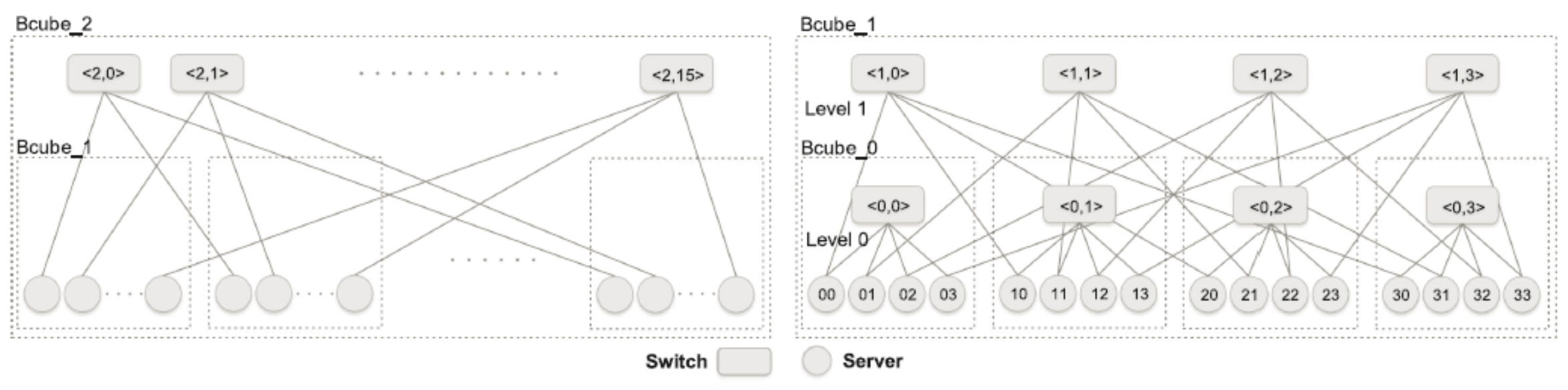
Publishing Scheme for Skewed Data Set

Team 34 – Zhenhao Cao & Yichen Zhu

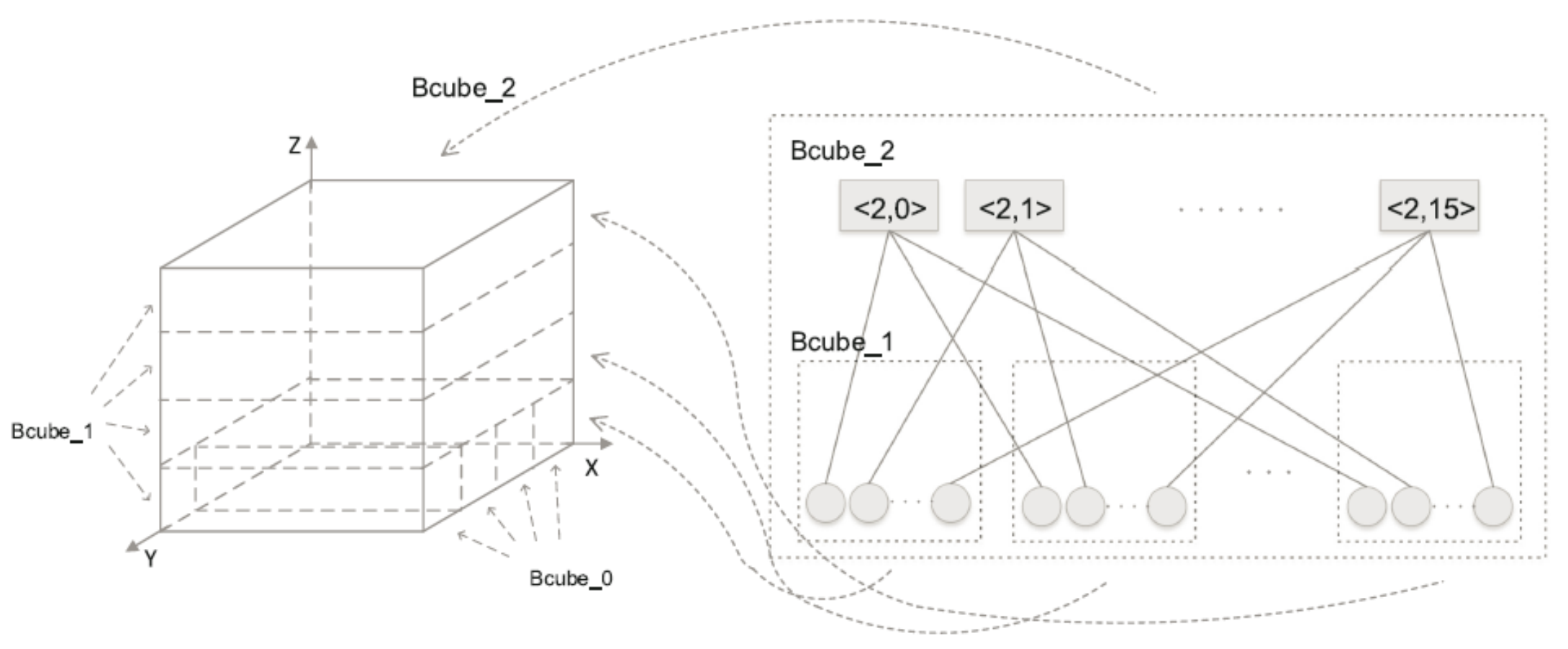
Hazelnut@sjtu.edu.cn & zyc_IEEE@sjtu.edu.cn

Background

Structure of BCube2 (n = 4)



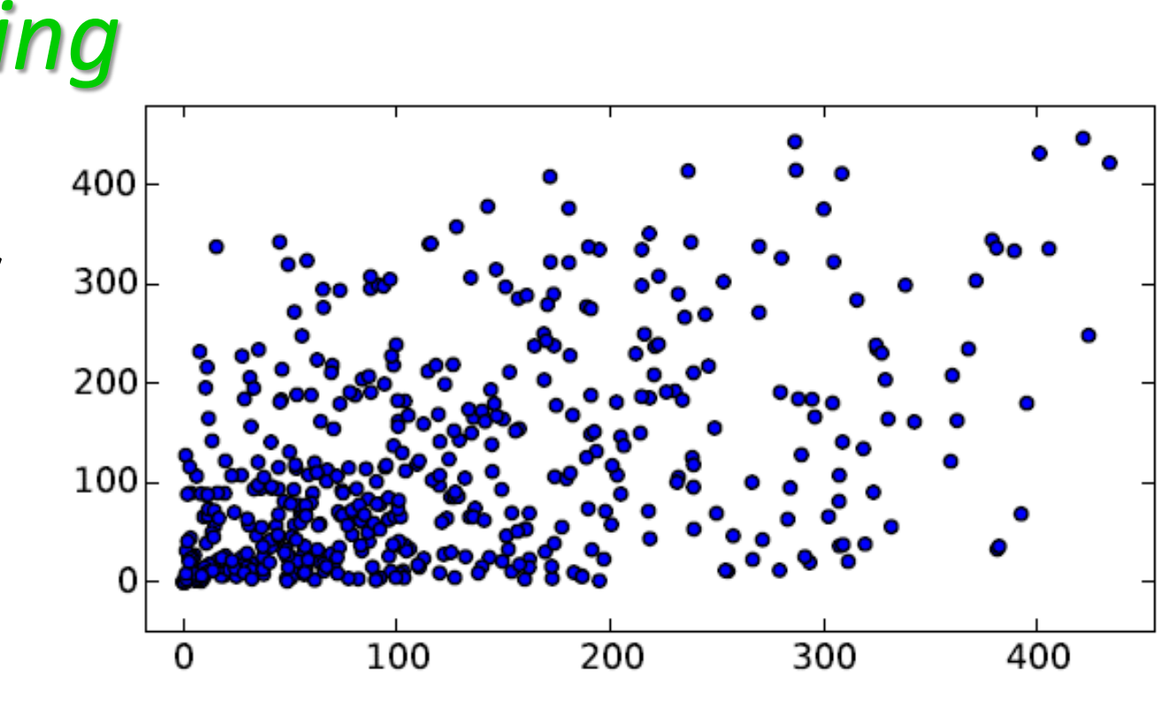
Indexing Space for BCube2 (n = 4)



Motivation

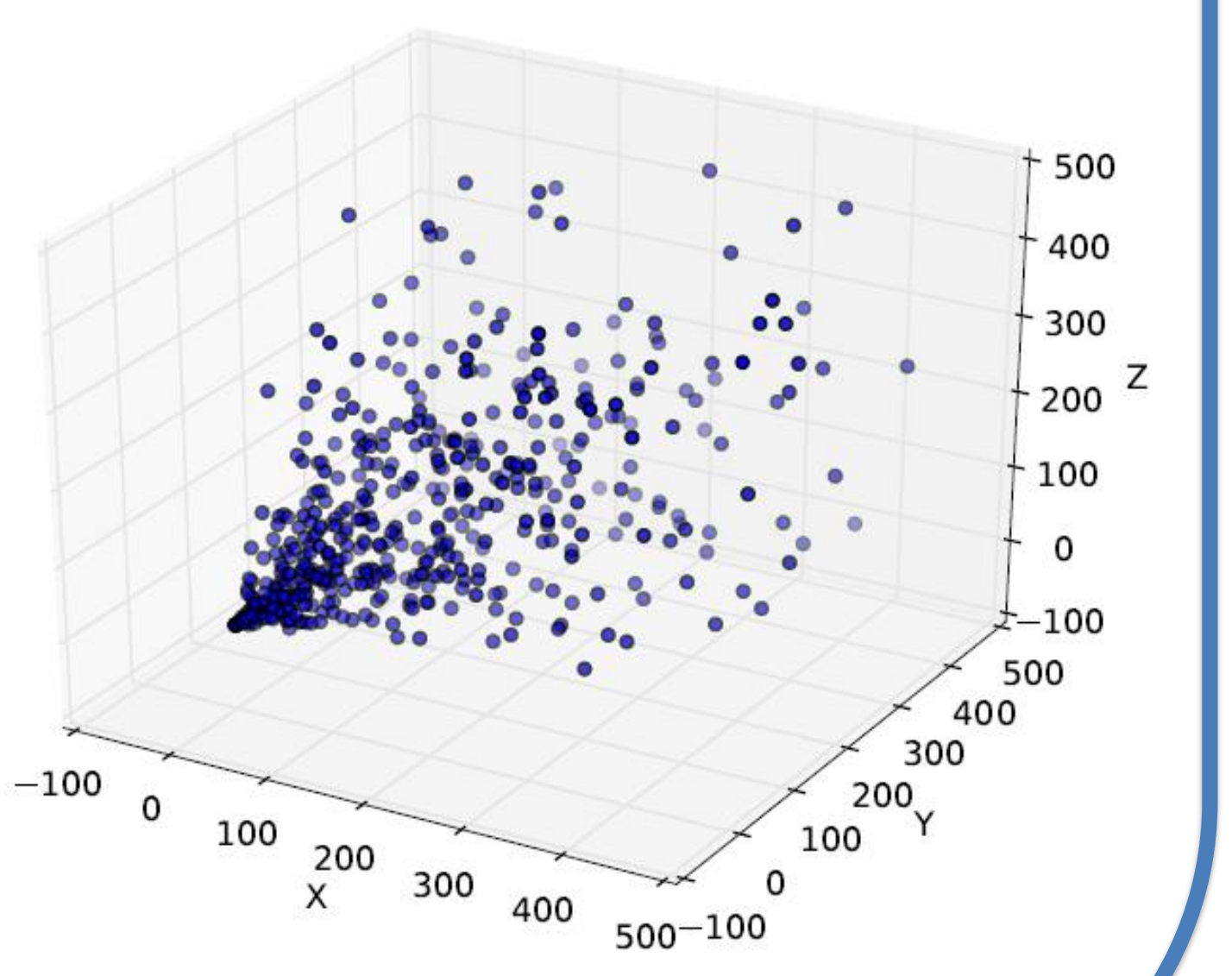
Existing Works: Direct-Mapping

In direct-mapping publishing scheme, data points are directly mapped to the indexing space without any pretreatment.



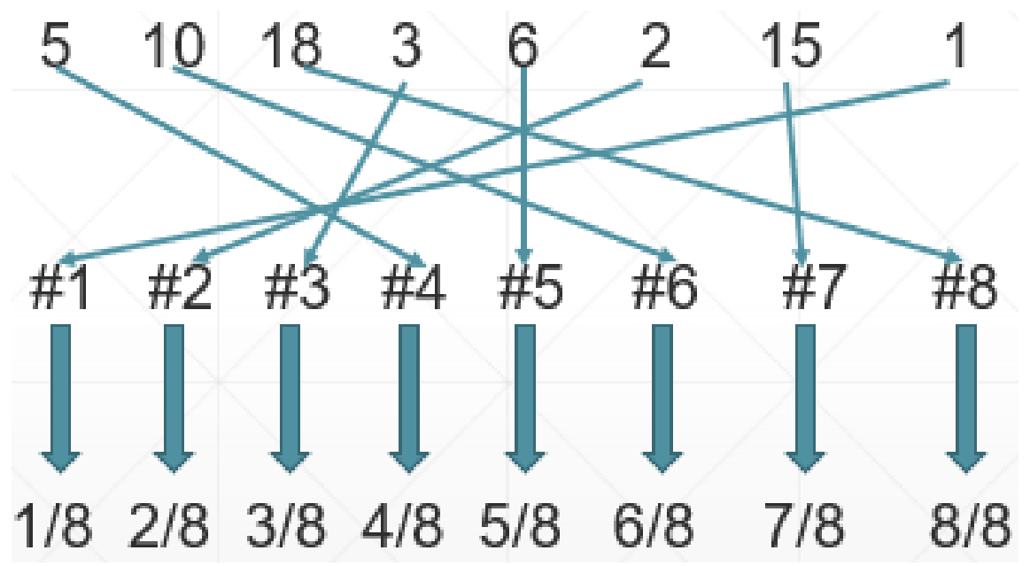
Drawback: Hot Spot Problem

For direct-mapping publishing scheme, if the data set is skewed, the distribution of points in the indexing space is unevenly distributed, which results in some servers being visited more frequently (too much workload).



Sorting-Based Mapping

In sorting-based mapping, for a single dimension, each data point is mapped to its corresponding sorting number.



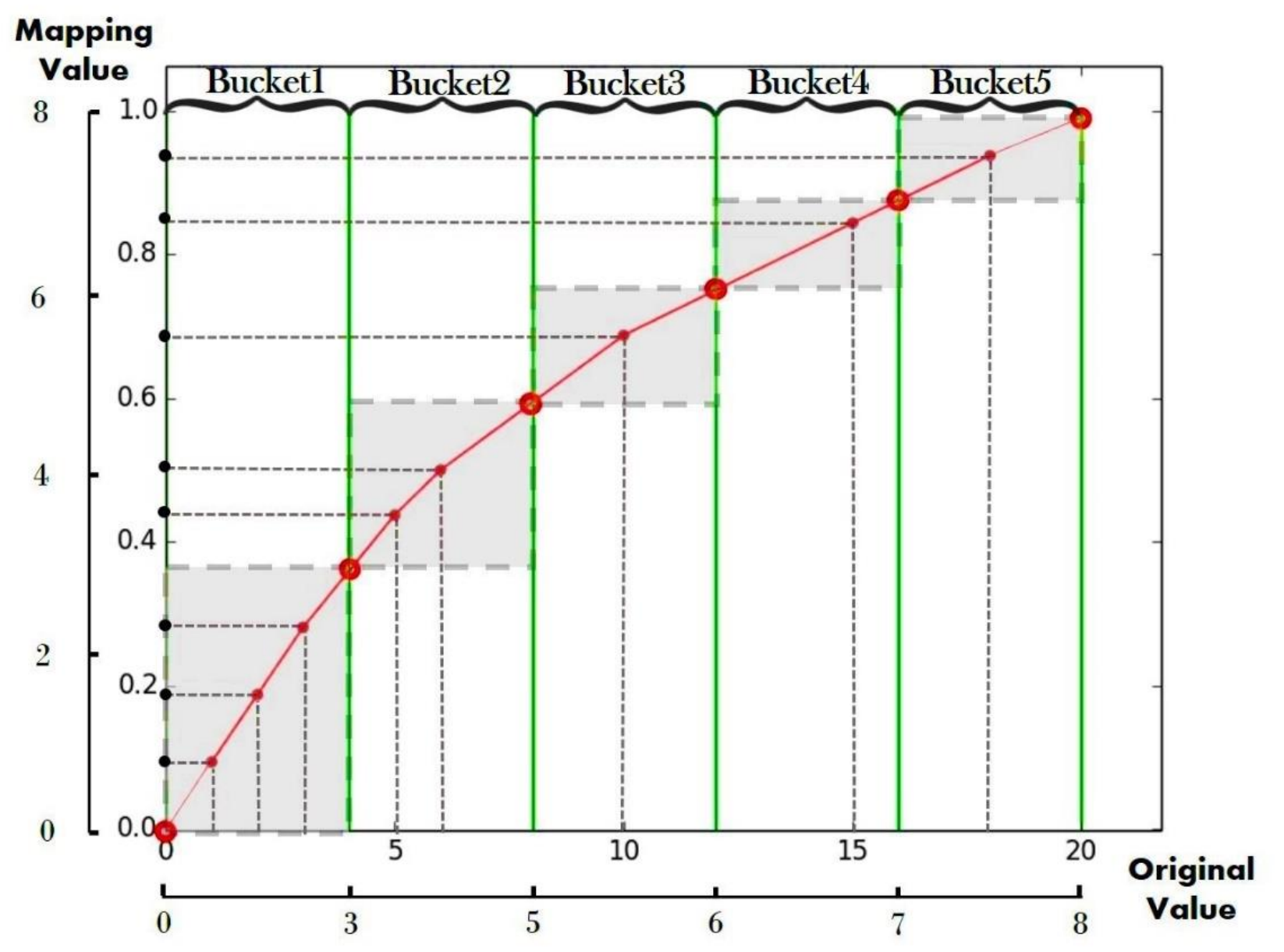
Drawback: Sorting Complexity

All the data points have to be sorted firstly, which is too expensive for a large data set.

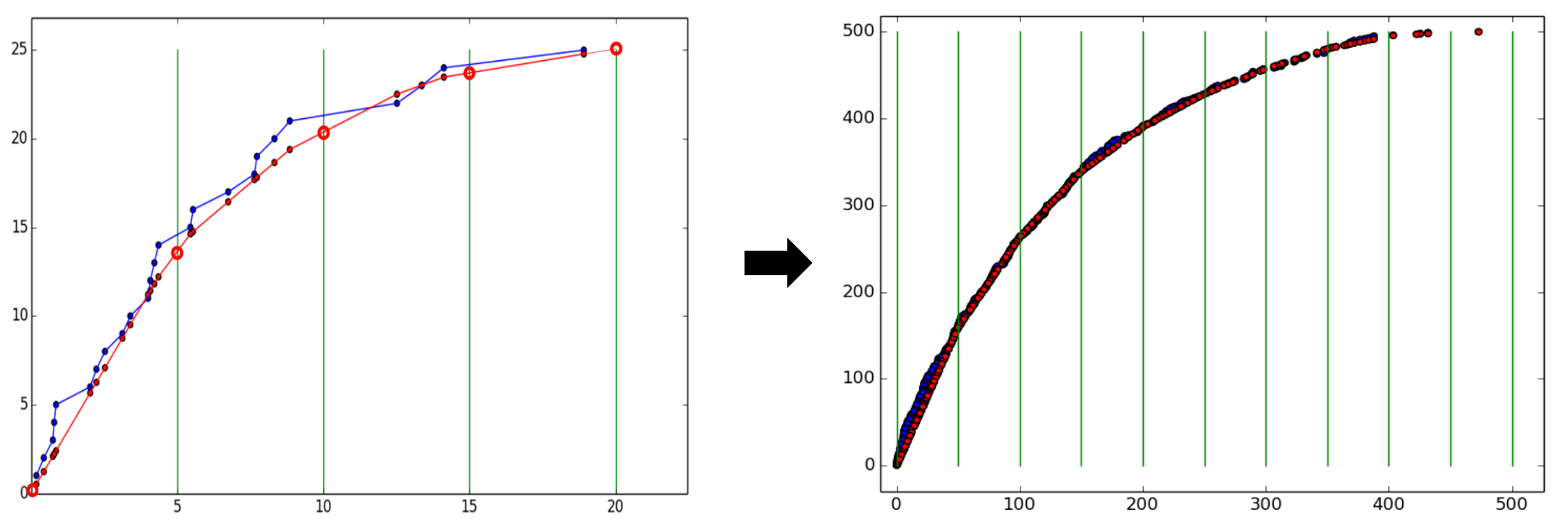
Casting-Based Mapping

Reduce Complexity by Approximation

We use one segment to fit several segments in order to achieve approximation.

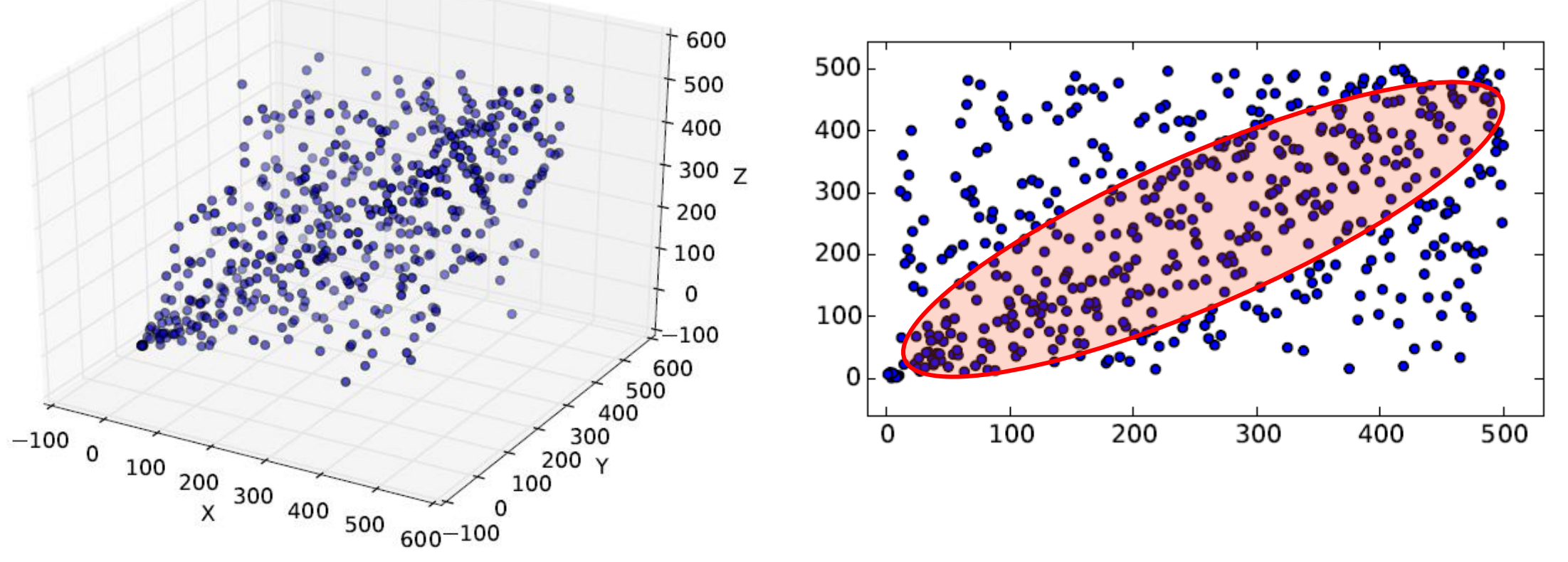


Massive Data



Drawback: Bad Uniformity in Space

For sorting-based and casting-based mapping, data points only evenly distributed in a single dimension, but tend to gather around diagonal line 'x=y=z'.



Modified Casting-Based Method

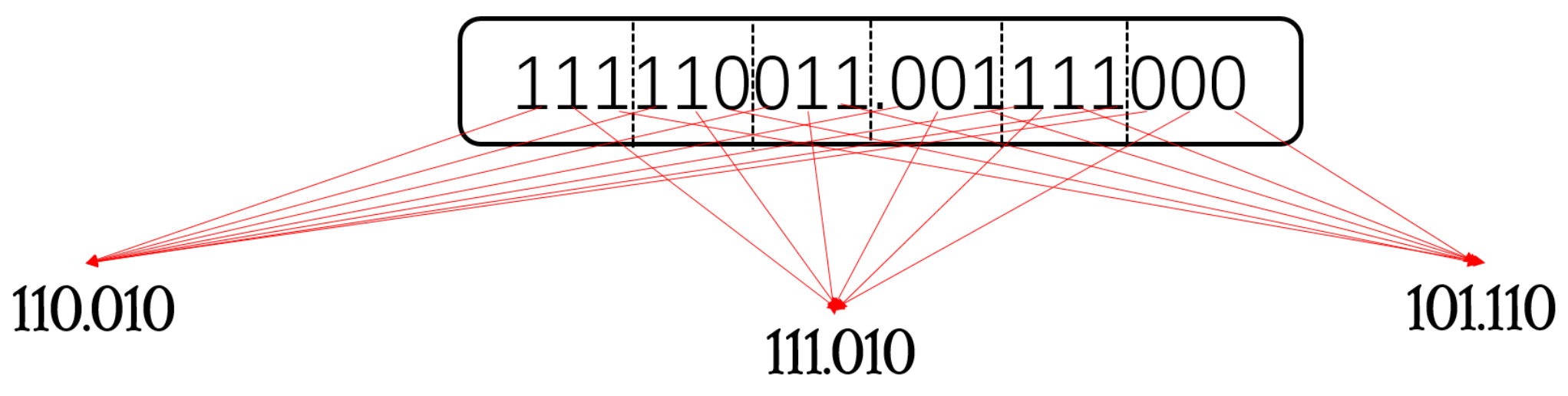
Dominant Dimension Selection

- ➢ Define $Relative\ Uniformity$ for single dimension

$$RU_i = \prod_{j=1}^{b} Slope_j$$

One-dimension to d-dimension

499.2345 → 111110011.001111000



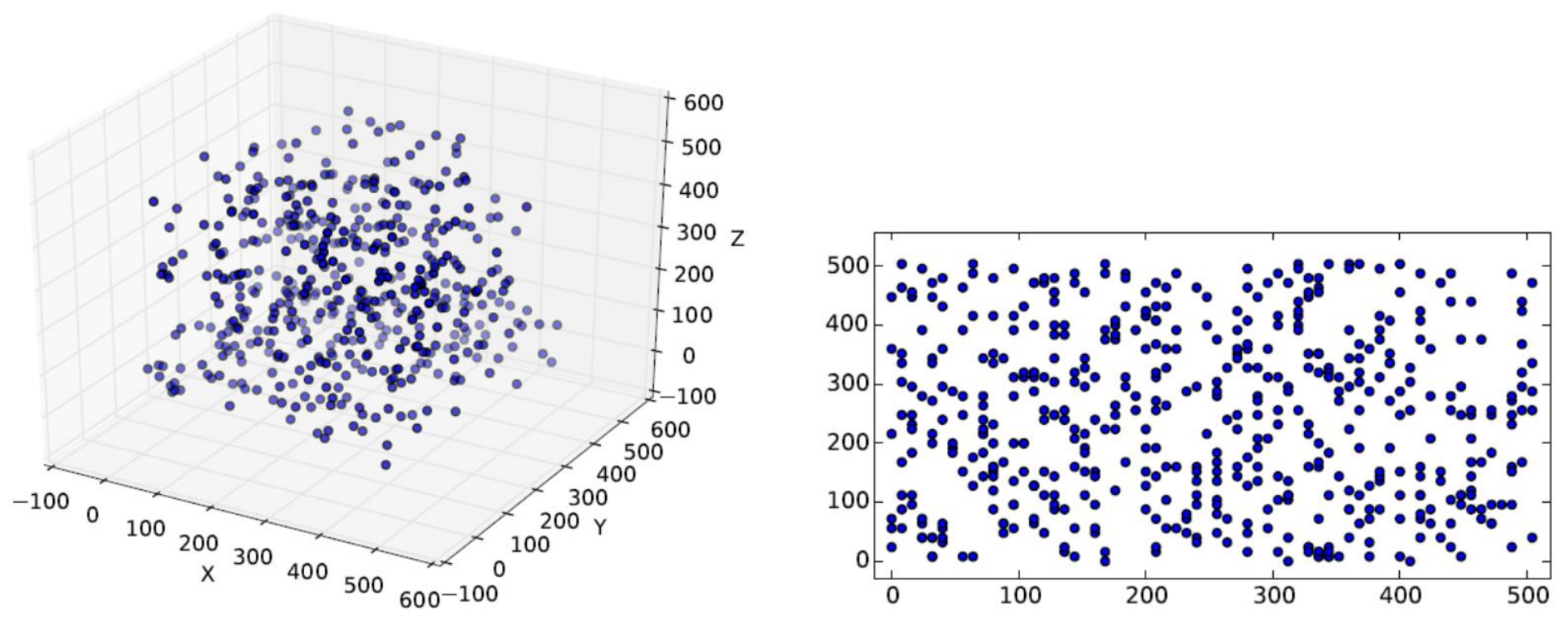
(110.010, 111.010, 101.110) * Scale

↓

(400, 464, 368)

Result scatter diagram

(Data Set in Overlapping Uniform Distributions)



(a) 3-D Mapping (b) 2-D Projection

Evaluation Metrics

Metric 1: Gap Ratio

- ➢ Let (M, δ) be a metric space and P be a set of k points sampled from M
- ➢ Define the $minimum\ gap$ as

$$r_p \coloneqq \min_{p,q \in P, p \neq q} \frac{\delta(p,q)}{2}$$

- ➢ Define the $maximum\ gap$ as

$$Rp \coloneqq \sup_{q \in M} \delta(q, P)$$

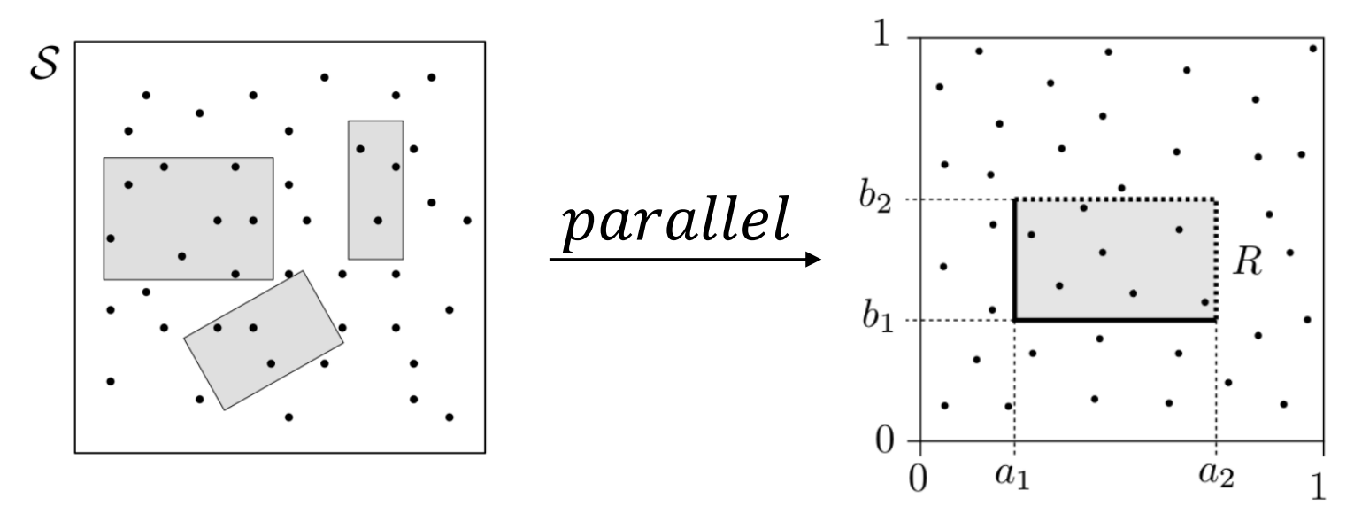
- ➢ Define the $Gap\ Ratio$ as

$$GR_p \coloneqq \frac{R_p}{r_p}$$

Metric 2: Discrepancy

- ➢ $Discrepancy$ definition

$$D(P) \coloneqq \sup_{B \in R} \left| n \cdot vol(B) - |P \cap B| \right|$$



- ➢ Three-dimensional generalization

$$D(P) \coloneqq \sup_{x,y,z \in [0,1]} \left| xyz - \frac{|([0,x] \times [0,y] \times [0,z]) \cap P|}{N} \right|$$

Experiment Outcomes

Experiment Outcomes

	Overlapping Uniform Distribution (cone-shaped)				Overlapping Normal Distribution			
	rp	Rp	Gap Ratio	Discrepancy	rp	Rp	Gap Ratio	Discrepancy
DMM	0.0665	865.5078	13007.14	0.5248	1.3111	840.3373	640.9543	0.4080
SBM	1.2247	860.8304	702.87	0.0979	2.5495	850.4511	333.5743	0.0241
CBM	0.2283	864.2498	3784.28	0.1067	2.1956	855.9634	389.8527	0.0537
✓ MCBM	4.0013	821.7348	205.43	0.0735	4.1059	866.0254	216.5064	0.0472