DATA-DRIVEN SIGNAL RECOVERY

Yury Polyanskiy (MIT)

joint with a (large) bunch of people

Talk plan

1. Thesis
2. Theory
3. Practice

1. Let's move beyond iid Gaussian noise
2. Likelihood-free hypothesis testing
3. Physics, Computer Vision, Comm

Classical detection and estimation

Adding a splash of 21st century

- How do we teach signal detection?

$$H_0 : Y_i \sim \mathcal{N}(1,\sigma^2) \qquad\qquad H_1 : Y_i \sim \mathcal{N}(-1,\sigma^2)$$

  and threshold the average $\frac{1}{m}\sum_i Y_i \gtrless 0$

- ... more generally:

$$H_0 : Y^m = s_0 + Z^m \qquad\qquad H_1 : Y^m = s_1 + Z^m$$

  and do matched filter: $(Y^m, s_1 - s_0) \gtrless 0$

- ... more generally:

$$H_0 : Y^m \sim P_{Y^m} \qquad\qquad H_1 : Y^m \sim Q_{Y^m}$$

  and do Neyman-Pearson

- ... more generally:

$$H_0 : Y^m \sim P, P \in \mathcal{P} \qquad\qquad H_1 : Y^m \sim Q, Q \in \mathcal{Q}$$

  and do what?..

- Try GLRT, otherwise search Annals of Stats

- Problem: if $\mathcal{P}$, $\mathcal{Q}$ are realistic (i.e. large), then sample complexity is **bad** (curse of dimensionality etc)

Thesis: often we have side information (prior knowledge) about P_{Y^m} Q_{Y^m} in the form of iid samples.

Science: simulations

Communication: RF captures

What is likelihood-free inference?

aka simulation-based inference

What is likelihood-free inference (LFI)?

aka simulation-based inference (SBI)

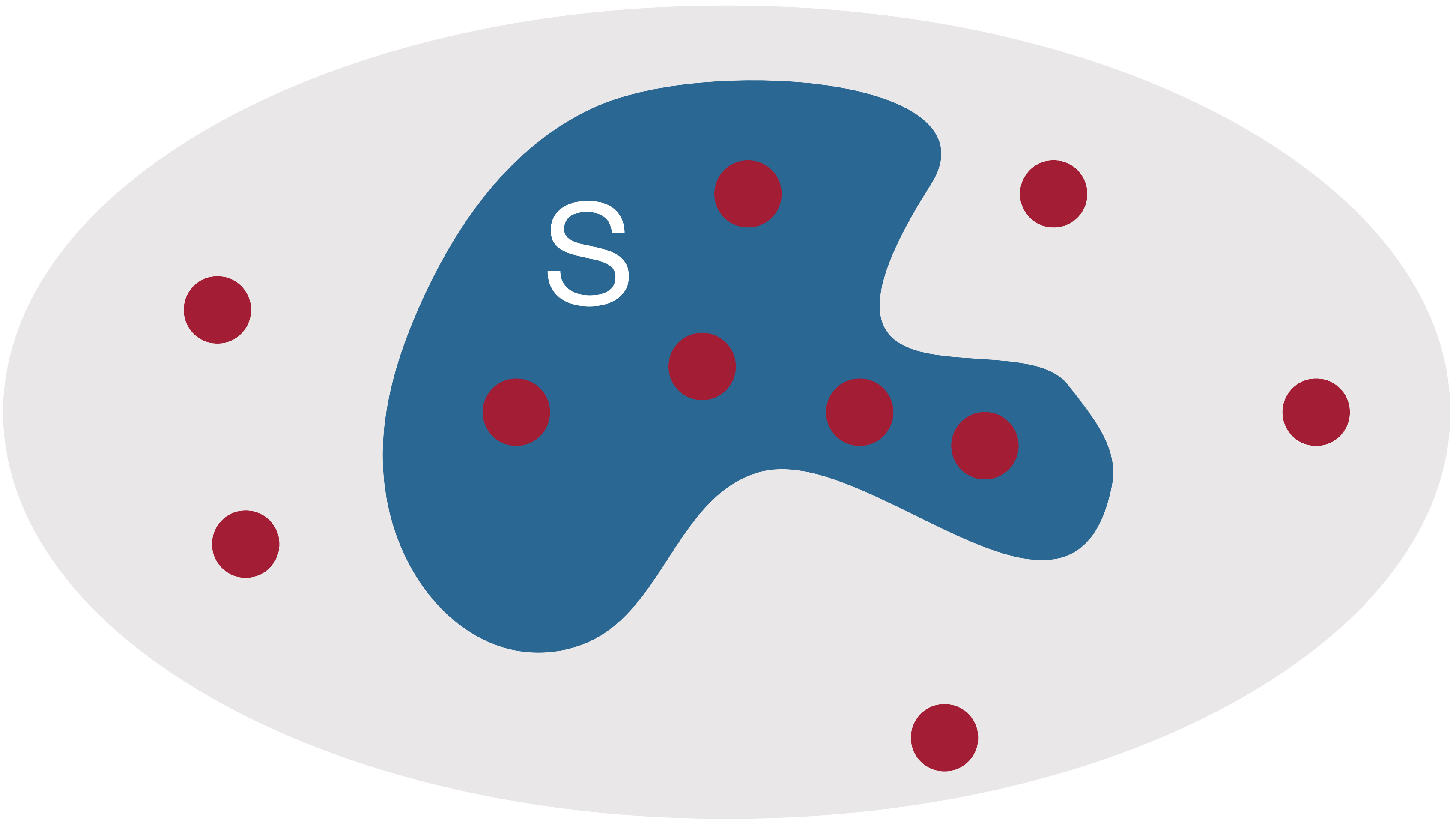
- Simulation access to **black-box** model $\theta \mapsto X \sim \mathbb{P}_\theta$
- Given true data $Z \sim \mathbb{P}_{\theta^\star}^{\otimes m}$, do **inference on** $\theta^\star$
- Intractable likelihood: do so **without learning** the map $\theta \mapsto \mathbb{P}_\theta$
- Examples: climate modeling and particle physics

Discovery of the Higgs boson

- Observe data $Z \sim \mathbb{P}^{\otimes m}$

$$H_0 : \mathbb{P} = \mathbb{P}_{\text{noHiggs}} \quad \text{versus} \quad H_1 : \mathbb{P} = \mathbb{P}_{\text{Higgs}}$$

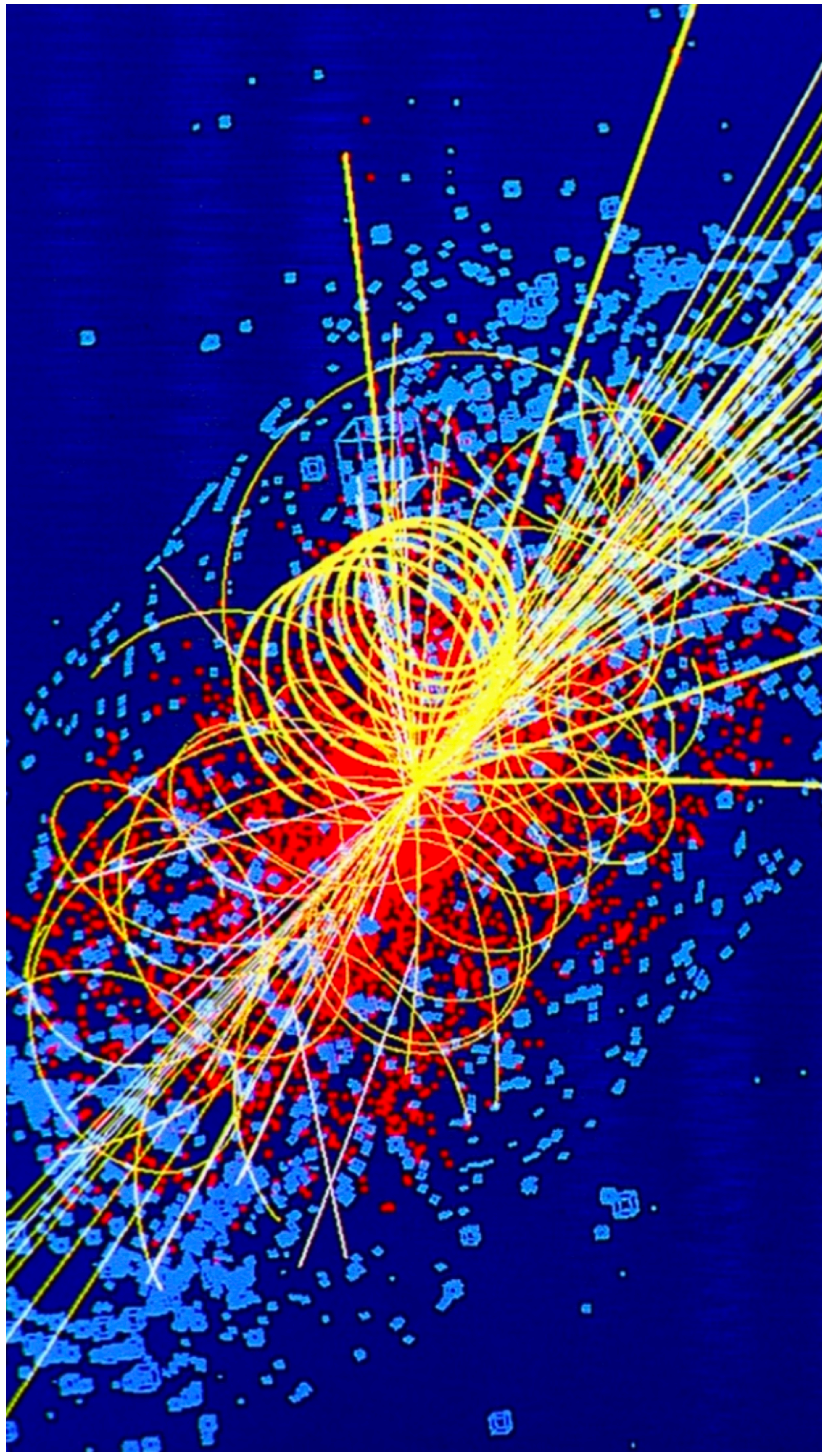
- Simulate $X \sim \mathbb{P}^{\otimes n}_{\text{noHiggs}}$ and $Y \sim \mathbb{P}^{\otimes n}_{\text{Higgs}}$



● $= Z_i$

S = classifies* X vs Y

Output = $\#\{$●$\in S\} \lessgtr \gamma$



Simulation of the birth of a Higgs boson (CERN, Lucas Taylor)

*Boosted decision trees in the case of the Higgs boson discovery

Minimax setup

Likelihood-free hypothesis testing (LFHT)

1) Fix

2) Let

3) Sim

Questions we will address:

- Can we avoid learning distributions P_X, P_Y?
- Is there a tradeoff m vs n?

4) Depending on H_0 or H_1 nature generates $Z \sim \mathbb{P}_X^{\otimes m}$ or $\mathbb{P}_Y^{\otimes m}$ respectively

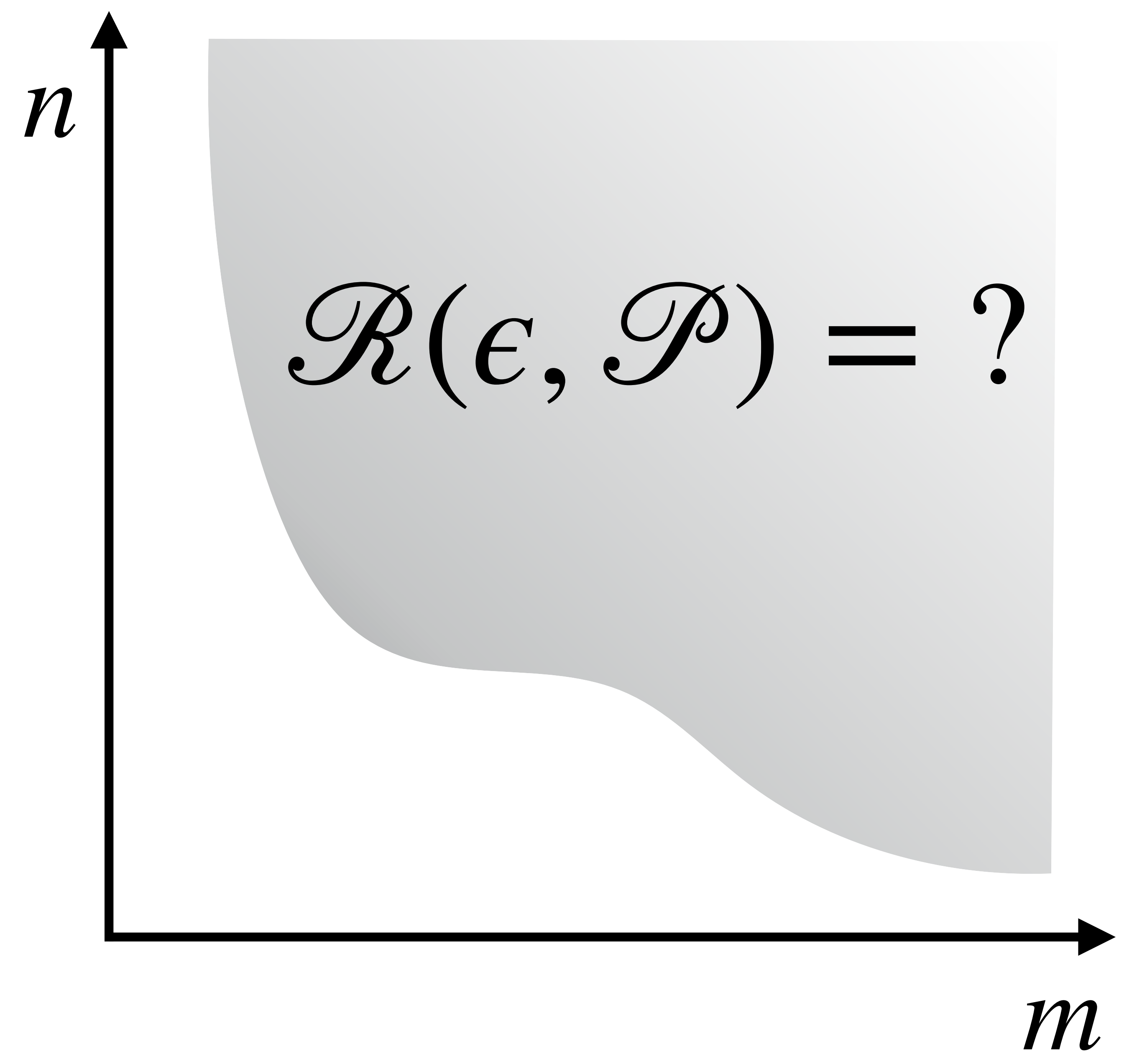
5) Statistician observes $(X, Y, Z, \mathscr{P}, \epsilon)$ and decides H_0 or H_1

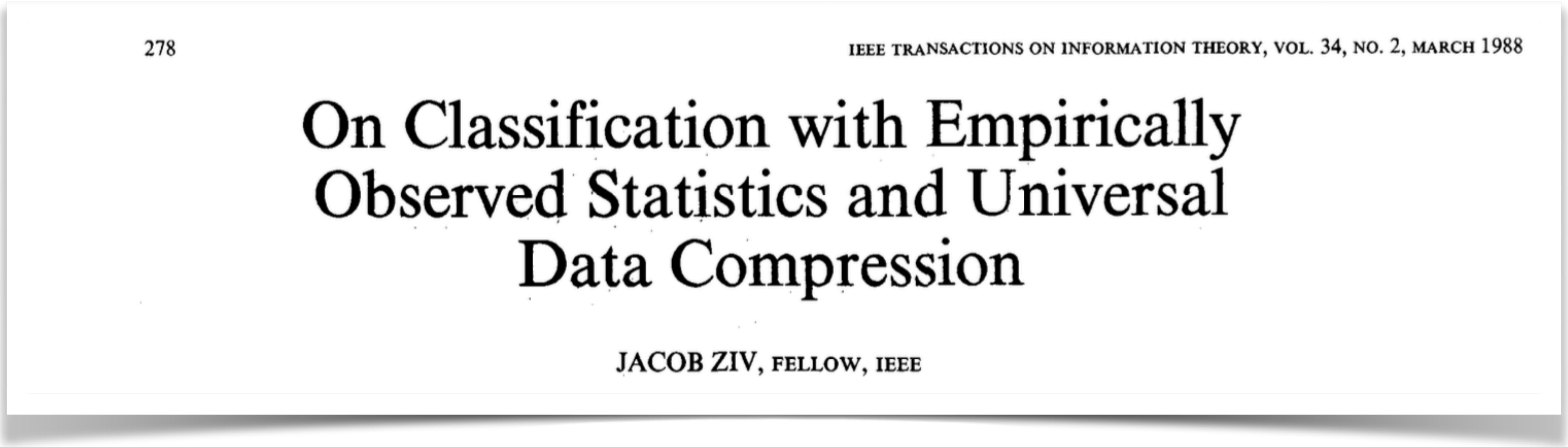
Likelihood-free hypothesis testing (LFHT)

$\mathscr{R}(\epsilon, \mathscr{P}) \subseteq \mathbb{N}^2$ is set of (m, n) s.t. exists test that given X, Y, Z performs

$$H_0 : \mathbb{P}_\mathsf{X} = \mathbb{P}_\mathsf{Z} \qquad \text{versus} \qquad H_1 : \mathbb{P}_\mathsf{Y} = \mathbb{P}_\mathsf{Z}$$

with (Type-I + Type II error) $< 1\ \%$.




278 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 34, NO. 2, MARCH 1988

On Classification with Empirically Observed Statistics and Universal Data Compression

JACOB ZIV, FELLOW, IEEE

Michael Gutman
Aaron Wagner
Sean Meyn
Ashish Khisti
Vincent Tan
....

- ... and other prior work. But only for discrete distributions and fixed $\mathrm{TV}(\mathbb{P}_X, \mathbb{P}_Y) \asymp 1, \quad m, n \to \infty$

Statistical Problems

$$X \sim \mathbb{P}_{\mathsf{X}}^{\otimes n}$$

$$Y \sim \mathbb{P}_{\mathsf{Y}}^{\otimes n}$$

$$Z \sim \mathbb{P}_{\mathsf{Z}}^{\otimes m}$$

unknown, $\mathbb{P}_0$ known, all in $\mathscr{P}$

The classes $\mathscr{P}$

Choices of $\mathscr{P}$ we considered:

- $\mathscr{P}_{\mathsf{H}}(\beta, d) = \{\beta$-Hölder densities over $[0,1]^d$ with $\|\cdot\|_{\mathscr{C}^\beta} \le C_{\mathsf{H}}\}$

  aka β times differentiable densities.

- $\mathscr{P}_{\mathsf{G}}(s) = \{\bigotimes_{j=}^{\infty}$ … $\theta_j^2 j^{2s} \le C_{\mathsf{G}}\}$

This talk: focus on $\mathscr{P}_H$ (smooth densities)

- $\mathscr{P}_{\mathsf{Db}}(k) = \{$dis … $\le C_{\mathsf{Db}}/k\}$
- $\mathscr{P}(k) = \{$all discrete distributions on $[k]\}$
- arbitrary densities on $[0,1]^d$ (with MMD separation instead of TV)

Rates vs sample complexity

Famous results for $\mathscr{P}_{\mathrm{H}}(\beta, d)$

	Rate	Sample complexity
Goodness-of-fit	$n^{-\frac{\beta}{2\beta + d/2}}$	$\epsilon^{-\frac{2\beta + d/2}{\beta}} = n_{\mathrm{GoF}}$
Estimation	$n^{-\frac{\beta}{2\beta + d}}$	$\epsilon^{-\frac{2\beta + d}{\beta}} = n_{\mathrm{Est}}$

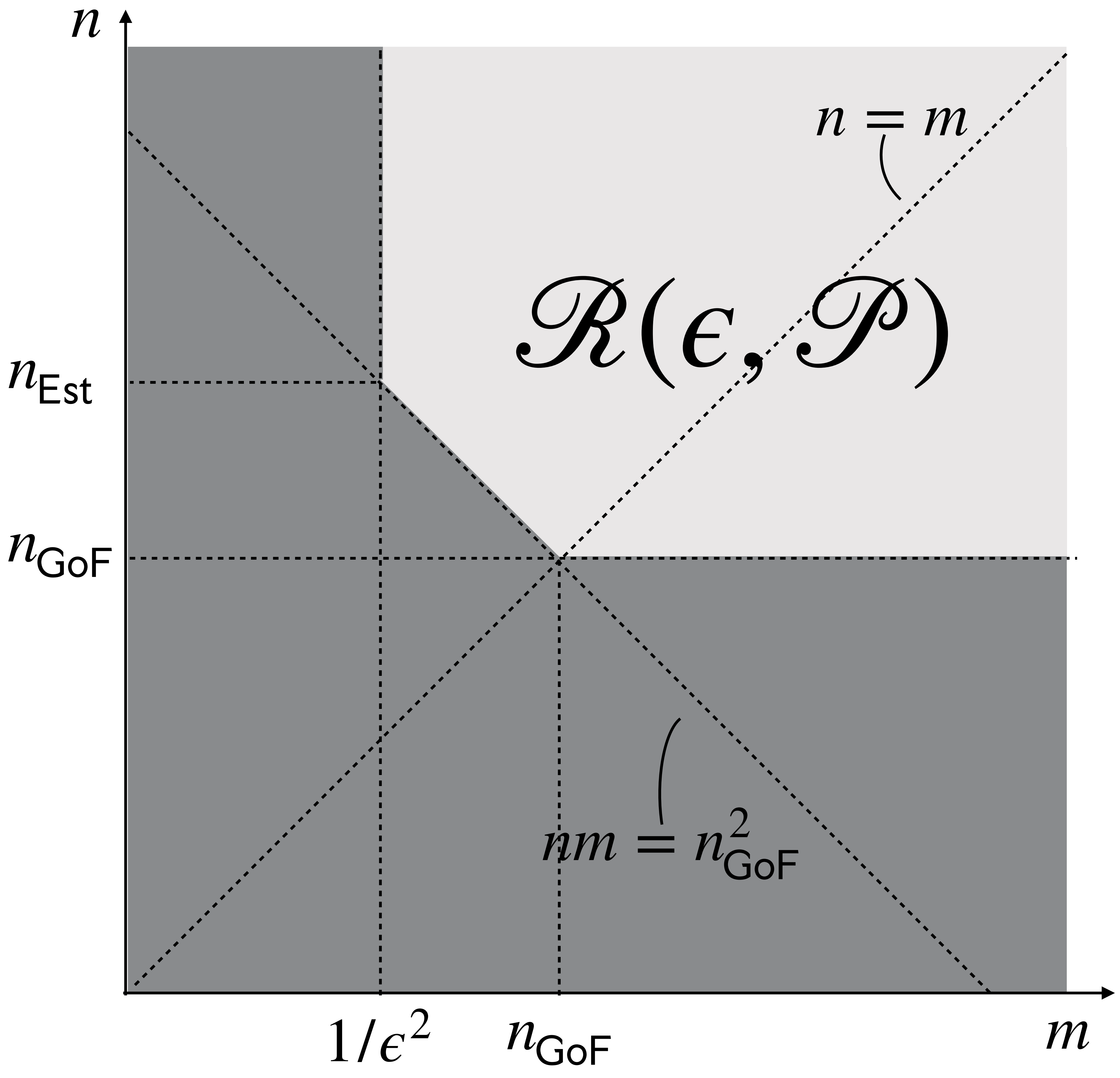
Results for $\mathscr{P}_{\mathsf{H}}$, $\mathscr{P}_{\mathsf{G}}$ and $\mathscr{P}_{\mathsf{Db}}$

Theorem (Gerber-P.'2022)

Up to constant factors:

$$\mathscr{R}(\epsilon, \mathscr{P}) \asymp \left\{ \begin{array}{c} m \geq 1/\epsilon^2 \ \text{ and } \ n \geq n_{\mathsf{GoF}} \\ \text{and} \quad n \cdot m \geq n_{\mathsf{GoF}}^2 \end{array} \right\}$$

Interpreting the results



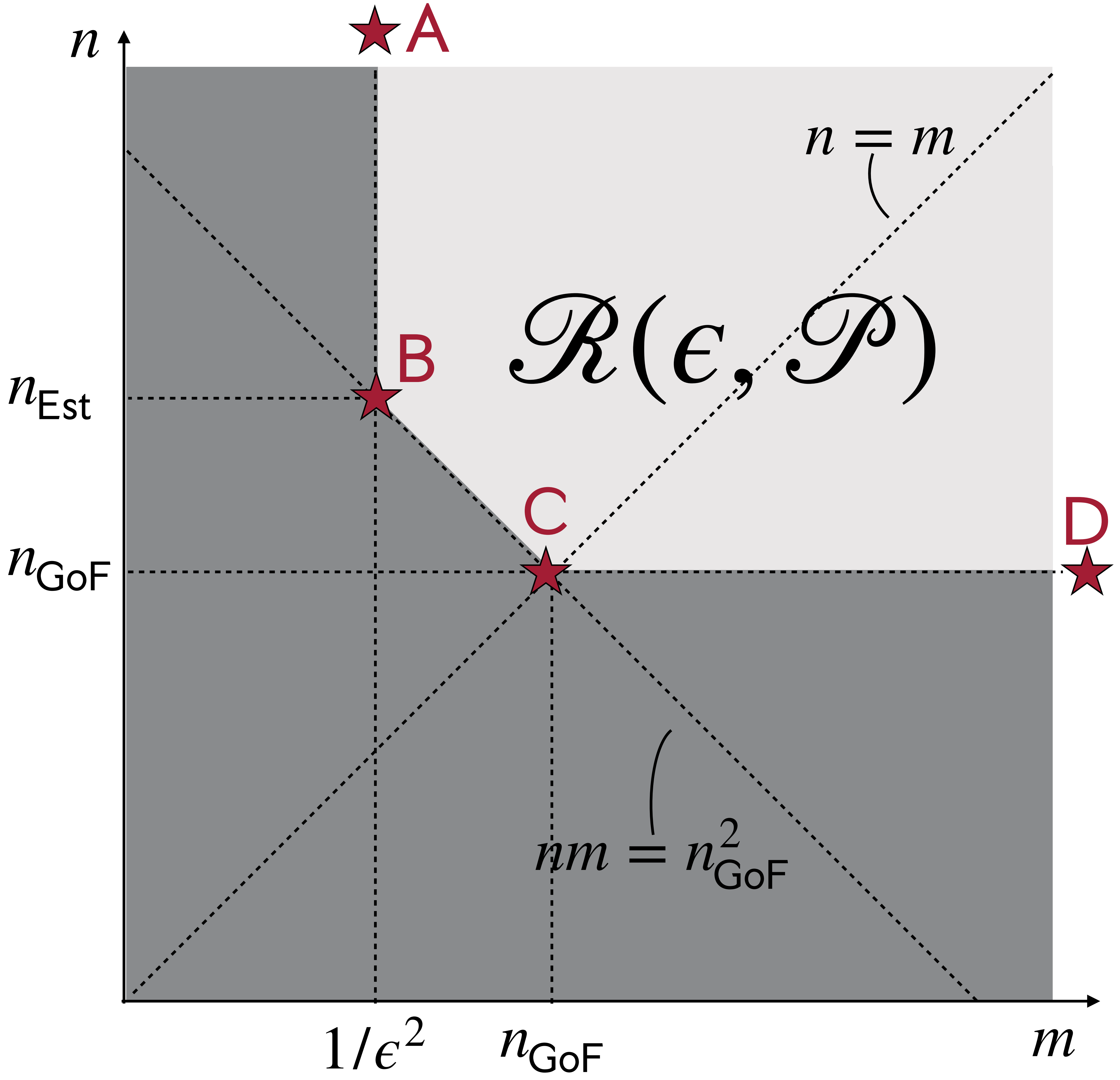
$$\mathscr{R}(\epsilon, \mathscr{P}) \asymp \left\{ \begin{array}{c} m \geq 1/\epsilon^2 \ \text{ and } \ n \geq n_{\text{GoF}} \\ \text{and} \quad n \cdot m \geq n_{\text{GoF}}^2 \end{array} \right\}$$

	n_{GoF}	n_{Est}
$\mathscr{P}_{\text{H}}(\beta, d)$	$\epsilon^{-\frac{2\beta + d/2}{\beta}}$	$\epsilon^{-\frac{2\beta + d}{\beta}}$

Target: minimal m (as in Higgs)

$$n_{\text{Est}} = n_{\text{GoF}}^2 \, \epsilon^2$$

Interpreting the results



$$\mathscr{R}(\epsilon, \mathscr{P}) \asymp \left\{ \begin{matrix} m \geq 1/\epsilon^2 \text{ and } n \geq n_{\mathsf{GoF}} \\ \text{and } n \cdot m \geq n_{\mathsf{GoF}}^2 \end{matrix} \right\}$$

Point	Algorithm	Lower bd
A $\leftrightarrow (1/\epsilon^2, \infty)$	Binary HT	Trivial
B $\leftrightarrow (1/\epsilon^2, n_{\mathsf{Est}})$	Est + robust HT	New
C $\leftrightarrow (n_{\mathsf{TS}}, n_{\mathsf{TS}})$	Two-sample*	Reduction to TS
D $\leftrightarrow (\infty, n_{\mathsf{GoF}})$	Goodness-of-fit	New but easy

$\boxed{\text{Can estimate } \mathbb{P}_{\mathsf{X}} \text{ and } \mathbb{P}_{\mathsf{Y}}}$

*$n_{\mathsf{GoF}} = n_{\mathsf{TS}}$ for each of these classes

The test statistic

- Based on Ingster's L^2-comparison idea
- Discretize $[0,1]^d$ cube into $k = \epsilon^{-\frac{d}{\beta}}$ bins
- Empirical pmfs $\hat{p}_{\mathsf{X}}, \hat{p}_{\mathsf{Y}}, \hat{p}_{\mathsf{Z}}$ based on (n, n, m) observations
- **Theorem:** All points on the optimal tradeoff are achieved by

$$T = \|\hat{p}_{\mathsf{X}} - \hat{p}_{\mathsf{Z}}\|_2^2 - \|\hat{p}_{\mathsf{Y}} - \hat{p}_{\mathsf{Z}}\|_2^2$$

$$\Psi = \mathbb{I}\{T \geq 0\}$$

Note: no training, distance estimates are both wrong.

Enter Machine Learning: Practical tests

Kernel-based L_2 test (MMD)

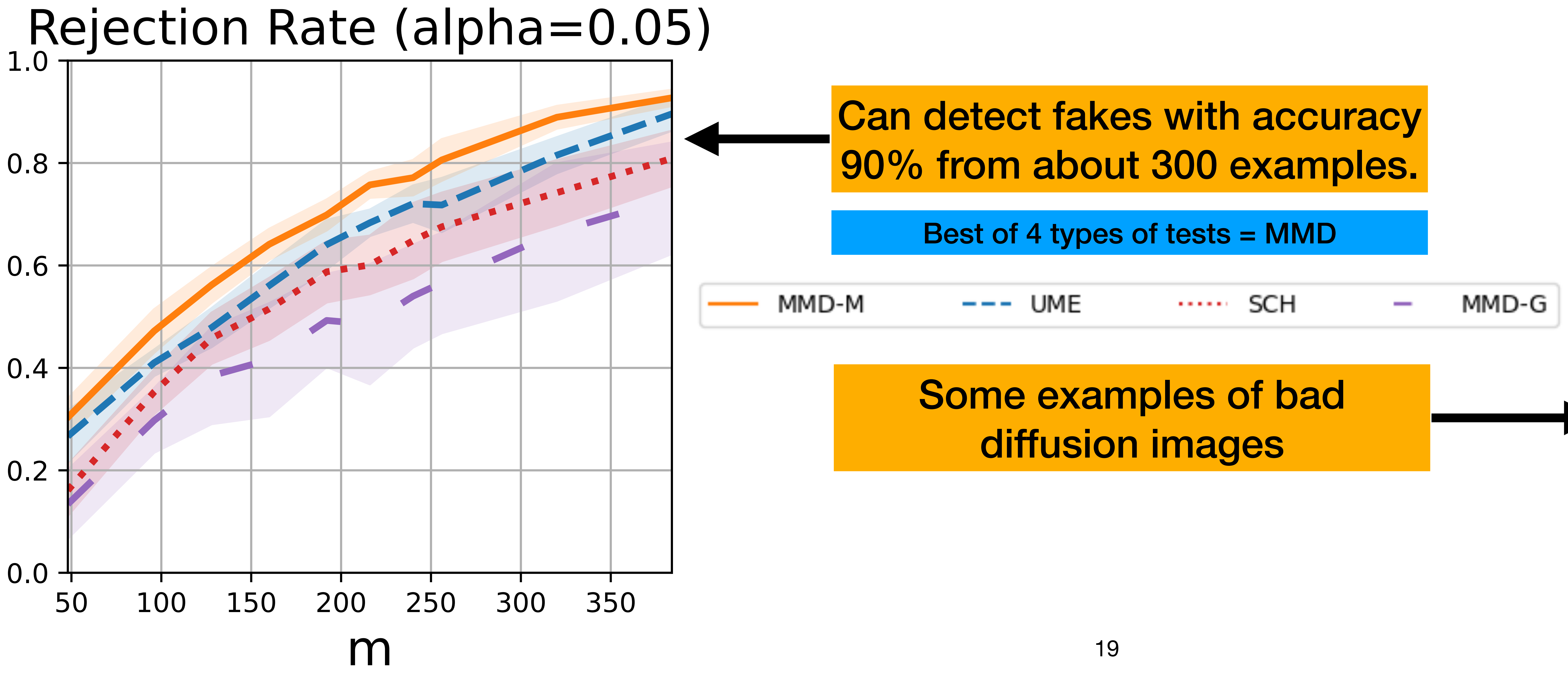
- Real-world distributions are high-dimensional $\implies$ discretization impractical.
- Given $\hat{\mathbb{P}}_X$, $\hat{\mathbb{P}}_Z$ measure distance after applying **feature map** ϕ:
  $\mathrm{MMD}^2(\hat{\mathbb{P}}_X, \hat{\mathbb{P}}_Z) = \|\hat{\mathbb{E}}\phi(X) - \hat{\mathbb{E}}\phi(Z)\|_2^2$
  (proposed for two-sample testing [Sutherland et al, ICLR'17])
- We adopt this to LFHT via the same mechanism:
  $T(X, Y, Z) = \|\hat{\mathbb{E}}\phi(X) - \hat{\mathbb{E}}\phi(Z)\|_2^2 - \|\hat{\mathbb{E}}\phi(Y) - \hat{\mathbb{E}}\phi(Z)\|_2^2 \quad \gtrless 0$
- Has the same LFHT region wrt $\mathrm{MMD}(P_X, P_Y) \geq \epsilon$ [Gerber, Jiang, Sun, P., NeurIPS'23]
- **Train feature map** to maximize $\dfrac{\mathbb{E}[T \mid H_0]}{\sqrt{\mathrm{Var}[T \mid H_0]}}$ ratio (gradient descent in kernel space)

LFHT for CIFAR [NeurIPS'23]

- So our test:

$$T(X,Y,Z) = \|\hat{\mathbb{E}}\phi(X) - \hat{\mathbb{E}}\phi(Z)\|_2^2 - \|\hat{\mathbb{E}}\phi(Y) - \hat{\mathbb{E}}\phi(Z)\|_2^2 \quad \gtrless 0$$

- Here is an example: **X=CIFAR10 vs Y=1/3 CIFAR + 2/3 Diffusion Model (DDPN)**
- $(n \approx 10^5, m \approx 10^1)$



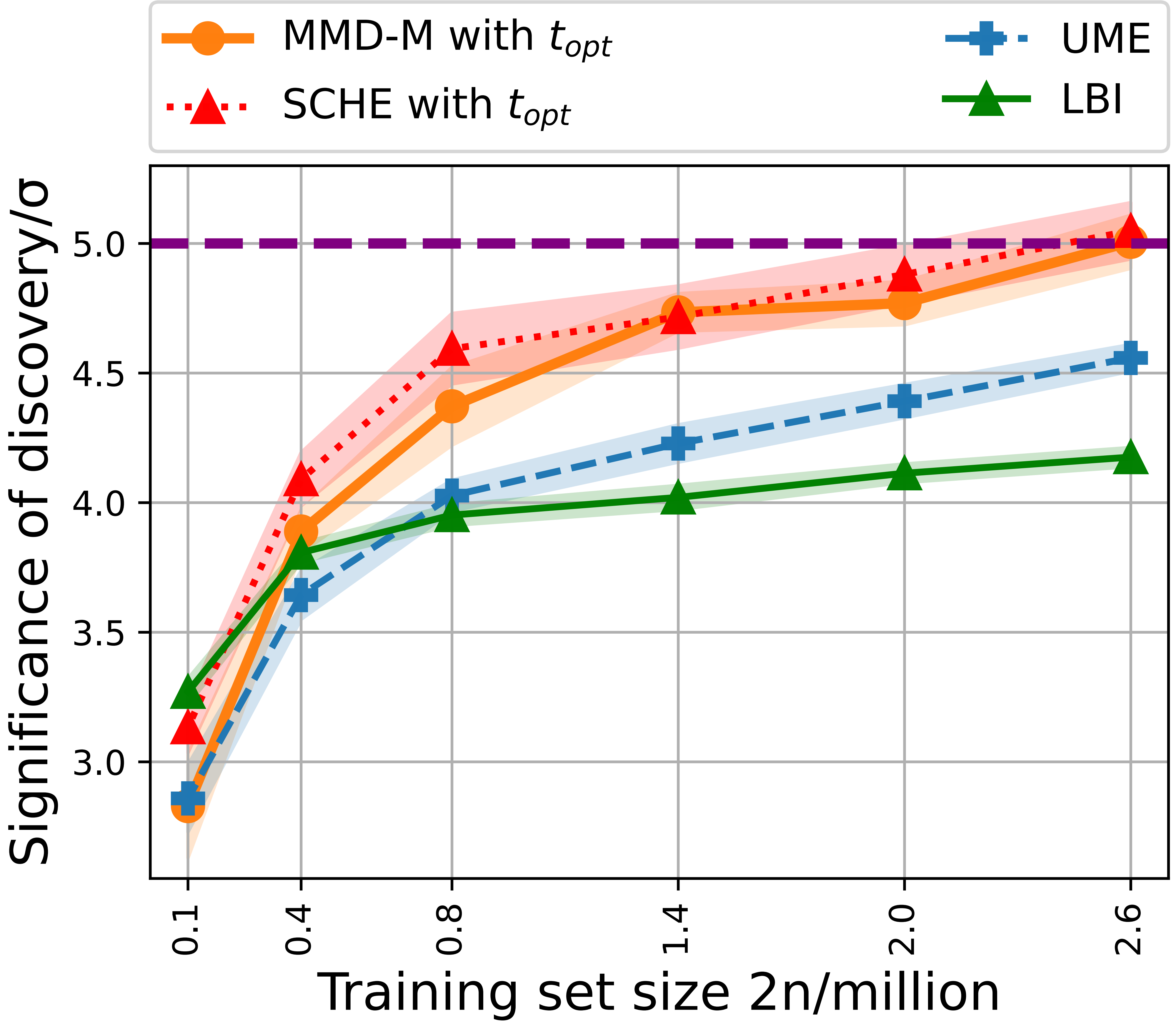
Can detect fakes with accuracy 90% from about 300 examples.

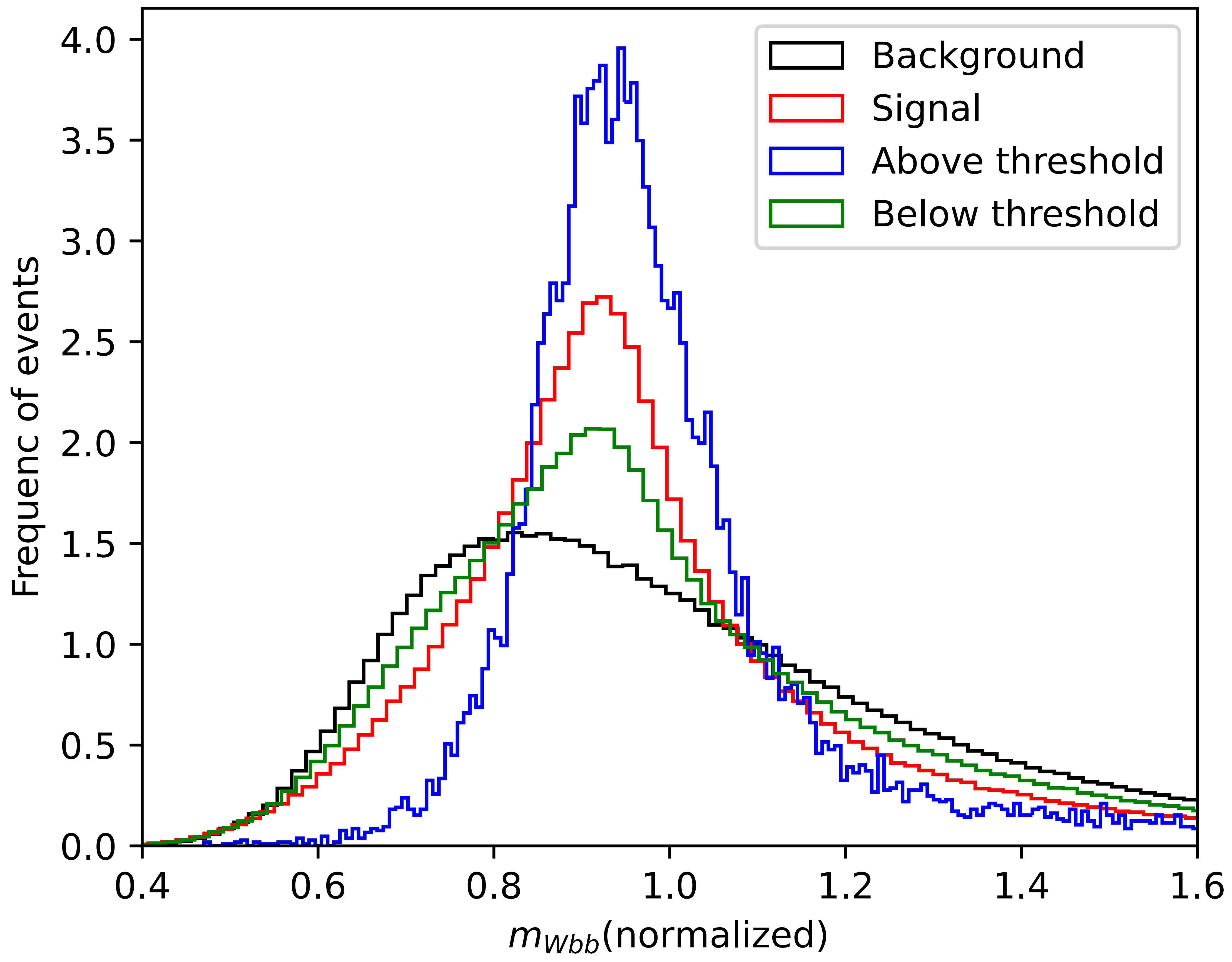
Best of 4 types of tests = MMD

Some examples of bad diffusion images

Back to Higgs [NeurIPS'23]

- Instead of fixed two-sided error physicists use significance of discovery
- Expressed in σ's. For the *new particle* need 5σ. Our road to 5σ...





Interference rejection

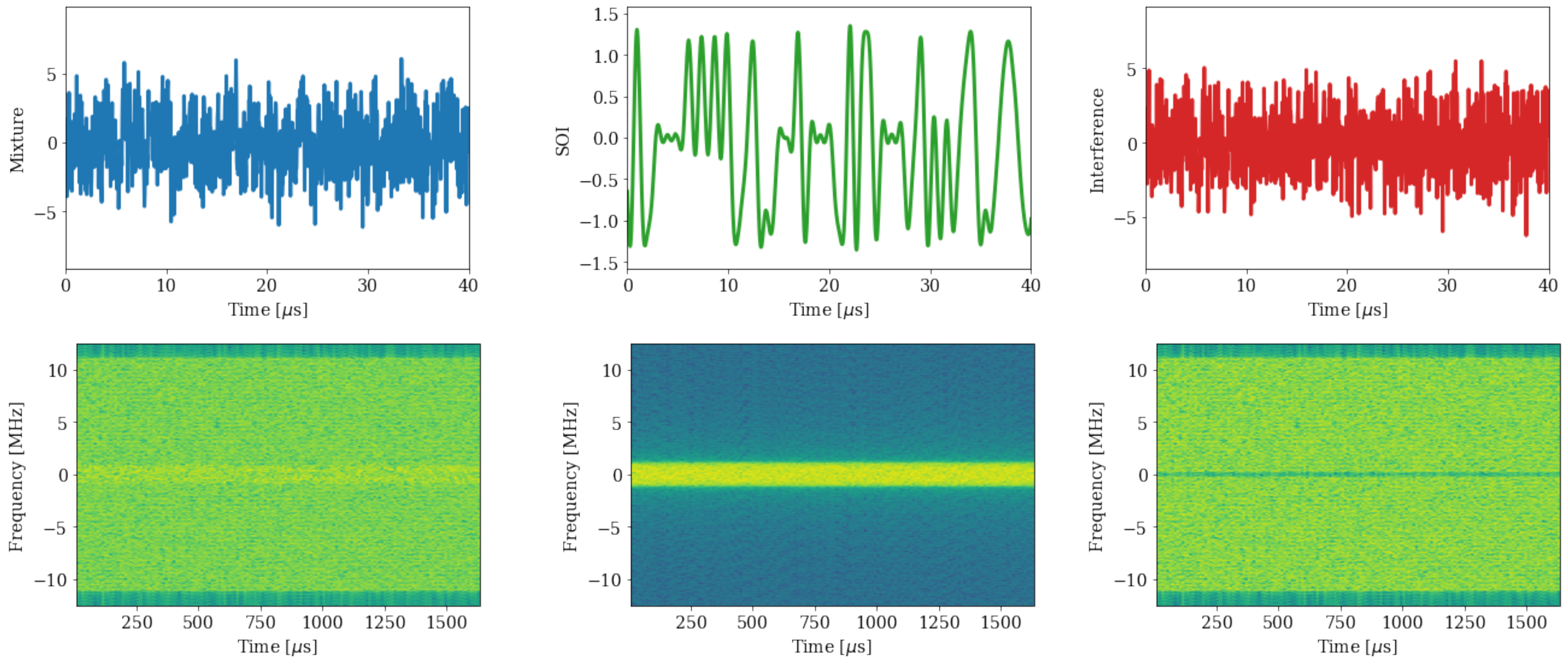
Demodulation task in communication

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{b}$$

Received signal

Signal of Interest (SOI)
e.g. BPSK/QPSK

Noise and interference



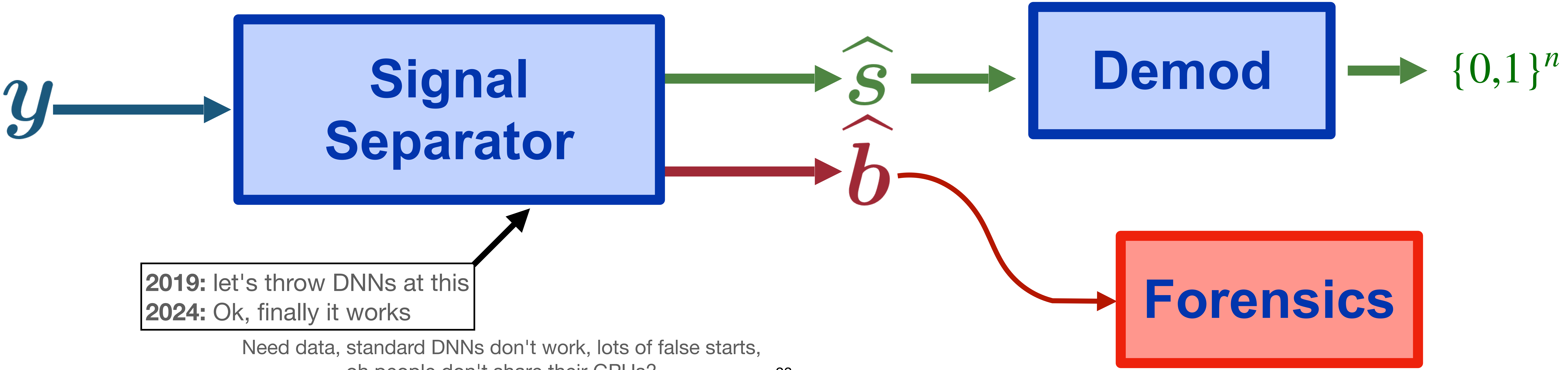
Example at -9 dB Signal-to-Interference Ratio (SIR)

OFDM interference is marginally Gaussian => need to exploit time-frequency structure of the interference. How?

Idea: Use signal (source) separation

$$y = s + b$$

Observed — *SOI* — *Interference*



Need data, standard DNNs don't work, lots of false starts, oh people don't share their GPUs?

Two types of architectures

$$\underbrace{y}_{\text{Observed}} = \underbrace{s}_{\text{SOI}} + \underbrace{b}_{\text{Interference}}$$

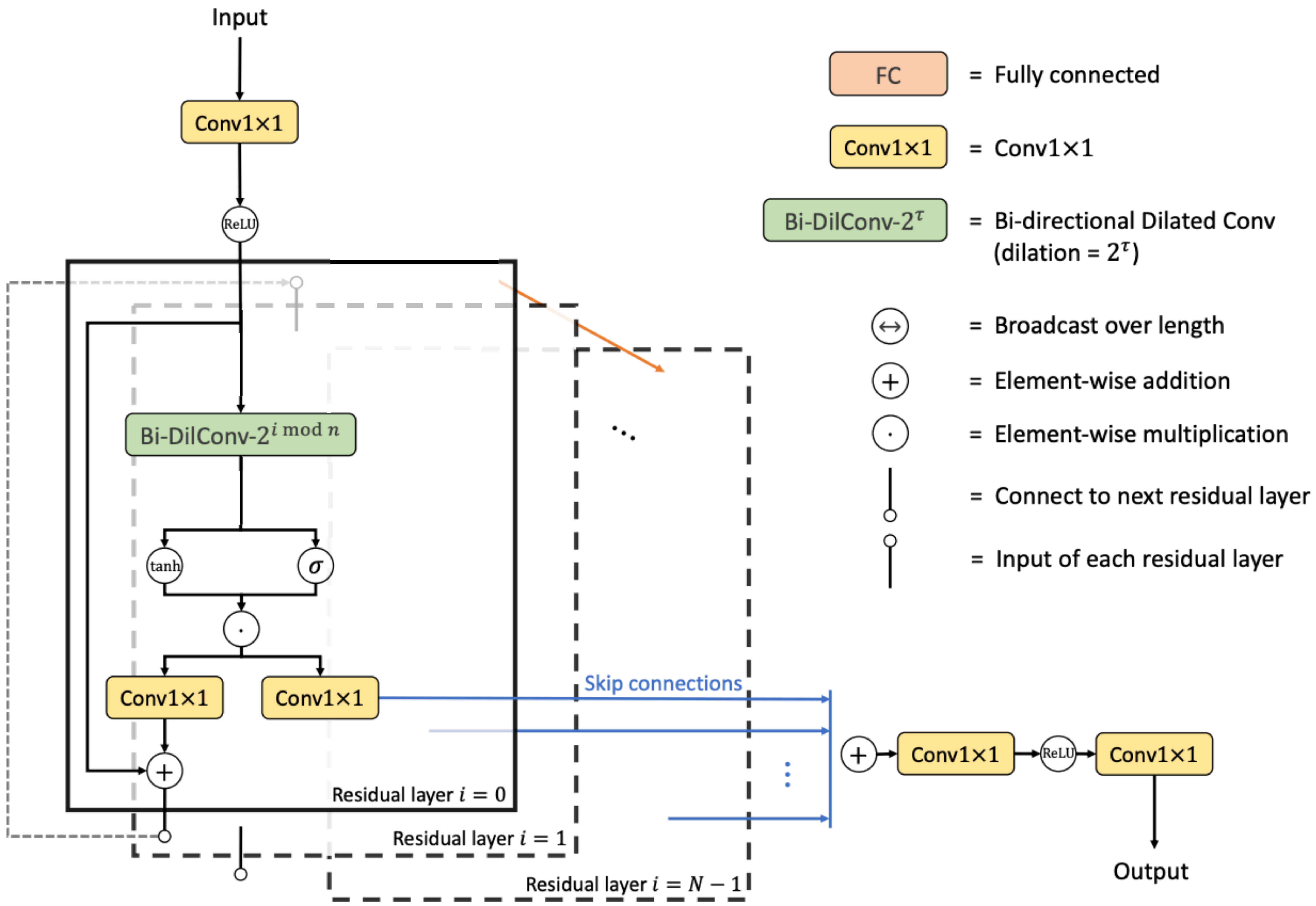
Supervised (end-to-end)

- Create many synthetic mixtures $s + b$
- Feed pairs (y, s) to DNN
- Force it to learn to recover s from y
- **Pros:** best performance
- **Cons:** need to retrain DNN for each signal-interference pair

Bayesian MAP

- Collect many samples of b
- Train a diffusion model to learn P_b
- Use MAP to recover s from y
- **Pros:** one model works for all SOI
- **Cons:** slow inference, performance

NeurIPS'2023: WaveNet (dilated CNN)



	Description
Number of layers	30 residual layers with dilation cycle of {1, 2, … 512} repeated three times
Total number of parameters	4M
GPU compute (training)	8 GPU days

Additional training tricks:

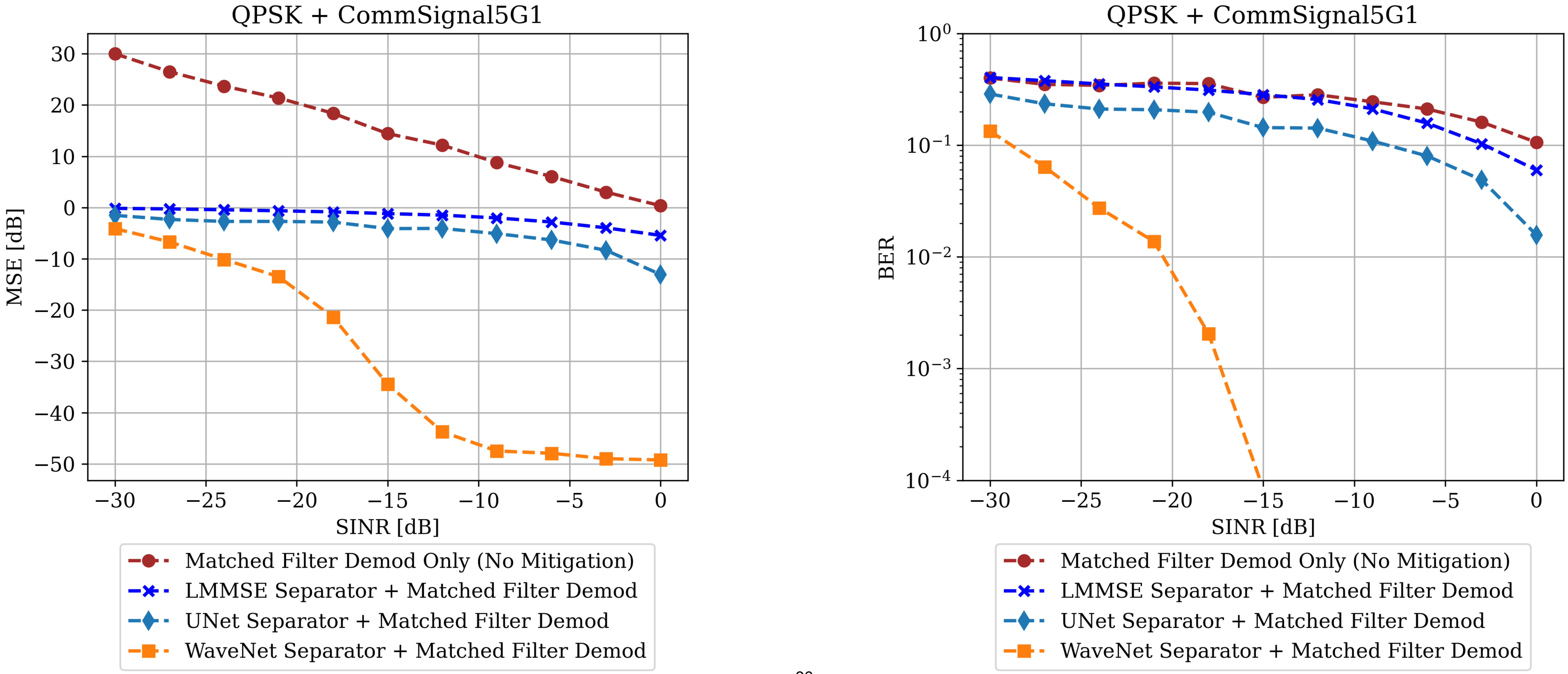
<u>Adaptive learning rate scheduler</u> based on validation loss

<u>Mixed precision training</u> with fp16 to speed up inference

T. Jayashankar, G. C. F. Lee, A. Lancho, A. Weiss, Y. Polyanskiy, and G. Wornell, "Score-based source separation with applications to digital communication signals," 36th for Advances in Neural Information Processing Systems, 2023.

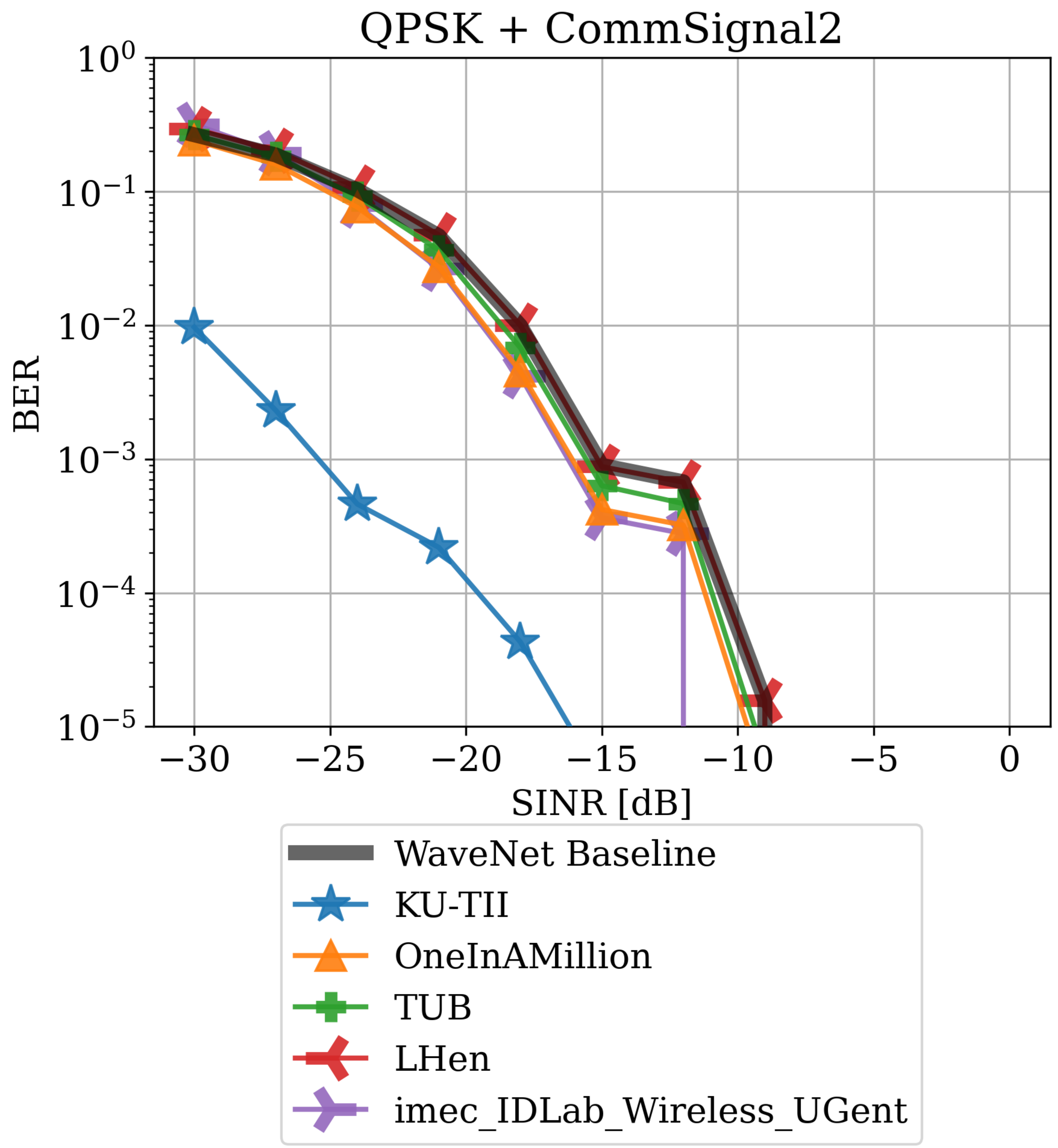
So does it work?

QPSK vs OFDM (5GNR) example



ICASSP'2024: Session on RF Challenge

Can we obtain further gains from other novel architectures?



Technology Innovation Institute

Learnable dilations and new data augmentation schemes

Number of parameters: **16M**
Number of GPUs: 4 x RTX 3090
GPU Compute: **13 GPU days**

Fraunhofer HHI

Attention-based UNet and fine-tuning of our WaveNet baseline

Number of parameters: **350M**
Number of GPUs: 4 x A100
GPU Compute: **8 GPU days**

TECHNISCHE UNIVERSITÄT BERLIN

New UNet architecture with bi-directional LSTM bottleneck layer

Number of parameters: **60M**
Number of GPUs: 1 x RTX 6000
GPU Compute: **4 GPU days**

Two types of architectures

$$y = s + b$$

Observed *SOI* *Interference*

Supervised (end-to-end)

- Create many synthetic mixtures $s + b$
- Feed pairs (y, s) to DNN
- Force it to learn to recover s from y
- **Pros:** best performance
- **Cons:** need to retrain DNN for each signal-interference pair

Bayesian MAP

- Collect many samples of b
- Train a **diffusion model** to learn P_b
- Use MAP to recover s from y
- **Pros:** one model works for all SOI
- **Cons:** slow inference, performance

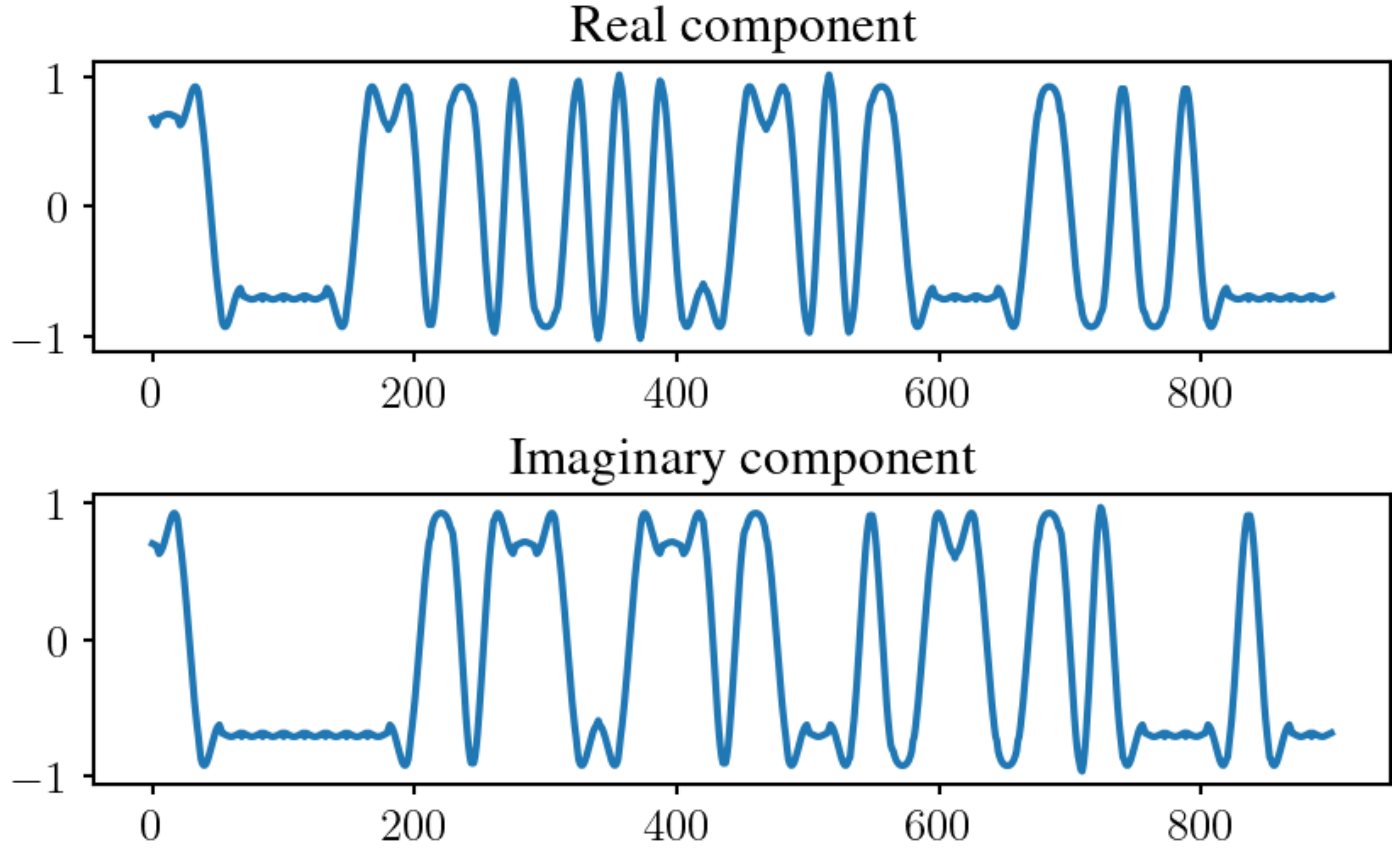
Diffusion models

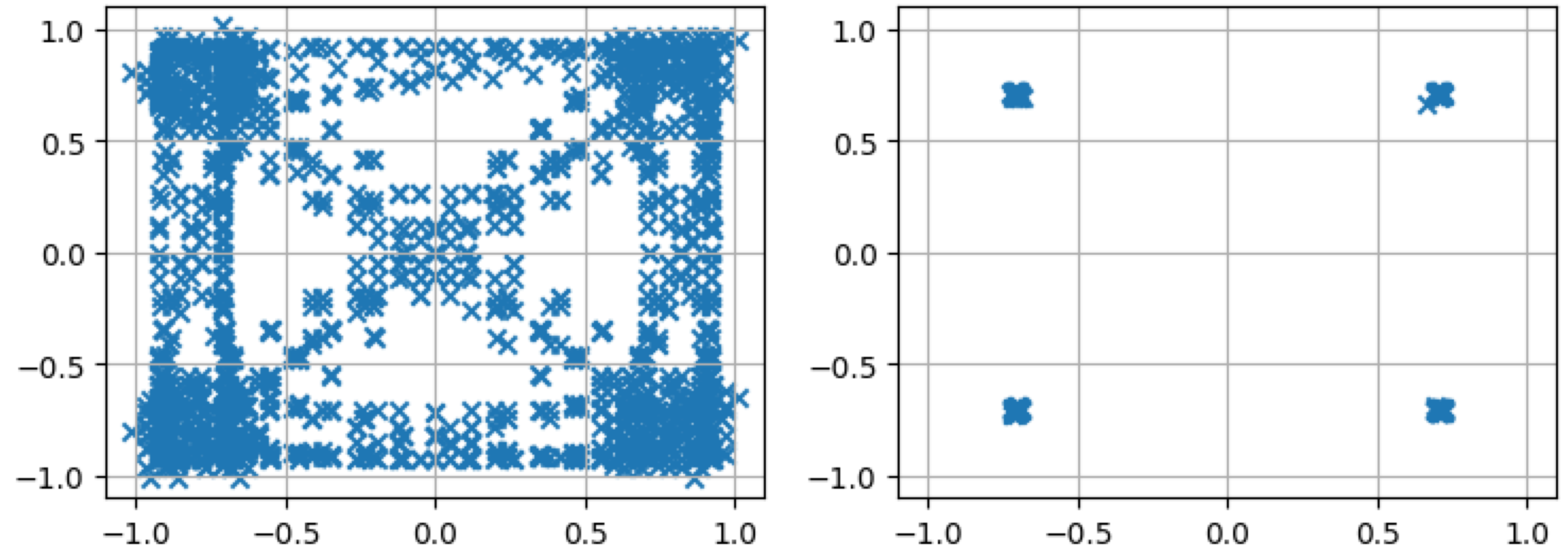
Images and RF

SOTA ***generative model*** that can learn complex structures from signal datasets

Can diffusion models capture the underlying discrete statistical structures of RF signals?

Sample from diffusion model trained on QPSK signals





Constellation of generated sample

After matched filtering

Score-based Source separation (α-RGS)

$\mathbf{s} \in \mathcal{S} \subset \mathbb{C}^D, \mathbf{b} \in \mathbb{C}^D$ statistically independent sources

Combinatorially hard

Non-differentiable

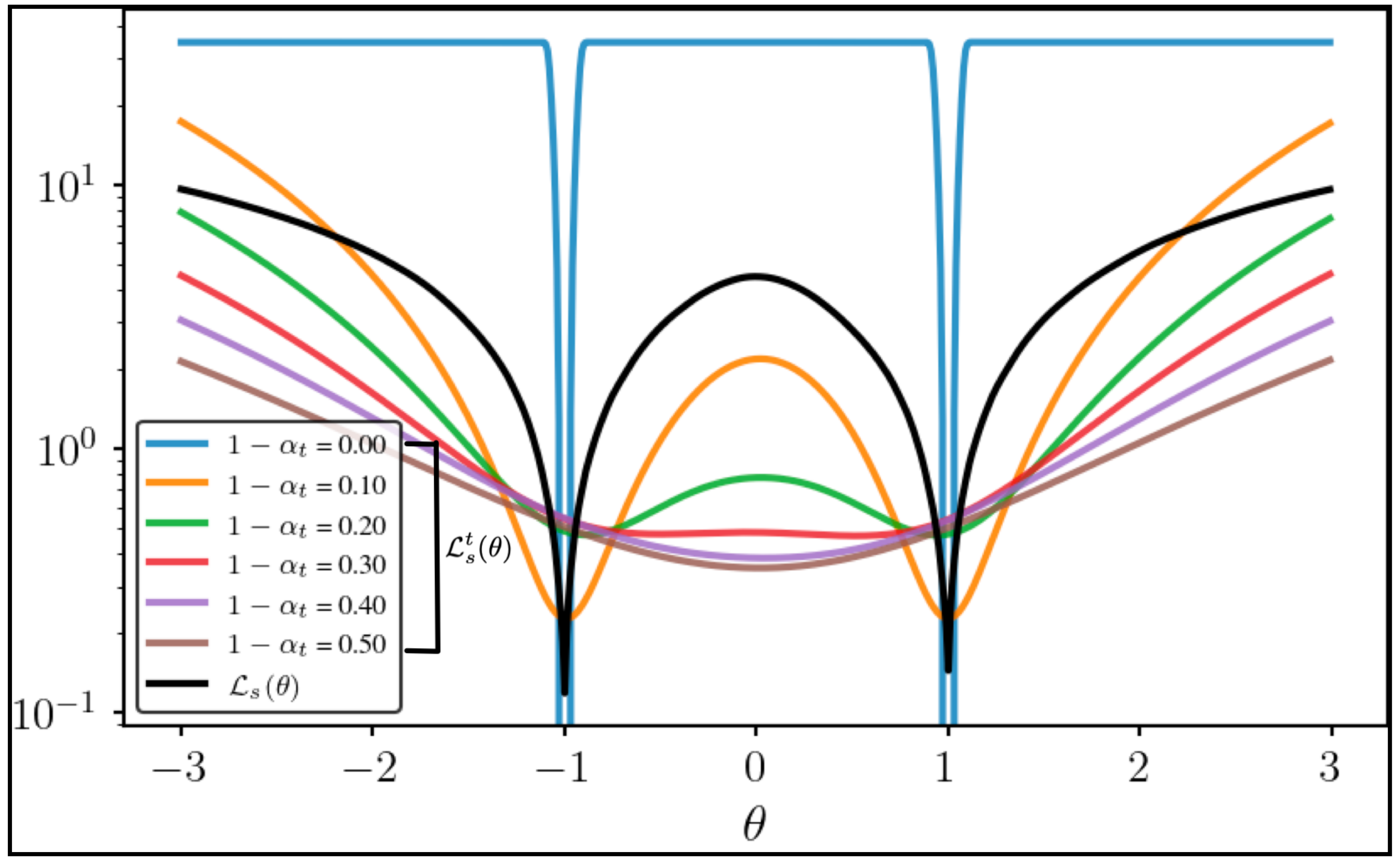
MAP Estimation Given $\mathbf{y} = \mathbf{s} + \mathbf{b}$

$$\hat{\mathbf{s}} = \underset{\mathbf{s}\in\mathcal{S}:\mathbf{y}=\mathbf{s}+\mathbf{b}}{\arg\max}\ p_{\mathsf{s}|\mathsf{y}}(\mathbf{s}|\mathbf{y}) = \underset{\mathbf{s}\in\mathcal{S}}{\arg\min} - \log P_{\mathsf{s}}(\mathbf{s}) - \log p_{\mathsf{b}}(\mathbf{y} - \mathbf{s})$$

Gradient Descent Estimate $\bar{\mathbf{s}} = \mathbf{s} + \epsilon,\ \epsilon \to 0$

$$\mathbf{s}_{i+1} \leftarrow \mathbf{s}_i + \underbrace{\nabla \log p_{\mathsf{s}}(\mathbf{s}_i)}_{\text{Score}} - \nabla \log p_{\mathsf{b}}(\mathbf{y} - \mathbf{s}_i)$$

Smoothed optimization landscape



<u>**Randomized Gaussian Smoothing with an α-posterior (α-RGS)**</u>

Diffusion Models Model **unknown priors (score functions)** over $\mathbf{s}$ and

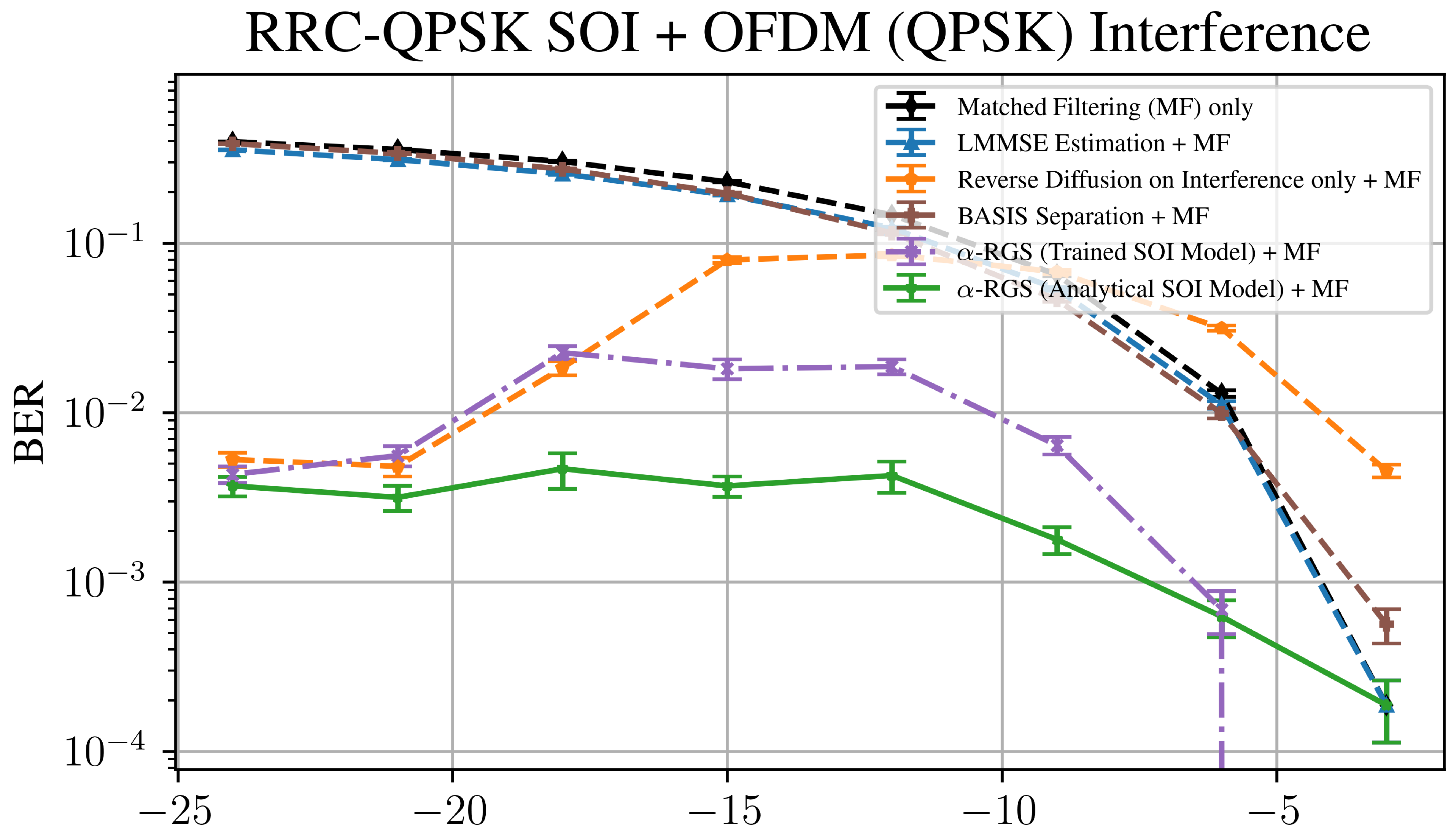
Gaussian Smoothing Use noise variance levels α_t and α_u

α-posterior Reweight likelihood with weight $\alpha = \omega$

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t,\mathbf{z}_s}\left[\log p_{\tilde{\mathsf{s}}_t}\left(\tilde{\mathbf{s}}_t\left(\theta\right)\right)\right] - \omega\mathbb{E}_{u,\mathbf{z}_b}\left[\log p_{\tilde{\mathsf{b}}_u}\left(\tilde{\mathbf{b}}_u\left(\theta, \mathbf{y}\right)\right)\right]$$

Results: improving SOTA

Other algos based on approximating MAP via score-learning



Averaging over regularization + α-posterior give us an edge

Conclusion

(i) We studied signal detection (hypothesis testing) when hypotheses are only specified through examples.

(ii) We saw minimax optimal bounds and practical algorithms

(iii) **Next :** Study notion of regret or instance-optimality.

(iv) **More generally:** Study parameter estimation, confidence intervals, channel coding, constellation design,...

Thank you!