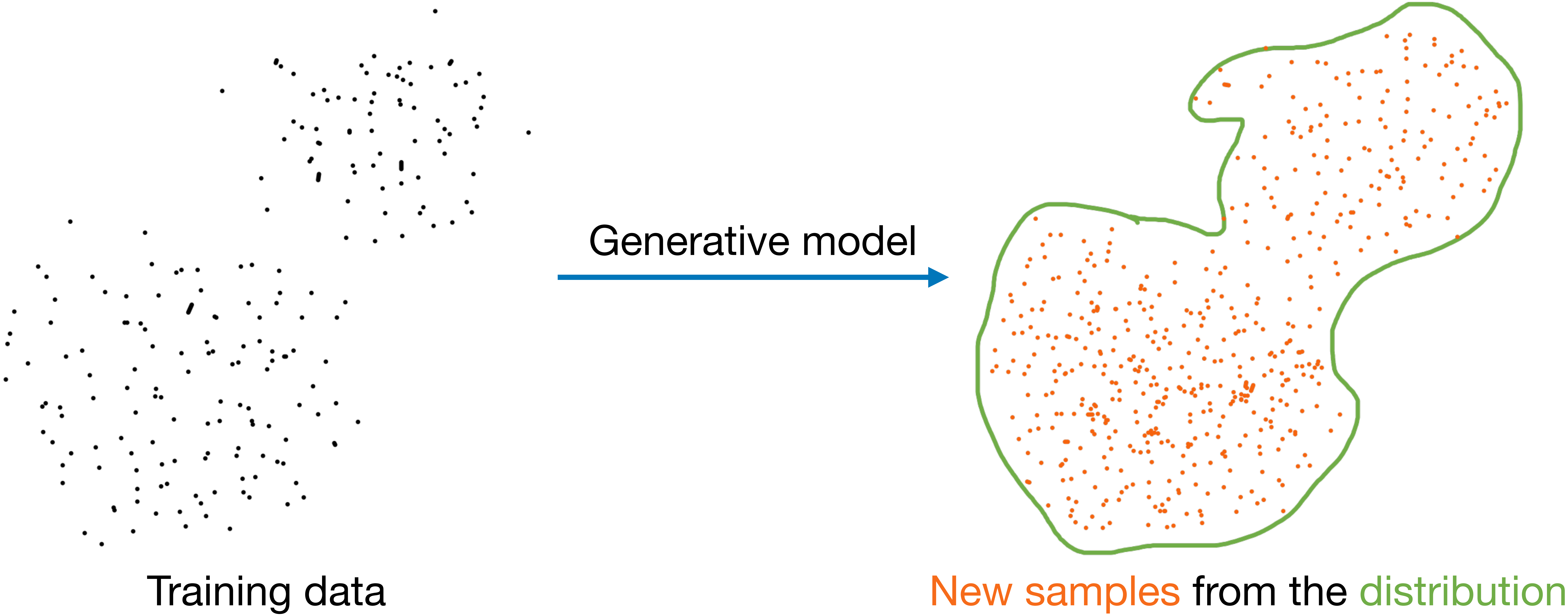
Evaluating Generative Models

Nicolas Salvy, 3rd year PhD student
under the supervision of Hugues Talbot and Bertrand Thirion

Monday 15th, December

Goal of Generative Models



Goal of Generative Models

Training data

Generative model

New samples from the distribution

How can we evaluate that?


Real **samples**

Unknown
distributions


Synthetic **samples**

How can we evaluate that? using distributions

Approximate the distributions

Real samples

Synthetic samples

Unknown **distributions**

Real Gaussian: $\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}})$

Fréchet distance

Synthetic Gaussian: $\mathcal{N}(\mu_{\text{syn}}, \Sigma_{\text{syn}})$

How can we evaluate that? using distributions

Real **samples**

Synthetic **samples**

Unknown
distributions

Approximate the distributions

Real Gaussian: $\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}})$

Fréchet
distance

Synthetic Gaussian: $\mathcal{N}(\mu_{\text{syn}}, \Sigma_{\text{syn}})$

How can we evaluate that? using distributions

Real **samples**

Synthetic **samples**

Encoder

"distances make sense"

Approximate the distributions

Real Gaussian: $\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}})$

Fréchet distance

Synthetic Gaussian: $\mathcal{N}(\mu_{\text{syn}}, \Sigma_{\text{syn}})$

Fréchet Inception Distance (FID) *(Heusel et al. 2017)*

Real **samples**

Synthetic **samples**

Encoder
Inception
v3

”distances make sense”

Approximate the distributions

Real Gaussian: $\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}})$

Fréchet distance

Synthetic Gaussian: $\mathcal{N}(\mu_{\text{syn}}, \Sigma_{\text{syn}})$

FD DINOv2 *(Stein et al. 2023)*

Real **samples**

Synthetic **samples**

Encoder
DINOv2

"distances make sense" →

Approximate the distributions

Real Gaussian: $\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}})$

$\Updownarrow$ Fréchet distance

Synthetic Gaussian: $\mathcal{N}(\mu_{\text{syn}}, \Sigma_{\text{syn}})$

Feature Likelihood Divergence

(Jiraleerspong et al. 2023)

Approximate the distributions

Real **samples**

Synthetic **samples**

Encoder
DINOv2

"distances make sense"

→ Likelihood score using Mixtures of Gaussians

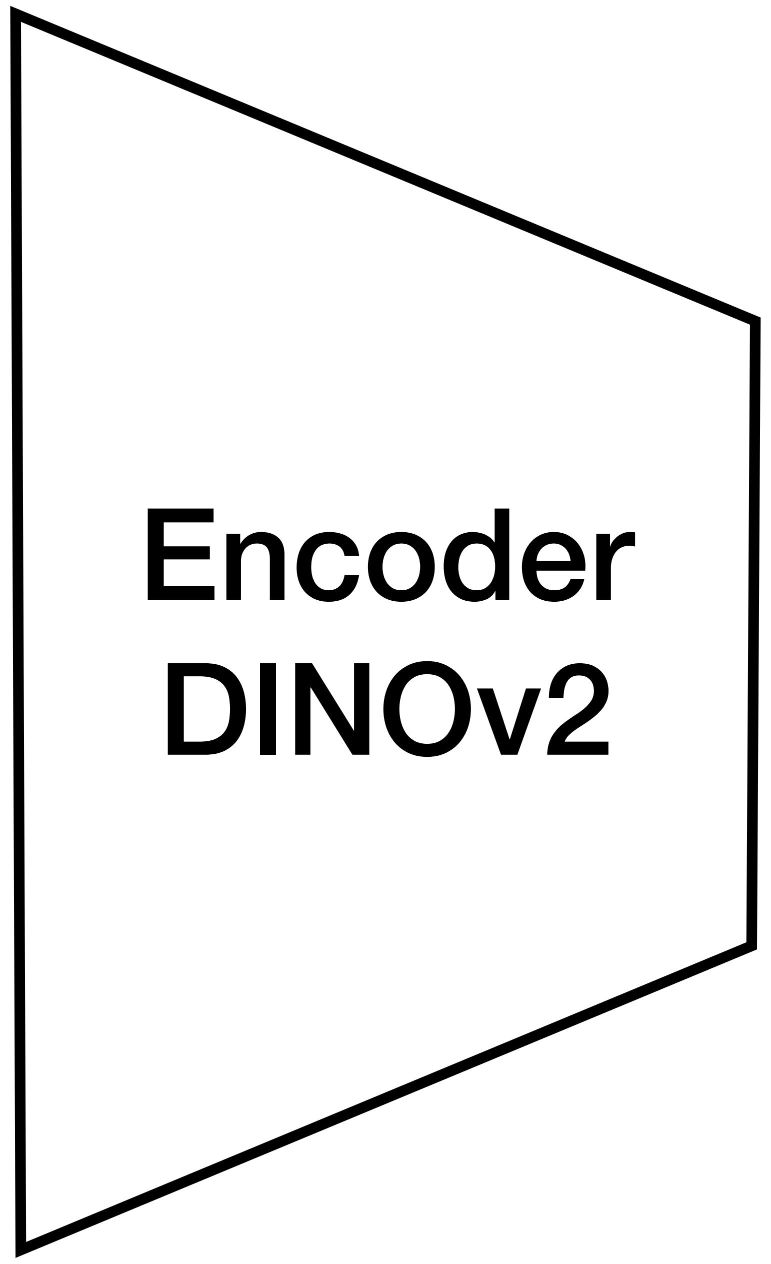
Feature Likelihood Divergence

(Jiralerspong et al. 2023)

Approximate the distributions

Real **samples**

Synthetic **samples**



"distances make sense"

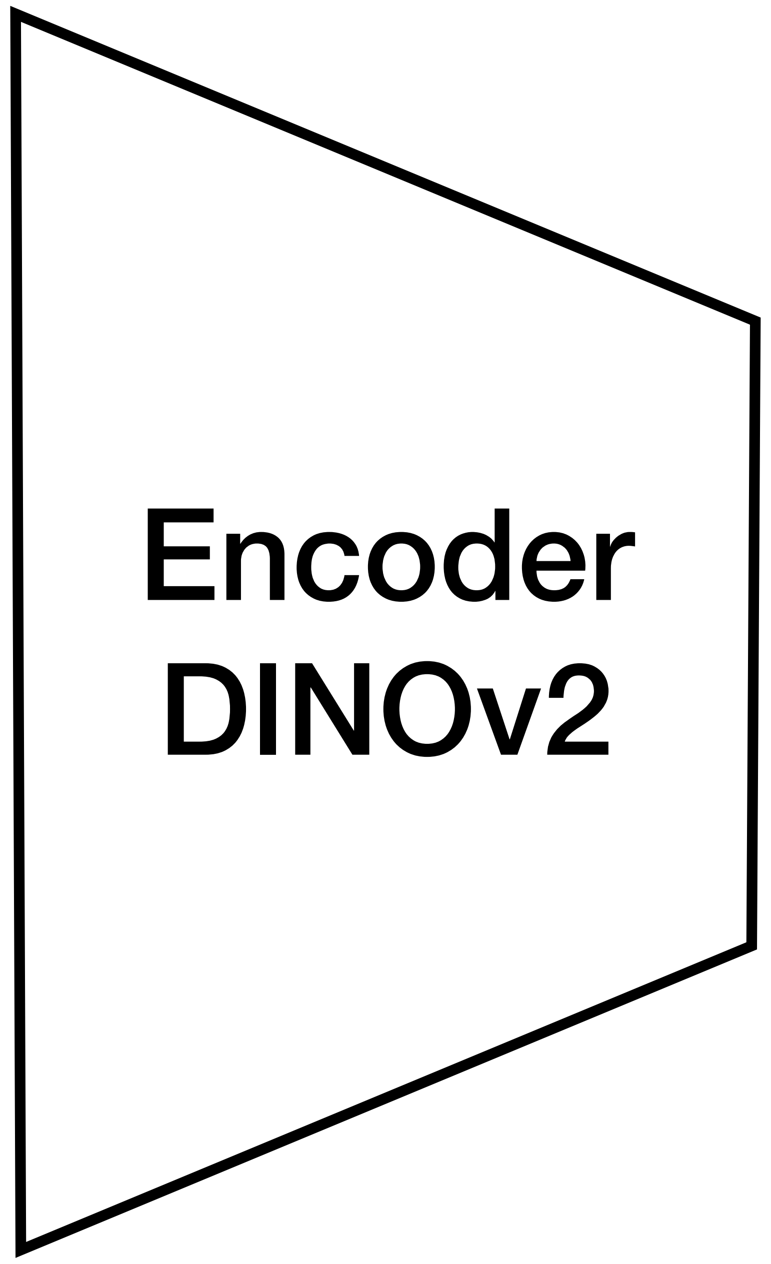
Likelihood score using Mixtures of Gaussians

- Simple approximations
- Yield one value ➡ Same value for a lack of **quality** and a lack of **diversity**

Disentangling quality & diversity

Real **samples**

Synthetic **samples**



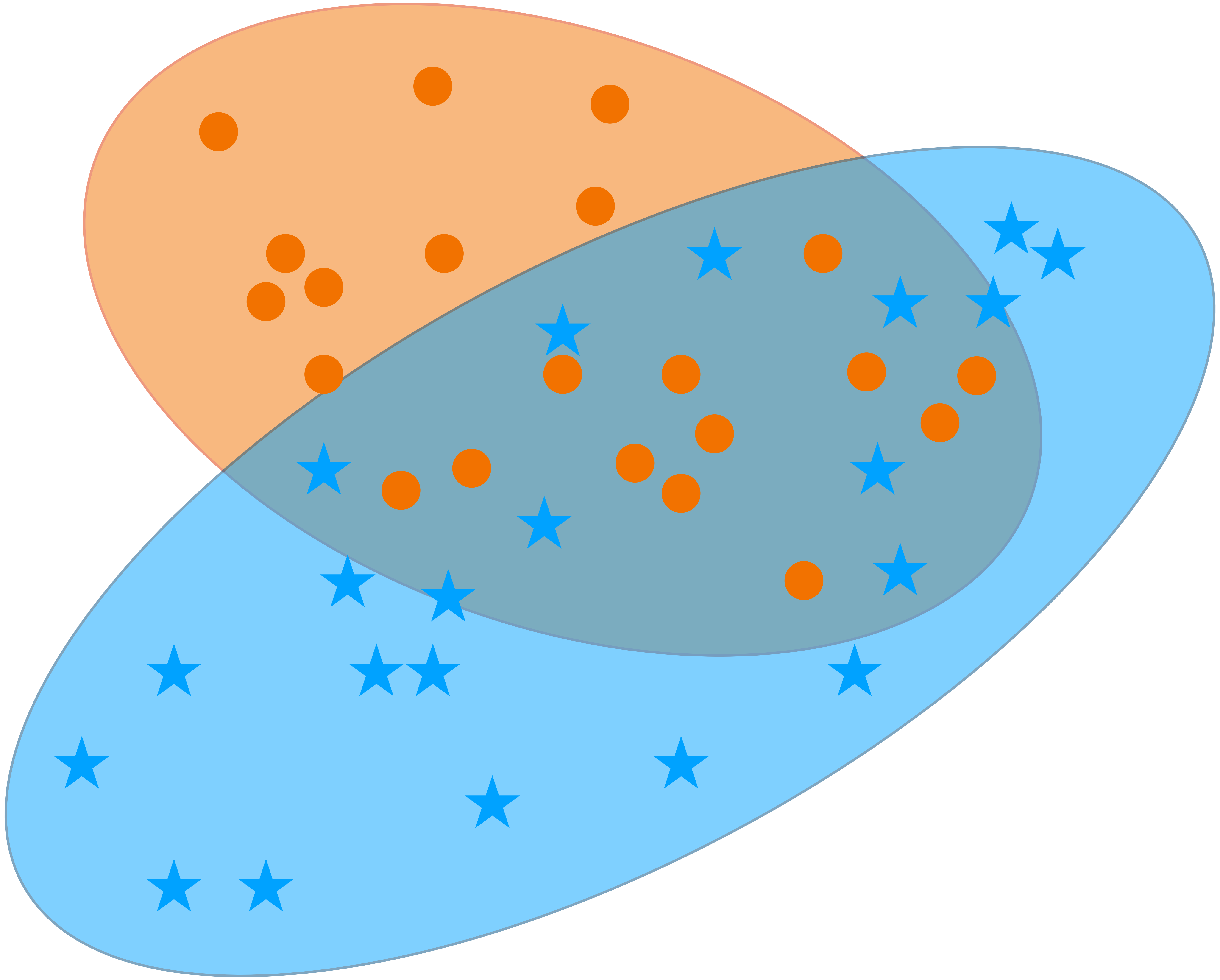
”distances make sense”

~~**Approximate** the distributions~~

~~Likelihood score using Mixtures of Gaussians~~

- Simple approximations
- Yield one value ➡ Same value for a lack of **quality** and a lack of **diversity**

Setup

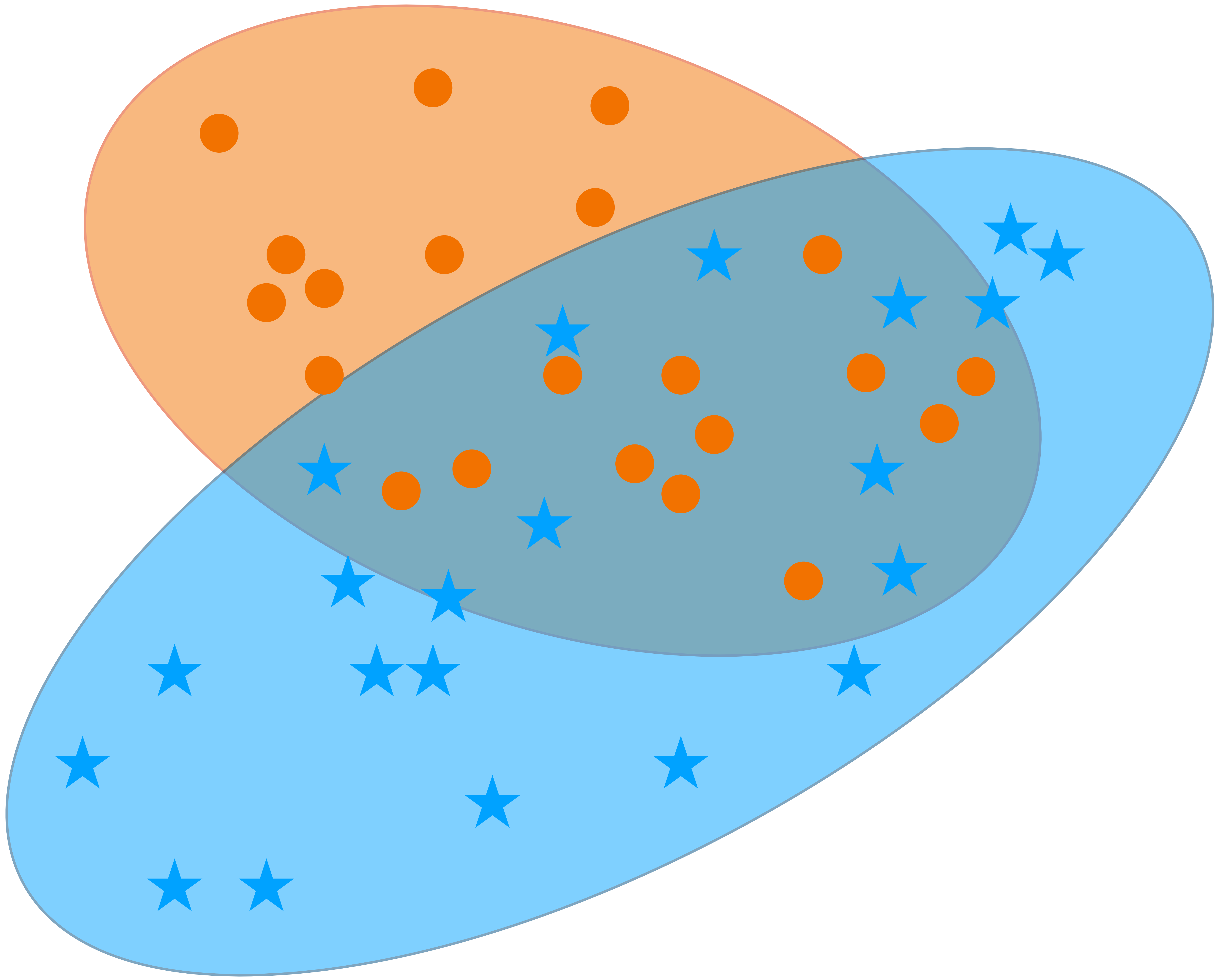


We have:

- real ★ and synthetic ● samples
- unknown distributions
- space where ”distances make sense”

Goal: disentangle quality and diversity

Goal



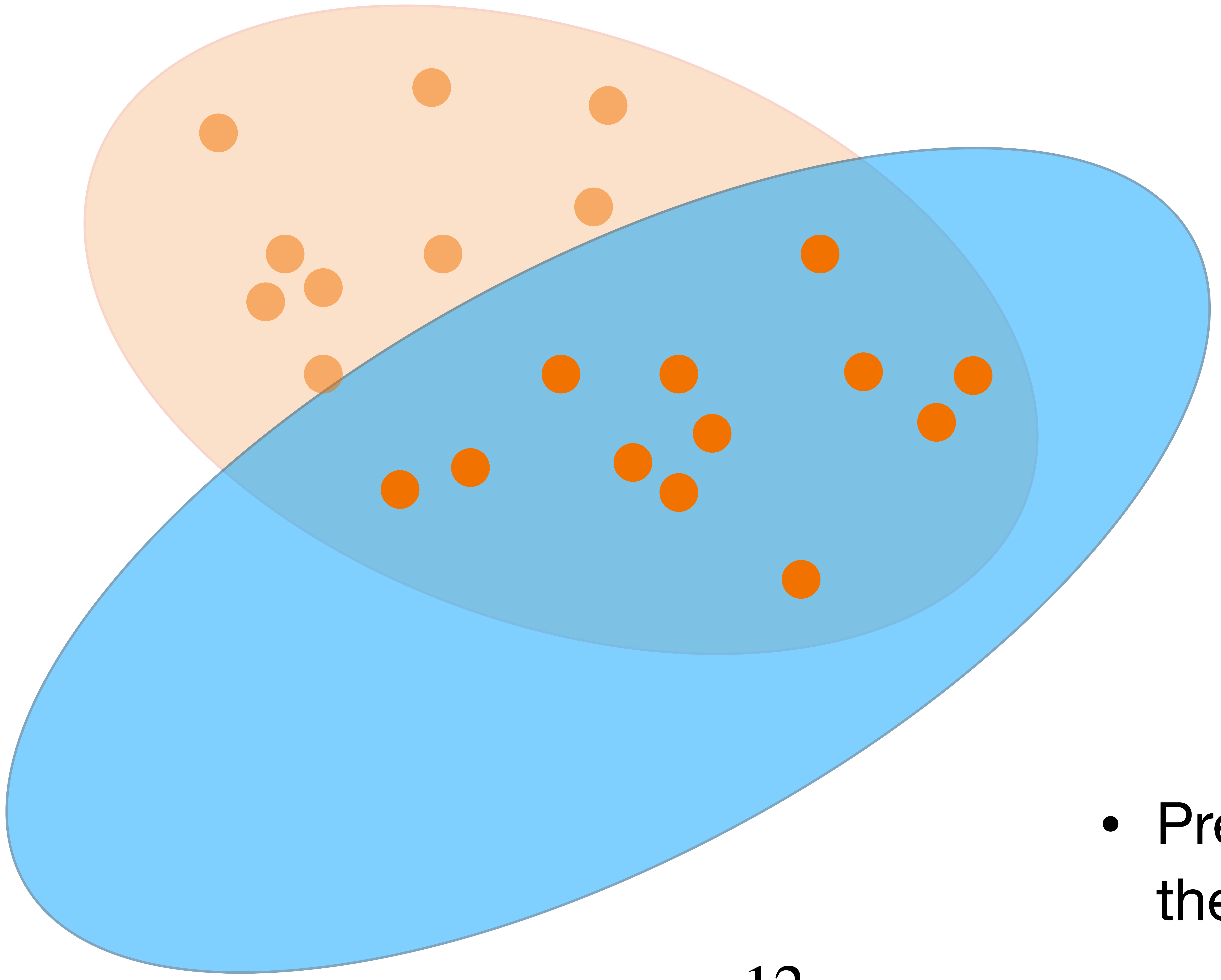
We have:

- real ★ and synthetic ● samples
- unknown distributions
- space where ”distances make sense”

Goal: disentangle quality and diversity

- Fidelity: How *realistic* each synthetic sample is?

First approach



$$\text{Precision} = \frac{12}{21}$$

We have:

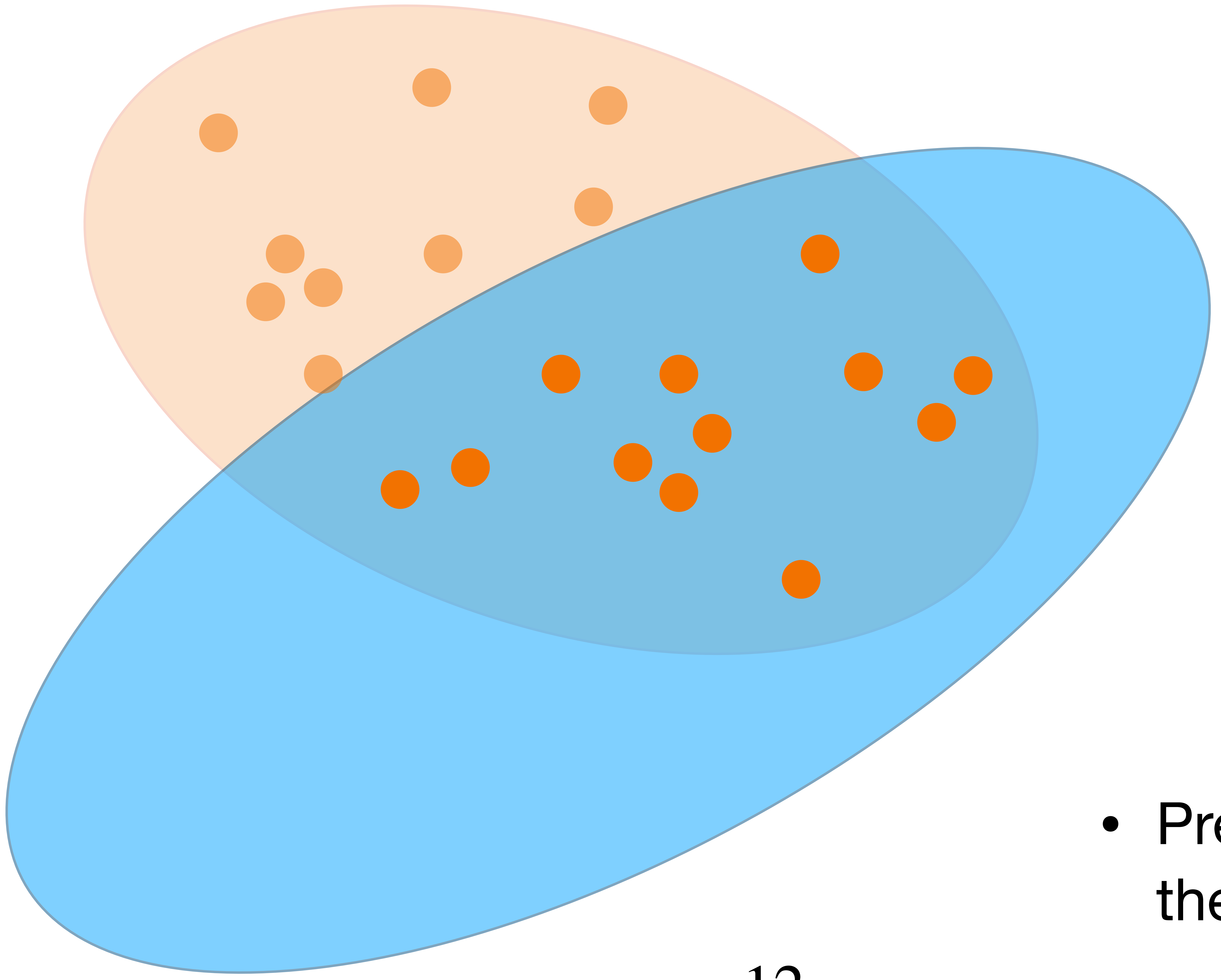
- real ★ and synthetic ● samples
- unknown distributions
- space where "distances make sense"

Goal: disentangle **quality** and diversity

- Fidelity: How *realistic* each synthetic sample is?

- Precision: proportion of synthetic points in the real support *(Sajjadi et al. 2018)*

First approach



$$\text{Precision} = \frac{12}{21}$$

We have:

- real ★ and synthetic ● samples
- unknown distributions
- space where ”distances make sense”

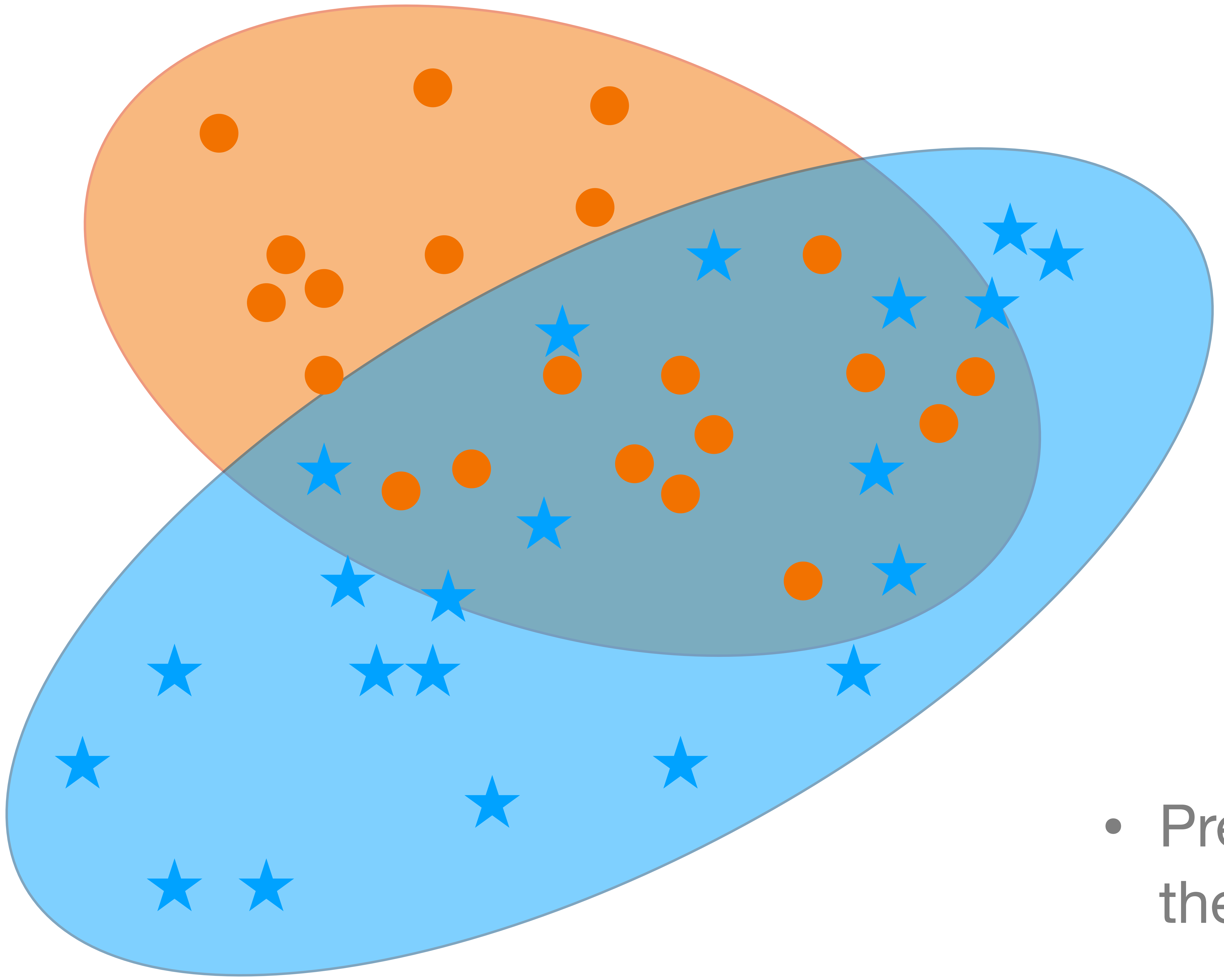
Goal: disentangle **quality** and diversity

- Fidelity: How *realistic* each synthetic sample is?

- Precision: proportion of synthetic points in the real support *(Sajjadi et al. 2018)*

Distribution-wise ≠ Point-wise
Realistic: could be sampled from the real distribution

Goal



We have:

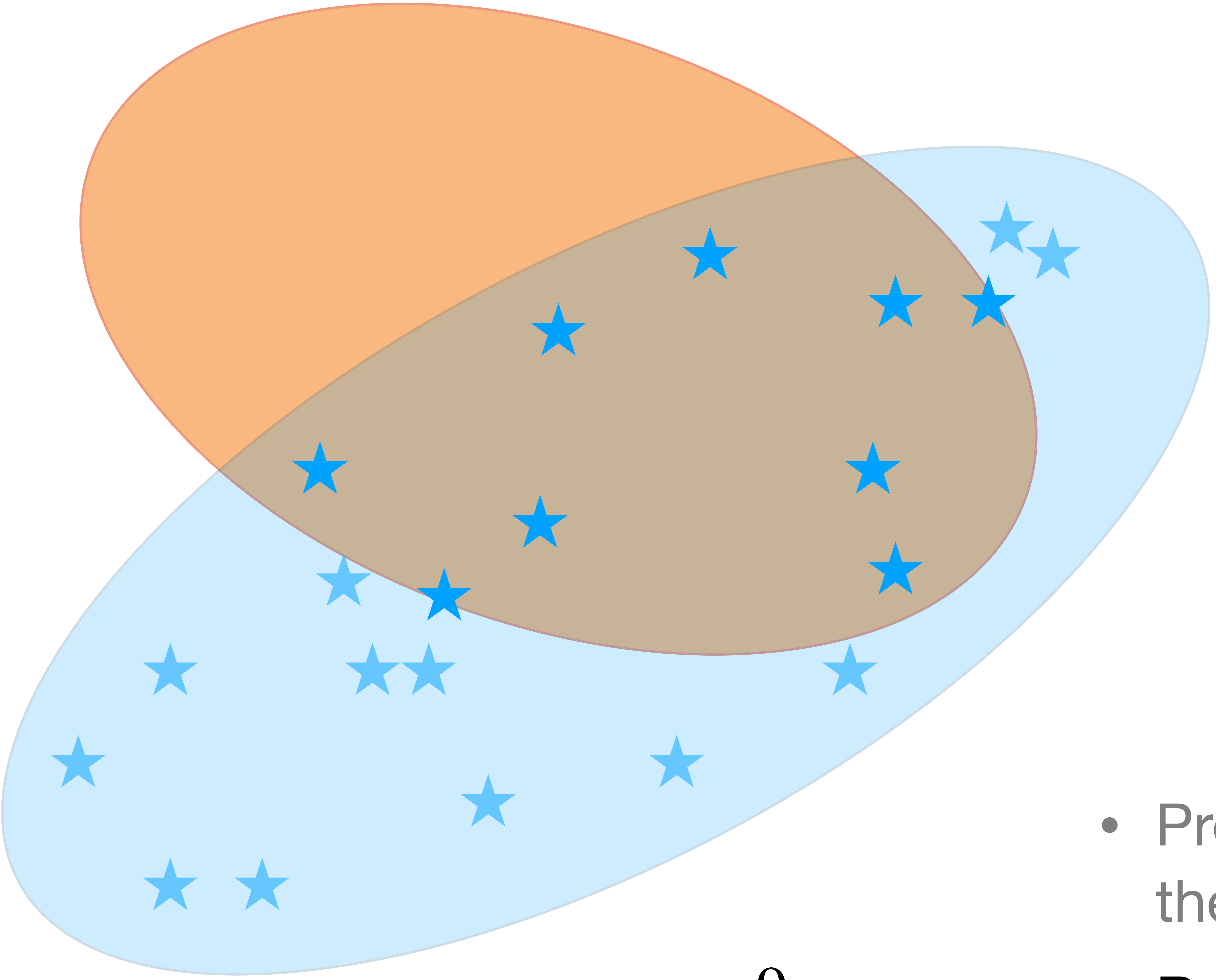
- real ★ and synthetic ● samples
- unknown distributions
- space where ”distances make sense”

Goal: disentangle quality and diversity

- Fidelity: How *realistic* each synthetic sample is?
- Coverage: To what extent do the synthetic samples represent the real *distribution*?

- Precision: proportion of synthetic points in the real support *(Sajjadi et al. 2018)*

First approach



$$\text{Recall} = \frac{9}{21}$$

We have:

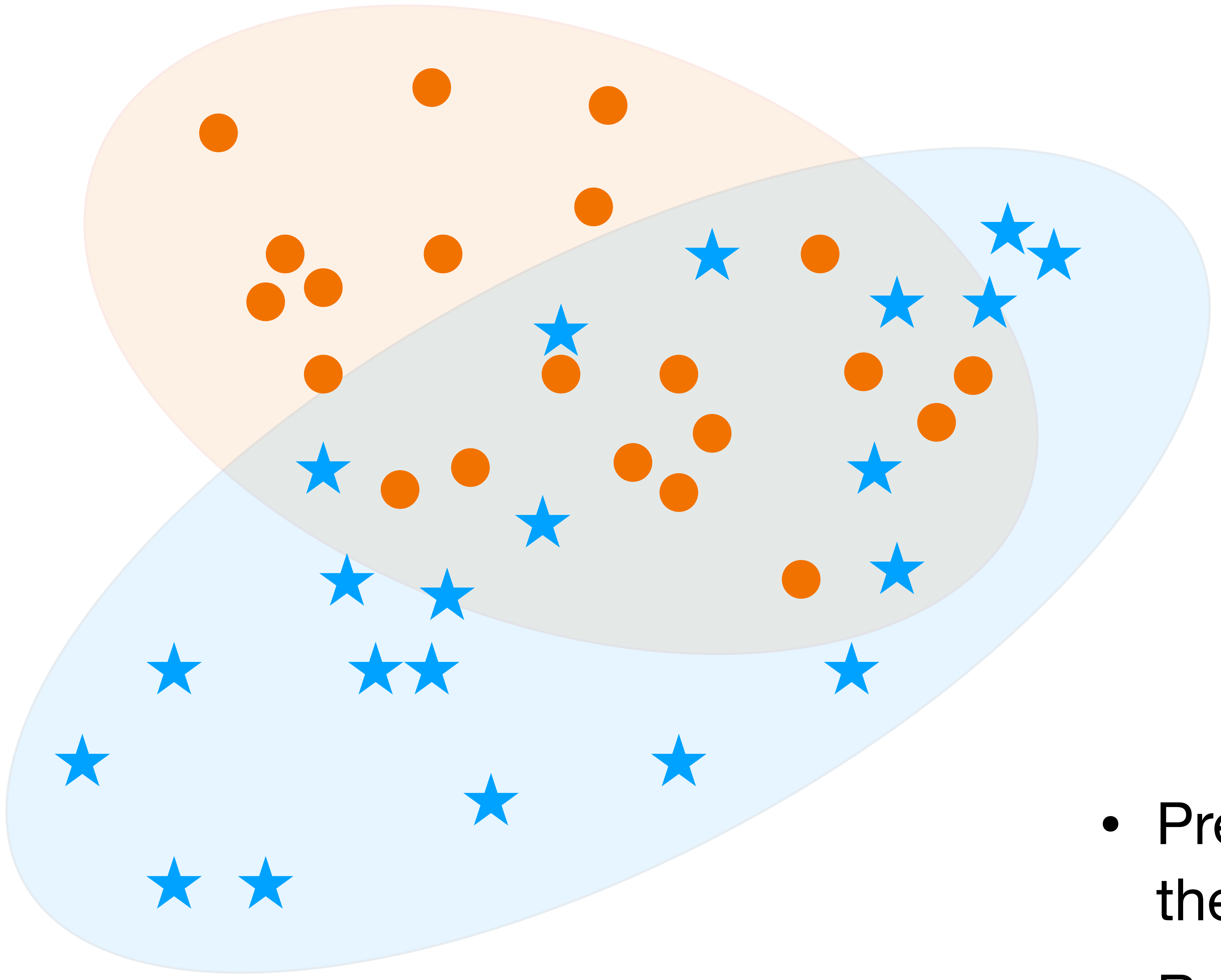
- real ★ and synthetic ● samples
- unknown distributions
- space where "distances make sense"

Goal: disentangle quality and diversity

- Fidelity: How *realistic* each synthetic sample is?
- Coverage: To what extent do the synthetic samples represent the real *distribution*?

- Precision: proportion of synthetic points in the real support *(Sajjadi et al. 2018)*
- Recall: proportion of real points in the synthetic support *(Sajjadi et al. 2018)*

First approach

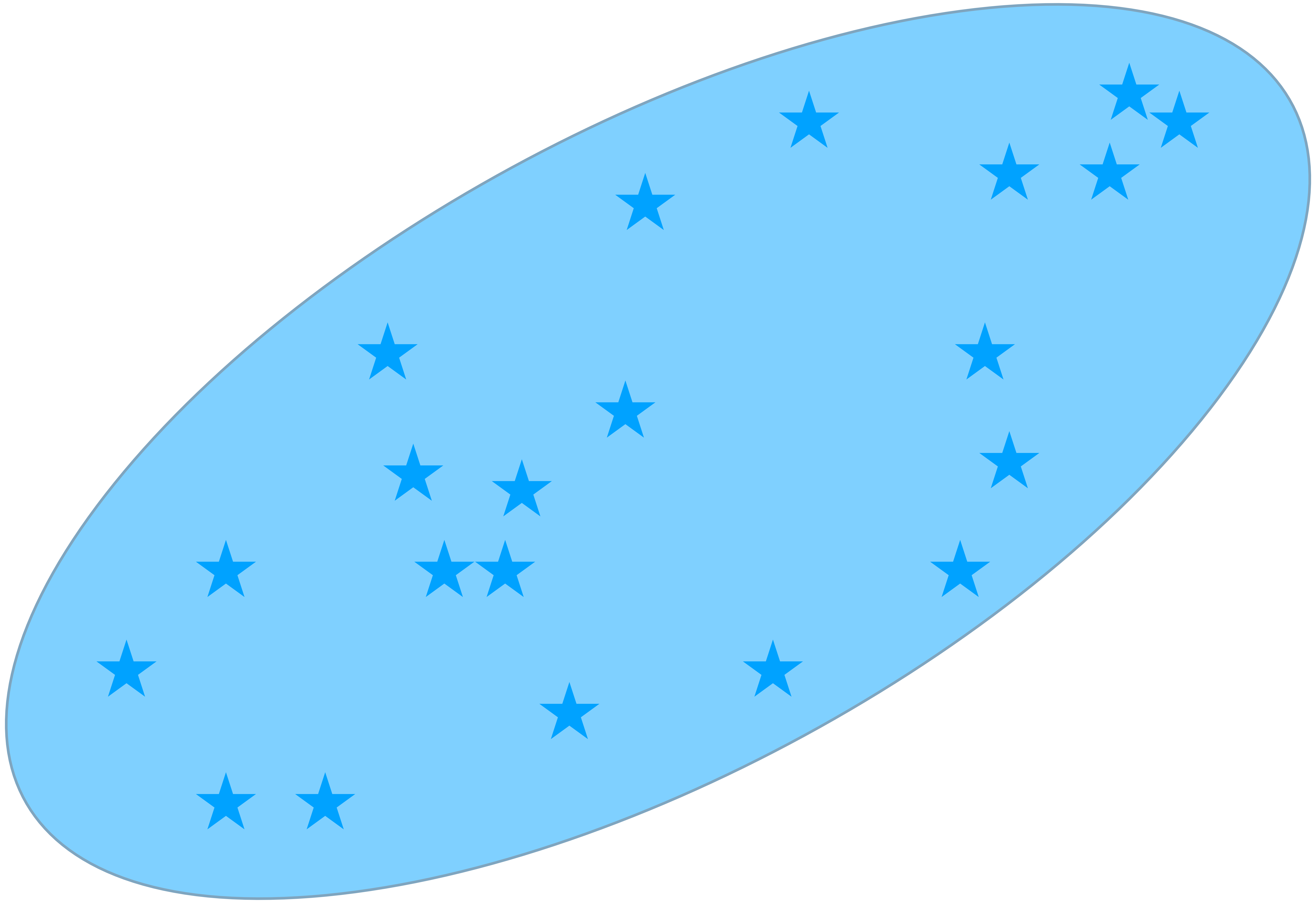


We have:

- real ★ and synthetic ● samples
- unknown distributions
- space where ”distances make sense”

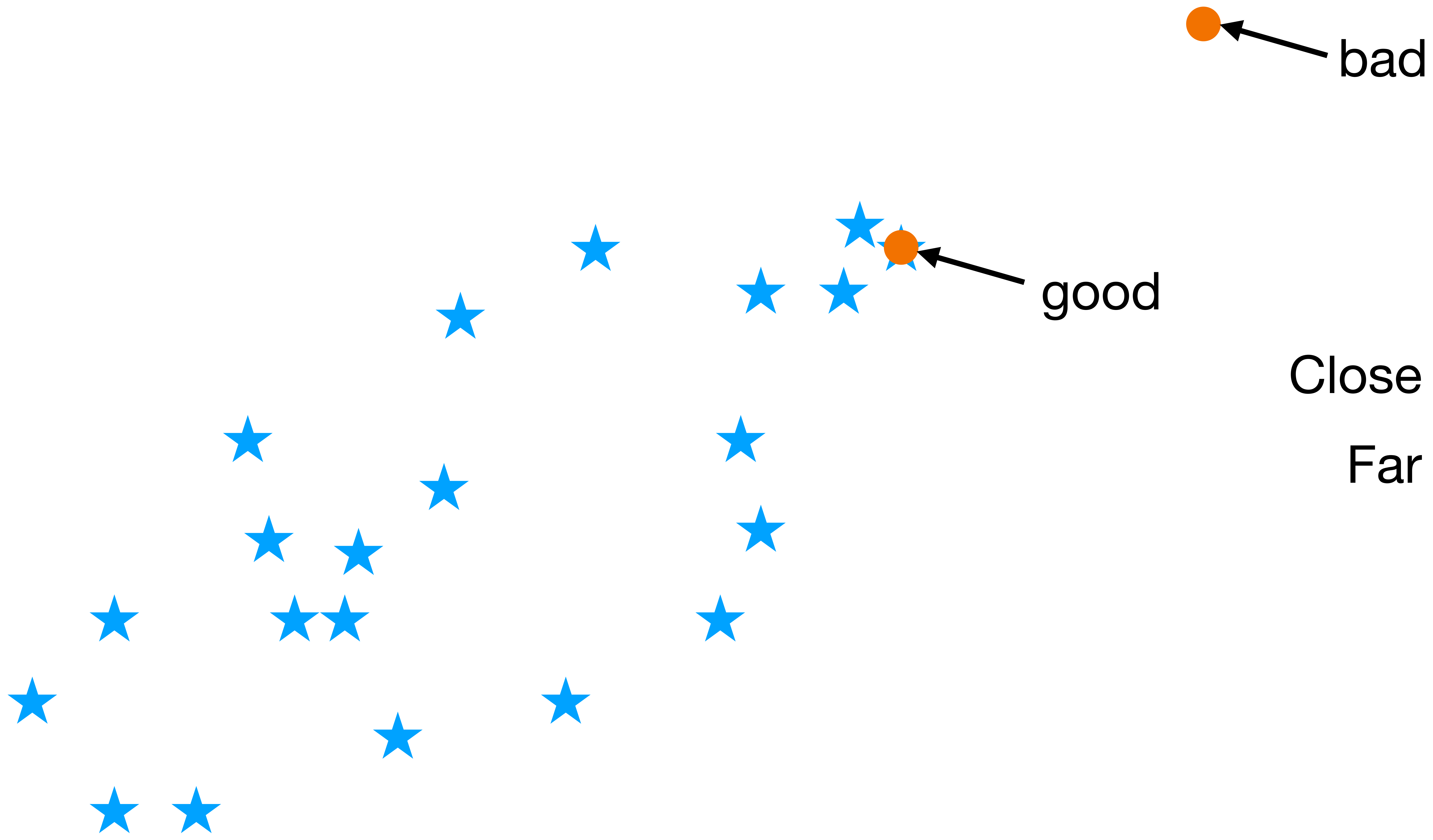
- Precision: proportion of synthetic points in the real support
- Recall: proportion of real points in the synthetic support

Unknown support



Unknown support

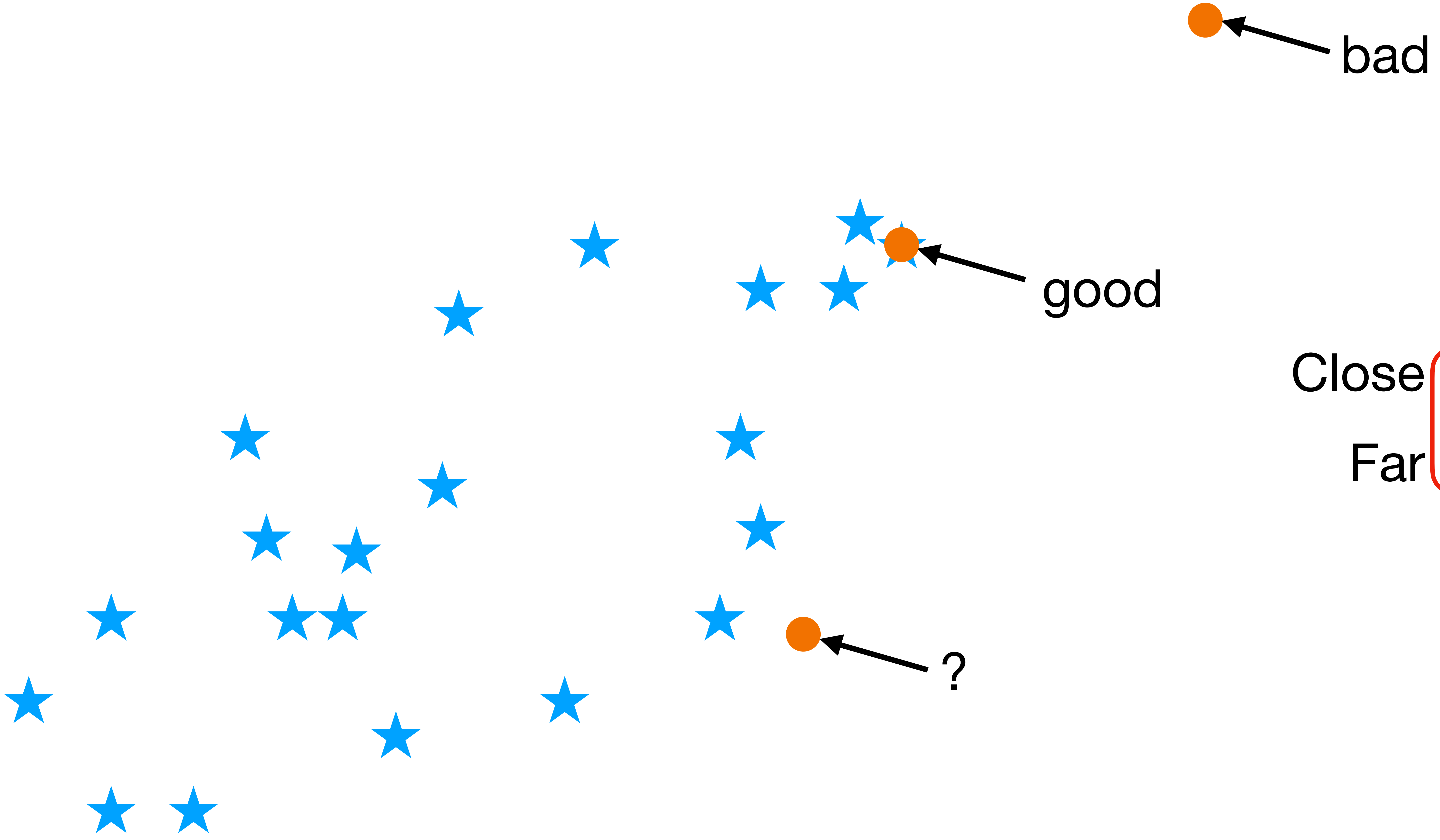
Relying on distances



Close enough ⇒ good

Far enough ⇒ bad

Relying on distances



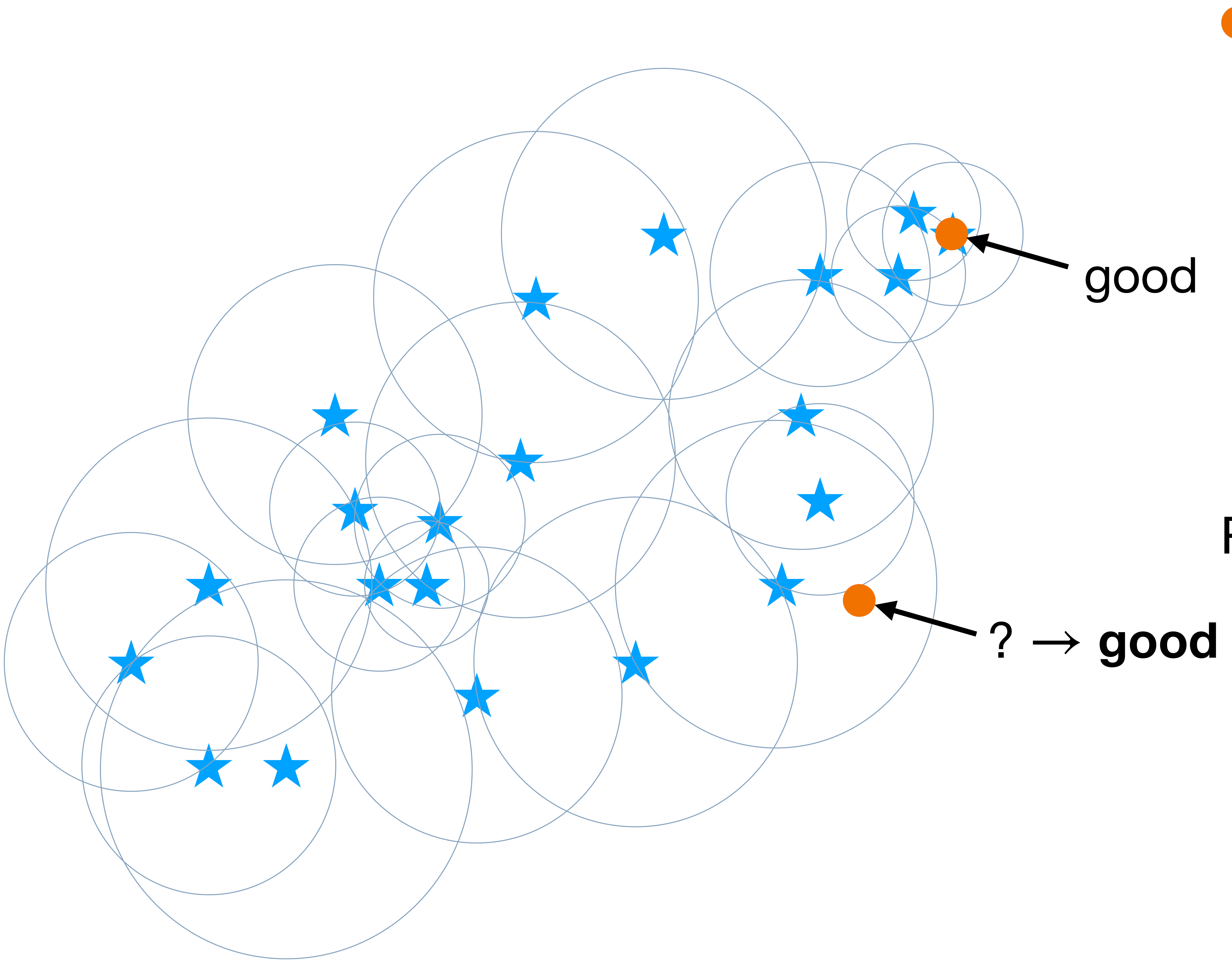
Close enough ⇒ good

?

Far enough ⇒ bad

Relying on distances

$k = 2$



Close $\boxed{\text{enough}}$ ⇒ good

?

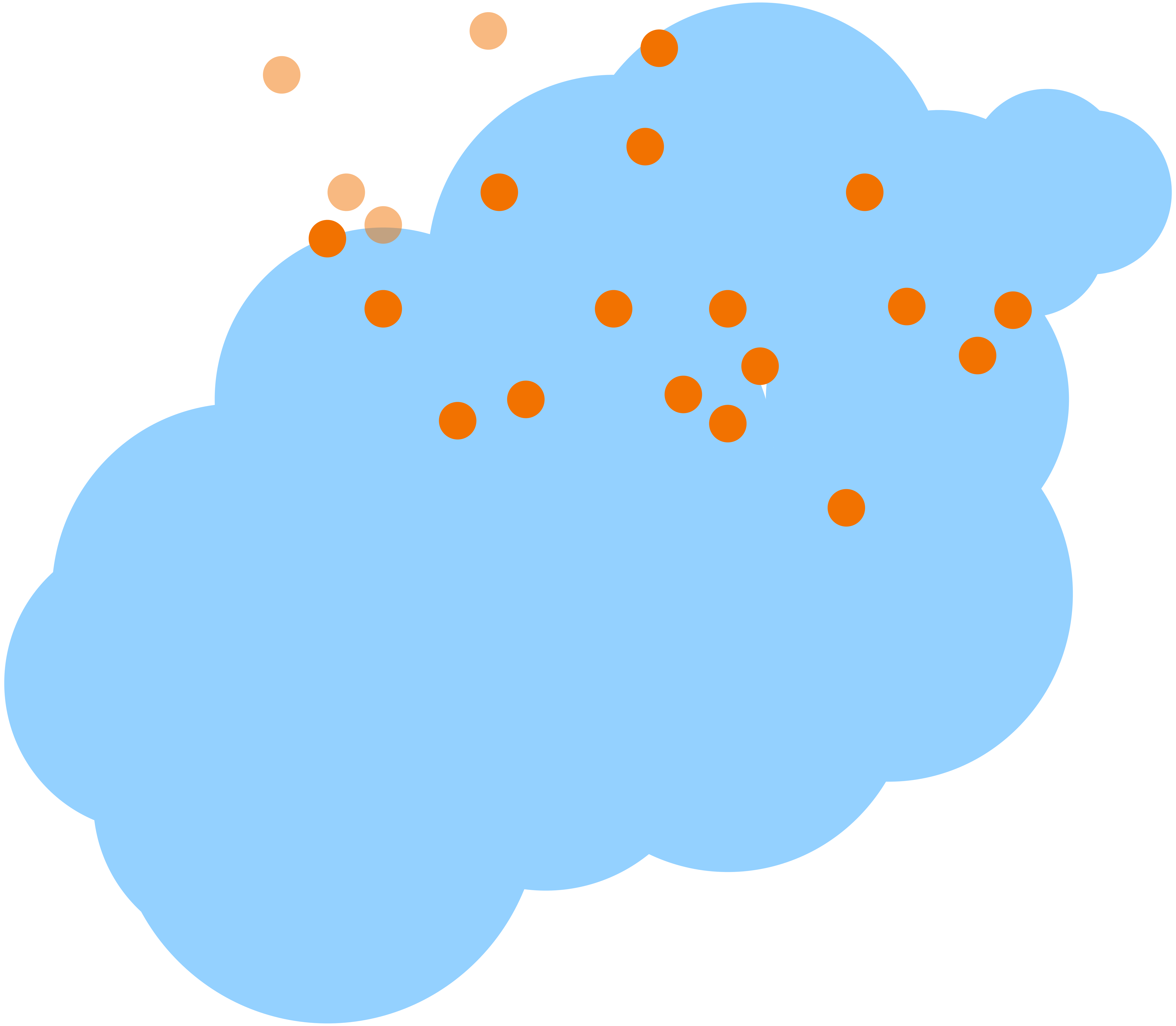
Far $\boxed{\text{enough}}$ ⇒ bad

Reference distance: $\mathrm{NND}_k^r(x_i^r)$

Approximating the support

First working approach

(Kynkäänniemi et al. 2019)



- **Fidelity**: How *realistic* each synthetic sample is?

- **Precision**: proportion of synthetic points in the real support

$$\text{iPrecision} = \frac{17}{21}$$

First working approach

(Kynkäänniemi et al. 2019)

- Fidelity: How *realistic* each synthetic sample is?
- **Coverage**: To what extent do the synthetic samples represent the real *distribution*?
- Precision: proportion of synthetic points in the real support

$$\text{iPrecision} = \frac{17}{21}$$

- **Recall**: proportion of real points in the synthetic support

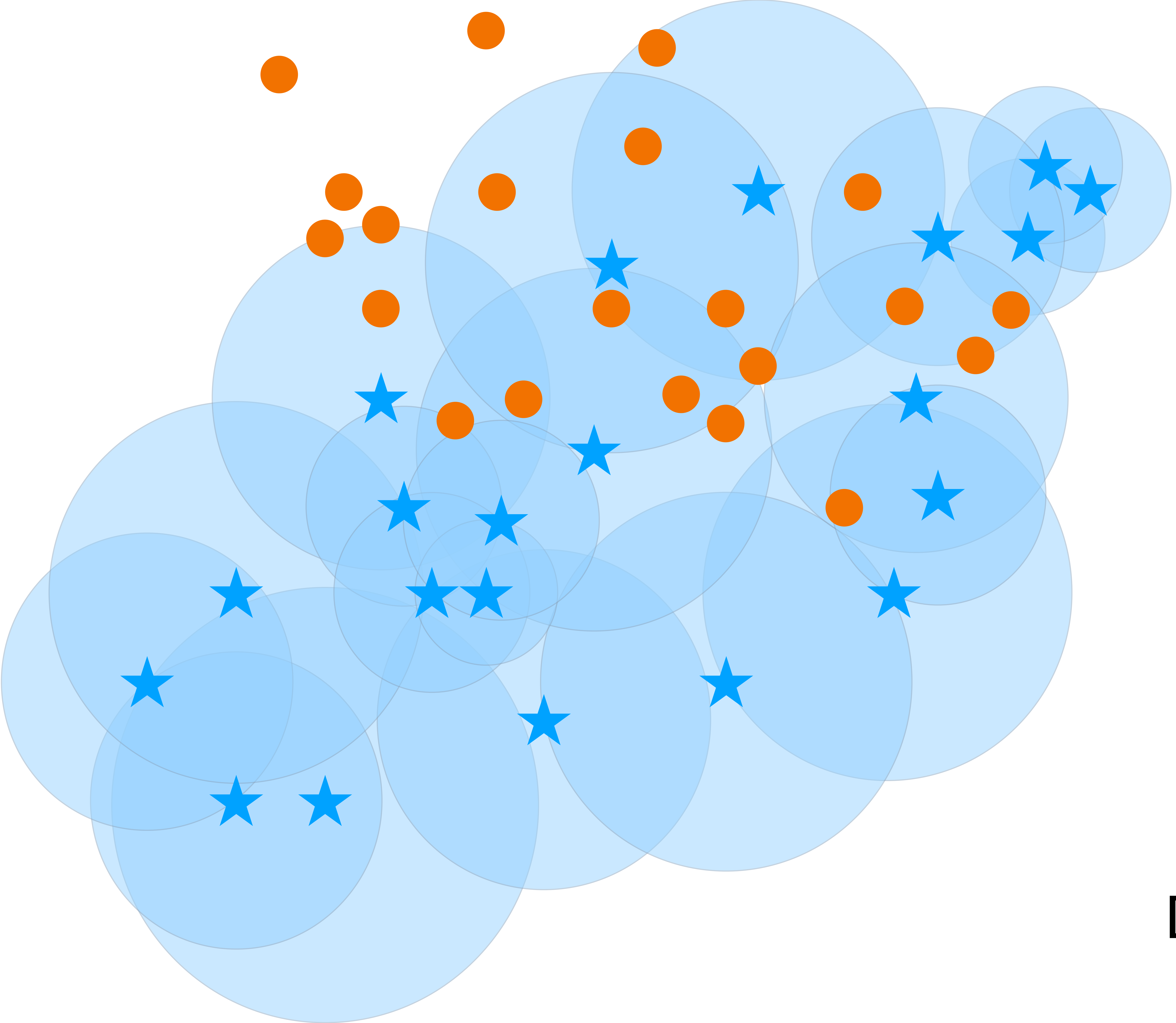
$$\text{iRecall} = \frac{10}{21}$$

Fidelity & Coverage metrics

- Fidelity: How *realistic* each synthetic sample is?
- Coverage: To what extent do the synthetic samples represent the real *distribution*?
- Most Fidelity & Coverage metrics follow these steps: approximation, simple properties

Density & Coverage

(Naeem et al. 2020)



$$\text{Precision} = \frac{1}{M}\sum_{j=1}^{M} 1_{x_j^s \in \cup_{i=1}^{N} B(x_i^r, NND_k^r(x_i^r))}$$

Is each synthetic point in a ball?

$$\text{Density} = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{k} \sum_{i=1}^{N} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

In how many balls each synthetic point is?

$$\text{Density} = \frac{1}{21}\left(\frac{1}{2} \times 12 + \frac{2}{2} \times 5\right) = \frac{11}{21}$$

Density & **Coverage**

(Naeem et al. 2020)

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N} 1_{x_i^r \in \cup_{j=1}^{M} B(x_j^s, NND_k^s(x_j^s))}$$

Is each real point in a synthetic ball?

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists j, x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

Is there a synthetic point in each real ball?

$$\text{Coverage} = \frac{9}{21}$$

Fidelity & Coverage metrics

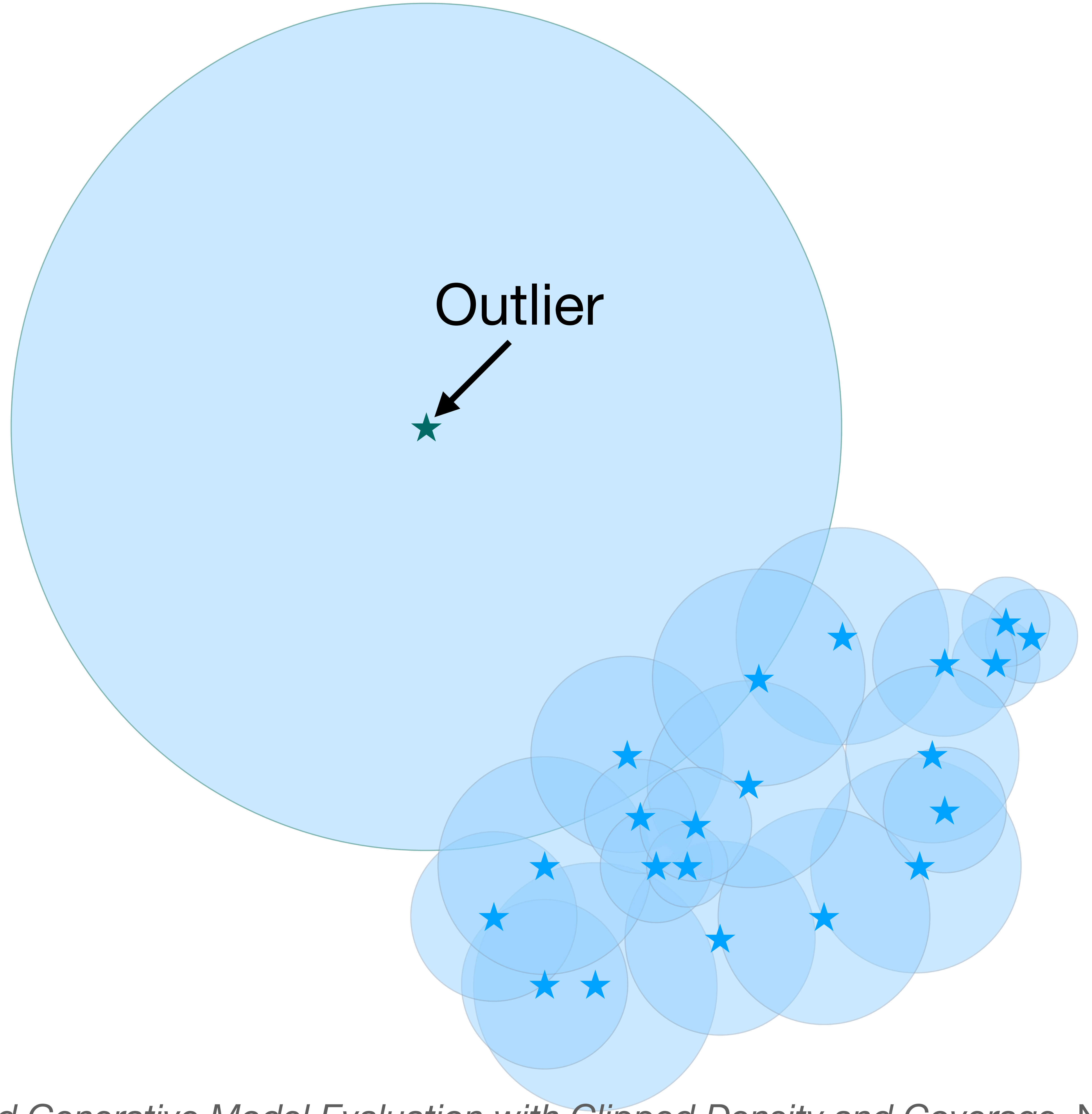
- Most Fidelity & Coverage metrics follow these steps: approximation, simple properties
- They also share their weaknesses: **robustness** and **interpretability** *(Raïsä et al., Position: All Current Generative Fidelity and Diversity Metrics are Flawed, ICML 2025)*

Enhanced Generative Model Evaluation with Clipped Density and Coverage

submitted to ICLR

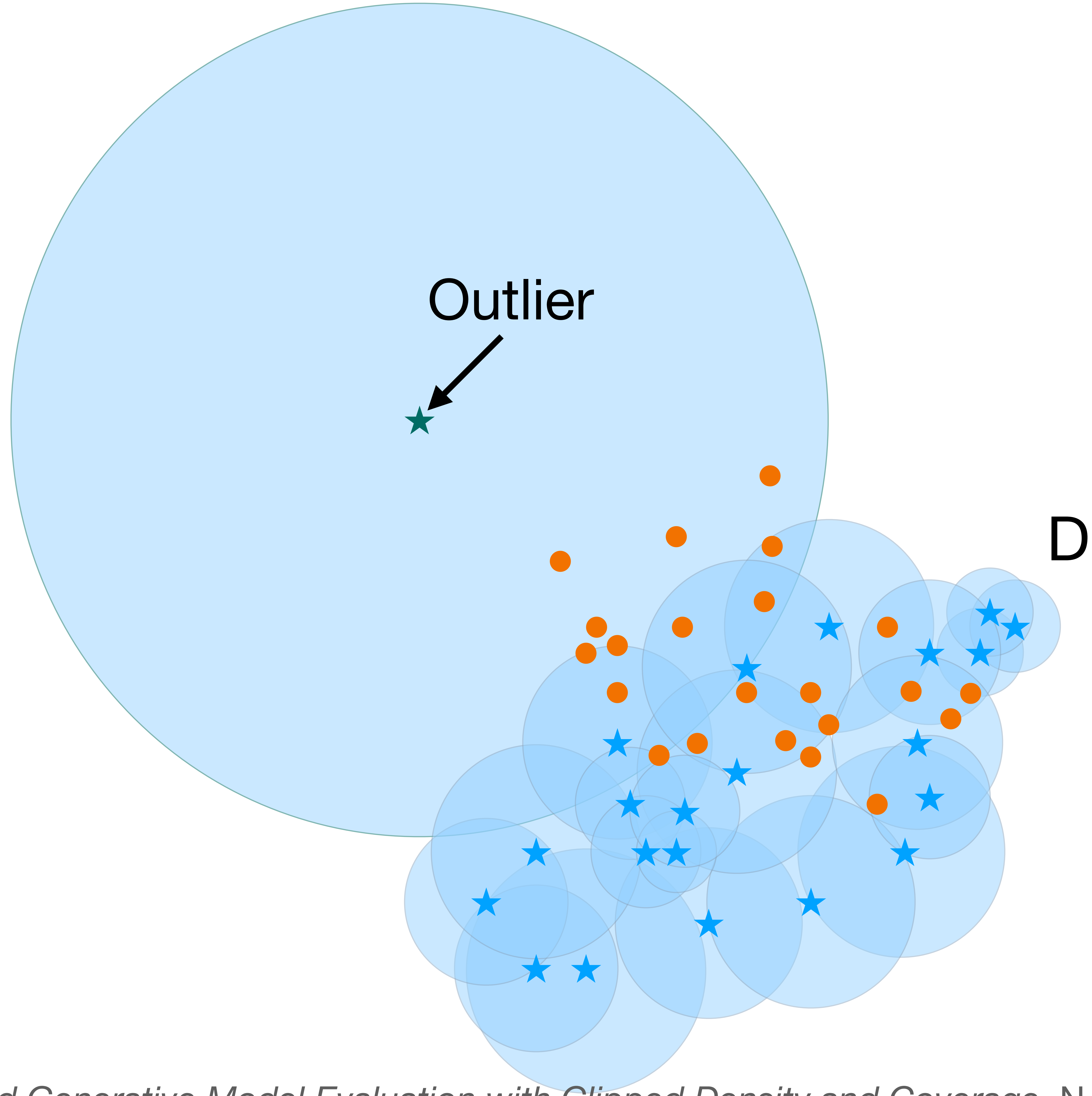
- Most Fidelity & Coverage metrics follow these steps: approximation, simple properties
- They also share their weaknesses: **robustness** and **interpretability** *(Raïsä et al., Position: All Current Generative Fidelity and Diversity Metrics are Flawed, ICML 2025)*

Robustness



Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Robustness: Density



$$\text{Density} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{k}\sum_{i=1}^{N}1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

In how many balls each synthetic point is?

$$\text{Density} = \frac{1}{22}\left(\frac{1}{2}\times 12 + \frac{2}{2}\times 8 + \frac{3}{2}\times 2\right)$$

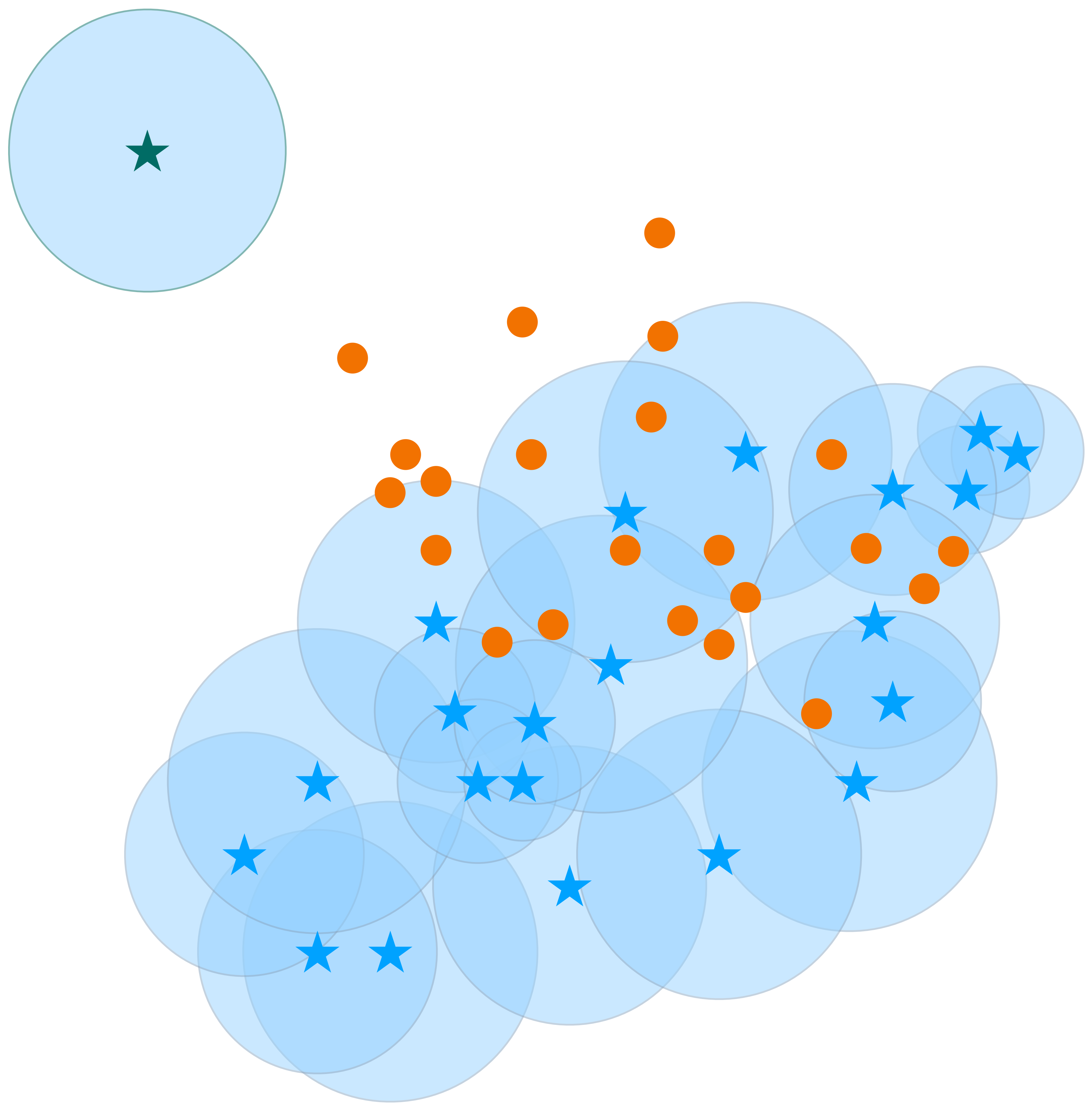
$$= \frac{17}{22} \qquad \leftarrow \frac{11}{21}$$

Robustness: Density

→ **Clip the radii** to the median: $R_k(x_i^r)$

$$\text{score} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{k}\sum_{i=1}^{N} 1_{x_j^s \in B(x_i^r, R_k(x_i^r))}$$

In how many clipped balls each synthetic point is?

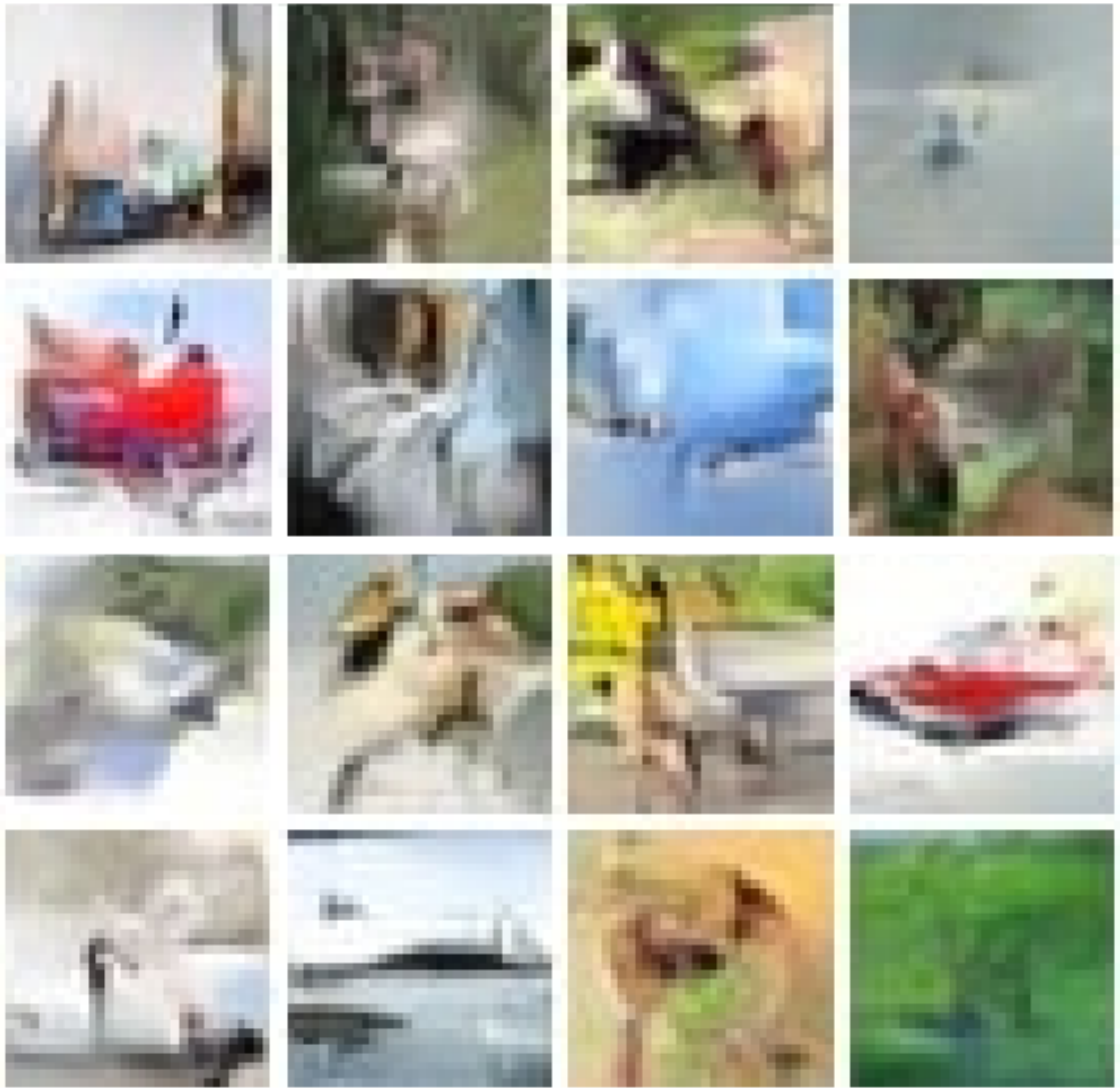


$$\text{score} = \frac{1}{22}\left(\frac{1}{2}\times 12 + \frac{2}{2}\times 5 + \frac{3}{2}\times 0\right)$$

$$= \frac{11}{22} \qquad \leftarrow \frac{11}{21}$$

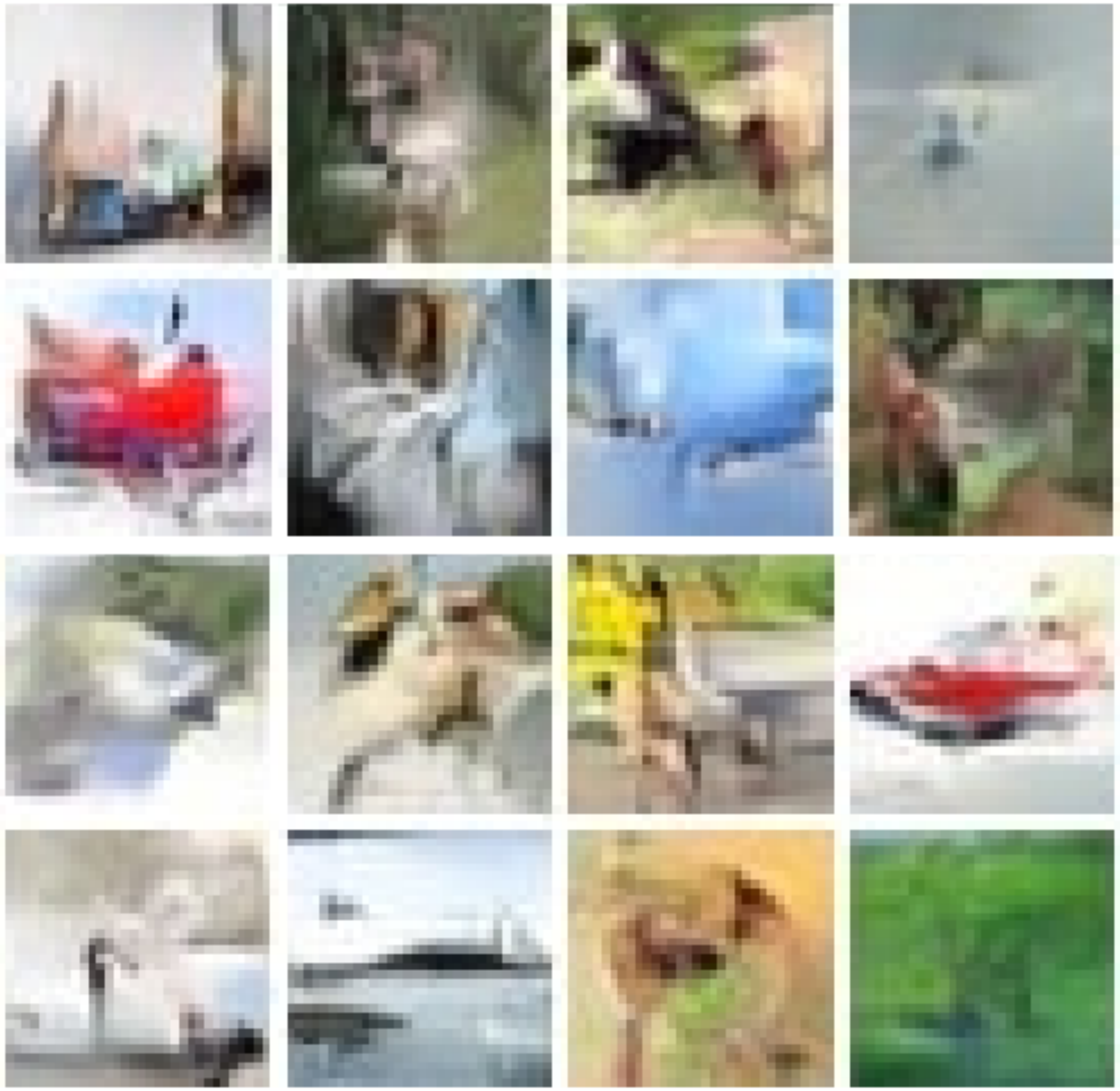
Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Outliers in practice: RESFLOW CIFAR-10



Density $= 2.47$

Outliers in practice: RESFLOW CIFAR-10

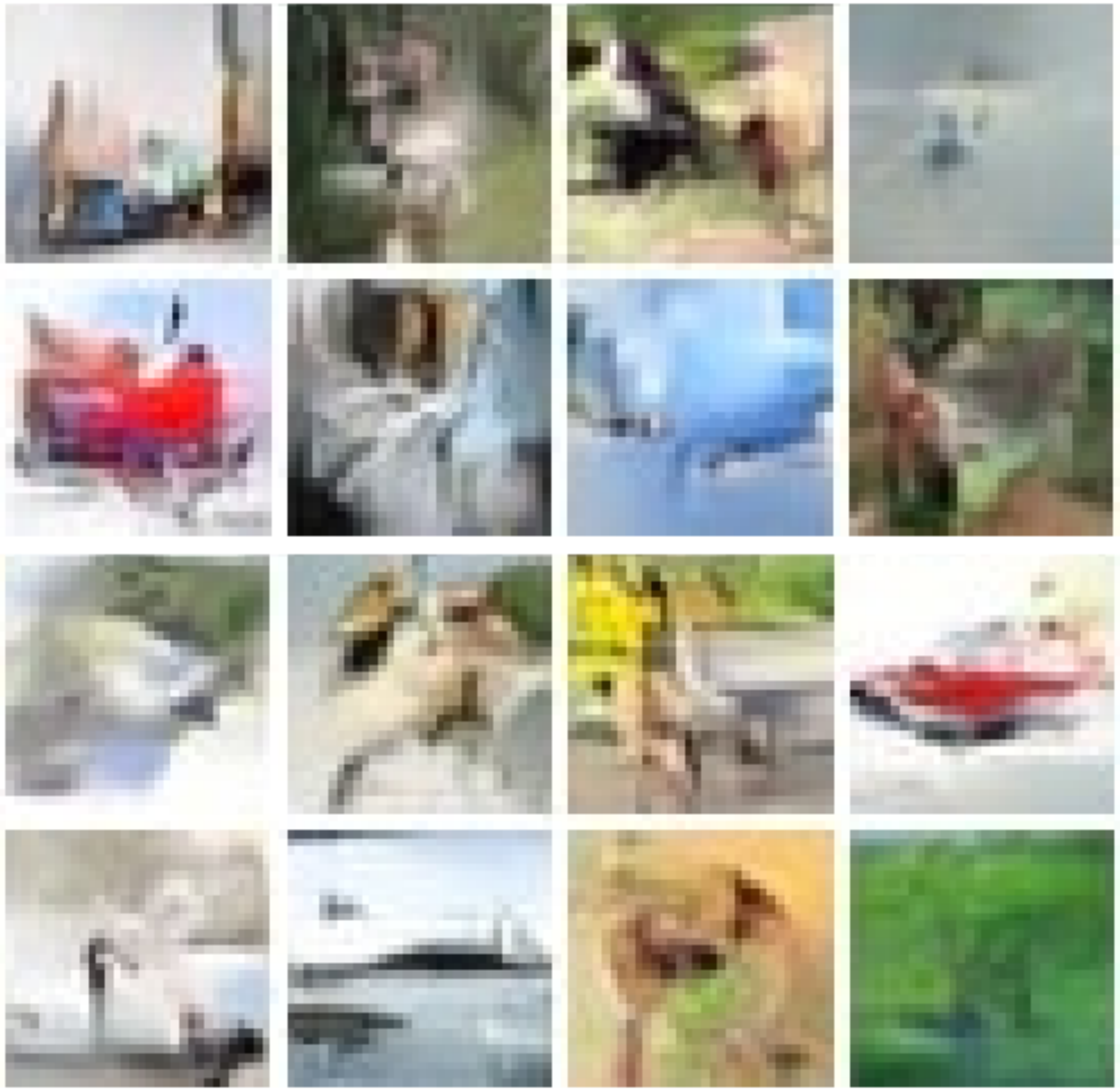


Density $= \textcolor{red}{2.47}$

4 real images have > 10000 synthetic images in their ball!

Very gray **ship** image	Stealthy **cat**	2-legged **horse**	Toy **ship** on a stand

Outliers in practice: RESFLOW CIFAR-10



Density $= 2.47$

4 real images have > 10000 synthetic images in their ball!

Very gray **ship** image

Stealthy **cat**

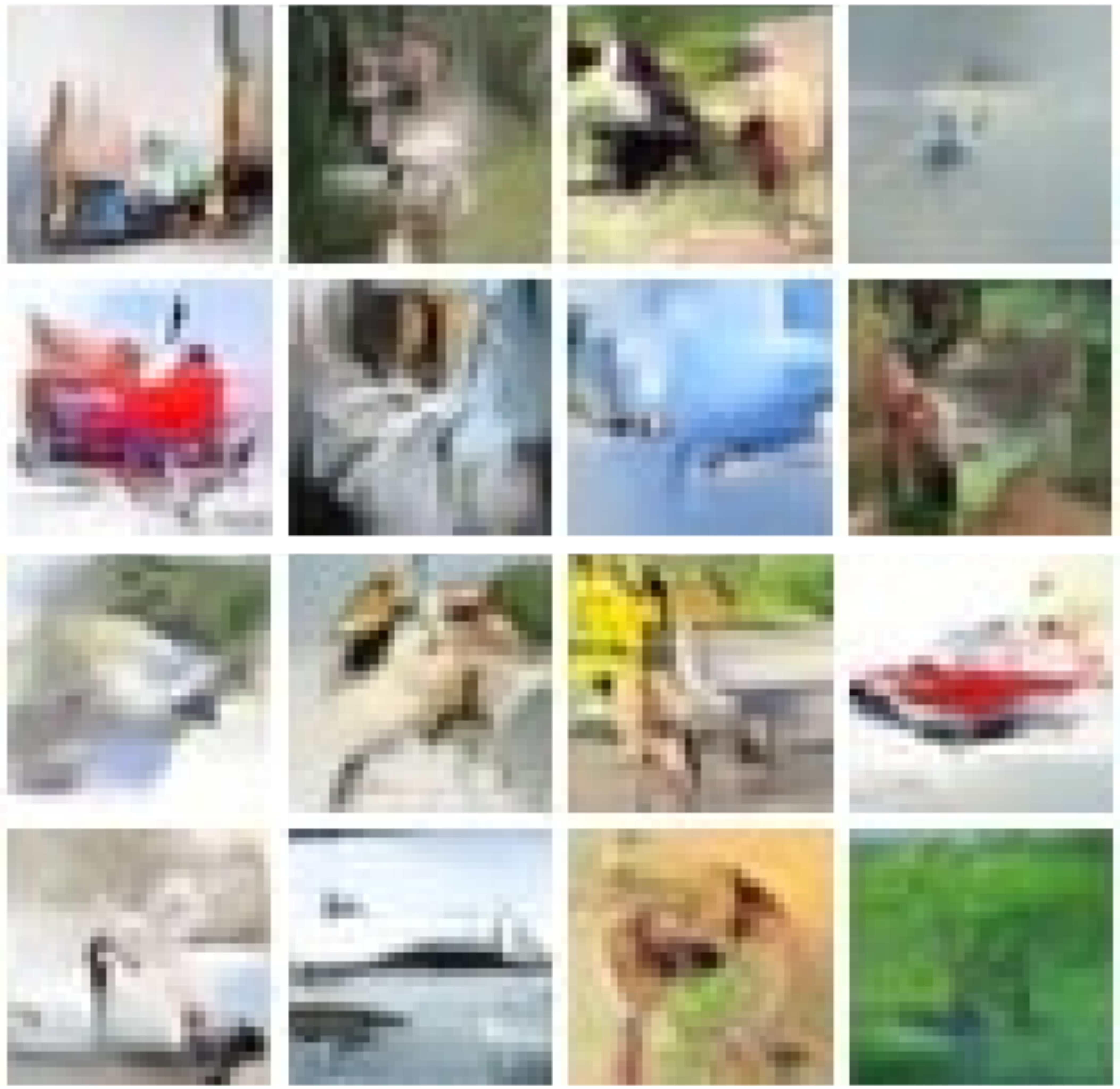
2-legged **horse**

Toy **ship** on a stand

Clipped radii $\rightarrow$ score $= 0.00$

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Outliers in practice: RESFLOW CIFAR-10



How metrics are evaluated:

know the answer

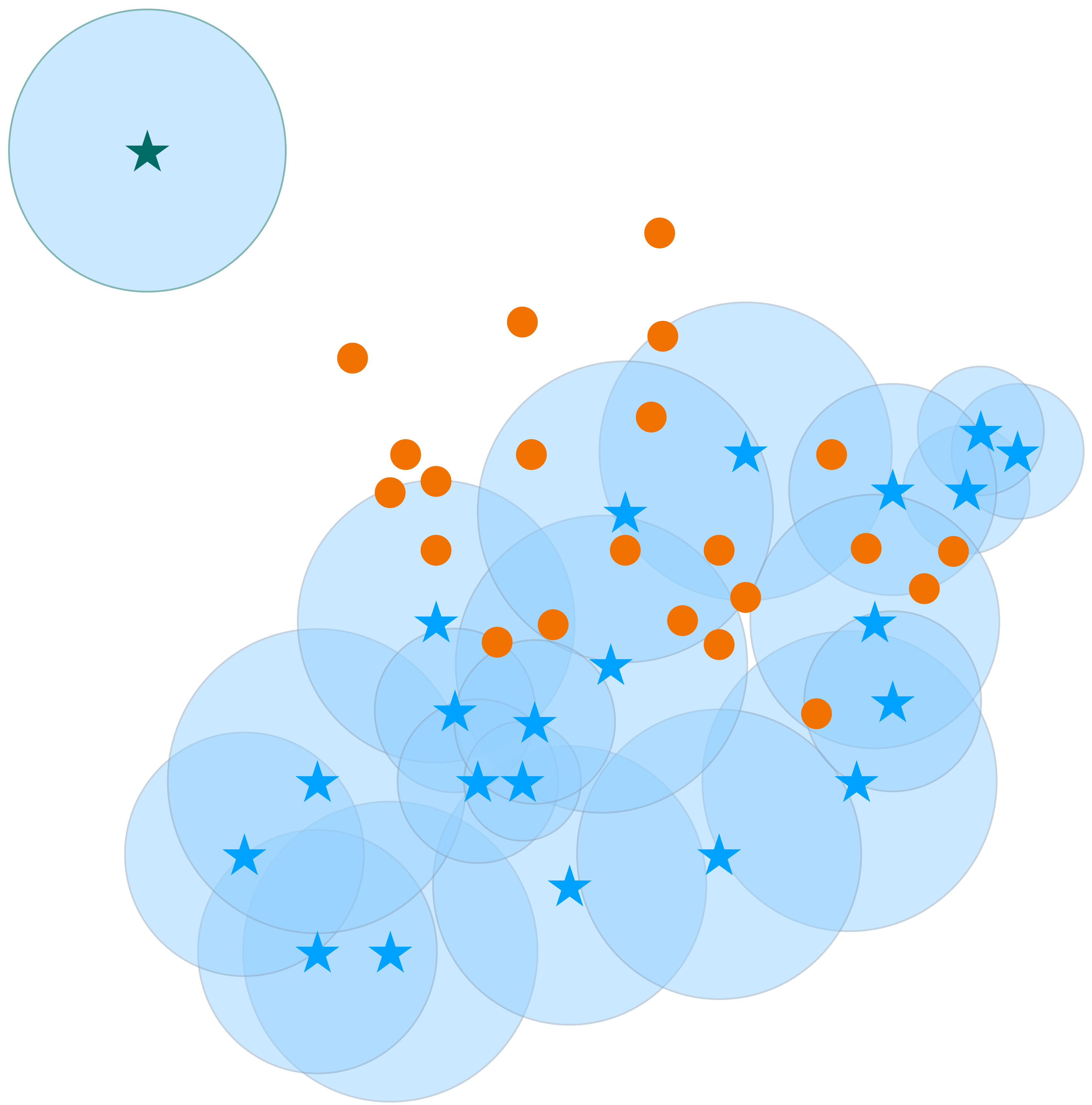
Density = 2.47 **failure**

Robustness: Density

→ **Clip the radii** to the median: $R_k(x_i^r)$

$$\text{score} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{k}\sum_{i=1}^{N}1_{x_j^s \in B(x_i^r, R_k(x_i^r))}$$

In how many clipped balls each synthetic point is?



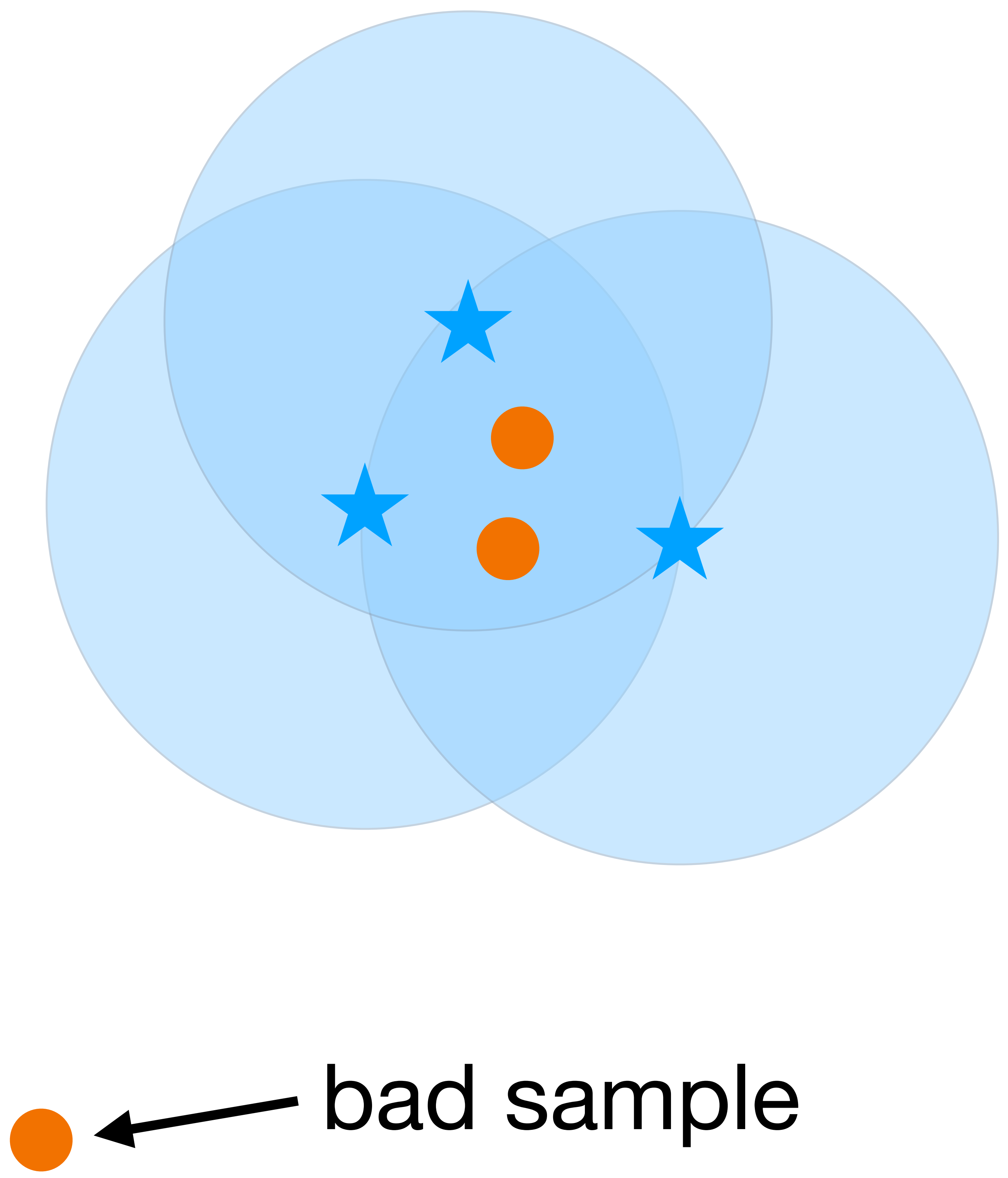
$$\text{score} = \frac{1}{22}\left(\frac{1}{2}\times 12 + \frac{2}{2}\times 5 + \frac{3}{2}\times 0\right)$$

$$= \frac{11}{22}$$

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Robustness: Density

→ **Clip the radii** to the median: $R_k(x_i^r)$



$$\text{score} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{k}\sum_{i=1}^{N} 1_{x_j^s \in B(x_i^r, R_k(x_i^r))}$$

In how many clipped balls each synthetic point is?

$$\text{score} = \frac{1}{3}\left(\frac{3}{2}\times 2\right) = \frac{3}{3} = 1$$

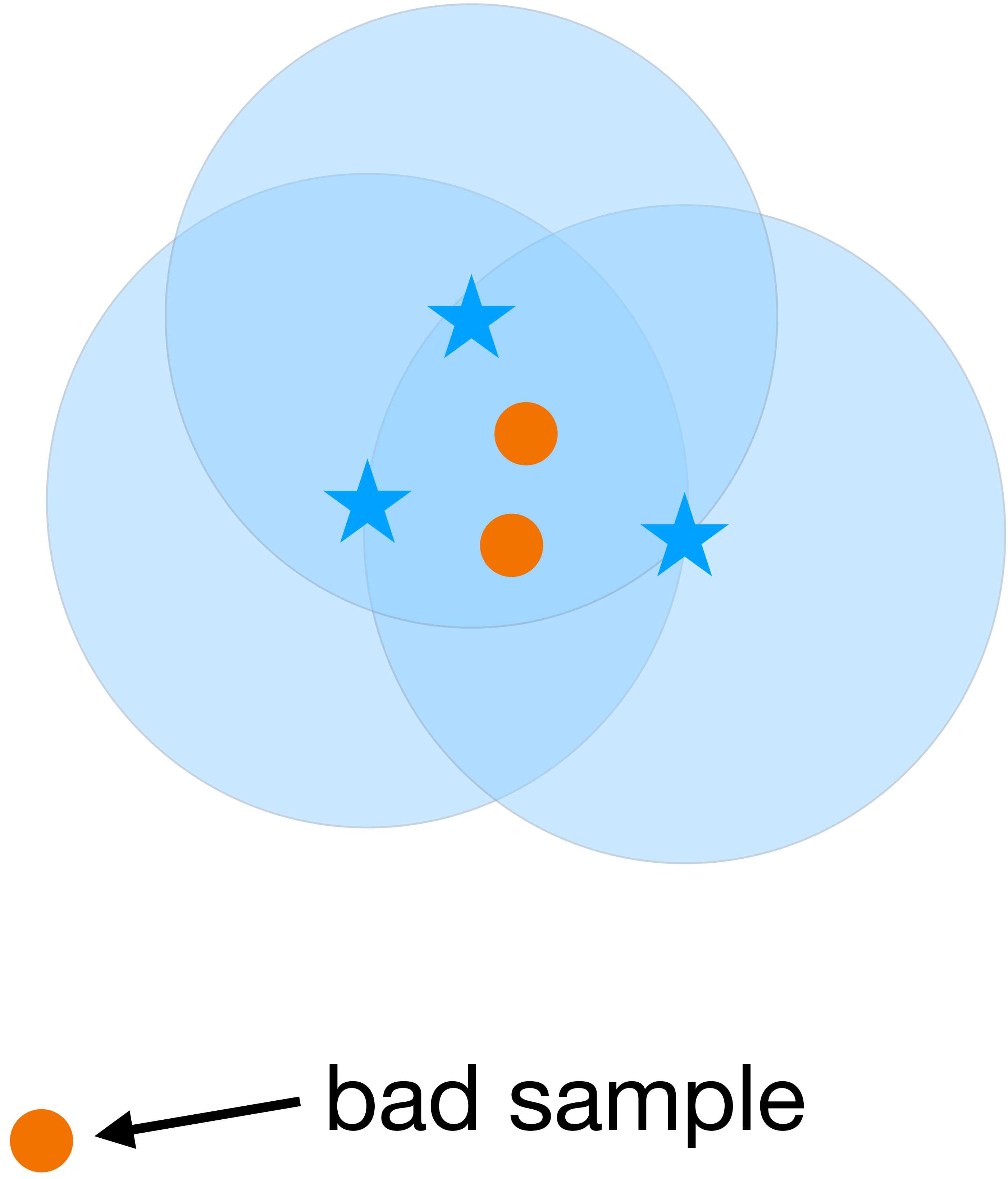
Robustness: Density

→ **Clip the radii** to the median: $R_k(x_i^r)$

→ **Clip the contributions**

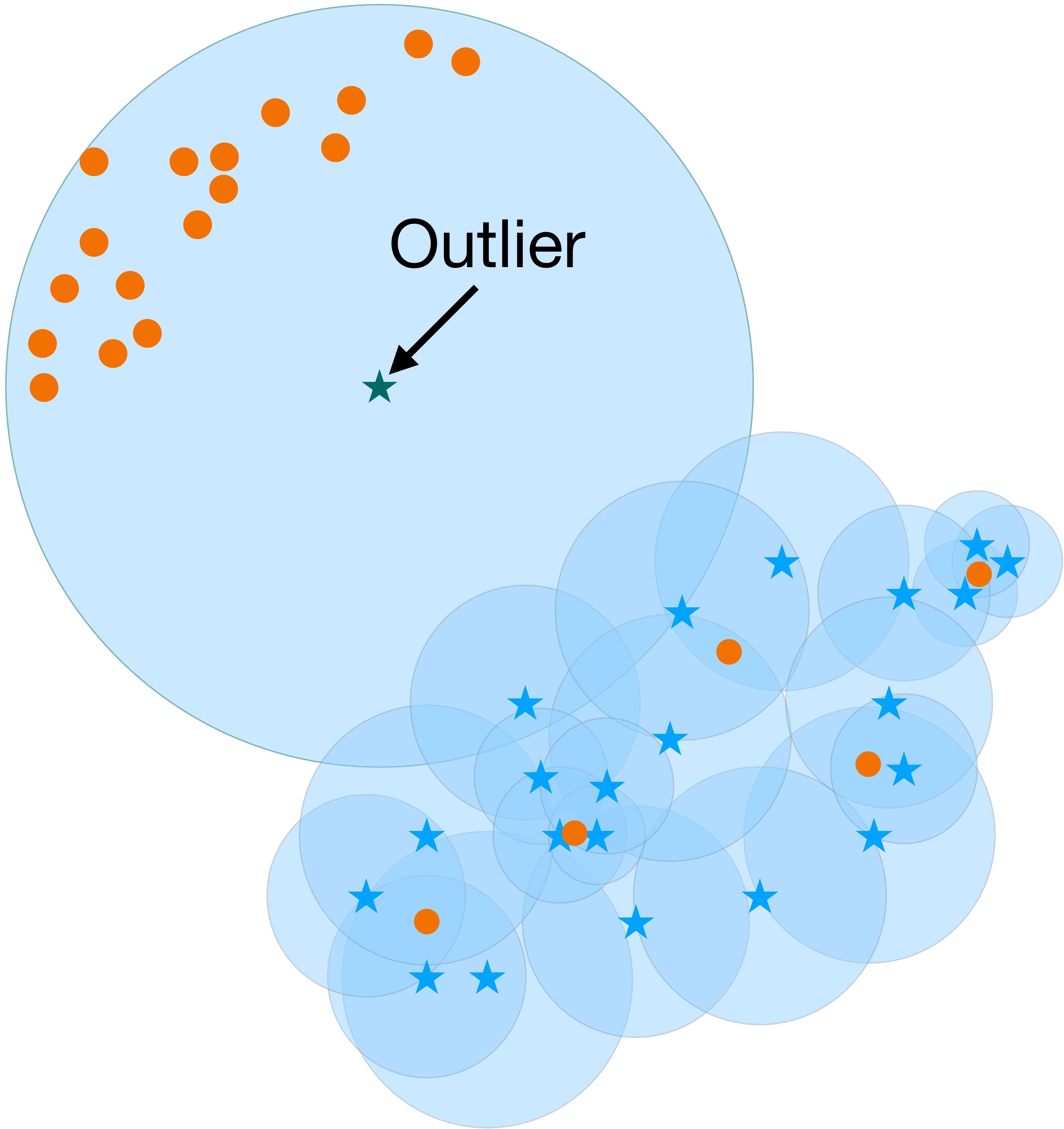
$$\text{ClippedDensity}_{\text{unnorm}} = \frac{1}{M}\sum_{j=1}^{M} \min\left(\frac{1}{k}\sum_{i=1}^{N} 1_{x_j^s \in B(x_i^r, R_k(x_i^r))}, 1\right)$$

In how many clipped balls each synthetic point is?



$$\text{ClippedDensity}_{\text{unnorm}} = \frac{1}{3}\left(\min\left(\frac{3}{2}, 1\right) \times 2\right) = \frac{2}{3}$$

Robustness: Coverage



$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists j, x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

Is there a synthetic point
in each real ball?

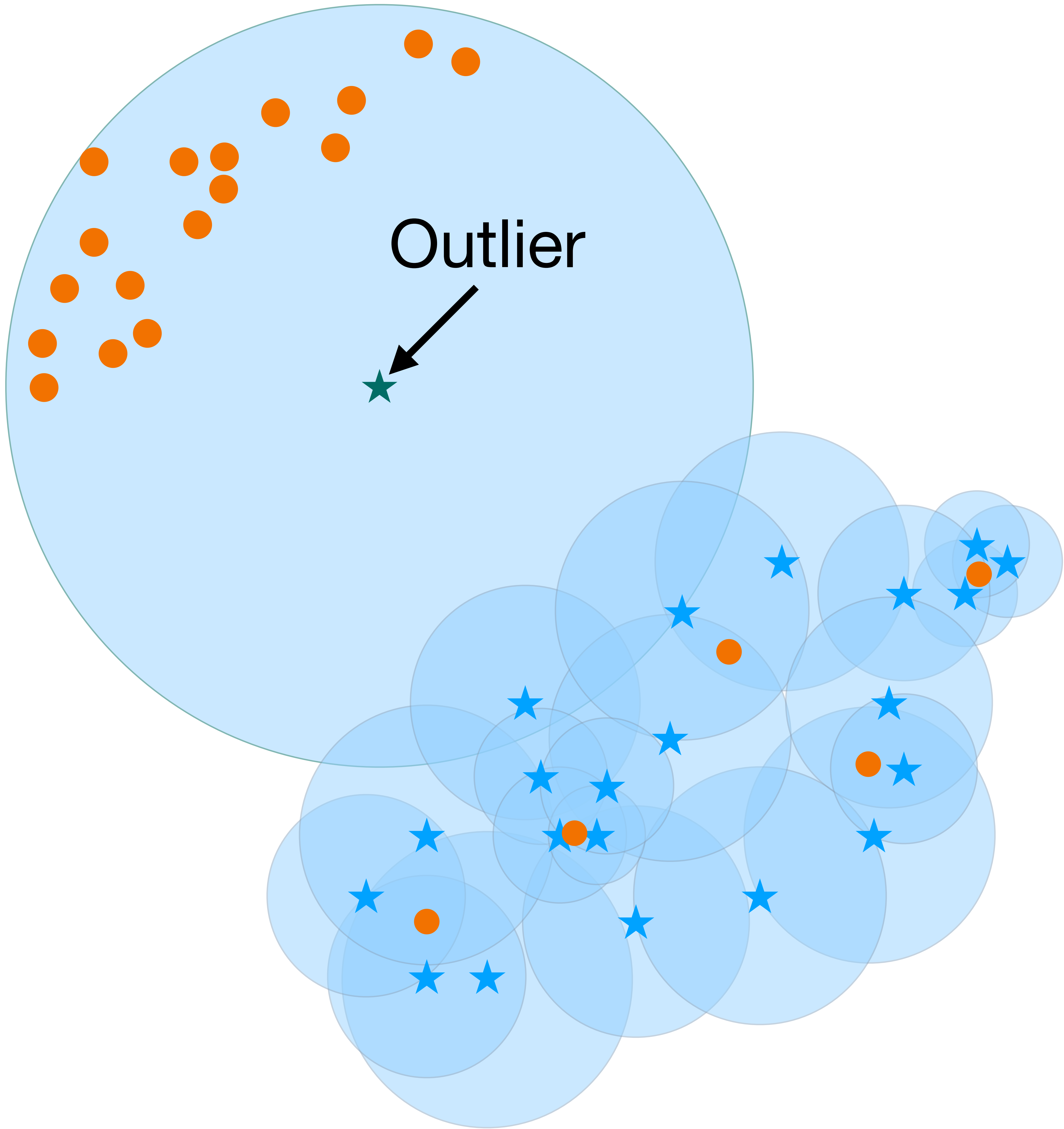
$$\text{score} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

How many synthetic points
are there in each real ball?

$$\text{score} = \frac{1}{22}\left(\frac{1}{2}\times 19 + \frac{17}{2}\times 1\right)$$

$$= \frac{36}{44}$$

Robustness: Coverage



$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists j, x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

Is there a synthetic point in each real ball?

$$\text{score} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

How many synthetic points are there in each real ball?

$$\text{score} = \frac{1}{22}\left(\frac{1}{2}\times 19 + \frac{17}{2}\times 1\right)$$

$$= \frac{36}{44}$$

Robustness: Coverage

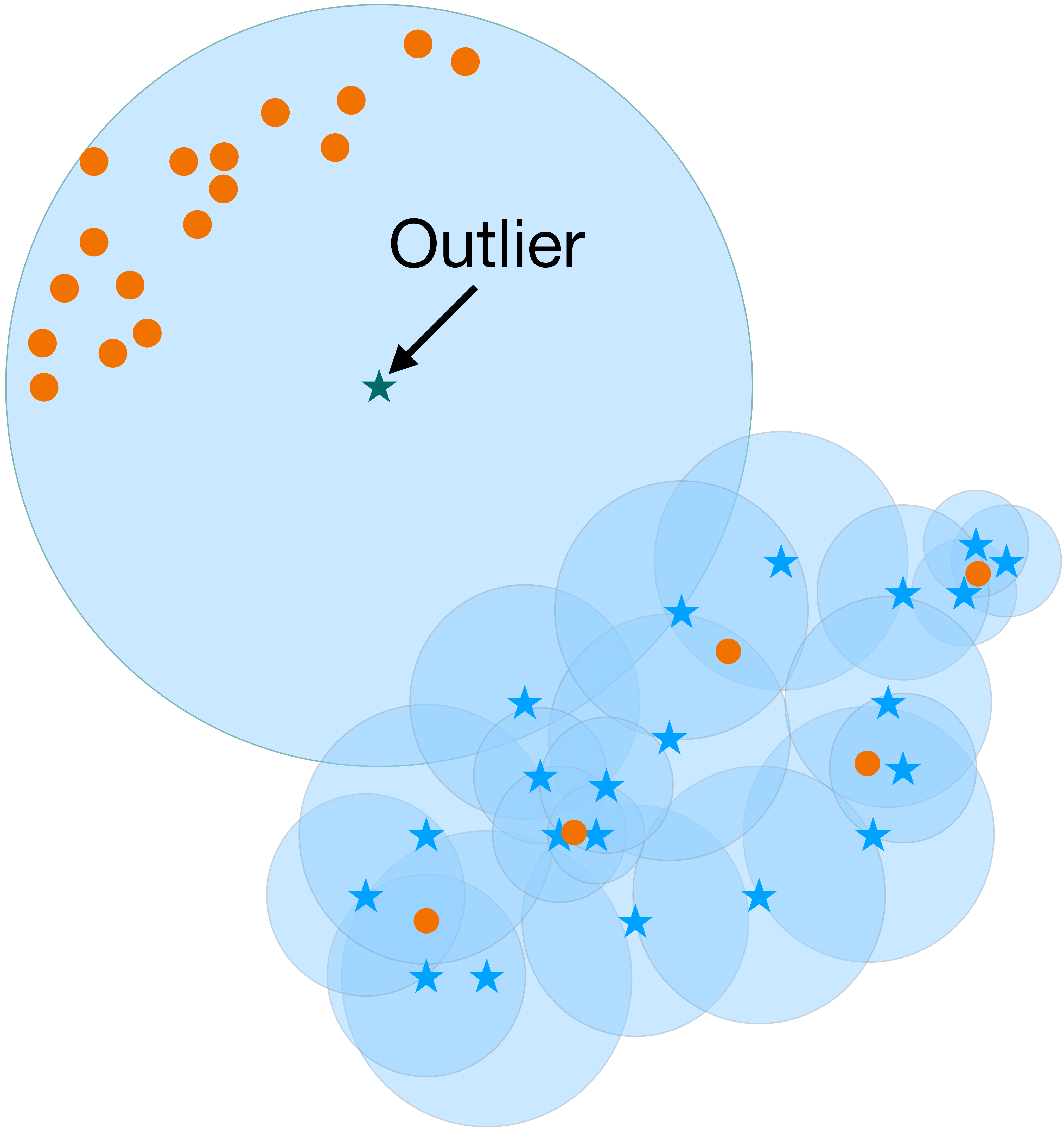
→ **Clip the contributions**

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists j, x_j^s \in B(x_i^r, NND_k^r(x_i^r))}$$

Is there a synthetic point in each real ball?

$$\text{score} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}, 1\right)$$

How many synthetic points are there in each real ball?



$$\text{score} = \frac{1}{22}\left(\frac{1}{2} \times 19 + \min\left(\frac{17}{2}, 1\right) \times 1\right)$$

$$= \frac{21}{44}$$

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Interpretability

0.6

- **Relative** results: model A is better than model B
  - Model A might be bad
  - There might be no other model to compare to

➡ Can my model be used?

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Interpretability: desideratum

? ? ? ? ? ? ? ? ? ?

? 0.6 ? ?

? ? ? ? ? ? ?

- **Relative** results: model A is better than model B
    - Model A might be bad
    - There might be no other model to compare to

➡ Can my model be used?

Absolute interpretability

Reference point:

0.6

$\Leftrightarrow$

60 % good samples

40 % bad samples

Interpretability: desideratum

? ? ? ? ? ? ? ? ? ? ? ? 0.6 ? ? ? ? ? ? ? ? ? ?

- **Relative** results: model A is better than model B
  - Model A might be bad
  - There might be no other model to compare to

➡ Can my model be used?

Absolute **interpretability**

Reference point:

$$0.6$$
$$\Leftrightarrow$$
$$60\ \%$$ good samples
$$40\ \%$$ bad samples

- linear score degradation: decrease by x if $x\ \%$ samples are ”bad” (0-score)
- normalisation between 0 and 1

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Interpretability: Clipped Density

Absolute **interpretability**

Reference point:

$$0.6$$

$$\Leftrightarrow$$

$60\ \%$ good samples

$40\ \%$ bad samples

$$\text{ClippedDensity}_{\text{unnorm}} = \boxed{\frac{1}{M}\sum_{j=1}^{M}} \min\left(\frac{1}{k}\sum_{i=1}^{N} 1_{x_j^s \in B(x_i^r, R_k(x_i^r))}, 1\right)$$

average over synthetic points

- linear score degradation: decrease by x if $x\ \%$ samples are "bad" (0-score)

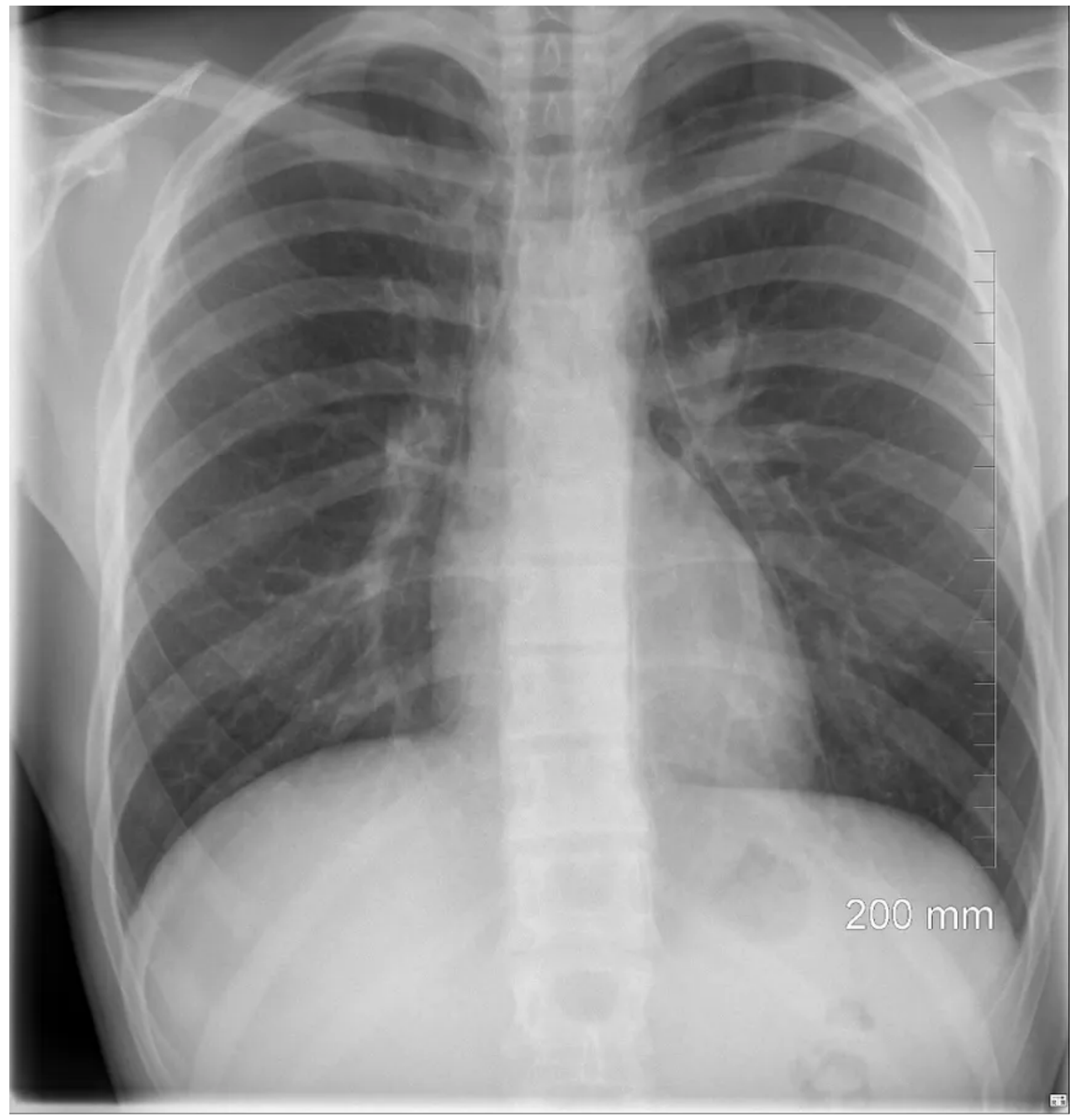
$$\text{ClippedDensity} = \frac{\text{ClippedDensity}_{\text{unnorm}}}{\text{ClippedDensity}_{\text{real}}}$$

- normalisation between 0 and 1

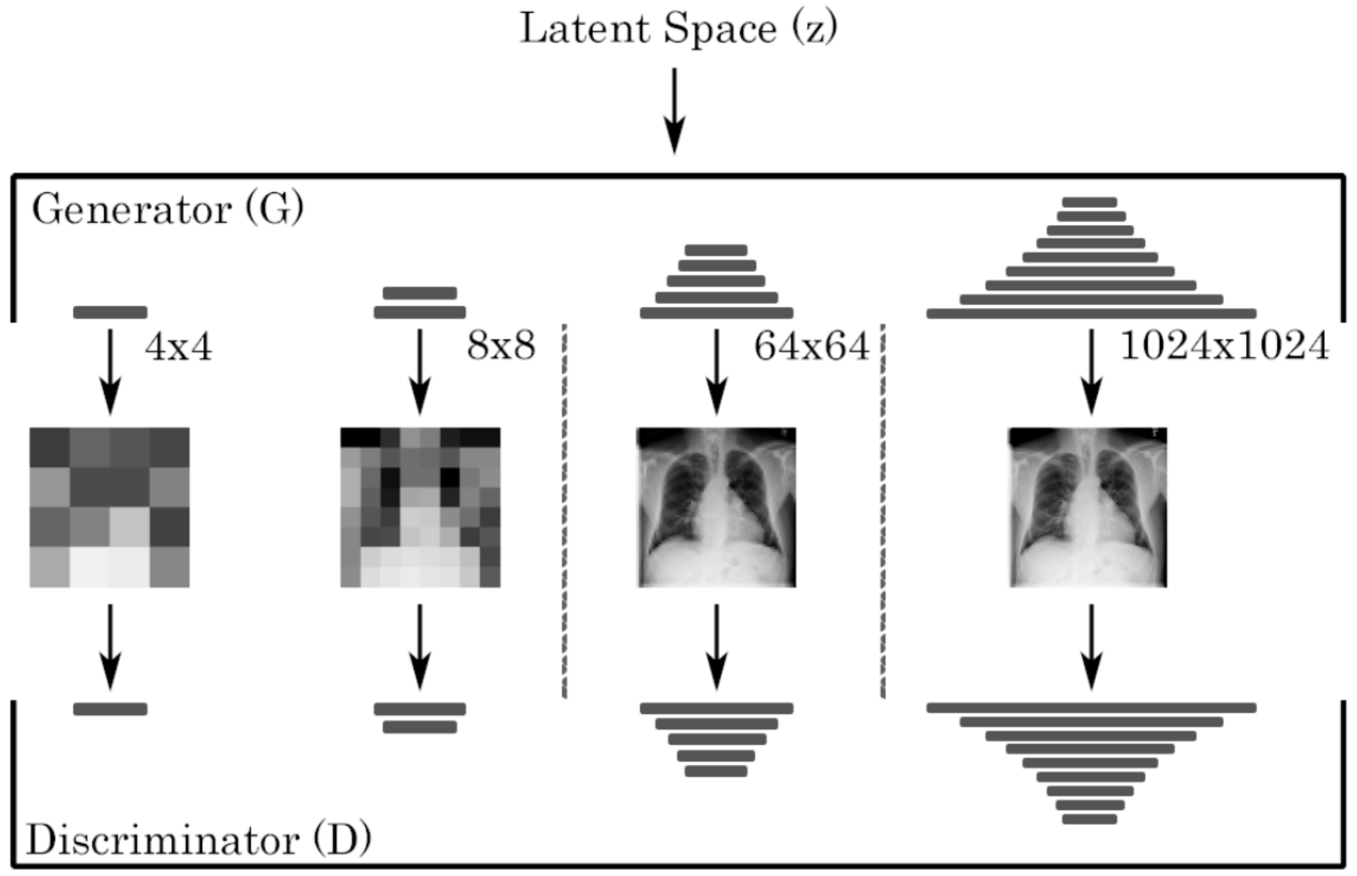
Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Absolute interpretability: Clipped Density

ChestX-ray18



Progressively Growing GAN
(Segal et al., 2021)



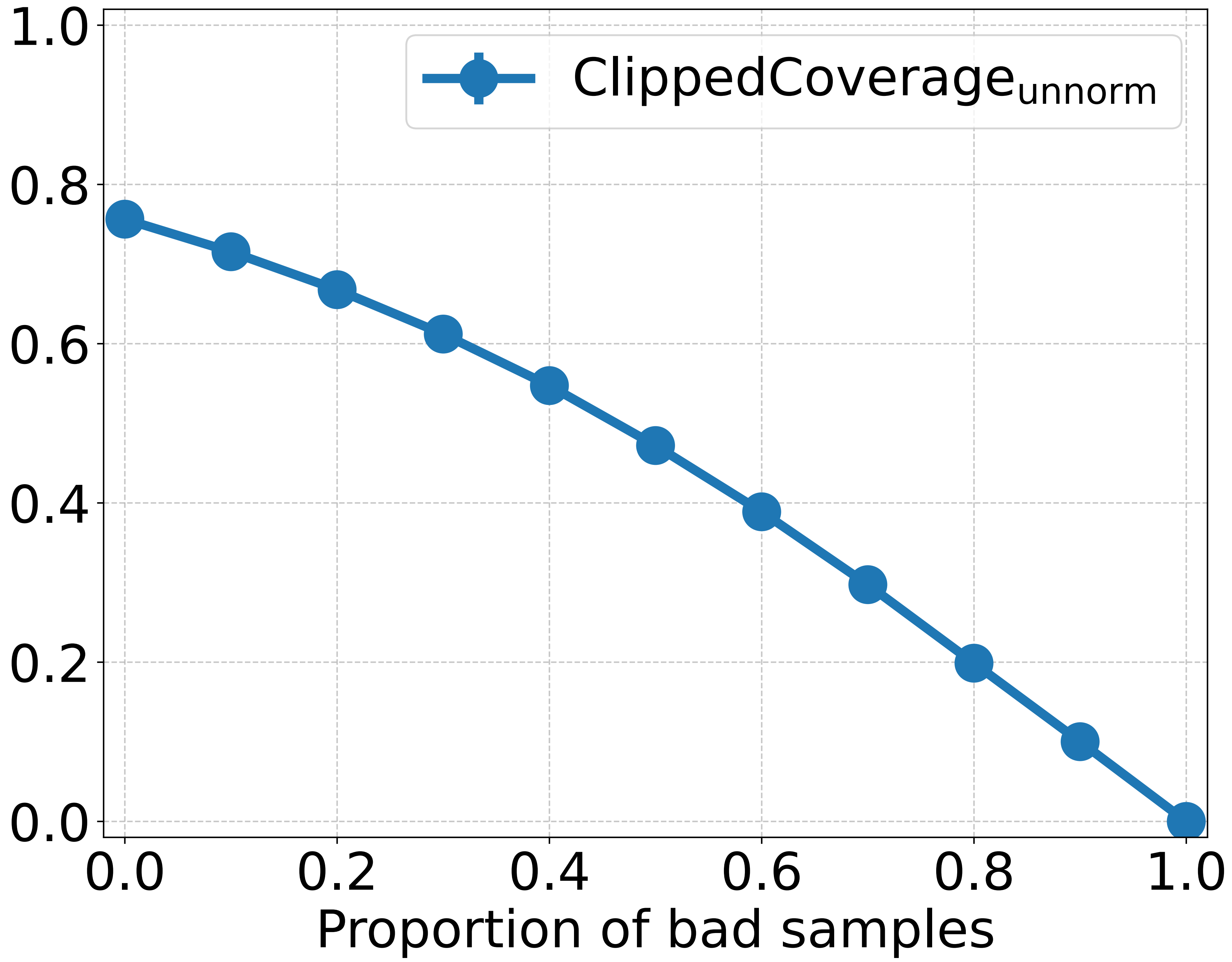
$\text{ClippedDensity} = 0.06$

Fidelity score equivalent to that of only $6\ \%$ of good samples → not useful in practice

Absolute interpretability: Coverage

- linear score degradation: decrease by x if $x\,\%$ samples are ”bad” (0-score)
- normalisation between 0 and 1

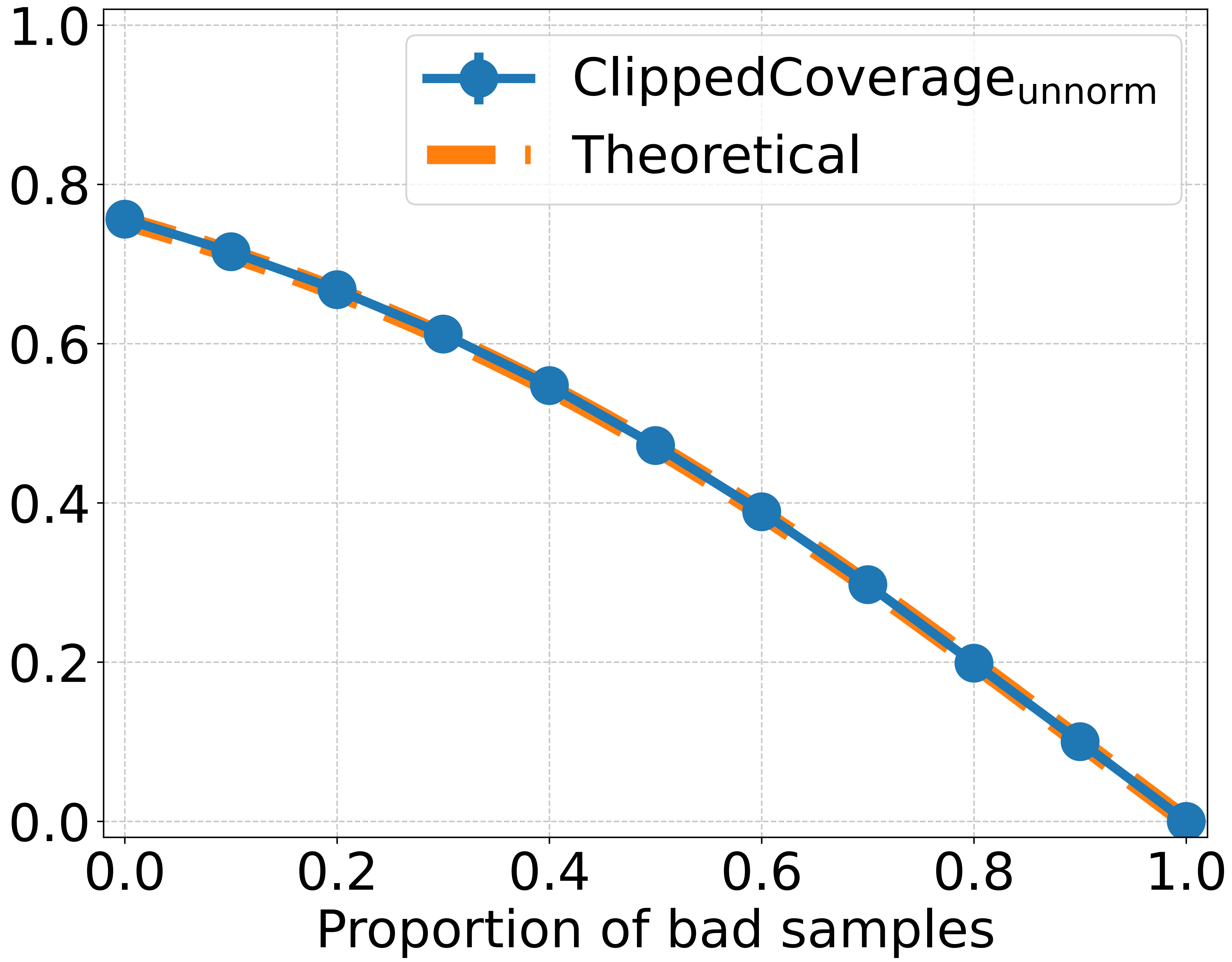
$$\text{ClippedCoverage}_{\text{unnorm}} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}, 1\right)$$



Absolute interpretability: Coverage

- linear score degradation: decrease by x if $x\,\%$ samples are ”bad” (0-score)
- normalisation between 0 and 1

$$\text{ClippedCoverage}_{\text{unnorm}} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}, 1\right)$$



$$\mathbf{E}\left[\text{ClippedCoverage}_{\text{unnorm}}\right]$$
$$= \sum_{j=1}^{M} \min\left(\frac{j}{k}, 1\right)\binom{M}{j}\frac{\beta(k+j, M-j+N-k)}{\beta(k, N-k)}$$

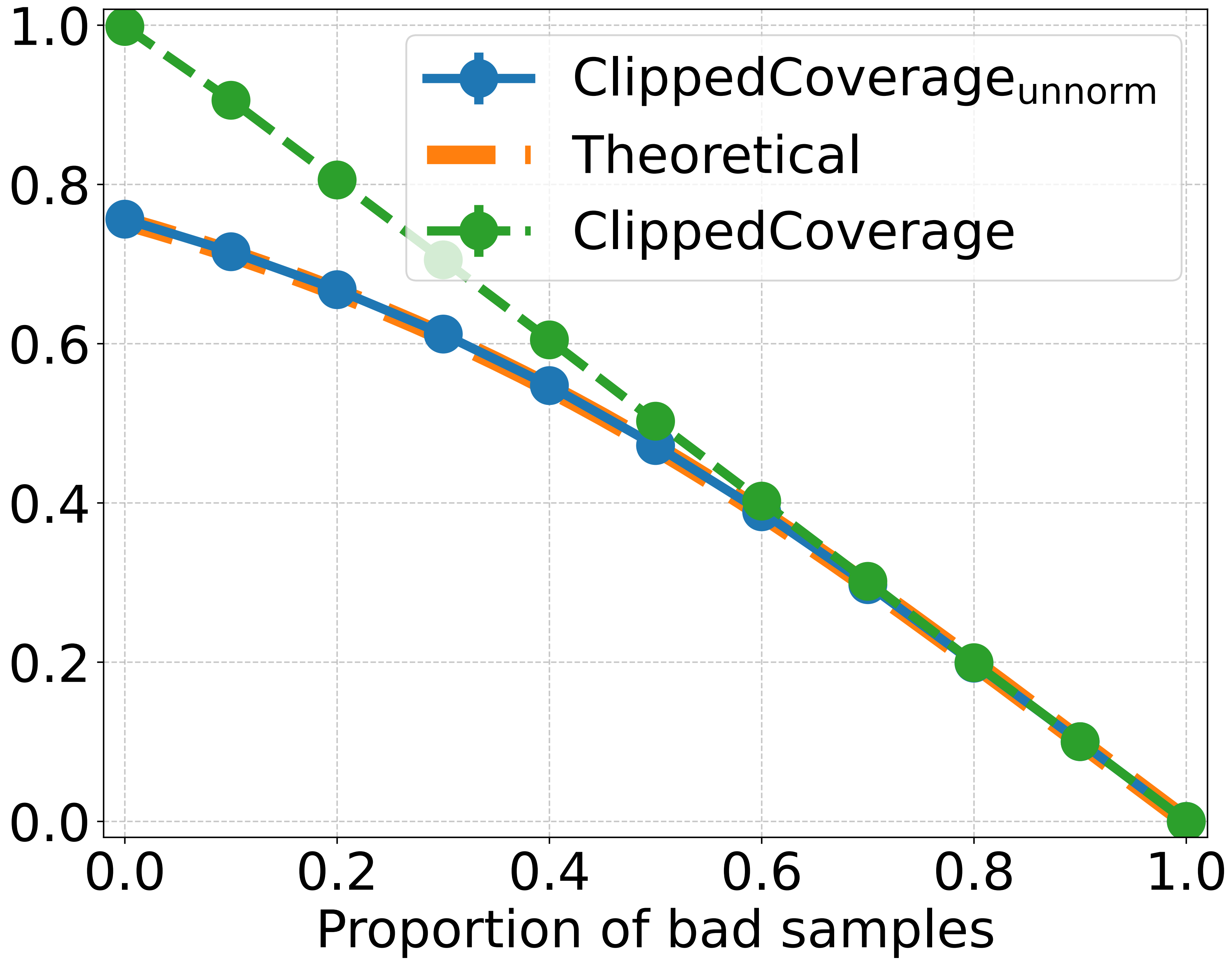
Let X and Y be N and M independent samples from p.

$\mathbb{P}\left(X_{(k)} < Y_{(l)}\right)$ is **independent** of p

Absolute interpretability: Coverage

- linear score degradation: decrease by x if $x\,\%$ samples are "bad" (0-score)
- normalisation between 0 and 1

$$\text{ClippedCoverage}_{\text{unnorm}} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{1}{k}\sum_{j=1}^{M} 1_{x_j^s \in B(x_i^r, NND_k^r(x_i^r))}, 1\right)$$



$$\mathbf{E}\left[\text{ClippedCoverage}_{\text{unnorm}}\right] = \sum_{j=1}^{M} \min\left(\frac{j}{k}, 1\right)\binom{M}{j}\frac{\beta(k+j, M-j+N-k)}{\beta(k, N-k)}$$

Let X and Y be N and M independent samples from p.

$\mathbb{P}\left(X_{(k)} < Y_{(l)}\right)$ is **independent** of p

$$\text{ClippedCoverage} = g \circ \text{ClippedCoverage}_{\text{unnorm}}$$

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Testing coverage metrics

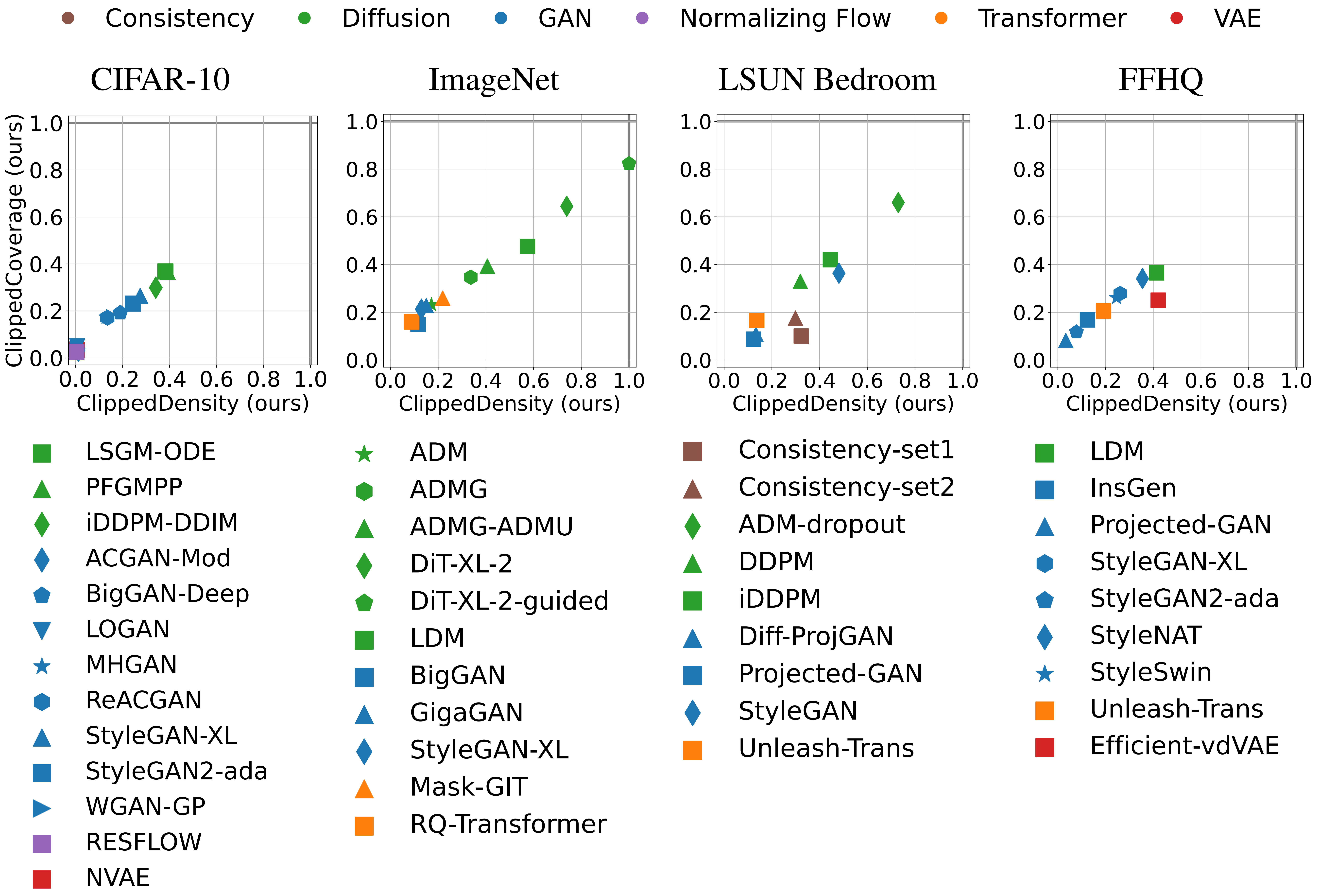
	Linear degradation	Stability	Robust to outliers
Recall	✗	✓	✗
Coverage	✗	✓	✓
symRecall	✗	✓	✗
β-Recall	✗	✗	✗
TopR	✗	✗	✓
P-recall	✗	✗	✗
RecallCover	✗	✓	✓
ClippedCoverage (ours)	✓	✓	✓

Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025

Testing fidelity metrics

	Linear degradation	Measures fidelity	Stability	Robust to outliers
Precision	✓	✓	✓	✗
Density	✓	✓	✓	✗
symPrecision	✓	✗	✓	✗
α-Precision	✓	✗	✗	✗
TopP	✗	✓	✗	✓
P-precision	✓	✓	✗	✗
PrecisionCover	✓	✗	✓	✓
ClippedDensity (ours)	✓	✓	✓	✓

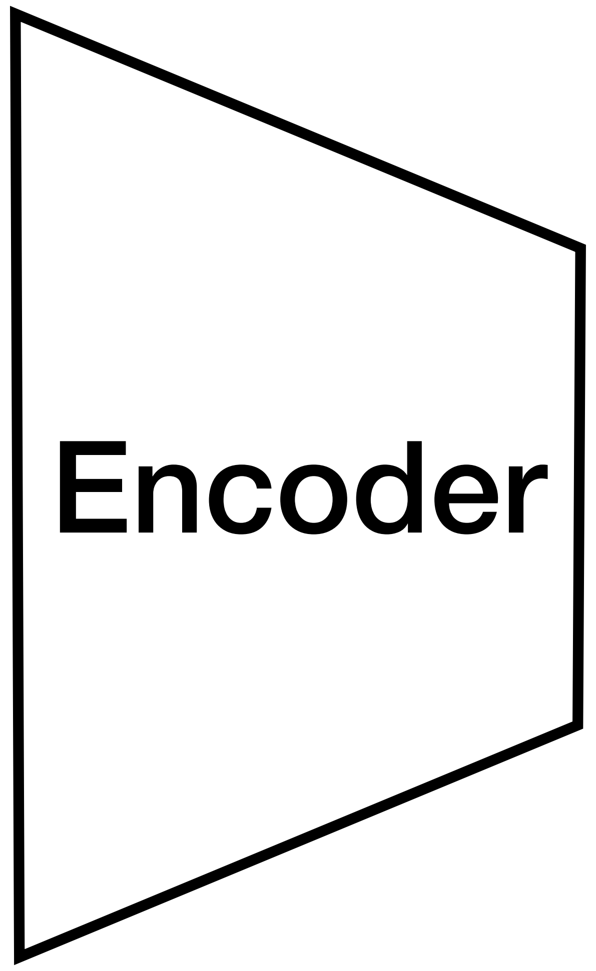
Enhanced Generative Model Evaluation with Clipped Density and Coverage, N. Salvy, H. Talbot, B. Thirion, preprint, 2025



Consistency
Diffusion
GAN
Normalizing Flow
Transformer
VAE
CIFAR-10
ImageNet
LSUN Bedroom
FFHQ
ClippedCoverage (ours)
ClippedDensity (ours)
0.0
0.2
0.4
0.6
0.8
1.0
LSGM-ODE
PFGMPP
iDDPM-DDIM
ACGAN-Mod
BigGAN-Deep
LOGAN
MHGAN
ReACGAN
StyleGAN-XL
StyleGAN2-ada
WGAN-GP
RESFLOW
NVAE
ADM
ADMG
ADMG-ADMU
DiT-XL-2
DiT-XL-2-guided
LDM
BigGAN
GigaGAN
StyleGAN-XL
Mask-GIT
RQ-Transformer
Consistency-set1
Consistency-set2
ADM-dropout
DDPM
iDDPM
Diff-ProjGAN
Projected-GAN
StyleGAN
Unleash-Trans
LDM
InsGen
Projected-GAN
StyleGAN-XL
StyleGAN2-ada
StyleNAT
StyleSwin
Unleash-Trans
Efficient-vdVAE

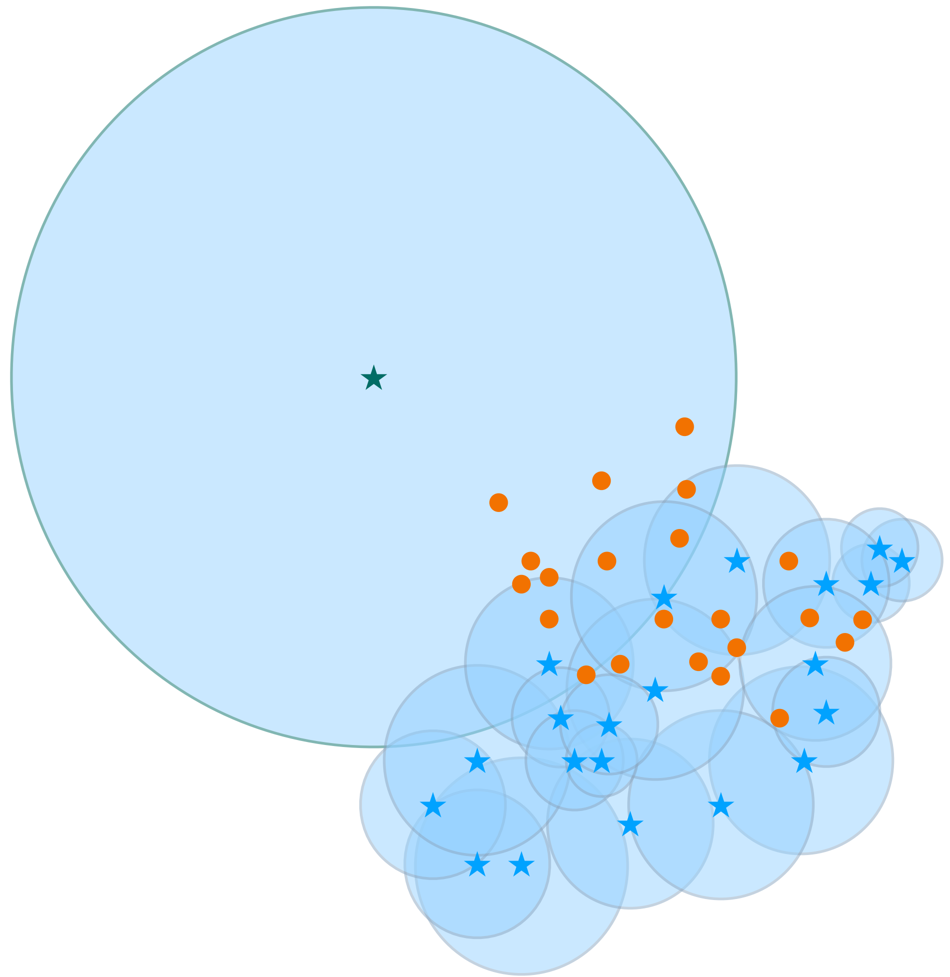
Real samples

Synthetic samples



”distances make sense”

Fidelity & **Coverage** **metrics**



Rely on distances (usually)

➡ two values

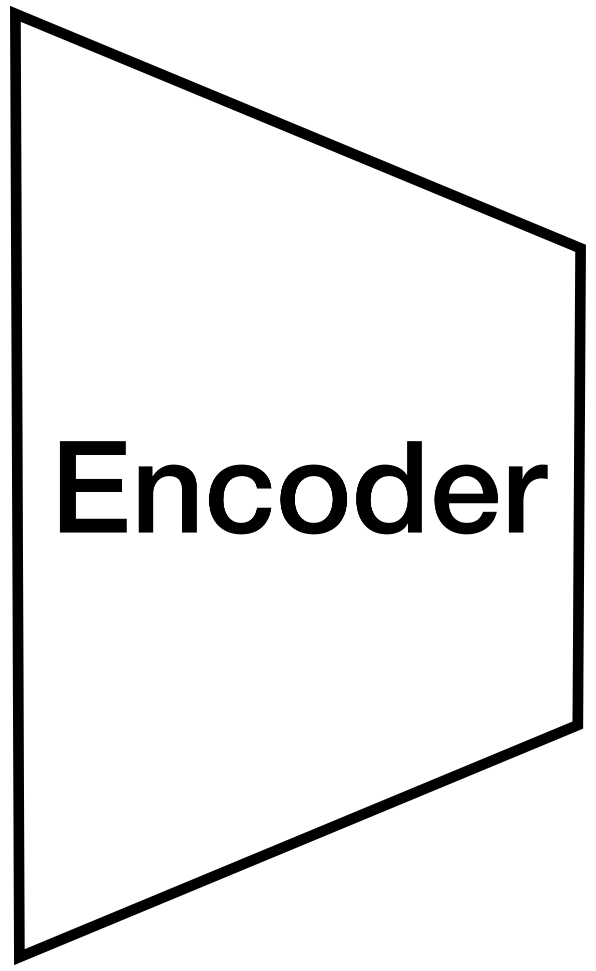
- lack **robustness**
- lack **interpretability**

Clipped Density

Clipped Coverage

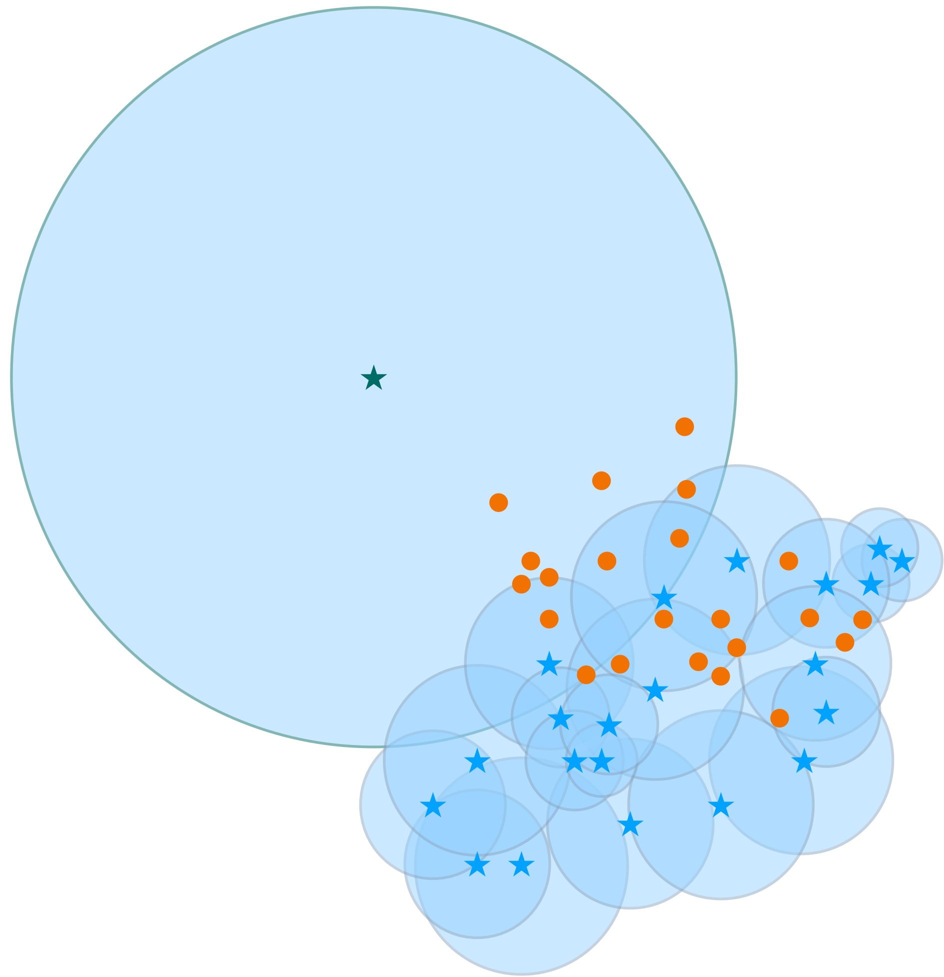
Real samples

Synthetic samples



”distances make sense”

Fidelity & Coverage metrics

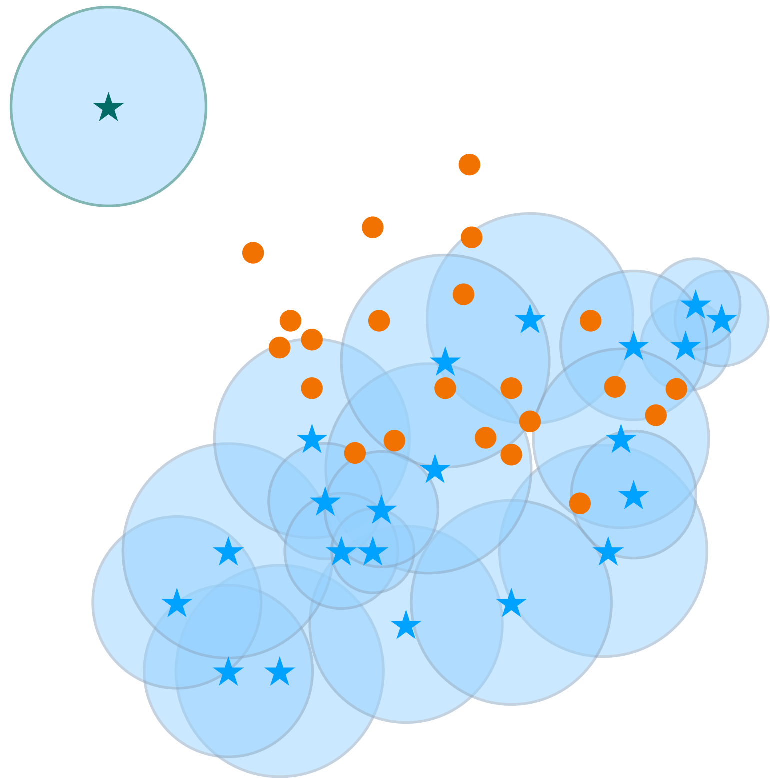


Rely on distances (usually)

➡ two values

- lack **robustness**
- lack **interpretability**

Clipped Density



→ **Clip** the radii

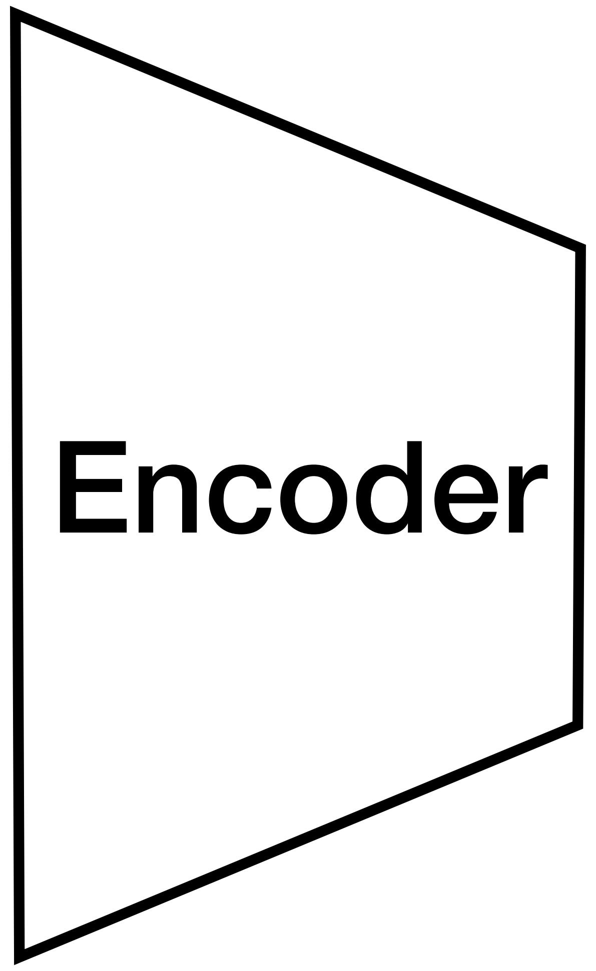
→ **Clip** the contributions

Clipped Coverage

→ **Clip** the contributions

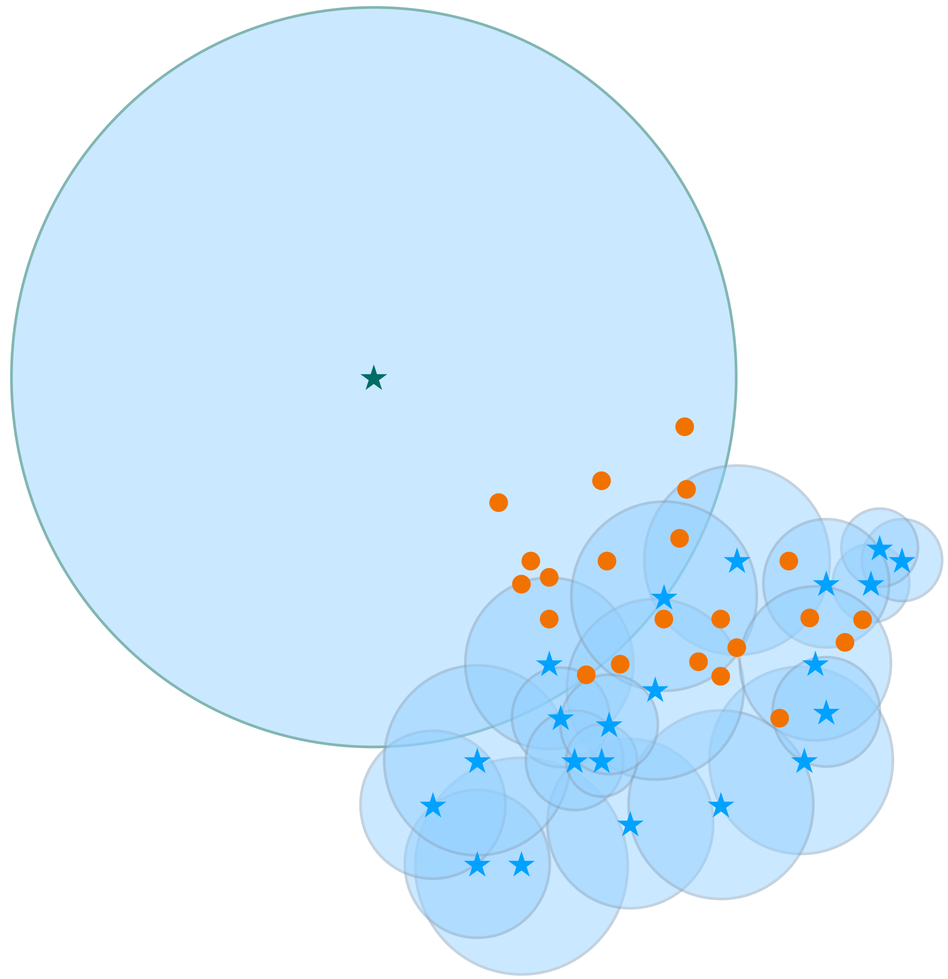
Real samples

Synthetic samples



”distances make sense”

Fidelity & **Coverage** **metrics**

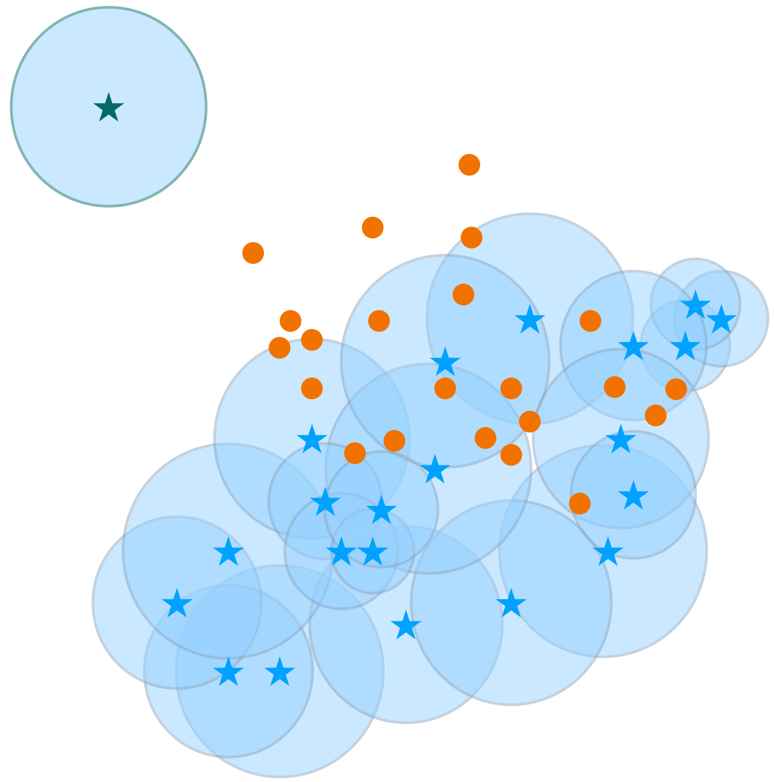


Rely on distances (usually)

➡ two values

- lack **robustness**
- lack **interpretability**

Clipped Density

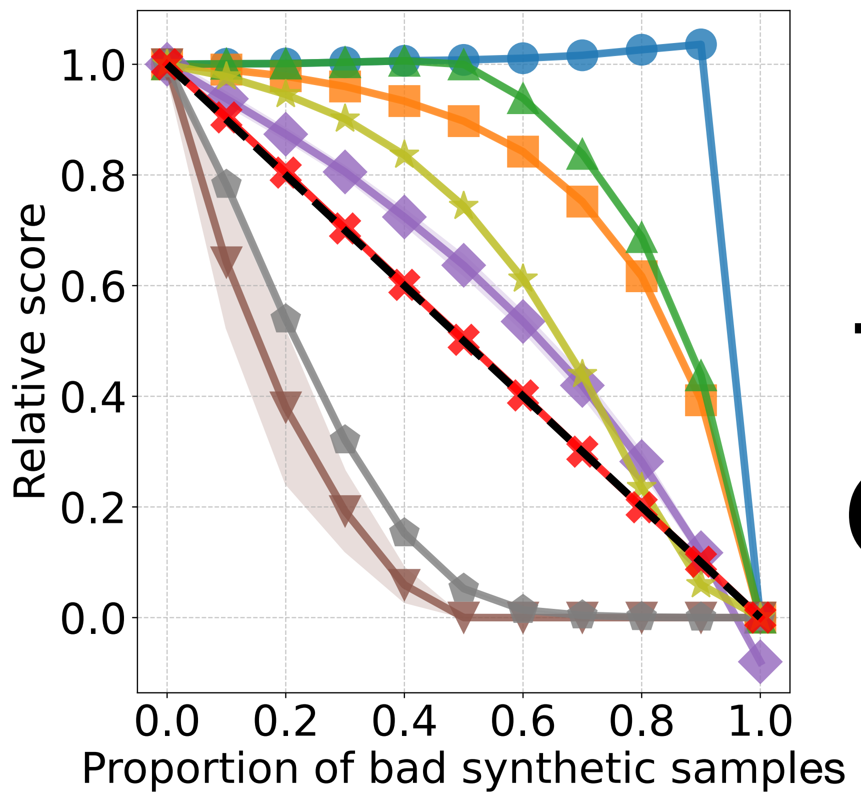


→ **Clip** the radii

→ **Clip** the contributions

Normalise empirically by $\text{ClippedDensity}_{\text{real}}$

Clipped Coverage

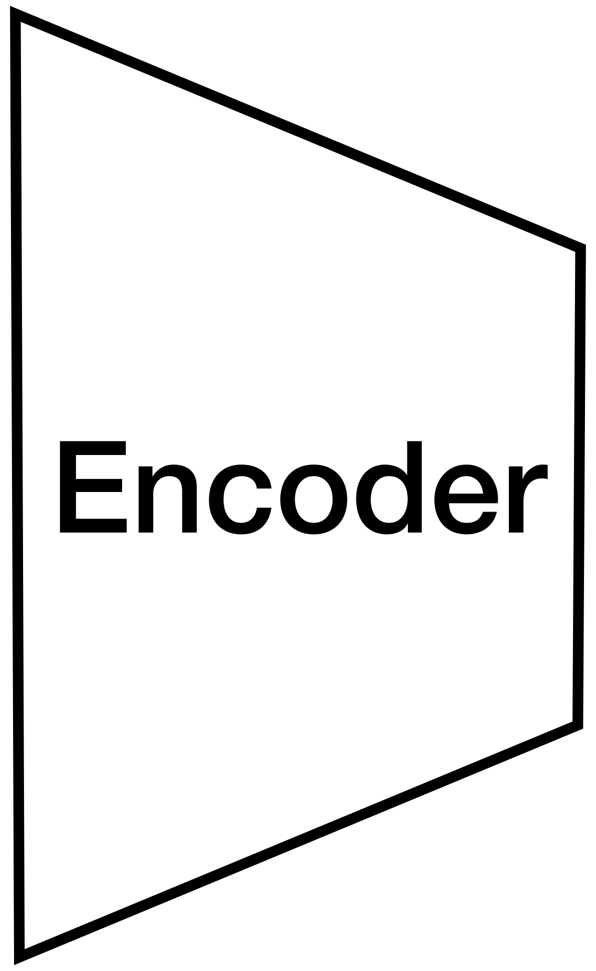


→ **Clip** the contributions

Calibrate theoretically

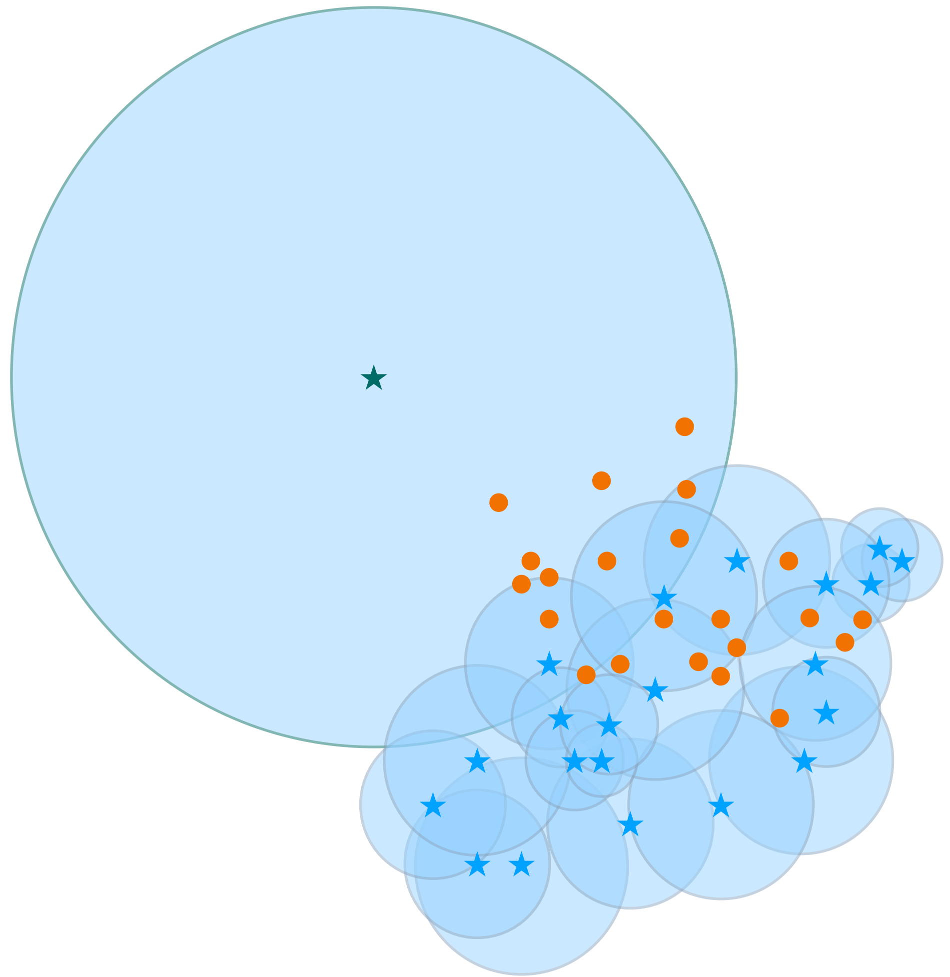
Real samples

Synthetic samples



”distances make sense”

Fidelity & Coverage metrics

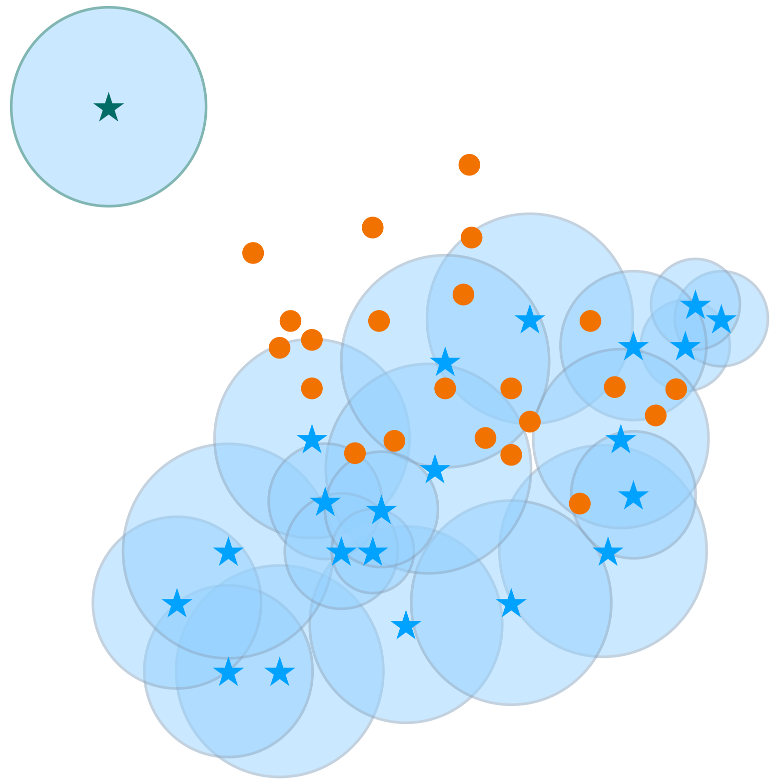


Rely on distances (usually)

➡ two values

- lack **robustness**
- lack **interpretability**

Clipped Density

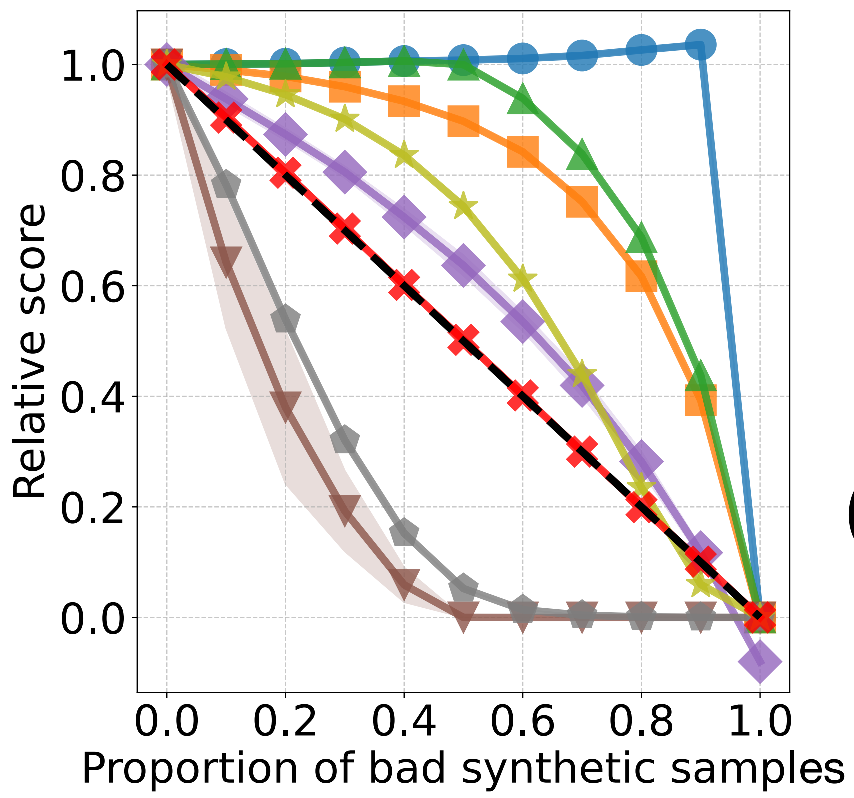


→ **Clip** the radii

→ **Clip** the contributions

Normalise empirically by $\text{ClippedDensity}_{\text{real}}$

Clipped Coverage



→ **Clip** the contributions

Calibrate theoretically

Use Clipped Density and Clipped Coverage!