Visual Causality Analysis Made Practical

Jun Wang and Klaus Mueller
Computer Science Department
Stony Brook University

Causality Analysis

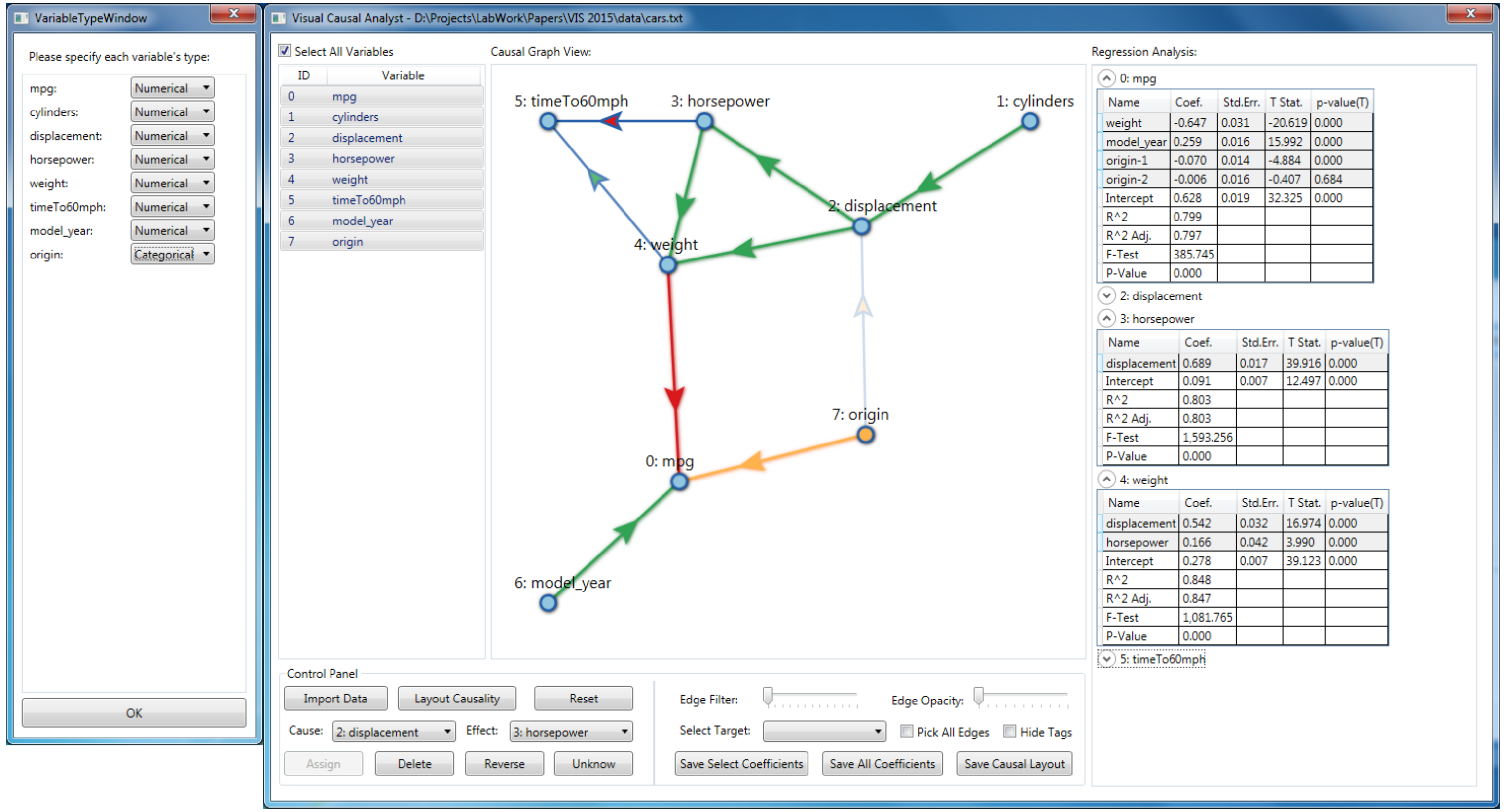
- Goal - Recover causal relations from observations
- Advantages
  - *More explicit than correlation analysis*
    - *"A causes B" vs. "A and B may be associated"*
  - *More practical than controlled experiments*
    - *The experiment for testing "smoking causes cancer"*

Visual Causality Analysis

- Why taking a Visual Analytics approach?
  - *Automated algorithms are not reliable*
  - *Get users involved with their domain knowledge*
  - *Make analysis more manageable*

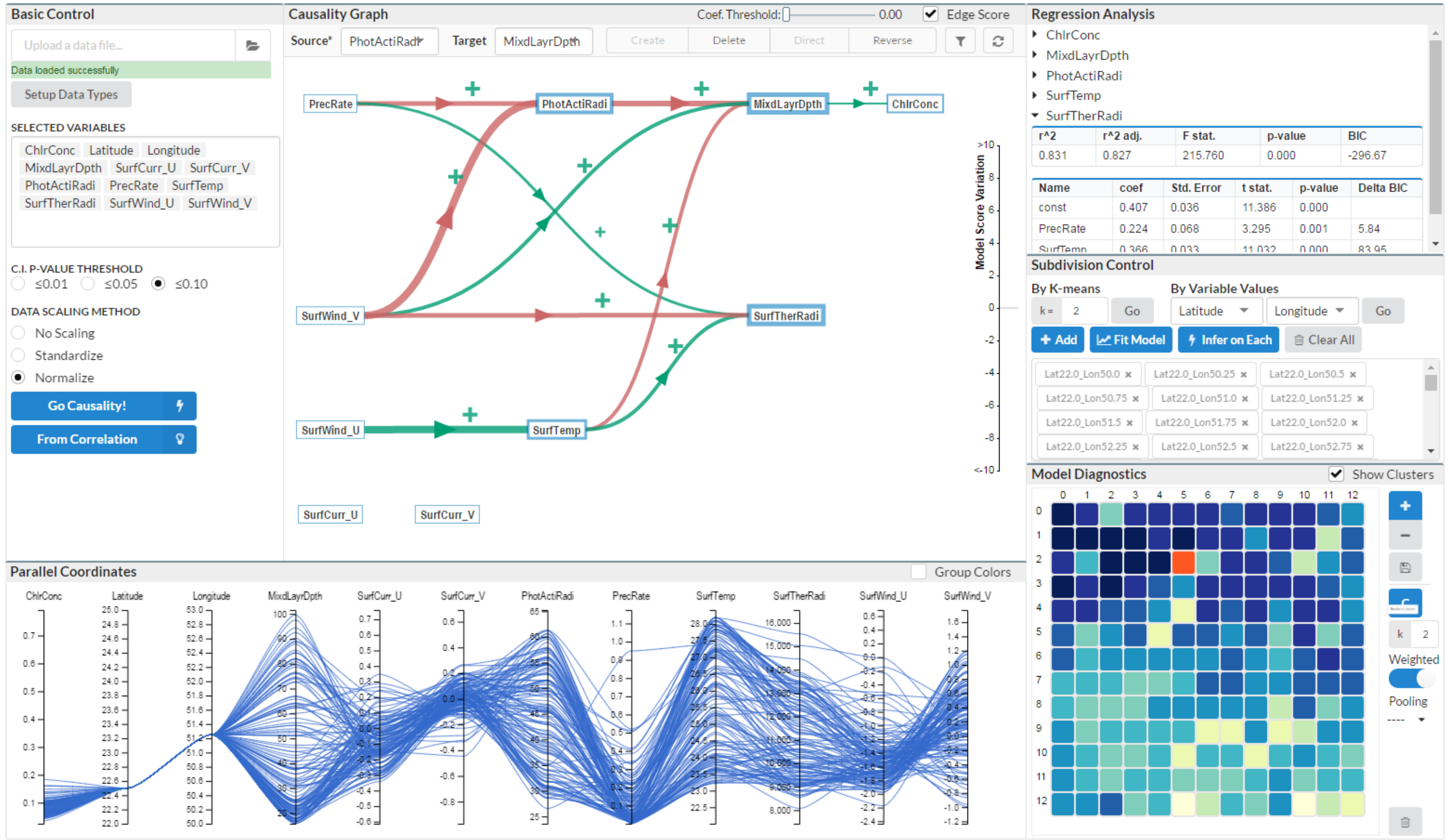
Previous Work

- Visual Causality Analyst
  - *Operating on a single model*
  - *Force-directed graph*
  - *Model refinement with statistical tables*
  - *Naïve method for processing heterogeneous data*

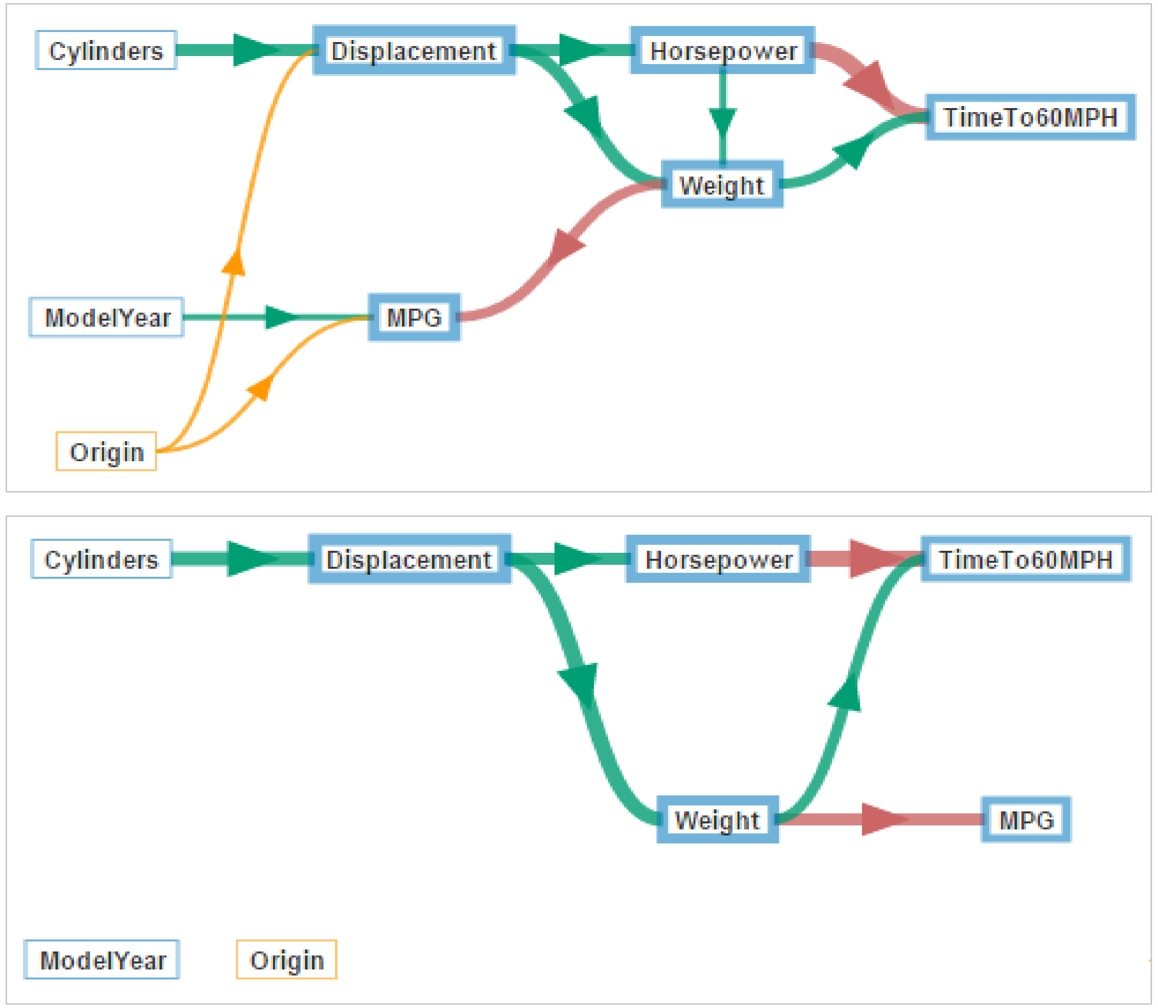


Current Work

- Causal Structure Investigator
  - *Visualizing causal flows*
  - *Visual model refinement*
  - *Interface for data subdivisions*
  - *Managing and pooling of the multiple models learned from data subdivisions*



Causal Flows



- Laid out by Breadth-first spanning tree
- Causal relations as paths flowing mostly from left to the right
- Color of path encodes relation type
- Width encodes strength of the relation

Causal Graph of the AutoMPG dataset

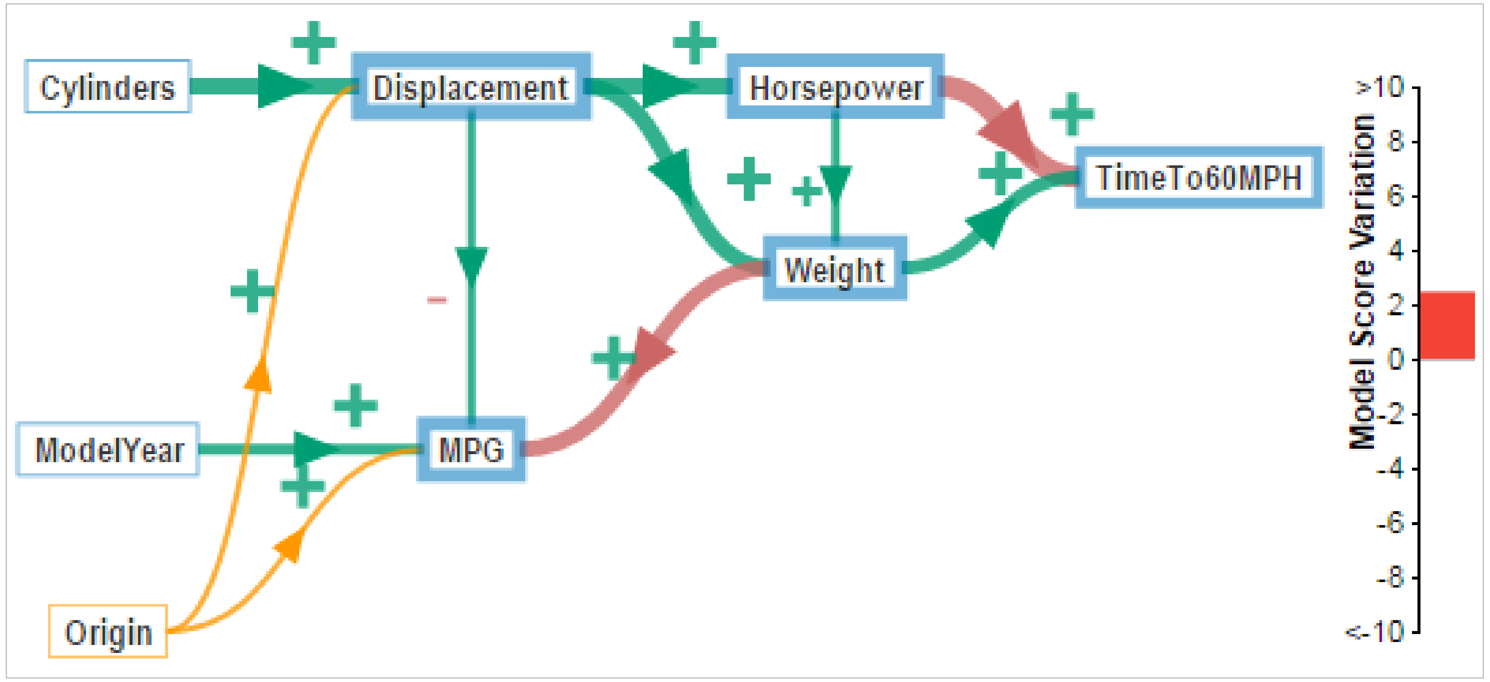
VIS 2017

Visual Model Refinement

- Measure model goodness with *Bayesian Information Criterion* (*BIC*)

$$\mathrm{BIC} = -2\ln\hat{L} + k\ln(n)$$

- An extra step in parameterization
- The heuristic – removing a good relation will lower the quality of the model



Causal Graph of the AutoMPG dataset

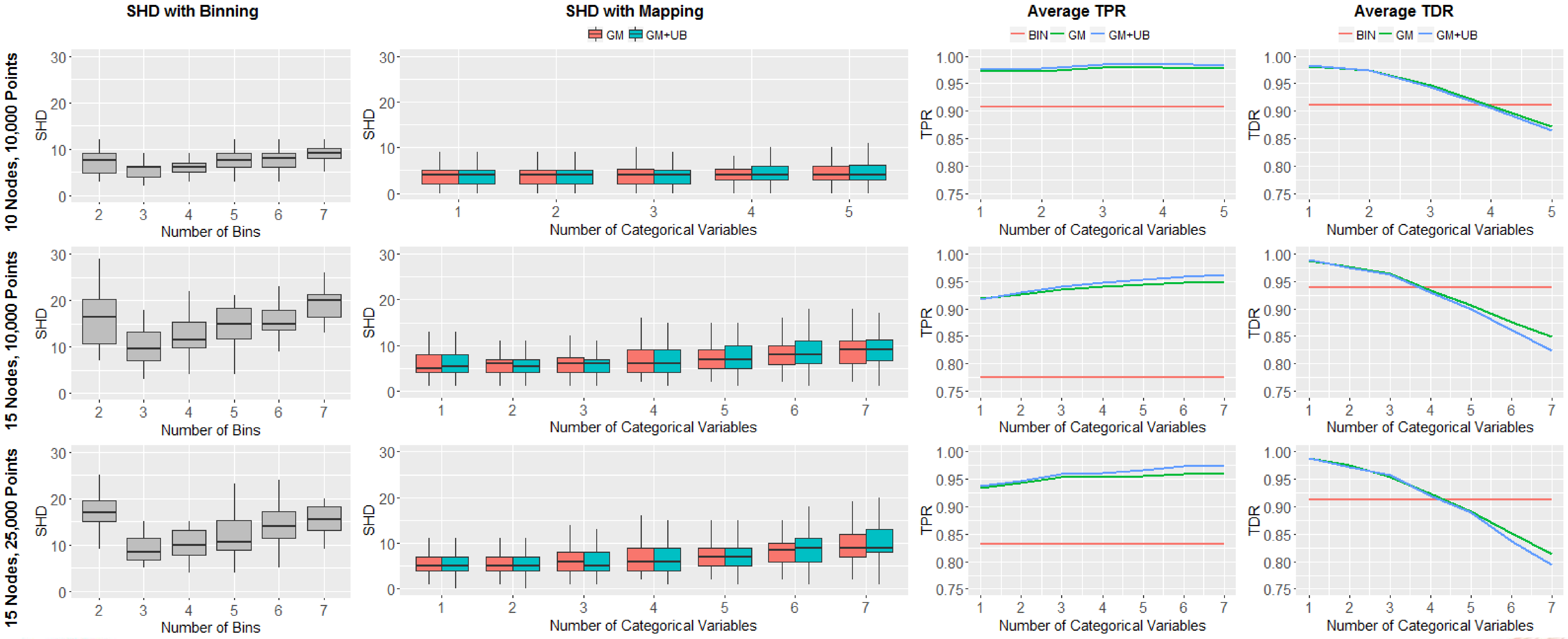
Handling Heterogeneous Data

- Global Mapping (GM) strategy (*previous method*)

$$v_c(j) \propto \sum_{i=1}^{D} \Theta_i \rho_i \mu\left(v_i(j)\right)$$

- GM + Un-binning (UB) strategy
  - *Random sample in the range to simulate continuous domain*
- Experiment evaluation comparing to *Binning*
  - *100 random DAGs and the according data*
  - *Measure the rebuilding error in Structure Hamming Distance (SHD), True Positive Rate(TPR),* and *True Discovery Rate (TDR)*

Handling Heterogeneous Data

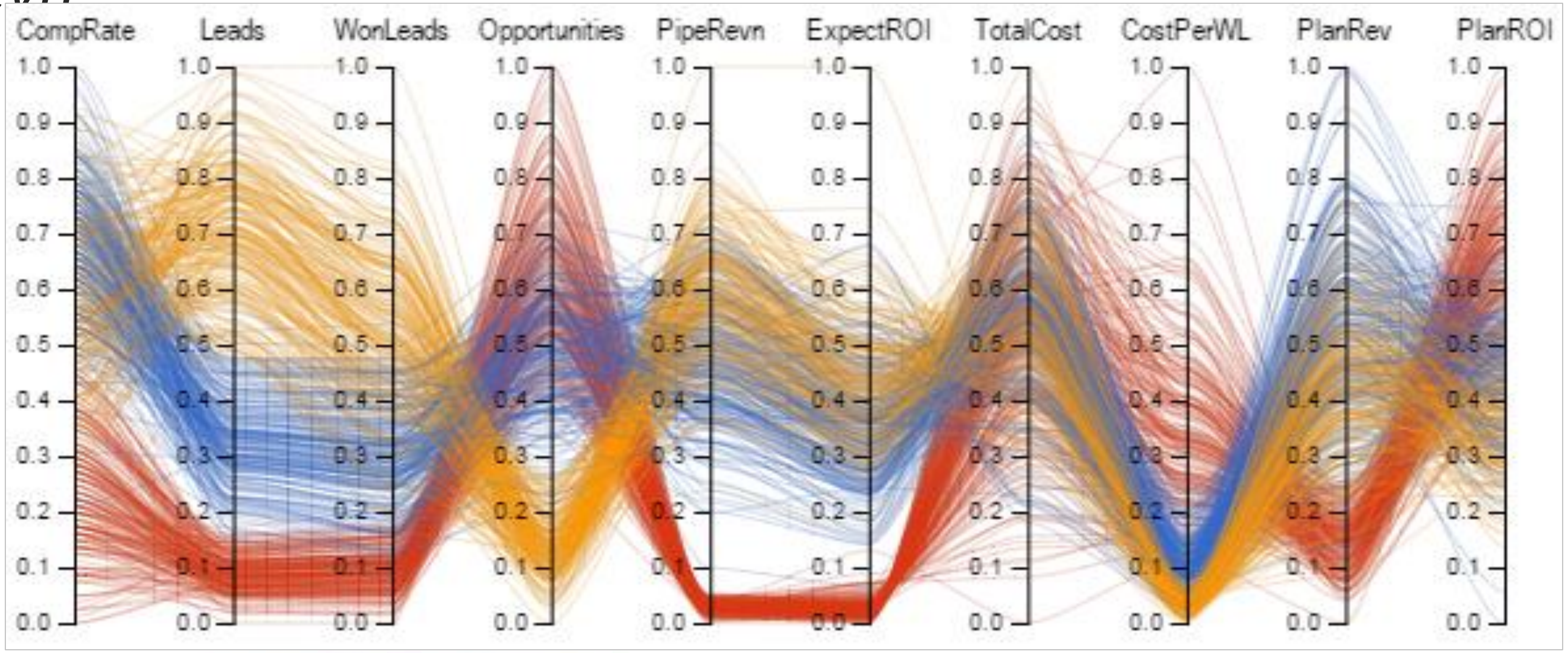


Data Subdivisions

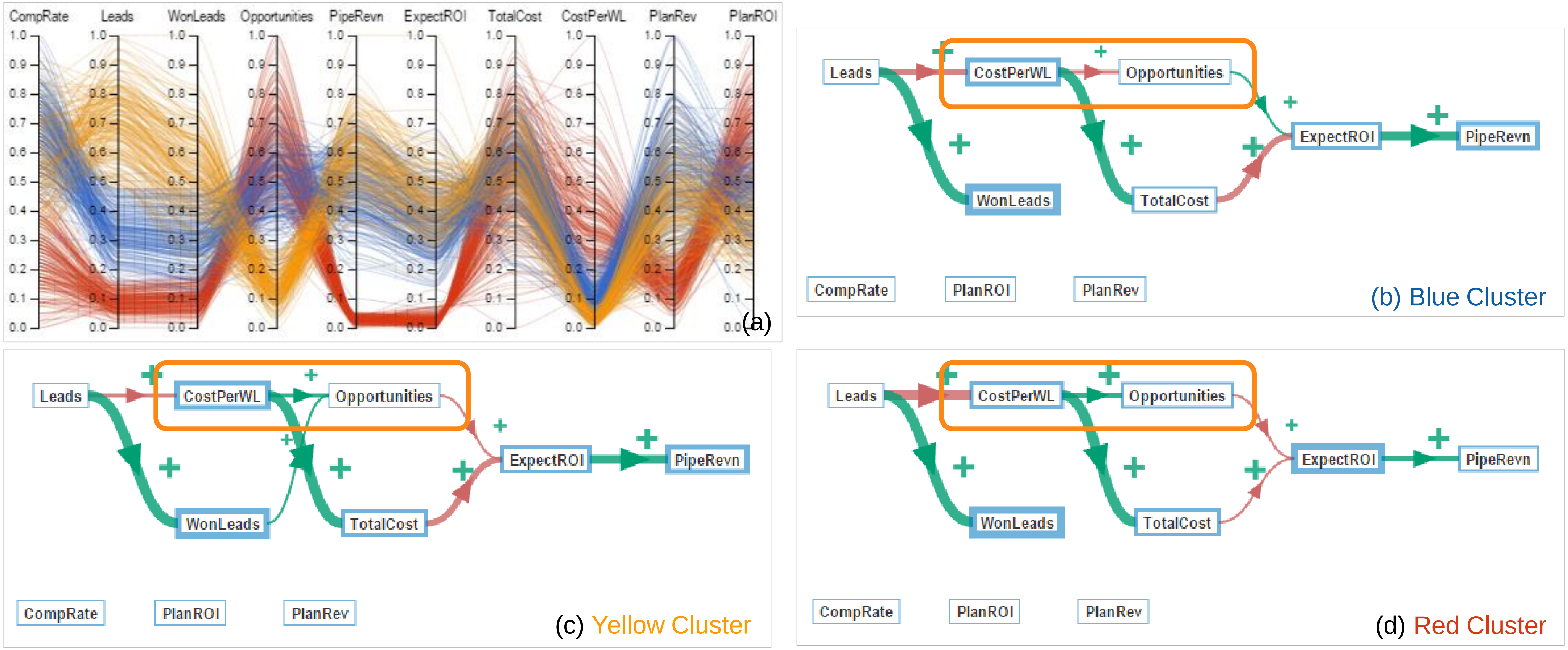
- *Simpson's Paradox*
  - *A relation found in the overall data may not hold in certain subdivisions*
- Subdivide data via the parallel coordinate interface
  - *Manual brushing*
  - *By values of dimensions*
  - *By clustering*

Data Subdivisions

- Example – the Sales Campaign dataset
  - 600 rows, each represents a salesman
  - Attributes - *Leads, WonLeads, Opportunities, CostPerWL, ExpectROI, PipeRevn*



Data Subdivisions – Multiple Models



Analyzing the Sales Campaign Dataset

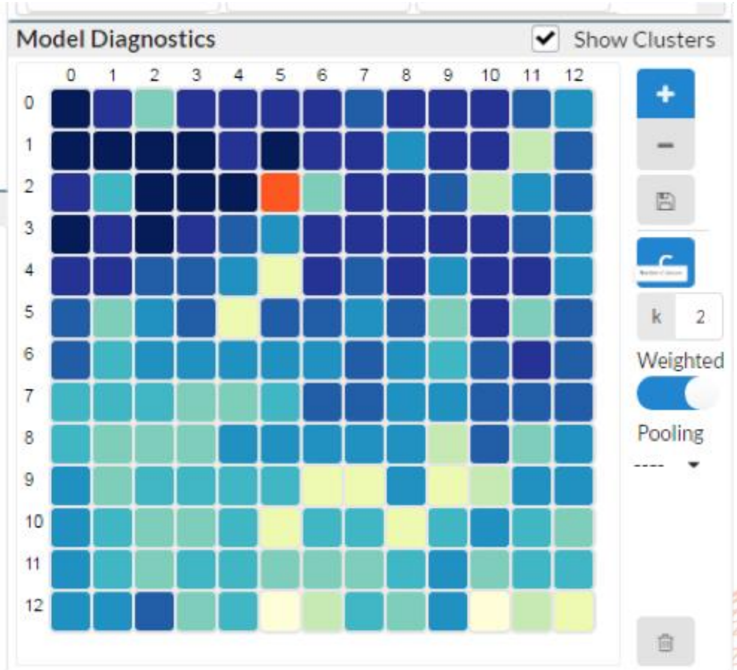
VIS 2017

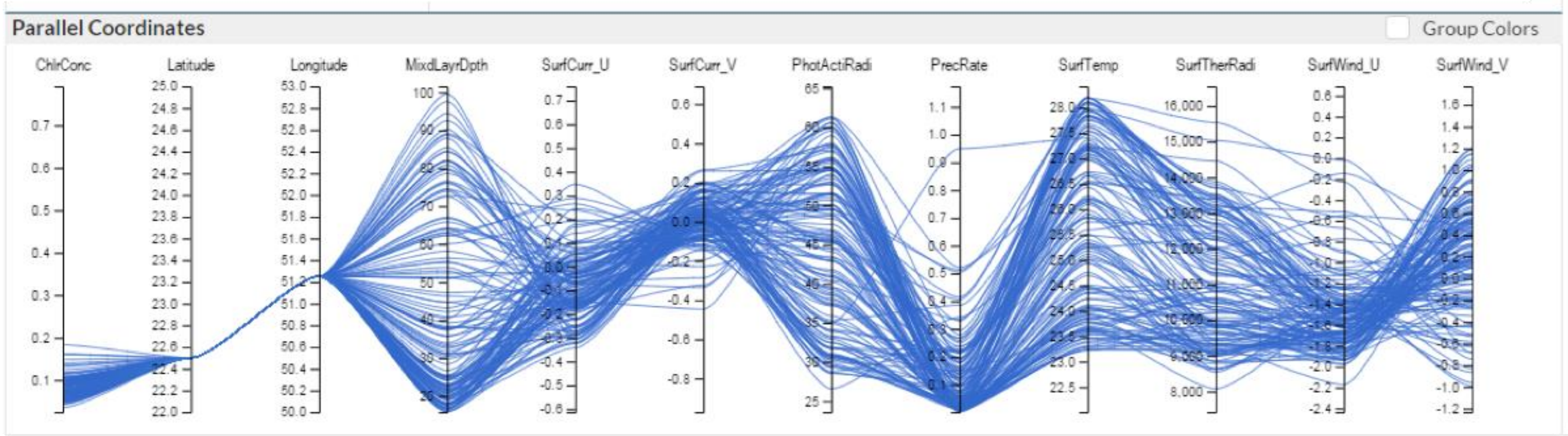
Model Pooling

- Purposes
  - Recognize the possible grouping of causal models
    - *Pooling by clustering*
  - Summarize the common relations from multiple models
    - *Pooling by frequency*
    - *Pooling by credibility*
      - $c_e(e_j) = \frac{\sum_i \delta_{ij}(F_{max} - F_i)}{N(F_{max} - F_{min})}$

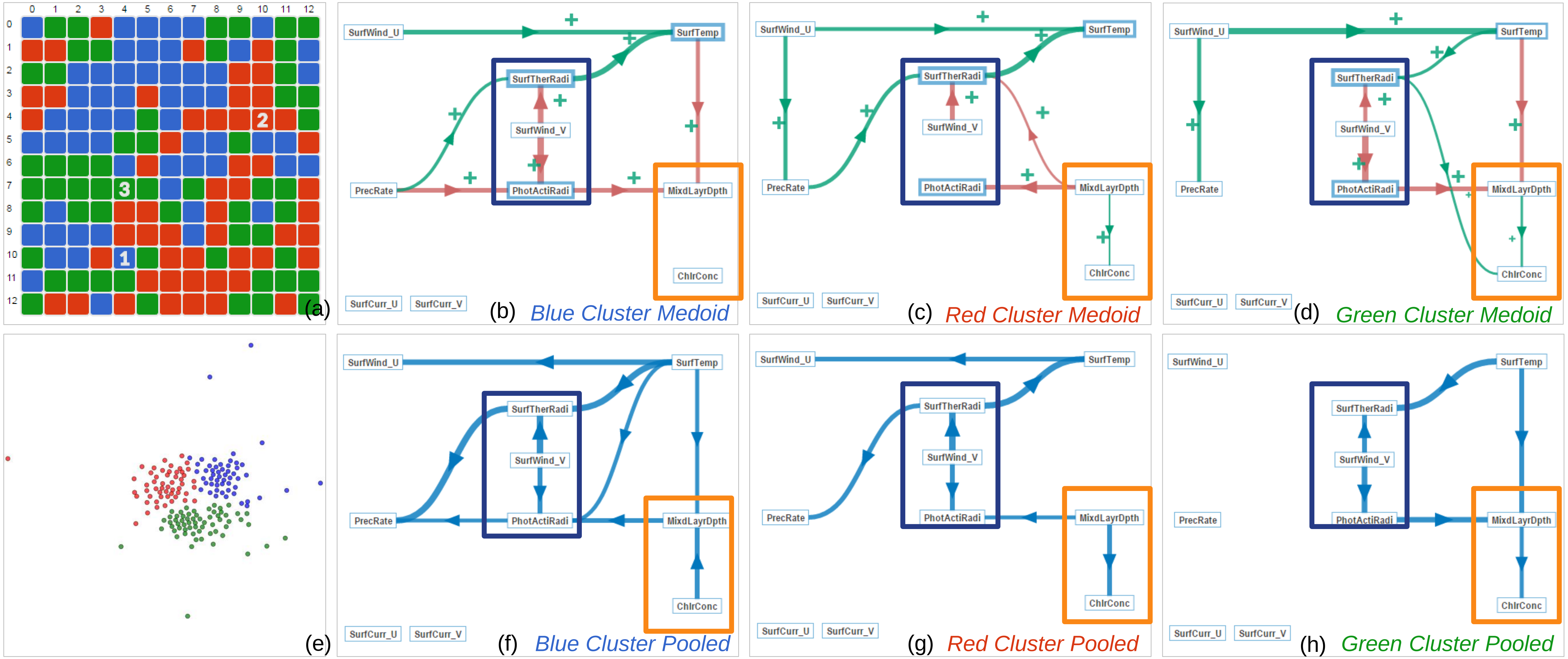
Model Pooling

- Example − the Ocean Chlorophyll dataset
  - Satellite data covers the South Madagascar sea, recording 10 attributes over more than 10 years
  - Rearranged into 13 by 13 (169) geo-locations, each a sub-dataset
  - Derive a model from each sub-dataset





Model Pooling

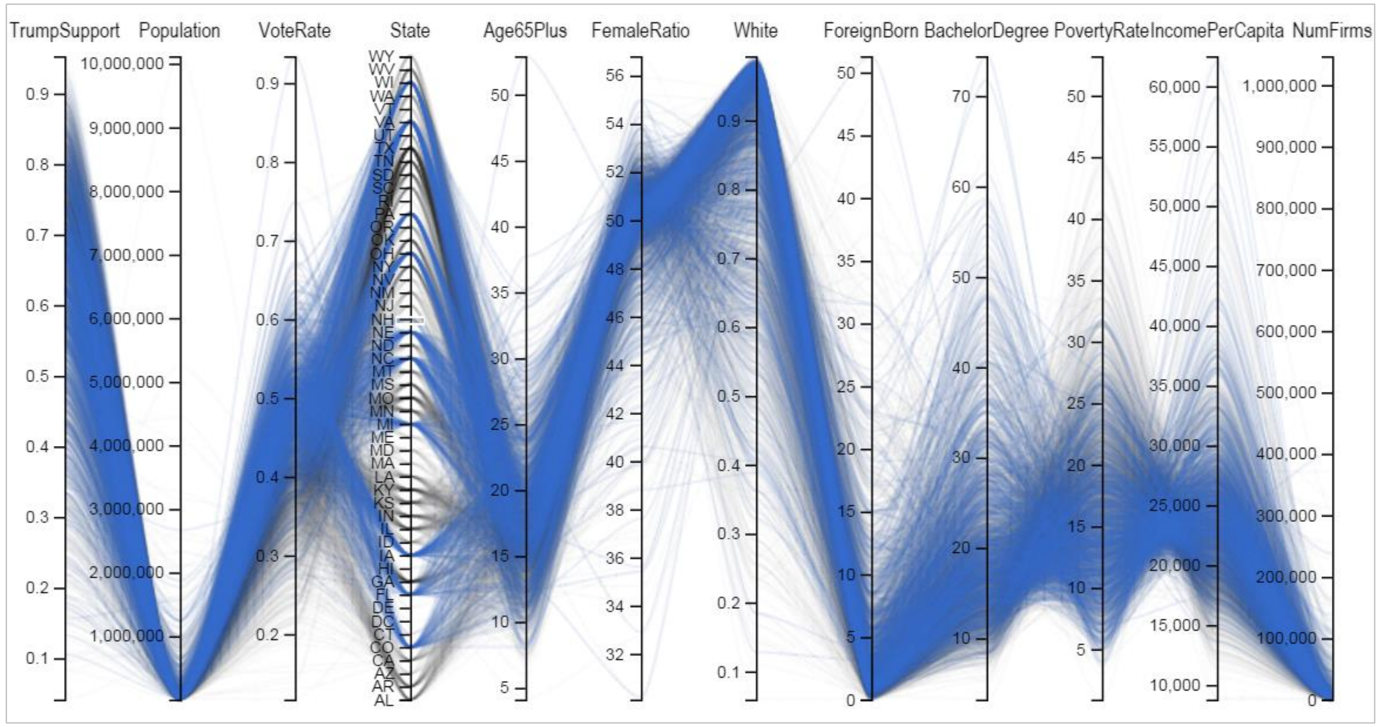


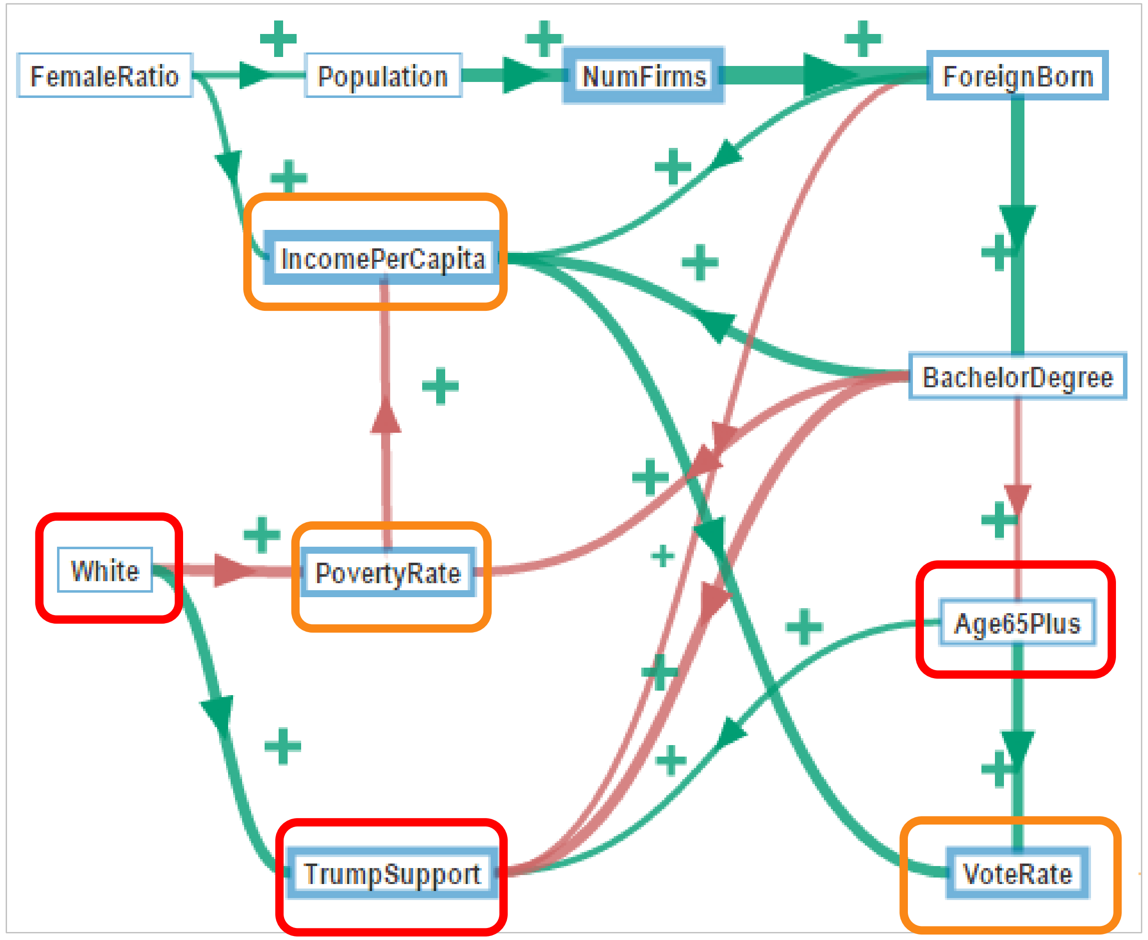
Analyzing the Ocean Chlorophyll Dataset

VIS 2017

One More Use Case

- The presidential election dataset – *county level statistics*





Visual Causality Analysis Made Practical

Jun Wang, Klaus Mueller
{junwang2, mueller}@cs.stonybrook.edu

Thanks for attending!