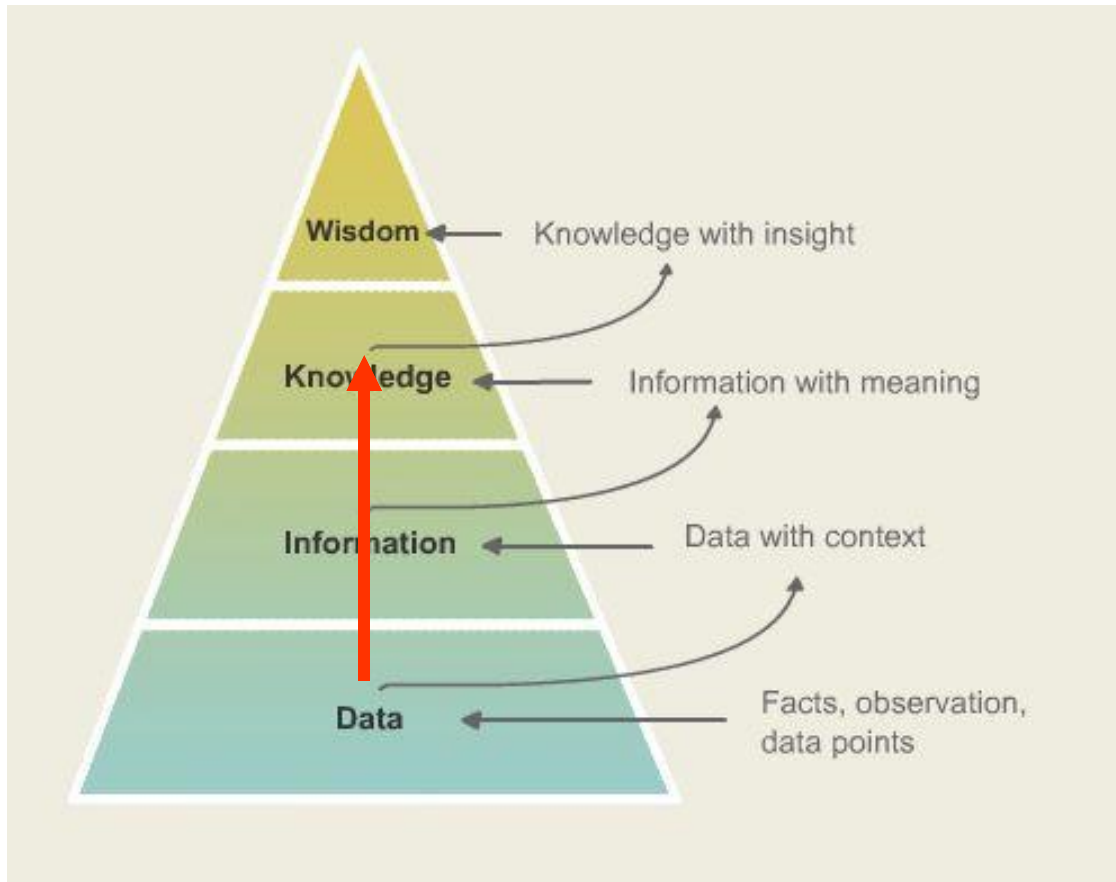
Data-Mining, Clustering and Cyberinfrastructure: An Information Science and Engineering Perspective

Xiaolong "Luke" Zhang

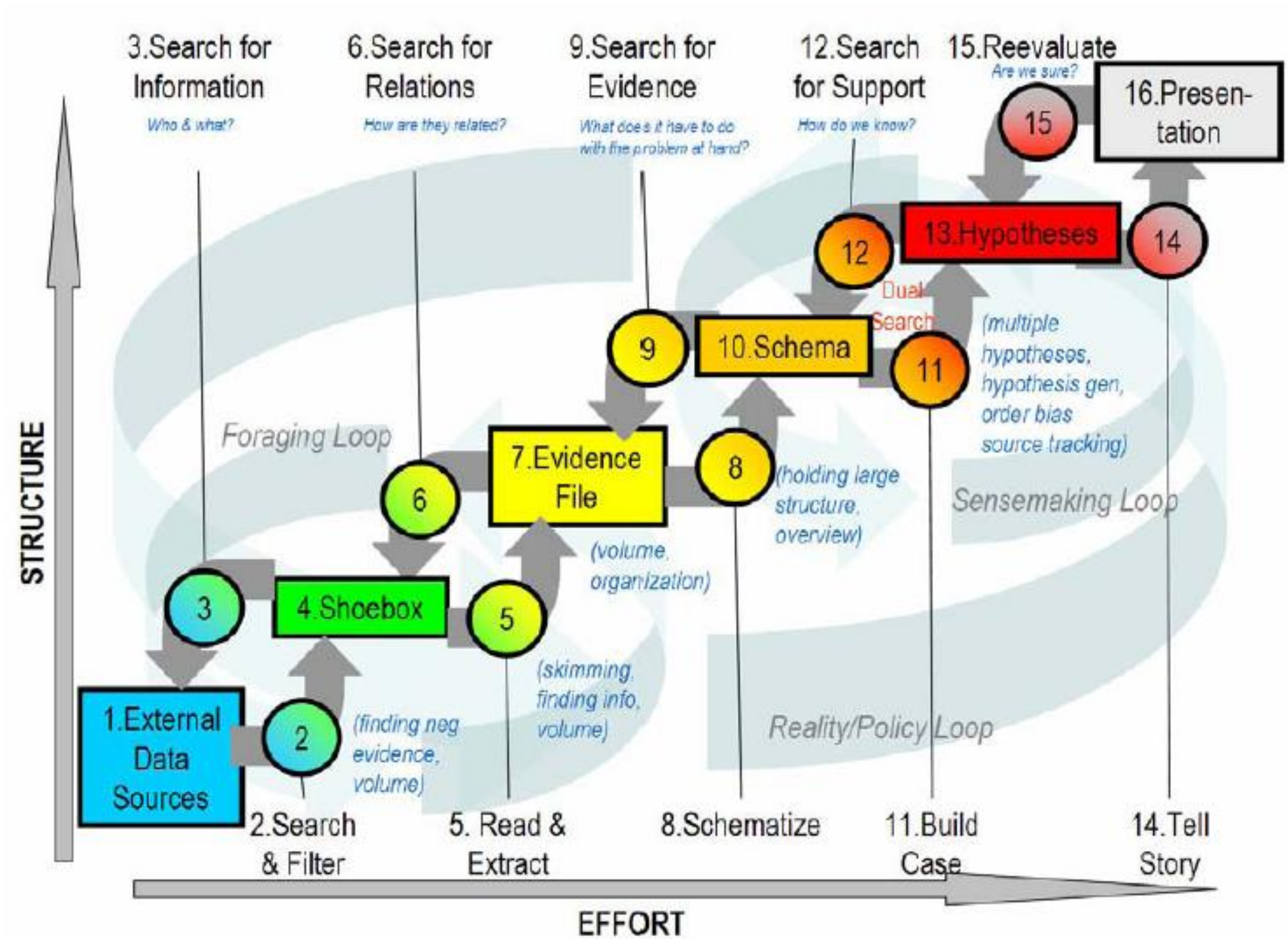
College of Information Science and Technology
Department of Industrial and Manufacturing Engineering
Penn State University

Core Research Question

- How to help people make sense of big data with interactive visualization?

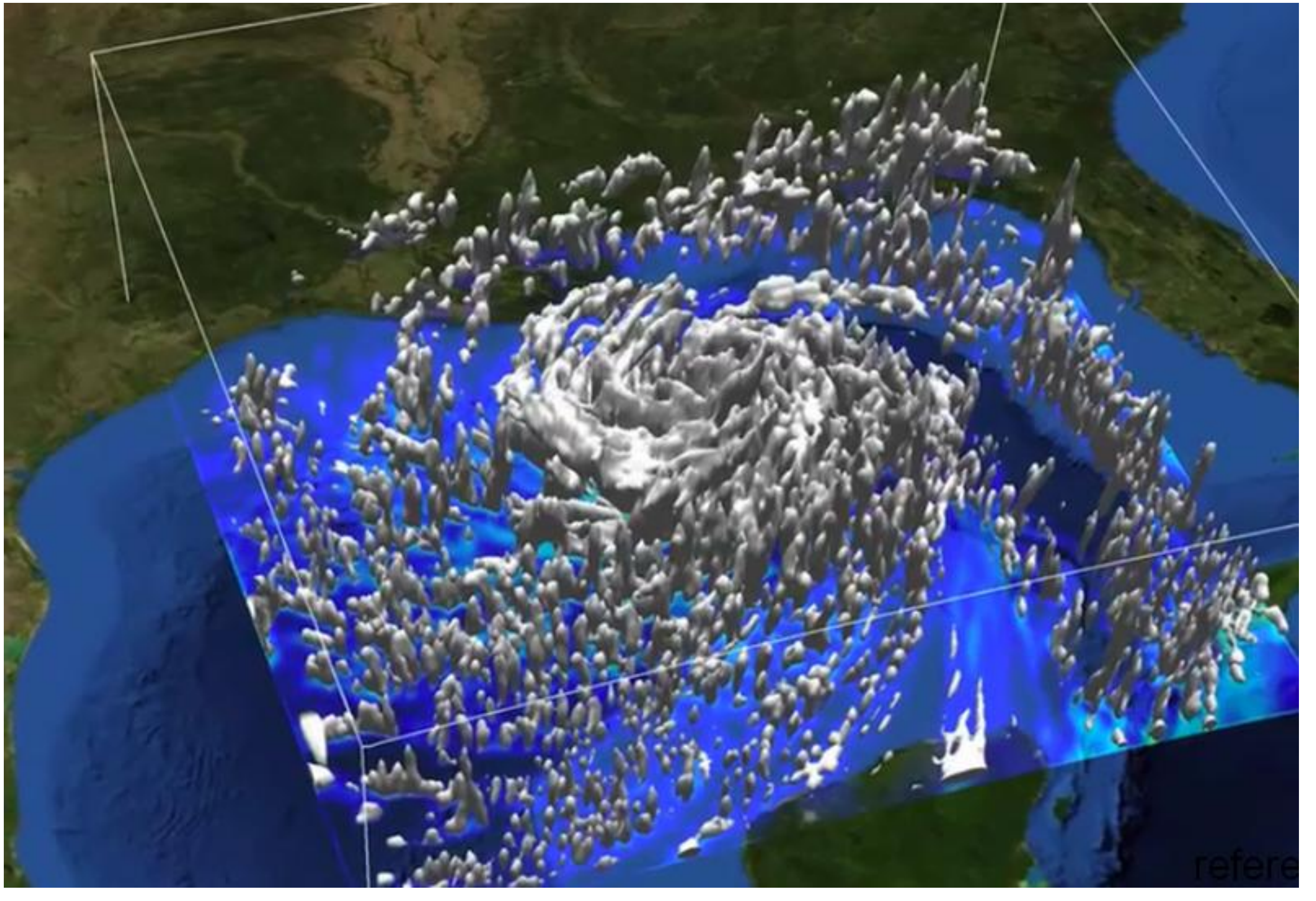


Sensemaking of Data



(Pirolli & Card, 2005)

A Motivation Scenario: Weather Forecasting



(Credit: F. Zhang, Department of Meteorology, Penn State &
Texas Advanced Computing Center)

How Did We Get Here?

- In-depth analysis
  - Comparison of models
  - Analysis at different levels of granularity
  - Explore “what-if” situations
  - .
  - .

One Challenge in Visual Analytics Involving Big Data

- Disconnection between data space and user space
  - Data space: complex models, large datasets
    - Hard for people to understand
    - Need tools to discover and present the hidden patterns of data
      - Data-mining: data-oriented
  - User space: limited cognition resources and specific tasks
    - Need design to consider the cognition and task features
      - Visualization design: user-centered

My Strategy

- A "Work-centered" Approach
  - Work: data, algorithms, user tasks
- Collaborative research effort
  - Experts in statistics and data mining
  - Researchers in Human-Computer Interaction
    - Visualization, interactive system design
  - Domain experts in science and engineering

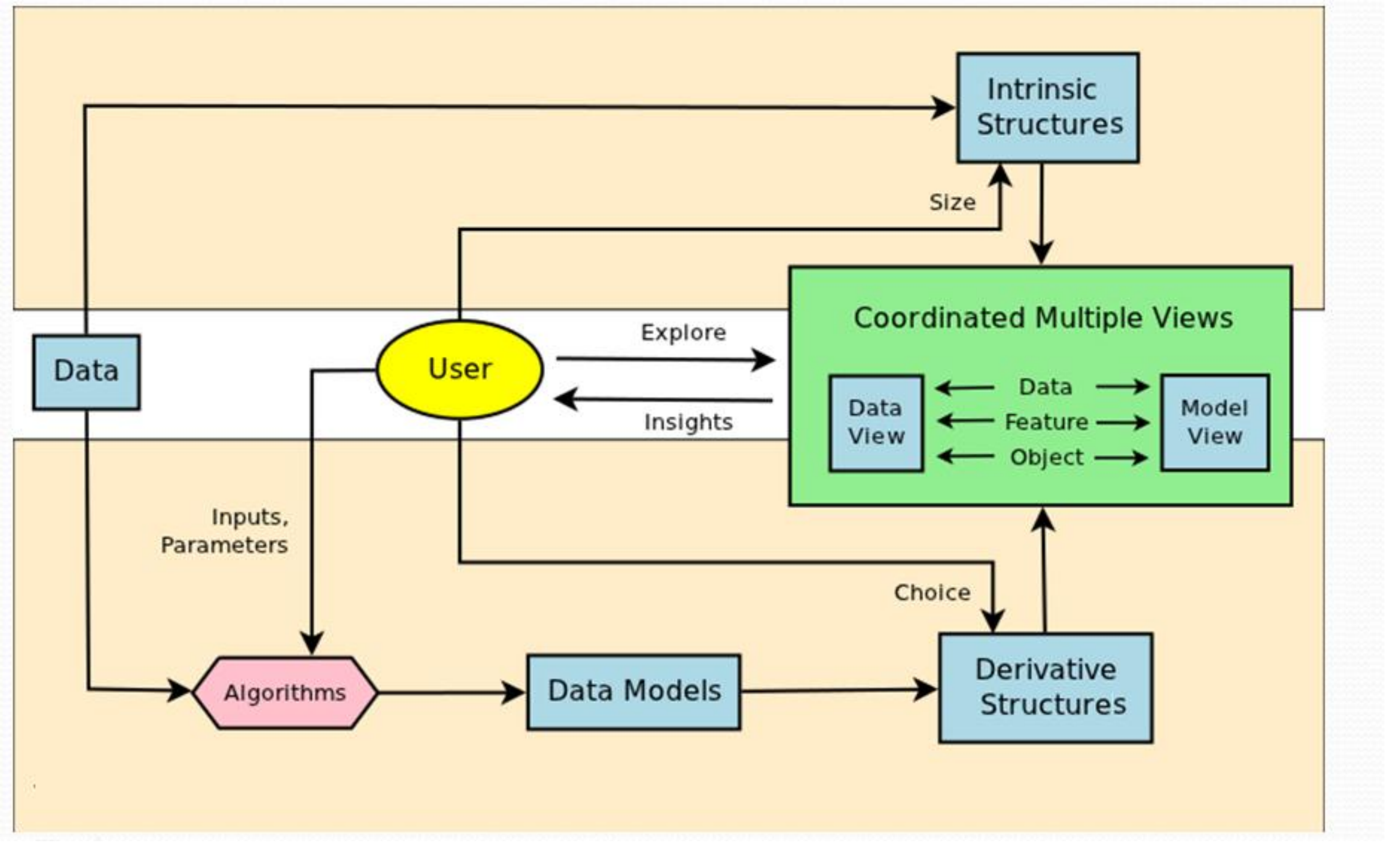
Our Goals

- Develop approaches to data clustering, dimension reduction, and variable selection based on geometric methods of mixture models
- Develop a technical infrastructure to support visual analytics empowered by a suite of statistical learning tools and interactive visualization tools
  - Visual analytics in science and engineering

Our Work

- Algorithms
  - Clustering methods based on mode association.
    - Hierarchical clustering to support analysis at different levels of detail.
- Technical infrastructure
  - Combine algorithms and interactive visualization tools

Architecture of the Infrastructure



Core Algorithm:
Hierarchical Mode Association Clustering

- Model expectation maximization (MEM)
  - Identify the modes of data clusters
- Mode association clustering (MAC)
  - Cluster data based on their distance to modes
- Hierarchical mode association clustering
  - Gradually change the bandwidth of distribution functions.

MEM

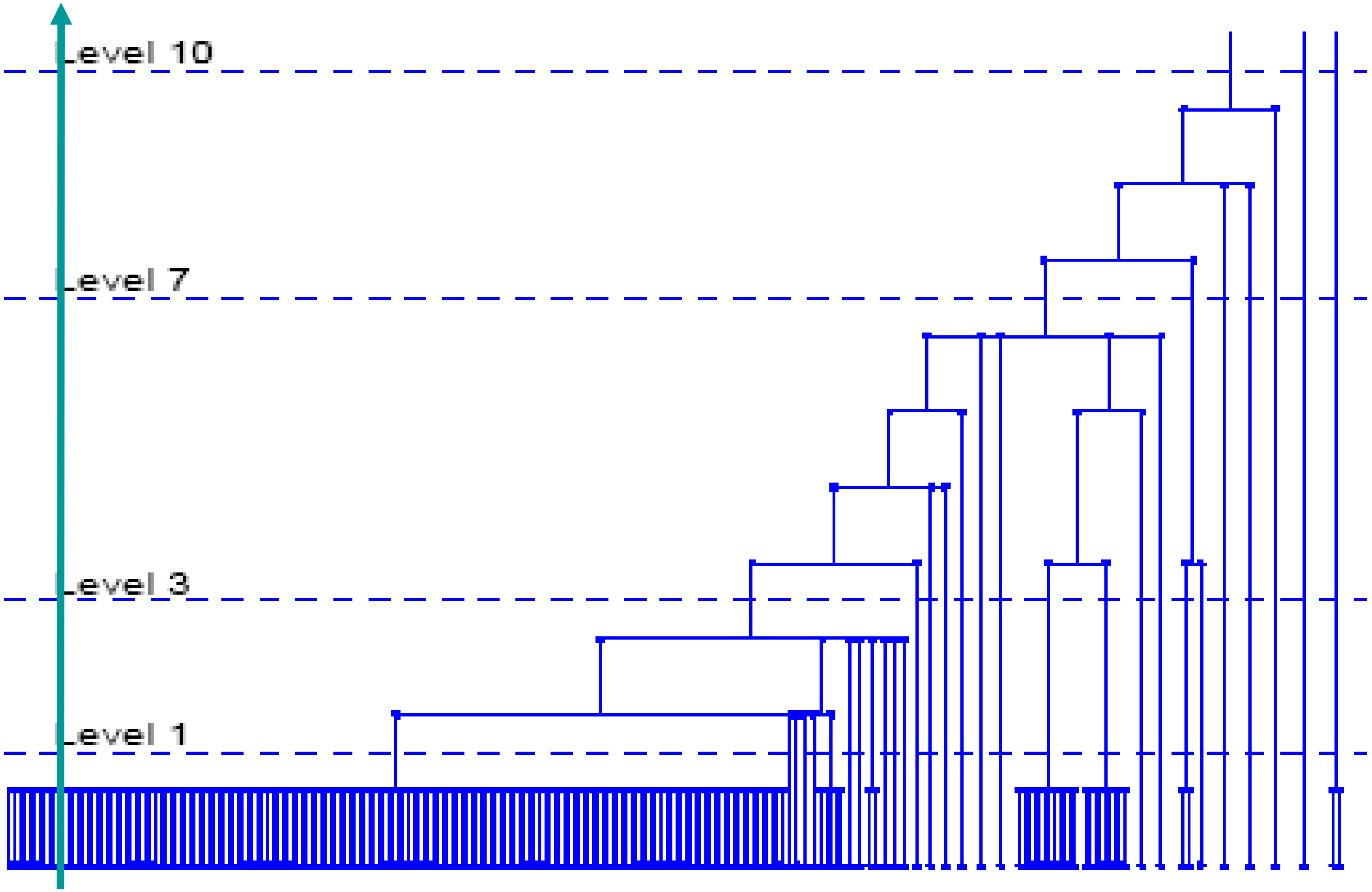
- Let a mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$.
  - $x \in \mathcal{R}^d$
  - π_k is the prior probability of mixture component k.
  - $f_k(x)$ is the density of component k.
- Given any initial value $x^{(0)}$, MEM solves a local maximum of the mixture by alternating two steps.

$$p_k = \frac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}, \; k = 1, \dots, K.$$

$$x^{(r+1)} = \underset{x}{\operatorname{argmax}} \sum_{k=1}^{K} p_k \log f_k(x) .$$

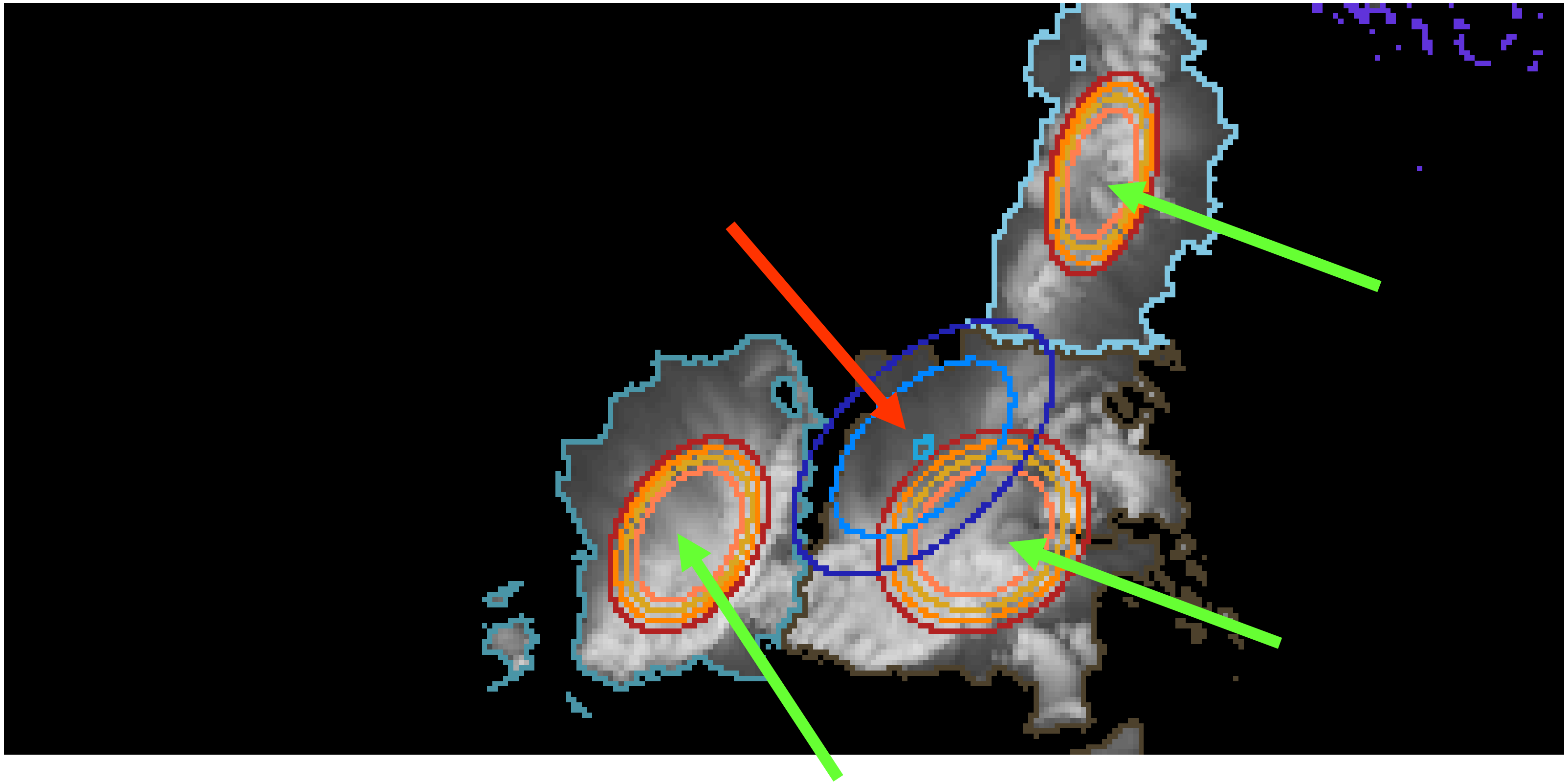
MAC

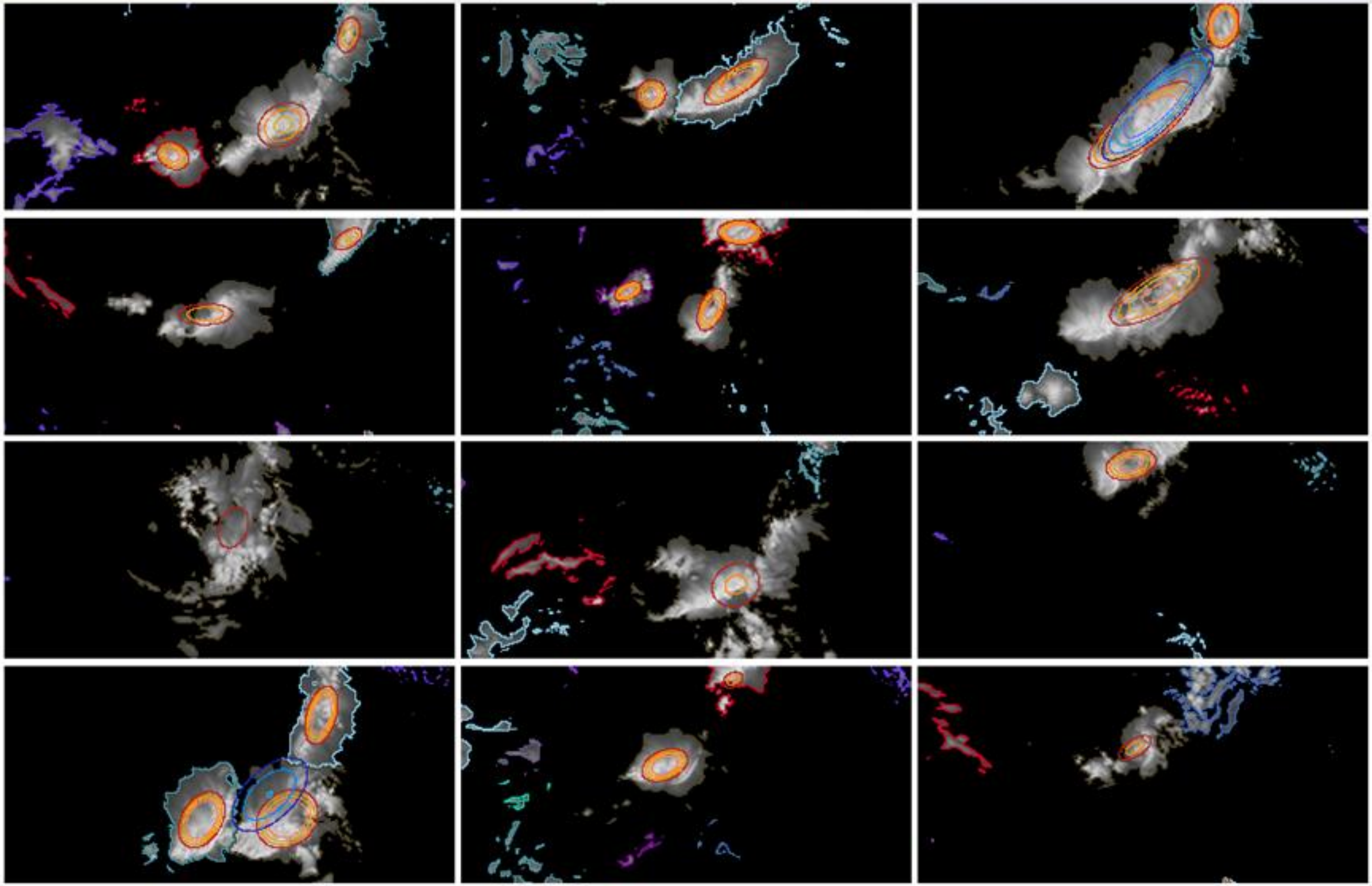
1. Form kernel density $f(x \mid S, \sigma^2) = \sum_{i=1}^{n} \frac{1}{n}\phi(x \mid x_i, D(\sigma^2))$, where $S = \{x_1, x_2, ..., x_n\}$.
2. Use $f(x|S, \sigma^2)$ as the density function. Use each x_i, $i = 1, 2, ..., n$, as the initial value in the MEM algorithm to find a mode of $f(x|S, \sigma^2)$. Let the mode identified by starting from x_i be $\mathcal{M}_\sigma(x_i)$.
3. Extract distinctive values from the set $\{\mathcal{M}_\sigma(x_i), i = 1, 2, ..., n\}$ to form a set G. Label the elements in G from 1 to $|G|$.
4. If $\mathcal{M}_\sigma(x_i)$ equals the kth element in G, x_i is put in the kth cluster.



Level 10
Level 7
Level 3
Level 1

Example:
Cloud Image Data Clustering





Interactive Visual Analytics

Example 1: Engineering Design

Design Task: Conceptual Ship Design

Design input variables:

Length (*L*), Beam (*B*), Depth (*D*), Draft (*T*), Block Coeff (C_B), and Speed (V_k).

Design output variables :

Transportation Cost (*TC*), Light Ship Weight (*LSM*) and Annual Cargo (*AC*).

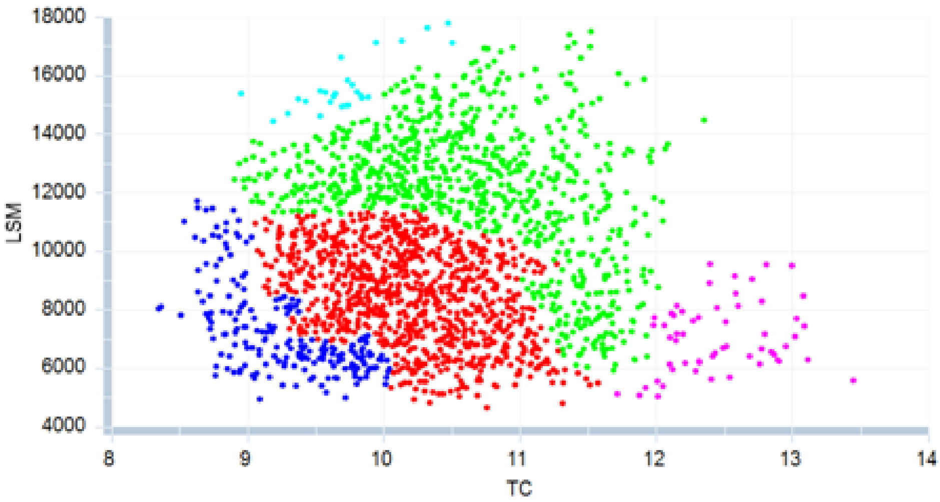
Goal

Minimize *TC*, minimize *LSM*, and maximize *AC*.

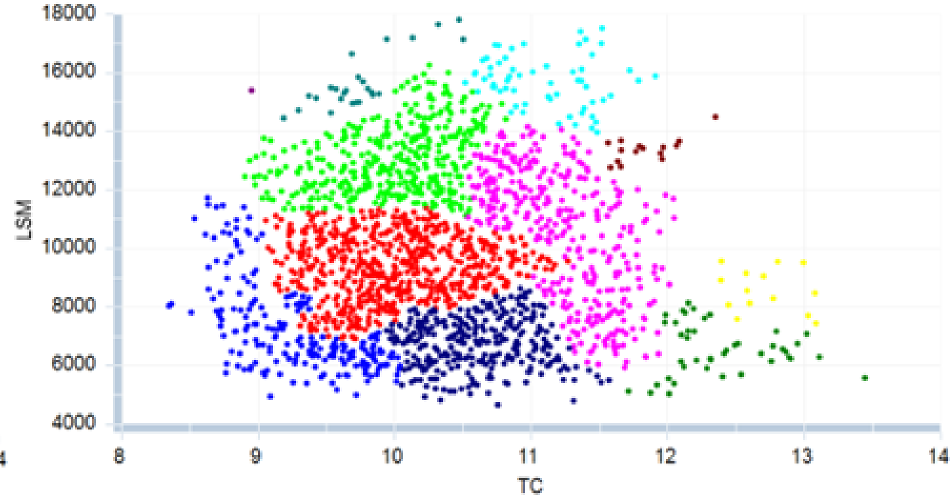
Constraints:

$L/B \geq 6$;
$L/D \leq 15$;
$L/T \leq 19$;
$F_n \leq 0.32$;
$25{,}000 \leq DWT \leq 50{,}000$;
$Const\_1 = T - 0.45DWT^{0.31} \leq 0$;
$Const\_2 = T - (0.7D + 0.7) \leq 0$;
$Const\_3 = 0.07B - GM_T \leq 0$;

Multi-Objective Optimization (MOO)



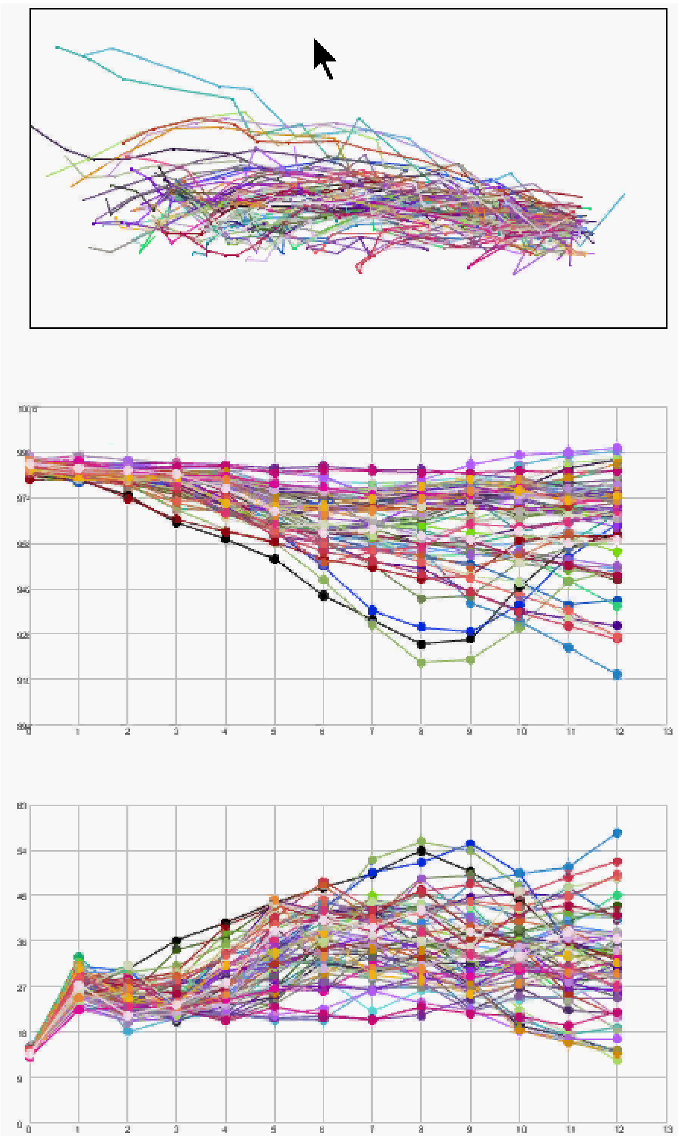
18000
16000
14000
12000
10000
8000
6000
4000
LSM
8
9
10
11
12
13
14
TC

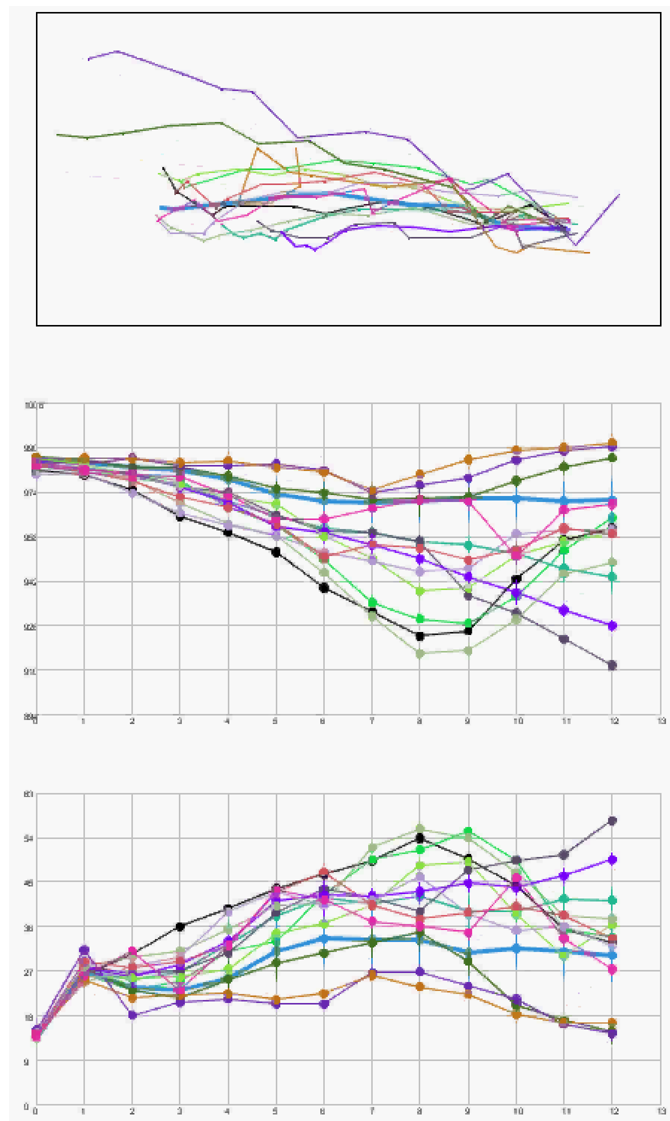


18000
16000
14000
12000
10000
8000
6000
4000
LSM
8
9
10
11
12
13
14
TC

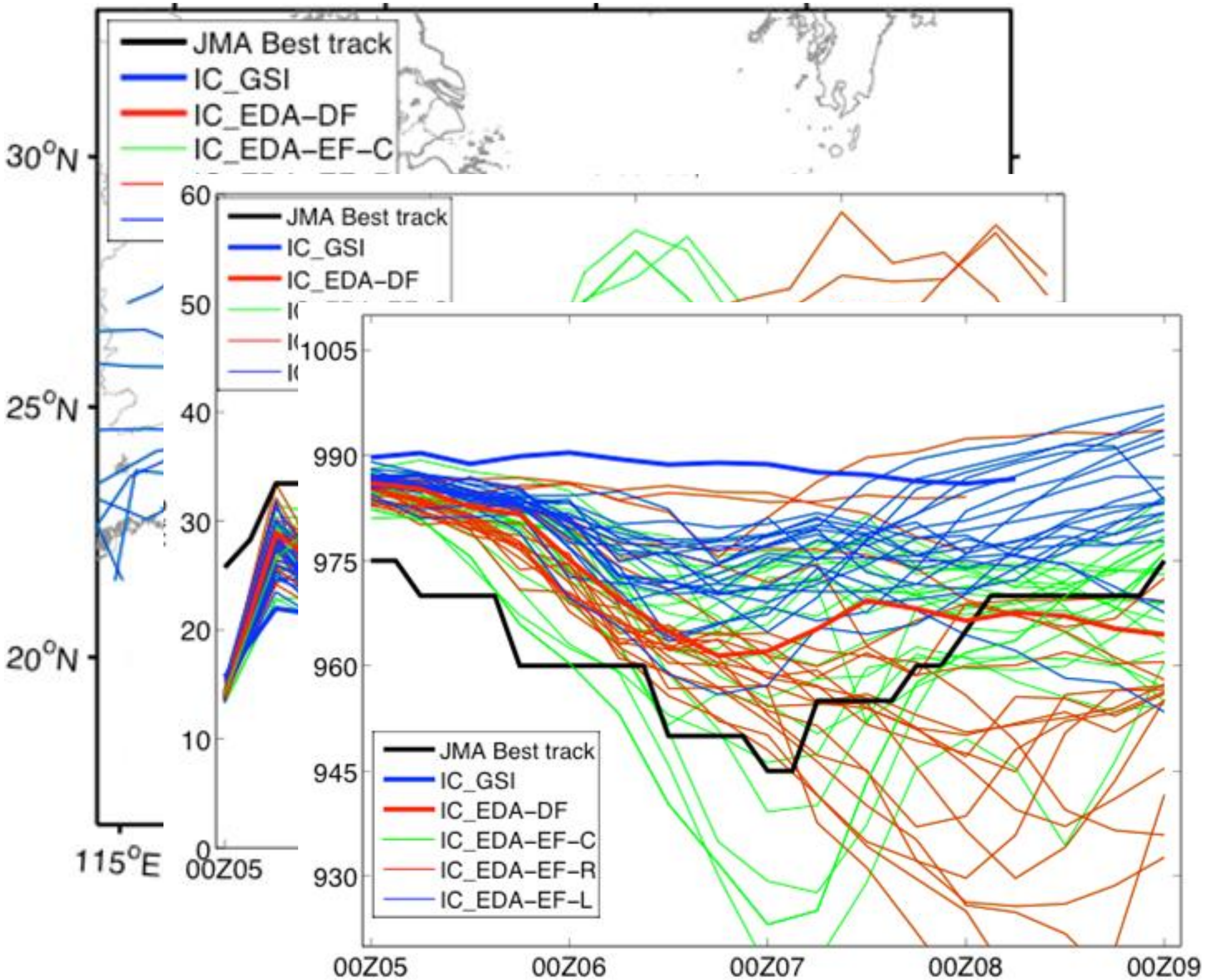
LSM
TC
D > 17.437906
T <= 11.52756
T <= 10.914416
B <= 28.314922
D > 18.416422
B > 25.840878
D > 13.718345
L <= 190.893982
B > 28.01489
D > 17.4715
T <= 11.20666
L <= 236.831
L <= 228.529
L <= 171.309
D > 19.8305

Example 2: Ensemble-based Analysis and Forecast





Scenario: Typhoon Morakot



Images from F. Zhang

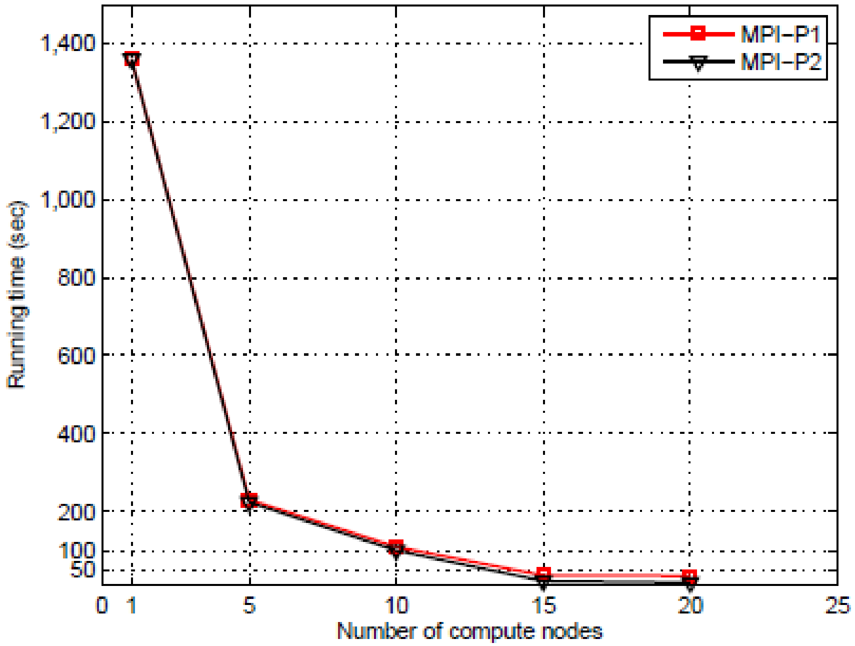
Can we support interactive analysis of these models (e.g., are tracks similar, how the tracks evolve)?

We use HMAC to cluster these models and provide interactive visualizaiton tools.

Some Challenges

- Enhance the cyberinfrastructure
  - Parallel computing to support interactive visual analytics
  - Collaborative analysis
    - Distributed users
- Increase the model transparency of clustering algorithms
  - Validation
- Support the evaluation of the analysis results
  - Verification

Parallelization of HMAC



Ship design data: 2,000 * 17

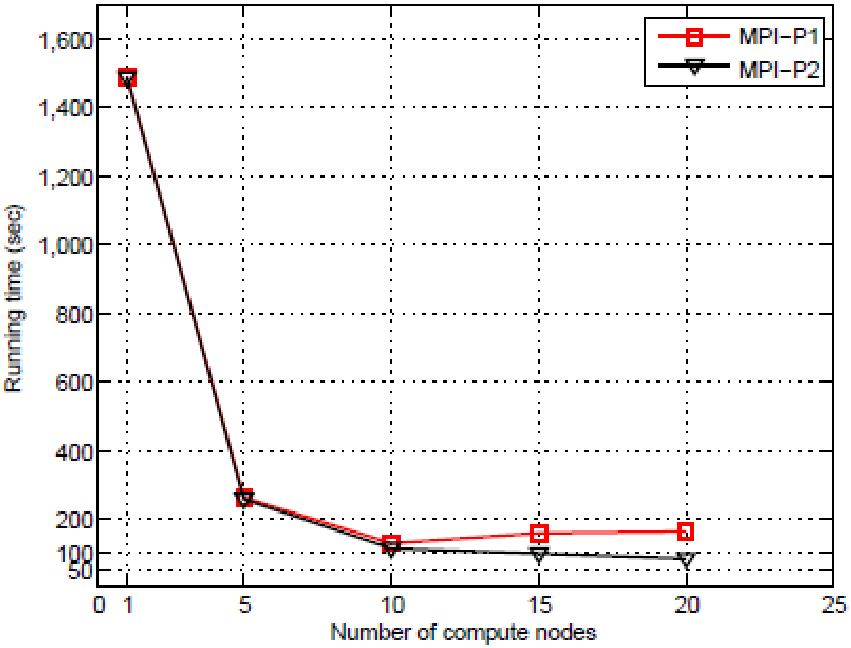
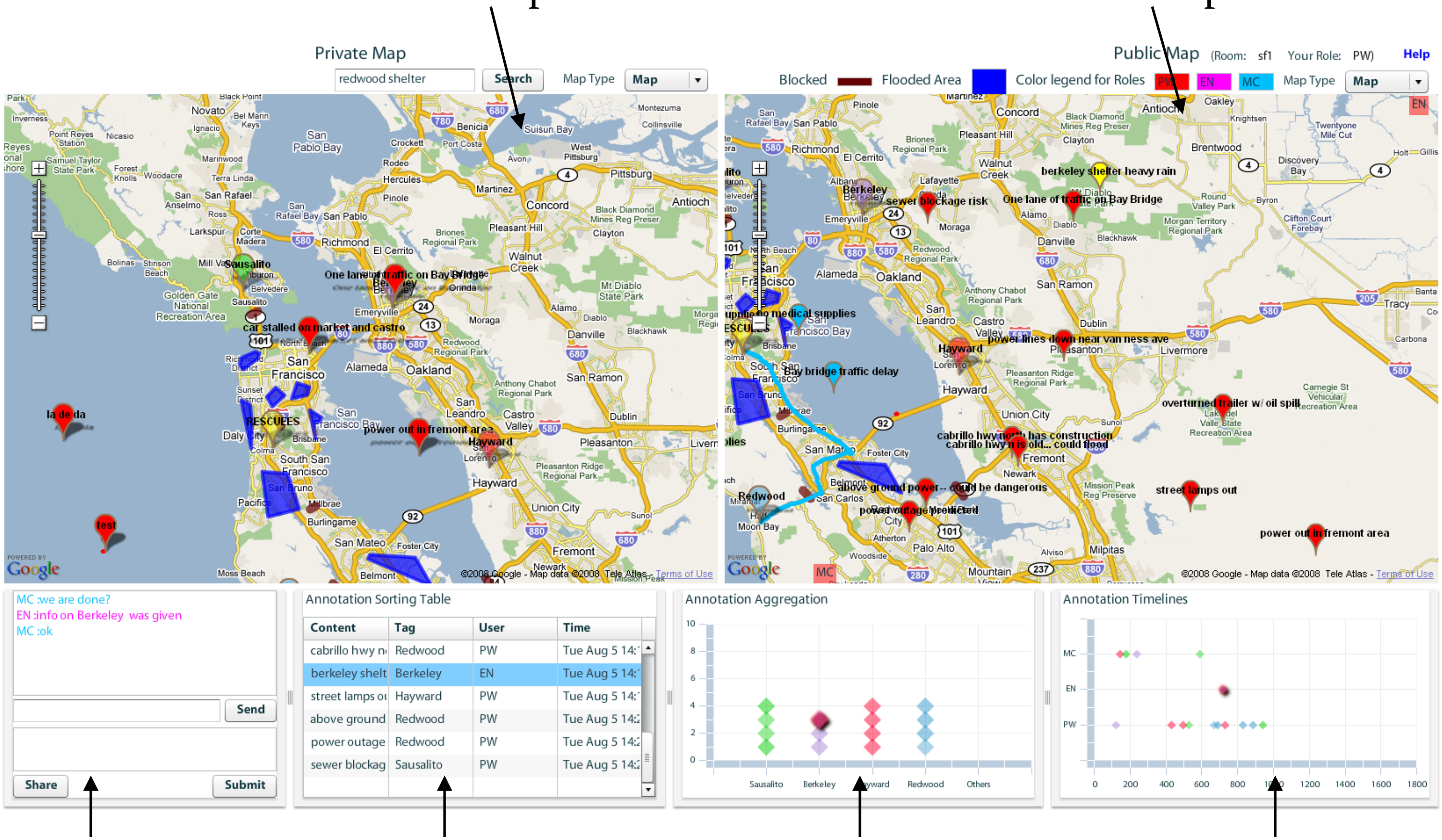


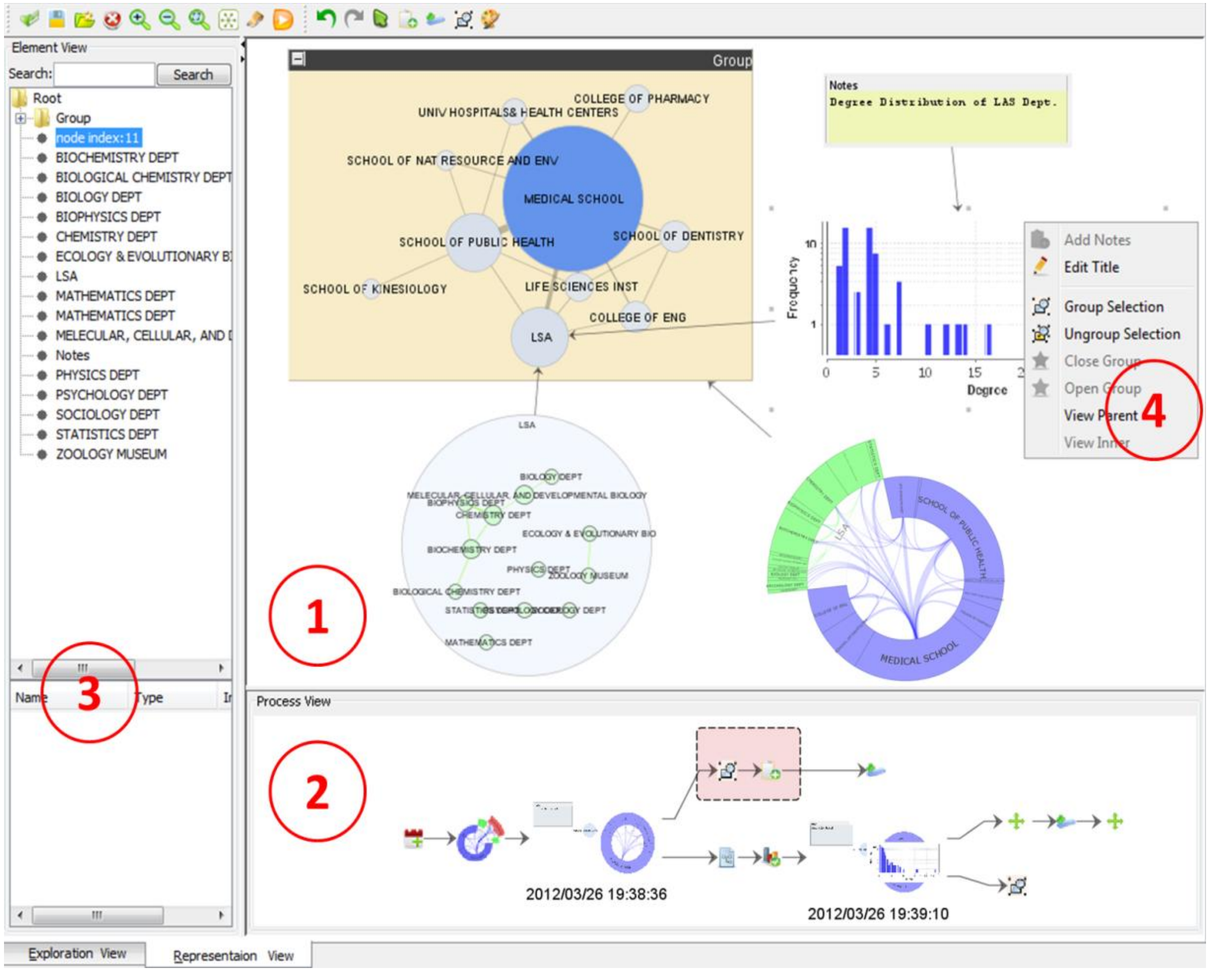
Image Data : 1,400 * 64

Collaborative Decision Making



(Wu, Convertino, Ganoe, Carroll & Zhang, 2012)

Sensemaking Process Visualization



(Gou & Zhang, 2012)

In Summary

- Analyzing big data needs both computer and human brain.
  - Advanced algorithms to reveal hidden data patterns.
    - E.g., clustering and classification methods
  - Human brain to interpret the meaning of data and patterns with domain knowledge.
    - Iterative sensemaking process
- Our efforts focus on building cyberinfrastructure to leverage the powers of both.
  - Developing algorithms and visual analytics systems.
    - Consider data, algorithms, and tasks.
    - Support domain-specific data analysis (multi-disciplinary efforts)
- Potential impacts
  - Scientific research, problem-solving, education, etc.