Topics in Large Data Analysis and Visualization (CS677)

Soumya Dutta, Preeti Malakar

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: {soumyad, pmalakar}@cse.iitk.ac.in

Acknowledgements

- Some of the slides are adapted from the excellent course materials made available by: Prof. Klaus Mueller (State University of New York at Stony Brook) and Prof. Tamara Munzner (University of British Columbia).

Course Website

- We have two pages for our course:
- https://www.cse.iitk.ac.in/users/cs677/index.html
- HelloIITK: https://hello.iitk.ac.in/studio/cs677sem12425/instructor/home

Assignments

- Assignments will be done in a group of 3
- Please form your group by **August 7th** and email the names and roll numbers of group members to TA: Nitesh Trivedi (nitesht@cse.iitk.ac.in)
  - One email form each group is sufficient
  - Thanks to those who have already formed groups

Grading Policy

<table>
<tr><td>Paper presentation</td><td colspan="3">35%</td></tr>
<tr><td></td><td colspan="2">Presentation 70%</td><td rowspan="2">Viva 30%</td></tr>
<tr><td></td><td>Instructor (70%)</td><td>Peer (30%)</td></tr>
<tr><td>Mid Sem</td><td colspan="3">10%</td></tr>
<tr><td>End Sem</td><td colspan="3">10%</td></tr>
<tr><td>Assignments</td><td colspan="3">40% (15% for Assignment 1 and 25% for Assignment 2)</td></tr>
<tr><td>Attendance</td><td colspan="3">5% (minimum 80% required)</td></tr>
</table>

Data Analysis

- "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

- Dr. John W. Tukey

The Future of Data Analysis, John Tukey, 1962

Typical Analysis Done on Data

- Retrieve Value
- Filter data
- Compute Derived Value
- Determine Range
- Find Anomalies
- Cluster
- Find correlation and dependency among variables
- Find summary statistics

Handling Data

What Do We Do After Getting the Raw Data?

- Do you think real world data is clean and perfect?
  - Not really!
- Real world data is often dirty

- Data cleaning (Wrangling)
  - Missing values
  - Noisy data
  - Deal with outliers
  - Standardize/normalize
  - Resolve inconsistency
  - Fuse/merge

Data Cleaning Cycle

https://blog.insycle.com/data-cleaning-hubspot

Missing Data: Why?

- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - many more other reasons …..

Missing Data: How to Handle?

- How would you estimate the missing value for a dataset?
  - Ignore or put in a default value (will decimate the usable data)
  - Manually fill in (can be tedious or infeasible for large data)
  - Use the available value of the nearest neighbor
  - Average over all the values
  - Use a probabilistic methods (Regression, Bayesian Methods, etc.)
  - Use a Neural Networks to predict missing data (Generative Models)

Data Normalization and Standardization

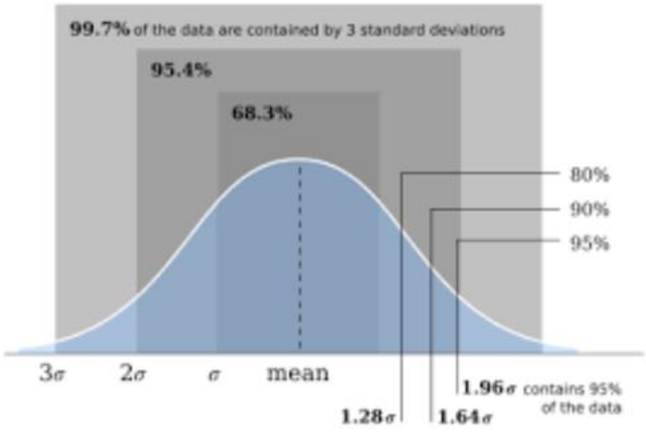
- Sometimes we like to have all variables on the same scale
  - Min-max normalization

$$v' = \frac{v - min}{max - min}$$

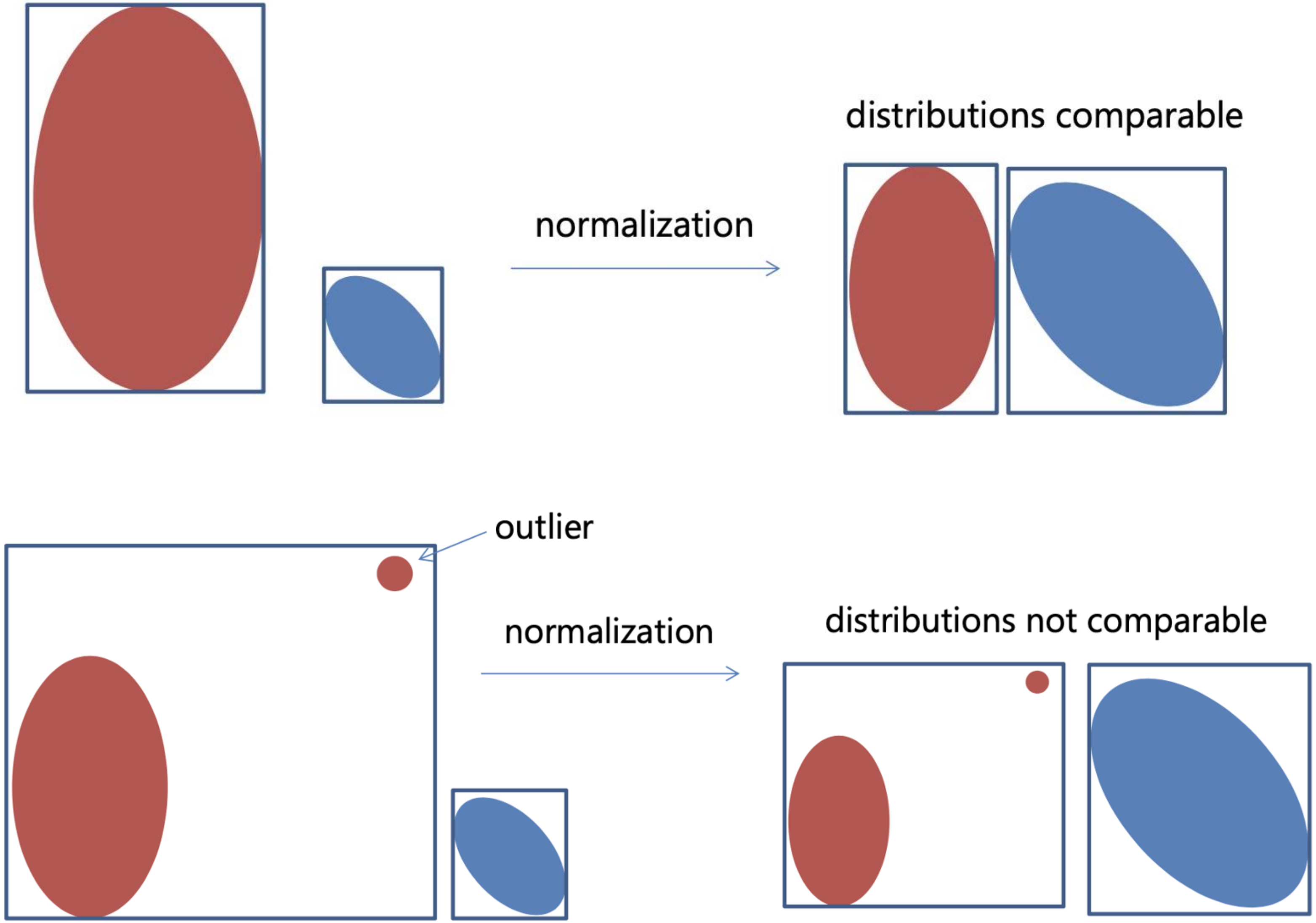
  - Standardization (Z-score)

$$v' = \frac{v - \overline{v}}{\sigma_v}$$

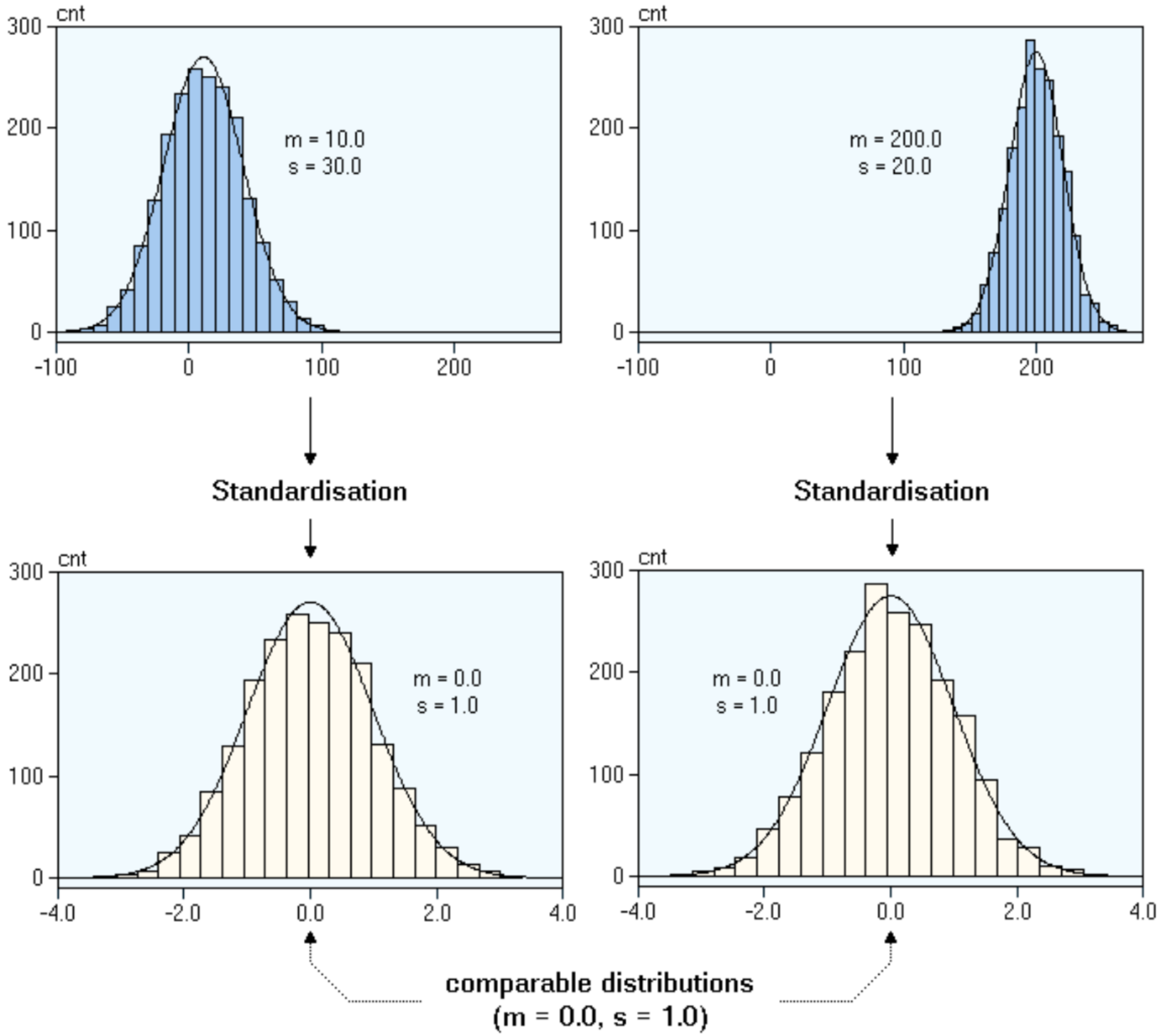
  - Clipping tails and outliers
    - set all values beyond $\pm 3\sigma$ to value at 3σ



Normalization

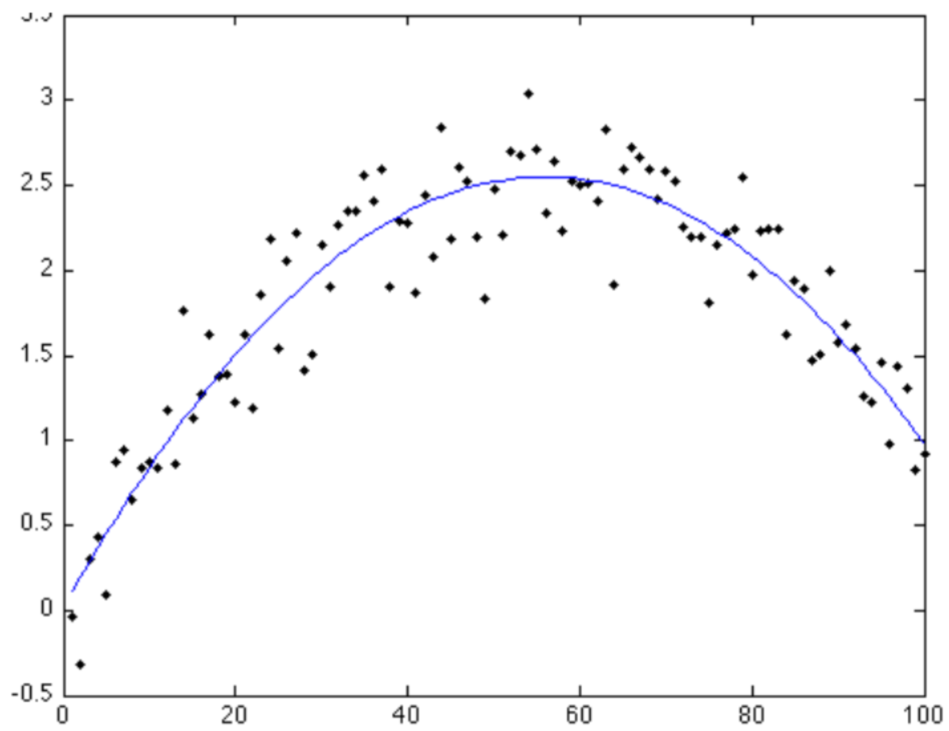


Standardization



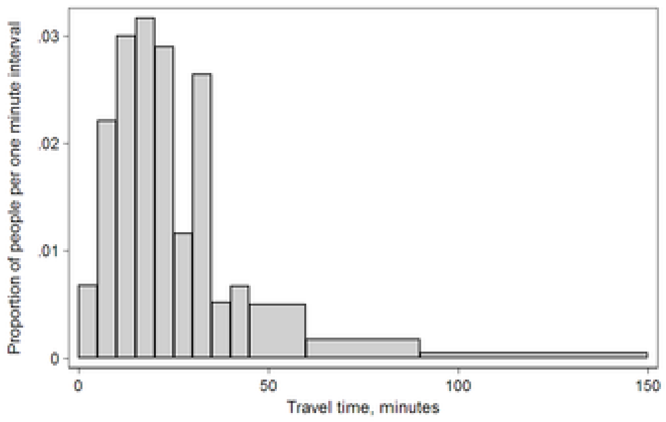
Noisy Data

- Noise = Random error in a measured variable
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention



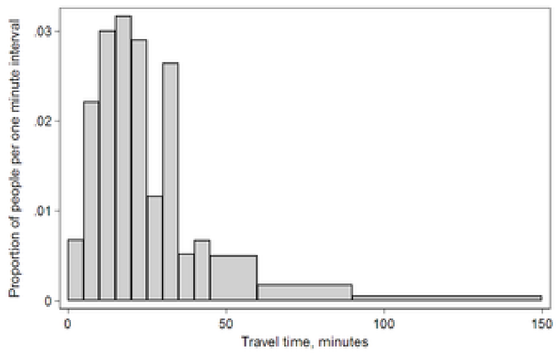
Noisy Data: What to Do?

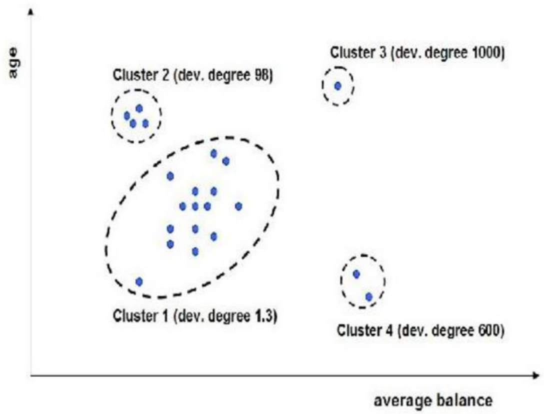
- Binning (quantization)
  - Replace data with bin centers



Noisy Data: What to Do?

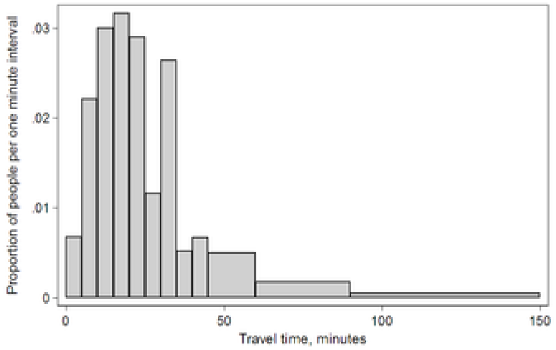
- Binning (quantization)
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers

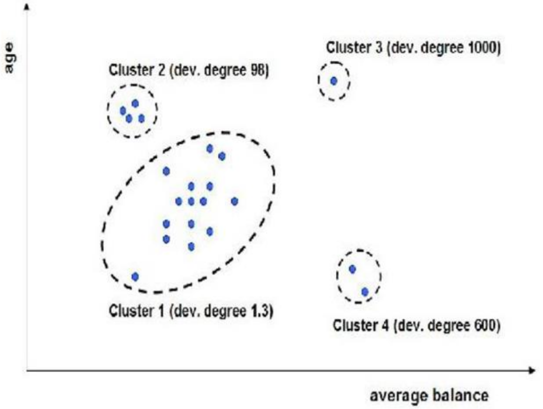




Noisy Data: What to Do?

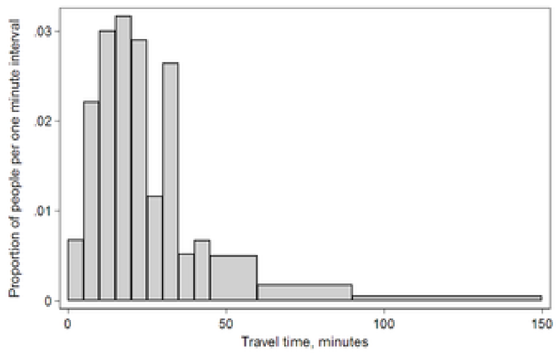
- Binning (quantization)
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually

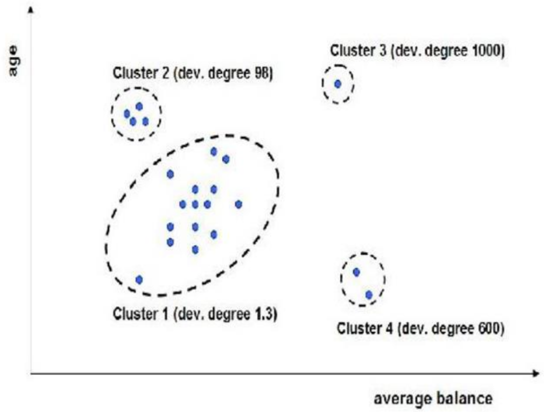


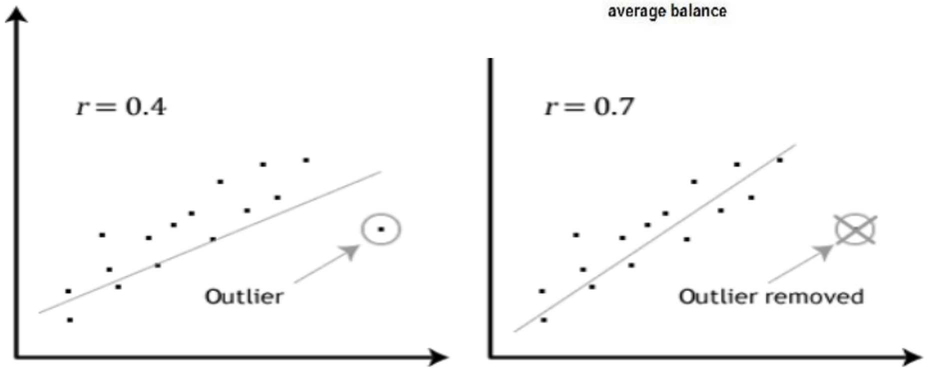


Noisy Data: What to Do?

- Binning (quantization)
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually
- Regression
  - Smooth data by fitting to a regression function

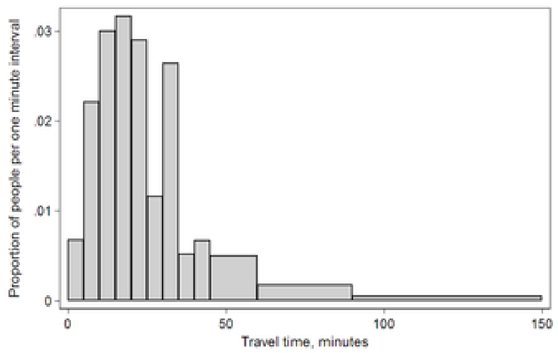


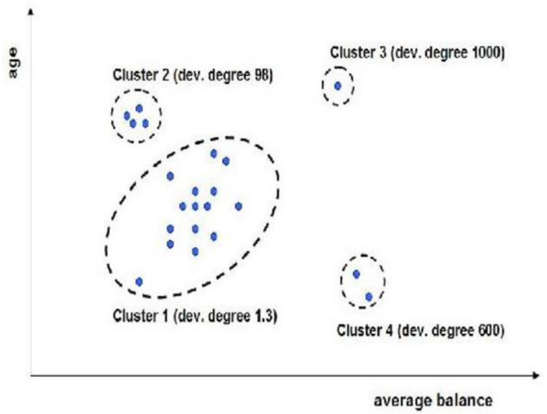


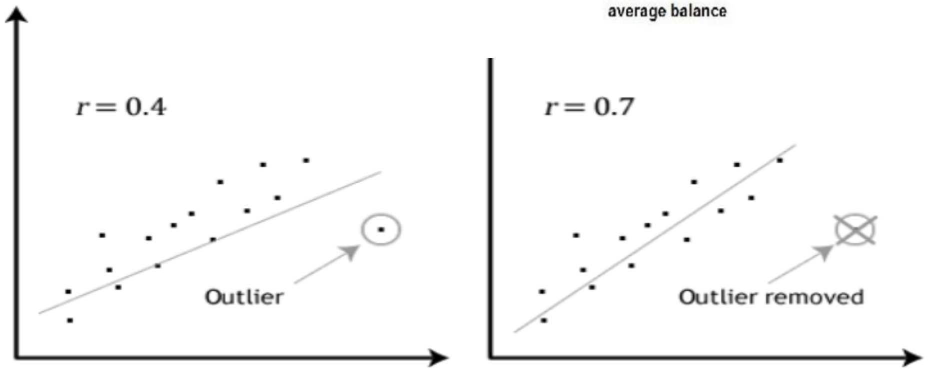


Noisy Data: What to Do?

- Binning (quantization)
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually
- Regression
  - Smooth data by fitting to a regression function
- <u>Outliers are not always noise! Be careful!</u>







Deal with Small Data

- Can you invent new data?

Deal with Small Data → Data Augmentation

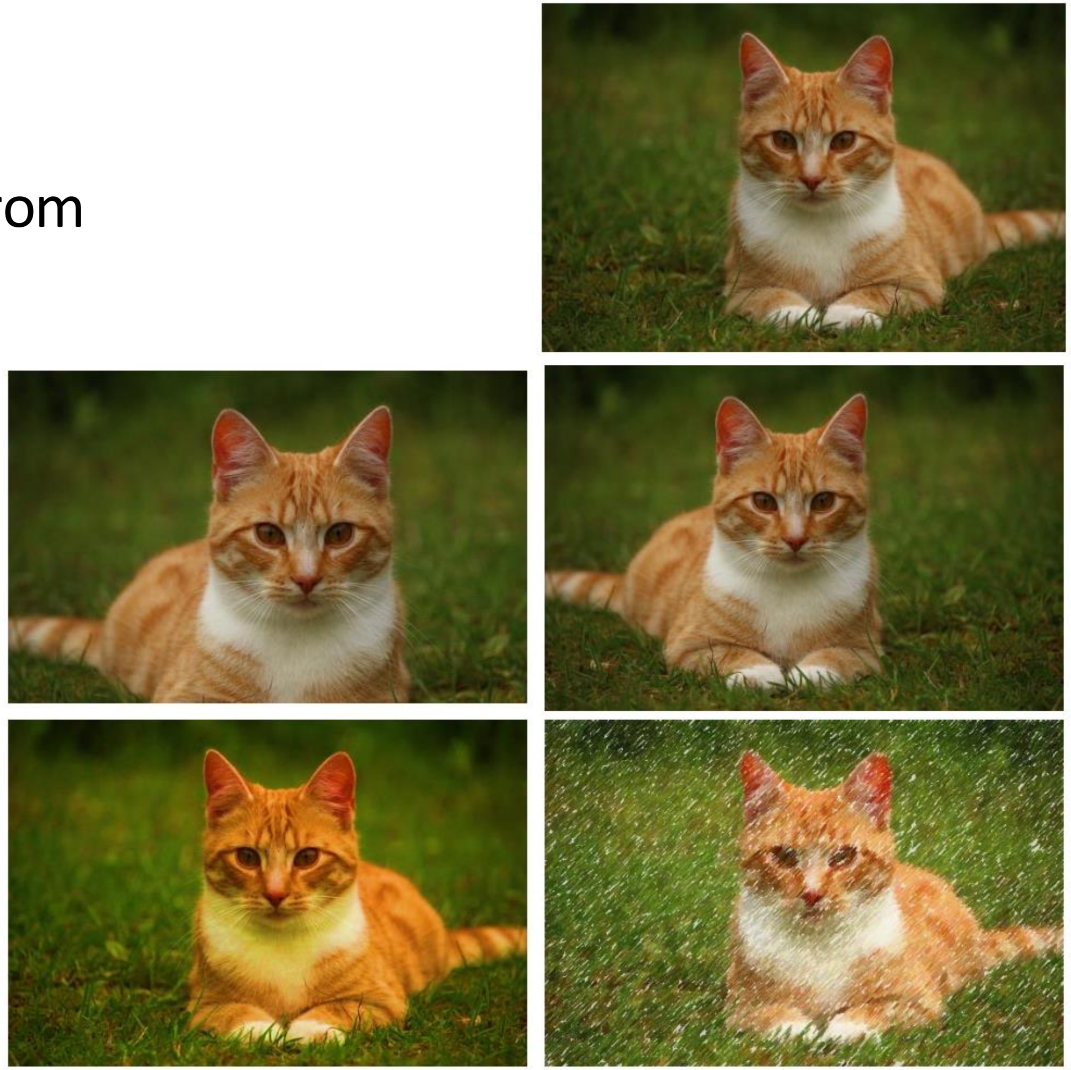
- Can you invent new data?
- Data Augmentation
  - Strategy to artificially synthesize new data from existing data

Deal with Small Data → Data Augmentation

- Can you invent new data?
- Data Augmentation
  - Strategy to artificially synthesize new data from existing data
- Common techniques are (for images)
  - Rotations
  - Translations
  - Zooms
  - Flips
  - Color perturbations
  - Crops
  - Add noise by jittering

Deal with Small Data → Data Augmentation

- Can you invent new data?
- Data Augmentation
  - Strategy to artificially synthesize new data from existing data
- Common techniques are (for images)
  - Rotations
  - Translations
  - Zooms
  - Flips
  - Color perturbations
  - Crops
  - Add noise by jittering



Deal with Big Data → Use Big Machines!

- Purpose
  - Use modern computing capabilities to process and analyze large data efficiently
  - Develop parallel data processing and analysis algorithms
  - Divide and Conquer

Param Sanganak at IITK
~1.6 Petaflops

Deal with Big Data → Use Big Machines!

- Purpose
  - Use modern computing capabilities to process and analyze large data efficiently
  - Develop parallel data processing and analysis algorithms
  - Divide and Conquer
- Sometimes you want to explore and analyze your data in your personal computer, but data might be too large to fit!

Param Sanganak at IITK
~1.6 Petaflops

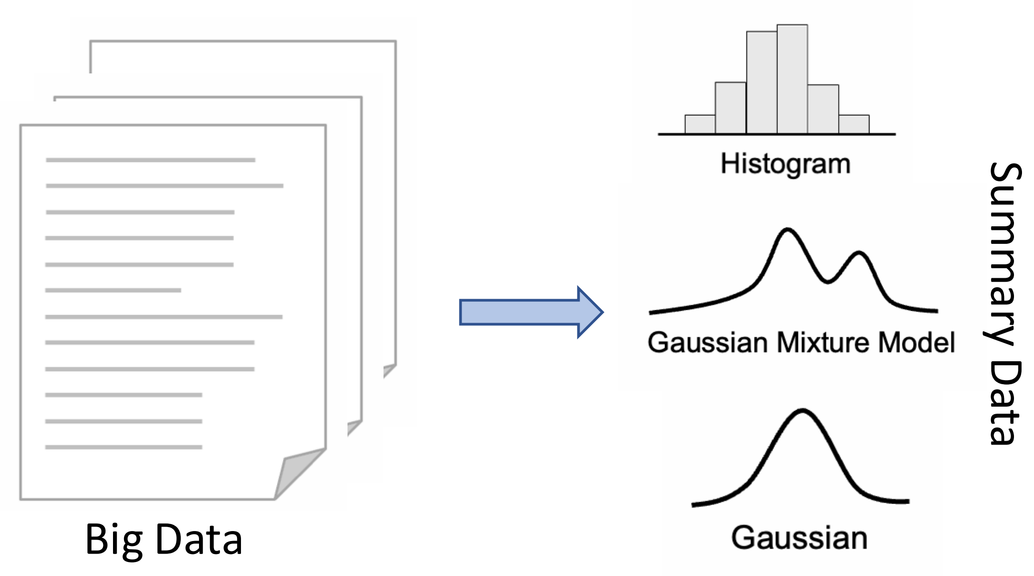
Deal with Big Data → Data Reduction!

- Purpose
  - Reduce the data to a size that can be feasibly stored
  - Reduce the data so an analysis algorithm can be feasibly run
- Alternatives?
  - Buy more storage
  - Access powerful computers
  - Develop more efficient algorithms
- In practice, all of this is happening at the same time

But the growth of data and complexities is faster and so data reduction is important !

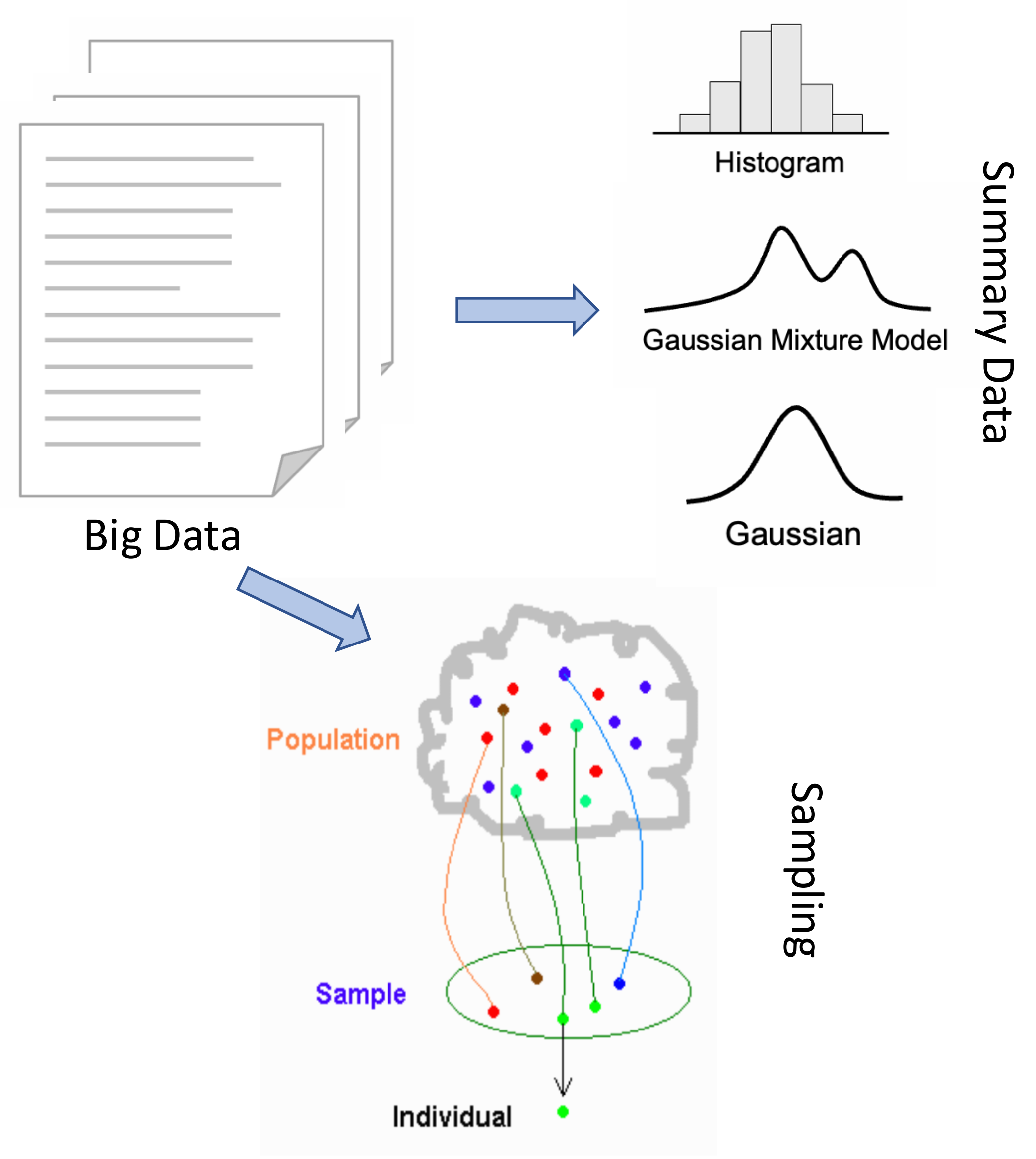
Data Reduction: How?

- Summarization
  - Binning
  - Distribution-based



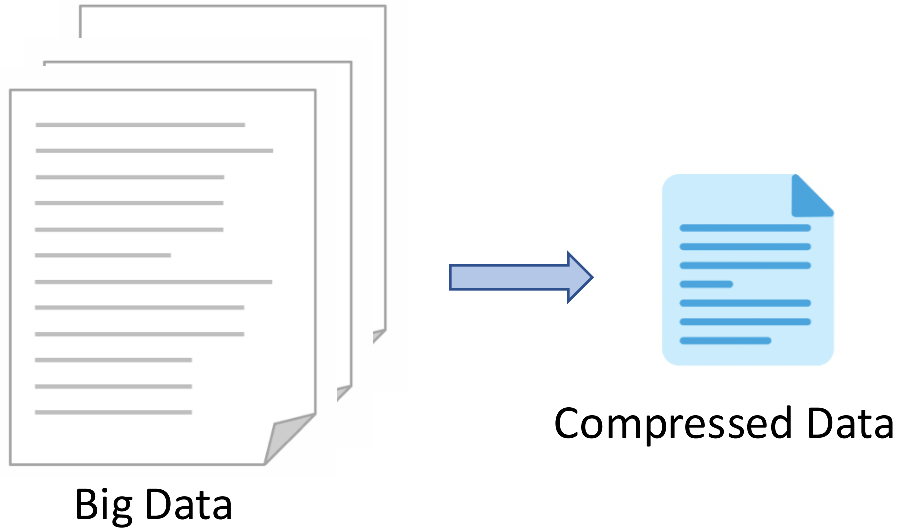
Data Reduction: How?

- Summarization
  - Binning
  - Distribution-based
- Sampling
  - Regular, Random, Stratified
  - Cluster-based, Adaptive/Data-driven
  - Importance-driven



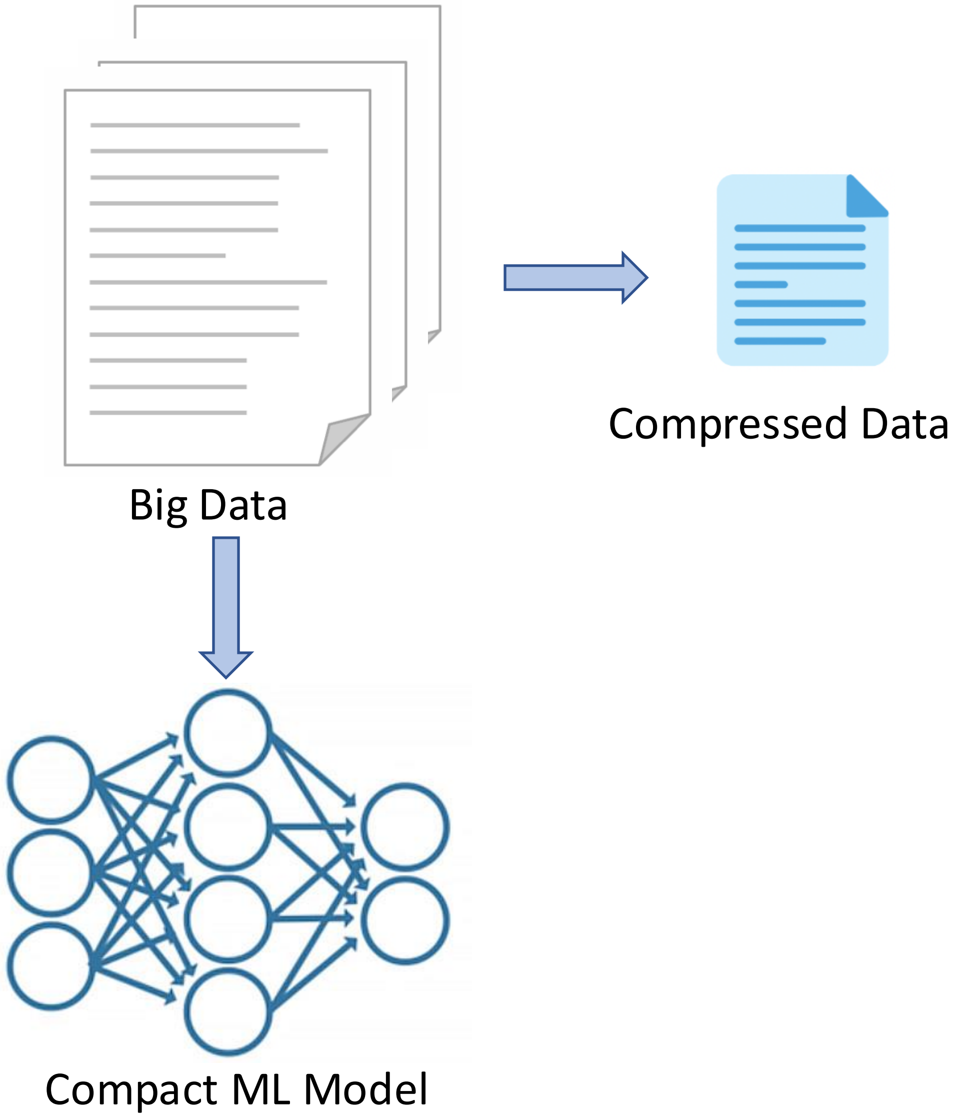
Data Reduction: How?

- Summarization
  - Binning
  - Distribution-based
- Sampling
  - Regular, Random, Stratified
  - Cluster-based, Adaptive/Data-driven
  - Importance-driven
- Data Compression
  - Compress floating points



Data Reduction: How?

- Summarization
  - Binning
  - Distribution-based
- Sampling
  - Regular, Random, Stratified
  - Cluster-based, Adaptive/Data-driven
  - Importance-driven
- Data Compression
- Machine/Deep Learning



Methodologies of Scientific Data Analysis

Different Methodologies of Scientific Data Analysis

- Exploratory data analysis
- Statistical data analysis
- Predictive data analysis
- Topological data analysis

Exploratory Scientific Data Analysis

Exploratory Data Analysis (EDA)

- Look at the data before making any assumptions or building sophisticated models
- Interact with the original form of raw data without any kind of transformations
- Often Visualization of the data is the key
- Enables Human-in-the-loop exploration
- A broad set of techniques are applicable in EDA
  - Basic statistical methods/plots

Objectives of EDA

- A preview of the overall data
- Spot visible patterns and trends
- Discover hidden relationships among data variables
- Identify outliers and anomalies
- Enable unexpected discoveries
- Enable formulation of new hypotheses
- Provide an indication of more sophisticated analysis tools/methods that can be applied to derive more detailed information

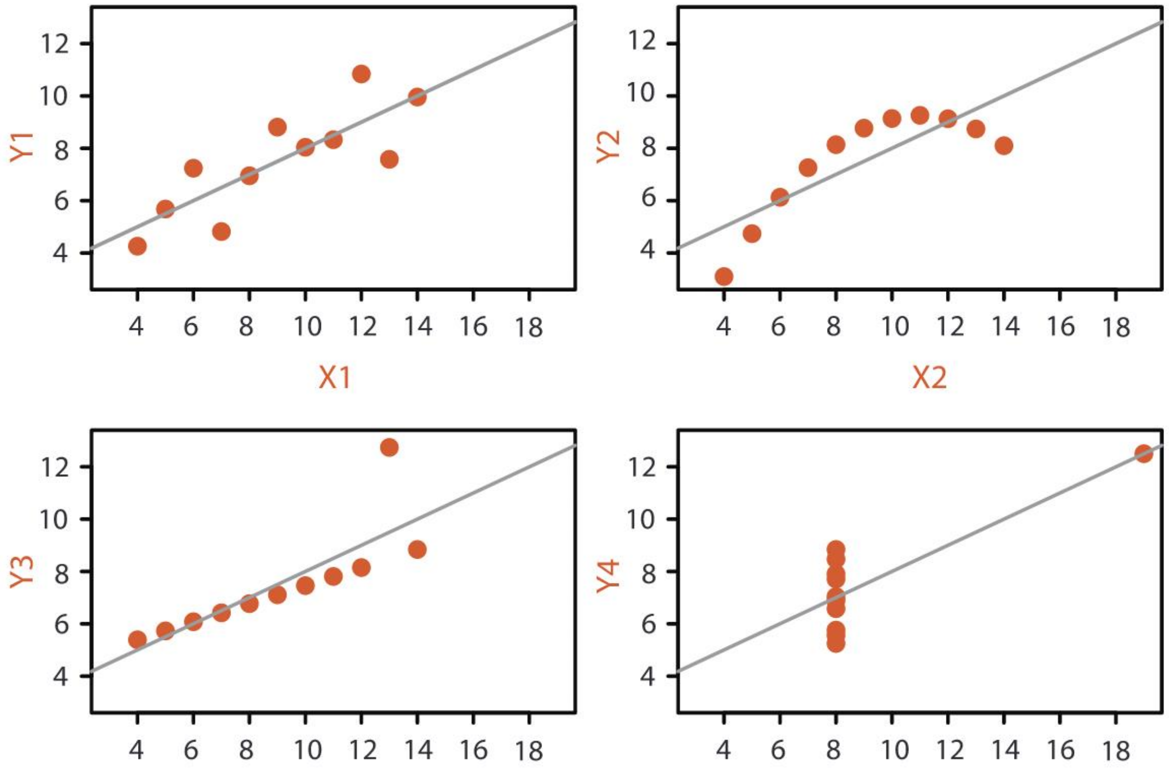
Essentially, during EDA we try to get a sense of our data when we see it for the first time

Example

Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

Example

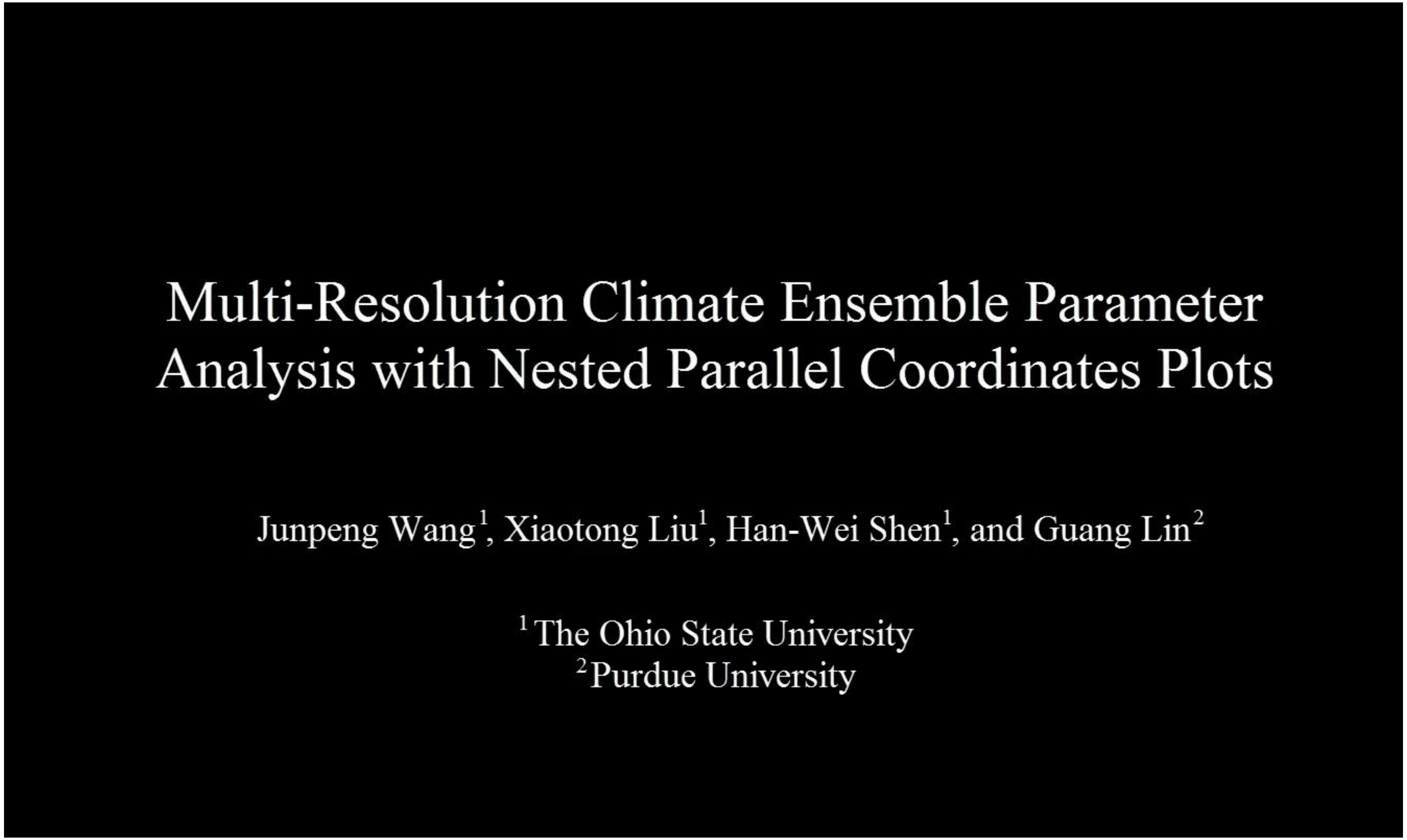


Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

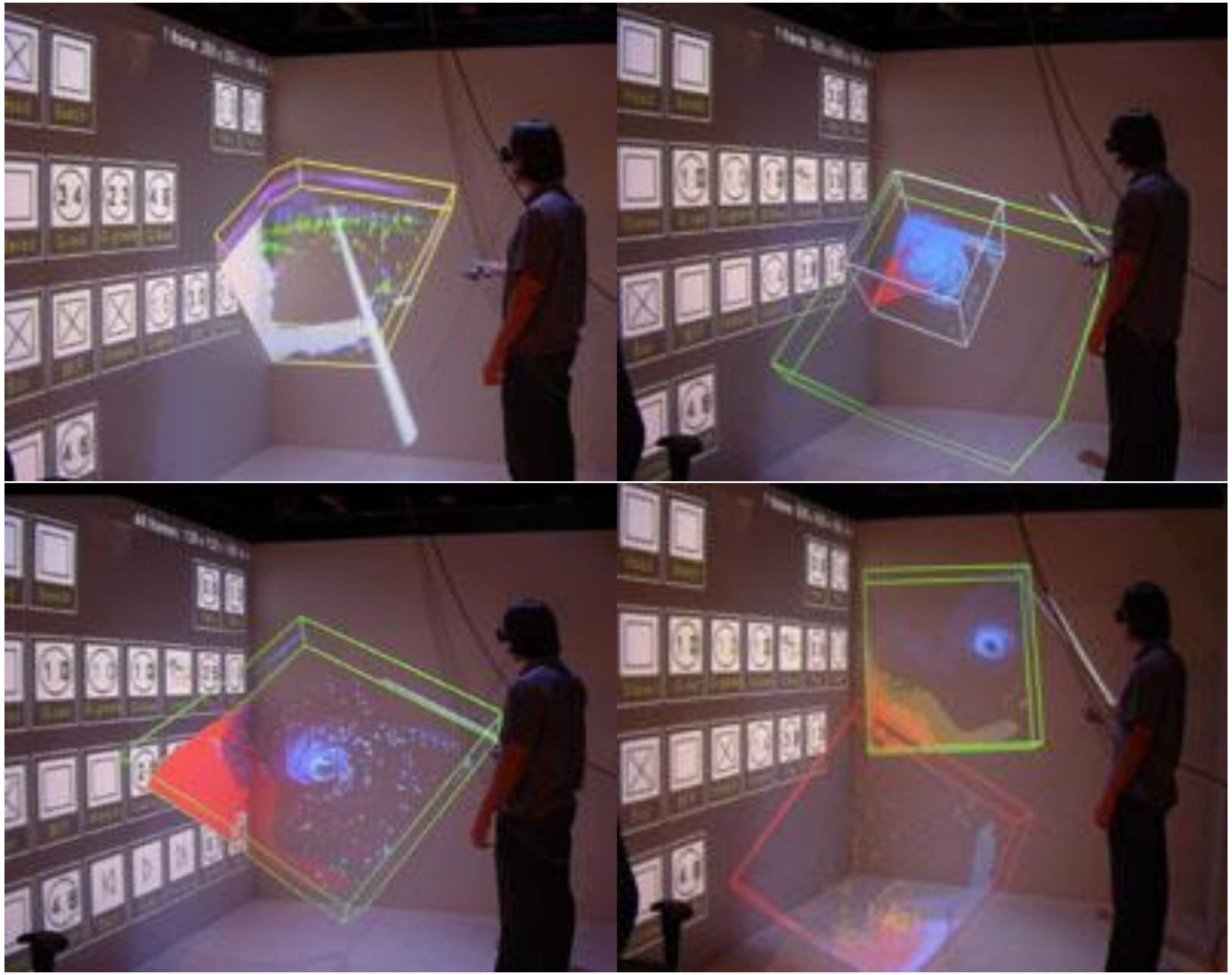
Lesson: Summaries may not be always enough; we need to see the data in visual form to comprehend the patterns in it

EDA for Scientific Data



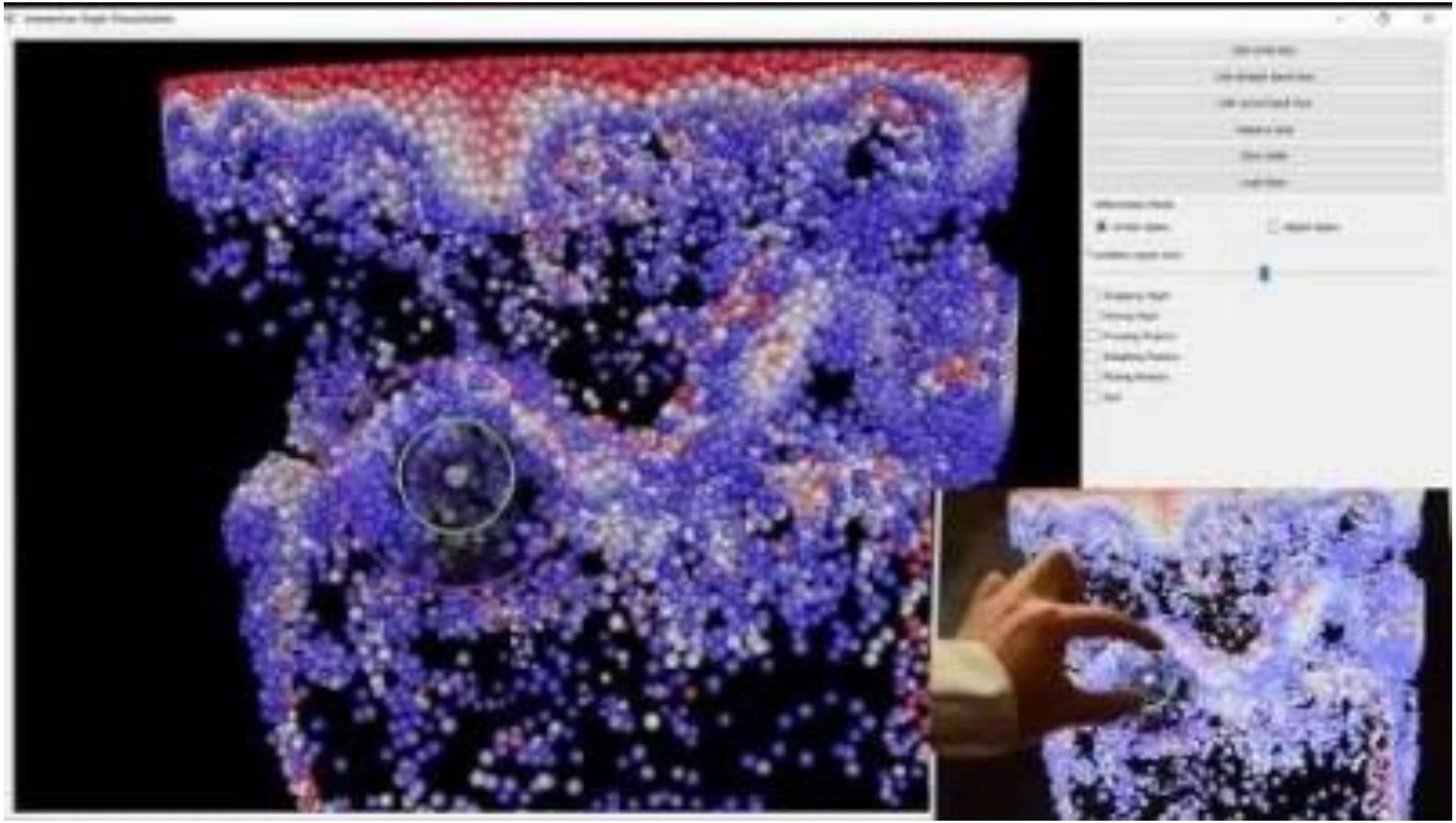
Multi-Resolution Climate Ensemble Parameter Analysis with Nested Parallel Coordinates Plots

EDA for Scientific Data



http://vis.computer.org/vis2004contest/brown/

EDA for Scientific Data



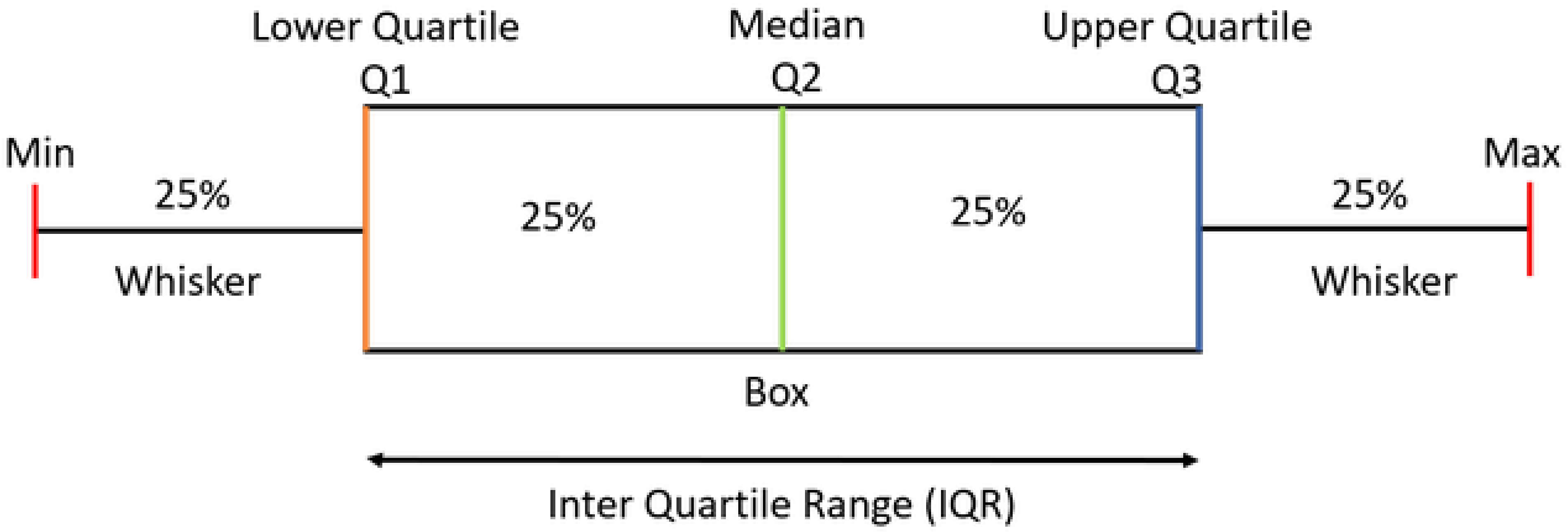
GlyphLens: View-dependent Occlusion Management in the Interactive Glyph Visualization

Statistical Scientific Data Analysis

Statistical Data Analysis

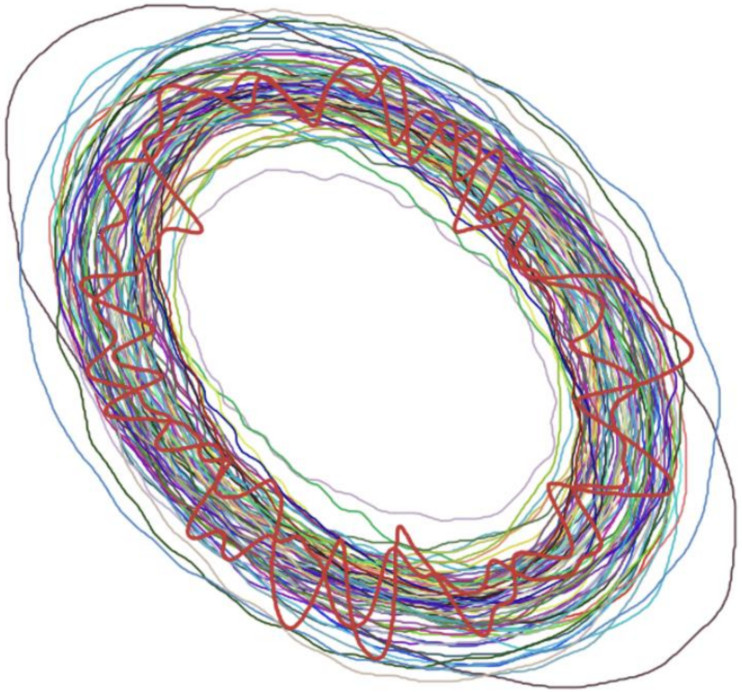
- Extracting meaningful insights from the data
- Descriptive statistics: provide a summary of the main characteristics of the data
  - Measures such as mean, median, standard deviation, and quartiles help understand the central tendency and spread of the data
- Build effective statistical representation of the large data and use it to perform feature extraction, visualization, and scientific discovery
  - Represent data as statistical distributions
  - Represent data as (sub)samples
  - Represent data as other form of statistical models

Statistical Data Analysis: Box and Whisker Plot



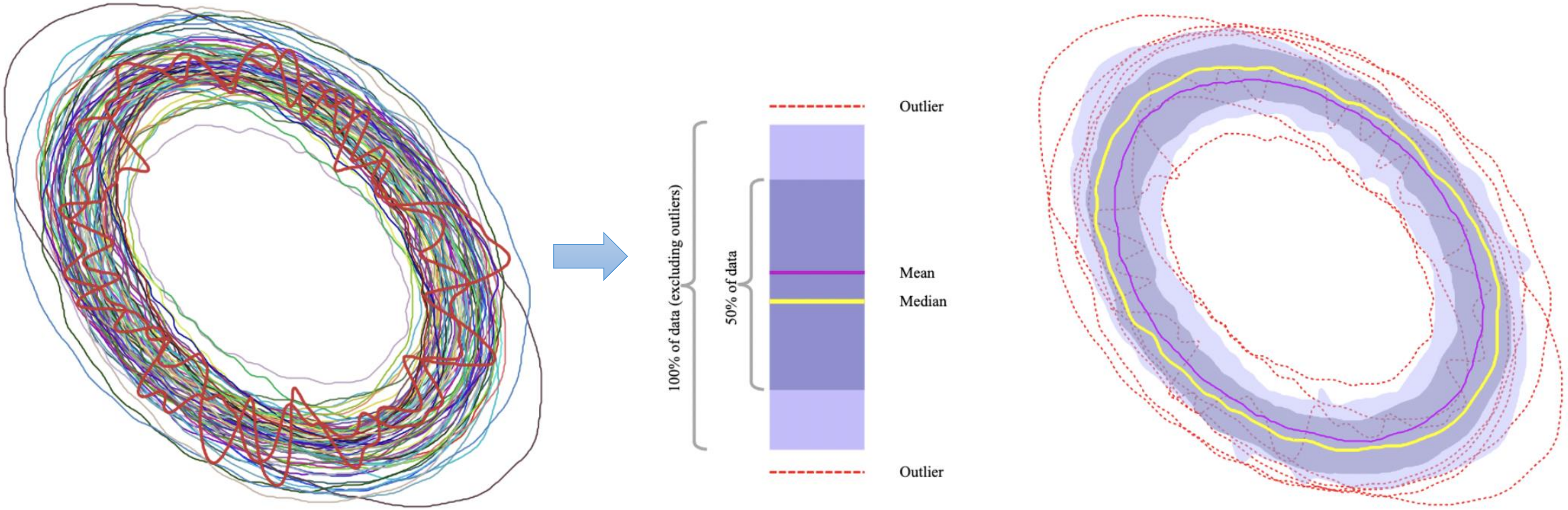
Source: Google

Statistical Data Analysis: Contour Box Plot



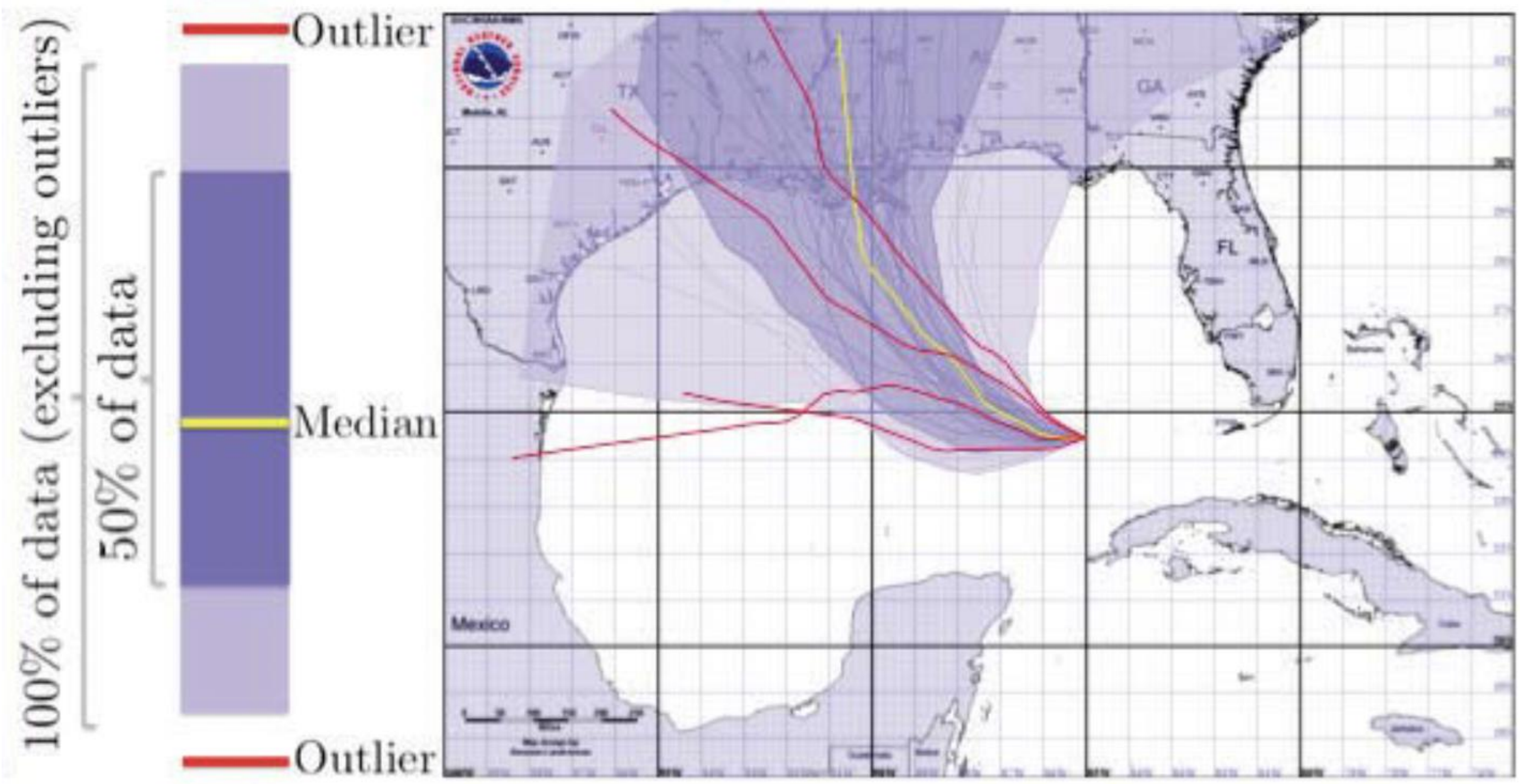
A collection of lines, but can you see what is going on?

Statistical Data Analysis: Contour BoxPlot

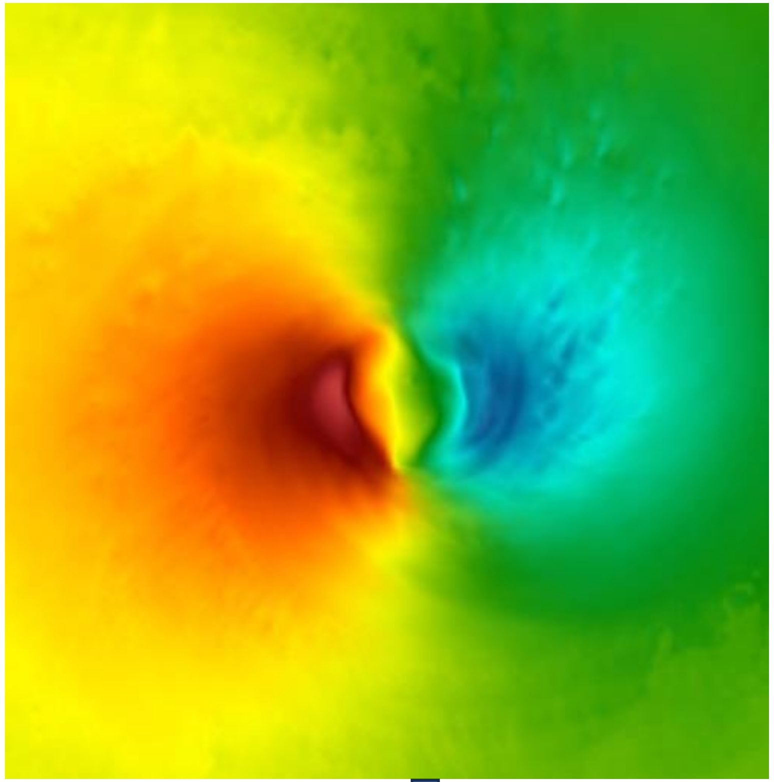


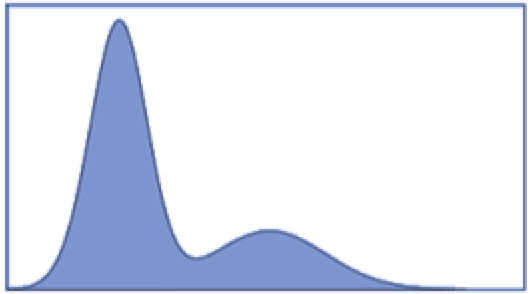
A collection of lines, but can you see what is going on?

Statistical Data Analysis: Curve BoxPlot

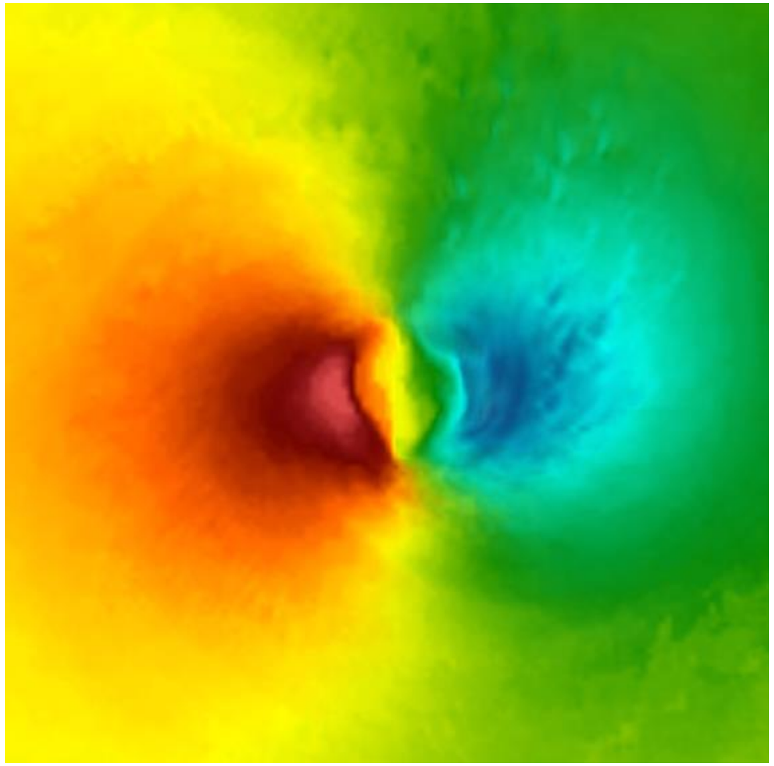


Statistical Data Analysis using Distributions



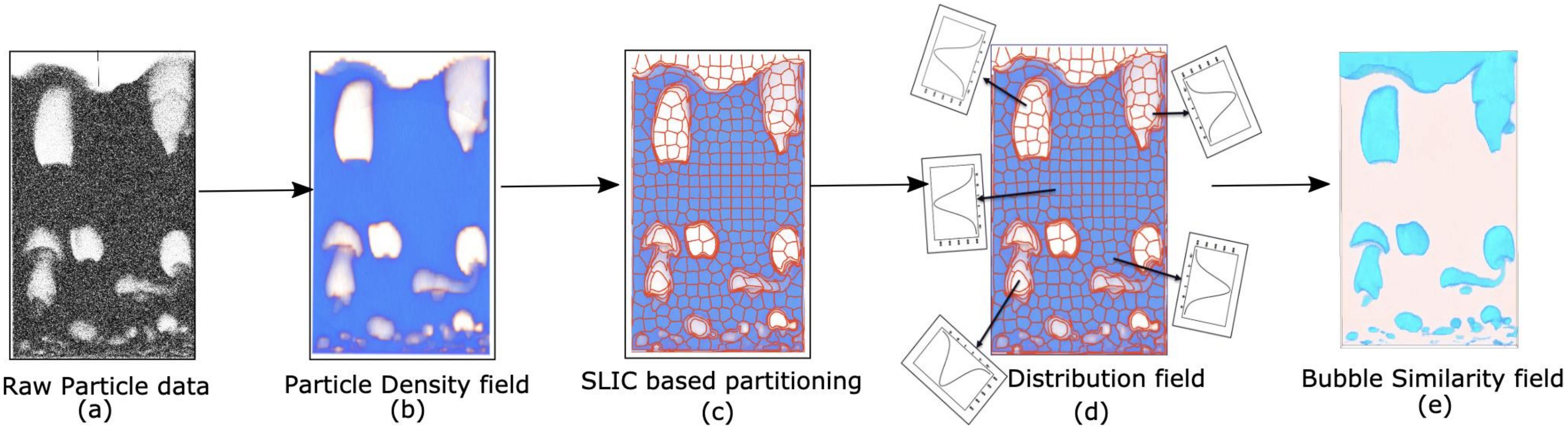


Convert raw data into compact distribution models



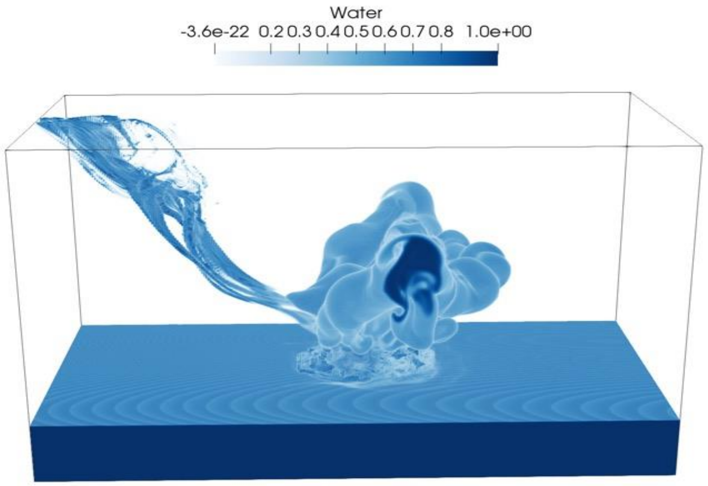
Reproduce data back from distribution model as needed during analysis

Statistical Feature Extraction

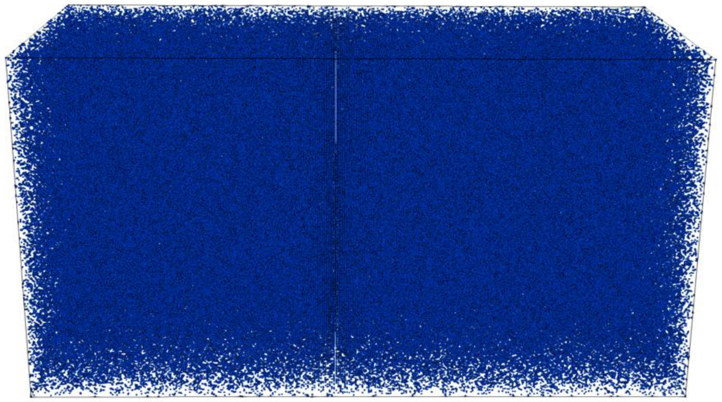


In Situ Feature Analysis for Large-scale Multiphase Flow Simulations, *Journal of Computational Science (Elsivier)*

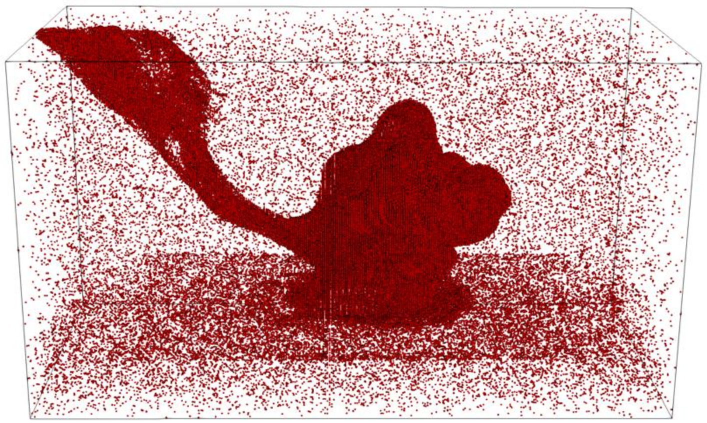
Statistical Data Analysis using Sampling



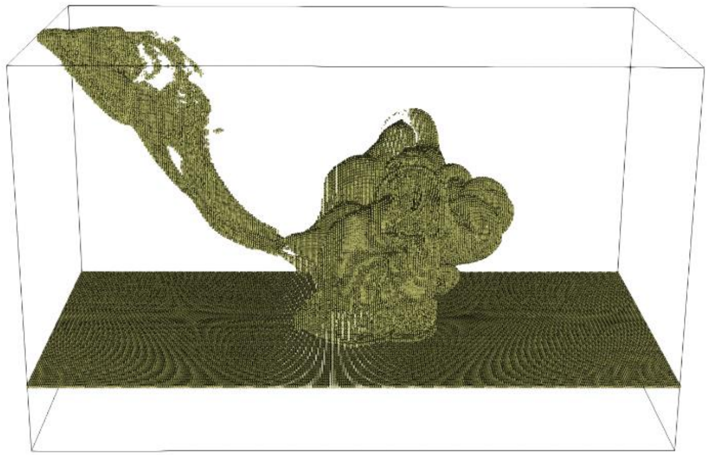
Original data showing the important region



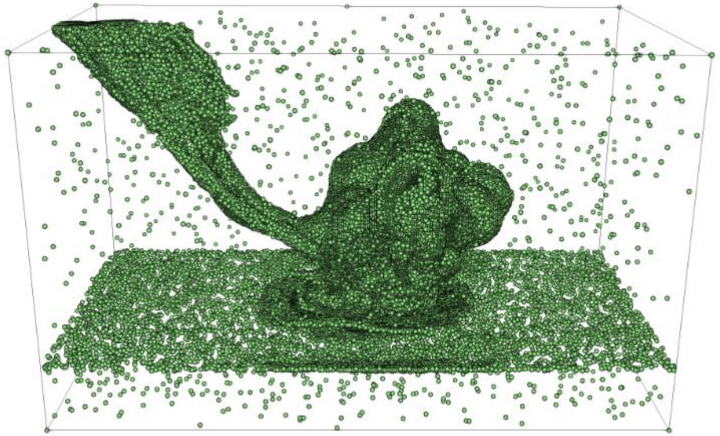
Random Sampling



Probability-based Sampling



Gradient-based Sampling



Joint Sampling

Predictive Scientific Data Analysis

Predictive Data Analysis using ML/DL Models

- Build efficient machine learning (can also be statistical) models of large scientific data
- Develop analysis and visualization techniques built on top of ML models
- Models are compact and smaller than the full-scale data
  - Easy to manage
  - Facilitates significant storage reduction
  - Allow interactive data analysis as if we have the entire data
- Image-based approaches
- Data space approaches

InSituNet: Image-Space Approach for Data Analysis

- In situ training data collection from ensemble simulations
- Offline training of InSituNet
- Interactive post-hoc exploration and analysis

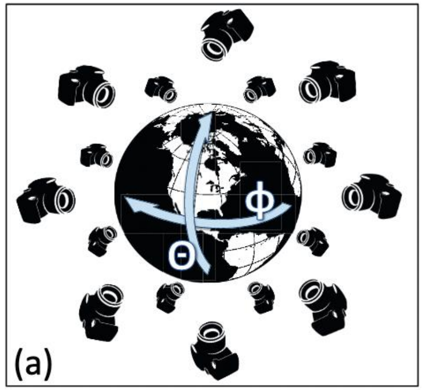
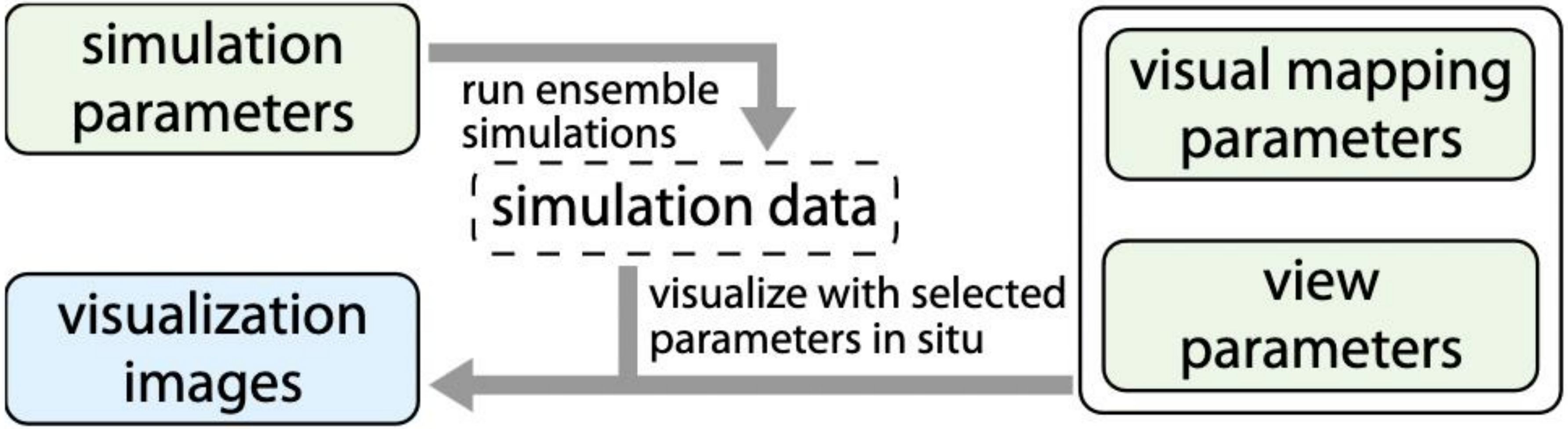
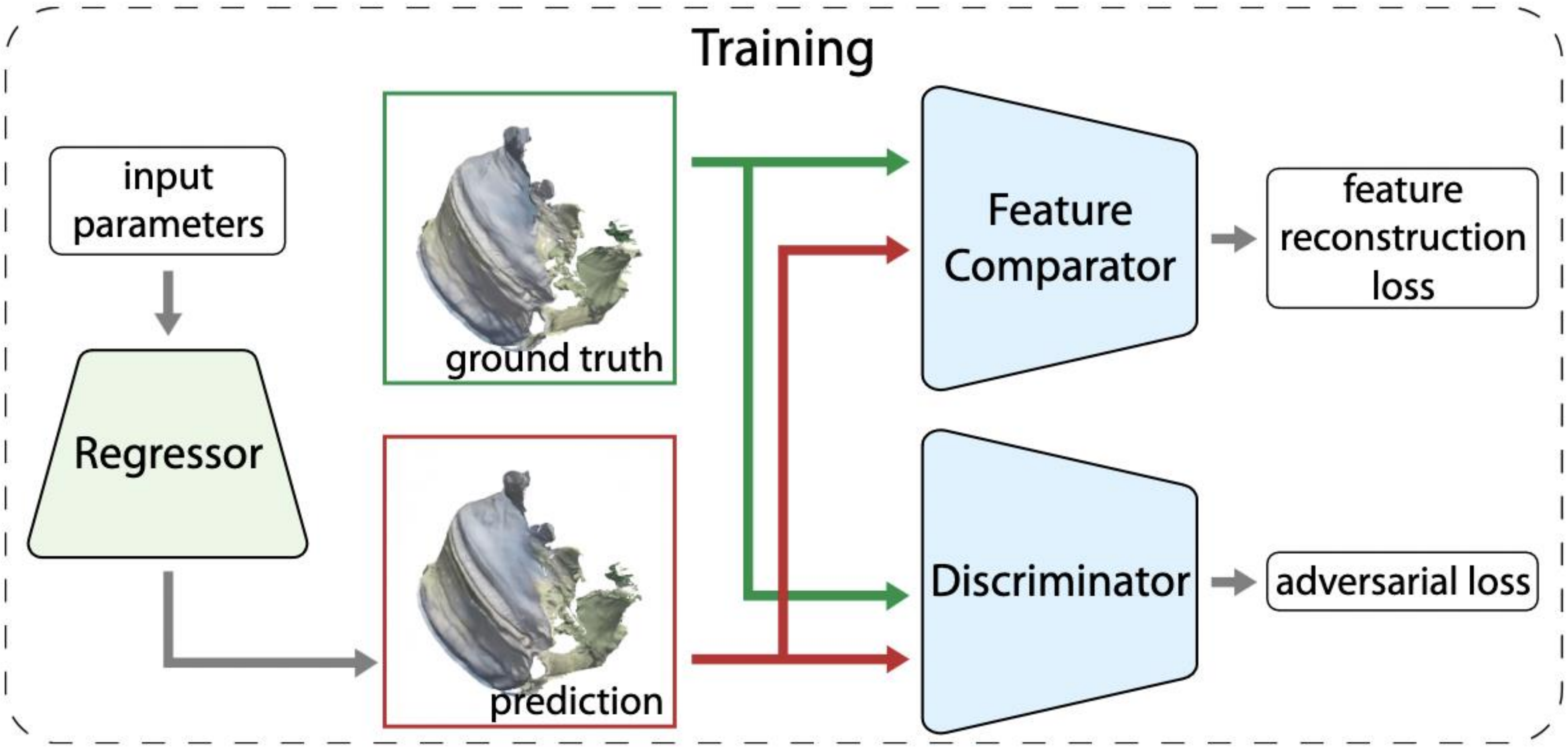


Image: D. Banesh et al.



InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations, He et al. TVCG

InSituNet: Image-Space Approach for Data Analysis



InSituNet: Image-Space Approach for Data Analysis

InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations

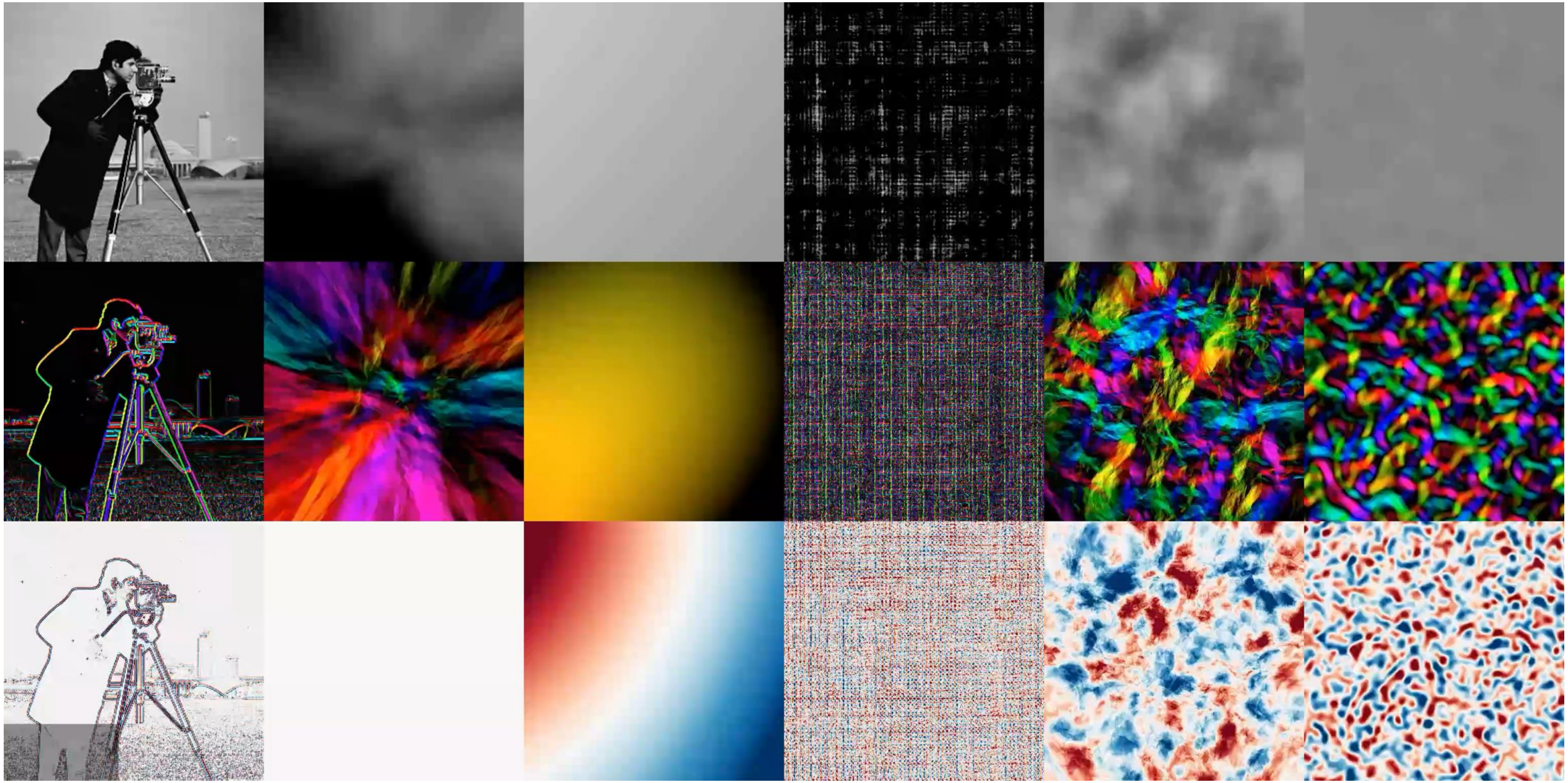
Online Submission ID: 1048

Data Space Approach: Scalar Data Compression

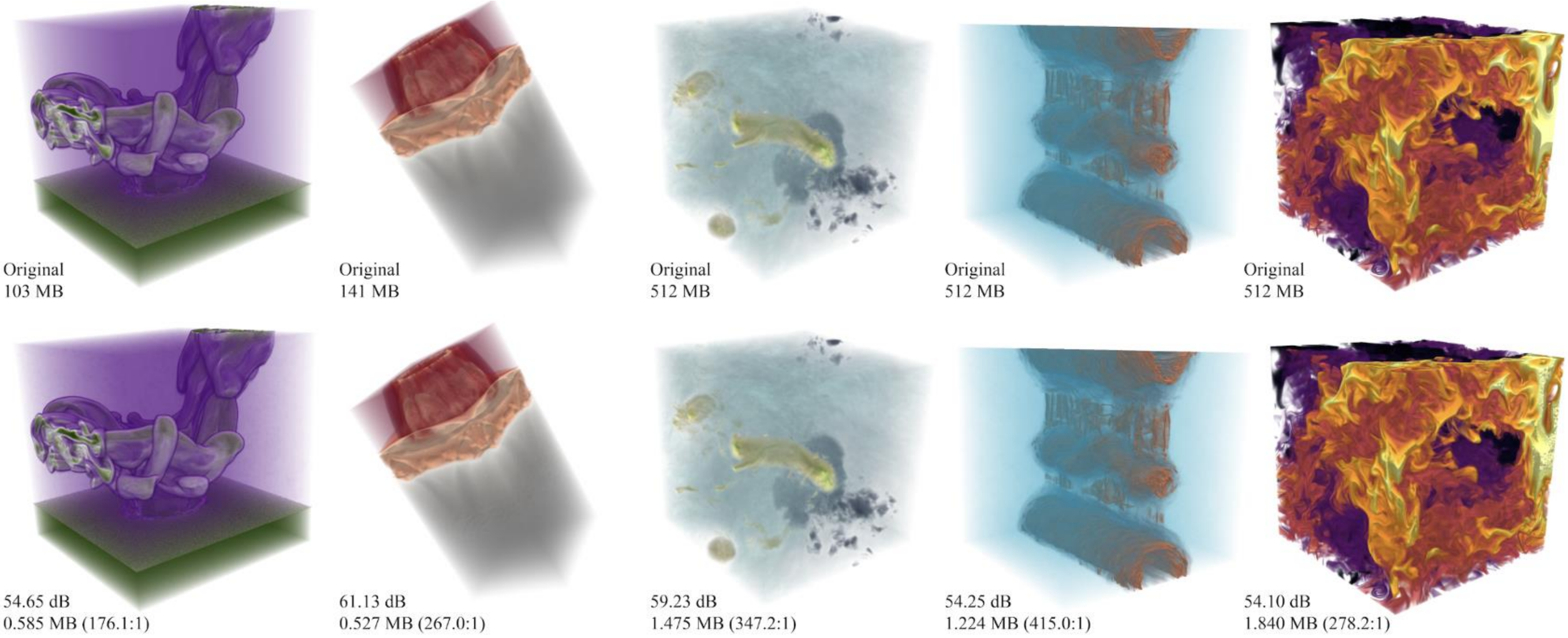
- A multi-layer perceptron network with **sinusoidal activation function**
- Excellent for modeling **coordinate-based** data sets
  - Images
  - Scientific data sets etc.
- Can learn higher order gradients of the data

https://www.vincentsitzmann.com/siren/

SIREN: Sinusoidal Representation Network



Data Space Approach: Scalar Data Compression



Compressive Neural Representations of Volumetric Scalar Fields, Lu et al., CGF

Topological Scientific Data Analysis

Topological Data Analysis

- Mathematical Topology + Computer Science for Scientific Data Analysis
- **Topology**: The study of geometrical properties and spatial relations unaffected by the continuous change of shape or size of figures
- Classic Example:

=

Topologically speaking

Source: google

Topological Data Analysis