Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

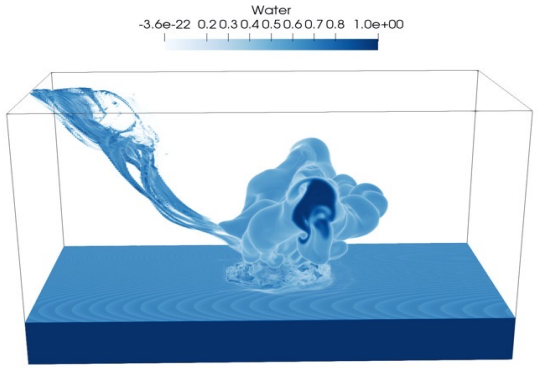
email: soumyad@cse.iitk.ac.in

CS661 Quiz

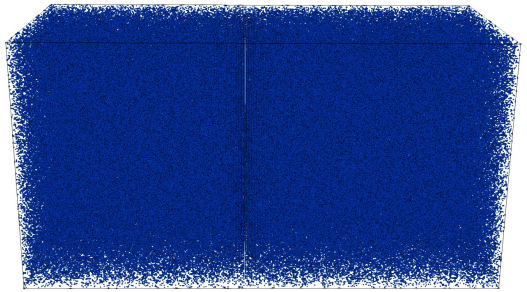
- Date: April 7, 2025
- Time: 2-3pm
- Location: LH-20
- Syllabus: Everything starting from Lecture 13 (Status Refresher) up to topics discussed until April 2nd

Data Driven Importance-based Sampling

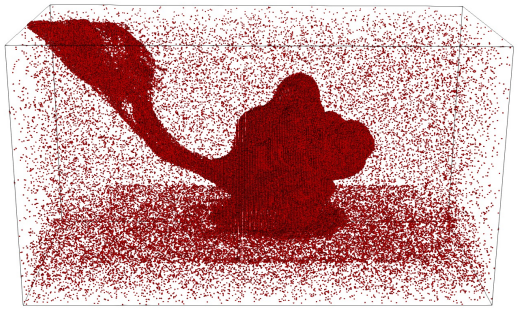
- Is SRS always the preferred choice?
  - No, depends on the application and goal
- We may need to preserve certain data more accurately than others
  - Features (regions of interest) in the data
  - All data points are NOT equally likely to get picked
- Sample data points based on an importance criterion to achieve the desired objective
- How to define and construct an importance function?



Original data showing the important region



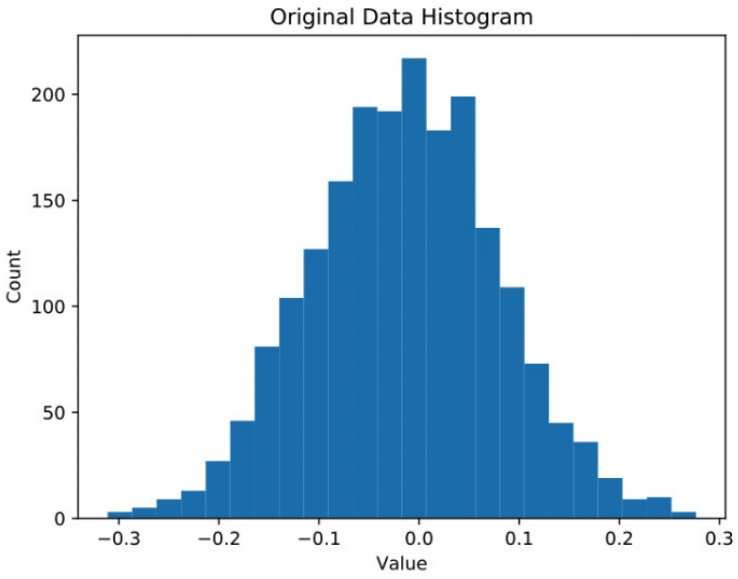
What SRS will give us



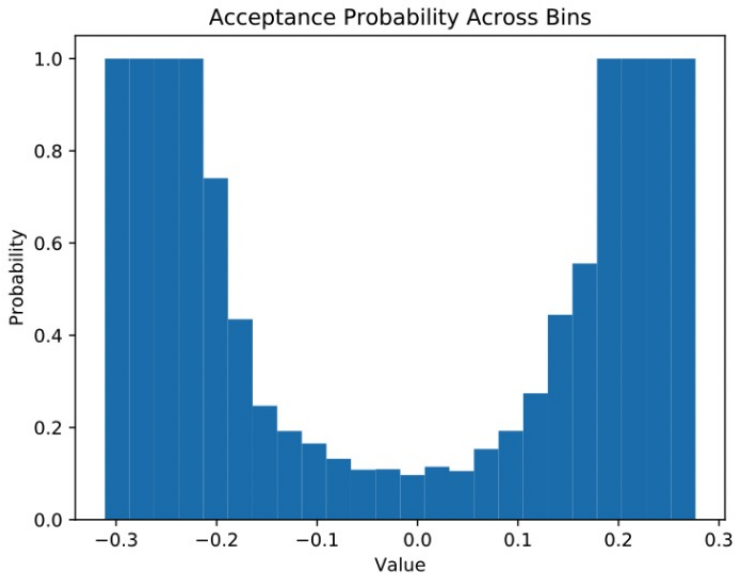
What we want, more samples from the important region

Data Driven Importance-based Sampling

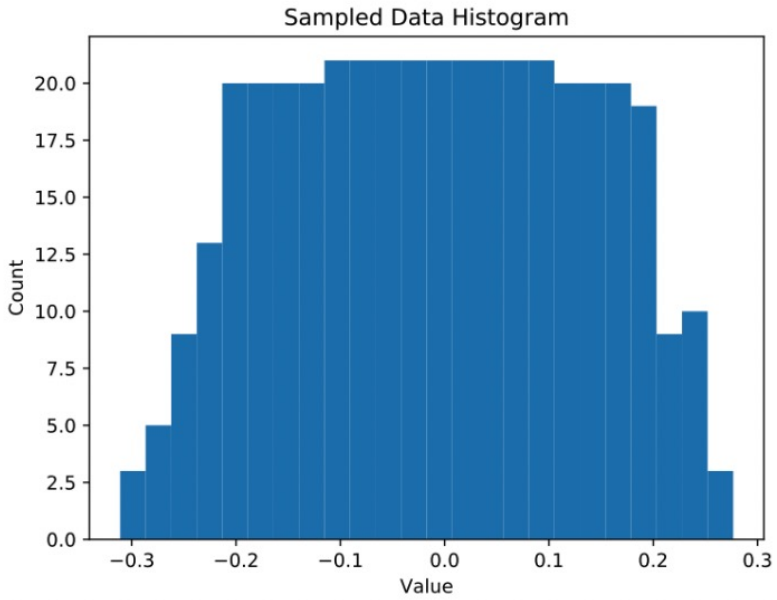
- How to design an appropriate importance function/criterion?
- Idea: Assign more importance to the data points with low value occurrence/probability



Data Histogram

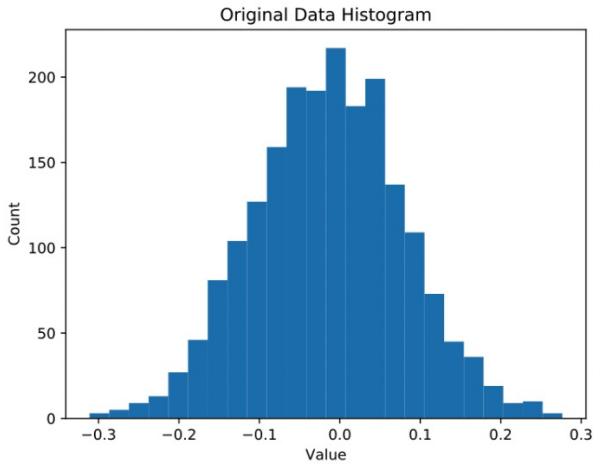


Acceptance Function

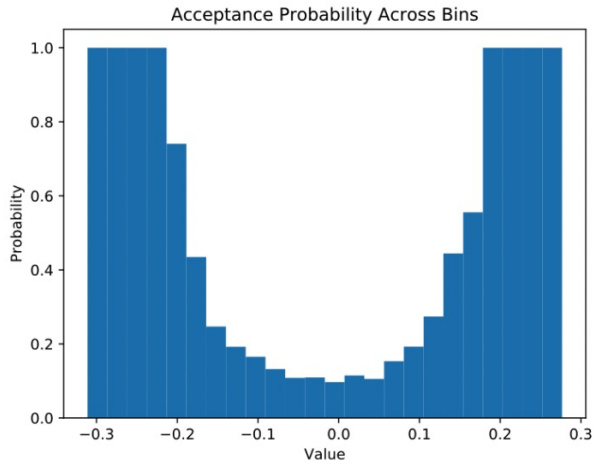


Data Histogram after Sampling

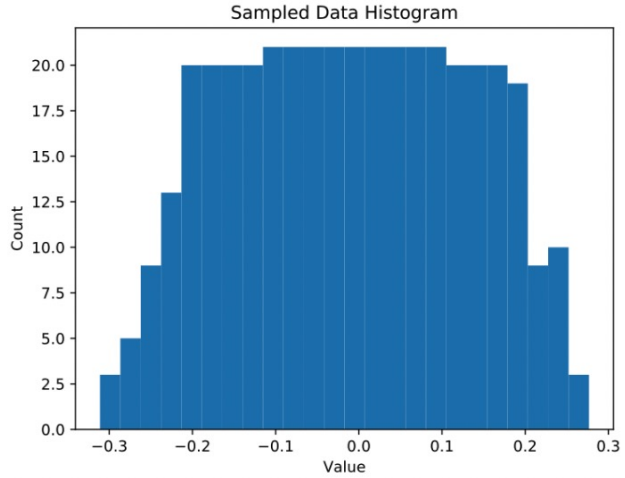
Data Driven Importance-based Sampling



Data Histogram



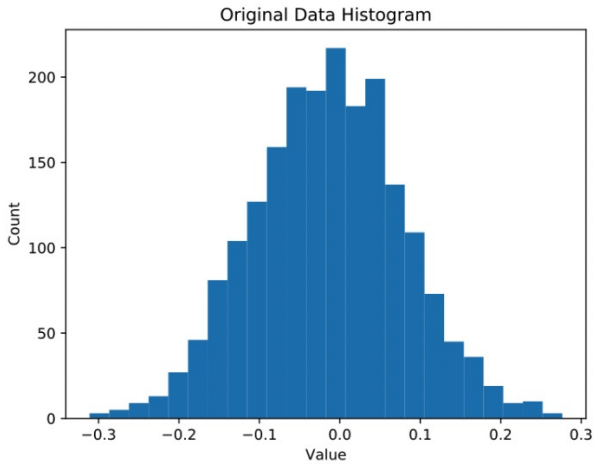
Acceptance Function



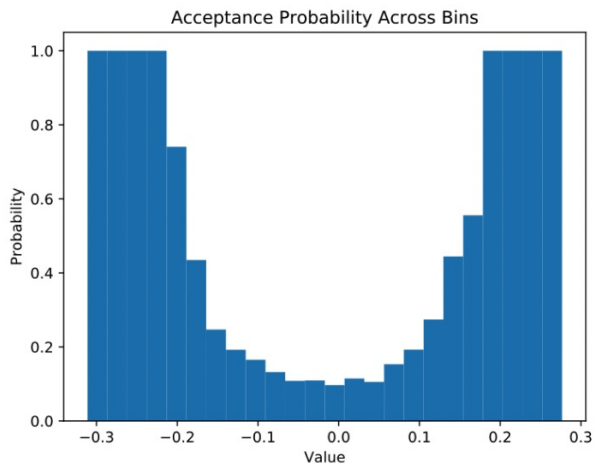
Data Histogram after Sampling

- Similar to performing Entropy maximization
- Entropy: In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

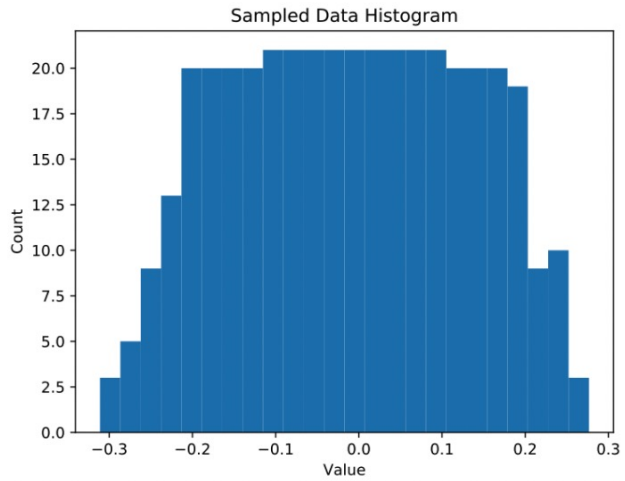
Data Driven Importance-based Sampling



Data Histogram



Acceptance Function

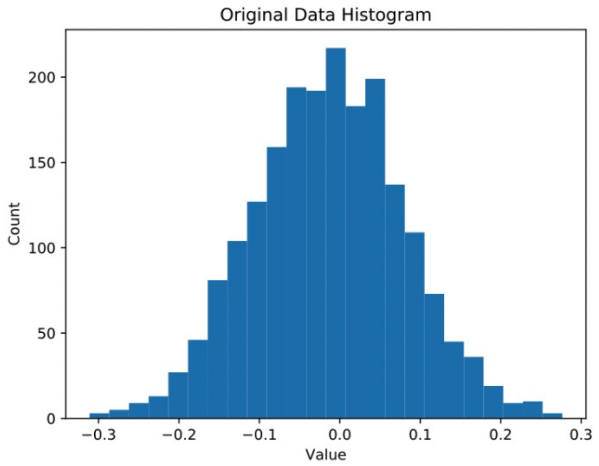


Data Histogram after Sampling

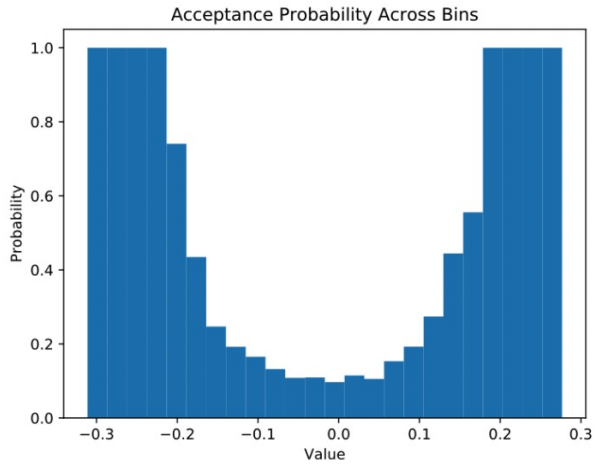
$$\text{Entropy} = H(X) = -\sum_x p(x) * \log(p(x)) = E[-\log(p(X))]$$

- Base 2 with log gives the result in units of *bits*

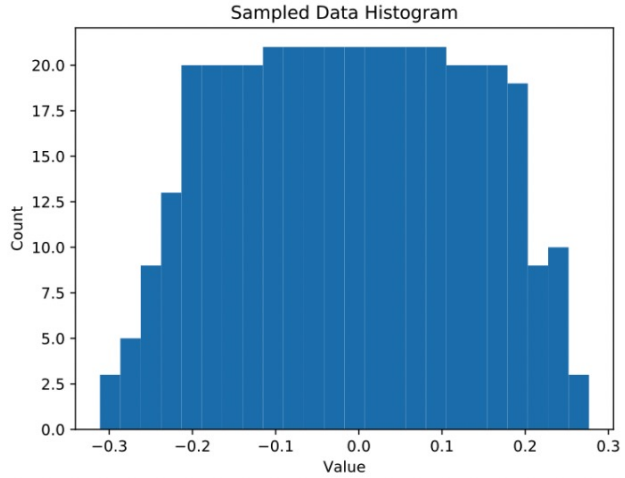
Data Driven Importance-based Sampling



Data Histogram



Acceptance Function



Data Histogram after Sampling

$$\text{Entropy} = H(X) = -\sum_x p(x) * \log(p(x)) = E[-\log(p(X))]$$

Uniform distribution has maximum entropy

Importance-based Sampling Pseudo Code

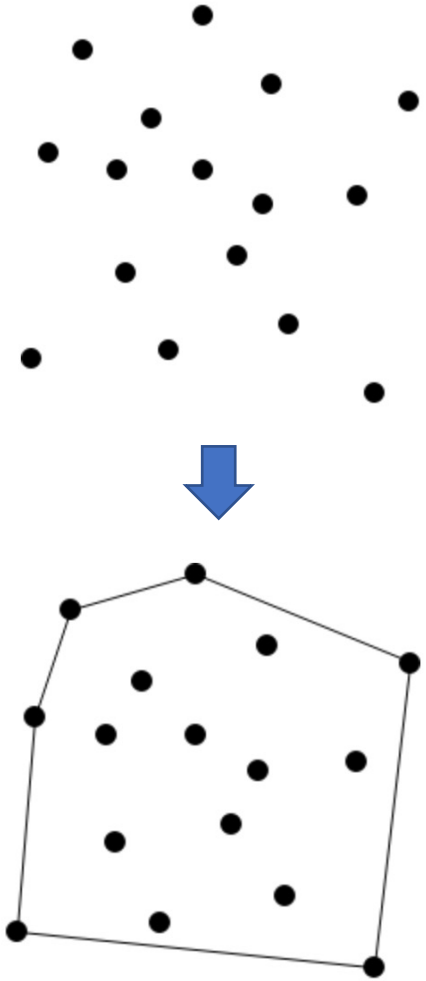
- How do we create an acceptance function that is implicitly going to maximize entropy of the data value distribution?

Input: Sim generated data at each time step
Output: Acceptance histogram
n_k = samples to be taken from full data
N = total points from full data
nbins = number of bins
$n_{samps} = n_k \div nbins$: samples to be taken from each bin
remaining-samples = N
H = CreateHistogram(Data,nbins)
count,bin-edges = SortBinsAccordingToTheirCount(H)

while *not all bins are visited* **do**
  i=0;
  if *count[i]* < n_{samps} **then**
    samples-taken = count[i];
  else
    samples-taken = n_{samps};
  end
  P_i = samples-taken ÷ count[i]
  remaining-samples = remaining-samples - samples-taken
  remaining-bins = nbins - i
  n_{samps} = remaining-samples ÷ remaining-bins
  i=i+1;
end
Use P_is as the probabilities for the corresponding bins of the acceptance histogram

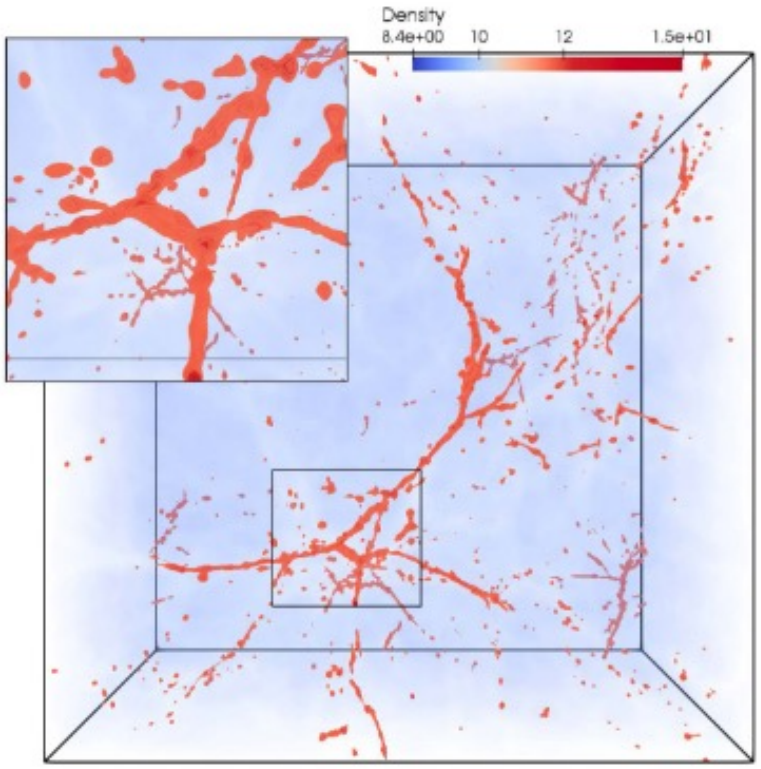
Reconstruction From Sampled Data

- A naïve reconstruction method is nearest neighbor interpolation
  - Fast execution but quality is not good
- Linear interpolation-based reconstruction
  - Create a 3D convex hull using all sampled points
    - a set of points in a plane (or higher-dimensional space) is the smallest convex shape that encloses all the given points.
  - Convert the points into a polygonal mesh using Delaunay triangulation
  - Next, for each grid point in the reconstruction grid, the value is obtained by linearly interpolating scalar values from the vertices of the simplex that encloses the current grid point

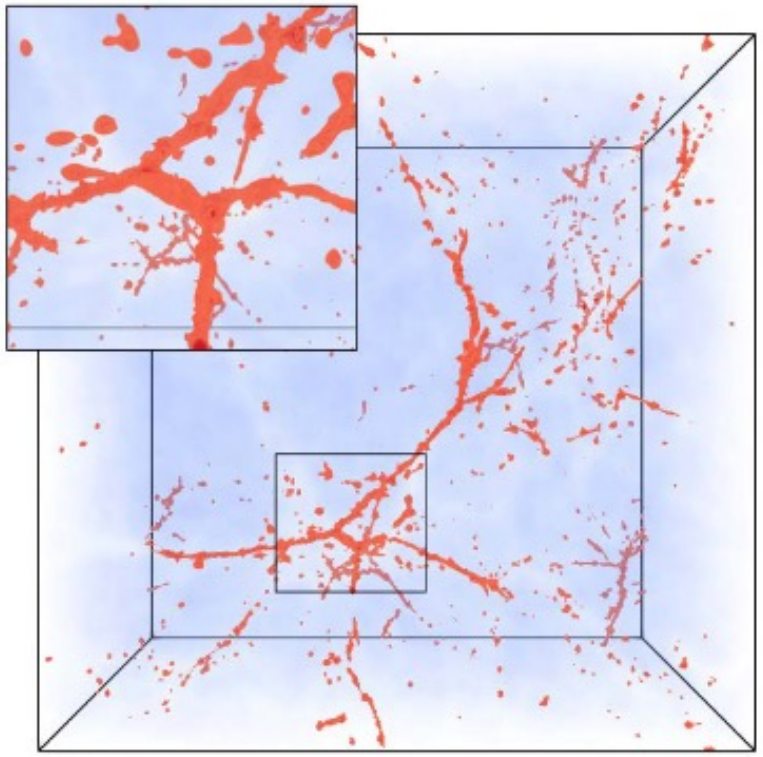


Convex hull

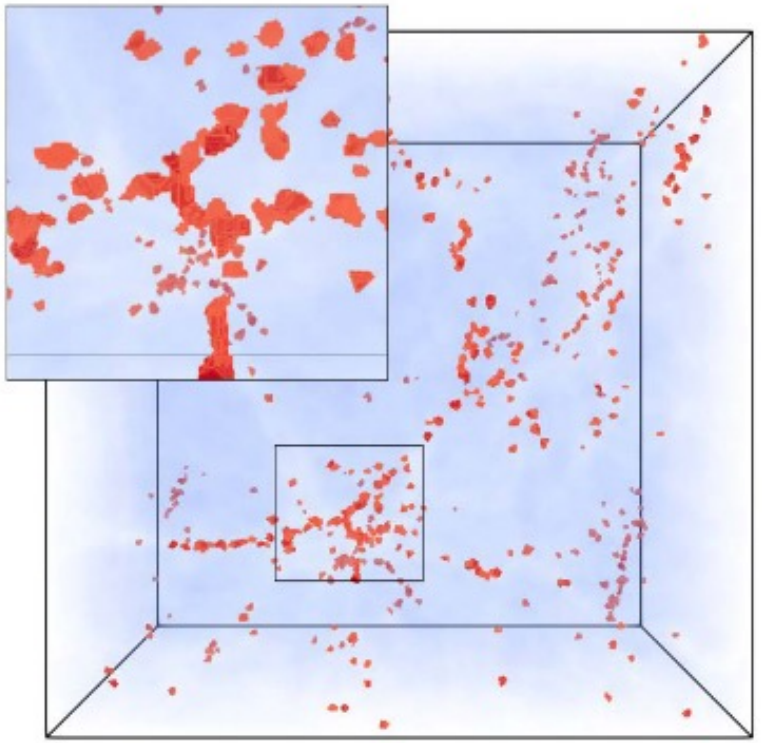
Results of Importance Sampling



Original data (volume rendered)

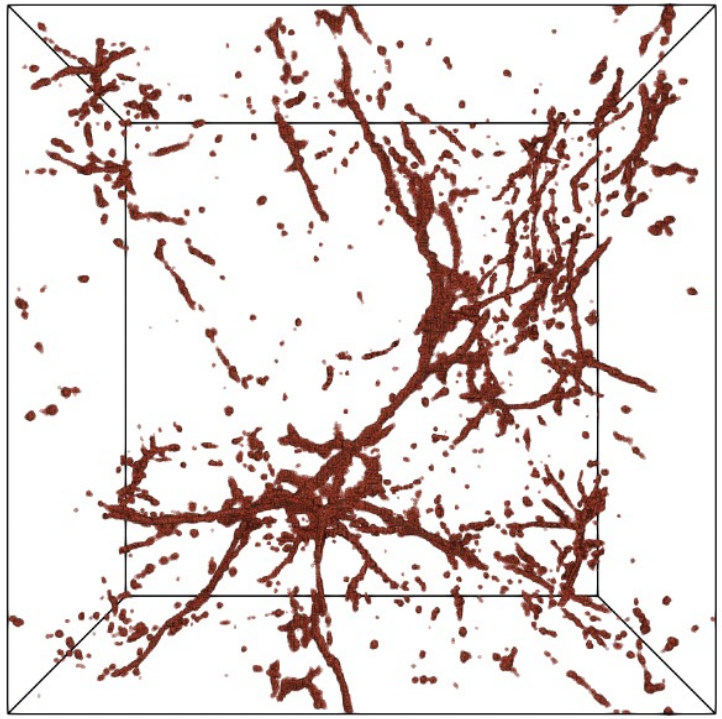


Reconstruction using Importance-based Sampling (1% points sampled)

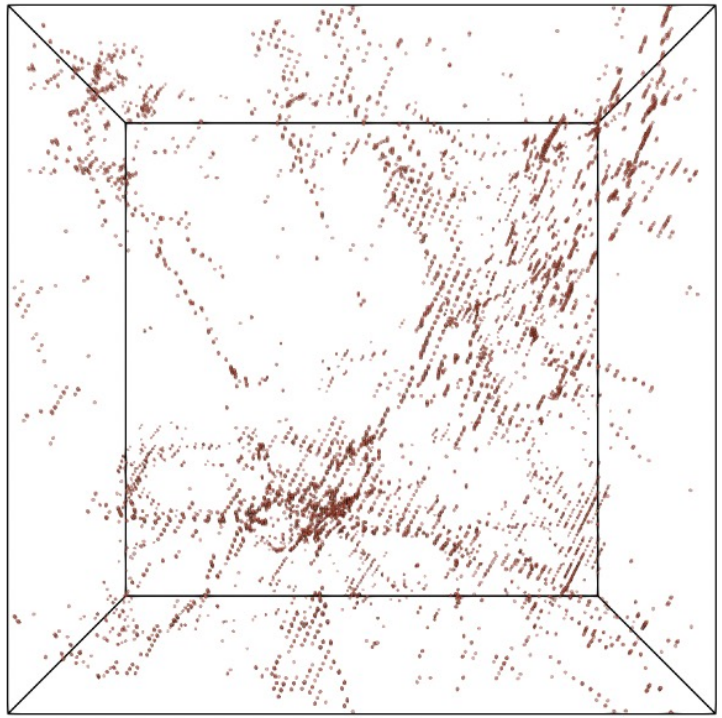


Reconstruction using stratified random Sampling (1% points sampled)

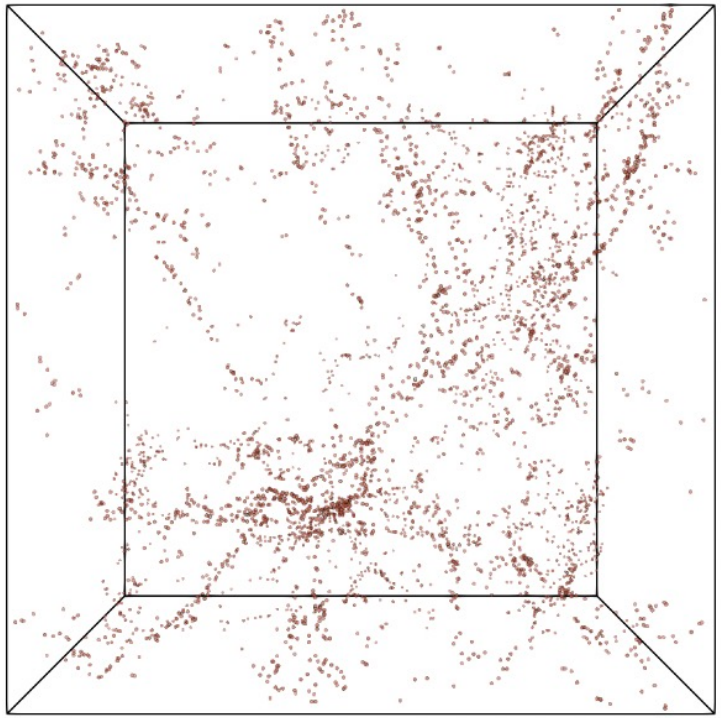
Results of Importance Sampling



Importance-based Sampling
(0.5% points sampled)



Systematic Sampling
(1.5% points sampled)

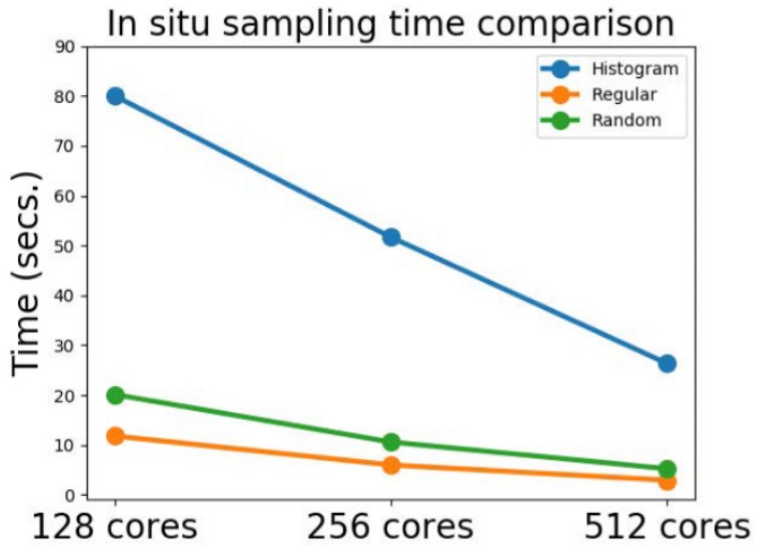


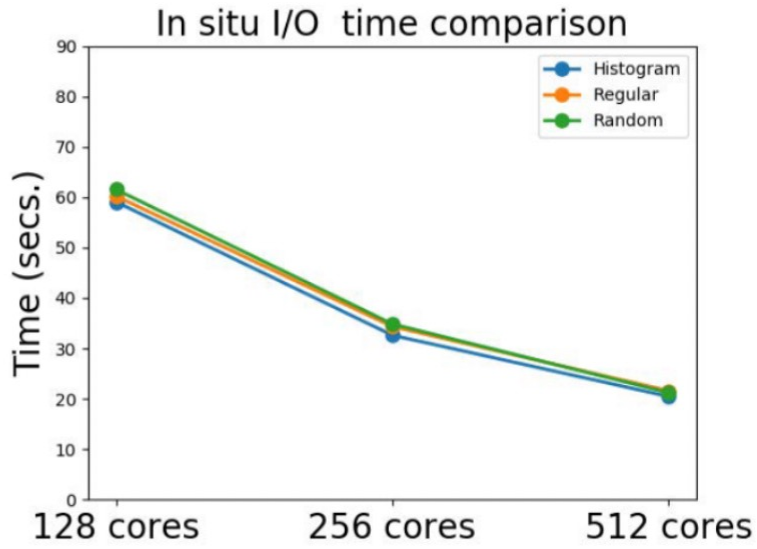
Stratified Random Sampling
(0.5% points sampled)

Quality and Performance Evaluation

	Hist. Samp	StRand. Samp	Regular Samp
Isabel data	0.978	0.921	0.961
Nyx data	0.987	0.865	0.881

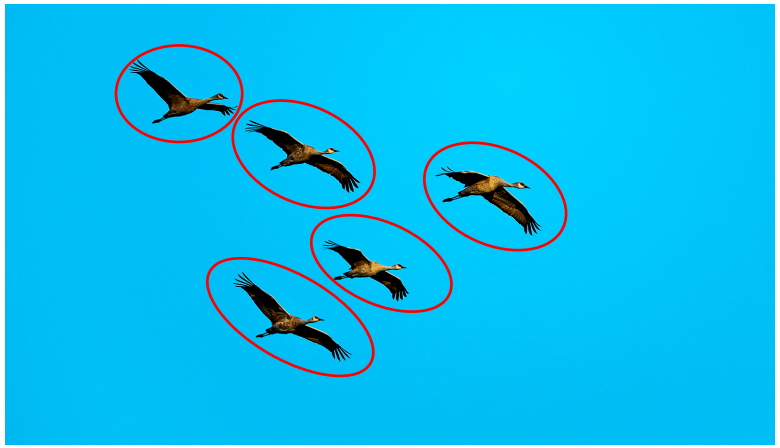
Showing linear correlation values with the original data
(Higher is better)



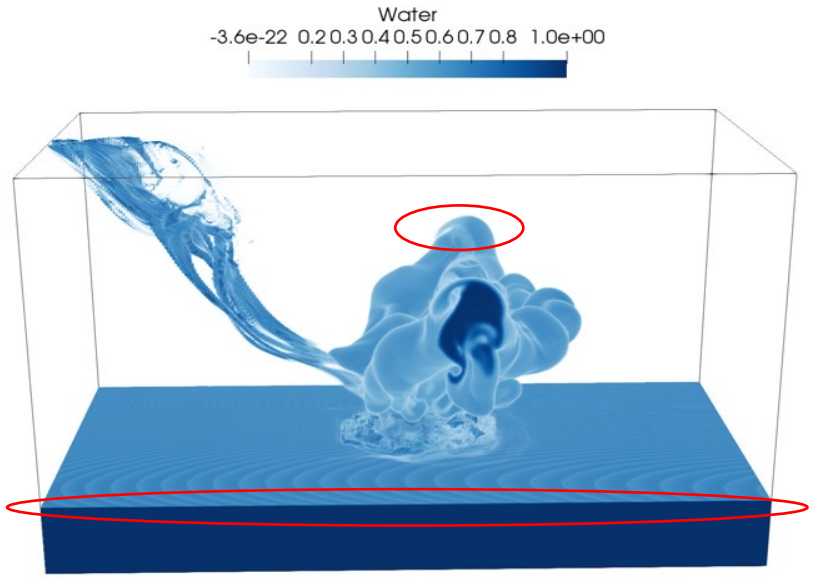


Improved Importance-based Sampling

- How to select a subset of data points that are informative and preserve features?



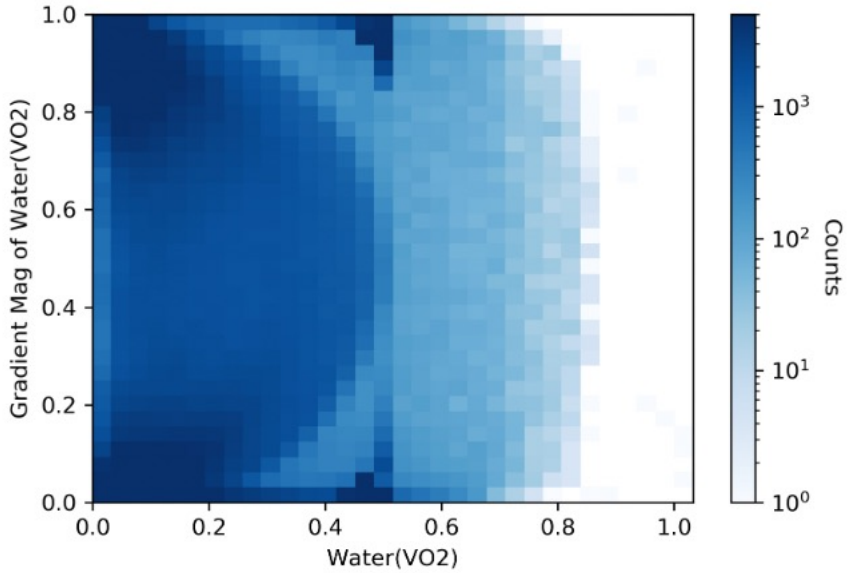
Informative regions are the birds, and the homogeneous sky is background.



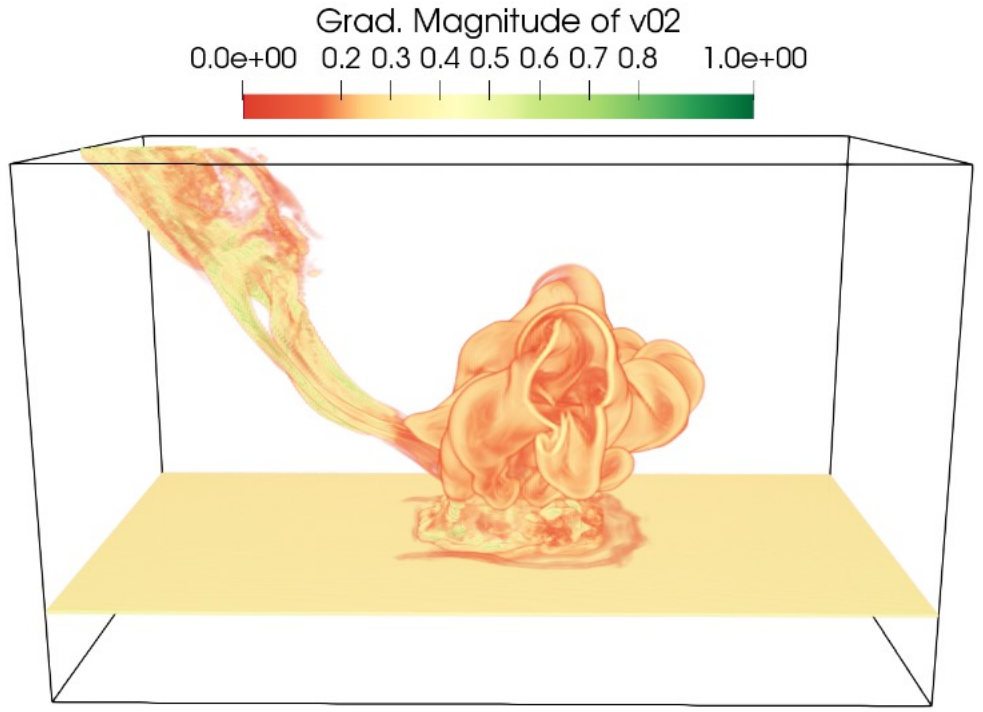
High gradient regions are important

- Analyze the data value occurrence and local smoothness/gradient

Smoothness/Gradient Visualization



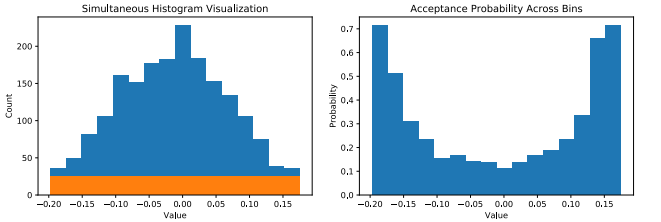
2D histogram of water fraction and its gradient magnitude



Volume visualization of gradient of water fraction.

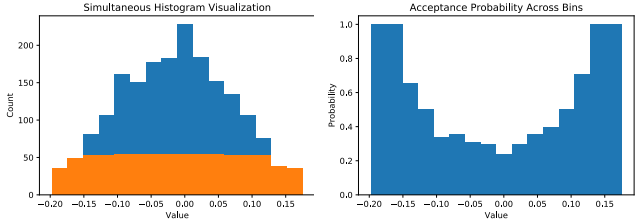
Joint Multi-Criteria Sampling

Value-based Sampling



Blue: Original Hist.
Orange: selected points

Acceptance function for the histogram



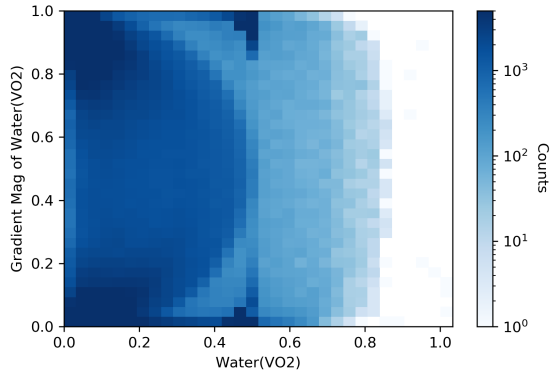
Blue: Original Hist.
Orange: selected points

Acceptance function for the histogram

```
while j < B do
    c_j ← H[j] ;
    if c_j < C then
        I_F[j] = c_j;
        M ← M − c_j;
        B ← B − 1;
        C ← M/B;
        j = j + 1;
    else
        for k to B by 1 do
            I_F[k] ← C;
        end
        break
    end
end
```

H – Data value Histogram
I_F – Importance function
M – Total num. of points to be selected
B – Num. of histogram Bins
C – Expected num. of samples per bin

Smoothness-based Sampling



2D Histogram of data value and Gradient Magnitude

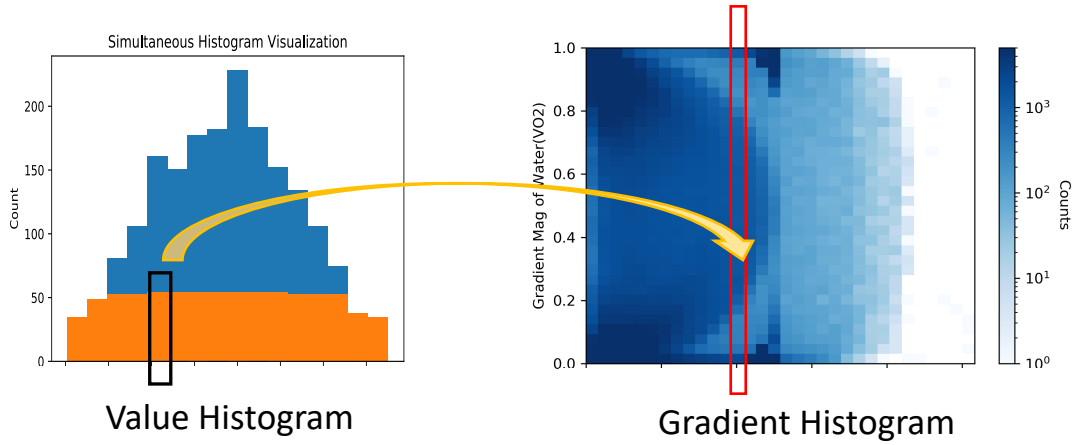
```
G ← computeGradient(D);
G_mag ← computeGradientMagnitude(G);
H ← histogram(G_mag, N, B);
I_F ← zeros(B);
C ← M/B ; // Expected number of samples
j = B − 1;
while j >= 0 and M > 0 do
    c_j ← H[j] ;          // Count in bin j
    I_F[j] = c_j;
    M = M − c_j;
    j = j − 1;
end
```

H – Data value, Grad. Mag. Histogram
I_F – Importance function
M – Total num. of points to be selected
B – Num. of histogram Bins
C – Expected num. of samples per bin

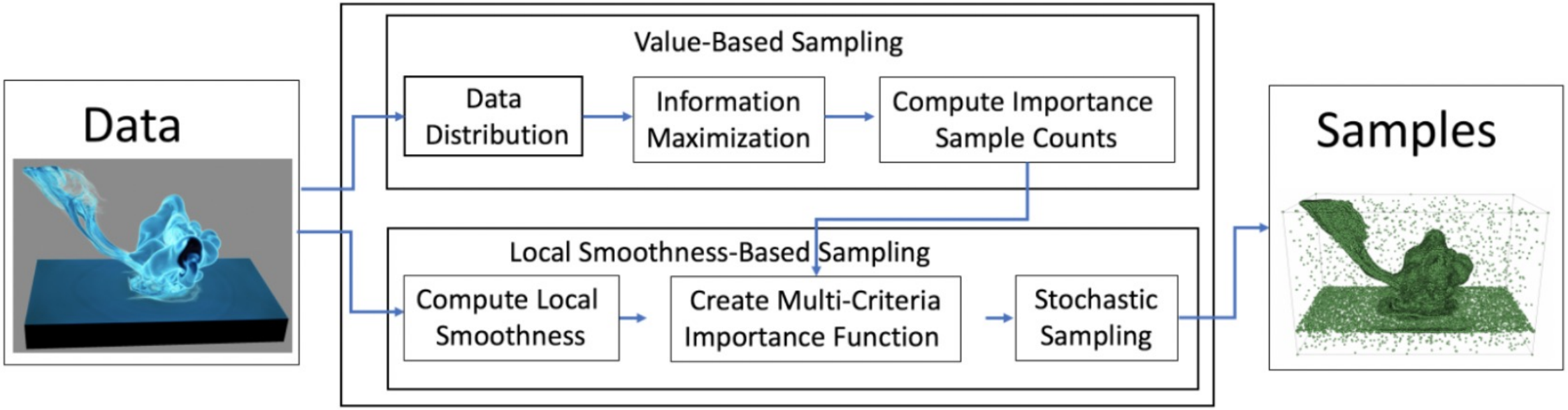
Joint Multi-Criteria Sampling

Joint Multi-Criteria Sampling

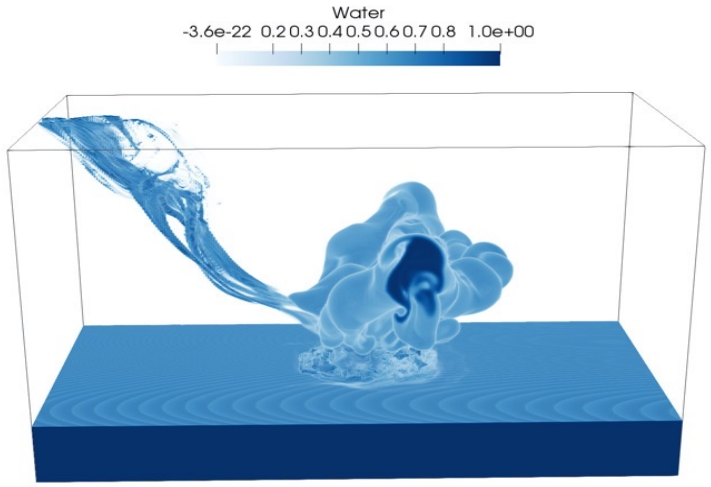
- First compute the Importance function (I_F) based on value-based sampling
- Use gradient to prioritize samples when selecting from each histogram bin of value-based histogram



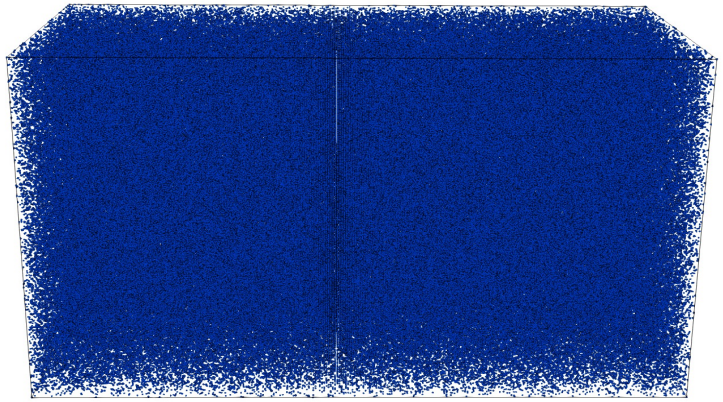
Overall Framework for Multi Criteria Sampling



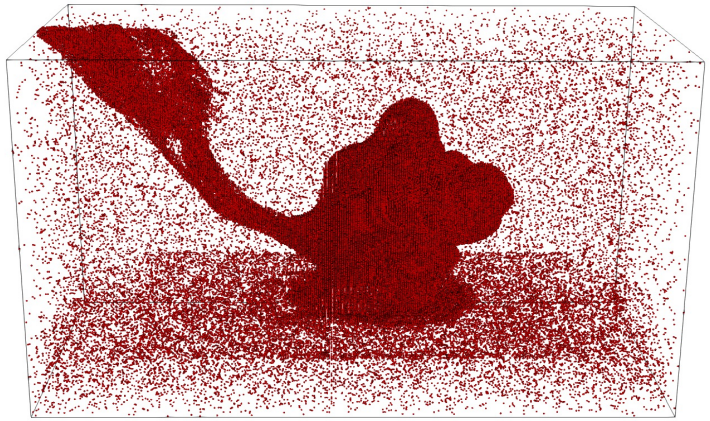
Results of Importance Sampling



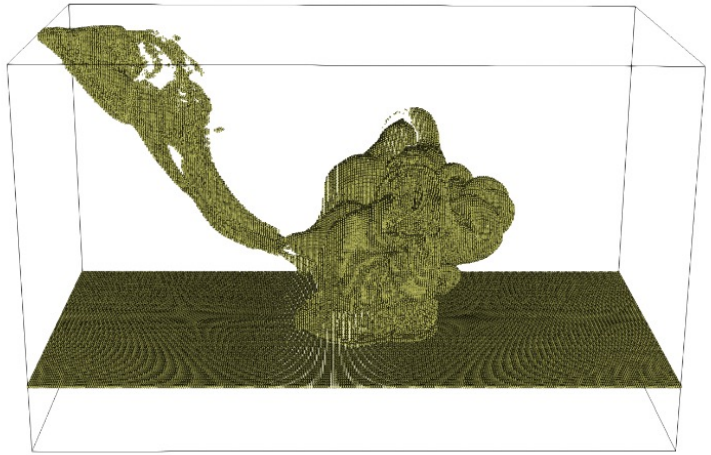
Original data showing the important region



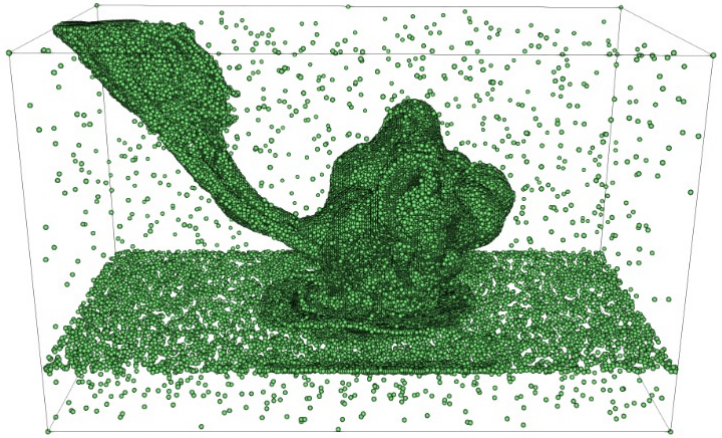
Random Sampling



Value-based Sampling

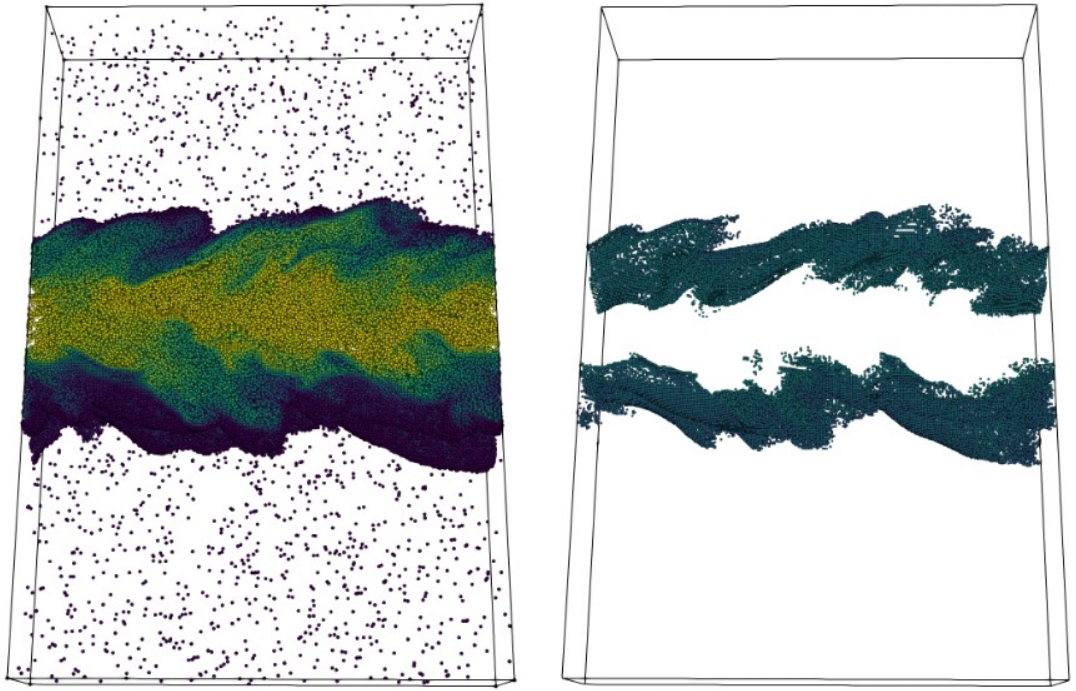


Smoothness-based Sampling

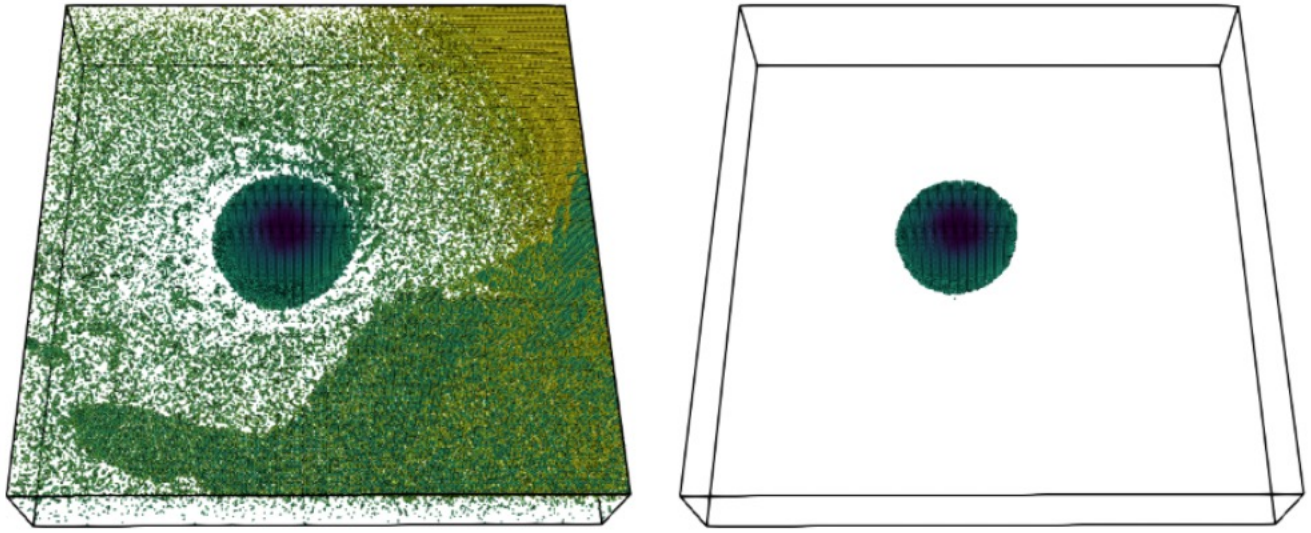


Joint Multi-Criteria Sampling

Applications: Query-driven Analysis

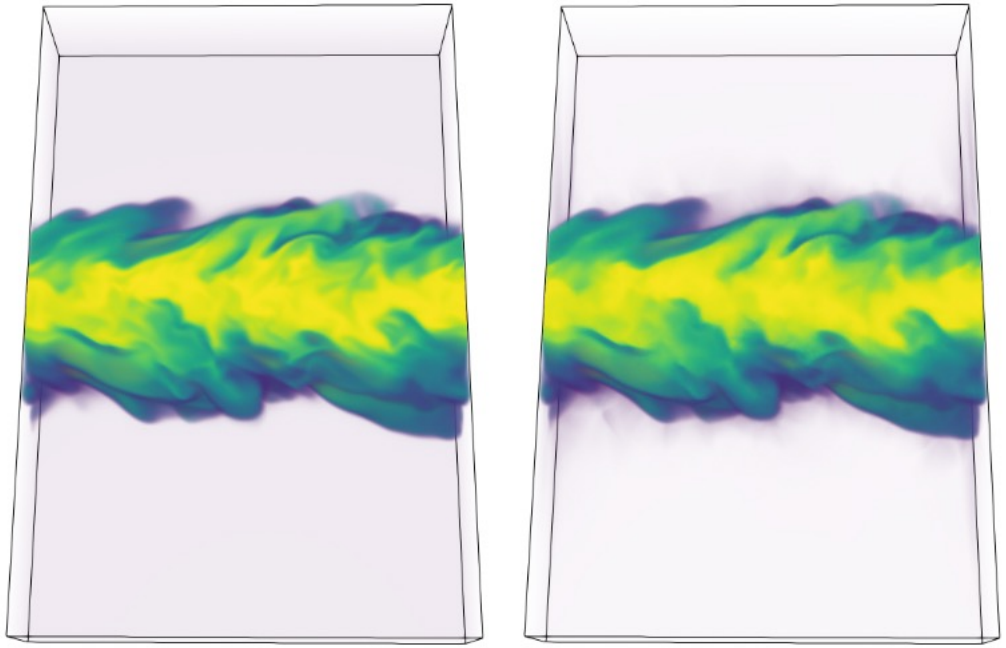


Result of a query of Mixfrac var >= 0.3 AND <= 0.5



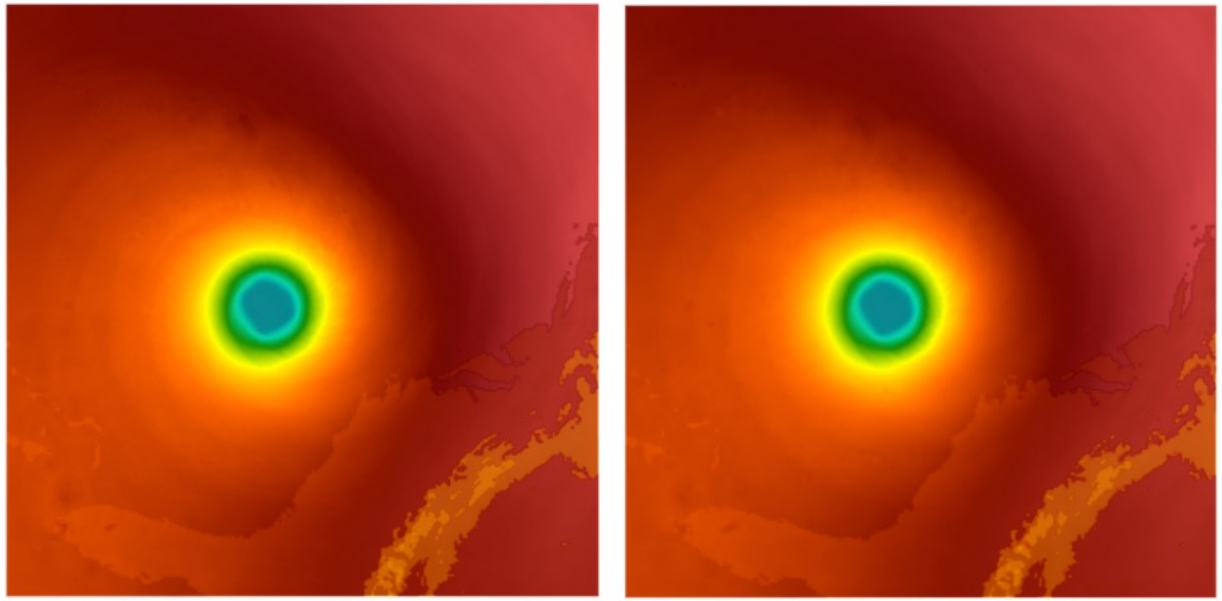
Result of a query of pressure ≥ -5000.0 AND ≤ -500.0

Applications: Reconstruction of Data



Original data

Reconstructed data

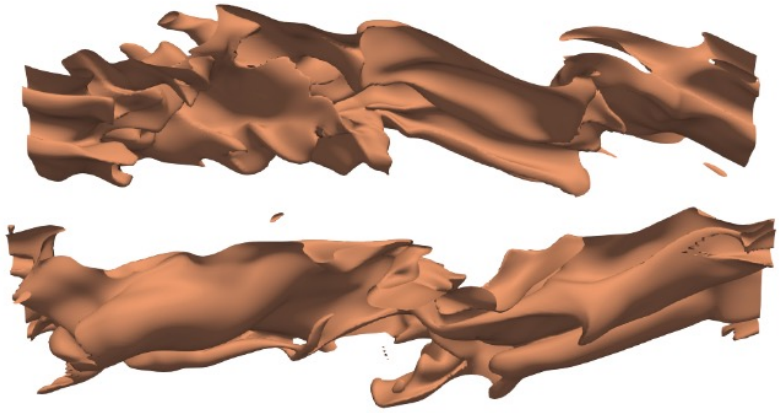


Original data

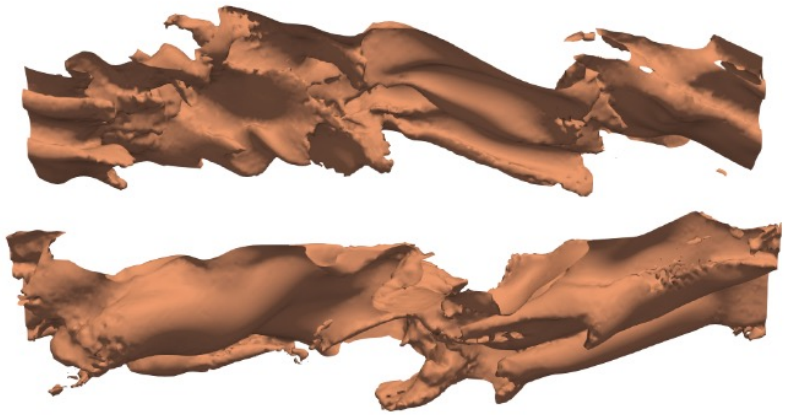
Reconstructed data

A. Biswas, S. Dutta et al., "Probabilistic Data-Driven Sampling via Multi-Criteria Importance Analysis", IEEE TVCG 2021

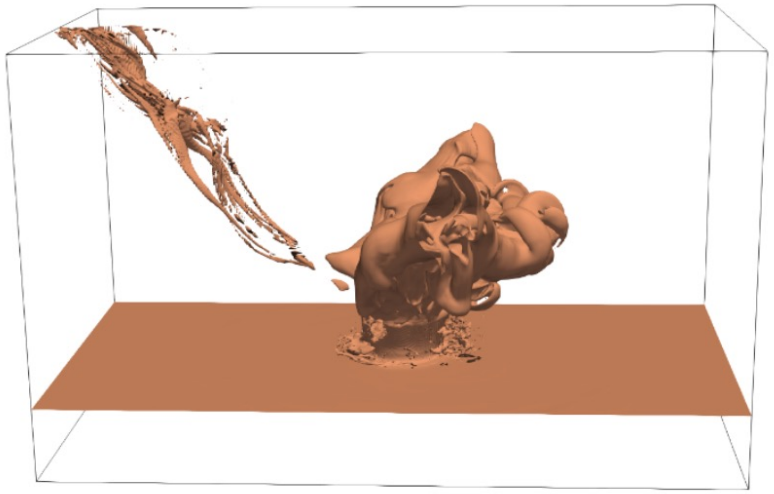
Applications: Isocontour of Data



Original contour



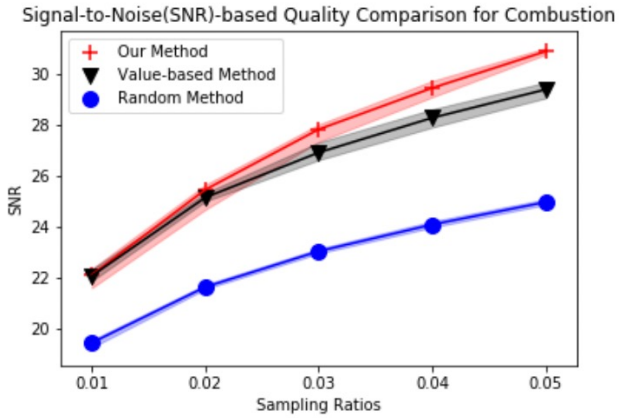
Reconstructed contour



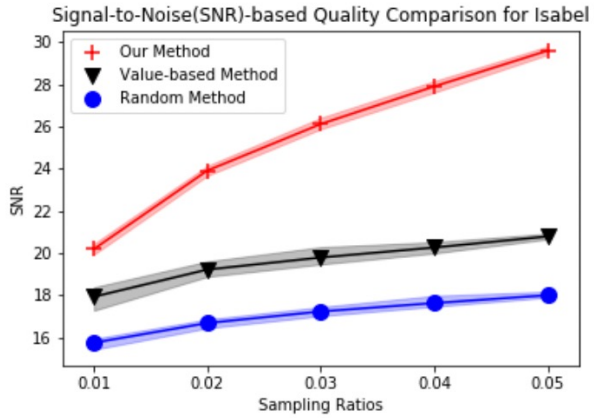
Original contour

Reconstructed contour

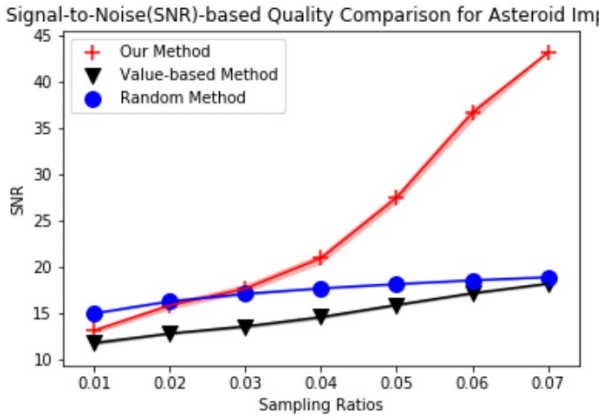
Quality Evaluation



(a) For Combustion.



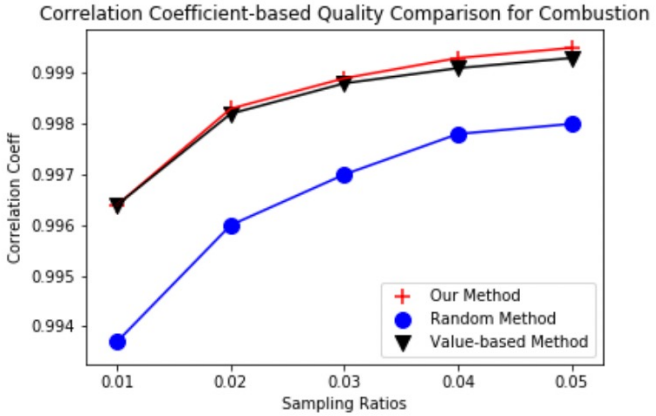
(b) For Isabel.



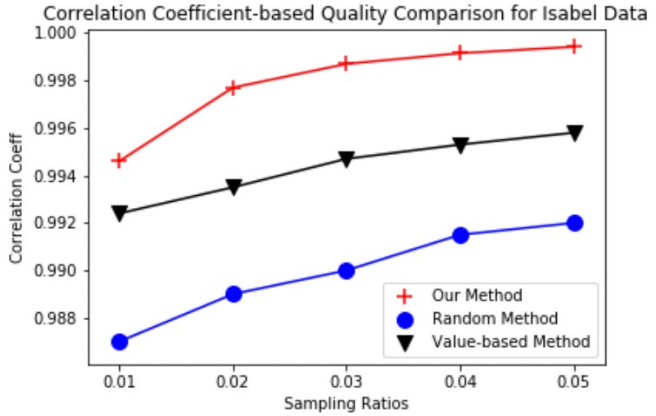
(c) For Asteroid

$$SNR = 20 * log_{10}\frac{\sigma_{raw}}{\sigma_{noise}}$$

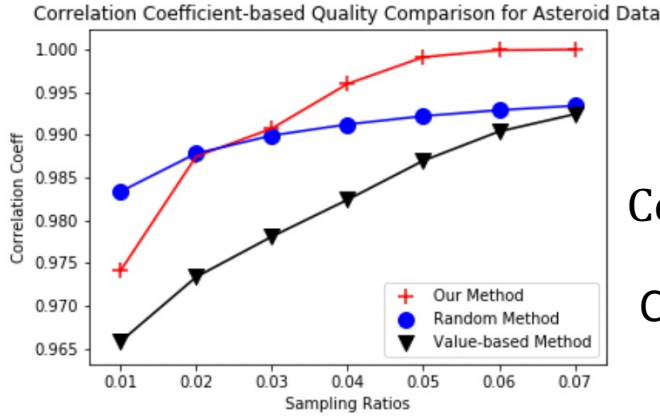
Signal to noise Ratio



(a) For Combustion.



(b) For Isabel.

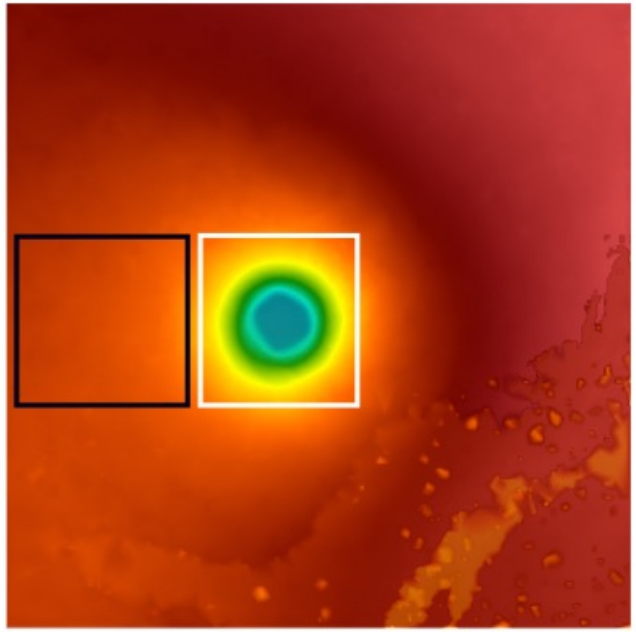


(c) For Asteroid

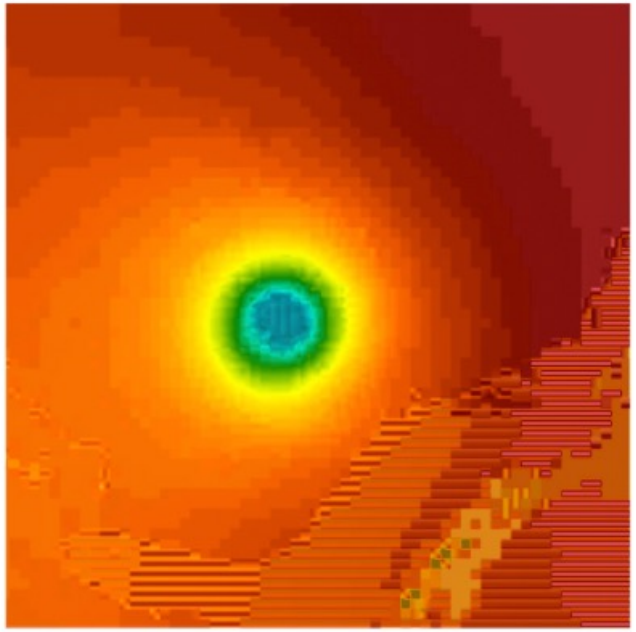
$$\mathrm{Corr(X,Y)} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Correlation coefficient

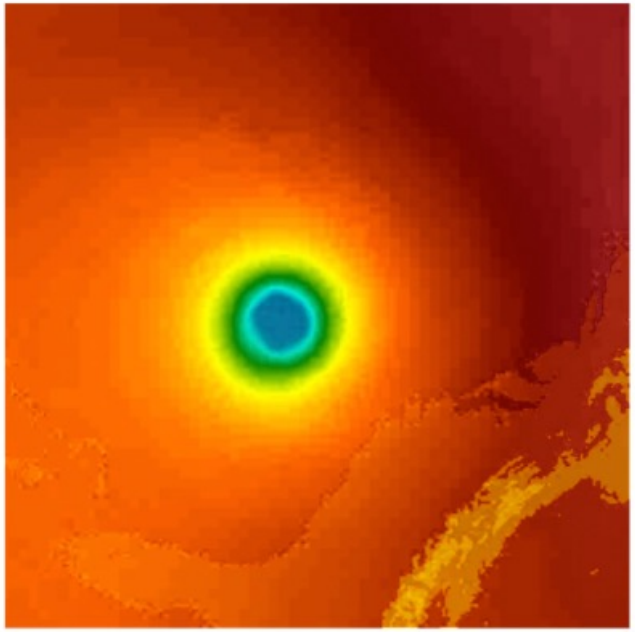
Comparison with Zfp Compression Method



Joint multi criteria sampling (1%)

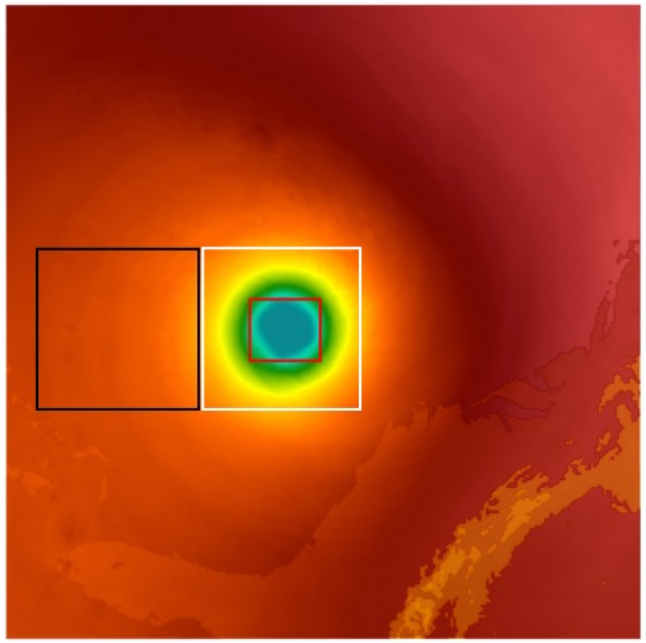


Zfp at fixed rate (1%)



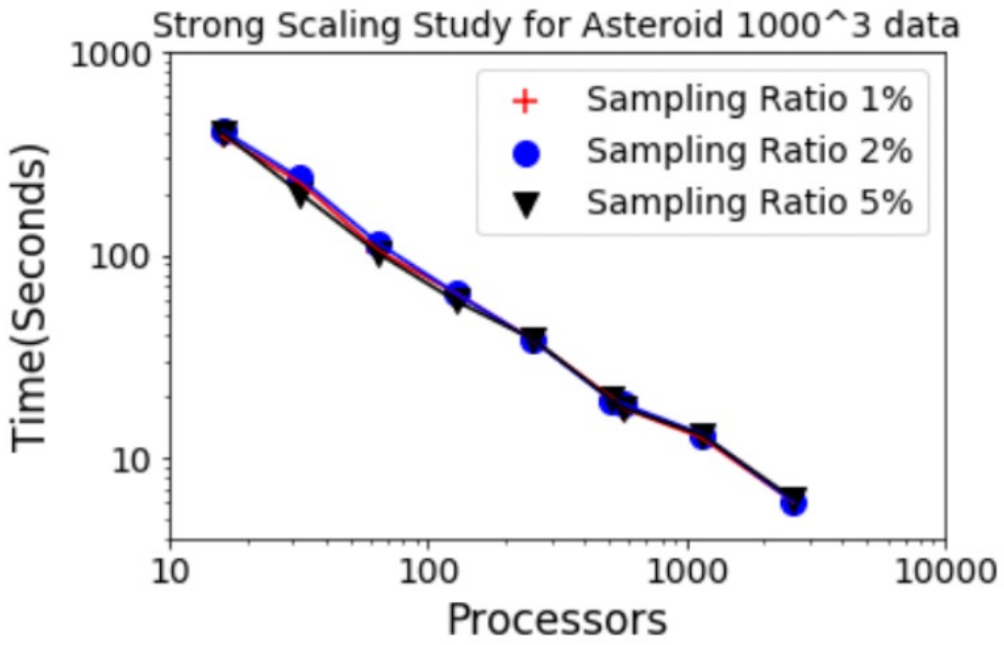
Zfp at fixed accuracy

Error Analysis

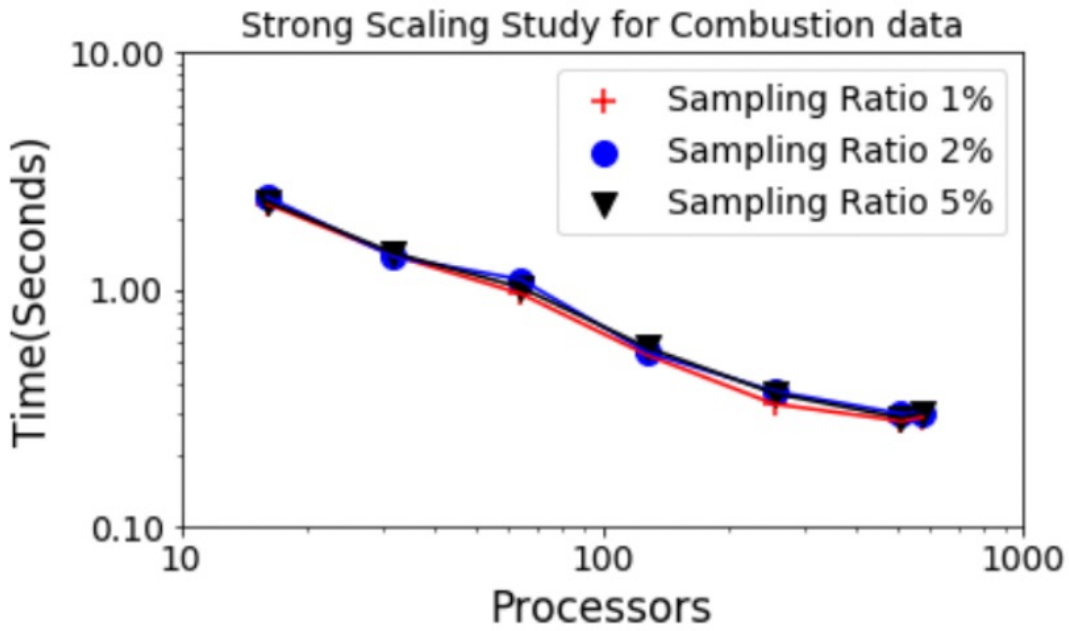


Signal-To-Noise Ratio for Isabel dataset				
Region	Sample Rate	Our Method	ZFP (Fixed Rate)	ZFP (Fixed Error)
White (Feature)	0.1%	20.50	-	-
	0.5%	24.87	0.91	17.76
	1%	28.66	21.13	26.38
	3%	34.15	35.33	47.87
	5%	35.89	37.97	53.19
Black (Non-feature)	0.1%	7.78	-	-
	0.5%	13.06	2.10	9.65
	1%	17.57	15.59	18.61
	3%	23.19	28.54	38.02
	5%	25.19	31.42	42.90
Overall	0.1%	11.42	-	-
	0.5%	16.63	1.85	18.37
	1%	20.0	11.87	27.42
	3%	26.25	23.11	49.13
	5%	29.7	25.52	54.15

Parallel Implementation Performance



(a) Asteroid data.



(b) Combustion data.