Visual Analytics for Emerging Data Intensive Problems

Kwan-Liu Ma
University of California at Davis

ICRI 2023, Academia Sinica

Visual Analytics

is the science of analytics reasoning

guided by

statistical analysis, machine learning

and interactive visualization

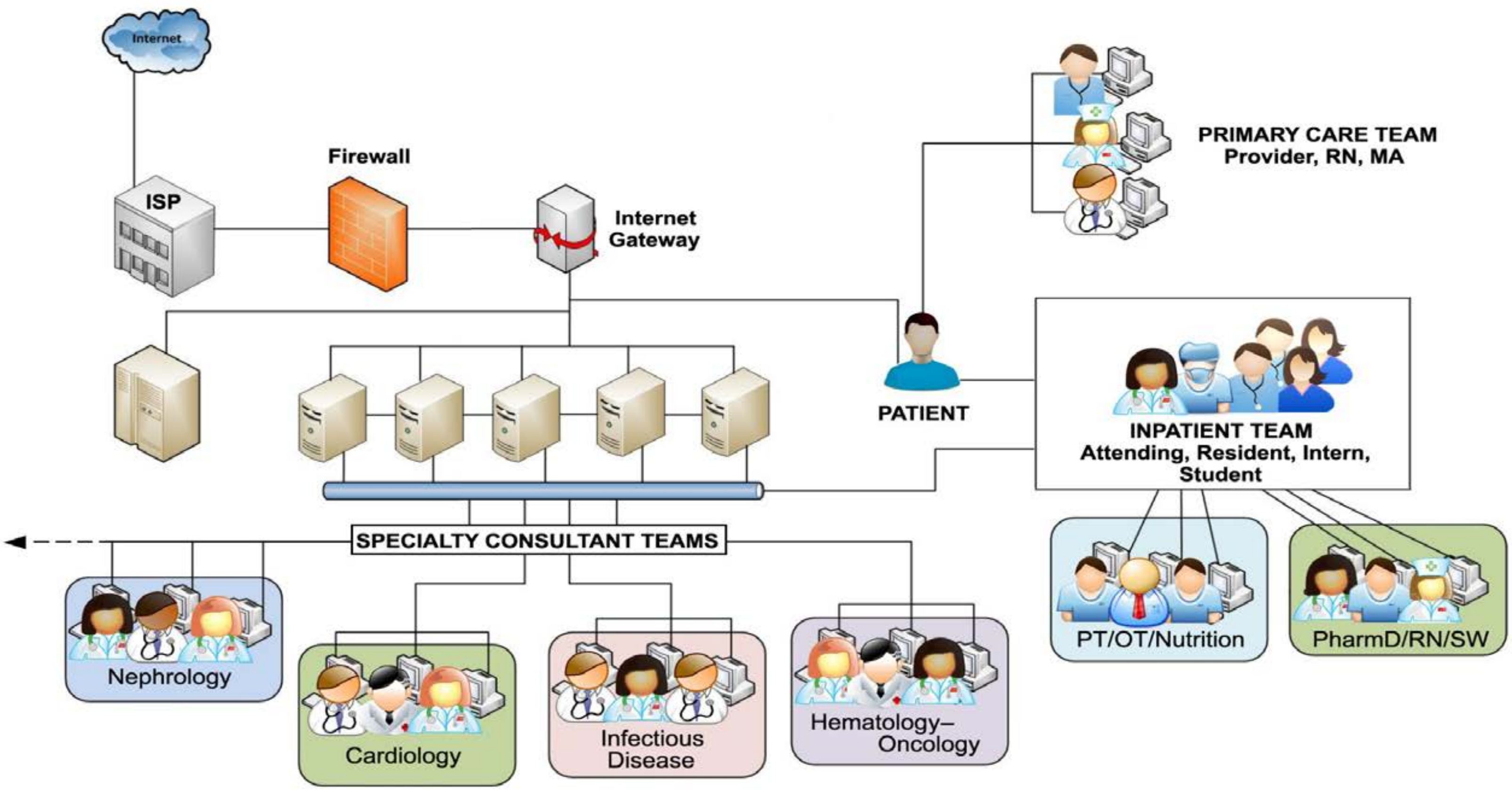
Outline

Visual analytics to enhance:

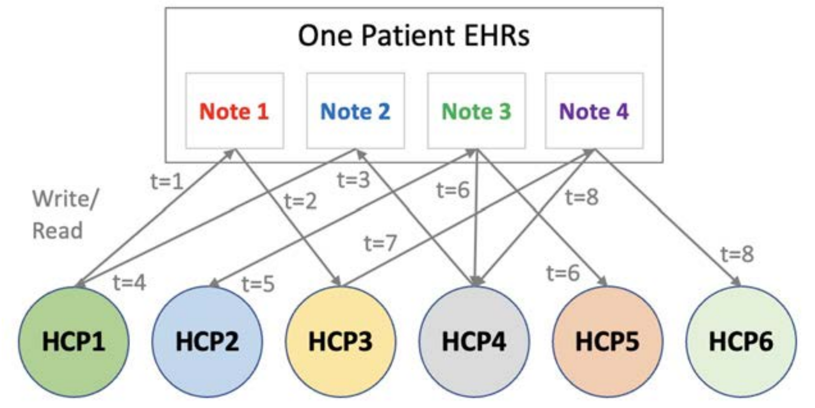
- Teamwork for healthcare
- Operation & performance of extreme-scale systems
- Comprehension of large collection of documents

EHR Data Analytics
to Enhance Patient Care Multiteam Systems

Multiteam Systems for Healthcare

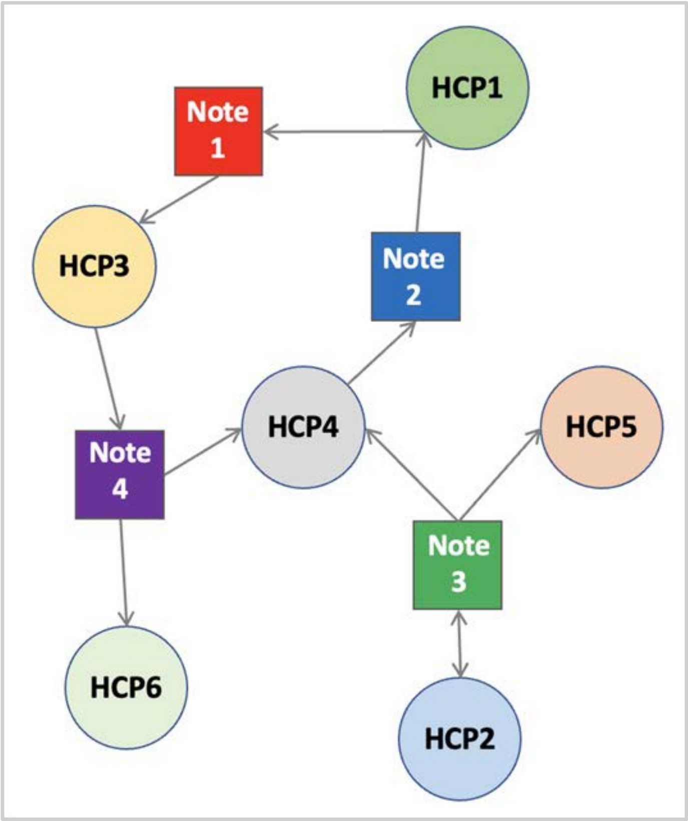


From Access Log to Networks



Stored information:

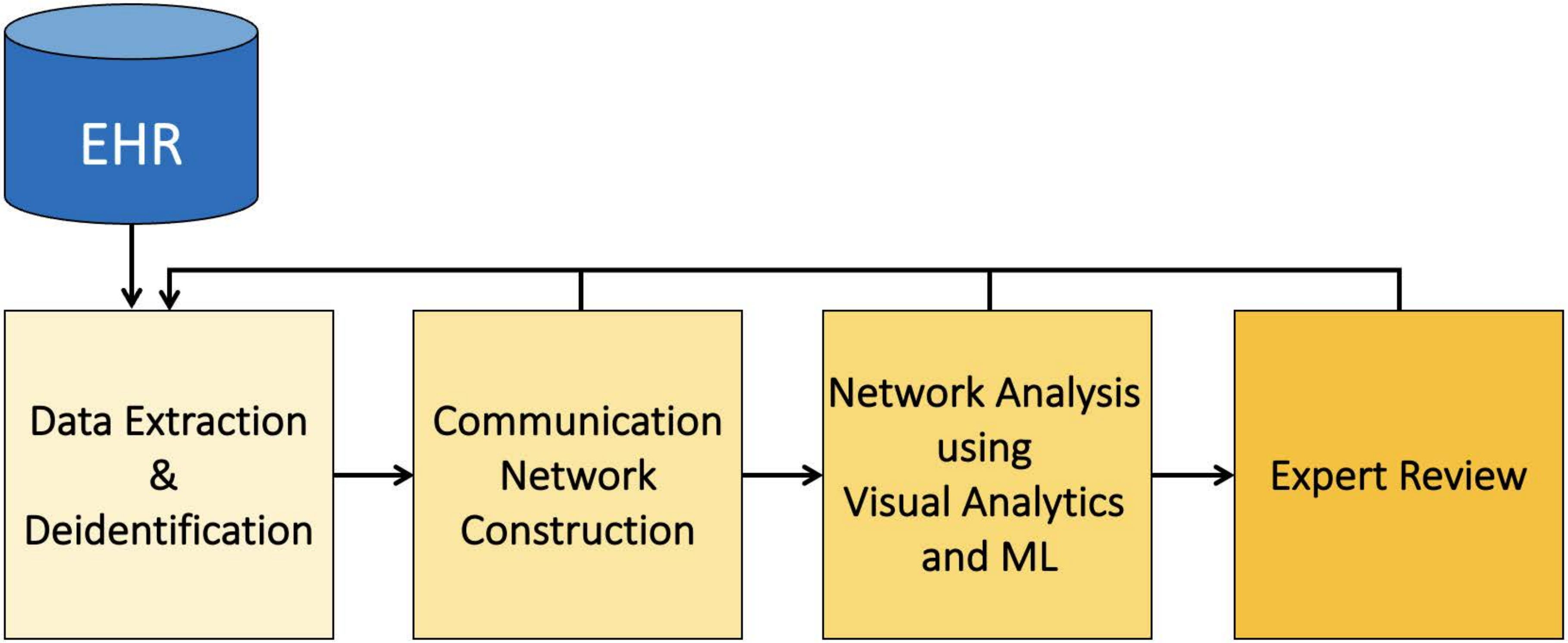
1. Healthcare Professionals (HCPs)
2. Notes
3. Access timestamps
4. Direction of interactions



ML Assisted Visual Analytics

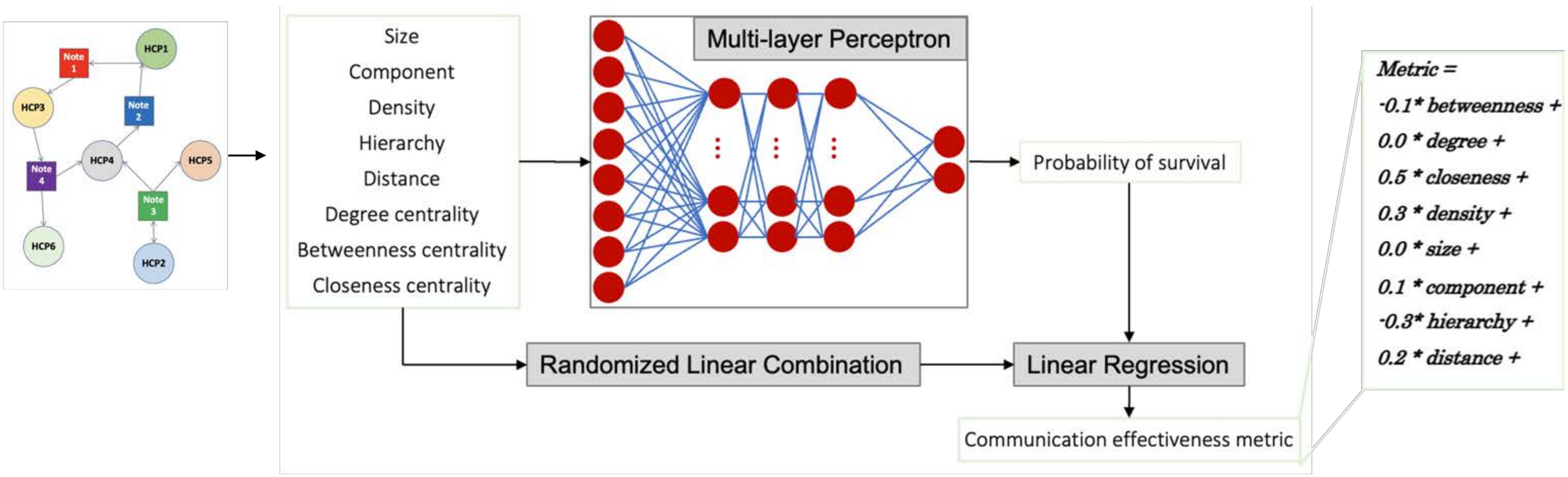
- **Informed by:**
  - Communication network theories: structure, position, and flow
  - Systems engineering principles: workflow, value, waste (time, cognition)
- **Designed to assist HCPs with visualization guided information-processing to:**
  - quickly browse an overview of their patients' medical care
  - identify MTS members involved at specific points and periods in time
  - drill down to discover important details (i.e., notes, messages, reports, etc.)
  - see who has accessed/reviewed specific EHR documentations
  - efficiently rectify gaps in information-sharing by members of the MTS

Workflow


EHR
Data Extraction
&
Deidentification
Communication
Network
Construction
Network Analysis
using
Visual Analytics
and ML
Expert Review

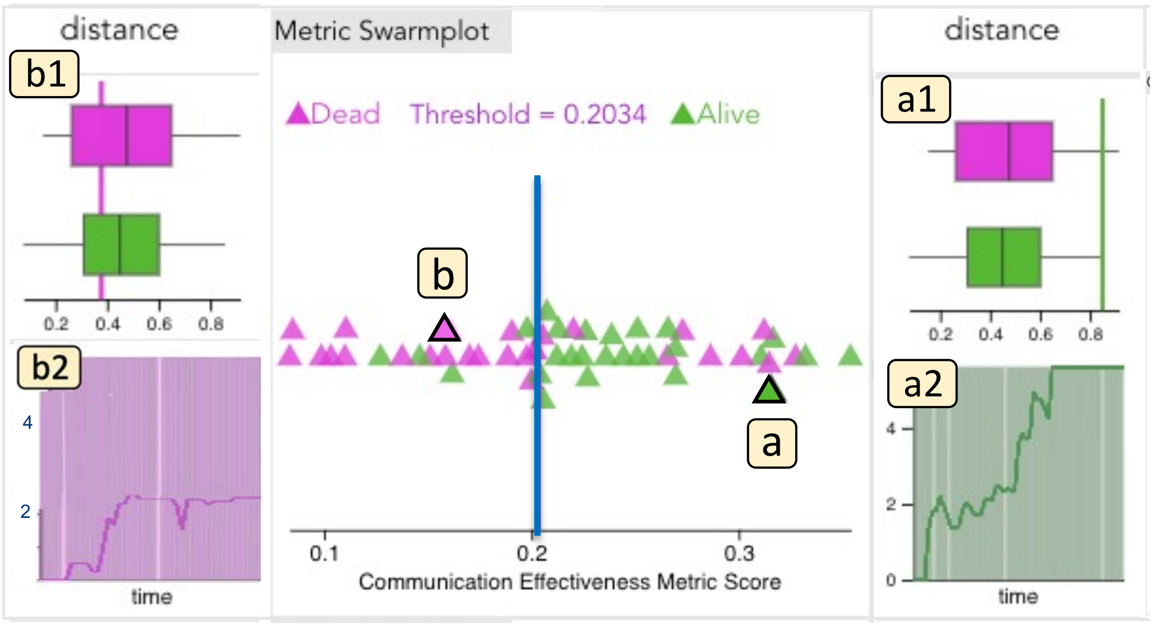
Case Study I

- Patient Group: 53 patients, between age 65-75 with Stage 3 lung cancer, in which 27 survived and 26 did not.
- Communication effectiveness metrics are weighted combinations of multiple network measures (e.g., degree, distance, closeness, betweenness, etc.)

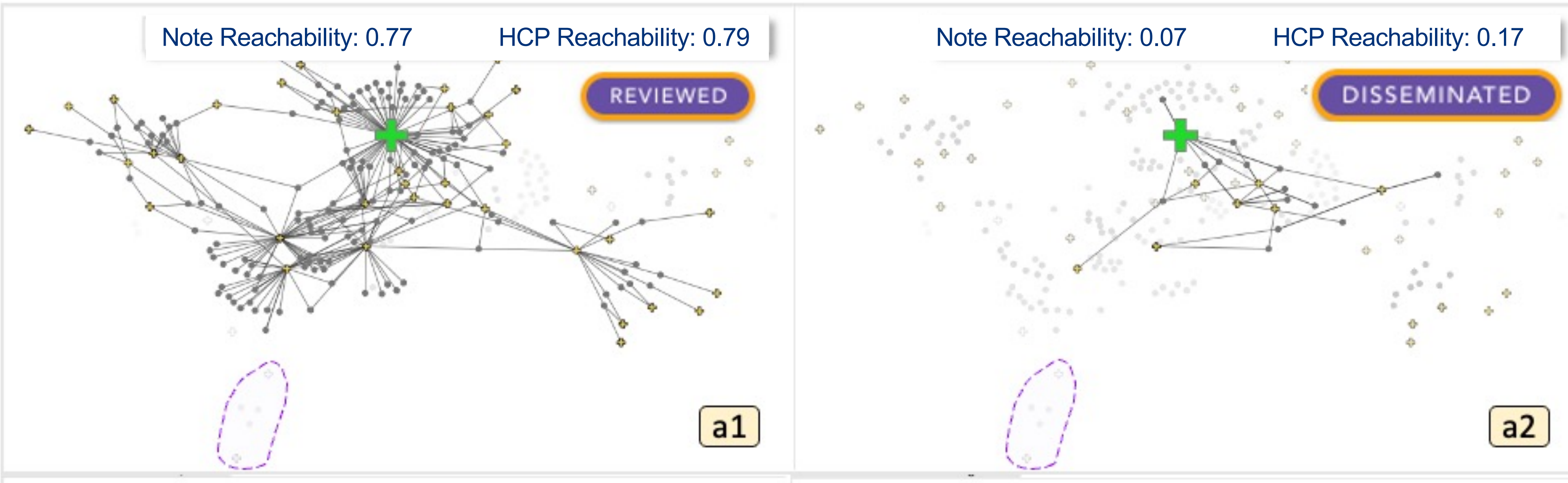


Case Study I

- Patient Group: 53 patients, between age 65-75 with Stage 3 lung cancer, in which 27 survived and 26 did not.
- Communication effectiveness metrics are weighted combinations of multiple network measures (e.g., degree, distance, closeness, betweenness, etc.)
- The distance measure is the average number of directed edges between every *HCP* node and *note* node in a network.
- A network with a higher distance value suggests the *HCPs* in the network are more connected.

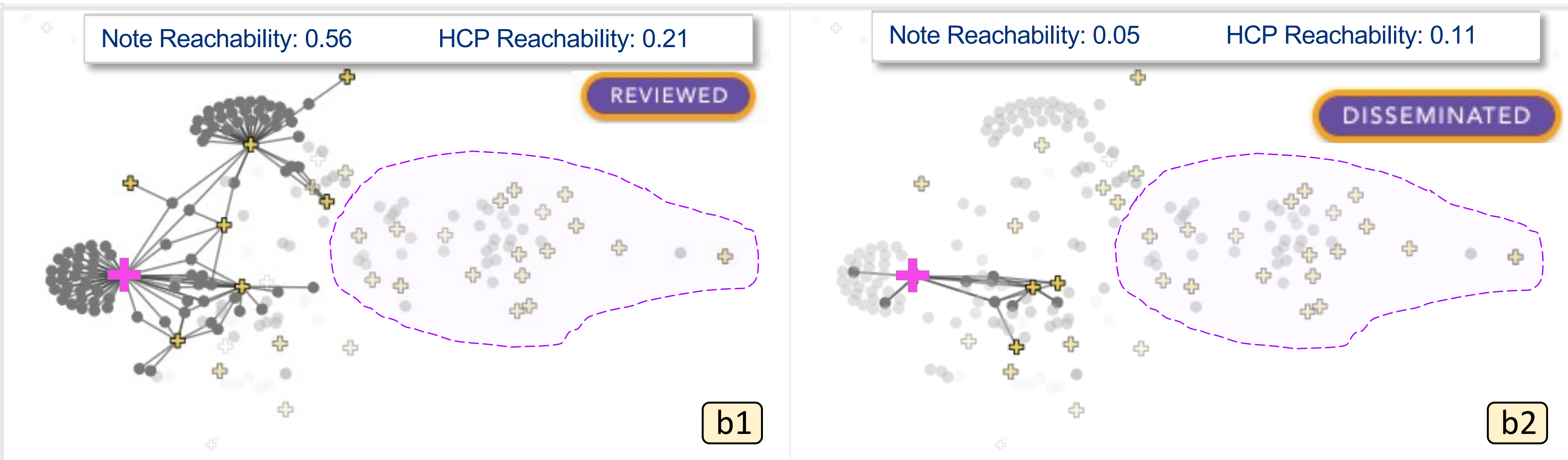


Information Reachability



Information reachability visualization for a patient who survived

Information Reachability

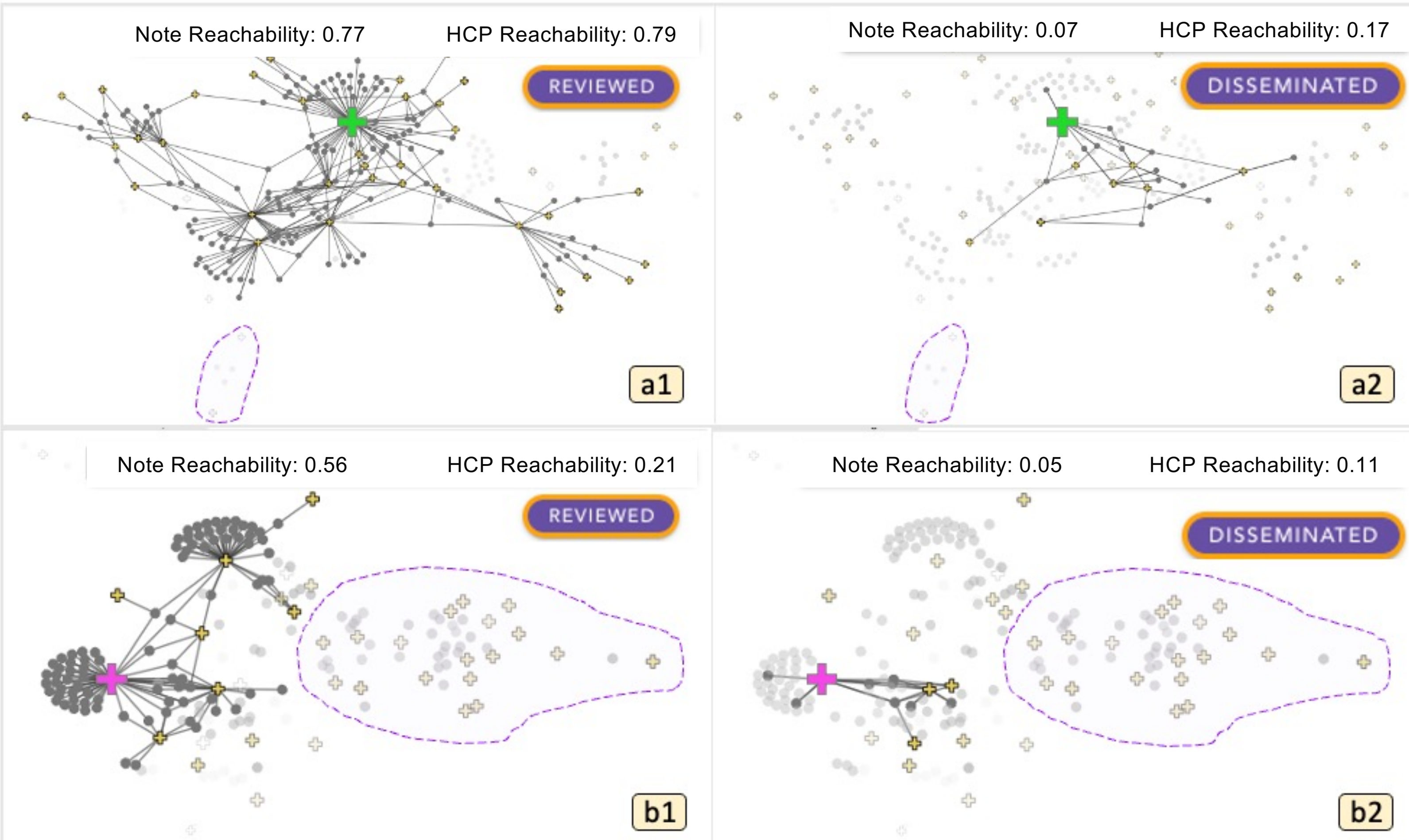


Information reachability visualization for a patient who did not survive

HCP

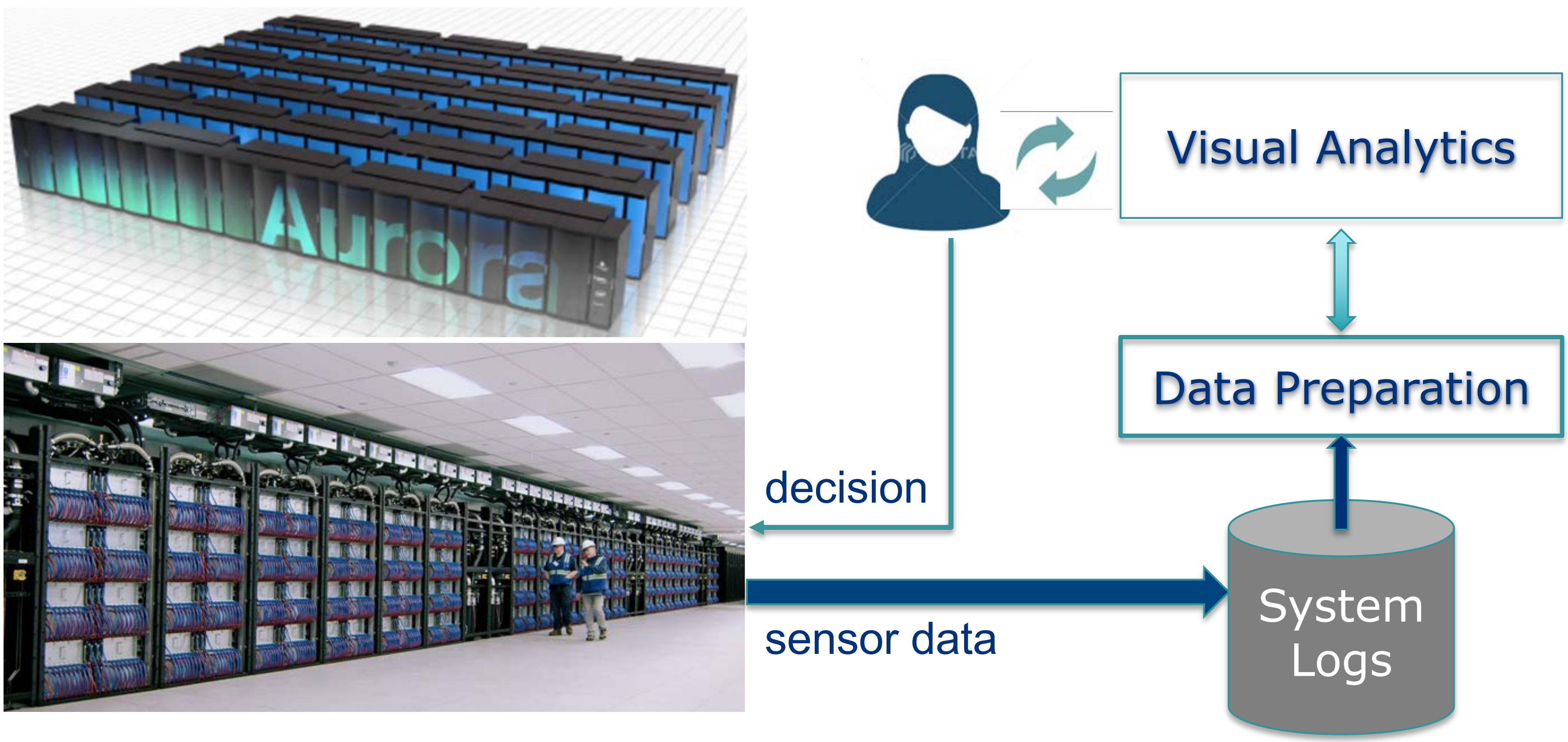
Note

Information Reachability



Streaming Data Analytics for Hardware System Monitoring

Monitoring Supercomputing Systems



Supercomputer Hardware Logs

- Streaming, multivariate, time series
- A representative dataset
  - Extracted from the K Computer, Riken, Japan
  - 864 Racks
  - 390 temperature and 480 voltage readings per rack
  - 1162 measures per rack
  - 288 time points per measure (every 5min)
  - Data volume: ~2GB per day

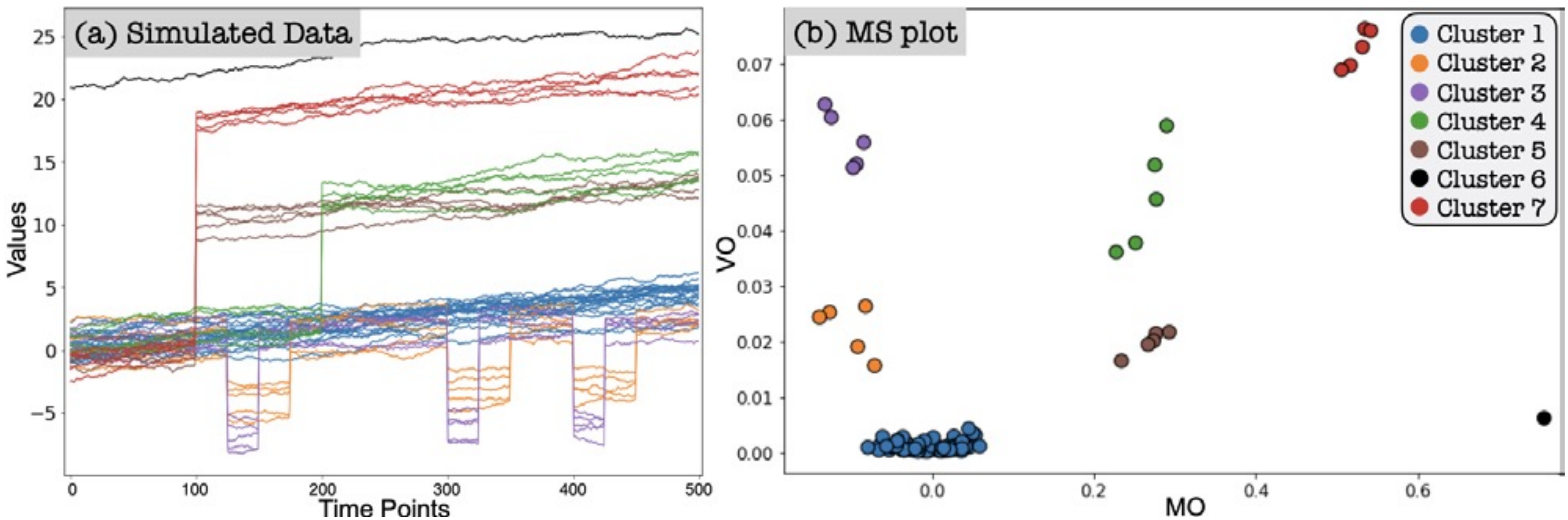
Technical Approach

- Expect to start with select readings and initial time points
- We adopt the magnitude-shape plot (MS plot by Dai and Genton 2018), which reveals both the functional magnitude and shape outlyingness of time-series data.
- **Incrementally** compute the MS plot with every new time point added to the continuum
- **Progressively** update the MS plot with every new time series added
- Employ FPCA together with MS plots to conduct further analysis of select outliers

Magnitude Shape Plot

- W. Dai and M. Genton, “Multivariate Functional Data Visualization and Outlier Dection”, J. Comput. Graph. Stat. 27(4) 2018
- Designed to visualize both the magnitude outlyingness (MO) and shape outlyingness (VO) of multivariate functional data based on measures of directional outlyingness (O)
- The plot depicts how much a time series has a different magnitude and shape with other time series, and thus visually reveals outliers.

Visual Outlier Detection with the Magnitude Shape Plot



- Simulated functional data and the MS plot with magnitude outlyingness (MO) and shape outlyingness (VO) along x-axis and y-axis, respectively.

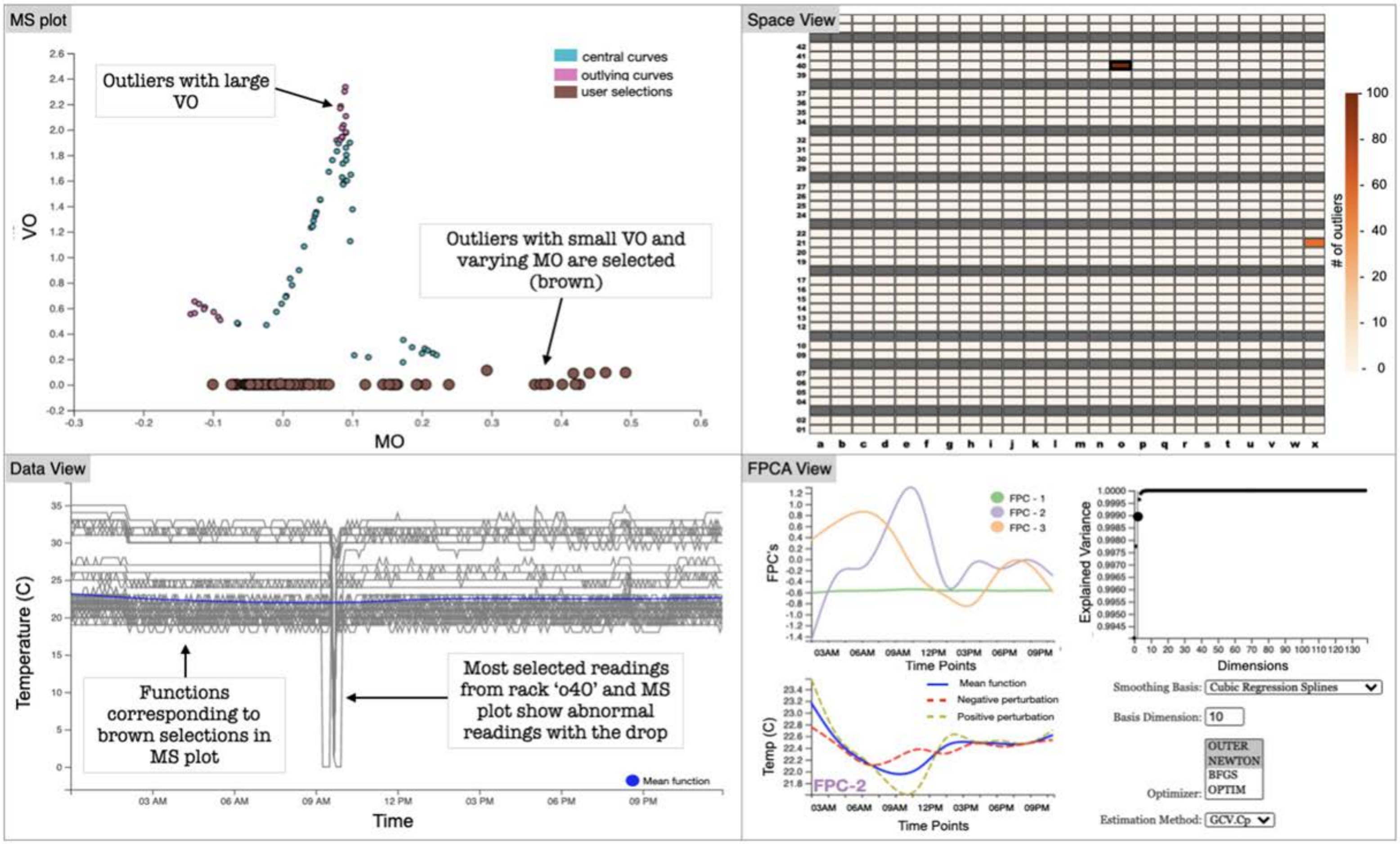
Incremental Update of the MS Plot

- Recomputing the MS Plot when new time points for each series arrive is costly, infeasible for real-time monitoring tasks.
- Incremental update of exact MO and VO!
- $\mathrm{MO}^{T+1}(\mathrm{X}^{T+1}) = \frac{1}{T+1}(T\mathrm{MO}^{T}(\mathrm{X}^{T}) + \mathrm{O}(\mathrm{X}^{T+1}[\mathrm{T}+1]))$
- The required memory space to save the previous results, MO and FO, for all N time series is O(N).

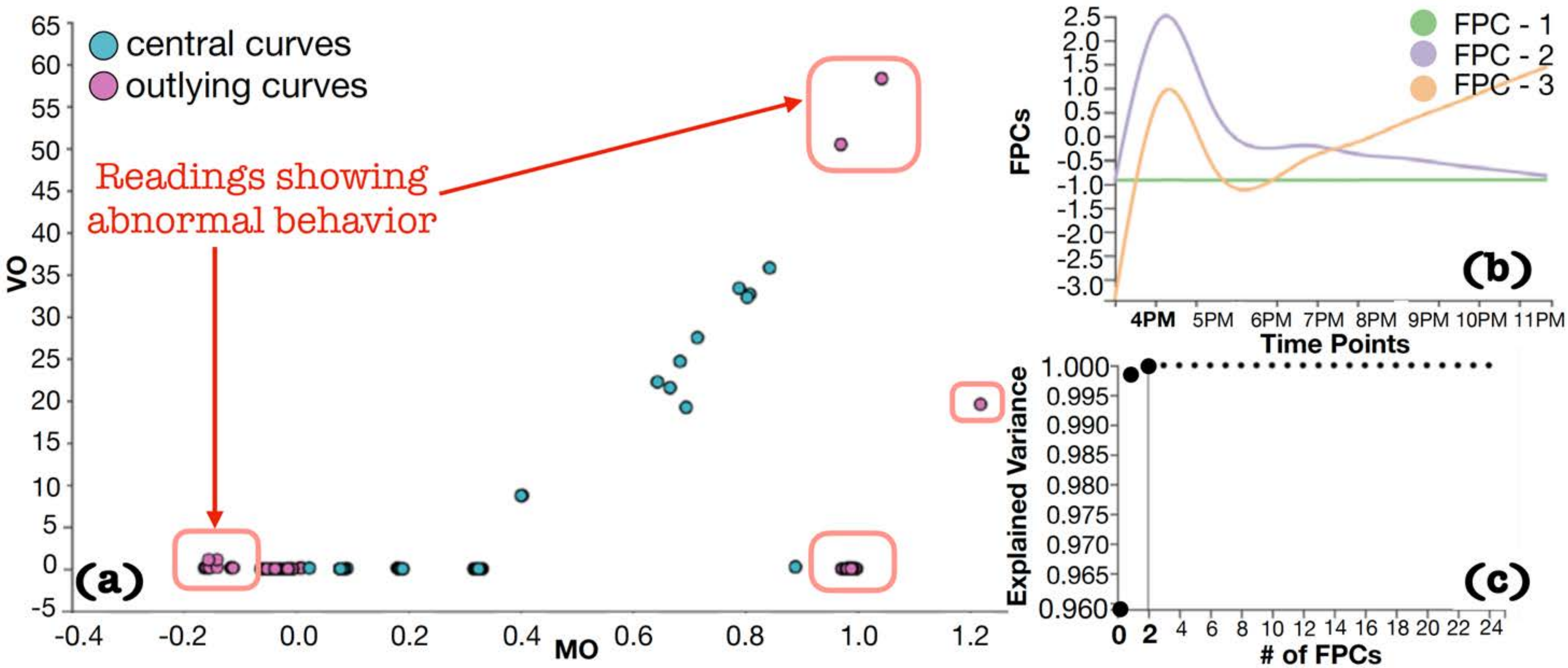
Progressive Update of the MS Plot

- The exact update for a new time series requires recomputation of the measures for all N time series at all time points!
- When the number of time series, N, in a system is large, this recomputation is computationally prohibit.
- To avoid recomputation, a progressive algorithm is designed to generate the MS plot with approximated O.
- Exact update is made only when the KL (Kullback-Leibler) divergence of the mean absolute deviation between the new and original sets of time series becomes larger than a set threshold value.

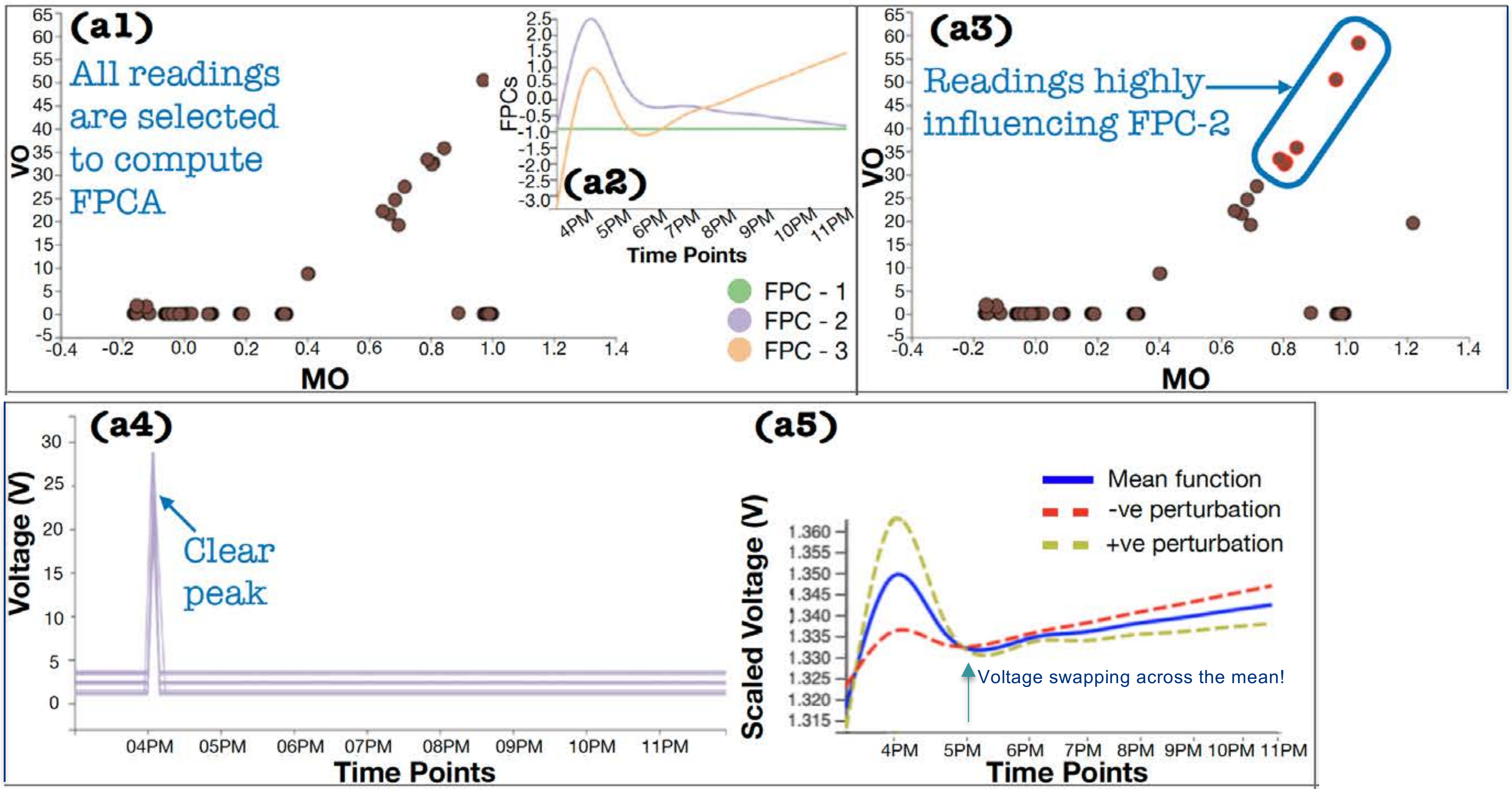
Visual Analytics Interface



Outlier Identification and Validation



Outlier Identification and Validation



Visual Analytics for Comprehension of Large Collections of Documents

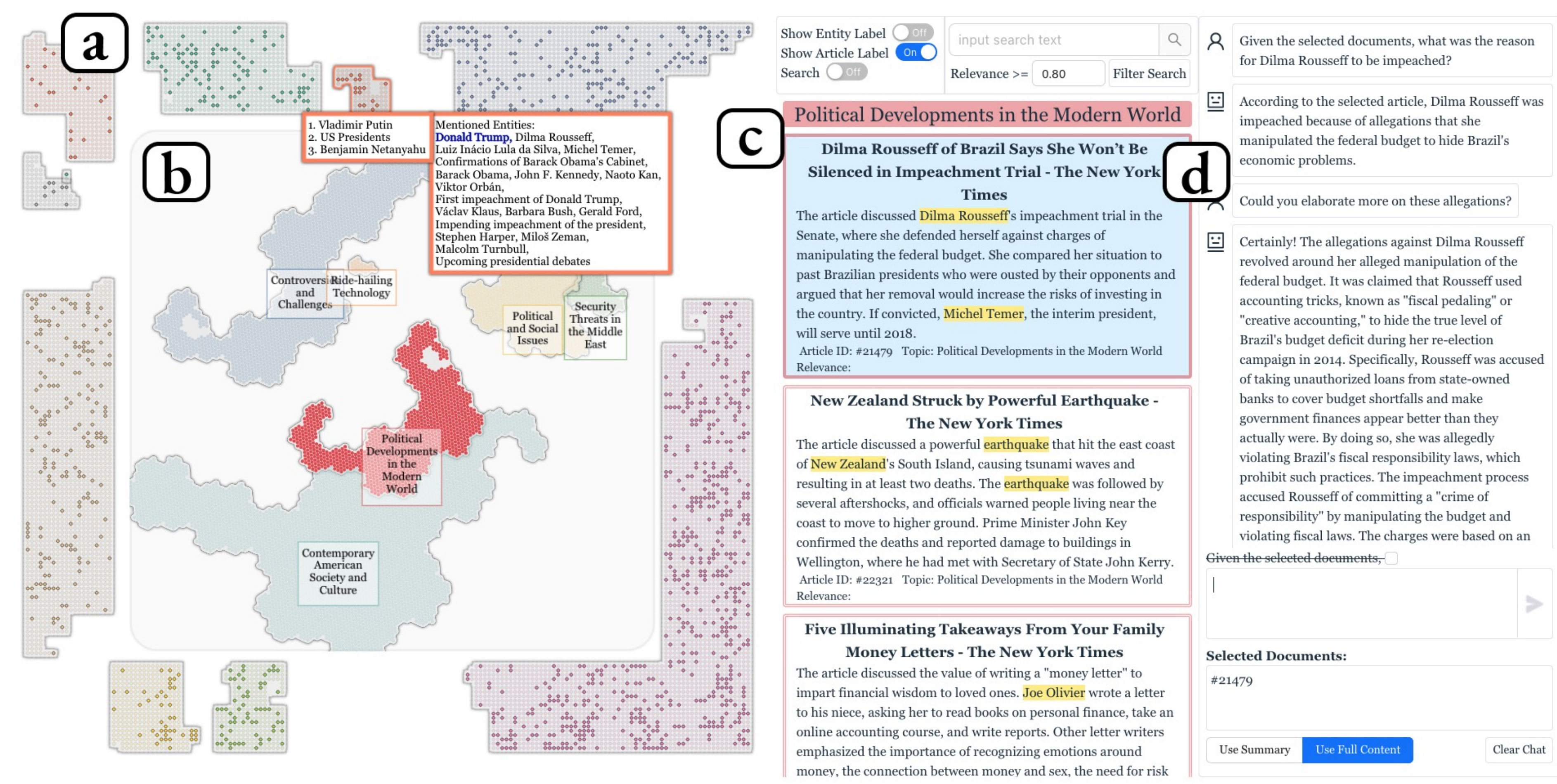
LLM in Document Analysis

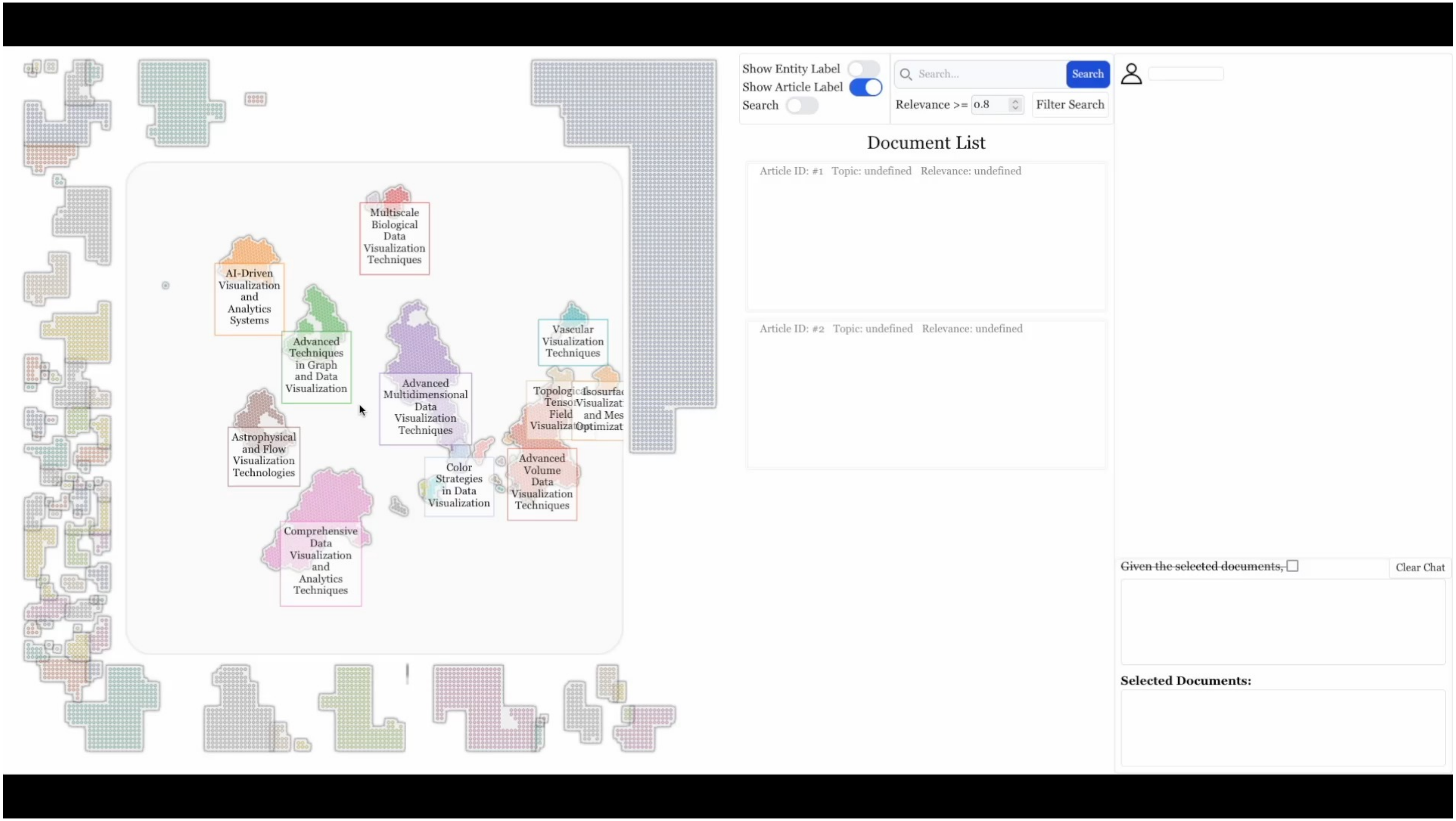
- LLM as Agents
  - A chatbot that directly answers questions
  - Easy to develop and highly generalizable
  - Lacking transparency & mechanisms to convey uncertainty/confidence
  - No robust, established ways of evaluating the quality of responses
- LLM as NLP (Natural Language Processing) Task Solvers
  - NLP Tasks: Topic analysis, sentiment analysis, relation extraction, etc.
  - Transparent processes, interpretable results, and customizable tasks
  - Prompt engineering is non-trivial and using LLMs is still a costly approach

A Combined Approach

- Sensemaking of large collections of unstructured text
- Design considerations:
  - Overview of article topics, keywords, and their connections
  - Support of drill-down, progressive disclosure
  - Transparency, interpretability, and direct manipulation
  - Detailed analysis of a specific target of interest
- A combined approach:
  - Processing data with LLMs followed by interactive visualization to find and keep track of *targets of interest*
  - Sensemaking on the targets of interest with LLM Chatbots

Example I: HyperMap




Show Entity Label
Show Article Label
Search
Search...
Search
Relevance >= 0.8
Filter Search
Document List
Article ID: #1 Topic: undefined Relevance: undefined
Article ID: #2 Topic: undefined Relevance: undefined
Multiscale Biological Data Visualization Techniques
AI-Driven Visualization and Analytics Systems
Advanced Techniques in Graph and Data Visualization
Advanced Multidimensional Data Visualization Techniques
Vascular Visualization Techniques
Astrophysical and Flow Visualization Technologies
Color Strategies in Data Visualization
Advanced Volume Data Visualization Techniques
Comprehensive Data Visualization and Analytics Techniques
~~Given the selected documents,~~
Clear Chat
Selected Documents:

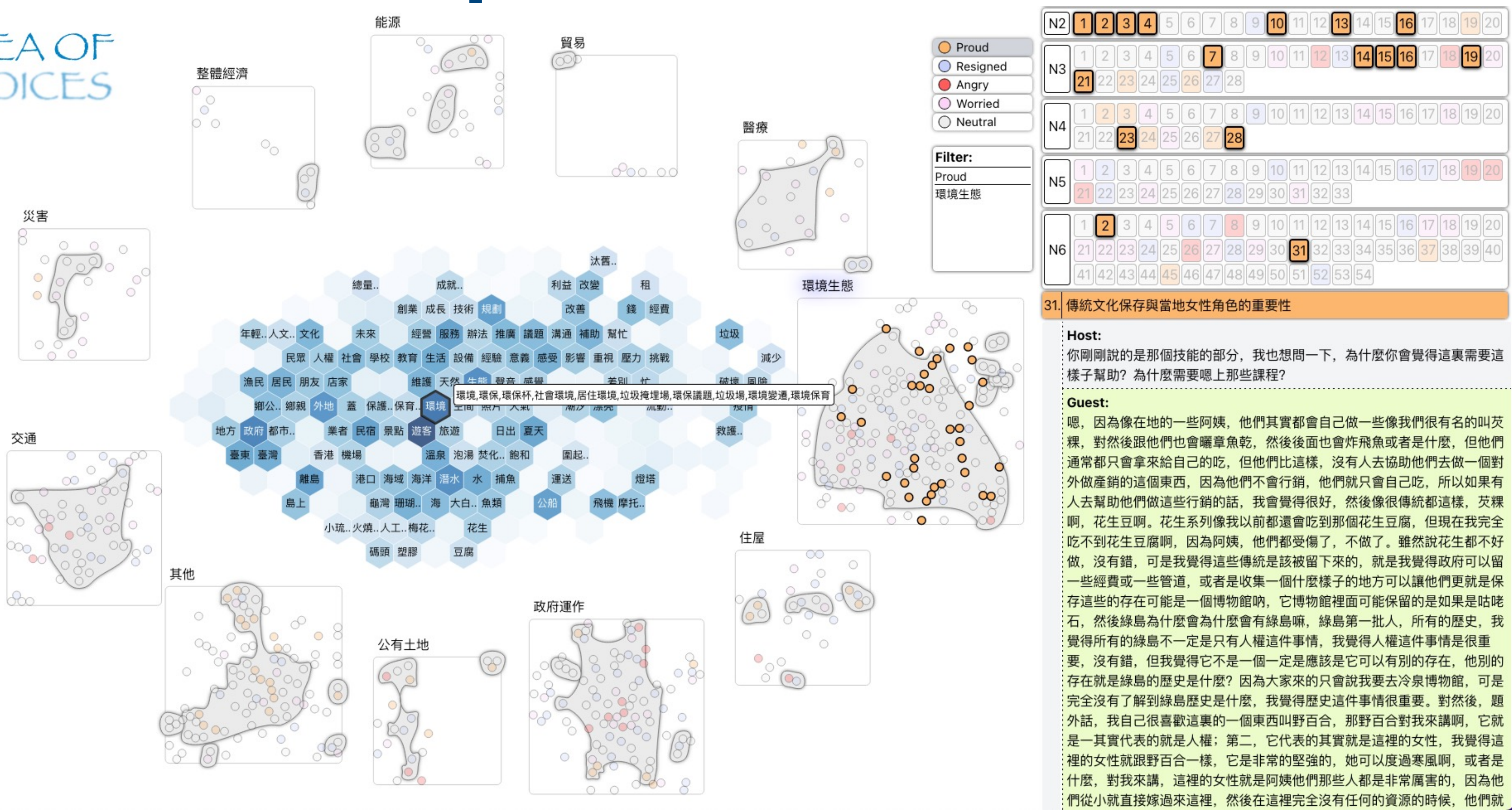
Example I: HyperMap

- User Study on literature review using HyperMap
  - Chatbot group vs HyperMap group
- Findings
  - Chatbot is good at explaining unknown keywords and summarizing articles or the course of conversation.
  - Users have less trust over the Chatbot for critical questions (e.g., if a statement/claim is true/justified)
  - The performance of the Chatbot group is less consistent (e.g., the user may get contradictory answers or be unsure what to ask next.)
  - Visualizations can help direct the process and track the analysis targets.

Example II: Sea of Voices

Example II: Sea of Voices

SEA OF VOICES



Visual Analytics for High Dimensional Data

- Hyeon Jeon, Yun-Hsin Kuo, Michaël Aupetit, Kwan-Liu Ma, Jinwook Seo: Classes are not Clusters: Improving Label-based Evaluation of Dimensionality Reduction. IEEE **Transactions on Visualization and Computer Graphics** 30 (1) (2024) (presented at **IEEE VIS** 2023)
- Takanori Fujiwara, Yun-Hsin Kuo, Anders Ynnerman, Kwan-Liu Ma: Feature Learning for Nonlinear Dimensionality Reduction toward Maximal Extraction of Hidden Patterns. **IEEE PacificVis** 2023: 122-131
- Shilpika, Takanori Fujiwara, Naohisa Sakamoto, Jorji Nonaka, Kwan-Liu Ma: A Visual Analytics Approach for Hardware System Monitoring with Streaming Functional Data Analysis. IEEE **Transactions on Visualization and Computer Graphics** 28(6): 2338-2349 (2022)
- Takanori Fujiwara, Xinhai Wei, Jian Zhao, Kwan-Liu Ma: Interactive Dimensionality Reduction for Comparative Analysis. IEEE **Transactions on Visualization and Computer Graphics** 28(1): 758-768 (2022) (presented at **IEEE VIS** 2021)
- Takanori Fujiwara, Shilpika, Naohisa Sakamoto, Jorji Nonaka, Keiji Yamamoto, Kwan-Liu Ma: A Visual Analytics Framework for Reviewing Multivariate Time-Series Data with Dimensionality Reduction. IEEE **Transactions on Visualization and Computer Graphics** 27(2): 1601-1611 (2021) (presented at **IEEE VIS** 2020)
- Takanori Fujiwara, Jia-Kai Chou, Shilpika, Panpan Xu, Liu Ren, Kwan-Liu Ma: An Incremental Dimensionality Reduction Method for Visualizing Streaming Multidimensional Data. **IEEE Transactions on Visualization and Computer Graphics** 26(1): 418-428 (2020)

Acknowledgments

- My students and collaborators
- National Institute of Health
- Department of Energy
- National Science Foundation
- University of California Climate Action Initiative
- Bosch Research
- Intel Corporation
- Institute of Statistical Science, Academia Sinica