} \sim \mathbb{P} \in \mathbb{B} = \{\mathbb{P} : \mathbb{G}(\mathbb{P}, \hat{\mathbb{P}}) \leq \rho\}$

Lower bound $\leq$ True validity $\leq$ Upper bound

$\inf_{\mathbb{Q} \in \mathbb{B}} \mathbb{Q}(\tilde{\theta} \in$  $) \leq$ $\mathbb{P}(\tilde{\theta} \in$  $) \leq \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{Q}(\tilde{\theta} \in$  $)$

True validity of plan $\{x_j\}_{j=1,\ldots,J}$

Uncertainty Quantification

Assume that the *future* parameter $\tilde{\theta} \sim \mathbb{P} \in \mathbb{B} = \{\mathbb{P} : \mathbb{G}(\mathbb{P}, \hat{\mathbb{P}}) \leq \rho\}$

Lower bound $\leq$ True validity $\leq$ Upper bound

$$L^\star \leq \inf_{\mathbb{Q} \in \mathbb{B}} \mathbb{Q}(\tilde{\theta} \in \blacktriangleleft) \leq \mathbb{P}(\tilde{\theta} \in \blacktriangleleft) \leq \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{Q}(\tilde{\theta} \in \blacktriangleleft) \leq U^\star$$

True validity of plan $\{x_j\}_{j=1,\ldots,J}$

Semi-definite program

Semi-definite program

Theorem: $L^\star$ and $U^\star$ can be computed by solving *linear* semidefinite programs.

- Computational tractability (solved effectively by MOSEK)
- Complementary information: either $L^\star = 0$ or $U^\star = 1$

Uncertainty Quantification

Theorem: We have

$$
L^\star = \left\{ \begin{array}{ll}
\inf & 1 - \sum_{j\in[J]} \lambda_j \\
\text{s.t.} & \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{S}_+^d,\ C \in \mathbb{R}^{d\times d},\ M \in \mathbb{S}_+^d \\
& \lambda_j \in \mathbb{R},\ z_j \in \mathbb{R}^d,\ Z_j \in \mathbb{S}^d \ \ \forall j \in [J] \\
& -x_j^\top z_j \geq 0, \ \begin{bmatrix} Z_j & z_j \\ z_j^\top & \lambda_j \end{bmatrix} \succeq 0 \qquad \forall j \in [J] \\
& \displaystyle\sum_{j\in[J]} \begin{bmatrix} Z_j & z_j \\ z_j^\top & \lambda_j \end{bmatrix} \preceq \begin{bmatrix} M & \mu \\ \mu^\top & 1 \end{bmatrix}, \ \begin{bmatrix} \Sigma & C \\ C^\top & \hat{\Sigma}_1 \end{bmatrix} \succeq 0, \ \begin{bmatrix} M-\Sigma & \mu \\ \mu^\top & 1 \end{bmatrix} \succeq 0 \\
& \|\hat{\mu}_1\|^2 - 2\hat{\mu}_1^\top \mu + \mathrm{Tr}\big[M + \hat{\Sigma}_1 - 2C\big] \leq \rho^2.
\end{array} \right.
$$

Proof ideas:

- Chebyshev bounds on an open polyhedron
- Rejoin a two-layer optimization problem

Uncertainty Quantification

Theorem: We have

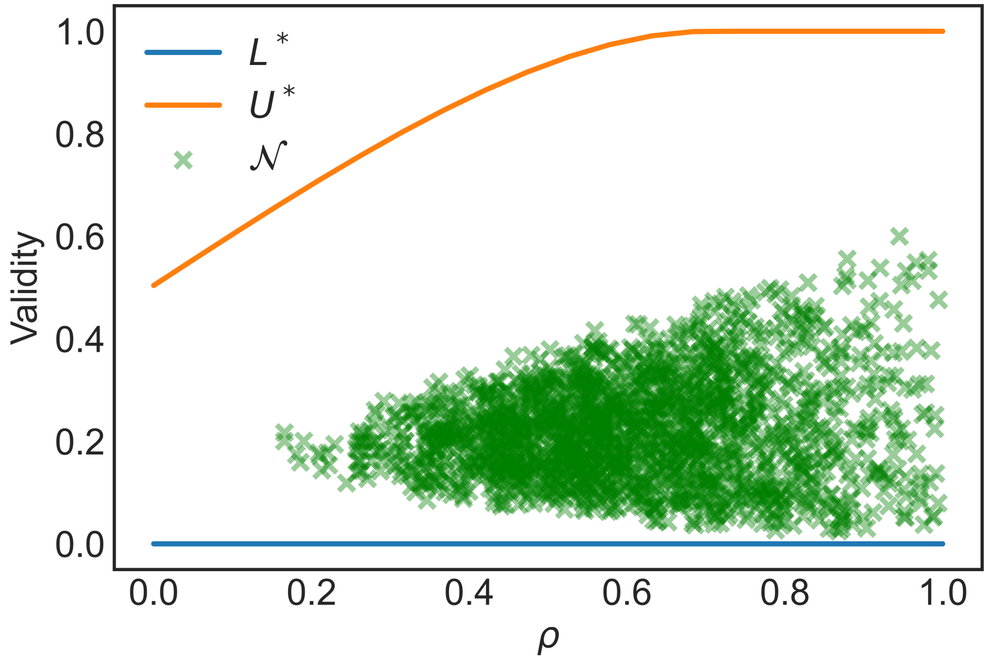
$$
U^\star = \left\{ \begin{array}{ll}
\inf & z_0 + \gamma(\rho^2 - \|\hat{\mu}_1\|_2^2 - \mathrm{Tr}\big[\hat{\Sigma}_1\big]) + q + \mathrm{Tr}\big[Q\big] \\
\text{s.t.} & \gamma \in \mathbb{R}_+,\ z_0 \in \mathbb{R},\ z \in \mathbb{R}^d,\ Z \in \mathbb{S}_+^d,\ q \in \mathbb{R}_+,\ Q \in \mathbb{S}_+^d,\ \lambda \in \mathbb{R}_+^J \\
& \begin{bmatrix} \gamma I - Z & \gamma \hat{\Sigma}_1^{\frac{1}{2}} \\ \gamma \hat{\Sigma}_1^{\frac{1}{2}} & Q \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma I - Z & \gamma \hat{\mu}_1 + z \\ \gamma \hat{\mu}_1^\top + z^\top & q \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} Z & z \\ z^\top & z_0 \end{bmatrix} \succeq 0, \quad \begin{bmatrix} Z & z \\ z^\top & z_0 - 1 \end{bmatrix} \succeq \sum_{j \in [J]} \lambda_j \begin{bmatrix} 0 & \frac{1}{2} x_j \\ \frac{1}{2} x_j^\top & 0 \end{bmatrix}.
\end{array} \right.
$$

Proof ideas:

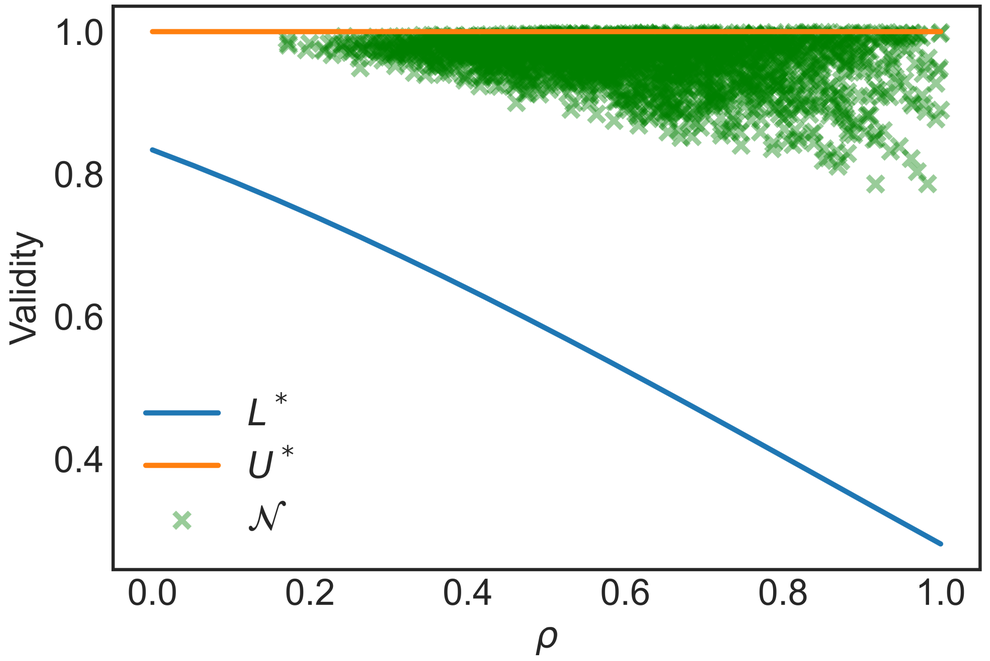
- Duality techniques from DRO
- Rejoin with the support function of the uncertainty set

Empirical Results

- fix a $\{x_j\}_J$
- fix $\hat{\mu} = \theta_0, \hat{\Sigma} = 0.5I$
- evaluate $L^\star$ and $U^\star$
- sample $\tilde{\theta} \sim \mathcal{N}(\mu_g, \Sigma_g)$, (μ_g, Σ_g) is chosen randomly



$\hat{\mu} \notin$



$\hat{\mu} \in$

Counterfactual Plan - Construction

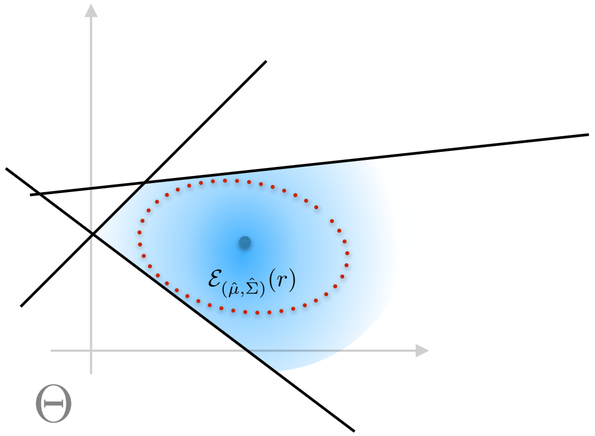
$$\min_{x_1,\ldots,x_J} \quad \text{Proximity}(\{x_j\}, x_0) - \lambda_1 \text{FutureValidity}(\{x_j\}) - \lambda_2 \text{Diversity}(\{x_j\})$$

Average L-1 distance

where $$\text{Proximity}(\{x_j\}, x_0) \triangleq \frac{1}{J}\sum_{j=1}^{J} \|x_j - x_0\|_1.$$

The maximum volume of the ellipsoid inscribed in a set of feasible parameters

$$\text{FutureValidity}(\{x_j\}) \triangleq \max\{r : r \geq 0,\ \mathcal{E}_{(\hat{\mu},\hat{\Sigma})}(r) \subseteq \Theta(\{x_j\})\}.$$



Determinantal point process (DPP)

$$\text{Diversity}(\{x_j\}) \triangleq \det(K), \text{ where } K_{i,j} = (1 + \|x_i - x_j\|_1)^{-1}\ \forall 1 \leq i, j \leq J.$$

Solved by projected gradient descent

Empirical Results

- each dataset contains two sets of data (present data and future data).
- generate counterfactuals w.r.t. the classifier trained on the present data.
- evaluate empirical validity w.r.t. the classifiers trained on the future data.

Dataset	Method	Proximity	Diversity	L^*	Empirical Validity
German credit	DiCE	0.986 ± 0.324	0.072 ± 0.050	0.649 ± 0.073	0.996 ± 0.008
	MahalanobisCrr	1.002 ± 0.323	0.064 ± 0.047	0.750 ± 0.064	0.999 ± 0.003
	COPA ($\lambda_1 = 0.2; \lambda_2 = 2.0$)	$\mathbf{0.916} \pm 0.178$	0.017 ± 0.058	0.944 ± 0.168	0.997 ± 0.018
	COPA ($\lambda_1 = 0.5; \lambda_2 = 5.0$)	1.154 ± 0.253	0.114 ± 0.101	$\mathbf{0.946} \pm 0.040$	$\mathbf{1.000} \pm 0.000$
	COPA ($\lambda_1 = 1.0; \lambda_2 = 10.0$)	1.351 ± 0.166	$\mathbf{0.225} \pm 0.045$	0.911 ± 0.022	$\mathbf{1.000} \pm 0.000$
SBA	DiCE	2.037 ± 0.470	0.089 ± 0.057	0.946 ± 0.014	0.801 ± 0.061
	MahalanobisCrr	2.014 ± 0.473	0.085 ± 0.055	0.966 ± 0.007	0.945 ± 0.062
	COPA ($\lambda_1 = 0.2; \lambda_2 = 2.0$)	$\mathbf{1.831} \pm 0.139$	0.253 ± 0.026	0.994 ± 0.000	1.000 ± 0.000
	COPA ($\lambda_1 = 0.5; \lambda_2 = 5.0$)	1.966 ± 0.112	0.363 ± 0.012	$\mathbf{0.995} \pm 0.000$	$\mathbf{1.000} \pm 0.000$
	COPA ($\lambda_1 = 1.0; \lambda_2 = 10.0$)	2.010 ± 0.124	$\mathbf{0.380} \pm 0.006$	0.995 ± 0.000	$\mathbf{1.000} \pm 0.000$
Student performance	DiCE	$\mathbf{1.486} \pm 0.325$	$\mathbf{0.136} \pm 0.044$	0.549 ± 0.307	0.408 ± 0.363
	MahalanobisCrr	1.497 ± 0.325	0.126 ± 0.044	0.864 ± 0.117	0.757 ± 0.284
	COPA ($\lambda_1 = 0.2; \lambda_2 = 2.0$)	1.779 ± 0.352	0.052 ± 0.047	$\mathbf{0.998} \pm 0.000$	$\mathbf{1.000} \pm 0.000$
	COPA ($\lambda_1 = 0.5; \lambda_2 = 5.0$)	1.882 ± 0.353	0.089 ± 0.032	0.998 ± 0.000	$\mathbf{1.000} \pm 0.000$
	COPA ($\lambda_1 = 1.0; \lambda_2 = 10.0$)	1.926 ± 0.349	0.109 ± 0.024	0.997 ± 0.000	$\mathbf{1.000} \pm 0.000$

Follow-up Questions

- Can we use the lower bound to construct a *robust* counterfactual plan?
  - No, we have to use a maximum ellipsoid to approximate it.
  - But, if there is only one counterfactual explanation, we can construct it efficiently -> ICLR'23

- Can we use it for nonlinear models?
  - Yes, but we must use a linear model (e.g., LIME) to construct a surrogate decision boundary before applying our technique.

- How does the linear surrogate affect the constructed explanation? (e.g., unfaithful to the decision boundary)
  - Analyzing and proposing a more faithful linear surrogate -> Rejected (4 times)

- I don't wanna use a linear surrogate.
  - Robust Bayesian Recourse -> UAI'22

The Shapley value

from game theory to machine learning

Cooperative games

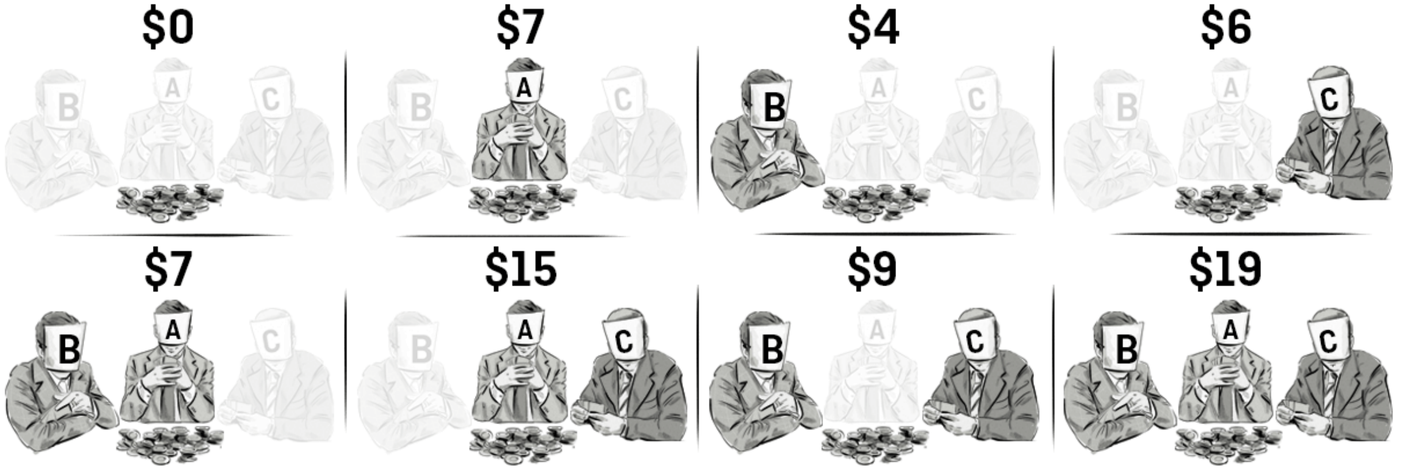
- Cooperative games model scenarios where
  - Agent can benefit by cooperating
  - Binding agreements are possible

Cooperative games

- How to construct a salary structure in a company or a football club?

Cooperative games

- How to construct a salary structure in a company or a football club?
- How to divide a surplus (profit) to shareholders so that everyone is satisfied?



Cooperative games

Cooperative game

- Players can benefit by cooperating
- Binding agreements are possible
- Answer the question: How to divide the surplus when joining the grand coalition?

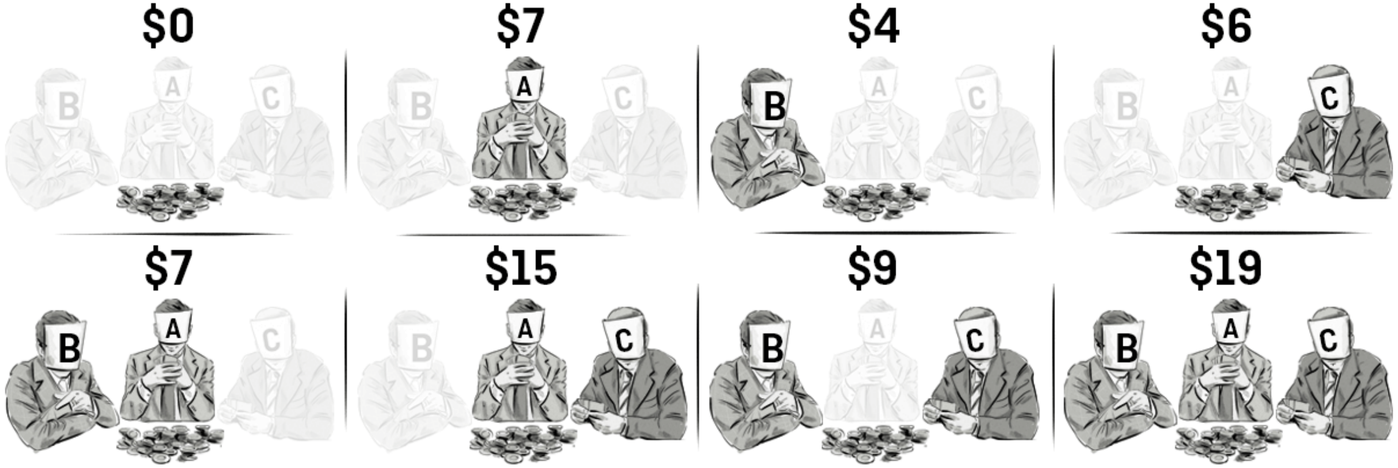
Non-cooperative game

- Players are independent
- No cooperation
- Answer the question: What is the “good” strategy for each player to maximize their return?

Nash equilibrium, zero-sum game

Cooperative games

- A cooperative game is described by a tuple (N, F) where $N = \{1, \ldots, n\}$ a set of players and $F : 2^N \to \mathbb{R}$, $F(\emptyset) = 0$ is a characteristic function.
  - By von Neumann and Morgenstern:
    - Superadditive game: $F(S \cup T) \geq F(S) + F(T)$ (relaxed by later works)
    - Transferable Utility (TU) game | non-Transferable Utility (NTU) game



John von Neumann and Oskar Morgenstern, *Theory of games and economic behavior. 1944*

Cooperative games

- A cooperative game is described by a tuple (N, F) where $N = \{1, \ldots, n\}$ a set of players and $F : 2^N \to \mathbb{R}$, $F(\emptyset) = 0$ is a characteristic function.
  - $\mathbb{G}$: set of all possible games

- How to divide a surplus (profit) to players?
  - $\psi : \mathbb{G} \to \mathbb{R}^n, \quad \psi((N, F)) \in \mathbb{R}^n$: a solution concept

When context is clear, I omit the player set *N* in *(N, F)*
and use *F* to denote the game and $\psi(F)$ to denote a solution concept of *F*

John von Neumann and Oskar Morgenstern, *Theory of games and economic behavior. 1944*

Cooperative games

- A cooperative game is described by a tuple (N, F) where $N = \{1, \ldots, n\}$ a set of players and $F : 2^N \to \mathbb{R}$, $F(\emptyset) = 0$ is a characteristic function.
  - $\mathbb{G}$: set of all possible games

- How to divide a surplus (profit) to players?
  - $\psi : \mathbb{G} \to \mathbb{R}^n, \quad \psi((N, F)) \in \mathbb{R}^n$: a solution concept

$\psi((N, F))$	$4	$11	$9

Payoff vector

John von Neumann and Oskar Morgenstern, *Theory of games and economic behavior. 1944*

Several Solution Concepts

- **Imputation set** (von Neumann & Morgenstern, 1944): Includes all payoff vectors that are *efficient,* i.e., share the whole utility, and *individually rational*, i.e., assign to each player at least her standing alone worth.

$$\psi_{IS}((N,F)) = \left\{x \in \mathbb{R}^n : \sum_{i \in N} x_i = F(N), x_i \geq v(i), \forall i \in N\right\}$$

- **Core set** (Gillies, 1953): similar to imputation set but considering *coalitional rational* instead of individual rational.

$$\psi_{CS}((N,F)) = \left\{x \in \mathbb{R}^n : \sum_{i \in S} x_i \geq F(S), \forall S \subseteq N\right\}$$

Could be empty and not unique!

Chapman and Hall, *Handbook of the Shapley value.* 2019

The Shapley Value

- A solution concept for cooperative games that assigns payoffs for each player in the game and that is
  - unique for each game
  - efficient
  - fair

$\psi((N, F))$

$4	$8	$7

Payoff vector

The Shapley Value

Def 1.1 (*dividend*): The value of the player *i* is the sum of its dividend share to all possible coalitions to which he could be joined. The dividend of a coalition is shared equally by its members.

$$\phi_i((N,F)) = \sum_{S \subseteq N \setminus i} \frac{\Delta_F(S \cup i)}{s+1}, \quad \forall i \in N,$$

where $\Delta_F(T)$ is Harsanyi dividend of the game F to the coalition T:

$$\Delta_F(T) = \sum_{R \subseteq T} (-1)^{t-r} F(R), \quad \forall T \subseteq N$$

Harsanyi, *A bargaining model for cooperative n-person games*. 1959.

The Shapley Value

Def 1.2 (weighted *marginal contribution*): The value of player *i* is computed by its weighted average marginal contribution over all possible coalitions to which player *i could be joined*.

$$\phi_i((N,F)) = \frac{1}{n} \sum_{S \subseteq N \setminus i} \frac{1}{\binom{n-1}{s}} [F(S \cup i) - F(S)], \quad \forall i \in N$$

L.S. Shapley, *A value for n-person games*. 1953.

The Shapley Value

Def 1.3 (*random order*): The grand coalition is formed in a way that each player joins the coalition one-by-one, the value of player *i* is the average of its marginal contribution over all possible order.

$$\phi_i((N,F)) = \frac{1}{n!} \sum_{\pi \in \Pi(N)} [F(\pi^i \cup i) - F(\pi^i)]$$

where π^i is the set of players that precede player i in the permutation π.

Chapman and Hall, *Handbook of the Shapley value.* 2019

Axiomatic Characterization

We want to design a solution concept ψ that satisfies the following axioms:

- **Axiom 1** (*Linearity, L*)**:** for any two games F_1 and F_2,
$$\psi(F_1 + \alpha F_2) = \psi(F_1) + \alpha\psi(F_2)$$
- **Axiom 2** (*Dummy, D*): if $i \in N$ is a dummy player,
i.e., $F(S \cup i) = F(S) + F(i), \forall S \subseteq N$, then $\psi_i(F) = F(i)$.
- **Axiom 3** (*Symmetry, S*): For any permutation $\pi \in \Pi(N)$, $\psi_{\pi i}(\pi F) = \psi_i(F)$
- **Axiom 4** (*Efficiency, E*): $\sum_{i \in N} \psi_i(F) = F(N) - F(\emptyset)$

Alvin E. Roth, *The Shapley value.* 1988.

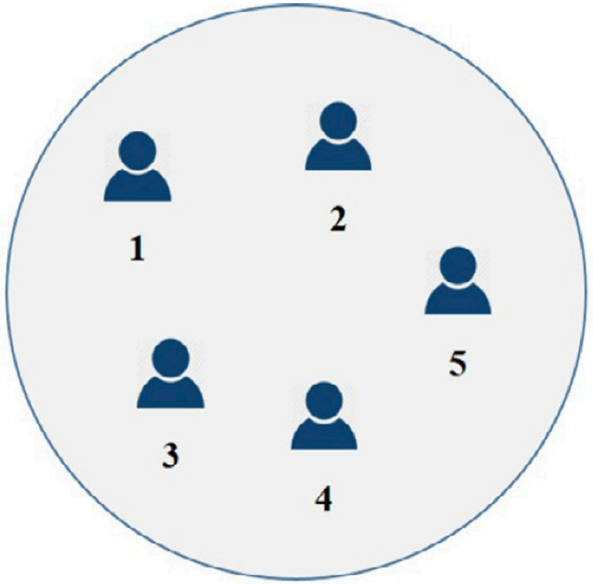
When context is clear, I omit the player set *N* in *(N, F)* and use *F* to denote the game

Main result

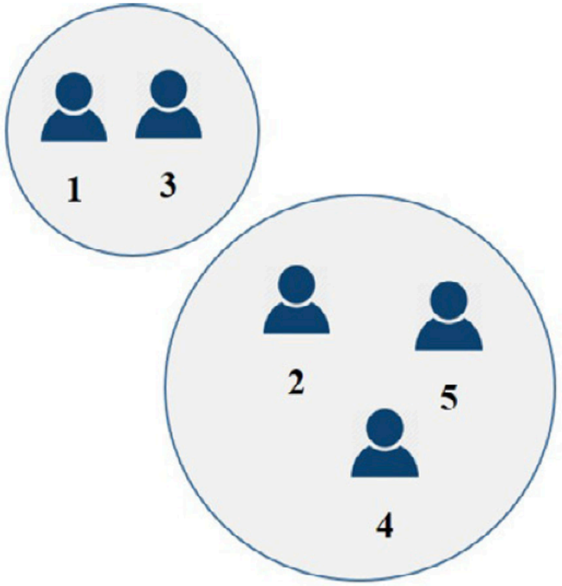
Theorem 1: Shapley value is a **unique** solution concept that satisfies four axioms *linearity, dummy, symmetry,* and *efficiency*.

L.S. Shapley, *A value for n-person games*. 1953.

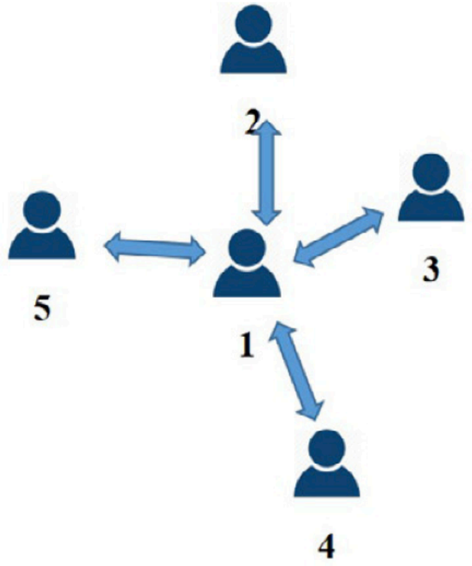
Types of Cooperative Games



Class I: Canonical Coalitional Games



Class II: Coalition Formation Games

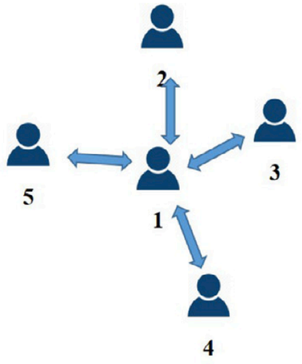


Class III: Coalitional Graph Games

Myerson, *Handbook of the Shapley value.* 1977

Coalitional Graph Games

- N is a set of nodes, $L = \{ij \mid i, j \in N\}$ is a set of edges. An edge represents a communication connection between two nodes.
- We omit the set N whenever the context is clear.
- $\zeta(N, L) = \{C_1, \ldots, C_r\}$ denotes a set of connected components of a graph (N, L).
- $L|_S$ is a set of edges induced by subset $S \subseteq N$, *i.e.*, $L|_S = \{ij \in L \mid i, j \in S\}$.



Class III: Coalitional Graph Games

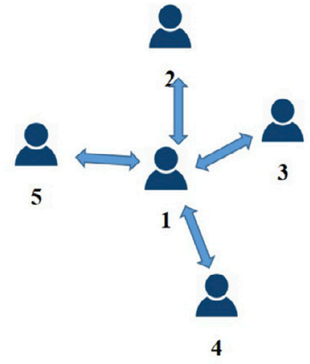
- A communication-restricted function $f|_L$ is defined as

$$f|_L(T) = \sum_{R \in \zeta(T, L|_T)} f(R) \quad \forall T \subseteq N.$$

Myerson, *Handbook of the Shapley value.* 1977

The Myerson Value

- N is a set of nodes, $L = \{ij \mid i, j \in N\}$ is a set of edges. An edge represents a communication connection between two nodes.
- We omit the set N whenever the context is clear.
- $\zeta(N, L) = \{C_1, \ldots, C_r\}$ denotes a set of connected components of a graph (N, L).
- $L|_S$ is a set of edges induced by subset $S \subseteq N$, *i.e.*, $L|_S = \{ij \in L \mid i, j \in S\}$.



Class III: Coalitional Graph Games

- A communication-restricted function $f|_L$ is defined as

$$f|_L(T) = \sum_{R \in \zeta(T, L|_T)} f(R) \quad \forall T \subseteq N.$$

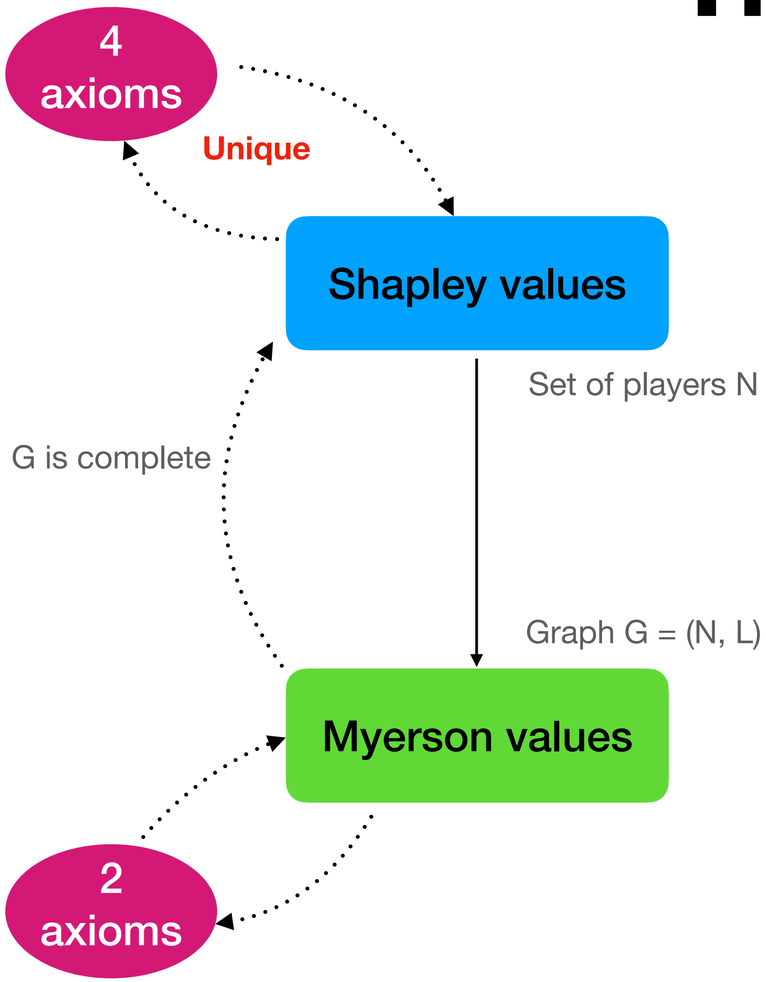
Definition 2.2 (Myerson value (Myerson, 1977))**.** *Given a graph* (N, L) *and a characteristic function* f*, Myerson value on a CO-game* (N, f, L) *is defined as*

$$\psi(f, L) = \phi(f|_L),$$

where $\phi(f|_L)$ *is Shapley value of a communication-restricted function* $f|_L$*.*

Myerson, *Handbook of the Shapley value.* 1977

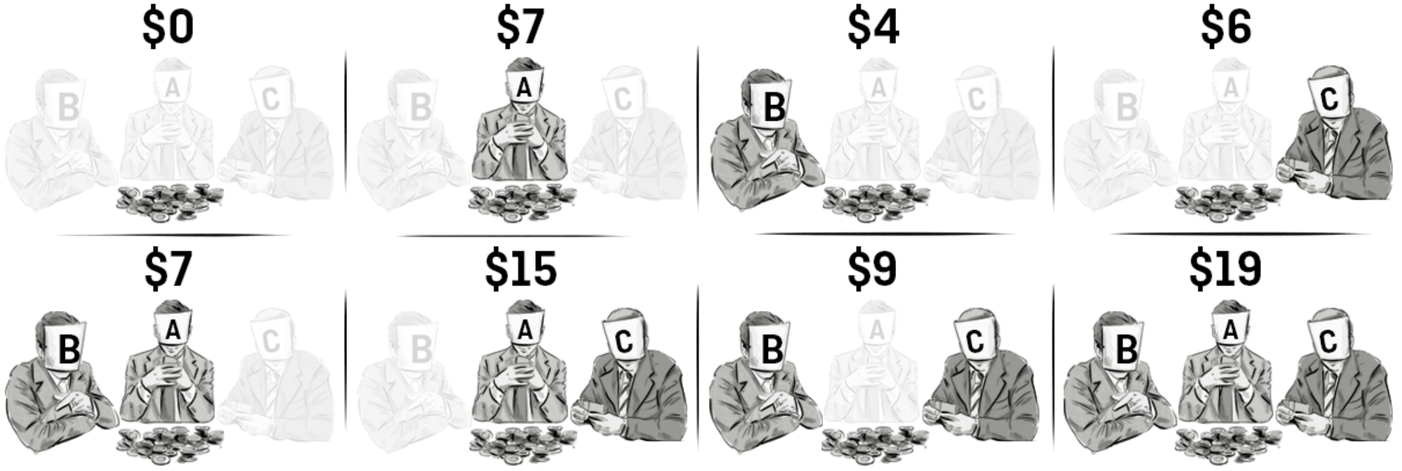
The Myerson Value



The Shapley Taylor Interaction Index

Definition 2.1 (Shapley-Taylor Interaction)**.** *The Shapley-Taylor index is defined as the expectation of* $\Phi^m_{S,\pi}(f)$ *over an ordering of* N *players chosen uniformly at random:*

$$\Phi^m_S(f) = \mathbb{E}_\pi(\Phi^m_{S,\pi}(f)) = \frac{m}{n} \sum_{T \subseteq N \setminus S} \delta_S f(T) \frac{1}{\binom{n-1}{t}}.$$



Myerson, *Handbook of the Shapley value.* 1977

The Myerson Value