

UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR IT SECURITY

Privacy-Preserving Data Synthesis

Yara Schütt

AnoMed II Seminar

09.04.2026

Overview

- Motivation
  - Why privacy-preserving?
  - How can we achieve privacy?
- Privacy-preserving data synthesis: DPScans
- Next steps
- Conclusion

Motivation

- Why privacy-preserving?
- How can we achieve privacy?

Motivation – Why privacy-preserving?

- Data that contains sensitive information
  - e.g., medical records of patients in a hospital (Project AnoMed)
- Learn something from the data without learning something about individuals
  - e.g., statistics about patients (reason for admission, diagnosis, …) ✓
    ‘Alice was admitted for abdominal pain and was diagnosed with cancer’ ✗

Motivation – Why privacy-preserving?

- Data that contains sensitive information
  - e.g., medical records of patients in a hospital (Project AnoMed)
- Learn something from the data without learning something about individuals
  - e.g., statistics about patients (reason for admission, diagnosis, …) ✓
    ‘Alice was admitted for abdominal pain and was diagnosed with cancer’ ✗
- How can we achieve privacy?
  - Remove identifiers like names
  - Only learn statistics
  - Synthetic data (no real data)

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Remove identifiers like names
  - Only learn statistics
  - Synthetic data (no real data)

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Remove identifiers like names, generalize attributes (Pseudonymisation)

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Remove identifiers like names, generalize attributes (Pseudonymisation)

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered

↓ Replace name with anonymised ID
Generalise age to range of 5 years

ID	Age	Sex	Admission reason	Diagnosis	Outcome
1	100-105	w	Abdominal pain	Pancreatic cancer	Died
2	25-30	m	Abdominal pain	Food poisoning	Recovered

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Remove identifiers like names, generalize attributes (Pseudonymisation) ✗

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered

↓ Replace name with anonymised ID
Generalise age to range of 5 years

ID	Age	Sex	Admission reason	Diagnosis	Outcome
1	100-105	w	Abdominal pain	Pancreatic cancer	Died
2	25-30	m	Abdominal pain	Food poisoning	Recovered

  - If we know either age range or sex, we still know diagnosis and outcome

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Only learn statistics
  - E.g. statistics about patients that were admitted for abdominal pain

Age	Sex	Diagnosis	Count
100-105	w	Pancreatic cancer	1
25-30	m	Food poisoning	20
…	…	…	…

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Only learn statistics
  - E.g. statistics about patients that were admitted for abdominal pain

Age	Sex	Diagnosis	Count
100-105	w	Pancreatic cancer	1
25-30	m	Food poisoning	20
…	…	…	…

- Only one female in dataset aged 100-105
- Attacker knows Alice 102 was admitted for abdominal pain
  → Alice has pancreatic cancer

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Only learn statistics
  - E.g. statistics about patients that were admitted for abdominal pain ✗

Age	Sex	Diagnosis	Count
100-105	w	Pancreatic cancer	1
25-30	m	Food poisoning	20
…	…	…	…

- Only one female in dataset aged 100-105
- Attacker knows Alice 102 was admitted for abdominal pain
  → Claire has pancreatic cancer

  - Statistics can also leak information about individuals!

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Synthetic data (no real data)

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Synthetic data (no real data)

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…



Name	Age	Sex	Admission reason	Diagnosis	Outcome
Lea	101	w	Abdominal pain	Pancreatic cancer	Died
Ross	25	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Synthetic data (no real data)

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

↓

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Lea	101	w	Abdominal pain	Pancreatic cancer	Died
Ross	25	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

Motivation – Why privacy-preserving?

- How can we achieve privacy?
  - Synthetic data (no real data) ✗

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Alice	102	w	Abdominal pain	Pancreatic cancer	Died
Bob	28	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

↓

Name	Age	Sex	Admission reason	Diagnosis	Outcome
Lea	101	w	Abdominal pain	Pancreatic cancer	Died
Ross	25	m	Abdominal pain	Food poisoning	Recovered
…	…	…	…	…	…

- Name and age not exact match
- Knowing Alice, attacker can still be quite certain about her diagnosis

→ Synthetic data can still leak information about real individuals!

Motivation – How can we achieve privacy?

- How can we achieve privacy?
  - Remove identifiers like names ✗
  - Only learn statistics ✗
  - Synthetic data (no real data) ✗

Motivation – How can we achieve privacy?

Differential Privacy (DP):

Formal Definition: A randomized mechanism $M: \mathcal{D} \rightarrow \mathcal{R}$ is $(\varepsilon, \delta(\varepsilon))$-DP if for all datasets $D_0, D_1 \in \mathcal{D}$ differing by one data record and all subsets of outputs $\mathcal{O} \in \mathcal{R}$, $\Pr[M(D_0) \in \mathcal{O}] \leq e^{\varepsilon} \Pr[M(D_1) \in \mathcal{O}] + \delta(\varepsilon)$

Motivation – How can we achieve privacy?

Differential Privacy (DP):

Formal Definition: A randomized mechanism $M: \mathcal{D} \rightarrow \mathcal{R}$ is $(\varepsilon, \delta(\varepsilon))$-DP if for all datasets $D_0, D_1 \in \mathcal{D}$ differing by one data record and all subsets of outputs $\mathcal{O} \in \mathcal{R}$, $\Pr[M(D_0) \in \mathcal{O}] \leq e^{\varepsilon} \Pr[M(D_1) \in \mathcal{O}] + \delta(\varepsilon)$

Intuition:

- Impact of individuals on the outcome must be bounded
- Privacy budget $\varepsilon, \delta(\varepsilon)$: higher values mean less privacy guarantees
- Privacy-utility tradeoff: adding uncertainty results in better privacy guarantees but less accuracy

DPScans

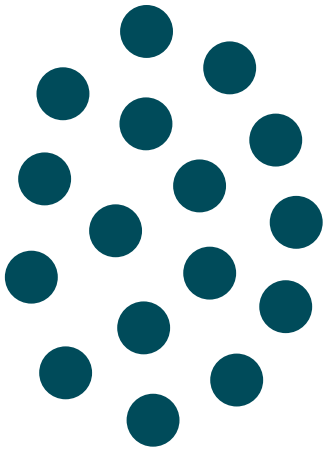
- Motivation
- Algorithm (+ DP)
- Experimental Results

DPScans – Privacy-Preserving Data Synthesis

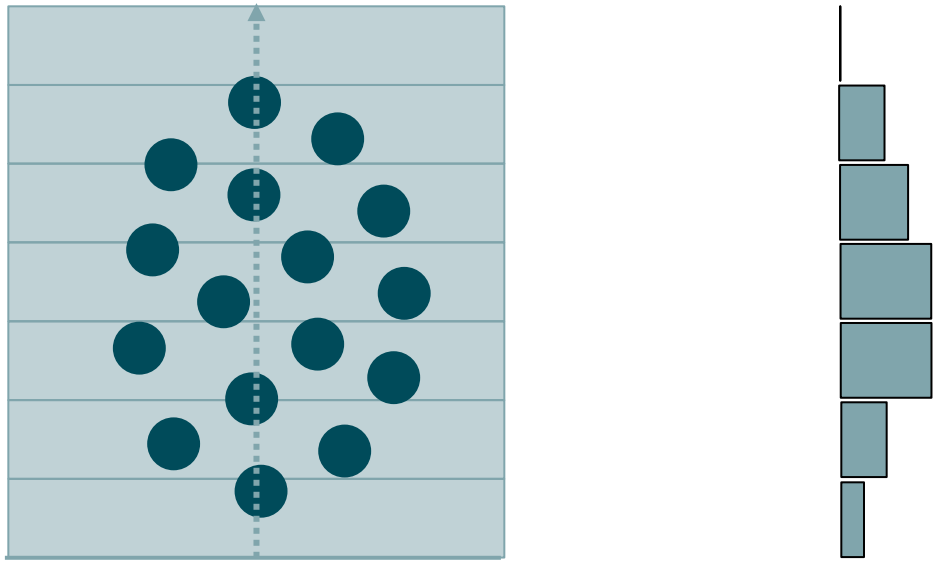
Shared work with Marven Kummerfeldt, Nele Brügge, Sebastian Meiser and Esfandiar Mohammadi

- Motivation
  - Generate synthetic data based on (sensitive) input
  - Differential Privacy → all following computations are privacy-preserving
- Overview
  1. Compute noisy 1-dimensional representations of the original data
  2. Find synthetic datapoints that result in the same representations

DPScans – Generate DP histograms

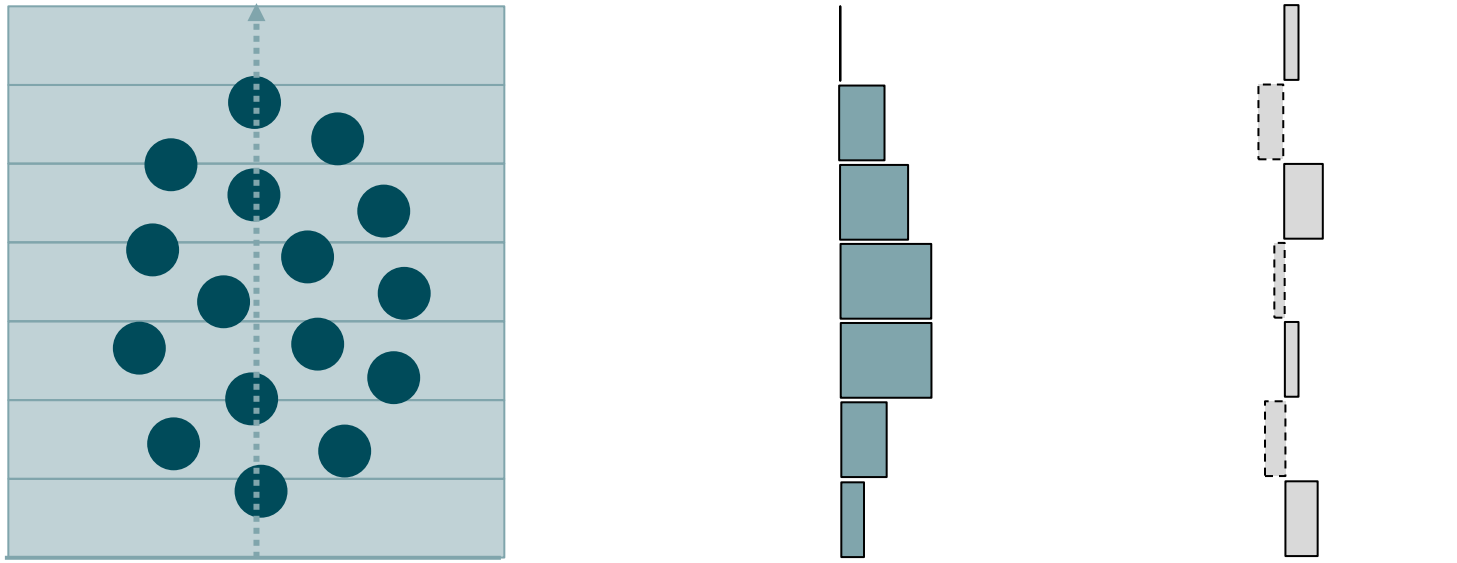


DPScans – Generate DP histograms



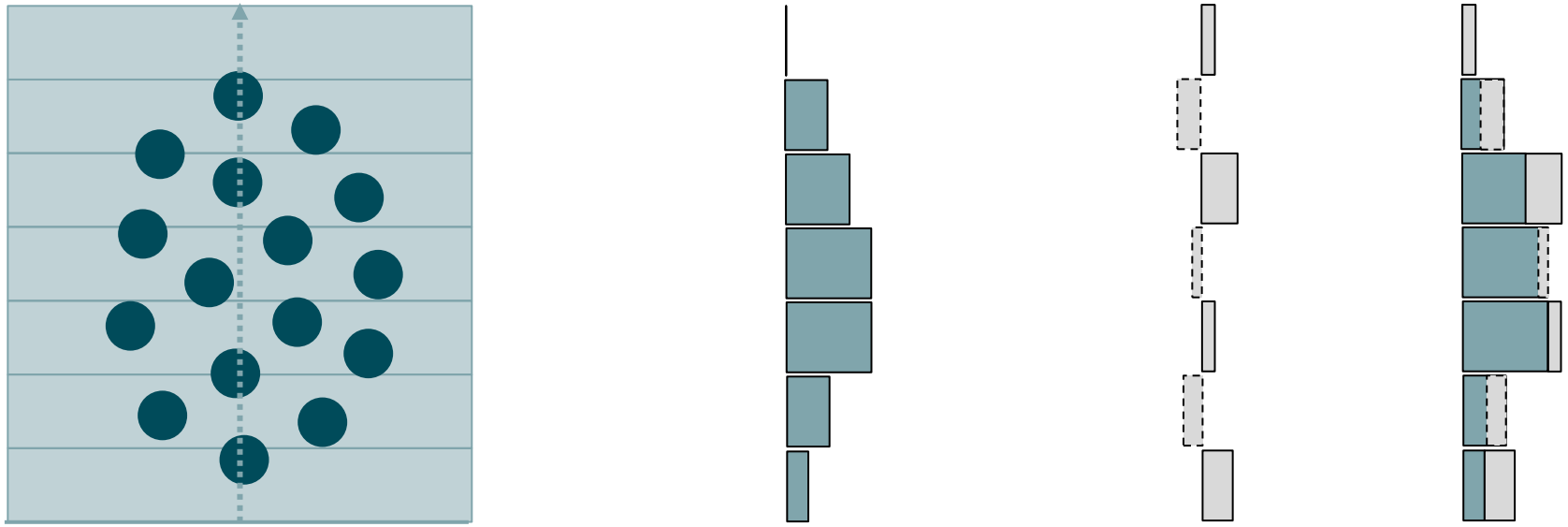
Scan datapoints → Generate histograms

DPScans – Generate DP histograms

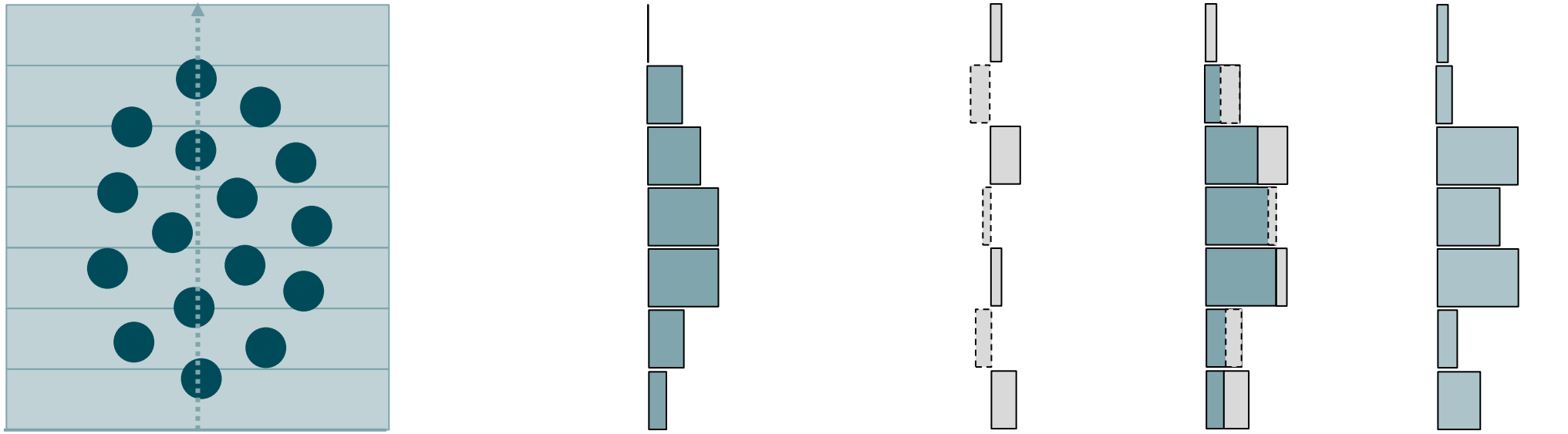


Scan datapoints → Generate histograms + Gaussian noise

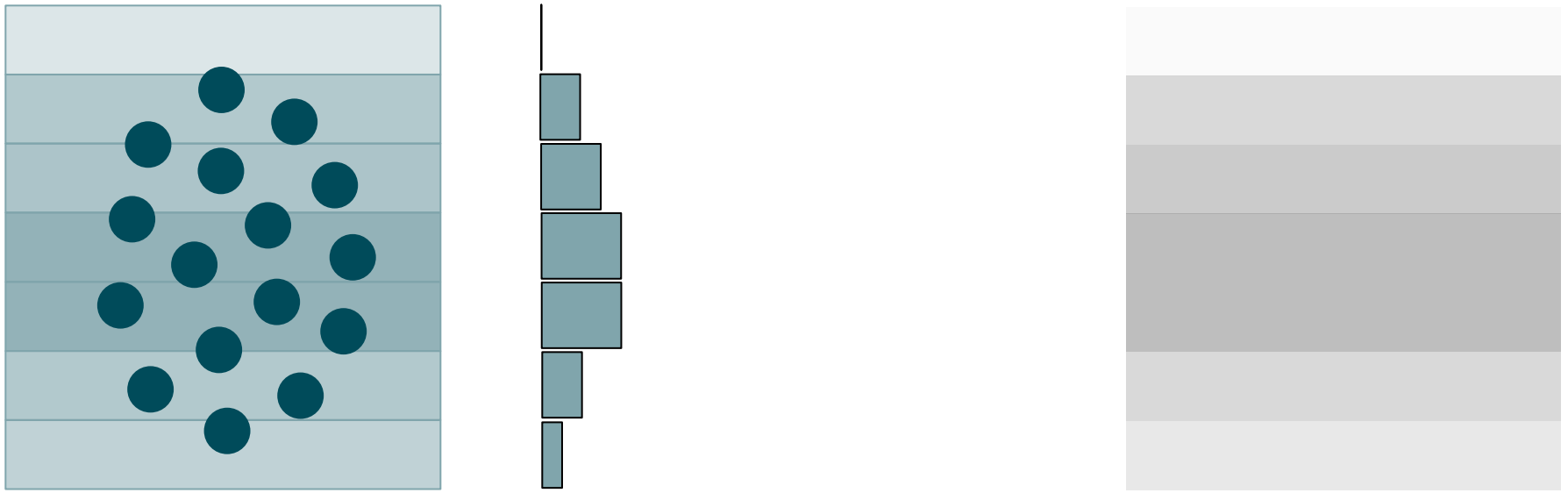
DPScans – Generate DP histograms



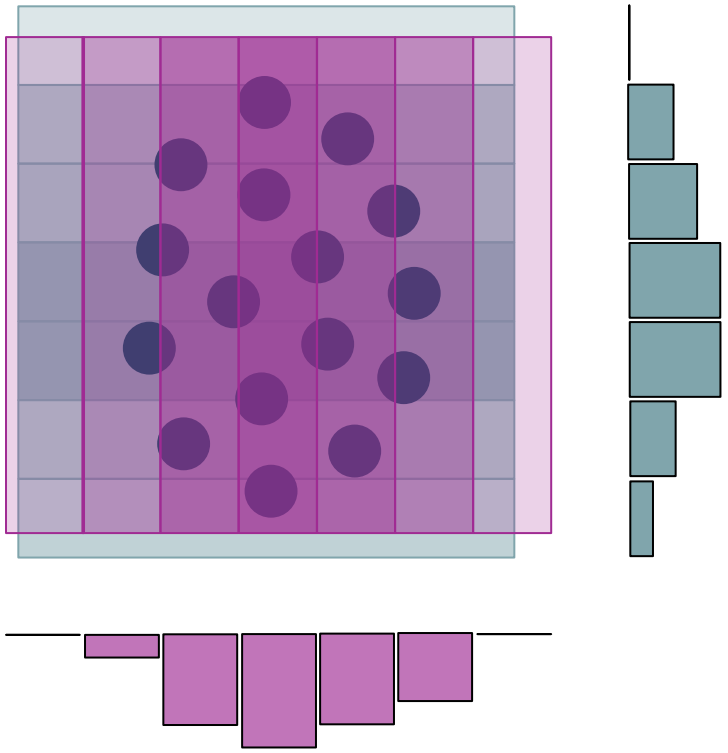
DPScans – Generate DP histograms

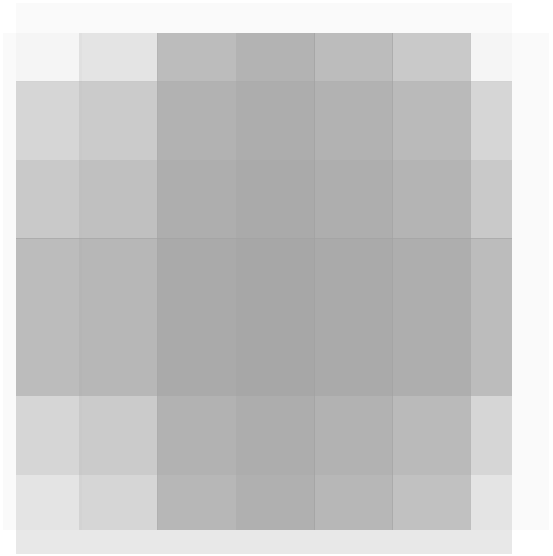


DPScans – Generate DP histograms

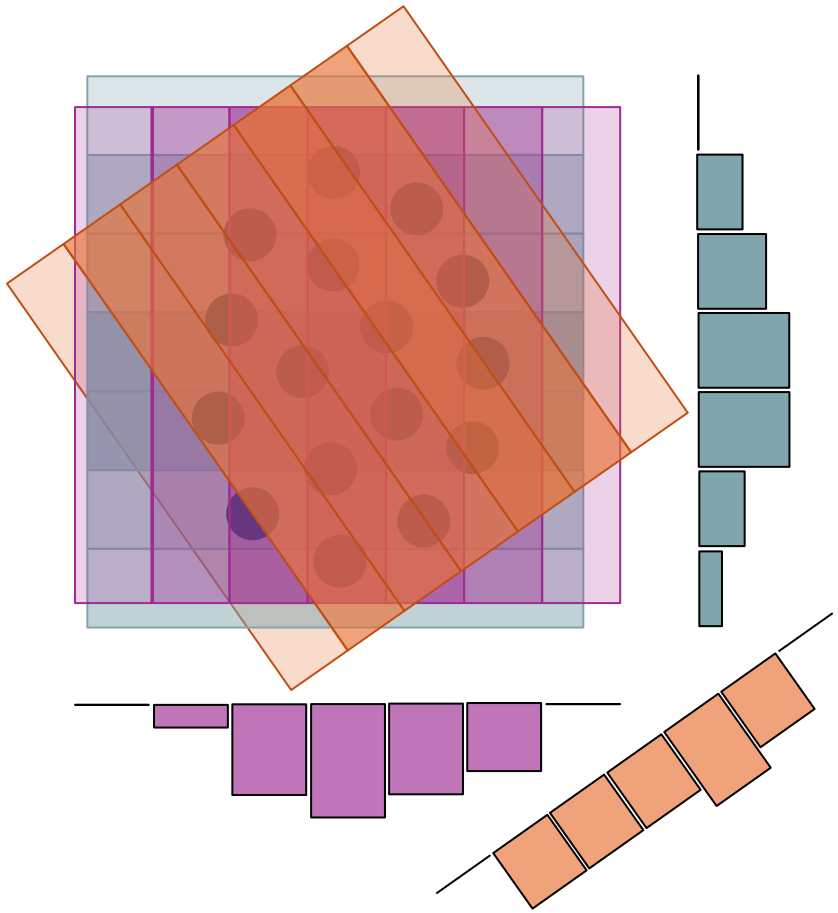


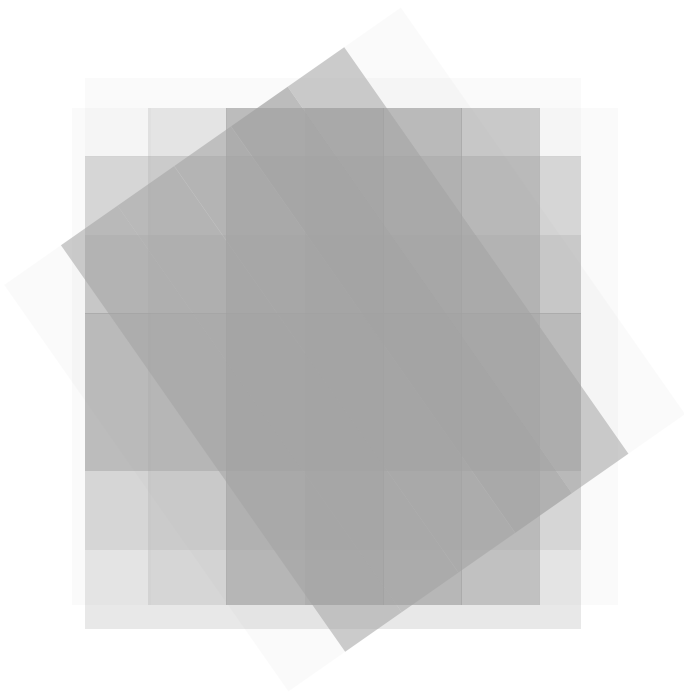
DPScans – Generate DP histograms



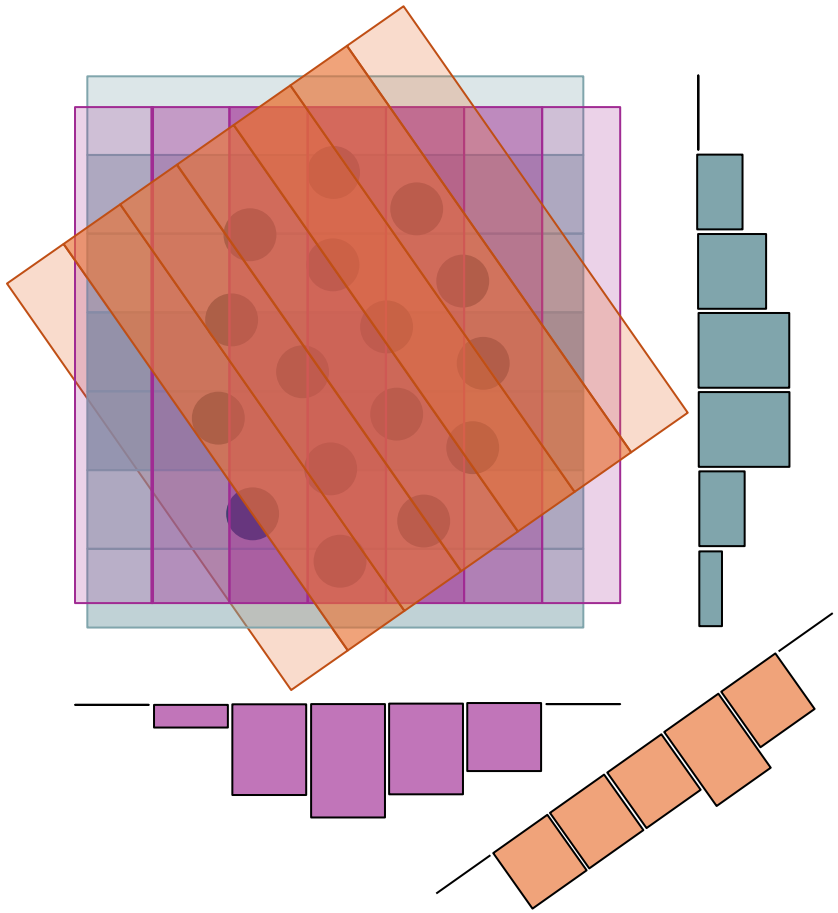


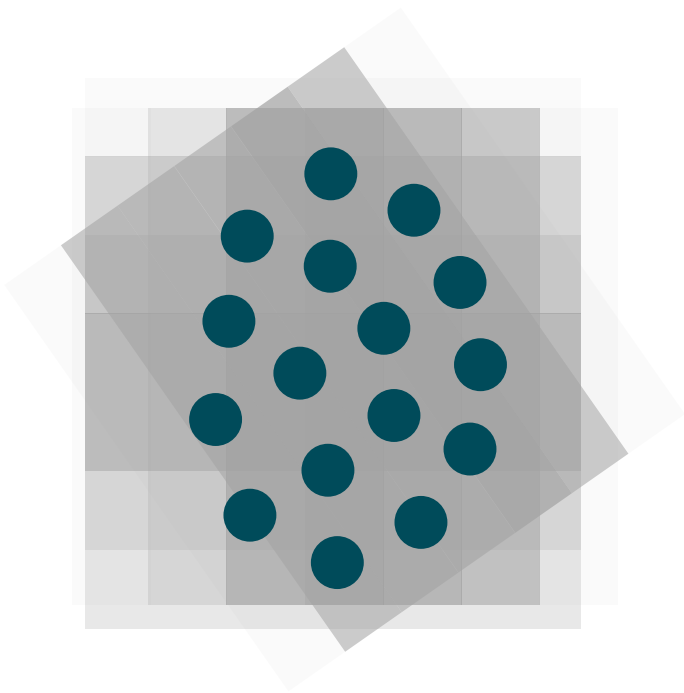
DPScans – Generate DP histograms



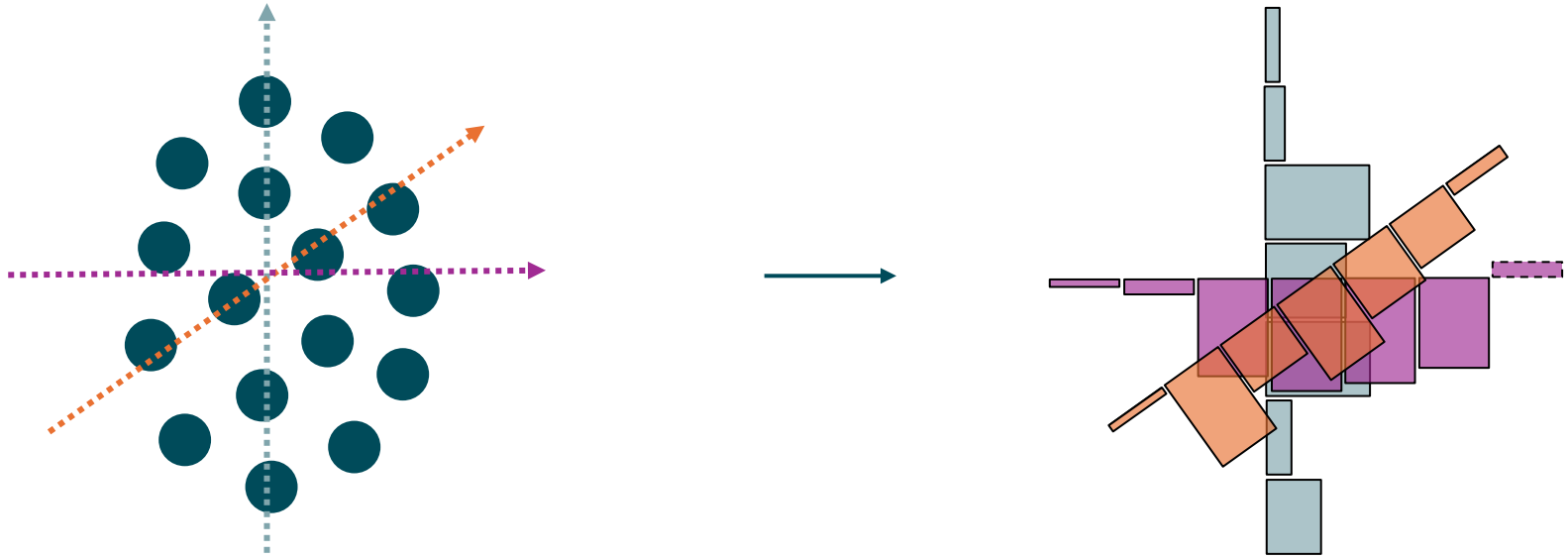


DPScans – Intuition multiple scans

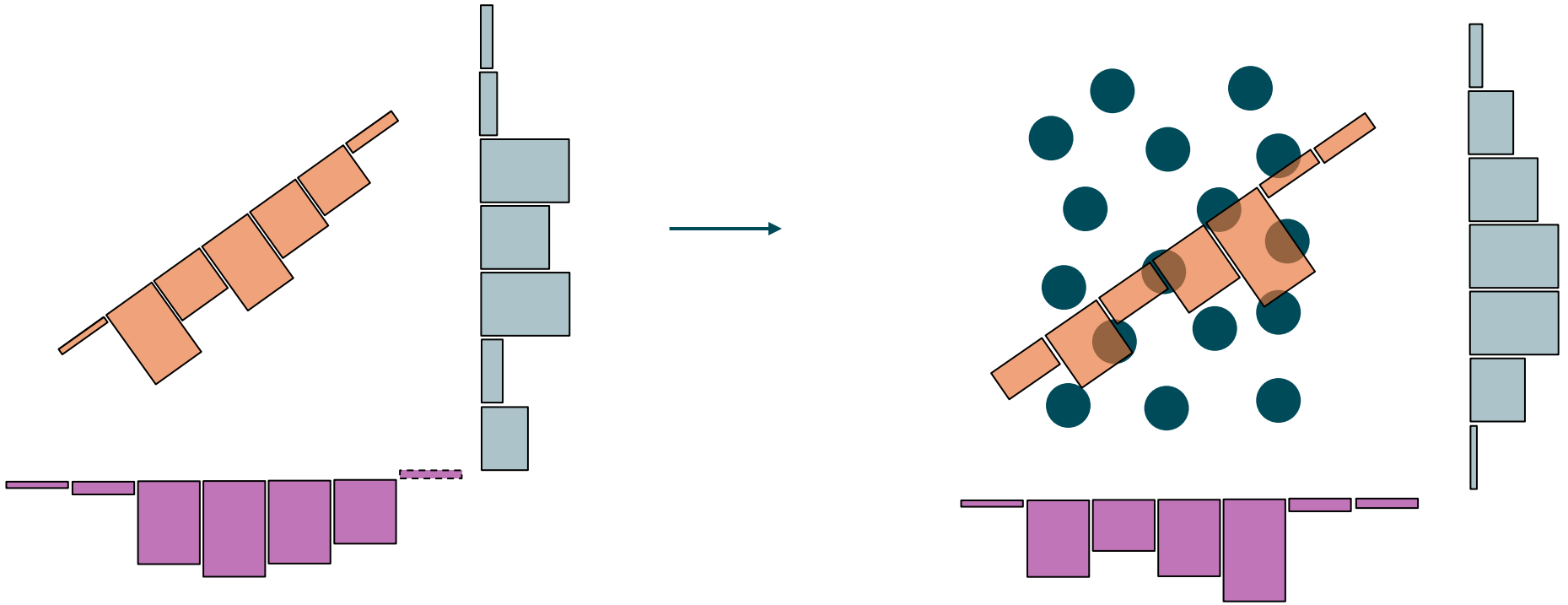




DPScans – Generate DP histograms

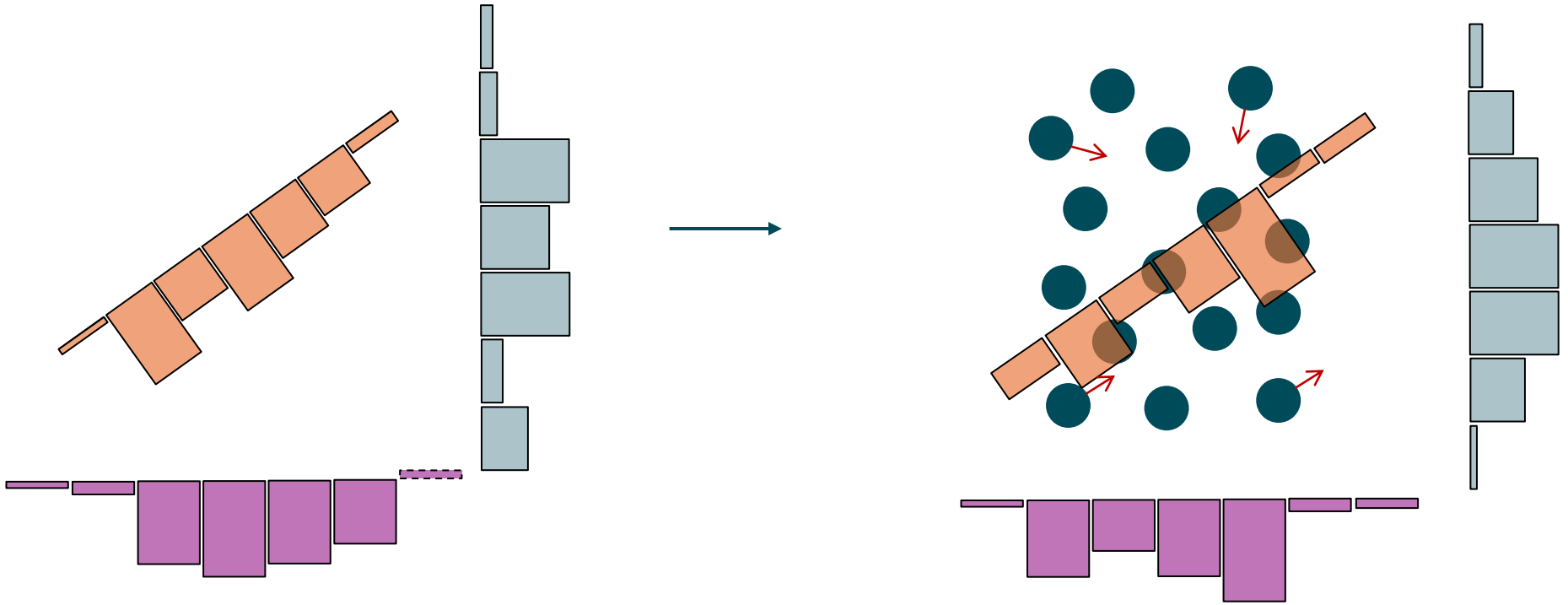


DPScans – Generate datapoints



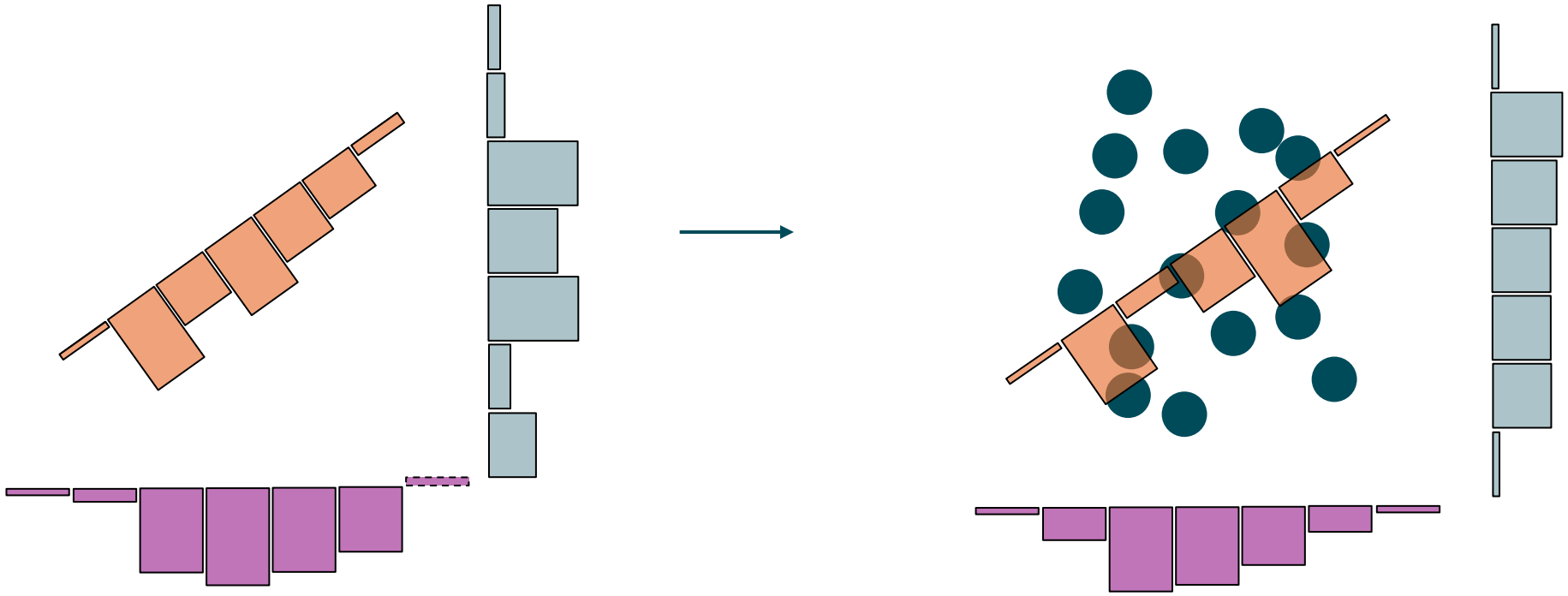
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



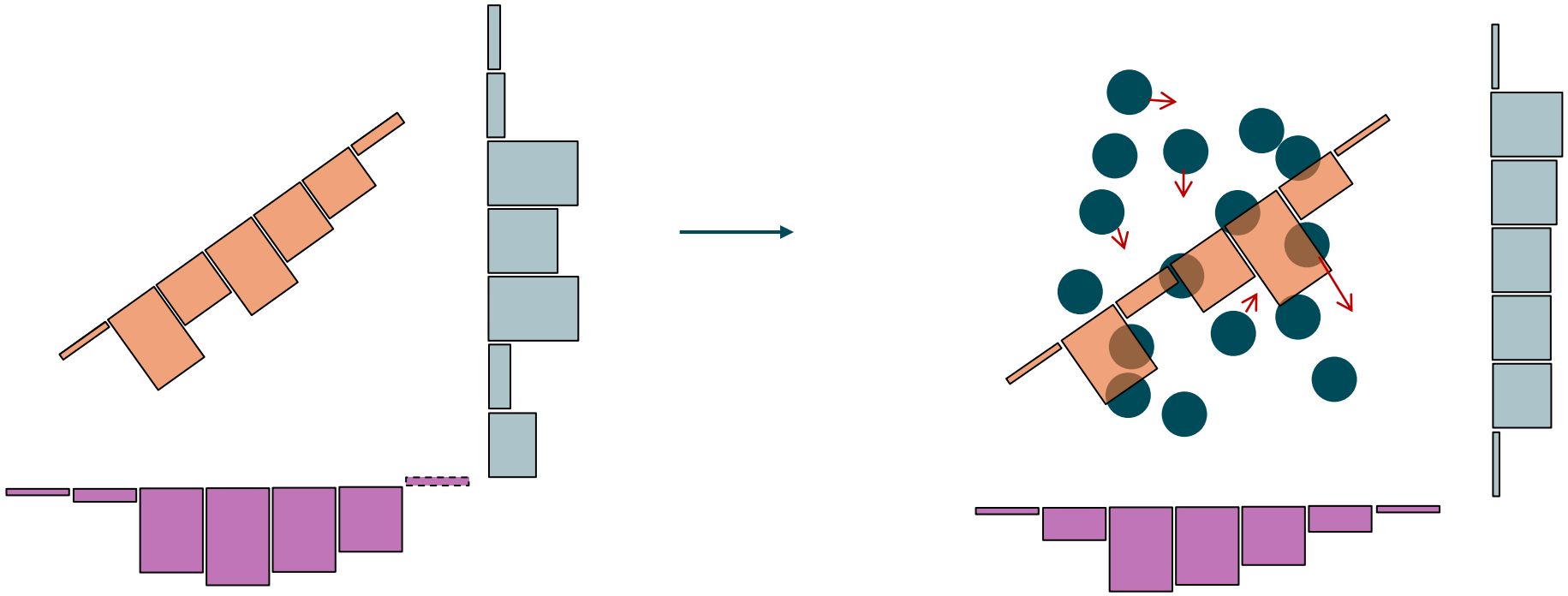
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



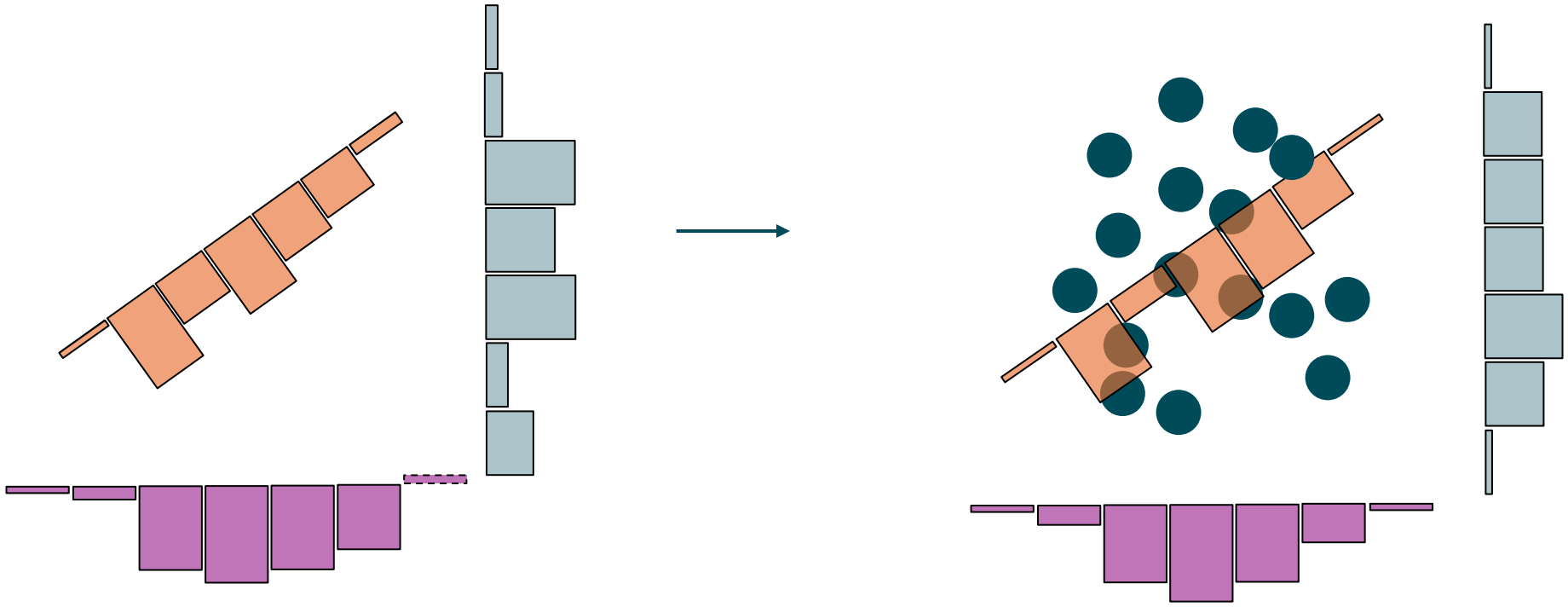
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



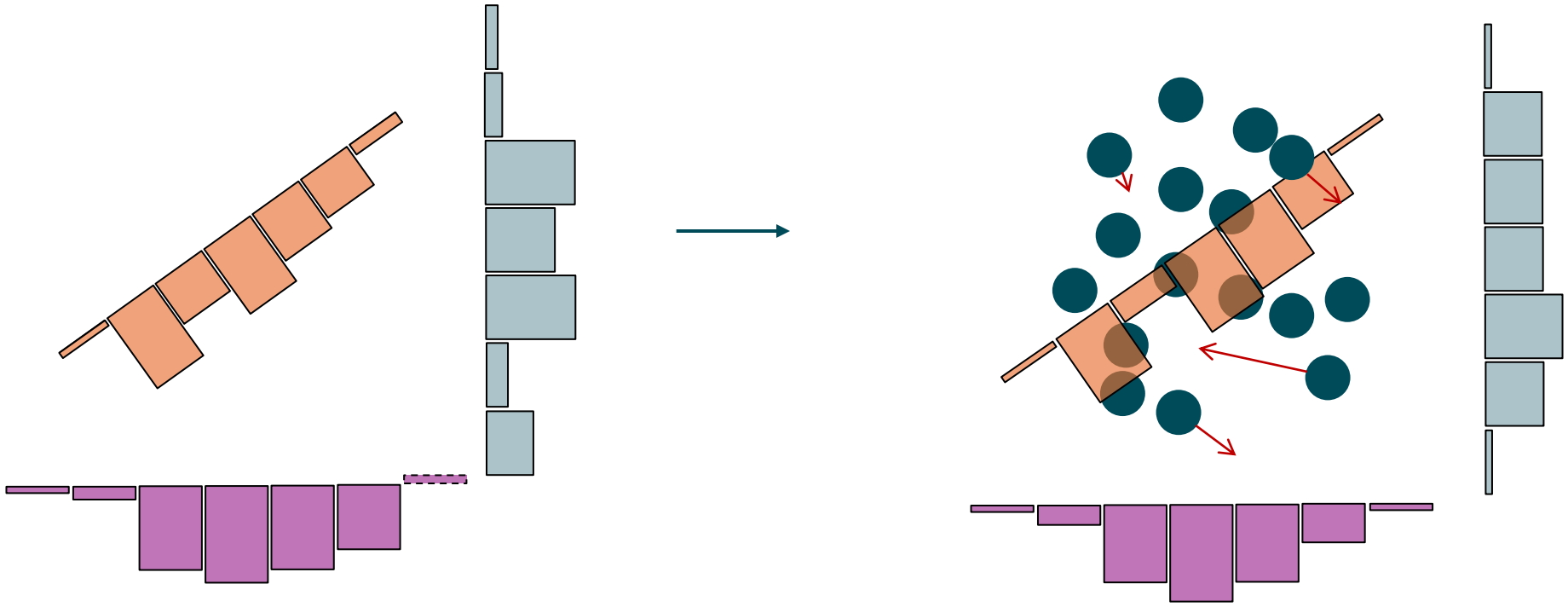
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



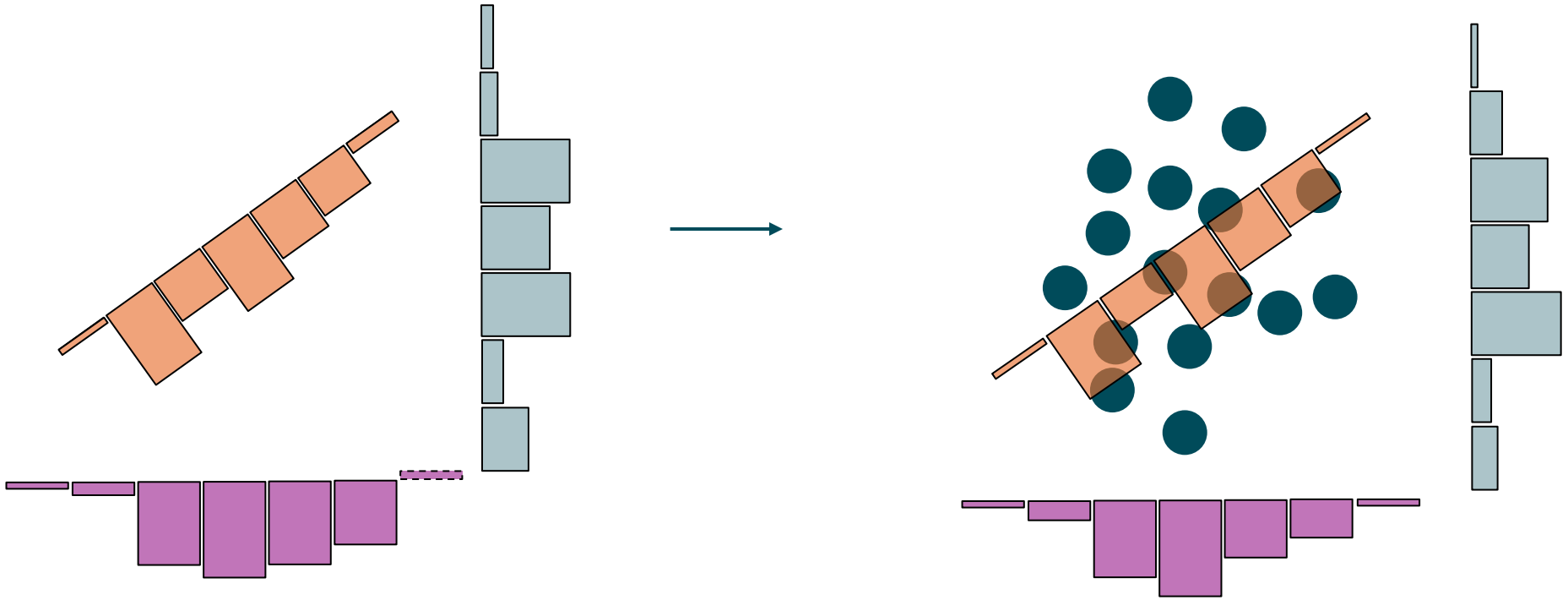
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



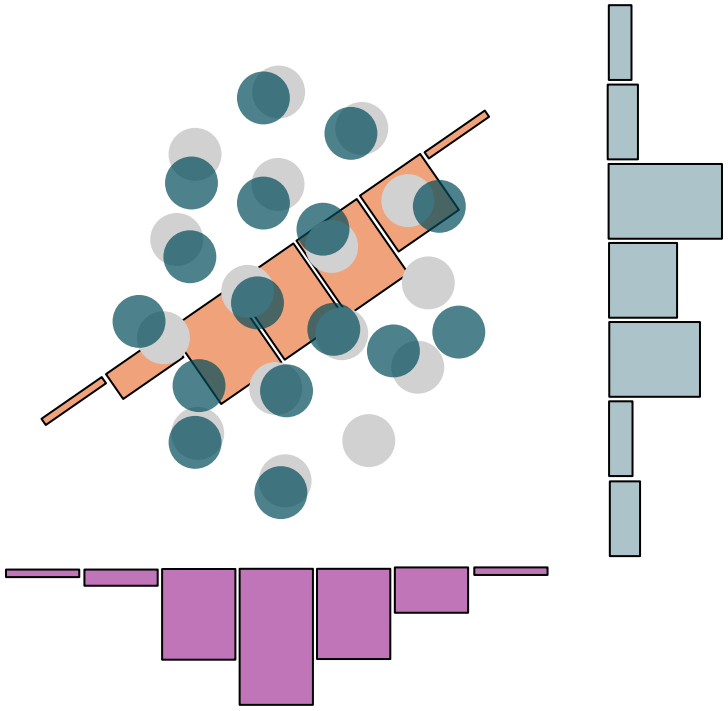
Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Generate datapoints



Iteratively:
Find datapoints that approximately generate same histograms

DPScans – Synthetic vs. Original data

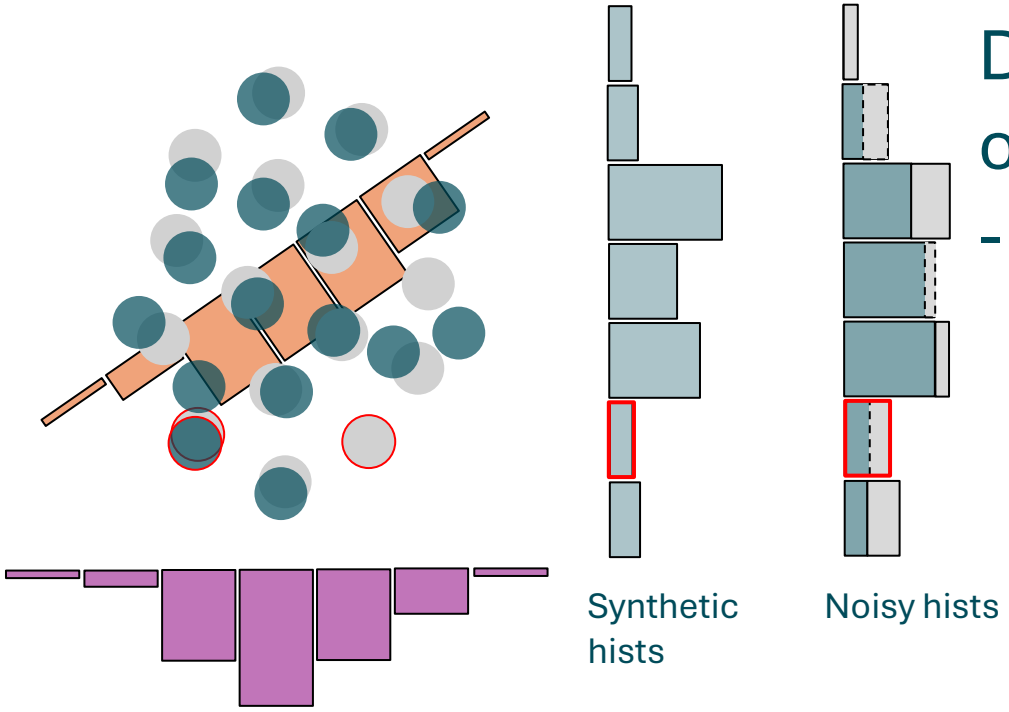


Original datapoints

Synthetic datapoints

Difference between synthetic and original data

DPScans – Synthetic vs. Original data



Synthetic hists

Noisy hists

Original datapoints

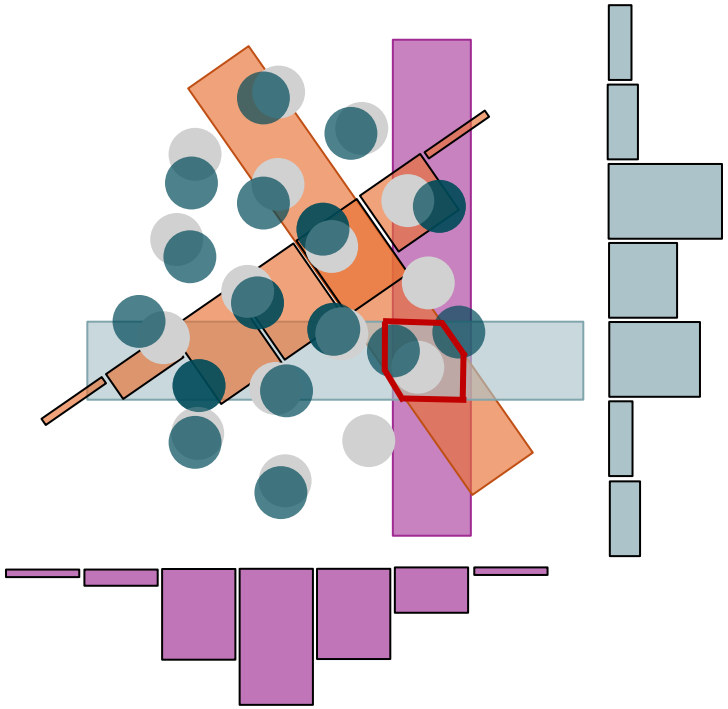
Synthetic datapoints

Difference between synthetic and original data

- Noisy histograms

Noisy hists ~ 1 datapoint → synthetic ~1 instead of real count of 2

DPScans – Synthetic vs. Original data



Original datapoints

Synthetic datapoints

Difference between synthetic and original data

- Noisy histograms
- Discretisation using histograms

DPScans – Experimental Results

		Adult			BreastCancer			Bank			Higgs			Loan		
		$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 10$
ACC ↑	PrivSyn	0.755	0.750	0.755	-	-	-	0.883	0.883	0.882	0.653	0.653	0.653	0.886	0.886	0.886
	PrivMRF	0.821	0.827	0.816	0.649	0.636	0.676	0.887	0.885	0.889	0.551	0.532	0.529	0.530	0.548	0.461
	RAP++	0.776	0.787	0.794	0.537	0.608	0.941	0.874	0.881	0.883	0.643	0.541	0.577	0.450	0.362	0.667
	AIM	0.816	0.815	0.823	0.560	0.609	0.524	0.884	0.886	0.886	0.750	0.764	0.734	0.886	0.884	0.886
	Private-GSD	0.782	0.791	0.797	0.573	0.765	0.933	0.882	0.883	0.883	0.503	0.564	0.573	0.886	0.886	0.886
	GEM	0.768	0.775	0.768	0.515	0.515	0.424	0.864	0.860	0.866	0.649	0.617	0.628	0.818	0.831	0.832
	DPScans (Ours)	0.796	0.827	0.821	0.616	0.680	0.942	0.883	0.884	0.831	0.716	0.748	0.748	0.886	0.886	0.886
	Real Data (non-private)		0.836			0.977			0.879			0.813			0.947	
Bal. ACC ↑	PrivSyn	0.501	0.514	0.501	-	-	-	0.500	0.500	0.500	0.505	0.503	0.504	0.500	0.500	0.500
	PrivMRF	0.746	0.756	0.722	0.609	0.593	0.634	0.654	0.690	0.700	0.530	0.521	0.513	0.506	0.512	0.499
	RAP++	0.571	0.573	0.591	0.582	0.673	0.943	0.511	0.506	0.503	0.588	0.576	0.560	0.526	0.535	0.546
	AIM	0.758	0.749	0.740	0.565	0.571	0.621	0.674	0.684	0.703	0.702	0.716	0.658	0.501	0.503	0.501
	Private-GSD	0.594	0.581	0.593	0.520	0.689	0.912	0.501	0.500	0.500	0.560	0.550	0.551	0.500	0.500	0.500
	GEM	0.642	0.647	0.670	0.566	0.492	0.405	0.555	0.543	0.565	0.547	0.580	0.557	0.501	0.515	0.522
	DPScans (Ours)	0.605	0.705	0.715	0.631	0.701	0.922	0.500	0.517	0.737	0.645	0.700	0.712	0.500	0.500	0.530
	Real Data (non-private)		0.727			0.969			0.788			0.802			0.933	
F1 score ↑	PrivSyn	0.437	0.465	0.438	-	-	-	0.469	0.469	0.469	0.420	0.413	0.418	0.470	0.470	0.470
	PrivMRF	0.751	0.760	0.734	0.581	0.572	0.619	0.678	0.703	0.713	0.524	0.510	0.503	0.419	0.435	0.386
	RAP++	0.566	0.566	0.597	0.443	0.598	0.937	0.496	0.482	0.475	0.565	0.488	0.527	0.352	0.324	0.478
	AIM	0.753	0.749	0.749	0.471	0.535	0.499	0.691	0.700	0.712	0.710	0.725	0.667	0.474	0.479	0.473
	Private-GSD	0.601	0.580	0.599	0.460	0.659	0.925	0.472	0.469	0.469	0.485	0.519	0.570	0.470	0.470	0.470
	GEM	0.649	0.659	0.671	0.502	0.396	0.363	0.557	0.548	0.577	0.530	0.576	0.551	0.491	0.502	0.515
	DPScans (Ours)	0.619	0.729	0.733	0.612	0.677	0.934	0.469	0.505	0.681	0.652	0.707	0.716	0.470	0.470	0.512
	Real Data (non-private)		0.750			0.975			0.743			0.806			0.947	

Open Questions – AnoMed II

- Theoretical utility guarantees of DPScans?
  - How many scans do we need for good reconstruction?
- How can the performance be improved?
  - Other generation models?
  - Adaptive Scans
  - Better pre- or post-processing?
- Other datasets (different tabular datasets, image datasets)
- Evaluation metrics?

Conclusion

- Pseudonymisation, statistics or synthetic data can still leak information about individuals
- Strong privacy guarantees required: Differential Privacy
- Differentially private synthetic data allows any post-processing
- DPScans: novel DP data synthesis algorithm with SOTA performance