Probabilistic Machine Learning

+

Deep Learning

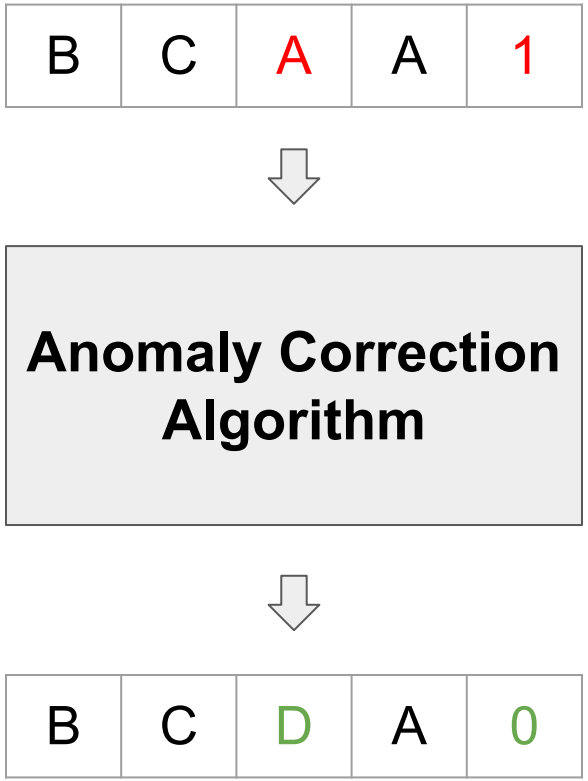
CATEGORICAL ANOMALY CORRECTION BASED ON INVERSE GRADIENT OF CLASSIFIER AND DIFFUSION INPAINTING

Omar Cusma Fait
Luis Fernando Palacios Flores

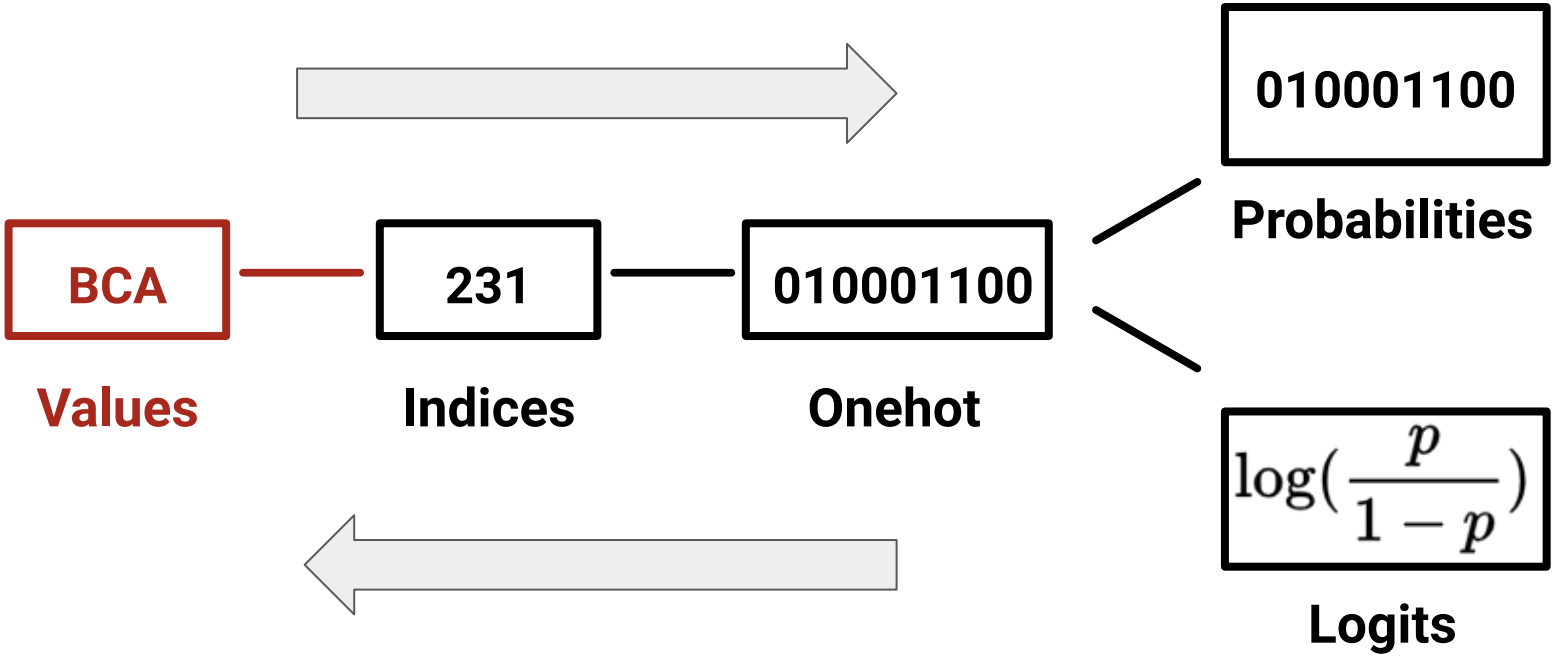
Problem

Data

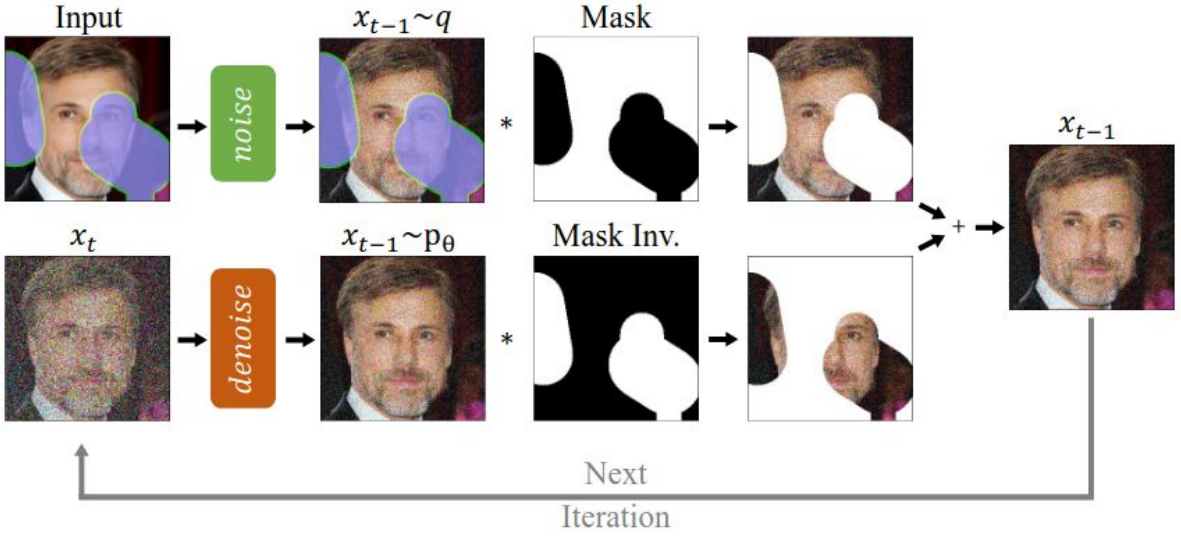
C1	C2	C3	C4	y
A	C	B	D	0
B	C	A	D	0
A	E	C	B	0
B	C	A	A	1



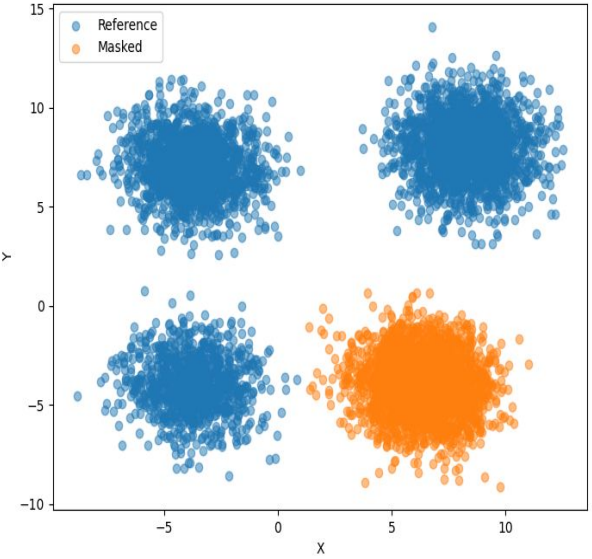
Data Domain

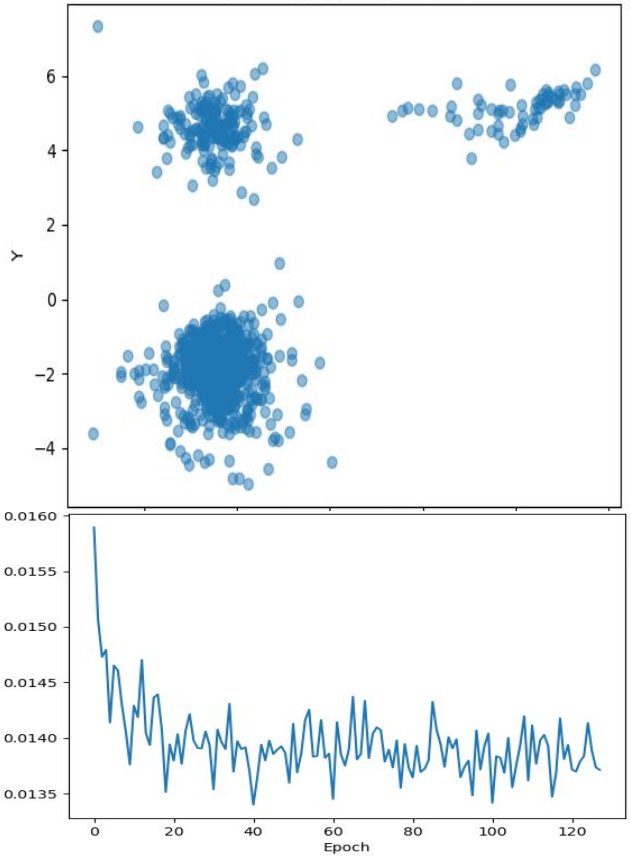

010001100
Probabilities
BCA
231
010001100
Values
Indices
Onehot
$\log(\frac{p}{1-p})$
Logits

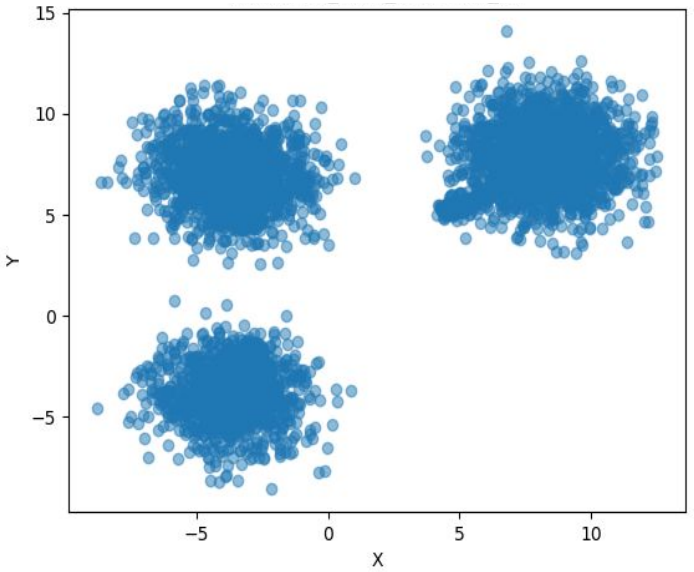
Inpainting with Denoising Diffusion Probabilistic Models



Toy Example with DDPM: Gaussian Data



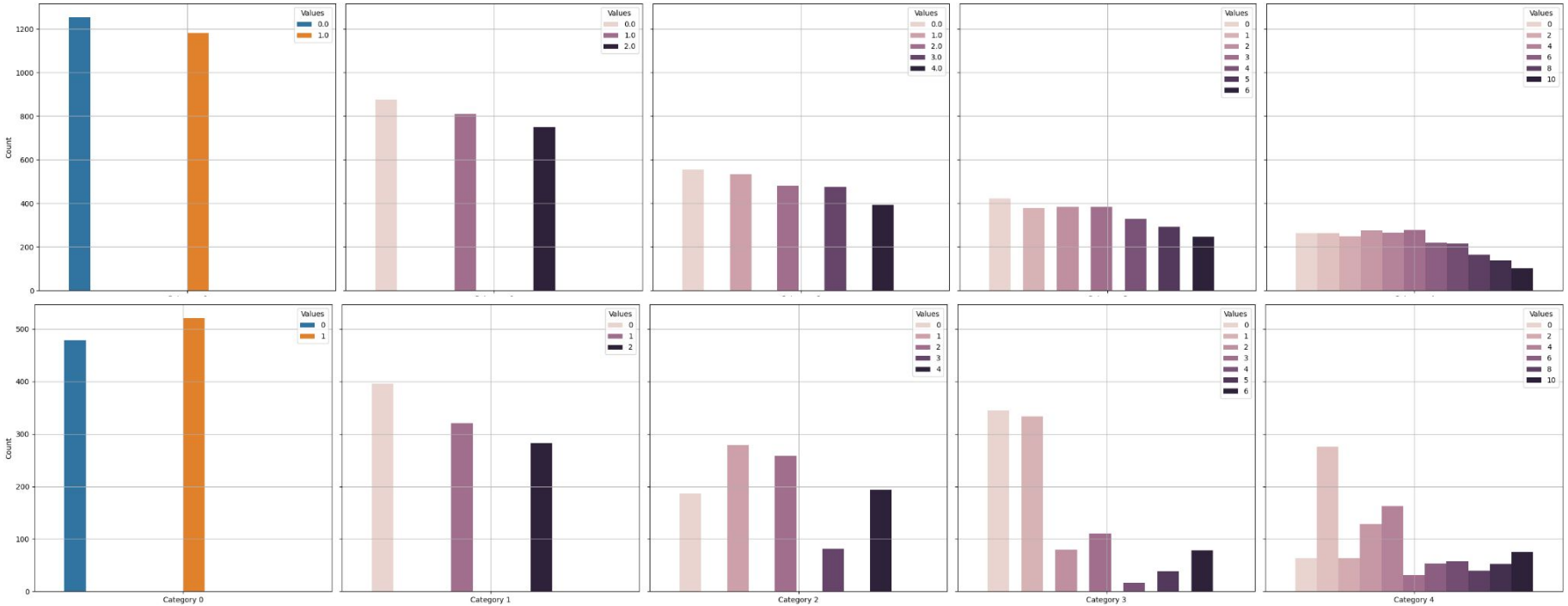





```
hidden_units =
[256, 512, 256]
```

Toy Example with DDPM

$$\sum_{j=1}^{K} \text{index}_j < \text{threshold}$$




```
concat_x_and_t = True
hidden_units = [92]
```

Dissimilarity: 73.34%

Anomalies Generated: 4.00%

Toy Example with DDPM

$$\sum_{j=1}^{K} \text{index}_j < \text{threshold}$$

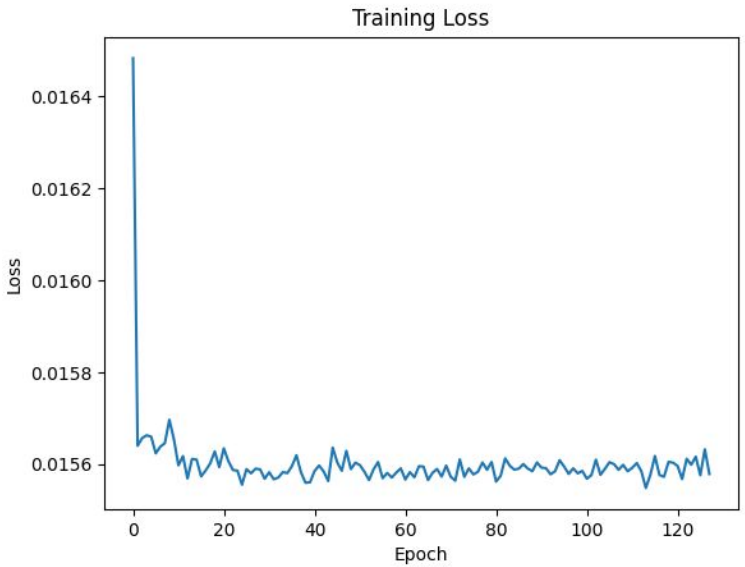
Metric	Entire Dataset	Anomalies feature mask	Anomalies all masked
Anomalies before inpainting / Total	268 / 1500	254 / 254	905 / 905
Remaining anomalies / Total	64 / 1500	72 / 254	124 / 905
Correct changes	204	182	781
Wrong changes	0	0	0
Percentage of anomalies before inpainting	17.87%	100.00%	100.00%
Percentage of anomalies after inpainting	4.27%	28.35%	13.70%
Percentage change (-100% desired)	**-76.12%**	**-71.65%**	**-86.30%**
Percentage of classified anomalies after inpainting	4.20%	26.77%	13.70%
Number of rows wrongly modified	0 (1500)	0 (254)	N/A
Number of known values wrongly modified	0 (7232)	0 (1016)	N/A
Dissimilarity original and the inpainted distribution	15.20%	**85.04%**	**87.85%**
Mean values agree per instance (features)	N/A	N/A	**1.21 (5)**
Median values agree per instance	N/A	N/A	**1.00**
Standard deviation values agree per instance	N/A	N/A	0.87

~4

Diffusion Inpainting Example: Bank Dataset

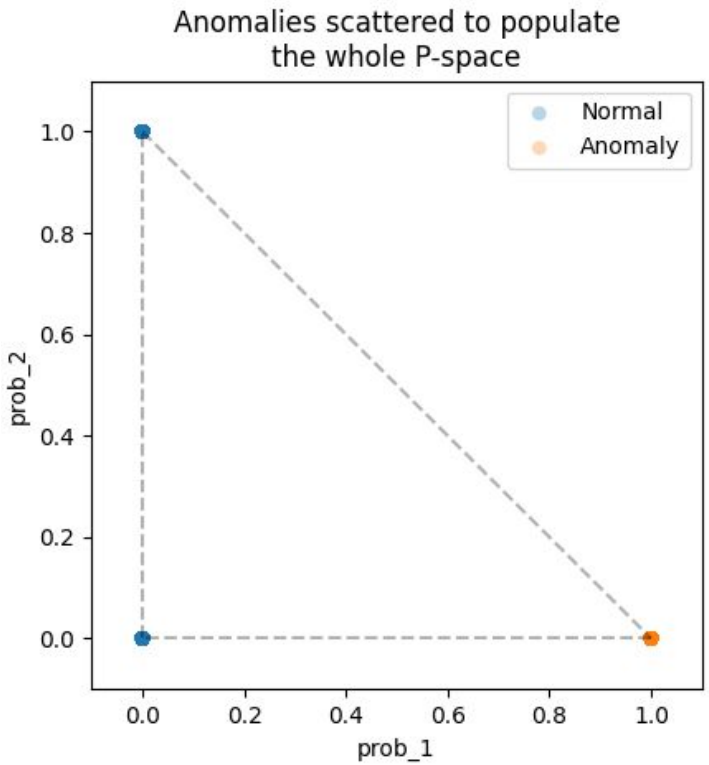
Dissimilarity original and sampled data: 99.95%
% anomalies generated: 7.60%

Dummy accuracy = 88.7%
Accuracy = 89.6%
Usefulness = 7.7%



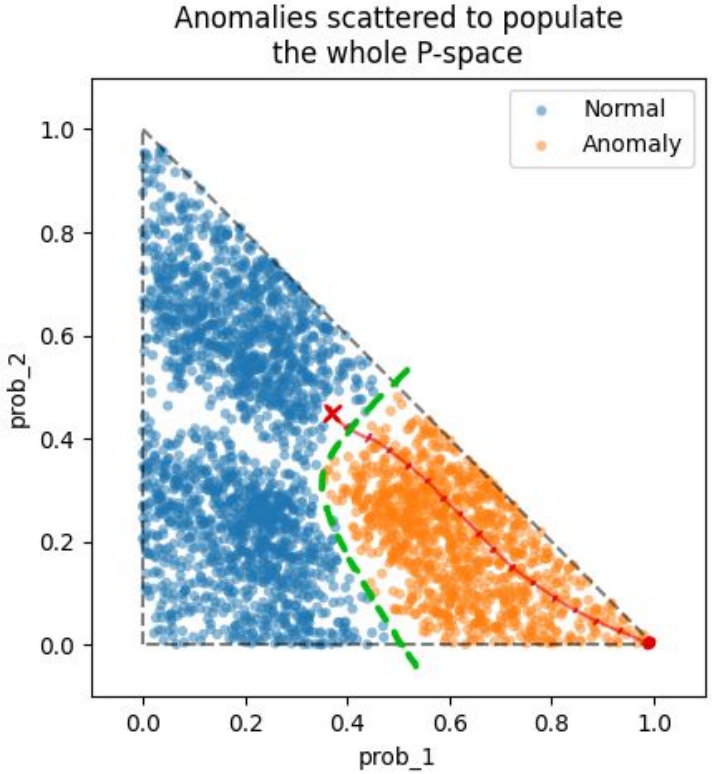
Metric	Value
Anomalies before inpainting / Total	4640 / 4640
Remaining anomalies / Total	300 / 4640
Percentage of anomalies before inpainting	100.00%
Percentage change (-100% desired)	**-93.53%**
Percentage of classified anomalies after inpainting	**6.47%**
Mean values agree per instance (columns)	**2.17 (10)**
Median values agree per instance	2.00
Standard deviation values agree per instance	1.33
Dissimilarity original and inpainted data	**99.87%**

The Inverse-Gradient Method

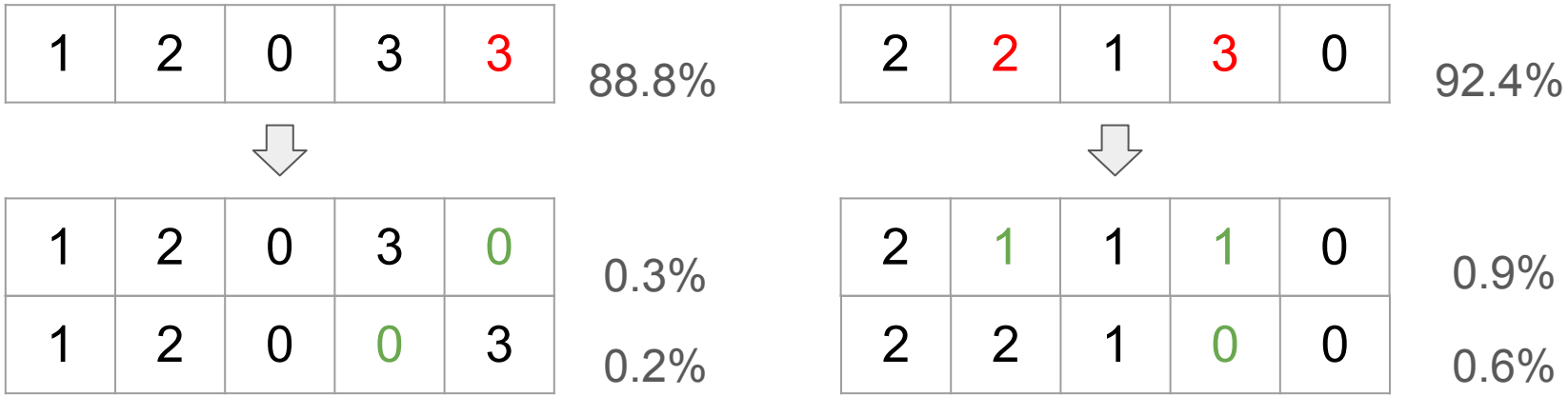


Populating the entire probability space

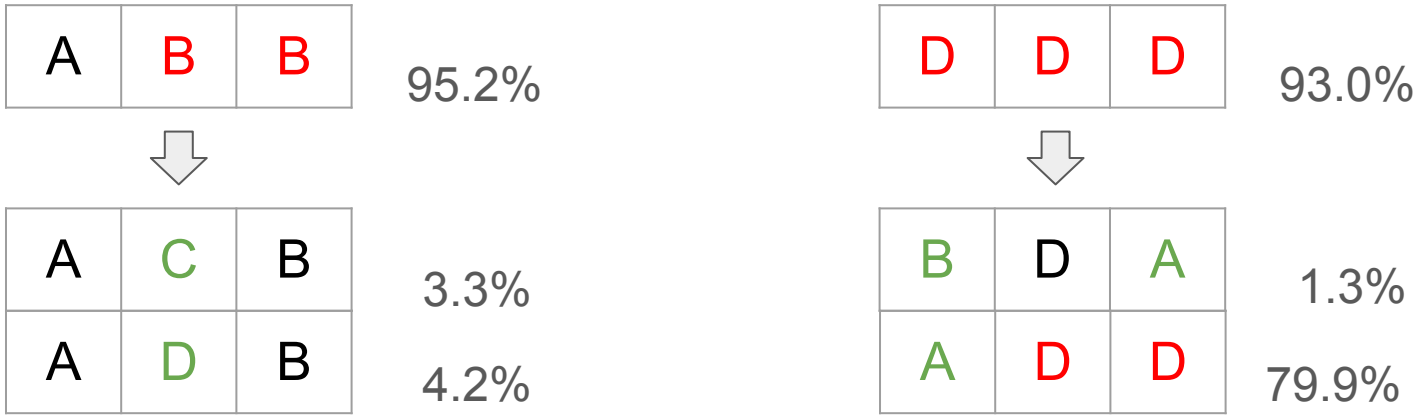
noise*



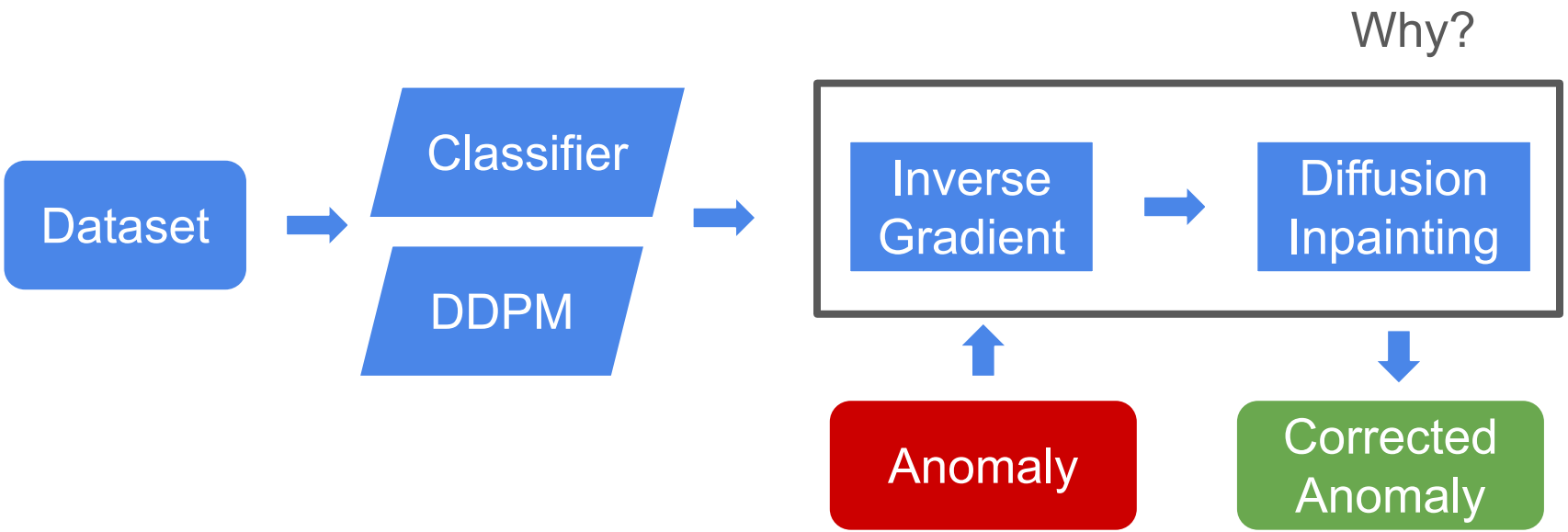
Toy Example with Inverse Gradient: Sum of Digits



Toy Example with Inverse Gradient: No Duplicates



Anomaly Correction Pipeline



upsides and downsides…

Results: Sum Limit Problem

Anomaly:

	x_0	x_1	x_2	x_3	x_4
4840	1	1	3	3	3

```
Iteration 116) loss is 0.1% < 10.0%
Iteration 126) loss is 0.0% < 10.0%
Iteration 129) loss is 0.0% < 10.0%
Iteration 122) loss is 0.0% < 10.0%
Iteration 119) loss is 0.1% < 10.0%
Iteration 120) loss is 0.0% < 10.0%
Iteration 126) loss is 0.0% < 10.0%
Iteration 124) loss is 0.1% < 10.0%
Iteration 119) loss is 0.0% < 10.0%
Iteration 118) loss is 0.1% < 10.0%
```

masks

```
[False False False  True False]
[False False  True False False]
[False False False  True False]
[False False False  True False]
[False False False  True False]
[False False False  True False]
[False False  True  True False]
[False False False  True False]
[False False  True False False]
[False False  True False False]
```

Corrected anomaly:

	x_0	x_1	x_2	x_3	x_4
0	1	1	3	0	3
1	1	1	2	3	3
2	1	1	3	2	3
3	1	1	3	0	3
4	1	1	3	3	3
5	1	1	3	2	3
6	1	1	2	2	3
7	1	1	3	1	3
8	1	1	1	3	3
9	1	1	1	3	3

Results: No Duplicates Problem

```
Anomaly:
      x_0 x_1 x_2
24436   C   C   B
Iteration 142) loss is 0.1% < 5.0%
Iteration 146) loss is 0.1% < 5.0%
Iteration 204) loss is 0.1% < 5.0%
Iteration 119) loss is 0.4% < 5.0%
Iteration 166) loss is 0.1% < 5.0%
Iteration 141) loss is 0.1% < 5.0%
Iteration 157) loss is 0.1% < 5.0%
Iteration 129) loss is 0.2% < 5.0%
Iteration 194) loss is 0.1% < 5.0%
Iteration 132) loss is 0.3% < 5.0%
```

```
masks
[False  True False]
[False  True False]
[ True False False]
[False  True False]
[ True False False]
[False  True False]
[False  True False]
[False  True False]
[False  True False]
[False  True False]
```

```
Corrected anomaly:
  x_0 x_1 x_2   p_anomaly
0   C   A   B   0.0%
1   C   A   B   0.0%
2   D   C   B   0.0%
3   C   A   B   0.0%
4   D   C   B   0.0%
5   C   A   B   0.0%
6   C   A   B   0.0%
7   C   A   B   0.0%
8   C   A   B   0.0%
9   C   A   B   0.0%
```