Trustworthy AI Autonomy

Lecture 6: Generative models - Adversarial

Ding Zhao

Assistant Professor

Carnegie Mellon University

Carnegie Mellon University

Safe AI Lab @CMU

Agenda

- Traditional ways to identify scenarios
- Data-based Scenario Generation
- Adversarial Scenario Generation
    - Adversarial Generative Network
    - Importance Sampling methods
- Knowledge-based Scenario Generation

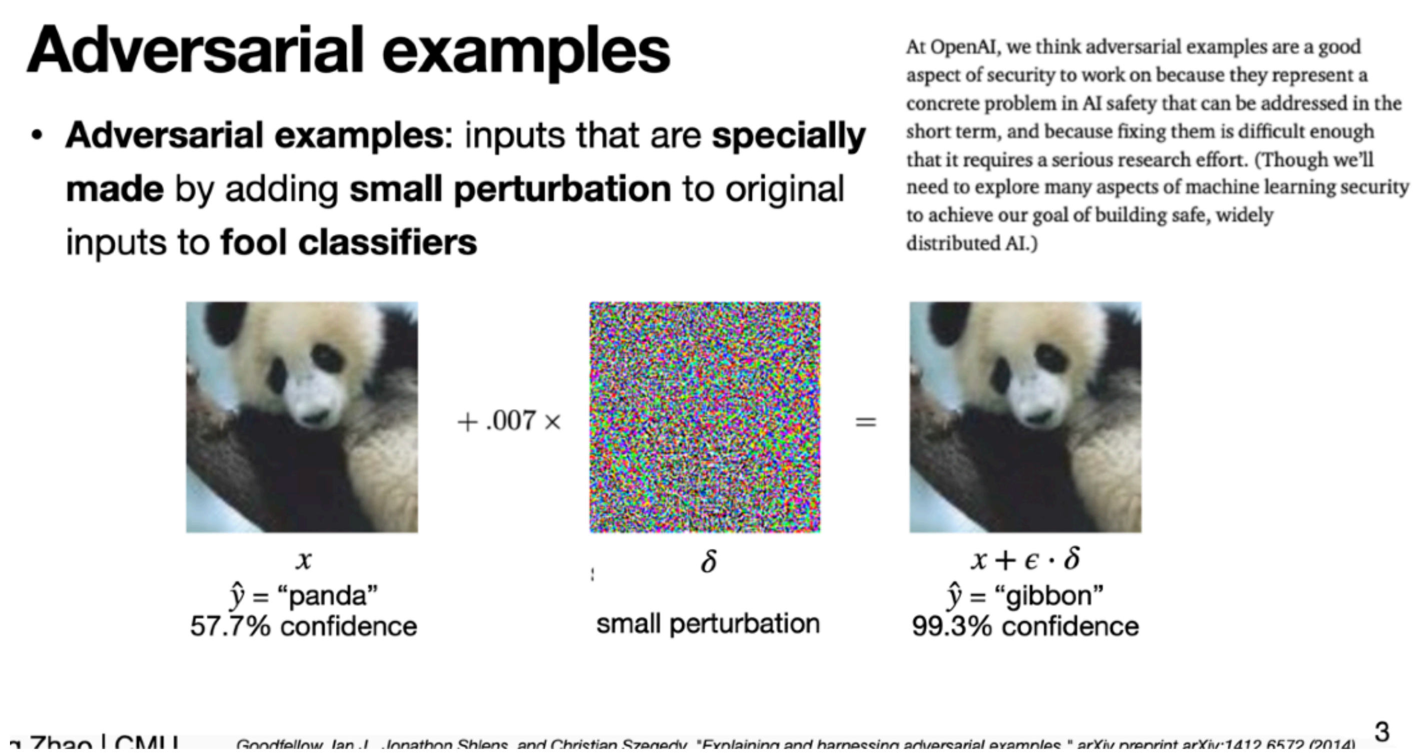
Adversarial Generative models?

Check This ..

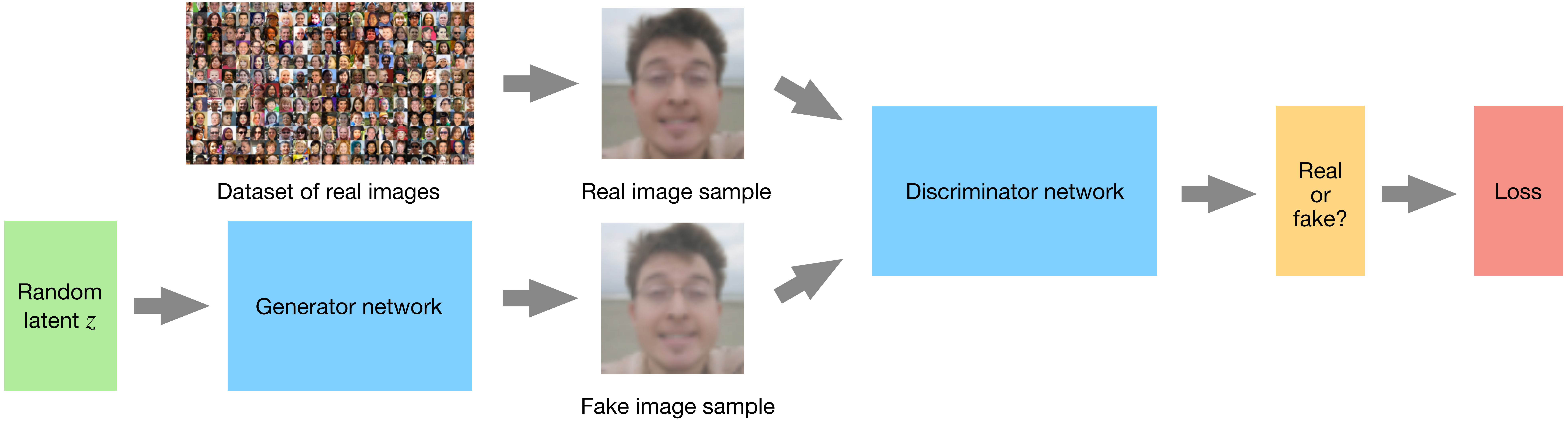
https://thispersondoesnotexist.com

Generative models

Generative Adversarial Network

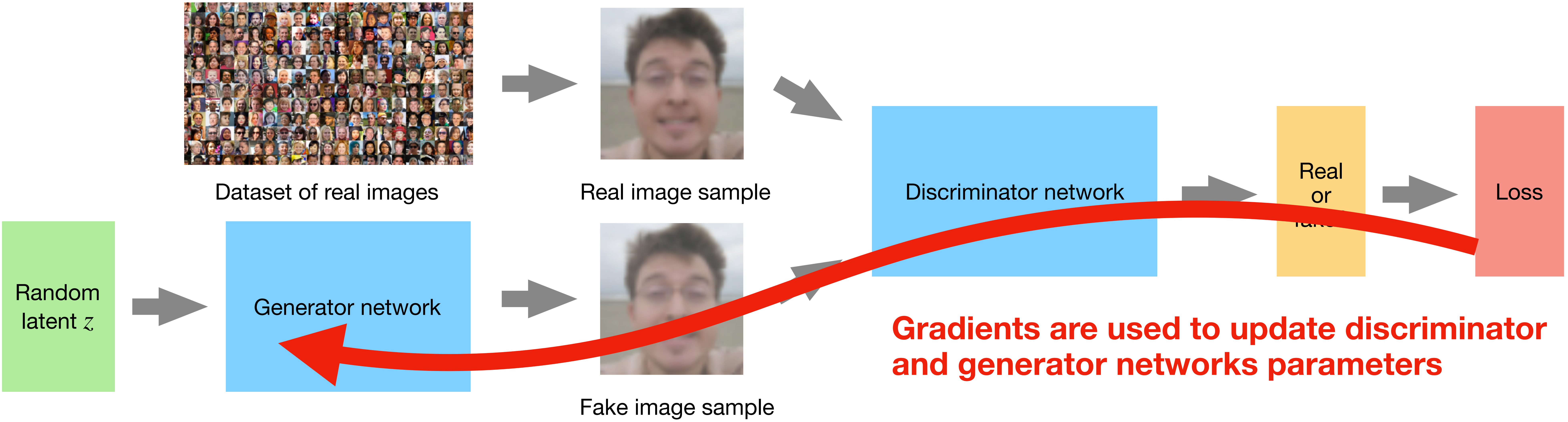


- Consists of two neural networks:
  - generator: generating fake high-quality images from random latent samples (e.g. Gaussian noise)
  - discriminator: classifying whether images are real (from datasets) or fake (generated by the generator)



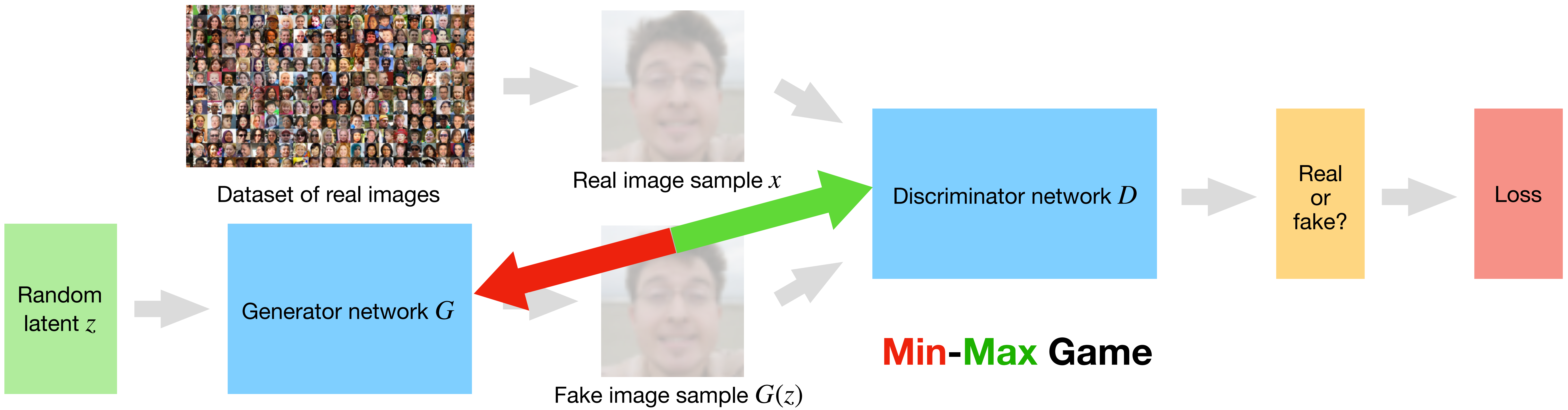
Generative Adversarial Network

- Training procedure
  - The parameters of both networks are updated by backpropagating the gradient of a mutual loss function
  - Key step: ensuring both networks are well-balanced (none dominating the other during training)



Generative Adversarial Network

- Training mechanism is a minimax game:
  - Generator (G): generating good images using latent samples $z \sim p_z$
  - Discriminator (D): discriminating real images $x \sim p_x$ from fake $G(z)$



Generative Adversarial Network

- Training mechanism is a minimax game:
  - Discriminator (D): discriminating real images $x \sim p_x$ from fake $G(z)$

Training objective: $\max_{D} \mathbb{E}_{x \sim p_x}[\log D(x)]$

Generative Adversarial Network

- Training mechanism is a minimax game:
  - Generator (G): generating good images using latent samples $z \sim p_z$

Training objective:

$$\max_{G} \; \mathbb{E}_{z \sim p_z}\left[\mathbb{1}(G(z) = real)\right] \approx \min_{G} \; \mathbb{E}_{z \sim p_z}\left[\log(1 - D(G(z)))\right]$$

Generative Adversarial Network

- Training mechanism is a minimax game:
  - Generator (G): generating good images using latent samples $z \sim p_z$
  - Discriminator (D): discriminating real images $x \sim p_x$ from fake $G(z)$
- Training goal: finding the best G and D simultaneously:

$$\min_G \max_D \left[\mathbb{E}_{x\sim p_x}[\log D(x)] + \mathbb{E}_{z\sim p_z}\left[\log(1 - D(G(z)))\right] \right]$$

Generative Adversarial Network

- Improving the convergence of the minimax optimization
  - Choosing an appropriate mutual loss function (similar idea, but different formulation)

GAN Type	Key Take-Away
GAN	The original (JSD divergence)
WGAN	EM distance objective
Improved WGAN	No weight clipping on WGAN
LSGAN	L2 loss objective
RWGAN	Relaxed WGAN framework
McGAN	Mean/covariance minimization objective
GMMN	Maximum mean discrepancy objective
MMD GAN	Adversarial kernel to GMMN
Cramer GAN	Cramer distance
Fisher GAN	Chi-square objective
EBGAN	Autoencoder instead of discriminator
BEGAN	WGAN and EBGAN merged objectives
MAGAN	Dynamic margin on hinge loss from EBGAN

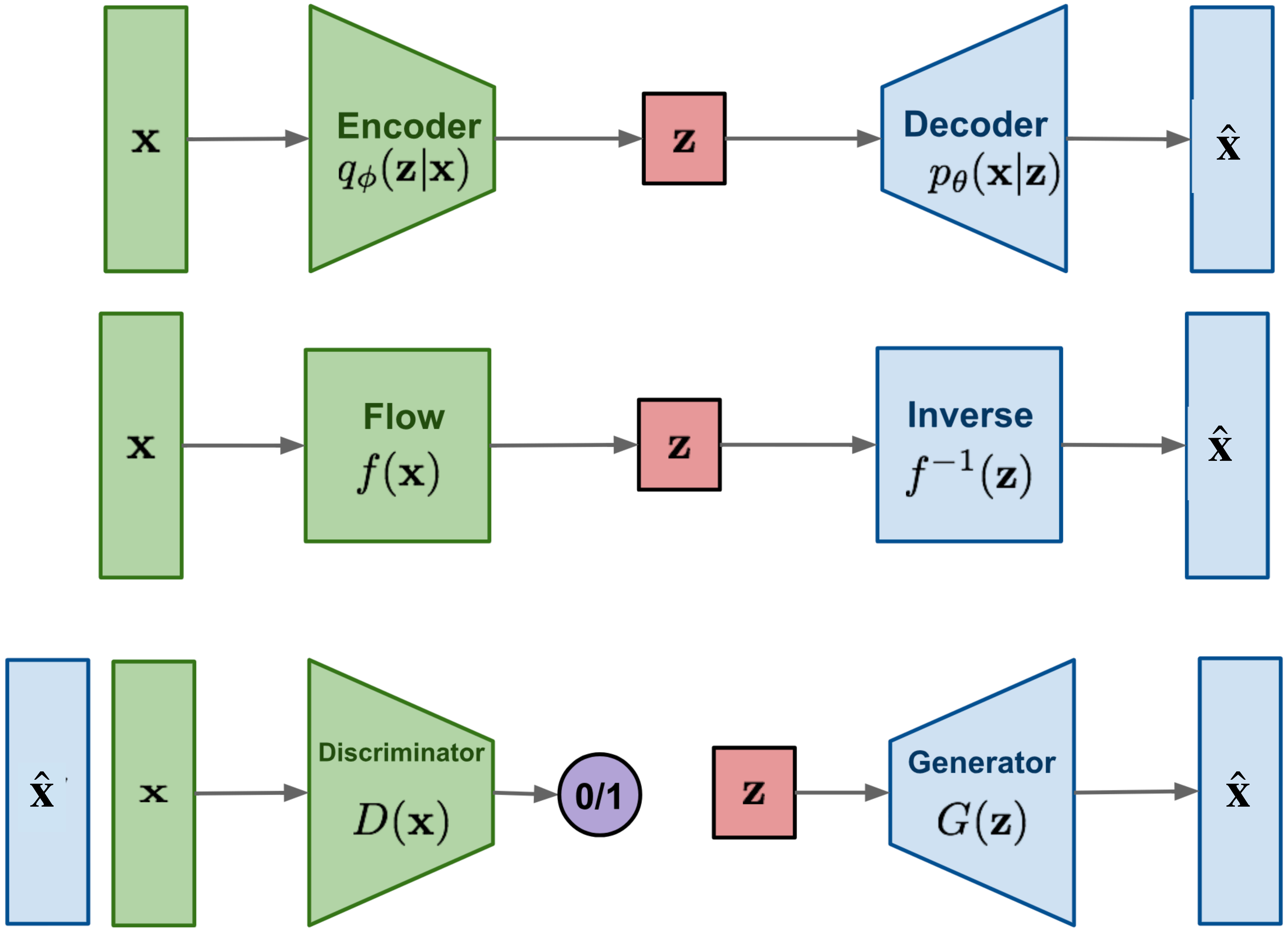
Source: https://towardsdatascience.com/gan-objective-functions-gans-and-their-variations-ad77340bce3c

Deep generative models

Approximate likelihood — **VAE:** maximize ELBO.

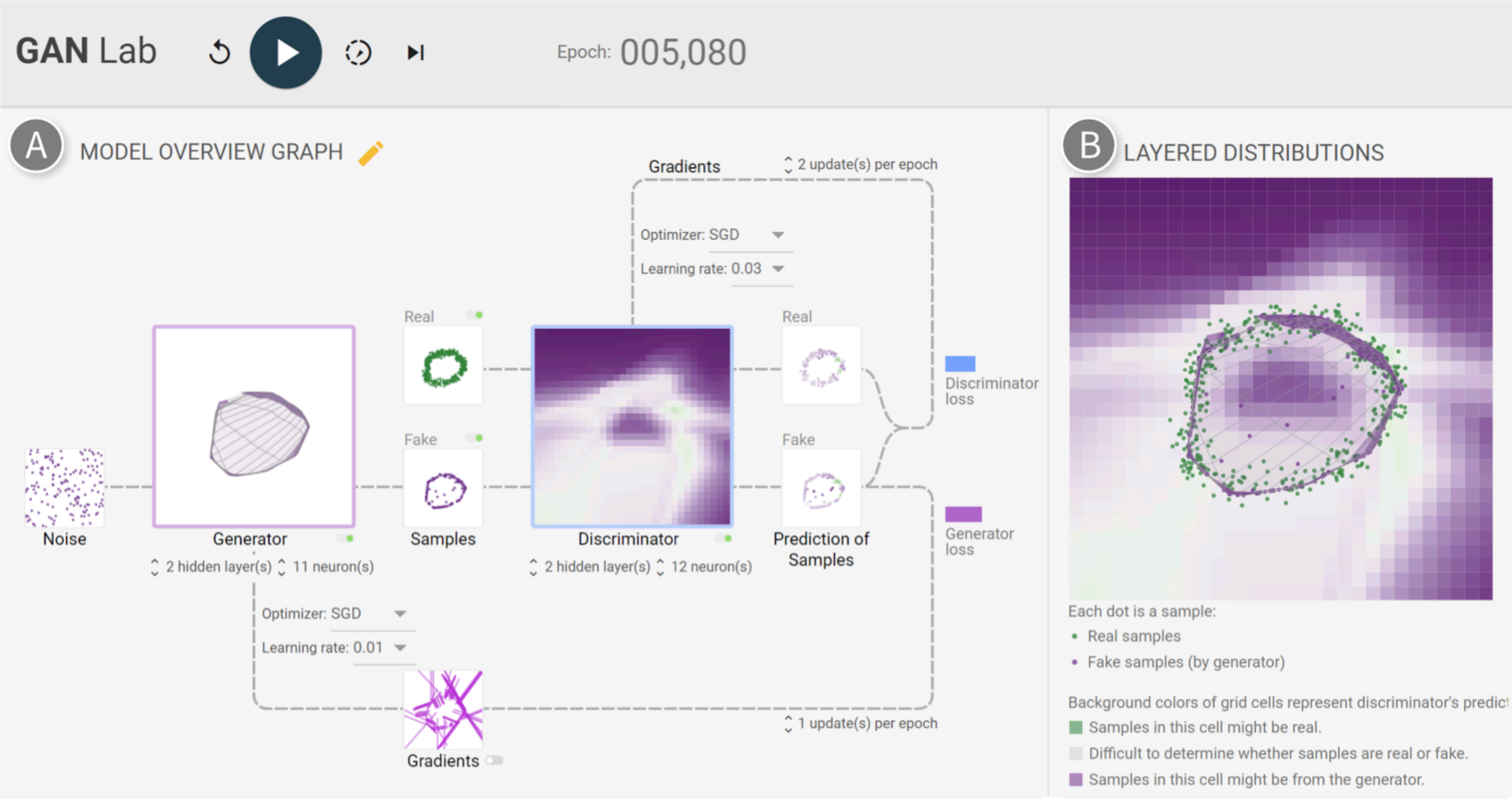
Exact likelihood — **Flow-based generative models:** minimize the negative log-likelihood

Likelihood free — **GAN:** minimax the classification error loss.



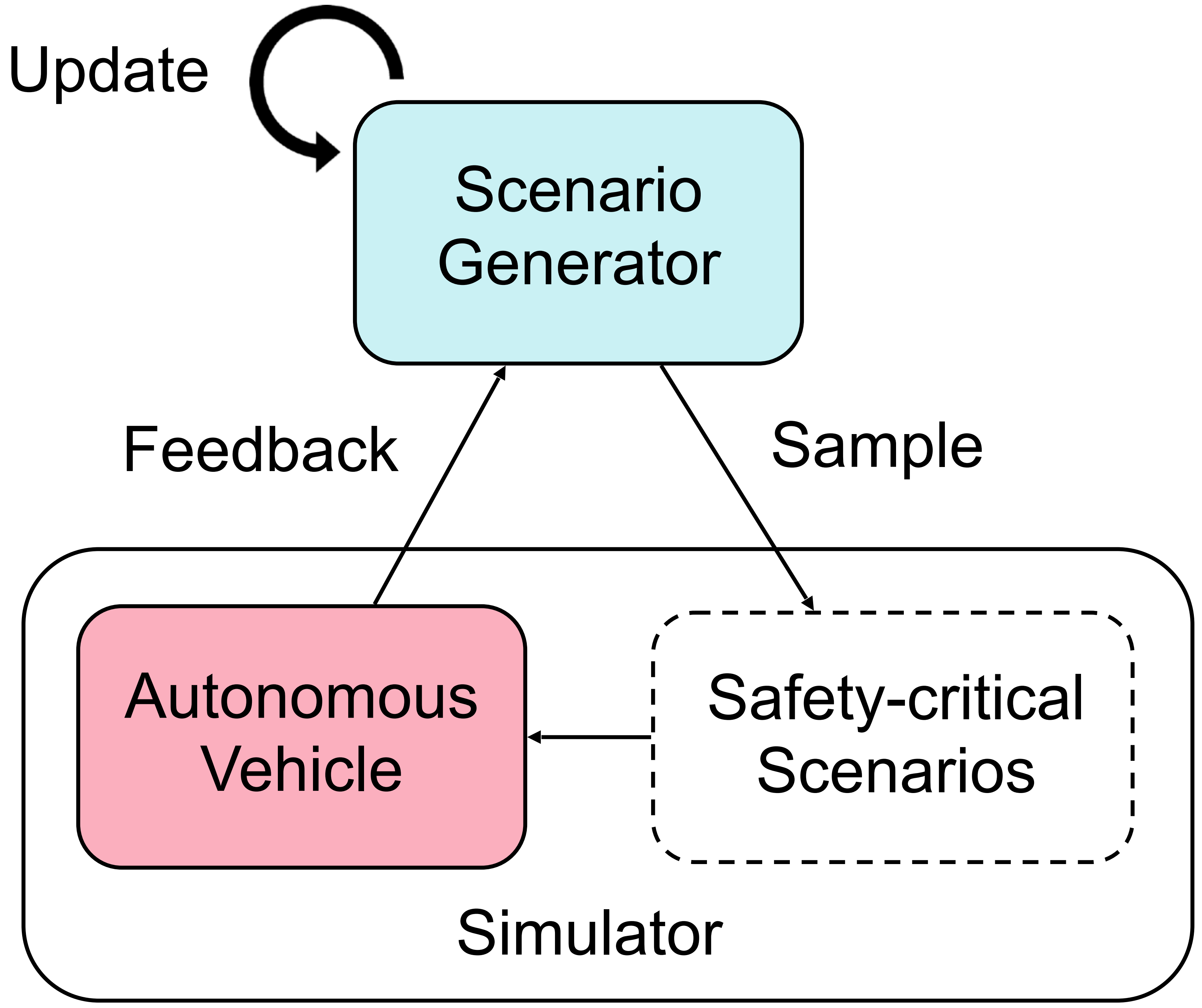
https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

Hands-on time: GAN Lab



https://poloclub.github.io/ganlab/

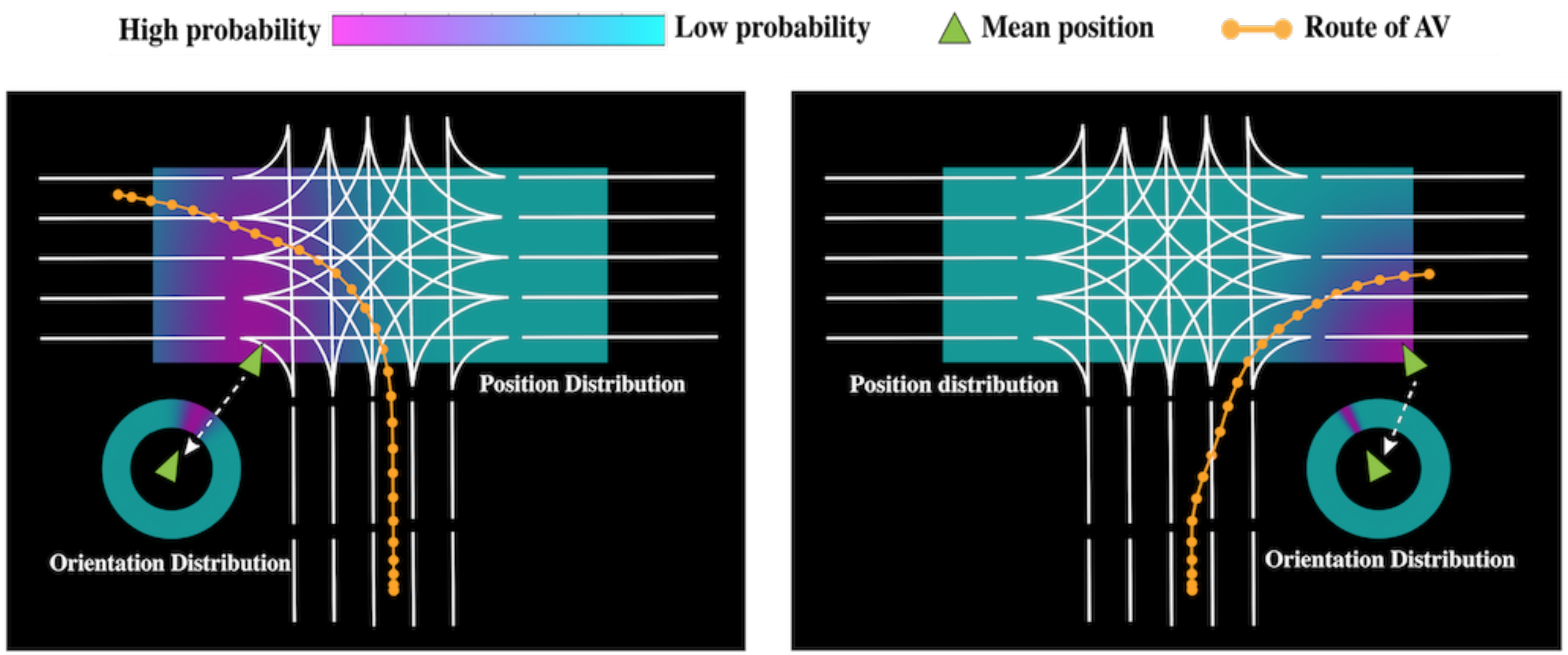
Adversarial Scenario Generation



- Put an autonomous vehicle into the loop to give feedback to the generator.

Adversarial Scenario Generation

Distribution of learned safety-critical scenarios (initial position and orientation)

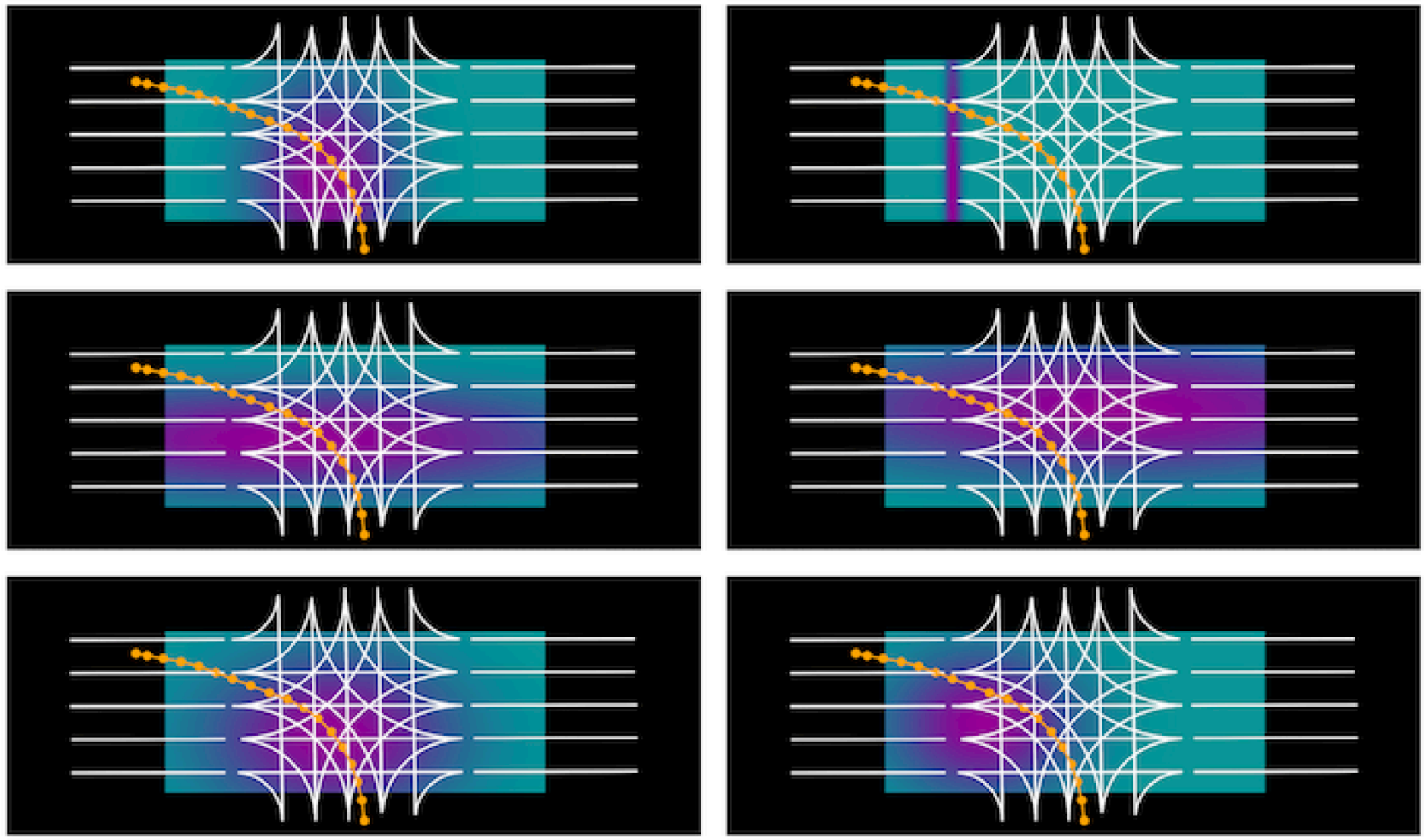


- Adaptive to different routes of AV

W. Ding, M. Xu, D. Zhao. Learning to collide: An adaptive safety-critical scenarios generating method, IROS 2020

Adversarial Scenario Generation

What’s the remaining problem ?

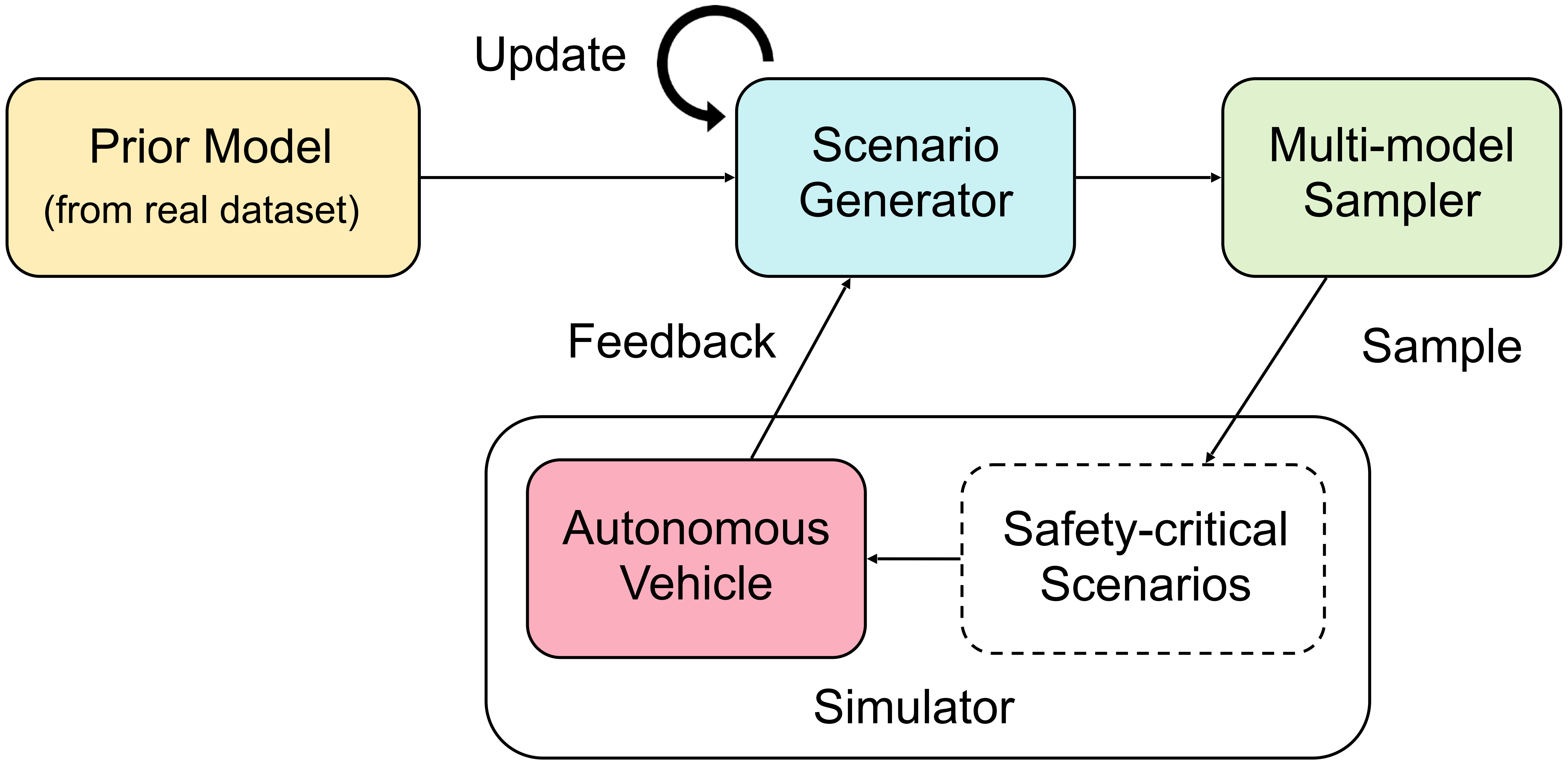


Different results with different initialization

- Safety-critical scenarios are diverse and follow a multi-modal distribution.
- Generated Safety-critical scenarios should be realistic.

W. Ding, M. Xu, D. Zhao. Learning to collide: An adaptive safety-critical scenarios generating method, IROS 2020

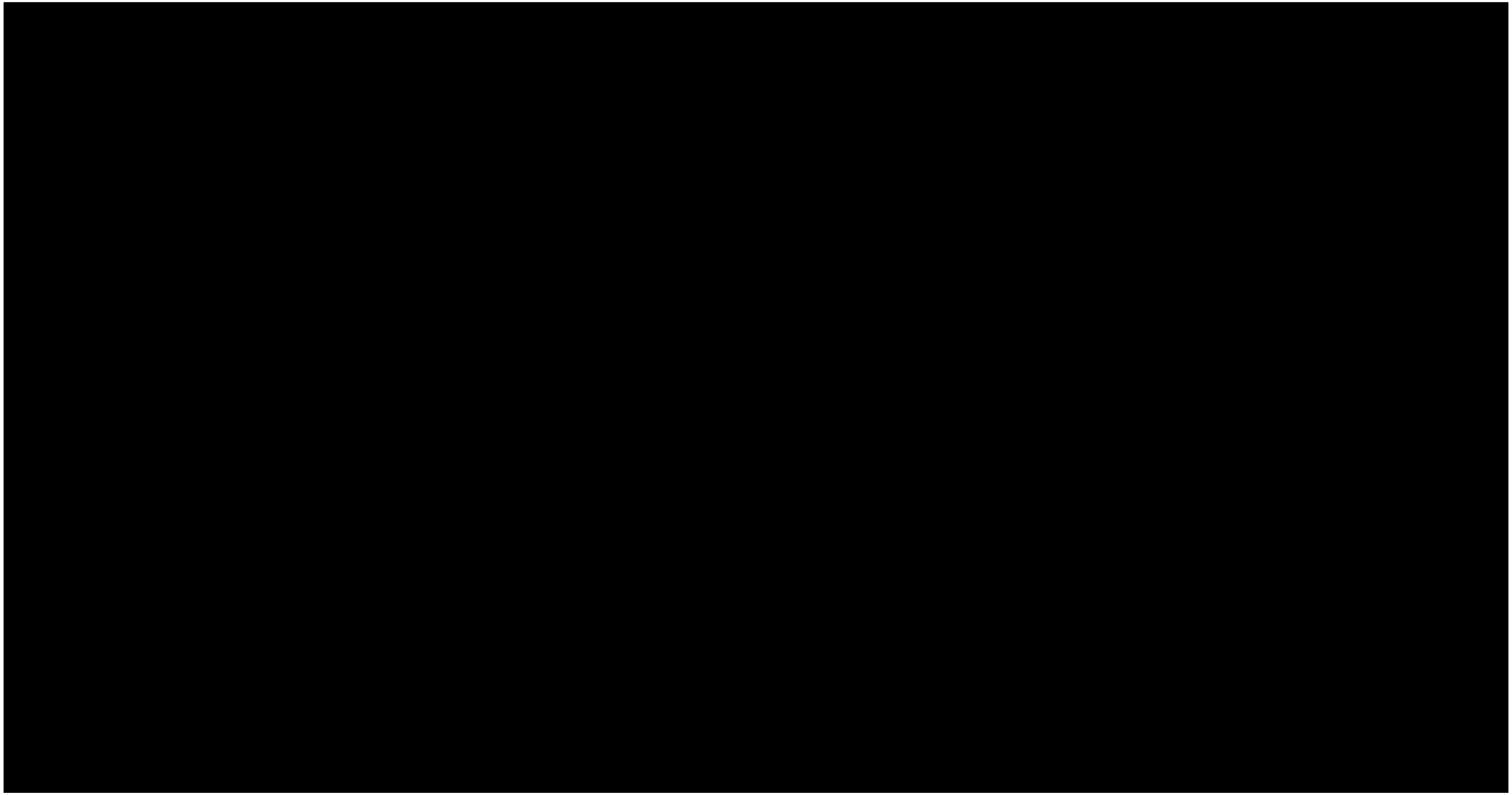
Adversarial Scenario Generation



- Use a prior model to represent the probability of a scenario happen in the real-world.
- Use a normalizing flow model to estimate the multi-modal distribution.

W. Ding, B. Chen, B. Li, D. Zhao, *Multimodal Safety-Critical Scenarios Generation for Decision-Making Algorithms Evaluation, Robotics and Automation Letters 2021*

Adversarial Scenario Generation



W. Ding, B. Chen, B. Li, D. Zhao, Multimodal Safety-Critical Scenarios Generation for Decision-Making Algorithms Evaluation, Robotics and Automation Letters 2021

Adversarial Scenario Generation

Summary

- Adaptivity, interact with downstream vehicle
- Considering the real-world data
- Diversity, multi-modal distribution

- Poor generalization, only works for target autonomous vehicle
- Sparse and inefficient, robust vehicle is hard to attack
- Traffic rule violation

11 billion miles

To prove an AV is safer than human drivers

Rare event analysis

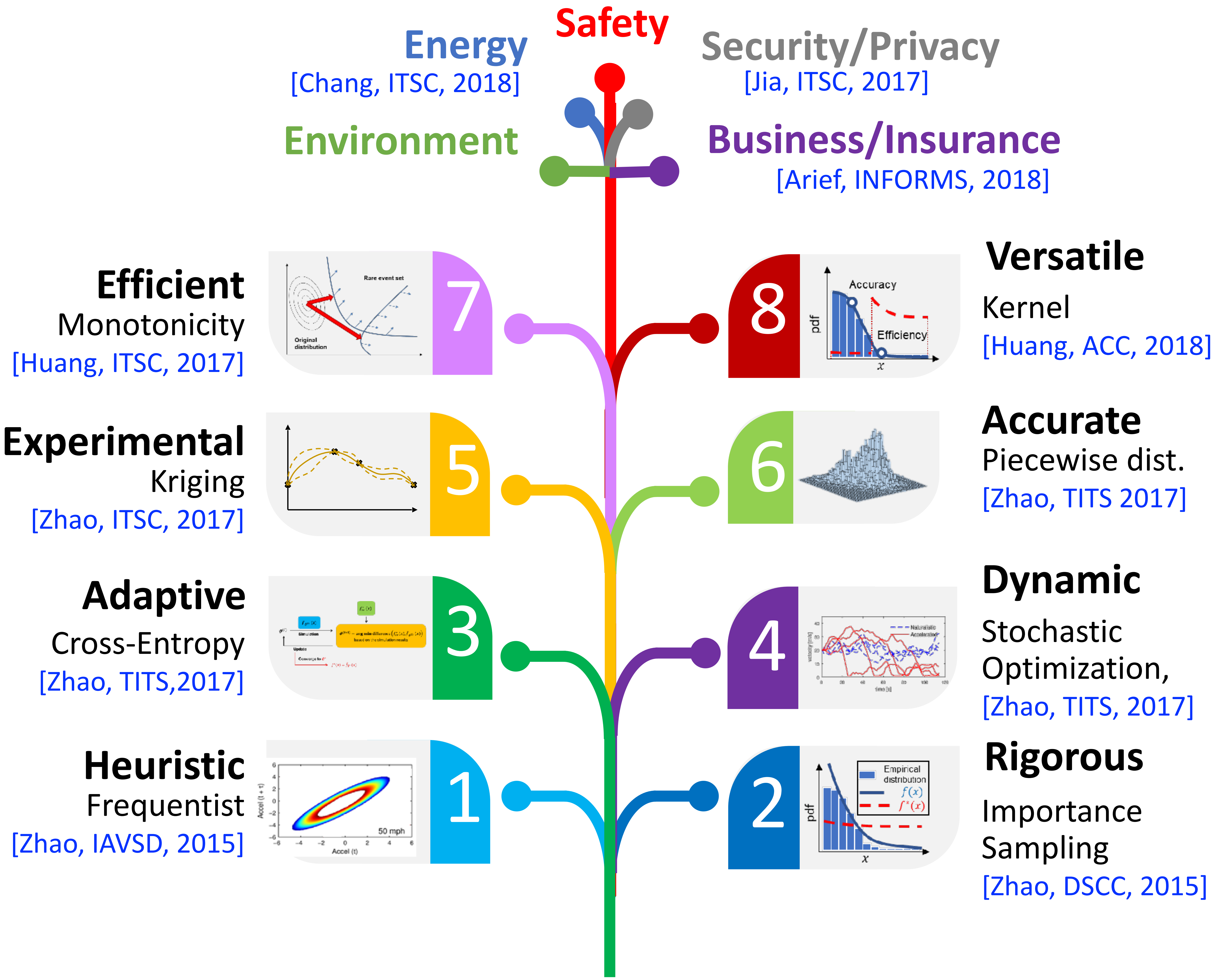
Nidhi Kalra, Susan M. Paddock, “How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? RAND report 2016

Accelerated Evaluation

"Development of provable autonomous vehicle **evaluation approaches** with efficient data collection, unsupervised analysis, and high-dimensional stochastic models of on-road driving environment" (Uber, PI)

"Development of efficient multi-model **annotation and checking tools** based on synthesized learning methods" (Bosch, PI)

"Development of a "primary other **test vehicle**" for the testing and evaluation of high-level automated vehicles" (Toyota, Co-PI)



M city
UNIVERSITY OF MICHIGAN

From the Lab to the Street:
Solving the Challenge of Accelerating Automated Vehicle Testing

DING ZHAO, PhD
Assistant Research Scientist
Mechanical Engineering
University of Michigan

HUEI PENG, PhD
Director, Mcity
Roger L. McCarthy Professor
of Mechanical Engineering
University of Michigan

EXECUTIVE SUMMARY

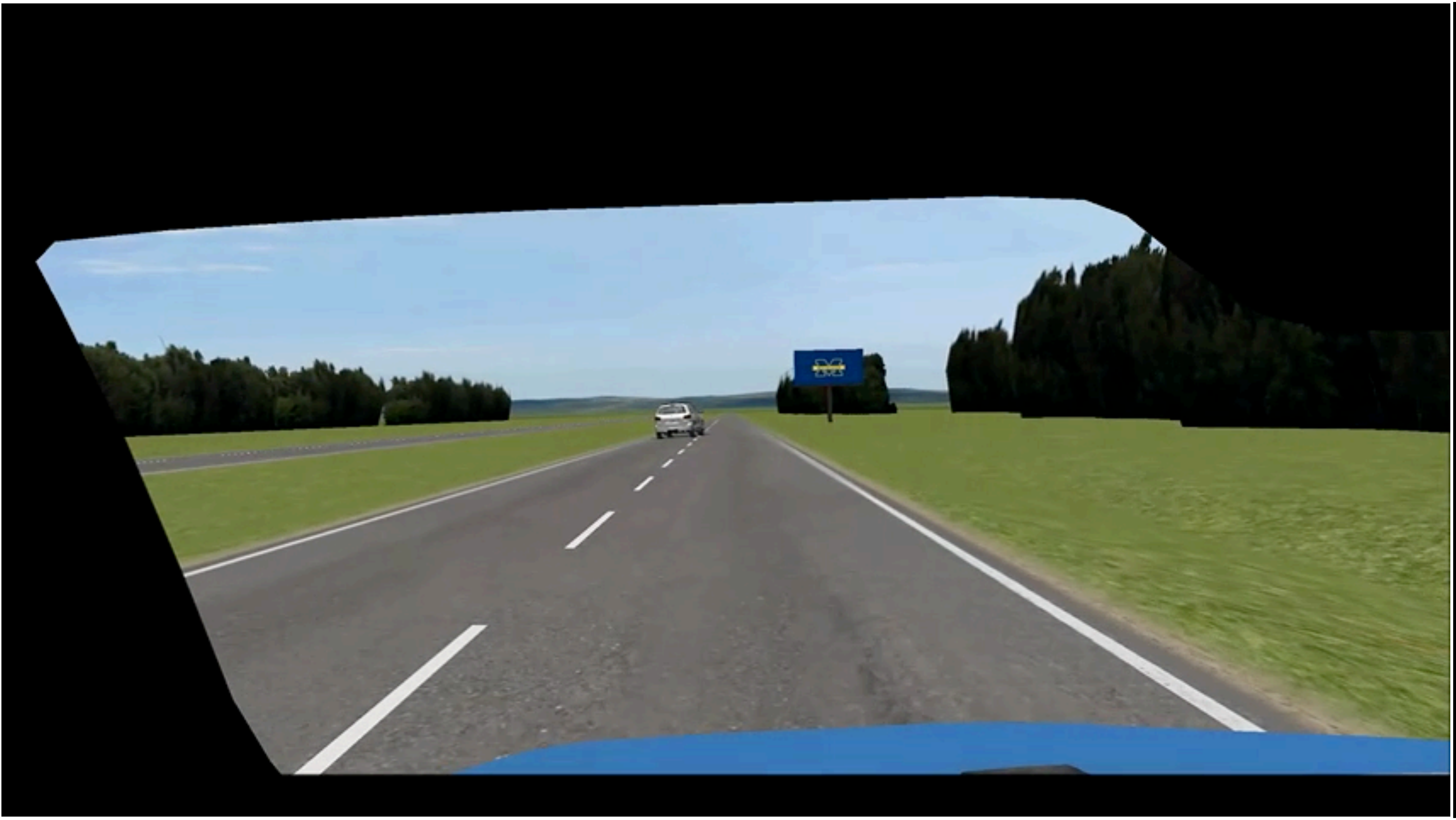
...hicles and their technology become more advanced and technically ... easure the safety and reliability of these ... accurate

Key idea

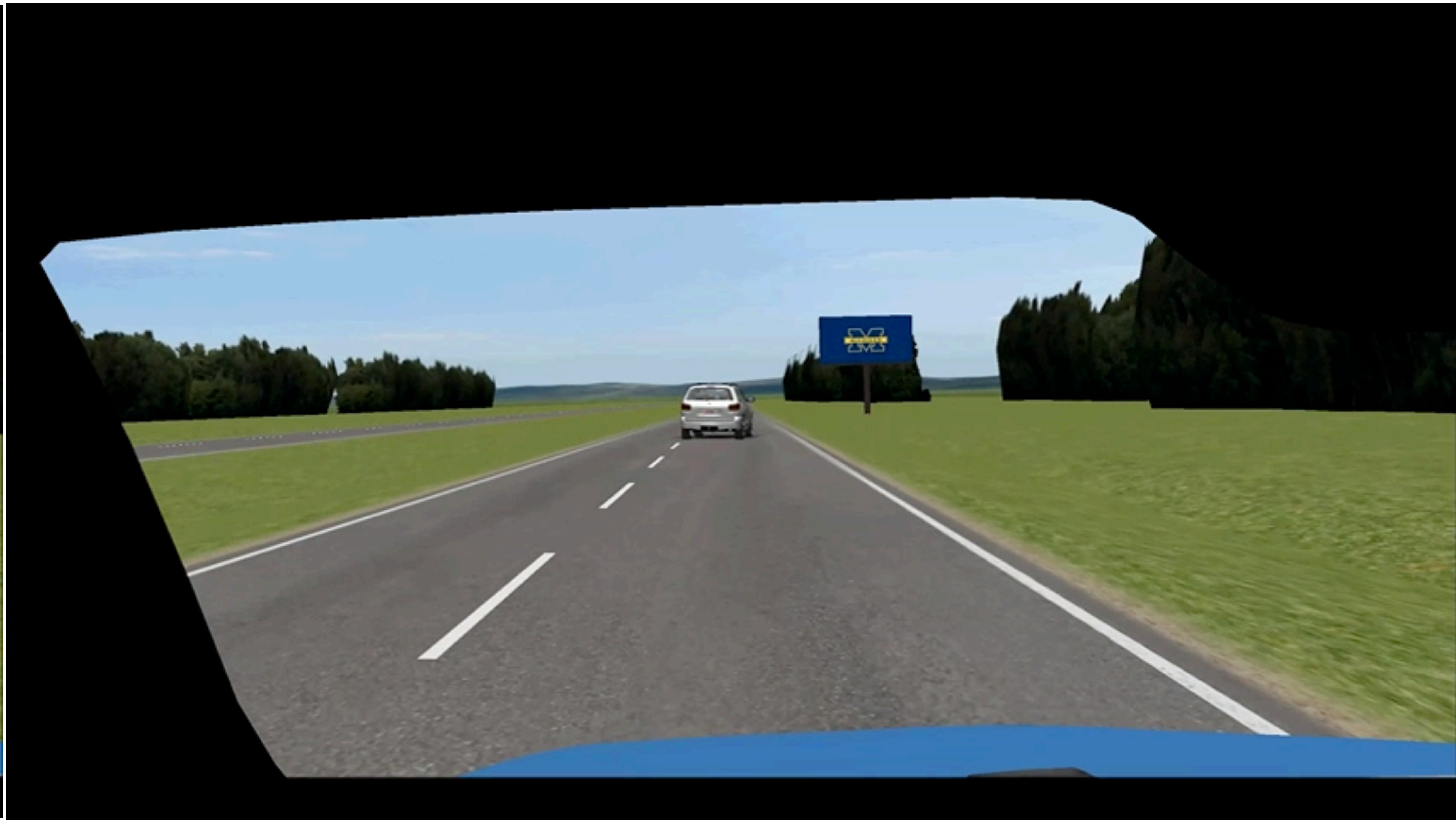
- Give more test budgets to scenarios that may most likely fail AVs and also most likely happen in the real world
  - Likelihood of scenario in the real world <- models of real world data
  - Likelihood of failure <- AV-in-the-loop tests (physical/simulation)

Naturalistic environment vs accelerated environment

Naturalistic Environment



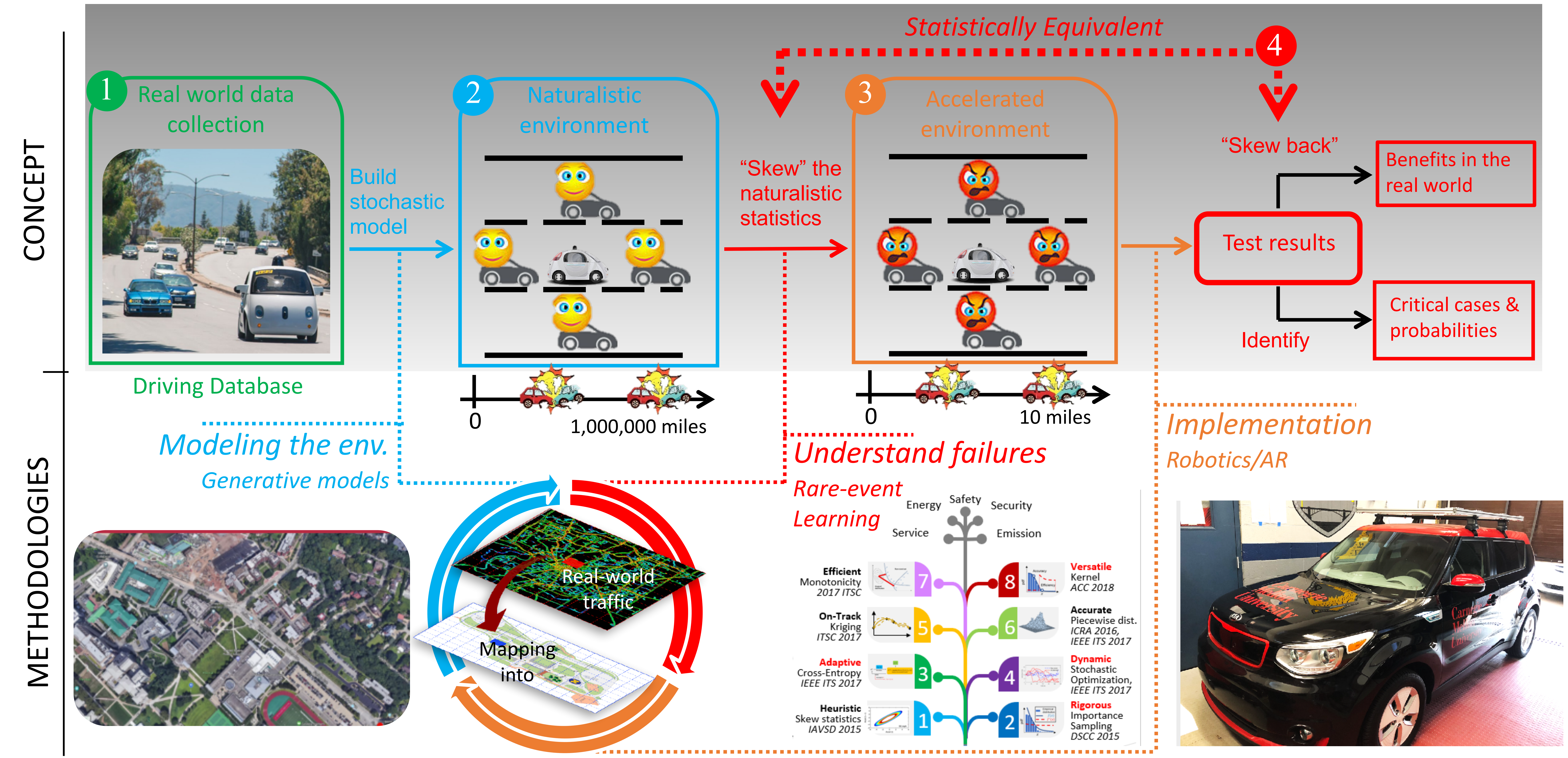
Accelerated Environment



TOYOTA
RESEARCH INSTITUTE

Zhao, ''Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques", IEEE ITS, 2017.

Accelerated Evaluation

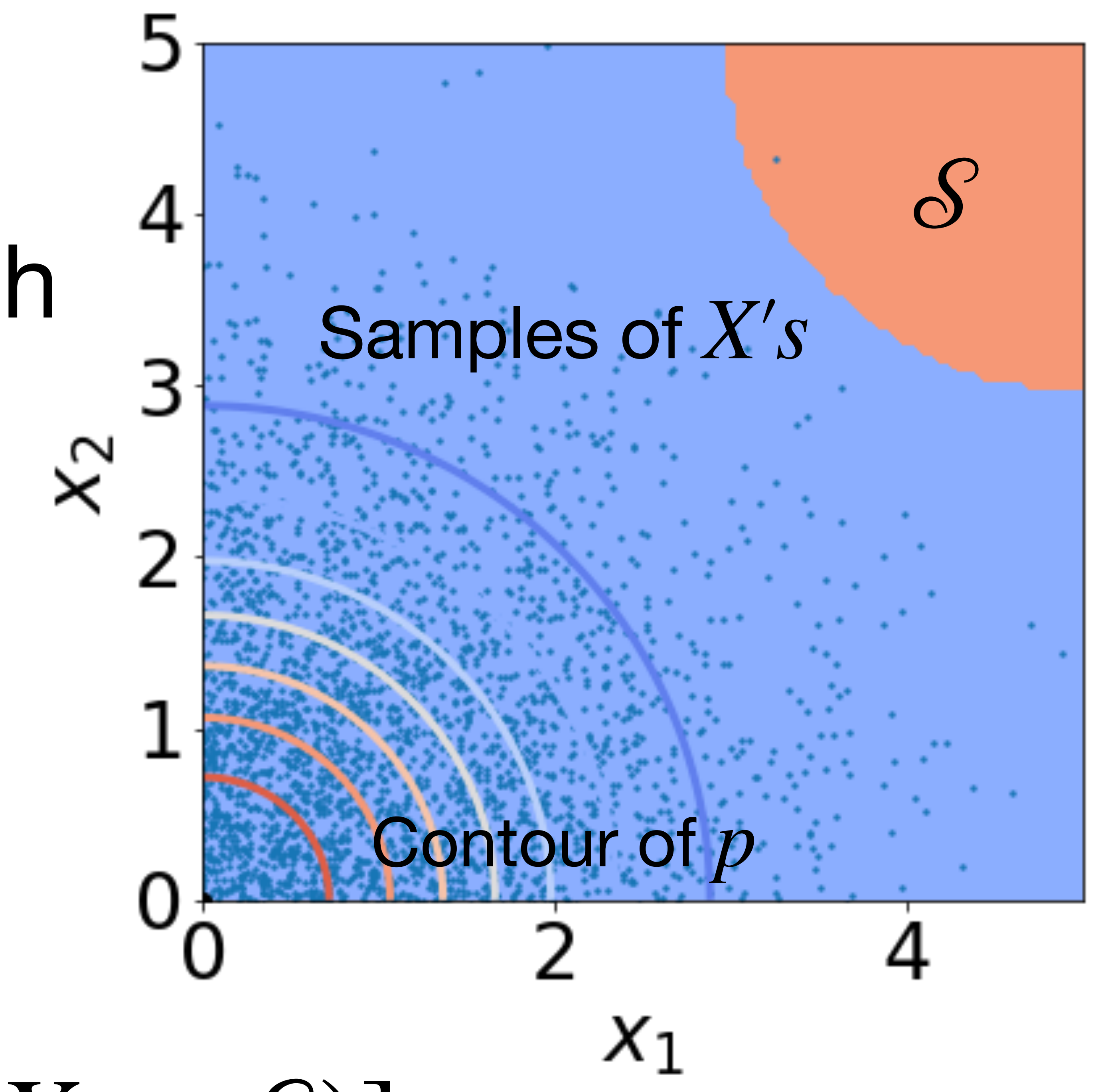


Zhao, "Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques", IEEE ITS, 2017.

Probabilistic Adversarial Sampling

- Suppose we want to estimate the probability of dangerous events $\mathcal{S}$
- Input: $X =$ random initial distance and relative velocity
- Output: $Y =$ simulation outcome, either crash or not crash

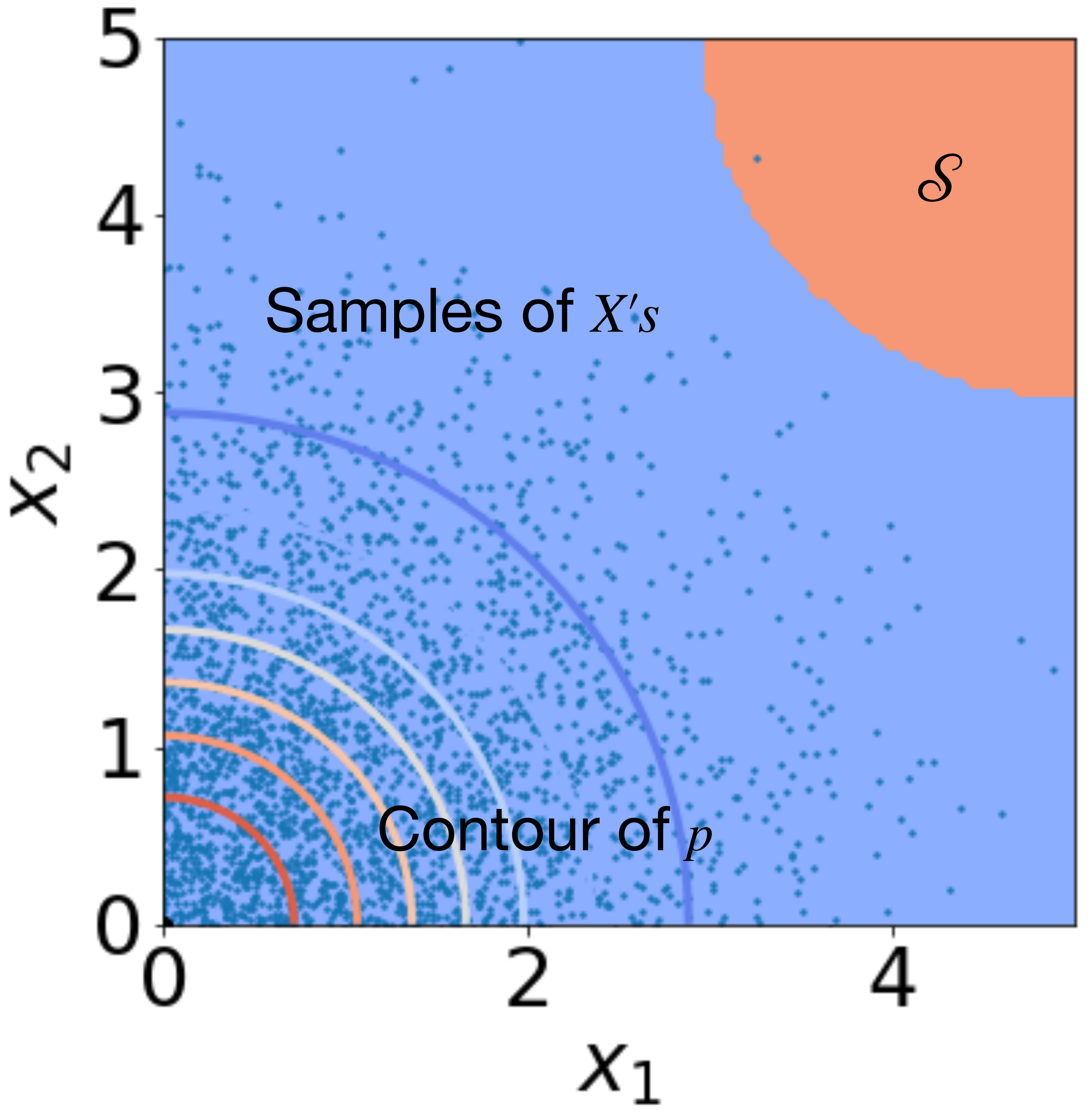
$$Y = f(X) = \begin{cases} 1, & \text{crash} \\ 0, & \text{not crash} \end{cases}$$



- Crash or dangerous set: $\mathcal{S} = \{X : f(X) = 1\}$
- Goal: Estimating $\mu = P(Y = 1) = P(X \in \mathcal{S}) = \mathbb{E}_{X\sim p}[1(X \in \mathcal{S})]$

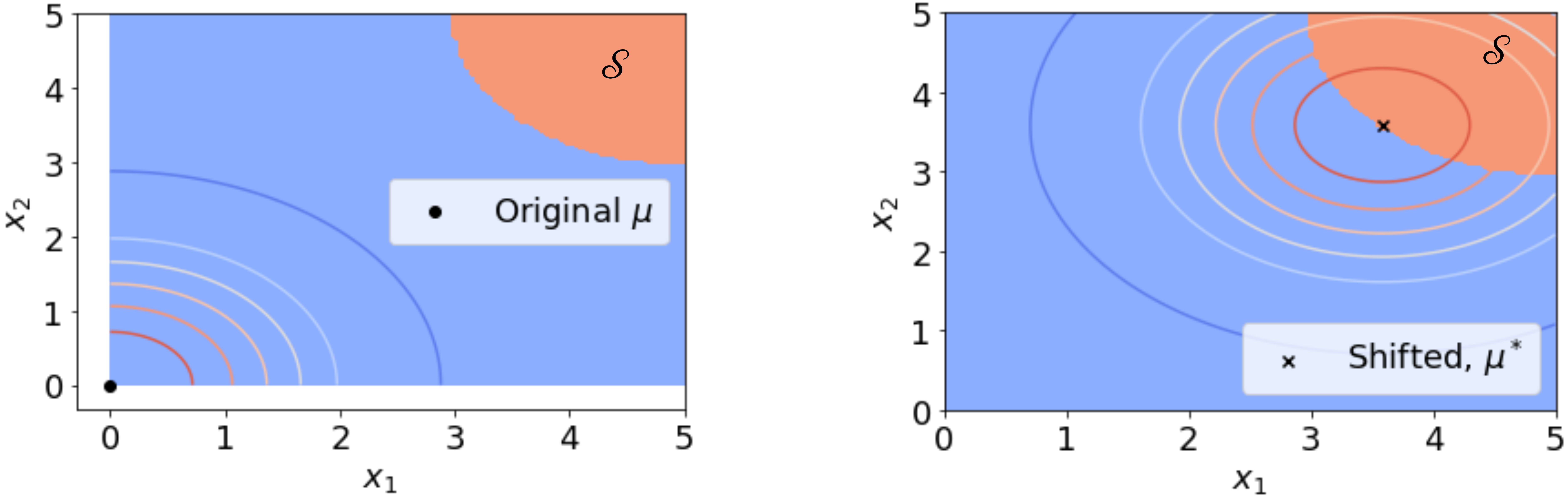
Monte Carlo (MC) sampling

- Monte Carlo procedure for estimating $\mu = \mathbb{E}_{X \sim p}[1(X \in \mathcal{S})]$:
  - generate n i.i.d samples $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$, where $X^{(i)} \sim p$
  - observe $Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)}$, where $Y^{(i)} = f(X^{(i)})$
  - compute sample average (MC estimator) $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)}$
- Note that
  - $\mathbb{E}_{X \sim p}[\hat{\mu}_n] = \int f(x)p(x)dx = \mu$ (unbiased)
  - Var$(\hat{\mu}_n) = \dfrac{\mu(1-\mu)}{n}$ (shrinking in n)



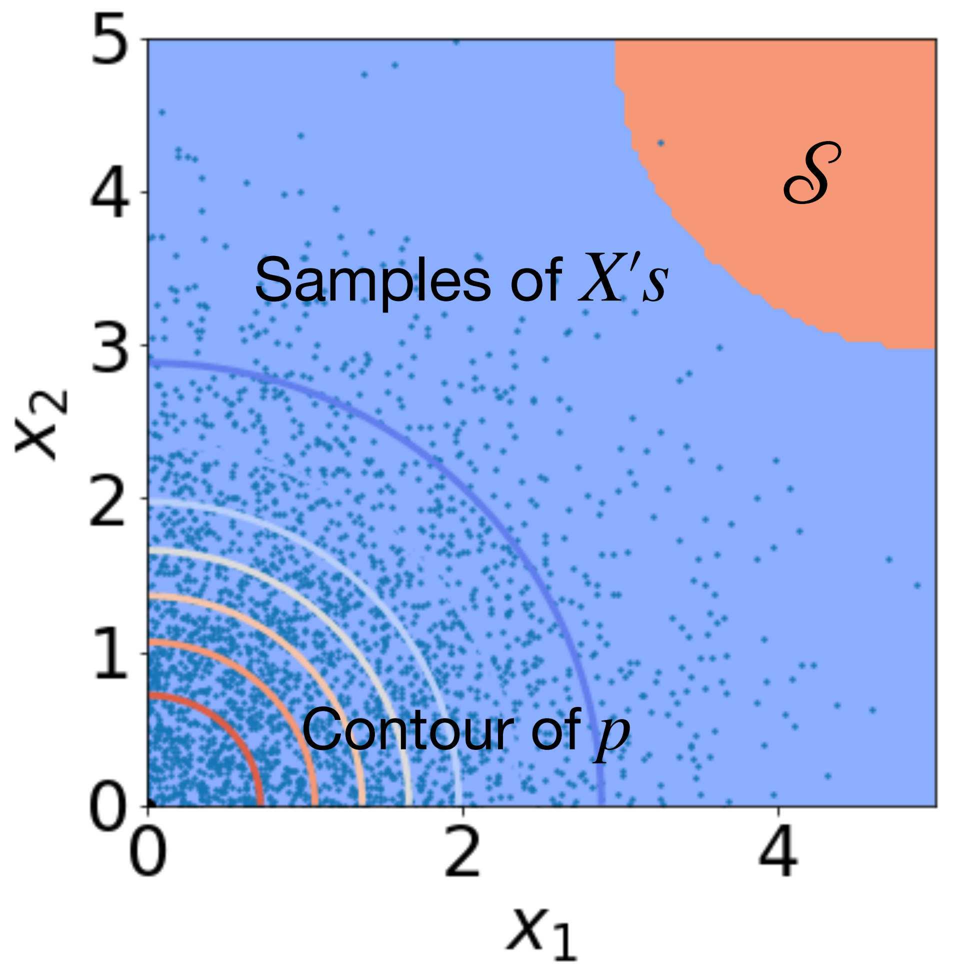
Probabilistic Accelerated evaluation: Framework

- Four elements: $< f, p, \mathcal{S}, q >$
  - Design of q is related to key characteristics of the problem $< f, p, \mathcal{S} >$
  - If $\mathcal{S}$ has a single dominating point, an analytical efficient solution can be found



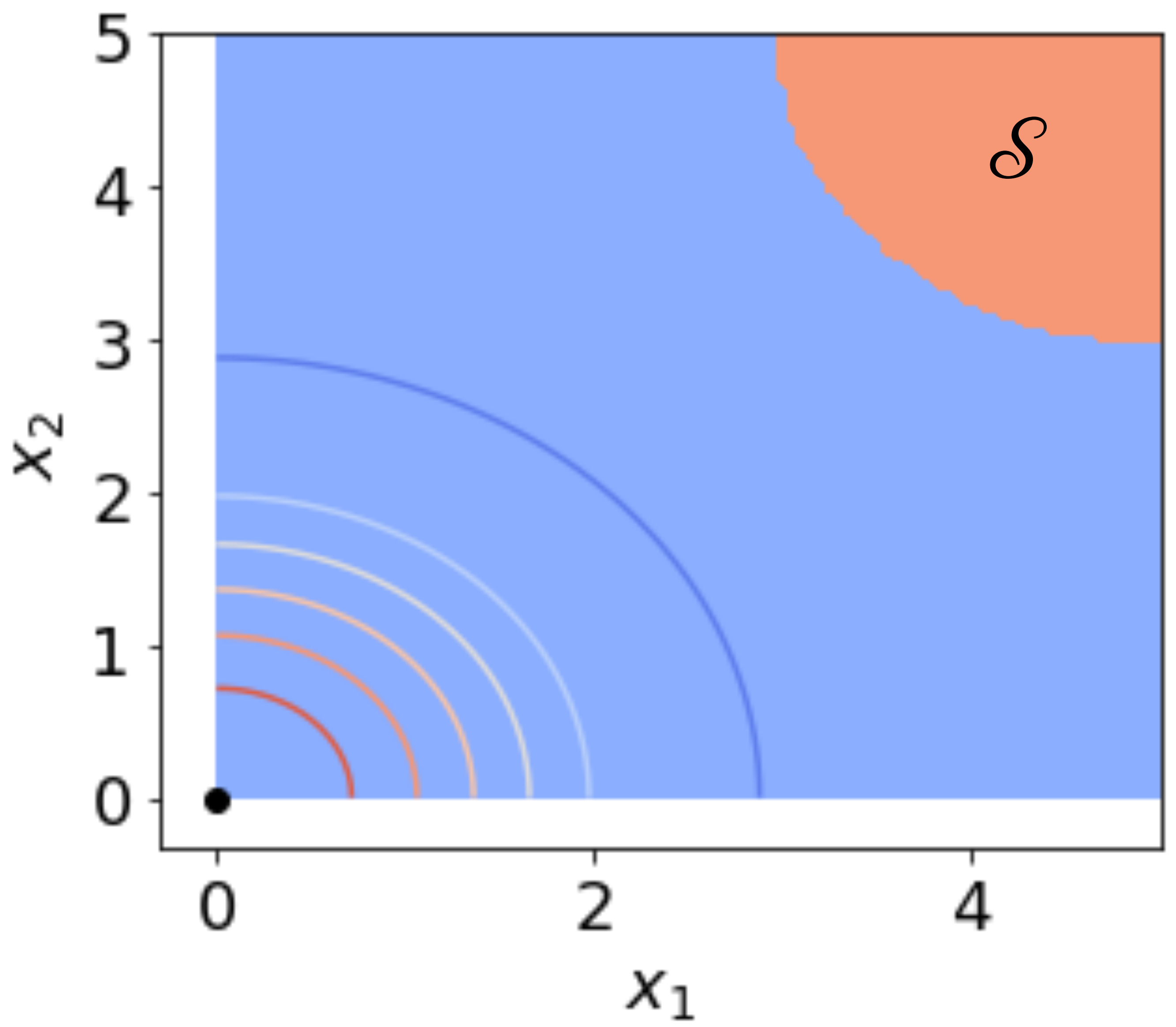
Importance Sampling (IS)

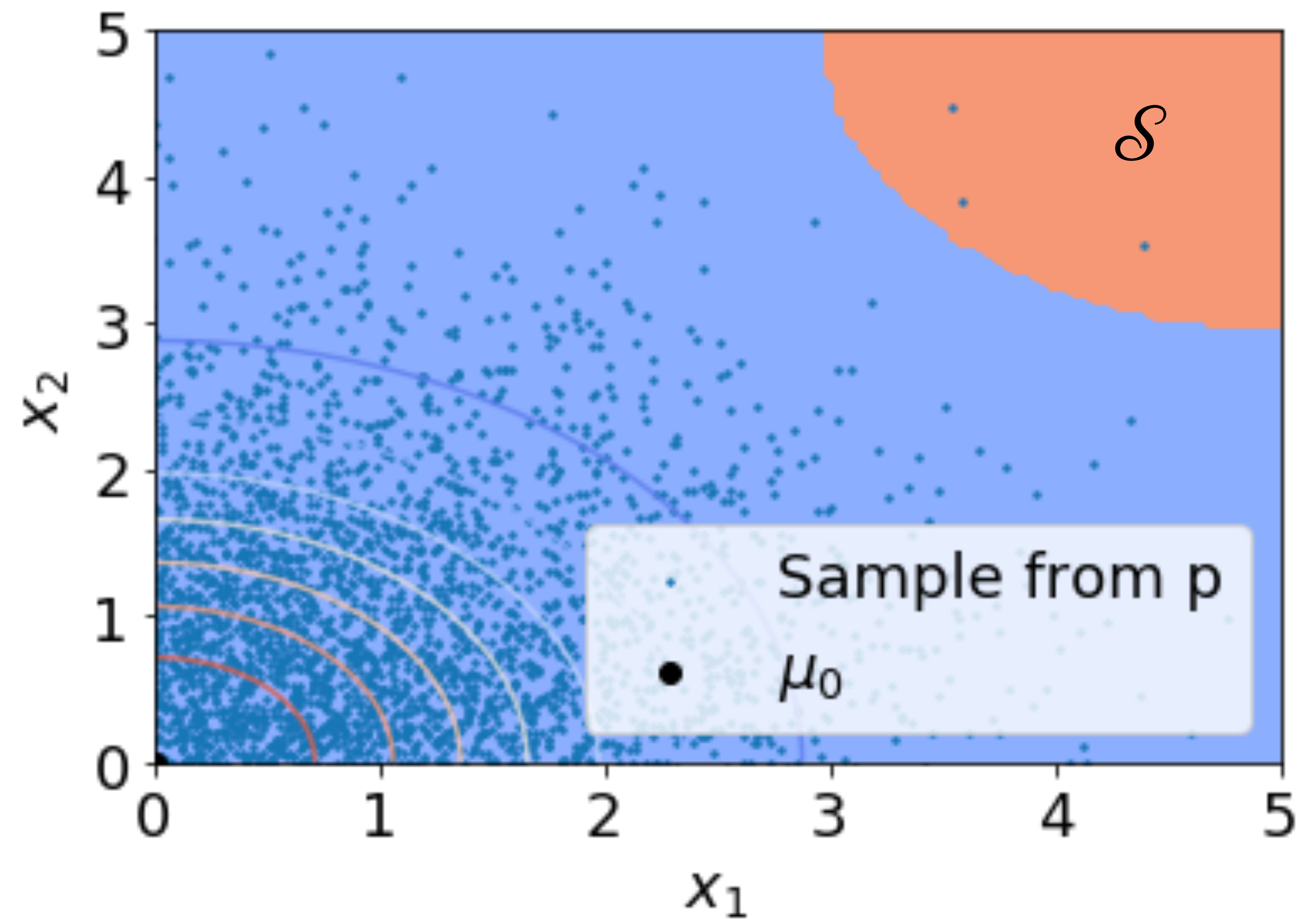
- IS procedure for estimating $\mu = \mathbb{E}_{X\sim p}[1(X \in \mathcal{S})]$:
  - generate n i.i.d samples $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$, from another distribution $X^{(i)} \sim \tilde{p}$
  - observe $Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)}$, where $Y^{(i)} = f(X^{(i)})$
  - compute likelihood ratio $W^{(1)}, W^{(2)}, \cdots, W^{(n)}$, where $W^{(i)} = \frac{p(X^{(i)})}{\tilde{p}(X^{(i)})}$
  - compute weighted average (IS estimator) $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y^{(i)} W^{(i)}$
- Note that $\mathbb{E}_{X\sim\tilde{p}}[\hat{\mu}_n] = \int f(x)\left(\frac{p(x)}{\tilde{p}(x)}\right)\tilde{p}(x)dx = \int f(x)p(x)dx = \mu$ (unbiased)

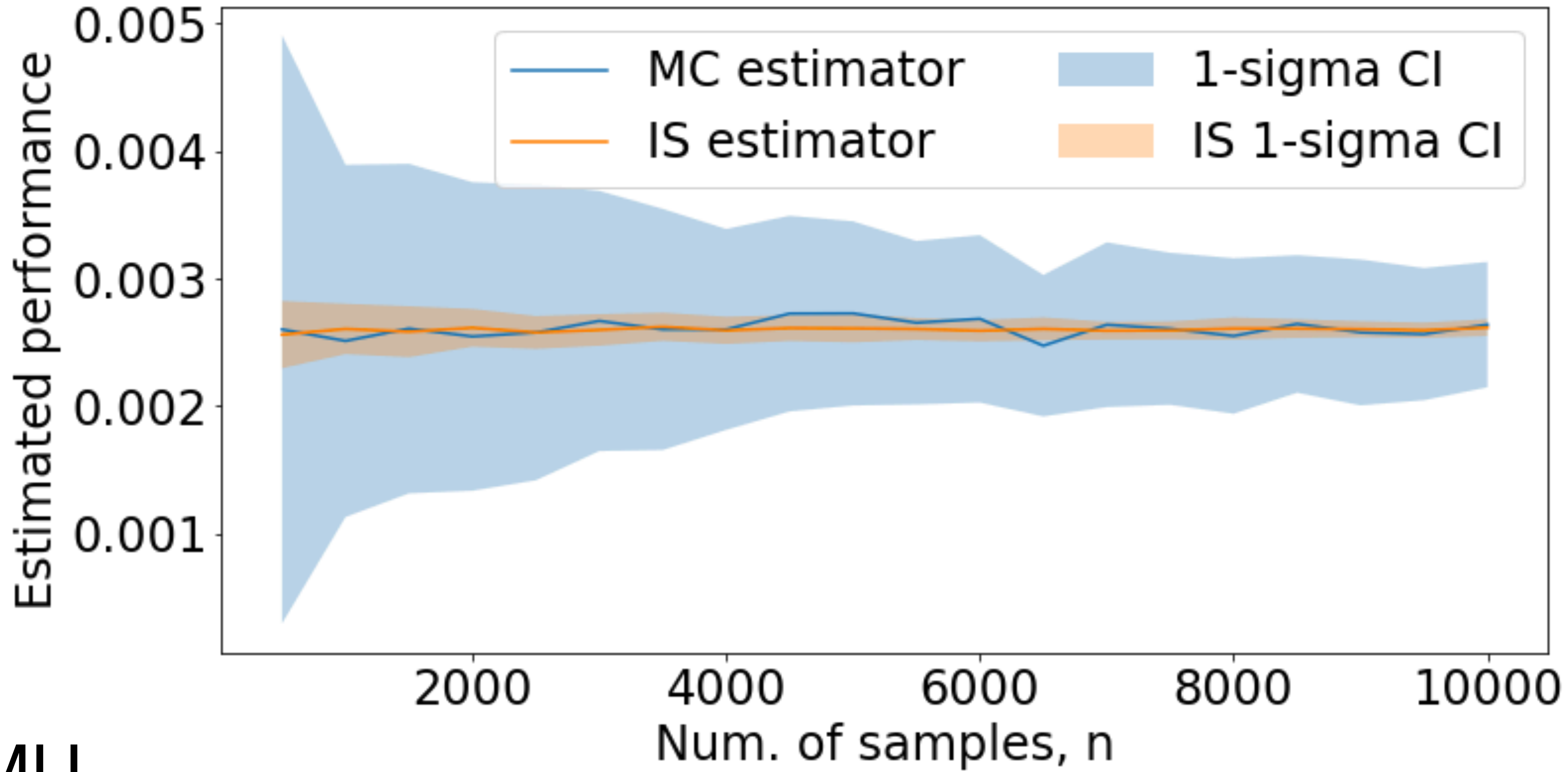


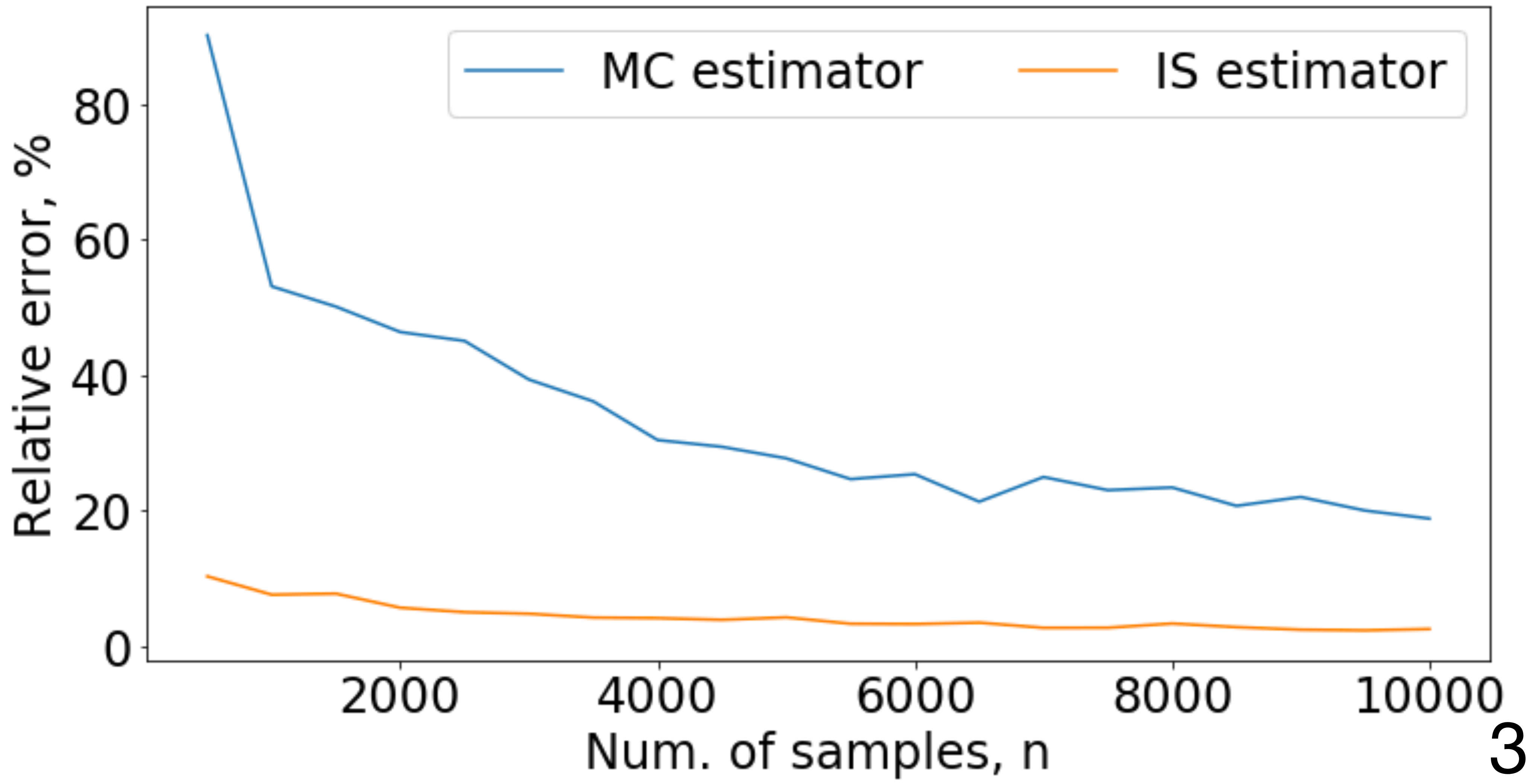
Deep IS: Toy Example

- **Example:** Suppose we want to estimate the probability $\mu_Y = \mathbb{E}[f(X)] = P(X \in \mathcal{S})$ for some set $\mathcal{S} \subset \mathbb{R}^2$ Suppose that $X \sim p$ where p is a Gaussian centered at [0, 0]



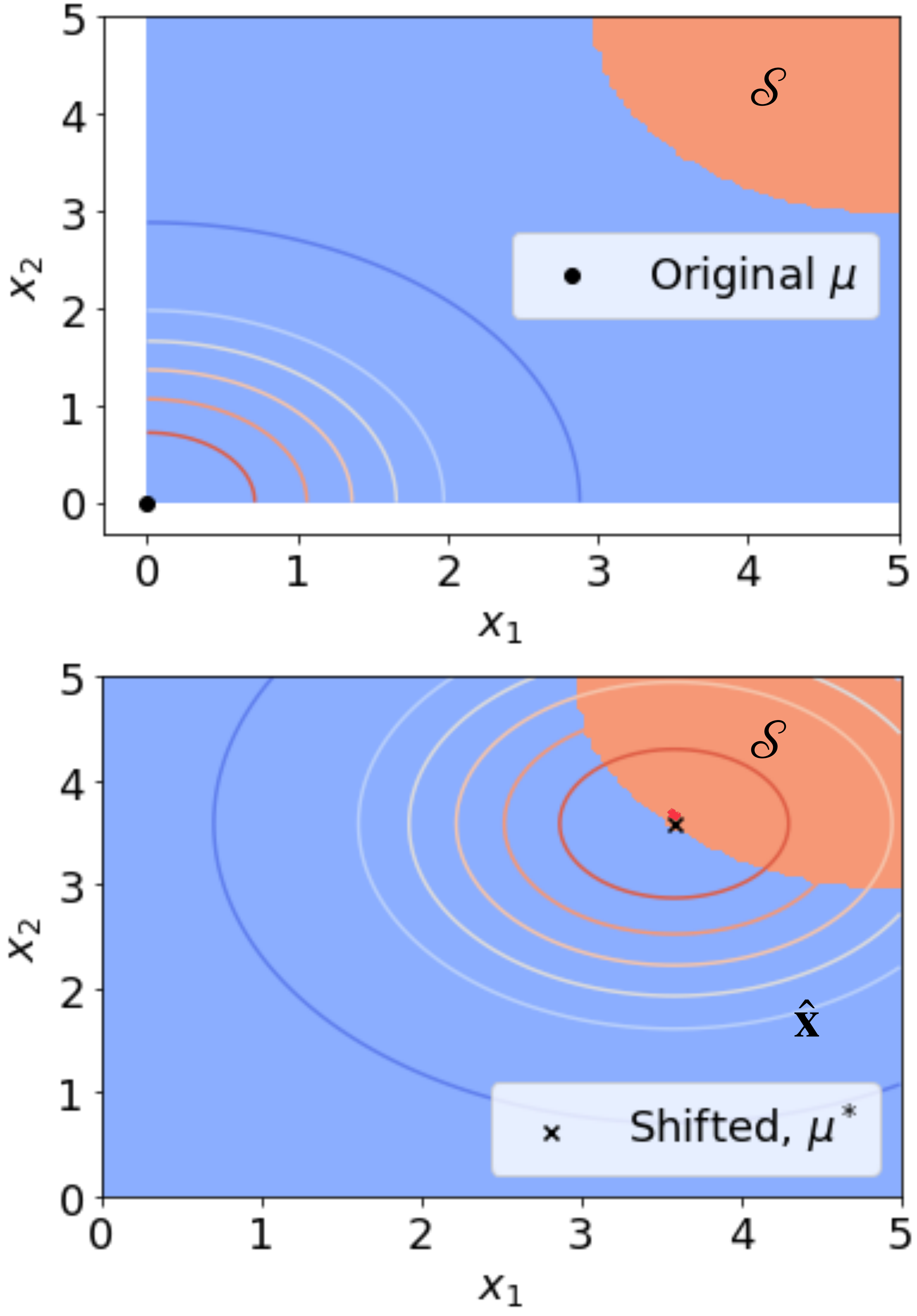






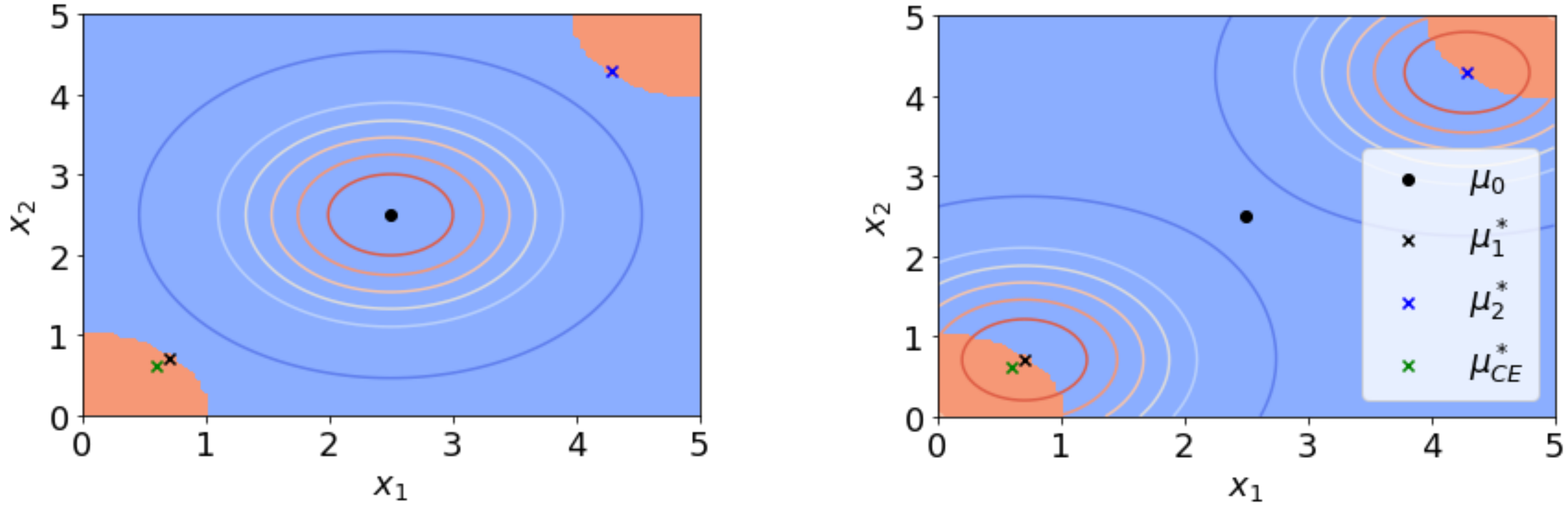
Dominant points

- Dominating point x^* of the set $\mathcal{S}$ with respect to density p is defined as $x^* = \arg\max_{x \in \mathcal{S}} p(x)$



Multiple dominating points issue with iterative methods (CE)

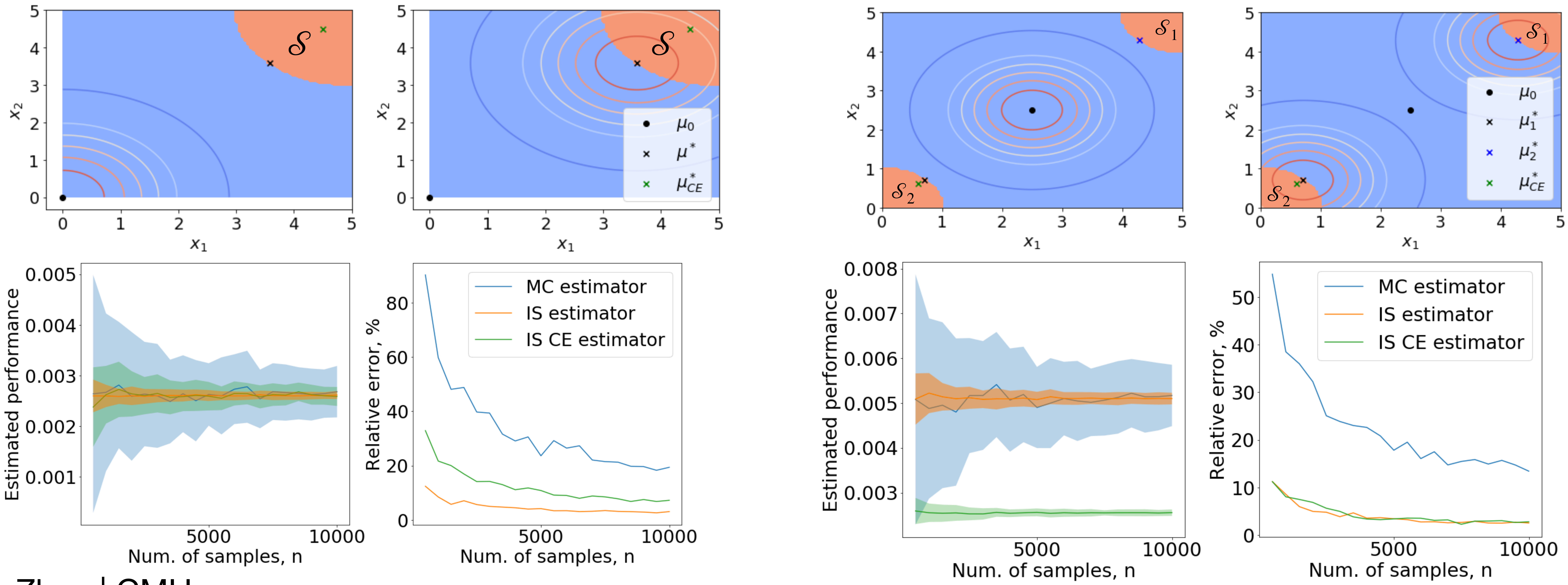
- One of the main challenges with the traditional iterative methods (Cross Entropy) is selecting and optimizing over the parametric class $\mathcal{Q} = \{q_\theta, \forall \theta \in \Theta\}$
- An overly simple $\mathcal{Q}$ may result in a biased estimator, e.g. in multiple dominating point $\mathcal{S}$ case



Arief, Mansur, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, Henry Lam, and Ding Zhao. "Deep Probabilistic Accelerated Evaluation: A Certifiable Rare-Event Simulation Methodology for Black-Box Autonomy." To appear in the Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, 2021.

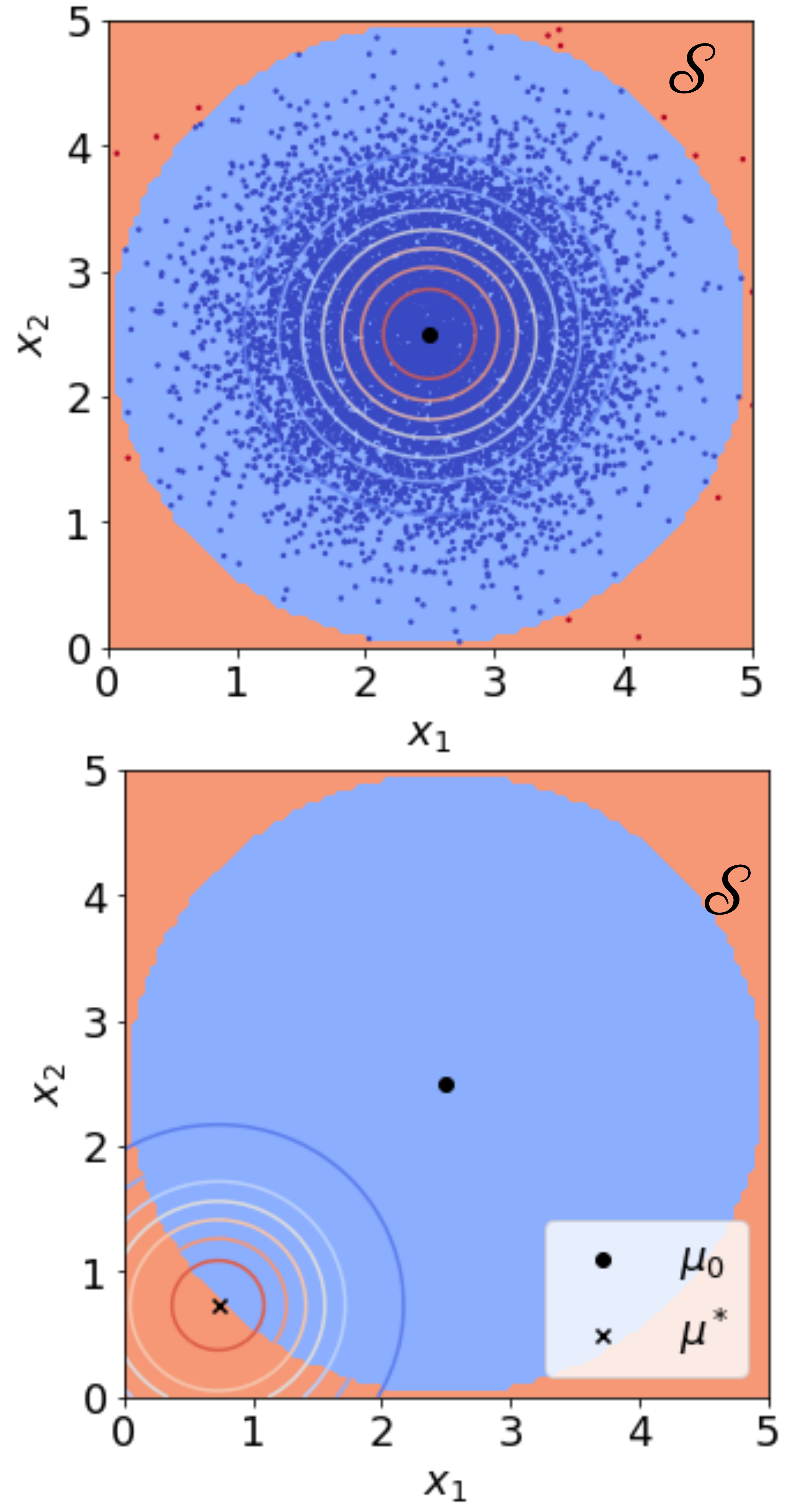
GMM-PrAE: Using GMM for the multiple dominating points case

- If $\mathcal{S} = \cup_{j=1}^{J} \mathcal{S}_j$ in which all $\mathcal{S}_j$'s are convex, then a Gaussian Mixture (GMM) with component means shifted to cover all $\mathcal{S}_j$'s is efficient



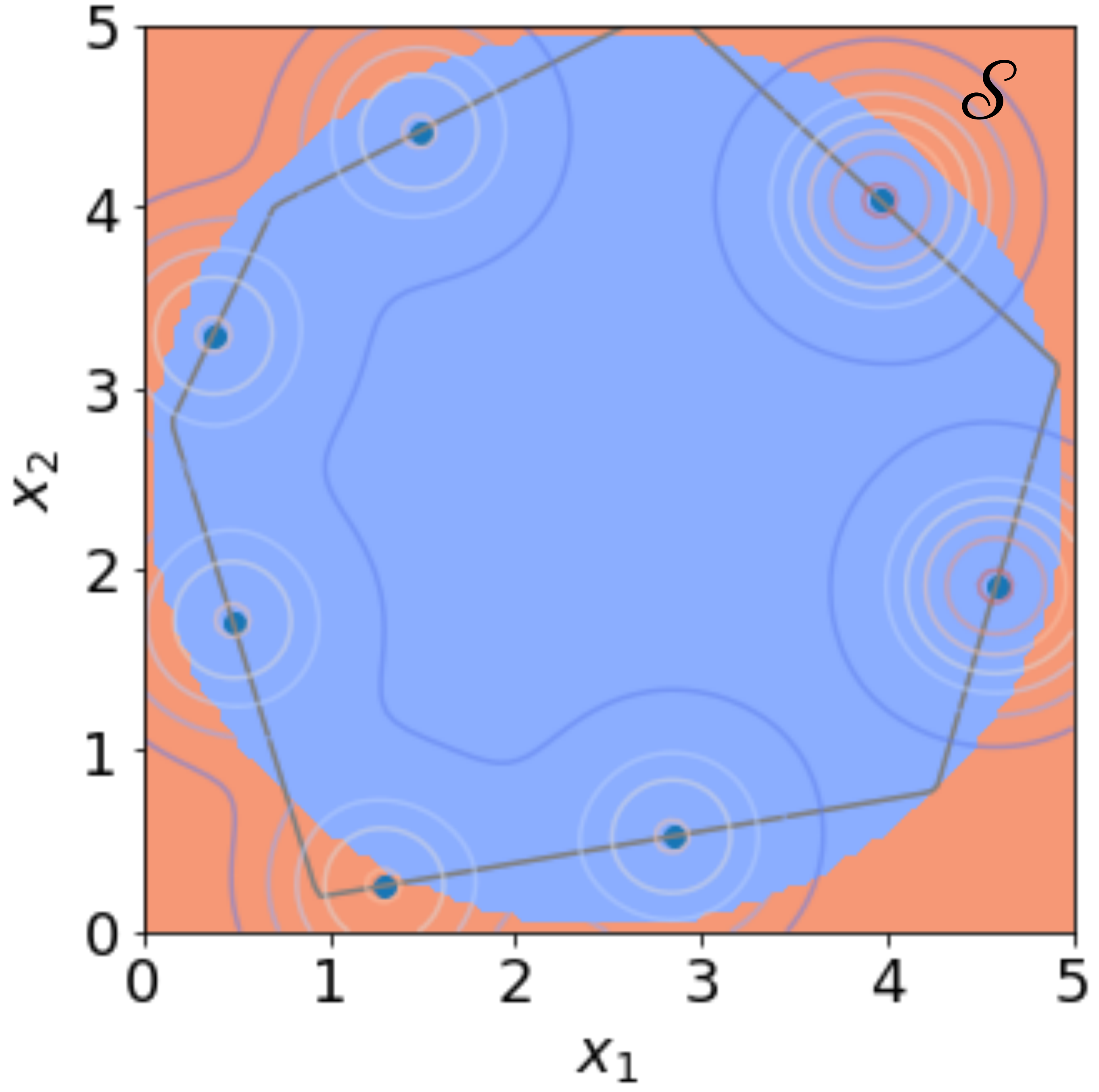
What if there exist a lot of (infinite) dominating points?

- What about other cases? $\mathscr{S}$ may have no or infinite dominating points
- Previous approach would suggest infinite-component GMM



Deep-IS: Deep learning based PrAE

- Designing q via deep learning classifier for monotonic rare-event set
  - Train a conservative classifier with piecewise linear decision boundary (ReLU)
  - Sufficiently prune or simplify the model
  - Find the dominating point w.r.t. classifier decision boundary and p
  - Construct GMM-based q with these dominating points



Arief, Mansur, Zhiyuan Huang, Guru Koushik Senthil Kumar, Yuanlu Bai, Shengyi He, Wenhao Ding, Henry Lam, and Ding Zhao. "Deep Probabilistic Accelerated Evaluation: A Certifiable Rare-Event Simulation Methodology for Black-Box Autonomy." To appear in the Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, 2021.

Summary

- Adversarial scenario generation
  - GAN
  - GAN+prior
  - IS-based method (Accelerated Evaluation)

Worth reading

- Zhao, Ding, and Huei Peng. "From the lab to the street: Solving the challenge of accelerating automated vehicle testing. https://mcity.umich.edu/wp-content/uploads/2017/05/Mcity-White-Paper_Accelerated-AV-Testing.pdf
- Waymo Safety Report, 2020. https://storage.googleapis.com/sdc-prod/v1/safety-report/2020-09-waymo-safety-report.pdf
- Corso, A., Moss, R.J., Koren, M., Lee, R. and Kochenderfer, M.J., 2020. A survey of algorithms for black-box safety validation. https://arxiv.org/abs/2005.02979