Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Project Group Formation

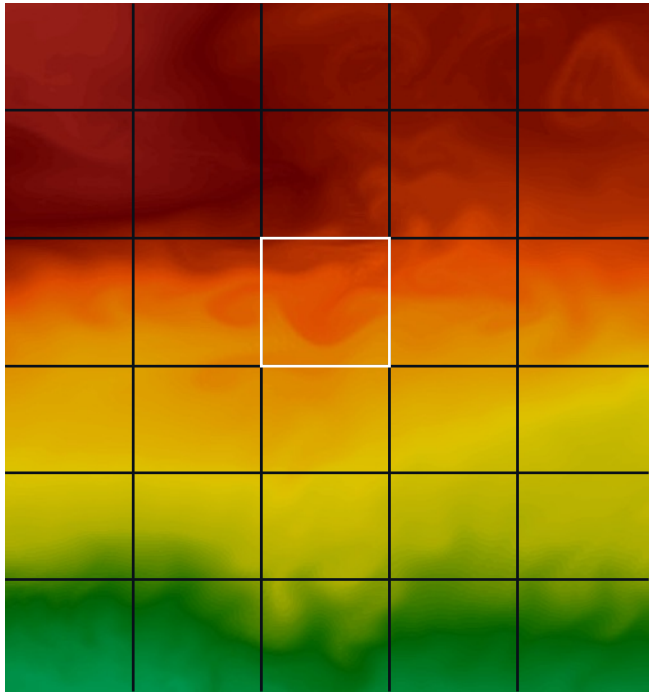
- Form groups by tonight and update the google spreadsheet
- An email was sent earlier today
- If you can’t find a group, I will start forming new groups

Study Materials for Lecture 15

- SLIC Superpixels Compared to State-of-the-Art Superpixel Methods, Achanta et al.
- Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets, Dutta et al. IEEE PacificVis.
- Statistical visualization and analysis of large data using a value-based spatial distribution, Wang et al., IEEE PacificVis.
- Distribution Driven Extraction and Tracking of Features for Time-varying Data Analysis, Dutta et al., IEEE TVCG.

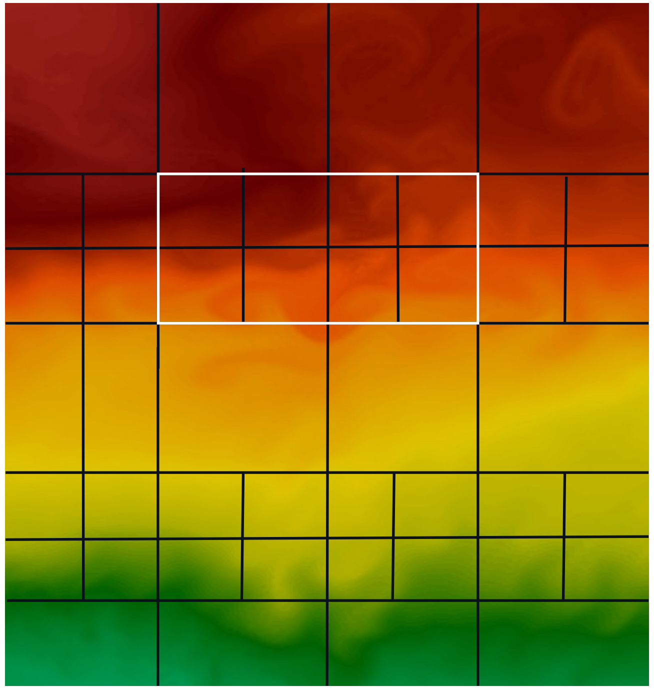
Goals of a Region-wise Statistical Summarization

- Produce coherent partitions
  - Similar data values are grouped together
  - Partitions are spatially contiguous
- Preserve the statistical properties of the data accurately
  - Minimize sampling errors
  - Efficient feature analysis
- Use appropriate distribution models for summarization
  - A compact storage representation



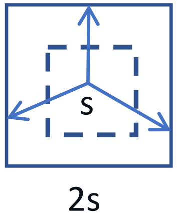
Goals of a Region-wise Statistical Summarization

- Produce coherent partitions
  - Similar data values are grouped together
  - Partitions are spatially contiguous
- Preserve the statistical properties of the data accurately
  - Minimize sampling errors
  - Efficient feature analysis
- Use appropriate distribution models for summarization
  - A compact storage representation



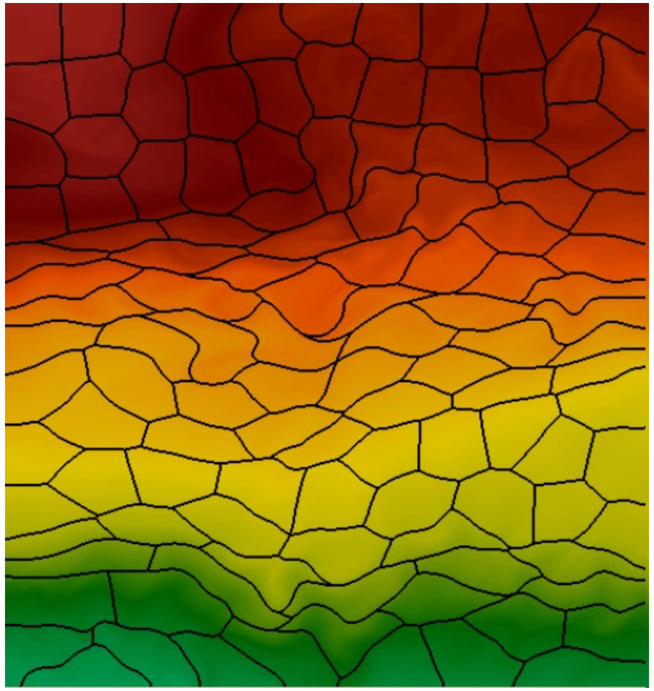
A Superior Solution for Region-wise Statistical Summarization

- Generate partitions based on data homogeneity
- Simple Linear Iterative Clustering (SLIC)
  - Produces irregular shaped partitions/clusters
  - Value variation inside partitions is minimized
  - Reduced sampling error



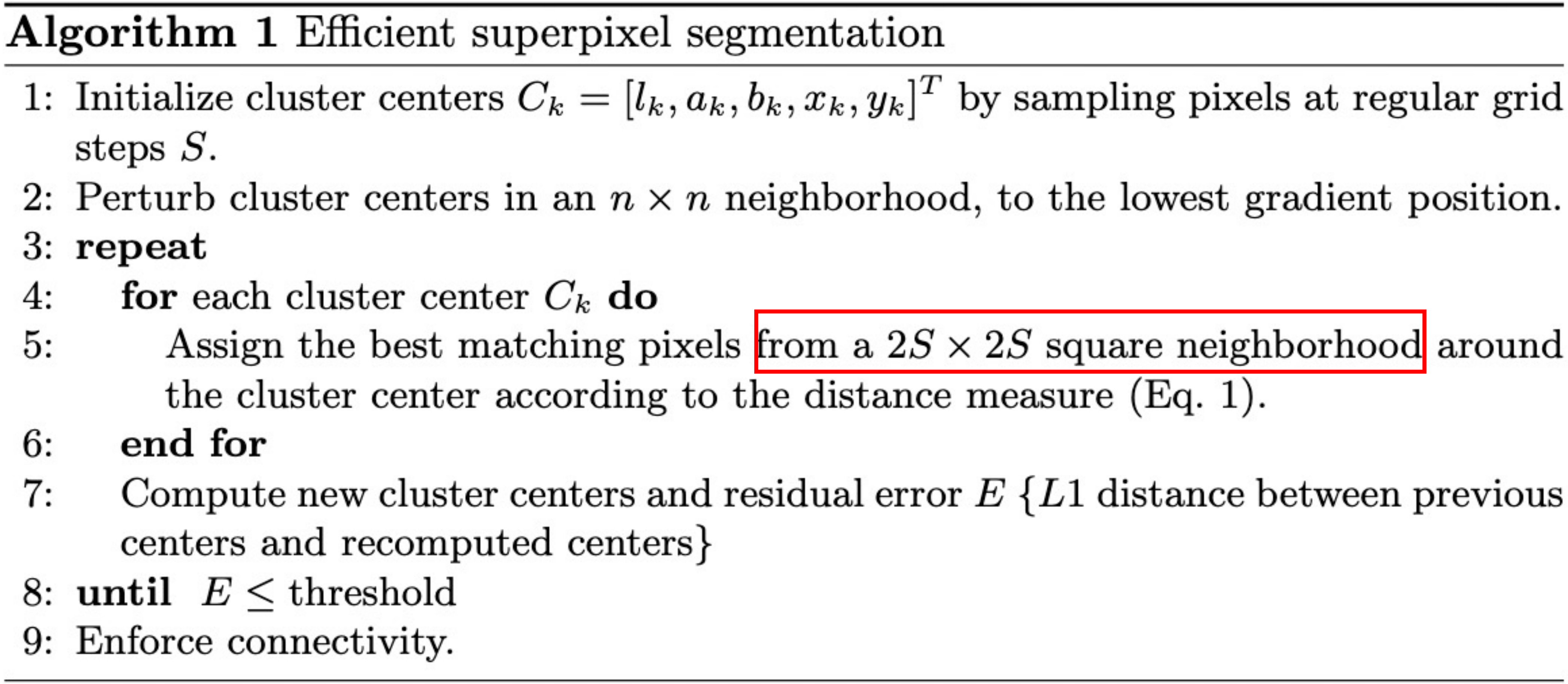
SLIC concept

$$dist(i,j) = \alpha . \| C_i - P_j \|_2 + (1-\alpha).|val_i - val_j|$$



https://www.iro.umontreal.ca/~mignotte/IFT6150/Articles/SLIC_Superpixels.pdf

SLIC Algorithm Steps



Algorithm 1 Efficient superpixel segmentation

```
Initialize cluster centers $C_k = [l_k, a_k, b_k, x_k, y_k]^T$ by sampling pixels at regular grid steps $S$.
Perturb cluster centers in an $n \times n$ neighborhood, to the lowest gradient position.
repeat
    for each cluster center $C_k$ do
        Assign the best matching pixels from a $2S \times 2S$ square neighborhood around the cluster center according to the distance measure (Eq. 1).
    end for
    Compute new cluster centers and residual error $E$ {$L1$ distance between previous centers and recomputed centers}
until $E \leq$ threshold
Enforce connectivity.
```

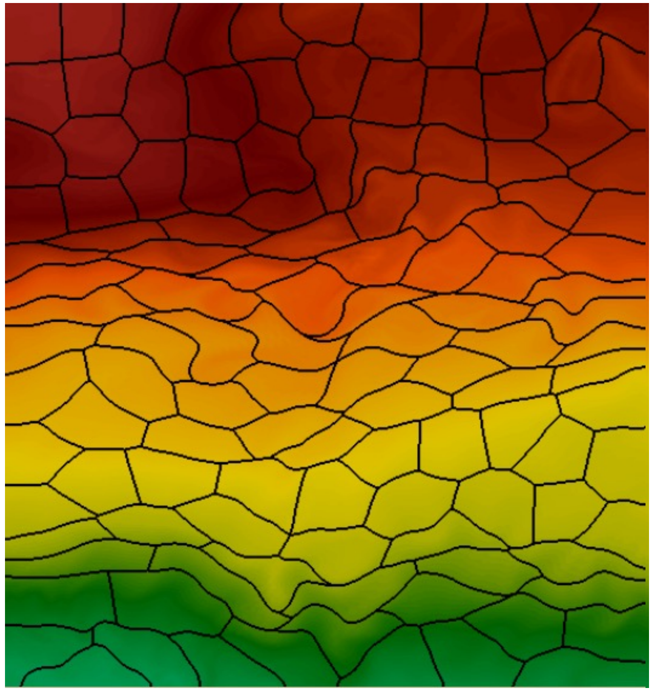
https://www.iro.umontreal.ca/~mignotte/IFT6150/Articles/SLIC_Superpixels.pdf

SLIC Partitioning and Distribution Modeling

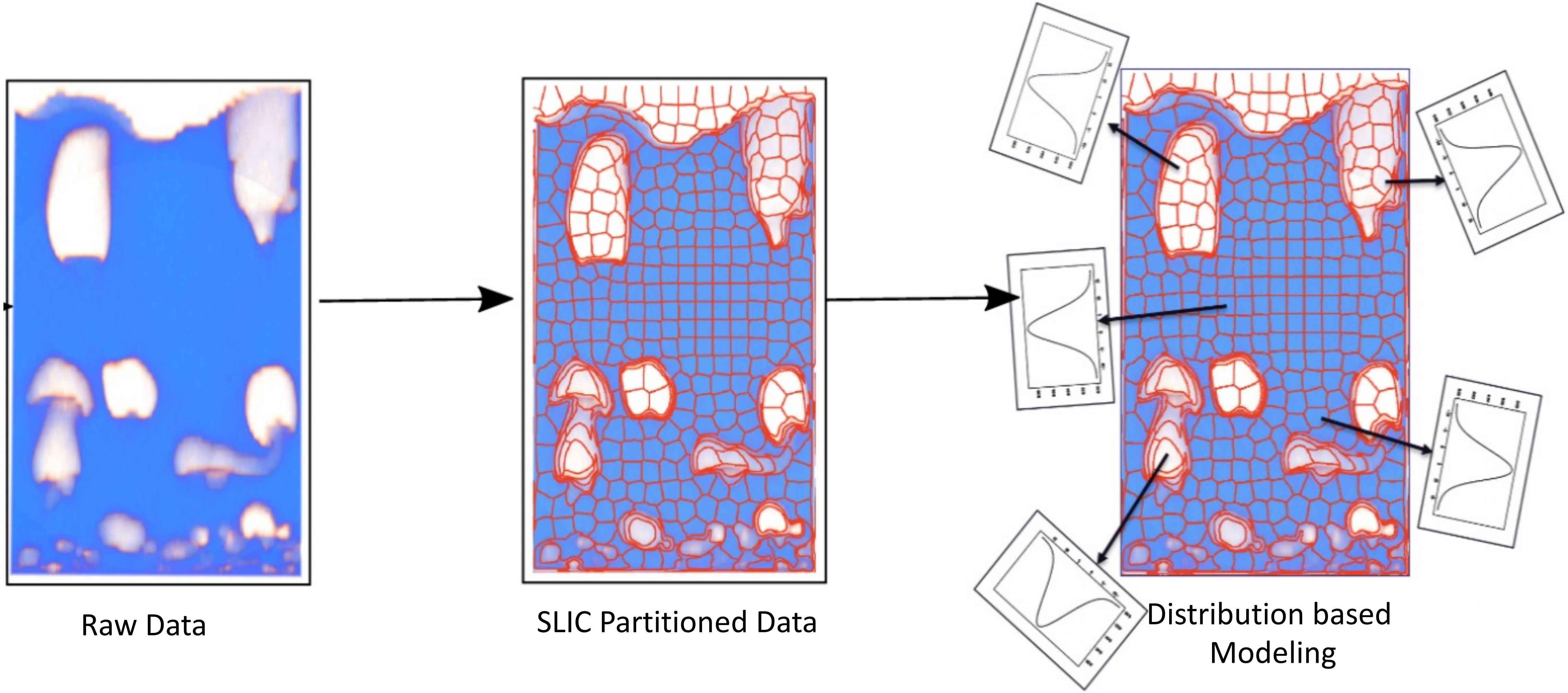
- Generate partitions based on data homogeneity
- Summarize each partition using a

hybrid distribution scheme

  - A single Gaussian or a mixture of Gaussians
  - Reduce data size and a compact representation



SLIC Partitioning and Distribution Modeling



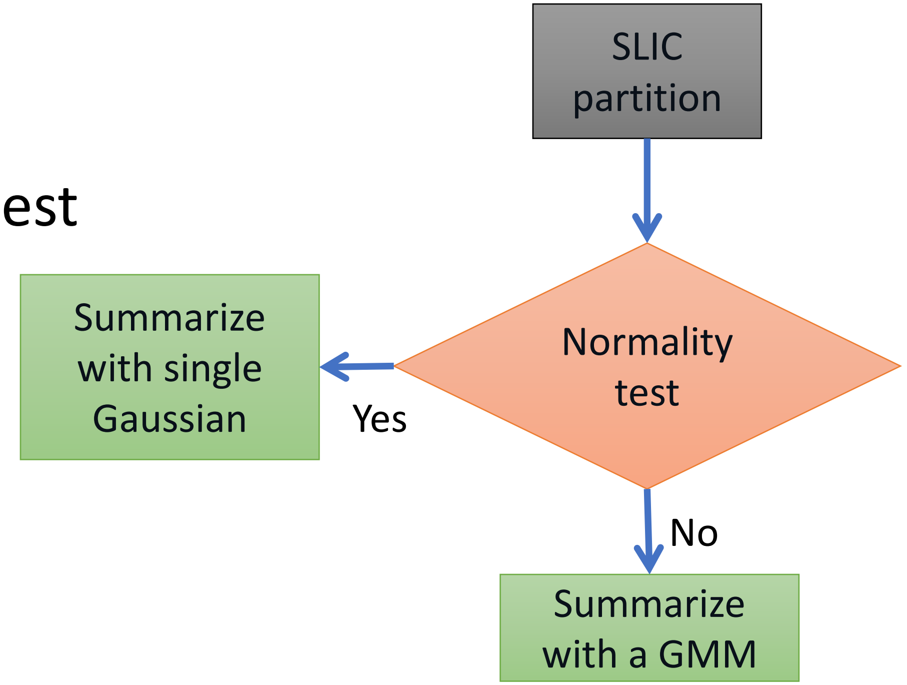
Raw Data

SLIC Partitioned Data

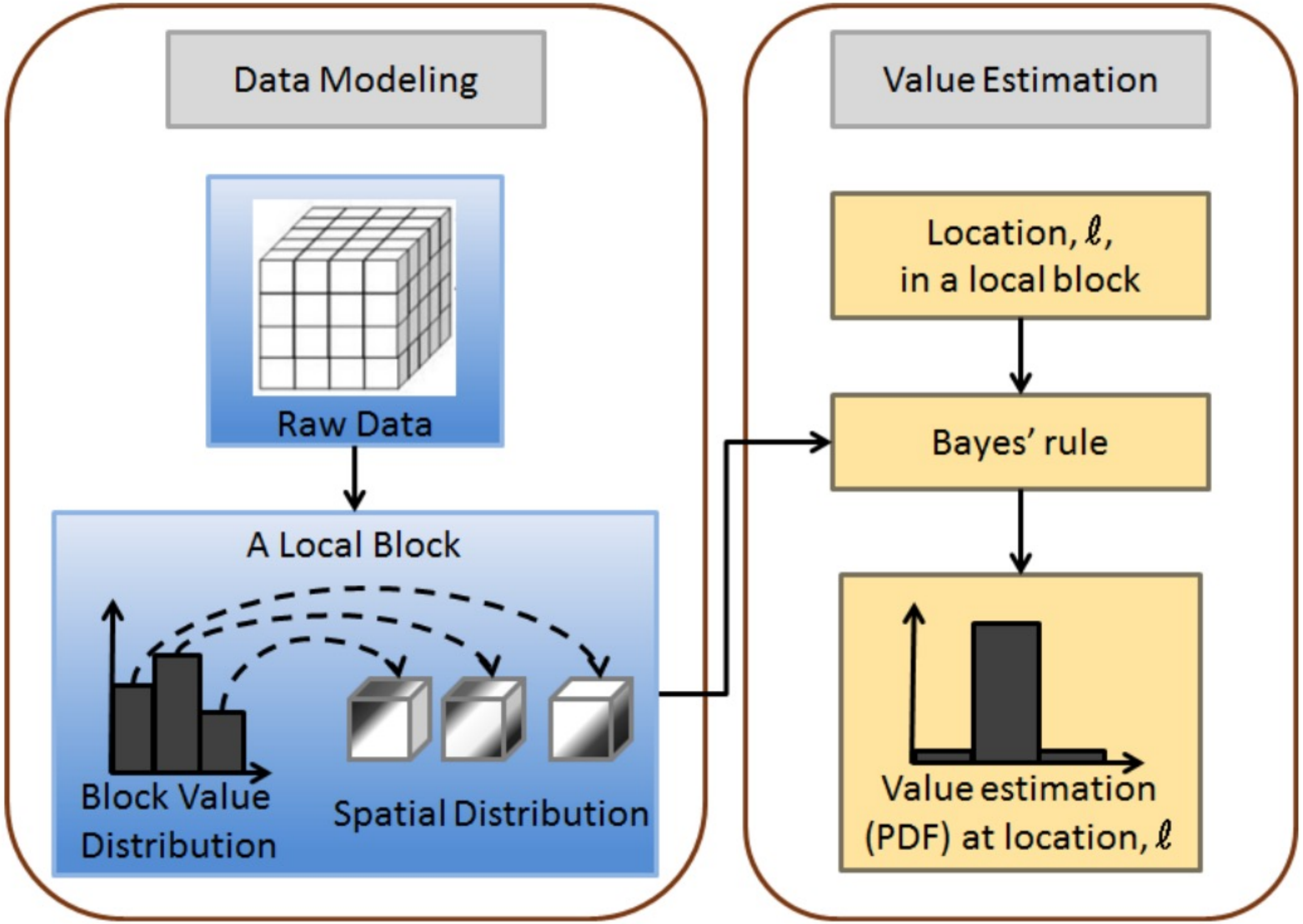
Distribution based
Modeling

Hybrid Distribution-based Summarization

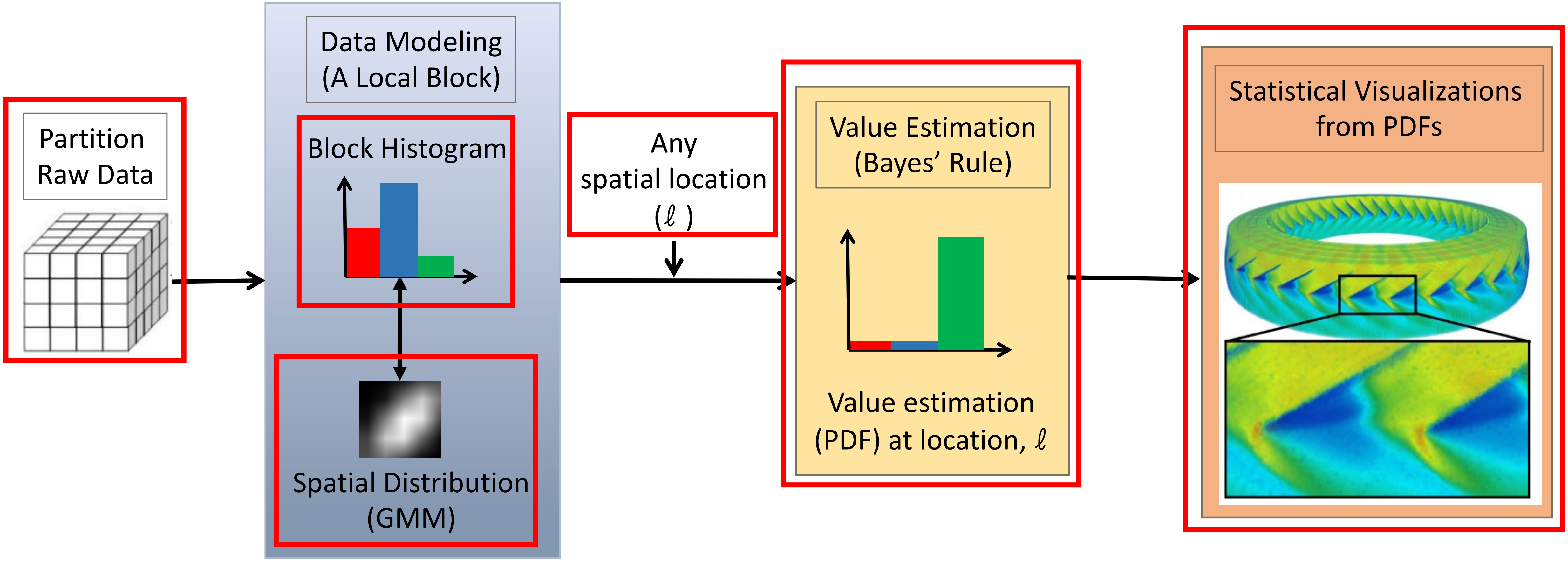
- Each SLIC partition is summarized using
  - Either a single Gaussian or a mixture of Gaussians (GMM)
  - The decision is made based on a Normality test
  - Expectation Maximization (EM) for GMM parameter estimation
- A hybrid distribution-based data set
  - Compact statistical representation
  - Significantly smaller than the raw data



Spatial Distribution-augmented Statistical Data Summarization

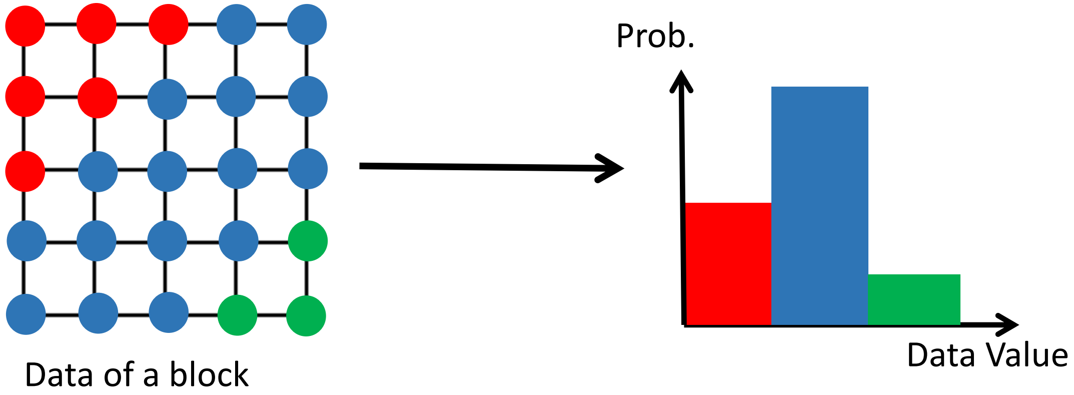


Spatial Distribution-augmented Statistical Data Summarization


Partition Raw Data
Data Modeling (A Local Block)
Block Histogram
Spatial Distribution (GMM)
Any spatial location (ℓ)
Value Estimation (Bayes' Rule)
Value estimation (PDF) at location, ℓ
Statistical Visualizations from PDFs

Spatial Distribution-augmented Statistical Data Summarization

- Block histogram summarizes data samples in a block
  - Bin b_i represents a continuous data value range $[L_{b_i}, U_{b_i}]$
  - $H(b_i) = \frac{N(b_i)}{\sum_{k=0}^{B-1} N(b_k)}$ (normalized frequency of bin $\mathrm{b_i}$)
    - $N(b_k)$: number of grid points whose values are in range $[L_{b_k}, U_{b_k}]$

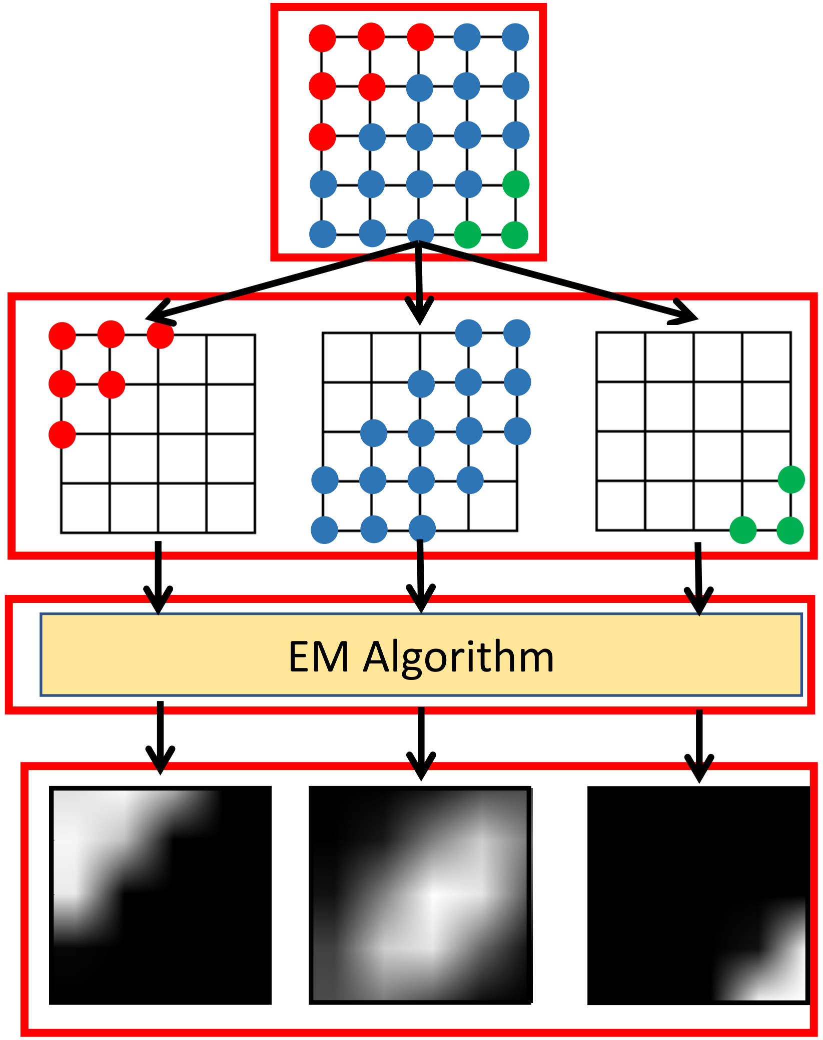


Spatial Distribution-augmented Statistical Data Summarization

- Block histogram does not retain samples' location information

- Each bin attaches a spatial distribution: $\{S_0, S_1, \ldots S_{B-1}\}$
  - S_{b_i} : maps a spatial location (ℓ) to a probability
  - Estimated by a multivariate GMM (Spatial GMM)

- Spatial GMM modeling
  - Collect coordinates of all grid points assigned to bin b_i
  - Use EM algorithm to estimate the parameters of the GMM
  - Repeat the process for each bin



Slides adapted from Prof. Ko-Chih Wang

Spatial Distribution-augmented Statistical Data Summarization

- The complexity of the spatial distribution of each bin are quite different
    - Increases storage overhead if we use too many Gaussian components
    - May have insufficient modeling quality if fewer Gaussian components are used

- Use an adaptive scheme to determine the best number of Gaussian components
    - Bayesian Information Criterion (BIC) evaluates the fitting quality
    - Given an upper bound of number of Gaussian components
    - The number of Gaussian components with the best BIC is selected

Bayesian Information Criterion (BIC)

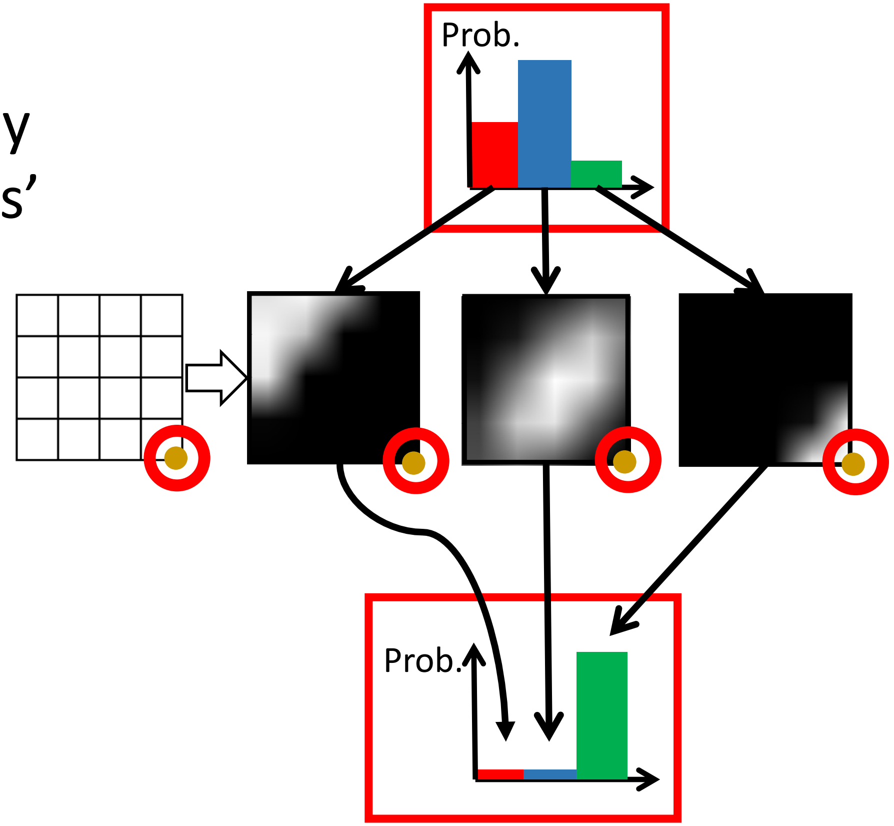
- BIC is a criterion of model selection given different model configurations
- Lower values of BIC are preferred
- Tries to balance between the number of model parameters (model complexity) and the amount of overfitting

$$BIC = k * ln(n) - 2 * \ln(L)$$

- L is the maximized value of likelihood function
- n = sample size
- k = number of model parameters

Spatial Distribution-augmented Statistical Data Summarization: Value Estimation

- A probability density function (PDF) at a given location ℓ
  - Possible values at ℓ and their associated probability
  - Combine the histogram and spatial GMMs by Bayes' rule



- Bayes' rule
  - The prior is adjusted by the related evidences
    - $P_{\ell}(b_i) = \frac{SGMM_{b_i}(\ell) * H(b_i)}{\sum_{k=0}^{B-1} SGMM_{b_k}(\ell) * H(b_k)}$
    - Prior: block histogram
    - Evidences: probabilities of spatial GMMs at ℓ
    - Posterior: estimated PDF at ℓ

Data/Storage Reduction

- SLIC-based partitioning + hybrid modeling Technique
  - 100 GB raw data could be reduced to 10.8 GB distribution-based data
  - Reduced data needs only ~10% of the raw data storage
- Spatial GMM + value distribution-based modeling
  - 10.61 GB raw data could be reduced to 0.152 GB distribution-based data (32^3 block size)
  - Reduced data needs only 1.39% of the raw data storage

What Can We Do With the Distribution-based Reduced Data?

- Reconstruct the full resolution data
  - Bayes' Rule-based method of reconstruction using Spatial distribution models
  - Monte Carlo sampling for SLIC-based hybrid distribution models
- Distribution-based feature search
- Feature extraction and tracking
- Many more ……

Box Muller Transform

- Box Muller Transform is used to generate pairwise independent Standard Normally distributed values
- Let U_1 and U_2 are independent samples chosen from a Uniform distribution defined over (0,1)
- Compute:

$$Z_0 = R\cos(\Theta) = \sqrt{-2\ln U_1}\,\cos(2\pi U_2)$$

$$Z_1 = R\sin(\Theta) = \sqrt{-2\ln U_1}\,\sin(2\pi U_2)$$

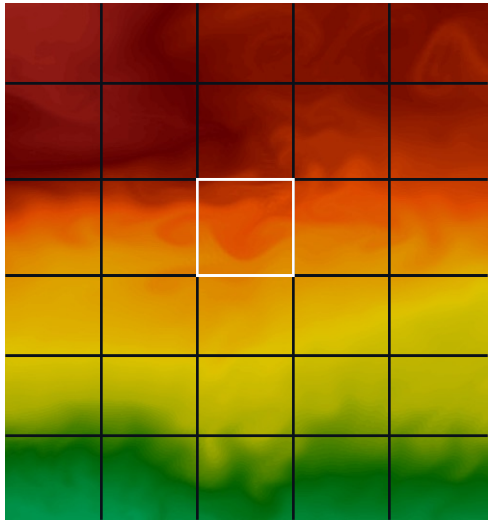
- We can show that Z_0 and Z_1 are coming from Standard Normal distributions

Box Muller Transform for GMM

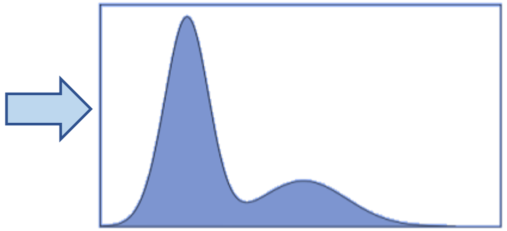
- Box Muller Transform gives us samples that follow Standard Normal distributions
- But we have Gaussian mixture models (GMM)
- So, apply the following steps to generate samples from a GMM:
  - Generate a random number r between 0 - 1
  - Find the Gaussian component for which $\sum_k weight_k \geq r$
  - Use Box Muller Transform to sample a value from the selected Gaussian component

Monte Carlo Sampling-based Reconstruction

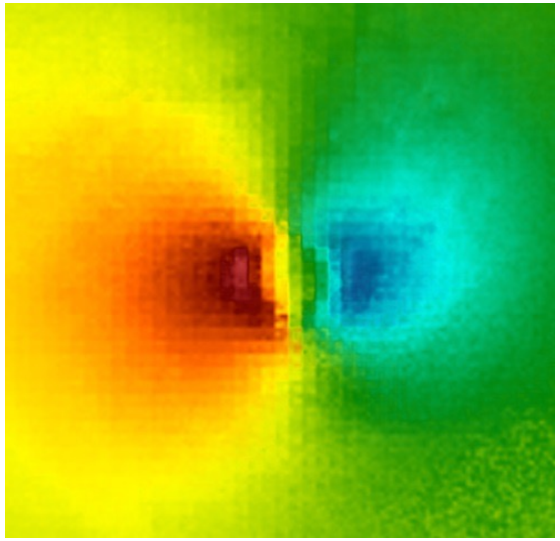
- Monte Carlo Sampling is used to reconstruct the full data from distribution models
  - Box Muller Transform on GMM and Gaussians
- Regular partitioning is not optimal
  - Does not consider inherent data coherency



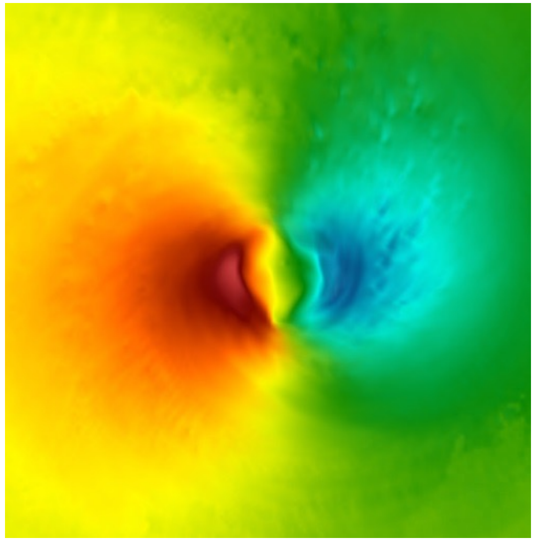
Regular partitioning



High value variation in partition distributions



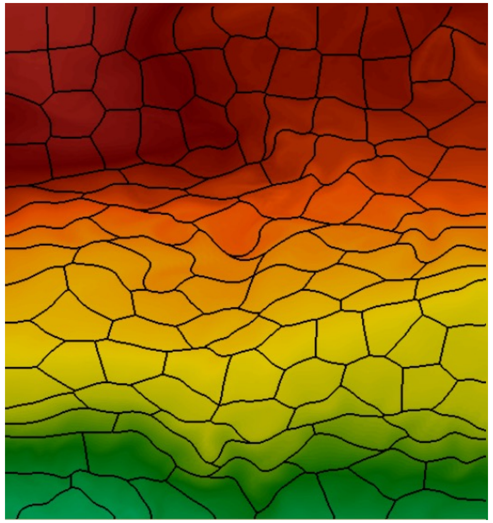
High sampling error in reconstructed data



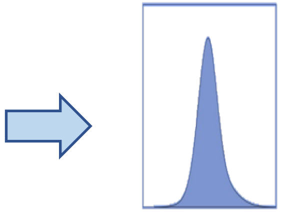
Ground truth data

Monte Carlo Sampling-based Reconstruction

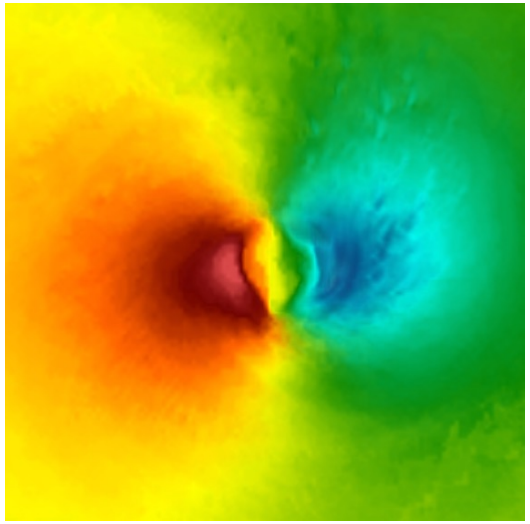
- Generate partitions based on data homogeneity
- Simple Linear Iterative Clustering (SLIC)
  - Produces irregular shaped partitions/clusters
  - Value variation inside partitions is minimized
  - Reduced sampling error



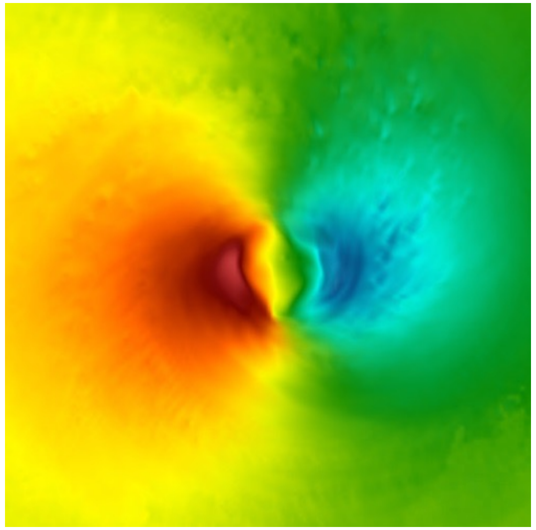
SLIC partitioning



Low value variation in partition distributions

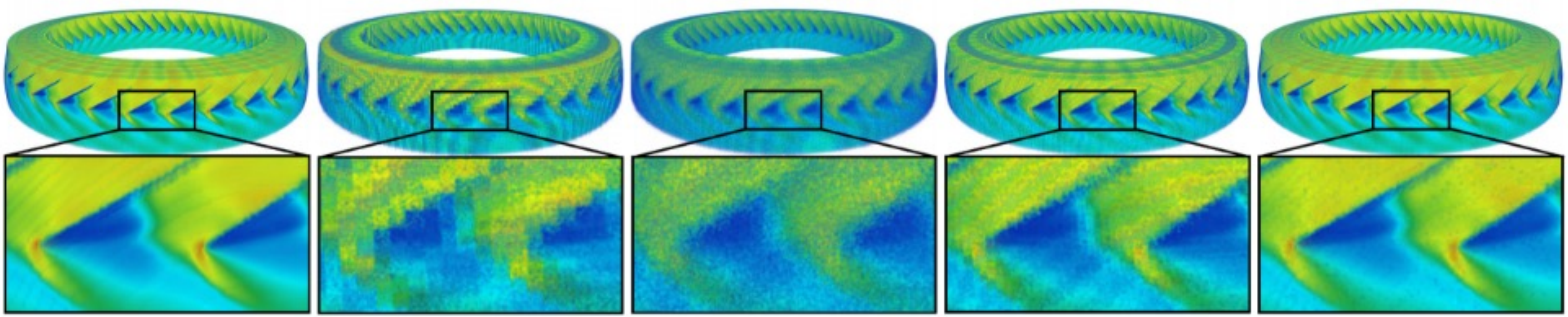


Smooth reconstruction using SLIC partitioning



Ground truth data

Bayes’ Rule-based Method of Reconstruction using Spatial Distribution



Raw data
Size: 10871MB
Resolution: 2545x2545x440

Block histogram
Size: 131.4MB
Block Size: 22^3

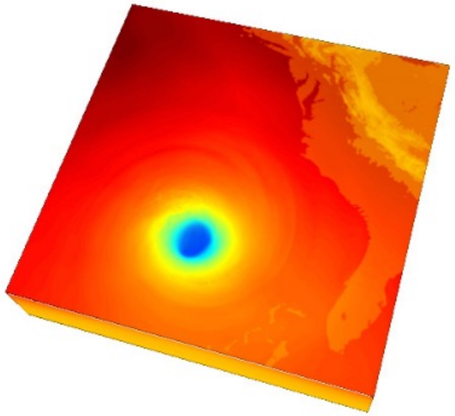
Block histogram w/ interpolation
Size: 131.4MB
Block Size: 22^3

Block GMM
Size: 163.71MB
Block Size: 10^3

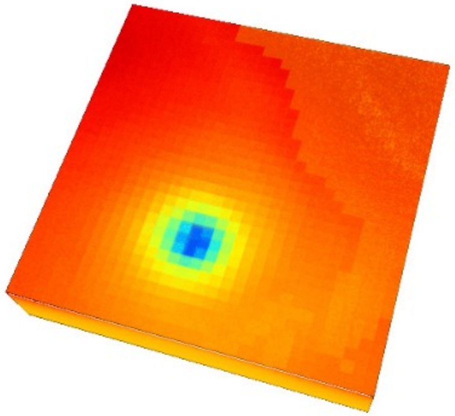
Our approach
Size: 151.54MB
Block Size: 32^3

Spatial Distribution

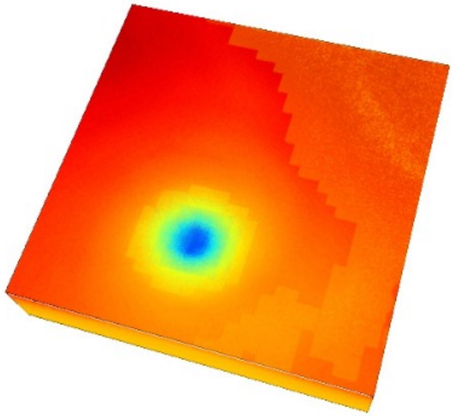
Bayes' Rule-based Method of Reconstruction using Spatial Distribution



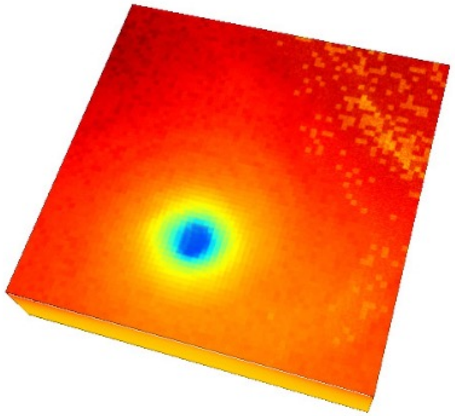
(a) Rendering from raw data (100MB)



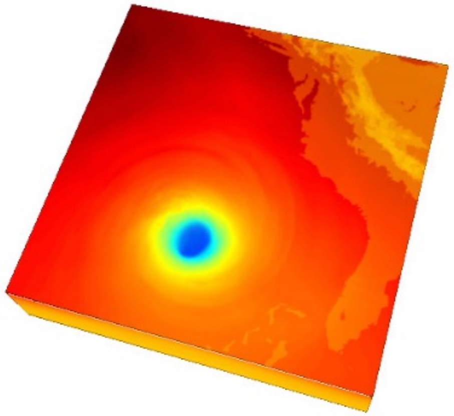
(b) Block histogram (7.93MB)



(c) Block Hisotgram w/ interpolation (7.93MB)

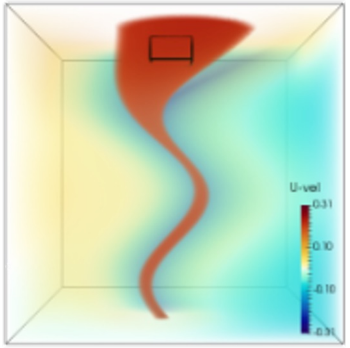


(d) Block GMM: (7.03MB)

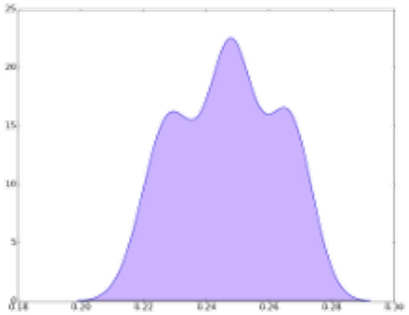


Spatial Distribution (6.76MB)

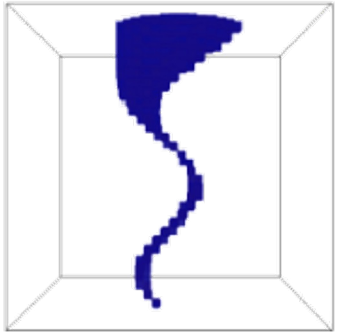
Distribution Similarity-based Feature Search



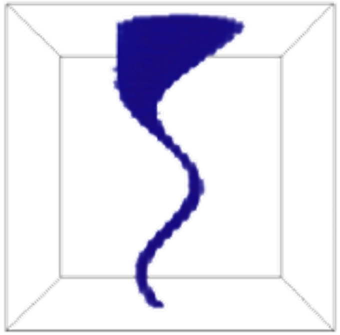
Selection of feature



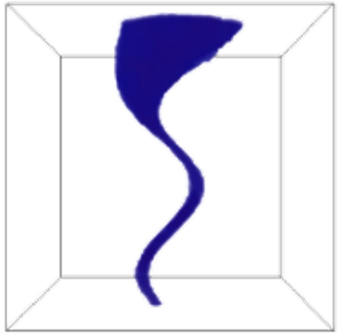
Feature distribution to be searched



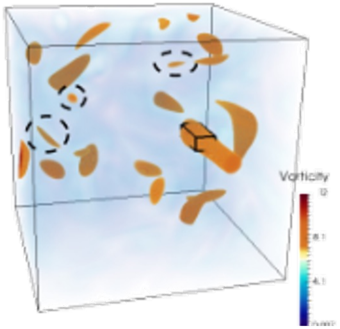
Regions with similar distributions for regular partitioning



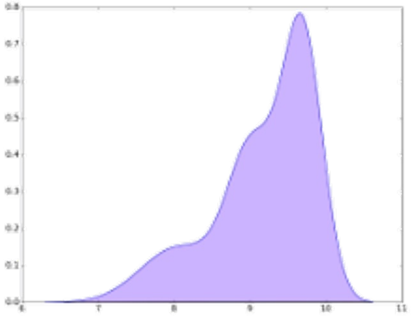
Regions with similar distributions for K-d tree partitioning



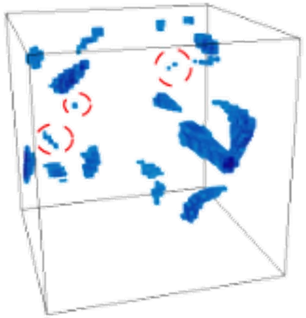
Regions with similar distributions for SLIC partitioning



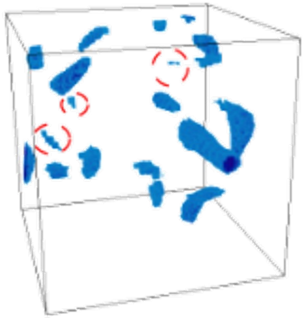
Selection of feature



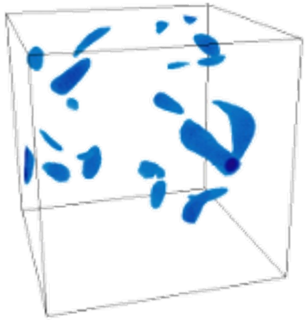
Feature distribution to be searched



Regions with similar distributions for regular partitioning



Regions with similar distributions for K-d tree partitioning



Regions with similar distributions for SLIC partitioning

How to Compute Similarity Between Distributions?

- **Kullback–Leibler divergence:** Amount of work needed to convert one distribution to other

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

- **Earth Mover's distance:** Minimum cost of turning mass of one distribution to another. For 1-D distributions, this can be shown as a sum of the differences between their cumulative distributions

$$EMD(X,Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)|\, dx$$

- **Bhattacharya distance:**

$$D_B(P,Q) = -\ln(BC(P,Q)) \quad \text{where} \quad BC(P,Q) = \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}$$

P and Q are probability distributions defined on the same domain $\mathcal{X}$