# Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

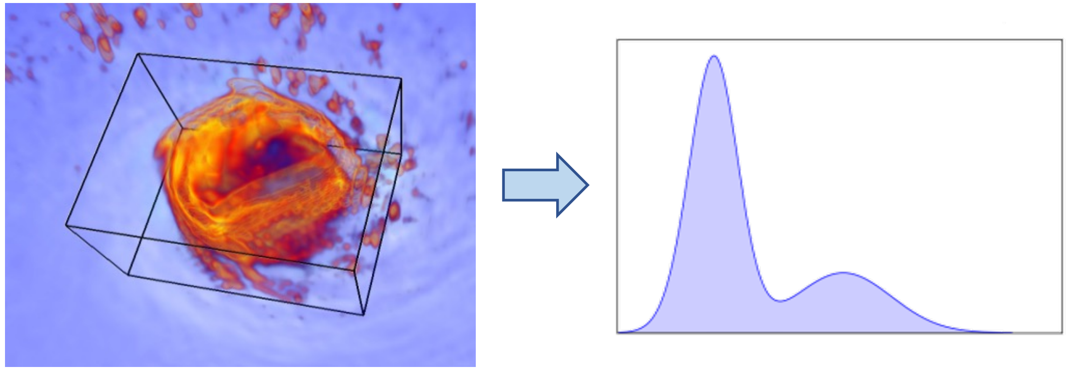
email: soumyad@cse.iitk.ac.in

# Study Materials for Lecture 16

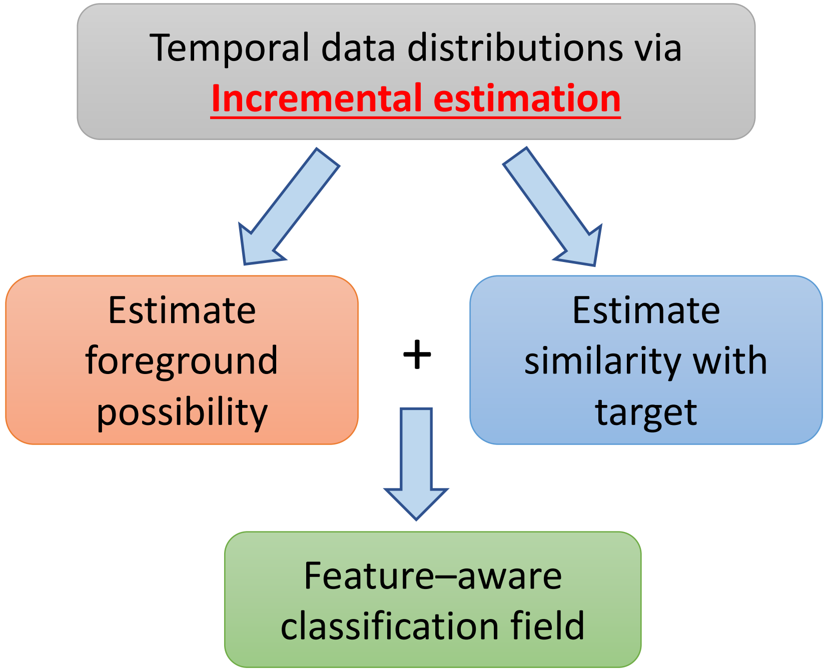
- Distribution Driven Extraction and Tracking of Features for Time-varying Data Analysis, Dutta et al., IEEE TVCG.
- CoDDA: A Flexible Copula-based Distribution Driven Analysis Framework for Large-Scale Multivariate Data, Hazarika et al., IEEE TVCG.

# Distribution-driven Feature Extraction and Tracking

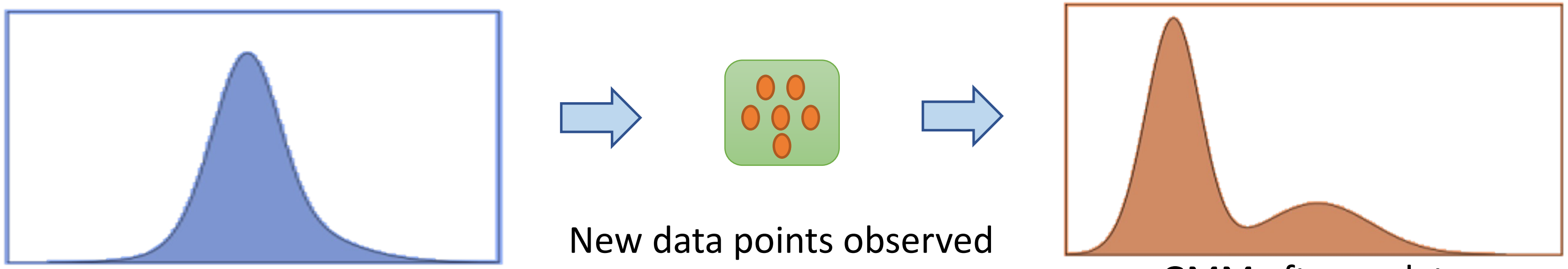
- Vaguely define features
  - Vortex core, hurricane eye, tumor in medical data
- Traditional feature tracking algorithms
  - Assume precise feature definition
- Distribution-based Methods
  - Can extract imprecisely defined features robustly
  - Track the extracted features over time



Model target feature as a distribution



# Incremental GMM Modeling for Time-varying Data



GMM before update

New data points observed

GMM after update

- Update weights as:

$$\omega_{k,t} = (1-\alpha)\omega_{k,t} + \alpha(M_{k,t}), \quad M_{k,t} = 1 \text{ for matched dist., 0 for others}$$

- Update means and covariances for the matched distribution as:

$$\mu_t = (1-\rho)\,\mu_{t\text{-}1} + \rho x_t$$

$$\sigma_t^2 = (1-\rho)\,\sigma_{t-1}^2 + \rho(x_t - \mu_t)^T(x_t - \mu_t), \quad \rho = \alpha * N(x_t | \mu_k, \sigma_k)$$

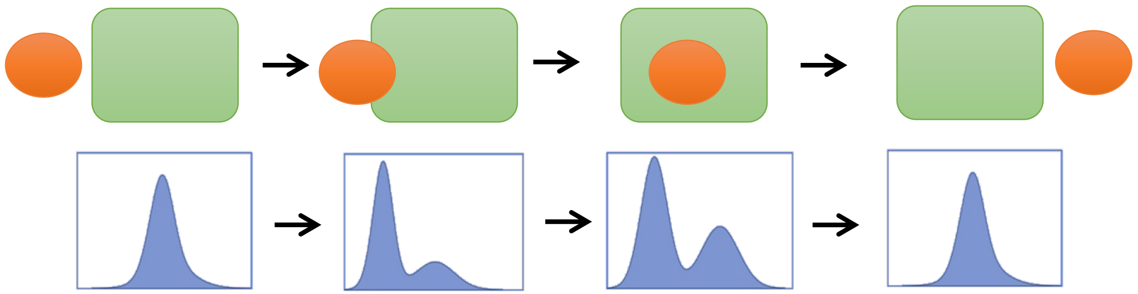
$$\Sigma_{k,t} = \sigma_k^2 \mathrm{I}, \text{ where I = Identity matrix, } \alpha = \text{learning rate}$$

# Distribution-driven Feature Extraction and Tracking

## Classification using Foreground Detection

- A block is classified as foreground if new data
  - do not match any existing Gaussians
  - match with a newly created Gaussian

$$Possibility_{foreground,t}(b_{i,t}) = q_{i,t} / n_{i,t}$$



Conceptual diagram for foreground estimation

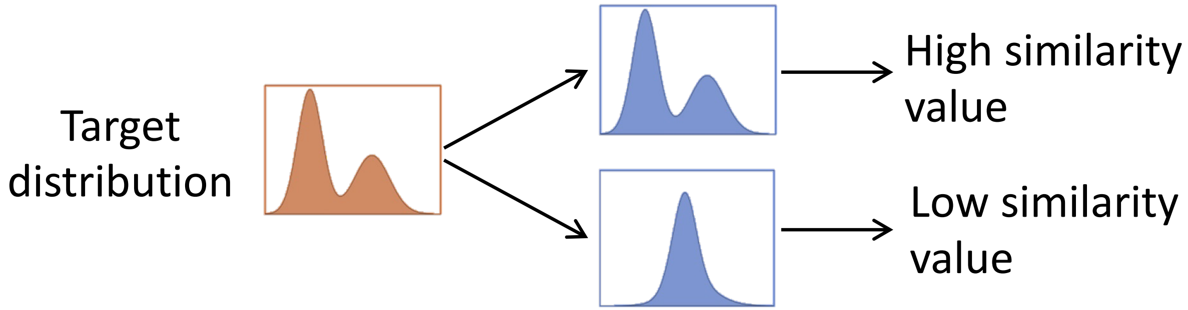
## Distribution Similarity Based Classification

- Similarity of a block with the target GMM is estimated by Bhattacharya distance

$$Possibility_{similarity,t}(b_{i,t}) = 1 - \psi_{norm}(b_{i,t}, f_t)$$

$$\psi(p, p') = \sum_{i=0}^{n} \sum_{j=0}^{m} \omega_i \omega'_j \xi(p_i, p'_j)$$

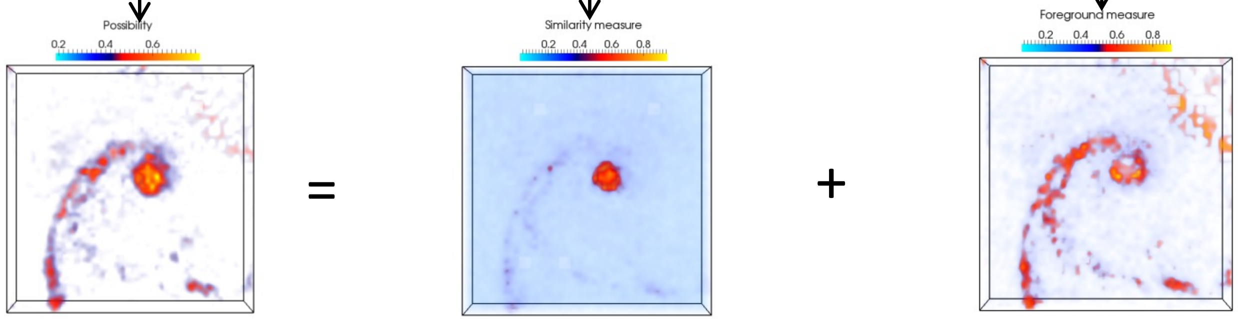
$$\xi(p,p') = \frac{1}{8}(\mu - \mu')^T \left(\frac{\Sigma+\Sigma'}{2}\right)(\mu - \mu') + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma+\Sigma'}{2}\right|}{\sqrt{|\Sigma||\Sigma'|}}\right]$$



# Distribution-driven Feature Extraction and Tracking

- Final Feature-aware Classification field:

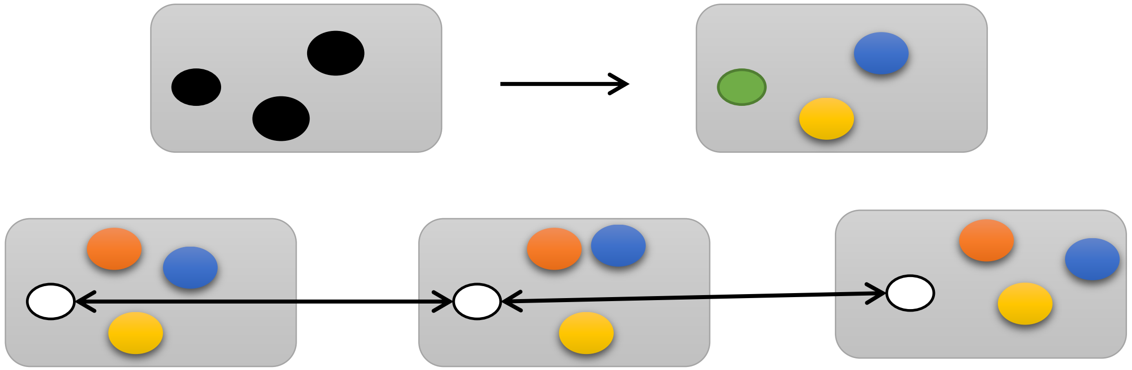
$$Possibility_{feature}(b_i) = \gamma * Possibility_{similarity}(b_i) + (1-\gamma) * Possibility_{foreground}(b_i)$$



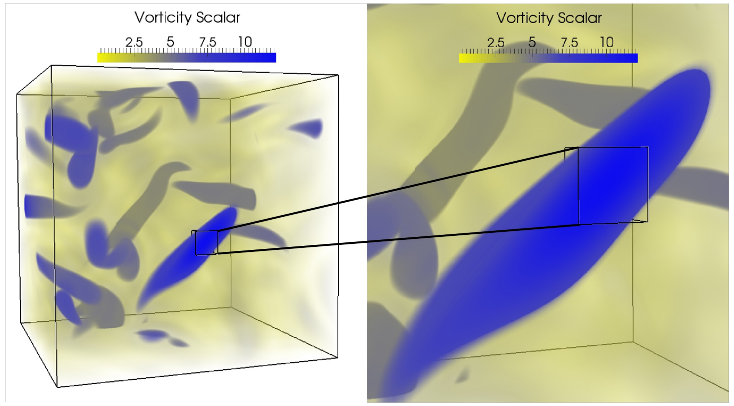

Final combined field | Similarity measure | Foreground measure

- Tracking in Classification Field: Segment the data using the threshold, apply Connected Component algorithm
- Attribute based tracking:

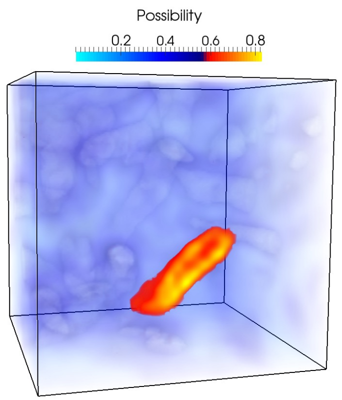
  Volume,
  mass,
  shape,
  CoG



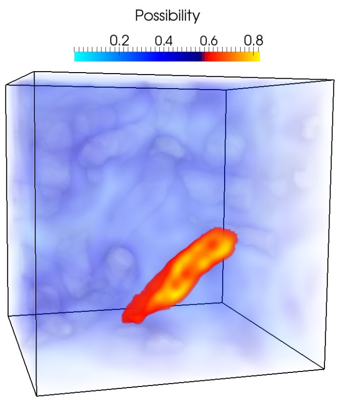
# Distribution-driven Feature Extraction and Tracking



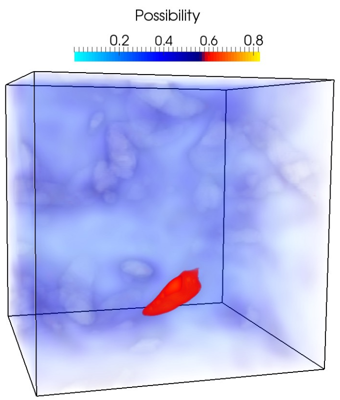

Selection of the target feature in Vortex data



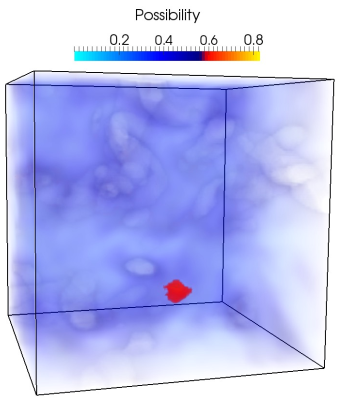

Feature at T = 3



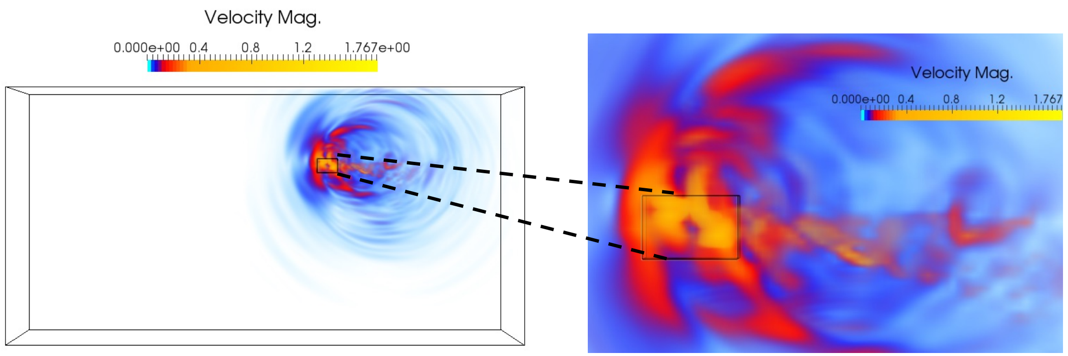

Feature at T = 6



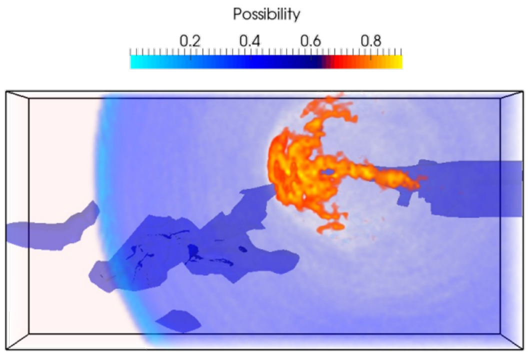

Feature at T = 15



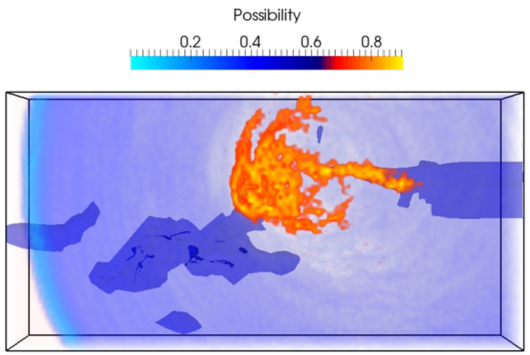

Feature at T = 20




Selection of the target feature in Earthquake data




Feature at T=10




Feature at T=20

# Multivariate Data Models using Distributions

- Multivariate Histogram
  - Storage footprint of a multivariate histogram can increase exponentially with the number of variables and the desired level of discretization
  - Sparse representations can be used by storing only non-zero bins
- Multivariate GMM
  - Estimation of multivariate GMM using Expectation-Maximization is computationally expensive
  - Model complexity increases with the number of variables
- Multivariate KDE
  - Often computationally expensive and storage inefficient
- Important to preserve the <u>dependency/correlations between variables</u> for downstream multivariate analysis

# Concept of Copula

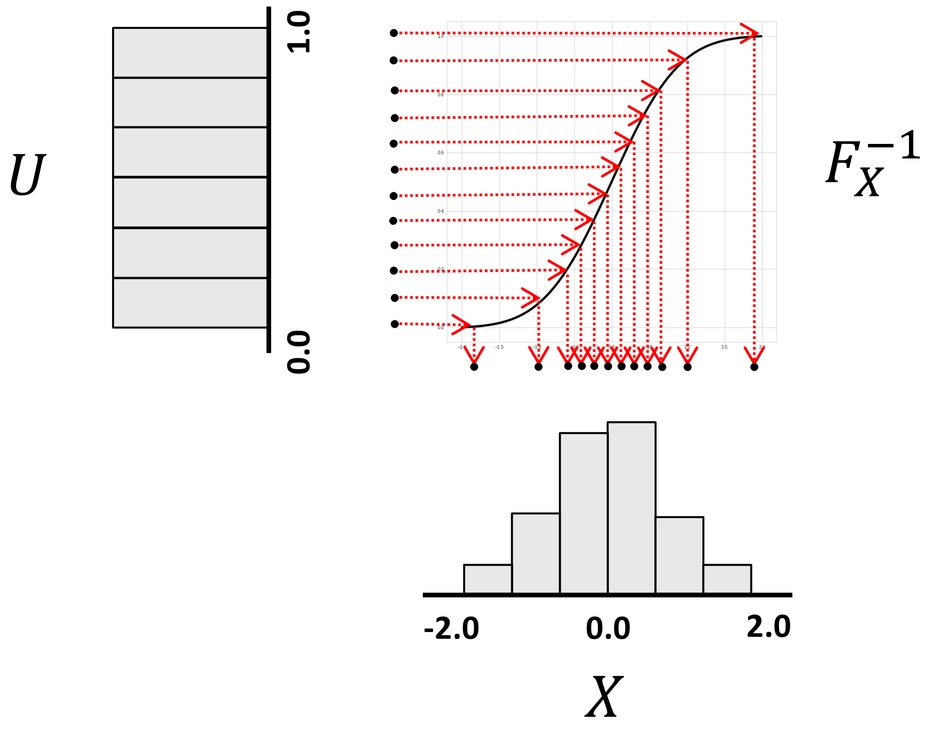
- **Definition**: A $d$-dimensional Copula is a CDF with *uniform marginals*. For $d$-uniform random variables, it can be denoted as $\boldsymbol{C}(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_d})$ .

- "Every joint CDF in $\boldsymbol{\mathcal{R}^d}$ inherently embodies a copula function": Sklar (1959)

- Sklar's Theorem provides the mathematical basis for splitting the problem of joint distribution estimation into two parts (using copula functions):
  1. Estimating individual univariate marginal distributions.
  2. Estimating the dependencies between the random variables.

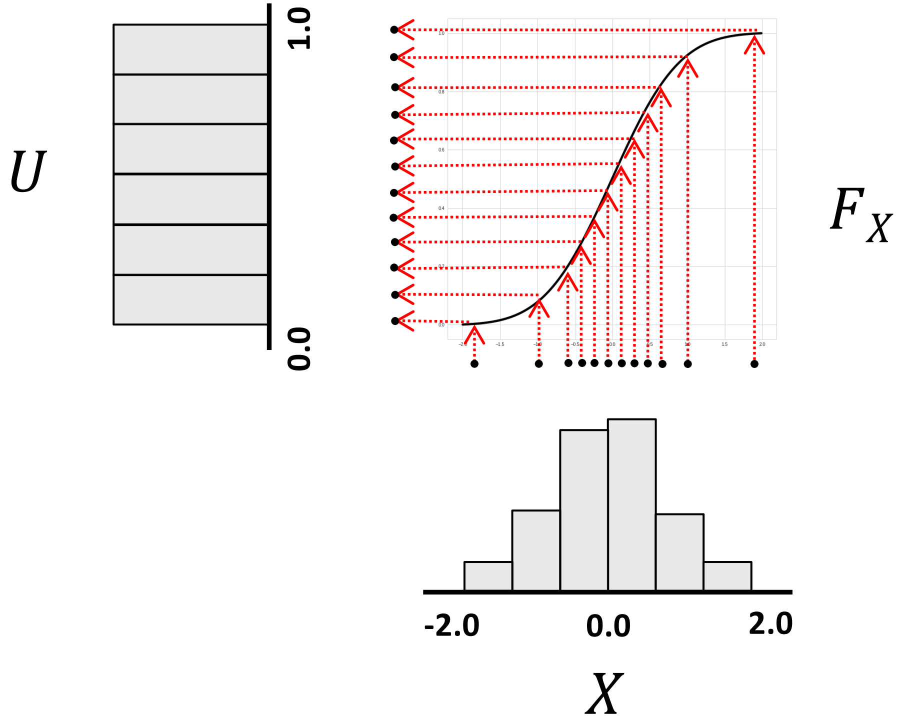
# Distribution Transformation Property

- **Property 1:** If $U$ is a uniform random variable (i.e., $U \sim Unif[0,1]$ ) and $F_X$ is a CDF of random variable $X$, then its inverse function $F_X^{-1}(U)$ corresponds to the random variable $X$ (i.e., $F_X^{-1}(U) \sim X$)



# Distribution Transformation Property

- **Property 2:** If a real-valued random variable $X$ has a continuous cumulative distribution function $F_X$, then $F_X(X) \sim Unif\,[0,1]$

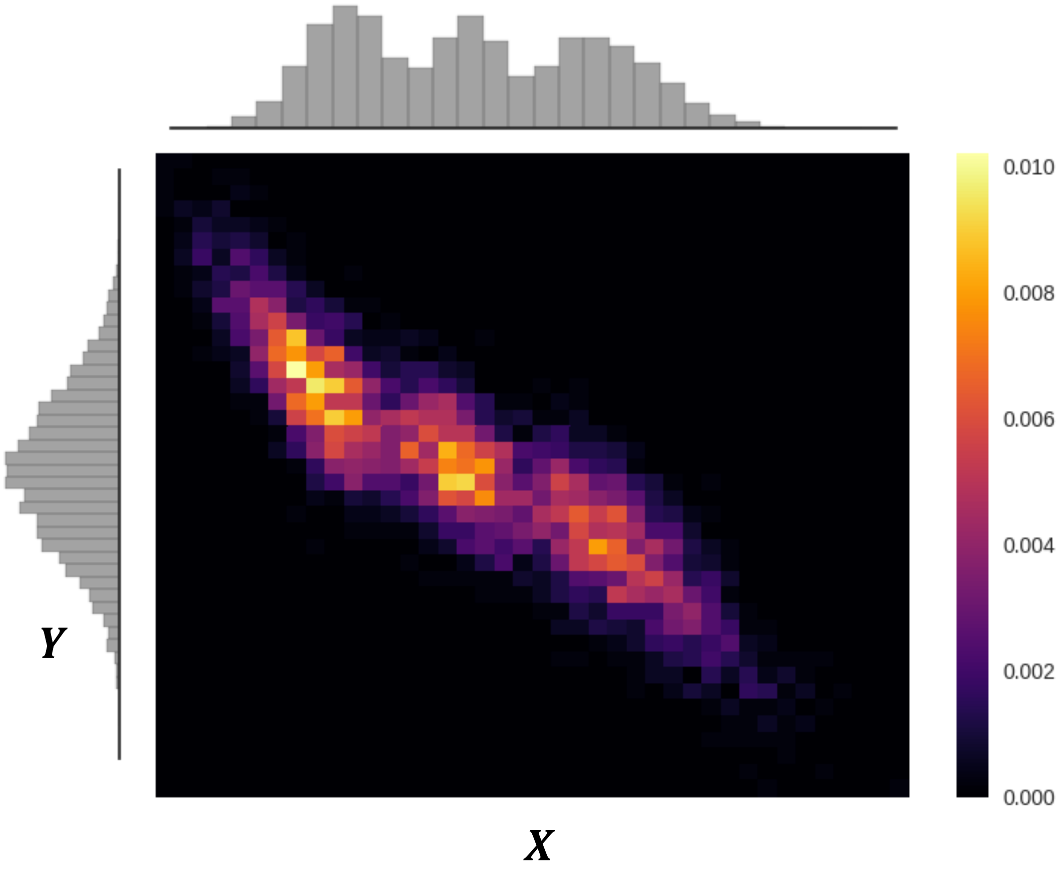


# Concept of Copula: Bivariate Example

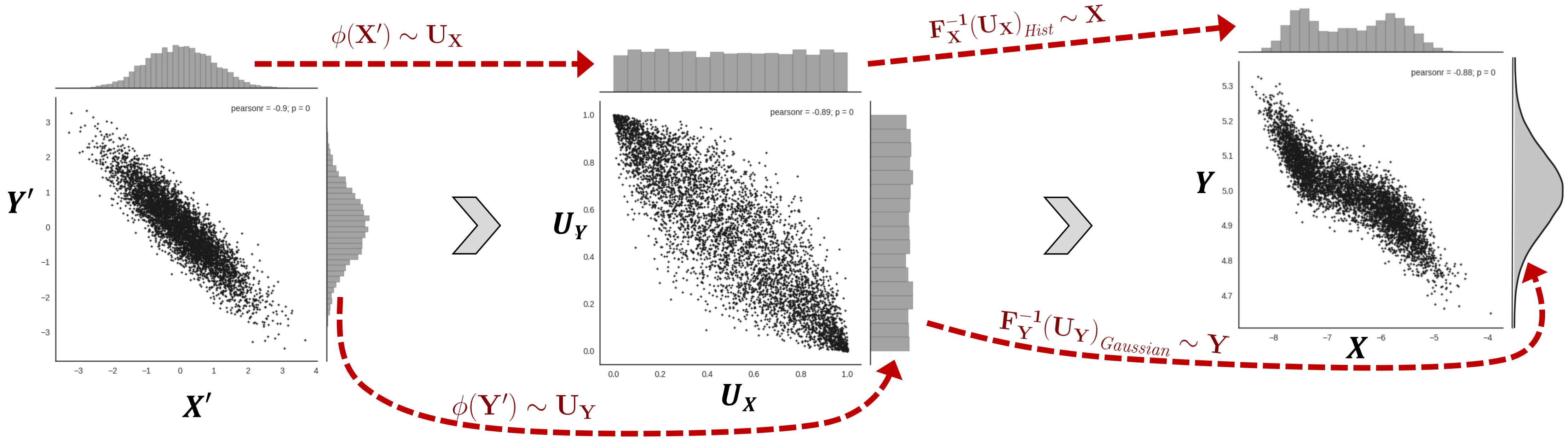
- Consider a bivariate distribution of 2 random variables $X$ and $Y$ with a strong negative correlation coefficient of $-0.9$

- Let, $F_X$ and $F_Y$ be their respective CDFs.

- So, we have: $F_X$ , $F_Y$ and $\rho = -0.9$

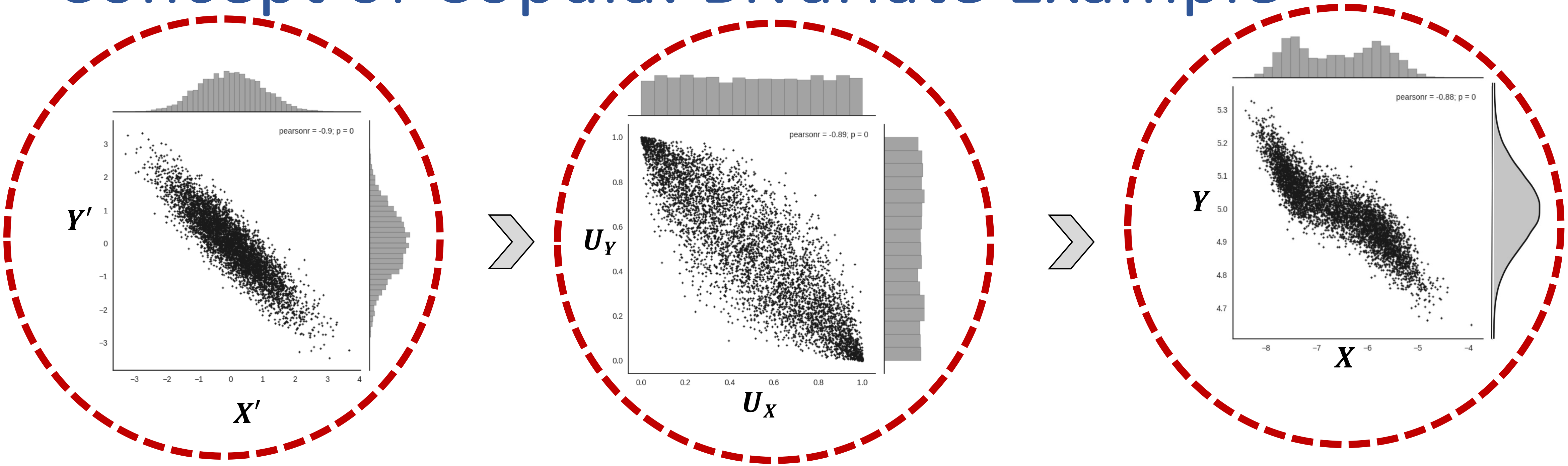


# Concept of Copula: Bivariate Example




- **Step 1:** Draw samples from a bivariate standard normal distribution with the desired correlation
- **Step 2:** Transform the Gaussian marginal distributions to Uniform distributions i.e., $U_X$ and $U_Y$ **[Property 2]**
- **Step 3:** Transform $U_X$ and $U_Y$ to the desired marginal distribution forms using the inverse function of $F_X$ and $F_Y$ **[Property 1]**

# Concept of Copula: Bivariate Example



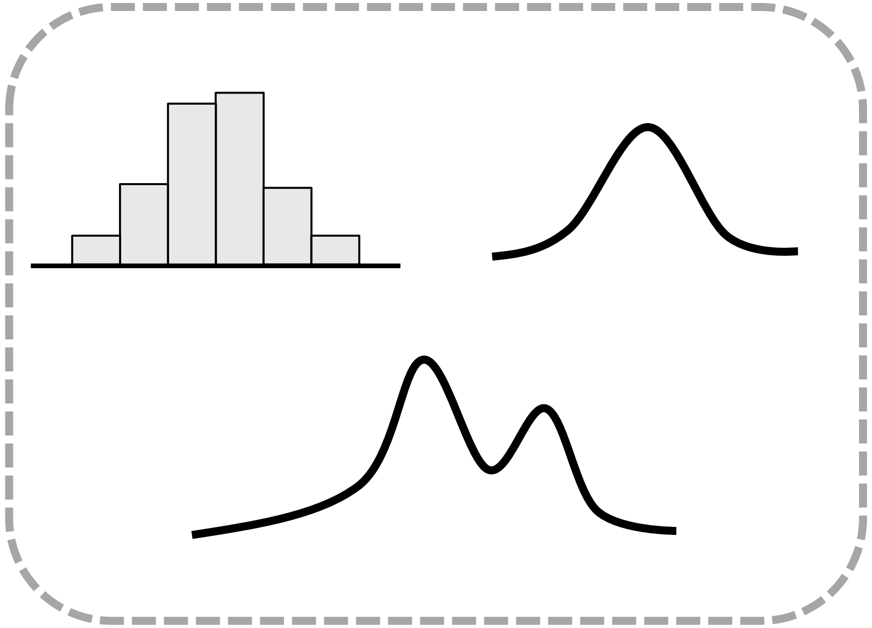

- Step 2 represents samples generated from a **Gaussian** Copula
- Gaussian Copula → standard normal distribution
- Final distribution → **Meta-Gaussian** distribution

# Multivariate Data Summary Using Copula

- What information to store in our multivariate data summaries?

# Multivariate Data Summary Using Copula

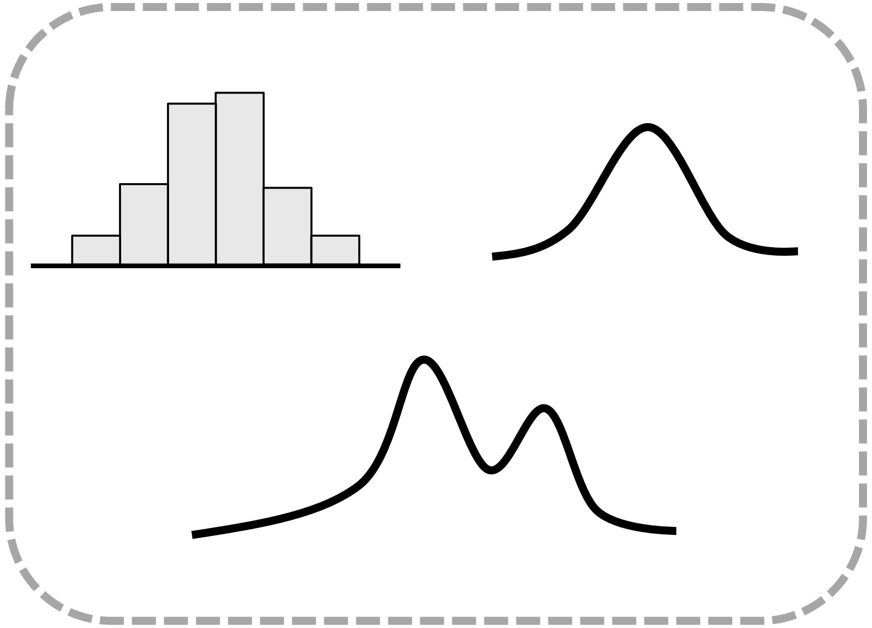
- Consider multivariate system of variables $\boldsymbol{v_1, v_2, v_3, v_4}$
- Univariate distributions, $\boldsymbol{\theta_i}$ → distribution parameters for i^th variable
- Dependency structure (correlation matrix for Gaussian Copula)



$$\begin{array}{c} \\ \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_3} \\ \boldsymbol{v_4} \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{v_1} & \boldsymbol{v_2} & \boldsymbol{v_3} & \boldsymbol{v_4} \end{array} \\ \begin{pmatrix} \mathbf{1} & \boldsymbol{\rho_{12}} & \boldsymbol{\rho_{13}} & \boldsymbol{\rho_{14}} \\ \boldsymbol{\rho_{21}} & \mathbf{1} & \boldsymbol{\rho_{23}} & \boldsymbol{\rho_{24}} \\ \boldsymbol{\rho_{31}} & \boldsymbol{\rho_{32}} & \mathbf{1} & \boldsymbol{\rho_{34}} \\ \boldsymbol{\rho_{41}} & \boldsymbol{\rho_{42}} & \boldsymbol{\rho_{43}} & \mathbf{1} \end{pmatrix} \end{array}$$

# Multivariate Data Summary Using Copula

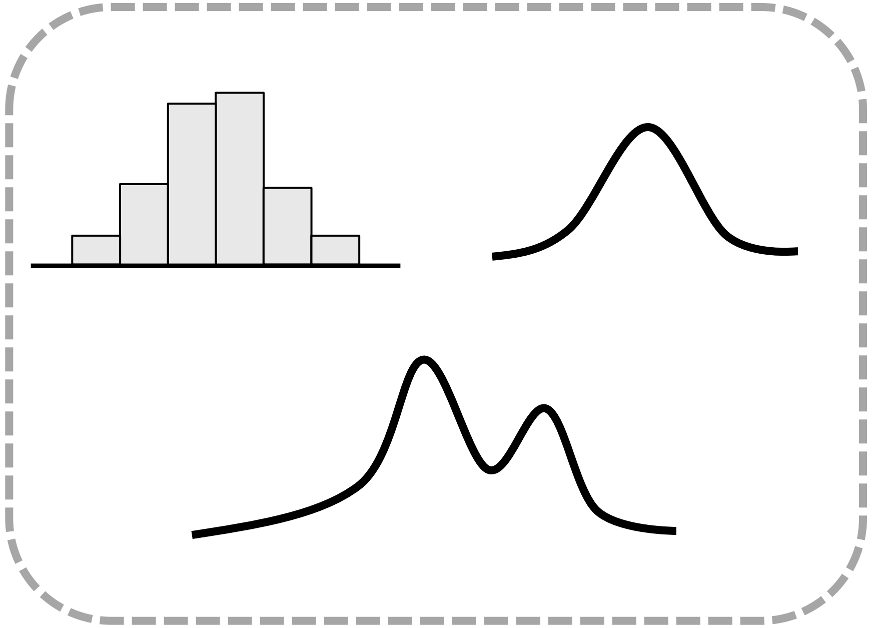
- Consider multivariate system of variables $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \boldsymbol{v_4}$
- Univariate distributions, $\boldsymbol{\theta_i}$ → distribution parameters for i<sup>th</sup> variable
- Dependency structure (correlation matrix for Gaussian Copula)



$$\begin{array}{c c} & \begin{array}{cccc} \boldsymbol{v_1} & \boldsymbol{v_2} & \boldsymbol{v_3} & \boldsymbol{v_4} \end{array} \\ \begin{array}{c} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_3} \\ \boldsymbol{v_4} \end{array} & \begin{pmatrix} 1 & \boldsymbol{\rho_{12}} & \boldsymbol{\rho_{13}} & \boldsymbol{\rho_{14}} \\ \rho_{21} & 1 & \boldsymbol{\rho_{23}} & \boldsymbol{\rho_{24}} \\ \rho_{31} & \rho_{32} & 1 & \boldsymbol{\rho_{34}} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix} \end{array}$$

# Multivariate Data Summary Using Copula

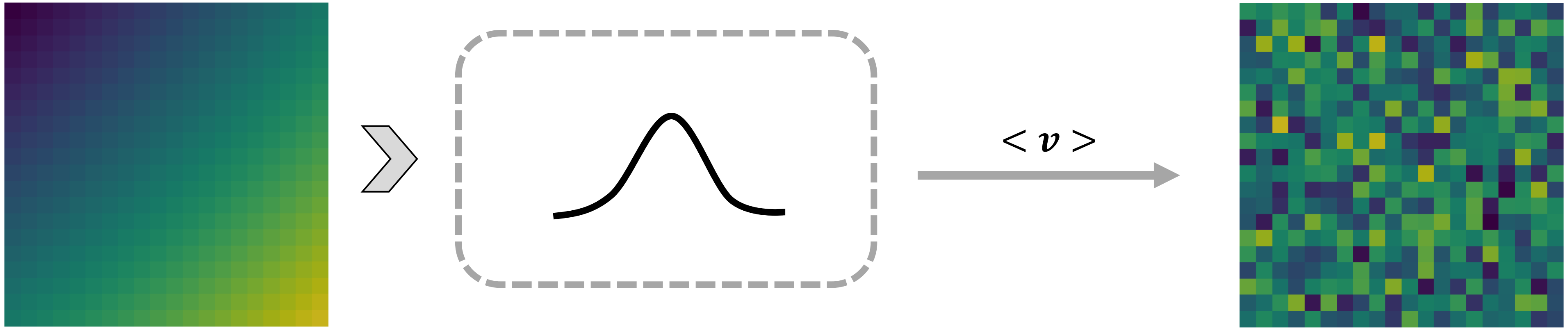
Storage footprint = $\sum_{i=1}^{\mathbf{n}} \boldsymbol{size}(\boldsymbol{\theta}_i) +$
elems from upper (or lower)triangular matrix)



$$
\begin{array}{c}
 \\ \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_3} \\ \boldsymbol{v_4}
\end{array}
\begin{array}{c}
\begin{array}{cccc} \boldsymbol{v_1} & \boldsymbol{v_2} & \boldsymbol{v_3} & \boldsymbol{v_4} \end{array} \\
\begin{pmatrix}
1 & \boldsymbol{\rho_{12}} & \boldsymbol{\rho_{13}} & \boldsymbol{\rho_{14}} \\
\rho_{21} & 1 & \boldsymbol{\rho_{23}} & \boldsymbol{\rho_{24}} \\
\rho_{32} & \rho_{32} & 1 & \boldsymbol{\rho_{34}} \\
\rho_{42} & \rho_{42} & \rho_{43} & 1
\end{pmatrix}
\end{array}
$$

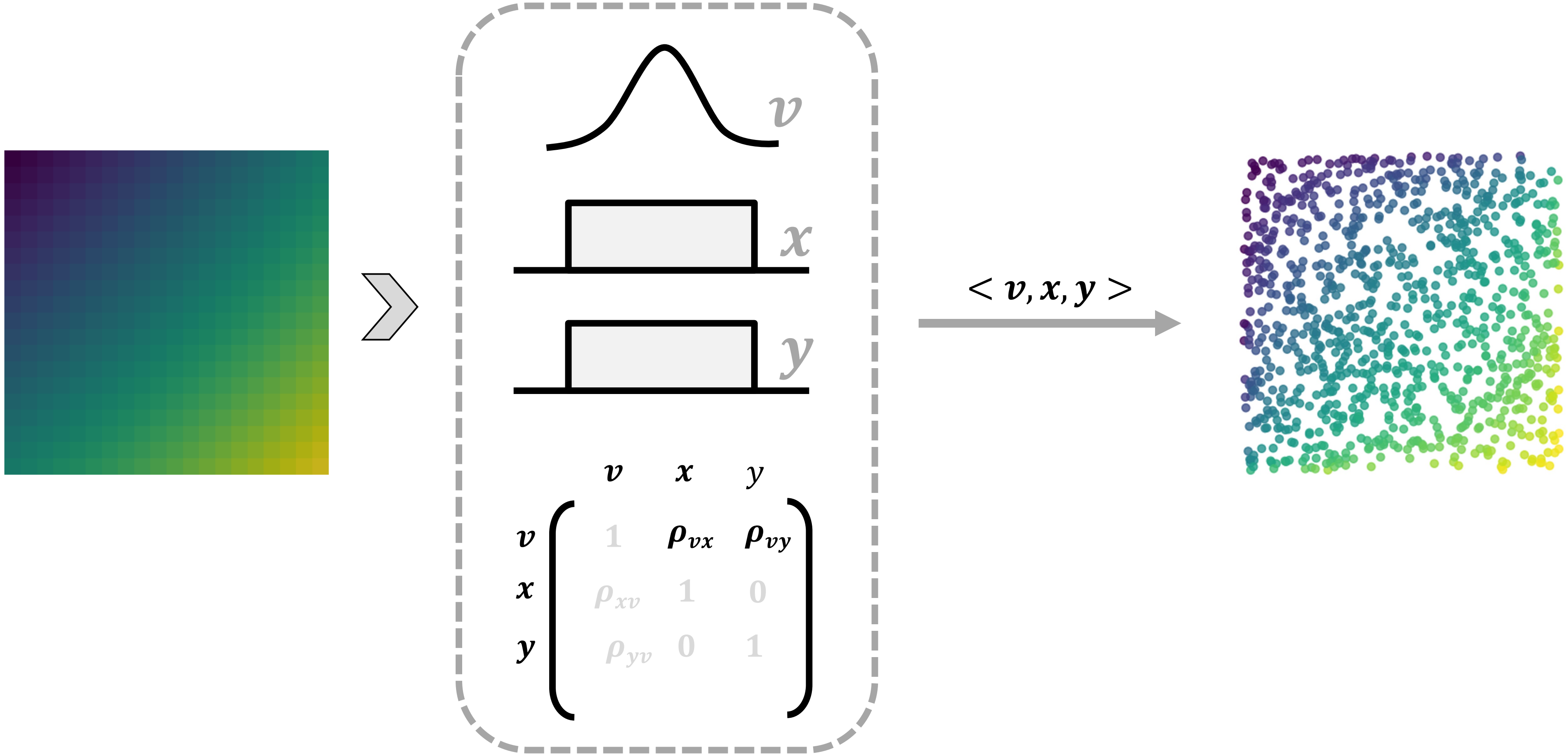
# Spatial Distributions

- Storing only the value distribution misses the spatial context of the data



# Spatial Distributions

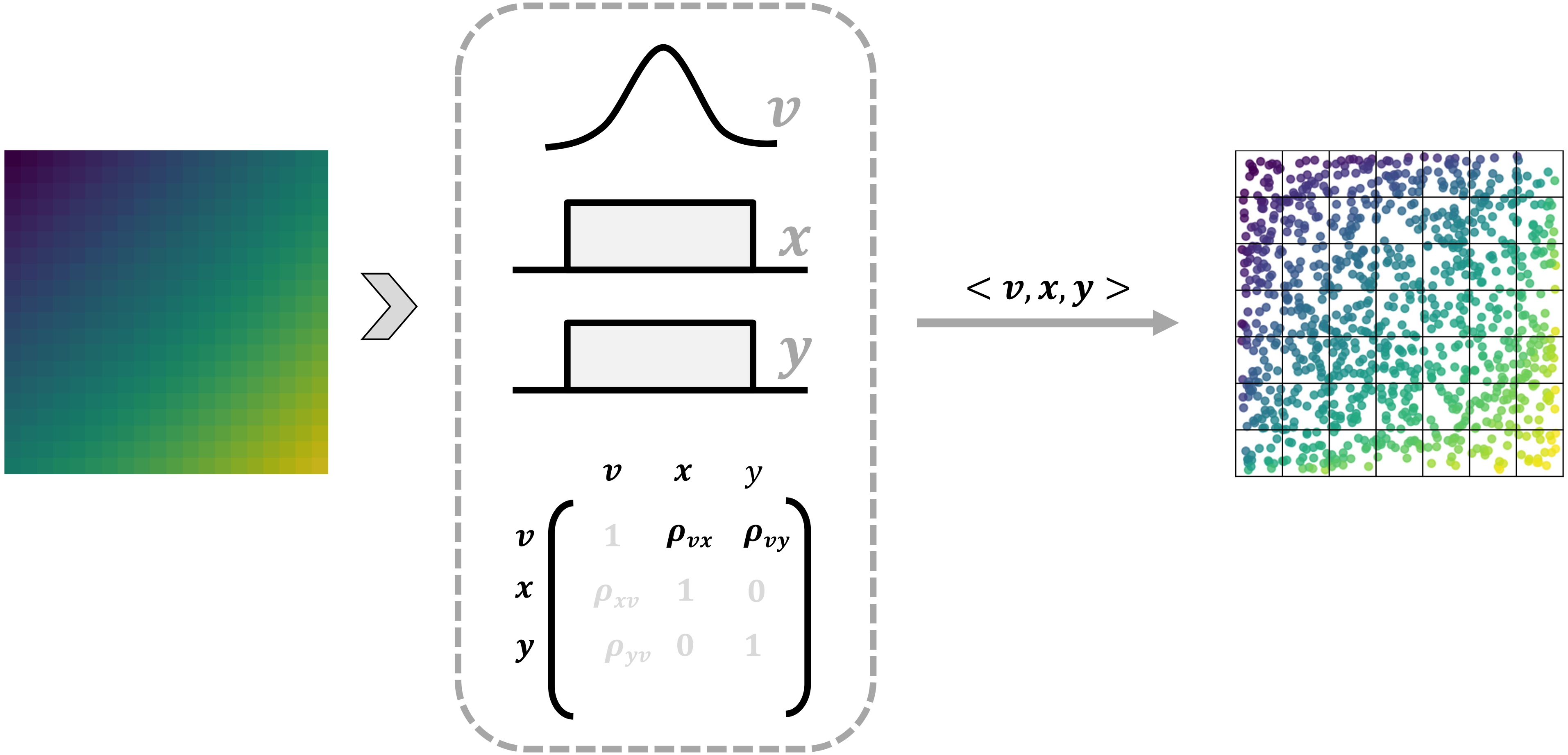
- Consider $\boldsymbol{x}$ and $\boldsymbol{y}$ distributions (uniform) along with $\boldsymbol{v}$




MV Data Summaries

# Spatial Distributions

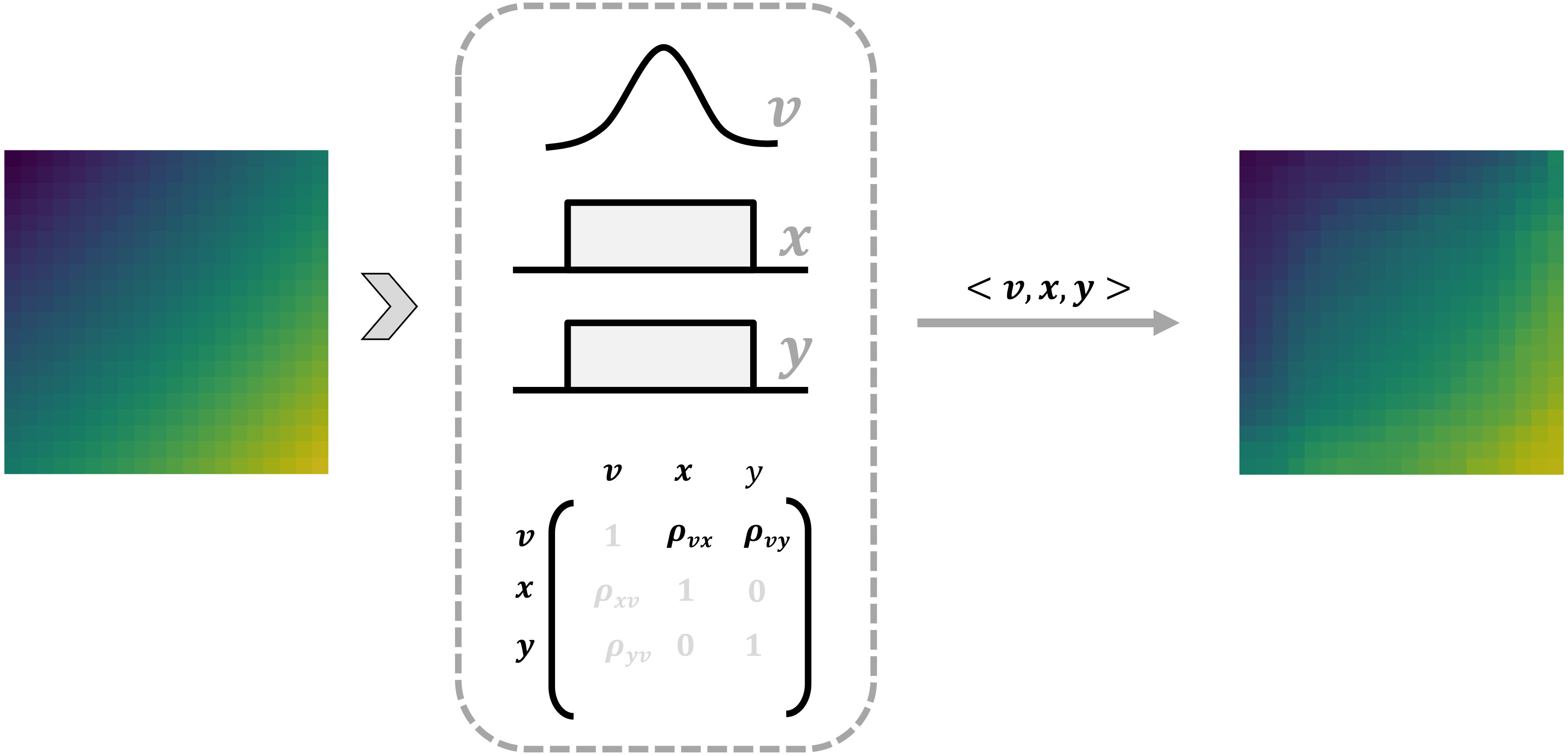
- Consider $x$ and $y$ distributions (uniform) along with $v$




MV Data Summaries

# Spatial Distributions

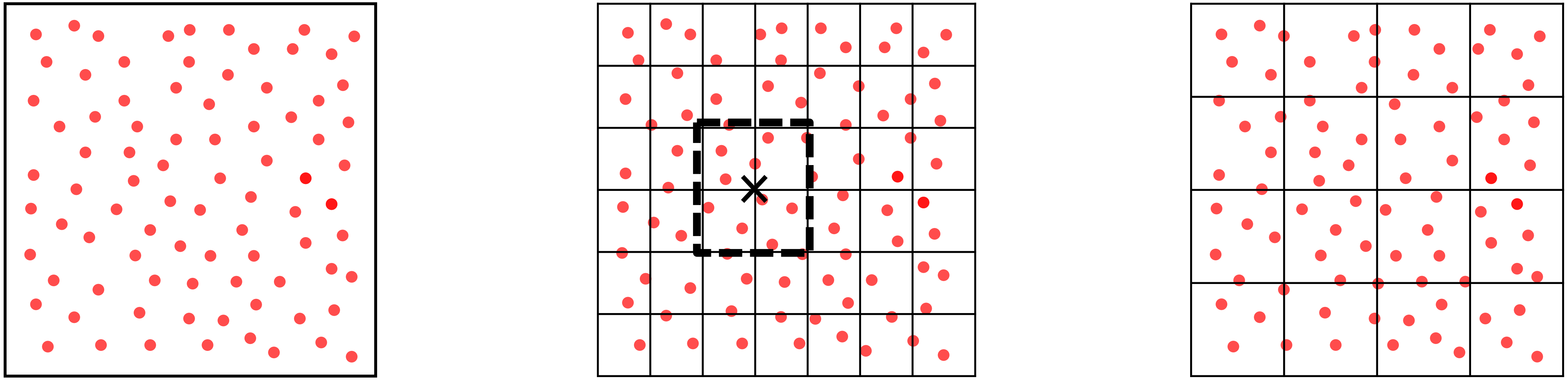
- Consider $x$ and $y$ distributions (uniform) along with $v$



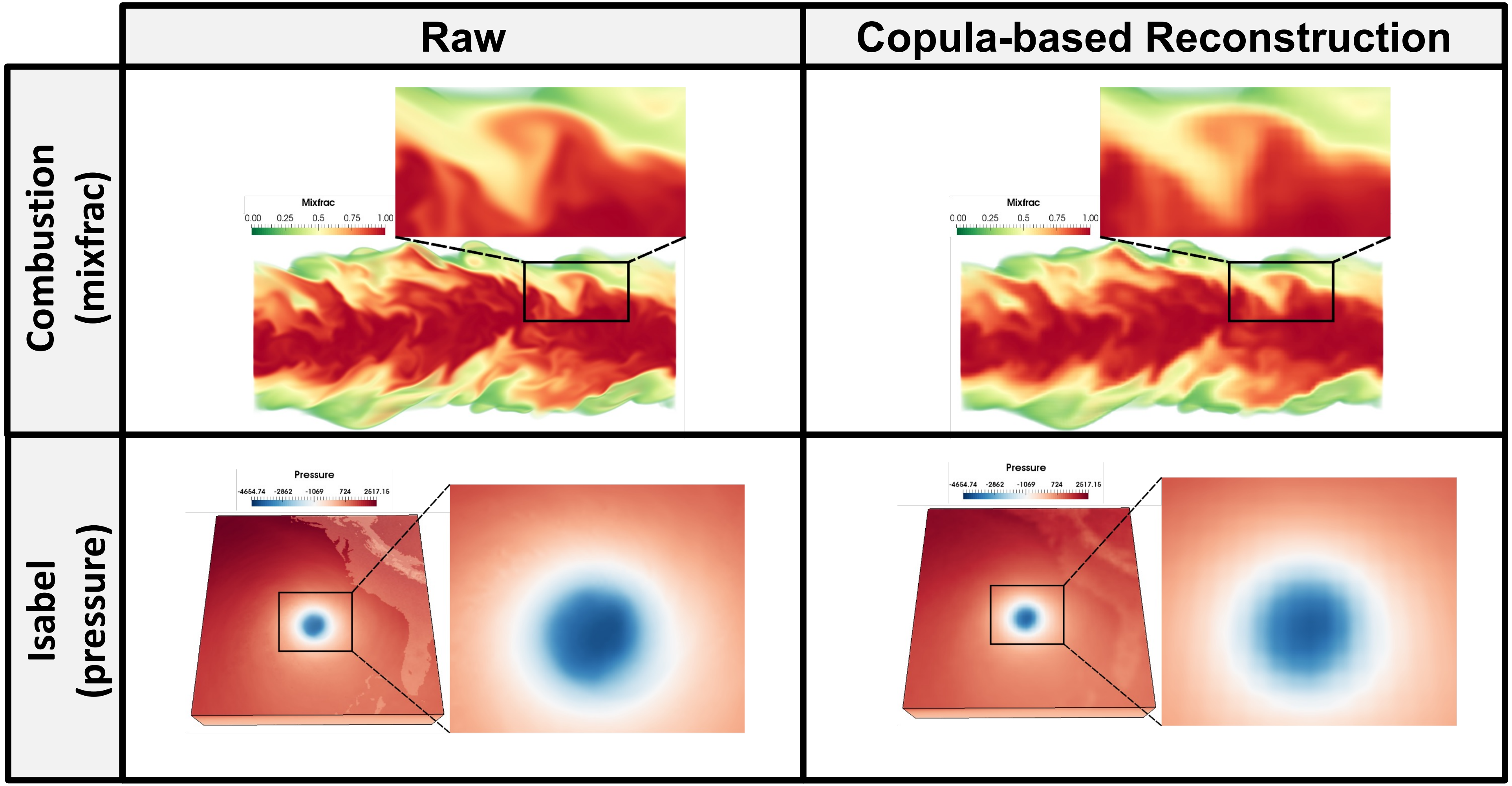

MV Data Summaries

# Multivariate Sampling For Reconstruction

- Create statistical realizations of the scalar fields from the multivariate data summaries
- Utilize the spatial information to reconstruct the scalar field by computing local particle density estimate for the grid locations
- The sample scalar field can be generated in arbitrary user-specified grid resolution



# Multivariate Sampling-based Visualization



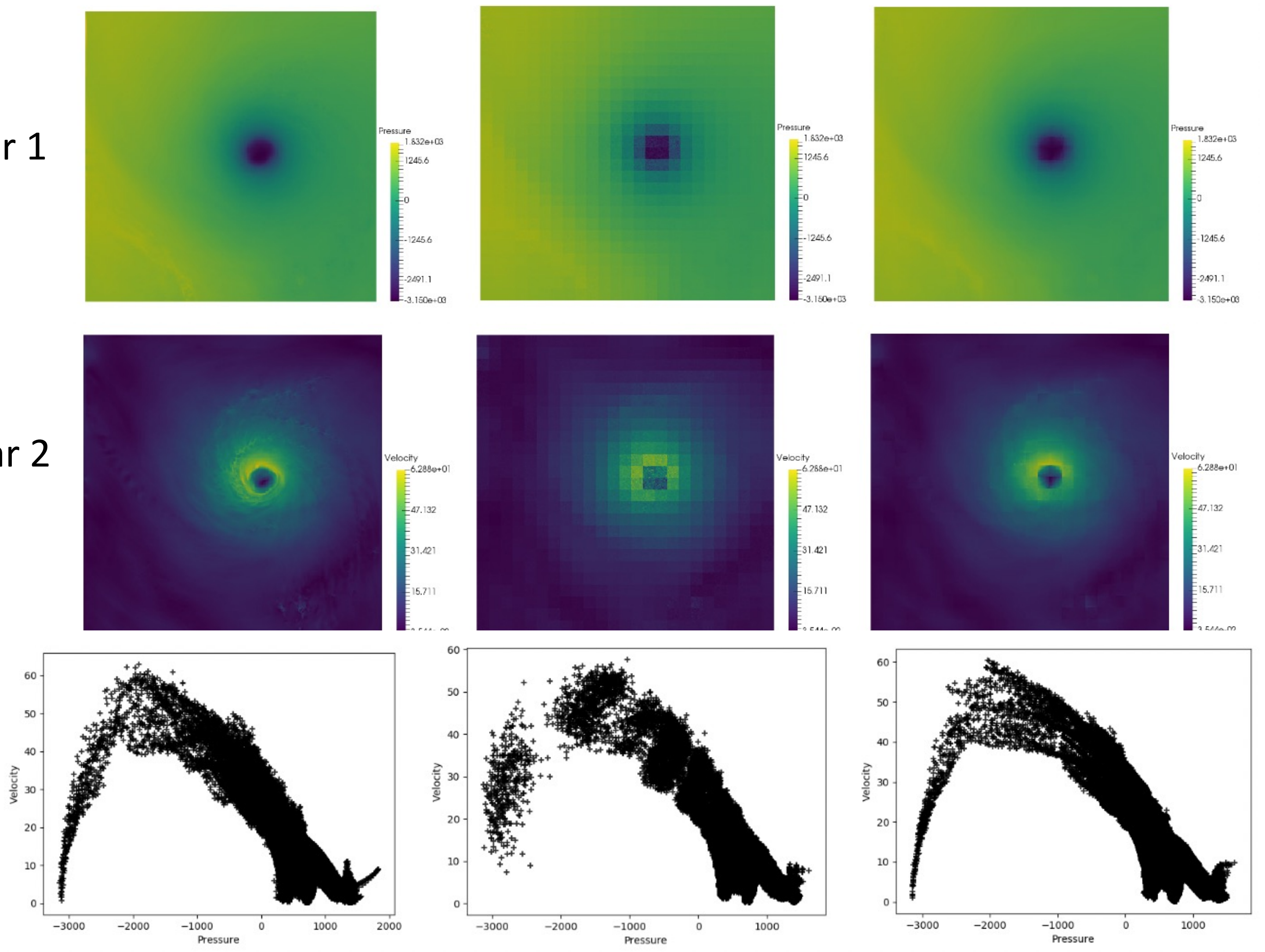
# Preserving Correlations

MV Histogram-based | Copula-based

Original var 1

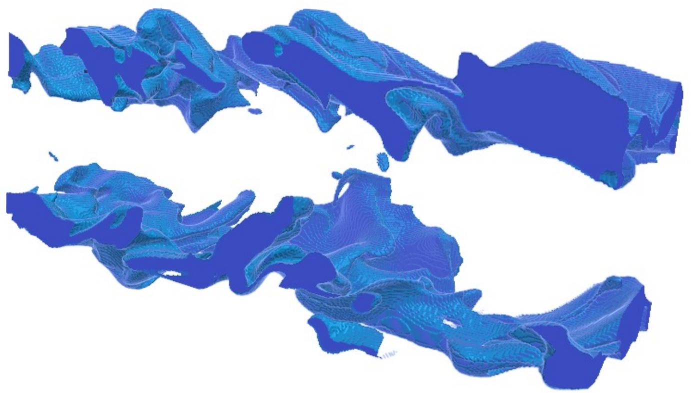
Original var 2

Original scatter plot

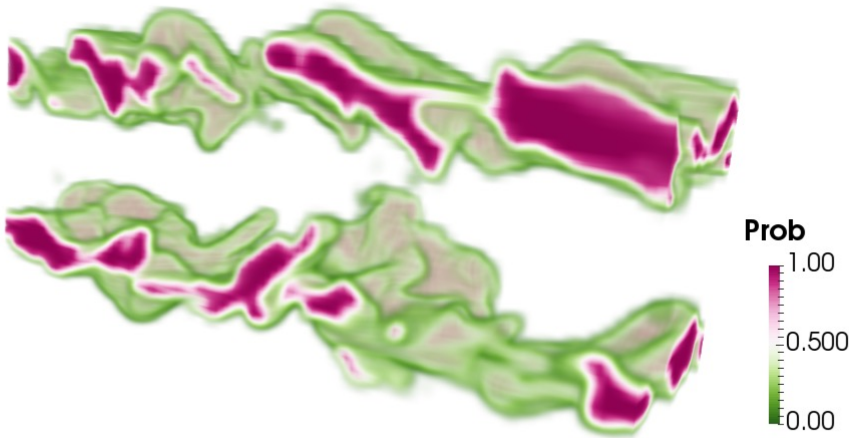


# Probabilistic Query-driven Analysis

- Query-driven methods are a class of highly effective visualization strategies
- Analysis tasks can be targeted only for the queried regions
- Selectively sample only from distributions which satisfy the user query



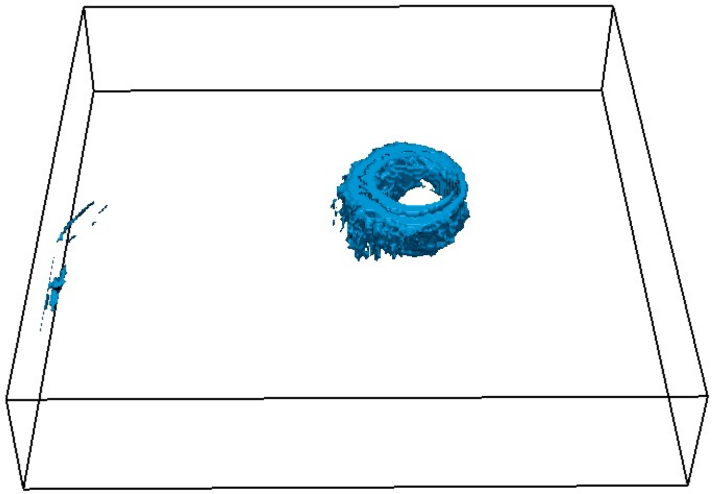
$0.3 < Mixfrac < 0.7 \textbf{ AND } y_{oh} > 0.0006$



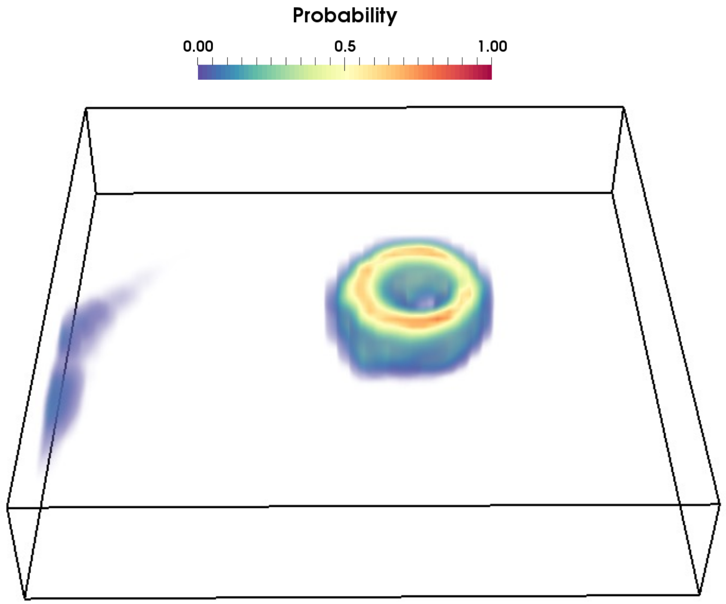

$\boldsymbol{P}(0.3 < Mixfrac < 0.7 \textbf{ AND } y_{oh} > 0.0006)$

# Probabilistic Query-driven Analysis

- Query-driven methods are a class of highly effective visualization strategies
- Analysis tasks can be targeted only for the queried regions
- Selectively sample only from distributions which satisfy the user query



$-2000 < Pressure < 500 \; \boldsymbol{AND} \; 40 < \; Velocity < 50$




$\boldsymbol{P}(-2000 < Pressure < 500 \; \boldsymbol{AND} \; 40 < \; Velocity < 50)$

# Performance and Storage

| Dataset (Resolution) | #variables | Raw Size (MB) | block size | MV Histogram | | MV GMM | | Hybrid + Copula | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Size (MB) | Est. Time (s) | Size (MB) | Est. Time (s) | Size (MB) | Est. Time (s) |
| Isabel (250x250x50) | 11 | 137.5 | 5x5x5 | 173.1 | 106.1 | 23.7 | 2623.6 | 16.2 | 203.9 |
| | | | 7x7x7 | 152.5 | 111.5 | 8.13 | 4671.6 | 5.8 | 205.4 |
| | | | 10x10x10 | 113.7 | 98.2 | 2.95 | 5006.2 | 2.2 | 230.2 |
| Combustion (480x720x120) | 3 | 497.7 | 5x5x5 | 579.4 | 311.7 | 55.7 | 4077.7 | 39.2 | 573.3 |
| | | | 7x7x7 | 509.1 | 322.4 | 39.7 | 5150.4 | 14.3 | 561.7 |
| | | | 10x10x10 | 434.2 | 305.7 | 27.8 | 9708.5 | 5.1 | 583.6 |

- Copula-based modeling + hybrid distribution-based summarization
  - 496.8 GB raw data could be reduced down to 19.8GB distribution-based data